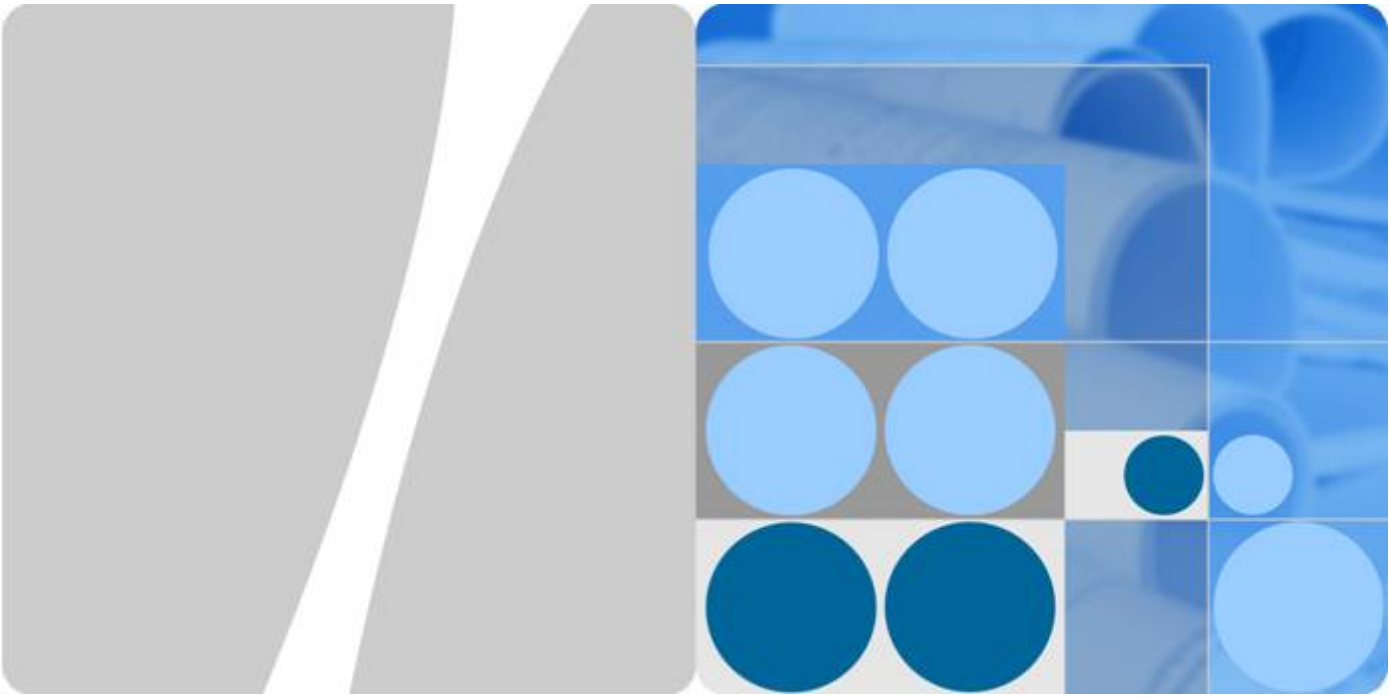


S7700 Smart Routing Switch

V200R001C00

Feature Description - Ethernet

Issue 05

Date 2013-04-10

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://enterprise.huawei.com

About This Document

Intended Audience

This document describes the Ethernet feature in terms of its overview, principle, and applications.

This document together with other types of document helps intended readers get a deep understanding of the Ethernet feature.

This document is intended for:

- Network planning engineers
- Commissioning engineers
- Data configuration engineers
- System maintenance engineers

Symbol Conventions

The symbols that may be found in this document are defined as follows.

Symbol	Description
DANGER	Indicates a hazard with a high level of risk, which if not avoided, will result in death or serious injury.
WARNING	Indicates a hazard with a medium or low level of risk, which if not avoided, could result in minor or moderate injury.
CAUTION	Indicates a potentially hazardous situation, which if not avoided, could result in equipment damage, data loss, performance degradation, or unexpected results.
TIP	Indicates a tip that may help you solve a problem or save time.
NOTE	Provides additional information to emphasize or supplement important points of the main text.

Command Conventions

The command conventions that may be found in this document are defined as follows.

Convention	Description
Boldface	The keywords of a command line are in **boldface**.
Italic	Command arguments are in *italics*.
[ ]	Items (keywords or arguments) in brackets [ ] are optional.
{ x | y | ... }	Optional items are grouped in braces and separated by vertical bars. One item is selected.
[x | y | ...]	Optional items are grouped in brackets and separated by vertical bars. One item is selected or no item is selected.
{ x | y | ... }*	Optional items are grouped in braces and separated by vertical bars. A minimum of one item or a maximum of all items can be selected.
[x | y | ...]*	Optional items are grouped in brackets and separated by vertical bars. Several items or no item can be selected.
&<1-n>	The parameter before the & sign can be repeated 1 to n times.
#	A line starting with the # sign is comments.

Change History

Updates between document issues are cumulative. Therefore, the latest document issue contains all updates made in previous issues.

Changes in Issue 05 (2013-04-10)

The fifth commercial release has the following updates:

- Some contents are modified according to updates in the product such as features.

Changes in Issue 04 (2012-10-20)

The fourth commercial release has the following updates:

- Some contents are modified according to updates in the product such as features.

Changes in Issue 03 (2012-07-03)

The third commercial release has the following updates:

- Some contents are modified according to updates in the product such as features.
- Some figures are optimized.

Changes in Issue 02 (2012-05-23)

The second commercial release has the following updates:

- Some contents are modified according to updates in the product such as features.
- Some figures are optimized.

Changes in Issue 01 (2012-03-15)

Initial commercial release.

Contents

1 Ethernet

About This Chapter

1.1 Introduction to Ethernet

Definition

The Ethernet technology originates from an experimental network with the purpose of connecting multiple PCs at the speed of 3 Mbit/s. In general, Ethernet refers to a standard for 10 Mbit/s Ethernet networks. The Digital Equipment Corporation (DEC), Intel, and Xerox (DIX) joined efforts to develop and then issued the standard in 1982. The IEEE 802.3 standard is developed on the basis of the Ethernet standard, and is compatible with it.

In TCP/IP, the encapsulation format of IP packets of the Ethernet is defined in RFC 894, and that of the IEEE 802.3 network is defined in RFC 1042. Currently, the most commonly-used encapsulation format is that defined in RFC 894, which is called Ethernet_II or Ethernet DIX.

NOTE

To distinguish Ethernet frames of those two types, in this document, Ethernet frames defined in RFC 894 are called Ethernet_II frames; Ethernet frames defined in RFC 1042 are called IEEE 802.3 frames.

Purpose

Ethernet is a universal communication protocol standard used for local area networks (LANs). This standard defines the cable type and signal processing method used for LANs.

Ethernet networks are broadcast networks established based on the Carrier Sense Multiple Access with Collision Detection (CSMA/CD) mechanism. Collisions restrict Ethernet performance. Early Ethernet devices such as hubs work at the physical layer, and cannot confine collisions to a particular scope. This restricts network performance improvement. Working at the data link layer, switches are able to confine collisions to a particular scope. Therefore, switches help improve Ethernet performance and gradually replace hubs to become mainstream Ethernet devices. Switches, however, do not restrict broadcast traffic on the Ethernet. This affects Ethernet performance. To resolve this problem, divide a LAN into virtual local area networks (VLANs) on switches or use Layer 3 switches.

As a simple, cheap, and easy-to-implement LAN technology, Ethernet has become the mainstream in the industry. The development of Fast Ethernet (FE) and Gigabit Ethernet (GE), which provide higher Ethernet performance, helps Ethernet become the most promising network technology.

1.2 References

The following table lists the references of this document.

Document	Description	Remarks
IEEE 802.3	Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications	-
IEEE 802.3ae	Media Access Control (MAC) Parameters, Physical Layers, and Management parameters for 10Gb/s Operation	-

Document	Description	Remarks
RFC 894	A Standard for the Transmission of IP Datagrams over Ethernet Networks	-
RFC 1042	A Standard for the Transmission of IP Datagrams over IEEE 802 Networks	-

1.3 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

1.4 Principles

1.4.1 Physical Layer of the Ethernet

Introduction to Ethernet Cable Standards

Currently, the well-developed Ethernet cabling standards are as follows:

- 10BASE-2
- 10BASE-5
- 10BASE-T
- 10BASE-F
- 100BASE-T4
- 100BASE-TX
- 100BASE-FX
- 1000BASE-SX
- 1000BASE-LX
- 1000BASE-CX
- 1000BASE-TX

In the preceding cabling standards, 10, 100, and 1000 stand for the transmission rate (the unit is Mbit/s), and BASE represents baseband.

- 10M Ethernet cable standards

  Table 1-1 shows the 10M Ethernet cabling standard defined in IEEE 802.3.

  Table 1-1 10M Ethernet cable standards

Name	Cable	Maximum Transmission Distance
10BASE-5	Thick coaxial cable	500 m
10BASE-2	Thin coaxial cable	200 m
10BASE-T	Twisted pair cable	100 m
10BASE-F	Fiber	2000 m

  NOTE

  The fatal defect of the coaxial cable is the fact that devices on the cable are connected in series and thus a single node failure can cause the breakdown of the entire network. As the physical standards of coaxial cables, 10BASE-2 and 10BASE-5 have fallen into disuse.

- 100M Ethernet cable standards

  The 100M Ethernet is also called Fast Ethernet (FE). Compared with the 10M Ethernet, the 100M Ethernet has faster transmission rate at the physical layer, but they have no difference at the data link layer.

  Table 1-2 lists the 100M Ethernet cable standards.

  Table 1-2 100M Ethernet cable standards

Name	Cable	Maximum Transmission Distance
100Base-T4	Four pairs of Category 3 twisted pair cables	100 m
100Base-TX	Two pairs of Category 5 twisted pair cables	100 m
100Base-FX	Single-mode fiber or multi-mode fiber	2000 m

  Both the 10Base-T and 100Base-TX are applied to Category 5 twisted pair cables. They have different transmission rates. The 10Base-T transmits data at a rate of 10 Mbit/s whereas the 100Base-TX transmits data at 100 Mbit/s.

  The 100Base-T4 is rarely adopted now.

- Gigabit Ethernet cable standards

The Gigabit Ethernet is developed on the basis of the Ethernet standard defined in IEEE 802.3. Based on the Ethernet protocol, the transmission rate of the FE is increased by 10 times and reaches 1 Gbit/s. **Table 1-3** lists the Gigabit Ethernet cable standards.

Table 1-3 Gigabit Ethernet cable standards

Interface Name	Cables	Maximum Transmission Distance
1000Base-LX	Single-mode fiber or multi-mode fiber	316 m
1000Base-SX	Multi-mode fiber	316 m
1000Base-CX	Balanced twisted pair copper wire cable	25 m
1000Base-TX	Category 5 twisted pair cable	100 m

Using the Gigabit Ethernet technology, you can upgrade the existing Fast Ethernet from 100 Mbit/s to 1000 Mbit/s.

The physical layer of a Gigabit Ethernet uses 8B10B coding. In the traditional Ethernet technology, the data link layer delivers 8-bit data sets to its physical layer. After proper processing, the data sets, still being 8 bit, are sent to the data link layer for transmission.

The situation is different on the Gigabit Ethernet of optical fibers, in which the physical layer maps the 8-bit data sets transmitted from the data link layer to 10-bit data sets and then sends them out.

- 10GE cable standards

  IEEE 802.3ae is the 10GE cable standard. For a 10GE, the cables are all optical fiber in full-duplex mode.

  The 10GE is under way, and will be widely deployed in future.

CSMA/CD

- Concept of CSMA/CD

  The Ethernet network was originally designed to connect computers and other digital devices on a shared physical line. The computers and digital devices can access the shared line only in half-duplex mode. Therefore, a mechanism of collision detection and avoidance is required to prevent multiple devices from contending for the line. Carrier Sense Multiple Access with Collision Detection (CSMA/CD) is thus introduced.

  The concept of CSMA/CD is described as follows:

  - CS: carrier sense

    Before transmitting data, a station monitors the line to check whether the line is idle. In this manner, chances of collision are decreased.

  - MA: multiple access

    The data sent by a station can be received by multiple stations.

  - CD: collision detection

If two stations transmit electrical signals at the same time, the signals are superimposed, and thus the voltage amplitude doubles the normal amplitude. The situation results in collision.

The stations, therefore, stop transmission after sensing the conflict, and resume the transmission after a random delay.

- Working process of CSMA/CD

  CSMA/CD works as follows:

  1. A station continuously detects whether the shared line is idle.
     - If the line is idle, the station sends data.
     - If the line is in use, the station waits until the line becomes idle.
  2. If two stations send data at the same time, a conflict occurs on the line, and the signal becomes unstable.
  3. After detecting the instability, the station immediately stops sending data.
  4. The station sends a series of disturbing pulse. After waiting for a period of time, the station resumes the data transmission.

     Sending the disturbing pulse is to inform other stations, especially the station that sends data at the same time, that a conflict occurs on the line.

     After detecting a conflict, the station waits for a random period of time, and then resumes the data transmission.

Minimum Frame Length and Maximum Transmission Distance

- Minimum frame length

  Due to the limitation of the CSMA/CD algorithm, an Ethernet frame cannot be shorter than a certain length. On the Ethernet, the minimum frame length is 64 bytes, which is determined jointly by the maximum transmission distance and the collision detection mechanism.

  The use of minimum frame length can prevent the situation where station A finishes sending the last bit, but the first bit does not arrive at station B, which is in the distance. Station B considers that the line is idle and begins to send data, leading to a conflict.

  The upper layer protocol must guarantee that the Data field contains at least 46 bytes. Therefore, the Data field plus 14-byte Ethernet frame header, and the 4-byte check code at the frame tail equals the minimum frame length. If the Data field is less than 46 bytes, the upper layer must fill up the field.

  The upper limit of the Data field is set to 1500 bytes, which is required by the memory cost and the buffer of low-cost LAN controller in 1979.

- Maximum transmission distance

  The maximum transmission distance is decided by the factors such as line quality and signal attenuation.

Duplex Modes of the Ethernet

The physical layer of an Ethernet can work in either half-duplex or full-duplex mode.

- Half-duplex mode

  The half-duplex mode has the following features:

  - Receiving data or sending data takes place in only one direction at a time.

- The CSMA/CD mechanism is adopted.
- The transmission distance is limited.

Hubs work in half-duplex mode.

- Full-duplex mode

  After Layer 2 switches replace Hubs in networking, the shared Ethernet changes to the switched Ethernet, and the half-duplex mode is replaced by the full-duplex mode. As a result, the transmission rate is drastically increased, and the maximum throughput reaches the double rate.

  The full-duplex mode solves the problem of conflicts once and for all. CSMA/CD, therefore, is no longer adopted by the Ethernet.

  The full-duplex mode has the following features:

  - Transmitting data and receiving data can take place simultaneously.
  - The maximum throughput doubles the transmission rate.
  - This mode does not have the limitation on the transmission distance.

  Except Hubs, the network cards, Layer 2 devices, and Layer 3 devices produced in recent 10 years all support the full-duplex mode.

  To realize the full-duplex mode, the hardware requirements are as follows:

  - Full-duplex network cards and chips
  - Physical media over which sending and receiving frames are separated
  - Point-to-point connection

Auto-Negotiation of the Ethernet

- Purpose of auto-negotiation

  The earlier Ethernet adopts the 10 Mbit/s half-duplex mode, thus, mechanisms such as CSMA/CD are required to guarantee the system stability. With the development of technology, the full-duplex mode and 100M Ethernet emerge. As a result, the Ethernet performance is greatly improved. A new problem about how to achieve the compatibility between the earlier Ethernet and the new-constructed Ethernet arises.

  The auto-negotiation technology is thus introduced. In auto-negotiation, the devices on two ends of a link can choose the same operation parameters by exchanging information. The main parameters to be negotiated are mode (half-duplex or full-duplex), speed, and flow control. After the negotiation succeeds, the devices on two ends operate in the negotiated mode and rate.

  The auto-negotiation of duplex is defined in the following standards:

  - 100M Ethernet standard: IEEE 802.3u

    In IEEE 802.3u, the auto-negotiation is defined as an optional function.

  - Gigabit Ethernet standard: IEEE 802.3z

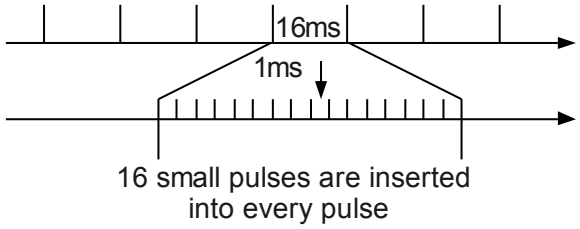
    In IEEE 802.3z, the auto-negotiation is defined as a mandatory and default function.

- Principle of auto-negotiation

  Auto-negotiation is an Ethernet procedure by which two connected devices choose common transmission parameters. It allows a network device to transmit the supported operating mode to the peer and receives the operating mode from the peer. In this process, the connected devices first share their capablities regarding these parameters and then choose the highest performance transmission mode they both support.

When no data is transmitted over twisted pair cables on an Ethernet network, pulses of high frequency are transmitted at an interval of 16 ms to form Normal Link Pulse (NLP) to maintain the connections at the link layer. Some pulses of high frequency can be inserted in the NLP to form Fast Link Pulse (FLP) to transmit more information, as shown in **Figure 1-1**. The basic mechanism of auto-negotiation is to encapsulate the negotiation information into FLP.

Figure 1-1 Schematic diagram of pulse insertion



Similar to an Ethernet network that uses twisted pair cables, an Ethernet network that uses optical modules and optical fibers also implements auto-negotiation by sending code streams. These code streams are called Configuration (C) code streams. Different from electrical interfaces, optical interfaces generally do not negotiate traffic transmission rates and work in duplex mode. Therefore, only flow control parameters are negotiated.

Auto-negotiation priorities of the Ethernet duplex link are listed as follows in a descending order:

- 1000M full-duplex
- 1000M half-duplex
- 100M full-duplex
- 100M half-duplex
- 10M full-duplex
- 10M half-duplex

If auto-negotiation succeeds, the Ethernet card activates the link. Then, data can be transmitted on the link. If auto-negotiation fails, the link is unavailable.

If one end does not support auto-negotiation, the other end that supports auto-negotiation adopts the default operating mode, which is generally at 10 Mbit/s and in half-duplex mode.

Auto-negotiation is implemented based on the chip design at the physical layer. As defined in IEEE 802.3, auto-negotiation is implemented in any of the following cases:

- A faulty link recovers.
- A device is re-powered on.
- Either of two connected devices resets.
- A renegotiation request packet is received.

In other cases, two connected devices do not always send auto-negotiation code streams. Auto-negotiation does not use special packets or bring additional protocol costs.

- Auto-negotiation rules for interfaces

  Two connected interfaces can communicate with each other only when they are in the same working mode.

- If both interfaces work in the same non-auto-negotiation mode, the interfaces can communicate.
- If both interfaces work in auto-negotiation mode, the interfaces can communicate through negotiation. The negotiated working mode depends on the interface with lower capability (specifically, if one interface works in full-duplex mode and the other interface works in half-duplex mode, the negotiated working mode is half-duplex). The auto-negotiation function also allows the interfaces to negotiate about the traffic control function.
- If a local interface works in auto-negotiation mode and the remote interface works in a non-auto-negotiation mode, the negotiated working mode of the local interface depends on the working mode of the remote interface.

  Table 1-4 describes the auto-negotiation rules for interfaces of the same type.

Table 1-4 Auto-negotiation rules for interfaces of the same type (the local interface works in auto-negotiation mode)

Interface Type	Working Mode of the Remote Interface	Auto-negotiation Result	Description
FE electrical interface	10 M half-duplex	10 M half-duplex	If the remote interface works in 10 M full-duplex or 100 M full-duplex mode, the working modes of the two interfaces are different after auto-negotiation, and packets may be dropped. Therefore, if the remote interface works in 10 M full-duplex or 100 M full-duplex mode, configure the local interface to work in the same mode.
	10 M full-duplex	10 M half-duplex	
	100 M half-duplex	100 M half-duplex	
	100 M full-duplex	100 M half-duplex	
GE electrical interface	FE auto-negotiation	100 M full-duplex	If the remote interface works in 10 M full-duplex or 100 M full-duplex mode, the working modes of the two interfaces are different after auto-negotiation, and packets may be dropped. Therefore, if the
	10 M half-duplex	10 M half-duplex	
	10 M full-duplex	10 M half-duplex	
	100 M half-duplex	100 M half-duplex	
	100 M full-duplex	100 M half-duplex	

Interface Type	Working Mode of the Remote Interface	Auto-negotiation Result	Description
	1000 M full-duplex	1000 M full-duplex	remote interface works in 10 M full-duplex or 100 M full-duplex mode, configure the local interface to work in the same mode.

Table 1-5 describes the auto-negotiation rules for interfaces of different types.

Table 1-5 Auto-negotiation rules for interfaces of different types

Interface Type	Working Mode of an FE Electrical Interface	Working Mode of a GE Electrical Interface	Auto-negotiation Result	Description
An FE electrical interface connecting to a GE electrical interface	10 M half-duplex	Auto-negotiation	10 M half-duplex	If the FE electrical interface works in 10 M full-duplex or 100 M full-duplex mode and the GE electrical interface works in auto-negotiation mode, the working modes of the two interfaces are different after auto-negotiation and packets may be dropped. Therefore, if the FE electrical interface works in 10 M full-duplex or 100 M full-duplex mode, configure the GE electrical interface to work in the same mode.
	10 M full-duplex		10 M half-duplex	
	100 M half-duplex		100 M half-duplex	
	100 M full-duplex		100 M half-duplex	
	Auto-negotiation	10 M half-duplex	10 M half-duplex	If the FE electrical interface works in auto-negotiation mode and the GE electrical interface works in 10 M full-duplex or 100 M full-duplex mode, the working modes of the two interfaces are different after auto-negotiation, and packets may be dropped. Therefore, if the GE electrical interface
		10 M full-duplex	10 M half-duplex	
		100 M half-duplex	100 M half-duplex	

Interface Type	Working Mode of an FE Electrical Interface	Working Mode of a GE Electrical Interface	Auto-negotiation Result	Description
		100 M full-duplex	100 M half-duplex	works in 10 M full-duplex or 100 M full-duplex mode, configure the FE electrical interface to work in the same mode. Do not configure the GE electrical interface to work in 1000 M full-duplex mode. If you configure the GE electrical interface to work in this mode, auto-negotiation fails.
		1000 M full-duplex	Failure	

According to the auto-negotiation rules described in **Table 1-4** and **Table 1-5**, if an interface works in auto-negotiation mode and the connected interface works in a non-auto-negotiation mode, packets may be dropped or auto-negotiation may fail. It is recommended that you configure two connected interfaces to work in the same mode to ensure that they can communicate properly.

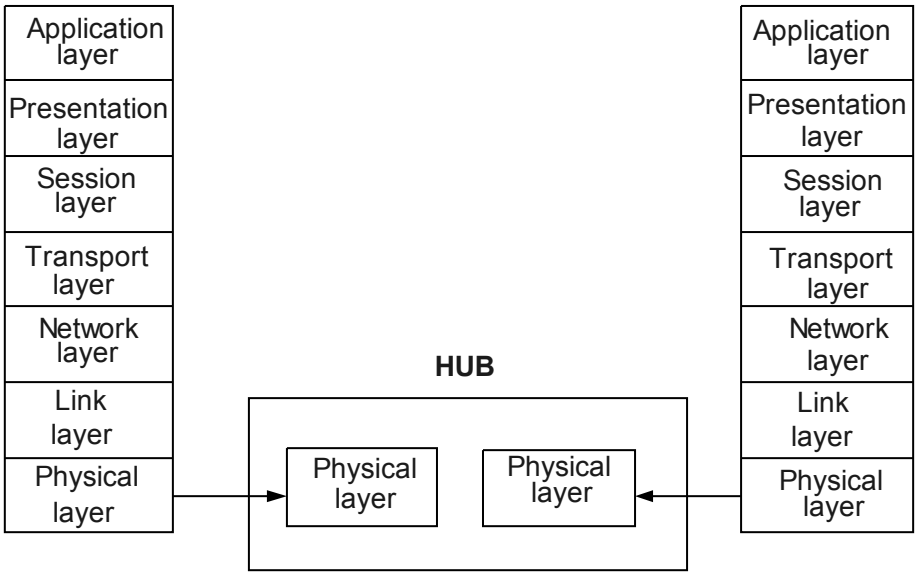
FE optical interfaces and higher-rate optical interfaces support only full-duplex mode. If these interfaces work in auto-negotiation mode, they can negotiate only about the traffic control function.

HUB

- Hub principle

  hen terminals are connected through twisted pair cables, a convergence device, which is called Hub, is required. Operating at the physical layer, Hubs connect devices. **Figure 1-2** shows a Hub operation model.

Figure 1-2 Hub operation model

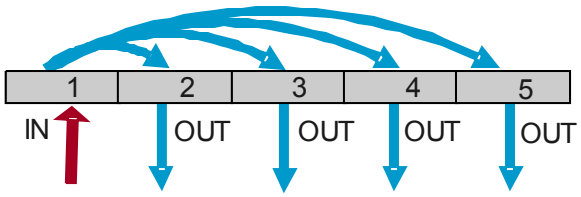


The appearance of a Hub is a box with multiple interfaces. Each interface can connect to a terminal. Thus, multiple devices can be connected through a Hub to form a star topology.

NOTE

Note that although the topology is physically a star shape, the Hub uses the bus and CSMA/CD technologies.

Figure 1-3 Hub operation principle



- According to the supported interfaces, Hubs can be classified into the following two types:
  - Category-I Hub: supports physical interfaces of one type.

    For example, a Category-I Hub provides only Category-5 twisted pair interfaces, Category-3 twisted pair interfaces, or optical fiber interfaces.
  - Category-II Hub: provides interfaces of different types. For example, a Category-II Hub can provide both Category-5 twisted pair interfaces and optical fiber interfaces.

    Both types have no difference in internal operation mode; however, they are used in different scenarios because they provide different types of interface. In practice, Category-I Hubs are commonly used.

1.4.2 Data Link Layer of the Ethernet

Hierarchical Structure of the Data Link Layer

In the Ethernet, according to different duplex modes, the following access modes are used:

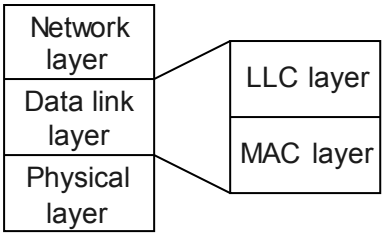
- CSMA/CD is used in half-duplex mode.
- Data is sent without detecting whether the line is idle in full-duplex mode.

Duplex mode, either half or full, refers to the operation mode of the physical layer. Access mode refers to the access of the data link layer. Therefore, in the Ethernet, the data link layer and physical layer are associated.

Thus, different access modes are required for different operation modes. This brings about some inconvenience to the design and application of the Ethernet.

Some organizations and vendors propose to divide the data link layer into two sub-layers: the Media Access Control (MAC) sub-layer and the Logical Link Control (LLC) sub-layer. Thus, different physical layers correspond to different MAC sub-layers, and the LLC sub-layer becomes totally independent, as shown in **Figure 1-4**.

Figure 1-4 Hierarchical structure of the data link layer of the Ethernet



Functions of the MAC sub-layer

The MAC sub-layer is responsible for the following:

- Providing the access to physical links.

  The MAC sub-layer is associated with the physical layer. That is, different MAC sub-layers provide access to different physical layers.

  In the Ethernet, two types of MAC sub-layers exist:

  - Half-duplex MAC: provides access to the physical layer in half-duplex mode.
  - Full-duplex MAC: provides access to the physical layer in full duplex mode.

  The two types of MAC are integrated in a network interface card. After the network interface card is initialized, auto-negotiation is performed to choose an operation mode, and then a MAC is chosen according to the operation mode.
- Identifying stations at the data link layer.

  The MAC sub-layer reserves a unique MAC address to identify each station.

  The MAC sub-layer uses a MAC address to uniquely identify a station.

  MAC addresses are managed by Institute of Electrical and Electronics Engineers (IEEE) and allocated in blocks. An organization, generally a vendor, obtains a unique address block

from IEEE. The address block is called the Organizationally Unique Identifier (OUI). Using the OUI, the organization can allocate addresses to 16777216 devices.

A MAC address consists of 48 bits, which are generally represented in 12-digit dotted hexadecimal notation. For example, the 48-bit MAC address 000000001110000011111100001110011000000000110100 is generally represented by 00e0.fc39.8034.

The first 6 digits in dotted hexadecimal notation stand for the OUI; the last 6 digits are allocated by the vendor. For example, in 00e0.fc39.8034, 00e0.fc is the OUI allocated by IEEE to Huawei; 39.8034 is the address number allocated by Huawei.

The second bit of a MAC address indicates whether the address is globally unique or locally unique. The Ethernet uses globally unique MAC addresses.

MAC addresses are divided into the following types:

- Physical MAC address

  A physical MAC address is burned into hardware (such as a network interface card) and is used to uniquely identify a terminal on the Ethernet.

- Broadcast MAC address

  A broadcast MAC address indicates all the terminals on a network.

  The 48 bits of a broadcast MAC address are all 1s, such as ffff.ffff.ffff.

- Multicast MAC address

  A multicast MAC address indicates a group of terminals on a network.

  The eighth bit of a multicast MAC address is 1, such as 00000001101110110011101010111010101111101010100**0**.

- Transmitting data over the data link layer. After receiving data from the LLC sub-layer, the MAC sub-layer adds the MAC address and control information to the data, and then transmits the data to the physical link. In the process, the MAC sub-layer provides other functions such as the check function.

