Huawei Certification

HCDP-IESN

Implementing Enterprise Switching Networks

Huawei Technologies Co.,Ltd

Huawei Certification

HCDP-IESN Implementing Enterprise Switching

Lab Guide

Edition 1.6

Huawei Certification System

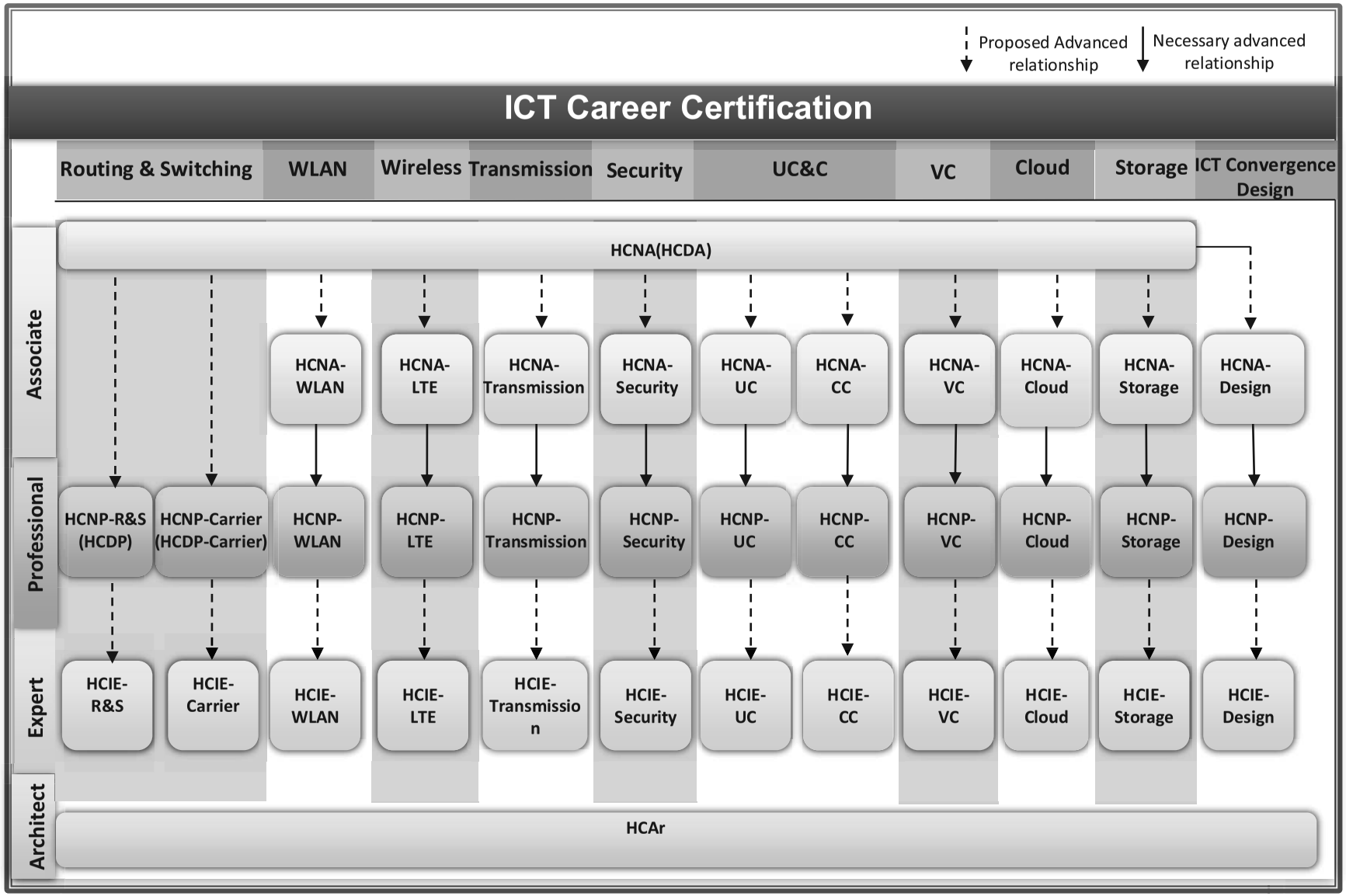
Relaying on its strong technical and professional training system, according to different customers at different levels of ICT technology, Huawei certification is committed to provide customs with authentic, professional certification.

Based on characteristics of ICT technologies and customers' needs at different levels, Huawei certification provides customers with certification system of four levels.

HCDA(Huawei Certification Datacom Assocaite) is primary for IP network maintenance engineers, and any others who want to learn the IP network knowledge.HCDA certification covers the TCP/IP basics, routing, switching and other common foundational knowledge of IP networks, together with Huawei communications products, versatile routing platform VRP characteristics and basic maintenance.

HCDP-Enterprise (Huawei Certification Datacom Professional-Enterprise) is aimed at enterprise-class network maintenance engineers, network design engineers, and any others who want to in depth grasp routing, switching, network adjustment and optimization technologies. HCDP-Enterprise is consist of IESN (Implementinging Enterprise Switch Networks), IERN (Implementing Enterprise Routing Networks), and IENP (Improving Enterprise Network performance), which includes advanced IPv4 routing and switching technology principle, IP technology of network security, high availability and QoS, as well as the implementation in Huawei products.

HCIE-Enterprise(Huawei Certified Internetwork Expert-Enterprise) is designed to endue engineers with a variety of IP network technology and proficiency in maintenance, diagnostics and troubleshooting of Huawei products, which equips the engineers with competence in planning, design and optimization of large-scale IP network.


Proposed Advanced relationship
Necessary advanced relationship
ICT Career Certification
Routing & Switching
WLAN
Wireless
Transmission
Security
UC&C
VC
Cloud
Storage
ICT Convergence Design
Associate
Professional
Expert
Architect
HCNA(HCDA)
HCNA-WLAN
HCNA-LTE
HCNA-Transmission
HCNA-Security
HCNA-UC
HCNA-CC
HCNA-VC
HCNA-Cloud
HCNA-Storage
HCNA-Design
HCNP-R&S (HCDP)
HCNP-Carrier (HCDP-Carrier)
HCNP-WLAN
HCNP-LTE
HCNP-Transmission
HCNP-Security
HCNP-UC
HCNP-CC
HCNP-VC
HCNP-Cloud
HCNP-Storage
HCNP-Design
HCIE-R&S
HCIE-Carrier
HCIE-WLAN
HCIE-LTE
HCIE-Transmission
HCIE-Security
HCIE-UC
HCIE-CC
HCIE-VC
HCIE-Cloud
HCIE-Storage
HCIE-Design
HCAr

Foreword

Outline

This course is applicable to the candidates who are preparing for the Huawei Certified Datacom Professional-Implementing Enterprise Switching Network (HCDP-IESN) exam, and the readers who want to understand the switching network technologies and Huawei Versatile Routing Platform (VRP) implementation.

Content

This course consists of six modules related to switching network protocols, MPLS and LDP, and MPLS implementation on Huawei VRP. In addition, this course provides network application of Huawei Ethernet switches.

Module 1 describes how the VLAN and GVRP technologies are implemented, including the knowledge about Layer 2 isolation, Layer 3 routing, and QinQ, and the VLAN configurations on the VRP.

Module 2 describes the mechanism and implementation of the STP protocols, including STP, RSTP, and MSTP.

Module 3 describes the mechanisms and configurations of access technologies, including 802.1x, DHCP.

Module 4 describes the implementation of MPLS and LDP.

Module 5 describes the characteristics of Huawei Ethernet switches

This course helps readers understand the enterprise network switching technologies and how these technologies are implemented on Huawei switches.

Readers' Knowledge Background

To fully understand this course, the readers should:

(1) Have learned the Huawei Certified Datacom Associate (HCDA) course.

(2) Have passed the HCDA exam.

(3) Be familiar with TCP/IP protocol suite and basic knowledge about Ethernet

technologies.

(4) Know the operating mechanisms of Ethernet switches.

Icons Used in This Book

IPv6 Router

SOHO Router

Voice Router

Low-end Router

High-end Router

Core Router

Hub

Convergence Switch

Socket switch

Core Switch

Edge Switch

Cascade Switch

AP

AP Amplifier

Wireless Bridge

Wireless Network Card

Access Server

Audio Gateway

Firewall

Internet Telephony

Table of Contents

Module 1 VLAN ..... Page 1

VLAN Technology Principles and Configuration ..... Page 3

QinQ Principles ..... Page 45

Module 2 STP ..... Page 67

STP Principles and Configuration ..... Page 69

RSTP Principles and Configuration ..... Page 115

MSTP Principles and Configuration ..... Page 151

Module 3 Access Layer Protocols ..... Page 185

802.1x Principles and Configuration ..... Page 187

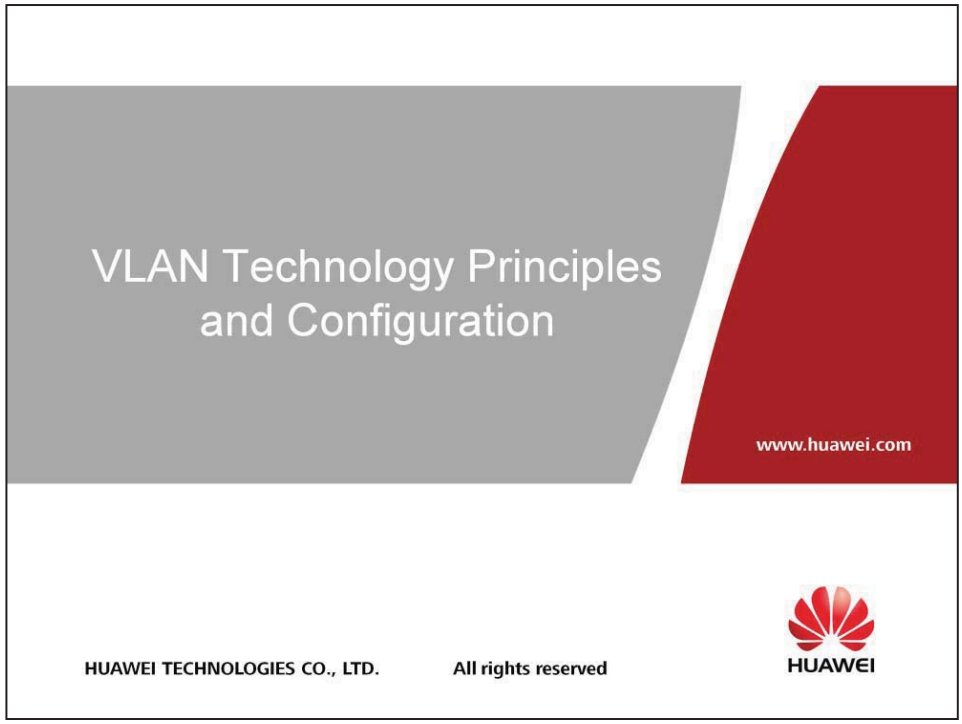
DHCP Principles and Application ..... Page 235

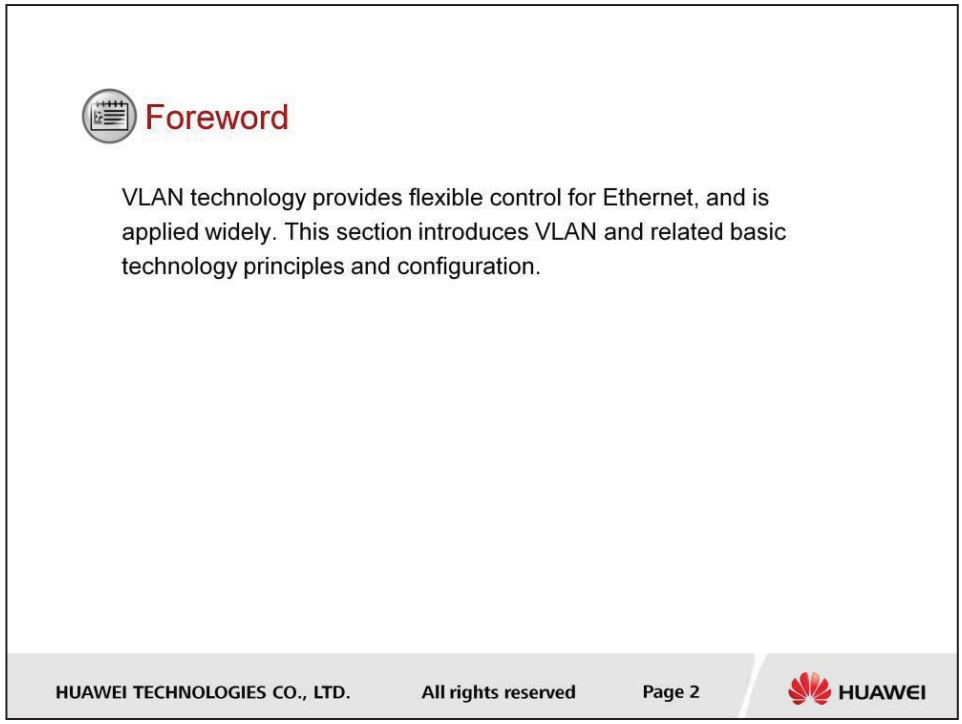
Module 4 MPLS ..... Page 287

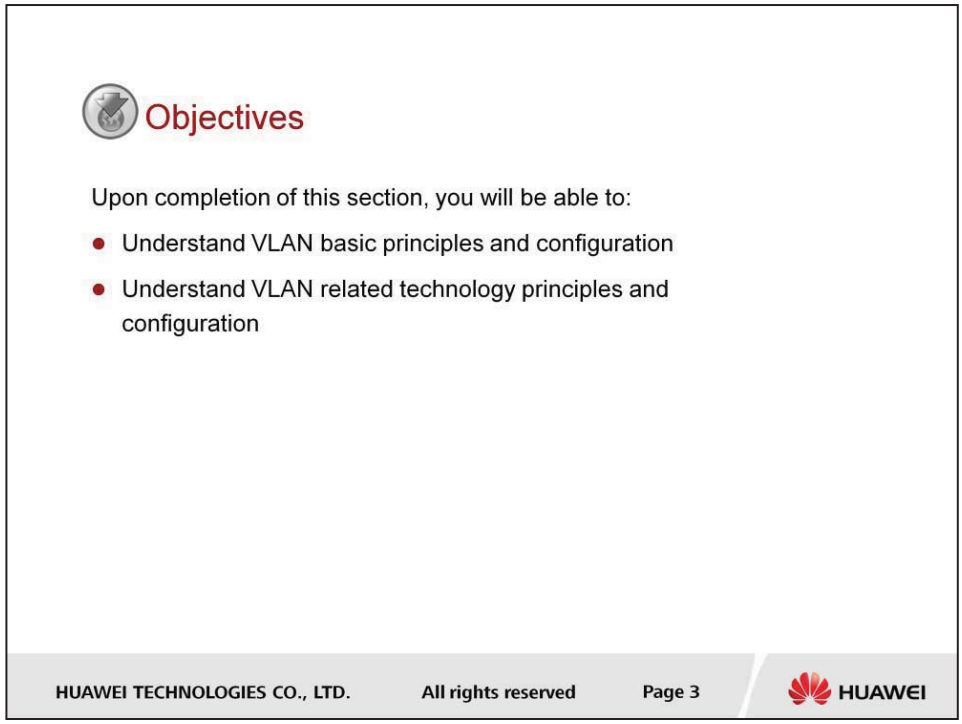
MPLS Principles ..... Page 289

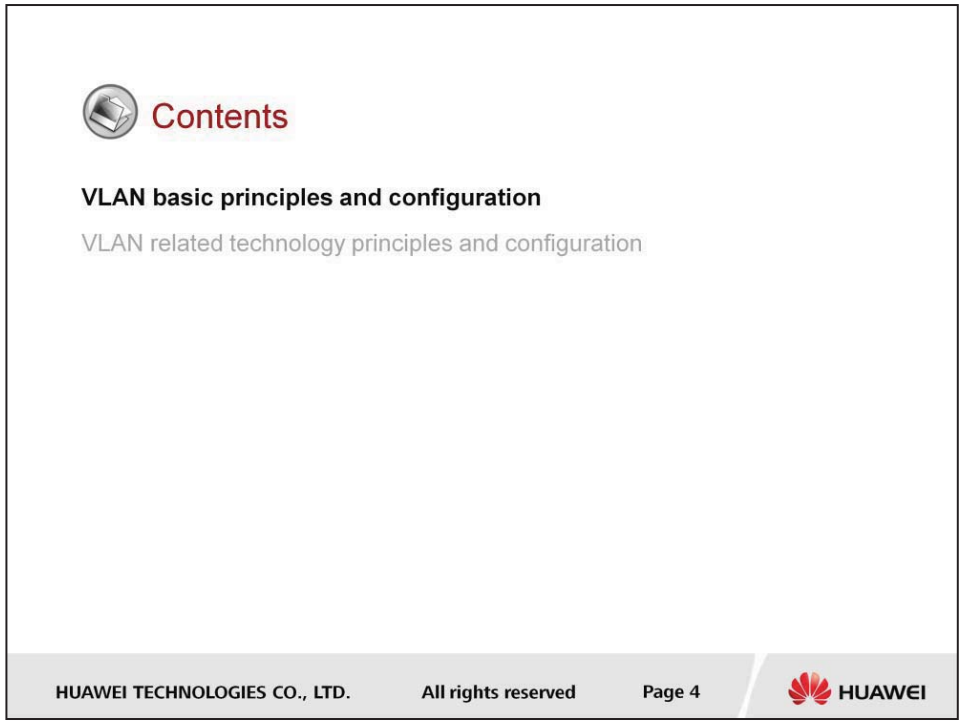
LDP Principles ..... Page 346

Module 1
VLAN


VLAN Technology Principles and Configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

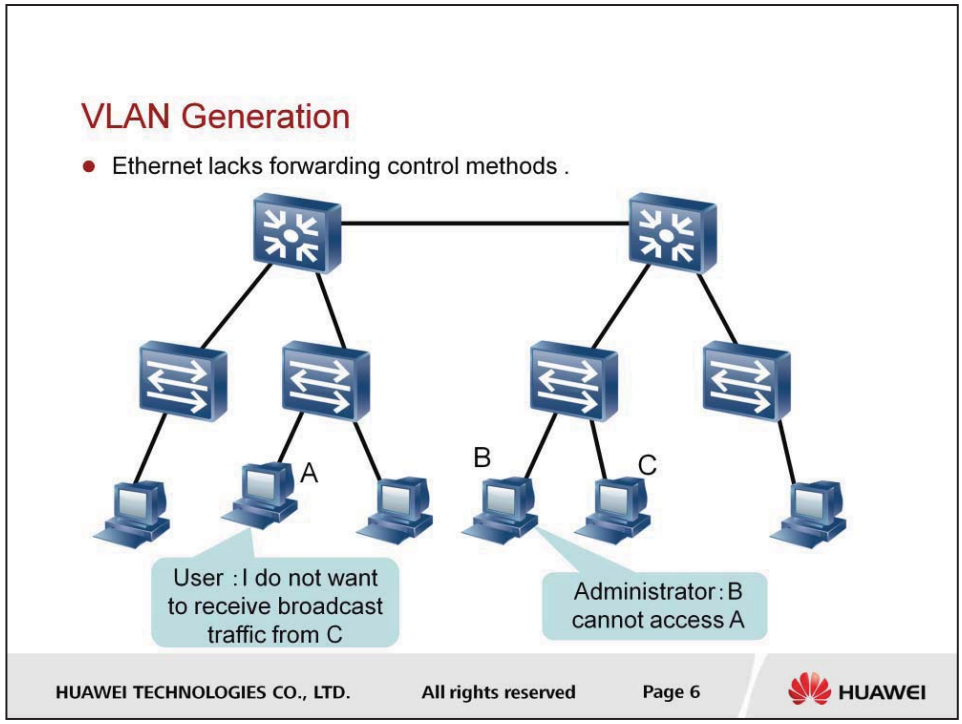

Foreword
VLAN technology provides flexible control for Ethernet, and is applied widely. This section introduces VLAN and related basic technology principles and configuration.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to:
Understand VLAN basic principles and configuration
Understand VLAN related technology principles and configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
VLAN basic principles and configuration
VLAN related technology principles and configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI

Contents

- VLAN basic principles and configuration
  - ⇨ VLAN origin
  - ⇨ VLAN data frame structure
  - ⇨ VLAN partitioning
  - ⇨ VLAN interfaces

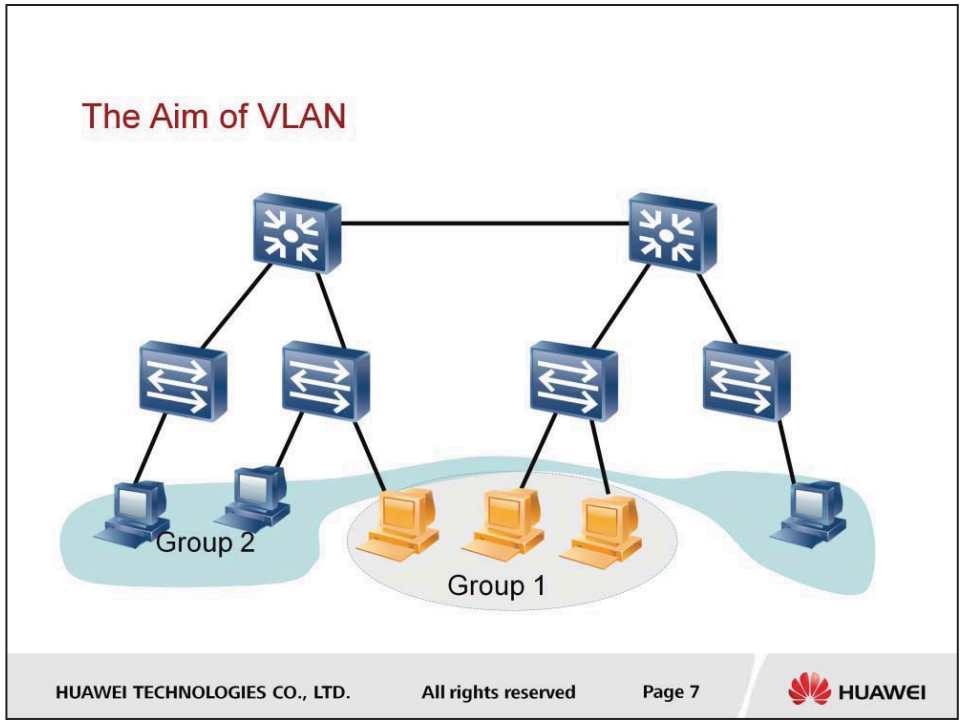


The traditional Ethernet switch adopts source address learning mode when it forwards data. It can automatically learn the MAC address of the host connecting to each port, form forwarding tables, and then forward Ethernet frames according to the table. The whole forwarding process is completed automatically, all the ports can communicate with each other, and maintenance personnel can not control the forwarding between any two ports. For example, they can not implement that host B can not access host A. Following disadvantages are exist in this kind of network:

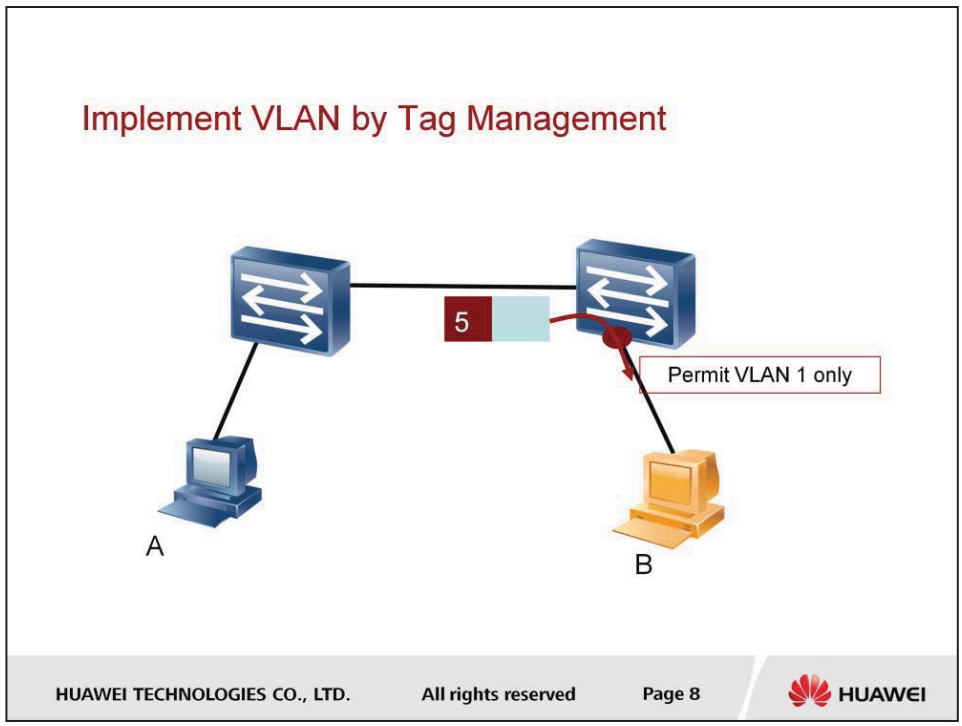
• Network Security is bad. All the ports can communicate with each other, which increases the possibility that users attack the network.

• Network efficiency is low. Users may receive abundant unnecessary packets, which is a waste of bandwidth resource and host CPU resource. For example, unnecessary broadcast packets.

• Service expansion capability is bad. The network can not implement differential service, for example, it can not forward Ethernet frame used for network management with higher priority.



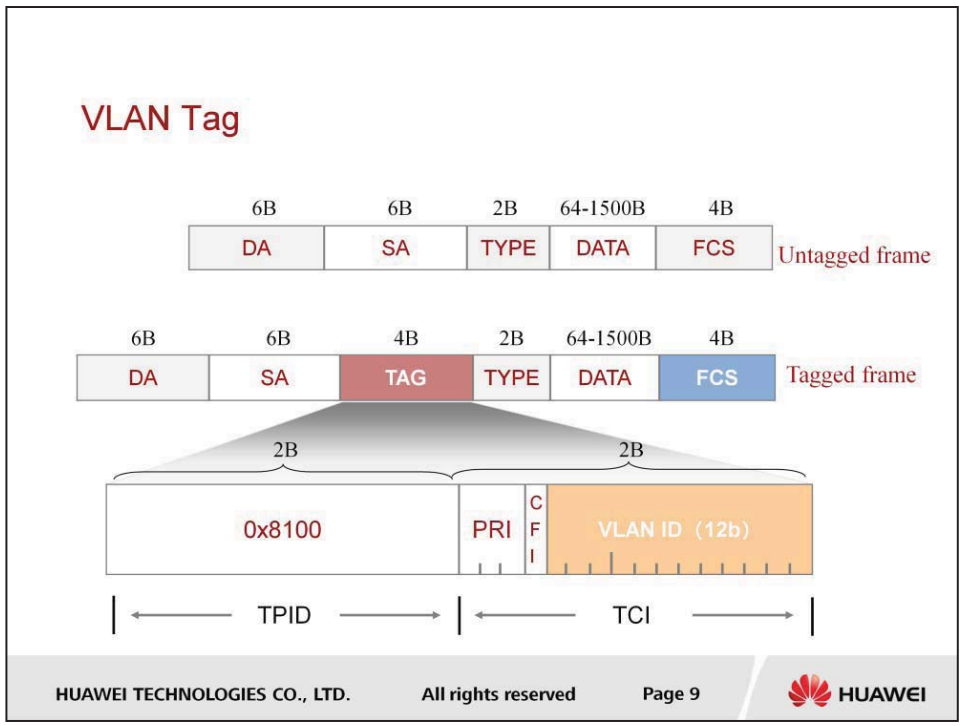
VLAN technology divides users into multiple logical networks (groups). The communication is allowed within a group, but it is prohibited among groups. Layer-2 unicast packet, layer-2 multicast packet and layer-2 broadcast packet can only be forwarded within a group. It is easy to add and delete group members by VLAN technology. VLAN technology provides a management method to control the intercommunication among terminals. In the figure above, PCs in group 1 and group 2 cannot communicate with each other.



In order to control the forwarding, the switch will add VLAN tag to Ethernet frame before forwarding it, and then decide how to deal with the tag and frame, including discarding of the frame, forwarding frames, adding tag and moving tags.

Before forwarding the frame, the switch will check VLAN tag of packet, whether the tag is allowed to pass the port, so as to decide whether the frame can be forwarded from the port. In the figure above, if the switch adds tag 5 to all the frames sent from A, and then look up the layer-2 forwarding table, and according to destination MAC address forward them to the port connected to B. But this port is configured to only allowed VLAN 1 to pass, so the frames sent by A will be discarded.

Hence, switch supporting VLAN will forward Ethernet frames not only according to destination MAC address but also VLAN configuration of ports, so as to implement layer-2 forwarding control.



4-byte VLAN tag is added to Ethernet frame header directly. Document IEEE802.1Q describes VLAN tagging.

TPID：Tag Protocol Identifier，2 bytes，fixed value，0x8100，new type defined by IEEE, it indicates that it is a frame with 802.1Q tag.

TCI：Tag Control Information，2 bytes.

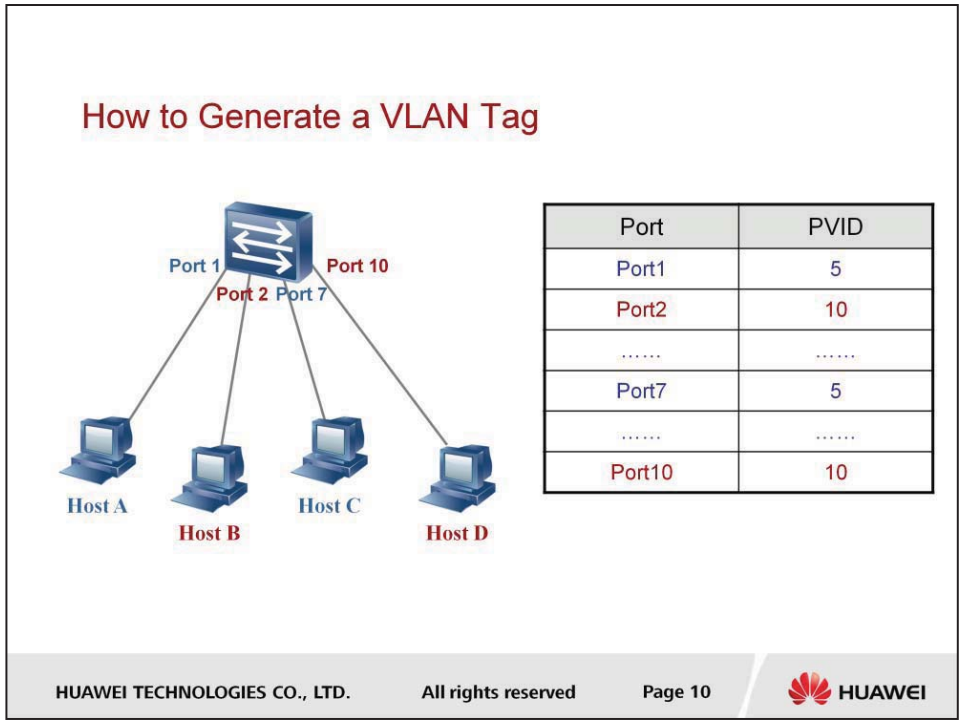
• Priority：3 bits, the priority of Ethernet frame. It has 8 priorities, 0—7, used to provide differential forwarding service.

• CFI：Canonical Format Indicator，1 bit. Used to indicate bit order of address information in token ring or source route FDDI media access, namely, whether the low bit is transmitted before high bit.

• VLAN Identifier：VLAN ID，12 bits, from 0 to 4095. Combined with VLAN configuration of port, it can control the forwarding of Ethernet frame.

Ethernet frame has two formats: the frame without tag is called untagged frame; the frame with tag is called tagged frame.

This section will only discuss the VLAN ID of VLAN tag.



All the Ethernet frames exist in switch in the form of tagged frames, maybe a certain port receives untagged frame from peer device, but the frame from the port of the local switch must be tagged frame. If the frame received is tagged, it will be forwarded; if it is untagged, tag will be added to it. The following methods can confirm the VLAN ID in a tag:

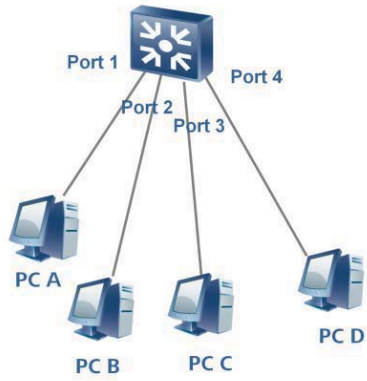
- Based-on port: Network manager configures a PVID for every port of switch, Port VLAN ID， it is also called port default VLAN. If an untagged frame is received, the VLAN ID will be PVID.
- Based-on MAC: Network manger configures the mapping relation between MAC address and VLAN ID, if an untagged frame is received, VLAN ID will be added according to the mapping relation table.
- Based-on protocol: Network manger configures the mapping relation between protocol filed of the Ethernet frame and VLAN ID, if an untagged frame is received, VLAN ID will be added according to the mapping relation table.
- Based-on subnet: add VLAN ID according to IP address information in packet.

If the device can support multiple methods at one time, in general, the priority order from high to low is : based-on subnet—based-on protocol—based-on MAC address—based-on port. Presently, based-on port is the most common method.

VLAN Partitioning

- According to the port: based on port VLAN
- According to the MAC: based on MAC VLAN
- According to the IP: based on IP subnet VLAN
- According to the protocol: based on protocol VLAN
- According to several ways division: based on policy VLAN

Port based VLAN



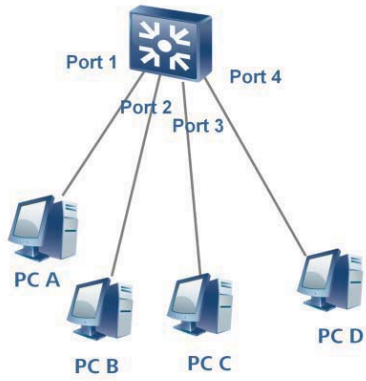
VLAN information

VLAN 10	VLAN 20	VLAN 30
Port1	Port 2 Port 3	Port4

 Page 13 HUAWEI

The VLAN distinguishes division based on port. This is the most concise, and the most widely used partition method, and can put more ports into a VLAN.

MAC based VLAN

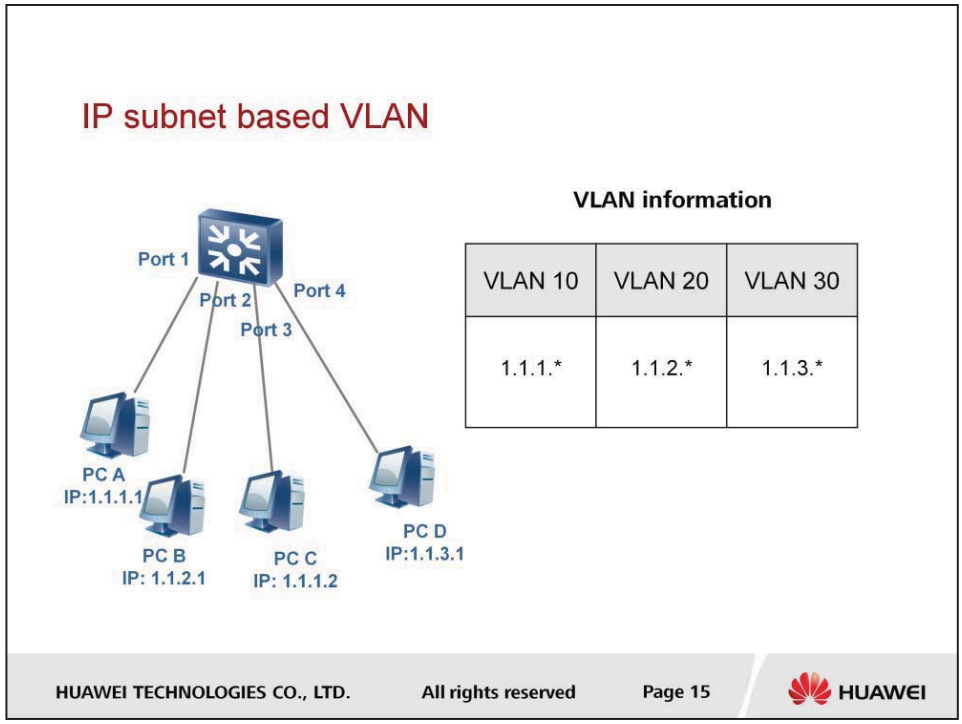


VLAN information

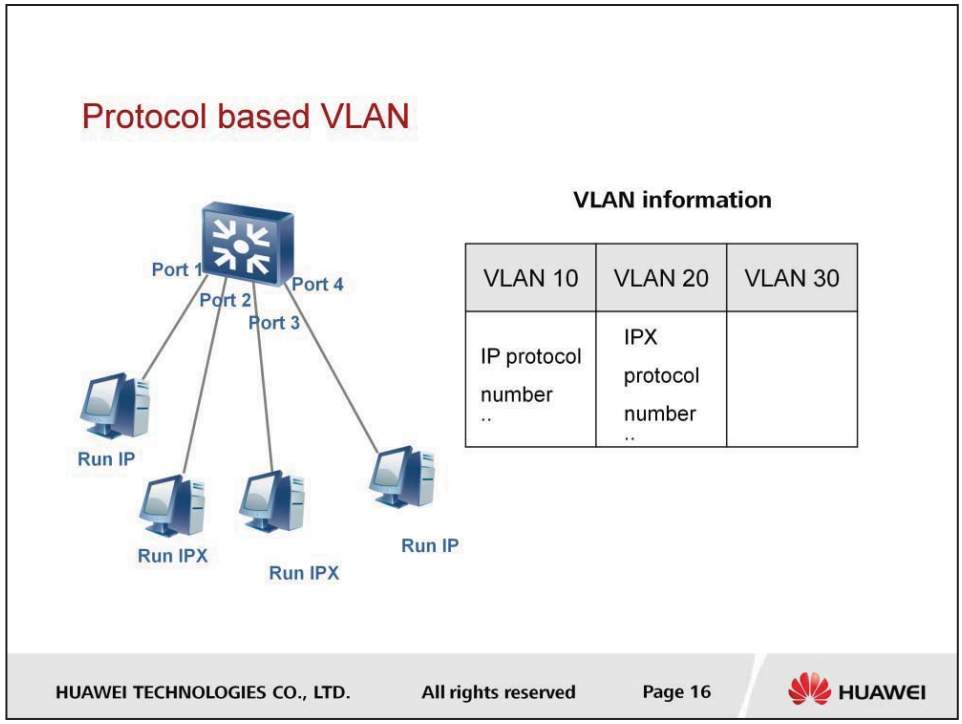
VLAN 10	VLAN 20	VLAN 30
PC-A MAC	PC-B MAC PC-C MAC	PC-D MAC

  HUAWEI

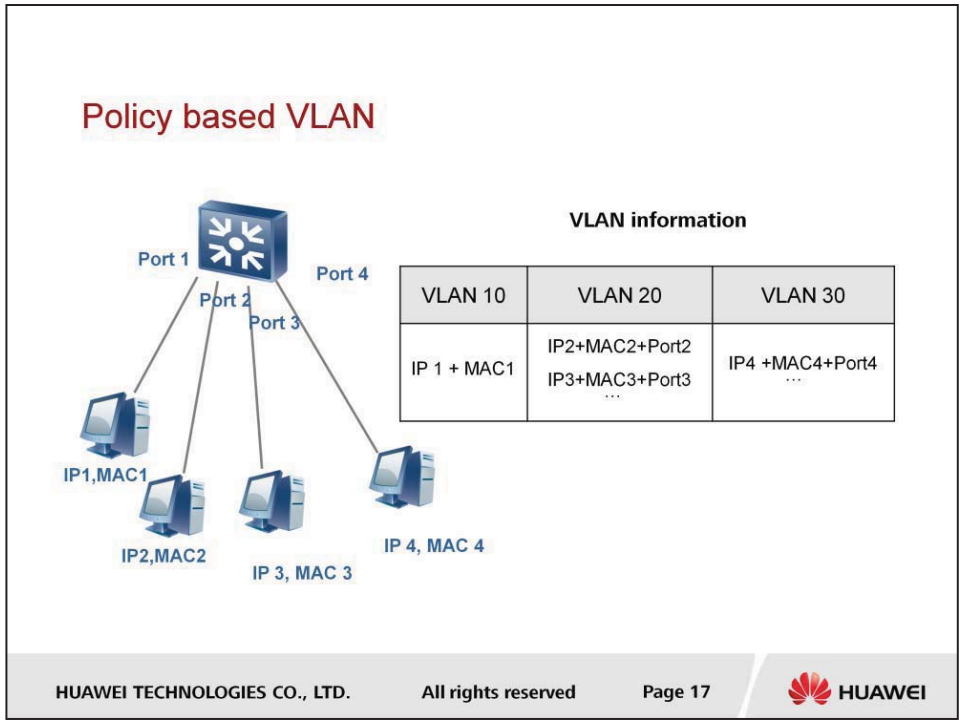
MAC based VLAN: switches according to the source MAC of the message to determine which VLAN the message should be in to forwarding. In fact is up to the MAC terminal equipment to differentiate VLANs.



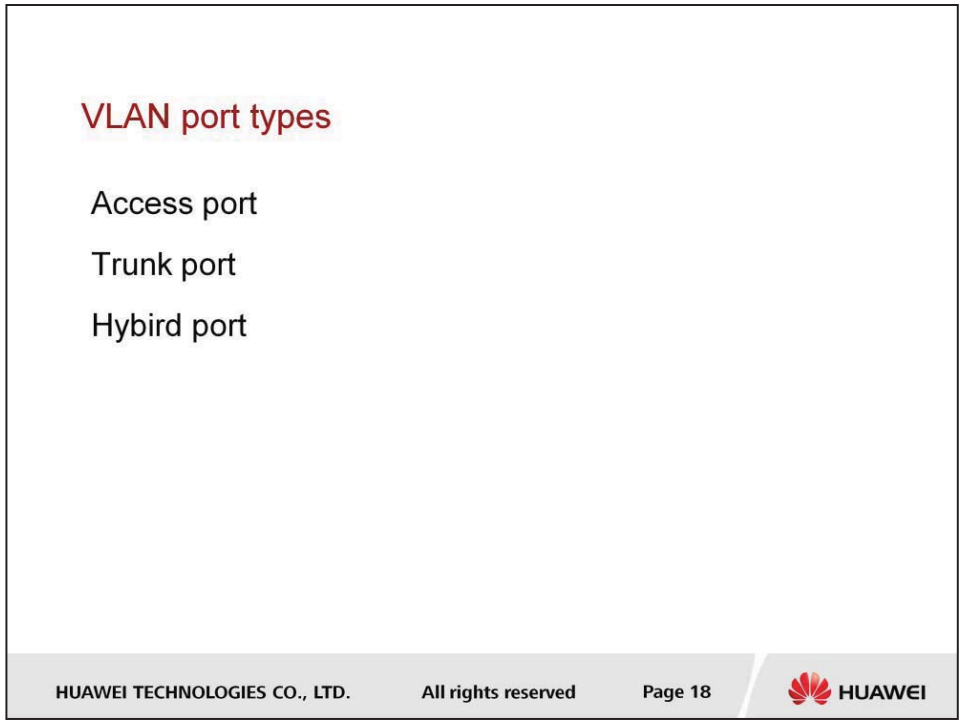
Based on IP network segment division VLAN, according to the source IP and mask of packet to determine to which VLAN a frame belongs. For example, can configure 1.1.1.0/24 into VLAN 10, 1.1.2.0/24 into VLAN 20, 1.1.3.0/24 into VLAN 30. and then configure some port belong to these VLANs. On this port, to the received message without VLAN tag: the packet's source IP among 1.1.1.0/24 will be put VLAN10 tag, among 1.1.2.0/24 will be put VLAN20 tag, among 1.1.3.0/24 will be put VLAN30 tag.



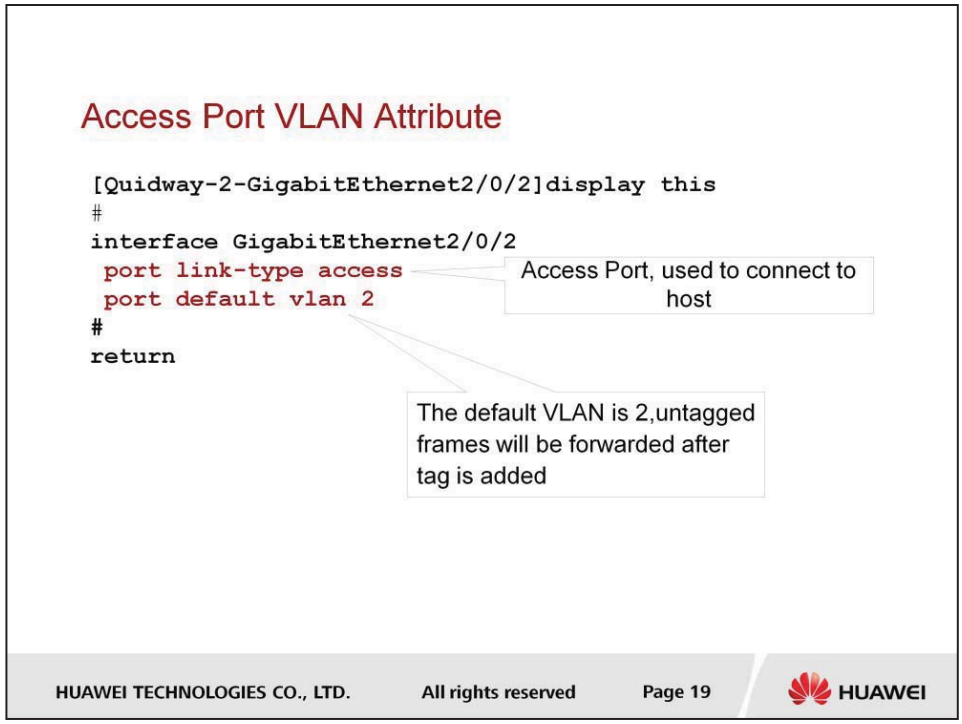
Protocol-based VLAN functions according to the type of port the received packet belongs protocol (family) and package format to which the packet is assigned a different VLAN ID. Such as IP, IPX, and AppleTalk protocol suite; Ethernet II, 802.3, / 802.2 LLC, 802.3/802.2 of SNAP and other package formats. Untagged packets received on port are determined its protocol type by the equipment, marked with the corresponding VLAN tag and forwards in this VLAN.



Policy-based VLAN: Dividing the VLAN for untagged packets that match with policy IP+MAC or IP+MAC+PORT.



The switch port has been divided into 3 types: the Access port, trunk port, and hybrid port types, after the introduction of VLAN function .



Access port, is used to connect host, it has features as followings:

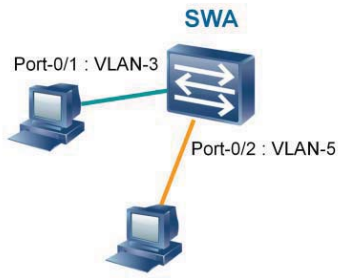
• Only permit unique VLAN ID to pass the port, the VLAN ID is the same with PVID of the port.

• If the frame received from peer device is untagged, the switch will add PVID to the frame by force.

• The frame sent by Access port is always untagged frame.

• The default port type of many types of switches are access, PVID is 1 by default, VLAN 1 is created by system and can not be deleted.

Configure Access Port Attribute




```
\\configure port type
[Switch-Ethernet0/1]port link-type access
[Switch-Ethernet0/2]port link-type access

\\create VLAN
[Switch]vlan 3
[Switch]vlan 5

\\set port PVID
[Switch-Ethernet0/1]port access vlan 3
[Switch-Ethernet0/2]port access vlan 5
```

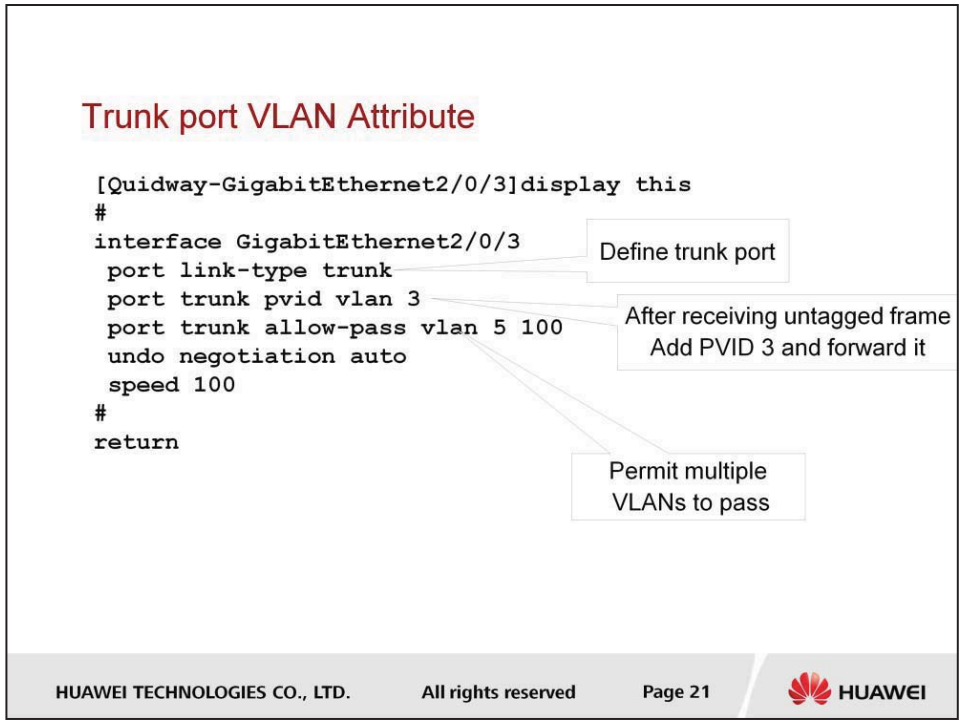

The following commands can be used to set PVID of access port: add access ports into the VLAN after creating the VLAN:

[Switch]vlan 3

[Switch-vlan3]port ethernet 0/1

[Switch]vlan 5

[Switch-vlan5]port ethernet 0/2

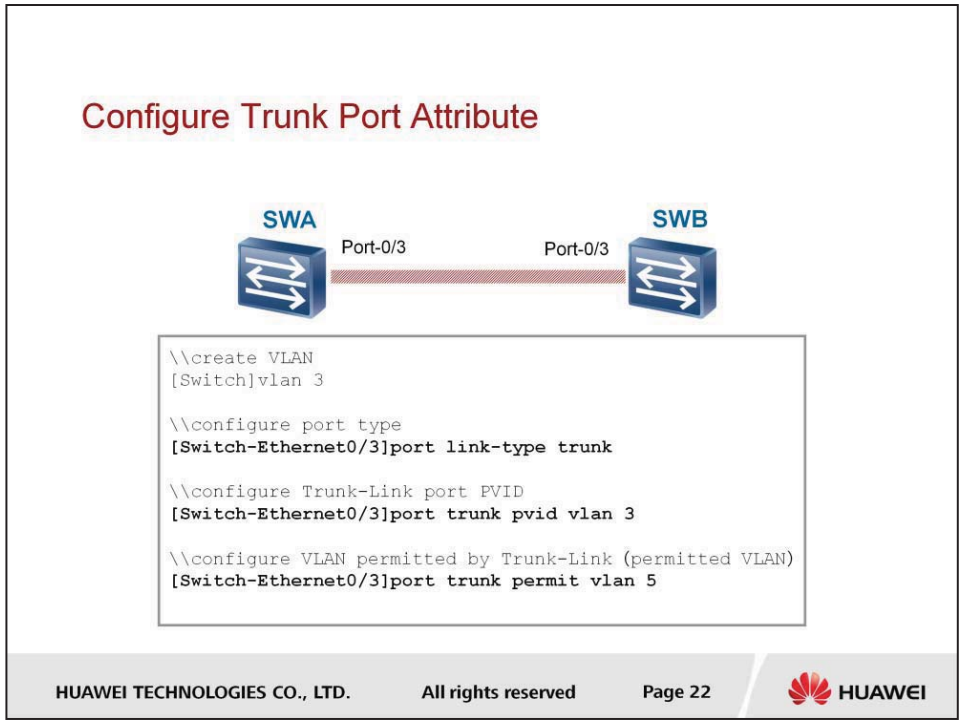


Trunk port: used to connect switches, transmit tagged frames among switches. It can be set to permit multiple VLAN IDs, these IDs can be the same with PVID, and also can be different. Trunk port is going to send Tagged frame to other devices, rules as followings:

If the VLAN ID of the tagged frame does not exist in VLAN permitted list, it will be discarded; if it exists, then:

• If the VLAN ID of the tagged frame is the same with PVID, the frame will be sent to another device after removing the tag. Because the PVID of each port is unique, only in this case, the frame is untagged sent by trunk port;

• If the VLAN ID of the tagged frame is different from PVID, the frame will be forwarded to peer device without modification. VLAN passing： In general, the query content of VLAN passing and VLAN permitted is the same. But if a VLAN which is registered by GVRP, if did not register on port, the VLAN ID will not exist in VLAN passing list, and the corresponding VLAN frame can not be forwarded from the port.



As shown in the figure above, the following commands can be used to configure Trunk port attribute:

\\create VLAN

[Switch]vlan 3

\\configure port type

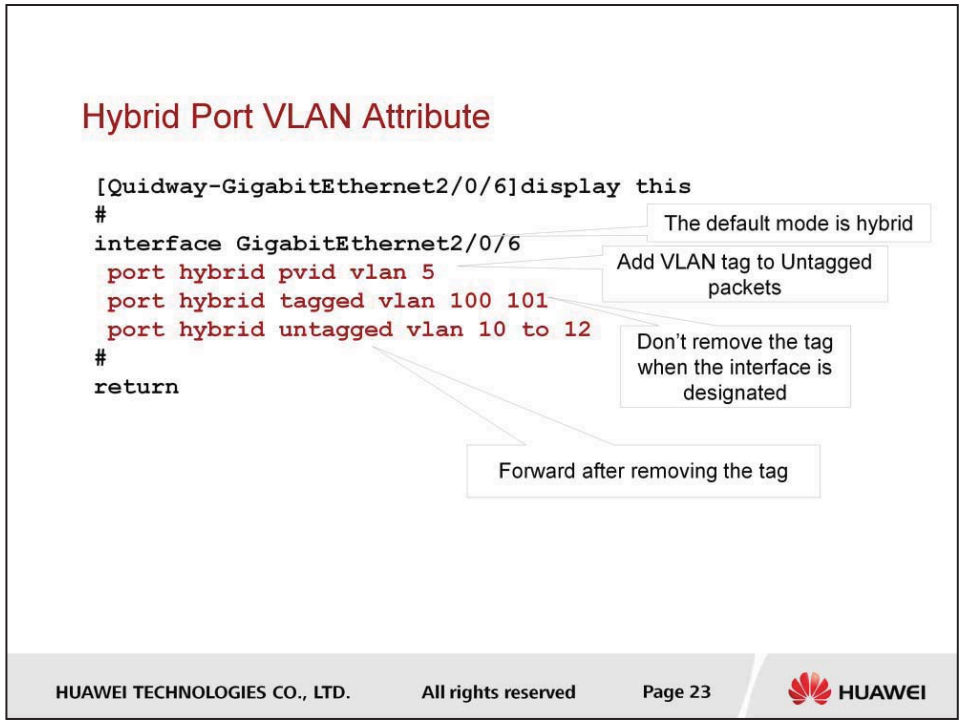
[Switch-Ethernet0/3]port link-type trunk

\\configure PVID of Trunk-Link port

[Switch-Ethernet0/3]port trunk pvid vlan 3

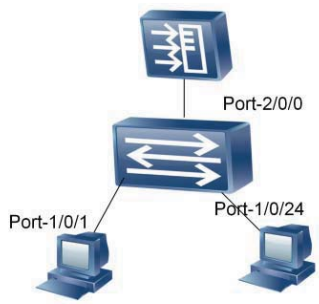
\\configure permitted VLAN of Trunk-Link

[Switch-Ethernet0/3]port trunk permit vlan 5



Access ports send packets to other devices in the form of untagged frame, and a trunk port can send out untagged frame only in one special situation. In other cases, it sends tagged frame. In some applications, it is hoped to neatly control VLAN tag. For example, the up-stream device of switch can not support VLAN, but the user ports can be isolated from each other. Hybrid port can flexibly control the VLAN tag. In this example, if the VLAN ID of frame is 3, then forward it according to the forwarding mode of trunk port. If it is 4, move away tag 4 and then forward it.

Configure Hybrid Port



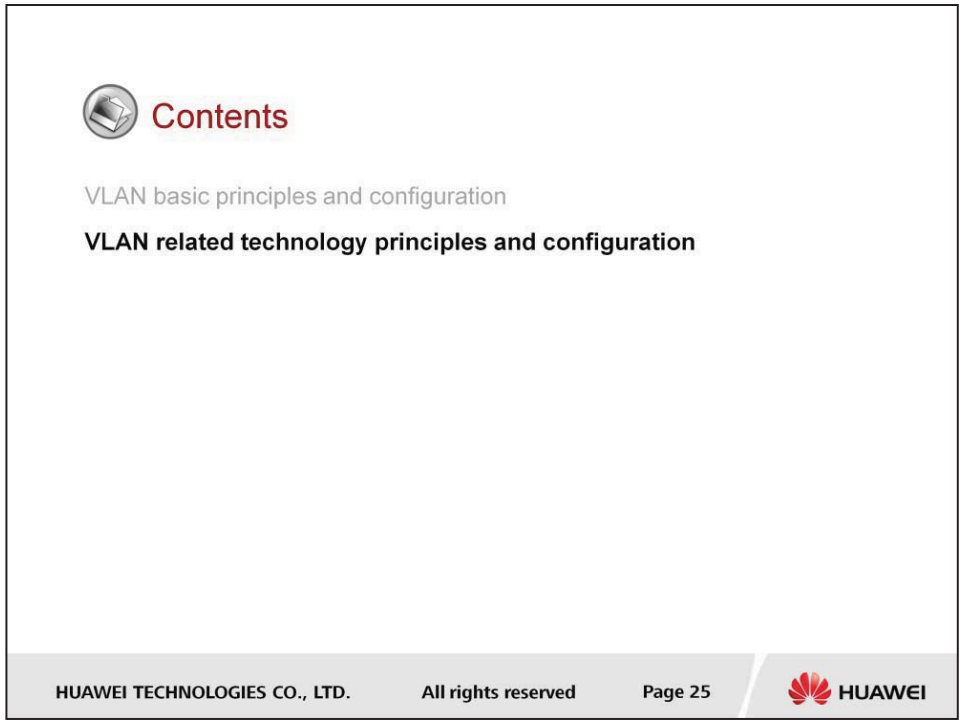

```
[Quidway-Ethernet1/0/1]port link-type hybrid
[Quidway-Ethernet1/0/1]port hybrid pvid vlan 2
[Quidway-Ethernet1/0/1]port hybrid vlan 2 untagged
[Quidway-Ethernet1/0/1]port hybrid vlan 99 untagged

[Quidway-Ethernet1/0/24]port link-type hybrid
[Quidway-Ethernet1/0/24]port hybrid pvid vlan 3
[Quidway-Ethernet1/0/24]port hybrid vlan 3 untagged
[Quidway-Ethernet1/0/24]port hybrid vlan 99 untagged

[Quidway-Ethernet2/0/0]port link-type hybrid
[Quidway-Ethernet2/0/0]port hybrid pvid vlan 99
[Quidway-Ethernet2/0/0]port hybrid vlan 2 to  3 untagged
```


If the tagged VLAN of Hybrid port is none, untagged VLAN has only one value, then the port has the same function with access port. if the port does not configure untagged VLAN, then it has the same function with trunk. If set all the switch ports to different VLAN attribute, for example 2, 3……24, and up-stream interface Untagged VLAN ID list is 2,3……24, so that it can implement users isolation from each other and enhance security. At the same time, upstream frame is untagged frame, it satisfies the requirement of communication with up-stream devices.

The configuration above can implement isolation between port 1/0/1 and port 1/0/24, but they can communicate with up-stream devices, and the up-stream frame is untagged frame.

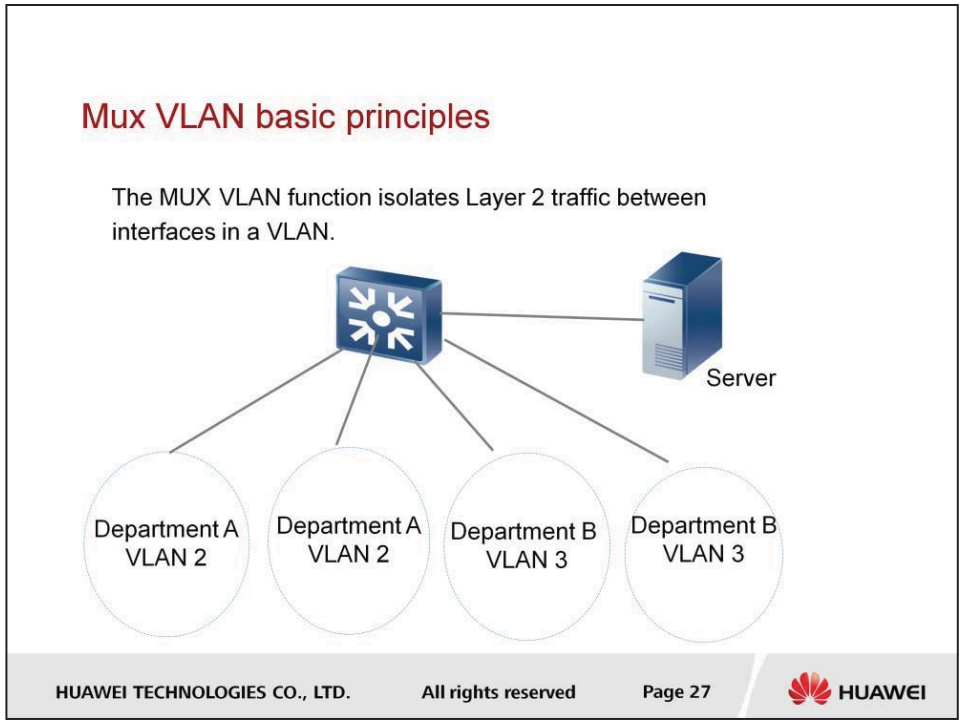
\\ tagged VLAN configuration example

[Quidway-Ethernet2/0/0] port hybrid vlan 3 tagged


Contents
VLAN basic principles and configuration
VLAN related technology principles and configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 25
HUAWEI

Contents

- VLAN related technology principles and configuration
  - ⇨ Mux VLAN basic principles and configuration
  - ⇨ Super VLAN principles
  - ⇨ ARP Proxy
  - ⇨ VLAN Mapping
  - ⇨ Port-Isolate



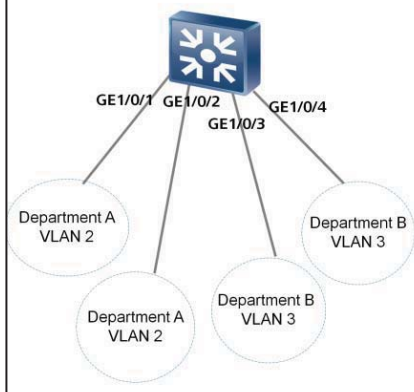
For example, in a enterprise network, the client port can communicate with the server port, client port can not communicate with each other.

MUX VLAN can be divided into principal VLAN and subordinate VLAN ,and subordinate VLAN can also be divided into separate VLAN and group VLAN.

Principal VLAN and subordinate VLAN can communicate with each other, group VLAN can communicate with each other through ports, separate VLAN can not communicate with each other through ports, different VLANs can not communicate with each other.

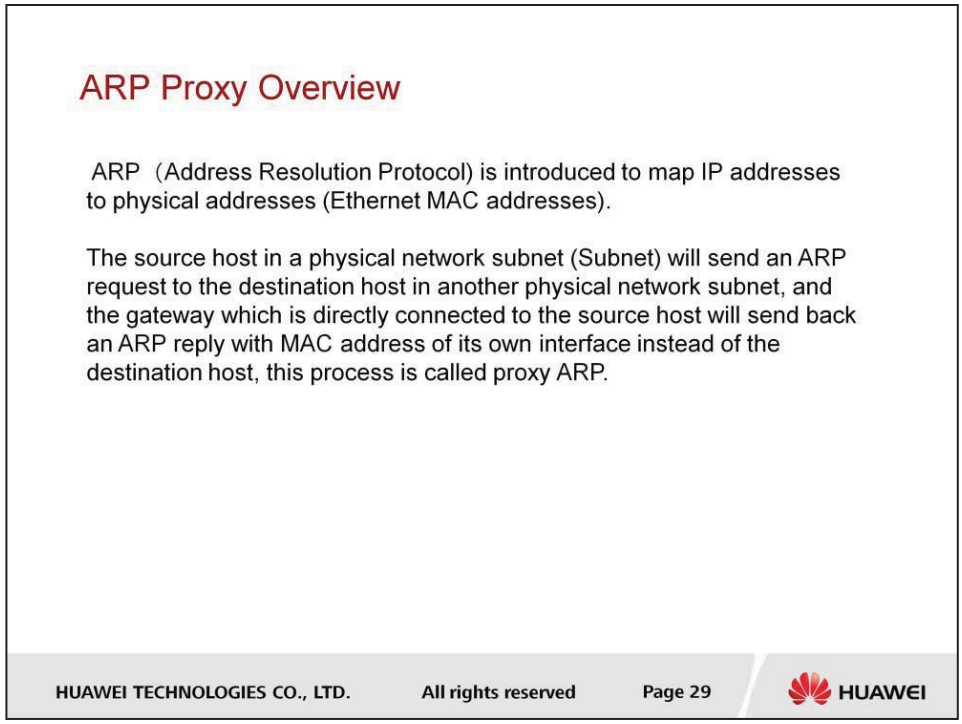
Employees of the Department A can not communicate with each other, and employees of the Department B can communicate with each other , but Department A and Department B can not communicate with each other. All employees can access the company's server. For this application, you can use the MUX VLAN to achieve. Employees of the department A add to the isolated slave VLAN, and staffs of Department B add to the interoperable slave VLAN ,the server add to the master VLAN.

Mux VLAN Configuration




```
[Quidway]vlan batch 2 3 10
//Create VLAN
[Quidway]vlan 10
[Quidway-vlan10]mux-vlan
//Configure principal VLAN
[Quidway-vlan10]subordinate group 3
//Configure subordinate group VLAN
[Quidway-vlan10]subordinate separate 2
//Configure subordinate separate VLAN
[Quidway]interface gigabitethernet1/0/1
[Quidway-GigabitEthernet1/0/1]port mux-vlan enable
[Quidway-GigabitEthernet1/0/2]port mux-vlan enable
[Quidway-GigabitEthernet1/0/3]port mux-vlan enable
[Quidway-GigabitEthernet1/0/4]port mux-vlan enable
```

  HUAWEI



The ARP Proxy basic process is as follows:

Source host send an ARP request to the destination host in another physical network subnet; Enable ARP PROXY function on the gateway connected with the source host, if there is the normal route to reach the destination host, instead of the MAC address of the destination host to respond with their own interface.

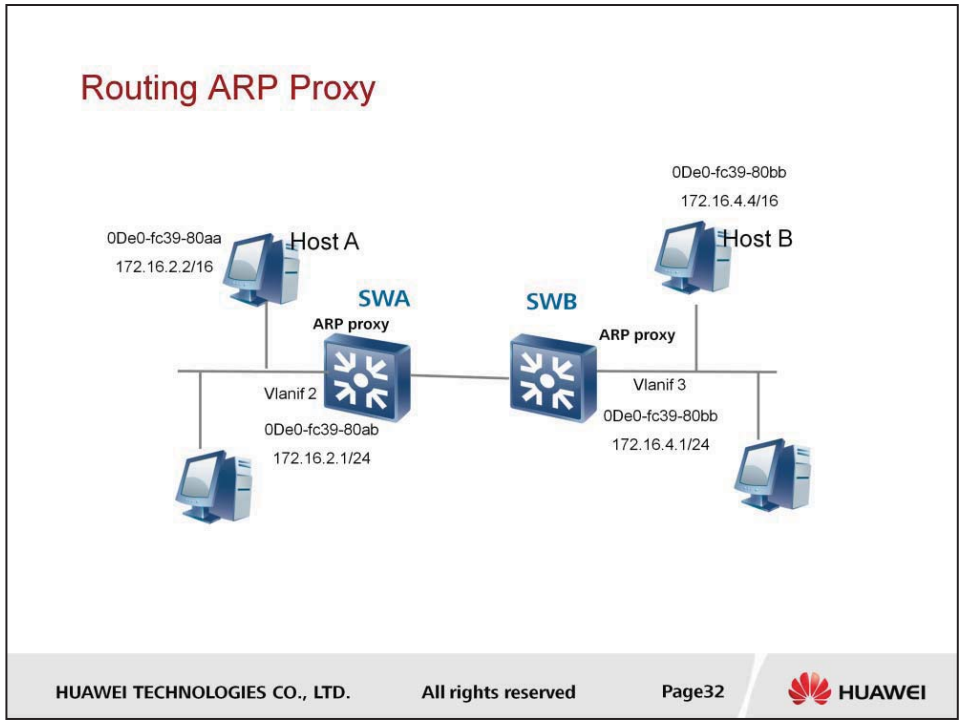
Source host to destination host to send IP packets are sent to the router. Packets are forwarding through normal IP routing. IP packets that sent to the destination host through the network, and ultimately arrive at the destination host.

The Basic Principles of ARP Proxy

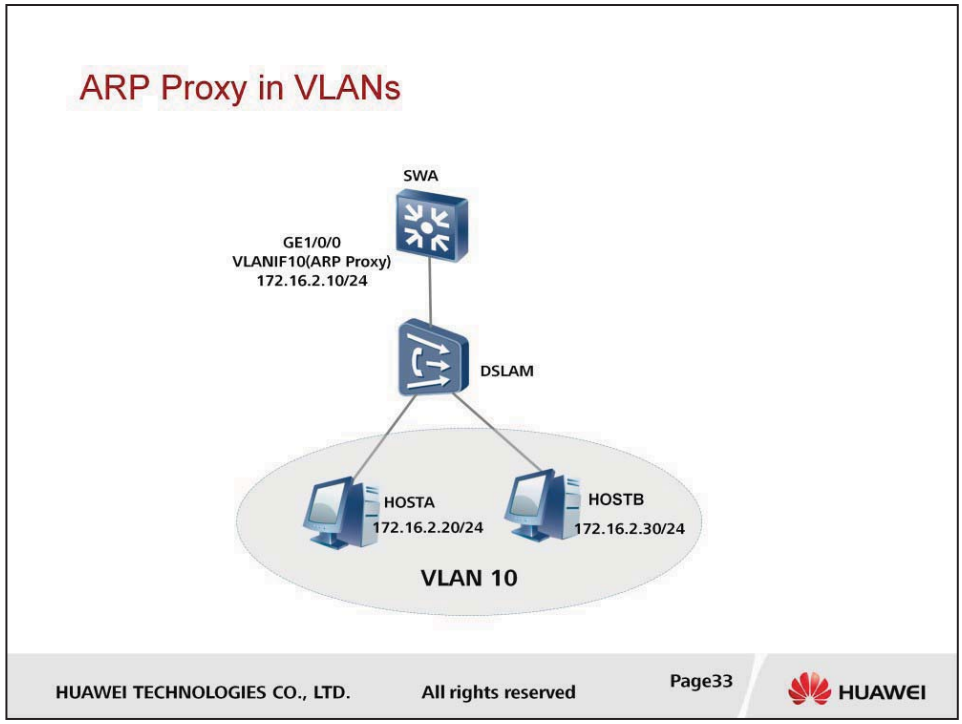
When the host has no configured default gateway address, it can send an ARP request to the MAC address of the requested destination host. The switch receiving such a request that enable the ARP proxy function will use its own MAC address in response as the ARP request, so hosts in a different physical networks but the same network number, among them can communicate with each other normally.

ARP Proxy Way

ARP Proxy Way	Problems solved
Routing ARP Proxy	Solve the interoperability issue of when computers on the same segment exist in different physical networks.
ARP Proxy in VLANs	Solving computer interoperability problems on the network within the same VLAN, but user isolation has been configured between VLANs.
ARP Proxy among VLANs	Solves the three-layer interoperability issues between different VLAN corresponding computers.



Route-ARP proxy is to solve the interoperability for those computers or switches that are not in the same segment but on the same physical network. In practical applications, if the host connected to the switch doesn’t configure the default gateway address, the data will not be forwarded. Routing ARP Proxy can solve this problem, the host sends an ARP request (request destination host's MAC address). After the switch that enables the ARP proxy function receives such a request, it will use its own MAC address in response to the ARP request, in order to deceive the host to forward data. The switch that enables the ARP proxy function can also hide the details of the physical network, so between the different physical networks, but the same network number for Ethernet A and B, the Ethernet internal host normally communicate with each other.

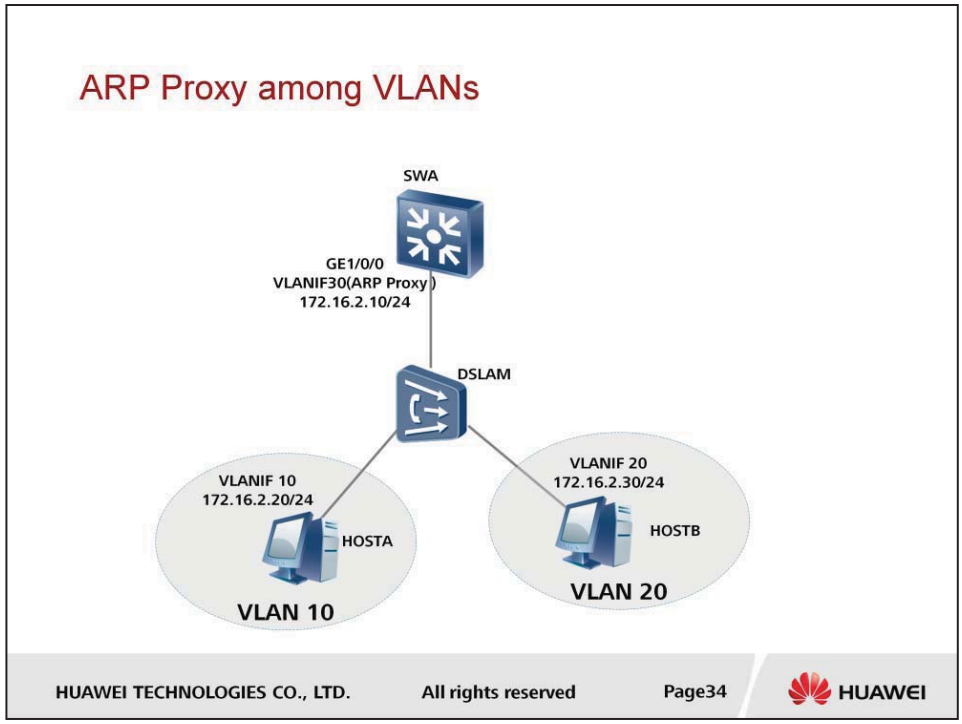


If two users belong to the same VLAN, but user isolation has been configured in VLAN. Interoperability between users need to associate the VLAN interface to enable VLAN proxy ARP.

If interface of the switch that enables the Proxy ARP in VLAN , the destination address received is not its own ARP request packet, the switch does not immediately discard the packet, but find the interface the ARP table. If you meet the proxy, the square is the switch MAC address send an ARP request.

VLAN within the Proxy ARP is primarily used to configure user isolation VLAN interoperability between users.

HOST A and HOST B is the two users in the DSLAM. Connect HOST A and HOST B, the two interfaces belong to the same VLAN 10 as the DSLAM. As in the DSLAM to configure the VLAN interface separated from each other, so HOST A and HOST B on the second floor can not be directly interoperable. If created in S5700 the interface VLANIF10. Enable Proxy ARP, VLAN, HOST A and HOST B can communicate at Layer in VLANIF10. IP address of VLANIF10 and the host IP address of VLAN10 must be in the same network segment.



If two users belong to different VLANs, the three-layer interoperability between users need to enable inter-VLAN proxy ARP on the associated VLAN interface .

If interface of the switch that enables the Proxy ARP in VLAN , the destination address received is not its own ARP request packet, the switch does not immediately discard the packet, but find the interface the ARP table. If you meet the proxy, the square is the switch MAC address send an ARP request.

VLAN between Proxy ARP is mainly used for:
Different VLAN users can communicate in layer 3.

Inter-VLAN proxy ARP enabled on the interface of the Super VLAN corresponding VLANIF can achieve the interoperability of the Sub VLANs between users.

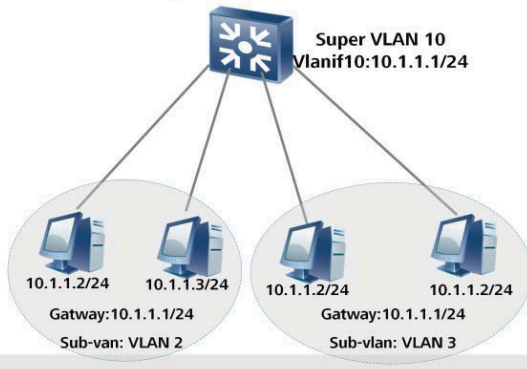
As shown in figures:

HOST A and HOST B is the two users in the DSLAM. Two interfaces to connect the HOST A and HOST B belong to different VLANs on the DSLAM HOST A and HOST B can not be directly interoperable in layer 2.

SWA on to create the Super VLANs 30 VLAN10 and the VLAN20 join VLAN30, and create the interface VLANIF 30, on in VLANIF 30 between the VLAN Proxy the ARP HOST A and HOST B can be exchange in layer 3. IP address of VLAN IF 30 and the host IP address of VLAN10 and VLAN20 are in the same network segment.

Super VLAN Principle

VLAN aggregation, configure the IP address only in the super-VLAN interface , without having to assign IP addresses for each of the sub-VLANs. All sub-VLANs share the IP network segment, to solve the problems of the of wasting IP address resources.



  HUAWEI

VLAN is widely applied to switching networks because of its flexible control of broadcast domains and convenient deployment. On a layer 3 switch, the interconnection between the broadcast domains is implemented by using one VLAN to correspond to one Layer-3 logic interface. However, this wastes IP addresses.

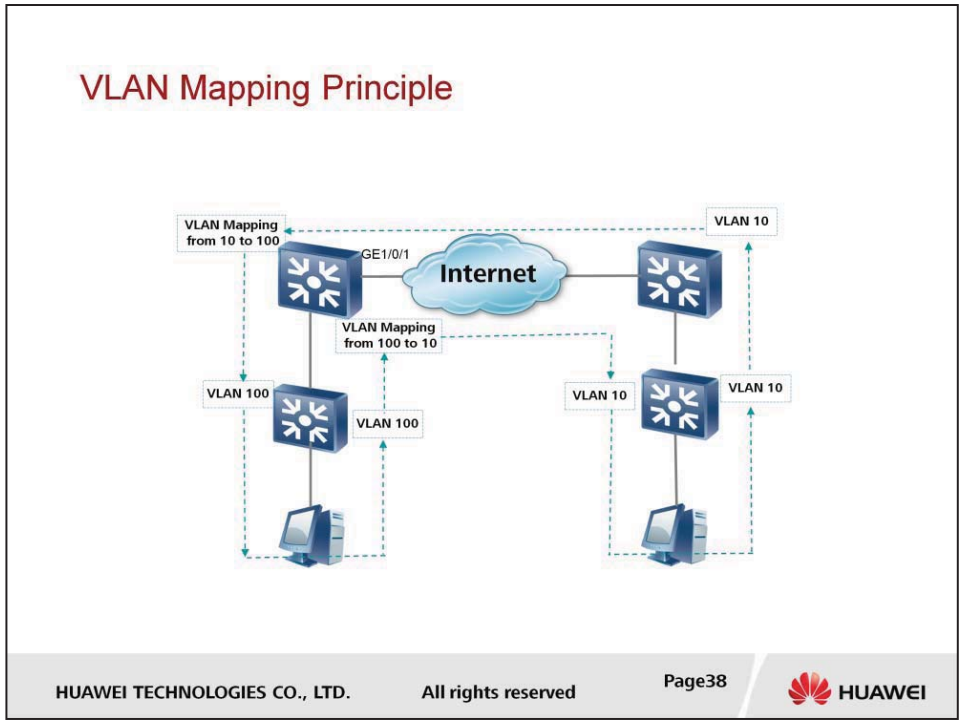
The VLAN aggregation technology, also known as the super-VLAN, provides a mechanism that partitions the broadcast domain by using multiple VLANs in a physical network so that different VLANs can belong to the same subnet. In VLAN aggregation, two concepts are involved, namely, super-VLAN and sub-VLAN.

Super-VLAN: It is different from the common VLAN. In the super-VLAN, only Layer 3 interfaces are created and physical ports are not contained. The super-VLAN can be viewed as a logical Layer-3 interface. The super-VLAN is a collection of many sub-VLANs.

As shown, the super-VLAN 10 consists of the sub-VLAN2 sub-VLAN3. sub-VLAN 2, sub-VLAN3 and the super-VLAN 10 belongs to the same subnet 10.1.1.0/24.
gateway address in the sub-VLANs 2 and sub-VLANs 3 is a host interface address of the Vlanif 10 in super-VLAN 10. Host in different the sub-VLANs can not be exchanged. Interoperability of the Super-VLAN should enable the ARP Proxy on VLAN interface .

VLAN Mapping

VLAN Mapping, also known as VLAN translation, can be achieved in the mutual conversion between the users' VLAN IDs (private VLAN) and the operators' VLAN ID (business VLANs can also be said to be a function of the mutual conversion between the public VLANs).



VLAN mapping occurs when frames are received by the inbound port and when frames are forwarded by the outbound port. If VLAN mapping is configured on a port, when the port sends frames from the local VLAN to the remote VLAN, the port replaces the local VLAN ID in the frames with the remote VLAN ID. When the port receives frames from the remote VLAN to the local VLAN, the port replaces the remote VLAN ID in the frames with the local VLAN ID. In this manner, inter-VLAN communication can be implemented.

As shown in figure, VLAN mapping between VLAN 100 and VLAN 10 is configured on GE 1/0/1. When GE 1/0/1 sends frames from VLAN 100, it replaces VLAN tag with VLAN 10. When GE 1/0/1 sends frames from VLAN 10,it replaces VLAN tag with VLAN 100. In this manner, VLAN 100 and VLAN 10 can communicate with each other.

VLAN Mapping Configuration

```
[Quidway] interface gigabitethernet 2/0/1
[Quidway-GigabitEthernet2/0/1]port link-type trunk
```

// Configure the type of the interface

```
[Quidway-GigabitEthernet2/0/1]port trunk allow-pass vlan
  100
```

// configure the VLAN that interface allows after the VLAN mapping

```
[Quidway-GigabitEthernet2/0/1]qinq vlan-translation enable
```

// Enable interface VLAN conversion function

```
[Quidway-GigabitEthernet2/0/1]port vlan-mapping vlan 1 to
  20 map-vlan 100
```

// configure packets on VLAN1 to 20 on interface 2/0/1 mapping into VLAN100

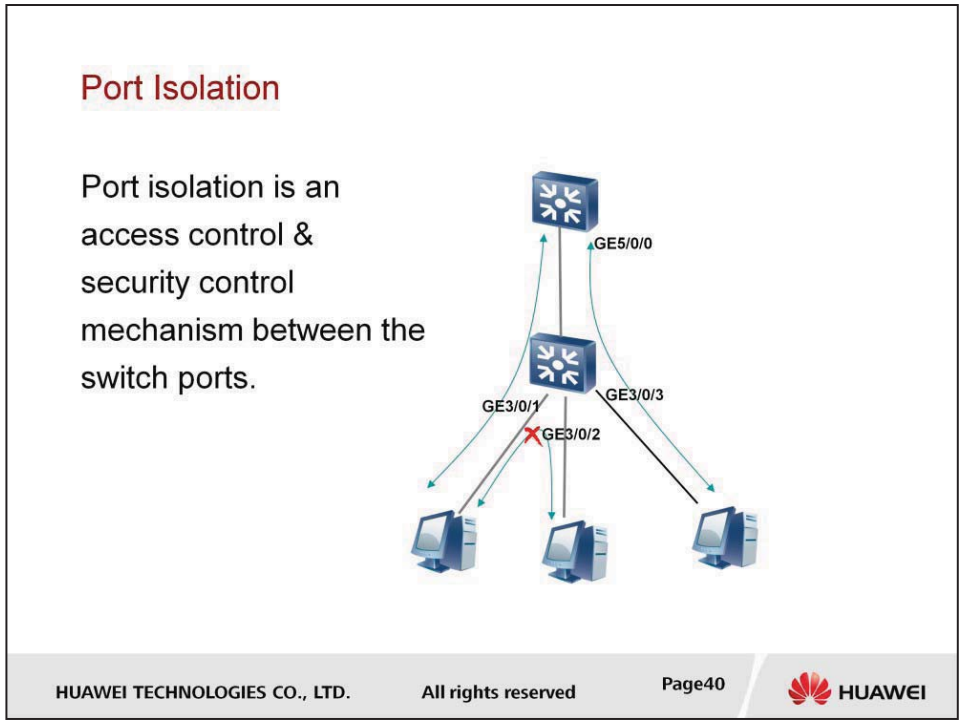
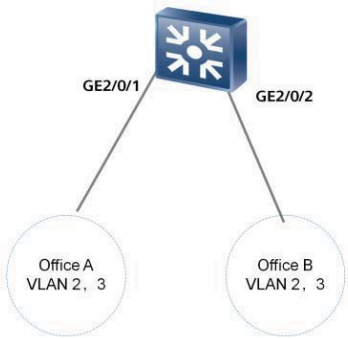


Figure above is to achieve that between the different port access PCs cannot communicate with each other (PC belongs to the same VLAN), all of the PC through the uplink switches to access the network, you can configure the port GE3/0/1 ,GE3/0/2,GE3/0/3 for port isolation, and GE5/0/0 and port GE3/0/1, GE3/0/2, GE3/0/3, don't isolate, to achieve this security control requirements.

Port isolation configuration



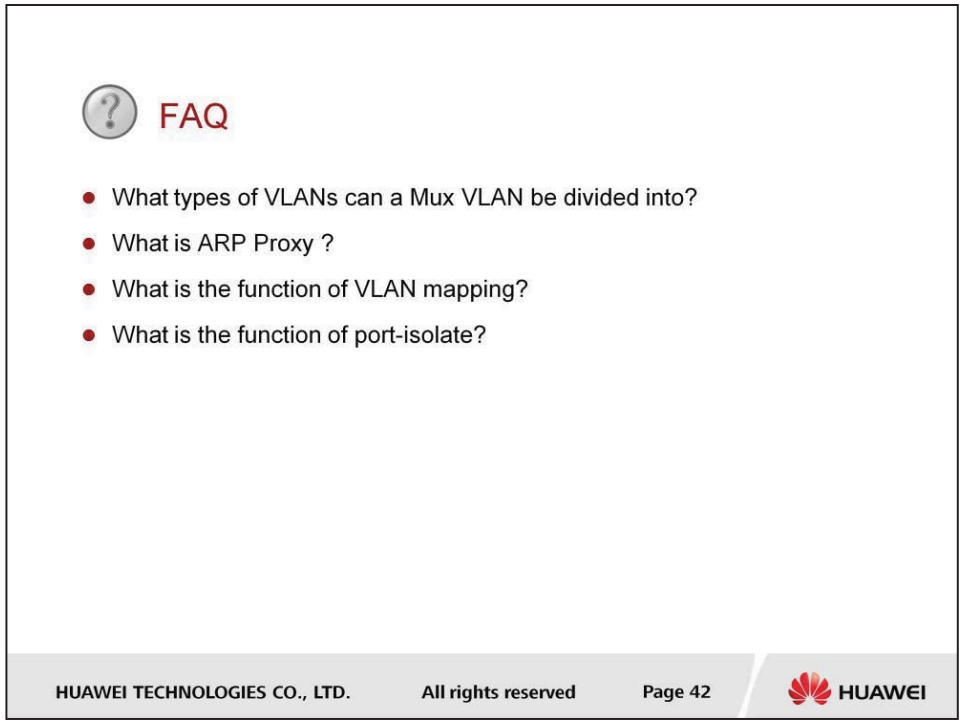

```
[Quidway] port-isolate mode all
```
// Configure port isolation mode

```
[Quidway] interface gigabitethernet2/0/1
[Quidway-GigabitEthernet2/0/1]port-
isolate enable
```
// enable port isolation in the port GE2/0/1

```
Quidway-GigabitEthernet2/0/1]interface
gigabitethernet2/0/2
[Quidway-GigabitEthernet2/0/2]port-
isolate enable
```
// enable port isolation in the port GE2/0/2

 Page41 HUAWEI

As shown, the need to configure office A and office B isolated from each other. Office A may have multiple VLAN users, office area B may also have multiple VLAN users. You can configure the port isolation to achieve A, B isolation. Mux VLAN can be configured between users within the VLAN, isolated from each other, or interoperability. Port Isolation is isolation on the physical level, and based on port. After configuring the port isolation between two ports, they can not communicate.



Q:What types of VLANs can Mux VLAN be divided into?

A:Mux VLAN can be divided into principal VLAN and subordinate VLAN ,and subordinate VLAN can also be divided into separate VLAN and group VLAN.

Q:What is ARP Proxy ?

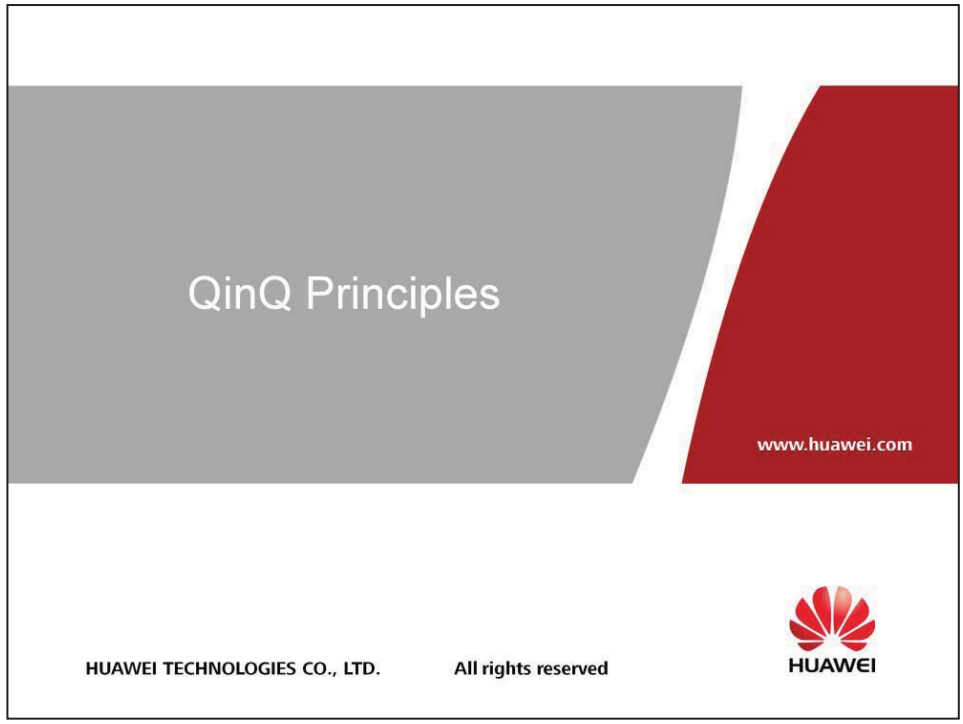
A:The source host in a physical network subnet (Subnet) sends an ARP request to the destination host in another physical network subnet, and the gateway which directly connected to the source host send back a ARP reply with MAC address of its own interface instead of the destination host, and the process is called proxy ARP.

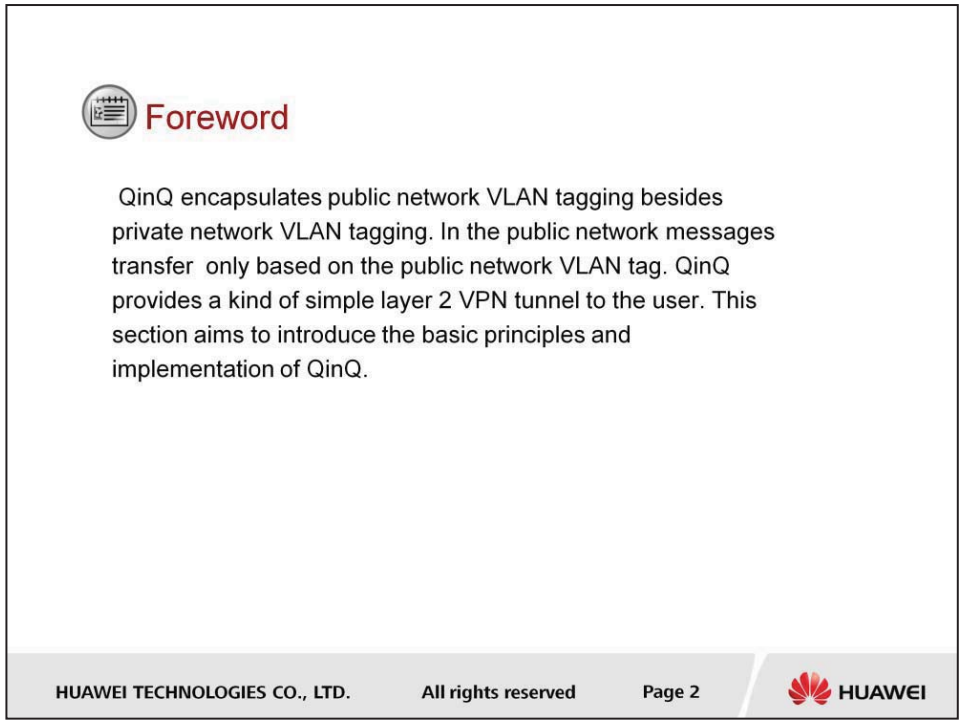
Q:What is the function of VLAN mapping?

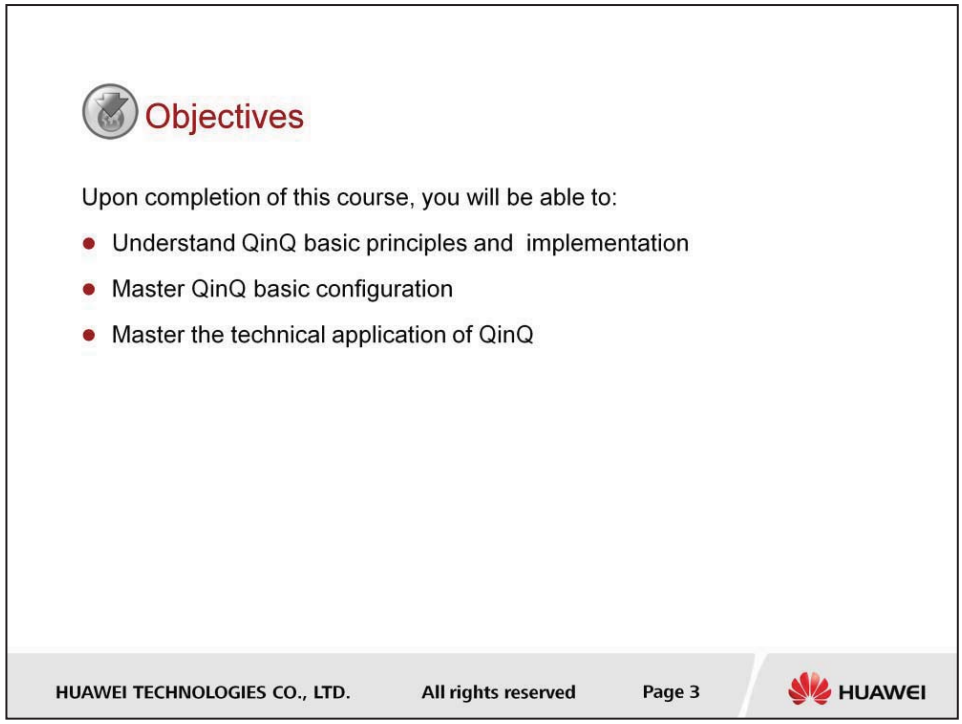
A:VLAN Mapping can be achieved in the mutual conversion between the users ' VLAN IDs (private VLAN) and the operators' VLAN ID (business VLANs can also be said to be a function of the mutual conversion between the public VLANs).

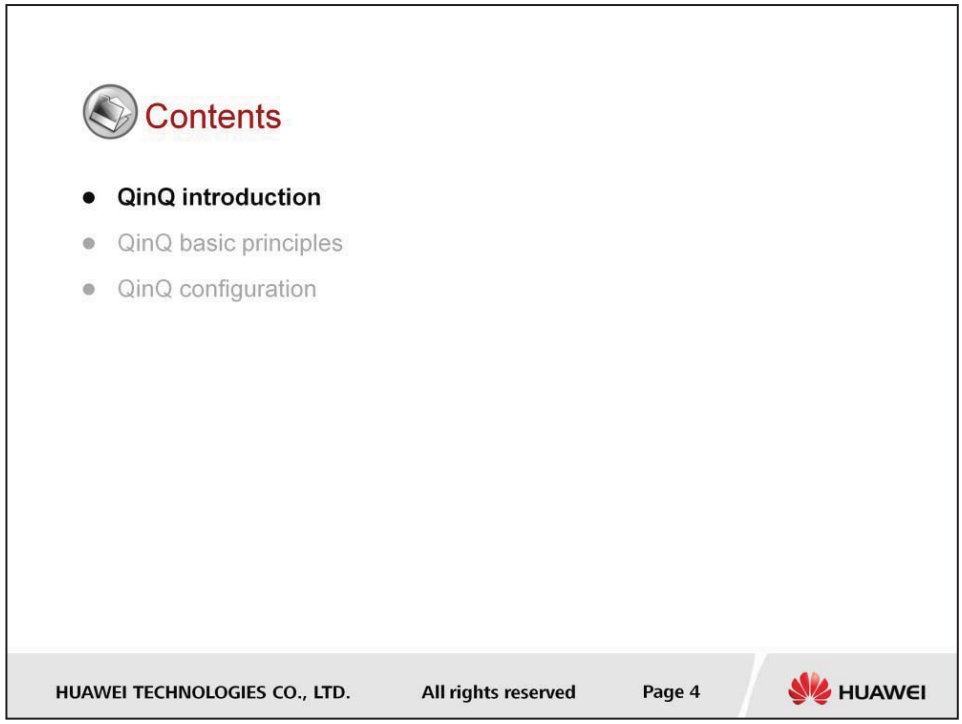
Q:What is the function of port-isolate?

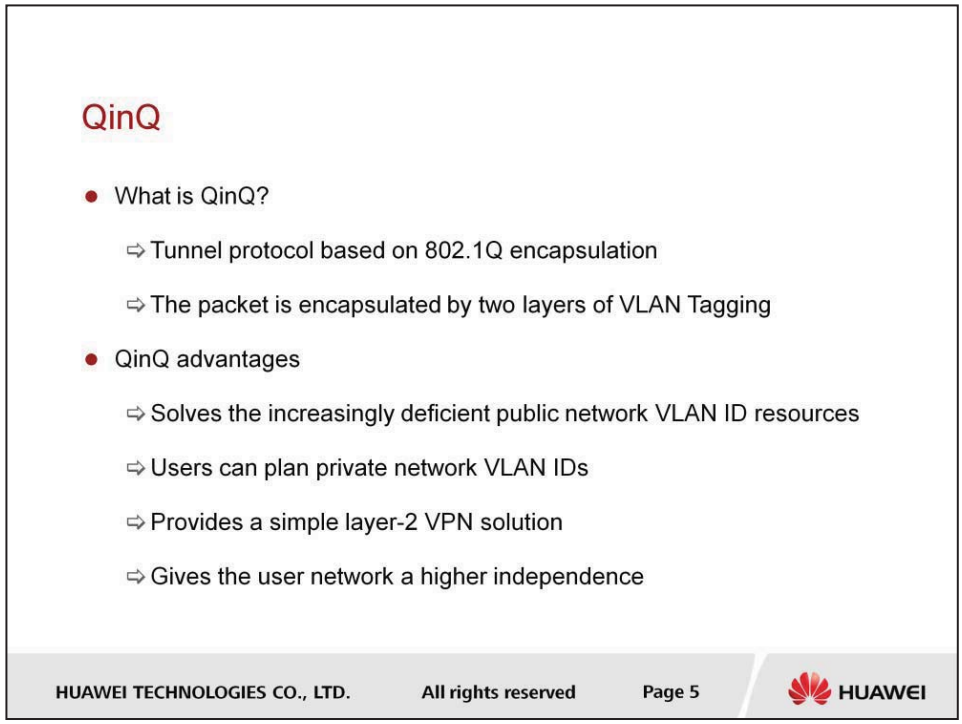
A:Port isolation is an access control security control mechanism between the switch port for control mutual access rights of two physical ports.


QinQ Principles
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
QinQ encapsulates public network VLAN tagging besides private network VLAN tagging. In the public network messages transfer only based on the public network VLAN tag. QinQ provides a kind of simple layer 2 VPN tunnel to the user. This section aims to introduce the basic principles and implementation of QinQ.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this course, you will be able to:
Understand QinQ basic principles and implementation
Master QinQ basic configuration
Master the technical application of QinQ
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Contents
QinQ introduction
QinQ basic principles
QinQ configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 4
HUAWEI



QinQ is a tunnel protocol based on 802.1 Q encapsulation, the main idea is to encapsulate public VLAN tag outside of the private VLAN tag, the packet will go across the public network with two layers of tags, so that it can provide the users with a kind of more simple layer-2 VPN tunnel. QinQ protocol is simple and easy to manage, it does not need signal and can be implemented by static configuration, it is adapt to small-scale enterprise network or MAN.

