Recommendations

- Huawei Learning Website
  - http://learning.huawei.com/en
- Huawei e-Learning
  - https://ilearningx.huawei.com/portal/#/portal/ebg/51
- Huawei Certification
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_31&lang=en
- Find Training
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_trainingsearch&lang=en

More Information

- Huawei learning APP

版权所有© 2018 华为技术有限公司

HCIE-Routing & Switching

— LAN Technologies

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- Ethernet is the most widely used local area network (LAN) technology that defines LAN cable types and signal processing methods. As a simple, cost-effective, and easy-to-implement LAN technology, Ethernet has become the mainstream in the industry. The emergence of Gigabit Ethernet and 10G Ethernet further makes Ethernet the most promising network technology.
- This document describes the following LAN concepts: MAC address table, ARP fundamentals, MSTP, iStack and CSS, and link aggregation.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

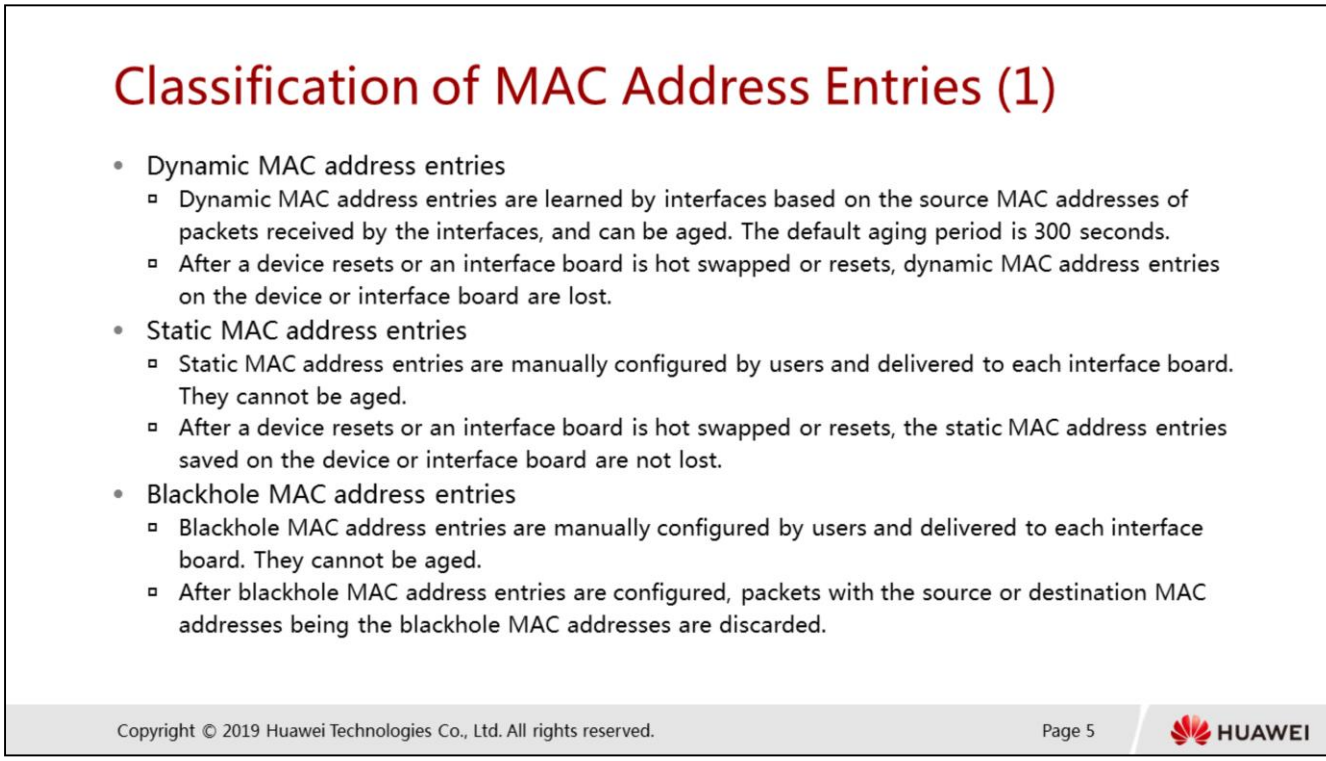
- Upon completion of this course, you will be able to:
  - Describe the classification of MAC address entries and what MAC address flapping is.
  - Describe the fundamentals and application scenarios of gratuitous ARP.
  - Understand MSTP fundamentals.
  - Understand iStack and cluster switch system (CSS).
  - Understand what link aggregation is and how it works.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. **MAC Address Table**
   - Classification of MAC Address Entries
   - Port Security
   - MAC Address Flapping
2. Gratuitous ARP
3. MSTP
4. iStack
5. CSS
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- **Definition of a MAC Address Table**
  - A MAC address table is used by a switch to record the mappings between learned MAC addresses of other devices and interfaces on which MAC addresses are learned, as well as VLANs to which the interfaces belong. When the switch receives a packet, it searches the MAC address table for the MAC address of the destination. If the switch finds the MAC address, it forwards the packet from the corresponding outbound interface in the MAC address entry. Otherwise, the switch broadcasts the packet to all interfaces (except the interface on which the packet was received) in the associated VLAN.
- **MAC address entries are classified into dynamic, static, and blackhole entries.**
  - Dynamic MAC address entries

Dynamic MAC address entries are learned by interfaces based on the source MAC addresses of packets received by the interfaces, and can be aged. After a device resets or an interface board is hot swapped or resets, dynamic MAC address entries on the device or interface board are lost. You can check whether data is forwarded between two connected devices by checking dynamic MAC address entries. You can obtain the number of communicating users connected to an interface by checking the number of specified dynamic MAC address entries.

  - Static MAC address entries

Static MAC address entries are manually configured by users and delivered to each interface board. They cannot be aged. After a device resets or an interface board is hot swapped or resets, the static MAC address entries saved on the device or interface board are not lost. A static MAC address entry can have only one outbound interface. Statically binding an interface to a MAC address does not affect the learning of dynamic MAC address entries on the interface. A static MAC address entry can be configured to allow an authorized user to access network resources and prevent other users from using the bound MAC address to initiate attacks.

  - Blackhole MAC address entries

Blackhole MAC address entries are manually configured by users and delivered to each interface

board. They cannot be aged. After a device resets or an interface board is hot swapped or resets, the static MAC address entries saved on the device or interface board are not lost. A blackhole MAC address entry can be configured to filter out unauthorized users.

Classification of MAC Address Entries (2)

```
[Huawei]dis mac-address
MAC address table of slot 0:
-------------------------------------------------------------------------------
MAC Address    VLAN/       PEVLAN CEVLAN Port            Type      LSP/LSR-ID
               VSI/SI                                              MAC-Tunnel
-------------------------------------------------------------------------------
00aa-bbcc-ddee -           -      -      -               blackhole -
0011-2233-4455 1           -      -      GE0/0/2         static    -
-------------------------------------------------------------------------------
Total matching items on slot 0 displayed = 2

MAC address table of slot 0:
-------------------------------------------------------------------------------
MAC Address    VLAN/       PEVLAN CEVLAN Port            Type      LSP/LSR-ID
               VSI/SI                                              MAC-Tunnel
-------------------------------------------------------------------------------
5489-986d-53e0 1           -      -      GE0/0/16        dynamic   0/-
5489-98fc-608b 1           -      -      GE0/0/1         dynamic   0/-
-------------------------------------------------------------------------------
Total matching items on slot 0 displayed = 2
```

- By default, the aging time of dynamic MAC address entries is 300 seconds.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- MAC address entries of a device can be checked using the **display mac-address** command. As shown in the preceding figure, MAC entries are classified into dynamic, static, and blackhole entries. The entries also show the VLAN and VSI information corresponding to MAC addresses.

Configuring MAC Address Entries

- Configure a static MAC address entry.

```
[Huawei]mac-address static 0011-2233-4455 GigabitEthernet 0/0/2 vlan  1
```

- Configure a blackhole MAC address entry.

```
[Huawei]mac-address blackhole 00aa-bbcc-ddee
```

- Set an aging time for a dynamic MAC address entry.

```
[Huawei]mac-address aging-time 400
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



Contents

1. **MAC Address Table**
   - Classification of MAC Address Entries
   - Port Security
   - MAC Address Flapping
2. Gratuitous ARP
3. MSTP
4. iStack
5. CSS
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Port Security

- Port security converts the dynamic MAC addresses learned on an interface into secure MAC addresses (including secure dynamic and static MAC addresses, and sticky MAC addresses). This function prevents unauthorized users from using this interface to communicate with the switch and therefore enhances device security.
- Classification of secure MAC addresses
  - Secure dynamic MAC address
    - MAC address that is converted on an interface with port security enabled but sticky MAC disabled
  - Secure static MAC address
    - MAC address that is manually configured on an interface with port security enabled
  - Sticky MAC address
    - MAC address that is converted on an interface with both port security and sticky MAC enabled

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- After port security is enabled on an interface, dynamic MAC address entries that have been learned on the interface are deleted. Subsequent MAC address entries are converted into secure dynamic MAC address entries.
- After the sticky MAC function is enabled on an interface, existing secure dynamic MAC address entries and subsequent MAC address entries are converted into sticky MAC address entries.
- After port security is disabled on an interface, the secure dynamic MAC address entries on the interface are deleted, and dynamic MAC address entries are re-learned.
- After the sticky MAC function is disabled on an interface, sticky MAC address entries on the interface are converted into secure dynamic MAC address entries.

Configuring Port Security

- Configure the secure MAC address function.

```
[Huawei-GigabitEthernet0/0/2]port-security enable
\\Enables the port security function.
[Huawei-GigabitEthernet0/0/2]port-security protect-action shutdown
\\Configures a protection action for port security.
[Huawei-GigabitEthernet0/0/2]port-security max-mac-num 5
\\Configures the maximum number of secure dynamic MAC addresses that can be learned by the interface.
[Huawei-GigabitEthernet0/0/2]port-security aging-time 1000
\\Configures an aging time for the dynamic secure MAC address entries learned by the interface.
```

- Configure the sticky MAC address function.

```
[Huawei-GigabitEthernet0/0/2]port-security enable
[Huawei-GigabitEthernet0/0/2]port-security mac-address sticky
\\Enables the sticky MAC function on the interface.
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Description
  - After the sticky MAC function is enabled on an interface, existing secure dynamic MAC address entries and subsequent MAC address entries are converted into sticky MAC address entries.
  - After the sticky MAC function is enabled on an interface, sticky MAC address entries are not aged even if the **port-security aging-time** command is run.
  - The saved sticky MAC address entries are not lost after a device restart.
- Port security actions
  - Restrict: Discards the packets whose source MAC addresses do not exist and reports an alarm. This action is recommended.
  - Protect: Discards the packets whose source MAC addresses do not exist but does not report alarms.
  - Shutdown: Sets the interface status to error-down and reports an alarm.

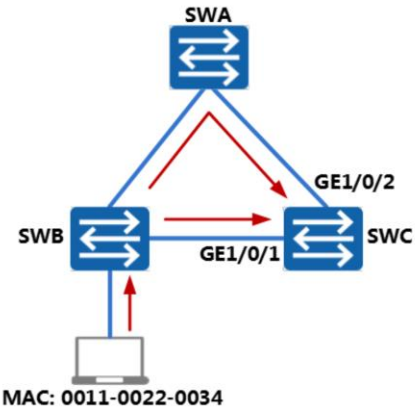
Contents

1. **MAC Address Table**
   - Classification of MAC Address Entries
   - Port Security
   - MAC Address Flapping
2. Gratuitous ARP
3. MSTP
4. iStack
5. CSS
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MAC Address Flapping

- MAC address flapping occurs when a MAC address is learned by two interfaces in the same VLAN and the MAC address entry learned later overrides the earlier one.
- MAC address flapping avoidance mechanism:
  - Increases the MAC address learning priority of an interface.
  - Prohibits MAC address entry overriding between interfaces with the same priority.



A MAC address entry is learned, with the outbound interface being GE 1/0/1.

MAC Address	Port
0011-0022-0034	GE1/0/1

The outbound interface of the MAC address entry is updated to GE 1/0/2, indicating that address flapping occurs.

MAC Address	Port
0011-0022-0034	GE1/0/2

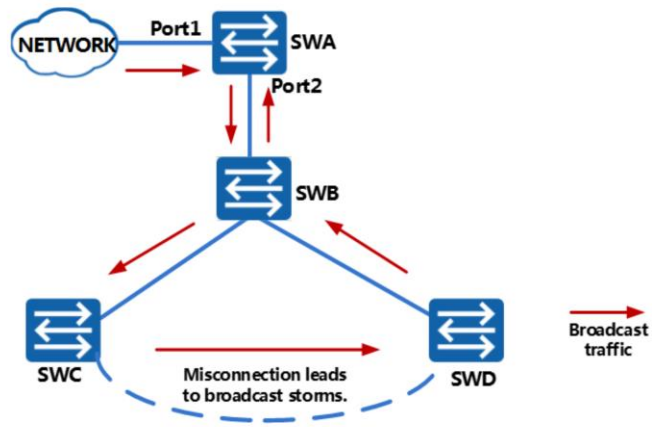
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In the preceding figure, in the MAC address entry with MAC address 0011-0022-0034, the outbound interface is changed from GE1/0/1 to GE1/0/2. This is MAC address flapping. MAC address flapping can cause an increase in the CPU usage on the switch. Generally, MAC address flapping does not occur unless a network loop occurs. If frequent MAC address flapping occurs on your network, alarms and MAC address flapping records provide insight for locating faults and eliminating loops.
- MAC address flapping occurs on a network when loops or attacks occur. During network planning, you can use the following methods to prevent MAC address flapping:
  - Increase the MAC address learning priority of an interface: If the same MAC address is learned on interfaces that have different priorities, the MAC address entry on the interface with the highest priority overrides that on the other interfaces.
  - Prevent MAC address entries from being overridden on interfaces with the same priority: If the interface connected to a bogus network device has the same priority as the interface connected to an authorized device, the MAC address entry of the bogus device learned later does not override the original correct MAC address entry. If the authorized device is powered off, the MAC address entry of the bogus device is learned. After the authorized device is powered on again, its MAC address cannot be learned.

MAC Address Flapping Detection

- MAC address flapping detection determines whether MAC address flapping occurs by checking whether outbound interfaces in MAC address entries change frequently.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- After MAC address flapping detection is enabled, the switch reports an alarm if MAC address flapping occurs (for example, due to a loop between the outbound interfaces). The alarm contains the flapping MAC address, VLAN ID, and outbound interfaces between which the MAC address flaps. The network administrator can locate the cause of the loop based on the alarm. As an alternative, the switch can perform the action specified in the configuration of MAC address flapping detection to remove the loop automatically. The action can be quit-vlan (remove the interface from the VLAN) or error-down (shut down the interface).
- In the preceding figure, a network cable is incorrectly connected between SwitchC and SwitchD, causing a loop between SwitchB, SwitchC, and SwitchD. When Port1 of SwitchA receives a broadcast packet, SwitchA forwards the packet to SwitchB. The packet is then sent to Port2 of SwitchA. After being configured with MAC address flapping detection, SwitchA can detect that the source MAC address of the packet flaps from Port1 to Port2. If the MAC address flaps between Port1 and Port2 frequently, SwitchA reports an alarm.

Configuring MAC Address Anti-flapping

- Configure a MAC address learning priority for an interface.

```
[Huawei-GigabitEthernet0/0/2]mac-learning priority 3
\\Configures a MAC address learning priority for an interface. By default, the priority of MAC address learning on an interface is 0. A larger value indicates a higher priority.
```

- Configure a device to prohibit MAC address flapping between interfaces with the

```
[Huawei]undo mac-learning priority 3 allow-flapping
\\Configures a device to prohibit MAC address flapping between interfaces with the same priority.
```

- Configure global MAC address flapping detection.

```
[Huawei]mac-address flapping detection
\\Configures global MAC address flapping detection.
```

- Configure VLAN-based MAC address flapping detection.

```
[Huawei]vlan 2
[Huawei-vlan2]loop-detect eth-loop block-time 100 retry-times 3
\\Configures VLAN-based MAC address flapping detection.
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- After different MAC address learning priorities are configured for interfaces, when two interfaces learn the same MAC address entry, the MAC address entry learned by the interface with a higher priority overrides that learned by the other interface to prevent MAC address flapping.
- Configuring a device to prohibit MAC address flapping between interfaces with the same priority also prevents MAC address flapping and improve network security.



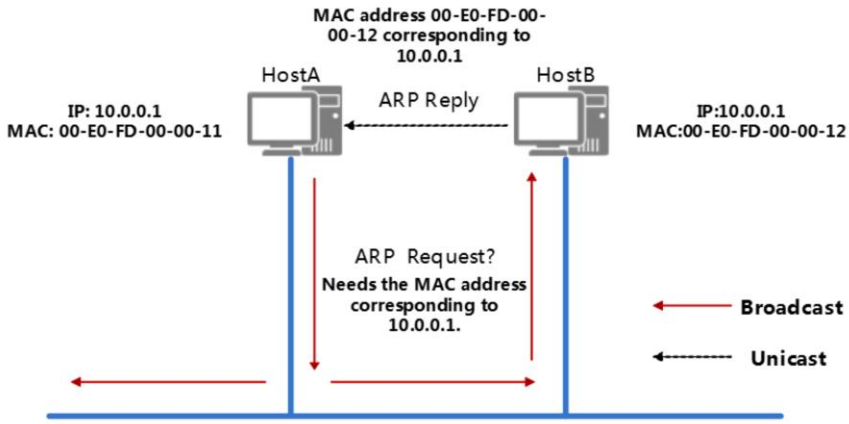
Contents

1. MAC Address Table
2. **Gratuitous ARP**
3. MSTP
4. iStack
5. CSS
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Gratuitous ARP

- Gratuitous ARP allows a device to send an ARP Request message that carries its own IP address as the destination IP address.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Gratuitous ARP has the following functions:
  - Checks for IP address conflicts.

    When the protocol status of a device interface changes to Up, the device broadcasts gratuitous ARP packets. If the device receives an ARP reply, another device is using the same IP address. When detecting an IP address conflict, the device periodically broadcasts gratuitous ARP Reply packets until the conflict is removed.
  - Advertises a new MAC address.

    If the MAC address of a device is changed because its network adapter is replaced, the device sends a gratuitous ARP packet to notify all devices of the change before the ARP entry is aged out.
  - Notifies other devices of a master/backup switchover in a VRRP group.

    After a master/backup switchover, the master device broadcasts a gratuitous ARP packet in the VRRP group to notify other devices of the switchover.



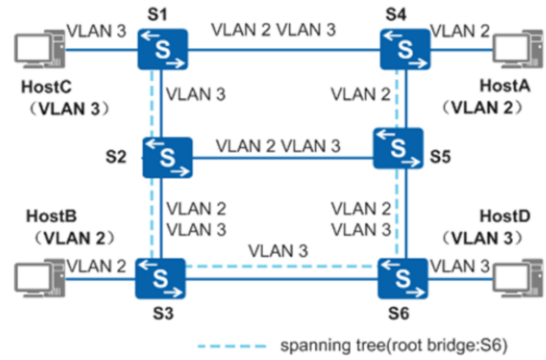
Contents

1. MAC Address Table
2. Gratuitous ARP
3. **MSTP**
   - STP/RSTP Review
   - MSTP Fundamentals
4. iStack
5. CSS
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

STP/RSTP Defect

- STP and RSTP both have a defect: All VLANs on a LAN share one spanning tree. As a result, inter-VLAN load balancing cannot be performed, and blocked links cannot transmit any traffic, which may lead to VLAN packet transmission failures.



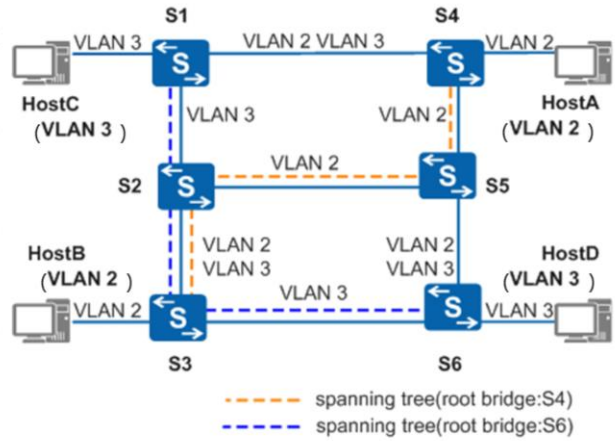
- In the preceding figure, the spanning tree is marked in dotted lines, with S6 being the root bridge. The links between S2 and S5 and between S1 and S4 are blocked. HostA and HostB both belong to VLAN 2. Because the link between S2 and S5 is blocked and the link between S3 and S6 denies packets from VLAN 2, HostA and HostB cannot communicate with each other.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- All VLANs share one spanning tree for STP and RSTP. As a result, some VLANs cannot communicate with each other, sub-optimal paths have to be used, and traffic cannot be balanced.
- To fix this defect, the IEEE released the 802.1s standard that defines MSTP in 2002. Compatible with STP and RSTP, MSTP can rapidly converge traffic and provides multiple paths to balance VLAN traffic.

MSTP Improvements Over STP and RSTP

- MSTP divides a switching network into multiple regions, each of which has multiple spanning trees that are independent of each other.
- Each spanning tree is called a multiple spanning tree instance (MSTI), and each region is called a multiple spanning tree (MST) region.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An MSTI is a collection of VLANs. Binding multiple VLANs to a single MSTI reduces communication costs and resource usage. The topology of each MSTI is calculated independently, and traffic can be balanced among MSTIs. Multiple VLANs with the same topology can be mapped to a single MSTI. The forwarding state of the VLANs for an interface is determined by the interface state in the MSTI.
- In the preceding figure, MSTP associates VLANs and MSTIs by mapping VLANs to MSTIs. Each VLAN can be mapped to only one MSTI. This means that traffic of a VLAN can be transmitted in only one MSTI. An MSTI, however, can correspond to multiple VLANs.
- After calculation, two MSTIs are generated:
  - MSTI 1 uses S4 as the root bridge to forward packets of VLAN 2.
  - MSTI 2 uses S6 as the root bridge to forward packets of VLAN 3.
- Devices within the same VLAN can then communicate with each other, and packets of different VLANs are then balanced along different paths.

Contents

1. MAC Address Table
2. Gratuitous ARP
3. **MSTP**
   - STP/RSTP Review
   - MSTP Fundamentals
4. iStack
5. CSS
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Basic MSTP Concepts

- MST region
- MSTI
- VLAN mapping table
- CST, IST, SST, CIST
- Regional root
- CIST root
- Master bridge
- Port role
- Port status

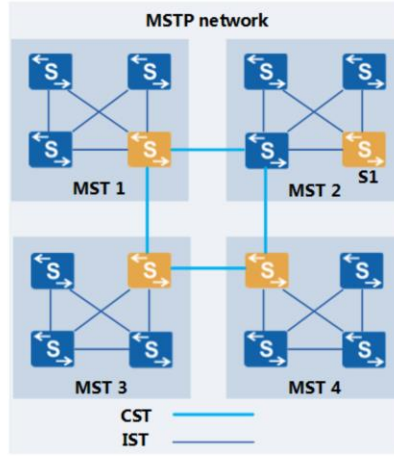


Figure 1 MSTP network

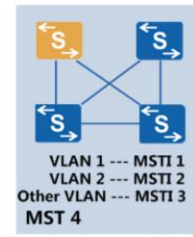


Figure 2 Mapping between VLANs and MSTIs

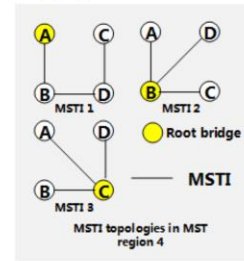


Figure 3 Independent spanning tree calculation within each MSTI

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An MST region contains multiple switches and their network segments. An MSTI is an instance in an MST region. An MST region can have multiple MSTIs.
- A VLAN mapping table describes the mappings between VLANs and MSTIs. As shown in Figure 2, in MST region 4, VLAN 1 is mapped to MSTI 1, VLAN 2 is mapped to MSTI 2, and other VLANs are mapped to MSTI 3.
- A common spanning tree (CST) connects all MST regions on a switching network. If each MST region is considered as a single node, the CST is a spanning tree calculated using STP or RSTP.
- An internal spanning tree (IST) resides within an MST region. An IST is a special MSTI with an ID of 0.
- A single spanning tree (SST) is formed when a switch running STP or RSTP belongs to only one spanning tree or an MST region has only one switch.
- The ISTs of all MST regions plus the CST form a complete spanning tree, that is, the CIST.
- Regional roots are classified into internal spanning tree (IST) and MSTI regional roots.
  - In Figure 1, the switches that are closest to the CIST root are IST regional roots.
  - An MST region can contain multiple spanning trees, each of which is called an MSTI. An MSTI regional root is the root of the MSTI. In Figure 3, each MSTI has its own regional root.
- The CIST root is the root bridge of the CIST. S1 in Figure 1 is the CIST root.
- Master bridges, also called IST masters, are the switches nearest to the CIST root. Orange switches in Figure 1 are master bridges. If the CIST root is in an MST region, the CIST root is the master bridge of the region.
- Port role: Similar to RSTP, MSTP defines the root port, designated port, alternate port, backup port, and edge port.
- Port status: Similar to RSTP, MSTP defines port status of forwarding, learning, and discarding.

MSTP Topology Calculation

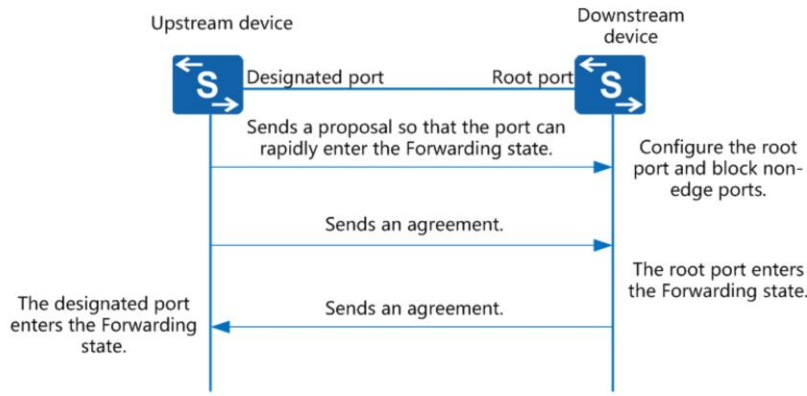
- CIST calculation
  - After comparing BPDUs, the switch with the highest priority on the entire network is selected as the CIST root. MSTP calculates an IST for each MST region, treats each MST region as a single device, and calculates a CST to interconnect MST regions. The CST and ISTs form a CIST for the entire network.
- MSTI calculation
  - In an MST region, MSTP independently calculates an MSTI for each VLAN based on mappings between VLANs and MSTIs. The calculation process is similar to that used by STP to calculate a spanning tree.
- MSTP responding to topology changes
  - MSTP responds to topology changes in a way similar to the way RSTP responds to topology changes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- MSTI characteristics
  - A spanning tree is calculated independently for each MSTI in
  - a way similar to the way STP calculates a spanning tree.
  - Spanning trees of MSTIs can have different roots and topologies.
  - Each MSTI sends BPDUs on its own spanning tree.
  - The topology of each MSTI is determined using commands.
  - The spanning tree parameters can be different on a port for different MSTIs.
  - A port can play different roles or have different status in different MSTIs.
- On an MSTP network, a VLAN packet is forwarded as follows:
  - Along MSTI in an MST region
  - Along CST among MST regions

Fast MSTP Convergence

- MSTP supports both ordinary and enhanced Proposal/Agreement (P/A) mechanisms.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- As shown in the preceding figure, the P/A mechanism works as follows for MSTP:
  1. The upstream device sends a proposal BPDU to the downstream device, requesting the port connecting to the downstream device to rapidly enter the Forwarding state. After receiving this BPDU, the downstream device sets its port connected to the upstream device as the root port and blocks all non-edge ports.
  2. The upstream device sends an agreement BPDU. After receiving this BPDU, the root port on the downstream enters the Forwarding state.
  3. The downstream device replies with an agreement BPDU. After receiving this BPDU, the upstream device sets its port connected to the downstream device as the designated port, and the port then enters the Forwarding state.
- By default, Huawei switches use fast transition in enhanced P/A. To enable a Huawei switch to communicate with a third-party device that uses fast transition in common P/A, configure the Huawei switch to use ordinary P/A.

Configuring MSTP

- Configure basic MSTP functions.

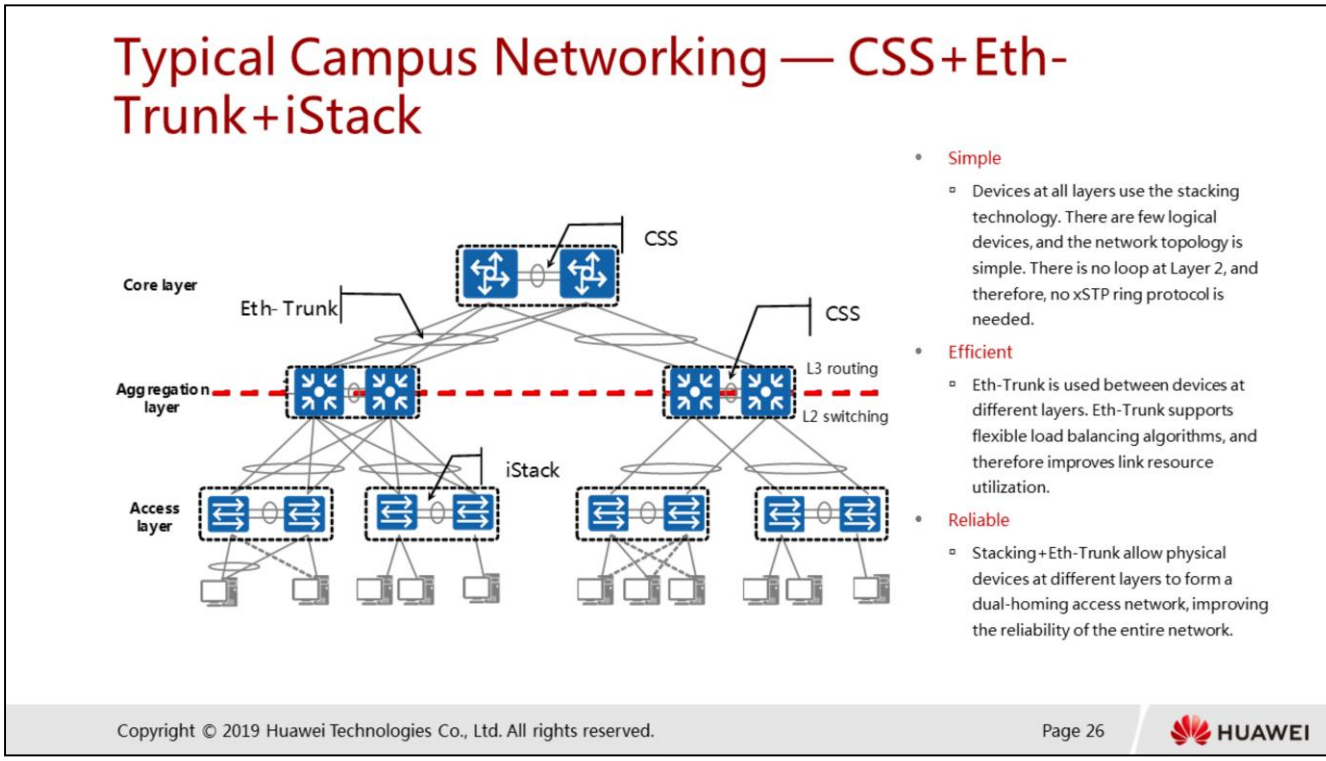
```
[SwitchA] stp region-configuration
\\Enters the MST region view.
[SwitchA-mst-region] region-name RG1
\\Configures a name for the MST region.
[SwitchA-mst-region] instance 1 vlan 2 to 10
\\Configures mappings between MSTIs and VLANs.
[SwitchA-mst-region] instance 2 vlan 11 to 20
[SwitchA-mst-region] active region-configuration
\\Activates the MST region configurations to make the domain name, VLAN mapping table, and
MSTP revision level take effect.
[SwitchA] stp instance 1 root primary
\\Configures the device as the root bridge for MSTI 1.
[SwitchA] stp instance 2 root secondary
\\Configures the device as the secondary root bridge for MSTI 2.
[SwitchA] stp pathcost-standard legacy
# Configures SwitchA to use Huawei's proprietary algorithm to calculate the path cost.
[SwitchA] stp enable
\\Enables MSTP on SwitchA.
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. **MAC Address Table**
2. Gratuitous ARP
3. MSTP
4. **iStack**
   - Typical Campus Networking
   - iStack Fundamentals
5. CSS
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- The preceding figure shows a CSS+iStack campus network, which is simple, efficient, and highly reliable.
- Simple
  - Devices at all layers use the stacking technology. There are few logical devices, and the network topology is simple. There is no loop at Layer 2, and therefore, no xSTP ring protocol is needed.
- Efficient
  - Eth-Trunk is used between devices at different layers. Eth-Trunk supports flexible load balancing algorithms, and therefore improves link resource utilization.
- Reliable
  - Servers and hosts can be configured with multi-NIC Teaming-based load balancing or active/standby redundancy links, improving server access reliability.
  - Stacking+Eth-Trunk allow physical devices at different layers to form a dual-homing access network, improving the reliability of the entire network.
- Disadvantages
  - There are high requirements on device performance. If a large number of fixed-configuration switches are stacked, the performance of the master switch's main control board may deteriorate.
  - If service ports are used for stacking or CSS, service port resources are occupied.

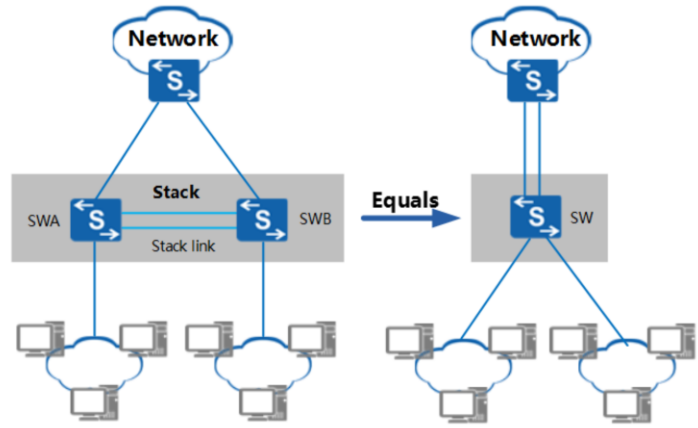
Contents

1. MAC Address Table
2. Gratuitous ARP
3. MSTP
4. **iStack**
   - Typical Campus Networking
   - iStack Fundamentals
5. CSS
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Device Stacking — iStack

- Intelligent stack (iStack) enables multiple stacking-capable switches to function as a single logical switch. In the right figure, after SwitchA and SwitchB are connected by stack cables and form a stack, they are considered as one switch by their upstream and downstream devices.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

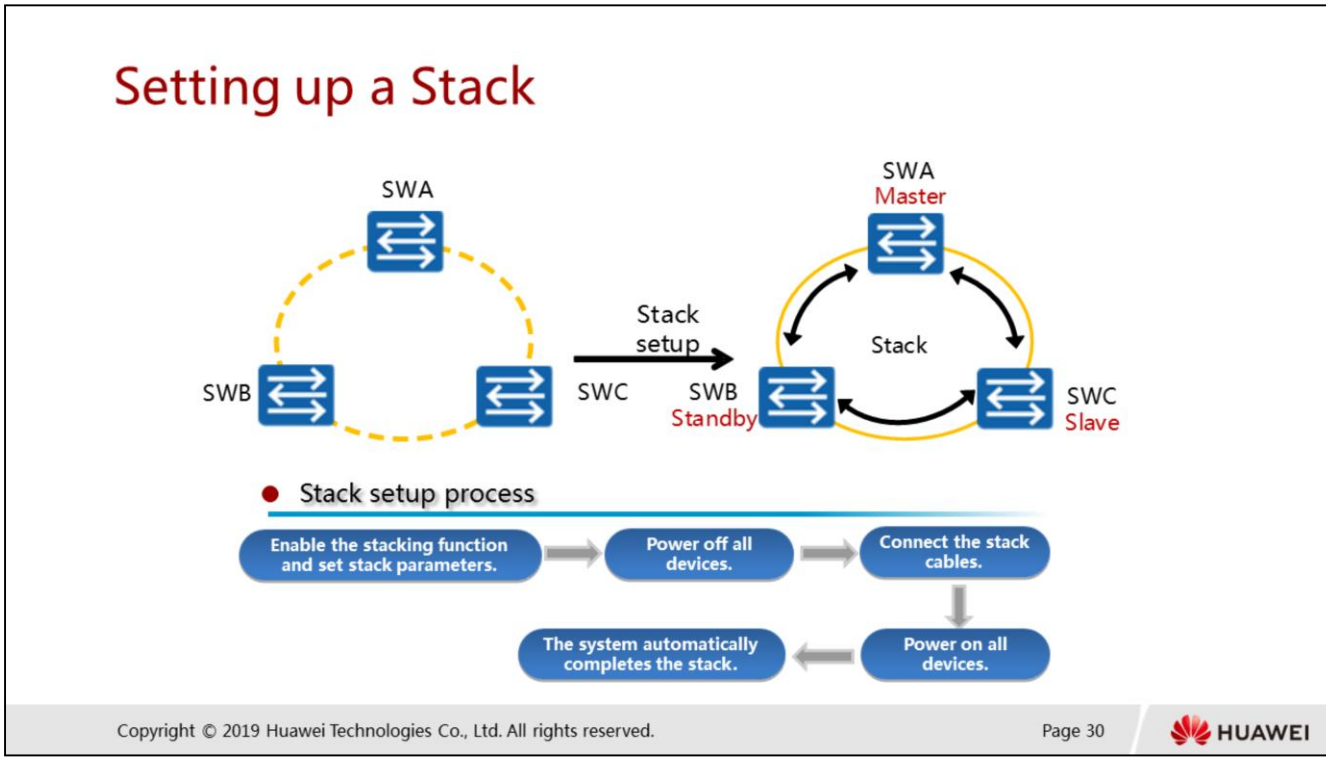
HUAWEI

- Stacking improves network reliability and scalability while simplifying network management.
  - High reliability. Member switches in a stack work in redundancy mode. Link redundancy can also be implemented between member switches through inter-device link aggregation.
  - High scalability. The number of interfaces, bandwidth, and processing capacity of a stack can be increased by adding switches to the stack. A switch can join or leave the stack without affecting other member switches. New switches automatically synchronize their configuration files and system software versions with the master switch.
  - Simplified management and configuration. A stack can be logged in from any member switch to centrally manage and configure all member switches in the stack. In addition, complicated Layer 2 loop prevention protocols or Layer 3 protection switching protocols are not required after switches set up a stack, simplifying network configuration.

Basic iStack Concepts

- Roles
  - Switches that join a stack are member switches. Member switches in a stack can play one of the following roles:
    - Master: A master switch manages the entire stack. A stack has only one master switch.
    - Standby: A standby switch acts as the backup of the master switch. If the master switch fails, the standby switch takes over all services from the master switch and assumes the master role. A stack has only one standby switch.
    - Slave: A slave switch forwards service traffic. A high number of slave switches can increase forwarding performance of a stack. Apart from the master and standby switches, all the other member switches are slave switches.
- Stack ID
  - A stack ID identifies a stack member and is the slot ID of the member. Each stack member must have a unique stack ID in the stack.
- Stack priority
  - A stack priority determines the role of a stack member in role election. A higher priority value for a stack member increases the probability of it being elected the stack master.

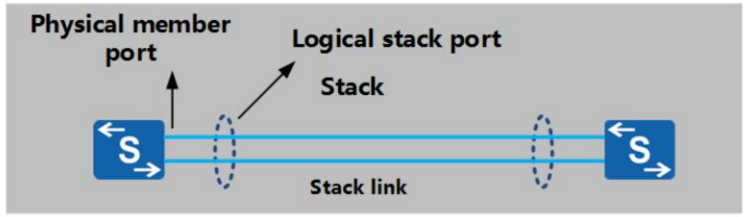
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- The last step "The system automatically completes the stack" entails the following three sub-steps:
  1. Master switch election
     ① Compares the switches' running status. The switch that has been running is preferentially elected the master switch over the switch that is starting.
     ② If multiple switches complete startup at the same time, the switch with the highest stack priority becomes the master switch.
     ③ If multiple switches complete startup at the same time and have the same stack priority, the switch with the smallest MAC address becomes the master switch.
  2. Topology information collection and standby switch election
     - After the master switch is elected, it collects topology information from all the other member switches, calculates forwarding entries and to-be-blocked interfaces, sends the calculated information to these member switches, and assigns them stack IDs. A standby switch is then elected as the backup of the master switch. The election rules are as follows: When all the switches except the master switch complete startup at the same time:
       - The switch with the highest stack priority becomes the standby switch.
       - If the switches have the same stack priority, the one with the smallest MAC address becomes the standby switch.
  3. Stable running
- After role election and topology collection are complete, other switches are added to the stack as slave switches. They automatically synchronize their system software versions and configuration files with the master switch.
  - A stack supports automatic software version loading. To-be-stacked switches can run different software versions and can set up a stack if these versions are compatible with one another. If the software version running on the master switch is different from those on the standby switch and slave switches, the standby and slave switches automatically download the system software from the master switch, restart with the new system software, and then rejoin the stack.
  - A stack supports configuration file synchronization. The standby and slave switches download and apply the configuration file of the master switch to ensure that all the member switches work like a single device and the rest switches continue working properly if the master switch fails.

Stack Modes

- Switches can set up a stack using stack cards or service ports.
  - Stack cards can be used in either of the following methods:
    - Stack cards are used to connect switches through stack cables.
    - Stack cards are integrated into the rear panels of switches, and switches are connected using integrated stack ports and stack cables.
  - When service ports are used to set up a stack, stack cards are not required. The service ports are physical ports that are configured as stack member ports and bundled into a logical stack port, as shown in the following figure. Switches are connected through the physical ports.

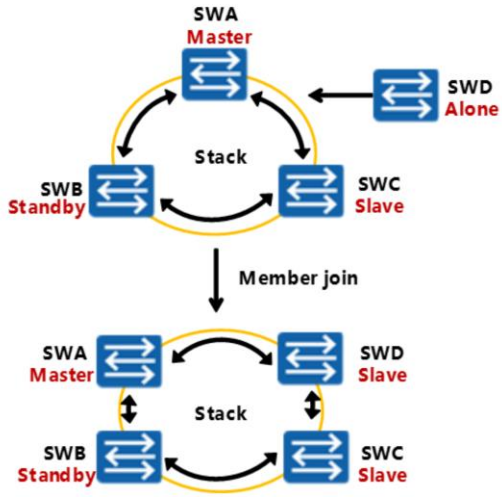


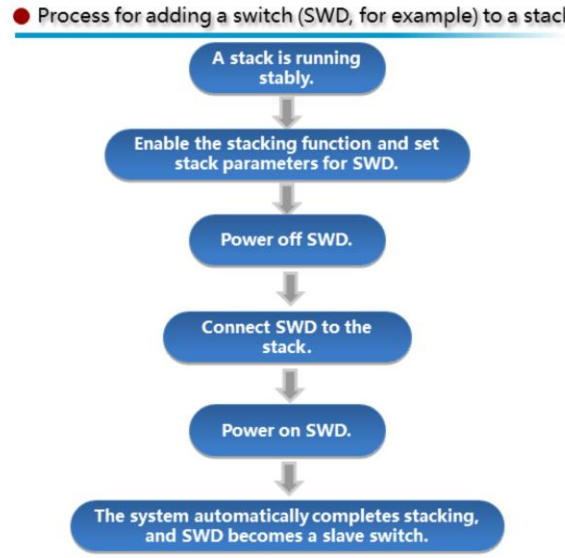
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Physical member port
  - A physical member port is a service port used to connect stack member switches. Physical member ports forward service packets or stack protocol packets between member switches.
- Logical stack port
  - A logical stack port is exclusively used for stacking and has the physical member ports bundled. Each member switch in a stack supports two stack ports: stack-port *n*/1 and stack-port *n*/2, where *n* is the stack ID of the member switch.
- Service port connections are classified into ordinary and dedicated cable connections based on cable types.
  - Ordinary cable connection
    - Ordinary stack cables include optical cables, network cables, and high-speed cables. When ordinary stack cables are used to set up a stack, logical stack ports must be manually configured. Otherwise, the stack cannot be set up.
  - Dedicated cable connection
    - A dedicate stack cable has two ends: master end with the Master tag and slave end without any tag. Switches can automatically set up a stack after dedicated cables are connected to ports according to connection rules.

Adding a Switch to a Stack



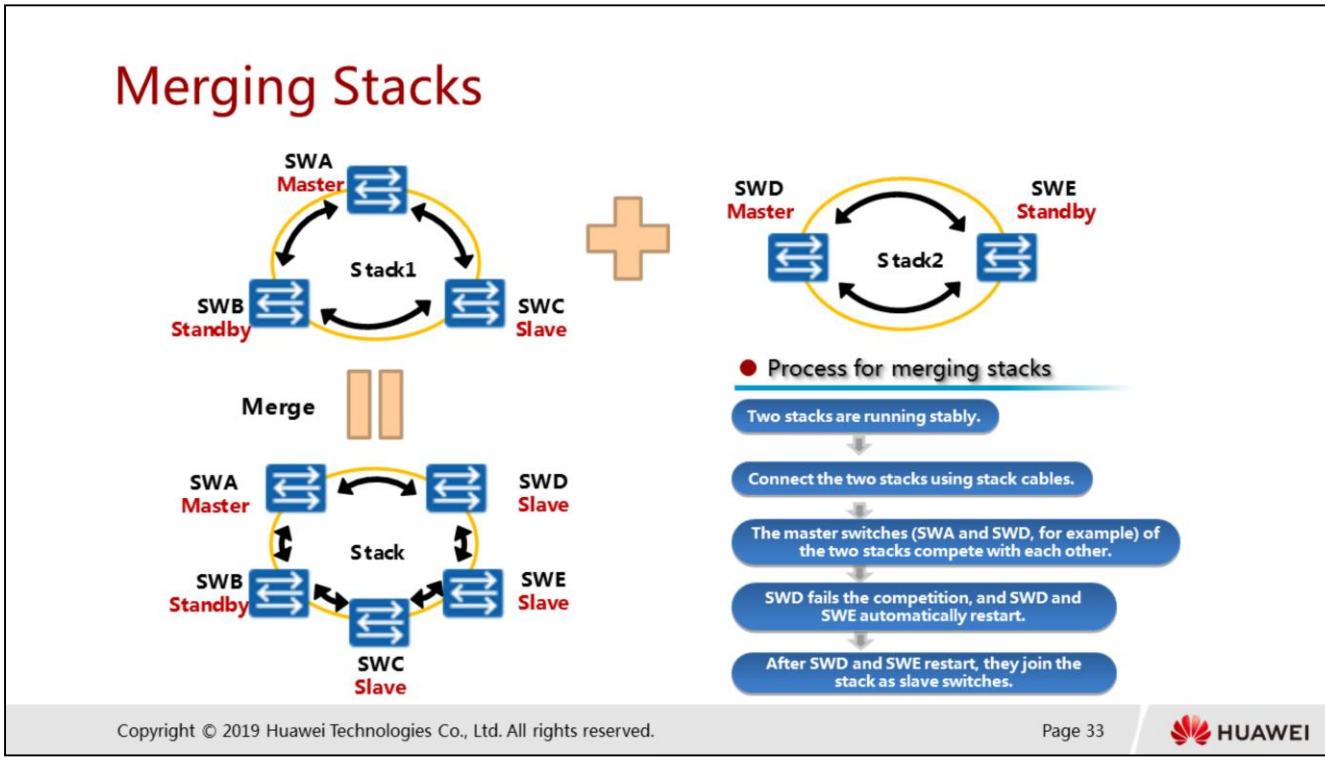


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

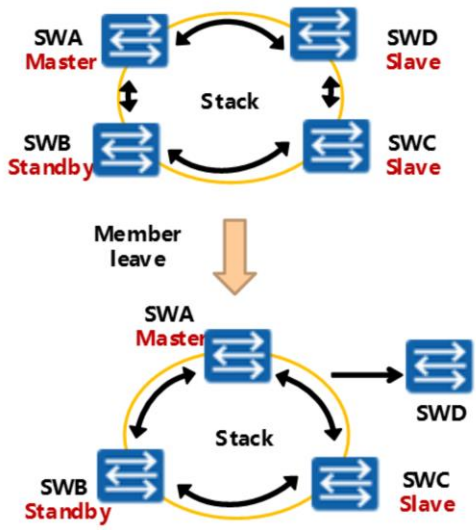
- Member switch addition means adding a switch to a stable stack. The following steps are involved in the process:
- Enable the stacking function and set stack parameters for SWD.
  - If service ports are used for stacking, the physical ports of the newly added switch must be added to the logical stack port as stack member ports. If the stack has a chain topology, perform this configuration at both ends (or one end) of the chain.
  - If stack cards are used for stacking, the stacking function must be enabled for the newly added switch.
  - To facilitate device management, configure a stack ID for the new member switch. If no stack ID is configured for the new member switch, the master switch assigns a stack ID to it.
- Connect SWD to the stack.
  - If the stack has a chain topology, add the new switch to either end of the chain to minimize the impact on running services.
  - If the stack has a ring topology, tear down a physical link to change the ring topology to a chain topology, and add the new switch to either end of the chain. Then connect the switches at two ends to form a ring if required.
- The system automatically completes the stack.
  1. After the switch is connected to the stack and is powered on, it is elected a slave switch. The roles of the other member switches in the stack remain unchanged.
  2. The master switch updates the stack topology information, synchronizes the stack topology information to the other member switches, and assigns a stack ID to the new member switch (if the new member switch has no stack ID configured or the configured stack ID conflicts with that of another member switch).
  3. The new member switch updates its stack ID and synchronizes its configuration file and system

software with the master switch. It then enters the stable running state.



- Stack merging means that two stable stacks are merged into one stack. In the preceding figure, the master switches SWA and SWD of the two stacks compete with each other for the final master role of the new merged stack. After SWA is elected the new master, the roles, configurations, and services of the member switches in the stack where SWA resides remain unaffected. In contrast, SWD and SWE in the other stack restart and join the new stack as slave switches. The master switch SWA assigns new stack IDs to SWD and SWE. SWD and SWE then synchronize their configuration files and system software with the master switch. During this process, services on SWD and SWE are interrupted.
- Stack merging occurs in either of the following situations:
  - A stack splits because a stack link or member switch fails. After the stack link or member switch recovers, the split stacks remerge into one.
  - After the stacking function is enabled on a switch to be added to a stack, the switch with power on is connected to a running stack through a stack cable. Using this method to merge the switch to a stack is not recommended because the running stack may restart during the merging process, affecting service running.

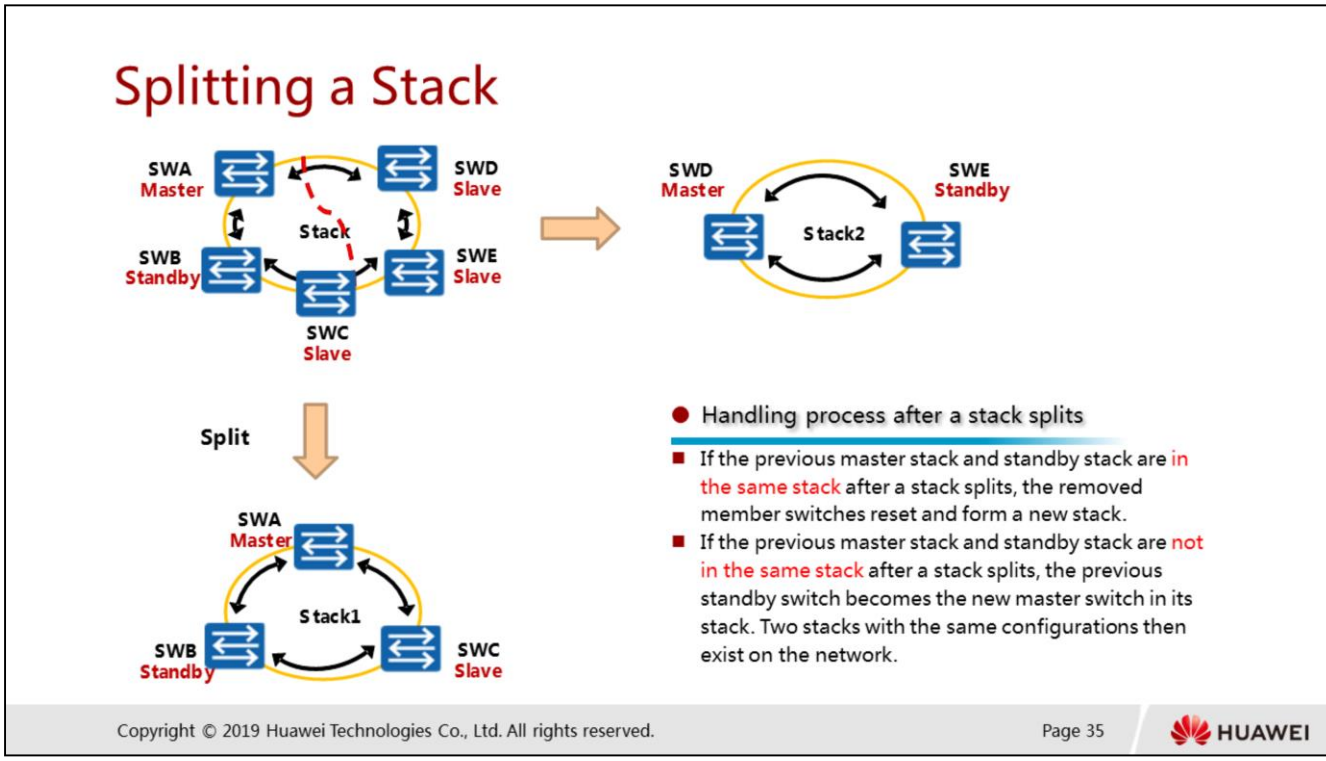
Removing a Member Switch from a Stack



- **Conditions for triggering member switch removal from a stack**
  - The stack cable is pulled out.
  - The involved stack port or physical member port is shut down.
  - The member switch is restarted.
  - The member switch fails.
- **Process for handling member switch removal**
  - If the master switch is removed, the standby switch becomes the master switch. The stack then updates its topology and continues to run properly.
  - If the standby switch is removed, the stack re-elects a standby switch. After the stack updates its topology, it continues to run properly.
  - If a slave switch is removed, the stack updates its topology and continues to run properly.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

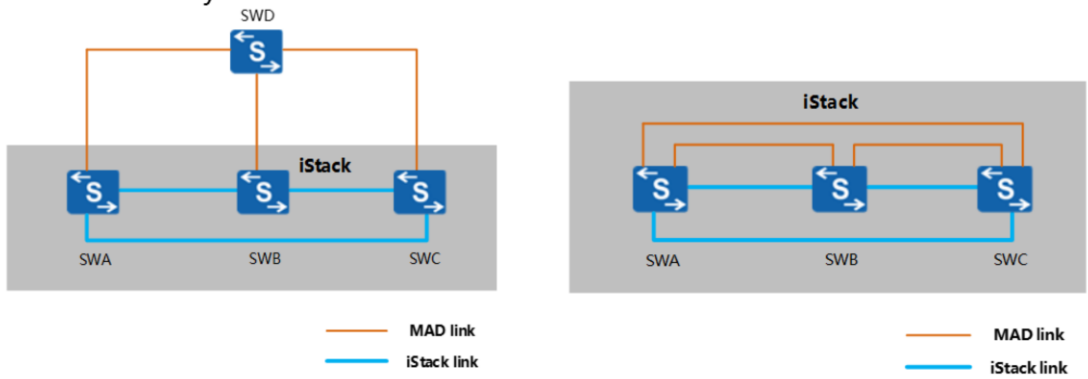
- Member switch removal means that a member switch leaves a stack. Depending on the role of the member switch that leaves a stack, the stack is affected in the following ways:
  - If the master switch leaves the stack, the standby switch becomes the new master switch. The new master switch then recalculates the stack topology, synchronizes updated topology information to the other member switches, and re-elects a new standby switch. Afterwards, the stack begins to run stably.
  - If the standby switch leaves the stack, the master switch selects a new standby switch, recalculates the stack topology, and synchronizes updated topology information to the other member switches. Afterwards, the stack begins to run stably.
  - If a slave switch leaves the stack, the master switch recalculates the stack topology and synchronizes updated topology information to the other member switches. Afterwards, the stack begins to run stably.
- A member switch leaves a stack after you disconnect its stack cables and remove it from the stack. When removing a member switch, pay attention to the following points:
  - After removing a member switch from a ring stack topology, use a stack cable to connect the two ports originally connected to this member switch to ensure network reliability.
  - In a chain topology, removing an intermediate switch causes the stack to split. Analyze the impact on services before doing so.



- Stack split means that a stack is split into multiple stacks because member switches are removed from a running stack without being powered off or because multiple points of failures occur on the stack cables. Depending on whether the previous master and standby switches are in the same stack after a stack splits, switch roles are elected in either of the following ways:
  - If the previous master and standby switches are in the same stack after a stack splits: The previous master switch recalculates the stack topology after deleting topology information related to the removed member switches, and synchronizes updated topology information to the other member switches. When removed member switches detect that the timeout timer for stack protocol packets has expired, they restart and begin a new master election.
  - If the previous master and standby switches are in different stacks after a stack splits: The previous master switch selects a new standby switch in its stack, recalculates stack topology information, and synchronizes updated topology information to the other member switches. The previous standby switch becomes the new master switch in its stack, recalculates stack topology information, synchronizes stack topology information to the other member switches, and selects a new standby switch.

Multi-Active Detection — Direct Mode

- In direct mode, stack members can be directly connected to either an intermediate device or fully meshed with each other:



Directly connected to an intermediate device

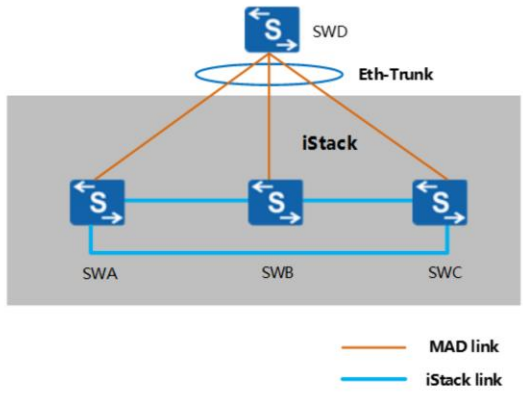
Fully meshed with each other

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

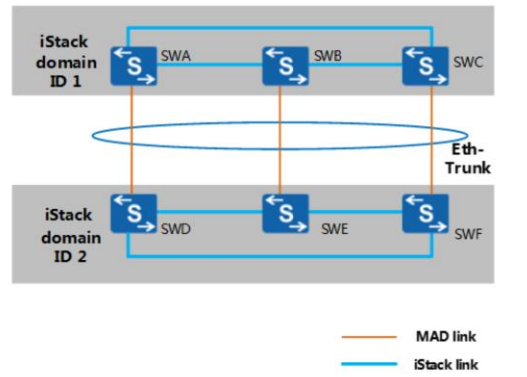
- All member switches in a stack use the same IP address and MAC address (stack MAC address). After a stack splits, more than one stack may use the same IP address and MAC address. To prevent a network fault caused by this situation, a mechanism is required to check for IP address collision and MAC address collision after a split. Multi-Active Detection (MAD) is a stack split detection and handling protocol. If a stack splits due to a link failure, MAD provides split detection, multi-active handling, and fault recovery mechanisms to minimize the impact of the stack split on services.
- MAD can be implemented in direct or relay mode. Direct and relay modes cannot be both configured in the same stack.
- In direct mode, stack members use MAD links over ordinary network cables. When the stack is running properly, member switches do not send MAD packets. After the stack splits, member switches each send a MAD packet every 1s over a MAD link to check whether more than one master switch exists.
- Directly connected to an intermediate device: Each member switch has at least one MAD link connected to the intermediate device.
- Fully meshed with each other: In the full-mesh topology, at least one MAD link exists between any two member switches.
- The use of an intermediate device can shorten the MAD links between member switches. This topology applies to stacks with a long distance between member switches. The full-mesh topology prevents MAD failures caused by intermediate device failures, but full-mesh connections occupy many interfaces on the member switches. Therefore, this topology applies to stacks with only a few member switches.

MAD — Relay Mode

- The relay mode can be implemented in two ways: a single switch as an MAD relay agent, and two stacks as MAD relay agents of each other.



A single switch as the MAD relay agent



Two stacks as MAD relay agents of each other

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In relay mode, MAD relay detection is configured on an Eth-Trunk interface in the stack, and the MAD detection function is enabled on an agent. Each member switch must have a link to the agent, and these links must be added to the same Eth-Trunk. In contrast to the direct mode, the relay mode does not require additional interfaces because the Eth-Trunk interface can run other services while performing MAD relay detection.
- In relay mode, when the stack is running properly, member switches send MAD packets at an interval of 30s over the MAD links and do not process received MAD packets. After the stack splits, member switches send MAD packets at an interval of 1s over the MAD links to check whether more than one master switch exists.
- Multi-active handling
  - After a stack splits, the MAD mechanism sets the new stacks to the Detect or Recovery state. The stack in Detect state still works, whereas the stack in Recovery state is disabled.
  - MAD handles a multi-active situation as follows: When multiple stacks in Detect state are detected by the MAD split detection mechanism, the stacks compete to retain the Detect state. The stacks that fail the competition enter the Recovery state, and all the physical ports except the reserved ports on the member switches in these stacks are shut down, so that the stacks in Recovery state no longer forward service packets.
- MAD fault recovery
- After the faulty link recovers, the stacks merge into one in either of the following ways:
  - The stack in Recovery state restarts and merges with the stack in Detect state, and the service ports that have been shut down are restored to Up state. The entire stack then recovers.
  - If the stack in Detect state becomes faulty before the faulty link recovers, you can remove this stack from the network and start the stack in Recovery state using a command to direct service traffic to this stack. Then rectify the stack fault and link fault. After the stack in Detect state recovers, merge it with the other stack.

Configuring a Stack

- Set up a stack using stack cards.

```
[SwitchA] stack slot 0 priority 200
\\Configures a stack priority for SwitchA. By default, the stack priority of a member switch is 100.
[SwitchB] stack slot 0 renumber 1
\\Configures a stack ID for SwtichB.
[SwitchC] stack slot 0 renumber 2
```

- Set up a stack using service ports.

```
[SwitchA] interface stack-port 0/1
[SwitchA-stack-port0/1] port interface gigabitethernet 0/0/27 enable
\\Configures a physical port as a service port and adds it to a logical stack port. Perform similar operations on SwitchB and SwitchC.
[SwitchA] interface stack-port 0/2
[SwitchA-stack-port0/2] port interface gigabitethernet 0/0/28 enable
[SwitchA] stack slot 0 priority 200
\\Sets the stacking priority of SwitchA to 200.
[SwitchB] stack slot 0 renumber 1
\\Sets the stacking ID of SwitchB to 1.
[SwitchC] stack slot 0 renumber 2
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

1. MAC Address Table
2. Gratuitous ARP
3. MSTP
4. iStack
5. **CSS**
6. Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

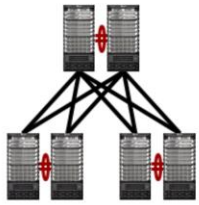
CSS



Physical form of a CSS



Logical form of a CSS



Physical form of a CSS network



Logical form of a CSS network

CSS definition

A cluster switch system (CSS), also called a cluster, is a logical switch consisting of two clustering-capable switches.

CSS features

- Many-to-one virtualization: Switches can be virtualized into one logical switch (CSS) that has a unified control plane for unified management.
- Unified forwarding plane: Physical switches in a CSS use a unified forwarding plane, and share and synchronize forwarding information in real time.
- Inter-device link aggregation: Links between physical switches in a CSS can be aggregated into a single Eth-Trunk interface for interconnection with downstream devices.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The difference between a CSS and iStack lies in that a CSS is a stack of modular switches while an iStack is a stack of fixed-configuration switches. They have different names and some unique implementations but provide similar functions.
- In addition to high forwarding performance, CSS technology provides high network reliability and scalability while simplifying network management.
  - High reliability: Member switches in a CSS work in redundancy mode. Link redundancy can also be implemented between member switches through inter-chassis link aggregation.
  - High scalability: Switches can set up a CSS to increase the number of ports, bandwidth, and packet processing capabilities.
  - Simplified configuration and management: After two switches set up a CSS, they are virtualized into a single switch. You can log in to the CSS from either member switch to configure and manage the CSS.

Basic CSS Concepts

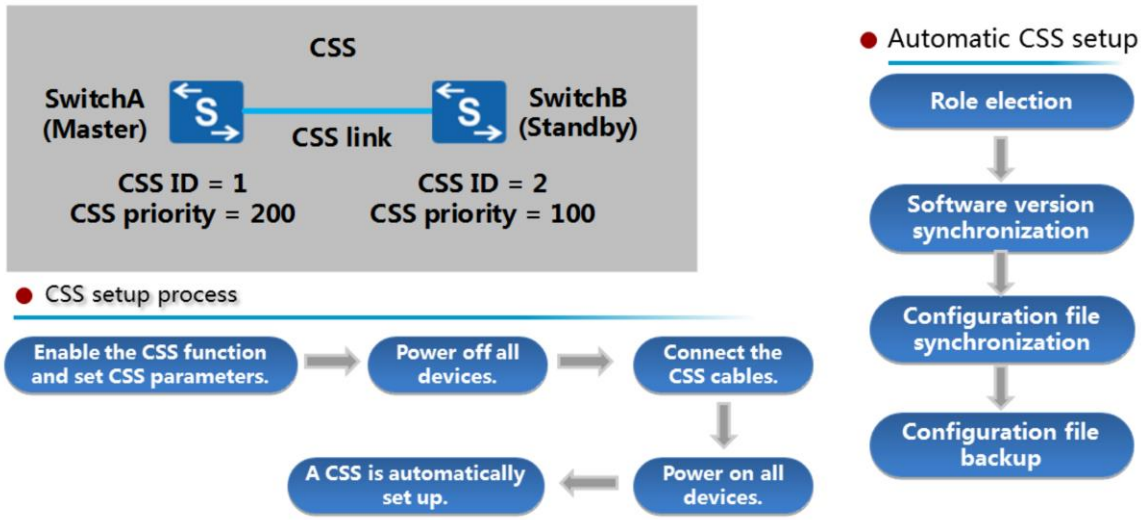
- Master switch
  - Manages the entire CSS. A CSS has only one master switch.
- Standby switch
  - A standby switch acts as a backup of the master switch. If the master switch fails, the standby switch takes over all services from the master switch and assumes the CSS master role. A CSS has only one standby switch.
- CSS ID
  - A CSS ID is used to identify and manage a CSS member switch. Each member switch has a unique CSS ID.
- CSS priority
  - A CSS priority determines the role of a CSS member in role election. A higher priority value for a CSS member increases the probability of it being elected the CSS master.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Different from iStack, which allows multiple switches to be stacked, a CSS has only one master switch and one standby switch.

Setting up a CSS

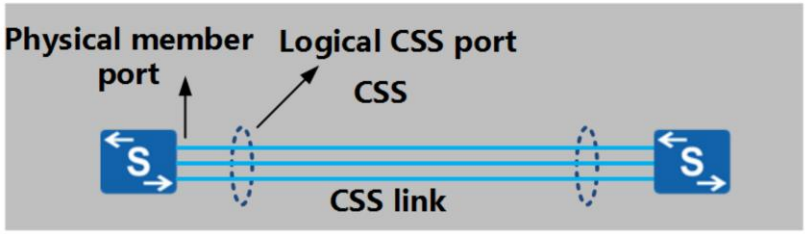


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- A CSS is set up automatically after you use cluster cables to connect two switches, enable the CSS function on the two switches, and restart them. The member switches then exchange CSS competition packets for role election. Through competition, one switch becomes the master switch to manage the CSS, and the other becomes the standby switch.
- Role election
  1. The switch that first starts up and enters the single-chassis CSS running state becomes the master switch.
  2. If the two switches start up at the same time, the switch with a higher CSS priority becomes the master switch.
  3. If the two switches start up at the same time and have the same CSS priority, the switch with a smaller MAC address becomes the master switch.
  4. If the two switches start up at the same time and have the same CSS priority and MAC address, the switch with a smaller CSS ID becomes the master switch.
- Software version synchronization
  - CSS technology provides an automatic software loading mechanism. Switches do not have to run the same software version and can set up a CSS if their software versions are compatible with one another. If the software version running on the standby switch is different from that on the master switch, the standby switch downloads the system software from the master switch, restarts with the new system software, and re-joins the CSS.
- Configuration file synchronization
  - CSS technology uses a strict mechanism to synchronize configuration files. This mechanism ensures that CSS member switches function as a single switch.
- Configuration file backup
  - After a switch enters the CSS state, it automatically adds the file name extension .bak to the name of its original configuration file and backs up the configuration file. In this way, the switch can restore the previous configuration if the CSS function is disabled. For example, if the original configuration file name extension is .cfg, the backup configuration file name extension becomes .cfg.bak. If you want to restore the original configuration of a switch after disabling the CSS function, delete the extension .bak from the backup configuration file name, specify the configuration file without .bak for next startup, and then restart the switch.

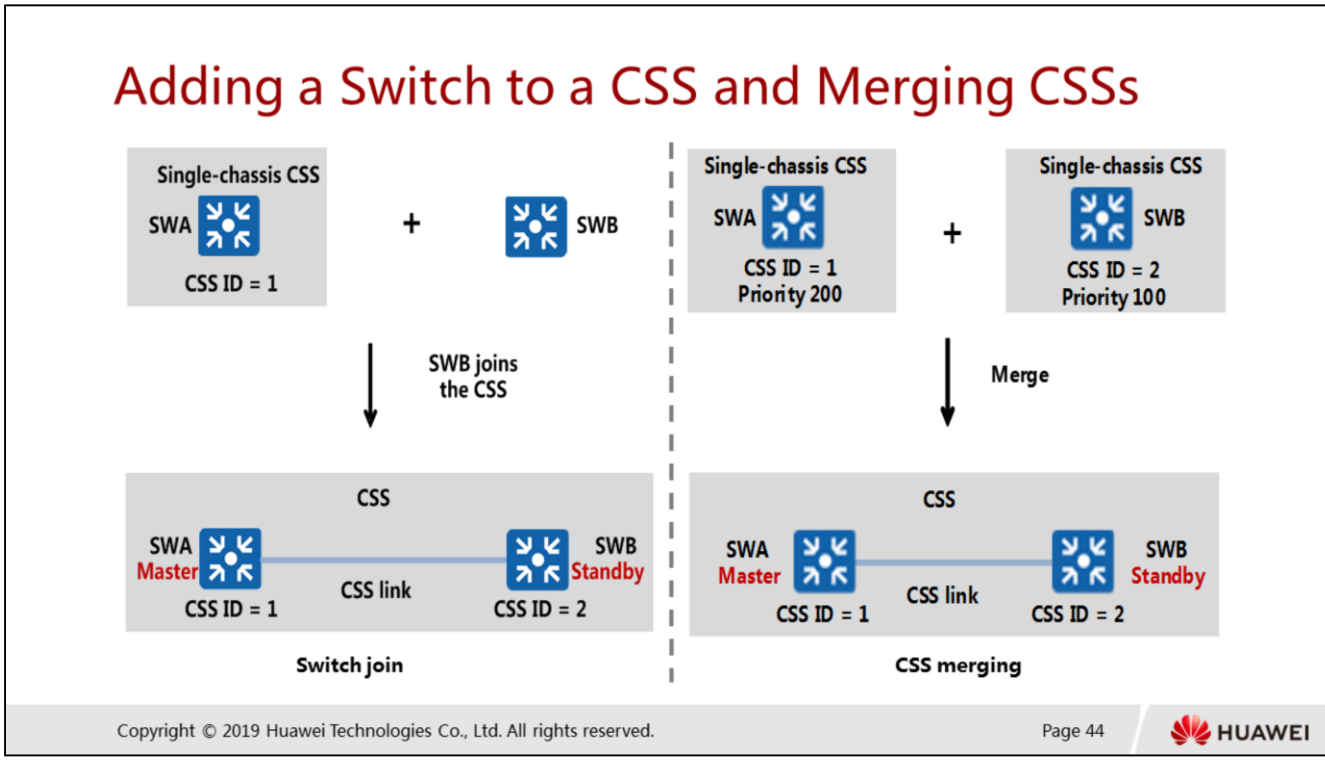
CSS Connection Modes

- Switches can set up a CSS using CSS cards or service ports.
  - In CSS card connection mode, switches are connected using dedicated CSS cards and CSS cables on the MPUs.
  - In service port connection mode, switches are connected using ordinary service ports on service cards, without a need for CSS cards. Similar to iStack, service port-based CSS involves two types of ports: physical member ports and logical CSS ports



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Physical member port
  - A physical member port is a service port used to set up a CSS link between CSS member switches. Physical member ports forward service packets or CSS protocol packets between member switches.
- Logical CSS port
  - A logical CSS port is exclusively used for CSS setup and must have physical member ports bundled. Each CSS member switch supports a maximum of two logical CSS ports.



- A single CSS-enabled switch is a single-chassis CSS.
- A switch can join a running single-chassis CSS. As shown in the left figure, SwitchA is a running single-chassis CSS. After SwitchB joins the CSS, the two switches set up a new CSS. SwitchA becomes the master switch, and SwitchB becomes the standby switch.
- A switch joins a single-chassis CSS in either of the following situations:
  - After one switch has the CSS function enabled and is restarted, the switch enters the single-chassis CSS state. After the other switch has the CSS function enabled and is restarted, it joins the CSS as the standby switch.
  - In a running two-chassis CSS, after one switch is restarted, it re-joins the CSS as the standby switch.
- Two single-chassis CSSs can merge into one CSS. As shown in the right figure, two single-chassis CSSs merge into one and elect a master switch. The master switch retains its original configuration but its standby MPU resets, without affecting services. The standby switch is restarted, joins the new CSS as the standby switch, and synchronizes its configuration file with the master switch. Existing services on this switch are interrupted.
- CSS merging occurs in either of the following situations:
  - After two switches are configured with the CSS function and restarted, they run as two single-chassis CSSs. After they are connected using cluster cables, they merge into one CSS.
  - A CSS splits due to a failure of a CSS link or member switch. After this link or switch recovers, the two single-chassis CSSs merge into one.

Splitting a CSS

- After a CSS is set up, the master and standby MPUs of the CSS periodically exchange heartbeat packets to maintain the CSS status. If a cluster cable, a CSS card, or an MPU fails or one switch is powered off or restarted, communication between the two switches is interrupted. When the heartbeat timeout timer (8s) expires, the CSS splits into two single-chassis CSSs, as shown in the following figure.

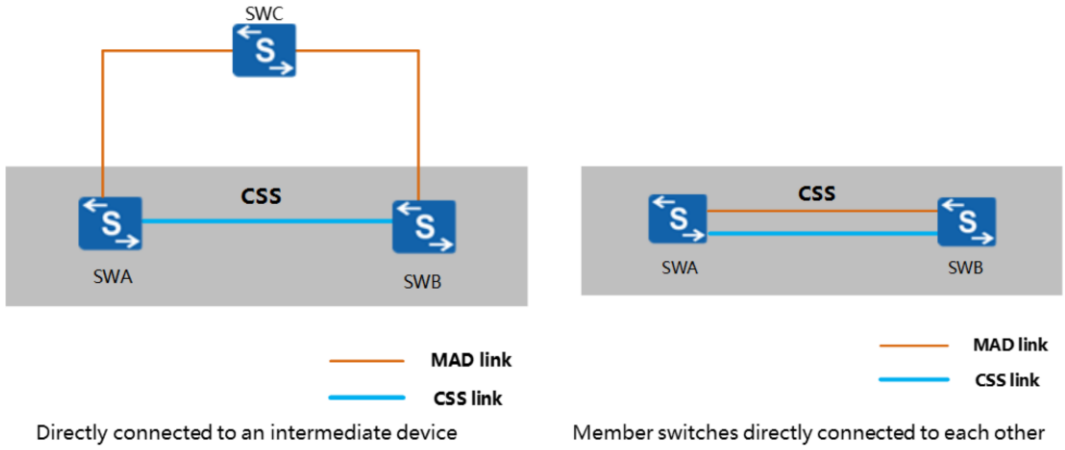


- After a CSS splits, two CSSs with the same IP address and MAC address are generated because member switches run the same configuration file. To prevent a network fault triggered by this situation, a mechanism is required to check for IP address and MAC address collision after a CSS split.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MAD— Direct Mode

- In direct mode, CSS member switches can be directly connected to an intermediate device or directly connected to each other:



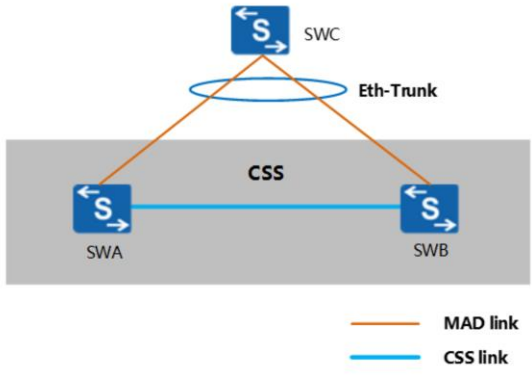
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

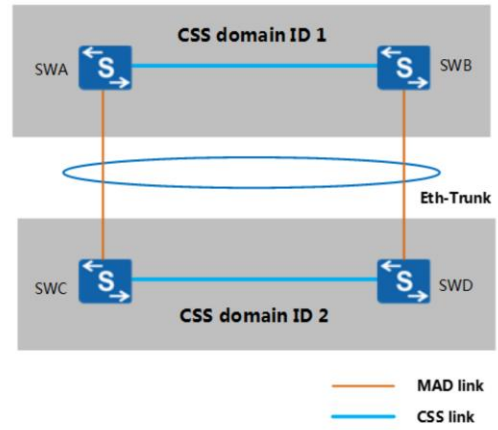
- Two member switches in a CSS use the same IP address and MAC address (CSS system MAC address). After the CSS splits, it becomes two single-chassis CSSs using the same IP address and same MAC address, because the two switches both run the configuration file of the previous CSS. To prevent this situation, a mechanism is required to check for IP address and MAC address collision after a CSS split.
- MAD is a CSS split detection and handling protocol. When a CSS splits due to a link failure, MAD provides split detection, multi-active handling, and fault recovery mechanisms to minimize the impact of a CSS split on services.
- MAD can be implemented in direct or relay mode. The direct and relay modes cannot both be configured in the same CSS.
- In direct mode, CSS member switches use MAD links over ordinary network cables. When the CSS is running properly, member switches do not send MAD packets. After the CSS splits, member switches periodically send MAD packets over MAD links to check whether more than one master switch exists.
- In direct mode, CSS member switches can be directly connected to an intermediate device or directly connected to each other:
    - Directly connected to an intermediate device: Each member switch has at least one MAD link connected to the intermediate device. This deployment can be used when member switches are far from each other.
    - Directly connected to each other: No intermediate device is deployed to prevent MAD from being affected by intermediate device failures.

MAD — Relay Mode

- In relay mode, an independent switch can function as a relay agent, or two CSSs can function as each other's relay agents.



A single switch as the MAD relay agent



Two CSSs as MAD relay agents of each other

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In relay mode, MAD relay detection is configured on an Eth-Trunk interface in the CSS, and the MAD detection function is enabled on an agent. Each member switch must have a link to the agent, and these links must be added to the same Eth-Trunk. In contrast to the direct mode, the relay mode does not require additional interfaces because the Eth-Trunk interface can run other services while performing MAD relay detection.
- In relay mode, when the CSS is running properly, member switches send MAD packets at an interval of 30s over the MAD links and do not process received MAD packets. After the CSS splits, member switches periodically send MAD packets over MAD links to check whether more than one master switch exists.
- Multi-active handling
  - After a CSS splits, the MAD mechanism sets the new single-chassis CSSs to Detect or Recovery state. The CSS in Detect state still works, whereas the CSS in Recovery state is disabled.
  - MAD handles a multi-active situation as follows: When detecting two CSSs (two switches) in Detect state, MAD allows only the switch with a higher CSS priority to work. (If the two switches have the same CSS priority, their MAC addresses and CSS IDs are compared in turn.) The other switch then enters the Recovery state, and all its physical ports except the reserved ones are shut down to prevent the switch from forwarding service packets.
- MAD fault recovery
- After the faulty link recovers, the CSSs merge into one in either of the following ways:
  - The CSS in Recovery state restarts and merges with the CSS in Detect state, and the service ports that have been shut down are restored to Up state. The entire CSS then recovers.
  - If the CSS in Detect state becomes faulty before the faulty link recovers, you can remove this CSS from the network and start the CSS in Recovery state using a command to direct service traffic to this CSS. Then rectify the CSS fault. After the CSS recovers, merge it with the other CSS.

Configuring a CSS

- Set up a CSS using CSS cards.

```
[SwitchA] set css mode css-card     \\Configures the CSS card connection mode.
[SwitchA] set css id 1               \\Sets a CSS ID for SwitchA.
[SwitchA] set css priority 100       \\Sets a CSS priority for SwitchA.
[SwitchA] css enable                 \\Enables the CSS function on SwitchA.
```

- Set up a CSS using service ports.

```
[SwitchA] set css mode lpu      \\Configures the service port connection mode.
[SwitchA] set css id 1          \\Sets a CSS ID for SwitchA.
[SwitchA] set css priority 100  \\Sets a CSS priority for SwitchA.
[SwitchA] interface css-port 1  \\Enters the logical CSS port view.
[SwitchA-css-port1] port interface xgigabitethernet 1/0/1 to xgigabitethernet 1/0/2 enable
                    \\Configures the service ports as physical member ports and add them to the
logical CSS port.
[SwitchA] interface css-port 2
[SwitchA-css-port2] port interface xgigabitethernet 2/0/1 to xgigabitethernet 2/0/2 enable
[SwitchA] css enable                \\Enables the CSS function on SwitchA.
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. MAC Address Table
2. Gratuitous ARP
3. MSTP
4. iStack
5. CSS
6. **Eth-Trunk**
   - Eth-Trunk Fundamentals
   - Eth-Trunk Implementation
   - E-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Eth-Trunk Fundamentals

- Ethernet link aggregation, also called Eth-Trunk, bundles multiple physical links into a logical link to increase available link bandwidth. The bundled links back up each other, improving link reliability.
- The Eth-Trunk link connected to Eth-Trunk interfaces can be considered as a point-to-point link. Traffic can be balanced among the Eth-Trunk member links. Various routing protocols and services can be configured on the Eth-Trunk interface.



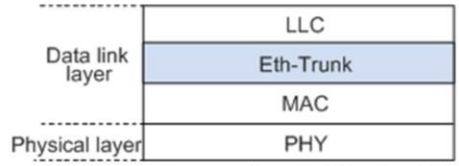
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Link aggregation increases the total available bandwidth, improves link reliability, and supports traffic load balancing among member links.
- Link aggregation group and member interface
  - A link aggregation group (LAG) is a logical link composed of multiple Ethernet links. The interfaces that are bundled into the Eth-Trunk interface (or LAG) are member interfaces.
- Active and inactive interfaces and links
  - Eth-Trunk member interfaces can be active or inactive. An interface that forwards data is active, while an interface that does not forward data is inactive.
  - The link connected to an active interface is an active link, whereas the link connected to an inactive interface is an inactive link.
- Upper threshold for the number of active member interfaces
  - When the number of active member interfaces in an Eth-Trunk interface reaches the upper threshold, newly added interfaces cannot be active but function as backup interfaces. The links of the newly added interfaces are set to Down.
- Lower threshold for the number of active member interfaces
  - When the number of active interfaces in an Eth-Trunk interface falls below the lower threshold, the Eth-Trunk interface goes Down. This ensures that the Eth-Trunk interface has the minimum required bandwidth.
- Link aggregation modes supported by a device
  - Intra-card: Member interfaces of an Eth-Trunk interface reside on the same card.
  - Inter-card: Member interfaces of an Eth-Trunk interface reside on different cards.
  - Inter-chassis: Member interfaces of an Eth-Trunk interface reside on member devices of a cluster.
  - Inter-device: Inter-device link aggregation refers to Enhanced Trunk (E-Trunk). E-Trunk extends LACP and allows link aggregation between different devices.

Forwarding Principles

- Eth-Trunk lies between the MAC and LLC sub-layers and belongs to the data link layer.



- The Eth-Trunk module maintains a forwarding table that is composed of the following items:
  - HASH-KEY: calculated using the hash algorithm based on the MAC address or IP address in a data packet.
  - PORT: outbound interface used to forward packets. The interface numbers vary with the maximum number of member interfaces supported by Eth-Trunk interfaces and can be mapped to different hash keys.

HASH-KEY	0	1	2	3	4	5	6	7
PORT	1	2	3	4	1	2	3	4

Example of an Eth-Trunk forwarding table

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The Eth-Trunk module forwards data frames as follows:
  1. After the Eth-Trunk module receives a data frame from the MAC sub-layer, it extracts the frame's source MAC address, source IP address, destination MAC address, or destination IP address according to the load balancing mode.
  2. The Eth-Trunk module obtains the hash keys using the hash algorithm.
  3. Based on the hash keys, the Eth-Trunk module searches for the interface number in the Eth-Trunk forwarding table and then sends the data frame from the corresponding interface.
- For example, an Eth-Trunk interface on a device supports a maximum of eight member interfaces. If physical interfaces 1, 2, 3, and 4 are bundled into an Eth-Trunk interface, the generated Eth-Trunk forwarding table contains eight entries, as shown in the preceding second figure. In the Eth-Trunk forwarding table, hash keys are 0, 1, 2, 3, 4, 5, 6, and 7, and the corresponding interface numbers are 1, 2, 3, 4, 1, 2, 3, and 4.
- To prevent out-of-order data frames, an Eth-Trunk interface uses per-flow load balancing. Data forwarding varies according to the load balancing mode.
- The load balancing modes based on the following fields of data frames are available and can be selected as required:
  - Source MAC addresses
  - Destination MAC address
  - Source IP addresses
  - Destination IP addresses
  - Source and destination MAC addresses
  - Source and destination IP addresses
  - VLAN IDs and source physical interface numbers (enhanced load balancing mode for Layer 2,

IPv4, IPv6, and MPLS packets)

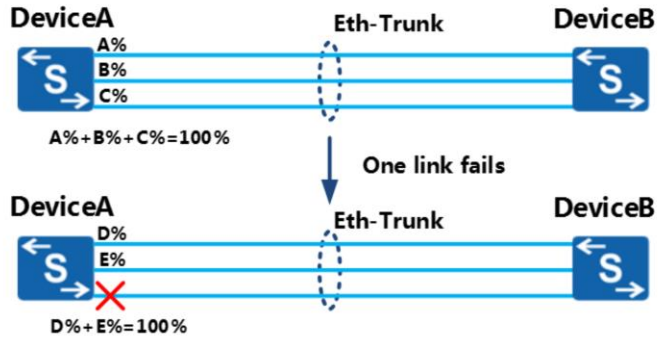
Contents

1. MAC Address Table
2. Gratuitous ARP
3. MSTP
4. iStack
5. CSS
6. **Eth-Trunk**
   - Eth-Trunk Fundamentals
   - Eth-Trunk Implementation
   - E-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Link Aggregation in Manual Mode

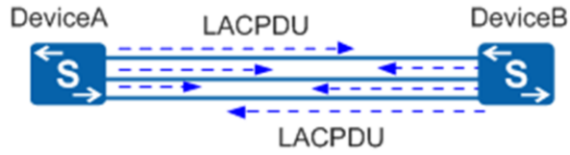
- In manual mode, an Eth-Trunk interface is manually created, and interfaces are manually added to the Eth-Trunk interface, without requiring LACP. In the following figure, after an Eth-Trunk interface is manually created on DeviceA and DeviceB, all the three active links as Eth-Trunk member links forward traffic. If one link becomes faulty, the remaining two Eth-Trunk member links continue to balance traffic.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Link Aggregation in LACP Mode (1)

- To improve Eth-Trunk fault tolerance and provide link backup for higher reliability, the Link Aggregation Control Protocol (LACP) was introduced. After an Eth-Trunk link is formed, LACP maintains or adjusts the link dynamically.
- An Eth-Trunk interface in LACP mode is set up as follows:
  1. Devices at both ends exchange LACPDUs. In the following figure, after an Eth-Trunk interface is created on DeviceA and DeviceB and configured to work in LACP mode, interfaces are added to the Eth-Trunk interface. LACP is then automatically enabled on the member interfaces, and devices at both ends begin to exchange LACPDUs.

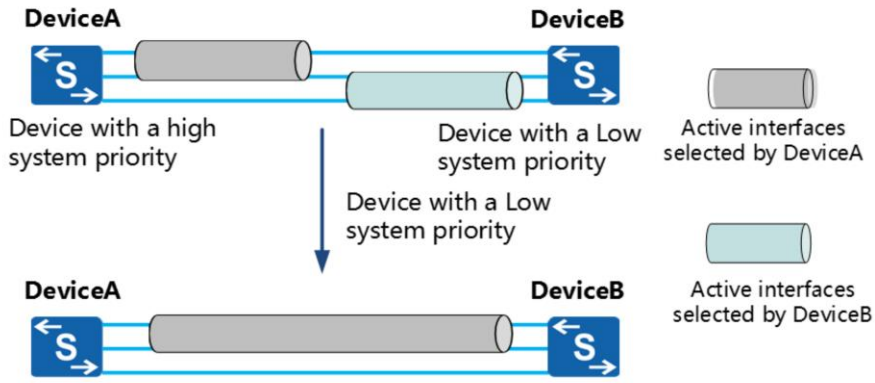


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Link aggregation in manual mode can increase the available bandwidth and detect a few faults like member link disconnections within an aggregation group. However, it cannot detect other faults, such as link layer faults and incorrect link connections.

Link Aggregation in LACP Mode (2)

2. Determine the Actor and active links.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- In the preceding figure, the devices at both ends receive LACPDUs from each other. DeviceB is used as an example. When DeviceB receives LACPDUs from DeviceA, DeviceB checks and records information about DeviceA and compares LACP system priorities. If the LACP system priority of DeviceA is higher than that of DeviceB, DeviceA becomes the Actor. If the two devices have the same system priority, the device with a smaller MAC address functions as the Actor.
- After the Actor is selected, both devices select active interfaces based on the interface priorities of the Actor. If priorities of interfaces on the Actor are the same, interfaces with smaller interface numbers are selected as active interfaces. After devices at both ends select consistent active interfaces, the Eth-Trunk interface begins to balance traffic among its member interfaces.

Configuring Link Aggregation

- Configure link aggregation in manual mode.

```
[SwitchA] interface eth-trunk 1
//Creates an Eth-Trunk interface and enters the Eth-Trunk interface view.
[SwitchA-Eth-Trunk1] mode manual load-balance
//Sets the manual link aggregation mode.
[SwitchA-Eth-Trunk1] trunkport gigabitethernet 0/0/1 to 0/0/3
//Adds interfaces to the Eth-Trunk interface.
```

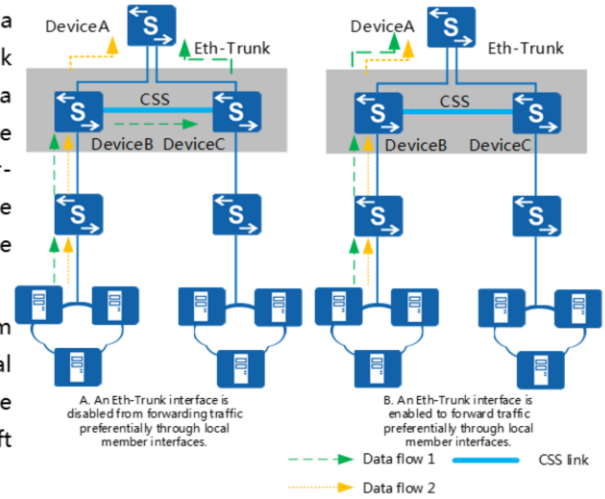
- Configure link aggregation in LACP mode.

```
[SwitchA] interface eth-trunk 1
[SwitchA-Eth-Trunk1] mode lacp
//Sets the LACP link aggregation mode.
[SwitchA-Eth-Trunk1] max active-linknumber 2  \\Sets the upper threshold for the number of active
member interfaces to 2.
[SwitchA] interface gigabitethernet 0/0/1  \\Adds an interface to the Eth-Trunk interface.
[SwitchA-GigabitEthernet0/0/1] eth-trunk 1
[SwitchA] interface gigabitethernet 0/0/2
[SwitchA-GigabitEthernet0/0/2] eth-trunk 1
[SwitchA] interface gigabitethernet 0/0/3
[SwitchA-GigabitEthernet0/0/3] eth-trunk 1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Link Aggregation in Cluster Scenarios

- Physical interfaces on different devices in a cluster can be aggregated into an Eth-Trunk interface. If one device in the cluster or a physical Eth-Trunk member interface fails, the data traffic can be transmitted along the inter-chassis cluster cable, thereby ensuring reliable data transmission and implementing device backup.
- When the network runs properly, traffic from DeviceB or DeviceC is forwarded through local member interfaces, instead of through the inter-chassis cluster cable (as shown in the left figure).



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- When devices form a cluster, an Eth-Trunk interface can be configured as the traffic outbound interface for reliable traffic transmission. In the Eth-Trunk, there must be member interfaces residing on different devices. When the cluster forwards traffic, the Eth-Trunk interface may select inter-chassis member interfaces to forward traffic after using the hash algorithm to calculate the outbound interfaces. The cable bandwidth between devices in the cluster is limited. Inter-chassis traffic forwarding further increases the bandwidth bearer pressure on the cluster cable and lowers the traffic forwarding efficiency. To resolve this issue, Eth-Trunk interface traffic can be preferentially forwarded by local devices.
- As shown in the preceding figure, DeviceB and DeviceC form a cluster, and the cluster connects to DeviceA through an Eth-Trunk interface. After the cluster is configured to preferentially forward traffic through local devices, the following two situations may occur:
  - Traffic entering a local device is directly forwarded by the local device.
    - If DeviceB has working Eth-Trunk member interface as the outbound interfaces, the Eth-Trunk forwarding table of DeviceB contains only the local member interfaces. In this way, only DeviceB's interfaces are selected as the outbound interfaces for the traffic from DeviceB to DeviceA using the hash algorithm, indicating that traffic is directly forwarded by DeviceB.
  - Traffic entering a local device is forwarded by another device.
    - If DeviceB has no Eth-Trunk member interfaces as outbound interfaces or all the outbound interfaces fail, the Eth-Trunk forwarding table of DeviceB contains all available member interfaces. In this way, the member interfaces on DeviceC are selected as the outbound interfaces for the traffic from DeviceB to DeviceA using the hash algorithm, indicating that traffic is forwarded through DeviceC.

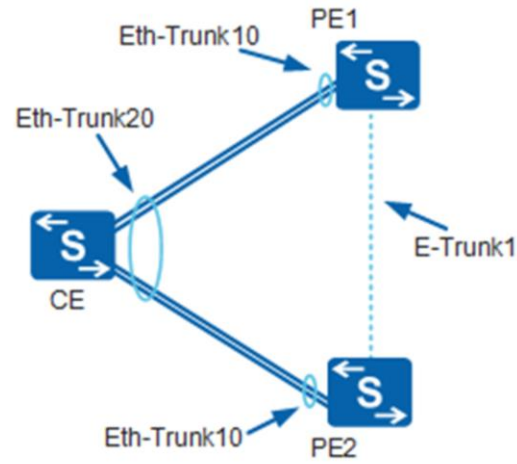
Contents

1. MAC Address Table
2. Gratuitous ARP
3. MSTP
4. iStack
5. CSS
6. **Eth-Trunk**
   - Eth-Trunk Fundamentals
   - Eth-Trunk Implementation
   - E-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

E-Trunk vs. Eth-Trunk

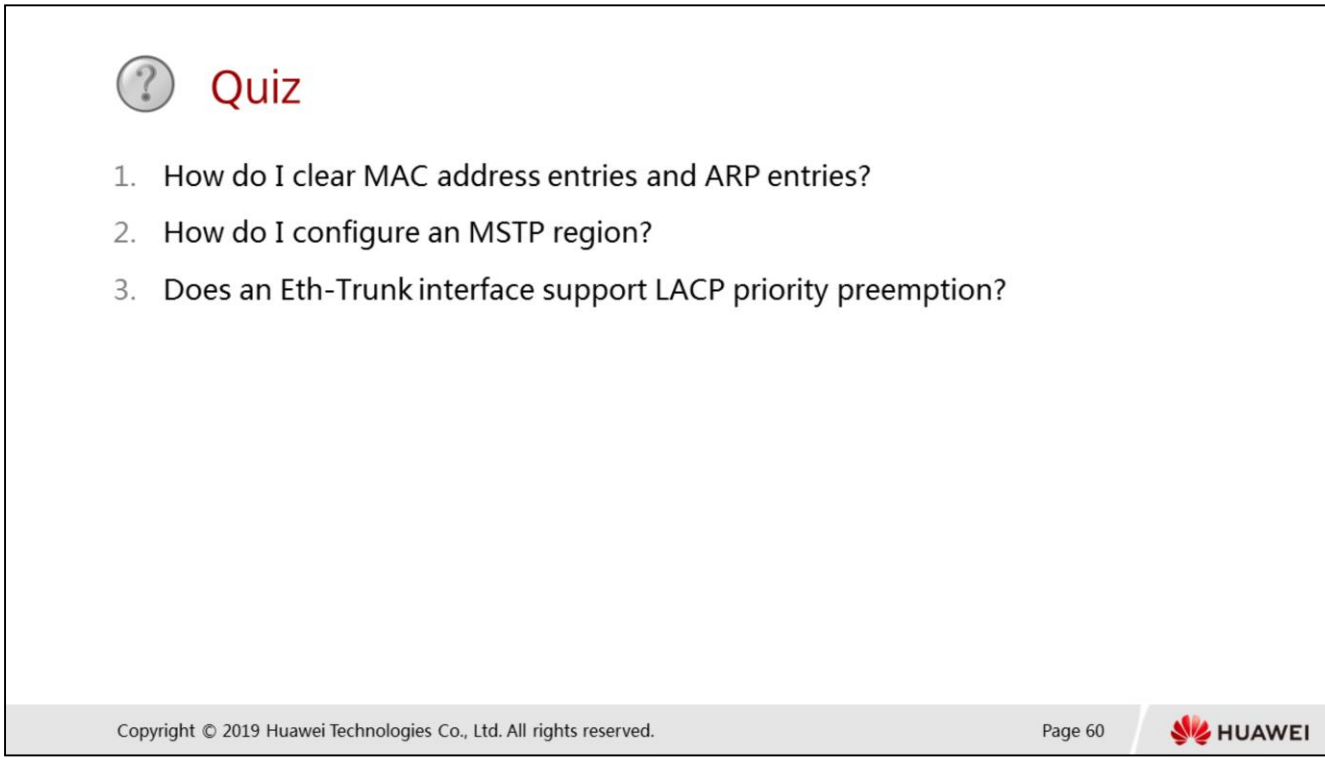
- Enhanced Trunk (E-Trunk) is an extension of LACP. It controls and implements link aggregation among multiple devices for device-level link reliability, instead of board-level link reliability.
- In the right figure, a CE is dual-homed to PE1 and PE2, and an E-Trunk is deployed on PE1 and PE2. If the link between the CE and PE1 fails or PE1 fails, traffic can be switched to the link between the CE and PE2, thereby implementing device-level protection.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- When a CE is dual-homed to a VPLS, VLL, or PWE3 network, E-Trunk is used to protect PEs and links between the CE and PEs. Without E-Trunk, a CE can connect to only one PE by using an Eth-Trunk link. If the Eth-Trunk link or PE fails, the CE cannot communicate with the PE. With E-Trunk, the CE can be dual-homed to PEs to protect PEs and links between the CE and PEs, enabling device-level protection.
- In the preceding figure, the CE is directly connected to PE1 and PE2. E-Trunk needs to run between PE1 and PE2. The configuration is as follows: Create E-Trunks with the same ID and Eth-Trunk interfaces with the same ID on PE1 and PE2 and add the Eth-Trunk interfaces to the E-Trunk. Configure an Eth-Trunk interface (Eth-Trunk 20) in LACP mode on the CE, and connect the Eth-Trunk interface to PE1 and PE2. The CE is unaware of the E-Trunk.
- PE1 and PE2 exchange E-Trunk packets to negotiate their E-Trunk master/backup status. After the negotiation, one PE functions as the master, and the other as the backup. The master/backup status of a PE depends on the E-Trunk priority and E-Trunk system ID carried in the PE's E-Trunk packets. The PE with a higher E-Trunk priority (smaller value) functions as the master device. If the PEs have the same E-Trunk priority, the PE with a smaller E-Trunk system ID functions as the master device. This example assumes that PE1 functions as the master. Eth-Trunk 10 of PE1 then stays in the master state with an Up link status. PE2 functions as the backup, and Eth-Trunk 10 of PE2 stays in the backup state with a Down link status.
- If the link between the CE and PE1 fails, PE1 sends an E-Trunk packet containing Eth-Trunk 10 failure information to PE2. Upon receipt, PE2 finds that Eth-Trunk 10 on PE1 is faulty and changes its Eth-Trunk 10 status to master. Through LACP negotiation, Eth-Trunk 10 on PE2 becomes Up. Traffic from the CE is then forwarded to PE2, preventing CE traffic interruption.
- If both PEs are configured with BFD and PE1 fails, after PE2 detects the Down BFD session status, it changes its state from backup to master. Eth-Trunk 10 on PE2 then enters the master state. If BFD is not configured on PEs and PE2 does not receive E-Trunk packets from PE1 before the timer expires, PE2 changes its state from backup to master. Eth-Trunk 10 on PE2 then enters the master state. Through LACP negotiation, Eth-Trunk 10 on PE2 becomes Up. Traffic from the CE is then forwarded to PE2, preventing CE traffic interruption.



- Answers
- How do I clear MAC address entries and ARP entries?
  - To clear all dynamic MAC addresses in the system view, run the **undo mac-address dynamic** command.
  - To clear all static MAC addresses in the system view, run the **undo mac-address static** command.
  - To clear one static ARP entry in the system view, run the **undo arp static** command.
  - To clear all ARP entries in the user view, run the **reset arp** command.
- How do I configure an MSTP region?
  - Run the **stp region-configuration** command to enter the MST region view to configure region information. The devices in the same MST region must have the same MST region configuration. Any difference will cause the devices to be in different regions. The following parameters can be set for an MST region:
    - Format selector: The default value is 0 and cannot be set using commands.
    - Region name: name of an MST region. The default value is the bridge MAC address.
    - Revision level: The default value is 0.
    - Instance/Vlans Mapped: mapping between MSTIs and VLANs. By default, all VLANs are mapped to instance 0.
- Does an Eth-Trunk interface support LACP priority preemption?
  - Only Eth-Trunk interfaces in LACP mode support LACP priority preemption. To enable LACP priority preemption, run the **lacp preempt enable** command. In LACP mode, if an active link fails, a device selects the link with the highest priority from backup links to replace the faulty one. With LACP priority preemption enabled, if the faulty link recovers and has a priority higher than that of the replacement link, the recovered link will preempt the replacement link and switches to active. LACP priority preemption configurations at both ends of an Eth-Trunk link must be the same, either enabled or disabled.

Section Summary

- MAC Address Table
- Gratuitous ARP
- MSTP
- iStack
- CSS
- Eth-Trunk

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— WAN Technology

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- This document describes principles and configurations of common interfaces, PPP, MP, POS interface, IP-Trunk, and PPPoE on the WAN.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- On completion of this course, you will be able to:
  - Describe common Ethernet and POS interfaces on the WAN.
  - Describe the principle and configuration of PPP and MP.
  - Describe the principle and configuration of IP-Trunk.
  - Describe the principle and configuration of PPPoE.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

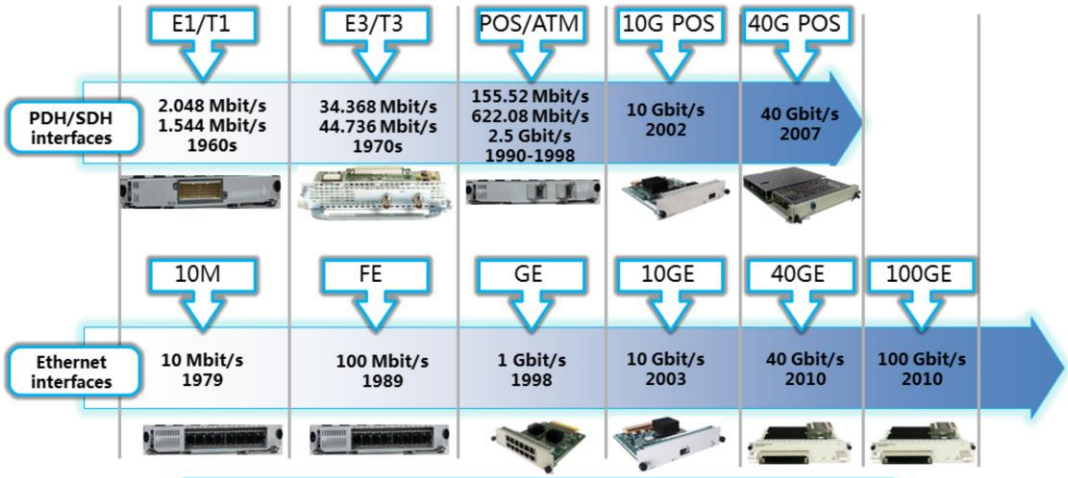
Contents

1. **WAN Interfaces**
   - Evolution of WAN Interfaces on Routers
   - Ethernet Interface
   - POS Interface
   - Comparison Between Ethernet and POS Interfaces
2. Working Mechanism and Configuration of PPP
3. IP-Trunk
4. Working Mechanism and Configuration of PPPoE

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Bandwidth Evolution of Physical Interfaces on Routers

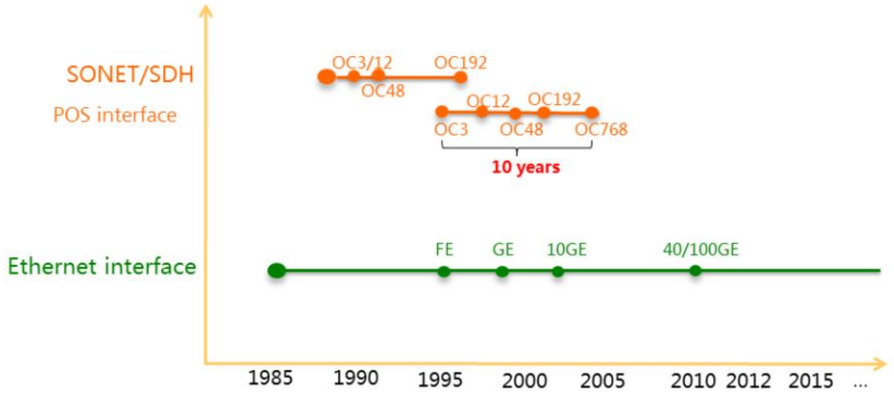


Evolution of router interfaces
Core elements of interfaces: bandwidth, transmission distance, cost, power consumption, density, and compatibility

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Evolution Trend of POS and Ethernet Interfaces



- POS interfaces are widely used on the live network.
- Ethernet interfaces continuously develop.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Contents

1. **WAN Interfaces**
   - Evolution of WAN Interfaces on Routers
   - Ethernet Interface
   - POS Interface
   - Comparison Between Ethernet and POS Interfaces
2. Working Mechanism and Configuration of PPP
3. IP-Trunk
4. Working Mechanism and Configuration of PPPoE

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Ethernet Interface

- Ethernet was first proposed by Xerox PARC Research Center on May 22, 1973.
- Ethernet types:
  - 10M Ethernet (standard Ethernet)
  - 100M Ethernet (fast Ethernet)
  - 1000M Ethernet (Gigabit Ethernet)
  - 10GE Ethernet
  - 100GE Ethernet
  - ...

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

10M/100M Ethernet

- The physical layer of the 10M/100M Ethernet is divided according to the rate level and transmission medium. Common interface types are as follows:
  - 10BASE-T: Uses two pairs of Category 3, 4, and 5 unshielded twisted pairs.
  - 100BASE-TX: Uses two pairs of shielded twisted pairs or high-quality Category 5 unshielded twisted pairs.
  - 100BASE-FX: Uses two optical fibers. One is used to transmit signals, and the other one is used to receive signals.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

1000M Ethernet

- Common interface types are as follows:
  - 1000BASE-T: Complies with IEEE 802.3ab and uses Category 5 unshielded twisted pairs.
  - 1000BASE-X: Complies with IEEE 802.3z and uses multi-mode optical fiber, single-mode optical fiber, and 150-ohm balanced shielded twisted pairs.
  - 1000BASE-CX: Uses transmission media that supports signal transmission within a maximum distance of 25 m, so it is rarely used.
  - 1000BASE-SX: Uses multi-mode optical fibers if the wavelength is 850 nm.
  - 1000BASE-LX: Uses single-mode and multi-mode optical fibers if the wavelength is 1310 nm.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

10GE Ethernet

- Standard: IEEE 802.3ae
- Common interface types are as follows:
  - 10GBASE-S/L/E-R/W in serial mode:
    - 10GBASE-S: uses multi-mode optical fibers if the wavelength is 850 nm.
    - 10GBASE-L: uses single-mode optical fibers if the wavelength is 1310 nm.
    - 10GBASE-E: uses single-mode optical fibers if the wavelength is 1550 m.
    - W = WAN PHY (physical layer on the WAN): The interface rate is 9.95328 Gbit/s and the interface is encapsulated using the SONET STS-192c and SDH VC-4-64C standards. The DWDM or SDH/SONET optical transmission network can be used to enable the 10G Ethernet to connect to the optical transmission network seamlessly.
    - R = LAN PHY (physical layer on the LAN): The interface rate is 10.3125 Gbit/s.
  - 10GBASE-LX4 with WDM in parallel mode (four wavelengths):
    - 10GBASE-LX4: uses multi-mode optical fibers if the wavelength is 1310 nm.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

100GE Ethernet

- In June, 2010, IEEE ratified IEEE 802.3ba after four years of working on 100 Gbit/s Ethernet technology standards.
- In the officially released 802.3ba standard, the following specifications of the physical layer are stipulated for the 100G MAC rate:
  - 40 km single-mode optical fiber (SMF): The corresponding PHY is 100GBASE-ER4, and the optical fiber consists of four 1310-nm WDM channels with 800 GHz channel spacing.
  - 10 km single-mode optical fiber: The corresponding PHY is 100GBASE-LR4, and the optical fiber consists of four 1310-nm WDM channels with 800 GHz channel spacing.
  - 10 m OM3 multi-mode optical fiber: The corresponding PHY is 100GBASE-SR10, and the optical fiber consists of 10 independent 850-nm multi-mode optical fiber channels.
  - 7 m copper cable: The corresponding PHY is 100GBASE-CR10, and the copper cable consists of 10 independent copper cable channels.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. **WAN Interfaces**
   - Evolution of WAN Interfaces on Routers
   - Ethernet Interface
   - POS Interface
   - Comparison Between Ethernet and POS Interfaces
2. Working Mechanism and Configuration of PPP
3. IP-Trunk
4. Working Mechanism and Configuration of PPPoE

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

POS Interface

- Packet over SONET/SDH (POS): POS is a technology that is used to transmit IP data packets and other data packets over a SONET/SDH network.
- The POS can map data packets of variable length directly into the payload of SONET. The POS uses the physical layer transmission standard of SONET, which offers a high-speed, reliable, and P2P data connection. Optical fibers are used for data transmission.
- Rate of common POS interfaces:
  - OC-3/STM-1: 155.52 Mbit/s
  - OC-12/STM-4: 622.08 Mbit/s
  - OC48/STM-16: 2488.32 Mbit/s
  - OC192/STM-64: 9953.28 Mbit/s
  - OC768/STM-256: 39813.12 Mbit/s
- Synchronous optical network (SONET) is the transmission standard of North America and Japan, and uses the format of OC-n.
- Synchronous digital hierarchy (SDH) is the transmission standard of Europe and China, and uses the format of STM-n.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Optical carrier level n (OC-n) is a unit of optical fiber transmission. The minimum unit is OC-1, and the data transmission rate is about 51.84 Mbit/s.
- Synchronous transport module (STM)

Contents

1. **WAN Interfaces**
   - Evolution of WAN Interfaces on Routers
   - Ethernet Interface
   - POS Interface
   - Comparison Between Ethernet and POS Interfaces
2. Working Mechanism and Configuration of PPP
3. IP-Trunk
4. Working Mechanism and Configuration of PPPoE

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Comparison Between Ethernet and POS Interfaces

- The rate of Ethernet and POS interfaces can reach 10 Gbit/s or 40 Gbit/s. After the 100G Ethernet standard is released, the rate of the Ethernet interface is higher than that of the POS interface.
- The POS interface uses PPP or HDLC as the Layer 2 encapsulation protocol to carry IP packets, and the maximum length of a Layer 2 header overhead is 9 or 7 bytes. 10G LAN and WAN interfaces use the Ethernet encapsulation protocol, and the length of a Layer 2 header overhead is 18 bytes. Therefore, the transmission efficiency of IP packets on the POS interface is higher.
- In terms of the cost, the price of the Ethernet interface or card is lower. For example, the price of a 10G Ethernet interface is approximately half of that of a 10G POS interface.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

LAN and WAN Modes of Interfaces

- A 10G XFP multi-mode optical transceiver works in either LAN or WAN mode. You can configure a proper mode as required.
- The difference between 10G LAN and WAN interfaces lies in the encapsulation mode. The 10G LAN interface uses only the Ethernet encapsulation protocol, and the 10G WAN interface uses the Ethernet encapsulation protocol at the link layer. However, the encapsulated data is directly mapped to the SDH sequence at the physical layer.
- The WAN interface can work in both LAN and WAN modes to implement long-distance transmission on the WAN. The interface modes of the two routers at both ends of the WAN must be the same.
- If there are intermediate devices between routers, pay attention to the interface mode of the routers. Generally, when the intermediate device uses a 10G POS interface, the router can use 10G POS and 10GE WAN interfaces for communication. When the intermediate device uses a 10G Ethernet interface, the router can only use the 10GE LAN interface. Note that new intermediate devices support both the 10GE LAN/WAN interface and 10G POS interface, and you only need to replace the software configuration.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

1. WAN Interfaces
2. **Working Mechanism and Configuration of PPP**
   - Basic Concepts of PPP
   - LCP
   - PAP Authentication
   - CHAP Authentication
   - NCP
   - MP Fundamentals
   - PPP and MP Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Basic Concepts of PPP - Three Components

- PPP works at the data link layer in the TCP/IP protocol stack and is the most widely used point-to-point link layer protocol.
- Three components of PPP:
  - Data encapsulation mode: defines how to encapsulate multi-protocol data packets.
  - Link Control Protocol (LCP): defines methods for establishing, negotiating, and testing connections at the data link layer.
  - Network Control Protocol (NCP): contains a set of protocols used to establish connections and negotiate parameters for different network layer protocols.

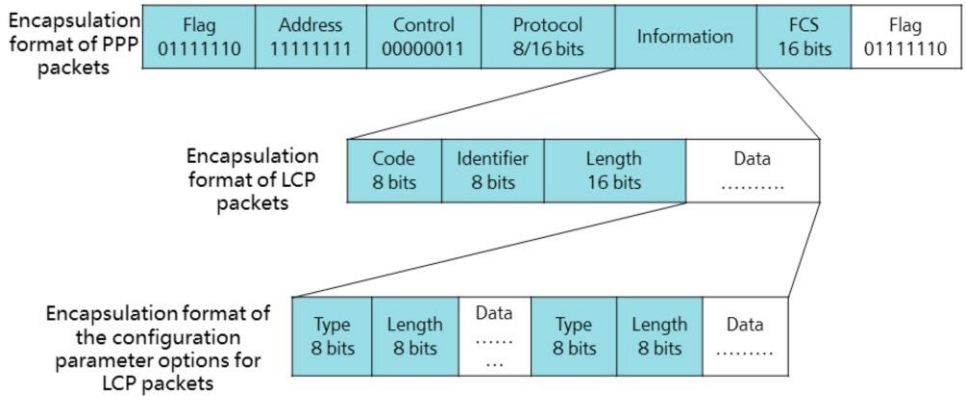
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- PPP defines three protocol components: data encapsulation mode, Link Control Protocol (LCP), and Network Control Protocol (NCP).
- The data encapsulation mode defines how to encapsulate multiple types of upper-layer protocol packets.
- PPP defines LCP so that LCP can be applied to various link types. LCP can automatically detect link environments (for example, detect whether a loop exists) and negotiate link parameters, such as the maximum packet length and authentication protocol. Compared with other data link layer protocols, PPP provides the authentication function. Both ends on a link negotiate the authentication protocol and implement the authentication. The connection is set up only after the authentication succeeds. With this function of PPP, carriers can allow the access of distributed users.
- PPP defines a group of NCPs. Each NCP corresponds to a network layer protocol and is used to negotiate parameters such as network layer addresses. For example, IPCP is used for IP address negotiation and control, and IPXCP is used for IPX negotiation and control.

Basic Concepts of PPP - Packet Structure

- PPP packet structure



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

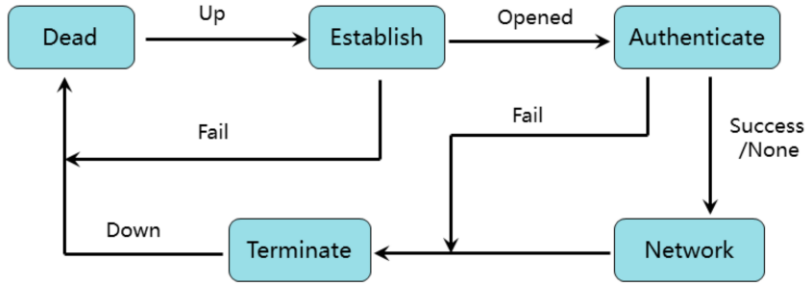
- Encapsulation format of a PPP packet
  - Flag field
    - The **Flag** field identifies the start and end of a physical frame and is always 0x7E.
  - Address field
    - The **Address** field identifies a peer. Two communication devices connected using PPP do not need to know the data link layer address of each other because PPP is used on P2P links. This field must be filled with a broadcast address of all 1s and is of no significance to PPP.
  - Control field
    - The default value of the **Control** field is 0x03, indicating an unsequenced frame. By default, PPP does not use sequence numbers or acknowledgement mechanisms to ensure reliable transmission.
    - The **Address** and **Control** fields identify a PPP packet, so the PPP packet header value is FF03.
  - Protocol field
    - The **Protocol** field identifies the datagram encapsulated in the **Information** field of a PPP data packet.

- Encapsulation format of an LCP packet
  - Code field
    - The **Code** field is 1 byte in length and identifies the LCP packet type.
  - Identifier field
    - The **Identifier** field is 1 byte in length and is used to match requests and replies. If a packet with an invalid **Identifier** field is received, the packet is discarded.
    - The sequence number of a Configure-Request packet usually begins with 0x01 and increases by 1 each time a Configure-Request packet is sent. After a receiver receives a Configure-Request packet, it must send a reply packet with the same sequence number as the received Configure-Request packet.
  - Length field
    - The **Length** field specifies the total number of bytes in the negotiation packet. It is the sum of the lengths of the **Code**, **Identifier**, **Length**, and **Data** fields.
    - The **Length** field value cannot exceed the MRU of the link. Bytes outside the range of the **Length** field are treated as padding and are ignored after they are received.
  - Data field
    - The **Type** field specifies the negotiation option type.

- The **Length** field specifies the total length of the **Data** field.
- The **Data** field contains the contents of the negotiation option.

Basic Concepts of PPP - Link Establishment Process

- PPP link establishment process



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Link establishment process
  - Dead: PPP starts and ends with the Dead phase. After the status of the physical layer becomes Up, PPP enters the Establish phase.
  - Establish: Devices perform LCP negotiation to negotiate link layer parameters in the Establish phase. If the negotiation fails, the PPP connection fails to be established and PPP returns to the Dead phase. If the negotiation succeeds, PPP enters the Authenticate phase.
  - Authenticate: Peer devices are authenticated in this phase. If the authentication fails, PPP enters the Terminate phase. If the authentication succeeds or no authentication is configured, PPP enters the Network phase.
  - Network: In this phase, devices use NCP to negotiate network layer parameters. If the negotiation succeeds, the PPP connection is successfully established and data packets at the network layer are transmitted. If the upper-layer application (for example, on-demand circuit) considers that the connection needs to be disabled or the administrator manually disables the PPP connection (Closing), the PPP enters the Terminate phase.
  - Terminate: LCP disables a PPP link in the Terminate phase. After the PPP link is disabled, PPP enters the Dead phase.
- Note: This part describes working phases of PPP, rather than the PPP protocol status. PPP is composed of a group of protocols. Therefore, PPP has no protocol status. Only specific protocols, such as LCP and NCP, have protocol status and status transition

(protocol status machine).



Contents

1. WAN Interfaces
2. **Working Mechanism and Configuration of PPP**
   - Basic Concepts of PPP
   - LCP
   - PAP Authentication
   - CHAP Authentication
   - NCP
   - MP Fundamentals
   - PPP and MP Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

LCP - Packet Type

Packet Type	Function Description
Configure-Request	Contains a list of LCP parameters with changes to default values.
Configure-Ack	Indicates that the values of all LCP parameters in a Configure-Request packet are accepted.
Configure-Nak	Indicates that the values of LCP parameters in a Configure-Request packet are invalid locally.
Configure-Reject	Indicates that LCP parameters in a Configure-Request packet cannot be identified locally.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- There are three types of LCP packets:
  1. Link configuration package is used to set up and configure links, including Configure-Request, Configure-Ack, Configure-Nak, and Configure-Reject packets.
  2. Link end package is used to end links, including Terminate-Request and Terminate-Ack packets.
  3. Link maintenance package is used to manage and commission links, including Code-Reject, Protocol-Reject, Echo-Request, Echo-Reply, and Discard-Request packets.

LCP - Parameters Used for Negotiation

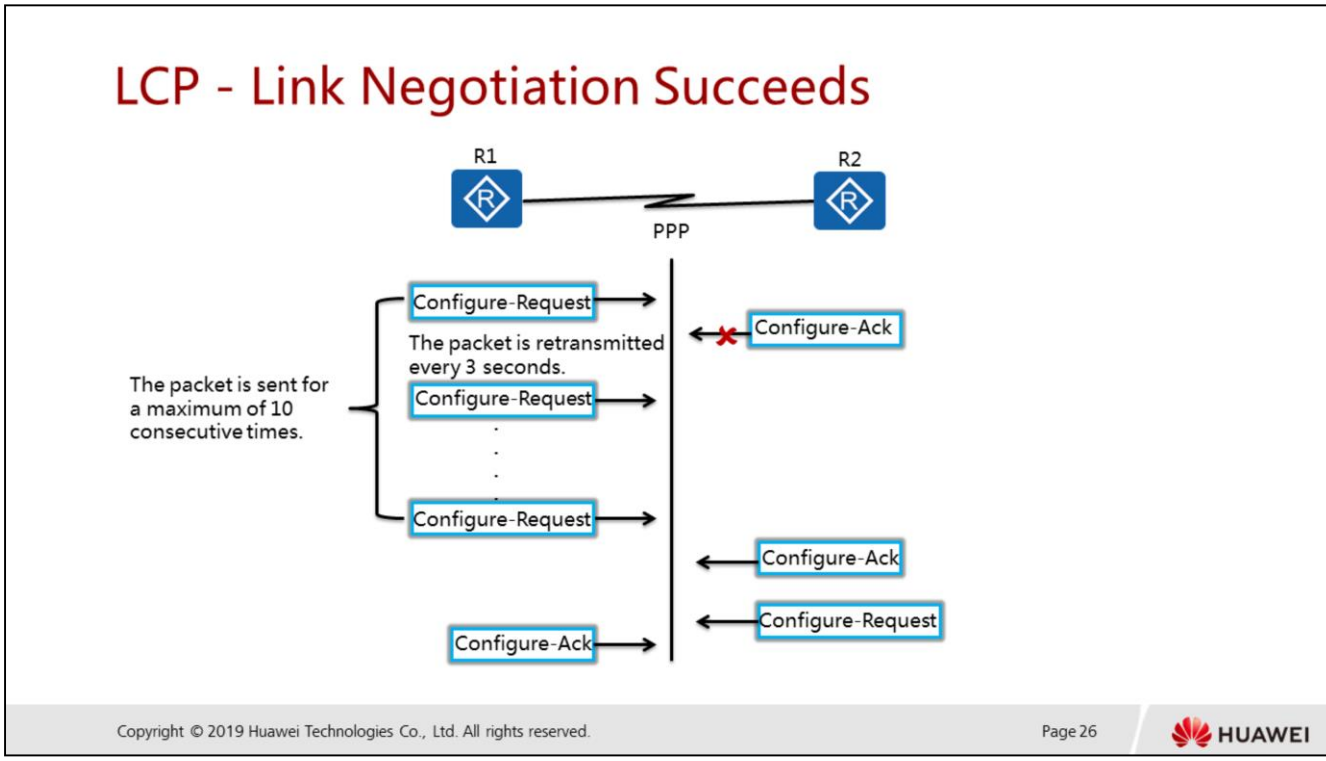
Parameter	Function Description	Negotiation Rule	Default Value
Maximum receive unit (MRU)	Indicates the total length of the **Information** field in a PPP data frame.	The smaller value of both ends is used.	1500
Authentication protocol	Authenticates the authentication protocol used by the peer end.	The authenticated device must support the authentication protocol used by the authenticating device and the protocol must be correctly configured. Otherwise, the negotiation fails.	Non-authentication
Magic number Magic-Number	Specifies a random number used to detect the link loop. If the magic number in the received LCP packet is the same as the local magic number, a loop occurs on the link.	If the magic number in the received LCP and the local magic number are different, no loop exists on the link and the negotiation succeeds. If magic numbers on both ends are the same, the detection mechanism is used to detect loops.	Enable

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

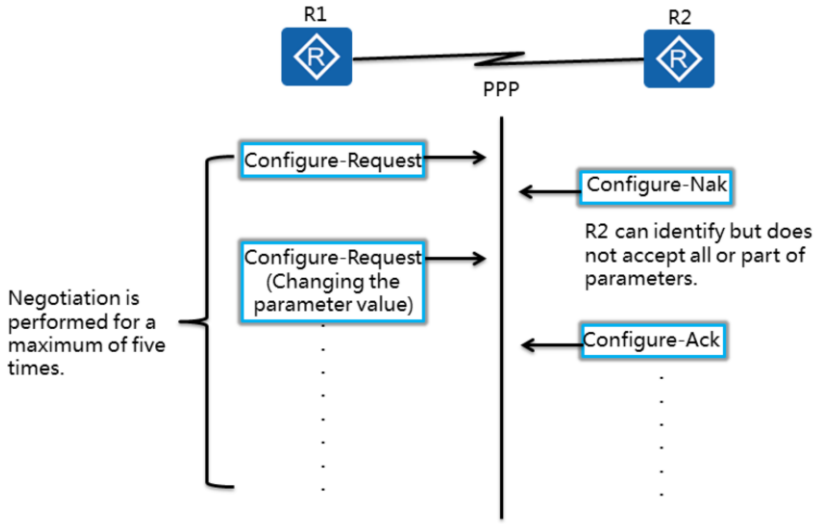
- Parameters used for negotiation
  - On the VRP, MRU is represented by the maximum transmission unit (MTU) configured on an interface.
  - The common PPP authentication protocols are PAP and CHAP. Devices at both ends of a PPP link can use different authentication protocols to authenticate the peer end. However, the device to be authenticated must support the authentication protocol used by the authenticator, and the authentication information such as the user name and password must be correctly configured.
  - LCP uses the magic number to detect link loops and other exceptions. The magic number is a random number, and the random mechanism must ensure that it is almost impossible that the two ends of a link generate the same magic number.
  - After receiving a Configure-Request packet, the system compares the magic number contained in the packet with the locally generated magic number. If they are different, no loop occurs on the link. In this case, the Configure-Ack packet is sent, indicating that the magic number and other parameters are successfully negotiated. If the subsequent packets contain the **Magic-Number** field, this field is set to the negotiated magic number. In this case, the LCP does not generate new magic numbers.
  - If the magic number contained in the Configure-Request packet is the same as the locally generated magic number, the system sends a Configure-Nak packet carrying a new magic number. LCP then sends a Configure-Request packet carrying a new magic number no matter whether the Configure-Nak packet carries the same magic

number. If a loop occurs on a link, the process will continue. If no loop occurs on the link, packet exchange will recover quickly.



- The link negotiation succeeds.
  - As shown in the figure, R1 and R2 are connected through serial links and run PPP. After the physical link becomes available, R1 and R2 use LCP to negotiate link parameters. In this example, R1 sends an LCP packet.
  - R1 sends R2 a Configure-Request packet that contains link layer parameters configured on the sender (R1). Each link layer parameter uses the structure of "type, length, value".
  - After receiving the Configure-Request packet, R2 returns a Configure-Ack packet to R1 if it can identify all link layer parameters in the packet and considers that the value of each parameter is acceptable.
  - If R1 does not receive the Configure-Ack packet, R2 retransmits the Configure-Request packet every 3s. If R1 does not receive the Configure-Ack packet after R2 sends the Configure-Request packet for 10 consecutive times, R2 considers that the peer end is unavailable and stops sending the Configure-Request packet.
- Note: The preceding process shows that R2 considers that the link parameter settings on R1 are acceptable. R2 also needs to send a Configure-Request packet to R1 to check whether the link parameter settings on R2 are acceptable.

LCP - Link Parameter Negotiation Fails

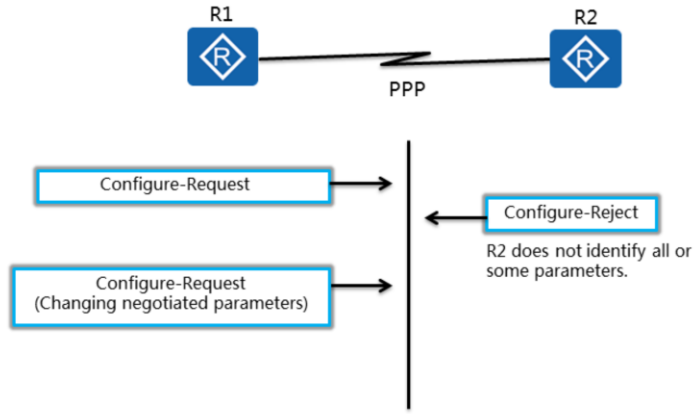


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Link parameter negotiation fails.
  - After receiving the Configure-Request packet sent from R1, R2 needs to send a Configure-Nak packet to R1 if it can identify all link layer parameters carried in the packet but considers that the values of some or all parameters are unacceptable, that is, parameter negotiation fails.
  - The Configure-Nak packet contains only unacceptable link layer parameters. The value of each link layer parameter in the packet is changed to the value (or value range) that can be accepted by the sender (R2).
  - After receiving the Configure-Nak packet, R1 reselects local parameters according to link layer parameters in the packet, and sends a new Configure-Request packet.
  - The parameters that fail to be negotiated for five consecutive times are disabled and no further negotiation is performed.

LCP - Link Negotiation Parameters Cannot be Identified

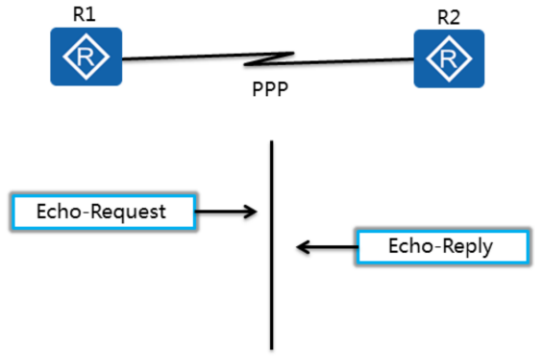


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Negotiated link parameters cannot be identified.
  - After receiving the Configure-Request packet sent from R1, R2 needs to return a Configure-Reject packet to R1 if R2 cannot identify some or all link layer parameters carried in the packet.
  - The Configure-Reject packet contains only the list of link layer parameters that are not identified.
  - After receiving the Configure-Reject packet, R1 needs to send a new Configure-Request packet to R2. The Configure-Request packet does not contain the parameters that are not identified by the peer (R2).

LCP - Checking Link Status

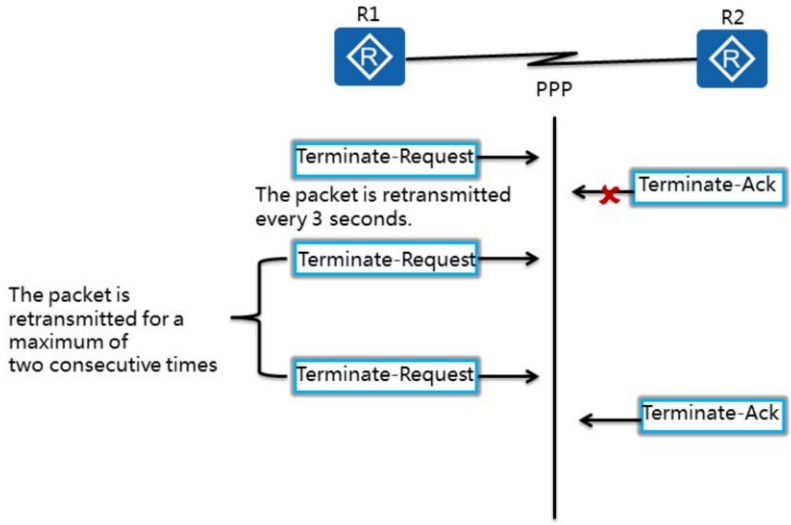


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Checking the link status
  - After the LCP connection is established, the Echo-Request and Echo-Reply packets can be used to detect the link status. After receiving an Echo-Request packet, the device responds with an Echo-Reply packet, indicating that the link is normal.
  - By default, the VRP platform sends an Echo-Request packet every 10 seconds.

LCP - The Connection Is Disabled



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The connection is disabled.
  - If the authentication fails or the administrator manually disables the connection, the established LCP connection may be disabled.
  - The Terminate-Request and Terminate-Ack packets are used in disabling the LCP connection. The Terminate-Request packet is used to request the peer end to disable the connection. Once a Terminate-Request packet is received, the LCP must respond with a Terminate-Ack packet to confirm that the connection is disabled.
  - If no Terminate-Ack packet is received, the Terminate-Request packet is retransmitted every three seconds. If no Terminate-Ack packet is received for two consecutive times, the peer end is considered unavailable and the connection is disabled.



Contents

1. WAN Interfaces
2. **Working Mechanism and Configuration of PPP**
   - Basic Concepts of PPP
   - LCP
   - PAP Authentication
   - CHAP Authentication
   - NCP
   - MP Fundamentals
   - PPP and MP Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

PAP Authentication - Packet Type

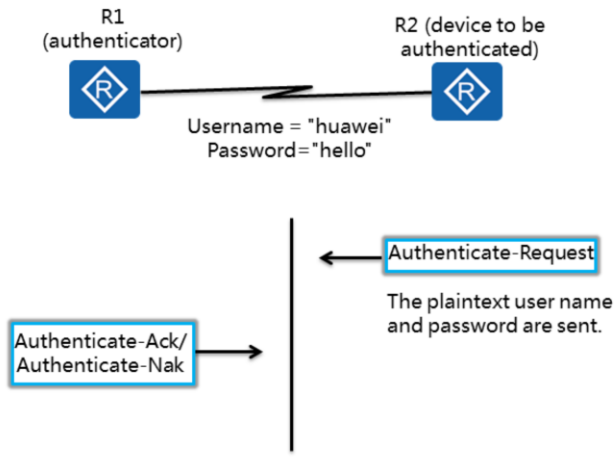
Packet Type	Function Description
Authenticate-Request	The device to be authenticated sends an Authenticate-Request message with the user name and password. The **Data** field contains the plaintext user name and password.
Authenticate-Ack	The authenticator sends the authentication success message. The **Data** field contains text information.
Authenticate-Nak	The authenticator sends the authentication failure message. The **Data** field contains text information.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The PAP packet is encapsulated in a PPP packet.

PAP Authentication - Implementation



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- PAP mode
  - The device to be authenticated sends the configured plaintext user name and password to the authenticator through Authenticate-Request packets. In this example, the user name is huawei and the password is hello.
  - After receiving the user name and password sent by the peer, the authenticator checks whether the user name and password are correct according to the locally configured database. If they are matched, the authenticator returns an Authenticate-Ack packet, indicating that the authentication succeeds. If they are not matched, the authenticator returns an Authenticate-Nak packet, indicating that the authentication fails.

Contents

1. WAN Interfaces
2. **Working Mechanism and Configuration of PPP**
   - Basic Concepts of PPP
   - LCP
   - PAP Authentication
   - CHAP Authentication
   - NCP
   - MP Fundamentals
   - PPP and MP Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

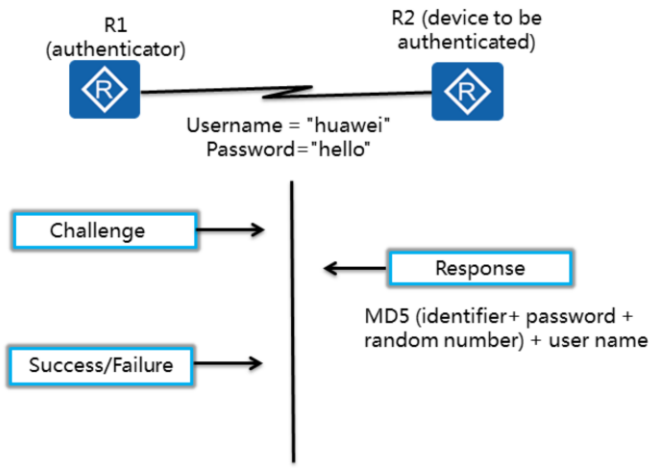
CHAP Authentication - Packet Type

Packet Type	Function Description
Challenge	The authenticator sends a Challenge message to the peer.
Response	The peer returns the user information through a Response message. The **Data** field contains the user name and encrypted password.
Success	If the authentication is successful, the authenticator sends a Success message to the peer. The **Data** field can contain text information.
Failure	If the authentication fails, the authenticator sends a Failure message to the peer. The **Data** field can contain text information.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

CHAP Authentication - Implementation



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- CHAP authentication requires three packet exchanges. The Identifier field in the packet is required to match the request and response packets, and the packets used in the authentication process use the same value of the Identifier field. Unidirectional CHAP authentication is applicable to two scenarios: the authenticator is configured with a user name and the authenticator is not configured with a user name. It is recommended that the authenticator be configured with a user name.
- If the authenticator is configured with a user name (the ppp chap user username command is configured on the interface), the authentication process is as follows:
  - The authenticator initiates an authentication request by sending Challenge packets that carry the local user name to the device to be authenticated.
  - After receiving the authentication request from the authenticator, the device to be authenticated checks whether the ppp chap password command is configured on the local interface. If the command is configured, the device sends the generated ciphertext (which is generated based on the identifier, password, and random number using MD5 algorithm) and its user name to the authenticator (Response). If the ppp chap password command is not configured on the interface, the device to be authenticated searches for the password in the local user table based on the user name of the authenticator, and sends the ciphertext (which is generated based on the identifier, password, and random number using MD5 algorithm) and the user name of the peer to the authenticator (Response).
  - The authenticator encrypts the saved password, identifier, and random number using the MD5 algorithm, and compares the encrypted value with the ciphertext in

the received response packet to check whether the authentication is correct.

- If the authenticator is not configured with a user name (the ppp chap user username command is not configured on the interface), the authentication process is as follows:
    - The authenticator initiates an authentication request by sending Challenge packets.
    - After receiving the Challenge packet, the device to be authenticated encrypts the Challenge packet with the identifier, password, and random number configured by the ppp chap password command using the MD5 algorithm to generate a ciphertext. It then sends a Response packet carrying the ciphertext and local user name to the authenticator.
    - The authenticator encrypts the saved password, identifier, and random number using the MD5 algorithm, and compares the encrypted value with the ciphertext in the received response packet to check whether the authentication is correct.

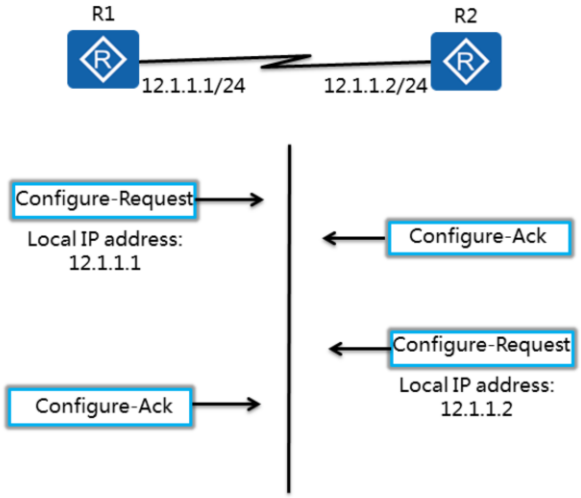
Contents

1. WAN Interfaces
2. **Working Mechanism and Configuration of PPP**
   - Basic Concepts of PPP
   - LCP
   - PAP Authentication
   - CHAP Authentication
   - NCP
   - MP Fundamentals
   - PPP and MP Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

NCP - Static Address Negotiation Using IPCP

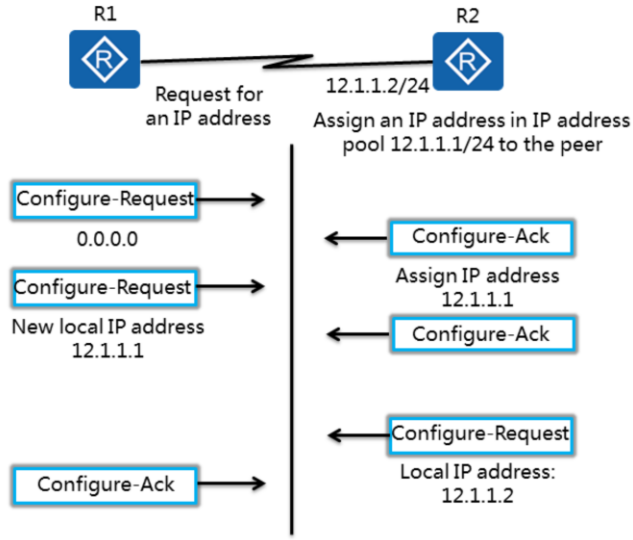


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IPCP is used for IP parameter negotiation so that PPP can be used to transmit IP packets.
- IPCP and LCP use the same negotiation mechanism and packet type. However, IPCP does not invoke LCP but only has the same working process and packet type as LCP.
  - The IP addresses of both ends are 12.1.1.1/24 and 12.1.1.2/24. If the IP addresses of both ends are not in the same network segment, IPCP negotiation is performed.
  - When IP addresses are statically configured at both ends, the negotiation process is as follows:
    - R1 and R2 send Configure-Request packets that contain the locally configured IP address.
    - After receiving the Configure-Request packet from the peer end, R1 and R2 check the IP address in the packet. If the IP address is a valid unicast IP address and is different from the locally configured IP address (no IP address conflict), the peer end can use this address and responds with a Configure-Ack packet.
    - Both ends on a PPP link obtain the 32-bit IP address used by the peer end from the message sent through IPCP.

NCP - Dynamic Address Negotiation Using IPCP



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- As shown in the figure, R1 is configured to request an IP address to the peer end. R2 is configured with an IP address pool 12.1.1.2/24, and is enabled to assign an IP address to the peer end.
- The two ends dynamically negotiate IP addresses as follows:
  - R1 sends a Configure-Request packet that contains IP address 0.0.0.0 to R2. The Configure-Request packet containing IP address 0.0.0.0 indicates that the local end requests an IP address from the peer end.
  - After receiving the Configure-Request packet, R2 considers that IP address 0.0.0.0 contained in the packet is invalid and sends a Configure-Nak packet with new IP address 12.1.1.1.
  - After receiving the Configure-Nak packet, R1 updates the local IP address and sends a new Configure-Request packet containing the new IP address 12.1.1.1.
  - After receiving the Configure-Request packet, R2 considers that the IP address contained in the packet is a valid IP address and returns a Configure-Ack packet.
  - In addition, R2 sends a Configure-Request packet to R1 to request IP address 12.1.1.2. R1 considers that the IP address is valid and returns a Configure-Ack packet.

Contents

1. WAN Interfaces
2. **Working Mechanism and Configuration of PPP**
   - Basic Concepts of PPP
   - LCP
   - PAP Authentication
   - CHAP Authentication
   - NCP
   - MP Fundamentals
   - PPP and MP Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

MP Fundamentals

- MP Purpose
  - To increase bandwidth by bundling multiple PPP links into an MP group.
- Link negotiation in MP mode
  - In the LCP phase, whether the peer interface works in MP mode is also checked.
  - In the NCP phase, devices at both ends perform NCP negotiation by using NCP parameters (such as IP addresses) of the MP group interface or virtual template interface.
- Implementation mode
  - Virtual template interface
  - MP group interface

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Multilink PPP (MP) supports packet fragmentation, and fragmented packets are transmitted to the same destination through multiple point-to-point links.



Contents

1. WAN Interfaces
2. **Working Mechanism and Configuration of PPP**
   - Basic Concepts of PPP
   - LCP
   - PAP Authentication
   - CHAP Authentication
   - NCP
   - MP Fundamentals
   - PPP and MP Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Configuring PPP

- On the interfaces connecting R1 and R2, set the encapsulation type to PPP, configure CHAP authentication, and set the user name to **Huawei** and the password to **Hello**.




```
aaa
 local-user Huawei password cipher Hello
 local-user Huawei service-type ppp
#
interface Serial1/0/0
 link-protocol ppp
 ppp authentication-mode chap
 ppp chap user Huawei
 ip address ppp-negotiate
```

```
interface Serial1/0/0
 link-protocol ppp
 remote address 12.1.1.1
 ppp chap user Huawei
 ppp chap password cipher Hello
 ip address 12.1.1.2 255.255.255.0
```

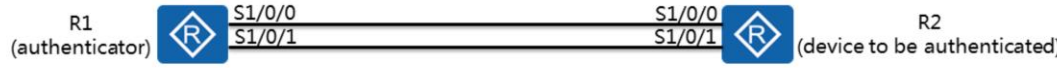
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Command usage
  - The **ppp authentication-mode** command configures the PPP authentication mode in which the local device authenticates the remote device.
  - The **ppp chap user** command configures a user name for CHAP authentication.
  - The **ppp chap password** command configures a password for CHAP authentication.
  - The **ip address ppp-negotiate** command configures IP address negotiation on an interface to allow the interface to obtain an IP address from the remote end.
  - The **remote address** command configures the local device to assign an IP address or specify an IP address pool for the remote device.
- ppp authentication-mode { chap | pap }
  - chap: Sets the PPP authentication mode to Challenge Handshake Authentication Protocol (CHAP) authentication.
  - ppp: Sets the PPP authentication mode to Password Authentication Protocol (PAP) authentication.
- ppp chap user username
  - username：username: Sets the user name in CHAP authentication.
- ppp chap password { cipher | simple } password
  - cipher: Displays the password in cipher text.

- simple: Displays the password in plain text.
- password: Sets the password for CHAP authentication.

Configuring MP

- To increase the interface bandwidth, bundle all the interconnected PPP interfaces between R1 and R2 into an MP group. Configure PAP authentication between R1 and R2, configure R1 as the authenticator and R2 as the device to be authenticated, and set the user name and password to **Huawei** and **Hello**, respectively.




```
aaa
 local-user Huawei password cipher Hello
 local-user Huawei service-type ppp
#
interface Mp-group0/0/0
 ip address 12.1.1.1 255.255.255.0
#
interface Serial1/0/0
 link-protocol ppp
 ppp authentication-mode pap
 ppp mp Mp-group 0/0/0
#
interface Serial1/0/1
 link-protocol ppp
 ppp authentication-mode pap
 ppp mp Mp-group 0/0/0
```

```
interface Mp-group0/0/0
 ip address 12.1.1.2 255.255.255.0
#
interface Serial1/0/0
 link-protocol ppp
 ppp pap local-user Huawei password simple Hello
 ppp mp Mp-group 0/0/0
#
interface Serial1/0/1
 link-protocol ppp
 ppp pap local-user Huawei password simple Hello
 ppp mp Mp-group 0/0/0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Command usage
  - The **interface mp-group** command creates an MP-group interface and displays the MP-group interface view.
  - The **ppp mp mp-group** command binds an interface to the MP-group interface so that the interface works in MP mode.
  - The **restart** command restarts an interface.



Contents

1. WAN Interfaces
2. Working Mechanism and Configuration of PPP
3. **IP-Trunk**
4. Working Mechanism and Configuration of PPPoE

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IP-Trunk

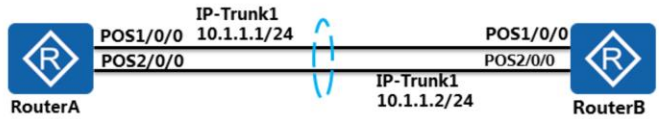
- Trunk interfaces are classified into Eth-Trunk and IP-Trunk.
  - An Eth-Trunk consists of only Ethernet links.
  - An IP-Trunk consists of POS interfaces.
- An IP-Trunk can implement load balancing in either of the following modes:
  - Flow-based load balancing: Packets with the same source and destination IP addresses are transmitted over the same member link.
  - Packet-based load balancing: Packets are transmitted over different member links.
- Member interfaces of an IP-Trunk must be encapsulated using HDLC. The working mechanism of IP-Trunk is similar to that of Eth-Trunk.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Trunk interfaces are classified into Eth-Trunk and IP-Trunk.
  - An Eth-Trunk consists of only Ethernet links.
  - An IP-Trunk consists of POS interfaces.

IP-Trunk Configuration




```
#
sysname RouterA
#
interface Ip-Trunk1
 undo shutdown
 ip address 10.1.1.1 255.255.255.0
#
interface Pos1/0/0
 link-protocol hdlc
 undo shutdown
 ip-trunk 1
#
interface Pos2/0/0
 link-protocol hdlc
 undo shutdown
 ip-trunk 1
#
```

```
#
sysname RouterB
#
interface Ip-Trunk1
 undo shutdown
 ip address 10.1.1.2 255.255.255.0
#
interface Pos1/0/0
 link-protocol hdlc
 undo shutdown
 ip-trunk 1
#
interface Pos2/0/0
 link-protocol hdlc
 undo shutdown
 ip-trunk 1
#
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



Contents

1. WAN Interfaces
2. Working Mechanism and Configuration of PPP
3. IP-Trunk
4. **Working Mechanism and Configuration of PPPoE**
   - Working Mechanism of PPPoE
   - PPPoE Configurations

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

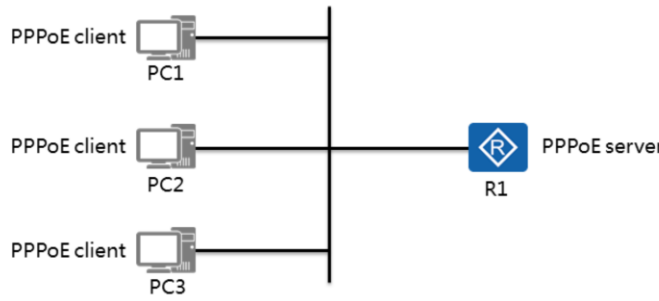
Overview of PPPoE (1)

- PPP over Ethernet (PPPoE) is a link layer protocol that encapsulates PPP frames into Ethernet frames. PPPoE enables multiple hosts on an Ethernet to connect to a remote broadband access server.
- Carriers want to connect multiple hosts at a site to the same remote access device. In addition, the access device provides access control and accounting for these hosts in a manner similar to dial-up access. Ethernet is the most cost-effective technology among all access technologies that connect multiple hosts to an access device. Moreover, PPP provides the access control and accounting functions. Therefore, PPPoE emerges and effectively achieves the goal of carriers.
- PPPoE allows a large number of hosts on an Ethernet to connect to the Internet using a remote access device and controls each host using PPP. PPPoE features a large application scale, high security, and convenient accounting.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Overview of PPPoE (2)

- PPP is applied to an Ethernet to connect multiple hosts on a broadcast network to a remote access server.
- PPPoE uses the client/server model.

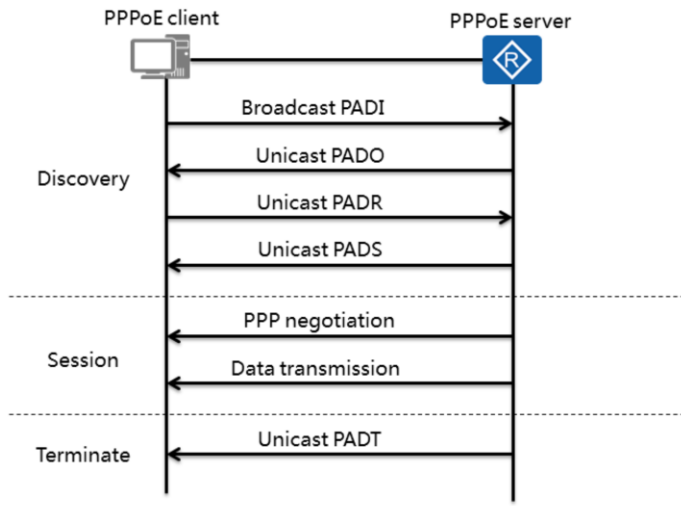


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Overview of PPPoE
  - PPPoE allows a large number of hosts on an Ethernet to connect to the Internet using a remote access device and controls each host using PPP. PPPoE features a large application scale, high security, and convenient accounting.

Process of Establishing a PPPoE Session



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Pepo involves three stages: Discovery stage, Session stage, and Terminate stage.
- Discovery stage
  - A PPPoE client broadcasts a PPPoE Active Discovery Initial (PADI) packet that contains service information required by the PPPoE client.
  - After receiving the PADI packet, all PPPoE servers compare the requested service with the services they can provide. The PPPoE servers that can provide the requested service sends PPPoE Active Discovery Offer (PADO) packets to the PPPoE client in unicast mode.
  - The PPPoE client may receive PADO packets from multiple PPPoE servers. The PPPoE client selects one from these PPPoE servers and sends a PPPoE Active Discovery Request (PADR) packet to the selected PPPoE server in unicast mode.
  - The PPPoE server generates a unique session ID to identify the PPPoE session with the PPPoE client, and then sends a PPPoE Active Discovery Session-confirmation (PADS) packet containing this session ID to the PPPoE client. When the PPPoE session is established, the PPPoE server and PPPoE client enter the PPPoE session stage.
  - After the PPPoE session is established, the PPPoE server and PPPoE client share the unique PPPoE session ID and learn peer MAC addresses.
- Session stage
  - PPP negotiation process at the PPPoE Session stage is the same as common PPP negotiation process. After PPP negotiation succeeds, PPP data packets can be forwarded over the established PPP link. At the PPPoE session stage, the PPPoE

server and client send all Ethernet data packets in unicast mode.

- Terminate stage
  - In the session stage, the PPPoE client and server exchange PADT packets to disable the PPPoE connection. The PADT packets can be sent in unicast mode anytime after a session is established. When a PADT packet is received, no PPP traffic can be sent using this session.



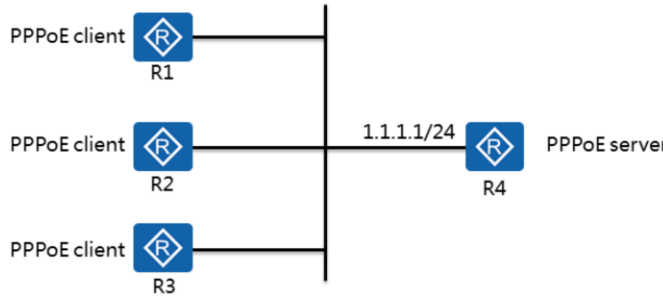
Contents

1. WAN Interfaces
2. Working Mechanism and Configuration of PPP
3. IP-Trunk
4. **Working Mechanism and Configuration of PPPoE**
   - Working Mechanism of PPPoE
   - PPPoE Configurations

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuring PPPoE (1)

- Company A wants to deploy PPPoE and has the following requirements:
  - R4 functions as the PPPoE server to assign IP addresses in the IP address pool 1.1.1.0/24 to clients, and uses PAP authentication.
  - R1 functions as a PPPoE client, and uses **HuaweiR1** and **R1** as the user name and password for authentication.
  - R2 functions as a PPPoE client, and uses **HuaweiR2** and **R2** as the user name and password for authentication.
  - R3 functions as a PPPoE client, and uses **HuaweiR3** and **R3** as the user name and password for authentication.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuring PPPoE (2)

- R1 configuration:

```
interface Dialer1
 link-protocol ppp
 ppp pap local-user HuaweiR1 password cipher
R1
 ip address ppp-negotiate
 dialer user HuaweiR1
 dialer bundle 1
 dialer-group 1
#
interface GigabitEthernet0/0/0
 pppoe-client dial-bundle-number 1
#
dialer-rule
 dialer-rule 1 ip permit
```

- R4 configuration:

```
ip pool POOL_1
 gateway-list 1.1.1.1
 network 1.1.1.0 mask 255.255.255.0
#
aaa
 local-user HuaweiR1 password cipher R1
 local-user HuaweiR2 password cipher R2
 local-user HuaweiR3 password cipher R3
#
interface Virtual-Template0
 ppp authentication-mode pap
 remote address pool POOL_1
 ip address 1.1.1.1 255.255.255.0
#
interface GigabitEthernet0/0/0
 pppoe-server bind Virtual-Template 0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In this example, R1 simulates as the PPPoE client to perform PPPoE dialup to access the Internet. R4 functions as the PPPoE server to authenticate and assign addresses.

Section Summary

- Common WAN Interfaces
- Ethernet Interface
- POS Interfaces/IP-Trunk
- PPP/MP
- PPPoE

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— IPv6 Protocol Basics

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- As is known to all, the shortage of IPv4 address resources restricts the further development of IP technologies. We urgently need a technology that can replace IPv4. This technology not only provides IPv4 functions, but also meets the IP address requirements of future industries. All industries, from government to market, are in pressing need of next-generation Internet technologies, and promote the emergence and development of the IPv6 technology, which can fundamentally solve this problem.
- This course focuses on IPv6 basics, including the background, packet format, address classification, basic protocol, and transition technologies.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- On completion of this course, you will be able to:
  - Describe the background of the IPv6 technology.
  - Distinguish IPv6 addresses.
  - Understand the IPv6 basic protocol.
  - Describe the IPv6 transition technologies.
  - Master the IPv6 configuration.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. **IPv6 Background**
2. IPv6 Implementation
3. IPv6 Configuration Commands
4. Suggestions for IPv6 Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

International IP Address Assignment Mode

Internet Assigned Number Authority (IANA)

The IANA department of Internet Corporation for Assigned Names and Numbers (ICANN) assigns IP addresses to five regional Internet registration organizations, such as APNIC, which is responsible for IP address assignment in the Asia Pacific region.

Regional Internet Registry (RIR)

RIR assigns IP addresses to local ISPs, for example, China Telecom and China Mobile.

Internet Service Provider (ISP)

An ISP assigns IP addresses to organizations or users. For example, an ISP assigns Class A IPv4 addresses to a super large organization and Class C IPv4 addresses to a medium-sized enterprise.



Five RIRs:
- Réseaux IP Européens (RIPE): European IP address registration center
- Latin American and Caribbean Internet Address Registry (LACNIC)
- American Registry for Internet Numbers (ARIN)
- Africa Internet Network Information Center (AFRINIC)
- Asia-Pacific Network Information Center (APNIC)

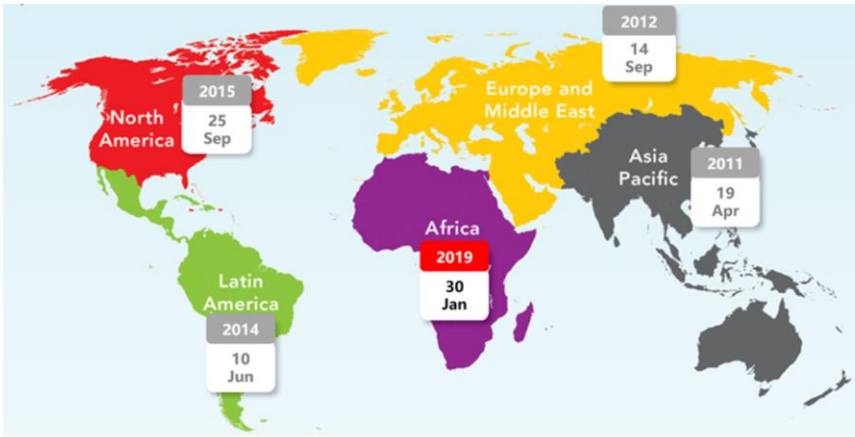
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- ICANN is the central management organization of the Internet.
- Not all addresses will be assigned. Some addresses are reserved for broadcasting, testing, and private networks. These addresses are called special-use addresses. You can query RFC5735 to know which addresses are special-use addresses.

Status Quo of International IPv4 Address Assignment

- On February 3, 2011, IANA announced to evenly assign the last 4.68 million IP addresses to Internet information centers in five regions around the world. Since then, no IPv4 addresses can be assigned.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- It has been proven that IPv4 is a very successful protocol, which stands the test of the Internet that a small number of computers evolve into the interconnection among hundreds of millions of computers. The protocol, however, was designed dozens of years ago based on the network scale at that time. From today's perspectives, designers of IPv4 did not make a full estimation on the Internet. With the expansion of the Internet and the emergence of new applications, IPv4 shows more of its limitations.
- The rapid expansion of the Internet scale is beyond people's expectation. Especially over the past decade, it has been increasing explosively. The Internet has connected thousands of households and penetrated into people's daily life. However, such rapid development brings about an urgent problem of IP address exhaustion.

Challenges to IPv4 Address Assignment

- IPv4 public address exhaustion. This is the biggest driving force for IPv6 to replace IPv4.
- Increasing number of Internet users. As technology industries develop, more users and various types of devices are connected to public networks.
- A lack of end-to-end communication model. NAT can solve the problem of mutual access between private address spaces and public networks, but destroys the integrity of end-to-end communication.
- Inability to adapt to new technologies, such as Internet of Things (IoT). All industries are potential IPv6 users.
- Broadcast mechanism and dependency on ARP, leading to frequent occurrence of operation problems related to the IPv4 LAN.
- Inadequate support for mobility.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Temporary Countermeasures

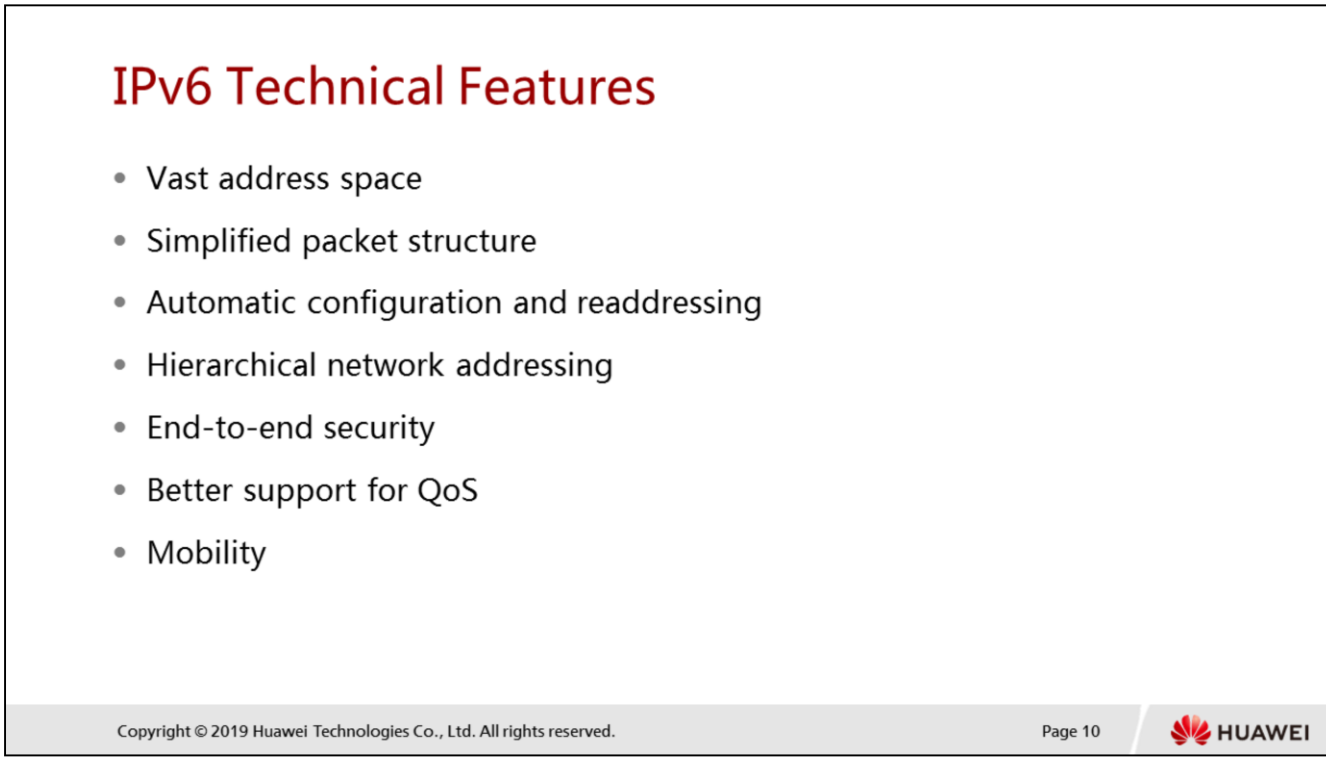
- In 1991, Internet Engineering Task Force (IETF) launched the classified network solution to delay the time when IPv4 addresses were exhausted.
- In 1993, network address translation (NAT) and classless inter-domain routing (CIDR) were launched.
- However, these transition schemes cannot prevent the occurrence of address exhaustion, but only slow down its speed. The problem fails to be solved fundamentally.
- IETF proposed the next-generation Internet protocol (IPv6) in the 1990s. Currently, IPv6 has become an acknowledged upgrade version of IPv4 in the future.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. IPv6 Background
2. **IPv6 Implementation**
   - **Packet Format**
   - Address Classification
   - Basic Protocol
   - Transition Technologies
3. IPv6 Configuration Commands
4. Suggestions for IPv6 Exam Preparation

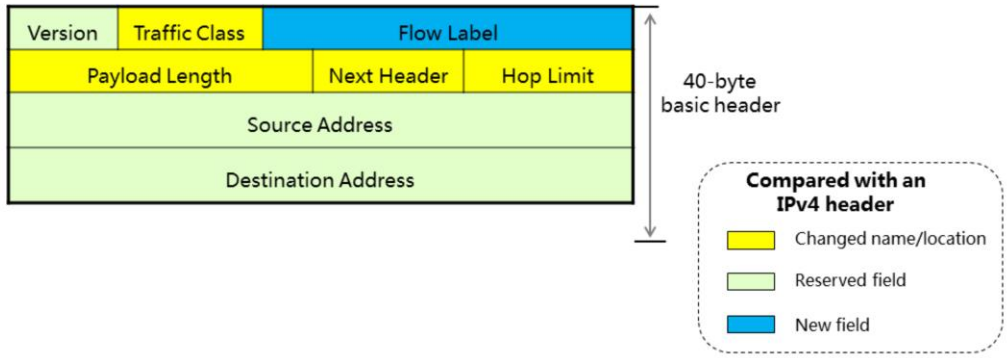
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- IPv6 features:
  - Vast address space. IPv6 addresses are 128 bits long. A 128 bit structure allows for an address space of 2^{128} (4.3 billion x 4.3 billion x 4.3 billion x 4.3 billion) possible addresses. This vast address space makes it very unlikely that IPv6 address exhaustion will ever occur.
  - Simplified packet structure. IPv6 uses a new protocol header format. That is, an IPv6 packet has a new header instead of simply expanding the address in the IPv4 packet header to 128 bits long. An IPv6 packet header includes a fixed header and extension headers. Some non-fundamental and optional fields are moved to extension headers following the fixed header. This improves the efficiency for intermediate routers in the network to process IPv6 protocol headers.
  - Automatic configuration and readdressing. IPv6 supports automatic address configuration to enable hosts to automatically discover networks and obtain IPv6 addresses, greatly improving the manageability of internal networks.
  - Hierarchical network architecture. The vast address space allows for the hierarchical network design in IPv6 to facilitate route summarization and improve forwarding efficiency.
  - End-to-end security. IPv6 supports IP Security (IPsec) authentication and encryption at the network layer, providing end-to-end security.
  - Better support for QoS. IPv6 defines a special field called flow label in the packet header. The IPv6 flow label field enables routers on a network to identify packets of the same data flow and provide special processing. Using this label, a router can identify a data flow without parsing the inner data packets. This ensures the support for QoS even if the payload of data packets is encrypted.
  - Mobility. Because extension headers such as Routing header and Destination option header are used, IPv6 provides built-in mobility.

IPv6 Packet Format — Basic Header

- Packet format
  - An IPv6 packet contains an IPv6 basic header, multiple IPv6 extension headers, and an upper-layer protocol data unit.
  - An IPv6 basic header has eight fields with a fixed size of 40 bytes. Each IPv6 data packet must contain a basic header.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

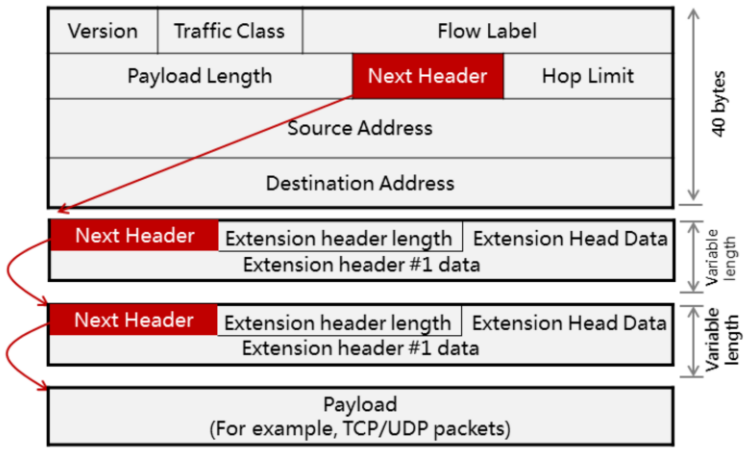
- An IPv6 packet consists of an IPv6 header, multiple extension headers, and an upper-layer protocol data unit (PDU).
- IPv6 header
  - Each IPv6 packet must contain a header with a fixed length of 40 bytes.
  - An IPv6 header provides basic packet forwarding information, and will be parsed by all routers on the forwarding path.
- Upper-layer PDU
  - An upper-layer PDU is composed of the upper-layer protocol header and its payload, which maybe an ICMPv6 packet, a TCP packet, or a UDP packet.
- An IPv6 header contains the following fields:
  - Version: 4 bits long. In IPv6, the value of the Version field is set to 6.
  - Traffic Class: 8 bits long. This field indicates the class or priority of an IPv6 packet. The Traffic Class field is similar to the TOS field in an IPv4 packet and is mainly used in QoS control.
  - Flow Label: 20 bits long. This field was added in IPv6 to differentiate real-time traffic. A flow label and source IP address identify a data flow. Intermediate network devices can effectively differentiate data flows based on this field.
  - Payload Length: 16 bits long. This field indicates the length of the IPv6 payload in bytes. The payload is the part of the IPv6 packet following the IPv6 basic header, including the extension header and upper-layer PDU.
  - Next Header: 8 bits long.
  - Hop Limit: 8 bits long. This field is similar to the Time to Live field in an IPv4 packet, defining the maximum number of hops that an IP packet can pass through. The value is decreased by 1 on each router that forwards the packet. The packet is discarded if Hop

Limit is decreased to 0.

- Source Address: 128 bits long. This field indicates the address of the packet originator.
- Destination Address: 128 bits long. This field indicates the address of the packet recipient.

IPv6 Packet Format — Extension Header

- The extension header is optional. The data sender adds the corresponding extension header only when the function of the extension header is required.



- **Next Header: Indicates the next header, with the length of 8 bits.** The function of this parameter is the same as that of Next Header in the basic header, specifying the next extension header (if any) or the type of the upper-layer protocol.
- **Extension Header Length: Indicates the length of an extension header, which is 8 bits** (excluding the Next Header field).
- **Extension Header Data: Indicates the extension header data, with a variable length.** This includes a series of options and padding fields.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An IPv4 packet header has an optional field (Options), which includes security, timestamp, and record route options. The variable length of the Options field makes the IPv4 packet header length range from 20 bytes to 60 bytes. When routers forward IPv4 packets with the Options field, many resources need to be used. Therefore, these IPv4 packets are rarely used in practice.
- To improve packet processing efficiency, IPv6 uses extension headers to replace the Options field in the IPv4 header. Extension headers are placed between the IPv6 basic header and upper-layer PDU. An IPv6 packet may carry zero or more extension headers. The sender of a packet adds one or more extension headers to the packet only when the sender requests routers or the destination device to perform special handling. Unlike IPv4, IPv6 has variable-length extension headers, which are not limited to 40 bytes. This facilitates further extension. To improve extension header processing efficiency and transport protocol performance, IPv6 requires that the extension header length be an integer multiple of 8 bytes.
- When multiple extension headers are used, the Next Header field of an extension header indicates the type of the next header following this extension header.

IPv6 Packet Format — Types of Extension Headers

- When an IPv6 packet uses more than one extension header, the headers must be listed in the following order:

Header Type	Next Header Field Value	Description
Hop-by-Hop Options header	0	This option specifies sending parameters for each hop on a transmission path. Each intermediate node on a transmission path needs to read and process the field. The application scenarios include: Jumbo payload Router prompt Resource reservation
Destination Options header	60	This header carries information that only the destination node of a packet processes. Currently, this header is used in mobile IPv6.
Routing header	43	An IPv6 source node uses this header to specify the intermediate routers that a packet must pass through on the way to its destination. This option is similar to the Loose Source and Record Route option in IPv4.
Fragment header	44	Like IPv4 packets, the length of IPv6 packets to be forwarded cannot exceed the maximum transmission unit (MTU). When the packet length exceeds the MTU, the packet needs to be fragmented. In IPv6, the Fragment header is used by an IPv6 source node to send a packet larger than the MTU.
Authentication header	51	IPsec uses this header to provide data origin authentication, data integrity check, and packet anti-replay functions. It also protects some fields in the IPv6 basic header.
Encapsulating Security Payload header	50	This header provides the same functions as the Authentication header plus IPv6 packet encryption.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Note:
  - Intermediate routers determine whether to process extension headers based on the Next Header field value in the IPv6 basic header. The intermediate routers do not need to examine or process all extension headers.
  - Each extension header can only occur once in an IPv6 packet, except for the Destination Options header which may occur twice (once before a Routing header and once before the upper-layer header).