  Data transmission at the data link layer is as follows:

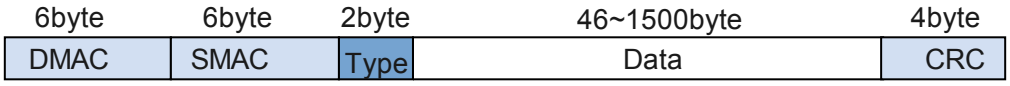
  1. The upper layer delivers data to the MAC sub-layer.
  2. The MAC sub-layer stores the data in the buffer.
  3. The MAC sub-layer adds the destination MAC address and source MAC address to the data, calculates the length of the data frame, and forms Ethernet frames.
  4. The Ethernet frame is sent to the peer according to the destination MAC address.
  5. The peer compares the destination MAC address with entries in the MAC address table.
     - If an entry is matched, the frame is accepted.
     - If no entry is matched, the frame is discarded.

  The preceding describes frame transmission in unicast mode. After an upper-layer application is added into a multicast group, the data link layer generates a multicast MAC address according to the application, and then adds the multicast MAC address to the MAC address table. The MAC sub-layer receives frames with the multicast MAC address and transmits the frames to the upper layer.

Frame Structure of the Ethernet

- Format of an Ethernet_II frame

Figure 1-5 Format of an Ethernet_II frame



The fields of the Ethernet_II frame are described as follows:

- DMAC

  It indicates the destination MAC address. DMAC specifies the receiver of the frame.
- SMAC

  It indicates the source MAC address. SMAC specifies the station that sends the frame.
- Type

  The 2-byte Type field identifies the upper layer protocol of the Data field. The receiver can know the meaning of the Data field according to the Type field.

  On the Ethernet, multiple protocols can coexist on a LAN. The hexadecimal values in the Type field of an Ethernet_II frame stand for different protocols.
  - Frames with the Type field value being 0800 are IP frames.
  - Frames with the Type field value being 0806 are Address Resolution Protocol (ARP) frames.
  - Frame with the Type field value being 8035 are Reverse Address Resolution Protocol (RARP) frames.
  - Frames with the Type field value being 8137 are Internetwork Packet Exchange (IPx) and Sequenced Packet Exchange (SPx) frames.
  - Frame with the Type field value being 8847 are Multiprotocol Label Switching (MPLS) frames.
- Data

  The minimum length of the Data field is 46 bytes, which guarantees that the frame is at least 64 bytes in length. The 46-byte Data field is required even if you attempt to transmit only 1-byte information.

  If the payload of the Data field is less than 46 bytes, the Data field must be padded to 46 bytes.

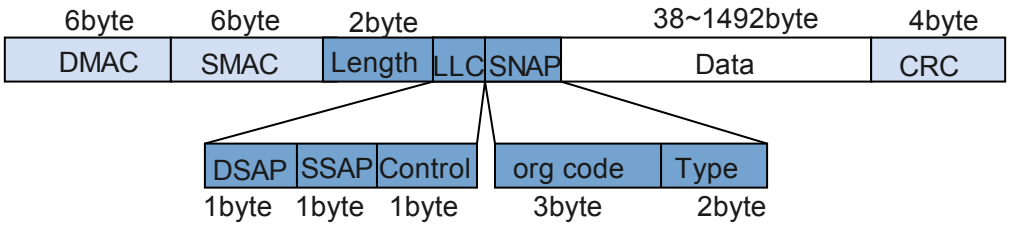
  The maximum length of the Data field is 1500 bytes.
- CRC

  The Cyclic Redundancy Check (CRC) field provides an error detection mechanism.

  Each sending device calculates a CRC code containing the DMAC, SMAC, Type, and Data fields. Then the CRC code is filled into the 4-byte CRC field.

- Format of an IEEE 802.3 frame

Figure 1-6 Format of an IEEE 802.3 frame



As shown in **Figure 1-6**, the format of an IEEE 802.3 frame is similar to that of an Ethernet_II frame except that in an IEEE 802.3 frame, the Type field is changed to the Length field, and the LLC field and the Sub-Network Access Protocol (SNAP) field occupy 8 bytes of the Data field.

- Length

  The Length field specifies the number of bytes of the Data field.

- LLC

  The LLC field consists of three sub-fields: Destination Service Access Point (DSAP), Source Service Access Point (SSAP), and Control.

- SNAP

  The SNAP field consists of the Org Code field and the Type field. Three bytes in the Org Code field are all 0s. The Type field functions the same as the Type field in Ethernet_II frames.

For descriptions about other fields, see the relevant description of Ethernet_II frames.

Based on the values of DSAP and SSAP, IEEE 802.3 frames can be divided into the following types:

- If DSAP and SSAP are both 0xff, the IEEE 802.3 frame changes to a Netware-Ethernet frame that bears NetWare data.
- If DSAP and SSAP are both 0xaa, the IEEE 802.3 frame changes to an Ethernet_SNAP frame.

  Ethernet_SNAP frames can be encapsulated with data of multiple protocols. The SNAP can be considered as an extension of the Ethernet protocol. SNAP allows vendors to invent their own Ethernet transmission protocols.

  The Ethernet_SNAP standard is defined by IEEE 802.1 to guarantee interoperability between IEEE 802.3 LANs and Ethernet networks.
- Other values of DSAP and SSAP indicate IEEE 802.3 frames.

LLC Sub-layer

As described, the MAC sub-layer supports two types of frame: IEEE 802.3 frames and Ethernet_II frames. In an Ethernet_II frame, the Type field identifies the upper layer protocol. Therefore, on a device, only the MAC sub-layer is required, and the LLC sub-layer need not be realized.

In an IEEE 802.3 frame, besides the traditional services of the data link layer, the LLC sub-layer defines additional useful features. All these features are provided by the sub-fields of DSAP, SSAP, and Control.

The following lists three types of point-to-point services:

- Connectionless service

  Currently, the Ethernet implements this service.
- Connection-oriented service

  The connection is set up before data is transmitted. The reliability of the data is guaranteed during the transmission.
- Connectionless data transmission with acknowledgement

The connection is not required before data transmission. The acknowledgement mechanism is adopted to improve the reliability.

The following is an example that describes the applications of SSAP and DSAP. Assume that terminals A and B use connection-oriented services. Data is transmitted in the following process:

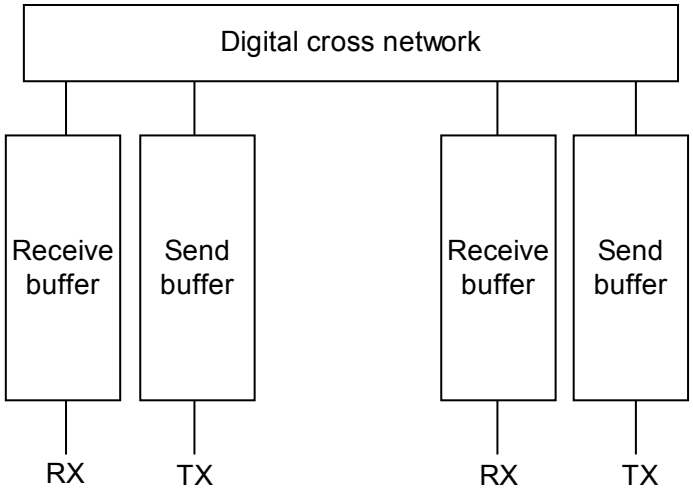
1. A sends a frame to B to require the establishment of a connection with B.
2. After receiving the frame, if B has enough resources, B returns an acknowledgement message that contains a Service Access Point (SAP). The SAP identifies the connection required by A.
3. After receiving the acknowledgement message, A knows that B has set up a local connection with A. After creating a SAP, A sends a message containing the SAP to B. The connection is set up.
4. The LLC sub-layer of A encapsulates the data into a frame. The DSAP field is filled in with the SAP sent by B; the SSAP field is filled in with the SAP created by A. Then the LLC sub-layer sends the data to the MAC sub-layer of A.
5. The MAC sub-layer of A adds the MAC address and the Length field into the frame, and then sends the frame to the data link layer.
6. After the frame is received at the MAC sub-layer of B, the frame is transmitted to the LLC sub-layer. The LLC sub-layer figures out the connection to which the frame belongs according to the DSAP field.
7. After checking and acknowledging the frame based on the connection type, the LLC sub-layer of B transmits the frame to the upper layer.
8. After the frame reaches its destination, A instructs B to release the connection by sending a frame. At this time, the communications end.

Ethernet Switches

- Structure of Ethernet Switches

  Similar to a hub in appearance, an Ethernet switch is a box with multi-interface. In internal structure, an Ethernet switch is much complicated than a hub. **Figure 1-7** shows the internal structure of an Ethernet switch.

Figure 1-7 Internal structure of an Ethernet switch



As shown in **Figure 1-7**, the internal structure of the Ethernet switch is a digital cross network. The digital cross network can temporarily connect terminals. On an Ethernet switch, each interface is configured with a buffer that temporarily stores the data. After the link becomes idle, the data is sent out.

An interface has two buffer: the receive buffer and the send buffer. When a terminal sends data to a switch, the data is temporarily stored in the receive buffer, waiting for further processing.

If the switch wants to transmit the data to another terminal, the switch sends the data to the send buffer of the switch interface that is directly connected with the destination terminal. Eventually, the data is sent to the terminal. If the terminal keeps busy, the data is stored in the send buffer all along.

The most significant difference between a switch and a hub is that the switch can perform interface-specific forwarding. After receiving a frame, a switch forwards the frame to another interface according to the MAC address in the frame. A hub, in this case, forwards the frame to all the interfaces.

- Operation of Ethernet Switches

  An Ethernet switch works as follows:

  1. The switch receives all the frames across the network.
  2. The switch sets up the MAC address table based on the source MAC addresses in the received frames. The address table is maintained through the address aging mechanism.
  3. The switch searches for the destination MAC addresses of the frames in the MAC address table.
     - If the matching addresses are found, the frames are forwarded to the corresponding interfaces except for the source interface.
     - If no matching address is found, the frames are sent to all the interfaces except for the source interface.
  4. The broadcast and multicast frames are sent to all the interfaces except for the source interface.

1.5 Applications

1.5.1 Computer Interconnection

Computer interconnection is the principal object and the major application of the Ethernet technology.

At the beginning, a lot of computers are connected through coaxial cables to access shared directories or access a file server located on the local network segment. All the computers, regardless of servers or hosts, are equal on the network.

The structure, however, cannot keep up with the development in applications. Currently, most traffic flows between clients and servers, this type of traffic model becomes a bottleneck on servers inevitably.

After the full-duplex Ethernet technology and Ethernet switches are introduced, the servers are connected to high-speed interfaces (100 Mbit/s) on Ethernet switches, and the clients are connected to low-speed interfaces on Ethernet switches. The traffic bottleneck can be alleviated. The modern operating system provides distributed services and database services. Servers based

on this operating system communicate with clients and other servers for data synchronization. 100M FE cannot meet the bandwidth requirement; therefore, the 1000M Ethernet technology emerges as the times require.

1.5.2 Interconnection Between High-Speed Network Devices

With the development of the Internet, bandwidth between some traditional network devices such as routers cannot meet the transmission requirements. As a higher-speed and more-efficient technology, 1000M Ethernet becomes the first choice to solve the problem. 100M FE can also solve this problem because after being converged, 100M FE networks can form FE channels whose speed ranges from 100 Mbit/s to 1000 Mbit/s.

1.5.3 Means to Access MANs

Nowadays, accessing the Metropolitan Area Network (MAN) to surf online, download files, and view Video on Demand (VoD) programs become more and more popular. The Ethernet technology is used as the means to access MANs because most computers support Ethernet network interface cards. Thus, users can go online without changing software and hardware configurations.

1.6 Terms and Abbreviations

Terms

Term	Description
10Base-T	Defined in IEEE 802.3i, it is an Ethernet specification that uses the twist pair with the maximum length of 100 meters (328.08 ft.) at 10 Mbit/s for each network segment.
100Base-T	Defined in IEEE 802.3u, it is a Fast Ethernet specification that uses the twist pair with the maximum length of 100 meters (328.08 ft.) at 100 Mbit/s for each network segment.
1000BaseT	Defined in IEEE 802.3ab, it is an Ethernet specification that uses the twist pair with the maximum length of 100 meters (328.08 ft.) at 1000 Mbit/s for each network segment.
E	
Ethernet	Created by Xerox and developed by Xerox, Intel, and Digital Equipment Corporation (DEC), it is a baseband LAN specification that uses CSMA/CD and transmits data over various cables at 10 Mbit/s. Ethernet-related standards are defined in IEEE 802.3 series.
Ethernet_II	It is an encapsulation format of Ethernet frames, which is the standard ARPA Ethernet Version 2.0 encapsulation that uses a 16-bit protocol type code.
Ethernet_SNAP	It is an encapsulation format of Ethernet frames. As specified in RFC 1042, it allows Ethernet frames to be transmitted through IEEE 802.2 media.

Term	Description
F	
FE	It is short for the Fast Ethernet. Complying with IEEE 802.3u, it is an extension and enhancement of the traditional media-sharing Ethernet standard and allows data to be transmitted at 100 Mbit/s.
Full-duplex	The full-duplex mode is an operation mode of Ethernet interfaces. In full-duplex mode, interfaces on both ends can send and receive data at the same time without interruption.
G	
GE	It is short for Gigabit Ethernet. Complying with IEEE 802.3z, the GE is compatible with the 10M Ethernet and the 100M Ethernet (FE).
H	
Half-duplex	The half-duplex mode is an operation mode of Ethernet interfaces. In half-duplex mode, an interface can only receive or send data at a time.
M	
MAC	It is short for Media Access Control. At the data link layer of the OSI model, the MAC sub-layer is adjacent to the physical layer.
Z	
Auto-negotiation	Through auto-negotiation, devices on both ends of a physical link exchange information to automatically select an operation mode. In auto-negotiation, the duplex mode and operation rate are negotiated. Once the negotiation result is approved, the operation mode is fixed until the device is restarted or the cable is removed.

Abbreviations

Abbreviation	Full Spelling
C	
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
G	

Abbreviation	Full Spelling
GE	Gigabit Ethernet
M	
MAC	Media Access Control
T	
TCP	Transmission Control Protocol

2 Interface Attributes

About This Chapter

2.1 Introduction to Basic Interface Attributes

Definition

The attributes of an interface refer to the inherent attributes of the interface. These attributes determine the operation mode and outward performance of the interface. The attributes of an interface can be configured.

Interfaces on the Switch support the following features:

- Auto-negotiation
- Traffic control
- Types of network cables supported by interfaces
- Jumbo frames
- Virtual cable test

Purpose

- Auto-negotiation

  With the development of network technologies, the devices on the network adopt various operation modes. To make the two ends on an Ethernet link adopt the same operation mode, a mechanism of automatic configuration is introduced. This mechanism is called auto-negotiation.
- Traffic control

  Interfaces can be configured with traffic control or auto-negotiation of traffic control to implement back pressure on congestion and control congestion.
- Types of network cables supported by interfaces

  This attribute is used to support the network cables with different wire sequences.
- Jumbo frames

  Generally, the length of a frame cannot exceed 1518 bytes. A large number of frames may enlarge the useless space between frames and lengthen frame headers. In this case, more bandwidths are occupied and bandwidth usage is reduced.

  Jumbo frames on the interface can reduce the number of frames and improve the bandwidth usage.
- Virtual cable test

  Currently, cables on LANs are laid inside walls or underground considering factors such as security and good looking. Tt is difficult to remove errors on cables. The Virtual Cable Test (VCT) technology is used. Generally, the technology of cable detection based on the Time Domain Reflect (TDR) is integrated on the existing PHY. By controlling the related hardware interfaces, the VCT function shows the status of cables through friendly interfaces. In this manner, users can conveniently and quickly locate faults and check lengths of cables.

2.2 References

Table 2-1 The references of this feature are as follows:

Document	Description	Remarks
IEEE 802.3	IEEE Std 802.3 - 2005 Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications	-

2.3 Availability

Involved Network Element

It is unnecessary to cooperate with other network elements.

License Support

No license is required.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

2.4 Principles

2.4.1 Auto-negotiation

An interface enabled with auto-negotiation sends signals to its connected interface and detects the signals sent from its connected interface to share transmission capabilities and negotiate transmission parameters. Once the interface receives signals from its connected interface and detects that the connected interface also receives the signals that it has sent, the two interfaces choose the fastest transmission mode they both support, which is automatically implemented by the link.

Acknowledgement is involved during auto-negotiation. That is, the local interface needs to acknowledge whether the information sent from the peer interface is received.

After obtaining the transmission capabilities from each other, the two interfaces negotiate and adopt the fastest operation mode they both share.

The priorities of operation modes are as follows:

- 1000 BaseT Full Duplex > 1000 BaseT Half Duplex

- 100 BaseT Full Duplex > 100 BaseT Half Duplex
- 10 BaseT Full Duplex > 10 BaseT Half Duplex

NOTE

- If two electrical interfaces are enabled with auto-negotiation, the two interfaces automatically negotiate the operation mode when the link is set up. The rate and duplex mode are automatically adopted according to the fastest transmission mode they both support.
- If only one interface is enabled with auto-negotiation, this interface automatically adopts the rate of the peer interface and half duplex.

2.4.2 Traffic Control

Traffic Control in Full Duplex Mode

As for devices in full-duplex communications, traffic control information is transmitted through pause frames.

Traffic Control in Half Duplex Mode

Devices in half-duplex communications implement traffic control through back pressure.

Summary of Traffic Control

In essence, both the transmission of pause frames in full-duplex mode and back pressure in half-duplex mode are methods of instructing the transmitter to stop sending frames when the space usage of the buffer on the receiver exceeds the threshold by sending messages that can be identified by both ends. In this manner, the buffer of the receiver is prevented from being overflowed and packet loss is prevented accordingly.

The traffic control mechanism is defined in IEEE 802.3. Most of the Ethernet devices support this mechanism.

2.4.3 Types of Network Cables

Generally, if two interfaces are connected with a twisted-pair cable, the receiving pins on the local end should be connected to the sending pins on the peer end and the sending pins on the local end should be connected to the receiving pins on the peer end so that a link can be up.

According to their wire sequence, twisted-pair cables can be classified into two categories: straight-through cables and cross-over cables. The device must support the negotiation and crossover of receiving and sending pins to support the two types of twisted-pair cables. It is required to set the type of network cables for Ethernet interfaces.

Medium Dependent Interfaces (MDIs) have three modes. They are displayed as follows:

- MDI Auto
- MDI Normal
- MDI Across

The following describes the three modes in detail.

MDI Auto

In MDI Auto mode, devices can automatically identify wire sequence and negotiate the sequence for sending and receiving packets. In this case, communications can be ensured, regardless of

whether cross-over or straight-through network cables are used and whether the peer device adopts the same MDI mode.

MDI Auto is the most popular mode. It is also the default mode on the switches. Its advantage is that devices communicate with each other normally regardless of the type of the twisted-pair cables and whether the peer end supports MDI.

MDI Normal

In MDI Normal mode, the receiving and sending pins are fixed and do not cross over. In this mode, no negotiation is implemented on the wire sequence of receiving and sending pins, regardless of whether the interfaces are enabled with auto-negotiation.

In MDI Normal mode, the sending pins correspond to pair A (and C) of the RJ-45 connector; the receiving pins correspond to pair B (and D). If both the connected ends are in this mode and use the straight-through network cable, the receiving pins on both ends are connected to each other and the sending pins on both ends are connected to each other. In this case, the devices cannot work normally.

The correct connection is as follows:

- If two devices work in MDI Normal mode, they should be connected with a cross-over network cable.
- If one device is in MDI Normal mode and the other device is in MDI Across mode, they should be connected with a straight-through network cable.

MDI Across

The MDI Across mode is the crossover mode. In this mode, the sending and receiving pins are fixed and do not cross over. In this mode, no negotiation is implemented on the wire sequence of receiving and sending pins, regardless of whether the interfaces are enabled with auto-negotiation.

Unlike the MDI Normal mode, in MDI Across mode, the sending pins correspond to pair B (and D) of the RJ-45 connector; the receiving pins correspond to pair A (and C). If both the connected ends are in this mode and use the straight-through network cable, the receiving pins on both ends are connected to each other and the sending pins on both ends are connected to each other. In this case, the devices cannot work normally.

The correct connection is as follows:

- If two devices work in MDI Across mode, they should be connected with a cross-over network cable.
- If one device is in MDI Normal mode and the other device is in MDI Across mode, they should be connected with a straight-through network cable.

Summary

A device in MDI Normal mode or MDI Across mode must use network cables with specific wire sequences. If a network cable does not match the connected interfaces, the interfaces cannot work normally. Therefore, the MDI Normal and MDI Across modes are not widely adopted. Generally, the MDI Auto mode is more popular.

The IEEE 802.3 standard recommends the MDI Auto mode so that the PHY can use its internal mechanism to implement the MDI or MDIX negotiation. Other modes are not recommended,

because the interface status may become abnormal in these modes. Especially in the case of 1000Base interfaces, MDI Auto mode is recommended.

2.4.4 Jumbo Frames

This attribute specifies the maximum size of a packet that is allowed to pass through an interface. The frames whose lengths are greater than 1518 bytes are called jumbo frames. Using jumbo frames can increase the bandwidth usage and decrease the useless space between frames and the transmission of frame headers.

2.4.5 VCT

Virtual cable test (VCT) uses the time-domain reflectometer (TDR) method for diagnosis. The TDR method takes advantage of different transmission speeds of electromagnetic waves in different media to test breakpoints in cables. Similar to a radar, a time-domain reflectometer sends a pulse and measures the pulse reflection time to analyze the characteristics of a transmission line. Time-domain signals are transmitted during a VCT test.

When the pulse is transmitted to the end of a cable or the faulty point of the cable, partial or all pulse energies are reflected to the transmission location. The VCT algorithm measures the time spent on transmitting pulses over cables, reaching the point of failure, and returning the pulses. The measured time is converted to the distance.

The status of links that can be detected by VCT is as follows:

- Ok
- Open
- Short
- Crosstalk

By using the VCT function, the fault type of a network cable can be detected and the point of failure can be located. In this manner, the network cable fault can be conveniently located.

Ok

When the twisted pair on a GE electrical interface is working properly, the VCT function can measure the lengths of pairs A, B, C, D in the twisted pair.

Open

Connectors at the end of twisted pairs lack continuity. The test result displays the position of a faulty line, that is, the distance between the interface and the point of failure.

Short

Short circuit occurs on two or more conducting wires. The test result displays the position of a faulty line, that is, the distance between the interface and the point of failure.

Crosstalk

Ends of twisted pairs are improperly connected, that is, the wire sequence is incorrect.

2.4.6 Automatic Port Sleeping

Ethernet electrical ports support automatic sleeping.

An Ethernet electrical port detects carrier signals to determine its working state. If no carrier signal is detected, the port enters low power mode (sleeping mode). The port in low power mode periodically sends carrier signals and detects incoming carrier signals. When the port receives carrier signals, it restores to normal mode.

2.5 Terms and Abbreviations

Terms

Term	Description
Auto-negotiation	Auto-negotiation is a mechanism that is applied to the Ethernet for negotiating the transmission parameters of two connected interfaces. In this process, the connected devices first share their capabilities as for these parameters and then choose the fastest transmission mode they both support.
FLP	FLPs are a series of pulses sent by a device to declare its transmission capabilities during auto-negotiation.
Jumbo frame	Ethernet frames longer than 1518 bytes and VLAN frames longer than 1522 bytes are called jumbo frames.
Full Duplex	Full duplex is a mode in which the transmitter and receiver are capable of transmitting and receiving data over the same channel simultaneously.
Half Duplex	Half duplex is a mode in which the transmission of data can be carried out in both directions, but only in one direction at a time.

Abbreviations

Acronym and Abbreviation	Full Spelling
MDI	Medium Dependent Interface
VCT	Virtual Cable Test
TDR	Time Domain Reflect

3 Trunk

About This Chapter

3.1 Introduction to Trunk

Definition

Trunks bind multiple physical interfaces together into a single logical interface, called a trunk interface. The bound physical interfaces are called member interfaces.

Trunk technology can increase the bandwidth, enhance reliability, and help implement load balancing.

Purpose

Without using trunk technology, the transmission rate between two network devices connected by fast Ethernet twisted pair cables is limited to 100 Mbit/s. Higher transmission rates can be achieved by changing the transmission media and replacing twisted pair cables with gigabit fiber cables, or upgrading the existing network to a Gigabit Ethernet network. These solutions, however, are costly and not suitable for small-and-medium size enterprises or institutions.

Therefore, trunk technology is an economical solution, which by binding multiple interfaces together, can increase interface bandwidth and achieve a higher transmission rate. For example, three 100 Mbit/s full-duplex interfaces can be bound together to provide a maximum bandwidth of 300 Mbit/s.

3.2 References

The following table lists the reference of this document.

Document	Description	Remarks
IEEE 802.3ad	IEEE Std 802.3ad - 2005 IEEE Standard for link aggregation operation, link aggregation control, link aggregation control protocol (LACP), marker protocol, and configuration capabilities and restrictions.	-

3.3 Availability

Involved Network Element

The Eth-Trunk and E-Trunk in static LACP mode need also be configured on the peer device.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

3.4 Principles

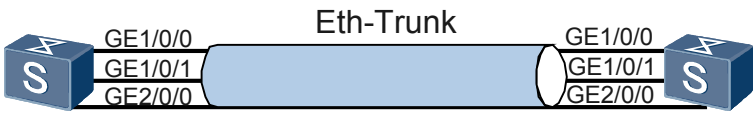
3.4.1 Basic Principles of Trunk

Trunking improves connection reliability, increases bandwidth, and implements load balancing across a trunk's member interfaces.

Additionally, a trunk interface can be configured to support various routing protocols and services.

Take the Eth-Trunk as an example. As shown in **Figure 3-1**, two switchs are directly connected through three interfaces. The interfaces are bound together into an Eth-Trunk interface to increase the bandwidth and improve the reliability.

Figure 3-1 Eth-Trunk network diagram



A trunk link can be considered a direct point-to-point link.

Trunks feature the following advantages over standard connections:

- Load balancing

  Load balancing can be implemented on a trunk interface. That is, on an Eth-Trunk interface, you can carry out load balancing.
- Higher reliability

  When the physical link of a member interface fails, the traffic on the member link is switched to another member link, ensuring uninterrupted service on the trunk link.
- Increased bandwidth

  The bandwidth of a trunk interface equals the sum of the bandwidth of all member interfaces.

Eth-Trunk link aggregation modes supported by the S7700 are shown in **Table 3-1**.

Table 3-1 Eth-Trunk link aggregation modes

Link Aggregation Modes	Network Requirements	Description
Eth-Trunk interfaces in manual load balancing mode	If one of the devices on one end of an Eth-Trunk link does not support the Link Aggregation Control Protocol (LACP), you can create an Eth-Trunk interface in load balancing mode on the S7700 and add multiple interfaces to the Eth-Trunk to increase bandwidth and enhance transmission reliability.	Manual load balancing is a basic link aggregation mode, in which you must manually create the Eth-Trunk interface, add interfaces to the Eth-Trunk interface, and specify active member interfaces. LACP is not involved. In manual load balancing mode, all active member interfaces forward data and perform load balancing. In this mode, Traffic can be evenly balanced among all member interfaces. If an active link of the link aggregation group fails, traffic is balanced among the remaining active links.
Eth-Trunk interfaces in static LACP mode	If both devices on either end of an Eth-Trunk link support LACP, you can create an Eth-Trunk interface in static LACP mode on the S7700. In this mode, both load balancing and redundancy backup can be implemented.	In static LACP mode, you must manually create an Eth-Trunk interface and add interfaces to the Eth-Trunk interface. Different from link aggregation in manual load balancing mode, active member interfaces are selected by sending LACP data units (LACPDUs) in static LACP mode. That is, when a group of interfaces are added to an Eth-Trunk interface, devices at both ends determine active and inactive interfaces by sending LACPDUs to each other. Static LACP mode is called M:N mode. In M:N mode, load balancing and redundancy backup can both be implemented. In the link aggregation group, M links actively forward data and perform load balancing and the other N links are inactive and do not forward data, functioning as standby links. When one active link fails, the system selects the highest priority standby link to replace the faulty link. The link then activates and begins forwarding data.

Link Aggregation Modes	Network Requirements	Description
Manual 1:1 active/standby mode	If the two ends of an Eth-Trunk are connected over intermediate devices, configure the manual 1:1 active/standby mode. **NOTE** The manual 1:1 active/standby mode is applicable only to Layer 2 Eth-Trunk interfaces.	An Eth-Trunk interface working in manual 1:1 active/standby mode contains only two member interfaces. Of the two member interfaces, one is active and the other standby. The active member interface forwards traffic when it functions properly. If the active member interface fails, the standby member interface takes over the traffic.

3.4.2 Restrictions on Trunk Interfaces

As a logical interface binding multiple physical interfaces and relaying upper-layer data, a trunk interface must comply with the following rules:

- Parameters of the physical interfaces (member interfaces) on both ends of the trunk link must be consistent. These parameters include:
  - Number of physical interfaces
  - Transmission rates of the physical interfaces
  - Duplex modes of the physical interfaces
  - Traffic-control modes of the physical interfaces
- Data sequence must be unchanged.

  A data flow can be considered as a group of frames with the same MAC address andIP address. For example, the Telnet or FTP connection between two devices can be considered as a data flow.

  If the trunk interface is not configured, frames that belong to a data flow can still reach their destination in the correct order because data flows are transmitted over only a physical link. When the trunk technology is used, multiple physical links are bound to the same trunk link, and frames are transmitted along these physical links. If the first frame is transmitted over a physical link, and the second frame is transmitted over another physical link, it is possible that the second frame reaches the destination earlier than the first frame.

  To prevent the disorder of frames, a frame forwarding mechanism is used to ensure that frames in the same data flow reach the destination in the correct sequence. This mechanism differentiates data flows based on their MAC addresses or IP addresses. In this manner, frames belonging to the same data flow are transmitted over the same physical link.

  After the frame forwarding mechanism is used, frames are transmitted in the following manners:
  - Frames with the same source MAC addresses are transmitted over the same physical link.
  - Frames with the same destination MAC addresses are transmitted over the same physical link.
  - Frames with the same source IP addresses are transmitted over the same physical link.