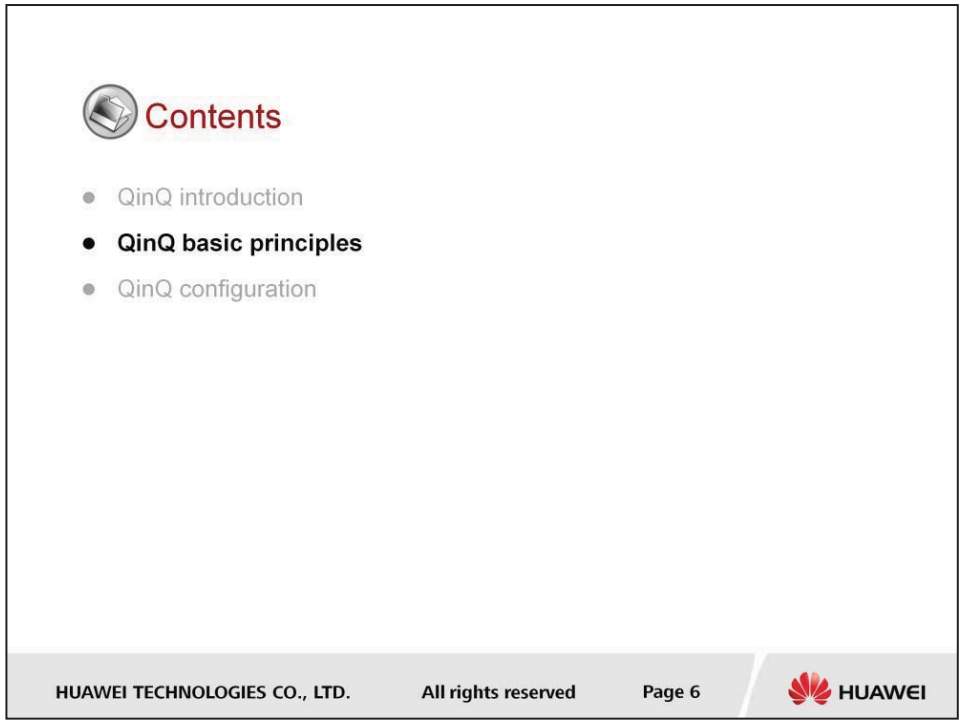
QinQ has the following advantages:

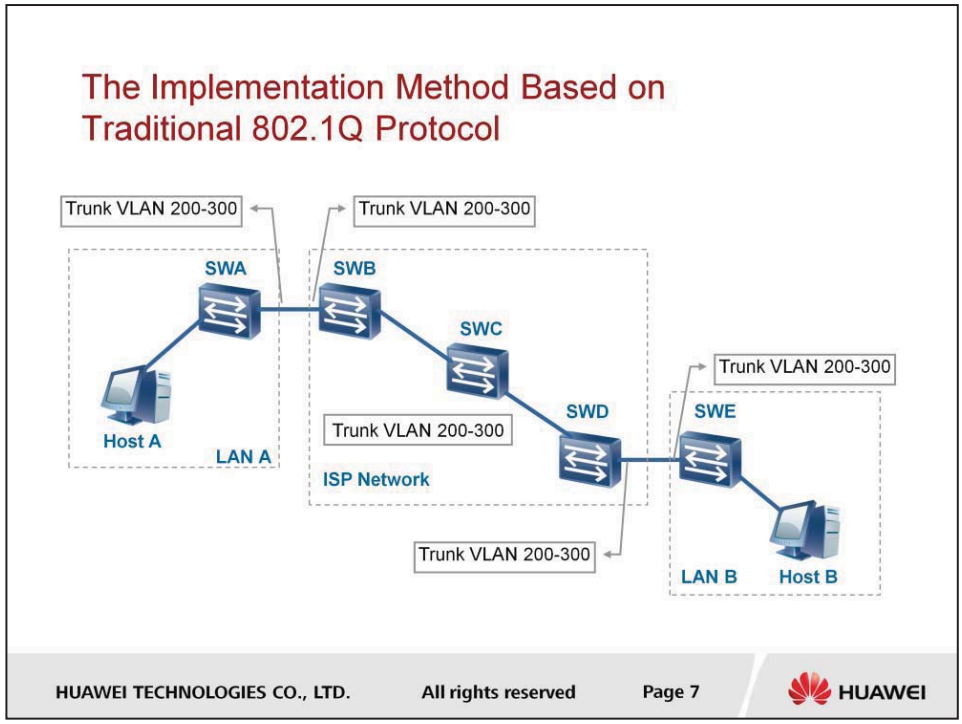
Solves the increasingly deficient public network VLAN ID resources.

Users can plan private network VLAN IDs, avoid collision with public VLAN ID.

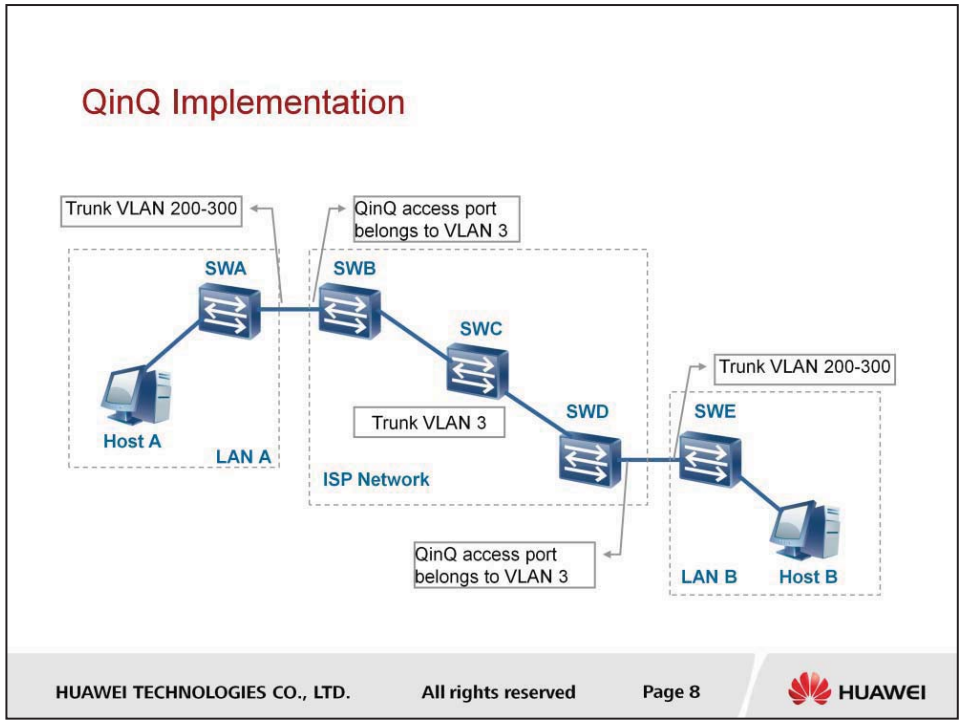
Provides a simple layer-2 VPN solution scheme.

Gives the user network a higher independence, the user network does not need to change the primary configuration when the service provider upgrades the network.

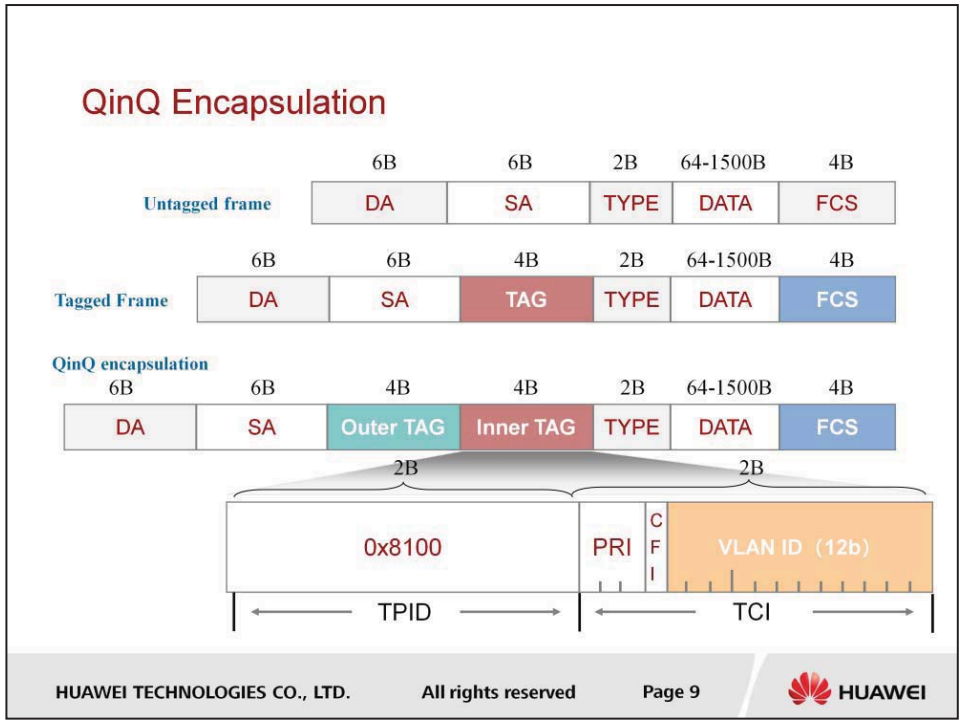

Contents
QinQ introduction
QinQ basic principles
QinQ configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 6
HUAWEI



As shown above, assume that LANA and LANB of the enterprise are located in two sites, they access to ISP Network though SWA and SWB separately. For traditional 802.1Q networks, if users need to connect VLAN200-300 of LANA and VLAN200-300 of LAN B together, then all the connection ports of SWA, SWB, SWC, SWD and SWE should be configured to trunk port, and permit VLAN200-300 to pass. This method makes user VLAN transparent in backbone network, which will waste VLAN ID (4094 VLAN ID resource) and users have no priority to plan VLANs, because the VLAN ID is managed by ISP.



To solve the problem above, QinQ protocol provides the user with unique public VLAN ID, it will encapsulate private VLAN tag inside of public VLAN ID, private VLAN ID is shielded in the public network and the packet is transmitted in the public network according to public VLAN ID, so as to save VLAN ID resources. As shown above, the port of SWA is configured as trunk port which permits VLANs200-300 to pass. The port of SWB that is connected to SWA is configured as an access port for VLAN 3, it does not need to be a trunk port and permits VLAN 200-300 to pass. Enable QinQ function on the port. ISP Network only needs to assign VLAN ID 3 to users, no matter how many VLANs are divided in the enterprise. When the VLAN packets enter ISP Network, they will be forced to insert a public VLAN ID and they will go cross the backbone network with the public VLAN ID. When the packets get to SWD, the public VLAN ID will be peeled off and become user packets again, then it will be transmitted to SWE of LANB. So, the packet in backbone network has two layers of 802.1Q tags: public tag and private tag.



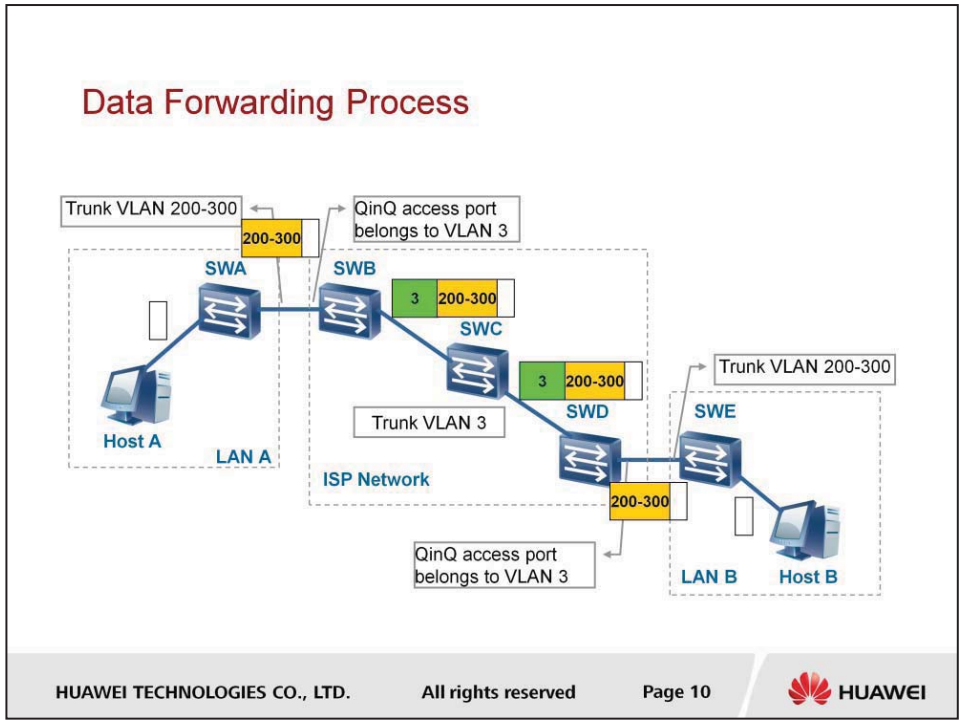
As shown above, ordinary Ethernet frame has no VLAN Tag; in 802.1Q frame, 4B VLAN TAG is inserted between source address and Type filed; For QinQ encapsulation, it inserts a 4B VLAN Tag before 802.1Q VLAN Tag, the two 4B VLAN Tags include the same field. TPID：Tag Protocol Identifier, 2 byte, fixed value, 0x8100, it is a new type defined by IEEE, it indicates that it is a frame with 802.1Q tag. The default TPID value of out tag in HUAWEI switch is 0x8100, some other manufactures set the value as 0x9100 or 0x9200. but the TPID value of port based QinQ packet in HUAWEI switch can be changed.

TCI：Tag Control Information, 2byte.

•Priority：3bit, indicate priority of Ethernet frame. 8 kinds of priority, 0－7,used to provide differential forwarding service.

•CFI：Canonical Format Indicator, 1bit. Used to indicate bit order of address information in token ring or source route FDDI media access, namely, low bit is transmitted before high bit

•VLAN Identified：VLAN ID,12bit, from 0 to 4095.



As shown above, the packet forwarding flow is as followings:

Host A sends standard Ethernet frame to SWA, SWA will add VLAN Tag 200－300 to it, and forward it from Trunk port. SWB receives frame with private VLAN Tag 200－300, because the input port is QinQ access port, SWB does not take care of the private VLAN tag, it will insert public VLAN Tag 3 to the frame. In ISP Network, the packet will be transmitted though Trunk VLAN 3 port, private Tag will not change until it gets to ISP network edge device: SWD. The SWD port connecting SWE is QinQ access port of VLAN 3, SWD will peel off public VLAN Tag 3 and forward the packet with private VLAN Tag from the port to SWE.

When SWE receives the frame with VLAN Tag 200-300 and forwards it to correlative out interface, the out interface will peal off VLAN Tag 200-300 and forward the standard Ethernet frame to host B. From the forwarding flow, we can see that QinQ is very simple, it does not need signal to establish tunnel and it can be implemented by static configuration.

QinQ classification

- QinQ can be classified based on the following:
- Based on port QinQ
  - ⇨ Basic port based QinQ
- Flexible QinQ
  - ⇨ VLAN stacking
- Based on flow flexible QinQ
  - ⇨ Based on ACL flexible QinQ

Based on port QinQ

If this function is configured on the port, equipment will add a VLAN ID for the packets which come from this port as outer layer VLAN tag of the port PVID.

Port based QinQ is achieved through the configuration of the dot1q-tunnel type.

⇨ When port type is dot1q-tunnel, the added VLAN of this port doesn't support layer 2 multicast functionality.

Flexible QinQ

Various conditions can allow flexible QinQ to add the service VLAN (S-VLAN) tagging for incoming packets.

- ⇨ Specified conditions include an incoming packets' outer tag VLAN, or VLAN with 802.1p.

Implemented by configuring the VLAN Stacking on the port.

Strengths:

- ⇨ Compared to port based QinQ, flexible QinQ can choose add S-VLAN tag or not according to the outer VLAN tag and 802.1p, and the S-VLAN tag can be configured.

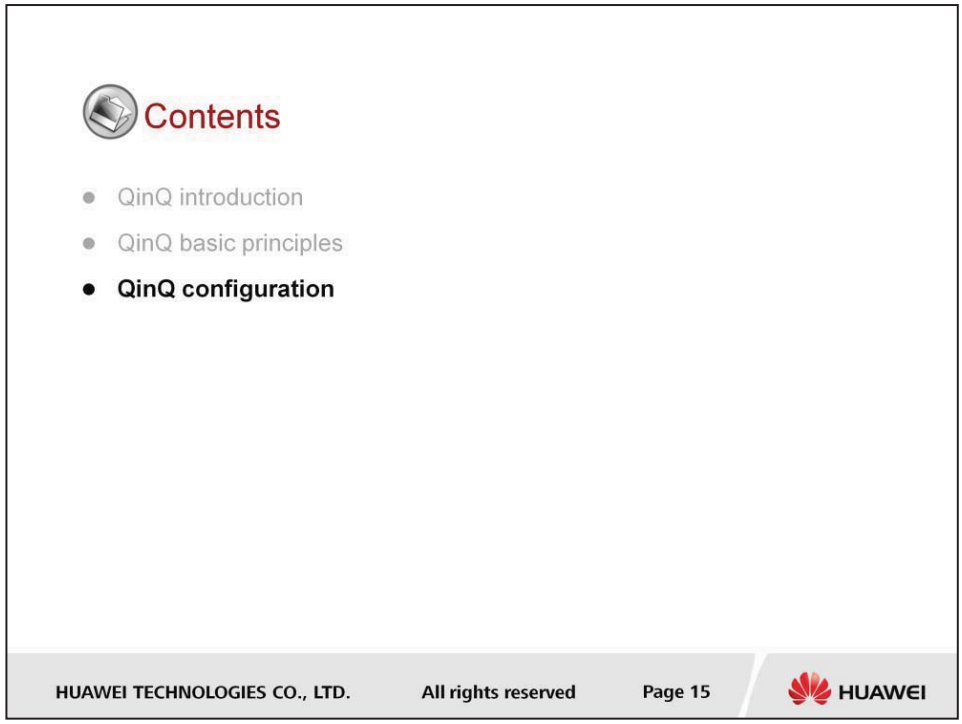
Flow based flexible QinQ

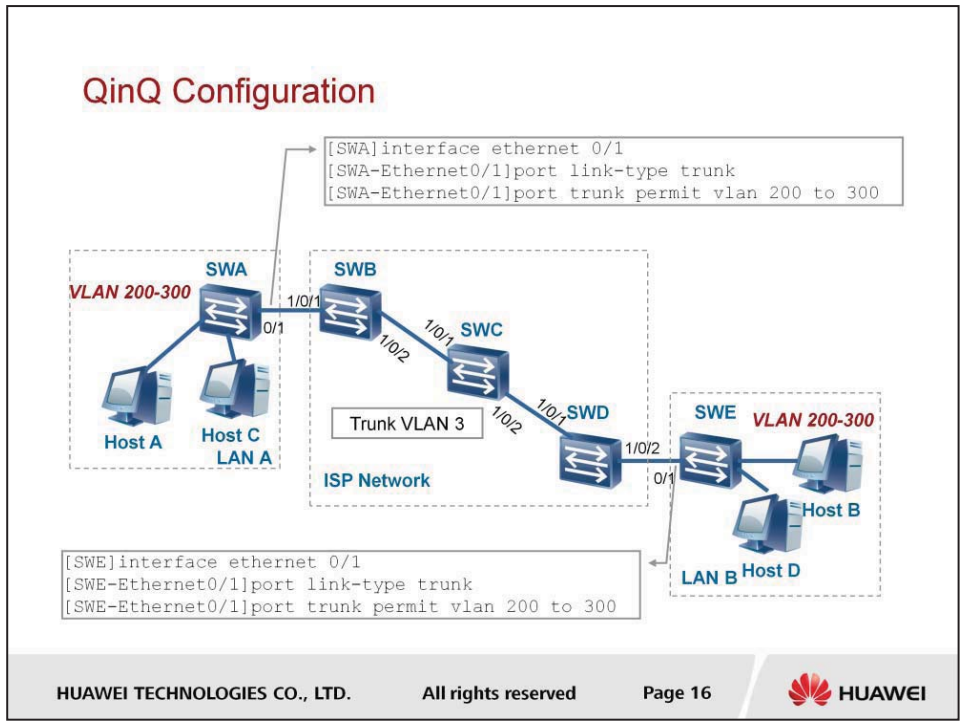
Flow based flexible QinQ is configured by global configuration of flow classification and flow behavior.

Strengths:

- ⇨ Compared to flexible QinQ, flow based flexible QinQ also can add S-VLAN tag according to the packets' inlayer VLAN, Configuration range is more flexible. Such as:
  - − Inlayer VLAN, inlayer VLAN+802.1p, out layer VLAN, out layer+802.1p and so on

To understand flow based flexible QinQ it it necessary to know QoS first, and some basic principles in this course.


Contents
QinQ introduction
QinQ basic principles
QinQ configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 15
HUAWEI



Take this network for example to introduce QinQ basic configuration.

As shown above, SWA in LAN A and SWB in LAN B both have the private network users of VLAN200 to 300, Ethernet1/0/1 of SWB is connected to SWA,

Ethernet1/0/2 of SWD is connected to SWE. Ethernet1/0/1 and Ethernet1/0/2 are QinQ access ports, and belong to VLAN3.

SWA and SWE are switch at user side, upstream Trunk port Ethernet0/1 will send the frame with VLAN ID 200-300 to SWB, the configuration of SWA upstream Trunk port is as followings:

[SWA]interface ethernet 0/1

//enter ethernet 0/1 interface view

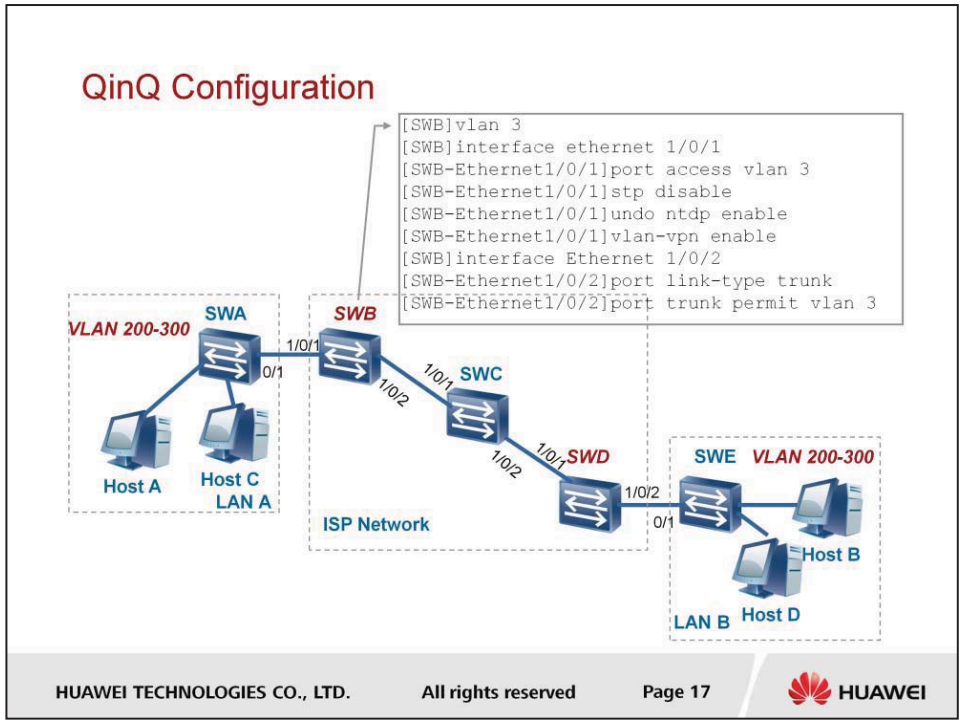
[SWA-Ethernet0/1]port link-type trunk

//configure link type as trunk link

[SWA-Ethernet0/1]port trunk permit vlan 200 to 300

//permit frame with VLAN ID 200 to 300 to pass

The configuration of SWE upstream Trunk port is similar to SWA



Ethernet1/0/1 of SWB is configured as QinQ access port and belongs to VLAN3， Ethernet1/0/2 is configured as Trunk port, the configuration is as followings:

[SWB]vlan 3

[SWB-vlan3]quit

//create VLAN 3 on SWB

[SWB]interface ethernet 1/0/1

[SWB-Ethernet1/0/1]port access vlan 3

//configure Ethernet1/0/1 as Access port and belonging to VLAN3

[SWB-Ethernet1/0/1]stp disable

[SWB-Ethernet1/0/1]undo ntdp enable

//prohibit STP and NTDP protocol. On QinQ access port, GVRP、GMRP、IRF、STP or 802.1x protocol must be disabled, or QinQ can not be enabled.

[SWB-Ethernet1/0/1]vlan-vpn enable

//enable VLAN-VPN feature of port

[SWB-Ethernet1/0/1]quit

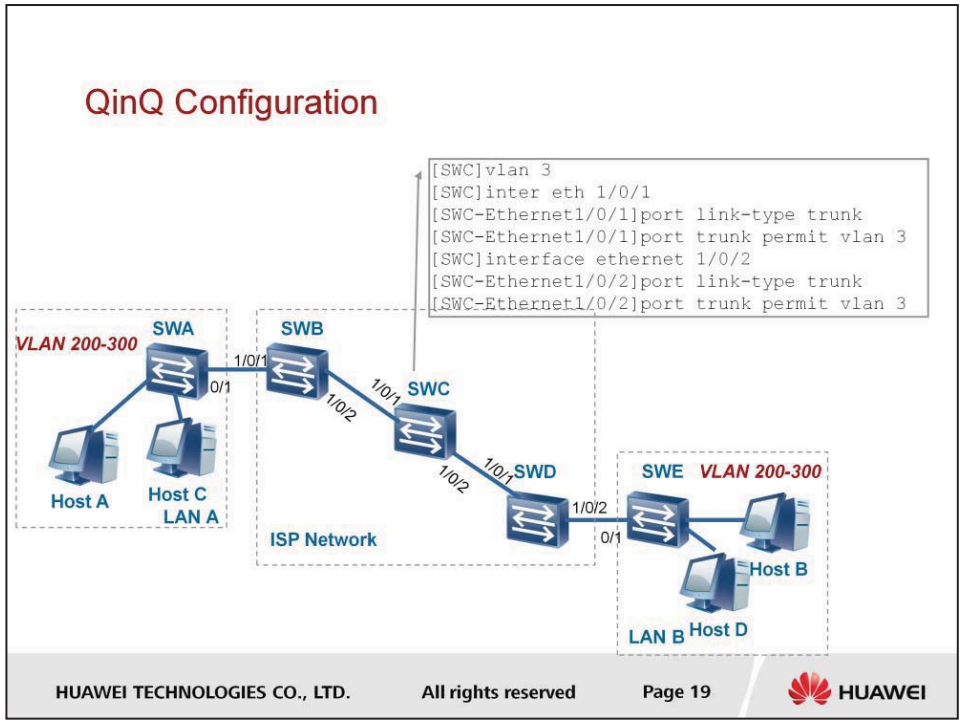
[SWB]interface Ethernet 1/0/2

[SWB-Ethernet1/0/2]port link-type trunk

[SWB-Ethernet1/0/2]port trunk permit vlan 3

//configure Ethernet1/0/2 as Trunk port and permit VLAN 3 to pass

The configuration of SWD is similar to SWB.



SWB can apperceive public VLAN ID, but can not apperceive user private VLAN ID 200-300. On SWB, configure Trunk port and permit frame with VLAN ID 3 to pass transparently. The configuration is as followings:

[SWB]vlan 3

[SWB-vlan3]quit

[SWC]inter eth 1/0/1

[SWC-Ethernet1/0/1]port link-type trunk

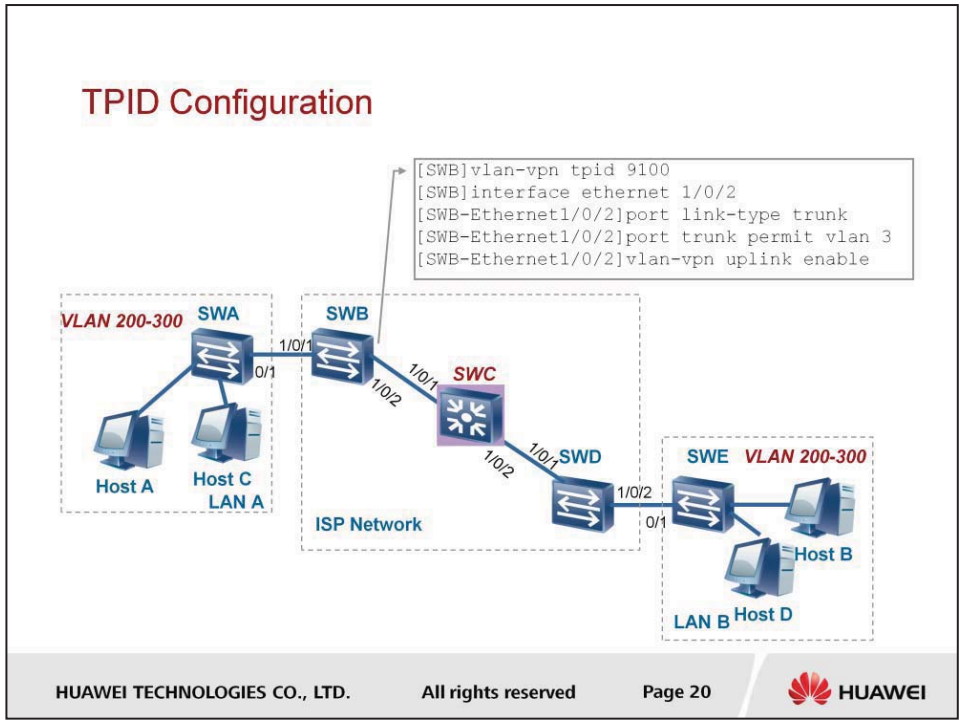
[SWC-Ethernet1/0/1]port trunk permit vlan 3

[SWC-Ethernet1/0/1]quit

[SWC]interface ethernet 1/0/2

[SWC-Ethernet1/0/2]port link-type trunk

[SWC-Ethernet1/0/2]port trunk permit vlan 3



Assume that SWC is made by other company, the TPID value is 0x9100, but for HUAWEI switch, the TPID value in outer Tag of QinQ packet is 0x8100. In this case, we can configure Ethernet1/0/2 of SWB and Ethernet1/0/1 of SWD as VLAN-VPN uplink port, the TPID of VLAN-VPN uplink port can be configured by manager. When the VLAN-VPN Uplink port receives packet, it will replace the TPID value in outer VLAN Tag with a configured value and then forward it, so that SWC can identify the QinQ packet. The configuration is as followings:

[SWB]vlan-vpn tpid 9100

//configure TPID of VLAN-VPN uplink port on SWB as 0x9100

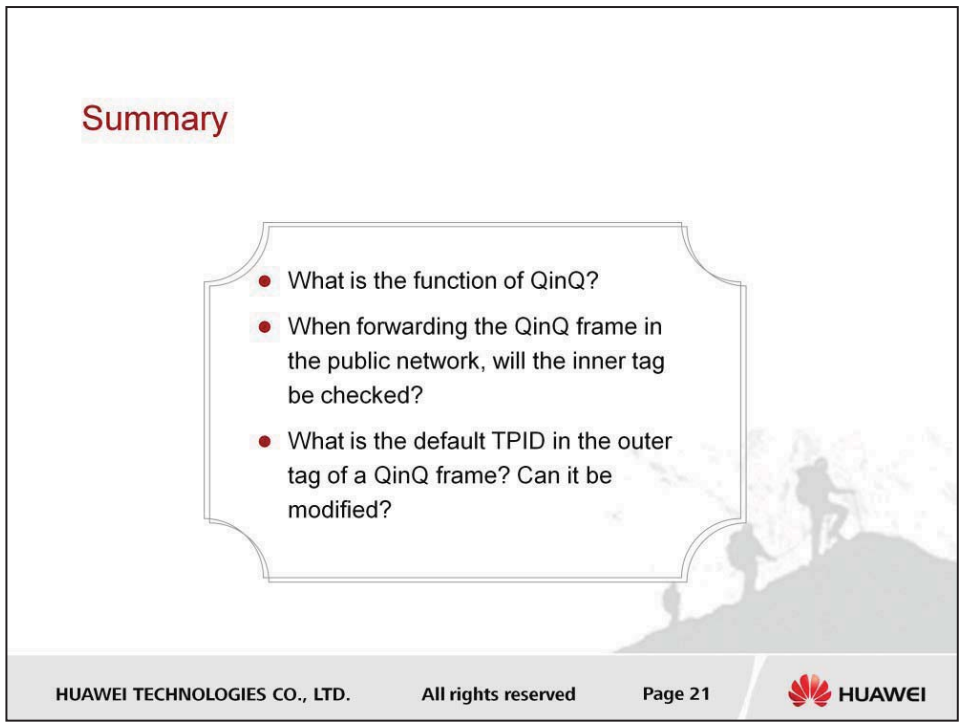
[SWB]interface ethernet 1/0/2

[SWB-Ethernet1/0/2]port link-type trunk

[SWB-Ethernet1/0/2]port trunk permit vlan 3

[SWB-Ethernet1/0/2]vlan-vpn uplink enable

//configure SWBEthernet1/0/2 as VLAN-VPN uplink port



Q：what is the function of QinQ?

A：QinQ can provides users with a simple layer-2 VPN tunnel, it encapsulates

public VLAN Tag outside of the private VLAN Tag.

Q：When forwarding the QinQ packet in the public network, will the inner tag be checked?

A：QinQ packet is transmitted in public network according to the public Tag, the

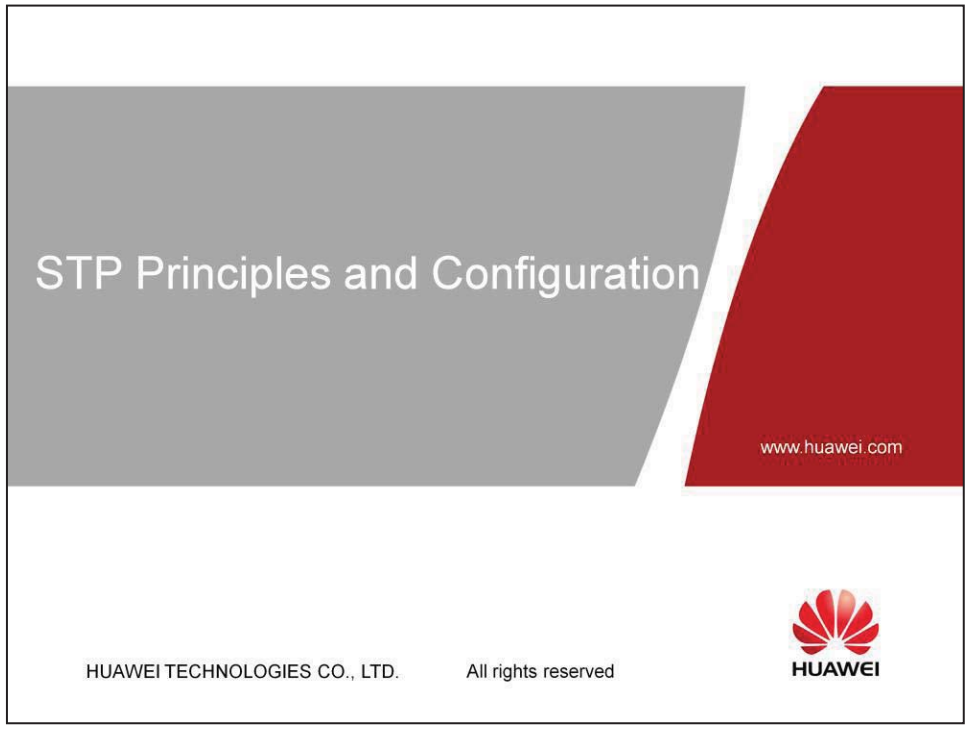
network will not check the private Tag.

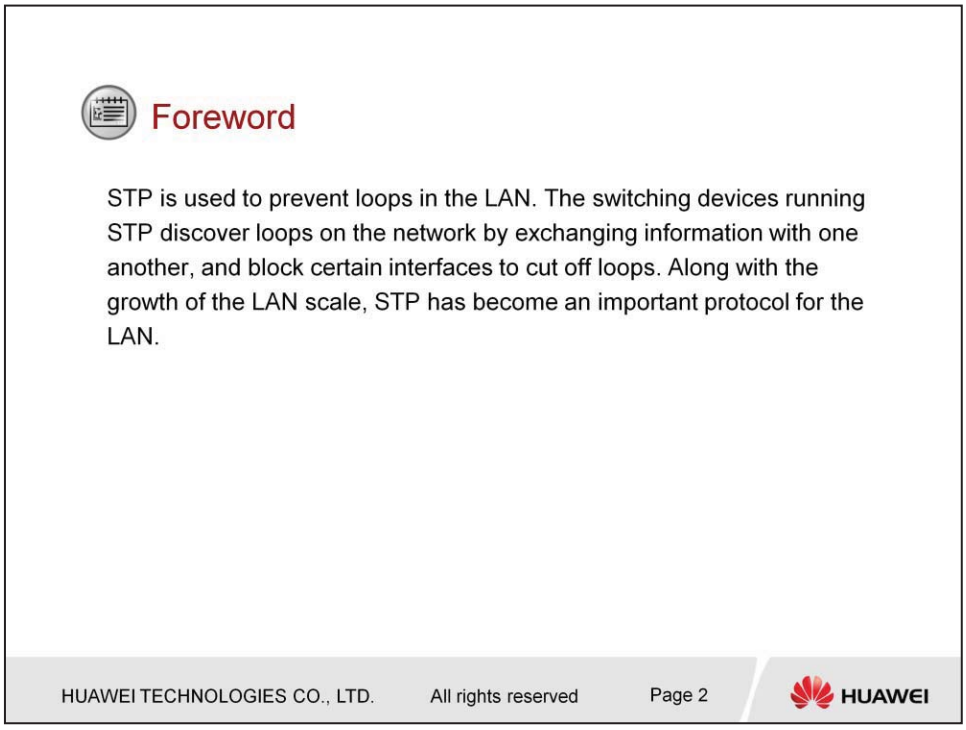
Q：what is the default TPID value of QinQ frame outer tag? Can it be modified?

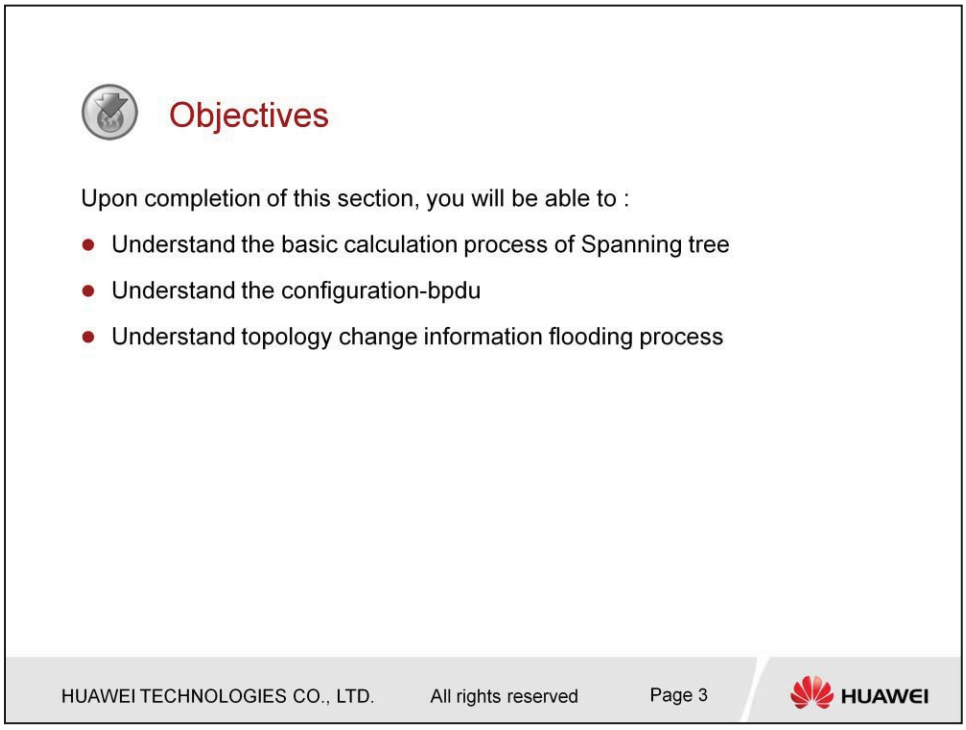
A： the default value of TPID 0x8100, command “qinq protocol” can be used to

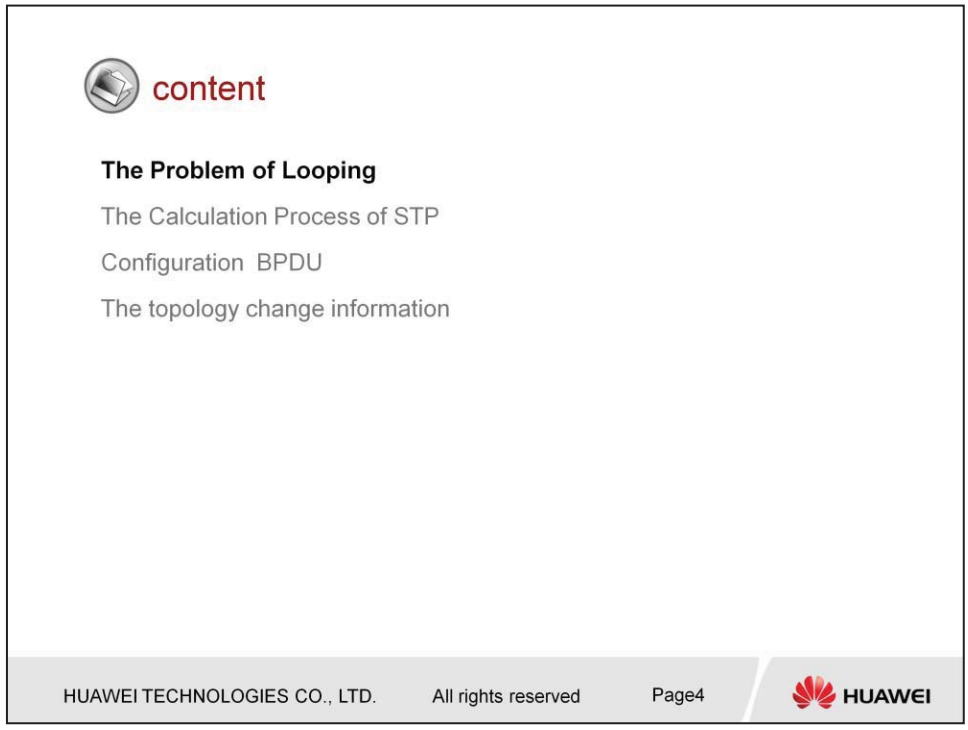
change it.

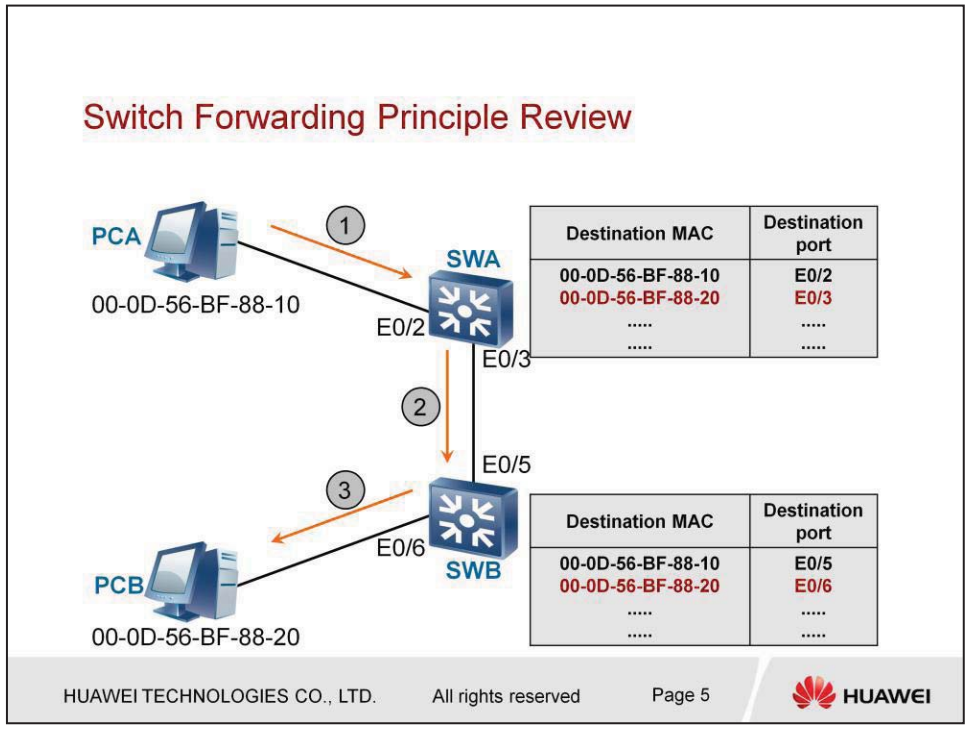
Module 2
STP


STP Principles and Configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
STP is used to prevent loops in the LAN. The switching devices running STP discover loops on the network by exchanging information with one another, and block certain interfaces to cut off loops. Along with the growth of the LAN scale, STP has become an important protocol for the LAN.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 2
HUAWEI


Objectives
Upon completion of this section, you will be able to :
Understand the basic calculation process of Spanning tree
Understand the configuration-bpdu
Understand topology change information flooding process
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


content
The Problem of Looping
The Calculation Process of STP
Configuration BPDU
The topology change information
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page4
HUAWEI



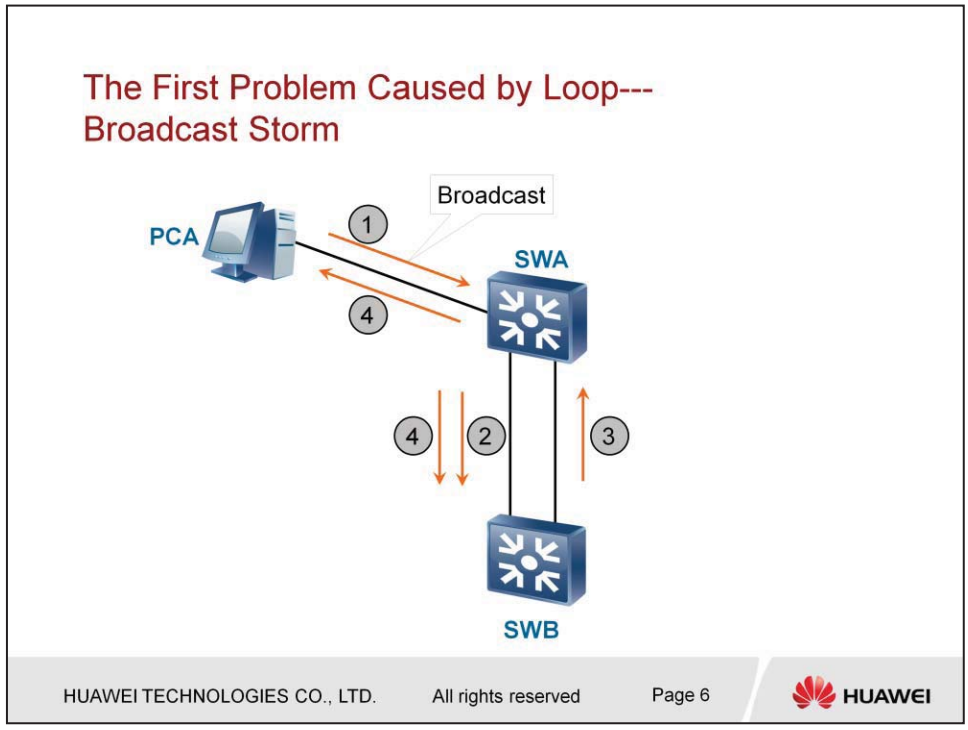
A switch forwards data frames based on the MAC address table. The MAC address table specifies the mapping between destination MAC addresses and destination ports.

1: Assume that PCA sends a data frame to PCB. The MAC address of this data frame is set to the MAC address of PCB, namely, 00-0D-56-BF-88-20. When SWA receives this frame, it searches the MAC address table. According to the entries in the MAC address table, SWA forwards the data frame through port E0/3.

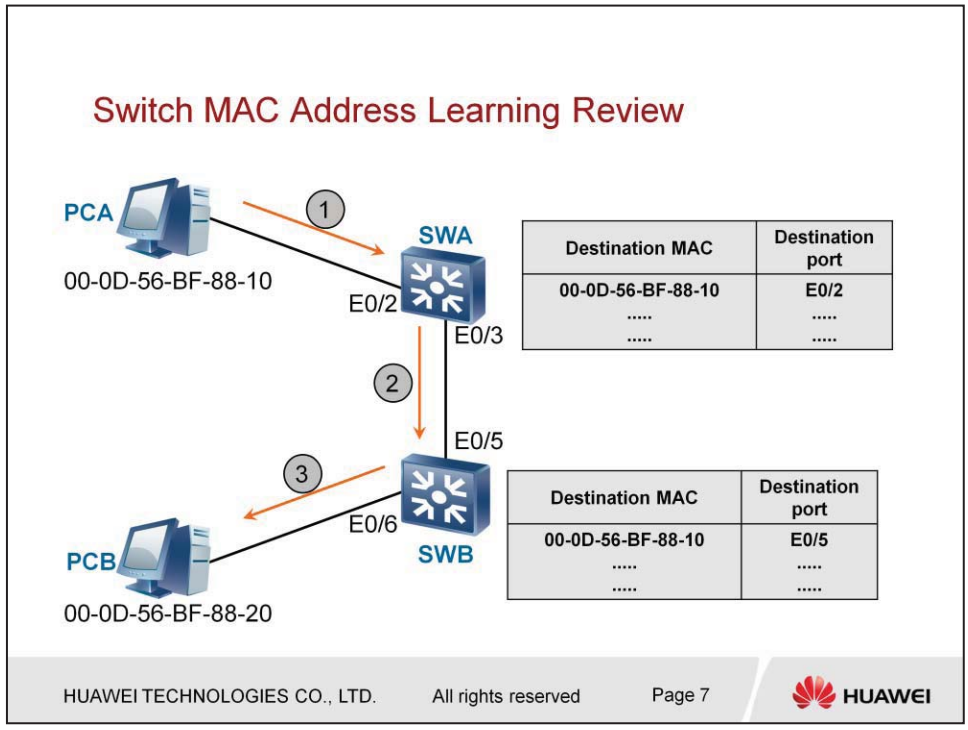
The switch does not make any modification to the data frame before forwarding it. If the switch receives a broadcast frame, it forwards the frame to all ports.

2: When SWB it searches the MAC address table. According to the entries in the MAC address table, SWB forwards this frame through port E0/6. SWB does not make any modification to the data frame.

3: PCB receives this frame, it searches the MAC address table and finds that the destination MAC address is the MAC address of its own. Then PCB processes this data frame and forwards the data generated when the upper layer protocol processes the data frame to the switch.



If a switch receives a broadcast data frame from a port, the switch forwards the data frame to all other ports. In addition, it does not make any modification to the data frame before forwarding it. Therefore, if a loop exists in the network, the broadcast frames are forwarded in the network infinitely, thus causing the broadcast storm.

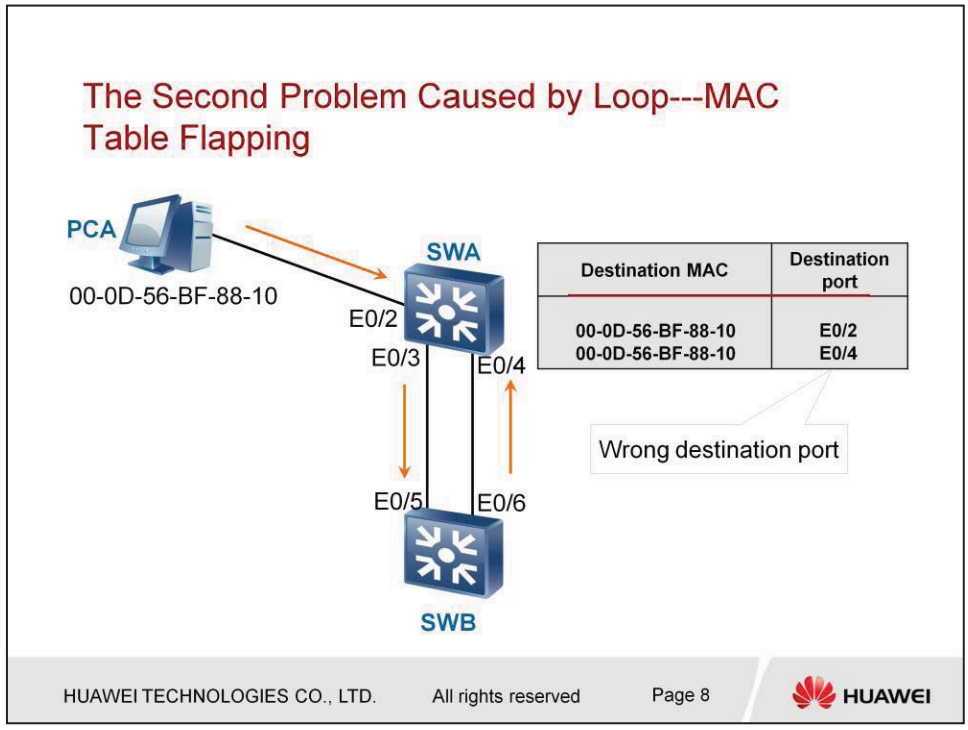


A switch forwards data frames based on the MAC address table, but the MAC address table is empty when the switch is started. Therefore, the switch needs to learn the MAC address table. A switch learns the MAC address table based on the mapping between the source address of the received data frame and the receiving port.

1: Assume that PCA sends a data frame to PCB. The destination address of the frame is the MAC address of PCB, namely, 00-0D-56-BF-88-20. The source address is the MAC address of PCA, namely 00-0D-56-BF-88-10. When SWA receives the data frame, it checks the source address of the frame and adds mapping between the source address and receiving port to the MAC address table. Thus, the mapping between the destination address and destination port is recorded in the table.

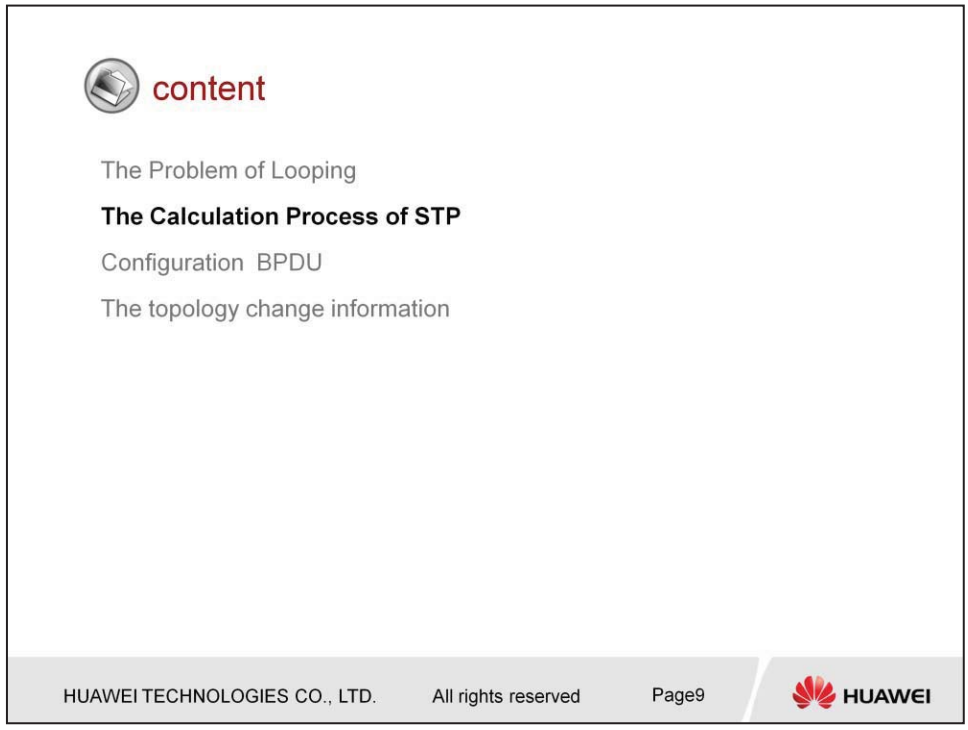
2: When SWB receives this frame, it also adds the mapping between the source address and receiving port to the MAC address table as a MAC address entry.

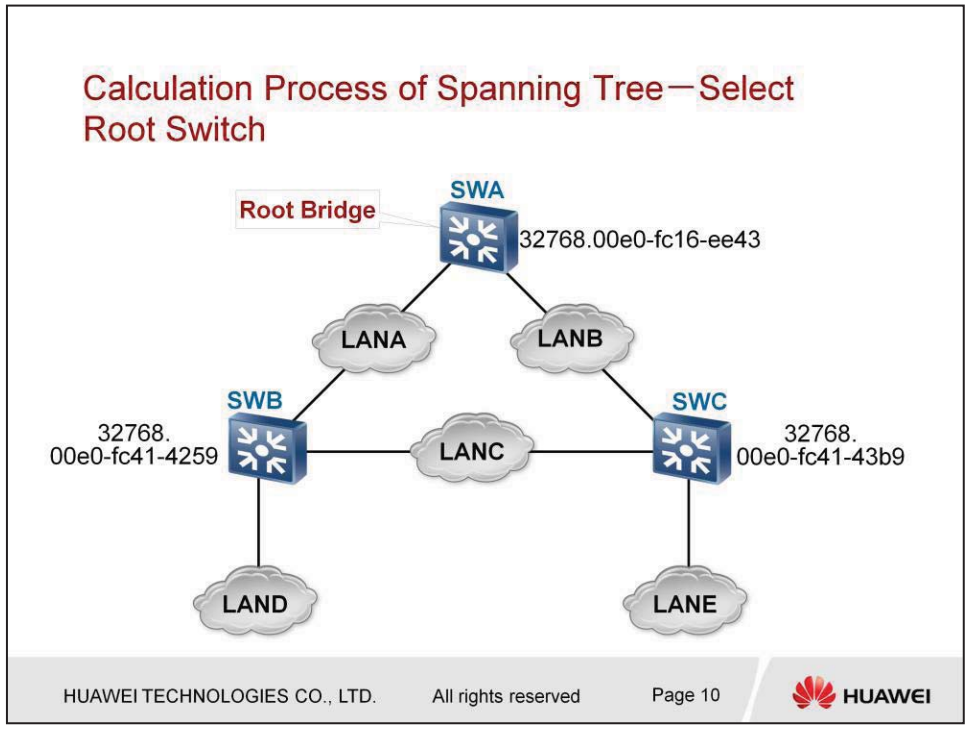
3: When PCB receives the frame, it processes this frame.



A switch generates a MAC address entry according to the source address and receiving port of the received data frame.

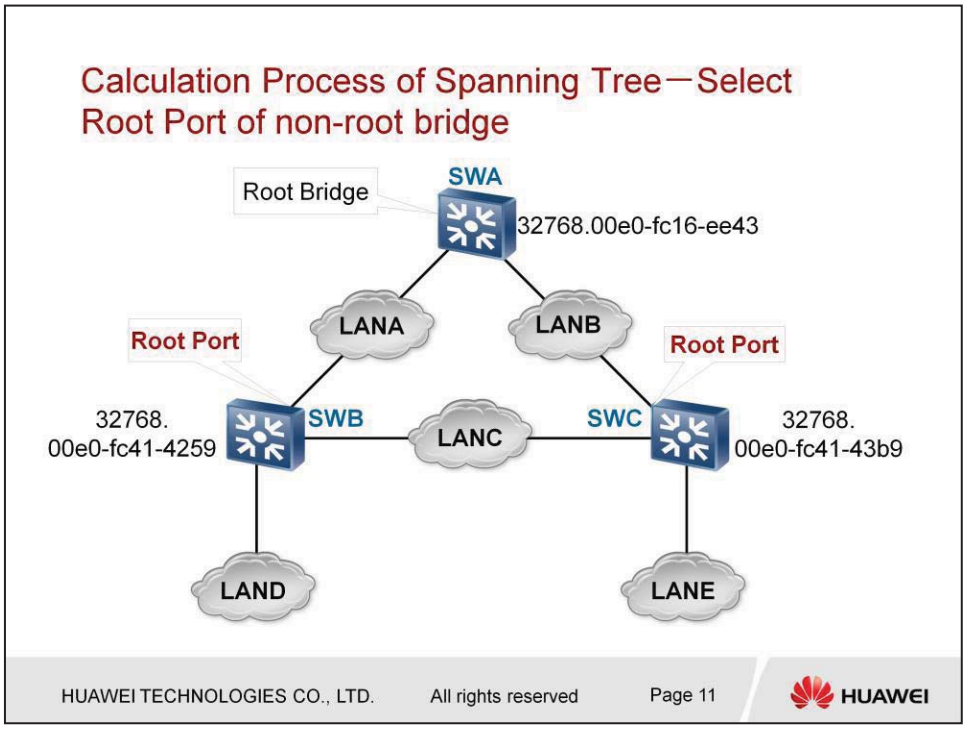
PCA sends a data frame. Assume that the destination MAC address of the data frame does not exist in any MAC address table of the switches in the network. When SWA receives this data frame, it generates a MAC address entry, in which the MAC address 00-0D-56-BF-88-10 maps port E0/2. Because the MAC address table of SWA does not contain any entry with this destination MAC address, SWA forwards the data frame to E0/3 and E0/4. The MAC address table of SWB also does not contain any entry with this destination MAC address. So, after SWB receives the data frame on E0/5, it forwards the frame to SWA through E0/6. After SWA receives this data frame on E0/4, it deletes the previous entry with this address and generates a new entry. In the new entry, MAC address 00-0D-56-BF-88-10 maps port E0/4. In this case, the MAC address table is unstable and wrong entries are generated.


content
The Problem of Looping
The Calculation Process of STP
Configuration BPDU
The topology change information
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page9
HUAWEI



To calculate the spanning tree, switches need to exchange information and parameters. The information and parameters are encapsulated in the Configuration Bridge Protocol Data Unit (BPDU) and transmitted between switches.

In a broad sense, a BPDU refers to a data unit used to exchange information between switches. The configuration BPDU is a type of the BPDU. Calculation of the spanning tree starts from election of the root bridge. The root bridge is elected based on the bridge identifier. A bridge identifier consists of a 2-byte bridge priority and a 6-byte MAC address. The bridge priority is configurable. The value ranges from 0 to 65535 and the default value is 32768. In the network, the switch with the smallest identifier becomes the root bridge. The system first compares the priority. If the switches have the same priority, the system compares their MAC addresses. The switch with the smallest MAC address is elected first. In this example, the three switches have the same priority. SWA has the smallest MAC address, so SWA is elected as the root bridge.



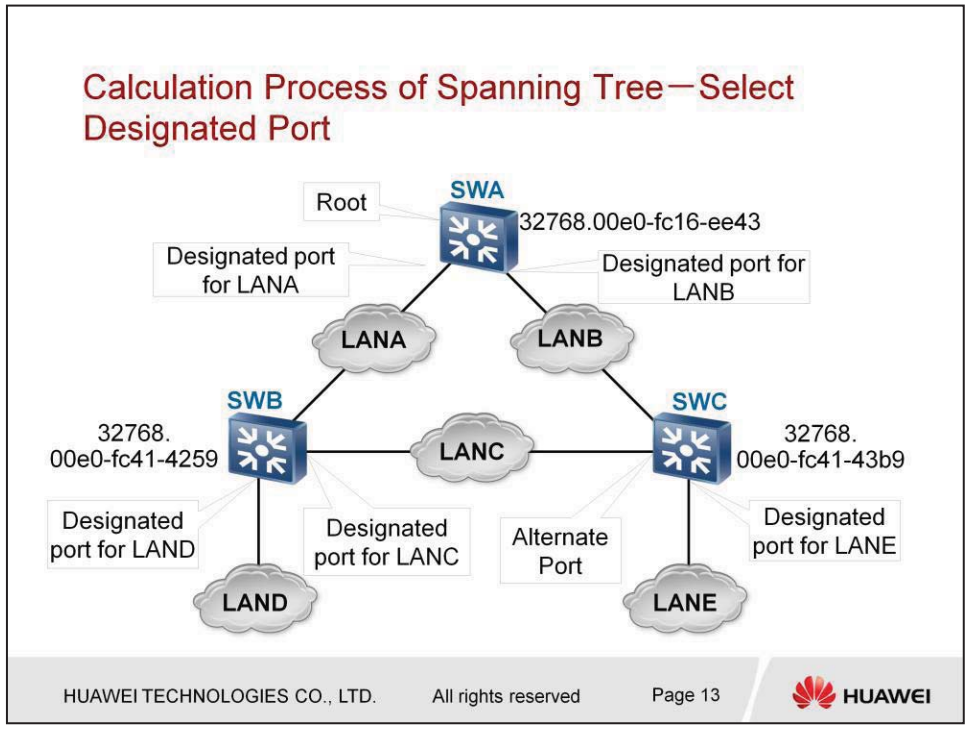
STP elects a root port for each non-root bridge. Each port of a switch has a port cost parameter. The port cost refers the cost for sending the data from this port, namely the cost of the outgoing port. STP considers that no cost is needed for receiving the data on a port. The port cost depends on the bandwidth of the port. The higher the bandwidth is, the smaller the port cost will be. On the VRP, the cost of a100 M port is 200. Multiple paths may exist between a non-root bridge and a root bridge. The cost of a path is the total cost of all outgoing ports on this path. Multiple paths may exist between a non-root bridge and a root bridge. The cost of a path is the total cost of all outgoing ports on this path.

A root port is a local port on the path with the least cost from a non-root bridge to the root bridge. The cost of this path is referred to as the root path cost. If multiple root ports exist, the system compares the identifiers of the upstream switches. The port whose upstream switch has the smallest identifier is elected. If the upstream switches have the same bridge identifier, the system compares the identifier of the upstream ports. The port whose upstream port has the smallest identifier is elected.

The port identifier consists of a 1-byte port priority and a 1-byte

port number. The port priority is configurable. The default value is 128.

In this example, we assume that all ports are 100 M ports and their cost values are all 200.



STP elects the designated port for each network segment. The designated port forwards the data transmitted between the root bridge and this network segment. The switch where the designated port is located is called the designated switch.

When electing the designated port and designated bridge for a network segment, STP compares the root path cost of the switch on which the port is connected to this network segment. If the switches have the same root path cost, RSTP compares their bridge identifiers. The port on the switch with the smallest identifier has the highest priority. If their identifiers are also the same, STP compares the identifiers of the ports connected to the network segment. The port with the smallest identifier has the highest priority.

On the root bridge, all ports are the designated ports of the connected network segments. Therefore, the designate ports of LANA and LANB are both on SWA. LAND and LANE are both connected to the port of only one switch, and the

connected ports are designated port for LAND and LANE respectively. LANC is connected to the ports of two switches and the two switches have the same root path cost. Therefore, the identifiers of the switches are compared.

SWB has a smaller identifier (Priority consistent with the two switches, but the MAC address of SWB smaller), so the designated port for LANC is on SWB.

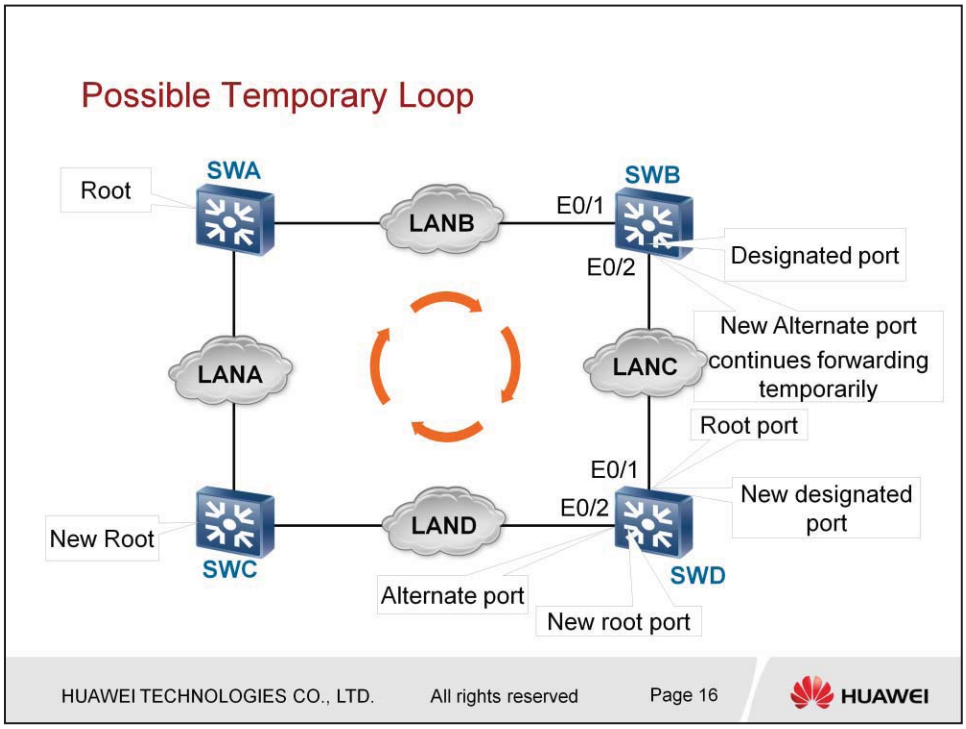
The port that is neither the root port nor the designated port is called the alternate port. The alternate port does not forwards data and is in Blocking state.

Switch Port Role

Port role	Description
Root Port	The root port is the nearest port to the root switch, it is in forwarding status.
Designated Port	Responsible for forwarding data to the downstream network segments or switches
Alternate Port	Does not forward any data to the network segment on which it is connected

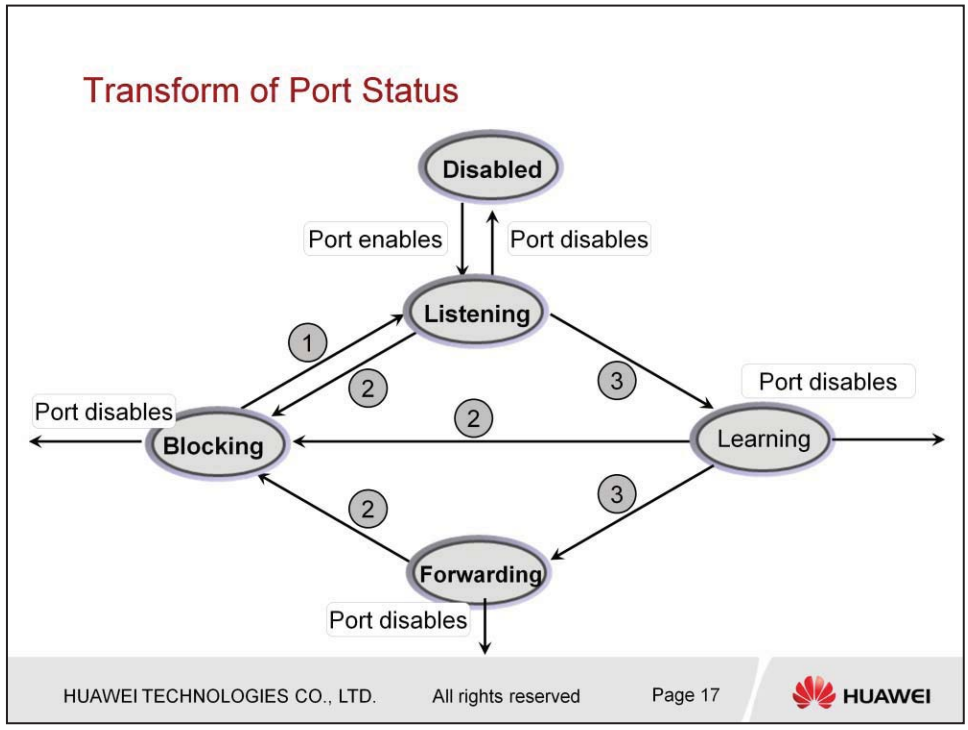
  HUAWEI

STP defines three roles for the STP-enabled port that works normally on the physical layer and data link layer. The root port and designated port are in Forwarding state. The port that is not enabled is called the Disable port.



When the port role and status changes, temporary loops may be formed. In this example, SWA is the root bridge initially. Among all switches, only SWD has an alternate port E0/2 and the port is in non-Forwarding status. Assume that the priority of SWC is changed so that SWC becomes the new root switch. In this case, E0/2 of SWD will become the new root port and switch to Forwarding state. E0/1 of SWD will become the new designated port and switch to Forwarding state. E0/2 of SWB should become the new alternate port and switch to non-forwarding state.

If E0/2 of SWD switches from non-Forwarding state to Forwarding state before E0/2 of SWB switches from Forwarding state to non-Forwarding state, temporary loop is formed in the network. To avoid temporary loops, a port (for example, E0/1 of SWC) must wait enough time before switching from non-Forwarding state to Forwarding state. Therefore, the ports that need to switch to non-Forwarding state have enough time to calculate the spanning tree and switch to non-Forwarding state.



1: The port is elected as the designated port or root port.

2: The port is elected as the alternate port.

3: The port waits a period of the forward delay. By default, the forward delay is 15 seconds.

When a port is disabled, it switches to Disable state.

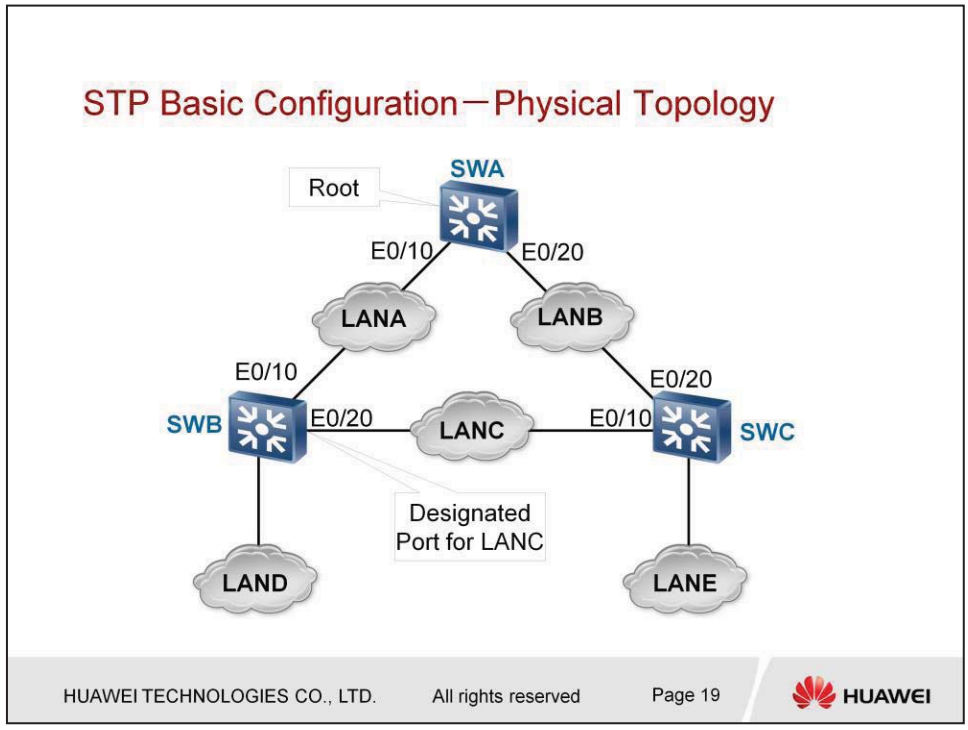
Before switching from non-Forwarding state to Forwarding sate, a port needs to wait two times as along as the forward delay (transition of port status will be described later). Thus, the potential risk of temporary loop is avoided.

Port Status Description

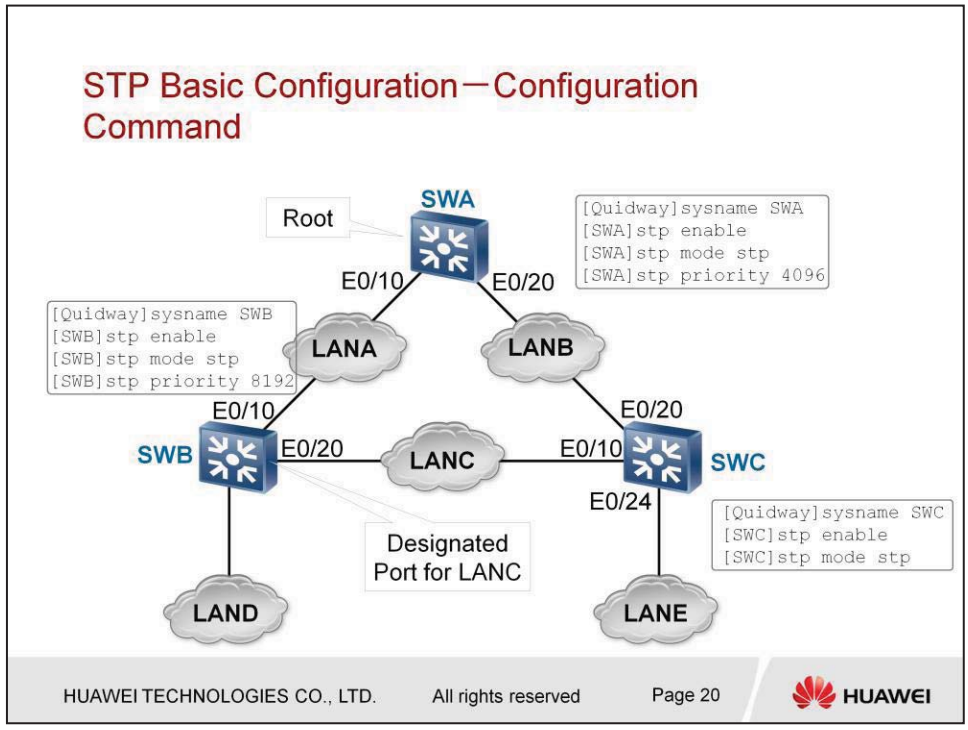
Port status	Description
Disabled	Port will not forward data, learn MAC addresses or calculate spanning tree
Blocking	Port will not forward data or learn MAC addresses; it will receive and process BPDUs but not send BPDUs.
Listening	Port will not forward data and learn MAC addresses, but it will calculate spanning tree, send and receive BPDUs
Learning	Port will not forward data, but it will learn MAC addresses, calculate spanning tree and send and receive BPDUs.
Forwarding	Port will forward data, learn MAC addresses, calculate spanning tree as well as send and receive BPDUs.

  HUAWEI

After enabled, a port switches to Listening state and begins to calculate the spanning tree. After the calculation, if the port is set to the alternate port, the port status changes to Blocking. If the port is set to the root port or designated port, the port status switches from Listening to Learning after a period of forward delay. After another period of forward delay, the port status switches from Learning to Forwarding, and the port can forward data frames.



This figure shows the physical topology. The priority of SWA is 4096; the priority of SWB is 8192; the priority of SWC is 32678. Therefore, SWA becomes the root bridge and SWB becomes the designated switch of LANC.



stp { enable | disable }

The stp command is used to enable or disable STP on a switch or on a port. By default, STP is disabled on the switch.

stp mode { stp | rstp | mstp }

The stp mode command is used to set the STP working mode on a switch. By default, the working mode of the switch is MSTP.

RTSP and MSTP will be described in later courses. This course only describe STP.

stp priority priority

priority: specifies the priority of a switch. The value ranges from 0 to 61440, with the step of 4096. That is, 16 priority values are available for a switch, for example, 0, 4096, 8192, and so on. The stp priority command is used to set the bridge priorities. By default, the bridge priority is 32768.

STP Basic Configuration—Verify STP Global Status

```
[SWC]display stp
 Protocol mode:  IEEE compatible STP
 "bridge ID  (Pri.MAC) :  32768.00e0-fc41-43b9
 The bridge times:  Hello Time 2 sec, Max Age 20 sec, Forward Delay 15
sec
 Root bridge ID (Pri.MAC) :  4096.00e0-fc41-4259
 Root path cost:  200
 Bridge bpdu-protection:  disabled
 Timeout factor:  3
```

  HUAWEI

In the global information, the root bridge identifier is different from the identifier of this switch. It indicates that the switch is a non-root switch.

STP Basic Configuration—Verify STP Port Information

```
[SWC]display stp interface Ethernet 0/20
 Port 20  (Ethernet0/20)  of bridge is Forwarding
   Port spanning tree protocol:  enabled
   Port role:  Root Port
   Port path cost:  200
   Port priority:  128
   Designated bridge ID (Pri.MAC) :  4096.00e0-fc41-4259
   The Port is a non-edged port
   Connected to a point-to-point LAN segment
   Maximum transmission limit is 3  Packets / hello time
   Times:  Hello Time 2 sec,    Max Age 20 sec
           Forward Delay 15 sec, Message Age 0
   BPDU sent:      4
           TCN:  2, RST:  2, Config BPDU:  0
   BPDU received:  806
           TCN:  0, RST:  11, Config BPDU:  795
```

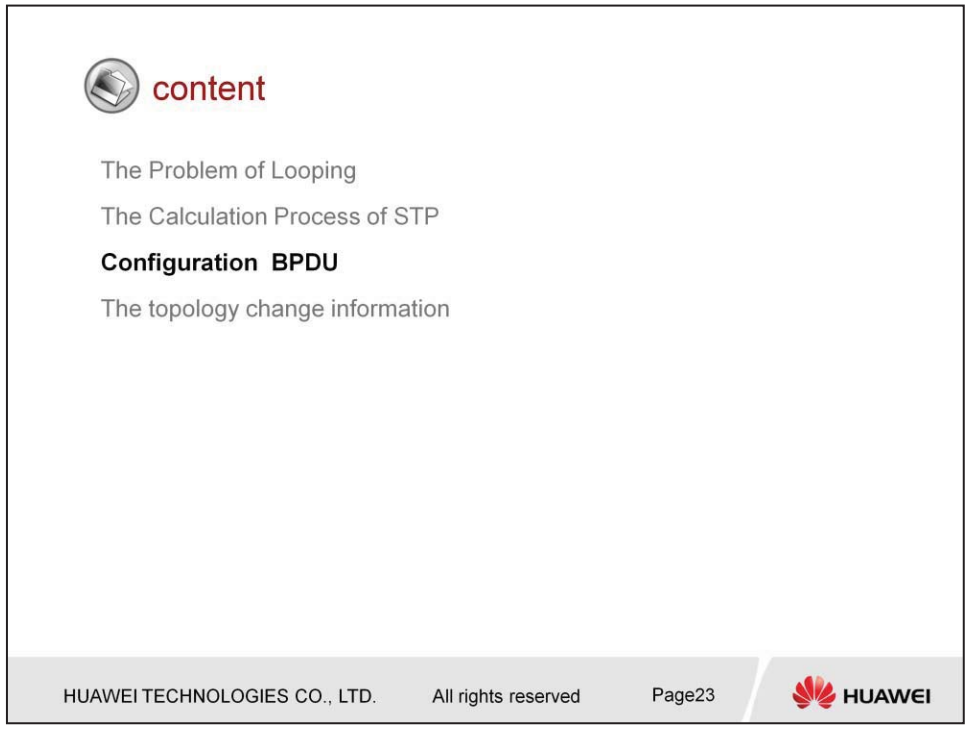

The information about the STP port indicates that:

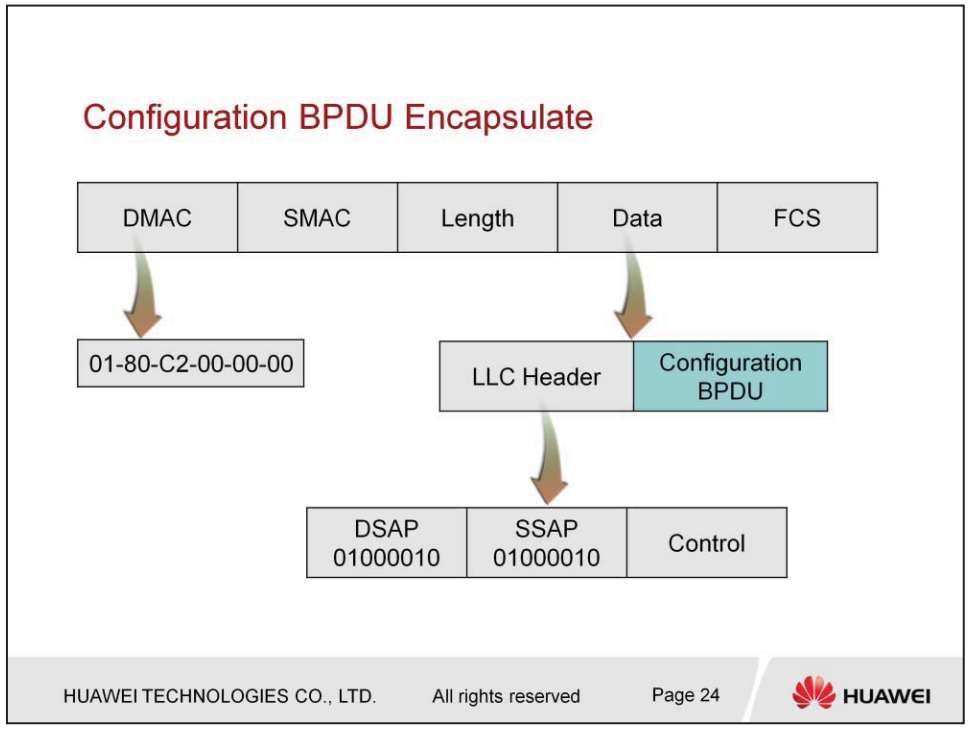
The port status is Forwarding.

The port is the root port.

The default port priority is 128.

The identifier of designated port for the network segment connected to this port is 4096.00e0-fc41-4259, which identifies SWA.


content
The Problem of Looping
The Calculation Process of STP
Configuration BPDU
The topology change information
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page23
HUAWEI



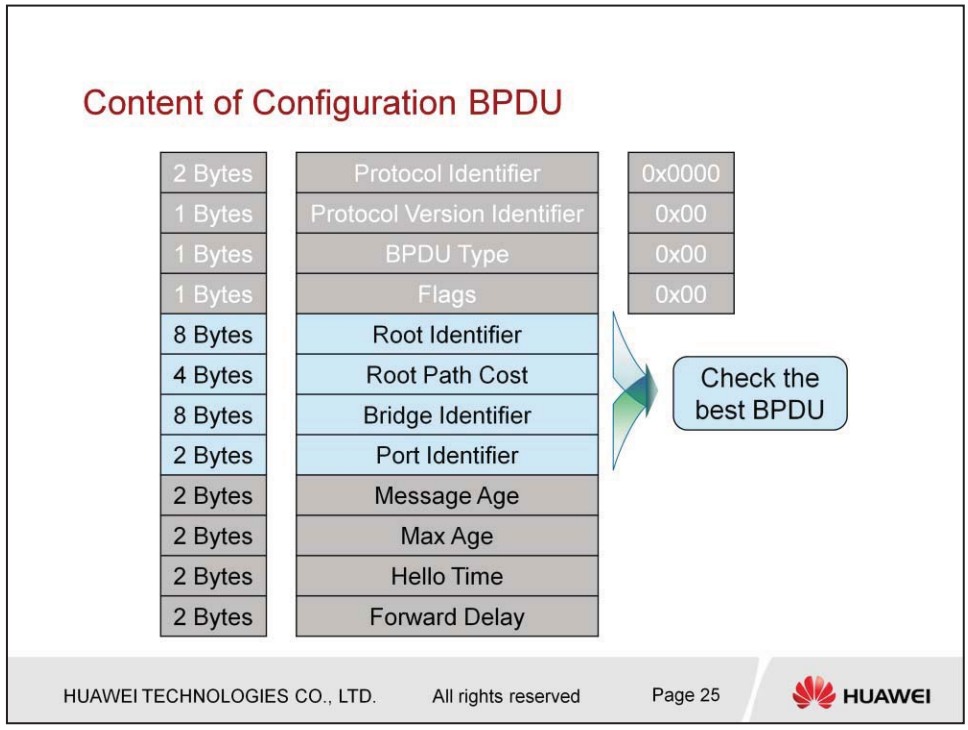
The information and parameters for calculating the spanning tree is encapsulated in the Configuration Bridge Protocol Data Unit (BPDU) and set between switches. The configuration BPDU is encapsulated in the Ethernet frame in the standard LLC format.

The configuration BPDU is sent only by the designated port.

DMAC: destination MAC address.

The Ethernet frame used to encapsulate the RST BPDU uses the reserved group MAC address 01-80-C2-00-00-00. This MAC address identifies all switches but cannot be forwarded by the switches. That is, this MAC address is valid only on the local link.

LLC Header: the LLC Service Access Point (SAP) used by the RST BPDU is a binary value 01000010.



When the configuration BPDU is used to calculate the spanning tree but is not used to transmit the topology change message (will be described in Chapter 4), the fields in the configuration BPDU are as follows:

The Protocol Identifier, Protocol Version Identifier, BPDU Type, and Flags fields are set to all 0s.

The Root Identifier, Root Path Cost, Bridge Identifier, and Port Identifier fields are used to detect the configuration BPDU with the highest priority and calculate the spanning tree.

The value of the Message Age field increases with time. The default value of the Max Age field is 20 seconds. If the value of Message Age reaches the Max Age, the configuration BPDU is regarded as expired.

The default value of the Hello Time field is 2 seconds. That is, the BPDU is sent every two seconds.

The default value of the Forward Delay field is 15 seconds.

Parameters in Configuration BPDU

Field	Description
Protocol Identifier	Always 0
Protocol Version Identifier	Always 0
BPDU Type	Indicates the type of a BPDU. The value is one of the following: 0x00: configuration BPDU 0x80: TCN BPDU
Flags	Indicates whether the network topology is changed. The rightmost bit is the Topology Change (TC) flag. The leftmost bit is the Topology Change Acknowledgement (TCA) flag.

Parameters in Configuration BPDU

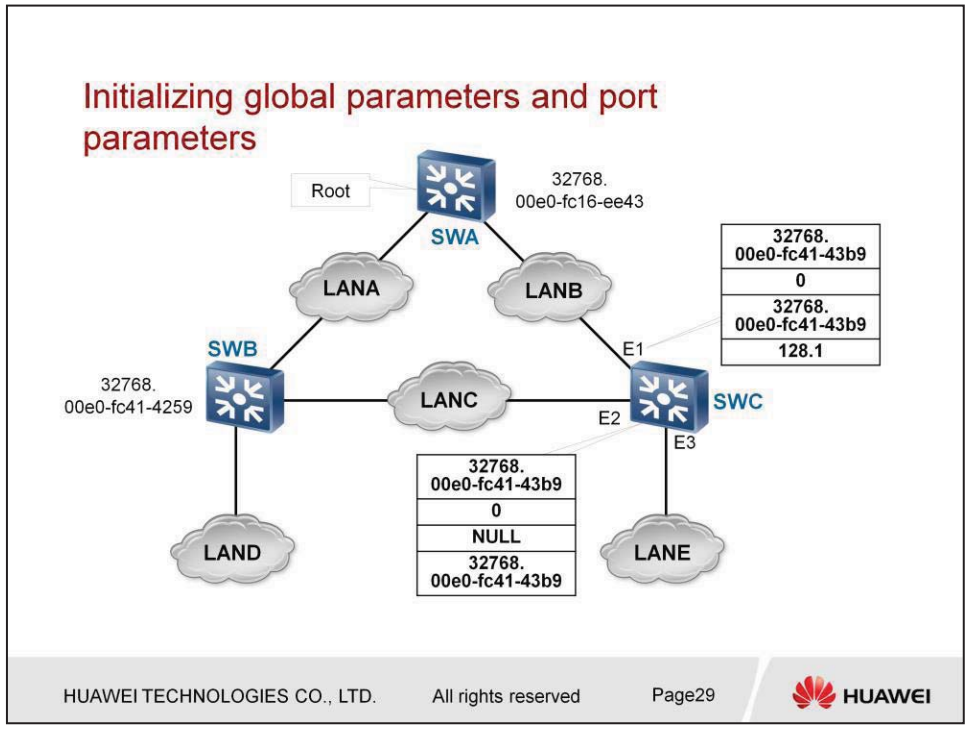
Parameters	Description
Root Identifier	The identifier of the switch which is taken as the root by the current switch sends the configuration BPDU.
Root Path Cost	The total cost of the shortest path from the switch that sends the configuration BPDU to root switch, it includes the cost of root port, but the cost of port sending configuration BPDU is not included.
Bridge Identifier	The identifier of the switch that sends the configuration BPDU
Port Identifier	The identifier of the port that sends the configuration BPDU

  HUAWEI

This table lists the parameters for detecting the configuration BPDU with the highest priority and provides the description.

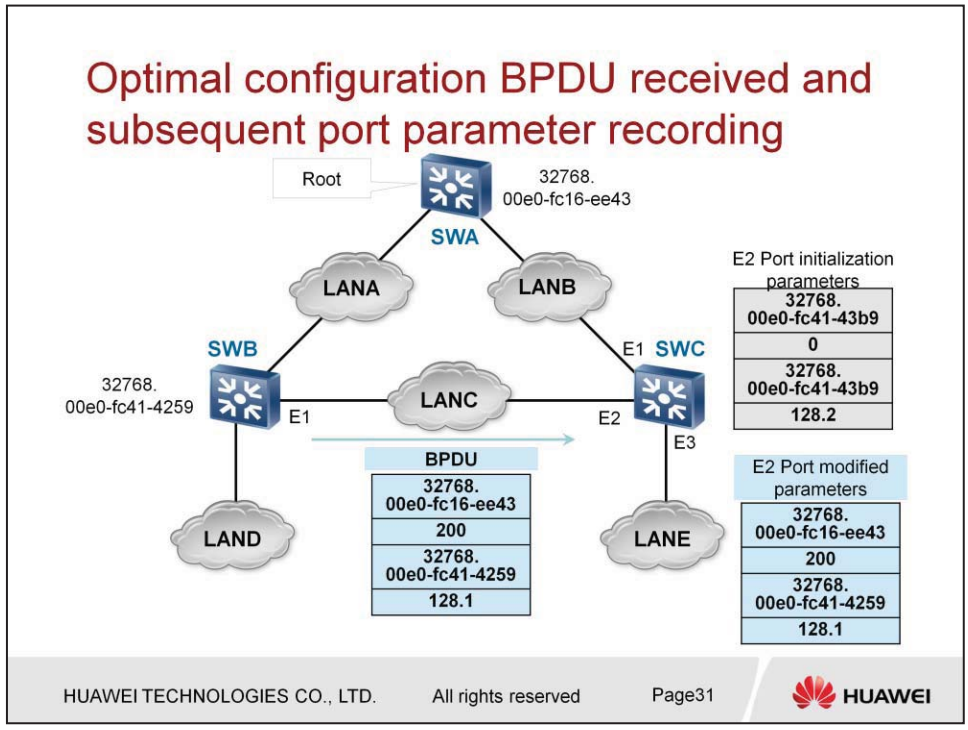
Parameters in Configuration BPDU

parameters	description
Port Identifier	Indicates the ID of the port sending a BPDU.
Message Age	Records the time since the root bridge originally generated the information that a BPDU is derived from.
Max Age	Indicates the maximum time that a BPDU is saved.
Hello Time	Indicates the interval at which BPDUs are sent.
Forward Delay	Indicates the time spent in the Listening and Learning states.

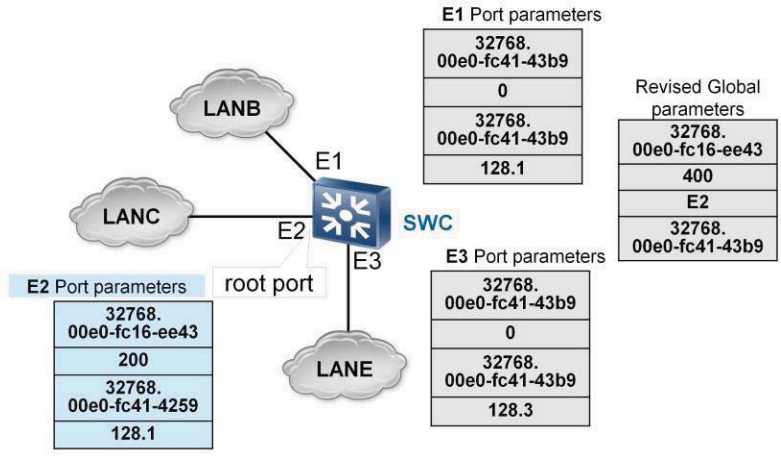

Initializing global parameters and port parameters
Root
32768.
00e0-fc16-ee43
SWA
LANA
LANB
32768.
00e0-fc41-43b9
0
32768.
00e0-fc41-43b9
128.1
E1
SWB
32768.
00e0-fc41-4259
LANC
SWC
E2
E3
32768.
00e0-fc41-43b9
0
NULL
32768.
00e0-fc41-43b9
LAND
LANE
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page29
HUAWEI

Optimal configuration BPDU received and subsequent port parameter recording

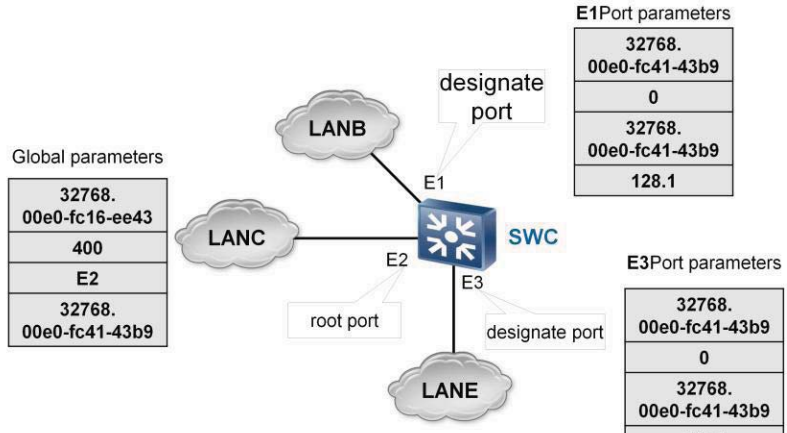
order	content
1	Designated Root and Root Identifier , if they are equal, proceed to the second step
2	Root Path Cost and Designated Cost,If they are equal, proceed to the third step
3	Bridge Identifier and Designated Bridge， If they are equal, proceed to the fourth step
4	Compare Port Identifier ,Designated Port

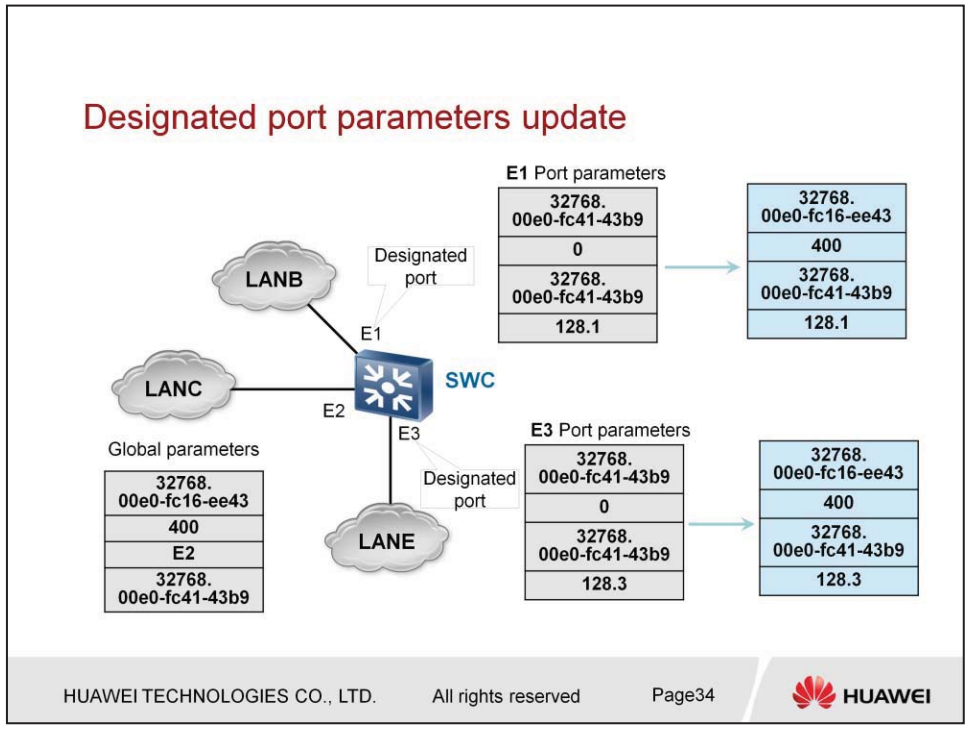

Optimal configuration BPDU received and subsequent port parameter recording
Root
32768.
00e0-fc16-ee43
SWA
LANA
LANB
E2 Port initialization parameters
32768.
00e0-fc41-43b9
0
32768.
00e0-fc41-43b9
128.2
SWB
E1
SWC
32768.
00e0-fc41-4259
E1
LANC
E2
E3
BPDU
32768.
00e0-fc16-ee43
200
32768.
00e0-fc41-4259
128.1
E2 Port modified parameters
32768.
00e0-fc16-ee43
200
32768.
00e0-fc41-4259
128.1
LAND
LANE
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page31
HUAWEI

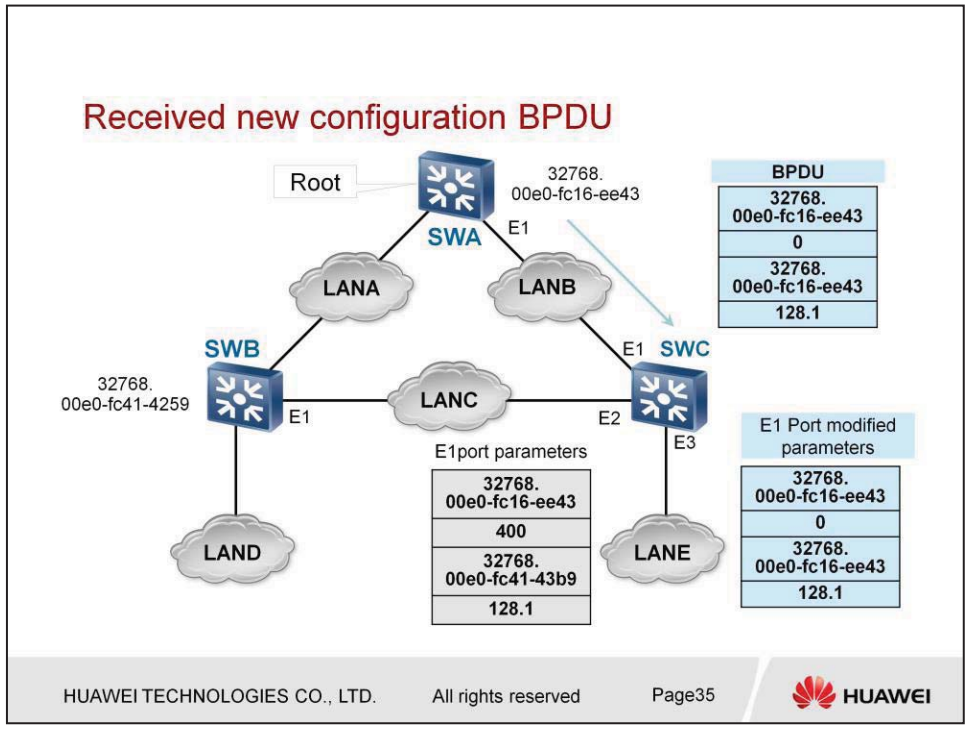
Root switch, root port and root path cost calculation



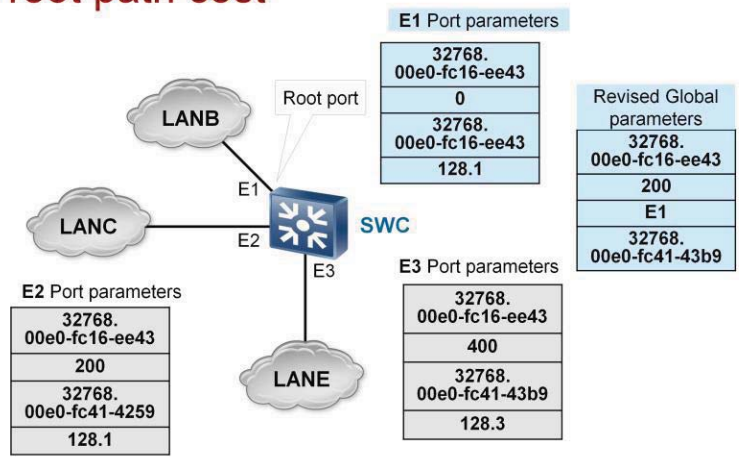
Designated port selection



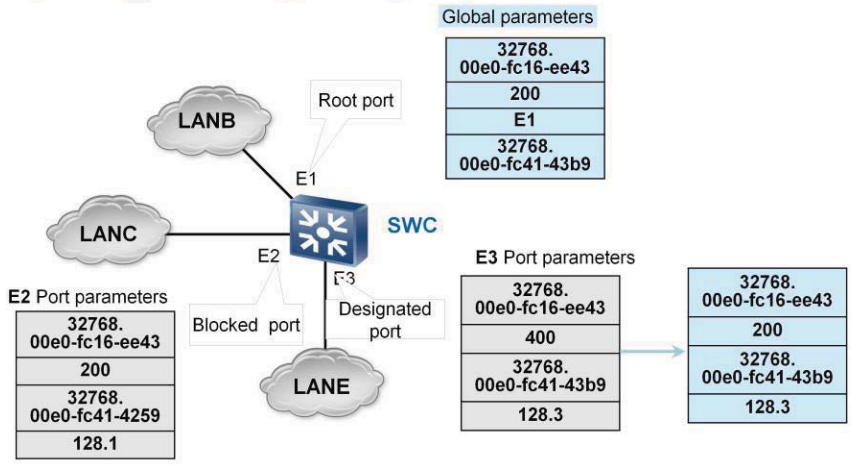

Designated port parameters
E1 Port parameters
32768.
00e0-fc41-43b9
0
32768.
00e0-fc41-43b9
128.1
32768.
00e0-fc16-ee43
400
32768.
00e0-fc41-43b9
128.1
LANB
Designated port
E1
LANC
SWC
E2
E3
Global parameters
32768.
00e0-fc16-ee43
400
E2
32768.
00e0-fc41-43b9
Designated port
LANE
E3 Port parameters
32768.
00e0-fc41-43b9
0
32768.
00e0-fc41-43b9
128.3
32768.
00e0-fc16-ee43
400
32768.
00e0-fc41-43b9
128.3
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page34
HUAWEI


Received new configuration BPDU
Root
32768.
00e0-fc16-ee43
SWA
E1
LANA
LANB
BPDU
32768.
00e0-fc16-ee43
0
32768.
00e0-fc16-ee43
128.1
SWB
E1
SWC
32768.
00e0-fc41-4259
E1
LANC
E2
E3
E1port parameters
32768.
00e0-fc16-ee43
400
32768.
00e0-fc41-43b9
128.1
E1 Port modified
parameters
32768.
00e0-fc16-ee43
0
32768.
00e0-fc16-ee43
128.1
LAND
LANE
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page35
HUAWEI

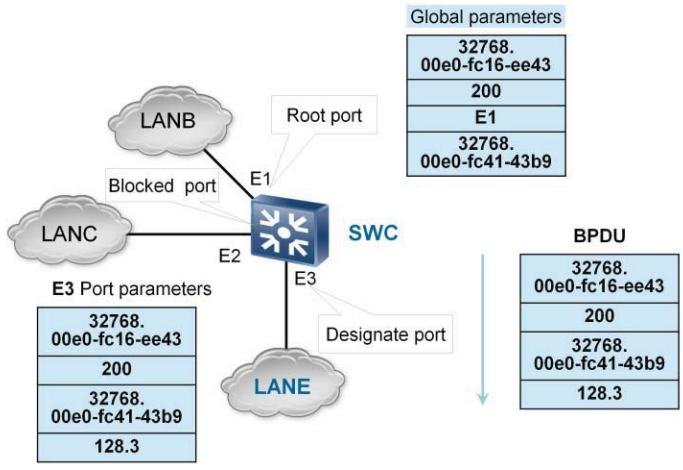
Re-calculation of the root switch, root port and root path cost

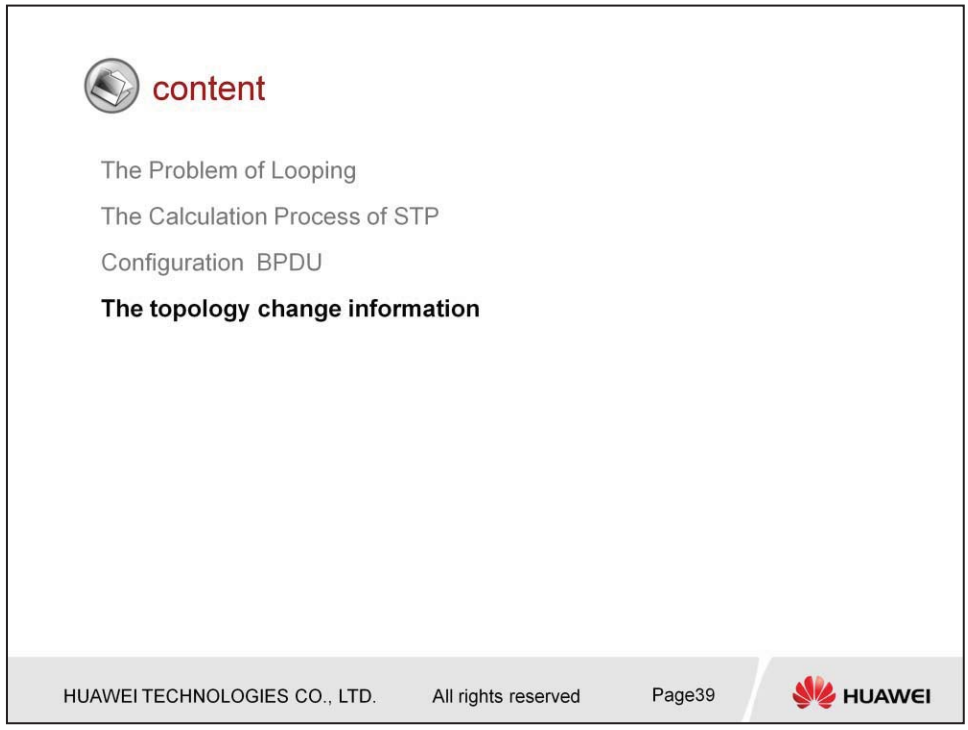


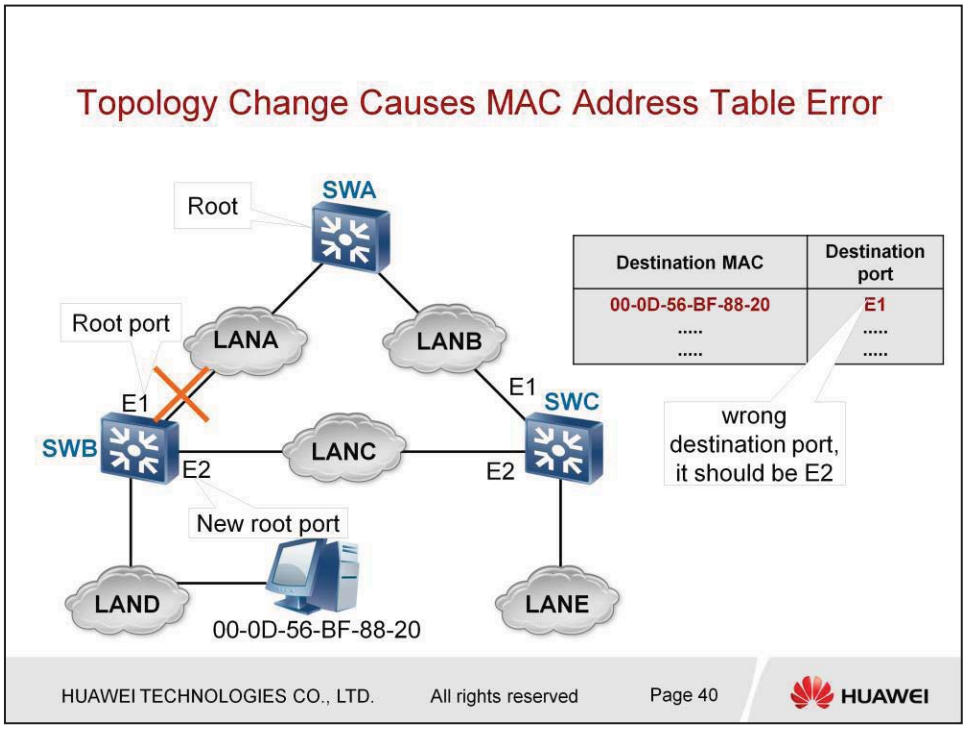
Recalculation of the designated port and updating the designated port parameters



Sends the configuration BPDU designate port



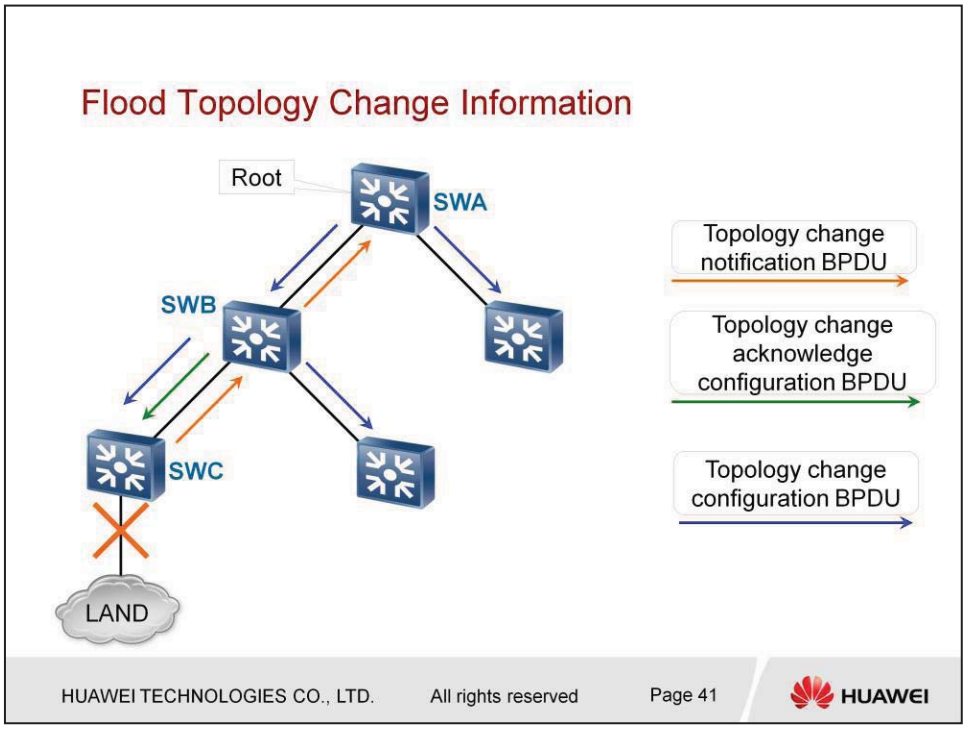

content
The Problem of Looping
The Calculation Process of STP
Configuration BPDU
The topology change information
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page39
HUAWEI



By default, the lifetime of dynamic entries in the MAC address table is 300 seconds (5 minutes).