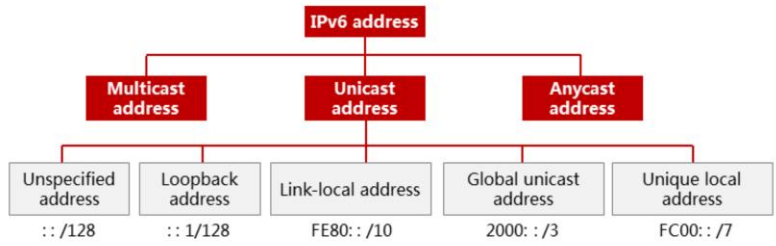
Contents

1. IPv6 Background
2. **IPv6 Implementation**
   - Packet Format
   - **Address Classification**
   - Basic Protocol
   - Transition Technologies
3. IPv6 Configuration Commands
4. Suggestions for IPv6 Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Types of IPv6 Addresses

- Unicast address: An identifier for a single interface. A packet sent to a unicast address is delivered to the interface identified by that address. In IPv6, it is common for an interface to have multiple IPv6 addresses.
- Multicast address: An identifier for a set of interfaces. A packet sent to a multicast address is delivered to all interfaces identified by that address. Only interfaces that are added to the corresponding multicast group listen to the packets destined for the multicast address.
- Anycast address: An identifier for a group of network interfaces, which usually belong to different nodes. A packet sent to an anycast address is delivered to the nearest interface that is identified by the anycast address, depending on routing protocols.
- No broadcast address is defined in IPv6.



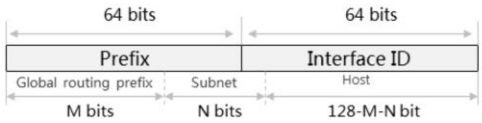
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IPv4 addresses are classified into the following types: Unicast address, multicast address, and broadcast address. IPv6 addresses are classified into unicast addresses, multicast addresses, and anycast addresses.
  - A unicast address identifies a single interface. A packet sent to a unicast address is delivered to the interface identified by that address.
  - A multicast address identifies multiple interfaces. A packet sent to a multicast address is delivered to all interfaces identified by that address.
  - An anycast address identifies multiple interfaces. A packet destined for an anycast address is sent to the interface that is nearest to the sender in the interface group identified by the anycast address. In fact, anycast addresses and unicast addresses use the same address space. That is to say, it is the router that determines whether to send a packet in unicast mode or anycast mode.

IPv6 Unicast Address — Aggregatable Global Unicast Address

- The global unicast address definition is used for IPv6 Internet. This address is globally unique and routable.
  - Similar to the IPv4 public address.
  - Composed of the prefix, subnet ID, and interface ID.



  - Global routing prefix: Assigned to an organization by a provider. Generally, the value is at least 48 bits. Currently, the first 3 bits of every assigned global routing prefix is 001. Therefore, the prefix is 2000: : /3.
  - Subnet ID: Used by an organization to construct a local network (site), which is similar to the function of the subnet ID in IPv4. There are a maximum of 64 bits for the subnet ID.
  - Host ID: Identifies a device (host), which is similar to the host ID in IPv4.

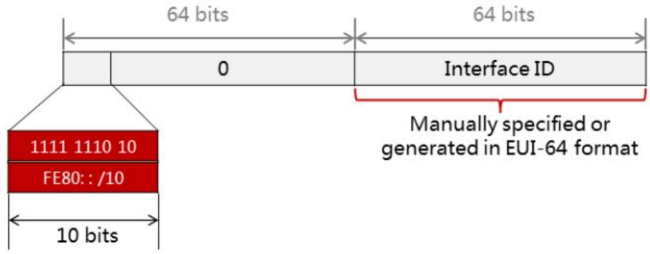
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An IPv6 global unicast address is an IPv6 address with a global unicast prefix, which is similar to an IPv4 public address. IPv6 global unicast addresses support route prefix summarization, helping limit the number of global routing entries.

IPv6 Unicast Address — Link-Local Address

- When IPv6 runs on a node, a link-local address is automatically assigned to each interface of the node. This address is used to communicate with other hosts on the same link.
  - An IPv6 link-local address can be used for communication between nodes on the same link and is widely used in neighbor discovery and stateless address configuration.
  - A link-local address uses a prefix of FE80: : /10 and an interface ID as the last 64 bits.
  - Each IPv6 interface must have a link-local address.



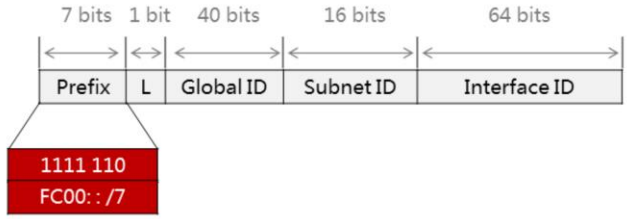
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Link-local addresses are used only in communication between nodes on the same local link. A link-local address uses a link-local prefix of FE80::/10 as the first 10 bits (1111111010 in binary) and an interface ID as the last 64 bits.
- When IPv6 runs on a node, a link-local address that consists of a fixed prefix and an interface ID in EUI-64 format is automatically assigned to each interface of the node. This mechanism enables two IPv6 nodes on the same link to communicate without any configuration, making link-local addresses widely used in neighbor discovery and stateless address configuration.
- Routers do not forward IPv6 packets with the link-local address as a source or destination address to devices on different links.

IPv6 Unicast Address — Unique Local Address

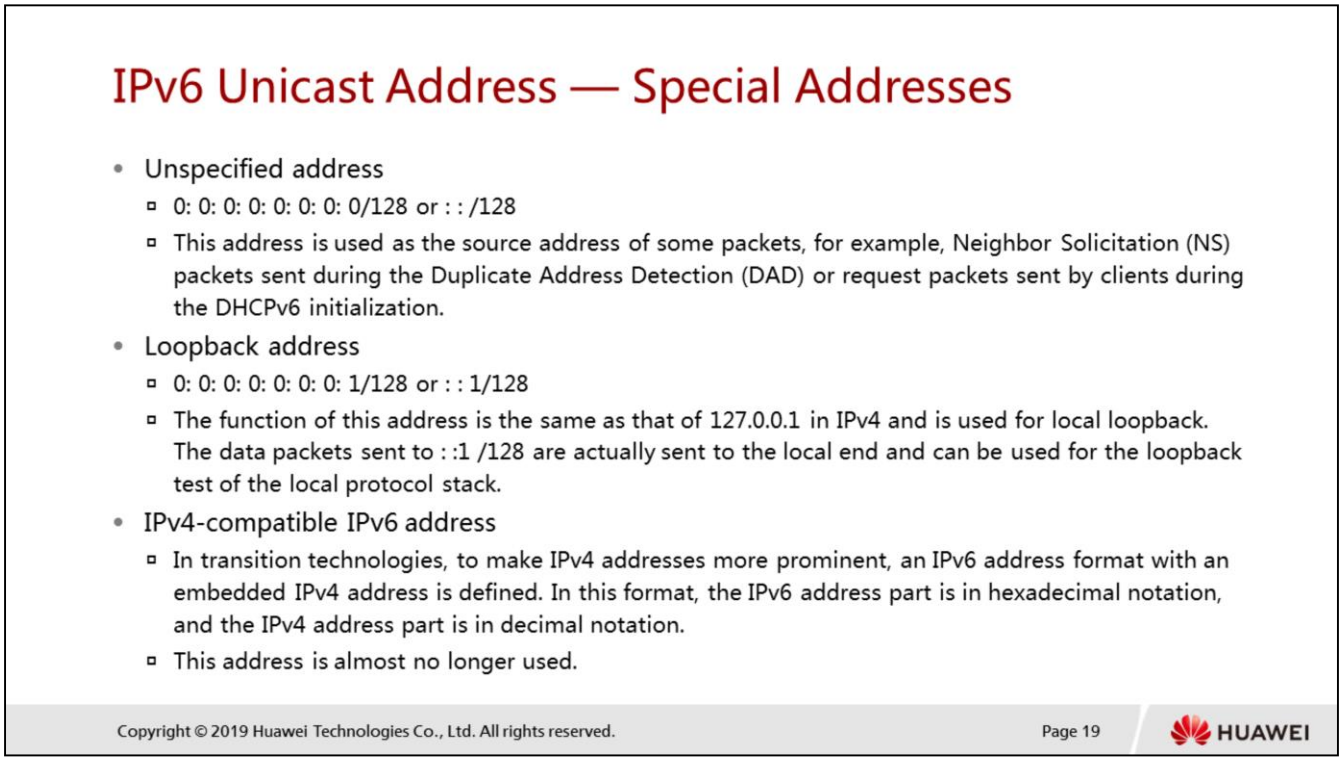
- To replace the site-local address and ensure local address uniqueness, RFC4193 defines the unique local address to prevent problems caused by the disclosure of IPv4 private addresses to the public network.
  - A unique local address, which is similar to a private network address in IPv4, can be used only on the local network and cannot be routed on the IPv6 Internet.
  - The fixed prefix of a unique local address is FC00: : /7. The prefix is divided into two parts. The FC00: : /8 part is not defined, and the other part is FD00: : /8, with the format as follows:



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Unique local addresses are used only within a site. Site-local addresses, according to RFC 3879, have been replaced by unique local addresses (RFC4193).
- Unique local addresses are similar to IPv4 private addresses. Any organization that does not obtain a global unicast address from a service provider can use a unique local address. However, unique local addresses are routable only within a local network, not the Internet as a whole.
- Description:
  - Prefix: is fixed as FC00::/7.
  - L: is set to 1 if the address is valid within a local network. The value 0 is reserved for future expansion.
  - Global ID: indicates a globally unique prefix, which is pseudo-randomly allocated (for details, see RFC 4193).
  - Subnet ID: identifies a subnet within the site.
  - Interface ID: identifies an interface.
- A unique local address has the following features:
  - Has a globally unique prefix that is pseudo-randomly allocated with a high probability of uniqueness.
  - Allows private connections between sites without creating address conflicts.
  - Has a well-known prefix (FC00::/7) that allows for easy route filtering at site boundaries.
  - Does not conflict with any other addresses if it is accidentally routed offsite.
  - Functions as a global unicast address to applications.
  - Is independent of Internet Service Providers (ISPs).



- Unspecified address
  - The IPv6 unspecified address is 0:0:0:0:0:0:0:0/128 or ::/128, indicating that an interface or a node does not have an IP address. It can be used as the source IP address of some packets, such as Neighbor Solicitation (NS) messages, in duplicate address detection (DAD). Routers do not forward packets with an unspecified address as the source IP address.
- Loopback address
  - The IPv6 loopback address is 0:0:0:0:0:0:0:1/128 or ::1/128. Similar to the IPv4 loopback address 127.0.0.1, the IPv6 loopback address is used when a node needs to send IPv6 packets to itself. This IPv6 loopback address is usually used as the IP address of a virtual interface, such as a loopback interface. The loopback address cannot be used as the source or destination IP address of packets needing to be forwarded.

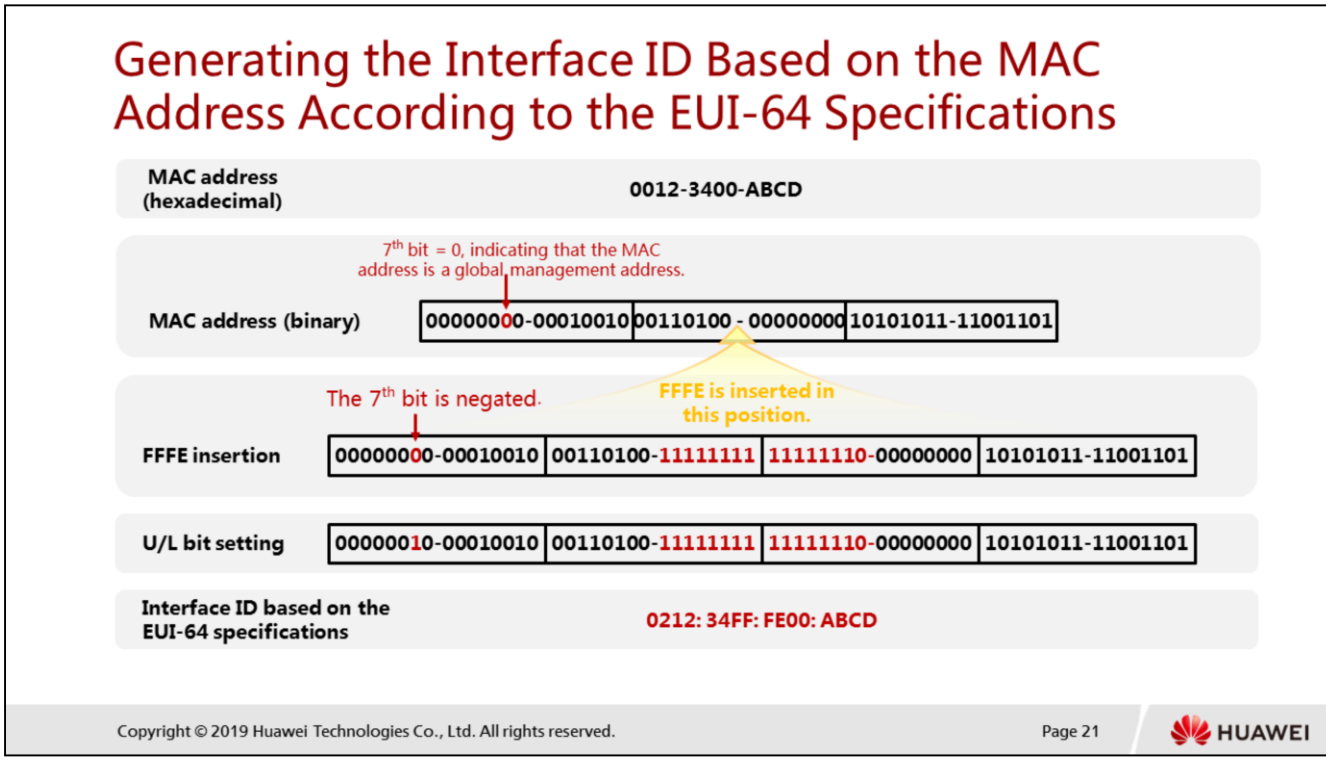
Methods of Generating an Interface ID

- Interface ID: It is 64-bit and is used to identify an interface on a link. The interface ID must be unique on each link.
- You can manually configure the interface ID, generate it through system software, or generate it in IEEE 64-bit Extended Unique Identifier (EUI-64) format.
  - Manual configuration: You are advised to configure the interface ID on servers and key network devices.
  - Automatic generation through the system software: This method intends to protect the privacy of a host.
  - Automatic generation in IEEE EUI-64 format: This method is most commonly used.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- If the first 3 bits of an IPv6 unicast address are not 000, the interface ID must contain 64 bits. If the first 3 bits are 000, there is no such limitation.
- An interface ID is 64 bits long and identifies an interface on a link. The interface ID must be unique on each link. An interface ID is used for many purposes, with the most common one being the attachment to the link-local address prefix, forming the link-local address of the interface. Or in stateless autoconfiguration, an interface ID can be attached to the IPv6 global unicast address prefix to form the global unicast address of the interface.
- IEEE EUI-64 criteria
  - Converting MAC addresses into IPv6 interface IDs reduces the configuration workload. When using stateless address autoconfiguration (described in detail in later sections), you only need an IPv6 network prefix to obtain an IPv6 address.
  - One defect of this method, however, is that an IPv6 address is easily calculable based on a MAC address, and could therefore be used for malicious attacks.



- Assume that the MAC address of an interface is shown in the preceding figure. According to the EUI-64 specifications, the interface ID can be calculated based on the MAC address. Like the MAC address, the interface ID is globally unique. The calculation process is as follows:
- EUI-64 inserts FFFE between the vendor identifier and extension identifier of the MAC address (separated in half), and then the higher seventh bit (U/L bit) 0 is changed to 1 to indicate that the interface ID is globally unique.
- In a unicast MAC address, the seventh bit of the first byte is U/L (Universal/Local, also called G/L, where G indicates Global) bit, which is used to indicate the uniqueness of the MAC address. If the U/L bit is 0, the MAC address is the global management address, which is allocated by a vendor with the OUI. If the U/L bit is 1, the MAC address is the local management address, which is customized by the network administrator based on the service purpose.
- In an EUI-64 interface ID, the indication of the seventh bit is opposite to that of an MAC address. 0 indicates local management and 1 indicates global management. Therefore, in an EUI-64 interface ID, if the U/L bit is 1, the address is globally unique; if the value is 0, the address is locally unique. This is why the bit needs to be reversed.

IPv6 Multicast Address

- Identifies a group of interfaces. The data sent to a multicast address is forwarded to multiple devices that listen to the address.
- Address range: FF00: : /8



- **Flags**
  - Indicates a permanent or temporary multicast group.
  - 0000 indicates a permanently-assigned or well-known address.
  - 0001 indicates a temporary address.
- **Scope**
  - Indicates the range of a multicast group.
- **Group ID**
  - Indicates the multicast group ID.

0: Reserved
1: Node-local scope
2: Link-local scope, for example, FF02: : 1
5: Site-local scope
8: Organization-local scope
E: Global scope
F: Reserved

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Like an IPv4 multicast address, an IPv6 multicast address identifies a group of interfaces, which usually belong to different nodes. A node may belong to any number of multicast groups. Packets sent to an IPv6 multicast address are delivered to all the interfaces identified by the multicast address.
- Flag:
  - 0000: permanent multicast address.
  - 0001: temporary multicast address.
  - Note: The first three bits are reserved as 0s.
- Application scope
  - 0001: indicates the local-interface scope. It is valid on a single interface, and is applicable to only loopback interfaces.
  - 0010: indicates the link-local scope.
  - 0100: indicates the local-administration scope, configured by the administrator.
  - 0101: indicates the site-local scope.
  - 1000: indicates the local-organization scope, that is, the range of sites of the same organization.
  - 1110: indicates the global scope.
- Group ID:

- multicast group ID.

IPv6 Address Category — Predefined Multicast Address

- Node-local
  - FF01: 0: 0: 0: 0: 0: 0: 1: multicast address of all nodes
  - FF01: 0: 0: 0: 0: 0: 0: 2: multicast address of all routers
- Link-local
  - FF02: 0: 0: 0: 0: 0: 0: 1: multicast address of all nodes
  - FF02: 0: 0: 0: 0: 0: 0: 2: multicast address of all routers
  - FF02: 0: 0: 0: 0: 1: FFXX: XXXX: multicast address of Solicited-Nodes
  - FF02: 0: 0: 0: 0: 0: 0: 5: multicast address of all OSPF routers
  - FF02: 0: 0: 0: 0: 0: 0: 6: multicast address of all OSPF DR routers
  - FF02: 0: 0: 0: 0: 0: 0: D: multicast address of all PIM routers

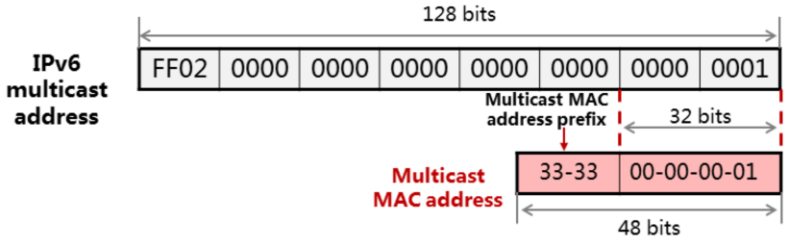
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Similar to IPv4, IPv6 has certain special multicast addresses. For example:
  - **FF01::1 (node-local scope all-nodes address)**
  - FF02: : 1 (link-local scope all-nodes address)
  - FF01: : 2 (node-local scope all-routers address)
  - FF02: : 2 (link-local scope all-routers address)
  - FF05: : 2 (site-local scope all-routers address)

Mapping Between a Multicast IPv6 Address and an IPv6 Multicast MAC Address

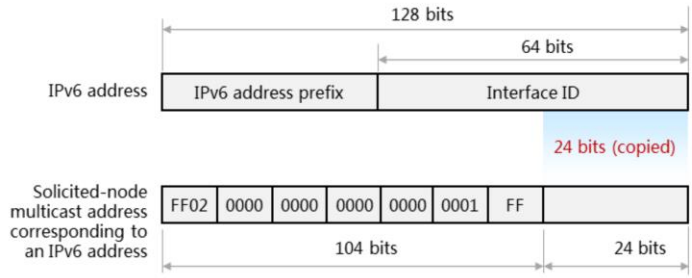
- On an Ethernet, a multicast IPv6 packet must be encapsulated in Ethernet mode.
- The destination IP address of a multicast IPv6 packet is a multicast IPv6 address, and the destination MAC address must be a multicast MAC address, which must correspond to the multicast IPv6 address.
- 33-33 is a MAC address prefix reserved for multicast IPv6 addresses. The last 32 bits of the MAC address are copied from the last 32 bits of the corresponding multicast IPv6 address.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Solicited-Node Multicast Address

- The solicited-node multicast address is generated based on the unicast or anycast address of a node. When a node has an IPv6 unicast or anycast address, a solicited-node multicast address is generated for the node, and the node joins the multicast group that corresponds to its IPv6 unicast or anycast address.
- A unicast address or anycast address corresponds to a solicited-node multicast address. This address is used for neighbor discovery (ND) and duplicate address detection (DAD).
- The solicited-node multicast address consists of the fixed prefix FF02::1:FF00: 0/104 and the last 24 bits of the corresponding IPv6 address. The valid scope of the solicited-node multicast address is the link-local scope.



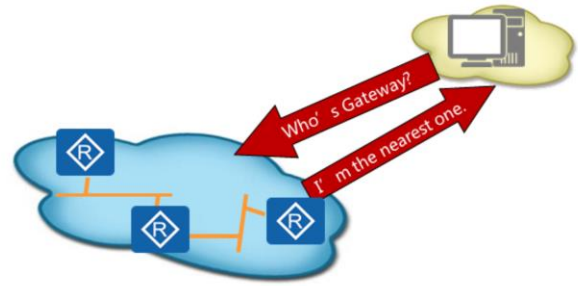
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

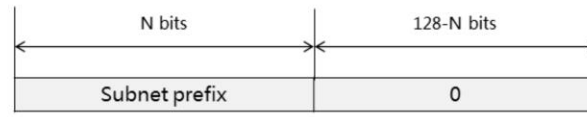
HUAWEI

- When a node has an IPv6 unicast or anycast address, a solicited-node multicast address is generated for the node, and the node joins the multicast group that corresponds to its IPv6 unicast or anycast address. Each unicast or anycast address corresponds to a solicited-node multicast address, which is often used in address resolution, neighbor discovery, and duplicate address detection.
- The solicited-node multicast address consists of the prefix FF02: : 1: FF00: 0/104 and the last 24 bits of the corresponding IPv6 address. The valid scope of the solicited-node multicast address is the link-local scope.
- What is the function of the solicited-node multicast address? Here, we use the example of ARP in IPv4 to explain. ARP is mainly used for address resolution. When a device needs to resolve an IP address to a MAC address, it sends a broadcast ARP Request frame so that all nodes in the broadcast domain can receive the broadcast frame. However, it is unnecessary for nodes except the destination node to parse the frame (till the ARP payload) because this action will waste device resources.
- In an IPv6 network, when a device needs the MAC address mapping an IPv6 address, the device sends a Request packet. This packet is a multicast packet whose destination IPv6 address is the solicited-node multicast address corresponding to the destination IPv6 unicast address. The destination MAC address of the Request packet is the multicast MAC address corresponding to the multicast address. Only the destination node listens to the solicited-node multicast address. Therefore, when other devices receive the frame, they identify it based on the destination MAC address at the network adapter layer and discard it.

IPv6 Anycast Address

- IPv6 anycast addresses are a type of peculiar addresses of IPv6. An anycast address identifies a group of interfaces that generally belong to different nodes.
- A packet sent to an anycast address is sent to the nearest interface that is identified by the address.
- Anycast addresses are allocated from the unicast address space, using any of the defined unicast address formats.
- Therefore, when configuring an anycast address, you need to specify that it is an anycast address.





Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IPv6 anycast addresses are a type of peculiar addresses of IPv6. An anycast address identifies a group of interfaces that generally belong to different nodes. Packets sent to an anycast address are delivered to the nearest interface that is identified by the anycast address, depending on the routing protocols. An anycast address is used for one-to-one-of-many communication. The receiver needs to be one of a group of interfaces. For example, a mobile subscriber accesses the nearest receiving station based on the physical location. Thereby, mobile subscribers are not strictly limited by physical locations.
- Anycast addresses are allocated from the unicast address space, using any of the defined unicast address formats. Thus, anycast addresses are syntactically indistinguishable from unicast addresses. The node to which an anycast address is assigned must be explicitly configured to know that it is an anycast address. Currently, anycast addresses are used only as destination addresses, and are assigned to only routers.
- The subnet-router anycast address is defined in RFC 3513 and the interface ID of a subnet-router anycast address is all 0s.
- Packets destined for a subnet-router anycast address are delivered to a certain router (the nearest router that is identified by the address) in the subnet specified by the prefix of the address. The nearest router is defined as being closest in terms of routing distance.

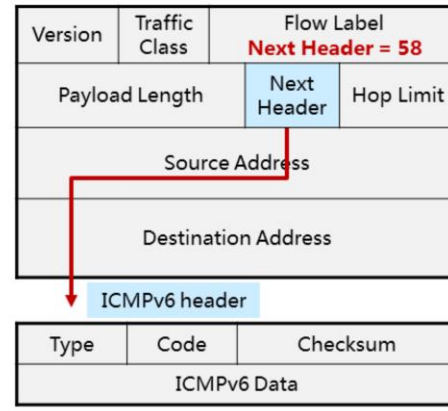
Contents

1. IPv6 Background
2. **IPv6 Implementation**
   - Packet Format
   - Basic Configuration
   - **Basic Protocol**
   - Transition Technologies
3. IPv6 Configuration Commands
4. Suggestions for IPv6 Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

ICMPv6 Overview

- ICMPv6 is one of the basic IPv6 protocols and is used by IPv6 nodes to report information about packet forwarding or errors to the source node.
- The protocol type number, that is, IPv6 Next Header, of ICMPv6 is 58.
- ICMPv6 provides additional mechanisms alongside the current ICMPv4 functions such as Neighbor Discovery (ND), stateless address configuration, duplicate address detection (DAD), and Path Maximum Transmission Unit (PMTU) discovery.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The protocol number of ICMPv6 (that is, the value of the Next Header field in an IPv6 packet) is 58.
- In IPv4, ICMP reports IP packet forwarding information and errors to the source node. ICMP defines certain messages such as Destination Unreachable, Packet Too Big, Time Exceeded, Echo Request, and Echo Reply to facilitate fault diagnosis and information management. In addition to the current ICMPv4 functions, ICMPv6 provides mechanisms such as neighbor discovery (ND), stateless address configuration (including duplicate address detection), and path MTU discovery.
- Therefore, ICMPv6 is a basic protocol for other IPv6 mechanisms.
- Packet description:
  - Type: specifies a message type. Values 0 to 127 indicate the error message type, and values 128 to 255 indicate the informational message type.
  - Code: indicates a specific message type.
  - Checksum: indicates the checksum of an ICMPv6 message.

ICMPv6

Message	Type	Detail	Code
Error message	1	Destination Unreachable	0: No route
			1: Communication with the destination node administratively prohibited
			2: Unspecified
			3: Unreachable address
			4: Unreachable port
	2	Packet Too Big	0
	3	Time Exceeded	0: Hop limit exceeded in transit
			1: Fragment reassembly time exceeded
	4	Parameter Problem	0: Error header field
			1: Unidentified Next Header type
			2: Unidentified IPv6 option
Informational message	128	Echo Request	0
	129	Echo Reply	0

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

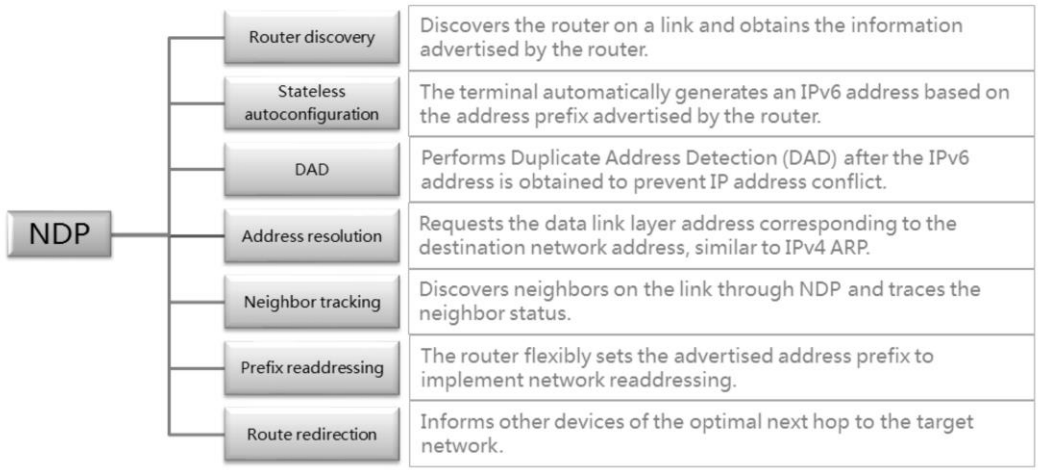
- Destination Unreachable message
  - When a data packet fails to be sent to the destination node or the upper-layer protocol, the router or the destination node sends an ICMPv6 Destination Unreachable message. In an ICMPv6 Destination Unreachable message, the value of the Type field is set to 1 and the value of the Code field ranges from 0 to 4. Each code has a specific meaning (defined in RFC 2463):
    - 0: no route to the destination node
    - 1: Communication with the destination node is administratively prohibited.
    - 2: unspecified
    - 3: unreachable address
    - 4: unreachable port
- Packet Too Big message
  - A Packet Too Big message is sent by a router in response to a packet that the router cannot forward because the packet is larger than the MTU of the outgoing link. The information in this message is used for the path MTU discovery process. In a Packet Too Big message, the value of the Type field is set to 2 and the value of the Code field is set to 0.
- Time Exceeded message
  - If a router receives a packet with a Hop Limit of zero, the router discards the packet, originates an ICMPv6 Time Exceeded message, and sends the message to the source node. In an ICMPv6 Time Exceeded message, the value of the Type field is set to 3 and the value of the Code field is set to 0 or 1.
    - 0: hop limit exceeded in transit
    - 1: fragment reassembly time exceeded

- Parameter Problem message
  - If an IPv6 node processing a packet finds a problem with a field in the IPv6 header or extension headers such that it cannot complete processing the packet, it discards the packet and originates an ICMPv6 Parameter Problem message to the packet source, indicating the type and location of the problem. In a Parameter Problem message, the value of the Type field is set to 4, the value of the Code field ranges from 0 to 2, and the 32-bit Point field indicates the location of the problem. The meaning of the Code field value is as follows:
    - 0: error header field
    - 1: unidentified Next Header type
    - 2: unidentified IPv6 option
- RFC2463 defines only two types of informational packets: Echo Request and Echo Reply messages
  - Echo Request message
    - Echo Request messages are sent to destination nodes. After receiving an Echo Request message, the destination node responds with an Echo Reply message. In an Echo Request message, the Type field value is 128 and the Code field value is 0. The Identifier and Sequence Number fields are specified on the source node. They are used to match the Echo Reply packet to be received with the sent Echo Request packet.

- Echo Reply message
  - After receiving an Echo Request message, the destination node responds with an Echo Reply message. In an Echo Reply message, the Type field value is 129 and the Code field value is 0. The values of the Identifier and Sequence Number fields in an Echo Reply message must be the same as those in an Echo Request message.

Overview of IPv6 Neighbor Discovery Protocol (NDP)

- NDP is defined in RFC2462 and RFC4861. NDP implements many important IPv6 mechanisms, as shown in the following figure.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Router discovery: This function helps a device to discover routers on a link and obtain the information advertised by routers.
- Stateless address autoconfiguration: Stateless autoconfiguration is a highlight function of IPv6. It enables IPv6 hosts to be easily connected to IPv6 networks and to be plug and play. Using this function, you do not need to manually configure complex IPv6 addresses or deploy application servers (such as DHCP servers) to assign addresses to hosts. The stateless autoconfiguration mechanism involves the Router Solicitation (RS) and Router Advertisement (RA) messages in ICMPv6.
- Duplicate Address Detection (DAD): DAD is an important mechanism. An IPv6 address can be used only after passing through the DAD process. DAD detects whether an IPv6 address conflict exists on a link.
- Address resolution: IPv6 does not use the ARP protocol in IPv4, but uses the Neighbor Solicitation (NS) and Neighbor Advertisement (NA) messages defined in NDP to implement address resolution.
- Neighbor tracking: IPv6 defines the neighbor state machine between nodes and maintains the mapping between IPv6 addresses and Layer 2 addresses (such as MAC addresses) of neighbors. The corresponding entries are stored in the IPv6 neighbor table of a device.
- Prefix readdressing: An IPv6 router advertises the IPv6 prefix information to the local link through ICMPv6 RA messages. In this manner, hosts can automatically generate their own IPv6 unicast addresses based on the prefix information in RA messages. These automatically obtained addresses have a specified period of time to live (TTL). Advertising the IPv6 address prefix in RA messages and flexibly setting the TTL of addresses enable an old prefix to smoothly migrate to a new one. This removes the need to reconfigure addresses on hosts, greatly reducing the manual labor.
- Router redirection: A router sends an ICMPv6 Redirection message to an IPv6 node to notify it that there is a better next hop to the destination on the local link. The

redirection function in IPv6 is the same as that in IPv4.

ICMPv6 Packets Used by NDP

ICMPv6 Packet / Mechanism	RS 133	RA 134	NS 135	NA 136	Redirection 137
Address resolution			•	•	
Prefix advertisement	•	•			
Prefix readdressing	•	•			
DAD			•	•	
Route redirection					•

- RS: Router solicitation packet
- RA: Router advertisement packet
- NS: Neighbor solicitation packet
- NA: Neighbor advertisement packet

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Address Resolution

- IPv6 does not support Address Resolution Protocol (ARP) or broadcast addresses.
- Layer 3 address resolution enables different link layer protocols to use the same address resolution protocol.
- Address resolution is implemented through ICMPv6 (NS packets of type 135 and NA packets of type 136).
- An NS packet is sent in multicast mode. The destination IPv6 address of the packet is the solicited-node multicast address corresponding to the solicited IPv6 address, and the destination MAC address of the packet is the multicast MAC address.
- Sending NS messages in multicast mode is more efficient than broadcasting, and also reduces the impact on other nodes and the pressure on the Layer 2 network performance.
- Layer 3 security mechanisms (such as IPsec) can be used to prevent address resolution attacks.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An IPv4 address can be resolved into a link-layer address through ARP. The ARP protocol works at Layer 2. IPv6 address resolution is defined in the Neighbor Discovery protocol (RFC2461), using ICMPv6 packets to resolve addresses at Layer 3. This mechanism brings the following benefits:
  - Enhanced media independence: This means that we do not need to define a new address resolution protocol for each link layer but use the same address resolution protocol at all link layers.
  - Layer 3 security mechanism: ARP spoofing (for example, forging ARP Reply packets to steal data flows) is a big security threat in IPv4. The Layer 3 standard security authentication mechanism (for example, IPsec) can be used to resolve this problem during address resolution.
  - If an ARP Request packet is sent in broadcast mode, it will be flooded to all hosts on the Layer 2 network, causing IPv4 performance deterioration. At Layer 3, an address resolution request packet will only be sent to the solicited-node multicast group to which the address to be resolved belongs. The transmission in multicast mode greatly reduces the performance pressure.

Address Resolution Packet

- Two types of ICMPv6 packets are used during address resolution: Neighbor Solicitation (NS) packet and Neighbor Advertisement (NA) packet

- **NS**
  - Type=135, Code=0
  - Target Address is the IPv6 address to be resolved, which cannot be a multicast address.
- **NA**
  - Type=136, Code=0
  - The R flag (Router flag) indicates whether the sender is a router. If the value is 1, the sender is a router.
  - The S flag (Solicited flag) indicates whether a Neighbor Advertisement (NA) packet is sent to respond to a Neighbor Solicitation (NS) packet. If the value is 1, the NA packet responds to an NS packet.
  - The O flag (Override flag) indicates whether information in an NA packet overrides the existing information in an address. If the value is 1, the existing information is overwritten.
  - Target Address indicates the IPv6 address corresponding to the link layer address.

Type	Code	checksum
Reserved		
Target Address		
Options…		

Type			Code	Checksum
R	S	O	Reserved	
Target Address				
Options…				

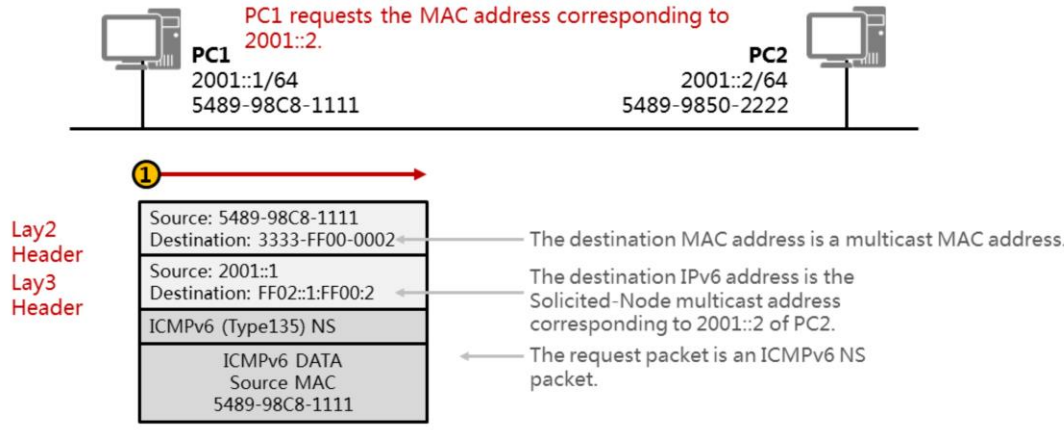
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Two types of ICMPv6 packets are involved during address resolution: Neighbor Solicitation (NS) and Neighbor Advertisement (NA) messages
- NS message
  - The ICMP Type field value is 135 and the Code field value is 0.
  - The Target Address field indicates the IPv6 address to be resolved, which cannot be a multicast address.
  - The link-layer address of the NS message sender is encapsulated in the Options field.
- NA message
  - The ICMP Type field value is 136 and the Code field value is 0.
  - The R flag (Router flag) indicates whether the sender is a router. If the value is 1, the sender is a router.
  - The S flag (Solicited flag) indicates whether an NA message is sent to respond to an NS message. If the value is 1, the NA message responds to an NS message.
  - The O flag (Override flag) indicates whether information in an NA message overrides the existing information in an address. If the value is 1, the existing information is overridden.
  - Target Address indicates the IPv6 address corresponding to the link-layer address carried in the NA message.

- The requested link-layer address is encapsulated in the Options field, in the TLV format. For details, see RFC2463.

Address Resolution (1/2)

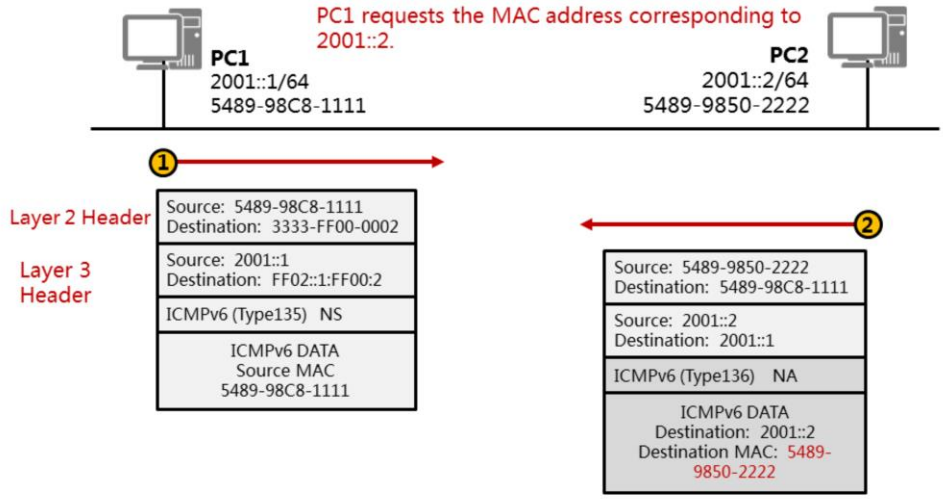


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- There are two types of messages: NS and NA. How can two hosts obtain the link-layer address of each other?
- In the scenario shown in the preceding figure, if PC1 requests the MAC address corresponding to 2001::2 of PC2, PC1 sends an NS message. The source address of the NS message is 2001::1, and the destination address is the solicited-node multicast address corresponding to 2001::2.
- Then, a frame header is encapsulated into the IPv6 packet. The source MAC address is the MAC address of PC1 and the destination MAC address is the mapping MAC address of the solicited-node multicast address corresponding to 2001::2. The destination MAC address is a multicast MAC address.
- In this way, a bidirectional exchange of link-layer addresses is completed.

Address Resolution (2/2)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Other nodes except R2 also receive the data frame. When a node extracts the frame header, it finds that the destination MAC address is a multicast MAC address, which is not listened on the local device. Therefore, the data frame is discarded by the network adapter.
- The local network adapter receives the data frame whose destination MAC address is 3333-FF00-0002. After PC2 receives and checks the data frame, the local network adapter detects that the packet is an IPv6 packet based on the Type field in the frame header. It then removes the frame header, and sends the IPv6 packet to the IPv6 protocol stack for processing. The IPv6 protocol stack detects that the packet that the packet is destined for the solicited-node multicast address FF02::1:FF00:2 based on the destination IPv6 address in the IPv6 header. The local network adapter has joined the multicast group. The Next Header field in the IPv6 packet header indicates that an ICMPv6 packet is encapsulated following the IPv6 packet header. Therefore, PC2 removes the IPv6 packet header and sends the ICMPv6 packet to the ICMPv6 protocol for processing. Finally, ICMPv6 finds that the packet is an NS message requesting the MAC address corresponding to 2001::2. In response, PC2 sends an NA message, containing the MAC address of PC2, to PC1.

Checking the IPv6 Neighbor Routing Table

- Unlike IPv4, IPv6 does not use the ARP table to cache the mapping between IP addresses and MAC addresses. Instead, it maintains an IPv6 neighbor routing table. On a Huawei datacom device, you can run the **display ipv6 neighbors** command to check the IPv6 neighbor routing table.

```
[R1] display ipv6 neighbors
------------------------------------------------------------------------------
IPv6 Address    : 2012::2
Link-layer      : 00e0-fcc2-13b6            State : STALE
Interface       : GE0/0/0                   Age   : 0
VLAN            : -                         CEVLAN: -
VPN name        :                           Is Router: TRUE
Secure FLAG     : UN-SECURE

IPv6 Address        : FE80::2E0:FCFF:FEC2:13B6
Link-layer      : 00e0-fcc2-13b6            State : STALE
Interface       : GE0/0/0                   Age   : 0
VLAN            : -                         CEVLAN: -
VPN name        :                           Is Router: TRUE
Secure FLAG     : UN-SECURE
------------------------------------------------------------------------------
Total: 2     Dynamic: 2     Static: 0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- On a device running the Windows 7 operating system, you can run the **netsh interface ipv6 show neighbors** command to check the cached neighbor information.

Neighbor State Type

- In actual communication, a neighbor table needs to be maintained to monitor the state of each neighboring device, and transition between states is allowed.
- There are five types of neighbor states:
  - INCOMPLETE: This state indicates that an NS packet has been sent to the solicited-node multicast address of the destination node, but the NA packet has not been received.
  - REACHABLE: This state indicates that the neighbor is reachable, and no packet needs to be sent for confirmation.
  - STALE: This state indicates that more than 30 seconds have passed after the last reachability acknowledgement was received.
  - DELAY: This state indicates that a packet is sent when the neighbor is in stale state, and no reachability acknowledgement is received within five seconds.
  - PROBE: This state indicates that an NS packet is retransmitted every one second to detect reachability until an acknowledgment is received.

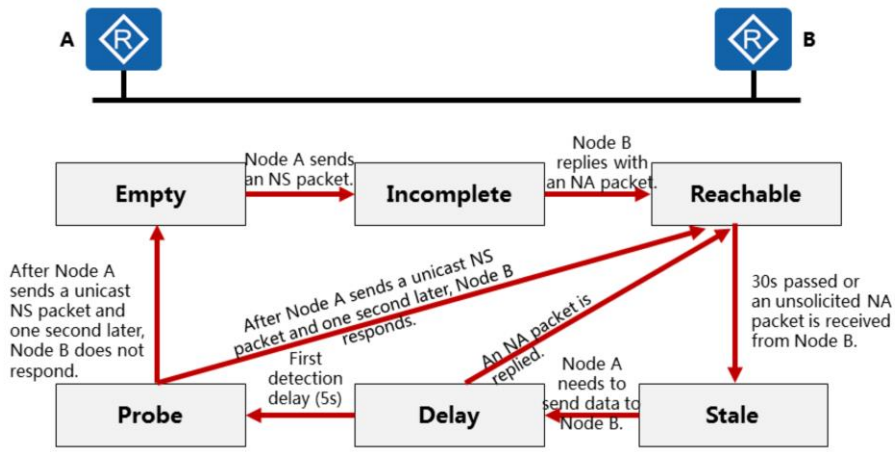
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The previous sections describe the process of address resolution. However, in the actual communication process, a neighbor table needs to be maintained. In the table, each neighbor is in its own state and can migrate between states.
- RFC2461 defines five neighbor states: Incomplete, Reachable, Stale, Delay, and Probe.

Neighbor State Change

- Example: Node A wants to access Node B, but the cache of node A does not contain the entry of node B. The following figure shows the change of the neighbor state machine.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The neighbor state transition is complex and is not described in detail here. The following example describes changes in neighbor state of Node A during its first communication with Node B.
  - Node A sends an NS message and generates a cache entry. The neighbor state of Node A is Incomplete.
  - If Node B replies with an NA message, the neighbor state of Node A changes from Incomplete to Reachable. Otherwise, the neighbor state changes from Incomplete to Empty after 10 seconds, and Node A deletes this entry.
  - After the neighbor Reachable time times out (30s by default), the neighbor state changes from Reachable to Stale.
  - If Node A in the Reachable state receives an unsolicited NA message from Node B, and the link-layer address of Node B carried in the message is different from that learned by Node A, the neighbor state of Node A changes to Stale.
  - Node A sends data to Node B. The state of Node A changes from Stale to Delay. Node A then sends an NS Request message.
  - After a period of Delay_First_Probe_Time (5s by default), the neighbor state changes from Delay to Probe. During this period, if Node A receives an NA Reply message, the neighbor state of Node A changes to Reachable.
  - Node A in the Probe state sends several (MAX_UNICAST_SOLICIT) unicast NS messages at the configured RetransTimer interval (1s by default). If Node A receives a Reply message, its neighbor state changes from Probe to Reachable. Otherwise, the state changes to Empty and Node A deletes the entry.
- The preceding mechanism shows that the IPv6 neighbor relationship is better than IPv4 ARP. The IPv6 neighbor state maintenance mechanism ensures that the neighbor is reachable before the communication is initiated, but the ARP maintains the neighbor

state only through the aging mechanism.

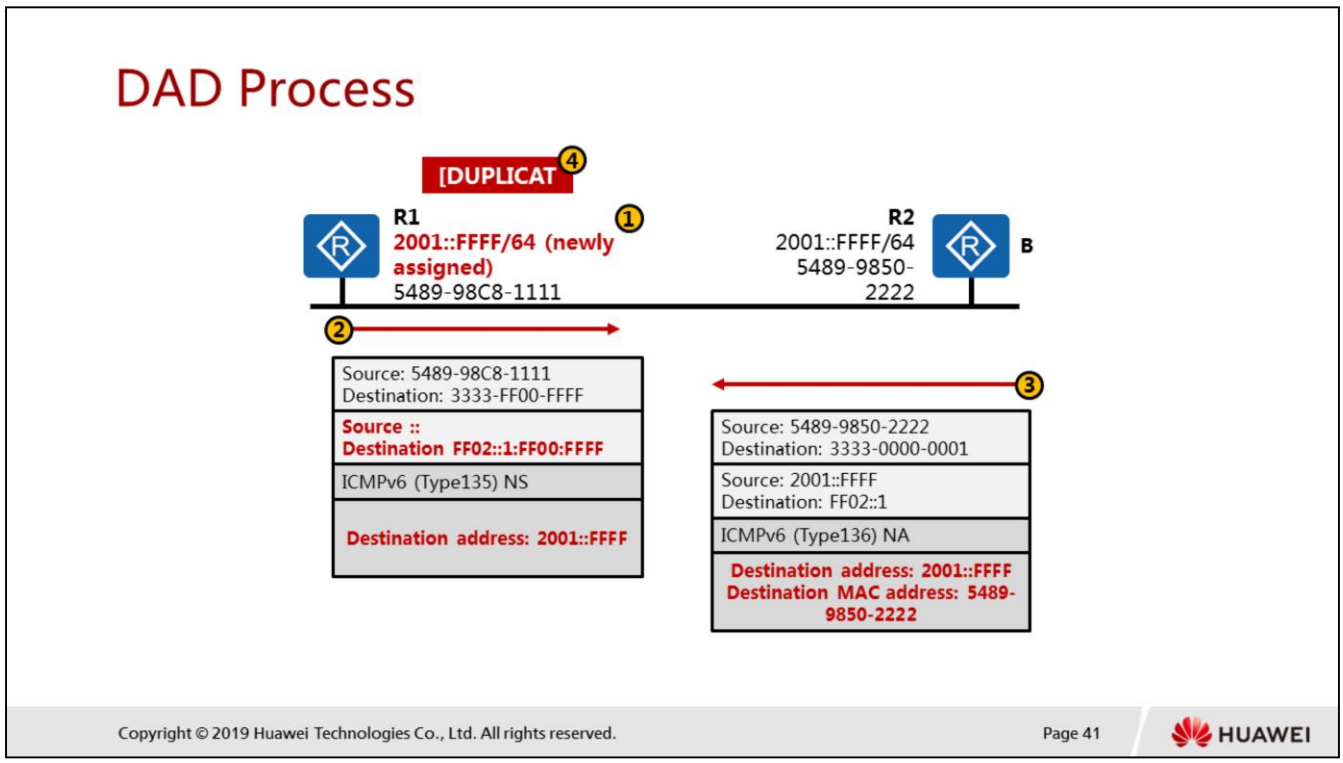
- For details about neighbor state maintenance and transition, see RFC2461.

Duplicate Address Detection (DAD)

- Overview
  - DAD ensures that there are no two identical unicast addresses on a network.
  - DAD is required for all addresses.
  - NS and NA packets are involved in the DAD process.
- Implementation
  - An IPv6 address assigned to an interface but not verified by DAD is called a tentative address. The interface cannot use the tentative address for IPv6 unicast communication but will join the Solicited-Node multicast group.
  - DAD: A node sends an NS packet to the Solicited-Node multicast address group where the tentative address belongs to. If the node receives an NA Reply packet from another site, the tentative address has already been used and this node will not use this tentative address for communication.
  - Before assigning any IPv6 unicast address to an interface, including the link-local address, you need to perform DAD on the interface.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



- R2 is an online device and has used the address as shown in the figure. Now, a new IPv6 address 2001: : FFFF/64 is configured for R1. After 2001: : FFFF/64 is configured on the interface of R1, this address enters the tentative state and is unavailable until the address passes DAD.
  - R1 sends an NS message to the local link in multicast mode, with the source IPv6 address being ": :" and the destination IPv6 address being the solicited-node multicast address corresponding to 2001: : FFFF, that is, FF02: : 1: FF00: FFFF for DAD. The NS message contains the destination address 2001: : FFFF for DAD.
  - All nodes on the link receive this NS message. Because the interfaces not configured with 2001: : FFFF are not added to the corresponding solicited-node multicast group, they discard the NS message. Because the interface of R2 is configured with 2001: : FFFF, it is added to the multicast group FF02::1:FF00:FFFF. After R2 receives the NS message destined for FF02::1:FF00:FFFF, it parses the message and finds that the destination address for DAD is the same as the local interface address. Then R2 replies with an NA message with the destination address being FF02: : 1, that is, the multicast address of all nodes. In addition, the destination address 2001:: FFFF and the MAC address of the interface on R2 are contained in the message.
  - After R1 receives the NA message, it knows that 2001:: FFFF has already been in use on the link. Therefore, R1 marks the address as Duplicate. Therefore, this address cannot be used for communication.

DAD Observed on R1

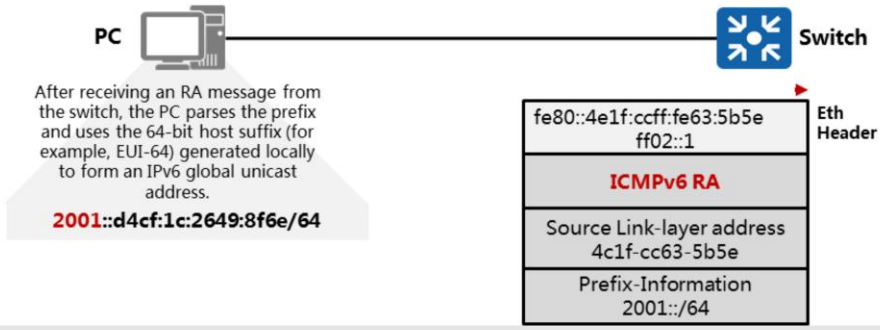



```
[R1] display ipv6 interface
GigabitEthernet0/0/0  current state: UP
IPv6 protocol current state: UP
IPv6 is enabled, link-local address is FE80::5489:98FF:FEC8:1111
  Global unicast address(es):
    2001: : FFFF, subnet is 2001::/64 [DUPLICATE]
  Joined group address(es):
    FF02::1:FF00:FFFF
    FF02::1:FFC8:1111
    FF02::2
    FF02::1
  MTU is 1500 bytes
  ND DAD is enabled, number of DAD attempts:  1
  ...
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Overview of IPv6 SLAAC

- Stateless Address Autoconfiguration (SLAAC) is a standard IPv6 function defined in RFC2462.
- In IPv6, a device can obtain its IP address dynamically or through manual configuration. In the dynamic address obtaining mode, DHCPv6 and stateless address autoconfiguration are available.
- Compared with DHCPv6, SLAAC does not require an application server and is more lightweight.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- After IPv6 stateless address autoconfiguration is enabled, the IPv6 address of a device does not need to be manually configured and the device is plug and play, reducing the burden on network management.
- The process is as follows:
  - The host automatically generates the link-local address of the network adapter based on the local interface ID.
  - The host performs DAD on the link-local address. If no address conflict exists, the link-local address can be used.
  - The host sends an RS message to discover any IPv6 router on the link. The source address of the message is the link-local address of the host.
  - The router replies with an RA message carrying the IPv6 prefix. The router can be configured to send an RA message even if it does not receive an RS message.
  - The host obtains the IPv6 address prefix based on the RA message replied by the router and generates a unicast IPv6 address by using the prefix and the locally generated interface ID.
  - The host performs DAD on the generated IPv6 address. If no conflict is detected, the address can be used.

Router Discovery Overview

- Router Discovery is the basis of IPv6 address autoconfiguration and is implemented through the following two types of packets:
  - Router Advertisement (RA) packet: Each router periodically sends multicast RA messages carrying network prefixes and identifiers on the network to declare its existence to Layer 2 hosts and devices. An RA message has an ICMP Type field value of 134.
  - Router Solicitation (RS) packet: After being connected to the network, a host immediately sends an RS message to obtain network prefixes. Devices on the network reply with RA messages. An RS message has an ICMP Type field value of 133.

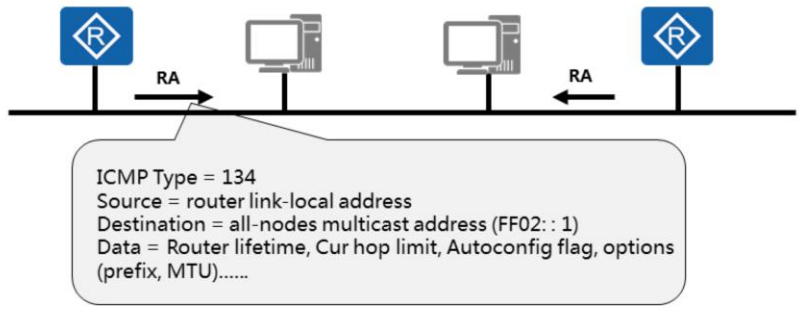
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Router discovery locates neighboring devices and learn their address prefixes and configuration parameters for address autoconfiguration.
- As mentioned above, we have learned that IPv6 addresses can be obtained through stateless autoconfiguration. That is, hosts obtain network prefixes through RA messages sent by routers, generate interface IDs, and automatically configure IPv6 addresses.

Router Discovery — Router Periodically Sending RA Messages

- Routers on the link periodically send RA messages.



- A host that receives the RA message is added to the default router list.
- A router that receives the RA message checks the RA consistency.

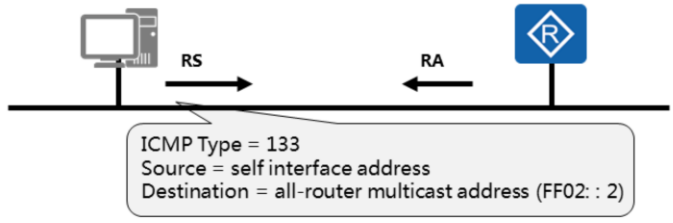
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- How does a host obtain information including the network prefix? Two methods are available: A host can directly obtain information in the Router Advertisement (RA) message received from a router or send an Router Solicitation (RS) message to a router and wait for the router to reply with an RA message, from which required information can be obtained.

Router Discovery — Router Replying with RA Messages

- When a host sends an RS message during interface initialization, the router responds with an RA message.



Note: The replied RA message can be sent to the requesting host in unicast mode or to all nodes in multicast mode.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Process for a Host to Obtain the Prefix and Other Parameters

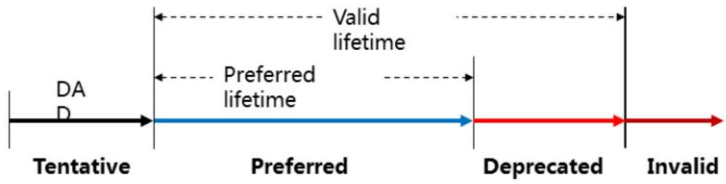
- If the following conditions are met, the prefix in an RA message can be ignored:
  - The **auto** field in the RA message is displayed as 0.
  - The prefix is the same as that of an existing address (including the link-local address).
  - In the RA message, the value of **preferred lifetime** is greater than the value of **valid lifetime**.
  - The sum of the prefix length and interface ID length is not 128 bits.
- Except the preceding conditions, a host obtains the following time parameters when obtaining the prefix:
  - **preferred lifetime** = validity period of the new communication
  - **valid lifetime** = validity period of the original communication
- The host periodically receives RA messages and updates its time parameters based on the received RA messages.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

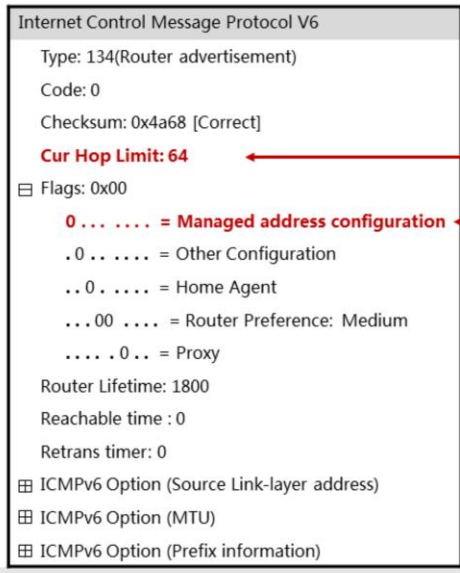
Lifetime Parameters

- When an address is in the Deprecated state, it cannot initiate a connection but only can be connected. This design ensures the upper-layer applications. However, the address becomes invalid after the valid lifetime, and all connections go Down.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Flags Field in an ICMPv6 RA Message (1/2)



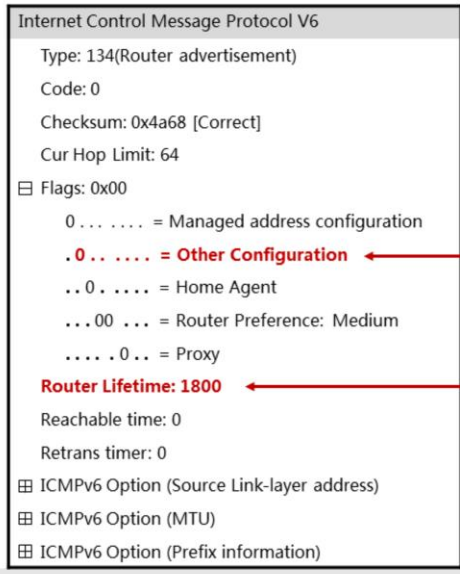
This field limits the number of hops on a host. After a PC uses the prefix advertised in the RA message to construct an IPv6 address, the hop limit of IPv6 packets sent by the PC is set to 64.

The default M bit value is 0. When 0 is used, the host that receives the RA message uses the IPv6 prefix contained in the RA message for stateless address autoconfiguration.

Run the following command to set this field to 1:
ipv6 nd autoconfig managed-address-flag
If this field is set to 1, the host that receives the RA message uses stateful autoconfiguration, that is, DHCPv6 is used to obtain the IPv6 address.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Flags Field in an ICMPv6 RA Message (2/2)



The default value of Other Configuration is 0, indicating that the host should not use the stateful autoconfiguration mechanism to configure other parameters except the IPv6 address.

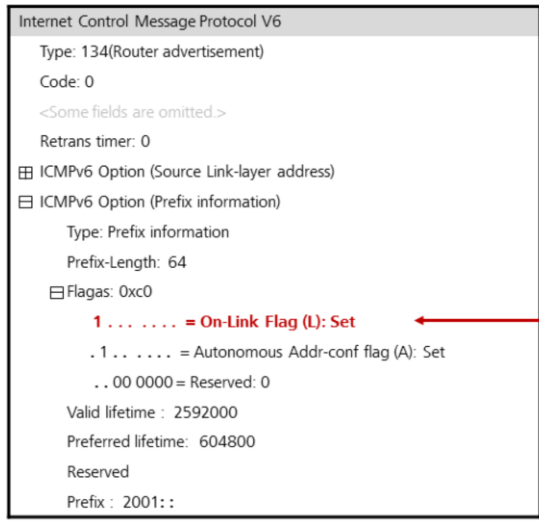
Run the following command to set this field to 1:
ipv6 nd autoconfig other-flag
The host then needs to use DHCPv6 to configure other information except the IPv6 address, such as DNS and domain name.

The unit is second. The host considers the value as the default router lifetime. When Router Lifetime is set to 0, the router will be excluded from the default gateway list of the host.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Flags Field of the Prefix in an ICMPv6 RA Message (1/2)



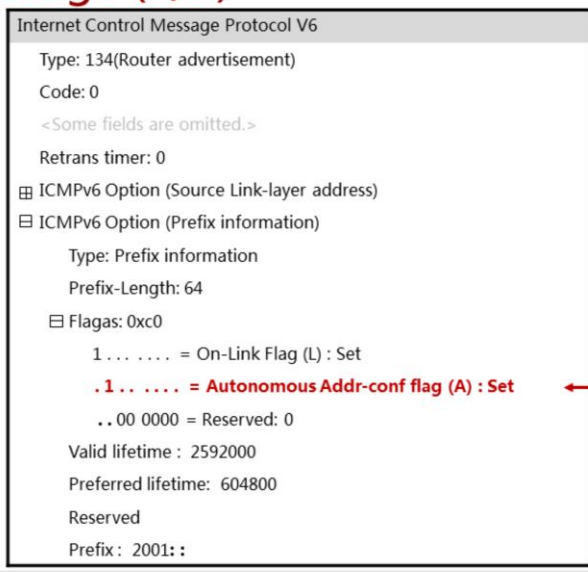
The default value of the L bit (defined in RFC 2461) is 1, indicating that the prefix in the RA message is assigned to the local link. Therefore, nodes that send data to addresses containing the specified prefix consider the local link as the destination.

You can run the following command to set this parameter to 0:

ipv6 nd ra prefix 2001: : 64 2592000 604800 off-link

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Flags Field of the Prefix in an ICMPv6 RA Message (2/2)



The default value of the A bit (defined in RFC 2461) is 1, indicating that the host on the local link can use this prefix for stateless autoconfiguration. If the value is 0, the prefix cannot be used for stateless autoconfiguration.

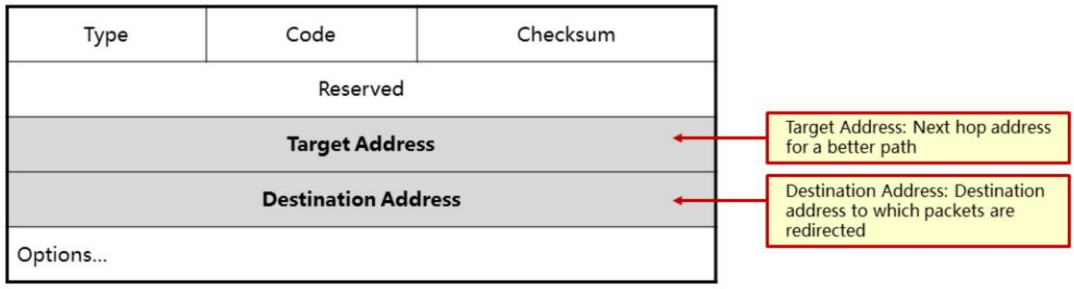
Run the following command to set this bit to 0:

ipv6 nd ra prefix 2001: : 64 2592000 604800 no-autoconfig

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Redirection Packet

- When a better forwarding path is available, the gateway router sends a Redirection message to notify the host.
- The structure of the Redirection packet is as follows:



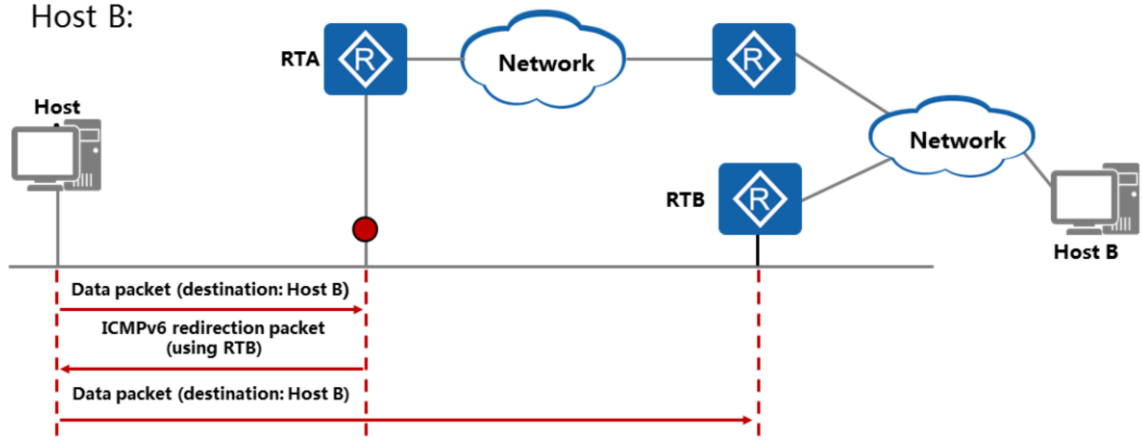
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- When a better forwarding path is available, the current gateway router sends a Redirection message to notify the sender that another gateway router can send packets.
  - In the packet format, the Type field value is 137 and the Code field value is 0.
  - Target Address indicates a better next hop address.
  - Destination Address is the destination address of the packet that needs to be redirected.

Redirection Process

- The default router of Host A is RTA. When Host A needs to send data to Host B:



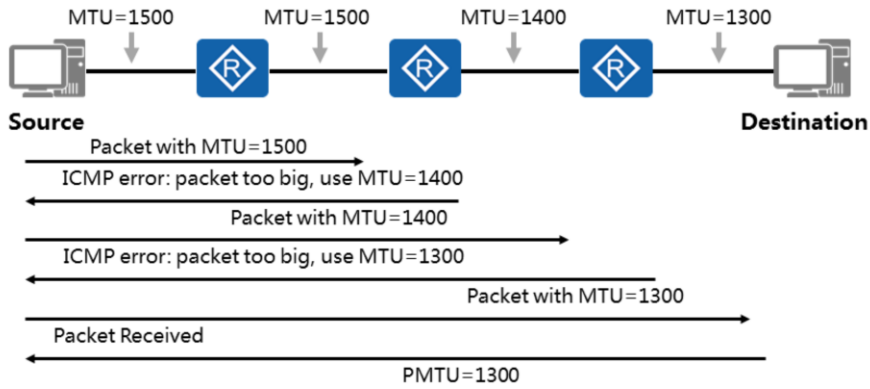
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The following is an example. Assume that Host A wants to communicate with Host B and the default gateway router of Host A is RTA. When Host A sends a packet to Host B, the packet is forwarded by RTA.
- RTA sends a Redirection message carrying the destination address of Host B to Host A to notify Host A that RTB is a better next hop address.
- After receiving the Redirection message, Host A adds a host route to the default routing table. Packets sent to Host B will be sent directly to RTB.
- This is a simple process of redirection. You may ask: How does RTA know that RTB is a better next hop address? This is simple because RTA finds that packets go in and out from the same interface. That is, the packets destined for Host B are actually forwarded to RTB after just passing through RTA. That's how RTA determines that the direct route to RTB is a better path.

PMTU Discovery (1/2)

- The path MTU (PMTU) is the minimum interface maximum transmission unit (MTU) on a path.
- The PMTU discovery protocol is defined in RFC 1981.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- After learning IPv6 packet forwarding in previous sections, we know that IPv6 packets are not fragmented or reassembled during forwarding. IPv6 packets are fragmented only on the source node and are assembled on the destination node. To ensure that all packets can be smoothly transmitted on a path, the size of fragmented packets cannot exceed the minimum MTU on the path, that is, path MTU (PMTU).
- RFC1981 defines the PMTU discovery mechanism, which is implemented through ICMPv6 Packet Too Big messages. A source node first uses the MTU of its outbound interface as the PMTU and sends a probe packet. If a smaller PMTU exists on the transmission path, the transit device sends a Packet Too Big message to the source node. The Packet Too Big message contains the MTU value of the outbound interface on the transit device. After receiving this message, the source node changes the PMTU value to the received MTU value and sends packets based on the new MTU. This process repeats until packets are sent to the destination address. The source node obtains the PMTU of the destination address.
- For example, packets are transmitted through four links with MTU values of 1500, 1500, 1400, and 1300 bytes. Before sending a packet, the source node fragments the packet based on a PMTU of 1500. When the packet is sent to the outbound interface with MTU 1400, the device returns a Packet Too Big message carrying MTU 1400. The source node then fragments the packet based on MTU 1400 and sends the fragmented packet again. The process repeats when the packet based on MTU 1400 is sent to the outbound interface with MTU 1300, the device returns another Packet Too Big message that carries MTU 1300. The source node receives the message and fragments the packet based on MTU 1300. In this way, the source node sends the packet to the destination address and

discovers the PMTU of the transmission path.

PMTU Discovery (2/2)

- The minimum PMTU is 1280 bytes. (The minimum MTU supported by the link layer in IPv6 is 1280 bytes.)
- The maximum PMTU is determined by the link layer. For example, a tunnel supports a large MTU.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

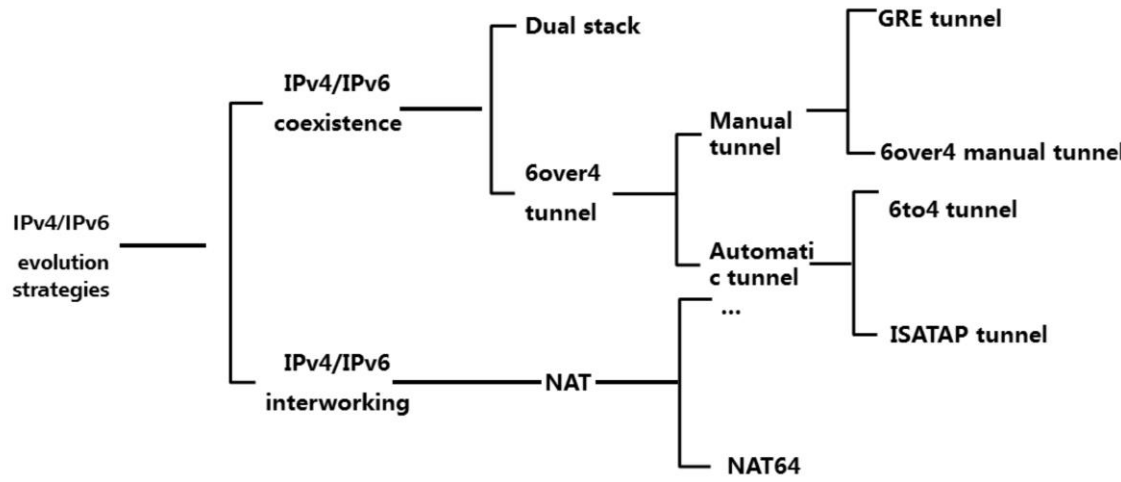
- Note that the PMTU discovery mechanism takes effect only when a transmitted data packet exceeds the minimum PMTU. If the packet size is smaller than the minimum PMTU, a Packet Too Big message cannot be generated.
- IPv6 allows a minimum MTU of 1280 bytes. Therefore, the PMTU cannot be smaller than 1280 bytes. The maximum PMTU is determined by the link layer. If the link layer is a tunnel, the PMTU value may be large.

Contents

1. IPv6 Background
2. **IPv6 Implementation**
   - Packet Format
   - Basic Configuration
   - Basic Protocol
   - **Transition Technologies**
3. IPv6 Configuration Commands
4. Suggestions for IPv6 Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Overview of IPv6 Transition Technologies



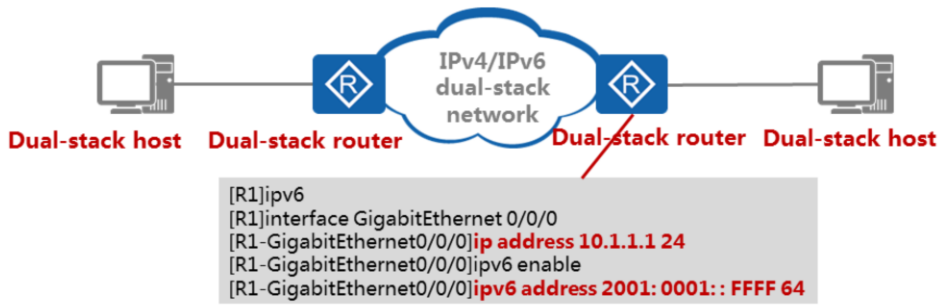
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IPv4/IPv6 coexistence technology:
  - Dual IPv4/IPv6 protocol stack:
    - IPv6 nodes support both IPv4 and IPv6 protocol stacks.
  - Tunnel:
    - IPv6 packets act as the IPv4 payload to connect multiple IPv6 islands on the IPv4 Internet.
- IPv4/IPv6 interworking technology:
  - Provides a technology for mutual access between IPv6 and IPv4.
  - Allows the IPv6 Internet and IPv4 Internet to coexist and communicate with each other.

Dual Stack

- Dual stack protocol
  - Devices must support the IPv4/IPv6 dual stack.
  - Interfaces connected to a dual-stack network must be configured with both an IPv4 address and an IPv6 address.



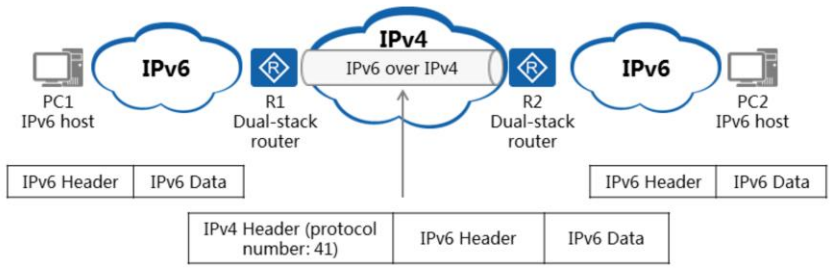
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Dual stack is a technology used for the transition from IPv4 to IPv6. Nodes on a dual stack network support both IPv4 and IPv6 protocol stacks. Source nodes select different protocol stacks based on different destination nodes. Network devices use protocol stacks to process and forward packets based on the protocol type of packets. You can implement dual stack on a unique device or a dual stack backbone network. On the dual stack backbone network, all devices must support both IPv4 and IPv6 protocol stacks. Interfaces connecting to a dual stack network must be configured with both IPv4 and IPv6 addresses.
- In a IPv4/IPv6 dual stack network, hosts or network devices support both IPv4 and IPv6 protocol stacks. If a node supports dual stack, it can use both IPv4 and IPv6 protocol stacks and process both IPv4 and IPv6 data. On a dual stack device, the upper-layer applications prefer the IPv6 protocol stack rather than the IPv4 protocol stack. For example, an application that supports IPv4/IPv6 dual stack firstly sends an Authentication, Authorization, Audit, Account (AAAA) request to the DNS server and turns to send an Authentication, Authorization, Audit, or Account request to the DNS server only when the AAAA request is not replied. IPv4/IPv6 dual stack is the basis of coexistence between IPv4 and IPv6 as well as the transition from IPv4 to IPv6.
  As shown in the preceding figure, routers are dual stack devices. By default, routers support IPv4. Their interfaces are configured with IPv4 addresses. Therefore, these routers can forward IPv4 packets. If you enable the IPv6 data forwarding capability of routers and assign IPv6 unicast addresses to their interfaces, the interfaces can forward IPv6 data. In this case, the IPv4 and IPv6 protocol stacks do not interfere with each other and work independently.

6over4 Manual Tunnel

- 6over4 manual tunnel
  - A type of 6over4 manual tunnel
  - The source address and destination address must be manually specified.
  - It is deployed between two border routers or between a host and a border router.



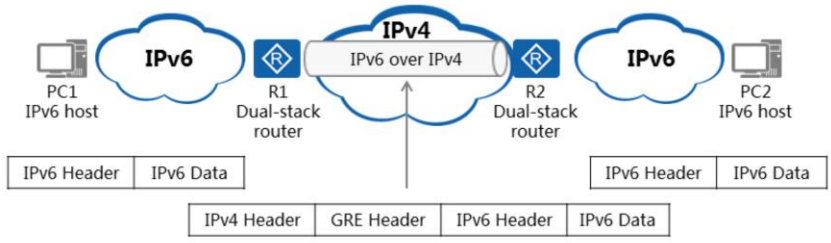
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- If you create multiple IPv6 over IPv4 manual tunnels between one border device and multiple devices, the configuration workload is heavy. Therefore, an IPv6 over IPv4 manual tunnel is commonly created between two border routers to connect IPv6 networks.
- Advantages and disadvantages of manual tunnels
  - Advantages: A manually configured tunnel applies to any environment in which IPv6 traverses IPv4.
  - Disadvantages: The tunnel must be manually configured.
- Forwarding mechanism
  - The forwarding mechanism of an IPv6 over IPv4 manual tunnel is as follows: After a border device receives a packet from the IPv6 network, it searches the destination address of the IPv6 packet in the routing table. If the packet is forwarded from a virtual tunnel interface, the device encapsulates the packet based on the source and destination IPv4 addresses configured on the interface. The IPv6 packet is encapsulated as an IPv4 packet and processed by the IPv4 protocol stack. The encapsulated packet is forwarded through the IPv4 network to the remote end of the tunnel. After the border router on the remote end of the tunnel receives the encapsulated packet, it decapsulates the packet and processes the packet using the IPv6 protocol stack.

6over4 GRE Tunnel

- 6over4 GRE tunnel
  - A type of 6over4 manual tunnel
  - The endpoint addresses of a 6over4 GRE tunnel are manually specified.
  - GRE carries the IPv6 protocol.

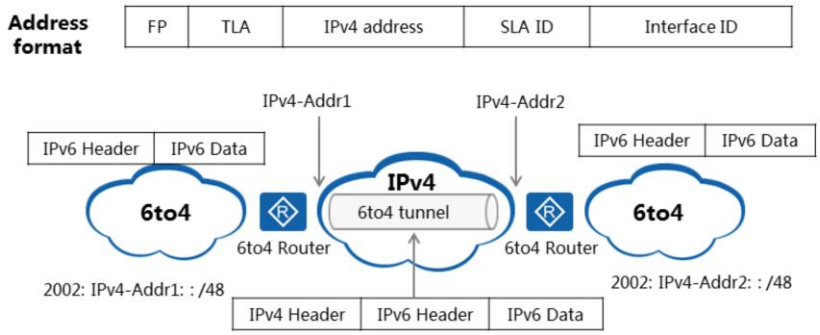


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- An IPv6 over IPv4 GRE tunnel uses the standard GRE tunneling technology to provide P2P connections. You must manually specify addresses for both ends of the tunnel. Any types of protocol packets that GRE supports can be encapsulated and transmitted through a GRE tunnel. The protocols may include IPv4, IPv6, Open Systems Interconnection (OSI), and Multiprotocol Label Switching (MPLS).
- The forwarding mechanism of an IPv6 over IPv4 GRE tunnel is the same as that of an IPv6 over IPv4 manual tunnel.

6to4 Tunnel (1/3)

- 6to4 tunnel
  - A type of 6over4 automatic tunnel
  - A 6to4 tunnel can be established between two routers, two hosts, or a host and a router.
  - A 6to4 dedicated address 2002:IPv4::/48 is used.



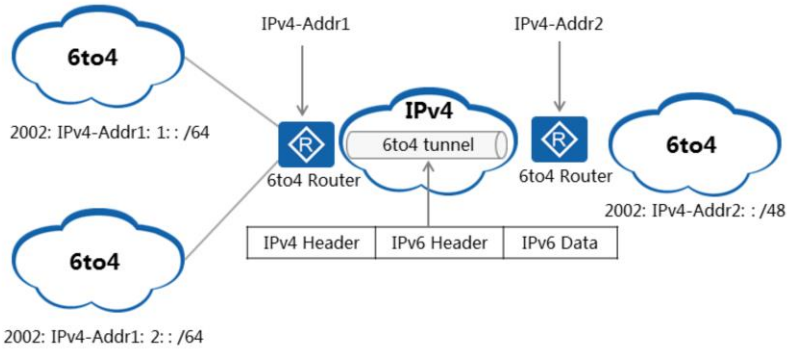
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An IPv6-to-IPv4 (6to4) tunnel is an automatic tunnel and uses an IPv4 address that is embedded in an IPv6 address. Unlike IPv4-compatible IPv6 tunnels, you can create 6to4 tunnels between two routers, a router and a host, and two hosts.
- Address format:
  - FP: format prefix of a global unicast address. The value is 001.
  - TLA ID: top level aggregation identifier. The value has 13 bits and is converted into 0 0000 0000 0010 in binary notation.
  - SLA ID: site level aggregation identifier.
- A 6to4 address is expressed in the format of 2002::/16. A 6to4 network is expressed as 2002:IPv4 address::/48. A 6to4 address has a 64-bit prefix composed of a 48-bit 2002:IPv4 address and a 16-bit SLA. A 2002:IPv4 address in the format of 2002:a.b.c.d is determined by the IPv4 address allocated to the router and the SLA is defined by the user.

6to4 Tunnel (2/3)

- 6to4 tunnel
  - A 6to4 tunnel can be connected to multiple 6to4 networks.
  - A 6to4 tunnel can be identified by SLA ID.



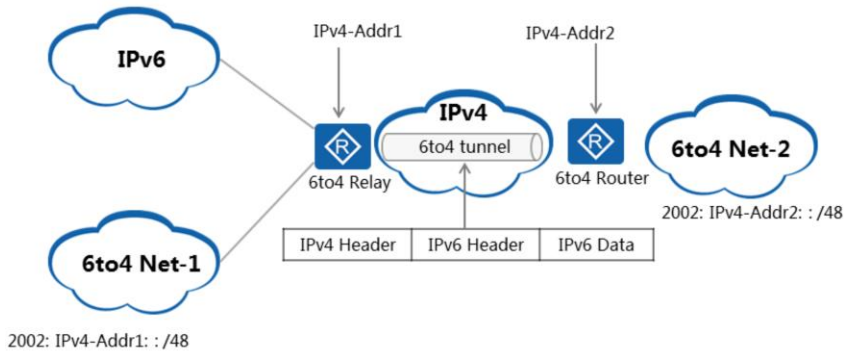
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- One IPv4 address can be used as the source address of only one 6to4 tunnel. When a border router connects to multiple 6to4 networks using the same IPv4 address as the source address of the tunnel, the 6to4 networks share a tunnel and are identified by SLA ID in the 6to4 address.

6to4 Tunnel (3/3)

- 6to4 relay
  - A 6to4 tunnel enables interworking between a 6to4 network and an IPv6 network.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

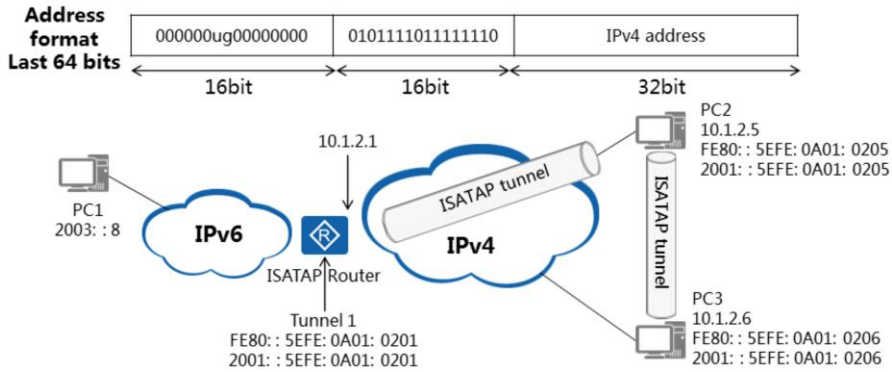
HUAWEI

- A 6to4 relay router is used when a common IPv6 network communicates with a 6to4 network over an IPv4 network. A 6to4 relay is a next-hop device that forwards IPv6 packets of which the destination address is not a 6to4 address but the next-hop address is a 6to4 address. The tunnel destination IPv4 address is obtained from the next-hop 6to4 address.

- If a host on the 6to4 network 2 needs to communicate with the IPv6 network, the next hop of the route must be configured as the 6to4 address of the 6to4 relay on the border router. The 6to4 address of the relay router matches the source address of the 6to4 tunnel of the relay router. A packet sent from the 6to4 network 2 to the IPv6 network is forwarded to the 6to4 relay router according to the next hop indicated by the routing table. The 6to4 relay router then forwards the packet to the IPv6 network. When a packet needs to be sent from the IPv6 network to the 6to4 network 2, the 6to4 relay router encapsulates the packet as an IPv4 packet according to the destination address (a 6to4 address) of the packet so that the packet can be successfully sent to the 6to4 network 2.

ISATAP Tunnel

- Intra-Site Automatic Tunnel Addressing Protocol (ISATAP) tunnel
  - A type of 6over4 automatic tunnel
  - A 6to4 tunnel can be established between hosts, between a host to a router, or between a router to a host.
  - An ISATAP tunnel dedicated address is used.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Intra-Site Automatic Tunnel Addressing Protocol (ISATAP) is another automatic tunneling technology. The ISATAP tunnel uses a specially formatted IPv6 address with an IPv4 address embedded into it. Different from the 6to4 address that uses the IPv4 address as the network prefix, the ISATAP address uses the IPv4 address as the interface ID.
- Address description
- The "u" bit in the IPv4 address that is globally unique is set to 1. Otherwise, the "u" bit is set to 0. "g" is the IEEE individual/group bit. An ISATAP address contains an interface ID and it can be a global unicast address, link-local address, ULA address, or multicast address. A device obtains the first 64 bits of an ISATAP address by sending Request packets to an ISATAP router. Devices on both ends of an ISATAP tunnel run the Neighbor Discovery (ND) protocol. The ISATAP tunnel considers the IPv4 network as a non-broadcast multiple access (NBMA) network.
- Description of the forwarding process:
- PC 2 and PC 3 are located on an IPv4 network. They both support dual protocol stacks and have private IPv4 addresses. You can perform the following operations to enable the ISATAP function on PC 2 and PC 3:
  - Configure an ISATAP tunnel interface to generate an interface ID based on the IPv4 address.
  - Generate a link-local IPv6 address based on the interface ID. When a host obtains the link-local IPv6 address, it can access the IPv6 network on the local link.
  - The host automatically obtains a global unicast IPv6 address and ULA address.
  - The host obtains an IPv4 address from the next hop IPv6 address as the destination address, and forwards packets through the tunnel interface to communicate with another IPv6 host. If the destination host is located on the same site as the source host, the next hop address is the address of the destination host. If the

destination host is not located on the local site, the next hop address is the address of the ISATAP router.

NAT64

- NAT64 is a protocol conversion technology that can flexibly convert packets between IPv4 and IPv6.
- NAT64 maps fields in IPv6 packets to fields in IPv4 packets, so that data packets are converted at the edge between the two types of networks.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Nodes on an IPv4 network cannot directly communicate with nodes on an IPv6 network by default, because the two protocol stacks are incompatible. However, this problem can be resolved if a device implements conversion between IPv6 and IPv4 protocols.



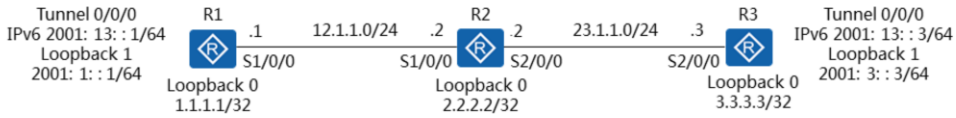
Contents

1. IPv6 Background
2. IPv6 Implementation
3. **IPv6 Configuration Commands**
4. Suggestions for IPv6 Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuring a GRE Tunnel (1/2)

- The following figure shows the network topology of company A. Complete the following configurations as required:
  - IPv4 addresses of R1, R2, and R3 are shown in the following figure and the three devices are deployed in Area 0 of OSPFv2. The configurations are complete.
  - The required IPv6 addresses are marked.
  - An IPv6 over IPv4 GRE tunnel is deployed to implement the interworking between Loopback1 interfaces of R1 and R3.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Case description:
- IPv6 and IPv4 addresses have been specified.

GRE Tunnel — Basic Configuration Commands

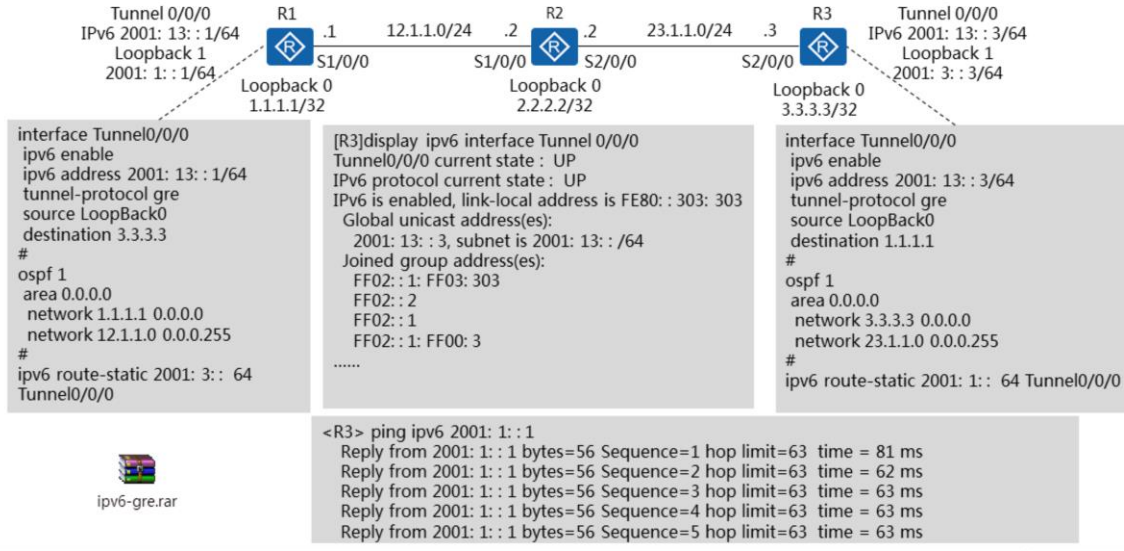
Step	Operation	Command
1	Enter the system view.	**system-view**
2	Create a tunnel interface.	**interface tunnel** *interface-number*
3	**Set the tunnel mode to GRE.**	**tunnel-protocol gre**
4	Specify the source interface of the tunnel.	**source** { *ipv4-address* | *interface-type interface-number* }
5	Specify the destination interface of the tunnel.	**destination ipv4-address**
6	**Configure the GRE packet header check.**	**gre checksum**
7	**Set the keyword of the GRE packet header.**	**gre key key-number**
8	Configure an IPv6 address for the tunnel interface.	**ipv6 address** *ipv6-address prefix-length*

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Checking the GRE packet header is optional. If the GRE packet header check is configured, the sender calculates the checksum based on information in the GRE header and the payload and sends the packet containing the checksum to the receiver. The receiver calculates the checksum based on information in the received packet and compares the calculated value with the checksum in the packet. If they are the same, the receiver forwards the packet. If they are different, the receiver discards the packet. If the checksum function is enabled on the local end and disabled on the remote end, the local end does not perform checksum on received packets.
- Setting keywords for the GRE packet header is optional. If the KEY field in the GRE packet header is set, the receiver checks the keyword of the received GRE packet header. If it is consistent with the keyword locally configured, the authentication is successful. Otherwise, the packet is discarded.

Configuring a GRE Tunnel (2/2)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Meaning of the commands:
  - The **interface tunnel** command creates a tunnel interface and enters the tunnel interface view.
  - The **tunnel-protocol gre** command sets the tunnel mode to manual.
  - The **source {ipv4-address | interface-type interface-number}** command specifies the source interface of a tunnel.
  - The **destination { ipv4-address }** command specifies the destination interface of a tunnel.
  - The **ipv6 address {ipv6-address prefix-length}** command sets the IPv6 address of the tunnel interface.



Contents

1. IPv6 Background
2. IPv6 Implementation
3. IPv6 Configuration Commands
4. **Suggestions for IPv6 Exam Preparation**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Suggestions for IPv6 Exam Preparation

- Have a good command of IPv6 protocol commands.
- Be familiar with the implementation and application of IPv6.
- Be familiar with the use of policy tools such as IPv6.
- Read the HedEx documents.
  - Including relevant cases.
- Understand the case scenarios designed in the course.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Recommendations

- Huawei E-Learning website:
  - http://support.huawei.com/learning/Index!toTrainIndex
- Huawei support case library:
  - http://support.huawei.com/enterprise/servicecenter?lang=en

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— OSPF

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- The Open Shortest Path First (OSPF) protocol, developed by the Internet Engineering Task Force (IETF), is a link-state Interior Gateway Protocol (IGP).
- OSPF features fast convergence, loop-free routing, and high scalability. It is being widely used.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to: Understand OSPF fundamentals.
- Improve your OSPF fault troubleshooting capability.
- Enhance your comprehensive capability of using OSPF.
- Understand OSPFv3 fundamentals.
- Understand the differences between OSPFv3 and OSPFv2.
- Understand OSPFv3 configuration methods and common commands.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

- **OSPFv2 Fundamentals**
  - OSPFv2 Basic Concepts and Features
  - Comparison Between OSPF and IS-IS
- OSPFv2 Fault Diagnosis
- OSPFv3 Introduction
- Differences Between OSPFv3 and OSPFv2
- OSPFv3 Configuration Methods and Common Commands
- Suggestions on OSPF Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

OSPFv2 Basic Concepts - LSA Types

LSA Type	Function
Router-LSA (Type 1)	Describes the link state and cost of a device. A router-LSA is generated by each device and advertised within the area to which the device belongs.
Network-LSA (Type 2)	Describes the states of all links on a network segment. A network-LSA is generated by a designated router (DR) and advertised within the area to which the DR belongs.
Network-summary-LSA (Type 3)	Describes routes on a network segment. A network-summary-LSA is generated by an area border router (ABR) and advertised to non-totally stub areas or not-so-stubby areas (NSSAs).
ASBR-summary-LSA (Type 4)	Describes the route to an autonomous system boundary router (ASBR). An ASBR-summary-LSA is generated by an ABR and advertised to all related areas except the area to which the ASBR belongs.
AS-external-LSA (Type 5)	Describes the route to a destination outside an autonomous system (AS). An AS-external-LSA is generated by an ASBR and advertised to all areas except stub areas and NSSAs.
NSSA LSA (Type 7)	Describes the route to a destination outside an AS. An NSSA LSA is generated by an ASBR and advertised only within NSSAs.
Opaque LSA (Type 9/Type 10/Type 11)	Provides a universal mechanism for OSPF extensions: A Type 9 LSA, such as a grace LSA used in graceful restart (GR), is advertised only on the network segment where the interface advertising the LSA resides. A Type 10 LSA, such as an LSA used to support TE, is advertised within an area. A Type 11 LSA is advertised within an AS. At present, there are no applications for Type 11 LSAs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- LSA header information (all OSPF packets, except Hello packets, carry LSA information):
  - LS age: indicates the time that has elapsed after the LSA is generated, in seconds.
  - Options: indicates the optional capabilities supported by a device.
  - LS type: indicates the format and function of an LSA. There are five types of commonly used LSAs.
  - Link State ID: This field's value varies according to the LSA.
  - Advertising Router: indicates the router ID of an LSA originator.
  - Sequence Number: detects old and duplicate LSAs. The LSA sequence number is incremented each time a router originates a new instance of the LSA. This update helps other routers identify the latest LSA instance.
  - Checksum: indicates the checksum of the complete content of an LSA. Due to the age field, the checksum is recalculated each time the aging time increases.
  - Length: indicates the length of an LSA, including the length of the LSA header.
- A router-LSA must describe the states of all interfaces or links of an LSA originating router.
  - Link State ID: indicates the router ID of an LSA originating router.
  - Flag:
    - V: If it is set to 1, an LSA originating router is an endpoint of one or more virtual links with complete adjacencies.
    - E: It is set to 1 if an originating router is an ASBR.
    - B: It is set to 1 if an originating router is an ABR
  - Number of links: indicates the number of router links.

    - Link Type:
        - If it is set to 1, a network is a point-to-point network. For common PPP links, point-to-point networks need to be used.
        - If it is set to 2, a link is connected to a transit network. A transit network segment contains the broadcast or NBMA network segments of at least two routers.
        - If it is set to 3, a link is connected to a stub network. Generally, the network has no neighbor relationships established, such as an Ethernet network that has only one outbound interface or has only loopback interfaces.
        - If it is set to 4, the link is a virtual link.
    - Link ID:
        - If Link Type is set to 1, this field indicates the router ID of a neighbor router.
        - If Link Type is set to 2, this field indicates the interface IP address of a DR router.
        - If Link Type is set to 3, this field indicates an IP network or subnet address.
        - If Link Type is set to 4, this field indicates the router ID of a neighbor router.
    - Link Data:
        - If Link Type is set to 1, this field indicates the IP address of the interface on the connected originating router.
        - If Link Type is set to 2, this field indicates the IP address of the interface on the connected originating router.
        - If Link Type is set to 3, this field indicates the subnet mask of a network.
        - If Link Type is set to 4, this field indicates the IP address of a virtual link interface on the originating router.
    - ToS: not supported currently.
    - Metric: indicates the cost of a link or interface.
- Network-LSA
    - Link State ID: indicates the interface address of a DR router.
    - Network Mask: specifies the address or subnet mask used on the network.
    - Attached router: lists the router IDs of all routers that have a complete adjacency relationship with the DR, including the router ID of the DR.
- Network-summary-LSA and ASBR-summary-LSA
    - Link State ID: For Type 3 LSAs, this field indicates the IP address of the advertised network or subnet. For Type 4 LSAs, this field indicates the router ID of the advertised ASBR.
    - Network Mask: For Type 3 LSAs, this field indicates the subnet mask of the advertised network. For Type 4 LSAs, this field is meaningless and is generally set to 0.0.0.0.
    - Metric: indicates the cost of the route from the originating router to the destination.

- AS-external-LSA
  - Link State ID: indicates the IP address of the advertised network or subnet.
  - Network Mask: indicates the subnet mask of the advertised network.
  - E: specifies the type of the external metric used by the route. If the E bit is set to 1, the metric type is E2. If the E bit is set to 0, the metric type is E1.
  - Metric: indicates the cost of a route. The value is determined by the ASBR.
  - Forwarding Address: indicates the address to which data packets are forwarded. If the forwarding address is 0.0.0.0, data packets will be forwarded to the originating ASBR.
  - External Route Tag: indicates an external route.
- NSSA LSA
  - Forwarding Address: If the next hop of an imported external route is in an OSPF routing domain, the forwarding address is set to the next hop of the imported external route. If the next hop of the imported external route is not in an OSPF routing domain, the forwarding address is set to the IP address of the stub network segment (for example, loopback 0 interface) in an OSPF routing domain on the ASBR. If there are multiple stub network segments, the IP address with the largest IP address is selected.

OSPFv2 Basic Concepts - Route Types

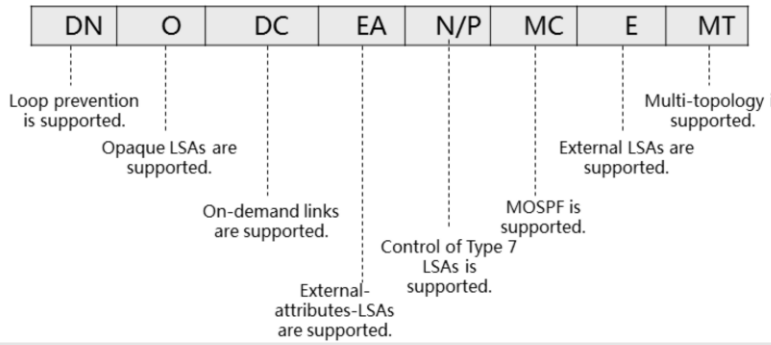
- AS inter-area and intra-area routes describe routes within an AS. AS external routes describe routes to destinations outside the AS.
- OSPF classifies imported AS external routes as Type 1 or Type 2 external routes.
- The following table describes OSPF routes in descending order of priority.

Route Type	Description
Intra Area	Indicates routes within an area.
Inter Area	Indicates routes between areas.
Type 1 external routes	Type 1 external routes have high reliability. The calculated cost of an external route is close to the route cost in the AS and comparable to the OSPF route cost. Cost of a Type 1 external route = Cost of the route from a device to an ASBR + Cost of the route from the ASBR to the destination.
Type 2 external routes	Type 2 external routes have low reliability. OSPF considers that the cost of a route from an ASBR to the destination outside the AS is much greater than the cost of any internal route to the ASBR. Cost of a Type 2 external route = Cost of the route from an ASBR to the destination outside of the AS.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

OSPFv2 Basic Concepts - Options Field

- Options field:
  - The Options field is carried in all Hello packets, DD packets, and LSAs.
  - The Options field defines the optional capabilities that a router supports.
- The Options field contains the following information:



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Description of bits in the Options field:
  - DN: This bit prevents loops on MPLS VPNs. The DN bit is set to 1 if a PE sends a Type 3, Type 5, or Type 7 LSA to a CE. The LSA does not participate in OSPF route calculation on another PE that receives this LSA from the CE.
  - O: This bit indicates the Opaque LSA type (Type 9, Type 10, or Type 11) supported by an originating router.
  - DC: This bit is set to 1 if an originating router supports on-demand links.
  - EA: This bit is set to 1 if an originating router has the capability of receiving and forwarding external-attributes-LSAs (Type 8 LSAs).
  - N: This bit is carried only in Hello packets. If the bit is 1, a router supports Type 7 LSAs. If the bit is 0, a router cannot send or receive NSSA LSAs.
  - P: This bit is carried only in NSSA LSAs. This bit is used to instruct the ABR of an NSSA to translate Type 7 LSAs into Type 5 LSAs.
  - MC: This bit is set to 1 if an originating router can forward multicast data packets.
  - E: This bit is set to 1 if an originating router can receive AS-external-LSAs (Type 5 LSAs). This bit is set to 1 in all Type 5 LSAs and the LSAs that are originated in backbone or non-stub areas. However, it is set to 0 in LSAs originated in stub areas. If the bit is set to 1 in a Hello packet, an interface can send and receive Type 5 LSAs.
  - MT: This bit indicates that an originating router supports OSPF multi-topology.

OSPFv2 Convergence Features

- Fast convergence:
  - Incremental shortest path first (I-SPF) algorithm:
    - Performs route calculation only for affected nodes.
    - Performs route calculation for all nodes only at the first time of calculation.
  - Partial route calculation (PRC) algorithm:
    - Performs route recalculation only for changed routes.
    - Updates routes based on the shortest path tree (SPT) calculated by I-SPF.
  - Intelligent timer:
    - Intelligently controls network instabilities, such as those caused by interface flapping, to reduce LSA flooding.
- Priority-based convergence:
  - Allows routes with a higher priority to converge first.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

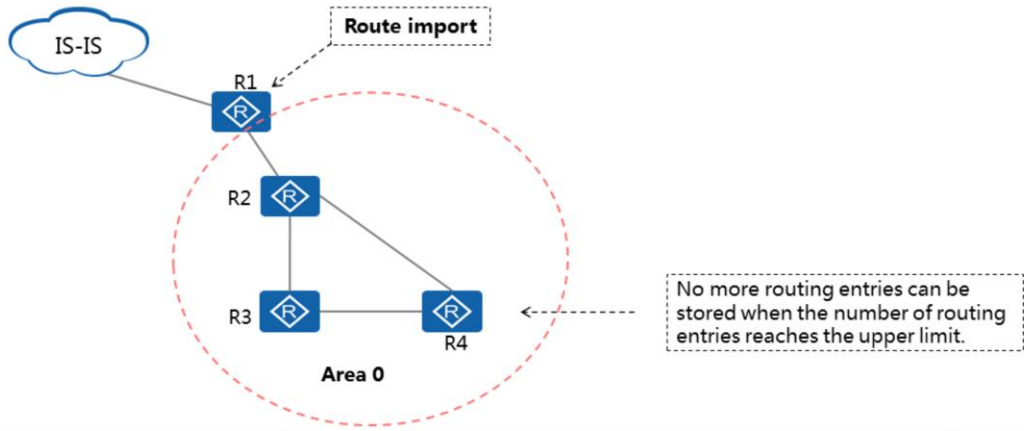
HUAWEI

- Fast convergence:
  - I-SPF performs route calculation only for affected nodes, except that the calculation is performed for the first time. The generated SPT is the same as the SPT generated when another conventional algorithm is used. Therefore, compared with SPF, I-SPF consumes less CPU resources and speeds up network convergence.
  - Similar to I-SPF, PRC calculates only changed routes. However, PRC does not calculate SPTs. Instead, it uses SPTs calculated by I-SPF to update routes. In route calculation, a leaf represents a route, and a node represents a router. Either an SPT or a leaf change causes a route change. The SPT change is irrelevant to the leaf change. PRC processes routing information as follows:
    - If the SPT changes, PRC processes the routing information of all leaves on a changed node.
    - If the SPT remains unchanged, PRC does not process the routing information on any node.
    - If a leaf changes, PRC processes the routing information on the leaf only.
    - If a leaf remains unchanged, PRC does not process the routing information on the leaf.
- Intelligent timer: OSPF uses an intelligent timer to control route calculation, LSA generation, and LSA receiving. This speeds up route convergence. The OSPF intelligent timer works as follows:
  - On a network where routes are calculated repeatedly, the OSPF intelligent timer dynamically adjusts route calculation based on user configuration and the exponential backoff technology to reduce the number of route calculations and CPU resource consumption. Routes are calculated after the network topology stabilizes.

- On an unstable network, if a router generates or receives LSAs due to frequent topology changes, the OSPF intelligent timer can dynamically adjust the route calculation interval. No LSAs are generated or processed within the interval, preventing invalid LSAs from being generated or advertised across the entire network.
- The functions of the intelligent timer for path computation are as follows:
  - According to the local LSDB, an OSPF router uses the SPF algorithm to calculate the shortest path tree with itself as the root, and determines the next hop to the destination network according to the shortest path tree. You can set a proper SPF calculation interval to prevent frequent network changes from exhausting bandwidth and router resources.
  - On a network where the convergence time of routes is required to be shorter, set millisecond as the unit of interval to increase the frequency of calculating routes. This increases route convergence.
  - When an OSPF LSDB changes, shortest paths must be recalculated. If the network changes frequently and the shortest path is calculated frequently, many system resources are occupied and thus the efficiency of routers is degraded. You can configure an intelligent timer and set a proper interval for calculating SPF to avoid the occupation of too many bandwidth and router resources.
  - Details about the interval for the SPF calculation are as follows:
  - The initial interval for the SPF calculation is specified by the parameter start-interval.
  - The interval for the SPF calculation for the nth ($n \geq 2$) time is equal to hold-interval x 2 x (n-1).
  - When the interval specified by hold-interval x 2 x (n-1) reaches the maximum interval specified by max- interval, OSPF performs SPF calculation at the maximum interval for three consecutive times, and then goes back to the first step. That is, OSPF performs SPF calculation at the initial interval specified by start-interval.
  - Priority-based convergence:
  - You can configure a device to filter specific routes based on an IP prefix list. You can configure different convergence priorities for different routes so that important routes can be converged first, improving network reliability.

OSPFv2 - Database Overflow

- Database overflow:
  - Limit the number of non-default external routes to prevent database overflow.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Setting the maximum number of non-default external routes on a router can prevent database overflow. All routers on the OSPF network must be configured with the same upper limit. If the number of external routes on a router reaches the upper limit, the router enters the overflow state and starts an overflow timer. The router automatically exits the overflow state after the timer (5 seconds by default) expires.
- OSPF database overflow process:
  - When a router enters the overflow state, it deletes all locally generated non-default external routes.
  - In the overflow state, the router does not generate non-default external routes or reply with acknowledgment packets when receiving non-default external routes. Instead, it discards newly received non-default external routes. When the overflow timer expires, the router checks whether the number of external routes exceeds the upper limit. If so, the router restarts the timer. Otherwise, the router exits the overload state.
  - When the router exits the overflow state, it deletes the overflow timer, and can generate non-default external routes, permit newly received non-default external routes, reply with acknowledgment packets in response to received non-default external routes, and prepare to enter the overload state next time.

OSPFv2 - Default Routes

- **Common area**:
  - Manually configure an ASBR to generate a Type 5 LSA and advertise it within the entire OSPF AS (except special areas in the AS).
- Stub area:
  - An ABR automatically generates a default Type 3 LSA and advertises it to the entire stub area.
- Totally stub area
  - An ABR automatically generates a default Type 3 LSA and advertises it to the entire stub area.
- NSSA:
  - An ABR automatically generates a default Type 7 LSA and advertises it within the entire NSSA.
  - Manually configure an ASBR to generate a Type 7 LSA and advertise it within the entire NSSA.
- Totally NSSA:
  - An ABR automatically generates a default Type 3 LSA and advertises it within the NSSA.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- OSPF default routes are used when:
  - An ABR advertises default summary LSAs (Type 3) to guide packet forwarding between areas.
  - An ASBR advertises default external ASE LSAs (Type 5) or default external NSSA LSAs (Type 7) to guide packet forwarding to other ASs.
- The principles for OSPF default route advertisement are as follows:
  - An OSPF router can advertise default route LSAs only when the router is connected to an external AS.
  - If an OSPF router has advertised a default route LSA, the router no longer learns the same type of default route advertised by other routers. That is, the router uses only the LSA advertised by itself to calculate routes. The LSAs advertised by other routers are still stored in the LSDB.
  - If a router must use a route to advertise an LSA carrying an external default route, the route cannot be a route learned by a local OSPF process. A router in an area uses an external default route to forward packets outside the area. If the next hops of routes in the area are routers in the area, packets cannot be forwarded outside the area.

OSPFv2 - Route Filtering

- Filtering of routes to be imported
- LSA filtering in the OSPF process view:
  - Filtering of Type 3 LSAs to be learned and advertised
  - Filtering of Type 5 and Type 7 LSAs to be generated
- LSA filtering in the interface view:
- Filtering of routes to be calculated
  - OSPF filters the LSAs in the database when calculating them as routes.
  - Only the routes that are calculated through the SPF algorithm are stored in the routing table. The LSAs used for generating the routes are flooded in the OSPF AS.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- OSPF supports route filtering using routing policies. By default, OSPF does not filter routes.
- These policies include route-policy, filter, filter-policy, filter-LSA-out, access-list, and prefix-list.
- OSPF route filtering can be used to:
  - Filter routes to be imported.
    - OSPF can import routes learned by other routing protocols. Routing policies can be configured to filter routes to be imported, allowing OSPF to import only routes that match specific conditions.
    - Imported routes in the routing table can be advertised.
  - Filter Type 3 LSAs to be learned and advertised.
    - The filter import and filter export commands can be run on an ABR to filter incoming and outgoing Type 3 LSAs. The commands can be run only on ABRs (only ABRs can advertise Type 3 LSAs).
  - Filter Type 5 and Type 7 LSAs to be generated.
    - After OSPF imports external routes, it generates Type 5 and Type 7 LSAs. The filter-policy export command can be run to filter Type 5 and Type 7 LSAs to be generated. This command can be run only on ASBRs.
  - Filter LSAs on specific interfaces. The ospf filter-lsa-out command can be run to filter all Type 3, Type 5, and Type 7 LSAs, except grace LSAs, based on the route prefixes specified in an ACL, so that the LSAs to be advertised can be filtered.
  - Filter LSAs for route calculation.
    - The filter-policy import command can be run to filter intra-area, inter-area, and external LSAs in the database that can be used in route calculation. The filtering function determines whether a route can be added to the local routing table. A route is added to the local routing table only if it matches the filtering rule. The LSA that generates the route is advertised in the OSPF AS.



Contents

- **OSPFv2 Fundamentals**
  - OSPFv2 Basic Concepts and Features
  - Comparison Between OSPF and IS-IS
- OSPFv2 Fault Diagnosis
- OSPFv3 Introduction
- Differences Between OSPFv3 and OSPFv2
- OSPFv3 Configuration Methods and Common Commands
- Suggestions on OSPF Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

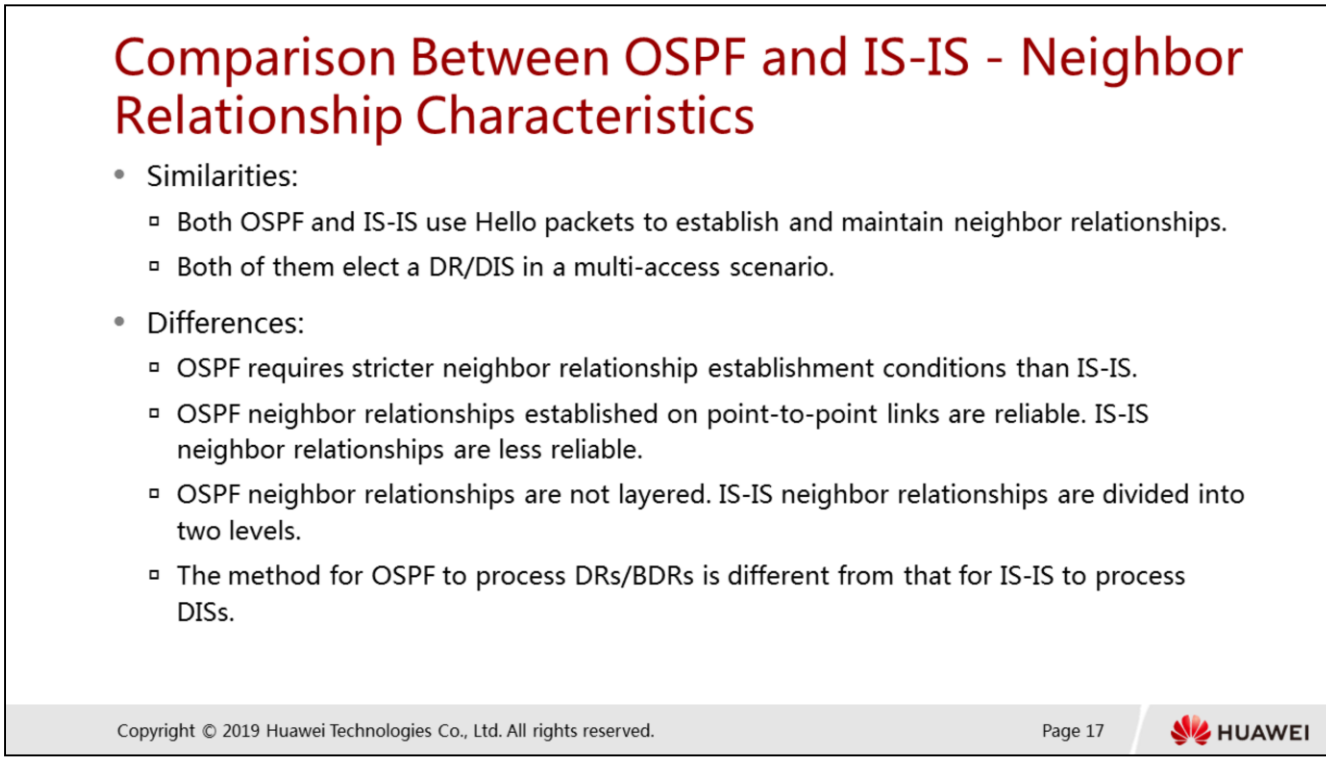
Comparison Between OSPF and IS-IS - Basic Characteristics

- Similarities:
  - Both OSPF and IS-IS are IGPs and widely used.
  - Both of them can be used on IP networks.
  - Both of them support hierarchy and multi-area designs.
- Differences:
  - OSPF can be used only on IP networks. IS-IS can be used on both IP and CLNP networks.
  - OSPF supports various network types. IS-IS supports only two network types.
  - OSPF supports virtual links. IS-IS has similar functions, but most vendors' IS-IS products do not support virtual links.
  - OSPF works at the IP layer. IS-IS works at the data link layer.
  - OSPF divides areas based on interfaces. IS-IS divides areas based on devices.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Related information:
  - OSPF supports P2P, P2MP, NBMA, and broadcast networks. IS-IS supports only P2P and broadcast networks.
  - OSPF works on the IP network and uses the protocol number 89.



- Related information:
  - OSPF checks the information such as the mask, authentication parameter, Hello/dead interval, and area in an Hello packet when establishing a neighbor relationship. The conditions for IS-IS to establish neighbor relationships are relatively loose.
  - OSPF requires three-way handshakes when establishing a neighbor relationship on a P2P link. IS-IS does not require three-way handshakes. However, Huawei devices require three-way handshakes on an IS-IS P2P network by default, improving the reliability of neighbor relationship establishment.
  - IS-IS neighbor relationships are classified into Level-1 or Level-2 relationships.
  - OSPF elects a DR/BDR based on election priorities and router IDs. After the election is complete, the DR/BDR role cannot be preempted. In OSPF, all DRother devices form complete neighbor relationships with the DR/BDR; all DRother devices form 2-way neighbor relationships, which are incomplete relationships. In OSPF, if the election priority of a router is 0, the router does not participate in the DR/BDR election.
  - IS-IS elects a DIS based on election priorities and router MAC addresses. After the election is complete, the DIS role can be preempted. In IS-IS, all routers form adjacencies. If the election priority of a router is 0, the router participates in DIS election with a low priority.

Comparison Between OSPF and IS-IS - Database Synchronization Characteristics

- Similarities:
  - Both OSPF and IS-IS need to form a unified LSDB.
- Differences:
  - There are various types of OSPF LSAs, but few types of IS-IS LSPs.
  - The OSPF and IS-IS database synchronization processes are different.
  - The lifetime of an OSPF LSA increases during transmission. The lifetime of an IS-IS LSP decreases during transmission.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Related information:
  - IS-IS provides few types of LSPs, but can extend functions using the TLV fields in LSPs.

Comparison Between OSPF and IS-IS - Other Characteristics

- Similarities:
  - Both OSPF and IS-IS use SPF to calculate routes.
  - Both of them can prevent loops and support fast convergence and large-scale networks.
- Differences:
  - The OSPF cost type is simple. The IS-IS cost type is complex.
  - OSPF supports on-demand dial-up networks, but IS-IS does not.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Related information:
  - OSPF route costs are based on bandwidth. IS-IS has four types of cost types: narrow, narrow-compatible, wide, and wide-compatible. However, only wide costs are used.



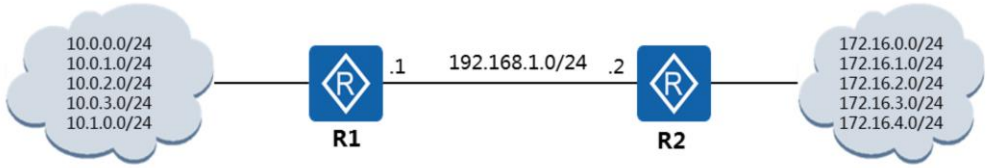
Contents

- OSPFv2 Fundamentals
- **OSPFv2 Fault Diagnosis**
- OSPFv3 Introduction
- Differences Between OSPFv3 and OSPFv2
- OSPFv3 Configuration Methods and Common Commands
- Suggestions on OSPF Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

OSPF Fault Diagnosis

- After OSPF is enabled across the network, R1 and R2 cannot ping each other.
- How do you analyze and rectify the fault?



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Troubleshooting Procedure

- Control plane:
  - If the neighbor relationship is Down:
    - Check whether the physical interfaces work normally.
    - Check whether the mask, authentication, and area ID configurations are consistent.
    - Check whether the corresponding interface information is advertised by OSPF.
  - If the neighbor relationship is in the Init state:
    - Check whether the local interface and the peer device are faulty.
  - If the neighbor relationship is in the 2-way state:
    - Check whether the interface priority is 0.
  - If the neighbor relationship is in the Exstart state:
    - Check whether the **ospf mtu-enable** command is run on the interfaces. If so, check whether the MTUs are consistent.
  - If the neighbor relationship is in the Exchange/Loading state:
    - Check whether the local and peer devices and interfaces are faulty.
- Data plane:
  - Check whether traffic filtering is configured on the interconnected interfaces.
  - Check whether policy-based routing is configured on the interconnected interfaces.
  - Check whether ping packets are discarded by the intermediate firewall.
  - If the neighbor relationship is in the full state:
    - Check whether the network types of the interfaces are consistent.
    - Check whether the **network** command's parameter settings are correct.
    - Check whether external routes to be imported are filtered based on the configured route-policy.
    - Check whether Type 5 and Type 7 LSAs to be advertised are filtered based on the **filter-policy export** configuration.
    - Check whether local routes to be used in route calculation are filtered based on the **filter-policy import** configuration.
    - Check whether inter-area LSA filtering is configured in the OSPF process view.
    - Check whether the devices are disabled from advertising summary Type 3, Type 5, and Type 7 LSAs in the OSPF process view.
    - Check whether the interfaces are configured to filter LSAs to be advertised.
    - Check whether a route that has a longer mask bit or higher protocol priority and different outbound interface from that of the OSPF route is generated by another protocol.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- By default, OSPF does not check the MTUs of DD packets.



Contents

- OSPFv2 Fundamentals
- OSPFv2 Fault Diagnosis
- **OSPFv3 Introduction**
- Differences Between OSPFv3 and OSPFv2
- OSPFv3 Configuration Methods and Common Commands
- Suggestions on OSPF Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

OSPFv3 Introduction

- OSPFv3 has a similar working mechanism as OSPFv2. To support the IPv6 address format, OSPFv3 has made some changes to OSPFv2.
- OSPFv3 is the enhanced version of OSPFv2 but independent of OSPFv2.
- OSPFv3 also uses 89 as the protocol number and a 32-bit unsigned integer as the router ID.

OSPFv2	OSPFv3
IPv4	IPv6
Data link layer	

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Impacts of IPv6 on OSPFv3

Changes in IPv6 Addresses	Impact on OSPFv3
An IPv6 address has 128 bits.	The LSA length is increased.
Link-local addresses are added. Link-local addresses are used to send packets on links, excluding VLinks.	
Multiple global unicast IPv6 addresses can be configured on an interface.	Communication is based on links, instead of subnets.
Extension header for IPv6 authentication	Extension headers are used for authenticating and encrypting packets.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Similarities Between OSPFv3 and OSPFv2

- Network and interface types
- Interface and neighbor state machines
- Link state database (LSDB)
- Flooding mechanism
- Five types of OSPF protocol packets: Hello, DD, LSR, LSU, and LSAck
- Route calculation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

- OSPFv2 Fundamentals
- OSPFv2 Fault Diagnosis
- OSPFv3 Introduction
- **Differences Between OSPFv3 and OSPFv2**
- OSPFv3 Configuration Methods and Common Commands
- Suggestions on OSPF Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

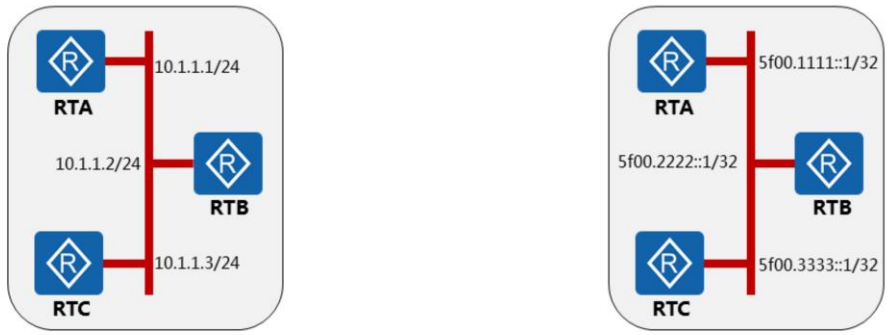
Differences Between OSPFv3 and OSPFv2

- Link-based operation
- Link-local address
- Multi-instance multiplexing
- Authentication
- Stub area
- Packet
- Content in the Options field
- LSA type and content

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Link-based Operation

- OSPFv2 operates based on networks. Two routers must reside on the same network segment if they want to establish a neighbor relationship.
- OSPFv3 is implemented based on links. A link can be divided into multiple subnets. Two nodes can communicate with each other if they are on the same link, regardless of whether they reside on the same subnet.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IPv6 emphasizes the link concept. Multiple IP subnets, that is, IPv6 prefixes, can be allocated to the same link. Different from IPv4, IPv6 allows two nodes on the same link to communicate even if they do not have the same IPv6 prefix. This greatly changes the OSPF behavior.
- OSPFv3 runs based on links rather than IP subnets. In OSPFv3, the concepts "link" and "prefix" are frequently used. However, the two concepts are separated, and there is no necessary mapping relationship between them. Two nodes on the same link can have different prefixes. Therefore, the concepts "network" and "subnet" need to be replaced by "link" when OSPFv3 is used. In addition, an OSPFv3 interface is connected to a link instead of an IP subnet. OSPFv3 made changes in the receiving of OSPF packets and the formats of Hello packets and LSAs.

Link-local Address

- An OSPFv3 router uses a link-local address as the source address for sending packets.
- On a virtual link, the global address or a local address of a site must be used as the source address of an OSPFv3 packet.
- A link-local address is valid only on a local link and can be flooded only on the local link. Therefore, a link-local address can be carried only in a link-LSA.

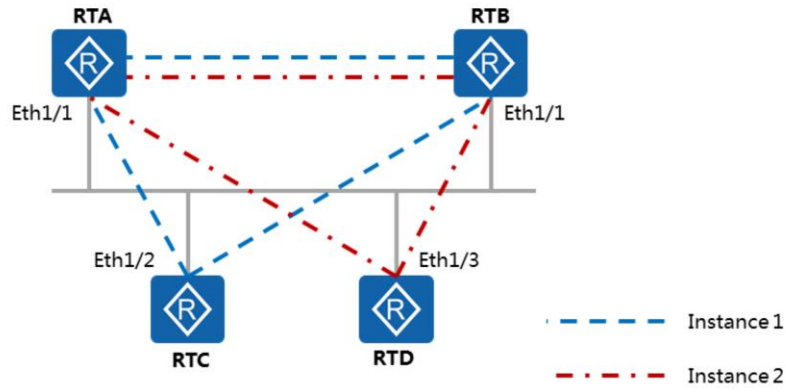
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A router can learn the link-local addresses of all other routers connected to the link and use the link-local addresses as the next hops to forward packets.
- As defined in RFC 2373 for IPv6, a link-local address is for use on a single link to implement functions such as neighbor discovery and auto-configuration. IPv6 routers do not forward packets that carry link-local source addresses. The unicast address range of a link-local address is within an IPv6 address range, FE80/10.

Multiplexing of Multiple Instances

- OSPFv3 allows multiple instances to run on the same link, multiplexing links and reducing costs.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Routers A, B, C, and D are connected to the same broadcast network. They share a link and can establish neighbor relationships. Instance 1 is created on Eth1/1 of router A, Eth1/1 of Router B, and Eth1/2 of router C. Instance 2 is created on Eth1/1 of Router A, Eth1/1 of router B, and Eth1/3 of router D. In this manner, routers A, B, and C can establish neighbor relationships. Routers A, B, and D can establish neighbor relationships.
- This is implemented by adding the Instance ID field to OSPFv3 packet headers. If the instance ID configured on an interface is different from the instance ID in a received OSPF v3 packet, the interface discards the packet and does not establish a neighbor relationship.

Authentication

- Verification change:
  - An OSPFv3 packet header does not contain the AuType or Authentication field. OSPFv3 generally depends on the IPv6 extension authentication header to implement verification.
- Checksum change:
  - OSPFv3 uses the standard IPv6 checksum.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- OSPFv3 does not provide the authentication function. Instead, it uses the security mechanism provided by IPv6 to check packet validity. Therefore, the authentication field in OSPFv2 packets is contained in OSPFv3 packet headers.

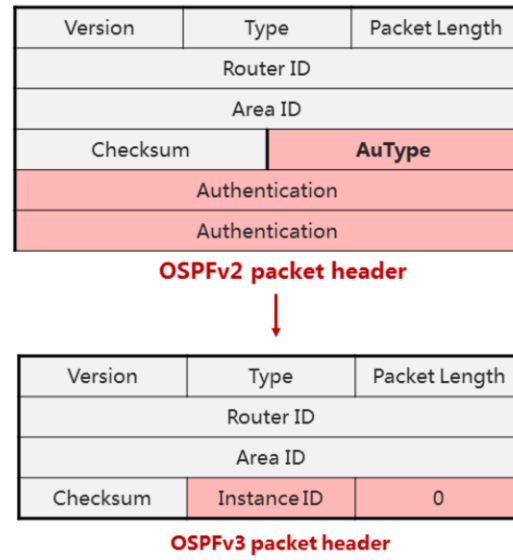
Stub Area

- OSPFv3 can flood unknown LSAs. To prevent a large number of unknown LSAs from being flooded to a stub area, OSPFv3 allows an unknown LSA to be flooded to a stub area only if the LSA can be flooded within an area or on a link and the U bit is not set to 1.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Packet Changes: Header Field Description

- Version: This field is 2 for OSPFv2 and 3 for OSPFv3.
- Type: This field is 1 for Hello packets, 2 for DD packets, 3 for LSR packets, 4 for LSU packets, and 5 for LSAck packets.
- Packet Length: An OSPFv3 packet has two bytes.
- Router ID
- Area ID
- Checksum
- Instance ID: This field differentiates OSPF instances running on the same link. Instance IDs are valid only for local links.
- Reserved: This field is always 0.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Similar to OSPFv2, OSPFv3 defines the same packet header, but different fields, for the five types of OSPFv3 packets.
- LSU and LSAck packets of OSPFv3 are almost the same as those of OSPFv2. However, the fields in OSPFv3 packet headers, Hello packets, DD packets, and LSR packets are slightly different from those in OSPFv2. Packet changes are as follows:
  - The version number is changed from 2 to 3.
  - Packet header changes: Compared with an OSPFv2 packet header, an OSPFv3 packet header contains only 16 bytes. The authentication field is deleted but the Instance ID field is added. The Instance ID field allows multiple instances to run on the same link and is valid only within the link-local range. If the instance ID of a received Hello packet is different from the instance ID configured for the packet receiving interface, the interface does not establish a neighbor relationship.

Packet Changes: Hello Packets

OSPFv2 Hello packet header

Version	Type	Packet Length
Router ID		
Area ID		
Checksum	AuType	
Authentication		
Authentication		
Network Mask		
HelloInterval	Options	Rtr Priority
RouterDeadInterval		
Designated Router ID		
Backup Designated Router ID		
Neighbor		

OSPFv3 Hello packet header

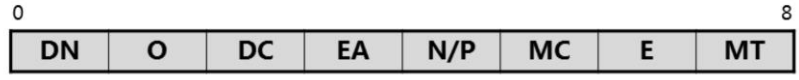
Version	Type	Packet Length
Router ID		
Area ID		
Checksum	Instance ID	0
Interface ID		
Rtr Priority	Options	
HelloInterval	RouterDeadInterval	
Designated Router ID		
Backup Designated Router ID		
Neighbor		

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

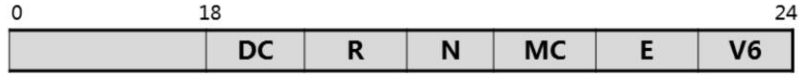
HUAWEI

- Compared with an OSPFv2 Hello packet, an OSPFv3 Hello packet does not have the Network Mask field but has the Interface ID field to identify the ID of the interface that sends the Hello packet.
- Instance ID: 4 bytes. It indicates the ID of the packet sending interface. This field is used to differentiate packet sending interfaces on the same router but does not contain address information.
- Rtr Pri: 1 byte. It indicates the router priority. The router with the highest priority becomes the DR.
- Options: 3 bytes. OSPFv3 extends the Options field to 24 bits.

Packet Changes: Options Field



Format of the Options field in an OSPFv2 packet



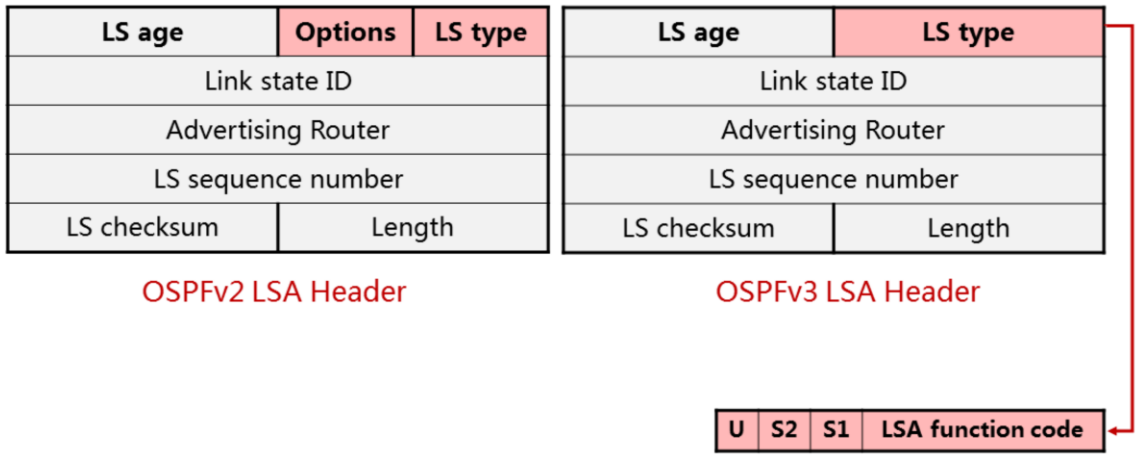
Format of the Options field in an OSPFv3 packet

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In OSPFv2, the Options field is contained in all Hello packets, DD packets, and LSAs.
- In OSPFv3, the Options field is contained only in Hello packets, DD packets, router LSAs, network LSAs, inter-area-router LSAs, and link-LSAs.
- As shown in the preceding figure, OSPFv3 adds the R and V6 bits in the Options field.
  - R bit: indicates whether a device is a router with the forwarding capability. If the R bit is set to 0, the routing information of the device does not participate in route calculation. If the current device is not expected to forward packets that carry non-local addresses, the R bit can be set to 0.
  - V6 bit: If the V6 bit is set to 0, the router or link does not participate in IPv6 route calculation.
  - E: If the value is 0, AS-external-LSA flooding is not supported.
  - MC: related to multicast.
  - N: indicates whether an area is an NSSA.
  - DC: indicates whether on-demand dialing is supported.
- The bit matching results of the Options field have different impacts:
  - If the Options field of a Hello packet is not matched, a neighbor relationship cannot be established.
  - If the E bit is 0, AS-external-LSAs cannot be flooded.
  - If the V6 bit is 0, a router does not participate in IPv6 route calculation.
  - The Options field enables an OSPF router to support optional capabilities and advertise its capabilities to other routers. With this mechanism, routers with different capabilities can work together in an OSPF routing domain.

LSA Format Differences - LSA Header



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The LS Type field specifies the LSA type. This field has 8 bits in an OSPFv2 LSA, but 16 bits in an OSPFv3 LSA.
  - U bit: indicates how a router processes unknown LSAs. The value 0 indicates that unknown LSAs are processed as LSAs with link-local addresses. The value 1 indicates that unknown LSAs are processed based the flooding scope identified by the S2 and S1 bits.
  - S2 and S1 bits: indicate the flooding scope of LSAs. The value 00 indicates that LSAs are flooded only on the local link that generates the LSAs. The value 01 indicates that LSAs are flooded in the area where the router that generates the LSA resides. The value 10 indicates that LSAs are flooded in the entire AS. The value 11 is reserved.
  - LSA Function Code: describes the LSA type.

LSA Type, Support for the Processing of Unknown LSAs

- U bit: indicates how a router processes the LSAs that cannot be identified.

U bit	Processing Method
0	Processes unknown LSAs as those carrying link-local addresses.
1	Stores and floods unknown LSAs.

- S2 and S1 bits: identify the flooding scope of LSAs.

S2	S1	Flooding Scope
0	0	Link-local scope
0	1	Area scope
1	0	AS scope
1	1	Reserved

- Different LSA types correspond to different U, S2, and S1 bits.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In OSPFv2, unknown LSAs are directly discarded.
- In OSPFv3, the U bit in the LS Type field of an unknown LSA identifies how the unknown LSA is processed.
  - If the U bit is set to 1, the unknown LSA is flooded in the scope defined in the LS Type field of the LSA.
  - If the U bit is set to 0, the unknown LSA is flooded only on the link.
- The LSA flooding scope is defined in the LS Type field of the LSA. Currently, there are three types of LSA flooding scopes.
  - Link-local scope
    - LSAs are flooded only on local links. Link-LSAs are added in OSPFv3.
  - Area scope
    - Router-LSAs, network-LSAs, inter-area-prefix-LSAs, inter-area-router-LSAs, and intra-area-prefix-LSAs (new in OSPFv3) are flooded within an area.
  - AS scope
    - LSAs (AS-external-LSAs) are flooded within an entire AS.

LSA Type - Function Code

LSA Function code	LS Type	Description
1	0x2001	Router-LSA
2	0x2002	Network-LSA
3	0x2003	Inter-Area-Prefix-LSA
4	0x2004	Inter-Area-Router-LSA
5	0x4005	AS-External-LSA
6	0x2006	Group-membership-LSA
7	0x2007	Type-7-LSA
8	0x0008	Link-LSA
9	0x2009	Intra-Area-Prefix-LSA

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Link State ID

- This field has 4 bytes. It does not contain address information. The following table describes the meaning of this field in different types of OSPFv2 and OSPFv3 LSAs.

LSA	OSPFv2 LS ID	OSPFv3 LS ID
Router-LSA	Router ID	Local unique 32-bit integer
Network-LSA	IPv4 address of DR	ID of a DR interface
Type 4 LSA	Router ID	Local unique 32-bit integer
Type 3, 5, or 7 LSA	IPv4 network segment	Local unique 32-bit integer
Link-LSA	---	Interface ID
Intra-Area-Prefix-LSA	----	Local unique 32-bit integer

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

LSA Type Differences - OSPFv3 LSA Types

OSPFv2 LSA	OSPFv3 LSA	Similarity and Difference Between OSPFv3 and OSPFv2 LSAs
Router-LSA	Router-LSA	The names are the same, and the functions are similar. However, this type of LSA does not describe address information in OSPFv3 and is used only to describe the topology of a routing domain.
Network-LSA	Network-LSA	
Network-summary-LSA	Inter-area-prefix LSA	The functions are similar, but the names are different.
ASBR-summary-LSA	Inter-area-router LSA	
AS-external-LSA	AS-External-LSA	The functions and names are the same.
None	Link-LSA	New-LSA
	Intra-area-prefix-LSA	New-LSA

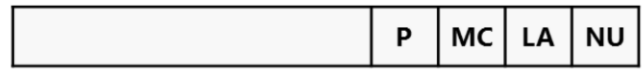
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- OSPFv3 adds link-LSAs and intra-area-prefix LSAs.
  - A router LSA does not contain address information. A router enabled with OSPFv3 generates an independent link-LSA for each link connected to the router. The router advertises the link-local address of the current interface and a series of IPv6 addresses of the router on the link to all other routers on the link.
  - In OSPFv3, router LSAs and network LSAs do not contain routing information. The routing information is described by intra-area-prefix LSAs, which are used to advertise one or more IPv6 address prefixes.

Prefix Presentation Method Changes: Prefix Option Field

- This field is used to describe the features of a prefix so that the system can perform corresponding judgment and processing when different routes are calculated.
  - NU: non-unicast bit
  - LA: local address bit
  - MC: multicast bit
  - P: propagation bit



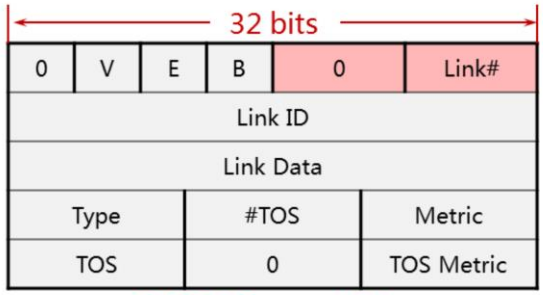
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

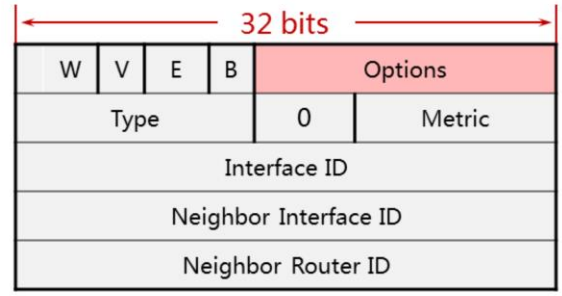
- In OSPFv2, an LSA uses the combination of an IP network segment and a mask to indicate the prefix information. The IP network segment and mask are in different locations of an LSA, and therefore the LSA structure is not clear. In OSPFv3, an LSA uses special triplet information (Prefix-Length, PrefixOptions, and Prefix) to indicate the prefix information. Each prefix advertised by an LSA has its own PrefixOptions field.
- Prefix-Length
  - 1 byte. It indicates the prefix length. The value of this field is 0 for a default route.
- PrefixOptions: 1 byte. It defines the prefix option, which is used to describe some special attribute fields of a prefix. It contains the following bits:
  - NU: non-unicast bit. If this bit is set to 1, the prefix is not considered in IPv6 unicast route calculation.
  - LA: local address bit. If this bit is set to 1, the prefix is an interface address of a router.
  - MC: multicast bit. If this bit is set to 1, the prefix is considered in multicast route calculation. Otherwise, the prefix is not considered in multicast route calculation.
  - P: propagation bit. This bit needs to be set to 1 if the prefix of an NSSA needs to be advertised by an ABR.
- Prefix
  - The length is an integral multiple of 4 bytes. It specifies the IPv6 address of a prefix.
- The prefix length is variable, but must be an integral multiple of 32 bits (4 bytes). It can be filled with 0s. Therefore, the length can be 0, 4, 8, 12, or 16 bytes.

Router LSA

- LS Type: 0x2001; flooding scope: within an area.
- Each router-LSA contains several link descriptions. Each link description is about the information of an interface.
- Multiple router-LSAs can be used, and they are differentiated by link state IDs.



OSPFv2 Router-LSA



OSPFv3 Router-LSA

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- W: multicast route.
- V: A router is an endpoint of a virtual connection.
- E: A router is an ASBR.
- B: A router is an ABR.
- Type: 1 byte. It indicates the type of a link.
- Metric: 2 bytes. It is the cost when a data packet is sent from the interface.
- Interface ID: 4 bytes. It identifies an interface but does not contain address information.
- Neighbor Interface ID: 4 bytes. It identifies the ID of a neighbor interface.
- Neighbor Router ID: 4 bytes. It identifies the router ID of a neighbor.

Router-LSA Link Type

Type	Description	Neighbor Router ID	Neighbor Interface ID
1	Link connected to another router in P2P mode	Router ID of a neighbor	Interface ID of a neighbor
2	Link connected to a transit network	Router ID of a DR	Interface ID of a DR
3	Reserved	---	---
4	Virtual link	Router ID of a neighbor	VLink interface ID of a neighbor

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A router-LSA generated by a router is flooded only within the area where the router resides. This LSA describes all neighbor relationships in the full state of the router. This indicates that a router-LSA does not describe stub links. (In OSPFv2, stub links are described in Type 3 LSAs.) A router-LSA needs to contain a description for each neighbor on a P2MP link. The length of each link description is fixed. Therefore, the number of links in a router-LSA can be determined by the length of the LSA in the LSA header.
- A router-LSA can contain multiple link descriptions. A router can generate multiple router-LSAs, which are differentiated by link state IDs. During SPF calculation, all router-LSAs generated by the same router must be combined.
- An OSPFv3 router-LSA does not contain prefix information, but only describes topology connections.

Router-LSA Example

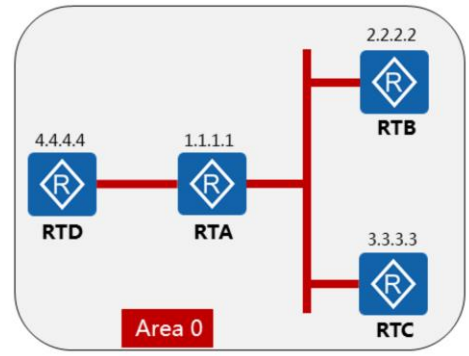
```
<RTA> display ospfv3 lsdb router

  Router-LSA (Area 0.0.0.0)

 LS Age: 872
 LS Type: Router-LSA
 Link State ID: 0.0.0.0
 Originating Router: 1.1.1.1
 LS Seq Number: 0x80000008
 Checksum: 0x0FC2
 Length: 56
 Flags: 0x00 (-|-|-|-)
 Options: 0x000013 (-|R|-|-|E|V6)

   Link connected to: a Transit Network
     Metric: 1
     Interface ID: 0x18000202
     Neighbor Interface ID:  0X18000202      Interface ID of a DR
     Neighbor Router ID:  2.2.2.2      Router ID of a DR

   Link connected to: another Router (point-to-point)
     Metric: 1562
     Interface ID: 0x18000286
     Neighbor Interface ID:  0X18000206      Interface ID of a neighbor
     Neighbor Router ID:  4.4.4.4      Router ID of a neighbor
```

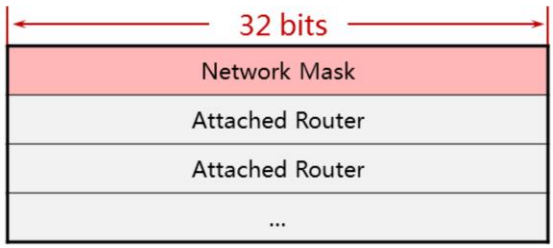


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

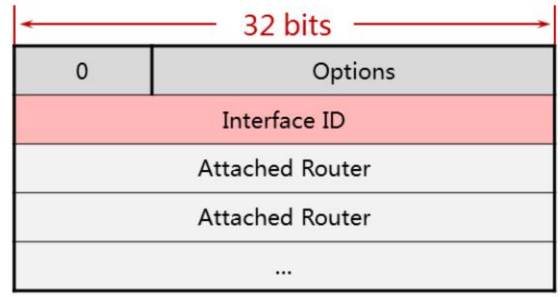
HUAWEI

Changes in a Network-LSA

- A network-LSA is generated by a DR and flooded within an area.
- A network-LSA describes all routers that have a full relationship with the DR on a link.



OSPFv2 Network LSA



OSPFv3 Network LSA

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Options: 3 bytes. This field is a set of Options fields of the link-LSAs of all routers on a link, that is, a set of capabilities that the routers support.
- Attached Router:
  - Four bytes for each router. This field describes the router IDs of all routers that have a full relationship with the DR on a link.
  - An OSPFv3 network-LSA is generated by a DR and flooded within an area. The link state ID is the interface ID of the DR, and the mask field existing in OSPFv2 is deleted. Therefore, an OSPFv3 network-LSA does not contain prefix information but only describes topology connections.
  - In addition, the Options field describes the capability set of all routers on a link. Therefore, the capability of the DR does not affect the LSA transmission of other routers.

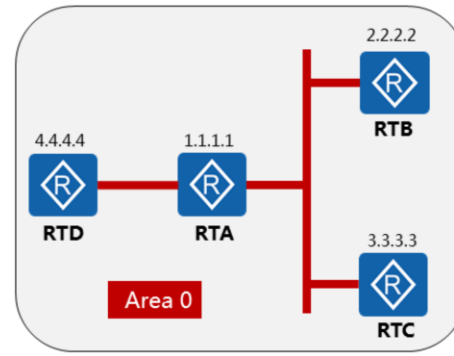
Network-LSA Example

```
<RTA>display ospfv3 lsdb network

OSPFv3 Router with ID (1.1.1.1) (Process 1)

Network-LSA (Area 0.0.0.0)

LS Age: 95
LS Type: Network-LSA
Link State ID:  24.0.2.2      DR Interface ID
Originating Router:  2.2.2.2      DR Router ID
LS Seq Number: 0x80000006
Checksum: 0xCB18
Length: 36
Options: 0x000013 (-|R|-|-|E|V6)
  Attached Router: 2.2.2.2
  Attached Router: 1.1.1.1
  Attached Router: 3.3.3.3
```



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

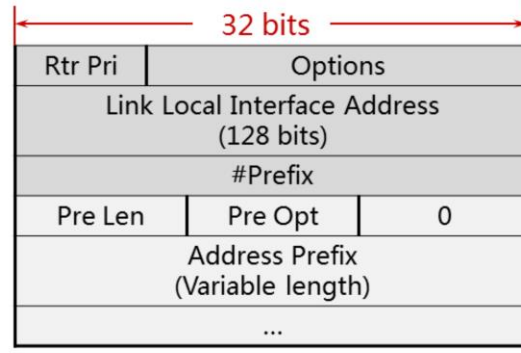
Link-LSA (New in OSPFv3)

- A link-LSA is a new type of LSA added to OSPFv3. A link-LSA is generated by a router for each OSPFv3 interface and flooded on a link. Its functions include:
  - Advertises a local-link address as a next-hop address to other routers on the link.
  - Advertises all IPv6 prefixes on the local-link to other routers on the link.
  - Provides the Options field value for the DR on a broadcast or NBMA network.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Link-LSA Structure

- Rtr Pri: indicates the priority of a router on a link.
- Options: describes the capabilities of a router.
- Link Local Interface Address: indicates the link-local address of an interface, which is used to calculate the next hop of a route.
- #Prefix: indicates the number of prefixes.
- Others: Prefix triple.



Link-LSA

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Link-LSA Example

```
[RTA]display ospfv3 lsdb link 24.0.2.134
Link-LSA (Interface Serial6/0/0)

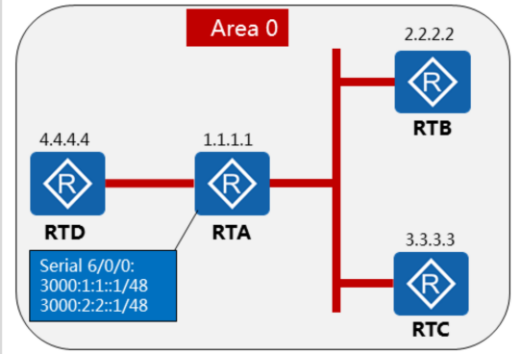
 LS Age: 17
 LS Type: Link-LSA
 Link State ID: 24.0.2.134      Interface index
 Originating Router: 1.1.1.1
 LS Seq Number: 0x80000024
 Checksum: 0x7F10
 Length: 68
 Priority: 1
 Options: 0x000013 (-|R|-|-|E|V6)
 Link-Local Address: FE80::5ECC:5200:1
 Number of Prefixes: 2

  Prefix: 3000:1:1::/48
  Prefix Options: 0 (-|-|-|-)

  Prefix: 3000:2:2::/48
  Prefix Options: 0 (-|-|-|-)
```

Displays the link-LSA of a specified link state ID.

Two prefixes are carried.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

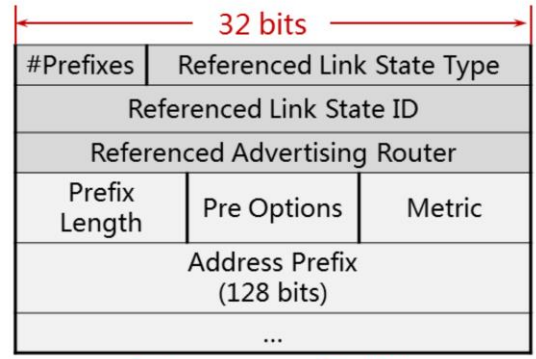
Intra-Area-Prefix-LSA

- Why intra-area-prefix-LSAs are introduced?
  - In OSPFv2, address information is contained in router-LSAs on stub networks and network-LSAs on transit networks. However, in OSPFv3, router-LSAs and network-LSAs do not contain address information. Intra-area-prefix-LSAs are then introduced.
- Intra-area-prefix-LSAs contain the following intra-area IPv6 prefix information:
  - Router prefixes
  - Prefixes of stub networks
  - Prefixes of transit networks
- Multiple intra-area-prefix-LSAs can be originated by a router or on a transit network.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Intra-Area-Prefix-LSA

- In OSPFv2, router-LSA and network-LSA are used to advertise intra-area routes. In OSPFv3, these two types of LSAs do not contain address information. Therefore, intra-area-prefix LSAs are introduced to advertise intra-area routes.
- #Prefixes: indicates the number of prefixes contained in an LSA.
- Referenced Link State Type:
  - =1: prefix of the connected router (including the stub network)
  - =2: prefix of the attached transit network
- Referenced Link State ID:
  - Type 1: 0;
  - Type 2: DR interface ID.
- Referenced Advertising Router:
  - Type 1: router ID of the attached router
  - Type 2: router ID of the DR
- Others: prefix triplet information



Intra-area-prefix-LSA

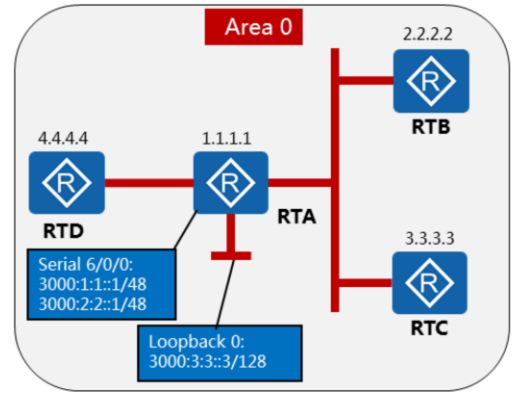
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An OSPFv3 intra-area-prefix-LSA is flooded within an area to advertise intra-area prefix information. According to different LSAs, there are two situations:
  - If router-LSAs are referenced, an intra-area-prefix-LSA is generated by each router to advertise the prefixes of point-to-point links and the prefixes of stub networks.
  - If network-LSAs are referenced, an intra-area-prefix-LSA is generated by a DR to advertise all prefixes on the network corresponding to the link. These prefixes are learned from the link-LSAs generated by all routers on the link. However, the link-local address information in link-LSAs and the prefixes of which the NU or LA bit is 1 are excluded.

Example for Intra-Area-Prefix-LSA Example (Referenced to Router-LSA)

```
<RTA>display ospfv3 lsdb intra-prefix
            Intra-Area-Prefix-LSA (Area 0.0.0.0)
 LS Age: 20
 LS Type: Intra-Area-Prefix-LSA
 Link State ID: 0.0.0.1
 Originating Router: 1.1.1.1
 LS Seq Number: 0x8000003E
 Checksum: 0xA932
 Length: 76
 Number of Prefixes: 3
 Referenced LS Type:  0x2001      A router-LSA is referenced.
 Referenced Link State ID:  0.0.0.0      The first LS ID in a router
LSA
 Referenced Originating Router:  1.1.1.1      Router ID.
  Prefix: 3000:2:2::/48
  Prefix Options: 0 (-|-|-|-)
  Metric: 1562
  Prefix: 3000:1:1::/48
  Prefix Options: 0 (-|-|-|-)
  Metric: 1562
  Prefix: 3000:3:3::3/128
  Prefix Options:  2 (-|-|LA|-)      Local host address
  Metric: 0
```



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

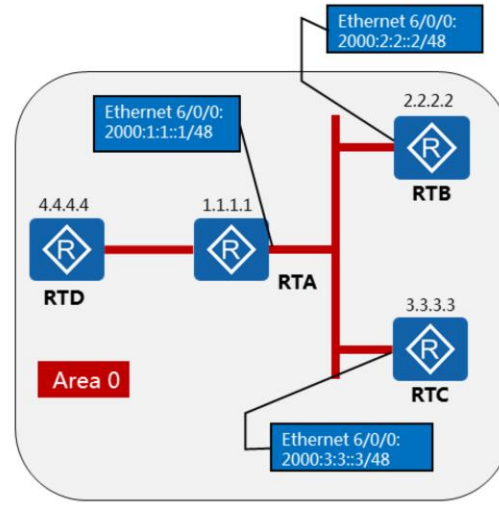
Intra-Area-Prefix-LSA (Referenced to Network-LSA)

```
<RTA>display ospfv3 lsdb intra-prefix
                Intra-Area-Prefix-LSA (Area 0.0.0.0)
 LS Age: 1407
 LS Type: Intra-Area-Prefix-LSA
 Link State ID: 0.0.0.2
 Originating Router: 2.2.2.2
 LS Seq Number: 0x80000028
 Checksum: 0xB1E7
 Length: 68
 Number of Prefixes: 3
 Referenced LS Type:  0x2002      A network-LSA is referenced.
 Referenced Link State ID:  24.0.2.2      DR interface ID
 Referenced Originating Router:  2.2.2.2      DR router ID

   Prefix: 2000:1:1::/48
   Prefix Options: 0 (-|-|-|-)
   Metric: 0

   Prefix: 2000:2:2::/48
   Prefix Options: 0 (-|-|-|-)
   Metric: 0

   Prefix: 2000:3:3::/48
   Prefix Options: 0 (-|-|-|-)
   Metric: 0
```



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

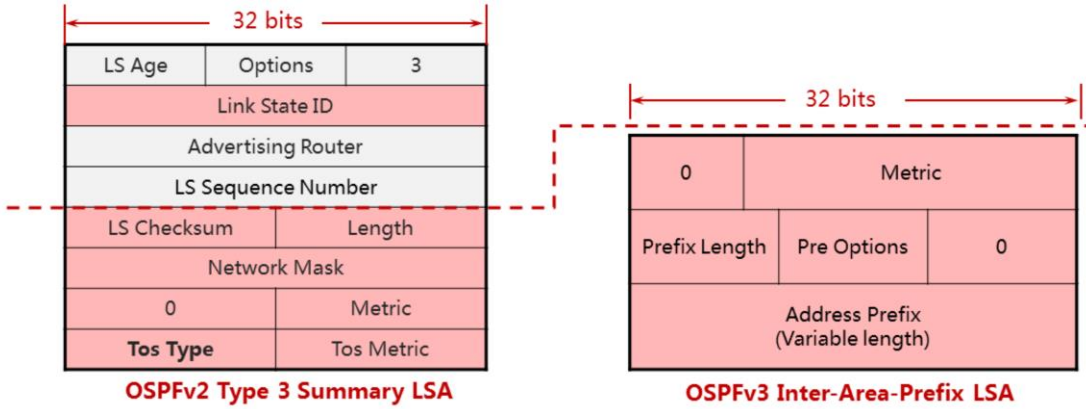
HUAWEI

Changes in Type 3 LSAs - Inter-Area-Prefix-LSA

- In OSPFv2, this type of LSA is called Type 3 summary-LSA. In OSPFv3, this type of LSA is called inter-area-prefix-LSA. It describes the prefix information of other areas.
  - A Type 3 LSA is generated by an ABR and flooded within an area.
  - This type of LSA describes the routing information of other areas in the local AS.
  - Each inter-area-prefix-LSA contains an address prefix.
  - The LSA does not contain link-local addresses.
  - The LSA uses a 32-bit integer as the link state ID to distinguish itself from other LSAs generated by the same router.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Inter-Area-Prefix-LSA Structure



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Metric: 20 bits. It indicates the cost of the route from an ABR to the prefix route.
- Prefix information: indicates prefix triple information.
- In OSPFv2, the Link State ID field in an LSA header indicates a network address. A mask is carried in the LSA.
- In an OSPFv3 inter-area-prefix-LSA, the Link State ID field in the LSA header does not contain prefix information. A link state ID is a 32-bit number used to differentiate LSAs generated by the same router. All prefixes are described using prefix triplets.
- An inter-area-prefix-LSAs is generated by an ABR and flooded within an area. Each inter-area-prefix-LSA contains an address prefix, but does not contain any link-local addresses.

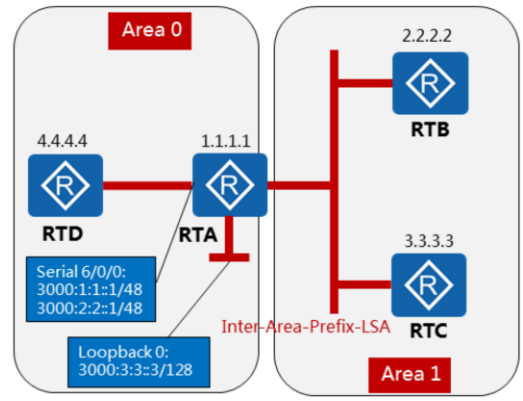
Inter-Area-Prefix-LSA Example

```
<RTC>display ospfv3 lsdb inter-prefix

     Inter-Area-Prefix-LSA (Area 0.0.0.1)

 LS Age: 89
 LS Type: Inter-Area-Prefix-LSA
 Link State ID: 0.0.0.3
 Originating Router: 1.1.1.1
 LS Seq Number: 0x80000002
 Checksum: 0xB5F6
 Length: 36

  Metric: 1562
  Prefix: 3000:1:1::/48
  Prefix Options: 0
```



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

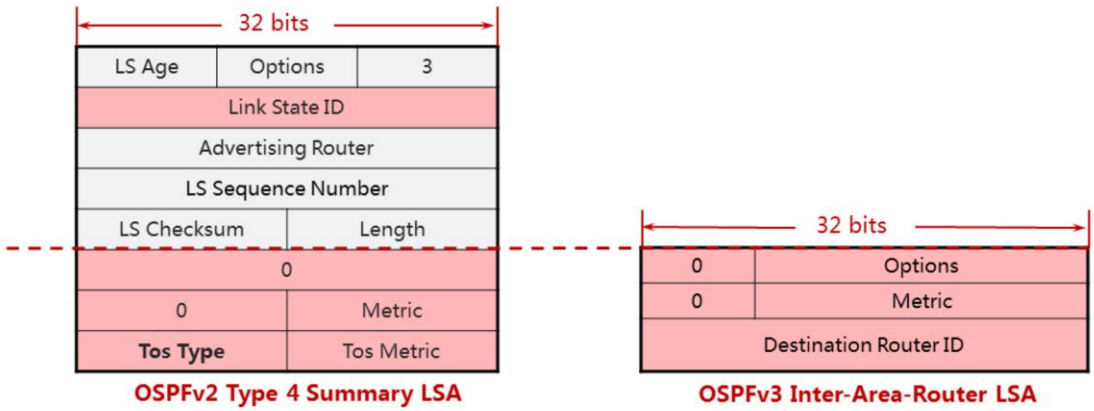
HUAWEI

Changes in Type 4 LSAs - Inter-Area-Router-LSA

- In OSPFv2, this type of LSA is called Type 4 summary-LSA. In OSPFv3, this type of LSA is called inter-area-router-LSA. It describes the information about an ASBR connected to other areas.
  - A Type 4 LSA is generated by an ABR and flooded within an area.
  - This type of LSA describes the ASBR information of other areas in a local AS.
  - Each inter-area-router-LSA contains the information about one ASBR.
  - The Options field in a Type 4 LSA describes the same capabilities as those described in a router LSA generated by the corresponding ASBR.
  - A Type 4 LSA uses a 32-bit integer as the link state ID to distinguish itself from other LSAs generated by the same router.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

OSPFv3 Inter-Area-Router-LSA Structure



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

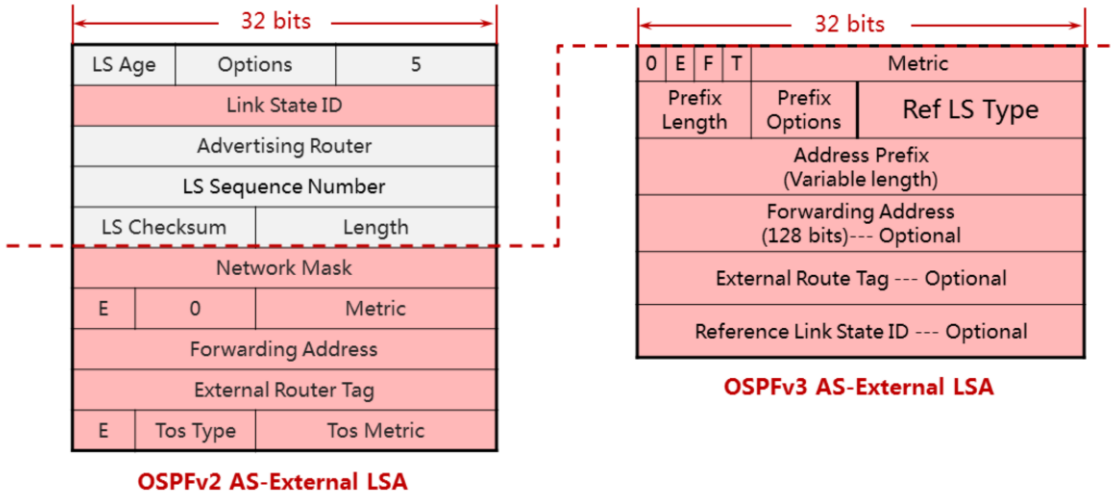
- Options: 3 bytes. It describes the capabilities of a destination ASBR, not the capabilities of the router that generates the LSA.
- Metric: 3 bytes. It indicates the cost of the route from an ABR to a destination ASBR.
- Destination Router ID: 4 bytes. It describes the router ID of a destination ASBR.
- In OSPFv2, the Link State ID field in an LSA header indicates the router ID of a destination ASBR. In an OSPFv3 inter-area-router-LSA, the Link State ID field in an LSA header does not have any specific meaning. It is a 32-bit number used to differentiate LSAs generated by the same router.
- An inter-area-router-LSA is generated by an ABR and flooded within an area. Each inter-area-router-LSA contains the information about a destination ASBR.

AS-External-LSA

- An AS-external-LSA is generated by an ASBR and describes the routing information of external areas.
- It is flooded within an AS.
- A link state ID in this type of LSA does not contain address information but is used to distinguish itself from other AS-external-LSAs.
- An AS-external-LSA does not contain link-local addresses.
- Optional fields:
  - Forwarding Address
  - External Router Tag
  - Referenced Link State ID: reserved field.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

AS-External-LSA Structure



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- E: metric type of an external route. The value 1 indicates the metric of a Type 2 external route. This metric does not increase during route transmission. The value 0 indicates the metric of a Type 1 external route. This metric increases during route transmission.
- F: The value 1 indicates that the Forwarding Address field is carried.
- T: The value 1 indicates that the External Route Tag field is carried.
- Prefix information: indicates prefix triple information.
- Ref LS Type: 2 bytes. If the value is not 0, the Referenced Link State ID field is carried.
- Forwarding Address: 16 bytes. This field is optional. It indicates a 128-bit IPv6 address. This field is carried if the F bit is set to 1. It indicates the address to which a packet needs to be forwarded before the packet reaches its destination. This address can be used if the advertising router is not the optimal next hop.
- External Route Tag: 4 bytes. This field is optional. It can be used for communication between ASBRs. Typically, routes that are imported by OSPF AS boundary routers can be filtered by setting this flag bit.

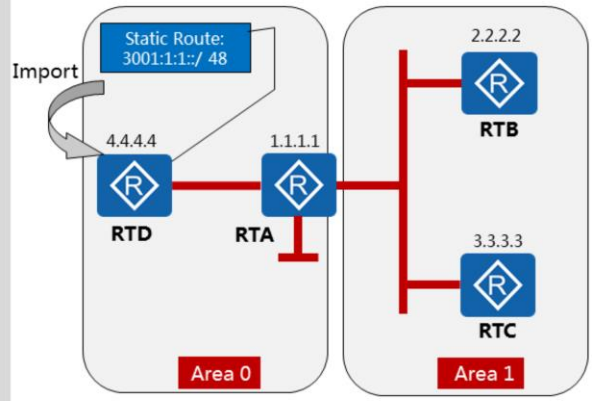
- Referenced Link State ID: 4 bytes. This field is carried if the Ref LS Type field is not set to 0. If this field exists, additional information concerning the advertised external route can be found in another LSA. The referenced information is as follows:
    - The LS type is the value of the Referenced LS Type field in an AS-external-LSA.
    - A link state ID is the value of the Referenced Link State ID field in an AS-external-LSA.
    - An advertising router is the value of the Advertising Router field in an AS-external-LSA.
    - This capability is used for extension and is not used currently.
- In OSPFv2, the Link State ID field in an LSA header indicates a network address. A mask is carried in the LSA.
- In an OSPFv3 AS-external-LSA, the Link State ID field in the LSA header does not contain the prefix information. It is a 32-bit number used to differentiate LSAs generated by the same router. All prefixes are described using prefix triplets.
- An AS-external-LSA is generated by an ASBR and flooded within an AS. Each AS-external-LSA contains an address prefix but does not contain link-local address information.