- Frames with the same destination IP addresses are transmitted over the same physical link.
- Frames with the same source + destination MAC addresses are transmitted over the same physical link.
- Frames with the same source + destination IP addresses are transmitted over the same physical link.

3.4.3 Classifications and Features of Trunk Interfaces

Classification

Trunk interfaces can be classified into two types, Eth-Trunk and IP-Trunk.

- An Eth-Trunk interface is composed of Ethernet interfaces.
- An IP-Trunk interface must be composed of POS interfaces.

Features of Trunk Interfaces

Eth-Trunk interfaces configured on the S7700 support the following features:

- Layer 2 forwarding and Layer 3 forwarding (unicast and multicast).
- Hash algorithm-based load balancing.
- QoS on the trunk interface.
- VPN instance binding.
- Hot backup and hot swapping.
- Interfaces from different boards can be added to a single Eth-Trunk interface.

Maximum/Minimum Number of Up Member Links

The number of Up member links affects the status and bandwidth of the trunk interface. To ensure that the trunk interface functions properly and is less affected by changes in member link status, set the following thresholds:

- Minimum number of Up member links

  When the number of Up member links falls below this threshold, the trunk interface goes Down. This guarantees the trunk interface a minimum available bandwidth.

  For example, if the trunk interface is required to provide a minimum bandwidth of 2 Gbit/s and each member link's bandwidth is 1 Gbit/s, the minimum number of Up member links must be set to 2 or a greater.

- Maximum number of Up member links

  When the number of Up member links reaches this threshold, the bandwidth of the trunk interface will not increase any further even if more member links go Up. This guarantees higher network reliability on the basis of sufficient bandwidth.

  For example, 8 trouble-free member links are bundled into a trunk link, each with a bandwidth of 1 Gbit/s. The trunk link, however, only needs to provide a maximum bandwidth of 5 Gbit/s. By setting the maximum number of Up member links to 5 or a greater, any unselected Up links automatically enter backup status, improving reliability.

NOTE

The maximum number of Up member links can be configured for only Eth-Trunk interfaces in static LACP mode.

The maximum number of Up member links is used to control the number of member links to go Up. That is, the number of Up member links cannot exceed this threshold, and additional Up member links are forcibly set Down.

In Layer 2 mode, the transmission rate of an Eth-Trunk interface is determined by the following conditions:

- Maximum number of Up member links
- Number of Up member interfaces

Load Balancing Carried Out Among Member Interfaces of a Trunk Interface

The S7700 uses per-flow load balancing.

Per-flow load balancing differentiates data flows based on the MAC address or the IP address in each packet, and then transmits the packets of the same data flow through one member link.

This load balancing mode guarantees orderly transmission, but not the bandwidth usage.

Trunk Member Interface Backup

To ensure high reliability on a trunk interface, you can configure a backup member interface in the Up state for another member interface.

If a member interface fails, its backup member interface (on the same trunk interface) takes over traffic transmitted along the faulty member interface. This is called trunk member interface backup or trunk fast switchover.

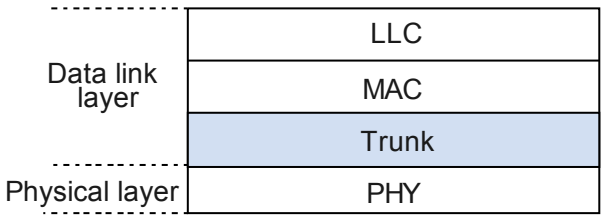
NOTE

Trunk member interface backup can only be configured for Eth-Trunk interfaces in static LACP mode.

3.4.4 Trunk Forwarding Principle

As shown in **Figure 3-2**, a trunk link is deployed on the data link layer, that is, between the physical layer and the MAC sub-layer.

Figure 3-2 Trunk interface in the Ethernet protocol stack



A trunk interface is assumed to be a physical interface on the MAC sub-layer. Therefore, frames transmitted in the MAC sub-layer only need to be delivered to the trunk module that maintains a trunk forwarding table.

The trunk forwarding table is composed of the following parts:

- HASH-KEY value

  The key value is calculated through the hash algorithm on the MAC address or IP address in the packet.

- Interface number

  The trunk forwarding table contains eight entries. The mappings between HASH-KEY values and interface numbers depend on the number of member interfaces in a trunk interface. Different HASH-KEY values are mapped to different outbound interfaces.

  For example, if four physical interfaces, 1, 2, 3, and 4, are bound into a trunk interface, the trunk forwarding table contains four entries, as shown in **Figure 3-3**.

  In the trunk forwarding table, the HASH-KEY values are 0, 1, 2, 3, 4, 5, 6, and 7, and the corresponding interface numbers are 1, 2, 3, 4, 1, 2, 3, and 4.

Figure 3-3 Example of a trunk forwarding table

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

The trunk module forwards a frame according to the trunk forwarding table. The forwarding process is as follows:

1. The trunk module receives a frame from the MAC sub-layer, and then extracts its source MAC address/IP address or destination MAC address/IP address.
2. The trunk module calculates the HASH-KEY value using the hash algorithm.
3. Based on the HASH-KEY value, the trunk module searches the trunk forwarding table for the interface number, and then sends the frame from the corresponding interface.

3.4.5 LACP

Introduction to Link Aggregation

As the Ethernet technology is more widely used on Metropolitan Area Networks (MANs) and Wide Area Networks (WANs), carriers pay more and more attention to Ethernet backbone network bandwidth and reliability. To increase bandwidth, a conventional solution uses high-speed interface cards and supported devices. This solution, however, is costly and inflexible. Link aggregation helps increase bandwidth by bundling a group of physical interfaces into a single logical interface, without having to upgrade hardware. In addition, link aggregation provides link backup mechanisms, greatly improving link reliability.

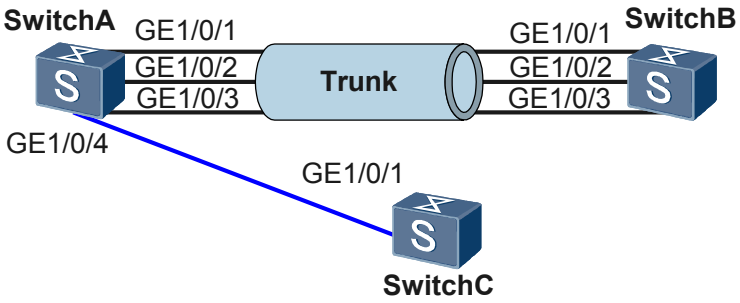
Trunking, as a link aggregation technique, helps increase bandwidth by bundling multiple physical interfaces into a single trunk interface. However, trunking is ineffective at fault detection, and can only detect link disconnection, but not other faults such as link layer faults or incorrect link connections. LACP is used to improve fault tolerance, provide M:N backup for trunks, and ensure high reliability of a trunk's member links.

LACP uses a standard negotiation mechanism for switching devices, ensuring that switching devices automatically create and enable aggregated links based on their configurations. After aggregated links are created, LACP maintains link status. If an aggregated link's status changes, LACP automatically adjusts or disables the link.

For example, in **Figure 3-4** a trunk link should be established between Switch A and Switch B by bundling four full-duplex GE interfaces on Switch A into a trunk interface and connecting it to the corresponding interfaces on Switch B. However, one of the GE interfaces is incorrectly connected to the interface on Switch C. As a result, the trunk interface cannot detect the fault in time and continues sending data to Switch C.

If LACP is enabled on Switch A, Switch B, and Switch C, and Switch A is configured with an LACP priority higher than that of Switch B, after the LACP negotiation, data will be correctly sent from Switch A to Switch B.

Figure 3-4 Incorrect trunk connection network diagram



Basic Concepts

- Link aggregation

  Link aggregation is a method of bundling a group of physical interfaces into a logical interface to increase bandwidth and improve reliability.

- Link aggregation group

  The Link Aggregation Group (LAG), also called a trunk link, is a logical link formed by bundling several physical links.

  If all bundled links are Ethernet links, the LAG is called an Ethernet LAG or an Eth-Trunk link. The LAG's interface is called an Eth-Trunk interface, and Ethernet interfaces that constitute an Eth-Trunk interface are called member interfaces.

  An Eth-Trunk interface is essentially a common Ethernet interface that needs to select one or more member interfaces to forward traffic. Therefore, Eth-Trunk interfaces are configured the same way as common Ethernet interfaces with the exception of a few parameters unique to Eth-Trunk member interfaces.

  NOTE

  The member interface of an Eth-Trunk interface cannot be the member interface of another Eth-Trunk interface.

- Active and inactive interfaces

  Member interfaces can be classified into active interfaces, which forward data and inactive interfaces, which do not.

  Links connected to active interfaces are called active links, and links connected to inactive interfaces are called inactive links.

  The link backup mechanism enhances link reliability, and if an active member link fails, a backup link will become active.

- Maximum number of active member interfaces

  If the number of active member interfaces reaches this threshold, no additional member interfaces will become active, regardless of whether there are more available.

- Minimum number of active member interfaces

  A minimum number active member interfaces are specified to ensure Eth-Trunk interface bandwidth. This threshold ensures the number of active member interfaces, preventing packet loss caused by heavy traffic across a few member links.

  If the number of active member interfaces falls below this threshold, the Eth-Trunk interface goes Down and none of its member interfaces forward data.

- LACP system priority

  LACP system priorities are set on devices at both ends of a trunk link. In static LACP mode, active member interfaces selected by both devices must be consistent; otherwise, the LAG cannot be established. To keep active member interfaces consistent at both ends, set a higher priority for one end. In this manner, the other end selects active member interfaces based on the selection of the peer.

  The smaller the LACP system priority value, the higher the LACP system priority. The default LACP system priority value is 32768.

- LACP interface priority

  The LACP interface priority is set for a member interface to determine whether it can be selected as an active member interface. The smaller the LACP interface priority value, the higher the LACP interface priority.

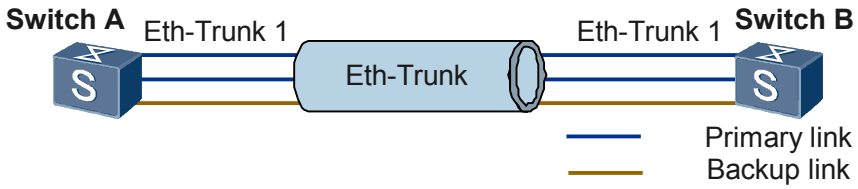
- M:N backup

  In static LACP mode, LACP is used to negotiate parameters to determine active member links in an LAG. This mode is also called the M:N mode, where M refers to the number of active links and N refers to the number of backup links. This mode guarantees high reliability and allows load balancing to be carried out across M active links.

  As shown in **Figure 3-5**, M+N links with the same attributes (in the same LAG) are set up between two devices. When data is transmitted over the aggregated link, load balancing is performed on the M active links; no data is transmitted over the N backup links. Therefore, the actual bandwidth of the aggregated link is the sum of the M links'bandwidth, and the maximum bandwidth of the aggregated link is the sum of the M+N links'bandwidth.

  If one of the M links fails, LACP selects a link from the N backup links to replace the faulty link. In such a situation, the actual bandwidth of the aggregated link is still the sum of M links'bandwidth; the maximum bandwidth of the aggregated link, however, becomes the sum of the M+N-1 links'bandwidth.

Figure 3-5 M:N backup network diagram



M:N backup is mainly applied in situations where the bandwidth of M links must be assured, and a fault tolerance mechanism in place. If an active link fails, the system can automatically select the backup link with the highest priority and add it to the current LAG.

If no available backup link is found, and the number of active links is smaller than the lower threshold, the system shuts down the LAG.

Link Aggregation Mode

Link aggregation is classified as one of the following two modes:

- Link aggregation in manual load balancing mode

  Manual load balancing is a basic link aggregation mode. In this mode, you need to manually create a trunk interface and add member interfaces to the trunk interface, without the assistance of the LACP protocol.

  In manual load balancing mode, all the member interfaces of an LAG share the traffic evenly.

  If an active link fails, the other active links share the traffic evenly.

- Link aggregation in static LACP mode

  In static LACP mode, you also need to manually create a trunk interface and add member interfaces into the trunk interface, the same as manual load balancing mode, but you must also specify active interfaces through LACP. That is, when a group of interfaces are added into the trunk interface, the status of each member interface (active or inactive) depends on LACP negotiation.

  A comparison of manual load balancing mode and static LACP mode is shown in **Table 3-2**.

Table 3-2 Comparison between the manual load balancing and static LACP modes

Difference/ Similarity	Manual Load Balancing Mode	Static LACP Mode
Difference	LACP is disabled. Does not check whether interfaces can be aggregated.	LACP is enabled. LACP checks whether interfaces can be aggregated.
Similarity	LAG is created and deleted manually. Member links are added and deleted manually.	

Link Aggregation in Manual Load Balancing Mode

Link aggregation in manual load balancing mode is widely applied. In this mode, you can manually add multiple interfaces to the LAG, and all the added interfaces forward data and perform load balancing. This mode is mainly applied to the scenario where wide link bandwidth is required and LACP is not supported by either or both devices. As shown in **Figure 3-6**, Switch A supports LACP, while Switch B does not support LACP.

Figure 3-6 Link aggregation in manual load balancing mode network diagram

In this mode, load balancing is implemented among all member interfaces. The S7700 supports the following load balancing types:

- Load balancing based on IP addresses or MPLS labels
- Load balancing based on MAC addresses

Link Aggregation in Static LACP Mode

LACP, as specified in IEEE 802.3ad, implements dynamic link aggregation and de-aggregation, allowing both ends to exchange LACPDUs.

After member interfaces are added to the trunk interface in static LACP mode, each end sends LACPDUs to inform its peer of its system priority, MAC address, member interface priorities, interface numbers, and keys. After being informed, the peer compares this information with that saved on itself, and selects which interfaces to be aggregated. Then, LACP negotiation occurs, selecting the active interfaces and links.

For detailed information about LACPDUs, see **Figure 3-7**.

Figure 3-7 LACPDU

Destination Address
Sourece Address
Length/Type
Subtype=LACP
Version Number
TLV_type=Actor Information
Actor_Information_Length=20
Actor_Port
Actor_State
Actor_System_Priority
Actor_System
Actor_Key
Actor_Port_Priority
Reserved
TLV_type=Partner Information
Partner_Information_Length=20
Partner_Port
Partner_State
Partner_System_Priority
Partner_System
Partner_Key
Partner_Port_Priority
Reserved
TLV_type=Collector Information
Collector_Information_Length=16
CollectorMaxDelay
Reserved
TLV_type=Terminator
Terminator_Length=0
Reserved
FCS

As show in **Figure 3-7**, explanation of main fields follow as:

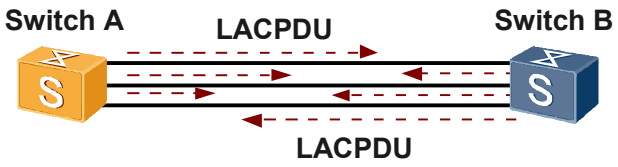
- Actor_Port/Partner_Port: interface of the Actor or Partner.
- Actor_State/Partner_State: status of the Actor or Partner.
- Actor_System_Priority/Partner_System_Priority: system priority of the Actor or Partner.
- Actor_State/Partner_System: system ID of the Actor or Partner.
- Actor_State/Partner_Key: operational Key of the Actor or Partner.
- Actor_Port_Priority/Partner_Port_Priority: interface priority of the Actor or Partner.
- The process of setting up an Eth-Trunk link in static LACP mode is as follows:

  1. Devices at both ends send LACPDUs to each other.

     As shown in **Figure 3-8**, you need to manually create an Eth-Trunk link in static LACP mode on Switch A and Switch B and add member interfaces to the Eth-Trunk. Then

the member interfaces are enabled with LACP, and devices at both ends can send LACPDUs to each other.

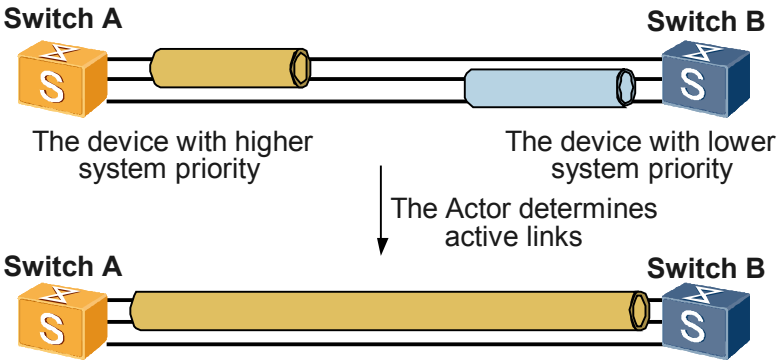
Figure 3-8 LACPDUs sent in static LACP mode network diagram



2. Devices at both ends determine the Actor based on the LACP system priority and system ID.

   As shown in **Figure 3-9**, devices at both ends receive LACPDUs from each other. For example, when When Switch B receives LACPDUs from Switch A, Switch B checks and records information about Switch A and compares system priorities. If the system priority of Switch A is higher than that of Switch B, Switch A acts as the Actor and Switch B selects the active interfaces based on the priorities of the corresponding interfaces on Switch A. In this manner, active interfaces of both switchs are determined.

   Figure 3-9 Actor selection process in static LACP mode network diagram

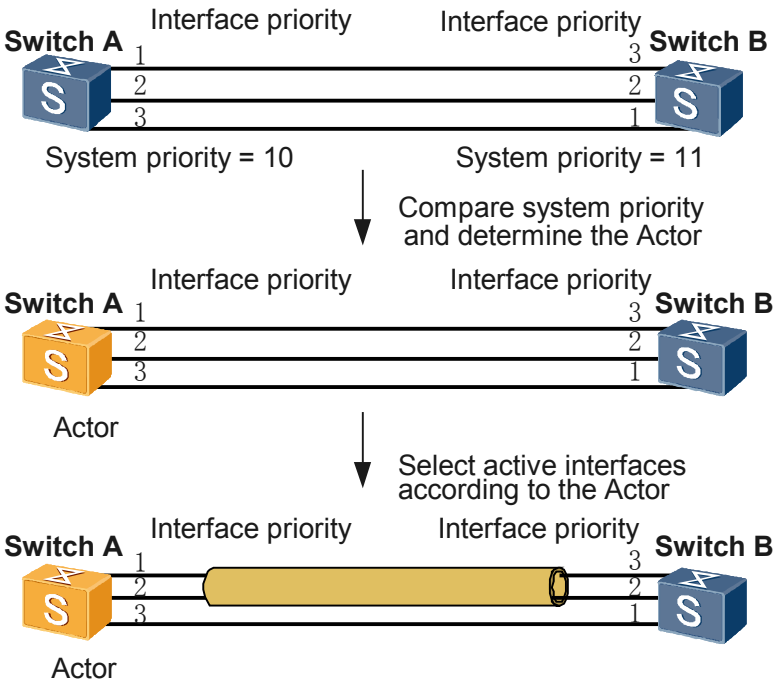


3. Devices at both ends determine active interfaces based on the Actor's LACP priorities and interface IDs.

   As shown in **Figure 3-10**, after devices at both ends select the Actor, they select active interfaces according to the priorities of the Actor's interfaces.

   Then active interfaces are selected, active links in the LAG are specified, and load balancing is implemented across these active links.

Figure 3-10 Active interface selection in static LACP mode network diagram



- Switchover between active links and inactive links

  In static LACP mode, a link switchover in the LAG is triggered if a device at one end detects one of the following events:

  - An active link goes Down.
  - Ethernet OAM detects a link fault.
  - LACP detects a link fault.
  - An active interface becomes unavailable.
  - If LACP preemption is enabled, the backup interface's priority is changed to be higher than that of the current active interface.

  When any of the preceding triggering conditions is met, the link switchover is performed in the following steps:

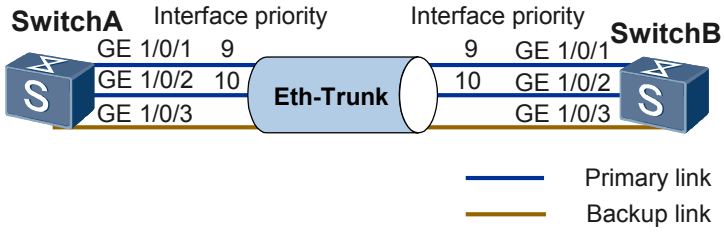
  1. The faulty link is disabled.
  2. The highest priority backup link is selected to replace the faulty active link.
  3. The highest priority backup link becomes the active link and begins forwarding data.

- LACP preemption

  After LACP preemption is enabled, interfaces with higher priorities in an LAG function as active interfaces.

  As shown in **Figure 3-11**, GE 1/0/1, GE 1/0/2, and GE 1/0/3 are member interfaces of Eth-Trunk 1. The upper threshold for the number of active interfaces is 2. LACP priorities of GE 1/0/1 and GE 1/0/2 are set to 9 and 10 respectively. The LACP priority of GE 1/0/3 is the default value. When the LACP negotiation is complete, GE 1/0/1 and GE 1/0/2 are selected as active interfaces because their LACP priorities are higher, and GE 1/0/3 is selected as the backup interface.

Figure 3-11 LACP preemption network diagram



LACP preemption is typically enabled in the following situations:

- GE 1/0/1 fails and then recovers. When GE 1/0/1 fails, GE 1/0/3 replaces it. After GE 1/0/1 recovers, if the LACP preemption is not enabled on Eth-Trunk 1, GE 1/0/1 remains to be the backup interface; if the LACP preemption is enabled on Eth-Trunk 1, GE 1/0/1 becomes the active interface and GE 1/0/3 becomes the backup interface.
- If the LACP preemption is enabled and GE 1/0/3 needs to replace GE 1/0/1 or GE 1/0/2 to become the active interface, you can set the LACP priority value of GE 1/0/3 to 8 or a smaller value. If the LACP preemption is not enabled, the system neither re-selects the active interface nor switches the active interface when the priority of a backup interface is higher than that of the active interface.

- LACP preemption delay

  After LACP preemption occurs, the backup link waits for a set period of time before switching to active status. This period is called LACP preemption delay. The LACP preemption delay can be configured, and can range from 10 to 180 seconds. The default delay is 30 seconds.

  The LACP preemption delay is set to prevent unstable data transmission along Eth-Trunk links caused by frequent status changes in member links.

  As shown in **Figure 3-11**, GE 1/0/1 becomes inactive because of a link failure. After a period, the link recovers. If LACP preemption is enabled and the period is shorter than the LACP preemption delay, GE 1/0/1 resumes as the active interface after the LACP preemption delay, and no status change of any backup interface occurs.

3.4.6 E-Trunk

Enhanced Trunk (E-Trunk), an extension from the Link Aggregation Control Protocol (LACP), is a mechanism that controls and implements link aggregation among multiple devices. E-Trunk implements device-level link reliability, instead of board-level link reliability implemented by LACP.

E-Trunk is mainly applied to a scenario where a CE is dual-homed to a VPLS, VLL, or PWE3 network. In this scenario, E-Trunk can be used to protect PEs and links between the CE and PEs. Without E-Trunk, a CE can be connected to only one PE by using an Eth-Trunk link. If the Eth-Trunk link or PE fails, the CE cannot communicate with the PE. By using E-Trunk, the CE can be dual-homed to PEs, establishing device-level protection.

Basic Concepts

- System LACP priority

  In LACP, the system LACP priority is used to differentiate the priorities of devices at both ends of an Eth-Trunk link. The smaller the value, the higher the priority.

In E-Trunk, to enable a CE to consider the peer PEs as a single device, you must configure the same system LACP priority and system ID for the PEs on both ends of an E-Trunk link, as shown in **Figure 3-12**.

- System ID

  In LACP, the system ID is used to determine the priorities of the two devices on both ends of an Eth-Trunk link if their LACP priorities are the same. The smaller the system ID, the higher the priority is. By default, the system ID is the MAC address of an Eth-Trunk interface.

  In E-Trunk, to enable a CE to consider the PEs as a single device, you must configure the same system LACP priority and system ID for the PEs on both ends of an E-Trunk link. As shown in **Figure 3-12**, the system ID is in the format of a MAC address.

- E-Trunk priority

  The E-Trunk priority determines the master/backup status of two devices in an aggregation group. As shown in **Figure 3-12**, PE1 has a higher E-Trunk priority than PE2, and therefore PE1 is the master device while PE2 is the backup device. The smaller the E-Trunk priority value, the higher the E-Trunk priority.

- E-Trunk ID

  An E-Trunk ID is an integer that uniquely identifies an E-Trunk link.

- Working mode

  The working mode is subject to the working mode of the Eth-Trunk interface added to the E-Trunk group. The Eth-Trunk interface works in one of the following modes:

  - Automatic
  - Forcible master
  - Forcible backup

- Timeout period

  Normally, the master and backup devices in an E-Trunk group periodically send Hello messages to each other. If the backup device does not receive any Hello message within the timeout period, it then becomes the master device.

  The timeout period is obtained through the formula: Timeout period = Sending period x Multiplier.

  The sending period ranges from 5 to 100, in 100 milliseconds. The default value is 10, or 1 second (ten 100 milliseconds). The multiplier value ranges from 3 to 300. The default value is 20.

  If the multiplier is 3, it indicates that the backup device becomes the master device if it does not receive any Hello message within three consecutive sending periods.

NOTE

Eth-Trunk interfaces mentioned in this document refers to the Eth-Trunk interfaces that are added to E-Trunk groups.

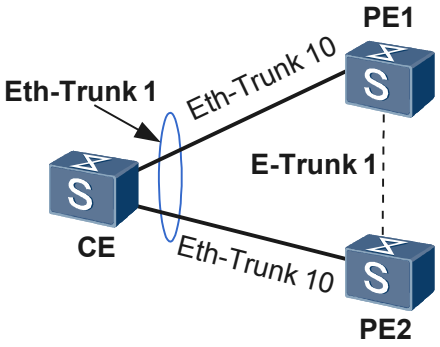
E-Trunk Working Principle

The E-Trunk working process is described as follows:

- Master/backup status negotiation

  As shown in **Figure 3-12**, the CE is directly connected to PE1 and PE2, and E-Trunk runs between PE1 and PE2.

Figure 3-12 Schematic diagram of E-Trunk



- PE end

  The same Eth-Trunk and E-Trunk interfaces are created on PE1 and PE2. In addition, the Eth-Trunk interfaces are added to the E-Trunk group.

- CE end

  Eth-Trunk interfaces in static LACP mode are configured on the CE. By using the Eth-Trunk interfaces, the CE is connected to PE1 and PE2.

  The E-Trunk group is invisible to the CE.

1. E-Trunk master/backup status

   PE1 and PE2 negotiate the E-Trunk master/backup status by exchanging E-Trunk packets. Normally, after the negotiation one PE functions as the master and the other as the backup.

   The master/backup status of a PE depends on the E-Trunk priority and E-Trunk ID carried in E-Trunk packets. The smaller the E-Trunk priority value, the higher the E-Trunk priority. The PE with the higher E-Trunk priority functions as the master. If the E-Trunk priorities of the PEs are the same, the PE with the smaller E-Trunk system ID functions as the master device.

2. Master/backup status of a member Eth-Trunk interface in the E-Trunk group

   The master/backup status of a member Eth-Trunk interface in the E-Trunk group is determined by its E-Trunk status and the peer Eth-Trunk interface status.

   As shown in **Figure 3-12**, PE1 and PE2 are on the two ends of the E-Trunk link. PE1 is considered as the local end and PE2 as the peer end.

   The status of each member Eth-Trunk interface in the E-Trunk group is determined, as shown in **Table 3-3**.

   Table 3-3 Master/backup status of an E-Trunk group and its member Eth-Trunk interfaces

Status of the Local E-Trunk	Working Mode of the Local Eth-Trunk Interface	Status of the Peer Eth-Trunk Interface	Status of the Local Eth-Trunk Interface
-	Forcible master	-	Master
-	Forcible backup	-	Backup

Status of the Local E-Trunk	Working Mode of the Local Eth-Trunk Interface	Status of the Peer Eth-Trunk Interface	Status of the Local Eth-Trunk Interface
Master	Automatic	Down	Master
Backup	Automatic	Down	Master
Backup	Automatic	Up	Backup

In normal situations:

- If PE1 functions as the master, Eth-Trunk 10 of PE1 functions as the master, and its link status is Up.
- If PE2 functions as the backup, Eth-Trunk 10 of PE2 functions as the backup, and its link status is Down.

If the link between the CE and PE1 fails, the following situations occur:

a. PE1 sends an E-Trunk packet containing information about the faulty Eth-Trunk 10 of PE1 to PE2.

b. After receiving the E-Trunk packet, PE2 finds that Eth-Trunk 10 on the peer is faulty. Then, the status of Eth-Trunk 10 on PE2 becomes master. Through the LACP negotiation, the status of Eth-Trunk 10 on PE2 becomes Up.

   The Eth-Trunk status on PE2 becomes Up, and traffic of the CE is forwarded through PE2. In this way, traffic destined for the peer CE is protected.

If PE1 is faulty, the following situations occur:

a. If the PEs are configured with BFD, the PE2 detects that the BFD session status becomes Down, then functions as the master and Eth-Trunk 10 of PE2 functions as the master.

b. If the PEs are not configured with BFD, PE2 will not receive any E-Trunk packet from PE1 before its timeout period runs out, after which PE2 will function as the master and Eth-Trunk 10 of PE2 will function as the master.

   Through the LACP negotiation, the status of Eth-Trunk 10 on PE2 becomes Up. The traffic of the CE is forwarded through PE2. In this way, destined for the peer CE is protected.

- Sending and receiving of E-Trunk packets

  E-Trunk packets carrying the source IP address and port number configured on the local end are sent through UDP. Factors triggering the sending of E-Trunk packets are as follows:

  - The sending timer times out.
  - The configurations change. For example, the E-Trunk priority, packet sending period, timeout period multiplier, addition/deletion of a member Eth-Trunk interface, or source/destination IP address of the E-Trunk group changes.
  - A member Eth-Trunk interface fails or recovers.

  E-Trunk packets contain the timeout period to be used as the timeout period for the peer.

- BFD fast detection

  A device cannot quickly detect a fault on its peer based on the timeout period of received packets. In this case, BFD can be configured on the device. The peer end needs to be

configured with an IP address. After a BFD session is established to detect whether or not the route destined for the peer is reachable, E-Trunk can sense any fault detected by BFD.