In this example:

In a stable topology, messages from SWC reach a PC in LAND through port E1 of SWB. When E1 of SWB is disconnected, E2 becomes the new root port. The destination port for messages from SWC should changes to E2. However, a switch cannot detect the change of topology, so the MAC address table is incorrect. The data forwarding error caused by this fault may last for up to 5 minutes. The solution to this problem is flooding the topology change messages in the entire network and changing the lifetime of MAC address table to a smaller value. When the topology becomes stable, the lifetime of the MAC address table is restored to the original value. STP changes the lifetime of the MAC address table to the forward delay of the switch. By default, the value is 15 seconds.



Three types of BPDUs are involved in the flooding of topology change messages:

1. Topology change notification BPDU, sent by the root port of the non-root switch to notify the upstream switch of topology change. The switch sends such BPDUs every 2 seconds (Hello timer), until it receives the topology change

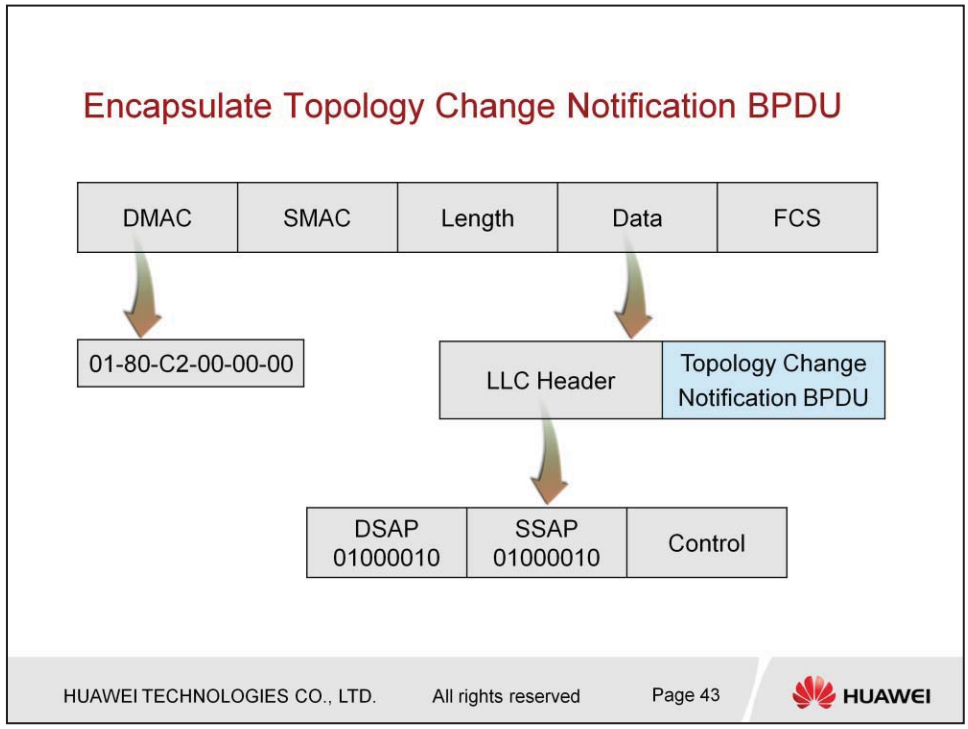
acknowledgment configuration BPDU or topology change configuration BPDU from the upstream switch.

2. Topology change acknowledgment configuration BPDU, which is a type of configuration BPDU. Different from a common configuration BPDU, this BPDU has a Flag field. This BPDU is used by the non-root switch to acknowledge the

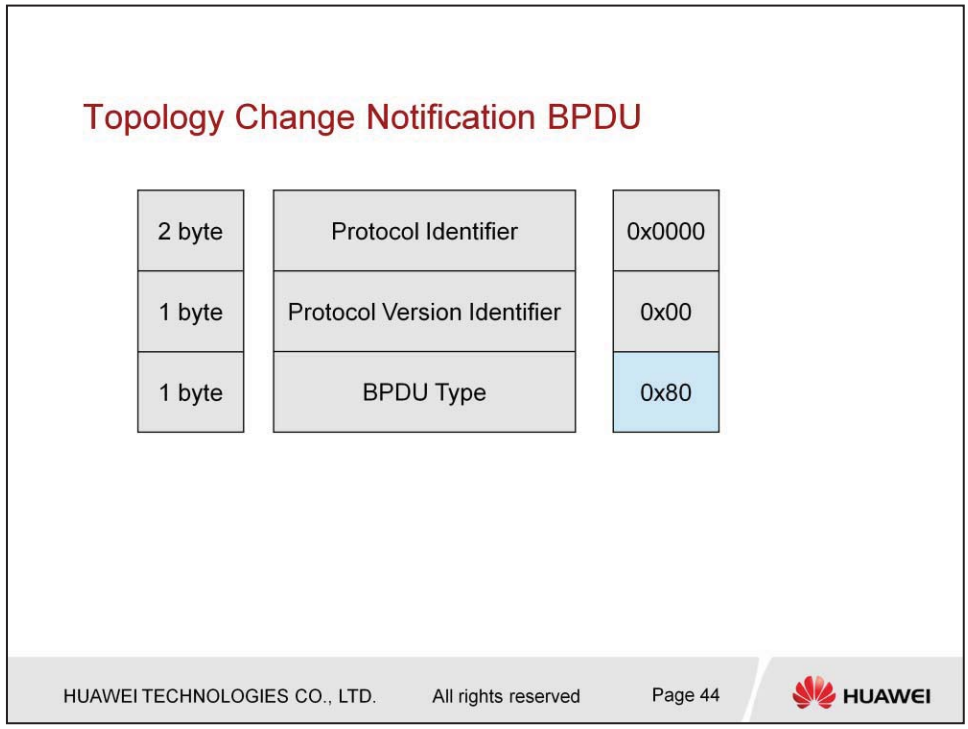
topology change notification sent by the downstream switch. The topology change acknowledgment configuration BPDU is sent by the port that receives the topology change notification BPDU.

3. Topology change configuration BPDU. This BPDU also has a Flag field and is used to flood the topology change messages in the entire network. Each switches floods such BPDUs from all its designated ports. After receiving the topology change notification BPDU from SWB, SWA sends the topology change configuration BPDU (configuration BPDU with a flag) every 2 seconds.

Thus all switches change the lifetime of their MAC address tables to the forward delay (15 seconds). After a period (max age plus forward delay, 35 seconds by default), SWA (the root bridge) clears the Flag field in the configuration BPDU. It indicates that the network topology is stable. At this time, the switches restore the lifetime of their MAC address tables to the original value.

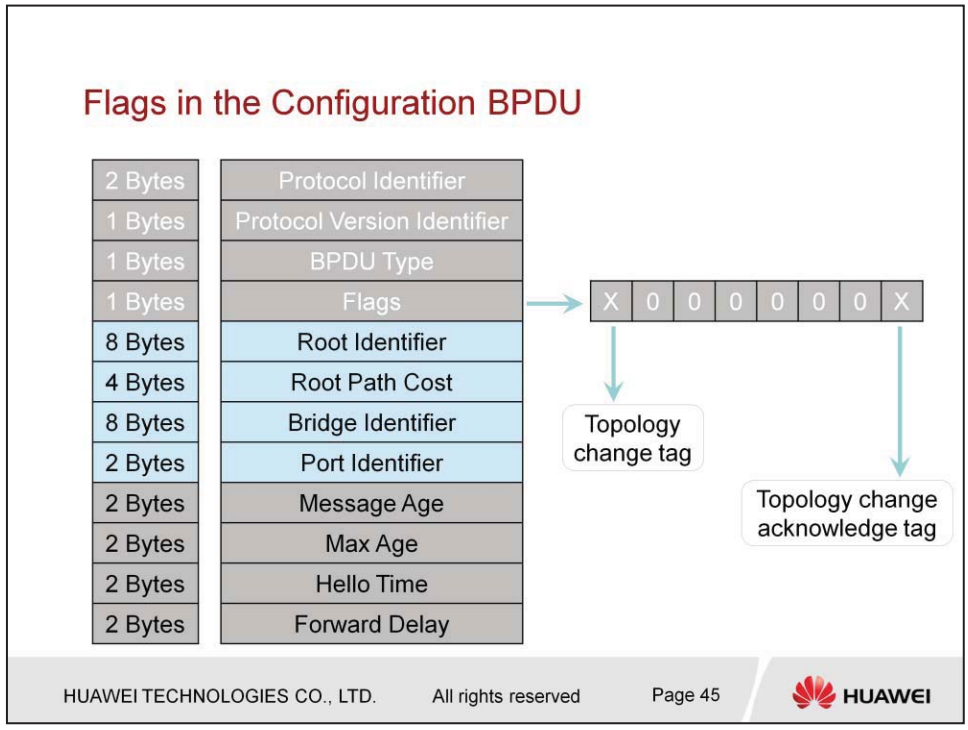


The encapsulation mode of the topology change notification BPDU is the same as the encapsulation mode of the configuration BPDU.



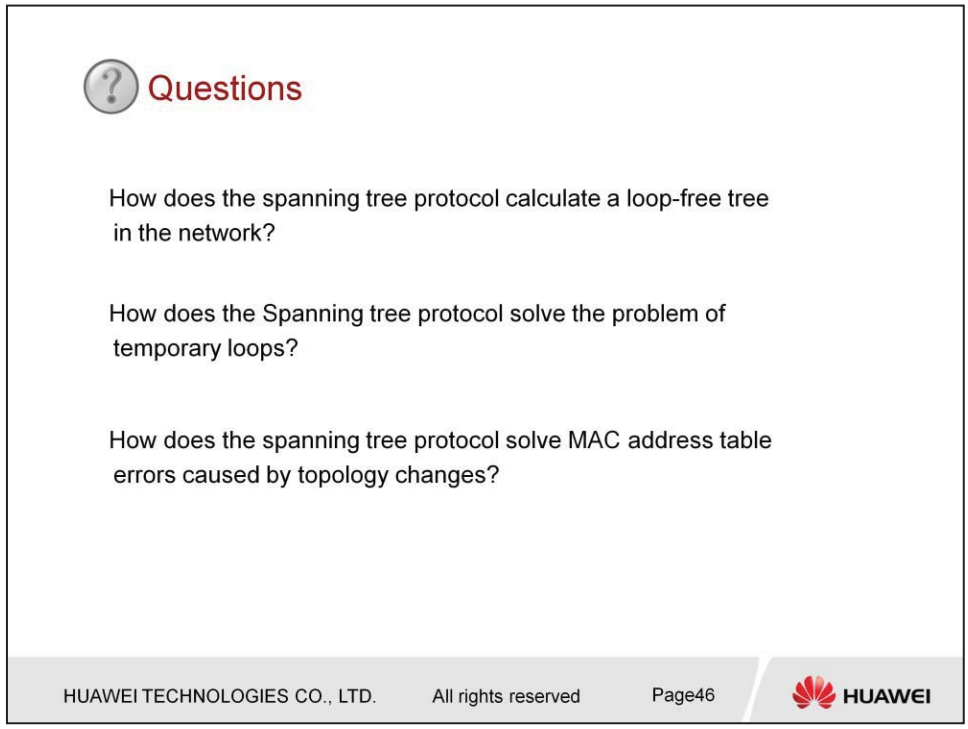
The format of the topology change notification is quite simple, because it does not contains as many parameters as the configuration BPDU does.

Protocol Identifier is 0. Protocol Version Identifier contains all 0s. BPDU Type is a binary value 1000 0000, which equals hexadecimal value 0x80.



The topology change acknowledgment configuration BPDU and the topology change configuration BPDU are both a type of configuration BPDU. The difference between such configuration BPDUs and the common configuration BPDU lies in:

In the common configuration BPDU, all bits in the Flag field are 0s. In the topology change acknowledgment configuration BPDU, the eighth bit of the Flag field is 1. In the topology change configuration BPDU, the first bit of the Flag field is 1.



How does the spanning tree protocol calculate a loop-free tree in the network?

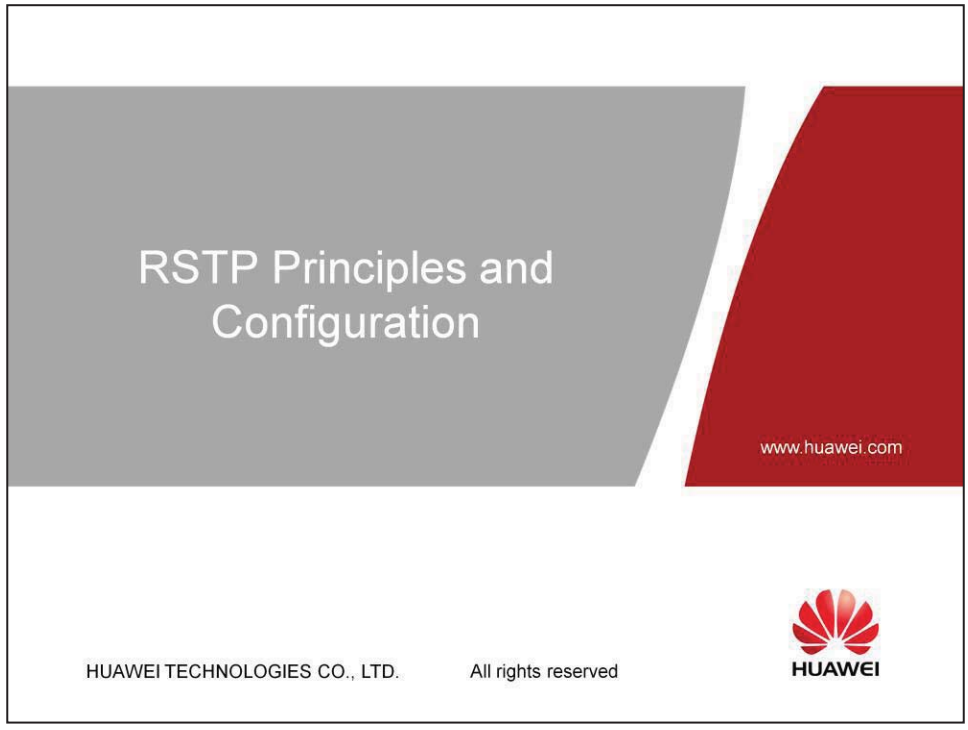
Elect a root switch in the network, select a root port for each non-root switch to select a specific port for each segment, is neither a root port is not specified port is set to blocking state.

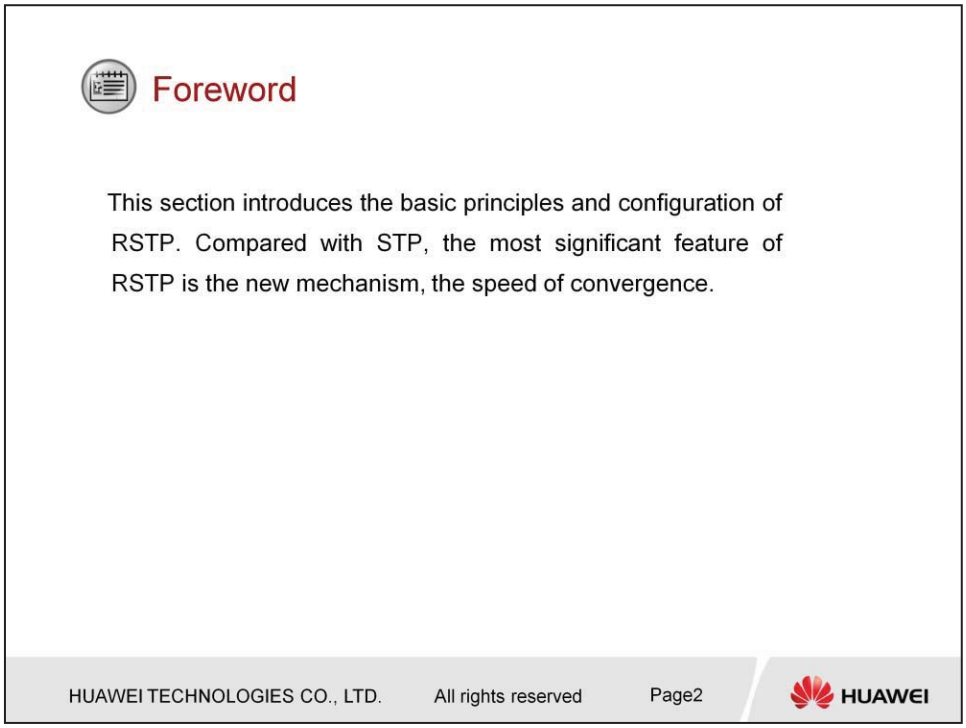
How does the Spanning tree protocol solve the problem of temporary loops?

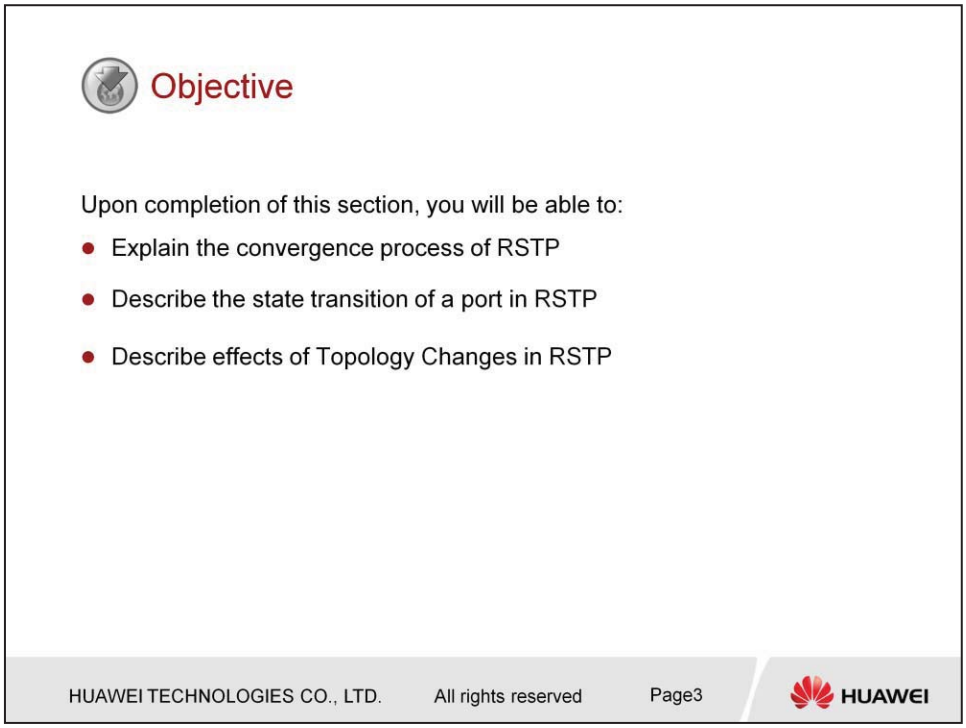
When a port never before forwarding state to the forwarding state to go through two Forward Delay interval to ensure that the other switches in the network to complete the spanning tree calculation.

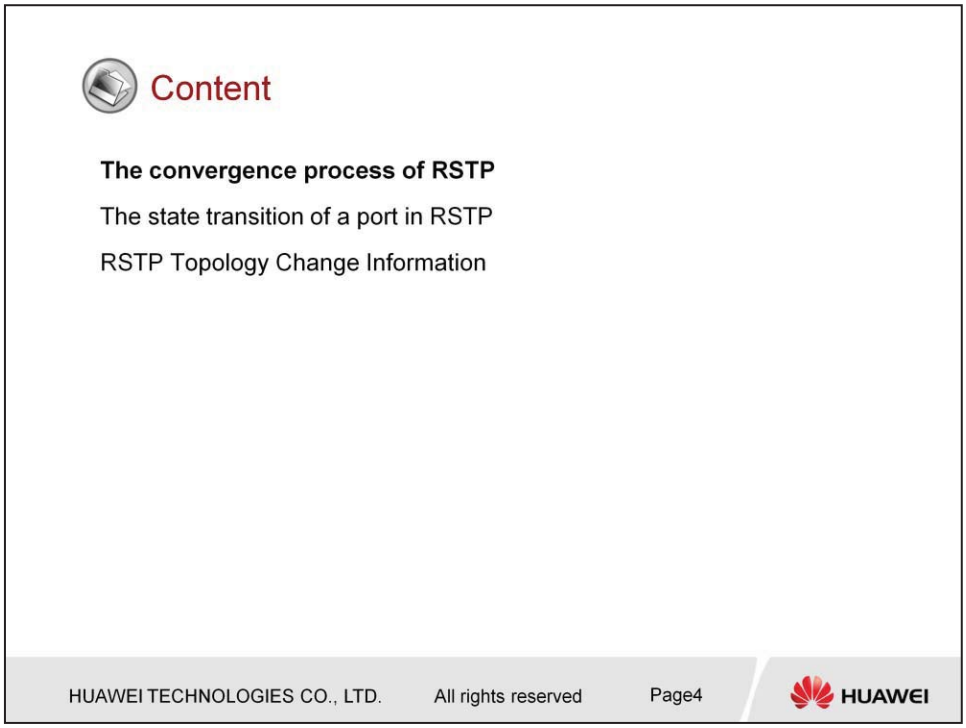
How does the spanning tree protocol solve MAC address table errors caused by topology changes?

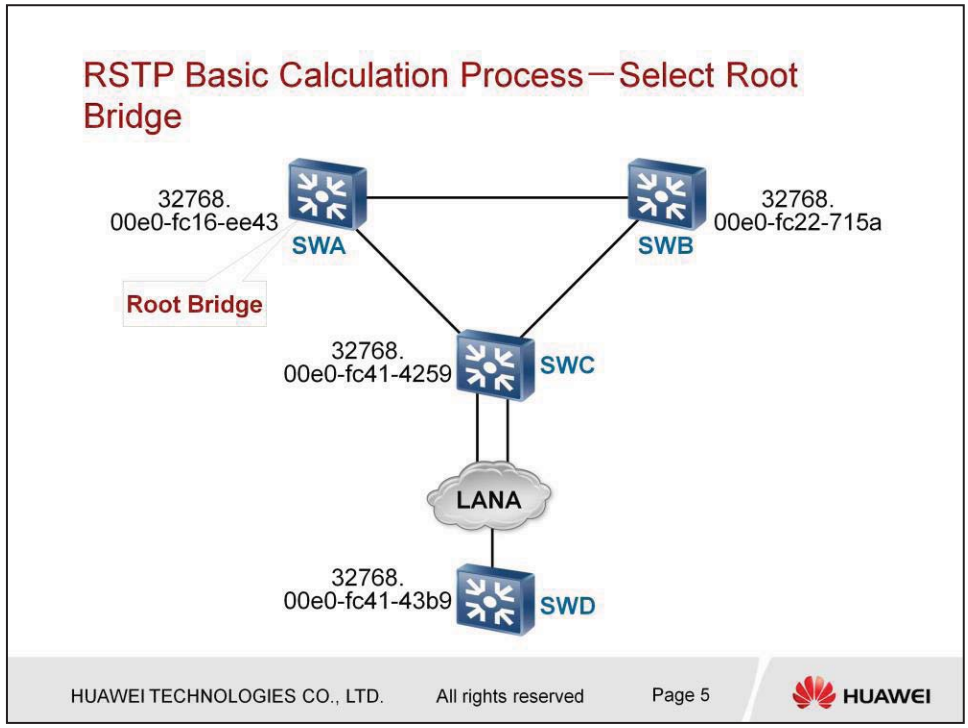
After the topology change, topology change information within the whole network flooding, the switch MAC address table survival is set to a shorter value, the stability of the topology, switch to restore the survival period of the MAC address table.


RSTP Principles and Configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
This section introduces the basic principles and configuration of RSTP. Compared with STP, the most significant feature of RSTP is the new mechanism, the speed of convergence.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page2
HUAWEI


Objective
Upon completion of this section, you will be able to:
Explain the convergence process of RSTP
Describe the state transition of a port in RSTP
Describe effects of Topology Changes in RSTP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page3
HUAWEI


Content
The convergence process of RSTP
The state transition of a port in RSTP
RSTP Topology Change Information
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page4
HUAWEI



Similar to the calculation process of STP, RSTP calculation also starts from electing the root bridge. The root bridge is elected based on the bridge identifier.

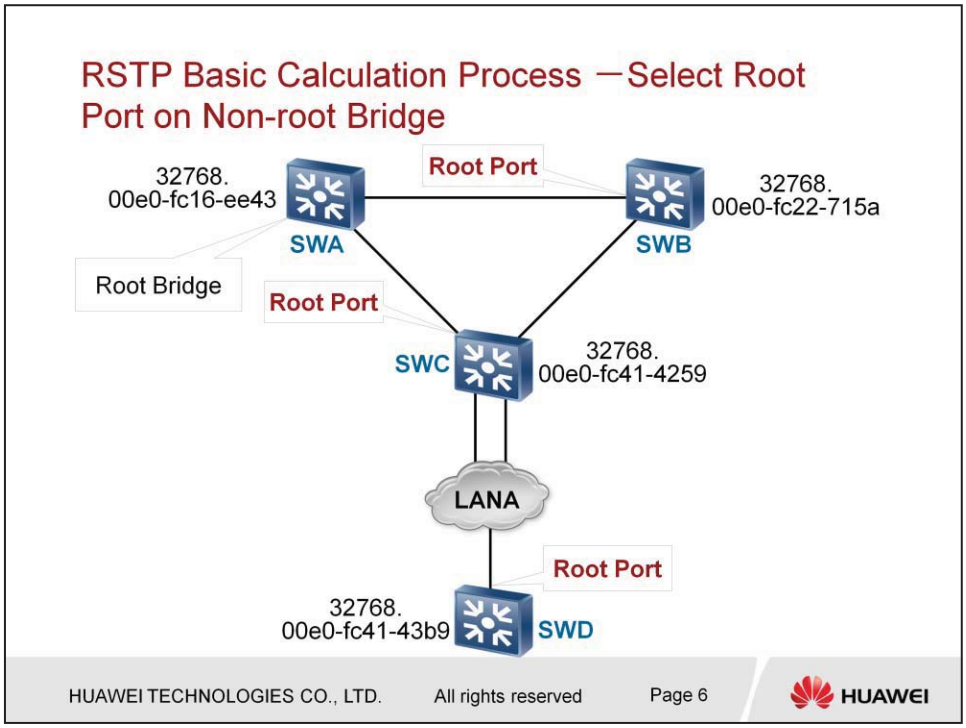
The bridge identifier consists of a 2-byte bridge priority and a 6-byte MAC address.

The bridge priority is configurable. The value ranges from 0 to 65535 and the default value is 32768.

In the network, the switch with the smallest identifier becomes the root bridge.

The system compares the priority first. If the switches have the same priority, the system compares their MAC addresses. The switch with the smallest MAC address is elected preferencetially.

In this example, the three switches have the same priority. SWA has the smallest MAC address, so SWA is elected as the root bridge.



RSTP elects a root port for each non-root bridge.

Each port of a switch has a port cost parameter. The port cost refers the cost for sending the data from this port, namely the cost of the outgoing port. RSTP considers that no cost is needed for receiving the data on a port..

The port cost depends on the bandwidth of the port. The higher the bandwidth is, the smaller the port cost will be. On the VRP, the cost of a100 M port is 200.

Multiple paths may exist between a non-root bridge and a root bridge. The cost of a path is the total cost of all outgoing ports on this path.

A root port is a local port on the path with the least cost from a non-root bridge to the root bridge. The cost of this path is referred to the root path cost. If multiple root ports exist, the system compares the identifiers of the upstream switches.

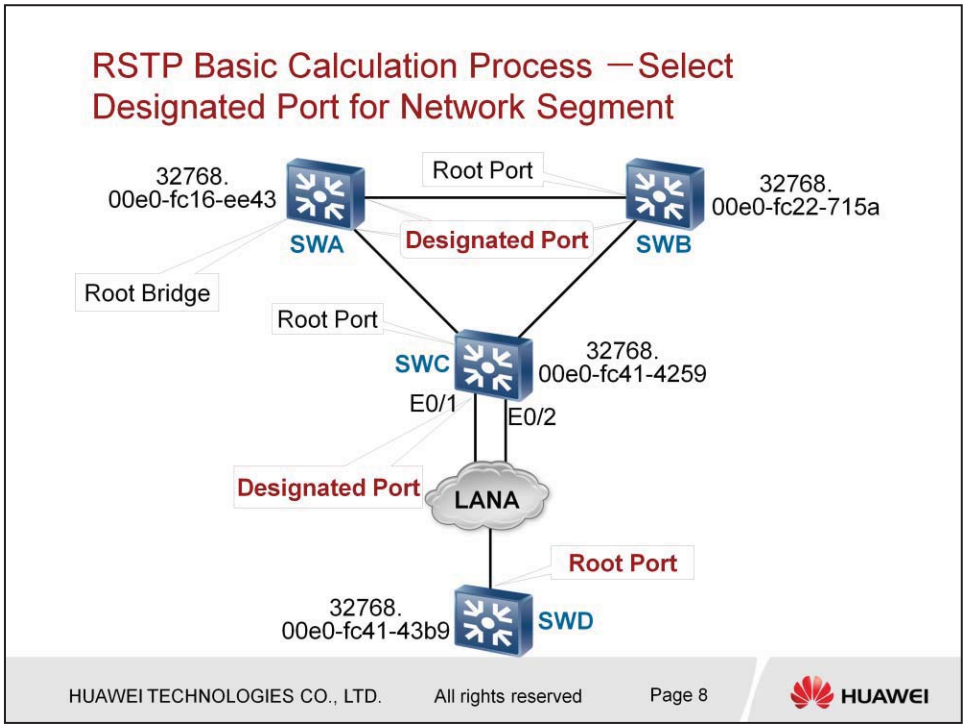
The port whose upstream switch has the smallest identifier is elected. If the upstream switches have the same bridge identifier, the system compares the identifiers of the upstream ports. The port whose upstream port has the smallest identifier is elected.

The port identifier consists of a 1-byte port priority and a 1-byte

port number.

The port priority is configurable. The default value is 128.

In this example, we assume that all ports are 100 M ports and their cost values are all 200.



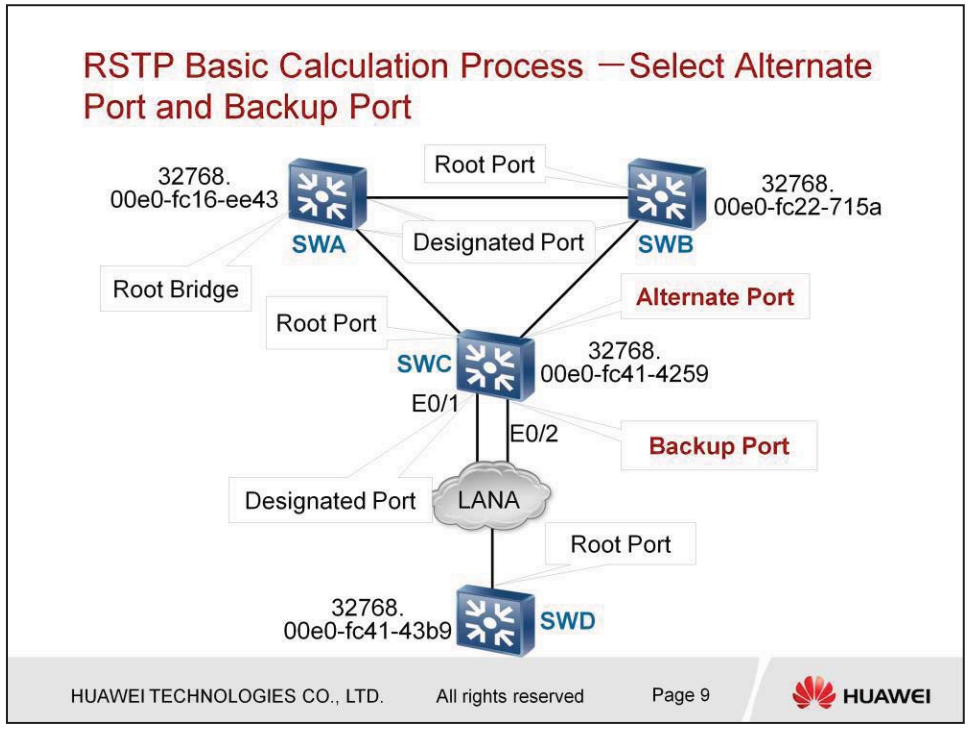
RSTP elects the designated port for each network segment. The designated port forwards the data transmitted between the root bridge and this network segment.

The switch where the designated port is located is called the designated switch.

When electing the designated port and designated switch for a network segment.

RSTP compares the root path cost of the switch on which the port is connected to this network segment. If the switches have the same root path cost, RSTP compares their bridge identifiers. The switch with the smallest identifier has the highest priority. If their identifiers are also the same, RSTP compares the identifiers of the ports connected to the network segment. The port with the smallest identifier has the highest priority.

In this example, SWA is the root bridge, so all its ports are designated ports. For the link between SWC and SWB, SWB has a smaller bridge identifier, so the designated port is on SWB. For LANA, the designated switch is SWC. SWC has two ports connected to LANA. RSTP compares the port identifiers. The ports have the same port priority, 128. Port E0/1 has a smaller port number, so E0/1 of SWC is the designated port of LANA.



For the ports that are neither the root port nor the designated port, the setting is as follows:

If a port is on the designated switch of the network segment, the port is set to the backup port.

If a port is not on the designated switch of the network , the port is set to the alternate port.

The alternate port is the backup of the root port, and the backup port is the backup of the designated port.

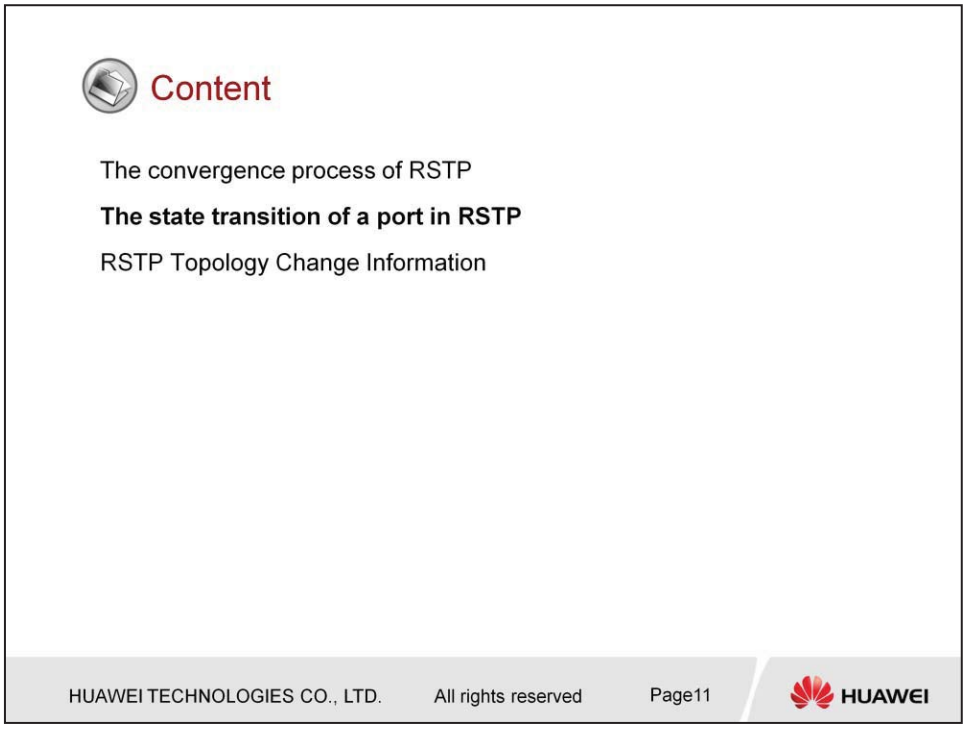
The alternate port and backup port are not in Forwarding states.

Bridge Port Role

Port role	Description
Root Port	Root port is the nearest port to the root bridge, it is in the forwarding state in a stable network .
Designated Port	Responsible for forwarding data to the downstream network segments or switches. It is in the forwarding state in a stable network.
Backup Port	Backup port is not in the forwarding state, the switch it belongs to is the designated bridge of network segment it connects.
Alternate Port	Alternate Port is not in the forwarding state, the switch it belongs to is not the designated bridge of network segment it connects.

As mentioned before, RSTP defines four roles for the RSTP-enabled port that works normally on the physical layer and data link layer. The root port and designated port are in Forwarding status when they are stable.

The port that is not enabled is called the Disabled port.


Content
The convergence process of RSTP
The state transition of a port in RSTP
RSTP Topology Change Information
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page11
HUAWEI

Bridge Port Status

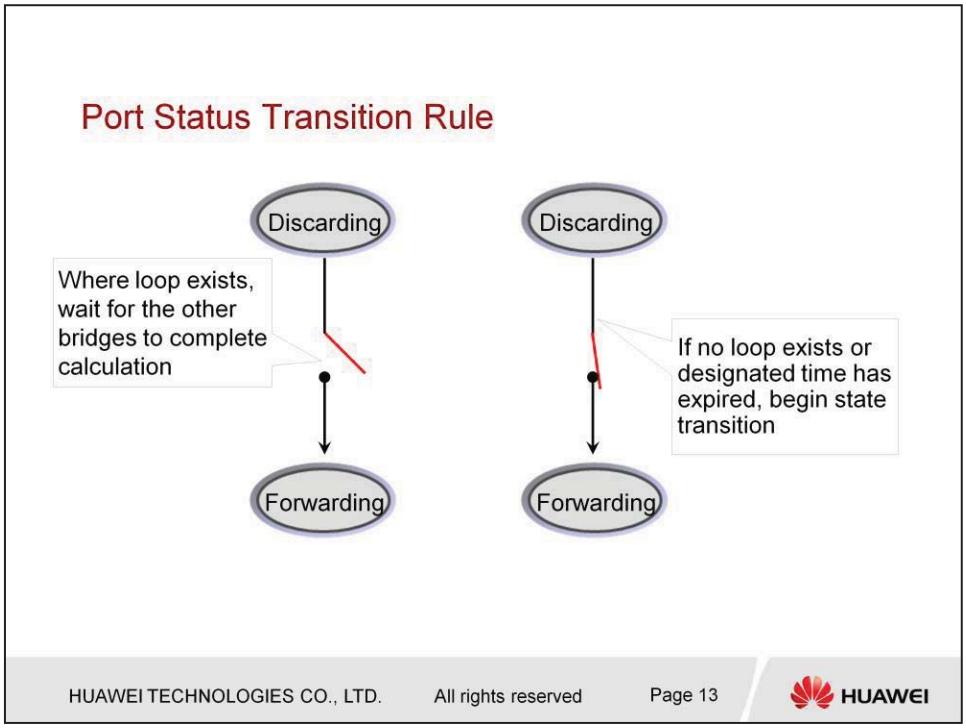
Port status	Description
Discarding	A port in the Discarding state can only receive BPDUs. Includes Alternate Port and Backup Port.
Learning	A port in the Learning state learns MAC addresses from user traffic to construct a MAC address table. In the Learning state, the port can send and receive BPDUs, but cannot forward user traffic.
Forwarding	A port in the Forwarding state can send and receive BPDUs as well as forward user traffic.

Different from STP, RSTP defines only three port statuses: Discarding, Learning, and Forwarding.

The alternate port and backup port are in Discarding state.

The designated port and root port is in Forwarding state when they are stable.

Learning is a temporary status of some designated ports and root ports before they switch to Forwarding state.



According to the election rule, the port status must be set after the port role is determined.

The risk of loop does not exist when the port status switches from Forwarding to Discarding (the root port or designated port changes to the alternate port or backup port). Therefore, the transition occurs immediately.

The risk of loop does not exist when the port status switches from Forwarding to Forwarding (the root port changes to the designated port or the designated port changes to the root port). Therefore, the transition also occurs immediately.

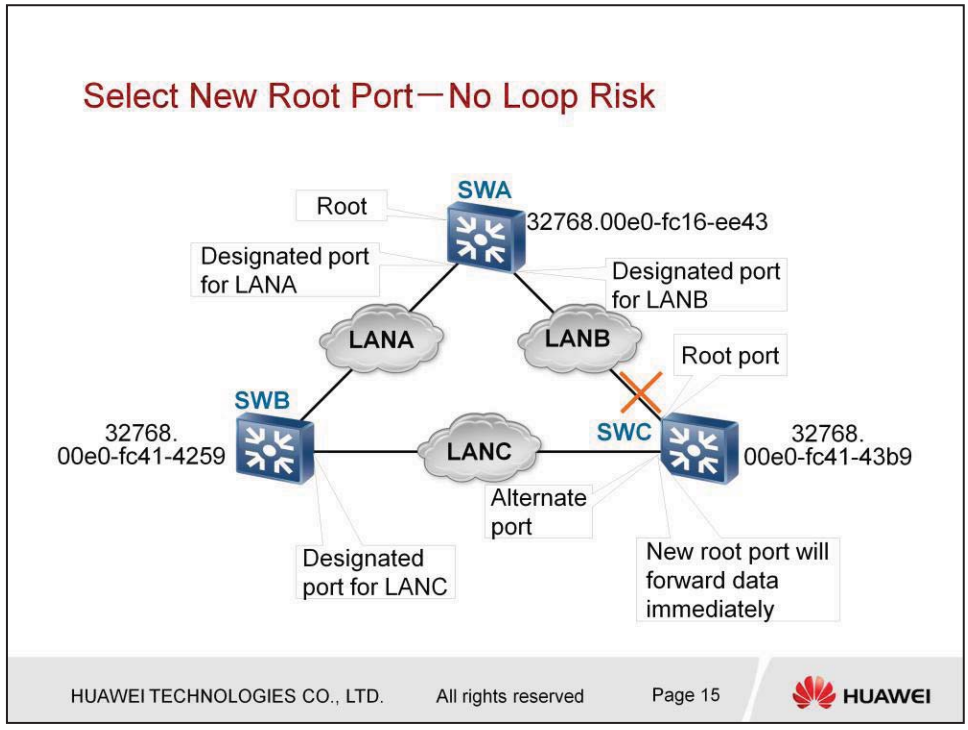
The risk of loop exists when the port status switches from Discarding to Forwarding (the alternate port or backup port changes to the root port or designated port). In STP, the port must wait two times of the forward delay before switching from the non-Forwarding status to Forwarding status.

Therefore, the ports that need to switch to non-Forwarding state have enough time to calculate the spanning tree. RSTP improves this feature.

The principle of RSTP is raises the convergence speed when no temporary risk exists. That is, make a port switch from the non-Forwarding status to Forwarding status as soon as possible after

it changes to the designated port or root port.

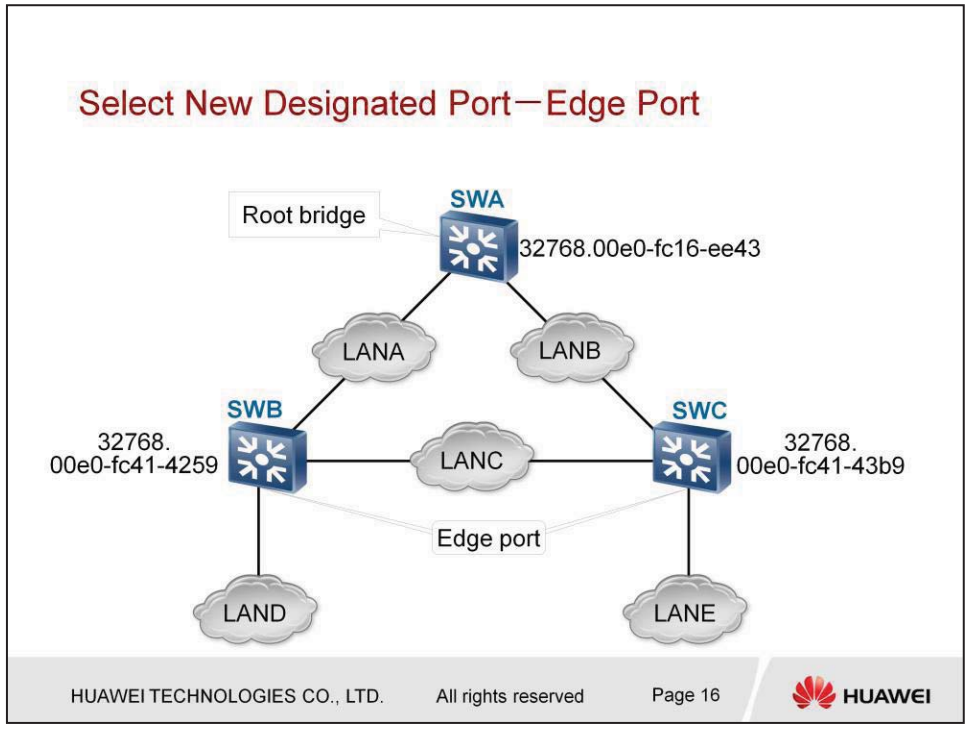
Therefore, ensuring that no loop risk exits is the core of RSTP.



When a new root port is elected on a non-root switch, the new root port switches to Forwarding state immediately if the previous root port is not in Forwarding state.

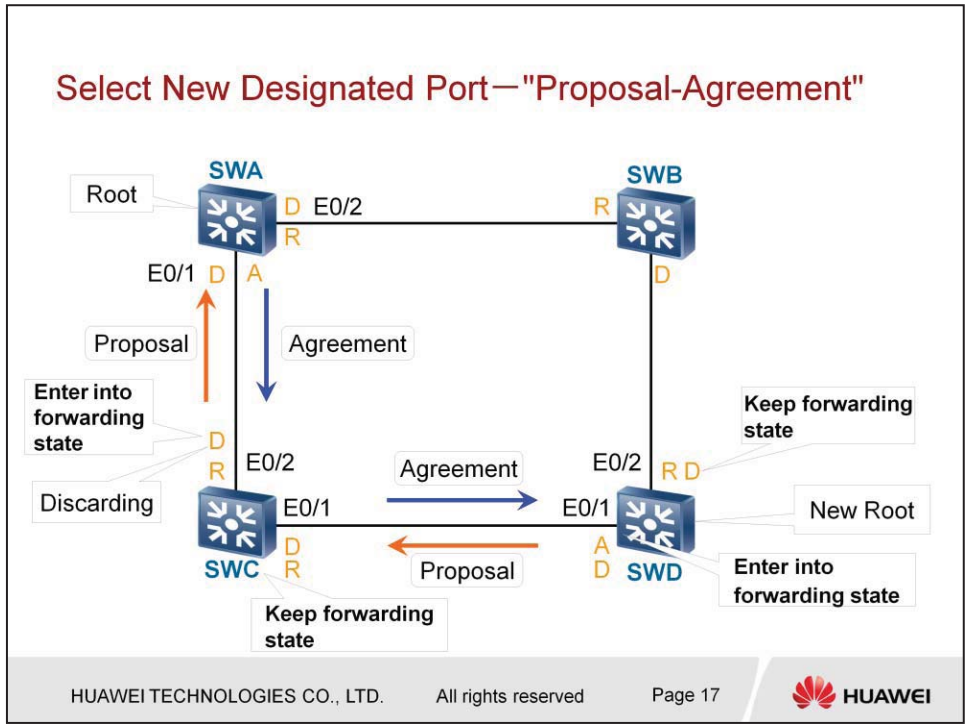
In this example, the port connected to LANB on SWC is the root port. If this root port is disconnected, that is, it is not in Forwarding state, SWC needs to elect a new root port. The port connected to LANC changes from the alternate port to the

root port. Since the previous root port is not in Forwarding state, no loop risk exists in the network. The new root port can switch to Forwarding state immediately.



An edge port is a port that is not connected to any switch.

When a switch port is configured as the edge port, the port becomes the designated port immediately after it is enabled and switches to Forwarding state.



RSTP adopts the “Proposal-Agreement” negotiation mechanism to speed up the process that a non-edge root switches from Discarding state to Forwarding state after the port becomes the root port.

In this example, assume that the priority sequence of the switches in the network is SWA>SWB> SWC>SWD. Thus, SWA is the root bridge; E0/1 of SWD is the alternate port and is in Discarding state. If the priority of SWD is changed and the priority sequence of the switches changes to SWD>SWA>SWB>SWC, the negotiation process is as follows:

1. SWD becomes the root bridge ,then E0/1 and E0/2 of SWD becomes the designated port immediately. E0/2 remains in Forwarding state. E0/1 sends out a Proposal message, which is an RST BPDU with a flag. This BPDU also contains the parameters for calculating the spanning tree.

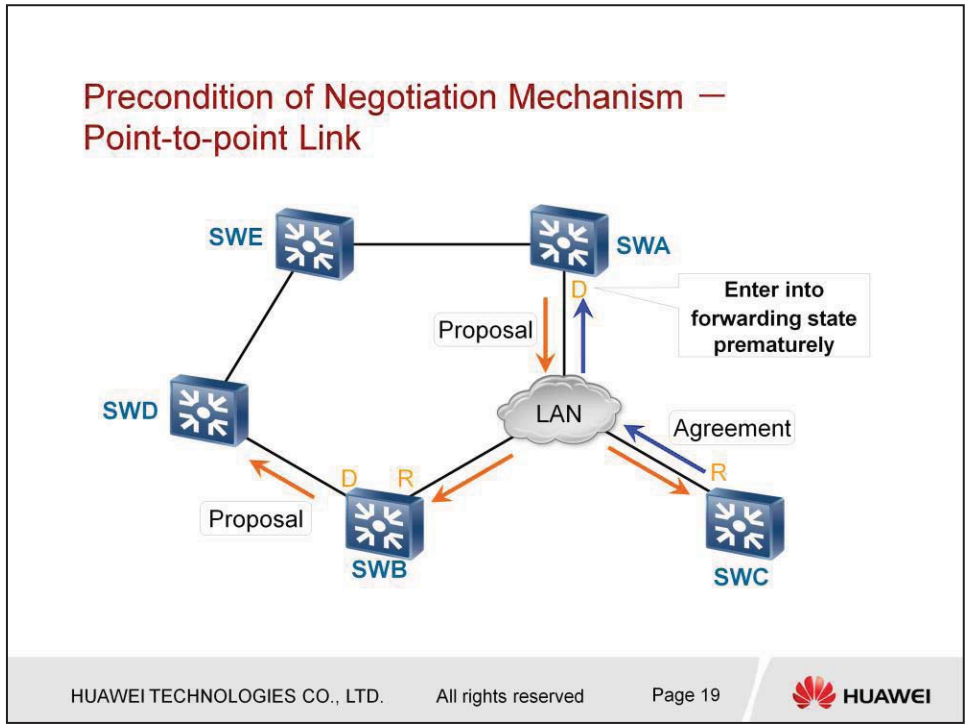
2. After SWC receives the Proposal, it calculates the spanning tree. E0/1 of SWC becomes the root port and remains in Forwarding state. E0/2 of SWC becomes the designated port. If the Proposal is received by the new root port, all non-edge designated ports switch to the Discarding state and send out new Proposal messages. If all non-root ports need to switch to

Discarding state or become the edge port, the root port that receives the Proposal message sends the Agreement message. In this example, E0/2 of SWC switches to Discarding state and sends a new Proposal message.

3. After SWA receives the Proposal, it calculates the spanning tree. E0/1 of SWA becomes the designated port and E0/2 becomes the root port. If the port that receives the Proposal needs to switch to Discarding state, this port sends the Agreement message after the status changes.

4. After E0/2 of SWC receives the Agreement, the port switches to Forwarding state immediately. After all non-edge designated ports receive the Agreement message, SWC sends the Agreement message through the root port.

5. The designated port of SWD switches to Forwarding state immediately after receiving the Agreement message.



The prerequisite for the “Proposal-Agreement” mechanism is that the all links that flood the Proposal and Agreement messages are point-to-point links. A point-to-point link is the link that directly connects two switches.

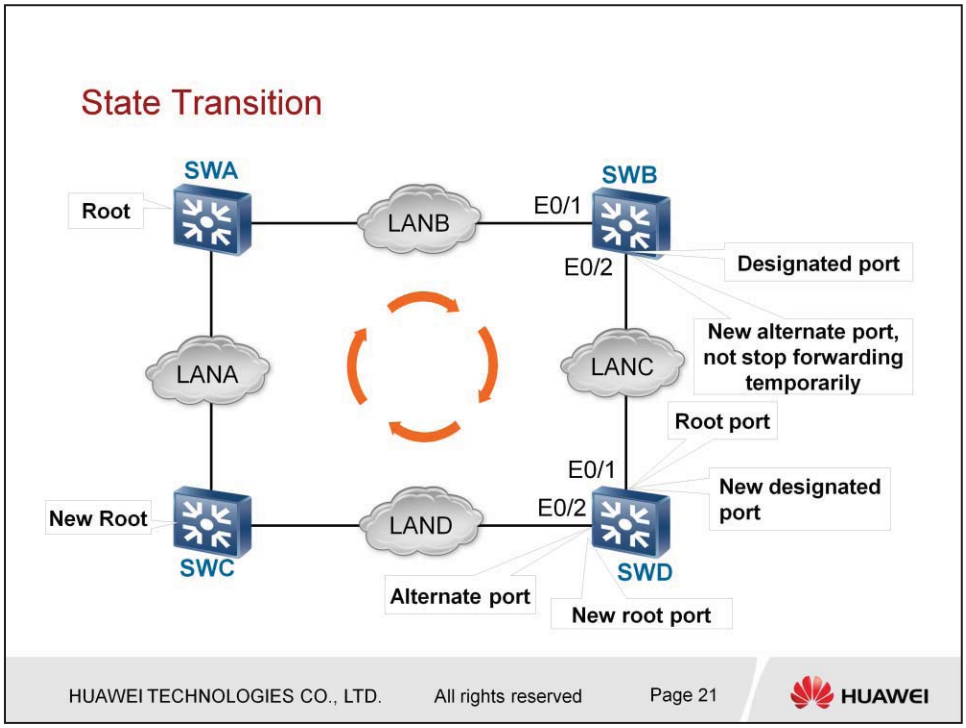
The point-to-point link is required because the point-to-multipoint link subjects to the loop risk. As shown in the figure, after SWA sends out a Proposal message, SWC immediately returns an Agreement message because SWC is the network edge. Thus, the new designated port of SWA switches to Forwarding state.

However, SWB, SWD, and SWE have not completed the flooding of “Proposal- Agreement”, so the loop may be generated. Therefore, the “Proposal-Agreement” mechanism requires that the link between switches be the point-to-point link.

In fact, if the link between two switches is not a point-to-point link, the flooding is automatically stopped. A port in Discarding state needs to wait enough time (two times as long as the forward delay) before switching to Forwarding state.

Actually, the “Proposal-Agreement” mechanism is the “triggering calculation-acknowledgement” mechanism on a point-to-point link. The “triggering calculation-acknowledgement” process is flooded on the point-to- point link to the end of the network (edge

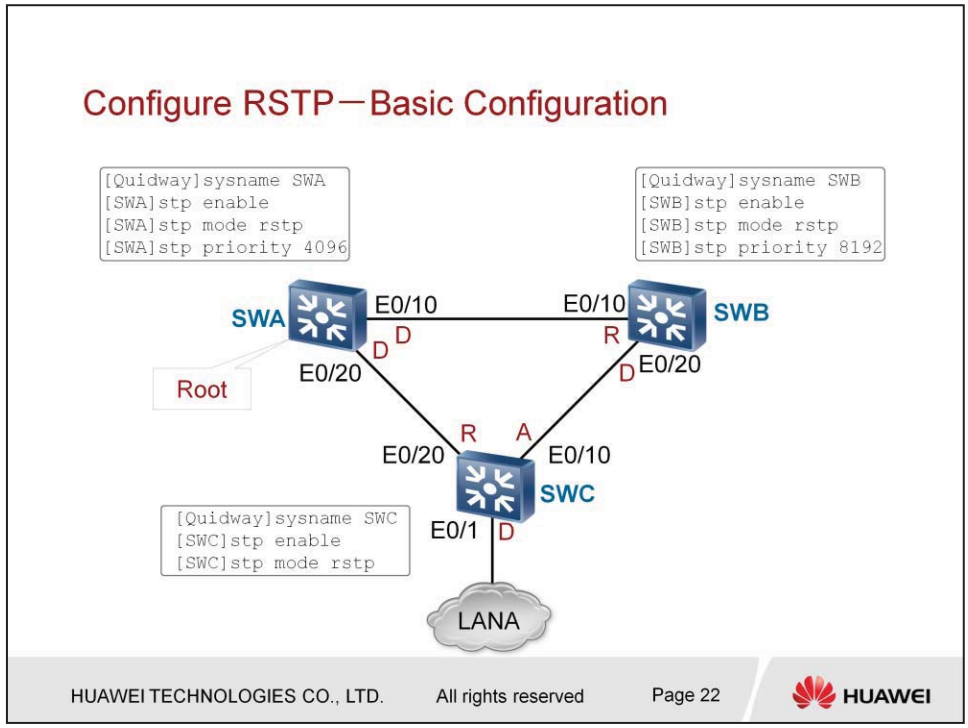
switch, on which all non-root ports are edge roots) or alternate port. If the alternate port is in Discarding state, it indicates that the loop is torn down.



As shown in the figure, after SWD elects a new root port, the previous root port becomes the new designated port and remains in Forwarding state. If E0/2 (the new root port) of SWD switches to Forwarding state before E0/2 (the new alternate port) of SWB switches to Discarding state, loop is generated in the network.

After the previous root port switches to Forwarding state (the previous root port becomes the new designated port), the new root port needs to wait a period of forward delay (15 seconds by default) before changing to the Learning state. After another period of forward delay, the new root port switches to Forwarding state.

This process is the same as that in STP.



In this example, you need to change the bridge priorities to make SWA become the root bridge.

Set the priority of SWA to 4096 and priority of SWB to 8192. Keep the default priority of SWC. Thus, SWA becomes the root bridge. The port roles on SWA are shown in the figure.

stp { enable | disable }

The stp command is used to enable or disable STP on a switch or on a port. By default, STP is disabled on the switch.

stp mode { stp | rstp | mstp }

The stp mode command is used to set the STP working mode on a switch. By default, the working mode of the switch is MSTP.

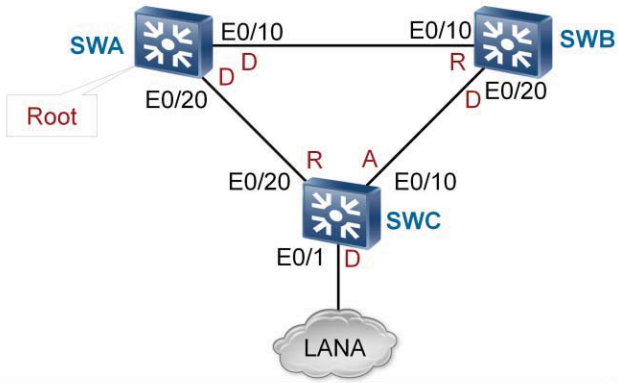
stp priority priority

priority: specifies the priority of a switch. The value ranges from 0 to 61440, with the step of 4096. That is, 16 priority values are available for a switch, for example, 0, 4096, 8192, and so on.

The stp priority command is used to set the bridge priorities. By default, the bridge priority is 32768.

Configure RSTP—Configure Point-to-point Link

```
[SWA]interface Ethernet 0/10
[SWA-Ethernet0/10]stp point-to-point force-true
[SWA]interface Ethernet 0/20
[SWA-Ethernet0/20]stp point-to-point force-true
```



To raise the convergence speed of RSTP, configure the links between switches to be point-to-point links.

stp point-to-point { force-true | force-false | auto }

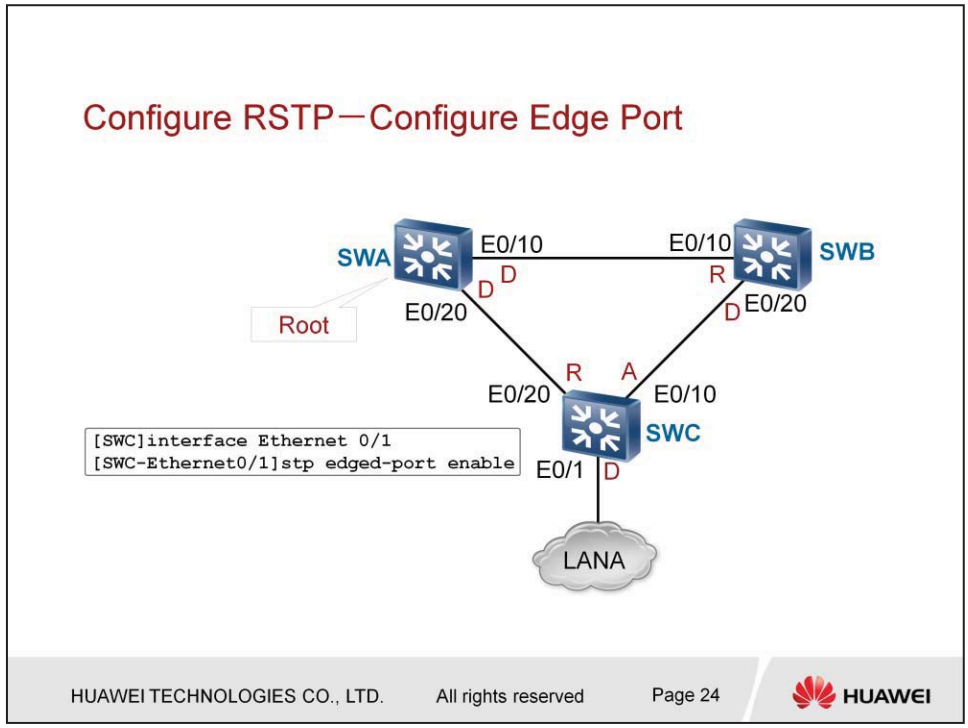
force-true: indicates that the link connected to the current Ethernet interface is a point-to-point link.

force-false: indicates that the link connected to the current Ethernet interface is not a point-to-point link.

auto: automatically checks whether the link connected to the current Ethernet interface is a point-to-point link.

By default, the auto keyword is selected. If the current Ethernet interface works in full-duplex mode, the link connected to this interface is regarded as a point-to-point link.

The configuration of SWB and SWC are similar to that of SWA.



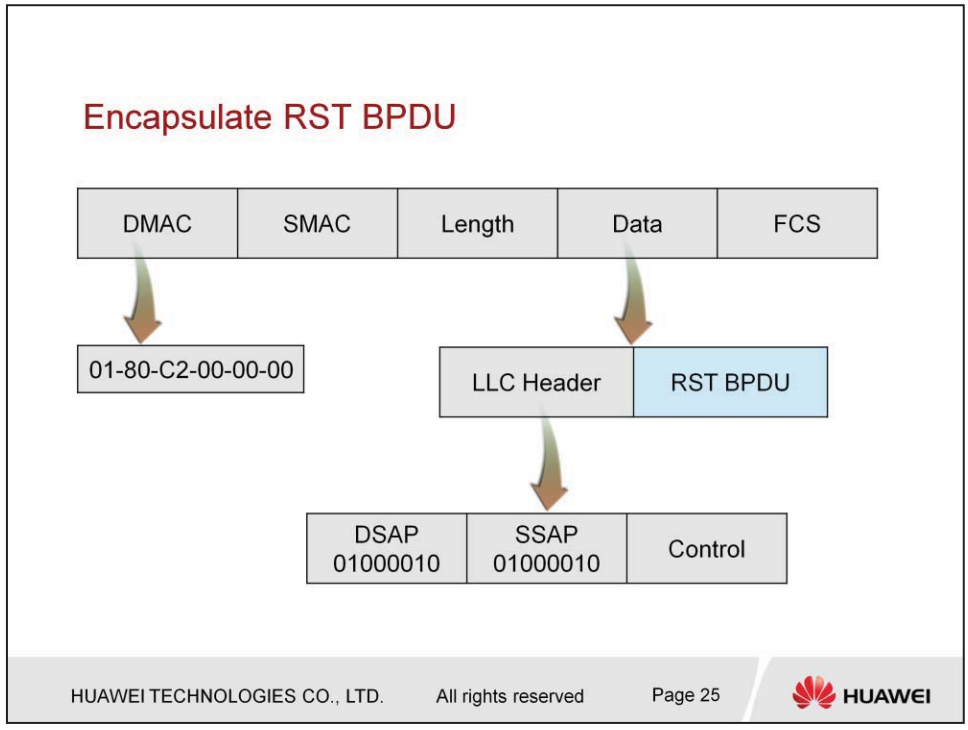
You can configure the ports connected to the network segment without other ports as the edge ports. In this way, you can raise the convergence speed of RSTP.

stp edged-port { enable | disable }

enable: configures the current Ethernet interface as the edge port.

disable: configures the current Ethernet interface as the non-edge port.

By default, all Ethernet interfaces of a switch are non-edge ports.



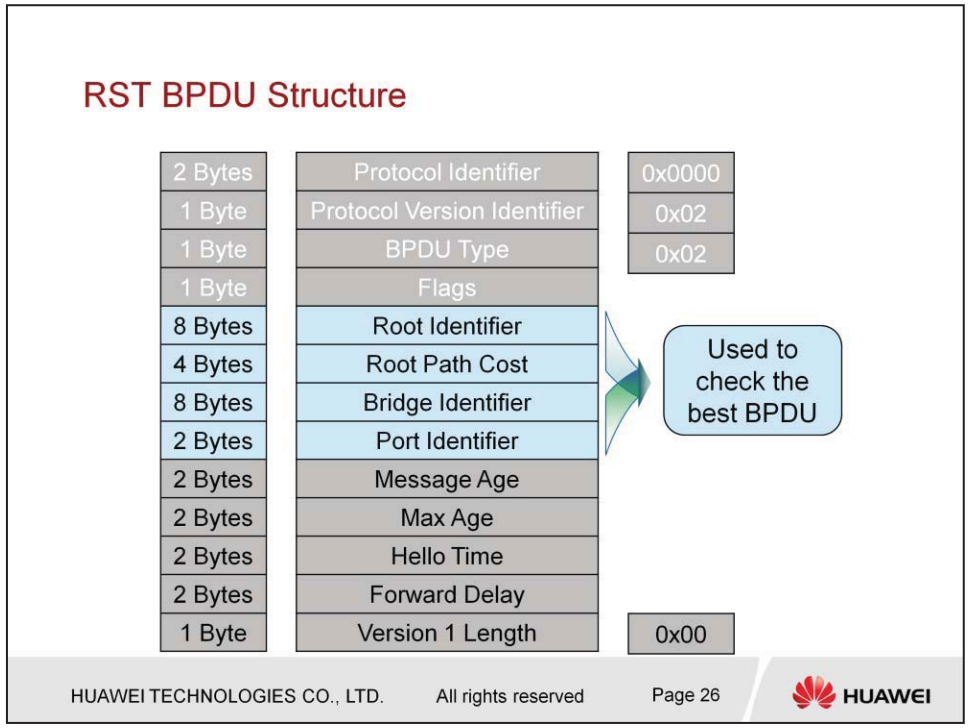
The information and parameters for calculating the spanning tree are encapsulated in the RST BPDU and transmitted between switches.

The RST BPDU is encapsulated in the Ethernet frame with the standard LLC format.

DMAC: destination MAC address.

The Ethernet frame used to encapsulate the RST BPDU uses the reserved malticast MAC address 01-80-C2-00-00-00. This MAC address identifies all switches but cannot be forwarded by the switches. That is, this MAC address is valid only on the local link.

LLC Header: the LLC Service Access Point (SAP) used by the RST BPDU is a binary value 01000010.



The structure of the RST BPDU differs from the configuration BPDU in STP in the following aspects:

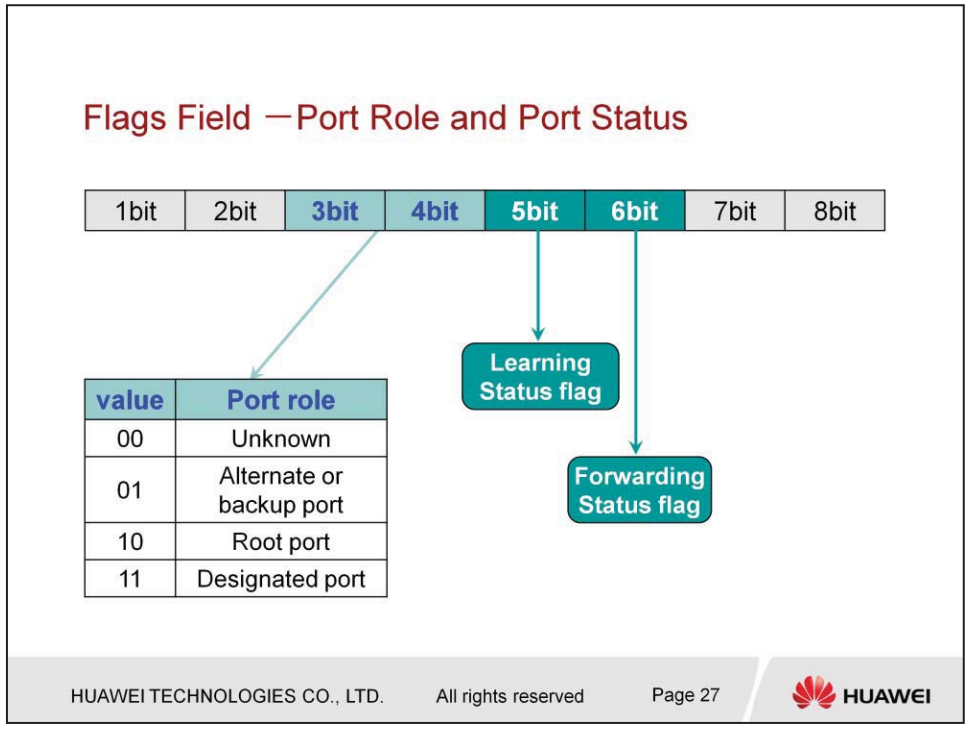
1. The Protocol Version Identifier field is 0x02, while the value in the configuration BPDU of STP is 0x00.

2. The BPDU type is 0x02, while the value in the configuration BPDU of STP is 0x00.

3. All the eight bits in the Flags field are defined (described later), while the configuration BPDU of STP uses only two bits to transmit the topology change message.

4. The RST BPDU has a new field Version 1 Length. The value is 0x00, indicating that the information about the BPDU in Version 1 is not included. The BPDU in Version 1 is defined by IEEE802.1G and is used for remote bridging.

The parameters used to check the BPDU with the highest priority and the timers in the RST BPDU are identical to equivalent parameters in the configuration BPDU of STP.

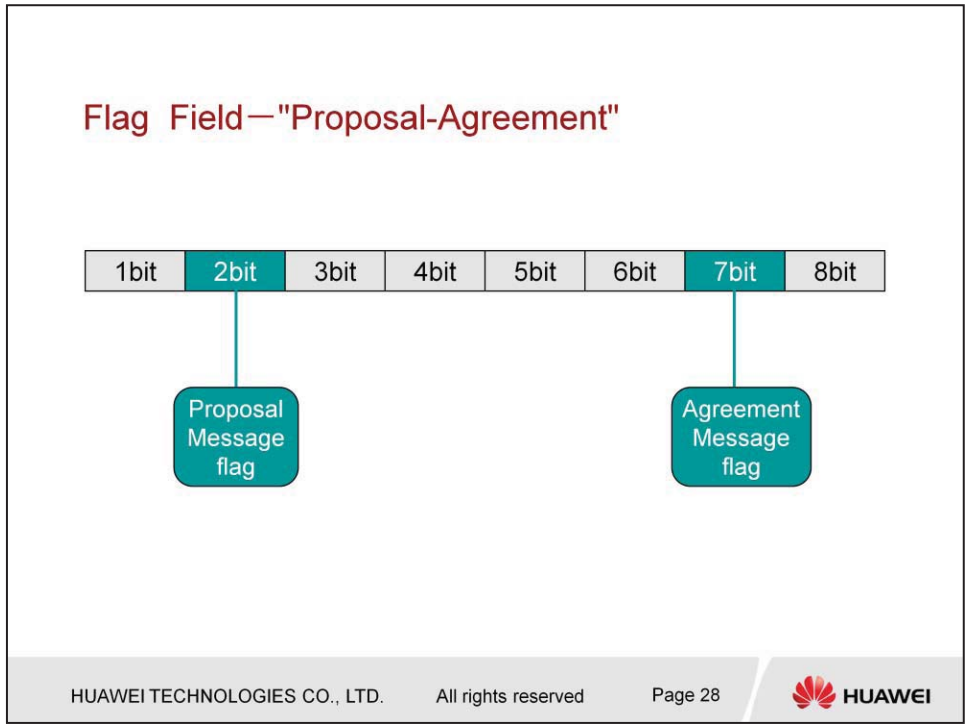


Different from STP, RSTP uses all the eight bits in the Flags field. The third and fourth bits identify the role of the port that sends this RST BPDU.

The fifth and sixth bits of the Flags field identify the status of the port that sends this RST BPDU.

When the fifth and sixth bits are both 0, it indicates that the port sending this RST BPDU is in Discarding state.

In STP, only the designated port can send the configuration BPDU. But in RSTP, non-designated ports can also send the RST BPDU, for example, the messages for the “Proposal-Agreement” negotiation mechanism.



Two types of messages, Proposal message and Agreement message, are used in the “Proposal-Agreement” negotiation mechanism. These messages are identified by the corresponding bits in the Flags field.

Apart from these flag bits, the bits identifying the port role and port status are also used in the “Proposal-Agreement” negotiation.

To initiate the negotiation, a new designated port sends an RST BPDU to the downstream port. The port role is identified as the designated port; the port status is identified as Discarding; the message type is identified as Proposal.

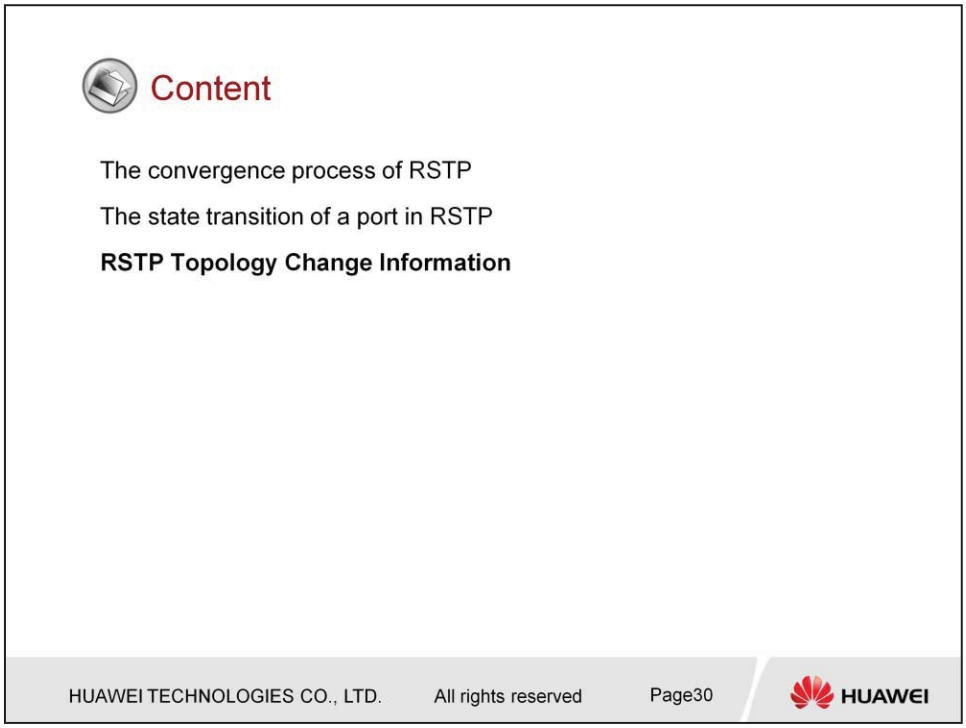
If the Proposal message sent by the designated port in Discarding state is received by a downstream root port, this message triggers further “Proposal- Agreement” flooding. That is, all non-edge designated ports switch to Discarding state and send Proposal messages to downstream ports.

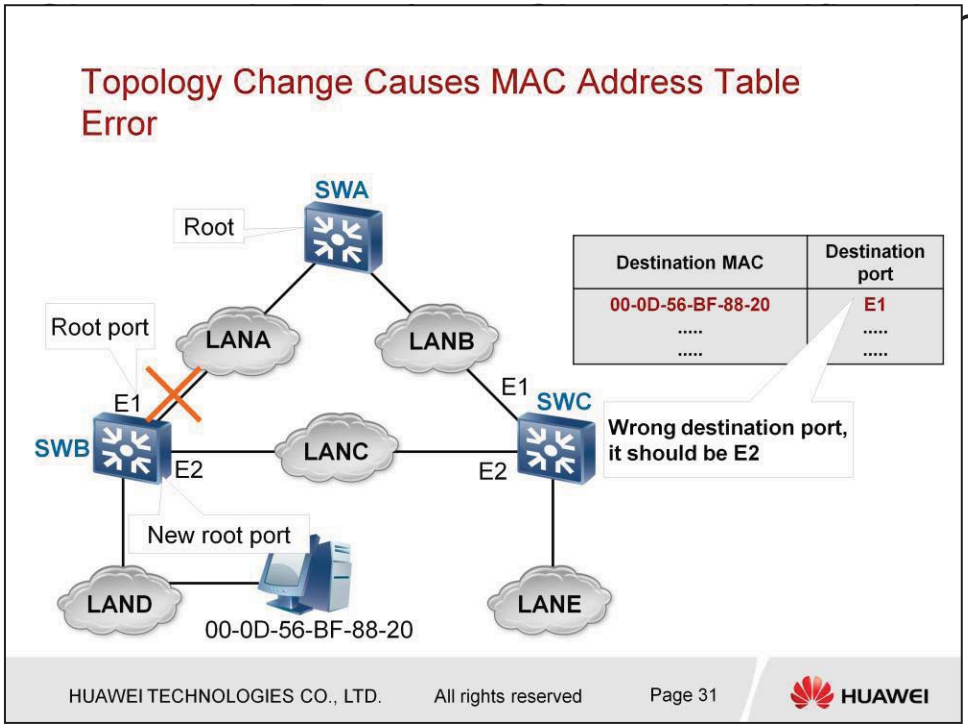
If the Proposal message sent by the designated port in Discarding state is received by a downstream alternate port or backup port, this downstream port returns an RST BPDU. The flag bits in the RST BPDU indicate that the port is an alternate port or backup port; the port status is Discarding; the message

type is Agreement.

When the designated port receives an Agreement message, it stops the “Proposal-Agreement” flooding process and switches to Forwarding state.

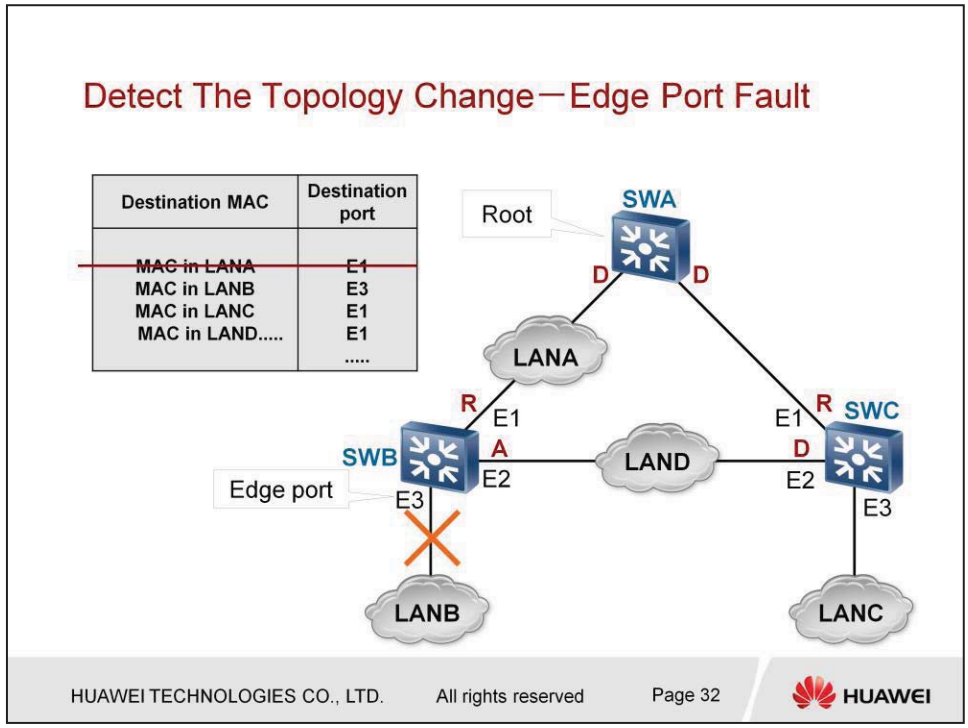
After all non-edge designated ports stop the flooding process, the root port (if any) sends the Agreement message to trigger the upstream designated port to stop flooding. Working principle of RSTP is more complicated than STP.RSTP uses more spanning tree priotity vector . Detail information is provided by Annex 2.


Content
The convergence process of RSTP
The state transition of a port in RSTP
RSTP Topology Change Information
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page30
HUAWEI



By default, the lifetime of dynamic entries in the MAC address table is 300seconds (5minutes).In this example: In a stable topology, messages from SWC reach a PC in LAND through port E1 of SWB.When E1 of SWB is disconnected, E2 becomes the new root port. The destination port for messages from SWC should changes to E2. However, a switch cannot detect the change of topology, so the MAC address table is incorrect. The data forwarding error caused by this fault may last for up to 5 minutes.STP responds to topology change by flooding the topology change messages in the entire network and changing the lifetime of MAC address entries to a smaller value (forward delay, 25 seconds by default).

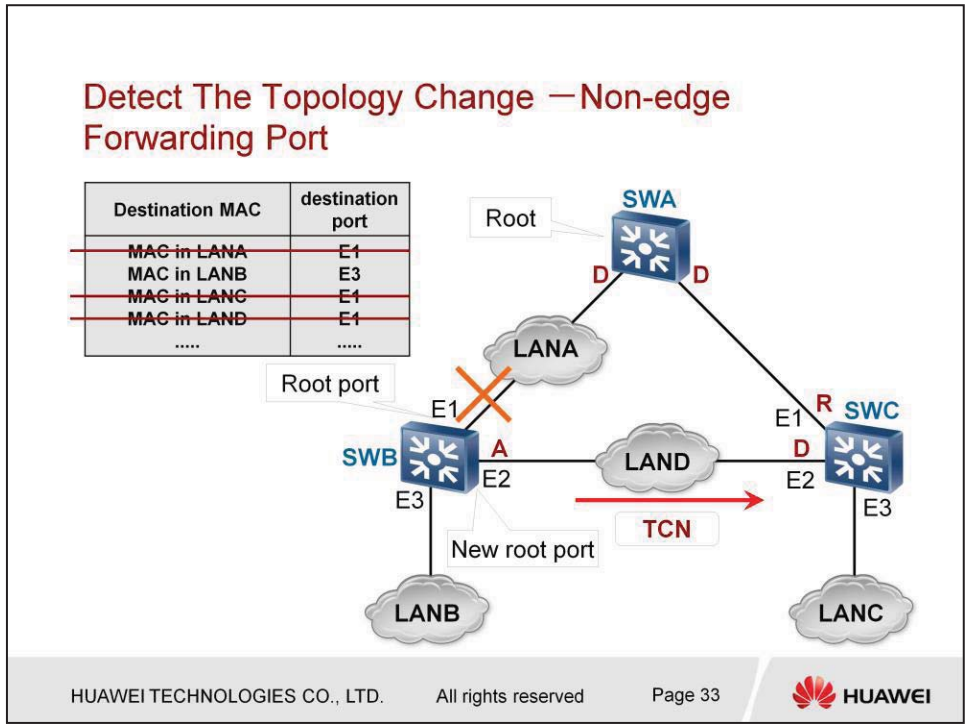
Different from STP, RSTP responds to topology change by deletingsome of MAC address entries.



When the port status changes, the topology change processing module of the switch treats the port based on the port role.

As shown in the figure, when SWB detects a fault on E3 of SWB, SWB deletes all entries with E3 as the destination port in the MAC address table.

If the status of an edge port changes, the switch does not send the topology change message to other switches.



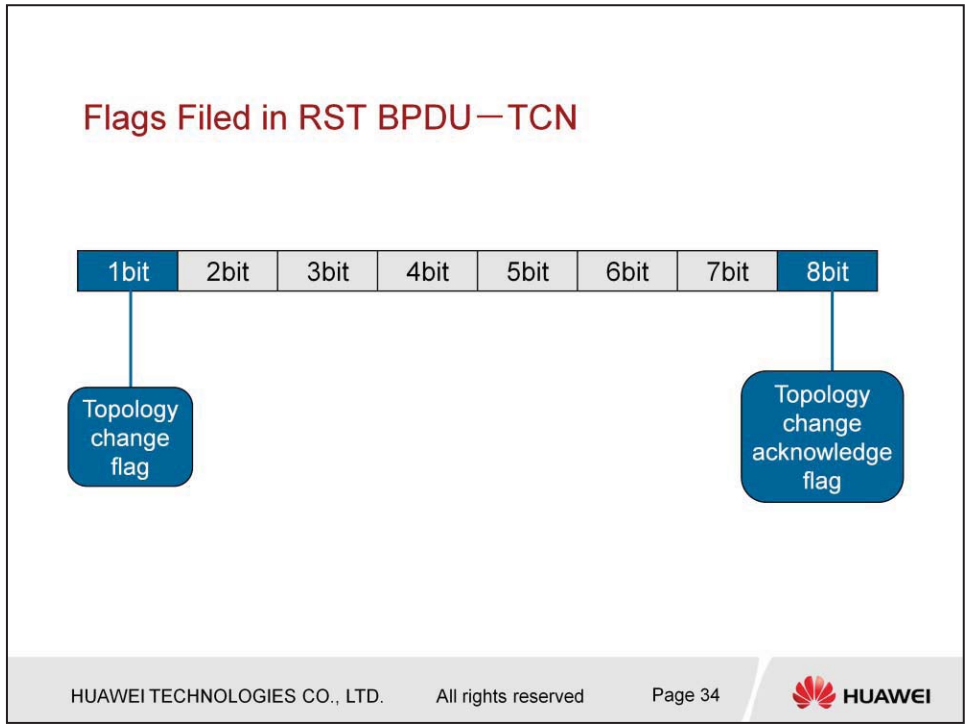
This figure shows the operation process of RSTP when a fault occurs to a nonedge port. As shown in the figure, E1 of SWB is the root port and E2 is the alternate port. When a fault occurs to E1, the operation process of SWB is as follows:

1. Deletes the entries with E1 as the destination port in the MAC address table.

2. Calculates a new spanning tree and elects E2 as the new root port.

3. After the new spanning tree is calculated (all ports that need to switch to Forwarding state complete status switchover), SWB sends the topology change notification through all non-edge ports in Forwarding state. Then other switches are notified that the network topology changes.

When a non-edge port switches from Discarding state to Forwarding state, the switch also considers that the network topology changes. In this case, the switch also sends the topology change notification through all non-edge ports in Forwarding state (including the port that just switches to Forwarding state). Actually, not only the fault on ports is regarded as topology change event.



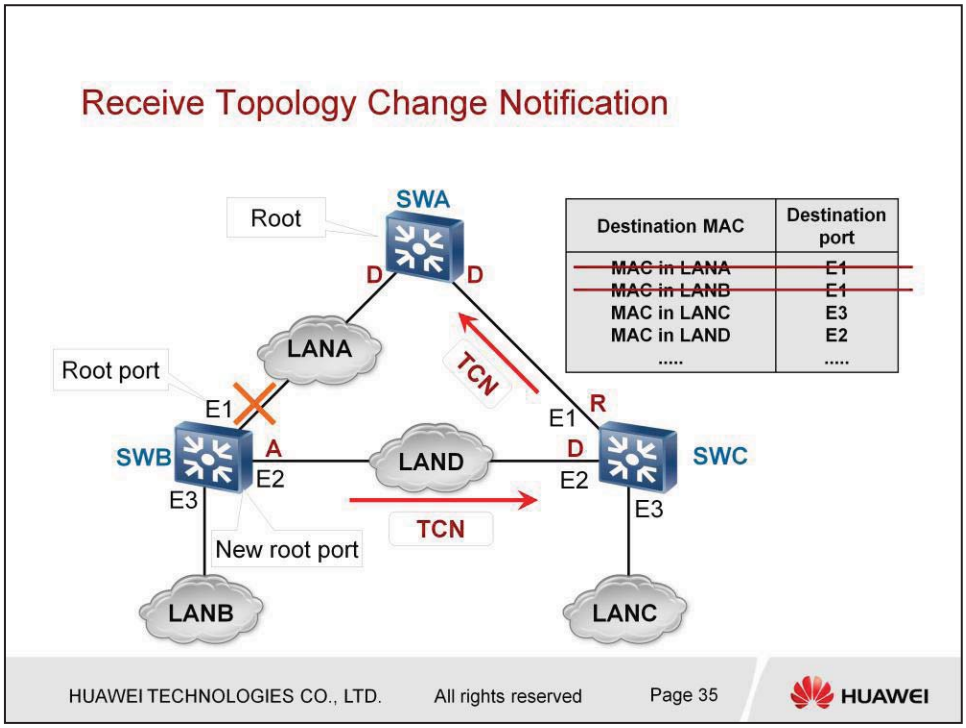
In the Flags field of RST BPDU, the first and eighth bits are identical to those in the configuration BPDU for STP. The first bit is the topology change flag, and the eighth bit is the topology change acknowledge flag.

In a network that contains only RSTP switches, an RSTP switch uses the RST BPDU with the topology change flag as the topology change notification to notify other RSTP switches that the topology changes.

In an STP-compatible network, an RSTP switch uses the topology change notification BPDU for STP to notify STP switches that the topology changes.

In RSTP, the eighth bit of the Flags field (topology change notification acknowledge flag) is not used.

When the RST BPDU with the topology change flag is used as the topology change notification, the topology change notification (TCN) timer is Hello timer plus 1 second. The default value is 3 seconds. Therefore, in an all-RSTP network,a switch generally sends only two TCNs.



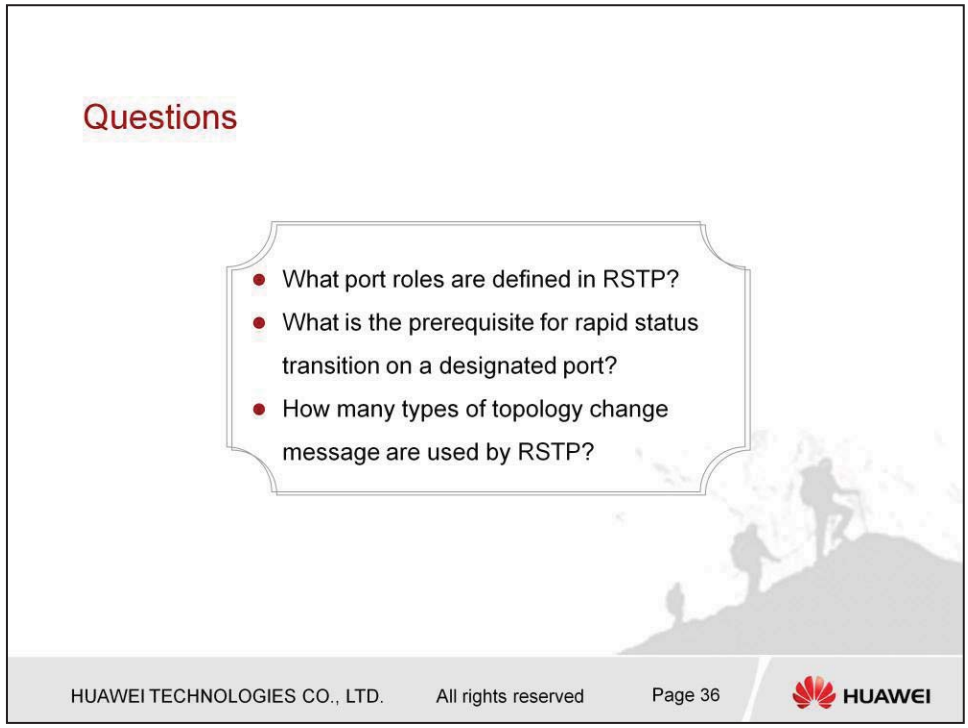
When an RSTP switch receives a TCN from a port in Forwarding state, it performs the following:

1. Calculates a new spanning tree (if needed). The reason is that the TCN is also an RST BPDU and it contains the parameters for calculating the spanning tree.

2. After the spanning tree is calculated and the port status is changed, the RSTP switch deletes the entries in which the destination port is in Discarding state. If the port that receives the TCN is in Forwarding state, the switch keeps the entries

with this port as the destination port and the entries with the edge designated port as the destination port. The entries with other ports in Forwarding state as the destination port are deleted.

3. Floods TCNs through other Forwarding ports that are neither the edge port nor the port receiving the TCN.



What port roles are defined in RSTP?

RSTP defines four port roles: root port, designated port, alternate port, and

backup port. The root port and designated port are in Forwarding state.

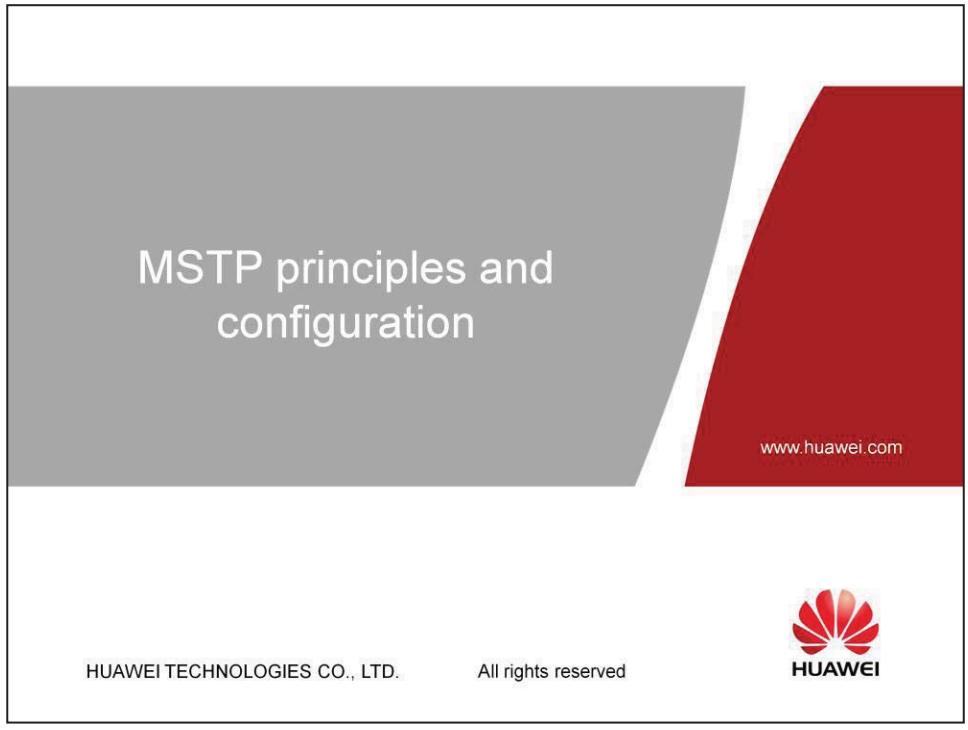
What is the prerequisite for rapid status transition on a designated port?