AS-External-LSA Example

```
<RTB>display ospfv3 lsdb external
OSPFv3 Router with ID (2.2.2.2) (Process 1)

        AS-External-LSA

 LS Age: 1027
 LS Type: AS-External-LSA
 Link State ID: 0.0.0.1
 Originating Router: 4.4.4.4
 LS Seq Number: 0x80000004
 Checksum: 0xD0C7
 Length: 36
  Metric Type: 2 (Larger than any link state path)
   Metric: 1
   Prefix: 3001:1:1::/48      Imported static route
   Prefix Options: 0 (-|-|-|-)
```



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

- OSPFv2 Fundamentals
- OSPFv2 Fault Diagnosis
- OSPFv3 Introduction
- Differences Between OSPFv3 and OSPFv2
- **OSPFv3 Configuration Methods and Common Commands**
- Suggestions on OSPF Exam Preparation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Maintenance and Debugging

Operation	Command
Enable debugging of OSPFv3 events.	debugging ospfv3 event { abr | asbr | vlink | all }
Enable debugging of OSPFv3 interface status.	debugging ospfv3 ifsm [status | event | timer]
Enable debugging of OSPFv3 LSAs.	debugging ospfv3 lsa { all | flooding | generate | install | maxage | refresh | verbose }
Enable debugging of OSPFv3 neighbor status.	debugging ospfv3 nfsm [status | event | timer]
Enable debugging of OSPFv3 packets.	debugging ospfv3 packet all [verbose] debugging ospfv3 packet { hello | dd | request | update | ack } * [verbose] debugging ospfv3 packet verbose
Enable debugging of OSPFv3 route calculation.	debugging ospfv3 route [ase | install | spf | ia]

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Displaying LSDB Information (1)

```
<RTA>display ospfv3 lsdb

        OSPFv3 Router with ID (1.1.1.1) (Process 1)
          Link-LSA (Interface Ethernet6/0/0)

Link State ID   Origin Router    Age  Seq#       CkSum  Prefix
24.0.2.2        1.1.1.1          0720 0x80000032 0x7314     1
24.0.2.2        2.2.2.2          0693 0x80000032 0xb1ce     1
24.0.2.2        3.3.3.3          0661 0x80000032 0xef89     1

          Link-LSA (Interface Serial6/0/0)

Link State ID   Origin Router    Age  Seq#       CkSum  Prefix
24.0.2.134      1.1.1.1           0004 0x80000061 0xa61d     1
24.0.2.6        4.4.4.4          0408 0x8000005f 0x5b11     1

          Router-LSA (Area 0.0.0.0)

Link State ID   Origin Router    Age  Seq#       CkSum   Link
0.0.0.0         1.1.1.1          0714 0x80000067 0x685c     1
0.0.0.0         4.4.4.4          0818 0x80000061 0x28a1     1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Displaying LSDB Information (2)

```
            Inter-Area-Prefix-LSA (Area 0.0.0.0)
Link State ID   Origin Router   Age  Seq#       CkSum
0.0.0.2         1.1.1.1         0645 0x80000032 0xf9b2
0.0.0.3         1.1.1.1         0645 0x80000032 0x0e9b
0.0.0.4         1.1.1.1         0645 0x80000032 0x2284

            Intra-Area-Prefix-LSA (Area 0.0.0.0)
Link State ID   Origin Router   Age  Seq#       CkSum  Prefix  Reference
0.0.0.1         1.1.1.1         0004 0x8000008b 0x0917     2  Router-LSA
0.0.0.1         4.4.4.4         0813 0x80000085 0x2cb2     1  Router-LSA

            Router-LSA (Area 0.0.0.1)
Link State ID   Origin Router   Age  Seq#       CkSum   Link
0.0.0.0         1.1.1.1         0645 0x80000036 0xc6d9     1
0.0.0.0         2.2.2.2         0633 0x80000036 0xa5f7     1
0.0.0.0         3.3.3.3         0620 0x80000034 0x8b10     1

            Network-LSA (Area 0.0.0.1)
Link State ID   Origin Router   Age  Seq#       CkSum
24.0.2.2        3.3.3.3          0620 0x80000033 0x337f
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Displaying LSDB Information (3)

```
 Inter-Area-Prefix-LSA (Area 0.0.0.1)

Link State ID   Origin Router   Age  Seq#       CkSum
0.0.0.3         1.1.1.1         0709 0x80000033 0x5328
0.0.0.5         1.1.1.1         0709 0x80000033 0x51e8

          Inter-Area-Router-LSA (Area 0.0.0.1)

Link State ID   Origin Router   Age  Seq#       CkSum
0.0.0.1         1.1.1.1         0815 0x80000031 0x8f32

          Intra-Area-Prefix-LSA (Area 0.0.0.1)

Link State ID   Origin Router   Age  Seq#       CkSum  Prefix  Reference
0.0.0.1         3.3.3.3         0620 0x80000032 0xbbcc     3  Network-LSA

          AS-External-LSA

Link State ID   Origin Router   Age  Seq#       CkSum
0.0.0.1         4.4.4.4         0818 0x80000031 0x76f4 E2
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Displaying Interface Information

```
<RTA>display ospfv3 interface ethernet6/0/0

Ethernet6/0/0 is up, line protocol is up
 Interface ID 0x18000202      Interface index
 Interface MTU 1500
 IPv6 Prefixes
   FE80::200:5EFF:FECC:5200 (Link-Local Address)      Interface link-local address
   2000:1:1::1/48
 OSPFv3 Process (1), Area 0.0.0.1, Instance ID 0
   Router ID 1.1.1.1, Network Type BROADCAST, Cost: 1
   Transmit Delay is 1 sec, State DROther, Priority 1
   Designated Router (ID) 3.3.3.3
     Interface Address FE80::200:5EFF:FECC:5202
   Backup Designated Router (ID) 2.2.2.2
     Interface Address FE80::200:5EFF:FECC:5201
   Timer interval configured, Hello 10, Dead 40, Wait 40, Retransmit 5
     Hello due in 00:00:01
   Neighbor Count is 2, Adjacent neighbor count is 2
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Displaying Neighbor Information

```
<RTA>display ospfv3 peer

OSPFv3 Process (1)

OSPFv3 Area (0.0.0.0)
Neighbor ID     Pri   State           Dead Time   Interface  Instance ID
4.4.4.4           1   Full/ -         00:00:39    S6/0/0           0

OSPFv3 Area (0.0.0.1)
Neighbor ID     Pri   State           Dead Time   Interface  Instance ID
2.2.2.2           1   Full/Backup     00:00:31    Eth6/0/0         0
3.3.3.3           1   Full/DR         00:00:37    Eth6/0/0         0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Displaying Topology Information

```
<RTA>display ospfv3 topology

OSPFv3 Process (1)
OSPFv3 Area (0.0.0.0) topology
Type  ID(If-Index)      Bits      Metric    Next-Hop       Interface
Rtr   1.1.1.1           B         --
Rtr   4.4.4.4           E         1562      4.4.4.4        Serial6/0/0

OSPFv3 Area (0.0.0.1) topology
Type  ID(If-Index)      Bits      Metric    Next-Hop       Interface
Rtr   1.1.1.1           B         --
Rtr   2.2.2.2                     1         2.2.2.2        Ethernet6/0/0
Rtr   3.3.3.3                     1         3.3.3.3        Ethernet6/0/0
Net   3.3.3.3(402653698)          1         0.0.0.0        Ethernet6/0/0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Displaying OSPFv3 Routing Table Information

```
<RTC>display ospfv3 routing

OSPFv3 Process (1)
 Destination                          Metric
   Next-hop
   2000:1:1::/48                        1
        via FE80::200:5EFF:FECC:5200, Ethernet6/0/0
   2000:2:2::/48                        1
        via FE80::200:5EFF:FECC:5201, Ethernet6/0/0
   2000:3:3::/48                        1
        directly-connected, Ethernet6/0/0
 IA 3000:1:1::/48                     1563
        via FE80::200:5EFF:FECC:5200, Ethernet6/0/0
 IA 3000:3:3::3/128                     1
        via FE80::200:5EFF:FECC:5200, Ethernet6/0/0
 E2 3001:1:1::/48                       1
        via FE80::200:5EFF:FECC:5200, Ethernet6/0/0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

- OSPFv2 Fundamentals
- OSPFv2 Fault Diagnosis
- OSPFv3 Introduction
- Differences Between OSPFv3 and OSPFv2
- OSPFv3 Configuration Methods and Common Commands
- **Suggestions on OSPF Exam Preparation**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Suggestions on OSPF Exam Preparation

- Be familiar with OSPF commands.
- Be familiar with OSPF policy configuration.
- Read HedEx documents.
  - Learn cases in the HedEx documents.
- Master display and debugging commands.
- Understand the case scenarios designed in the course.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Recommendations

- Huawei learning website
  - http://support.huawei.com/learning/en/newindex.html
- Huawei support case library
  - http://support.huawei.com/enterprise/servicecenter?lang=en

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— IS-IS

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- Intermediate System-to-Intermediate System (IS-IS) is an important internal gateway protocol (IGP) on IP networks. Thanks to the type-length-value (TLV) structure it uses, IS-IS has high scalability and strong vitality. As an IGP with high scalability, the application of IS-IS is not limited to traditional IP networks. IS-IS can also run in data centers, on IPv6 networks, and in various other IP network scenarios. This chapter focuses on the advanced features of IS-IS and IPv6 IS-IS features and configurations. In addition, it describes the deployment and implementation of dual IS-IS on the live network.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to:
  - Master the advanced IS-IS features.
  - Have a good command of IPv6 IS-IS features and configurations.
  - Understand the application of IS-IS.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

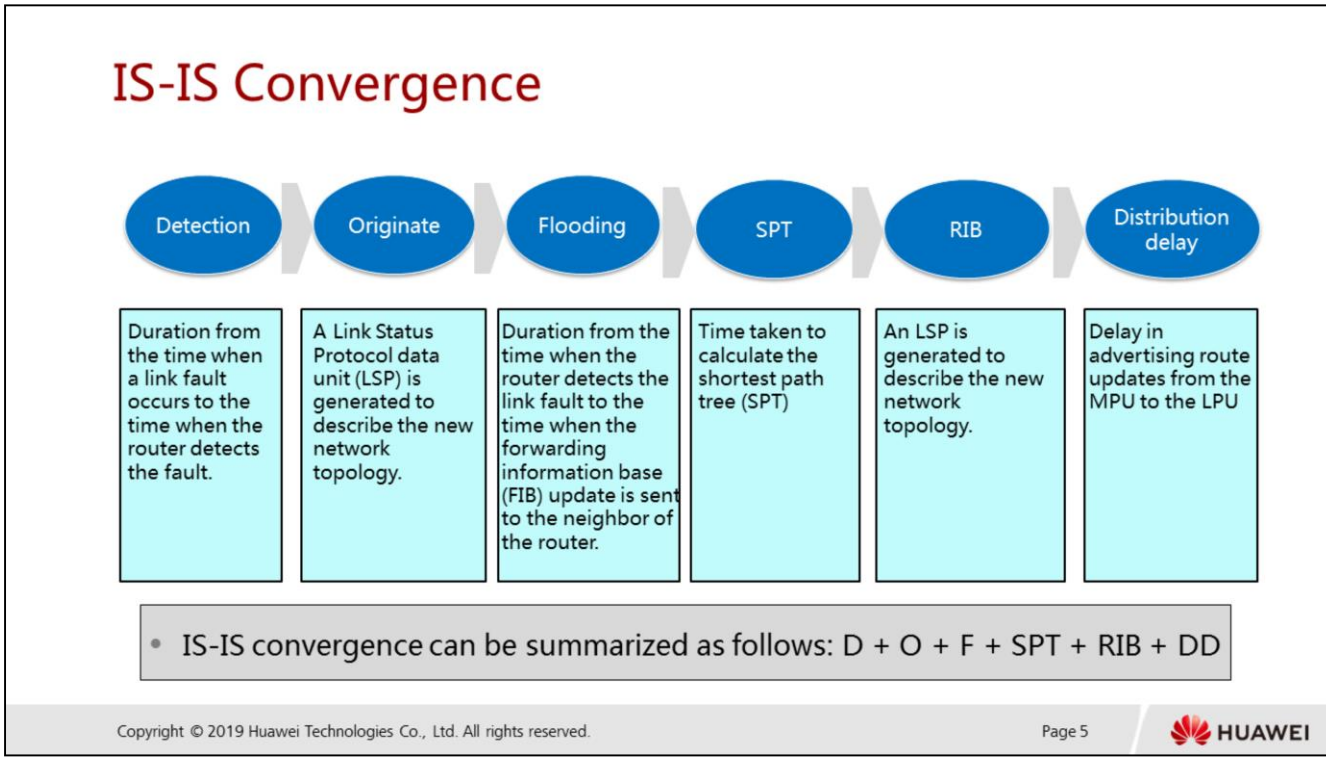


Contents

1. **Advanced IS-IS Features**
   - IS-IS fast convergence
   - IS-IS administrative tag
   - IS-IS LSP fragmentation
   - IS-IS route filtering
2. Basic Principles and Configurations of IPv6 IS-IS
3. Application Cases of Dual Stack IS-IS

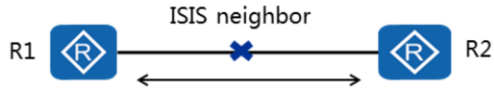
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



- IS-IS convergence is performed as follows: D + O + F + SPT + RIB + DD
  - D phase: period from the time when a fault occurs on a link to the time when a router senses the link fault
  - O phase: time taken to generate an LSP to describe the new network topology
  - F phase: period from the time when the router senses the link fault to the time when the router advertises FIB updates to neighbors
  - SPT phase: time taken to calculate the shortest path tree
  - RIB phase: time taken by the main CPU to update RIB and FIB entries
  - DD phase: delay in advertising route updates from the system control board to the service board
- The RIB and DD phases are related to hardware of the router, such as the CPU of the MPU, CPU of the LPU, memory, and network processor. The two phases have slight impact on the convergence time. Therefore, the following section describes only the first four phases.

Link Fault Detection



IS-IS link fault detection in normal cases

- In normal cases, IS-IS link status detection depends only on IS-IS, that is, the IS-IS Hello (IIH) packets of IS-IS. IIH packets support second-level fault detection, which is inapplicable to networks demanding fast convergence.
- To reduce the detection time, use a small interval for transmitting IIH packets.

IS-IS link fault detection with the assistance from other protocols

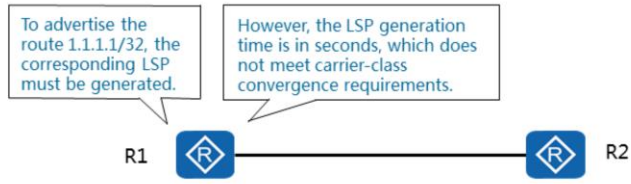
- SDH/SONET (POS) and Ethernet optical interface support millisecond-level link status detection, providing a higher network convergence speed. However, this detection solution is dependent on the interface type.
- BFD is usually applied to IP networks to detect link status. BFD supports millisecond-level detection and is independent from the interface type.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The current fault detection mechanisms include:
  - Hardware detection: For example, the Synchronous Digital Hierarchy (SDH) alarms are used to detect faults on links. The hardware detection can fast detect a fault; however, not all media can provide the hardware detection mechanism.
  - Slow Hello mechanism: It usually refers to the Hello mechanism offered by a routing protocol. The Hello mechanism can detect a fault in seconds. In high-speed data transmission, for example, at gigabit rates, the detection time longer than one second causes the loss of a large amount of data. In delay-sensitive services such as voice services, the delay longer than one second is also unacceptable. In addition, this mechanism relies on routing protocols. Generally, IS-IS uses IIH packets to detect neighbors and detect faults. The detection takes seconds to complete.
  - Other detection mechanisms: Specific detection mechanisms may be provided by different protocols or device vendors. If a network has devices from multiple vendors, these detection mechanisms, such as BFD, are difficult to implement.

LSP Generation Interval



LSP Generation Interval

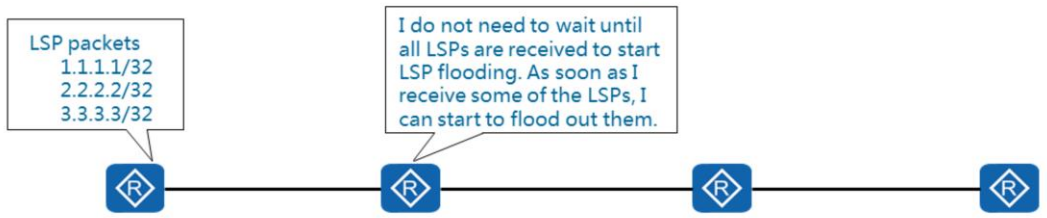
- For IS-IS, if the local routing information changes, the router should generate a new LSP to notify this change. On the other hand, if the delay in generating an LSP is too long, the router cannot advertise the changes to neighboring nodes immediately. This slows down network convergence.
- Therefore, an intelligent timer is used in LSP generation to speed up network convergence and maintain system performance. In this way, the delay can be automatically adjusted based on the change frequency of routing information.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- When the LSP generation intelligent timer expires, the system generates a new LSP based on the current topology. In the original implementation mechanism, a timer with a fixed interval is used, which, however, cannot meet the requirements on fast convergence and low CPU usage.
- Therefore, an intelligent timer is used in LSP generation to speed up network convergence and maintain system performance. In this way, the delay can be automatically adjusted based on the change frequency of routing information. The timer can quickly respond to emergencies (for example, an interface goes Up or Down) and as a result speed up network convergence. In addition, when the network topology changes frequently, the interval of the intelligent timer automatically prolongs, which reduces CPU resource consumption.

Speeding Up LSP Flooding



LSP Fast Flooding

- According to RFC protocols, when a router receives an LSP that is generated later than the LSP stored in the local LSDB, the router updates the local LSDB and periodically floods the updated LSP using a timer. Therefore, the LSDB synchronization is slow.
- LSP fast flooding improves the preceding method. When a router configured with this feature receives one or more new LSPs, it floods out the LSPs with a number smaller than the specified number before route calculation to speed up LSDB synchronization. This mechanism speeds up network convergence to a large extent.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- You can set the number of LSPs for each flooding. The number is valid to all interfaces. If the number of LSPs to be sent is larger than the specified number, *lsp-count* LSPs are sent. If a timer is configured and the configured timer does not expire before the route calculation, the LSPs are flooded immediately when being received; otherwise, the LSPs are sent when the timer expires.
- If the level is not specified in the command, the function is enabled for both Level-1 and Level-2.

Speeding Up SPF Calculation



SPF Intelligent Timer

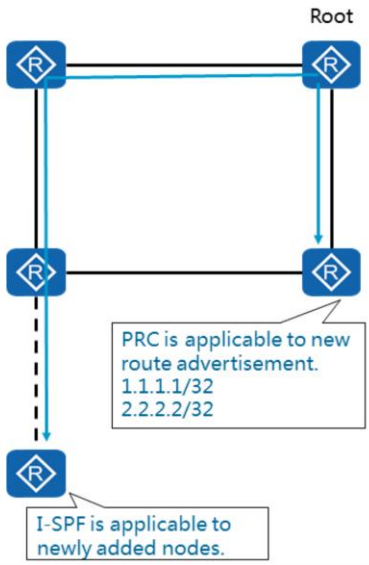
- Although the route calculation algorithm is improved, the long interval for triggering the route calculation also affects the convergence speed. Millisecond-level timer can shorten the interval. However, if frequent network changes occur, excessive CPU resources would be consumed.
- The SPF intelligent timer addresses both of these problems.
- If the topology of the network changes frequently, the intelligent timer increases the interval for the calculation to avoid too much CPU consumption.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In general, an IS-IS network running normally is stable. The probability of the occurrence of many network changes is very small, and the IS-IS router does not frequently calculate routes. The period for triggering the route calculation is very short (millisecond level). If the topology of the network changes very often, the intelligent timer increases the interval for the calculation times to avoid too much CPU consumption.

Adjusting the SPF Calculation Method



- Incremental SPF (I-SPF) calculates only changed routes at a time rather than all routes.
- During its first calculation, I-SPF calculates routes on all nodes. After that, only routes on the affected nodes are calculated each time.

- Similar to I-SPF, only changed parts are considered in PRC. However, PRC does not calculate the shortest path but updates leaf routes based on the SPT calculated by I-SPF.

- Modifying a broadcast network to a point-to-point network reduces the number of pseudo nodes on the network and shortens the SPF calculation time.

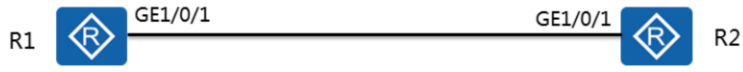
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Based on ISO-10589, the Dijkstra algorithm is used to calculate routes. When a node changes on the network, this algorithm is used to recalculate all routes. The calculation takes a long time and consumes too many CPU resources, affecting the convergence speed.
- I-SPF improves this algorithm. Except for the first time, only changed nodes instead of all nodes are involved in calculation. The SPT generated at last is the same as that generated by the Dijkstra algorithm. This decreases the CPU usage and speeds up network convergence.
- In route calculation, a route represents a leaf, and a router represents a node. If the SPT changes after I-SPF calculation, PRC processes all the leaves only on the changed node. If the SPT remains unchanged, PRC processes only the changed leaves.
- For example, if IS-IS is enabled on an interface of a node, the SPT calculated by I-SPF remains unchanged. PRC updates only the routes of this interface, consuming less CPU resources.
- PRC working with I-SPF further improves the convergence performance of the network. It has now replaced the original SPF algorithm.
- By default, Huawei routers use I-SPF and PRC for router calculation, which does not require configuration using commands.

IS-IS Fast Convergence Configuration

- R1 and R2 are interconnected, and they communicate with each other through IS-IS.
- It is required to improve the convergence speed of the two routers through BFD.




```
isis 1
is-level level-2
timer lsp-generation 1 1 120 level-2
flash-flood level-2
network-entity 10.0000.0000.0001.00
timer spf 1 20 100
#
bfd
bfd atob bind peer-ip 10.1.1.2 interface GigabitEthernet1/0/1
 discriminator local 1
discriminator remote 2
Commit
#
interface GigabitEthernet1/0/1
isis bfd static
```

```
isis 1
is-level level-2
timer lsp-generation 1 1 120 level-2
flash-flood level-2
network-entity 10.0000.0000.0002.00
timer spf 1 20 100
#
bfd
bfd atob bind peer-ip 10.1.1.1 interface GigabitEthernet1/0/1
 discriminator local 2
discriminator remote 1
Commit
#
interface GigabitEthernet1/0/1
isis bfd static
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Contents

1. **Advanced IS-IS Features**
   - IS-IS fast convergence
   - IS-IS administrative tag
   - IS-IS LSP fragmentation
   - IS-IS route filtering
2. Basic Principles and Configurations of IPv6 IS-IS
3. Application Cases of Dual Stack IS-IS

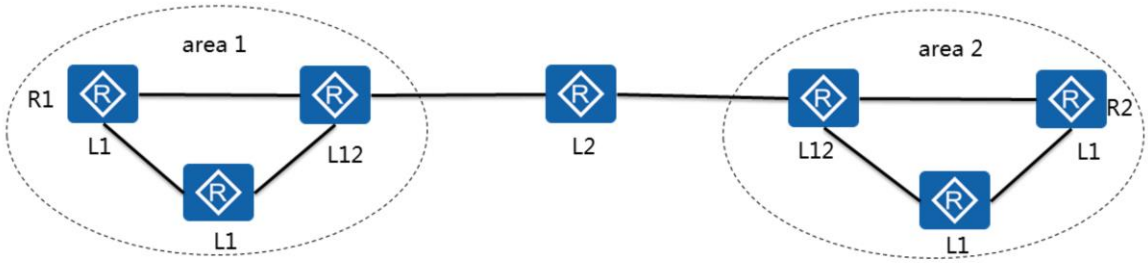
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Basic Features

- Administrative tags control the advertisement of IP prefixes in an IS-IS routing domain to simplify route management. You can use administrative tags to control the import of routes of different levels and different areas and control IS-IS multi-instances running on the same router.
- The value of an administrative tag is associated with certain attributes. When the cost-style is wide, wide-compatible or compatible and IS-IS advertises an IP address prefix with these attributes, IS-IS adds the administrative tag to the TLV in the prefix. The tag is flooded along with the prefix throughout the routing domain.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

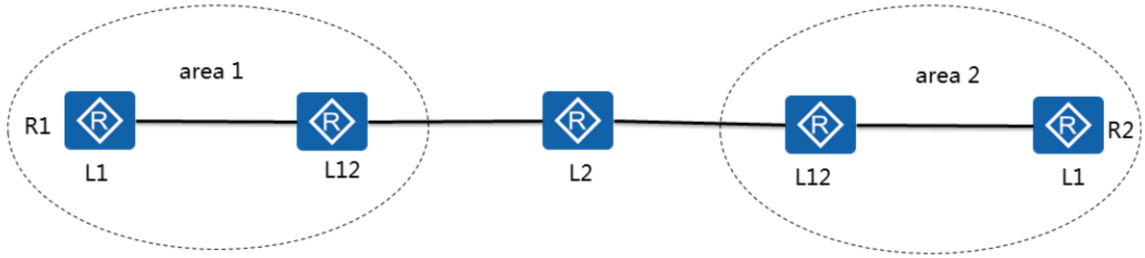
Working Principle (1)



- R1 needs to communicate with R2 in area 2. However, to ensure information security, routers in area 2 cannot receive packets sent by R1. First, you can configure the same administrative tag for IS-IS interfaces on R1 and R2. Then, the Level-1-2 router of area 1 leaks only routes matching the configured administrative tag from Level-2 to Level-1 areas. In this way, among all routers in area 2, R1 only communicates with R2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Working Principle (2)



- After the administrative tag is applied, R1 in area 1 can communicate only with R2 in area 2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



Contents

1. **Advanced IS-IS Features**
   - IS-IS fast convergence
   - IS-IS administrative tag
   - IS-IS LSP fragmentation
   - IS-IS route filtering
2. Basic Principles and Configurations of IPv6 IS-IS
3. Application Cases of Dual Stack IS-IS

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Basic Terms

- Originating system: A router that runs the IS-IS protocol. A single IS-IS process can function as multiple virtual routers to advertise LSPs, and the originating system refers to the IS-IS process.
- Normal system ID: The system ID of the originating system.
- Virtual system: is the system identified by the additional system ID to generate extended LSP fragments. These fragments carry additional system IDs in their LSP IDs.
- Additional system ID: Assigned by network administrators to identify a virtual system. A maximum of 256 extended LSP fragments can be generated for each additional system ID.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Working Principles

- In IS-IS, each system ID identifies a system, which can generate a maximum of 256 LSP fragments. With more additional system IDs (up to 50 virtual systems can be configured), an IS-IS process can generate a maximum of 13,056 LSP fragments.
- After LSP fragment extension is configured, the system prompts you to restart the IS-IS process if information is lost because LSPs overflow. After being restarted, the originating system loads as much routing information to LSPs, adds the overloaded information to the LSPs of the virtual system for transmission, and uses TLV 24 to notify other routers of its relationship with the virtual system.

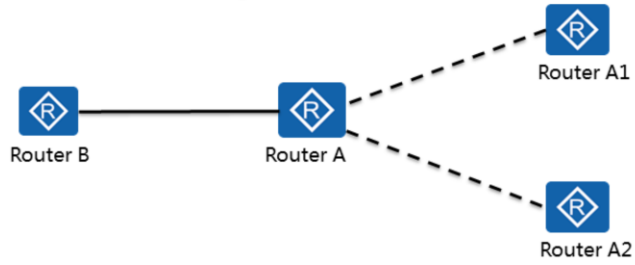
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- When an IS-IS router needs to advertise the LSPs that contain much information, the IS-IS router generates multiple LSP fragments to carry more IS-IS information.
- IS-IS LSP fragments are identified by the LSP Number field in their LSP IDs. This field is of 1 byte. An IS-IS process can generate a maximum of 256 LSP fragments; therefore, only a limited number of routes can be carried. As defined in RFC 3786, virtual system IDs can be configured and virtual LSPs that carry routing information can be generated for IS-IS.

Working Modes

- An IS-IS router can run the LSP fragment extension feature in two modes.



Mode	Application Scenario
Mode 1	Some routers on the network do not support LSP fragment extension.
Mode 2	All the routers on the network support LSP fragment extension.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Mode 1: is used when some routers on the network do not support LSP fragment extension.
- In Mode 1, virtual systems participate in the SPF calculation. The originating system advertises LSPs containing information about links to each virtual system. Similarly, each virtual system advertises LSPs containing information about links to the originating system. Virtual systems look like the physical routers that connect to the originating system. Mode 1 is a transitional mode for the earlier versions that do not support LSP fragment extension. In earlier versions, IS-IS cannot identify the IS Alias ID TLV and processes the received LSP that is advertised by a virtual system as an LSP advertised by an IS-IS process.
- Mode 2: is used when all routers on the network support LSP fragment extension.
- In Mode 2, virtual systems do not participate in the SPF calculation. All the routers on the network know that the LSPs generated by the virtual systems actually belong to the originating system. An IS-IS router working in Mode 2 can identify the IS Alias ID TLV, which is used as a reference for calculating the SPT and routes.
- Note: When the originating system and virtual system send the LSPs with fragment number 0, the LSPs must carry the IS Alias ID TLV to indicate the originating system regardless of the working mode (Mode 1 or Mode 2).



Contents

1. **Advanced IS-IS Features**
   - IS-IS fast convergence
   - IS-IS administrative tag
   - IS-IS LSP fragmentation
   - IS-IS route filtering
2. Basic Principles and Configurations of IPv6 IS-IS
3. Application Cases of Dual Stack IS-IS

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

IS-IS Filters Routes Calculated Based on LSDB (1)

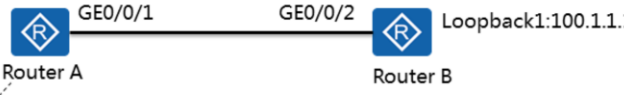
- By configuring IS-IS to filter the routes calculated based on LSDB, you can control the number of IS-IS routes to be added to the IP routing table, and reduce the size of the IP routing table.
- Configure an ACL rule:
- **acl** { **name** *basic-acl-name* { **basic** | [**basic**] **number** *basic-acl-number* } | [**number**] *basic-acl-number* } [**match-order** { **config** | **auto** }]

  rule [*rule-id*] [**name** *rule-name*] { **deny** | **permit** } [**fragment-type** { **fragment** | **non-fragment** | **non-subseq** | **fragment-subseq** | **fragment-spe-first** } | **source** { *source-ip-address* { *source-wildcard* | **0** | *src-netmask* } | **any** } | **time-range** *time-name* | **vpn-instance** *vpn-instance-name*]
- Configure the filter:
- **filter-policy** { *acl-number* | **acl-name** *acl-name* } **import**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Filters Routes Calculated Based on LSDB (2)

- Router A and Router B establish an IS-IS neighbor relationship. Router A can learn a route to loopback1 of Router B through IS-IS.



- Before the filtering:

Route Flags: R - relay, D - download to fib

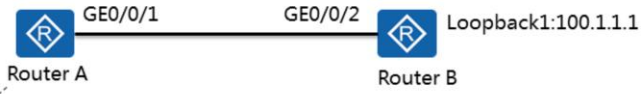
Routing Tables: Public
Destinations : 9 Routes : 9

Destination/Mask	Proto	Pre	Cost	Flags	NextHop	Interface
1.1.1.1/32	Direct	0	0	D	127.0.0.1	LoopBack0
10.1.1.0/30	Direct	0	0	D	10.1.1.1	GigabitEthernet0/0/1
10.1.1.1/32	Direct	0	0	D	127.0.0.1	GigabitEthernet0/0/1
10.1.1.3/32	Direct	0	0	D	127.0.0.1	GigabitEthernet0/0/1
100.1.1.1/32	ISIS-L2	15	10	D	10.1.1.2	GigabitEthernet0/0/1

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Filters Routes Calculated Based on LSDB (3)

- Router A and Router B establish an IS-IS neighbor relationship. A traffic policy is deployed for IS-IS so that the routing table of Router A does not contain a route to loopback1 of Router B, but the route to loopback1 of Router B exists in IS-IS LSDB of Router A.



- Configure a filtering policy.

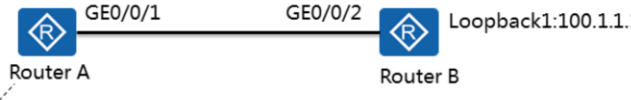
```
#
acl name huawei 2000
  rule 5 deny source 100.1.1.1 0
#
isis 1
is-level level-2
network-entity 01.0000.0000.0001.00
filter-policy acl-name huawei import
#
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

IS-IS Filters Routes Calculated Based on LSDB (4)

- After the filtering policy is configured, the routing table of Router A does not contain a route to loopback1 of Router B, but the route to loopback1 of Router B exists in IS-IS LSDB of Router A.



- View the routing table.

```
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 9        Routes : 9

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

        1.1.1.1/32  Direct  0    0           D   127.0.0.1       LoopBack0
       10.1.1.0/30  Direct  0    0           D   10.1.1.1        GigabitEthernet0/0/1
       10.1.1.1/32  Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
       10.1.1.3/32  Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Note:
- The prefix of the filtered route still exists in the IS-IS LSDB LSP.

IS-IS Filters Imported Routes (1)

- You can configure IS-IS to filter imported external routes to effectively control the number of IS-IS routes on the network.
- Configure an IP prefix list.
- **ip ip-prefix** *ip-prefix-name* [**index** *index-number*] { **permit** | **deny** } *ip-address mask-length* [**greater-equal** *greater-equal-value*] [**less-equal** *less-equal-value*]
- Configure the routing policy:
- **route-policy** *route-policy-name* { **permit** | **deny** } **node** *node*

  if-match acl { *acl-number* | *acl-name* }

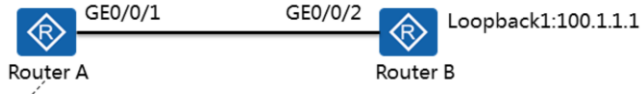
  if-match ip-prefix *ip-prefix-name*

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

IS-IS Filters Imported Routes (2)

- Router A and Router B establish an IS-IS neighbor relationship. Router B can import a direct route to loopback1 to its IS-IS LSDB and advertise the route to Router A.



- Before the filtering:

```
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 9        Routes : 9

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

        1.1.1.1/32  Direct  0    0           D   127.0.0.1       LoopBack0
       10.1.1.0/30  Direct  0    0           D   10.1.1.1        GigabitEthernet0/0/1
       10.1.1.1/32  Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
       10.1.1.3/32  Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
     100.1.1.1/32   ISIS-L2 15   10          D   10.1.1.2        GigabitEthernet0/0/1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Filters Imported Routes (3)

- Router A and Router B establish an IS-IS neighbor relationship. When Router B imports the direct route to loopback1 to its IS-IS LSDB, the filtering policy of imported routes takes effect.



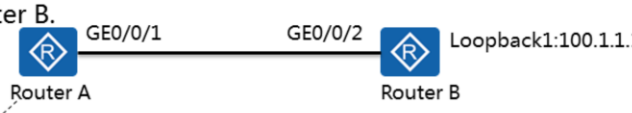
- Configure a filtering policy.

```
#
ip ip-prefix huawei index 10 deny 100.1.1.1 32
#
route-policy huawei deny node 10
if-match ip-prefix huawei
#
isis 1
is-level level-2
network-entity 01.0000.0000.0002.00
import-route direct route-policy huawei
#
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Filters Imported Routes (4)

- After the filtering policy is configured, the direct route to loopback1 of Router B is not imported to the IS-IS LSDB of Router B. Therefore, Router A does not learn the route to loopback1 of Router B.



- View the routing table.

```
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 9        Routes : 9

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

        1.1.1.1/32  Direct  0    0           D   127.0.0.1       LoopBack0
       10.1.1.0/30  Direct  0    0           D   10.1.1.1        GigabitEthernet0/0/1
       10.1.1.1/32  Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
       10.1.1.3/32  Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

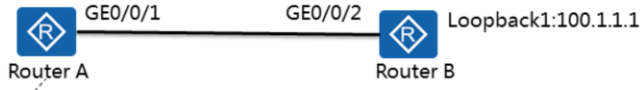
IS-IS Filters External Routes When They Are Advertised (1)

- IS-IS filters imported external routes when they are advertised.
- Configure a filtering policy for external routes that are to be advertised.
- **filter-policy** { *acl-number* | **acl-name** *acl-name* | **ip-prefix** *ip-prefix-name* | **route-policy** *route-policy-name* } **export** [*protocol* [*process-id*]]

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Filters External Routes When They Are Advertised (2)

- Router A and Router B establish an IS-IS neighbor relationship. Router B can import the direct route to loopback1 to its IS-IS LSDB and advertise the route to Router A. Router A learns the route to loopback1 of Router B.



- Before the filtering:

```
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 9        Routes : 9

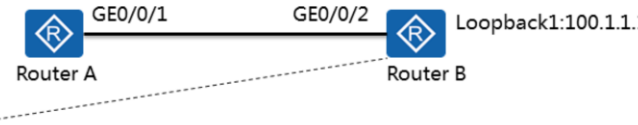
Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

        1.1.1.1/32  Direct  0    0           D   127.0.0.1       LoopBack0
       10.1.1.0/30  Direct  0    0           D   10.1.1.1        GigabitEthernet0/0/1
       10.1.1.1/32  Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
       10.1.1.3/32  Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
     100.1.1.1/32   ISIS-L2 15   10          D   10.1.1.2        GigabitEthernet0/0/1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Filters External Routes When They Are Advertised (3)

- Router A and Router B establish an IS-IS neighbor relationship. The direct route to loopback1 of Router B is imported to the IS-IS LSDB of Router B. When Router B advertises the LSP to Router A, the policy that filters out the route to loopback1 of Router B takes effect.



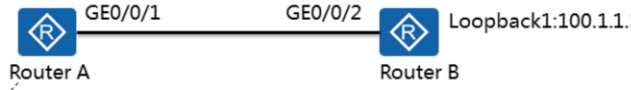
- Configure a filtering policy.

```
#
ip ip-prefix huawei index 10 deny 100.1.1.1 32
#
isis 1
is-level level-2
network-entity 01.0000.0000.0002.00
import-route direct
filter-policy acl-name huawei export
#
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Filters External Routes When They Are Advertised (4)

- After the filtering policy is configured, the IS-IS LSDB of Router A does not contain the route to loopback1 of Router B. As a result, Router A does not learn the route to loopback1 of Router B.



- View the routing table.

```
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 9        Routes : 9

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

        1.1.1.1/32  Direct  0    0           D   127.0.0.1       LoopBack0
      10.1.1.0/30   Direct  0    0           D   10.1.1.1        GigabitEthernet0/0/1
      10.1.1.1/32   Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
      10.1.1.3/32   Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
```

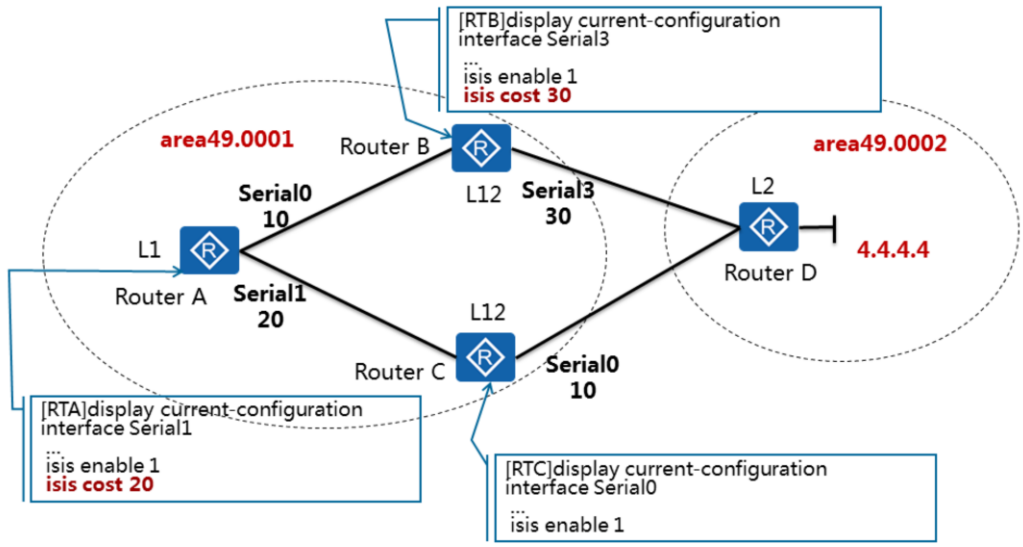
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. **Advanced IS-IS Features**
   - IS-IS fast convergence
   - IS-IS administrative tag
   - IS-IS LSP fragmentation
   - IS-IS route filtering
   - IS-IS route leaking
2. Basic Principles and Configurations of IPv6 IS-IS
3. Application Cases of Dual Stack IS-IS

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Route Leaking (1)

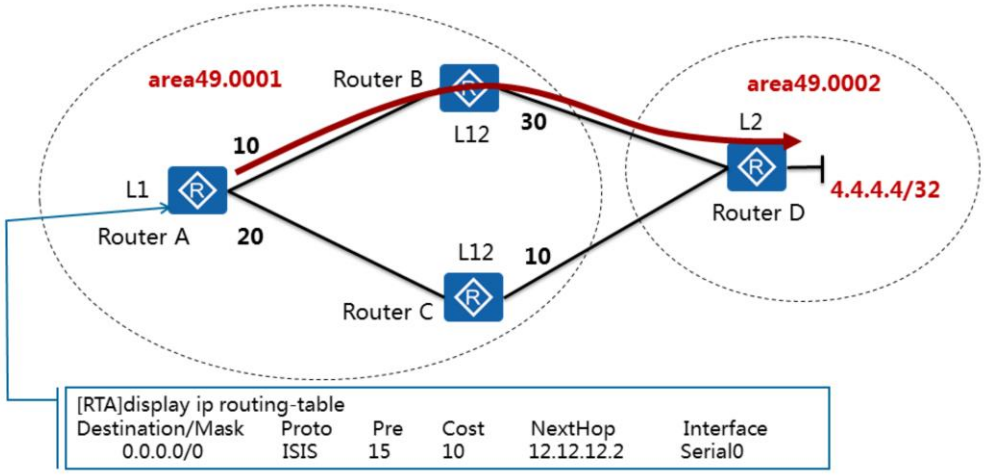


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Introduction:
  - The routing information of a Level-1 area is advertised to a Level-2 area through a Level-1-2 router; thus, Level-1-2 and Level-2 routers know the routing information of the entire IS-IS domain. A Level-2 router, by default, does not inform a Level-1 area of the learned routing information of other Level-1 areas and that of the backbone area. Therefore, Level-1 routers do not know the routing information outside the local area. As a result, the Level-1 routers cannot select the optimal routes to the destination outside the area.
- IS-IS route leaking can solve this problem.

IS-IS Route Leaking (2)



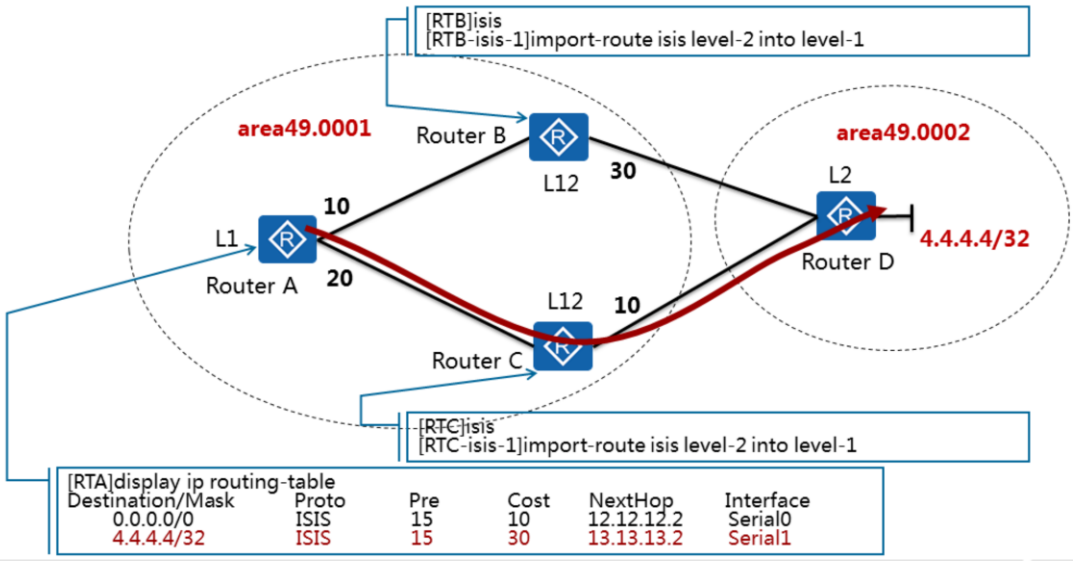
[RTA]display ip routing-table

Destination/Mask	Proto	Pre	Cost	NextHop	Interface
0.0.0.0/0	ISIS	15	10	12.12.12.2	Serial0

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

IS-IS Route Leaking (3)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- When receiving two identical routes, a Level-1 router prefers the route from the local area rather than the one from the Level-2 area, even though the cost of the Level-2 route is smaller.

Contents

1. Advanced IS-IS Features
2. **Basic Principles and Configurations of IPv6 IS-IS**
   - IS-IS Protocol Extension
   - IS-IS ST and MT
   - IPv6 IS-IS Configuration
3. Application Cases of Dual Stack IS-IS

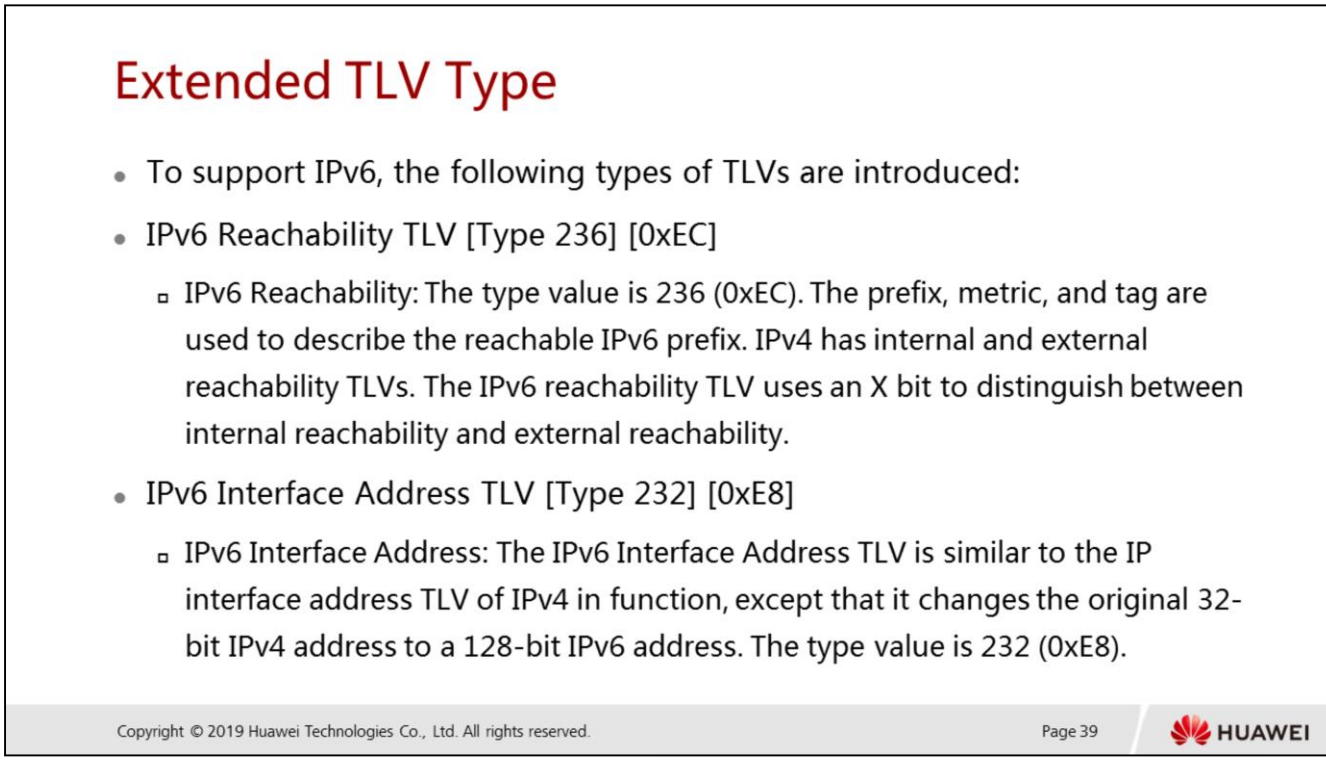
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Overview of IS-IS

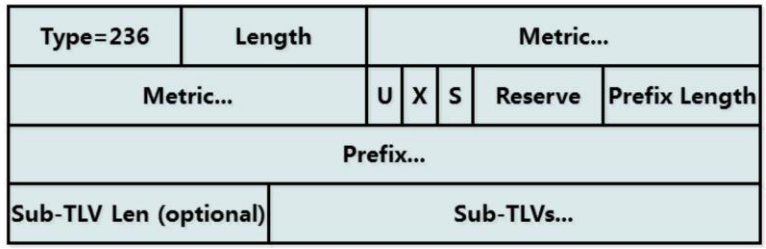
- IS-IS is an IGP protocol.
- IS-IS uses the SPF algorithm to calculate routes.
- IS-IS design is based on TLVs, which allows high scalability.
  - To support new protocols and features, you only need to extend the TLV format by using new TLVs or sub-TLVs.
  - The structure can be easily extended to support IPv6, TE, MT, and other protocols and features.
  - IS-IS does not need to change much to support IPv6. Therefore, it features high inheritance. The development of OSPFv3 is required to support IPv6 in the case of OSPF.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- Extended IS-IS for IPv6 is defined in the draft-ietf-isis-ipv6-05 of IETF. The draft introduces two TLVs and a Network Layer Protocol Identifier (NLPID) for extended IS-IS to support IPv6.
- The two TLVs are as follows:
  - IPv6 reachability: The type value is 236 (0xEC). Prefix, metric, and tag are used to describe the reachable IPv6 prefix. IPv4 has internal and external reachability TLVs. The IPv6 reachability TLV uses an X bit to distinguish between internal reachability and external reachability.
- IPv6 Interface Address: The IPv6 Interface Address TLV is similar to the IP interface address TLV of IPv4 in function, except that it changes the original 32-bit IPv4 address to a 128-bit IPv6 address. The type value is 232 (0xE8).

IPv6 Reachability TLV



- U (up/down bit): indicates whether the prefix is advertised from a high level (used for loop prevention).
- X (external original bit): indicates whether the prefix is imported from other routing protocols.
- S (subtlv present bit): sub-TLV flag (optional)

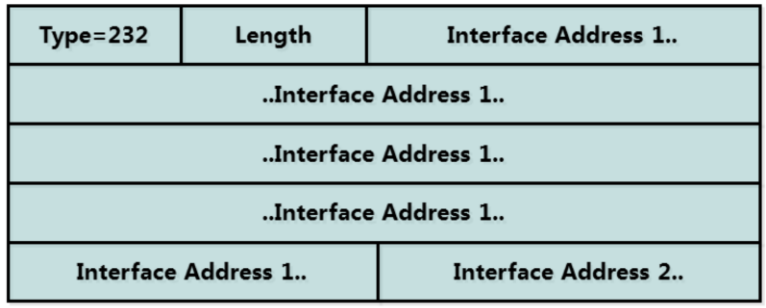
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- This data structure may be repeated multiple times (when there are multiple route prefixes).
- The Metric field has been redefined, and MAX_PATH_METRIC (1023) is changed to MAX_V6_PATH_METRIC (0xFE000000). If the Metric field value of a prefix is greater than MAX_V6_PATH_METRIC, it is not used to construct a routing table but is used for special purposes.
- TLV128: IP internal reachability information; TLV130: IP external reachability information. In TLV236, the X bit is used to distinguish between external reachability and internal reachability.

IPv6 Interface Address TLV (1)

- IPv6 Interface Address TLV [type 232] [0xE8]
  - TLV 232 is similar to TLV 132.
  - The address length of each interface is changed to 128 bits.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Note: In Hello packets, the interface address TLV contains only the link-local address of the interface that sends the Hello packets. For an LSP, the interface address TLV contains only the non-link-local IPv6 address of the IS.

IPv6 Interface Address TLV (2)

- The structure of this TLV is directly mapped from TLV132. Since TLV132 supports a maximum of 64 IP addresses (32 bits), for TLV232, a maximum of 16 IPv6 addresses (128 bits) are supported.
- The field content varies in different PDUs. In a Hello PDU, the Interface Address TLV can only contain the link-local address of the interface that sends the Hello packet. For an LSP, the Interface Address TLV can contain only the non-link-local IPv6 address of the IS.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

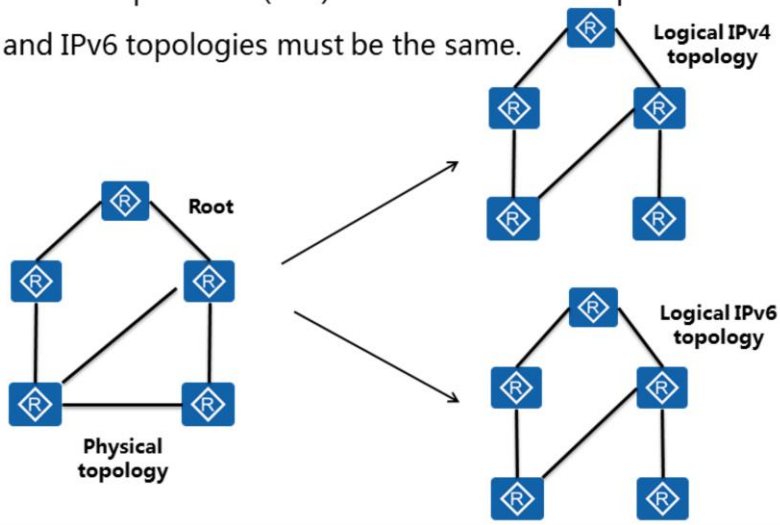
1. Advanced IS-IS Features
2. **Basic Principles and Configurations of IPv6 IS-IS**
   - IS-IS Protocol Extension
   - IS-IS ST and MT
   - IPv6 IS-IS Configuration
3. Application Cases of Dual Stack IS-IS

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

IS-IS Single Topology (ST)

- The same shortest path tree (SPT) is constructed for all protocols using IS-IS.
- The IPv4 and IPv6 topologies must be the same.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

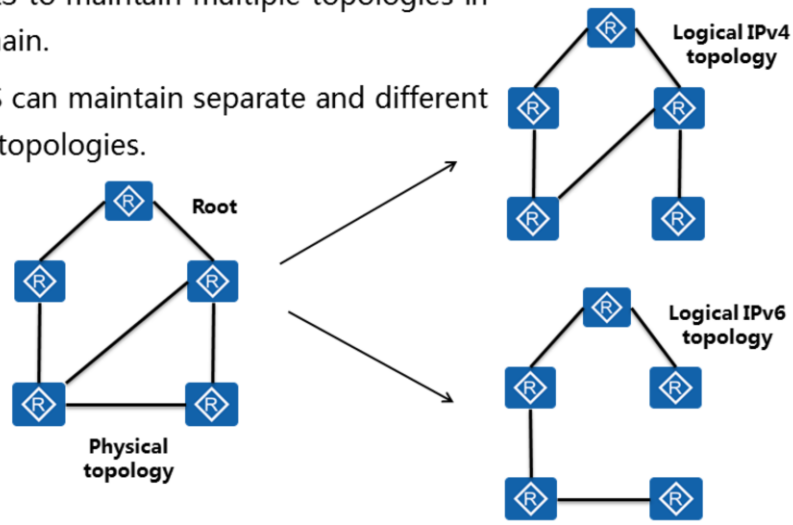
Disadvantages of ST

- AS-IS
  - As the requirement for network maintainability is growing, carriers want to have a maintenance network with an independent topology, that is, an in-band maintenance network.
  - IS-IS maintains the same SPT for all protocols, so IPv4 and IPv6 topologies must be the same.
- Disadvantages
  - ST is not applicable to a network with multiple topologies.
  - To maintain the same topology, all interfaces must run IS-IS IPv4 and IS-IS IPv6 simultaneously, leading to inflexible deployment.
  - IPv6 areas cannot be connected through an IPv4 area because IPv6 traffic will be discarded in the IPv4 area.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Multi-topology (MT)

- MT allows IS-IS to maintain multiple topologies in a routing domain.
- With MT, IS-IS can maintain separate and different IPv4 and IPv6 topologies.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MT: Enabling Topology Separation

- IS-IS involves the following processes:
  - Establishment of a neighbor relationship
  - Advertisement of reachable routes (Prefix Reachable) and reachable routers (IS Reachable)
  - SPF calculation and route calculation
- To implement isolation of multiple topologies, the MT parameter must be carried in all the preceding processes. (SPF calculation and route calculation must be completed on a router.)
- To fulfill such a requirement, draft-ietf-isis-wg-multi-topology-11 defines four new TLVs to implement IS-IS MT interaction in a general environment.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Four TLVs are added:
  - TLV 229 – Multi-Topology Identifier
  - TLV 222 – Multi-Topologies Intermediate System
  - TLV 235 – Multi-Topologies Reachable IPv4 Prefixes
  - TLV 237 – Multi-Topologies Reachable IPv6 Prefixes
- Reserved MT ID Values

Contents

1. Advanced IS-IS Features
2. **Basic Principles and Configurations of IPv6 IS-IS**
   - IS-IS Protocol Extension
   - IS-IS ST and MT
   - IPv6 IS-IS Configuration
3. Application Cases of Dual Stack IS-IS

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

IPv6 IS-IS Configuration

- Enable IS-IS MT.

```
[Huawei-isis-1]ipv6  enable topology ipv6
```

- Enable IS-IS on an interface.

```
[Huawei-GigabitEthernetX/Y/Z]isis  ipv6 enable
```

- Check the IS-IS neighbor relationship.

```
[Huawei]dis isis peer
```

- Check the IS-IS routing information.

```
[Huawei]dis isis route
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

1. Advanced IS-IS Features
2. Basic Principles and Configurations of IPv6 IS-IS
3. **Application Cases of Dual Stack IS-IS**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Planning (1)

- Network entity title (NET) planning
- An NET is in the format of AA.BBBB.CCCC.DDDD.SSSS.SSSS.SSSS.00.
  - Area ID = AA.BBBB.CCCC.DDDD
    - AA: represents an AFI. The value is 49.
    - BBBB: represents a country code. For example, the code for China is 0086.
    - CCCC: represents a province code. For example, the code for Liaoning province is 0003.
    - DDDD: represents the code of a site where devices are located. For example, the code for the Dalian site in Liaoning province is 3821.
  - System ID = SSSS.SSSS.SSSS
    - A system ID is derived from a loopback 0 address.
  - NSEL is fixed at 00.
- Layered and region-based planning
  - Level-2 mode (preferred): All IS-IS routers are located in the backbone.
  - Level-1 mode: All IS-IS routers are located in the same area.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Planning (2)

- Cost planning
- Flexible cost planning for IS-IS
  - The cost can be designed based on the bandwidth of the link.
  - The cost design determines the direction of network traffic.
  - Before the cost design, be clear about directions of network traffic in different end-to-end scenarios.
- IS-IS supports three methods for setting the link cost value:
  - Setting a link cost value in the interface view
  - Setting a cost value for all interfaces in an IS-IS instance in the system view
  - Automatic calculation of link cost values based on the link bandwidth in the system view

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IS-IS Planning (3)

- Reliability planning
- IS-IS fast convergence
  - BFD for IS-IS: rapidly detects link faults.
  - I-SPF: speeds up route convergence (supported by default).
  - Partial Route Calculation (PRC): speeds up route convergence (supported by default).
  - LSP Fast Flooding: speeds up LSDB synchronization.
  - Intelligent timer: speeds up route convergence and improves network stability.

Timer Parameter	Reference Value
Hello interval	Default value
Dead interval	Default value
BFD For ISIS	Default value
Fast Flooding	Default value
Timer spf { max | init | increment }	1 50 100
Timer lsp-generation { max | init | increment }	2 50 100

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- It is recommended that all IS-IS fast convergence features be deployed.

IS-IS Planning (4)

- Security planning
  - IS-IS supports three authentication modes: interface authentication, area authentication, and domain authentication
  - The three types of authentication modes support both simple password authentication and MD5 authentication. MD5 offers higher security and is therefore recommended.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- It is recommended that all IS-IS fast convergence features be deployed.

IS-IS Planning (5)

- Live network configuration

```
isis 1
 is-level level-2
 cost-style wide
 circuit-cost 100000 level-2
 network-entity 86.4680.0551.0611.3313.7003.00
 is-name city-equipment room-device name-
network type.device type
 preference 155
 log-peer-change
 maximum load-balancing 16
 timer lsp-max-age 65500
 timer lsp-refresh 32768
 #
 ipv6 enable topology ipv6
 ipv6 preference 155
 ipv6 maximum load-balancing 16
 ipv6 circuit-cost 100000
 #
```

```
interface GigabitEthernet1/0/0
 description uT:city name-equipment room
name-network element name-serial
number.network.device model:port number
 undo shutdown
 ipv6 enable
 ip address X.X.X.X X.X.X.X
 ipv6 address XXXX:X:XXXX:XXXX::X/127
 ipv6 address auto link-local
 isis enable 551
 isis ipv6 enable 551
 isis circuit-type p2p
 isis circuit-level level-2
 isis ipv6 cost 3000 level-2
 isis cost 3000 level-2
 isis small-hello
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- It is recommended that all IS-IS fast convergence features be deployed.

Summary

- Advanced IS-IS Features
- Basic Principles and Configurations of IPv6 IS-IS
- Application Cases of Dual Stack IS-IS

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— BGP and BGP4+ Basics

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- This chapter describes BGP and BGP4+ basics. Border Gateway Protocol (BGP) is a dynamic routing protocol used to transmit IPv4 and IPv6 routes between Autonomous Systems (ASs), and controls the route advertisement and selects the optimal route between ASs based on routing policies. To implement the MPLS VPN service, Multiprotocol BGP (MP-BGP) needs to be configured inside an AS and between ASs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- On completion of this course, you will be able to:
  - Understand BGP fundamentals.
  - Understand BGP4+ fundamentals.
  - Master BGP basic configuration commands.
  - Understand BGP troubleshooting roadmaps.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

1. **BGP Fundamentals**
2. BGP4+ Fundamentals
3. Basic Configuration of BGP and BGP4+
4. BGP Fault Diagnosis

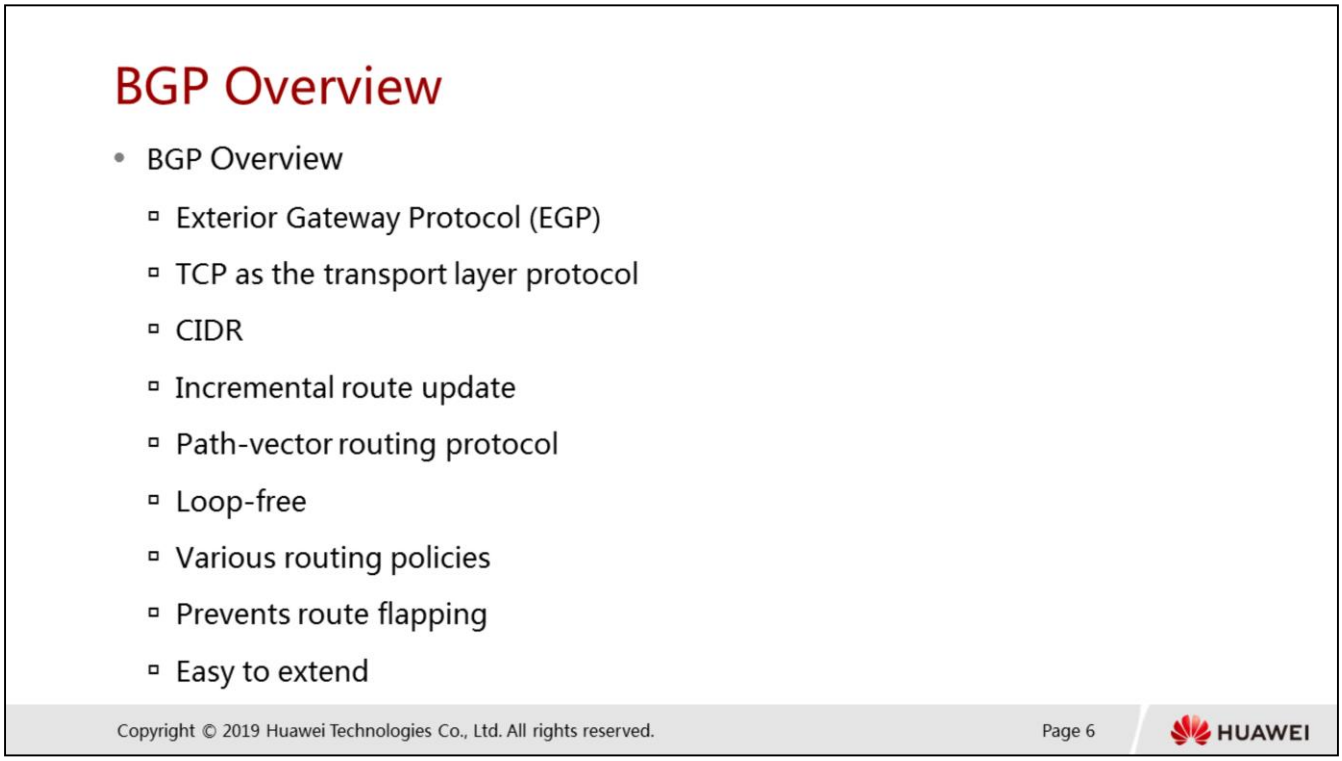
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

1. BGP Fundamentals
   - BGP Overview
   - BGP Concepts
   - BGP Implementation
   - BGP Attributes
   - BGP Route Calculation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- BGP is a dynamic routing protocol used between autonomous systems (ASs). BGP-1 (defined in RFC 1105), BGP-2 (defined in RFC 1163), and BGP-3 (defined in RFC 1267) are three earlier-released versions of BGP. BGP exchanges reachable inter-AS routes, establishes inter-AS paths, avoids routing loops, and applies routing policies between ASs. The version currently used is BGP-4 which is defined in RFC 4271.
- As an exterior routing protocol on the Internet, BGP is widely used among Internet Service Providers (ISPs).

- BGP Overview
  - Different from the Interior Gateway Protocol (IGP) such as Open Shortest Path First (OSPF) and Routing Information Protocol (RIP), BGP is an Exterior Gateway Protocol (EGP), which controls the route advertisement and selects the optimal route between ASs rather than discover network topologies.
  - BGP uses Transport Control Protocol (TCP) with the listening port number being 179 as the transport layer protocol. This enhances BGP reliability and does not need any additional mechanism to guarantee connection controllability.

- BGP selects inter-AS routes, which requires high stability. TCP has high reliability and is used to enhance BGP stability.
  - BGP peers must be logically connected and establish TCP connections. The destination port number is 179 and the local port number is a random value.
  - During route updates, BGP transmits only updated routes, greatly reducing bandwidth consumption. Therefore, BGP applies to the Internet where many routes need to be transmitted.
- BGP is designed to prevent loops.
  - Inter-AS: BGP routes carry information about the ASs along the path. The routes that carry the local AS number are discarded, thereby preventing inter-AS loops.
  - Intra-AS: BGP does not advertise the routes learned in an AS to BGP peers in the AS, thus avoiding intra-AS loops.
- BGP uses various routing policies to filter and select routes flexibly.
- BGP provides a mechanism for preventing route flapping, improving Internet network stability.
- BGP is easy to extend and adapts to the network development. It is an extensible protocol thanks to the Type-Length-Value (TLV) format of attributes.

BGP Implementation: Messages

- Open message

  Negotiate BGP parameters
- Update message

  Exchange routes between BGP peers
- Keepalive message

  Maintain peer relationships
- Notification message

  Used to notify errors
- Route-refresh message

  Request the peer to retransmit routes if routing policies are changed



Package

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- BGP runs by sending five types of BGP messages: Open, Update, Notification, Keepalive, and Route-refresh.
  - Open message: is the first message that is sent after a TCP connection is set up, and is used to set up BGP peer relationships. After a peer receives an Open message and peer negotiation succeeds, the peer sends a Keepalive message to confirm and maintain the peer relationship. Then, peers can exchange Update, Notification, Keepalive, and Route-refresh messages.
  - Update message: is used to exchange routes between BGP peers. Update messages can be used to advertise reachable routes with the same attributes or withdraw multiple unreachable routes.
    - An Update message can advertise multiple reachable routes with the same route attributes. These routes can share a group of route attributes. Route attributes contained in an Update message are applicable to all destination addresses (expressed by IP prefixes) contained in the Network Layer Reachability Information (NLRI) field of the Update message.
    - An Update message can be used to withdraw multiple unreachable routes. Each route is identified by its destination address (using the IP prefix), which identifies the routes previously advertised between BGP speakers.
    - An Update message can be used only to withdraw routes. In this case, it does not need to carry the route attributes or NLRI. In addition, an Update message can be used only to advertise reachable routes. In this case, it does not need to carry information about the withdrawn routes.

    - Keepalive message: is sent periodically to the peer to maintain the peer relationship.
    - Notification message: is sent to its peer when BGP detects an error. The BGP connection is then torn down immediately.
    - Route-refresh message: is used to notify the peer of the capability to refresh routes. If route-refresh is enabled on all BGP peers and the import policy of the local router is changed, the local router sends a Route-refresh message to peers or peer groups. After receiving the message, the peers or peer groups resend routing information to the local BGP router. In this manner, BGP routing tables are dynamically refreshed and new routing policies are applied without tearing down BGP connections.

- BGP message application:
    - BGP uses TCP to establish connections. The local listening port number is 179. Similar to the establishment of a TCP connection, a BGP connection also requires a series of sessions and handshakes. By means of handshake negotiation, TCP advertises its parameters, for example, port number. Handshake negotiation parameters used in BGP are BGP version, BGP connection hold time, local router ID, and authorization information. The information is carried in Open messages.
    - After establishing a connection, BGP sends an Update message to the peer end when a route is to be sent. When advertising a route, the Update message carries the route attributes of the route to help the BGP peer to select the optimal route. When a local BGP route changes, an Update message is sent to notify the BGP peer of the change.
    - After routing information is exchanged for a period of time, neither the local BGP router nor the BGP peer has any new route to advertise, and the BGP connection becomes stable. In this case, Keepalive messages are periodically sent to check the BGP connection validity. If the local BGP router does not receive any Keepalive message from the peer, the local BGP router considers the BGP connection as down, tears down the BGP connection, and deletes all the BGP routes learnt from the peer.
    - On detecting an error, for example, the peer BGP version is not supported locally or the local BGP router receives an invalid Update message, the local BGP router send a Notification message to the BGP peer. When the local BGP router exits a BGP connection, a Notification message is also sent.

- BGP message header
  - Marker: The value length is 16 octets. The field is set to all 1s.
  - Length: This 2-octets unsigned integer indicates the total length of a BGP message (including the header).
  - Type: This 1-octet unsigned integer indicates the type of a BGP message:
    - Open
    - Update
    - Keepalive
    - Notification
    - Route-refresh

- Open message format:
  - Version: BGP version number. For BGPv4, the value is 4.
  - My Autonomous System: Local AS number. The EBGP connection and IBGP connection have different AS numbers on BGP peers.
  - Hold Time: Holdtime that the BGP peers need to negotiate and keep it consistent when establishing the peer relationship. If the value of Holdtime of the peers is not the same, BGP selects the smaller value. If a router does not receive any Keepalive or Update message from its peer within this time, the BGP is considered as disconnected. If the value of Holdtime is 0, Keepalive messages are not sent.
  - BGP Identifier: Router ID of a BGP router. The field is in the form of the IP address and identifies a BGP router.
  - Opt Parm Len (Optional Parameters Length): Length of an optional parameter. If the value is 0, no optional parameter is available.
  - Optional Parameters: Optional parameter used for BGP authentication or multi-protocol extension (Multiprotocol Extensions). Each parameter is a triplet (parameter type, parameter length, and parameter value).

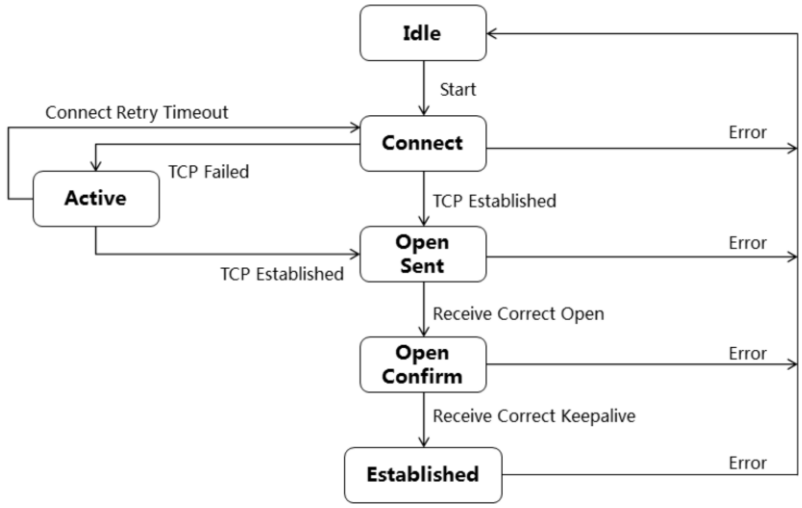
- Update message format:
  - Withdrawn Routes Length: This 2-octets unsigned integer indicates the total length of the Withdrawn Routes field. A value of 0 indicates that no routes are being withdrawn from service, and that the Withdrawn Routes field is not present in this Update message.

    - Withdrawn Routes: This is a variable-length field that contains a list of IP address prefixes for the routes that are being withdrawn from service. Each IP address prefix is encoded in the form <length, prefix>. For example, <19, 198.18.160.0> represents a network 198.18.160.0/255.255.224.0.
    - Path Attribute Length: This 2-octets unsigned integer indicates the total length of the Path Attribute field. A value of 0 indicates there is no data in the Path Attribute field, and that the Path Attribute field is not present in this Update message.
    - Network Layer Reachability Information: This variable-length field contains a list of IP address prefixes. Each IP address prefix is encoded in the form <length, prefix>, same as the Withdrawn Routes field.

- Keepalive message format
    - A Keepalive message consists of only a BGP message header.
    - The default interval for sending Keepalive messages is 60 seconds, and the default value for the hold time of a BGP session is 180 seconds. Upon the reception of the Keepalive message by the BGP peer, the hold time for the BGP session is reinitialized to 180 seconds. If the hold time timer expires, the peer is considered Down.

- Notification message format
    - Errorcode: This 1-octet unsigned integer indicates the error type. Each type of error is represented with a unique error code, and each error code may have one or more error subcodes associated with it. If no appropriate error subcode is defined, a zero value is used for the Error Subcode field.
    - Errsubcode: Error subcode.

BGP Implementation: State Machine



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The BGP Finite State Machine (FSM) has six states: Idle, Connect, Active, OpenSent, OpenConfirm, and Established.
  - Initially, BGP is in the Idle state. In Idle state, a BGP device refuses all incoming BGP connections. The BGP device initiates a TCP connection with its BGP peer and changes its state to Connect only after receiving a Start event from the system.
    - The Start event occurs when an operator configures a BGP process or resets an existing BGP process or when the router software resets a BGP process.
    - If an error occurs at any state of the FSM, for example, the BGP device receives a Notification packet or TCP connection termination notification, the BGP device changes its state to Idle.
  - In Connect state, the BGP device starts the Connect Retry timer (the default interval is 32 seconds and waits to establish a TCP connection.
    - In this state, the BGP device initiates TCP connection requests.
    - If the TCP connection is established, the BGP device sends an Open message to the peer and changes to the OpenSent state.
    - If the TCP connection fails to be established, the BGP device moves to the Active state.
    - If the BGP device does not receive a response from the peer before the Connect Retry timer expires, the BGP device attempts to establish a TCP connection with another peer and stays in Connect state.

        - In response to any other event (initiated by either the system or operator), the BGP device changes its state to Idle.
    - In Active state, the BGP device keeps trying to establish a TCP connection with the peer.
        - In this state, the BGP device waits for the peer to initiate a TCP connection.
        - If the TCP connection is established, the BGP device sends an Open message to the peer, closes the Connect Retry timer, and changes to the OpenSent state.
        - If the TCP connection fails to be established, the BGP device stays in the Active state.
        - If the BGP device does not receive a response from the peer before the Connect Retry timer expires, the BGP device returns to the Connect state.
    - In OpenSent state, the BGP device waits for an Open message from the peer and then checks the validity of the received Open message, including the AS number, version, and authentication password.
        - If the received Open message is valid, the BGP device sends a Keepalive message and changes to the OpenConfirm state.
        - If the received Open message is invalid, the BGP device sends a Notification message to the peer and returns to the Idle state.
    - In OpenConfirm state, the BGP device waits for a Keepalive or Notification message from the peer. If the BGP device receives a Keepalive message, it changes to the Established state. If it receives a Notification message, it returns to the Idle state.
    - In Established state, the BGP device exchanges Update, Keepalive, Route-refresh, and Notification messages with the peer.
        - If the BGP device receives a valid Update or Keepalive message, it considers that the peer is working properly and maintains the BGP connection with the peer.
        - If the BGP device receives an invalid Update or Keepalive message, it sends a Notification message to the peer and returns to the Idle state.
        - If the BGP device receives a Route-refresh message, it does not change its status.
        - If the BGP device receives a Notification message, it returns to the Idle state.
        - If the BGP device receives a TCP connection termination notification, it tears down the TCP connection with the peer and returns to the Idle state.

BGP Implementation: Databases

- IP routing table (IP-RIB)

  Global routing information base, which stores all IP routing information

- BGP routing table (Loc-RIB)

  Stores the local routing information that the local BGP speaker has selected

- BGP peer database

  List of BGP peers

- Adj-RIB-In

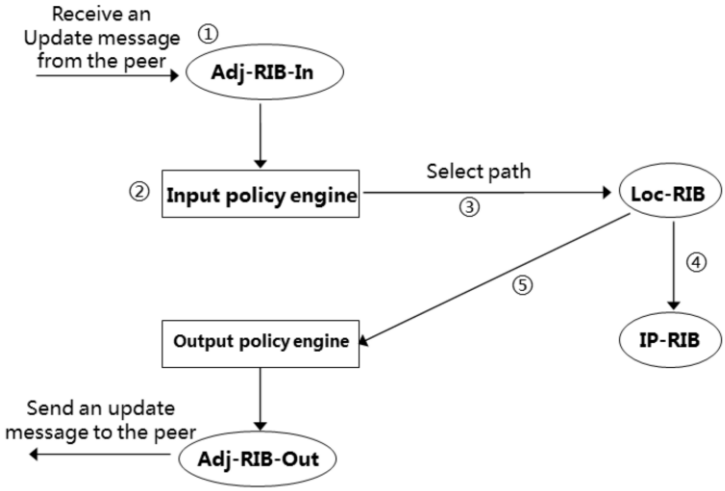
  Stores unprocessed routing information advertised by the peer to the local BGP speaker

- Adj-RIB-Out

  Stores the routing information advertised by the BGP speaker to a specified peer

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

BGP Implementation: BGP Route Information Processing



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- BGP route information processing:
  - When receiving an update message from a peer, the BGP router stores the update message in Adj-RIB-In, and specifies the peer from which it learns the route. After the received update messages are filtered by the input policy engine, the BGP router determines the optimal path for each IP address prefix based on the path selection algorithm.
  - The optimal paths are stored in Loc-RIB, and installed in the local IP-RIB.
  - In addition to the optimal path received from the peer, Loc-RIB also stores the prefixes of the BGP routes that are injected by the local router (also called locally originated routes) and selected as the optimal paths. The routes stored in Loc-RIB must be processed by the output policy engine before being advertised to other peers. Only the routes that are successfully processed by the output policy engine can be installed in Adj-RIB-Out.

BGP Implementation: Interaction Between BGP and IGPs

- Interaction principles are as follows:
  - Advertise the routes learned from IBGP peers only to EBGP peers.
  - Advertise the routes obtained from EBGP peers to all EBGP and IBGP peers.
  - Advertise only the optimal BGP routes to its peers.
  - Send only the updated BGP routes.
  - Synchronize routes between IBGP and IGP.

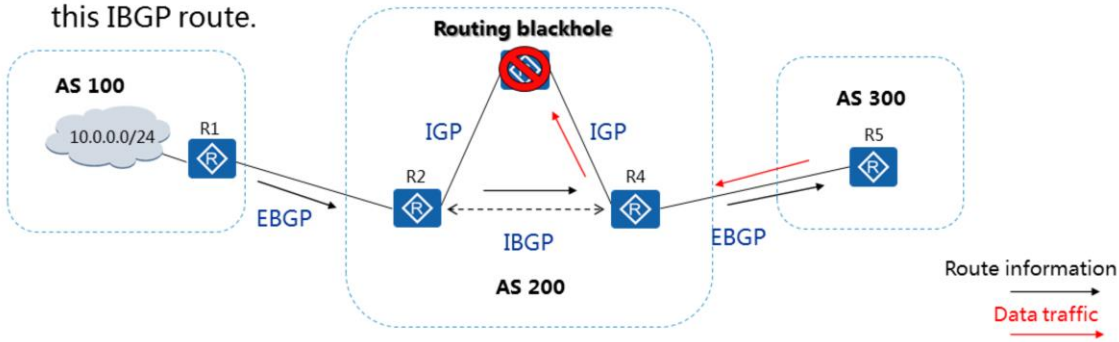
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A BGP device adds optimal routes to the BGP routing table to generate BGP routes.
  - A BGP device advertises the BGP routes received from its IBGP peers only to its EBGP peers.
  - A BGP device advertises the BGP routes received from its EBGP peers to its EBGP peers and IBGP peers.
  - A BGP device advertises the optimal route to its peers when there are multiple valid routes to the same destination.
  - A BGP device sends only updated BGP routes when BGP routes change.
  - A BGP device advertises the routes learned from its IBGP peers to its EBGP peers only when the same routes exist in the IGP.

BGP Implementation: Synchronization Between IBGP and IGP

- Synchronization
  - Before an IBGP route is added to a routing table and advertised to an EBGP peer, the BGP router first checks the IGP routing table. The IBGP route can be added to the routing table and advertised to EBGP peers only when IGP obtains this IBGP route.



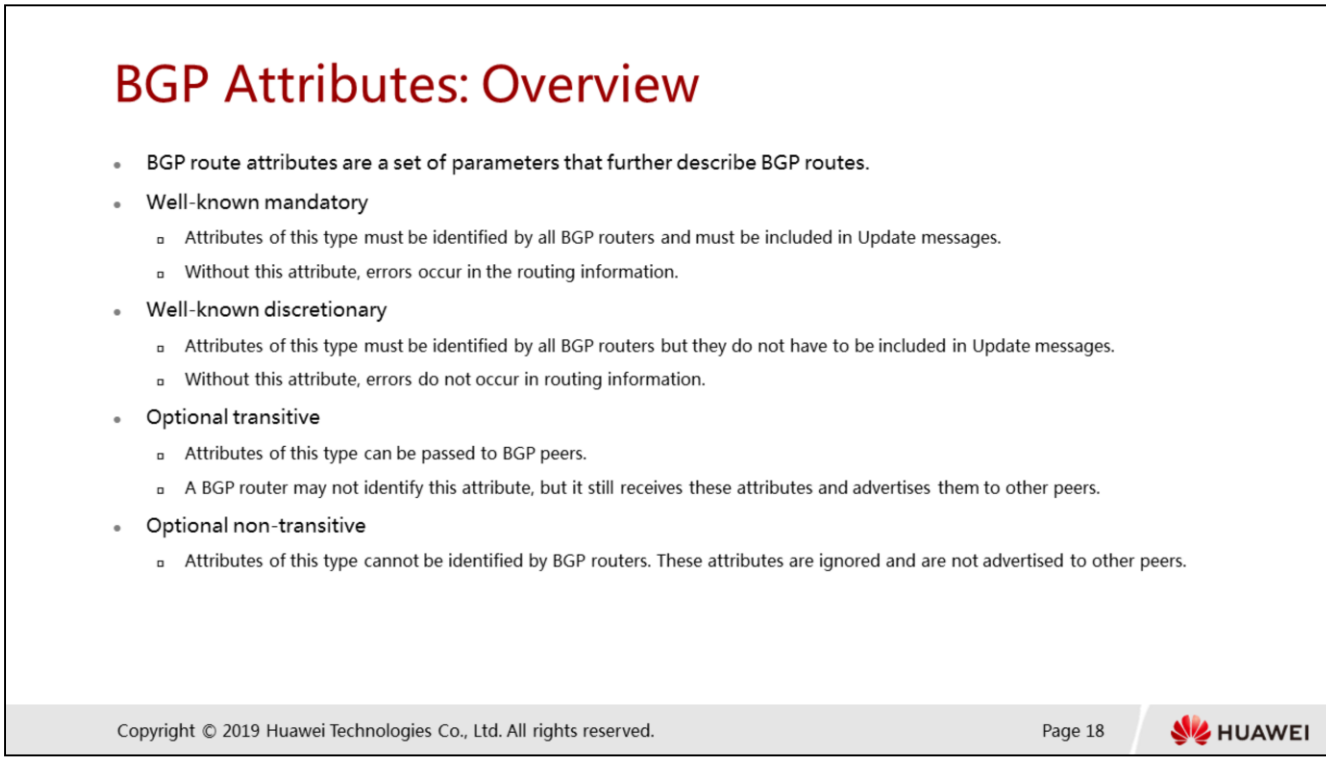
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IBGP and IGP are synchronized to prevent unreachable routes being imported to the external AS devices.

- Topology description (when synchronization is enabled)
  - R4 learns the network segment 10.0.0.0/24 advertised by R1 through BGP. Before advertising the network segment to R5, R4 checks whether this network segment exists in its IGP routing table. If so, R4 advertises the network segment to R5. If no, R4 cannot advertise it to R5.

- Precautions
  - By default, the synchronization mechanism of BGP and IGP is disabled on VRP and cannot be changed. However, synchronization can be canceled in either of the following scenarios:
  - The local AS is not a transit AS.
  - All routers in the AS establish IBGP full connections.



- BGP route attributes are a set of parameters that further describe BGP routes. With BGP route attributes, BGP can filter and select routes.

- The common attributes are as follows:
  - Origin: Well-known mandatory
  - AS_Path: Well-known mandatory
  - Next_Hop: Well-known mandatory
  - Local_Pref: Well-known discretionary
  - community: Optional transitive
  - Multi-exit-discriminator (MED): Optional non-transitive
  - Originator_ID: Optional non-transitive
  - Cluster_List: Optional non-transitive

BGP Attributes: Origin

- The Origin attribute is a well-known mandatory attribute that defines the origin of route information.
  - IGP
    - Routes obtained through an IGP in the AS from which the routes originate. For example, the Origin attribute of the routes imported to the BGP routing table using the **network** command is IGP.
    - A route of this type is identified with an 'i' in the BGP table.
  - EGP
    - Routes obtained through EGP.
    - A route of this type is identified with an 'e' in the BGP table.
  - Incomplete
    - Routes learned by other means, such as the routes imported by using the **import-route** command.
    - A route of this type is identified with an '?' in the BGP table.

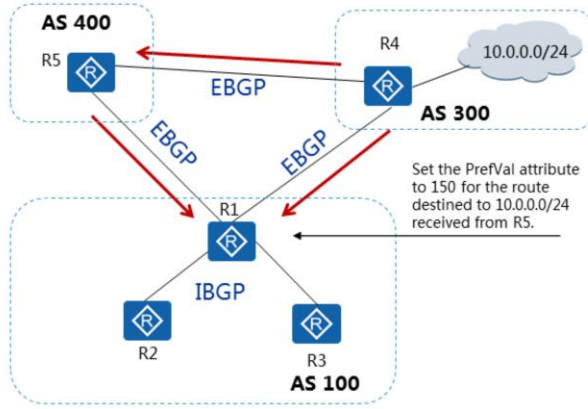
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The Origin attribute defines the origin of a route and marks the path of a BGP route. The Origin attribute has three possible values:
  - IGP: A route with EGP as the Origin attribute has the highest priority. IGP is the Origin attribute for the routes obtained through an IGP in the AS from which the routes originate. For example, the Origin attribute of the routes imported to the BGP routing table using the network command is IGP.
  - EGP: A route with EGP as the Origin attribute has the secondary highest priority. EGP is the Origin attribute for the routes obtained through EGP.
  - Incomplete: A route with Incomplete as the Origin attribute has the lowest priority. Incomplete is the Origin attribute for the routes learned by other means. For example, the Origin attribute of the routes imported by using the import-route command is Incomplete.

BGP Attributes: PrefVal

- PrefVal is Huawei-specific and valid only on the device where it is configured. The PreVal attribute is user-defined. Therefore, BGP first compares the PreVal values during route selection.



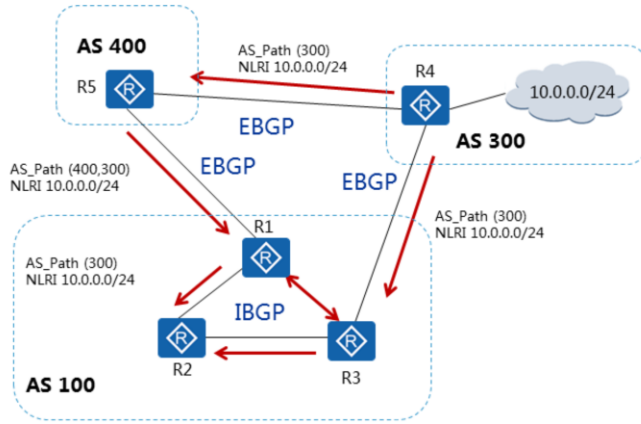
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- BGP first compares the PrefVal values during route selection. The default value is 0. A numerically larger value indicates a higher priority.

BGP Attributes: AS_Path

- The AS_Path attribute records the numbers of all ASs that a route passes through, from the source to the destination, in the vector order. This is a well-known mandatory attribute.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

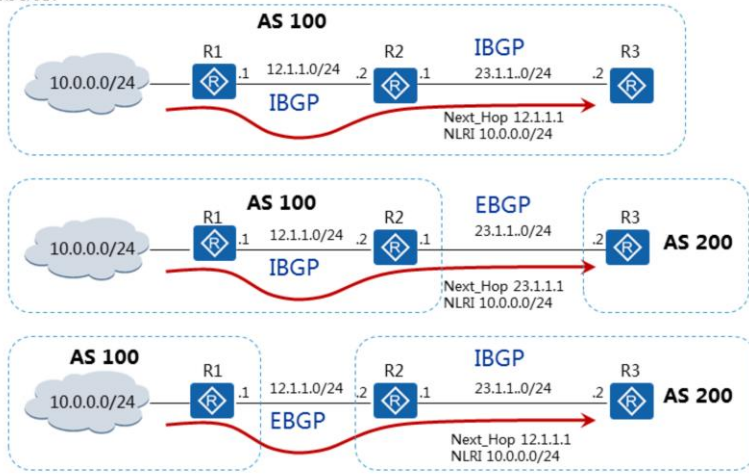
HUAWEI

- The AS_Path attribute can be used for BGP route selection. A shorter AS_Path length indicates a higher priority. In addition, to prevent inter-AS routing loops, a BGP router does not accept the routes whose AS_Path list contains the local AS number advertised from EBGP peers.
  - When a BGP speaker advertises a local route:
    - When advertising the route beyond the local AS, the BGP speaker adds the local AS number to the AS_Path list and then advertises it to the neighboring routers through Update messages.
    - When advertising the route to the local AS, the BGP speaker creates an empty AS_Path list in an Update message.
  - When a BGP speaker advertises a route learned from Update messages sent by another BGP speaker:
    - When advertising the route to other ASs, the BGP speaker adds the local AS number to the beginning of the AS_Path list. According to the AS_Path attribute, a BGP router that receives the route can know which ASs the route passes through before reaching the destination address. The number of the AS that is nearest to the local AS is placed on the top of the AS_Path list. The other AS numbers are listed according to the sequence in which the route passes through ASs.
    - When the BGP speaker advertises the route to the local AS, it does not change the AS_Path.

- Network topology:
  - When R4 advertises the network segment 10.0.0.0/24 to AS 400 and AS 100, it adds its local AS number to the AS-Path attribute. When R5 advertises the network segment 10.0.0.0/24 to AS 100, it adds its own AS number to the AS-Path attribute as well. When R1, R2, and R3 in AS 100 advertise network segment 10.0.0.0/24 to each other, the AS_PATH attributes of the routes do not change. If other conditions for BGP route selection are the same, BGP selects the route with the shortest AS_Path, that is, the route from R3 to R4.

BGP Attribute: Next_Hop

- The Next_Hop attribute records the next hop that a route passes through. This is a well-known mandatory attribute.



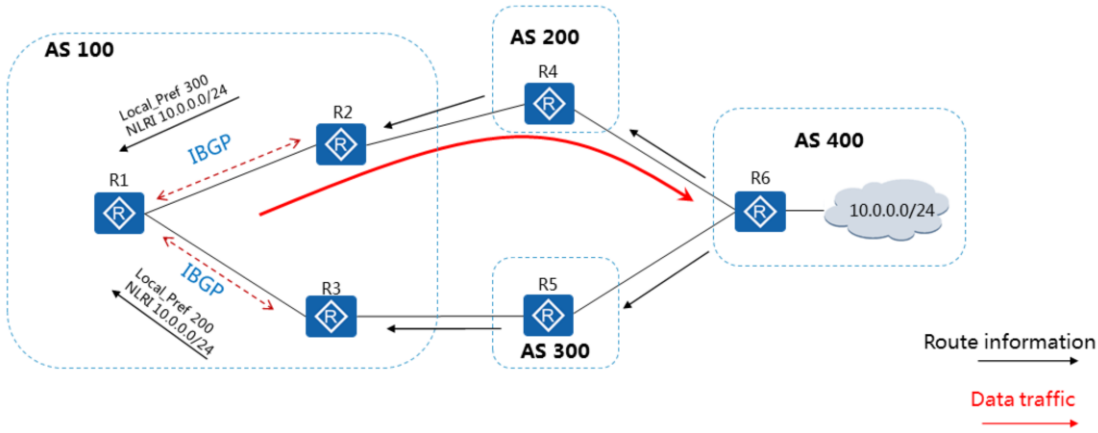
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The Next_Hop attribute records the next hop that a route passes through. The Next_Hop attribute of BGP is different from that of an IGP because it may not be an IP address of a BGP peer. A BGP speaker processes the Next_Hop attribute based on the following rules:
  - When advertising a locally originated route to an IBGP peer, the BGP speaker sets the Next_Hop attribute of the route to the address of the local interface through which the BGP peer relationship is established.
  - When advertising a route to an EBGP peer, a BGP speaker sets the Next_Hop attribute of the route to the address of the local interface through which the BGP peer relationship is established.
  - When advertising a route learned from an EBGP peer to an IBGP peer, the BGP speaker does not change the Next_Hop attribute of the route.

BGP Attributes: Local_Pref

- The Local_Pref attribute specifies the priority of a BGP router. A numerically larger value indicates a higher priority. This is a well-known discretionary attribute.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

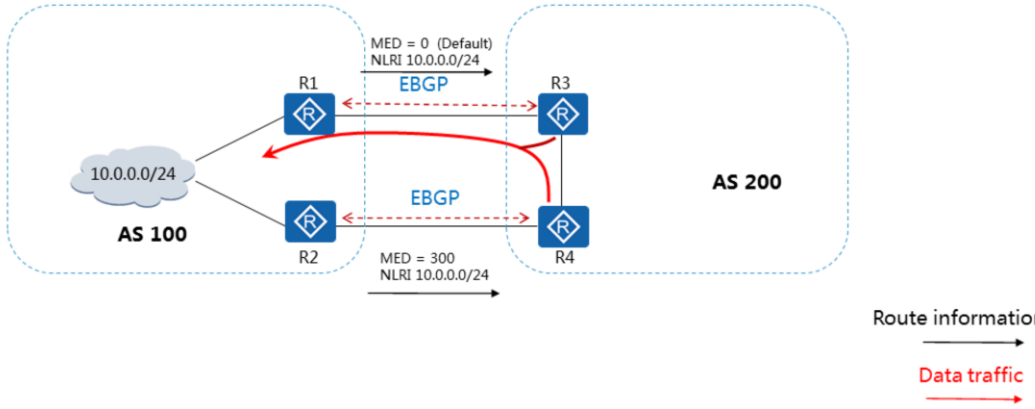
HUAWEI

- Local_Pref
  - It is exchanged only between IBGP peers and is not advertised to other ASs. It indicates priorities of BGP routers.
  - After a BGP router obtains multiple routes with the same destination address but different next hops from different IBGP peers, the route with a higher Local-Pref attribute value is selected.

- Topology description
  - IBGP peer relationships are established between R1, R2, and R3 in AS 100. R2 and R3 establish EBGP peer relationships with the routers in AS 200 and AS300 respectively. In this case, both R2 and R3 receive the route 10.0.0.0/24 from their EBGP peers. To enable the three routers in AS 100 to preferentially select R2 as egress of the 10.0.0.0/24 route in the local AS, you only need to modify the Local Pref attribute of the route on R2 and R3.

BGP Attributes: MED

- The MED attribute is similar to the Cost value in IGP, and is used for route selection between ASs. This is an optional non-transitive attribute.

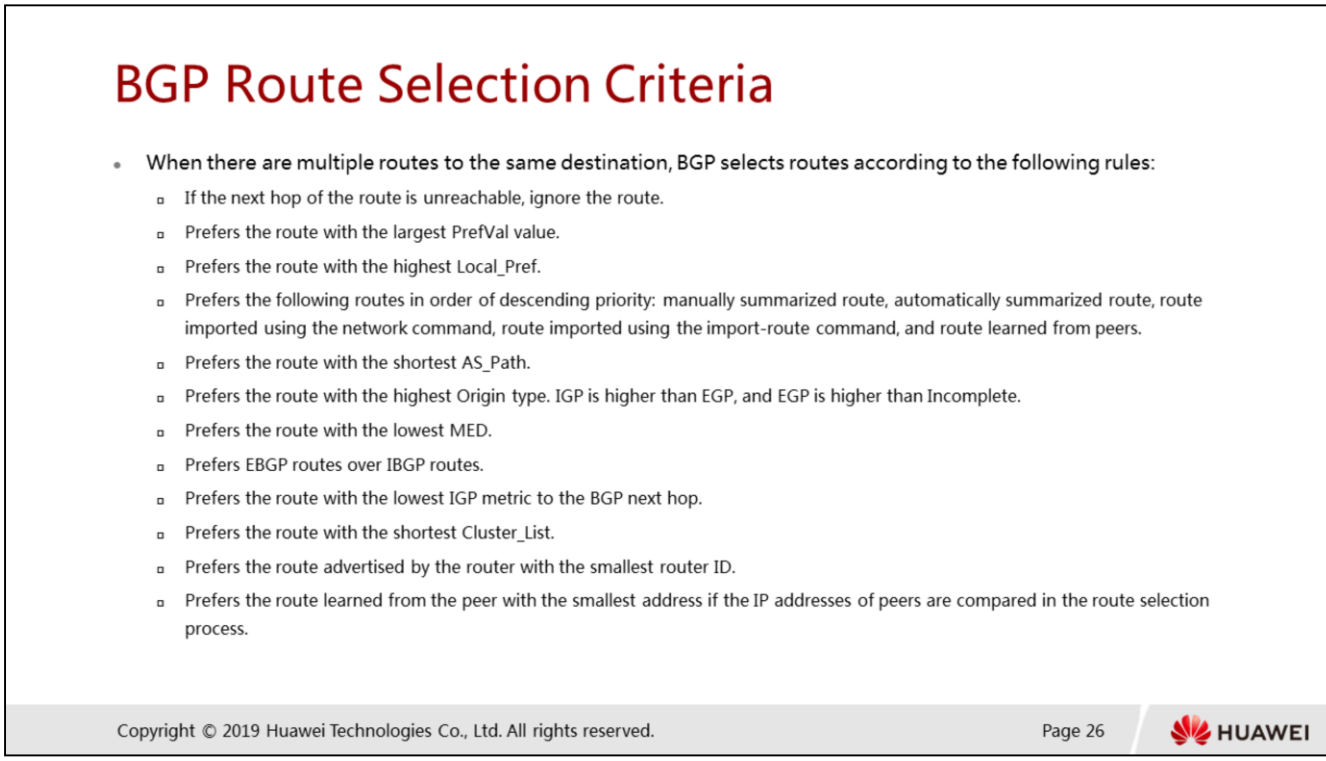


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- When a BGP device obtains multiple routes to the same destination address but with different next hops from different EBGP peers in one AS, the BGP device selects the route with the smallest MED value as the optimal route.

- The MED attribute is exchanged only between two neighboring ASs. The AS that receives the MED attribute does not advertise it to any other ASs. The MED attribute can be manually configured. If no MED attribute is configured for a route, the MED attribute of the route uses the default value 0.

- Topology description
  - R1 and R2 advertise the network segment 10.0.0.0/24 to their EBGP neighbors R3 and R4. R3 and R4 prefers the route with a lower MED value when other conditions are the same. That is, R3 and R4 accesses 10.0.0.0/24 through R1.



- BGP route selection criteria:
- The next hop IP address specified for a BGP route must be reachable.
- The PrefVal attribute is a Huawei-specific attribute and is valid only on the device where it is configured.
- If a route does not carry the Local_Pref attribute, the Local_Pref attribute of the route uses the default value 100. You can run the default local-preference command to set the default Local-Pref value of a BGP route.
- Locally originated routes include routes imported using the network command or the import-route command, manually summarized routes, and automatically summarized routes.
  - A summarized route is preferred. A summarized route takes precedence over a non-summarized route.
  - A route manually summarized using the aggregate command is preferred over a route automatically summarized using the summary automatic command.
  - A route imported using the network command is preferred over a route imported using the import-route command.
- Prefers the route with the shortest AS_Path.
  - The AS_Path attribute length contains the AS_Confed_Sequence and AS_Confed_Set.

    - A BGP router assumes that an AS_SET carries only one AS number regardless of the actual number of ASs it is carrying.
    - After you run the bestroute as-path-ignore command, the AS_Path attributes of routes are not compared in the route selection process.
- Prefers the route with the lowest MED.
    - BGP compares MEDs of only routes from the same AS, but not a confederation sub-AS. That is, MEDs of two routes are compared only when the first AS number in the AS_SEQUENCE (excluding AS_CONFED_SEQUENCE) is the same for the two routes.
    - A route without MED is assigned a MED of 0, unless the bestroute med-none-as-maximum command is run. If you run the bestroute med-none-as-maximum command, the route is assigned the highest MED of 4294967295.
    - After you run the compare-different-as-med command, MEDs in the routes received from peers in different ASs are compared. Do not use this command unless you confirm different ASs use the same IGP and route selection mode. Otherwise, a loop can occur.
    - If you run the bestroute med-confederation command, MEDs are compared for routes when the AS_Path attributes of the routes do not carry external AS numbers (not a confederation sub-AS) and the first AS number in the AS_CONFED_SEQUENCE are the same.
    - After you run the deterministic-med command, routes are not selected in the sequence in which routes are received.

- Load balancing
    - When there are multiple equal-cost routes to the same destination, you can perform load balancing among these routes to load balance traffic.
    - Equal-cost BGP routes can be used for traffic load balancing only when the attributes described before the "Prefers the route with the lowest IGP metric to the BGP next hop" rule are the same.



Contents

1. BGP Fundamentals
2. **BGP4+ Fundamentals**
3. Basic Configuration of BGP and BGP4+
4. BGP Fault Diagnosis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

2. BGP4+ Fundamentals
   - BGP4+ Overview
   - BGP4+ Capabilities Advertisement
   - Extended Attributes of BGP4+

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

BGP4+ Overview

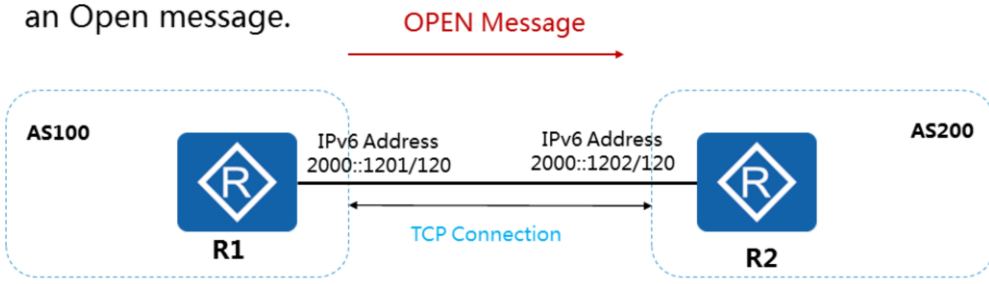
- BGP4+ Overview
  - An extension of BGP4
  - Provides capability negotiation
  - Supports multiple address families, including IPv6, VPNv4 and VPNv6
  - Transmits routing information about multiple address families via extended attributes

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

BGP4+ Capabilities Advertisement (1)

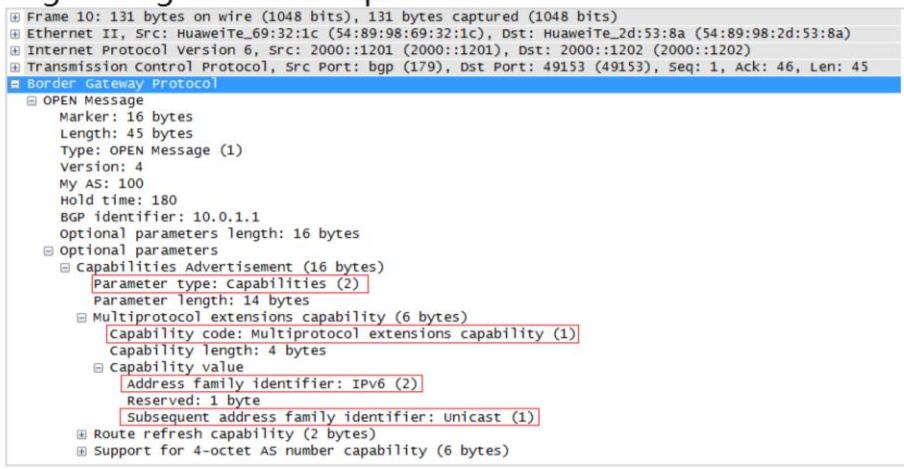
- When IPv6 addresses need to be transmitted between two BGP peers, the two BGP routers need to communicate the IPv6 capability to each other in an Open message.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

BGP4+ Capabilities Advertisement (2)

- The Capabilities Advertisement field in an Open message is used for negotiating extended capabilities.

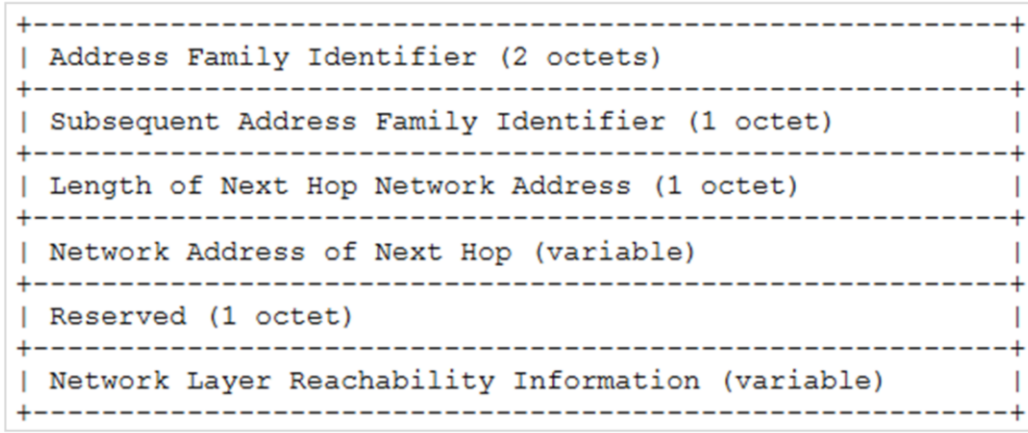


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In addition to the capability negotiation of multiple address families, the following capabilities can be negotiated in the Capabilities Advertisement field:
  - 4-byte AS number
  - Route-refresh capability
  - Multi-layer label capability

BGP4+ Extended Attribute: MP_REACH_NLRI (1)

- MP_REACH_NLRI (Type Code = 14)
  - This attribute is used to advertise IPv6 routes.
- 

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Address Family Information: consists of a 2-octets Address Family Identifier (AFI) and a 1-octet Subsequent Address Family Identifier (SAFI).
- Length of Next Hop Network Address: consists of 1 octet, indicating the length of the next hop address. Generally, the value is 16.
- Network Address of Next Hop: The length is variable and depends on Length of Network Hop Network Address. Generally, the value is a global unicast address.
- Reserved: consists of 1 octet. The value must be 0.
- Network Layer Reachability Information: lists the routes containing same attributes. If the value of this field is 0, the route is a default route.

BGP4+ Extended Attribute: MP_REACH_NLRI (2)

- When IPv6 routes are transmitted:
  - AFI = 2, SAFI = 1 (Unicast), SAFI = 2 (Multicast)
  - The Next Hop Address Length field determines the number of next hop addresses.
    - If the value is 16, the next hop address is a global unicast address of the next-hop router.
    - If the value is 32, the next hop address is the global unicast address and link-local address of the next-hop router.
  - Reserved field. The value is 0.
  - Network Layer Reachability Information (NLRI), which is a variable-length field and specifies the route prefix and mask

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

BGP4+ Extended Attribute: MP_REACH_NLRI (3)



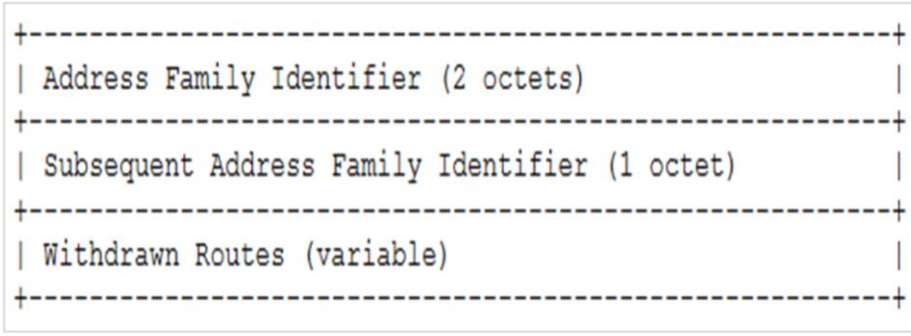

```
Border Gateway Protocol
  UPDATE Message
    Marker: 16 bytes
    Length: 77 bytes
    Type: UPDATE Message (2)
    Unfeasible routes length: 0 bytes
    Total path attribute length: 54 bytes
  Path attributes
    ORIGIN: IGP (4 bytes)
    AS_PATH: 100 (9 bytes)
    MULTI_EXIT_DISC: 0 (7 bytes)
    MP_REACH_NLRI (34 bytes)
      Flags: 0x90 (Optional, Non-transitive, Complete, Extended Length)
      Type code: MP_REACH_NLRI (14)
      Length: 30 bytes
      Address family: IPv6 (2)
      Subsequent address family identifier: Unicast (1)
      Next hop network address (16 bytes)
        Next hop: 2000::1201 (16)
      Subnetwork points of attachment: 0
      Network layer reachability information (9 bytes)
        2000:0:0:1::/64
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

BGP4+ Extended Attribute: MP_UNREACH_NLRI (1)

- MP_UNREACH_NLRI (Type Code = 15)
  - This attribute is used to withdraw routes.



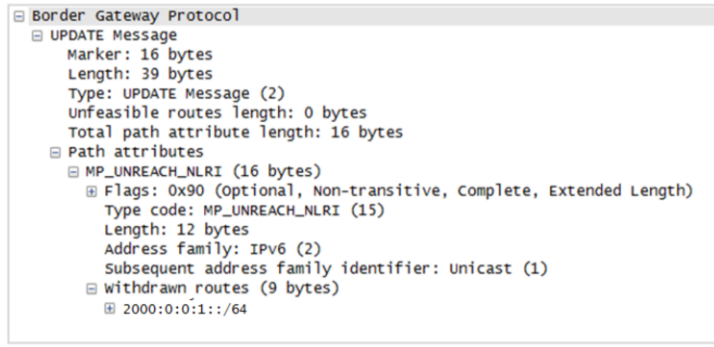
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Address Family Information: consists of a 2-octets AFI and a 1-octet SAFI.
- Withdrawn Routes: indicates the route to be withdrawn. The format is <mask length, route prefix>. If the mask length is 0, the route to be withdrawn is a default route.

BGP4+ Extended Attribute: MP_UNREACH_NLRI (2)

- When an IPv6 route is to be withdrawn:
  - AFI = 2, SAFI = 1 (Unicast), SAFI = 2 (Multicast)
  - The Withdrawn Routes field indicates the route prefix and mask of the route to be withdrawn.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



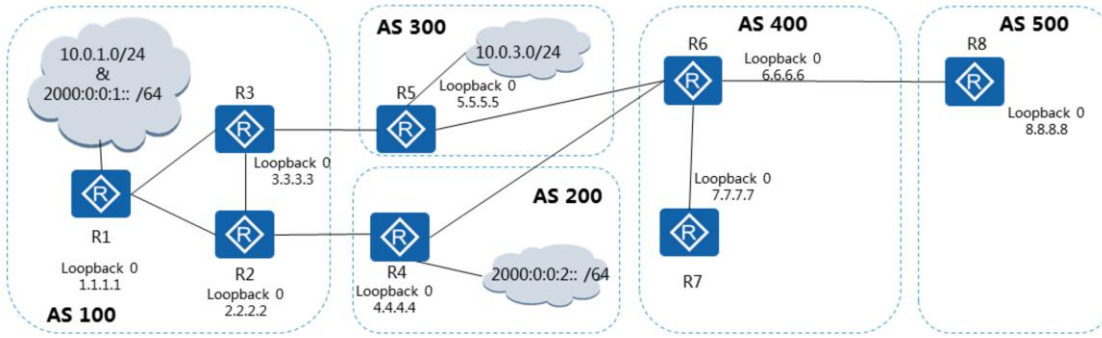
Contents

1. BGP Fundamentals
2. BGP4+ Fundamentals
3. **Basic Configuration of BGP and BGP4+**
4. BGP Fault Diagnosis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Basic Configuration of BGP (1)

- Configuration requirements:
  - An EBGP peer relationship needs to be established between ASBRs in each AS.
  - Stable IBGP peer relationships need to be established in each AS.
  - IPv4 and IPv6 networks can communicate with each other.
  - Routers in each AS can communicate with each other through IGP.

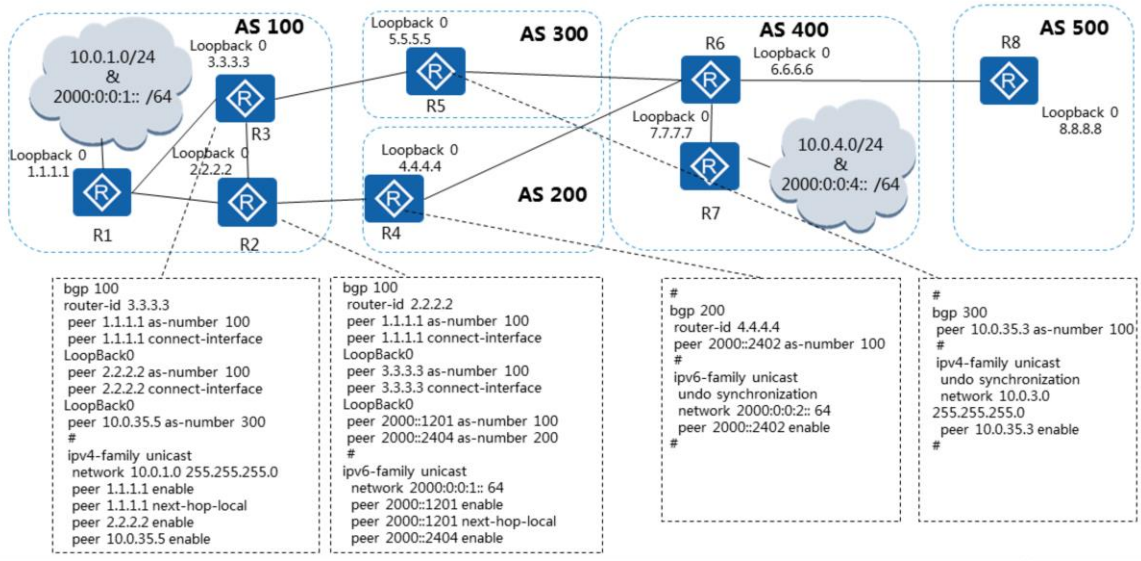


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IP address configuration rules:
  - The IPv4 network segment of the interfaces directly connecting Rx and Ry (X < Y) is 10.0.xy.0/24. The IPv4 address of the corresponding interface on Rx is 10.0.xy.x, and that on Ry is 10.0.xy.y.
  - The IPv6 network segment of the interfaces directly connecting Rx and Ry (X < Y) is 2000::xy00/120. The IPv6 address of the corresponding interface on Rx is 2000::xy0x, and that on Ry is 2000::xy0y.
  - The IPv6 address of loopback interface 0 on each router is 2000::z (z is the router ID).
- Notes:
  - OSPF and IS-IS can run in an AS to ensure routers in the AS can communicate with each other.
  - Stable IBGP relationships can be established through loopback interfaces.
  - EBGP neighbor relationships can be established directly through physical interfaces.

Basic Configuration of BGP (2)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Command description:
  - The peer as-number command specifies an AS number for a peer or peer group.
  - The peer connect-interface command specifies a source interface from which BGP packets are sent, and a source address used for initiating a connection.
  - The peer next-hop-local command configures a BGP device to set its IP address as the next hop of routes when the BGP device advertises routes to an IBGP peer or peer group.

- Command usage:
  - The preceding commands are run in the BGP process view.

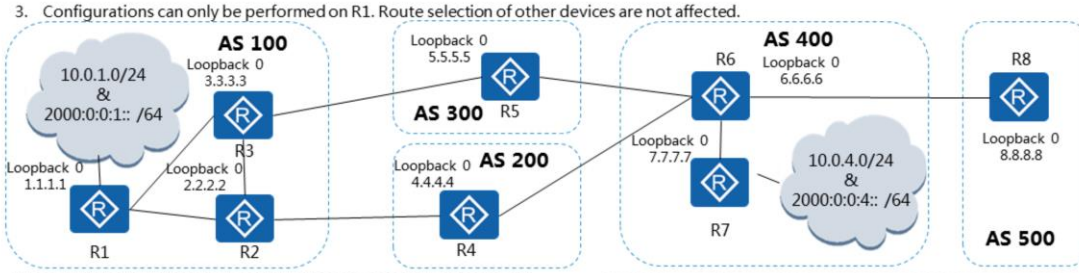
- Parameter description
  - peer ipv4-address as-number as-number
    - ip-address: specifies the IPv4 address of a peer.
    - as-number: specifies the AS number of a peer.
  - peer ipv4-address connect-interface interface-type interface-number [ipv4-source-address]
    - ip-address: specifies the IPv4 address of a peer.

        - interface-type interface-number: specifies the type and number of an interface.
        - Ipv4-source-address: specifies the source IPv4 address for establishing a BGP connection.
    - peer ipv4-address next-hop-local
        - ip-address: specifies the IPv4 address of a peer.

- Precautions
    - When configuring a device to use a loopback interface as the source interface of BGP messages, note the following points:
        - The IP address of the loopback interface must be reachable.
        - To establish an EBGP connection, you also need to run the peer ebgp-max-hop command to enable the two devices to establish an indirect peer relationship.
- The peer next-hop-local and peer next-hop-invariable commands are mutually exclusive.
- PrefRcv in the display bgp peer command output indicates the number of route prefixes that a BGP router receives from its peer.
- The configuration on an IPv6 nework is similar to that of IPv4. The difference is that after the peer address and AS number are specified, you need to enter the ipv6 unicast family view and run the peer peer-ip-address enable command to activate BGP.

Configuring BGP: PrefVal Attribute (1)

- Configuration requirements:
  1. The route from R1 to the network segment 10.0.4.0/24 in AS 400 passes through R1, R2, R4, R6, and R7 in sequence.
  2. The route from R1 to the network segment 2000:0:0:4::/64 in AS 400 passes through R1, R3, R5, R6, and R7 in sequence.
  3. Configurations can only be performed on R1. Route selection of other devices are not affected.




```
<R1>display bgp ipv6 routing-table 2000:0:0:4::64
BGP routing table entry information of 2000:0:0:4::/64:
  From: 2000::1202 (2.2.2.2)
  Original nexthop: 2000::1202
  AS-path 200 400, origin igp, localpref 100, pref-val 0, valid, internal, best,
select, active, pre 255
  From: 2000::1303 (3.3.3.3)
  Original nexthop: 2000::1303
  AS-path 300 400, origin igp, localpref 100, pref-val 0, valid, internal, pre 25
5, not preferred for router ID
```

```
<R1>display bgp routing-table
 Network            NextHop        MED        LocPrf    PrefVal Path/Ogn

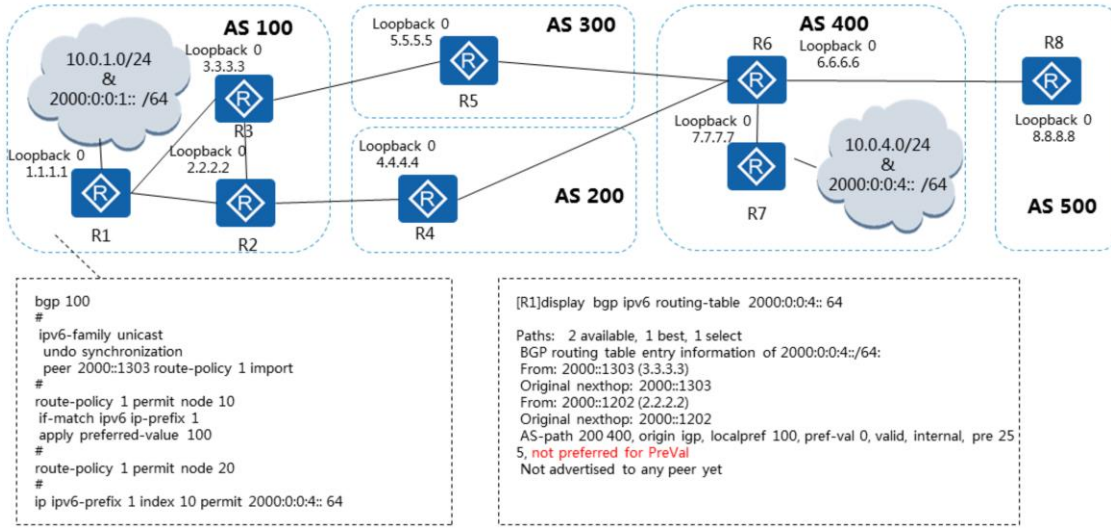
*>i 10.0.4.0/24     2.2.2.2                   100        0       200 400i
 *i                 3.3.3.3                   100        0       300 400i
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The topology is the same as that in BGP basic configuration. BGP peer relationships have been established.

Configuring BGP: PrefVal Attribute (2)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

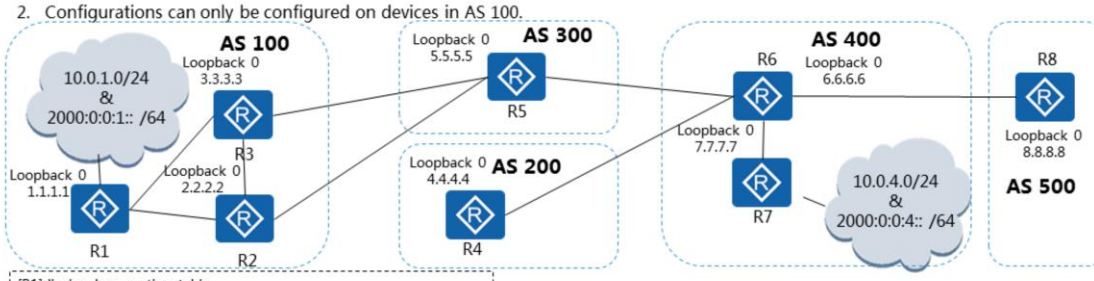
HUAWEI

- Command description:
  - The peer route-policy command specifies a routing policy for filtering routes received from a peer or peer group, or filtering routes to be advertised to a peer or peer group.
  - The apply preferred-value preferred-value command sets the action for changing the preferred value of BGP routes in a routing policy.
- Command usage:
  - The peer route-policy command is run in the BGP view.
- Parameter description
  - peer ipv4-address route-policy route-policy-name { import | export }
    - ipv4-address: specifies the IPv4 address of a peer.
    - route-policy-name: specifies the name of a routing policy.
    - import: applies a routing policy to routes received from a peer or peer group.
    - export: applies a routing policy to the route to be advertised to a peer or peer group.
    - preferred-value: specifies the preferred value of BGP routes. In route selection, the BGP route with the largest preferred value is preferred. The value is an integer in the range from 0 to 65535, and the default value is 0.

- Test result
  - By running the display bgp routing-table and display bgp ipv6 routing-table commands, you can check the BGP routing table.
- Precautions
  - The preferred value is a proprietary attribute of BGP, and this command takes effect only on BGP routes. The preferred value specifies the weight of a BGP route in BGP route selection. It is not a standard RFC-defined attribute and is valid only on the local device. The preferred value is inapplicable to export policies of BGP.

Configuring BGP: Local_Pref Attribute (1)

- Configuration requirements:
  1. Both the routes from R1 to 10.0.4.0/24 and 2000:0:0:4::/64 in AS 400 pass through R1, R3, R5, R6 and R7 in sequence.
  2. Configurations can only be configured on devices in AS 100.




```
[R1]display bgp routing-table
Network            NextHop        MED        LocPrf    PrefVal Path/Ogn

*>i 10.0.4.0/24    2.2.2.2                   100          0     300 400i
 *i                3.3.3.3                   100          0     300 400i
[R1]display bgp ipv6 routing-table

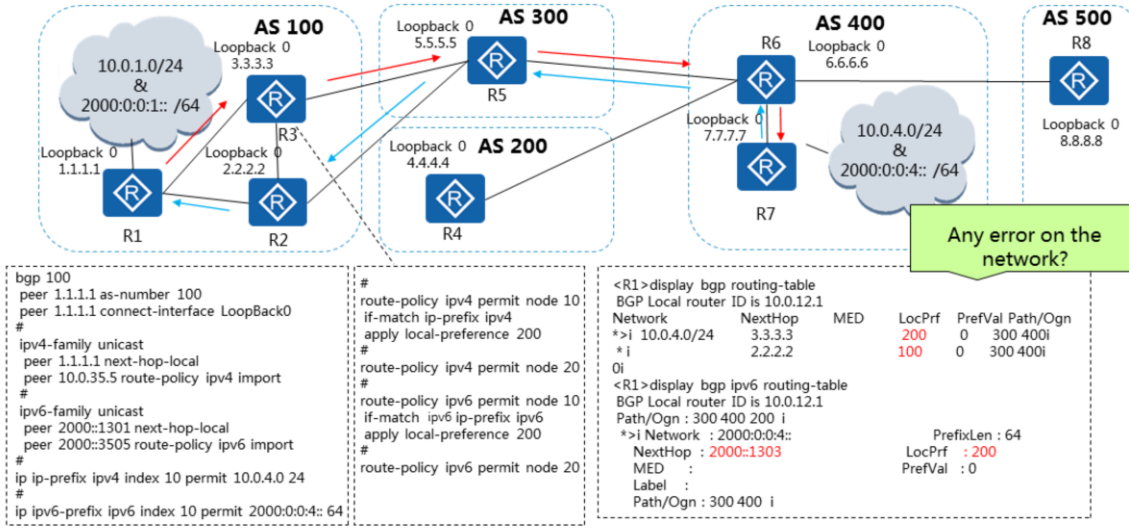
 *>i Network  : 2000:0:0:4::                PrefixLen : 64
     NextHop  : 2000::1202                  LocPrf    : 100
     MED      :                             PrefVal   : 0
     Label    :
     Path/Ogn : 300 400 i
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The topology is the same as that in BGP basic configuration. BGP peer relationships have been established.

Configuring BGP: Local_Pref Attribute (2)



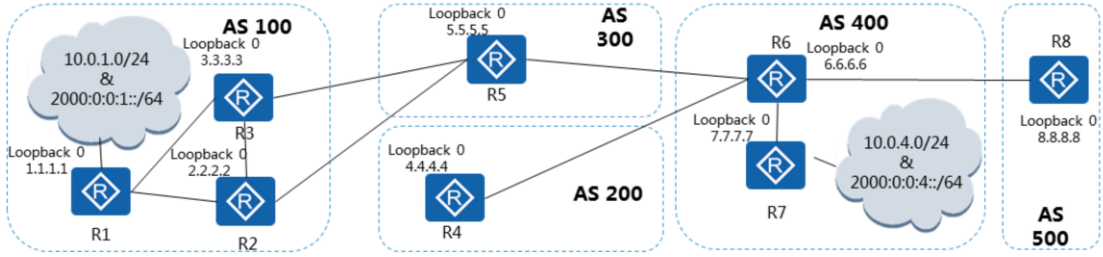
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Command description:
  - The apply local-preference preference command sets the local priority of a BGP route.
- Parameter description:
  - Preference: specifies the local priority of a BGP route. The value is an integer in the range from 0 to 4294967295. The default value is 100.
- Precautions
  - When a routing policy takes effect, it affects BGP route selection.
  - The Local_Pref attribute applies to the route selection within an AS, and is not advertised to the outside of the AS. In this case, the apply local-preference command does not take effect when an export routing policy for EBGP peers is configured.

Configuring BGP: MED Attribute (1)

- Configuration requirements:
  1. The network problem occurred when configuring the Local-Pref attribute is solved.
  2. Configurations can only be performed on devices in AS 100.




```
<R5>display bgp routing-table
Network                 NextHop        MED        LocPrf    PrefVal
Path/Ogn

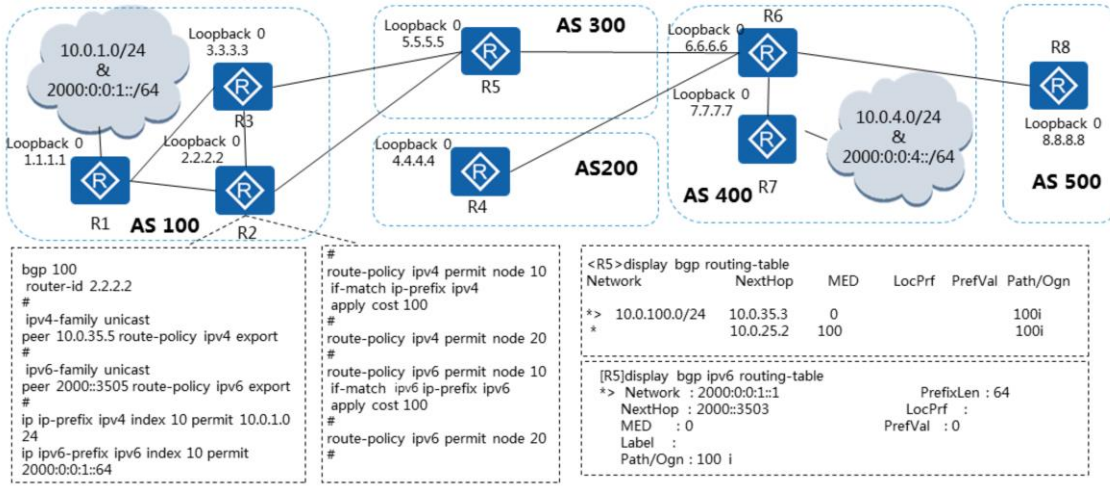
*>  10.0.100.0/24       10.0.25.2                 0         100i
*                       10.0.35.3                 0         100i
```

```
[R5]display bgp ipv6 routing-table
*>  Network  : 2000:0:0:1::1                    PrefixLen : 64
    NextHop  : 2000::2502                      LocPrf    :
    MED      : 0                               PrefVal   : 0
    Label    :
    Path/Ogn : 100 i
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- To solve the problem of inconsistent incoming and outgoing traffic paths, you can configure R2 to advertise routes with a higher MED attribute value so that R5 selects the routes advertised by R3.

Configuring BGP: MED Attribute (2)




```
bgp 100
 router-id 2.2.2.2
#
 ipv4-family unicast
peer 10.0.35.5 route-policy ipv4 export
#
 ipv6-family unicast
peer 2000::3505 route-policy ipv6 export
#
ip ip-prefix ipv4 index 10 permit 10.0.1.0
24
ip ipv6-prefix ipv6 index 10 permit
2000:0:0:1::64
```

```
#
route-policy ipv4 permit node 10
 if-match ip-prefix ipv4
 apply cost 100
#
route-policy ipv4 permit node 20
#
route-policy ipv6 permit node 10
 if-match ipv6 ip-prefix ipv6
 apply cost 100
#
route-policy ipv6 permit node 20
#
```

```
<R5>display bgp routing-table
Network                 NextHop        MED        LocPrf    PrefVal Path/Ogn

*>  10.0.100.0/24       10.0.35.3       0                             100i
*                       10.0.25.2       100                           100i
```

```
[R5]display bgp ipv6 routing-table
 *>  Network  : 2000:0:0:1::1                PrefixLen : 64
     NextHop  : 2000::3503                   LocPrf    :
     MED      : 0                            PrefVal   : 0
     Label    :
     Path/Ogn : 100 i
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

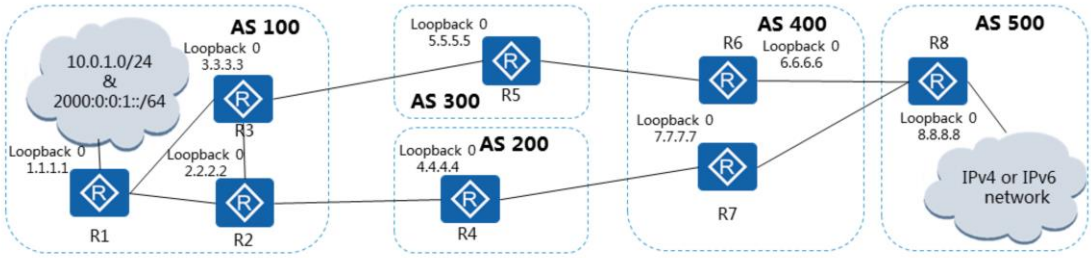
HUAWEI

- Command description:
  - The apply cost [+ | -] cost command sets the action for changing the cost of routes in a routing policy.
- Parameter description:
  - +: increases the route cost.
  - -: reduces the route cost.
  - cost: specifies the route cost. To control route selection, you need to modify the route cost to a specified value to prevent routing loops.
- Precautions
- By default, BGP compares the MED values of routes that come from the same AS only, excluding sub-ASs in a confederation. To enable BGP to compare MED values of routes in a confederation when selecting the optimal route, run the bestroute med-confederation command.

- After the bestroute med-confederation command is run, BGP compares MED values only when AS_Path does not contain an external AS (AS that is not in the confederation) number.
    - For example, ASs 65000, 65001, 65002, and 65004 belong the same confederation. Routes to the same destination are listed as follows:
        - path1: AS_Path = 65000 65004, MED = 2
        - path2: AS_Path = 65001 65004, MED = 3
        - path3: AS_Path = 65002 65004, MED = 4
        - path4: AS_Path = 65003 65004, MED = 1
- After the bestroute med-confederation command is run, the AS_Path attributes of paths 1, 2, and 3 do not contain the numbers of ASs that belong to other confederations, but the AS_Path attribute of path 4 contains the number of an AS that belongs to another confederation. Therefore, when selecting routes based on MED values, BGP compares the MED values of paths 1, 2, and 3 only.

Configuring BGP: AS_Path Attribute 1 (1)

- Configuration requirements
  1. Users in AS 500 need to preferentially select the route passing through AS 300 to AS 100.
  2. Configurations can only be performed on the router in AS 200.

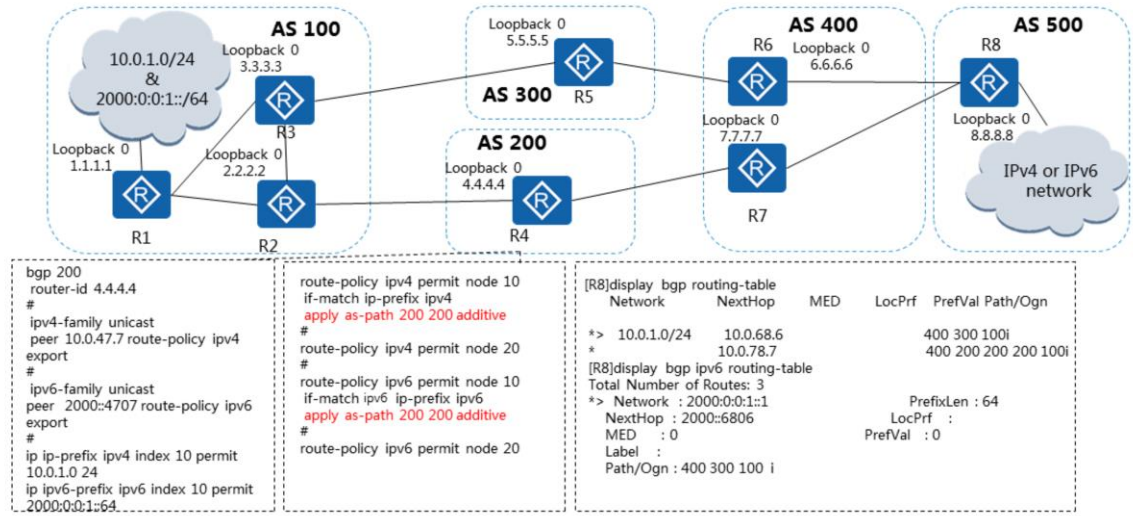


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The topology and configurations are the same as those in BGP basic configuration. Basic BGP peer relationships have been established.

Configuring BGP: AS_Path Attribute 1 (2)



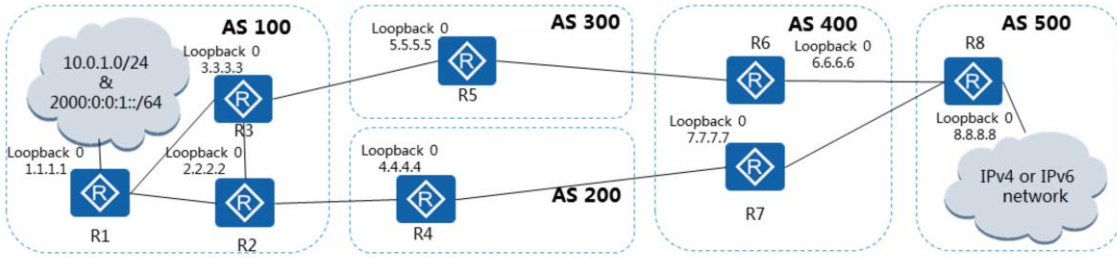
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Command description:
  - apply as-path { { as-number-plain | as-number-dot } &<1-10> { additive } | none overwrite }
- Parameter description:
  - as-number-plain: specifies an integral AS number to be added to the AS_Path list or to replace an AS number in the existing AS_Path list. A maximum of 10 AS numbers can be specified in one command.
  - as-number-dot: specifies an AS number in dotted notation to be added to the AS_Path list or to replace an AS number in the existing AS_Path list. A maximum of 10 AS numbers can be specified in one command.
  - additive: adds the specified AS number to the existing AS_Path list.
  - overwrite: replaces an AS number in the existing AS_Path list with a specified AS number.
  - None: clears the existing AS_Path list.
- Precautions
  - When a routing policy takes effect, it affects BGP route selection.
  - Running the apply as-path command changes the path through which network traffic passes through, or causes routing loops and incorrect route selection. Use this command only when you are familiar with the network topology and impact of the command on services.

Configuring BGP: AS_Path Attribute 2 (1)

- Configuration requirements
  1. Users in AS 500 need to preferentially select the route passing through AS 300 to AS 100.
  2. Configurations can only be performed on the router in AS 500.
  3. The router in AS 500 does not know the topology of the entire network.

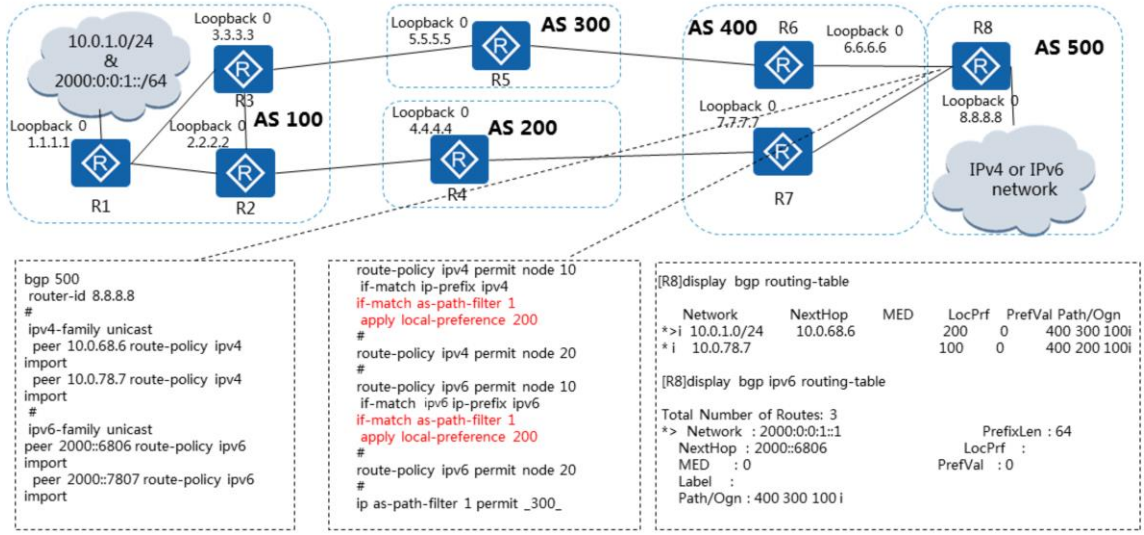


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The topology and configurations are the same as those in BGP basic configuration. Basic BGP peer relationships have been established.

Configuring BGP: AS_Path Attribute 2 (2)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Command description:
  - if-match as-path-filter { as-path-filter-number &<1-16> | as-path-filter-name }
- Parameter description:
  - as-path-filter-number: specifies the number of an AS_Path filter. The value is an integer in the range from 1 to 256. A maximum of 16 AS_Path filters can be specified in a command.
  - as-path-filter-name: specifies the name of an AS_Path filter. The value is a string of 1 to 51 case-sensitive characters, spaces not supported. The string cannot be all digits.
- Precautions
  - A maximum of 16 AS_Path filters can be specified in a command. The relationship between these AS_Path filters is OR. Specifically, if a route matches one of these AS_Path filters, it matches the matching rules of the command.

Configuring BGP: AS_Path Filter

- AS_Path filter
  - An AS_Path filter is used to filter BGP routes based on AS_Path attributes contained in the BGP routes.
  - The default action of an AS_Path filter is deny. Specifically, a route that has not been permitted in a filtering is denied by this filter.
  - The matching conditions of the AS _ Path filter are specified by regular expressions.
    - Examples:
      1. **^10_** indicates that the routes whose AS_Path attribute starts with AS 10 are to be matched.
      2. **_20_** indicates that the routes whose AS_Path attribute contains AS 20 are to be matched.
      3. **.*** indicates that all AS_Path attributes are to be matched.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Command description:
  - This ip as-path-filter { as-path-filter-number | as-path-filter-name } { deny | permit } regular-expression command creates an AS_Path filter.
- Parameter description:
  - as-path-filter-number: specifies the number of an AS_Path filter. The value is an integer in the range from 1 to 256.
  - as-path-filter-name: specifies the name of an AS_Path filter. The value is a string of 1 to 51 case-sensitive characters, spaces not supported. The string cannot be all digits. Spaces are allowed only when the string is enclosed in double quotation marks (" ").
  - deny: sets the action of the AS_Path filter to deny.
  - permit: sets the action of the AS_Path filter to permit.

Configuring BGP: Regular Expressions for AS_Path Filters (1)

Special Character	Function	Example
\	An escape character. It converts a special or common character following to it to a common character.	* matches the asterisk (*).
^	Matches the start of a line.	^10 matches "10,20,30", "100,200,300", and so on.
$	Matches the end position of the string.	1$ matches 101, 11, and so on.
*	Matches a sub-regular expression before the asterisk (*) for zero or multiple times.	10* matches 1, 10, 100, 1000, and so on. (10)* matches null, 10, 1010, 101010, and so on.
+	Matches a sub-regular expression before the asterisk (*) for once or multiple times.	10+ matches 10, 100, 1000, and so on. (10)+ matches 10, 1010, 10101, and so on.
?	Matches a sub-regular expression before the asterisk (*) for zero times or once.	10? matches 1 or 10. (10)? matches null or 10.
.	Matches any single character.	.0 matches 20, 30, and so on.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

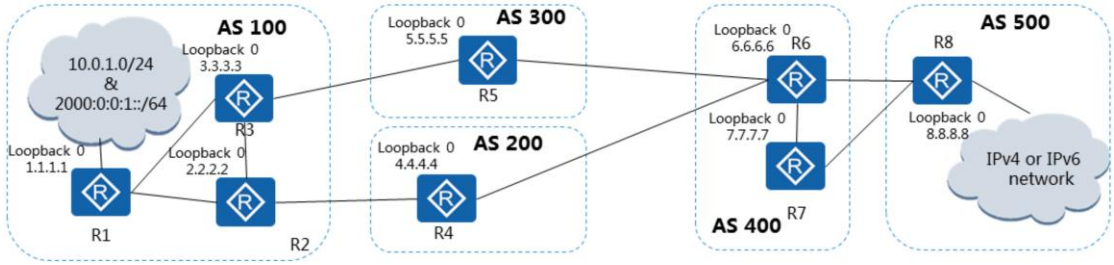
Configuring BGP: Regular Expressions for AS_Path Filters (2)

Special Character	Function	Example
()	Defines a sub-regular expression, which can be empty. Both the expression and the sub-regular expression must match.	100(2)+ matches 1002 and 10022.
_	Matches regular expressions with a special character, such as a comma (,), left brace ({), right brace (}), left parenthesis ((), right parenthesis ()), or space. Special characters can be used at the beginning of a regular expression with the same function as the caret (^) or at the end of a regular expression with the same function as the dollar sign ($).	_65001_ matches 65001, 20 65001, 65001 30, 65001, and so on.
x|y	Matches x or y.	100|200 matches 100 or 200. 1(2|3)4 matches 124 or 134.
[xyz]	Matches any character contained in a regular expression.	[123] matches the character 2 in 255.
[^xyz]	Matches any characters, except those listed in the regular expression.	[^123] matches any character except 1, 2, and 3.
[a-z]	Matches any character within the specified range.	[0-9] matches any character in the range from 0 to 9.
[^a-z]	Matches any character beyond the specified range.	[^0-9] matches all non-digit characters.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuring BGP: Load Balancing (1)

- Configuration requirements
  1. The router in AS 500 can access AS 100 through two links.
  2. Configurations can only be performed on the router in AS 500.



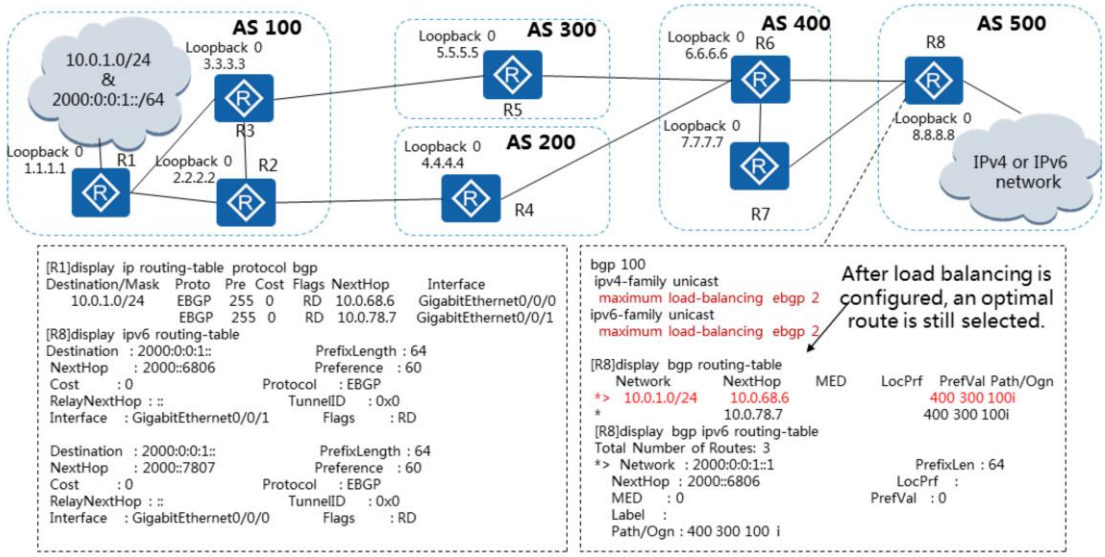

```
[R8]display ip routing-table protocol bgp
Destination/Mask   Proto   Pre  Cost      Flags NextHop         Interface
      10.0.1.0/24  EBGP    255  0           RD  10.0.68.6       GigabitEthernet0/0/0

[R8]display ipv6 routing-table protocol bgp
 Destination  : 2000:0:0:1::                PrefixLength : 64
 NextHop      : 2000::6806                  Preference   : 255
 Cost         : 0                           Protocol     : EBGP
 RelayNextHop : ::                          TunnelID     : 0x0
 Interface    : GigabitEthernet0/0/0        Flags        : D
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The topology and configurations are the same as those in BGP basic configuration. Basic BGP peer relationships have been established.

Configuring BGP: Load Balancing (2)




```
[R1]display ip routing-table protocol bgp
Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface
      10.0.1.0/24   EBGP    255  0           RD  10.0.68.6       GigabitEthernet0/0/0
                    EBGP    255  0           RD  10.0.78.7       GigabitEthernet0/0/1
[R8]display ipv6 routing-table
Destination : 2000:0:0:1::                PrefixLength : 64
NextHop     : 2000::6806                  Preference   : 60
Cost        : 0                           Protocol     : EBGP
RelayNextHop : ::                         TunnelID     : 0x0
Interface   : GigabitEthernet0/0/1        Flags        : RD

Destination : 2000:0:0:1::                PrefixLength : 64
NextHop     : 2000::7807                  Preference   : 60
Cost        : 0                           Protocol     : EBGP
RelayNextHop : ::                         TunnelID     : 0x0
Interface   : GigabitEthernet0/0/0        Flags        : RD
```

```
bgp 100
 ipv4-family unicast
  maximum load-balancing ebgp 2
 ipv6-family unicast
  maximum load-balancing ebgp 2