- Switchback mechanism

  The local device is in master state. In such a situation, if the physical status of the Eth-Trunk interface on the local device goes Down or the local device fails, the peer device becomes the master and the physical status of the member Eth-Trunk interface becomes Up.

  When the local end recovers, the local end needs to function as the master. Therefore, the local Eth-Trunk interface enters the LACP negotiation state. After being informed by LACP that the negotiation ability is Up, the local device starts the switchback delay timer. After the switchback delay timer times out, the local Eth-Trunk interface becomes the master. After LACP negotiation, the Eth-Trunk interface becomes Up.

E-Trunk Restrictions

To improve the reliability of CE and PE links, and to ensure that traffic can be automatically switched between these links, the configurations on both ends of the E-Trunk link must be consistent. Use the networking in **Figure 3-12** as an example.

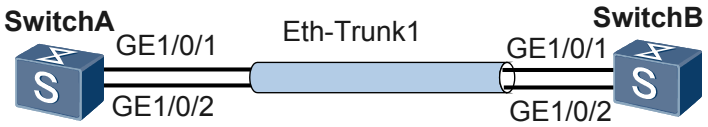
- The Eth-Trunk link directly connecting PE1 to the CE and the Eth-Trunk link directly connecting PE2 to the CE must be configured with the same working rate, and duplex mode. This ensures that both Eth-Trunk interfaces have the same key and join the same E-Trunk group. After the Eth-Trunk interfaces are added to the E-Trunk group, both PEs must contain the system LACP priorities and IDs. The interfaces connecting the CE to PE1 and PE2 must be added to the same Eth-Trunk interface. Note that the Eth-Trunk interface can have a different ID from that of the PEs. For example, the CE is configured with Eth-Trunk 1, whereas both PEs are configured with Eth-Trunk 10.
- Proper IP addresses must be specified for the two PEs to ensure Layer 3 connectivity. The address of the local PE is the peer address of the peer PE, and the address of the peer PE is the peer address of the local PE. Here, it is recommended that the addresses of the PEs are configured as loopback interface addresses.
- The E-Trunk group must be bound to a BFD session.
- The two PEs must be configured with the same security key (if necessary).

3.5 Application Environment

3.5.1 Eth-Trunk

As shown in **Figure 3-13**, an Eth-Trunk link is established between Switch A and Switch B, and two full-duplex interfaces are added to the Eth-Trunk interface. In this manner, the total bandwidth of the Eth-Trunk interface doubles that of each interface.

Figure 3-13 Networking diagram of Eth-Trunk



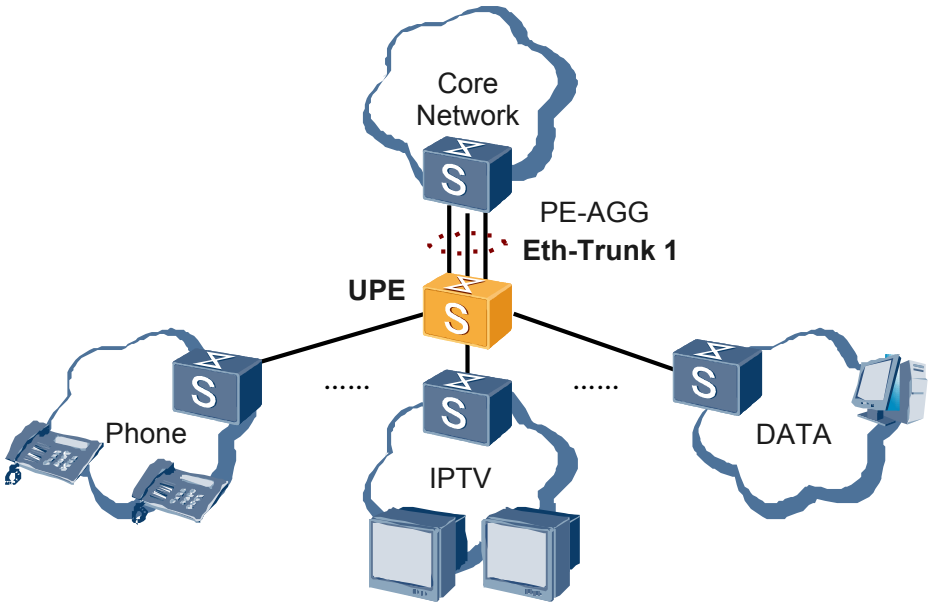
If the Eth-Trunk interface's backup function is enabled, traffic will be switched to the backup member link if an active member link fails. This improves link reliability.

In addition, network congestion can be avoided since traffic is balanced between two member links.

3.5.2 Link Aggregation Group

As shown in **Figure 3-14**, traffic of different services is sent to the core network through the UPE and PE-AGG, different services having varied priorities. To ensure the bandwidth and reliability of the link between the UPE and the PE-AGG, a link aggregation group, Eth-Trunk 1, is established.

Figure 3-14 Networking diagram of the link aggregation group



You can select the operation mode for the Eth-Trunk according to the following situations:

- If devices at both ends of the Eth-Trunk link support LACP, Eth-Trunk interfaces in static LACP mode are recommended.
- If the device at either end of the Eth-Trunk does not support LACP, Eth-Trunk interfaces in manual load balancing mode are recommended.

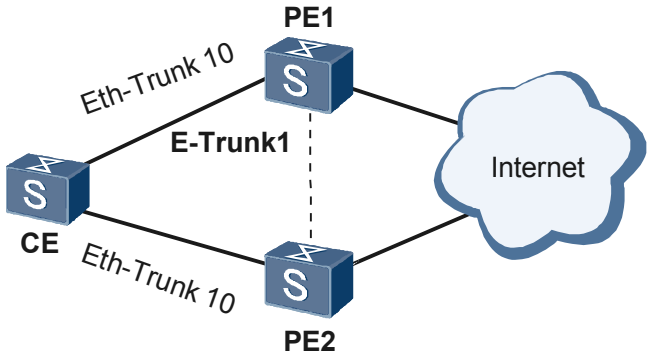
After an Eth-Trunk interface is created, QoS can be implemented on it as a common interface.

On both ends of Eth-Trunk 1 (the UPE and PE-AGG), you can implement traffic shaping, congestion management, and congestion avoidance for outgoing traffic. This ensures that packets of high priorities are sent in time.

3.5.3 E-Trunk

As shown in **Figure 3-15**, the E-Trunk is used to protect the links between a CE and two PEs when the CE is dual-homed to the two PEs. The CE is connected to PE1 and PE2 through a static LACP Eth-Trunk respectively. The two Eth-Trunks form an E-Trunk to implement backup of link aggregation groups between PE1 and PE2 and enhance the network reliability.

Figure 3-15 Network diagram of the E-Trunk



3.6 Terms and Abbreviations

Terms

Term	Explanation
LA	Link Aggregation (LA) refers to the method of binding a group of physical ports into a single logical port to increase bandwidth.
LAG	A Link Aggregation Group (LAG), also called a Load Sharing Group, is deployed between two devices to ensure higher bandwidth and provide redundancy protection (higher reliability) between them. LAGs require no hardware upgrade.
LACP	Link Aggregation Control Protocol (LACP) is a standard negotiation method devices use to exchange data.
BFD	Bidirectional Forwarding Detection (BFD) is a unified detection mechanism used to quickly detect and track connectivity of network links and/or IP routing. To improve network performance, adjacent systems must be able to quickly detect communication failures and establish backup channels to resume communication.
ETH-OAM	Ethernet-Operation Administration Maintenance (ETH-OAM) refers to the operation, administration, and maintenance of Ethernet networks.

Abbreviations

Abbreviation	Full Spelling
LACP	Link Aggregation Control Protocol
LA	Link Aggregation
LAG	Link Aggregation Group
BFD	Bi-directional Forwarding Detection
ETH-OAM	Ethernet-Operation Administration Maintenance

4 VLAN

About This Chapter

4.1 Introduction to VLAN

Definition

The Virtual Local Area Network (VLAN) technology logically divides a physical LAN into multiple VLANs (broadcast domains). As a result, hosts within the same VLAN can communicate directly, while hosts in different VLANs cannot. In this manner, messages are broadcast in each VLAN, inter-VLAN communication is restricted, and network security is enhanced.

Purpose

The traditional LAN technology, as based on the bus structure, has the following shortcomings:

- Conflict is inevitable if multiple nodes send messages simultaneously.
- Messages are broadcast to all nodes.
- Network security is not guaranteed since all hosts share the same transmission channel.

The traditional LAN can be considered a collision domain, where as more hosts added to the LAN, the more severe a conflict becomes, hindering network efficiency. Meanwhile, the traditional LAN can also be considered a broadcast domain, and as more hosts broadcast messages simultaneously, higher bandwidth is consumed by broadcast traffic.

In summary, the traditional LAN suffers the disadvantages of the collision domain and broadcast domain, and cannot ensure the network security.

Therefore, bridges and Layer 2 switches, which can be used to add more hosts to a LAN and effectively isolate the collision domain, can compensate for the shortcomings of traditional LAN solutions.

Bridges and Layer 2 switches can forward data from the inbound interface to the outbound interface in switching mode. This solves the access conflict problem experienced with shared media, and limits the collision domain to the port level. However, bridge and Layer 2 switch networking can only solve the problem of the collision domain, but not broadcast domain and network security issues.

To reduce the broadcast traffic, you should enable broadcast only among hosts that need to communicate with each other, and isolate hosts that do not. The router can select routes based on IP addresses and effectively reduce broadcast traffic between two connected network segments. The router solution, however, is costly, and most users choose to construct multiple logical LANs, namely, VLANs on the physical LAN.

In this manner, a physical LAN is divided into multiple broadcast domains, that is, multiple VLANs. The intra-VLAN communication is not restricted, while the inter-VLAN communication is restricted. As a result, broadcast packets are always confined in each VLAN, and the network security is also enhanced.

For example, different companies located in the same building may share the same LAN, each with their own secure VLAN, instead of incurring significantly more cost by building their own independent LANs separately or the security risks that come with an insecure shared LAN.

In such a situation, the VLAN technology can be adopted. Then these companies can not only share the LAN facility, but also guarantee the network security.

Figure 4-1 Typical VLAN topology network diagram

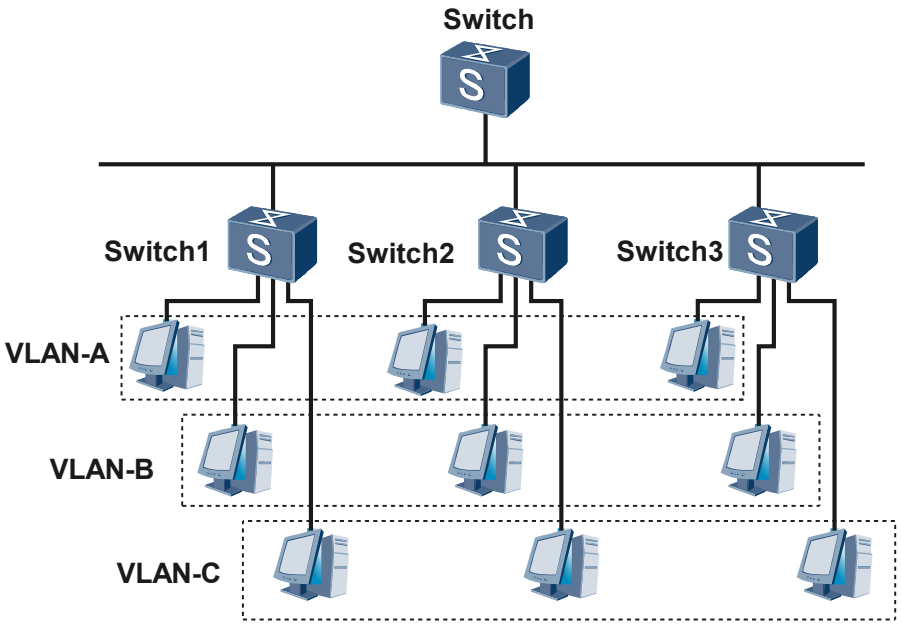


Figure 4-1 shows a typical VLAN topology. Three switches are placed in different locations (for example, different floors of a building); each switch is connected to three hosts that respectively belong to different VLANs (for example, different companies). In the diagram, a dotted box indicates a VLAN.

4.2 References

The following table lists the references of this document.

Document	Description	Remarks
RFC 3069	VLAN Aggregation for Efficient IP Address Allocation	-
IEEE 802.1q	IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks	-
IEEE 802.1ad	IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks- Amendment 4	-
IEEE 802.10	IEEE Standards for Local and Metropolitan Area Networks: Standard for Interoperable LAN/MAN Security	-
YD/T 1260-2003	Technical and Testing Specification of Virtual LAN Based on Port	-

4.3 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

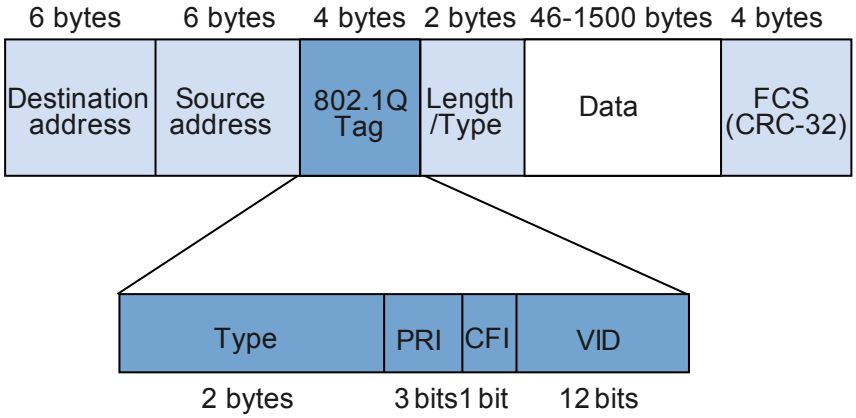
4.4 Principles

4.4.1 Basic Concepts of VLAN

Format of a VLAN Frame

The Ethernet frame format is modified in IEEE 802.1Q, with a 4-byte 802.1Q tag added between the source MAC address field and the protocol type field, as shown in **Figure 4-2**.

Figure 4-2 VLAN frame format as defined in IEEE 802.1Q



An 802.1Q tag contains four fields, described as follows:

- Type

The 2-byte Type field indicates the frame type. If the value of the field is 0x8100, it indicates an 802.1Q frame. When a device that does not support 802.1Q frames receives an 802.1Q frame, it discards the frame.

- PRI

  The 3-bit Priority field indicates the frame priority. The value of the field ranges from 0 to 7. The greater the value, the higher the frame priority. When a switch is congested, the frames of the higher frame priority are sent preferentially.

- CFI

  The 1-bit Canonical Format Indicator (CFI) field indicates whether the MAC address is in canonical format. If the CFI field is 0, it indicates that the MAC address is in canonical format. If the CFI field value is 1, it indicates that the MAC address is in non-canonical format. This field is mainly used to differentiate Ethernet frames, Fiber Distributed Digital Interface (FDDI) frames, and token ring frames. The CFI field value in Ethernet frames is 0.

- VID

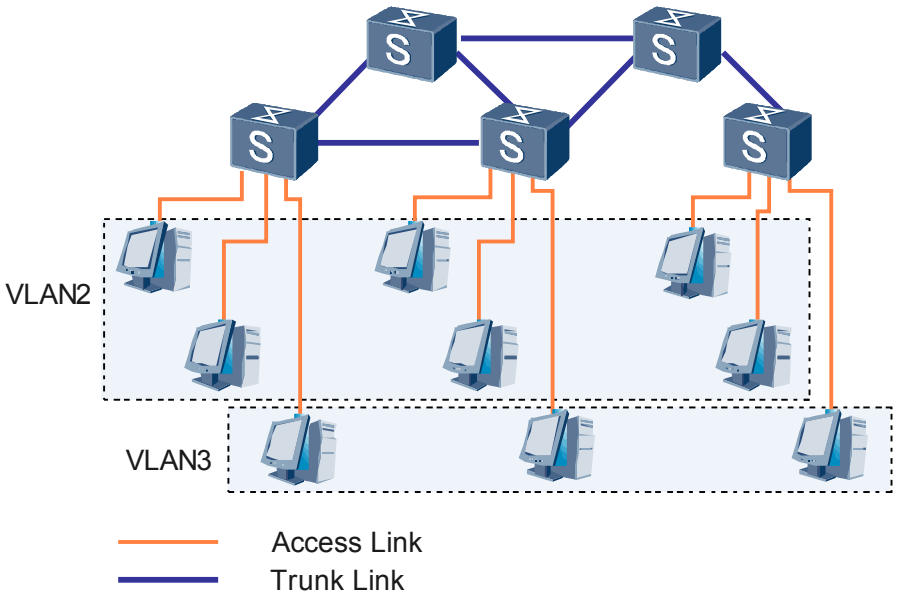
  The 12-bit VLAN ID (VID) field indicates to which VLAN the frame belongs. In the S7700, the VLAN ID ranges from 0 to 4095. Note that 0 and 4095 are reserved VLAN IDs and unavailable to users.

Link Types

VLAN links can be classified into the following types:

- Access link: refers to the link between a host and a switch. As shown in **Figure 4-3**, the link between PCs and switches are all access links.
- Trunk link: refers to the link between switches. As shown in **Figure 4-3**, the links between switches are trunk links. Frames transmitted over trunk links carry VLAN tags.

Figure 4-3 Link types network diagram



Port Types

After IEEE 802.1Q defines VLAN frames, some ports of a device can identify VLAN frames, whereas others cannot. Ports can be classified into four types based on whether or not they can identify VLAN frames according to the IEEE 802.1Q VLAN frame format:

- Access port

  As shown in **Figure 4-3**, the access port is used to connect the user host and it can connect to only the access link. An access port has the following features:

  - Only frames tagged with the port default VLAN ID (PVID) of the port can pass through an access port.
  - When receiving an untagged frame from an access port, the switching device adds a VLAN tag contains the PVID of the port to the frame.
  - Ethernet frames sent from an access port to the peer device never carry VLAN tags.

- Trunk port

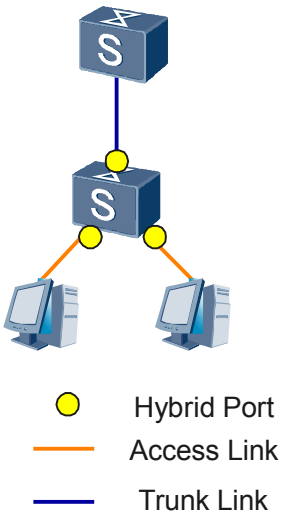
  As shown in **Figure 4-3**, the trunk port is used to connect to other switches and it can connect to only the trunk link. A trunk port has the following features:

  - A trunk port allows tagged frames from multiple VLANs to pass through.
  - When sending a tagged frame from a trunk port, if the VLAN ID in the tag is the PVID, the switching device removes the tag from the frame. Because the PVID of each port is unique, frames sent from a trunk port do not carry tags only in this case.
  - When sending a tagged frame from a trunk port, if the VLAN ID in the tag is different from the PVID, the switching device directly sends the frame out.

- Hybrid port

  As shown in **Figure 4-4**, the hybrid port is used to connect to either hosts or switches. That is, the hybrid port can connect to either access links or trunk links. A hybrid port allows frames from multiple VLANs to pass through and can remove VLAN tags from some outgoing VLAN frames.

  Figure 4-4 Port types network diagram

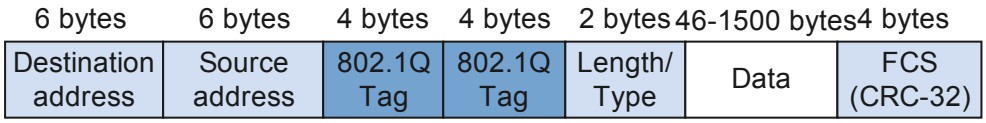


- QinQ port

An 802.1Q-in-802.1Q (QinQ) port refers to a QinQ-enabled port. A QinQ port adds double VLAN tags to a frame. That is, a QinQ port adds an outer tag to a single-tagged frame. In this manner, a maximum of 4094 x 4094 VLANs can be supported, which ensures enough VLANs on MANs.

Figure 4-5 shows the format of an QinQ frame. The outer tag is usually called the public network tag for carrying the public network VLAN ID. The inner tag is usually called the private network tag for carrying the private network VLAN ID.

Figure 4-5 Format of a QinQ frame



For details on the QinQ protocol, see **QinQ**.

Classification of VLANs

Table 4-1 shows the classification of VLANs.

Table 4-1 Differences between VLAN classification modes

VLAN Classification Mode	Principle	Advantage	Disadvantage
Classification of VLANs based on port numbers	In this mode, VLANs are classified based on the numbers of ports on a switching device. The network administrator configures a port default VLAN ID (PVID), that is, the default VLAN ID, for each port on the switching device. That is, a port belongs to a VLAN by default. When a data frame reaches a port, it is marked with the PVID if the data frame carries no VLAN tag and the port is configured with a PVID. If the data frame carries a VLAN tag, the switching device will not add a VLAN tag to the data frame even if the port is configured with a PVID. Different types of ports process VLAN frames in different manners.	It is simple to define VLAN members.	VLANs must be re-configured when VLAN members change locations.

VLAN Classification Mode	Principle	Advantage	Disadvantage
Classification of VLANs based on MAC addresses	In this mode, VLANs are classified based on the MAC addresses of network interface cards (NICs). The network administrator configures the mappings between MAC addresses and VLAN IDs. In this case, when a switching device receives an untagged packet, it searches the MAC-VLAN table for a VLAN tag to be added to the packet according to the MAC address of the packet.	When the physical locations of users change, you do not need to re-configure VLANs for the users. This improves the security of users and increases the flexibility of user access.	This mode is applicable to only a simple networking environment where the NIC seldom changes. In addition, all members on the network must be pre-defined.
Classification of VLANs based on IP subnets	When receiving an untagged packet, a switching device adds a VLAN tag to the packet based on the IP address of the packet.	Packets sent from specific network segments or IP addresses are transmitted in specific VLANs. This decreases burden on the network administrator and facilitates management.	This mode is applicable to the networking environment where users are distributed in an orderly manner and multiple users are on the same network segment.
Classification of VLANs based on protocols	VLAN IDs are allocated to packets received on an interface according to the protocol (suite) type and encapsulation format of the packets. The network administrator configures the mappings between types of protocols and VLAN IDs. In this case, when a switching device receives an untagged packet, it searches the Protocol-VLAN table for a VLAN tag to be added to the packet according to the protocol of the packet. **NOTE** At present, VLANs can be classified based on IPv4, IPv6, IPX, or AppleTalk (AT), and the encapsulation format can be Ethernet_II, raw defined in IEEE 802.3, and LLC and SNAP defined in IEEE 802.2.	The classification of VLANs based on protocols binds the type of services to VLANs. This facilitates management and maintenance.	The network administrator must initially configure the mappings between types of protocols and VLAN IDs.

VLAN Classification Mode	Principle	Advantage	Disadvantage
Classification of VLANs based on policies (MAC addresses, IP addresses, and interfaces)	In this mode, VLANs are classified based on MAC addresses and IP addresses configured on switched and associated with VLANs. Only users matching a policy can be added to a specific VLAN. After users are added to the VLAN, if their IP addresses or MAC addresses are changed, they no longer belong to the VLAN.	Policy-based VLAN classification is of high security. Do not change MAC addresses or IP addresses of users that have been added to VLANs based on MAC addresses and IP addresses. Compared with other VLAN classification modes, MAC address and IP address-based VLAN classification has the highest priority.	Each policy has to be manually configured.

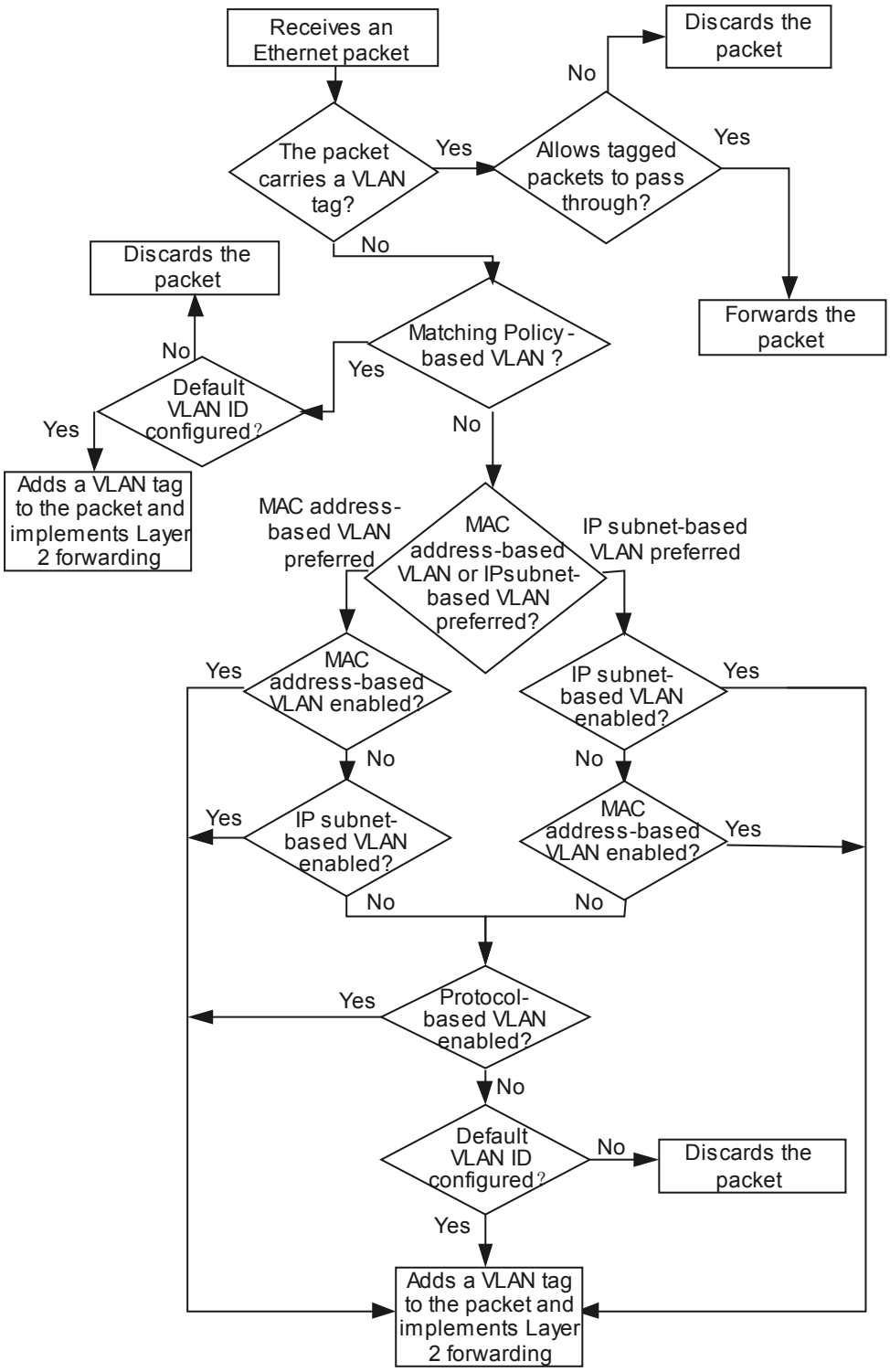
If the S7700 supports multiple VLAN classification modes, it adopts a mode in a descending order of policy-based VLAN classification, MAC address-based VLAN classification, IP subnet-based VLAN classification, protocol-based VLAN classification, and port-based VLAN classification.

- MAC address-based VLAN classification and IP subnet-based VLAN classification have the same priority.

  By default, MAC address-based VLAN classification is preferentially adopted. Alternatively, you can run commands to change priorities of these two VLAN classification modes to select a VLAN classification mode.
- Port-based VLAN classification has the lowest priority and is the most common VLAN classification mode.
- policy-based VLAN classification has the highest priority and is the least useful VLAN classification mode.

Figure 4-6 shows the process of classifying VLANs on the S7700.

Figure 4-6 Process of classifying VLANs



Default VLAN

On a switching device, every access, trunk, and hybrid port can be configured with a default VLAN. However, the meaning of default VLAN' varies with port types, as follows:

- Default VLAN of an access port
  - When receiving an untagged frame from an access port, the switching device adds a VLAN tag to the frame and sets the VID tag to match the port's PVID.
  - When sending a frame from an access port, if the VID tag matches the port's PVID, the switching device removes the VLAN tag from the frame. Ethernet frames sent from an access port to the peer device never carry VLAN tags.
- Default VLAN of a trunk port
  - When receiving an untagged frame from a trunk port, the switching device adds a VLAN tag to the frame and sets the VID tag to match the port's PVID.
  - When sending a frame from a trunk port, note the following:
    - If the VID tag matches the port's PVID, the switching device removes the tag from the frame since the PVID of each port is unique.
    - If the VID tag differs from the port's PVID, the switching device directly forwards the frame on.
- Default VLAN of a hybrid port
  - When receiving an untagged frame from a hybrid port, the switching device adds a VLAN tag to the frame and sets the VID tag to match the port's PVID.
- Default VLAN for the QinQ port
  - When a switching device receives a frame from a QinQ port, it adds a tag to the frame and sets the VID in the tag as the PVID of the QinQ port, irrespective of whether the frame carries a VLAN tag or not.
  - If the frame sent from a QinQ port contains an outer VLAN tag in which the VID is the PVID, the switching device removes the outer tag off the frame because the PVID of each port is unique.

4.4.2 Principle of VLAN Communication

Basic Principle of VLAN Communication

To improve the efficiency in processing frames, frames within a device all carry VLAN tags for uniform processing. When a data frame reaches a port of the switch, if the frame carries no VLAN tag and the port is configured with a PVID, the frame is marked with the port's PVID. If the frame has a VLAN tag, the device will not mark a VLAN tag for the frame regardless of whether the port is configured with a PVID.

The device processes frames differently according to the type of port receiving the frames. The following describes the frame processing according to the port type.