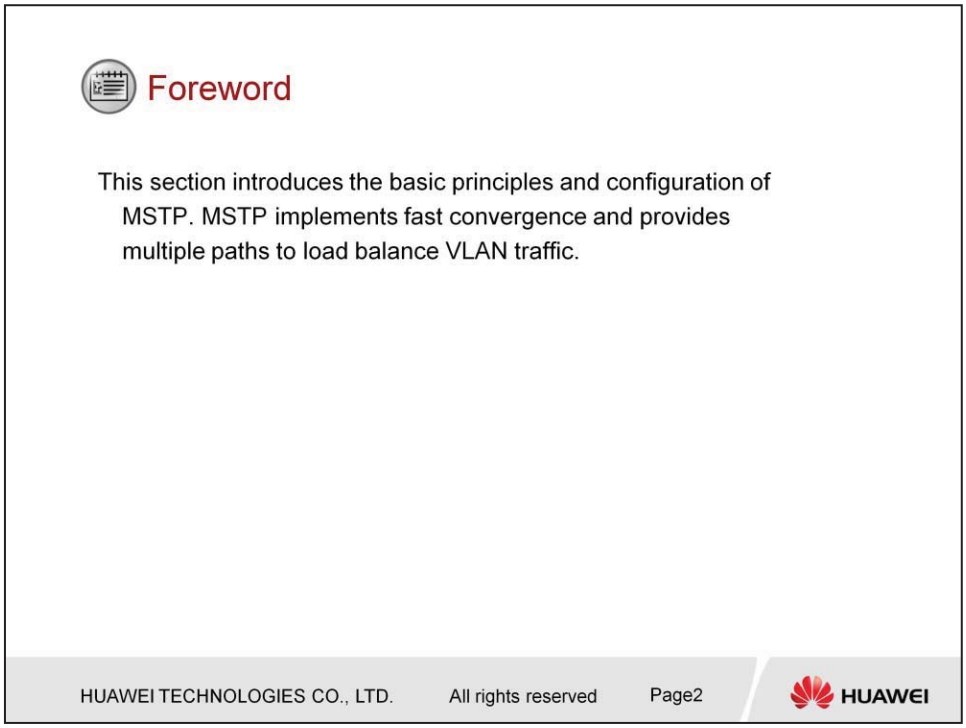
The link connected to the designated port is a point-to-point link.

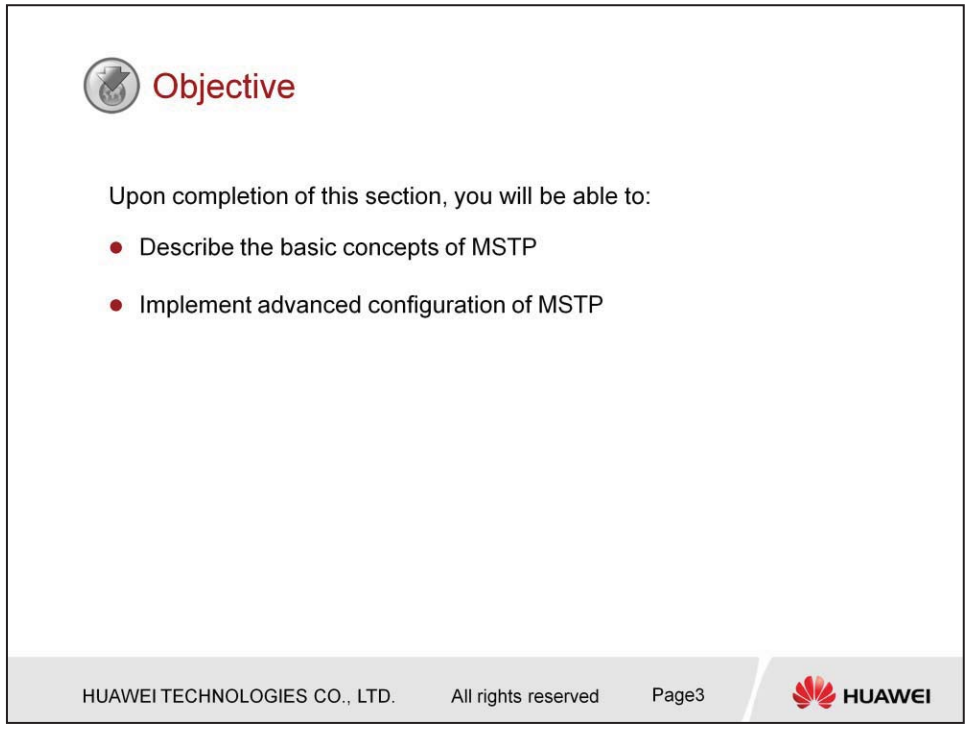
How many types of topology change messages are used by RSTP?

RSTP uses only one type of topology change message, namely, the RST BPDU

with the topology change flag.

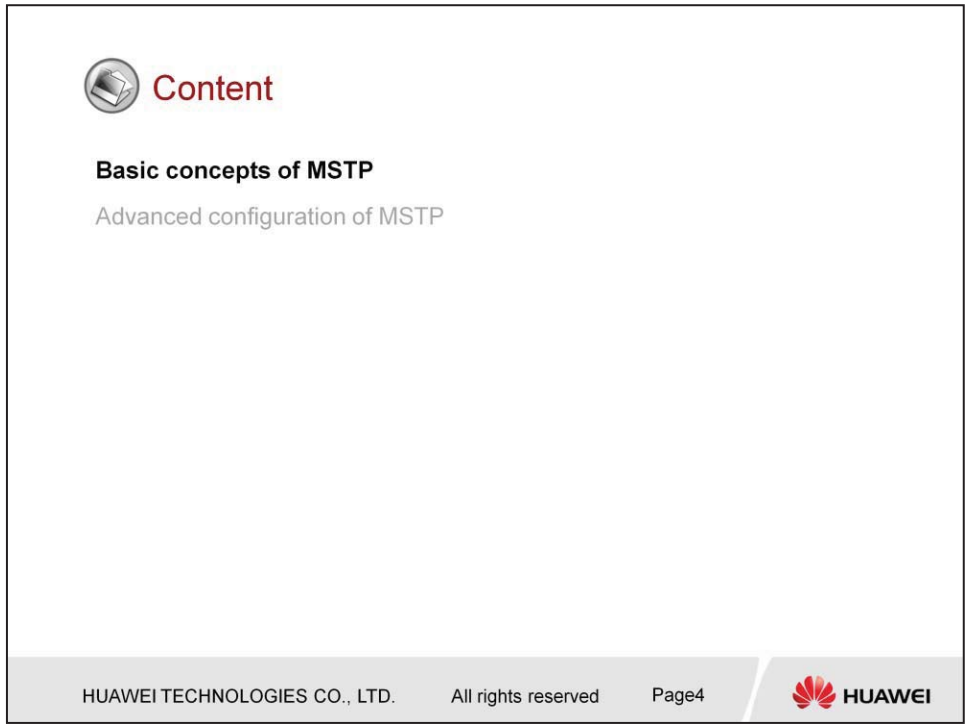

MSTP principles and configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

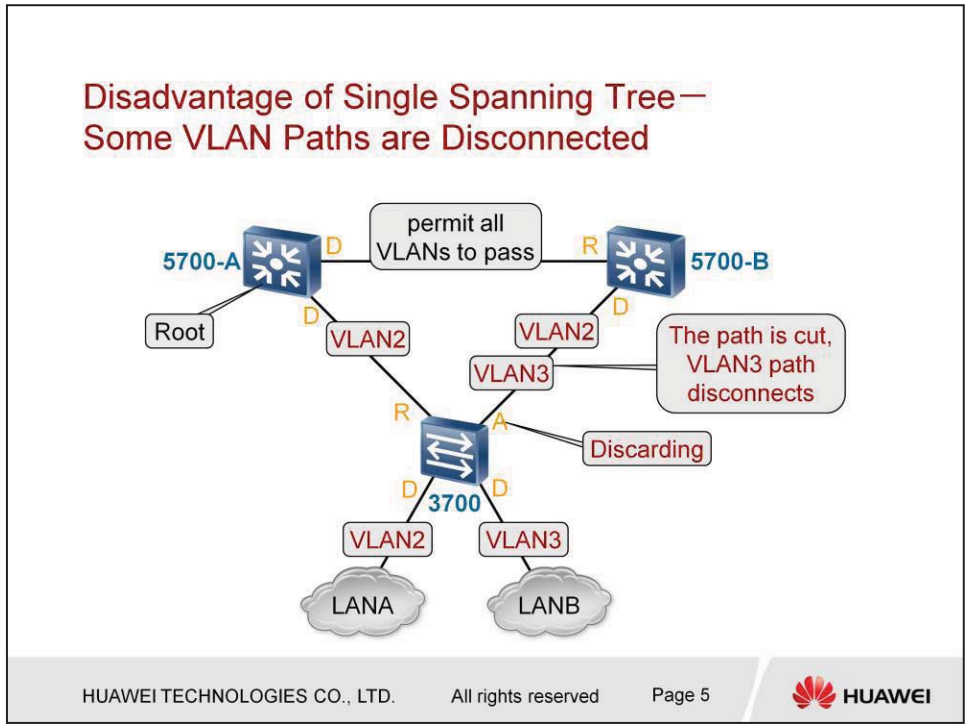

Foreword
This section introduces the basic principles and configuration of MSTP. MSTP implements fast convergence and provides multiple paths to load balance VLAN traffic.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page2
HUAWEI



Upon completion of this section, you will be able to:

- Describe the basic concepts of MSTP
- Implement advanced configuration of MSTP


Content
Basic concepts of MSTP
Advanced configuration of MSTP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page4
HUAWEI

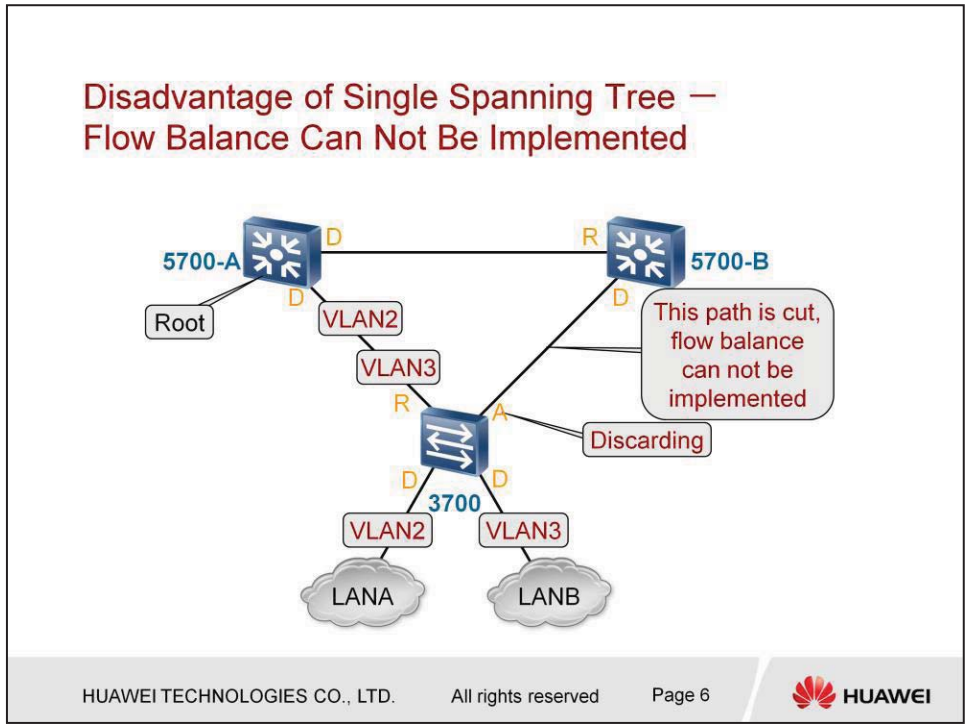


As shown in the figure, two terminal network segments are connected by a 3700 device. The 3700 device is connected to two 5700 devices through two links.

Packets from VLAN2 are sent to the upstream devices through two links, and packets from VLAN3 are sent to the upstream devices through only one link.

To avoid the loop problem of VLAN2, the system runs the spanning tree. In the case of the single spanning tree, assume that the 3700 interface connected to 5700-B becomes the alternate port and enters the Discarding status. In this case,

the link for upstream packets from VLAN3 is terminated and the packets cannot reach 5700-B.



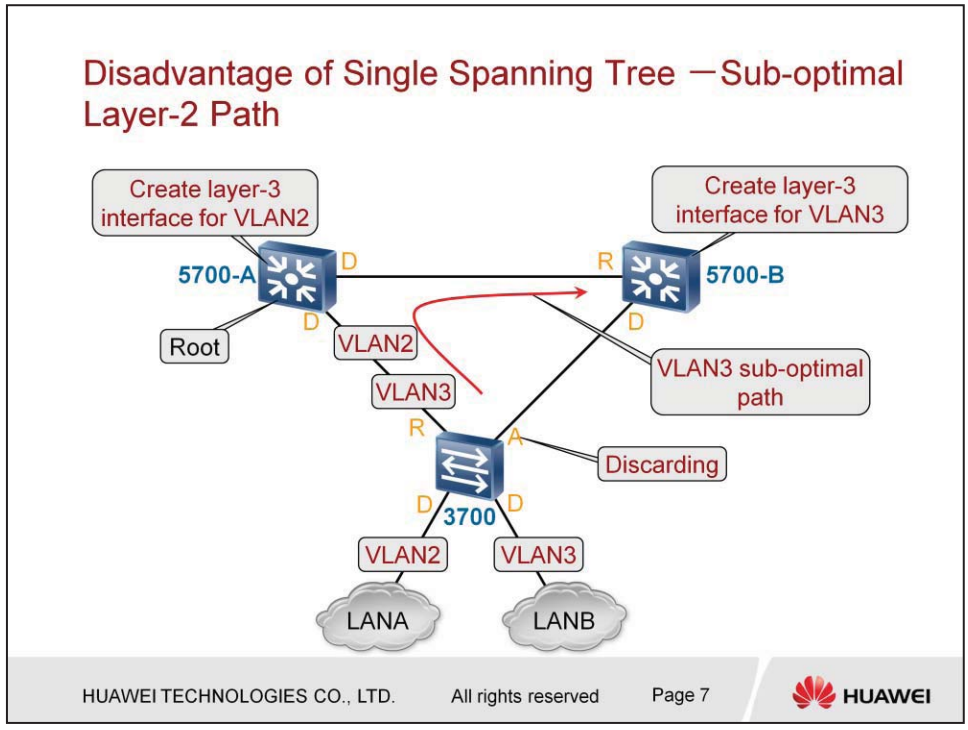
As shown in the figure, two terminal network segments are connected by a 3700 device. The 3700 device is connected to two 5700 devices through two links. The 5700 devices support hot backup and perform load balancing.

To implement hot backup between the 5700 devices, you need to configure the two upstream links of the 3700 as trunk links. Configure the trunk links to permit packets from all VLANs to pass through. Configure the link between two 5700

devices as a trunk link and configure the link two to permit packets from all VLANs to pass through.

Configure an interface on 5700-A as the layer-3 interface for VLAN2 and configure an interface on 5700-B as the layer-3 interface for VLAN3.

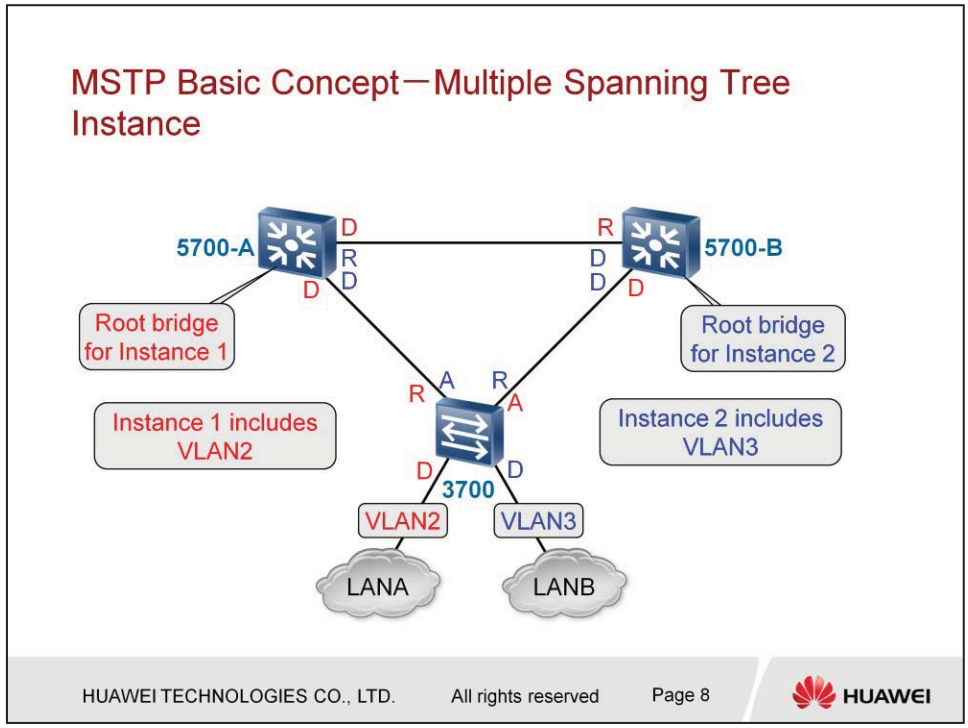
Packets from VLAN2 and VLAN3 should be sent to different layer-3 interfaces through different links. However, if only one spanning tree exists in the network, the loop formed by the two 5700 devices is torn down. Assume that the 3700 interface connected to 5700-B becomes the alternate port and enters the Discarding status. The data from VLAN2 and VLAN3 is sent to 5700-A through only one link, so load balancing cannot be implemented.



As shown in the figure, the link between the 3700 and two 5700 devices are configured as trunk links and permit packets from all VLANs to pass through. The link between two 5700 devices is also configured as a trunk link and permits packets from all VLANs to pass through.

After s single spanning tree is enabled, the loop formed by 3700 and two 5700 devices is torn down, and packets from VLAN2 and VLAN3 are all sent to 5700-A.

Configure the layer-3 interface for VLAN2 on 5700-A and configure the layer-3 interface for VLAN3 on 5700-B. For VLAN3, the link from 3700 to the layer-3 interface is the sub-optimal link. The optimal link is the one directly connecting 3700 to 5700-B.



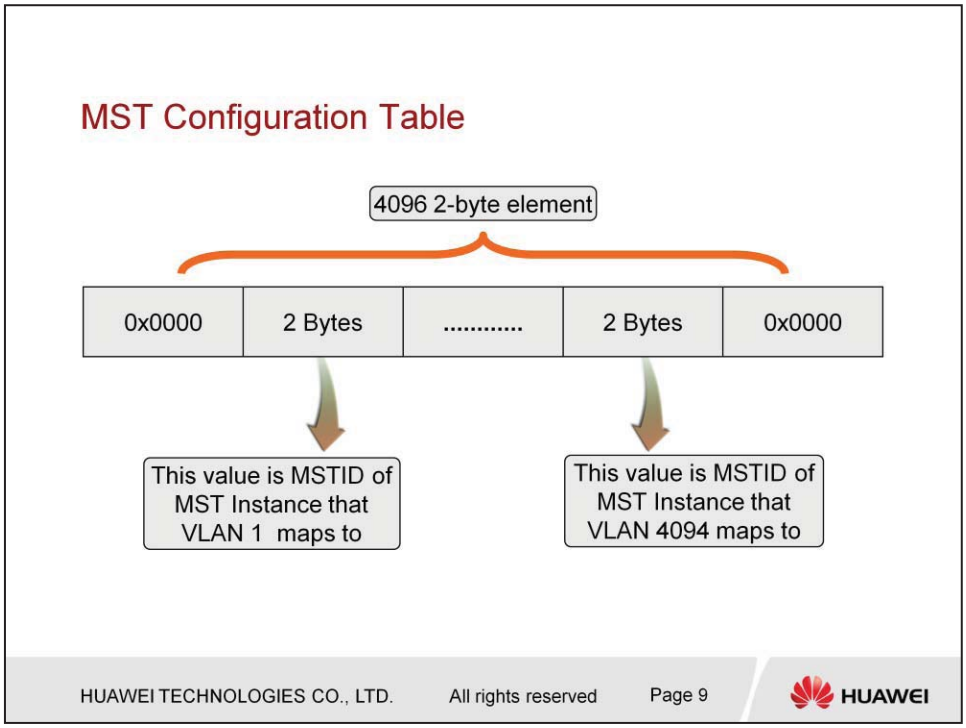
MSTP can map one or more VLANs to an MST instance. MSTP calculates the root bridge and sets the interface status for each MST instance individually. That is, MSTP calculates multiple spanning trees in the network.

The RSTP algorithm is used for each MST instance to calculate an individual spanning tree. Each MST instance has an MST ID, which is an integer of two bytes. The VRP supports 16 MST instances. The MST ID ranges from 0 to 15. By default, all VLANs are mapped to MST instance 0. In this example, two MST instances are configured in the network. VLAN2 is mapped to MST instance 1 and VLAN3 is mapped to MST instance 2.

By setting the priority of a switch for different MST instances, you can configure the root bridge for different MST instances.

In this example, 5700-A is configured as the root bridge for MST instance 1, and 5700-B is configured as the root bridge for MST instance 2.

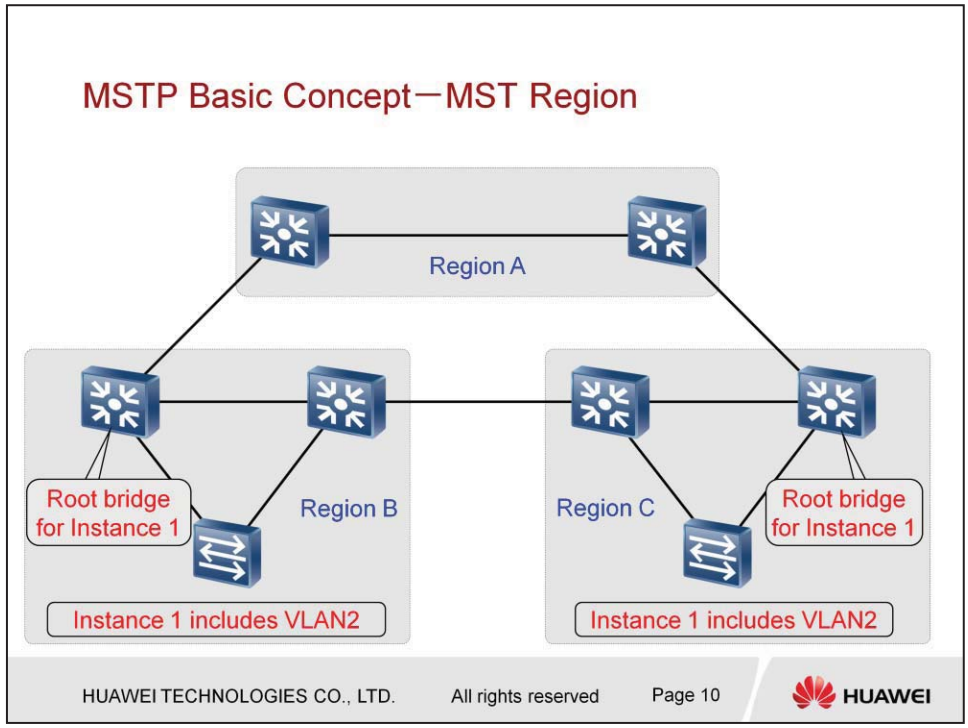
After multiple spanning trees are enabled, data from VLAN2 is sent to 5700-A directly and data from VLAN3 is sent to 5700-B directly. Thus, load balancing is implemented. The fault of unreachable path is removed and the problem of suboptimal path is also solved.



To identify the mapping between the VLAN and MST instance, the switch maintains an MST configuration table.

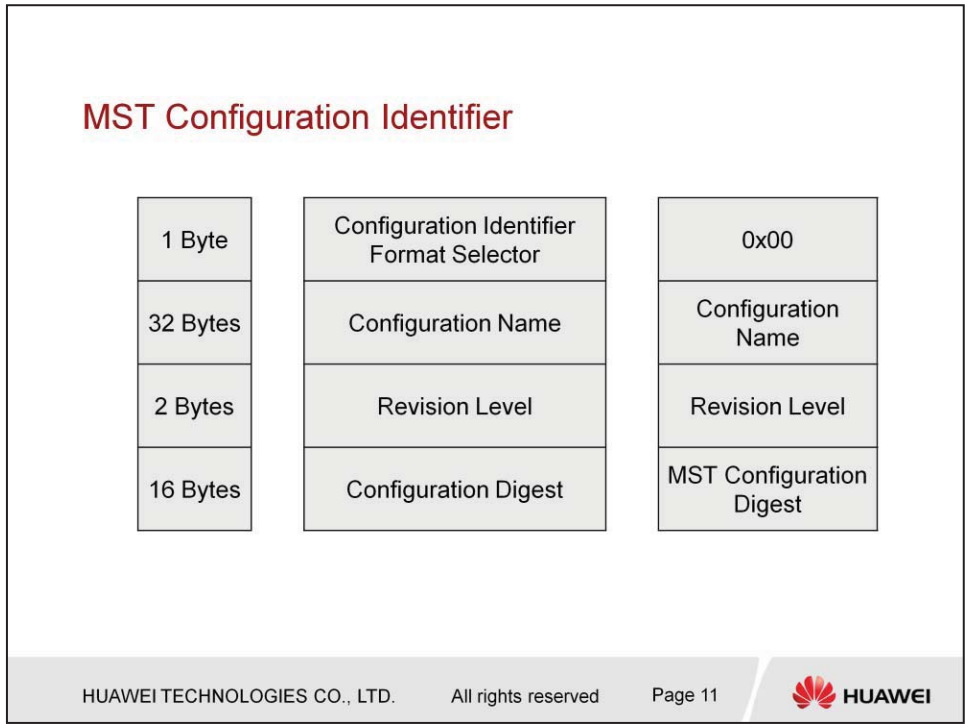
The MST configuration table consists of 4096 2-byte elements, which represent 4096 VLANs. The first and last elements are both 0. The second element indicates the MST ID to which the VLAN1 is mapped; the third element indicates the MST ID to which VLAN2 is mapped; the penultimate element (4095th element) indicates the MST ID to which VLAN4094 is mapped.

When the switch is initialized, all fields in the MST configuration table are set to 0, indicating that all VLANs are mapped to MST instance 0.



MSTP allows a group of adjacent switches to form an MST region. The switches in the same MST region have the same mapping entries between VLANs and MST instances.

Except instance 0 (described later), MST instances in each MST region calculate the spanning tree independently, no matter whether the instances are mapped to the same VLAN or whether the packets of the VLAN pass the link between MST regions. The calculation of the spanning tree for each MST instance does not affect one another.



A switch identifies the region that it belongs to through the MST configuration identifier.

The MST configuration identifier is encapsulated in the BPDU transmitted between switches. As shown in the figure, the MST configuration identifier consists of four parts. Switches with identical values in the four parts are regarded as in the same region.

Configuration Identifier Format Selector: It contains one byte and has a fixed value 0.

Configuration Name: It the MST region name assigned to a switch. It contains 32 bytes. Each switch is assigned an MST regions name. By default, the MST region name is the MAC address of the switch.

Configuration Digest: It contains 16 bytes. Switches in the same region must maintain the same mapping table for the VLAN and MST instance. However, the size of the MST configuration table is large (8192 bytes) and cannot be transmitted between switches. This filed is the digest calculated from the MST

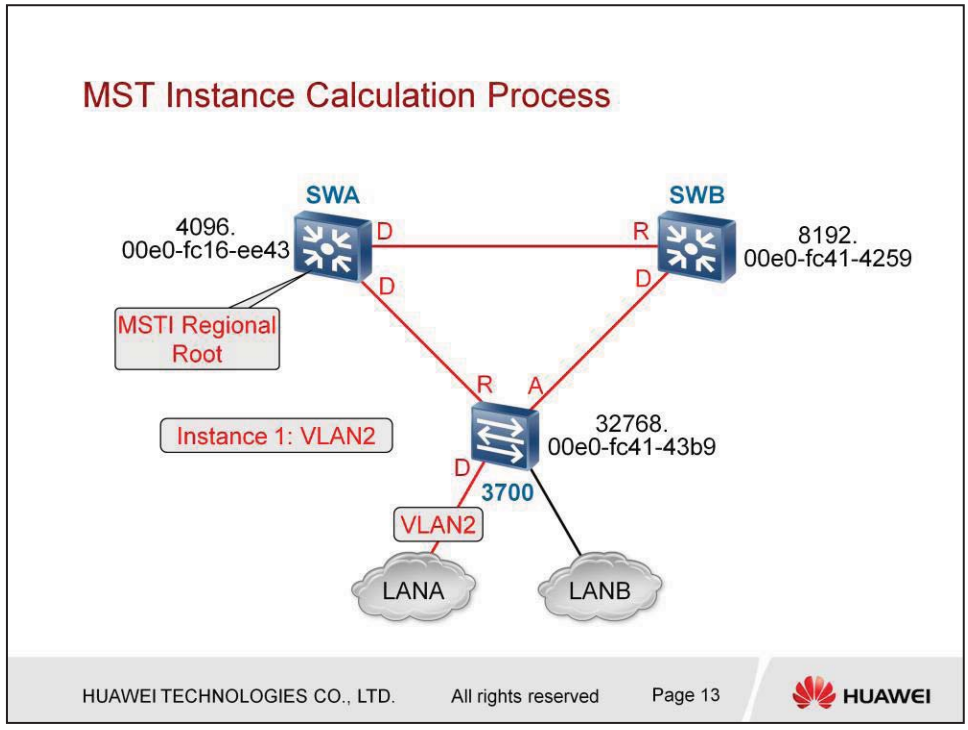
configuration table through the MD5 algorithm.

Revision Level: It contains two bytes and the default value

contains all 0s.

Different MST configuration tables may have the same configuration digest, although the probability is small. In this case, switches in different regions may be regarded as in the same region. The Revision Level field is an additional identifier.

You are recommended to set different values for different regions to avoid the above-mentioned error.



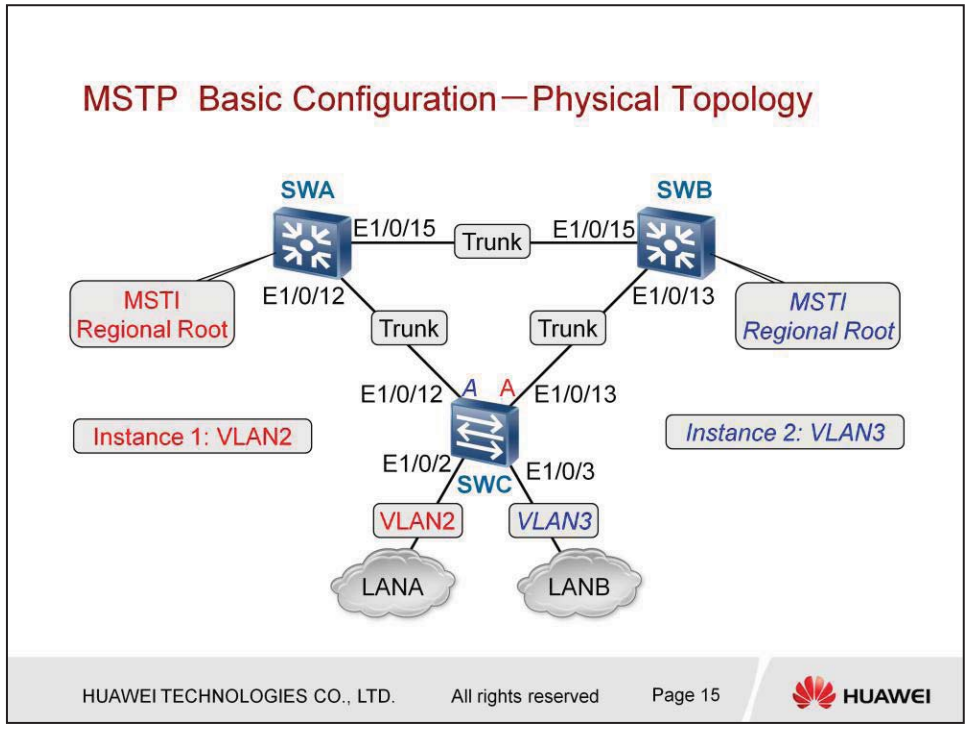
Here, the MST instance refers to the MST instance from 1 to 15, namely, the MST instances except MST instance 0. Instance 0 will be described in later Chapters.The calculation process of each MST instance is the same as the RSTP calculation process. The difference is that they use different terms.

1. First, MSTI regional root is elected for the MST instance. The MSTI regional root is equivalent to the root bridge in RSTP calculation. The election is based on the bridge identifiers configured in the MST instance. Similar to the bridge identifier in RSTP, the MST bridge identifier consists of the priority and MAC address of the switch. A smaller value indicates a higher priority.

2. The non-root bridge in this MST instance elects a root port, which provides the optimal path to the MSTI regional root of this MST instance. The root port election is based on the internal root path cost, which is the cost from a switch to MST region root within an MST region. If the paths provided by multiple ports have the same cost, the system compares the identifiers of the upstream switches, identifiers of the upstream switch ports, and the identifiers of the receiving ports in sequence, and then elects the optimal path.

3. The designated port of each network segment provides the optimal path to the corresponding MSTI regional root for this network segment.

4. The basis for electing the alternate port and backup port is the same as the election basis used for RSTP.

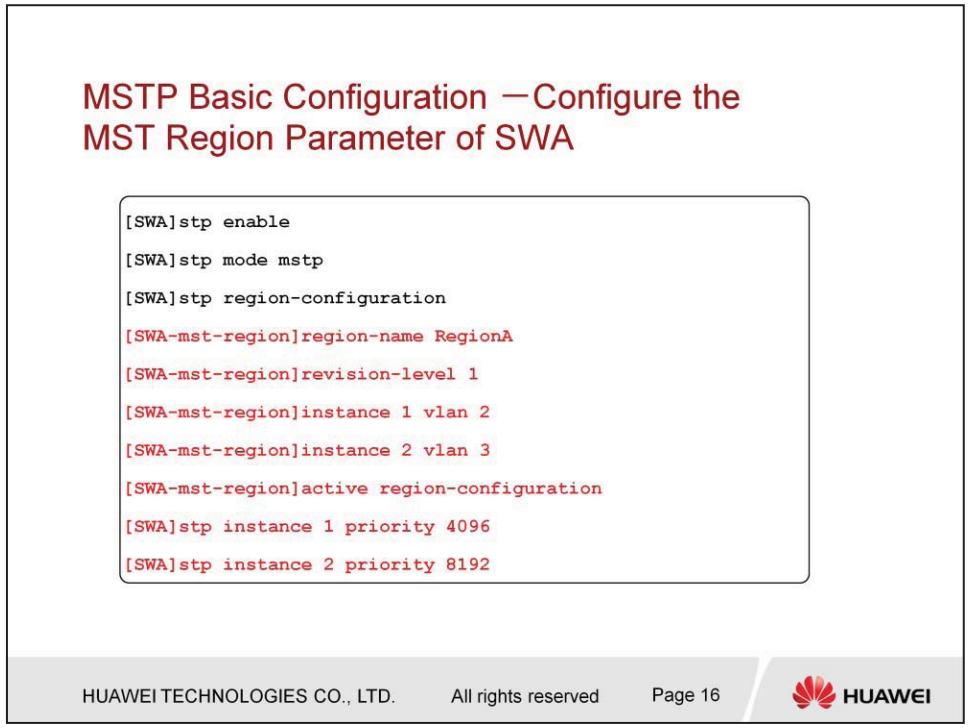


As shown in the figure, all links between the switches are configured as the trunk link. Each trunk link permits packets of all VLANs to pass through.

The three switches belong to the same MST region. The configuration name (MST region name) is “RegionA”; the revision level is 1. Two MST instances are created in this region. Instance 1 is mapped to VLAN 2 and Instance 2 is mapped to VLAN 3.

Change the priority of a switch in different instances to configure SWA as the root bridge of Instance 1 and configure E1/0/13 of SWC as the alternate port of Instance 1. Similarly, configure SWB as the root bridge of Instance 2 and configure E1/0/12 of SWC as the alternate port of Instance 2.

After the configuration, packets from VLAN2 and packets from VLAN3 are sent through different links. Thus, load balancing is implemented and the two upstream links function as the backup of each other.



stp { enable | disable }

The stp command is used to enable or disable the MSTP feature in the entire network or on an interface. By default, the MSTP feature is disabled on the switch.

stp mode { stp | rstp | mstp }

The stp mode command is used to configure the working mode of the switch. By default. The switch works in MSTP mode.

The stp region-configuration command is used to enter into the MST region view.

region-name name

name: specifies the configuration name of the switch. The value is a string of 1 to 32 characters. By default, the configuration name of a switch is the MAC address of this switch.

revision-level level

level: specifies the MSTP revision level. The value ranges from 0 to 65535. By default, the MSTP revision level is 0.

instance instance-id vlan vlan-list

The instance command is used to map the specified VLAN list to an MST instance. By default, all VLANs are mapped to Instance

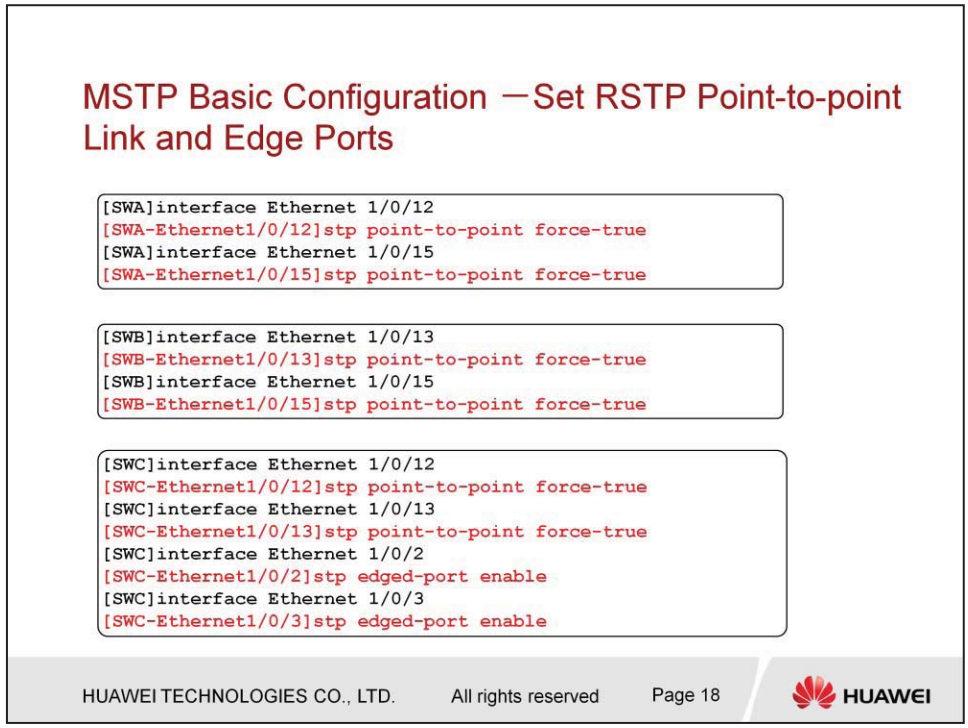
0.

active region-configuration

The active region-configuration command is used to activate the configuration of the MST region.

stp [instance instance-id] priority priority

The stp priority command is used to set the priority of a switch in the specified MST instance. The value is the integer multiple of 4096. The default priority for each instance is 32768.



Each MST instance uses the RSTP algorithm to calculate the spanning tree independently. The rapid convergence mechanism of RSTP is valid for each MST instance.

Similar to the configuration of RSTP, configure all links between the switches as the point-to-point link. Configure E1/0/2 and E1/0/3 of SWC as the edge port.

stp point-to-point { force-true | force-false | auto }

force-true: indicates that the link connected to the current Ethernet interface is a point-to-point link.

force-false: indicates that the link connected to the current Ethernet interface is not a point-to-point link.

auto: automatically checks whether the link connected to the current Ethernet interface is a point-to-point link.

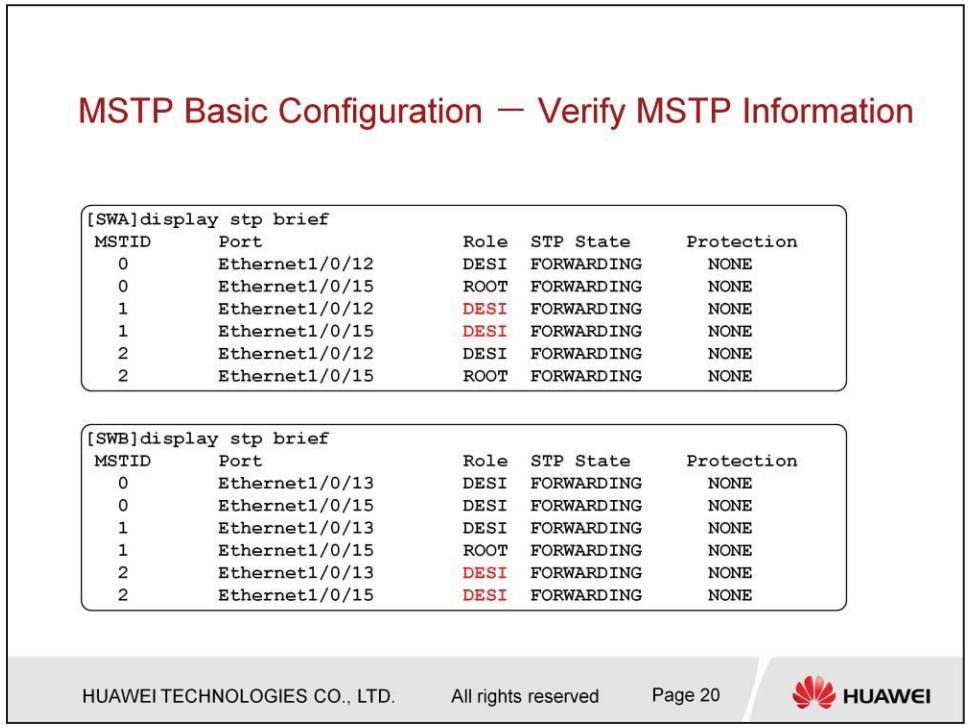
By default, the auto keyword is used. When the system detects that the interface works in full-duplex mode, it regards the link connected to this interface as a point-to-point link. When the system detects that the interface works in half-duplex mode, it considers that the link connected to the interface is not a point-to-point link. Here, the link is configured as a point-to-point link by command.

stp edged-port { enable | disable }

The stp edged-port enable command is used to configure the current Ethernet interface as the edge port.

The stp edged-port disable command is used to configure the current Ethernet interface as a non-edge port.

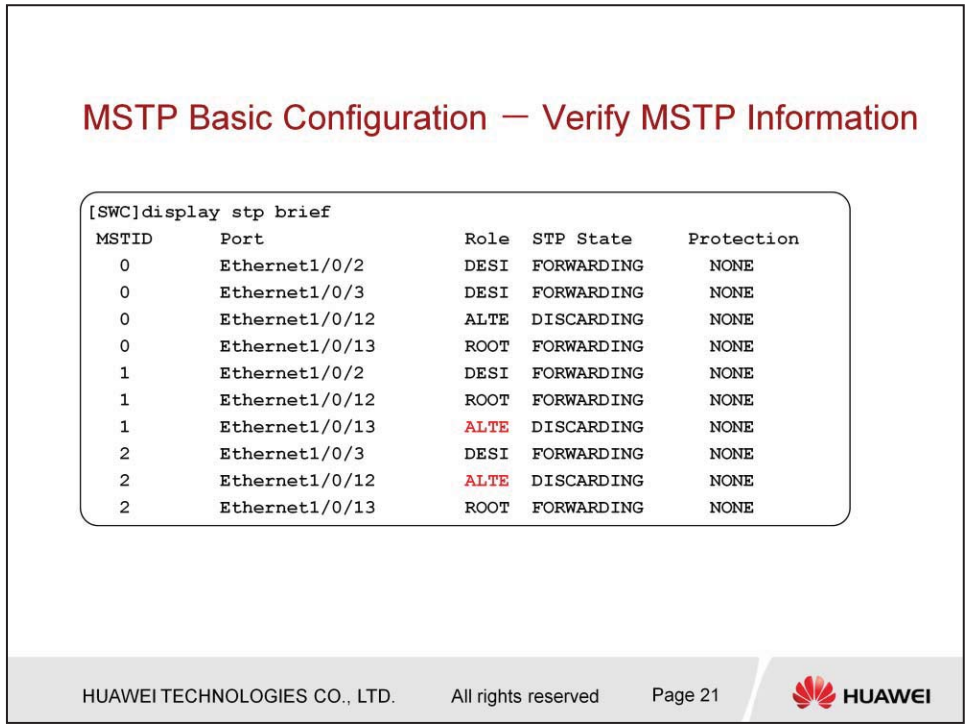
By default, all Ethernet interfaces of a switch are configured as non-edge interfaces.



The information about Instance 0 is omitted here (it will be described later).

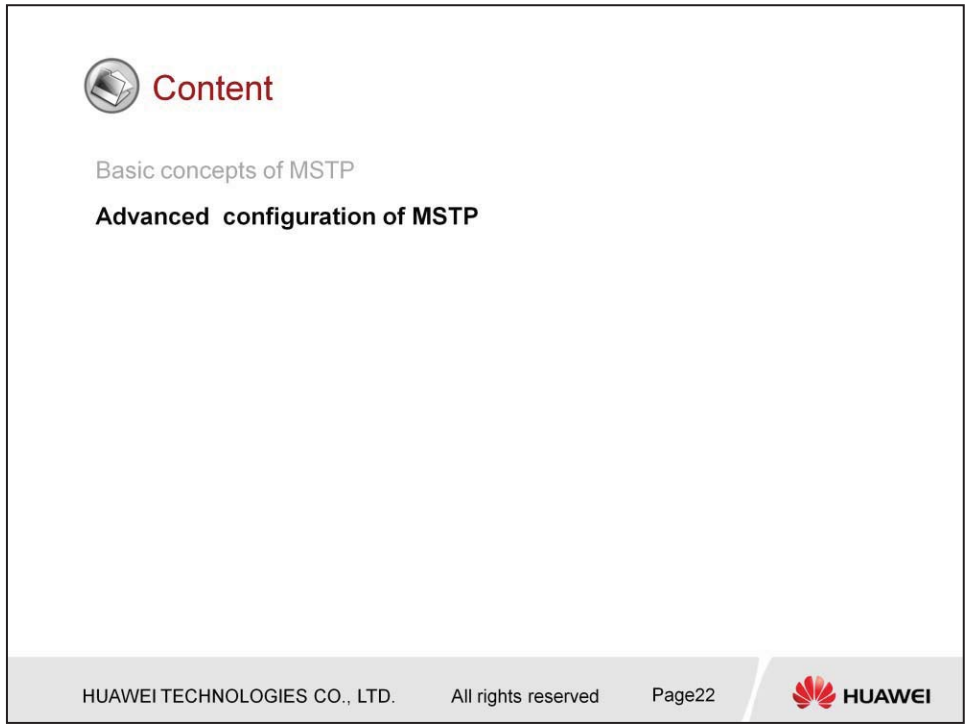
On SWA, two ports in Instance 1 are both the designated ports, which indicates that the SWA is the root bridge of Instance 1.

On SWB, two ports in Instance 2 are both the designated ports, which indicates that the SWB is the root bridge of Instance 2.



The information about Instance 0 is omitted here (it will be described later).

In Instance 1, E1/0/13 of SWC is the alternate port and the status is in the Discarding status. In Instance 2, E1/0/12 of SWC is the alternate port and the status is in the Discarding status.


Content
Basic concepts of MSTP
Advanced configuration of MSTP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page22
HUAWEI

STP Mode

```
[SWB]stp mode stp
[SWB]stp mode rstp
[SWB]stp mode mstp
```

Work mode	Description
STP	Can only communicate with STP bridge, receive and transfer configuration BPDUs.
RSTP	If it detects that peer bridge is running STP, it will run STP.
MSTP	If it detects that peer bridge is running RSTP, it will run RSTP; if it detects that peer bridge is running STP, it will run STP.

```
[SWB]stp mcheck
```

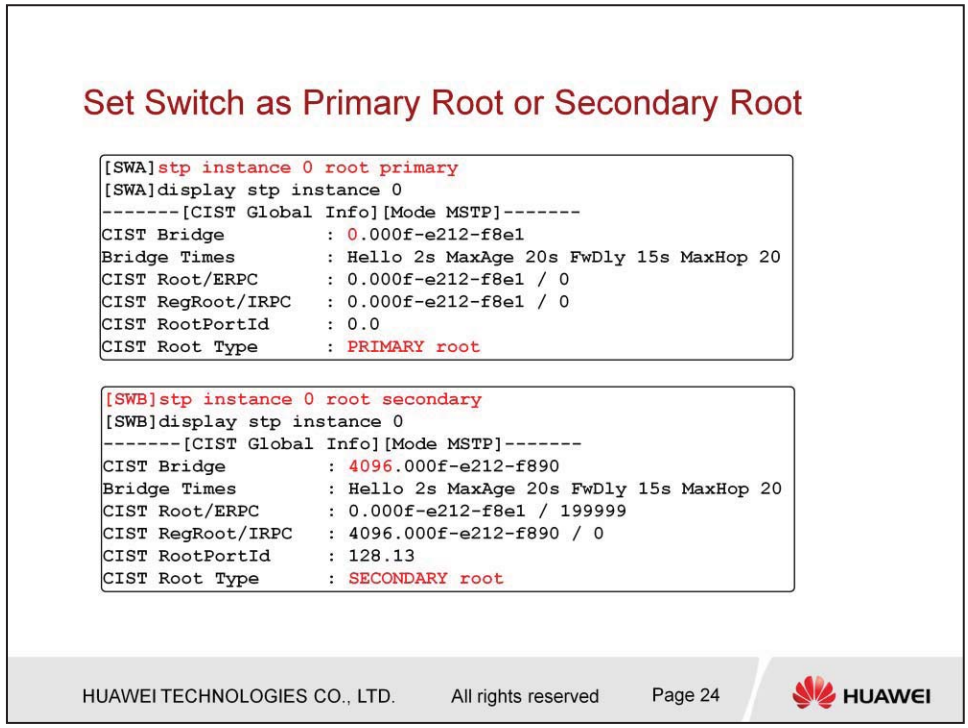
  HUAWEI

The working mode can be configured in the system view or interface view.

MSTP is compatible with RSTP, and RSTP is compatible with STP.

If the port of the MSTP switch was connected to the STP/RSTP switch, the port switched to the STP/RSTP compatible working mode. If the STP/RSTP switch is shut down or removed from this port, the port cannot automatically switch to the MSTP working mode. If you perform the mcheck operation of the port, the port will switch to the MSTP mode again.

Using the stp mcheck command, you can perform the mcheck operation on the current port.



The VRP platform can configure the switch as the primary root bridge or secondary root bridge, so you need not configure the priority manually.

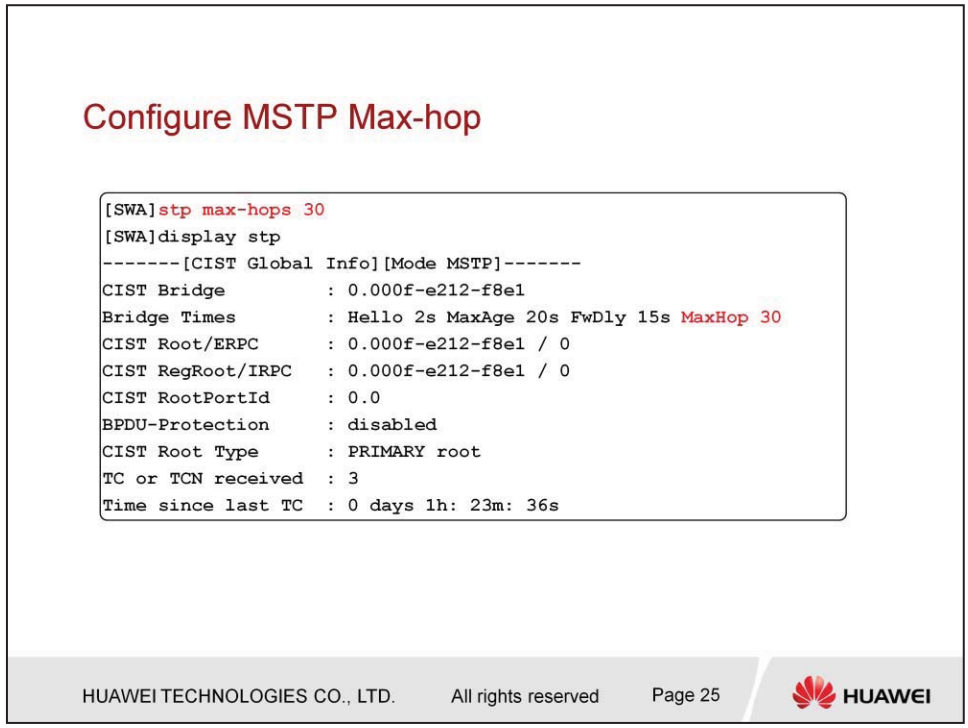
stp [instance instance-id] root primary

After you run this command, the switch priority is automatically changed to 0 in the specified instance. The switch becomes the primary root bridge of this instance.

stp [instance instance-id] root secondary

After you run this command, the switch priority is automatically changed to 4096 in the specified instance. The switch becomes the secondary root bridge of this instance. When the primary root bridge is Down, the secondary root bridge becomes the primary one immediately.

These commands provide another method of changing the priority.

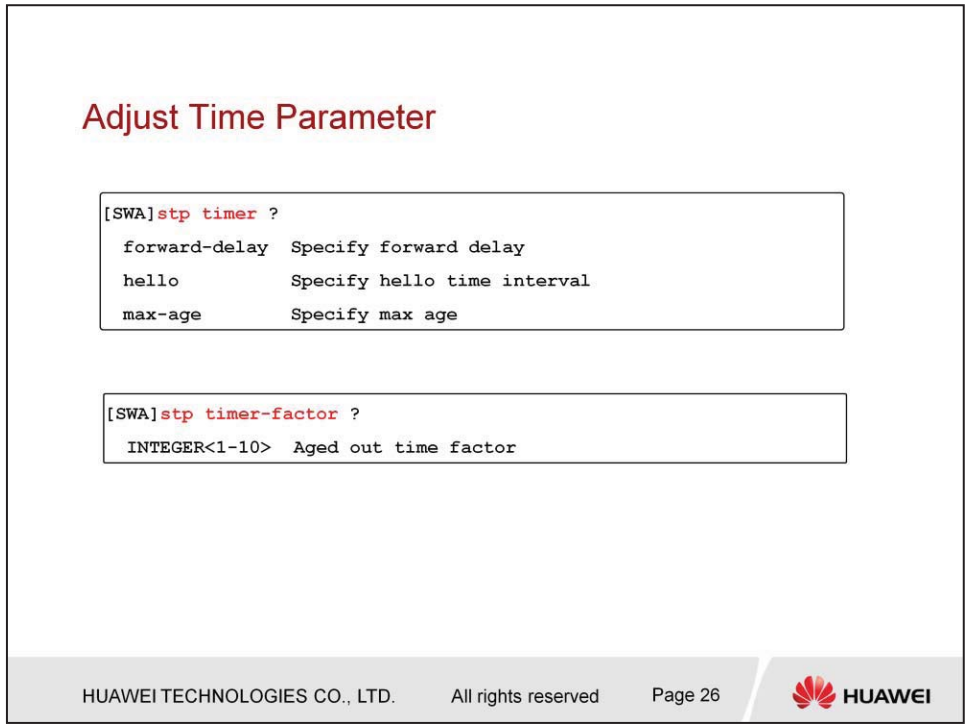


MSTP supports the setting of max hops. The MST BPDU contains a CIST remaining hops field, which is similar to the TTL field in the IP packet.

When a MST BPDU is sent from the MST root, the CIST remaining hops field is set to the max hops set on this MST root. When a non-root switch receives an MST BPDU from the upstream switch, it reduces the value of CIST remaining hops by 1, and then generates its own MST BPDU and sends it to the downstream switch.

When a switch receives an MST BPDU in which the value of CIST remaining hops is 0, it discards this MST BPDU. Therefore, the switches beyond the max hops are not included in the spanning tree, thus, the size of the MST region is Limited.

The stp max-hops command is used to set the max hops of the MST region on the switch. This command is applicable to the MST regional root. The VRP supports 1 to 40 hops. By default, the max hops is 20.



stp timer forward-delay centi-seconds centi-seconds: specifies the forward delay. The value ranges from 400 to 3000, in

centiseconds. The default value is 15 seconds.

stp timer hello centi-seconds

centi-seconds: specifies the Hello timer. The value ranges from 100 to 1000, in centiseconds. The default value is 2 seconds.

stp timer max-age centi-seconds

centi-seconds: specifies the max age. The value ranges from 600 to 4000, in centiseconds. The default value is 20

seconds.

stp timer-factor number

number: specifies the STP timeout factor. Multiplying the Hello timer by this factor, you can set the timeout duration. The value ranges from 1 to 10, and the default value is 3. That is, if the switch does not receive the BPDU from the specified port of the connected network segment within 3 times of the Hello timer

(6 seconds), the switch regards the port as Down and calculates a new spanning tree.

The Relation between the Network Diameter and Time Parameter

Network diameter	Hello Timer	Max Age	Forward Delay
2	2s	10s	7s
3	2s	12s	9s
4	2s	14s	10s
5	2s	16s	12s
6	2s	18s	13s
7	2s	20s	15s

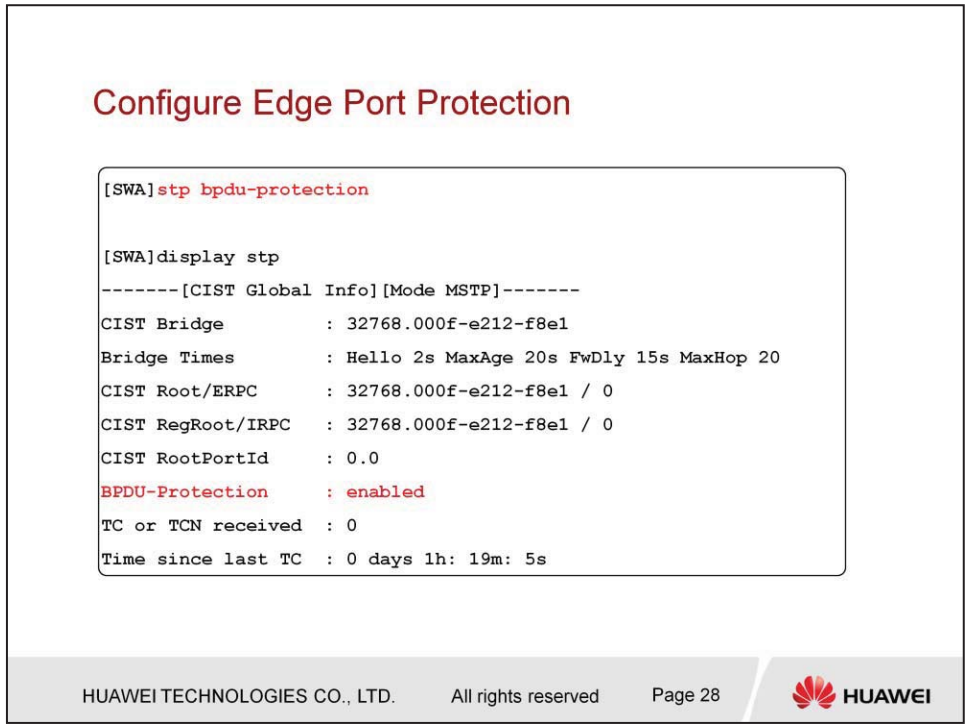
```
[SWA]stp bridge-diameter ?
  INTEGER<2-7>  Bridge diameter
```


If the max age or forward delay is improper, the loop may be generated in the network, or the network will be Down for a long time after the topology changes.

The VRP can automatically calculate the rational max age and forward delay according to the network diameter and the Hello timer. This table lists the max age and forward delay calculated by the VRP according to the network diameter when the Hello timer is 2 seconds.

stp bridge-diameter bridgenum

bridgenum: specifies the diameter of the switching network. The value ranges from 2 to 7. The default value is 7.



In general, the edge port cannot receive the BPDU. If the edge port receives a forged BPDU with higher priority, the switch calculates the new spanning tree, thus causing the topology flapping.

The stp bpdu-protection command is used to enable edge port protection. If the edge port receives a BPDU, the port considers that the switch is attacked. In this case, the edge port is shut down automatically and need be started by the network administrator manually.

By default, edge port protection is disabled.

Configure Designated Port Protection of Root Bridge

```
[SWA]stp interface Ethernet 1/0/13 root-protection

[SWA]display stp interface Ethernet 1/0/13

----[CIST][Port13(Ethernet1/0/13) ][FORWARDING]----
 Port Protocol        : enabled
 Port Role            : CIST Designated Port
 Port Priority        : 128
 Port Cost(Dot1T)      : Config=auto / Active=199999
 Desg. Bridge/Port    : 0.000f-e212-f8e1 / 128.13
 Port Edged           : Config=disabled / Active=disabled
 Point-to-point       : Config=auto / Active=true
 Transit Limit        : 3 packets/hello-time
 Protection Type      : Root
 Num of Vlans Mapped : 1
 PortTimes            : Hello 2s MaxAge 20s FwDly 15s RemHop 0
 BPDU Sent            : 18
          TCN: 0, Config: 0, RST: 0, MST: 18
 BPDU Received        : 0
          TCN: 0, Config: 0, RST: 0, MST: 0
```

  HUAWEI

In network design, the CIST root bridge and the secondary root bridge are located in a core region with high bandwidth.

However, due to incorrect configuration by the maintenance personnel or malicious attack to the network, the root bridge may receive the configuration message with higher priority. In this case, the root bridge becomes the non-root switch and the network topology also changes. This illegal change leads the traffic that should be transmitted through a high-speed link to a low-speed link. Thus, network congestion occurs.

The root protection function can avoid this problem. When this function is enabled on a port, the role of this port is always the designated port in all instances. When the port receives the configuration message with higher priority, that is, the port will be selected as a non-designated port, the status of this port is set to Discarding. Then this port cannot forward packets (equivalent to the situation that the link connection to this port is disconnected). If the port does not receive the configuration message with higher priority with a certain period (max age, by default, 20 seconds), this port becomes the designated port again and restores to the Forwarding status.

Therefore, the root protection function can avoid the attack to the designated port of the root bridge.

Configure Loop Protection Function

```
[SWB]stp interface Ethernet 1/0/13 loop-protection

[SWB]display stp interface Ethernet 1/0/13

----[CIST][Port13(Ethernet1/0/13) ][FORWARDING]----
 Port Protocol        : enabled
 Port Role            : CIST Root Port
 Port Priority        : 128
 Port Cost(Dot1T)      : Config=auto / Active=199999
 Desg. Bridge/Port    : 0.000f-e212-f8e1 / 128.13
 Port Edged           : Config=disabled / Active=disabled
 Point-to-point       : Config=auto / Active=true
 Transit Limit        : 3 packets/hello-time
 Protection Type      : Loop
 Num of Vlans Mapped : 1
 PortTimes            : Hello 2s MaxAge 20s FwDly 15s RemHop 0
 BPDU Sent            : 6
          TCN: 0, Config: 0, RST: 0, MST: 6
 BPDU Received        : 705
          TCN: 0, Config: 0, RST: 0, MST: 705
```

  HUAWEI

By receiving the BPDUs continuously from the upstream switch, a switch can maintain the status of the root port and other blocked ports.

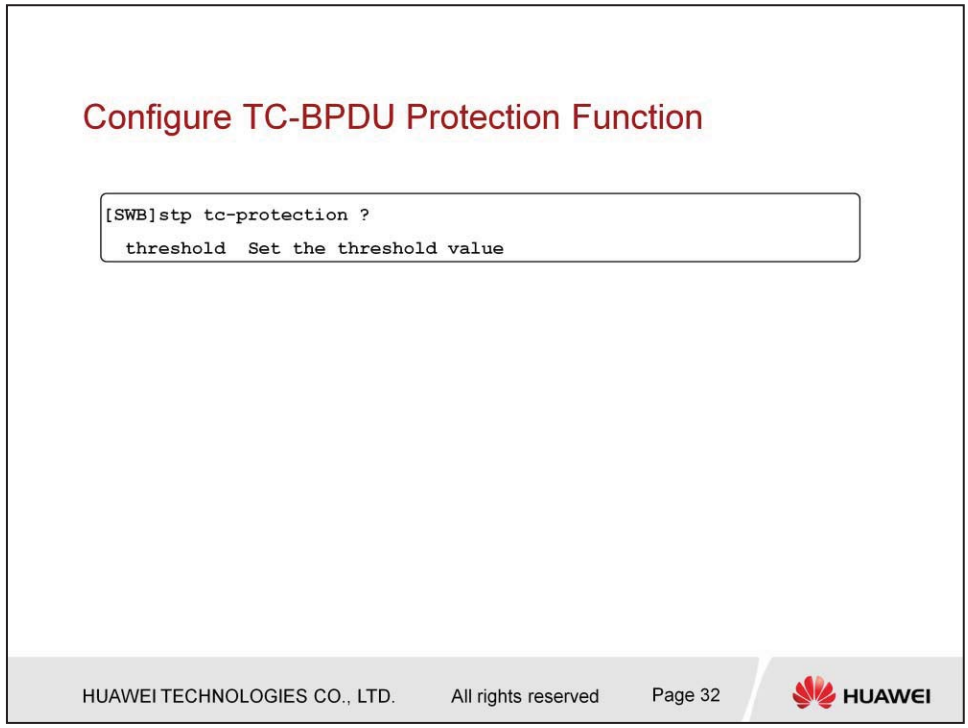
However, upon link congestion or unidirectional link fault, some ports may fail to receive the BPDUs from the designated port of the upstream switch. In this case, the switch selects the new root port. The previous root port becomes the designated port ,and the blocked port switches to the Forwarding status. Thus, loop

is generated in the switching network.

The protection function against the loop can restrict the generation of loop. When this function is enable, the role of the root port does not change, and the blocked port remains in the Discarding status and does not forward packets.

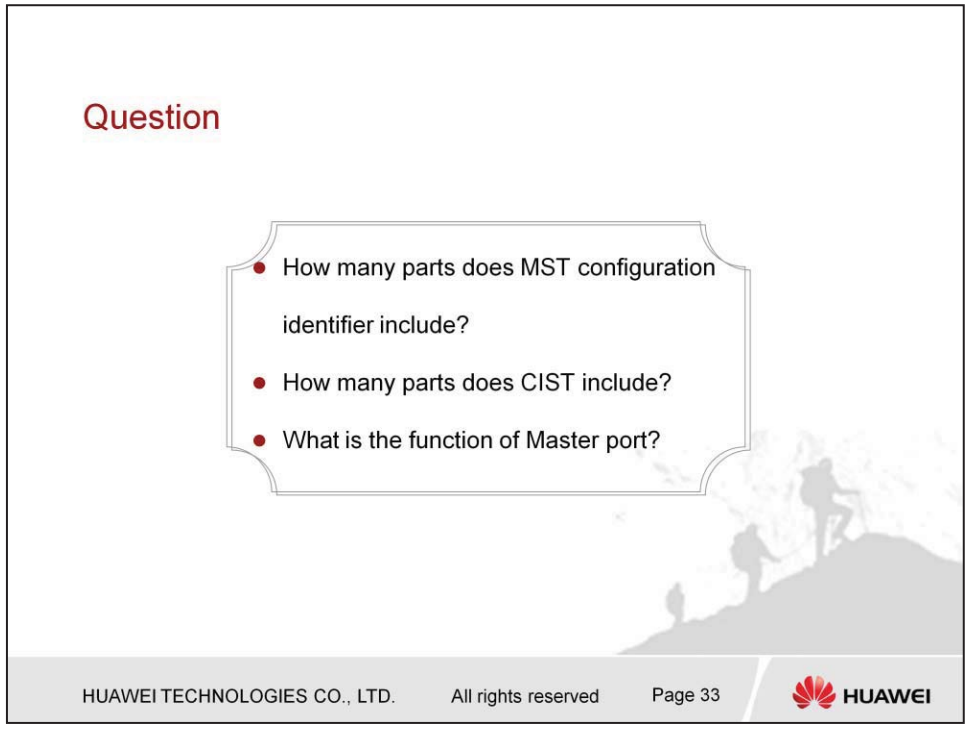
Thus, the loop cannot be generated in the network.

Therefore, this protection function is used to protect the non-designated ports.



When a switch receives the TC-BPDU, the switch deletes the MAC address entries and ARP entries. When an attacker forges the TC-BPDUs to attack the switch, the switch will receive many TC-BPDUs in a short time. The switch has to frequently delete the MAC address entries and ARP entries, which increases the burden of the network and threatens network stability.

When the protection function against TC-BPDU attack is enabled, the switch deletes the MAC address entries and ARP entries within 10 seconds only once after it receives the TC-BPDU. This function prevents the switch from frequently deleting the MAC address entries and ARP entries.



How many parts does the MST configuration identifier include?

The MST configuration identifier consists of the configuration identifier format selector, configuration name, configuration digest, and revision level. Only the switches with identical values in the four parts are regarded as in the same region

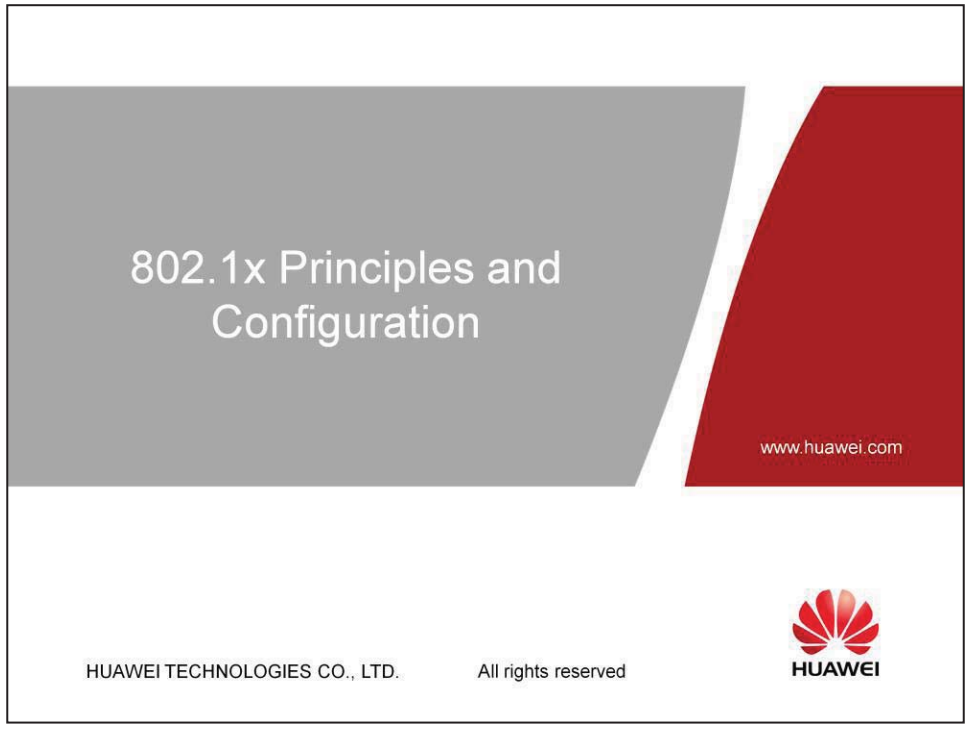
How many parts does the CIST include?

The CIST consists of the CST and the IST in the MST region. The CIST connects all switches and network segments in a network.

What is the function of the master port?

The master port is used to connect the MST instance to the CIST root.

Module 3
Access Layer Protocols

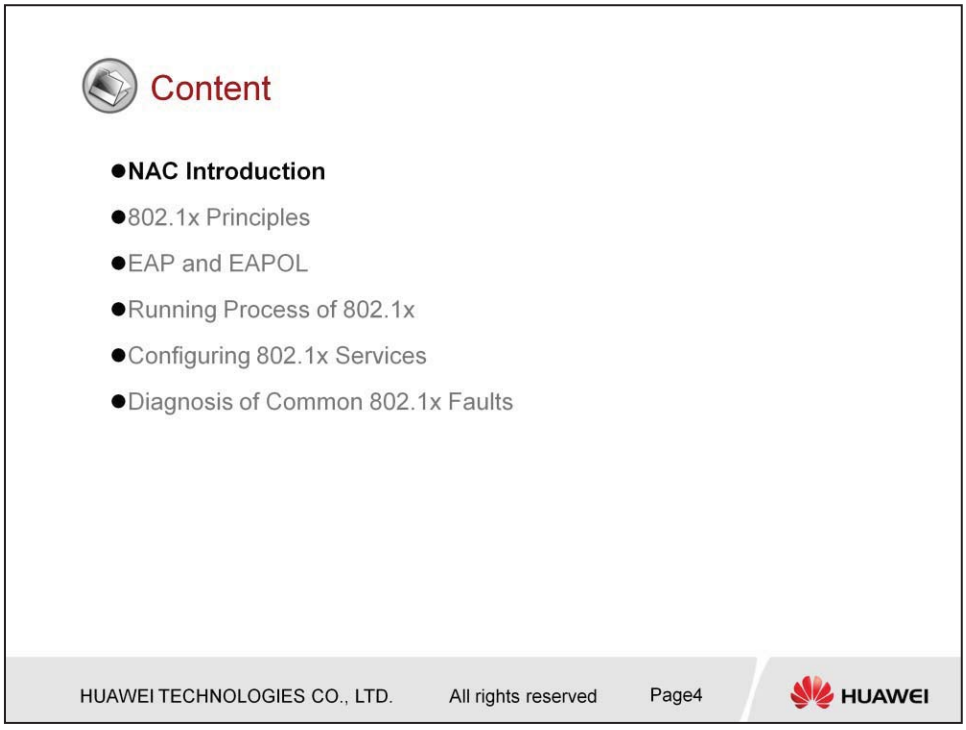

802.1x Principles and Configuration
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

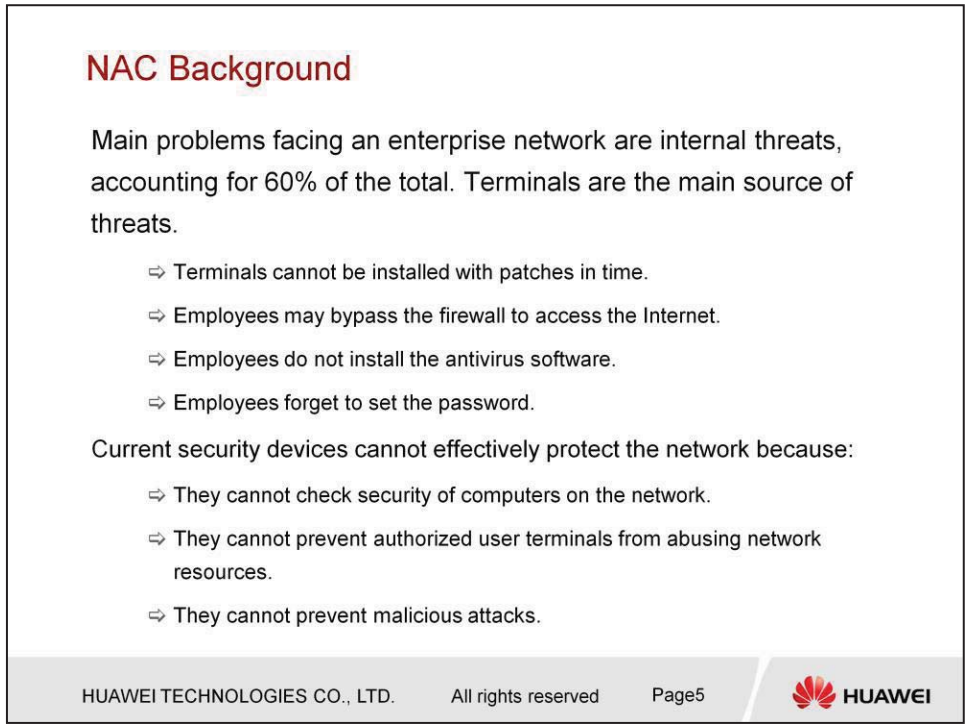
Foreword

- At present, network security has become one of the greatest concerns of corporate users.
  In a corporate network, the security status of a terminal will directly affect the security of the entire network. While the traditional system of defense against viruses is based on isolated single point of defense-based, such a decentralized management can not avoid many security threats.

Objective

- Upon completion of this section, you will be able to :
- Explain the concept of 802.1x
- Describe the 802.1x system components and operation
- Master the ability to diagnose basic 802.1x failures


Content
NAC Introduction
802.1x Principles
EAP and EAPOL
Running Process of 802.1x
Configuring 802.1x Services
Diagnosis of Common 802.1x Faults
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page4
HUAWEI



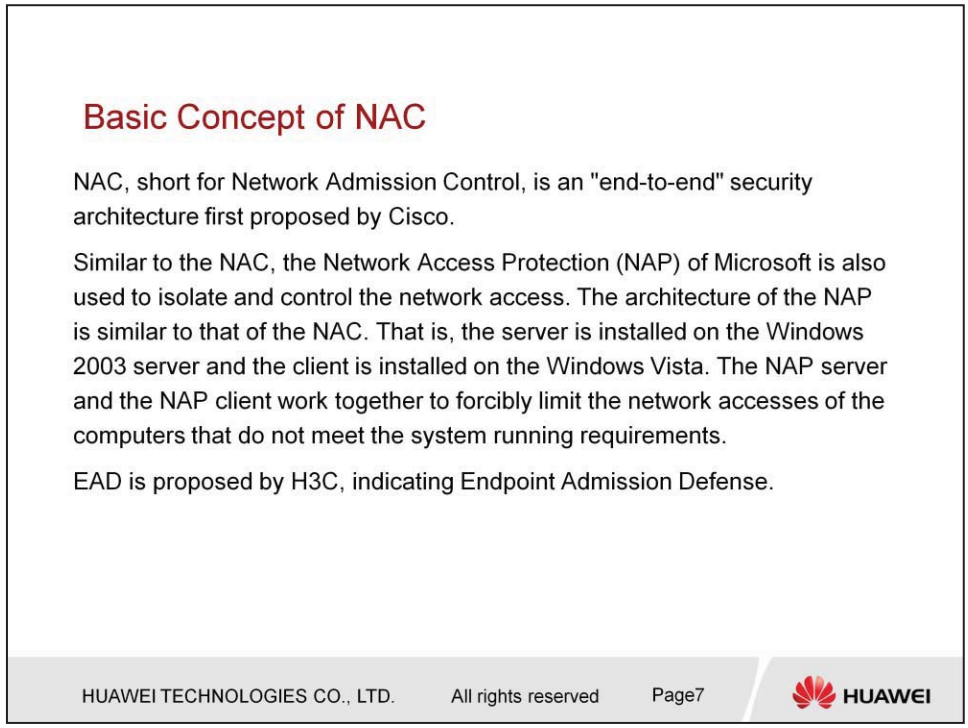
Most Enterprise network use the identity management and authentication, authorization, and accounting (AAA) mechanisms to authenticate users and assign network access rights to them. However, these make hardly any effect on the authentication of security of terminal devices. If no accurate method is used to evaluate the “security status“ of the devices or terminals , Whole enterprise network may be exposed to dangerous threats through being access by authenticated user with his/her unsecure device that has been infected with viruses or is not properly protected.

NAC based on the infrastructure of network to implement security policy-based check on all devices attempting to access network resources. In this way, new security threats such as viruses, worms, and spy-software are prevented from threaten of network security. Customers that implement the NAC mechanism would permit only trusted terminals (PC, server, and PDA) that observe security policies to access the network and restrict devices that do not conform to security policies or cannot be managed to access of the network.

NAC Background (Con.)

Preventing and controlling internal threats must be strengthened and security management of terminals must be strengthened.

- ⇨Terminal access control: Accesses of illegal terminals must be prevented, reducing the threats of insecure terminals.
- ⇨Terminal access authorization: Unauthorized accesses of legal terminals are prevented, protecting core resources of the enterprise.
- ⇨Robustness check and policy management of the terminal security: The security management rules are implemented for the enterprise.
- ⇨Staff behavior management and violation audit: The behavior audit is strengthened, preventing malicious terminal damages.



NAP also supports Windows 2008 and Windows XP SP3.

The Network Access Protection (NAP) technology is a new set of operating system components designed for the next-generation Windows Vista and Windows Server Longhorn. It implements health check on the system platform for access of a private network. The NAP platform provides a set of integrity check methods to determine the health status of clients connecting to the network. The clients that do no conform to health policies are restricted for network access.

To check the health of hosts connecting to a network, the network architecture needs to provide the following functions:

Health policy authentication: determines whether computers comply with the health policies.

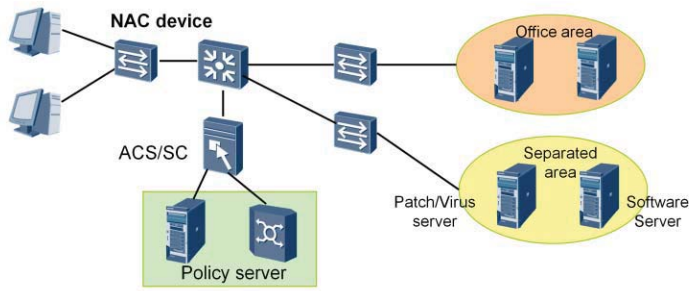
Network access restriction: restricts the access of computers that do not comply with the policies.

Automatic remedy: provides necessary upgrade of non-compliant computers, provide a way to them to comply with health policies.

Dynamic adaptation: automatically upgrades health policy for compliant computers to keep them up with the latest health policies.

Key Components of NAC

The NAC technology involves three core components, that is, communication agent, NAC device, and policy server,



The S series switches of Huawei can be used for network access control.

  HUAWEI

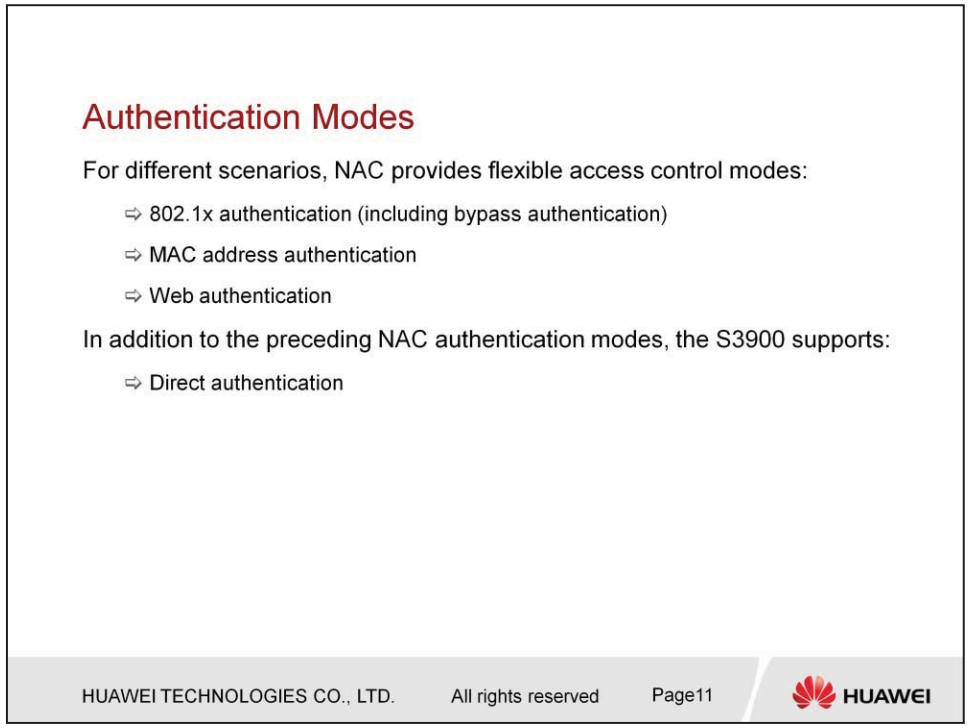
The communication agent (the security client), which is the software installed on the user terminal system. It authenticates the user's terminal, evaluates security status, and performs the security policy. The main functions are as follows:

The NAC device is the point where security policies are performed on the enterprise campus network. It constrainedly authenticates the access users, isolates invalid terminals, and provides network services for authorized users. The NAC device may implement endpoint access control through different authentication means such as 802.1x, MAC, and Portal authentication.

The security policy server is also called management server. The core of the NAC scheme is combination and association .The security policy server is the management and control center of the NAC scheme, providing the following functions:

- Managing users
- Managing security policies
- Evaluating the security status
- Controlling security association
- Auditing security events

The S series switches of Huawei support 802.1x, Portal, and MAC authentication technologies. They can serve as NAC devices to implement NAC control in cooperation with mainstream communication agents and security policy servers, thus providing reliable access control for enterprise networks and campus networks.



802.1x authentication

The Institute of Electrical and Electronics Engineers (IEEE) 802.1X standard, 802.1X in brief, is an interface-based network access control protocol. Interface-based network access control indicates that the authentication and control implemented for access devices is based on an interface of a LAN access control device. User devices connected to the interface can access the resources on the LAN only after passing the authentication.

The 802.1X protocol is concerned about only the status of an access interface. When a legal user accesses an interface by using the correct account and password, the interface is enabled; when an illegal user (such as who cannot provide correct username and password) accesses an interface or no user accesses an interface, the interface is disabled. The authentication result is only about the change of the interface status but is not involved with the IP address negotiation and assignment that need to be considered in common authentication technologies. 802.1X authentication is the most simplified implementation solution among various authentication

technologies.

Supports MAC address authentication:

MAC address authentication is used to control the network access permission of a user based on the access interface and MAC address of the user. The user does not need to install any client software. Both user name and password for authentication are set to the MAC address of the user device. After detecting the MAC address of the user for the first time, a network access device starts authenticating the user.

Supports Web authentication.

Web authentication is also called Portal authentication. The basic principle is that when a user opens a browser for the first time and enters any Website address, the user is forcibly redirected to the authentication page on the Web authentication server and the user can access network resources only after passing the authentication. Unauthenticated users can access only some specified site servers. Web authentication uses the Portal protocol to finish the authentication process after a user name and password are entered on a Web page.

The Portal protocol is mainly used to exchange the message between Web servers and other devices. The Portal protocol is based on the client/server model and uses the User Datagram Protocol (UDP) as the transmission protocol. During Web authentication, the Web authentication server communicates with the S9300 acting as a client through the Portal protocol. When obtaining the user name and password submit by the user on the authentication page, the Web authentication server transfers them to the S9300.

Direct authentication exclusively provided by S9300:

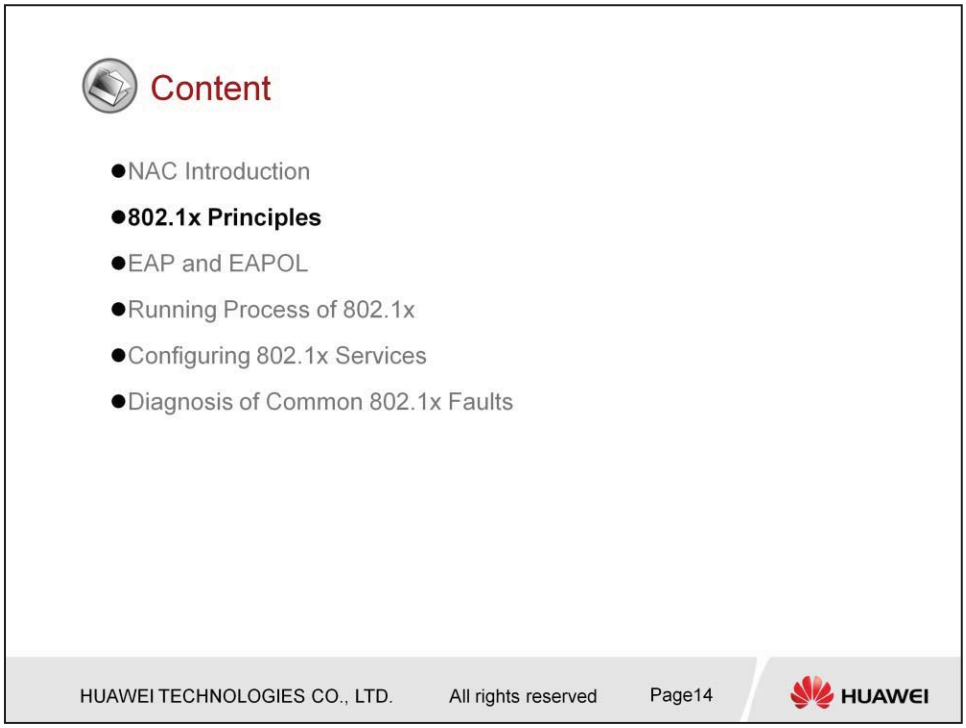
After direct authentication feature is enabled, users who connect to the network through this interface pass the authentication directly.

Direct authentication cannot be enabled globally. It is only enabled on the interface. After direct authentication is enabled, the forwarding status of the interface cannot be changed.

After receiving ARP or DHCP packets, the authentication module sends authentication requests to the server. The user is authorized without the user name and password.

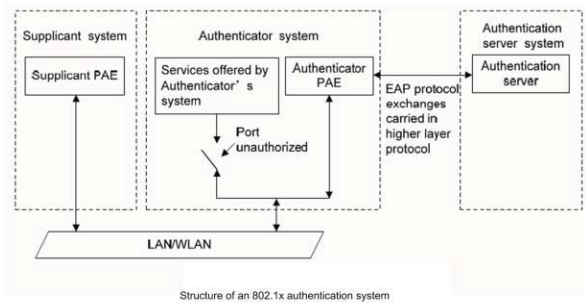
In direct authentication, the VLAN ID and ACL can be sent dynamically.

This course dwells upon the 802.1x authentication mode. Other authentication modes are comparatively easy and corresponding manuals are available for your reference.


Content
NAC Introduction
802.1x Principles
EAP and EAPOL
Running Process of 802.1x
Configuring 802.1x Services
Diagnosis of Common 802.1x Faults
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page14
HUAWEI

Structure of an 802.1x System

The system where 802.1x authentication is enabled adopts the typical client/server model and involves three entities, that is, the supplicant, authenticator, and authentication server.



Structure of an 802.1x authentication system

The supplicant is an entity on one end of point-to-point LAN network segment. It is authenticated by the device on the other end of the link. The supplicant is usually a user terminal. The user initiates 802.1X authentication by starting the supplicant software. The client software must support the EAP over LANs (EAPOL) protocol.

The authenticator is an entity on the other end of point-to-point LAN network segment. It is used to authenticate supplicant. The authenticator is usually a network device supporting the 802.1X protocol. The authenticator provides the interface, either physical interface or logical interface, for LAN access of the supplicant.

The authentication server is an entity that provides the authentication service for the authenticator. The authentication server implements user authentication, authorization, and accounting, and is usually called RADIUS server. This server can store user information, for example, user account and password, VLAN where users belong, priority, and user access control list.

Port Access Entity (PAE):

PAE is an entity that implements algorithms and protocol operations in the authentication mechanism. The authenticator PAE depend on the authentication server

to authenticate the supplicant that accesses the LAN and according to the authentication result to control the controlled interface on authorized or unauthorized status .The supplicant PAE is responsible for responding to the authentication requests from the authenticator and submitting the user's authentication information to the authenticator. The supplicant PAE can also send authentication requests and offline requests to the authenticator.

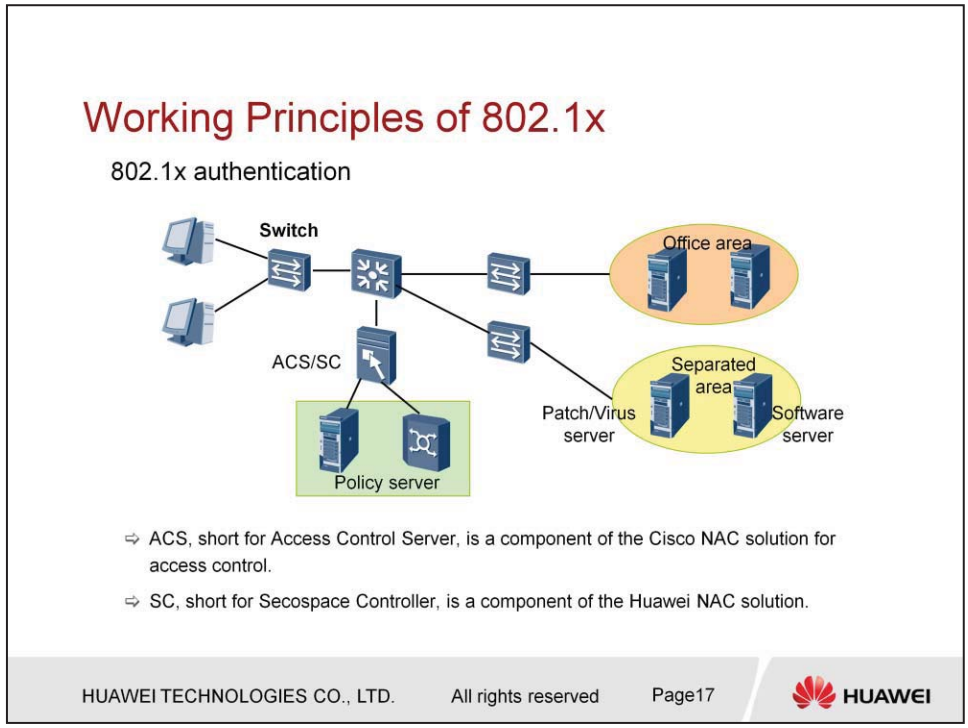
Controlled port:

The authenticator provides an interface for accessing the LAN for the supplicant. The interface is classified into two logical interfaces, that is, the controlled interface and the uncontrolled interface.

The uncontrolled interface is mainly used to transmit EAPOL frames in both directions to ensure that the supplicant can sends and receives authentication packets at any time.

In authorized mode, the controlled interface transmits service packets in both directions; in unauthorized mode, the controlled interface refuse to receive packets from the supplicant.

The controlled interface and the uncontrolled interface are two parts of the same physical interface. All frames entering the interface can be seen on the controlled interface and the uncontrolled interface.



The switch initiates an EAP authentication request after detecting a new MAC address.

If the client does not respond to the request:

The switch thinks that the client software is not installed after a certain number of requests are sent. In this case, user rights are restricted to permit the PC to access the specific separated-area.

Detection would be initiated again after a period of time (the period is configurable).

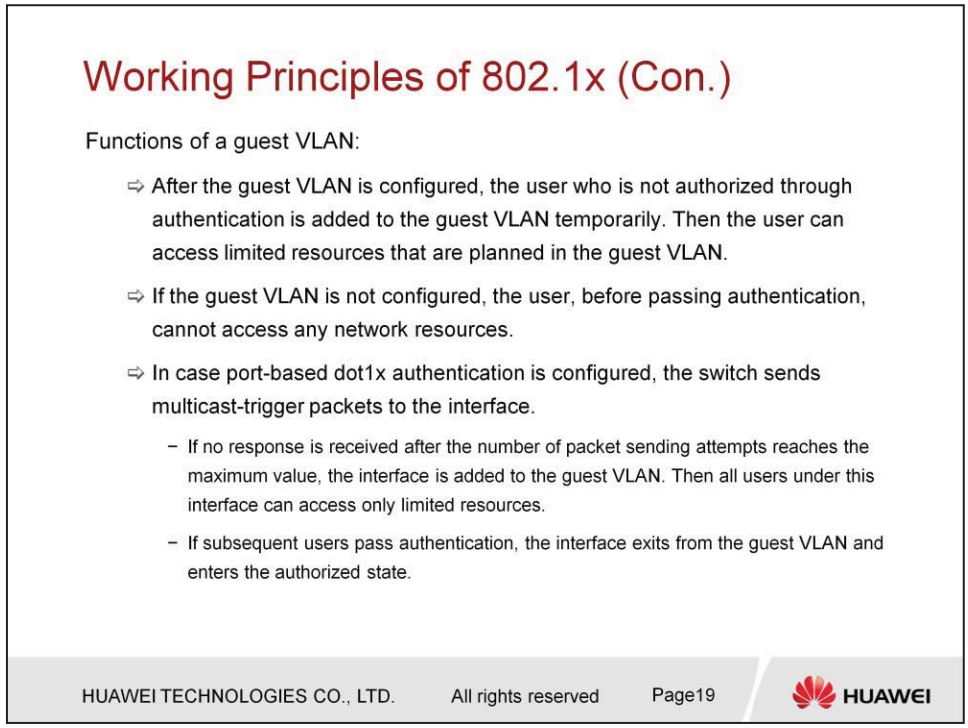
If the client software is installed, the switch implements 802.1x authentication. After the PC passes authentication, an HTTP link is established between the PC and the ACS. After the PC passes the security check implemented by the ACS, the ACL is updated and then the PC can access the office area.

If the client software is not installed, but MAC bypass authentication is configured, the switch takes the MAC address of the PC as the user name and password for authentication. If the PC fails the authentication, the switch forces the PC offline and does not initiate authentication and detection requests within a given period. After the timer expires, the detection process is re-initiated. At this time, if the

user has downloaded the client software, the switch re-initiates EAP authentication. Upon receipt of the request, the ACS forces the PC offline and then responds to the EAP authentication request.

After the ACS detects that the PC is infected with virus, the ACS delivers an ACL to allow the user to access the isolation area only, and redirects the user to the URL to update its virus library or install a corresponding patch.

After the user updates its virus library, the ACS detects that the user is secured, and the switch updates the ACL through the COA interface on the RADIUS server. In this case, the user is allowed to access the working area.



The guest VLAN must be planned, and does not conflict with other special VLANs. Ensure that the user can access only limited resources such as the virus library, client downloading, promotion page, and DHCP server if the user has no access authority.

Note the following issues when configuring a guest VLAN:

The VLAN set as the guest VLAN must already exist in the system and cannot be the default VLAN on interface.

After the guest VLAN is configured on an interface, the interface cannot be added to the guest VLAN, and the guest VLAN cannot be deleted.

Different interfaces can be configured with different guest VLANs.

If you run the dot1x port-method command repeatedly in the same view, the latest configuration takes effect.

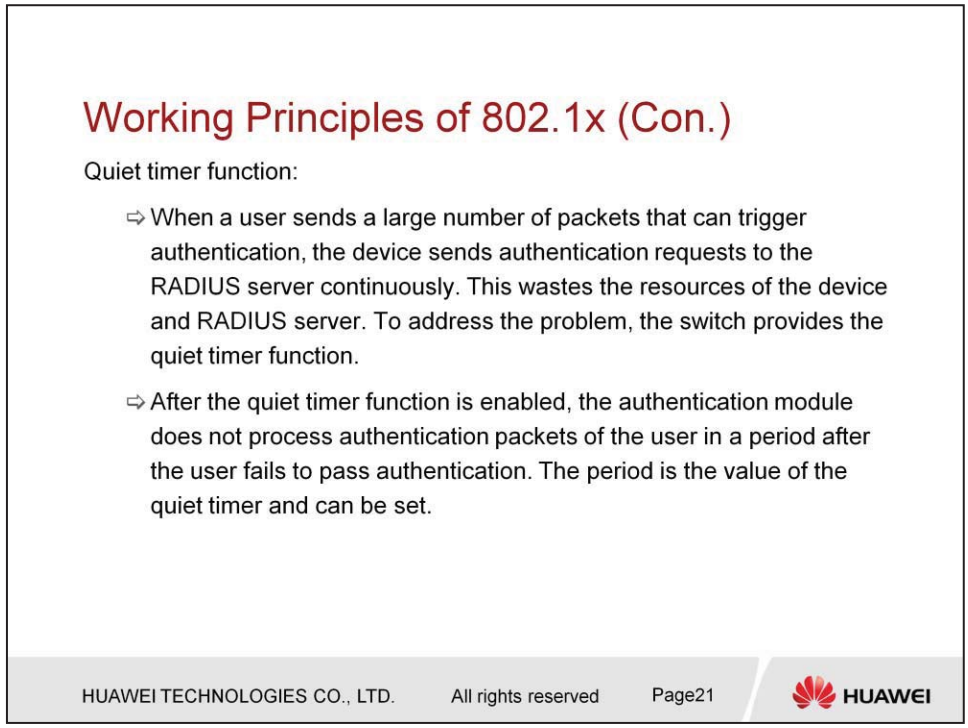
After the guest VLAN function is enabled:

The switch sends multicast-trigger packets to all ports enabled with the 802.1x function.

If the number of packet sending attempts reaches the maximum value and there are still a port that fails to

respond, the switch adds this port to the guest VLAN.

Then users in the guest VLAN can access resources in the guest VLAN without 802.1x authentication. Authentication, however, is required when such users access other special VLANs.

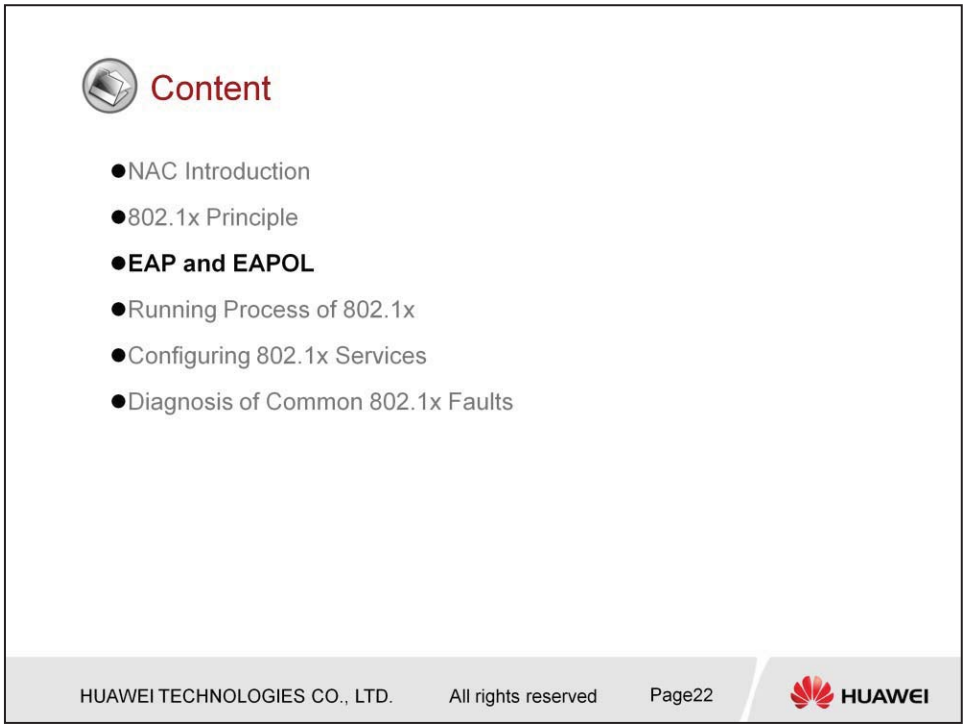


The 802.1x quiet timer function is enabled on the S9300 by default and can be disabled by using a command.

In the case that the quiet timer function is enabled, to prevent the 802.1x user from entering the silent state after the first authentication failure, you can set the number of authentication failures before the 802.1x user enters the silent state to be greater than 1 on the S9300.

By default, an 802.1x user enters the quiet state after three authentication failures within 60 seconds.

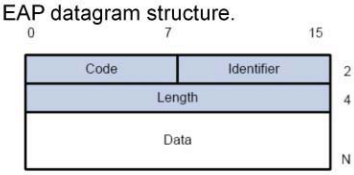
During the quite period, the S9300 discards 802.1x authentication request packets from the user. The value of the quiet timer is set through the command.


Content
NAC Introduction
802.1x Principle
EAP and EAPOL
Running Process of 802.1x
Configuring 802.1x Services
Diagnosis of Common 802.1x Faults
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page22
HUAWEI

EAP Datagram

Format of an EAP datagram

⇨ When the Type field of an EAPOL datagram is **EAP-Packet**, the packet body is in the EAP datagram structure.