[R8]display bgp routing-table
     Network          NextHop        MED        LocPrf    PrefVal Path/Ogn
*>   10.0.1.0/24      10.0.68.6                             400 300 100i
*                     10.0.78.7                             400 300 100i
[R8]display bgp ipv6 routing-table
Total Number of Routes: 3
*>  Network  : 2000:0:0:1::1                 PrefixLen : 64
    NextHop  : 2000::6806                     LocPrf    :
    MED      : 0                              PrefVal   : 0
    Label    :
    Path/Ogn : 400 300 100 i
```

After load balancing is configured, an optimal route is still selected.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Command description:
  - The maximum load-balancing command configures the maximum number of equal-cost routes for load balancing.
- Command usage:
  - The maximum load-balancing command is run in the BGP view.
- Parameter description:
  - ebgp: Only EBGP routes take part in load balancing.
  - ibgp: Only IBGP routes take part in load balancing.
  - number: specifies the maximum number of equal-cost routes in the BGP routing table.

- Precautions:
  - The maximum load-balancing number command cannot be configured together with the maximum load-balancing ebgp number or maximum load-balancing ibgp number command.
  - Routes that have the same AS_Path length and AS_Path sequence can be used to balance loads. The load-balancing as-path-ignore command prevents a router from comparing the AS_Path attributes of routes when selecting routes for load balancing.
- Test result:
  - After running the display ip routing-table protocol bgp command, you can check the equal-cost routes learned through BGP.



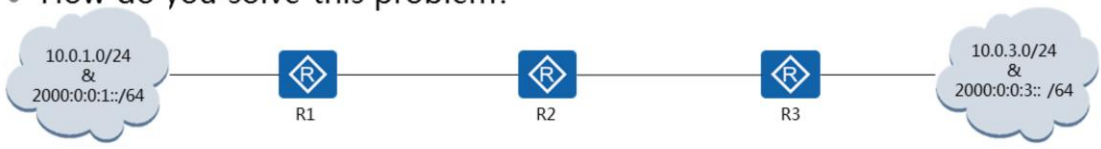
Contents

1. BGP Fundamentals
2. BGP4+ Fundamentals
3. **Basic Configuration of BGP and BGP4+**
4. BGP Fault Diagnosis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

BGP Fault

- BGP peer relationships are established between R1 and R2, and between R2 and R3.
- After the configuration is complete, the networks connected to R1 and R3 respectively cannot communicate with each other.
- How do you solve this problem?



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

BGP Troubleshooting Procedure (1)

- Assume the fault is caused by BGP:
  - If a BGP peer relationship cannot enter the Established state, it indicates that:
    - IGP is unavailable.
    - An ACL is configured to filter the packets whose destination port is TCP port 179.
    - The peer router ID conflicts with the local router ID.
    - The peer AS number is incorrect.
    - Loopback interfaces are used to establish BGP peer relationships, but the **peer connect-interface** command is not configured.
    - The **peer ebgp-max-hop** command is not configured when the EBGP peer relationship is established.
    - The configurations of the **peer valid-ttl-hops** command are incorrect.
    - The number of routes sent by the remote router exceeds the upper limit configured using the **peer route-limit** command.
    - The **peer ignore** command is configured on the peer.
    - Address families of devices on both ends do not match.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

BGP Troubleshooting Procedure (2)

- If the BGP peer relationship is normal, but the BGP routing table does not contain this route entry, check:
  - Whether the next hop address is reachable.
  - Whether a control policy is configured on the ingress.
  - Whether the prefixes of the routes that can be received are limited.
  - Whether a policy is configured on the egress of the peer.
  - Whether the prefix is the optimal one in the peer's BGP routing table.
  - Whether the peer is configured with the **active-route-advertise** command.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

BGP Troubleshooting Procedure (3)

- If the BGP peer relationship is normal but some entries in the BGP routing table are not optimal, check:
  - Which entries are not optimal based on route selection rules.
  - Whether the routes with some certain prefixes are suppressed.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Quiz

1. BGP negotiates the address family capability in Open messages.
   A. True
   B. False

2. If routes have the same AS_Path length but with different AS_Path sequences, load balancing is not affected.
   A. True
   B. False

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Answers:
  - True.
  - False. Only routes that have the same AS_Path length and AS_Path sequence can be used to balance loads.

Summary

- BGP 4 & BGP4+ Fundamentals
  - BGP fundamentals
  - BGP attributes and features
  - BGP4+ fundamentals and attributes
- BGP4 & BGP4+ Configuration
  - Configuration commands on Huawei devices
  - Configurations in different scenarios
- BGP Troubleshooting
  - BGP troubleshooting procedure

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— BGP Advance & Internet

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- How to understand various Border Gateway Protocol (BGP) advanced features and technologies
- How to properly use various features and technologies to improve network performance
- How to troubleshoot possible problems on a BGP network

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to:
  - Be familiar with common BGP tools and technologies when a great number of routes exist on a network.
  - Understand the extended and enhanced features of BGP.
  - Understand the concept of the Internet design.
  - Master BGP attribute configuration and troubleshooting.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

- **Technologies Regarding Mass BGP Routes**
- Extended BGP Features
- Enhanced BGP Features
- Internet Design Fundamentals
- BGP Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

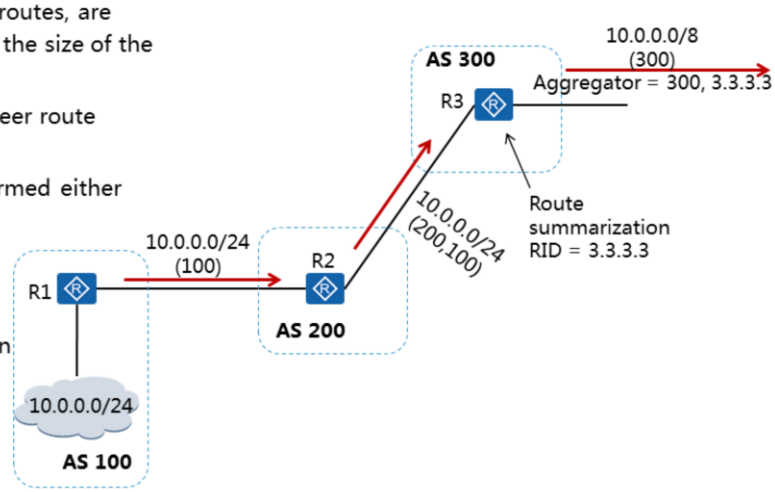
Technologies Regarding Mass BGP Routes

- The following tools or technologies are used to simplify configurations, reduce routing entries, and improve device performance on a large-scale network or a network with a large number of routing entries:
  - Route summarization
  - Peer group
  - Community attributes
  - Route reflection
  - BGP confederation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Route Summarization

- Route summarization
  - Summary routes, not specific routes, are advertised to peers to reduce the size of the routing table of the peers.
  - If a specific route flaps, BGP peer route calculation is not triggered.
  - Route summarization is performed either automatically or manually.
    - IPv4 routes:
      - Manual summarization
      - Automatic summarization
    - IPv6 routes:
      - Manual summarization



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

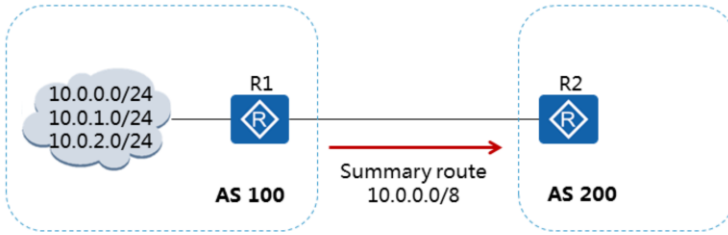
- On a large-scale network, the BGP routing table is in a huge size, which greatly burdens a device, increases the probability of route flapping, and reduces network stability.

- Route summarization is the mechanism that combines multiple routes into one. It reduces the size of the routing table by advertising only summary routes to peers without advertising each specific route. If a summary route flaps, the network is no longer adversely affected, which improves network stability.

- The Aggregator attribute (optional transitive) is used in route summarization. This attribute identifies the node on which routes are summarized and carries the router ID and AS number of the summarization node.

Route Summarization - Automatic Summarization

- Automatic summarization
  - Performs classful summarization on routes imported by BGP.
  - Suppresses the advertisement of member specific routes after route summarization is configured.
  - Summarizes the routes that BGP imports by protocol types (**import-route**), not by prefixes or masks (**network**).



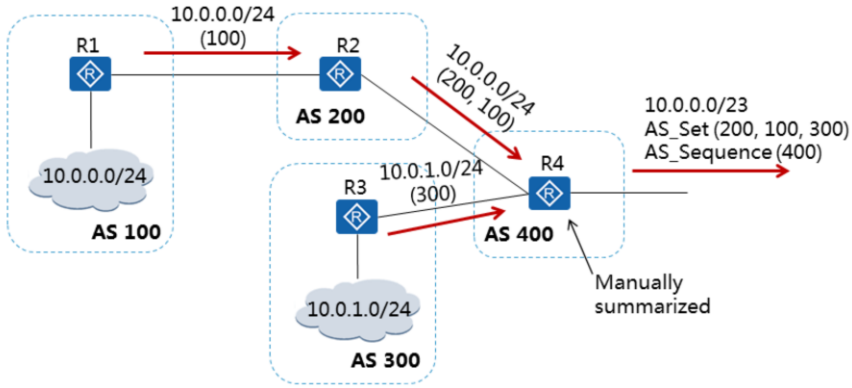
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Precautions for Automatic Summarization
  - This command summarizes the routes imported by BGP. The imported routes can be direct routes, static routes, OSPF routes, or IS-IS routes. With route summarization enabled, BGP summarize routes of each natural network segment into one route. Specific route information is no longer carried in BGP Update messages. This command does not take effect on routes imported using the network command.
  - BGP advertises only summarized routes to peers.
  - Automatic summarization is disabled for BGP by default.
  - A summary route carries the Atomic_Aggregate and Aggregator attributes.

Route Summarization - Manual Summarization (1)

- Manual summarization
  - Summarizes routes in the local BGP routing table. Controls the attributes of summarized routes and determines whether to advertise specific routes.

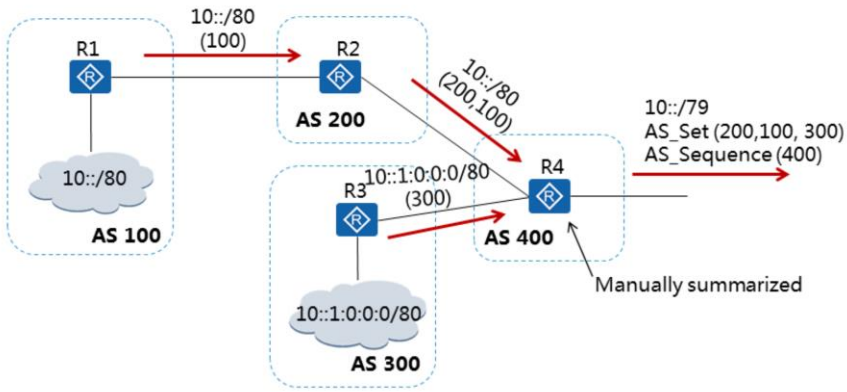


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Manual summarization
  - You can run a command to determine whether to suppress specific routes. After the suppression, the summarized routes carry the Atomic_Aggregate attribute.
  - The summary route does not carry AS-Path attributes of specific routes.
  - The AS_Set attribute is used to carry an AS number to prevent loops. The difference between AS_Set and AS_Sequence is as follows: The AS_Set option is an unordered list of AS numbers used for route summarization. The AS_Sequence option is an ordered list of AS numbers. Each time a message passes through an AS, an AS number is added. The AS numbers are listed in descending order.

Route Summarization - Manual Summarization (2)



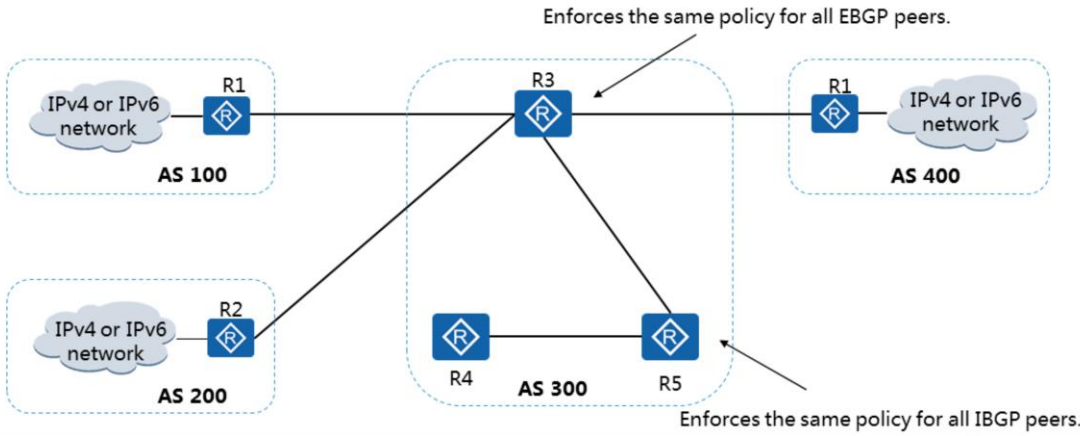
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Manual summarization
  - You can run a command to determine whether to suppress specific routes. After the suppression, the summary routes carry the Atomic_Aggregate attribute.
  - The summary route does not carry AS-Path attributes of member specific routes.
  - The AS_Set attribute is used to carry an AS number to prevent loops. The difference between AS_Set and AS_Sequence is as follows: The AS_Set option is an unordered list of AS numbers used for route summarization. The AS_Sequence option is an ordered list of AS numbers. Each time a message passes through an AS, the AS numbers are listed in descending order.

Peer Group

- A set of peers with the same policy configured.
- Simplifies BGP configuration and reduces route performance loss.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

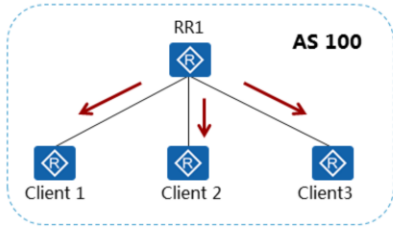
- A set of peers with the same policy configured. When a peer is added to a peer group, the peer obtains the same configuration as the peer group. If the configuration of a peer group is changed, the configurations of group members are also changed.

- A large BGP network has a large number of peers, many of which use the same policy. Some commands are repeatedly used when such peers are configured. In this situation, a peer group can be used to simplify the configuration.

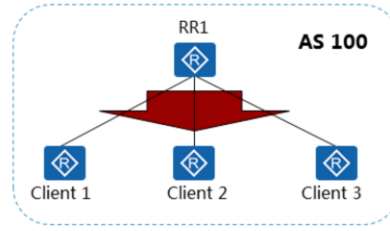
- A peer in a peer group can also have its own policy configured to advertise and receive routes.

Dynamic Update Peer-Group Enabled

- BGP dynamic update peer-group
  - The dynamic update peer-group function treats all BGP peers with the same export policies as an update peer-group.
  - Each route to be advertised is grouped once and then advertised to all peers in the update peer-group.



BGP dynamic update peer-group disabled



Dynamic update peer-group enabled

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

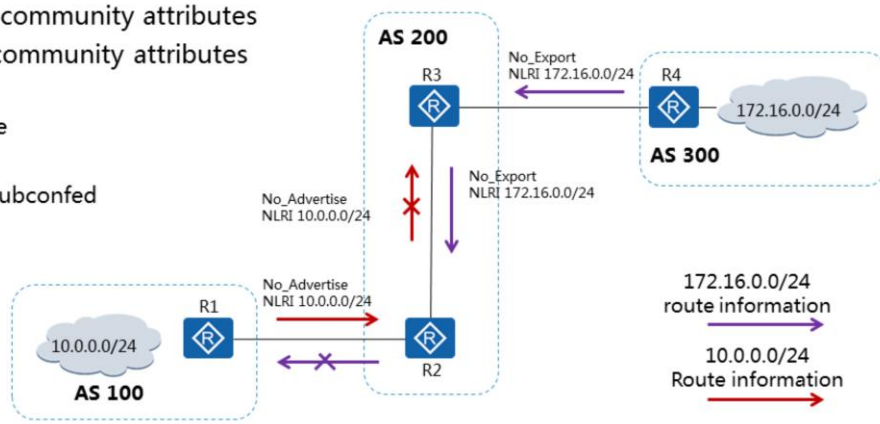
HUAWEI

- BGP Dynamic Update Peer-Groups
  - By default, BGP groups routes for each peer (even though the peers share an export policy).
  - After the dynamic update peer-group function is enabled, each route to be advertised is grouped only once and then advertised to all peers in the update peer-group, which increases grouping efficiency.

- Topology Description
  - A reflector has three clients, and 100,000 routes need to be reflected. If an RR groups a route for each peer, the total number of times that all routes are grouped is 100,000 x 3 before the RR advertises the routes to the three clients. The dynamic update peer-groups function reduces the total number of times that all routes are grouped to 100,000 x 1, which improves performance to three times.

BGP Attribute Characteristics - Community

- A community attribute is used to identify BGP routes with the same characteristics. This attribute is optional transitive.
  - Community attribute types:
    - User-defined community attributes
    - Well-Known community attributes
      - Internet
      - No_Advertise
      - No_Export
      - No_Export_Subconfed



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Community Attribute - Well-Known Attributes

- A community attribute is expressed in a 4-byte list. The format is as follows:
  - aa:nn: are integers ranging from 0 to 65535.
  - Community number: is an integer ranging from 0 to 4294967295.
  - As defined in RFC 1997, the ranges of 0 (0x00000000) through 65535 (0x0000FFFF) and 4294901760 (0xFFFF0000) through 4294967295 (0xFFFFFFFF) are reserved.

Community Name	Community ID	Description
Internet	0 (0x00000000)	By default, all routes belong to the Internet community. Routes with this attribute can be advertised to all BGP peers.
No_Export	4294967041 (0xFFFFFF01)	After a route with this attribute is received, it can be advertised within the local AS, not out of the local AS.
No_Advertise	4294967042 (0xFFFFFF02)	After a route with this attribute is received, it cannot be advertised to the other BGP peers.
No_Export_Subconfed	4294967043 (0xFFFFFF03)	After a route with this attribute is received, it cannot be advertised out of the local AS or to the other sub-ASs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

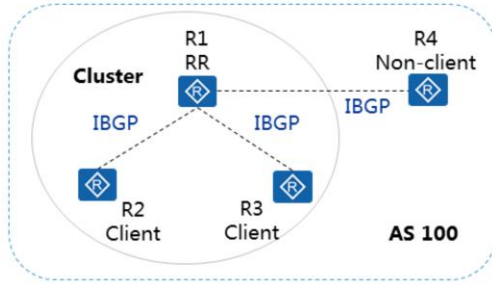
HUAWEI

- A community attribute is a set of destination addresses with the same characteristics. A community attribute is expressed in a 4-byte list. The community attribute on a device is in the format of aa:nn or identified by a community number.
  - aa:nn: The values of aa and nn are integers ranging from 0 to 65535. You can set a value as desired. The aa value identifies an AS number, and the nn value identifies the ID of a community attribute defined by an administrator. For example, for a route from AS 100, if the community attribute ID defined by an administrator is 1, the community attribute of the route is 100:1.
  - Community number: is an integer ranging from 0 to 4294967295. As defined in RFC 1997, the ranges of 0 (0x00000000) through 65535 (0x0000FFFF) and 4294901760 (0xFFFF0000) through 4294967295 (0xFFFFFFFF) are reserved.
- The community attribute simplifies the application of routing policies and maintenance and management. A community can be used to enable a group of BGP devices in multiple ASs to share the same policy. A community is a route attribute. It is transmitted between BGP peers and is not restricted by the ASs. Before advertising a route with a community attribute to the other peers, a BGP device can change the original community attribute of the route.

- Well-known community attributes
  - Internet: By default, all routes belong to the Internet community. Routes with this attribute can be advertised to all BGP peers.
  - No_Advertise: After a route with this attribute is received, it cannot be advertised to the other BGP peers.
  - No_Export: After a route with this attribute is received, it cannot be advertised to the other ASs, except for the local AS. If a confederation is used, a route with this attribute cannot be advertised outside the confederation but can be advertised to the other sub-ASs in the confederation.
  - No_Export_Subconfed: After a route with this attribute is received, it cannot be advertised outside the local AS or to other sub-ASs in the confederation.

Route Reflector

- Route reflector (RR)
  - Routes learned from IBGP peers can be advertised to specific IBGP peers. This prevents IBGP peer relationships from being fully meshed and reduces the number of IBGP sessions to be established.
  - Routers function as RRs and clients.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

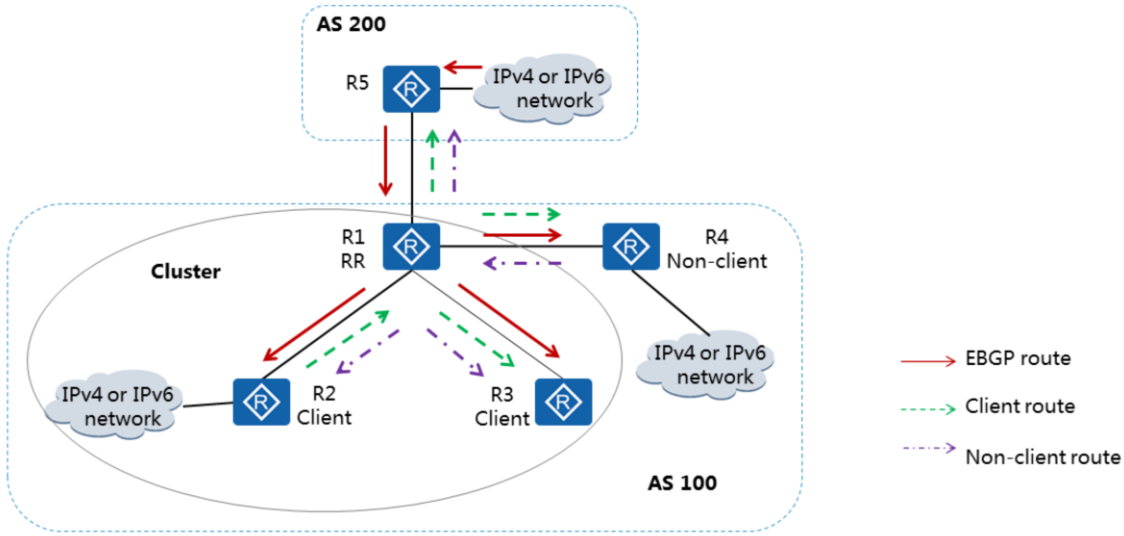
HUAWEI

- To implement connectivity between IBGP peers, establish a full-mesh relationship between IBGP peers. If there are n routers in an AS, the number of IBGP connections to be established is n x (n – 1)/2. When there are a large number of IBGP peers, network resources and CPU resources are consumed. This problem can be solved using route reflection.

- In an AS, one or two routers function as RRs, and the other routers function as clients. An IBGP connection is created between the client and each RR. The RR and its clients form a cluster. The RR reflects route information between clients, and no BGP connection needs to be established between clients.

- Concept of Route Reflector
  - Route reflector (RR): reflects routes learned from an IBGP peer to the other IBGP peers.
  - Client: is an IBGP device that establishes a reflection peer relationship with an RR. The client in an AS is merely directly connected to the RR.
  - Non-client: is an IBGP device that is neither an RR nor a client. Full-mesh connections must be established between non-clients and RRs in an AS and between non-clients.
  - Originator: is a device that originates a route in an AS. The Originator_ID attribute is used to prevent routing loops in a cluster.
  - Cluster: is a collection of RRs and their clients. The Cluster_List attribute is used to prevent routing loops between clusters.

Route Reflector - Reflection Rules



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An RR uses the following rules to advertise learned routes to IBGP peers:
  - Routes learned from EBGP peers are advertised to all non-clients and clients.
  - Routes learned from non-client IBGP peers are advertised to all clients of the RR.
  - A route learned from a client is advertised to all non-clients and the other clients of the RR (except the client that advertises the route).
- The RR is easy to configure. You only need to configure the router that functions as a reflector. The client does not need to know that itself is a client by configuration.
- On some networks, full-mesh connections have been established between clients of an RR, and they can directly exchange routing information. In this case, route reflection between the clients is unnecessary and wastes bandwidth resources. The VRP supports the undo reflect between-clients command to disable the RR from reflecting routes received from a client to the other clients.

Route Reflector - Anti-Loop Mechanism

- Anti-loop mechanism of the route reflector
  - Originator_ID attribute
    - This attribute is optional non-transitive.
    - It is used to prevent loops within a cluster.
    - It is generated by an RR and carries the router ID of the device that advertises the route in the local AS.
  - Cluster_List attribute
    - This attribute is optional non-transitive.
    - It is used to prevent loops between clusters.
    - It is generated by an RR and records the clusters through which a reflected route passes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

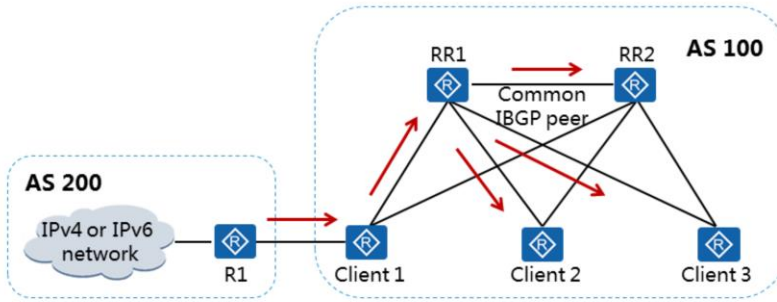
HUAWEI

- The Originator_ID attribute is generated by an RR and carries a router ID to identify the originator of the route to prevent routing loops in a cluster.
  - When a route is reflected by an RR for the first time, the RR adds the Originator_ID attribute to the route to identify the initiating device of the route. If the Originator_ID attribute is already contained in a route, the RR does not create another Originator_ID attribute.
  - When receiving the route carrying the Originator_ID attribute, the device compares the received router ID with the local one. If the two match, the device is the originator and does not accept the route.

- The RR and its clients form a cluster. Within an AS, each RR uses a unique cluster ID.
  - To prevent routing loops between clusters, an RR uses the Cluster_List attribute to record the cluster IDs of all clusters through which the route passes.
  - When the RR reflects routes between its clients or between clients and non-clients, the RR adds the local Cluster_ID to the top of the Cluster_List. If the Cluster_List is empty, the RR creates one.
  - When the RR receives an updated route, the RR matches the local Cluster_ID against the Cluster_List. If a match is found, the route is discarded. If no match is found, the RR adds the local Cluster_ID to the Cluster_List and then reflects the updated route.

Route Reflector - Backup RR

- Backup RR
  - RRs in the same cluster share a Cluster_ID.
  - Cluster_List ensures that no routing loop occurs between RRs in the same AS.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

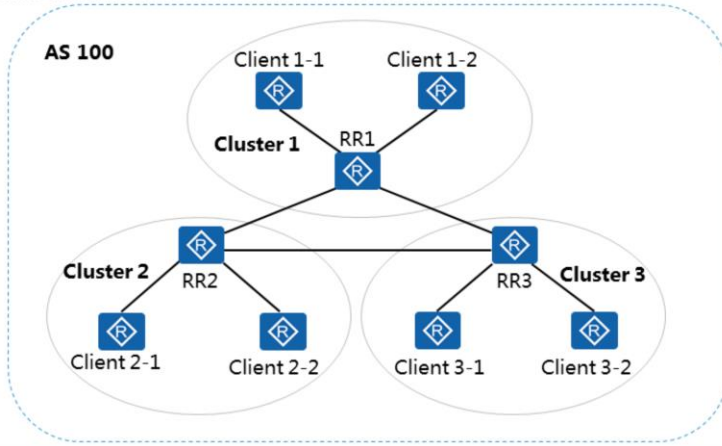
- The backup RR function is used to solve a single point of failure (SPOF).

- Backup RR
  - On the VRP, run the reflector cluster-id command to set the same Cluster_ID for all RRs in a cluster.
  - In a redundant environment, a client receives multiple routes with the same destination after the routes are reflected by different RRs. In this situation, the client selects the optimal route based on a BGP route selection policy.
  - Cluster_List ensures that no routing loop occurs between RRs in the same AS.

- Topology Description
  - After receiving an updated route (10.0.0.0/24) advertised by an EBGP peer, Client 1 advertises the route to RR1 and RR2 using IBGP.
  - After receiving the updated route, RR1 reflects it to the other clients (Client 2 and Client 3) and a non-client (RR2), and adds the local Cluster_ID to the top of the Cluster_List.
  - After receiving the reflected route, RR2 checks the Cluster_List and finds that its Cluster_ID is included in the Cluster_List. Therefore, RR2 discards the updated route and does not reflect it to its clients.

Route Reflector - RRs of the Same Level

- RRs of the same level
  - An AS may consist of multiple clusters, and RRs establish IBGP peer relationships with one another.



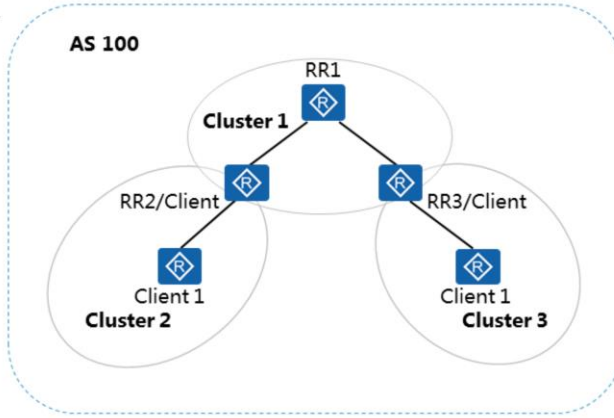
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A backbone network is divided into multiple reflection clusters. Each RR is a non-client of the other RRs in the other clusters, and all RRs establish full-mesh connections. Each client establishes IBGP connections with the RRs only in a local cluster. In this way, all BGP routers in the AS receive the reflected route information.

Route Reflector - Hierarchical RR

- Hierarchical RR
  - An RR at a lower network layer is configured as a client of another RR at a higher network layer.



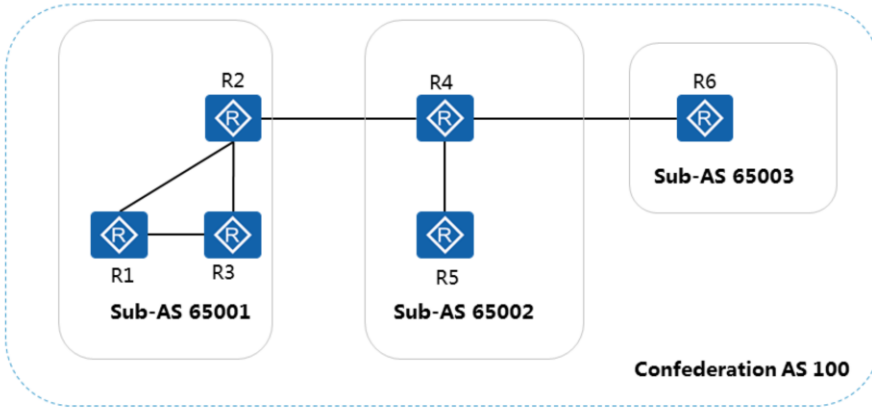
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A level-1 RR (RR-1) is deployed in Cluster1. RRs (RR-2 and RR-3) in Cluster 2 and Cluster 3 function as RR-1's clients.

BGP Confederation

- An AS is divided into several sub-ASs. Each sub-AS establishes fully connected IBGP peer relationships. An EBGP connection is established between sub-ASs.



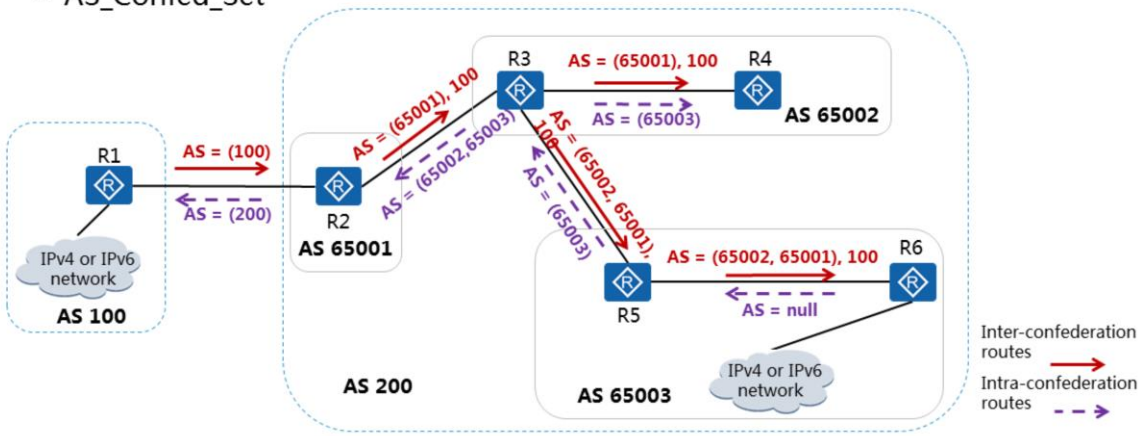
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Confederation
  - An AS in a confederation is divided into several sub-ASs. An AS is divided into several sub-ASs. Full-mesh IBGP connections are established within each sub-AS. An EBGP connection is established between each pair of sub-ASs. However, the ASs outside a confederation consider the confederation as an AS.
  - After a confederation is configured, the original AS number is used as the confederation ID of each router.
  - Original IBGP attributes include the Local_Pref, MED, and Next_Hop attributes. The confederation-related attributes are automatically deleted when routes are sent out of the confederation. That is, an administrator does not need to configure information, such as a sub-AS number, at the egress of the confederation.

BGP Confederation - Anti-Loop Mechanism

- Anti-loop mechanism of BGP confederations
  - AS_Confed_Sequence
  - AS_Confed_Set



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The AS-Path attribute is well-known mandatory and consists of AS numbers. There are four AS-Path types:
  - AS_Set: consists of a list of unordered AS numbers, which are contained in an Update message. When routes are summarized, you can use a proper policy to use the AS_Set attribute to prevent path information loss.
  - AS_Sequence: consists of a series of ordered AS numbers, which are contained in an Update message. Generally, the AS-Path type is set to AS_Sequence.
  - AS_Confed_Sequence: consists of a list of unordered sub-AS numbers in a local confederation, which are contained in an Update message. The AS_Confed_Set attribute is used in the same way as the AS_Set attribute and is transmitted only in a local confederation.
  - AS_Confed_Set: consists of a list of unordered sub-AS numbers in a local confederation, which are contained in an Update message. The AS_Confed_Set attribute is used in the same way as the AS_Set attribute and is transmitted only in a local confederation

- The sub-AS numbers in a confederation are invisible to other non-confederation ASs. The sub-AS numbers are stripped off from the routes before the routes are advertised from a confederation to the other non-confederation ASs.

Comparison Between BGP RR and Confederation

RR	Confederation
The existing network topology does not need to be modified. Compatibility is good.	The logical topology needs to be modified.
The configuration is simple. A client does not know that itself is a client.	All devices must be reconfigured and support the confederation function.
Full-mesh connections must be established between clusters.	The sub-ASs in a confederation establish special EBGP connections and do not need to be fully meshed.
It is used widely on large-scale networks.	There are a few applications.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

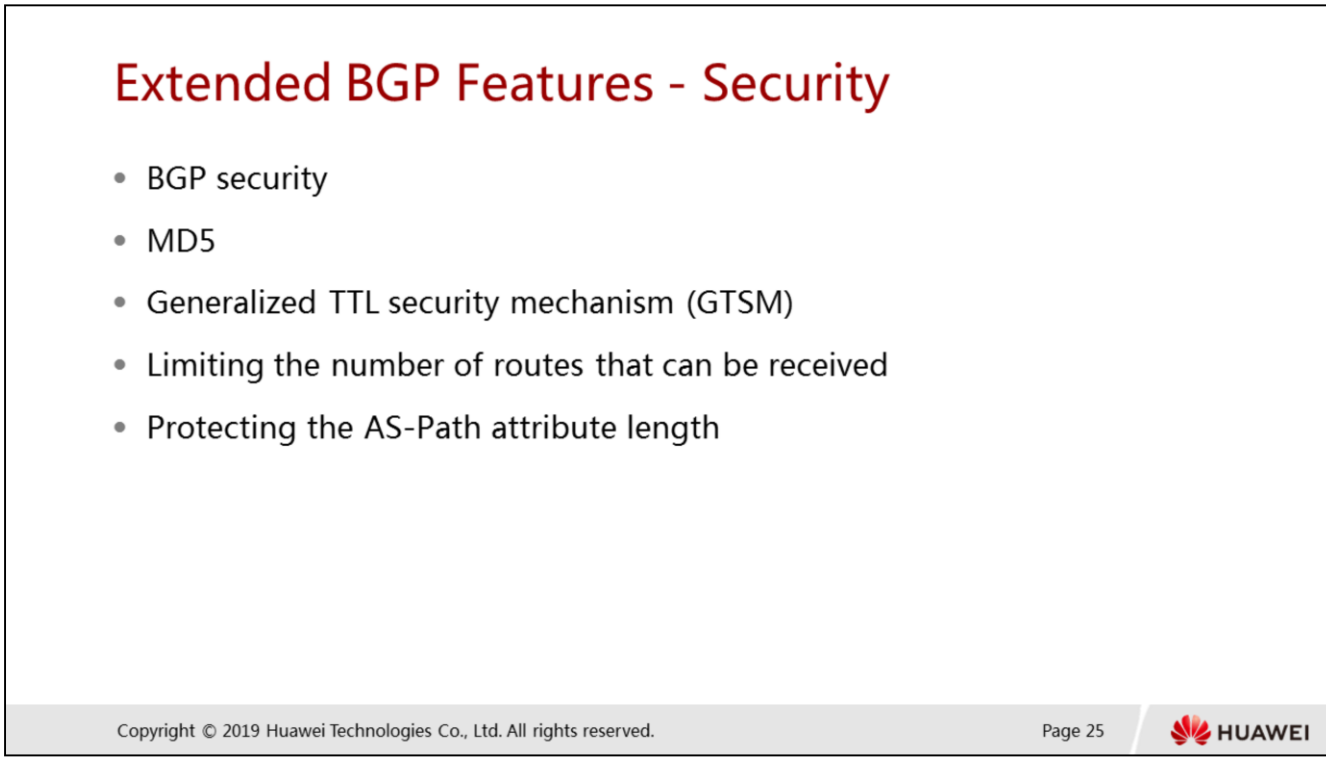
- Comparison between RR and confederation
  - A confederation needs to be re-partitioned, which results in a great number of modifications on the live network.
  - You need to only configure an RR, and no action is required for clients. The confederation function, however, must be configured on all routers in a confederation.
  - Full-meshed IBGP connections must be established between RRs.
  - RRs are widely used. There are a few applications for confederation scenarios.



Contents

- Technologies Regarding Mass BGP Routes
- **Extended BGP Features**
- Enhanced BGP Features
- Internet Design Fundamentals
- BGP Troubleshooting

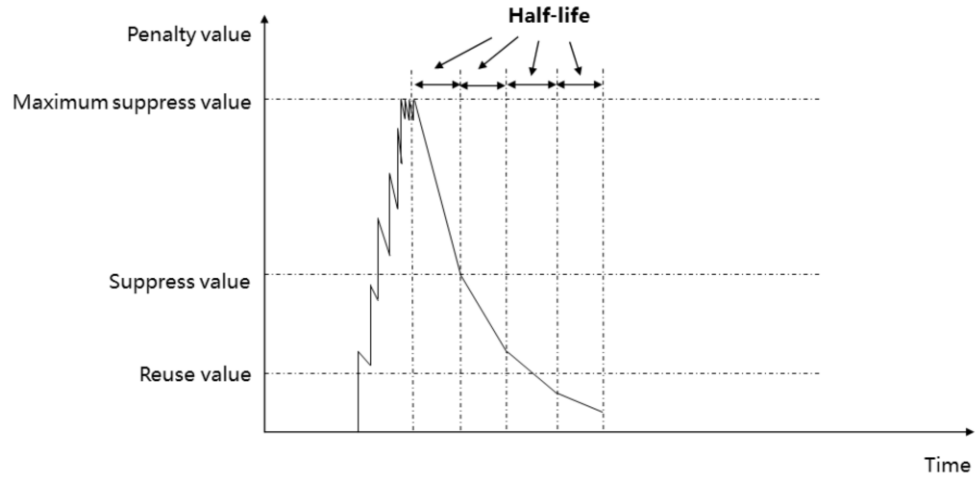
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- BGP security features:
  - MD5: BGP uses TCP as a transport layer protocol. To improve BGP security, perform MD5 authentication when establishing a TCP connection. MD5 authentication of BGP, however, does not authenticate BGP messages. In MD5 authentication, a password merely needs to be set for a TCP connection, and TCP completes the authentication process. If authentication fails, no TCP connection is established.
  - Generalized TTL Security Mechanism (GTSM): checks whether a TTL value in the IP message header is within a defined range, which helps protect services at the IP layer and enhance system security. After the GTSM is enabled for BGP, an interface board checks the TTL value carried in each BGP message. Based on actual networking requirements, a GTSM policy can be configured to permit or discard messages in which TTL values are out of a specified range. When the default GTSM action is set to "discard", you can select a proper TTL value range based on the network topology. The messages that do not match the TTL value range are discarded directly by an interface board. This prevents "valid" BGP messages simulated by network attackers from consuming CPU resources. This function is mutually exclusive with the EBGP multi-hop function.
  - Limits the number of routes that can be received, which prevents resource exhaustion attacks.
- Protects the AS-Path attribute length. The AS-Path attribute length is limited on the inbound and outbound interfaces. The messages whose AS-Path attribute lengths exceed a specified limit are discarded.

Extended BGP Features - Route Dampening

- Route dampening is used to solve the problem of unstable routes.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Route dampening is used to solve the problem of unstable routes. In most cases, BGP applies to complex networks on which routes change frequently. To minimize the adverse impact caused by continuous route flapping, BGP uses route dampening to suppress unstable routes.
- In BGP dampening, a penalty value measures the stability of a route. A higher penalty value indicates a more unstable route. Each time a route flaps (the route alternates between active and inactive), BGP adds a penalty value (1000) for the route. If the penalty value exceeds a specified suppression value, the route is suppressed and not added to the routing table or advertised to the other BGP peers.
- If the penalty value of a route reaches a specified maximum suppression value, the penalty value does not increase any more. This ensures that the penalty value is not accumulated to a high value, at which a route remains in the suppressed state, in case the route is changed a dozen of times within a short period.
- The penalty value of the suppressed route decreases by half at an interval. This interval is called half-life. When the penalty value decreases to a specified reuse value, the route becomes available and is added to the routing table again. In addition, the route is advertised to the other BGP peers. The penalty value, suppression value, and half-life values can be manually set.
- Route dampening applies only to EBGP routes. IBGP routes cannot be dampened because local AS routes are configured as IBGP routes. Intra-AS route information in the forwarding tables must be consistent as much as possible. If route dampening takes effect on IBGP routes and dampening parameter settings differ on devices, information in the forwarding tables become inconsistent.



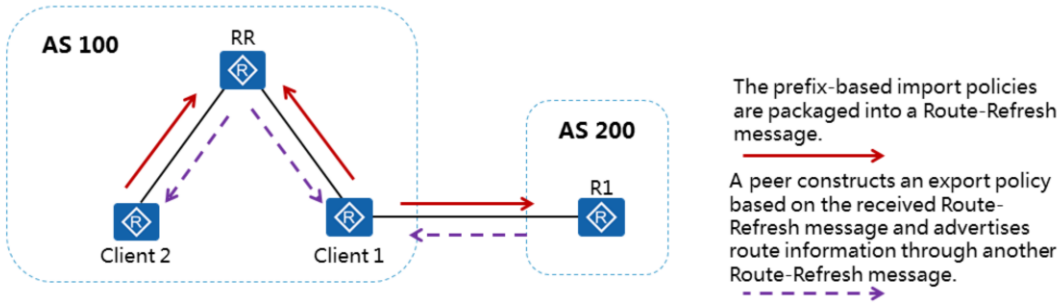
Contents

- Technologies Regarding Mass BGP Routes
- Extended BGP Features
- **Enhanced BGP Features**
- Internet Design Fundamentals
- BGP Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Enhanced BGP Features - BGP ORF

- BGP prefix-based outbound route filtering (ORF):
  - An export policy of a peer is created based on a local import policy to advertise BGP routes on demand.
  - BGP ORFs include the prefix-based ORF and VPN ORF.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

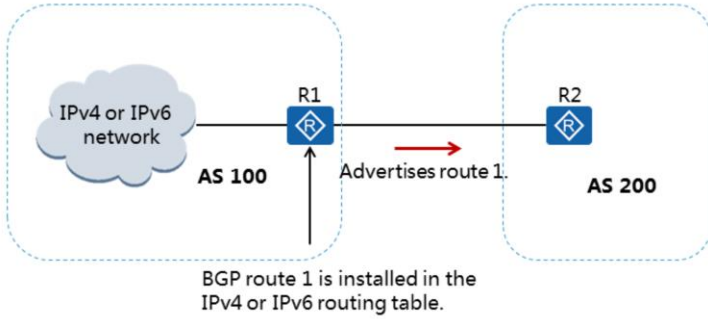
HUAWEI

- RFC 5291 and RFC 5292 define the prefix-based ORF capability of BGP. This capability can be used to send a prefix-based import policy configured on a local device to BGP peers through Route-Refresh messages. Each BGP peer constructs an export policy based on the received policy and filters out routes before advertising the routes. This prevents a local device from receiving a large number of unwanted routes and reduces CPU usage of the local device, BGP peer configuration workload, and link bandwidth usage.

- Topology description
  - In a directly connected EBGP peer relationship, after Client 1 and R1 negotiate the prefix-based ORF capability, Client 1 encapsulates a locally configured prefix-based import policy into a Route-Refresh message and sends the message to R1. Upon receipt of the message, R1 constructs an export policy and sends a Route-Refresh message to Client 1. Client 1 accepts only the routes that it needs. R1 does not need to maintain a routing policy, which reduces configuration workload.
  - Clients 1 and 2 are RR's clients, Clients 1 and 2 negotiate with the RR on the prefix-based ORF capability. Clients 1 and 2 encapsulate locally configured prefix-based import policies into Route-Refresh messages and send them to the RR. The RR constructs an export policy based on the received prefix-based import policies sent by Clients 1 and 2 and reflects the routes to Clients 1 and 2 through Route-Refresh messages. Clients 1 and 2 accept only the required routes. The RR does not need to maintain a routing policy, which reduces configuration workload.

Enhanced BGP Features - Active-Route-Advertise

- Active-Route-Advertise
  - A BGP route can be advertised to a peer only after the BGP route is successfully installed into the IPv4 or IPv6 routing table.



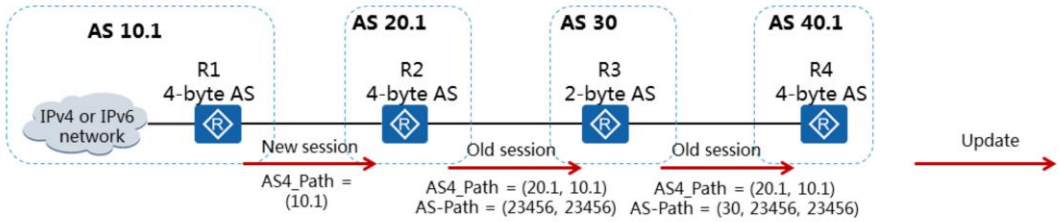
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Active-Route-Advertise
  - By default, routes can be advertised to peers only when they are preferred BGP routes. After the Active-Route-Advertise feature is configured, a device only advertises preferred BGP routes that are active on the route management plane.
  - This function is mutually exclusive with the routing-table rib-only command (used to prevent BGP routes from being installed into an IP routing table).

Enhanced BGP Features - 4-Byte AS Number

- 4-byte AS number
  - An AS number is lengthened from 2 bytes to 4 bytes.
- Protocol extension
  - A new type of open capability code is defined for BGP connection capability negotiation.
  - Two new optional transitive attributes: AS4_Path and AS4_Aggregator
  - The AS_Trans attribute (reserved value: 23456) is used to represent a non-mappable 4-byte AS number as a 2-byte AS number.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

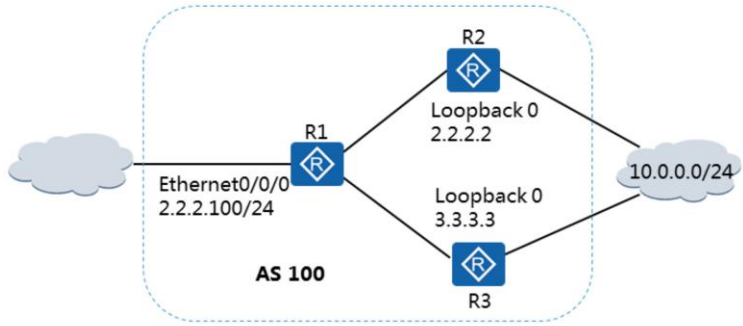
- Roles defined based on the support for the 4-byte AS number function
  - New speaker: a peer that supports 4-byte AS numbers
  - Old speaker: a peer that does not support 4-byte AS numbers
  - New session: a BGP connection established between new speakers
  - Old session: a BGP connection established between a new speaker and an old speaker or between old speakers

- Protocol extension
  - Two new optional transitive attributes, that is, AS4_Path (attribute code: 0x11) and AS4_Aggregator (attribute code: 0x12), are defined to transmit 4-byte AS numbers over an old session.
  - If a new speaker establishes a peer relationship with an old speaker, the AS_Trans (reserved value: 23456) attribute is defined to represent a non-mappable 4-byte AS number as 2-byte AS number.
  - A new AS number can be in any of the following formats:
    - Splain: a decimal number.
    - asdot+: in the format of 2-byte-value.2-byte-value. Therefore, a 2-byte ASN123 can be written as 0.123, and ASN65536 as 1.0. The maximum value is 65535.65535.
    - asdot: An old 2-byte AS number remains its format, and a new 4-byte AS number is in the asdot+ format. (A 2-byte AS number ranges from 1 to 65535; a 4-byte AS number ranges from 1.0 to 65535.65535.)
    - Huawei devices support the asdot format.

- Topology description:
  - R2 receives a route containing a 4-byte AS number from R1. The AS number is 10.1.
  - If R2 establishes a peer relationship with R3, R3 needs to consider that R2's AS number is represented by the AS_Trans attribute.
  - Before R2 advertises a route to R3, R2 records the AS_Trans value in the AS-Path attribute and adds 10.1 and its AS number 20.1 to the AS4_Path attribute in a desired order.
  - R3 does not process the unknown AS4_Path attribute and retains it. R3 advertises the route to R4 based on BGP rules.
  - R3 considers that the AS number of R4 is also represented by the AS_Trans attribute.
  - In this way, when R4 receives the route from R3, R4 replaces the AS_Trans value in the AS-Path with AS numbers recorded in the AS4_Path attribute and restores the AS-Path attribute to 30, 20.1, and 10.1.

Enhanced BGP Features - Policy-based Next-Hop Recursion

- Policy-based next-hop recursion
  - A route-policy is used to limit recursive routes. If a recursive route fails to match against a route-policy, route recursion fails.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Policy-based next-hop recursion
  - BGP performs route recursion for routes that contain indirect next hops. If recursive routes are not filtered out, traffic may be recursively forwarded to an incorrect forwarding path. The policy-based next-hop recursion is to limit the recursive routes using a route-policy. If a recursive route fails to match against a route-policy, route recursion fails.

- Topology Description
  - R1, R2, and R3 establish IBGP peer relationships using loopback addresses. R1 receives BGP routes with the prefix 10.0.0.0/24 from R2 and R3. The original next hop of the BGP route advertised by R2 is 2.2.2.2. In addition, the IP address and mask of Ethernet 0/0/0 on R1 are 2.2.2.100/24.
  - When R2 is running properly, R1 receives the route with the prefix 10.0.0.0/24 from R2 and recurses this route to an IGP route 2.2.2.2/32. When an IGP becomes faulty on R2, the IGP route of 2.2.2.2/32 is withdrawn. As a result, next-hop recursion is triggered again. On R1, the original next-hop 2.2.2.2 is used to perform longest match recursion in the IP routing table. The original route recurses to the route 2.2.2.0/24. However, a user expects that when the route 2.2.2.2 is unavailable, the route 3.3.3.3 can be preferentially selected. The route-withdrawal-triggered re-recursion actually stems from BGP convergence. As a result, a transient black hole is generated.
  - Configure a route-policy for next-hop recursion to filter out recursive routes based on mask lengths of the routes mapped to the original next hops. You can configure a next-hop recursion policy so that the original next hop 2.2.2.2 can only depend on the IGP route 2.2.2.2/32.



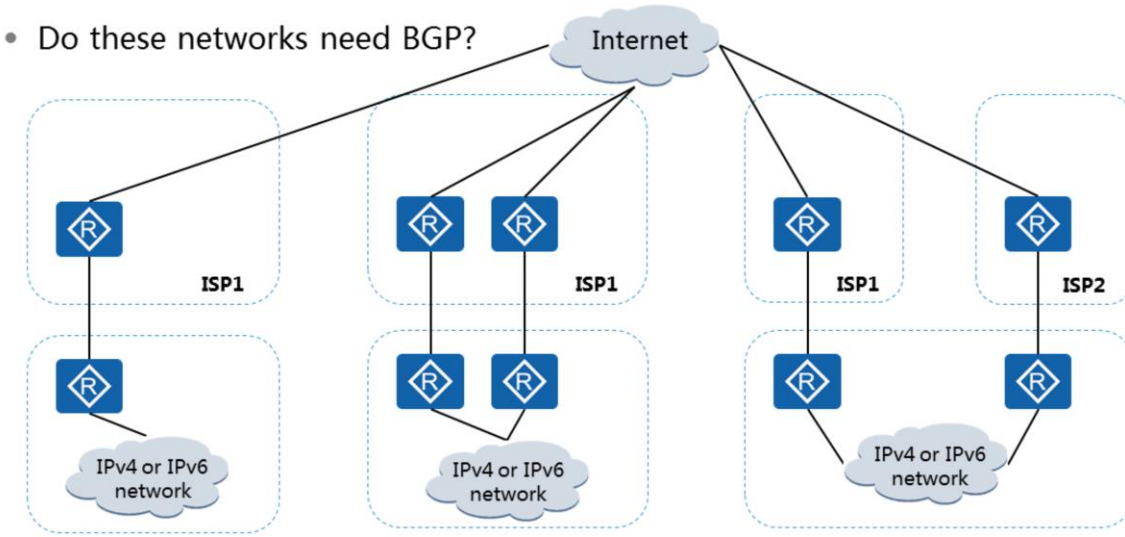
Contents

- Technologies Regarding Mass BGP Routes
- Extended BGP Features
- Enhanced BGP Features
- **Internet Design Fundamentals**
- BGP Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Which Types of Networks Need BGP?

- Do these networks need BGP?



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Common enterprise network topology types:
  - Single-homed AS (Each egress connects to a single ISP.)
  - Multi-homed single-AS (Multiple egresses connect only to one ISP.)
  - Multi-homed multi-AS (Multiple egresses connect to multiple ISPs.)
- Single-homed AS: A single egress connects only to a single ISP.
  - In this case, you do not need to configure BGP. You can add a default route to a user edge device that advertises the default route to the user AS.
- Multi-homed single-AS: Redundancy is implemented on links and network devices. In this situation, private AS numbers are used on user networks.
  - If two links are working in active/standby mode, BGP is not required. The two egresses advertise the default routes with different cost values to devices in the local AS. (If OSPF is used as an IGP, the cost of an external route is calculated in E2 mode, and the external cost is only used.)
  - If two routers are working in load balancing mode:
    - Method 1: The two routers advertise the default routes whose cost type is E1 to the local AS (OSPF used as an IGP) so that other routers in the AS can select the nearest egress router to reach the external network. In this case, BGP is not required. However, when the physical distance between the two egresses is long and the delay time is sensitive, BGP can be used to obtain more specific routing entries.
    - Method 2: A BGP connection is established between a device and an ISP device. The device receives more specific routing entries from BGP and uses a route-policy to map each particular destination IP address to a specific egress route.

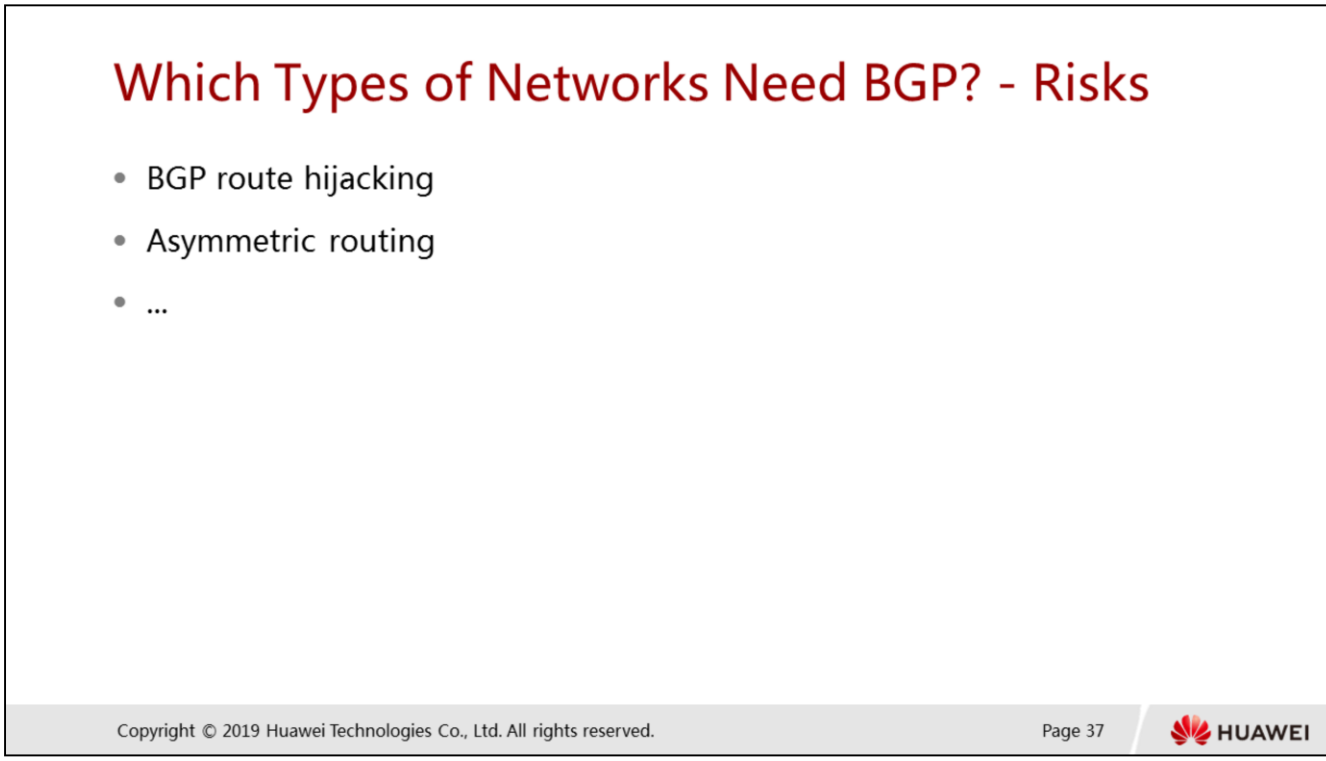
- Multi-homed multi-AS: Redundancy is implemented on links and network devices, and ISP redundancy is also implemented.
    - For such an AS, determine whether the address space is independent of ISPs and whether public AS numbers are available.
    - Ideally, three deployment methods can be used when a user network has the address space and public AS numbers independent of the ISPs.
        - Method 1: In active/standby mode, the egress routers advertise the default routes with different costs to internal devices.
        - Method 2: In load balancing mode, the egress routers advertise the default routes to the internal network. The IGP cost calculation mechanism is only used. The IGP determines which egress router is selected.
        - Method 3: Deploy BGP.
    - Sign a contract with an ISP. Based on service traffic characteristics of an enterprise, use various routing policy tools. If necessary, advertise default routes.
    - Efficiently control the incoming and outgoing traffic of the enterprise.
    - Generally, BGP is deployed on a network that belongs to multiple ASs, because methods 1 and 2 are not suitable for route control. Nevertheless, you have to make a considerable tradeoff between benefits and increasing routing complexity.

Which Types of Networks Need BGP? - Some Considerations

- A large number of routes exist and cannot be summarized.
- Cross-operating entities (Communication is not convenient and information is confidential.)
- Multiple IGPs run in a routing domain.
- An IGP cannot provide tools to execute the desired policies.
- A network belongs to multiple ISPs.
- ...

- BGP can be deployed if the preceding conditions are met and the project implementation and network planning are fully considered.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

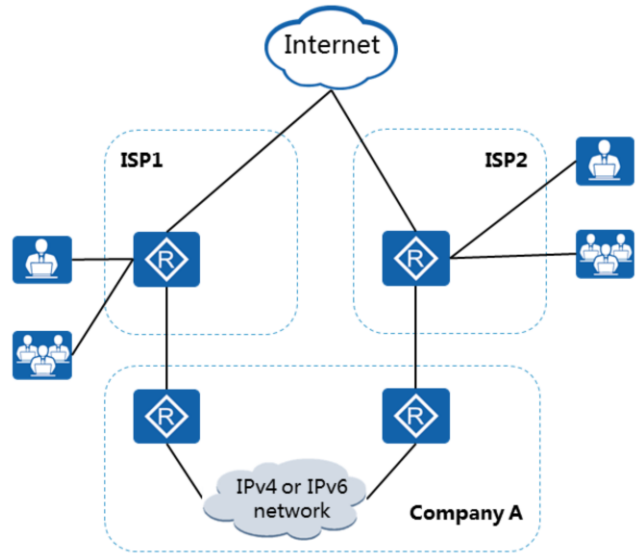


- BGP route hijacking
  - Cause: BGP provides some simple security authentication functions. If two ASs have established BGP connections, the two ASs unconditionally trust information sent by each other, including the IP address range claimed by the peer AS.
  - Risk: If a device in an AS unconditionally trust the Update message sent by a BGP peer in the other AS. A malicious AS may advertise a nonexistent IP network segment, the BGP attributes, such as AS-Path, are tampered with so that the other AS considers the path as the shortest one to the destination network segment. In this case, the malicious AS can intercept data traffic.

- Asymmetric routing
  - Cause: Inappropriate attributes or route summarization causes insufficient route precision. As a result, the paths of outbound and inbound traffic are different.
  - Risk: First, asymmetric traffic makes the traffic model of the Internet difficult to predict. Consequently, the network benchmarking, capacity planning, fault detection, and troubleshooting become difficult. Second, asymmetric traffic causes a link usage imbalance. The bandwidth of some links is saturated, but the bandwidth of the other links cannot be effectively used. Third, asymmetric traffic causes a great delay inconsistency between the outgoing and incoming traffic. This delay variation (jitter) may compromise some delay-sensitive applications (such as voice and live video).

Precautions for BGP Network Deployment and Maintenance

- A multi-ISP-homing AS is used as an example:
  - Company A connects to two ASs. A route advertised by ISP1 is incorrectly advertised to ISP2 because of a configuration error.
  - Filter out and select routes advertised by the ISPs based on service.
  - Pay attention to operations performed by the peer AS to prevent the local network from being adversely affected by mis-operations in the peer AS.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Internet Design Fundamentals - Optimizing BGP Capabilities

- Optimizing BGP capabilities
  - Establishing a peer session
  - Updating the origin of routes
  - Optimizing a route-policy
  - Filtering routes and controlling attributes
  - Route summarization

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Internet Design Fundamentals - Improving BGP Availability

- Improving BGP availability
  - Redundancy
  - Symmetric traffic
  - Load balancing

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Internet Design Fundamentals - Controlling Intra-AS Routes

- Controlling intra-AS routes
  - Interaction between non-BGP routes and BGP routes
  - Default route
  - Policy-based routing

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Interaction between non-BGP routes and BGP routes
  - Generally, an IGP and BGP import routes. Proper filter policies must be used to enable proper routes to be imported between an IGP and BGP.
- Default route control
  - A policy can be configured to deliver default routes based on specific conditions.
- Policy-based routing
  - Policy-based routing is used to optimize traffic paths.

Internet Design Fundamentals - Limiting Large ASs

- Limiting large ASs
  - BGP Dynamic Update Peer-Groups
  - Route Reflector
  - Confederation

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Internet Design Fundamentals - Designing the Stable Internet

- Designing the stable Internet
  - Reducing the generation of unstable routes
  - Improving BGP stability

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



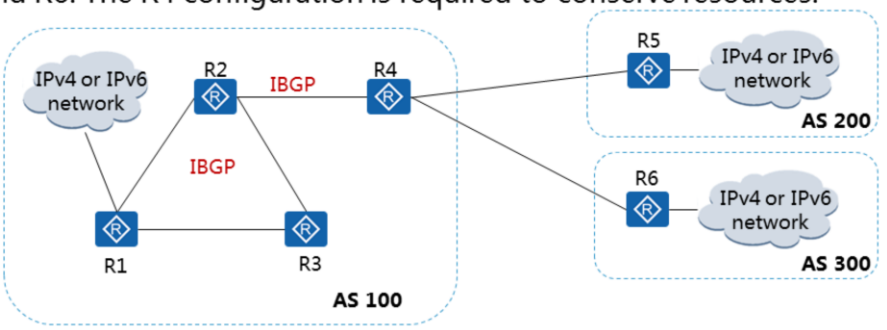
Contents

- Technologies Regarding Mass BGP Routes
- Extended BGP Features
- Enhanced BGP Features
- Internet Design Fundamentals
- **Advanced BGP Configurations**
- BGP Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuring a Peer Group (1)

- Configuration requirements:
  - An IGP protocol has been deployed, and all devices are in the OSPF backbone area.
  - Loopback addresses are used to establish EBGP peer relationships between R4, R5, and R6. The R4 configuration is required to conserve resources.

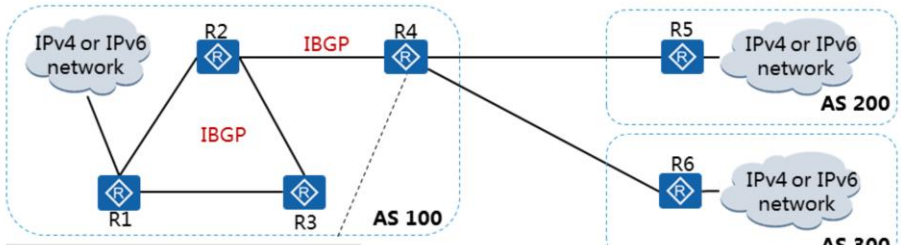


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Case description
  - In this case, device interconnection rules are as follows:
    - If the RX and RY are interconnected, the interconnection addresses are XY.1.1.X and XY.1.1.Y, respectively, and the mask length is 24 bits.
  - OSPFv2 and OSPFv3 are running properly, and the device interconnection addresses and loopback addresses have been advertised to OSPFv2 or OSPFv3.
- Case analysis
  - EBGP peers use loopback interfaces to establish peer relationships.

Configuring a Peer Group (2)




```
bgp 100
group 100 external
 peer 100 ebgp-max-hop 2
 peer 100 connect-interface LoopBack0
 peer 5.5.5.5 as-number 200
 peer 5.5.5.5 group 100
 peer 6.6.6.6 as-number 300
 peer 6.6.6.6 group 100
 #
 ipv4-family unicast
  undo synchronization
  peer 100 enable
  peer 5.5.5.5 enable
  peer 5.5.5.5 group 100
  peer 6.6.6.6 enable
  peer 6.6.6.6 group 100
```

```
[R4]display bgp peer
BGP local router ID : 4.4.4.4
 Local AS number : 100
 Total number of peers : 2                    Peers in established state : 2
  Peer        V    AS  MsgRcvd  MsgSent  OutQ  Up/Down       State PrefRcv
  5.5.5.5     4   200        3        3     0  00:01:06 Established       0
  6.6.6.6     4   300        2        2     0  00:00:50 Established       0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

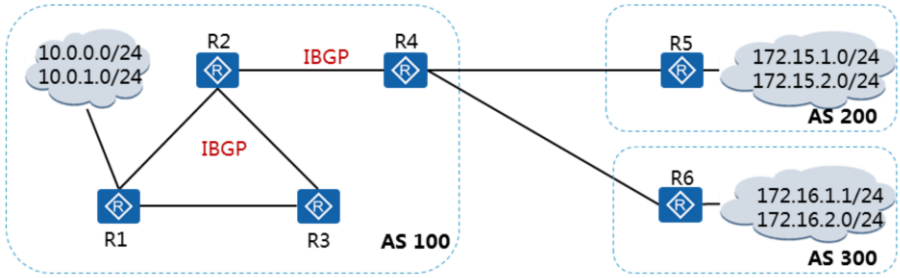
- Command description
  - The peer as-number command sets a peer AS number for a specified peer or peer group.
  - The peer connect-interface command specifies the name of the source interface used to send BGP messages and a source address used to initiate a connection.
  - The peer next-hop-local command enables a device to set the next-hop IP address to a local IP address before the device advertises routes to an IBGP peer or peer group.
  - The group command creates a peer group.
- Usage guidelines
  - The preceding commands are run in the BGP process view.
- Parameter description
  - peer ipv4-address as-number as-number
    - ip-address: specifies the IPv4 address of a peer.
    - as-number: specifies the peer AS number of a peer.
  - peer ipv4-address connect-interface interface-type interface-number [ipv4-source-address]
    - ip-address: specifies the IPv4 address of a peer.
    - interface-type interface-number: specifies the type and number of an interface.
    - ipv4-source-address: specifies an IPv4 source address used to establish a connection.

- peer ipv4-address next-hop-local
  - ip-address: specifies the IPv4 address of a peer.
- group group-name [external | internal]
  - group-name: specifies the name of a peer group.
  - external: creates an EBGP peer group.
  - internal: creates an IBGP peer group.

- Precautions
  - When a loopback interface is used as the source interface of BGP messages, note the following points:
  - Ensure that the loopback interface address of a BGP peer is reachable.
  - For EBGP connections, run the peer ebgp-max-hop command to allow EBGP to establish peer relationships over indirect connections.
  - The peer next-hop-local and peer next-hop-invariable commands are mutually exclusive with each other.
  - In the display bgp peer command output, Rec indicates the number of route prefixes received by the local end from the peer.
  - The IPv6 configuration is similar to the IPv4 configuration.

Configuring Automatic BGP Summarization and Default Routes (1)

- Optimize paths in AS 300. The requirements are as follows:
  - Import the network segment 172.16.X.0/24 to BGP and perform automatic summarization.
  - When the route 172.16.0.0/16 exists on R4, R4 needs to advertise a default route to AS300.

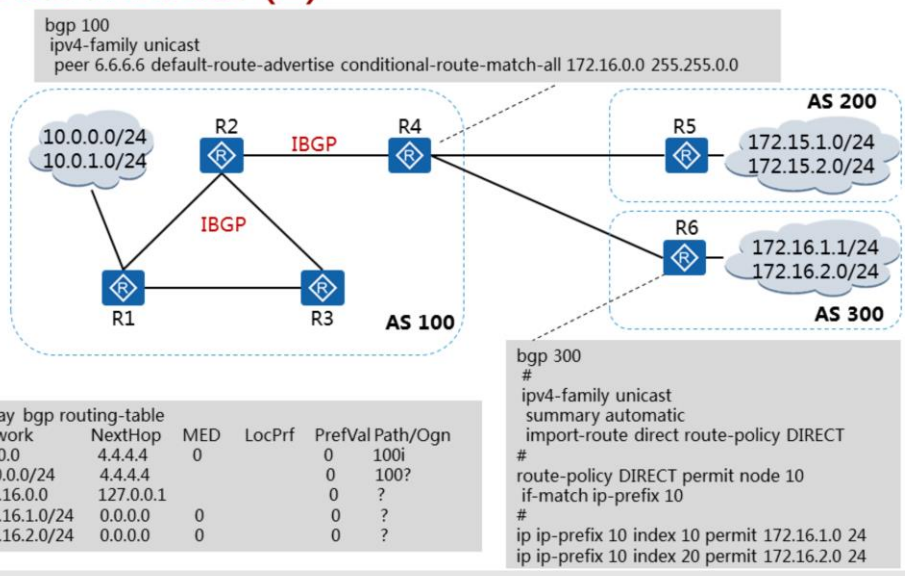


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Case description
  - This case demonstrates a requirement extension of the previous case, and the related configuration is based on the original case.
  - In requirement 2, the default route needs to be associated with the route 172.16.0.0/16. If the route 172.16.0.0/16 disappears, the default route also disappears.

Configuring Automatic BGP Summarization and Default Routes (2)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

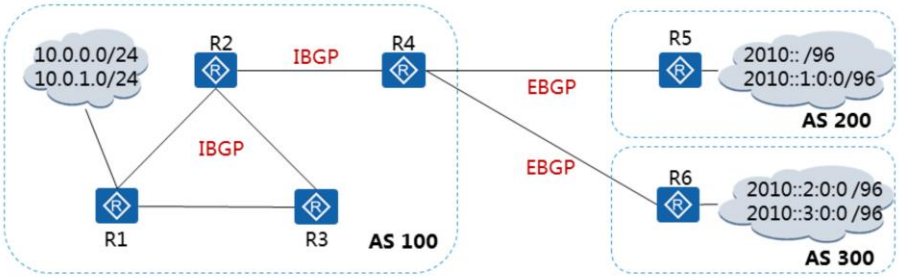
HUAWEI

- Command description
  - The **peer route-policy** command specifies a route-policy used to accept routes from or advertise routes to a peer or peer group.
  - The **peer default-route-advertise** command configures a device to advertise default routes to a peer or peer group.
- Usage guidelines
  - The **peer route-policy** command is run in the BGP view.
  - The **peer default-route-advertise** command is run in the BGP view.
- Parameter description
  - **peer** *ipv4-address* **route-policy** *route-policy-name* { **import | export** }
    - *ipv4-address*: specifies the IPv4 address of a peer.
    - *route-policy-name:* specifies the name of a route-policy.
    - **import**: applies a route-policy to routes advertised by a peer or peer group.
    - **export**: applies a route-policy to routes advertised to a peer or peer group.

- **peer** { *group-name* | *ipv4-address* } **default-route-advertise** [**route-policy** *route-policy-name*] [**conditional-route-match-all**{ *ipv4-address1 { mask1* | *mask-length1* } } &<1-4> | **conditional-route-match-any** { *ipv4-address2 { mask2* | *mask-length2* } } &<1-4>]
  - *ipv4-address*: specifies the IPv4 address of a BGP peer.
  - **route-policy** *route-policy-name*: specifies the name of a route-policy.
  - **conditional-route-match-all** *ipv4-address1* { *mask1* | *mask-length1* }: specifies an IPv4 address and a mask or a mask length of a route.
  - If all conditions are matched, a default route is sent.
  - **conditional-route-match-any** *ipv4-address2 {mask2* | *mask-length2*}: specifies an IPv4 address and a mask or a mask length of a route. If any condition is met, a default route is advertised.
- Experiment symptom
  - You can run the **display ip routing-table** command to view information in the routing table.

Configuring Manual BGP Summarization (1)

- Configuration requirements:
  - R1 imports a network segment 10.0.X.0/24 and marks its community as 200:200.
  - On R1, use the community 200:200 to manually summarize routes. After the summarization, the route mask is /23 and specific routes are suppressed. In addition, loop prevention must be taken into consideration.
  - On R4, summarize the routes belonging to network segment 2010::x:0:0/96 and set the subnet mask to 94 for the summary route.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Case description
  - This case demonstrates a requirement extension of the previous case, and the related configuration is based on the original case.

Configuring Manual BGP Summarization (2)

```
[R1]
bgp 100
 ipv4-family unicast
  aggregate 10.0.0.0 255.255.254.0 as-set detail-suppressed origin-policy COMM
  import-route direct route-policy DIRECT
#
route-policy DIRECT permit node 10
 if-match ip-prefix 10
 apply community 200:200
#
route-policy COMM permit node 10
 if-match community-filter 200:200
#
ip ip-prefix 10 index 10 permit 10.0.1.0 24
ip ip-prefix 10 index 20 permit 10.0.0.0 24
```

```
[R4]
bgp 100
#
 ipv6-family unicast
 aggregate 2010:: 94 detail-suppressed
 peer 2011::24:2 next-hop-local
#
```

```
[R4]dis bgp routing-table

 Total Number of Routes: 5
     Network          NextHop        MED        LocPrf    PrefVal Path/Ogn

*>   10.0.0.0/23      1.1.1.1          0                    0                ?
```

```
[R2]dis ipv6 routing-table protocol bgp
 Destination  : 2010::                    PrefixLength : 94
 NextHop      : 2011::24:4                 Preference   : 255
 Cost         : 0                          Protocol     : IBGP
 RelayNextHop : ::                         TunnelID     : 0x0
 Interface    : GigabitEthernet0/0/0       Flags        : D
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Command description:
  - The **aggregate** command creates a summary route in the BGP routing table.
- Usage guidelines
  - The **aggregate** command is run in the BGP view.
- Parameter description
  - **aggregate** *ipv4-address { mask | mask-length }* [**as-set | attribute-policy** *route-policy-name1* | **detail-suppressed | origin-policy** *route-policy-name2* | *suppress-policyroute-policy-name3*] *
    - *ipv4-address*: specifies the IPv6 address of a summary route.
    - *mask*: only uses IBGP routes in load balancing.
    - *mask-length*: specifies the network mask length of a summary route.
    - **as-set**: generates routes with the AS_Set attribute.
    - **attribute-policy** *route-policy-name1*: specifies the name of an attribute policy for a summary route.
    - **detail-suppressed**: advertises summary routes.
    - **origin-policy** *route-policy-name2*: specifies the name of a policy for generating summary routes.
    - **suppress-policy** *route-policy-name3*: specifies the name of a policy for suppressing the advertisement of specific routes.
- Precautions
  - In both manual summarization and automatic summarization, a route uses NULL0 as the out-interface is generated locally .
  - The IPv6 configuration is similar to the IPv4 configuration.
- Experiment result
  - You can run the **display ip routing-table protocol bgp** command to view the routes learned by BGP.



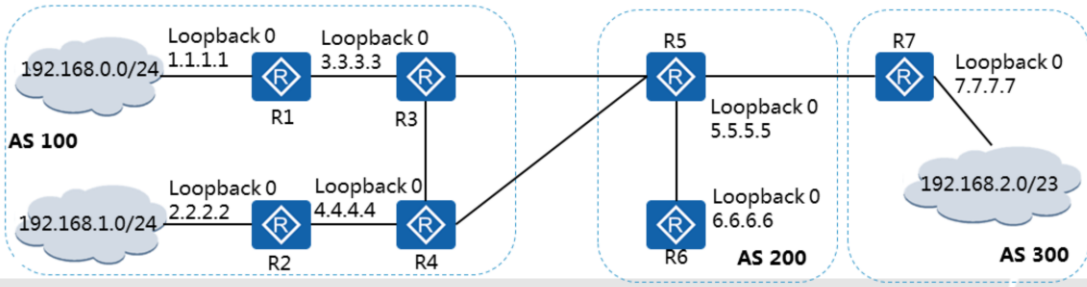
Contents

- Technologies Regarding Mass BGP Routes
- Extended BGP Features
- Enhanced BGP Features
- Internet Design Fundamentals
- Advanced BGP Configurations
- **BGP Requirement Implementation and Troubleshooting**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

BGP Requirement Implementation and Troubleshooting - Case 1 (1)

- Requirement:
  - OSPF runs in AS 100. All routers belong to area 0. IBGP peer relationships are established between directly connected routers using addresses of loopback 0 interfaces. The number of BGP connections to be established is minimized.
  - R5 and R6 use loopback 0 interfaces to establish IBGP peer relationships. R5 uses direct interfaces to establish EBGP peer relationships with R3 and R4 and uses loopback 0 interface to establish an EBGP peer relationship with R7.
  - R1 is directly connected to 192.168.0.0/24, R2 is directly connected to 192.168.1.0/24, and R7 is directly connected to 192.168.2.0/23. The three network segments must be able to communicate with one another, and the number of routes on R6 must be minimized.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The network segment between Rx and Ry (X < Y) is 10.0.xy.0/24. Rx's interface IP address is 10.0.xy.x, and Ry's interface IP address is 10.0.xy.y.
- All interface addresses have been configured.

BGP Requirement Implementation and Troubleshooting - Case 1 (2)

- Configurations of devices in AS 100:

```
[R1]
ospf 1
 area 0.0.0.0
  network 1.1.1.1 0.0.0.0
  network 10.0.12.1 0.0.0.0
#
bgp 100
 peer 2.2.2.2 as-number 100
 peer 2.2.2.2 connect-interface LoopBack0
 #
 ipv4-family unicast
 network 192.168.0.0
  peer 2.2.2.2 enable
```

```
[R4]
ospf 1
 area 0.0.0.0
  network 10.0.34.4 0.0.0.0
#
bgp 100
 peer 3.3.3.3 as-number 100
 peer 3.3.3.3 connect-interface LoopBack0
 peer 10.0.45.5 as-number 200
 #
 ipv4-family unicast
 network 192.168.1.0
  peer 3.3.3.3 enable
```

```
[R2]
ospf 1
 area 0.0.0.0
  network 2.2.2.2 0.0.0.0
  network 10.0.12.2 0.0.0.0
  network 10.0.23.2 0.0.0.0
#
bgp 100
 peer 1.1.1.1 as-number 100
 peer 1.1.1.1 connect-interface
LoopBack0
 peer 3.3.3.3 as-number 100
 peer 3.3.3.3 connect-interface
LoopBack0
 #
 ipv4-family unicast
  undo synchronization
  peer 1.1.1.1 enable
  peer 1.1.1.1 reflect-client
  peer 1.1.1.1 next-hop-local
  peer 3.3.3.3 enable
  peer 10.0.25.5 enable
#
```

```
[R3]
ospf 1
 area 0.0.0.0
  network 3.3.3.3 0.0.0.0
  network 10.0.34.3 0.0.0.0
  network 10.0.23.3 0.0.0.0
#
bgp 100
 peer 2.2.2.2 as-number 100
 peer 2.2.2.2 connect-interface
LoopBack0
 peer 4.4.4.4 as-number 100
 peer 4.4.4.4 connect-interface
LoopBack0
 peer 10.0.35.5 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 2.2.2.2 enable
  peer 4.4.4.4 enable
  peer 4.4.4.4 reflect-client
  peer 4.4.4.4 next-hop-local
  peer 10.0.35.5 enable
#
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Run the **display bgp peer** command to check whether a BGP peer relationship has been established.
- Run the **display bgp routing-table** command to check whether routing information has been obtained.

BGP Requirement Implementation and Troubleshooting - Case 1 (3)

- Configurations of devices in AS 200:

```
[R5]
#
bgp 200
 peer 10.0.25.2 as-number 100
 peer 10.0.35.3 as-number 100
 peer 10.0.56.6 as-number 200
peer 7.7.7.7 as-number 300
peer 7.7.7.7 ebgp-max-hop 2
 #
 ipv4-family unicast
  undo synchronization
  aggregate 192.168.0.0 255.255.252.0 as-set detail-suppressed
  peer 10.0.45.4 enable
  peer 10.0.35.3 enable
  peer 10.0.56.6 enable
  peer 7.7.7.7 enable
#
ip route-static 7.7.7.7 255.255.255.255 10.0.57.7
```

```
[R6]
#
bgp 200
peer 10.0.56.6 as-number 200
#
 ipv4-family unicast
  undo synchronization
  peer 10.0.56.5 enable
#
```

- Configurations of devices in AS 300:

```
[R7]
#
bgp 300
peer 5.5.5.5 as-number 200
peer 5.5.5.5 ebgp-max-hop 2
#
 ipv4-family unicast
  undo synchronization
  peer 5.5.5.5enable
#
ip route-static 5.5.5.5 255.255.255.255 10.0.57.5
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The network segment between Rx and Ry (X < Y) is 10.0.xy.0/24. Rx's interface IP address is 10.0.xy.x, and Ry's interface IP address is 10.0.xy.y.
- All interface addresses have been configured.

BGP Requirement Implementation and Troubleshooting - Case 1 (4)

- After the configuration is complete, all BGP peer relationships are established. Two network segments in AS 100, however, cannot access the network segment in AS 300.
  - Check the routing tables of R1 and R3.

```
<R1>dis bgp routing-table
Network          NextHop        MED        LocPrf    PrefVal Path/Ogn

 *>  192.168.0.0     0.0.0.0         0                        0      i
 *>i 192.168.1.0     4.4.4.4         0          100           0      I

[R3-bgp]dis bgp routing-table
Network          NextHop        MED        LocPrf    PrefVal Path/Ogn

 *>i 192.168.0.0     1.1.1.1         0          100           0      i
 *>i 192.168.1.0     4.4.4.4         0          100           0      i
```

```
[R5]dis bgp routing-table

 Total Number of Routes: 6
      Network          NextHop        MED        LocPrf    PrefVal Path/Ogn

 *>   192.168.0.0/22   127.0.0.1                                0      {100 300}i
 s>   192.168.0.0      10.0.25.2                                0      100i
 *                     10.0.35.3                                0      100i
 s>   192.168.1.0      10.0.25.2                                0      100i
 *                     10.0.35.3                                0      100i
 s>   192.168.2.0/23   10.0.56.6       0                        0      300i
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The command output shows that the AS_Set attribute contains the local AS number. As a result, the summary route cannot be accepted. In this situation, disable specific route suppression or delete the AS_Set configuration.

BGP Requirement Implementation and Troubleshooting - Case 1 (5)

- The modified R5 configuration is as follows:

```
#
bgp 200
#
 ipv4-family unicast
  undo synchronization
  aggregate 192.168.0.0 255.255.252.0 detail-suppressed
```

- Check the routing tables of R1 and R3.

```
<R1>dis bgp routing-table
      Network          NextHop      MED    LocPrf    PrefVal    Path/Ogn
*>i   192.168.0.0/22   2.2.2.2               100       0          200i
 *>   192.168.0.0      0.0.0.0      0                 0          i
 *>i  192.168.1.0      4.4.4.4      0        100      0          I

[R3-bgp]dis bgp routing-table
      Network          NextHop      MED    LocPrf    PrefVal    Path/Ogn
*>i   192.168.0.0/22   10.0.35.5             100       0          200i
 *>i  192.168.0.0      1.1.1.1      0        100      0          i
 *>i  192.168.1.0      4.4.4.4      0        100      0          i
```

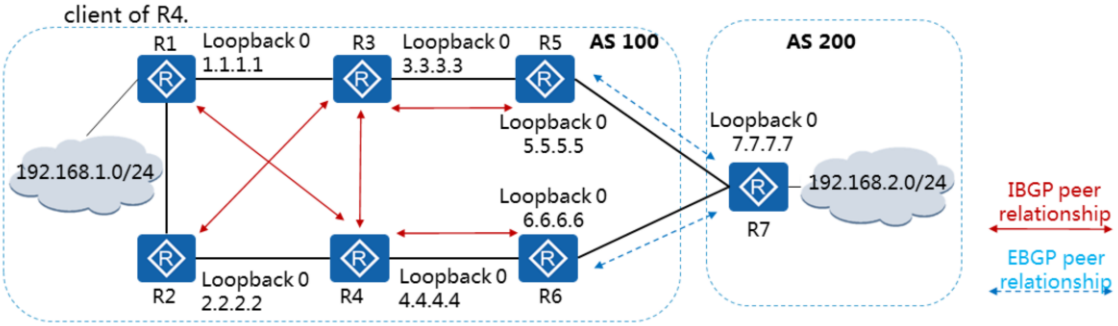
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Case summary:
  - Exercise caution when configuring route summarization. Improper summarization configuration may cause the following problems:
    - Correct routes cannot be learned.
    - Loops may occur.

BGP Requirement Implementation and Troubleshooting - Case 2 (1)

- Requirements:
  - OSPF runs in AS 100. All routers belong to area 0. IBGP peer relationships are established between routers using loopback 0 addresses. The peer relationships are shown in the figure.
  - An EBGP peer relationship is established between R7 and R5 and between R7 and R6 through directly connected interfaces.
  - To enable all routers to learn the routes of AS 200, configure RRs. R5 is the client of R3, and R6 is the client of R4.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The network segment between Rx and Ry (X < Y) is 10.0.xy.0/24. Rx's interface IP address is 10.0.xy.x, and Ry's interface IP address is 10.0.xy.y.
- All interface addresses have been configured.
- R5 is the client of R3, and R6 is the client of R4.

BGP Requirement Implementation and Troubleshooting - Case 2 (2)

- R1 and R4 configurations:

```
[R1]
ospf 1 router-id 1.1.1.1
 area 0.0.0.0
  network 1.1.1.1 0.0.0.0
  network 10.0.12.1 0.0.0.0
  network 10.0.13.1 0.0.0.0
#
bgp 100
 peer 4.4.4.4 as-number 100
 peer 4.4.4.4 connect-interface LoopBack0
 #
 ipv4-family unicast
  undo synchronization
  network 192.168.1.0
 255.255.255.0
  peer 4.4.4.4 enable
```

```
[R4]
#
ospf 1 router-id 4.4.4.4
 area 0.0.0.0
  network 4.4.4.4 0.0.0.0
  network 10.0.24.4 0.0.0.0
  network 10.0.46.4 0.0.0.0
#
#
bgp 100
 peer 1.1.1.1 as-number 100
 peer 1.1.1.1 connect-interface LoopBack0
 peer 3.3.3.3 as-number 100
 peer 3.3.3.3 connect-interface LoopBack0
 peer 6.6.6.6 as-number 100
 peer 6.6.6.6 connect-interface LoopBack0
 #
 ipv4-family unicast
  undo synchronization
  peer 1.1.1.1 enable
  peer 6.6.6.6 enable
  peer 6.6.6.6 reflect-client
#
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- After the configuration is complete, all BGP peer relationships are established, and all routes are learned by OSPF.
- The configuration of R2 is similar to that of R1. The configuration of R3 is similar to that of R4. The configuration of R5 is similar to that of R6.
- After the establishment is complete, R1 advertises the direct route 192.168.1.0/24 to BGP. R7 advertises the direct route 192.168.2.0/24 to BGP.

BGP Requirement Implementation and Troubleshooting - Case 2 (3)

- R6 and R7 configurations:

```
[R6]
ospf 1 router-id 6.6.6.6
 area 0.0.0.0
  network 6.6.6.6 0.0.0.0
  network 10.0.46.6 0.0.0.0
  network 10.0.67.6 0.0.0.0
#
bgp 100
 peer 4.4.4.4 as-number 100
 peer 4.4.4.4 connect-interface LoopBack0
 peer 4.4.4.4 next-hop-local
 peer 10.0.67.7 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 4.4.4.4 enable
  peer 10.0.67.7 enable
```

```
[R7]
#
bgp 200
 router-id 7.7.7.7
 peer 10.0.57.5 as-number 100
 peer 10.0.67.6 as-number 100
 #
 ipv4-family unicast
  undo synchronization
  peer 10.0.57.5 enable
  network 192.168.2.0 255.255.255.0
  peer 10.0.67.6 enable
#
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- After the configuration is complete, all BGP peer relationships are established, and all routes are learned by OSPF.
- The configuration of R2 is similar to that of R1. The configuration of R3 is similar to that of R4. The configuration of R5 is similar to that of R6.
- After the establishment is complete, R1 advertises the direct route 192.168.1.0/24 to BGP. R7 advertises the direct route 192.168.2.0/24 to BGP.

BGP Requirement Implementation and Troubleshooting - Case 2 (4)

- Check routing entries of R1 and R7.

```
[R1]dis bgp routing-table

 Total Number of Routes: 2
      Network         NextHop       MED        LocPrf    PrefVal Path/Ogn
*>i   192.168.2.0     6.6.6.6         0          100       0       200i

[R1]display ip routing-table
Destination/Mask    Proto   Pre  Cost      NextHop         Interface
192.168.2.0/24      IBGP    255  0           6.6.6.6       Ethernet0/0/1
```

```
[R1]ping -a 192.168.1.1 192.168.2.1
  PING 192.168.2.1: 56  data bytes, press CTRL_C to break
    Request time out
    Request time out
    Request time out
    Request time out
    Request time out

  --- 192.168.2.1 ping statistics ---
    5 message(s) transmitted
    0 message(s) received
    100.00% message loss
```

```
<R7>dis bgp routing-table

 Total Number of Routes: 2
      Network         NextHop       MED        LocPrf    PrefVal Path/Ogn

 *>   192.168.1.0     10.0.67.6                  0                 100i
 *>   192.168.2.0     0.0.0.0           0        0                 I

<R7>display ip routing-table
Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface
127.0.0.1/32        Direct  0    0           D   127.0.0.1       InLoopBack0
192.168.1.0/24      EBGP    255  0           D   10.0.67.6       Ethernet0/0/1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- After the configuration is complete, all BGP peer relationships are established, and all routes are learned by OSPF.
- The configuration of R2 is similar to that of R1. The configuration of R3 is similar to that of R4. The configuration of R5 is similar to that of R6.
- After the BGP peer relationships are established, each router advertises its own loopback 0 address.

BGP Requirement Implementation and Troubleshooting - Case 2 (5)

- Check routing entries of R1 and R7.

```
[R1]display ip routing-table
Destination/Mask  Proto  Pre  Cost      NextHop         Interface
192.168.2.0/24    IBGP   255  0         6.6.6.6         Ethernet0/0/1
6.6.6.6/32        OSPF   10   3         10.0.12.2       Ethernet0/0/1
```

```
[R2]display ip routing-table
Destination/Mask  Proto  Pre  Cost      NextHop         Interface
192.168.2.0/24    IBGP   255  0         5.5.5.5         Ethernet0/0/1
5.5.5.5/32        OSPF   10   3         10.0.12.1       Ethernet0/0/1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

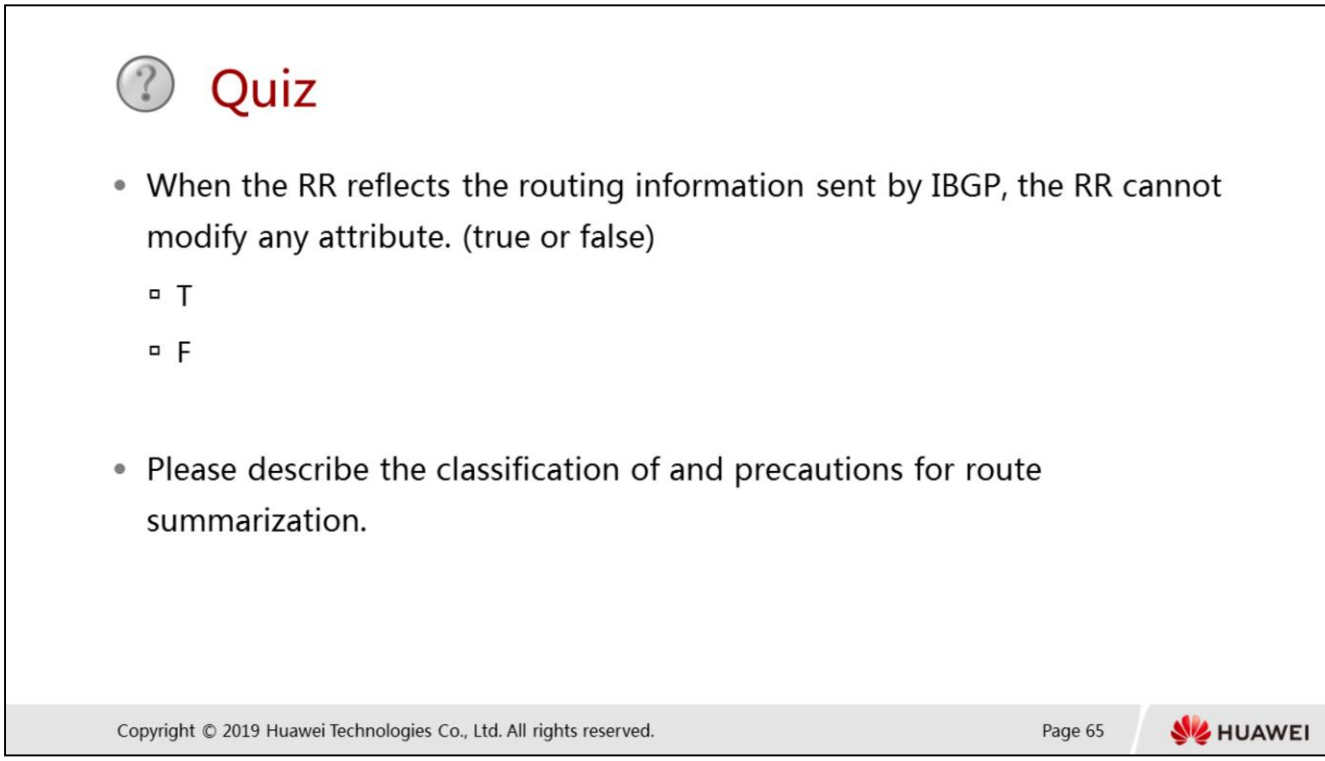
- Fault analysis:
  - R7 advertises the route prefix 192.168.2.0/24 to R5 and R6.
  - R5 and R6 receive the packets and advertise the prefix to their IBGP peers R3 and R4, respectively.
  - This section analyzes R4. A path selection process is performed onR4. R3 also sends the route prefix 192.168.2.0/24 to R4. Based on the preceding 13 BGP path selection rules, R4 selects the route with the smallest IGP cost. Consequently, R6 is selected as a next hop. R4 then sends information about the optimal path to R3 and R1.
  - Similarly, R3 selects R5 as a next hop.
  - The key lies in R1 and R2. R1 can receive the route update sent only by R4. Therefore, the next hop to 192.168.2.0/24 is R4. Similarly, R2's next hop to 192.168.2.0/24 is R5.
  - After the recursive query of IGP routes, packets from 192.168.1.1 to 192.168.2.1 are forwarded between R1 and R2 until the TTL in IP packets is reduced to 0.

BGP Requirement Implementation and Troubleshooting - Case 2 (6)

- Case summary:
  - Use the following rules to configure an RR:
    - Do not configure a peer relationship between an RR and a client through a non-client.
    - Do not configure a peer relationship between an RR and a non-client through a client.
    - Do not establish an IBGP peer relationship between the RR's client and a non-client.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



- Answer:
  1. T
  2. Route summarization includes automatic and manual modes.
     - Automatic summarization: Only routes imported using the import-route command can be summarized. Only natural masks can be used for summarization. IPv6 does not support automatic summarization.
     - Manual summarization: Both IPv4 and IPv6 routes can be summarized. Specific route suppression and the AS_Set option can be configured.

Section Summary

- Deployment solution when mass BGP routes are used
  - Route summarization
  - Peer Group
  - RR
  - Confederation
  - Community attributes
- Internet design fundamentals
  - Which types of network need BGP?
  - BGP optimization
- Typical BGP configuration and troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

More Information

- RFC 3065: RFC 3065 Autonomous System Confederations for BGP
- RFC 4271: RFC 4271 A Border Gateway Protocol 4 (BGP-4)
- RFC 4360: RFC 4360 BGP Extended Communities Attribute
- RFC 4456: RFC 4456 BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)
- RFC 4760: RFC 4760 Multiprotocol Extensions for BGP-4
- RFC 4893: RFC 4893 BGP Support for Four-octet AS Number Space

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Recommendations

- Huawei Learning website
  - http://support.huawei.com/learning/en/newindex.html
- Huawei Support case library
  - http://support.huawei.com/enterprise/servicecenter?lang=en

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank you

www.huawei.com

HCIE-Routing & Switching

— Route Import and Control

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- On complex IP networks, routing policies usually need to be implemented based on actual networking requirements to change forwarding paths of network traffic by modifying route attributes (including the route reachability). Routing policies can be used to filter routes, set route attributes, and control route receiving, advertisement, and importing.
- This document describes the implementation and configurations of commonly used route filters, routing policies, and policy-based routing (PBR).

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

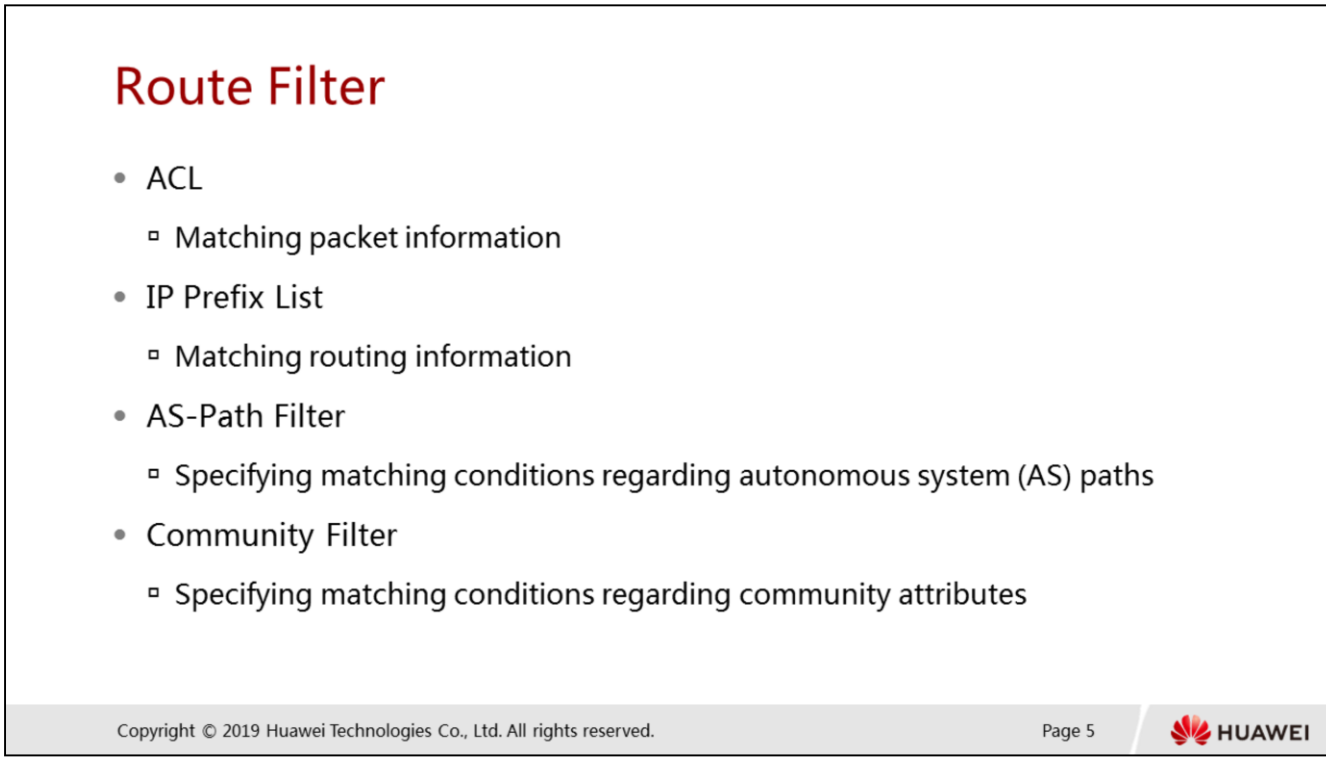
- Upon completion of this course, you will be able to:
  - Have a good command of route filters such as access control lists (ACLs) and IP prefix lists.
  - Understand the principles and configuration methods of routing policies.
  - Understand the principles and configuration methods of PBR.
  - Use routing policies and PBR for route control and packet forwarding.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. Route Filter
   - ACL
   - IP Prefix List
   - AS-Path Filter
   - Community Filter
2. Routing Policy and PBR
3. Routing Policy and PBR Configuration Commands
4. Case Analysis

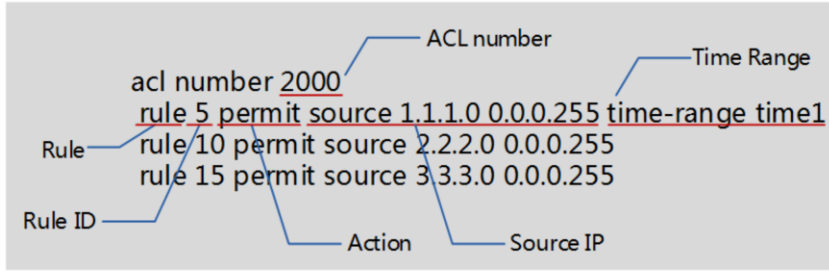
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- ACL
  - An ACL is composed of a list of rules. Each rule contains a permit or deny clause. These rules classify packets based on information in the packets. After ACL rules are applied, the routers determine the packets to be received and rejected.
  - An ACL defines a series of rules and identifies data packets that need to be filtered. Then, routers permit or deny data packets according to the configured rules. In addition, an ACL can be referenced by other service modules as a basic configuration.
- IP Prefix List
  - An IP prefix list matches routes with each entry in the list to filter routes based on the defined matching mode.
  - An IP prefix list can be used to filter only routing information, but cannot filter data packets.
- AS-Path Filter
  - Information about each Border Gateway Protocol (BGP) route contains an AS path domain. AS-Path filters specify matching rules regarding AS path domains. An AS-Path filter is used to filter only BGP.
- Community Filter
  - Information about each BGP route can carry one or more community attributes. A community filter specifies matching conditions regarding community attributes.

Route Filter: ACL Implementation

- An ACL is a collection of one or more rules. Rules describe packet matching conditions, such as the source IP address, destination IP address, and port number of packets.
- An ACL is a packet filter that filters packets based on rules. A device with an ACL configured matches packets against ACL rules to filter packets of a certain type, and then permits or denies these packets according to the policies used by the service module to which the ACL is applied.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- ACL number: identifies a numbered ACL.
  - Depending on functions, ACLs are classified into basic ACL, advanced ACL, Layer 2 ACL, and user ACL. These ACLs have different number ranges.
  - You can also define the name of an ACL to help you remember the ACL's purpose. In this situation, an ACL name is like a domain name that represents an IP address. Such an ACL is called named ACL.
  - An ACL number can be part of an ACL name. That is, you can also specify an ACL number when you define a named ACL. If you do not specify an ACL number, the system will automatically allocate a number to an ACL.
- Rule: describes packet matching conditions.
  - Rule ID: identifies an ACL rule. The rule IDs can be manually set or automatically allocated by the system. The ACL rule IDs range from 0 to 4294967294. The rule IDs in an ACL are allocated in an ascending order. Therefore, in the above figure, rule 5 is in the first line of an ACL and rule 15 is in the bottom line. The system matches packets against the rules from the first line to the bottom line, and stops matching if the packets match a rule.
  - Action: includes permit and deny.
  - Matching condition: ACLs support many matching conditions. In addition to the source IP address and time range shown in the above figure, an ACL supports many other matching conditions, for example, Layer 2 Ethernet frame header information (such as the source MAC address, destination MAC address, and Ethernet protocol

type), Layer 3 packet information (such as the destination IP address and protocol type), and Layer 4 packet information (such as the TCP/UDP port number).

- If an ACL contains rules, the system matches packets against the rules in ascending order of rule IDs. If the packets match a permit rule, the system stops matching and returns the result "positive match (permit)." If the packets match a deny rule, the system stops matching and returns the result "positive match (deny)." If the packets do not match a rule in the ACL, the system continues matching the packets against the next rule. If the packets do not match any rule in the ACL, the system returns the result "negative match.

Route Filter: ACL Classification

- ACLs are classified in different ways.
  - Based on ACL naming methods:
    - Numbered ACL
    - Named ACL
  - Based on ACL rule definition methods:
    - Basic ACLs (number range: 2000–2999)
    - Advanced ACLs (number range: 3000–3999)
    - Layer 2 ACLs (number range: 4000–4999)
    - User ACLs (number range: 6000–6031)

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Basic ACL
  - A basic ACL defines rules based on source IP addresses, fragmentation information, and time range of packets.
- Advanced ACL
  - An advanced ACL defines rules based on the source IPv4 address, destination IPv4 address, IPv4 protocol type, Internet Control Message Protocol (ICMP) type, TCP source/destination port numbers, UDP source/destination port numbers, and time range of packets.
- Layer 2 ACL
  - A Layer 2 ACL defines rules based on information in Ethernet frame headers of packets, such as the source MAC address, destination MAC address, and Layer 2 protocol type.
- User ACL
  - A user ACL defines rules based on the source IPv4 address, destination IPv4 address, IPv4 protocol type, ICMP type, TCP source/destination port number, and UDP source/destination port number of packets.
- In addition, there are IPv6 ACLs (ACL6s), including basic ACL6s and advanced ACL6s.
  - Basic ACL6: defines rules based on the source IPv6 address, fragmentation information, and time range.
  - Advanced ACL6: defines rules based on the source IPv6 address, destination IPv6

address, IPv6 protocol type, ICMPv6 type, TCP source/destination port numbers, UDP source/destination ports, and time range.

Rout Filter: Matching Order of ACL Rules

- Configuration order (config)
  - Packets match ACL rules in ascending order of rule IDs.
- Automatic order (auto)
  - Packets match ACL rules according to the precision degree of the rules (depth-first principle). The following table describes details about the matching order o

ACL Type	Matching Order
Basic ACL	1. The rule that defines a virtual private network (VPN) instance is matched first. 2. The rule that defines the smallest source IP address range is matched. 3. The rule with the smallest ID is matched.
Advanced ACL	1. The rule that defines a VPN instance is matched first. 2. The rule that defines the type of the protocol over IP is matched first. 3. The rule that defines the smallest source IP address range is matched. 4. The rule that defines the smallest destination IP address range is matched. 5. The rule that defines the smallest Layer 4 port number range is matched. 6. The rule with the smallest ID is matched.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Matching order of ACL rules
  - An ACL consists of multiple deny and permit clauses, each of which describes a rule. These rules may repeat or conflict. One rule can contain another rule, but two rules must be different.)
  - Two matching orders of ACL rules are supported: configuration order (config) and automatic order (auto). When the system matches a data packet against rules in an ACL, the rule matching order decides the rule priorities. The ACL processes rule overlapping or conflict based on rule priorities. The default matching order is **config**.
- Configuration order (config)
  - If a smaller rule ID is manually specified for a rule, the rule is inserted in one of the front lines of an ACL. This rule is matched earlier.
  - If no ID is manually specified for a rule, the system automatically allocates an ID to the rule. The rule ID is greater than the largest rule ID in the ACL and is the minimum integer of the integer multiple of the increment; therefore, this rule is matched last.
- Automatic order (auto)
  - The system matches packets against ACL rules according to the precision degree of the rules (depth-first principle).
  - The system matches packets against the rules in descending order of precision. A rule with the highest precision defines strictest conditions (such as the protocol

type and source and destination IP address ranges). For example, an ACL rule can be configured based on the wildcard mask of an IP address. A smaller wildcard identifies a smaller network segment and stricter matching conditions.

- If the ACL rules are of the same depth-first order, they are matched in ascending order of rule IDs.

IPv6 ACL

- IPv6 ACLs (also called ACL6s) filter IPv6 packets based on configured rules. The ACL6 implementation is similar to that of ACLs.
- ACL6 classification

Category	Number Range	Usage Scenario
Basic ACL6	2000–2999	A basic ACL6 defines rules based on source IPv6 addresses, VPN instances, fragmentation flags, and time ranges of IPv6 packets.
Advanced ACL6	3000–3999	An advanced ACL defines rules based on source and destination IPv6 addresses, IPv6 protocol types, and protocol properties (such as source and destination TCP port numbers, ICMPv6 protocol type, and ICMPv6 Code).

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

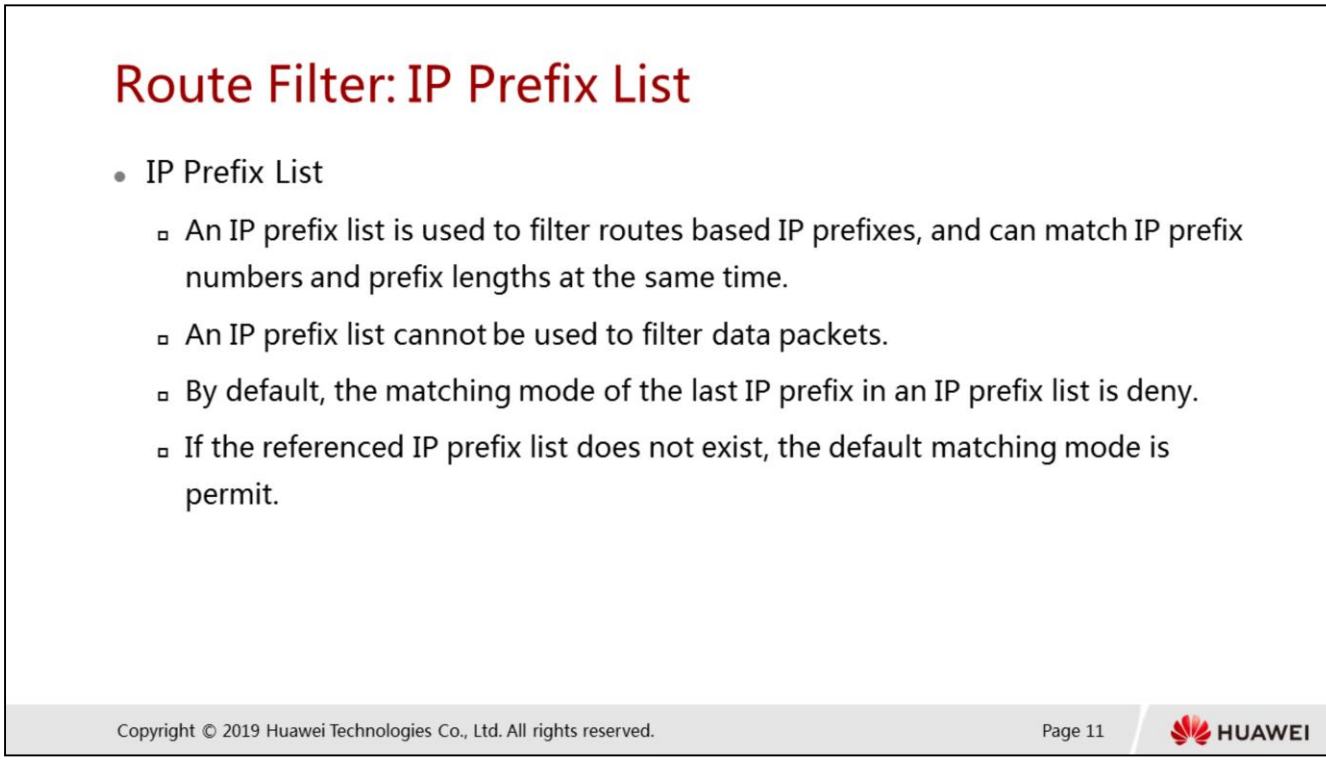
- ACL6s and ACLs are configured using different commands. ACL6s and ACLs can have the same number and do not affect each other.
- Example:
  - [RouterA] acl ipv6 number 3001
  - [RouterA-acl6-adv-3001] rule deny ipv6 source 3001::2/64
  - [RouterA] acl 3001
  - [Router-acl-adv-3001] rule permit ip source 202.169.10.5 0.0.0.0



Contents

1. **Route Filter**
   - ACL
   - **IP Prefix List**
   - AS-Path Filter
   - Community Filter
2. Routing Policy and PBR
3. Routing Policy and PBR Configuration Commands
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- IP prefix list
  - Each IP prefix list can contain multiple IP prefixes, and each IP prefix entry corresponds to an index. The system matches the prefix of a route against IP prefixes in the IP prefix list in ascending order of indexes. If any IP prefix is matched, the system stops matching against other IP prefixes. If no IP prefix in the list is matched, the route is filtered out.
  - An IP prefix list supports accurate matching or matching within a specified mask length.
  - An IP prefix list can match a specific route or match the routes within a specified mask length. The prefix mask length can also be specified using the keywords **greater-equal** or **less-equal**. If the keyword **greater-equal** or **less-equal** is not specified, accurate matching is used. That is, only the route with the same mask length as that in the IP prefix list is matched. If only the keyword **greater-equal** is specified, the routes whose mask length ranges from the **greater-equal** value to 32 bits are matched. If only the keyword **less-equal** is specified, the routes whose mask length ranges from the specified value to the **less-equal** value are matched.
  - *greater-equal-value* and *less-equal-value* must meet the following requirements: *mask-length* ≤ *greater-equal-value* ≤ *less-equal-value* ≤ 32.

- IP prefix list characteristics
  - If no IP prefix is matched in an IP prefix list, the default matching mode of the last

IP prefix in the list is deny by default.

- If the referenced IP prefix list does not exist, the default matching mode is permit.

IP Prefix List Examples

- ip ip-prefix FILTER index 10 permit 1.1.1.0 24
  - This is an accurate matching case. Only the route to 1.1.1.1/24 is permitted.
- ip ip-prefix FILTER index 10 permit 1.1.1.0 24 less-equal 32
  - Only the routes to 1.1.1.0 with the mask length ranging from 24 bits to 32 bits are permitted
- ip ip-prefix FILTER index 10 permit 1.1.1.0 24 greater-equal 26
  - Only the routes to 1.1.1.0 with the mask length ranging from 26 bits to 32 bits are permitted.
- ip ip-prefix FILTER index 10 permit 1.1.1.0 24 greater-equal 26 less-equal 32
  - Only the routes to 1.1.1.0 with the mask length ranging from 26 bits to 32 bits are permitted.
- ip ip-prefix FILTER index 10 permit 0.0.0.0 0 greater-equal 8 less-equal 32
  - Routes with the mask length ranging from 8 bits to 32 bits are all permitted.
- ip ip-prefix FILTER index 20 permit 0.0.0.0 0 less-equal 32
  - All routes are permitted.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

IPv6 Prefix List

- IPv6 prefix lists filter IPv6 packets based on configured rules. The implementation is similar to that of IP prefix lists.
- An IPv6 prefix list is used to filter routes based on IPv6 addresses. An IPv6 prefix list may contain multiple entries with each entry specifying an IPv6 prefix range. The relationship between the entries is "OR". That is, if a route matches one entry, the route matches the IPv6 prefix list; if a route does not match any entry, the route fails to match the IPv6 prefix list.
- All unmatched routes are denied by default. If all entries are in deny mode, all routes are filtered out. Therefore, after multiple entries in deny mode, a **permit :: 0 less-equal 128** entry needs to be defined to permit all the other IPv6 routes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

IPv6 Prefix List Examples

- Permit the routes with the mask length ranging from 32 bits to 64 bits.
  - [Huawei] ip ipv6-prefix abc permit :: 0 greater-equal 32 less-equal 64
- Deny the routes with the IP prefix FC00:3:0:D00::/32 and with prefixes longer than 32 bits, and permit the other IPv6 routes.
  - [Huawei] ip ipv6-prefix abc deny 3FFE:D00:: 32 less-equal 128
  - [Huawei] ip ipv6-prefix abc permit :: 0 less-equal 128
- Configure the IPv6 prefix list named p3 to deny the routes to IPv6 addresses ranging from ::1 to ::FFFF:FFFF.
  - [Huawei] ip ipv6-prefix p3 index 10 deny :: 96 match-network
  - [Huawei] ip ipv6-prefix p3 index 20 permit :: 0 less-equal 128

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

1. **Route Filter**
   - ACL
   - IP Prefix List
   - **AS-Path Filter**
   - Community Filter
2. Routing Policy and PBR
3. Routing Policy and PBR Configuration Commands
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Route Filter: AS-Path Filter

- AS-Path Filter
  - AS-Path filters use the AS-Path attribute of BGP to filter routes.
  - AS-Path filters use regular expressions to define matching rules.
- Examples
  - ip as-path-filter 10 permit .*
    - Configure AS-Path filter 10 to permit all routes.
  - ip as-path-filter 10 permit _100$
    - Configure AS-Path filter 10 to permit routes originating from AS100.
  - ip as-path-filter 10 permit ^100_
    - Configure AS-Path filter 10 to permit routes received from AS100.
  - ip as-path-filter 10 permit _100|200$
    - Configure AS-Path filter 10 to permit routes originating from AS100 or AS200.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An AS-Path filter uses the AS-Path attribute of BGP to filter routes. It is used only when BGP advertises and receives routes.

- The AS-Path attribute records the number of the AS that a route passes through to the leftmost of the AS-Path list. Therefore, pay special attention when configuring an AS-Path filter.
  - If a route originates from AS100 and passes through AS300, AS200, and AS500, and finally reaches AS600, the AS-Path attribute of the route is 500 200 300 100 in AS600.

Route Filter: Commonly Used Regular Expressions

Expression	Meaning
^$	Routes originating from the local AS
.*	All routes
_10_	Routes passing through AS10
^10_	Routes received only from AS10
_10$	All routes originating from AS10

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

1. **Route Filter**
   - ACL
   - IP Prefix List
   - AS-Path Filter
   - **Community Filter**
2. Routing Policy and PBR
3. Routing Policy and PBR Configuration Commands
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Route Filter: Community Filter

- Community Filter
  - Community filters use community attributes of BGP to filter routes.
- Examples
  - ip community-filter 1 permit 100:1
    - **Set the community attribute for route matching to 100:1.**
  - ip community-filter 1 permit no-export
    - **Set the community attribute for route matching to no-export.**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A community filter uses community attributes of BGP to filter routes. It is used only when BGP advertises and receives routes.

- Community attributes include basic and extended community attributes.
  - User-defined community attributes and well-known community attributes are basic community attributes.
  - The route target (RT) and Site of Origin (SoO) attributes in MPLS VPN scenarios are extended community attributes.



Contents

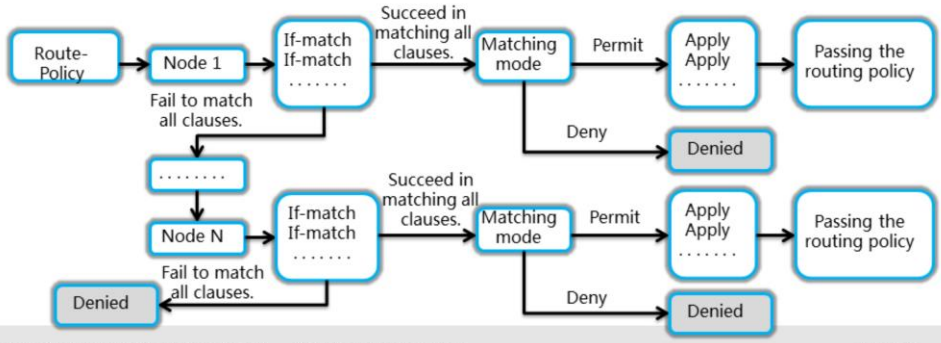
1. Filter
2. **Routing Policy and PBR**
   - **Routing Policy**
   - PBR
   - Comparison Between Routing Policy and PBR
3. Policy Routing and PBR Configuration Commands
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Routing Policy

- Routing policy
  - Routing policies are used to filter routes and set route attributes. By changing the attributes of a route, a routing policy can change network traffic paths.
  - A routing policy is configured using the **route-policy** command.
- Working mechanism of a routing policy



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Routing policies change the forwarding path of packets by filtering routes and setting route attributes. For example, route attributes (including reachability) can be set to change the forwarding path of network traffic.
- Routing policies are mainly used in the following scenarios:
  - Control route importing.
    - A routing policy can be used to prevent sub-optimal paths or loops when routes are imported.
  - Control route receiving and advertising.
    - Only required and valid routes are received or advertised.
  - Modify attributes of specified routes.
    - Attributes of routes that are filtered by a routing policy can be modified to adjust and optimize networks.

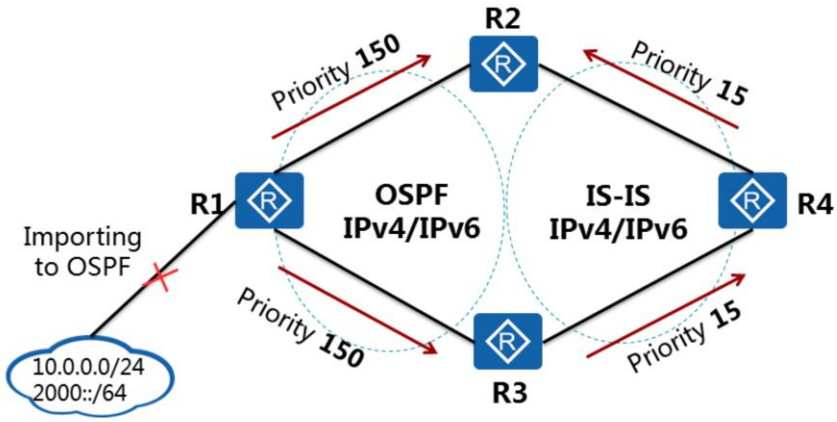
- Working Mechanism of Routing Policies
  - A routing policy consists of more than one node. The system checks routes in the nodes of a routing policy in ascending order of node IDs. One node can be configured with multiple **if-match** and **apply** clauses. The **if-match** clauses define matching rules for this node, and the **apply** clauses define actions for the routes

that match the rules. The relationship between **if-match** clauses is "AND". That is, a route must match all the **if-match** clauses. The relationship between the nodes of a routing policy is "OR". That is, if a route matches one node, the route matches the routing policy. If a route does not match any node, the route fails to match the routing policy.

- The relationship between the **if-match** clauses in a node of a routing policy is "AND". A route must match all the rules before the action defined by the **apply** clause is taken. The relationship between the **if-match** clauses in the **if-match route-type** and **if-match interface** commands is "OR", but the relationship between the **if-match** clauses in these two commands and other commands is "AND".

Routing Policy: Controlling Route Importing

- Control route importing.
  - Control route importing to prevent loops or sub-optimal routes



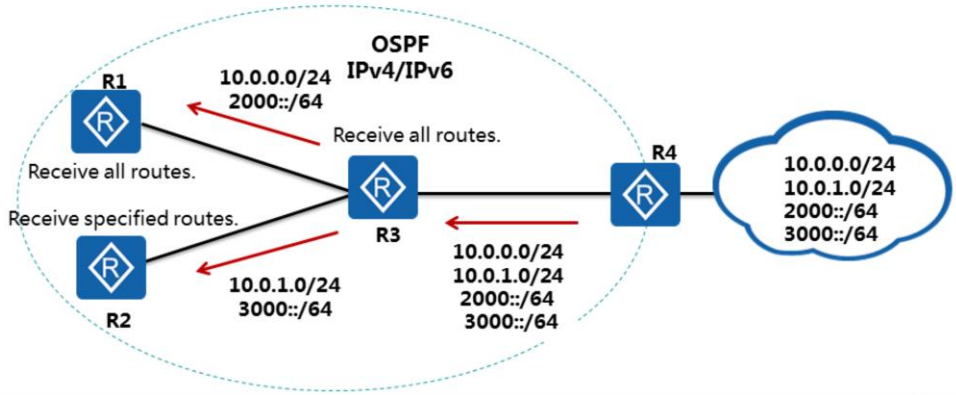
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The preceding topology shows dual-point bidirectional route re-advertisement. If no control is performed, sub-optimal routes and loops may occur. The fault analysis is as follows:
  - R1 imports routes destined for the network segments 10.0.0.0/24 and 2000::/64 into OSPF. R2 and R3 imports the routes to IS-IS respectively. Assume that R2 imports the routes to IS-IS earlier than R3. Then R2 learns the routes destined for 10.0.0.0/24 and 2000::/64 from both OSPF and IS-IS. R2 preferentially selects the routes learned from IS-IS according to the preference of routing protocols. (The preference of external routes in OSPF process is 150, and the preference of IS-IS routes is 15.) Therefore, when R2 accesses the network segments 10.0.0.0/24 and 2000::/64, the sub-optimal route R4-R3-R1 is used. To prevent this issue, run the **route-policy** command on R2 to change the preference of the OSPF ASE route to be higher than the route learned from IS-IS, so that R2 selects the correct route.
  - If the interface on R1 connected to the network segments 10.0.0.0/24 and 2000::/64 is shut down, the external LSAs are aged in the OSPF area. R2 imports the route to OSPF because it has learned the network segments 10.0.0.0/24 and 2000::/64 from IS-IS. In this way, R1 and R3 can learn the routes destined for network segments 10.0.0.0/24 and 2000::/64 from R2. When R2 accesses network segments 10.0.0.0/24 and 2000::/64, traffic is transmitted along the path R4->R3->R1->R2, causing a routing loop. To prevent the loop, you can add tags to routing entries and filter routes with specific tags to prevent routing loops.

Routing Policy: Controlling Route Receiving and Advertising

- Control route receiving and advertising.
  - Specified routes are received to optimize routing entries.
  - Generally, this function is implemented using the **filter-policy** command.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Control route receiving and advertising.
  - Only the required and valid routes are received. This reduces the size of the routing table and improves network security.

- Topology Description
  - R4 imports routes destined for 10.0.X.0/24, 2000::/64, and 3000::/64 into OSPF. According to service requirements, R1 can receive only the routes destined for 10.0.0.0/24 and 2000::/64, and R2 can receive only the routes destined for 10.0.1.0/24 and 3000::/64. This requirement can be met using the **filter-policy** command.

Routing Policy: Filtering Policy (OSPF)

- Filter-policy Import
  - The **filter-policy import** command is used to configure a filtering policy to filter received routes. Only the routes that pass the filtering can be added to the routing table. The routes that fail to pass the filtering cannot be added to the routing table but can be advertised.
  - OSPF routing information is recorded in the link state database (LSDB). Instead of filtering the received or sent LSAs, the device filters routes calculated by OSPF using the **filter-policy import** command.
- Filter-policy Export
  - After importing external routes to OSPF using the **import-route** command, you can use the **filter-policy export** command to filter the imported routes to be advertised to prevent loops. Only the external routes that pass the filtering can be converted into Type 5 and Type 7 LSAs and advertised.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The **filter-policy import** command configures a filtering policy to filter routes received by OSPF.
- The **filter-policy export** command configures a filtering policy to filter imported routes to be advertised.
  - The **protocol** or **process-id** parameter can be specified to determine a specified protocol or process. If the **protocol** or **process-id** parameter is not specified, OSPF filters all imported routes.
  - This command can be configured only on autonomous system boundary routers (ASBRs) because Type 5 and Type 7 LSAs are generated by ABSRs.

Routing Policy: Filtering Policy (IS-IS)

- Filter-policy Import
  - IS-IS routing entries need to be added to an IP routing table to guide IP packet forwarding. If an IS-IS routing table has routes destined for a specific network segment, but these routes are not expected to be added to an IP routing table, run the **filter-policy import** command with specified conditions (such as a basic ACL, an IP prefix, and a routing policy) to allow only the needed IS-IS routes to be added to the IP routing table.
- Filter-policy Export
  - When IS-IS and other routing protocols are running on a network, and a boundary router in the IS-IS routing domain has imported routes of other routing protocols, the boundary router will advertise all the imported routes to its IS-IS neighbors by default. To advertise only some of the imported routes to neighbors, use the **filter-policy export** command.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The **filter-policy import** command configures a filtering policy to allow IS-IS to filter received routes to be added to the IP routing table.
  - Running this command on a router does not affect LSP flooding and LSDB synchronization on the router, but affects the local IP routing table.
- The **filter-policy export** command configures a filtering policy to allow IS-IS to filter the imported routes to be advertised.
  - Running this command does not affect the routes on the local device, but advertises only specific imported routes to IS-IS neighbors.

Routing Policy: Filtering Policy (BGP)

- Filter-policy Import
  - The **filter-policy import** command is used to filter the routes received by a BGP-enabled device to determine whether to add them to the BGP routing table.
- Filter-policy Export
  - The filter-policy export command is used to filter routes to be globally advertised.
  - After the filter-policy export command is run, BGP filters the routes that are imported by using the import-route (BGP) command. Only the routes that pass the filtering can be added to the local BGP routing table and advertised by BGP

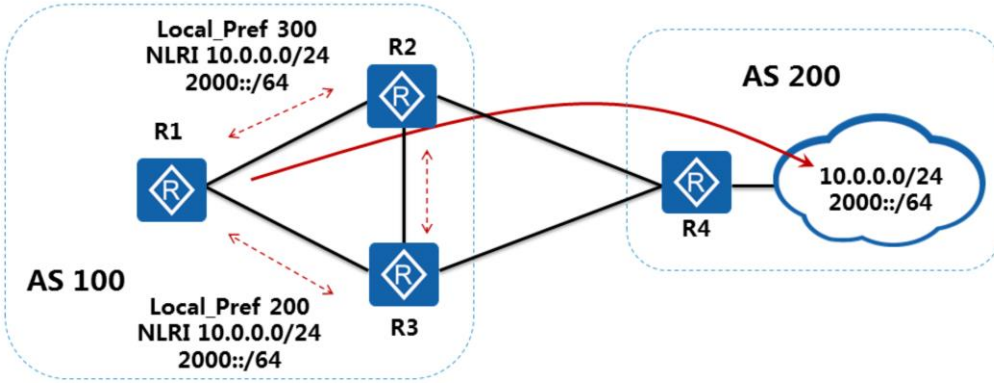
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The **filter-policy import** command configures a device to filter received routes.
- The **filter-policy export** command configures a device to filter the routes to be advertised. BGP advertises only the routes that pass filtering.
  - If the **protocol** parameter is specified in the command, only the routes imported from the specified protocol will be filtered. If the **protocol** parameter is not specified, the routes imported from all protocols will be filtered.

Routing Policy: Setting Attributes of Specified Routes

- Set attributes of specified routes.
  - You can set attributes of specified routes using routing policies.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Topology Description
  - Run the **route-policy** command to modify the **Local_Pref** attribute of BGP routes, which affects the traffic forwarding direction. On R2, the **Local_Pref** attribute of the routes destined for 10.0.0.0/24 and 2000::/64 learned from EBGP to 300. On R3, set the **Local_Pref** attribute of the routes learned from EBGP to 200. R1, R2, and R3 exchange routes with each other through IBGP. Finally, R2 is selected as the egress of the local AS to 10.0.0.0/24 and 2000::/64.



Contents

1. Rout Filter
2. **Routing Policy and PBR**
   - Routing Policy
   - **PBR**
   - Comparison Between Routing Policy and PBR
3. Policy Routing and PBR Configuration Commands
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

PBR

- Traditionally, devices search their IP routing tables for routes of packets based on packets' destination addresses and then forward the packets. PBR allows network administrators to define policies to change packet routes based on packets' source addresses, packet size, and link quality in addition to packets' destination addresses.
- PBR has the following advantages:
  - Allows network administrators to define policies for routing packets, which improves route selection flexibility and controllability.
  - Enables different data flows to be forwarded on different links, which increases link efficiency.
  - Uses cost-effective links to transmit service data without compromising service quality, which reduces the cost of enterprise data services.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- PBR is different from routing policies as follows:
  - PBR applies to data packets and provides a means to change the forwarding path of data packets, in accordance with predefined policies instead of following the routes in an existing routing table.
  - Routing policies implement routing based on routing information. Routing policies change the forwarding path of packets by filtering routes and setting route attributes. For example, route attributes (including reachability) can be set to change the forwarding path of network traffic.

PBR Classification

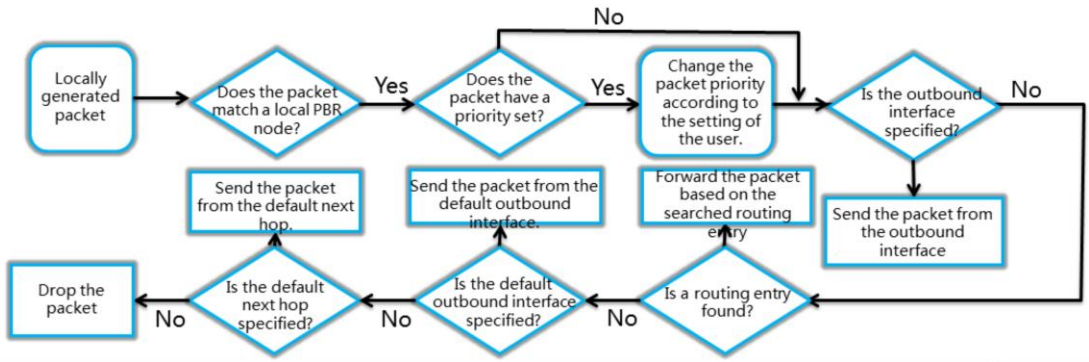
- BR can be categorized into local PBR, interface PBR, and smart policy routing (SPR).
- Local PBR
  - Local PBR applies only to locally generated packets, but not to forwarded packets.
- Interface PBR
  - Interface PBR applies only to forwarded packets, but not to locally generated packets (such as local ping packets).
- SPR
  - SPR allows devices to select routes based on link quality and requirements for link quality.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Local PBR

- Local PBR
  - Local PBR applies only to locally generated packets, but not to forwarded packets.
  - Local PBR on a device can have multiple local PBR nodes. Packets are matched based on the local PBR node sequence.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Matching order
  - If the device finds a matching local PBR node, it performs the following steps:
    1. Checks whether the priority of the packets has been set. If so, the device applies the configured priority to the packets and performs the next step. If not, the device performs the next step.
    2. Checks whether an outbound interface has been configured for local PBR. If so, the device sends the packets out from the outbound interface. If not, the device performs the next step.
    3. Checks whether next hops have been configured for local PBR (two next hops can be configured for load balancing). If so, the device sends the packets to the next hop. If not, the router searches the routing table for a route based on the destination addresses of the packets. If no route is available, the device performs the next step.

  1. Checks whether a default outbound interface has been configured for local PBR. If so, the device sends the packets out from the default outbound interface. If not, the device performs the next step.
  2. Check whether default next hops have been configured for local PBR. If so, the device sends the packets to the default next hops. If not, the device performs the next step.
  3. Discards the packets and generates **ICMP_UNREACH** messages.

  - If the device does not find any matching local PBR node, it searches the routing

table for a route based on the destination addresses of the packets and then sends the packets.



Contents

1. Rout Filter
2. **Routing Policy and PBR**
   - Routing Policy
   - PBR
   - **Comparison Between Routing Policy and PBR**
3. Policy Routing and PBR Configuration Commands
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Comparison Between Routing Policy and PBR

Feature	Characteristics
Routing policy	1. Routing policies enable packet forwarding according to the destination address in the routing table. 2. Routing policies are based on the control plane and serve routing protocols and routing tables. 3. Routing policies are combined with a routing protocol to form a policy. 4. The route-policy command is used to create a routing policy.
PBR	1. Packets are forwarded based on a policy. If packets fail to be forwarded based on the policy, the device searches the routing table for packet forwarding. 2. PBR is based on the forwarding plane and serves forwarding policies. 3. PBR needs to be manually configured hop by hop to ensure that packets are forwarded according to policies. 4. The policy-based-route command is used to configure PBR.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



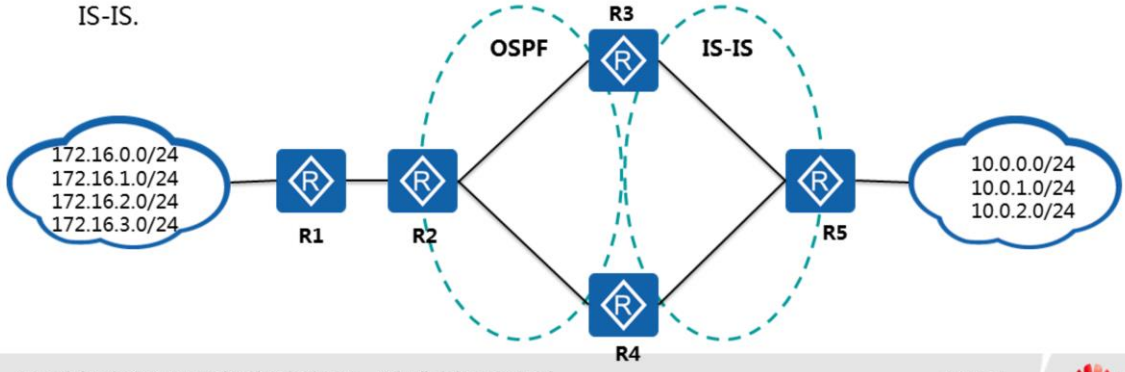
Contents

1. Rout Filter
2. Routing Policy and PBR
3. **Policy Routing and PBR Configuration Commands**
   - **Configuring Route Import**
   - Configuring Policy-based Route Optimization
   - Configuring PBR
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

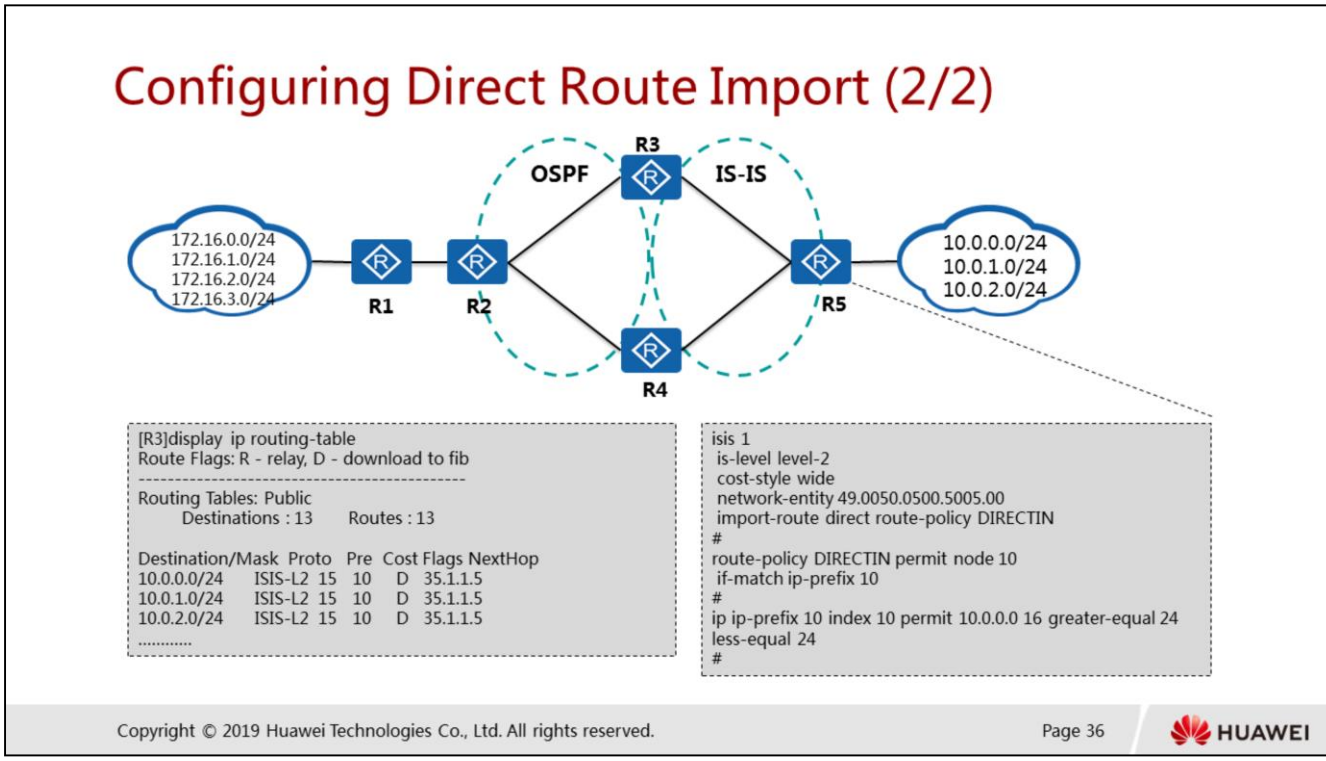
Configuring Direct Route Import (1/2)

- The following figure shows the network topology of company A. Complete the following configurations as required:
  - Run OSPF on R2, R3, and R4, and IS-IS on R3, R4, and R5.
  - Import routes destined for the network segment 10.0.X.0/24 directly connected to R5 to IS-IS.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- In this case, the addresses for interconnecting devices are as follows:
  - For example, if RTX is interconnected with RTY, the interconnection addresses are XY.1.1.X and XY.1.1.Y, and the mask length is 24 bits.

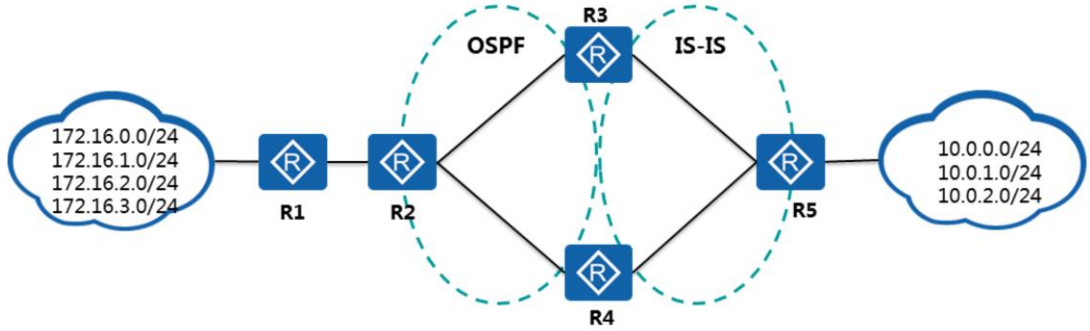


- Command Usage
  - The **route-policy** command creates a routing policy and displays the routing policy view.
- Parameter Description
  - route-policy route-policy-name { permit | deny } node node
    - route-policy-name: specifies the name of a routing policy.
    - permit: specifies the matching mode of the routing policy as permit. If a route matches all the if-match clauses of a node, the route matches the node and all the actions defined by the apply clause are performed on the route. Otherwise, the route continues to match the next node.
    - deny: specifies the matching mode of the routing policy as deny. If a route matches all the if-match clauses of a node, the route is denied and does not match the next node
    - node node: specifies the index of the node in the routing policy.
- Precautions
  - A routing policy is used to filter routes and set route attributes for the routes that match the routing policy. A routing policy consists of multiple nodes. One node can be configured with multiple **if-match** and **apply** clauses. The **if-match** clauses define matching rules for this node, and the **apply** clauses define actions for the routes that match the rules. The relationship between **if-match** clauses is "AND".

That is, a route must match all the **if-match** clauses. The relationship between **if-match** clauses is "AND". That is, a route must match all the **if-match** clauses. If a route does not match any node, the route fails to match the routing policy.

Configuring Static Route Import (1/2)

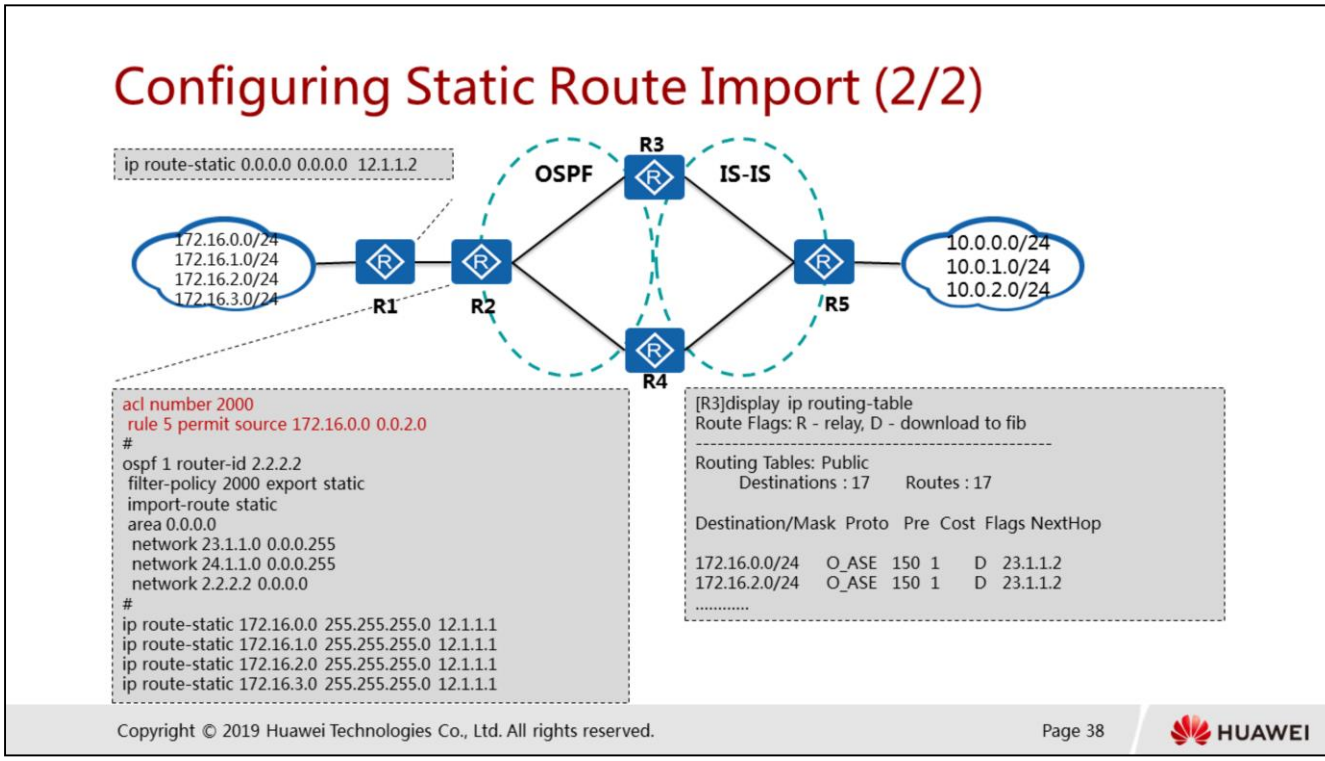
- Company A needs to deploy a network between R1 and R2, and has the following requirements:
  - Static routes are configured between R1 and R2 so that R1 and R2 can communicate with each other.
  - Routes destined for 172.16.X.0/24 are imported to OSPF. The OSPF domain can learn only the routes destined for 172.16.0.0/24 and 172.16.2.0/24. Run the filter-policy and ACL commands to achieve the optimal configuration.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- These requirements are expanded on the basis of those in the previous case. Perform configurations based on those in the previous case.
- This requirement is provided to help you understand filtering policies and ACLs. The optimal configuration means that you can use the fewest commands to meet the desired effect.

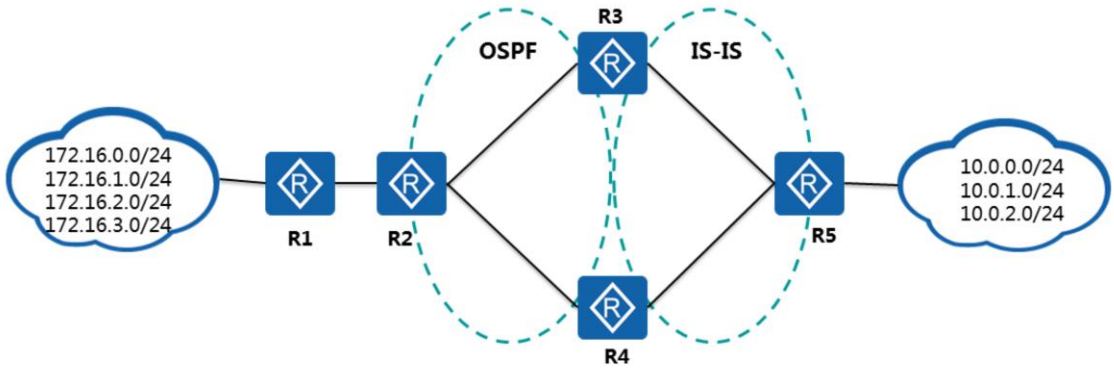


- Command Usage
  - The **filter-policy export** command configures a filtering policy to filter imported routes to be advertised.
- Parameter Description
  - filter-policy { acl-number | acl-name acl-name | ip-prefix ip-prefix-name } export [protocol [process-id]]
    - acl-number: specifies the number of a basic ACL.
    - acl-name acl-name: specifies the name of an ACL.
    - ip-prefix ip-prefix-name: specifies the name of an IP prefix list.
    - Protocol: specifies the name of a routing policy.
    - process-id: specifies the process ID when the advertised protocol is RIP, IS-IS, or OSPF.
- Precautions
  - After OSPF imports external routes using the **import-route** command, you can use **the filter-policy export** command to filter the imported routes to be advertised. Only the external routes that pass the filtering can be converted into Type 5 LSAs (AS-external LSAs) and advertised.
  - The *protocol* or *process-id* parameter can be specified to filter routes of a specified protocol or process. If the *protocol* or *process-id* parameter is not specified, OSPF

filters all imported routes.

Configuring Two-Way Route Import (1/2)

- Company A needs to implement network interworking, and has the following requirements:
  - OSPF and IS-IS on R3 and R4 import routes from each other.

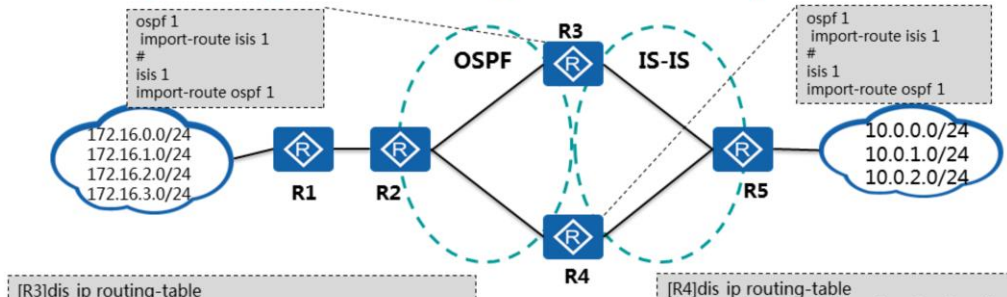


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The network topology in this case is the same as that in the previous case. After complete configurations as required, you need to check whether sub-optimal routes exist or routing loops occur.

Configuring Two-Way Route Import (2/2)



[R3]dis ip routing-table
Route Flags: R - relay, D - download to fib

Routing Tables: Public
Destinations : 17 Routes : 17

Destination/Mask	Proto	Pre	Cost	Flags	NextHop
10.0.0.0/24	ISIS-L2	15	10	D	35.1.1.5
10.0.1.0/24	ISIS-L2	15	10	D	35.1.1.5
10.0.2.0/24	ISIS-L2	15	10	D	35.1.1.5
24.1.1.0/24	OSPF	10	2	D	23.1.1.2
45.1.1.0/24	ISIS-L2	15	20	D	35.1.1.5
172.16.0.0/24	O_ASE	150	1	D	23.1.1.2
172.16.2.0/24	O_ASE	150	1	D	23.1.1.2

[R4]dis ip routing-table
Route Flags: R - relay, D - download to fib

Routing Tables: Public
Destinations : 17 Routes : 17

Destination/Mask	Proto	Pre	Cost	Flags	NextHop
10.0.0.0/24	ISIS-L2	15	10	D	45.1.1.5
10.0.1.0/24	ISIS-L2	15	10	D	45.1.1.5
10.0.2.0/24	ISIS-L2	15	10	D	45.1.1.5
23.1.1.0/24	OSPF	10	2	D	24.1.1.2
35.1.1.0/24	ISIS-L2	15	20	D	45.1.1.5
172.16.0.0/24	ISIS-L2	15	20	D	45.1.1.5
172.16.2.0/24	ISIS-L2	15	20	D	45.1.1.5

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- After OSPF and IS-IS on R3 and R4 import routes from each other, sub-optimal routes exist from R4 to the network segment 172.16.X.0/24. This is because R3 distributes OSPF routes to the IS-IS routing domain first. Therefore, R4 learns routes destined for 172.16.X.0/24 from both OSPF and IS-IS. The preference value of OSPF external routes is 150, and that of IS-IS routes is 15. Therefore, R4 selects the IS-IS route destined for 172.16.X.0/24, which is a sub-optimal route.



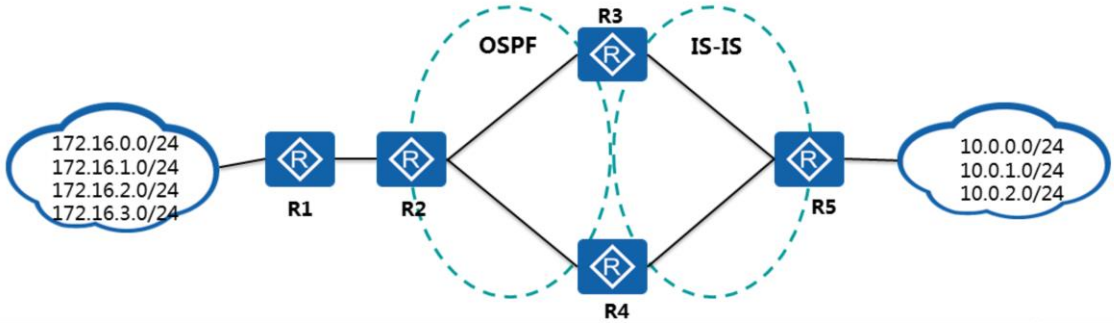
Contents

1. Rout Filter
2. Routing Policy and PBR
3. **Policy Routing and PBR Configuration Commands**
   - Configuring Route Import
   - **Configuring Policy-based Route Optimization**
   - Configuring PBR
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuring a Routing Policy for Route Optimization (1/3)

- Company A wants to optimize its network, and has the following requirements:
  - No routing loop occurs in the event of a network change.
  - R4 can access the network segment 172.16.X.0/24 through the optimal route.

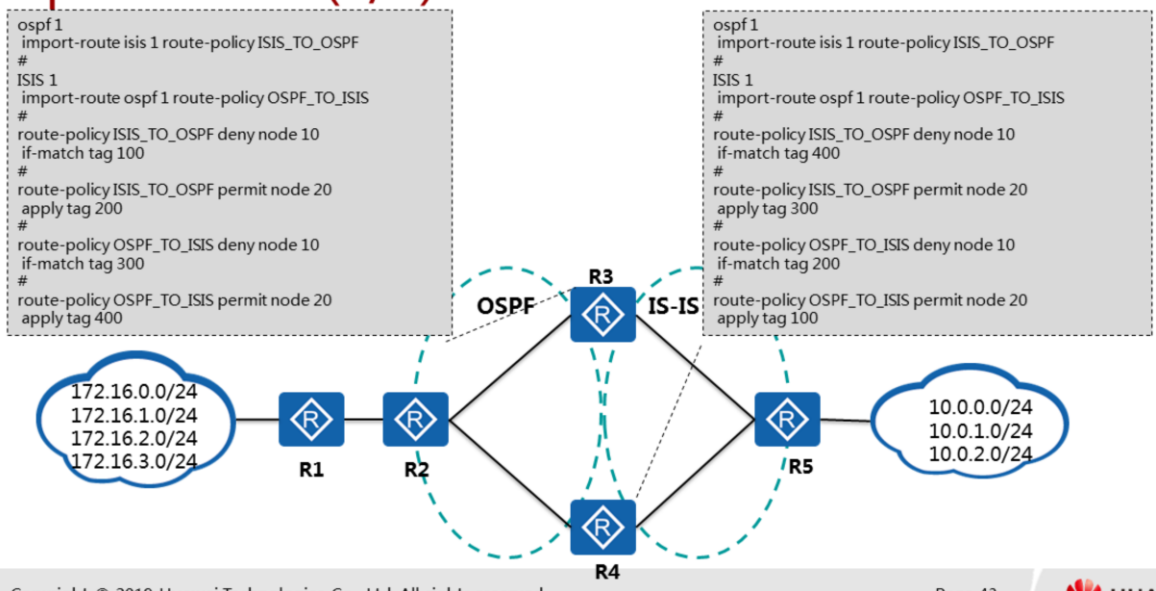


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- These requirements are expanded on the basis of those in the previous case. Perform configurations based on those in the previous case.
  - The route for R4 to access the network segment 172.16.X.0/24 needs to be modified to prevent a sub-optimal route that passes through the IS-IS domain.
  - The tag can be used to control OSPF and IS-IS from importing routes from each other, thereby preventing routing loops.

Configuring a Routing Policy for Route Optimization (2/3)

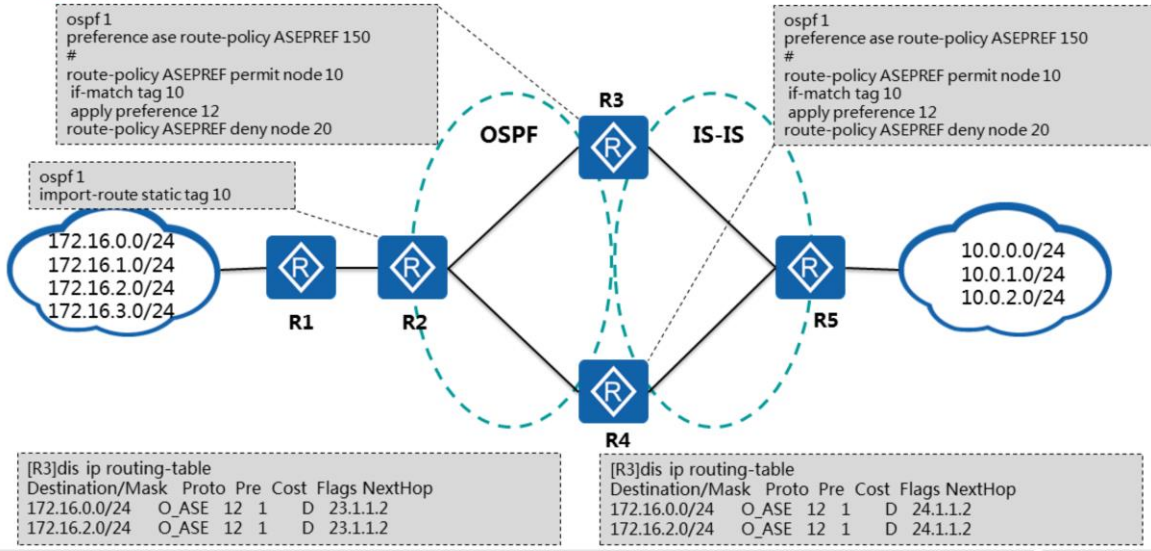


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- If routes to be imported are not filtered, routing loops may occur on the network when the network changes. To prevent routing loops, ensure that only the routes of each routing domain are imported when routing protocols import routes from each other. In the preceding configuration scenario, the tag is used to control route import. When the tag is used, no routing entry needs to be specified. When the number of routing entries in the routing domain increases or decreases, the tag value of imported routes changes without manual intervention. This offers good scalability.
- Although configuring a routing policy in the preceding scenario can prevent routing loops, it cannot solve the problem of sub-optimal routes.

Configuring Routing Policy for Route Optimization (3/3)



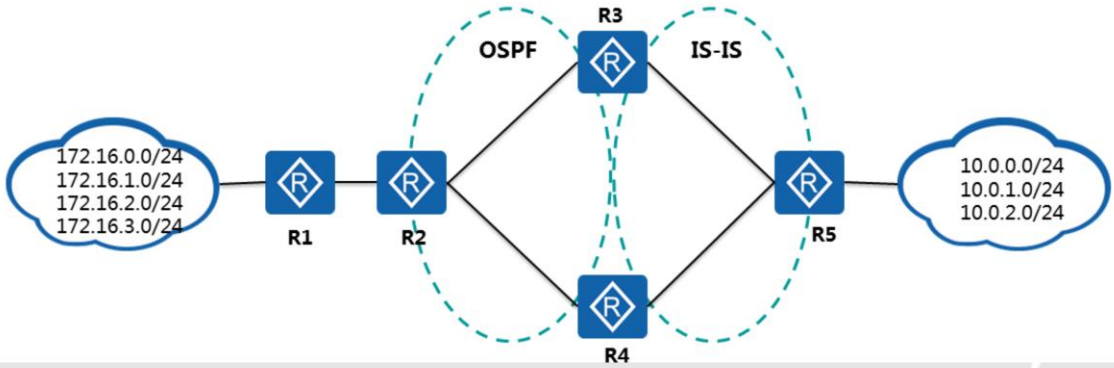
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Sub-optimal routes are generated mainly because R3 or R4 obtains routes destined for 172.16.X.0/24 from both the OSPF and IS-IS domains when importing routes from each other. The preference value of OSPF external routes is greater than that of IS-IS routes (a smaller preference value indicates a higher priority). As a result, R3 or R4 selects the sub-optimal route. To solve this problem, change the preference value of OSPF external routes. This issue is addressed as long as the preference value of the OSPF_ASE routes is smaller than that of the IS-IS routes.
- You are not advised to set the preference value of the OSPF_ASE routes to be smaller than the preference value (10) of OSPF internal routes.

Optimizing Routes (1/3)

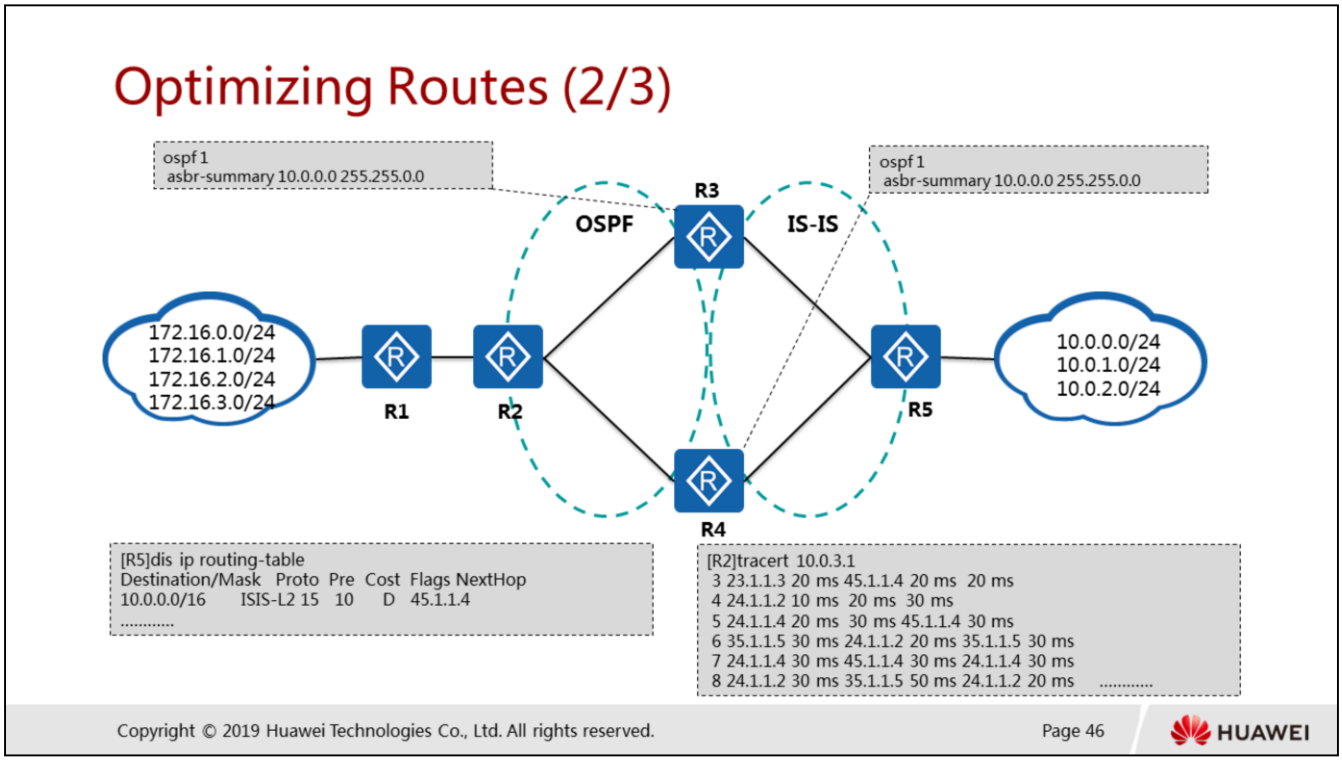
- Company A wants to optimize its network, and has the following requirements:
  - R3 and R4 are in an OSPF domain. The routes destined for 10.0.X.0/24 are summarized as 10.0.0.0/16. No routing loop occurs.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- These requirements are expanded on the basis of those in the previous case. Perform configurations based on those in the previous case.

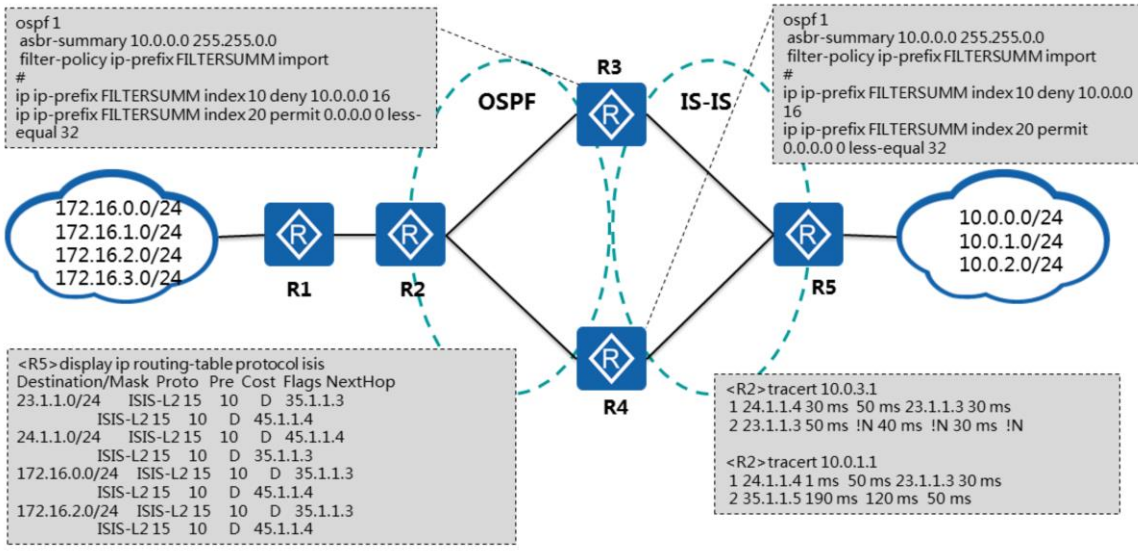


- When only route summarization is performed, two problems exist. The first problem is that R5 learns summary routes. The second problem is that a routing loop occurs when a nonexistent IP address is pinged from R2.
- The cause of the first problem is that R3 and R4 learn summary routes from each other and then import the summary routes to the IS-IS domain. OSPF summarization is first performed on R3. The generated summary route is then transmitted to R4 through R2. R4 import this summary route to IS-IS and advertises it to R5.
- Then come to the second problem. Two equal-cost routes destined for 10.0.0.0/16 exist on R2, and their next hops are R3 and R4 respectively. When the tracert destination port number changes, the tracert packet is sent to R3 or R4.
  - When the tracert packet is sent to R4: R4 performs OSPF route summarization later than R3. In this case, R4 has only one OSPF summary route advertised by R3. The next hop of the route from R4 to 10.0.0.0/16 is R2. As a result, a routing loop occurs.
  - When the tracert packet is sent to R3: After an OSPF summary route is generated on R4, R4 advertises this summary route to R3. After an OSPF summary route is generated on R3, R3 advertises this summary route to R4. R4 then imports this summary route to IS-IS and then advertises it to R3 through R5. Finally, R3 has two routes with 16-bit subnet masks. R3 compares the routing protocol of these two routes and selects the IS-IS route with a higher priority and R4 as the next hop. Since R4 performs route summarization later than R3, R4 has the OSPF summary route advertised by R3. The next hop of the route from R4 to 10.0.0.0/16 is R2. As a

result, a routing loop occurs.

- To solve the problems mentioned above, ensure that R3 and R4 cannot learn summary routes from each other and cannot import summary routes to the IS-IS routing domain. Therefore, you only need to filter out the summary routes on R3 and R4 learned from each other.

Optimizing Routes (3/3)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Create filtering policies on R3 and R4 to prevent them from receiving specified summary routes from OSPF. This ensures that the summary route will not be imported to the IS-IS routing domain again and loops are avoided.



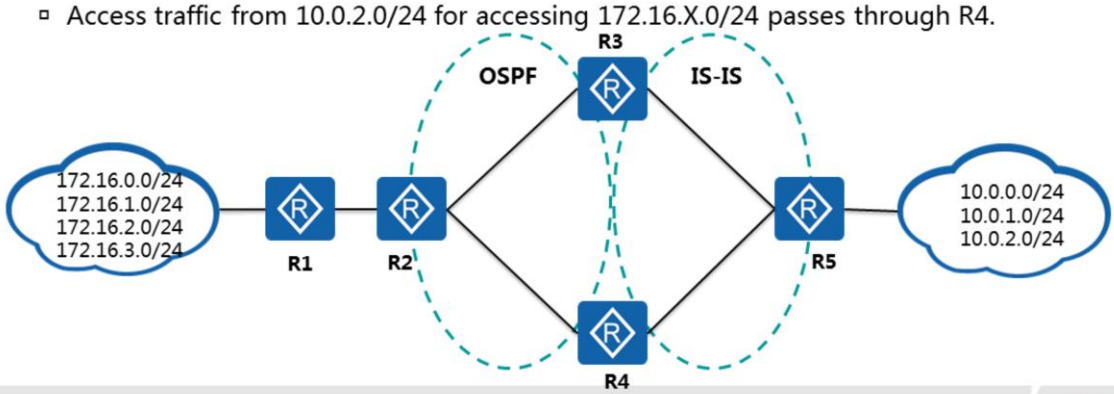
Contents

1. Rout Filter
2. Routing Policy and PBR
3. **Policy Routing and PBR Configuration Commands**
   - Configuring Route Import
   - Configuring Policy-based Route Optimization
   - **Configuring PBR**
4. Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuring PBR (1/2)

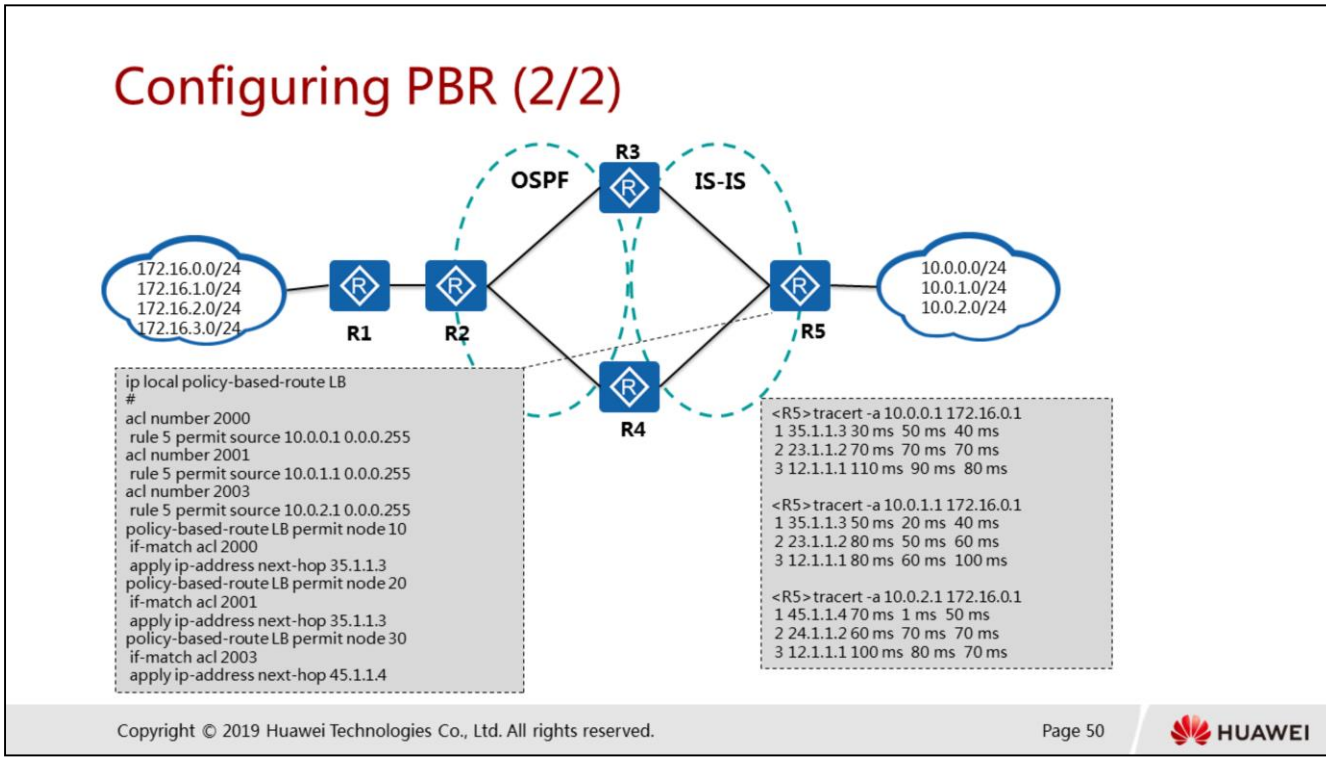
- Company A wants to optimize its network, and has the following requirements:
  - Access traffic from 10.0.0.0/24 and 10.0.1.0/24 to 172.16.X.0/24 passes through R3.
  - Access traffic from 10.0.2.0/24 for accessing 172.16.X.0/24 passes through R4.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- These requirements are expanded on the basis of those in the previous case. Perform configurations based on those in the previous case.



- Command Usage
  - The **policy-based-route** command creates or modifies a policy and a PBR node.
  - The **ip local policy-based-route** command enables local PBR.
- Parameter Description
  - policy-based-route *policy-name* { permit | deny } node *node-id*
    - policy-name: specifies the name of a policy.
    - Permit: indicates a PBR mode in which PBR is enabled for matched packets.
    - Deny: indicates a PBR mode in which PBR is disabled for matched packets.
    - node-id: specifies the sequence number of a PBR node.
  - ip local policy-based-route *policy-name*
    - policy-name: specifies the name of a local policy.
- Precautions
  - If the outbound interface needs to be specified for packets when PBR is configured, the outbound interface cannot be a broadcast interface such as an Ethernet interface.
- Test Result
  - When different source addresses are specified on R5 to trace the packets with the

same destination, it is found that the packets are forwarded along different paths.

- Note: The routing policy applied in the **ip local policy-based-route** command takes effect only on the data packets originated on the local router.

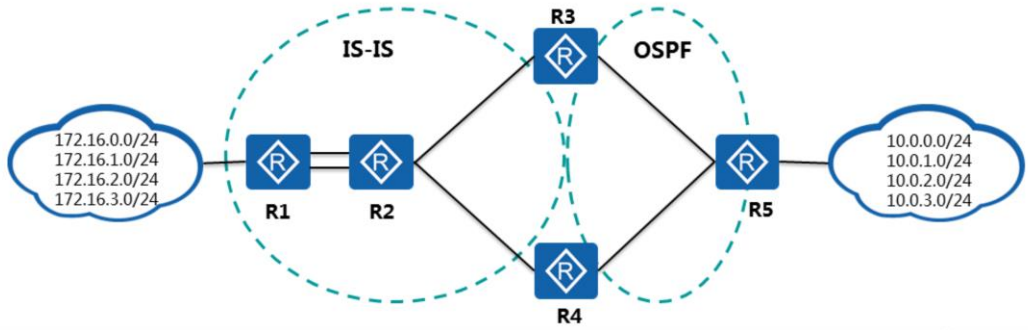
Contents

1. Rout Filter
2. Routing Policy and PBR
3. Policy Routing and PBR Configuration Commands
4. **Case Analysis**
   - **Case 1**
   - Case 2

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Case 1

- The following figure shows the network topology of company B. The company has the following requirements:
  - R3, R4, and R5 belong to OSPF area 0. Only routes destined for 10.0.X.0/24 are imported to the OSPF area.
  - On R5, routes destined for 10.0.X.0/24 are summarized into a summary route destined for 10.0.0.0/16. R3 and R4 can deliver default routes to the OSPF area.
  - No loop exists on the network.



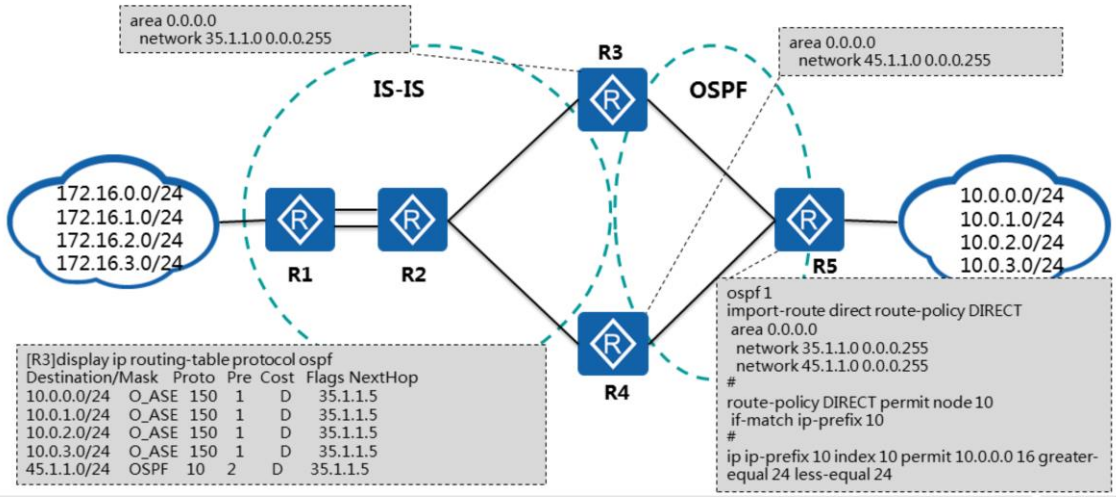
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In this case, the addresses for interconnecting devices are as follows:
  - For example, if RTX is interconnected with RTY, the interconnection addresses are XY.1.1.X and XY.1.1.Y, and the mask length is 24 bits.

Case 1: Requirement 1

- R3, R4, and R5 belong to OSPF area 0. Only routes destined for 10.0.X.0/24 are imported to the OSPF area.



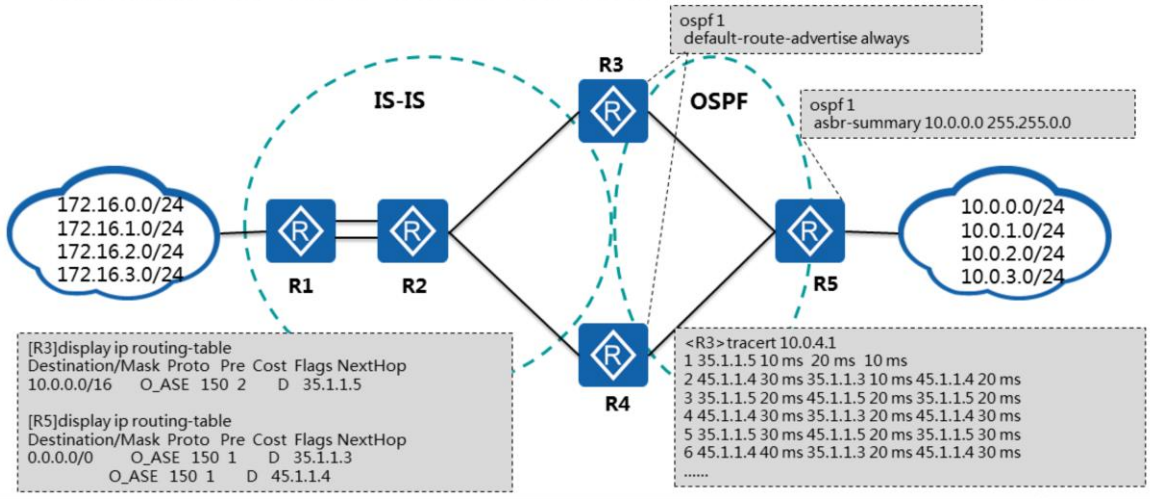
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Note that accurate matching is required when routes are imported to R5.

Case 1: Requirement 2

- On R5, routes destined for 10.0.X.0/24 are summarized into a summary route destined for 10.0.0.0/16. R3 and R4 can deliver default routes to the OSPF area.



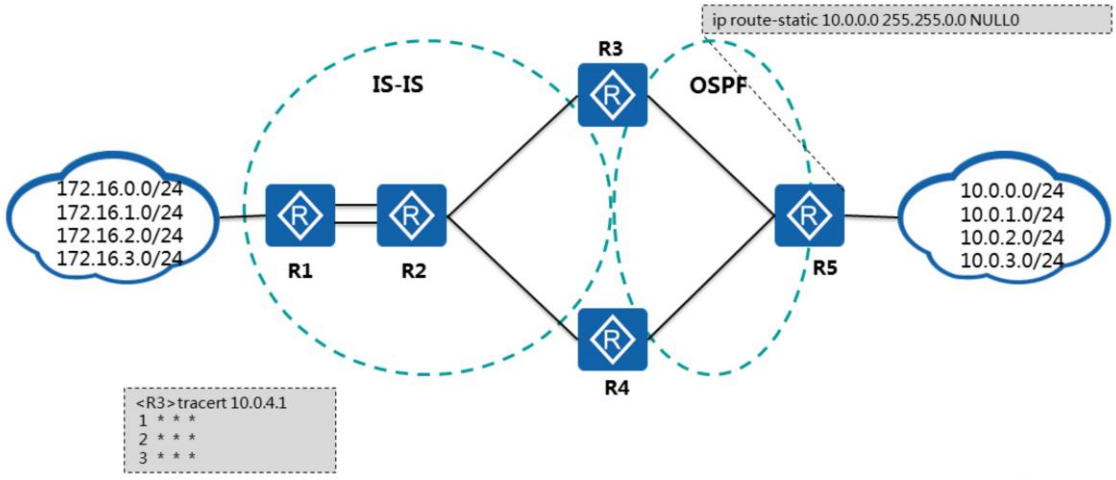
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A loop occurs when the **tracert** command is run to trace a nonexistent IP address on the network segment 10.0.0.0/16. This loop occurs because no route pointing to Null 0 is automatically generated when the OSPF summary route is generated.

Case 1: Requirement 3

- No loop occurs on the network.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- To eliminate loops, use a command on R5 to configure a static route pointing to Null 0.

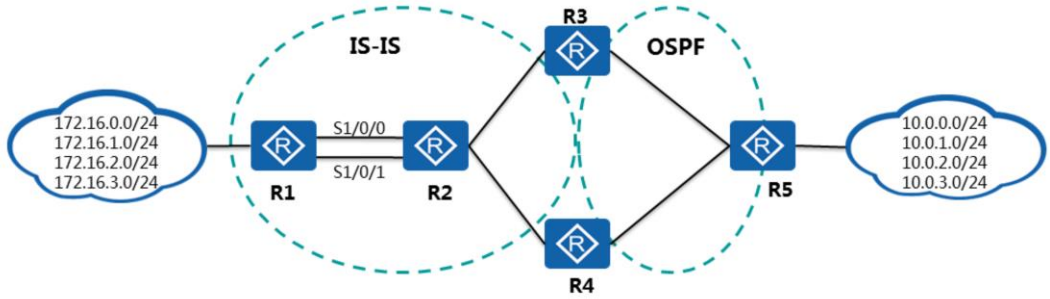
Contents

1. Rout Filter
2. Routing Policy and PBR
3. Policy Routing and PBR Configuration Commands
4. **Case Analysis**
   - Case 1
   - **Case 2**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Case 2

- Company B wants to expand its network based on the original configurations. The requirements are as follows:
  - Basic IS-IS configurations have been completed.
  - Based on a filtering policy, routes destined for 172.16.0.0/24 and 172.16.2.0/24 are imported to IS-IS.
  - R3 and R4 import routes from each other. No loop occurs and no sub-optimal route exists.
  - Access traffic from 172.16.0.0/24 and 172.16.2.0/24 to the OSPF network passes through S1/0/0.



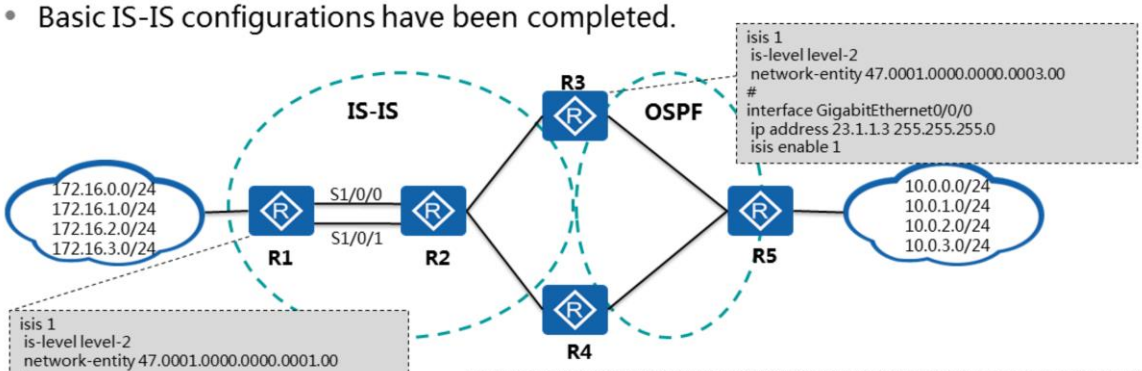
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- These requirements are expanded on the basis of those in the previous case. Perform configurations based on those in the previous case.
- In this case, the addresses for interconnecting devices are as follows:
  - For example, if RTX is interconnected with RTY, the interconnection addresses are XY.1.1.X and XY.1.1.Y, and the mask length is 24 bits.
    - The IP address of S1/0/0 is 12.1.1.1/24 on R1 and 12.1.1.2/24 on R2. The IP address of S1/0/1 is 21.1.1.1/24 on R1 and 21.1.1.2/24 on R2.

Case 2: Pre-configurations

- Basic IS-IS configurations have been completed.




```
isis 1
 is-level level-2
 network-entity 47.0001.0000.0000.0003.00
#
interface GigabitEthernet0/0/0
 ip address 23.1.1.3 255.255.255.0
 isis enable 1
```

```
isis 1
 is-level level-2
 network-entity 47.0001.0000.0000.0001.00
#
interface Serial1/0/0
 link-protocol ppp
 ip address 12.1.1.1 255.255.255.0
 isis enable 1
#
interface Serial1/0/1
 link-protocol ppp
 ip address 21.1.1.1 255.255.255.0
 isis enable 1
```

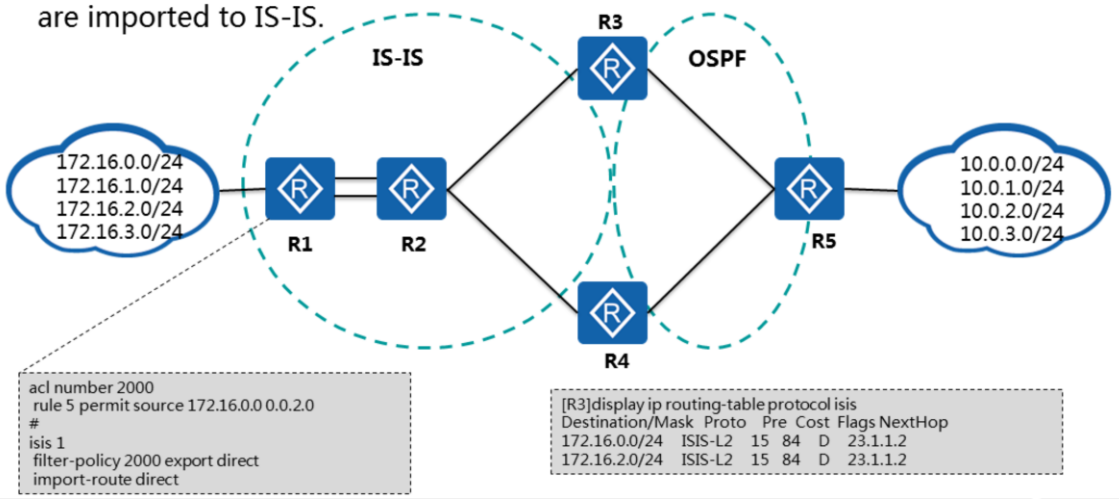
```
[R2]display isis peer
                Peer information for ISIS(1)
 System Id     Interface   Circuit Id      State HoldTime Type  PRI
------------------------------------------------------------------------------
0000.0000.0001 S1/0/0      0000000001      Up   22s       L2  --
0000.0000.0001 S1/0/1      0000000002      Up   24s       L2  --
0000.0000.0003 GE0/0/0     0000.0000.0002.01 Up  28s       L2  64
0000.0000.0004 GE0/0/1     0000.0000.0004.01 Up  7s        L2  64
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Case 2: Requirement 1

- Based on a filtering policy, routes destined for 172.16.0.0/24 and 172.16.2.0/24 are imported to IS-IS.

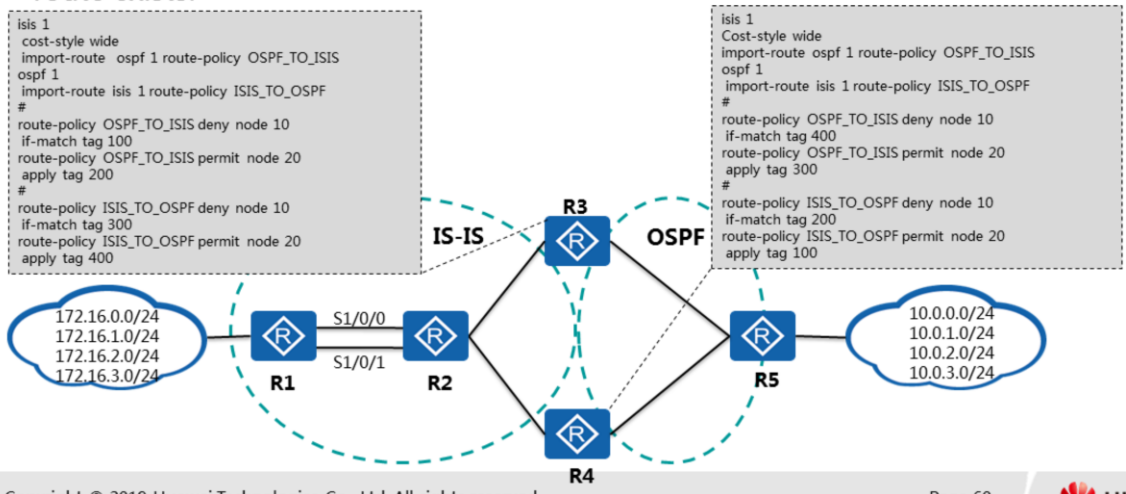


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Use a filtering policy with an ACL to import the routes destined for the two network segments specified in the requirement to IS-IS.
- Note: To filter imported routes of a routing protocol using a filtering policy, use the **filter-policy export** command.

Case 2: Requirement 2 (1/2)

- R3 and R4 import routes from each other. No loop occurs and no sub-optimal route exists.

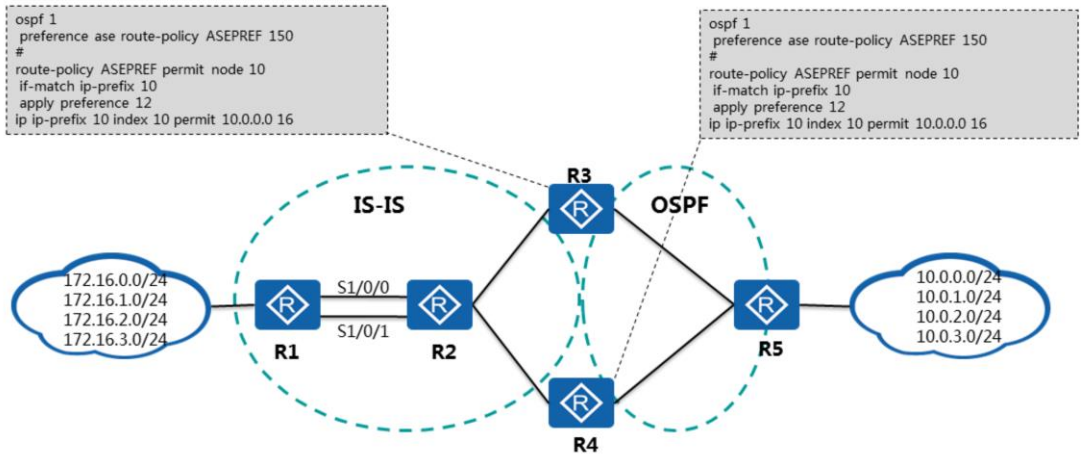


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- In this case, the tag can be added to routes during route importing to prevent routing loops. If the IS-IS protocol needs to support the tag, the cost type must be wide; otherwise, IS-IS routes do not carry the tag.
- The tag is used to prevent routing loops but cannot prevent generation of sub-optimal routes. To prevent generation of sub-optimal routes, change the preference value of corresponding routes.

Case 2: Requirement 2 (2/2)

- R3 and R4 import routes from each other. No loop occurs and no sub-optimal route exists.



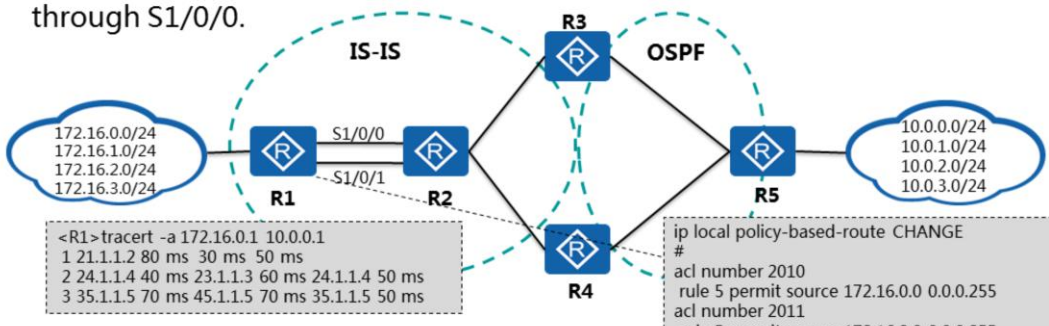
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The configurations in this example prevent generation of sub-optimal routes on the network segment 10.0.0.0/16 on R3 and R4. The route import speeds are inconsistent on R3 and R4 are different. As a result, R3 or R4 will learn routes destined for 10.0.0.0/16 from both IS-IS and OSPF at the same time. If R3 imports routes first, R4 will learn the routes destined for 10.0.0.0/16 from both IS-IS and OSPF at the same time. When selecting routes, R4 compares preference values of these routes. The preference value of OSPF external routes is 150, and that of IS-IS routes is 15. Therefore, R4 selects the route destined for 10.0.0.0/16 through the IS-IS domain. This route is a sub-optimal route. On R4, change the preference value of OSPF external routes destined for 10.0.0.0/16 to be smaller than the preference value of IS-IS routes. In this way, sub-optimal routes are eliminated. It is recommended that the preference value of OSPF external routes be greater than the internal preference value (10) of OSPF internal routes.

Case 2: Requirement 3

- Access traffic from 172.16.0.0/24 and 172.16.2.0/24 to the OSPF network passes through S1/0/0.




```
<R1>tracert -a 172.16.0.1 10.0.0.1
 1 21.1.1.2 80 ms  30 ms  50 ms
 2 24.1.1.4 40 ms 23.1.1.3 60 ms 24.1.1.4 50 ms
 3 35.1.1.5 70 ms 45.1.1.5 70 ms 35.1.1.5 50 ms
```

```
<R1>tracert -a 172.16.2.1 10.0.0.1
 1 21.1.1.2 60 ms  30 ms  30 ms
 2 24.1.1.4 60 ms 23.1.1.3 60 ms 24.1.1.4 40 ms
 3 35.1.1.5 60 ms 45.1.1.5 70 ms 35.1.1.5 60 ms
```

```
ip local policy-based-route CHANGE
#
acl number 2010
 rule 5 permit source 172.16.0.0 0.0.0.255
acl number 2011
 rule 5 permit source 172.16.2.0 0.0.0.255
#
policy-based-route CHANGE permit node 10
 if-match acl 2000
 apply output-interface Serial1/0/0
policy-based-route CHANGE permit node 20
 if-match acl 2001
 apply output-interface Serial1/0/1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Quiz

1. Which of the following are involved in routing control? ( )
   A. Route advertisement
   B. Route acceptance
   C. Filtering and control of imported routes
   D. Setting attributes of specified routes
2. What does the prefix list configured using the ip ip-prefix Prefix1 permit 160.0.0.0 8 command mean? ( )
   A. The first three bits of the prefix must be "101" and the mask length must be between 8 and 32.
   B. The first three bits of the prefix must be "101" and the mask length must be 8.
   C. The prefix number must be "160" and the mask length must be between 8 and 32.
   D. The prefix number must be 160 and the mask length must be 8.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- ABCD
- D

Quiz

3. Which statements about routing policy are true? ( )
   A. A routing policy is composed of **if-match** clauses or **if-match** and **apply** clauses.
   B. A routing policy can have multiple nodes. If a route matches all the conditions in one node of a routing policy, the route matches the routing policy.
   C. A routing policy can have multiple nodes. If a route matches any condition in one node of a routing policy, the route matches the routing policy.
   D. If a route matches the first node in a routing policy and is denied and it also matches the second node, the action defined in the second node is also performed.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- AB



Summary

- ACL
- IP Prefix List
- AS-Path Filter
- Community Filter
- Routing Policy
- PBR

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank you

www.huawei.com

HCIE-Routing & Switching

— MLD

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- Multicast Listener Discovery (MLD) enables each IPv6 router to discover multicast listeners (that is, the nodes that expect to receive multicast data) on its directly connected network and identify the multicast addresses that the neighbor nodes are interested in.
- This document describes the fundamentals and configuration commands of various MLD versions.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to:
  - Understand the functions and fundamentals of multicast listener discovery (MLD).
  - Understand the differences between MLDv1 and MLDv2.
  - Understand the functions of source-specific multicast (SSM) mapping.
  - Master the basic configuration commands of MLD.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

- MLD Overview
- MLDv1
- MLDv2
- SSM Mapping

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MLD Overview

- MLD manages IPv6 multicast group members.
- MLD has two versions:
  - MLDv1 defined in RFC 2710: corresponds to IGMPv2.
  - MLDv2 defined in RFC 3810: corresponds to IGMPv3.
- MLDv1 uses a query-report mechanism to manage IPv6 multicast group members. MLDv2 is similar to MLDv1 and has the following new features:
  - IPv6 multicast source filtering
  - IPv6 multicast group status tracking
  - Receiver host status listening

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- MLD manages IPv6 multicast group members, and its fundamentals and functions are similar to those of IGMP. MLD enables each IPv6 router to discover multicast listeners (that is, the nodes that expect to receive multicast data) on its directly connected network and identify the multicast addresses that the neighbor nodes are interested in. The messages are then offered to the multicast routing protocol used by the router to ensure that multicast data is forwarded to all links where receivers exist.
- MLD is an asymmetric protocol that specifies the behaviors of multicast listeners and routers. For the multicast address that a router is listening to, the router acts as the protocol's two roles, including responding to its own message.
- If a router has more than one interface on the same network, it only needs to run the protocol on one of the interfaces. Additionally, listeners must run this protocol on all interfaces so that upper-layer protocols receive required multicast data from the interfaces.
- Both MLD versions support the any-source multicast (ASM) model. MLDv2 can be independently used in the SSM model, whereas MLDv1 must be used with SSM mapping.

MLD Functions

- MLD enables IPv6 routers to discover multicast listeners on their directly connected network segments.
- Routers can use MLD to check whether multicast group listeners exist on their directly connected network segments and record the corresponding information in their databases. Routers also maintain information about the timers associated with these multicast addresses.
- Routers with MLD configured use IPv6 unicast link-local addresses as source addresses to send MLD messages.
- MLD uses ICMP to carry messages. All MLD messages are limited on the local link, and the number of hops is 1.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A multicast listener is a host that wants to receive multicast data.