Port Type	Untagged Frame Processing	Tagged Frame Processing	Frame Transmission
Access	Accepts an untagged frame and adds a tag with the default VLAN ID to the frame.	● Accepts the tagged frame if the frame's VLAN ID matches the default VLAN ID. ● Discards the tagged frame if the frame's VLAN ID differs from the default VLAN ID.	After the PVID tag is stripped, the frame is transmitted.
Trunk	● Adds a tag with the default VLAN ID to the untagged frame and then transmits it if the default VLAN ID is permitted by the port. ● Adds a tag with the default VLAN ID to the untagged frame and then discards it if the default VLAN ID is denied by the port.	● Accepts the tagged frame if the frame's VLAN ID is permitted by the port. ● Discards the tagged frame if the frame's VLAN ID is denied by the port.	● If the frame's VLAN ID matches the default VLAN ID and the VLAN ID is permitted by the port, the switch removes the tag and transmits the frame. ● If the frame's VLAN ID differs from the default VLAN ID, but the VLAN ID is still permitted by the port, the switch will directly transmit the frame.
Hybrid	● Adds a tag with the default VLAN ID to an untagged frame and accepts the frame if the port permits the default VLAN ID. ● Adds a tag with the default VLAN ID to an untagged frame and discards the frame if the port denies the default VLAN ID.	● Accepts a tagged frame if the VLAN ID carried in the frame is permitted by the port. ● Discards a tagged frame if the VLAN ID carried in the frame is denied by the port.	If the frame's VLAN ID is permitted by the port, the frame is transmitted. The port can be configured whether or not to transmit frames with tags.
QinQ	QinQ ports are enabled with the IEEE 802.1QinQ protocol. A QinQ port adds a tag to a single-tagged frame, and supports a maximum of 4094 x 4094 VLAN tags, which meets the requirement of a Metropolitan Area Network (MAN) for the number of VLANs.		

Intra-VLAN Communication

Sometimes VLAN hosts are connected to different switches, in which case the VLAN spans multiple switches. Since ports between these switches must recognize and send packets belonging to the VLAN, the trunk link technology becomes very helpful in simplifying this solution.

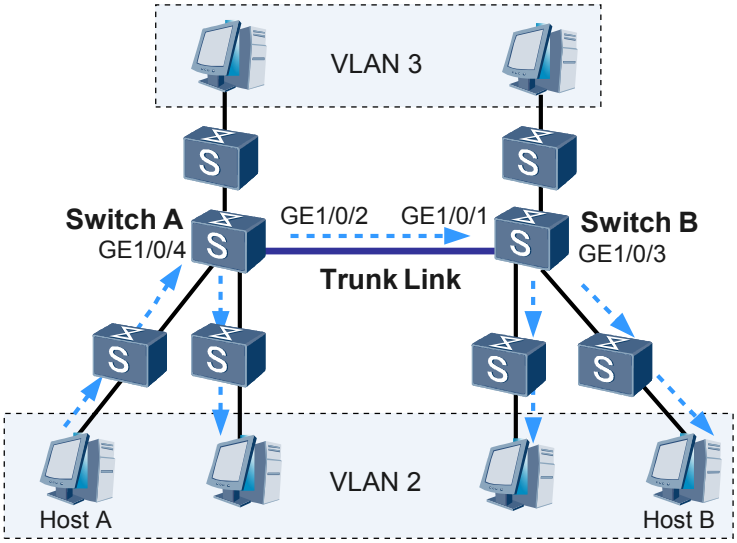
The trunk link plays the following two roles:

- Trunk line

  The trunk link transparently transmits VLAN packets between switches.
- Backbone line

  The trunk link transmits packets belonging to multiple VLANs.

Figure 4-7 Trunk link communication



As shown in **Figure 4-7**, the trunk link between Switch A and Switch B must both support the intra-communication of VLAN 2 and the intra-communication of VLAN 3. Therefore, the ports at both ends of the trunk link must be configured to belong to both VLANs. That is, GigabitEthernet1/0/2 on Switch A and GigabitEthernet1/0/1 on Switch B must belong to both VLAN 2 and VLAN 3.

Host A sends a frame to Host B in the following process:

1. The frame is first sent to GigabitEthernet1/0/4 on Switch A.
2. A tag is added to the frame on GigabitEthernet1/0/4. The VID field of the tag is set to 2, that is, the ID of the VLAN to which GigabitEthernet1/0/4 belongs.
3. Switch A queries its MAC address table for the MAC forwarding entry with the destination MAC address of Host B.
   - If this entry exists, Switch A sends the frame to the outbound interface GE 1/0/2.
   - If this entry does not exist, Switch A sends the frame to all interfaces bound to VLAN 2 except for GE 1/0/4.

4. GigabitEthernet1/0/2 sends the frame to Switch B.
5. After receiving the frame, Switch B queries its MAC address table for the MAC forwarding entry with the destination MAC address of Host B.
   - If this entry exists, Switch B sends the frame to the outbound interface GE 1/0/3.
   - If this entry does not exist, Switch B sends the frame to all interfaces bound to VLAN 2 except for GE 1/0/1.
6. GigabitEthernet1/0/3 sends the frame to Host B.

The intra-communication of VLAN 3 is similar, and is not mentioned here.

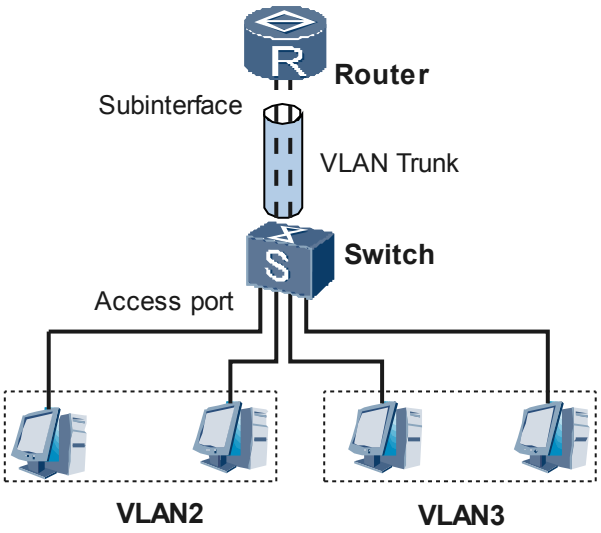
Inter-VLAN Communication

After VLANs are configured, hosts in different VLANs cannot directly communicate with each other at Layer 2. To implement communication between VLANs, you must create routes between these VLANs. The specific implementations schemes are shown as follows:

- Layer 2 switch + Router

  Generally, VLANs can communicate through the connection between the routers' Ethernet interfaces (routed Ethernet interface) and the switches' Ethernet interfaces (switched Ethernet interface), as shown in **Figure 4-8**.

  Figure 4-8 Inter-VLAN communication implemented through Layer 2 switch + Router



  Assuming that VLAN 2 and VLAN 3 are configured on the switch to communicate between VLAN 2 and VLAN 3, you need to create two sub-interfaces corresponding to VLAN 2 and VLAN 3 on the Ethernet interface of the router connected to the switch.

  Then you must enable 802.1Q encapsulation and assign IP addresses to the sub-interfaces.

  On the switch, you need to configure the Ethernet port type that connects to the router to the Trunk or Hybrid port, allowing VLAN 2 and VLAN 3 frames to pass.

  Layer 2 switch + router mode has the following shortcomings:

  - Multiple devices are needed, and the networking is complex.

- Inter-VLAN communication is implemented through a router, which is expensive and has a low transmission rate.

- Layer 3 switch

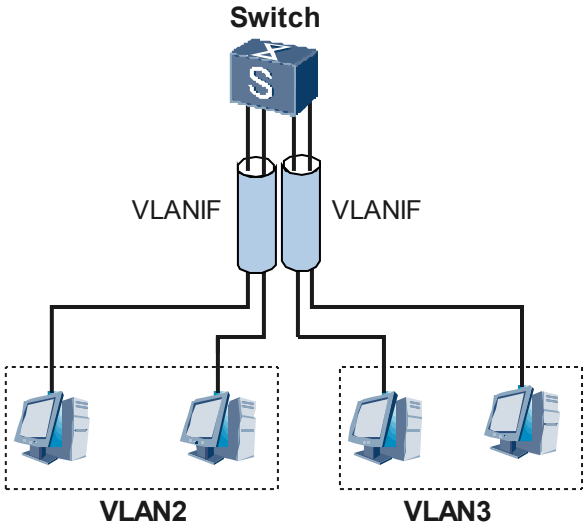
  Layer 3 switching combines routing and switching techniques to implement routing on a switch, improving the overall performance of the network. After sending the first data flow, a Layer 3 switch generates a mapping table on which it records the mapping between the MAC address and the IP address for the data flow. If the switch needs to send the same data flow again, it directly sends the data flow at Layer 2 (not Layer 3) based on the mapping table. In this manner, network delays caused by route selection are eliminated, and data forwarding efficiency is improved.

  In order for new data flows to be correctly forwarded the routing table must have the correct routing entries. Therefore, VLANIF interfaces are used to configure routing protocols on Layer 3 switches in order to reach Layer 3 routes.

  A VLANIF interface is a Layer 3 logical interface, which can be configured on either a Layer 3 switch or a router.

  As shown in **Figure 4-9**, two VLANs, VLAN 2 and VLAN 3, are configured on the switch. To implement communication between the two VLANs create two VLANIF interfaces on the switch and assign IP addresses to and configure routes for them.

Figure 4-9 Inter-VLAN communication implemented through Layer 3 switch



The Layer 3 switch scheme addresses the shortcomings in the Layer 2 switch + Router scheme, and provides faster traffic forwarding at a lower cost. Nevertheless, the Layer 3 switch has the following shortcomings of its own:

- The Layer 3 switch scheme is only applicable to networks whose interfaces are almost Ethernet interfaces.
- The Layer 3 switch scheme is only applicable to networks with stable routes and few change in network topology.

4.4.3 VLAN Aggregation

Background of VLAN Aggregation

VLAN is widely applied to switching networks because of its flexible control of broadcast domains and convenient deployment. On a Layer-3 switch, interconnection between broadcast domains is implemented using one VLAN to correspond to a singleLayer-3 logic interface. However, this can waste IP addresses. **Figure 4-10** shows a typical VLAN division in the device.

Figure 4-10 VLAN division network diagram

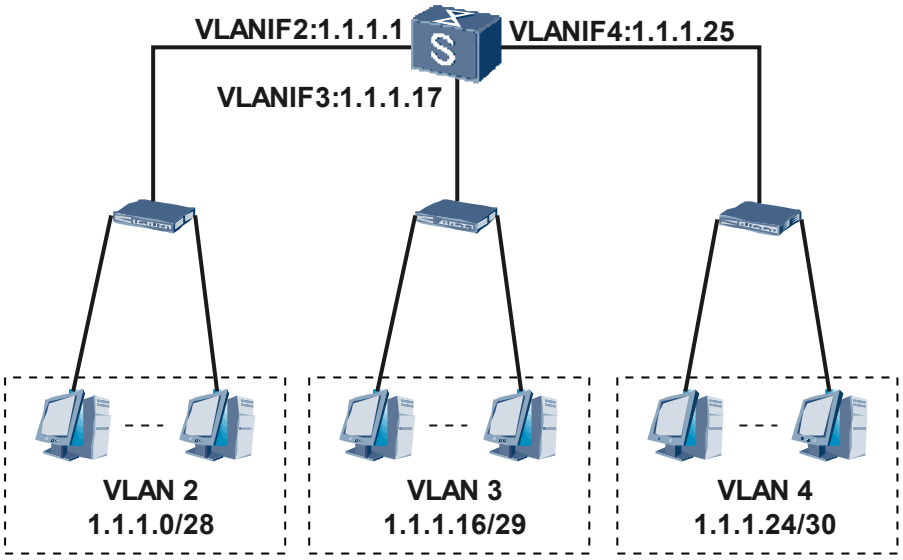


Table 4-2 Example of host address assignment on a typical VLAN

VLAN	Sub-network	Gateway Address	Number of Available Addresses	Number of Available Hosts	Practical Requirements
2	1.1.1.0/28	1.1.1.1	14	13	10
3	1.1.1.16/29	1.1.1.17	6	5	5
4	1.1.1.24/30	1.1.1.25	2	1	1

As show in **Table 4-2**, VLAN 2 requires 10 host addresses. Subnet 1.1.1.0/28 with mask length 28 bits is assigned to VLAN 2. 1.1.1.0 is the subnet address, and 1.1.1.15 is the directed broadcast address. These two addresses cannot be used as the host address. In addition, as the default address of subnet, 1.1.1.1's network gateway cannot be used as the host address. The other 13 addresses ranging from 1.1.1.2 to 1.1.1.14 can be used by the hosts. In this way, although VLAN 2 needs only ten addresses, 13 addresses need to be assigned to it according to the division of the subnet.

VLAN 3 requires five host addresses, and subnet 1.1.1.16/29 with mask length 29 bits needs to be assigned to VLAN 3. VLAN 4 requires only one address, and subnet 1.1.1.24/30 with mask length 30 bits needs to be assigned to VLAN 4.

In the above example, 16 (10+5+1) addresses are required for all the VLANs, however 28 (16 +8+4) addresses will be used according to the common VLAN addressing mode even if the optimal scheme is used. Therefore, nearly half of the addresses will be wasted. In addition, if later on VLAN 2 is accessed by only three hosts instead of ten, the extra addresses will also be wasted.

This division is inconvenient for future network upgrade and expansion. If VLAN 4 needs an additional two hosts and does not want to change the assigned IP addresses, and the addresses after 1.1.1.24 has been assigned to others, a new subnet with mask length 29 bits and a new VLAN need to be assigned to VLAN 4's new customers. As a result, VLAN 4's customers only have three hosts, but the customers are assigned to two different subnets in separate VLANs, which becomes inconvenient for network management.

In the above example, several IP addresses are used as subnet addresses, subnet directional broadcast addresses, and default addresses of subnet network gateways, meaning these IP addresses cannot be used as host addresses in the VLAN. VLAN aggregation is used to eliminate this limitation on address assignment that reduces the addressing flexibility, and wastes so many addresses.

Principle

VLAN aggregation, also known as a super-VLAN, partitions broadcast domains by using multiple VLANs in a physical network so different VLANs can belong to the same subnet. In VLAN aggregation, two basic concepts are involved, super-VLAN and sub-VLAN.

- Super-VLAN: Super-VLANs differ from common VLANs. In super-VLANs, only Layer 3 interfaces are created and physical ports are not contained. The super-VLAN can be regarded as a logical Layer 3 collection of many sub-VLANs.
- Sub-VLAN: Sub-VLANs are used to isolate broadcast domains. In sub-VLANs, only physical ports are contained and Layer 3 VLAN interfaces cannot be created. The Layer 3 switching with the external network is implemented through the super-VLAN Layer 3 interface.

A super-VLAN can contain one or more sub-VLANs each with different broadcast domains. The sub-VLAN does not occupy an independent subnet segment. In the same super-VLAN, IP addresses of hosts belong to the super-VLAN's subnet segment, regardless of the mapping between hosts and sub-VLANs.

The same Layer 3 interface is shared by sub-VLANs. Some subnet IDs, default gateway addresses of the subnet, and directed broadcast addresses of the subnet are saved; meanwhile, different broadcast domains can use the unused addresses in the same subnet segment. As a result, subnet differences are eliminated, addressing becomes flexible and previously wasted addresses can be used.

Take the **Table 4-2** to explain the implementation theory. Suppose that user demands are unchanged. In VLAN 2, 10 host addresses are demanded; in VLAN 3, 5 host addresses are demanded; in VLAN 4, 1 host address is demanded.

Create VLAN 10 and configure VLAN 10 as a super-VLAN. Then assign subnet address 1.1.1.0/24 with mask length being 24 bits to VLAN 10, where 1.1.1.0 is the subnet ID and 1.1.1.1 is the gateway address of the subnet, as shown in **Figure 4-11**. The corresponding sub-VLAN address assignment of VLAN 2, VLAN 3, and VLAN 4 is shown in **Table 4-3**.

Figure 4-11 VLAN aggregation schematic diagram

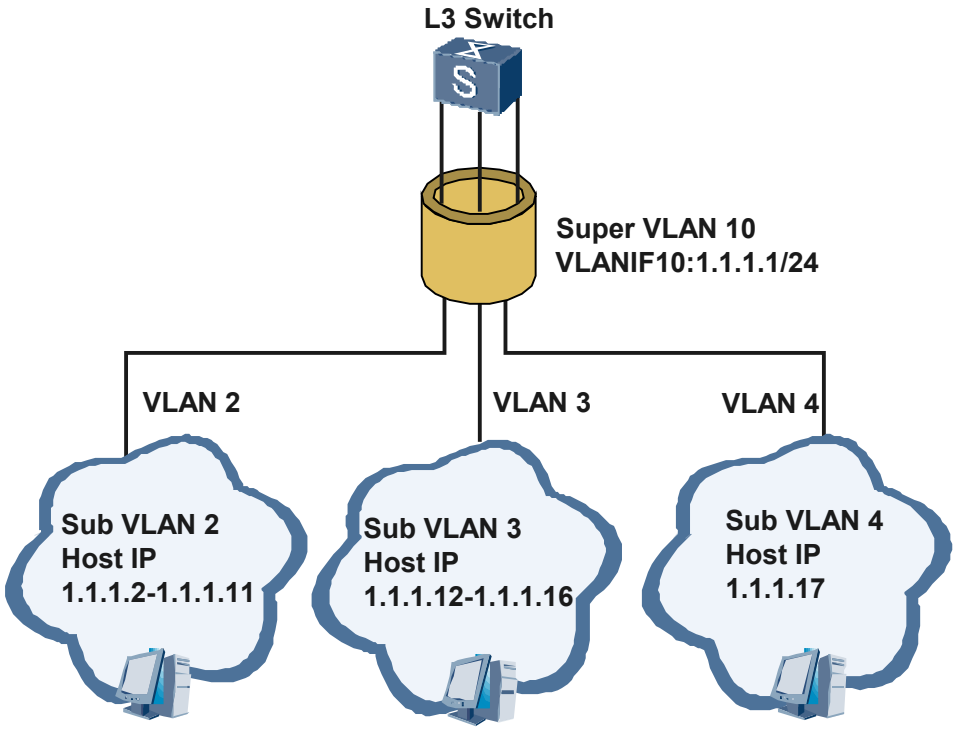


Table 4-3 Example of host address assignment in VLAN aggregation mode

VLAN	Subnet	Gateway Address	Number of Available Addresses	Available Addresses	Address Requirements
2	1.1.1.0/24	1.1.1.1	10	1.1.1.2-1.1.1.11	10
3			5	1.1.1.12-1.1.1.16	5
4			1	1.1.1.17	1

In VLAN aggregation implementation, sub-VLANs are not divided according to the previous subnet border. Instead, their addresses are flexibly assigned in the super-VLAN's subnet according to the required number of hosts.

Table 4-3 shows that VLAN 2, VLAN 3, and VLAN 4 share a subnet (1.1.1.0/24), a default gateway address of the subnet (1.1.1.1), and a directed broadcast address of the subnet (1.1.1.255). In this manner, the subnet ID (1.1.1.16, 1.1.1.24), the default gateway of the subnet (1.1.1.17, 1.1.1.25), and the directed broadcast address of the subnet (1.1.1.5, 1.1.1.23, and 1.1.1.24) can be used as host IP addresses.

In total, 16 addresses (10 + 5 + 1 = 16) are required for the three VLANs. In practice, in this subnet, a total of 16 addresses are assigned to the three VLANs (1.1.1.2 to 1.1.1.17). A total of 19 IP addresses are used, that is, the 16 host addresses together with the subnet ID (1.1.1.0), the

default gateway of the subnet (1.1.1.1), and the directed broadcast address of the subnet (1.1.1.255). In the network segment, 236 addresses (255 - 19 = 236) are available, which can be used by any host in the sub-VLAN.

Communications Between VLANs

- Introduction

  VLAN aggregation ensures that different VLANs use the IP addresses in the same subnet segment; however, this leads to the problem of Layer 3 forwarding between sub-VLANs.

  In common VLAN mode, the hosts of different VLANs can communicate with each other based on Layer 3 forwarding through their respective gateways. In VLAN aggregation mode, however, hosts in a super-VLAN use IP addresses in the same network segment and share the same gateway address. Since hosts in different sub-VLANs belong to the same subnet, they communicate with each other based on Layer 2 forwarding, not Layer 3 forwarding through a gateway. Therefore, hosts in different sub-VLANs are separated in Layer 2 and are incapable of communicating with each another. To resolve this issue, the AR2200 ARP Proxy solution ensures sub-VLANs are capable of communicating with each other.

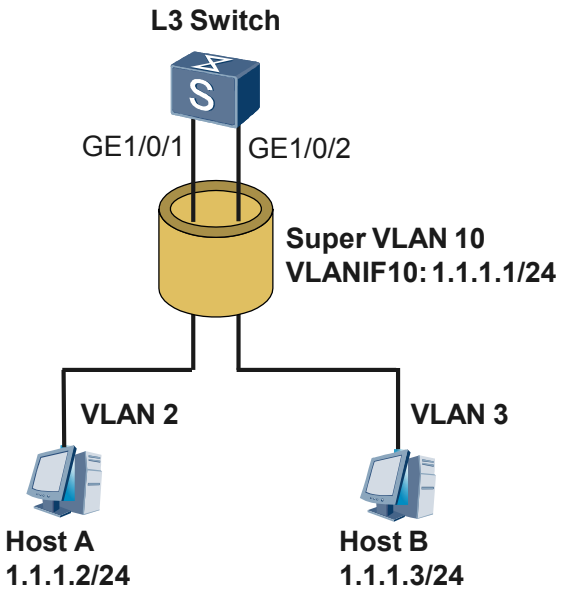
  NOTE

  For ARP proxy details, refer to the chapter "ARP" in the *S7700 Feature Description - IP Services*.

- Layer 3 Communications Between Different Sub-VLANs

  As shown in **Figure 4-12**, the super-VLAN, namely, VLAN 10, contains the sub-VLANs, namely, VLAN 2 and VLAN 3.

  Figure 4-12 Networking diagram of Layer 3 communication between different sub-VLANs based on ARP proxy



  Suppose that Host A's ARP table has no corresponding entry for Host B, and the gateway between sub-VLANs is enabled with the ARP proxy. In this case, communication between Host A in VLAN 2 and Host B in VLAN 3 proceeds as follows:

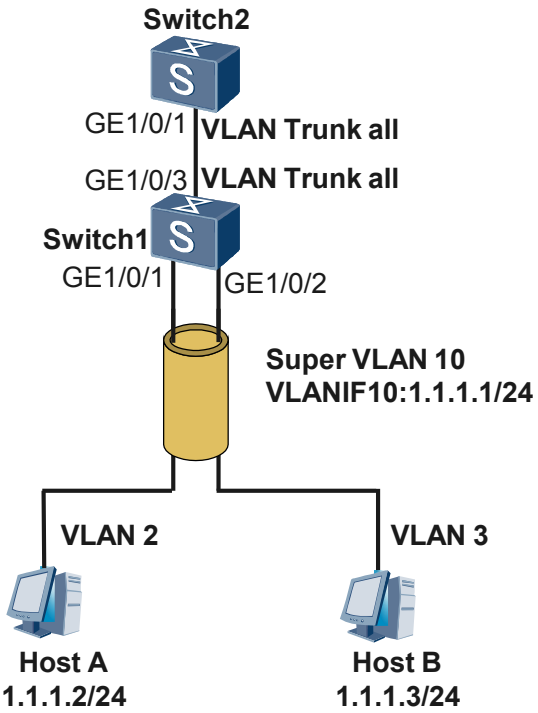
1. After comparing the IP address of Host B 1.1.1.3 with its IP address, Host A finds that both IP addresses are in the same network segment 1.1.1.0/24, and its ARP table has no entry corresponding to Host B.
2. Host A initiates an ARP broadcast to request for Host B's MAC address.
3. Host B is not in the broadcast domain of VLAN 2, and cannot receive the ARP request.
4. Since the gateway's ARP proxy is enabled between sub-VLANs, after receiving Host A's ARP request, the gateway discovers that the IP address of Host B 1.1.1.3 is the IP address of a directly-connected interface. The gateway then initiates an ARP broadcast to all other sub-VLAN interfaces to request Host B's MAC address.
5. After receiving an ARP request, Host B offers an ARP response.
6. After receiving Host B's ARP response, the gateway replies with Host A's MAC address.
7. The ARP tables in both the gateway and Host A have entries corresponding to Host B.
8. To send packets to Host B, Host A initially sends packets to the gateway, and then the gateway carries out Layer 3 forwarding.

The process that Host B uses to send packets to Host A functions in the same way.

- Layer 2 communication between a Sub-VLAN and an external network

  As shown in **Figure 4-13**, in the Layer 2 VLAN communications based on ports, the received or sent frames are not tagged with the super-VLAN ID.

Figure 4-13 Networking diagram of Layer 2 communication between a sub-VLAN and an external network



The frame that accesses Switch 1 through GE 1/0/1 on Host A is tagged with VLAN 2's ID. The VLAN ID, however, is not changed to VLAN 10's ID on Switch 1 even if VLAN

2 is the sub-VLAN of VLAN 10. After passing through GE 1/0/3 (a trunk port) this frame still carries VLAN 2's ID.

That is, Switch 1 does not send VLAN 10's frames itself.

A super-VLAN has no physical port. This limitation is obligatory, as shown below:

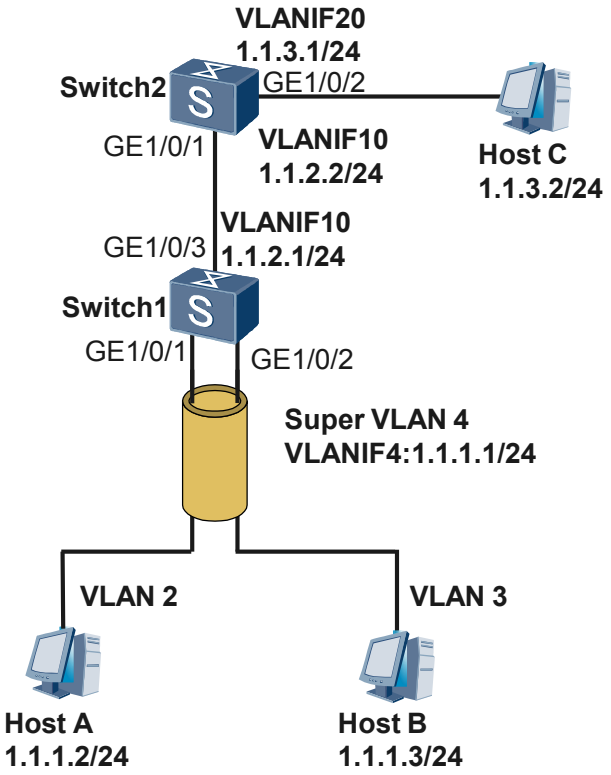
- If you configure the super-VLAN and then the trunk interface, the frames of a super-VLAN are filtered automatically according to the VLAN range set on the trunk interface.

  As shown in **Figure 4-13**, no frame of the super-VLAN 10 passes through GE 1/0/3 on Switch 1, even though the interface allows frames from all VLANs to pass through.
- If you finish configuring the trunk interface and allow all VLANs to pass through, you still cannot configure the super-VLAN on Switch 1. The root cause is that any VLAN with physical ports cannot be configured as the super-VLAN, and the trunk interface allows only the frames tagged with VLAN IDs to pass through. Therefore, no VLAN can be configured as a super-VLAN.

As for Switch 1, the valid VLANs are just VLAN 2 and VLAN 3, and all frames are forwarded in these VLANs.

- Layer 3 Communication Between a Sub-VLAN and an External Network

Figure 4-14 Networking diagram of Layer 3 communication between a sub-VLAN and an external network



As shown in **Figure 4-14**, Switch 1 is configured with super-VLAN 4, sub-VLAN 2, sub-VLAN 3, and an common VLAN 10. Switch 2 is configured with two common VLANs, VLAN 10 and VLAN 20. Suppose that Switch 1 is configured with the route to network segment 1.1.3.0/24, and Switch 2 is configured with the route to network segment 1.1.1.0/24. Then Host A in sub-VLAN 2 that belongs to the super-VLAN 4 will have to access Host C in Switch 2.

1. By comparing the IP address of Host C 1.1.3.2 with its IP address, Host A determines that two IP addresses are not in the same network segment 1.1.1.0/24.
2. Host A initiates an ARP broadcast to its gateway, requesting the gateway's MAC address.
3. After receiving the ARP request, Switch 1 identifies the correlation between the sub-VLAN and the super-VLAN, and offers an ARP response to Host A through sub-VLAN 2. The source MAC address in the ARP response packet is the MAC address of VLANIF4 for super-VLAN 4.
4. Host A learns the gateway's MAC address.
5. Host A sends the packet to the gateway, with the destination MAC address as the MAC address of VLANIF4 for super-VLAN 4, and the destination IP address of 1.1.3.2.
6. After receiving the packet, Switch 1 performs Layer 3 forwarding and sends the packet to Switch 2, with the next hop address as 1.1.2.2, the outgoing interface as VLANIF10.
7. After receiving the packet, Switch 2 performs Layer 3 forwarding and sends the packet to Host C through the directly-connected interface VLANIF20.
8. The response packet from Host C reaches Switch 1 after Switch 2 carries out Layer 3 forwarding.
9. After receiving the packet, Switch 1 performs Layer 3 forwarding and sends the packet to Host A through the super-VLAN.

4.4.4 VLAN Mapping

VLAN mapping, also known as VLAN translation, enables the user VLAN ID to be translated to the operator VLAN ID and the operator VLAN ID to be translated to the user VLAN ID.