- **Code**: indicates the type of EAP data packets, which can be **Request**, **Response**, **Success**, **and Failure.**
- **Identifier**: matches Request and Response messages.
- **Length**: indicates the length of EAP data packets, including the fields of Code, Identifier, Length, and Data. The value is expressed in bytes.
- **Data**: indicates the contents of EAP data packets, which is determined by the Code field.
  - Success and Failure data packets do not have the **Data** field and their length is 4 bytes.

  HUAWEI

Code: This field contains one byte, used to identify the type of an EAP packet.Identifier: This field contains one byte, used to match Request and Response packets.

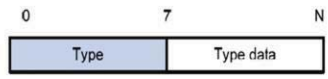
Length: This field contains two bytes, used to indicate the total length of an EAP packet, including the fields of Code, Identifier, Length, and Data. The maximum total length is 65536 bits.

Data: This is a field with variable length, whose content is determined by the Code field.

EAP Datagram (Con.)

EAP Request and Response packets:

⇨ When the code of an EAP packet is **Request** or **Response**, the format of the **Data** field is as follows:



- When the **Type** field is **EAP**, the authentication mode may be:
  - 1, indicating **Identity**, used for querying the identity of the other party.
  - 4, indicating **MD5-Challenge**. This is similar to the PPP CHAP protocol.
- The **Type-Data** field varies according to the type.

  HUAWEI

The authentication type and authentication information are included in the **Type** field and **Type-Data** field.

Request packets are sent from the authenticator to the authenticated.

The request packets to be retransmitted must be marked with the same Identifier values to distinguish with other request packets.

The authenticated must respond to a Request packet with a Response packet.

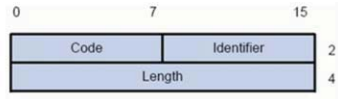
The Response packet is used only for replying to a received Request packet and cannot be retransmitted.

The Identifier values of response packets must be the same as those of corresponding request packets.

EAP Datagram (Con.)

EAP Success and Failure packets:

⇨ When the code of an EAP packet is **Success** or **Failure**, the packet does not have the **Data** field. The value of the Length field is **4**.



- A Success packet is sent by the authenticator to the authenticated, indicating a successful authentication result.
- If the authentication process fails, the authenticator sends a Failure packet to the authenticated, to notify an unsuccessful authentication result. The Identifier values of the Success packet and the Failure packet must be the same as those of the response packets.

EAPOL Datagram

Layer 2 header of an EAPOL datagram:

DMAC	SMAC	TYPE	EAPOL	FCS

Field	Description
DMAC	01-80-C2-00-00-03
SMAC	Physical MAC address of an interface
TYPE	**PAE Ethernet Type**, indicating the protocol type

When an EAPOL data frame is sent, the destination address is the group MAC address 01-80-C2-00-00-03.

This group address is one of the group addresses reserved in IEEE802.1D that cannot be forwarded by switches.

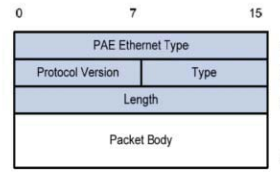
The source MAC address is the physical MAC address of the supplicant connecting the switch port.

The **Type** field is set to **88-8E**, it indicates that an EAPOL datagram is encapsulated.

EAPOL Datagram (Con.)

Format of an EAPOL datagram

- PAE Ethernet Type: indicates the protocol type; the value is **0x888E**.
- Protocol version: indicates the protocol version number supported by the sender of EAPOL frames.
- Type: indicates the type of an EAPOL datagram.
- Length: indicates the length of the data, that is, the length of the **Packet Body** field. The value is expressed in bytes. **0** indicates no **Data** field.
- Packet Body: indicates the contents of the data. The format varies according to the type.



Version: This field contains one byte, used to indicate the EAPOL version number used by the EAPOL datagram sender. The current version number is **1**.

Type: This field contains one byte, used to identify the type of a transmitted EAPOL datagram.

Length: indicates the length of an encapsulated EAP datagram. **0** indicates that no EAP datagram is encapsulated.

Packet Body: The EAP datagram is encapsulated in this field.

EAPOL Datagram (Con.)

Type of packets

Type Value	Packet Type
0x00	EAP packet (EAP-Packet), used for carrying authentication information
0x01	EAPOL starting packet (EAPOL-Start), used for initiating authentication
0x02	EAPOL logoff packet (EAPOL-Logoff), a request for logging off
0x03	EAPOL information packet (EAPOL-Key)
0x04	EAPOL alert packet (EAPOL-Encapsulated-ASF-Alert)

EAP-Packet: An EAP packet is contained in this EAPOL datagram.

EAPOL-Start: sent by the supplicant to the authenticator, which then starts the authentication protocol.

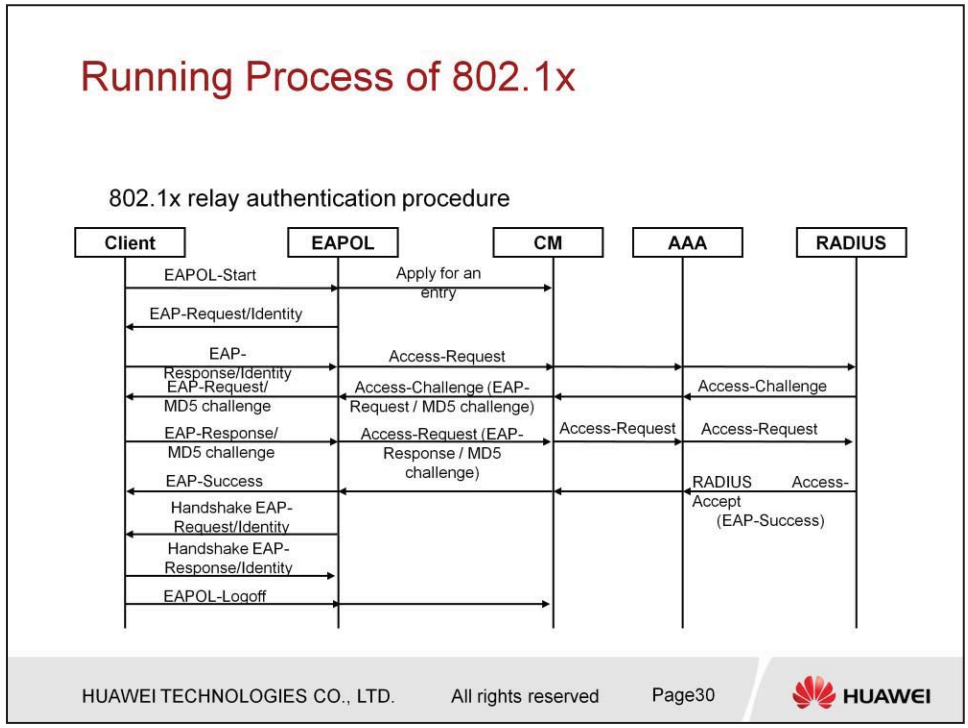
EAPOL-Logoff: sent by the supplicant to the authenticator, to identify an explicit logoff request.

EAPOL-Key: used by the authenticator and the supplicant to exchange information about optional 802.1x capabilities.

EAPOL-Encapsulated-ASF-Alert: used to transmit alert information such as SNMP Trap messages, defined by the Alert Standard Forum (ASF).

Content

- NAC Introduction
- 802.1x Principles
- EAP and EAPOL
- **Running Process of 802.1x**
- Configuring 802.1x Services
- Diagnosis of Common 802.1x Faults



The 802.1x authentication procedure consists of relay authentication and termination authentication.

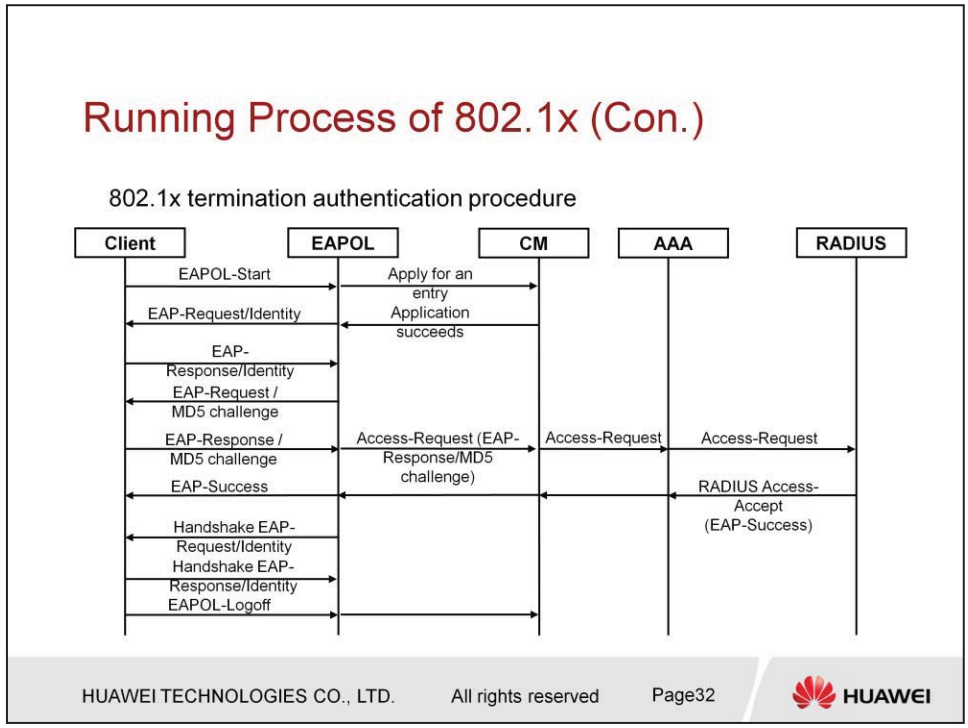
In relay authentication mode, the switch transparently transmits EAP responses and MD5 challenge requests. That is why we call it relay authentication and also transparent authentication.

The process of EAP termination authentication is as follows (A device indicates a network access device):

1. The EAP client sends an EAP-Start packet to a device.
2. The device sends an EAP-Request/Identity packet to the EAP client.
3. The EAP client responds with an EAP-Response/Identity packet and the device transparently transmits the packet to the RADIUS server.
4. After receiving an RADIUS challenge packet, the device sends an EAP-Request/MD5-Challenge packet to the EAP client.
5. The EAP client responds with an EAP-Response/MD5-Challenge packet and the device transparently transmits the packet to the RADIUS server.

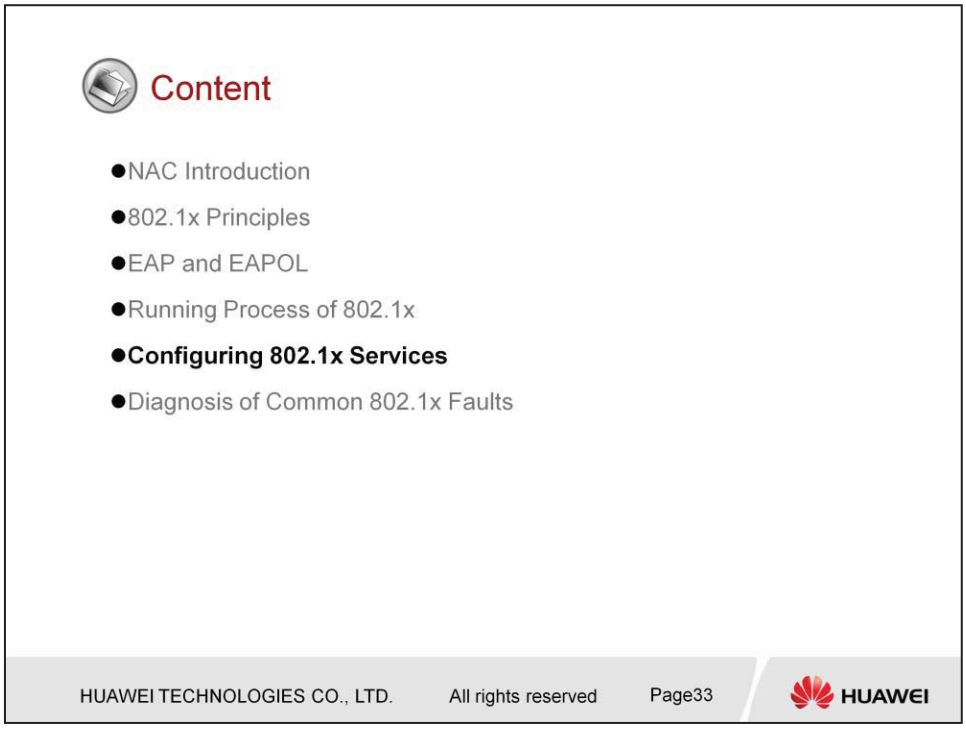
6. After passing the authentication, the device notifies the EAP client of the authentication success and the interface is enabled.

7. During the login of the EAP client, the device detects whether the EAP client remains online according to EAP handshake packets.

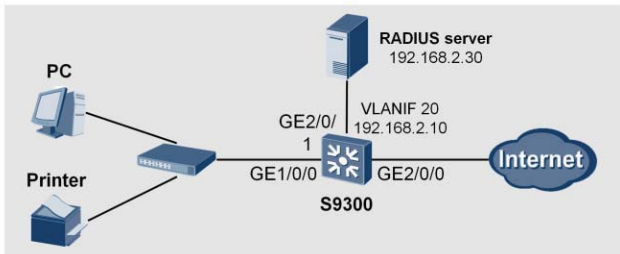


The process of EAP termination authentication is as follows (A device indicates a network access device):

1. The EAP client sends an EAP-Start packet to a device.
2. The device sends an EAP-Request/Identity packet to the EAP client.
3. The EAP client responds with an EAP-Response/Identity packet carrying user name information.
4. The device sends an EAP-Request/MD5-Challenge packet to the EAP client.
5. The client responds with an EAP-Response/MD5-Challenge packet and the device obtains password information about the client.
6. The device carries user account information for authentication in the AAA system.
7. After passing the authentication, the device notifies the client of the authentication success and the interface is enabled.
8. The device detects whether the EAP client remains online according to EAP detection.


Content
NAC Introduction
802.1x Principles
EAP and EAPOL
Running Process of 802.1x
Configuring 802.1x Services
Diagnosis of Common 802.1x Faults
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page33
HUAWEI

Basic Networking for 802.1x Services



The NAC feature needs to interact with multiple neighboring modules. Incorrect configuration of a module often results in authentication failure.

Service requirements:

802.1x authentication is performed for the user connected to GE 1/0/0 to control the user's access to the Internet. The default access control mode is adopted, that is, control of user access based on the MAC address of the user.

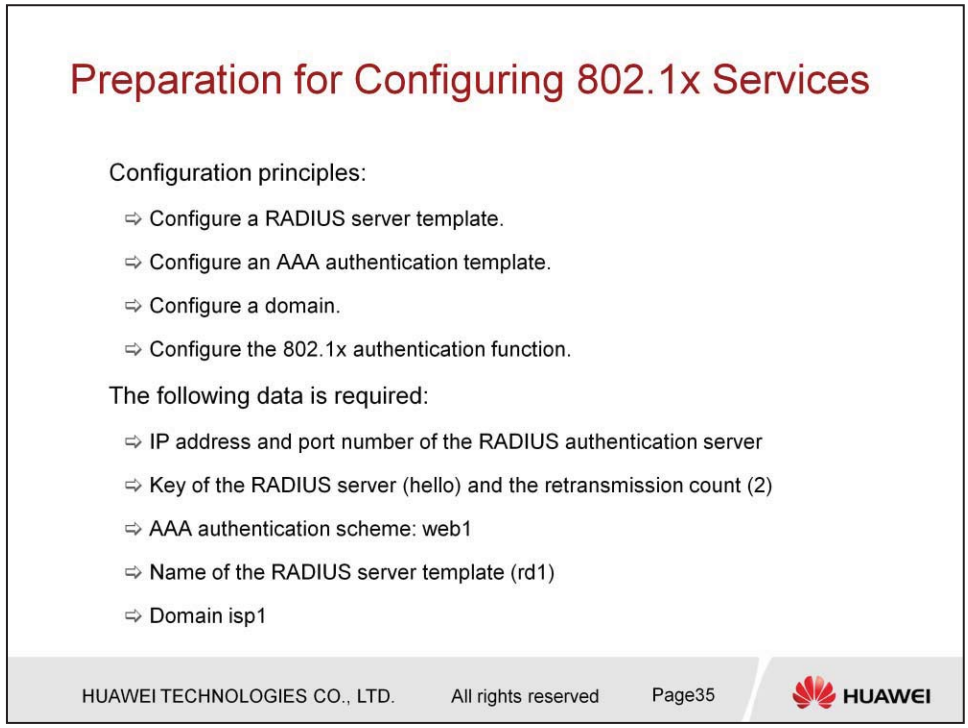
The authentication is performed by the RADIUS server.

The maximum number of users on GE 1/0/0 is 100.

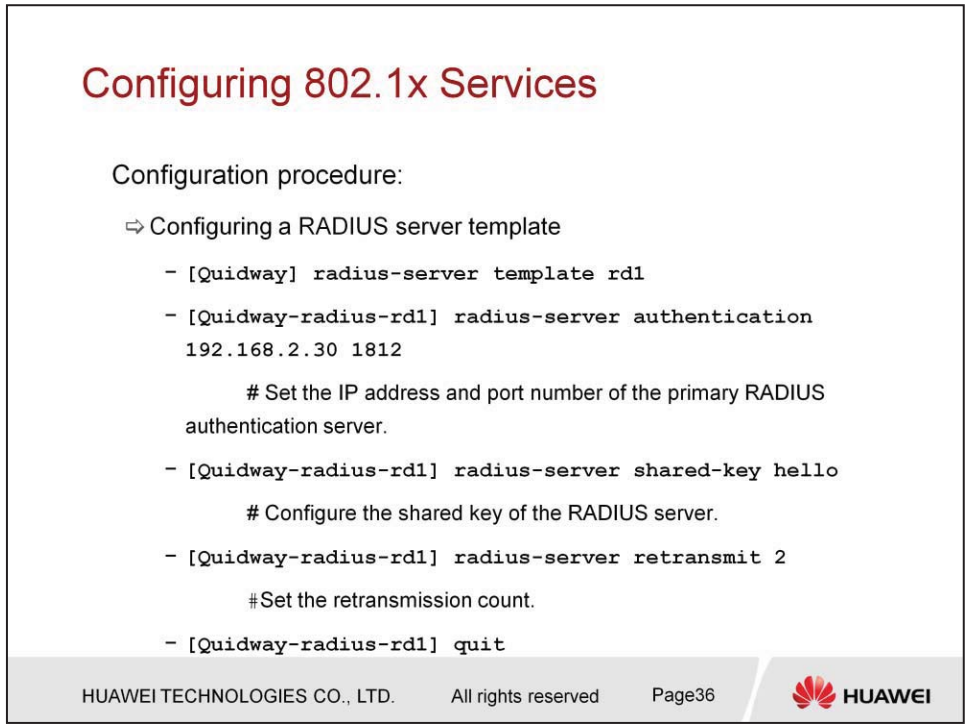
MAC address bypass authentication is performed for the printer connected to GE 1/0/0.

The 802.1x client software cannot be installed or used on some special terminals, such as printers. In this case, the MAC bypass authentication can be adopted.

MAC bypass function means that If the 802.1x authentication on the terminal fails, the access device sends the user name and password, namely, the MAC address of the terminal, to the RADIUS server for authentication.



Before configuring 802.1x services, you need to determine the configuration roadmap and prepare data.



Cautions:

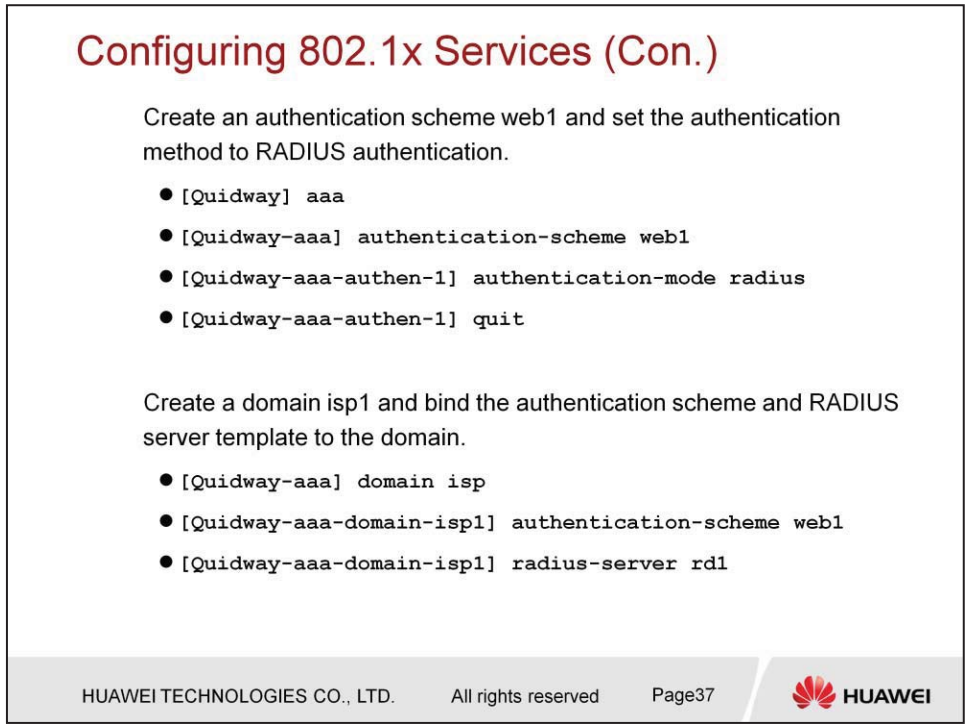
If only local authentication is used, the RADIUS configuration is of no concern.

The switch functions as the client of the RADIUS server; thus, the preceding configurations are performed on the client. To make the configurations effective, you need to configure the server.

The shared key needs to be negotiated with the RADIUS server; otherwise, the switch fails to communicate with the RADIUS server.

If the domain name is inconsistent with that recorded in the server, It can be removed from the username with following system command , and then be transmited to the RADIUS server.

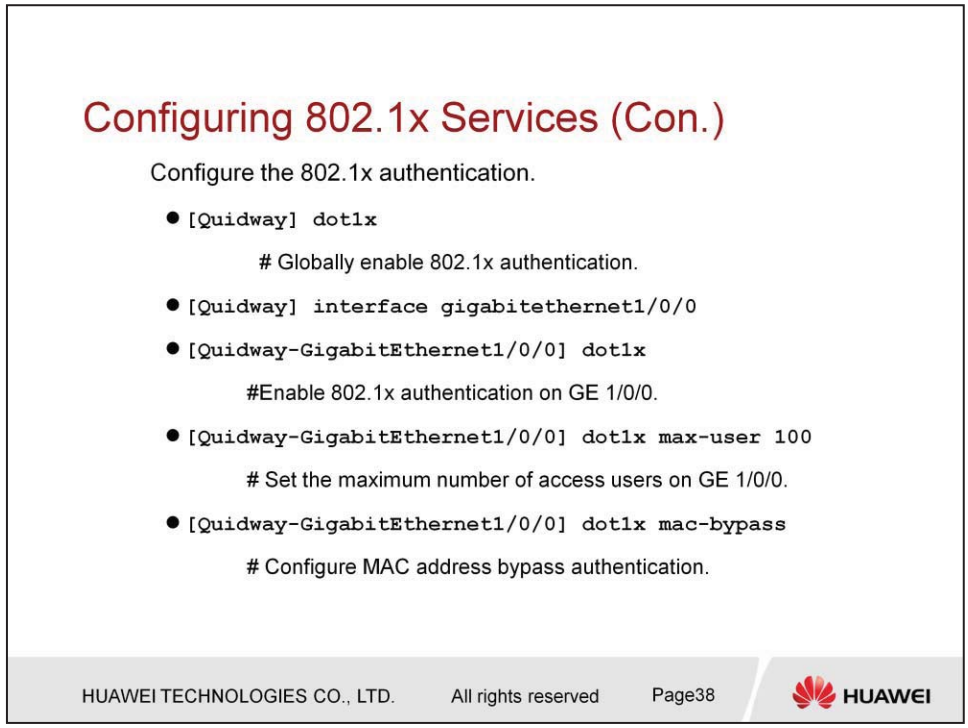
[Quidway-radius-rd1] user-name-format without-domain



Bind the authentication scheme and authentication template to the domain:

The authentication scheme specifies RADIUS as the authentication type.

Query the corresponding RADIUS through the authentication template.



By default, 802.1x authentication is disabled.

If there are online users who log in through 802.1x authentication, disabling 802.1x authentication is prohibited.

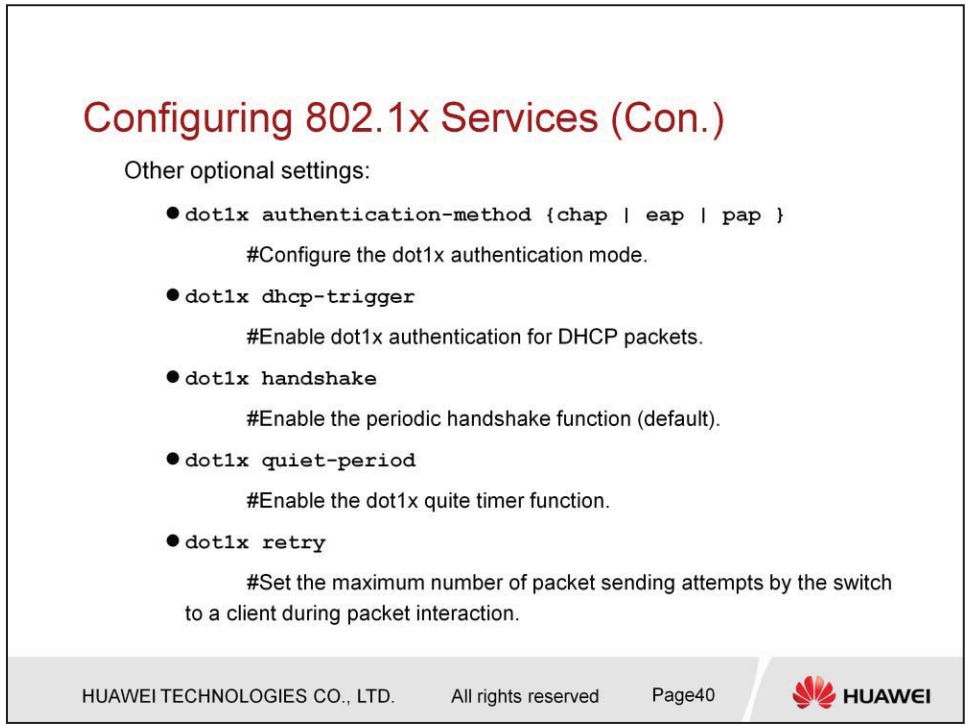
Set the maximum number of access users

The maximum number of online users depends on the model of the S9300. The value is (8192 x Number of LPU slots).

If you have run the dot1x port-method command with the port parameter specified (that is, users are authenticated based on interfaces), the maximum number of users on an interface changes to 1. In this case, to set the maximum number of users, run the undo dot1x port-method command first, and then run the dot1x max-user command.

If the number of users already existing on the interface is larger than the maximum number that you set, all the users are disconnected from the interface. The system displays a message for you: "Warning: The total number of online users is greater than the limit, so all the online users will go offline. are you sure to continue?[Y/N]: "

If you run the dot1x max-user command repeatedly in the same view, the latest configuration takes effect.



In the actual process of service configuration, you need to select from optional configurations as required by your customer. Therefore, you must know the default values of these options.

Configure the dot1x authentication method: dot1x authentication-method {chap | eap | pap }

PAP uses the two-way handshake mechanism and transmits the password in plain text.

CHAP uses the three-way handshake mechanism. It transmits only the user name but not the password on the network; Therefore, compared with PAP, CHAP is more secure and reliable and protects user privacy better.

In EAP authentication, the switch sends the authentication information of an 802.1x user to the RADIUS server through EAP packets without converting EAP packets into standard RADIUS packets.

dot1x dhcp-trigger enables the S9300 to trigger 802.1x authentication through DHCP messages.

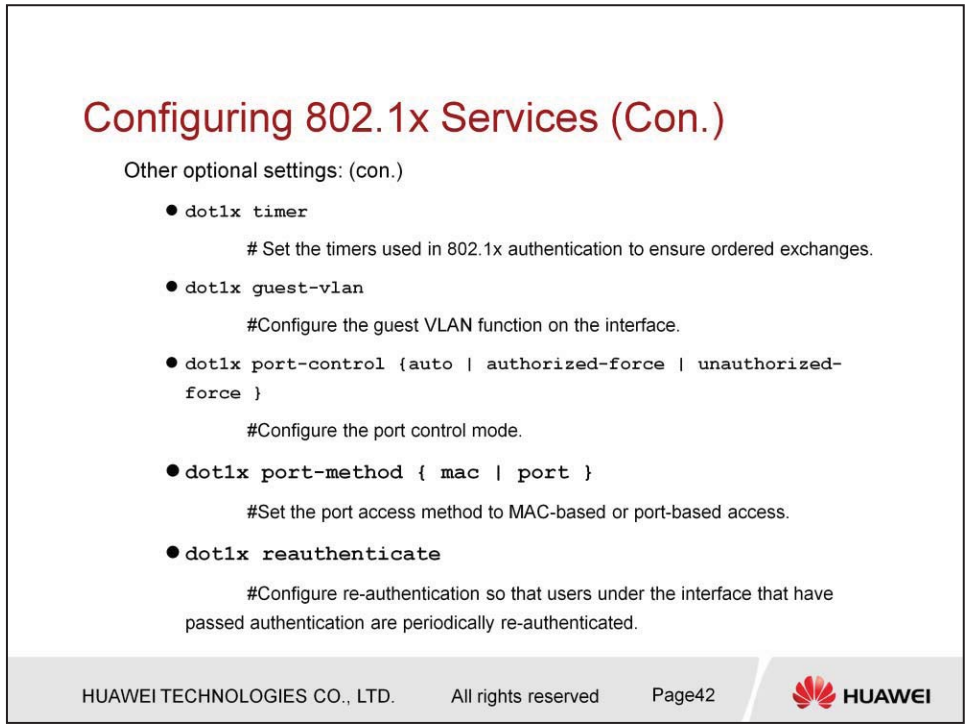
By default, 802.1 authentication cannot be triggered by DHCP.

After you run the dot1x dhcp-trigger enable command,

users cannot obtain IP addresses through DHCP server if they do not pass the authentication.

Enable the quiet timer function (dot1x quiet-period).

By default, an 802.1x user enters the quiet state after three authentication failures within 60 seconds.



dot1x port-control { auto | authorized-force | unauthorized-force }: used to configure the port control mode.

Default is auto. An interface is initially in unauthorized state and sends and receives only EAPOL packets. Therefore, users cannot access network resources. If a user passes the authentication, the interface is in authorized state and allows users to access network resources.

Authorized-force mode: An interface is always in authorized state and allows users to access network resources without authentication.

Unauthorized-force mode: An interface is always in unauthorized state and does not allow users to access network resources.

dot1x reauthenticate: allows periodic re-authentication of users, without the involvement of users.

dot1x guest-vlan:

When the guest VLAN is enabled on a switch, the switch sends authentication request packets to all the

802.1x-enabled interface. If an interface does not return a response after certain number of authentication packets are sent, the switch adds this interface to the guest VLAN. Then users in the guest VLAN can access resources in the guest VLAN without 802.1x authentication. Authentication, however, is required when such users access external resources. Thus certain resources are available for users without authentication.

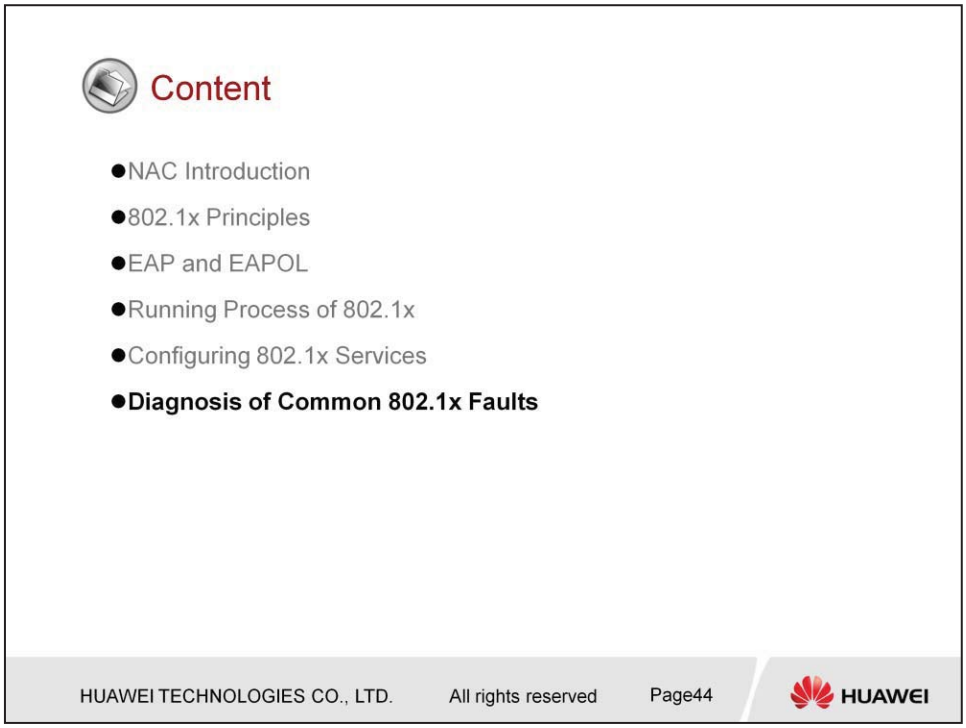
The configured guest VLAN cannot be the default VLAN of the interface.

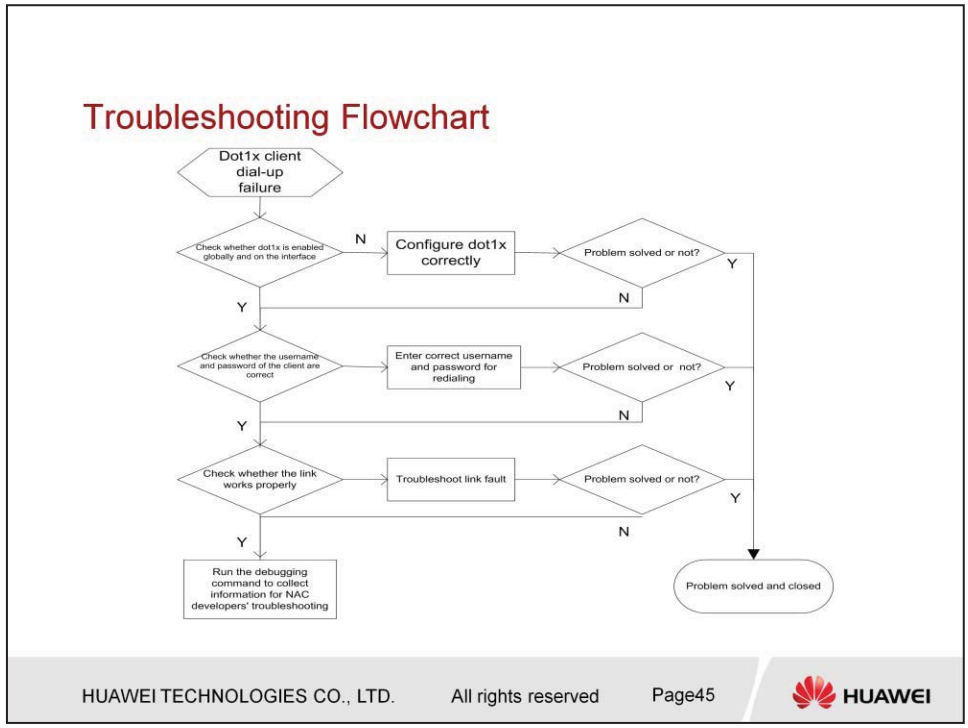
By default, no guest VLAN is configured on an interface.

dot1x port-method { mac | port }:

By default, the system adopts the MAC-based access control method.

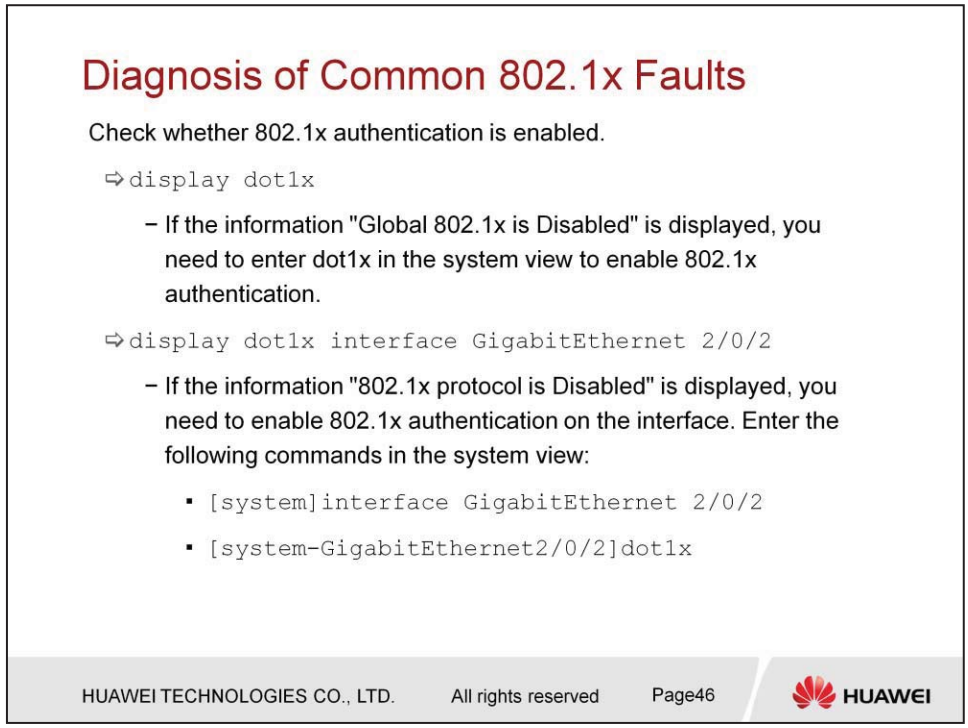
When 802.1x users are online, you cannot use this command to change the access mode of an interface.


Content
NAC Introduction
802.1x Principles
EAP and EAPOL
Running Process of 802.1x
Configuring 802.1x Services
Diagnosis of Common 802.1x Faults
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page44
HUAWEI



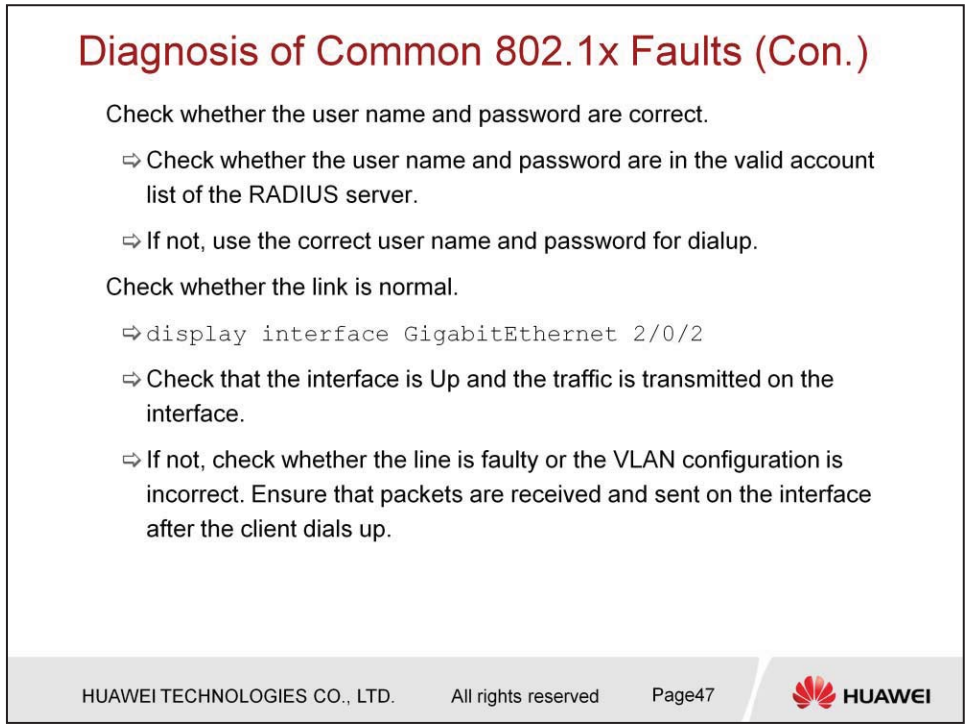
In this troubleshooting flow, RADIUS is deemed to work properly.

If RADIUS works improperly, you need to check whether the RADIUS server can be successfully pinged and whether RADIUS configuration is correct. If these operations cannot solve the problem, run debugging radius to collect information and then send such information to corresponding engineers of Huawei for troubleshooting.



802.1x faults may be diagnosed step by step according to this procedure.

You may directly view the configuration. If the current configuration does not contain dot1x, add it.



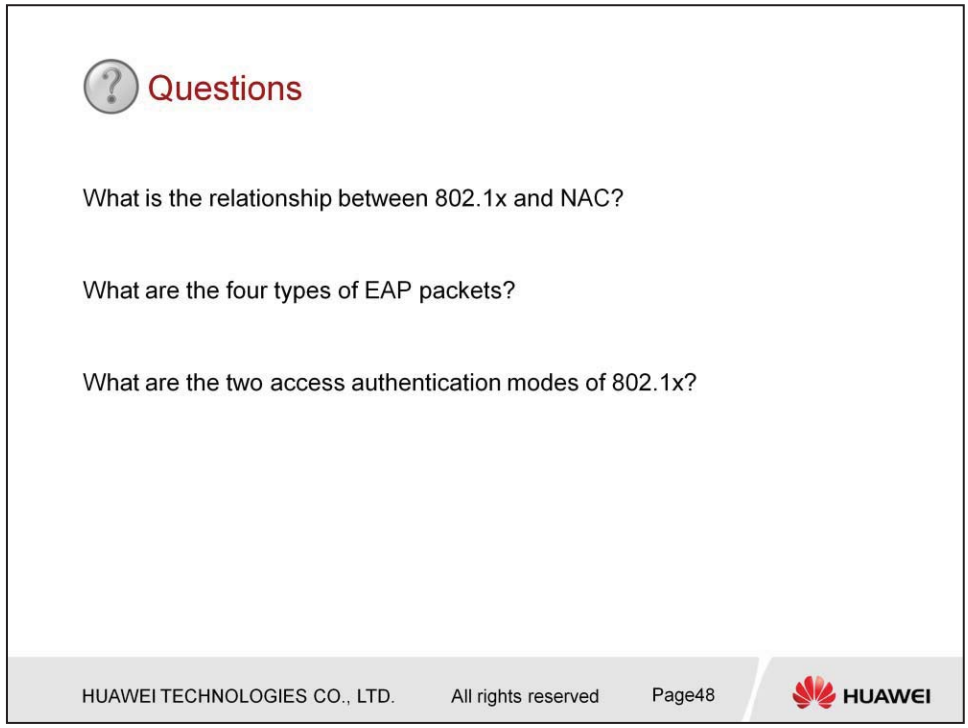
802.1x faults may be diagnosed step by step according to this procedure.

When the physical status of the port goes Down, the data link layer must be faulty.

For an optical interface, test the receiving and transmitting optical power.

For an RJ45 interface, use a cable tester to test the interface.

If the cable works properly, check whether the working modes of ports at both ends are consistent. If not, make the working modes consistent. It is recommended that you disable auto-negotiation.



What is the relationship between 802.1x and NAC?

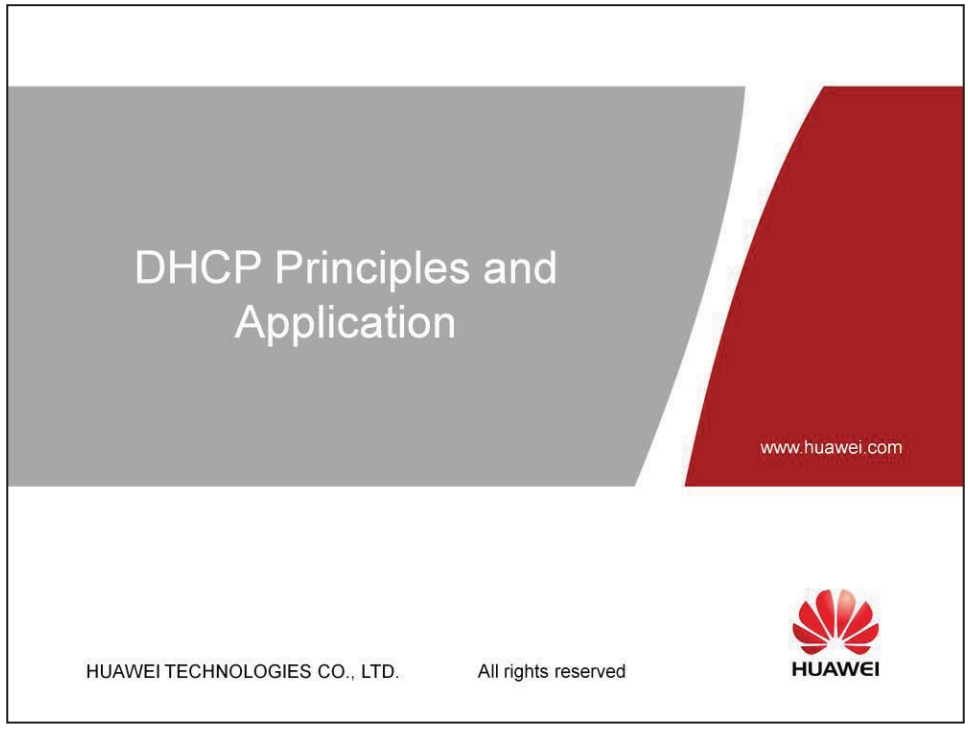
802.1x is one of the NAC authentication modes. NAC authentication modes also include MAC address authentication, Web authentication, and direct authentication.

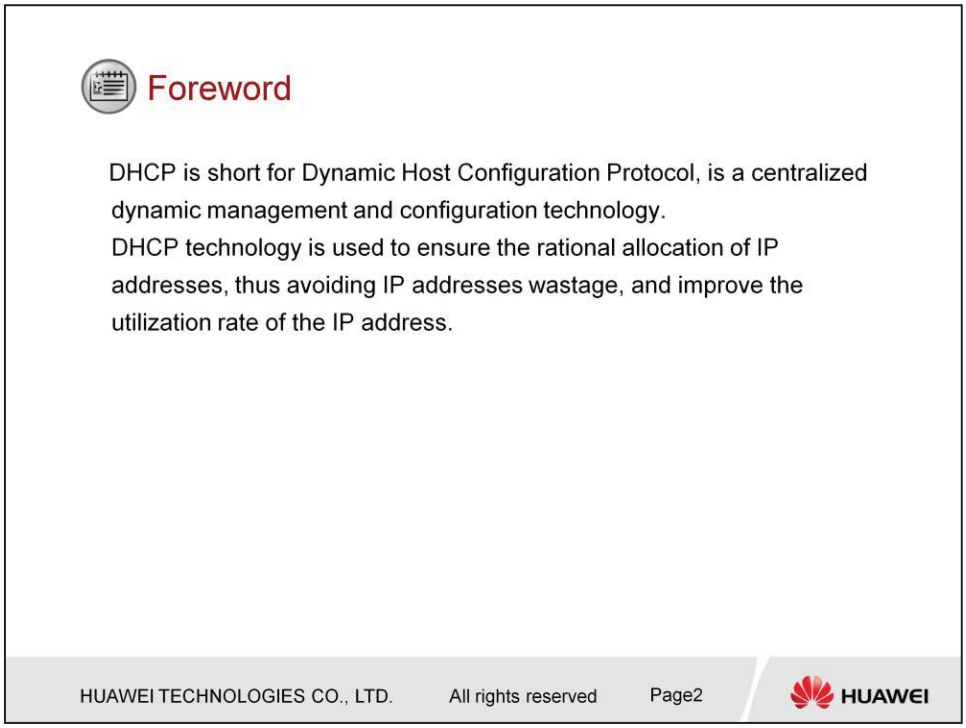
What are the four types of EAP packets?

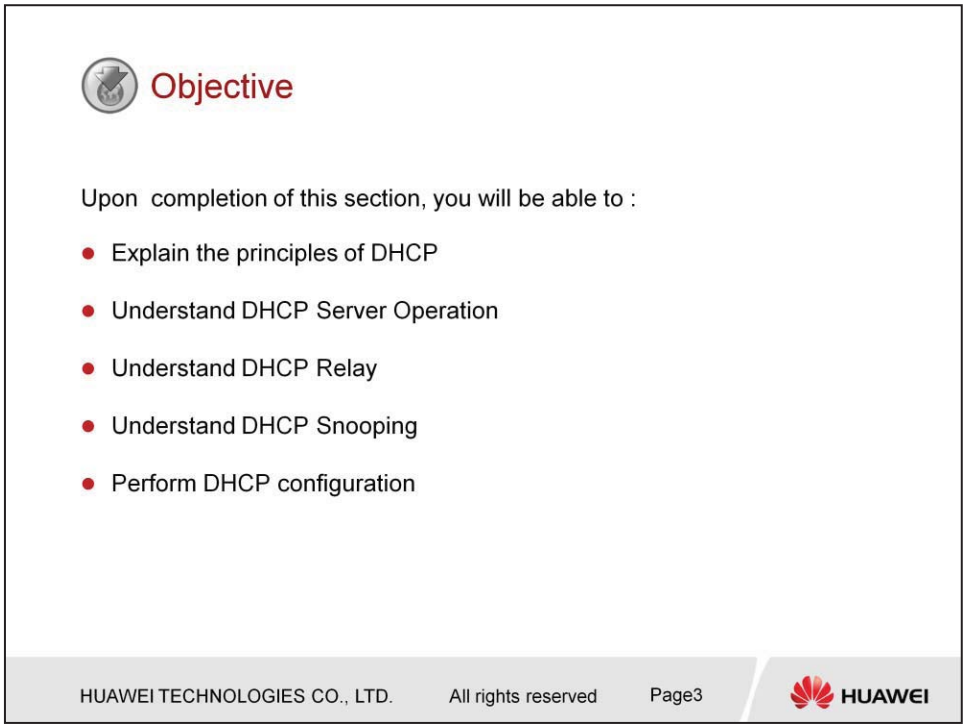
EAP packets include Request, Response, Success, and Failure packets, identified through the **Code** field.

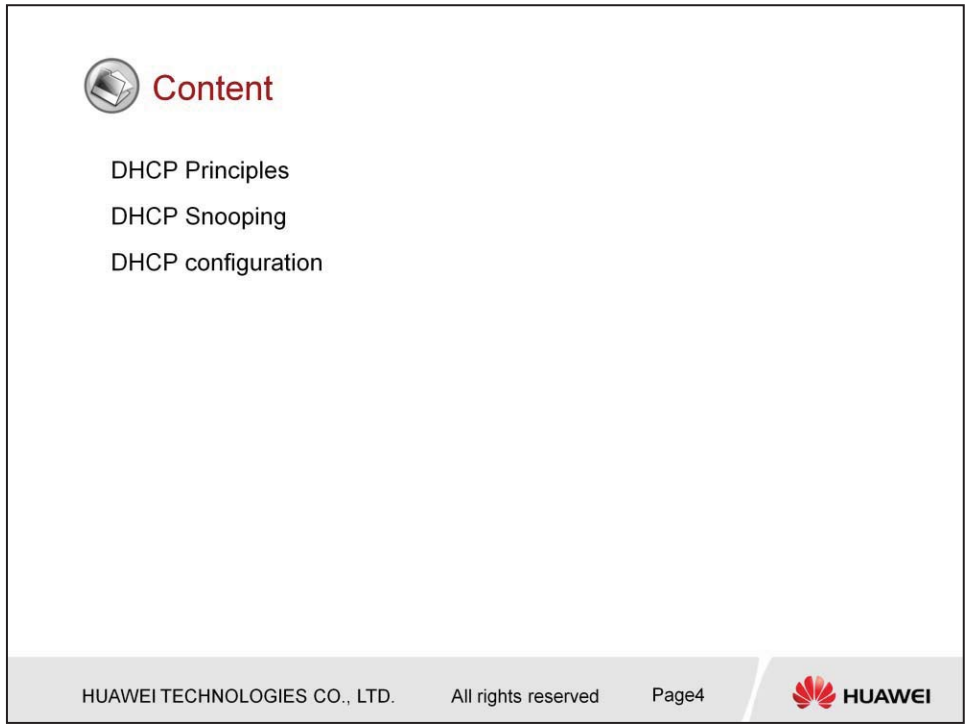
What are the two access authentication modes of 802.1x?

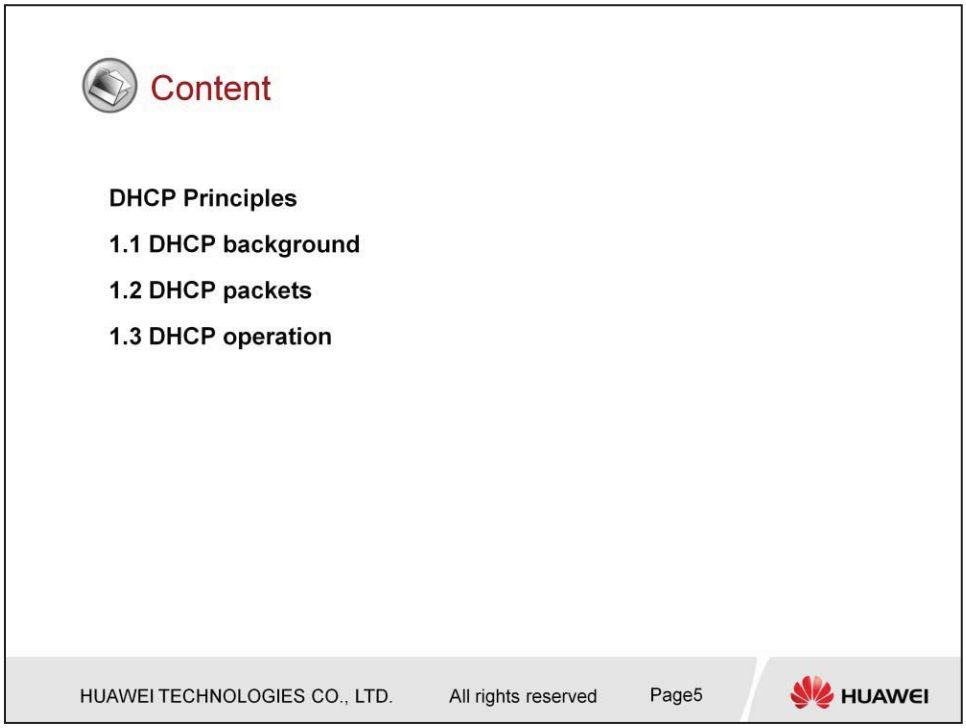
Relay authentication and termination authentication.

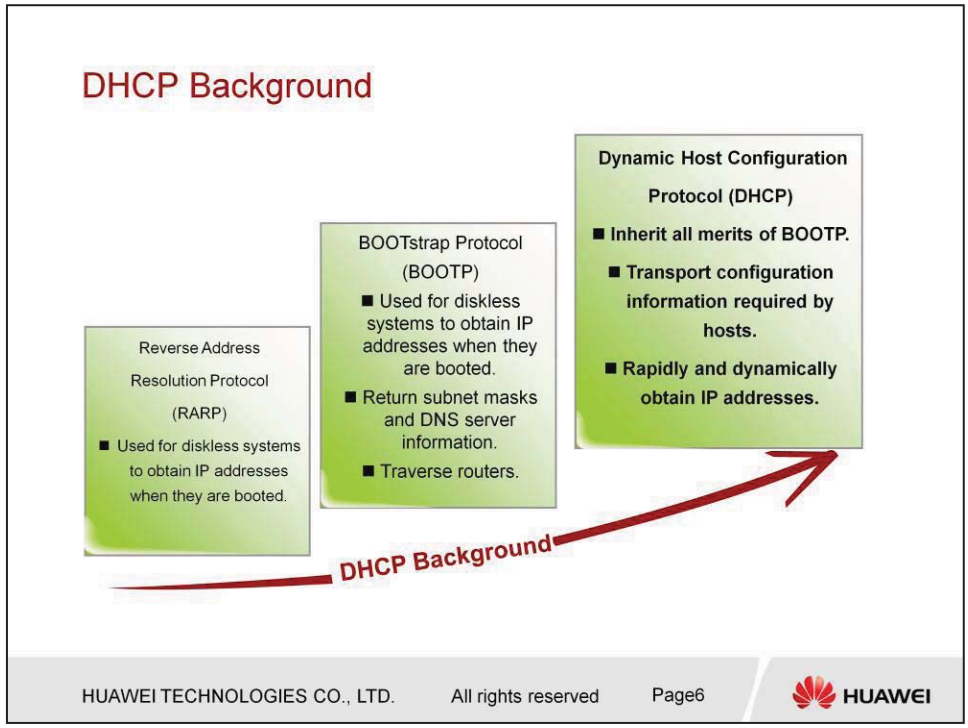

DHCP Principles and Application
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI


Foreword
DHCP is short for Dynamic Host Configuration Protocol, is a centralized dynamic management and configuration technology.
DHCP technology is used to ensure the rational allocation of IP addresses, thus avoiding IP addresses wastage, and improve the utilization rate of the IP address.
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page2
HUAWEI


Objective
Upon completion of this section, you will be able to :
Explain the principles of DHCP
Understand DHCP Server Operation
Understand DHCP Relay
Understand DHCP Snooping
Perform DHCP configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page3
HUAWEI


Content
DHCP Principles
DHCP Snooping
DHCP configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page4
HUAWEI


Content
DHCP Principles
1.1 DHCP background
1.2 DHCP packets
1.3 DHCP operation
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page5
HUAWEI



Origin of DHCP: RARP→BOOTP →DHCP

Reverse Address Resolution Protocol (RARP):

RARP is used for diskless systems to obtain IP addresses when they are booted. An RARP request is broadcast across the network. The hardware address of the sender is indicated in the request packet so that an IP address is allocated. The response is usually returned in unicast mode. The format of RARP packets is almost the same as that of ARP packets.

BOOTstrap Protocol (BOOTP)

The BOOTP packet is enclosed in a standard UDP datagram. BOOTP allows a diskless system to discover its own IP address. It also returns other types of information such as the IP address of a router, the subnet mask of a client, and the IP address of a DNS server. Compared with RARP, it enables the discovery of more information and traversal of routers.

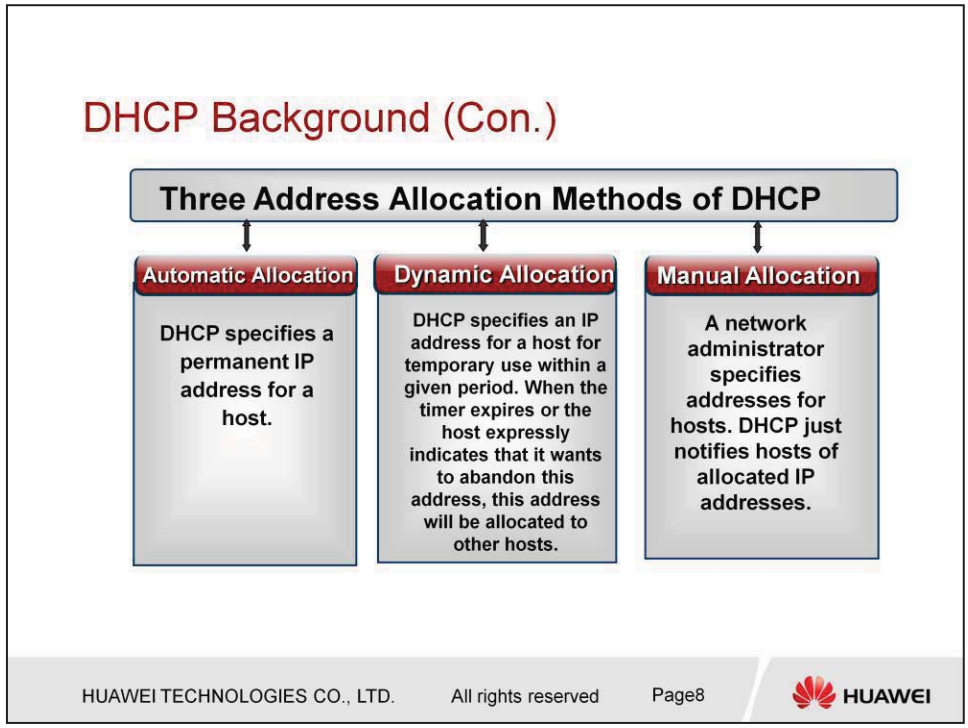
Dynamic Host Configuration Protocol (DHCP):

DHCP evolves from BOOTP. It inherits the merits of BOOTP. The main improvement lies in dynamic allocation.

of IP addresses to users. It is applicable to not only diskless systems but also systems moving between networks.

DHCP is an extension to BOOTP. First, DHCP enables a computer to obtain all desired configuration information through a message. That is why DHCP is known as a configuration transport protocol. Second, DHCP allows a computer to rapidly and dynamically obtain the IP address, which is known as the dynamic IP address assignment mechanism.

DHCP is based on the client-server model. The specified DHCP server allocates network addresses and transport configuration parameters to dynamically configured hosts. A host can function as a server only when the system administrator configures the host as a DHCP server.



Address allocation modes supported by the DHCP server:

Automatic allocation: In this mode, IP addresses are not manually allocated. When a DHCP client obtains an IP address from the DHCP server for the first time, this address permanently belongs to this DHCP client and will not be assigned to other clients.

Dynamic allocation: When a DHCP client leases an IP address from the DHCP server, the DHCP server assigns a temporary IP address to the client. When the lease term expires, this address is returned to the DHCP server and available for other clients. If the DHCP client needs an IP address later, maybe it need to request another IP address.

Manual allocation: In this mode, the network administrator assign specific IP addresses manually for specific DHCP clients on the DHCP server. When the DHCP client requests a IP address for network access, the DHCP server will transmits the manually configured IP address to the DHCP client.

Among the three modes, only dynamic address allocation allows reuse of IP addresses that have been allocated to hosts but are currently not in use. An address needs to be allocated to the host connected to the network temporarily or a group of limited IP

addresses need to be allocated to the hosts that do not require permanent IP addresses. When a new host needs to access a network permanently and there are limited IP address resources, dynamic allocation is a good choice because the IP address of the host can be recovered when this host is out of service in future.

Sequence of IP address assignment

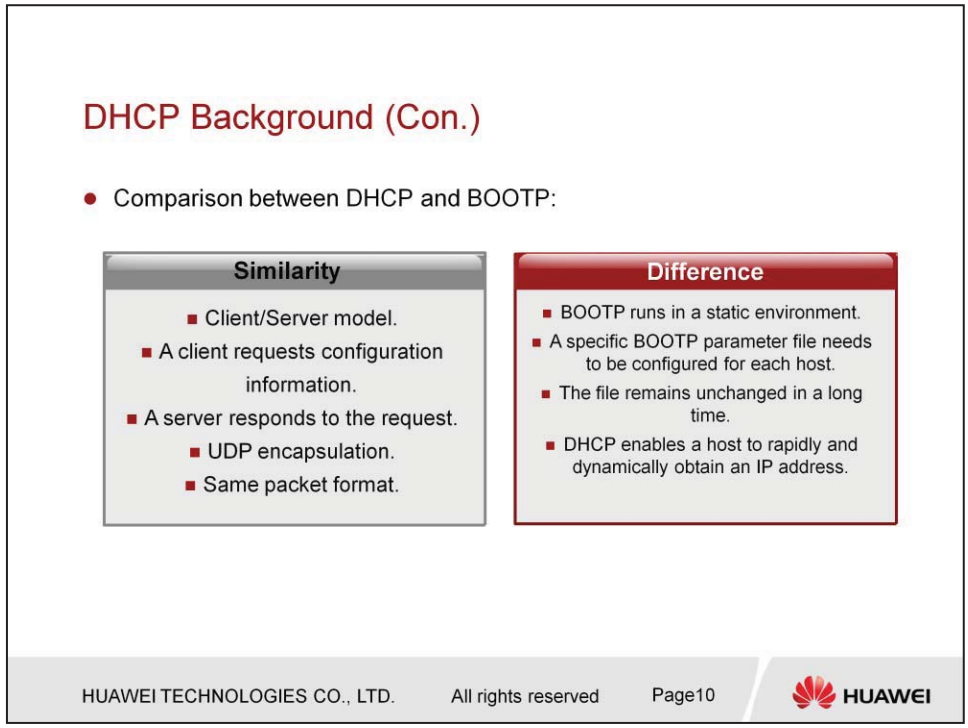
The DHCP server assigns IP addresses to a client in the following sequence:

IP address that is in the database of the DHCP sever and is statically bound to the MAC address of the client

IP address assigned to the client before, that is, the IP address in the requested IP address option of the DHCPDISCOVER packet sent by the client

IP address that is first found when the DHCP server searches the DHCP address pool for available IP addresses

If the DHCP address pool has no available IP address, the DHCP server searches the timeout IP addresses and conflicting IP addresses for an unused IP address, and then assigns the IP address to the client. If all the IP addresses are used, an error is reported.



Comparison between DHCP and BOOTP:

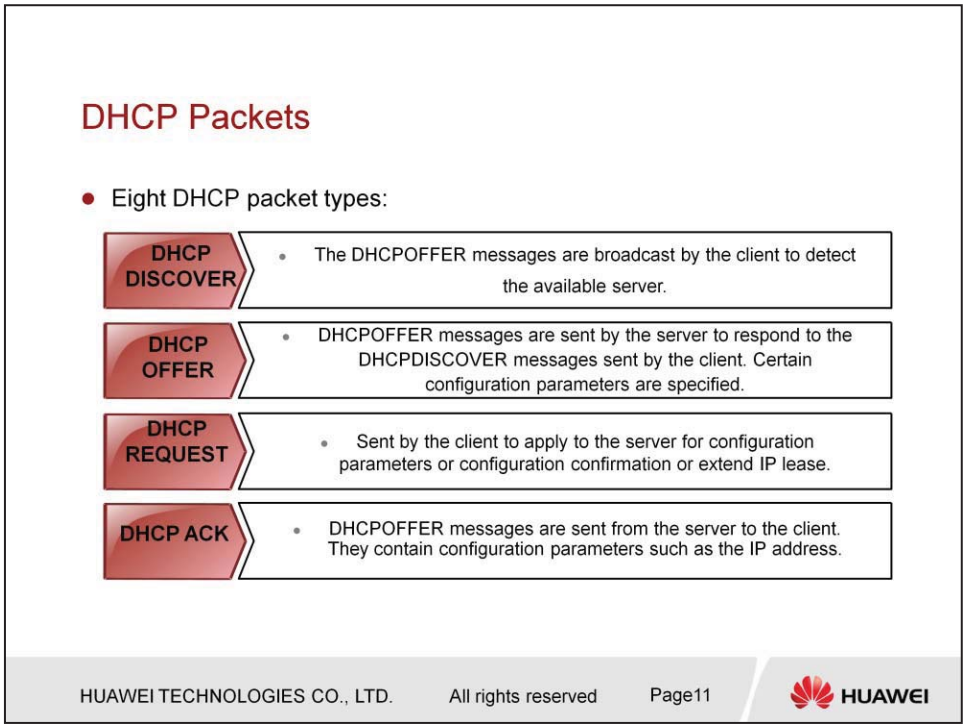
Similar to BOOTP, DHCP also works in client/server mode. A DHCP client requests configuration information from a DHCP server, including such parameters as the assigned IP address, the subnet mask, and the default gateway. The DHCP server replies with corresponding configuration information based on the routing policy. BOOTP and DHCP packets are both encapsulated in UDP packets and are of the same packet structure. Both DHCP and BOOTP use two well-known port numbers: the port number of the server is 67 and the port number of the client is 68.

BOOTP runs in a relatively static environment where each host's physical location is fixed. The administrator configures a specific BOOTP parameter file for each host. This file remains unchanged in a long period of time. DHCP enables a host to obtain an IP address dynamically rather than specifies an IP address for each host.

Compared with BOOTP, DHCP has the following extensions:

DHCP can help a host to obtain all the required configuration information by sending only one message.

DHCP enables a host to obtain an IP address dynamically rather than specifies an IP address for each host.



Any dynamic protocol has a set of specified languages, that is, protocol. DHCP is no exception.

DHCP packets have the following types:

DHCP DISCOVER: It is the first packet used to search for a DHCP server when a DHCP client accesses the network for the first time.

DHCP OFFER: Sent by the server to respond to the DHCP Discover messages sent by the client. Certain configuration parameters are specified.

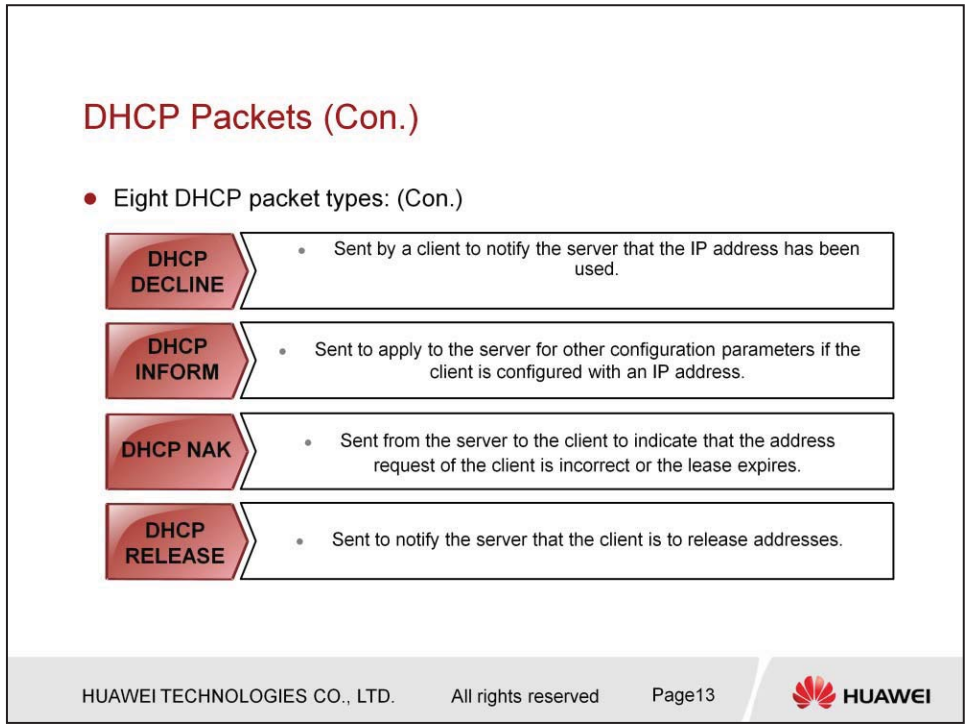
DHCP REQUEST: Sent by the client to apply to the server for configuration parameters or configuration confirmation or extend IP lease. The functions are as follows:

After being initialized, a DHCP client broadcasts a DHCPREQUEST packet to respond to the DHCPOFFER packet sent by a DHCP server.

After being restarted, a DHCP client broadcasts a DHCPREQUEST packet to confirm the correctness of the previously allocated IP address.

After being bound to an IP address, a DHCP client sends a unicast DHCPREQUEST packet to extend the lease of the IP address.

DHCP ACK: It is sent by a DHCP server to acknowledge the DHCPREQUEST packet from a DHCP client. After receiving a DHCPACK packet, the DHCP client obtains the configuration information, including the IP address.



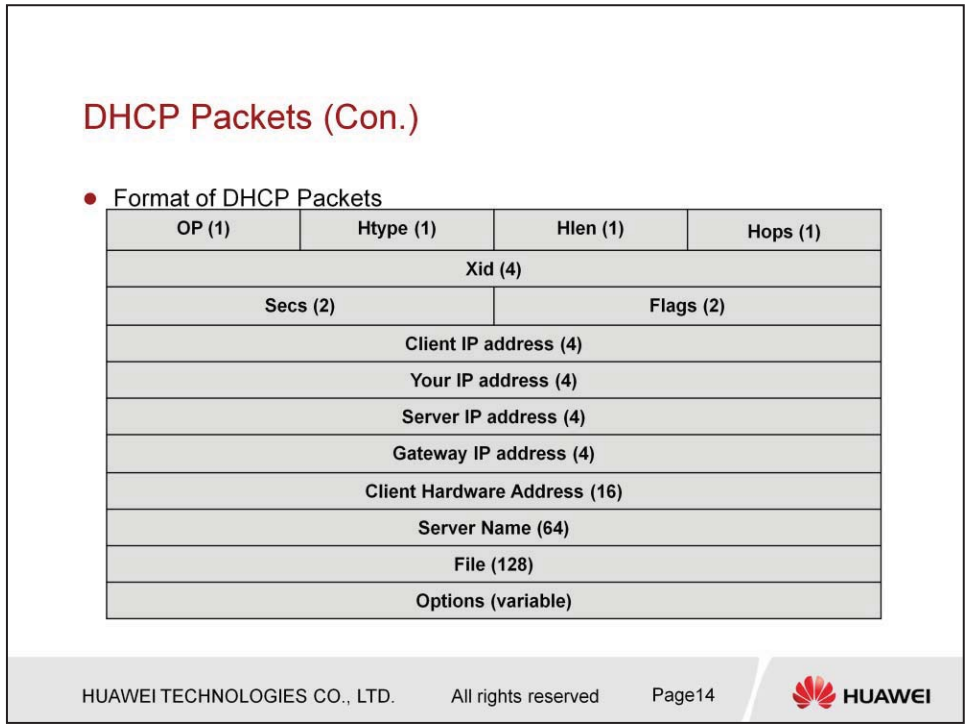
Four other types of DHCP packets:

DHCP DELINE: It is sent by a DHCP client to notify the DHCP server that the assigned IP address conflicts with the other IP addresses. Then, the DHCP client applies to the DHCP server for another IP address.

DHCP NAK: It is sent by a DHCP server to refuse the DHCPREQUEST packet sent by a client. Generally, the DHCPNAK packet is used in the following cases:

DHCP INFORM: After obtaining an IP address, a client sends this packet to obtain other network configuration information, such as the gateway address and DNS server address, from the server.

DHCP RELEASE: It is sent by a DHCP client to actively release the IP address assigned by a DHCP server. After receiving a DHCPRELEASE packet, the DHCP server assigns this IP address to another DHCP client.



OP: operation code (1 = bootrequest, 2 = bootreply)

Htype: type of hardware address (1 = 10mb ethernet)

Hlen: length of hardware address (10 for Ethernet)

Hops: indicates the number of DHCP relay agents that a DHCP packet passes through. This field is set to 0 by the client. The value of this field is increased by 1 each time the DHCP packet passes a DHCP relay. This field is used to limit the number of DHCP relay agents that a DHCP packet can pass through. The number of DHCP relays between a server and a client must be not more than 4. That is, the number of hops must be no greater than 4. Otherwise, the DHCP packet is discarded.

Xid: transmission ID, which is selected by a client for interaction with the server.

Secs: time that has elapsed since the IP address used by a client was last obtained or updated.

Flags: This field is reserved and not used in BOOTP. In DHCP, it indicates a flag. Only the most significant bit of the Flags field is of significance and other bits are set to 0. The leftmost bit of the Flags field is the broadcast response flag. The meanings of the values are as follows:

0: indicates that the client requests the server to send a response packet in unicast mode.

1: indicates that the client requests the server to send a response packet in broadcast mode.

Client IP address: It is used when the client is in bound, renew, or rebinding state and is used to respond to ARP request packets.

Client IP Address: indicates the IP address of a client. It may be an IP address allocated by the server or an existing IP address of the client. In initialization state, the client does not have an IP address. In this case, this field is 0.0.0.0. The IP address 0.0.0.0 is used only for temporary communication during system start-up in DHCP mode. It cannot be an effective destination address.

Your IP address: Client IP address allocated by the DHCP server.

Server IP address: indicates the IP address of a server.

Gateway IP address: indicates the IP address of the first DHCP relay agent. After a client sends a DHCPREQUEST packet, the first DHCP relay agent fills in its IP address in this field when forwarding this DHCPREQUEST packet if the server and the client are on different network segments. The server determines the network segment address based on this field, and then selects the address pool for assigning an IP address to the client. The server also returns a DHCPREPLY packet to the first DHCP relay agent. The DHCP relay agent then forwards the DHCPREPLY packet to the client. If the request passes more than one DHCP relay before arriving at the DHCP server, the non-first DHCP relays increment the hops by 1 rather than change this field.

Client hardware address: indicates the MAC address of a client. This field must be consistent with the fields Hardware Type and Hardware Length. When sending a DHCP Request packet, the client fills in its hardware address in this field. For an Ethernet, this field must be filled in with a 6-byte Ethernet MAC address when the values of the Hardware Type and Hardware Length fields are 1 and 6 respectively.

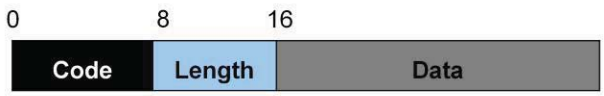
Server name: indicates the name of the server from which a client obtains configuration information. This field is optional and is filled by the DHCP server. If the field is filled in, it must be a character string that ends with 0. By default, the value is empty.

File: indicates the name of the startup file. The full name is provided in the DHCPOFFER packet.

Options: indicates the DHCP Options field. It must be no less than 312 bytes. This field contains configuration information assigned by a server to a client, including the IP addresses of the gateway and DNS server, and the lease of the IP address available to the client.

DHCP Packets (Con.)

- Format of DHCP Packets
  - ⇨ The option field in DHCP messages uses the CLV mode.
    - − Code: identifier of one byte to uniquely identify the information content.
    - − Length: indicates the length of the information content, containing one byte.
    - − Value: indicates the information content, whose length is determined by the **Length** field. It is expressed in bytes.



The option field of a DHCP packet is constructed in CLV mode. Specifically:

Code: identifier of one byte to uniquely identify the information content.

Length: indicates the length of the information content, containing one byte.

Value: indicates the information content, whose length is determined by the Length field. It is expressed in bytes.

This is a flexible format. When new information is needed, just apply for a new option in this mode. The CLV mode is extensible and therefore widely used in protocols.

Common option types:

Packet type: C = 53, L = 1, V = 1–8, indicating a DHCP packet type.

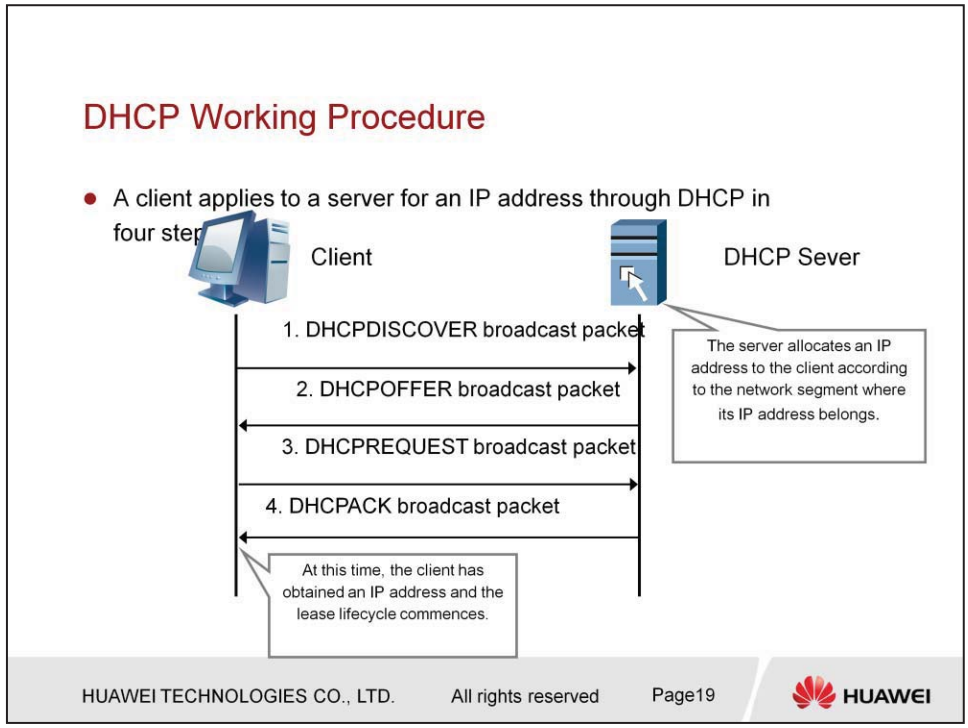
Router IP: C = 3, L = IP address length, V = IP address of the default gateway of a client

DNS IP: C = 6, L = multiple of IP address length, V = IP address sequence of the DNS server of a client

WINS IP: C = 44, L = multiple of IP address length, V = IP

address sequence of the WINS server of a client

DHCP provides a framework for transmitting configuration parameters over a TCP/IP network. Between DHCP clients and the DHCP server, configuration parameters and control information agreed by both parties are transmitted through option codes.



DHCP working procedure:

The client sends a DHCPDISCOVER broadcast packet, with 255.255.255.255 as the destination address, to find a DHCP server on the network.

All DHCP servers receiving the DHCPDISCOVER packet on the network respond to the request in broadcast mode. They select one from available IP addresses and allocate it to the DHCP client by sending a DHCPOFFER packet that contains the IP address for lease and other configuration information.

Upon receipt of the DHCPOFFER packets, the client sends a DHCPREQUEST packet, which contains the IP address it requests from a particular DHCP server. Because there are multiple DHCP servers, the client broadcasts the DHCPREQUEST packet and notifies all DHCP servers that it will accept the IP address provided by a particular DHCP server.

Upon receipt of the DHCPREQUEST packet, the DHCP server sends an ACK response packet to the client.

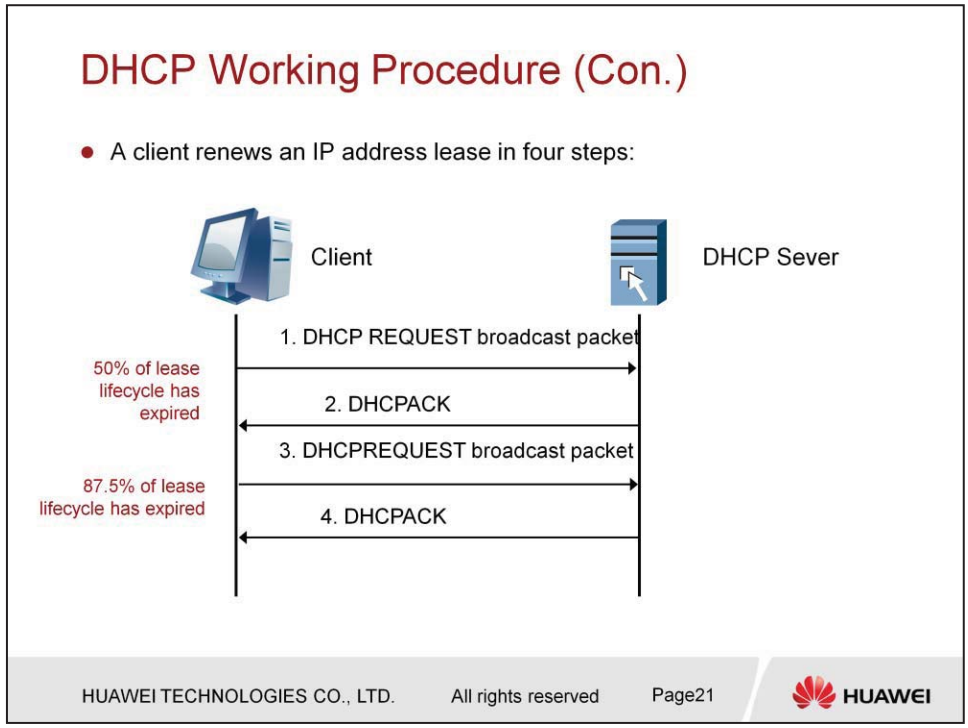
Subsequently, when the DHCP client logs in to the network again, it does not need to send a DHCPDISOVER message again. Instead, it sends the DHCPREQUEST message containing the

IP address last allocated. When the DHCP server receives this message, it attempts to make the DHCP client continue using the original IP address and responds with a DHCPACK message. If this IP address is unavailable (for example, it has been allocated to another DHCP client), the DHCP server responds with a DHCPNAK message to the client. After receiving the DHCPNAK message, the client must re-send a DHCPDISOVER message to request a new IP address.

Compared with that of a DHCP client, the behavior of a DHCP server is simpler, totally driven by the DHCP client. The server just responds to different DHCP packets according to the request packets received from a DHCP client.

IP address allocation principle: Upon receipt of a DHCP request packet, the DHCP server first checks whether the value of the **giaddr** (gateway IP address) field is **0**. If not, it allocates an IP address from the corresponding address pool according to the network segment where the gateway IP address belongs. If the value of the **giaddr** (gateway IP address) field is **0**, the DHCP server thinks that the client is in the same subnet and will then allocates an IP address to the client from the corresponding address pool according to the network segment where its own IP address belongs.

A DHCP server can also implement address pool management.



All IP addresses obtained by clients have a lease lifecycle. If a client does not renew a lease that has expired, the IP address of the client will be recovered by the server.

The process is as follows:

When 50% of the lease lifecycle has expired, the client sends a DHCPREQUEST packet for lease renewal. Because information about the DHCP server has been obtained earlier, this time the DHCPREQUEST packet is unicast.

Upon receipt of the DHCPREQUEST packet, the server sends a response message and resets the lease lifecycle.

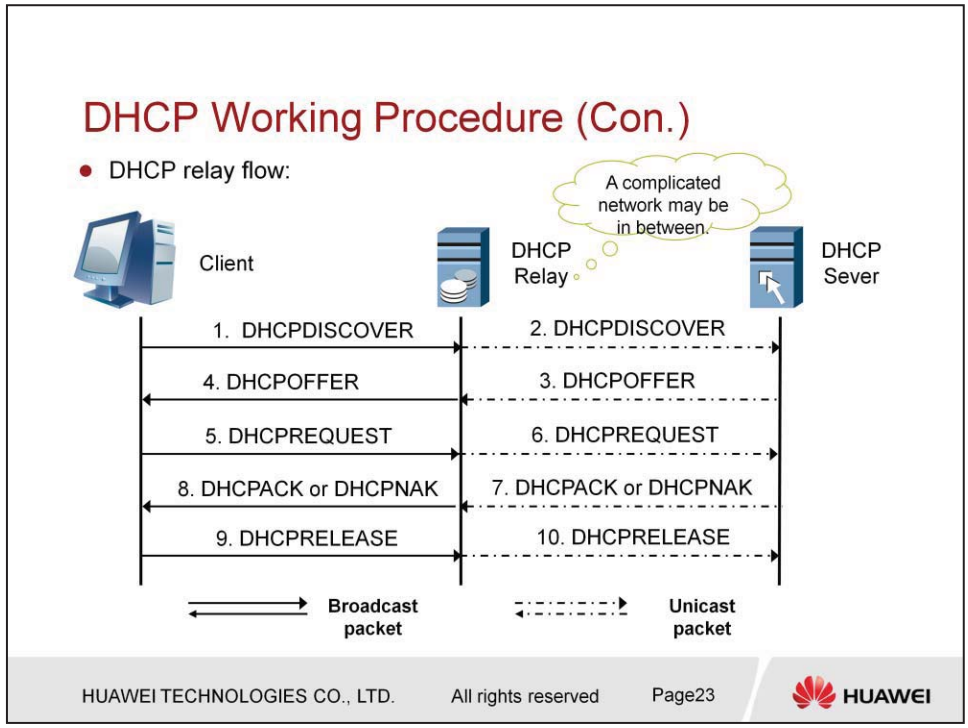
If the server fails to receive a DHCPREQUEST packet when 50% of the lease lifecycle has expired, the lease lifecycle of the IP address will not be reset. If no response to the DHCPREQUEST packet (sent when 50% of the lease lifecycle has expired) is received when 87.5% of the lease lifecycle has expired, the client assumes that the original DHCP server is unavailable and then begins to broadcast a DHCPREQUEST packet. Any DHCP server on the network can reply to this request with a DHCPACK or DHCPNAK packet.

If the client receives a DHCPACK packet, it returns to the binding state and resets the lease update and server rebinding timers.

If all the packets received by the client are DHCPNAK packets, the client returns to the initialization state. At the same time, the client must stop using this IP address immediately. After going back to the initializing state, the client can reapply for an IP address.

If the client does not receive any response before the lease expiration timer expires, the client must stop using the current IP address immediately and return to the initializing state to apply for a new IP address.

After receiving the lease renewal request, the DHCP server sends a DHCPACK message as a response and resets the lease lifecycle of the client.



Application environment of DHCP relay:

Earlier DHCP protocols are applicable only to the situation where DHCP clients and the DHCP server are in the same network segment. Many fields such as **giaddr** and **hop** are unavailable in DHCP packets. Subsequently, the DHCP protocol was extended. Thus, it is necessary to configure a DHCP server in each network segment for the dynamic host configuration. This operation, however, is uneconomical.

The DHCP relay function is thus introduced to address this problem. Through a DHCP relay agent, a DHCP client can apply to the DHCP server on another network segment for a valid IP address. In this manner, DHCP clients on multiple network segments can share one DHCP server. This saves costs and facilitates centralized management.

DHCP packets are generally broadcast and cannot traverse multiple subnets. When a DHCP packet needs to traverse multiple subnets, DHCP relays are needed.

The DHCP relay may be a router or a host. In a word, the DHCP relay must listen to all UDP packets whose destination

port number is 67.

When receiving such a packet, the DHCP relay first determines whether it is a request packet from a client. If it is and the value of the **giaddr** (gateway IP address) field is **0**, the relay fills its own IP address in this field and unicasts the packet to a real DHCP server, thus allowing the DHCP packet to traverse multiple subnets.

If the DHCP relay finds that it is a response packet from the DHCP server, the DHCP relay broadcasts or unicasts the encapsulated packet to the DHCP client, depending on the broadcast flag bit in the **Flags** field.

The working procedure of the DHCP relay is as follows:

A client broadcasts a DHCPDISCOVER packet to find a DHCP server.

Upon receipt of the DHCPDISCOVER packet, the DHCP relay fills its own IP address in the **giaddr** field and then fills the server IP address in the **DHCP Server** field before unicasting the request to the DHCP server.

Upon receipt of the DHCPDISCOVER packet from the DHCP relay, the DHCP server sends a DHCPOFFER packet that contains the IP address of the DHCP server to the DHCP relay.

The DHCP relay sends the received DHCPOFFER packet to the client.

After finding the DHCP server, the DHCP client sends a DHCPREQUEST packet to request an IP address.

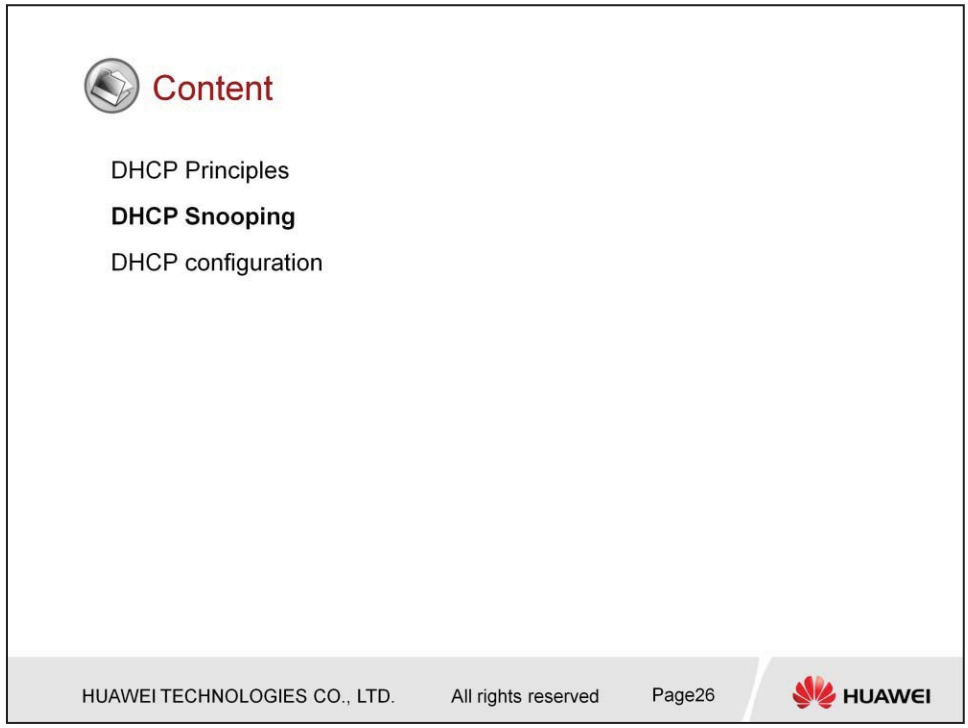
Upon receipt of this request, the DHCP relay fills its own IP address in the **giaddr** address and unicasts the DHCPREQUEST packet to the DHCP server.

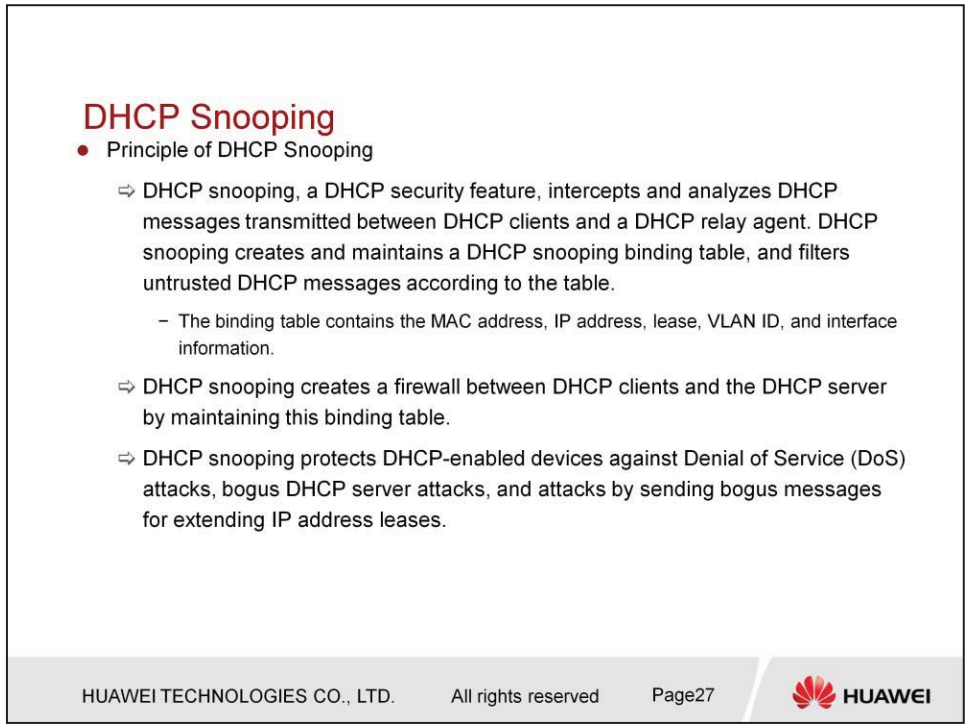
Upon receipt of the request, the DHCP server allocates an IP address to the client by sending a DHCPACK packet. If the requested IP address is unavailable, the server sends a DHCPNAK packet to the DHCP relay. In the packet sent to the DHCP relay, the server fills the IP address allocated to the client in the **Your IP address** field.

Upon receipt of the response packet from the DHCP server, the DHCP relay forwards this packet to the DHCP client.

If the client wants to release the IP address, it sends a DHCPRELEASE packet, whose **Client IP Address** field is filled with the obtained IP address.

Upon receipt of the DHCPRELEASE packet, the DHCP relay fills its own IP address in the **giaddr** field and then sends the packet to the DHCP server. After receiving the request, the DHCP server releases the IP address previously allocated.


Content
DHCP Principles
DHCP Snooping
DHCP configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page26
HUAWEI

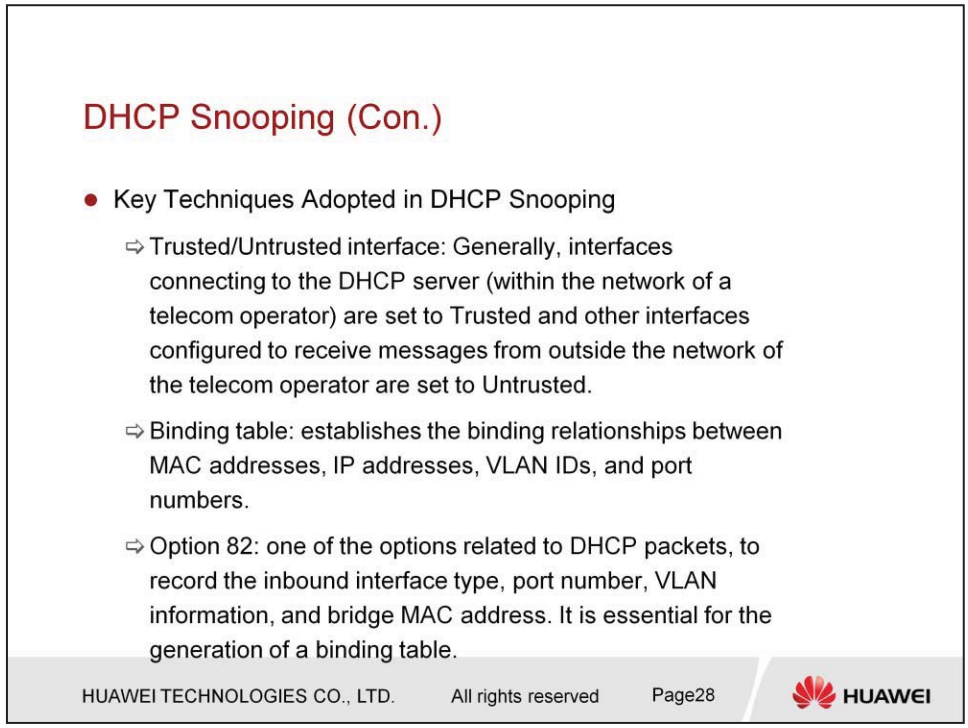


DHCP snooping records a DHCP binding table, which contains the client's IP and MAC addresses, port number, and VLAN ID of the interface that received the client's request in the address allocation process. When a client gets connected, the binding table is created for the client. When the client goes offline, this binding table is deleted.

When or after a binding table is generated, DHCP snooping checks DHCP packets and compares the fields in the packet with those in the DHCP binding table. If a DHCP attack is detected, this packet will be discarded, thus preventing the DHCP attack.

In addition, DHCP snooping distinguishes trusted interfaces and untrusted interfaces. Interfaces connecting to the DHCP server (within the network of a telecom operator) are set to Trusted and other interfaces configured to receive messages from outside the network of the telecom operator are set to Untrusted. Relay IP Address is added for only DHCP relay devices. DHCP relay packets whose giaddr filed is not 0 from untrusted interfaces are not processed to prevent bogus DHCP server attacks.

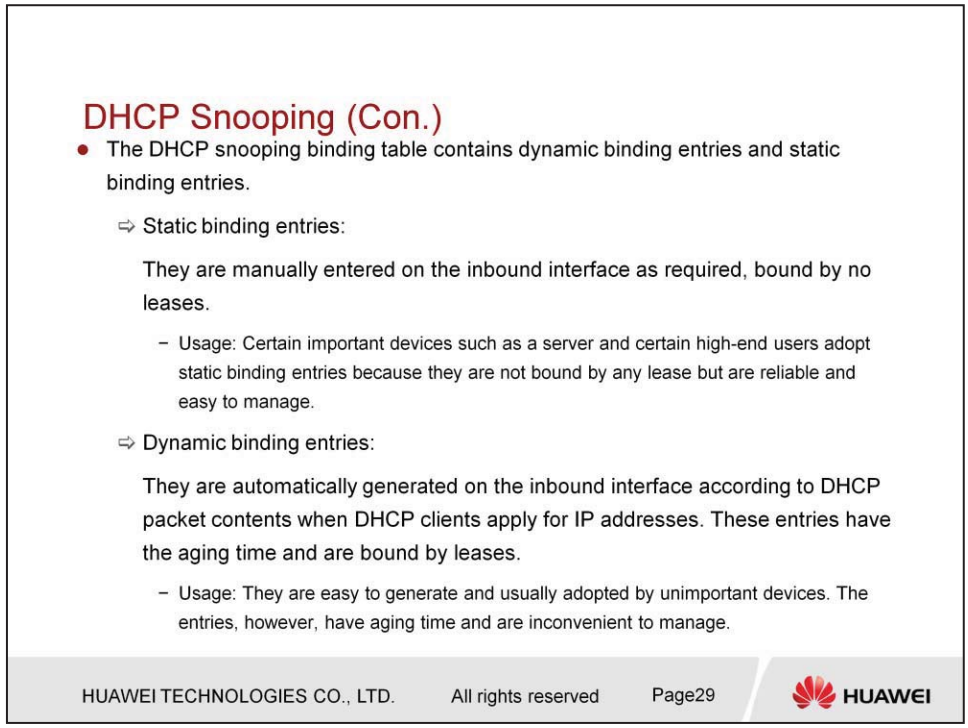
DHCP snooping entries can help prevent IP address spoofing and ARP spoofing.



The **giaddr** field in the DHCP packets destined to trusted zones is not **0**. The **giaddr** field in the DHCP packets received from untrusted zones is **0**.

DHCP snooping is mainly used to prevent DHCP Denial of Service (DoS) attacks, bogus DHCP server attacks, ARP middleman attacks, and IP/MAC spoofing attacks when DHCP is enabled on the device.

DHCP snooping can be applied to both layer 2 and layer 3 network devices.



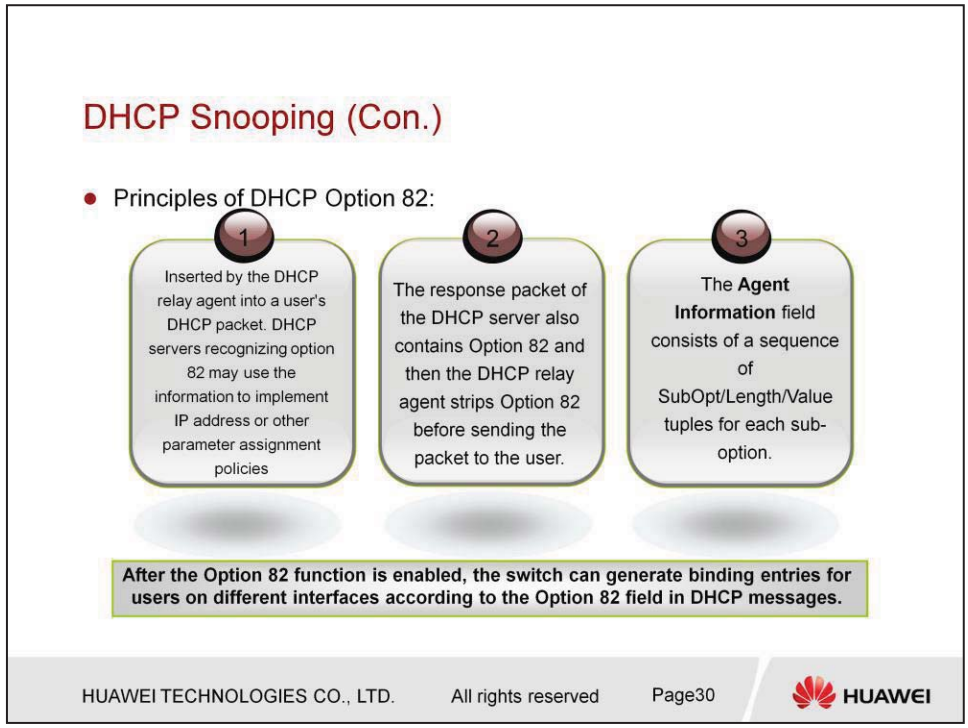
Static binding:

If static IP addresses are allocated to clients, you can configure static binding entries for these allocated IP addresses to prevent certain users from stealing these static IP addresses. In case a great number of clients need static IP addresses, you need to configure IP addresses for them one by one. Otherwise, illegal users cannot be prevented from stealing static IP addresses.