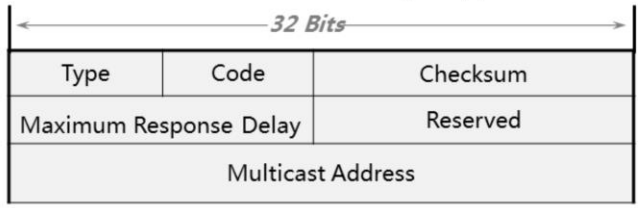
Contents

- MLD Overview
- MLDv1
- MLDv2
- SSM Mapping

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MLDv1 Message Format

- Type: There are three types of MLD messages.
  - Multicast Listener Query message (type value = 130), which can be classified into the following sub-types:
    - General Query message: used to obtain the multicast addresses of listeners on a connected network.
    - Multicast-Address-Specific Query message: used to obtain a listener for a specific multicast address on a connected network.
  - Multicast Listener Report message (type value = 131)
  - Multicast Listener Done message (type value = 132)

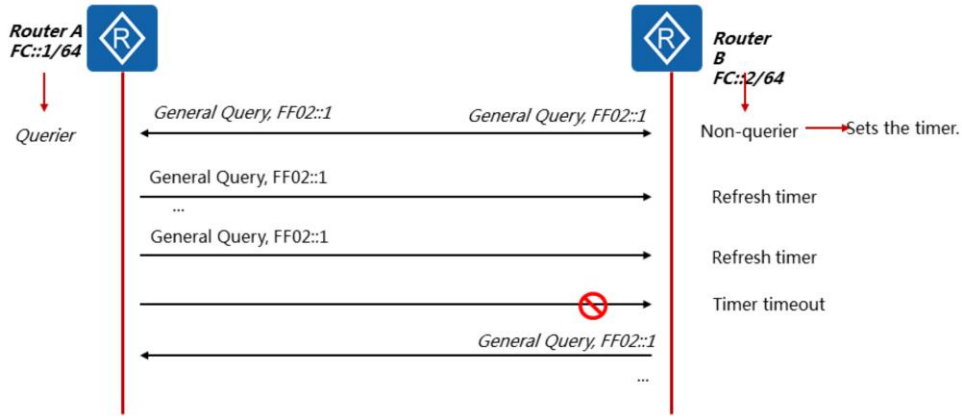


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Type: There are three types of MLD messages.
  - Multicast Listener Query message (type value = 130), which can be classified into the following sub-types:
    - General Query message: used to obtain the multicast addresses of listeners on a connected network.
    - Multicast-Address-Specific Query message: used to obtain a listener for a specific multicast address on a connected network.
  - Multicast Listener Report message (type value = 131)
  - Multicast Listener Done message (type value = 132)
- Code
  - Set to 1 during transmission or ignored during reception.
- Checksum
  - Standard ICMPv6 checksum, covering the entire MLD message plus a pseudo header of IPv6 header fields
- Maximum Response Delay
  - Maximum delay for sending a response message, in milliseconds. It is valid only in query messages. In other messages, this field is set to 0 during transmission or is ignored during reception.
- Reserved
  - Set to 0 for senders or ignored for receivers.
- Multicast Address
  - If a General Query message is sent, the multicast address is set to 0. If a Group-Specific Query message is sent, the multicast address is set to a specific IPv6 multicast address. In a Report or Leave message, the multicast address is set to a specific IPv6 multicast address that a sender needs to listen to or stops listening to.

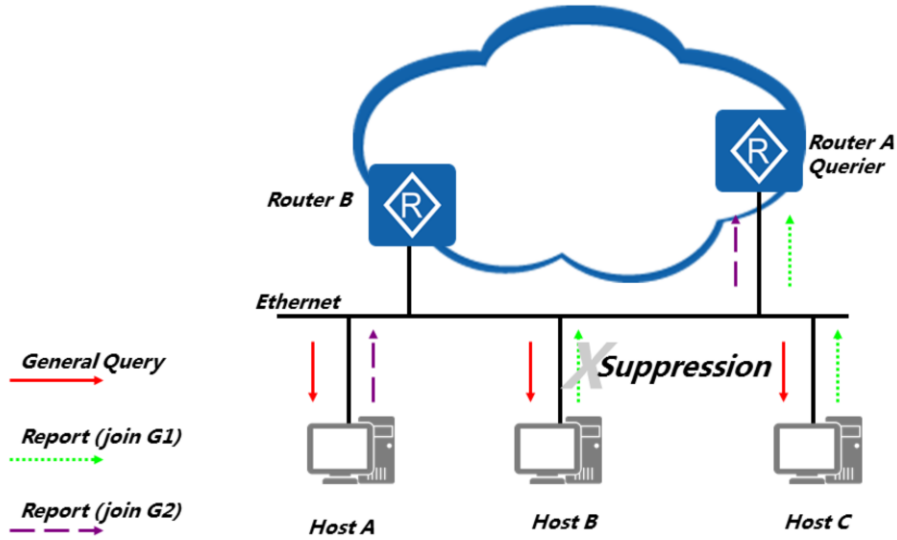
Querier Election Mechanism



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- 1. Each MLDv1 router considers itself as a querier when it starts and sends a General Query message with destination address FF02::1 to all hosts and routers on the local network segment.
- 2. When other routers receive a General Query message, they compare the source IPv6 address of the message with their own interface IP addresses. The router with the smallest IPv6 address becomes the querier, and the other routers are non-queriers.
- 3. All non-queriers start a timer (Other Querier Present Timer). If non-queriers receive a Query message from the querier before the timer expires, they reset the timer. If non-queriers receive no Query message from the querier when the timer expires, they trigger election of a new querier.

Group Member Join Process



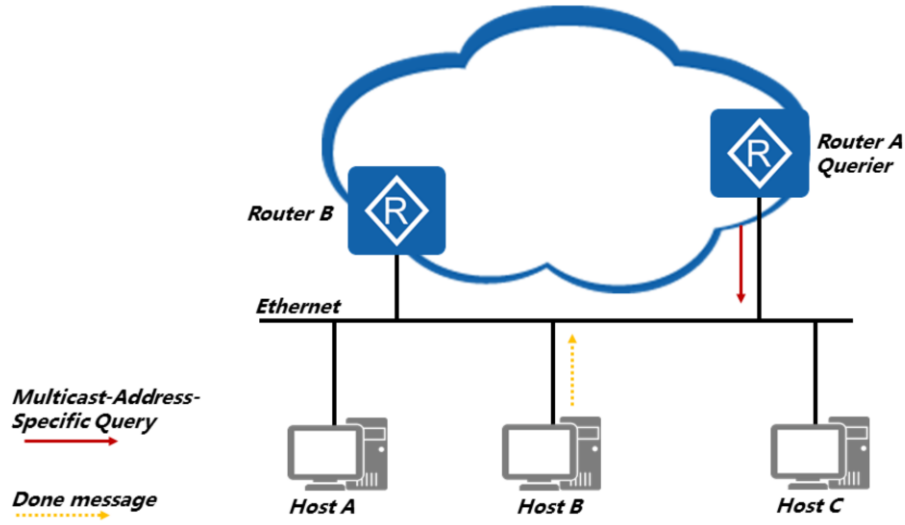
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The VRP implements MLDv1 according to RFC 2710. MLDv1 manages multicast group members based on the query/response mechanism.
- MLDv1 has two types of query messages:
  - General Query message: used to query whether there is any listener of a multicast group on a direct link.
  - Multicast-Address-Specific Query message: used to query whether there is any listener of a specified multicast address on a direct link.
- If multiple multicast routers with MLD configured exist on the shared network segment, the querier election mechanism is triggered. The router with the smallest IPv6 address on the network segment functions as the querier (also called MLD querier), and other routers function as non-queriers.
- The basic process that a host joins a multicast group is as follows (General Query messages are used as an example):
  - 1. The MLD querier periodically sends a General Query message with destination address FF02::1 to all link-local hosts on the shared network segment in multicast mode.
  - 2. All hosts on the network segment receive the General Query message. If Host B and Host C want to join the multicast group G1, set a timer delay to respond.
  - 3. After the timer expires, the host that wants to join the multicast group sends a Report message to all hosts and routers on the network segment in multicast mode to respond to the query message. This Report message contains the address of G1.
  - 4. After receiving the Report message, all hosts and routers on the network segment obtain the multicast information about G1. In this case, other hosts that want to join the multicast group G1 do not send the same Report message. If Host A wants to join another multicast group G2, it sends a Report message containing the G2 address to respond to the General Query message.
  - 5. After the query/report process is complete, the MLD querier can learn whether receivers of G1 exist on its directly connected network segment and generates (*, G1) multicast

routing entries, where * indicates any multicast source.

- 6. Through the multicast routing mechanism, the MLD querier receives multicast information from multicast sources. If there are receivers on the directly connected network segment, the data is forwarded on the network segment, and the hosts that join the multicast group receive the data.

Group Member Leave Process



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- If a host wants to leave the multicast group, it sends a Done message to the link using the multicast address (destination address FF02::2) and carries the address that it needs to stop listening to in the multicast address field.
- When the querier receives the Done message from the link, if the address of the multicast group that the host wants to leave is in the listener address list of the querier on the link, the querier sends Multicast-Address-Specific Query messages of Last Listener Query Count. The interval is Last Listener Query Interval. Generally, Last Listener Query Interval is set to Maximum Response Delay in Multicast-Address-Specific Query messages. If the last query response delay expires and no Report message containing the multicast address is sent to the querier on the link, the address is deleted from the listener address list.

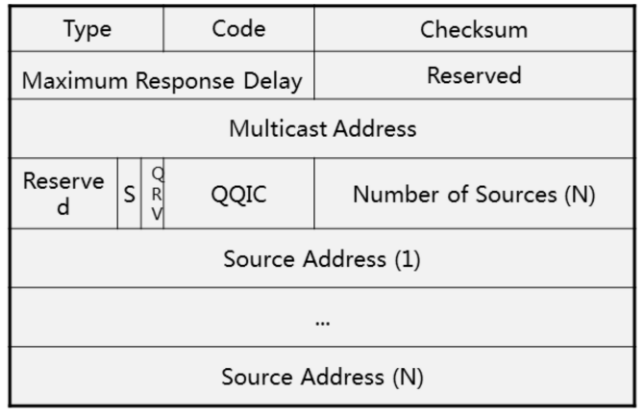


Contents

1. MLD Overview
2. MLDv1
3. **MLDv2**
4. SSM Mapping

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MLDv2 Query Message Format

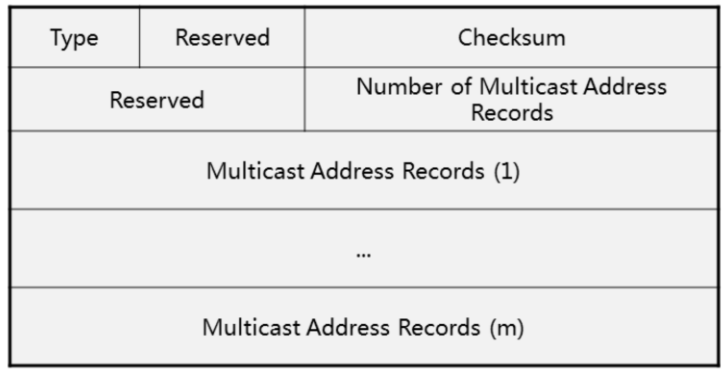


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The first 192 bits in an MLDv2 message is the same as that in an MLDv1 message.
- Flag S (Suppress Router-side Processing):
  - Indicates whether a router suppresses the timer update after receiving a Query message.
- Querier's Robustness Variable (QRV):
  - The QRV is the default value of Last Listener Query Count, that is, the number of times that a router sends Multicast-Address-Specific Query messages before determining that no remaining listener exists.
- Querier's Query Interval Code (QQIC):
- Number of Sources:
  - This field is set to 0 in General Query or Multicast-Address-Specific Query messages.
  - This field indicates the number of source addresses contained in a Group-Source-Specific Query message.
- Source Address:
  - Multicast source address

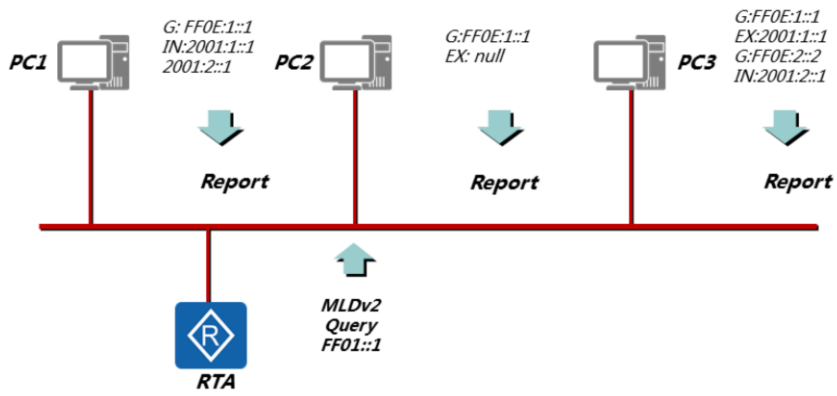
MLDv2 Report Message Format



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- A host sends an MLD Report message to report the current multicast listening status.
- Type:
  - Type = 143
- Reserved:
  - Set to 0 during transmission or ignored during reception.
- Checksum:
  - Standard ICMPv6 checksum, covering the entire MLD message plus a pseudo header of IPv6 header fields
- Number of Multicast Address Records
- Multicast Address Records:
  - Indicates information about each multicast address listened by a host on an interface. The information includes the record type, multicast address, and source address.

Group-Source-Specific Query



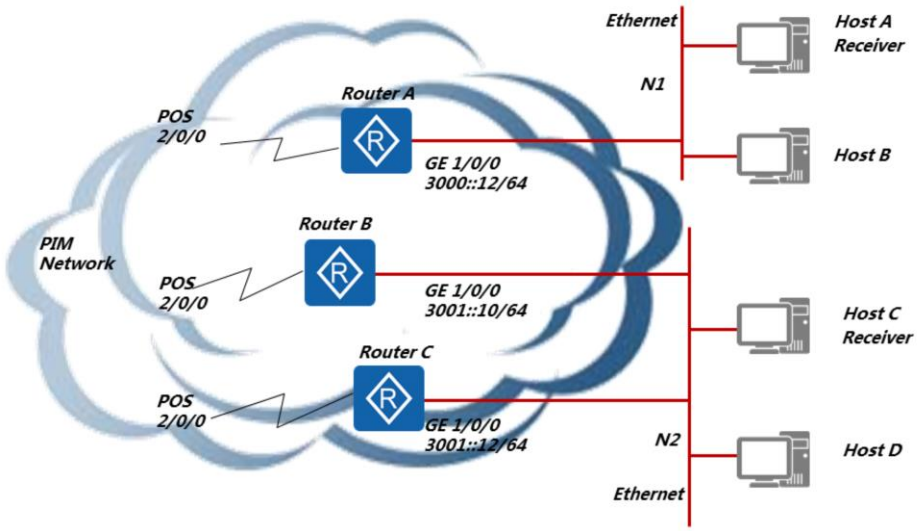
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- MLDv2 is compatible with MLDv1. The fundamentals of MLDv2 are the same as those of MLDv1. MLDv2 supports source lists and filter modes. You can specify source addresses to join a multicast group, implementing SSM.
- IPv6 multicast source filtering: Besides the group-specific query, MLDv2 adds the following filter modes for multicast sources: Include or Exclude.
  - When a host joins a multicast group, if the host only needs to receive multicast packets from specified sources, such as S1 and S2, Include Sources (S1, S2, ...) can be set in MLD Report messages.
  - When a host joins a multicast group, if the host does not want to receive multicast packets from specified sources, such as S1 and S2, Exclude Sources (S1, S2, or ...) can be set in MLD Report messages.
- IPv6 multicast group status tracking: Multicast routers running MLDv2 keep IPv6 multicast group state based on per multicast address per attached link. The IPv6 multicast group state includes:
  - Filter mode: The MLD querier tracks the Include or Exclude state.
  - Source list: The MLD querier tracks the sources that are added or deleted.
  - Timers: include a filter timer when the MLD querier switches to the Include mode after its IPv6 multicast address expires and a source timer about source records.
- Receiver host status listening: Multicast routers running MLDv2 listen to the receiver host status to record and maintain information about hosts that join IPv6 multicast groups on

the local network segment.

MLD Configuration Example

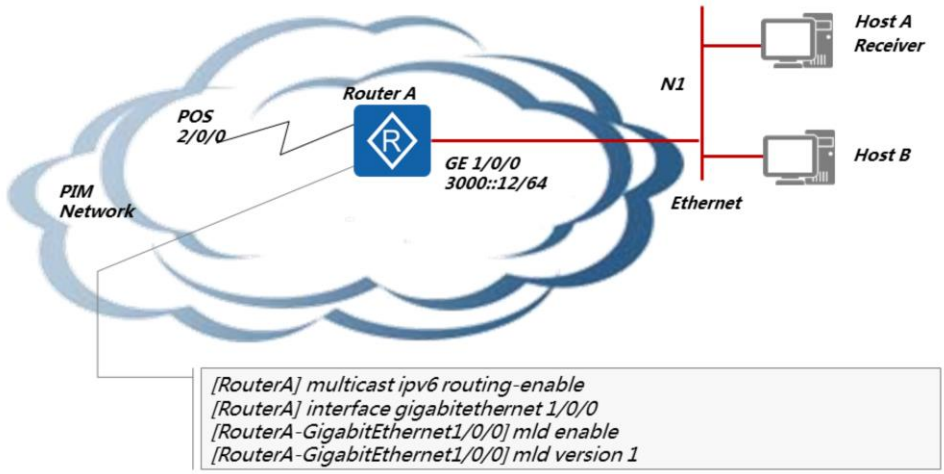


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Receivers receive video on demand (VoD) information in multicast mode. Receivers of different organizations form edge networks. Each edge network has one or more receiver hosts.
- Host A and Host C are multicast receivers on two edge networks. Router A on the PIM network connects to the edge network N1 through GE 1/0/0 and to another device on the PIM network through POS 2/0/0. Router B and Router C connect to the edge network N2 through their respective GE 1/0/0 interfaces, and to other devices on the PIM network through POS 2/0/0 interfaces.
- MLDv1 runs between Router A and the edge network N1.
- MLDv2 runs between Router B/Router C and the edge network N2.

Configuring Router A



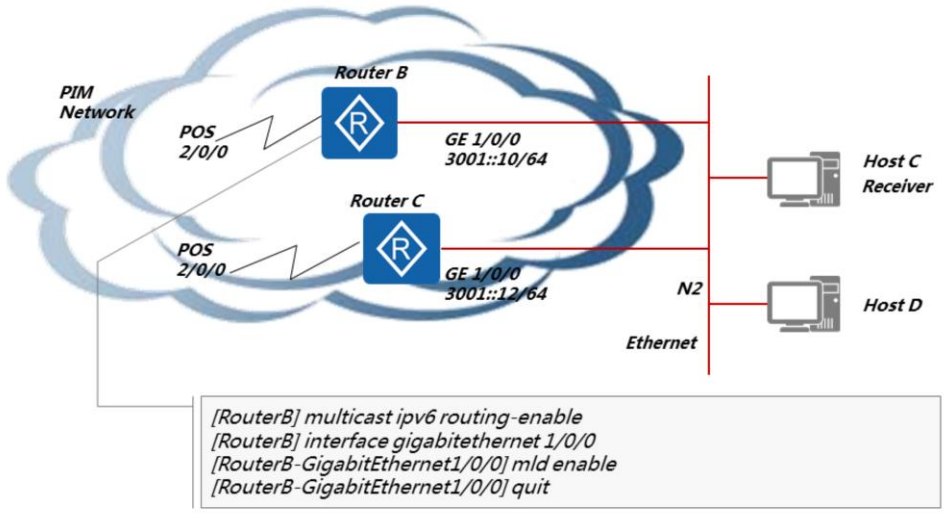
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Enter the system view.
  - **system-view**
- Enable IP multicast routing.
  - **multicast ipv6 routing-enable**
- Enter the interface view.
  - **interface** *interface-type interface-number*
- Enable MLD.
  - **mld enable**
  - MLD must be enabled on an interface of a router with MLD configured to listen to all IPv6 multicast addresses.
  - Enable MLD on the interfaces that need to establish and maintain multicast group memberships.
  - The querier periodically sends MLD Query messages on the directly connected network segment to maintain multicast listener information. When receiving a Report message from a group member, the multicast router updates the group member's information.
- Enter the MLD view.
  - **mld**
- Set an MLD version globally.
  - **version** { **1** | **2** }
- Enter the interface view.
  - **interface** *interface-type interface-number*
- Configure an MLD version on the interface.
  - **mld version** { **1** | **2** }
  - This configuration is optional. By default, MLDv2 is used.
  - If no MLD version is configured on an interface, the MLD version configured in the MLD view is used by default. If an MLD version is configured on an interface, the MLD version configured in the interface

view is preferred.

Configuring Router B




```
[RouterB] multicast ipv6 routing-enable
[RouterB] interface gigabitethernet 1/0/0
[RouterB-GigabitEthernet1/0/0] mld enable
[RouterB-GigabitEthernet1/0/0] quit
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- By default, MLDv2 is used.
- The configuration of Router C is similar to that of Router B.
- A querier needs to be elected on the network. Which router will be elected as a querier?

Displaying the Querier

```
[RouterB] display mld interface gigabitethernet 1/0/0
  GigabitEthernet1/0/0(FE80::200:5EFF:FE66:5100):
    MLD is enabled
    Current MLD version is 2
    MLD group policy: none
    Value of query interval for MLD(in seconds): 125
    Value of other querier timeout for MLD(in seconds): 0
    Value of maximum query response time for MLD(in seconds): 10
    Querier for MLD: FE80::200:5EFF:FE66:5100 (this router)
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The command output shows that Router B is the querier, because the IPv6 address of Router B's GE 1/0/0 on the same network segment is smaller.



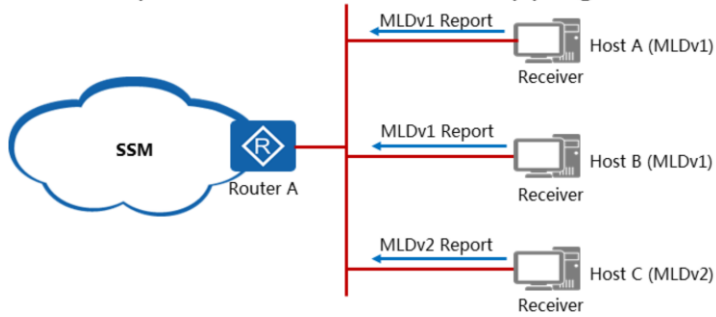
Contents

- MLD Overview
- MLDv1
- MLDv2
- **SSM Mapping**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MLD SSM Mapping

- SSM requires multicast routers to know multicast sources that hosts specify when they join a multicast group. A host running MLDv2 can specify multicast source addresses in MLDv2 Report messages. Some hosts can run only MLDv1. To enable such hosts to obtain the SSM service, multicast routers need to provide the MLD SSM mapping function.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- With SSM mapping entries configured, Router A checks the IPv6 multicast group address G in each received MLDv1 Report message, and processes the message based on the check result:
  - If G is out of the IPv6 SSM group address range, Router A provides the ASM service.
  - If G is within the IPv6 SSM group address range:
  - If the router has no MLD SSM mapping entry matching G, it does not provide the SSM service and drops the Report message.
  - If the router has an MLD SSM mapping entry matching G, it converts (*, G) information in the Report message into (G, INCLUDE, (S1, S2...)) information and provides the SSM service for the hosts. SSM mapping enables hosts running MLDv1 to receive SSM data packets without upgrading the MLD version. This function does not affect hosts running MLDv2.
- Mapping policies can be configured multiple times to map from one group to multiple sources. A router forwards only Group-Source-Specific Query messages in the mapping table.

Configuring MLD SSM Mapping (1)

- Before configuring SSM mapping, complete the following tasks:
  - Configure a unicast routing protocol so that the entire multicast domain is reachable.
  - Enable IPv6 multicast routing.
- Data preparation
  - To configure SSM mapping, you need the following data.

No.	Data
1	Interfaces on which SSM mapping needs to be enabled
2	Name of an IPv6 SSM mapping policy
3	Multicast group address and mask; multicast source address and mask

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuring MLD SSM Mapping (2)

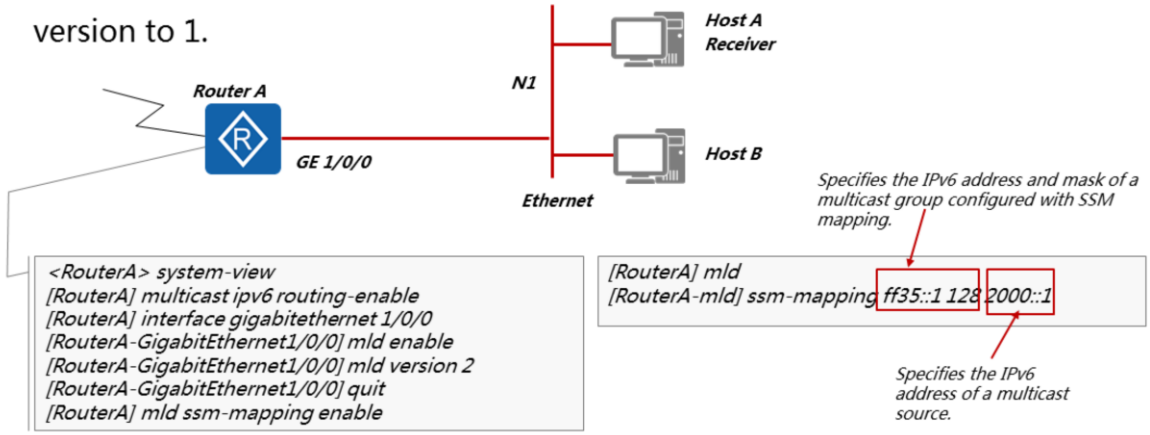
- Run the **system-view** command to enter the system view.
- Run the **interface** *interface-type interface-number* command to enter the interface view.
- Run the **mld enable** command to enable MLD.
- Run the **mld version 2** command to set the MLD version to 2. To ensure that hosts running any MLD version on the network segment can obtain SSM services, it is recommended that you configure MLDv2 on the interface.
- Run the **mld ssm-mapping enable** command to enable SSM mapping.

- Run the **system-view** command to enter the system view.
- Run the **mld** command to enter the MLD view.
- Run the **ssm-mapping** *ipv6-group-address ipv6-group-mask-length ipv6-source-address* command to configure a mapping from a multicast group to a source. To map a multicast group to multiple sources, run the command multiple times.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring MLD SSM Mapping

- Enable multicast, and enable MLD on a host-side interface.
- # Enable multicast on Router A, enable MLD on GE 1/0/0, and set the MLD version to 1.




```
<RouterA> system-view
[RouterA] multicast ipv6 routing-enable
[RouterA] interface gigabitethernet 1/0/0
[RouterA-GigabitEthernet1/0/0] mld enable
[RouterA-GigabitEthernet1/0/0] mld version 2
[RouterA-GigabitEthernet1/0/0] quit
[RouterA] mld ssm-mapping enable
```

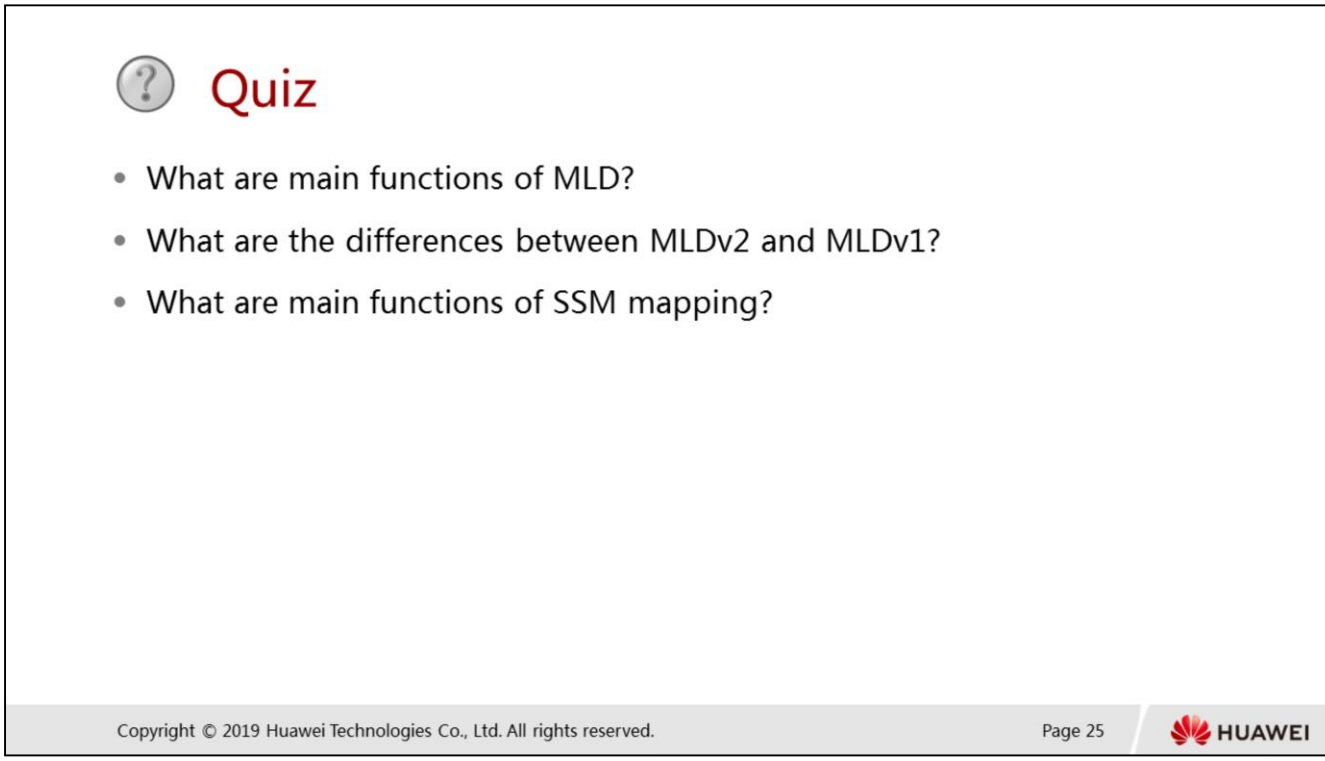
```
[RouterA] mld
[RouterA-mld] ssm-mapping ff35::1 128 2000::1
```

Specifies the IPv6 address and mask of a multicast group configured with SSM mapping.

Specifies the IPv6 address of a multicast source.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



- What are main functions of MLD?
  - MLD manages IPv6 multicast group members, and its fundamentals and functions are similar to those of IGMP. MLD enables each IPv6 router to discover multicast listeners (that is, the nodes that expect to receive multicast data) on its directly connected network and identify the multicast addresses that the neighbor nodes are interested in. The messages are then offered to the multicast routing protocol used by the router to ensure that multicast data is forwarded to all links where receivers exist.
- What are the differences between MLDv2 and MLDv1?
  - MLDv2 is compatible with MLDv1. The fundamentals of MLDv2 are the same as those of MLDv1. MLDv2 supports source lists and filter modes. You can specify source addresses to join a multicast group, implementing SSM.
- What are main functions of SSM mapping?
  - Hosts that send MLDv1 Report messages cannot receive data packets in the SSM group range. SSM mapping enables hosts running MLDv1 to receive SSM data packets without upgrading the MLD version. This function does not affect hosts running MLDv2.

Section Summary

- MLD Functions
- MLDv1 Fundamentals
- MLDv2 Improvement
- SSM Mapping Functions

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Recommendations

- Huawei E-Learning website:
  - http://support.huawei.com/learning/en/newindex.html
- Huawei support case library:
  - http://support.huawei.com/enterprise/en/knowledge/enterprises-common-pid-8221819

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

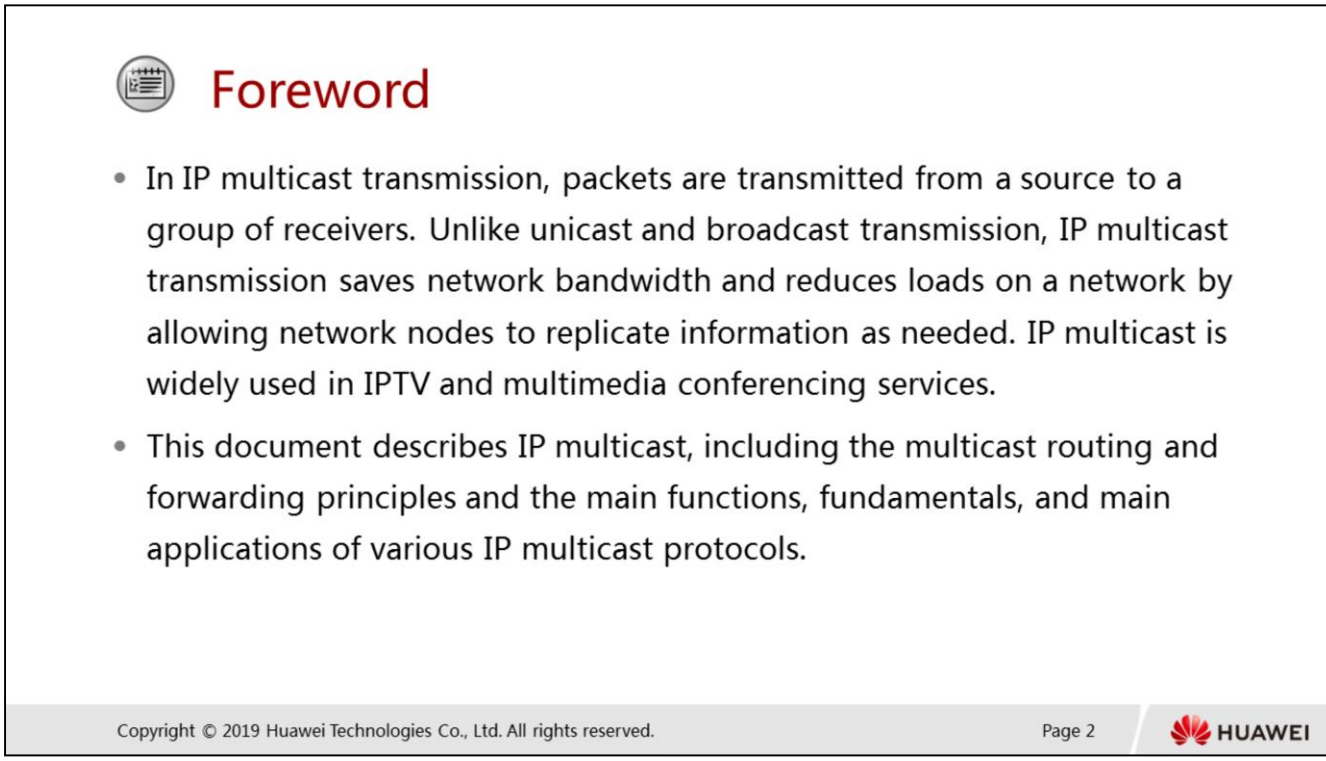
HCIE-Routing & Switching

— PIM

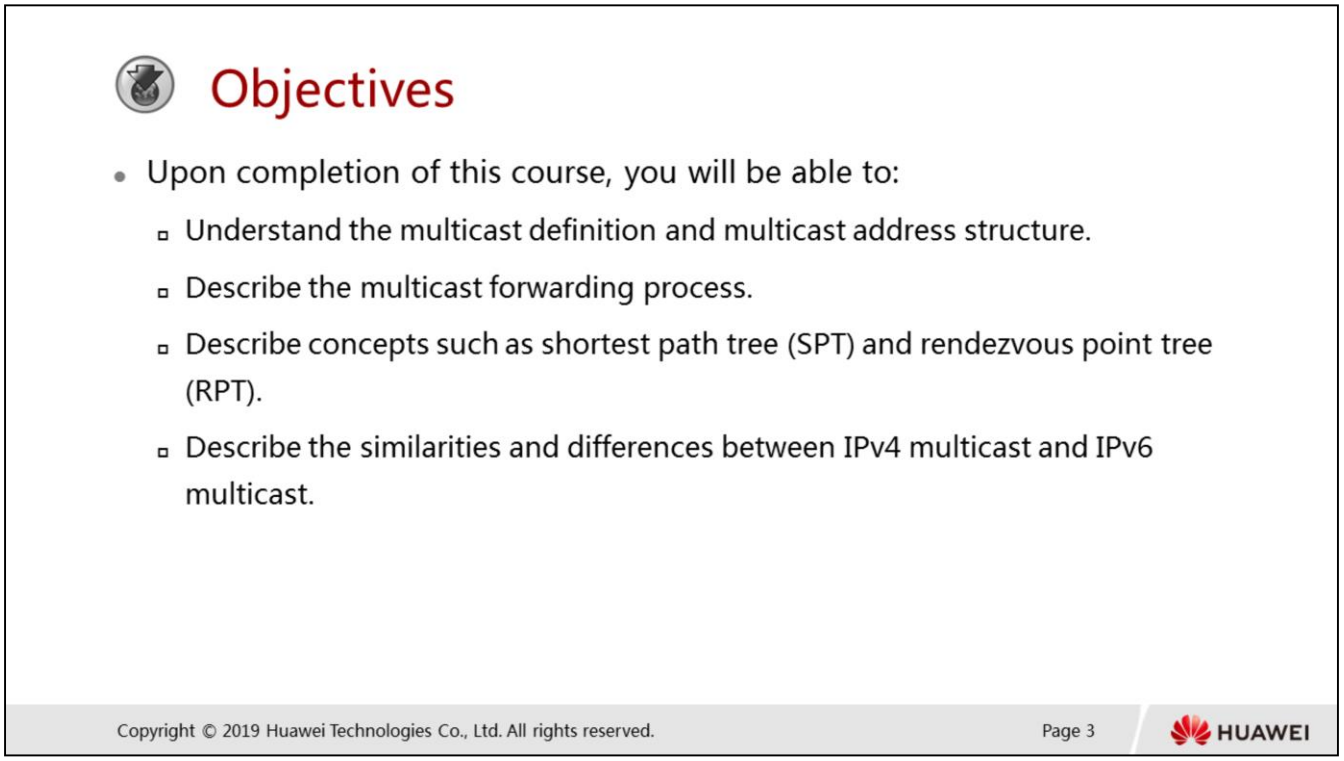
www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



- The modern network transmission technology pays more attention to the following two objectives:
  - Resource discovery
  - Point-to-multipoint transmission
- There are three solutions to achieve these two objectives: unicast, broadcast, and multicast.
- By comparing the data transmission modes of the three solutions, we can conclude that multicast is more suitable for point-to-multipoint IP transmission.



- Upon completion of this course, you will be able to understand the differences among multicast, unicast, and broadcast transmission modes, master the multicast address structure and multicast packet forwarding process, and master related multicast concepts, such as SPT and RPT.



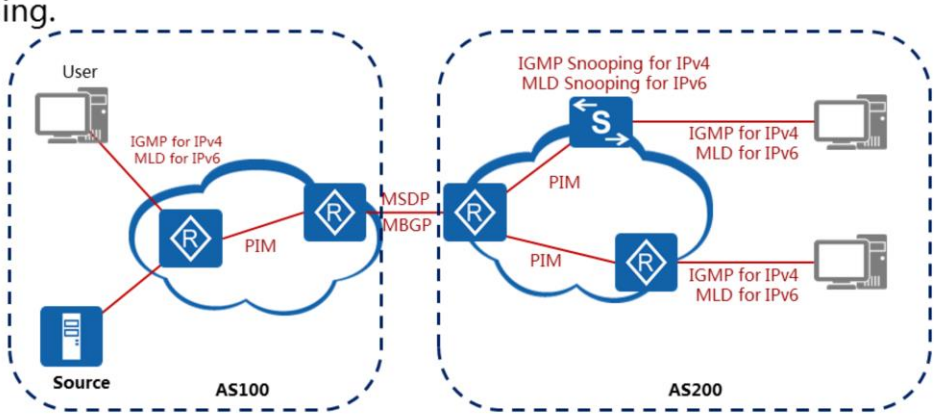
Contents

- **PIM-SM Mechanism**
- IPv6 PIM-SM Configuration
- IPv6 PIM-SSM Fundamentals
- Multicast Route Management
- Typical IPv6 Multicast Applications

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Multicast Protocols

- Multicast protocols include multicast group management protocols for host registration and multicast routing protocols for multicast routing and forwarding.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Multicast protocols include multicast group management protocols for host registration and multicast routing protocols for multicast routing and forwarding. The figure shows various multicast protocols on the network.
- Internet Group Management Protocol (IGMP) runs between receiver hosts and multicast routers, and defines the mechanism for creating and maintaining group membership between them.
- Multicast routing protocols, which run between multicast routers, are used to establish and maintain multicast routes and correctly and efficiently forward multicast data.
- In the ASM model, multicast routes are classified as intra-domain or inter-domain multicast routes.
  - Intra-domain multicast routing protocols discover multicast sources and establish multicast distribution trees in an autonomous system (AS) to deliver information to receivers. Intra-domain multicast routing protocols include Distance Vector Multicast Routing Protocol (DVMRP), multicast open shortest path first (MOSPF), and Protocol Independent Multicast (PIM).
    - DVMRP is a dense mode protocol. It defines a route hop count limit of 32.
    - MOSPF is an extended protocol of OSPF. It defines new LSAs to support multicast.
    - PIM is a typical intra-domain multicast routing protocol and can operate in dense mode (DM) or sparse mode (SM). DM is applicable when receivers are densely distributed on a network, whereas SM is applicable when receivers are sparsely distributed on a network. PIM must work with a unicast routing protocol.
  - Inter-domain multicast routing protocols are used to transmit multicast information between ASs.

- Multicast Source Discovery Protocol (MSDP) can transmit multicast source information across ASs.
- Multicast BGP (MBGP) of Multiprotocol Border Gateway Protocol (MPBGP) can transmit multicast routes across ASs.

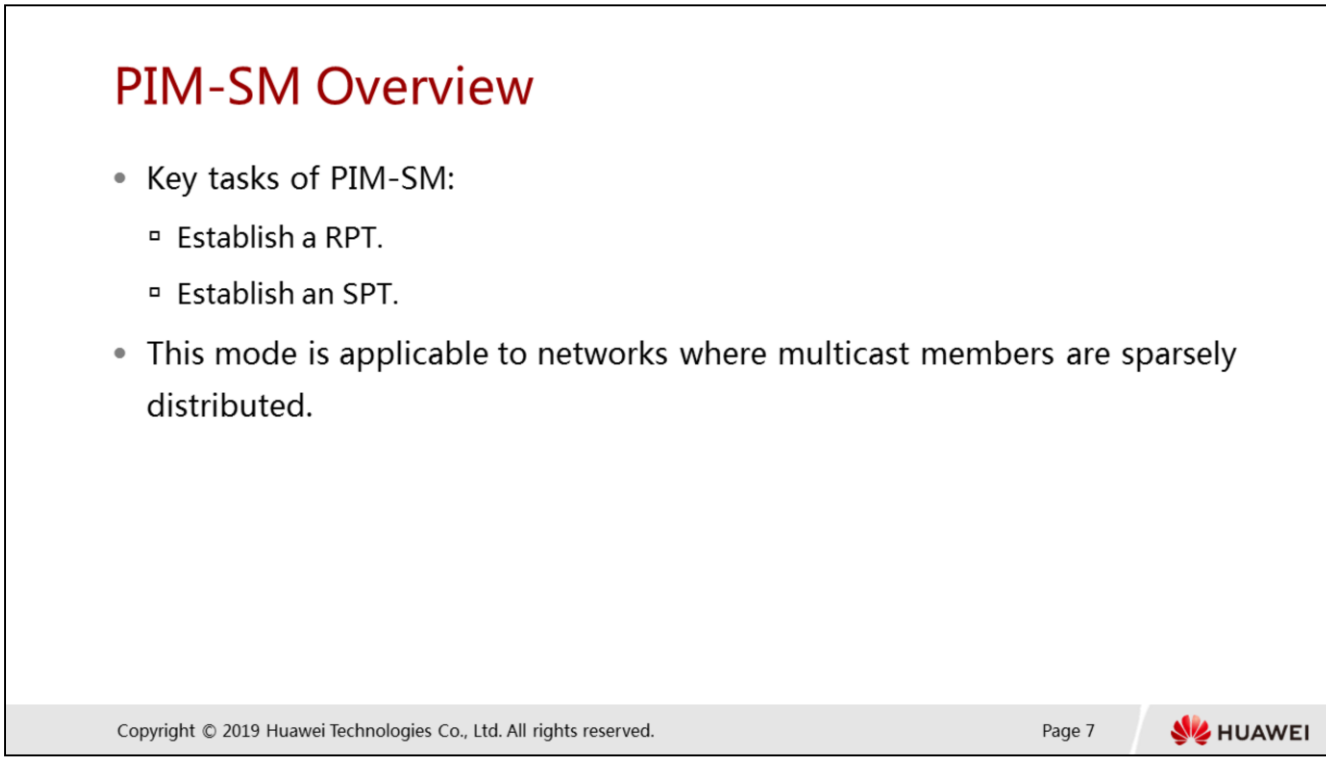
Protocols Used on a Multicast Network

Protocol	Deployment Location	Function
IGMP for IPv4 MLD for IPv6	Between multicast routers and hosts	On the host side, IGMP/MLD allows hosts to dynamically join and leave multicast groups. On the router side, IGMP/MLD exchanges information with upper-layer multicast routing protocols and manages and maintains multicast group member relationships.
PIM	All interfaces on all multicast routers	It provides multicast routing and forwarding, and maintains the multicast routing table based on network topology changes.
IGMP snooping for IPv4 MLD snooping for IPv6	VLANs on Layer 2 switches between multicast routers and hosts	It listens on IGMP/MLD messages exchanged between routers and hosts to create and maintain a Layer 2 multicast forwarding table. In this manner, multicast data can be forwarded on a Layer 2 network.
MSDP for IPv4 only MBGP	AS boundary router	MSDP is required when multicast source information is transmitted across PIM domains. MBGP is required when multicast routes are created across ASs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

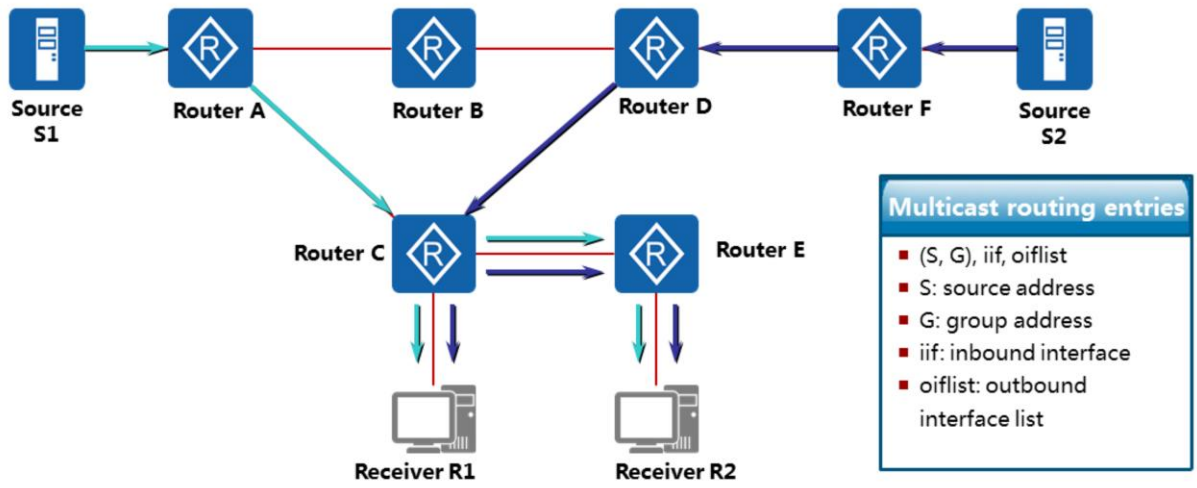
- In the SSM model, domains are not classified as intra-domains or inter-domains. Receivers know the location of the multicast source domain; therefore, multicast transmission paths can be directly established with the help of partial PIM-SM functions.
- MSDP must be deployed between PIM-SM domains to enable the domains to exchange multicast data. An MSDP peer relationship is established between PIM-SM domains, and MSDP peers exchange SA messages to obtain each other's multicast information. Receiver hosts in one PIM-SM domain can then receive data from a multicast source in another PIM-SM domain. MSDP applies only to IPv4 networks and is useful only in the ASM model. Within a PIM domain, IGMP manages group memberships, and PIM-SM maintains multicast forwarding routes.
- PIM forwards multicast data based on a unicast routing table; therefore, multicast forwarding paths are the same as unicast forwarding paths. When a multicast source and receivers are located in different ASs, a multicast distribution tree needs to be set up between the ASs. In this scenario, MBGP can be used to create a multicast routing table independent of the unicast routing table. Multicast data is then transmitted based on the multicast routing table.



- Compared with PIM-DM that uses the push mode, PIM-SM uses the pull mode to forward multicast packets. PIM-SM assumes that group members are distributed sparsely on a network, and almost all network segments have no group members. Multicast routes are created for data forwarding to a network segment only when group members appear on the network segment. PIM-SM is usually used for networks with a large number of sparsely distributed group members.
- Devices on the PIM-SM network work as follows:
  - A Rendezvous Point (RP), an important PIM router, is available to provide services for group members or multicast sources that appear at any time. All PIM routers on the network know the RP's position.
  - When a user host joins a multicast group G using IGMP, the last-hop router sends a Join message to the RP. A (*, G) entry is created hop by hop, and a RPT with the RP as the root is generated.
  - When a multicast source sends the first multicast packet to a multicast group G, the first-hop router encapsulates the multicast data in a Register message and sends the Register message to the RP in unicast mode. The RP then creates an (S, G) entry and registers multicast source information.
- PIM-SM uses the neighbor discovery, DR election, RP discovery, RPT setup, multicast source registration, SPT switchover, and assert mechanisms. A Bootstrap router (BSR) can also be configured to implement fine-grained management in a single PIM-SM domain. The neighbor discovery and assert mechanisms in PIM-SM are the same as those in PIM-DM.

SPT

- An independent SPT is set up between each multicast source and receiver.



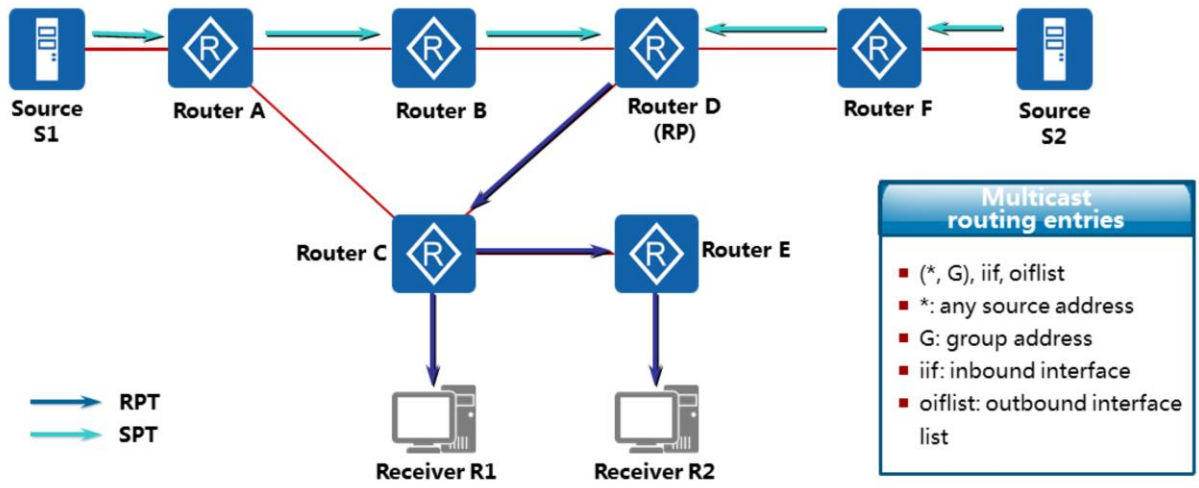
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An SPT is rooted at a multicast source and combines the shortest paths from the source to receivers.
- For a multicast group, routers need to establish an SPT from each multicast source that sends packets to the group.
- In this example, there are two multicast sources (S1 and S2) and two receivers (R1 and R2). Two SPTs are established on the network.
- S1-Router A-Router C (R1)-Router E (R2)
- S2-Router F-Router D-Router C (R1)-Router E (R2)

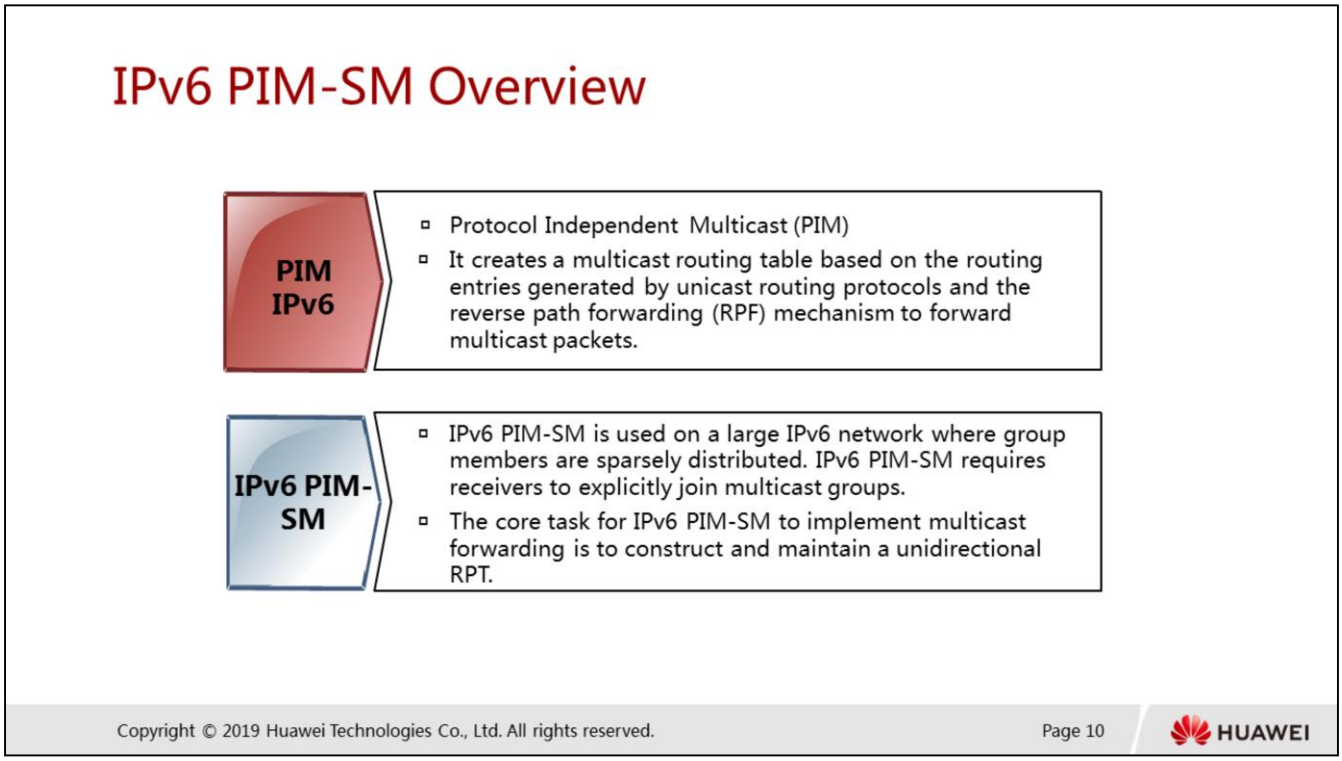
RPT

- Only one RPT is set up for each multicast group.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

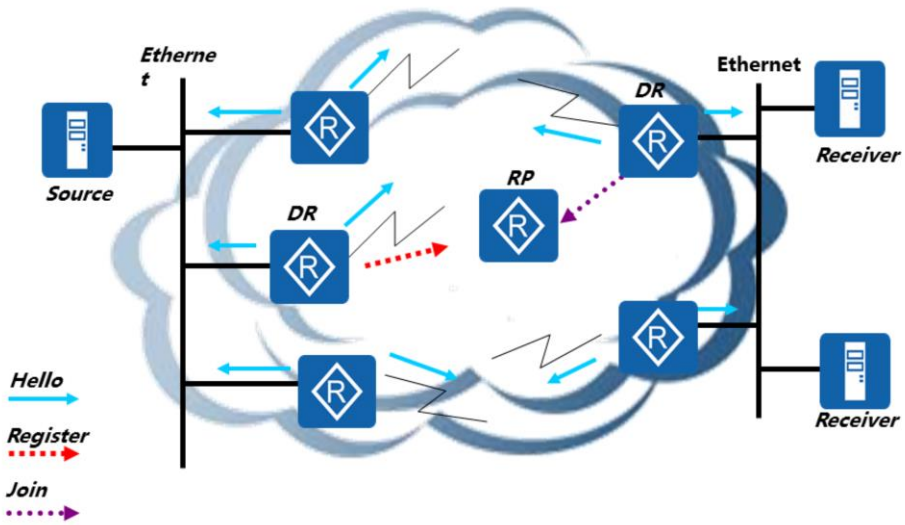
- A RPT is rooted at a rendezvous point (RP) and combines the shortest paths from the RP to all receivers. Only one RPT is set up for each multicast group. All multicast sources and receivers of a group send and receive multicast data packets along the RPT. A multicast source first sends data packets to the RP, which then forwards the packets to all receivers.
- In this example, multicast sources S1 and S2 share one RPT: Router D (RP)-Router C (R1)-Router E (R2)



- PIM IPv6 is a multicast routing protocol independent of IPv6 unicast routing protocols, such as static routes, RIPng, OSPFv3, IS-ISv6, and BGP4+. It creates a multicast routing table based on the routing entries generated by unicast routing protocols and the RPF mechanism to forward multicast packets. A PIM IPv6 domain is a network that consists of multicast routers supporting PIM IPv6.
- Currently, there are two multicast models: any-source multicast (ASM) and source-specific multicast (SSM). In IPv6, the ASM model involves IPv6 PIM-DM and IPv6 PIM-SM. The SSM model is implemented using MLDv2 and some mechanisms of IPv6 PIM-SM.
- IPv6 PIM-SM is used on a large IPv6 network where group members are sparsely distributed. IPv6 PIM-SM requires receivers to explicitly join multicast groups. By default, IPv6 PIM-SM assumes that all nodes on a network do not need to receive multicast packets. Upstream nodes forward multicast data only after receiving Join messages from downstream nodes.
- In IPv6 PIM-SM, the RP forwards multicast information only to downstream branches that have receivers, reducing network bandwidth consumption by data and control packets and lowering routers' processing overheads.
- When a host wants to receive data from a specified multicast group, the router connected to the host sends a Join message to the RP in the group. A RPT rooted at the RP is established along the path. The RPT enables this shared path to be used when different multicast sources forward multicast data to the same multicast group.
- When a multicast source sends data to a multicast group, the DR connected to the source encapsulates the multicast data into a Register message and sends the Register message to the RP in unicast mode. After the Register message reaches the RP, the RP decapsulates the multicast data and sends it to a receiver along the RPT. When the multicast data sent using Register messages reaches a certain rate, the RP sends a Join message to the multicast source to establish a multicast distribution tree between the multicast source and RP. The RP then sends a Register-Stop message to the DR on the

source side, instructing the DR to directly send multicast data in non-encapsulation mode according to the multicast forwarding information base (MFIB).

DR Election



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

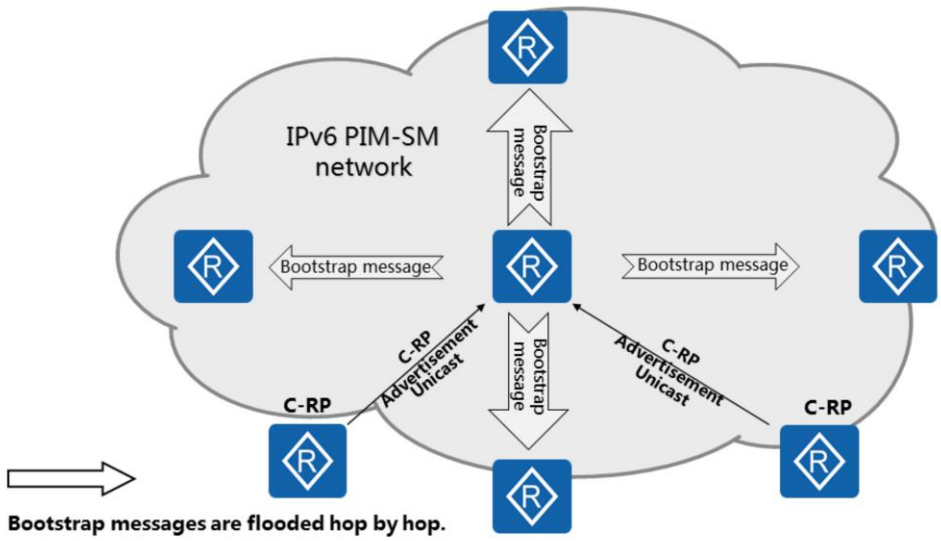
- A DR is elected on a shared network segment using Hello messages, and is the only multicast data forwarder on the network segment. A DR needs to be elected on the shared network connecting to the multicast source and receivers. The DR on the receiver side sends a Join message to the RP, and the DR on the source side sends a Register message to the RP.
- The DR election process is as follows:
  - Routers on the shared network segment send Hello messages with the DR priority to each other.
  - The router with the highest priority is elected as the DR on the network segment. If the routers have the same priority, the router with the largest IPv6 address is elected as the DR.
- If the DR is abnormal, other routers cannot receive Hello messages from the DR. After the DR expires, a new round of DR election is triggered on the shared network segment.
- If at least one router on the network does not allow Hello messages to carry the DR priority, the router with the largest IPv6 link-local address serves as the DR.

- How to discover a RP? For a small network, a RP is sufficient to forward information on the entire network and its location can be statically specified. You can manually specify the IP address of a RP on the DR, leaf routers, and all the routers that multicast data streams pass through. However, in most applications, an IPv6 PIM-SM network covers a large area, and a large amount of multicast traffic needs to be forwarded through RPs. Therefore, different multicast groups should have their own RPs. To reduce the workload of configuring multiple static RPs and better adapt to real-time network changes, use the bootstrap mechanism to dynamically elect RPs.
- A BSR is the management core of an IPv6 PIM-SM network. The BSR collects Advertisement messages from each candidate RP (C-RP), and selects proper C-RPs to form the RP-Set information of multicast groups. A RP-Set is the database for each multicast group and its corresponding C-RP. The BSR notifies the entire IPv6 PIM-SM network of the RP-Set information through a bootstrap message. After learning the C-RPs for each multicast group, all routers including the DR calculate the unique RP for each multicast group based on the hash algorithm.
- A network (or a management domain) can have only one BSR, but can have multiple candidate BSRs (C-BSRs). Once the BSR is faulty, a new BSR can be elected from the C-BSRs through the bootstrap mechanism to prevent service interruptions. Multiple C-RPs can be configured in an IPv6 PIM-SM domain. The BSR collects and sends the RP-Set information of each multicast group.
    - RP configuration recommendations:
        - On small- and medium-scale networks, configure a static RP because of its stability and low requirements on network devices.
        - If there is only one multicast source, use the router directly connected to the multicast source as a static RP so that the source DR does not need to register with the RP.
        - If a static RP is deployed, all routers including the RP in the same domain must be configured with the same RP information and the same range of multicast

groups.

- On large-scale networks, use a dynamic RP because of its high reliability and ease of maintenance.
- If multiple multicast sources are densely distributed on the network, configure the core routers close to the multicast sources as C-RPs. If multiple group members are densely distributed on the network, configure the core routers close to group members as C-RPs.

RP Election



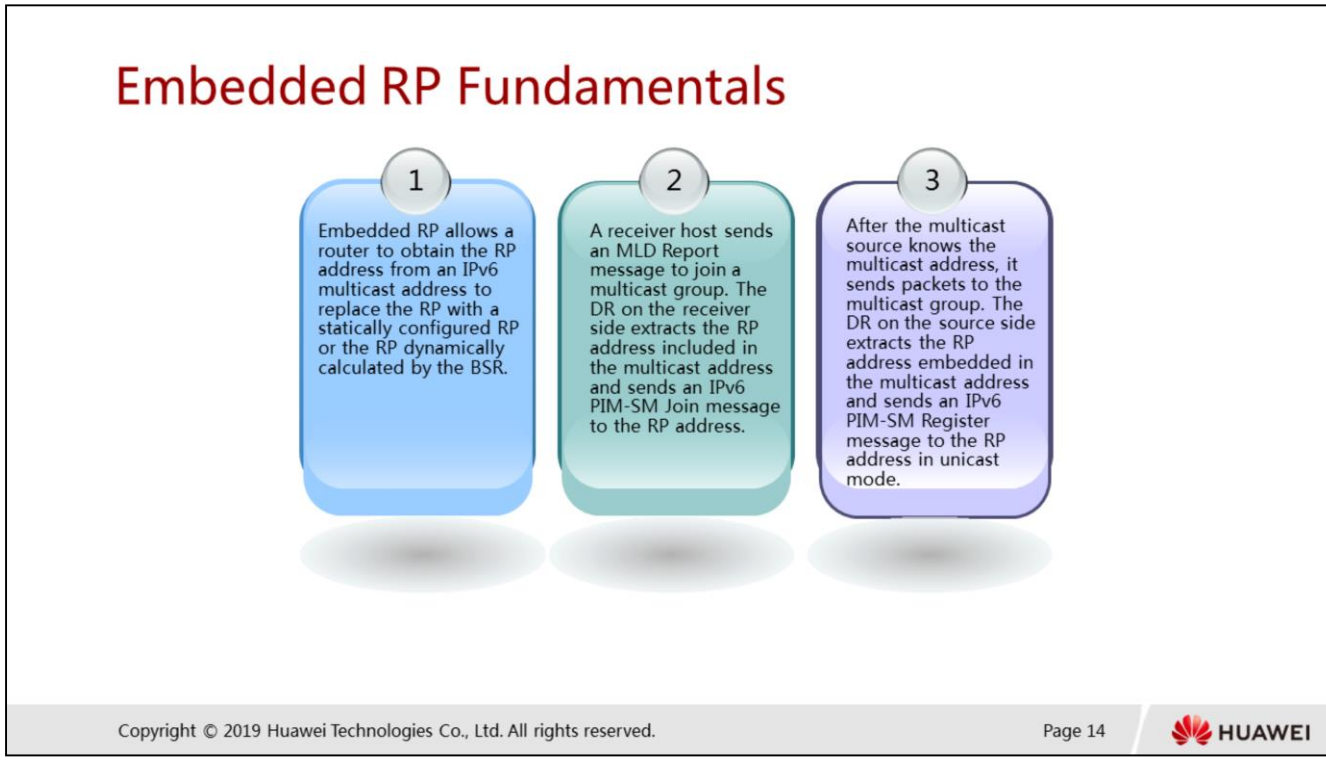
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The working process of the BSR is as follows: Suitable routers on a network are configured as C-BSRs. Each C-BSR has a priority. After a router is configured as a C-BSR, it starts a timer (of 150s by default) to monitor bootstrap messages on the network. The first bootstrap message sent from a C-BSR carries the priority and IPv6 address of the C-BSR. After receiving a bootstrap message, a C-BSR compares its priority with the priority in the message. If the priority in the message is higher, the C-BSR resets its timer and continues to listen to bootstrap messages. If the C-BSR checks that its own priority is higher, it sends a bootstrap message to declare that it is the BSR. If the priorities are the same, the C-BSR compares the IPv6 addresses. A C-BSR with a larger IPv6 address is elected as the BSR. The destination address of each bootstrap message is FF02::13 and the TTL is 1. All PIM IPv6 routers can receive the message and send it out of all PIM IPv6-enabled interfaces so that all PIM IPv6 devices on the network can receive the bootstrap message.
- RPs must be manually configured on devices. Configure C-RPs first, including the RP IPv6 addresses, priorities, and groups that the C-RPs can serve. As mentioned above, a RP can provide services for some or all IPv6 multicast groups. After receiving a bootstrap message, each C-RP learns the BSR on the network from the message. The C-RP then unicasts the multicast groups that it can serve to the BSR through a Candidate-RP-Advertisement message. In this way, the BSR collects information about all C-RPs on the network and sorts the information into a RP-Set. The BSR then sends the RP-Set information to all routers on the entire network through a bootstrap message.
- The RP election rules are as follows:
  - If the RP-Set has only one C-RP for the IPv6 group address, the DR selects the C-RP as the RP.
  - If the RP-Set has multiple C-RPs for the IPv6 address group, the DR selects the C-RP with the highest priority as the RP (a smaller value indicates a higher priority).
  - If the priorities are the same, the DR starts the hash algorithm and uses the group addresses, hash masks, and C-RP addresses as input parameters. The DR then outputs a

number for each C-RP and selects the C-RP with a higher number as the RP of the group.

- If the hash results are also the same, the C-RP with the largest IPv6 address becomes the RP of the group.



- Embedded RP allows a router to obtain the RP address from an IPv6 multicast address to replace the RP with a statically configured RP or the RP dynamically calculated by the BSR.
- The multicast address range in an embedded RP scenario is FF7*x*::/16 to FFF*x*::/16, where *x* indicates any hexadecimal number ranging from 0 to F.
- Receiver side:
  - The receiver host sends an MLD Report message to join the multicast group.
  - The DR on the receiver side extracts the RP address embedded in the multicast address and sends an IPv6 PIM-SM Join message to the RP address.
- Source side:
  - After the multicast source knows the multicast address, it sends packets to the multicast group.
  - The DR on the source side extracts the RP address embedded in the multicast address and sends an IPv6 PIM-SM Register message to the RP address in unicast mode.

Multicast Address into Which the RP Address Is Embedded

- How to embed a 128-bit RP address into a 128-bit IPv6 multicast address?
- A special multicast address is defined as follows:



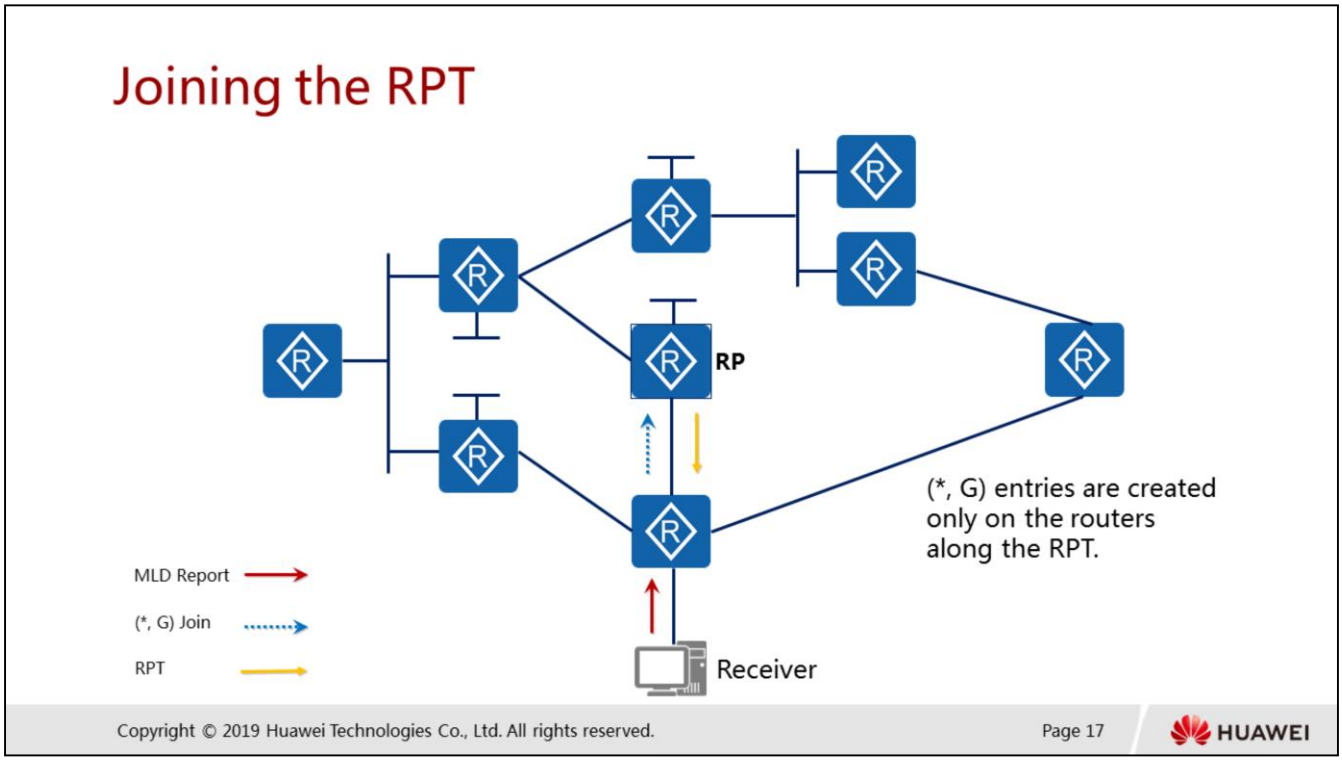
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The first 8 bits is FF, indicating an IPv6 multicast address.
- The value range of the Flags field is 7 to F, indicating an IPv6 multicast address into which the RP address is embedded.
- RIID: RP interface ID, which is filled in the last 4 bits of the RP address.
- Plen: prefix length of the RP address, which cannot be 0 or greater than 64 after being converted to a decimal number.
- Network Prefix: prefix of the RP address.

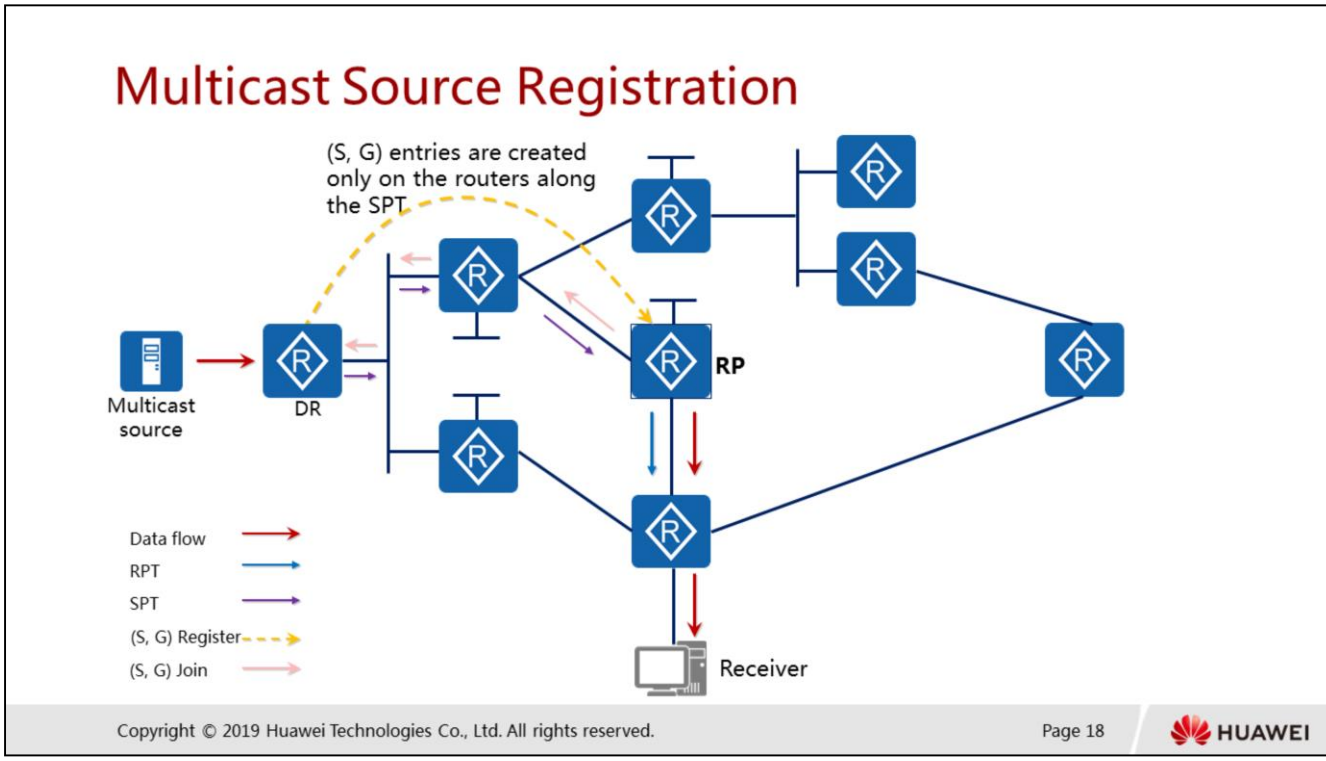
RP Address Conversion

- How to obtain the RP address from a multicast address?
  - Extract the 'Plen' field and convert it into a decimal number.
  - Extract the first 8 bits of the 'Network Prefix' field as the prefix of the RP address.
  - Extract the 'RIID' field as the last 4 bits of the interface ID of the RP address, and use 0s to supplement other parts of the interface ID.
- Embedded RP example
  - If a multicast address is FF70:140:2001:DB8:BEEF:FEED::/96, the RP address obtained from the multicast address is 2001:DB8:BEEF:FEED::1/64.

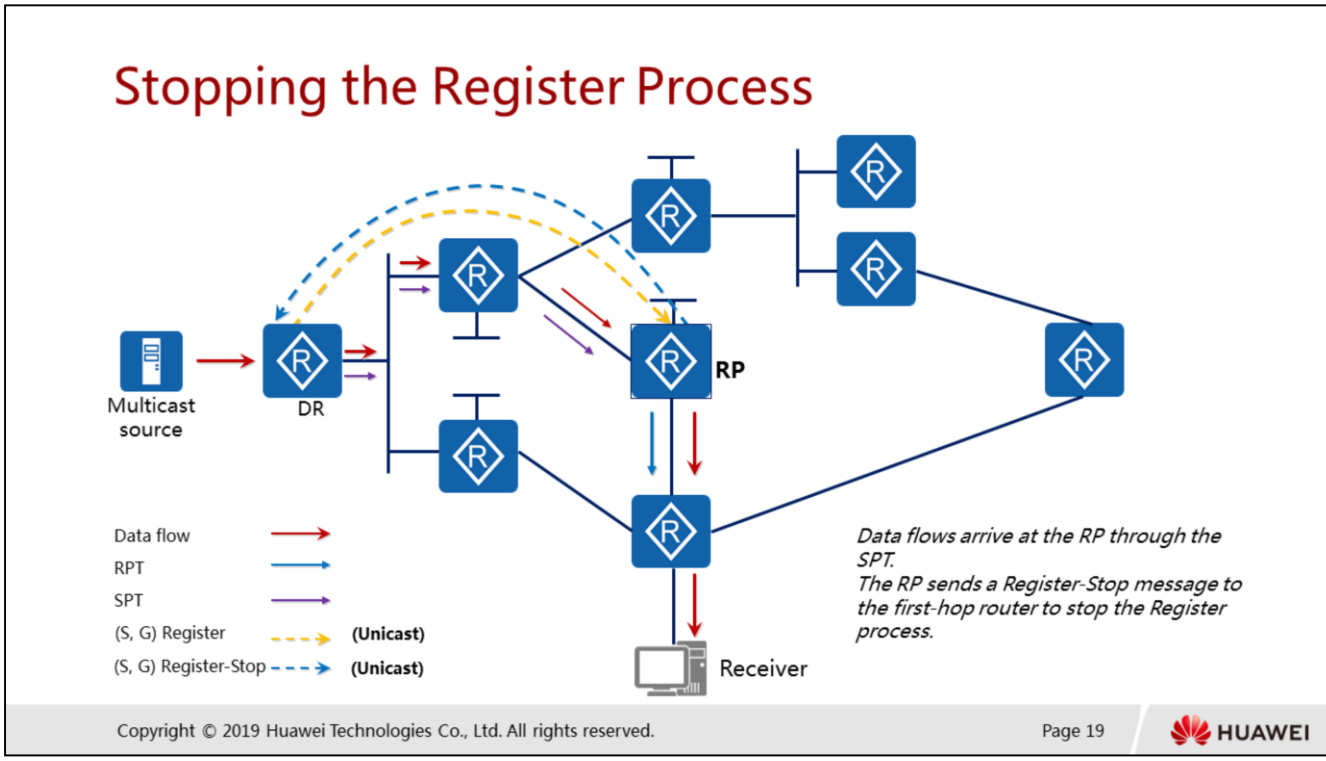
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



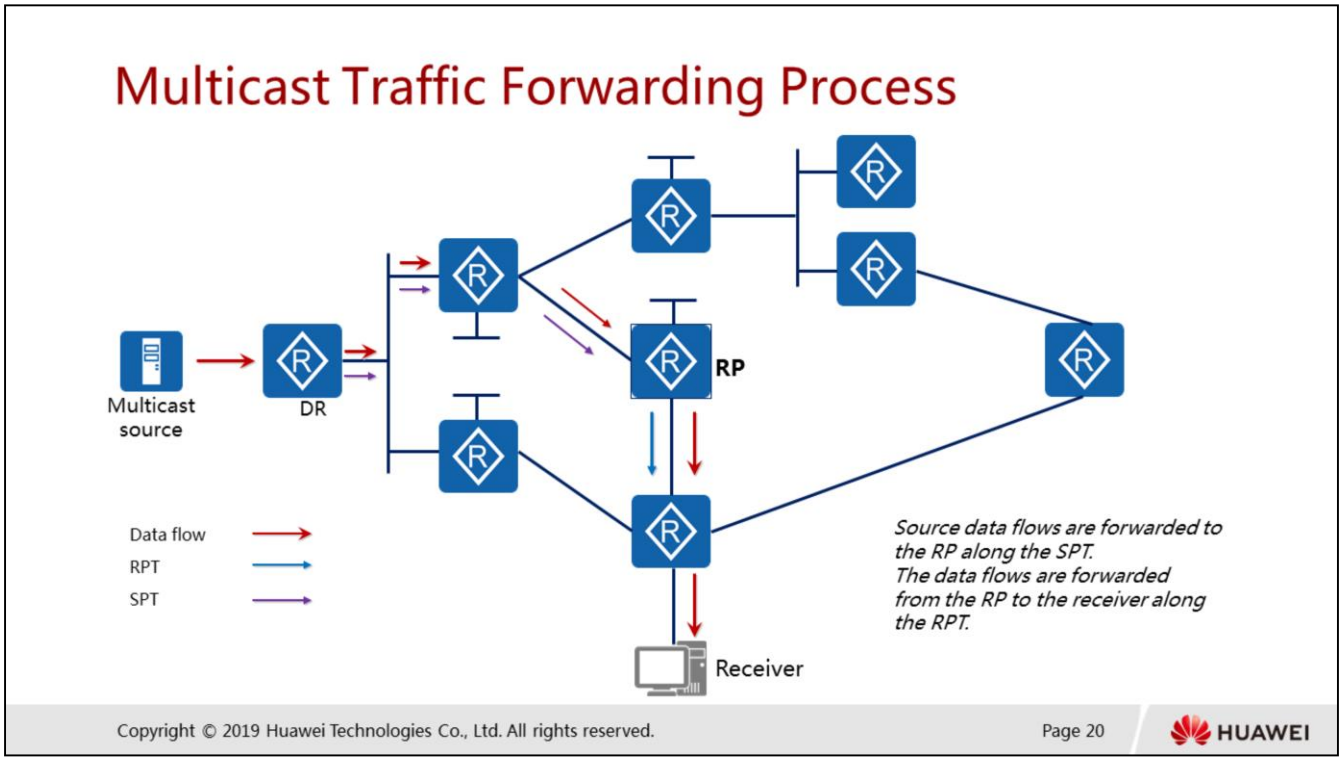
- When the receiver host joins the multicast group G, it sends an MLD message to the leaf router directly connected to the host. The leaf router keeps the receiver information of the multicast group G, and sends a Report message through the upstream nodes hop by hop to the RP.
- Each router that the message passes through from the leaf router to the RP creates (*, G) entries in the forwarding table. These routers constitute a branch of the RPT. (*, G) indicates the information from any source to the multicast group G. The RPT uses the RP and receiver as its root and leaf, respectively.
- When packets from the multicast source S to the multicast group G passes through the RP, the packets reaches the leaf router along the established RPT before arriving at the receiver host.
- When the receiver is no longer interested in the information from the multicast group G, the multicast router directly connected to the receiver sends a Prune message hop by hop to the RP corresponding to the group in a direction reverse to the RPT. Upon receiving the message, the first upstream router deletes the interface connected to the downstream router from its interface list and checks whether it has the receiver of the multicast group. If no, the upstream router forwards the Prune message to its upstream router.



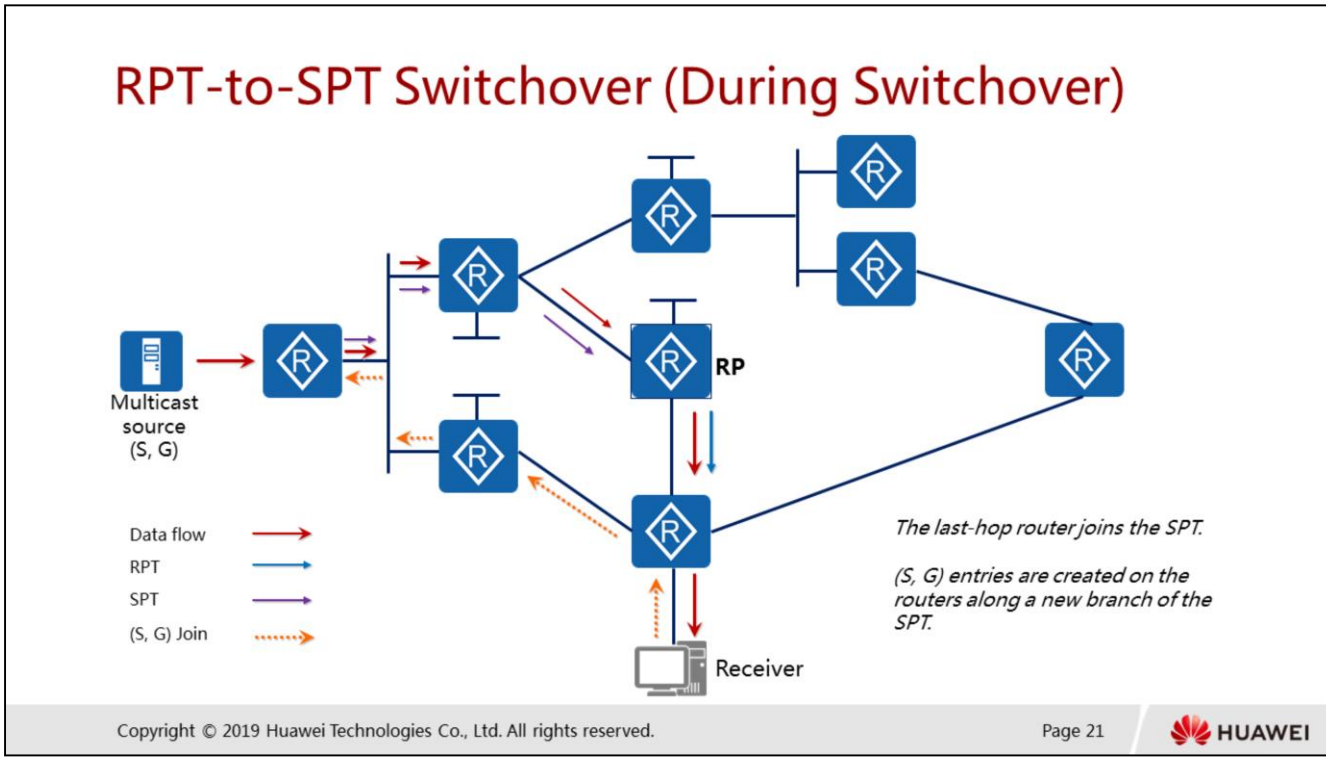
- When the multicast source S sends a multicast packet to the multicast group G, the router directly connected to S encapsulates the packet into an IPv6 PIM Register message after receiving the packet and unicasts the message to the RP.
- After receiving the Register message from the multicast source S, the RP decapsulates the Register message and forwards the multicast information to the receiver along the RPT. Additionally, the RP sends an (S, G) Join message to the multicast source S hop by hop, so that all routers between the RP and the multicast source S generate (S, G) entries. The routers along the path form a branch of the SPT. The SPT uses the multicast source S and RP as its root and destination, respectively.



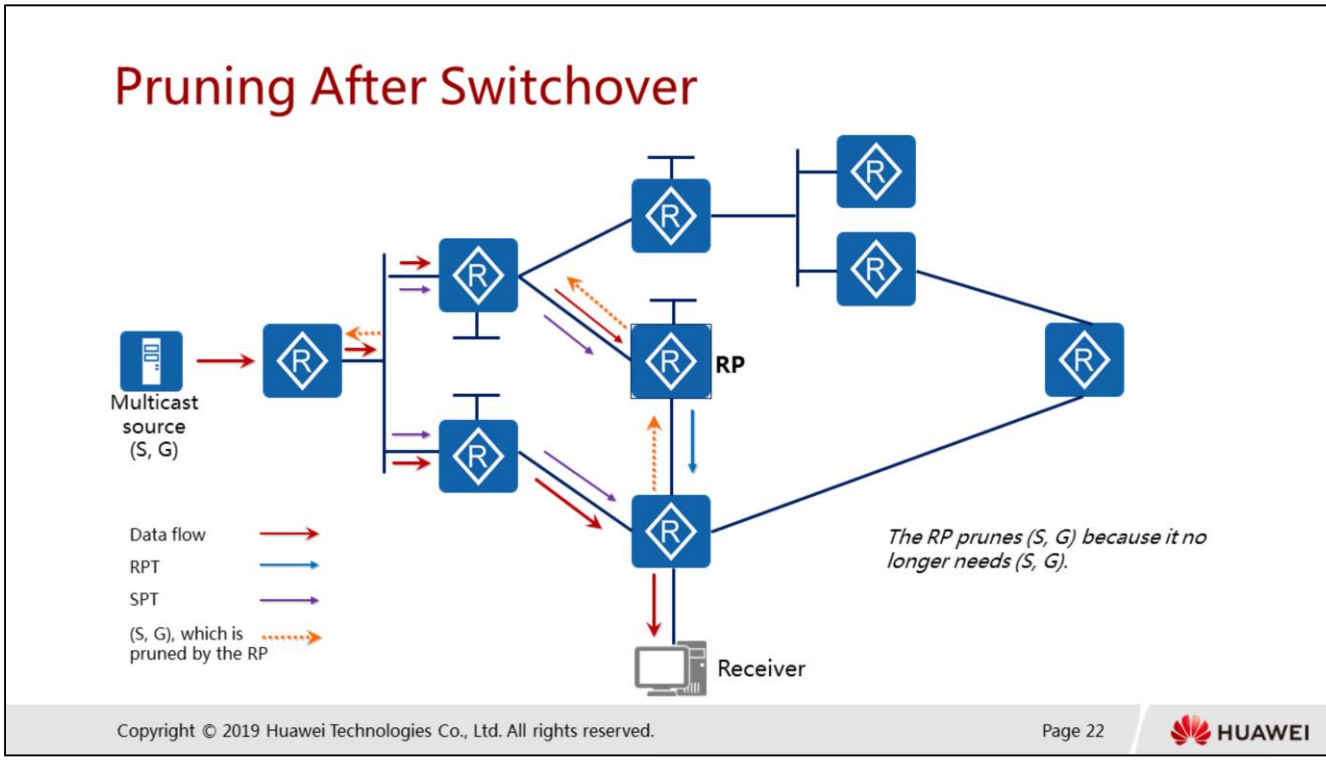
- The multicast information sent by the multicast source S reaches the RP along the established SPT, and then the RP forwards the information along the RPT. After receiving the multicast information forwarded along the SPT, the RP unicasts a Register-Stop message to the router that is directly connected to the multicast source S. Up to now, the registration of the multicast source is complete.



- Source data flows are forwarded to the RP along the SPT, and then the RP forwards them to the receiver along the RPT.



- By specifying a threshold for the rate of multicast packets from a specific source, PIM-SM enables the last-hop router (the DR on the receiver side) to switch from the RPT to the SPT. When the last-hop router finds that the rate of multicast packets from the RP to the multicast group G exceeds the threshold, it sends an (S, G) Join message to the next-hop router of the multicast source S based on the unicast routing table. The Join message reaches the first-hop router hop by hop. All routers along the path have the (S, G) entry, and a branch of the SPT is established.
- The DR on the receiver side periodically checks the rate of multicast packets. If the DR on the receiver side finds that the rate of multicast packets sent from the RP to the multicast group G exceeds the threshold, it triggers an SPT switchover.
  - The DR on the receiver side sends an (S, G) Join message to the DR on the source side and creates an (S, G) entry. The Join message is transmitted hop by hop, and routers along the path all create the corresponding (S, G) entry. Finally, an SPT is set up from the DR on the source side to the DR on the receiver side.
  - After the SPT is set up, the DR on the receiver side sends a Prune message to the RP. The Prune message is transmitted hop by hop along the RPT. After receiving the Prune message, the routers on the RPT convert the (*, G) entry into the (S, G) entry, and prune their downstream interfaces. After the prune action is complete, the RP no longer forward multicast packets along the RPT.
  - Because the SPT does not pass through the RP, the RP continues to send Prune messages along the RPT to the DR on the source side, which then deletes the downstream interface connected to the RP from the (S, G) entry. After the prune action is complete, the DR on the source side no longer forwards multicast packets along the SPT to the RP.
- According to default configuration of the VRP, routers connected to receivers join the SPT immediately after receiving the first multicast data packet from a multicast source, triggering a RPT-to-SPT switchover.



- When a router receives the same multicast data along the RPT and SPT on different interfaces, it discards the data received along the RPT and sends a Prune message to the RP hop by hop. After receiving the Prune message, the RP updates the forwarding status and stops forwarding (S, G) multicast traffic along the RPT. Additionally, the RP sends a Prune message to the multicast source to delete or update the (S, G) entry. In this way, multicast data is switched from the RPT to the SPT.

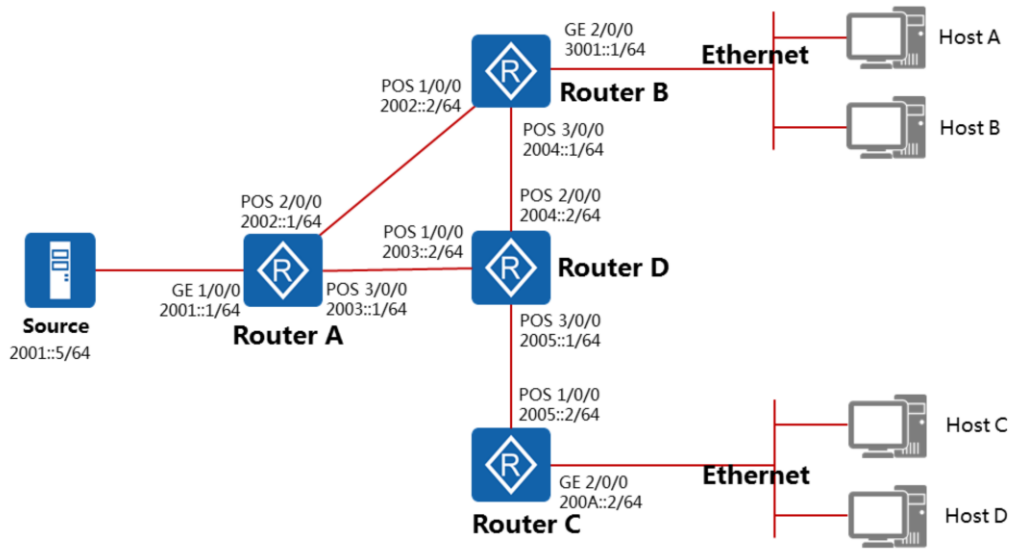


Contents

- IPv6 PIM-SM Mechanism
- **IPv6 PIM-SM Configuration**
- IPv6 PIM-SSM Fundamentals
- Multicast Route Management
- Typical IPv6 Multicast Applications

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IPv6 PIM-SM Configuration Example

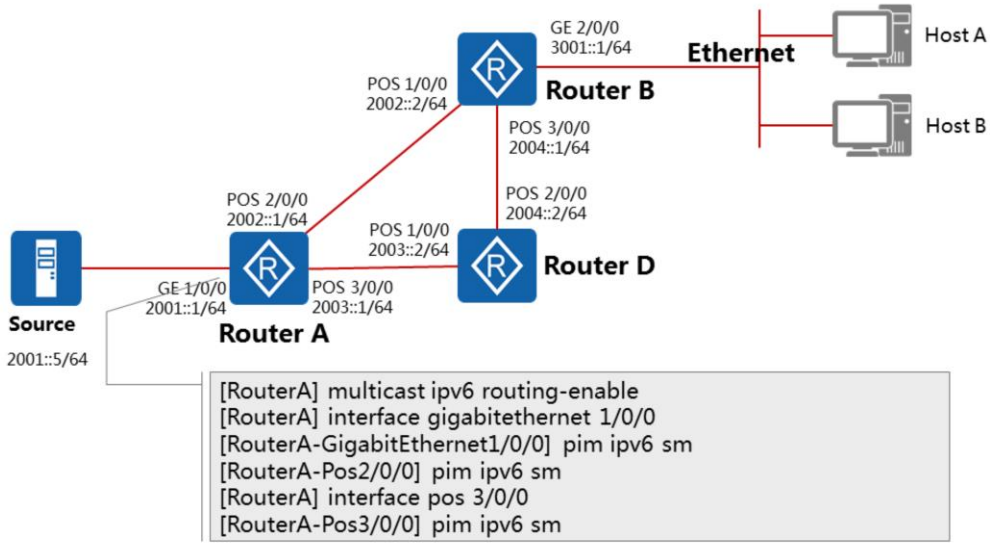


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Host A and Host C are multicast receivers on two leaf networks. These receivers connect to the multicast source through Router A, Router B, Router C, and Router D.
- Configuration roadmap:
  - Configure an IPv6 address for each router interface and an IPv6 unicast routing protocol.
    - Configure an IPv6 address and mask for each router interface.
    - Configure OSPFv3 on each router and set the process ID to1 and the area ID to 0 to ensure that Router A, Router B, Router C, and Router D can communicate at the network layer.
  - Enable IPv6 multicast on each router, enable IPv6 PIM-SM on each router interface, and configure MLD on the interfaces connected to hosts (the default version 2 is used).
  - Configure a C-BSR and C-RP (in this example, the IPv6 global unicast addresses of the C-BSR and C-RP are both 2004::2 on Router D).
  - Verify the configuration.

Configuring Router A

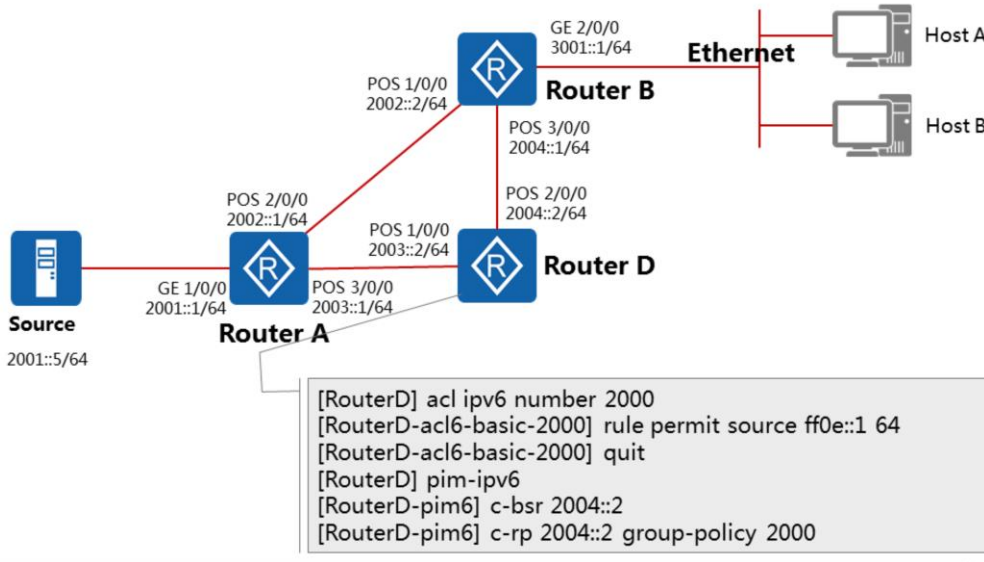


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Enter the system view.
  - **system-view**
- Enable IPv6 multicast routing.
  - **multicast ipv6 routing-enable**
- Enter the interface view.
  - **interface** *interface-type interface-number*
- Enable IPv6 PIM-SM.
  - **pim ipv6 sm**
- IPv6 PIM-SM can be configured only after IPv6 multicast is enabled. After IPv6 PIM-SM is configured on a router's interface, the router periodically sends Hello messages to discover PIM IPv6 neighbors and processes messages from the neighbors. When a router joins an IPv6 PIM-SM domain, you are advised to enable IPv6 PIM-SM on all interfaces of a non-border router.
- IPv6 PIM-SM and IPv6 PIM-DM cannot be enabled on an interface at the same time. The PIM IPv6 modes of all interfaces on a router must be the same.
- The configurations on Router B, Router C, and Router D are similar to the configuration on Router A.
- MLD must be enabled on the router connected to receivers.

Configuring Router D



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- One or more C-BSRs can be configured in a PIM domain. The BSR elected from C-BSRs collects and advertises C-RP information. The BSR and other devices in the domain need to exchange a great deal of information, and therefore sufficient bandwidth should be reserved between the C-BSR and other devices in the domain. Routers on a backbone network are used as C-BSRs.
- When you specify an interface address as the address of the C-BSR, you must enable IPv6 PIM-SM on the interface. The process of electing a BSR from C-BSRs is described as follows:
  - Initially, each C-BSR considers itself as a BSR in the PIM-SM domain and uses its interface IPv6 address as the BSR address to send bootstrap messages.
  - When a C-BSR receives a bootstrap message from another router, it compares the BSR address in the message with its own BSR address. The comparison items include the priorities and BSR addresses. If the priorities are the same, the C-BSR with a larger BSR address is preferred. That is, if a C-BSR checks that the BSR address in a received bootstrap message is greater than its own BSR address, the C-BSR uses the BSR address in the message to replace its own BSR address and no longer thinks itself as a BSR. If a C-BSR checks that the BSR address in a received bootstrap message is not greater than its own BSR address, the C-BSR continues to consider itself as a BSR.
- Enter the PIM IPv6 view.
  - **pim-ipv6**
- Configure an interface address as the address of the C-BSR.
  - **c-bsr** ipv6-address [hash-length] [priority-value]
- Configure a C-RP.
  - **c-rp** *ipv6-address* [**priority** *priority*]
- Configure a static RP.
  - **static-rp** *rp-address* [*basic-acl6-number*] [**preferred**]
- Configure an embedded RP.
  - **embedded-rp [** *basic-acl6-number* **]**
- The **c-bsr** command configures an interface address as a C-BSR address on a router that wants to become a BSR.
  - *ipv6-address*: specifies the global IPv6 unicast address of a C-BSR.
  - *hash-length*: specifies the mask length of the hash function for calculating a RP. The value is an integer ranging from 0 to 128.
- The **pim ipv6 bsr-boundary** command configures an interface as the BSR boundary. After this command is run on an interface, bootstrap messages cannot pass through this interface but other PIM packets can.

- The **c-rp** command configures a router to notify the BSR that it is a C-RP. A large bandwidth must be reserved between a router that is configured as a C-RP and other devices.
    - *ipv6-address*: specifies the global IPv6 unicast address of a C-RP.
    - If there is only one dynamic RP on a network, configuring a static RP prevents communication interruptions caused by a single RP's fault. When a static RP is used to forward multicast data, the same **static-rp** command must be configured on all routers in an IPv6 PIM-SM domain.
- The **static-rp** command configures a static RP.
    - *rp-address*: specifies the address of a static RP. This address must be a valid IPv6 global unicast address.
    - *basic-acl6-number*: specifies the number of a basic ACL for controlling the range of multicast groups that a static RP serves. The value ranges from 2000 to 2999.
    - **preferred**: prefers the configured static RP if this RP is different from the RP selected by the BSR mechanism. If this parameter is not specified, the RP selected by the BSR mechanism is preferred.
- A router uses an embedded RP to obtain the RP address from a multicast

address to replace a static RP or a dynamic RP elected using the BSR mechanism. The multicast address range in an embedded RP scenario is FF7*x*::/16 to FFF*x*::/16, where *x* indicates any hexadecimal number ranging from 0 to F.

Displaying the PIM Configurations and BSR Election Information on Interfaces

```
[RouterB] display pim ipv6 interface
Vpn-instance: public net
Interface    NbrCnt HelloInt  DR-Pri     DR-Address
Pos1/0/0        1     30        1      FE80::A01:10E:1(local)
GEth2/0/0       0     30        1       FE80::200:AFF:FE01:10E(local)
Pos3/0/0        1     30        1      FE80::9D62:0:FDC5:2
```

```
[RouterB] display pim ipv6 bsr-info
  Vpn-instance: public net
    Elected BSR Address: 2004::2
      Priority: 0
      Hash mask length: 126
      State: Accept Preferred
      Uptime: 00:04:22
      Expires: 00:01:46
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Display PIM IPv6 information on interfaces.
  - **display pim ipv6 interface** [*interface-type interface-number*]
- Display information about the BSR in an IPv6 PIM-SM domain.
  - **display pim ipv6 bsr-info**
  - The BSR on the network is Router D's POS 2/0/0.

Displaying RP Information

```
[RouterB] display pim ipv6 rp-info
  Vpn-instance: public net
   PIM-SM BSR RP information:
   Group/MaskLen: FF0E::1/64
     RP: 2004::2
     Priority: 0
     Uptime: 00:05:19
     Expires: 00:02:11
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Display RP information in an IPv6 PIM-SM domain.
  - **display pim ipv6 rp-info** [ipv6-group-address]
- The RP on the network is Router D's POS 2/0/0.

Displaying IPv6 Multicast Routing Information - Source-Side DR

```
[RouterA] display pim ipv6 routing-table
Vpn-instance: public net
 Total 0 (*, G) entry; 1 (S, G) entry
 (2001::5, FF0E::1)
    RP: 2004::2
    Protocol: pim-sm, Flag: SPT LOC ACT
    UpTime: 00:02:15
    Upstream interface: GigabitEthernet1/0/0
       Upstream neighbor: FE80::200:AFF:FE01:10D
       RPF prime neighbor: FE80::200:AFF:FE01:10D
    Downstream interface(s) information:
    Total number of downstreams: 3
       1: Register
          Protocol: pim-sm, UpTime: 00:02:15, Expires:  -
       2: Pos2/0/0
          Protocol: pim-sm, UpTime: 00:02:15, Expires: 00:03:15
       3: Pos3/0/0
          Protocol: pim-sm, UpTime: 00:02:15, Expires: 00:03:15
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Assume that Host A joins the multicast group G (FF0E::1), a RPT is set up between Router D and Router B, and a (*, G) entry is generated on Router D and Router B along the RPT. After the multicast source S (2001::5) sends a multicast packet to the multicast group G, Router A and Router D along the SPT generate an (S, G) entry.

Displaying IPv6 Multicast Routing Information - RP

```
[RouterD] display pim ipv6 routing-table
Vpn-instance: public net
 Total 1 (*, G) entry; 1 (S, G) entry

 (*, FF0E::1)
     RP: 2004::2 (local)
     Protocol: pim-sm, Flag: WC
     UpTime: 00:16:56
     Upstream interface: Register
         Upstream neighbor: NULL
         RPF prime neighbor: NULL
     Downstream interface(s) information:
     Total number of downstreams: 2
         1: Pos2/0/0
             Protocol: pim-sm, UpTime: 00:16:56, Expires: 00:02:34
         2: Pos3/0/0
             Protocol: pim-sm, UpTime: 00:07:56, Expires: 00:02:35
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Displaying IPv6 Multicast Routing Information - RP (Continued)

```
(2001::5, FF0E::1)
    RP: 2004::2 (local)
    Protocol: pim-sm, Flag: SWT ACT
    UpTime: 00:02:54
    Upstream interface: Register
        Upstream neighbor: NULL
        RPF prime neighbor: NULL
    Downstream interface(s) information:
    Total number of downstreams: 2
        1: Pos2/0/0
            Protocol: pim-sm, UpTime: 00:02:54, Expires:  -
        2: Pos3/0/0
            Protocol: pim-sm, UpTime: 00:02:54, Expires: 00:02:36
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

- PIM-SM Mechanism
- IPv6 PIM-SM Configuration
- **IPv6 PIM-SSM Fundamentals**
- Multicast Route Management
- Typical IPv6 Multicast Applications

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

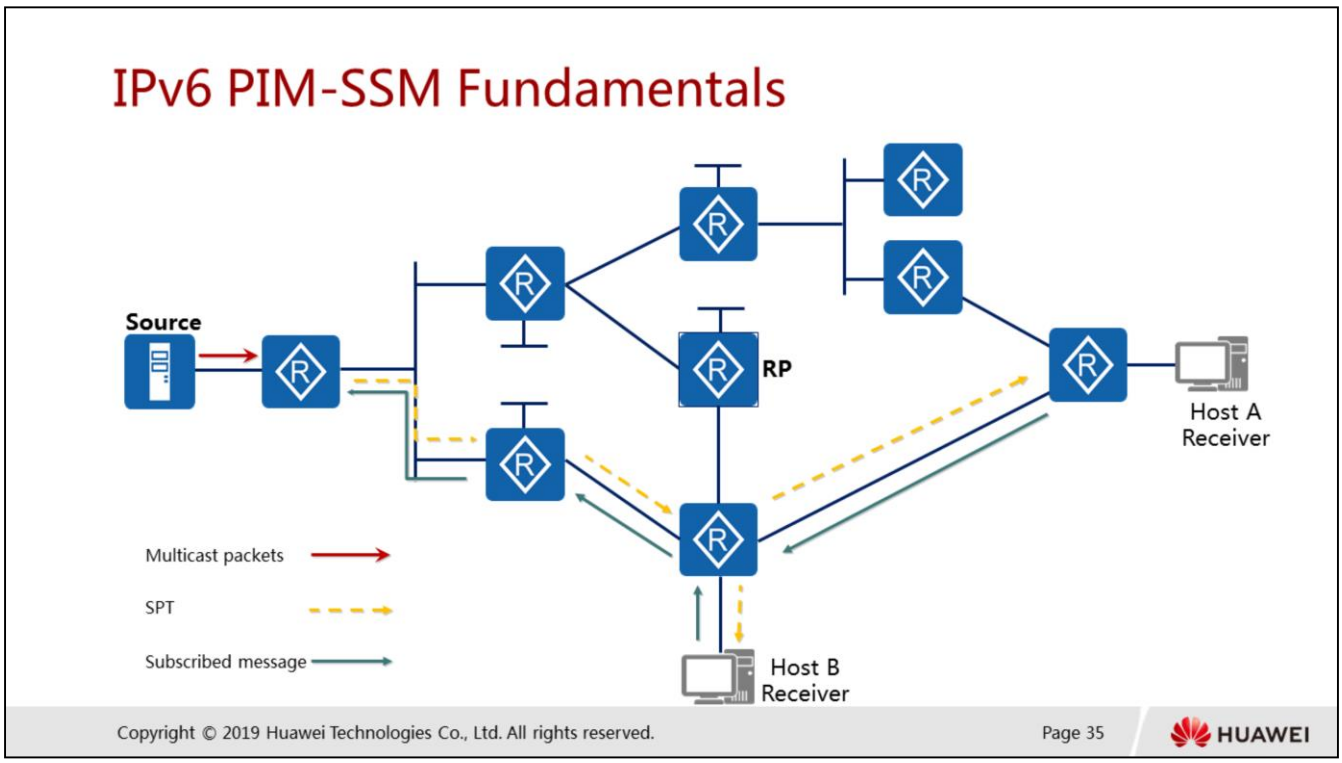
IPv6 PIM-SSM Overview

- The SSM model provides a source-specific multicast solution and is implemented using MLDv2 and some mechanisms of IPv6 PIM-SM. The last-hop router already knows the multicast source address through MLDv2. Therefore, the router can directly initiate the process of joining a group with a specific source to create the SPT from the multicast source to receivers on the SSM network.
  - A special multicast address of FF3x::/32 is defined, and no source discovery issue exists.
  - IPv6 PIM-SSM needs to work with MLDv2.
  - Compared with PIM-SM, PIM-SSM does not involve such complicated mechanisms as RP, BSR, RPT generation, and multicast source registration.
  - An SPT is directly generated based on a multicast source's unicast route to implement inter-domain multicast.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

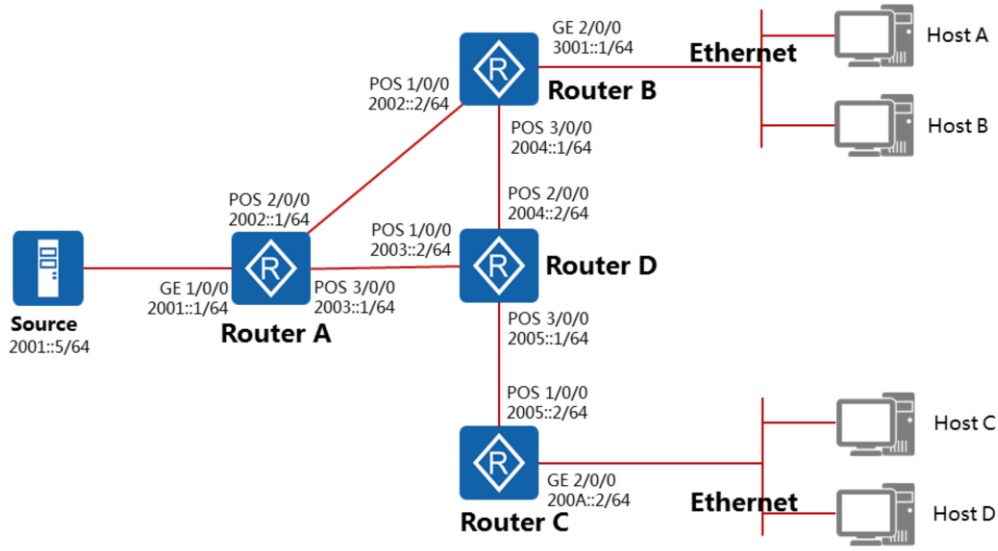
HUAWEI

- The implementation of IPv6 PIM-SSM involves neighbor discovery, DR election, and SPT generation.
  - Similar to IPv6 PIM-SM, IPv6 PIM-SSM completes neighbor discovery and DR election through Hello messages sent among multicast routers.
  - PIM-SSM also uses PIM-SM. The last-hop router determines whether to generate a RPT or SPT based on whether the multicast address is within the SSM group address range.



- In the SSM model, a channel is used to represent (S, G), and a Subscribed message is used to represent a Join message.
- If User A and User B need to receive information from the multicast source S, they send a Report message labeled with (include S, G) to the nearest querier through MLDv2. If User A and User B do not need to receive information from the multicast source S, they send a Report message labeled with (exclude S, G) or containing other multicast sources. The multicast source S is specified for receivers, no matter which Report message is used.
- After receiving a Report message, the querier checks whether the multicast address in the message is within the SSM group address range. If yes, the router sets up a multicast distribution tree based on the SSM model, and then sends a Subscribed message (also called Join message) to the specified source hop by hop. All the routers along the path create (S, G) entries, generating an SPT with the source S as the root and the receivers as the leaves. The SSM model uses this SPT as the transmission path.
- If the querier finds that the multicast address is beyond the SSM group range, it creates a multicast distribution tree based on IPv6 PIM-SM.

IPv6 PIM-SSM Configuration Example

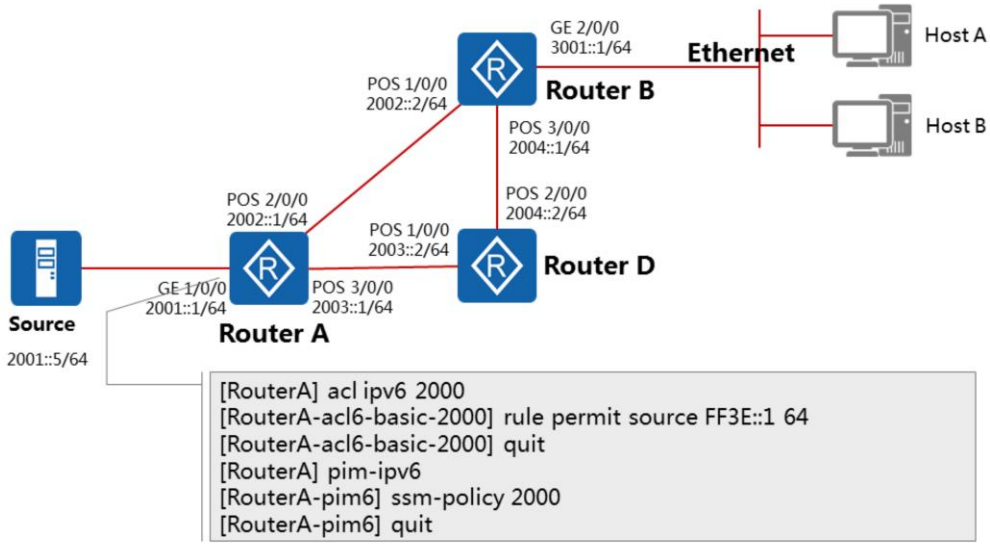


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Host A and Host C are multicast receivers on two leaf networks. These receivers connect to the multicast source through Router A, Router B, Router C, and Router D.
- MLDv2 must be run on the interfaces connected to hosts between Router B and N1 and between Router C and N2.
- Configuration roadmap:
  - Configure an IPv6 address for each router interface and an IPv6 unicast routing protocol.
    - Configure an IPv6 address and mask for each router interface.
    - Configure OSPFv3 on each router and set the process ID to1 and the area ID to 0 to ensure that Router A, Router B, Router C, and Router D can communicate at the network layer.
  - Enable IPv6 multicast routing on each router and IPv6 PIM-SM on each router interface.
  - Configure an IPv6 PIM-SSM multicast address range on each router.
  - Configure MLDv2 on interfaces connecting routers to hosts.
  - Verify the configuration.

Configuring Router A



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Set the IPv6 PIM-SSM multicast address range to FF3E::1 on Router A.
- The configurations on Router B, Router C, and Router D are similar to the configuration on Router A.
- The SSM model uses the subset of IPv6 PIM-SM. You must enable IPv6 PIM-SM on all routers on a network. An SSM group address range also needs to be configured. By default, the SSM group address range defined by IANA is used.
- If you want to receive information from the specified source S or other sources, an MLDv2 Report message containing the channel (S, G) must be sent. After receiving the message, the DR on the receiver side checks whether the multicast address G in the message is within the SSM group address range. If yes, the DR sends a Join message to multicast source S and creates an (S, G) entry on each router along the path to establish an SPT. Up to now, the SSM model is established. If the multicast address G is outside the SSM group address range or the source address S is not explicitly specified, the DR triggers the establishment of the ASM model based on IPv6 PIM-SM.
- The SSM model is implemented through the subset of IPv6 PIM-SM. Routers with IPv6 PIM-SM enabled have SSM processing capabilities. A router periodically sends Hello messages to discover PIM IPv6 neighbors and processes messages from neighbors. When a router joins an IPv6 PIM-SSM domain, you are advised to enable IPv6 PIM-SSM on all interfaces of a non-border router.
- Whether the multicast source's information is transmitted to receivers in IPv6 PIM-SSM or IPv6 PIM-SM mode depends on whether the multicast address of the channel (S, G) is within the SSM group address range. In the IPv6 PIM-SSM mode, the group address information is very important.
- If the SSM group address range is not specified, the system uses the FF3x::/12 network segment reserved by IANA for SSM as the default address range.
- Enter the system view.
  - **system-view**
- Enable IPv6 multicast routing.
  - **multicast ipv6 routing-enable**
- Enter the interface view.
  - **interface** *interface-type interface-number*
- Enable IPv6 PIM-SSM.
  - **pim ipv6 sm**
- Enter the PIM IPv6 view.
  - **pim-ipv6**
- Configure an IPv6 PIM-SSM group address range.
  - **ssm-policy** *basic-acl6-number*

Displaying IPv6 Multicast Routing Information

```
[RouterA] display pim ipv6 routing-table
Vpn-instance: public net
Total 1 (S, G) entry

 (2001::5, FF3E::1)
    Protocol: pim-ssm, Flag: LOC
    UpTime: 00:00:11
    Upstream interface: GigabitEthernet1/0/0
        Upstream neighbor: FE80::200:AFF:FE01:10D
        RPF prime neighbor: FE80::200:AFF:FE01:10D
    Downstream interface(s) information:
    Total number of downstreams: 1
        1: Pos2/0/0
            Protocol: pim-ssm, UpTime: 00:00:11,  Expires: 00:03:19
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- If Host A needs to receive information from the multicast source S (2001::5) to the multicast group G (FF3E::1), Router B sets up an SPT to the source. Router A and Router B on the SPT generate an (S, G) entry. No (S, G) entry exists on Router D outside the SPT.



Contents

- PIM-SM Mechanism
- IPv6 PIM-SM Configuration
- IPv6 PIM-SSM Fundamentals
- **Multicast Route Management**
- Typical IPv6 Multicast Applications

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Multicast route management refers to the control of multicast packet forwarding by creating or changing multicast routes, as well as checking and maintaining multicast forwarding paths.

Multicast Route Management (IPv6)

- The following table describes the multicast route management functions that allow you to create and maintain multicast forwarding paths.

Feature	Description
RPF check	Ensures that multicast data is forwarded through correct paths.
Multicast load splitting	Distributes multicast data to multiple equal-cost routes to reduce the load on individual paths.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Multicast route management ensures that multicast packets are forwarded efficiently through the correct paths. In multicast routing and forwarding, each multicast routing protocol creates and maintains its own routing table. The routing information from these tables is then used to create a general multicast routing table. Multicast routers use this general multicast routing table to determine optimal routes, according to multicast routing and forwarding policies. The optimal route information is then delivered to the MFIB, where multicast data forwarding is controlled.
- The MFIBs of network devices maintain a point-to-multipoint forwarding tree for the entire network, with a multicast source as the root and group members as leaves. Multicast route management provides a series of features to create and maintain multicast forwarding paths.

IPv6 Multicast Protocol Routing Table

- Multicast routing protocols maintain their own routing tables to guide multicast route management. The following is an example of a PIM routing table:

```
<HUAWEI> display pim ipv6 routing-table
VPN-Instance: public net
 Total 0 (*, G) entry; 1 (S, G) entry
 (FC00::2, FF3E::1)
    Protocol: pim-sm, Flag: SPT LOC ACT
    UpTime: 00:04:24
    Upstream interface: Vlanif20
        Upstream neighbor: FE80::A01:100:1
        RPF prime neighbor: FE80::A01:100:1
    Downstream interface(s) information:
    Total number of downstreams: 1
        1: Vlanif10
            Protocol: pim-sm, UpTime: 00:04:24, Expires: 00:02:47
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- **(FC00::2, FF3E::1)**: (S, and G) entry.
- **Protocol: pim-sm**: protocol type. The first Protocol field in an entry indicates the protocol that generates the entry, and the second Protocol field indicates the protocol that generates the downstream interfaces.
- **Flag: SPT LOC ACT**: flag of the PIM routing entry.
- **UpTime: 00:04:24**: The first UpTime field in an entry indicates how long the entry has existed, and the second UpTime field indicates how long a downstream interface has existed.
- **Upstream interface: Vlanif20**: upstream interface.
- **Upstream neighbor: FE80::A01:100:1**: upstream neighbor. NULL indicates that no upstream neighbor is available.
- **RPF prime neighbor: FE80::A01:100:1**: RPF neighbor. NULL indicates that no RPF neighbor is available.
- **Downstream interface(s) information**: downstream interface information.
- **Total number of downstreams: 1**: number of downstream interfaces.
- **Expires: 00:02:47**: aging time of the downstream interface.

IPv6 Multicast Routing Table

- A multicast routing table is generated and maintained by the multicast route management module of a router. If a router supports multiple multicast protocols, its multicast routing table contains the optimal routes selected from the routing tables of these protocols.

```
<HUAWEI> display multicast ipv6 routing-table
IPv6 multicast routing table
 Total 1 entry

 00001. (FC00::2, FF3E::1)
      Uptime: 00:00:14
      Upstream Interface: Vlanif10
      List of 1 downstream interface
        1:  Vlanif20
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- **00001. (FC00::2, FF3E::1)**: entry 00001, which is in the format of (S, G).
- **Uptime: 00:00:14**: time when the multicast routing entry was updated.
- **Upstream Interface: Vlanif10**: upstream interface.
- **List of 1 downstream interface**: list of downstream interfaces.

MFIB

- An MFIB is generated by the routing management module according to the multicast routing table and is used to forward multicast data. An MFIB has the same functions as a unicast FIB.

```
<HUAWEI> display multicast ipv6 forwarding-table
IPv6 Multicast Forwarding Table
Total 2 entries, 2 matched

00001. (FC00:1::3, FF1E::1)
    MID: 10, Flags: ACT
    Uptime: 02:54:43, Timeout in: 00:02:27
    Incoming interface: Vlanif10
    List of 1 outgoing interfaces:
      1: LoopBack0
        Activetime: 00:23:15
    Matched 0 packets(0 bytes), Wrong If 0 packets
    Forwarded 0 packets(0 bytes)
```

Multicast data is forwarded according to the MFIB. Each multicast forwarding entry records statistics about packets that are forwarded according to the entry.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

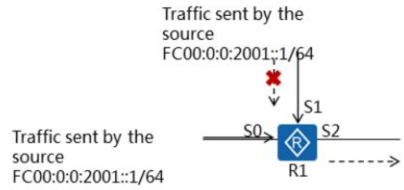
HUAWEI

- **00001. (FC00:1::3, FF1E::1)**: entry 00001, which is in the format of (S, G).
- **MID: 10**: used to rapidly search the MFIB.
- **Flags: ACT**: flag of the multicast forwarding entry.
- **UpTime: 02:54:43**: time when the multicast forwarding entry already exists.
- **Timeout in: 00:03:26**: timeout period of the multicast forwarding entry.
- **Incoming interface: Vlanif10**: inbound interface of the entry.
- **List of 1 outgoing interfaces**: list of outbound interfaces of the entry.
- **Activetime: 00:23:15**: time when the outbound interface already exists.
- **Matched 38264 packets(1071392 bytes)**: number of packets matching the entry.
- **Wrong If 0 packets**: number of packets from the incorrect interface.
- **Forwarded 38264 packets(1071392 bytes)**: number of forwarded packets.

RPF Check

- RPF check
  - Ensures that multicast traffic is forwarded along a correct path to prevent loops.
  - Based on a unicast routing table.
  - Check process:
    - A router forwards multicast packets only from its interface connected to a multicast source, and discards multicast packets received in other cases.

IPv6 routing table	
Network Segment	Interface
FC00:0:0:2001::1/64	S0
FC00:0:0:3001::1/64	S1
FC00:0:0:4001::1/64	S2

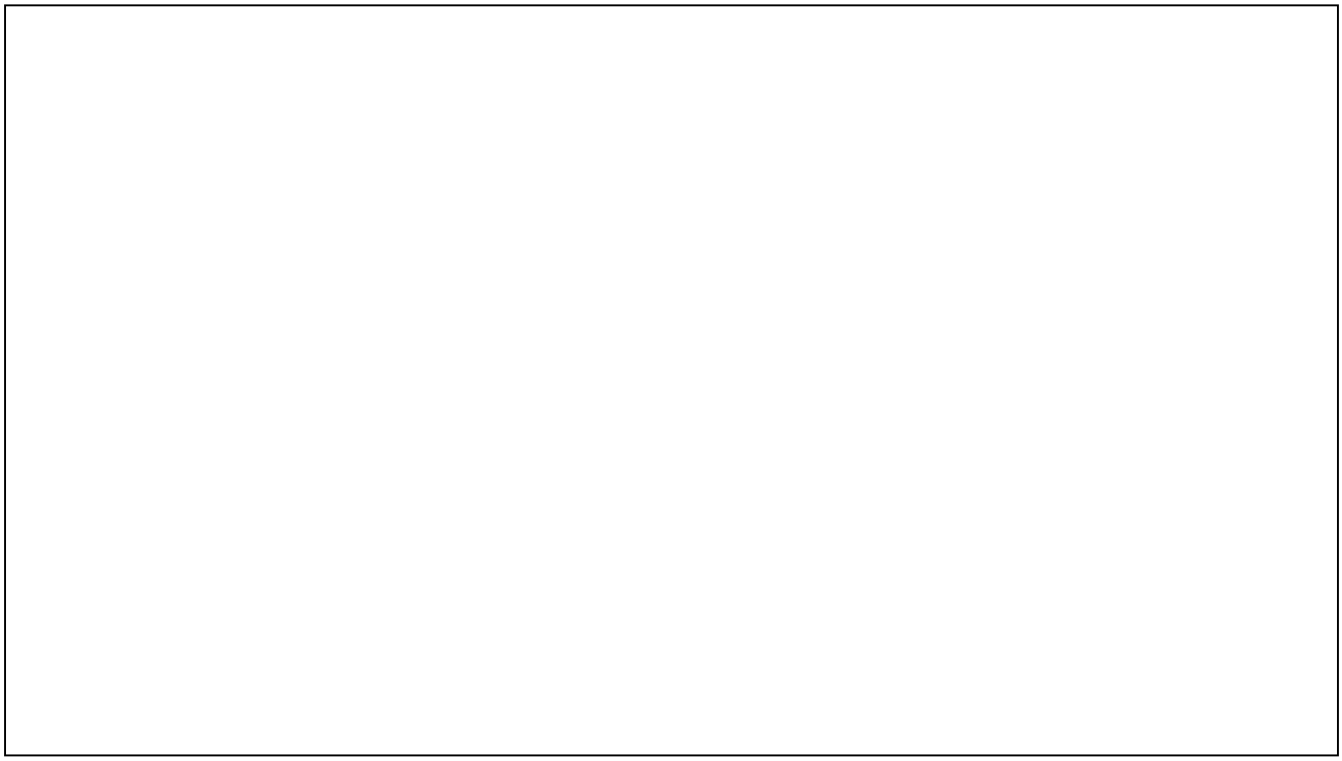


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- RPF check principles
  - When a router receives a multicast packet, it searches the unicast routing table for the route to the source address of the packet. After finding the route, the router checks whether the outbound interface for the route is the same as the inbound interface of the multicast packet. If they are the same, the router considers that the multicast packet has been received from a correct interface. This ensures correct forwarding paths for multicast packets.
  - If the equal-cost routes are in the same routing table, the router selects the route with the largest next-hop address as the RPF route.
  - RPF checks can be performed based on unicast, Multiprotocol Border Gateway Protocol (MBGP), or static multicast routes. The priority order of these routes is static multicast routes > MBGP routes > unicast routes.

- Topology description
  - A multicast stream sent from the source FC00:0:0:2001::1/64 arrives at interface S1 of the router. The router checks the routing table and finds that the multicast stream from this source should arrive at interface S0. The RPF check fails, and the router drops the multicast stream.
  - A multicast stream sent from the source FC00:0:0:2001::1/64 arrives at interface S0 of the router. The router checks the routing table and finds that the RPF interface is also S0. The RPF check succeeds, and the multicast stream is correctly forwarded.



- Multicast routing protocols determine upstream and downstream neighbors and create multicast routing entries based on unicast routes. The RPF check mechanism enables multicast data streams to be transmitted along the multicast distribution tree and prevents loops on forwarding paths.
- If a router searches the unicast routing table to perform a RPF check on every multicast data packet received, many system resources are consumed. To save system resources, a router first searches for the matching (S, G) entry after receiving a data packet sent from a source to a group.
- If no matching (S, G) entry is found, the router performs a RPF check to find the RPF interface for the packet. The router then creates a multicast route with the RPF interface as the upstream interface and delivers the route to the multicast forwarding information base (MFIB). If the RPF check succeeds, the inbound interface of the packet is the RPF interface, and the router forwards the packet to all the downstream interfaces in the forwarding entry. If the RPF check fails, the packet has been forwarded along an incorrect path, so the router drops the packet.
- If a matching (S, G) entry is found and the inbound interface of the packet is the same as the upstream interface in the entry, the router replicates the packet to all downstream interfaces specified in the entry.
- If a matching (S, G) entry is found but the inbound interface of the packet is different from the upstream interface in the entry, the router performs a RPF check on the packet. Based on the RPF check result, the router processes the packet as follows:
    - If the RPF interface is the same as the upstream interface in the entry, the (S, G)

entry is correct and the packet has been forwarded along an incorrect path.

- If the RPF interface is different from the upstream interface in the entry, the (S, G) entry is outdated, and the router changes the upstream interface in the entry to be the same as the RPF interface. The router then compares the RPF interface with the inbound interface of the packet. If the inbound interface is the RPF interface, the router replicates the packet to all downstream interfaces specified in the (S, G) entry.

Multicast Load Splitting

- Load splitting is different from load balancing.
- Load splitting provides a way to distribute data streams destined for the same destination to multiple equal-cost paths, which may not result in a balanced traffic load on the paths.
- Load balancing is a special form of load splitting and distributes even data traffic loads on multiple equal-cost paths.

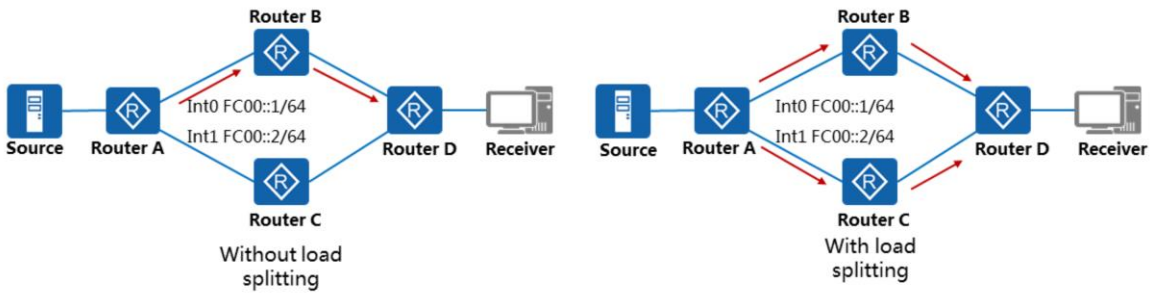
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- By default, if multiple equal-cost routes exist during multicast packet forwarding, a router selects the route with the largest next-hop address only from the IGP routing table as the RPF route.

Multicast Forwarding Without and with Multicast Load Splitting

- Multicast load splitting enables a router to distribute multicast traffic to multiple equal-cost routes, instead of selecting only one route based on the RPF check policy.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The multicast source (Source) sends multicast streams to group G. Router A and Router D run an Interior Gateway Protocol (IGP), OSPF for example, to implement IP interworking. Two equal-cost paths are available: Router A -> Router B -> Router D and Router A -> Router C -> Router D.
- Based on the default RPF check policy, multicast streams are forwarded through interface Int1 of Router A because Int1 has a larger IP address than Int0. After multicast load splitting is configured on Router A, Router A does not select forwarding paths by comparing the next-hop IP addresses. Instead, multicast streams are forwarded along both of the two equal-cost paths.

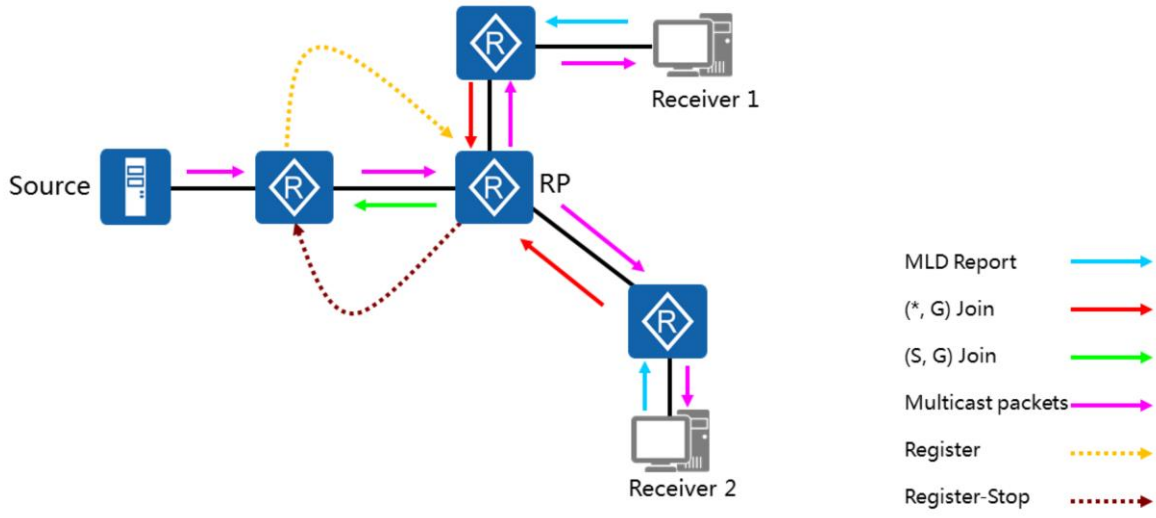


Contents

- PIM-SM Mechanism
- IPv6 PIM-SM Configuration
- IPv6 PIM-SSM Fundamentals
- Multicast Route Management
- **Typical IPv6 Multicast Applications**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Intra-domain Multicast: PIM-SM+MLD



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

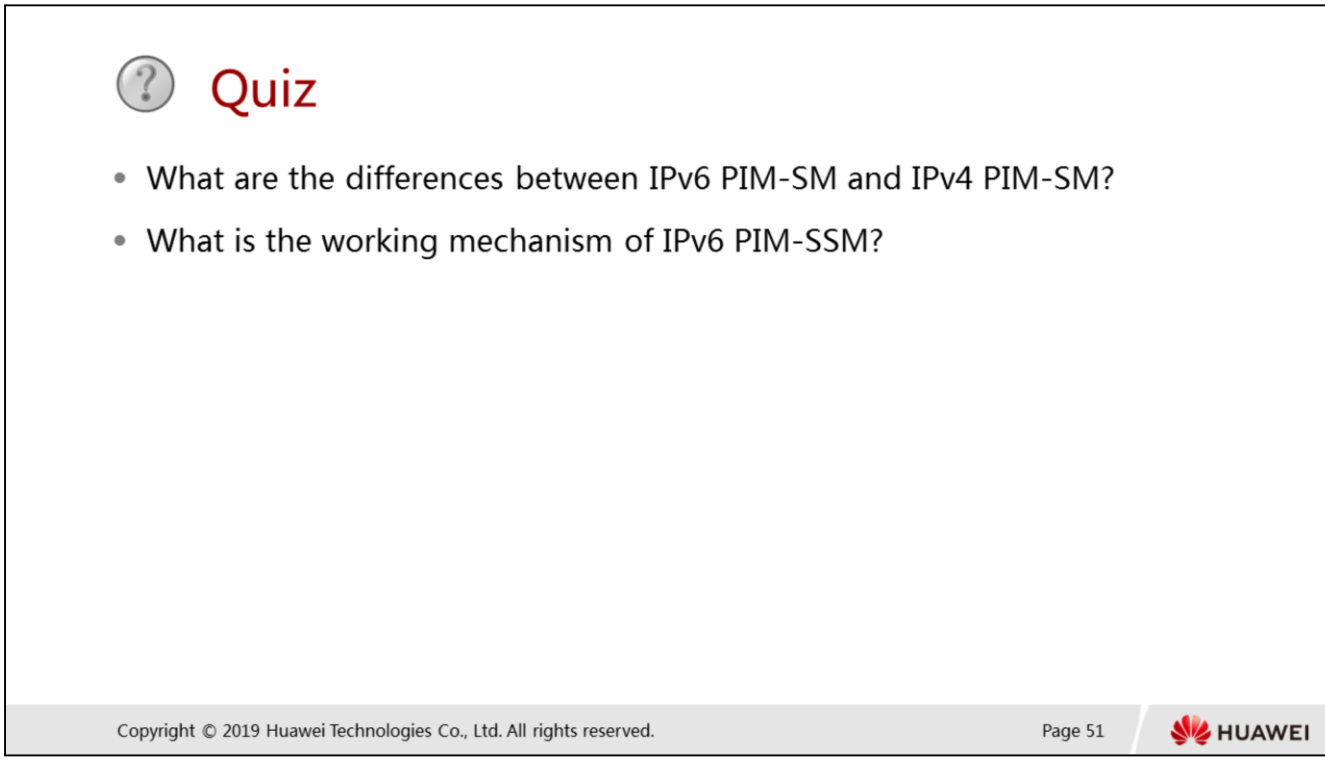
HUAWEI

- As shown in the figure, the routers in the domain run PIM-SM and the interfaces connected to the receivers run MLDv1. MLDv1 in IPv6 multicast is equivalent to IGMPv2 in IPv4 multicast, and is used to obtain multicast group member information and notify upper-layer protocols.
- All routers in the domain obtain RP information through static configuration, dynamic election the BSR mechanism, or automatic discovery.
- The penultimate-hop router connected to an IPv6 receiver receives an MLD Report message from the receiver, and sends a (*, G) Join message to the upstream routers through the RPF neighbors until the RP receives the (*, G) Join message. All routers along the path create a (*, G) entry, generating a RPT with the RP as the root.
- The multicast source sends multicast data. The first-hop router sends a PIM Register message to the RP. After receiving the message, the RP replies with a Register-Stop message.
- The RP sends an (S, G) Join message to the first-hop router through the RPF neighbors. The routers along the path create an (S, G) entry, generating an SPT with the first-hop router as the root.
- Multicast data arrives at the RP along the SPT and is forwarded based on a (*, G) entry. The routers along the path generate an (S, G) entry, and multicast data reaches a receiver.

Section Summary

- PIM-SM Mechanism
- IPv6 PIM-SM Configuration
- IPv6 PIM-SSM Fundamentals
- Multicast Route Management
- Typical IPv6 Multicast Applications

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- What are the differences between IPv6 PIM-SM and IPv4 PIM-SM?
  - Their addresses are different, but their protocol mechanisms are the same.
- What is the working mechanism of IPv6 PIM-SSM?
  - The implementation of IPv6 PIM-SSM involves neighbor discovery, DR election, and SPT generation.
    - Simliar to IPv6 PIM-SM, IPv6 PIM-SSM completes neighbor discovery and DR election through Hello messages sent among multicast routers.
    - PIM-SSM also uses PIM-SM. The last-hop router determines whether to generate a RPT or SPT based on whether the multicast address is within the SSM group address range.

Thank You

www.huawei.com

HCIE-Routing & Switching

— Inter-AS Option Modes for MPLS BGP VPN

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- MPLS seamlessly integrates the flexibility of IP routing and simplicity of Asynchronous Transfer Mode (ATM) label switching. MPLS introduces a connection-oriented control plane into the connectionless IP network, adding a means of managing and operating the IP network. Therefore, the VPNs (or MPLS VPNs) using MPLS IP networks as backbone networks are highly evaluated by carriers, and become an important means of providing value-added services.
- This document describes the fundamentals and configurations of the Option A, Option B, and Option C solutions in Inter-AS MPLS BGP VPN scenarios.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to:
  - Understand the fundamentals and configurations of Option A
  - Understand the fundamentals and configurations of Option B
  - Understand the fundamentals and configurations of Option C

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

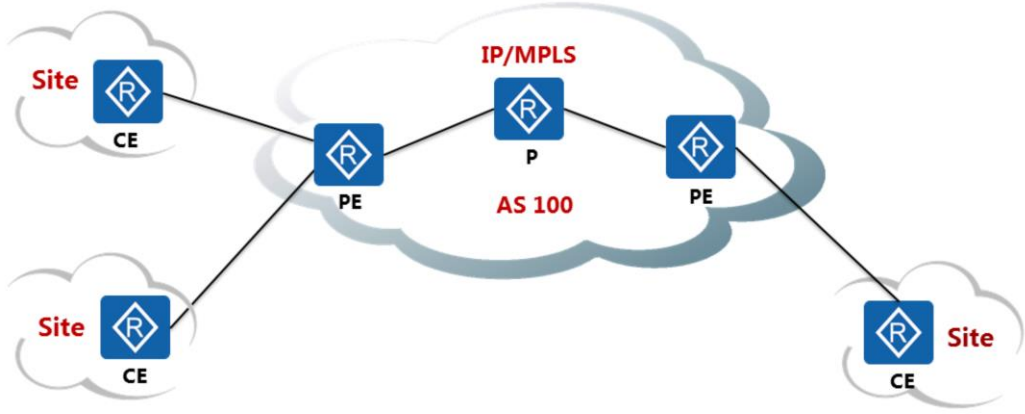
Contents

- **Background**
- Fundamentals
- Configurations

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Background of the Inter-AS MPLS BGP VPN Solution (1/3)

- Traditional MPLS BGP VPN Solution

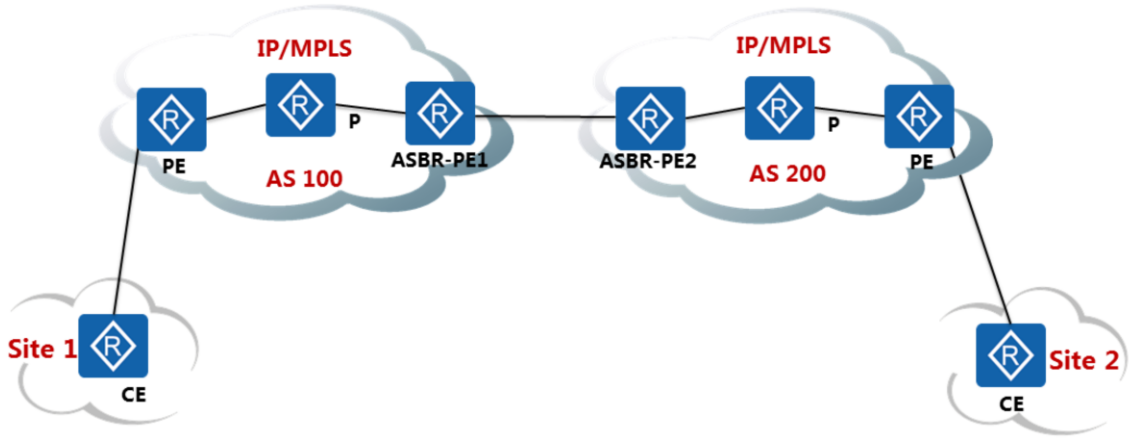


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Introduction:
- With the wide application of MPLS VPN solutions, branches of a large enterprise or networks of collaborative enterprises span multiple ASs.
- For example:
- Generally, an MPLS VPN architecture runs within an AS in which the routing information in any VPN instance is flooded on demand. However, the VPN routing information within the AS cannot be flooded to a different AS.
- On the network shown in this slide, an MPLS-based VPN connects various branches of a private network, forming a unified network. Also, it provides interconnection control for different VPNs. A customer edge (CE) is a user edge device. A provider edge (PE) is a service provider's router located on the edge of the backbone network. A provider (P) is a backbone router on the service provider's network and is not directly connected to a CE.

Background of the Inter-AS MPLS BGP VPN Solution (2/3)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- If two sites of the same VPN are located in different ASs, is the traditional MPLS BGP VPN solution still suitable for service deployment?
- The answer is no. In this case, the PEs with the same VPN instance configured cannot establish an IBGP peer relationship or establish peer relationships with an RR. Instead, the PEs need to establish an EBGP peer relationship to transmit VPNv4 routes.
- To enable the exchange of VPN routes between different ASs, the inter-AS MPLS VPN model is introduced. This model is an extension of the existing protocol and MPLS VPN framework. Through this model, the route prefix and label information can be advertised over the links between different ASs.

Background of the Inter-AS MPLS BGP VPN Solution (3/3)

- This course will explain the following three inter-AS VPN solutions:
  - Inter-AS MPLS BGP VPN Option A (inter-AS VPN Option A): VPN instances spanning multiple ASs are bound to dedicated interfaces of ASBRs to manage their own VPN routes. This solution is also called VRF-to-VRF.
  - Inter-AS MPLS BGP VPN Option B (inter-AS VPN Option B): ASBRs advertise labeled VPN-IPv4 routes to each other through MP-EBGP. This solution is also called EBGP redistribution of labeled VPN-IPv4 routes.
  - Inter-AS MPLS BGP VPN Option C (inter-AS VPN Option C): PEs or RRs advertise labeled VPN-IPv4 routes to each other through multi-hop MP-EBGP. This solution is also called multi-hop EBGP redistribution of labeled VPN-IPv4 routes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

- Background
- **Fundamentals**
  - Option A Fundamentals
  - Option B Fundamentals
  - Option C Fundamentals
- Configurations

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

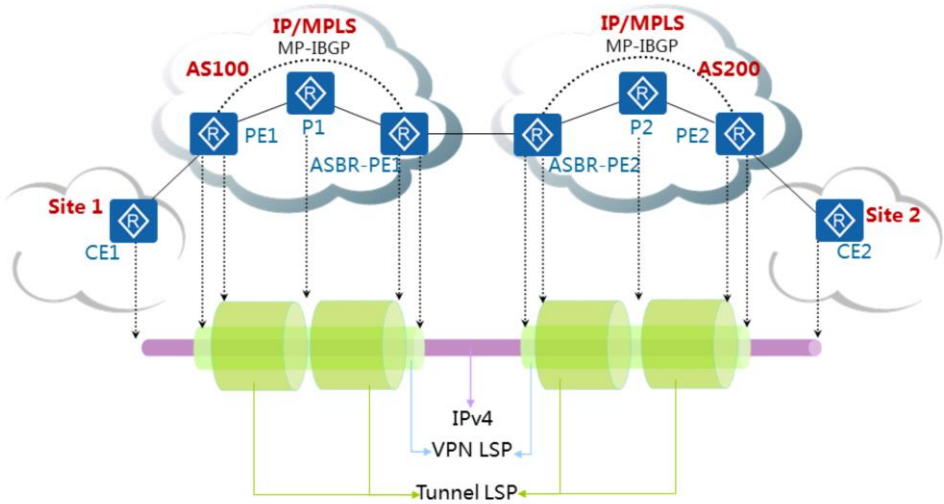
Inter-AS MPLS BGP VPN Option A

- As a basic BGP/MPLS IP VPN application in the inter-AS scenario, Option A does not need special configurations and MPLS does not need to run between ASBRs. In this mode, ASBRs of the two ASs are directly connected, and they act as PEs in their own ASs. Each ASBR takes the peer ASBR as its CE and advertises IPv4 routes to the peer ASBR through EBGP.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Inter-AS MPLS BGP VPN Option A — Topology

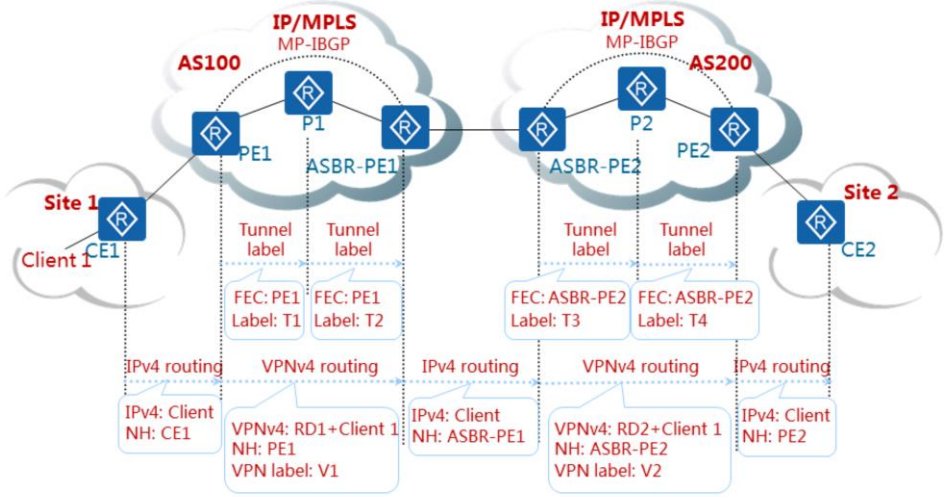


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In this solution, ASBR-PEs are directly connected. Two ASBR-PEs are connected to each other through multiple interfaces, including sub-interfaces. Each interface is associated with a VPN, and each ASBR-PE regards its peer as a CE. Therefore, the interfaces (including sub-interfaces) that connect the ASBR-PEs need to be bound to VRFs. In addition, VPNv4 routes need to be converted into common IPv4 routes and then advertised from one AS to the other through an EBGP peer relationship. There is no need to enable MPLS on the connected ASBR-PEs. This solution does not extend the service attributes in MPLS BGP VPN.

Inter-AS MPLS BGP VPN Option A — Control Plane



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

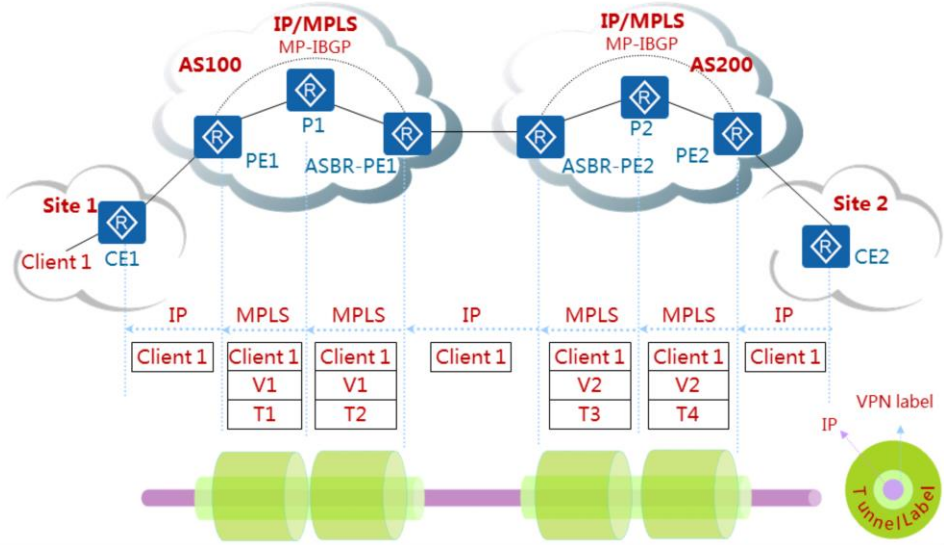
HUAWEI

- Let's take route advertisement in one direction on the control plane as an example. Suppose there is a host named Client 1 on Site 1. The route to Client 1 needs to be advertised from CE1 to CE2 through AS100 and AS200.
  - In AS100, PE1 uses LDP to assign P1 an outer tunnel label T1, which is associated with the route to PE1.
  - In AS100, P1 uses LDP to assign ASBR-PE1 an outer tunnel label T2, which is associated with the route to PE1.
  - In AS200, ASBR-PE2 uses LDP to assign P2 an outer tunnel label T3, which is associated with the route to ASBR-PE2.
  - In AS200, P2 uses LDP to assign PE2 an outer tunnel label T4, which is associated with the route to ASBR-PE2.
  - CE1 advertises the route destined for Client 1 to PE1, and the next hop of the route is CE1's interface address.
  - PE1 encapsulates the received IPv4 route to Client 1 into a VPNv4 route, changes the Next_Hop to PE1 in the related MP-BGP message, allocates the VPN label V1 to the route, and then advertises the route to ASBR-PE1.
  - ASBR-PE1 restores the received VPNv4 route to an IPv4 route and advertises it to ASBR-PE2 with the next hop being ASBR-PE1.
  - ASBR-PE2 encapsulates the received IPv4 route to Client 1 into a VPNv4 route, changes the Next_Hop to ASBR-PE2 in the related MP-BGP message, allocates the

VPN label V2 to the route, and then advertises the route to PE2.

- PE2 restores the received VPNv4 route to the IPv4 route to Client 1 and advertises it to CE2, with the next hop being PE2.

Inter-AS MPLS BGP VPN Option A — Forwarding Plane



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Now, Let's look at packet transmission on the forwarding plane. The packet transmission process from CE2 to CE1 is used as an example to illustrate the work flow on the forwarding plane.
  - CE2 sends an IP packet destined for Client 1 to PE2.
  - Upon receipt, PE2 encapsulates the IP packet with the VPN label V2 and then the outer label T4 and forwards the packet to P2.
  - P2 swaps the outer label T4 for T3 and forwards the IP packet to ASBR-PE2.
  - ASBR-PE2 removes both labels from the received packet and forwards the unlabeled IP packet to ASBR-PE1.
  - ASBR-PE1 encapsulates the received IP packet with the VPN label V1 and then the outer label T2 and forwards the packet to P1.
  - P1 swaps the outer label T2 for T1 and forwards the packet to PE1.
  - PE1 removes both labels from the received packet and forwards the unlabeled IP packet to CE1.

Features of Inter-AS MPLS BGP VPN Option A

- Easy configuration: MPLS is not required between ASBRs, and no special configuration is required for inter-AS connections.
- Poor scalability: ASBRs need to manage all VPN routes and a VPN instance needs to be configured for each VPN. This results in numerous VPN-IPv4 routes on the ASBRs. In addition, because common IP forwarding is performed between ASBRs, each inter-AS VPN instance must be bound to a different AC interface. This poses a high requirement on PEs. If a VPN spans multiple ASs, the intermediate ASs must support VPN services. This requires complex configurations and greatly affects the operations in the intermediate ASs. If only a few inter-AS VPN instances are used, Option A is recommended.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

- Background
- **Fundamentals**
  - Option A Fundamentals
  - Option B Fundamentals
  - Option C Fundamentals
- Configurations

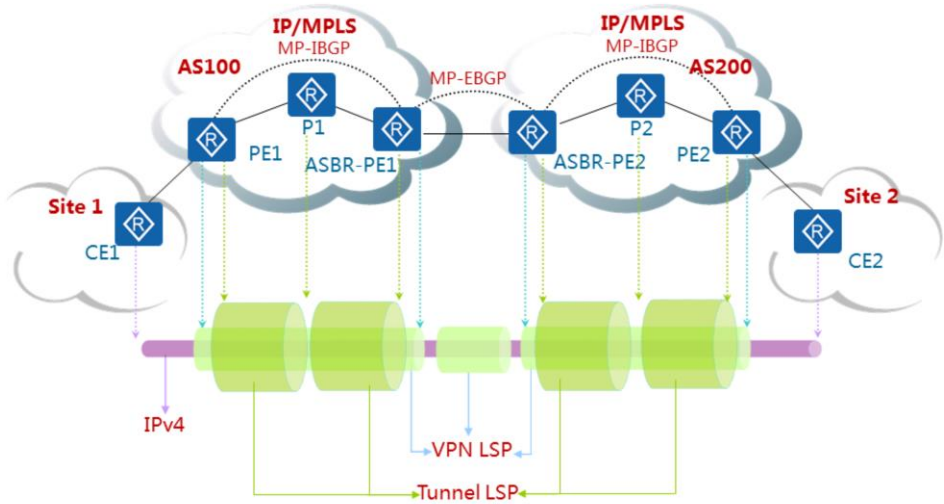
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Inter-AS MPLS BGP VPN Option B

- In Option B, two ASBRs use MP-EBGP to exchange labeled VPN-IPv4 routes that have been received from the PEs in their own ASs.
- In Option B, an ASBR advertises all the VPN-IPv4 routes received from an external AS to its local AS. It also advertises all the VPN-IPv4 routes received from the other devices in the same AS to the external AS. In the basic MPLS VPN implementation, a PE stores only the VPN routes that match the VPN targets of the VPN instance configured on the PE. You can disable RT-based route filtering on the ASBRs that advertise routes. Therefore, you do not need to create VPN instances on ASBRs or bind the VPN instances to any interfaces.
- RRs can be deployed on the network to transmit VPN routes from the user side.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Inter-AS MPLS BGP VPN Option B — Topology

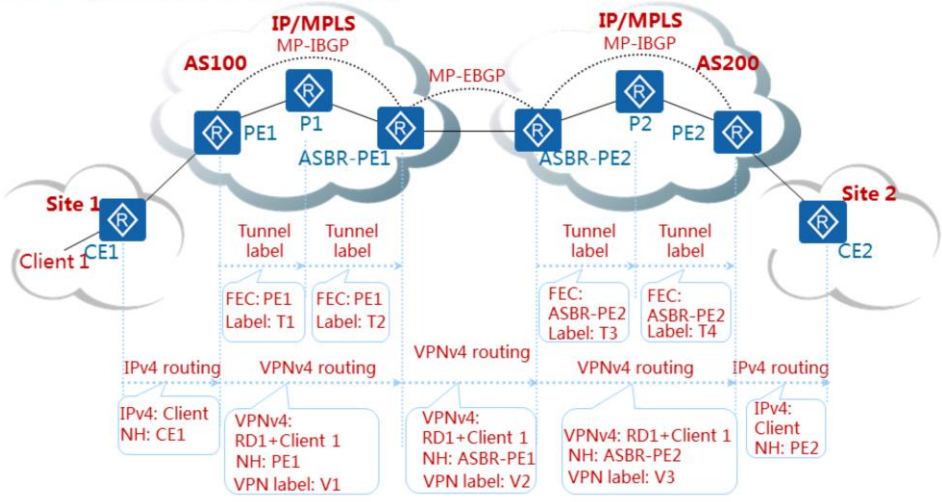


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In the Option B solution, each PE advertises VPNv4 routes to its connected ASBR-PE or VPN RR through MP-IBGP. The ASBR-PE is the client device of the PE. The ASBR-PE in one AS advertises the VPNv4 routes to the ASBR-PE in the other AS through MP-EBGP. The ASBR-PE that receives the VPNv4 routes then advertises the routes to the PE in the same AS.

Inter-AS MPLS BGP VPN Option B — Control Plane (Without RRs)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Let's take route advertisement in one direction on the control plane as an example. Suppose there is a host named Client 1 on Site 1.

1. In AS100, PE1 uses LDP to assign P1 an outer tunnel label T1, which is associated with the route to PE1.

2. In AS100, P1 uses LDP to assign ASBR-PE1 an outer tunnel label T2, which is associated with the route to PE1.

3. In AS200, ASBR-PE2 uses LDP to assign P2 an outer tunnel label T3, which is associated with the route to ASBR-PE2.

4. In AS200, P2 uses LDP to assign PE2 an outer tunnel label T4, which is associated with the route to ASBR-PE2.

5. CE1 advertises the route destined for Client 1 to PE1, and the next hop of the route is CE1's interface address.

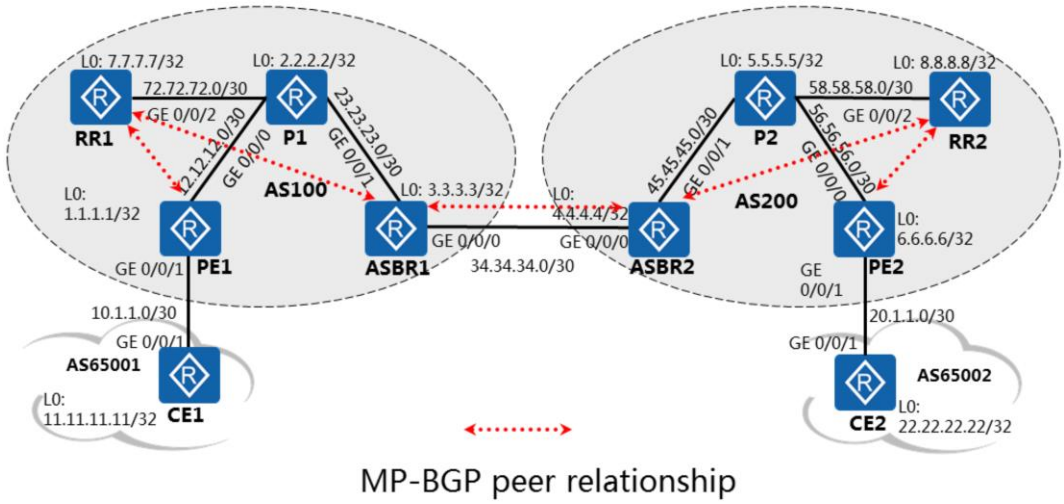
6. PE1 encapsulates the received IPv4 route to Client 1 into a VPNv4 route, changes the Next_Hop to PE1 in the related MP-IBGP message, allocates the VPN label V1 to the route, and then advertises the route to ASBR-PE1.

7. ASBR-PE1 advertises the VPNv4 route destined for Client 1 to ASBR-PE2 through MP-EBGP, changes the route's next hop to ASBR-PE1, and allocates a new VPN label V2 to the route.

8. ASBR-PE2 advertises the received VPNv4 route to PE2 through MP-IBGP, changes the route's next hop to itself, and allocates a new VPN label V3 to the route.

9. PE2 restores the received VPNv4 route to the IPv4 route to Client 1 and advertises it to CE2, with the next hop being PE2.

Inter-AS MPLS BGP VPN Option B — Control Plane (With RRs)

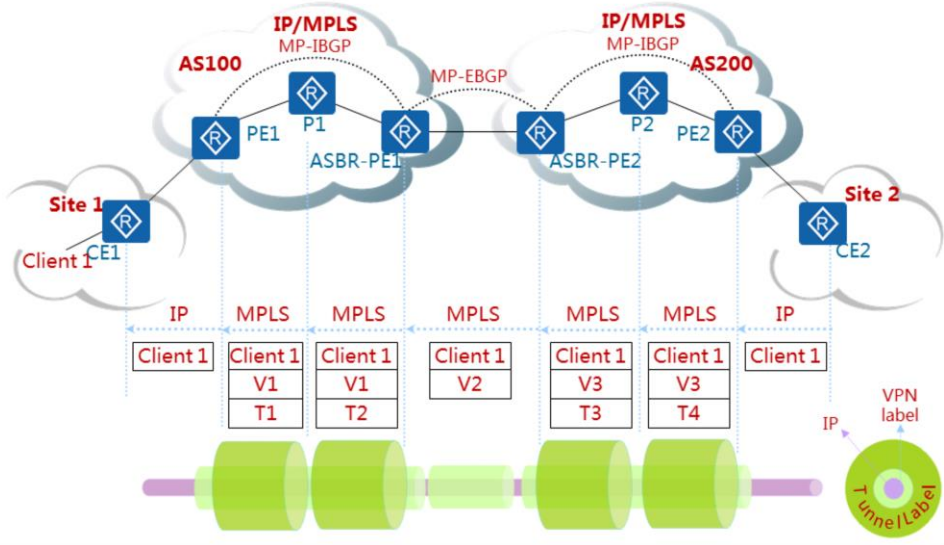


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- If a large number of VPN instances are required, standalone RRs can be deployed. As shown in this figure, the PE and ASBR in each AS establish MP-BGP peer relationships only with the RR. The RR in each AS reflects routes, avoiding the need to establish a BGP peer relationship between the PE and ASBR.
- An RR transmits only VPNv4 routes on the control plane and does not forward data traffic on the forwarding plane.

Inter-AS MPLS BGP VPN Option B — Forwarding Plane



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The packet transmission process from CE2 to CE1 is used to illustrate the work flow on the forwarding plane.
  - CE2 sends an IP packet destined for Client 1 to PE2.
  - Upon receipt, PE2 encapsulates the IP packet with the VPN label V3 and then the outer label T4 and forwards the packet to P2.
  - P2 swaps the outer label T4 for T3 and forwards the IP packet to ASBR-PE2.
  - ASBR-PE2 removes the outer label from the received packet, swaps the VPN label V3 for V2, and forwards the packet carrying only the VPN label V2 to ASBR-PE1.
  - Upon receipt, ASBR-PE1 swaps the VPN label V2 for V1, adds the outer tunnel label T2, and then forwards the packet to P1.
  - P1 swaps the outer label T2 for T1 and forwards the packet to PE1.
  - PE1 removes both labels from the received packet and forwards the unlabeled IP packet to CE1.

Features of Inter-AS MPLS BGP VPN Option B

- Unlike Option A, Option B is not restricted by the number of links between ASBRs.
- VPN routing information is stored on and forwarded by ASBRs. If a large number of VPN routes exist, the overloaded ASBRs tend to become fault points. Therefore, the ASBRs that establish an MP-EBGP peer relationship maintain VPN routing information only and do not forward IP packets on the public network.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. Background
2. **Fundamentals**
   - Option A Fundamentals
   - Option B Fundamentals
   - Option C Fundamentals
3. Configurations

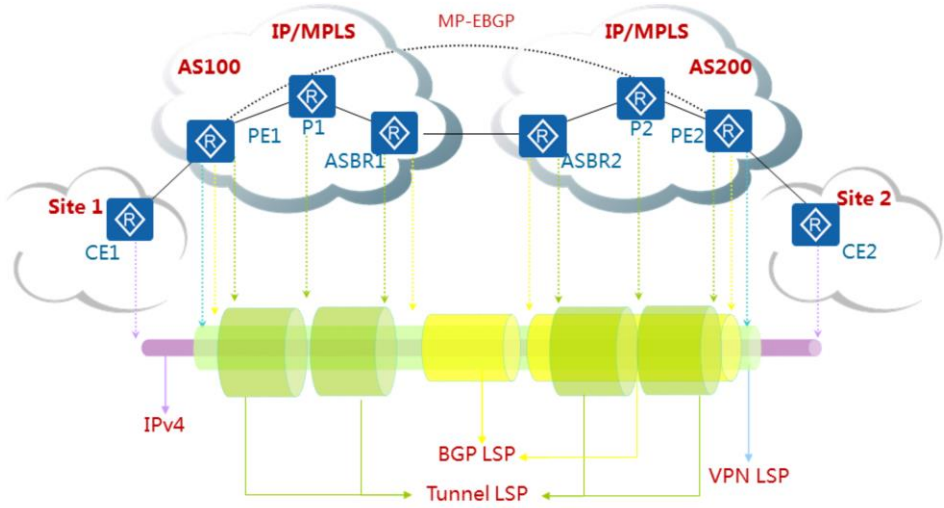
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Inter-AS MPLS BGP VPN Option C — Solution 1

- In Option C mode, the ASBR in each AS advertises labeled IPv4 routes to its connected PE through MP-IBGP, and advertises the labeled IPv4 routes received from the connected PE to the peer ASBR in the other AS. ASBRs in the transit AS also advertise labeled IPv4 routes. Therefore, a BGP LSP is established between the ingress PE and egress ASBR.
- PEs in different ASs establish a multi-hop EBGP connection with each other to exchange VPNv4 routes.
- ASBRs do not store VPN-IPv4 routes or advertise VPNv4 routes to each other.
- RRs can be deployed on a large-scale network to transmit user-side routes. Specifically, PEs and RRs establish MP-IBGP peer relationships. RR1 and RR2 establish an MP-EBGP peer relationship. Routes are transmitted through the PE1 <-> RR1 <-> RR2 <-> PE2 paths, and no BGP peer relationship needs to be established between the PEs. If a large number of VPNs exist, RRs help reduce the workload of PEs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Option C Solution 1— Topology

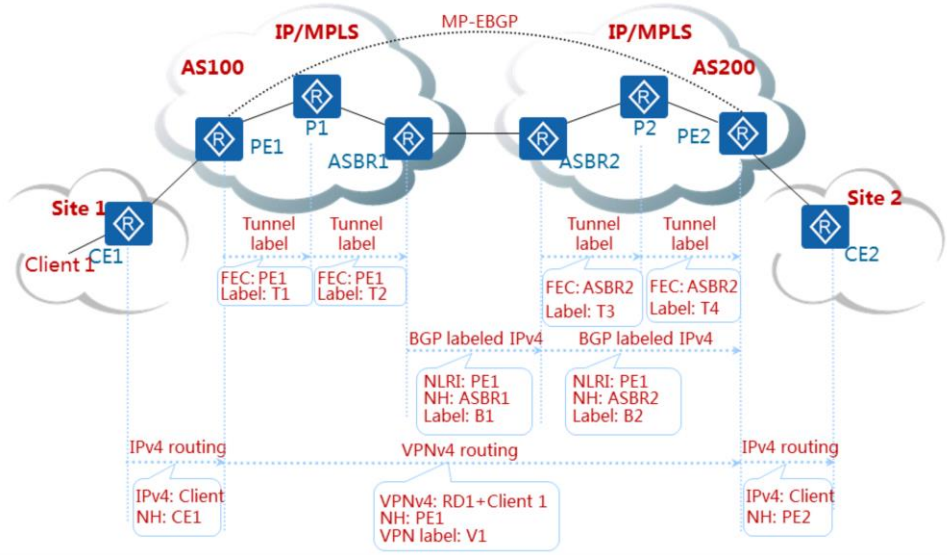


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In the Option C solution, ASBRs do not maintain or advertise VPNv4 routes. Therefore, the ASBR-PE routers are changed to ASBRs, as shown in the figure. An ASBR only needs to maintain all the labeled routes to a PE and use EBGP to advertise these labeled routes to its peer in the other AS. The ASBRs in a transit AS also need to use EBGP to advertise the labeled IPv4 routes. Therefore, a BGP LSP is established between the PEs in different ASs so that a multi-hop MP-EBGP connection can be established between the PEs for them to advertise VPNv4 routes.
- If the P router in each AS knows the routes to the PE in the other AS, data forwarding is simple. However, if the P router does not know the routes, the PE adds a Layer 3 label to the VPN data received from the CE. The inner label is the VPN label associated with the VPN route and is allocated by the peer PE, the intermediate label is the label allocated by the ASBR and is associated with the route to the peer PE, and the outer label is the label associated with the route to the next hop ASBR.
- To further expand performance, a multi-hop MP-EBGP session can be established between the VPN RRs in different ASs. The VPN RRs do not change the Next_Hop attribute when advertising VPNv4 routes. In each AS, the PE establishes an MP-IBGP session only with the VPN RR.
- Note: To facilitate illustration, a symmetric LSP is used in this example, as shown in the figure. However, the LSP structures in different ASs are not symmetric. For details, see the following slides.