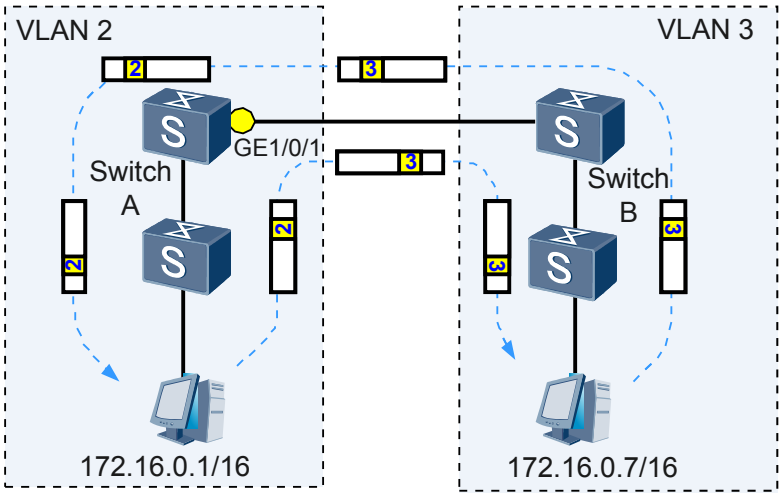
VLAN mapping occurs when frames are received by the inbound port and when frames are forwarded by the outbound port.

- 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 4-15**, VLAN mapping between VLAN 2 and VLAN 3 is configured on GE 1/0/1. When GE 1/0/1 sends frames from VLAN 2 to VLAN 3, it replaces VLAN 2 with VLAN 3. When GE 1/0/1 sends frames from VLAN 3 to VLAN 2, it replaces VLAN 3 with VLAN 2. In this manner, VLAN 2 and VLAN 3 can communicate with each other.

Figure 4-15 VLAN mapping



If devices in two VLANs need to communicate through VLAN mapping, the IP addresses of these devices must be on the same network segment. Otherwise, the devices communicate through Layer 3 routes, and VLAN mapping does not take effect.

Currently, the S7700 supports the following VLAN mapping modes:

- 1 to 1 VLAN mapping

  When the main interface configured with VLAN mapping receives a single-tagged frame, it maps the tag of the frame to the specified tag.
- 2 to 1 VLAN mapping

  When the main interface configured with VLAN mapping receives a double-tagged frame, it maps the outer tag of the frame to the specified tag and transparently transmits the inner tag as the data.
- 2 to 2 VLAN mapping

  When the main interface of the device configured with VLAN mapping receives a double-tagged frame, it maps the double tags of the frame to the specified double tags.

4.4.5 VLAN Damping

Assuming that a specific VLAN has been configured with a VLANIF interface, when the VLAN goes Down after all interfaces in the VLAN have gone Down, the VLAN reports the Down event to the VLANIF interface, changing the VLANIF interface status. To avoid network flapping due to changes in VLANIF interface status, you can enable VLAN damping on the VLANIF interface and set up a delay before the VLANIF interface goes Down.

When VLAN damping is enabled, and the VLAN's last Up interface goes Down, the Down event will be reported to the VLANIF interface only after a pre-set delay. If an interface in the VLAN goes Up during the delay, the VLANIF interface's status remains unchanged. That is, the VLAN damping function postpones when the VLAN reports a Down event to the VLANIF interface, avoiding unnecessary route flapping.

4.4.6 MUX VLAN

Multiplex VLAN (MUX VLAN) supports VLAN over network resources.

For example, on an enterprise network, enterprise employees and enterprise customers can access the enterprise server. The enterprise requires employees to communicate with each other, and customers to be isolated and unable to communicate with each other. MUX VLAN Layer 2 traffic isolation allows enterprises to meet these requirements.

Basic Concepts

As shown in **Table 4-4**, a MUX VLAN is classified into principal VLANs and subordinate VLANs; a subordinate VLAN is classified into separate VLANs and group VLANs.

Table 4-4 Classification of a MUX VLAN

MUX VLAN	VLAN Type	Associated Port	Access Authority
Principal VLAN	-	Principal port	A principal port can communicate with all ports in a MUX VLAN.
Subordinate VLAN	Separate VLAN	Separate port	A separate port can only communicate with principal ports and is isolated from other port types. A separate VLAN must be bound to a principal VLAN.
	Group VLAN	Group port	A group port can communicate with principal ports and other ports in the same group, but cannot communicate with ports in other groups or separate ports. A group VLAN must be bound to a principal VLAN.

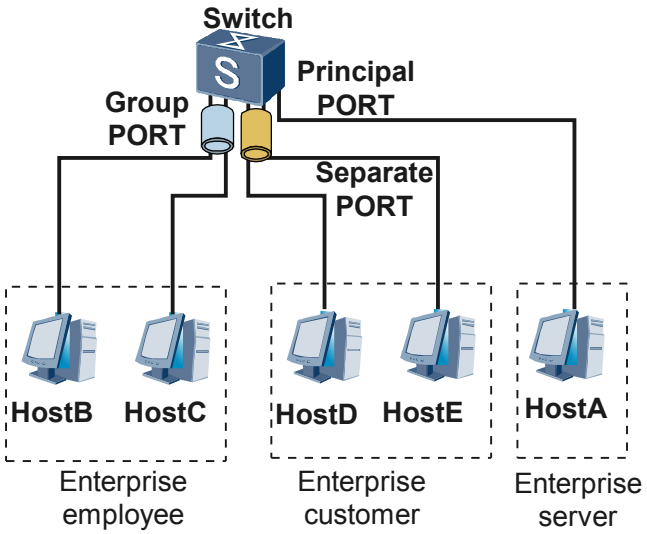
NOTE

- The principal VLAN ID cannot be applied to VLANIF interfaces, super VLANs, or sub VLANs.
- A MUX VLAN-enabled port cannot be configured with VLAN mapping or VLAN stacking.

Principle of Communication in MUX VLAN

As shown in **Figure 4-16**, the principal port connects to the enterprise server, separate ports connect to customers, and group ports connect to employees. These connections enable customers and employees to access the enterprise server, and communicate with each other. However, customers cannot communicate with each other, and customers and employees cannot communicate with each other.

Figure 4-16 Application scenario of MUX VLAN



4.4.7 VLAN Switch

VLAN switch is a forwarding technology based on VLAN tags. VLAN switch requires a pre-configured static forwarding path along the switching nodes on the network. After receiving VLAN-tagged frames that meet forwarding requirements, a switching node directly forwards the frames to corresponding interfaces by searching the VLAN switch table rather than the MAC address table. This improves forwarding efficiency and network security, and prevents MAC address attacks and broadcast storms.

The S7700 supports the following VLAN switch functions:

- Adding an outer VLAN tag (that is, the VLAN switch stack-vlan function)
- Switching the outer VLAN tags between interfaces (that is, the VLAN switch switch-vlan function)

VLAN Switch stack-vlan

Similar to VLAN stacking, VLAN switch stack-vlan is a Layer 2 technology used to encapsulate the outer VLAN tag to frames according to user VLANs. **Table 4-5** lists the comparison between VLAN stacking and VLAN switch stack-vlan.

NOTE

For VLAN stacking functions, see **VLAN Stacking**.

Table 4-5 Comparison between VLAN stacking and VLAN switch

Function	Similarity	Difference	Advantage/Disadvantage
VLAN switch stack-vlan	● You can add another VLAN tag to a received frame with an outer tag. ● Frames are processed as follows: – An interface can be configured with multiple VLANs and add different outer VLAN tags to frames from different VLANs. – When receiving a frame, the interface adds a VLAN tag to the frame; when sending a frame, the interface removes the outmost VLAN tag.	VLAN switch requires a pre-configured static forwarding path along the switching nodes on the network. After receiving VLAN-tagged frames that meet forwarding requirements, the switching node directly forwards the frames by searching the VLAN switch table rather than the MAC address table. The VLAN IDs specified in the vlan-switch command should not conflict with the global VLAN. If a specified VLAN ID has been applied in VLAN switch, the VLAN cannot be configured as the global VLAN.	● Advantage: Switching nodes can forward frames without searching the MAC address table, which improves forwarding efficiency and network security, and prevents MAC address attacks and broadcast storms. ● Disadvantage: In the case that a large number of users access a switching node, you need to configure each user in advance to establish a static forwarding path. This increases the workload of the network administrator and is inconvenient for network management.
VLAN stacking		After VLAN stacking is configured, frames are forwarded according to the MAC address table.	● Advantage: It is convenient for user access without any pre-configuration. Frames are forwarded according to the MAC address table. ● Disadvantage: Frame forwarding efficiency is low, which easily results in broadcast storms or MAC address attacks.

VLAN Switch switch-vlan

Similar to VLAN mapping, VLAN switch switch-vlan realizes communications between VLANs. **Table 4-6** lists comparison between VLAN mapping and VLAN switch.

NOTE

For VLAN mapping functions, see **VLAN Mapping**.

Table 4-6 Comparison between VLAN mapping and VLAN switch

Function	Similarity	Difference	Advantage/Disadvantage
VLAN switch switch-vlan	● After receiving VLAN-tagged frames, an interface replaces the outer VLAN tag. ● After an interface is configured with either VLAN mapping or VLAN switch, it replaces the local VLAN tag with the external VLAN tag when sending local frames to an external VLAN. ● The interface replaces the VLAN tag of the frames with the local VLAN tag when receiving frames from an external VLAN.	VLAN switch requires a pre-configured static forwarding path along the switching nodes on the network. After receiving VLAN-tagged frames that meet forwarding requirements, the switching node directly forwards the frames by searching the VLAN switch table rather than the MAC address table. The VLAN IDs specified in the vlan-switch command should not conflict with the global VLAN. If a specified VLAN ID has been applied in VLAN switch, the VLAN cannot be configured as the global VLAN.	● Advantage: Switching nodes can forward frames without searching the MAC address table, which improves forwarding efficiency and network security, and prevents MAC address attacks and broadcast storms. ● Disadvantage: In the case that a large number of users access a switching node, you need to configure each user device in advance to establish a static forwarding path. This increases the workload of the network administrator and is inconvenient for network management.

Function	Similarity	Difference	Advantage/Disadvantage
VLAN mapping		After VLAN Mapping is configured, frames are forwarded according to the MAC address table. When using VLAN Mapping to implement inter-VLAN communications, you need to ensure that the IP addresses of devices in both VLANs are in the same network segment.	● Advantage: It is convenient for user access without any pre-configuration. Frames are forwarded according to the MAC address table. ● Disadvantage: Frame forwarding efficiency is low, which easily results in broadcast storms or MAC address attacks.

4.4.8 Voice VLAN

Introduction to Voice VLAN

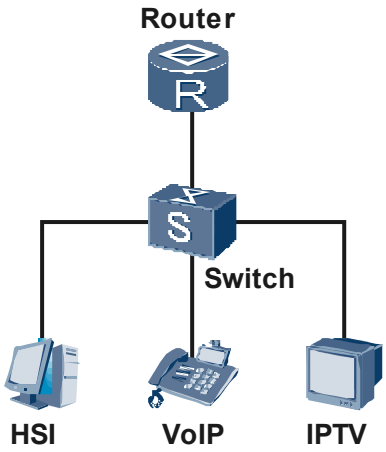
Voice data and non-voice data are often transmitted on the same network, but voice data requires a higher priority than other service data during transmission to shorten the packet delay and reduce packet loss during transmission.

In the voice VLAN system, voice data is differentiated from other data through Access Control Lists (ACLs) to ensure preferential transmission, and transmission quality is ensured through Quality of Service (QoS).

An interface enabled with voice VLAN determines whether incoming data is voice data by examining the data packet's source MAC addresses. If the source MAC address matches the Organizationally Unique Identifier (OUI), data with that source MAC address is considered voice data. The interface receiving voice data is then automatically added to the voice VLAN, effectively simplifying configurations and allowing users to manage voice data more conveniently.

As shown in **Figure 4-17**, High Speed Internet (HSI) services, Voice over IP (VoIP) services, and Internet Protocol Television (IPTV) services are all transmitted to Switch. To differentiate voice data from other data, VoIP traffic is isolated through different VLANs and is assigned a higher priority to ensure voice quality. Therefore, when voice VLAN is configured on the Switch, the Switch adds a pre-configured VLAN ID and assigns a higher priority to VoIP traffic.

Figure 4-17 Typical voice VLAN networking diagram



On different interfaces of a Switch, you can specify multiple VLANs as voice VLANs, however on an interface you can specify only one VLAN as a voice VLAN.

Basic Concepts

- OUI

  The OUI indicates a MAC address segment.

  You can perform an AND operation between a 48-bit MAC address and a mask to obtain the OUI. The length of all 1s in the mask determines the number of matched bits between a device's MAC address and the OUI. For example, if the specified MAC address is 1-1-1 and the mask is FFFF-FF00-0000, the OUI is 0001-0000-0000. In this example, if the first 24 bits of the MAC address of the device match the first 24 bits of the OUI, the interface enabled with voice VLAN considers the data from the access device as voice data, and the device as a voice device.

- Mode used when adding an interface to a voice VLAN

  Table 4-7 describes the mode in which an interface is added to a voice VLAN.

Table 4-7 Modes in which interfaces are added to voice VLANs

Mode	Description
Automatic mode	A voice VLAN-enabled interface determines whether incoming data is voice data by examining the data packet's source MAC addresses. If the source MAC address matches the OUI of a voice device, it is considered voice data. The interface receiving voice data is automatically added to a voice VLAN, and the number of such interfaces in the voice VLAN is controlled through the aging mechanism. During the aging time: ● If the switching device configured with voice VLAN does not receive any voice data from the voice device, the interface connected to the voice device will be automatically deleted from the voice VLAN. ● If the switching device configured with voice VLAN receives voice data again from the voice device, the interface connected to the voice device will be automatically added to the voice VLAN again.
Manual mode	When an interface is enabled with voice VLAN, you must manually add/remove the interface connected to the voice device to/from the voice VLAN.

Different interfaces can be added to voice VLANs in different modes, each of which are independent of each other.

- Working mode of a voice VLAN

  Table 4-8 shows the working mode of a voice VLAN.

Table 4-8 Working mode of a voice VLAN

Mode	Description	Application Scenario
Security mode	An interface enabled with voice VLAN checks whether the source MAC address of each packet entering the voice VLAN matches the OUI. ● If the source MAC address matches the OUI, the packet enters the voice VLAN and is forwarded. ● If the source MAC address does not match the OUI: – The packet is forwarded through a specified VLAN if the interface enabled with voice VLAN allows other common VLAN packets to pass through. – The packet is discarded if the interface enabled with voice VLAN does not allow other common VLAN packets to pass through.	Security mode is used when multiple services (HSI, VoIP, and IPTV) are accessed on a Layer 2 network through a single interface, and the interface transmits only voice data. The security mode can protect the voice VLAN against the attacks by invalid packets, but checking packets occupies certain system resources.
Normal mode	The interface enabled with voice VLAN can transmit both voice data and service data, and is vulnerable to attacks by invalid packets.	The normal mode is used when multiple services (HSI, VOIP, and IPTV) are transmitted to a Layer 2 network through one interface, and the interface transmits both voice data and service data.

- Aging time of a voice VLAN

  In automatic mode, the device configured with voice VLAN automatically adds the interface that connects to a voice device to the voice VLAN after learning the source MAC address of the voice data sent from the voice device, and controls the number of the interfaces in the voice VLAN through the aging mechanism.

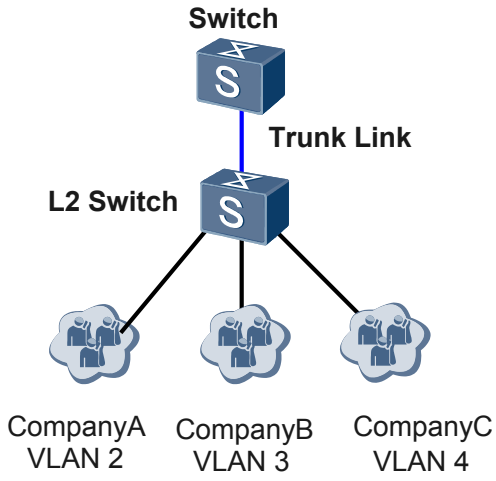
  When the aging time expires, if the interface enabled with voice VLAN does not receive any voice data from the voice device, the interface that connects to the voice device will be deleted from the voice VLAN. If the interface enabled with voice VLAN receives voice data again from the voice device, the interface that connects to the voice device will be automatically added to the voice VLAN again.

  In manual mode, the voice VLAN is not affected by the aging time.

4.5 Application

Port-Based VLAN Division

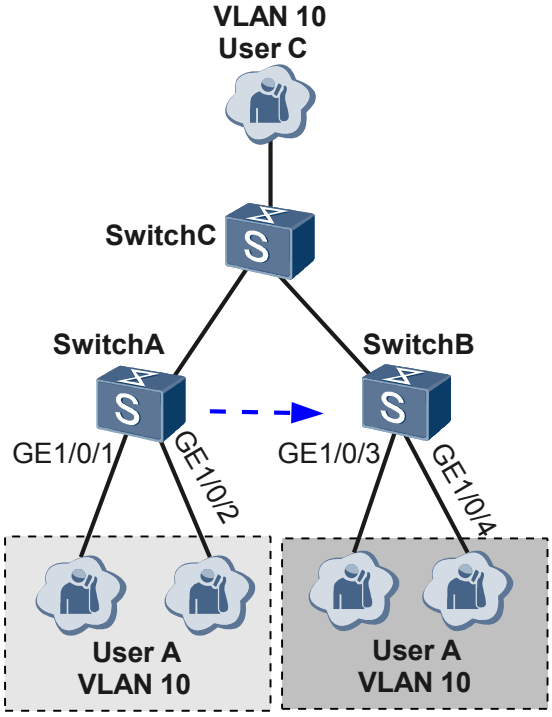
Figure 4-18 Networking diagram of Port-based VLAN division



Different companies residing in the same business premises may need to isolate service data from each other. Therefore, according to the port requirement of each company, VLANs are created on the core switch of the business premise, and ports of each company are assigned into the corresponding VLAN. This ensures that each company can have a "virtual switch" or say a "virtual workstation".

MAC Address-Based VLAN Classification

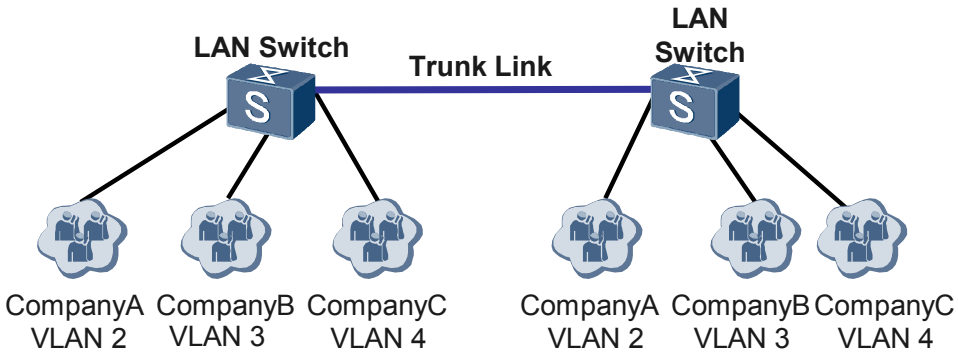
Figure 4-19 Networking diagram of MAC address-based VLAN classification



As shown in **Figure 4-19**, User A is initially connected to Switch A. Now, it is required that User A be connected to Switch B. To ensure that User A can still communicate with User C, you can configure the classification of VLANs based on MAC addresses on Switch C. As long as the MAC address of User A remains unchanged, no configuration needs to be changed for User A to communicate with User C.

Application of VLAN Trunk

Figure 4-20 Networking diagram of VLAN trunk application



A company may have departments scattered in different business premises. In such a situation, the trunk link can be utilized to interconnect core switches of different business premises, In this manner, data of different companies can be isolated, and the inter-department communication within the company can be implemented.

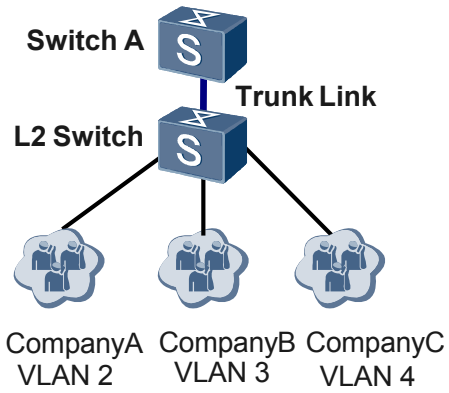
Application of Inter-VLAN Communication

Inter-VLAN communication ensures that different companies can communicate with each other.

The inter-VLAN communication can be classified into two types, as shown as follows:

- Multiple VLANs belongs to the same Layer 3 device.

Figure 4-21 Networking diagram of communications between multiple VLANs on the same Layer 3 device

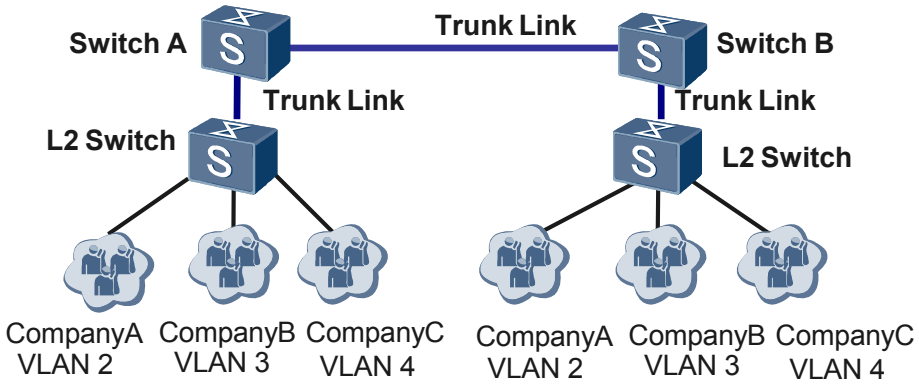


As shown in **Figure 4-21**, if VLAN 2, VLAN 3, and VLAN 4 only belong to Switch A, these VLANs are not VLANs across different switches. In such a situation, you can configure a VLANIF interface for each VLAN on Switch A to implement the communications between these VLANs.

The Layer 3 device shown in **Figure 4-21** can be a router or a Layer 3 switch.

- Multiple VLANs belongs to different Layer 3 devices.

Figure 4-22 Networking diagram of communications between multiple VLANs on different Layer 3 devices

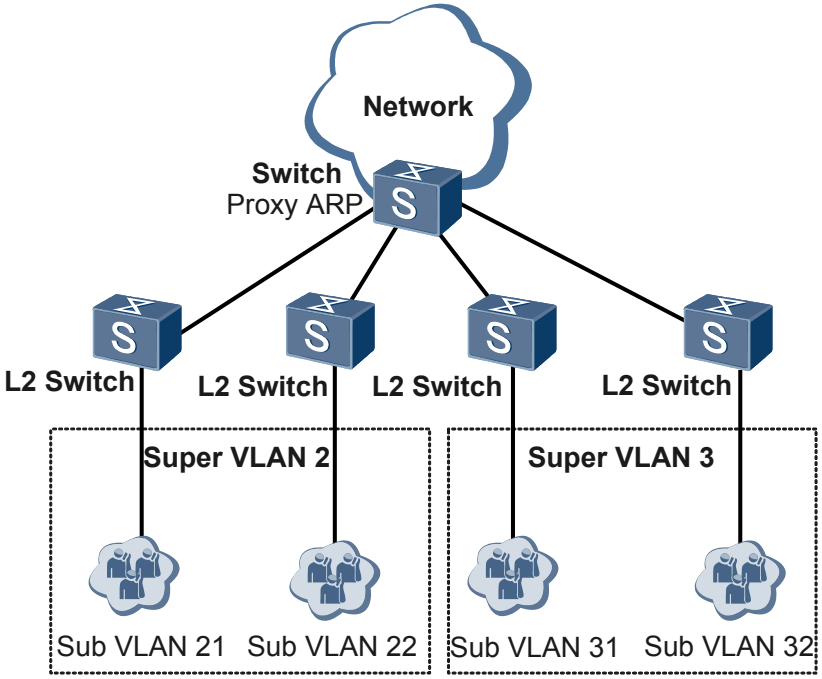


As shown in **Figure 4-22**, VLAN 2, VLAN 3, and VLAN 4 are VLANs across different switches. In such a situation, you can configure a VLANIF interface respectively on Switch A and Switch B for each VLAN, and then configure the static route or run a routing protocol between Switch A and Switch B.

The Layer 3 device shown in **Figure 4-22** can be a router or a Layer 3 switch.

Application of VLAN Aggregation

Figure 4-23 Networking diagram of VLAN aggregation application



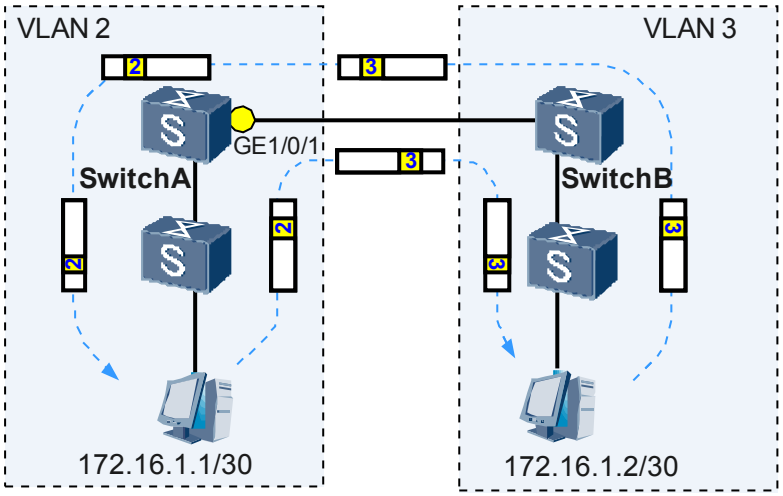
As shown in **Figure 4-23**, four VLANs, namely, VLAN 21, VLAN 22, VLAN 31, and VLAN 32, are configured. If these VLANs need to communicate with each other, you should configure an IP address for each VLAN on the Switch.

As an alternative, you can enable VLAN aggregation to aggregate VLAN 21 and VLAN 22 into super VLAN 2, and VLAN 31 and VLAN 32 into super VLAN 3. In this manner, you can save IP addresses by only assigning IP addresses to the super VLANs.

After ARP proxy is configured on Switch, the sub-VLANs in each super VLAN can communicate with each other.

VLAN Switch

Figure 4-24 Networking diagram of inter-VLAN communications through VLAN Switch

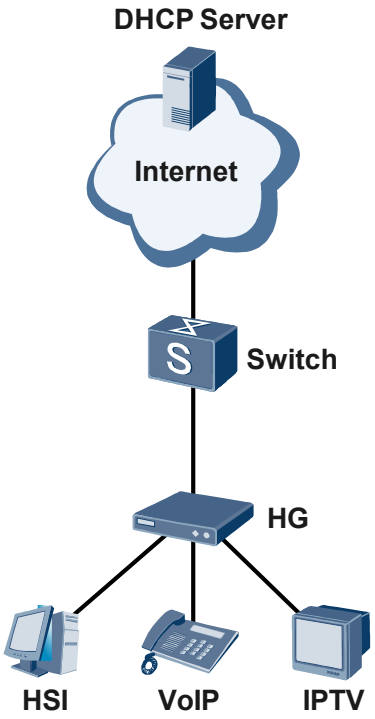


As shown in **Figure 4-24**, after the VLAN Switch switch-vlan function is configured on Switch A, frames are forwarded along the specified path. When Switch A receives frames tagged with VLAN 2, VLAN 2 is replaced with VLAN 3 on GE 1/0/1, and then the frames are sent to PCs in VLAN 3.

When Switch A receives frames tagged with VLAN 3, VLAN 3 is replaced with VLAN 2 on GE 1/0/1, and then the frames are sent to PCs in VLAN 2. In this manner, PCs in VLAN 2 and VLAN 3 can communicate with each other.

Application of Voice VLAN

Figure 4-25 Networking diagram of Voice VLAN application



As shown in **Figure 4-25**, terminals of the High Speed Internet (HSI), Voice over IP (VoIP), and Internet Protocol Television (IPTV) services are connected to Switch. Users require high quality of calls, so voice data flows must have a high priority.

To ensure high quality of calls, you can configure a voice VLAN on Switch.

After the voice VLAN is configured, Switch checks source MAC addresses of incoming data flows. If the source MAC address of a data flow matches the OUI address configured for voice devices, Switch considered the flow as a voice data flow. Switch changes the priority of voice data flows and transmit them in the voice VLAN. The call quality is thus ensured.

4.6 Terms and Abbreviations

Abbreviations

Abbreviation	Full Spelling
VLAN	Virtual Local Area Network
PVID	Port Default VLAN ID

5 QinQ

About This Chapter

5.1 Introduction to QinQ

Definition

The 802.1Q-in-802.1Q (QinQ) technology improves the utilization of VLANs by adding another 802.1Q tag. In this manner, services in the private VLAN can be transparently transmitted to the public network. The packet transmitted in the backbone network carries double 802.1Q tags (a public VLAN tag and a private VLAN tag), that is, 802.1Q-in-802.1Q. It is also called the QinQ protocol.

Purpose

As the metro Ethernet is widely used, the application of 802.1Q VLANs is restricted in terms of isolating and identifying users. The 12-bit VLAN tag defined in IEEE 802.1Q identifies only a maximum of 4096 VLANs, which are insufficient for massive users in the metro Ethernet. The QinQ technology is developed to solve this problem.

QinQ was originally designed to expand the number of VLANs by adding an 802.1Q tag to an 802.1Q packet. With this extra tag, the number of VLANs is increased to 4096 x 4096.

As the metro Ethernet grows and the refined operation requires, double tags of QinQ can be applied in other scenarios. The inner tag indicates the user; the outer tag indicates the service. In addition, when QinQ packets that carry double tags traverse the Internet Service Provider (ISP) network, the inner tag is transmitted transparently. Such an implementation mode can also be regarded as a simple and practical VPN technology. Therefore, QinQ extends services of a core MPLS VPN in the metro Ethernet, and thus the end-to-end VPN is formed.