Before forwarding the data of the users who are assigned IP addresses statically, a switch cannot automatically learn the MAC addresses of the users or generate binding table entries for these users. You need to create the binding table manually.

Dynamic binding:

Dynamic entries in the DHCP snooping binding table do not need to be configured. They are automatically generated when DHCP snooping is enabled. For IP addresses dynamically allocated to clients, the DHCP-enabled device automatically learns the MAC addresses of clients and creates a binding relationship table. In this case, a binding table does not need to be configured.



RFC 3046 (DHCP Relay Agent Information Option) proposes the application of Option 82. It is inserted by the DHCP relay agent when forwarding client-originated DHCP packets to a DHCP server. Servers recognizing the Relay Agent Information option may use the information to implement IP address or other parameter assignment policies. The DHCP server echoes the option back verbatim to the relay agent in server-to-client replies, and the relay agent strips the option before forwarding the reply to the client. The Agent Information field consists of a sequence of SubOpt/Length/Value tuples for each sub-option. Currently, two sub-options are defined as follows:

1 Agent Circuit ID Sub-option: identifies the circuit of a user.

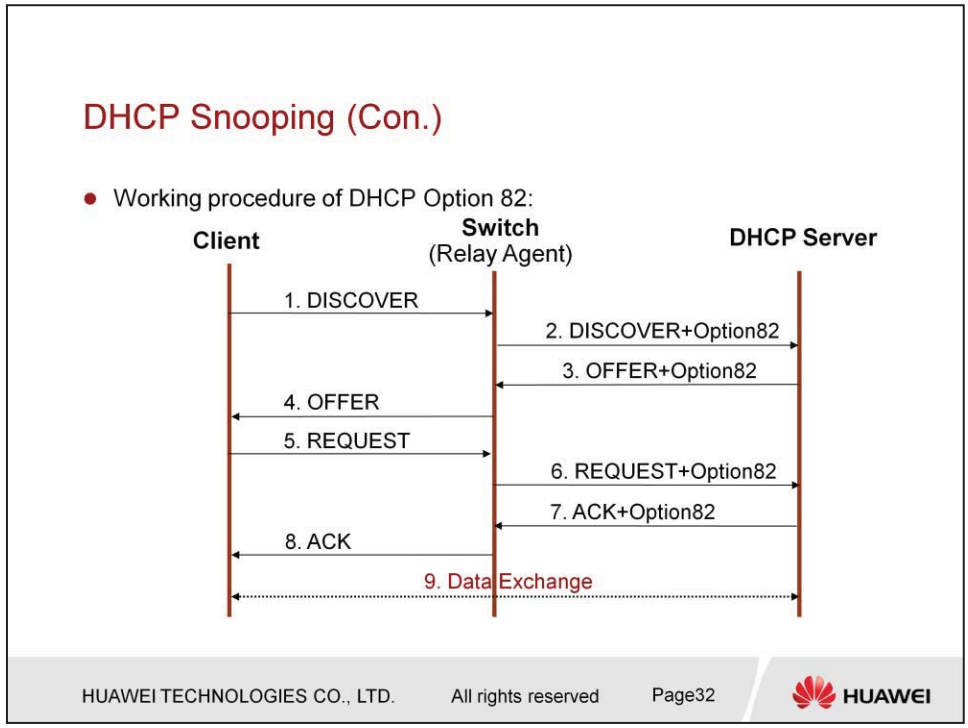
2 Agent Remote ID Sub-option: identifies the host at the remote end of the circuit.

When the Option 82 function is enabled on a DHCP relay, if Option 82 constructed by a user does not contain interface information, the binding table generated does not have such information. This may cause:

DHCP reply packets of servers to be listened by users under other interfaces in the same VLAN.

After the user goes online, if another user under a different interface in the same VLAN fabricates an IP address and MAC address, this bogus user can disguise himself as the legal user.

When DHCP snooping is used at Layer 2, the interface information can be obtained for the binding table if the Option 82 function is not configured.



A client sends a DHCPDISCOVER packet to find a DHCP server.

Upon receipt of the DHCPDISCOVER packet, the switch functioning as the DHCP relay agent adds Option 82 information to the packet and then forwards it to the DHCP server.

Upon receipt of the DHCPDISCOVER packet, the DHCP server responds with a DHCPOFFER packet that contains Option 82 information previously added.

Upon receipt of the response from the DHCP server, the switch functioning as the DHCP relay agent strips Option 82 information and forwards such a packet to the client.

Now the client has discovered the DHCP server and begins to send a DHCPREQUEST packet to request an IP address.

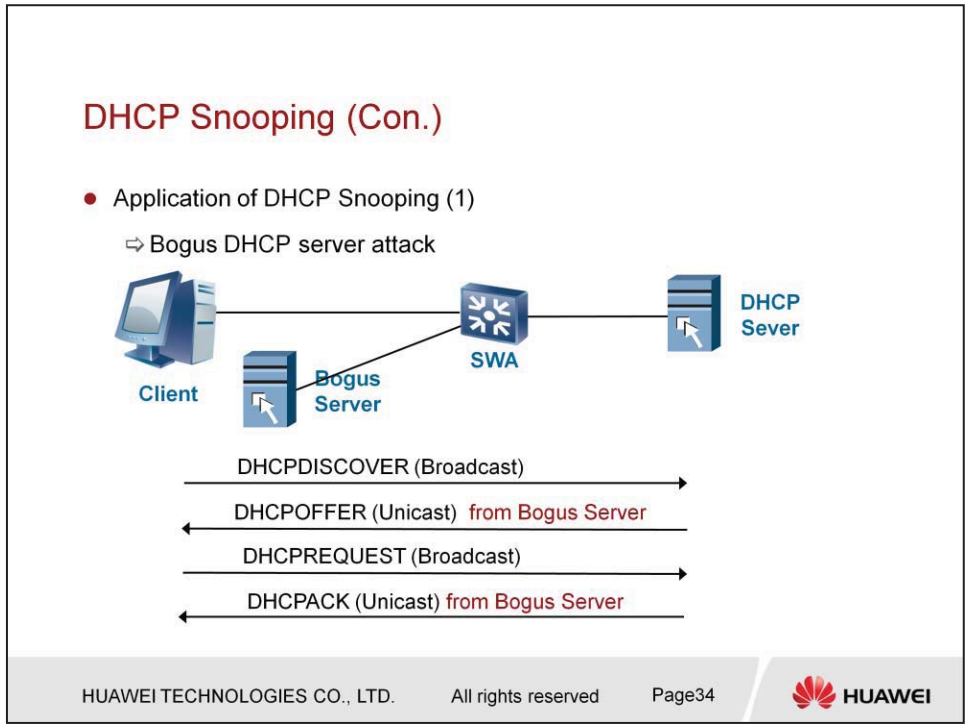
Upon receipt of the DHCPREQUEST packet, the switch functioning as the DHCP relay agent adds Option 82 information to the packet and then forwards it to the DHCP server.

Upon receipt of the DHCPREQUEST packet that contains Option 82 information, the DHCP server allocates an IP address to the client. If the requested IP address is unavailable or its lease lifecycle expires, the server responds with a DHCPNAK packet. Otherwise, the server responds with a DHCPACK packet, which

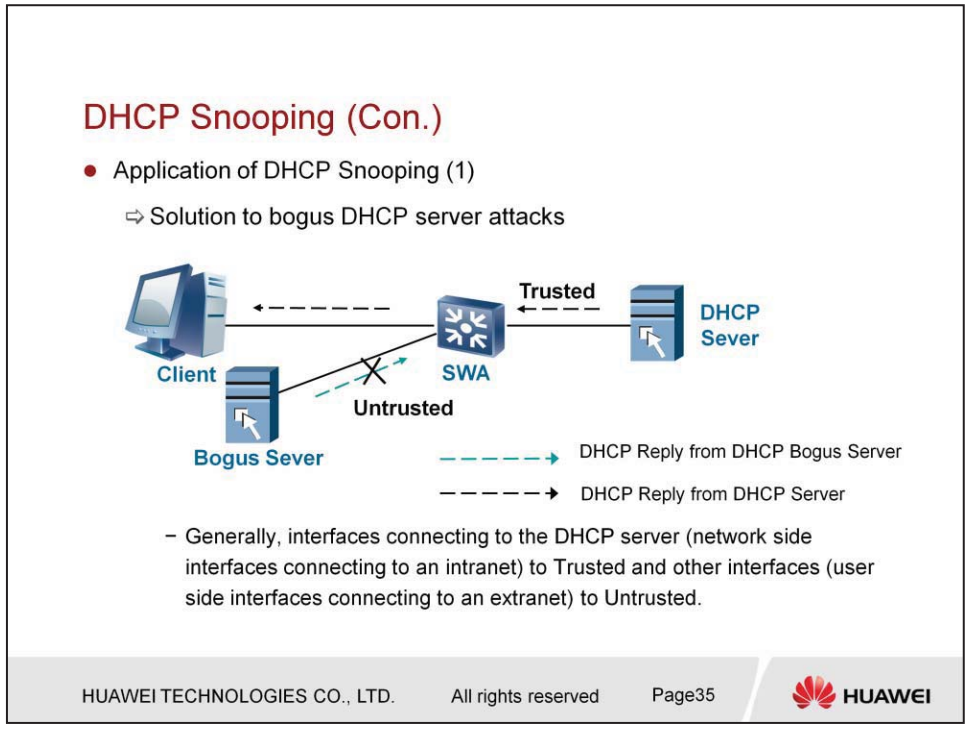
contains Option 82 information.

Upon receipt of the DHCPACK packet that contains Option 82 information, the switch functioning as the DHCP relay agent strips Option 82 information and then forwards the DHCPACK packet to the client.

After the address request process is completed, the client can exchange data with others.

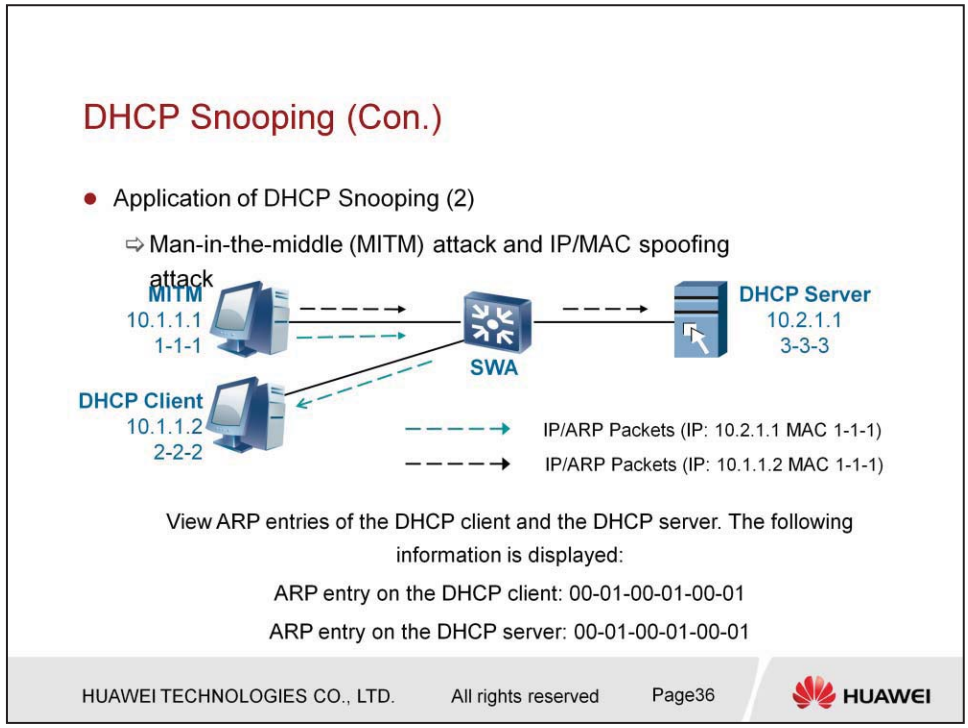


DHCP REQUEST packets are sent in broadcast mode. The bogus DHCP server can thus listen to the DHCP REQUEST packets. The bogus DHCP server then replies with incorrect packets with the incorrect IP address of the gateway, incorrect DNS server, and incorrect IP address to the DHCP client. This causes the Denial of Service (DoS).



To prevent bogus DHCP server attacks, you can configure trusted and untrusted interfaces.

You can set a physical interface or VLAN to Trusted or Untrusted. The DHCP reply messages (OFFER, ACK, or NAK messages) received from an untrusted interface are directly discarded so that the attacks from the bogus DHCP server can be prevented.



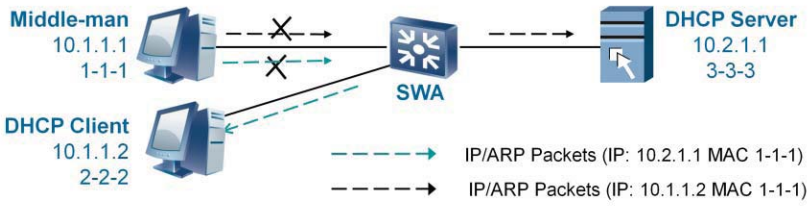
As shown in the figure, an MITM makes the DHCP server learn the IP address of the DHCP client, 10.1.1.2, and its own MAC address, 1-1-1, by sending an IP or ARP packet. From the perspective of the DHCP server, all packets are from or to the DHCP client although all packets are processed by the MITM.

The MITM re-sends an IP or ARP packet, making the DHCP client learn the IP address of the DHCP server, 10.2.1.1, and its own MAC address, 1-1-1. From the perspective of the DHCP client, all packets are from or to the DHCP server although all packets are processed by the MITM.

In this manner, the MITM disguises itself as the DHCP server and DHCP client and obtains interaction information between the DHCP server and the DHCP client.

DHCP Snooping (Con.)

- Application of DHCP Snooping (2)
  - ⇨ Solutions to MITM attack and IP/MAC spoofing attacks

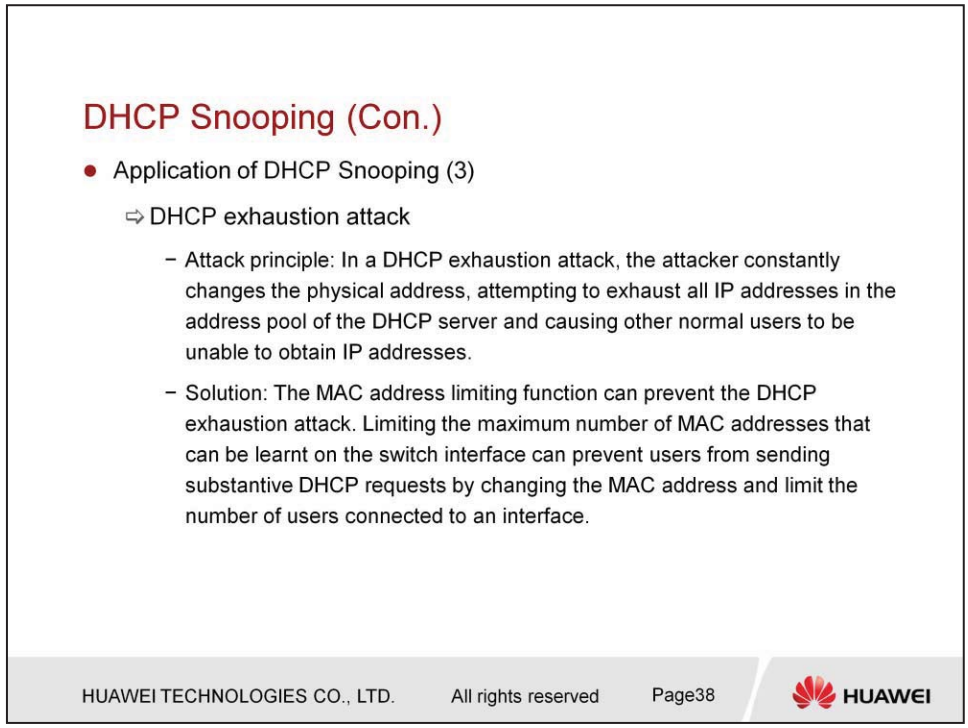


To prevent the MITM attack or IP/MAC spoofing attack, you can configure the DHCP snooping function on the switch. The binding table function of DHCP snooping allows the forwarding of a received packet only when its content is consistent with that in the binding table. If the contents are inconsistent, the packet is discarded.

When the switch receives an ARP or IP packet on an interface, the switch matches the source IP address and source MAC address of the packet with entries in the DHCP snooping binding table. When the strong policy is configured, the switch discards the IP packet if no matching entry is found.

A user that uses a static IP address does not have a matching DHCP snooping binding entry on the switch because the user does not obtain the IP address by sending a DHCP Request message. Therefore, ARP or IP packets sent from this user are discarded to prevent the user from using network resources. To allow the users with statically allocated IP addresses to access the network, you must configure a static DHCP snooping binding table.

A user that embezzles a valid IP address of another client also obtains the IP address without sending a DHCP Request message; therefore, the MAC address and interface of this user do not match the entry corresponding to the IP address in the DHCP snooping binding table. Then the IP packets sent by the embezzler are discarded and the embezzler cannot access the network.



The MAC address limiting function is generally deployed on a layer 2 device. Limiting the maximum number of MAC addresses that can be learnt on the switch interface can prevent users from sending substantive DHCP requests by changing the MAC address and limit the number of users connected to an interface.

MAC address limiting in QinQ mode:

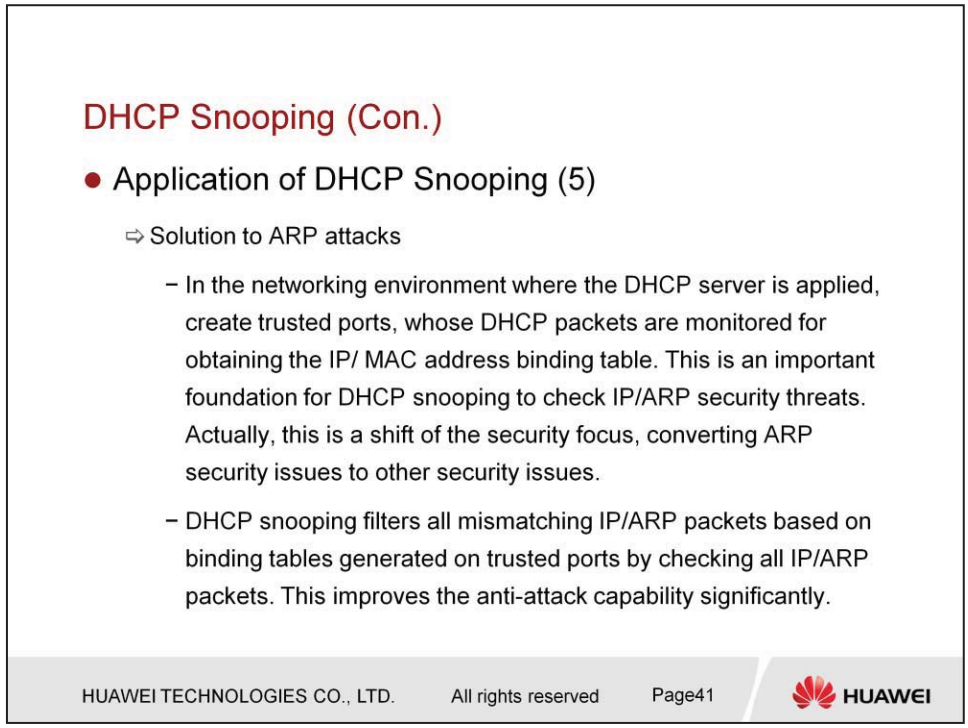
In practice, a client connects to a network through a Digital Subscriber Line Access Multiplexer (DSLAM). The DSLAM isolates clients by configuring different VLANs for clients. In addition, to avoid the limitation of total VLAN number (4094), the QinQ feature is adopted to encapsulate two layers of tags in packets from clients. After that, if the MAC address limiting function is deployed on the gateway, the gateway should be able to limit the number of MAC addresses based on the two-layer tags.

DHCP Snooping (Con.)

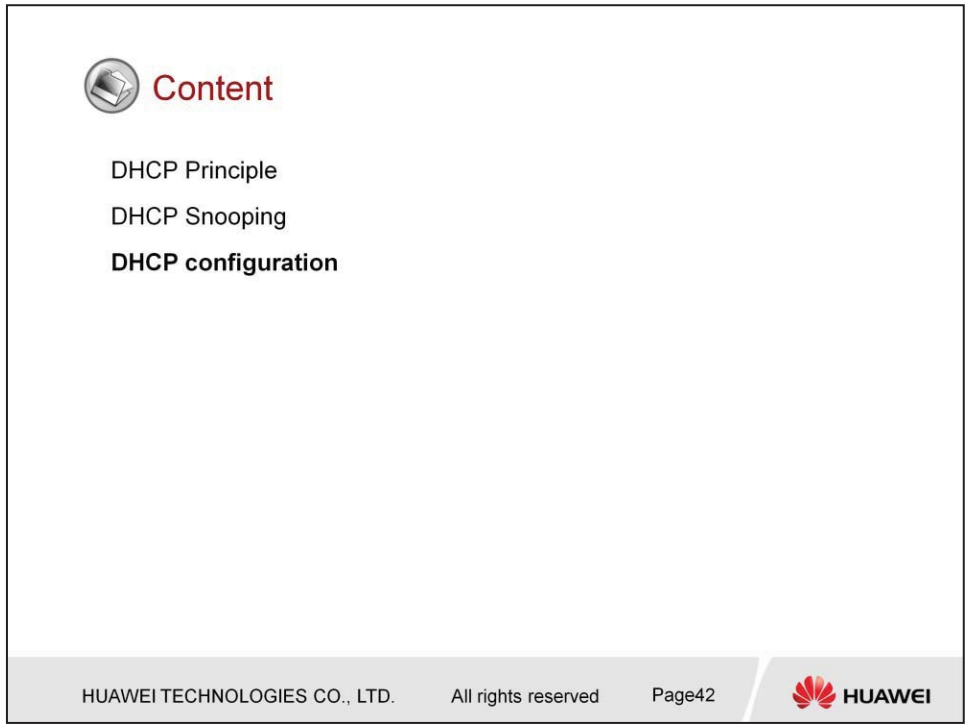
- Application of DHCP Snooping (4)
  - ⇨ DHCP exhaustion attack by changing the Client Hardware Address (CHADDR)
    - − The attacker may change the CHADDR field carried in DHCP messages but not the source MAC address in the frame header to apply for IP addresses continuously. If the device only checks validity of packets based on the source MAC address in the frame header, MAC address limit may not take effect.
    - − Solution: You can configure DHCP snooping on the switch to check the CHADDR field carried in a DHCP Request message. If the value of the field in a packet matches the source MAC address in the data frame header, the packet is forwarded. Otherwise, the packet is discarded.

DHCP Snooping (Con.)

- Application of DHCP Snooping (5)
  - ⇨ Principle of ARP attacks
    - The attacks to ARP are of several types and in multiple modes. The attacks may target at a host or a gateway. The attacks may be performed through address spoofing or violent attacks. The attacks may originate from viruses or illegitimate software.
    - Cause of ARP attacks: The ARP protocol is inherently too simple and open, without any security means, leaving a great many of opportunities for hacker attacks.
    - Impacts of ARP attacks: ARP address spoofing attacks generally target at individual hosts or hosts in a specified scope. Therefore, the impact is comparatively small. ARP DDoS attacks targeting at gateway devices, however, would force a great number of users offline due to the special location of gateways on a network.



In a networking environment where the DHCP server is applied, DHCP snooping achieves a better effect in attack prevention. This is because DHCP snooping shifts the security object from APR that has no security means to DHCP. The DHCP server application environment is more secure than the host application environment. Therefore, trusted ports may be created and DHCP packets of these ports are monitored for obtaining the IP/ MAC address binding table. This is a shift of the security focus, converting ARP security issues to other security issues.


Content
DHCP Principle
DHCP Snooping
DHCP configuration
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page42
HUAWEI

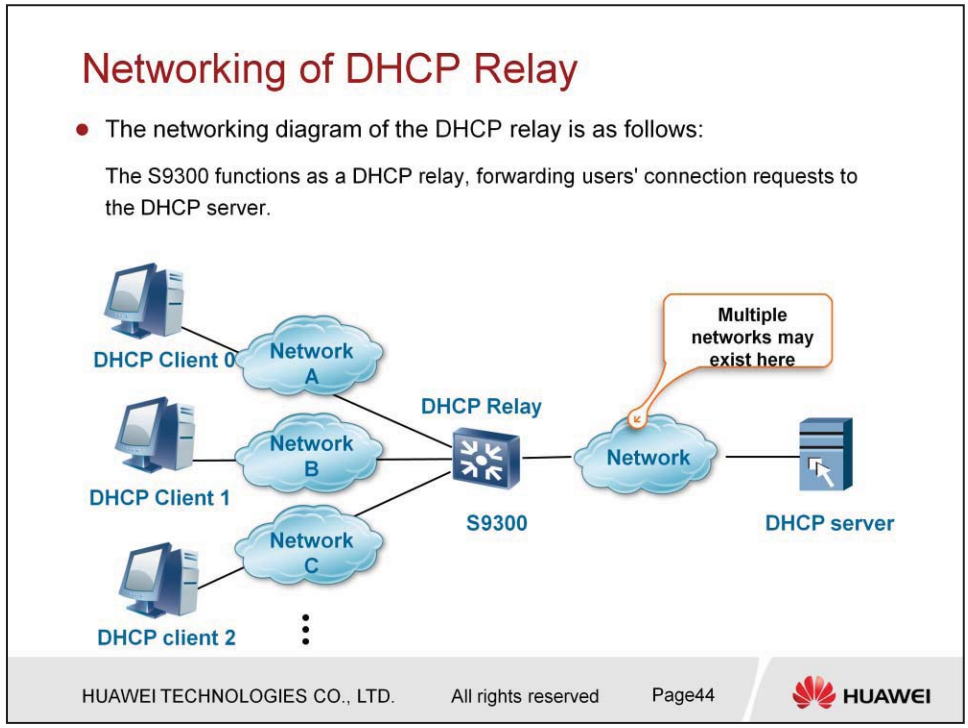
Content

DHCP configuration

3.1 DHCP Relay configuration

3.2 DHCP Server configuration

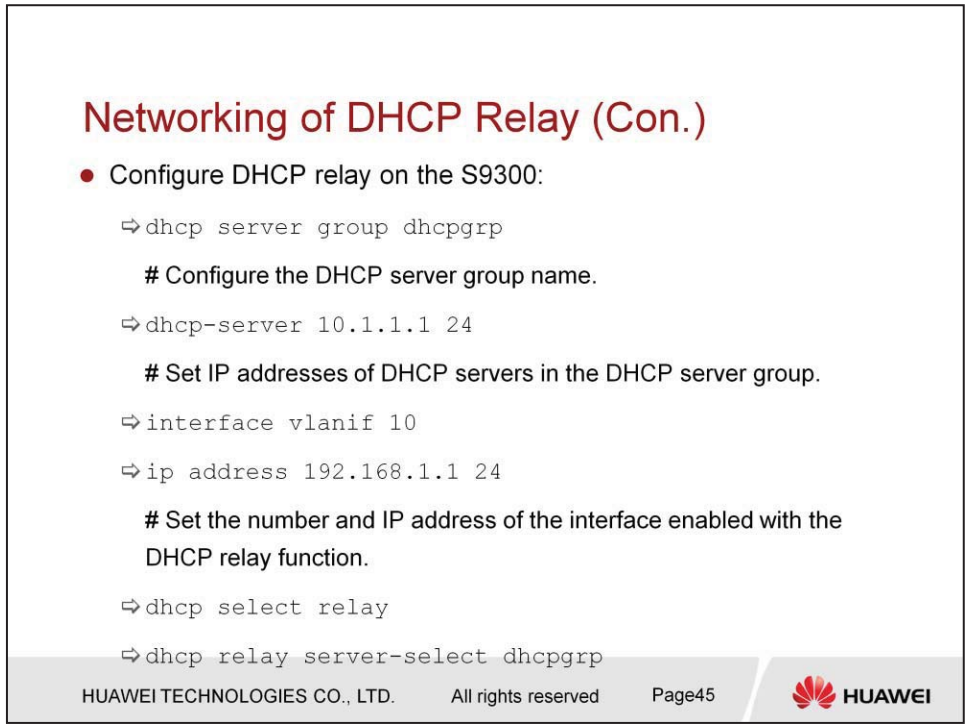
3.3 DHCP Snooping configuration



Usage Scenario of the DHCP Relay

If no DHCP server is configured on a network, the DHCP relay function can be enabled on an S9300. In this manner, the DHCP Request packet from clients can be transmitted to the DHCP server on another network through the DHCP relay agent. To enable clients to obtain IP addresses, the DHCP server must use a global address pool. That is, the interface of the server connected to the DHCP relay agent cannot be configured with any address pool.

The interface address pool takes precedence over the global address pool. If an address pool is configured on an interface, clients obtain IP addresses preferentially from the interface address pool even if a global address pool is configured.



You can configure up to 64 DHCP server groups in the system,

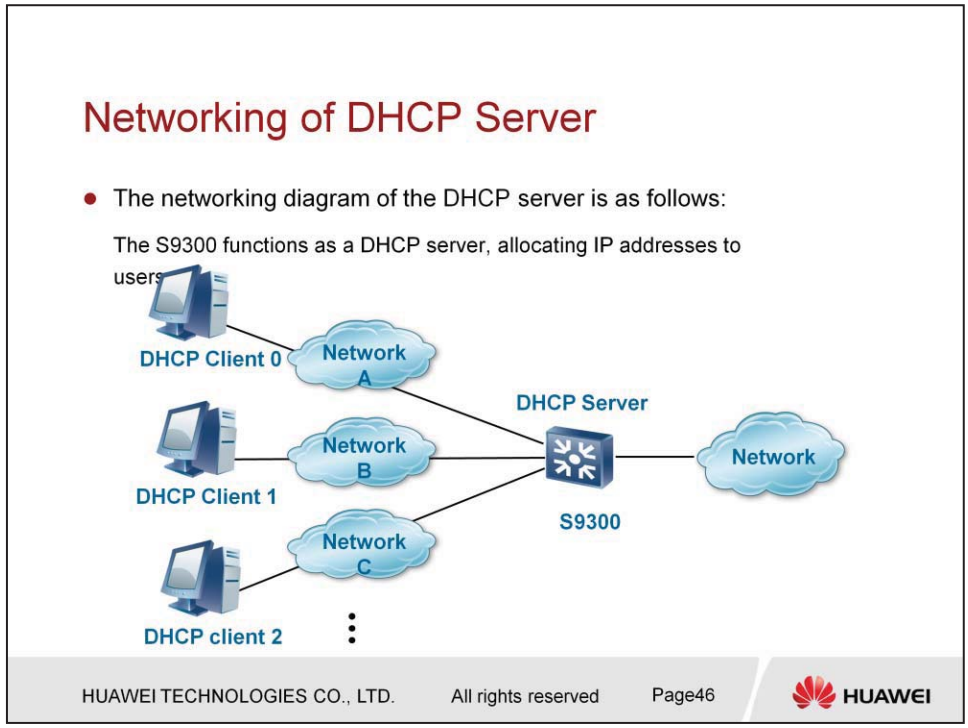
You can configure a maximum of eight DHCP servers in a DHCP server group. If no index is specified, the system automatically allocates an idle index.

After the DHCP relay function is enabled on the VLANIF interface of the S9300, the VLANIF interface relays the packets to the DHCP relay agent.

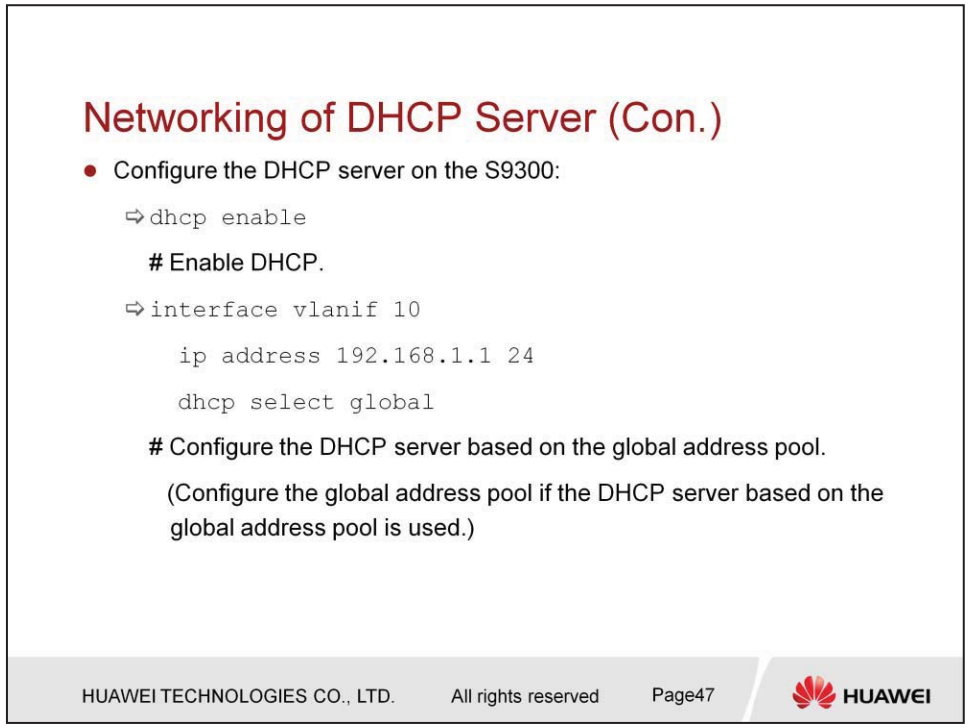
The number of DHCP packet relays between a DHCP server and a DHCP client cannot exceed 4. Otherwise, the DHCP packet would be discarded.

If DHCP relay is enabled in a super VLAN, DHCP snooping cannot be enabled in this super VLAN.

A DHCP server group can correspond to multiple VLANIF interfaces, whereas only one DHCP server group can be specified for a VLANIF interface. That is, DHCP Request messages on a VLANIF interface can be relayed to only one DHCP server.

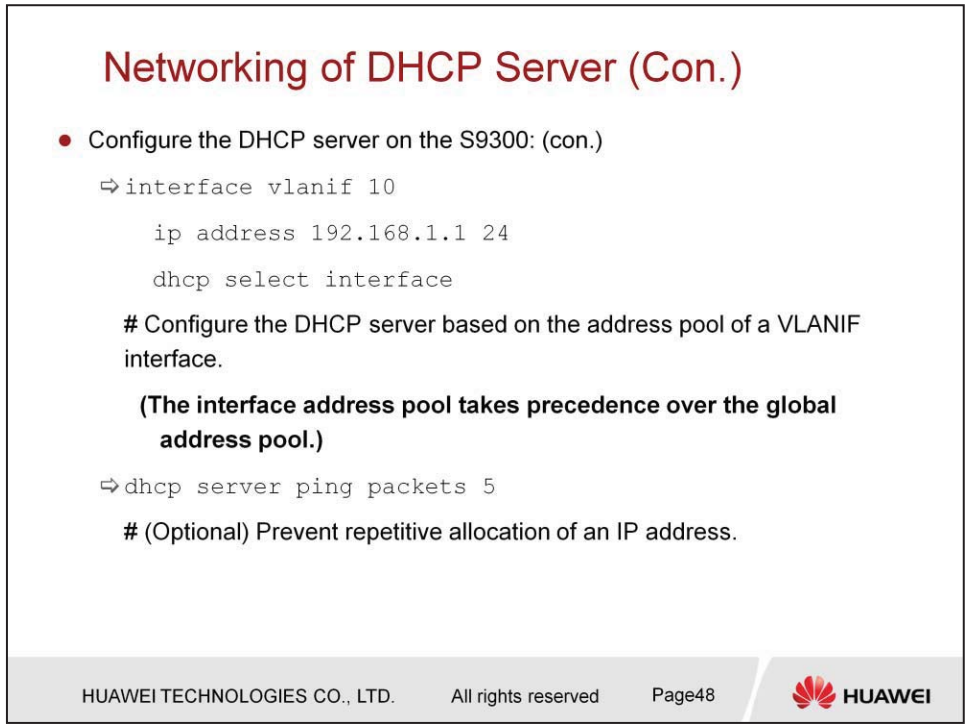


Deploy services near users on the distributed network. As a convergence point, the S9300 assigns IP addresses to users through the local DHCP server or the remote DHCP server to terminate clients on the S9300.



When the S9300 functions as a DHCP server, you can set the mode of processing the DHCP packets whose destination addresses are the IP address of the local device. The S9300 can assign IP addresses through the global address pool or the interface address pool.

The client and the S9300 need to be located on the same subnet and each interface can be configured with only one processing mode.



If a DHCP server based on a global address pool is configured, users going online through any interface can obtain IP addresses from the global address pool.

Configure the global address pool:

ip pool 2

ip pool 1

gateway-list 10.1.1.126

network 10.1.1.0 mask 255.255.255.128

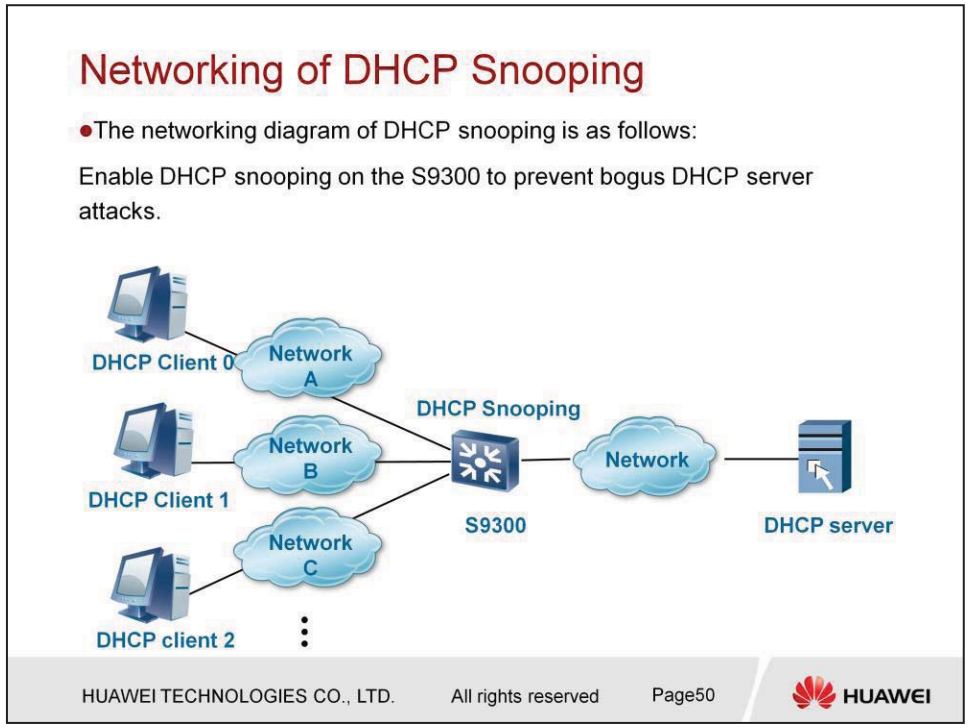
dns-list 10.1.1.2

If a DHCP server based on an interface address pool is configured, all the users going online through this interface obtain IP addresses from the interface address pool. The gateway address is the IP address of this interface.

The interface address pool takes precedence over the global address pool. If an address pool is configured on an interface, clients obtain IP addresses preferentially from the interface address pool even if a global address pool is configured.

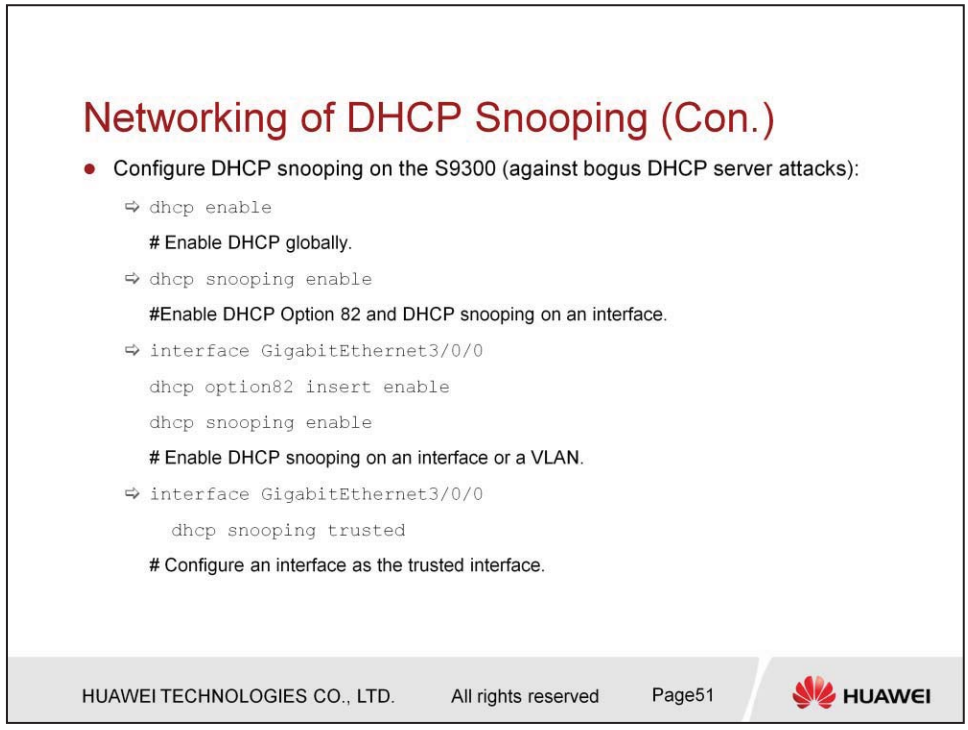
The DHCP server can send ping packets to check whether an IP address is in use. This prevents the repetitive allocation of an IP

address. By default, the maximum number of ping packets is 5 and the longest response-wait time of each ping packet is 0.



Both layer 2 switching function and layer 3 routing feature are supported. In both application scenarios, DHCP snooping can be configured.

When the S9300 is deployed in a layer 2 network or functions as a DHCP relay, enabling DHCP snooping can prevent DHCP attacks. The only difference in configuration is that the S9300 as a DHCP relay supports ARP and DHCP interaction, which is unavailable when the S9300 is deployed at layer 2.

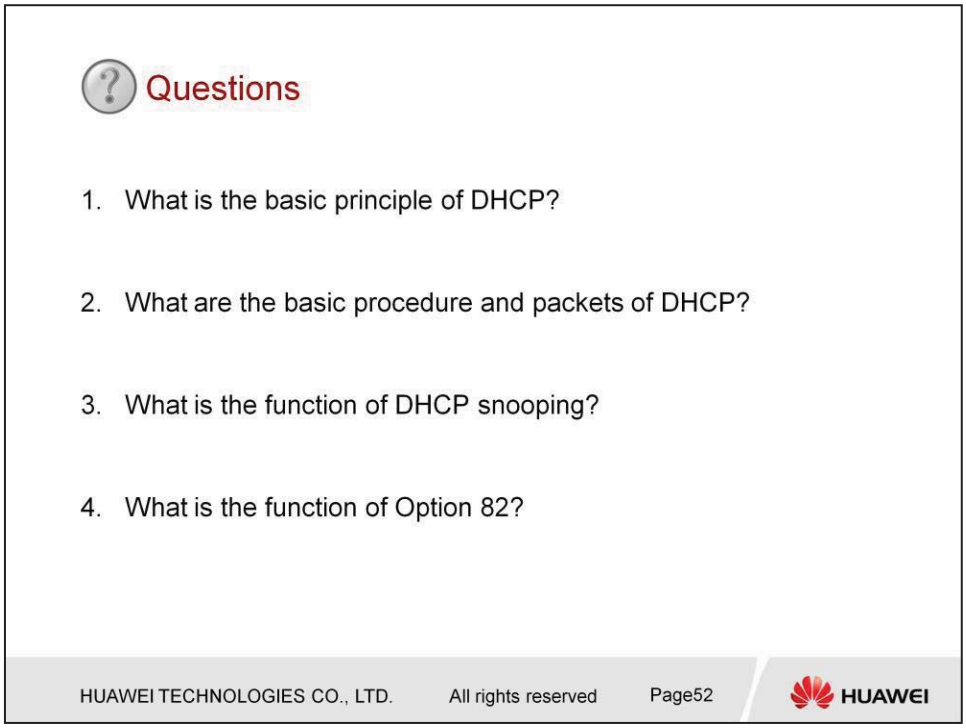


To enable DHCP snooping, you need to comply with the following sequence:

Enable DHCP globally.

Enable DHCP snooping globally.

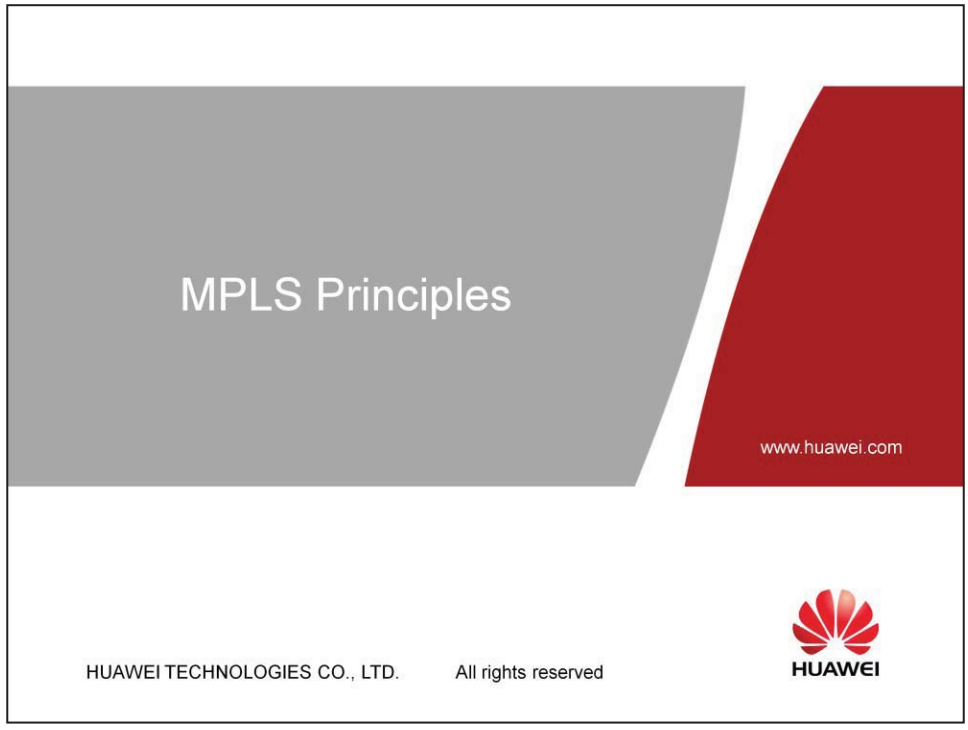
Enable DHCP snooping on an interface or in a VLAN.

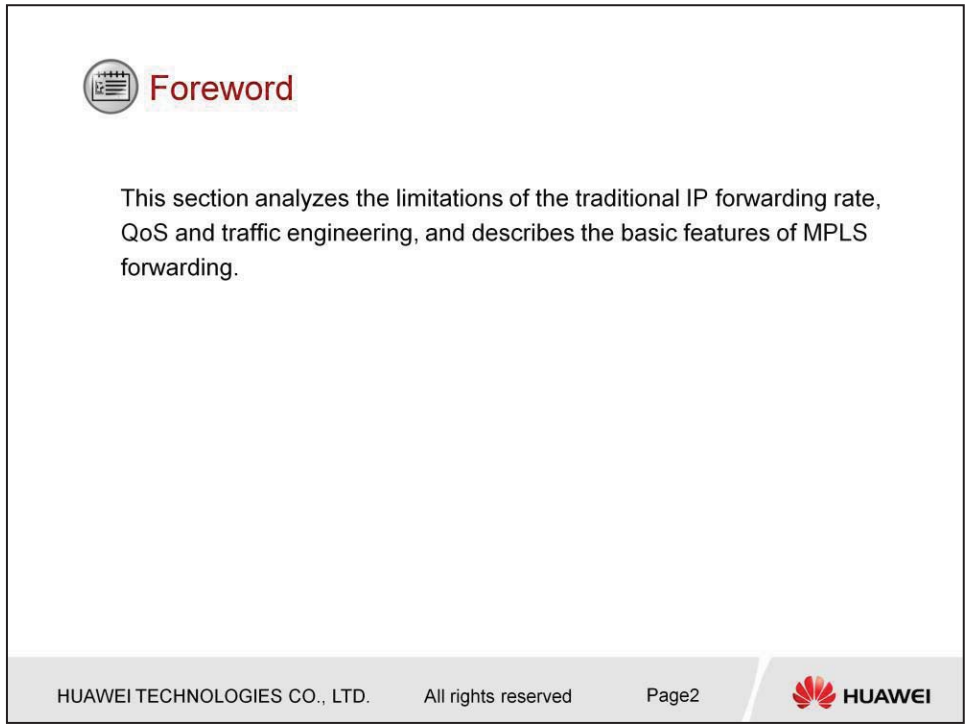


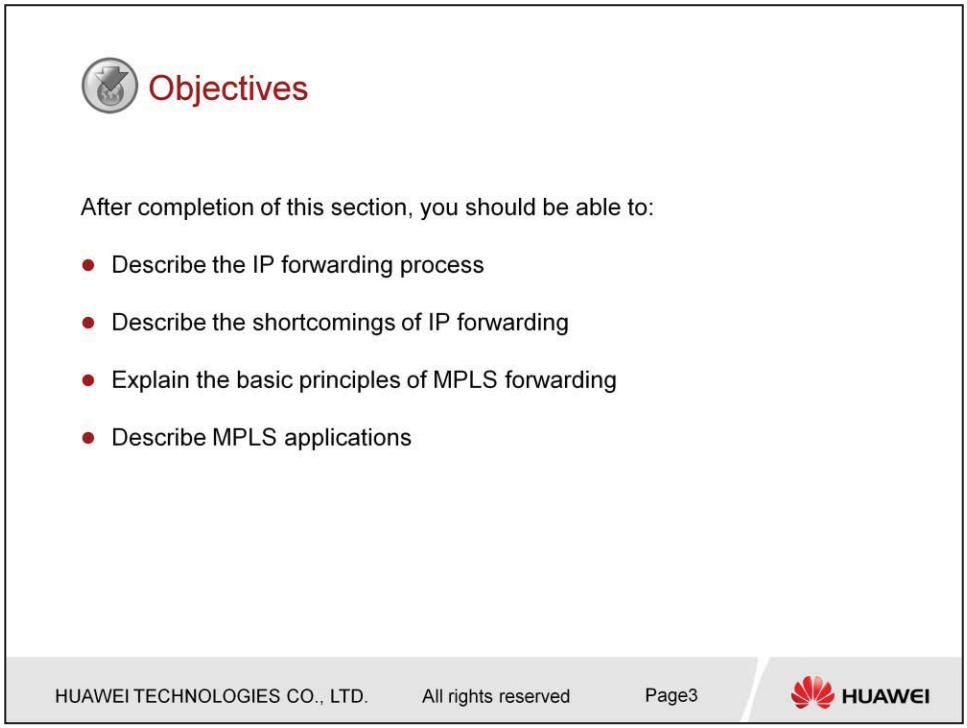
Answer

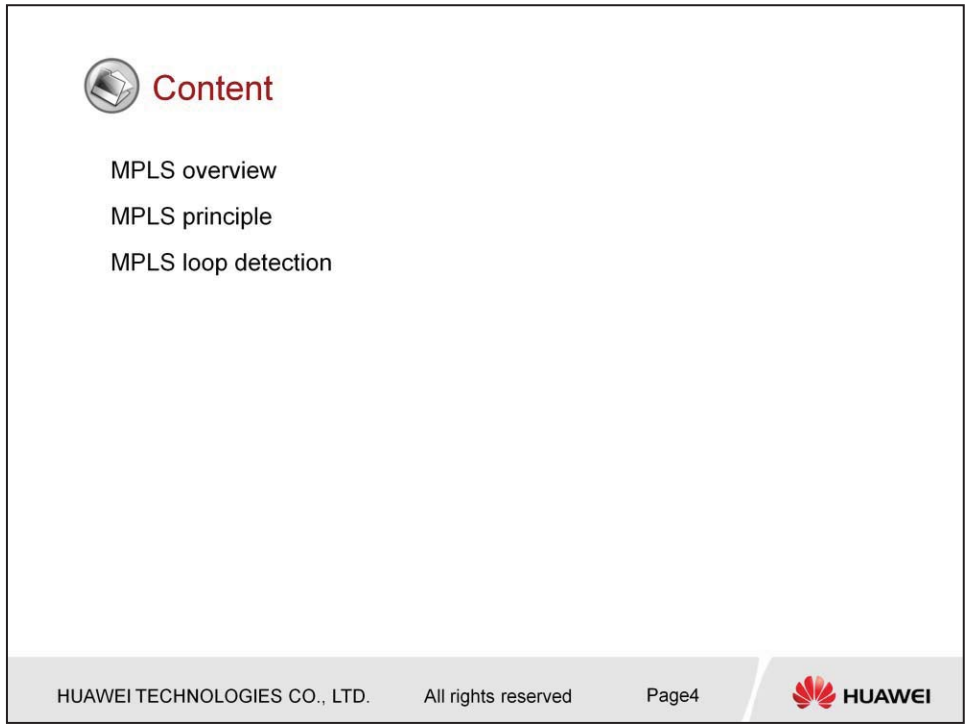
1. The basic principle of DHCP is that the host obtains the corresponding network configuration and IP address through dynamic packet interaction.
2. Basic procedure of DHCP: A host sends a DHCPDISCOVER packet to find a DHCP server. The DHCP server responds with a DHCPOFFER packet. Then the host sends a DHCPREQUEST packet to request an IP address. Upon receipt of the request, the server responds with a DHCPACK packet. When 50% of the lease lifecycle has expired, the host sends a DHCPREQUEST packet to renew the lease. If the lease renewal request is rejected, it sends a DHCPREQUEST packet again to renew the lease when 87.5% of the lease lifecycle has expired.
3. See P21.
4. The function of Option 82 is to encapsulate location information for the implementation of security policies and QoS.

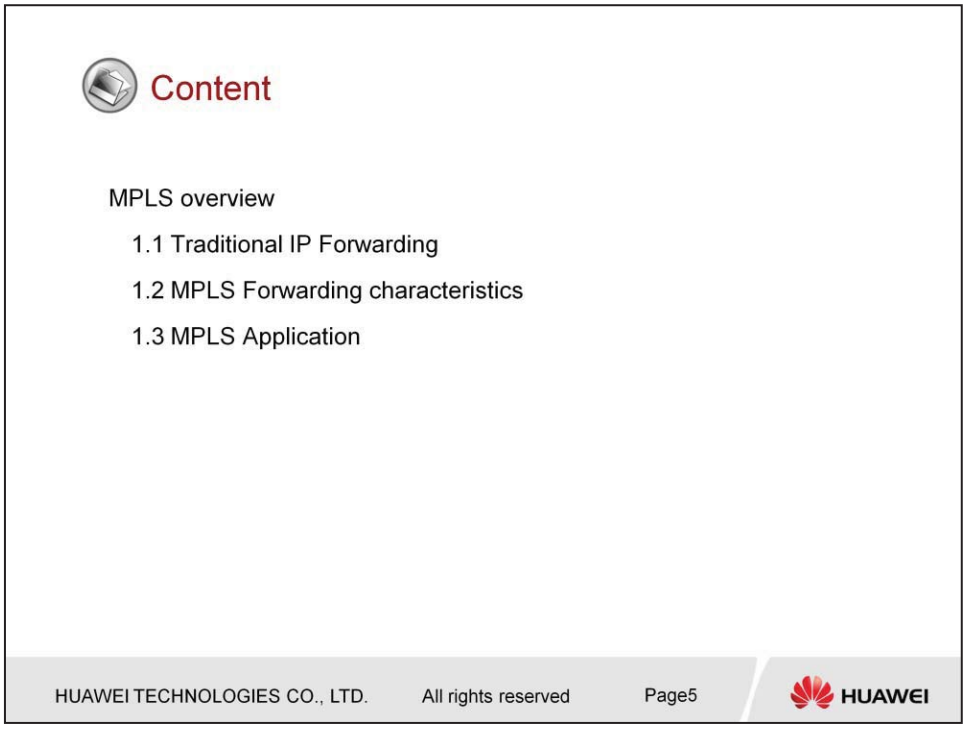
Module 4
MPLS


MPLS Principles
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

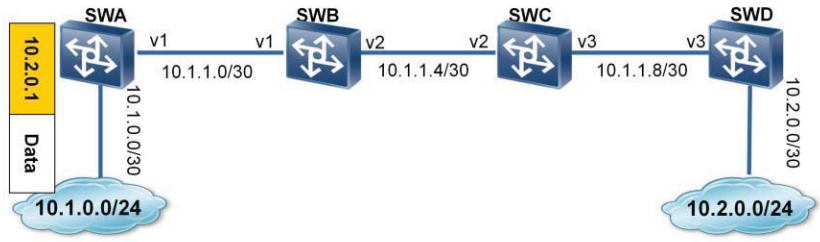

Foreword
This section analyzes the limitations of the traditional IP forwarding rate, QoS and traffic engineering, and describes the basic features of MPLS forwarding.
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page2
HUAWEI


Objectives
After completion of this section, you should be able to:
Describe the IP forwarding process
Describe the shortcomings of IP forwarding
Explain the basic principles of MPLS forwarding
Describe MPLS applications
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page3
HUAWEI

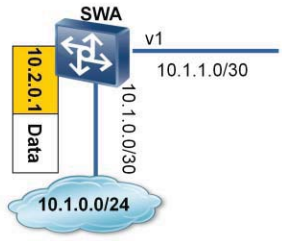

Content
MPLS overview
MPLS principle
MPLS loop detection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page4
HUAWEI


Content
MPLS overview
1.1 Traditional IP Forwarding
1.2 MPLS Forwarding characteristics
1.3 MPLS Application
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page5
HUAWEI

Traditional IP Forwarding

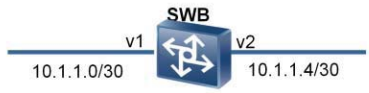


Traditional IP Forwarding(SWA)



Network	Nexthop
10.1.0.0/24	10.1.0.2
10.1.0.1/32	10.1.0.1
10.1.1.0/30	10.1.1.1
10.1.1.2/32	10.1.1.2
10.1.1.4/30	10.1.1.2
10.1.1.8/30	10.1.1.2
10.2.0.0/24	10.1.1.2

Traditional IP Forwarding(SWB)



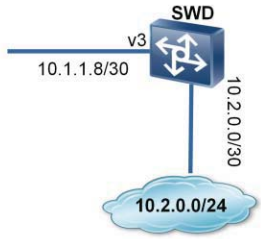
Network	Nexthop
10.1.0.0/24	10.1.1.1
10.1.1.0/30	10.1.1.2
10.1.1.1/32	10.1.1.1
10.1.1.4/30	10.1.1.5
10.1.1.6/32	10.1.1.6
10.1.1.8/30	10.1.1.6
10.2.0.0/24	10.1.1.6

 Page8 HUAWEI

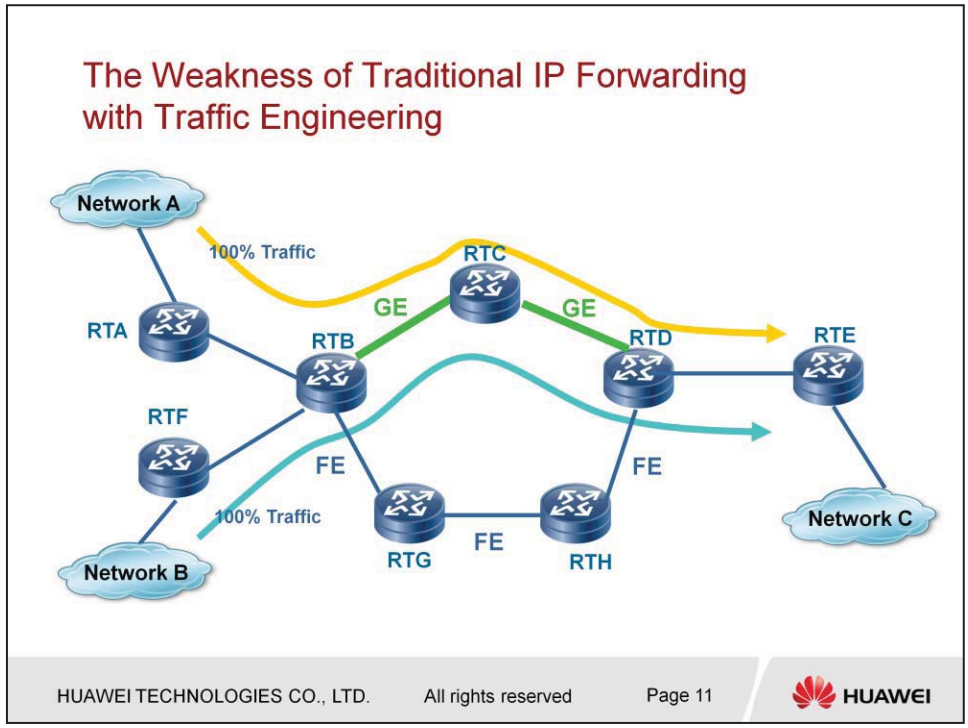
Traditional IP Forwarding(SWC)

Network	Nexthop
10.1.0.0/24	10.1.1.5
10.1.1.0/30	10.1.1.5
10.1.1.4/30	10.1.1.6
10.1.1.5/32	10.1.1.5
10.1.1.8/30	10.1.1.9
10.1.1.10/32	10.1.1.10
10.2.0.0/24	10.1.1.10

Traditional IP Forwarding(SWD)

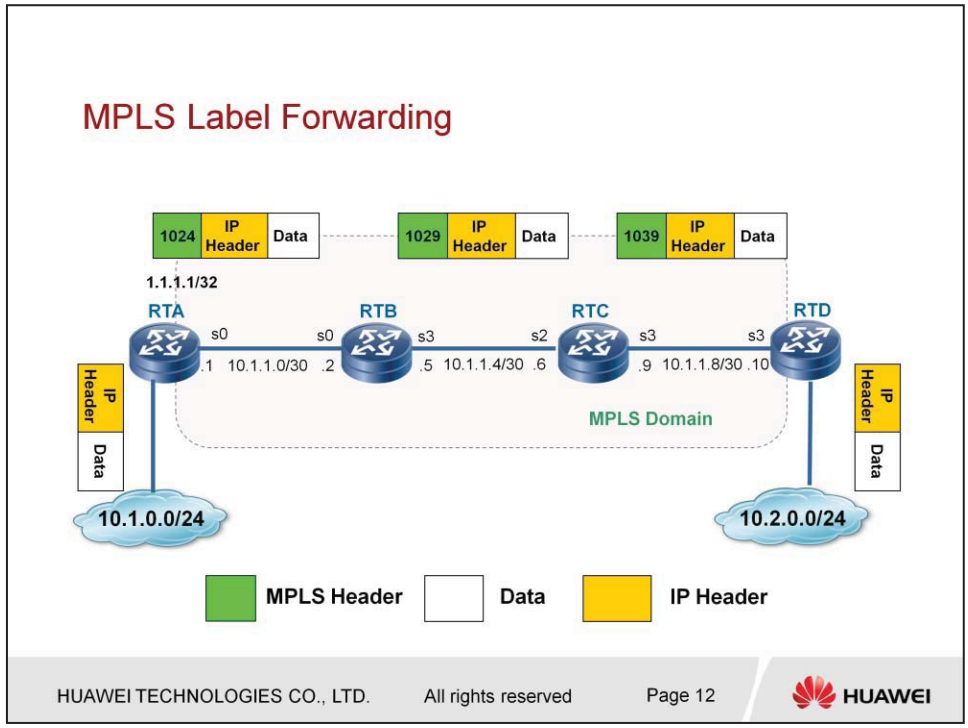


Network	Nexthop
10.1.0.0/24	10.1.1.9
10.1.1.0/30	10.1.1.9
10.1.1.4/30	10.1.1.9
10.1.1.8/30	10.1.1.10
10.1.1.9/32	10.1.1.9
10.2.0.0/24	10.2.0.2
10.2.0.1/32	10.2.0.1

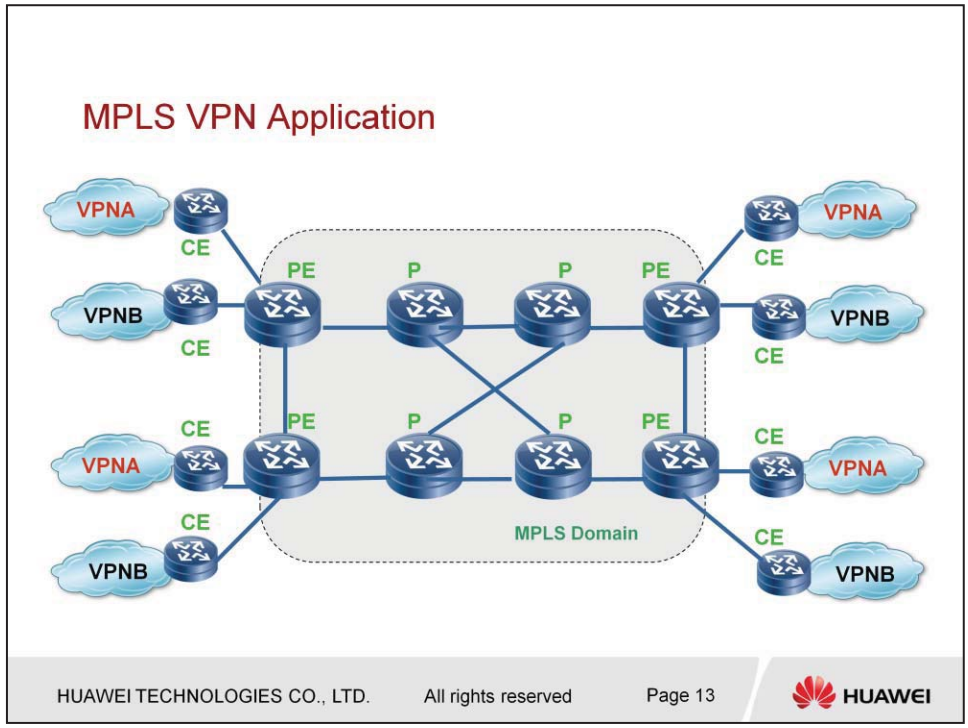


Traditional IP network calculates the optimal route based on IGP Metric, we should also consider bandwidth, link attributes and other factors; Traffic engineering based on IP is IGP destination-based forwarding, it uses hop-by-hop forwarding, can't control traffic forwarding by source. Additionally traffic engineering based on IP is connectionless, and can't implement explicit routing.

In the figure, there are two paths between RTB and RTD. IGP selects optimal route according to Metric and forwards all the IP packets from Network A and Network B to Network C in traditional IP forwarding, RTB-RTG-RTH-RTD link is idle. When the traffic is too large, it may cause congestion on the optimal path, while suboptimal path is idle and underutilized.



MPLS is a kind of label forwarding technology, it adopts connectionless control plane and connection oriented data plane, connectionless control plane implements routing transmission and label distribution, connection oriented data plane implements packet transmission along LSP (label switch path) established before. In an MPLS network domain, a router does not need to analyze every packet's destination IP address, but just forwards by label that are added before the IP header (as the figure shows that RTB receives labeled packet from RTA, then forwards by label, RTC is similar). Comparing to traditional IP forwarding, MPLS label forwarding greatly improves forwarding efficiency.



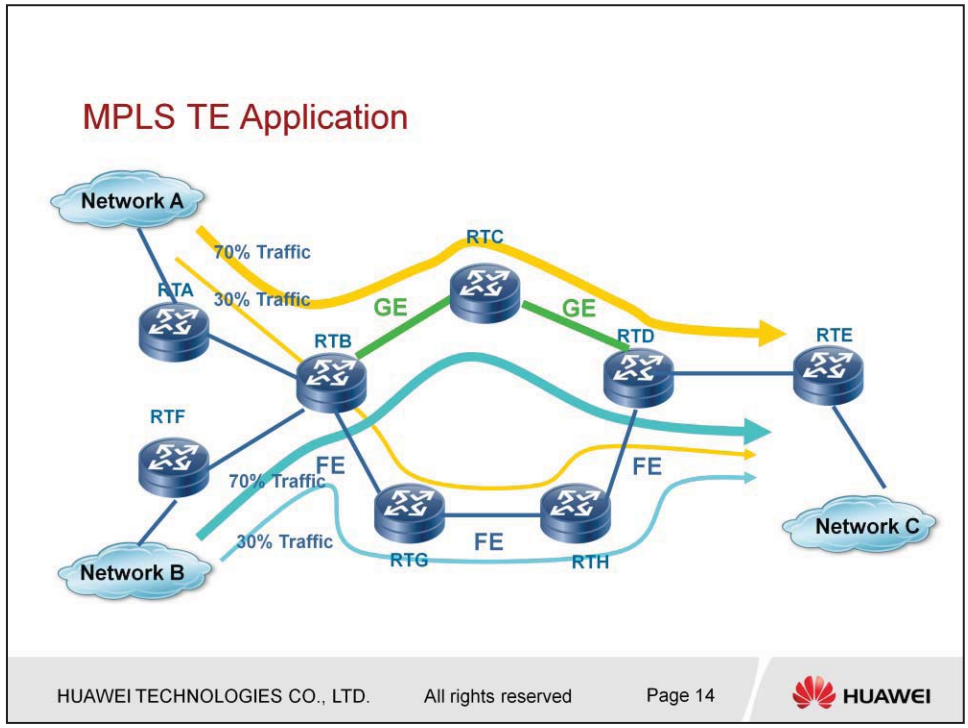
However, with the development of ASIC technology, routing lookup speed does not generate a bottleneck on the network any more. Improving forwarding speed is no longer the obvious advantage of MPLS.

MPLS integrates the advantage of the two forwarding technologies, powerful layer 3 routing function of IP network and high efficiency forwarding mechanism of traditional layer 2 network, its forwarding plane adopts connection oriented, it is

very similar to layer 2 network forwarding method in existence. It makes MPLS easy to implement seamless combination of IP and ATM, frame relay and other layer 2 network, and provide better solution for TE (Traffic Engineering), VPN (Virtual Private Network), QoS (Quality of Service) and other applications.

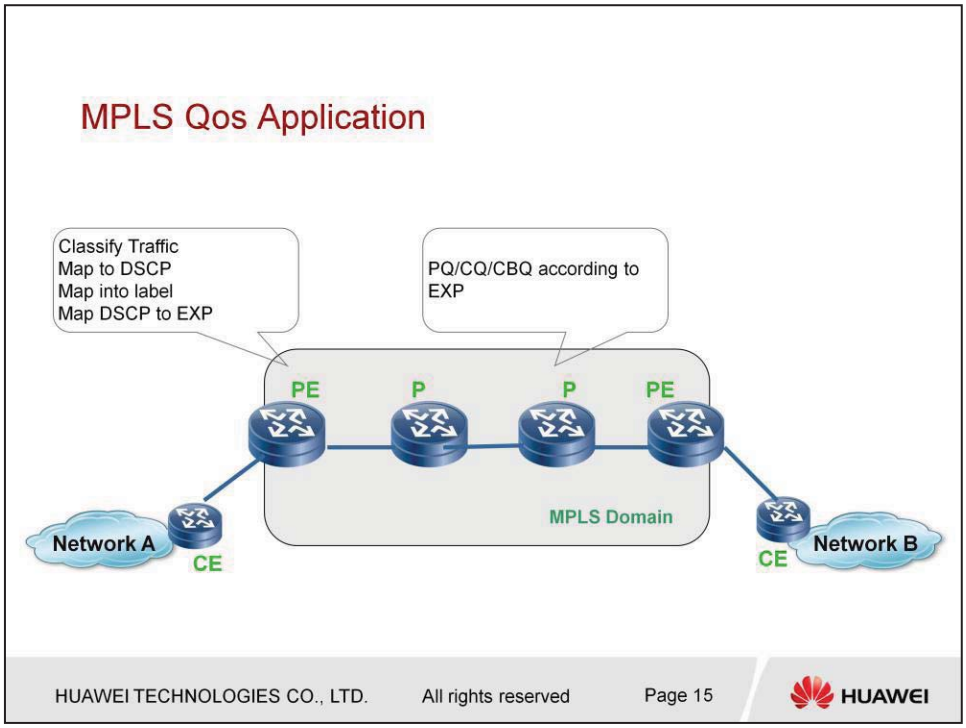
VPN based on MPLS can combine different embranchment of private network, form a uniform network, VPN based on MPLS also supports communication control between different VPN. As the figure shows, CE is user edge device; PE is service

provider edge router, which is located in backbone network. P is backbone router in the service provider network, it does not directly connect with CE. VPN data is transmitted along LSP (label switch path) encapsulated with MPLS label.



MPLS TE integrates MPLS technology and TE, reserves resource via establishing LSP tunneling towards appointed path, makes traffic steer clear of congestion node, reaches the objective that balance network traffic. As shown in the figure, 70% traffic from Network A to Network B is transmitted via the path of RTB-RTC-RTD, 30% of traffic is transmitted via the path of RTB-RTG-RTH-RTD.

The traffic from Network B to Network C is similar.



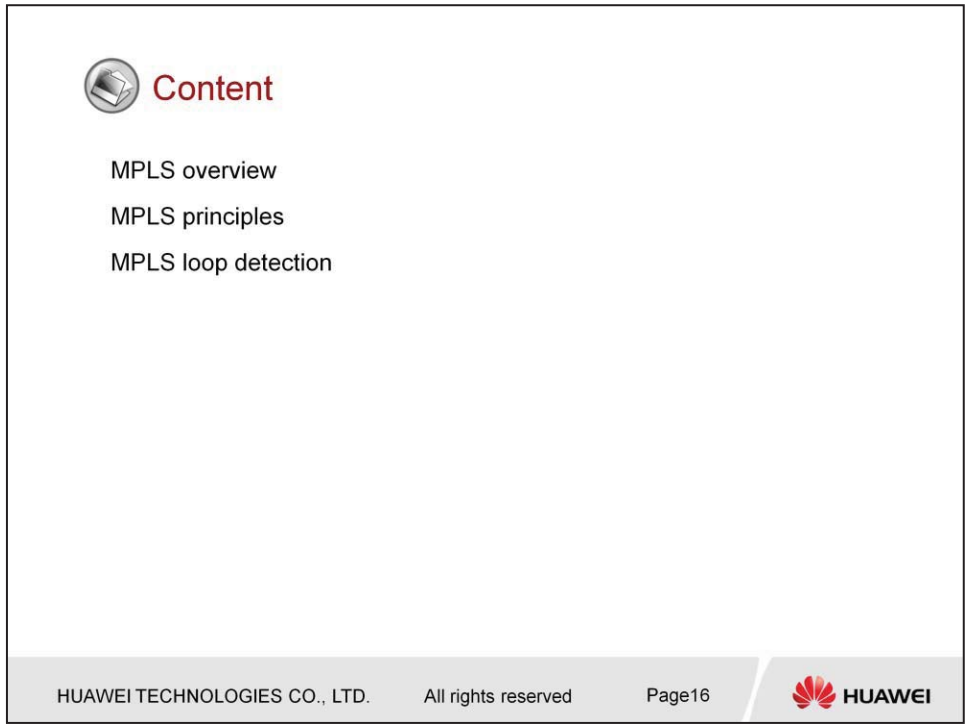
MPLS and DifferServ cooperate perfectly to provide Qos function.

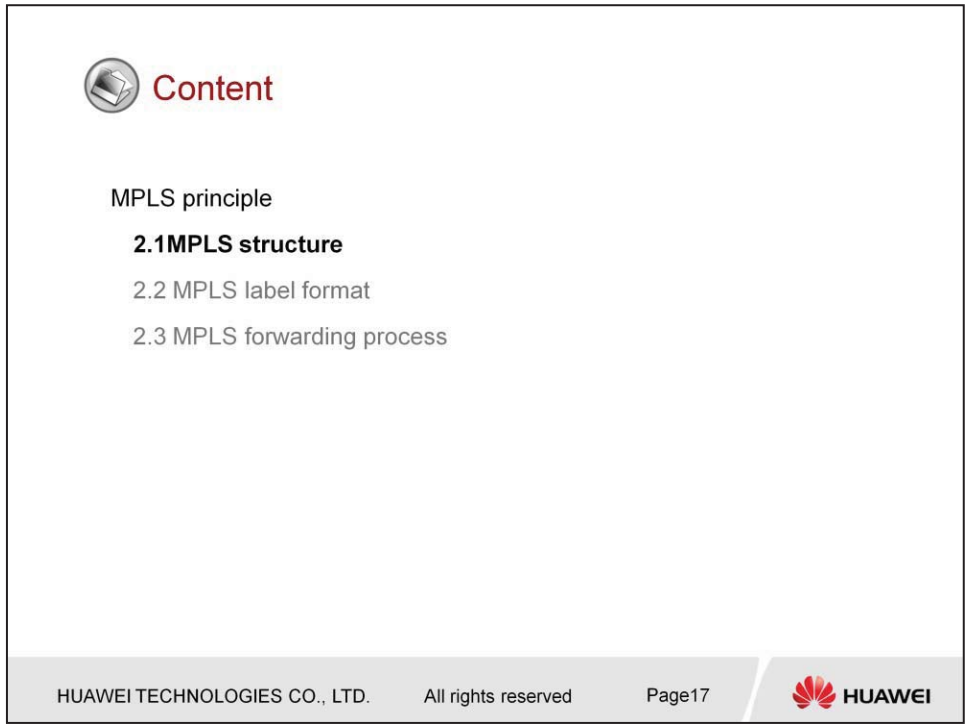
According to the requirement, classify the data flow on CE or PE, for example, classify the flow whose DSCP value is 2 as one type, and classify the flow whose DSCP value is 3 as another type, then the classified flow can be implemented by

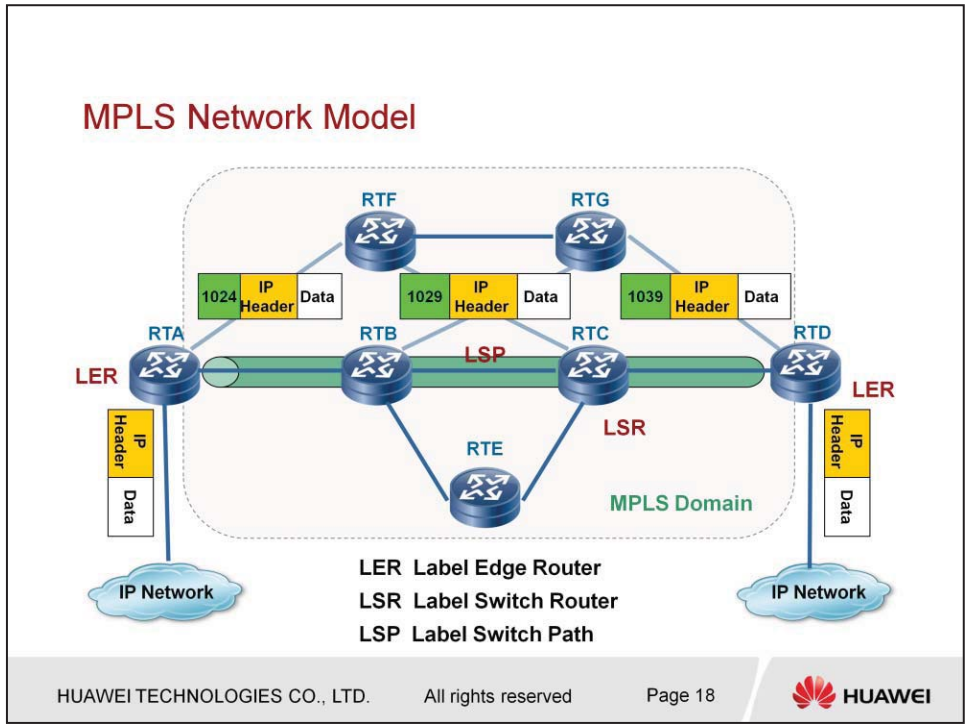
traffic policing, EXP remarked and so on.

When PE inserts Label into the packet, IP preference carried by IP packet will be mapped to EXP field of the label, thus the information of type of service carried by original IP will be carried by label now.

Between the P router and PE router, according to EXP field of label, implement different queue scheduling (such as PQ, CQ, CBQ), namely, transmit labeled flow along the same label switch path with different QoS.


Content
MPLS overview
MPLS principles
MPLS loop detection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page16
HUAWEI


Content
MPLS principle
2.1MPLS structure
2.2 MPLS label format
2.3 MPLS forwarding process
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page17
HUAWEI



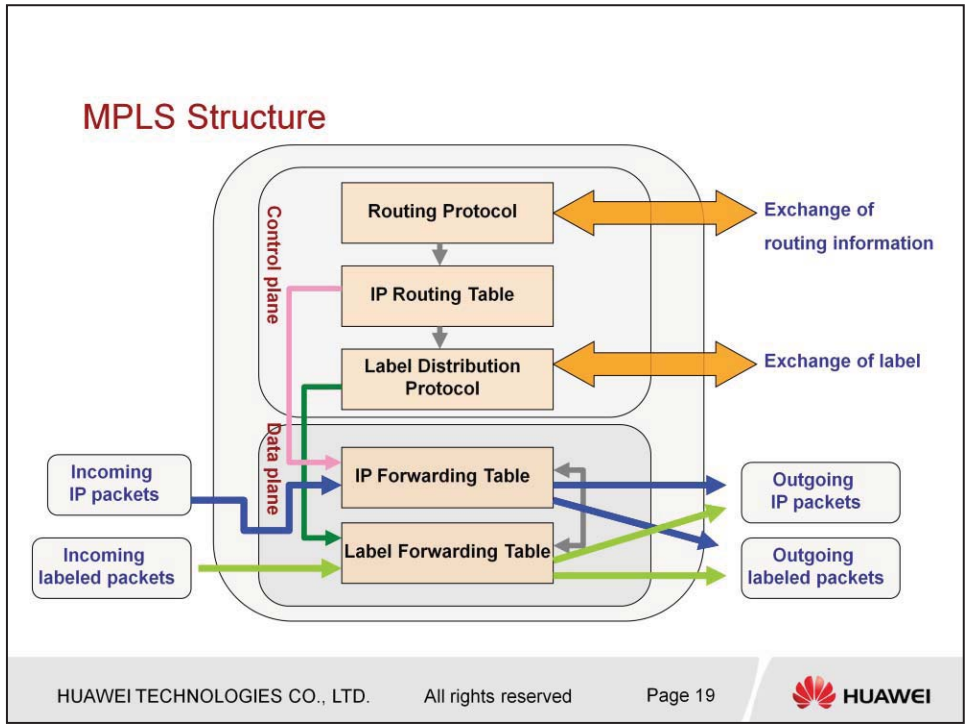
The typical structure of MPLS network is shown in this slide: the router and ATM switch located inside of MPLS domain are called LSR, router and ATM switch located at the edge of MPLS domain that used to connect IP network or other kinds of network are called LER.

In IP network, it implements traditional IP forwarding; in MPLS domain, it implements label forwarding.

Both of LER and LSR have the ability of label forwarding, but they are located in different position, the packet processing is different. LER's responsibility is to receive IP packet from IP network and insert label into the packet, then transmit it to LSR,

whereas, its responsibility is also to receive labeled packet from LSR and remove label, transmit it to IP network; LSR's responsibility is to forward according to the label.

The path that packet passes through in MPLS domain is called Label Switch Path (LSP), this path is already confirmed and established by kinds of protocols before packet forwarding, packet will be transmitted along the specified LSP.



MPLS network forwards packet according to the label. But how is the label generated? What mechanism does MPLS adopt to implement data forwarding?

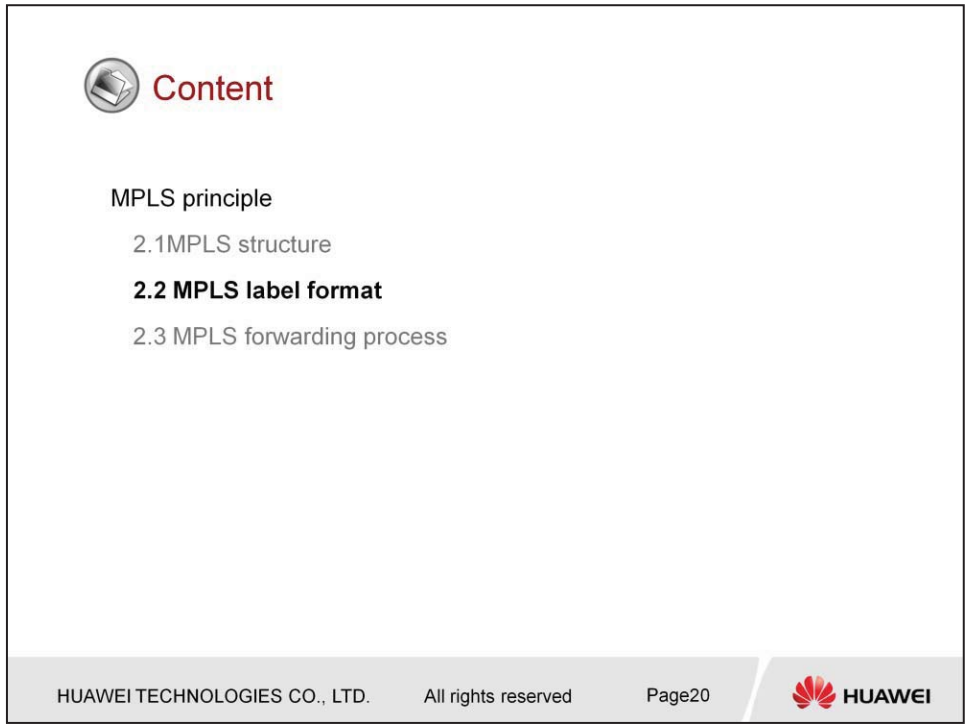
MPLS includes two plane: control plane and data plane.

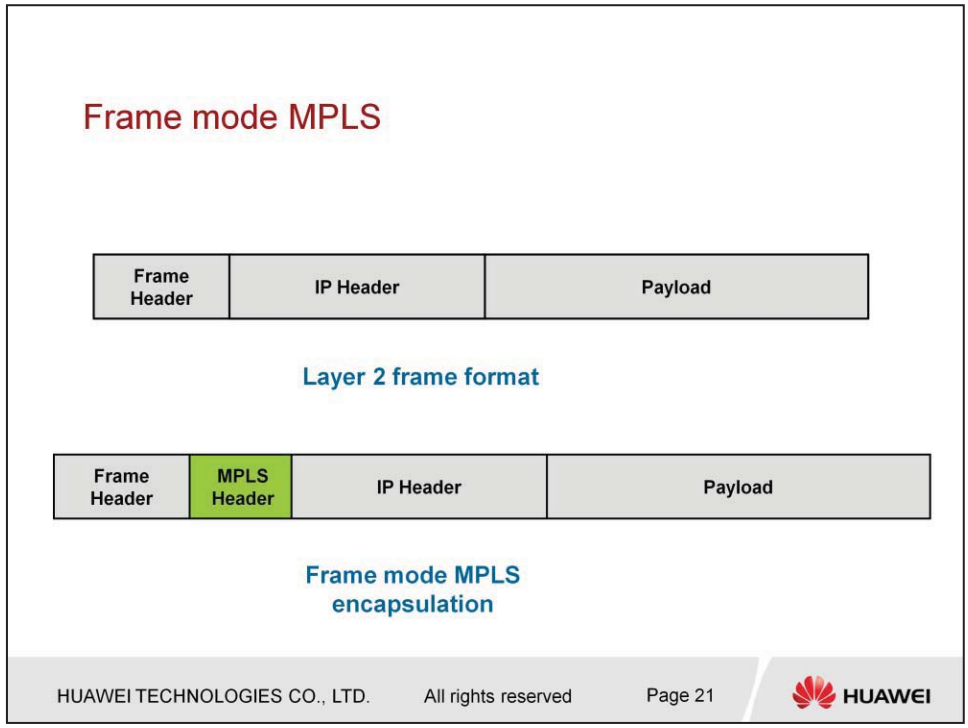
Control plane's responsibility is to generate and maintain routing information and label information. Data plane's charge is conventional IP packet forwarding and labeled packet forwarding.

In control plane, routing protocol module is used to transmit routing information, generate routing table; label distribution protocol is used to complete exchange of label and establish label switch path.

Data plane includes IP forwarding table and label forwarding table, when receiving conventional IP packets, if it is conventional IP forwarding, it should lookup routing table and forward, if it is label forwarding, it should forward by the

label forwarding table; when receiving labeled packets, if it needs to forward by label, it should forward by label forwarding table, if it needs to transmit to IP network, it should remove the label and forward using the IP routing table.

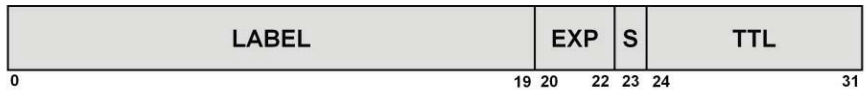

Content
MPLS principle
2.1MPLS structure
2.2 MPLS label format
2.3 MPLS forwarding process
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page20
HUAWEI



There are two MPLS encapsulation modes: frame mode and cell mode.

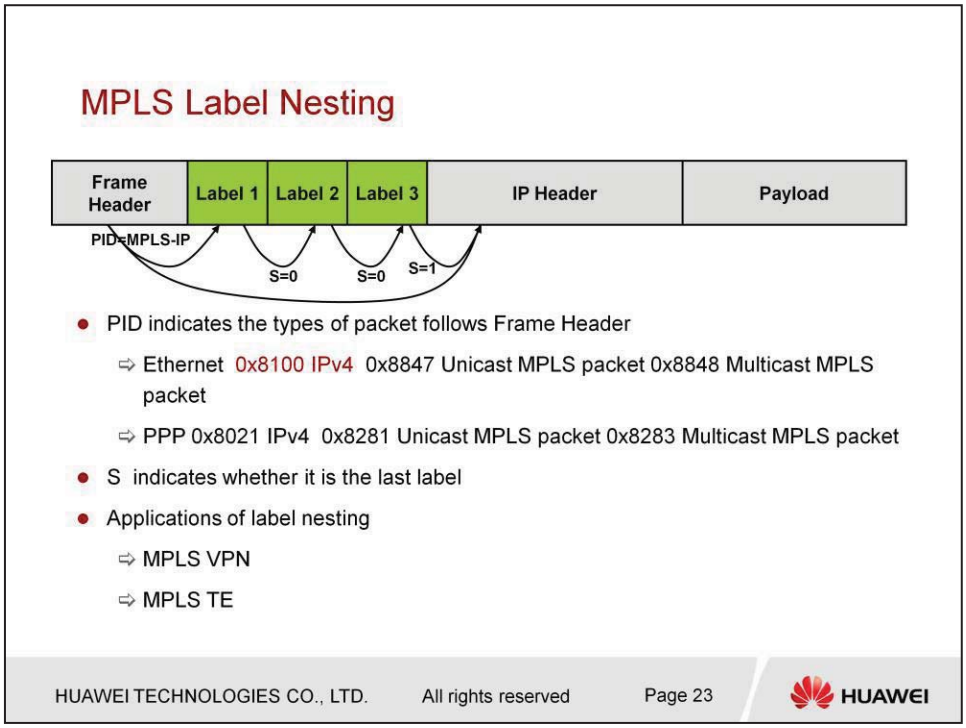
Frame encapsulation mode inserts a MPLS label header between layer 2 header and layer 3 header. Ethernet and PPP protocol adopt this mode.

MPLS Header



- The total length of MPLS header is 4bytes (32bits)
- The length of Label field is 20bits
- The length of EXP（Experimental Use）field is 3bits
- The length of S（Bottom of Stack）field is1bit
- The length of TTL field is 8bits

The length of MPLS header is 32 bits, it includes 20-bit label field, this field is used for data forwarding; 3-bit EXP is used to carry precedence of IP packet; 1-bit S is at the bottom of stack and is used to indicate whether it is the last label (MPLS label can support label nesting); 8-bit TTL, its function is similar to TTL of IP header, it is used to prevent data from looping.



The protocol filed PID in layer 2 header specifies that payload starts with packet with label encapsulated or IP header. For example, in Ethernet protocol, PID=0x8847 identifies that the frame payload is a multicast MPLS packet.

PID=0x8848 identifies that the frame payload is a unicast MPLS packet.

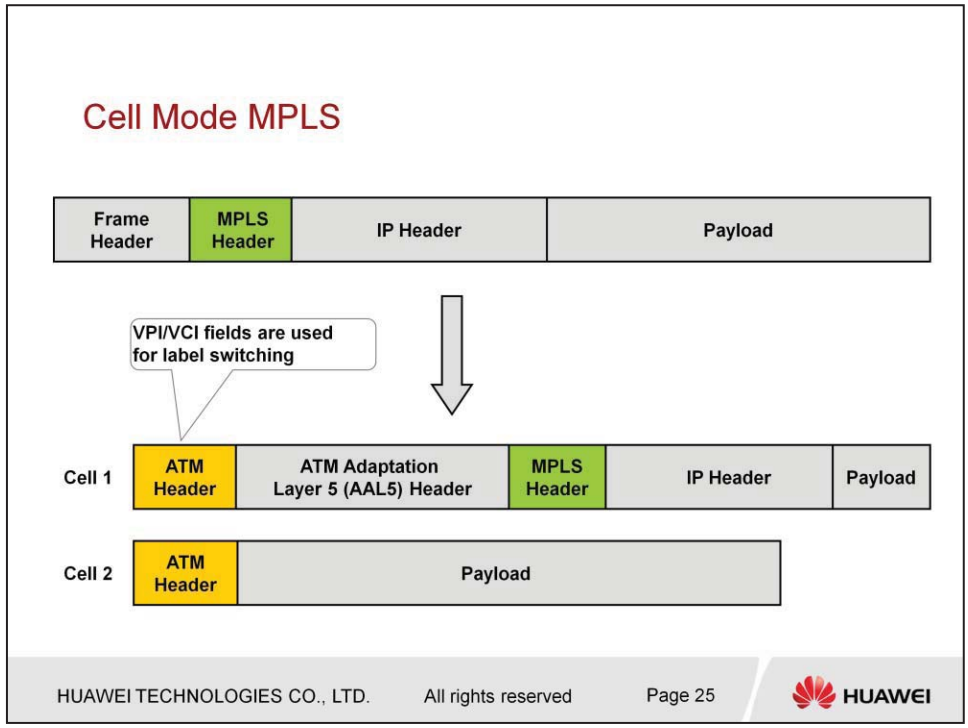
PID=0x0800 identifies that the frame payload is a unicast IP packet. In PPP protocol, PID=0x8281 identifies that the frame payload is a unicast MPLS packet.

PID=0x8283 identifies that the frame payload is a multicast MPLS packet. S bit in MPLS header indicates whether the next header is another label or a layer 3 IP header.

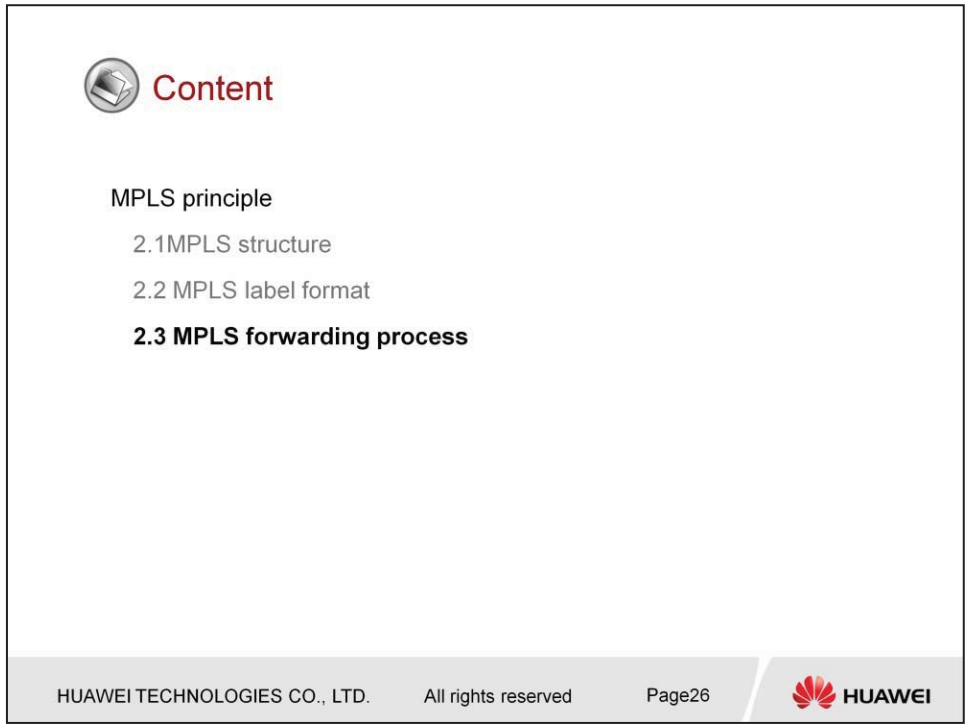
Usually MPLS only allocates one label for a packet. But some advanced applications of MPLS use multiple labels. For example, MPLS VPN will use 2 layers of labels (in complex situation, it even uses 3 layers of labels), out-label is used for public network forwarding, in-label is used to indicate that which VPN the packet belongs to; MPLS TE also uses two or more labels, the outmost label is used to indicate TE tunneling, in-label indicates the destination of packet.

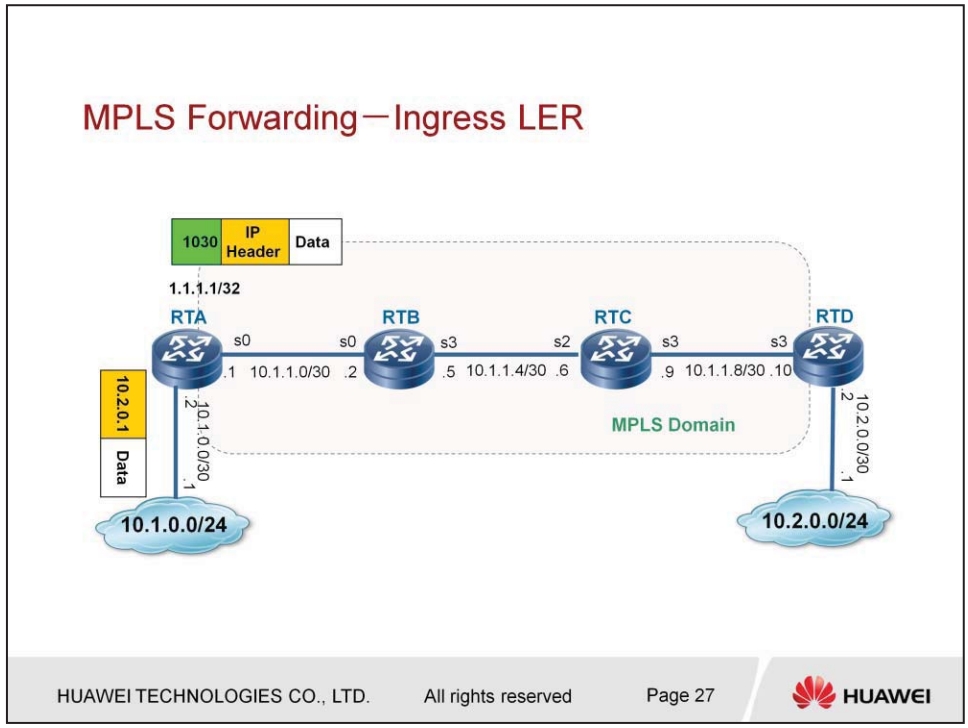
Note:

The Label1, Label2, Label3 all mean 4 Bytes MPLS header in last slide, it includes 20-bit label information.

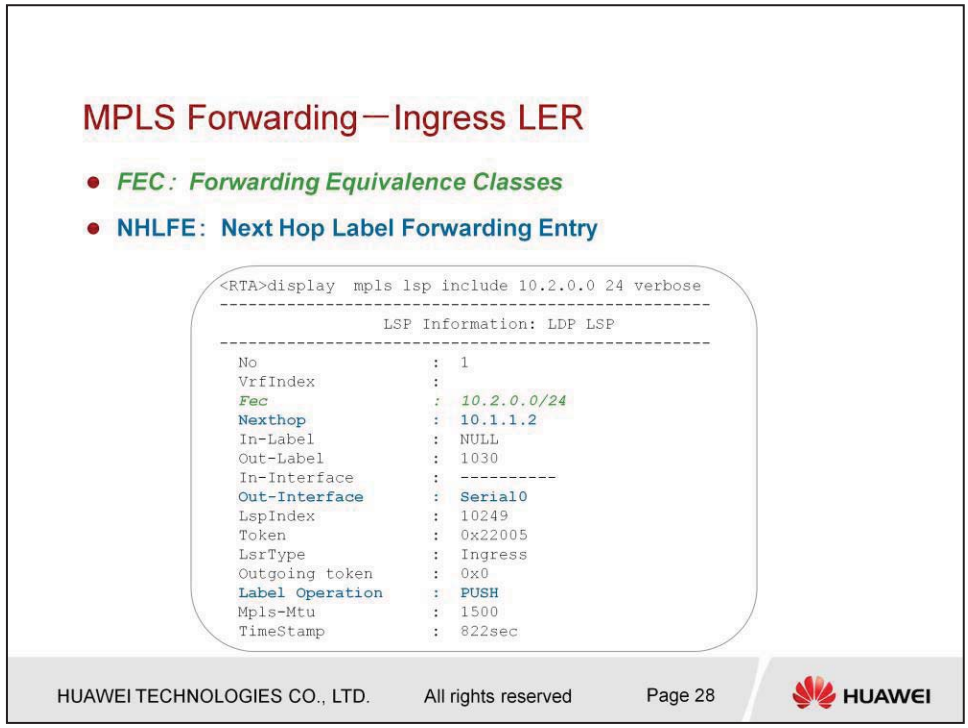


ATM adopts cell mode MPLS encapsulation, VPI/VCI fields in ATM cell header are used for label switching. If there is a MPLS Header in the packet, this MPLS Header will be reserved, but not used for forwarding, and only the first cell reserves MPLS Header.


Content
MPLS principle
2.1MPLS structure
2.2 MPLS label format
2.3 MPLS forwarding process
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page26
HUAWEI



When a IP packet enters MPLS domain, ingress LER (RTA) will analyze the packet, determine which label to encapsulate packet according to packet characteristic (generally by prefix analysis of destination IP address), and determine to transmit to which next hop from which interface.