Option C Solution 1— Control Plane (Without RRs)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

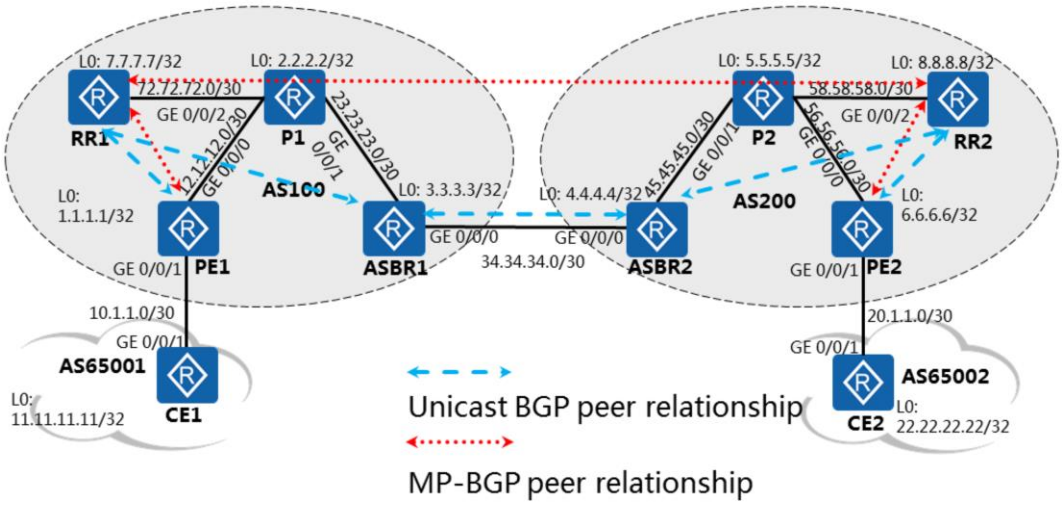
HUAWEI

- Let's take route advertisement in one direction on the control plane as an example. Suppose there is a host named Client 1 on Site 1, and the P router in each AS does not have routes to the peer PE in the other AS.
  - In AS100, PE1 uses LDP to assign P1 an outer tunnel label T1, which is associated with the route to PE1.
  - In AS100, P1 uses LDP to assign ASBR-PE1 an outer tunnel label T2, which is associated with the route to PE1.
  - In AS200, ASBR-PE2 uses LDP to assign P2 an outer tunnel label T3, which is associated with the route to ASBR-PE2.
  - In AS200, P2 uses LDP to assign PE2 an outer tunnel label T4, which is associated with the route to ASBR-PE2.
  - ASBR1 advertises a labeled IPv4 route destined for PE1 to ASBR2 through an EBGP session. The next hop is ASBR1, and the label is a BGP label with the value being B1.
  - ASBR2 advertises a labeled IPv4 route destined for PE1 to PE2 through a BGP session. The next hop is ASBR2, and the label is a BGP label with the value being B2.
    - Note: Assume that tunnel labels (or public network labels) have been allocated to the routes to PE2 and ASBR1, and the labeled routes to PE2 have been advertised.
  - PE1 and PE2 establish an MP-EBGP session with each other.
  - CE1 advertises the route destined for Client 1 to PE1, and the next hop of the route

is CE1's interface address.

- PE1 encapsulates the received IPv4 route to Client 1 into a VPNv4 route, changes the Next_Hop to PE1 in the related MP-EBGP message, allocates the VPN label V1 to the route, and then advertises the route to PE2.
- PE2 restores the received VPNv4 route to the IPv4 route to Client 1 and advertises it to CE2, with the next hop being PE2.

Option C Solution 1— Control Plane (With RRs)

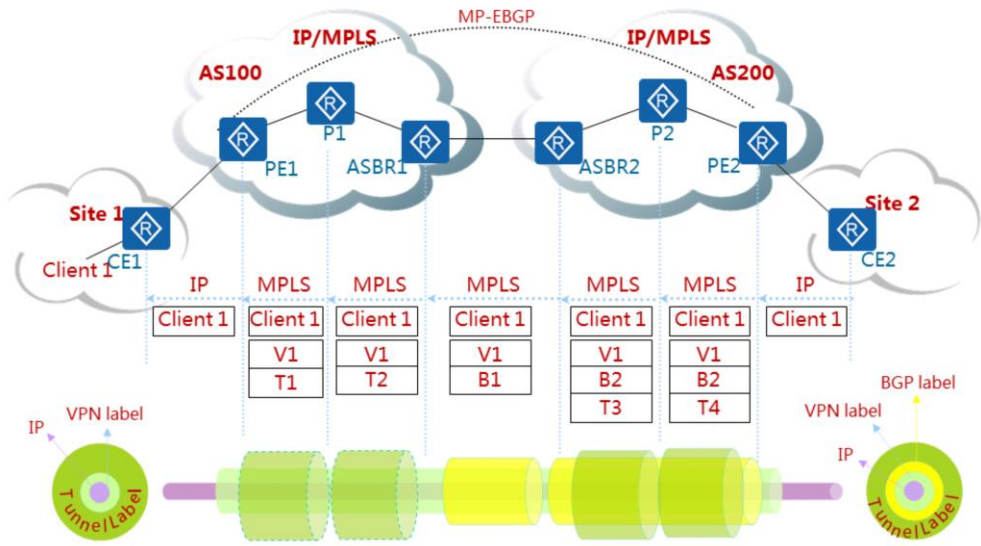


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- VPNv4 peer relationships:
  - A PE establishes a VPNv4 peer relationship only with the RR in the same AS. The local RR establishes a VPNv4 peer relationship with the peer RR to transmit inter-AS VPN routes.
- The ASBR, P and PE in one AS establish BGP unicast IPv4 peer relationships with the RR in the same AS.
  - The local ASBR learns the peer RR's loopback route from the peer ASBR through an IPv4 peer relationship and advertises the loopback route to the local RR so that the local RR can establish a VPNv4 peer relationship with the peer RR.
  - The local ASBR learns the loopback routes of the peer RR and PE from the peer ASBR through IPv4 peer relationships and advertises them to the local RR. The local RR then reflects the loopback routes to the local PE so that the PEs in different ASs can establish a BGP LSP.
- The RRs reflect IPv4 routes and transmit VPNv4 routes on the control plane, but do not forward traffic on the forwarding plane.

Option C Solution 1— Forwarding Plane



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

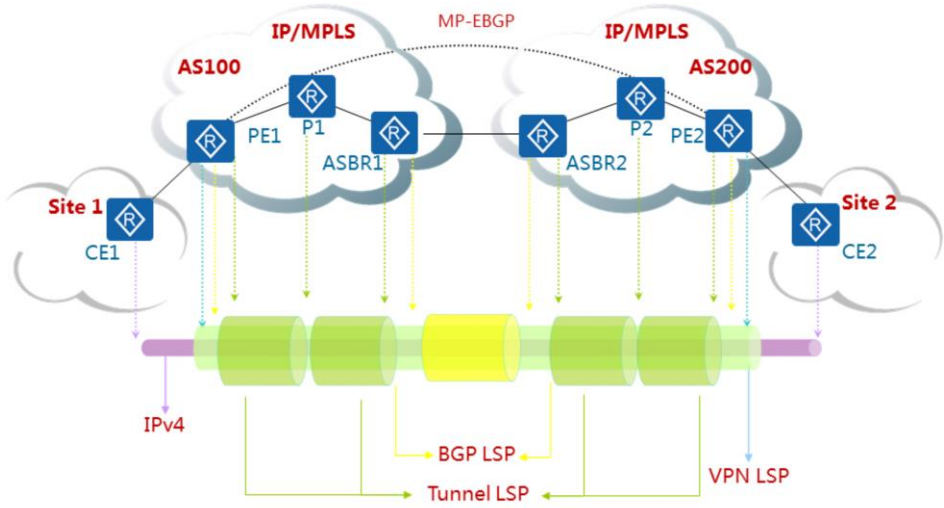
- The packet transmission process from CE2 to CE1 is used to illustrate the work flow on the forwarding plane.
  - CE2 sends an IP packet destined for Client 1 to PE2.
  - PE2 encapsulates the received IP packet with the VPN label V1 first. Because the next hop PE1 of the packet is not a directly connected peer, PE2 searches the routing table, finds the labeled BGP route to PE1, and then adds the BGP label B2 as the intermediate label to the packet. Because the next hop ASBR2 of the route to PE1 is not a directly connected peer either, PE2 searches the routing table and finds the label T4 that is associated with the route to ASBR2. As a result, PE2 adds the outer label T4 to the packet.
  - P2 swaps the outer label T4 for T3 and forwards the IP packet to ASBR-PE2.
  - ASBR2 removes the outer label from the received packet, swaps the BGP label B2 for B1, and forwards the packet to ASBR1.
  - Upon receipt, ASBR1 finds the self-assigned label B1, removes it, and searches the routing table. ASBR1 finds the label T2 associated with the route to PE1, adds the label T2 to top of stack, and then forwards the packet to P1.
  - P1 swaps the outer label T2 for T1 and forwards the packet to PE1.
  - PE1 removes both labels from the received packet and forwards the unlabeled IP packet to CE1.

Inter-AS MPLS BGP VPN Option C — Solution 2

- Solution 2 is similar to solution 1.
- Three labels (VPN label, BGP LSP label, and tunnel LSP label) are used in solution 1 to carry traffic, whereas only two labels (VPN label and tunnel label) are used in solution 2.
- Solution 1: A policy needs to be configured on the local ASBR to generate a new label. After the local ASBR receives a labeled BGP route from the peer ASBR, the local ASBR allocates the new label to the BGP route and advertises the route to the PE or RR in the same AS so that a complete BGP LSP can be established. Solution 2: MPLS needs to be configured on an ASBR to allocate labels to labeled BGP routes. On the PE in the same AS as the ASBR, you can see the LDP LSP to the peer PE instead of a BGP LSP.
- RR deployment is supported for both solution 1 and solution 2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Option C Solution 2— Topology

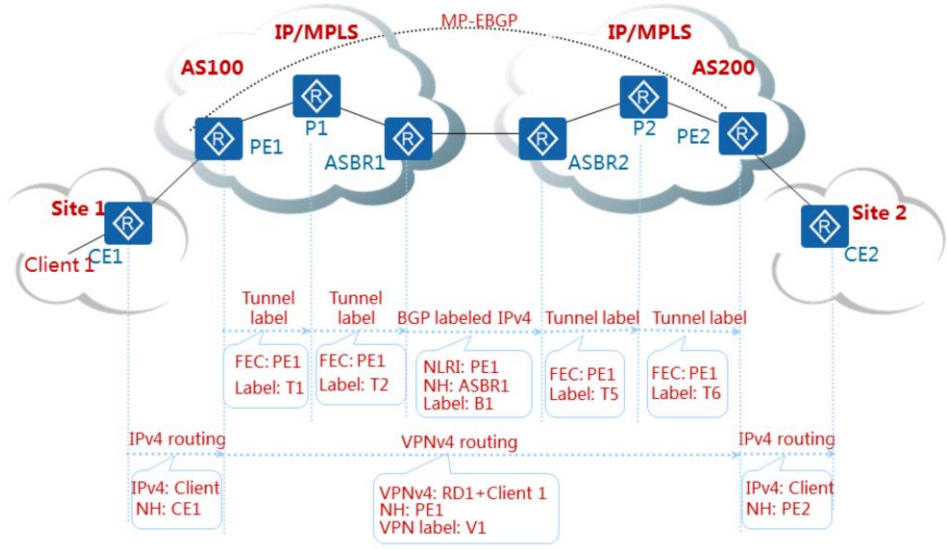


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In this solution, ASBRs do not maintain or advertise VPNv4 routes. An ASBR only needs to maintain all the labeled routes to a PE and use EBGP to advertise these labeled routes to the peer ASBR.
- After receiving a labeled BGP route, MPLS LDP on the peer ASBR triggers the generation of a label for the labeled BGP route and transmits the label to the LDP peer in the AS. Therefore, on the local PE, you can see the LDP LSP to the peer PE.
- To further expand performance, a multi-hop MP-EBGP session can be established between VPN RRs in different ASs. The PE in the local AS needs to establish an MP-IBGP peer relationship only with the RR in the same AS. The VPN RRs advertise VPNv4 routes without changing the Next_Hop attribute of the routes, so that the peer PE recurses the routes to the correct tunnel during traffic forwarding.

Option C Solution 2— Control Plane (Without RRs)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

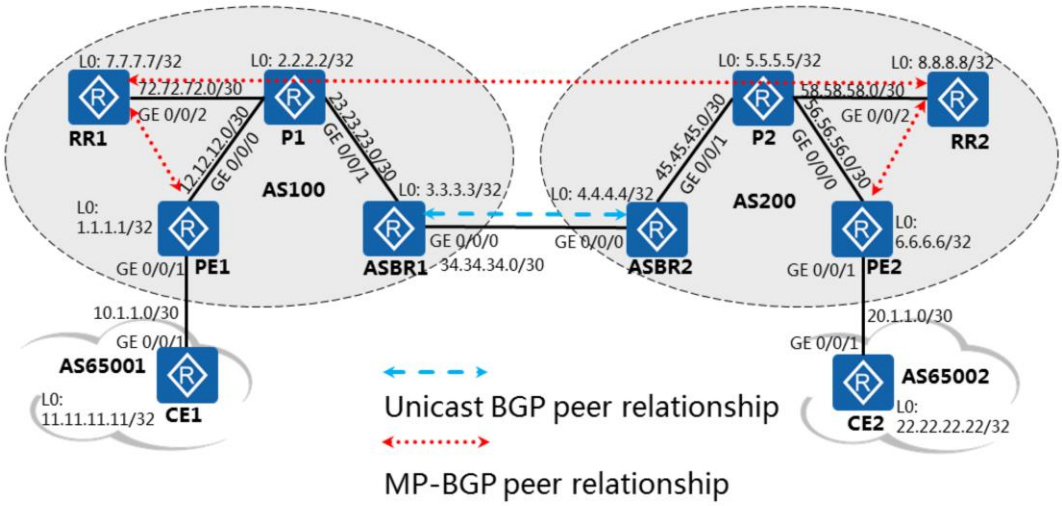
HUAWEI

- Let's take route advertisement in one direction on the control plane as an example. Suppose there is a host named Client 1 on Site 1, and the P router in each AS does not have routes to the peer PE in the other AS.
  - In AS100, PE1 uses LDP to assign P1 an outer tunnel label T1, which is associated with the route to PE1.
  - In AS100, P1 uses LDP to assign ASBR1 an outer tunnel label T2, which is associated with the route to PE1.
  - In AS200, ASBR2 uses LDP to assign P2 an outer tunnel label T3, which is associated with the route to ASBR2.
  - In AS200, P2 uses LDP to assign PE2 an outer tunnel label T4, which is associated with the route to ASBR2.
  - ASBR1 advertises a labeled IPv4 route destined for PE1 to ASBR2 through an EBGP session. The next hop is ASBR1, and the label is a BGP label with the value being B1.
  - ASBR2 sets up an LSP for the labeled BGP route and assigns an LDP label T5 to P2. P2 then assigns an LDP label T6 to PE2.
  - PE1 and PE2 establish an MP-EBGP session with each other.
  - CE1 advertises the route destined for Client 1 to PE1, and the next hop of the route is CE1's interface address.
  - PE1 encapsulates the received IPv4 route to Client 1 into a VPNv4 route, changes the Next_Hop to PE1 in the related MP-EBGP message, allocates the VPN label V1

to the route, and then advertises the route to PE2.

- PE2 restores the received VPNv4 route to the IPv4 route to Client 1 and advertises it to CE2, with the next hop being PE2.

Option C Solution 2— Control Plane (With RRs)

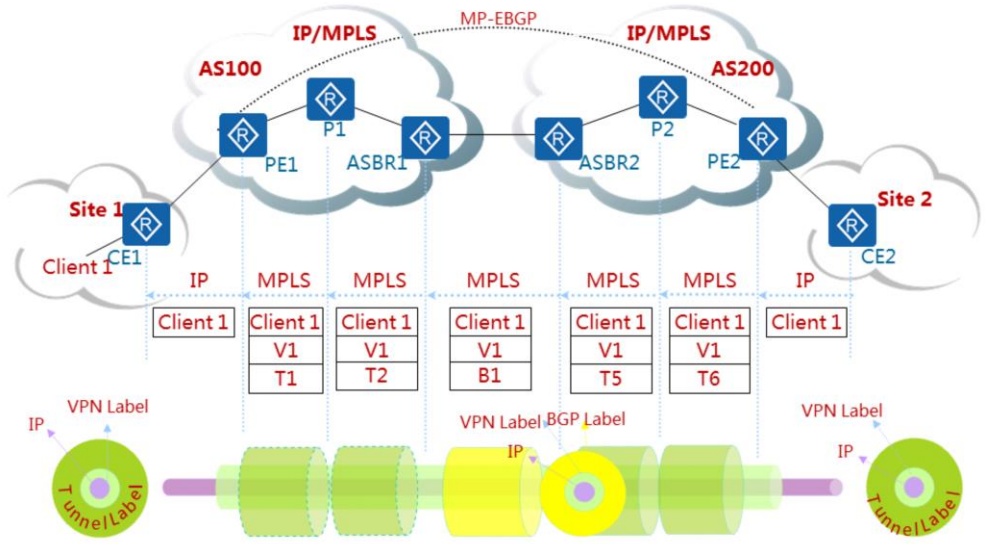


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- VPNv4 peer relationships:
  - A PE establishes a VPNv4 peer relationship only with the RR in the same AS. The local RR establishes a VPNv4 peer relationship with the peer RR to transmit inter-AS VPN routes.
- The RRs only transmit VPNv4 routes on the control plane and do not forward traffic on the forwarding plane.

Option C Solution 2— Forwarding Plane



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The packet transmission process from CE2 to CE1 is used to illustrate the work flow on the forwarding plane.
  - CE2 sends an IP packet destined for Client 1 to PE2.
  - PE2 encapsulates the received IP packet with the VPN label V1 first. Because the next hop PE1 of the packet is not a directly connected peer, PE2 searches the routing table, finds the label T6 associated with the route to PE1, and adds the label T6 to the packet.
  - P2 swaps the outer label T6 for T5 and forwards the packet to ASBR2.
  - ASBR2 removes the outer label from the received packet, swaps the label T5 for B1, and forwards the packet to ASBR1.
  - Upon receipt, ASBR1 finds the self-assigned label B1, removes it, and searches the routing table. ASBR1 finds the label T2 associated with the route to PE1, adds the label T2 to top of stack, and then forwards the packet to P1.
  - P1 swaps the outer label T2 for T1 and forwards the packet to PE1.
  - PE1 removes both labels from the received packet and forwards the unlabeled IP packet to CE1.

Features of Inter-AS MPLS BGP VPN Option C

- VPN routes are directly exchanged between the ingress PE and the egress PE. The routes do not need to be stored or forwarded by intermediate devices.
- Only PEs exchange VPN routing information, whereas Ps and ASBRs only forward packets. The intermediate devices only need to support MPLS forwarding instead of MPLS VPN services. Therefore, ASBRs are no longer bottlenecks. Option C, therefore, is suitable for the VPN that spans multiple ASs.
- MPLS VPN load balancing is easy to carry out in Option C.
- The disadvantage of this mode is that it costs too much to manage an end-to-end connection between PEs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

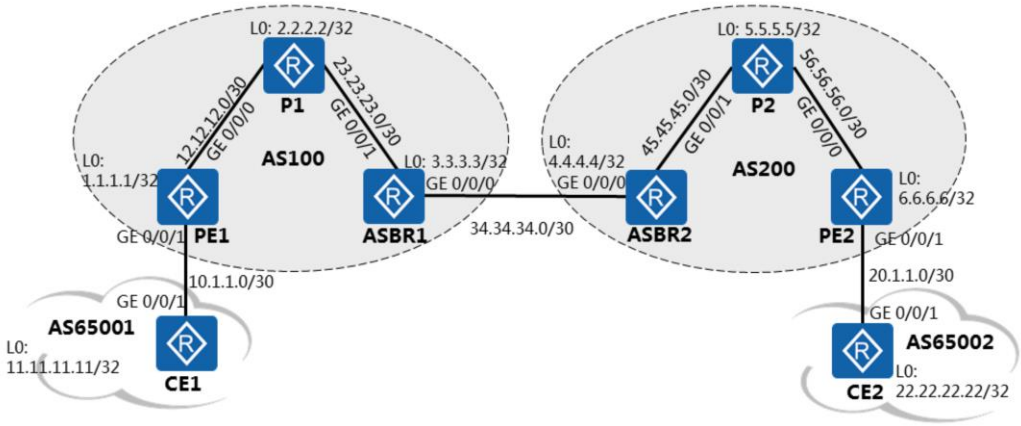
Contents

- Background
- Fundamentals
- **Configurations**
  - Option A Solution Configuration
  - Option B Solution Configuration
  - Option C Solution 1 Configuration
  - Option C Solution 2 Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Topology in Inter-AS MPLS BGP VPN Option A

- Configuration task: Use the Option A mode to implement mutual access between CE1 and CE2. Note: OSPF is used in the ASs in this example.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- On the network shown in the figure, AS100 and AS200 are used for the ISP, whereas the other two ASs are used for the customer. PE1 and ASBR1 belong to AS100, and PE2 and ASBR2 belong to AS200. CE1 and CE2 belong to the same VPN. CE1 is connected to PE1 in AS100, and CE2 is connected to PE2 in AS200.
- The IP addresses planned on each router are shown in this topology.

Configuration Procedure in Inter-AS MPLS BGP VPN Option A

Step	Description	Reference Command
1	Assign an IP address and mask to each interface.	See "Basic IPv4 Configuration" in the related product manual.
2	Configure OSPF on each PE, P and ASBR over the public network.	See "Configuring Basic OSPF Functions" in the related product manual.
3	Configure MPLS and LDP on each PE, P and ASBR over the public network.	mpls mpls lsr-id lsr-id mpls ldp
4	Configure IBGP peer relationships between PE1 and ASBR1, and between PE2 and ASBR2. Configure MP-IBGP peer relationships between PE1 and ASBR1, and between PE2 and ASBR2.	See "Configuring Basic BGP Functions" in the related product manual. ipv4-family vpnv4 peer { group-name | ipv4-address | ipv6-address } enable
5	Configure a VPN instance with an RD and RTs on each PE and ASBR.	ip vpn-instance vpn-instance-name route-distinguisher route-distinguisher vpn-target vpn-target [both |export-extcommunity | importextcommunity]
6	Bind the AC interface to the VPN instance on each PE and ASBR.	ip binding vpn-instance vpn-instance-name
7	Configure an EBGP peer relationship in the VPN instance on each PE and ASBR.	See "Configuring Basic BGP Functions" in the related product manual.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Example for Configuring Inter-AS MPLS BGP VPN Option A (1)

1. Set up a topology and configure IP addresses.
   - Set up the test topology based on the topology requirements and configure IP addresses.
2. Configure OSPF.
   - Configure OSPF on PE1, P1, ASBR1, PE2, P2 and ASBR2.
3. Configure MPLS and LDP on each PE, P and ASBR over the public network.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Set up a topology as shown in the previous slide and assign IP addresses to interfaces based on the topology.
  - On CE1, assign IP addresses to Loopback 0 and GE 0/0/1.
  - On PE1, assign IP addresses to Loopback 0 and GE 0/0/0.
  - On P1, assign IP addresses to Loopback 0, GE 0/0/0, and GE 0/0/1.
  - On ASBR1, assign IP addresses to Loopback 0 and GE 0/0/1.
  - On ASBR2, assign IP addresses to Loopback 0 and GE 0/0/1.
  - On P2, assign IP addresses to Loopback 0, GE 0/0/0, and GE 0/0/1.
  - On PE2, assign IP addresses to Loopback 0 and GE 0/0/0.
  - On CE2, assign IP addresses to Loopback 0 and GE 0/0/1.
- Configure OSPF on PE1, P1, ASBR1, PE2, P2 and ASBR2.
  - Enable PE1 to advertise routes to the network segments 1.1.1.1/32 and 12.12.12.0/30.
  - Enable P1 to advertise routes to the network segments 2.2.2.2/32, 12.12.12.0/30 and 23.23.23.0/30.
  - Enable ASBR1 to advertise routes to the network segments 3.3.3.3/32 and 23.23.23.0/30.
  - Enable ASBR2 to advertise routes to the network segments 4.4.4.4/32 and

45.45.45.0/30.

- Enable P2 to advertise routes to the network segments 5.5.5.5/32, 45.45.45.0/30 and 56.56.56.0/30.
- Enable PE2 to advertise routes to the network segments 6.6.6.6/32 and 56.56.56.0/30.

Example for Configuring Inter-AS MPLS BGP VPN Option A (2)

3. Configure MPLS and LDP on each PE, P and ASBR over the public network.
   - Take PE1 as an example:

```
<PE1>system-view
[PE1]mpls lsr-id 1.1.1.1
[PE1]mpls
Info: Mpls starting, please wait... OK!
[PE1-mpls]mpls ldp
[PE1-mpls-ldp]quit
[PE1]interface GigabitEthernet0/0/0
[PE1-GigabitEthernet0/0/0]mpls
[PE1-GigabitEthernet0/0/0]mpls ldp
[PE1-GigabitEthernet0/0/0]quit
```

- Repeat the preceding configuration on P1, P2, PE2, ASBR1 and ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option A (3)

- After the configurations are complete, LDP peer sessions are set up between the P and PE, and between the P and ASBR in the same AS. Run the **display mpls ldp session** command on each P. The command output shows that the session state is **Operational**.
- The following uses the command output on P1 as an example:

```
[P1]display mpls ldp session
 LDP Session(s) in Public Network
 Codes: LAM(Label Advertisement Mode), SsnAge Unit(DDDD:HH:MM)
 A '*' before a session means the session is being deleted.
 ------------------------------------------------------------------------------
 PeerID          Status      LAM  SsnRole  SsnAge       KASent/Rcv
 ------------------------------------------------------------------------------
 1.1.1.1:0       Operational DU   Active   0000:02:21   568/568
 3.3.3.3:0       Operational DU   Passive  0000:02:21   567/567
 ------------------------------------------------------------------------------
 TOTAL: 2 session(s) Found.
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option A (4)

4. Configure IBGP and MP-IBGP peer relationships between PE1 and ASBR1, and between PE2 and ASBR2.

```
[PE1]bgp 100
[PE1-bgp]peer 3.3.3.3 as-number 100
            //Configures an IBGP peer relationship between PE1 and ASBR1.
[PE1-bgp]peer 3.3.3.3 connect-interface loopback 0
            //Sets the interface used to establish the peer relationship to
Loopback 0.
[PE1-bgp]ipv4-family vpnv4
            //Displays the BGP VPNv4 view.
[PE1-bgp-af-vpnv4]peer 3.3.3.3 enable
            //Enables the MP-IBGP peer relationship between PE1 and ASBR1.
[PE1-bgp-af-vpnv4]quit
```

- Repeat the preceding configuration on PE2, ASBR1 and ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option A (5)

- Verify the configuration:

```
[PE1] display bgp vpnv4 all peer
BGP local router ID : 1.1.1.1
Local AS number : 100 Total number of peers : 1
Peers in established state : 1
Peer        V    AS  MsgRcvd  MsgSent  OutQ  Up/Down       State PrefRcv  3.3.3.3        4  100
3       7    0 00:01:36 Established      0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option A (6)

5. Configure a VPN instance on each PE and ASBR.

```
[PE1]ip vpn-instance huawei
[PE1-vpn-instance-huawei]route-distinguisher 100:1
[PE1-vpn-instance-huawei]vpn-target 100:1 both
[PE1-vpn-instance-huawei]quit
```

```
[ASBR1]ip vpn-instance huawei
[ASBR1-vpn-instance-huawei]route-distinguisher 100:1
[ASBR1-vpn-instance-huawei]vpn-target 100:1 both
[ASBR1-vpn-instance-huawei]quit
```

- Repeat the configuration of PE1 on PE2, and repeat the configuration of ASBR1 on ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Example for Configuring Inter-AS MPLS BGP VPN Option A (7)

6. Bind the AC interface to the VPN instance on each PE and ASBR.

```
[PE1]interface GigabitEthernet0/0/1
[PE1-GigabitEthernet0/0/1]ip binding vpn-instance huawei
[PE1-GigabitEthernet0/0/1]ip address 10.1.1.2 30
[PE1-GigabitEthernet0/0/1]quit
```

```
[ASBR1] interface GigabitEthernet0/0/0
[ASBR1-GigabitEthernet0/0/0]ip binding vpn-instance huawei
[ASBR1-GigabitEthernet0/0/0]ip address 34.34.34.1 30
[ASBR1-GigabitEthernet0/0/0]quit
//Binds the VPN instance to the interface that connects to ASBR2 (ASBR1 considers ASBR2 as its CE).
```

- Repeat the configuration of PE1 on PE2, and repeat the configuration of ASBR1 on ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Example for Configuring Inter-AS MPLS BGP VPN Option A (8)

7. Configure an EBGP peer relationship in the VPN instance on each PE and ASBR.

```
[PE1]bgp 100
[PE1-bgp]ipv4-family vpn-instance huawei
[PE1-bgp-huawei]peer 10.1.1.1 as-number 65001
//Configures an EBGP peer relationship between PE1 and CE1.
```

```
[CE1]bgp 65001
[CE1-bgp]peer 10.1.1.2 as-number 100
[CE1-bgp]network 11.11.11.11 32
```

```
[ASBR1]bgp 100
[ASBR1-bgp]ipv4-family vpn-instance huawei
[ASBR1-bgp-huawei]peer 34.34.34.2 as-number 200
//Configures an EBGP peer relationship between ASBR1 and ASBR2.
```

- Repeat the configuration of PE1 on PE2, repeat the configuration of CE1 on CE2, and repeat the configuration of ASBR1 on ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Example for Configuring Inter-AS MPLS BGP VPN Option A (9)

- Verify the configuration:

```
[PE1]display bgp vpnv4 vpn-instance huawei peer
BGP local router ID : 1.1.1.1
Local AS number : 100
Total number of peers : 1
Peers in established state : 1
Peer     V  AS  MsgRcvd  MsgSent  OutQ  Up/Down     State      PrefRcv  10.1.1.1   4 65001
10       10     0 00:07:10 Established       0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option A (1)

- After the preceding configurations are complete, CE1 and CE2 can learn the routes to each other's loopback interfaces and ping each other successfully.
  - Take the command output on CE1 as an example:

```
[CE1]dis ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 6        Routes : 6

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

10.1.1.0/30         Direct  0    0           D   10.1.1.1        GigabitEthernet0/0/1
10.1.1.1/32         Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
11.11.11.11/32      Direct  0    0           D   127.0.0.1       LoopBack0
22.22.22.22/32      EBGP    255  0           D   10.1.1.2        GigabitEthernet0/0/1
127.0.0.0/8         Direct  0    0           D   127.0.0.1       InLoopBack0
127.0.0.1/32        Direct  0    0           D   127.0.0.1       InLoopBack0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option A (2)

- Run the **display ip routing-table vpn-instance** command on an ASBR to check the VPN routing table.

```
[ASBR1]dis ip routing-table vpn-instance huawei
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: huawei
         Destinations : 4        Routes : 4

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

11.11.11.11/32      IBGP    255  0         RD    1.1.1.1         GigabitEthernet0/0/1
22.22.22.22/32      EBGP    255  0         D     34.34.34.2      GigabitEthernet0/0/0
34.34.34.0/30       Direct  0    0         D     34.34.34.1      GigabitEthernet0/0/0
34.34.34.1/32       Direct  0    0         D     127.0.0.1       GigabitEthernet0/0/0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option A (3)

- Run the **display bgp vpnv4 all routing-table** command on an ASBR to check for VPNv4 routes.

```
[ASBR1]dis bgp vpnv4 all routing-table

BGP Local router ID is 3.3.3.3
Status codes: * - valid, > - best, d - damped,
              h - history,  i - internal, s - suppressed, S - Stale
              Origin : i - IGP, e - EGP, ? – incomplete

Total number of routes from all PE: 2
Route Distinguisher: 100:1
     Network            NextHop    MED    LocPrf    PrefVal    Path/Ogn

*>i  11.11.11.11/32     1.1.1.1    0      100       0          65001i
*>   22.22.22.22/32     34.34.34.2                  0          200 65002i

VPN-Instance huawei, Router ID 3.3.3.3:
Total Number of Routes: 2
     Network            NextHop    MED    LocPrf    PrefVal    Path/Ogn
*>i  11.11.11.11/32     1.1.1.1    0      100       0          65001i
*>   22.22.22.22/32     34.34.34.2                  0          200 65002i
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

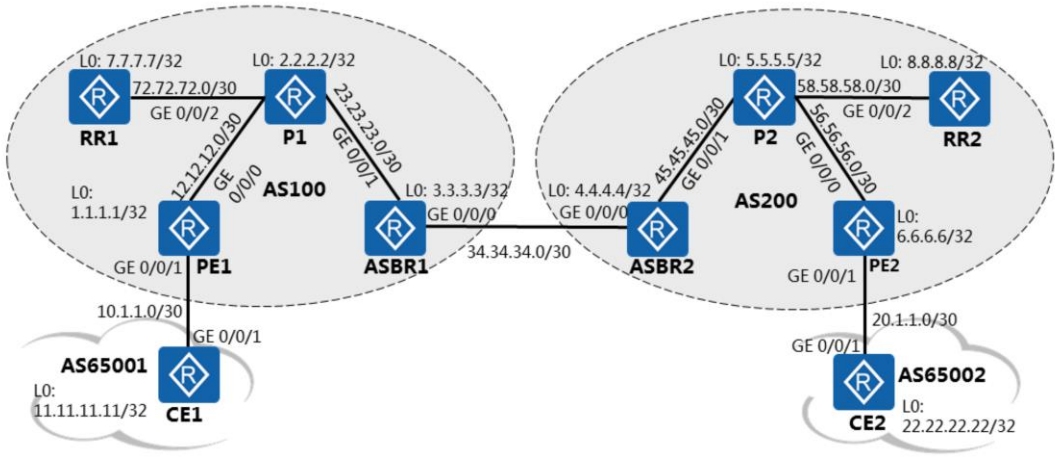


Contents

- Background
- Fundamentals
- **Configurations**
  - Option A Solution Configuration
  - Option B Solution Configuration
  - Option C Solution 1 Configuration
  - Option C Solution 2 Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Topology in Inter-AS MPLS BGP VPN Option B



- Configuration task: Use the Option B mode to implement mutual access between CE1 and CE2. Note: OSPF is used in the ASs in this example.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In this example, RR1 and RR2 serve AS100 and AS200, respectively. In each AS, the PE and ASBR establish BGP peer relationships with the RR, and the RR reflects VPN routes.

Configuration Procedure in Inter-AS MPLS BGP VPN Option B (1)

Step	Description	Reference Command
1	Assign an IP address and mask to each interface.	See "Basic IPv4 Configuration" in the related product manual.
2	Configure OSPF on each PE, P, ASBR and RR over the public network.	See "Configuring Basic OSPF Functions" in the related product manual.
3	Configure MPLS and LDP on each PE, P and ASBR over the public network.	mpls mpls lsr-id lsr-id mpls ldp
4	In each AS, configure IBGP peer relationships between the PE and RR, and between ASBR and RR. Enable the MP-IBGP peer relationship in the VPNv4 view on each PE, RR and ASBR.	See "Configuring Basic BGP Functions" in the related product manual. ipv4-family vpnv4 peer { group-name | ipv4-address | ipv6-address } enable
5	Enable RR1 and RR2 to reflect VPN routes.	ipv4-family vpnv4 peer { group-name | ipv4-address | ipv6-address } reflect-client
6	Configure a VPN instance with an RD and RTs on each PE.	ip vpn-instance vpn-instance-name route-distinguisher route-distinguisher vpn-target vpn-target [both |export-extcommunity | importextcommunity]

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuration Procedure in Inter-AS MPLS BGP VPN Option B (2)

Step	Description	Reference Command
7	Bind the AC interface to the VPN instance on each PE.	ip binding vpn-instance vpn-instance-name
8	In the VPN instance on each PE, configure an EBGP peer relationship with the connected CE.	See "Configuring Basic BGP Functions" in the related product manual.
9	Enable MPLS on the interfaces that connect ASBR1 and ASBR2.	mpls
10	Configure an MP-EBGP peer relationship between ASBR1 and ASBR2 in the VPNv4 view.	ipv4-family vpnv4 peer { group-name | ipv4-address | ipv6-address } enable undo policy vpn-target
Remarks	1. Enable MPLS instead of LDP on ASBR1 and ASBR2. 2. In Option B mode, the VPN targets of the VPN instances on the PEs in different ASs must match each other. 3. There is no need to configure VPN instances or bind them to the interfaces on ASBR1 and ASBR2. 4. The following special configuration is required on ASBR1, ASBR2, RR1 and RR2: Run the **undo policy vpn-target** command in the VPNv4 address family view.	

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Example for Configuring Inter-AS MPLS BGP VPN Option B (1)

- For the configurations in steps 1 to 4, see the related configurations in Option A.

5. Enable each RR to reflect VPN routes and disable each RR from filtering received VPN routes based on VPN targets.

```
[RR1-bgp]ipv4-family vpnv4
[RR1-bgp-af-vpnv4]undo policy vpn-target
[RR1-bgp-af-vpnv4]peer 1.1.1.1 reflect-client
[RR1-bgp-af-vpnv4]peer 3.3.3.3 reflect-client
```

6. Configure a VPN instance with an RD (100:1) and RTs (100:1) on each PE.

```
[PE1]ip vpn-instance huawei
[PE1-vpn-instance-huawei]route-distinguisher 100:1
[PE1-vpn-instance-huawei]vpn-target 100:1 both
[PE1-vpn-instance-huawei]quit
```

- Repeat the configuration of RR1 on RR2, and repeat the configuration of PE1 on PE2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option B (2)

7. Bind the AC interface to the VPN instance on each PE.

```
[PE1]interface GigabitEthernet0/0/1
[PE1-GigabitEthernet0/0/1]ip binding vpn-instance huawei
[PE1-GigabitEthernet0/0/1]ip address 10.1.1.2 30
[PE1-GigabitEthernet0/0/1]quit
```

- Repeat the configuration of PE1 on PE2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option B (3)

8. In the VPN instance on each PE, configure an EBGP peer relationship with the connected CE.

```
[PE1]bgp 100
[PE1-bgp]ipv4-family vpn-instance huawei
[PE1-bgp-huawei]peer 10.1.1.1 as-number 65001
//Configures an EBGP peer relationship between PE1 and CE1.
```

```
[CE1]bgp 65001
[CE1-bgp]peer 10.1.1.2 as-number 100
[CE1-bgp]network 11.11.11.11 32
```

- Repeat the configuration of PE1 on PE2, and repeat the configuration of CE1 on CE2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option B (4)

9. Enable MPLS on the interfaces that connect ASBR1 and ASBR2.

```
[ASBR1]interface GigabitEthernet0/0/0
[ASBR1-GigabitEthernet0/0/0]ip address 34.34.34.1 24
[ASBR1-GigabitEthernet0/0/0]mpls
//Enables MPLS on the ASBR1 interface that connects to ASBR2.
```

- Repeat the configuration of ASBR1 on ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Example for Configuring Inter-AS MPLS BGP VPN Option B (5)

10. Configure an MP-EBGP peer relationship between the ASBRs in the VPNv4 view.

```
[ASBR1]bgp 100
[ASBR1-bgp]peer 34.34.34.2 as-number 200
[ASBR1-bgp]ipv4-family vpnv4
[ASBR1-bgp-af-vpnv4]peer 34.34.34.2 enable
//Establishes an MP-EBGP peer relationship with ASBR2.
[ASBR1-bgp-af-vpnv4]undo policy vpn-target
//Disables the ASBR from filtering received VPNv4 routes based on VPN targets.
```

- Repeat the configuration of ASBR1 on ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option B (1)

- After the preceding configurations are complete, CE1 and CE2 can learn the routes to each other's loopback interfaces and ping each other successfully.
- Take the command output on CE1 as an example:

```
[CE1]dis ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 6        Routes : 6

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

10.1.1.0/30         Direct  0    0           D   10.1.1.1        GigabitEthernet0/0/1
10.1.1.1/32         Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
11.11.11.11/32      Direct  0    0           D   127.0.0.1       LoopBack0
22.22.22.22/32      EBGP    255  0           D   10.1.1.2        GigabitEthernet0/0/1
127.0.0.0/8         Direct  0    0           D   127.0.0.1       InLoopBack0
127.0.0.1/32        Direct  0    0           D   127.0.0.1       InLoopBack0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option B (2)

- Run the **display bgp vpnv4 all routing-table** command on an ASBR to check for VPNv4 routes.

```
[ASBR1]dis bgp vpnv4 all routing-table
 BGP Local router ID is 23.23.23.2
 Status codes: * - valid, > - best, d - damped,
               h - history,  i - internal, s - suppressed, S - Stale
               Origin : i - IGP, e - EGP, ? - incomplete
 Total number of routes from all PE: 2
 Route Distinguisher: 100:1

      Network            NextHop        MED        LocPrf    PrefVal    Path/Ogn
 *>i  11.11.11.11/32     1.1.1.1        0          100       0          65001i
 *>   22.22.22.22/32     34.34.34.2                           0          200  65002i
```

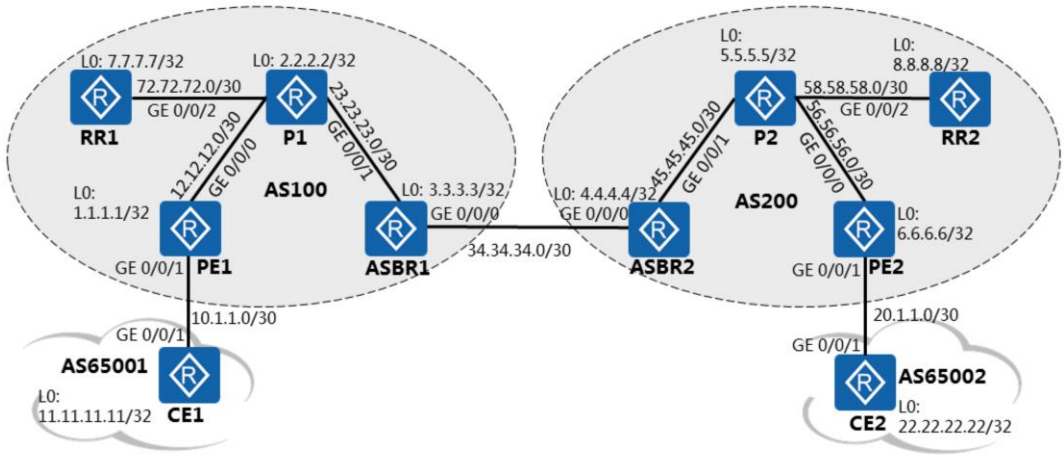
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

- Background
- Fundamentals
- **Configurations**
  - Option A Solution Configuration
  - Option B Solution Configuration
  - Option C Solution 1 Configuration
  - Option C Solution 2 Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Topology in Inter-AS MPLS BGP VPN Option C (Solution 1)



- Configuration task: Use Option C solution 1 to implement mutual access between CE1 and CE2. Note: OSPF is used in the ASs in this example.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- On the network shown in the figure, AS100 and AS200 are used for the ISP, whereas the other two ASs are used for the customer. PE1, P1, RR1 and ASBR1 belong to AS100. PE2, P2, RR2, and ASBR2 belong to AS200. CE1 and CE2 belong to the same VPN. CE1 is connected to PE1 in AS100 and CE2 is connected to PE2 in AS200.
- The IP addresses planned on each router are shown in this topology.
- Option C solution 1 is used in this example. PE1 and PE2 can establish an MP-EBGP peer relationship with each other to transmit inter-AS VPN routes, avoiding the need to use RRs. Alternatively, RR1 and RR2 establish an MP-EBGP peer relationship to transmit inter-AS VPN routes. In this case, MP-IBGP peer relationships are established between PE1 and RR1, and between PE2 and RR2. In this example, RRs are used to implement Option C solution 1.

Configuration Procedure in Inter-AS MPLS BGP VPN Option C Solution 1 (1)

Step	Description	Reference Command
1	Assign an IP address and mask to each interface.	See "Basic IPv4 Configuration" in the related product manual.
2	Configure OSPF on each PE, P, RR and ASBR over the public network.	See "Configuring Basic OSPF Functions" in the related product manual.
3	Configure MPLS and LDP on each PE, P, RR and ASBR over the public network.	mpls mpls lsr-id lsr-id mpls ldp
4	In each AS, configure an IBGP peer relationship between each of the PE, P and ASBR and the RR, and configure the PE, P and ASBR as RR clients.	See "Configuring Basic BGP Functions" in the related product manual.
5	Configure an EBGP peer relationship between ASBR1 and ASBR2.	See "Configuring Basic BGP Functions" in the related product manual.
6	Configure a VPN instance with an RD (100:1) and RTs (100:1) on each PE.	ip vpn-instance vpn-instance-name route-distinguisher route-distinguisher vpn-target vpn-target [both |export-extcommunity | importextcommunity]
7	Bind the AC interface to the VPN instance on each PE.	ip binding vpn-instance vpn-instance-name
8	In the VPN instance on each PE, configure an EBGP peer relationship with the connected CE.	See BGP Configuration.
9	Configure a multi-hop MP-EBGP peer relationship between RR1 and RR2 in the VPNv4 view. Configure an MP-IBGP peer relationship between PE1 and RR1 in the VPNv4 view. Configure an MP-IBGP peer relationship between PE2 and RR2 in the VPNv4 view. Disable each RR from changing the Next_Hop of the routes being advertised.	peer { group-name | ipv4-address | ipv6-address } ebgp-max-hop [hop-count] ipv4-family vpnv4 peer { group-name | ipv4-address | ipv6-address } enable peer next-hop-invariable

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Configuration Procedure in Inter-AS MPLS BGP VPN Option C Solution 1 (2)

Step	Description	Reference Command
10	Enable the capability of exchanging labeled IPv4 routes between the PE and RR in each AS, between the RR and ASBR in each AS, and between the ASBRs in different ASs.	peer { group-name | ipv4-address } label-route-capability
11	Create a route-policy and apply it to the routes that the local ASBR will advertise to the remote ASBR and the routes that the local ASBR will advertise to the local RR.	route-policy route-policy-name { permit | deny } node node peer { group-name | ipv4-address | ipv6-address } route-policy route-policy-name { import | export }
Remarks	1. Enable MPLS on ASBR1 and ASBR2. LDP is optional. 2. There is no need to configure VPN instances or bind them to the interfaces on ASBR1 and ASBR2. 3. Configure the capability of exchanging labeled IPv4 routes: Between PE1 and RR1 Between RR1 and ASBR1 Between PE2 and RR2 Between RR2 and ASBR2 Between ASBR1 and ASBR2 4. Ensure that the local RR does not change the Next_Hop of the VPN routes when advertising them to the remote RR. Also ensure that the remote RR does not change the Next_Hop of the received VPN routes when advertising them to the PE in the same AS. 5. Configure a route-policy on each of ASBR1 and ASBR2 to ensure that: The local ASBR can assign MPLS labels to the routes received from the RR in the same AS when advertising them to the ASBR in the peer AS. An ASBR can assign new MPLS labels to the labeled IPv4 routes when advertising them to the RR in the same AS.	

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 1 (1)

- For the configurations in steps 1 to 3, see the related configurations in Option A.

4. In each AS, configure an IBGP peer relationship between each of the PE, P and ASBR and the RR, and configure the PE, P and ASBR as RR clients.

```
[RR1-bgp]peer 1.1.1.1 as-number 100
[RR1-bgp]peer 1.1.1.1 connect-interface Loopback0
[RR1-bgp]peer 1.1.1.1 reflect-client
[RR1-bgp]peer 2.2.2.2 as-number 100
[RR1-bgp]peer 2.2.2.2 connect-interface Loopback0
[RR1-bgp]peer 2.2.2.2 reflect-client
[RR1-bgp]peer 3.3.3.3 as-number 100
[RR1-bgp]peer 3.3.3.3 connect-interface Loopback0
[RR1-bgp]peer 3.3.3.3 reflect-client
```

5. Configure an EBGP peer relationship between ASBR1 and ASBR2.

```
[ASBR1-bgp]peer 34.34.34.2 as-number 200
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In step 4, repeat the configuration of RR1 on RR2. For the configuration of the PE, P, and ASBR, see "Configuring Basic BGP Functions" in the related product manual.
- In step 5, repeat the configuration of ASBR1 on ASBR2.
- Establish a unicast EBGP peer relationship between the ASBRs so that the local ASBR can advertise routes to the loopback interface addresses of the local RR and PE to the peer ASBR.
- When advertising routes to the loopback interface addresses of RR1 and PE1 to ASBR2, the local ASBR allocates MPLS labels to these routes. When advertising the routes to the loopback interface addresses of RR1 and PE1 to RR2, ASBR2 allocates new MPLS labels to these routes.
- After establishing IBGP peer relationships between the ASBR and RR, and between the PE and RR in the same AS, enable the IBGP peers to exchange labels.
- In the same AS, establish an IPv4 peer relationship between each of the ASBR, P and PE and the RR.
  - The local ASBR learns the peer RR's loopback route from the peer ASBR through an IPv4 peer relationship and advertises the loopback route to the local RR so that the local RR can establish a VPNv4 peer relationship with the peer RR.
  - The local ASBR learns the loopback routes of the peer RR and PE from the peer ASBR through IPv4 peer relationships and advertises them to the local RR. The local RR then reflects the loopback routes to the local P router for recursive lookup of BGP routes.

- The local ASBR learns the loopback routes of the peer RR and PE from the peer ASBR through IPv4 peer relationships and advertises them to the local RR. The local RR then reflects the loopback routes to the local PE so that the PEs in different ASs can establish a BGP LSP.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 1 (2)

6. Configure a VPN instance with an RD (100:1) and RTs (100:1) on each PE.

```
[PE1]ip vpn-instance huawei
[PE1-vpn-instance-huawei]route-distinguisher 100:1
[PE1-vpn-instance-huawei-af-ipv4]vpn-target 100:1 both
```

7. Bind the AC interface to the VPN instance on each PE.

```
[PE1]interface GigabitEthernet0/0/1
[PE1-GigabitEthernet0/0/1]ip binding vpn-instance huawei
[PE1-GigabitEthernet0/0/1]ip address 10.1.1.2 30
[PE1-GigabitEthernet0/0/1]quit
```

- Repeat the configuration of PE1 on PE2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 1 (3)

8. In the VPN instance on each PE, configure an EBGP peer relationship with the connected CE.

```
[PE1]bgp 100
[PE1-bgp]ipv4-family vpn-instance huawei
[PE1-bgp-huawei]peer 10.1.1.1 as-number 65001
//Configures an EBGP peer relationship between PE1 and CE1.
```

```
[CE1]bgp 65001
[CE1-bgp]peer 10.1.1.2 as-number 100
[CE1-bgp]network 11.11.11.11 32
```

- Repeat the configuration of PE1 on PE2, and repeat the configuration of CE1 on CE2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 1 (4)

9. Configure an MP-EBGP peer relationship between RR1 and RR2 in the VPNv4 view, and disable the RRs from changing the Next_Hop of the VPN routes being advertised.

```
[RR1]bgp 100
[RR1-bgp]peer 8.8.8.8 as-number 200
[RR1-bgp]peer 8.8.8.8 connect-interface LoopBack 0
[RR1-bgp]peer 8.8.8.8 ebgp-max-hop 10
[RR1-bgp]ipv4-family vpnv4
[RR1-bgp-af-vpnv4]peer 8.8.8.8 enable
[RR1-bgp-af-vpnv4]peer 8.8.8.8 next-hop-invariable
[RR1-bgp-af-vpnv4]undo policy vpn-target
```

- In each AS, configure an MP-IBGP peer relationship between the PE and RR in the VPNv4 view, and disable the RR from changing the Next_Hop of the VPN routes being advertised.

```
[RR1]bgp 100
[RR1-bgp]peer 1.1.1.1 as-number 100
[RR1-bgp]peer 1.1.1.1 connect-interface LoopBack 0
[RR1-bgp]ipv4-family vpnv4
[RR1-bgp-af-vpnv4]peer 1.1.1.1 enable
[RR1-bgp-af-vpnv4]peer 1.1.1.1 next-hop-invariable
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- For the establishment of an MP-IBGP peer relationship between PE2 and RR2, see the configuration between PE1 and RR1.
- The **undo policy vpn-target** command configuration in Option C functions the same as that in Option B. They both disable RRs from filtering routes based on RTs.
- The **peer** *X.X.X.X* **next-hop-invariable** command configuration ensures that the peer PE can recurse routes to the BGP LSP destined for the local PE during traffic forwarding.
- Establish an MP-EBGP peer relationship between the RRs in the VPNv4 view, and disable the local RR from changing the Next_Hop attribute of the routes being advertised to the peer RR. That is, the next hop of a VPNv4 route learned by the peer PE is the local PE.
- Establish an MP-IBGP peer relationship between the RR and PE in the VPNv4 view, and disable the RR from changing the Next_Hop attribute of routes being advertised to the local PE. That is, the next hop of a VPNv4 route learned by the local PE is the peer PE.
- A PE establishes a VPNv4 peer relationship only with the RR in the same AS. The local RR establishes a VPNv4 peer relationship with the peer RR to transmit inter-AS VPN routes.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 1 (5)

10. Enable the capability of exchanging labeled IPv4 routes between the PE and RR in each AS, between the RR and ASBR in each AS, and between the ASBRs in different ASs.

```
[PE1]bgp 100
[PE1-bgp]peer 7.7.7.7 as-number 100
[PE1-bgp]peer 7.7.7.7 label-route-capability
```

```
[RR1]bgp 100
[RR1-bgp]peer 1.1.1.1 as-number 100
[RR1-bgp]peer 1.1.1.1 label-route-capability
[RR1-bgp]peer 3.3.3.3 as-number 100
[RR1-bgp]peer 3.3.3.3 label-route-capability
```

```
[ASBR1]bgp 100
[ASBR1-bgp]peer 7.7.7.7 as-number 100
[ASBR1-bgp]peer 7.7.7.7 label-route-capability
[ASBR1-bgp]peer 34.34.34.2 as-number 200
[ASBR1-bgp]peer 34.34.34.2 label-route-capability
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- For configurations on PE2, RR2 and ASBR2, see the configurations on PE1, RR1, and ASBR1, respectively.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 1 (6)

11. Create a route-policy and apply it to the routes that the local ASBR will advertise to the remote ASBR and the routes that the local ASBR will advertise to the local RR.

```
[ASBR1]interface GigabitEthernet0/0/0
[ASBR1-GigabitEthernet0/0/0]ip address 34.34.34.1 30
[ASBR1-GigabitEthernet0/0/0]mpls
[ASBR1-GigabitEthernet0/0/0]quit
[ASBR1]route-policy policy1 permit node 10
[ASBR1-route-policy]apply mpls-label //Assigns MPLS labels to the routes that hit the filtering rule.
[ASBR1-route-policy]quit
[ASBR1]route-policy policy2 permit node 10
[ASBR1-route-policy]if-match mpls-label
[ASBR1-route-policy]apply mpls-label //Assigns MPLS labels to the labeled IPv4 routes so that the original MPLS labels are replaced with new ones.
[ASBR1-route-policy]quit
//Creates two route-policies on ASBR1.
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 1 (7)

```
[ASBR1]bgp 100
[ASBR1-bgp]peer 7.7.7.7 route-policy policy2 export
//On ASBR1, applies the route-policy to the routes to be advertised to RR1. If the routes to be advertised are labeled IPv4 routes, ASBR1 will assign new MPLS labels to them.

[ASBR1-bgp]peer 34.34.34.2 as-number 200
[ASBR1-bgp]peer 34.34.34.2 route-policy policy1 export
[ASBR1-bgp]peer 34.34.34.2 label-route-capability
[ASBR1-bgp]quit
//On ASBR1, applies the route-policy to the routes to be advertised to ASBR2. When advertising the routes received from RR1 to ASBR2, ASBR1 will assign MPLS labels to the routes.

[ASBR1]bgp 100
[ASBR1-bgp]network 1.1.1.1 32
[ASBR1-bgp]network 7.7.7.7 32
//Configures ASBR1 to advertise routes destined for the loopback interfaces of PE1 and RR1 to ASBR2 so that ASBR2 can advertise the routes to RR2 and PE2.
[ASBR1-bgp]quit
```

- Repeat the configuration of ASBR1 on ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option C Solution 1 (1)

- After the preceding configurations are complete, CE1 and CE2 can learn the routes to each other's loopback interfaces and ping each other successfully.
- Take the command output on CE1 as an example:

```
[CE1]dis ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 6        Routes : 6

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

10.1.1.0/30         Direct  0    0           D   10.1.1.1        GigabitEthernet0/0/1
10.1.1.1/32         Direct  0    0           D   127.0.0.1       GigabitEthernet0/0/1
11.11.11.11/32      Direct  0    0           D   127.0.0.1       LoopBack0
22.22.22.22/32      EBGP    255  0           D   10.1.1.2        GigabitEthernet0/0/1
127.0.0.0/8         Direct  0    0           D   127.0.0.1       InLoopBack0
127.0.0.1/32        Direct  0    0           D   127.0.0.1       InLoopBack0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option C Solution 1 (2)

- Run the **display bgp routing-table label** command on an ASBR to check the label information of the routes.

```
[ASBR1]dis bgp routing-table label

BGP Local router ID is 23.23.23.2
Status codes: * - valid, > - best, d - damped,
              h - history,  i - internal, s - suppressed, S - Stale
              Origin : i - IGP, e - EGP, ? - incomplete

Total Number of Routes: 4
      Network         NextHop         In/Out Label

*>    1.1.1.1         23.23.23.1      1080/NULL
*>    6.6.6.6         34.34.34.2      1077/1071
*>    7.7.7.7         23.23.23.1      1081/NULL
*>    8.8.8.8         34.34.34.2      1076/1070
```

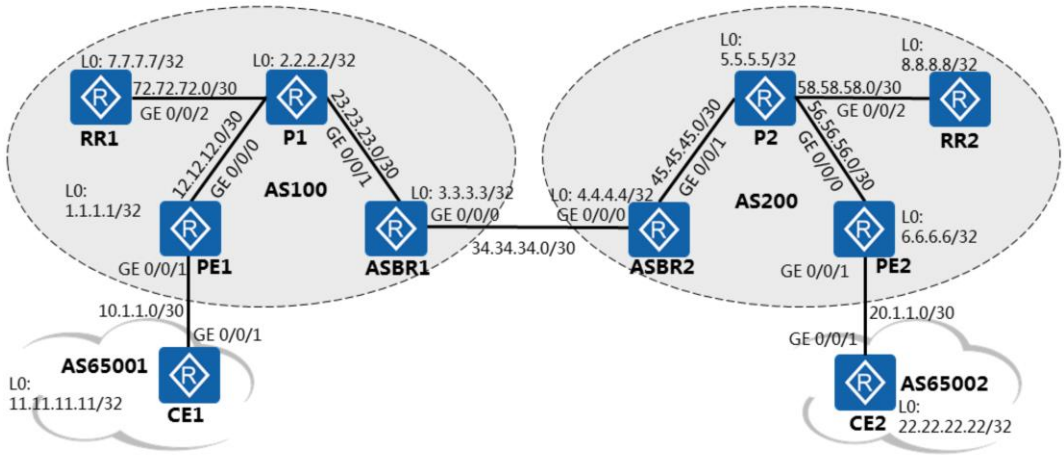
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

- Background
- Fundamentals
- **Configurations**
  - Option A Solution Configuration
  - Option B Solution Configuration
  - Option C Solution 1 Configuration
  - Option C Solution 2 Configuration

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Topology in Inter-AS MPLS BGP VPN Option C (Solution 2)



- Configuration task: Use Option C solution 2 to implement mutual access between CE1 and CE2. Note: OSPF is used in the ASs in this example.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- On the network shown in the figure, AS100 and AS200 are used for the ISP, whereas the other two ASs are used for the customer. PE1, P1, RR1 and ASBR1 belong to AS100. PE2, P2, RR2, and ASBR2 belong to AS200. CE1 and CE2 belong to the same VPN. CE1 is connected to PE1 in AS100, and CE2 is connected to PE2 in AS200.
- The IP addresses planned on each router are shown in this topology.
- Option C solution 2 is used in this example. The implementation of solution 2 is similar to that of solution 1. The difference is that when the local ASBR receives the labeled IPv4 routes from the peer ASBR, LDP is triggered and can allocate labels to labeled BGP public network routes.

Configuration Procedure in Inter-AS MPLS BGP VPN Option C Solution 2 (1)

Step	Description	Reference Command
1	Assign an IP address and mask to each interface.	See "Basic IPv4 Configuration" in the related product manual.
2	Configure OSPF on each PE, P, RR and ASBR over the public network.	See "Configuring Basic OSPF Functions" in the related product manual.
3	Configure MPLS and LDP on each PE, P and ASBR over the public network.	mpls mpls lsr-id lsr-id mpls ldp
4	Configure an EBGP peer relationship between ASBR1 and ASBR2. In each AS, configure the ASBR to advertise the route to the loopback interface of the local RR&PE and import the route to the IGP process.	See "Configuring Basic BGP Functions" in the related product manual. import-route bgp
5	Configure a VPN instance with an RD (100:1) and RTs (100:1) on each PE.	ip vpn-instance vpn-instance-name route-distinguisher route-distinguisher vpn-target vpn-target [both |export-extcommunity | importextcommunity]
6	Bind the AC interface to the VPN instance on each PE.	ip binding vpn-instance vpn-instance-name
7	In the VPN instance on each PE, configure an EBGP peer relationship with the connected CE.	See "Configuring Basic BGP Functions" in the related product manual.
8	Configure a multi-hop MP-EBGP peer relationship between RR1 and RR2 in the VPNv4 view. Configure an MP-IBGP peer relationship between PE1 and RR1 in the VPNv4 view. Configure an MP-IBGP peer relationship between PE2 and RR2 in the VPNv4 view. Disable each RR from changing the Next_Hop of the routes being advertised.	peer { group-name | ipv4-address | ipv6-address } ebgp-max-hop [hop-count] ipv4-family vpnv4 peer { group-name | ipv4-address | ipv6-address } enable peer next-hop-invariable

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Configuration Procedure in Inter-AS MPLS BGP VPN Option C Solution 2 (2)

Step	Description	Reference Command
9	Enable the capability of exchanging labeled IPv4 routes between ASBRs.	peer { group-name | ipv4-address } label-route-capability
10	On each ASBR, enable LDP to allocate labels to labeled BGP public network routes.	lsp-trigger bgp-label-route
11	Apply a route-policy to the routes to be advertised by the local ASBR to the remote ASBR.	route-policy route-policy-name { permit | deny } node node peer { group-name | ipv4-address | ipv6-address } route-policy route-policy-name { import | export }
Remarks	1. Enable MPLS on ASBR1 and ASBR2. LDP is optional. 2. There is no need to configure VPN instances or bind them to the interfaces on ASBR1 and ASBR2. 3. Configure ASBR1 and ASBR2 to exchange labeled IPv4 routes with each other. 4. Ensure that the local RR does not change the Next_Hop of the VPN routes when advertising them to the remote RR. Also ensure that the remote RR does not change the Next_Hop of the received VPN routes when advertising them to the PE in the same AS. 5. Configure a route-policy on each of ASBR1 and ASBR2 so that an ASBR can allocate MPLS labels when advertising routes to the ASBR in the peer AS. 6. On each ASBR, enable LDP to allocate labels to labeled BGP public network routes.	

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 2 (1)

- For the configurations in steps 1 to 3, see the related configurations in Option A.

4. Configure an EBGP peer relationship between ASBR1 and ASBR2. In each AS, configure the ASBR to advertise the route to the loopback interface of the local RR&PE and import the route to the IGP process.

```
[ASBR1]bgp 100
[ASBR1-bgp]peer 34.34.34.2 as-number 200
[ASBR1-bgp]network 7.7.7.7 255.255.255.255
[ASBR1-bgp]network 1.1.1.1 255.255.255.255
[ASBR1-bgp]quit
[ASBR1]ospf
[ASBR1-ospf-1]import-route bgp
```

```
[ASBR2]bgp 200
[ASBR2-bgp]peer 34.34.34.1 as-number 100
[ASBR2-bgp]network 8.8.8.8 255.255.255.255
[ASBR2-bgp]network 6.6.6.6 255.255.255.255
[ASBR2-bgp]quit
[ASBR2]ospf
[ASBR2-ospf-1]import-route bgp
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The purpose of importing BGP routes to the OSPF process is to ensure that RR1 and RR2 can establish an EBGP peer relationship to transmit VPN routes. It is recommended that a route-policy be configured to precisely import BGP routes to the OSPF process. This prevents unwanted routes from entering the IGP area.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 2 (2)

5. Configure a VPN instance with an RD (100:1) and RTs (100:1) on each PE.

```
[PE1]ip vpn-instance huawei
[PE1-vpn-instance-huawei]route-distinguisher 100:1
[PE1-vpn-instance-huawei-af-ipv4]vpn-target 100:1 both
```

6. Bind the AC interface to the VPN instance on each PE.

```
[PE1]interface GigabitEthernet0/0/1
[PE1-GigabitEthernet0/0/1]ip binding vpn-instance huawei
[PE1-GigabitEthernet0/0/1]ip address 10.1.1.2 30
[PE1-GigabitEthernet0/0/1]quit
```

- Repeat the configuration of PE1 on PE2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 2 (3)

7. In the VPN instance on each PE, configure an EBGP peer relationship with the connected CE.

```
[PE1]bgp 100
[PE1-bgp]ipv4-family vpn-instance huawei
[PE1-bgp-huawei]peer 10.1.1.1 as-number 65001
//Configures an EBGP peer relationship between PE1 and CE1.
```

```
[CE1]bgp 65001
[CE1-bgp]peer 10.1.1.2 as-number 100
[CE1-bgp]network 11.11.11.11 32
```

- Repeat the configuration of PE1 on PE2, and repeat the configuration of CE1 on CE2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 2 (4)

8. Configure an MP-EBGP peer relationship between RR1 and RR2 in the VPNv4 view, and disable the RRs from changing the Next_Hop of the VPN routes being advertised.

```
[RR1]bgp 100
[RR1-bgp]peer 8.8.8.8 as-number 200
[RR1-bgp]peer 8.8.8.8 connect-interface LoopBack 0
[RR1-bgp]peer 8.8.8.8 ebgp-max-hop 10
[RR1-bgp]ipv4-family vpnv4
[RR1-bgp-af-vpnv4]peer 8.8.8.8 enable
[RR1-bgp-af-vpnv4]peer 8.8.8.8 next-hop-invariable
[RR1-bgp-af-vpnv4]undo policy vpn-target
```

- In each AS, configure an MP-IBGP peer relationship between the PE and RR in the VPNv4 view, and disable the RR from changing the Next_Hop of the VPN routes being advertised.

```
[RR1]bgp 100
[RR1-bgp]peer 1.1.1.1 as-number 100
[RR1-bgp]peer 1.1.1.1 connect-interface LoopBack 0
[RR1-bgp]ipv4-family vpnv4
[RR1-bgp-af-vpnv4]peer 1.1.1.1 enable
[RR1-bgp-af-vpnv4]peer 1.1.1.1 next-hop-invariable
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- For the establishment of an MP-IBGP peer relationship between PE2 and RR2, see the configuration between PE1 and RR1.
- The **undo policy vpn-target** command configuration in Option C functions the same as that in Option B. They both disable RRs from filtering routes based on RTs.
- The **peer** *X.X.X.X* **next-hop-invariable** command configuration ensures that the peer PE can recurse routes to the BGP LSP destined for the local PE during traffic forwarding.
- Establish an MP-EBGP peer relationship between the RRs in the VPNv4 view, and disable the local RR from changing the Next_Hop attribute of the routes being advertised to the peer RR. That is, the next hop of a VPNv4 route learned by the peer PE is the local PE.
- Establish an MP-IBGP peer relationship between the RR and PE in the VPNv4 view, and disable the RR from changing the Next_Hop attribute of routes being advertised to the local PE. That is, the next hop of a VPNv4 route learned by the local PE is the peer PE.
- A PE establishes a VPNv4 peer relationship only with the RR in the same AS. The local RR establishes a VPNv4 peer relationship with the peer RR to transmit inter-AS VPN routes.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 2 (5)

9. Enable the capability of exchanging labeled IPv4 routes between ASBRs.

```
[ASBR1]bgp 100
[ASBR1-bgp]peer 34.34.34.2 as-number 200
[ASBR1-bgp]peer 34.34.34.2 label-route-capability
```

```
[ASBR2]bgp 200
[ASBR2-bgp]peer 34.34.34.1 as-number 100
[ASBR2-bgp]peer 34.34.34.1 label-route-capability
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 2 (6)

10. On each ASBR, enable LDP to allocate labels to labeled BGP public network routes.

```
[ASBR1]mpls
[ASBR1-mpls]lsp-trigger bgp-label-route
```

```
[ASBR2]mpls
[ASBR2-mpls]lsp-trigger bgp-label-route
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Example for Configuring Inter-AS MPLS BGP VPN Option C Solution 2 (7)

11. Apply a route-policy to the routes to be advertised by the local ASBR to the remote ASBR.

```
[ASBR1]interface GigabitEthernet0/0/0
[ASBR1-GigabitEthernet0/0/0]ip address 34.34.34.1 30
[ASBR1-GigabitEthernet0/0/0]mpls
[ASBR1-GigabitEthernet0/0/0]quit
[ASBR1]route-policy policy1 permit node 10
[ASBR1-route-policy]apply mpls-label //Assigns MPLS labels to the routes that hit the filtering rule.
[ASBR1-route-policy]quit
[ASBR1]bgp 100
[ASBR1-bgp]peer 34.34.34.2 as-number 200
[ASBR1-bgp]peer 34.34.34.2 route-policy policy1 export
[ASBR1-bgp]network 1.1.1.1 32
```

- Repeat the configuration of ASBR1 on ASBR2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option C Solution 2 (1)

- After the preceding configurations are complete, CE1 and CE2 can learn the routes to each other's loopback interfaces and ping each other successfully.
- Take the command output on CE1 as an example:

```
[CE1]dis ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 6        Routes : 6

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

10.1.1.0/30         Direct  0    0         D     10.1.1.1        GigabitEthernet0/0/1
10.1.1.1/32         Direct  0    0         D      127.0.0.1      GigabitEthernet0/0/1
11.11.11.11/32      Direct  0    0         D     127.0.0.1       LoopBack0
22.22.22.22/32      EBGP    255  0         D     10.1.1.2        GigabitEthernet0/0/1
127.0.0.0/8         Direct  0    0         D     127.0.0.1       InLoopBack0
127.0.0.1/32        Direct  0    0         D     127.0.0.1       InLoopBack0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option C Solution 2 (2)

- Run the **display bgp routing-table label** command on an ASBR to check the label information of the routes.

```
[ASBR1]dis bgp routing-table label

BGP Local router ID is 34.34.34.1
Status codes: * - valid, > - best, d - damped,
              h - history,  i - internal, s - suppressed, S - Stale
              Origin : i - IGP, e - EGP, ? - incomplete

Total Number of Routes: 4
      Network         NextHop        In/Out Label

*>    1.1.1.1         23.23.23.1     1047/NULL
*>    6.6.6.6         34.34.34.2     NULL/1041
*>    7.7.7.7         23.23.23.1     1043/NULL
*>    8.8.8.8         34.34.34.2     NULL/1042
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Verifying the Configuration of Inter-AS MPLS BGP VPN Option C Solution 3 (2)

- Run the **display mpls lsp** command on a PE to check for the LDP LSP to the peer PE.

```
[PE1]dis mpls lsp
------------------------------------------------------------------------------
             LSP Information: BGP  LSP
------------------------------------------------------------------------------
FEC                 In/Out Label       In/Out IF             Vrf Name
11.11.11.11/32      1042/NULL          -/-                   huawei
------------------------------------------------------------------------------
             LSP Information: LDP LSP
------------------------------------------------------------------------------
FEC                 In/Out Label       In/Out IF             Vrf Name
2.2.2.2/32          NULL/3             -/GE0/0/0
2.2.2.2/32          1037/3             -/GE0/0/0
6.6.6.6/32          NULL/1033          -/GE0/0/0
6.6.6.6/32          1039/1033          -/GE0/0/0
.............
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Quiz

- Which of the following solutions does not require an ASBR to save CE-side routing information? ( )
  - Option A
  - Option B
  - Option C
- When only LDP or BGP is used to allocate labels, which of the following solutions may use three labels during packet forwarding? ( )
  - Option A
  - Option B
  - Option C

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

1. C
2. C

Section Summary

- Fundamentals and Configurations of Option A
- Fundamentals and Configurations of Option B
- Fundamentals and Configurations of Option C

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— Security

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

Foreword

- Internet development has made the TCP/IP protocol suite the most widely used network interconnection protocols. However, because the protocol suite is initially designed for a small number of trusted user groups in a research environment, network security is not the top consideration. This leaves some security risks in the protocols. Considering that security has become a trending topic nowadays in the information age, this document provides some basic network security technologies.

Copyright © 2019 Huawei Technologies Co., Ltd.

Objectives

- Upon completion of this course, you will be able to:
  - Describe common attacks on a network.
  - Deploy basic attack defense technologies on a network.

Copyright © 2019 Huawei Technologies Co., Ltd.



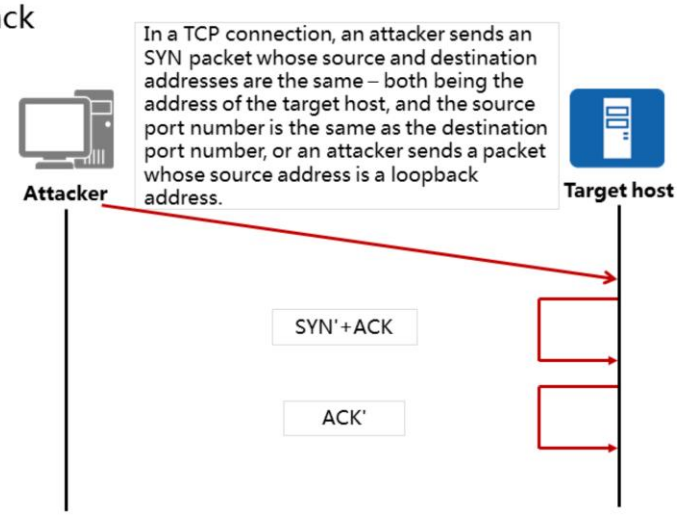
Contents

1. **Single-Packet Attack Defense**
2. Flood Attack Defense
3. Source IP Address Spoofing Attack Defense
4. Man-in-the-Middle Attack Defense
5. Data Transmission Security

Copyright © 2019 Huawei Technologies Co., Ltd.

Example of a Single-Packet Attack — LAND Attack

- LAND attack



Copyright © 2019 Huawei Technologies Co., Ltd.

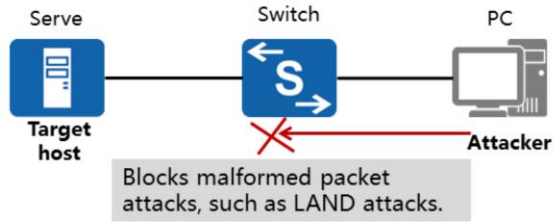
HUAWEI

- Single-packet attacks are a type of denial of service (DoS) attack and classified into the following types:
  - Scan attack: a potential attack behavior that has not produced direct damage. It is usually a network detection behavior prior to a real attack. Examples of such attacks include IP address scan attacks and port scan attacks.
  - Malformed packet attack: An attacker sends a large number of malformed packets to crash the host or server that processes these packets. Examples of such attacks include LAND attacks and smurf attacks.
  - Special control packet attack: Normal packets are used to snoop on a network structure or attack a system or network, leading to a system breakdown or network disconnection. Examples of such attacks include oversized ICMP packet attacks and ICMP destination unreachable packet attacks.
- Information about LAND attacks
  - Exploiting the defect in the TCP three-way handshake mechanism, an attacker sends an SYN packet to a target host. In this packet, the source and destination addresses both are the address of the target host, and the source port number is the same as the destination port number. After the target host receives the packet, it creates an empty TCP connection where the source and destination addresses are both itself until the connection times out. In this attack mode, the target host creates a large number of useless empty TCP connections, consuming a lot of resources until the host crashes.

Example of Single-Packet Attack Defense — LAND Attack Defense

- How LAND attack defense works
  - After defense against malformed packet attacks is enabled on a device, the device checks the source and destination addresses of TCP SYN packets to prevent LAND attacks. If the source address and destination address in a TCP SYN packet are the same, the TCP SYN packet is considered a malformed packet and is discarded.
  - Enable defense against malformed packet attacks (system view):

anti-attack abnormal enable



Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- All attack defense functions, including malformed packet attack defense, can be enabled using the **anti-attack enable** command in the system view.

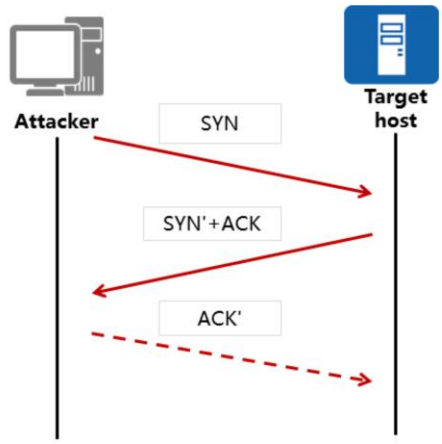
Contents

1. Single-Packet Attack Defense
2. **Flood Attack Defense**
3. Source IP Address Spoofing Attack Defense
4. Man-in-the-Middle Attack Defense
5. Data Transmission Security

Copyright © 2019 Huawei Technologies Co., Ltd.

Example of a Flood Attack — TCP SYN Flood

- TCP SYN attack

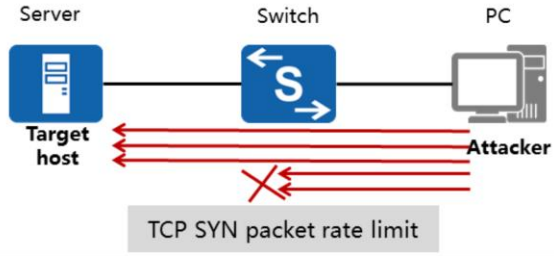


Copyright © 2019 Huawei Technologies Co., Ltd.

- Flood attacks are also a type of DoS attack.
- Information about TCP SYN attacks
  - A TCP SYN attack exploits the vulnerability in the TCP three-way handshake mechanism. During the TCP three-way handshake, when a receiver receives the first SYN packet from a sender, it returns an SYN+ACK packet to the sender and keeps waiting for the final ACK packet from the sender. In this process, the connection is always in the half-open state. If the receiver does not receive the final ACK packet, it retransmits a SYN+ACK packet to the sender. If the sender does not return an ACK packet after multiple times, the receiver closes the session and refreshes the session in the memory. During this period, the attacker may send hundreds of thousands of SYN packets to an open port and does not respond to the SYN+ACK packets from the receiver. The receiver soon becomes overloaded, cannot process any new connection requests, and disconnects all active connections.

Example of Flood Attack Defense — TCP SYN Flood Attack Defense

- How TCP SYN flood attack defense works
  - After TCP SYN flood attack defense is enabled on a device, the device limits the rate of TCP SYN packets to ensure that the resources of the target host are not exhausted.
  - **anti-attack tcp-syn enable**
    anti-attack tcp-syn car cir *cir*



Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- Flood attack defense commands
  - The **anti-attack tcp-syn enable** command enables defense against TCP SYN flood attacks.
  - The **anti-attack tcp-syn car** command sets a rate limit for TCP SYN flood attack packets. If the receiving rate of TCP SYN flood packets exceeds the limit, the device discards excess packets to ensure that the CPU works properly.



Contents

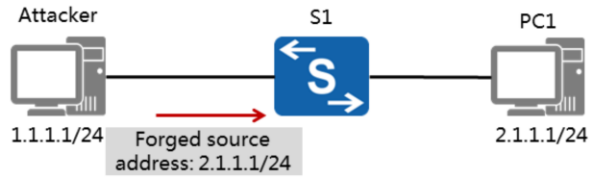
1. Single-Packet Attack Defense
2. Flood Attack Defense
3. **Source IP Address Spoofing Attack Defense**
4. Man-in-the-Middle Attack Defense
5. Data Transmission Security

Copyright © 2019 Huawei Technologies Co., Ltd.

URPF

- Source IP address spoofing: Hosts on a network are usually trusted based on their IP addresses. Source IP address spoofing uses forged source addresses to win trust for stealing network information or damaging system communication.
- About unicast reverse path forwarding (URPF)
  - Defends against network DoS attacks that are launched based on spoofed source IP addresses.
  - Working modes
    - Strict mode
    - Loose mode
  - Command (interface view)

URPF { **loose** | **strict** } [**allow-default-route**] [**acl** *acl-number*]



Copyright © 2019 Huawei Technologies Co., Ltd.

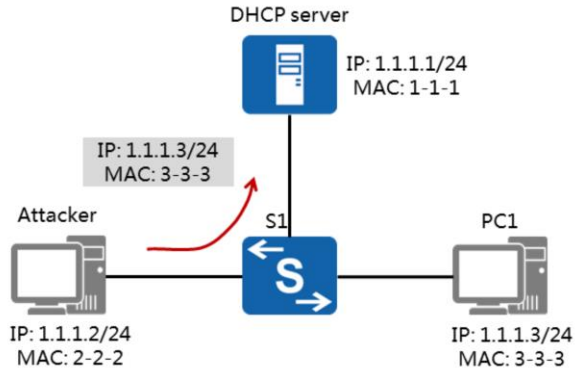
HUAWEI

- URPF works in either of the following modes:
  - Strict mode
    - In strict mode, a packet passes the URPF check only when a device has a route to the source IP address of the packet in its FIB table and the inbound interface of the packet is the same as the outbound interface of the route. In the preceding figure, an attacker forges a packet with the source address being 2.1.1.1 to initiate a request to S1. After receiving the request, S1 sends a packet to the real host (PC1) that possesses 2.1.1.1. The forged packet is an attack on both S1 and PC1. If URPF is enabled on S1, when S1 receives a packet with the source address being 2.1.1.1, URPF checks that the outbound interface corresponding to the source address of the packet does not match the interface that receives the packet and therefore discards the packet.
    - You are advised to use the strict URPF mode in an environment with symmetric routes. For example, if there is only one path between two network edge devices, the URPF strict mode can be used to maximize network security.
  - Loose mode
    - In loose mode, a packet passes the check as long as the device has a route to the source IP address of the packet in its FIB table, and the inbound interface of the packet is not required to be the same as the outbound interface of the route.

- You are advised to use the URPF loose mode in an environment where routes are not symmetric. For example, if there are multiple paths between two network border devices, the URPF loose mode can be used to improve network security and prevent the packets transmitted along the correct path from being discarded.

IPSG

- About IP source guard (IPSG)
  - IPSG checks IP packets against a DHCP dynamic or static binding table to prevent source IP address spoofing.



Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- Information about IPSG
  - IPSG checks IP packets against a static binding table or DHCP dynamic binding table. Before forwarding an IP packet, a device compares the source IP address, source MAC address, port number, and VLAN ID in the IP packet with the information in the binding table. If the information matches, the packet is from an authorized user, and the device permits the packet; otherwise, the device considers the packet an attack and discards it. In the preceding figure, IPSG is configured on S1 to check the incoming IP packets against a binding table. Information about the packets sent by authorized users is the same as the information in the binding table, the packets are permitted. Information about forged packets from attackers is inconsistent with the information in the binding table, and the packets are discarded.
- IPSG commands
  - There are DHCP dynamic binding tables and static binding tables. A static binding

table is manually configured using the **user-bind static** command.

- The **ip source check user-bind enable** command enables the IP packet check function.
- The **ip source check user-bind check-item** command configures an IP packet check item against a VLAN ID or an interface. This command takes effect only for dynamic binding entries.

Contents

1. Single-Packet Attack Defense
2. Flood Attack Defense
3. Source IP Address Spoofing Attack Defense
4. **Man-in-the-Middle Attack Defense**
5. Data Transmission Security

Copyright © 2019 Huawei Technologies Co., Ltd.

Man-in-the-Middle Attack Defense — DAI

- Man-in-the-middle attack: As the name implies, an attacker is located between a client and a server. The attacker impersonates the server for the client and impersonates the client for the server.
- Dynamic ARP inspection (DAI)
  - Uses DHCP snooping binding tables to defend against man-in-the-middle attacks. When a device receives an ARP packet, it compares the source IP address, source MAC address, VLAN ID, and interface information in the ARP packet with those in a DHCP snooping binding table. If the information matches, the device considers that the packet is from an authorized user and permits the packet. Otherwise, the device considers that
  - Command

arp anti-attack check user-bind enable

ARP entry of PC1

IP address	MAC address	Type
10.1.1.3	3-3-3	Dynamic

Updated ARP entry

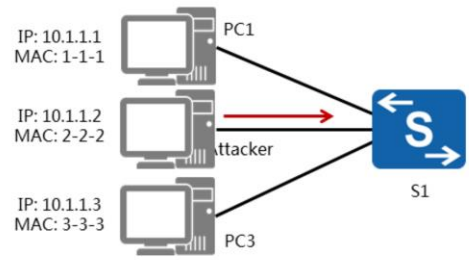
IP address	MAC address	Type
10.1.1.3	2-2-2	Dynamic

ARP entry of PC3

IP address	MAC address	Type
10.1.1.1	1-1-1	Dynamic

Updated ARP entry

IP address	MAC address	Type
10.1.1.1	2-2-2	Dynamic



Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- The preceding figure shows a man-in-the-middle attack. The attacker impersonates PC3 and proactively sends a forged ARP packet to PC1. As a result, the ARP table of PC1 records the incorrect mapping between PC3's IP address and MAC address. The attacker then easily obtains the data that PC1 desires to send to PC3. Similarly, the attacker can easily obtain the data that PC3 desires to send to PC1. In this way, information security between PC1 and PC3 cannot be ensured.
- To defend against the man-in-the-middle attack, DAI can be enabled on S1.
- After DAI is enabled on S1, if an attacker connecting to S1 attempts to send forged ARP packets, S1 will detect the attack against the DHCP snooping binding table and discard the ARP packets. If the DAI-based alarm function is also enabled on S1, S1 sends an alarm to an administrator when the number of ARP packets discarded due to mismatching the DHCP snooping binding entries exceeds the alarm threshold.
- DAI command

- The **arp anti-attack check user-bind enable** command enables DAI for an interface or a VLAN. After this command is run, ARP packets are checked against binding table entries.

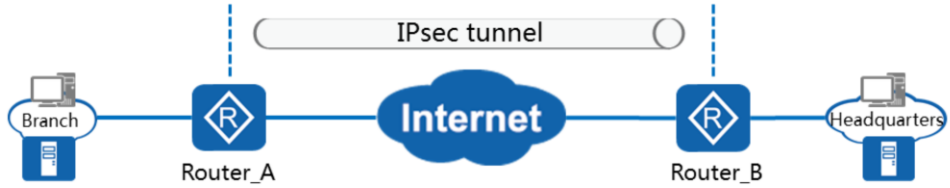
Contents

1. Single-Packet Attack Defense
2. Flood Attack Defense
3. Source IP Address Spoofing Attack Defense
4. Man-in-the-Middle Attack Defense
5. **Data Transmission Security**

Copyright © 2019 Huawei Technologies Co., Ltd.

IPsec VPN Overview

- Internet Protocol Security (IPsec) is an easy-to-use secure network protocol suite. It provides transparent security services for IP network communication by protecting TCP/IP communication against eavesdropping and tampering, and by effectively defending against network attacks.
- IPsec establishes a bidirectional security association (SA) between IPsec peers to form a secure IPsec tunnel for secure data transmission on the Internet.



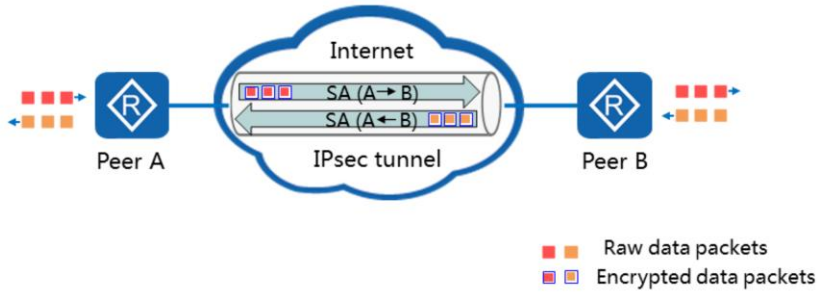
Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- IPsec deployed on a network can perform encryption, integrity check, and source authentication on transmitted data to mitigate information leakage risks.

IPsec SA

- An IPsec security association (SA) needs to be established between IPsec peers (two IPsec endpoints) before IPsec securely transmits data.
- An SA is the establishment of shared security attributes (such as algorithms and keys) between IPsec peers.



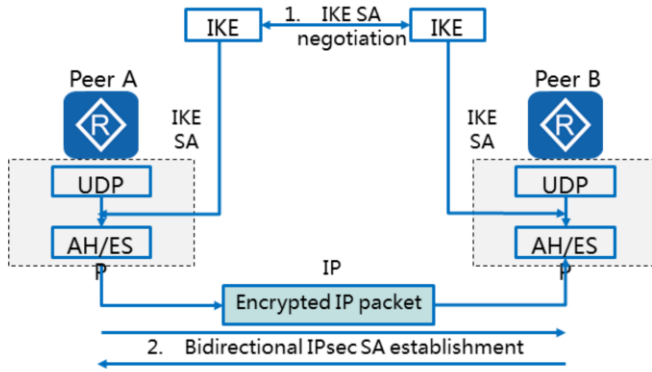
Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- IPsec peers establish shared security attributes in an SA for data transmission. The attributes include the security protocol, characteristics of data flows to be protected, data encapsulation mode, encryption algorithm, authentication algorithm, key exchange method, IKE, and SA lifetime.
- An SA is uniquely identified by three parameters: security parameter index (SPI), destination IP address, and security protocol ID (AH or ESP).

IKE SA

- An IKE SA serves an IPsec SA by providing automatic key negotiation.

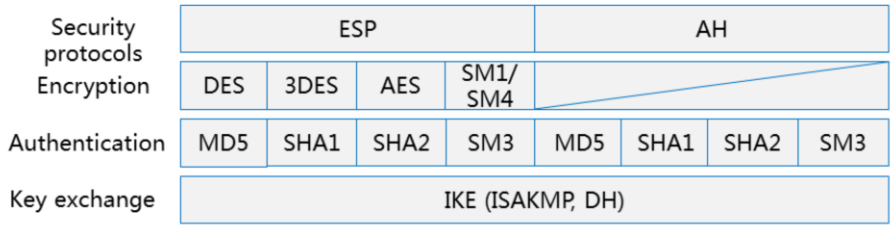


Copyright © 2019 Huawei Technologies Co., Ltd.

- Internet Key Exchange (IKE) works on the framework provided by the Internet Security Association and Key Management Protocol (ISAKMP). IKE is a UDP-based application layer protocol that provides keys for data encryption. It simplifies the use, management, configuration, and maintenance of IPsec.
- After an IKE SA is established between IPsec peers to complete identity authentication and key exchange, a pair of IPsec SAs is negotiated based on configured parameters, such as the AH or ESP protocol. Data is then encrypted and transmitted between the peers in an IPsec tunnel.

IPsec Security Protocols

- IPsec protects IP packets using two security protocols: Authentication Header (AH) and Encapsulating Security Payload (ESP).



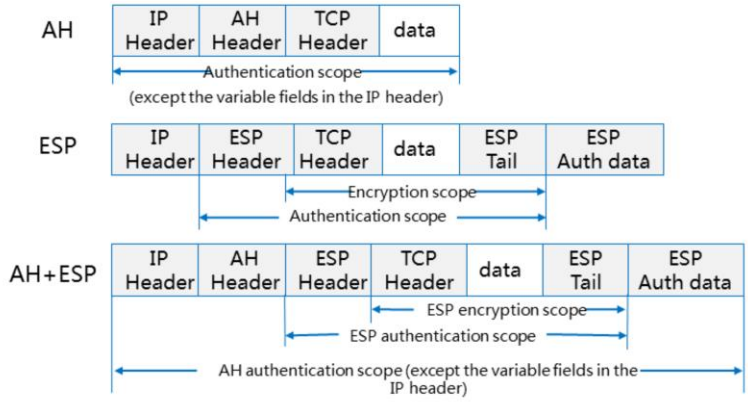
Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- AH provides data origin authentication, data integrity check, and protection against replay attacks, but does not provide encryption.
- ESP provides encryption, data origin authentication, data integrity check, and protection against replay attacks.
- Security functions provided by AH and ESP depend on the authentication and encryption algorithms used by IPsec.
- The keys used for IPsec encryption and authentication can be manually configured or dynamically negotiated using the IKE protocol. This course describes how to establish an IPsec tunnel manually.

Encapsulation Mode — Transport Mode

- In transport mode, an AH or ESP header is added between an IP header and a transport-layer protocol (TCP/UDP/ICMP) header to protect the TCP, UDP, or ICMP payload.



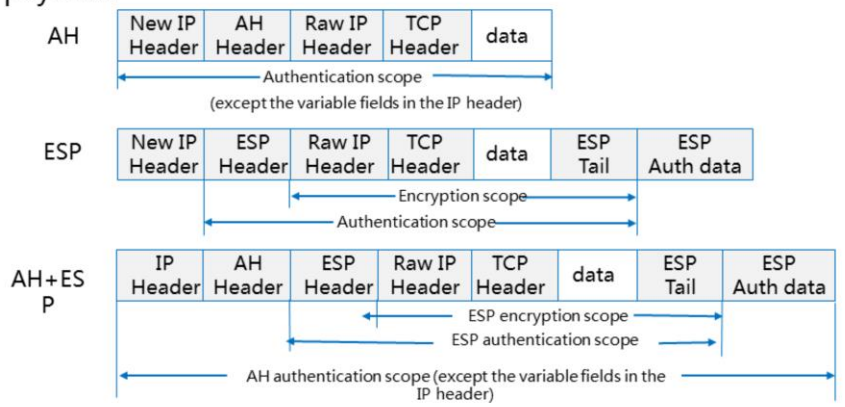
Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- The transport mode does not change the IP header, so the source and destination addresses of an IPsec tunnel must be the same as those in the IP header. This encapsulation mode applies only to communication between two hosts or between a host and a VPN gateway.

Encapsulation Mode — Tunnel Mode

- In tunnel mode, an AH or ESP header is added prior to the raw IP header and then encapsulated into a new IP packet with a new IP header to protect the IP header and payload.

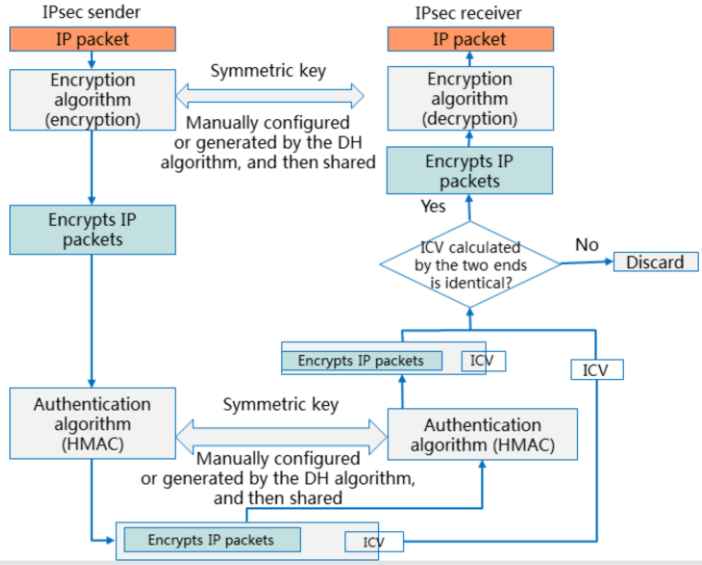


Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- The tunnel mode applies to communication between two VPN gateways or between a host and a VPN gateway.
- The two encapsulation modes differ in the following:
  - In terms of security, the tunnel mode is more secure than the transport mode. The tunnel mode can completely authenticate and encrypt original IP packets. It hides the IP addresses, protocol types, and port numbers in original IP packets.
  - In terms of performance, the tunnel mode occupies more bandwidth resources because of an extra IP header.

IPsec Encryption, Decryption, and Authentication



Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- IPsec uses symmetric encryption algorithms to encrypt and decrypt data.
- IPsec provides the authentication mechanism to allow a data receiver to authenticate a data sender in IP communication and determines whether data is modified during transmission.
- IPsec uses the keyed-hash message authentication code (HMAC) function for authentication. The HMAC function verifies the integrity and authenticity of data packets by comparing digital signatures.

Quiz

1. When IPsec uses the AH protocol to encapsulate data, it does not support data integrity check. ( )
   A. T
   B. F
2. Which of the following attacks are DoS attacks? ( )
   A. Malformed packet attack
   B. Source IP address spoofing
   C. Man-in-the-middle attack
   D. Flood attack

Copyright © 2019 Huawei Technologies Co., Ltd.

HUAWEI

- Answers:
  - B
  - A D

Section Summary

- Single-Packet Attack and Defense Method
- Flood Attack and Defense Method
- Source IP Address Spoofing Attack
- URPF
- IPSG
- DAI
- IPsec VPN

Copyright © 2019 Huawei Technologies Co., Ltd.

Thank You

www.huawei.com

HCIE-Routing & Switching

— HA and Network Management

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- A network must meet service continuity requirements in addition to data connectivity requirements. In this case, the network must be highly reliable and available to ensure that services are not affected when a fault occurs, and the fault should be quickly located and rectified. To meet the preceding requirements, this section describes the network high availability (HA) technologies and network management protocols.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to:
  - Deploy HA technologies based on application scenarios.
  - Manage networks and quickly locate network faults.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. **HA Technologies**
   - HA Overview
   - BFD
   - NSF
   - NSR
2. Network Management Protocols

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HA Overview

- High availability (HA)
  - Definition: A product or system has high availability.
  - HA characteristics
    - Low fault occurrence
    - Quick recovery after a fault occurs
  - HA standards
    - MTBF refers to mean time between failures of a component or device.
    - MTTR refers to the average time required for a component or device to recover from a fault.

$$\text{System availability} = \frac{\text{MTBF}}{\text{MTBF} + \text{MTTR}}$$

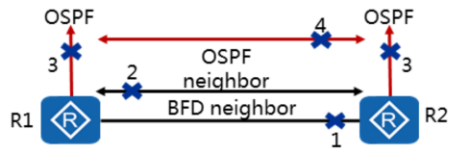
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

BFD - Workflow

- Bidirectional Forwarding Detection (BFD)
  - BFD is used to quickly detect and identify connectivity of links and IP forwarding on networks.
- BFD session setup



- BFD fault detection



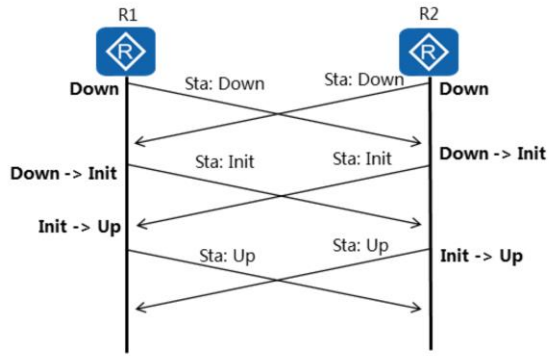
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Basic BFD concepts
  - Two network devices establish a BFD session to monitor the path between them and serve upper-layer applications. BFD does not provide neighbor discovery. Instead, BFD obtains information about neighbors from the upper-layer applications it serves. After two devices establish a BFD session, devices periodically send BFD packets. If a device does not receive a response within a set time limit, the device considers the forwarding path faulty. BFD will then notify the upper-layer protocol.
  - BFD control packets are encapsulated using UDP. The destination port number is 3784, and the source port number ranges from 49152 to 65535.
- BFD session setup
  1. OSPF uses the Hello mechanism to discover neighbors and establishes a neighbor relationship.
  2. OSPF notifies BFD of neighbor information including source and destination addresses.
  3. BFD sets up a BFD session based on received neighbor information.
  4. After the BFD session is established, BFD starts to monitor link faults, responding quickly to faults.
- BFD fault detection
  1. The detected link fails.
  2. The BFD session is considered Down.
  3. BFD notifies the local OSPF process that the neighbor is unreachable.

4. The local OSPF process ends the OSPF neighbor relationship.

BFD - Session Management

- BFD session management
  - The BFD session state can be Down, Init, Up, or AdminDown.
  - BFD state machine setup and release use the three-way handshake mechanism.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- State of a BFD session
  - Down: A BFD session is in the Down state or a request has been sent.
  - Init: The local end can communicate with the remote end and wants the session state to be Up.
  - Up: A BFD session is successfully established.
  - AdminDown: A BFD session is in the AdminDown state.
- BFD session migration process
  - BFD configured on both R1 and R2 independently starts state machines. The initial state of BFD state machines is Down. R1 and R2 send BFD control packets with the State field set to Down.
  - After receiving a BFD control packet with the State field set to Down, R2 switches the session state to Init and sends a BFD control packet with the State field set to Init.
  - After the local BFD session state of R2 changes to Init, R2 no longer processes the received BFD control packets with the State field set to Down.
  - The BFD state change of R1 is the same as that of R2.
  - After receiving a BFD control packet with the State field set to Init, R2 changes the local session state to Up.
  - The BFD state change of R1 is the same as that of R2.

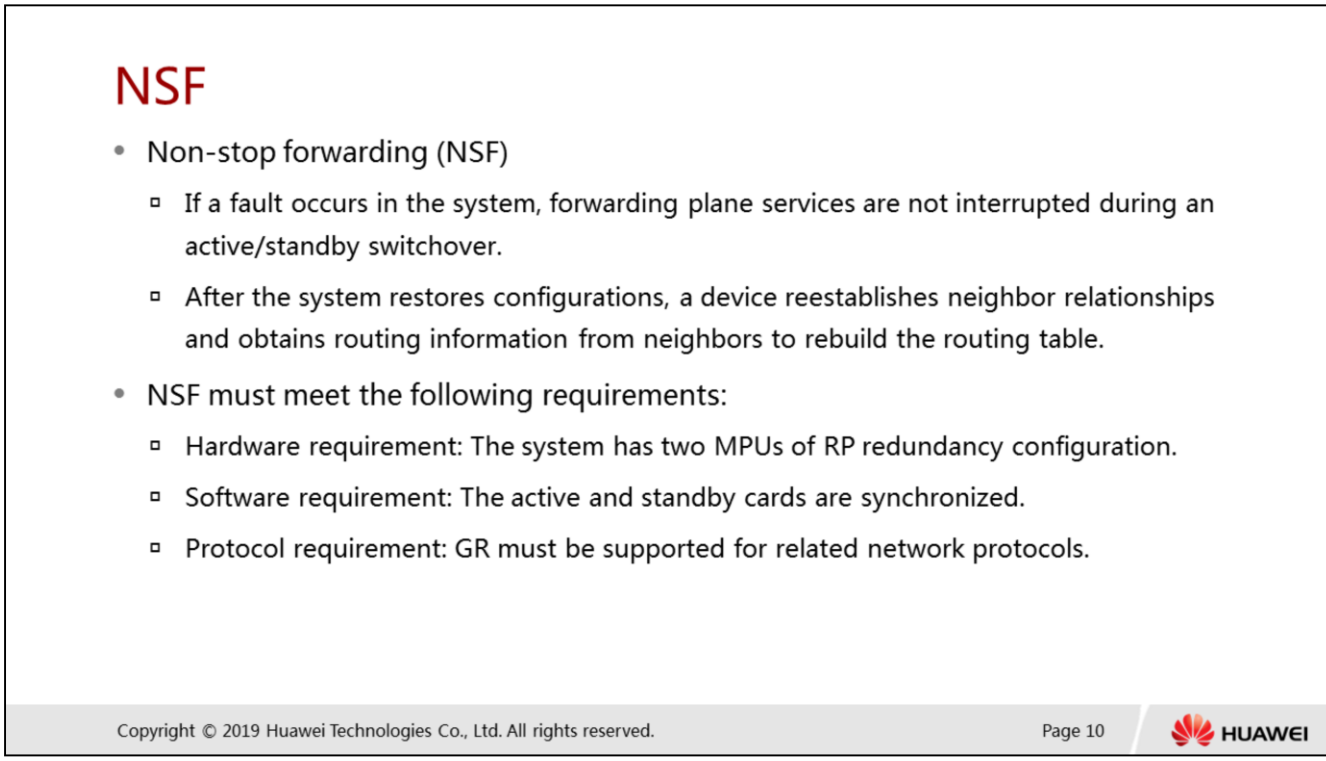
BFD - Application Scenario

- BFD application scenarios
  - BFD for IP links
  - Association between the BFD session status and the interface status
  - BFD for static routes
  - BFD for OSPF
  - BFD for IS-IS
  - BFD for BGP
  - BFD for MPLS LSPs
  - BFD for VRRP
  - BFD for PIM

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Common commands
  - IP link detection is classified into single-hop detection and multi-hop detection.
    - The **bfd** command enables BFD globally in the system view and displays the BFD global view.
    - The **bfd bind peer-ip** command creates a BFD binding relationship and sets up a BFD session.
    - The **discriminator** command sets the local and remote discriminators for the current BFD session.
    - The **commit** command commits the BFD session configuration.
    - Based on whether the peer device supports BFD, there are two scenarios: 1. When the peer device supports BFD, create a BFD session that can be established only when BFD parameters are negotiated at both ends and both ends sends packets to the MPU. 2. When the peer device does not support BFD, create a BFD one-arm echo session.
  - Association between the BFD session status and the interface status
    - The **bfd** command enables BFD globally in the system view and displays the BFD global view.

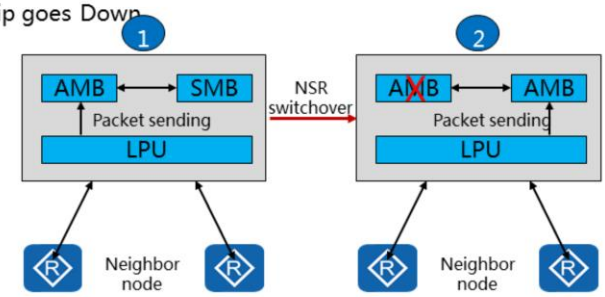
- The **bfd bind peer-ip default-ip** command creates a BFD binding relationship for detecting the physical status of a link.
- The **discriminator** command sets the local and remote discriminators for the current BFD session.
- The **process-interface-status** command associates a BFD session with an interface to which the BFD session is bound.
- The configuration of interface status association is similar to that of route association, and therefore is not described here.



- When a router fails, neighbors at the routing protocol layer detect that their neighbor relationships are Down and then become Up again after a period of time. This is the flapping of neighbor relationships. The flapping of neighbor relationships causes route flapping, which leads to black hole routes on the restarted router or causes data services from the neighbors to bypass the restarting router. This decreases the reliability on the network. The objective of the NSF technology is to solve the problem of route flapping. Therefore, the following requirements must be met:
  - Hardware requirement: The system has two MPUs of RP redundancy configuration. One is the active MPU, and the other is the standby MPU. When the active MPU restarts, the standby MPU becomes active. The distributed structure is used, so data forwarding and control are separated. Dedicated LPUs (interface cards) are used for data forwarding.
  - System software requirement: When the active MPU is running normally, it synchronizes configuration and interface state information to the standby MPU. When an active/standby switchover occurs, interfaces remain Up and the LPUs do not reset or withdraw forwarding entries.
  - Protocol requirement: GR must be supported for related network protocols, such as routing protocols OSPF, IS-IS, and BGP, and other protocols such as Label Distribution Protocol (LDP).

NSR

- Non-stopping routing (NSR): The protocol backup mechanism ensures that the control plane (routing) and forwarding plane (services) are not interrupted during an active/standby switchover.
  - During the device switchover, route processing is not interrupted because neighbor and topology information is not lost.
  - No neighbor relationship goes Down.
- NSR workflow
  1. Batch backup
  2. Real-time backup
  3. AMB/SMB switchover



1. The routing protocol on the control plane backs up routing information in real time.
2. The hardware channel detects the exception on the AMB, the SMB becomes the new AMB, and LPUs send packets to the new AMB.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- NSR advantages
  - NSR does not affect or rely on the peer device.
  - The route convergence speed of NSR is higher than that of NSF.
- NSR workflow
  1. Batch backup: After NSR is enabled and the SMB restarts, the service process on the AMB receives a message indicating that the SMB goes online. After receiving the message, the ACP backs up its data to the SCP in batches.
     - After batch backup is complete, the device enters the redundancy protection state. If the AMB fails, the SMB can become the new AMB and restore data.
     - If the AMB fails before batch backup is complete, the SMB cannot become the new AMB. The fault can be rectified after the device restarts.
  2. After batch backup is complete, the device enters the real-time backup phase. If the neighbor status or routing information changes on the AMB, the AMB backs up the updated information to the SMB in real time.
  3. If the AMB's software or hardware fails, the SMB detects the failure and automatically becomes the new AMB. The new AMB uses the backup data to forward traffic. The LPU sends the information that has been updated during the AMB/SMB switchover to the new AMB. Routes are reachable and traffic forwarding is uninterrupted during the switchover.
- During an AMB/SMB switchover, the system supports two types of HA protection: NSR

and GR. They are mutually exclusive. That is, for a specific protocol, after the system switchover, only one of NSR and GR processing can be used.



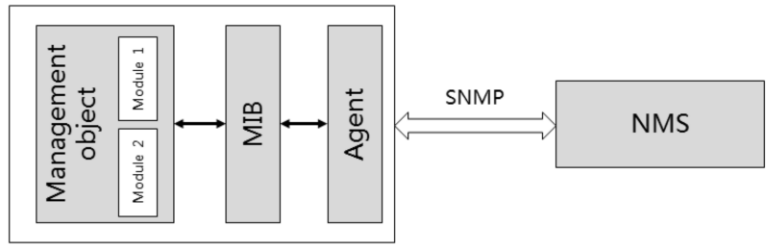
Contents

1. HA Technologies
2. **Network Management Protocols**
   - SNMP
   - NTP

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

SNMP - Basic Concept

- Basic concept of Simple Network Management Protocol (SNMP)
  - SNMP uses a central computer (a network management station) that runs network management software to manage devices.
- SNMP versions
  - SNMPv1, SNMPv2c, and SNMPv3
- SNMP model



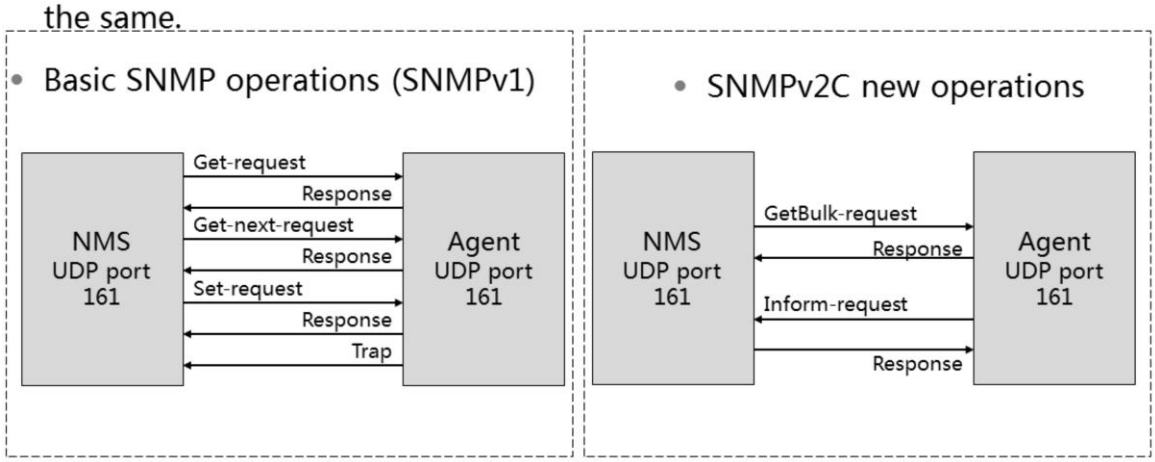
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- SNMP model
  - Network Management System (NMS): The NMS is usually an independent device that runs network management applications. The network management application provides at least one man-machine interface through which the network administrator manages the network.
  - SNMP agent: The agent is the software installed on the managed device. It receives and handles the request packets from the NMS, and returns responses to the NMS. In some urgent cases, the agent sends a trap packet to the NMS.
  - SNMP protocol: As an application layer protocol in the TCP/IP suite, SNMP exchanges management information between the NMS and the managed device.
  - Management information base (MIB): The MIB is a set of managed objects. It is a bridge between the NMS and the agent and enables the NMS software to interconnect with devices. Each agent maintains an MIB. The NMS reads or sets the value of an object contained in an MIB.
  - An MO is an object to be managed on a network device. A managed device contains multiple MOs, for example, a hardware component (such as an LPU) and a set of parameters configured for the hardware or software (such as a route selection protocol).

SNMP - Working Principles of SNMPv1/SNMPv2c

- The working principles of SNMPv1, SNMPv2c, and SNMPv3 are basically the same.



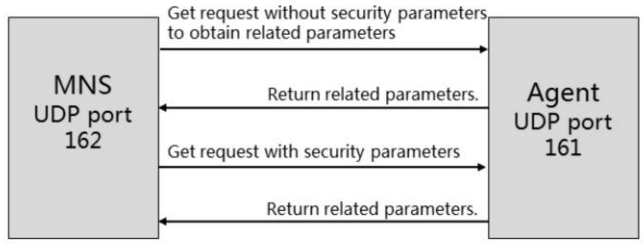
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Basic operations:
  - get-request: The NMS wants to fetch one or more parameters from the MIB of the agent.
  - get-next-request: The NMS wants to fetch the next parameter from the MIB of the agent.
  - set-request: That the NMS wants to set one or more parameters in the MIB of the agent.
  - get-response: It returns one or more parameters. It is generated by an agent and transmitted in reply to any of the preceding operations.
  - trap: It is sent by the agent to inform the NMS of some important events.
- SNMPv2C new operations:
  - getbulk-request: It queries information about managed devices in batches. A GetBulk operation is equal to consecutive GetNext operations. You can set the number of GetNext operations to be included in one GetBulk operation.
  - Inform-request: A managed device proactively sends alarms to the NMS. After a managed device sends an Inform packet, the NMS must send an InformResponse packet to the managed device.
- SNMPv1 and SNMPv2c have weak security.

SNMP - Working Principles of SNMPv3

- Authentication and encryption are added to SNMPv3.

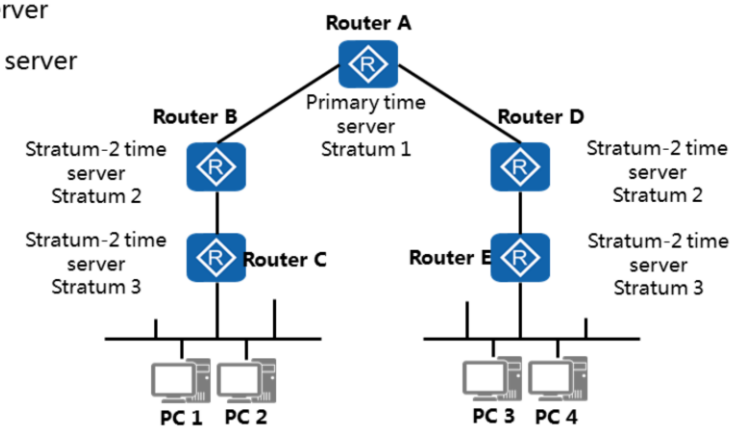


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The implementation principles of SNMPv3 are similar to those of SNMPv1 and SNMPv2c.
- Working principles of SNMPv3
  - The NMS sends a Get request message without security parameters to the agent and obtains security parameters (such as the SNMP entity engine information, user name, authentication parameters, and encryption parameters) from the agent.
  - The agent responds to the request from the NMS and sends the requested parameters to the NMS.
  - The NMS sends a Get request message with security parameters to the agent. (Security parameters are the authentication parameters used for identity authentication and encryption parameters used for packet encryption, and these parameters are calculated by the algorithms configured on the NMS.)
  - The agent authenticates the message and decrypts the message information. Then it encrypts the response message and sends the message to the NMS.

NTP - Network Architecture

- NTP network architecture
  - Synchronization subnet
  - Primary time server
  - Stratum-2 time server
  - Stratum



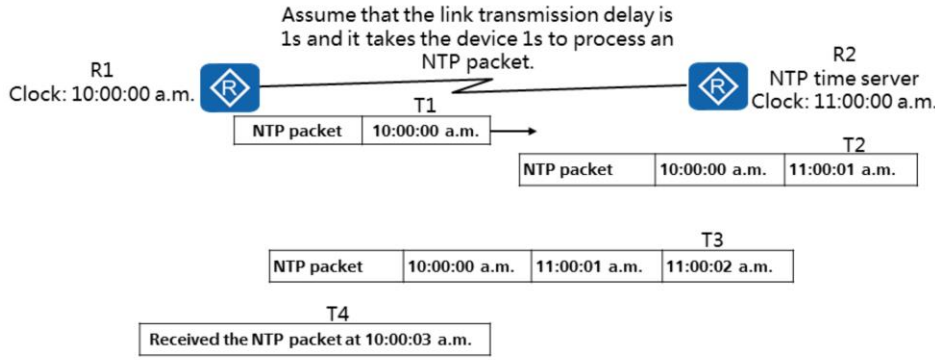
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Key concepts of NTP architecture and their functions include the following:
  - Synchronization subnet: consists of the primary time server, stratum-2 time servers, PC clients, and interconnecting transmission paths.
  - Primary time server: directly synchronizes its clock with a standard reference clock through a cable or radio. Typically, the standard reference clock is either a radio clock or the Global Positioning System (GPS).
  - Stratum-2 time server: synchronizes its clock with either the primary time server or other stratum-2 time servers within the network. A stratum-2 time server transmits the time to other hosts within the local area network (LAN) through NTP.
  - Stratum: is a hierarchical standard for clock synchronization. It represents the precision of a clock. The value of a stratum ranges from 1 to 16. A smaller value indicates higher precision. The value 1 indicates the highest precision, and 16 indicates that the clock is not synchronized.
- Under typical circumstances within a synchronization subnet, the primary time server and stratum-2 time servers are arranged in a hierarchical-active-standby structure. In this structure, the primary time server is located at the root, and stratum-2 time servers are located near leaf nodes. As their strata increase, their precision decreases accordingly. The decreased precision of the stratum-2 time servers varies with both the network path and local clock stability.

NTP - Working Principles

- NTP
  - It synchronizes time between time servers and clients.
  - NTP packets are transmitted by UDP over port 123.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- NTP synchronization process
  - R1 sends an NTP packet to R2. When the packet leaves R1, it carries a timestamp of 10:00:00 a.m. (T1).
  - When the NTP packet reaches R2, R2 adds a receive timestamp of 11:00:01 a.m. (T2) to the packet.
  - When the NTP packet leaves R2, R2 adds a transmit timestamp of 11:00:02 a.m. (T3) to the packet.
  - When R1 receives the response packet, it adds a new receive timestamp of 10:00:03 a.m. (T4) to the packet. R1 uses the received information to calculate the following important values:
    - Roundtrip delay for the NTP packet: Delay = (T4 - T1) - (T3 - T2)
    - Time difference between R1 and R2: Offset= ((T2 - T1) + (T3 - T4))/2
  - After the calculation, R1 knows that the roundtrip delay is 2 seconds and the clock offset is 1 hour. According to the delay and offset, R1 sets its own clock to synchronize with the clock of R2.



Quiz

1. NSF and NSR are used together to ensure service continuity. ( )
   A. T
   B. F

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Answer:
  - B

Summary

- HA Technologies
  - BFD
  - NSF
  - NSR
- Network Management Technologies
  - SNMP
  - NTP

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— QoS

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- As networks rapidly develop, services on the Internet become increasingly diversified. Apart from traditional applications such as WWW, email, and File Transfer Protocol (FTP), the Internet has expanded to encompass other services such as IP phones, e-commerce, multimedia games, e-learning, telemedicine, videophones, videoconferencing, video on demand (VOD), and online movies. In addition, enterprise users use Virtual Private Network (VPN) technologies to connect their branches in different areas so that they can access each other's corporate databases or manage remote devices through Telnet.
- Diversified services enrich users' lives but also increase the possibility of traffic congestion on the Internet. In the case of traffic congestion, services can encounter long delays or even packet loss. As a result, services deteriorate or even become unavailable. Therefore, a solution to resolve traffic congestion on the IP network is urgently needed. Quality of service (QoS) that uses a policy to manage traffic congestion at a low cost has been deployed. This document describes the advanced QoS features.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to:
  - Master the QoS multi-field classification technology.
  - Master the QoS token bucket algorithm.
  - Understand the QoS deployment solution on the live network.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



Contents

- **QoS MQC Configuration**
- QoS Token Bucket Algorithm
- QoS Application

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

QoS MQC Configuration

- MQC concept:
  - Modular QoS command line interface (MQC) allows you to classify packets based on packet characteristics and specify the same service for packets of the same type. This allows different types of packets to be provided differentiated services.
- MQC entities:
  - MQC involves three entities: traffic classifier, traffic behavior, and traffic policy.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Traffic Classifier

- Layer 2 traffic classification rules
  - Source MAC address, destination MAC address, VLAN ID, MPLS EXP, Layer 2 protocol-related field, and ACL 4000–4999 rule-related field
- Layer 3 traffic classification rules
  - IP DSCP, IP precedence, IP protocol type, RTP port number, ACL 2000–3999 rule-related field, and ACL6 2000–3999 rule-related field

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A traffic classifier defines a group of matching rules to classify packets.

Traffic Classifier Configuration

- **traffic classifier** *classifier-name* [**operator** { **and** | **or** }]
  - The **and** parameter indicates that the relationship between rules in a traffic classifier is AND, which means:
    - If a traffic classifier contains ACL rules, packets match the traffic classifier only when the packets match one ACL rule and all the non-ACL rules.
    - If a traffic classifier does not contain any ACL rules, packets match the traffic classifier only when they match all the rules in the classifier.
  - The **or** parameter indicates that the relationship between rules in a traffic classifier is OR. That is, packets match the traffic classifier if the packets match one or more rules in the traffic classifier.

  By default, the relationship between rules in a traffic classifier is OR.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- **traffic classifier** *classifier-name* [**operator** { **and** | **or** }]
  - ***classifier-name:*** Specifies the name of a traffic classifier.
  - **operator**: Indicates the relationship between rules in a traffic classifier. If this parameter is not specified, the relationship between rules is OR by default.
  - **and**: Indicates that the relationship between rules is AND.
  - **or**: Indicates that the relationship between rules is OR. After this parameter is specified, packets match a traffic classifier if the packets match one or more rules.

Traffic Behavior Configuration

- MQC traffic behaviors include:
  - Packet filtering
  - Priority re-marking
  - Redirection
  - Traffic policing
  - Traffic statistics collection
- You can run the **traffic behavior** *behavior-name* command to create a traffic behavior.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

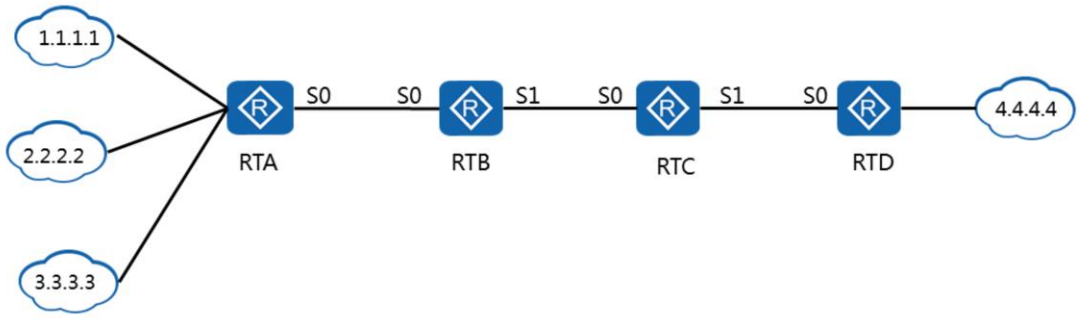
Traffic Policy Configuration

- Traffic classifiers and traffic behaviors can be associated in a traffic policy.
  - Run the **traffic policy** *policy-name* command to create a traffic policy and enter the traffic policy view.
  - Run the **classifier** *classifier-name* **behavior** *behavior-name* [**precedence** *precedence-value*] command to associate a traffic classifier and a traffic behavior.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

MQC Configuration Example



- On RTA, classify traffic into three types: AF11, AF21, and EF.
- On RTB, re-mark AF21 traffic as AF22 traffic.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- This is a class-based QoS configuration example. Traffic classification is performed on RTA, and policies, such as rate limiting and priority re-marking, are implemented on RTB.

MQC Configuration on RTA

```
acl number 2001
 rule 0 permit source 1.1.1.1 0
acl number 2002
 rule 0 permit source 2.2.2.2 0
acl number 2003
 rule 0 permit source 3.3.3.3 0

traffic classifier Class_af11 operator and
 if-match acl 2001
traffic classifier Class_af21 operator and
 if-match acl 2002
traffic classifier Class_ef operator and
if-match acl 2003                          //Traffic classifier

traffic behavior Behavior_af11
 remark dscp af11
traffic behavior Behavior_af21
 remark dscp af21
traffic behavior Behavior_ef
remark dscp ef                             //Traffic behavior

traffic policy Mark
 classifier Class_af11 behavior Behavior_af11
 classifier Class_af21 behavior Behavior_af21
classifier Class_ef behavior Behavior_ef    //Traffic policy

interface Serial0
 ip address 12.12.12.1 255.255.255.252
 traffic-policy Mark outbound
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Traffic classification is performed on RTA so that traffic is marked as AF11, AF21, and EF traffic based on the source address.