Since the QinQ technology is easy-to-use, it has been widely applied in the ISP network. For example, it is used in conjunction with multiple services in the metro Ethernet. The introduction to selective QinQ (VLAN stacking) makes QinQ more popular among ISPs. It can isolate different user VLANs and public VLANs and save VLAN resources of the ISP network to the maximum extent. As the metro Ethernet develops, different vendors propose their own metro Ethernet solutions. In virtue of simplicity and flexibility, QinQ plays important roles in metro Ethernet solutions.

5.2 References

The references of this feature are as follows:

Document	Description	Remarks
IEEE 802.1q	IEEE standard for local and metropolitan area networks: Virtual Bridged Local Area Networks	-
IEEE 802.1ad	IEEE 802.1ad, "Virtual Bridged Local Area Networks:Provider Bridges"	-

5.3 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

5.4 Principles of QinQ

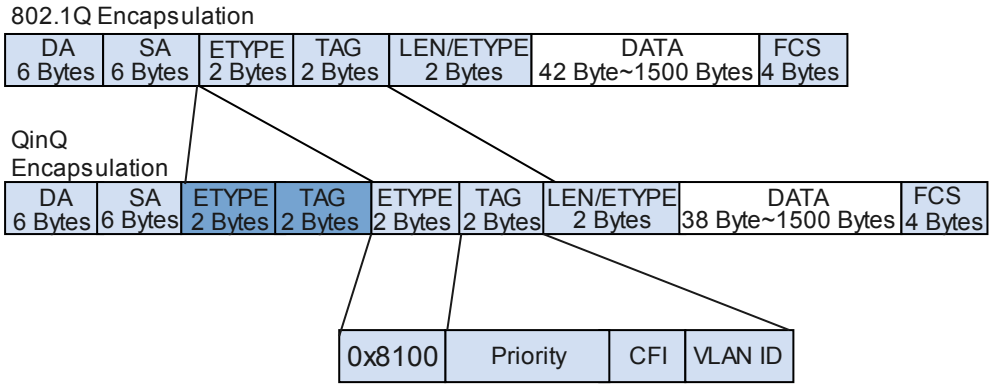
5.4.1 Principle

QinQ is a technology used to expand the VLAN space by adding an 802.1Q VLAN tag to an 802.1Q packet. To adapt to the development of the metro Ethernet, QinQ encapsulation and termination modes become diversified. Meanwhile, QinQ is further applied in ISP's refined operation.

Format of QinQ Packet

The QinQ packet is of fixed format. That is, the packet is added with another 802.1Q tag in addition to an 802.1Q tag. QinQ packets have more 4 bytes than common 802.1Q packets.

Figure 5-1 802.1Q encapsulation



QinQ Encapsulation

QinQ encapsulation is to convert an 802.1Q packet with a single tag into a QinQ packet with double tags. Encapsulation is mainly implemented on the switched port of the UPE.

According to encapsulation modes, QinQ encapsulation can be divided into port-based QinQ encapsulation, traffic-based QinQ encapsulation, and special QinQ encapsulation on the route sub-interface. The details are as follows:

- Port-based QinQ encapsulation

  Port-based encapsulation, also referred to as the QinQ tunnel, means that all traffic entering one port is encapsulated with an outer tag. This encapsulation mode is inflexible and fails to distinguish users and services specifically.

- Traffic-based QinQ encapsulation

  Traffic-based QinQ encapsulation means that the device classifies the traffic entering one port and then decides whether to encapsulate traffic with the outer tag and which outer tag is encapsulated. Therefore, this encapsulation mode is also referred to as selective QinQ.

  When a user uses different VLAN IDs for different services, traffic can be classified according to the VLAN ID range. For example, the VLAN ID for PC access ranges from 101 to 200; the VLAN ID for IPTV services ranges from 201 to 300; the VLAN ID for VIP ranges from 301 to 400. After receiving user package, the UPE encapsulates traffic of PC access with outer tag 100, traffic of IPTV services with outer tag 300, and traffic of VIP with outer tag 500.

- QinQ encapsulation on the route sub-interface

  In general, QinQ encapsulation is performed on the switched port. In a special situation, however, QinQ encapsulation can be performed on the route sub-interface.

  When user package is transmitted transparently over the MPLS/IP core network by PWE3/VLL/VPLS, the route sub-interface on the NPE can encapsulate packets with the user VLAN ID and access VLL/PWE3 through the outer VLAN. In this mode, services of multiple user VLANs can be transmitted transparently through one sub-interface, which is called a QinQ stacking sub-interface.

  The encapsulation is traffic-based encapsulation, but the QinQ stacking sub-interface can be integrated into L2VPN (PWE3/VLL/VPLS) only; otherwise, it does not support Layer 3 forwarding.

Sub-interface for QinQ/Dot1q VLAN Tag Termination

QinQ termination refers to identifying one tag or double tags of QinQ packets and then stripping one tag or double tags or sending the packets according to the subsequent forwarding.

When the QinQ technology is applied in the MPLS/IP core network, different termination methods are used in different situations.

QinQ termination is usually conducted on the route sub-interface, that is, the sub-interface for QinQ/dot1q VLAN tag termination.

- The route sub-interface that terminates a single tag is called a sub-interface for dot1q VLAN tag termination.
- The route sub-interface that terminates double tags is called a sub-interface for QinQ VLAN tag termination.

According to the values of terminated VLAN tags, the sub-interface for QinQ VLAN tag termination is classified into the following types:

- Explicit QinQ termination sub-interface: Double VLAN tags specify two VLANs.
- Implicit QinQ termination sub-interface: Double VLAN tags specify two ranges of VLANs.

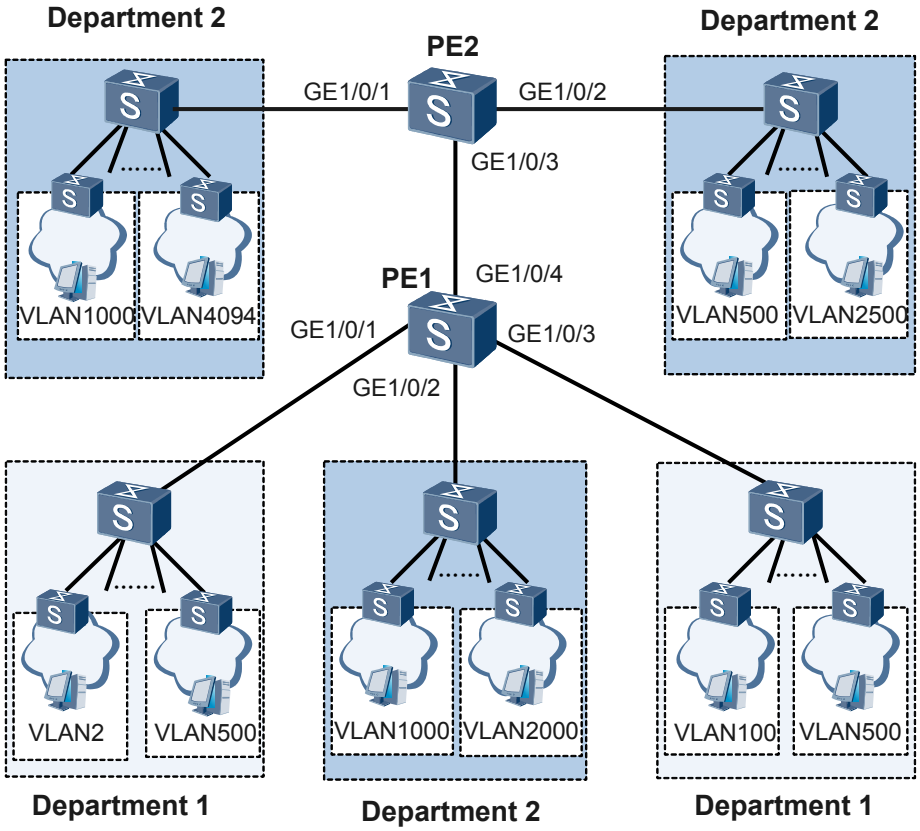
The detailed implementation and function of the sub-interface for QinQ VLAN tag termination are related with the specific scenario. The following explains it in different scenarios.

5.4.2 QinQ Tunnel

When multiple VLANs are required, the QinQ tunnel need be configured. The QinQ tunnel, by adding an outer tag to the VLAN tag, extends the range of available number of VLANs. In QinQ mode, packets with double tags are transmitted in the tunnel.

In the network as shown in **Figure 5-2**, department 1 has two offices and department 2 has three offices; offices of department 1 and department 2 connect to PE1 and PE2 respectively. Department 1 and department 2 can plan their own VLANs as desired.

Figure 5-2 Networking diagram of the QinQ tunnel



It is required to configure the QinQ tunnel on PE1 and PE2. Thus, office networks in department 1 or department 2 can interwork but office networks of department 1 cannot interwork with office networks of department 2.

- On PE1, the user packets entering GE1/0/1 and GE1/0/3 are encapsulated with outer VLAN 10 and the user packets entering GE 1/0/2 are encapsulated with outer VLAN 20.
- On PE2, the user packets entering GE1/0/1 and GE1/0/2 are encapsulated with outer VLAN 20.

- GE1/0/4 on PE1 and GE1/0/3 on PE2 allow the packets tagged with VLAN 20 to pass.

Table 5-1 shows planning of outer VLAN tags of department 1 and department 2.

Table 5-1 Planning of outer VLAN tags of department 1 and department 2

Department Name	VLAN ID Range	Outer VLAN ID
Department 1	2 to 500	10
Department 2	500 to 4094	20

5.4.3 Layer 2 Selective QinQ

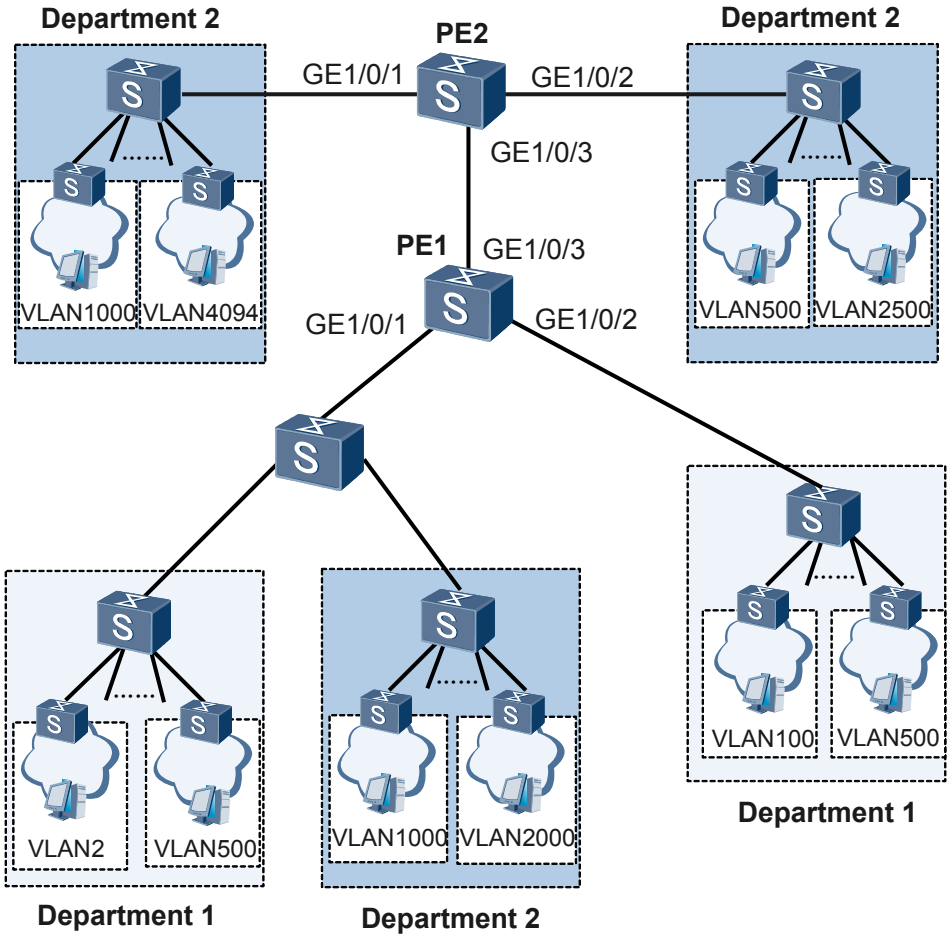
Layer 2 selective QinQ is an extension of the QinQ tunnel. Layer 2 selective QinQ is more flexible than the QinQ tunnel. The major difference is as follows:

- The QinQ tunnel attaches the same outer tag to all the frames entering the Layer 2 QinQ interface.
- Layer 2 selective QinQ attaches different outer tags to the frames entering the Layer 2 QinQ interface according to different inner tags.

As shown in Figure 5-3, department 1 and department 2 have many offices.

- VLAN 2 to VLAN 500 are used in the networks of department 1.
- VLAN 500 to VLAN 4094 are used in the networks of department 2.
- GE1/0/1 on PE1 receives the packets from different VLANs of department 1 and department 2 simultaneously.

Figure 5-3 Typical networking diagram of Layer 2 selective QinQ



It is required to configure Layer 2 selective QinQ on PE1 and PE2. Office networks in department 1 or department 2 can interwork but office networks of department 1 cannot interwork with office networks of department 2.

- Table 5-2 shows the planning of outer VLAN tags in the packets entering different interfaces on PE1 and PE2.

Table 5-2 Planning of outer VLAN tags on PE1 and PE2

Device Name	Interface Name	VLAN ID Range	Outer VLAN ID
PE1	GE1/0/1	2 to 500	10
	GE1/0/1	1000 to 2000	20
	GE1/0/2	100 to 500	10
PE2	GE1/0/1	1000 to 4094	20
	GE1/0/2	500 to 2500	20

- GE1/0/3 on PE1 or PE2 allows the packets tagged with VLAN 20 to pass.

5.4.4 VLAN Stacking

VLAN stacking is a Layer 2 technology that encapsulates different outer VLAN tags for different VLANs.

In the carrier network-accessing environment, user packets usually need to be differentiated according to user's applications, access points, or access devices. Therefore, VLAN stacking is adopted to realize packet differentiation by adding different outer VLAN tags to user packets according to the inner VLAN tags, IP addresses, or MAC addresses of these packets.

The VLAN stacking port has the following features:

- The VLAN stacking port can be configured with multiple outer VLAN tags. Then different outer VLAN tags can be assigned to different VLAN frames.
- The VLAN stacking port can add outer VLAN tags to the received frames, and strip the outer VLAN tags from the sent frames.

5.4.5 QinQ Mapping

Principle of QinQ Mapping

QinQ mapping occurs between when frames are received by an inbound interface and when frames are forwarded by an outbound interface.

- When sending a local VLAN frame to the external VLAN, the sub-interface replaces the VLAN tag of the frame with the VLAN tag of the external VLAN.
- When receiving an external VLAN frame, the port replaces the VLAN tag of the frame with the VLAN tag of the local VLAN.

In actual networking applications, QinQ mapping can be used to map the C-VLAN tag to the S-VLAN tag so that different C-VLAN tags are shielded.

QinQ mapping is often deployed on edge devices of an ME network to map the C-VLAN tag carried in a frame to the S-VLAN tag before the frame is transmitted on the public network. QinQ mapping can be applied but not limited to the following scenarios:

- The VLAN IDs deployed at new sites and old sites conflict, but new sites need to communicate with old sites.
- The VLAN ID planning of each site on the public network is different. As a result, the VLAN IDs conflict. The sites, however, do not need to communicate.
- The VLAN IDs on both ends of the public network are different.

Currently, the S7700 supports the following mapping modes:

- 1 to 1 QinQ mapping

  When a sub-interface of the device configured with QinQ mapping receives a single-tagged frame, it maps the tag of the frame to the specified tag.

- 1 to 2 QinQ mapping

  When a sub-interface of the device configured with QinQ mapping receives a single-tagged frame, it maps the tag of the frame to the specified double tags.

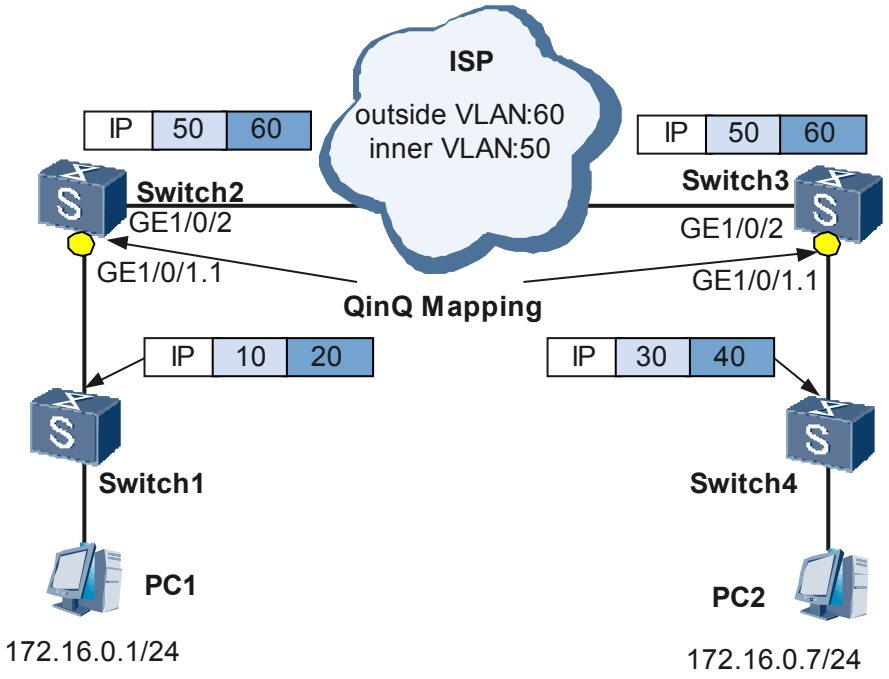
- 2 to 1 QinQ mapping

When a sub-interface of the device configured with QinQ mapping receives a double-tagged frame, it maps the double tags of the frame to the specified tag.

- 2 to 2 QinQ mapping

  When a sub-interface of the device configured with QinQ mapping receives a double-tagged frame, it maps the double tags of the frame to the specified double tags.

Figure 5-4 QinQ mapping



As shown in **Figure 5-4**, 2 to 2 QinQ mapping is configured on GE 1/0/1.1 of Switch 2 and Switch 3. In this example, PC1 sends a frame to PC2.

1. GE 1/0/2 of Switch 2 sends a double-tagged frame with outer VLAN 60 and inner VLAN 50.
2. The frame sent by Switch 2 is transparently transmitted over the ISP network.
3. After receiving the double-tagged frame sent by Switch 2, GE 1/0/1.1 of Switch 3 replaces outer VLAN 60 with outer VLAN 40 and inner VLAN 50 with inner VLAN 30.

PC2 sends a frame to PC1 in the same manner.

PC1 can thus communicate with PC2.

Comparison Between QinQ Mapping and VLAN Mapping

Table 5-3 lists the comparison between QinQ mapping and VLAN mapping.

Table 5-3 Comparison between QinQ mapping and VLAN mapping

Mapping Type	Similarity	Difference
1 to 1	An interface maps the tag of a received single-tagged frame to the specified tag.	● QinQ mapping – The mapping is performed on the sub-interface. – QinQ mapping is mainly used for VPLS access. ● VLAN mapping – The mapping is performed on the main interface. – VLAN mapping is mainly applied to Layer 2 networks where VLAN frames are forwarded.
1 to 2	An interface maps the tag of a received single-tagged frame to the specified double tags.	● QinQ mapping – The mapping is performed on the sub-interface. – QinQ mapping is mainly used for VPLS access. ● VLAN mapping This mapping mode is not supported.
2 to 1	The inbound interface receives double-tagged frames.	● QinQ mapping – The mapping is performed on the sub-interface. – The sub-interface maps the double tags of the frame to the specified tag. – QinQ mapping is mainly used for VPLS access. ● VLAN mapping – The mapping is performed on the main interface. – The main interface maps the outer tag of the received double-tagged frame to the specified tag and transparently transmits the inner tag as the data. – VLAN mapping is mainly applied to Layer 2 networks where VLAN frames are forwarded.

Mapping Type	Similarity	Difference
2 to 2	The interface maps double tags of the received double-tagged frame to the specified double tags.	● QinQ mapping – The mapping is performed on the sub-interface. – QinQ mapping is mainly used for VPLS access. ● VLAN mapping – The mapping is performed on the main interface. – VLAN mapping is mainly applied to Layer 2 networks where VLAN frames are forwarded.

5.4.6 IP Forwarding on the Termination Sub-interface

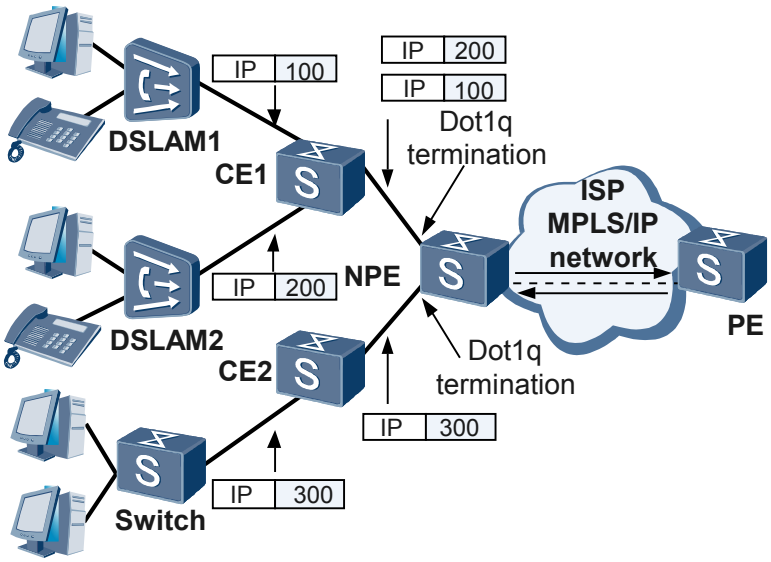
As shown in **Figure 5-5** and **Figure 5-6**, when the NPE at the edge of the MPLS/IP core network acts as a gateway device of users, the termination sub-interface supports IP forwarding.

Whether the sub-interface for dot1q VLAN tag termination or the sub-interface for QinQ VLAN tag termination supports IP forwarding depends on the packet received by the NPE:

- If one tag is contained in the packet, the sub-interface for dot1q VLAN tag termination supports IP forwarding.
- If double tags are contained in the packet, the sub-interface for QinQ VLAN tag termination supports IP forwarding.

IP Forwarding on the Sub-interface for Dot1q VLAN Tag Termination

Figure 5-5 Networking diagram of IP forwarding on the sub-interface for dot1q VLAN tag termination

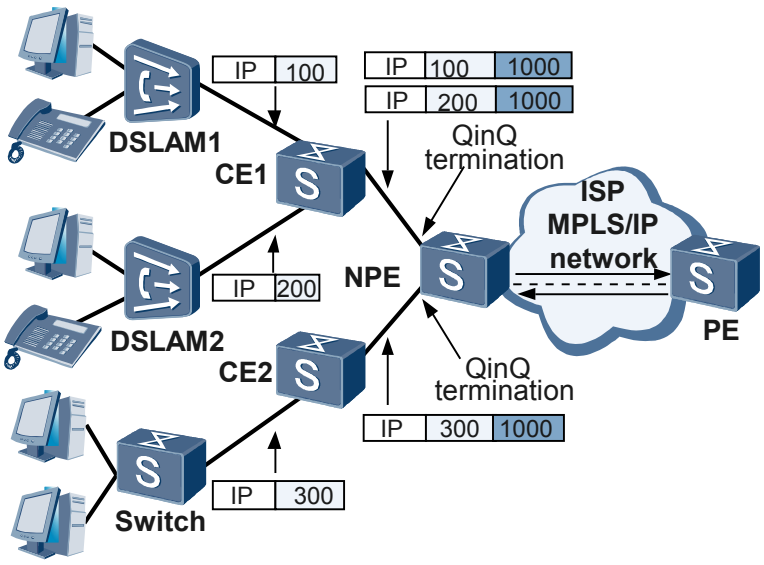


The sub-interface for dot1q VLAN tag termination first identifies the outer VLAN tag and then generates an ARP entry containing the IP address, MAC address, and outer VLAN tag.

- For the upstream traffic, the termination sub-interface strips the MAC address and the outer VLAN tag, and searches the routing table to perform Layer 3 forwarding according to the destination IP address.
- For the downstream traffic, the termination sub-interface encapsulates IP packets with the MAC address and outer VLAN tag according to ARP entries and then send IP packets to the target user.

IP Forwarding on the Sub-interface for QinQ VLAN Tag Termination

Figure 5-6 Networking diagram of IP forwarding on the sub-interface for QinQ VLAN tag termination



The sub-interface for QinQ VLAN tag termination first identifies double VLAN tags and then generates an ARP entry containing the IP address, MAC address, and double VLAN tags.

- For the upstream traffic, the termination sub-interface strips the MAC address and double VLAN tags, and searches the routing table to perform Layer 3 forwarding according to the destination IP address.
- For the downstream traffic, the termination sub-interface encapsulates IP packets with the MAC address and double VLAN tags according to ARP entries and then sends IP packets to the target user.

5.4.7 ARP Proxy on the Termination Sub-interface

As shown in **Figure 5-7** and **Figure 5-8**, the termination sub-interface supports a range of VLANs instead of a VLAN. Thus, users on the same network segment but in different VLANs fail to communicate at Layer 2 without the help of IP forwarding. Therefore, the termination sub-interface needs to support ARP proxy.

Whether the sub-interface for dot1q VLAN tag termination or the sub-interface for QinQ VLAN tag termination supports ARP proxy depends on the packet received by the PE:

- If one tag is contained in the packet, the sub-interface for dot1q VLAN tag termination supports ARP proxy.
- If double tags are contained in the packet, the sub-interface for QinQ VLAN tag termination supports ARP proxy.

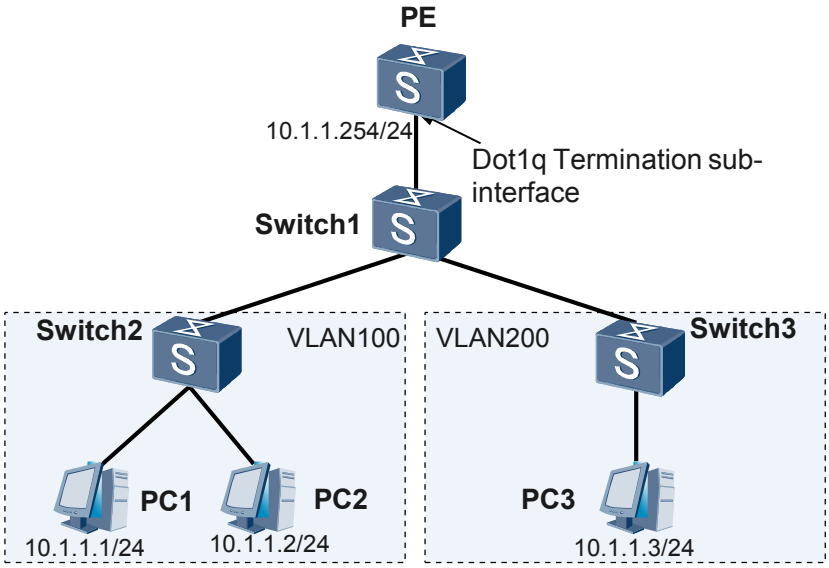
ARP Proxy on the Sub-interface for Dot1q VLAN Tag Termination

As shown in **Figure 5-7**, PC1 and PC2 belong to VLAN 100; PC3 belongs to VLAN 200; Switch 1 is an ordinary Layer 2 switch, which allows any VLAN packet to pass; PC1, PC2, and PC3 are on the same network segment.

When PC1 and PC3 intend to communicate with each other, PC1 sends an ARP request to PC3 since PC1 and PC3 are on the same network segment. However, PC1 and PC3 are in different VLANs, so PC3 fails to receive the ARP request from PC1.

This problem can be solved by enabling ARP proxy on the sub-interface for dot1q VLAN tag termination.

Figure 5-7 ARP proxy on the sub-interface for dot1q VLAN tag termination



ARP Proxy on the Sub-interface for QinQ VLAN Tag Termination

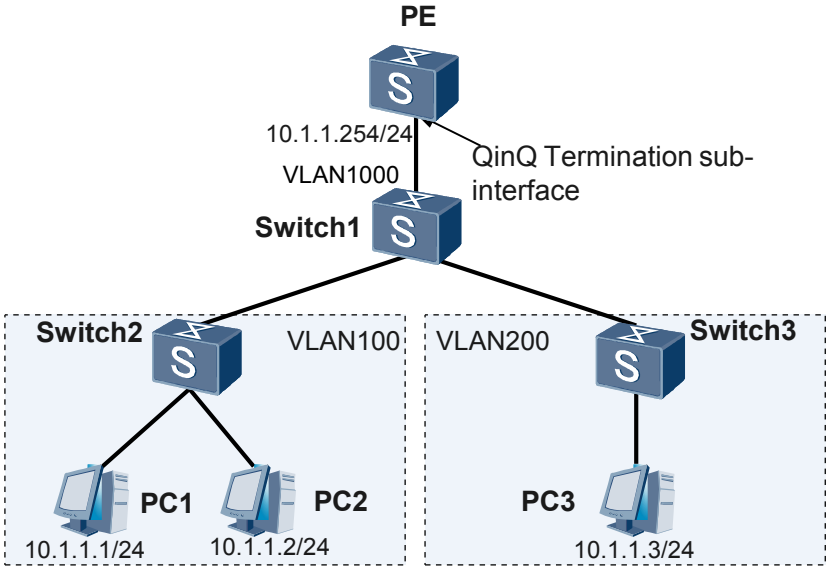
The sub-interface for QinQ VLAN tag termination supports a range of VLANs instead of a VLAN. Thus, users on the same network segment but in different VLANs fail to communicate at Layer 2 without the help of IP forwarding. Therefore, the sub-interface for QinQ VLAN tag termination needs to support ARP proxy.

As shown in **Figure 5-8**, PC1 and PC2 belong to VLAN 100; PC3 belongs to VLAN 200; Switch 1 is enabled with selective QinQ and attaches outer VLAN tag 1000 to the packets that are sent from Switch 2 and Switch 3 to the PE; PC1, PC2, and PC3 are on the same network segment.