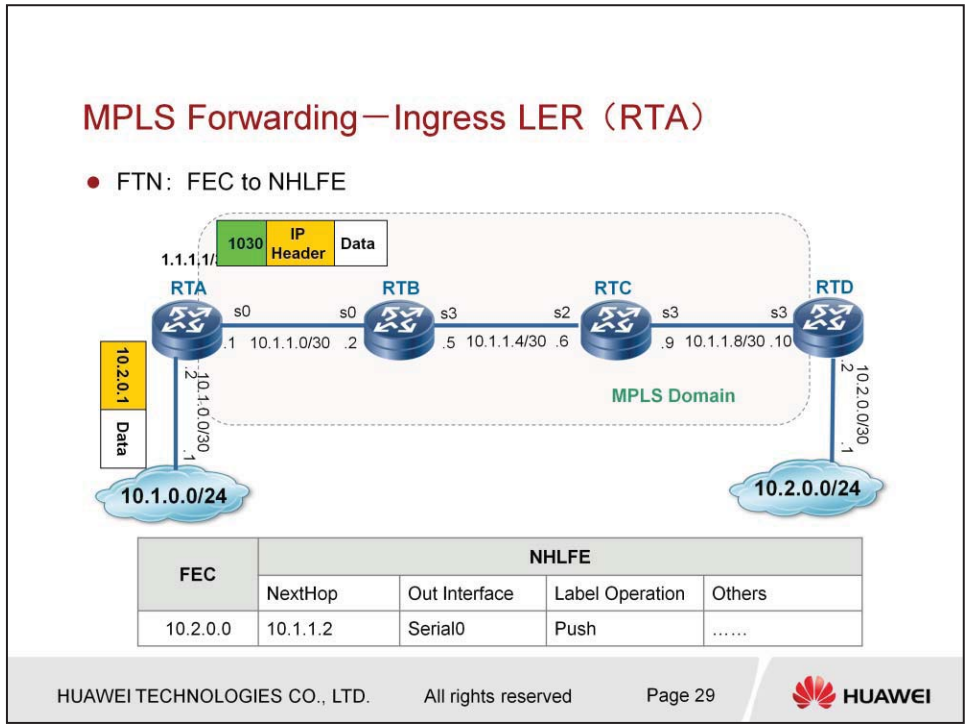


FEC (Forwarding equivalence class) means a group of IP packets which are forwarded in the equipollence method, for example, a group of IP packets with same destination IP prefix will be allocated a unique label. In this case, the packet that destination IP prefix is 10.2.0.0/24 belongs to a FEC, the label allocated for this FEC is 1030.

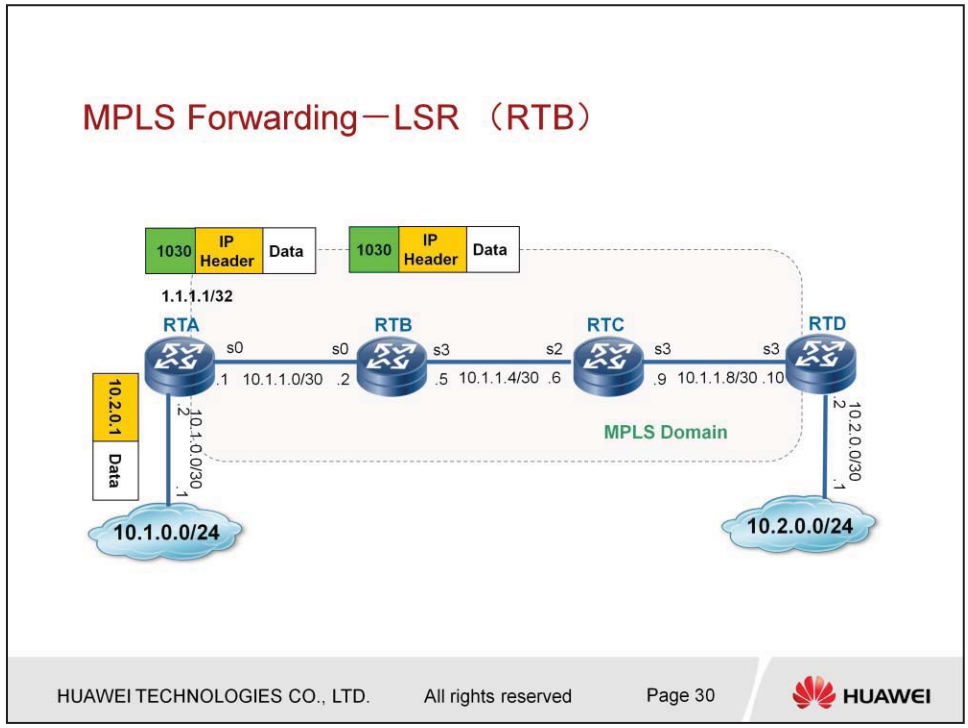
NHLFE is used when forwarding a labeled packet, It contains the following information:

1. the packet's next hop;

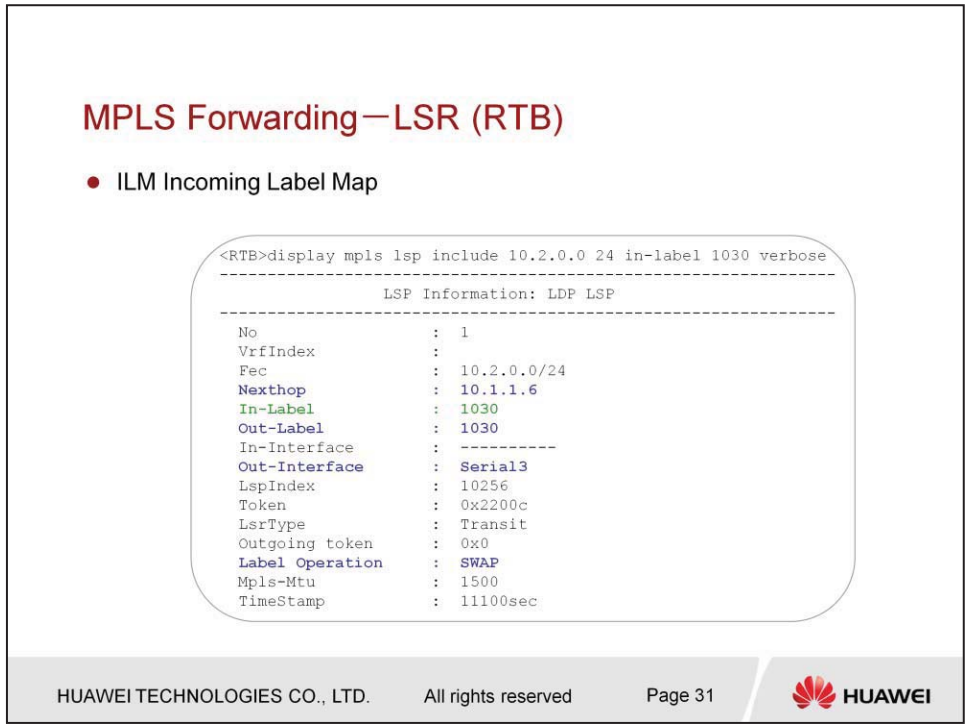
2.the operation to perform on the packet's label stack (it contains pushing new label, popping label, replacing the original label with new label). It may also contain other information, such as the data link encapsulation to use when transmitting the packet. In this case, next hop is 10.1.1.2, label operation is “push”.



FEC represents the same kind of packets, NHLFE contains next hop, label operation and other information. Only associating FEC with NHLFE, it can implement particular label forwarding for the same kinds of packets, FTN can implement this function, FTN (FEC-to-NHLFE) indicates the mapping for an FEC to NHLFE, if there are multiple cost-equal paths, one FEC maybe map to multiple NHLFE.



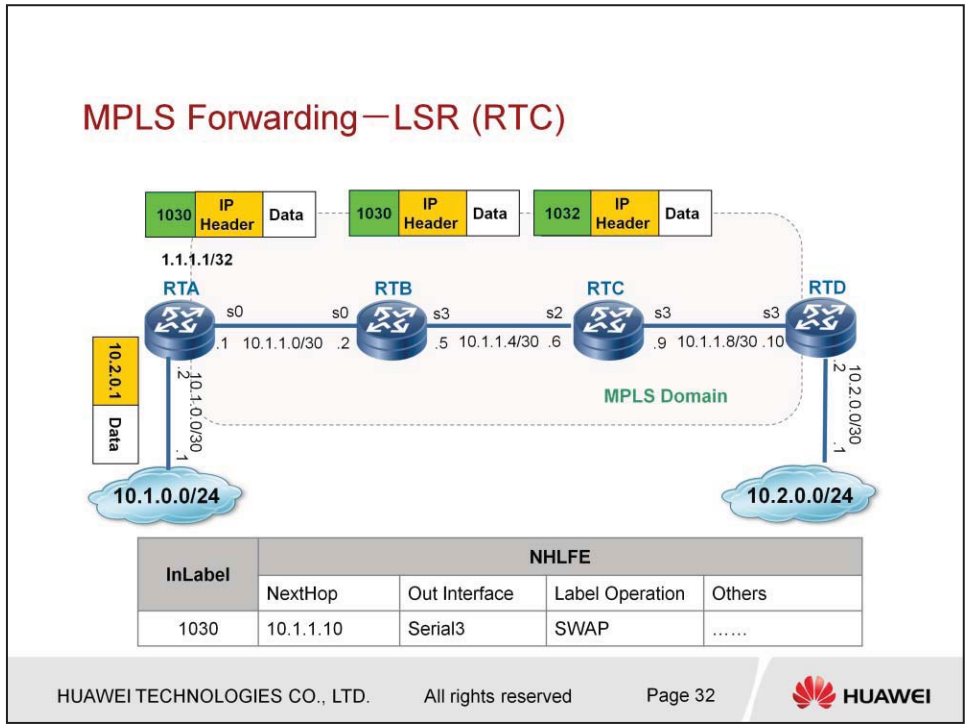
RTB receives message with MPLS label 1030 from RTA, and forwards it according to the MPLS label.



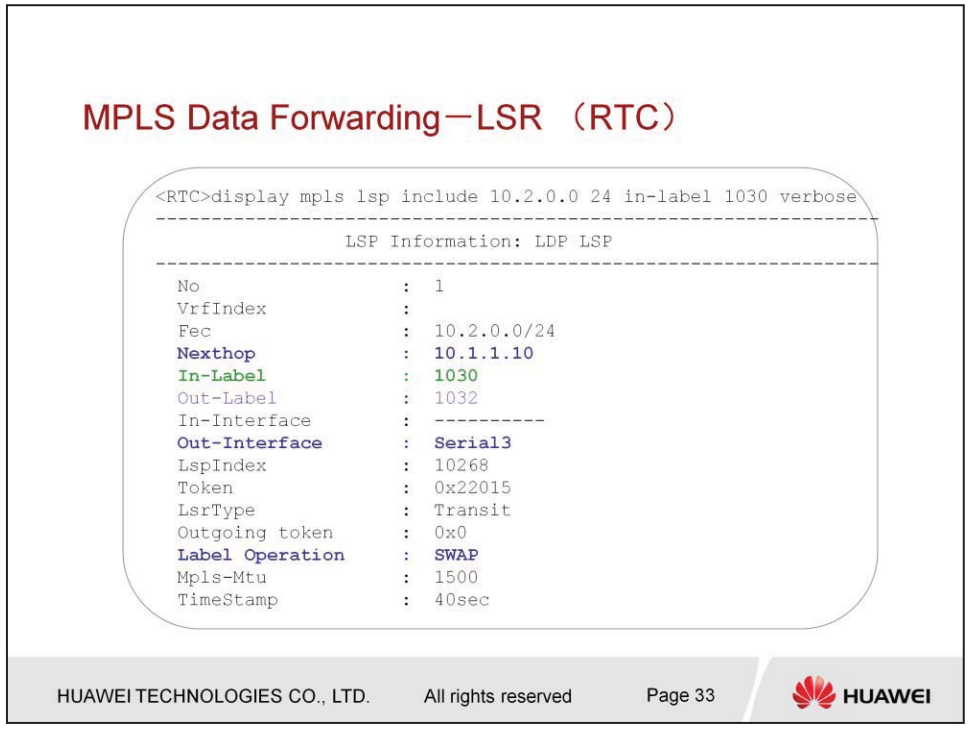
In the case, after RTB receives data packets with label 1030, it will forward it by label; first , it will find the next hop 10.1.1.6, use outgoing label to swap incoming label, and then continue forwarding. (this case is special, outgoing label and

incoming label are the same.)

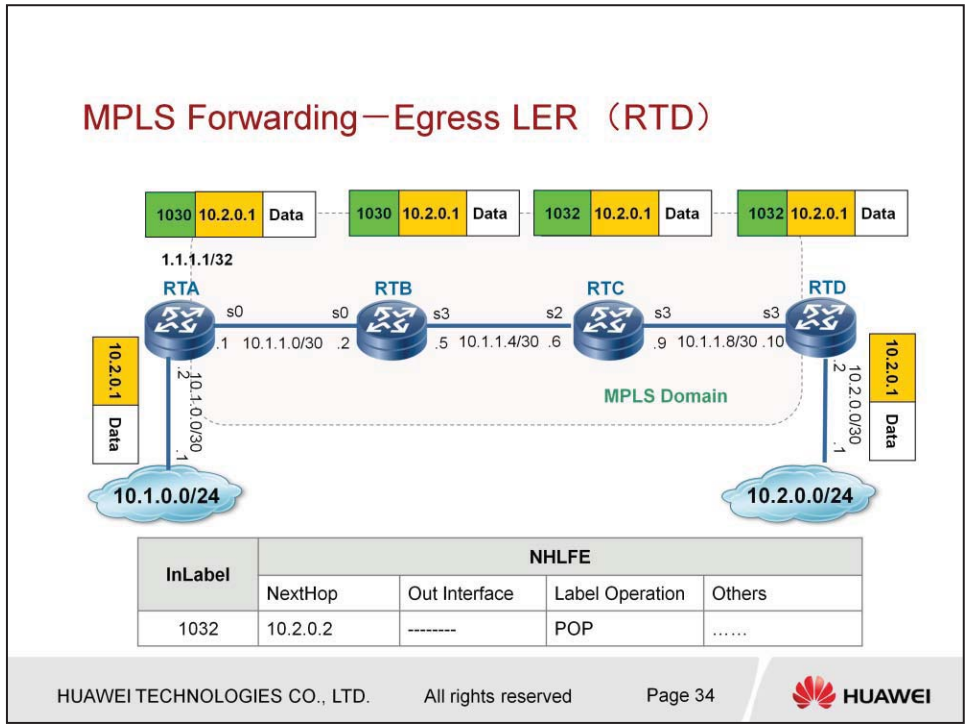
ILM maps each incoming label to a set of NHLFE, It is used when forwarding packets that arrive as labeled packets. If there are multiple equal-cost paths, one incoming label maps to multiple NHLFE.



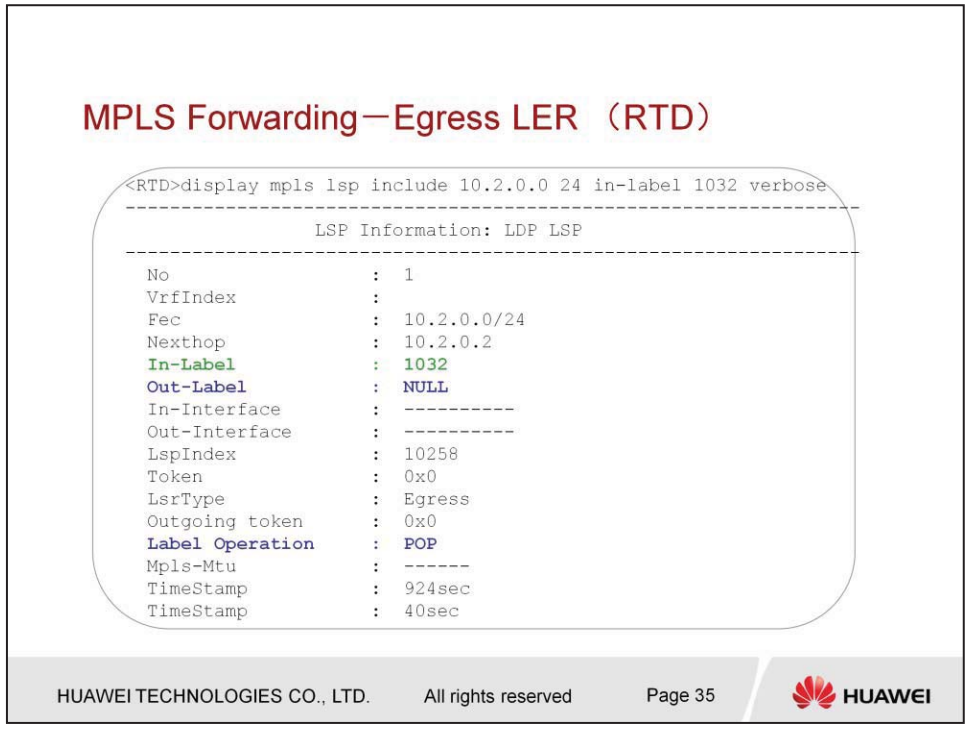
Similar to RTB, when RTC receives message with label 1030, it forwards packets by label, and uses a new outgoing label to swap original label.



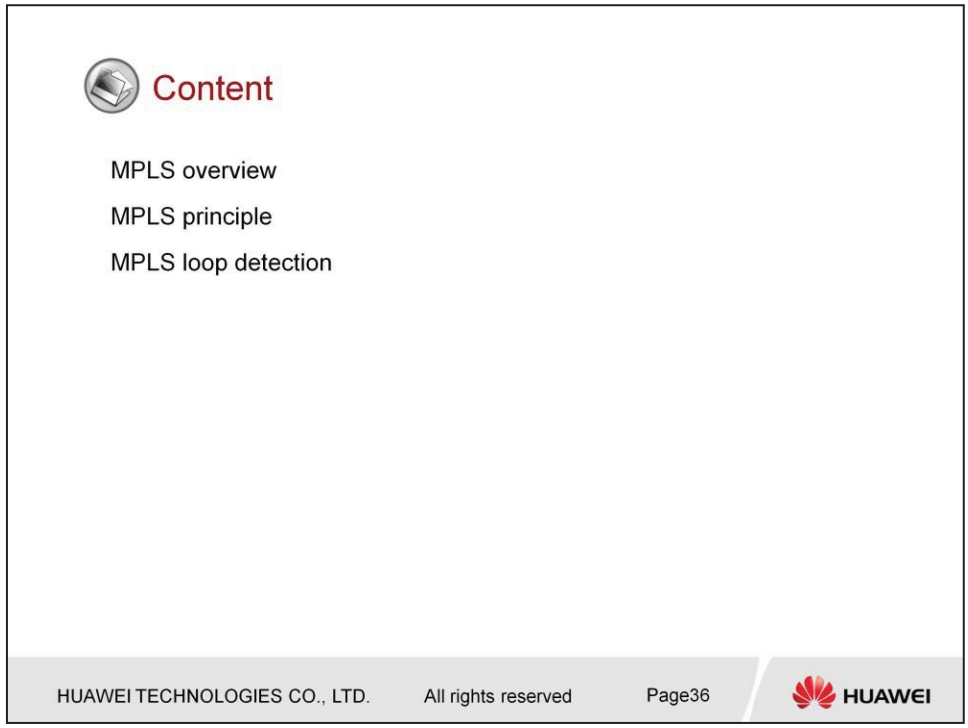
In this case, RTC uses outgoing label 1032 to swap incoming label, then transmits packet from outgoing interface Serial3, the next hop is 10.1.1.10.

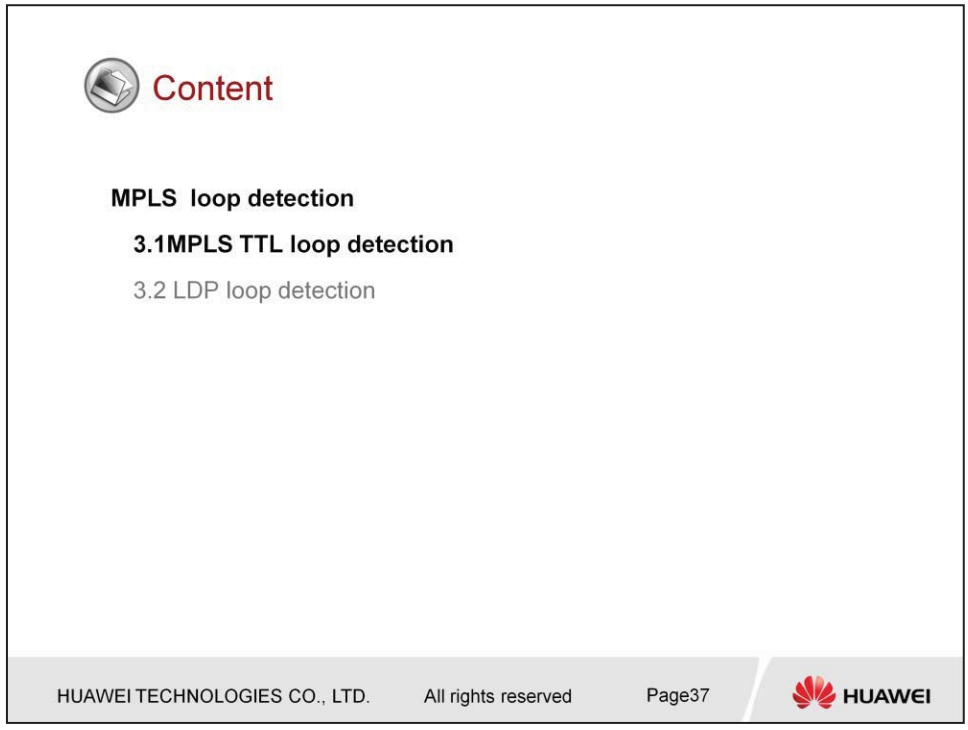


Egress LSR RTD receives message with label 1032, Pops the label, lookups IP routing table and forwards it.



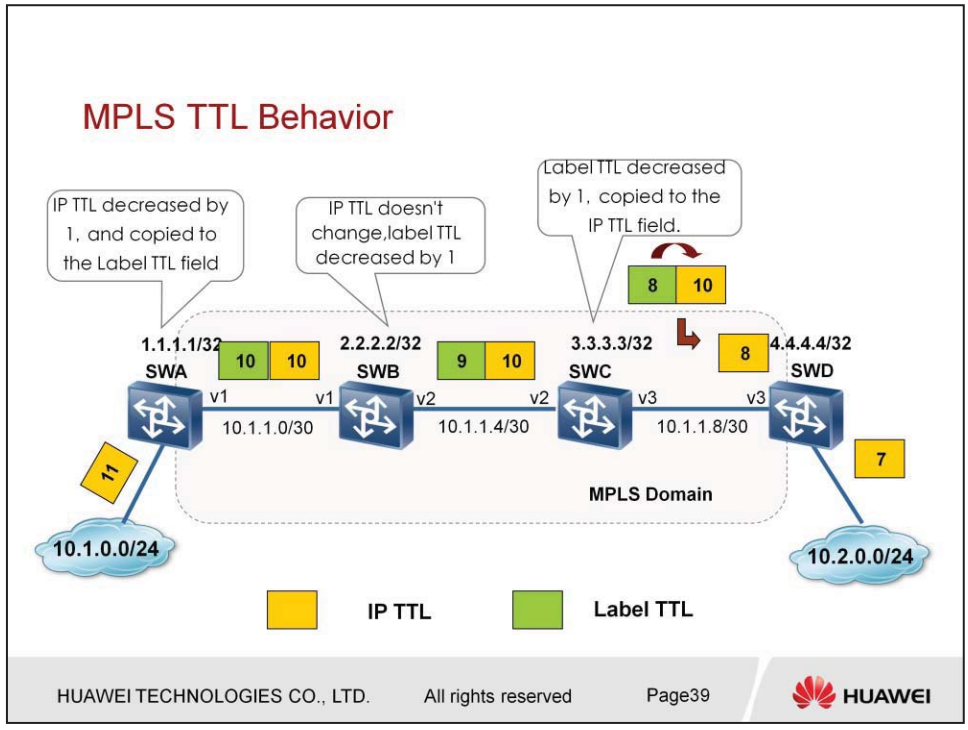
In the case, RTD pops label 1032 and forwards the message to the next hop 10.2.0.2.


Content
MPLS overview
MPLS principle
MPLS loop detection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page36
HUAWEI

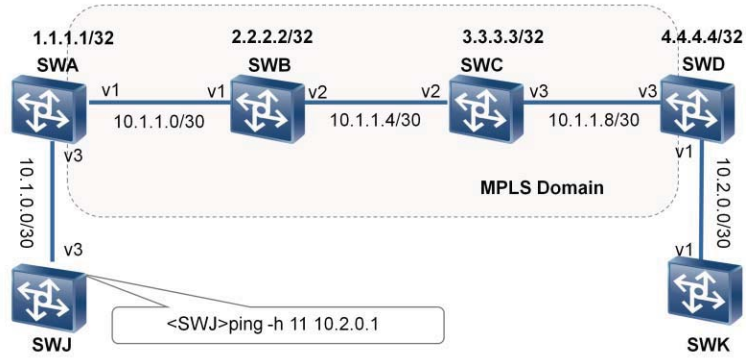

Content
MPLS loop detection
3.1MPLS TTL loop detection
3.2 LDP loop detection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page37
HUAWEI

MPLS Loop Detection

- IGP Loop Detection
- TTL Loop Detection
- LDP loop detection


MPLS TTL Behavior
IP TTL decreased by 1, and copied to the Label TTL field
IP TTL doesn't change,label TTL decreased by 1
Label TTL decreased by 1, copied to the IP TTL field.
8
10
1.1.1.1/32
SWA
10
10
2.2.2.2/32
SWB
9
10
3.3.3.3/32
SWC
8
4.4.4.4/32
SWD
v1
v1
v2
v2
v3
v3
10.1.1.0/30
10.1.1.4/30
10.1.1.8/30
11
7
MPLS Domain
10.1.0.0/24
10.2.0.0/24
IP TTL
Label TTL
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page39
HUAWEI

MPLS TTL Configuration



 Page40 HUAWEI

MPLS TTL Configuration

```
<SWA>debug mpls packet
<SWA>debug ip packet acl 3000
<SWA>terminal monitor
<SWA>terminal debugging

*0.86297391 SWA IP/8/debug_case:
Receiving, interface = Serial3, version = 4, headlen = 20, tos = 0,
pktlen = 84, pktid = 2273, offset = 0, ttl = 11, protocol = 1,
checksum = 37572, s = 10.1.0.1, d = 10.2.0.1
prompt: Receiving IP packet from Serial3

*0.86297391 SWA IP/8/debug_case:
Sending, interface = Serial3, version = 4, headlen = 20, tos = 0,
pktlen = 84, pktid = 2273, offset = 0, ttl = 10, protocol = 1,
checksum = 37572, s = 10.1.0.1, d = 10.2.0.1
prompt: Sending the packet by lsp

*0.86297391 SWA MFW/8/MPLSFW PACKET:
PUSH Label=1030, EXP=0, TTL=10
Sending to V1, PktLen=88, Label(s)=1030, EXP=0, TTL=10
```


MPLS TTL Configuration

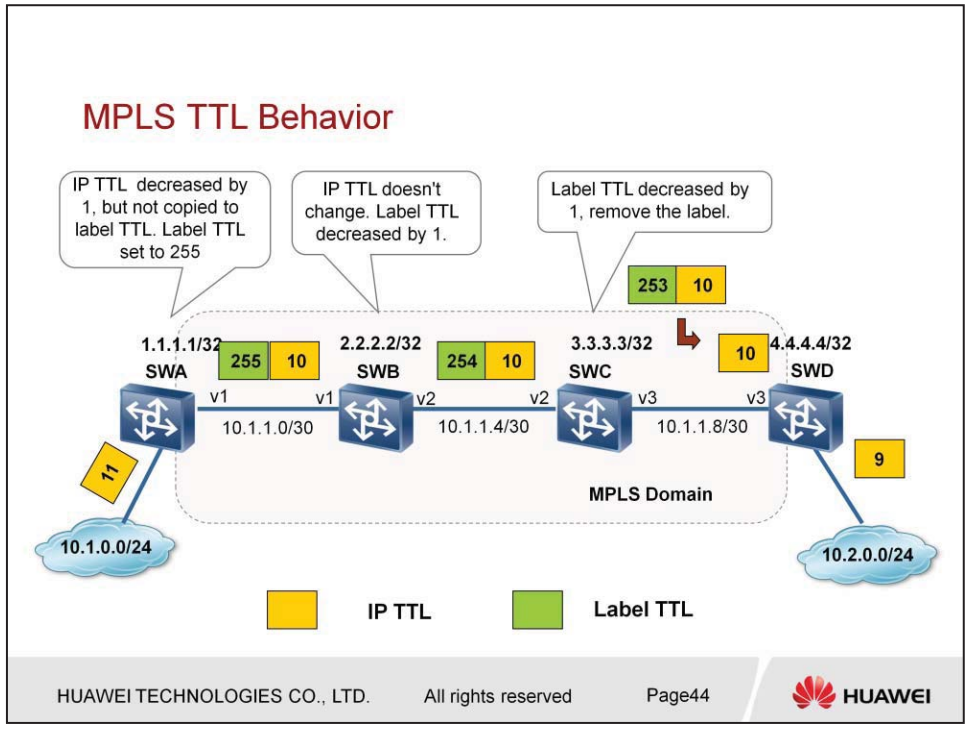
```
<SWB>debug mpls packet
<SWB>debug ip packet acl 3000
<SWB>terminal monitor
<SWB>terminal debugging

*0.189653734 SWB MFW/8/MPLSFW PACKET:
Receiving from V1, PktLen=88, Label(s)=1030, EXP=0, TTL=10
SWAP Label=1029, EXP=0, TTL=9
Sending to V2, PktLen=88, Label(s)=1029, EXP=0, TTL=9
```

```
<SWC>debug mpls packet
<SWC>debug ip packet acl 3000
<SWC>terminal monitor
<SWC>terminal debugging

*0.189533719 SWC MFW/8/MPLSFW PACKET:
Receiving from V2, PktLen=88, Label(s)=1029, EXP=0, TTL=9
SWAP Label=3, TTL=8
Sending to V3, Dest=10.2.0.1, Nexthop=10.1.1.10
```

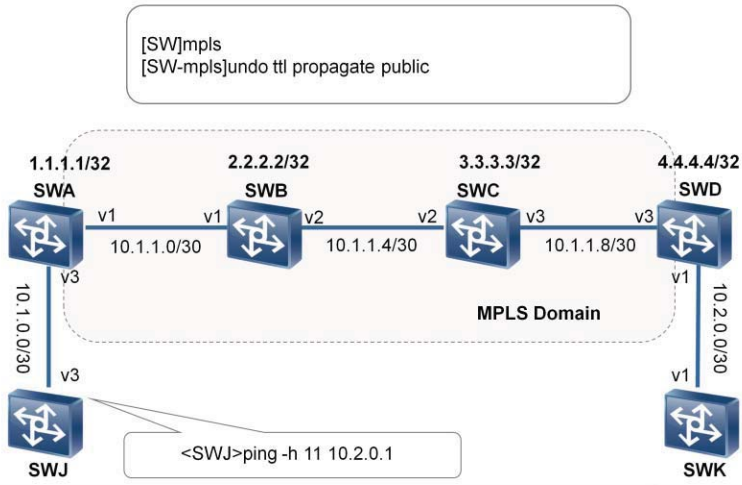

MPLS TTL Configuration

```
<SWD>debug mpls packet
<SWD>debug ip packet acl 3000
<SWD>terminal monitor
<SWD>terminal debugging
*0.64991297 SWD IP/8/debug_case:
Receiving, interface = Serial3, version = 4, headlen = 20, tos = 0,
pktlen = 84, pktid = 2273, offset = 0, ttl = 8, protocol = 1,
checksum = 38340, s = 10.1.0.1, d = 10.2.0.1
prompt: Receiving IP packet from Serial3
*0.64991297 SWD IP/8/debug_case:
Sending, interface = Serial1, version = 4, headlen = 20, tos = 0,
pktlen = 84, pktid = 2273, offset = 0, ttl = 7, protocol = 1,
checksum = 38596, s = 10.1.0.1, d = 10.2.0.1
prompt: Sending the packet from Serial3 at Serial1
```

```
<SWJ>tracert 10.2.0.1
traceroute to  10.2.0.1(10.2.0.1) 30 hops max,40 bytes packet
1 10.1.0.2 31 ms  32 ms  1 ms
2 10.1.1.2 62 ms  94 ms  62 ms
3 10.1.1.6 94 ms  94 ms  94 ms
4 10.1.1.10 125 ms  125 ms  125 ms
5 10.2.0.1 156 ms  156 ms  156 ms
```



MPLS TTL Behavior
IP TTL decreased by 1, but not copied to label TTL. Label TTL set to 255
IP TTL doesn't change. Label TTL decreased by 1.
Label TTL decreased by 1, remove the label.
253
10
1.1.1.1/32
SWA
255
10
2.2.2.2/32
SWB
254
10
3.3.3.3/32
SWC
10
4.4.4.4/32
SWD
v1
v1
v2
v2
v3
v3
10.1.1.0/30
10.1.1.4/30
10.1.1.8/30
11
9
MPLS Domain
10.1.0.0/24
10.2.0.0/24
IP TTL
Label TTL
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page44
HUAWEI

MPLS TTL Configuration



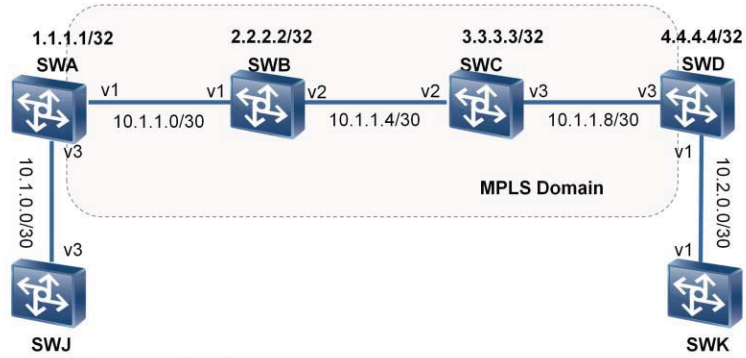
MPLS TTL Configuration

```
<SWA>debug mpls packet
<SWA>debug ip packet acl 3000
<SWA>terminal monitor
<SWA>terminal debugging

*0.81886516 SWA IP/8/debug_case:
Receiving, interface = Serial3, version = 4, headlen = 20, tos = 0,
pktlen = 84, pktid = 1318, offset = 0, ttl = 11, protocol = 1,
checksum = 38527, s = 10.1.0.1, d = 10.2.0.1
prompt: Receiving IP packet from Serial3

*0.81886516 SWA IP/8/debug_case:
Sending, interface = Serial3, version = 4, headlen = 20, tos = 0,
pktlen = 84, pktid = 1318, offset = 0, ttl = 10, protocol = 1,
checksum = 38527, s = 10.1.0.1, d = 10.2.0.1
prompt: Sending the packet by lsp

*0.81886516 SWA MFW/8/MPLSFW PACKET:
PUSH Label=1030, EXP=0, TTL=255
Sending to V1, PktLen=88, Label(s)=1030, EXP=0, TTL=255
```


MPLS TTL Configuration

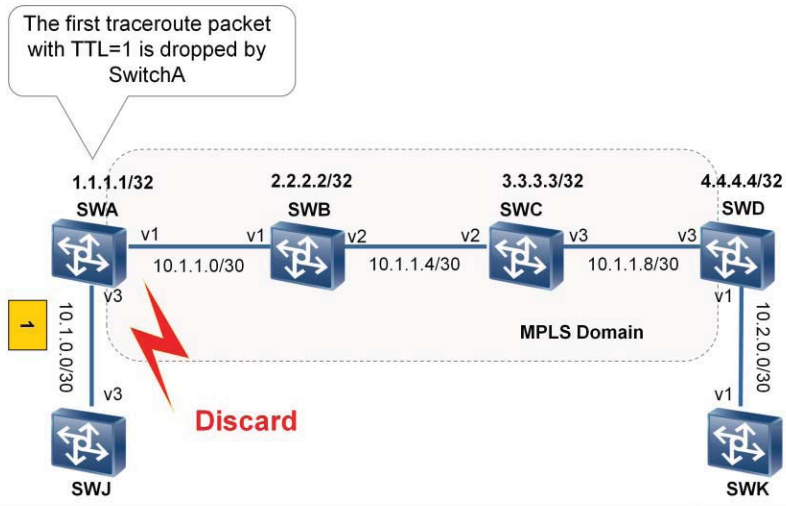
```
<SWD>debug mpls packet
<SWD>debug ip packet acl 3000
<SWD>terminal monitor
<SWD>terminal debugging

*0.99910344 SWD IP/8/debug_case:
Receiving, interface = Serial3, version = 4, headlen = 20, tos = 0,
pktlen = 84, pktid = 9625, offset = 0, ttl = 10, protocol = 1,
checksum = 30476, s = 10.1.0.1, d = 10.2.0.1
prompt: Receiving IP packet from Serial3

*0.99910344 SWD IP/8/debug_case:
Sending, interface = Serial1, version = 4, headlen = 20, tos = 0,
pktlen = 84, pktid = 9625, offset = 0, ttl = 9, protocol = 1,
checksum = 30732, s = 10.1.0.1, d = 10.2.0.1
prompt: Sending the packet from Serial3 at Serial1
```


MPLS TTL Configuration

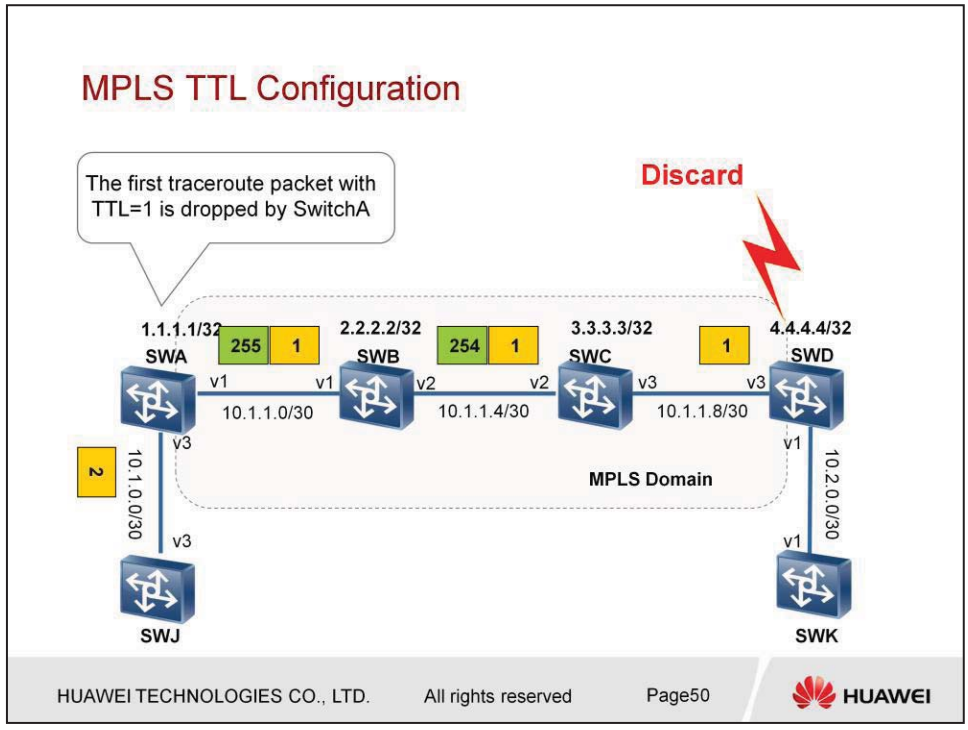


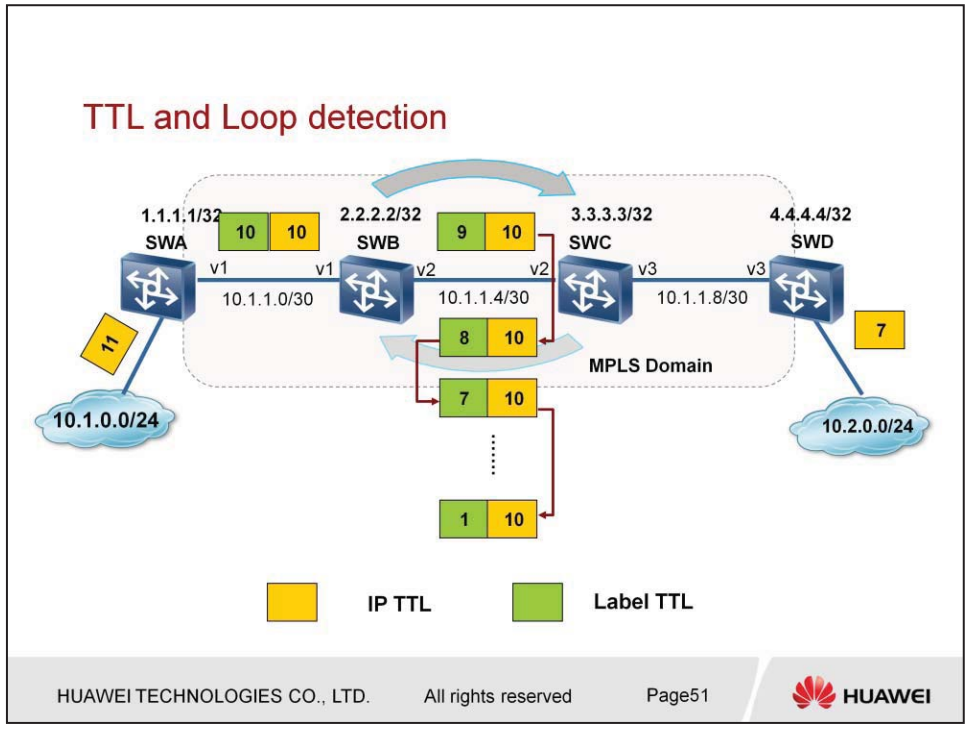

```
<SWJ>tracert 10.2.0.1
traceroute to  10.2.0.1(10.2.0.1) 30 hops max,40 bytes packet
1 10.1.0.2 31 ms  31 ms  32 ms
2 10.1.1.10 156 ms  94 ms  125 ms
3 10.2.0.1 125 ms  125 ms  125 ms
```

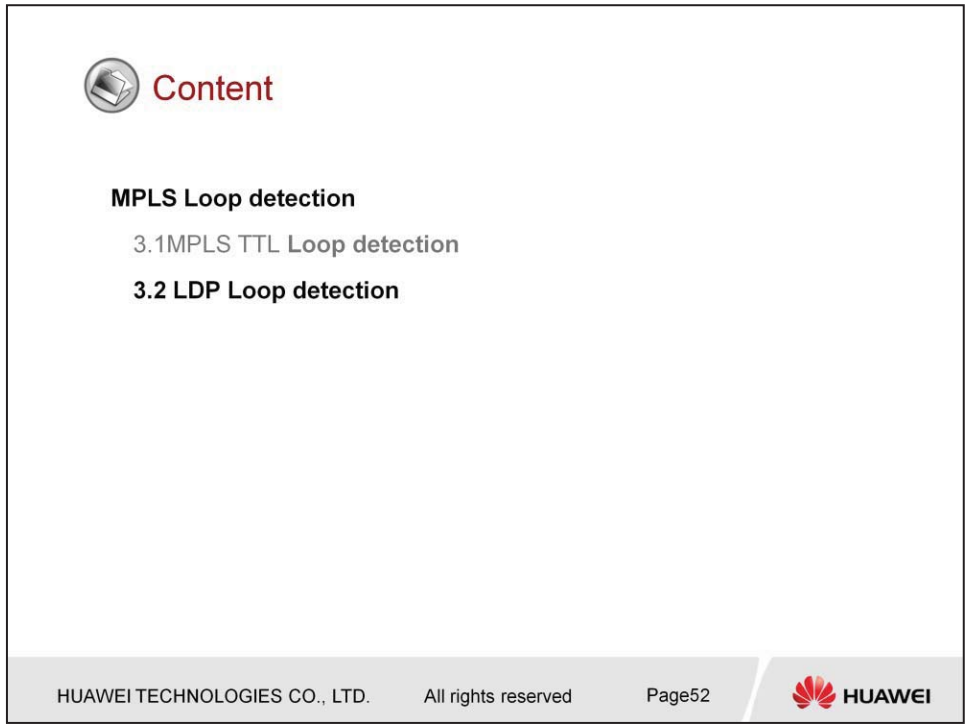

MPLS TTL Configuration



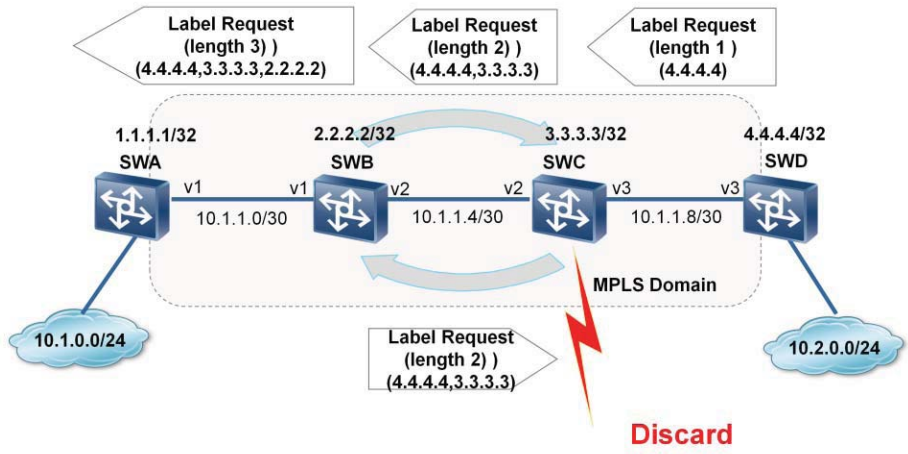
 Page49 HUAWEI

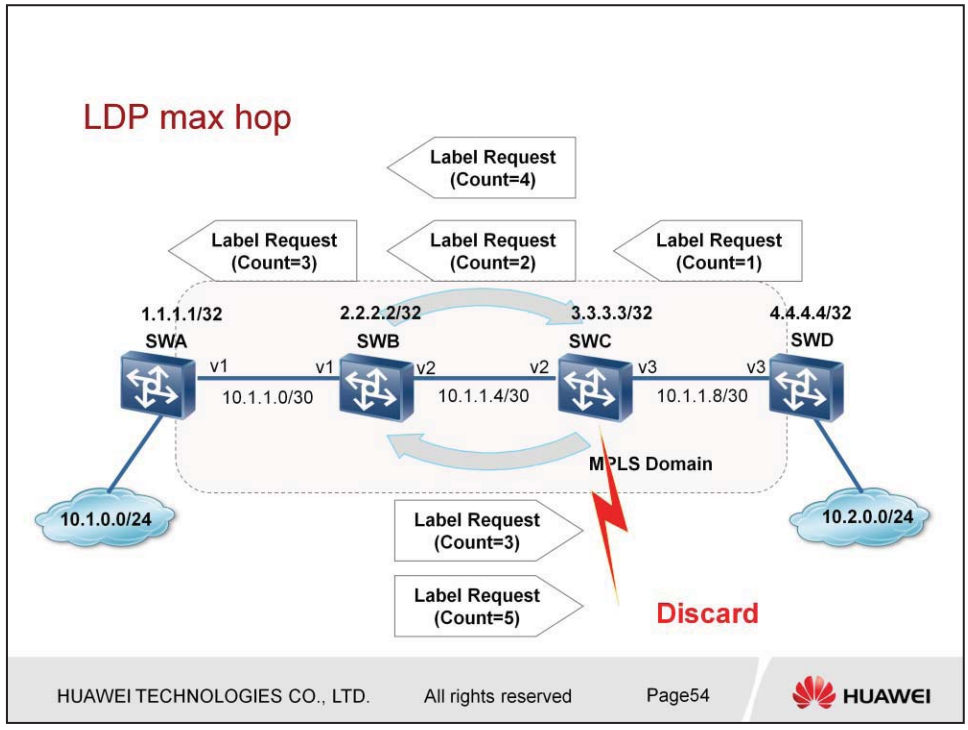

MPLS TTL Configuration
The first traceroute packet with TTL=1 is dropped by SwitchA
Discard
1.1.1.1/32
SWA
255
1
2.2.2.2/32
SWB
254
1
3.3.3.3/32
SWC
1
4.4.4.4/32
SWD
v1
v1
v2
v2
v3
v3
10.1.1.0/30
10.1.1.4/30
10.1.1.8/30
v3
v1
2
10.1.0.0/30
MPLS Domain
10.2.0.0/30
v3
v1
SWJ
SWK
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page50
HUAWEI


TTL and Loop detection
1.1.1.1/32
SWA
10
10
2.2.2.2/32
SWB
9
10
3.3.3.3/32
SWC
4.4.4.4/32
SWD
v1
v1
v2
v2
v3
v3
10.1.1.0/30
10.1.1.4/30
10.1.1.8/30
11
8
10
MPLS Domain
7
10
7
10.1.0.0/24
10.2.0.0/24
1
10
IP TTL
Label TTL
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page51
HUAWEI


Content
MPLS Loop detection
3.1MPLS TTL Loop detection
3.2 LDP Loop detection
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page52
HUAWEI

LDP Path vector



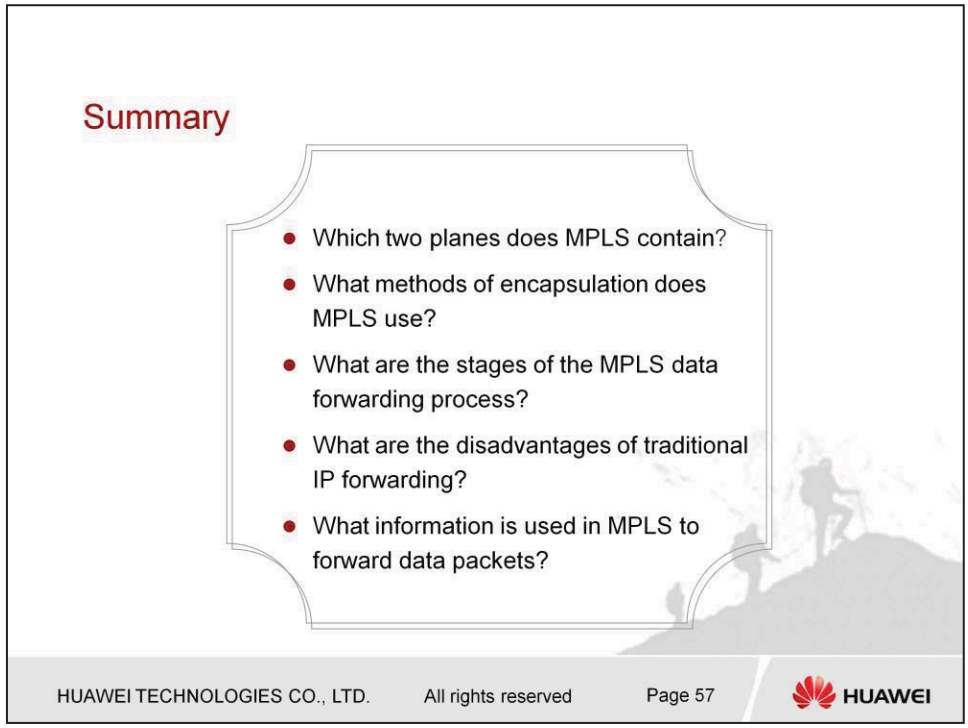

LDP max hop
Label Request (Count=4)
Label Request (Count=3)
Label Request (Count=2)
Label Request (Count=1)
1.1.1.1/32
SWA
2.2.2.2/32
SWB
3.3.3.3/32
SWC
4.4.4.4/32
SWD
v1
v1
v2
v2
v3
v3
10.1.1.0/30
10.1.1.4/30
10.1.1.8/30
MPLS Domain
10.1.0.0/24
10.2.0.0/24
Label Request (Count=3)
Label Request (Count=5)
Discard
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page54
HUAWEI

LDP Loop detection configuration

```
[SWC-mpls-ldp]display mpls ldp

                LDP Global Information
------------------------------------------------------------------------------
Protocol Version      : V1      Neighbor Liveness    : 600 Sec
Graceful Restart      : Off     FT Reconnect Timer   : 300 Sec
MTU Signaling         : On      Recovery Timer       : 300 Sec

                LDP Instance Information
------------------------------------------------------------------------------
Instance ID           : 0       VPN-Instance         :
Instance Status       : Active  LSR ID               : 3.3.3.3
Hop Count Limit       : 32      Path Vector Limit    : 32
Loop Detection        : Off
DU Re-advertise Timer : 10 Sec  DU Re-advertise Flag : On
DU Explicit Request   : Off     Request Retry Flag   : On
Label Distribution Mode : Ordered Label Retention Mode : Liberal

------------------------------------------------------------------------------
```


LDP Loop detection configuration

```
[SWC-mpls-ldp]
[SWC-mpls-ldp]loop-detect
Warning: Loop-Detection cannot be configured after enabling LDP on an interface
[SWC-mpls-ldp]quit
[SWC]undo mpls ldp
[SWC]mpls ldp
[SWC-mpls-ldp]loop-detect
[SWC-mpls-ldp]hops-count ?
  INTEGER<1-32>  Value of the maximum Hop-Count
[SWC-mpls-ldp]path-vectors ?
  INTEGER<1-32>  Value of the Path-Vector limit
```

Q: Which two plane does MPLS contain?

A: MPLS includes control plane and data plane. Control plane's charge is routing information transmission and label distribution, data plane's charge is data forwarding.

Q: What methods of encapsulation does MPLS use?

A: Frame mode and cell mode. Ethernet and PPP adopt frame mode encapsulation, ATM adopts cell mode encapsulation.

Q: What are the stages of the MPLS data forwarding process?

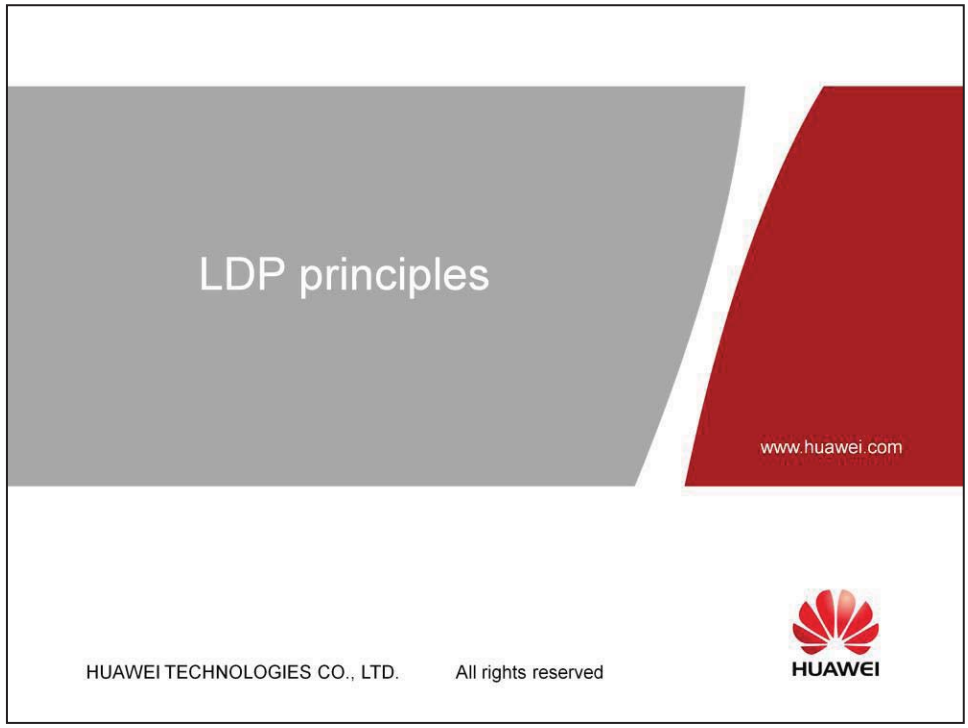
A: Ingress LER pushes label, LSR swaps label, Egress LER pops label.

Q: What are the disadvantages of traditional IP forwarding?

A: The speed of traditional IP forwarding is low, and all the routers should maintain the routes of the entire network, so it is hard to deploy QoS.

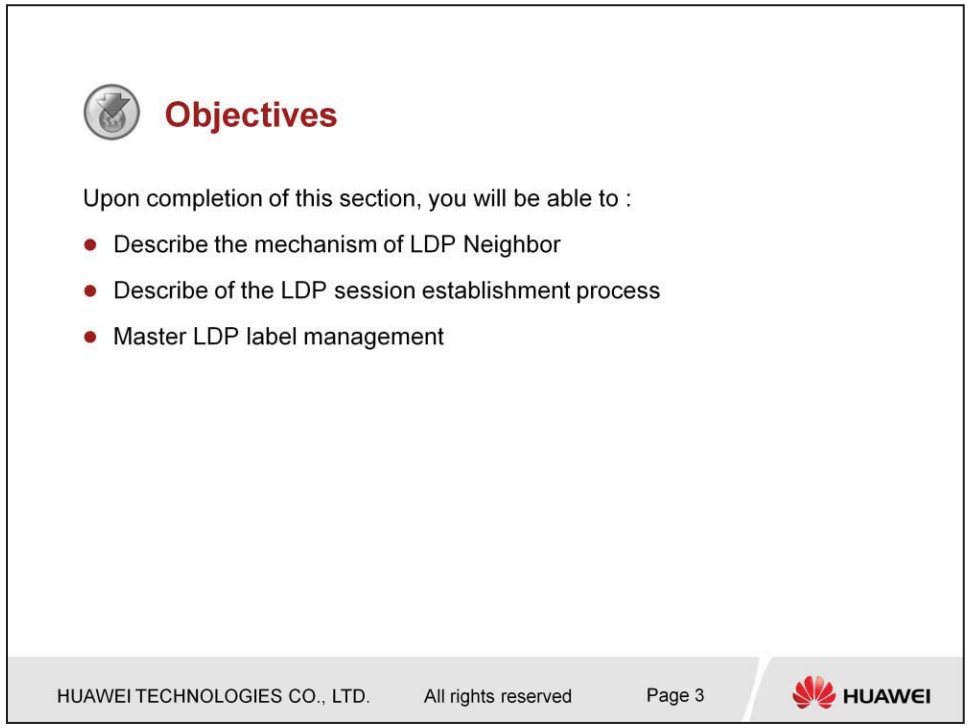
Q: What information is used in MPLS to forward data packets?

A: MPLS completes data forwarding by label.

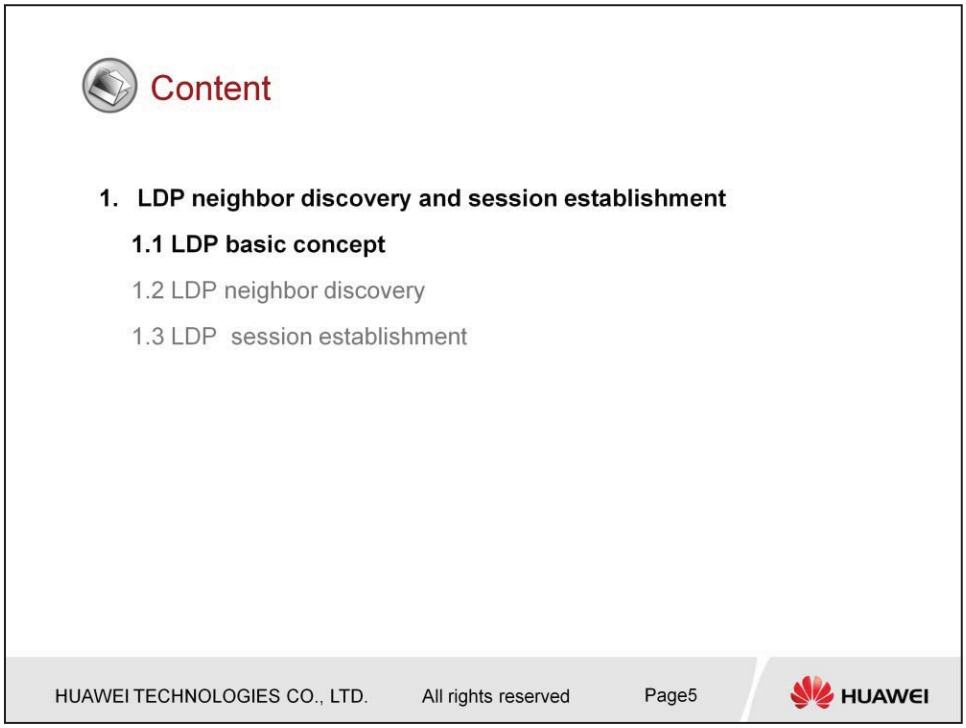

LDP principles
www.huawei.com
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
HUAWEI

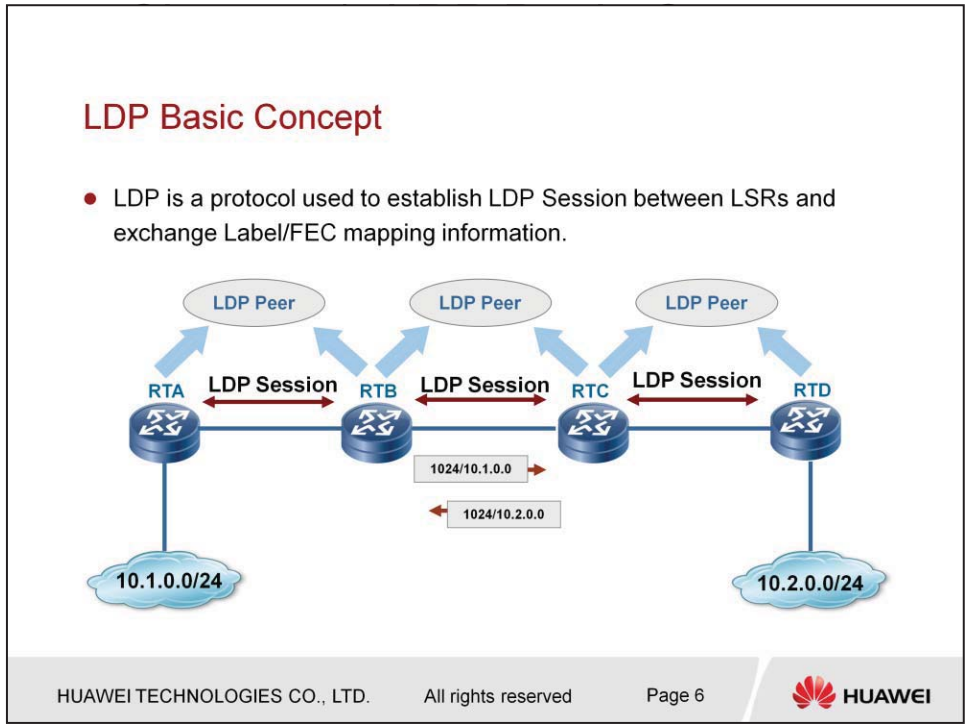
Foreword

This section introduces the LDP label space, Label Distribution Protocol, LDP neighbor agreement works, and in the VPN, QoS and traffic engineering.


Objectives
Upon completion of this section, you will be able to :
Describe the mechanism of LDP Neighbor
Describe of the LDP session establishment process
Master LDP label management
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page 3
HUAWEI


Content
LDP neighbor discovery and session establishment
LDP label management
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page4
HUAWEI


Content
1. LDP neighbor discovery and session establishment
1.1 LDP basic concept
1.2 LDP neighbor discovery
1.3 LDP session establishment
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page5
HUAWEI



MPLS needs label distribution protocol to complete label distribution control and retention, there are many label distribution protocols, LDP is one of them, LDP protocol can be used to exchange label information between LSRs.

As shown in the figure above, RTA，RTB，RTC，RTD are configured as LSRs.

Two LSRs which run LDP establish LDP Session and exchange Label/FEC mapping information, two routers who have established LDP Session are called LDP Peers.

LDP Message Types

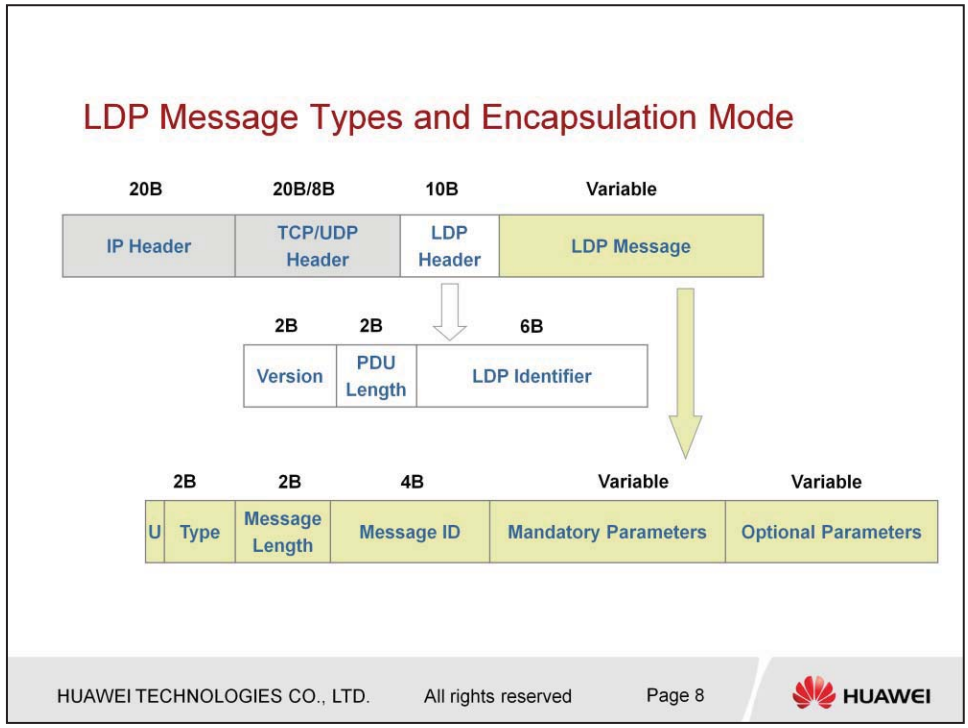
- Discovery message: announce and maintain the presence of an LSR in a network.
- Session message: establish, maintain, and terminate sessions between LDP peers.
- Advertisement message: create, change, and delete label mapping for FECs.
- Notification message: announce advisory information and error information.

LSRs which run LDP protocol exchange LDP message to discover neighbor, establish and maintain LDP Session, manage label. LDP message is carried by UDP or TCP, the port number is 646. Let's simply introduce some common used LDP messages and the main function of each message. According to the function of message, LDP message can be divided into 4 types: Discovery Message, Session Message, Advertisement Message and Notification Message. Discovery message is used to announce and maintain the presence of an LSR in a network ;

Session message is used to establish, maintain and terminate LDP Session between LDP peer. Advertisement messages is used to create, change, and delete label mappings for FECs. Notification message is used to announce alarm and error information.

Discovery Message is used to discover neighbor, carried by UDP packet. LDP requires to transmit message reliably and orderly, so LDP uses TCP to establish Session; Session Message，Advertisement Message，Notification Message are

all transmitted based on TCP.



LDP PDU includes two parts: LDP Header and LDP Message.

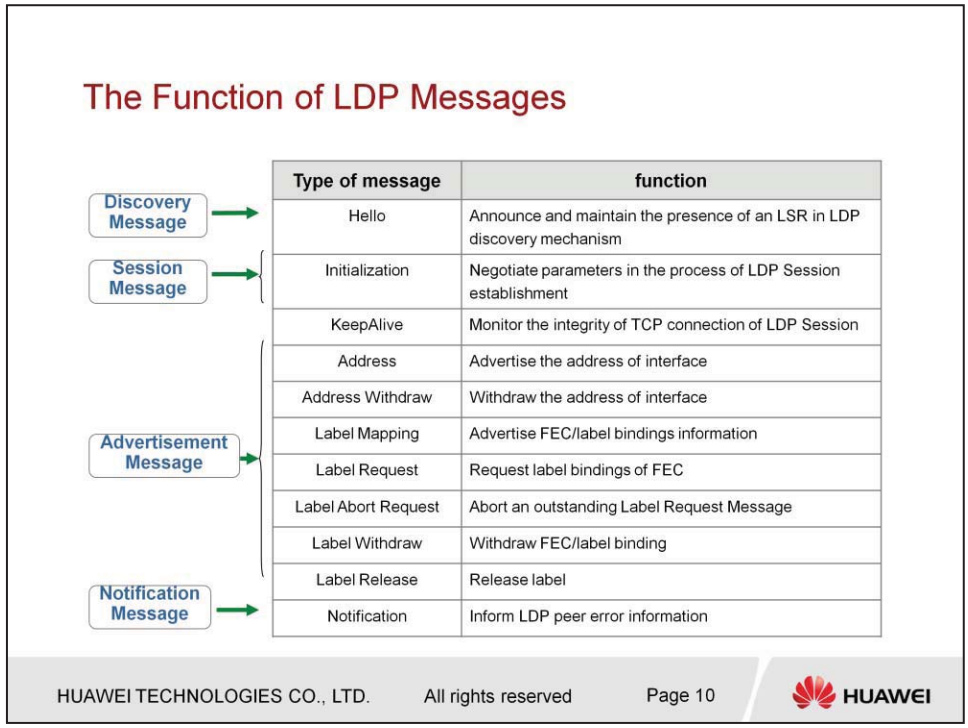
The length of LDP Header is 10 bytes, it includes Version, PDU length and LDP identifier. Version field uses two Bytes, indicates the version number of LDP protocol, and present version number is 1. the length of PDU length field is 2 Bytes, indicates the total length of this PDU in octets, excluding the Version and

PDU Length fields. The length of LDP identifier is 6 Bytes, the first four octets identify a LSR uniquely, the last two Bytes identify a label space within the LSR, there is particular introduction about label space in LDP label management.

LDP message includes 5 parts. U uses 1 bit, it is Unknown Message bit. Upon receipt of Unknown message, if U=0, a notification must be returned to the message originator, if U=1, the unknown message must be ignored and no notification is sent to originator. Message Length used 4 bytes, specifies the total length of Message ID， Mandatory Parameters and Optional Parameters in octets.

Message uses 32 bits to identify this message. Mandatory Parameters and Optional Parameters respectively are variable length set of required message parameters and optional parameters. Now, the following message types are defined of

LDP: Notification， Hello， Initialization， KeepAlive，
Address ,Address Withdraw，Label Mapping，Label Request，
Label Abort Request,Label Withdraw，Label Release.



The function of each message is as follows :

Notification Message is used to inform an LDP peer of error or other advisory information such as the state of the LDP Session.

Hello Message is used to announce this LSR and discover neighbor as a part of LDP discovery mechanism.

Initialization Message is used to negotiate parameters in the process of LDP Session establishment.

KeepAlive Message is used to monitor the integrity of TCP connection of LDP Session.

Address Message is used to announce the address of interface.

Address Withdraw Message is used to withdraw the address of interface.

Label Mapping is used to advertise FEC/label bindings information.

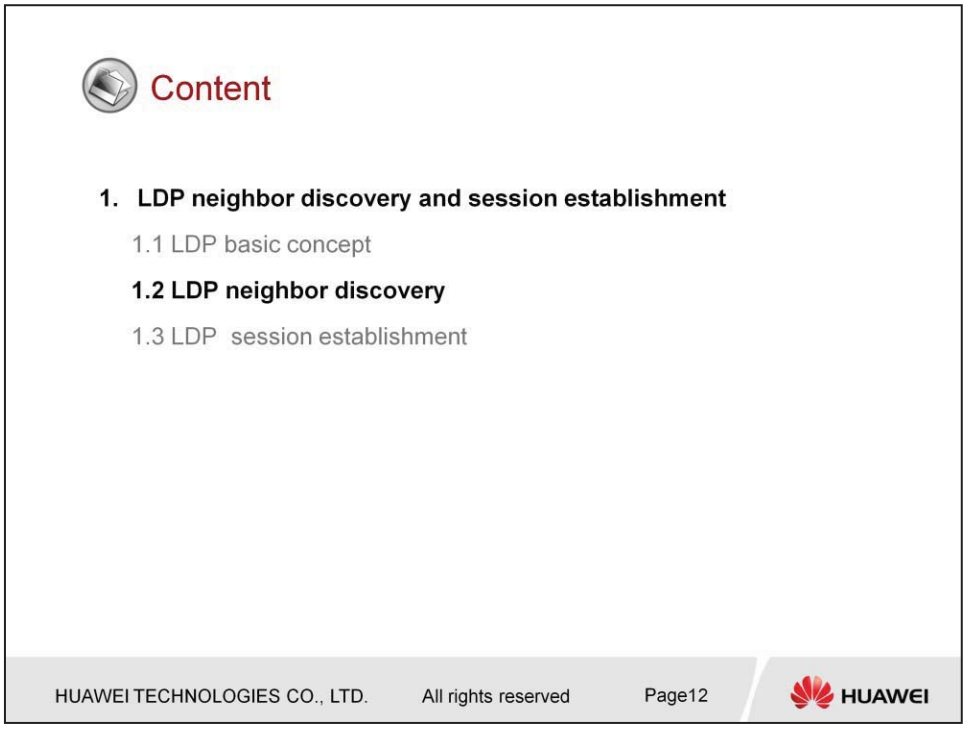
Label Request is used to request label bindings of FEC.

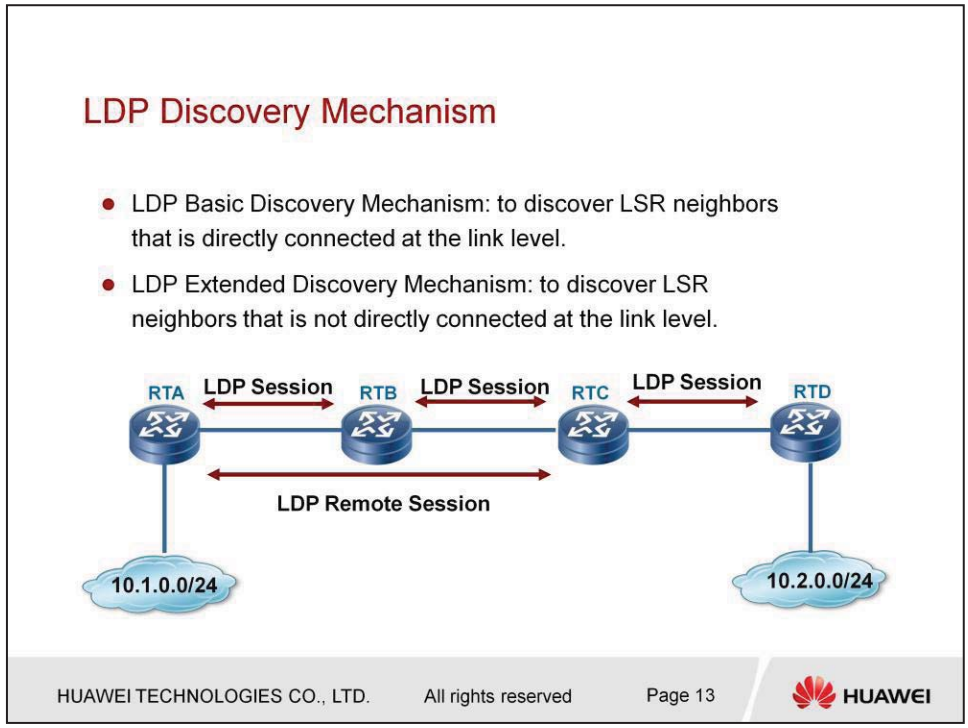
Label Abort Request is used to abort an outstanding Label Request Message.

Label Withdraw is used to withdraw FEC/label binding. LSR informs peers that this peer can’t use label advertised

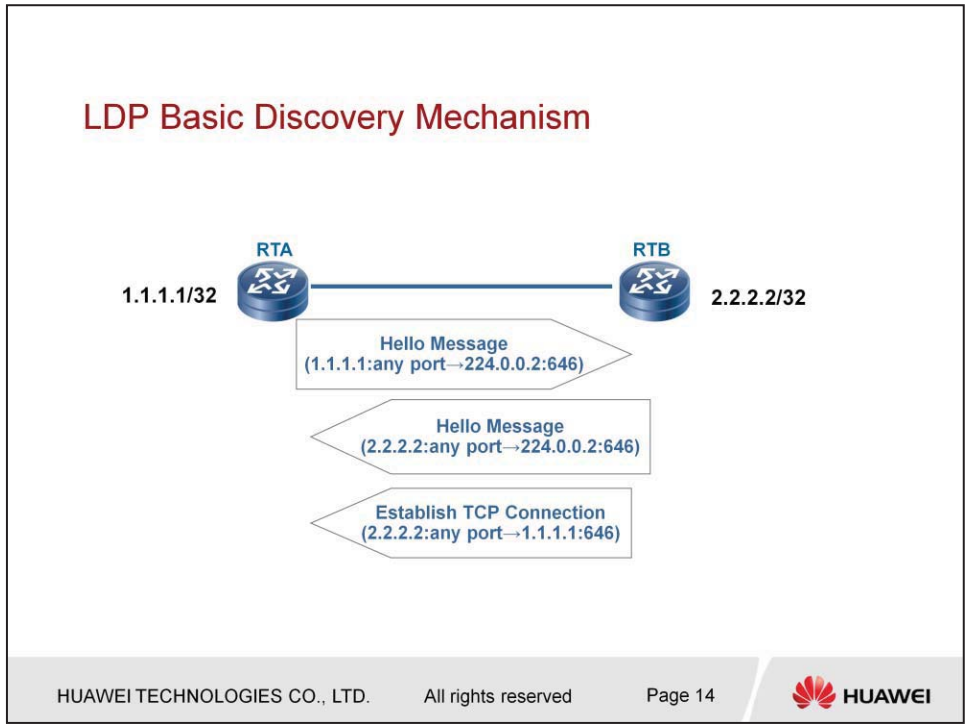
continuously via sending Label Withdraw Message.

Label Release is used to release label. when a LSR doesn't need the label received from LDP Peer before, a Label Release Message should be sent to the LDP Peer.


Content
1. LDP neighbor discovery and session establishment
1.1 LDP basic concept
1.2 LDP neighbor discovery
1.3 LDP session establishment
HUAWEI TECHNOLOGIES CO., LTD. All rights reserved Page12
HUAWEI



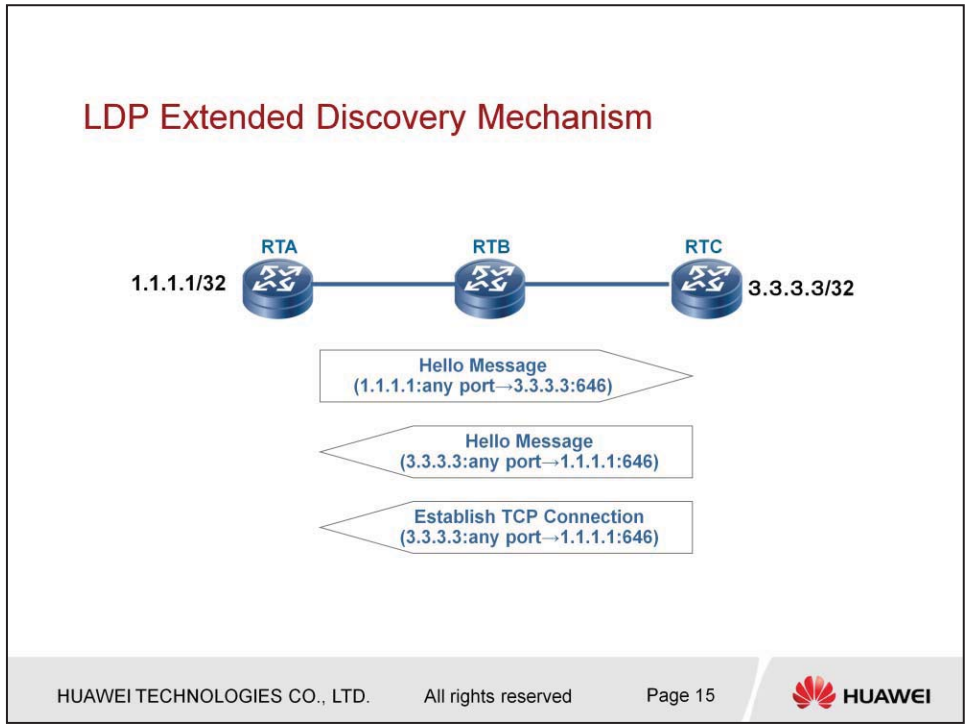
LSR discovers LDP Peers via LDP Discovery Mechanism. LDP Discovery Mechanism includes LDP Basic Discovery Mechanism and LDP Extended Discovery Mechanism. LDP Basic Discovery Mechanism can automatically discover LDP Peers that are directly connected to the same link, so LDP Peer does not need to be appointed specifically in this case. LDP Extended Discovery Mechanism can discover LDP Peers that are not directly connected to the link.



Discovery message is used for neighbor discovery, it provides the mechanism:

LSR indicates its presence in a network by sending a Hello Message periodically.

This message is encapsulated in UDP packet, destination port is 646. this message's destination IP address is multicast IP address of 224.0.0.2, namely, this message is sent to all the routers in the subnet (the figure shows that RTA and RTB send Hello Message on multicast address 224.0.0.2 periodically). Hello Message carries LDP identifier for the label space the LSR use in order to inform peer, then LSR with higher IP address attempts to establish TCP connection as active role. After TCP connection is established, LSR continuously sends hello message in order to discover new neighbor and detect error.



Different from LDP Basic Discovery Mechanism, LDP Extended Discovery Mechanism means LSR (such as RTA) which runs LDP protocol periodically sends Hello Message to specific destination IP, so it needs to appoint LDP Peer to establish Session via configuration, another LSR (such as RTB) determines whether it replies this message, if it replies, Hello Message will be sent to specified LSR (such as RTA).

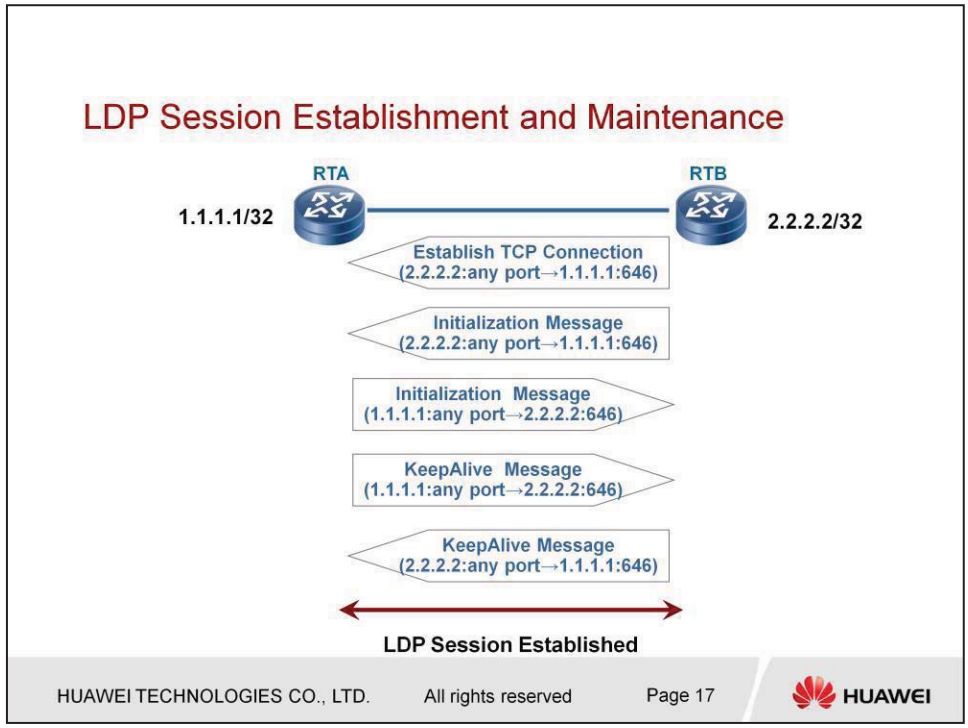
Content

1. **LDP neighbor discovery and session establishment**

1.1 LDP basic concept

1.2 LDP neighbor discovery

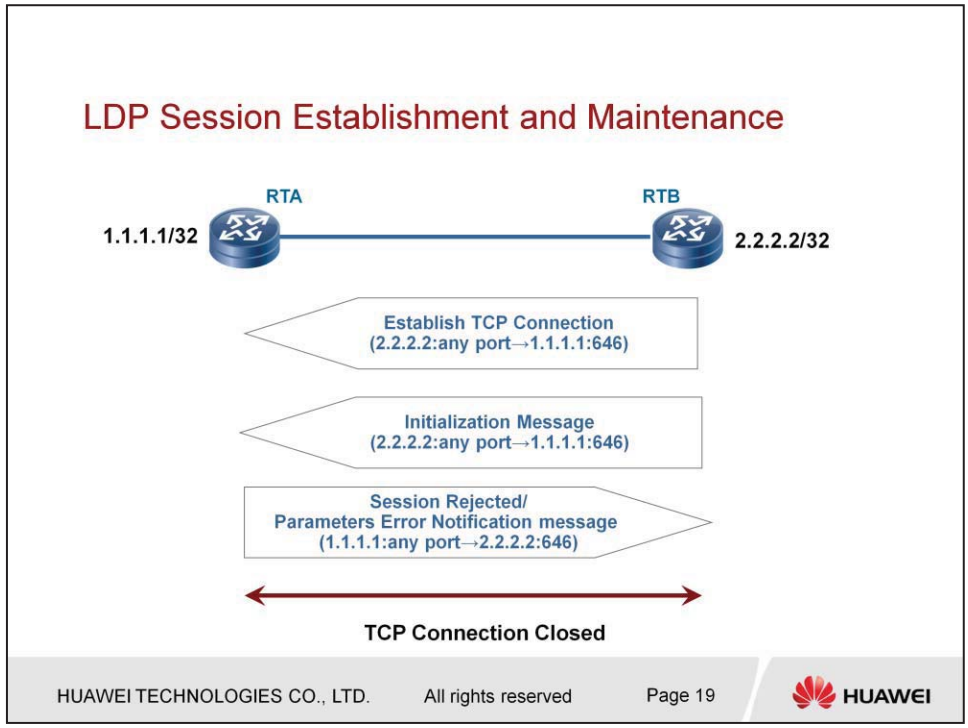
1.3 LDP session establishment



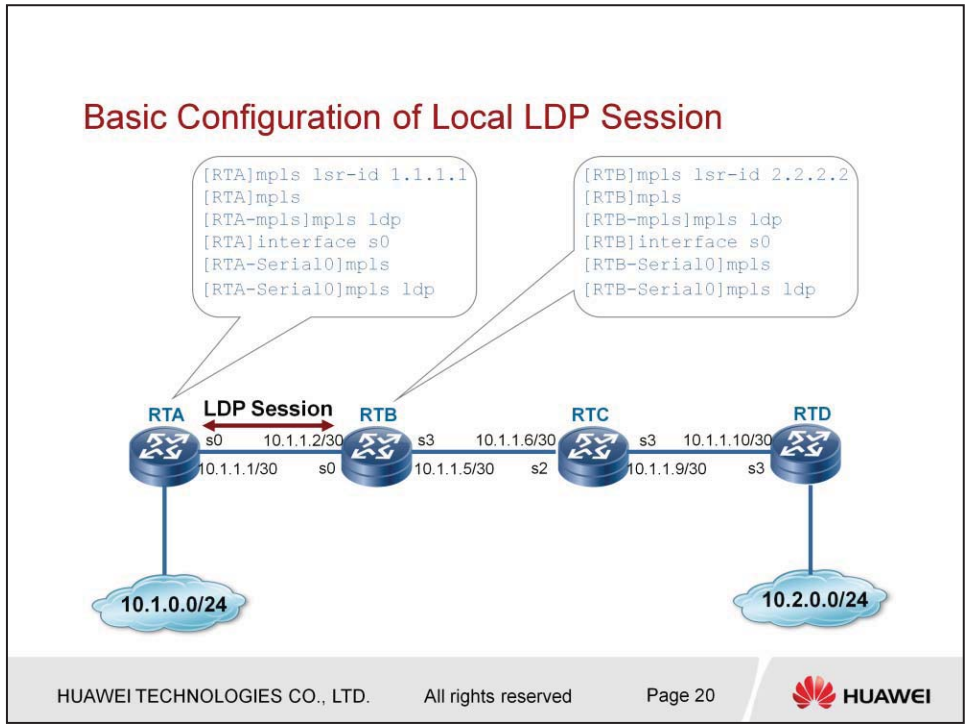
Before TCP connection is established, two LSRs (such as RTA and RTB) first determine which IP address is used to establish TCP connection and which plays the active role or passive role, then the active role attempts to establish connection. If the Hello Message carries Transport Address, this Transport Address is used to establish TCP connection, if not, thus the source IP address of Hello Message is used to establish TCP connection. Both of two LSRs obtain address which the peer is used to establish TCP connection from Hello Message sent by peer, then it compares the two address, the LSR with higher address will be active role and it attempts to establish TCP connection. The figure shows that RTB is active and attempts to establish TCP connection.

After TCP connection is established, the active role (RTB) initiates negotiation of session parameters by sending an Initialization Message, includes LDP protocol number, label distribution mode, etc. the passive role (RTA) checks whether the session parameters are acceptable. If they are, it replies with an Initialization Message of its own to propose the parameters it wishes to use and a KeepAlive Message to signal acceptable of active's parameters. After peers received KeepAlive Message of

each other, session is established. Two LSRs become LDP Peers and exchange Advertisement Message.



If the passive role (RTA) can not accept parameters, it will send Error Notification Message to peer to close the TCP connection.



RTA's loopback1 address is 1.1.1.1/32, RTB's Loopback1 address is 2.2.2.2/32, RTC's Loopback1 address is 3.3.3.3/32, RTD's Loopback1 address is 4.4.4.4/32.

Configure RTA and RTB to establish LDP Session in order to distribute label.

Basic Configuration Procedure:

1. First configure MPLS lsr-id for RTA and RTB, it is used to establish and maintain LDP Session. By default, there is no LSR ID on VRP platform. It must be configured manually; usually, the LSR ID is IPv4 address of a loopback interface and it should be unique in the MPLS domain.

2. Lsr-id is used to establish LDP Session, by default, lsr-id is used to establish LDP Session on VRP platform, so

3routing protocol should be configured in order that lsr-id of RTA and RTB are reachable.

3. Enable MPLS and MPLS LDP in system view.

4. Enable MPLS and MPLS LDP on corresponding interface.

Configuration introduction:

[RTA]mpls lsr-id 1.1.1.1

Configure lsr-id

[RTA]mpls

Enable mpls in system-view

[RTA-mpls]mpls ldp

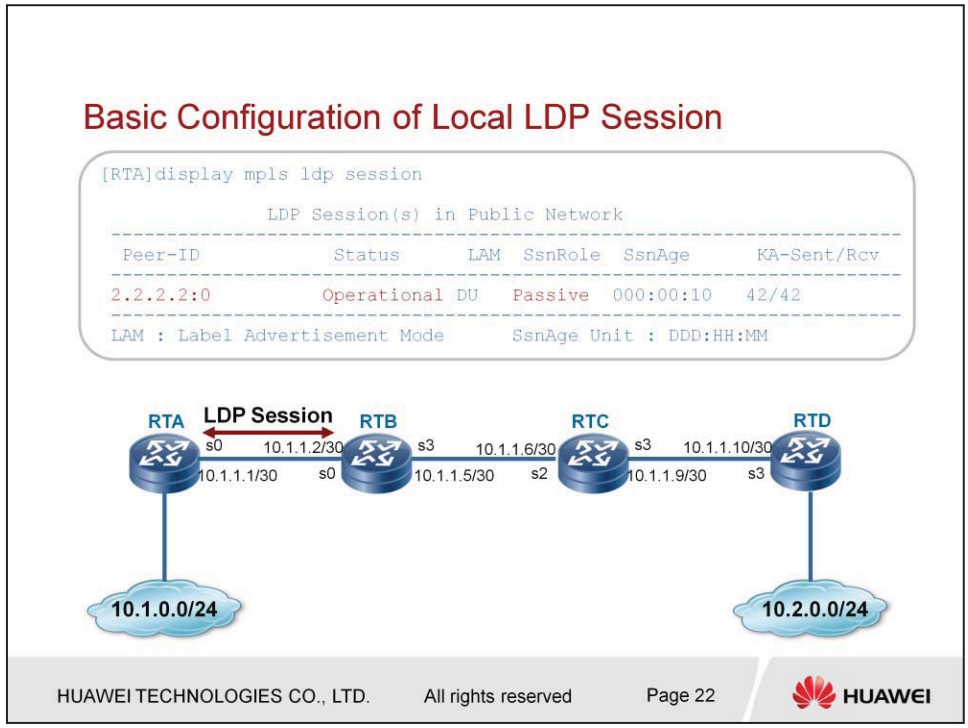
Enable mpls ldp in system-view or mpls-view

[RTA]inter s0

[RTA-Serial0]mpls

[RTA-Serial0]mpls ldp

Enable mpls and ldp in interface-view



[[RTA]dis mpls ldp session

This command shows the state of mpls ldp session, operational state indicates that Session has been established. Peer-ID is LDP Peer's LDP Identifier, it consists of LDP Peer's lsr-id (here is 2.2.2.2) and 2 Bytes which indicates label space (here is 0, indicates platform-based label space).

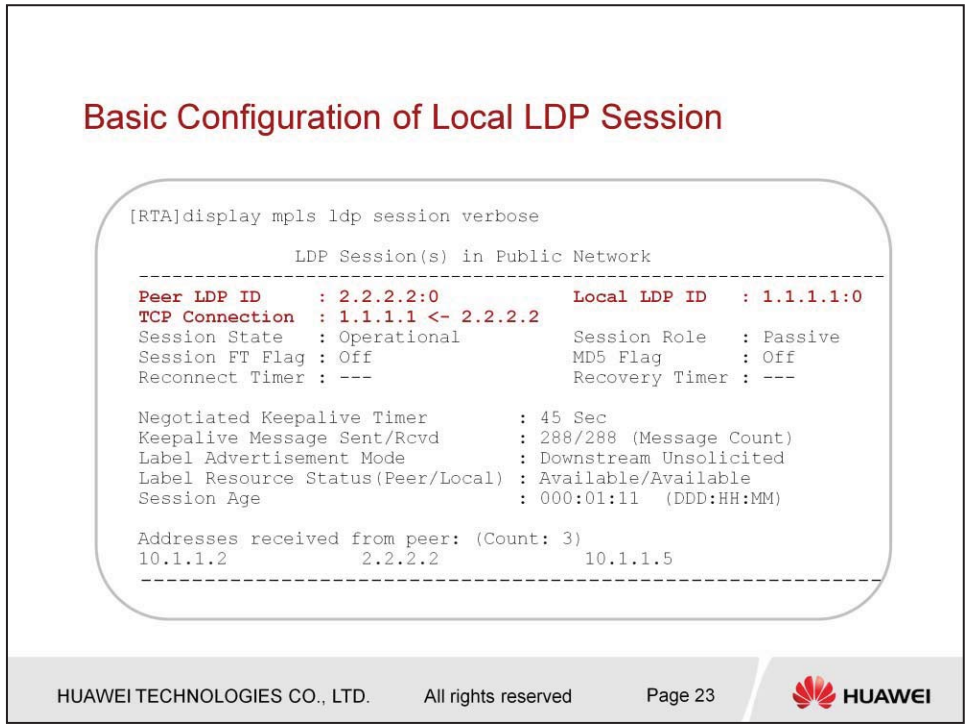
RTA's lsr-id is smaller than RTB's, so RTA is passive role, SSnRole is Passive. It can also use following command to show LDP Peer's information.

[RTA-Serial0]dis mpls ldp peer

LDP Peer Information in Public network

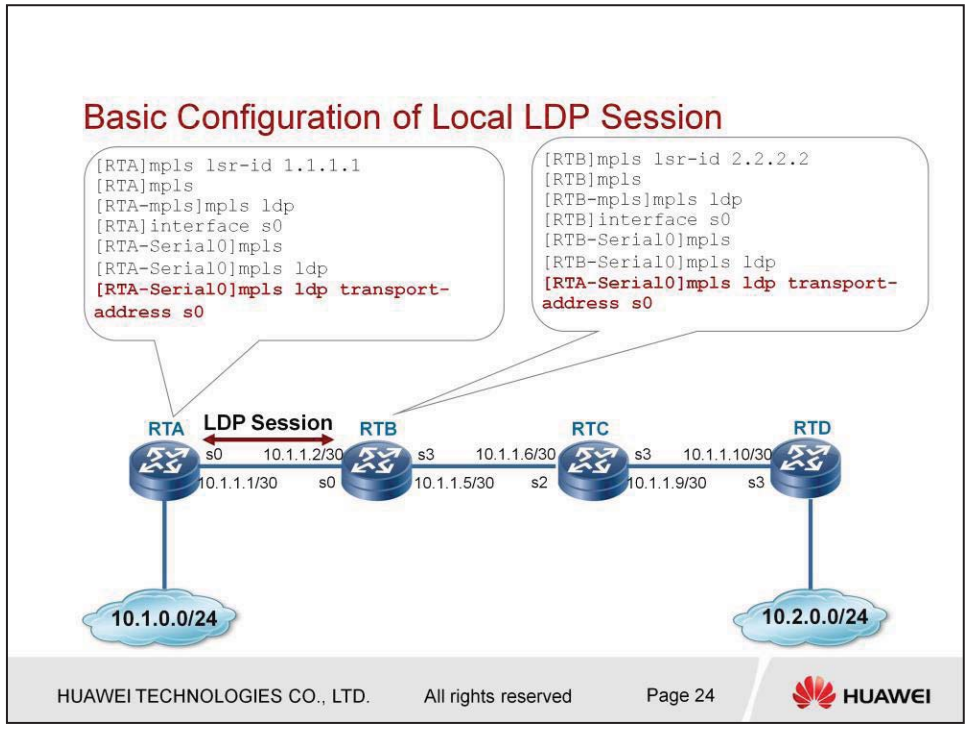
Peer-ID Transport-Address Discovery-Source

2.2.2.2:0 10.1.1.2 Serial0



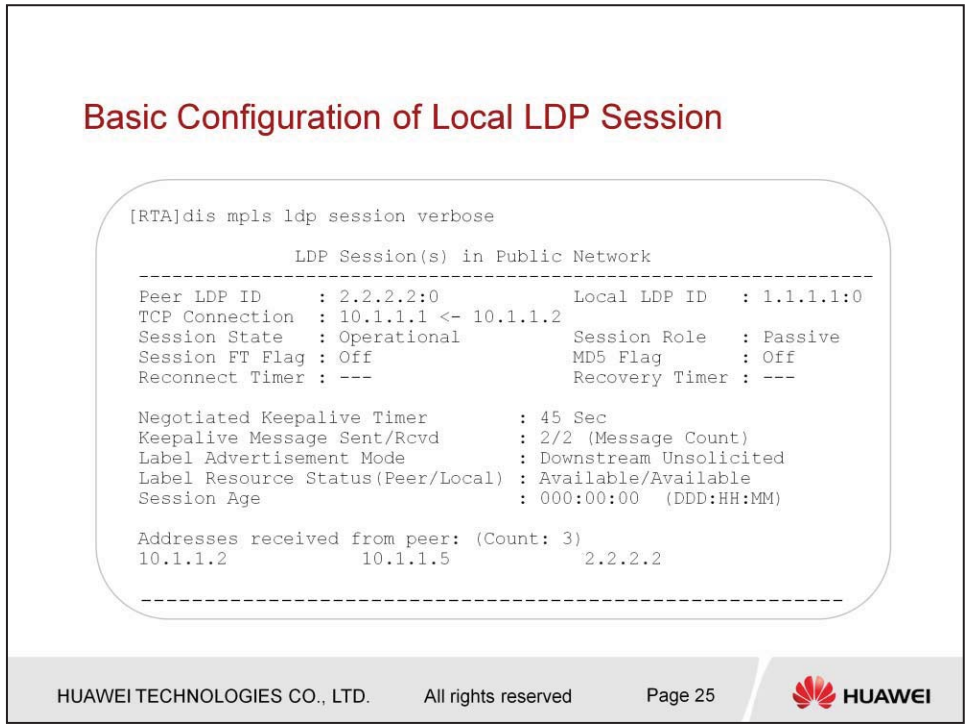
[RTA]dis mpls ldp session verbose

It shows detailed information of LDP Session. From the table, it shows that lsr-id is used to establish TCP connection by default on platform VRP .

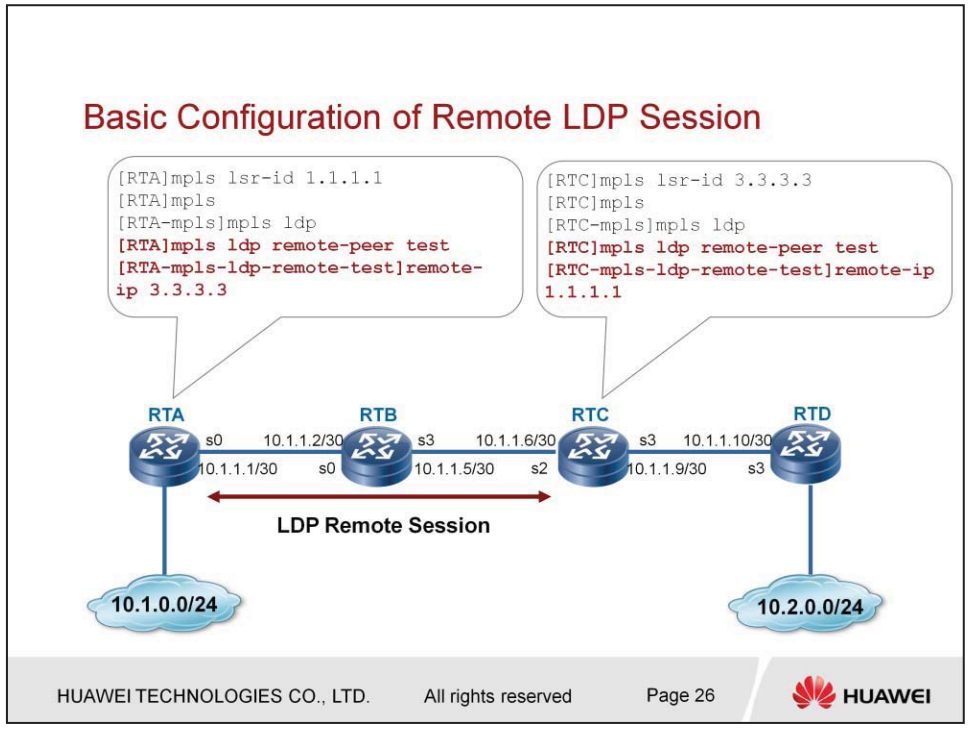


[RTA-Serial0]mpls ldp transport-address s0

Configure that directly interface S0 is used to establish TCP connection.



As shown of the detailed information of LDP Session, after Transport Address is configured, LDP uses IP address which is configured by Transport Address to establish TCP connection.



Configure Remote LDP Session between RTA and RTC.

Configuration procedure:

1. It is the same with local LDP Session, it needs to configure lsr-id and ensure the reachable of lsr-id of RTA and RTB.

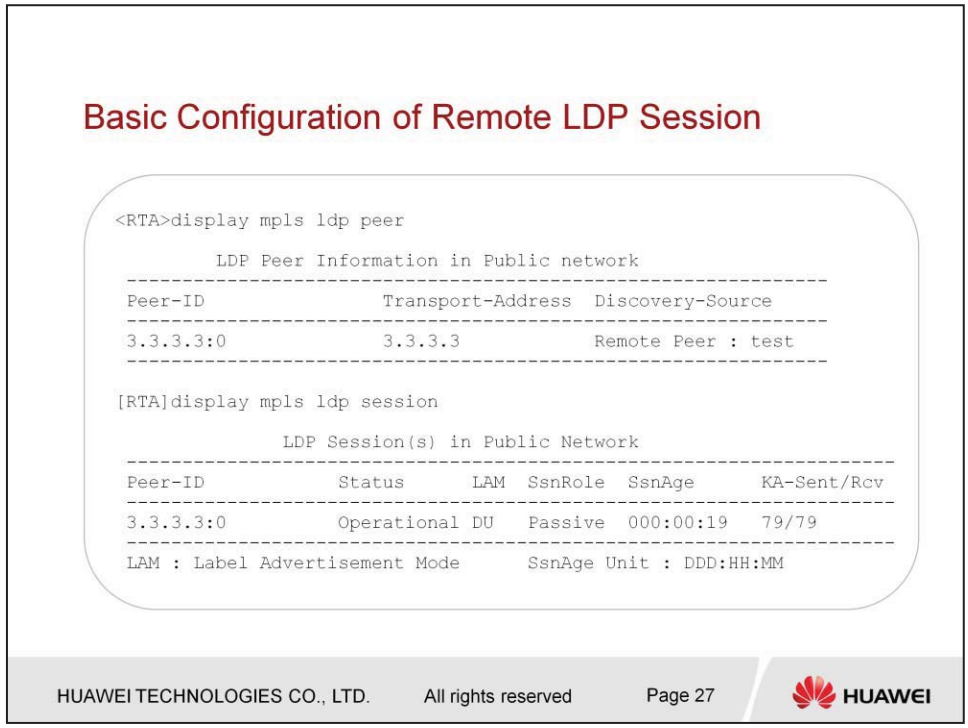
2. Configure LDP remote peer.