MQC Configuration on RTB

```
traffic classifier Class_af21 operator and
if-match dscp af21                                  //Traffic classifier

traffic behavior Behavior_af21
remark dscp af22                                    //Traffic behavior

traffic policy dscp
classifier Class_af21 behavior Behavior_af21        //Traffic policy

interface Serial0
 ip address 23.23.23.1 255.255.255.252
 traffic-policy dscp inbound
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Different QoS policies are implemented for traffic that is marked differently on RTB.

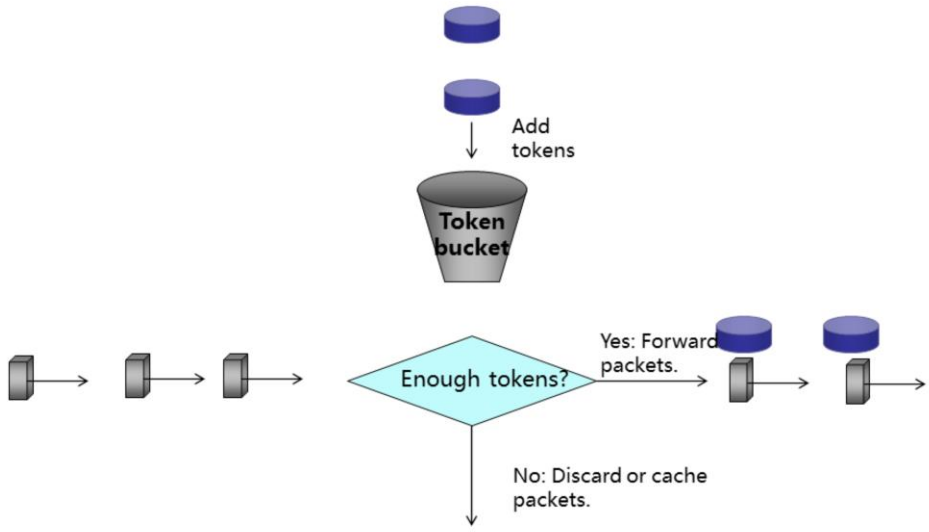


Contents

- QoS MQC Configuration
- **QoS Token Bucket Algorithm**
- QoS Application

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Token Bucket



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- To implement traffic control, a mechanism that measures the traffic passing through a device is required. A token bucket is a commonly used mechanism that measures such traffic.
- When packets reach a device, the device obtains enough tokens from the token bucket for packet transmission. If the token bucket does not have enough tokens to send a packet, the packet waits for enough tokens or is discarded. This feature limits packets to be sent at a rate less than or equal to the rate at which tokens are generated.

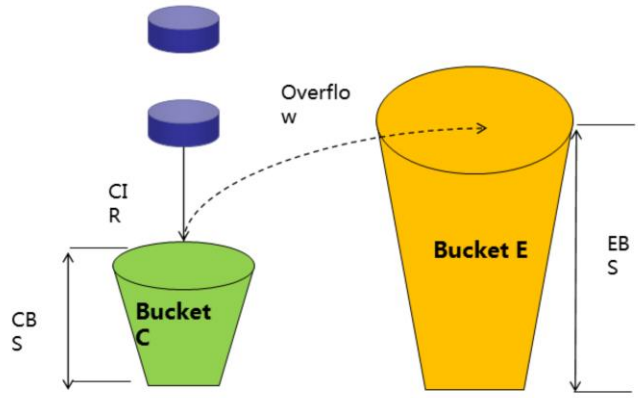
Single-Rate Traffic Policing Parameters

- The following parameters are involved in single-rate traffic policing:
  - Committed information rate (CIR)
  - Committed burst size (CBS)
  - Extended burst size (EBS)

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Single-Rate Traffic Policing Mechanism

- Two token buckets, buckets C and E, are used.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A Huawei router uses two token buckets for single-rate traffic policing.
- Two token buckets, buckets C and E, are used. The capacity of bucket C is the CBS, and the capacity of bucket E is the EBS. Therefore, the total capacity of the two token buckets is the CBS plus EBS. To prevent burst traffic, users can set the EBS to 0.
- When the EBS is not 0, two token buckets are used for single-rate traffic policing. When the EBS is 0, no token is added in bucket E. Therefore, only bucket C is used for single-rate traffic policing. When only bucket C is used, packets are marked either green or red.
- What are CIR, CBS, and EBS?
  - CIR: indicates the rate at which an interface allows packets to pass through, also the rate at which tokens are put into a token bucket. The CIR is expressed in kbit/s.
  - CBS: indicates the committed volume of traffic that an interface allows to pass through, also the depth of a token bucket. The CBS is expressed in bytes. The CBS must be greater than or equal to the size of the largest possible packet entering a device. Note that sometimes a single packet can consume all the tokens in the token bucket. The larger the CBS is, the greater the traffic burst can be.
  - EBS: indicates the maximum volume of burst traffic before the rate of all traffic exceeds the CIR.

Rules for Single-Rate Traffic Policing

- Color-Blind mode
  - Single bucket at a single rate
  - Dual buckets at a single rate
- Color-Aware mode
  - Single bucket at a single rate
  - Dual buckets at a single rate

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- **Method of Adding Tokens for Single-Rate Traffic Policing**
  - In single-rate traffic policing, both buckets C and E are full of tokens at the beginning. Tokens are put into bucket C and then bucket E, for possible burst traffic whose traffic rate exceeds the CIR, after bucket C is full of tokens. After both buckets C and E are filled with tokens, subsequent tokens are dropped.
- **Rules for Single-Rate Traffic Policing**
  - When a packet arrives at an interface, the length of the packet is compared with the number of tokens in the token buckets (one token is generally required for one bit). If the number of tokens is less than the length of the packet, the packet is dropped or buffered.
  - Tc and Te refer to the numbers of tokens in buckets C and E, respectively. The initial values of Tc and Te are respectively the CBS and EBS.
  - In Color-Blind mode, the following rules apply when a packet of size B arrives at time t:
  - When a token bucket is used for single-rate traffic policing:
    - If $Tc(t) - B \geq 0$, the packet is marked green, and Tc is decremented by B.
    - If $Tc(t) - B < 0$, the packet is marked red, and Tc remains unchanged.
  - When two token buckets are used for single-rate traffic policing:
    - If $Tc(t) - B \geq 0$, the packet is marked green, and Tc is decremented by B.
    - If $Tc(t) - B < 0$ but $Te(t) - B \geq 0$, the packet is marked yellow, and Te is decremented by B.
    - If $Te(t) - B < 0$, the packet is marked red, and neither Tc nor Te is decremented.

- In Color-Aware mode, the following rules apply when a packet of size B arrives at time t:
- When a token bucket is used for single-rate traffic policing:
  - If the packet has been marked green and Tc(t) – B ≥ 0, the packet is re-marked green, and Tc is decremented by B.
  - If the packet has been marked green and Tc(t) – B < 0, the packet is re-marked red, and Tc remains unchanged.
  - If the packet has been marked yellow or red, the packet is re-marked red regardless of the packet length. The Tc value remains unchanged.
- When two token buckets are used for single-rate traffic policing:
  - If the packet has been marked green and Tc(t) – B ≥ 0, the packet is re-marked green, and Tc is decremented by B.
  - If the packet has been marked green and Tc(t) – B < 0 but Te(t) – B ≥ 0, the packet is marked yellow, and Te is decremented by B.
  - If the packet has been marked yellow and Te(t) – B ≥ 0, the packet is re-marked yellow, and Te is decremented by B.
  - If the packet has been marked yellow and Te(t) – B < 0, the packet is re-marked red, and Te remains unchanged.
  - If the packet has been marked red, the packet is re-marked red regardless of the packet length. The Tc and Te values remain unchanged.

Parameters for Two-Rate Traffic Policing

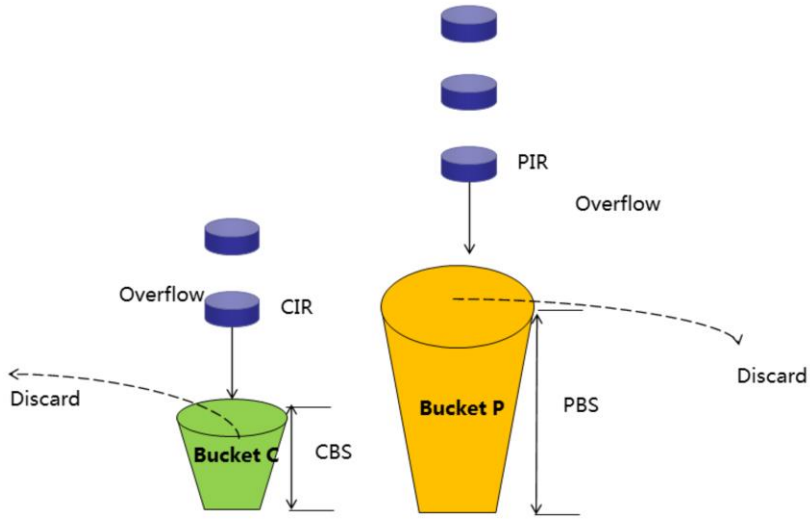
- Committed information rate (CIR)
- Committed burst size (CBS)
- Peak information rate (PIR)
- Peak burst size (PBS)

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- CIR: indicates the rate at which an interface allows packets to pass through, also the rate at which tokens are put into a token bucket. The CIR is expressed in kbit/s.
- CBS: indicates the committed volume of traffic that an interface allows to pass through, also the depth of a token bucket. The CBS is expressed in bytes. The CBS must be greater than or equal to the size of the largest possible packet entering a device.
- PIR: indicates the maximum rate at which an interface allows packets to pass and is expressed in kbit/s. The PIR must be greater than or equal to the CIR.
- PBS: indicates the maximum volume of traffic that an interface allows to pass through in a traffic burst.

Two-Rate Traffic Policing Mechanism



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The two rate three color marker uses two token buckets and focuses on the burst traffic rate. The single rate three color marker puts excess tokens beyond the capacity of the first token bucket into the second bucket, whereas the two rate three color marker uses two token buckets that separately store tokens. Therefore, the two rate three color marker has two rates at which tokens are put into token buckets. These two token buckets are called buckets C and P. The capacity of bucket C is the CBS, and the capacity of bucket P is the PBS. Tokens are put into bucket C at the rate of CIR and into bucket P at the rate of PIR.
  - "Two rate" in the two rate three color markers refers to the two rates at which tokens are put into the two token buckets.
- **Method of Adding Tokens for Two-Rate Traffic Policing**
  - Buckets C and P are full of tokens at the beginning. Tokens are put into buckets C and P at the rate of CIR and PIR, respectively. Buckets C and P work separately. When one bucket is full of tokens, any subsequent tokens for the bucket are dropped, but tokens continue being put into the other bucket if it is not full.
- **Rules for Two-Rate Traffic Policing**
  - The two rate three color marker focuses on the traffic burst rate and checks whether the traffic rate is conforming to the specifications. Therefore, traffic is measured based on bucket P and then bucket C.
  - The two rate three color marker works in either Color-Blind or Color-Aware mode. Tc and Tp refer to the numbers of tokens in buckets C and P, respectively. The initial values of Tc and Tp are respectively the CBS and PBS.
  - In Color-Blind mode, the following rules apply when a packet of size B arrives at time t:
    - If $Tp(t) - B < 0$, the packet is marked red, and The Tc and Tp values remain

unchanged.

        - If Tp(t) – B ≥ 0 but Tc(t) – B < 0, the packet is marked yellow, and Tp is decremented by B.
        - If Tc(t) – B ≥ 0, the packet is marked green and both Tp and Tc are decremented by B.
    - In Color-Aware mode, the following rules apply when a packet of size B arrives at time t:
        - If the packet has been marked green and Tp(t) – B < 0, the packet is re-marked red, and neither Tp nor Tc is decremented.
        - If the packet has been marked green and Tp(t) – B ≥ 0 but Tc(t) – B < 0, the packet is re-marked yellow, and Tp is decremented by B, and Tc remains unchanged.
        - If the packet has been marked green and Tc(t) – B ≥ 0, the packet is re-marked green, and both Tp and Tc are decremented by B.
        - If theIf the packet has been marked yellow and Tp(t) – B < 0, the packet is re-marked red, and neither Tp nor Tc is decremented.
        - packet has been marked yellow and Tp(t) – B ≥ 0, the packet is re-marked yellow, and Tp is decremented by B and Tc remains unchanged.
        - If the packet has been marked red, the packet is re-marked red regardless of what the packet length is. The Tp and Tc values remain unchanged.

- What are CIR, CBS, and EBS?
    - **cir** *cir-value* specifies the committed rate of traffic that an interface allows to pass. The value is an integer ranging from 0 to 4294967295, in kbit/s.
    - **pir** *pir-value* specifies the peak rate of traffic that an interface allows to pass. The value is an integer ranging from 0 to 4294967295, in kbit/s. *pir-value* must be greater than or equal to the configured *cir-value*.
    - **cbs** *cbs-value* specifies the committed volume of traffic that an interface allows to pass and the depth of the first bucket (assuming it is bucket C). The value is an integer ranging from 0 to 4294967295, in bytes. The CBS value must be greater than the CIR value. The default value varies with the value of *cir-value*.
        - **pbs** *pbs-value* specifies the peak volume of traffic that an interface allows to pass and the depth of the second token bucket (assuming it is bucket P). The value is an integer ranging from 0 to 4294967295, in bytes. The default value varies with the value of *pir-value*.



Contents

- QoS MQC Configuration
- QoS Token Bucket Algorithm
- **QoS Application**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

QoS Application on the Live Network — Planning and Deployment

- In principle, traffic classification, traffic re-marking, and rate limiting are implemented on the inbound user-side interface, and traffic shaping is implemented on the outbound user-side interface (if packets of various levels are involved, queue scheduling and congestion avoidance must be configured on the outbound user-side interface). Queue scheduling and congestion avoidance are configured on the outbound network-side interface.
- Configuring rate limiting for priority queues (PQs) based on service requirements is recommended.
- Traffic classification (multi-field classification) is recommended at the network edge (close to users and service sources) to simplify subsequent traffic classification. Behavior aggregate classification is recommended on the equipment at the aggregation or upper layers to reduce equipment loads.
- In principle, packets of lower priority services are dropped more and earlier than those of higher priority services. Packets of real-time services are processed using the tail drop policy. Signaling protocol packets are not dropped.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

QoS Application on the Live Network — Service Priority Planning

- Management control messages: include important packets such as the protocol packets used to communicate with the NMS and other devices, and are crucially important for the running network. Such packets must be allocated the highest priority.
- Voice streams: require low delay, low jitter, and low packet loss rate. Such packets must be placed in the EF queue.
- Real-time video streams (such as video conferences and video calls): require low delay, low jitter, and low packet loss rate. Such packets must be placed in the EF queue.
- Office services: have low requirements on delay and jitter and are generally placed in AF and BE queues.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

QoS Application on the Live Network — Traffic Policing and Traffic Shaping

- Traffic policing must be implemented for the traffic sent from some NMS servers, which are vulnerable to virus infection, to avoid attacks on voice and video services. Traffic sent from the NMS servers has the highest priority and is scheduled using PQ.
- To prevent some office service from launching DDoS attacks to affect other services due to virus infection or bandwidth overuse, configure traffic policing (CAR) for each access flow based on SLA. Set the priority of the excess traffic to the lowest level so that the excess traffic is forwarded in best effort mode instead of being directly discarded. This configuration helps to leverage network resources and improve customer satisfaction.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

QoS Application on the Live Network — Queue Scheduling and Packet Drop Policy

- Enable behavior aggregate classification for incoming packets on edge routers on the backbone network so that queue scheduling is performed based on the priorities carried by the packets on non-edge routers on the backbone network. Mappings between IP DSCP/precedence and MPLS EXP values are generally configured on edge routers on the MPLS network connected to the IP network. Generally, management control messages and packets of voice calls, video conferences, and video calls are of high priority and are forwarded preferentially. Packets of other services can be scheduled using WFQ.
- Tail drop is used for PQ services, such as voice and video calls, in an effort to forward PQ packets without dropping them. Packets are lost only when PQ resources are exhausted. To prevent global TCP synchronization, WRED is used to schedule WFQ services. Set a high packet drop threshold for high-priority traffic and set a low packet drop threshold for low-priority traffic so that low-priority traffic faces a higher probability of being dropped.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— SDN VXLAN

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- VXLAN is a very important overlay technology. It is widely used in SDN scenarios, such as DC scenarios with the integrated cloud and network and CloudVPN overlay scenarios.
- This material helps you understand motivations behind VXLAN emergence and basic VXLAN concepts.
- By analyzing the VXLAN network traffic, you can understand how services are provisioned from end to end on an SDN DCN network.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to understand:
  - Basic SDN concepts
  - Why VXLAN is required on a DC network
  - Basic VXLAN concepts
  - How VXLAN packets are processed
  - VXLAN applications

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



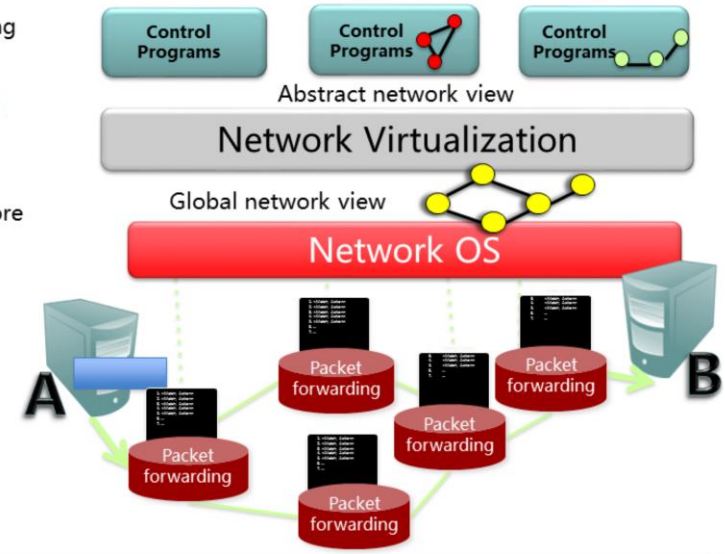
Contents

- **SDN Overview**
- Cloudification Challenges Facing Traditional DC Networks
- Basic VXLAN Concepts
- VXLAN Applications
- VXLAN Concepts and Applications in Cloud DCs

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

SDN Definition

Software-defined networking (SDN) is a brand-new network architecture that separates the control plane from the data plane for flexible network traffic control. It provides an outstanding platform for core network and application innovations.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- SDN was born on campus networks in 2006. 2012 is considered the first year of SDN commercial use. In 2012, significant events, such as Google's deployment of SDN, pushed SDN to the spotlight, and then SDN was extended to telecom networks.
- The following describes the major events involved in the SDN development (you only need to know the key points).

- In 2006, SDN was born in the Clean Slate Program of Stanford University funded by the U.S GENI project. Led by Professor Nick McKeown of Stanford University, the research team proposed the OpenFlow concept for experimental innovation on campus networks. Later, based on OpenFlow's programbility characteristics brought to networks, the concept of SDN emerges. The ultimate goal of the Clean Slate Program is to reinvent the Internet, aiming at changing the existing network infrastructure that is slightly outdated and difficult to evolve.
- In 2007, Stanford student Martin Casado led a project Ethane on network security and management. The project attempts a centralized controller, which allows network administrators to easily define security control policies based on network flows and to apply these security policies to various network devices, thereby implementing security control over the entire network communication.
- In 2008, inspired by the Ethane project and its predecessor project Sane, Professor Nick McKeown and others proposed the concept of OpenFlow. In the paper entitled "OpenFlow: Enabling Innovation in Campus Networks" published in ACM SIGCOMM, Nick McKeown introduced in detail the concept of OpenFlow for the first time. In addition to describing how OpenFlow works, this paper lists several application scenarios

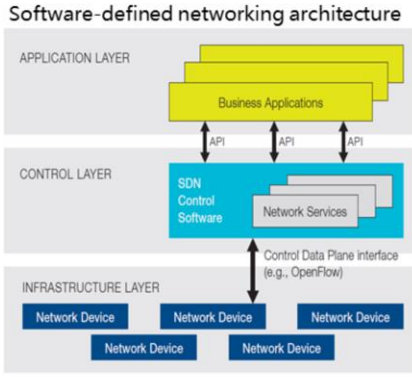
of OpenFlow.

- Based on the programmability that OpenFlow brings for networks, Nick McKeown and his team further proposed the concept of software defined network (SDN) (currently more literally translated as "software-defined networking"). In 2009, SDN was shortlisted as one of the top ten frontier technologies by Technology Review. This concept was then widely recognized and supported by the academic and industrial sectors.
- In December 2009, OpenFlow 1.0, a milestone version that can be used on commercial products, was released. Along with this, the plug-in for obtaining OpenFlow packet headers on the Wireshark, OpenFlow debugging tool (liboftrace), OpenFlow virtual computer emulation (OpenFlowVMS), and more OpenFlow functions gradually became mature. OpenFlow versions 1.1, 1.2, 1.3, and 1.4 have been released so far. The current version of OpenFlow is 1.5.1.

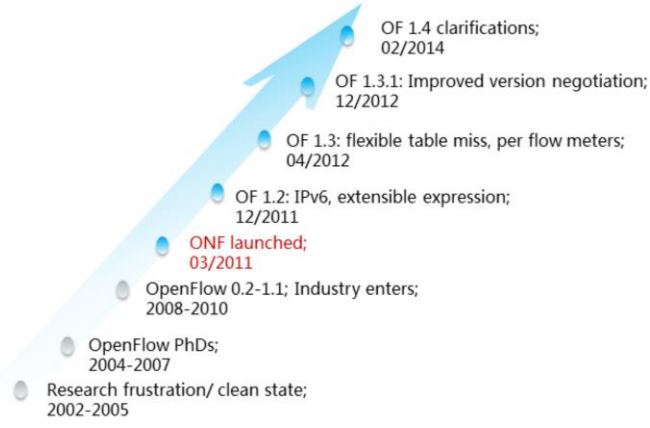
- In March 2011, with the help of Professor Nick Mckeown, the Open Network Foundation (ONF) was established to promote the standardization and development of SDN architecture and technologies. The ONF has 96 ONF members, including the seven founders: Google, Facebook, NTT, Verizon, Deutsche Telekom, Microsoft, and Yahoo.
- In May 2011, NEC launched the first commercial OpenFlow switch.
- In April 2012, Google announced that its backbone network had been fully running on OpenFlow and connected to 12 DCs across the globe through 10 Gbit/s networks, improving WAN link utilization from 30% to nearly 100%.
- This proved that OpenFlow is no longer just a research model in academy, but is technologically ready for the commercial use.
- In July 2012, Nicira, a company focused on SDN and network virtualization, was acquired by VMware for $1.26 billion. Nicira is a startup that subverts DCs. It creates a network virtual platform (NVP) based on OpenFlow. OpenFlow is an open source project created by Martin Casado during his pursing of PHD in Stanford. He co-founded Nicira with his two Stanford University professors, Nick McKeown and Scott Shenker. VMware's acquisition has transformed Casado's technology research of over a decade from paper into reality. Network software is stripped from hardware servers, which is also the first step for SDN going to market.
- At the end of 2012, AT&T, BT, Deutsche Telekom, Orange, Italy Telecom, Spain Telecom, and Verizon jointly launched the Network Functions Virtualization (NFV) Industry Alliance, aiming to introduce SDN to the telecom industry. The alliance consists of 52 network operators, telecom equipment suppliers, IT equipment suppliers, and technology suppliers.
- In April 2013, Cisco and IBM jointly established Open Daylight with Microsoft, Big Switch, Brocade, Citrix, Dell, Ericsson, Fujitsu, Intel, Juniper Networks, Microsoft, NEC, HP, Red Hat, and VMware. In cooperation with the Linux Foundation, the organization developed SDN controllers, southbound APIs, and northbound APIs, aiming to break the monopoly of large vendors on network hardware, drive network technology innovation, and make network management easier and cheaper. In this organization, there are only SDN vendors but no SDN users (Internet users or carriers). The Open Daylight project covers SDN controller development and API proprietary extension, and announced to launch an industrial-grade open source SDN controller.
- More background knowledge:
- Clean Slate Program
    - Pain points: Constantly patching the existing network architecture is difficult to solve the fundamental problems. Redefining the network architecture may be the ultimate solution.
    - The ultimate goal of the Clean Slate Program is to reinvent the Internet, aiming at changing the existing network infrastructure that is slightly outdated and difficult to evolve.
    - Clean Slate Program in a broad sense and narrow sense:
        - Broad sense: Refers to various next-generation network (NGN) projects.
        - Narrow sense: Lab research plan led by Professor Nick McKeown, Stanford University (birth place of SDN)

- Ethane project (sub-subject of the Clean Slate Program)
  - Ethane is a project on network security and management led by Stanford student Martin Casado. The project attempts a centralized controller, which allows network administrators to easily define security control policies based on network flows and to apply these security policies to various network devices, thereby implementing security control over the entire network communication.
  - Inspired by this project, Martin and his mentor, Nick McKeown, proposed the concept of OpenFlow.

ONF-Defined SDN Architecture



Source: ONF white paper, April 13, 2012

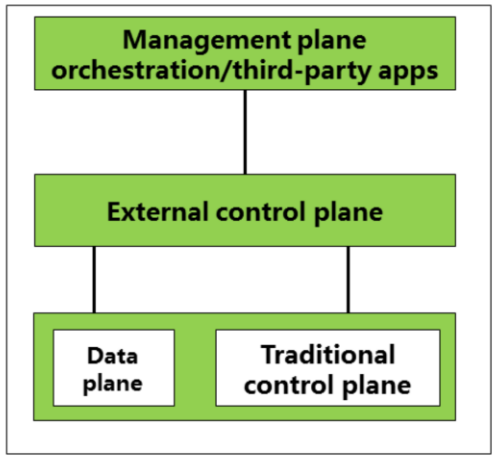


ONF emphasizes OpenFlow-based SDN and the separation of control and forwarding planes to standardize the forwarding equipment. The key relies in OpenFlow protocol standardization.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

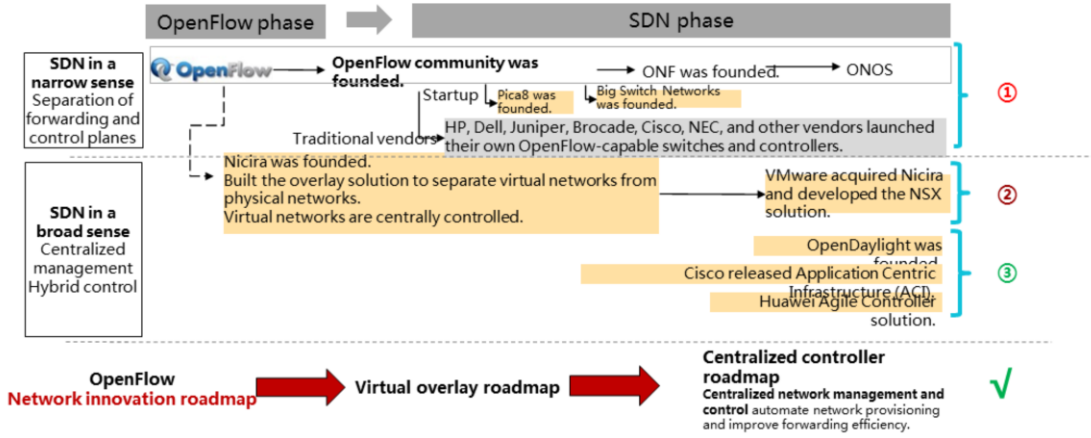
SDN Architecture in the Eyes of Traditional CT Vendors



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Major SDN Technical Evolution Roadmaps

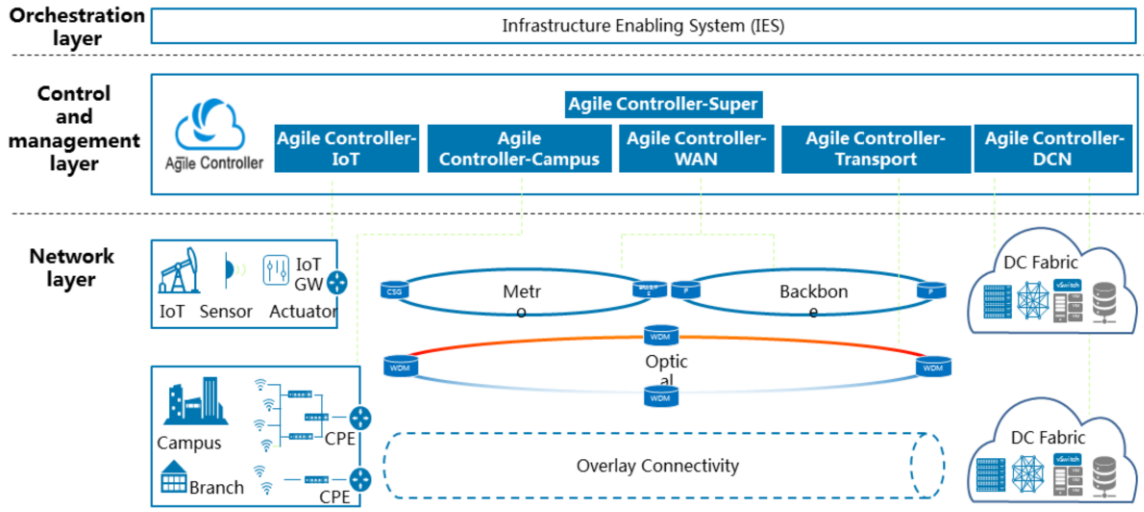
- In 2006, Stanford University released OpenFlow. This protocol separates network devices' forwarding plane from the control plane, and the control plane can then flexibly control network traffic.
- In addition to the moderate separation of control and forwarding planes, the core feature of Huawei DC SDN is separating management and control planes to enable automated provisioning of network services, facilitating



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

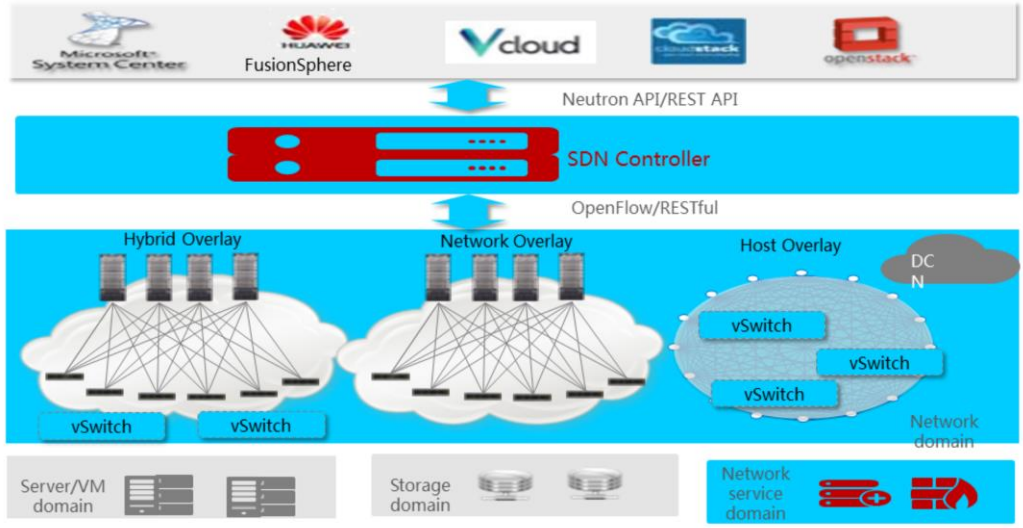
Huawei SDN Solution



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

SDN DCN Solution



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

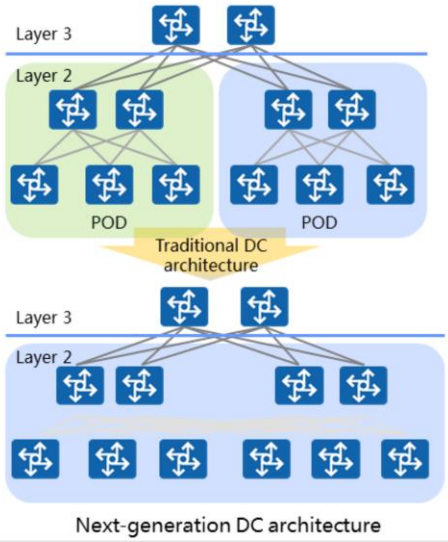


Contents

1. SDN Overview
2. **Cloudification Challenges Facing Traditional DC Networks**
3. Basic VXLAN Concepts
4. VXLAN Applications
5. VXLAN Concepts and Applications in Cloud DCs

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

DC Development Trend (1)



Next-generation DC architecture

Traditional DC architecture

- On traditional DC networks, Layer 2 only extends to access or aggregation switches, and virtual machines (VMs) can only be migrated within a Layer 2 domain. If VMs need to be migrated to another Layer 2 domain, their IP addresses must be changed, and services are interrupted during VM migration.

Next-generation DC architecture

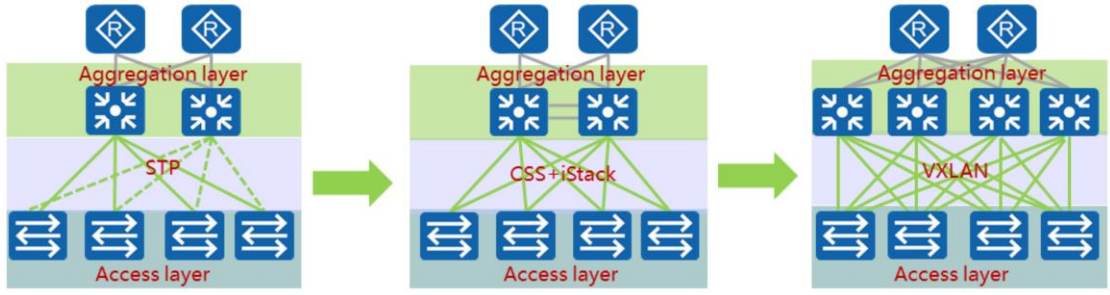
- In the cloud computing era, IDC carriers require VM migration within a larger domain to better utilize DC resources.
- Because a large amount of east-west traffic needs to be transmitted between servers, non-blocking forwarding is required to fully utilize network link resources.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The VM scale is limited by network specifications.
  - On a large-scale Layer 2 network, data packets are forwarded at Layer 2 based on MAC entries. However, the MAC table capacity is limited, which subsequently limits the number of VMs.

DC Development Trend (2)

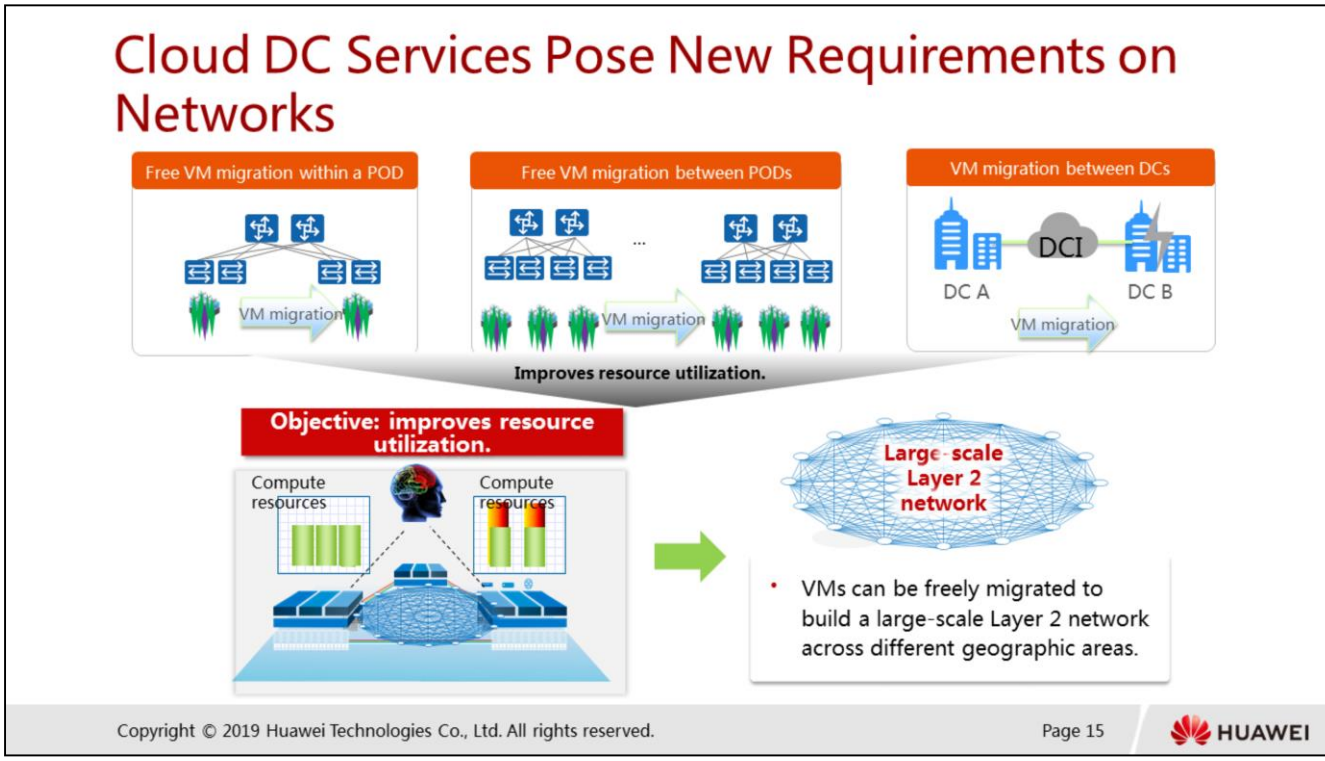


- Traditional Layer 2 technologies such as STP and CSS+iStack are unsuitable for building a large-scale Layer 2 network. In contrast, VXLAN can be used to construct a large-scale Layer 2 network and supports flat fat-tree network topology with high bandwidth utilization.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

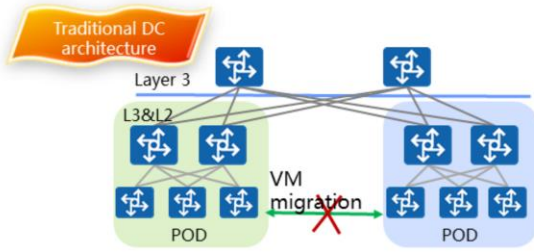
HUAWEI

- Network isolation capability limitations:
  - Currently, the mainstream network isolation technology is VLAN or VPN. The VLAN Tag field defined in IEEE 802.1Q has only 12 bits and can represent only up to 4096 VLANs, which cannot meet the requirement of identifying numerous user groups on a large-scale Layer 2 network.
  - VLAN or VPN on traditional Layer 2 networks does not support dynamic network adjustment.

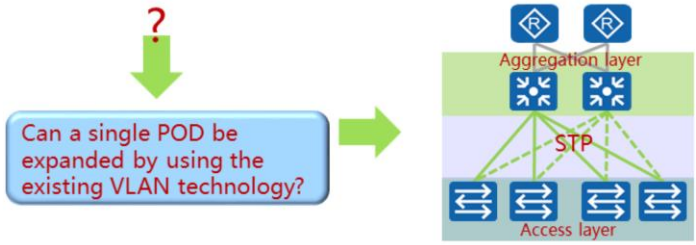


- Google is the practitioner of large-scale server clusters. A large amount of communication between servers requires non-blocking networks.
- The number of Google servers in a single cluster has reached 10,000.
- Internet vendors in China are planning 20,000 servers in a single cluster.
- The number and capacity of network interfaces are the core factors that determine the cluster scale.
- A cross-region server cluster improves the system disaster recovery capability.
- Large-scale Layer 2 network:
  - Requires the non-blocking forwarding capability for east-west network traffic.

Why Cannot Traditional Networks Be Large?



- **A VLAN cannot cross the Layer 3 boundary.** Therefore, in traditional DC networking, gateways are deployed on aggregation switches, which communicate with each other through Layer 3 core switches. VMs can be migrated only within single PODs. If VMs need to be migrated to another POD within the Layer 2 domain, their IP addresses must be changed, and services will be interrupted during VM migration.



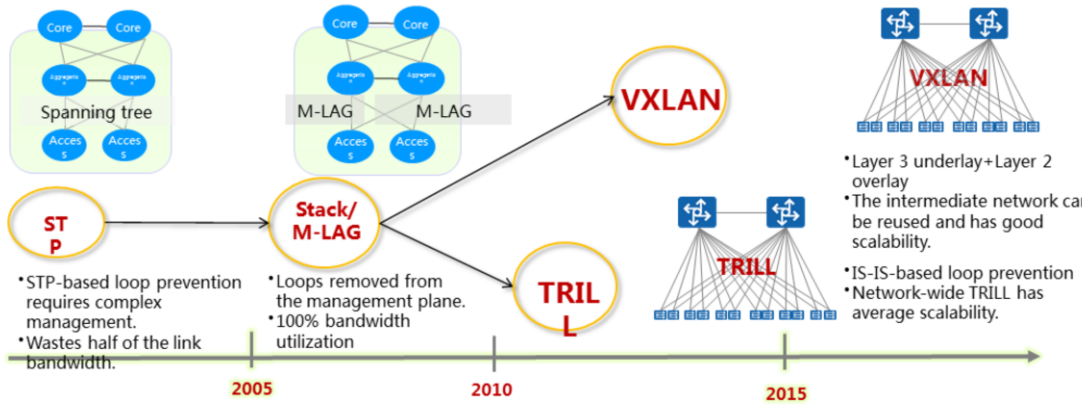
The STP technology has the following defects while solving the network loop problem:

- STP convergence takes a long time. It applies only to a network with less than 50 network nodes and is not suitable for large-scale cloud DC networks.
- When STP is used to construct a loop-free network, half of the links need to be blocked, which lowers bandwidth utilization.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

DC Network Architecture Development Trends



- TRILL solves the STP problems. Through the mature link-state routing algorithm, TRILL extends the IS-IS protocol and can construct a loop-free network for multi-path load balancing.
- Virtual Extensible LAN (VXLAN) provides better availability, scalability, and easy O&M, and therefore is preferred by IT&CT vendors.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IT+CT form a joint force.
- TRILL is a revolutionary technology. VXLAN is an improved technology.

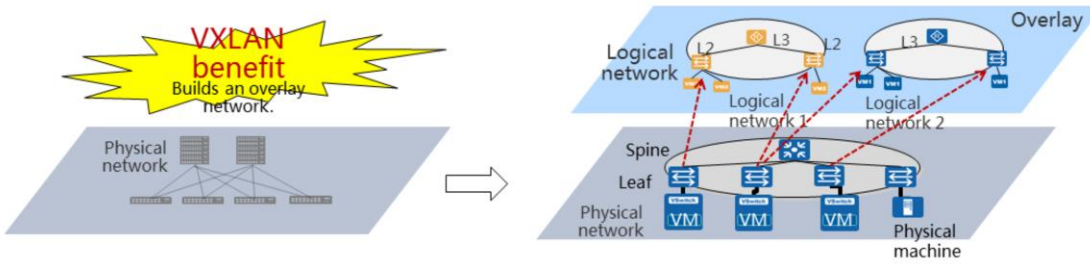
VXLAN — De Facto Standard of the Overlay Technology in the Industry

Requirements of High-End Cloud DC Networks		VXLAN	CSS/SVF	TRILL
> 4K tenants	Leased DC, must be able to support a large number of tenants.	**16M**	4K	4K (The latest standard allows the number to be upgraded to 16M.)
Protecting existing network investment	A new fabric can be built based on the existing network.	**The existing network only needs to support common Layer 3 features.**	New construction is required.	New construction is required.
SDN capability	Can be smoothly upgraded to an SDN network.	**Overlay is an important roadmap of SDN.**	Not supported	Not supported
Standard protocol	Allows different vendors to communicate with each other.	**Standard**	Vendor-specific	Standard
Cross-DC capability	A large-scale Layer 2 network can be constructed across an IP WAN.	**Supported**	Not supported	Not supported

- VXLAN has stronger capabilities in supporting SDN and multi-tenant. Therefore, VXLAN has become a hot technology in the industry.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

VXLAN Benefits



VXLAN concepts:

- An overlay network builds a compute network on another network.
- The core is to encapsulate traffic to decouple network services from underlying infrastructure.
- VXLAN uses the encapsulation technology.
- VXLAN tunnel encapsulation endpoints are called network virtualization edges (NVEs), which are responsible for VXLAN encapsulation and decapsulation of original Ethernet packets.

VXLAN advantages:

- VXLAN is a standard overlay technology in the industry.
- Compared with STP, which uses active and standby paths for traffic forwarding, VXLAN uses ECMP of the underlay network to improve network forwarding performance.
- VXLAN is constructed based on UDP and is an overlay technology that unifies IT and CT.
- VXLAN expands the maximum number of supported subnets from 4K to 16 million and supports multi-tenant.
- VXLAN provides a forwarding plane for SDN.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Physical network
  - A physical network has high bandwidth and large capacity.
  - The large-scale Layer 2 network requires STP to solve the loop problem.
  - Supports isolation of only up to 4K VLANs.
  - VM migration is not flexible and requires change of the physical network configurations.

- Overlay network
  - An overlay network has better mobility because VNIs are separated from geographical locations, meeting the elasticity requirement of Layer 2 networks.
  - An overlay network supports on-demand service deployment. When services change, the underlay network does not need to be changed.
  - An overlay network is decoupled from a physical network.



Contents

- SDN Overview
- Cloudification Challenges Facing Traditional DC Networks
- **Basic VXLAN Concepts**
- VXLAN Packet Forwarding Process
- VXLAN Concepts and Applications in Cloud DCs

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

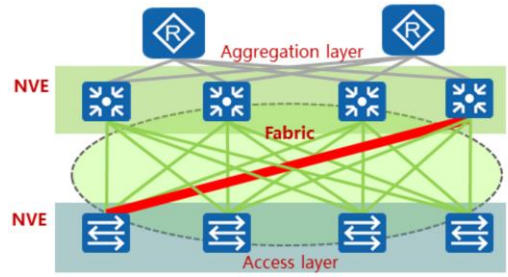
Basic VXLAN Concepts

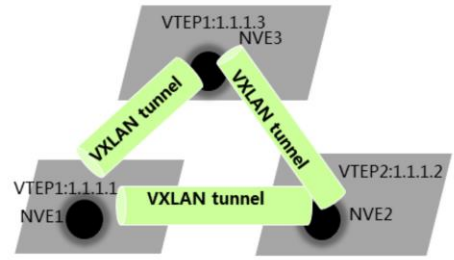
Layer 2 fabric networking based on NVo3

Network virtualization over Layer 3 (NVO3) technologies build virtual networks based on Layer 3 IP overlay networks. Typical NVO3 technologies include VXLAN, Network Virtualization using Generic Routing Encapsulation (NVGRE), and Stateless Transport Tunneling (STT).
An NVO3-running device is called a network virtualization edge (NVE). An NVE is at the border of an overlay network and implements Layer 2 and Layer 3 virtualization functions.

VXLAN is the most influential NVO3 technology. It uses the MAC-in-UDP packet encapsulation mode to build IP overlay-based virtual LANs.

- VTEPs are used to identify NVEs on a VXLAN network.
- Each NVE has at least one VTEP that is identified by the NVE's IP address.
- Two VTEPs can determine a VXLAN tunnel.



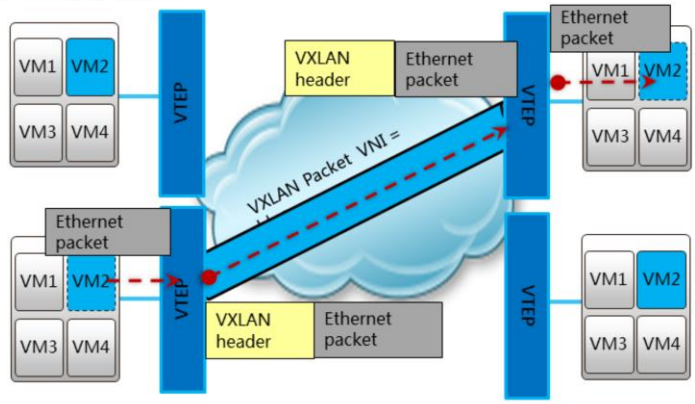


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

VXLAN Concept — VTEP

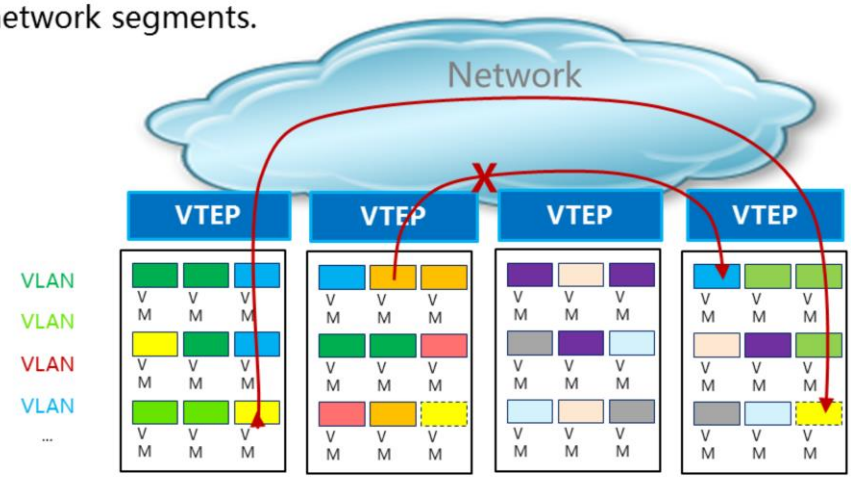
- VTEPs are used to identify NVEs on a VXLAN network.
- Each NVE has at least one VTEP that is identified by the NVE's IP address.
- Two VTEPs determine a VXLAN tunnel, which is shared by all VXLAN network identifiers (VNIs) between the two NVEs.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

VXLAN — VNI

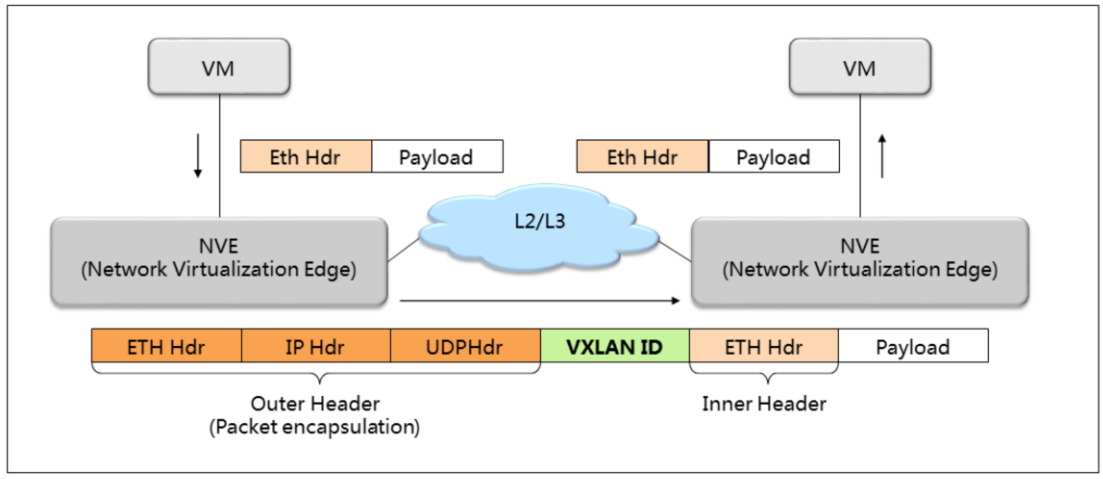
- A VNI (24 bits) identifies a virtual network. The VNI field supports up to 16M network segments.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

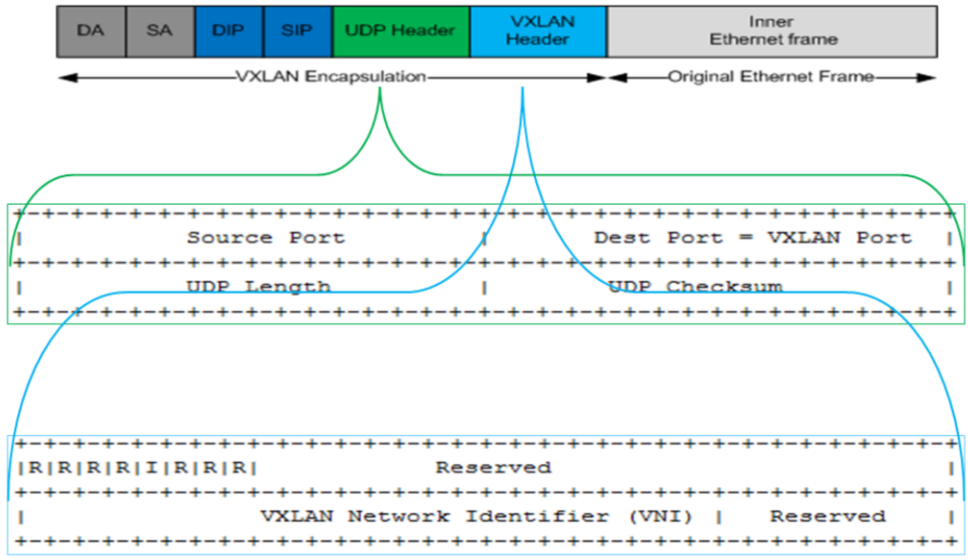
VXLAN Packet Format (1)



VXLAN packet encapsulation process

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

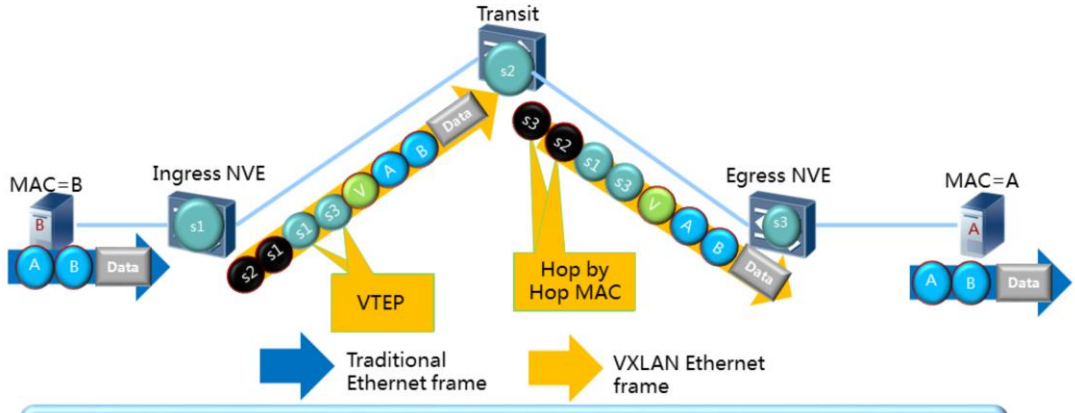
VXLAN Packet Format (2)



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

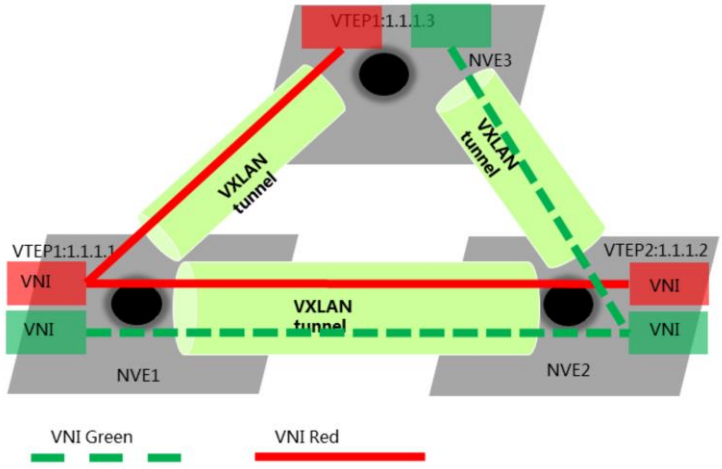
VXLAN Forwarding Data Encapsulation



Layer 2 packets from a source host can arrive at the destination host over the IP network. The VXLAN network functions as a bridge fabric for the hosts.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Relationship Between Tunnels and VNIs



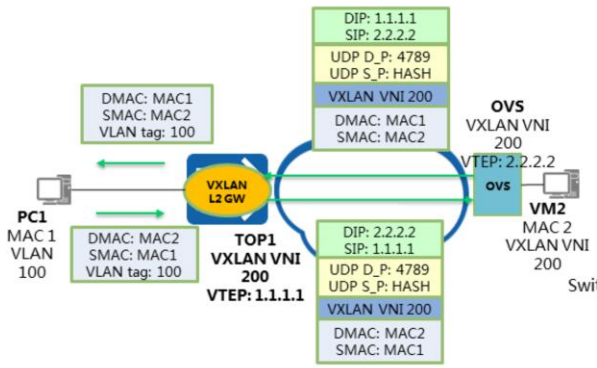
VNI Concept

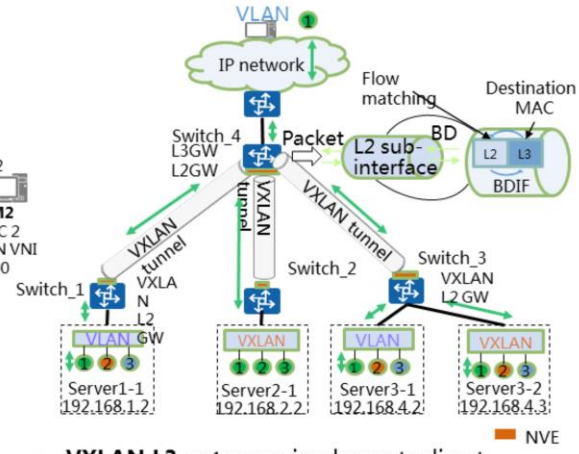
- A VNI identifies a Layer 2 domain on a VXLAN network.
- Two VTEPs determine a VXLAN tunnel, which is shared by all VNIs between the two NVEs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

VXLAN Gateways





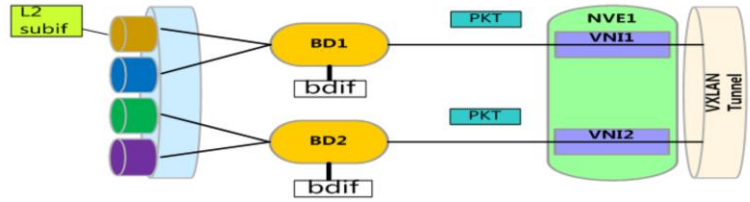
- **VXLAN L2 gateway:** allows tenants to access a VXLAN network to implement intra-VXLAN traffic exchange.
- **VXLAN L3 gateway:** implements direct traffic exchange between different VXLANs or between VXLANs and non-VXLANs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

VXLAN Access Service Model (1)

- A VXLAN gateway uses the EVC service model. The model components include the bridge domain (BD), VNI, NVE, Layer 2 sub-interface (L2 subif), and VXLAN tunnel.
  - L2 subif: used for user access. A Layer 2 sub-interface can be configured with one VLAN tag or no tag for access.
  - BD: identifies a Layer 2 broadcast domain and maps to a VNI in 1:1. All broadcast domains are supported based on BDs, such as MAC address learning, Layer 2 table lookup, and broadcast replication.
  - NVE: used for local VTEP address management, VXLAN tunnel management, and ingress replication list management.
  - VXLAN tunnel: used to forward VXLAN packets and is identified by the local VTEP address and remote VTEP address.
  - BDIF: Layer 3 routing interface of a BD, which is used to forward Layer 2 traffic to a Layer 3 domain for route-based forwarding.

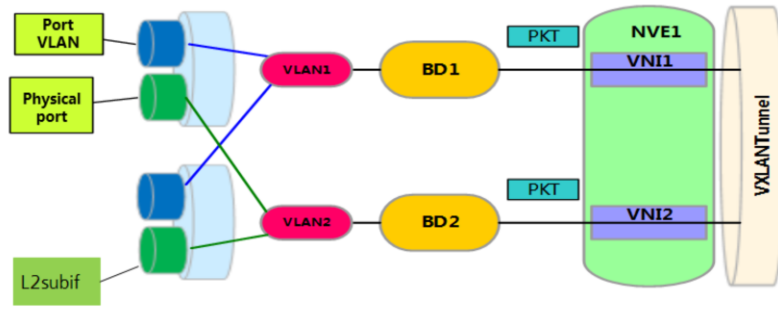


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

VXLAN Access Service Model (2)

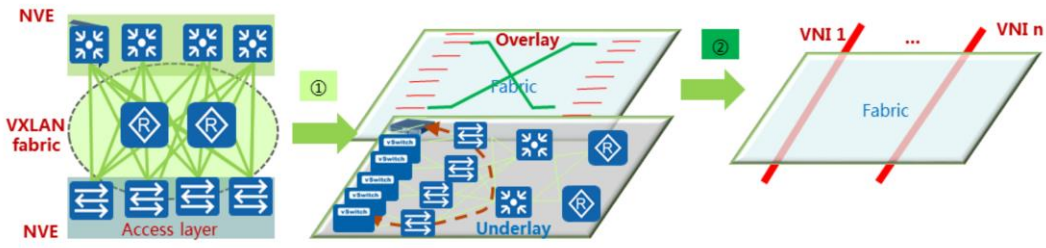
- Global VLAN access model: mainly used in an L2VPN service scenario. In this scenario, a VLAN is bound to a BD to connect the traditional port+VLAN interface to the VXLAN network. A Layer 2 sub-interface needs to be bound to a BD.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

VXLAN Logic Abstraction



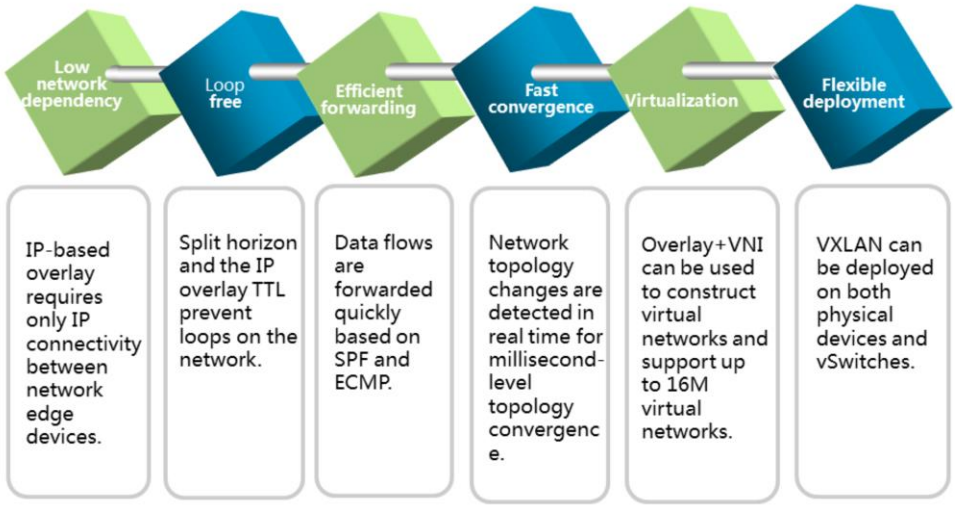
- Simplified VXLAN understanding: virtualization twice

1. First virtualization: The tunneling technology is used to connect edge devices to transparently transmit Layer 2 packets. The entire network is abstracted as an ultra-large LAN switch with a large number of interfaces.
2. Second virtualization: VNIs are used to virtualize the ultra-large switch into multiple Layer 2 broadcast domains, which are similar to VLANs. A VNI is similar to a VLAN ID. The VNI field in the VXLAN header extends the subnet number from 4K to 16M.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

VXLAN Advantages



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

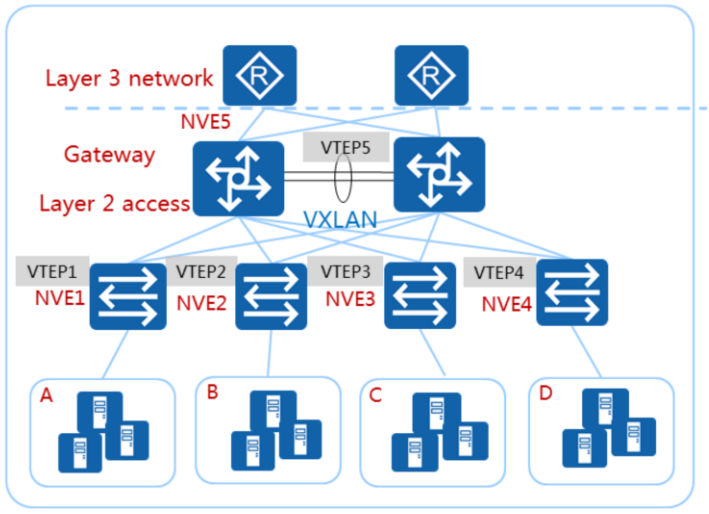


Contents

1. SDN Overview
2. Cloudification Challenges Facing Traditional DC Networks
3. Basic VXLAN Concepts
4. **VXLAN Packet Forwarding Process**
5. VXLAN Concepts and Applications in Cloud DCs

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

VXLAN Packet Forwarding Process (Intra-subnet)



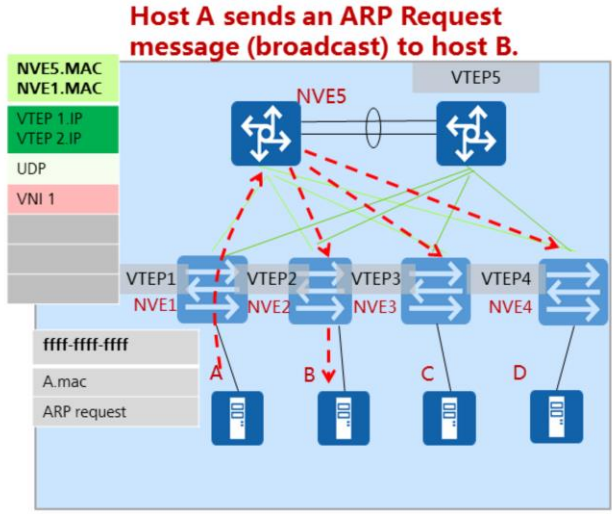
Overall Process

- Host A sends an ARP Request message to host B.
- Host B returns an ARP Reply message to host A.
- Host A sends a unicast data packet to host B.

Note: Hosts A, B, C and D all belong to VNI 1. ARP broadcast suppression is disabled.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MAC Address Table Lookup for Layer 2 Forwarding on the Same Network Segment (1)

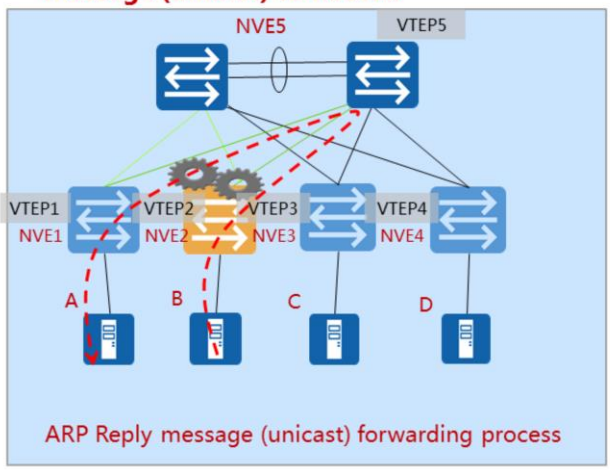


1. After NVE1 finds that the received ARP Request message is a broadcast packet, it broadcasts this packet in VNI 1 and encapsulates it into a VXLAN packet.
2. The intermediate nodes transparently transmit the encapsulated VXLAN packet.
3. After NVEs 2,3, 4, and 5 receive the packet, they decapsulate the packet and learn the MAC address of host A.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

MAC Address Table Lookup for Layer 2 Forwarding on the Same Network Segment (2)

Host B returns an ARP Reply message (unicast) to host A.



ARP Reply message (unicast) forwarding process

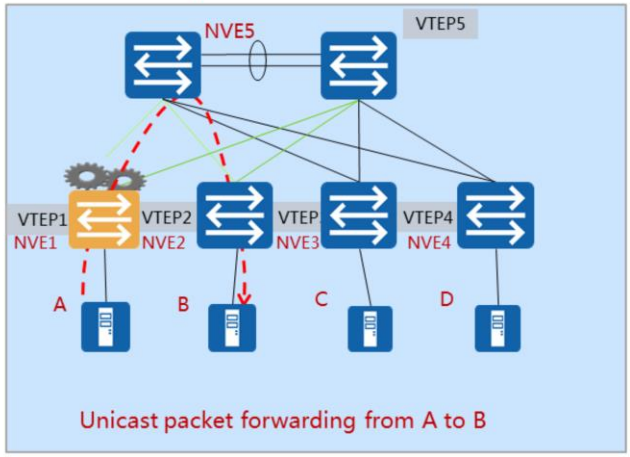
1. NVE2 searches for host A's MAC address entry and finds that the outbound interface is an NVE2-to-NVE1 tunnel interface. NVE2 then encapsulates the packet into a VXLAN packet and sends it at Layer 3.
2. The intermediate nodes transparently transmit the encapsulated VXLAN packet.
3. After NVE1 receives the packet, it decapsulates and forwards it locally. NVE1 learns the MAC address of host B.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

MAC Address Table Lookup for Layer 2 Forwarding on the Same Network Segment (3)

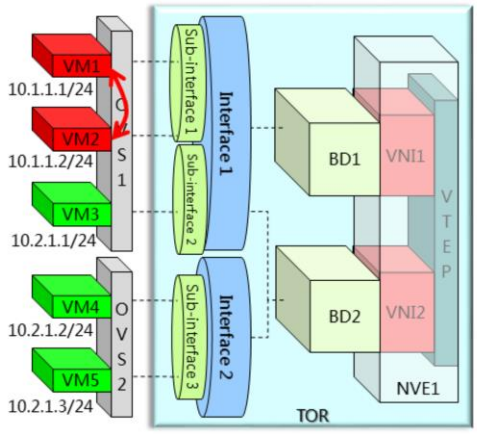
Unicast data packet from host A to host B



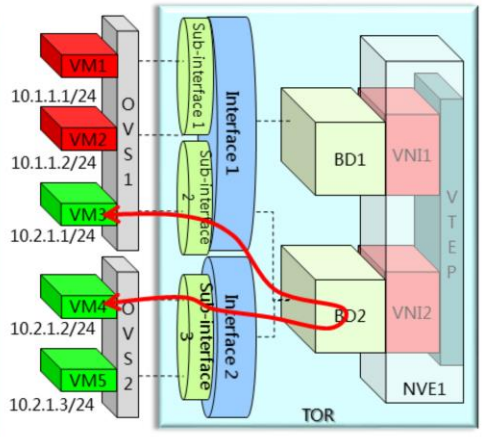
1. Both NVE1 and NVE2 have learned the MAC addresses of hosts A and B. When they search their MAC address tables, they will find matching entries and forward the packet in a process similar to that of VLAN packet forwarding.
2. The only difference is that the outer layer of the data packet is encapsulated with a VXLAN header. The underlay network still forwards the packet based on IP routes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

VXLAN Forwarding Model: Mutual Access Between VMs on the Same Network Segment (1)



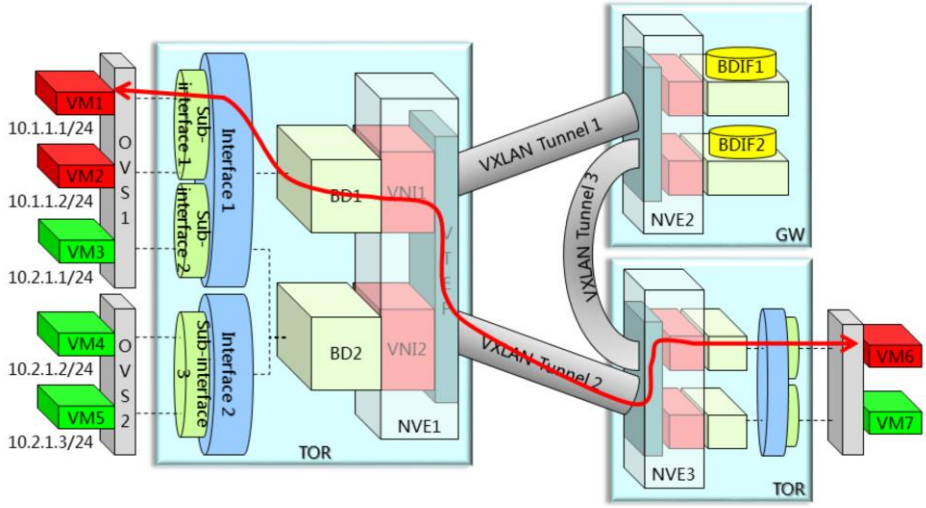
Scenario 1: Both VMs located at the same vSwitches connected to same TOR



Scenario 2: Both VMs located at different vSwitches connected to same TOR

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

VXLAN Forwarding Model: Mutual Access Between VMs on the Same Network Segment (2)

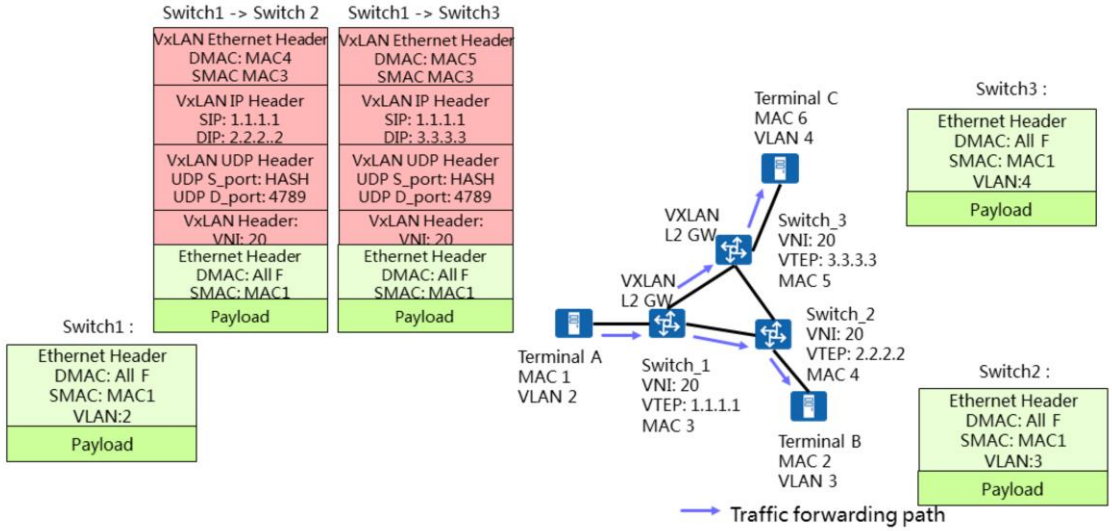


Scenario 3: Both VMs located at different vSwitches connected to different TOR

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

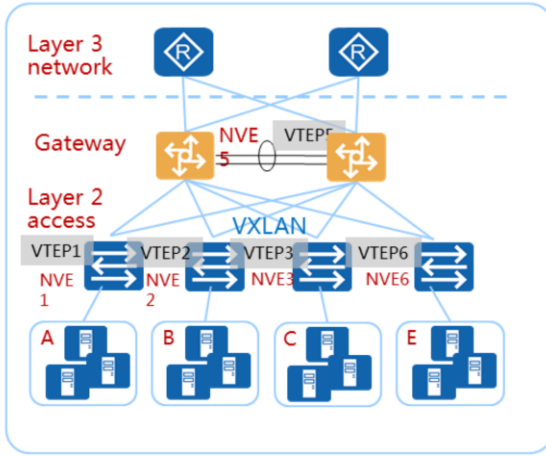
BUM Packet Forwarding Process on a VXLAN Network



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

VXLAN Packet Forwarding Process (Inter-subnet)



Overall Process

- Host A sends a unicast data packet to host E.

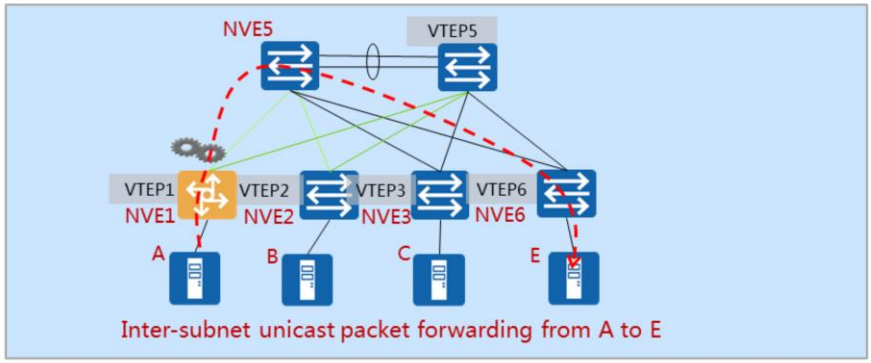
Note: NVE5 functions as a Layer 3 gateway. Host A belongs to VNI 1, and host E belongs to VNI 2. This example assumes that hosts and the gateways have learned the MAC addresses of all nodes through ARP broadcast.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Host A sends a unicast data packet to host E.
- Note: NVE5 functions as a Layer 3 gateway. Host A belongs to VNI 1 and host E belongs to VNI 2. This example assumes that hosts and the gateways have learned the MAC addresses of all nodes through ARP broadcast.

Unicast Packet Forwarding from A to E

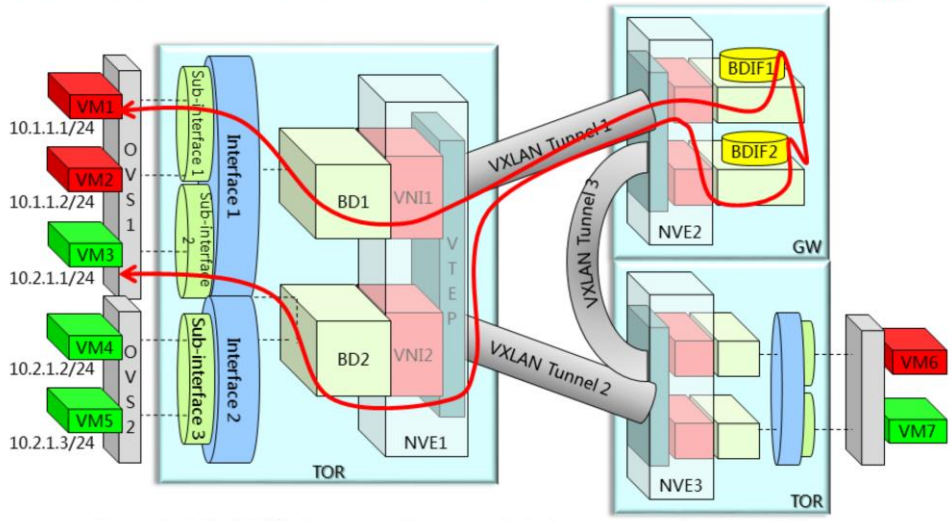


1. After NVE1 finds the MAC address entry of its gateway, it encapsulates the packet into a VXLAN packet, with the VNI being 1 and sends the packet to the gateway.
2. Upon receipt of the packet, the gateway decapsulates the packet, looks for the desired route based on the inner IP header, replaces the inner Ethernet header, encapsulates the packet into a VXLAN packet, with the VNI being 2, and finally sends the packet along the VXLAN tunnel to NVE6.
3. Upon receipt of the packet, NVE6 decapsulates the packet and forwards the inner packet based on the destination MAC address.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

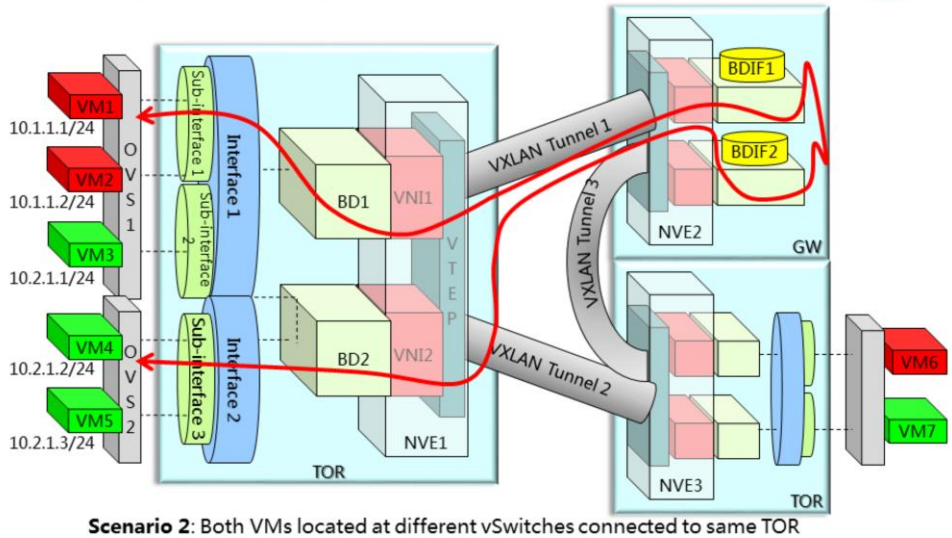
VXLAN Forwarding Model: Mutual Access Between VMs on Different Network Segments (1)



Scenario 1: Both VMs located at the same vSwitches connected to same TOR

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

VXLAN Forwarding Model: Mutual Access Between VMs on Different Network Segments (2)

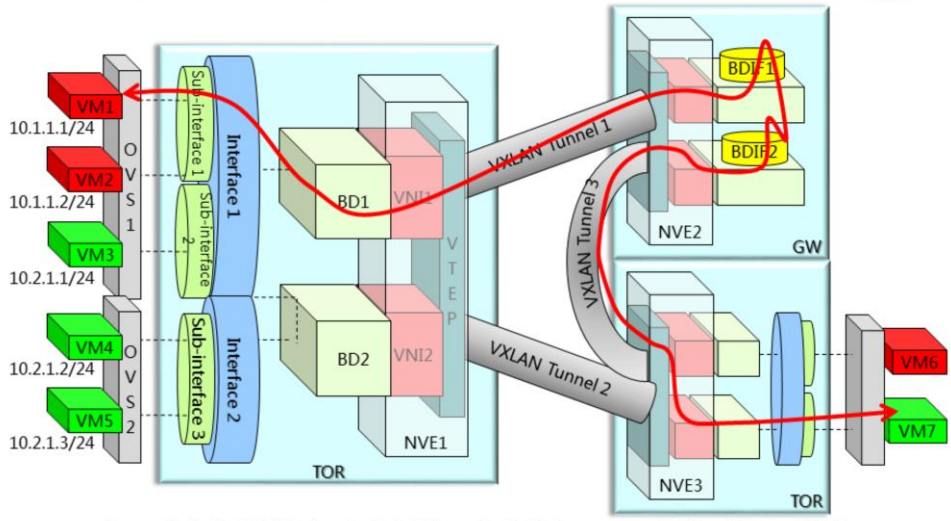


Scenario 2: Both VMs located at different vSwitches connected to same TOR

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

VXLAN Forwarding Model: Mutual Access Between VMs on Different Network Segments (3)



Scenario 3: Both VMs located at different vSwitches connected to different TOR

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

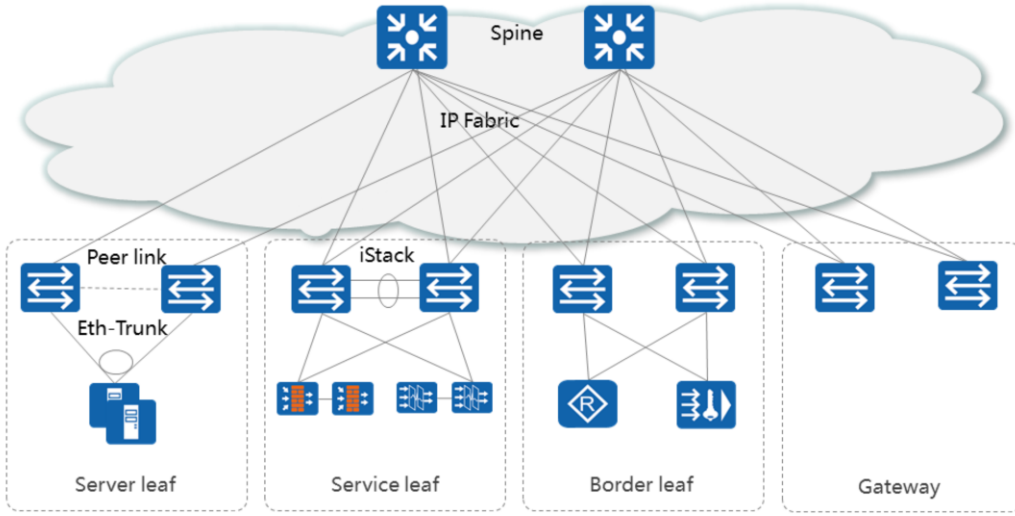


Contents

1. SDN Overview
2. Cloudification Challenges Facing Traditional DC Networks
3. Basic VXLAN Concepts
4. VXLAN Packet Forwarding Process
5. **VXLAN Concepts and Applications in Cloud DCs**

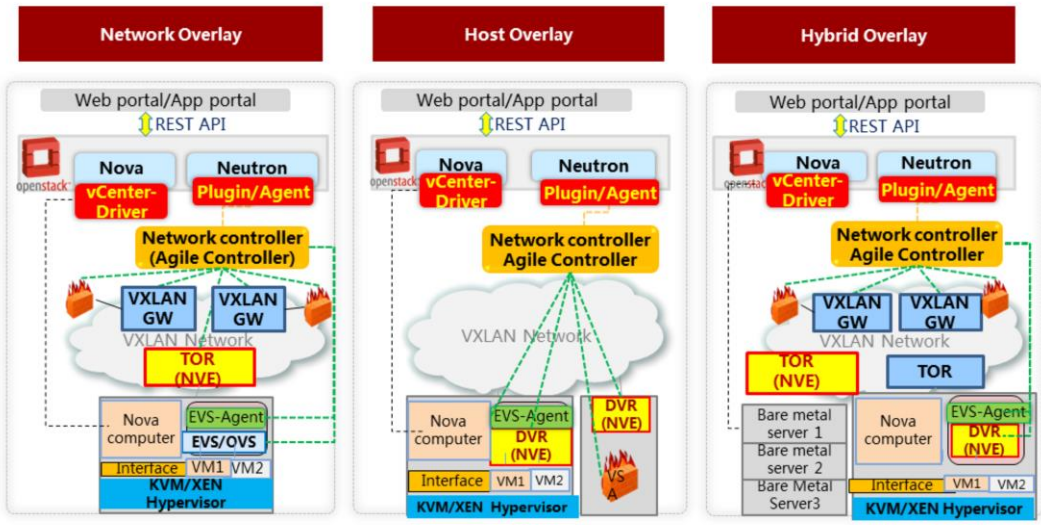
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Spine-Leaf Networking Architecture for VXLAN



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Three VXLAN Overlay Networking Solutions

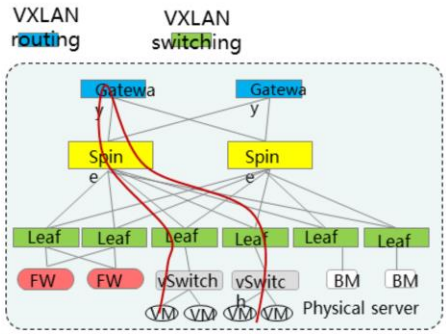


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

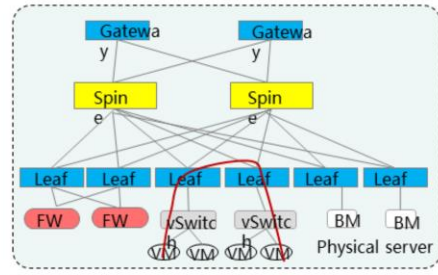
- Host overlay
  - Logical Layer 2 networks can be automatically created without involving physical network reconstruction and configuration. This solution is decoupled from hardware devices.
- Network overlay
  - A new physical network is required for automated service provisioning over the VXLAN overlay network.
- Hybrid overlay
  - SDN is used to configure and manage virtual and physical networks consisting of switches, firewalls, and F5 load balancers and to automate service provisioning.

Distributed and Centralized VXLAN Overlay



Centralized network overlay gateway solution

- VXLAN Layer 2 VTEPs: deployed on leaf nodes
- VXLAN Layer 3 gateways: deployed at the core layer
- Spine: common route



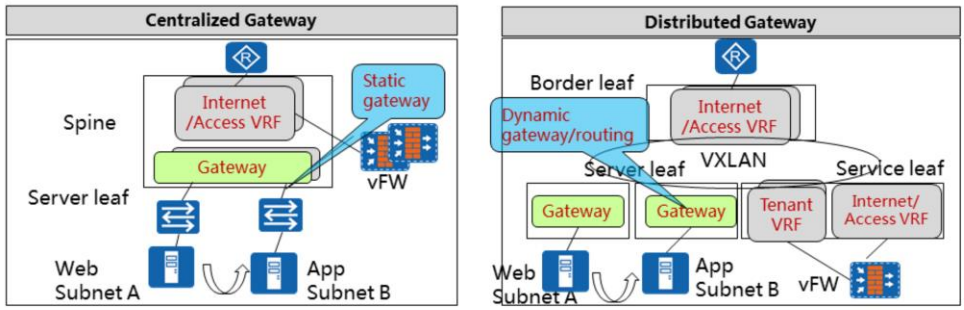
Distributed network overlay gateway solution

- Leaf nodes function as both Layer 3 VXLAN gateways for east-west traffic and Layer 2 VXLAN VTEP gateways.
- The gateway for north-south traffic is deployed at the core layer.
- Spine: common route

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Gateway Deployment Comparison

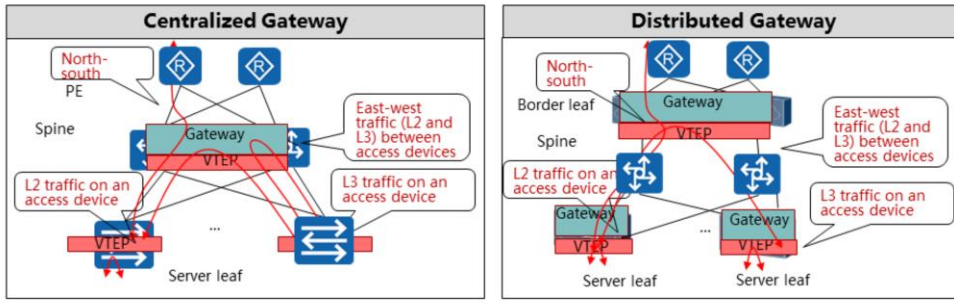


Item	Centralized Deployment	Distributed Deployment
Gateway deployment change upon VM migration	The gateway remains unchanged when VMs are migrated.	The gateway changes accordingly when VMs are migrated. Specifically, the gateway is deleted for the source access device, and a new one is created for the destination access device.
Entries affected by VM migration	The centralized gateway updates ARP entries. Access devices update MAC address entries.	All devices update host routes. Access devices update MAC address entries.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

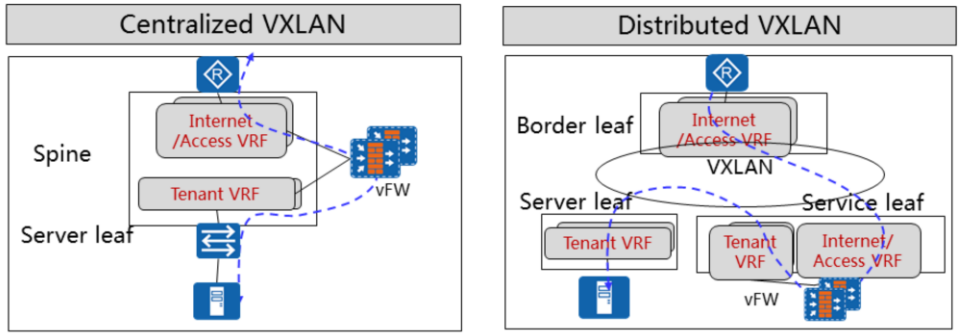
Comparison of Forwarding Paths Before and After Optimization



Item	Centralized Deployment	Distributed Deployment
North-south traffic (customer to servers)	Passing through the core layer	Passing through the core layer
East-west traffic (L2 and L3) between different access devices	Passing through the core layer	Passing through the core layer
East-west traffic (L2) between hosts connected to the same access device	Local forwarding	Local forwarding
East-west traffic (L3) between hosts connected to the same access device	Passing through the core layer	Local forwarding

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

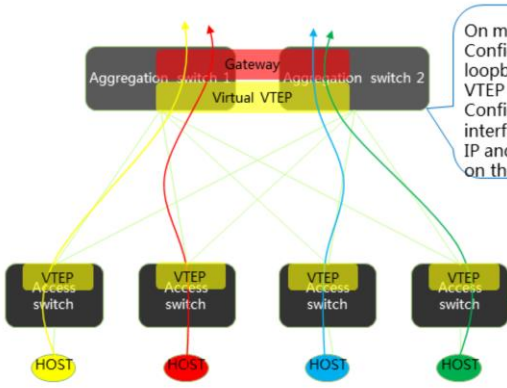
Comparison of Firewall Traffic Filtering



Item	Centralized Deployment	Distributed Deployment
Traffic diversion to firewalls	Firewalls are connected to the centralized gateway in bypass mode. Policies are deployed on the centralized gateway to divert traffic to firewalls.	Firewalls are not directly connected to gateways. Multiple gateways are connected to firewalls through VXLAN tunnels. Policies are deployed on multiple distributed gateways.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Highly Reliable Centralized Gateway



On multiple gateways:
Configure the same loopback address as the VTEP IP address.
Configure same interfaces with the same IP and MAC addresses on the gateways.

1. When a gateway is recovering from a fault, it advertises low-priority routes (adjusted on the controller) to ensure that uplink traffic and downlink traffic are sent to other gateways. After ARP entries on the gateway recovered, route priorities recovered, and the gateway participates in load balancing again.
2. When downlink interfaces of a gateway fail, routes are converged on the underlay network.
3. If all uplink interfaces of a gateway fail, the OPS lowers the priority of the gateway VTEP address so that uplink traffic will not be sent to the faulty gateway.

MAC address table and IP routing table on an access device

Destination MAC address	Remote VTEP
Gateway virtual MAC	vVTEP

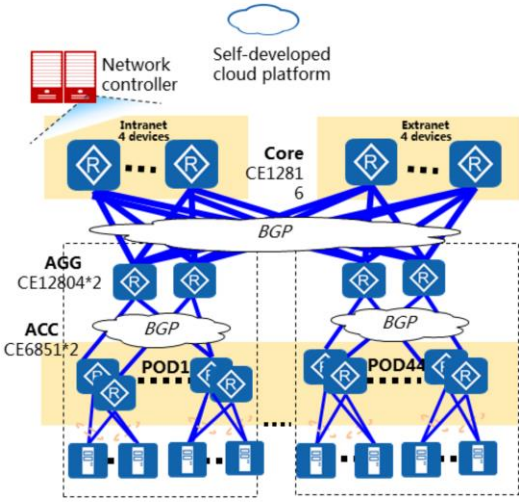
Destination VTEP route	Next hop
vVTEP IP	AGG1&AGG2

Compared with traditional stacked gateways or gateway configured with VRRP, the centralized gateway has the following advantages:
Gateways do not need to run any subnet-based heartbeat protocols, such as VRRP and GLBP, greatly reducing the workload of signaling packet processing.
Compared with VRRP Layer 3 gateways, flow-based load balancing can be implemented among physical gateways, and the number of these physical gateways can be extended.
Advertisement of routes can be controlled on the routing protocol controller vVTEP to implement lossless gateway capacity expansion or upgrade.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Super Large-Scale Public Cloud for Internet Company A in China



Customer Requirements

- Large scale, high performance, and low oversubscription ratio
- DC-wide migration
- SDN controller-based automatic and refined O&M

Solution

- The underlay network uses the spine-leaf architecture, and the overlay network uses VXLAN for large-scale Layer 2 connection. VTEPs are deployed on leaf switches.
- VXLAN Layer 3 gateways in active-active mode are centrally deployed for load balancing to improve the throughput.
- The network controller provides customized northbound APIs to connect to the private cloud platform for automated network service provisioning.
- The network controller is used to achieve path and traffic visualization for easy O&M.

Customer Benefits

- The multi-active centralized hardware gateway ensures network scale and performance, and complies with the O&M habits of management engineers in traditional DCs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Quiz

1. Which of the following statements about VXLAN are incorrect? ( )
   A. VXLAN uses the MAC-in-MAC encapsulation mode.
   B. Similar to MPLS label-based forwarding, VXLAN packets are forwarded based on VNIs.
   C. VXLAN is an NVO3 technology.
   D. In VXLAN packets, the source port number in the outer UDP header is usually calculated using the hash algorithm based on the inner Ethernet frame header.
2. Which of the following are VXLAN networking solutions? ( )
   A. Network overlay
   B. Host overlay
   C. Hybrid overlay

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Answers
  - AB
  - ABC

Section Summary

- Requirements and Challenges of Cloud DCs
- VXLAN Fundamentals
- VXLAN Packet Forwarding Process
- Common VXLAN Networking
- Application Case

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

More Information

- Huawei official websites:
  - Enterprise service: http://enterprise.huawei.com/en/
  - Technical support: https://support.huawei.com/enterprise/en
  - Online learning: http://learning.huawei.com/en/
- Documentation tool:
  - HedEx Lite
- Simulator:
  - eNSP

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Recommendations

- Huawei E-Learning website:
  - http://support.huawei.com/learning/en/newindex.html
- Huawei support case library:
  - http://support.huawei.com/enterprise/servicecenter?lang=en

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— BGP EVPN

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- This chapter describes Ethernet VPN (EVPN) which is used in the network virtualization overlay (NVO) solution. EVPN can be used together with VXLAN technology to implement tunnel encapsulation over the IP layer and function as the control plane of routes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to:
  - Understand the origin and development of EVPN technology.
  - Master the fundamentals of EVPN.
  - Understand Huawei's combined use of EVPN and VXLAN.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. **EVPN Overview**
2. EVPN Fundamentals
3. EVPN Application in NVO

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

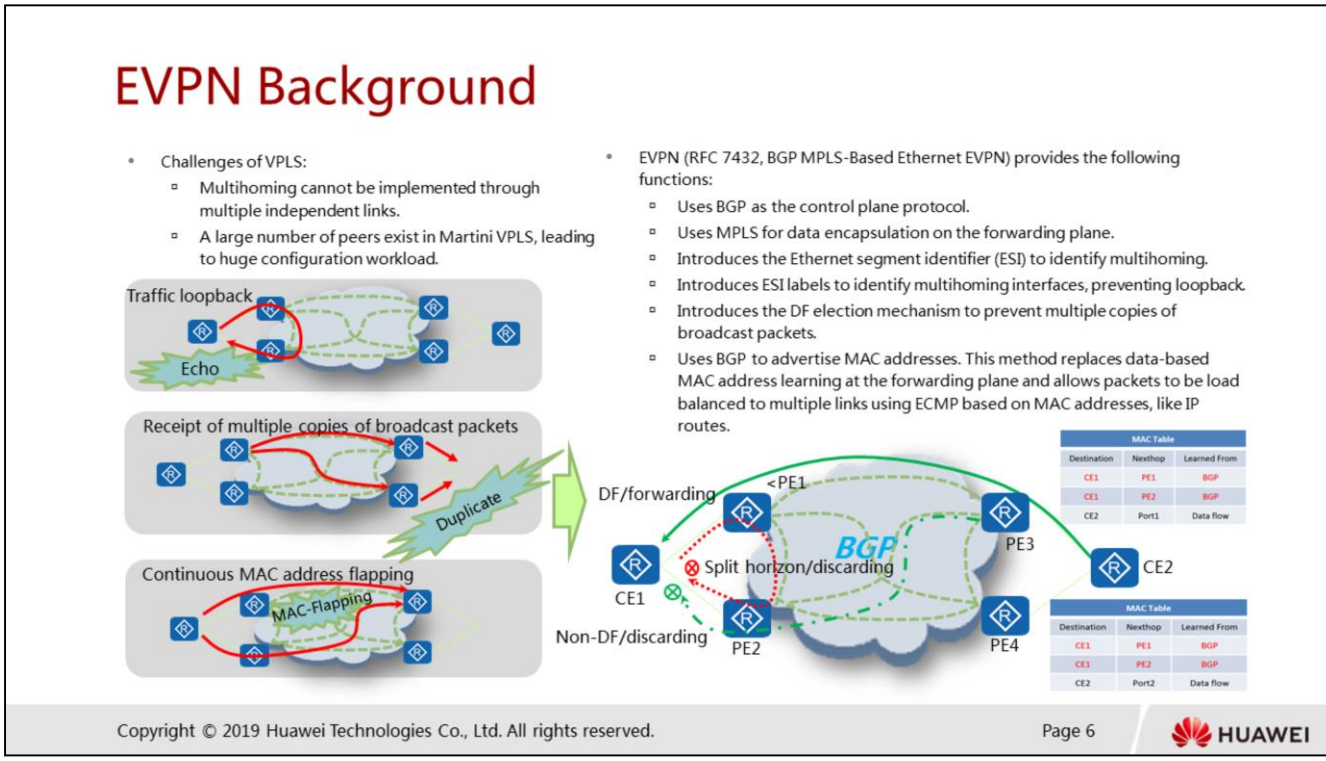
Ethernet L2VPN

- Ethernet L2VPN services are widely deployed on the live network. For example, gateways of services, such as voice over Internet Protocol (VoIP), high-speed Internet (HSI), and broadcast TV (BTV) services, are deployed on BRASs. Layer 2 networks are constructed over a metro network between enterprises to achieve data center (DC) interworking.
- VPLS is widely used on the live network to transparently transmit Layer 2 packets. However, VPLS is in face of the following restrictions and challenges:
  - Multihoming
  - Multicast optimization
  - Complex configuration
  - Multi-tenant DCI
  - Fast convergence
  - Broadcast suppression
- EVPN is introduced to address these challenges.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- HSI: high-speed Internet
- BTV: broadcast TV
- Multihoming: Currently, VPLS supports multihoming only in single-active redundancy mode, instead of multi-path all-active forwarding.
- Multicast optimization: Multicast LSPs can be used together with VPLS but can only be used for P2MP LSPs. VPLS does not support MP2MP LSPs.
- Complex configuration: Currently, VPLS provides BGP-based single-sided access for auto-discovery. This requires heavy workload in configuring access-side Ethernet configurations.
- Multi-tenant DCI: In addition to supporting Layer 2 networks between DCs, DCI links require extension of Layer 2 networks for tenants.



- We have mentioned that the disadvantages of VXLAN call for new control plane protocols. Let's take a look at the EVPN protocol. EVPN, which is short for Ethernet VPN, is defined in RFC 7432 and used to solve some existing problems of VPLS. For example, VPLS does not support multihoming through multiple independent links. In some cases, multiple broadcast packets may be received or MAC address flapping occurs. A large number of peers exist in Martini VPLS, leading to huge configuration workload.
- EVPN uses BGP as the control plane protocol and uses MPLS to implement forwarding-plane data encapsulation to resolve the problems of loops, multiple broadcast packets, and MAC address learning in VPLS scenarios.

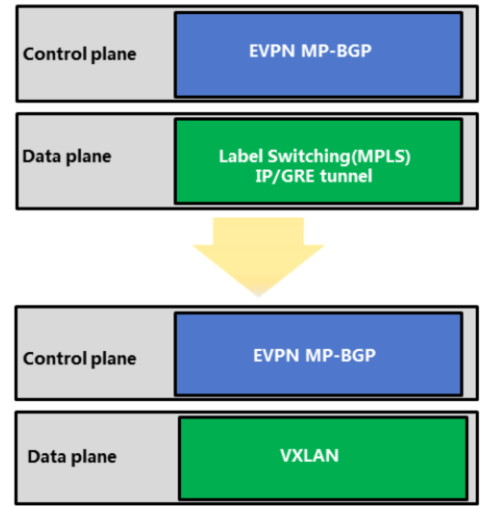
BGP-EVPN Development and Application

EVPN Development

- Ethernet VPN was originally defined in RFC 7432 to meet requirements for high bandwidth and complex QoS in MPLS-based VPNs. The MP-BGP protocol is applied to the control plane to introduce the address family concept.
- Characteristics of EVPN: The control plane and data plane are abstracted and isolated from each other. The MP-BGP control plane transmits MAC/IP routes. There are several methods for encapsulation on the data plane.

EVPN for VXLAN

- Since VXLAN has no control plane, the MP-BGP control plane of EVPN is suitable for VXLAN.
- VXLAN extends the standard EVPN.
- EVPN drafts:
  - draft-ietf-bess-evpn-overlay
  - draft-ietf-bess-evpn-inter-subnet-forwarding
  - draft-ietf-bess-evpn-prefix-advertisement

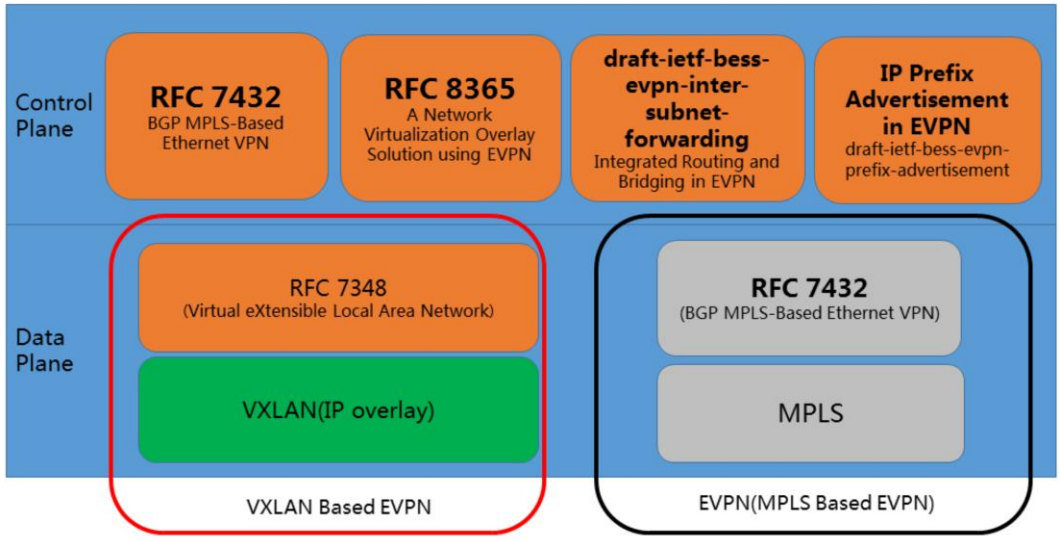


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- EVPN is the VPN technology used for Layer 2 interworking. EVPN uses a mechanism similar to BGP/MPLS IP VPN. EVPN defines a new type of BGP Network Layer Reachability Information (NLRI) called EVPN NLRI. EVPN NLRI defines new BGP EVPN routes to implement MAC address learning and advertisement between different sites on a Layer 2 network.
- The original VXLAN implementation solution does not have a control plane. VTEP discovery and host information (including IP addresses, MAC addresses, VNIs, and gateway VTEP IP addresses) learning are performed through traffic flooding on the data plane. As a result, there is a lot of flooded traffic on the DC network. To solve this problem, VXLAN adopts EVPN as the control plane. BGP EVPN routes are exchanged between VTEPs to implement automatic discovery of VTEPs and advertisement of host information, avoiding unnecessary traffic flooding.
- In addition to RFC 7432, there are three EVPN drafts. The draft-ietf-bess-evpn-overlay has evolved to RFC 8365, A Network Virtualization Overlay Solution Using Ethernet VPN (EVPN). The other two drafts under the way of becoming standards.

EVPN Implementation



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- VXLAN is used as the data plane.

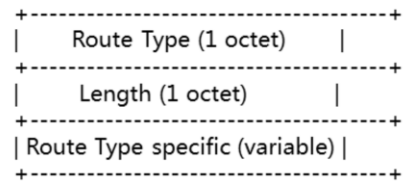
Contents

1. EVPN Overview
2. **EVPN Fundamentals**
3. EVPN Application in NVO

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

EVPN NLRI

- In RFC 7432, EVPN defines a new NLRI, which is called EVPN NLRI. The format of EVPN NLRI is as follows:



- RFC 7432 defines four types of routes:
  - 1 - Ethernet Auto-Discovery (A-D) Route
  - 2 - MAC/IP Advertisement Route
  - 3 - Inclusive Multicast Ethernet Tag Route
  - 4 - Ethernet Segment Route

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Type 1: Ethernet Auto-Discovery Route

- EVPN NLRI format:



- Prefix indexes:
  - Ethernet Segment Identifier
  - Ethernet Tag ID

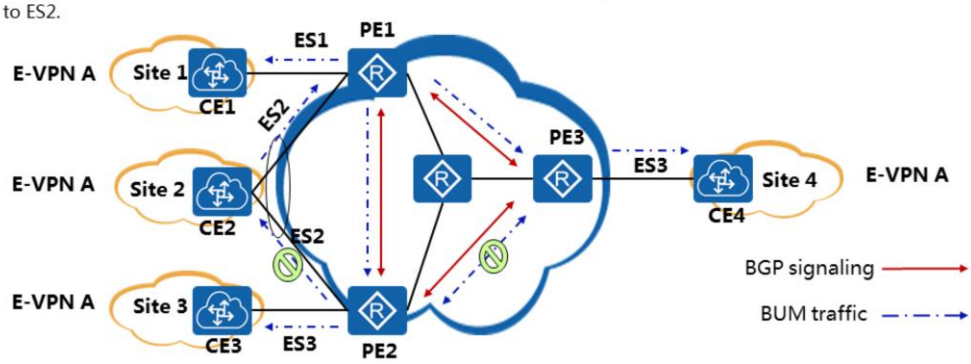
- Route type 1 is also called the Ethernet Auto-Discovery (A-D) route.
- It is required only when ESI is used to implement multihoming access.
- It provides following functions:
  - Split horizon
  - Fast convergence
  - Aliasing

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Split horizon (ESI label assignment)
- Fast convergence (Other PEs implement batch fast switchover of specific routes, such as MAC advertisement routes based on RT1 routes.)
- Alias (Multihoming PEs can advertise specific routes, such as MAC advertisement routes. Other PEs can form ECMP links to all multihoming PEs based on RT1 routes.)
- M-LAG and stacking technologies can be used to replace such routes.

Split Horizon

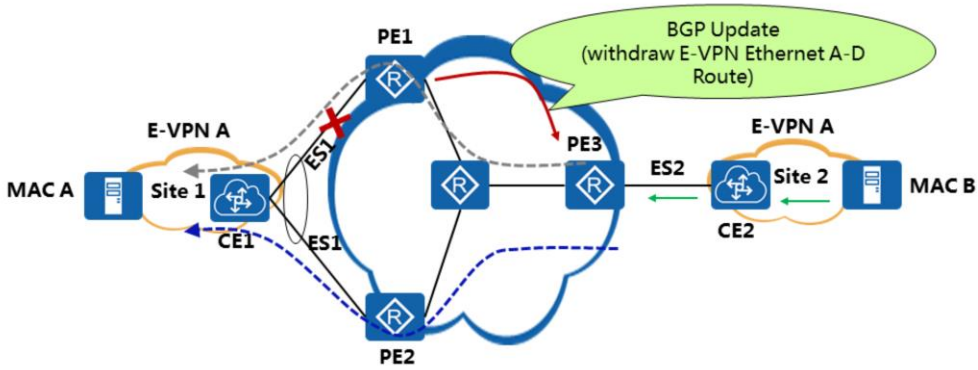
- As shown in this figure, CE2 is dual-homed to PE1 and PE2. After receiving the BUM traffic from CE2, PE1 replicates and forwards the traffic to PE2. PE2 does not send the received packets to CE2. This prevents loops on the CE side. Similarly, the multicast traffic received by PE2 from PE1 is not forwarded to the PE side, preventing loops on the public network side.
- Split horizon is implemented using the ESI label carried in Ethernet A-D per ES routes. As shown in the figure, PE2 assigns an ESI label to identify ES2 and advertises the ESI label to PE1 through Ethernet A-D routes. PE1 adds this label to the BUM traffic before sending the BUM traffic to PE2, and PE2 identifies the label after receiving the BUM traffic. That is, the BUM traffic is not forwarded to ES2.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fast Convergence

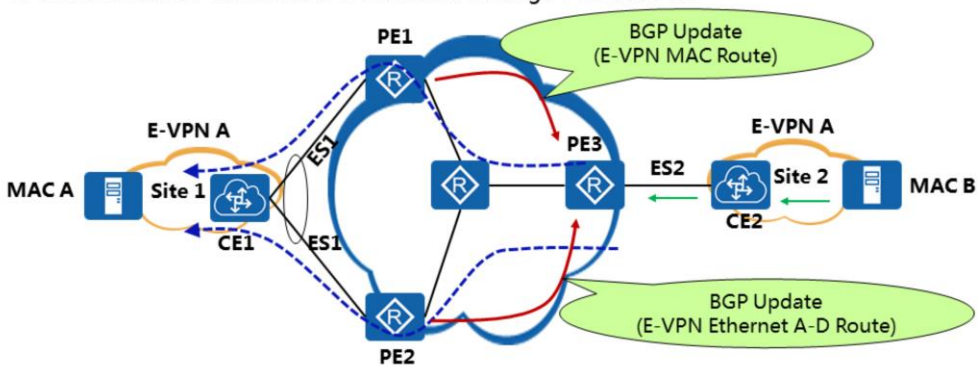
- Ethernet A-D per ES routes are used to quickly withdraw MAC routes to achieve fast convergence.
- As shown in this figure, CE1 is dual-homed to PE1 and PE2. PE1 and PE2 learn the MAC address of CE1 and advertise MAC routes to PE3. When the link between CE1 and PE1 fails, PE1 sends an Ethernet A-D per ES withdrawal message to PE3, so that PE3 can rapidly update the MAC routes of ES1 and switch traffic to PE2.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Aliasing

- In CE multihoming all-active scenarios, some PEs may fail to learn the CE's MAC address. If this problem occurs, the remote PEs do not participate in load balancing or work as a backup. To address this problem, EVPN introduces the concept of aliasing through Ethernet A-D per EVI routes.
- As shown in this figure, CE1 is dual-homed to PE1 and PE2. Assume that PE1 learns MAC A and advertises MAC routes to PE3. PE2 does not learn MAC A, but PE2 can advertise Ethernet A-D per EVI routes to PE3. PE3 knows that MAC A is reachable through PE1 and PE2.

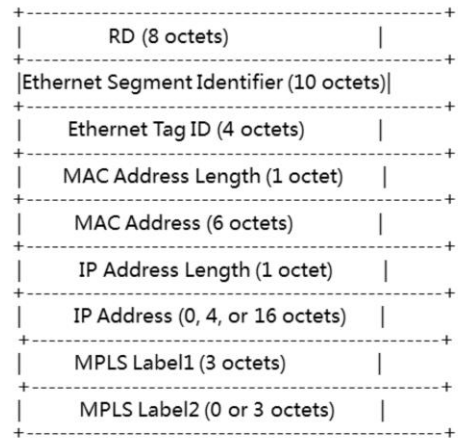


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Type 2: MAC/IP Advertisement Route

- EVPN NLRI format:
- Functions:
  - Guide unicast traffic forwarding.
  - Advertise MAC/IP addresses of VMs.
- Prefix indexes:
  - Ethernet Tag ID, MAC Address, and IP Address are used as prefix indexes. ESI and MPLS Label are used as route attributes.

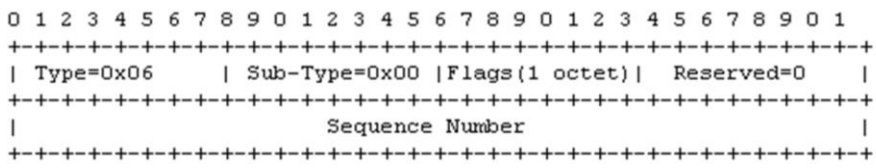


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

MAC Mobility Extended Community

- Description:
  - This extended community is a transitive extended community having a Type field value of 0x06 and a Sub-Type field value of 0x00. It may be advertised along with MAC/IP Advertisement routes.
  - Flags: Low-order bit. A value of 1 means that the MAC address is static and cannot move.
- Format:



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Ethernet Segment Identifier (ESI)

- Format:

```
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Type=0x06   | Sub-Type=0x01 | Flags (One Octet)  |Reserved=0  |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
| Reserved = 0|          ESI Label                                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

- Description:
  - This extended community is a transitive extended community having a Type field value of 0x06 and a Sub-Type field value of 0x01. It may be advertised along with Ethernet A-D routes.
  - Flags: Least significant bit. A value of 0 means the all-active redundancy mode, and a value of 1 means the single-active redundancy mode.
  - ESI Label: This 3-byte ESI label field is used for split horizon. In single-active redundancy mode, the value of this field is 0.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

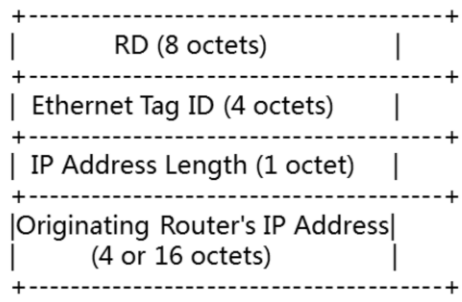
Redundancy Modes

- Single-active redundancy mode:
  - For a given ES, if one Flags field is set to Single-Active in a set of Ethernet A-D per ES routes received by a PE, the PE works in single-active redundancy mode. In this case, traffic can be transmitted only over the primary PE.
  - If the primary PE fails, Ethernet A-D per ES routes can be withdrawn and the remote PE switches traffic to another PE in the group.
- All-active redundancy mode:
  - For a given ES, if all Flags fields are set to All-Active in a set of Ethernet A-D per ES routes received by a PE, the PE works in all-active redundancy mode. In this case, traffic can be transmitted over any PE.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Type 3: Inclusive Multicast Ethernet Tag Route

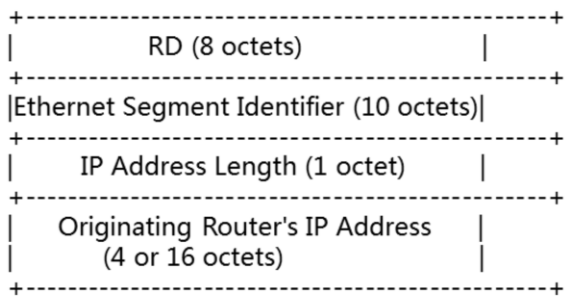
- Type 3 route is also called the Inclusive Multicast Ethernet Tag route.
- It is used for automatic tunnel establishment and automatic join of VNI broadcast members.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Type4: Ethernet Segment Route

- NLRI format:

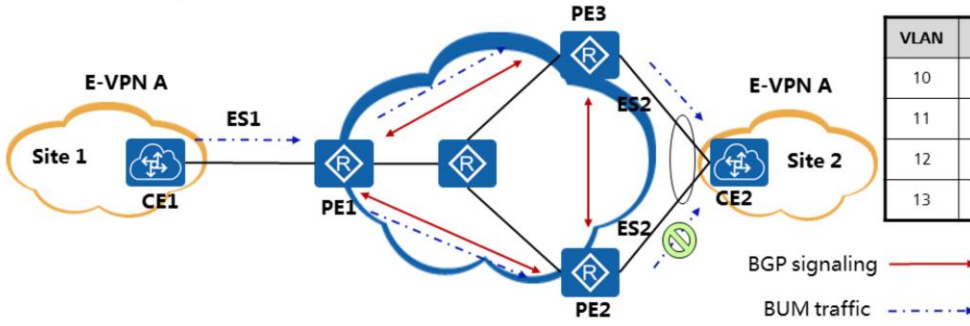


- This type of route is required only in multihoming access scenarios and can be used for DF election between multihoming links.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

DF Election

- When a CE is multi-homed to several PEs, only one PE needs to forward BUM traffic to the CE. The process of selecting the PE is called DF election.
- As shown in this figure, CE2 is dual-homed to PE2 and PE3. The BUM packets originating at CE1 will be sent to PE2 and PE3. DF election is performed between PE2 and PE3. If PE3 is elected as the primary PE and PE2 as the backup PE, PE2 does not forward BUM traffic to CE2.
- DF election is implemented through ES routes. When the DF election timer (3s by default) expires, each PE performs election according to the ES routes of an EVI. The election algorithm is as follows: The same ES routes are sorted according to the originator IP address in ascending order. After a PE list is generated, the modulo operation is performed based on the number of PEs in the VLAN. If the modulus value is equal to the position of the PE in the PE list, the PE is elected as the primary PE and other PEs are backup PEs.



VLAN	PE2	PE3
10	DF	
11		DF
12	DF	
13		DF

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



Contents

1. EVPN Overview
2. EVPN Fundamentals
3. **EVPN Application in NVO**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Background of Type 5 Routes

- BGP EVPN defines four types of routes and provides the flexible control plane to achieve the goal of Layer 3 interworking on the same network segment. In some cases, cross-tenant systems need to directly access the network without protocol interaction.
- EVPN defines Type 5 routes to advertise IP prefix routes.
- IP prefix routes are used to allow connections between different subnets and host access to external networks.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- For details, see the RFC draft "IP Prefix Advertisement in EVPN."
- draft-ietf-bess-evpn-prefix-advertisement-11

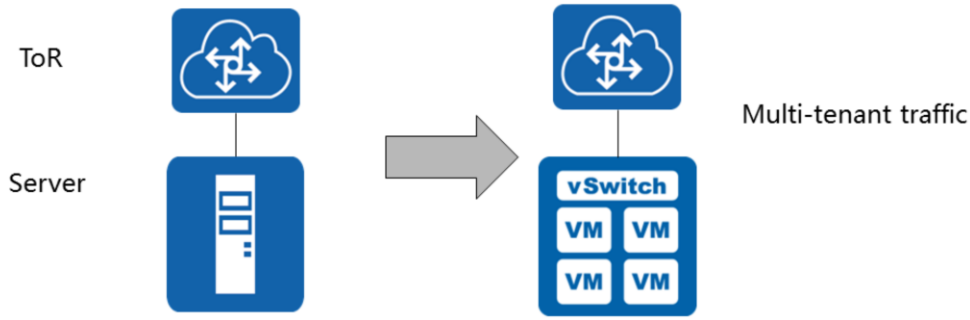
Huawei's Application of EVPN

- EVPN is an emerging technology. The current RFCs have defined Type 1 to Type 4 routes and support MPLS LSPs on the data plane.
- Huawei applies EVPN technology to wider IP networks and combines the use of EVPN technology and IP tunneling technology.
- IP tunnels use VXLAN technology. BGP EVPN functions as the control plane, and a VXLAN tunnel functions as the forwarding plane. BGP EVPN is used in NVO scenarios.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

NVO

- The virtualization technology is introduced in the cloud DC so that one host can carry multiple VMs and each VM belongs to a different tenant. This poses new requirements for the network. The NVO solution is introduced to solve this problem.

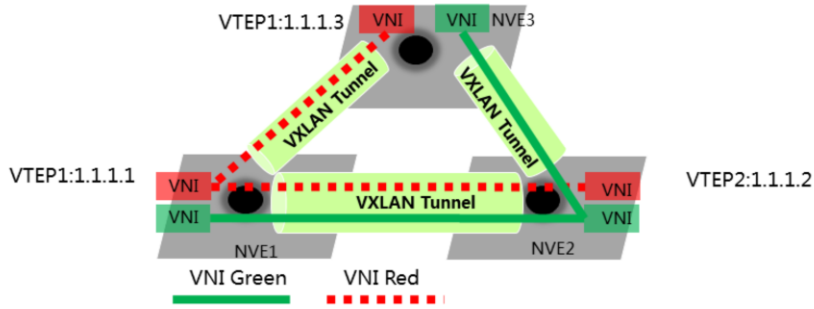


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- NVO allows traffic of each tenant to be carried over an independent overlay tunnel.
- Multiple overlay tunnels can be carried over an underlay network.

VXLAN Encapsulation

- Currently, the mainstream NVO technologies are VXLAN, NVGRE, and MPLS over GRE.
- VXLAN provides the data-plane encapsulation capability that allows data packet transmission between two NVEs on a common IP network.
- VXLAN encapsulation is based on UDP and provides an 8-byte VXLAN header next to the UDP header. The VNI consists of 24 bits. Each VNI identifies a tenant. Data frames encapsulated on the ingress VTEP do not contain internal VLAN tags. The egress VTEP discards the frames carrying the inner VLAN tag.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

EVPN and NVO (1)

- NVO resolves the multi-tenant problem but that applies only to tenants on the data plane.
- EVPN is used as the control plane of NVO. BGP is extended to define NLRI based on EVPN, which is used to transmit control signaling over data forwarding.
- The combined use of EVPN and VXLAN meets the following requirements:
  - Multi-tenant network isolation is achieved.
  - Massive tenants are supported.
  - The tenant network supports large Layer 2 connections across physical networks.
  - VM networks can be migrated across physical networks.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

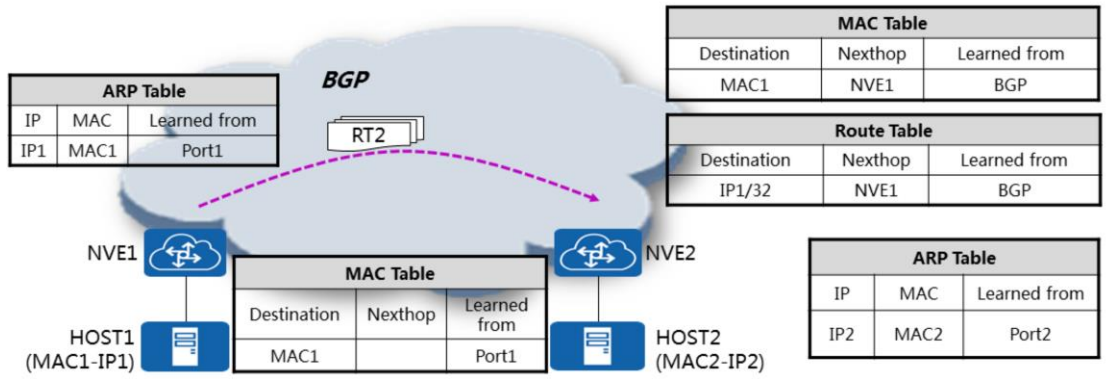
- VXLAN does not provide the control plane, and traffic is flooded on the data plane to discover VTEPs and host information (IP and MAC addresses, VNIs, and gateway VTEP IP address), resulting in high traffic volumes in the DC. To solve this problem, VXLAN adopts EVPN as the control plane. BGP EVPN routes are exchanged between VTEPs to implement automatic discovery of VTEPs and advertisement of host information, avoiding unnecessary traffic flooding.

EVPN and NVO (2)

- EVPN is used to solve the following problems:
  - The control plane is distributed by BGP, and broadcast and multicast packets are replicated by the shared tree or ingress.
  - MAC learning is implemented on the control plane.
  - Route reflection replaces full interconnection.
- Because NVO needs to allow millions of instances to be implemented on a physical device, the scalability of the control plane is very important. EVPN and its extended functions are designed to meet this requirement.
- Huawei combines EVPN with VXLAN, with Type 2, Type 3, and Type 5 routes mainly used.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

EVPN Routes — Type 2 Routes



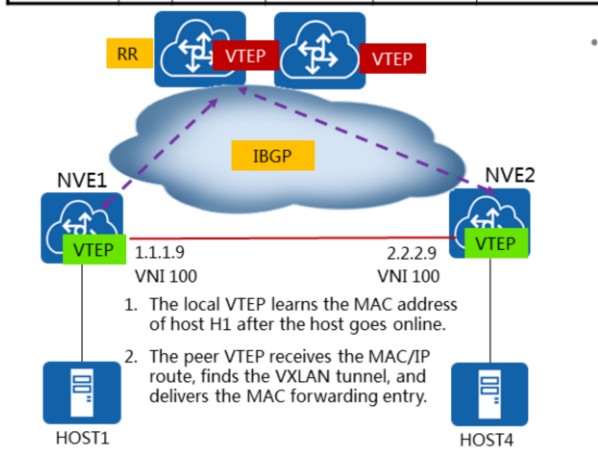
- Type 2 routes are used to advertise the MAC addresses and MAC+IP information of hosts.
- After receiving Type 2 routes that contain the MAC addresses and ARP entries learned by Ethernet interfaces, devices generate MAC forwarding tables and host route forwarding tables.
- The IP field in Type 2 routes is optional. Without this field, a device, such as an L2 VXLAN-capable switch, advertises only MAC addresses but not IP addresses.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- EVPN extends BGP to define several types of BGP EVPN routes, which can be used to transmit VTEP addresses and host information. EVPN is applied to the VXLAN network to move VTEP discovery and host information learning from the data plane to the control plane. Now, let's take a closer look at BGP EVPN routes.
- VXLAN uses the Type 2 routes (also called MAC/IP Advertisement routes) specified by the EVPN protocol to advertise the MAC address or MAC+IP of a host. BGP-EVPN allows the MAC addresses and ARP entries learned by Ethernet interfaces to be converted into Type 2 routes. After Type 2 routes are advertised to other devices, these devices generate MAC forwarding tables and host route forwarding tables.
- Traditionally, MAC address learning is triggered by packets. BGP-EVPN uses Type 2 routes that carry MAC information to advertise MAC addresses, which saves a lot of ARP traffic.

EVPN VXLAN — MAC/IP Route Advertisement

Route Type	RD	MAC	Label	Tunnel Type	Next-hop
2	1:1	H1-MAC	L2 VNI:100	8(VXLAN)	L2 VTEP: 1.1.1.9



MP BGP EVPN
VXLAN Tunnel
VTEP VXLAN L3 GW
VTEP VXLAN L2 Bridging

- EVPN defines MAC/IP routes to advertise host MAC addresses, reducing traffic floods on the VXLAN network.
  - The local VTEP collects the MAC address of the local host, converts it into an EVPN MAC/IP route, and sends the route to the remote VTEP through the EVPN control plane.
  - After receiving a MAC/IP route, the peer VTEP delivers the route to the corresponding EVPN instance and finds the matching VXLAN tunnel based on the next hop in the route. If the tunnel is reachable, the VTEP delivers the MAC forwarding entry.

MAC Table

MAC	VNI	DIP
H1-MAC	100	1.1.1.9

Tunnel Table

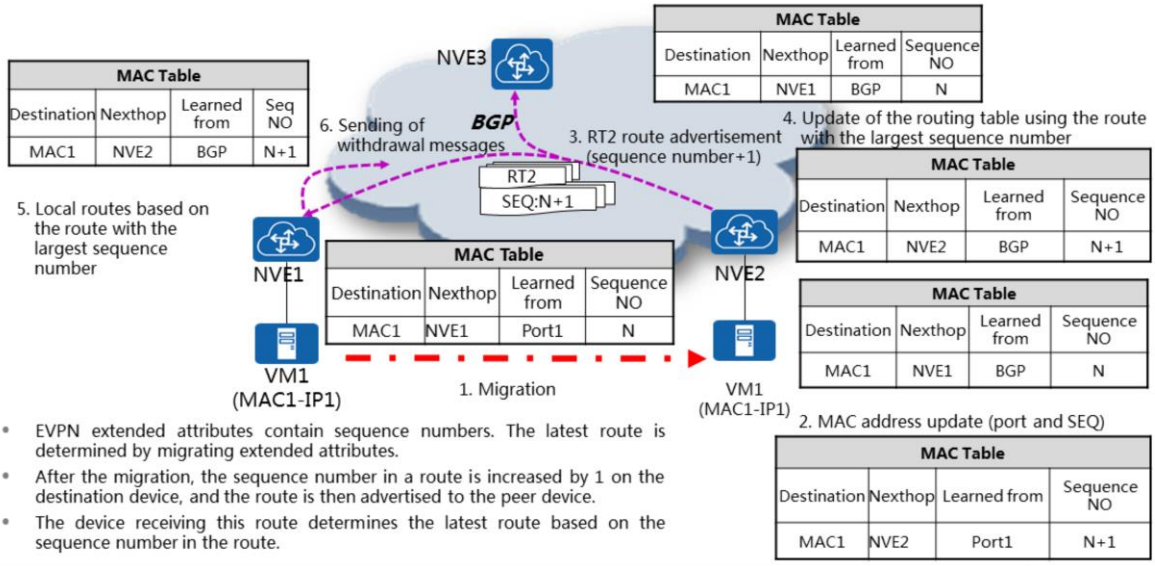
Tunnel	SIP	DIP
1	2.2.2.9	1.1.1.9

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Let's take a look at the application scenarios of Type 2 routes.
  - First up, MAC route advertisement. In this example, we can see that after the local host H1 goes online, the local NVE learns the MAC address of the host and sends the MAC address to the remote device through BGP-EVPN.
  - After receiving a MAC/IP route, the peer VTEP delivers the route to the corresponding EVPN instance and finds the matching VXLAN tunnel based on the next hop in the route. If the tunnel is reachable, the VTEP delivers the MAC forwarding entry.

VM Migration



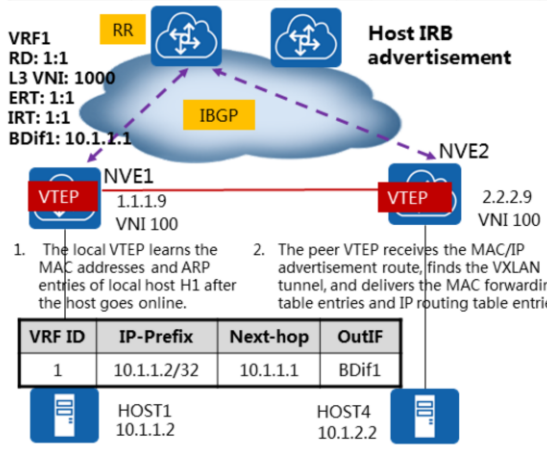
- EVPN extended attributes contain sequence numbers. The latest route is determined by migrating extended attributes.
- After the migration, the sequence number in a route is increased by 1 on the destination device, and the route is then advertised to the peer device.
- The device receiving this route determines the latest route based on the sequence number in the route.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

EVPN VXLAN - Host Route Learning

Route Type	RD	MAC	IP	Label1	Label2	Tunnel Type	Nexthop
2	1:1	H1-MAC	10.1.1.2	L2 VNI: 100	L3 VNI: 1000	8(VXLAN)	VTEP: 1.1.1.9



VRF ID	IP-Prefix	Next-hop	OutIF
1	10.1.1.2/32	10.1.1.1	BDif1



- EVPN uses Type 2 routes (MAC/IP advertisement routes) to transmit host MAC addresses and IP addresses.
  - The local VTEP collects the MAC address and ARP information about the local host, generates an EVPN Type 2 route, and advertises the route to the remote VTEP through the EVPN control plane.
  - After receiving the MAC/IP route, the peer VTEP delivers the route to the corresponding EVPN instance and VPN instance, and finds the matching VXLAN tunnel based on the next hop in the route. If the tunnel is reachable, the VTEP delivers the MAC forwarding entry and IP routing entry.

MAC Table

MAC	VNI	DIP
H1-MAC	100	1.1.1.9

Tunnel Table

Tunnel	SIP	DIP
1	2.2.2.9	1.1.1.9

FIB table

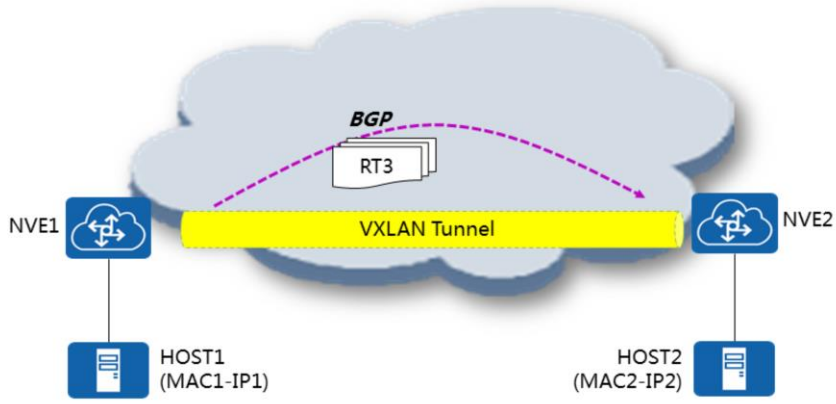
VRF ID	IP	L3 VNI	RouterMAC	Nexthop
1	10.1.1.2/32	1000	VTEP1 MAC	1.1.1.9

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Type 2 routes are also called MAC/IP advertisement routes. After the local host H1 goes online, the local VTEP learns the MAC address and ARP entry of the host and generates EVPN Type 2 routes, and the routes are sent to the remote device through BGP-EVPN.
- After receiving MAC/IP advertisement routes, the peer VTEP delivers the routes to the corresponding EVPN instance and finds the matching VXLAN tunnel based on the next-hop address in the routes. If the tunnel is reachable, the VTEP delivers the MAC forwarding table and IP routing table.

EVPN Routes - Type 3 Routes



- Inclusive Multicast Ethernet Tag Route:
  - Type 3 routes are used for automatic tunnel establishment and automatic join of VNI broadcast members.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

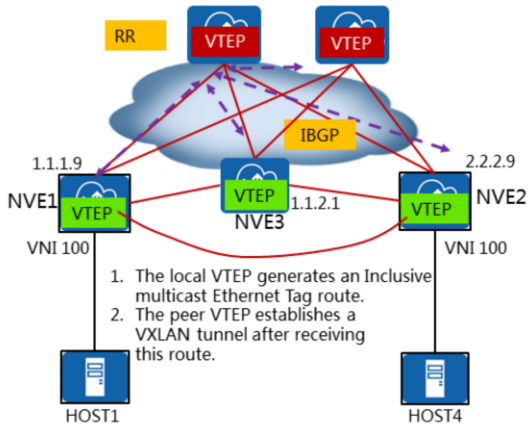
HUAWEI

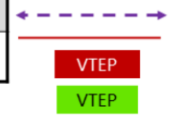
- Type3 routes are also called Inclusive Multicast Ethernet Tag routes. This type of route consists of the prefix and PMSI attributes, which is used for automatic tunnel establishment and automatic join of VNI broadcast members.
- This type of route is used for automatic VTEP discovery and dynamic VXLAN tunnel establishment on the VXLAN control plane. After a BGP EVPN peer relationship is established between VTEPs, they exchange inclusive multicast routes to transmit Layer 2 VNIs and VTEP IP addresses to each other.
- The Originating Router's IP Address and MPLS Label fields carried in the routes indicate the local VTEP's IP address and Layer 2 VNI, respectively. If a route destined for the peer VTEP's IP address is reachable, a VXLAN tunnel is established from the local VTEP to the peer VTEP. Additionally, if the local and peer VNIs are the same, an ingress replication list is created for subsequent BUM packet forwarding.

EVPN VXLAN - Tunnel Establishment

Route Type	RD	Originating IP	Label	Tunnel ID
3	1:1	1.1.1.9	L2 VNI: 100	L2 VTEP: 1.1.1.9

Inclusive route





- Inclusive multicast Ethernet Tag routes are used to transmit Layer 2 VNIs and VTEP addresses and dynamically establish Layer 2 VXLAN tunnels.
  - After the BD, L2 VNI, and VTEP address are configured on the local VTEP, Inclusive multicast Ethernet Tag routes are generated and sent to the remote VTEP.
  - After receiving the Inclusive multicast Ethernet Tag routes, the remote VTEP compares parameters in the routes with the local BD and L2 VNI. If the parameters are the same, the two VTEPs set up a VXLAN tunnel and an ingress replication list for this VXLAN tunnel.

Tunnel Table

Tunnel	SIP	DIP
1	2.2.2.9	1.1.1.9

Ingress Replication List of the VXLAN Tunnel

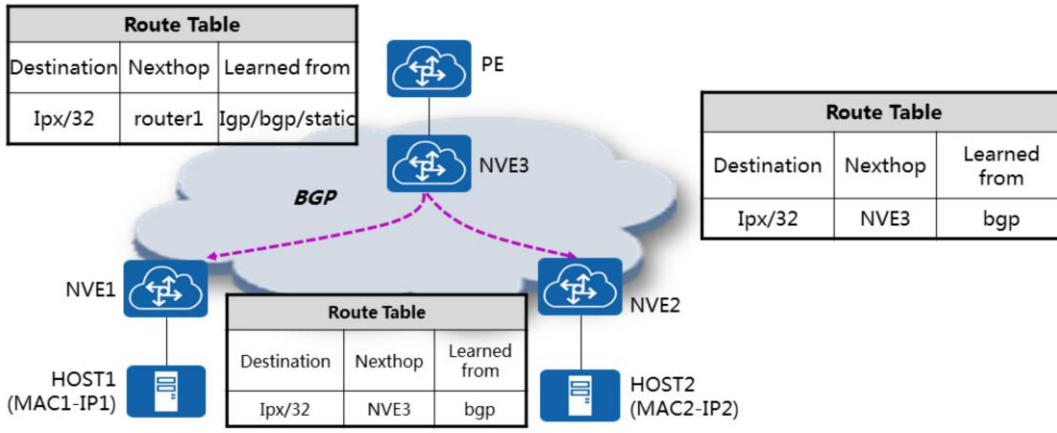
VNI	Peer
100	1.1.1.9

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- You can manually create a VXLAN tunnel by specifying the VTEP addresses and VNIs on both ends. In dynamic BGP EVPN, a VXLAN tunnel is created through Type 3 routes. The local VTEP address and VNI are contained in the Type 3 routes sent to the remote VTEP. After the remote VTEP receives the routes, it creates a VXLAN tunnel with the local VTEP and an ingress replication list of the VXLAN tunnel.

EVPN Routes - Type 5 Routes



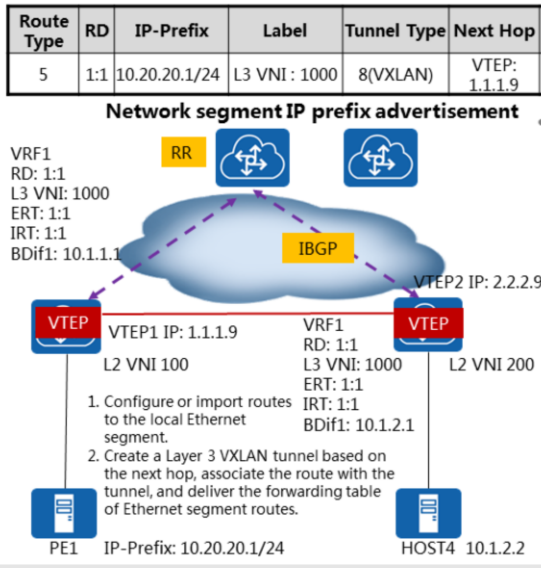
- Type 5 routes are also called IP Prefix routes.
  - This type of route is used to import subnets outside an EVPN to the EVPN and advertise host routes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Type5 routes are also called IP Prefix routes. This type of route is used to import subnets outside an EVPN to the EVPN. The subnet mask can be 32 bits. Type 5 routes are used to advertise host routes.

EVPN VXLAN - Network Segment Route Learning

Route Type	RD	IP-Prefix	Label	Tunnel Type	Next Hop	Router MAC	ERT
5	1:1	10.20.20.1/24	L3 VNI : 1000	8(VXLAN)	VTEP: 1.1.1.9	VTEP1 MAC	1:1



- EVPN uses type 5 routes to transmit Layer 3 VNIs and network segment IP routes:
  - ① The local VTEP summarizes routes of the local IP network segment and sends the routes to its neighbors through EVPN. These routes carry the Layer 3 VNI of the corresponding VRF, the local VTEP address as the next hop, and the router MAC of the local NVE.
  - ② After learning the network segment route, the remote VTEP adds the route to the corresponding VPN instance, creates a dynamic Layer 3 VXLAN tunnel according to the next hop specified in the route, associates the VPN route with the tunnel, and delivers a routing forwarding table.

IP routing table

VRF ID	IP-Prefix	L3 VNI	Router MAC	Next Hop
1	10.20.20.1/24	1000	VTEP1 MAC	1.1.1.9

Tunnel table

Tunnel	SIP	DIP
1	2.2.2.9	1.1.1.9

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Type 5 routes can be used to transmit network segment IP routes and carry the L3 VNI of the corresponding VRF.
- It can also be used to transmit an L3 VNI that represents a VRF. Then, what is an L3 VNI?
  - In the distributed gateway environment, the subnets that need to communicate each represent a VRF. However, a packet does not contain VRF information. Therefore, a VNI is mapped to the VRF through a specific VNI. Such a VNI is called L3 VNI.
  - After learning the network segment route, the remote VTEP adds the route to the corresponding VPN instance, creates a dynamic Layer 3 VXLAN tunnel according to the next hop specified in the route, and delivers a routing table.



Quiz

1. Please briefly describe the functions of Type 2 routes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Answer:
  - Type 2 routes are also called MAC/IP advertisement routes, which are used for VM migration in the distributed gateway environment.

Summary

- EVPN Overview
- EVPN Fundamentals
- EVPN Application in NVO

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

More Information

- https://datatracker.ietf.org/doc/rfc8365/
- https://datatracker.ietf.org/doc/rfc7209/
- https://datatracker.ietf.org/doc/rfc7432/

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Recommendations

- Huawei E-learning website
  - http://support.huawei.com/learning/Index!toTrainIndex
- Huawei support case library
  - http://support.huawei.com/enterprise/servicecenter?lang=zh

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

–Network Design and Product Introduction

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- IP networks have become an integral part of people's production and life.
- How to design an IP network suitable for an enterprise according to service needs of the enterprise is an important topic.
- This course introduces design principles of enterprise IP networks and best practices of enterprise IP network design.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- Upon completion of this course, you will be able to:
  - Describe basic principles of enterprise network design.
  - Describe the basic process of enterprise network design.
  - Analyze network design requirements of common enterprises.
  - Describe best practices of general enterprise network design.
  - Describe types of Huawei enterprise network products.
  - Design common enterprise networks.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

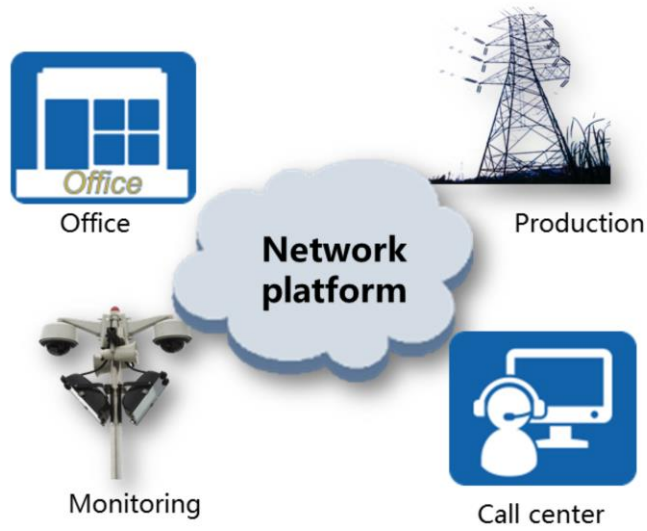


Contents

1. **Enterprise Network Design Principles**
2. Enterprise Network Design Overview
3. Introduction to Enterprise Network Products
4. Enterprise Network Design Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Enterprise Network Platform

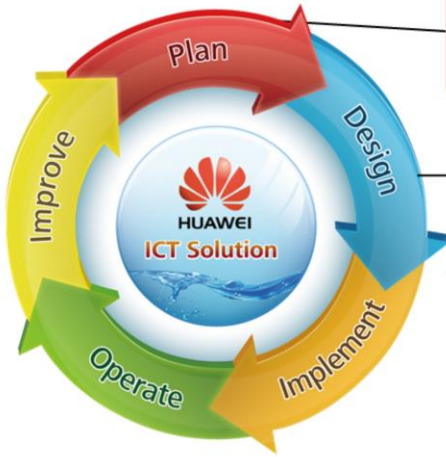


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The enterprise network is the support platform for enterprise services and the information center for enterprises.
  - An enterprise may have many services, such as office, production, monitoring, and customer service (call center).
  - Different services are connected through the enterprise network platform, so that enterprises can operate efficiently.
  - The network is the basic platform that supports enterprise services.
  - The design of enterprise networks is very important for enterprise services.

Network Lifecycle



- Organization policy
- Service policy
- Finance decision-making
- High-level design (HLD)

- Requirement analysis
- Project plan
- Model selection
- Low-level design (LLD)

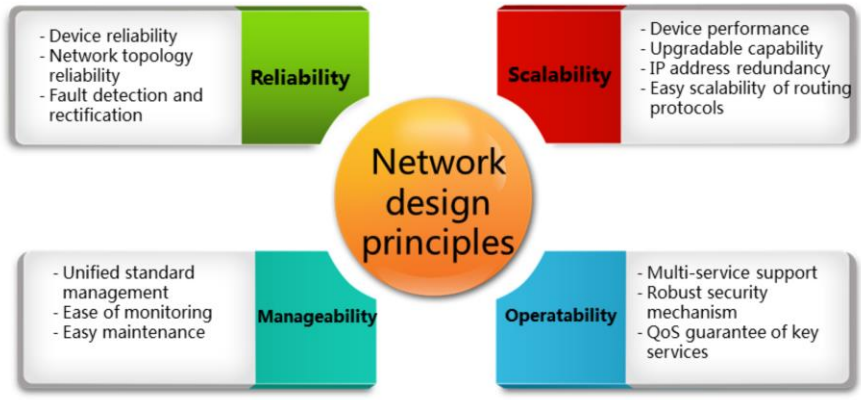
- Network design plays an important role in the entire network lifecycle:
  - It covers the preparation phase and design phase.
  - Good network design is the prerequisite for smooth network implementation and normal operation.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Plan: Help enterprise users plan enterprise networks, information security, unified communications, and IT systems based on long-term objectives of enterprises, thereby effectively managing and monitoring enterprise network risks and system problems.
- Design: Based on the understanding of the existing network, system, and application, make detailed design to meet enterprise users' requirements for current technologies and services, and support the capabilities, reliability, security, scalability, and performance of enterprises in the IT and service domains.
- Implement: Help enterprise users develop, install, and test services, networks, and IT systems based on design specifications to meet customers' service and technical requirements.
- Operate: Help enterprise users maintain continuous and healthy service operations, proactively monitor and manage the system, and maximize the performance, capacity, availability, reliability, and security of system devices and application systems.
- Improve: Help enterprise customers continuously improve service quality, regularly update and upgrade services based on customer service requirements, and implement targeted policies and operations to continuously improve services.

Basic Principles of Network Design

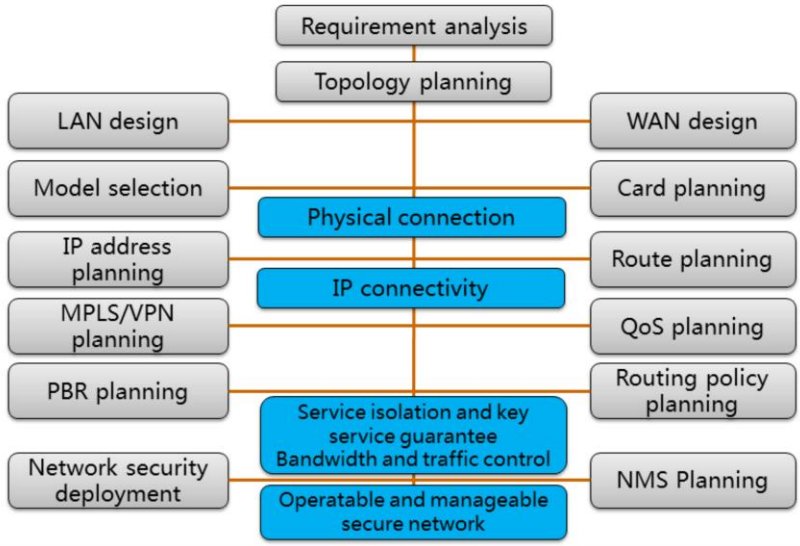


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Basic principles of network design:
  - Reliability: When a fault occurs on a network, the services carried on the network are not interrupted.
  - Scalability: The network can support the increasing service volume and facilitate capacity expansion.
  - Operatability: The network must support multiple services and provide secure and hierarchical service assurance.
  - Manageability: The network is required to provide standard management methods to facilitate monitoring and maintenance.
- In addition, the cost must be considered during network design. We should select the most cost-effective design solution when service requirements are met.

Enterprise Network Design Process

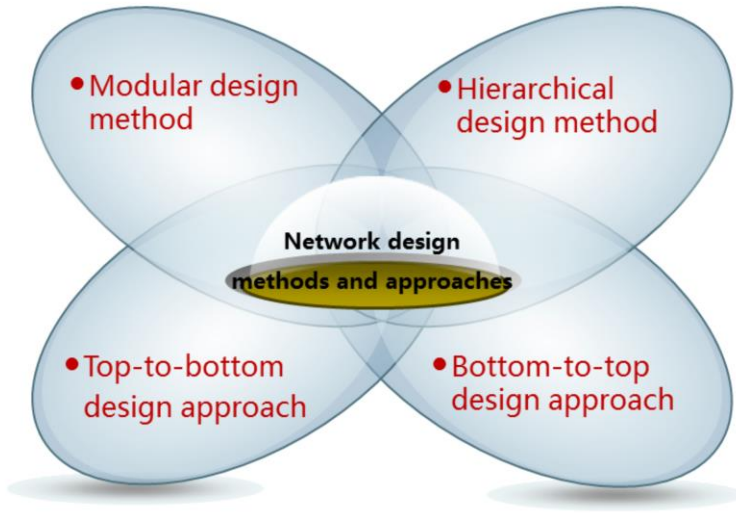


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The network design includes multiple modules. According to the service requirements of the enterprise, not every module is required.

Network Design Methods and Approaches

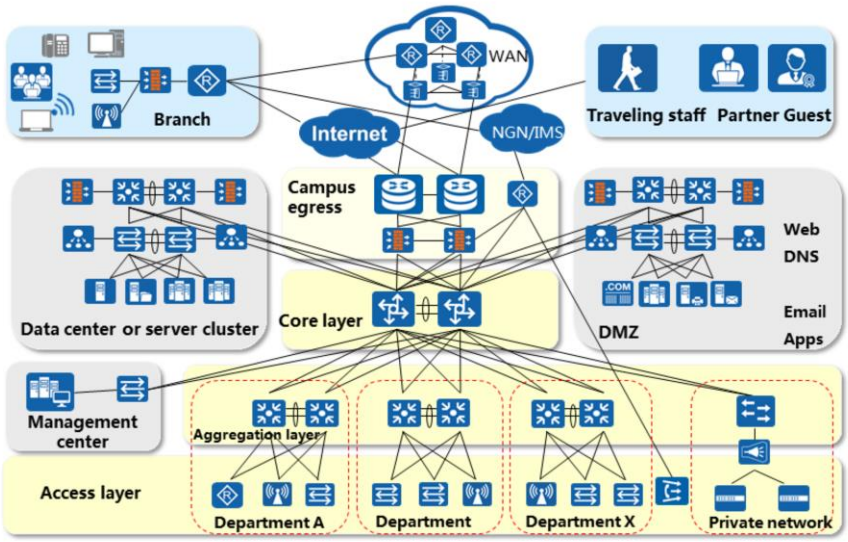


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Common network design methods and approaches:
  - Modular design method
  - Hierarchical design method
  - Top-to-bottom design approach
  - Bottom-to-top design approach

Modular Network Design

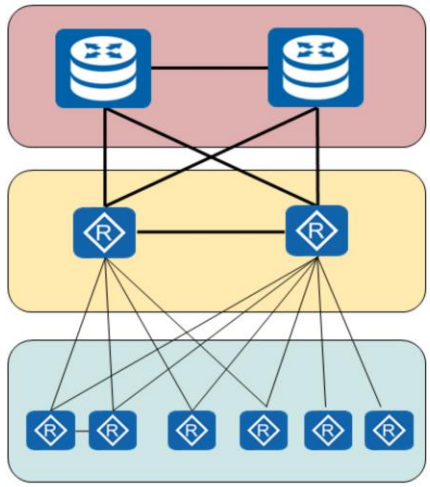


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A large enterprise network can be divided into different modules based on functions. Different modules have different requirements and features. The common classification is as follows:
  - Campus network: all LANs of the enterprise headquarters
  - DMZ network
  - WAN
  - Data center network
  - Branch network
  - Remote user network
- The modular network design has the following advantages:
  - The functions are independent of each other. Each module can be designed separately.
  - This facilitates capacity expansion. For example, adding a module does not affect the entire network.
  - Management is easy. For example, different security policies are defined for different modules.
- In actual deployment, the egress router of the campus network is often integrated with the IP PBX function.

Hierarchical Network Design



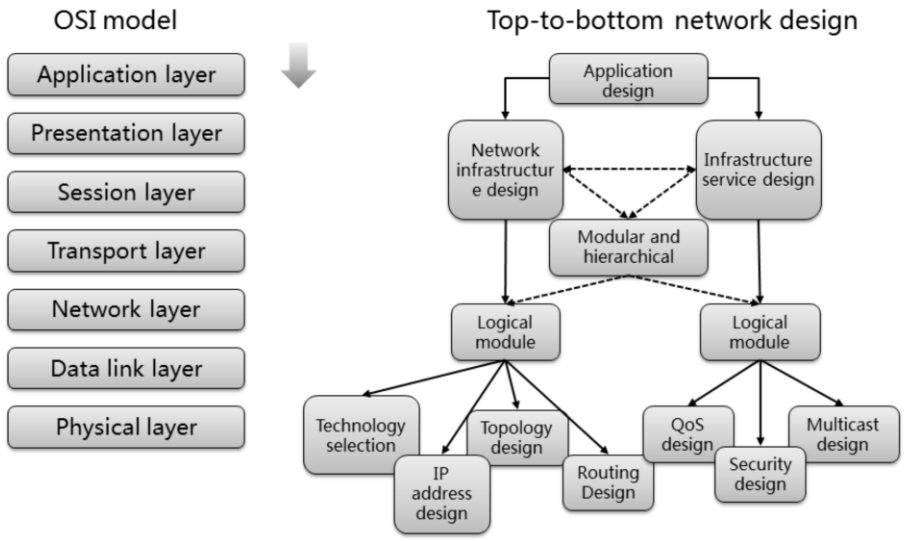
- Core Layer
  - High-speed data forwarding
  - High reliability
- Aggregation Layer
  - Route summarization
  - Traffic convergence
- Access Layer
  - Terminal access
  - Access control

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Hierarchical network design brings the following benefits:
  - Devices at different levels can be used at different layers to reduce costs. Devices at different levels can be used at different levels.
  - Easy to understand: Hierarchical network functions are differentiated, which is easy to understand.
  - Modular: The hierarchical architecture is also a modular design, which facilitates design, management, and expansion.
  - Fault isolation: The layered structure can effectively control the impact scope of a fault.

Top-to-Bottom Design Approach



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The top-to-bottom design is based on the application layer of the OSI model. The network needs to support upper-layer applications.
- The top--to-bottom design is to analyze the application requirements first, and then design the network architecture and basic services from the application requirements.
  - For example, if an enterprise needs to deploy a video conference system (using multicast), the network topology, multicast, and QoS must be considered during network design.

Top-to-Bottom and Bottom-to-Top

Design Approach	Advantage	Disadvantage	Usage Scenario
Top-to-bottom	• Application requirements can be well supported from the perspective of applications. • Current and future requirements can be met from an overall perspective .	• Network designers invest more time and energy.	• Large-sized new network
Bottom-to-top	• Based on experiences • Quick implementation	• Overall requirements are not considered. • Long-term requirements may be not met.	• Small-sized new network • Network reconstruction

- Top-to-bottom and bottom-to-top design approaches are not contradictory:
  - They are applied to different scenarios.
  - They can also be considered in the same network project.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The top-to-bottom design approach corresponds to the bottom-to-top design approach.
- The bottom-to-top design approach does not analyze specific application requirements from the service perspective. Instead, it designs networks based on experiences.
  - For example, when an office network is expanded, the network architecture remains unchanged. Only access switches are added.



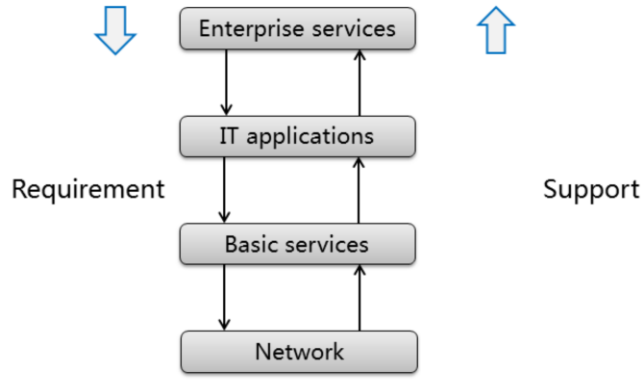
Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - User Requirement Analysis
   - Topology Design
   - LAN Design
   - WAN Design
   - Network Device Selection
   - IP Address Planning and Device Naming
   - Routing Design

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

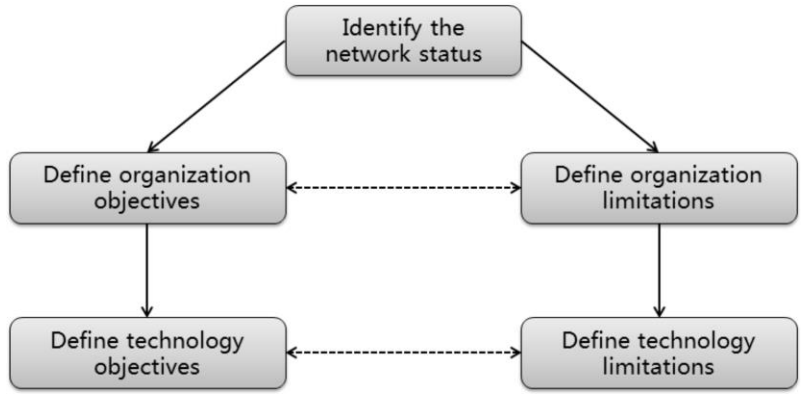
Why Do We Need to Analyze User Requirements



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- For common enterprises, the enterprise network is a technical platform that supports the development of enterprise services.
- Therefore, the network to be constructed must be considered based on enterprise services.

How Do We Analyze User Requirements



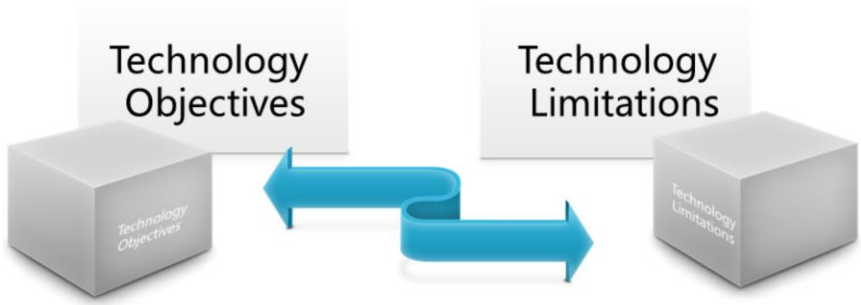
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The first step in analyzing user requirements is to identify the network status. The specific methods are as follows:
  - Querying documents
  - Consulting parties
  - Network monitoring
  - Traffic analysis
- The second step in analyzing user requirements is to define organizational goals. Network design is of no use if organizational goals are not followed. Common organizational objectives are as follows:
  - Improving customer satisfaction
  - Adding service items
  - Enhancing competitiveness
  - Lowering costs
- As opposed to organizational goals, defining organizational constraints is essential because organizational constraints directly affect network solution design. Common organizational constraints are as follows:
  - Budget
  - Labor

- Policy
- Time arrangement

Technology Considerations



- Network capacity expansion
- Simplified network management
- Improved network security
- Enhanced network reliability
- ...

- The current hardware and software do not support new technologies and applications.
- Device reuse
- Compatibility of original applications
- ...

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Common technical objectives:
  - Increase the network throughput.
  - Reduce the service interruption time.
  - Simplify network management.
  - Enhance network security.
  - Improve reliability of key services.
- Define technical constraints:
  - Cables on the live network do not support new technologies.
  - The bandwidth on the live network cannot support new applications.
  - The network must be compatible with old devices.
  - Old applications must be supported.

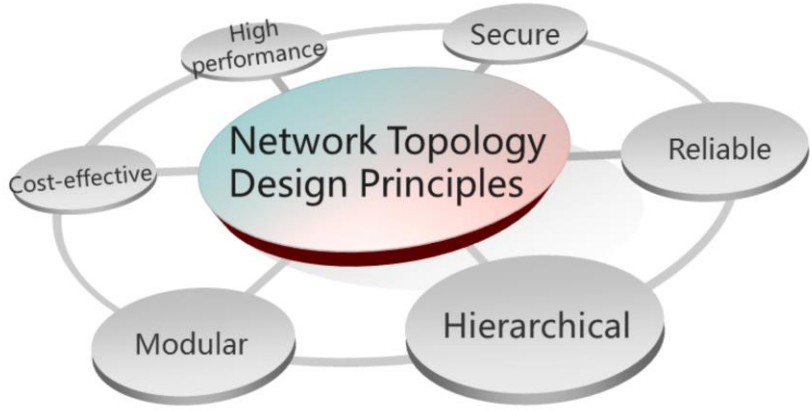


Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - User Requirement Analysis
   - Topology Design
   - LAN Design
   - WAN Design
   - Network Device Selection
   - IP Address Planning and Device Naming
   - Routing Design

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Network Topology Design Principles

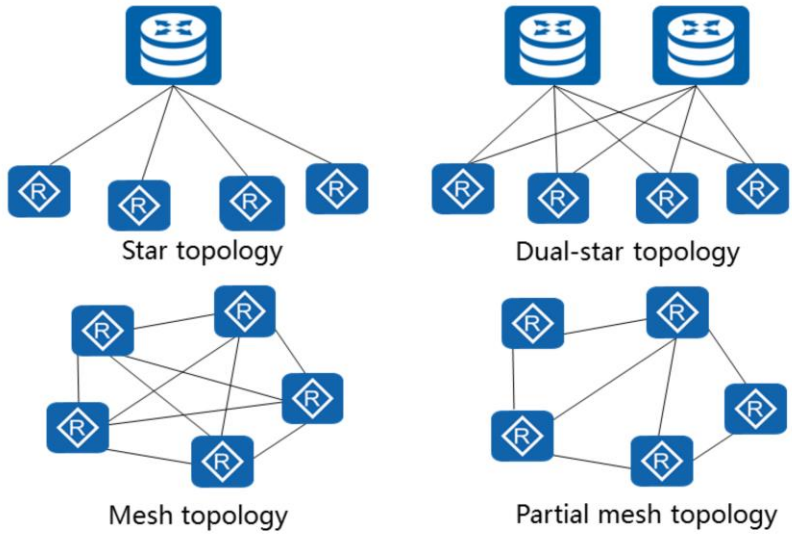


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Modular
  - As mentioned earlier, the modular design simplifies network design, and network management and expansion.
- Hierarchical
  - Hierarchical design reduces costs and isolates network faults.
- Reliable
  - Network reliability reduces the impact of single-point failures on the network, implements load balancing, and improves network performance.
  - However, increasing network reliability often increases network complexity and costs.
- Secure
  - Core routers, boundary points, switches, and servers need to be protected..
  - The firewall protects the network against external attacks.
- High performance
  - No performance bottleneck exists in the network topology design.
- Cost-effective
  - The cost is considered.

Common Network Topologies

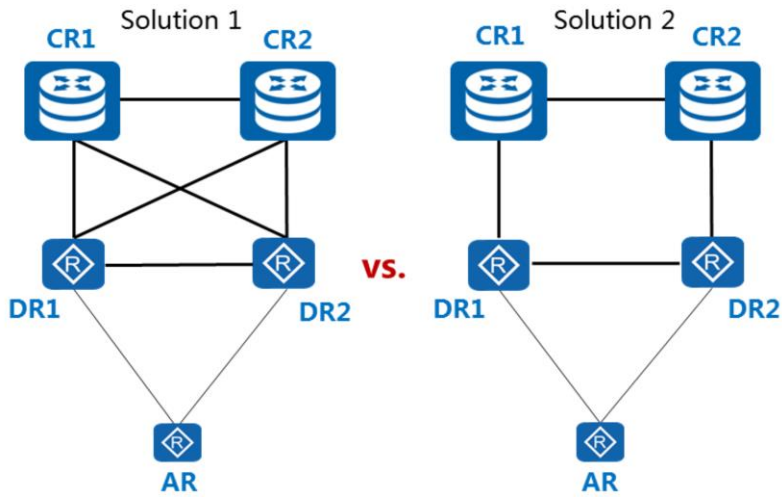


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Few enterprise networks use only the topology structure. In most cases, multiple topologies are used.
  - The star topology is used when the lower-level network is connected to the upper-level network.
  - In the case of network interconnection at the same level, a mesh topology or a partially mesh topology is used.

Case Study: Which Topology Is Better



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- When aggregation devices are interconnected with core devices, there are two popular topologies.
  - Solution 1:
    - The aggregation router is connected to two core routers, but the two aggregation routers are not connected.
  - Solution 2:
    - The aggregation router is connected to a core router and another aggregation router.
- The two solutions ensure that no single point of failure occurs on the access router. Which solution is better?
- Technical consideration:
  - If a node or link on the network is faulty, is the data forwarding path from the access layer to the core layer optimal?
  - Solution 1: No matter which single point of failure occurs, data is transmitted from the access layer to the core layer through three hops.
  - Solution 2: If the link between CR1 and DR1 is faulty, data may reach the core layer through four hops (AR -> DR1 -> DR2 -> CR2).
  - From a technical perspective, solution 1 is better.
- Cost consideration:
  - If the core layer is far away from the aggregation layer, the cost of solution 1 is twice that of solution 2.
  - From a cost perspective, solution 2 is better.
- Comprehensive consideration:
  - In the LAN, the line cost is low. Solution 1 or the combination of the two solutions can

be used.

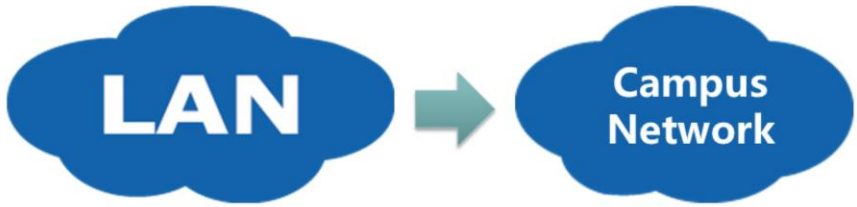
- On the WAN, solution 2 is recommended because the line cost is high. In addition, the optimal sub-optimal route is avoided in routing design.

Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - Topology Design
   - LAN Design
   - WAN Design
   - Network Device Selection
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

LAN and Campus Network



A Local Area Network (LAN) is a computer network covering a local area from the technology perspective.

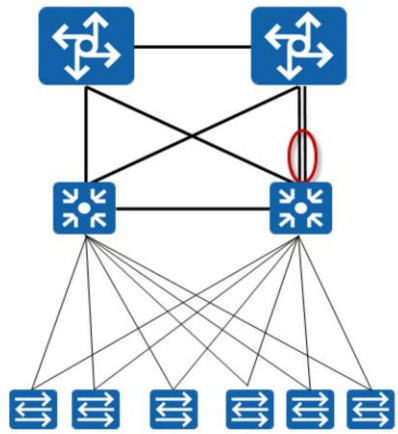
A campus network is a computer network inside an enterprise campus from the service perspective.
The campus network is also a LAN.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- From the technical perspective, a LAN is a computer network covering a local area.
  - Modern LANs often use only single link-layer technology, that is, Ethernet technology.
  - Therefore, Ethernet technology is a recognized LAN technology standard.
- From the service perspective, a campus network is a computer network inside an enterprise campus.
  - A campus network is also a LAN or a combination of multiple LANs.
  - Therefore, the LAN and campus network can be discussed together.

Technology Application on the Hierarchical Network



Layer 3 routing:
- Cross-connect
- Link aggregation

Layer 2 switching network
- VLAN
- Spanning Tree Protocol, Smart Link, CSS, and iStack
- VRRP

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Best practice at the access layer
  - Do not assign VLANs on access switches. Connect the same services to the same access switch and use the same VLAN.
  - Use RSTP or MSTP to prevent loops, and enable the edge port function on the port connected to hosts.
  - Connect an access interface to a host, and connect a trunk interface to an aggregation switch and configure the trunk interface to allow packets from the specified VLAN (do not allow all VLANs to pass through).
  - For the dual-uplink topology, the Smart Link technology can be used.
- Best practice at the aggregation layer
  - Use VRRP to provide gateway redundancy for hosts.
  - Deploy Layer 3 routes between aggregation land core devices to implement load balancing and fast convergence.
  - Aggregation and core devices use the full-mesh topology but not the square-shaped topology.
  - When an aggregation switch advertises routes, use route summarization first.
  - Layer 2 interconnection is used between aggregation and access devices.
  - Use link aggregation technology for important links to increase bandwidth and improve redundancy.

Several Common LANs

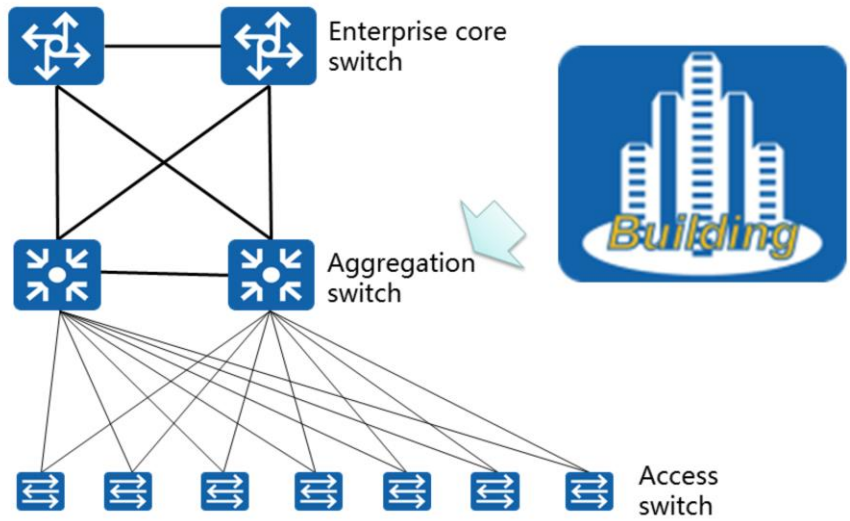
- Building LAN
- Campus network
- DC LAN
- Medium-sized LAN
- Small-sized LAN

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The following describes common LANs.

Building LAN

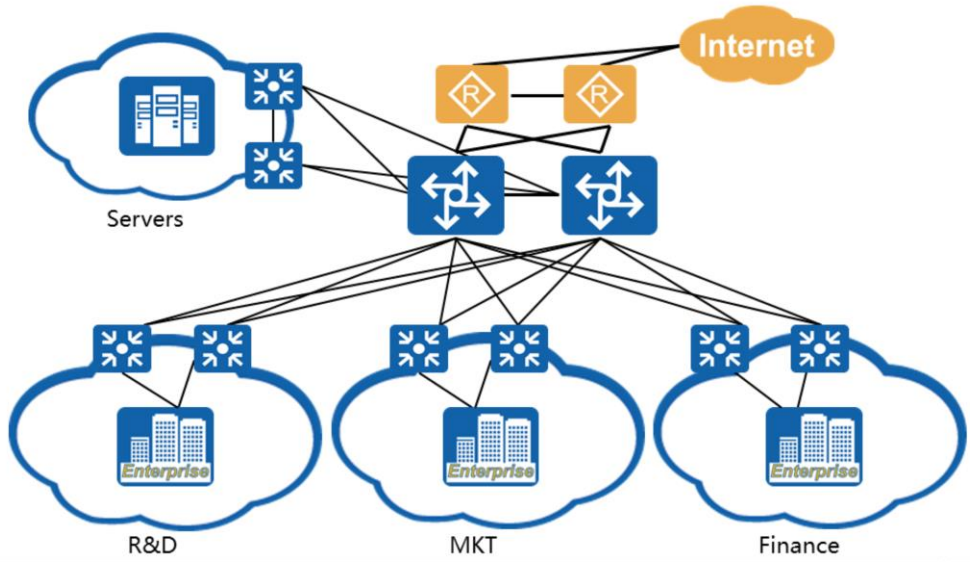


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The building LAN is the most typical LAN.
  - Generally, the building LAN is divided by floor or department.
  - Access switches connect to one or more floors or departments.
  - The aggregation switch is connected to multiple access switches.
  - Building aggregation switches are connected to other modules (such as data centers) on the enterprise campus network through core switches.
  - In most cases, Layer 2 interconnection is used between the access layer and aggregation layer to reduce costs.
  - Typically, an access switch is dual-homed to two aggregation switches. RSTP or MSTP is run to prevent loops.
  - Layer 3 interconnection is used between the aggregation layer and core layer to implement fast convergence and load balancing.
  - Policies can be used on the aggregation switch to control routes.
  - In small buildings, the core layer and aggregation layer may be combined.

Enterprise Campus Network

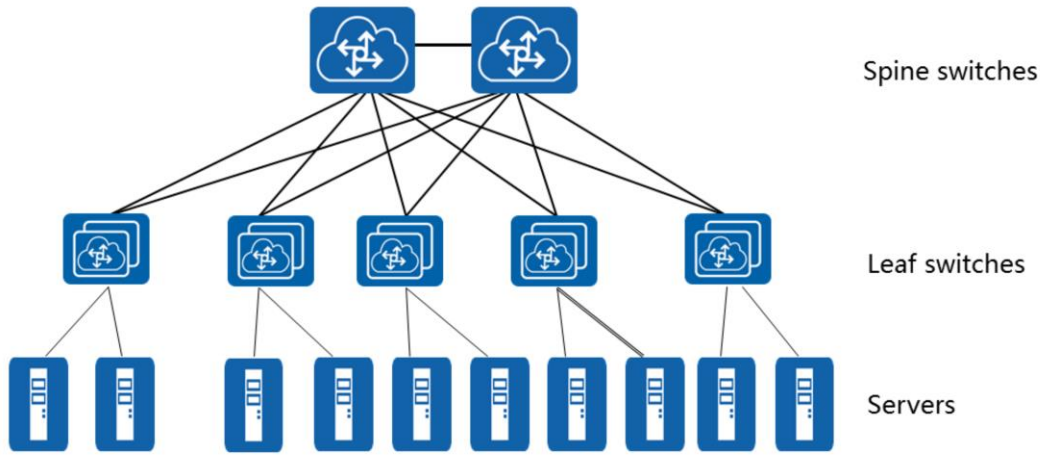


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An enterprise campus network can be regarded as the interconnection between multiple building LANs.
  - The enterprise campus network uses high-speed links for interconnection. If the network is newly constructed, it is recommended that links with 10 Gbit/s or higher bandwidth be used.
  - The physical distance between networks is not very long (generally within several kilometers), so infrastructures such as links are usually built by enterprises.
  - On a campus network, aggregation switches in all buildings can be connected to two core switches that back up each other.
  - A large-scale campus network can have more than two core devices. The core layer uses ring, partial full-mesh, or full-mesh topology.
  - During address planning, a segment of contiguous IP addresses is used in a building to facilitate route aggregation.
  - At the access layer, security needs to be considered. It is recommended that NAC be used to authenticate and authorize access users.

DC LAN

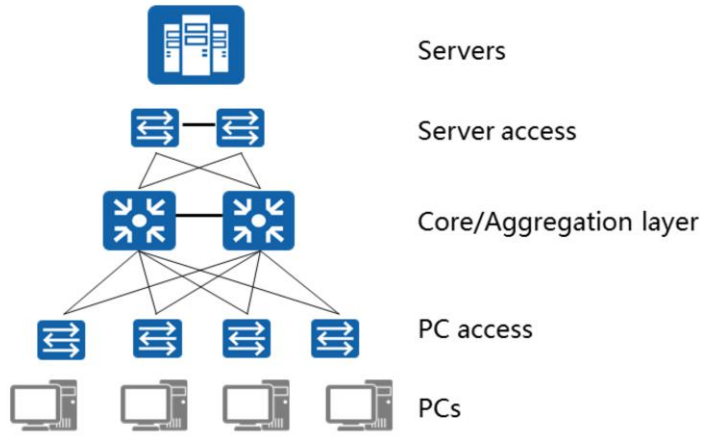


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The differences between the data center LAN and the common LAN are as follows:
  - As the enterprise information center, the data center LAN has heavy external traffic. Therefore, high-performance switches are used.
  - Due to the use of technologies such as server clusters and VMs, the volume of internal traffic on the data center LAN is large. Therefore, the total bandwidth is gradually lowered from the access layer to the aggregation layer and to the core layer. However, the data center requires no convergence.
  - To implement features such as load balancing and redundancy, servers provide multiple network adapters, and access switches must support features such as link bundling.
  - Some new technologies, such as FCoE, require the support of switches.
  - To meet application requirements such as VM migration, the data center uses Layer 2 networking.
  - To address loops on a large-scale Layer 2 network, switch virtualization technologies such as CSS, stacking, M-LAG, and SVF can be used.

Medium-sized LAN

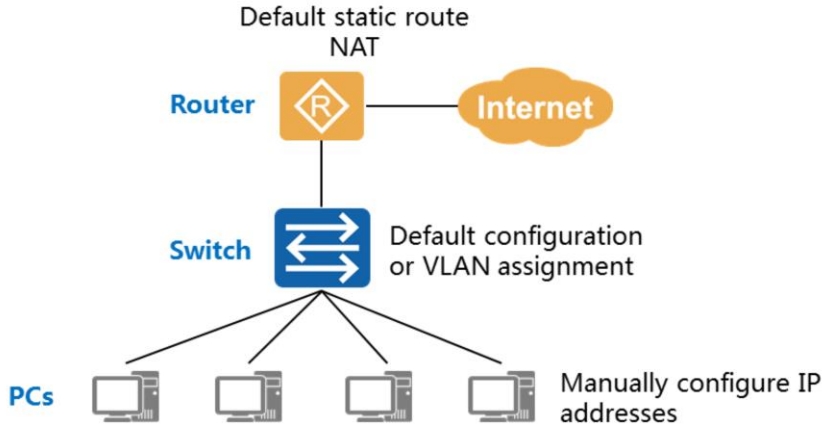


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The architecture of a medium-sized LAN is similar to that of a building LAN.
  - The medium-sized LAN has a small scale. Functions of the core layer and aggregation layer are combined on a group of devices.

Small-sized LAN

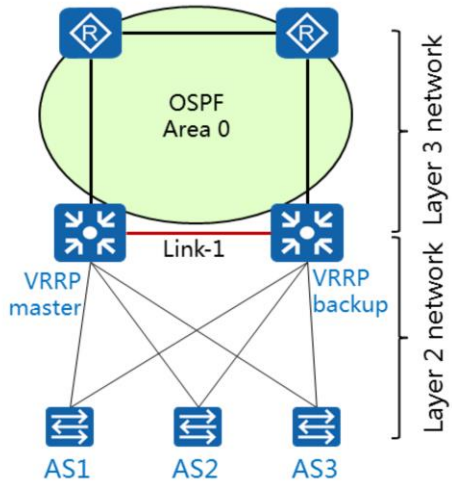


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The small-sized LAN has the simplest architecture.
  - Typically, Layer 2 switches are connected to downlink hosts and uplink egress router.

Case Study: VRRP+STP Networking



Question:
A Layer 3 network is deployed between core and aggregation devices. Considering the cost, core and aggregation devices are connected in square-shaped mode, OSPF is used, and routes of service network segments are advertised by importing routes.
If a Layer 2 network is deployed between aggregation and access devices, is Link-1 a Layer 2 link or Layer 3 link?

If Link-1 is a Layer 3 link, what will happen?

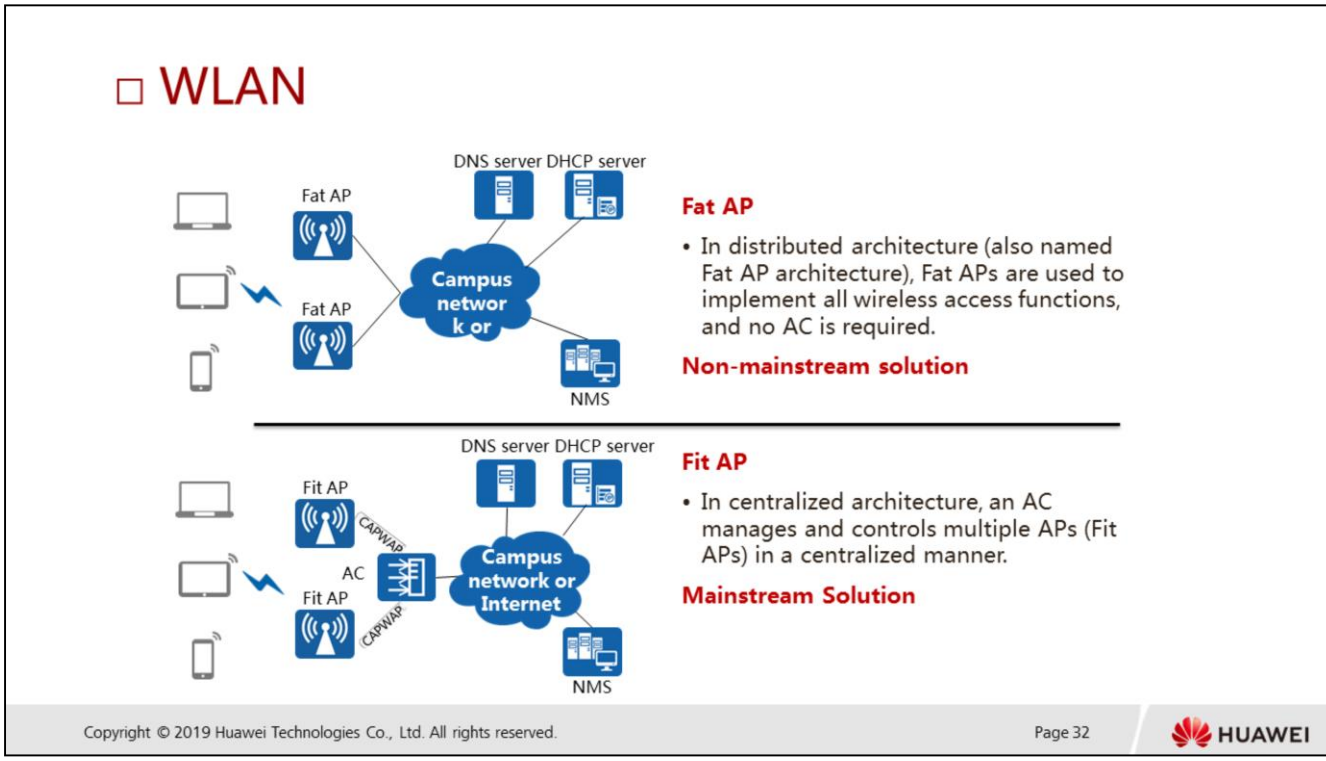
If Link-1 is a Layer 2 link, what will happen?

Is there a good comprehensive solution?

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- If link 1 is a Layer 3 link, what will happen?
  - The root switch and secondary root switch of STP are configured on the master and backup switches, respectively. If link 1 is a Layer 3 link, all interfaces on an access switch will not be blocked by STP (it is difficult to determine this switch, for example, AS1). However, one interface on other access switches is blocked by STP.
  - In this way, AS1 becomes a key node on the network, and all VRRP traffic passes through AS1. If AS1 is faulty, the entire network below the convergence layer flaps.
- If link 1 is a Layer 2 link, what will happen?
  - Similarly, the root switch and secondary root switch of STP are configured on the master and backup switches, respectively. If link 1 is a Layer 2 link and allows all VLANs used by the access layer, all the interfaces of access switches connected to the backup switch are blocked by STP.
  - However, if link 1 is a pure Layer 2 link, the Layer 3 network between core and aggregation layers is a chain topology. Any device or link fault may cause OSPF area 0 to split.
- The comprehensive solution is as follows:
  - Link 1 is a Layer 2 link that allows all VLANs used by the access layer.
  - Enable a VLAN and create a VLANIF interface on the two aggregation switches for establish an OSPF neighbor relationship between them. In this way, the Layer 3 network uses a ring architecture and has redundancy.
  - Considering the importance of link 1, link bundling can be used to enhance reliability.
  - You can also use MSTP and deploy master devices of multiple VRRP groups to achieve load balancing.



- Fat AP
  - In distributed architecture (also called FAT AP architecture), Fat APs are used to implement all wireless access functions, and no AC is required.
  - The distributed architecture was widely applied on WLANs in early days. With an increasing number of deployed APs, the management work such as AP configuration and software upgrade brings high costs. Therefore, this architecture is applied less now.
  - It is not the mainstream solution.
- Fit AP
  - In centralized architecture (also called Fit AP architecture), an AC manages and controls multiple Fit APs in centralized manner.
  - In the centralized architecture, the AC and APs implement wireless access. The centralized architecture is the mainstream architecture of enterprise WLANs and carrier WLANs because it allows for centralized management, authentication, and security management. This solution is a general solution for the enterprise network.
  - It is the mainstream solution.
- WLAN design requires professional knowledge and tools. We have a professional course to introduce the WLAN design.



Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - LAN Design
   - WAN Design
   - Network Device Selection
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Enterprise WAN



- An enterprise WAN is an enterprise private network that uses leased carrier links or self-built links to interconnect nodes such as the headquarters, DCs, branches, offices, and mobile offices.
- The enterprise WAN refers to the enterprise private network constructed by enterprises or leasing links.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

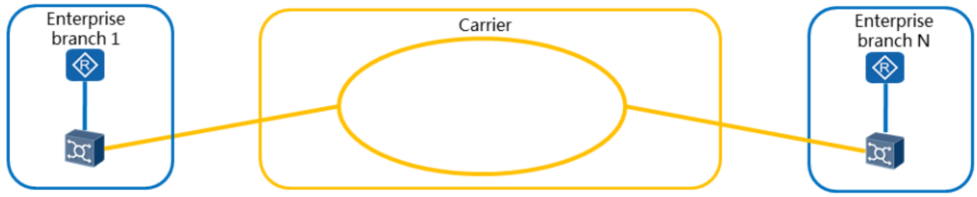
- Enterprise WAN = Egress border of the enterprise network + WAN link leased from the carrier or self-built line

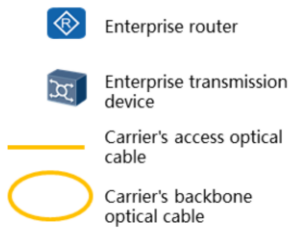
WAN Design Roadmap

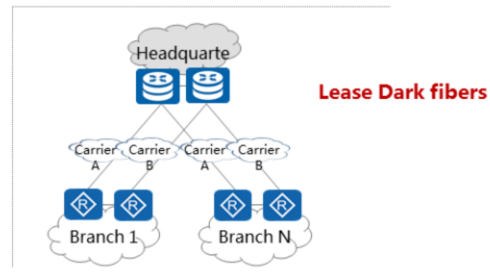
- Multi-level enterprise WAN
  - The multi-level management structure of enterprises determines the multi-hierarchy structure of enterprise WANs. The service bearer, topology model, and access mode between network nodes at different levels may be the same or different.
- Various network topology models
  - Different topology models apply to different traffic models.
- Service bearer modes
  - IP and VPN bearer modes are available and different WAN interworking or isolation policies are used.
- WAN solution roadmap
  - Technical and industry solutions are available. The technical solution focuses on WAN interconnection between two levels of network nodes. The industry solution is targeted for specific industries (such as electric power and government) or scenarios (such as large enterprises).
  - Technical solution: Scenarios are classified based on WAN interconnection technologies, for example, SDH/MSTP/WDM private line scenarios.
  - Industry solution: Scenarios are classified by industry. The technologies used in different industry solutions can be the same.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Common Enterprise WAN Deployment Technology - Leasing Dark Fibers





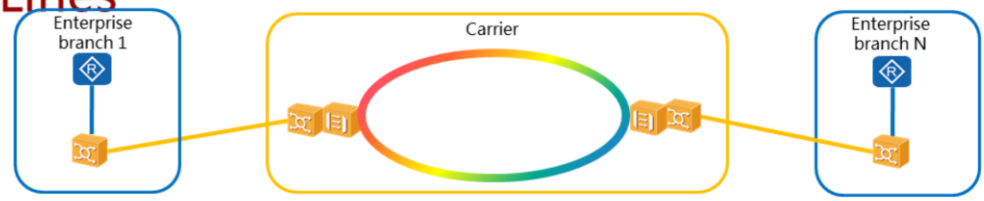


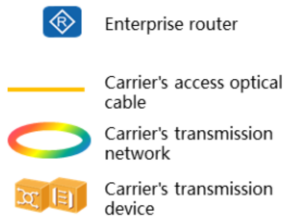
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

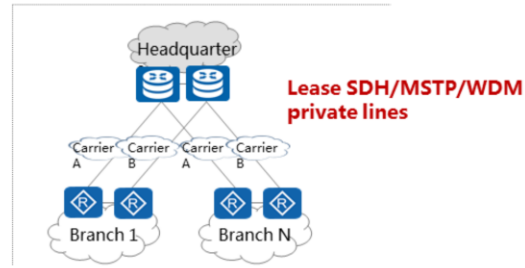
HUAWEI

- Private line types are classified based on the lease range.

Common Enterprise WAN Deployment Technology - Leasing SDH/MSTP/WDM Private Lines



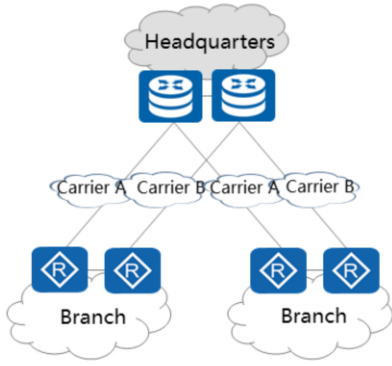




Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Here, MSTP refers to Multi-Service Transport Platform.

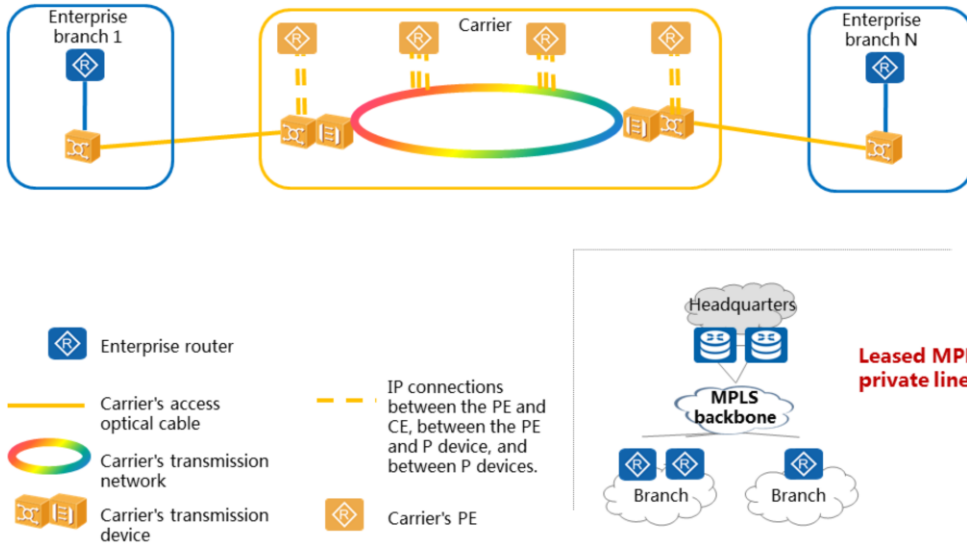
Architecture Based on the SDH/MSTP/WDM Private Line



- Application Scenario
  - Enterprises demanding high quality of WAN lines
  - Enterprises that have high requirements on communication security and privacy
  - Enterprises that want to have independent backbone networks but do not have the strength to build their own networks
- Advantages
  - The private line is divided into physical timeslots. The latency is small and the privacy is good.
  - High-quality service bearing and QoS guarantee
  - Accurate bandwidth (network cable extension)
- Disadvantages
  - The cost of the SDH/MSTP/WDM private line is high, which is only lower than that of dark fiber private line.
  - The provisioning period is long.
- Situation
  - The WAN bearer network of large enterprises in China is widely used.
  - Financial enterprises widely use the SDH/MSTP/WDM private line.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Common Enterprise WAN Deployment Technology - Leasing MPLS VPN Private Lines

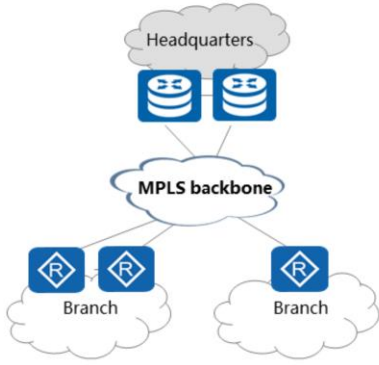


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Multi-Service Transport Platform (MSTP)
- The carrier's transmission devices in the enterprise equipment room are optional. Determine whether to deploy the solution based on access optical cables and service access of the enterprise equipment room.

Architecture Based on MPLS VPN Private Lines



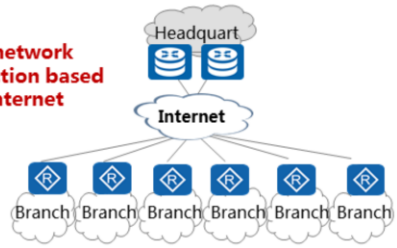
- Application Scenario
  - Low requirements on the WAN line quality
  - High-bandwidth private line access between the headquarters and branches
  - High cost of SDH/MSTP/WDM private lines
- Advantages
  - Carriers' Layer 3 VPNs are isolated, and enterprises do not need to pay attention to route deployment.
  - Flexible mutual access or isolation control between sites
  - The cost is much lower than that of dark fibers and SDH/MSTP/WDM private lines.
- Disadvantages
  - The carrier VPN service is leased. Different enterprises are isolated from each other at Layer 3. The carrier controls the mutual access relationship, and the privacy depends on the carrier.
  - QoS assurance depends on carriers.
- Situation
  - Almost all carriers can provide MPLS VPN private line services.
  - The deployment volume of MPLS VPN private lines increases rapidly.

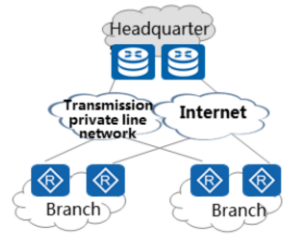
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Common Deployment and Application of Enterprise WANs

DSVPN network construction based on the Internet





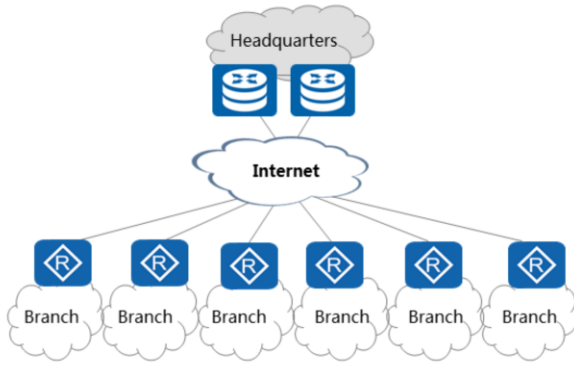
Highly reliable network: SDH/MSTP/WDM private line + overlay VPN backup

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- This page displays abstract types of WAN topologies.

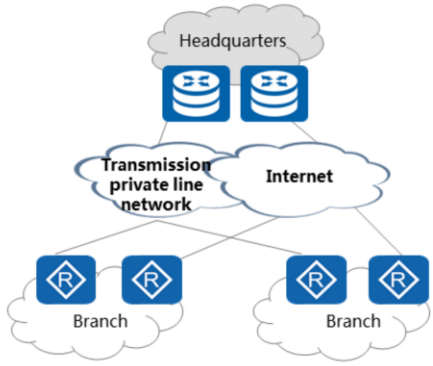
DSVPN Network Based on the Internet



- Application Scenario
  - The number of branches is large and dynamic expansion is required.
  - There are many service interactions between branches.
  - Low requirements on link quality
  - Sensitivity to network O&M costs
- Solution Characteristics
  - Traffic exchange between the headquarters and branches is implemented based on the Internet DSVPN solution.
  - The branch does not require a static public IP address.
  - It is easy to expand branches, and the headquarters is easy to deploy.
- Situation
  - Enterprise customers have little knowledge about DSVPN.
  - DSVPN becomes the commonly used overlay VPN technology, and the deployment is increasing.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Highly Reliable Network: SDH/MSTP/WDM Private Line + Overlay VPN Backup



- Application Scenario
  - Enterprises that require active/standby links
  - Enterprises that have high requirements on the quality of the primary link and have good privacy
  - Both the headquarters and branches can access the Internet.
- Solution Characteristics
  - The SDH/MSTP/WDM private line provides a high-quality and private primary transmission channel. In addition to Internet access requirements, the overlay VPN tunnel running on the Internet link is used as the backup link.
  - Overlay VPN includes DSVPN, IPSec VPN, and GRE VPN.
  - The combination gives full play to favorable conditions.
- Situation
  - Widely deployed solution

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI



Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - WAN Design
   - Network Device Selection
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet
   - QoS Design

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Common Network Devices

- Common network devices include routers, switches, hubs, and firewalls.
- Device selection here refers to selection of device types but not device models, for example, routers, Layer 2 switches, and Layer 3 switches.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Common network devices
  - The switch is a mainstream device on a LAN.
  - The router achieves WAN interconnection of an enterprise network and functions as the edge device.
  - The hub is not used.
  - Firewalls are playing an important role at the enterprise edge or in important zones.
  - In addition, some devices, such as the LB and IPS, have their professional functions and the deployment does not involve the change of the network topology. Therefore, these devices are not described here.
- Trend
  - Convergence of routing and switching: The Layer 3 switch provides the routing function, and the switching router supports the switching module.
  - Integration of VAS functions: More and more network devices support additional functions such as firewalls. For example, Huawei AR G3 routers support firewall functions, and S7700 switches provide functions such as firewall and AC when being equipped with specific cards.

Network Device Comparison

Device	Description	Price	Usage Scenario
Router	A Layer 3 device that uses the CPU architecture, forwards packets based on the routing table, and supports various interface types.	High	WAN and LAN edge
Layer 3 switch	A Layer 2 or Layer 3 device. Layer 2 traffic is forwarded based on the MAC address table, and Layer 3 traffic is forwarded based on the routing table. It supports only the Ethernet.	Medium	LAN aggregation layer and core layer
Layer 2 switch	A Layer 2 device that forwards packets based on the MAC address table. It supports only the Ethernet.	Low	LAN access layer
Firewall	A Layer 4 device that forwards packets based on session tables. It can work in Layer 2 or Layer 3 mode. It supports only the Ethernet.	High	Network edge and edge of the network area

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Devices are selected based on service requirements, considering device functions and prices.
  - Layer 2 switches are recommended at the access layer of an enterprise network.
  - Layer 3 switches are recommended at the aggregation layer and core layer.
  - Routers are recommended for the WAN of the enterprise network.
  - The router or firewall is recommended for the egress of the enterprise network.

Network Device Model

Device	Description	Usage Scenario	Example
Fixed device	The device interface type and quantity are fixed. Generally, the appearance is small. Users can select devices of different models as required.	Network access layer Core or aggregation layer of small-sized networks	S3700 series switches
Modular devices	The device consists of the chassis and optional modules. The appearance of the device is large. Users can flexibly select modules (interface type, number of interfaces, and supported functions) based on site requirements.	Core or aggregation layer	S7700 series switches



S3700 series switches



S7700 series switches

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Devices fall into fixed and modular ones.
  - Fixed devices can be added to a stack (for example, Huawei S5700 switch) to increase interface bandwidth, simplify management, and improve reliability.
  - Modular switches (such as Huawei S9700 switch) can also form a cluster through CSS.
  - Modular switches can be virtualized into multiple logical devices through VSs (such as Huawei CE12800 switches).



Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - Network Device Selection
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet
   - QoS Design
   - Network Security Deployment

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

IP Address Planning Rules

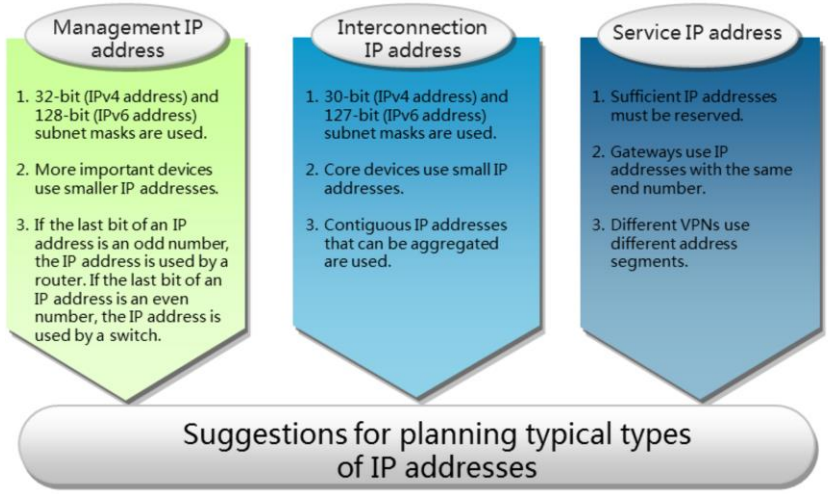


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Unique:
  - Hosts on an IP network must use unique IP addresses. Allocate different IP addresses to hosts even they support MPLS VPN address overlapping.
- Contiguous:
  - The routes with contiguous IP addresses can be aggregated easily on a hierarchical network. This reduces routing table size and speeds up route calculation and route convergence.
- Scalable:
  - IP addresses need to be reserved at each layer to ensure continuity of addresses when the network is expanded.
- Meaningful:
  - A well-planned IP address denotes the role of the device to which the IP address belongs. IP address planning is an art of skill. An ideal way is to use a formula as well as related parameters and coefficients to calculate every IP address.

Suggestions for Planning Important IP Addresses



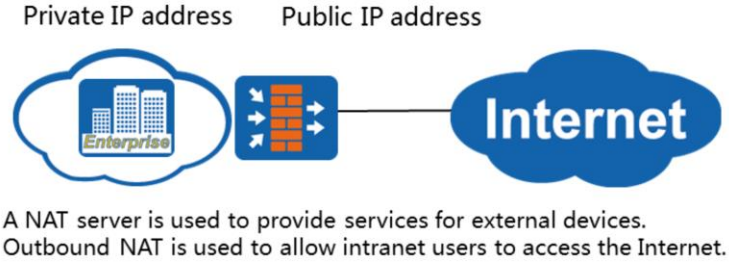
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Generally, typical IP addresses as mentioned above are involved in IP address planning. Although there is no mandatory standard, some experiences in the industry are available, as described above.

Public and Private IP Addresses

- On the Internet, public IP addresses become insufficient.
  - Available IPv4 addresses have been allocated.
  - Enterprises have to pay to use public IP addresses.
- Private IP addresses can be used randomly within an enterprise.



A NAT server is used to provide services for external devices.
Outbound NAT is used to allow intranet users to access the Internet.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Private IP address
  - An enterprise usually uses private IP addresses, that is, those on the network segments 10.0.0.0/8, 172.16.0.0/12, and 192.168.0.0/16.
  - Although private IP addresses can be used randomly, they must be planned.
- Public IP address
  - If an enterprise does not provide services to external devices but needs to access the Internet, the enterprise can use dynamic public IP address allocated by carriers.
  - If an enterprise provides services to external devices (for example, the self-built WWW service), the enterprise must apply for fixed public IP addresses.
  - Fixed public IP addresses are costly. NAT server technology can be used to use one public IP address to provide multiple services, saving costs.
  - In the long-term plan for the IT system, an enterprise should consider the IPv4-to-IPv6 transition.

Device Naming

- Naming devices facilitates network management.
- Device naming involves the following:
  - Name (sysname) of devices
  - IDs of logical interfaces, VLANs, and ACLs
  - Description of interfaces, VLANs, and ACLs

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The general name of a device includes the device name, configuration description, as well as the IDs of interfaces and VLANs.
- Device naming seems insignificant. Actually, it is essential to network O&M and management. Proper device naming rules greatly improve network management efficiency.
- An enterprise should establish its own device naming rules and strictly enforce them.

Naming Devices

- Naming devices does not affect running of the devices.
- To facilitate management, enterprises should use unified specifications to name devices.
- A-B-C-D-E, for example: HQ-CS-HW-S7706-1
  - A: Geographical location of a device. For example, HQ indicates that the device resides at the headquarters of an enterprise.
  - B: Location of a device on the logical network. For example, CS indicates that the device is a core switch.
  - C: Device vendor. For example, HW indicates that the device vendor is Huawei.
  - D: Device model, for example, S7706.
  - E: Device ID, for example, 1 or 2.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- No industry standards or regulations are available for naming of network devices.
- Enterprises can define naming rules based on service requirements.

Numbering Logical Configurations

Configuration	Suggestion	Example
Ethernet sub-interfaces	A sub-interface number is similar to a VLAN ID. Use description.	Ethernet 0/0/1.10 is bound to VLAN 10.
MP group interface	The first two digits in the ID of an MP group interface indicate the slot ID and sub-card ID respectively. The last two digits in the ID of an MP group interface are the same for the two devices in an MP group. Use description.	N/A
VLAN ID	VLAN IDs must be planned and used by segment. Use description.	VLAN 10 to VLAN 100 are used for network interconnection. VLAN 101 to VLAN 200 are used for production services.
ACL number	Use ACL names. ACL numbers must be planned and used by segment. Use description.	ACLs 3000–3100 are used for routing policies. ACLs 3101–3200 are used for packet filtering.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Numbering logical configurations of a device does not affect running of the device. Orderly numbering rules facilitate network management.

Adding Configuration Description

Configuration	Suggestion	Example
Interface	Describe an interface connected to the peer device using the name and port number of the peer device.	Interface connected to GE1/0/1 of core switch 1 at the headquarters: To_HQ-CS-HW-S7706-1_GE1/0/1
VLAN ID	Describe a VLAN based on services in the VLAN.	VLAN used for office services of the marketing department: MKT-OA VLANIF interface used for device interconnection at Layer 3 (described based on the local interface, peer device, and peer interface): GE0/0/1_to_HQ-CS-HW-S7706-1_GE1/0/1
ACL number	Describe the usage of an ACL.	ACL used for routing policy **rp1**: For_RP_rp1

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Configuration descriptions do not participate in any protocol calculation process. Descriptions greatly improve the readability of device configurations and enable new network administrators to quickly get familiar with device configurations.

Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet
   - QoS Design
   - Network Security Deployment
   - Network Management Planning

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Routing Protocols

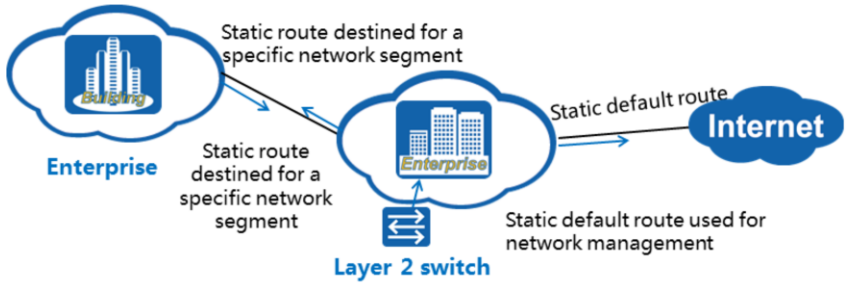
Routing Protocol	Algorithm	Convergence Speed	Deployment	Application Scenario
Direct route	Generated by link layer protocols	N/A	N/A	Default
Static route	Manually configured by administrators	N/A	Easy	Small-scale network (< 3 routers) Supplement to dynamic routing protocols
OSPF	Link state algorithm Route costs are measured based on bandwidth.	Fast	Fairly easy	Medium- and large-sized network (hundreds of routers)
IS-IS	Link state algorithm Route costs are measured based on multiple parameters, such as the bandwidth and delay.	Very fast	Fairly easy	Medium- and large-sized network (hundreds of routers)
BGP	Path vector algorithm Routes are measured based on multiple route attributes.	Medium	Fairly easy	Ultra-large network (> 1000 routers) BGP MPLS VPN

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- In addition to direct routes and static routes, most enterprises select IGP.
  - Due to the efficiency and network scale issues, few enterprises use RIP.
  - In the enterprise network market, more network engineers are familiar with OSPF.
  - Therefore, most enterprises use OSPF as IGP.
- Some enterprises use BGP in the following situations:
  - IGP is incapable of processing all the routes on a large-scale network.
  - A network needs to use a large number of complex routing policies.
  - MPLS VPN is deployed.
- In a complex network environment, multiple routing protocols may be used.

Static Route Planning

- Static routing is an indispensable routing mode for enterprise networks. Typically, static default routes are:
  - Applied to the egress of an enterprise network.
  - Used by devices that do not support dynamic routing protocols,
  - Used for simple interconnection of different networks,



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Static routing is the simplest routing mode, but also needs to be planned when static routes are used.
  - Generally, static routes should be avoided on enterprise networks. In normal cases, devices learn required routes using dynamic routing protocols. Configured static routes are not preferred, but hinder troubleshooting.
  - The static default route is mainly used at the egress of an enterprise network. The static default route is advertised to the intranet through a dynamic routing protocol.
  - Some low-end devices do not support dynamic routing protocols, but devices that can be managed by the NMS generally support static routes.
  - Dynamic routing is recommended for simple network interconnection.
  - Static routes cannot respond to network changes. BFD and NQA can be used to associate static routes with interfaces and links, so that static routes can respond to network changes.

OSPF and ISIS

- Consistencies
  - Both OSPF and IS-IS are based on the link state algorithm.
  - Both OSPF and IS-IS support fast route convergence.
  - Both OSPF and IS-IS can be deployed on large-scale networks.
- Differences
  - IS-IS has a higher scalability to support IPv6.
  - Enterprise network engineers are more familiar with OSPF.
- Conclusion
  - Both OSPF and IS-IS are popular IGP protocols.
  - OSPF is recommended in enterprise scenarios.
  - IS-IS is recommended in scenarios where IPv4 and IPv6 are deployed at the same time.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- OSPF and IS-IS use similar algorithms and provide similar performance.
- Due to various reasons, more enterprises deploy OSPF and many engineers are more familiar with OSPF in the enterprise network market.
  - In the carrier network market, more customers use IS-IS.

BGP Planning Overview

- When BGP is recommended for enterprises?
  - The network scale is large and IGP is incapable of processing routes.
  - A large number of complex routing policies are required for traffic distribution and filtering on the network.
  - MPLS VPN needs to be deployed on the network.
- Which devices can have BGP deployed?
  - BGP does not generate routes, but manages and advertises routes.
  - BGP does not consume a large number of device resources.
  - All devices can run BGP as long as proper planning is made.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Generally, large enterprises use BGP. If required, any enterprise can deploy BGP.
- Unlike IGP, BGP does not generate routes, but manages and advertises routes. Therefore, BGP does not require devices to have high performance. Therefore, BGP can run on any device as long as proper planning is made.



Contents

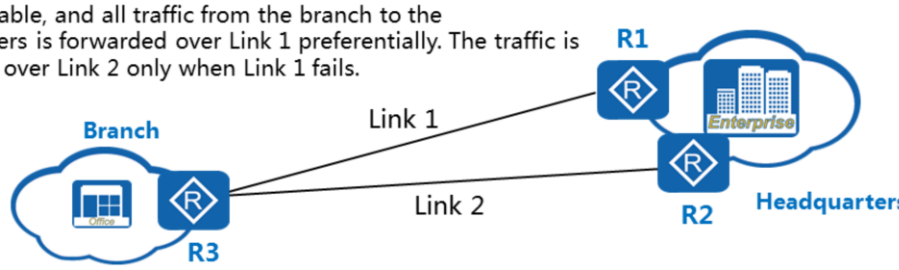
1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet
   - QoS Design
   - Network Security Deployment
   - Network Management Planning

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Routing Policy Planning Case

Requirement:
Link 1 is stable, and all traffic from the branch to the headquarters is forwarded over Link 1 preferentially. The traffic is forwarded over Link 2 only when Link 1 fails.



- Routing policies affect route calculation and therefore change the routing table.
  - A common routing policy is to filter out routes that are not be advertised. Metric values of routes can be adjusted to change the route selection sequence.
  - Different routing protocols have different routing policy tools.
  - Routing policies must be considered on a global basis.

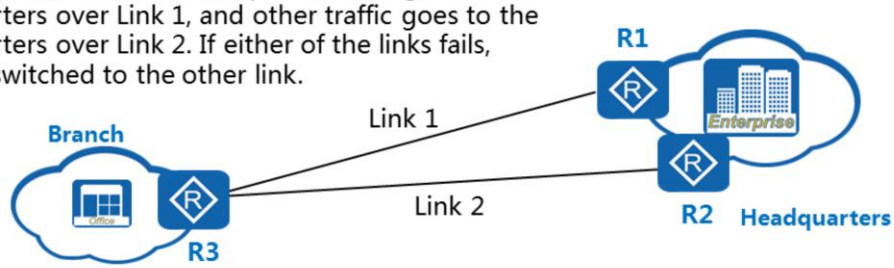
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- To meet the above mentioned requirement, plan as follows:
  - Assume that R1, R2, and R3 are in the same OSPF area.
  - Adjust the cost of Link 2 between R2 and R3 to a larger value.
  - In this way, OSPF prefers Link 1 with a lower cost value.

PBR Planning Case

Requirement:
All HTTP traffic from the enterprise branch goes to the headquarters over Link 1, and other traffic goes to the headquarters over Link 2. If either of the links fails, traffic is switched to the other link.



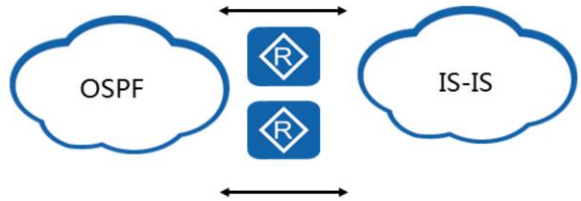
- PBR directly applies to the forwarding planes of devices, but has no impact on the routing table.
  - Common PBR is to change the next hop for data forwarding.
  - Based on different source addresses
  - Based on other conditions

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- To meet the above mentioned requirement, plan as follows:
  - Adjust the cost value (routing policy) to allow all traffic of enterprise branches to go to the headquarters over Link 2.
  - Configure PBR on R3. Identify HTTP traffic (TCP port number = 80) and specify R1 as the next hop of HTTP traffic.
  - Similarly, use a similar policy on R1 and R2 to allow HTTP traffic to go to R3 over Link 1.

Route Import Policy



- Key Points of Route Import
  - Before importing routes, add a routing control policy to prevent all routes from being imported.
  - The default route cannot be imported and needs to be advertised in a special mode.
  - When two dynamic routing protocols on two devices import routes from each other, routing loops may occur.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- In actual networking, static or direct routes are imported to advertise the static or direct routes.
  - Generally, static default routes cannot be imported by dynamic routing protocols.
  - Special commands are available for common dynamic routing protocols to import and advertise default routes.
  - If no routing policy is configured, all routes of the same type are imported by default when default routes are imported. As a result, some problems, such as routing loops, occur.
  - Especially when two dynamic routing protocols import routes to each other, there is a higher probability that routing loops occur.
  - Therefore, before importing routes, you need to add a routing policy to filter out the routes that do not need to be advertised, or add some identifiers (such as tags) to facilitate subsequent route control.

Use of Default Routes



- Default routes play an important role on a network.
  - Configure a static default route on the network egress.
  - Advertise the static default route to the intranet through a dynamic routing protocol.
  - It is not recommended that static default routes be configured on the enterprise intranet (except Layer 2 devices).

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- It is not recommended that static default routes be used on the intranet.
  - Only routes learned by dynamic routing protocols (including default routes) exist on the enterprise intranet.
  - If static default routes are configured on the intranet, routing loops may occur and faults cannot be located. (Why?)
  - Static default routes can be used for temporary emergency situations. In special cases, for example, when dynamic routing protocols fail, you can configure static default routes.

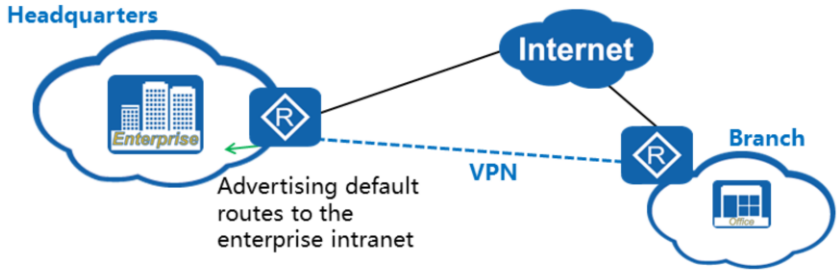
Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet
   - QoS Design
   - Network Security Deployment
   - Network Management Planning

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Enterprise Network and Internet



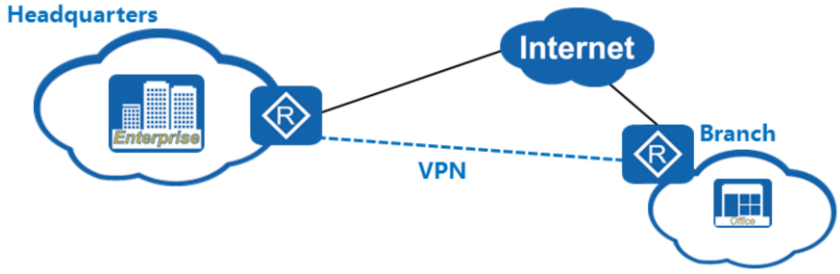
- Almost all enterprise networks are connected to the Internet.
  - Enterprise users need to access the Internet.
  - Enterprises provide Internet access services (such as web and email services).
  - Enterprise branches set up VPNs to connect to the headquarters through the Internet.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Nowadays, few enterprise networks are not connected to the Internet (except for confidential networks).
- Enterprise users need to access the Internet, enterprises need to provide Internet access, and enterprise VPNs also need to connect with the Internet.
- The Internet egress design of enterprise networks is also very important.
- The enterprise intranet needs to learn only the routes directing to egress devices, but does not need to learn specific routes directing to the Internet.

Internet Egress Device

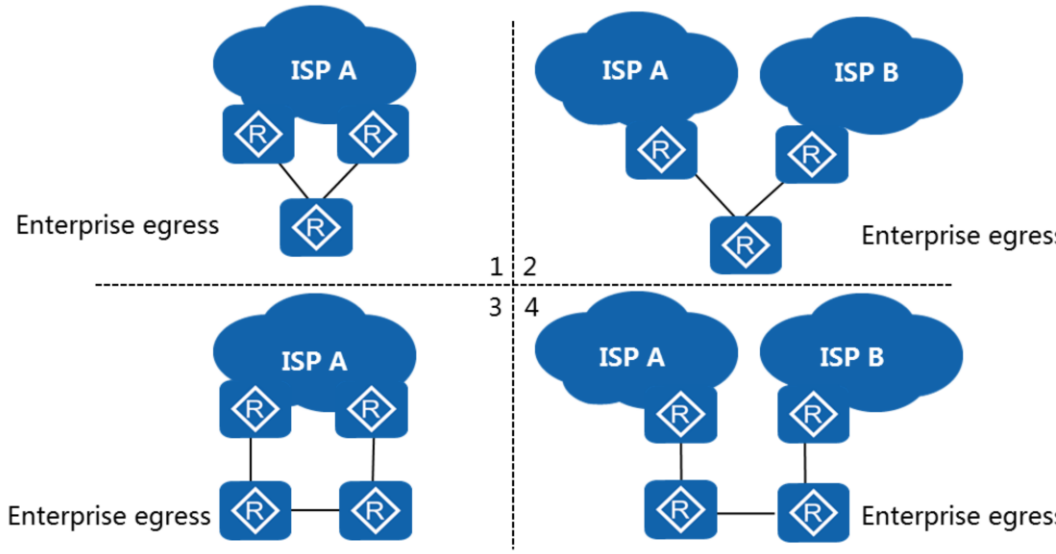


- Internet egress devices of enterprise networks must support the following functions:
  - NAT and NAT server functions
  - Routing function (interconnection with ISPs)
  - Security functions (such as attack defense and VPN)

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Generally, a high-end router, transparent firewall, or a firewall supporting the routing function can serve as the egress device of an enterprise network.
- VPN devices can be routers or firewalls, and can also be deployed separately.
  - Many routers or firewalls support the VPN function.
  - If an enterprise has a large number of branches or VPN users, you are advised to deploy dedicated VPN devices.

Internet Egress Backup

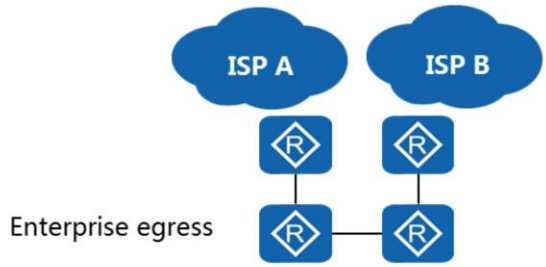


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Internet egress connections are critical to enterprise networks. Therefore, backup must be considered in the Internet egress design.
  - In solution 1, egress links are backed up.
  - In solution 2, egress links and ISPs are backed up.
  - In solution 3, egress devices and egress links are backed up.
  - In solution 4, egress devices, egress links, and ISPs are backed up.
- Enterprises should select appropriate Internet egress backup solutions based on service reliability requirements and budgets.

Multiple Internet Egresses



- When multiple Internet egresses are deployed, the following should be determined:
  - Outbound interface of Internet access traffic sent from intranet users
  - Inbound interface of the traffic sent from extranet users to access internal servers

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Determining the outbound interface of Internet access traffic sent from intranet users:
  - Running BGP between enterprise networks and ISP networks is the best way. However, dynamic routing protocols are not run between ISP networks and enterprise networks.
  - When only one ISP exists, enterprises can configure static egress routes on the egress routers to solve the problem.
  - When multiple egresses exist, enterprises need to configure specific static routes based on the Internet segments covered by different ISPs and advertise these static routes at least between egress routers.
  - In this way, enterprise users get higher Internet access rates and backup is achieved.
- Determining the inbound interface of the traffic sent from extranet users to access internal servers
  - Actually, this problem cannot be solved at the network layer. If an enterprise advertises only one public IP address, external users can access internal servers only through the interface from which the public IP address is advertised.
  - This problem can be solved at the application layer. Generally, users use domain names instead of IP addresses to access Internet services. An enterprise network can use two public IP addresses to provide Internet services and bind public IP addresses to the same domain name. You can set some parameters on the DNS server to enable domain name requests sent from different networks to obtain

different IP addresses.

Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet
   - QoS Design
   - Network Security Deployment
   - Network Management Planning

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

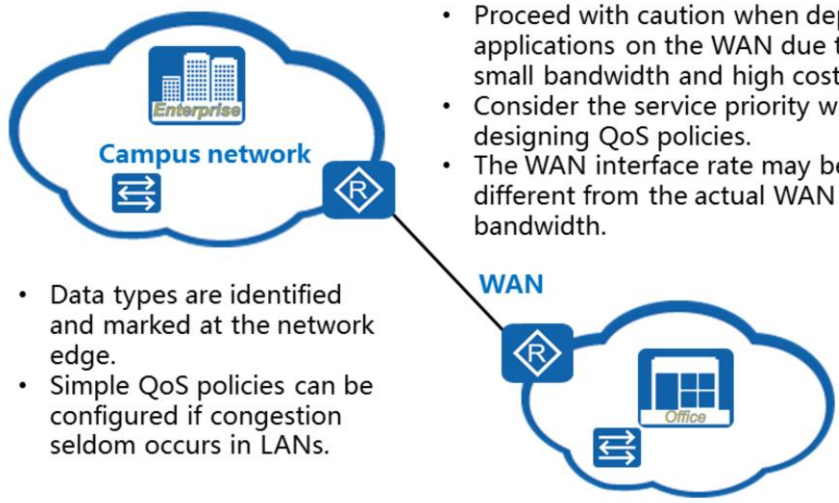
QoS Requirement Analysis

- QoS policies must be designed and deployed based on application requirements.
- Enterprise applications vary greatly. The application data is typically classified into the following types:
  - Audio
  - Video
  - Office
  - Network control information
  - Others

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Next, we will analyze different types of data.

QoS Deployment



- Proceed with caution when deploying applications on the WAN due to its small bandwidth and high cost.
- Consider the service priority when designing QoS policies.
- The WAN interface rate may be different from the actual WAN link bandwidth.

- Data types are identified and marked at the network edge.
- Simple QoS policies can be configured if congestion seldom occurs in LANs.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Different tags are added to different data. Devices then apply different QoS policies to the data according to the tags.

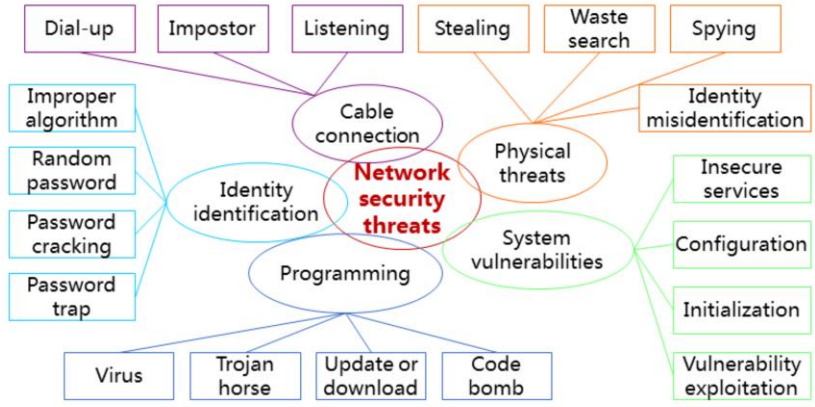
Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet
   - QoS Design
   - Network Security Deployment
   - Network Management Planning

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Network Security Threats



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Networks are under various threats, such as cable connection, identity identification, and system vulnerability threats. Network security involves any object on the network, which can be attacked or be used to attack others. There is no absolute security. Therefore, you need to evaluate the security risks and find the security vulnerabilities for security improvement.
- IP networks are mainly constructed based on the TCP/IP protocol stack. Therefore, our analysis is also based on the TCP/IP protocol stack.

Security Capabilities

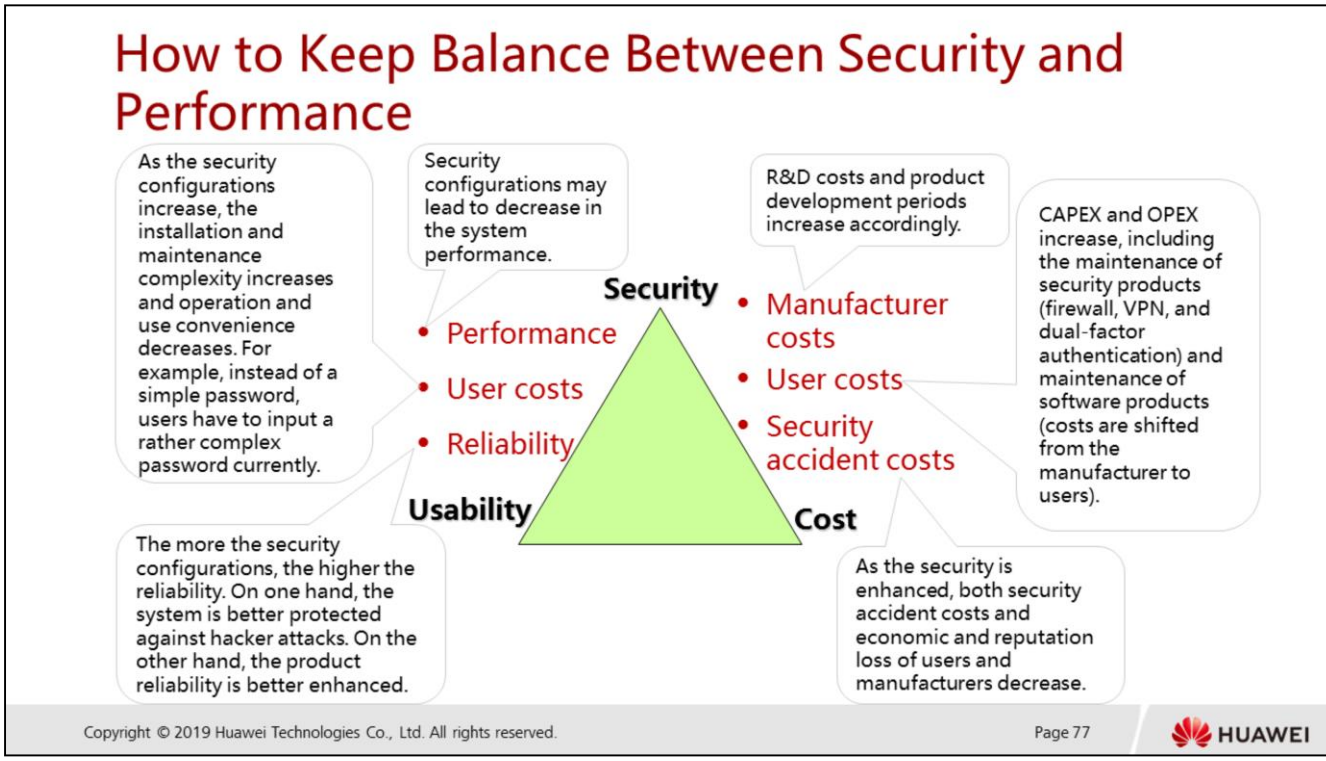


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

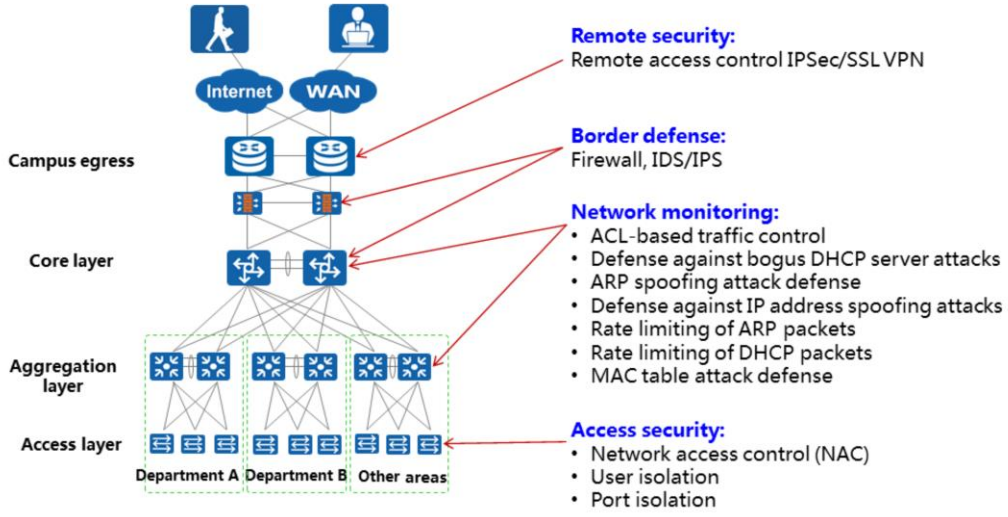
- Security inevitably involves people. If we put the most confidential things on Mars, they will be quite safe. In the current security environment, the security management capability is especially important because all information is processed by people. If strict security management is not performed on persons who can access the information, any security technology becomes ineffective. With the help of security management strategies and policies, related technical means can be used to improve security capabilities. Security capabilities are classified into six types:

- Protection, Detection, Response, Recovery, Counterattack, and Warning. These capabilities are also referred to as PDRRCW.
- Currently Protection, Detection, Response, Recovery (PDRR) is more widely applied and focuses on passive attack defense. PDRRCW covers two more items, Counterattack and Warning, and boasts more powerful security capabilities to some extent.



- Security, usability, and costs form a triangle. As security is enhanced, the system usability decreases, and maintenance costs increase. However, if security is not enhanced, security problems are prone to occur, causing additional costs.
- Security guarantee is a time-consuming and resource-consuming task. To provide the best security services, you must keep a good balance between security, usability, and costs.

Hierarchical End-to-End Security



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- NAC: Network Access Control
  - Verifies users' identities and controls network access rights, thus protecting security for the enterprise network.
  - Monitors attack traffic on the network, preventing man-in-the-middle attacks.
  - Uses external network attack defense methods to ensure access security of external users and implement secure and reliable data transmission.

Contents

1. Enterprise Network Design Principles
2. **Enterprise Network Design Overview**
   - IP Address Planning and Device Naming
   - Routing Design
   - PBR and Routing Policy
   - Connection with the Internet
   - QoS Design
   - Network Security Deployment
   - Network Management Planning

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Common Network Management Modes

- In actual network management, multiple modes are used.

Mode	Description	Applications
CLI	• Manages single devices through command lines. • Lacks the capability of managing the entire network. • Requires the administrator to be familiar with command lines. • Provides high efficiency but is not easy to use.	• Initial device configuration • Complex configuration • Solving of complex problems
Web	• Manages single devices on the web page. • Lacks the capability of managing the entire network. • Is easy to use and applicable to simple and repeated configurations.	• Observing of device status • Simple configuration
SNMP	• Manages devices on the entire network through SNMP. • Manages network-wide devices, links, and even services. • Provides a graphical user interface (GUI) and is easy to learn and use. • Lacks the capability of solving some complex problems.	• Management of network-wide devices, links, and services • Used with the CLI and web modes

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- CLI and web modes are device-based management modes.
  - Some devices provide only one of them, and some devices provide both.
  - For either CLI or web management, the commands or operation pages for devices from different vendors or devices of different models from the same vendor may be different.
- SNMP is a standard network management protocol.
  - Most network devices support SNMP.
  - SNMP manages different devices in the same way.

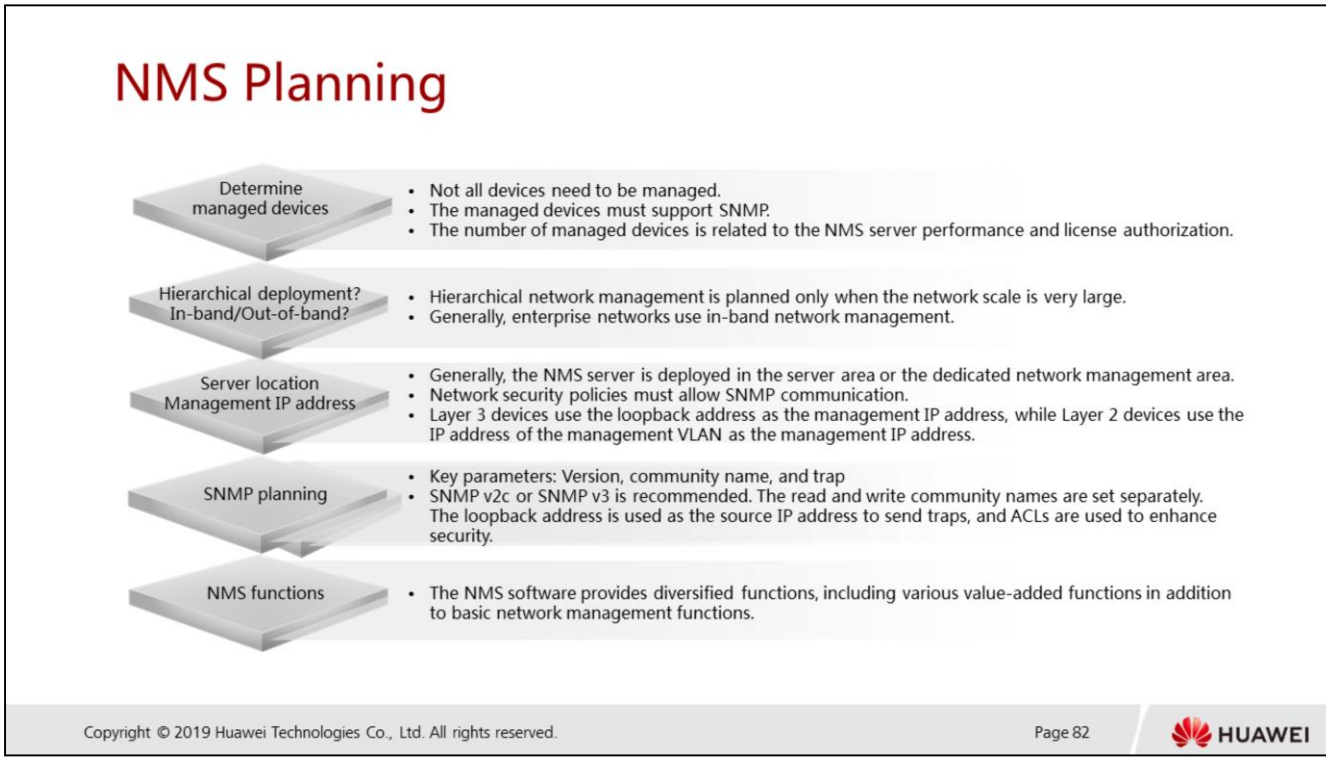
Secure Network Management

- Network management traffic includes key information, which is critical to the network.
- Traditional network management protocols lack a security mechanism:
  - Telnet
  - HTTP
  - SNMP v1/v2
- Secure network management protocols are recommended:
  - SSH
  - HTTPS
  - SNMP v3

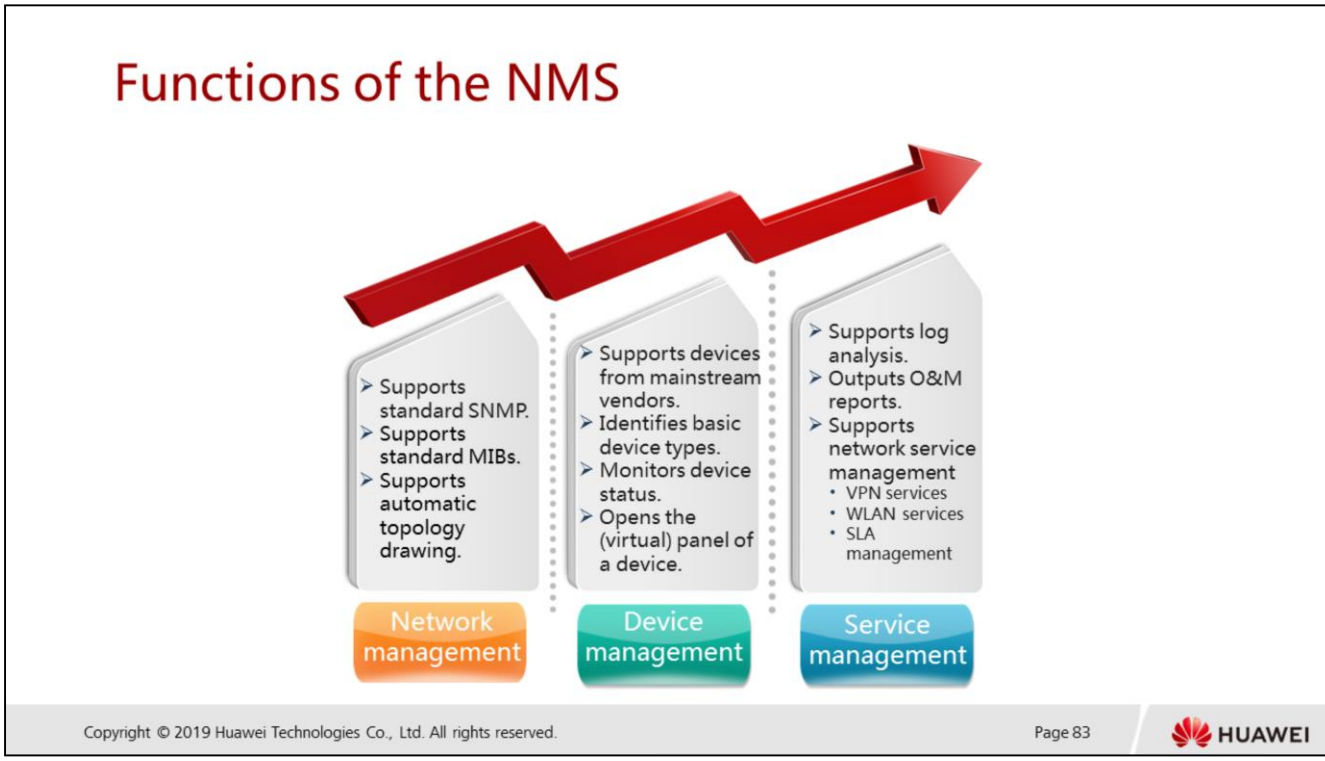
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Network management traffic includes key information, which is critical to the network.
  - The administrator password is an example.
- Because of the dedicated purposes and limitations of early networks, many management protocols transmit data in plaintext, which cannot ensure security. With the penetration of networks into various fields, network security is becoming more important. The traditional management protocols that transmit data in plaintext cannot meet security requirements. Therefore, management protocols that transmit data in encrypted mode are developed:
  - Telnet -> SSH
  - HTTP -> HTTPS
  - SNMP v1/v2 -> SNMP v3



- NMS: Network Management System
  - The NMS software is the main component of the NMS.
  - With the development of technologies, the functions of NMS software are changed from pure network management to ICT full-service management.



- Network management
  - Network management is a basic function of the NMS.
  - The NMS should support standard SNMP and standard MIBs.
  - The NMS can automatically discover and draw network topologies.
- Device management
  - Device management is a convenient function for users. However, the management of devices from multiple vendors imposes higher NMS requirements.
  - The NMS should be able to identify the main types of devices from mainstream vendors.
  - The NMS can monitor the status of these devices.
  - The NMS can open the (virtual) panels of these devices.
- Service management
  - As a network O&M tool, the NMS should be able to analyze logs and generate reports.
  - Additionally, the NMS should be able to manage network services, such as VPN services, WLAN services, and SLA.

Contents

1. Enterprise Network Design Principles
2. Enterprise Network Design Overview
3. **Product Selection**
   - How to Select Products
   - AR G3 Series Routers
   - Sx700 Series Switches
   - Other Products
4. Enterprise Network Design Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

How to Select Products



Price

The overall price of the product is reasonable.

Technology

- The vendor has technical strengths.
- The product meets technical specifications.
- The product supports common standards.
- The product is compatible with other mainstream products.

Consider all possible factors such as technology, price, and service to select **the most suitable** product.

Service

The vendor can provide professional technical services.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

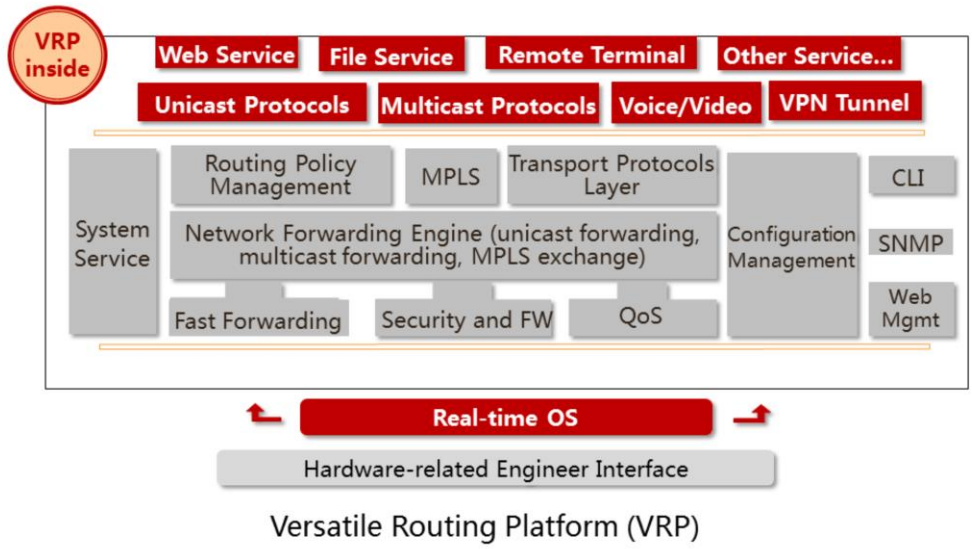
- Consider all possible factors such as technology, price, and service to select the most suitable product rather than the best product.
  - Meeting service requirements is still the foremost element for device selection.
  - On the premise of meeting service requirements, select products with leading technologies, excellent services, and reasonable prices.

Contents

1. Enterprise Network Design Principles
2. Enterprise Network Design Overview
3. **Product Selection**
   - How to Select Products
   - AR G3 Series Routers
   - Sx700 Series Switches
   - Other Products
4. Enterprise Network Design Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Huawei VRP



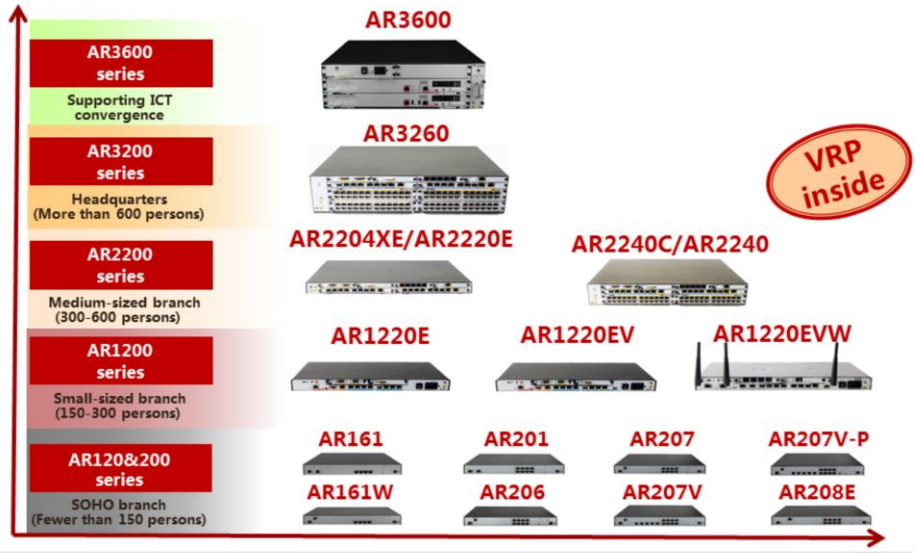
Versatile Routing Platform (VRP)

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The VRP is a network operating system used by Huawei data communications products. The network operating system is a system software that runs on certain devices and provides network access and interconnection services. The VRP is the core engine of Huawei routers, Ethernet switches, and service gateways. It provides a unified user interface and management interface, implements the functions of the control plane and defines the interface specifications of the forwarding plane to implement the interaction between the forwarding plane of each product and the VRP control plane. It also implements the network interface layer to shield the differences between the link layer of each product and the network layer.

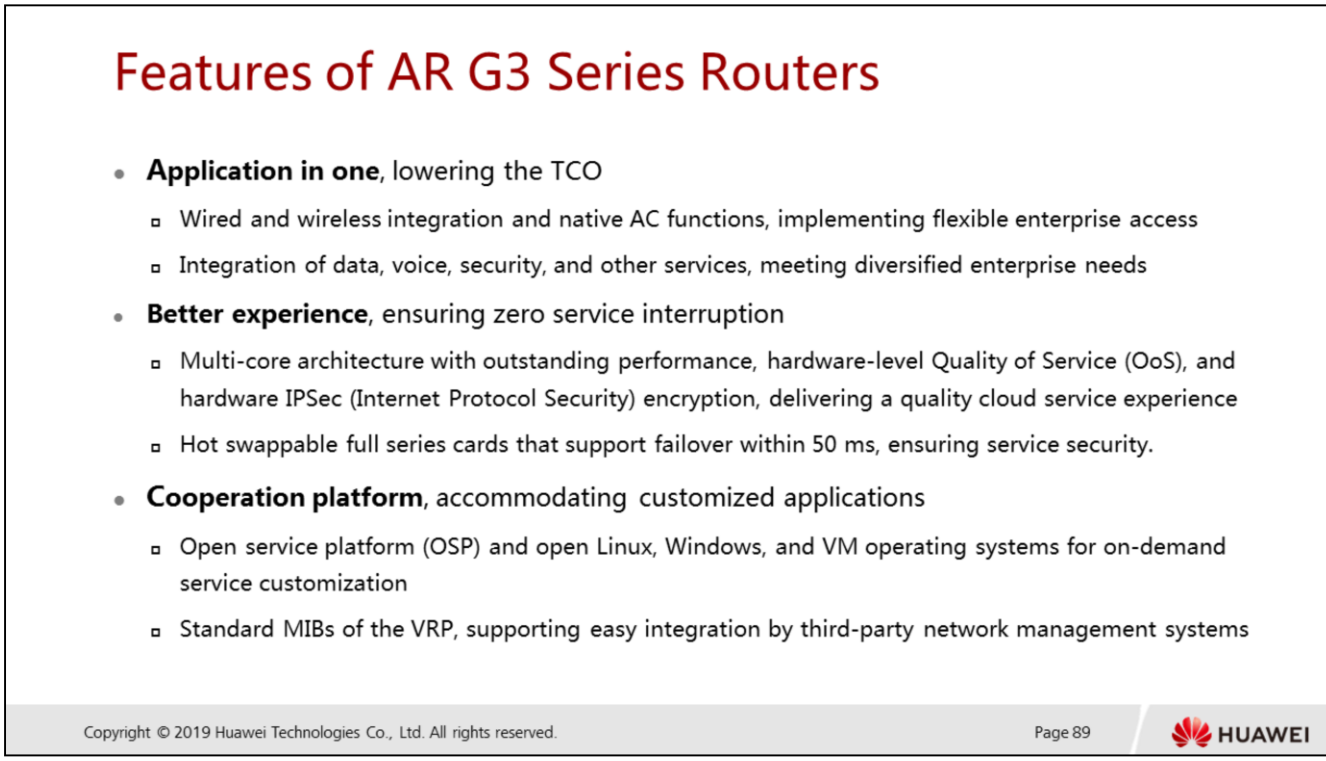
Huawei AR G3 Series Routers



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- AR G3 series routers are next-generation enterprise-class routers based on Huawei VRP. Huawei developed AR G3 series routers with the technologies accumulated in the data communication, wireless, access network, and core network fields. AR G3 routers integrate routing, switching, WLAN, 3G/LTE, voice, and security functions. They use the multi-core CPU and non-blocking switching structure, and provide industry-leading system performance and extensibility, meeting diversified application development requirements of enterprises in the future. These routers provide an integrated solution for enterprise networks, speed up multi-service provisioning, and protect customer investments.



- AR series enterprise routers use the multi-core CPU and non-blocking switching structure, and provide industry-leading system performance, meeting network extension requirements and service development requirements of enterprises and branches.
- AR G3 routers support 3G and LTE network modes as well flexible access through optical fibers and copper cables.
- AR series routers interconnect with mainstream third-party IT systems by using the OSP to provide a unified communication service experience for enterprise users. The customers, agents, third-party vendors, and manufacturers can develop and use AR series routers to create more value.
- AR series routers provide various voice functions for enterprise data networks, allowing enterprises to communicate flexibly and efficiently.
- During service provisioning, AR series routers ensure security of enterprise networks. They provide a complete set of security protection mechanisms including user access control, packet detection, and proactive attack defense. These mechanisms maximize customer investments.
- Intelligent service deployment and simple service management are achieved.

Appearance of AR2200 Series Routers



Front view of the AR2204-27GE

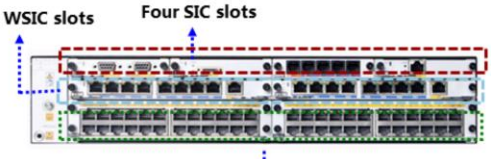


Front view of the AR2220E



3GE (2GE combo)

Front view of the AR2240



Rear view of the AR2240

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- AR2200 series routers use the embedded hardware encryption technique and support the voice Digital Signal Processor (DSP), firewall functions, call processing, voice mail, and various application programs. AR2200 series routers support various wired and wireless access modes, such as E1/T1, xDSL, xPON, CPOS, 3G, and LTE.
- AR2200 series routers include the following models: AR2204-27GE, AR2204-27GE-P, AR2204-51GE, AR2204-51GE-P, AR2204E, AR2220E, AR2240C, AR2240, and AR2204XE.

Specifications of AR2200 Series Routers

Specification	AR2220E	AR2240
Forwarding performance	9 Mpps	10–25 Mpps
Firewall performance	3 Gbit/s	5.5–15 Gbit/s
Ports	3*GE (1*Combo)	3*GE (2*Combo)/4*GE Combo + 2*GE SFP/4*GE Combo + 2*10GE SFP+
SIC slot	4	4
WSIC slot	2	4
XSIC slot	0	4
ESIC slot	0	1
DSP slot	1	3
Memory	1 GB	4 GB
Flash (default/maximum)	512 MB/4 GB	2 GB/8 GB
Micro SD card (default/maximum)	2 GB/4 GB	2 GB/4 GB
USB port	2	2
Dimensions (W x D x H)	442 mm x 420 mm x 44.5 mm	442 mm x 470 mm x 88.2 mm

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

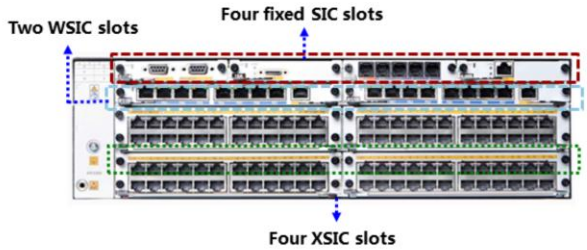
- The AR2240 supports multiple pluggable SRUs. The SRUs differ in the forwarding performance and traffic management functions. The SRUs provide hardware-level traffic management and hardware H-QoS.
- AR2200 series routers support multiple types of interface cards, including Ethernet interface cards, E1/T1/PRI/VE1 interface cards, synchronous/asynchronous interface cards, ADSL2+/G.SHDSL interface cards, FXS/FXO voice cards, ISDN interface cards, CPOS interface cards, EPON/GPON interface cards, and 3G/LTE interface cards. The cards are classified into SIC cards, WSIC cards, and XSIC cards depending on the slot type.

Appearance of AR3200 Series Routers



SRU80 high-performance SRUs. V200R003 does not support dual SRUs.

Front view of the AR3260



Rear view of the AR3260

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- AR3200 series routers use the embedded hardware encryption technique and support the voice DSP. They also support firewall functions, call processing, voice mail, and various application programs. AR3200 series routers support various wired and wireless access modes, such as E1/T1, xDSL, xPON, CPOS, and 3G.

Specifications of AR3200 Series Routers

Specification	AR3260
Forwarding performance	10–40 Mpps
Service forwarding capability	1–5.5 Gbit/s
System switching capacity	640 Gbit/s
SIC slot	4
WSIC slot	4
XSIC slot	6
Ports	3*GE (2*Combo)/4*GE Combo + 2*GE SFP/4*GE Combo + 2*10GE SFP+
DSP slot	3
Memory	4 GB
Flash	8 GB
ESIC slot	1
USB port	2
Dimensions (W x D x H)	442 mm x 470 mm x 130.5 mm

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The AR3260 supports multiple types of pluggable SRUs. The SRUs differ in the forwarding performance and traffic management functions. The SRUs provide hardware-level traffic management and hardware H-QoS.
- AR3200 series routers support multiple types of interface cards, including Ethernet interface cards, E1/T1/PRI/VE1 interface cards, synchronous/asynchronous interface cards, ADSL2+/G.SHDSL interface cards, FXS/FXO voice cards, ISDN interface cards, CPOS interface cards, EPON/GPON interface cards, and LTE interface cards. The cards are classified into SIC cards, WSIC cards, and XSIC cards depending on the slot type.

SRU40, SRU60, and SRU80 Specifications



SRU40/SRU60/SRU80

SRU specifications:

SRU Name	Performance	Function
SRU40	8-core CPU, 600 MHz frequency	Does not support traffic management.
SRU60	8-core CPU, 600 MHz frequency	Does not support traffic management.
SRU80	12-core CPU, 750 MHz frequency	Supports traffic management.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- An SRU integrates the control and management functions, and provides the control plane, management plane, and switching plane for the system. Control plane: provides functions such as protocol processing, service processing, route calculation, forwarding control, service scheduling, traffic statistics collection, and system security.
- Management plane: provides functions such as system monitoring, environment monitoring, log and alarm processing, system software loading, and system upgrade.
- Switching plane: provides high-speed, non-blocking data channels for service switching between service modules.

SRU100, SRU200, SRU400, and SRUX5 Specifications



SRU100



SRU200



SRU400

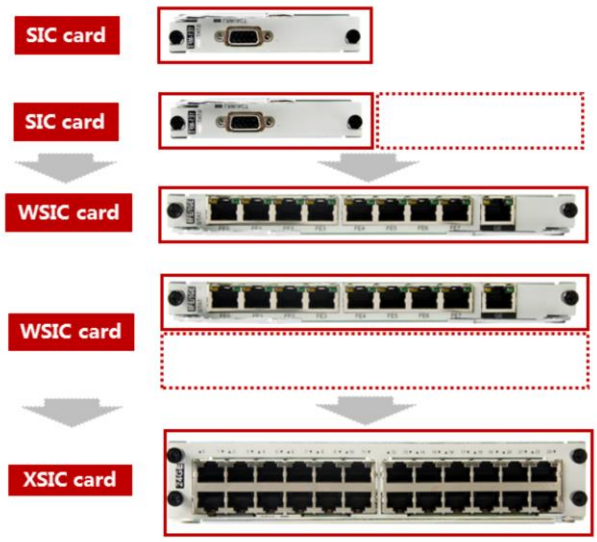


SRUX5

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- An SRU integrates the control and management functions and provides the control plane, management plane, and switching plane for the system. Control plane: provides functions such as protocol processing, service processing, route calculation, forwarding control, service scheduling, traffic statistics collection, and system security.
- Management plane: provides functions such as system monitoring, environment monitoring, log and alarm processing, system software loading, and system upgrade.
- Switching plane: provides high-speed, non-blocking data channels for service switching between service modules.

Interface Cards Supported by AR G3 Series Routers



AR G3 series routers support four types of interface cards:
- **SIC card**
- **WSIC card**
- **XSIC card**
- **EXSIC card**

Two SIC slots can be combined into **one WSIC slot** by removing the guide rail between them.

Two WSIC slots can be combined into **one XSIC slot** by removing the guide rail between them.

Switching capacity of each slot
SIC and WSIC slots: 2 Gbit/s XSIC and EXSIC slots: 20 Gbit/s

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

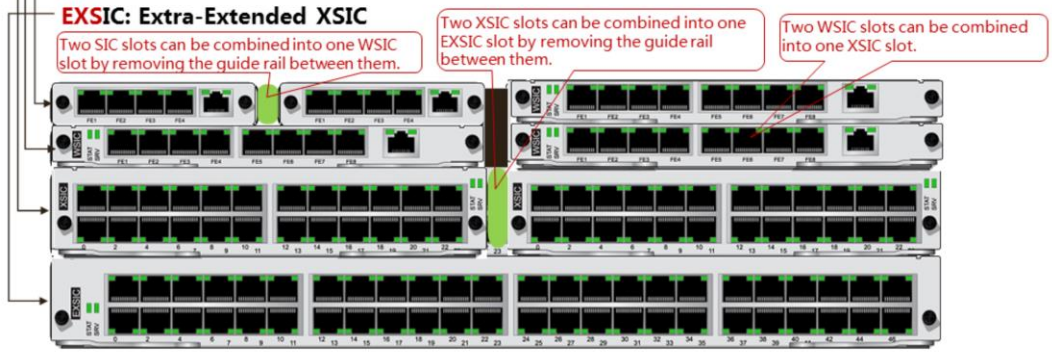
HUAWEI

- Two SIC slots can be combined into one WSIC slot by removing the guide rail between them.
- Two SIC slots and the WSIC slot below them can be combined into one XSIC slot by removing the guide rails.
- Two XSIC slots can be combined into one EXSIC slot by removing the guide rail between them.
- Slots can be combined into one, but one slot cannot be divided into multiple slots.
- The new slot ID is the larger one between the two original slot IDs.
- In V200R002C00 and later versions, a WSIC card can be inserted into an XSIC slot in the lower side of the slot, and uses the XSIC slot ID as its own slot ID.
- MFS stands for Multiple Function Slot.

Combination of Interface Cards on AR G3 Series Routers

With a standard design, AR G3 series routers support flexible combination of various service cards, making full use of slot resources and protecting investment.

- **SIC: Smart Interface Card**
- **WSIC: Wide SIC**
- **XSIC: Extended WSIC**
- **EXSIC: Extra-Extended XSIC**



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Slots can be combined into one, but one slot cannot be divided into multiple slots.
- The new slot ID is the larger one between the two original slot IDs.

WAN Interface Modules Supported by AR G3 Series Routers (1/2)

E1 SIC card:



1E1/T1-M 1/2-port channelized E1/T1/PRI/VE1 multifunctional interface module



2E1/T1-M It receives, sends, and processes data flows on E1 interfaces. This module can be configured with VE1 interfaces to transmit voice and data services as well as fax signals.



1E1-F 1/2-port unchannelized E1/T1 multifunctional interface module



2E1-F

Ethernet SIC



2FE 2-port 10M/100M electrical WAN interface module



1GEC 1-port GE Ethernet combo WAN interface module

Synchronous/Asynchronous SIC



1SA 1-port enhanced synchronous/asynchronous serial interface module
The following protocols are supported: V.24, V.35, X.21, RS-449, and RS-530.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- 1/2: one or two interfaces E1: E1 interface T1: T1 interface M: multiflex trunk Primary Rate Interface (PRI): ISDN primary rate interface
- VE1: E1 voice interface

WAN Interface Modules Supported by AR G3 Series Routers (2/2)

ADSL2+ SIC card:



1ADSL-A/M 1-port ADSL over POTS interface module, supporting ANNEX A/M



1ADSL-B/M 1-port ADSL over ISDN interface module, supporting ANNEX B

4G.SHDSL SIC card:



4G.SHDSL 1-port 4-pair G.SHDSL interface module

PON SIC card:



1PON 1-port EPON/GPON interface card

CPOS SIC card:



1CPOS 1-port 155M CPOS interface module

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The 4G.SHDSL provides 4-channel G.SHDSL access and the independent CPU as well as management interfaces.
- The 1PON is an EPON/GPON auto-sensing module used on an AR router. It works with SRUs and supports two PON SFP uplink interfaces.
- 1CPOS-155M (1-port channelized POS interface card): **C** indicates channelized; **POS** indicates packet over SDH/SONET; **155M** indicates that the rate is 155.52 Mbit/s.

Ethernet Switching Interface Modules Supported by AR G3 Series Routers

Ethernet switching WSIC card:



8FE/1GE

8-port 100M (RJ45) and 1-port 1000M (RJ45)-L2/L3 Ethernet electrical interface module It supports line-speed transmission of Layer 2 and Layer 3 data. Each interface supports auto-sensing and can work in half duplex or full duplex mode.

Ethernet switching XSIC card:



24GE

24-port 1000M (RJ45)-L2/L3 Ethernet electrical interface card It supports transmission of Layer 2 and Layer 3 data. Each interface supports auto-sensing and can work in half duplex or full duplex mode.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The 8FE1GE can be installed into the WSIC slot on the AR1200, AR2200, and AR3260 chassis. On the AR1200 and AR2204, two SIC slots need to be combined into one WSIC slot.
- The 24GE can be installed into the XSIC slot on the AR2220, AR2240, and AR3260 chassis. On the AR2220, two WSIC slots need to be combined into one XSIC slot.

Voice Interface Modules Supported by AR G3 Series Routers

Analog voice SIC card:



4FXS1FXO

It provides four FXS interfaces and one FXO interface to transmit voice signals over a data communication network.

ISDN S/T voice SIC card:



1BST

1-port ISDN data access module, which sends, receives, and processes ISDN data flows



2BST

2-port ISDN voice access module, which implements access, compression, and decompression of ISDN voice signals on a data communication network

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Foreign Exchange Station (FXS) interfaces are standard RJ-11 interfaces. FXS interfaces connect to devices such as ordinary telephones and fax machines through telephone lines, and exchange signaling with the devices through level changes of tip and ring lines to provide ringing, voltage, and dial tones.
- A Foreign Exchange Office (FXO) is a two-wire loop trunk. An FXO interface is an RJ-11 interface, and connects a local call to the central office of the Public Switched Telephony Network (PSTN) or a small user switch (PBX) through a telephone line. Similar to FXS interfaces, FXO interfaces also exchange signaling through level changes of tip and ring lines. FXO interfaces can connect only to other FXS interfaces.

- The 2BST is an ISDN service access module for AR series routers. It provides two ISDN S/T interfaces to transmit voice services.
- The 2BST offers the ISDN BRI function and provides the bandwidth of two B channels and one D channel:
- B channel is a voice channel and provides 64 kbit/s bandwidth.
- D channel is a signaling channel and provides 16 kbit/s bandwidth.
- The total bandwidth of two B channels and one D channel is 144 kbit/s.
- The S/T interface on the 2BST provides a line rate of 192 kbit/s, including 144 kbit/s for data transmission (two B channels and one D channel) and 48 kbit/s for maintenance information transmission.

Ethernet Electrical Interfaces on AR G3 Series Routers

- Ethernet electrical interface
  - Interface type: RJ45
  - Cable: UTP/STP
  - Rate: 10 Mbit/s, 100 Mbit/s, 1 Gbit/s
  - Link protocol: Ethernet
  - Card:
    - 8FE1GE
    - 24GE
    - 2FE



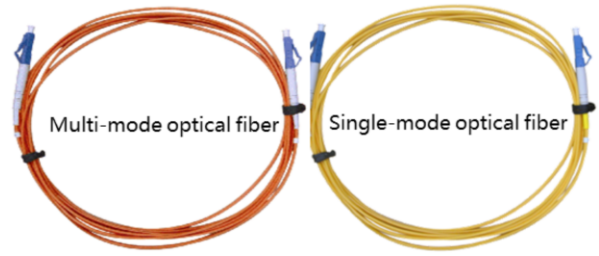
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Network cables connect network devices to each other to enable the devices to communicate or to allow local maintenance and remote access.

Ethernet Optical Interfaces on AR G3 Series Routers

- Ethernet optical interface
  - Interface type: LC/PC
  - Cable: single-mode and multi-mode
  - Rate: 100 Mbit/s, 1 Gbit/s
  - Link protocol: Ethernet
  - Card
    - SRU
    - 1GEC





SFP optical interface module



LC/PC header

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A single-mode optical fiber and a multi-mode optical fiber have the same appearance but different colors. A single-mode optical fiber is in yellow and a multi-mode optical fiber is in orange.
- Optical module types:
- An optical module of the multitranverse mode must be connected to a multi-mode optical fiber.
- An optical module of the single-longitudinal or multi-longitudinal mode must be connected to a single-mode optical fiber.

E1/T1 Interfaces on AR G3 Series Routers

- E1/T1 interface
  - Interface type: DB9
  - Cable: trunk cable
  - Rate: 2.048 Mbit/s, 1.544 Mbit/s
  - Link protocol: PPP/HDLC
  - Card:
    - 1E1T1-M
    - 2E1T1-M



E1 75-ohm unbalanced coaxial cable



E1 120-ohm or T1 100-ohm balanced twisted pair cable

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- E1 cables are classified into the following types:
- 75-ohm unbalanced coaxial cable (DB9 to BNC), which is connected as follows:
  - One end provides a DB9 connector.
  - The other end provides two BNC connectors.
- 120-ohm balanced twisted pair cable (DB9 to RJ45), which is connected as follows:
  - One end provides a DB9 connector.
  - The other end provides an RJ45 connector.
- A T1 trunk cable is a 100-ohm balanced twisted pair cable. Its appearance is the same as that of an E1 120-ohm balanced twisted pair cable.

xDSL Interfaces on AR G3 Series Routers

- xDSL interface
  - Interface type: RJ11/RJ45 (4G.SHDSL)
  - Cable: common telephone cable
  - Rate: 5.696 Mbit/s, 8 Mbit/s, 12 Mbit/s, 24 Mbit/s
  - Link protocol: analog modulation and demodulation
  - Card:
    - ADSL-A/M
    - ADSL-B
    - 4G.SHDSL



4G.SHDSL cable

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- A 4G.SHDSL cable is connected as follows:
  - One end provides an RJ45 connector.
  - The other end provides four RJ11 connectors.

Synchronous/Asynchronous Serial Interfaces on AR G3 Series Routers

- Synchronous/Asynchronous serial interface
  - Interface type: DB28
  - Cable: DCE/DTE
  - Rate: 64 kbit/s to 2.048 Mbit/s
  - Link protocol: Protocol/Flow
  - Card:
    - 1SA



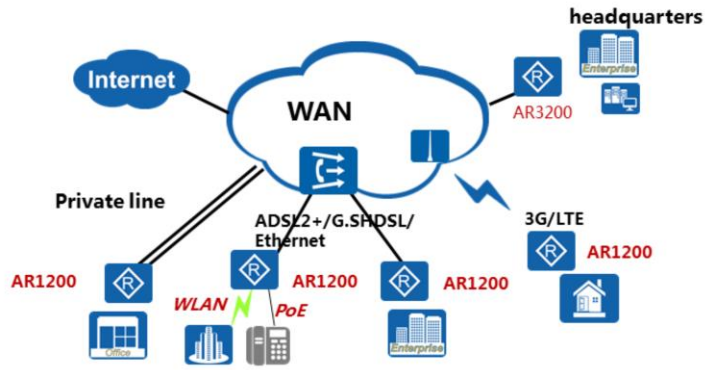
V.35 DCE cable



V.35 DTE cable

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

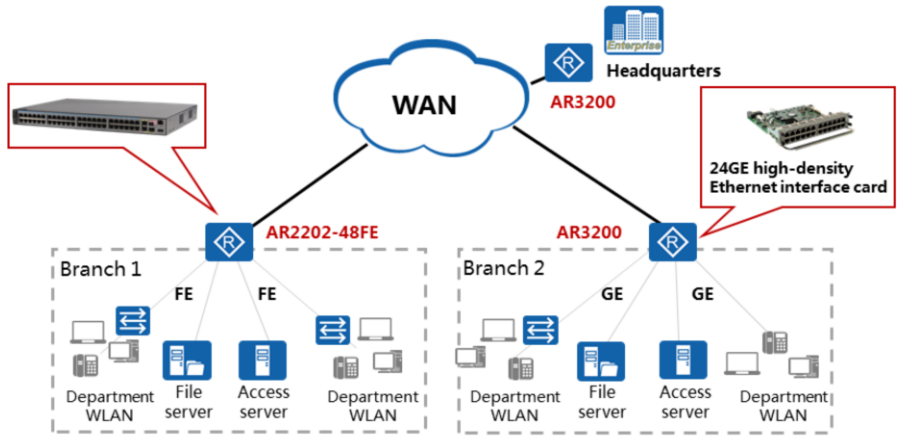
Broadband Access Scenarios of Enterprise Routers



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- AR G3 series enterprise routers function as the egress routers on enterprise branch networks and provide flexible access modes to implement remote network interconnection.
- AR series routers meet various access requirements, including private line, Ethernet, xDSL, 3G, and WLAN. This saves deployment and O&M costs and provides more benefits for customers.
- 100M Ethernet interfaces of the AR1220V and AR1220W (V200R001C01) support the PoE function in compliance with IEEE 802.3af and 802.3at; therefore, these routers can provide PoE power for remote powered devices (PDs), such as IP phones. An 802.3at interface provides more than 30 W power, meeting power supply requirements of large-power PDs.

Integration of the Access, Routing, and Switching Functions for Enterprise Branches

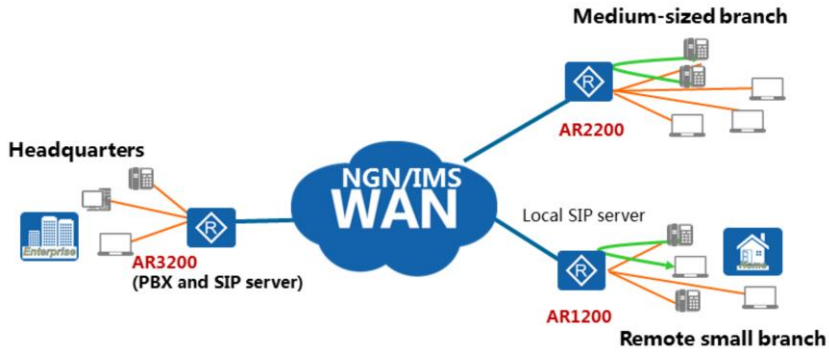


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- AR2200 and AR3200 series enterprise routers provide cards with eight FE ports and one GE combo port as well as cards with twenty-four GE ports to implement inter-card VLAN switching, mirroring, spanning tree, and link bundling, as well as Layer 2 and Layer 3 data exchange.

Enterprise VoIP



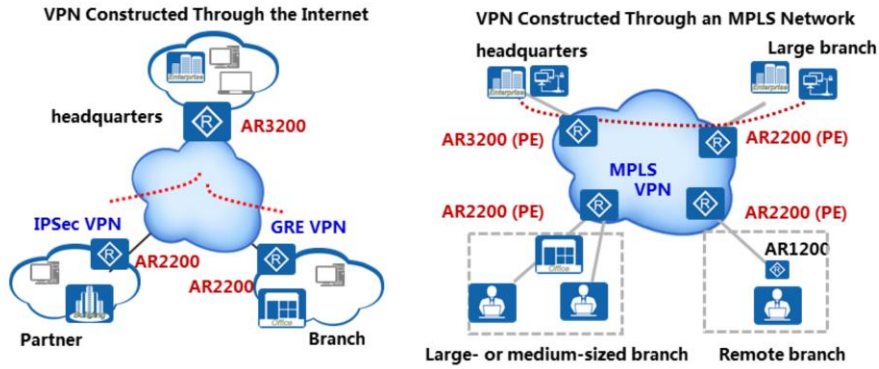
- An AR router has a built-in PBX and provides voice communication services such as the switchboard, IVR, and bill query.
- An AR router has a built-in SIP server, which ensures reliability of enterprise voice services.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- AR G3 series routers have a built-in PBX, and provide voice communication services such as the switchboard, IVR, and bill query to enhance enterprise image and improve enterprise communication efficiency.
- If the SIP server at the headquarters is unreachable, the built-in SIP server of the AR router implements communication between branches and between branches and NGN/IMS. This ensures reliability of voice services.
- Note: AR2200 and AR3200 series routers running V200R001C01 support enterprise VoIP.

Enterprise VPN Gateway Scenario

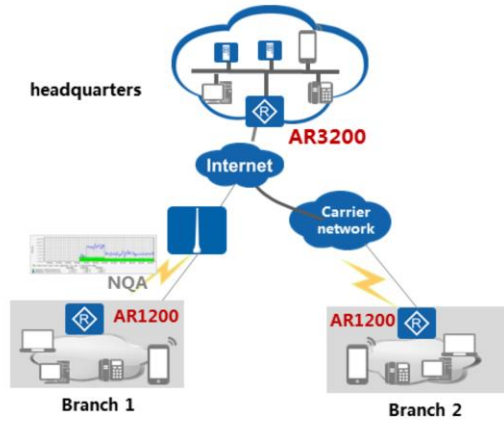


- AR G3 series enterprise routers provide GRE VPN and IPSec VPN security tunnels to implement secure data access and transmission, as well as fast deployment of tunnels and tunnel authentication for branches.
- AR G3 series routers can be deployed at branches as PEs on the MPLS network. Different services are isolated by the Layer 3 MPLS VPN, implementing virtualized operation of enterprise services.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- AR G3 series enterprise routers provide multiple security access functions, including GRE VPN and IPSec VPN security tunnels, to implement secure data access and transmission, as well as fast deployment of tunnels and tunnel authentication for branches. Through remote tunnel access, partners can access internal resources of the enterprise. Security authentication and authorization for users are supported.
- AR G3 series routers can also be deployed at branches as PEs on the MPLS network. Different services are isolated by the Layer 3 MPLS VPN to implement flexible deployment, fast forwarding, and secure transmission of VPN services, implementing virtualized operation of enterprise services.

3G/LTE Wireless Access Scenario



- AR G3 series routers comply with 3G and LTE standards including CDMA2000 EV-DO, WCDMA, and TD-SCMDA, LTE FDD, and LTE TDD, meeting the wireless communication requirements between branches and between the headquarters and branches.
- Generally, wireless access is a backup for wired access.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

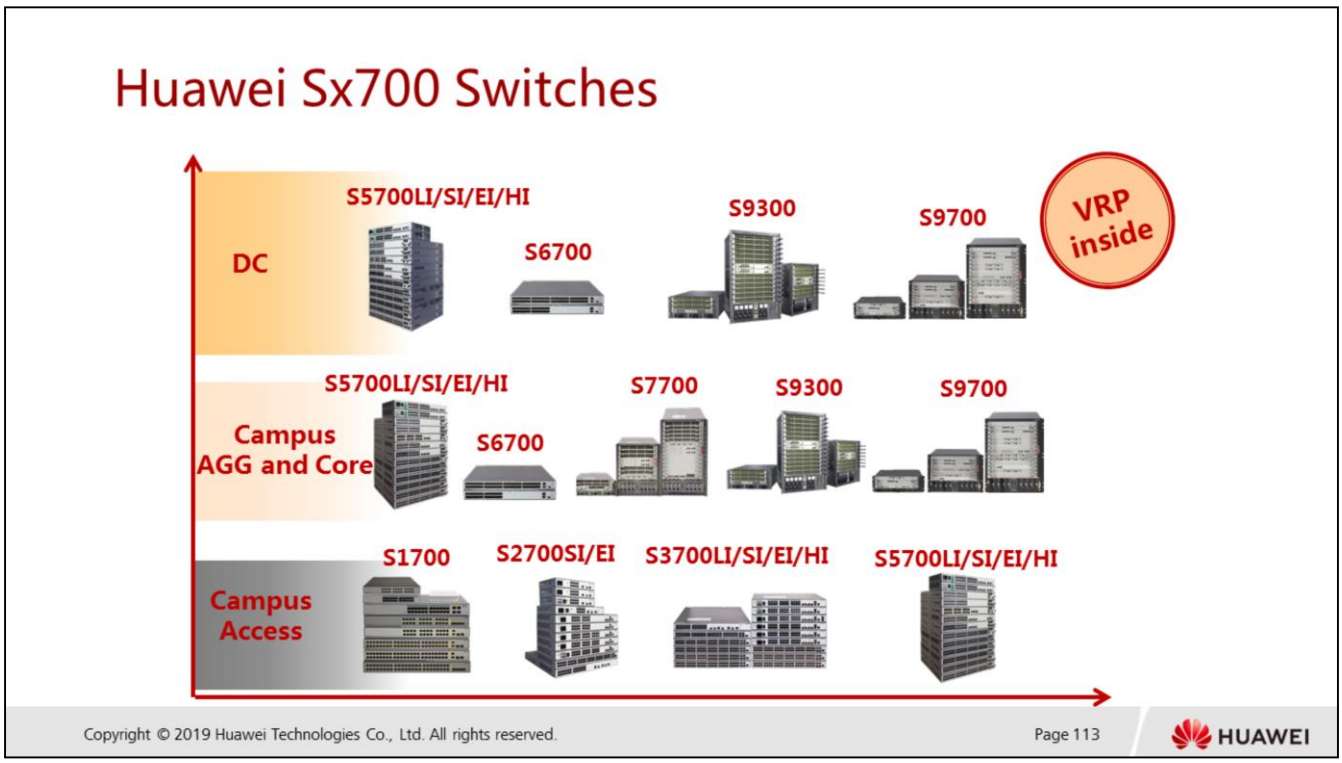
HUAWEI

- AR G3 series enterprise routers provide 3G and LTE wireless access functions, and support 3G standards including CDMA2000 EV-DO, WCDMA, and TD-SCDMA, meeting wireless interconnection requirements between enterprise branches and between the headquarters and branches. In addition, wireless data links can be used as a backup for wired links to protect the xDSL, FE/GE, GPON, and POS uplinks. Link backup improves network stability and reduces network construction costs. AR G3 series routers adopt NQA technology to detect quality of 3G and LTE links in real time, ensuring the SLA.

Contents

1. Enterprise Network Design Principles
2. Enterprise Network Design Overview
3. **Product Selection**
   - How to Select Products
   - AR G3 Series Routers
   - Sx700 Series Switches
   - Other Products
4. Enterprise Network Design Case Analysis

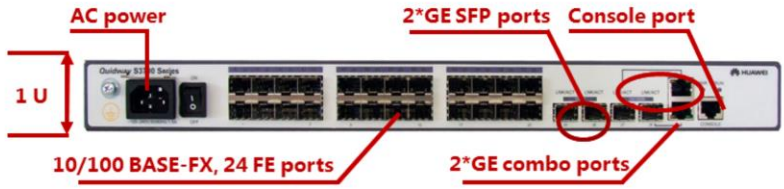
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



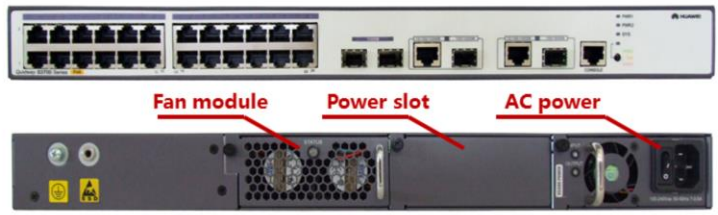
- Huawei Sx700 series switches are next-generation intelligent switches designed for enterprise campus networks. They can be deployed at the core, aggregation, and access layers, meeting flexible networking requirements of enterprises.
- Sx700 series switches include:
  - S9700 series terabit core routing switches
  - S7700 series intelligent routing switches
  - S6700 series data center 10GE switches
  - S5700 series gigabit enterprise switches
  - S3700 series Layer 3 100M enterprise switches
  - S2700 series Layer 2 100M enterprise switches
  - S1700 series SMB switches

S3700 Series Switches (1/4)

- **S3700-28TP-EI-24S-AC**



- **S3700-28TP-PWR-EI**



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

S3700 Series Switches (2/4)

S3700-26C-HI



- Twenty-two 10/100BASE-TX ports, two GE combo ports (10/100/1000BASE-T or 100/1000BASE-X), and 2-port 100BASE-X card (optional)
- Dual power modules, pluggable, DC or AC power supply
- Packet forwarding capacity: 9.3 Mpps
- Switching capacity: 64 Gbit/s

S3700-28TP-SI/EI



- Twenty-four 10/100BASE-TX ports, two 1000BASE-X SFP ports, and two GE combo ports (10/100/1000BASE-T or 100/1000BASE-X)
- AC or DC power supply
- Packet forwarding capacity: 9.6 Mpps
- Switching capacity: 64 Gbit/s

S3700-28TP-EI-MC-AC



- Twenty-four 10/100BASE-TX ports, two 1000BASE-X SFP ports, and two GE combo ports (10/100/1000BASE-T or 100/1000BASE-X)
- AC power supply
- Packet forwarding capacity: 9.6 Mpps
- Switching capacity: 64 Gbit/s

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

S3700 Series Switches (3/4)

S3700-28TP-PWR-EI



- Twenty-four 10/100BASE-TX ports, two 1000BASE-X SFP ports, and two GE combo ports (10/100/1000BASE-T or 100/1000BASE-X)
- AC power supply
- PoE+ supported
- Packet forwarding capacity: 9.6 Mpps
- Switching capacity: 64 Gbit/s

S3700-28TP-EI-24S-AC



- Twenty-four 100BASE-FX SFP ports, two 1000BASE-X SFP ports, and two GE combo ports (10/100/1000BASE-T or 100/1000BASE-X)
- AC power supply
- Packet forwarding capacity: 9.6 Mpps
- Switching capacity: 64 Gbit/s

S3700-52P-SI-AC



- Forty-eight 10/100BASE-TX ports, two 100/1000BASE-X SFP ports, and two 1000BASE-X SFP ports
- AC power supply
- Packet forwarding capacity: 13.2 Mpps
- Switching capacity: 64 Gbit/s

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

S3700 Series Switches (4/4)

S3700-52P-EI



- Forty-eight 10/100BASE-TX ports, two 100/1000BASE-X SFP ports, and two 1000BASE-X SFP ports
- AC or DC power supply
- Packet forwarding capacity: 13.2 Mpps
- Switching capacity: 64 Gbit/s

S3700-52P-PWR-EI



- Forty-eight 10/100BASE-TX ports, two 100/1000BASE-X SFP ports, and two 1000BASE-X SFP ports
- AC power supply
- PoE+ supported
- Packet forwarding capacity: 13.2 Mpps
- Switching capacity: 64 Gbit/s

S3700-52P-EI-48S



- Forty-eight 100BASE-FX SFP ports, two 100/1000BASE-X SFP ports, and two 1000BASE-X SFP ports
- AC or DC power supply
- Packet forwarding capacity: 13.2 Mpps
- Switching capacity: 64 Gbit/s

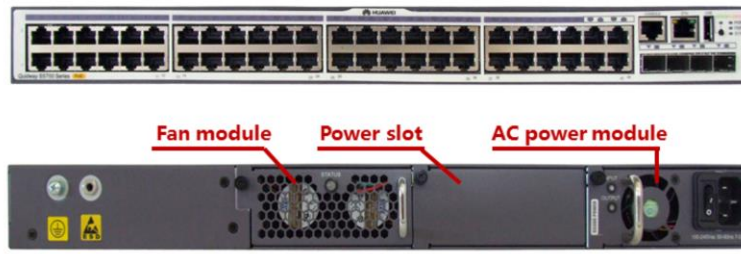
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

S5700 Series Switches (1/4)

- **S5700-24TP-SI-AC**



- **S5700-48TP-PWR-SI**



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The height of an S5700 series chassis is 1 U (1 U = 44.45 mm).
- The dimensions (W x D x H) of the S5700-24TP-SI-AC, S5700-24TP-SI-DC, S5700-28C-HI, and S5700-28C-HI-24S chassis are 442.0 mm × 220.0 mm × 43.6 mm.
- The dimensions (W x D x H) of the S5700-6TP-LI-AC chassis are 250.0 mm × 180.0 mm × 43.6 mm.
- Except the preceding chassis, the dimensions (W x D x H) of other chassis are 442.0 mm × 420 mm x 43.6 mm.
- S5700-EI series switches support uplink cards to provide high-density and flexible GE/10GE uplink ports.
- An S5710-EI series switch provides four fixed 10GE SPF+ ports. It can use uplink cards to implement a combination of 64*GE+4*10GE, 48*GE+8*10GE, or 56*GE+6*10GE, meeting different bandwidth upgrade requirements of customers and protecting customers' investment.

S5700 Series Switches (2/4)

S5700-24TP-SI



- Twenty 10/100/1000BASE-T ports and four 100/1000BASE-X GE combo ports
- AC or DC power supply, RPS supported (12 V redundant power supply)
- USB port
- Packet forwarding capacity: 36 Mpps
- Switching capacity: 256 Gbit/s

S5700-24TP-PWR-SI



- Twenty 10/100/1000BASE-T ports and four 100/1000BASE-X GE combo ports
- Dual pluggable AC power modules
- PoE+ supported
- USB port
- Packet forwarding capacity: 36 Mpps
- Switching capacity: 256 Gbit/s

S5700-48TP-SI



- Forty-four 10/100/1000BASE-T ports and four 100/1000BASE-X GE combo ports
- AC or DC power supply, RPS supported (12 V redundant power supply)
- USB port
- Packet forwarding capacity: 72 Mpps
- Switching capacity: 256 Gbit/s

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

S5700 Series Switches (3/4)

S5700-48TP-PWR-SI



- Forty-four 10/100/1000BASE-T ports and four 100/1000BASE-X GE combo ports
- AC power supply
- PoE+ supported
- USB port
- Packet forwarding capacity: 72 Mpps
- Switching capacity: 256 Gbit/s

S5700-28C-SI



- Twenty 10/100/1000BASE-T ports, four 100/1000BASE-X GE combo ports, 4-port 1000BASE-X SFP card, 2-port 10GE SFP+ card, and 4-port 10GE SFP+ card
- Dual pluggable power modules
- USB port
- Packet forwarding capacity: 96 Mpps
- Switching capacity: 256 Gbit/s

S5700-28C-EI



- Twenty-four 10/100/1000BASE-T ports, 4-port 1000BASE-X SFP card, 2-port 10GE SFP+ card, and 4-port 10GE SFP+ card
- Dual pluggable power modules
- Packet forwarding capacity: 96 Mpps
- Switching capacity: 256 Gbit/s

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

S5700 Series Switches (4/4)

S5700-28C-EI-24S



- Twenty 100/1000BASE-X ports, four 10/100/1000BASE-T GE combo ports, 4-port 1000BASE-X SFP card, 2-port 10GE SFP+ card, and 4-port 10GE SFP+ card
- Dual pluggable power modules
- Packet forwarding capacity: 96 Mpps
- Switching capacity: 256 Gbit/s

S5700-28C-PWR-EI



- Twenty-four 10/100/1000BASE-T ports, 4-port 1000BASE-X SFP card, 2-port 10GE SFP+ card, and 4-port 10GE SFP+ card
- Dual pluggable AC power modules
- PoE+ supported
- Packet forwarding capacity: 96 Mpps
- Switching capacity: 256 Gbit/s

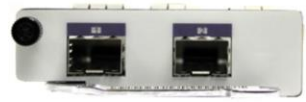
S5700-52C-SI



- Forty-eight 10/100/1000BASE-T ports, 4-port 1000BASE-X SFP card, 2-port 10GE SFP+ card, and 4-port 10GE SFP+ card
- Dual pluggable power modules
- USB port
- Forwarding performance: 132 Mpps
- Switching capacity: 256 Gbit/s

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Card Supported by S3700 Series Switches



Card Type	Description	Device Model
Front card	G2S 2-port GE SFP optical card	S3700-26C-HI

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The G2S card provides two 1000M SFP optical ports to implement data access and line-rate switching.
- The G2S card is controlled by the main control board of the S3700-HI. It supports power-on and power-off control, in-position detection, PHY and optical port management, and enhanced service features such as OAM and BFD.
- The G2S card can be inserted into the front card slot of the S3700-HI and is hot swappable.

Cards Supported by S5700 Series Switches



E2XX 2-port 10GE XFP optical card



E2XY 2-port 10GE SFP+ optical card



E4XY 4-port 10GE SFP+ optical card



X4SA 4-port GE SFP/10G SFP+ optical card



E4GF/E4GFA 4-port GE optical card

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

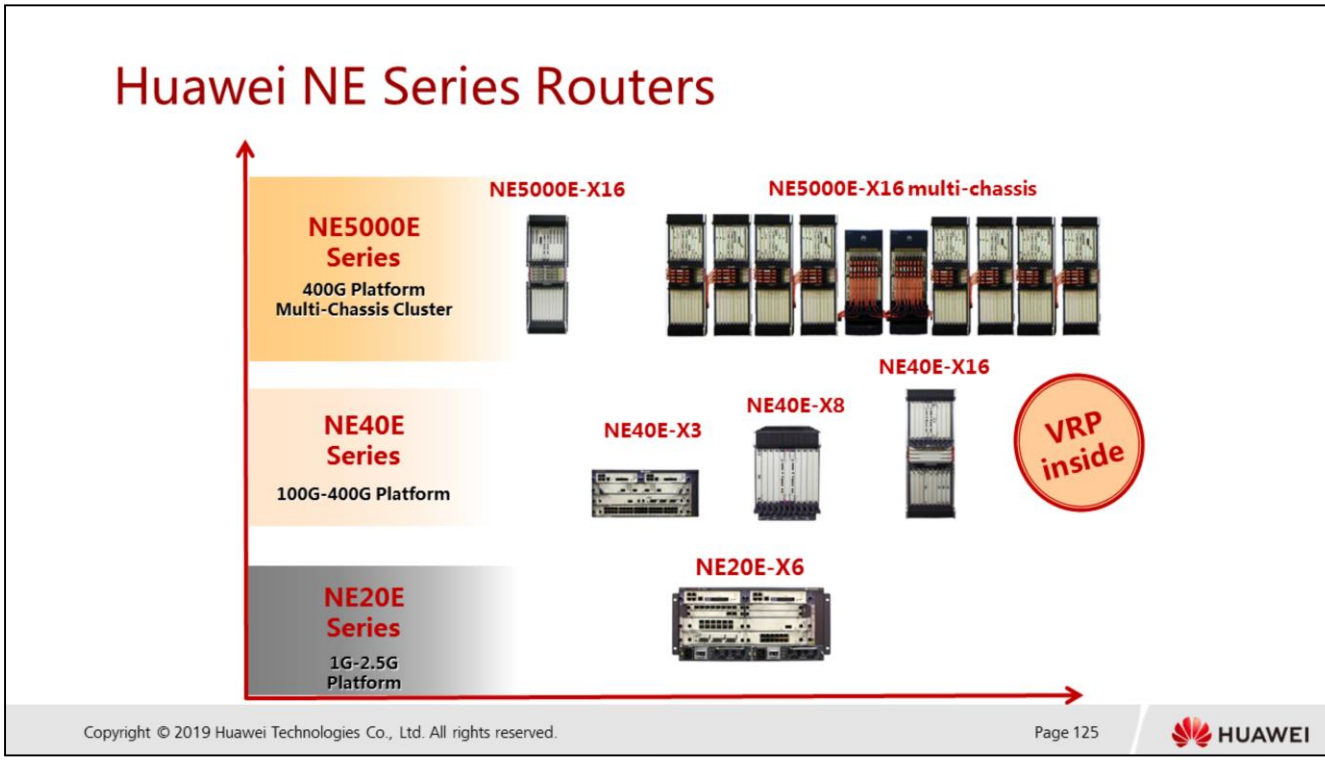
HUAWEI

- The E2XX card is applicable to the S5700-28C-EI, S5700-52C-EI, S5700-28C-EI-24S, S5700-28C-SI, S5700-52C-SI, and S5700-28C-PWR-EI.
- The E2XY card is applicable to the S5700-52C-PWR-EI.
- The E4XY card is applicable to the S5700-28C-EI, S5700-52C-EI, S5700-28C-EI-24S, S5700-28C-SI, and S5700-52C-SI.
- The E4GFA card is applicable to the S5700SI.
- The E4GF card is applicable to the S5700EI.
- Among S5700C series switches, only the S5700-28C-HI, S5700-28C-HI-24S, S5710-28C-EI, and S5710-52C-EI support hot swappable cards.

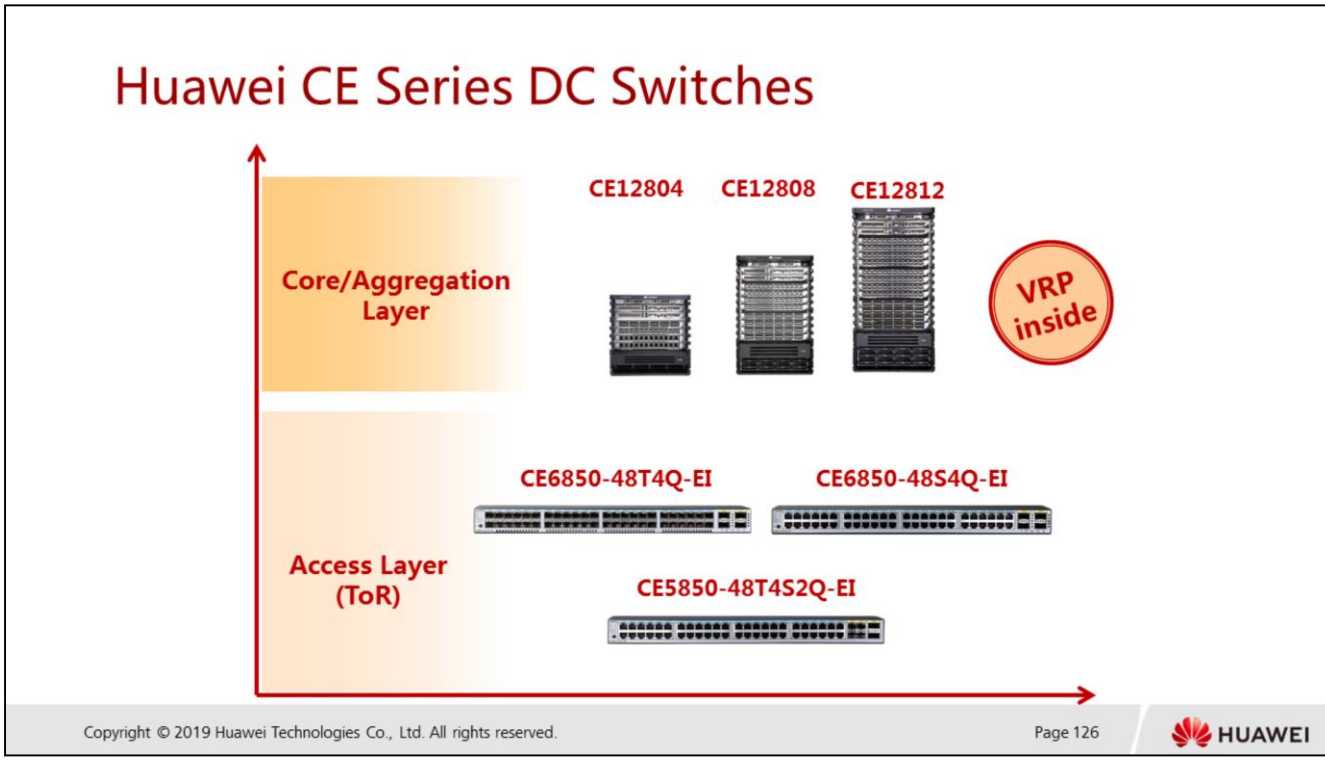
Contents

1. Enterprise Network Design Principles
2. Enterprise Network Design Overview
3. **Product Selection**
   - How to Select Products
   - AR G3 Series Routers
   - Sx700 Series Switches
   - Other Products
4. Enterprise Network Design Case Analysis

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

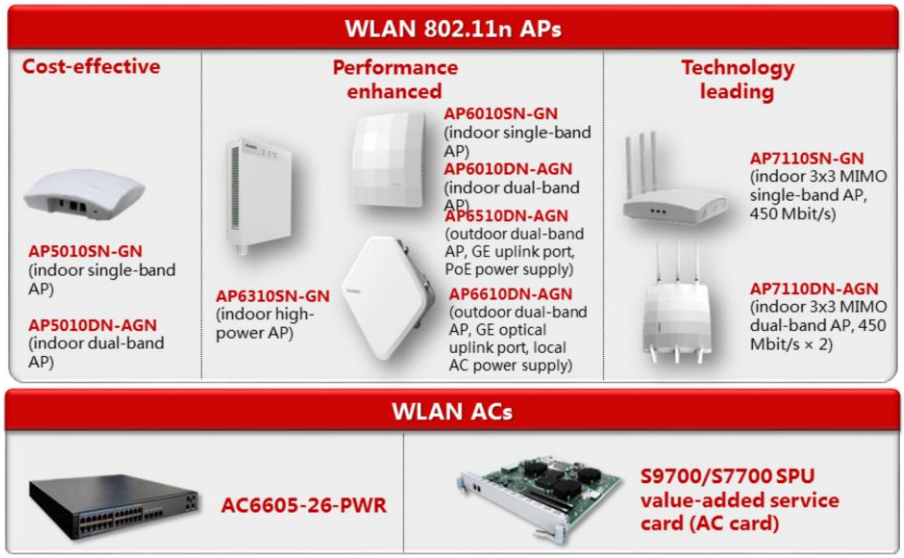


- The NetEngine5000E cluster router (NE5000E for short) is a super core router developed by Huawei for Internet backbone nodes, MAN core nodes, DCI nodes, and Internet bearer nodes. The NE5000E adopts the Huawei-developed Solar series chips, advanced fabric architecture, and a distributed, scalable software platform to deliver a mass switching capacity and ultra-high forwarding performance, meeting bandwidth performance, service quality, and capability needs for the next generation of the Internet.
- The NetEngine40E series universal service router (NE40E for short) is a high-end network product provided by Huawei. NE40Es are usually deployed at the edges of IP backbone networks, IP MANs, and other large-scale IP networks. The NE40E, NE5000E, and ME60 together provide a complete, layered IP network solution.
- The NetEngine20E-X6 router (NE20E-X6 for short) is a high-end service router developed by Huawei for enterprises and users in the financial, power, government, and education industries to meet high reliability and availability requirements on aggregation and access networks.



- Huawei CloudEngine series switches are high-performance cloud switches designed for next-generation data centers and high-end campus networks, which include CloudEngine 12800 flagship core switches with the highest performance in the world, and CloudEngine 6800 and 5800 high-performance fixed switches for 10GE/GE access. CloudEngine series switches use Huawei's next-generation VRP8 software platform to support extensive service features for data center networks and campus networks.
- Huawei has launched the CloudFabric solution for next-generation cloud-computing data centers by combining CloudEngine series switches with transmission, routing, security, and network management products of Huawei. This solution helps customers build a stable network architecture that supports development for the next 10 years.

Huawei WLAN Products

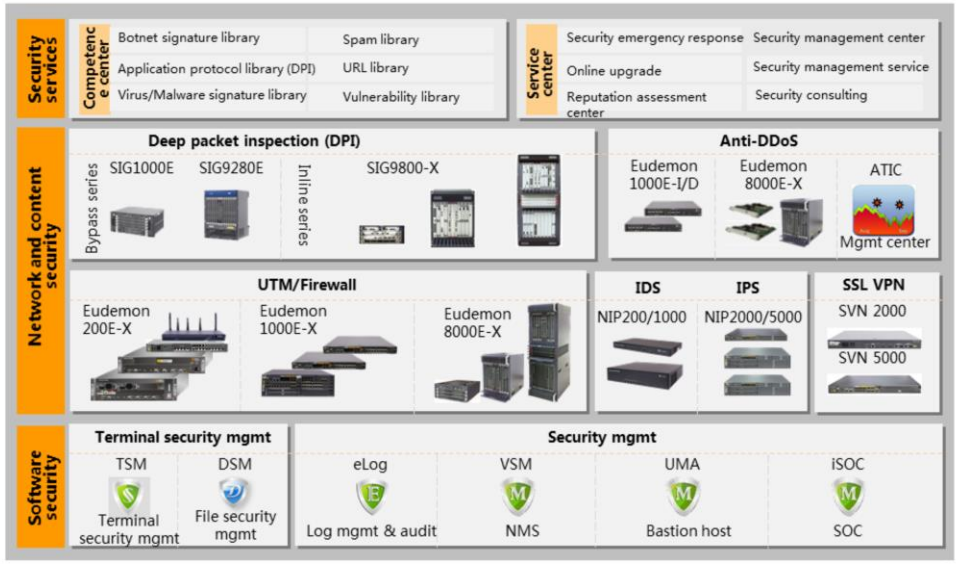


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Huawei next-generation 802.11n AP series is designed for enterprise users. These APs are compatible with 802.11a, 802.11b, and 802.11g standards. To meet the requirements of different types and sizes of enterprises, Huawei provides technology-leading APs, performance-enhanced APs, and cost-effective APs.
  - Technology-leading APs adopt the latest and most mature WLAN technologies to provide highest-performance WLAN services in high-density scenarios of medium- and large-sized enterprises.
  - Performance-enhanced APs adopt the latest generation of chip design to provide multiple services and applications for medium- and large-sized enterprises.
  - Cost-effective APs provide basic 802.11n WLAN access for small- and medium-sized enterprises and enterprise branches.
- Huawei provides two types of ACs: fixed AC and AC card. They are applicable to different wired network architectures of large campuses, enterprise branches, and small enterprises. These ACs are secure, reliable, easy-to-manage, and efficient.

Huawei Network Security Products

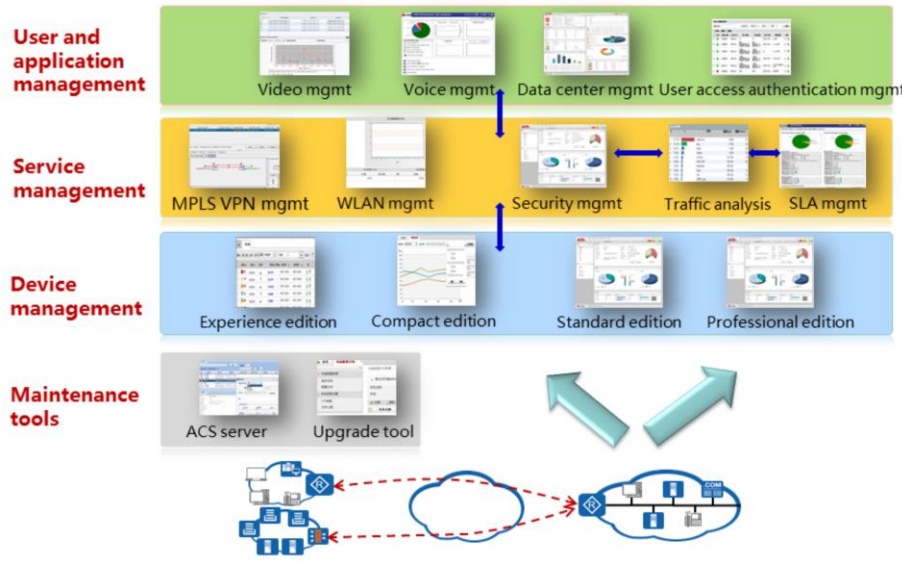


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Deep packet inspection (DPI): SIG series products
- IDS and IPS: NIP series products
- Anti-DDoS products: Eudemon 1000E-I/D and Eudemon 8000E-X
- UTM and firewall: Eudemon 200E-X Eudemon 1000E-X, and Eudemon 8000E-X
- SSL VPN: SVN series products
- Terminal security management: TSM and DSM
- Security management: eLog, VSM, UMA, and iSOC

Huawei eSight



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- eSight is a next-generation IP and IT unified network O&M solution developed by Huawei for enterprise markets. It complies with ITIL standards and implements centralized management of enterprise resources, services, and users, providing an integrated and open O&M platform for enterprises and partners.

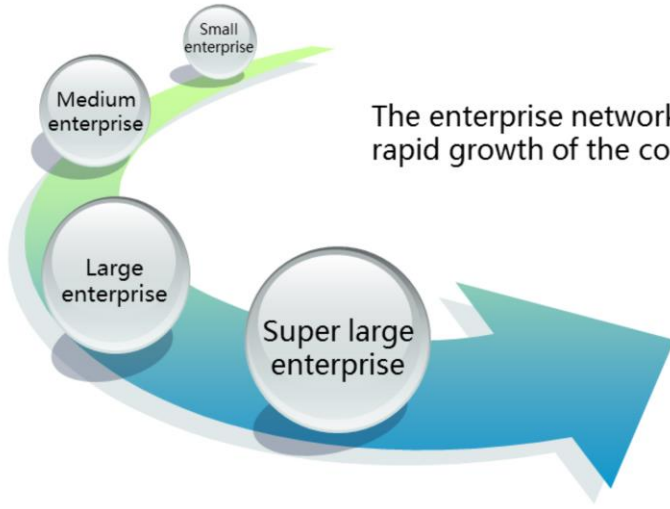


Contents

1. Enterprise Network Design Principles
2. Enterprise Network Design Overview
3. Product Selection
4. **Enterprise Network Design Case Analysis**

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Development of Company A



The enterprise network empowers rapid growth of the company.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

When A Was a Small Company...

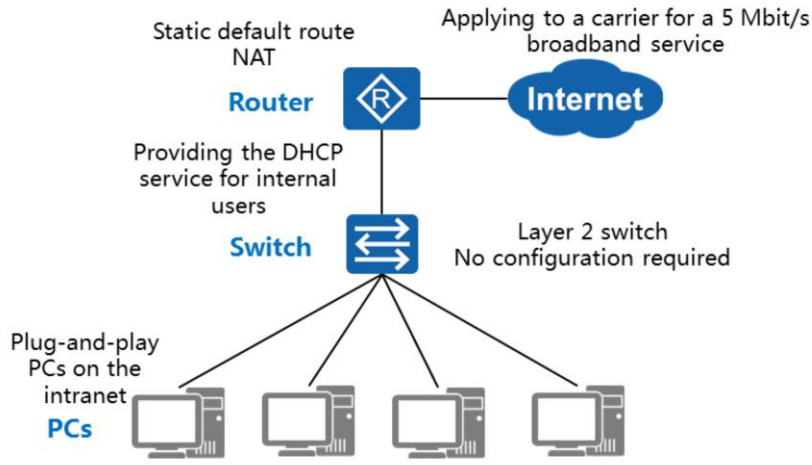
- Company A had only over 20 employees and rent an office.
- Each employee handled business affairs on PCs. Employees needed to share files frequently and surf the Internet to query information.
- Jack was responsible for network management, who was not a full-time network administrator.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Characteristics of small- and medium-sized enterprises
  - Single services
    - Small- and medium-sized enterprises usually provide single services or a mix of multiple services.
  - No isolation
    - Since these enterprises provide single services, service isolation is not required.
    - In addition, employees of these enterprises often have multiple job duties, so services cannot be isolated.
  - Centralized services
    - The personnel and fixed assets of small- and medium-sized enterprises are often concentrated in a small area, for example, an office or a building.
  - Simple requirements
    - Small- and medium-sized enterprises have simple requirements for networks as long as basic service interworking is achieved.

Jack's Solution



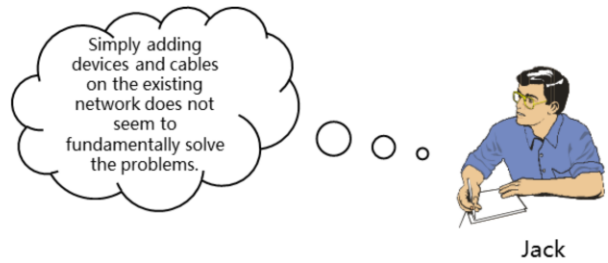
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The network architecture of small enterprises is simple.
- An office network is a typical small network.
  - The network devices include only one router and one switch, or a routing and switching device (such as an AR G3).
  - A static default route and NAT are configured on the router for connection to the Internet.
  - The switch uses the default configuration or has simple VLAN assignment.
  - QoS is not guaranteed.
  - There are a small number of PCs, so their IP addresses can be manually configured.
  - No dedicated personnel are available for network maintenance.

When the Company Scale Expanded...

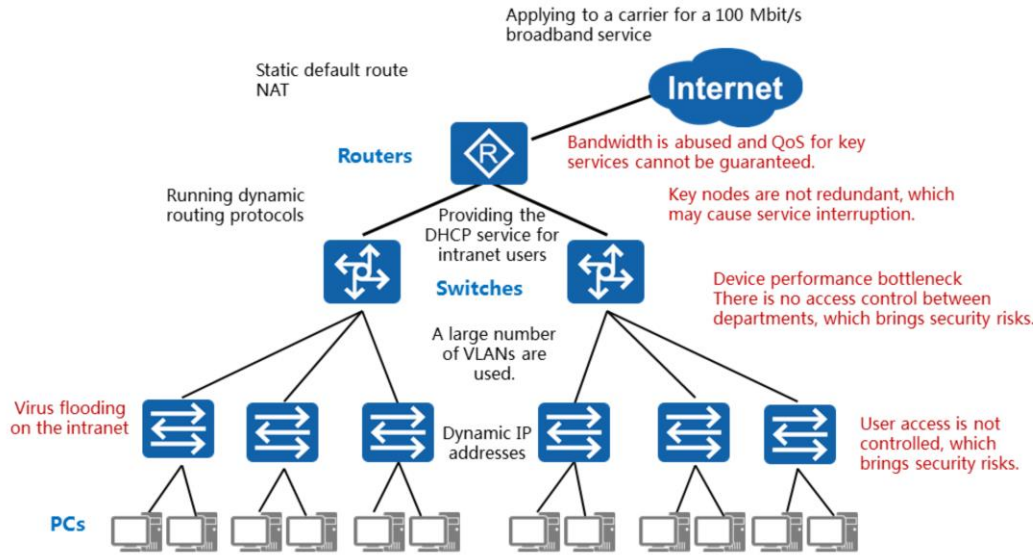
- With the development of the company's businesses, the requirements for the network are increasing.
- To quickly meet and respond to business development requirements, Jack gradually expanded the network.
- However, when the network reached a certain scale, more problems occurred.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- In the early stage of network development, it is convenient to add devices and cables to the existing network to increase the number of access users.
- However, with the expansion of the network scale, various problems are gradually exposed.

Status of the Existing Network



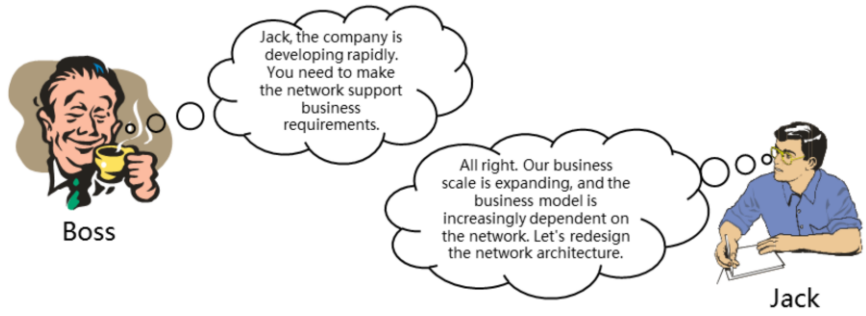
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The development of a network is not just adding devices and connecting cables. There will be many problems if a large-scale network is constructed according to the structure of a small network.

When Company A Has Grown into a Medium-Sized Enterprise...

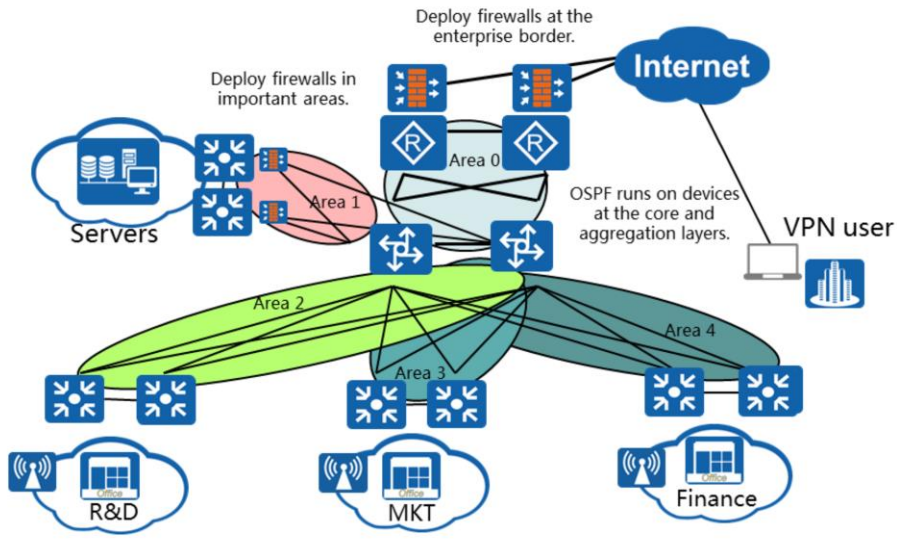
- The enterprise network construction greatly improves the operation efficiency of the company. Company A rapidly grows into a medium-sized enterprise with hundreds of employees.
- Jack has reconstructed the network, but it still cannot meet the requirements of enterprise business development. It is a must to change the network architecture.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- The enterprise network construction greatly improves the operation efficiency of the company. Company A rapidly has grown into a medium-sized enterprise with hundreds of employees.
- Jack has reconstructed the network as follows:
  - As the number of access users increases, Jack adds a large number of Layer 2 switches to the network for user access.
  - Because of the existence of many access switches, Jack deploys aggregation switches and assigns VLANs.
  - As the service volume increases, the egress router is upgraded and a larger bandwidth is leased.
- However, the network still cannot meet requirements of enterprise business development.

Jack's Solution



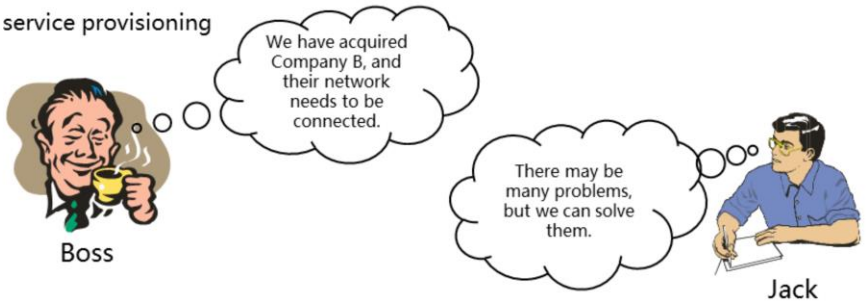
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Different service departments of an enterprise may be located in different campuses. Each campus is a small- and medium-sized network, which has the following characteristics:
  - A large number of switches are deployed on the network.
  - The entire network has a clear layered structure, and each subnet also has a layered structure.
  - Because the network structure is complex, static routes cannot meet the requirements, and a dynamic routing protocol (such as OSPF) is used.
  - Different services are distributed in different areas, and firewalls are used to isolate the areas.
  - As the IT system is becoming more and more important for enterprises, a dedicated server zone is built.
  - Enterprise services rely more on networks, so the redundancy design is adopted for important nodes.
  - Firewalls are deployed at enterprise borders and in important server areas.
  - To meet the requirements of mobile office, a WLAN is deployed using Fit APs in the office area. The AC centrally manages the APs.
  - Dedicated personnel are responsible for network maintenance.

For Years to Come

- With the development of Company A, the enterprise network is also faced with great challenges.
  - More complex control policies
  - Stricter security policies
  - More detailed QoS policies
  - Faster service provisioning



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

- Open discussions:
  - What problems may occur when two networks are integrated?
  - How can we solve these problems?

Summary

- Enterprise networks must be designed based on enterprise services.
- The enterprise network design needs to comply with general rules.
- Future expansion or trends need to be considered during the design of enterprise networks.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Thank You

www.huawei.com

HCIE-Routing & Switching

— Trouble-Shooting Case Analysis

www.huawei.com

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- This chapter describes the troubleshooting procedure and methods for IP routing protocols of routers.
- Based on the cases, the common faults of OSPF, IS-IS, BGP, and MPLS VPN are discussed and analyzed in depth, helping you further understand the troubleshooting of routing protocols.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

- On completion of this course, you will be able to:
  - Describe typical troubleshooting methods of OSPF, IS-IS, BGP, and MPLS BGP VPN.
  - Describes troubleshooting roadmaps of OSPF, IS-IS, BGP, and MPLS BGP VPN.
  - Analyze and rectify faults of OSPF, IS-IS, BGP, and MPLS BGP VPN.

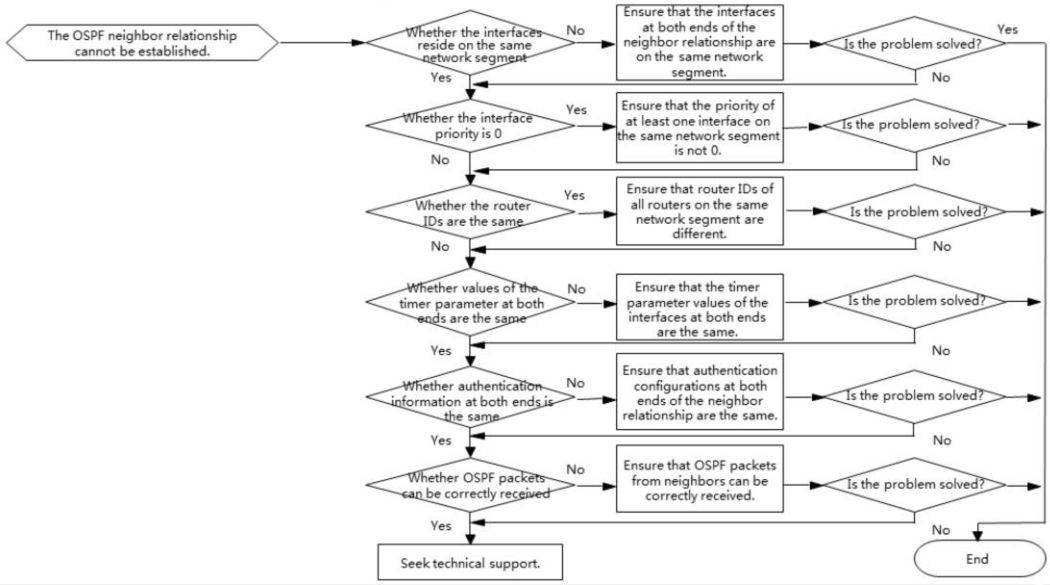
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Contents

1. **OSPF Troubleshooting**
   - Typical OSPF Troubleshooting Roadmap
   - OSPF Troubleshooting Cases
2. IS-IS Troubleshooting
3. BGP Troubleshooting
4. MPLS BGP VPN Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Troubleshooting Process

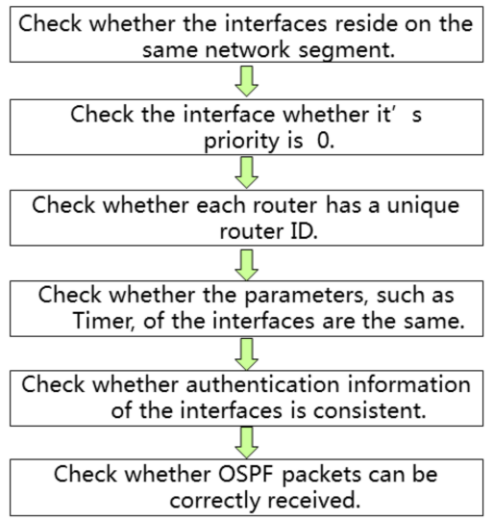


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- OSPF neighbor relationship establishment failures are the most common OSPF fault.
- If the OSPF neighbor relationship fails to be established after the routers are configured, follow the troubleshooting process as shown in the flowchart.

Troubleshooting Procedure



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Step 1: Check whether the interfaces reside on the same network segment.
  - To establish an OSPF neighbor relationship, ensure that the broadcast and NBMA interfaces are on the same network segment. The two firewalls must be able to ping each other, and the interfaces must have the same area ID and area type (including NSSA, Stub, and Normal Area).
- Step 2: Check the interface whether it' s priority is 0.
  - On broadcast and NBMA networks, there must be at least one interface whose priority is not 0, ensuring that the DR can be elected correctly. Otherwise, the neighbor relationship can only reach the two-way state. You can run the display ospf interface command to check the priority of each interface.
- Step 3: Check whether each router has a unique router ID.
  - The router ID of each router in the same autonomous system (AS) should be different from each other. Otherwise, unexpected route flapping will occur. You can run the display ospf brief command to check the router ID of each router.

- Step 4: Check whether the parameters, such as Timer, of the interfaces are the same.
  - Run the ospf timer hello command to set the interval for the interfaces to send Hello packets. By default, the interval for sending Hello packets on a Point-to-Point (P2P) or broadcast interface is 10 seconds. The interval for sending Hello packets on a Point-to-Multipoint (P2MP) or NBMA interface is 30 seconds.
  - Run the ospf timer dead command to set the dead interval of OSPF neighbor relationships. By default, the dead interval of OSPF neighbors on P2P and broadcast interfaces is 40 seconds, and the dead interval of OSPF neighbors on P2MP and NBMA interfaces is 120s.
  - Before setting up the OSPF neighbor relationship, make sure that the interval parameters must be consistent on the related interfaces; otherwise, the OSPF neighbor relationship cannot be set up. You can run the display ospf interface verbose command to check the interval parameter.
- Step 5: Check whether the authentication information of the interfaces is consistent.
  - Authentication information can be configured for an area or an interface.
  - The basic principle of OSPF authentication is as follows:
    - If authentication is set on the interface, the authentication configuration takes effect.
    - If Null is set on the interface, the interface is not authenticated.
    - If authentication is not configured on interfaces (Null is an authentication

mode), area authentication is used.

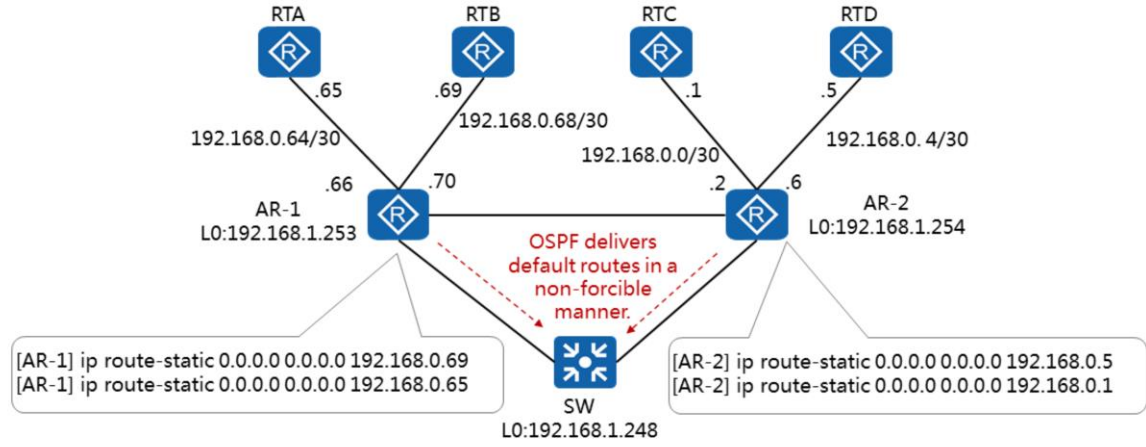
        - If neither interface or area authentication is configured, no authentication is performed.
    - During the establishment of the OSPF neighbor relationship, the OSPF neighbor relationship can enter the Full state only when the authentication configurations of the two ends are consistent.
- Step 6: Check whether OSPF packets can be received correctly.
    - OSPF packets sometimes cannot be received correctly. In this case, check the connectivity at the link layer first. You can run related debugging commands such as the **debugging ospf packet** and **debugging ospf event** commands to view the sending and receiving of OSPF packets. You can also run the **display ospf error** command to view OSPF error statistics.
    - If all OSPF packets are normal, check whether the GTSM configuration on the interface is correct. If only the private policy or the public policy is configured, and the default action of the packets that do not match the GTSM policy is pass, OSPF packets of other instances may be discarded incorrectly.
    - Run the **debugging ip packet** command to open the debug information of the IP packets and check whether IP forwarding is successful. You can also add the ACL Filter to filter the debug information.

Contents

1. **OSPF Troubleshooting**
   - Typical OSPF Troubleshooting Roadmap
   - OSPF Troubleshooting Cases
2. IS-IS Troubleshooting
3. BGP Troubleshooting
4. MPLS BGP VPN Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Topology Overview



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- AR-1 and AR-2 are the egress routers on an enterprise network. They both have two upstream GE interfaces and two default routes to guide the forwarding of upstream traffic.
- Each router uses OSPF to deliver the default routes in a non-forcible manner to the aggregation switch SW.

Fault Description 1

- Normally, the switch (SW) has two OSPF default external routes directing to the two ARs respectively.
- In the following situation, there is only one OSPF default route on the SW directing to one of the two ARs:
  - The default route with the next hop address 192.168.0.69 on AR-1 is deleted, and AR-1 has only one default route with the next hop address 192.168.0.65. AR-2 still has two static default routes to guide the forwarding of upstream traffic.

```
[AR-1]undo ip route-static 0.0.0.0 0.0.0.0 192.168.0.69
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Description 2

- The default route with the next hop address 192.168.0.65 on AR-1 is deleted, and AR-1 has only one default route with the next hop address 192.168.0.69. AR-2 still has two static default routes to guide the forwarding of upstream traffic. In this case, the preceding problem does not occur. On the SW, there are still two default routes directing to the two ARs respectively.
- Similarly, if the default route with the next hop address 192.168.0.5 on AR-2 is deleted, leaving only one default route with the next hop address 192.168.0.1 on AR-2, and AR-1 still has two static default routes to guide the forwarding of upstream traffic, there is only one OSPF default route on the SW directing to one of the two ARs.

```
[AR-2]undo ip route-static 0.0.0.0 0.0.0.0 192.168.0.5
```

- If the default route with the next hop address 192.168.0.1 on AR-2 is deleted, leaving only one default route with the next hop address 192.168.0.5 on AR-2, and AR-1 still has two static default routes to guide the forwarding of upstream traffic, the preceding problem does not occur. On the SW, there are still two default routes directing to the two ARs respectively.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Fault Analysis 1

- You can check the route learning of the SW by viewing its link state database (LSDB). Check the LSA of OSPF default routes on the SW.

```
[SW]display ospf lsdb ase
            OSPF Process 1 with Router ID 192.168.1.248
                  Link State Database
  Type      : External
  Ls id     : 0.0.0.0
  Adv rtr   : 192.168.1.254
  Ls age    : 482
  Len       : 36
  Options   :  E
  seq#      : 80000004
  chksum    : 0xeb78
  Net mask  : 0.0.0.0
  TOS 0  Metric: 1
  E type    : 2
  Forwarding Address : 0.0.0.0
  Tag       : 1
  Priority  : Low
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Analysis 2

- Check the LSA of OSPF default routes on the SW.
- The command output shows that the forwarding address (FA) is incorrectly set by AR-1. In this case, only one default OSPF route directing to AR-2 (IP address: 192.168.1.254) exists on the SW. If the AS-external (ASE) Type 5 LSAs have the same E type and cost, the intra costs of the LSAs are compared. The route with the lower intra cost is preferentially selected.

```
Type       : External
Ls id      : 0.0.0.0
Adv rtr    : 192.168.1.253
Ls age     : 149
Len        : 36
Options    :  E
seq#       : 8000000b
chksum     : 0xa50e
Net mask : 0.0.0.0
TOS 0  Metric: 1
E type     : 2
Forwarding Address : 192.168.0.65
Tag        : 1
Priority   : Low
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Based on RFC 2328 route selection rules, if the AS-external (ASE) Type 5 LSAs have the same E type and cost, the intra costs of the LSAs are compared. Specifically, the OSPF costs to the autonomous system boundary router (ASBR) or forwarding address (FA) are compared. If the FA is 0, the route is iterated to the ASBR. If the FA is not 0, the route is iterated to the FA. The route with the lower intra cost is preferentially selected, even though the intra cost will not be added to the costs in the routing table.

Fault Analysis 3

- According to the preceding analysis, the root cause of incorrect OSPF route learning on the SW is that the AR incorrectly sets the FA.
- The following describes the rules of filling in the FA in a Type 5 LSA and calculating routes on the Versatile Routing Platform (VRP):
  - When the FA is 0.0.0.0
    - When the value of the FA field of a Type 5 LSA is 0.0.0.0, the router that receives the LSA considers that the device sending the LSA is an ASBR, and calculates the next hop.
  - When the FA field is set to a value other than 0.0.0.0 and the following conditions are met, an ASBR fills in an address other than 0.0.0.0 in the FA field of a Type 5 LSA, and the router that receives the LSA calculates the next hop based on the value of the FA field.
    - The interface of the ASBR's next hop is routable.
    - The next-hop interface that connects the ASBR to an external network is not configured as a passive interface.
    - The next-hop interface that connects the ASBR to an external network is not an OSPF P2P or P2MP interface.
    - The IP address of the next-hop interface that connects the ASBR to an external network is within the OSPF network range.
  - If any one of the preceding conditions is not satisfied, the FA is set to 0.0.0.0.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- According to the preceding analysis, the root cause of incorrect OSPF route learning on the SW is that the AR incorrectly sets the FA.
- The following describes the rules of filling in the FA in a Type 5 LSA and calculating routes on the Versatile Routing Platform (VRP):
  - When the FA is 0.0.0.0
    - When the value of the FA field of a Type 5 LSA is 0.0.0.0, the router that receives the LSA knows that the device sending the LSA is an advertising router (that is, an ASBR), and calculates the next hop.
- When the FA field is set to a value other than 0.0.0.0 and the following conditions are met, an ASBR fills in an address other than 0.0.0.0 in the FA field of a Type 5 LSA, and the router that receives the LSA calculates the next hop based on the value of the FA field.
  - OSPF is enabled on the next-hop interface that connects the ASBR to an external network.
  - The next-hop interface that connects the ASBR to an external network is not configured as a passive interface.
  - The next-hop interface that connects the ASBR to an external network is not an OSPF P2P or P2MP interface.
  - The IP address of the next-hop interface that connects the ASBR to an external network is within the OSPF network range.

- If any one of the preceding conditions is not satisfied, the FA is filled with 0.0.0.0.

Troubleshooting

- Now, we find out the reason why the OSPF route calculation on the SW does not reach the expected result, and we learn about the rules of filling in the FA in a Type 5 LSA. We come to realize that an easy way to solve the problem is to interrupt one of the conditions of filling in a non-0.0.0.0 address in the FA. There are three methods to interrupt one of the conditions:
  1. The data configuration of AR-1 (IP address: 192.168.1.253) shows that the **network** command was configured to specify the interface that runs OSPF on the network segment 192.168.0.65/30, but not 192.168.0.68/30. Similarly, the data configuration of AR-2 (IP address: 192.168.1.254) shows that the **network** command was configured to specify the interface that runs OSPF on the network segment 192.168.0.0/30, but not 192.168.0.4/30. Therefore, run the **undo network** command to remove the interfaces that run OSPF from the network segments 192.168.0.65/30 and 192.168.0.4/30.
  2. You do not need to run the **undo network** command, and the configuration in the OSPF view does not need modifying. Run the **ospf network-type p2p** command on the AR interfaces that run OSPF and on their peer interfaces.
  3. Configure the corresponding interfaces on the ARs to be silent interfaces or configure routes on the SW to the interface addresses of the uplinks on the ARs.
- After you implement one of the preceding methods, the default OSPF route calculation functions properly on the SW.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Quiz

1. What should I do if the imported external routes are not displayed in the LSDB?
2. What should I do if an area border router (ABR) cannot aggregate regional network addresses?
3. What should I do if the OSPF-related LSAs are included in the LSDB but cannot be found in the routing table?

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Question 1: What should I do if the imported external routes are not displayed in the LSDB?
- Answer: The possible causes are as follows:
  - Run the **display ospf interface** command to check the OSPF interface. Ensure that the interface is not in the Down state.
  - Run the **display ospf brief** command to check whether the router that imports external routes belongs to the Stub area.
  - Run the **display ospf peer** command to check whether the neighbor status is Full if external routes are learned from neighbors.
  - Check whether the **lsdb-overflow-limit** command is configured and whether the total number of external routes exceeds the maximum value allowed by the Over-Flow-Limit.
  - Run the **display ospf asbr-summary** command to check whether the **asbr-summary** command has been configured to aggregate external routes.
- Question 2: What should I do if an area border router (ABR) cannot aggregate regional network addresses?
- Answer: The possible causes are as follows:
  - Run the **display current configuration** command to check whether the network segment addresses of the area are continuous.

- If the network segment addresses are discontinuous, divide them into several groups of continuous network segment addresses.

    - Run the **abr-summary** command to summarize each group of continuous networks into a single network on the area border router (ABR).
    - Run the **filter {acl | ip-prefix** *prefix* | route-policy route-policy-name**} {import | export}** command in the area view and ensure that the LSAs aggregated by the ABR are not filtered out.
- Question 3: What should I do if the OSPF-related LSAs are included in the LSDB but cannot be found in the routing table?
- Answer: The possible causes are as follows:
    - Check whether the IP address is correctly configured.
    - Check whether the FA is known.
    - Check whether the routes are summarized or redistributed correctly.
    - Check whether route lists are advertised.
    - Check whether the backbone area is interrupted.

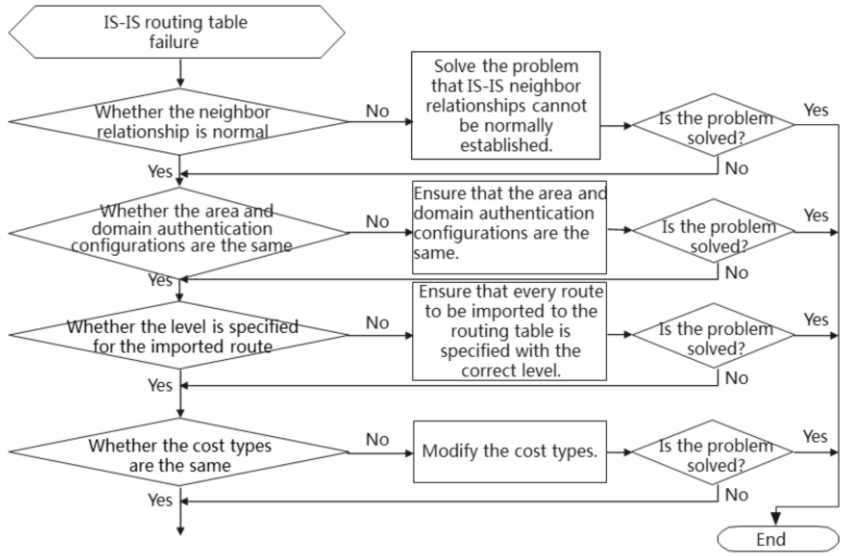


Contents

1. OSPF Troubleshooting
2. **IS-IS Troubleshooting**
   - Typical IS-IS Troubleshooting Roadmap
   - IS-IS Troubleshooting Cases
3. BGP Troubleshooting
4. MPLS BGP VPN Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Diagnosis Process 1

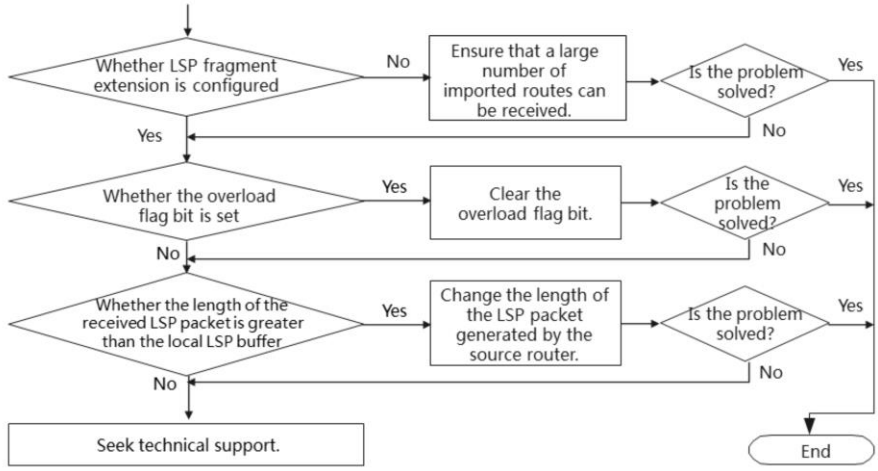


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IS-IS and OSPF are both Interior Gateway Protocols (IGPs), but IS-IS has obvious advantages in scalability (for example, IPv6 is supported). Therefore, IS-IS has been widely used.
- For details about the IS-IS fault diagnosis process, see the troubleshooting flowchart.

Fault Diagnosis Process 2

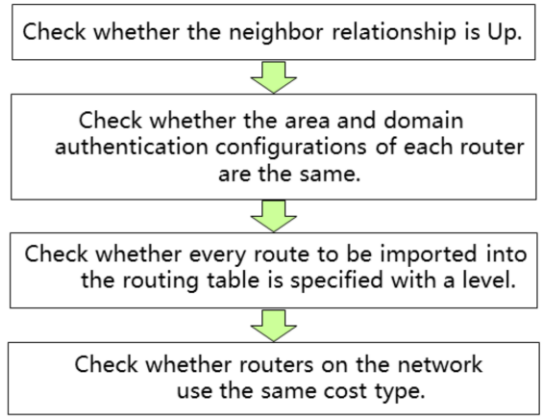


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- IS-IS and OSPF are both Interior Gateway Protocols (IGPs), but IS-IS has obvious advantages in scalability (for example, IPv6 is supported). Therefore, IS-IS has been widely used.
- For details about the IS-IS fault diagnosis process, see the troubleshooting flowchart.

Troubleshooting Procedure 1

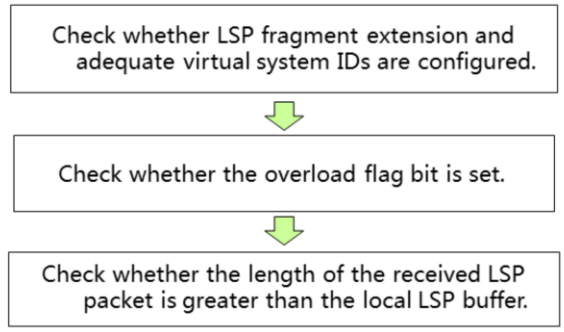


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Step 1: Check whether the neighbor relationship is Up.
  - Run the **display isis peer** command to check whether the neighbor relationship is Up.
  - If the neighbor relationship is Down, see the guide to processing the fault that the IS-IS neighbor relationship cannot be established.
- Step 2: Check whether the area and domain authentication configurations of each router are the same.
  - Run the **display isis lsdb** command to check whether the LSDB contents on two neighbors are consistent.
  - If the LSDB is not synchronized, check whether the area and domain authentication configurations are the same.
- Step 3: Check whether every route to be imported into the routing table is specified with a level.
  - If routes are imported to the Level-1 or Level-1-2 routing table, run the **display this** command in the IS-IS view to check whether the route is specified with the level.
- Step 4: Check whether routers on the network use the same cost type.

Troubleshooting Procedure 2



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

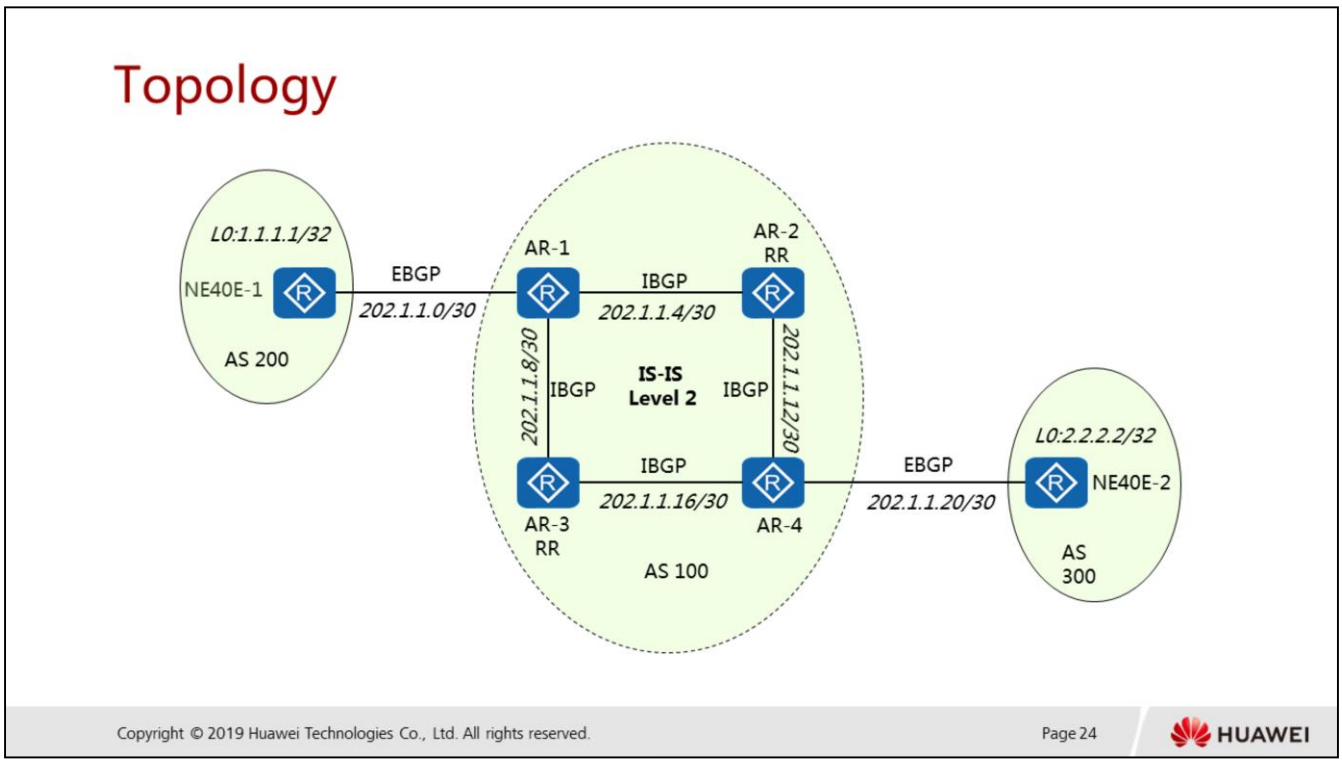
HUAWEI

- Step 5: Check whether LSP fragment extension and adequate virtual system IDs are configured.
  - Run the **display isis statistics** command to check the number of used LSP fragments in the initial system. If the number reaches 256, you need to configure LSP fragment extension and adequate virtual system IDs.
- Step 6: Check whether the overload flag bit is set.
  - If the overload flag bit is set, the LSP generated by the device notifies other devices its system database is overloaded and cannot forward packets. Other devices then do not send the packets that need to be forwarded by the device to it unless the destination address of the packets is the address of an interface directly connected to the device.
  - You can run the **undo set-overload** command to clear the overload flag bit.
- Step 7: Check whether the length of the received LSP packet is greater than the local LSP buffer.
  - If the length of the LSP packets sent by the peer is greater than the local LSP buffer, the local IS-IS discards these packets.
  - Run the **lsp-length** command to change the length of the generated LSP packets or the length of the received LSP packets.

Contents

1. OSPF Troubleshooting
2. **IS-IS Troubleshooting**
   - Typical IS-IS Troubleshooting Roadmap
   - IS-IS Troubleshooting Cases
3. BGP Troubleshooting
4. MPLS BGP VPN Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



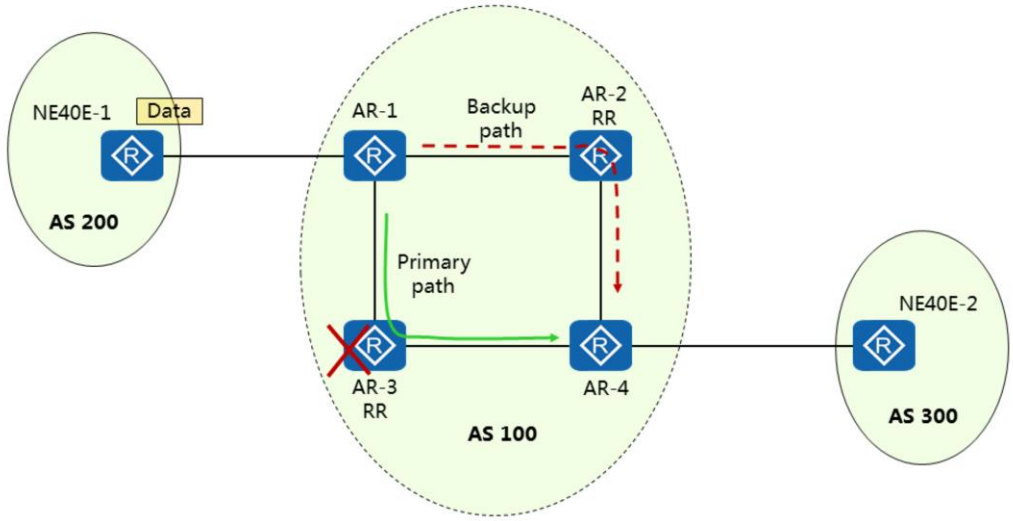
- The figure shows the networking of a large-sized enterprise campus network.
- NE40E-1 belongs to AS 200, and NE40E-2 belongs to AS 300.
- IBGP peer relationships are established between the four routers in AS 100. AR-2 and AR-3 are Border Gateway Protocol (BGP) route reflectors (RRs) that reflect routes for AR-1 and AR-4.
- AR-1 and AR-4 do not have a direct route between them, and their BGP packets must be forwarded by an RR.
- NE40E-1 in AS 200 sends data to the destination NE40E-2 through the primary path AR-1 – AR-3 – AR-4.
- The path AR-1 – AR-2 – AR-4 is the standby path.
- Adjust the cost value of the IGP so that the path AR-1 – AR-3 – AR-4 is preferentially selected to forward BGP traffic.

Fault Description

- After a fault occurred on AR-3, AR-3 restarted. IS-IS routes were fast converged, and AR-1 forwarded BGP packets to AR-2.
- After AR-3 restarted, traffic from AR-1 to AR-4 was interrupted and recovered after several minutes.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Simulation



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Analysis 1

- When services on AR-3 are interrupted, check the routing table on AR-1. In the command output, AR-1 selects the standby IS-IS router AR-2 (IP address: 202.1.1.6).

```
[AR-1]display ip routing-table 2.2.2.2 verbose
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1

    Destination:  2.2.2.2/32
       Protocol:  BGP              Process ID: 0
     Preference:  255                    Cost: 0
        NextHop:  10.1.1.4         Neighbour: 10.1.1.2
          State:  Active Adv Relied      Age: 00h00m01s
            Tag:  0                  Priority: low
          Label:  NULL               QoSInfo: 0x0
     IndirectID:  0x2
   RelayNextHop:  202.1.1.6        Interface: Ethernet0/0/0
       TunnelID:  0x0                  Flags: RD
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Fault Analysis 2

- After the AR-3 has restarted, check the routing table on AR-1 again. In the command output, AR-1 selects the active IS-IS router AR-3 (IP address: 202.1.1.10).

```
[AR-1]display ip routing-table 2.2.2.2 verbose
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1

Destination: 2.2.2.2/32
        Protocol: BGP                 Process ID: 0
      Preference: 255                       Cost: 0
         NextHop: 10.1.1.4            Neighbour: 10.1.1.2
           State: Active Adv Relied         Age: 00h04m34s
             Tag: 0                    Priority: low
           Label: NULL                 QoSInfo: 0x0
      IndirectID: 0x2
RelayNextHop: 202.1.1.10              Interface: Ethernet0/0/4
        TunnelID: 0x0                       Flags: RD
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Analysis 3

- However, BGP traffic convergence is not completed on AR-3. As a result, data packets are discarded on AR-3.

```
[AR-3]display isis peer

                Peer information for ISIS(1)

System Id       Interface  Circuit Id          State  HoldTime Type    PRI
------------------------------------------------------------------------------
0001.0001.0001  Eth0/0/2   0003.0003.0003.01 Up   2s       L2      64
0004.0004.0004  Eth0/0/3   0003.0003.0003.02 Up   1s       L2      64

Total Peer(s): 2

<AR-3>display bgp peer

 BGP local router ID : 10.1.1.3
 Local AS number : 100
 Total number of peers : 2              Peers in established state : 0

  Peer        V  AS  MsgRcvd  MsgSent  OutQ  Up/Down      State PrefRcv

  10.1.1.1  4  100     0        0        0   00:00:26      Idle     0
  10.1.1.4  4  100     0        0        0   00:00:26      Idle     0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Analysis 4

- The cause of the fault is as follows:
  - After AR-3 restarted and recovered, within seconds, the IS-IS neighbor relationships between AR-1, AR-4, and AR-3 were established and the database synchronization was completed within seconds. The forwarding information base (FIB) of AR-1 was updated, and the traffic sent to NE40E-2 was sent to AR-3 by AR-1. However, the BGP route convergence is so slow that in such a short time, AR-3 cannot learn the BGP route information about NE40E-2. As a result, AR-3 discards the packets destined for NE40E-2, and a temporary route black hole is generated.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- After AR-3 recovered, the IS-IS neighbor relationships between AR-1, AR-4, and AR-3 were established and the database synchronization was completed within seconds. The forwarding information base (FIB) of AR-1 was updated, and the traffic sent to NE40E-2 was sent to AR-3 by AR-1. However, the BGP route convergence is so slow that in such a short time, AR-3 cannot learn the BGP route information about NE40E-2. As a result, AR-3 discards the packets destined for NE40E-2, and a temporary route black hole is generated.

Troubleshooting

- On Huawei devices, you can run the following command to set the overload bit to prevent temporary route black holes:
  - **set-overload** [**on-startup** [**wait-for-bgp** [*timeout1*]] [**allow** { **interlevel** | **external** } *]
    - wait-for-bgp: sets the overload bit on system startup and the period for keeping the overload bit according to the status of the BGP convergence. If BGP does not send a signal to IS-IS indicating BGP convergence is finished, IS-IS cancels the overload bit after a specified period or default 10 minutes.
    - interlevel: allows IP prefixes learned from different IS-IS levels to be advertised when **allow** is configured.
    - external: allows IP prefixes learned from other protocols to be advertised when **allow** is configured.
- To rectify the fault, run the preceding command on AR-3.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- On Huawei devices, you can run the following command to set the overload bit to prevent temporary route black holes:
  set-overload [**on-startup** [**wait-for-bgp** [*timeout1*]] [**allow** { **interlevel** | **external** } *]
  - wait-for-bgp: sets the overload bit on system startup and the period for keeping the overload bit according to the status of the BGP convergence. If BGP does not send a signal to IS-IS indicating BGP convergence is finished, IS-IS cancels the overload bit after a specified period or default 10 minutes.
  - interlevel: allows IP prefixes learned from different IS-IS levels to be advertised when **allow** is configured.
  - external: allows IP prefixes learned from other protocols to be advertised when **allow** is configured.
- To rectify the fault, run the preceding command on AR-3.

Quiz

1. The router and other routers are connected through physical links. However, in the output of the **display isis peer** command, the peer neighbor is not displayed. How to solve this problem?
2. A Level-1 router cannot generate default routes to other areas. How to analyze the problem?
3. IS-IS cannot learn routes correctly. What are the possible causes?

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Question 1: The router and other routers are connected through physical links. However, in the output of the **display isis peer** command, the peer neighbor is not displayed. How to solve this problem?
- Answer: The possible causes are: If routers at the two ends work at different levels, or have different area IDs, interface authentication types, or passwords, the neighbor relationship may fail to be established. If they have the same system ID, the failure may also occur.
- Question 2: A Level-1 router cannot generate default routes to other areas. How to analyze the problem?
- Answer: The Level-1 router can generate the route to other areas only after it sets up the Level-1 neighbor relationship with the Level-1-2 router in the local area. If the Level-1-2 router at the area border has Level-2 neighbors in different areas, the Attachment (ATT) flag bit is set in the generated LSP. This indicates that the Level-1-2 router is connected to other areas and has routes to other areas. After receiving the LSP, all Level-1 routers in the same area generate a default route pointing to 0.0.0.0 0.
- Question 3: IS-IS cannot learn routes correctly. What are the possible causes?
- Answer: The possible causes are as follows:
  - The neighbor relationship cannot be set up normally.
  - The cost types of the two ends are different.
  - There is no next hop because the IPv4 and IPv6 topologies are different.
  - The route is filtered out by the routing policy and cannot be added to the unicast routing table (URT).
  - The LSP-ID is used up, causing the loss of the Neighbor TLV. If the number of imported routes is too large and the number of used LSP fragments reaches

255, LSP fragment extension must be configured.

- The area or domain configured on the router fails to pass the authentication. As a result, the LSDB is not synchronized.

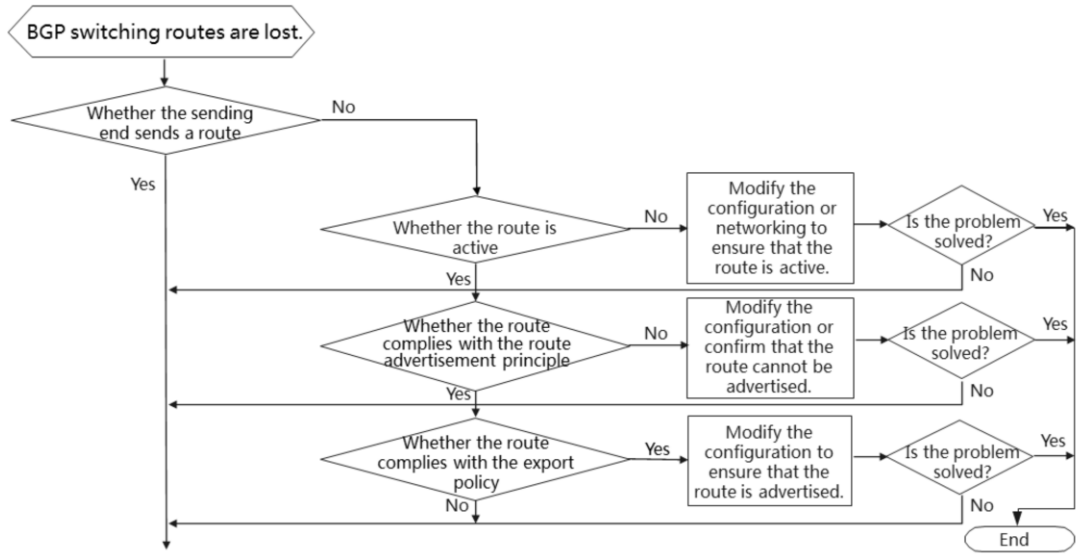


Contents

1. OSPF Troubleshooting
2. IS-IS Troubleshooting
3. **BGP Troubleshooting**
   - Typical BGP Troubleshooting Roadmap
   - BGP Troubleshooting Cases
4. MPLS BGP VPN Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Diagnosis Process 1



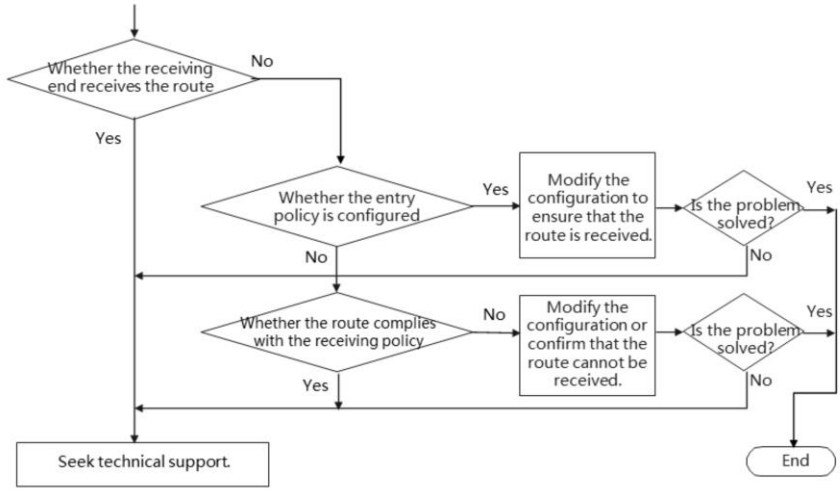
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- BGP is a dynamic routing protocol used between ASs. BGP exchanges information about reachable routes between ASs. BGP has the following advantages over IGP:
  - As an exterior gateway protocol, BGP selects the optimal routes between ASs and advertises these routes.
  - BGP uses various routing policies to filter and select routes flexibly.
  - BGP provides a mechanism for preventing route flapping, which improves Internet network stability.
  - BGP can be more easily extended to adapt to the development of networks.

- BGP faults can be classified into BGP neighbor faults and BGP route learning faults. The flowchart shows the process of BGP route learning.
  - BGP is a protocol that controls route transfer. Its detection of route leaning faults can be divided into two parts: detecting route sending problems and detecting route receiving problems.
  - If the sending end works normally, check the receiving end.

Fault Diagnosis Process 2

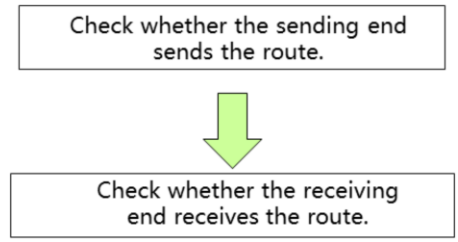


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- If both the sending and receiving ends work normally but BGP routes still cannot be learned, contact Huawei for technical support.

Troubleshooting Procedure



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Step 1: Check whether the sending end sends the route.
  - Run the **display bgp routing-table peer***peer-address***advertised-routes** command on the sending end to check whether a route is sent.
  - If the sender does not send any route, perform the following operations:
    - Check whether the local route is in active state. Run the **display bgp routing-table** command to check whether the route is in active state. That is, check whether tag *> is labeled on the route. If the local route is in inactive state, the next hop is unreachable or other preferred routes exist on the local.
    - Check whether the principle of advertising routes fails to be met. The convergence suppressed routes will not be released externally. Run the **display bgp routing-table** command and it is found that the routes have the s tag. The Dampening suppressed routes will not be released externally either. Run the **display bgp routing-table** command and it is found that the routes have the d flag. The routes learned from the IBGP peer will not be forwarded to the IBGP peer.
    - Check whether an export policy is configured to filter out the routes to be advertised. BGP can use the following filters: IP-Prefix List (prefix list filter), AS_Path Filter (path list filter), and Community Filter and Route-Policy (Community attributes list filters). These filters are applicable to routing information received from IBGP peers or

advertised to IBGP peers.

- Run the **display current-configuration configuration bgp** command to view the configuration information.

- Step 2: Check whether the receiving end receives the route.
  - Run the **display bgp routing-table peer** *peer-address* **received-routes** command on the receiving end to check whether a route is received.
  - If the receiving end does not receive any route, perform the following operations:
    - Check whether an import policy is configured to filter out the routes to be received. Run the **display current-configuration configuration bgp** command to view the configuration information.
    - Check whether the principle of receiving routes fails to be met. The following routes will be rejected: 1. The **peer allow-as-loop** command is not configured, and the local AS number is carried in the AS_Path attributes of the received route. 2. The **peer allow-as-loop [** *number* **]** command is configured. The repeated times of AS in the AS_Path attributes are greater than the configured number value (default value: 1). 3. The first AS number in the AS_Path attributes received from the EBGP peer is not the AS number of the peer end. 4. The Originator_ID and the local Router-ID are the same, or it is the invalid value 0.0.0.0. 5. The Cluster-List in the route received by the reflector contains the local Cluster-ID. 6. The aggregator is the invalid value 0.0.0.0. 7. The Next_Hop is the local interface address. 8. The Next_Hop of the route received from the direct EBGP peer is not reachable. 9. If the **peer route-limit alert-**

only command is used, all the received routes are denied after the threshold is reached.

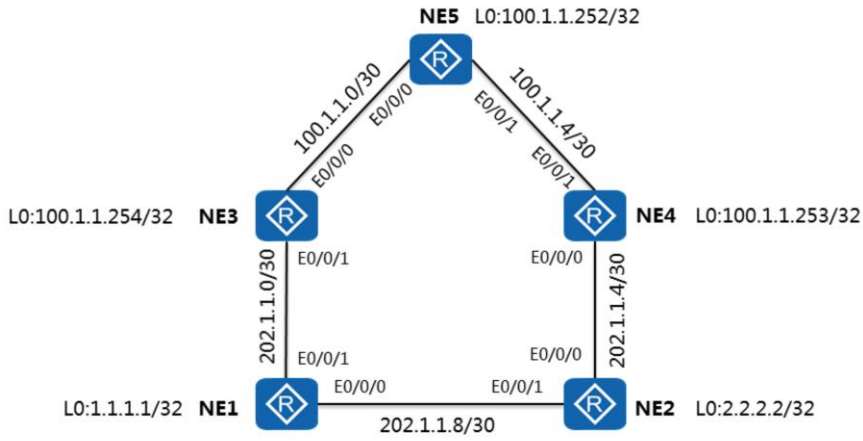
- If the fault persists after the preceding operations are complete, contact Huawei technical support personnel.

Contents

1. OSPF Troubleshooting
2. IS-IS Troubleshooting
3. **BGP Troubleshooting**
   - Typical BGP Troubleshooting Roadmap
   - BGP Troubleshooting Cases
4. MPLS BGP VPN Troubleshooting

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Topology



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The figure shows a network topology of edge routes on an IP MAN and a backbone network. NE1 and NE2 are edge routers in AS 200 on the IP MAN, and NE3, NE4, and NE5 in AS 100 are edge routers on the provincial backbone network. NE1 and NE2 use the **network** command to advertise routes to their EBGP peers NE3 and NE4. NE3 and NE4 establish IBGP peer relationships with NE5. NE5 functions as an RR, and NE3 and NE4 are its clients. Configure the virtual next hop address 202.105.0.5 on NE3 and NE4 so that NE3 and NE4 change the next hop of BGP routes to 202.105.0.5 before they advertise routes to their IBGP peer NE5.

Fault Description

- If the connection between NE1 and NE3 is interrupted, a loop occurs between NE3 and NE5.

```
[NE3]tracert 202.1.1.11
traceroute to  202.1.1.11(202.1.1.11), max hops: 30 ,packet length: 40
1 100.1.1.2 40 ms  30 ms  50 ms
2 100.1.1.5 60 ms 100.1.1.1 50 ms  50 ms
3 100.1.1.2 60 ms  60 ms  80 ms
4 * * *
5  *  * 100.1.1.2 90 ms
6 100.1.1.1 80 ms  90 ms  100 ms
7 100.1.1.2 110 ms  110 ms  *
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- When the connection between NE1 and NE3 is interrupted, a loop occurs when NE3 accesses an IP address in the network segment 202.1.1.0/24（exclude the network segment used by routers in the topology）. Assuming that NE3 accesses 202.1.1.11, the figure displays the consequence.

Fault Analysis 1

- Check the BGP configuration on the device to learn about the route transfer process.
- Check the configuration of NE1. It is found that NE1 uses the black hole route and network mode to advertise the route to its EBGP neighbor NE3. The configurations on NE2 are similar to those on NE1.

```
[NE1]display  current-configuration  configuration bgp
bgp 200
 ......
 ipv4-family unicast
  undo synchronization
  network 202.1.1.0
  peer 2.2.2.2 enable
  peer 2.2.2.2 next-hop-local
  peer 202.1.1.1 enable

[NE1]display current-configuration  | include  route-static
 ip route-static 202.1.1.0 255.255.255.0 NULL0
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- 202.1.1.0/24 is a simulated user address pool on a network segment (exclude the network segment used by routers in the topology) .

Fault Analysis 2

- Check the BGP configuration of NE3.

```
[NE3]display current-configuration  configuration bgp
#
bgp 100
 peer 100.1.1.252 as-number 100
 peer 100.1.1.252 connect-interface LoopBack0
 peer 202.1.1.2 as-number 200
 #
 ipv4-family unicast
  undo synchronization
  peer 100.1.1.252 enable
  peer 100.1.1.252 route-policy vir-add-bgp export
  peer 202.1.1.2 enable
#
Return
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Analysis 3

- On NE3, a policy is configured to forcibly change the BGP next hop to the virtual address 202.105.0.5.

```
[NE3] display  route-policy  vir-add-bgp
Route-policy : vir-add-bgp
  permit : 10
    Apply clauses :
      apply ip-address next-hop 202.105.0.5
[NE3]display  this
#
 sysname NE3
isis 1
Import-route static
preference 100
......
 ip route-static 202.105.0.5 255.255.255.255 100.1.1.2
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The next hop after the virtual address is the interconnection IP address 100.1.1.2 of NE5 and NE3.

Fault Analysis 4

- Check the routing table on NE3.

```
[NE3]display ip routing-table 202.1.1.0 verbose
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1

Destination: 202.1.1.0/24
    Protocol: BGP              Process ID: 0
  Preference: 255                  Cost: 0
     NextHop: 202.105.0.5      Neighbour: 100.1.1.252
       State: Active Adv Relied      Age: 00h00m00s
         Tag: 0                  Priority: low
       Label: NULL               QoSInfo: 0x0
  IndirectID: 0x5
RelayNextHop: 100.1.1.2          Interface: Ethernet0/0/1
    TunnelID: 0x0                Flags: RD
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Because the connection between NE1 and NE3 is interrupted, the routes on NE3 are advertised by NE4, reflected by the RR, and have an outbound interface on NE5.

Fault Analysis 5

- Check the routing table on NE5. NE5 has a route iterated to NE3. Therefore, a routing loop occurs.

```
[NE5]display  ip routing-table 202.1.1.0
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1
Destination/Mask    Proto  Pre  Cost      Flags NextHop         Interface

        202.1.1.0/24  BGP    255  0           RD  202.105.0.5     Ethernet0/0/0
                      BGP    255  0           RD  202.105.0.5     Ethernet0/0/1
[NE5]display  ip routing-table  202.105.0.5
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Table : Public
Summary Count : 1
Destination/Mask  Proto  Pre  Cost      Flags NextHop         Interface

     202.105.0.5/32  ISIS    15   74          D  100.1.1.5       Ethernet0/0/0
                     ISIS    15   74          D  100.1.1.1       Ethernet0/0/1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Check the routes to NE3 on NE5. The command output shows that the next hop of these routes is the virtual IP address 202.105.0.5. Check the routes from 202.105.0.5 on NE5. The command output shows that there are two equal-cost routes destined for NE3 and NE4 respectively.
- NE3 has routes to NE5, and NE5 has a route iterated to NE3. Therefore, a routing loop occurs.

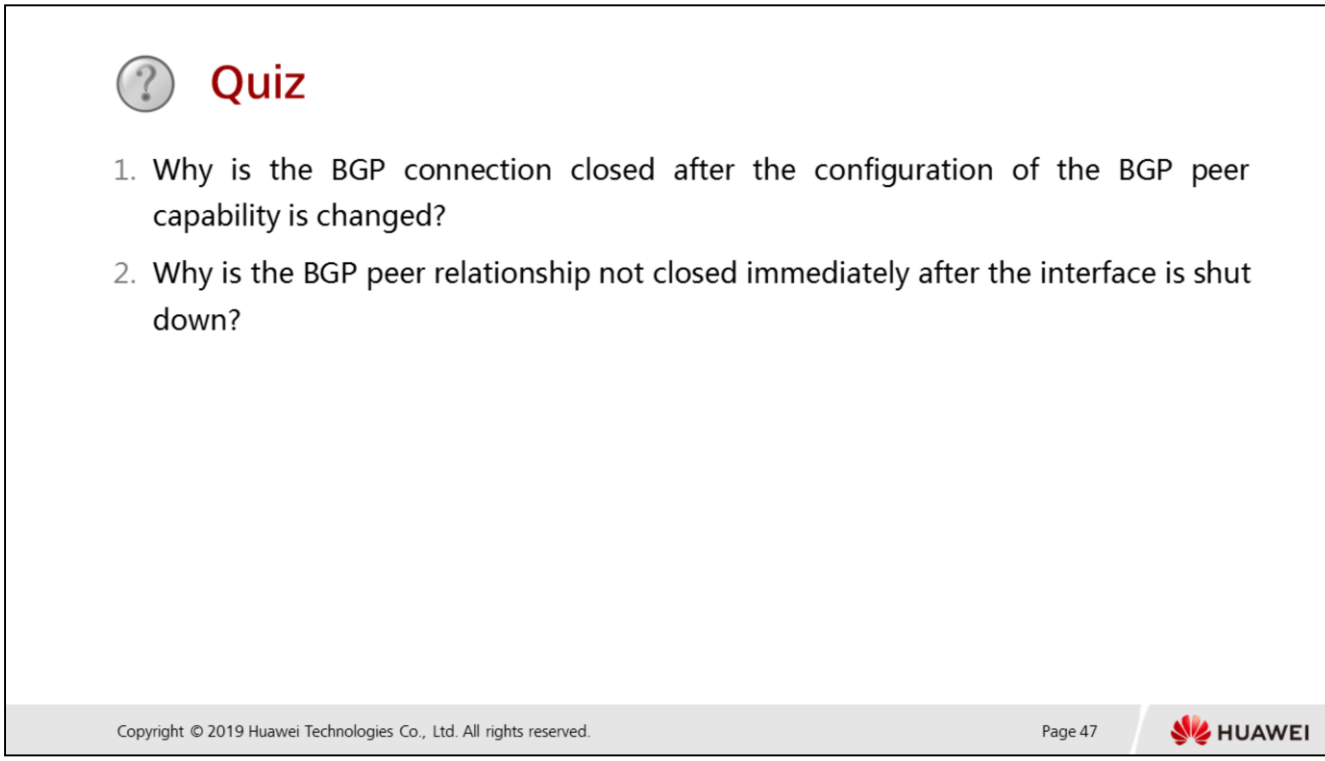
Troubleshooting

1. According to the routing information on the devices, the root cause of the loop is that the static route directing to the virtual address 202.105.0.5 is advertised by NE3 through IS-IS to NE5.
2. When a static route is configured on NE3, the next hop address is 100.1.1.2. Therefore, when Ethernet0/0/0 of the link is DOWN, the static route is still imported to IS-IS.
3. Change the next hop address of the static route directing to the virtual address on NE3 to 202.1.1.2.

```
[NE3]ip route-static 202.105.0.5  255.255.255.255  202.1.1.2
```



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- Question 1: Why is the BGP connection closed after the configuration of the BGP peer capability is changed?
- A: The BGP connection is closed automatically when the configuration of the BGP capability is changed. This is because BGP does not support dynamic capability negotiation. The neighbor capability is then negotiated again. The BGP connection is closed automatically when:
  - Label-Route-Capability is enabled or disabled.
  - The BGP peer in the address family is enabled or disabled. For example, if the **peer enable/undo peer enable** command is used in the VPNv4 address family, the BGP connection of the peer in other address family is closed automatically.
  - GR capability is enabled.
- Question 2: Why is the BGP peer relationship not closed immediately after the interface is shut down?
- Answer: The EBGP peer relationship is disconnected immediately after the interface is shut down only when EBGP peers are directly connected and the **ebgp-interface-sensitive** command is used in the BGP view. By default, the command is used. Otherwise, the BGP peer relationship is not torn down until the hold time expires.



Contents

1. OSPF Troubleshooting
2. IS-IS Troubleshooting
3. BGP Troubleshooting
4. **MPLS BGP VPN Troubleshooting**
   - MPLS BGP VPN Troubleshooting Roadmap
   - MPLS BGP VPN Troubleshooting Cases

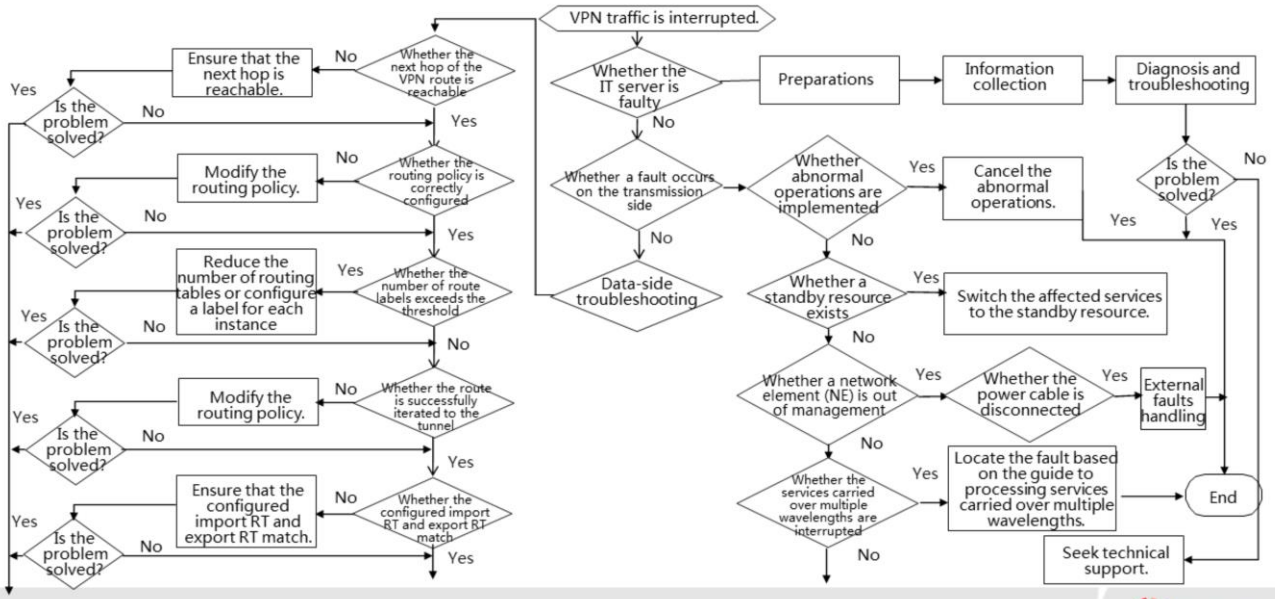
Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Common Fault Causes

- Routes are inactive because their next hops are unreachable.
- Routes fail to be advertised or received because routing policies are configured incorrectly.
- VPN routes fail to be advertised because the number of labels exceeds the upper limit.
- VPN routes are inactive because they fail to be iterated to tunnels.
- Routes fail to be added to the VPN routing table because the configured import RT and export RT do not match.
- The received routes are dropped because there is an upper limit on the number of routes on the device.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

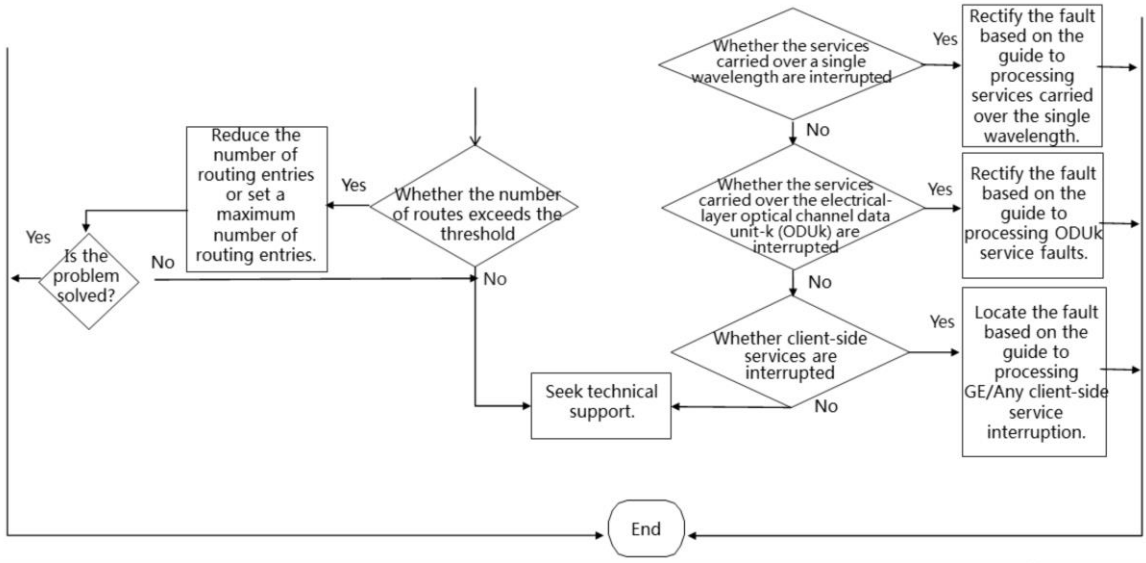
Fault Diagnosis Process



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Fault Diagnosis Process



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

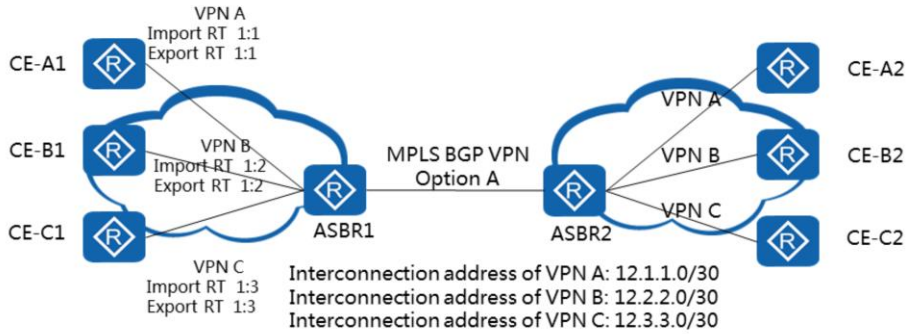


Contents

1. OSPF Troubleshooting
2. IS-IS Troubleshooting
3. BGP Troubleshooting
4. **MPLS BGP VPN Troubleshooting**
   - MPLS BGP VPN Troubleshooting Roadmap
   - MPLS BGP VPN Troubleshooting Cases

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Topology

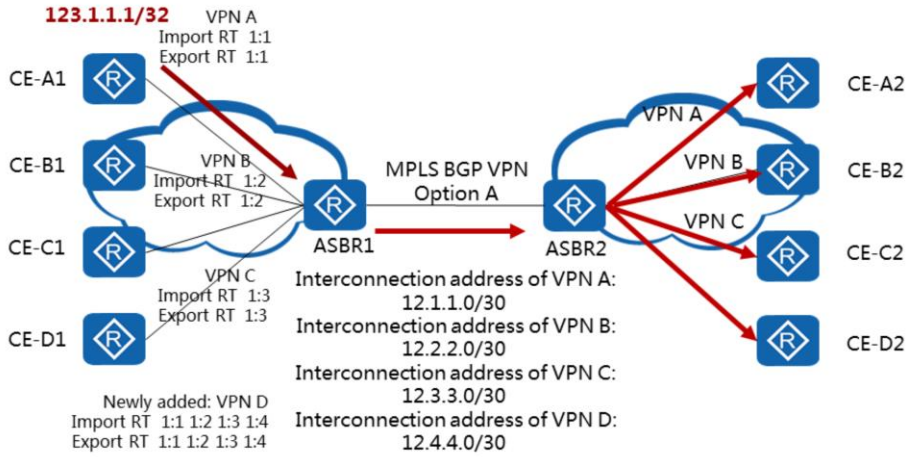


Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A company's network has three L3 VPN instances: VPN A, VPN B, and VPN C. The route distinguishers of the instances are 1:1, 1:2, and 1:3 respectively, and the VPN targets are 1:1, 1:2, and 1:3 respectively. The three VPNs are therefore isolated from each other and cannot communicate with each other.
  As shown in the figure, CE-A1, CE-B1, and CE-C1 are connected to VPN A, VPN B, and VPN C of ASBR1 respectively. CE-A2, CE-B2, and CE-C2 are connected to VPN A, VPN B, and VPN C of ASBR2 respectively. Inter-AS MPLS BGP VPN Option A is configured between ASBR1 and ASBR2. In this case, only CE-A2 can receive routes advertised by CE-A1, achieving the isolation between the VPN instances.

Fault Description



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Due to service expansion, the company configures a VPN D to the network. It is required that while VPN A, VPN B, and VPN C should remain isolated from each other, VPN D should be able to communicate with each of them. Therefore, the route distinguisher of VPN D is set to 1:4, and the VPN target is set to 1:1 1:2 1:3 1:4. Inter-AS MPLS BGP VPN Option A is configured between ASBR1 and ASBR2.
  However, in this case, CE-B2 and CE-C2 can also learn routes from CE-A1. In fact, after Inter-AS MPLS BGP VPN Option A is configured, every VPN can learn routes from other VPNs. The previously designed isolation becomes invalid.

Fault Analysis (1/6)

- Select a faulty route (123.1.1.1/32) and check whether the local crossing of the route is normal on ASBR1.

```
<ASBR1>display bgp vpnv4 all routing-table 123.1.1.1 32
 BGP local router ID : 1.1.1.1
Local AS number : 100
Total routes of Route Distinguisher(1:1): 1
BGP routing table entry information of 123.1.1.1/32:
From: 10.1.1.1 (10.1.1.1)
Route Duration: 00h02m09s
Direct Out-interface: Ethernet0/0/0
Original nexthop: 10.1.1.1
Qos information : 0x0
Ext-Community:RT <1 : 1>
AS-path 1000, origin igp, MED 0, pref-val 0, valid, external, best, select, pre
255
Not advertised to any peer yet
VPN-Instance vpna, Router ID 1.1.1.1:
Total Number of Routes: 1
BGP routing table entry information of 123.1.1.1/32:
From: 10.1.1.1 (10.1.1.1)
.....
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Fault Analysis (2/6)

```
Route Duration: 00h15m07s
 Direct Out-interface: Ethernet0/0/0
 Original nexthop: 10.1.1.1
 Qos information : 0x0
 AS-path 1000, origin igp, MED 0, pref-val 0, valid, external, best, select, act
ive, pre 255
 Advertised to such 2 peers:
   10.1.1.1
   12.1.1.2       // The route is advertised to ASBR2 through the Option A peer (12.1.1.2) of VPN A.
 VPN-Instance vpnd, Router ID 1.1.1.1:
 Total Number of Routes: 1
 BGP routing table entry information of 123.1.1.1/32:
 From: 10.1.1.1 (10.1.1.1)
 Route Duration: 00h15m07s
 Direct Out-interface: Ethernet0/0/0
 Original nexthop: 10.1.1.1
 Qos information : 0x0
 Ext-Community:RT <1 : 1>
 AS-path 1000, origin igp, MED 0, pref-val 0, valid, external, best, select, act
ive, pre 255
 Advertised to such 2 peers:
   40.1.1.1
   12.4.4.2       // The route is advertised to ASBR2 through the Option A peer (12.4.4.2) of VPN D.
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- The export RT (outbound VPN target) of VPN A is 1:1, and the import RT (inbound VPN target) of VPN D contains 1:1. Therefore, the route can be locally crossed to VPN D. For ASBR1, its Option A peer ASBR2 is equivalent to a customer edge (CE) device, so the route locally crossed to VPN D can be advertised to ASBR2 through the Option A peer (12.4.4.2) of VPN D.

Fault Analysis (3/6)

- Check the transfer process of route 123.1.1.1/32 on ASBR2.

```
<ASBR2>display bgp vpnv4 all routing-table 123.1.1.1 32
 BGP local router ID : 2.2.2.2
Local AS number : 200
Total routes of Route Distinguisher(1:1): 1
BGP routing table entry information of 123.1.1.1/32:
From: 12.1.1.1 (1.1.1.1)
Route Duration: 00h19m33s
Direct Out-interface: GigabitEthernet0/0/1
Original nexthop: 12.1.1.1
Qos information : 0x0
Ext-Community:RT <1 : 1>
AS-path 100 1000, origin igp, pref-val 0, valid, external, best, select, pre 255
Not advertised to any peer yet
Total routes of Route Distinguisher(1:4): 1
BGP routing table entry information of 123.1.1.1/32:
From: 12.4.4.1 (1.1.1.1)
Route Duration: 00h12m23s
Direct Out-interface: GigabitEthernet0/0/3
Original nexthop: 12.4.4.1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Fault Analysis (4/6)

```
Qos information : 0x0
 Ext-Community:RT <1 : 1>, RT <1 : 2>,
           RT <1 : 3>, RT <1 : 4>
 AS-path 100 1000, origin igp, pref-val 0, valid, external, best, select, pre 255
 Not advertised to any peer yet
 VPN-Instance vpna, Router ID 2.2.2.2:
 Total Number of Routes: 2
 BGP routing table entry information of 123.1.1.1/32:
 From: 12.1.1.1 (1.1.1.1)
 Route Duration: 00h19m33s
 Direct Out-interface: GigabitEthernet0/0/1
 Original nexthop: 12.1.1.1
 Qos information : 0x0
 AS-path 100 1000, origin igp, pref-val 0, valid, external, best, select, active
, pre 255
 Advertised to such 2 peers:
   10.1.2.1              // The route is advertised to CE-A2.
   12.1.1.1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- ASBR2 learns the VPN A route 123.1.1.1/32 through the Option A peer (12.1.1.1) of VPN A and advertises the route to CE-A2.

Fault Analysis (5/6)

```
BGP routing table entry information of
123.1.1.1/32:
 From: 12.4.4.1 (1.1.1.1)
 Route Duration: 00h12m24s
 Direct Out-interface: GigabitEthernet0/0/3
 Original nexthop: 12.4.4.1
 Qos information : 0x0
 Ext-Community:RT <1 : 1>, RT <1 : 2>,
               RT <1 : 3>, RT <1 : 4>
// The route of VPN D is locally crossed
to VPN A.
 AS-path 100 1000, origin igp, pref-val 0,
valid, external, pre 255, not preferred for
peer type
 Not advertised to any peer yet
```

```
 VPN-Instance vpnd, Router ID 2.2.2.2:

 Total Number of Routes: 2
 BGP routing table entry information of
123.1.1.1/32:
 From: 12.4.4.1 (1.1.1.1)
 Route Duration: 00h11m08s
 Direct Out-interface:
GigabitEthernet0/0/3
 Original nexthop: 12.4.4.1
 Qos information : 0x0
 AS-path 100 1000, origin igp, pref-val
0, valid, external, best, select, active
, pre 255
 Advertised to such 1 peers:
    40.1.2.1 // The route is advertised
to CE-D2.
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- ASBR2 learns the VPN D route 123.1.1.1/32 through the Option A peer (12.4.4.1) of VPN D. The route is locally crossed to VPN A (not preferred), CE-B2, CE-C2, and CE-D2.

Fault Analysis (6/6)

```
VPN-Instance vpnb, Router ID 2.2.2.2:

Total Number of Routes: 1
BGP routing table entry information of
123.1.1.1/32:
From: 12.4.4.1 (1.1.1.1)
Route Duration: 00h11m08s
Direct Out-interface: GigabitEthernet0/0/3
Original nexthop: 12.4.4.1
Qos information : 0x0
Ext-Community:RT <1 : 1>, RT <1 : 2>,
          RT <1 : 3>, RT <1 : 4>  // The
route of VPN D is locally crossed to VPN B.
AS-path 100 1000, origin igp, pref-val 0,
valid, external, best, select, active
, pre 255
Advertised to such 2 peers:
   20.1.2.1          // The route is advertised
to CE-B2.
   12.2.2.1
```

```
VPN-Instance vpnc, Router ID 2.2.2.2:

Total Number of Routes: 1
BGP routing table entry information of
123.1.1.1/32:
From: 12.4.4.1 (1.1.1.1)
Route Duration: 00h11m08s
Direct Out-interface: GigabitEthernet0/0/3
Original nexthop: 12.4.4.1
Qos information : 0x0
Ext-Community:RT <1 : 1>, RT <1 : 2>,
          RT <1 : 3>, RT <1 : 4>
// The route is locally crossed to VPN C.
AS-path 100 1000, origin igp, pref-val 0, valid,
external, best, select, active
, pre 255
Advertised to such 2 peers:
   30.1.2.1 // The route is advertised to CE-C2.
   12.3.3.1
```

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- ASBR2 learns the VPN D route 123.1.1.1/32 through the Option A peer (12.4.4.1) of VPN D. The route is locally crossed to VPN A (not preferred), CE-B2, CE-C2, and CE-D2.

Troubleshooting

- The export RT of VPN D is 1:1 1:2 1:3 1:4. The import RT of VPN A is 1:1, the import RT of VPN B is 1:2, and the import RT of VPN C is 1:3. Therefore, the local crossing of the faulty route 123.1.1.1/32 is normal. The route tables locally crossed to VPN B and VPN C are successfully advertised to CE-B2 and CE-C2. The advertisement is normal.
- Route crossing and route table advertisement on ASBR1 and ASBR2 are both normal. The problem is that VPN D can communicate with other three VPNs. As a result, routes learned by ASBR2 from VPN D are crossed to the other three VPNs. Under this mechanism, routes from any one of the four VPNs can be crossed to any other VPN. That is to say, every VPN can learn routes from other VPNs, and the previously designed isolation becomes invalid.



Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

Solution 1

- Solution 1: On ASBR1, configure an export policy for the Option A peer of VPN D on ASBR2.

```
ip extcommunity-filter 1 permit rt 1:4
ip extcommunity-filter 2 permit rt 1:1 rt 1:2 rt 1:3
#
route-policy rp-asbr2-vpnd permit node 10
 if-match extcommunity-filter 1
#
route-policy rp-asbr2-vpnd deny node 20
 if-match extcommunity-filter 2
#
route-policy rp-asbr2-vpnd permit node 100
#
bgp 100
  peer 12.4.4.2 route-policy rp-asbr2-vpnd export
#
```

- Perform the same configuration on ASBR2.
- Note: In local crossing scenarios where there are no provider edge (PE) devices, the configurations of **route-policy rp-asbr2-vpnd** and **extcommunity-filter 1** on **node 10** can be omitted.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.

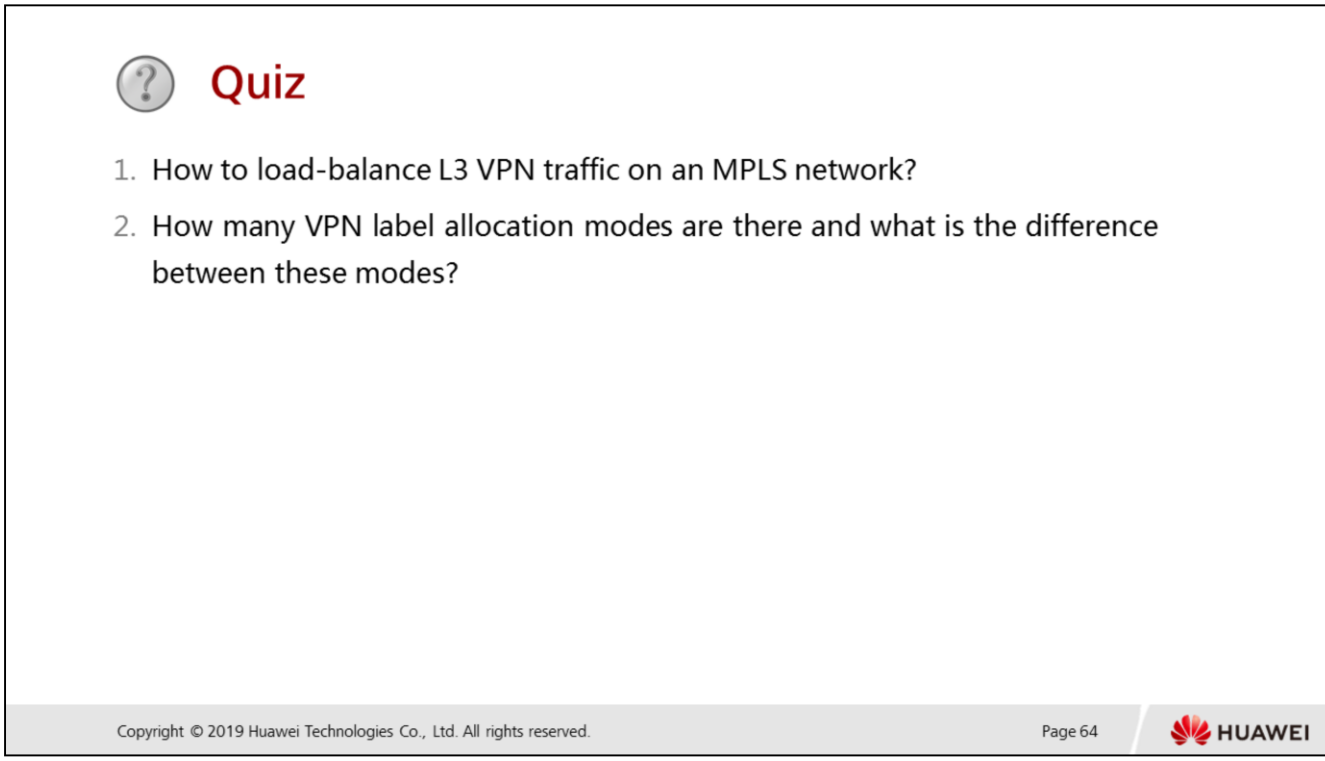
HUAWEI

- On ASBR1, configure an export policy for the Option A peer of VPN D. Only the routes originated from VPN D (including VPNv4 routes that are crossed to VPN D through the import RT 1:4, and routes that are received from other private network neighbors of VPN D) are advertised. Routes originated from other VPNs (including VPNv4 routes that are crossed to VPN D through the import RT 1:1, 1:2, or 1:3, and routes that are locally crossed to VPN D from other VPN instances) are not advertised. In this way, ASBR2 does not receive the route from CE-A1 through the Option A peer of VPN D, and therefore, it cannot cross the route to other VPN instances.
- The routes originated from VPN D on ASBR1 include VPNv4 routes that are crossed to VPN D through the import RT 1:4 (carrying extcommunity <1:4>) and routes received from other private network peers of VPN D.

Solution 2

- Solution 2: Configure Inter-AS MPLS BGP VPN Option B, in which ASBRs advertise routes through VPNv4 peers. In this way, when ASBR1 advertises routes received from CE-A1 to ASBR2, it advertises only the VPNv4 routes whose extcommunity is 1:1. ASBR2 receives only the VPNv4 routes with the extcommunity 1:1. Based on the import RT policies, these routes can be crossed only to VPN A and VPN D.

Copyright © 2019 Huawei Technologies Co., Ltd. All rights reserved.



- Question 1: How to load-balance L3 VPN traffic on an MPLS network?
- Answer: By default, on an MPLS network, L3 VPN traffic is not load-balanced. To implement load balancing, run the following commands:
  - **tunnel select-seq** { **cr-lsp** | **lsp** } * **load-balance-number** *load-balance-number*
- Question 2: How many VPN label allocation modes are there and what is the difference between these modes?
- Answer: The VPN labels are allocated in either of the two modes:
  - Apply-label per-route (default mode)
  - Apply-label per-instance
- Differences:
  - Route-based label allocation consumes a lot of product resources. Allocating labels to a lot of routes may overload a product, causing a forwarding failure.
  - Instance-based label allocation efficiently uses product resources.
- Generally, the two modes have the same effect, but instance-based label allocation is recommended.

Thank You

www.huawei.com

Recommendations

- Huawei Learning Website
  - http://learning.huawei.com/en
- Huawei e-Learning
  - https://ilearningx.huawei.com/portal/#/portal/ebg/51
- Huawei Certification
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_31&lang=en
- Find Training
  - http://support.huawei.com/learning/NavigationAction!createNavi?navId=_trainingsearch&lang=en

More Information

- Huawei learning APP

版权所有© 2018 华为技术有限公司