Configuration introduction:

[RTA]mpls ldp remote-peer test

[RTA-mpls-ldp-remote-test]remote-ip 3.3.3.3

It first creates a remote peer, then appoints the lsr-id of peer.



As shown in the figure, LDP Remote Session is established between RTA and RTC. It also can view LDP Remote Peer's information via following command.

[RTA]dis mpls ldp remote-peer test

LDP Remote Entity Information

Remote Peer Name: test

Remote Peer IP: 3.3.3.3 LDP ID: 1.1.1.1:0

Transport Address: 1.1.1.1 Entity Status: Active

Configured Keepalive Timer: 45 Sec Configured Hello Timer: 45 Sec Negotiated Hello Timer: 45 Sec Hello Packet sent/received: 100/98

As shown in the table, by default, LDP also uses lsr-id to establish TCP connection so as to establish LDP Remote Session (as shown in the figure, Transport Address is 1.1.1.1).

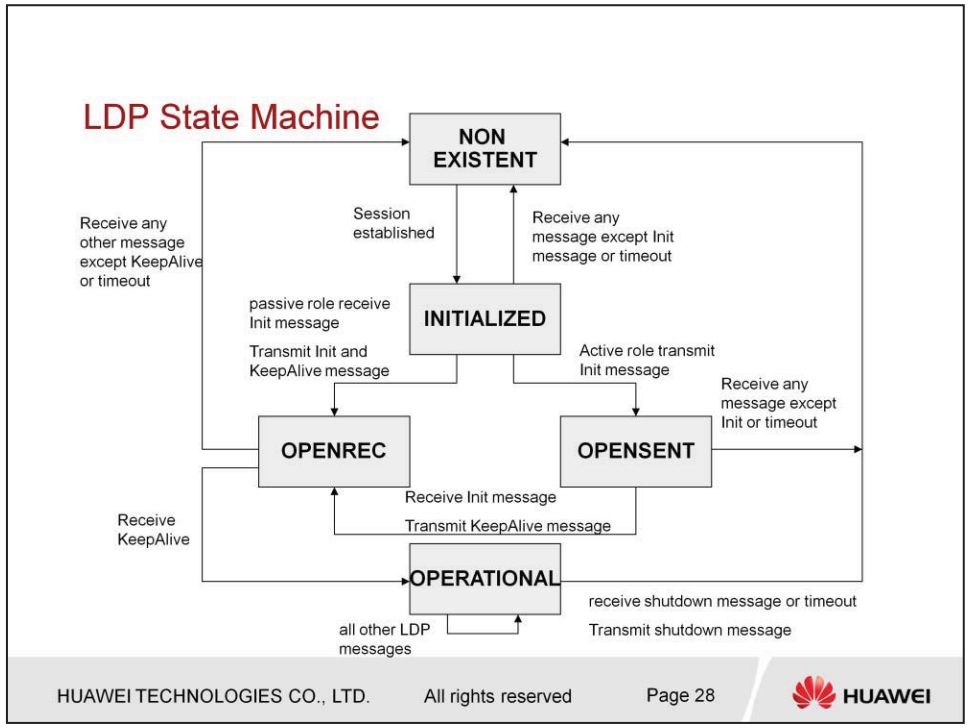
Transport Address can be modified via following command.

[RTA]mpls ldp remote-peer test

[RTA-mpls-ldp-remote-test]mpls ldp transport-address ?

LoopBack LoopBack interface

Serial Serial interface



LDP session negotiation process can be described by state machine. As shown in the figure, there are five states: NON-EXISTENT, INITIALIZED, OPENREC, OPENSENT and OPERATIONAL.

NON-EXISTENT state: this state is the first state of LDP Session, in this state, the adjacency routers exchange Hello Message, elect active role, after receive triggering of event of TCP connection establishment, the state turns to INITIALIZED.

INITIALIZED state: this state should be divided into two situation: active and passive, active role sends Initialization Message actively, then turns to OPENSENT state and waits for Initialization in response; passive role waits for

Initialization Message which is sent by active role, if the parameters carried by Initialization are acceptable, it replies with an Initialization Message and KeepAlive Message, and the state turns to OPENREC. When active and passive receives any other message except Initialization Message or timeout, the state

will turns to NON-EXISTENT.

OPENSENT state: after sending Initialization Message, the active role will change to this state, in this state, it waits for

Initialization Message and KeepAlive Message replied by passive role, if the parameters carried by Initialization are acceptable, the state will turn to OPENREC, if the parameter are not acceptable

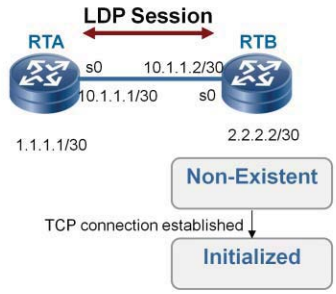
or timeout, the TCP connection will close and the state will turn to NONEXISTENT.

OPENREC state: after sending KeepAlive Message, both of active and passive become this state, wait for KeepAlive replied by peer, as long as receiving KeepAlive Message, the state will turn to OPERARIONAL; if receiving any other message or timeout, the state will turn to NON-EXISTENT.

OPERATIONAL state: this state indicates establishment of LDP Session. In this state, it can receive and send all other LDP message. If timeout or receiving Notification Message (Shutdown Message) or sending Shutdown message by itself to close session on its initiative, the state will turn to NON-EXISTENT.

Case Analysis of LDP State Machine

```
<RTB>terminal monitor
<RTB>terminal debugging
<RTB>debug mpls ldp session
*0.12902062 RTB LDP/8/Session: Serial0
 Link Hello message received on interface:
   Serial0
*0.12902062 RTB LDP/8/Session:
 Created session with LSR: 1.1.1.1
*0.12902062 RTB LDP/8/Session: Serial0
 Link Hello message sent on interface:
   Serial0
*0.12902062 RTB LDP/8/Session: Serial0
 Session(1.1.1.1,Active role) start to
   open TCP connection.
*0.12902062 RTB LDP/8/Session: Serial0
 Session(1.1.1.1)'s state changed from
   Non-existent to Initialized.
```



Use command “mpls ldp session” to enable debugging information on VRP platform, and learn more of the transferred process of LDP State Machine.

<RTB>terminal monitor

<RTB>terminal debugging

<RTB>debug mpls ldp session

Use command “mpls ldp session” to enable debugging and put out debugging information to console.

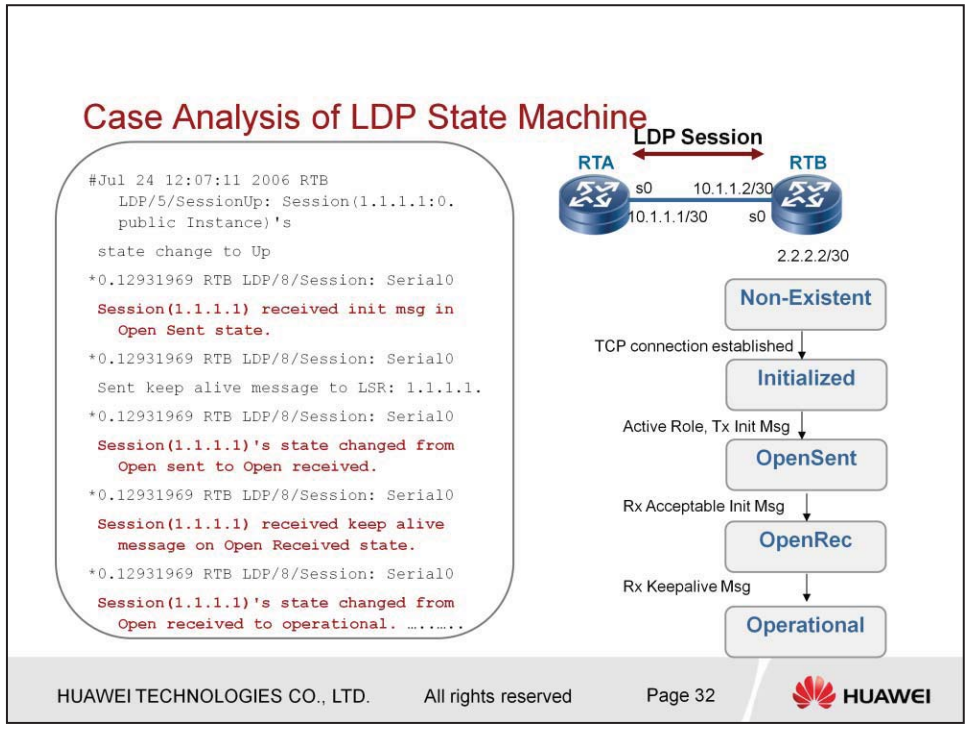
From the debugging information, you can find that RTB attempts to establish TCP connection as active role. The state will turn from NON-EXISTENT to INITIALIZED.

Case Analysis of LDP State Machine

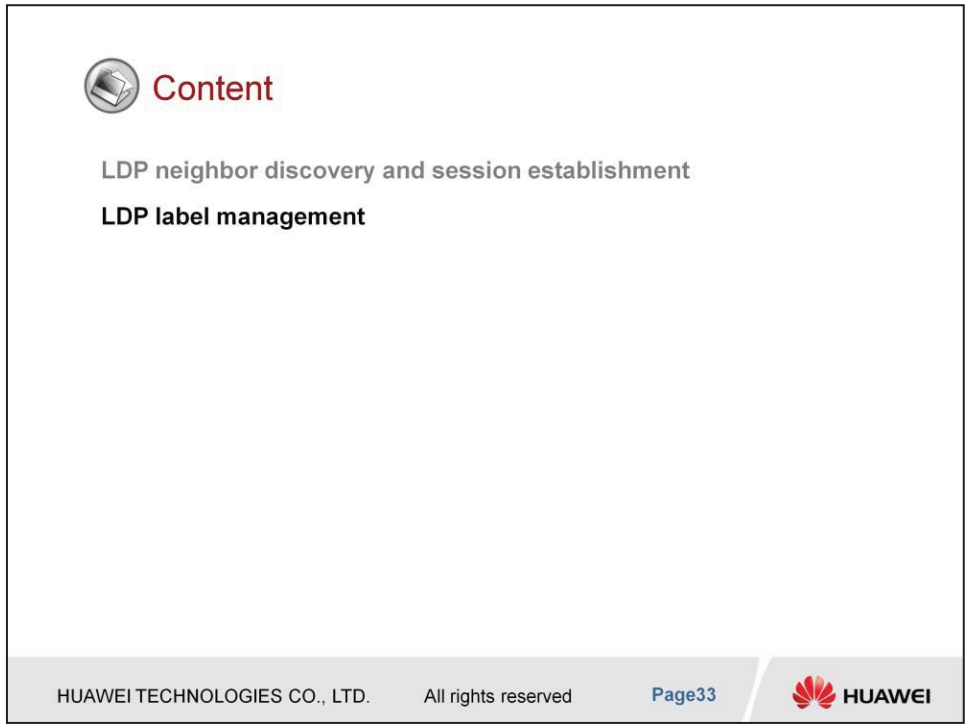
```
*0.12906969 RTB LDP/8/Session: Serial0
 Link Hello message received on
   interface: Serial0
.....
%Jul 24 12:07:11 2006 RTB LDP/5/LOG:
 Received TCP Up Event for TCP SockId 2
*0.12931844 RTB LDP/8/Session:
  TCP up event received for socket Id: 2
*0.12931844 RTB LDP/8/Session: Serial0
 Session(1.1.1.1) start to send init msg
   on Initialized state.
*0.12931844 RTB LDP/8/Session:
 Session Init message sent to LSR:
   1.1.1.1
*0.12931844 RTB LDP/8/Session: Serial0
 Session(1.1.1.1)'s state changed from
   Initialized to Open Sent.
```

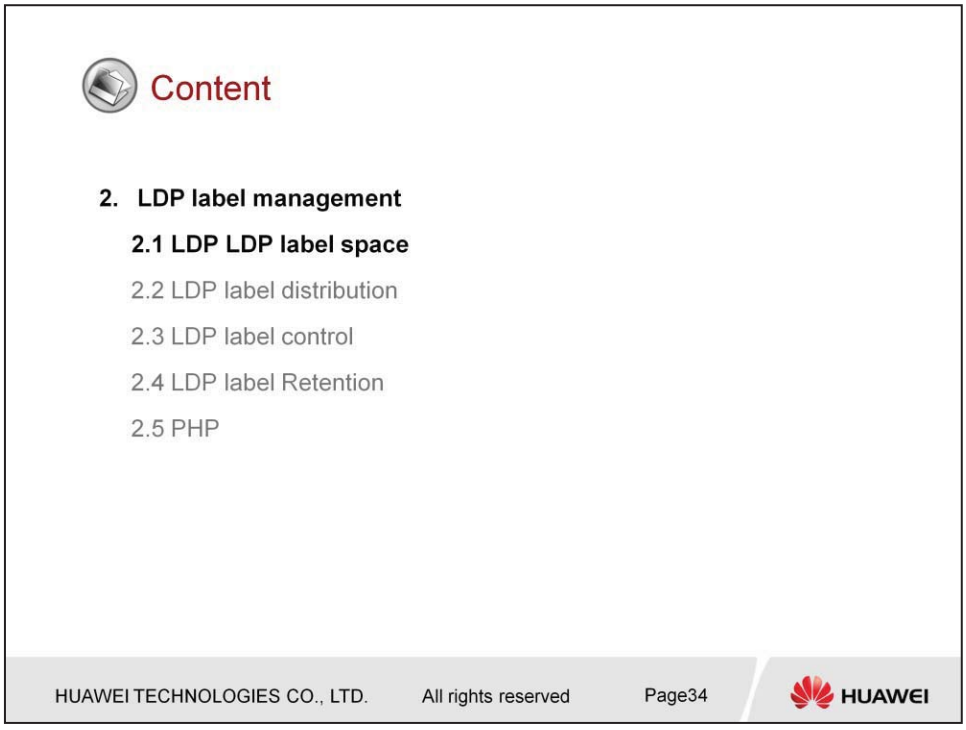
  HUAWEI

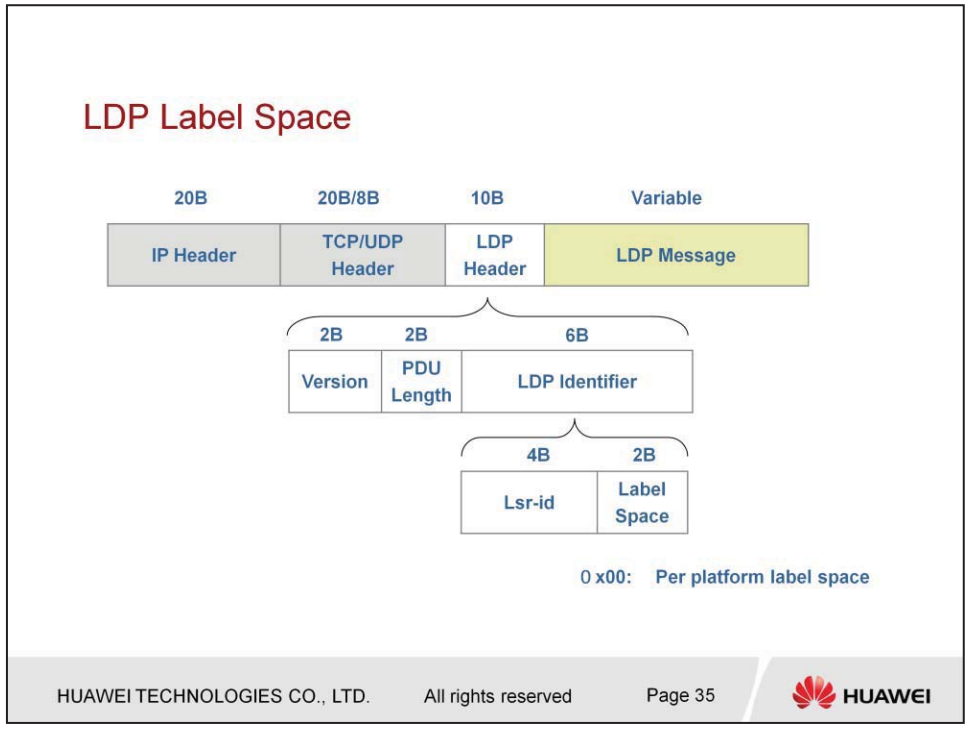
In INITIALIZED state, it sends INITIALIZATION Message and the state turns to OPENSent.



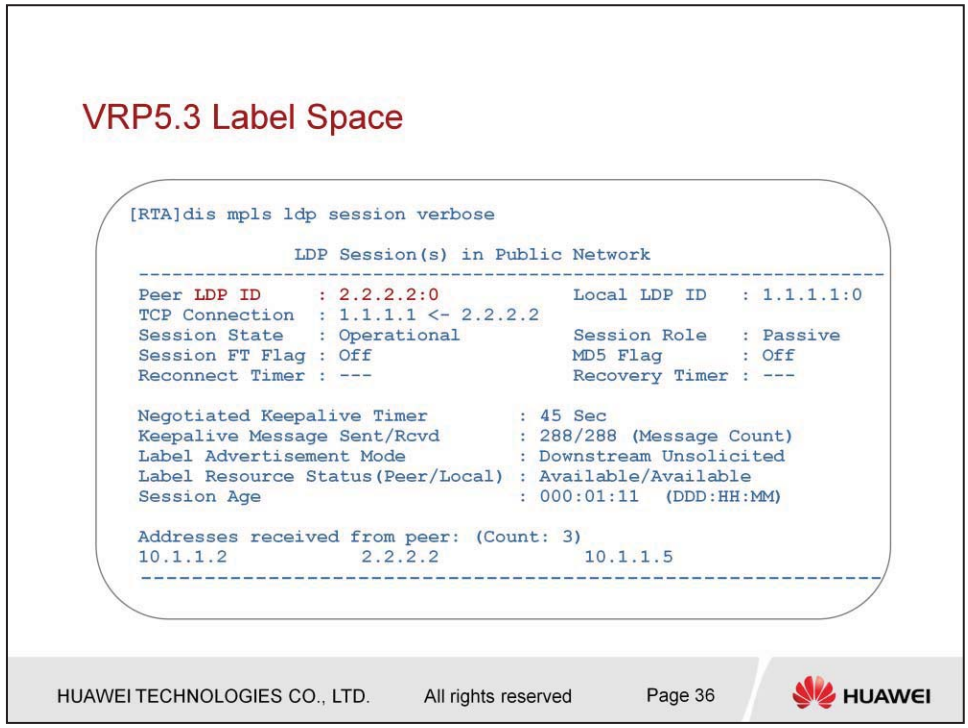
In OPENSent state, it receives INITIALIZATION Message and sends KeepAlive Message, the state turns to OPENRec. In OPENRec state, it receives KeepAlive Message, then the state turns to OPERATIONAL.


Content
LDP neighbor discovery and session establishment
LDP label management
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page33
HUAWEI

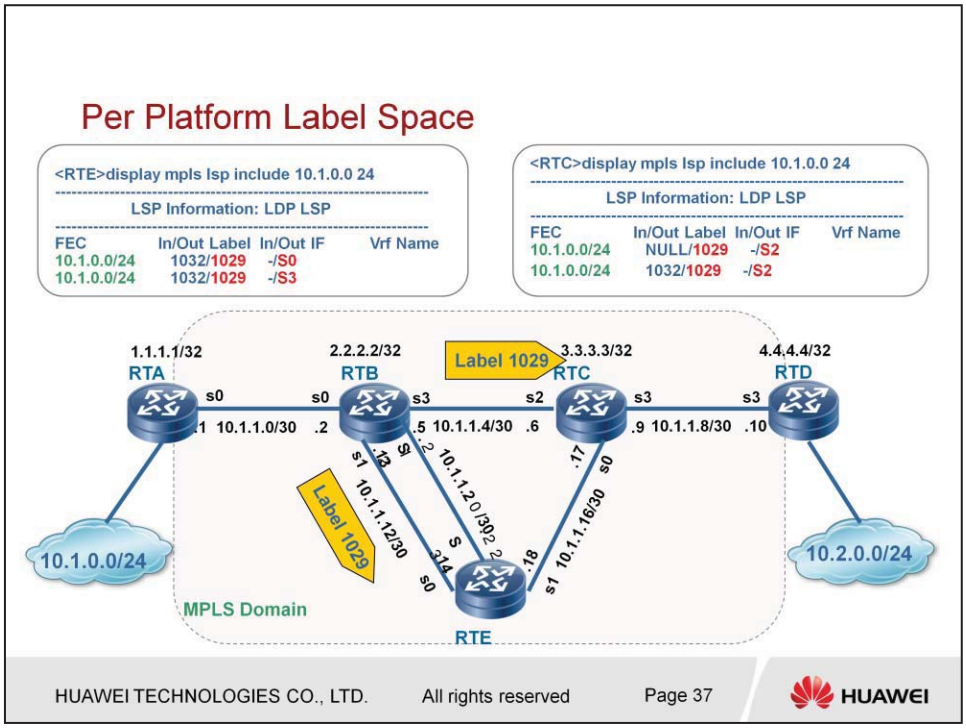

Content
2. LDP label management
2.1 LDP LDP label space
2.2 LDP label distribution
2.3 LDP label control
2.4 LDP label Retention
2.5 PHP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page34
HUAWEI



LDP Header includes 6 Bytes LDP Identifier, the first 4 Bytes identify Lsr-id, the last 2 Bytes identify a label space. There are two types of label space: platform based label space and interface-based label space. For platform-based label space, the last 2 Bytes should be zero. Frame mode encapsulation MPLS adopts platform-based label space, cell mode encapsulation MPLS adopts interface based label space.



Platform-based label space is adopted on VRP platform.



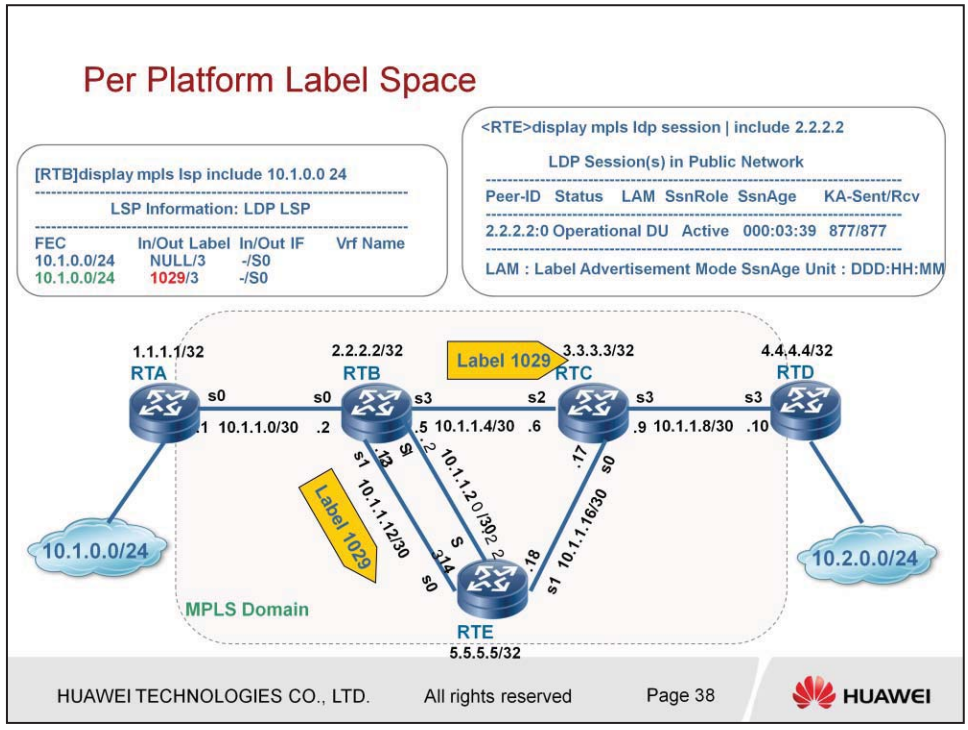
For platform-based label space, LSR only allocates one label for one destination network, and sends this label to all LDP Peers. So this label is based on platform and can be used for any incoming interface. This mode can save labels.

Frame mode MPLS adopts platform-based label space by default.

As show in the figure, RTB allocates a label 1029 to its LDP Peers, RTC and RTE.

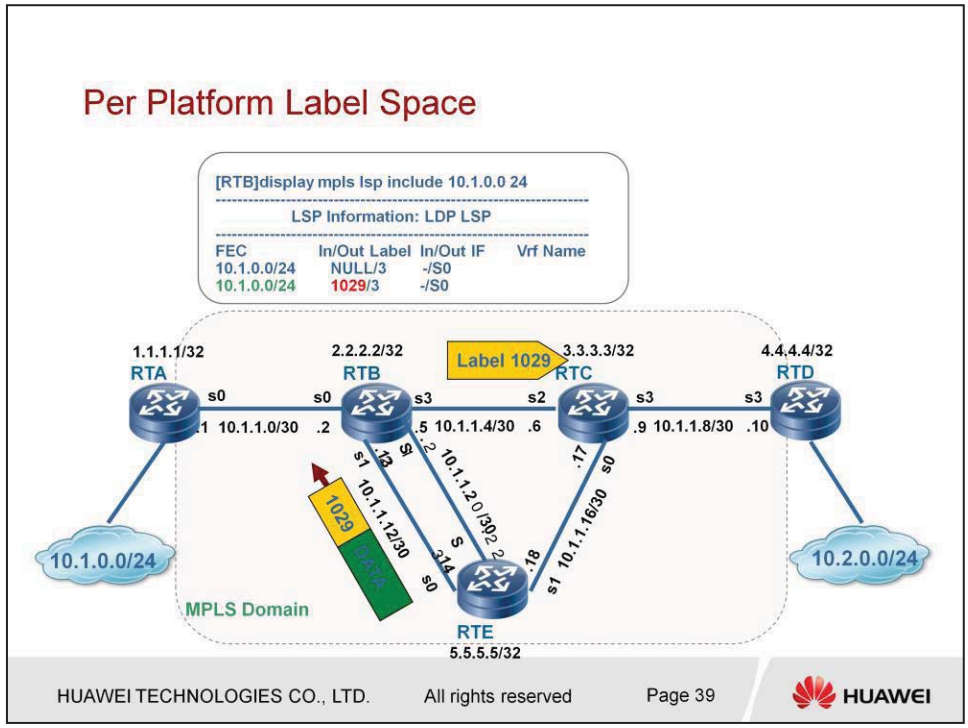
So the data packet with destination 10.1.0.0 on RTC is encapsulated with label 1029 and sent to next hop RTB via interface S2, the data packet with destination 10.1.0.0 on RTC is also encapsulated with label 1029 and sent to next hop RTB

via interface S2 and S3 (there are two links between RTE and RTB).



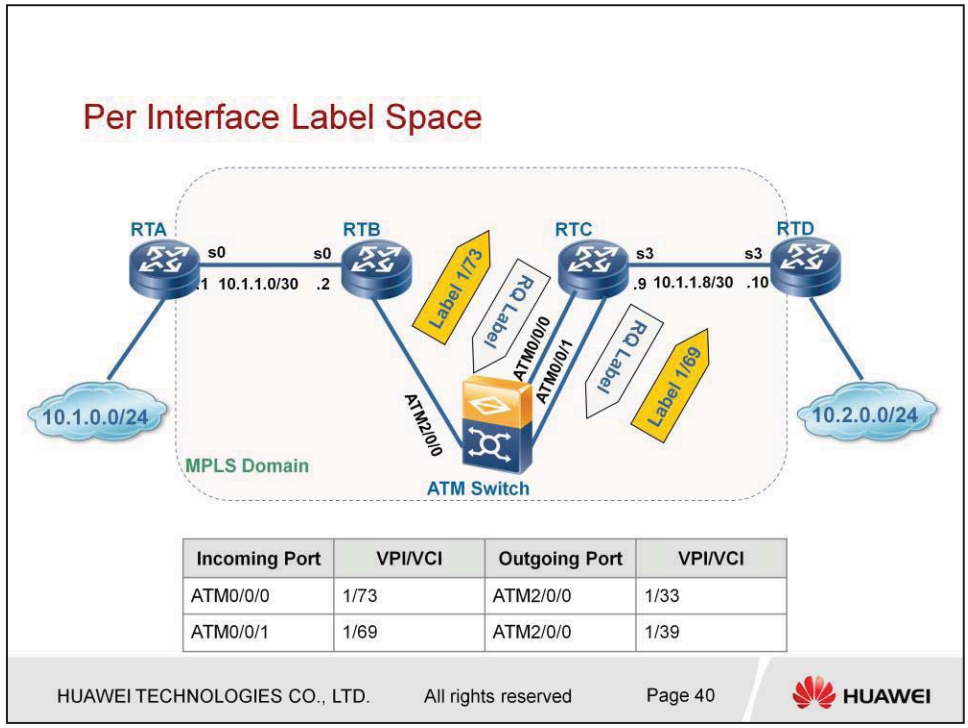
RTB forwards packets by incoming label, the packet with incoming label 1029 will be sent from interface S0.

Besides, although there are two links between RTE and RTB, LDP only establishes one Session to transmit label. So platform-based label space can minimize the amount of LDP session.



There is weakness of security of platform-based label space, assuming that RTB only sends label of 1029 to RTC but not sends to RTE, namely, RTC doesn't want to forward data packet from RTE. But if attacker counterfeits a packet with label

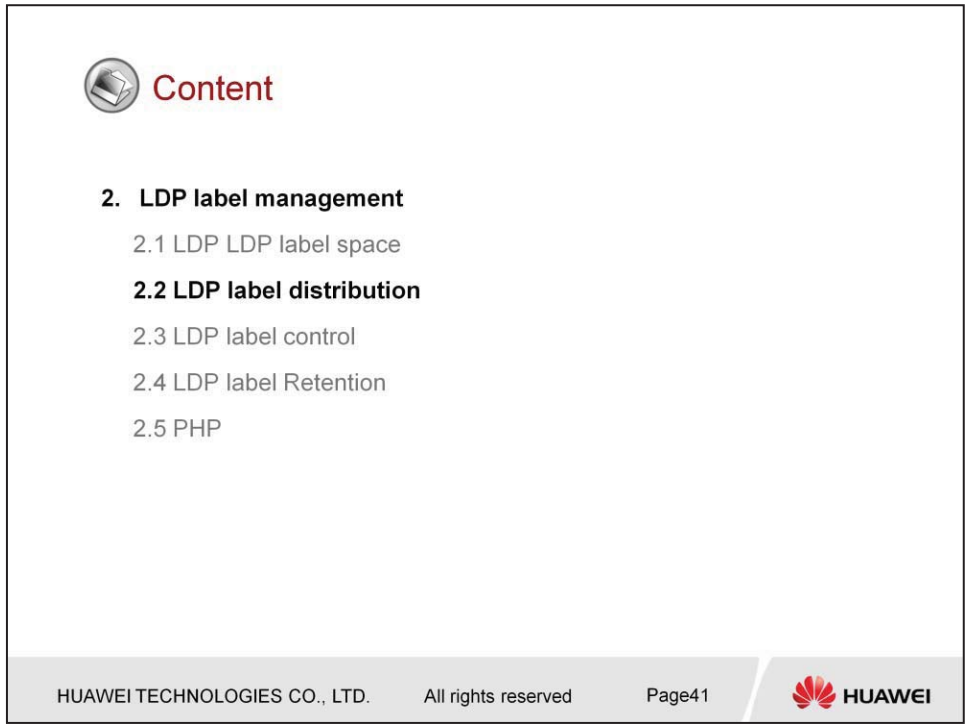
1029 on RTE and sends to RTB, RTB just forwards packets by incoming label, it doesn't check incoming interface, so RTB will also forward this lawless packet according to label forwarding table on RTB.

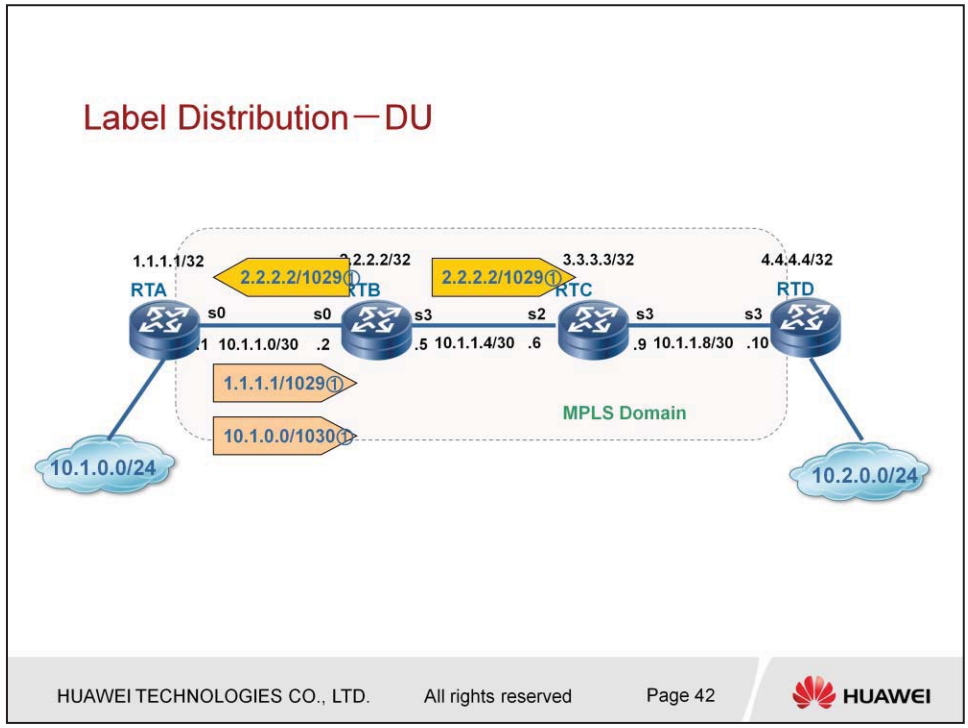


Interface-based label space allocates label for different destination IP network based on different interface, these labels are only unique in specific interface. Cell mode MPLS adopts interface-based label space by default.

It is different from platform-based label space, if there are two parallel links between a router (as RTC in the figure) and ATM switch (as ATM switch in the figure), it requires two LDP Session, RTC requests two labels towards 10.1.0.0 from ATM switch.

The table in the figure is ATM Switch label forwarding table, ATM switch adopts interface-based label space, the forwarding is according to incoming interface and incoming label, so relative to platform-based label space, interface-based label space can prevent forwarding packet with counterfeited label for security.

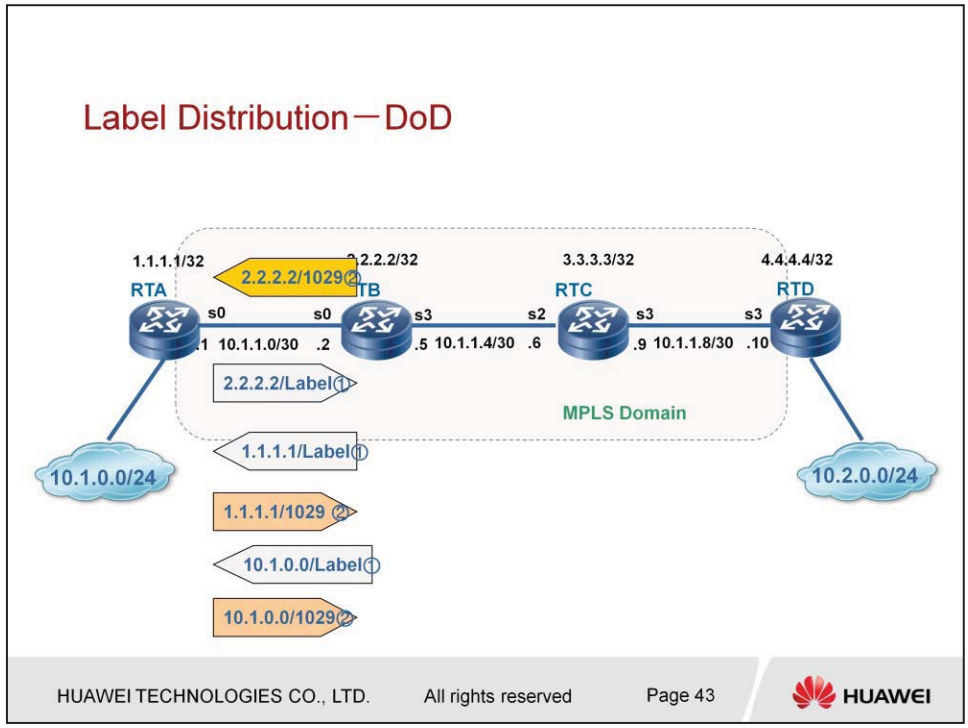

Content
2. LDP label management
2.1 LDP LDP label space
2.2 LDP label distribution
2.3 LDP label control
2.4 LDP label Retention
2.5 PHP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page41
HUAWEI



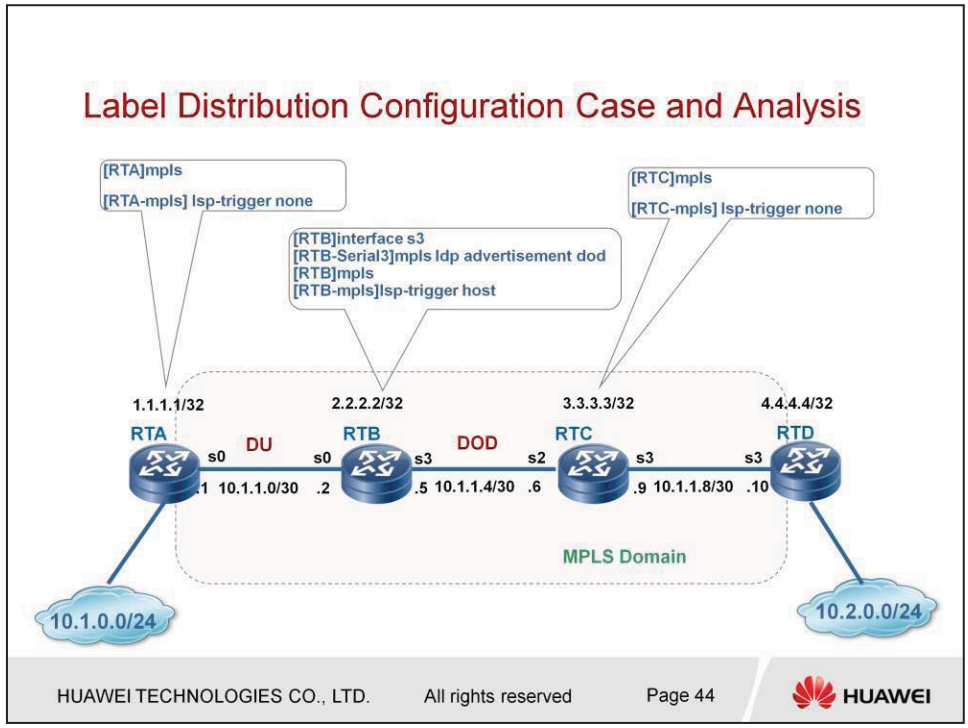
There are two label distribution modes: DU (Distribution Unsolicited) and DOD (Distribution on Demand). In DU mode, there is no request from upstream LSR, the downstream LSR will distribute label mapping message of corresponding

network (it can implement that allocating label for all the routes in the routing table, the host route in the routing table or specific IP Prefix route via configuration) to its upstream LSR. Let's simply introduce the concepts of upstream and downstream:

for 2.2.2.2/32, RTB is downstream LSR, RTA and RTC are upstream LSRs; for 1.1.1.1/32, RTA is downstream LSR, RTB is upstream LSR.



In DoD mode, it only sends Label Mapping Message to upstream router when LSR receives Label Request Message for specific route from upstream LSR. As shown in the figure, when RTB receives Label Request Message (2.2.2.2/label) from RTA, it will send Label Mapping Message to RTA (2.2.2.2/1029).



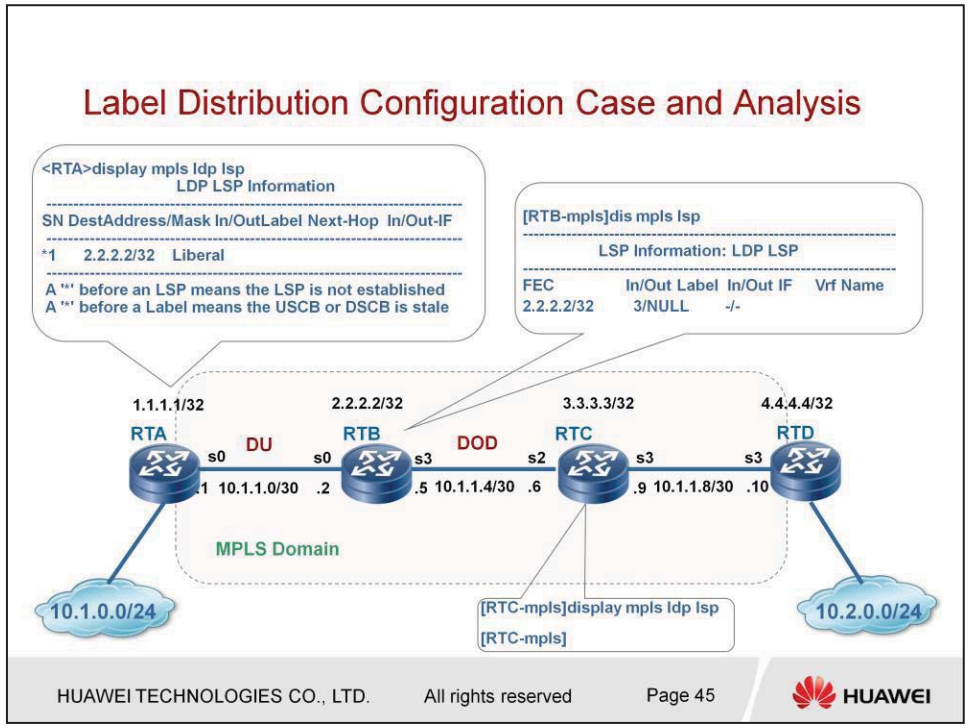
In the figure, between RTA an RTB, it adopts DU mode by default on VRP , between RTB and RTC, it is configured as DOD mode.

Configure route-triggered policy for establishing LSP and respectively observing the label distribution on RTB, RTA and RTC.

Configuration explain:

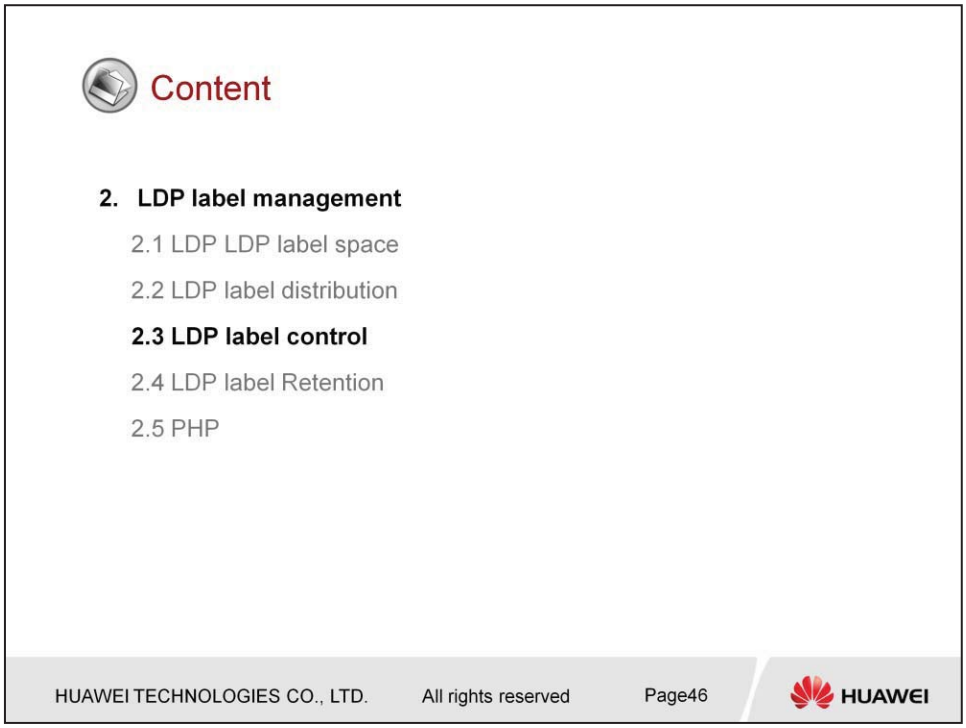
[RTB-mpls] lsp-trigger host LSP is triggered by 32-bit host IP route.

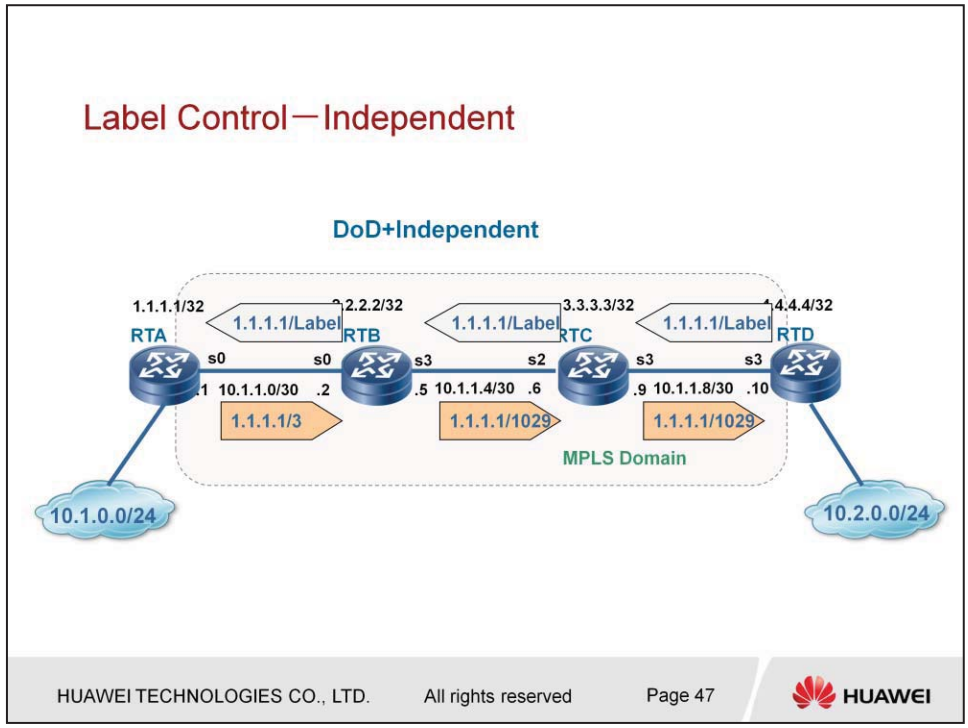
[RTC-mpls] lsp-trigger none not to trigger LDP to set up LSP.



Because It adopts DU mode to allocate label between RTB and RTA, although upstream LSR RTA doesn't request label from RTB, RTB still sends label mapping information to RTA, label allocated by RTB can be see on RTA.

However, it adopts DoD mode to allocate label between RTB and RTC, so upstream LSR RTC doesn't request label, RTB doesn't send label mapping information to RTC, there is no corresponding label mapping on RTC.

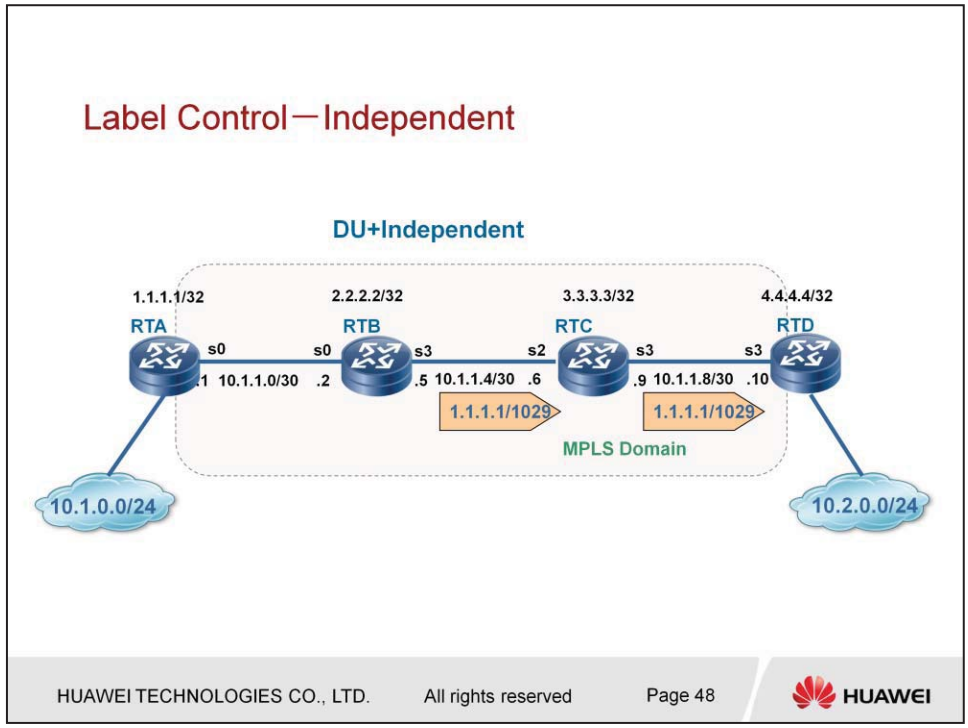

Content
2. LDP label management
2.1 LDP LDP label space
2.2 LDP label distribution
2.3 LDP label control
2.4 LDP label Retention
2.5 PHP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page46
HUAWEI



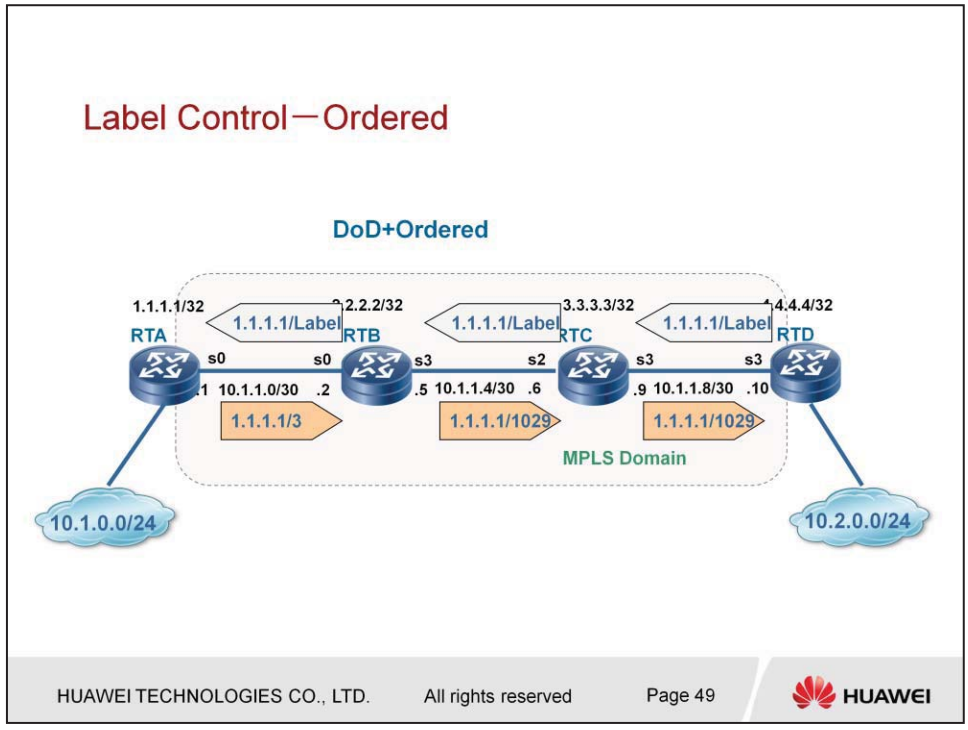
There are two label control modes: order and independent.

When LSRs adopts Independent control mode, every LSR sends Label Mapping message to peers, such as in DoD mode, when LSR (RTC in the figure) receives label request message from upstream LSR (RTD in the figure), it does not need to

wait label mapping from downstream (RTB in the figure), it can response with its own label mapping message for this label request message to upstream LSR (RTD in the figure)

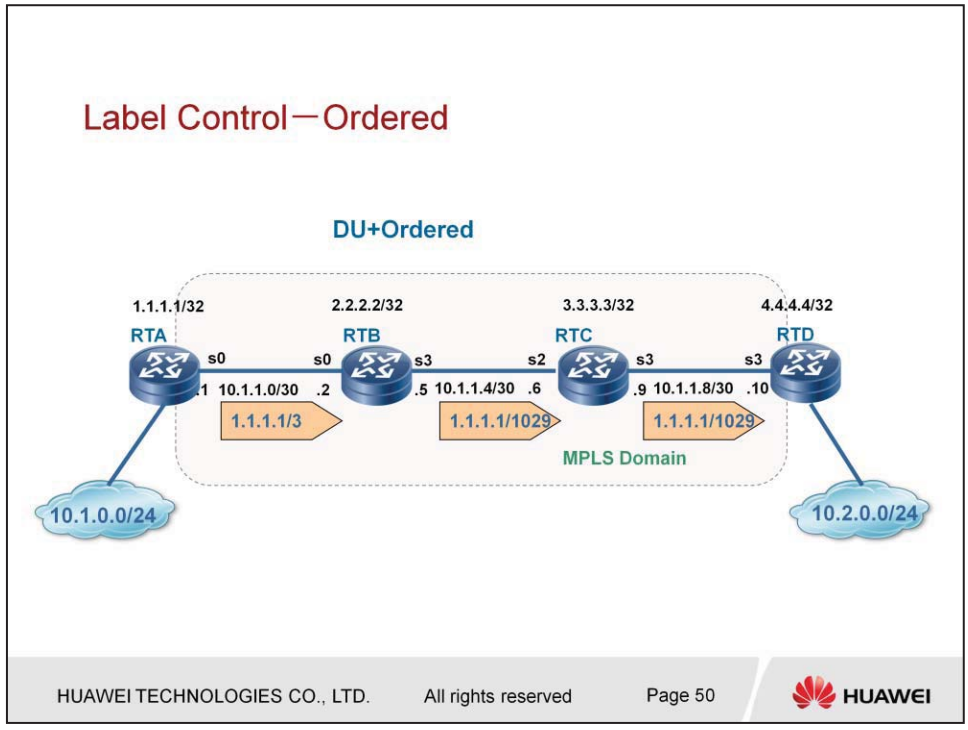


When the label distribution mode is DU, whenever, as long as LSR is ready to forward by label for corresponding FEC, it will send label mapping message to peers, as shown in the figure, RTA is configured not to trigger to establish LSP, RTB is configured that LSP is triggered by all the IP route, because of RTB adopts independent label control mode, although its downstream LSR (RTA) doesn't allocate label to it, RTB still sends label mapping message to its upstream LSR (RTD).

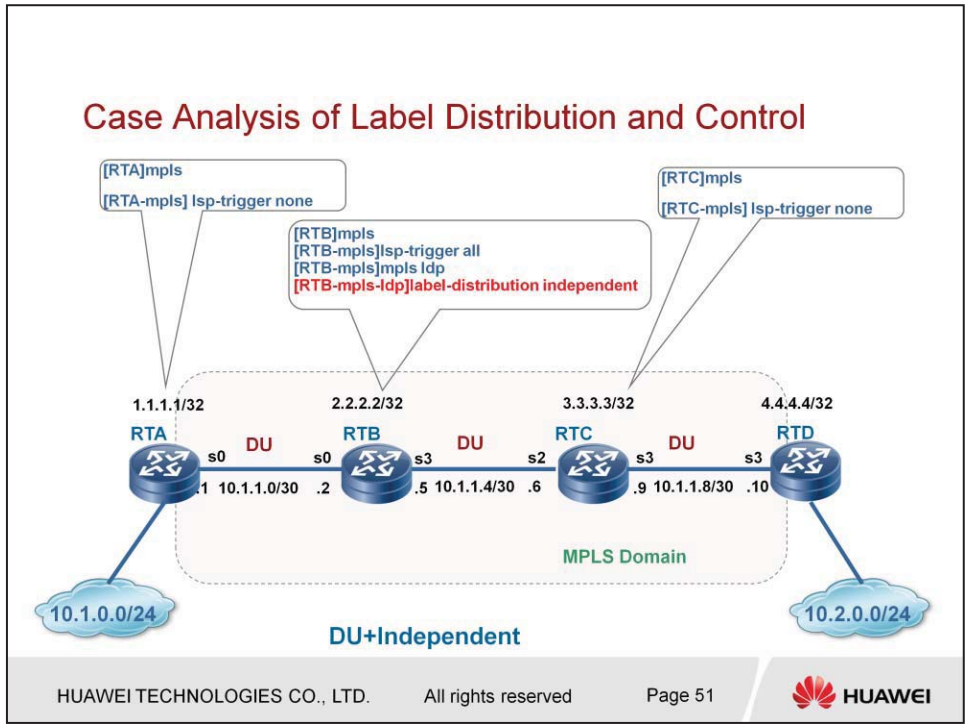


In Ordered LSP Control , an LSR only sends label mapping message to upstream if it has already received a label binding for that FEC from its next hop for that FEC, or if it is the egress LSR for that FEC. As shown in the figure, the label control mode is Ordered, label distribution is DoD, in this case, RTD request label for 1.1.1.1 from downstream LSR (particular next hop for the FEC), RTC only allocates label to RTD after it receives label mapping message from its downstream LSR RTD, so before RTC allocates label to RTD, it first sends label request message to RTB, and RTB sends label request message to RTA,

because RTA is the egress of the LDP, RTA sends label to RTB, after RTB receives label from RTA, it sends label to RTC, and after RTC receives label from RTB, it sends label to RTD.

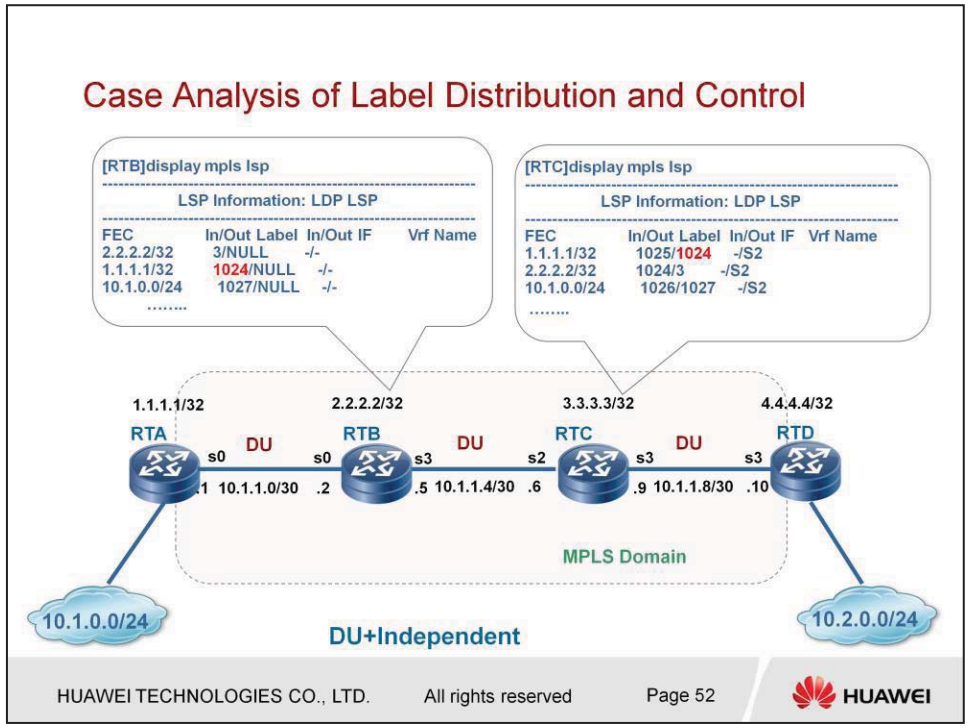


The figure shows the situation that label control mode is Ordered and label distribution mode is DU. Downstream LSR RTA sends label mapping information for 1.1.1.1 to RTB, after RTB receives label from downstream LSR RTA, it sends label to RTC, after RTC receives label from RTB, it sends label to RTD.

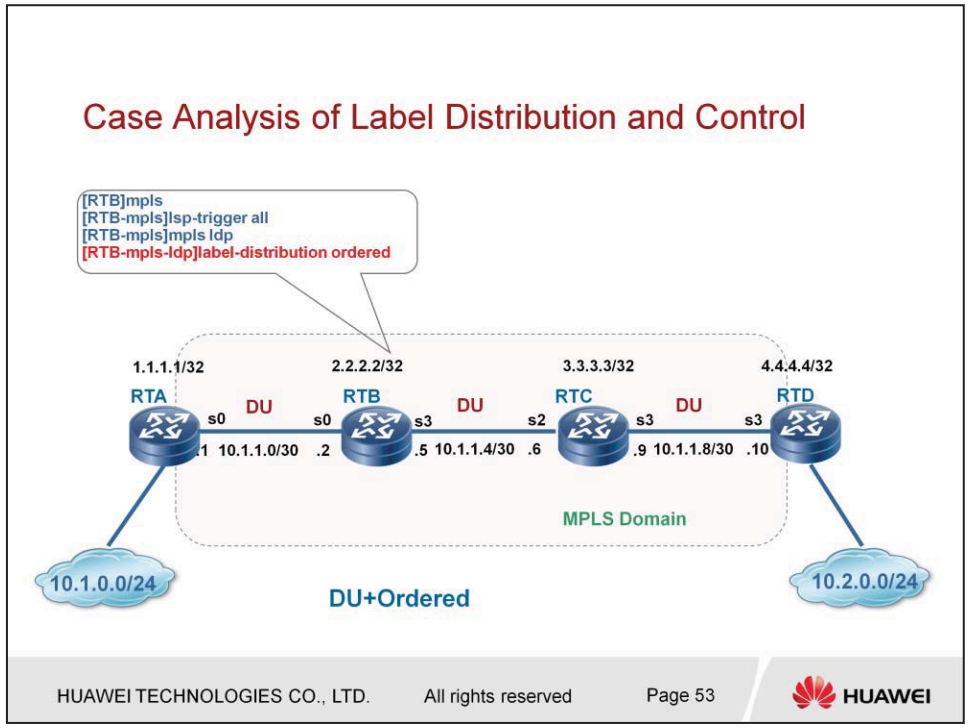


In the figure, it adopts DU distribution mode between RTA and RTB, RTB adopts independent control mode, RTB learn routes that connected to RTA via routing protocol, and RTB is configured that LSP is triggered by static route and IGP route,

RTA is configured that it is not to trigger to establish LSP. Observe the label information on RTB and RTC.



Although RTA (downstream LSR) doesn't allocates label to RTB, RTB still sends label mapping message to RTC (upstream LSR), because RTB adopts Independent control mode.

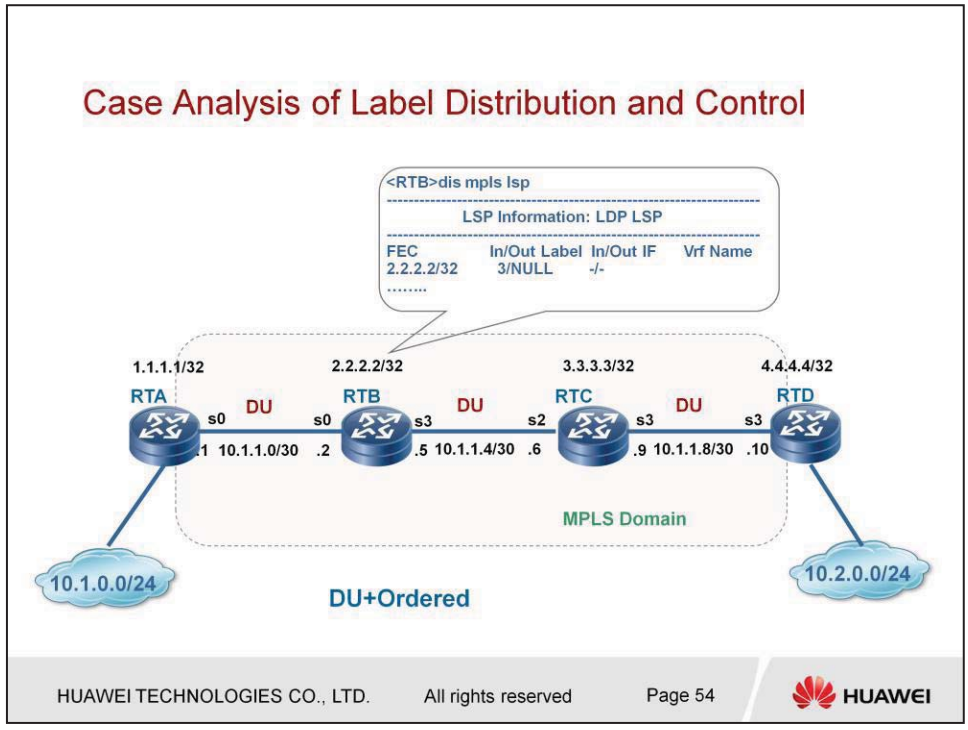


If the label control mode is changed to Ordered. Does RTB send label for 1.1.1.1/32?

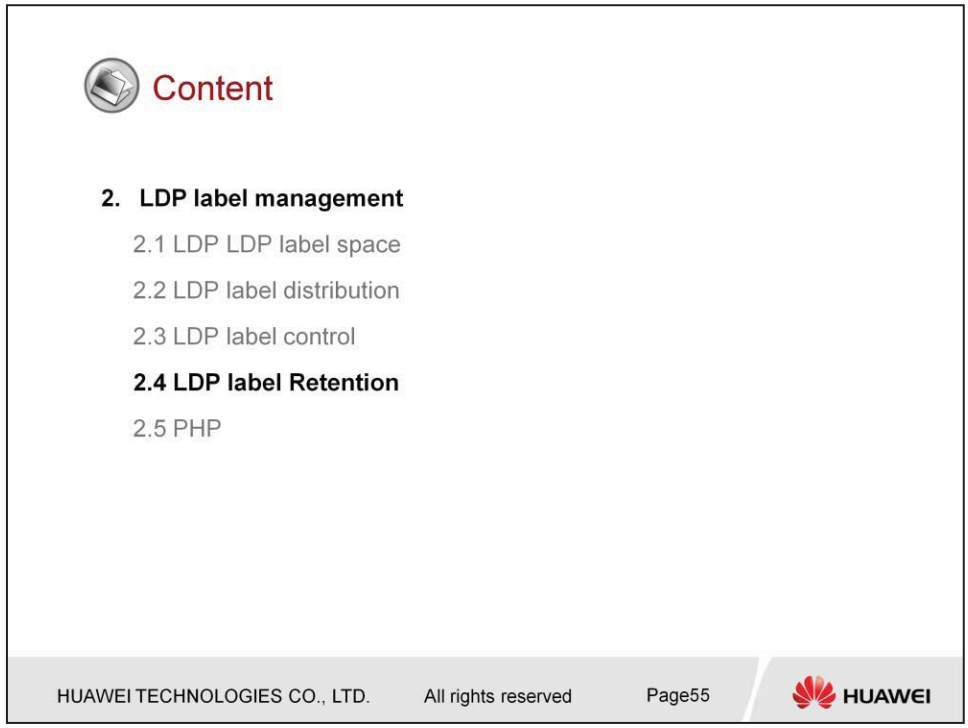
Configuration explained:

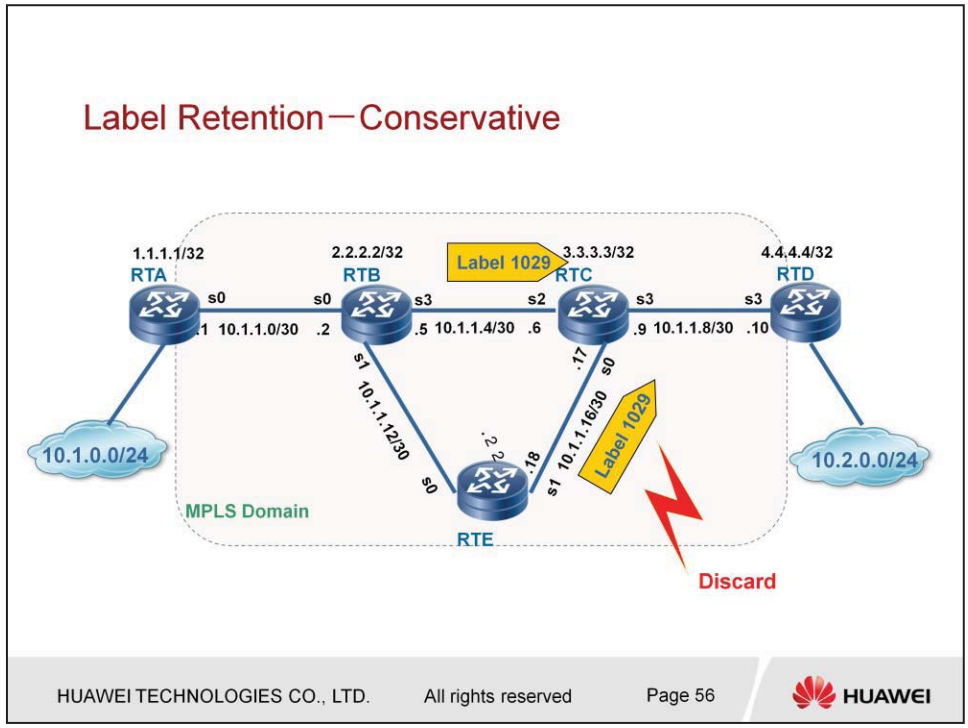
[RTB-mpls-ldp]label-distribution ordered

Label control mode is configured as ordered.



Using command “display mpls lsp” and observing LSP on RTB, after RTB is configured Ordered control mode, RTB doesn’t send label mapping message for 1.1.1.1/32 to upstream LSR RTC, namely, it isn’t to establish LSP, because RTB doesn’t receives label for 1.1.1.1/32 from downstream RTA (using command “lsptrigger none” indicates that the establishment of an LSP is not triggered for local routes on RTA).

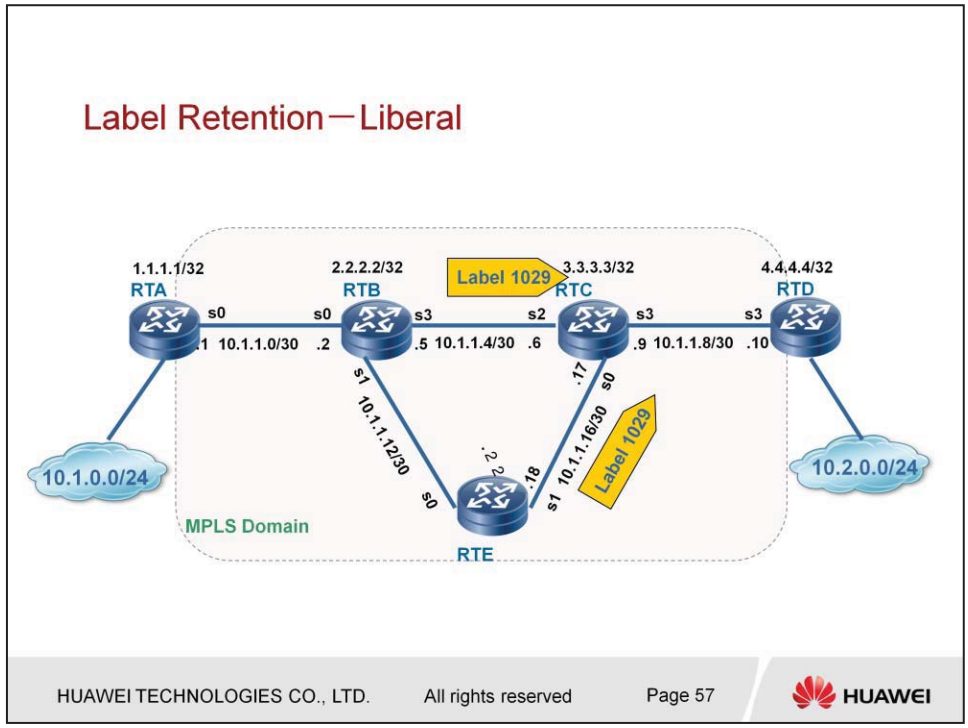

Content
2. LDP label management
2.1 LDP LDP label space
2.2 LDP label distribution
2.3 LDP label control
2.4 LDP label Retention
2.5 PHP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page55
HUAWEI



In DU distribution mode, LSR maybe receive Label Mapping Messages for same destination from multiple LDP Peers, as shown in the figure, RTC receives Label Mapping messages for 10.1.0.0/24 from RTB and RTE. If in conservative retention mode, RTC only keeps the label from next hop RTB, and the label from RTE which is not the next hop will be discarded.

In DoD distribution mode, if it adopts conservative retention mode, LSR only sends label to the next hop according to routing information.

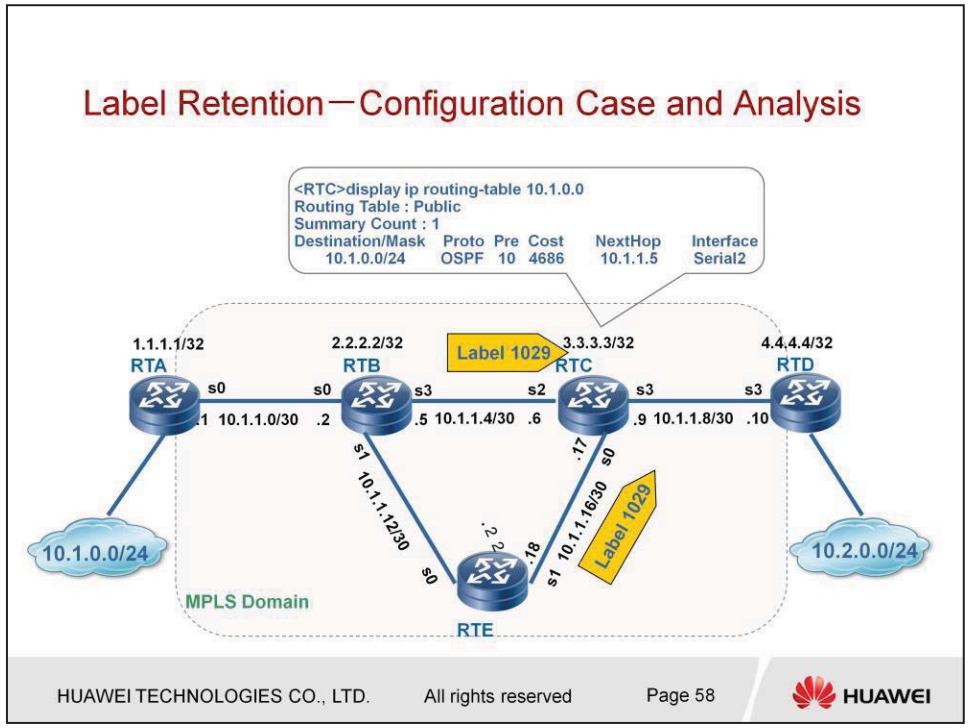
The advantage of conservative mode just needs to keep and maintain the label which is used for data forwarding, when the label space is limited, this mode is useful, such as ATM switch; but its weakness is that if the next hop towards the destination in the routing table is changed, it must obtain the label from the new next hop, then it can forward by label.



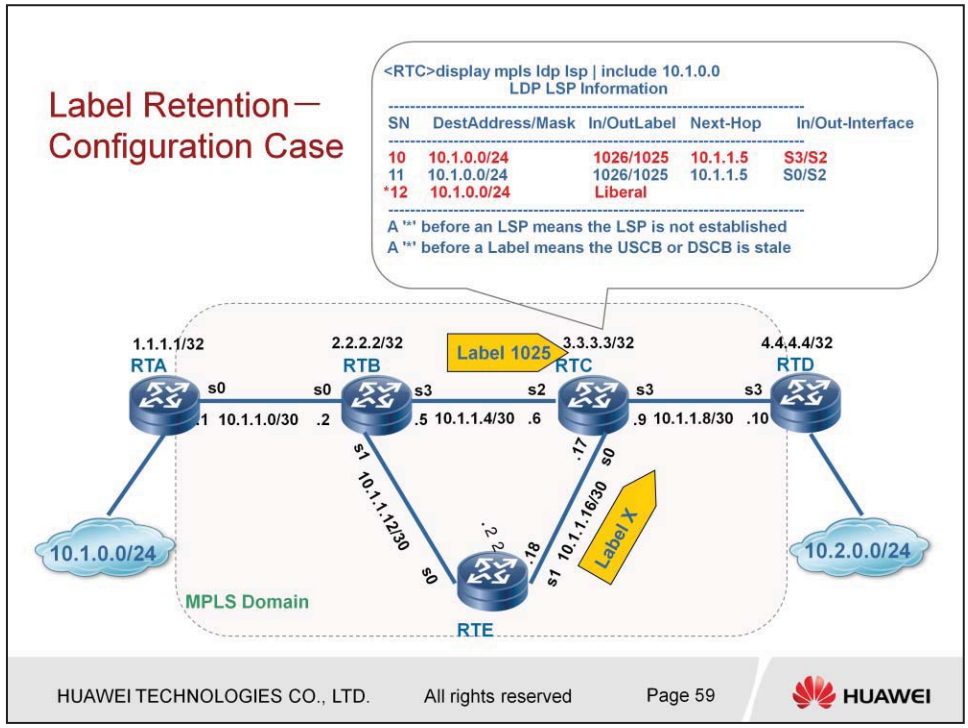
In DU distribution mode, if it adopts Liberal retention mode, RTC will keep the label from all the LDP peers RTB and RTD, no matter that the LDP Peer is the next hop towards the destination or not.

In DoD distribution mode, if it adopts Liberal retention mode, LSR will request label from all LDP peers. But generally, DoD distribution mode is always used with conservative retention mode. The most advantage of Liberal mode is that it is

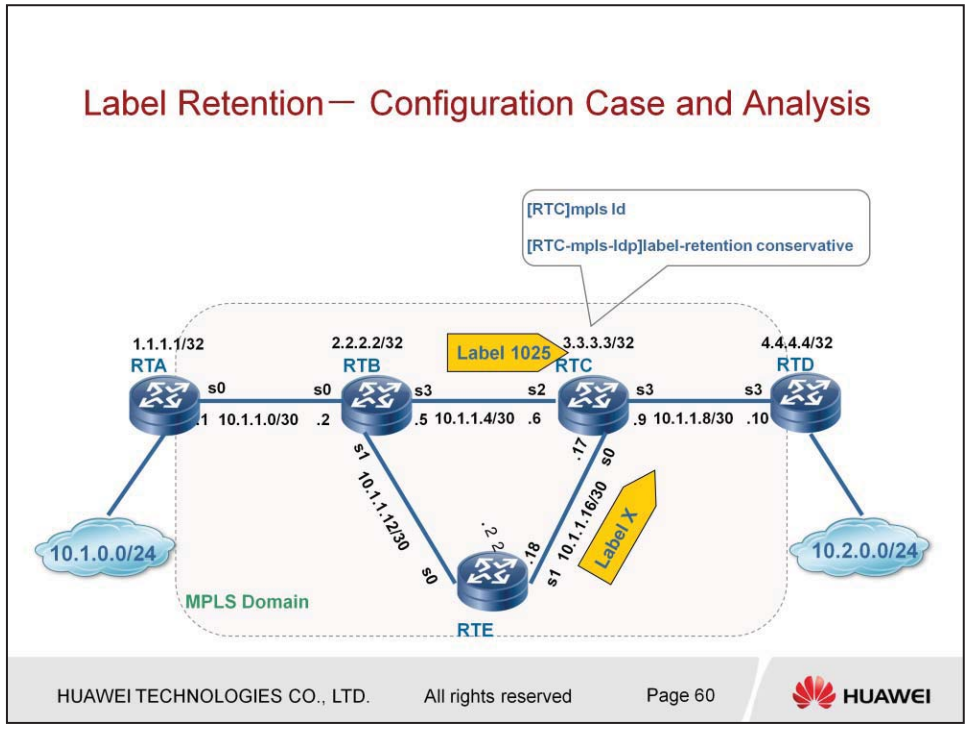
able to establish LSP to forward packet when route changed, because Liberal mode keeps all the labels. But the weakness is that it needs to allocate and maintain unnecessary label mapping information.



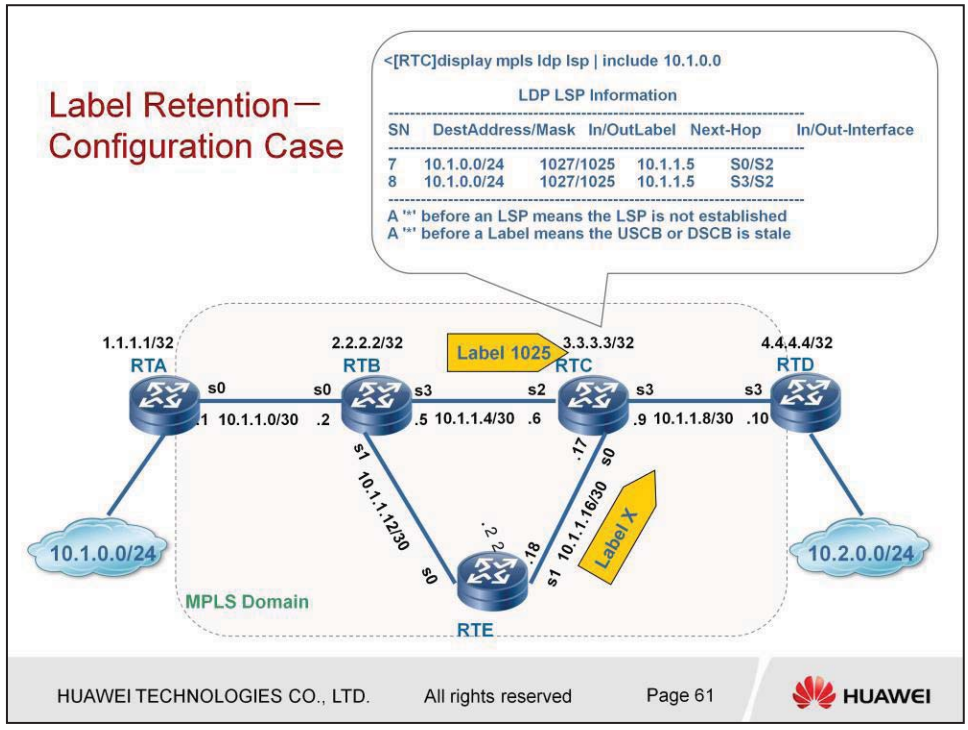
RTB is the next hop of 10.1.0.0/24 for RTC. Besides, the default Label Retention mode is Liberal on VRP platform.



RTC keeps the labels from RTB and RTE.



Configure that RTC adopts conservative retention mode.

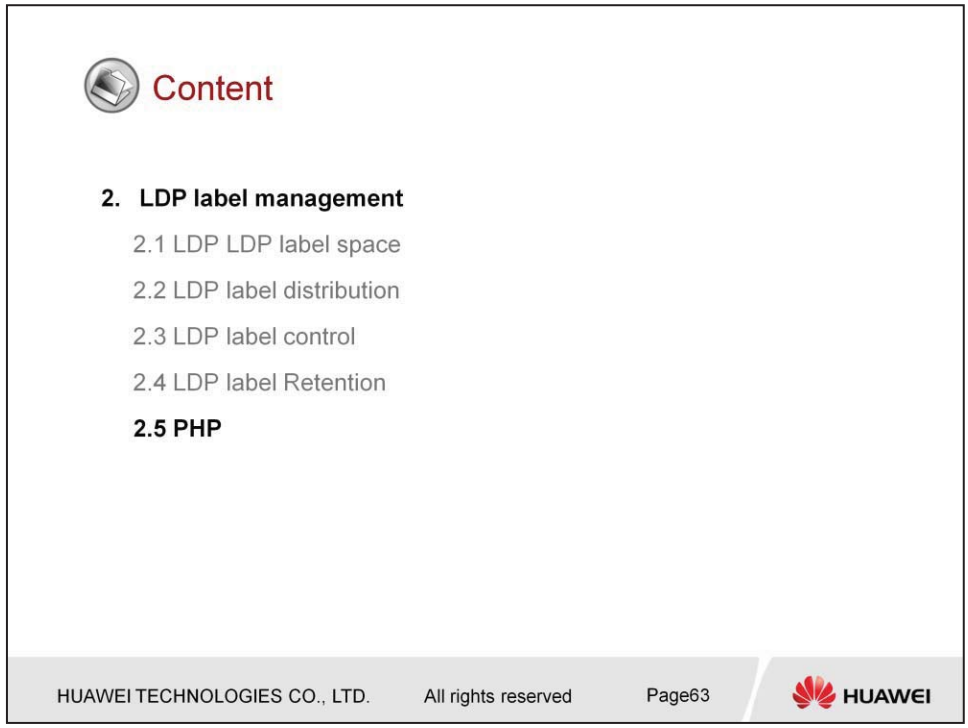


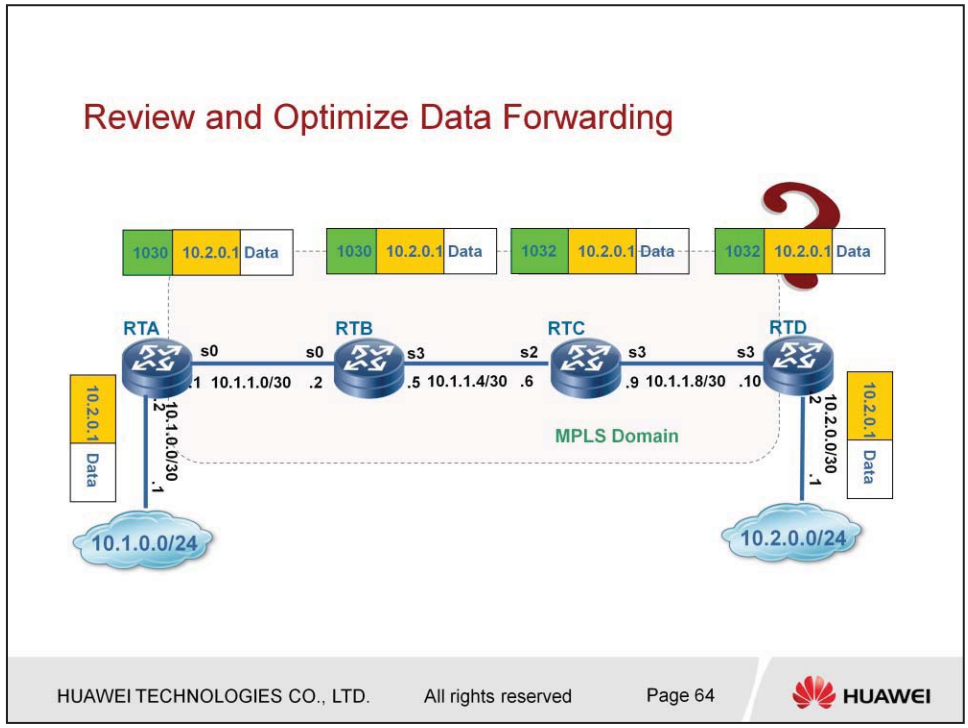
RTC only keeps the label from RTB.

VRP5.30 Recommended Combination

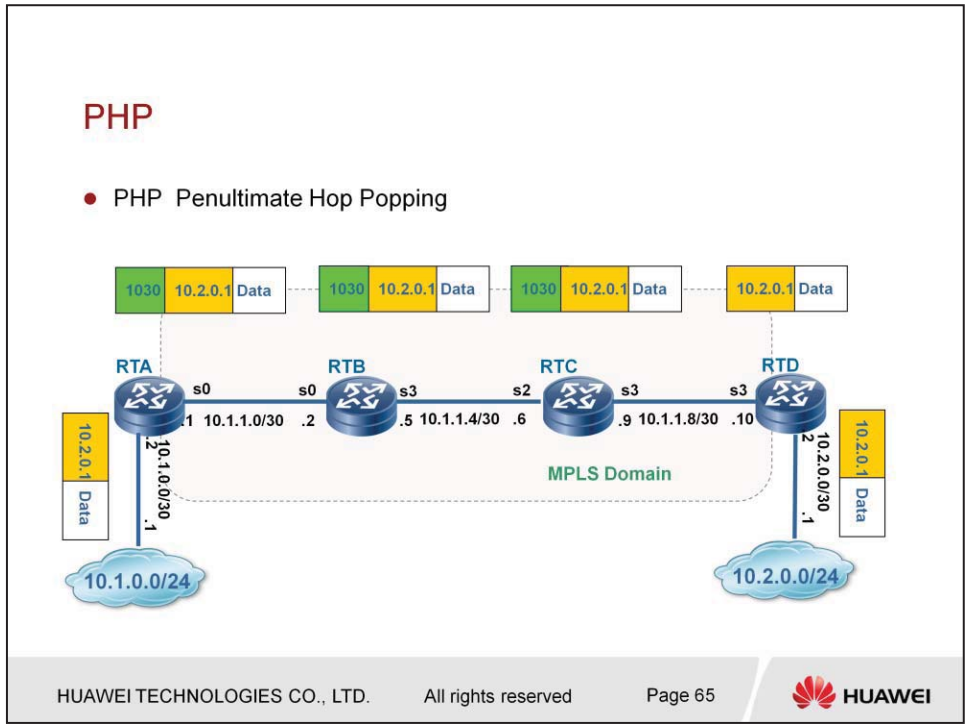
```
[RTC]display mpls ldp
                        LDP Global Information
------------------------------------------------------------------------------
Protocol Version         : V1            Neighbor Liveness     : 600 Sec
Graceful Restart         : Off           FT Reconnect Timer    : 300 Sec
MTU Signaling            : On            Recovery Timer        : 300 Sec
                         LDP Instance Information
------------------------------------------------------------------------------
Instance ID              : 0             VPN-Instance          :
Instance Status          : Active        LSR ID                : 3.3.3.3
Hop Count Limit          : 32            Path Vector Limit     : 32
Loop Detection           : Off
DU Re-advertise Timer    : 10 Sec        DU Re-advertise Flag  : On
DU Explicit Request      : Off           Request Retry Flag    : On
Label Distribution Mode  : Ordered       Label Retention Mode  : Liberal
------------------------------------------------------------------------------
[RTC]display mpls ldp session
      LDP Session(s) in Public Network
------------------------------------------------------------------------------
Peer-ID            Status      LAM  SsnRole  SsnAge       KA-Sent/Rcv
------------------------------------------------------------------------------
2.2.2.2:0          Operational DU   Active   000:00:10    44/44
------------------------------------------------------------------------------
LAM : Label Advertisement Mode      SsnAge Unit : DDD:HH:MM
```


The recommended combination is DU+Ordered+Liberal, it is also the default configuration on VRP platform.


Content
2. LDP label management
2.1 LDP LDP label space
2.2 LDP label distribution
2.3 LDP label control
2.4 LDP label Retention
2.5 PHP
HUAWEI TECHNOLOGIES CO., LTD.
All rights reserved
Page63
HUAWEI

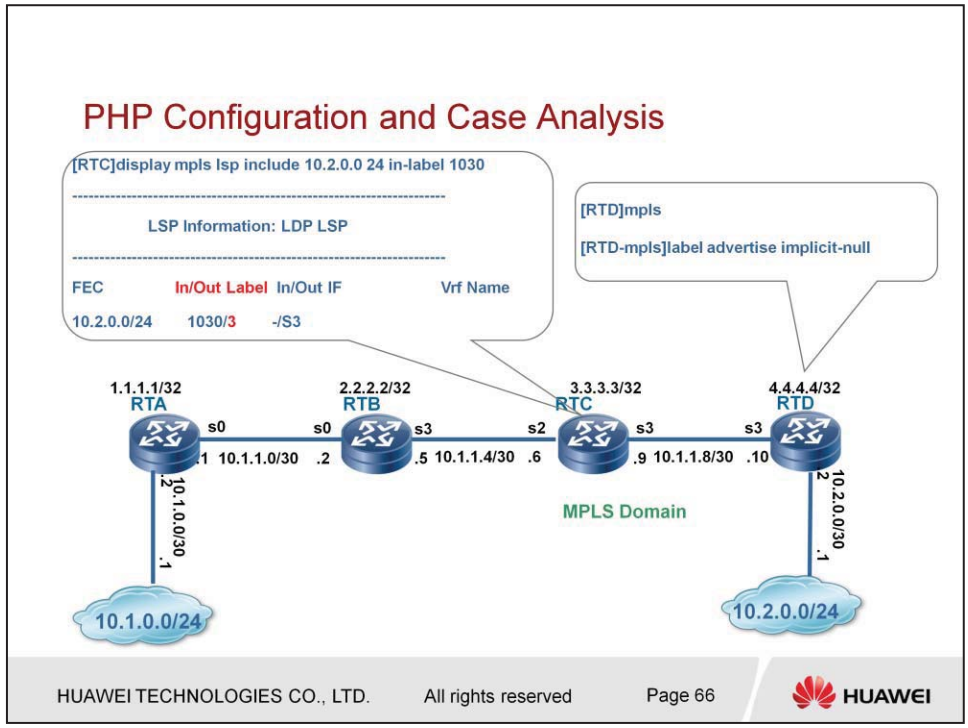


After LDP completes label exchanging and establishes LSP, the data packet will be forwarded along the LSP. Let's review the process of label forwarding. In this process, when Egress LER RTD receives packet with label 1032, first pops label, then lookups routing table according to destination IP address and implement traditional IP forwarding. Actually, for Egress LER, label 1032 has no meaning. If egress LER doesn't analyze the label and pop label, just implements IP forwarding, thus the efficiency will be improved. So penultimate LSR RTC can pop label and send IP packet to Egress LER RTD, thus Egress LER RTD doesn't operate the label, just transmits IP packet to corresponding destination network. It can reduce the burden of the last hop.



PHP (Penultimate Hop Popping), it makes label pop on the penultimate LSR.

When it adopts penultimate hop popping, penultimate LSR determines where the packet transmit to according to the label from upstream LSR, then pops the label and forwards it, when the last hop LSR (Egress LER) receives this packet, it is the traditional IP packet, it can implement traditional IP forwarding. However, how does LSR know it is the penultimate hop? The last hop LSR will allocate a special label 3.



Three different label distribution modes can be configured on LER on VRP platform to inform whether the penultimate hop LSR should pop the label or not.

[RTD-mpls]label advertise ?

explicit-null explicit-null

implicit-null implicit-null

non-null non-null

explicit-null means explicit null label, the value is 0. This value is valid only at the bottom of label stack, it indicates that the label stack must be popped and forwarded by the IPv4 header.

Label value 3 indicates implicit null label, this value can be placed in label stack.

When LSR (the penultimate hop) is allocated a implicit null label, it only needs to pop the label, but can’t use this value to replace the original label.

Non-null means that it doesn’t adopt PHP. The Egress node allocates a normal label to the penultimate hop.

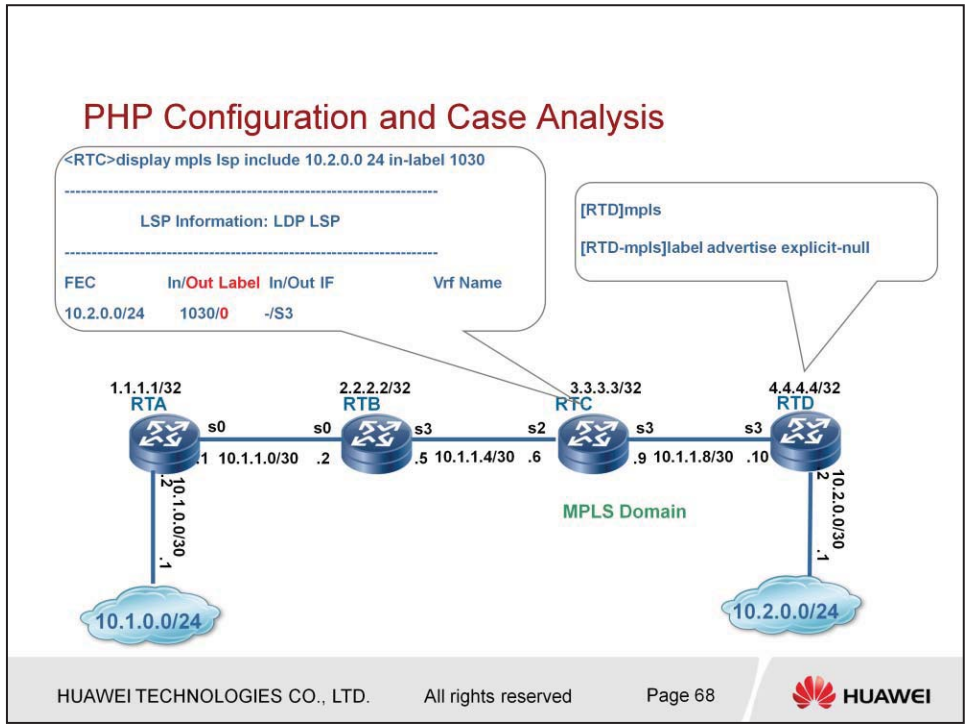
Configuration explain:

[RTD-mpls]label advertise implicit-null

Configure that the Egress node allocates a implicit label to the penultimate hop.

this is the default configuration on VRP platform.

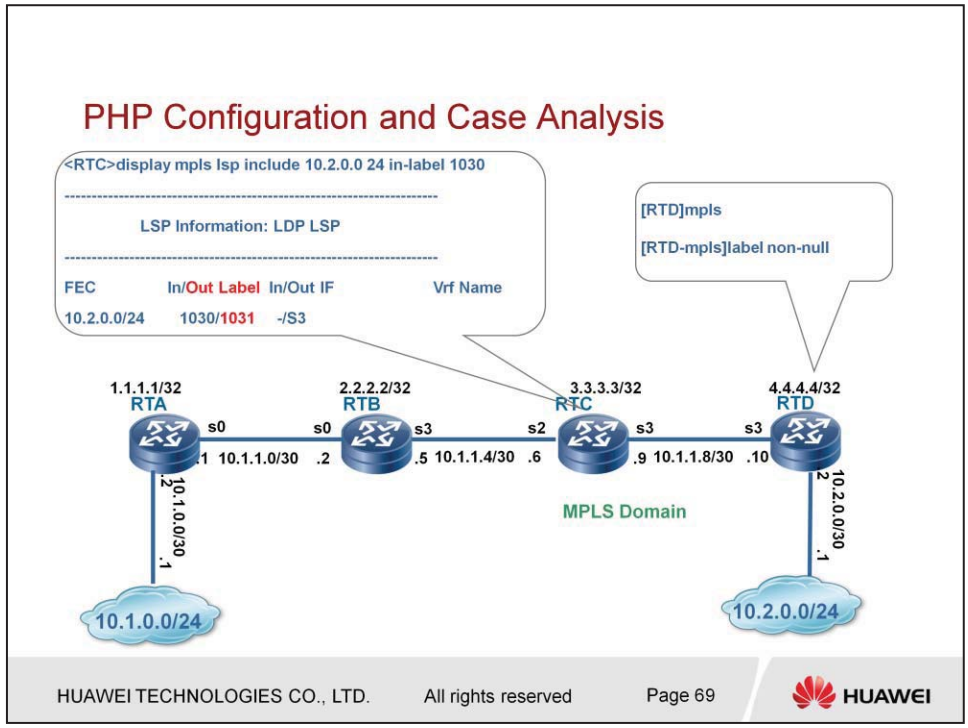
RTD allocates a implicit label 3 to RTC, When RTD receives packet with incoming label 1030, it will pop the label and forward IP packet to RTC.



Configuration explain:

[RTD-mpls]label advertise explicit-null

Configure that Egress node allocates explicit-null label to penultimate hop. RTD allocates explicit-null label valued 0 to RTC, when RTD receives packet with incoming label 0, it pops label and forwards by IP header, because label 0 only appears at the bottom of label stack. if it is other normal label, it will also judge whether it is the last label, if it is not, it will implement MPLS forwarding by in-label.

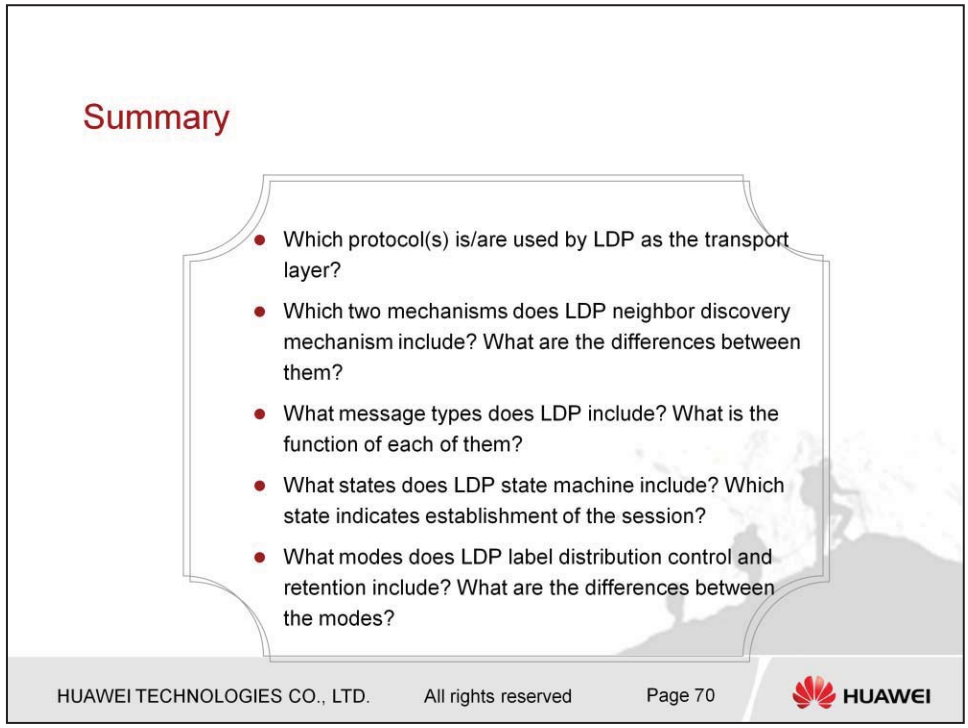


Configuration explain:

[RTD-mpls]label advertise non-null

It means that RTD doesn't adopt PHP, Egress node allocates normal label to penultimate hop.

The figure shows that RTD allocates the label 1031 to RTC.



Q: Which protocol(s) is/are used by LDP as the transport layer?

A: UDP and TCP

Q: Which two mechanisms does LDP neighbor discovery mechanism include?

What are the differences between them?

A: Basic discovery mechanism and extended discovery mechanism, Basic Discovery Mechanism is used to discover LSR neighbors that are directly connected at the link level. Extended Discovery Mechanism is used to discover

LSR neighbors that are not directly connected at the link level.

Q: What message types does LDP include? What is the function of each of them?

A: Discovery message: announce and maintain the presence of an LSR in a network.

Session message: establish, maintain, and terminate sessions between LDP peers.

Advertisement message: create, change, and delete label mapping for FECs.

Notification message: announce advisory information and error

information.

Q: What states does LDP state machine include? Which state indicates establishment of the session?

A: There are five states: NON-EXISTENT， INITIALIZED，OPENREC，

OPENSENT，OPERATIONAL. OPERATIONAL indicates that session has been established.

Q: What modes does LDP label distribution control and retention include? What are the differences between the modes?

A: There are two label distribution modes: DU (Distribution Unsolicited) and DoD (Distribution on Demand). In DU mode, no request message from upstream LSR, downstream LSR actively sends label mapping message for corresponding network to upstream LSR; in DoD mode, only the downstream LSR receives label for specific network, it sends label mapping message to upstream LSR.

There are two label control modes: Order and Independent. When it adopts Independent control mode, every LSR sends label mapping message to its adjacency neighbor. When it adopts Ordered control mode, LSR only sends label mapping message to upstream if it has already received a label binding for that FEC from its next hop for that FEC, or if it is the egress LSR for that FEC.

There are two label retention modes: Conservative and Liberal. When it adopts Conservative mode, it only keeps the label from next hop, discards the label from not next hop; when it adopts Liberal mode, it keeps all the label from LDP Peers.