When PC1 and PC3 intend to communicate with each other, PC1 sends an ARP request to PC3 since PC1 and PC3 are on the same network segment. However, PC1 and PC3 are in different VLANs, so PC3 fails to receive the ARP request from PC1.

This problem can be solved by enabling ARP proxy on the sub-interface for QinQ VLAN tag termination.

Figure 5-8 ARP proxy on the sub-interface for QinQ VLAN tag termination



5.4.8 Access of the Termination Sub-interface to L3VPN

As shown in **Figure 5-9** and **Figure 5-10**, the termination sub-interface's access to L3VPN means that the termination sub-interface is configured with L3VPN functions.

Whether the sub-interface for dot1q VLAN tag termination or the sub-interface for QinQ VLAN tag termination accesses the L3VPN depends on the packet received by the PE:

- If one tag is contained in the packet, the sub-interface for dot1q VLAN tag termination accesses the L3VPN.
- If double tags are contained in the packet, the sub-interface for QinQ VLAN tag termination accesses the L3VPN.

Access of the Sub-interface for Dot1q VLAN Tag Termination to L3VPN

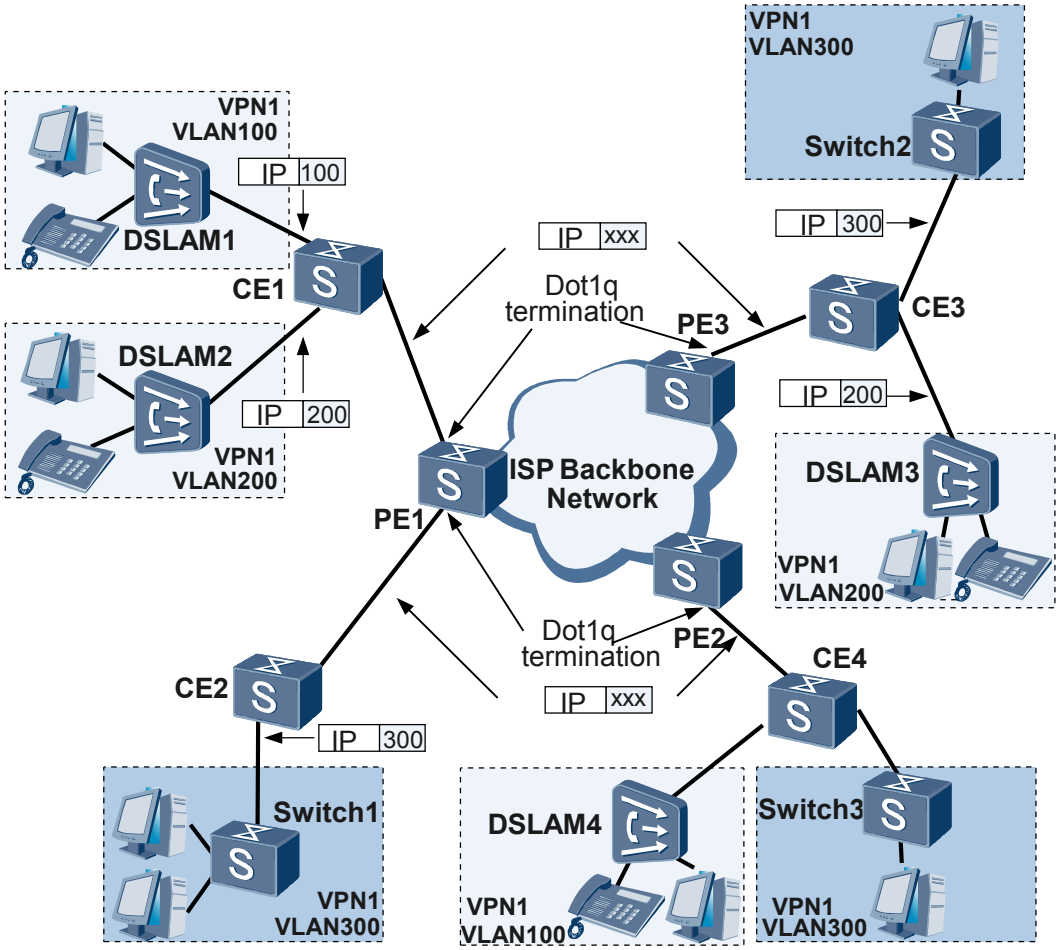
Figure 5-9 shows a typical networking for the access of the sub-interface for dot1q VLAN tag termination to L3VPN.

A user packet is attached with a customer-based VLAN tag on the Digital Subscriber Line Access Multiplexer (DSLAM) and then transmitted transparently from the CE to the PE. On the PE, the sub-interface for dot1q VLAN tag termination is configured, the outer VLAN tag is specified, and the sub-interface for dot1q VLAN tag termination is bound to a VPN instance according to the outer VLAN tag.

After receiving a user packet, the PE strips off the outer VLAN tag and then accesses the L3VPN. At the same time, the PE needs to add the correct outer VLAN tag to the packet returned to the CE.

When the PE is terminating the outer tag of a user packet, the ARP learning based on the outer VLAN tag of the user packet is required.

Figure 5-9 Networking diagram of the access of the sub-interface for dot1q VLAN tag termination to L3VPN



Access of the Sub-interface for QinQ VLAN Tag Termination to L3VPN

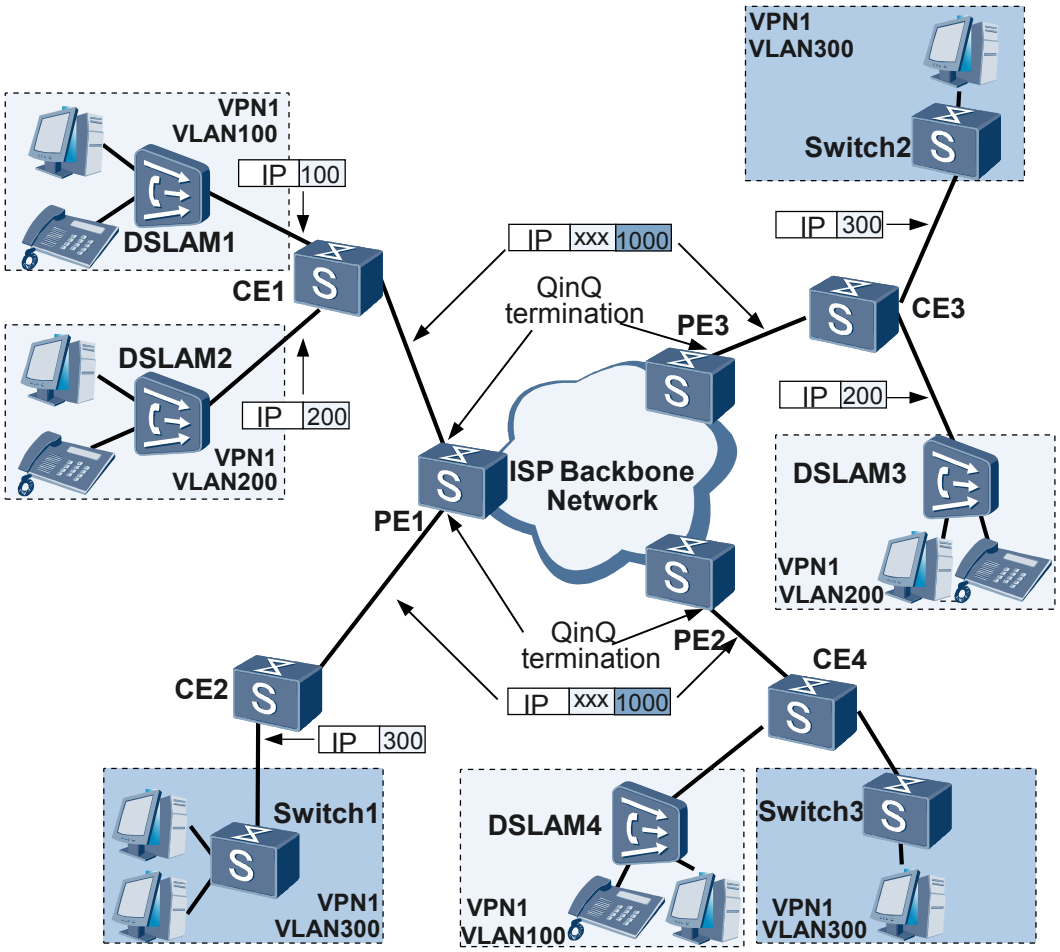
Figure 5-10 shows a typical networking for the access of the sub-interface for QinQ VLAN tag termination to L3VPN.

A user packet is attached with a customer-based VLAN tag on the DSLAM and then attached with a service-based VLAN tag on the CE. On the PE, the sub-interface for QinQ VLAN tag termination is configured, inner and outer VLAN tags are specified, and the sub-interface for QinQ VLAN tag termination is bound to a VPN instance according to double VLAN tags.

After receiving a QinQ packet from the user, the PE strips off double VLAN tags and then accesses the L3VPN. At the same time, the PE needs to attach a correct outer VLAN tag to the packet returned to the CE.

When the PE is terminating double tags of a user packet, the ARP learning based on double VLAN tags of the user packet is required.

Figure 5-10 Networking diagram of the access of the sub-interface for QinQ VLAN tag termination to L3VPN



5.4.9 Access of the Termination Sub-interface to PWE3/VLL

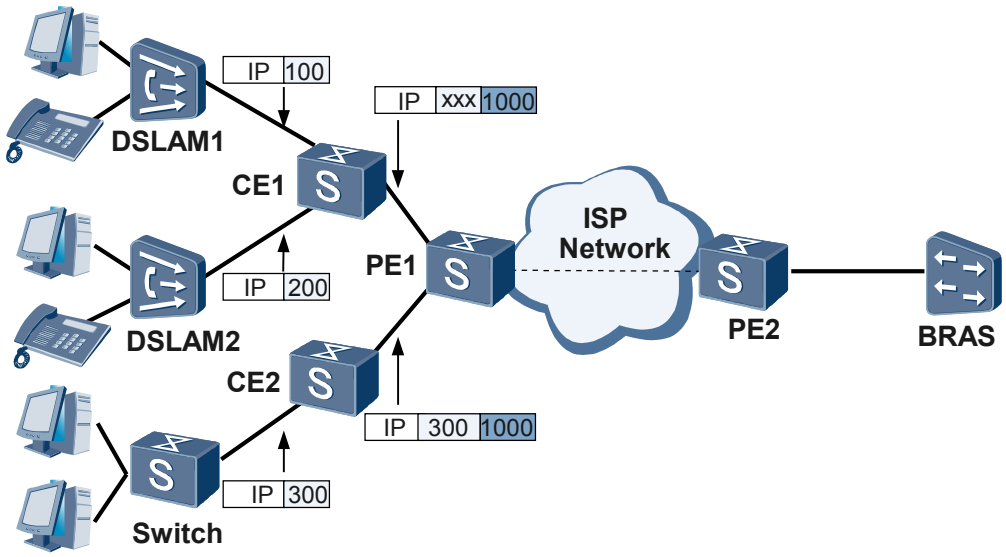
NOTE

At present, only the sub-interface for QinQ VLAN tag termination supports the access to Pseudo Wire Emulation Edge-to-Edge (PWE3)/Virtual Leased Line (VLL).

The access of the sub-interface for QinQ VLAN tag termination to PWE3/VLL means that the sub-interface for QinQ VLAN tag termination is configured with PWE3/VLL functions. By configuring the range of double VLAN tags on the sub-interface for QinQ VLAN tag termination of the PE, users within the VLAN tag range are allowed to access PWE3/VLL. Double VLAN tags of the packet, as the Layer 2 data, are transparently transmitted to the remote end. The remote end is often a Broadband Remote Access Server (BRAS). The double VLAN tags are identified on the remote BRAS and users are authenticated.

Figure 5-11 shows a typical networking for the access of the sub-interface for QinQ VLAN tag termination to PWE3/VLL.

Figure 5-11 Networking diagram of the access of the sub-interface for QinQ VLAN tag termination to PWE3/VLL



5.4.10 Access of the Termination Sub-interface to VPLS

The termination sub-interface supporting Virtual Private LAN Service (VPLS) refers to configuring VPLS on the termination sub-interface. By configuring the range of double VLAN tags on the sub-interface for QinQ VLAN tag termination of the PE, the local Virtual Switching Instance (VSI) can communicate with the remote VSI. It is often used for intercommunication of Layer 2 enterprise networks of QinQ users.

VPLS defines that one VC link connects only two VLANs that are distributed in different places. If the users need connect multiple VLANs that are distributed in remote places, multiple VCs are required.

The termination sub-interface supports a VLAN range instead of a VLAN. Therefore, only one VC is required to connect the users in the specified VLAN range by configuring the termination sub-interface's access to the VPLS. In addition, users can plan their own VLANs, disregarding ISP's VLANs.

Actually, traffic of all the VLANs in the specified range is transmitted on one VC. This greatly saves VC resources of the public network and the configuration workload.

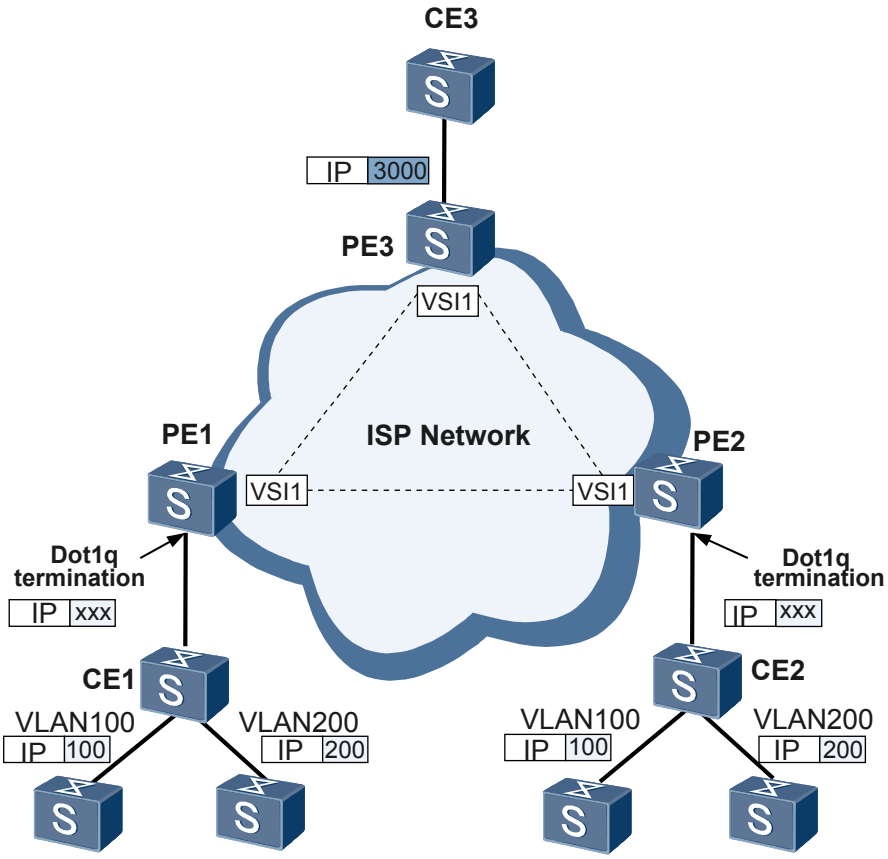
Whether the sub-interface for dot1q VLAN tag termination or the sub-interface for QinQ VLAN tag termination accesses the VPLS depends on the packet received by the PE:

- If one tag is contained in the packet, the sub-interface for dot1q VLAN tag termination accesses the VPLS.
- If double tags are contained in the packet, the sub-interface for QinQ VLAN tag termination accesses the VPLS.

Access of the Sub-interface for Dot1q VLAN Tag Termination to VPLS

Figure 5-12 shows a typical networking for the access of the sub-interface for dot1q VLAN tag termination to VPLS.

Figure 5-12 Networking diagram of the access of the sub-interface for dot1q VLAN tag termination to VPLS

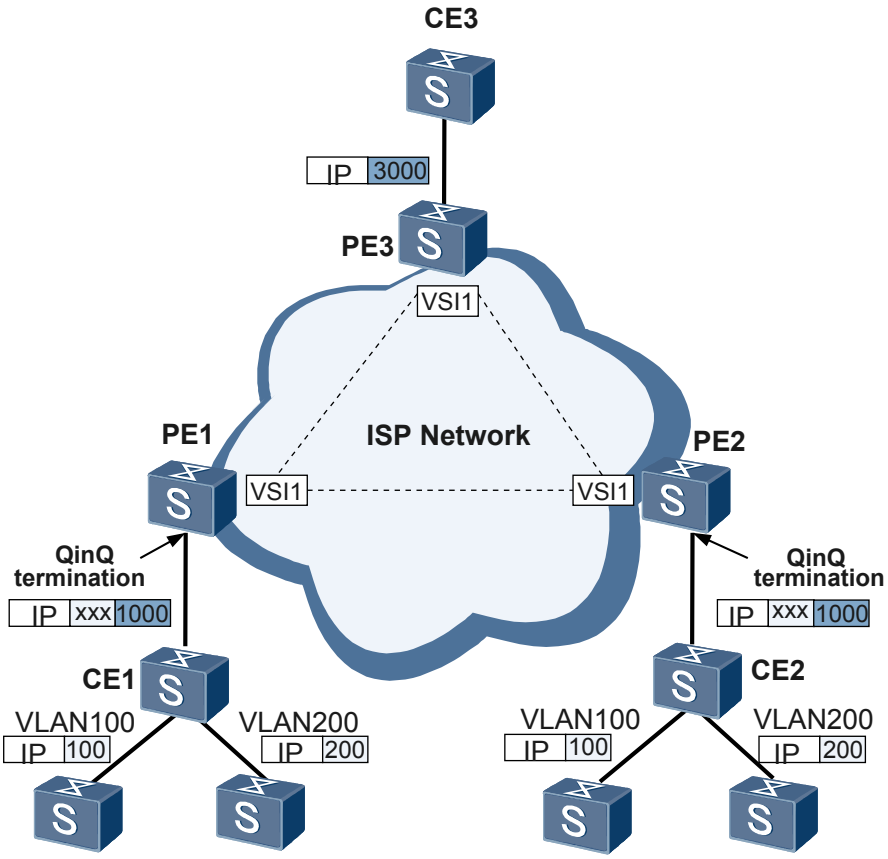


VPLS supports the Point-to-Multipoint Protocol (P2MP) and forwards data by learning MAC addresses. In this case, the access of the sub-interface for dot1q VLAN tag termination to VPLS can be achieved by MAC address learning on the basis of a single VLAN tag. Note that VLAN tags can be configured without limit for VPLS access.

Access of the Sub-interface for QinQ VLAN Tag Termination to VPLS

Figure 5-13 shows a typical networking for the access of the sub-interface for QinQ VLAN tag termination to VPLS.

Figure 5-13 Networking diagram of the access of the sub-interface for QinQ VLAN tag termination to VPLS



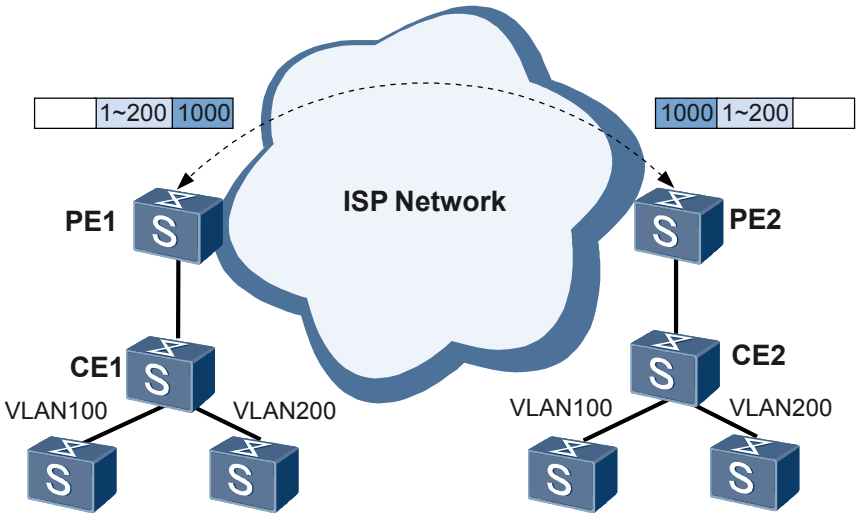
VPLS supports the P2MP and forwards data by learning MAC addresses. In this case, the access of the sub-interface for QinQ VLAN tag termination to VPLS can be achieved by MAC address learning on the basis of double VLAN tags. Note that VLAN tags can be configured without limit for VPLS access.

5.4.11 QinQ Stacking Sub-interfaces Support the Access to a PWE3 or VLL

The VLL is a point-to-point L2VPN. The VLANIF interface does not support VLL, and therefore you have to access a VPN through a main interface. Such a configuration is not flexible because multiple users cannot access the same physical interface. To ensure the access of multiple to the same physical interface, you can use the VLAN-based QinQ stacking function on different sub-interfaces. This requires that CE-VLANs on PE1 and PE2 must be the same.

Figure 5-14 shows that stacking is performed for the packets whose VLAN tags ranges from 1 to 200 on sub-interfaces. These packets are added with outer VLAN tags of the ISP network and then sent to the VLL and PWE3.

Figure 5-14 Typical networking of the access to a PWE3 or VLL supported by QinQ stacking sub-interfaces

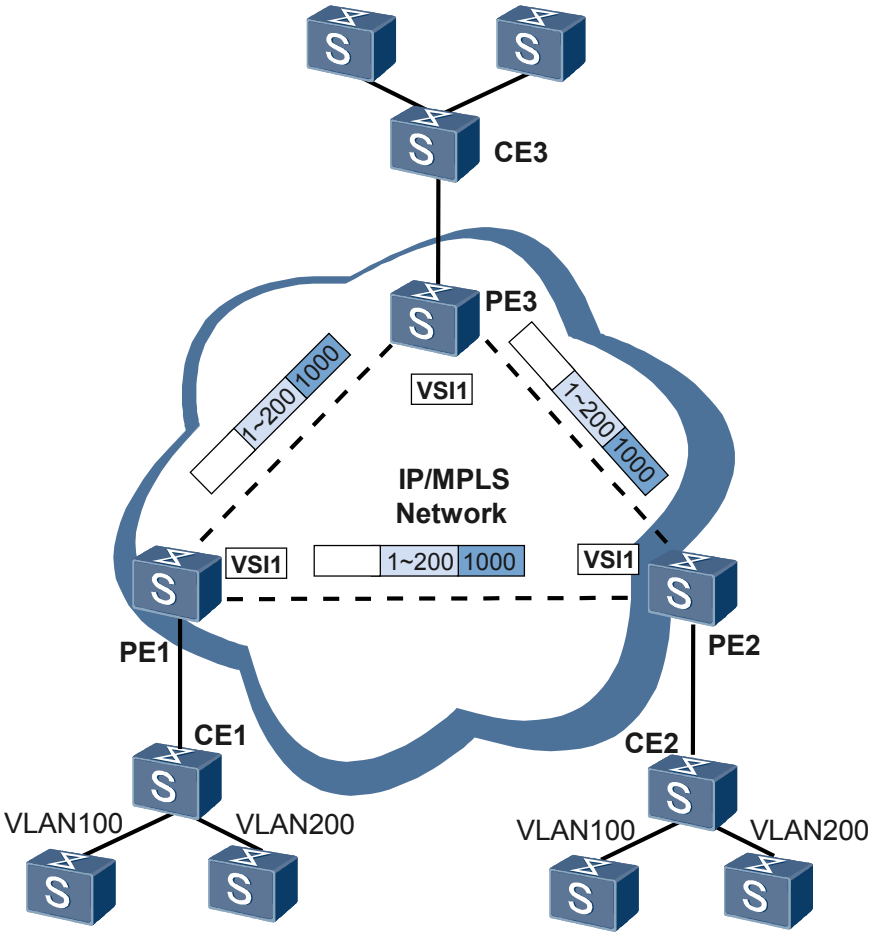


5.4.12 QinQ Stacking Sub-interfaces Support the Access to a VPLS

To access an ISP network in VPLS mode, you can enable VLAN-based VLAN stacking on the switched interface and bind VLANIF interfaces to a VSI. In this manner, user VLANs can be transparently transmitted on the ISP network.

Alternatively, you can access sub-interfaces to a VPLS through routing-based sub-interfaces. You can enable VLAN stacking on sub-interfaces. Then, as shown in **Figure 5-15**, you can implement stacking operation over packets whose VLAN tags range from 1 to 200 on sub-interfaces and add outer VLAN tags of the ISP network to these packets. In this manner, sub-interfaces are accessed to the VPLS by means of binding sub-interfaces to a VSI.

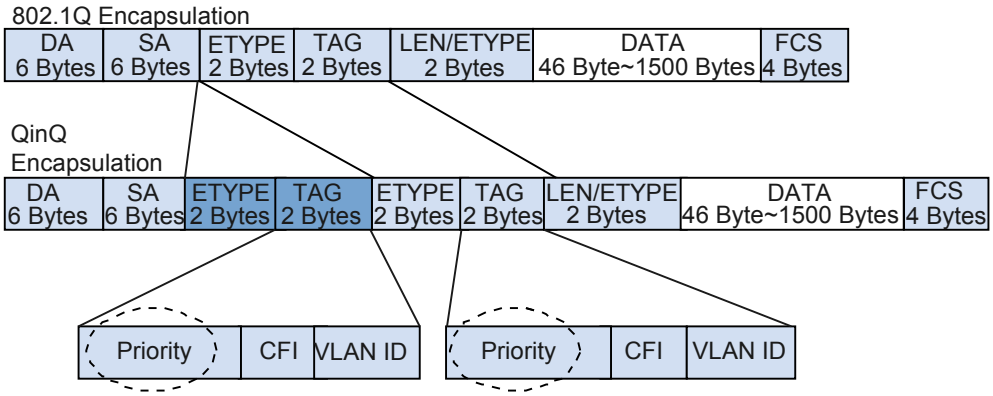
Figure 5-15 Networking diagram of the access to a VPLS supported by QinQ stacking sub-interfaces



5.4.13 QinQ Supports 802.1p Remark

After QinQ encapsulation, the 802.1p value in the inner VLAN tag is shielded and therefore not transmitted. That is, during QinQ encapsulation, the system only adds an outer VLAN tag to the packet and does not sense the 802.1p value in the inner VLAN tag. This results that after QinQ encapsulation, critical services and non-critical services are not differentiated. **Figure 5-16** shows the 802.1p remark supported by QinQ.

Figure 5-16 Typical networking of the 802.1p remark supported by QinQ



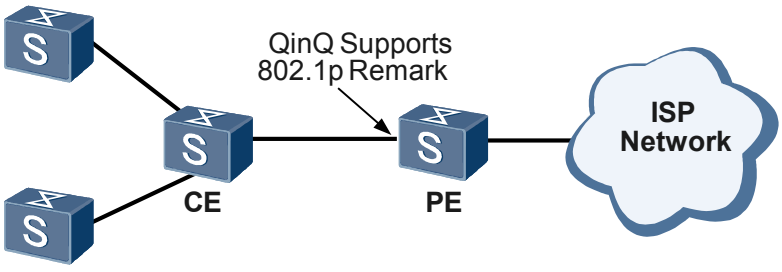
During the QinQ implementation, the 802.1p value in the inner VLAN tag needs to be sensed. You can set rules to sense the 802.1p value through commands as follows:

- Ignores the 802.1p value in the inner VLAN tag and sets a new 802.1p value for the outer VLAN tag.
- Automatically maps the 802.1p value in the inner VLAN tag as the 802.1p value in the outer VLAN tag.
- Sets the 802.1p value in the outer VLAN tag according to the 802.1p value in the inner VLAN tag.

As shown in **Figure 5-17**, QinQ supports 802.1p remark in following modes:

- Pipe mode: indicates that a 802.1p value is set.
- Uniform mode: indicates that the 802.1p value in the inner VLAN tag is adopted.
- Maps the 802.1p value in the inner VLAN tag to the 802.1p value in the outer VLAN tag. Multiple 802.1p value in the inner VLAN tag can be mapped to the 802.1p value in the outer VLAN tag; one 802.1p value in the inner VLAN tags cannot be mapped to the multiple 802.1p value in the outer VLAN tag.

Figure 5-17 Networking diagram of 802.1p remark supported by QinQ

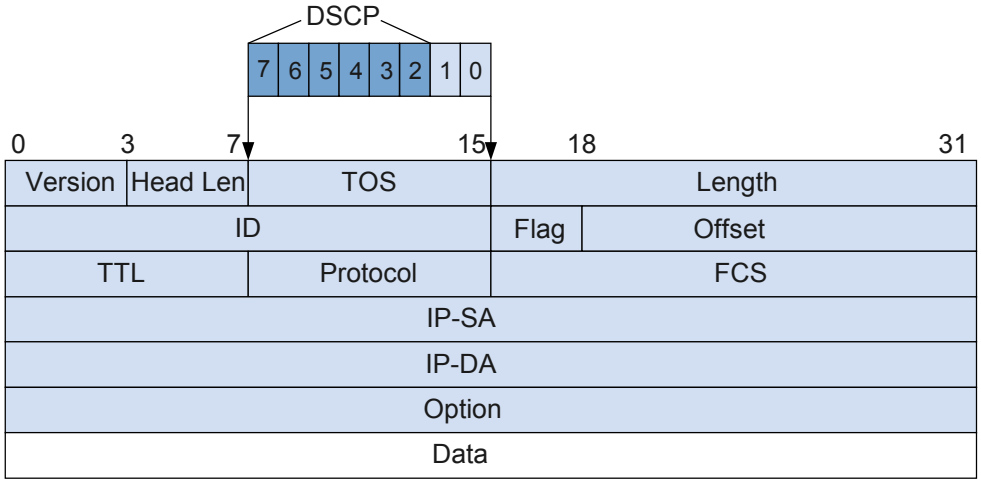


5.4.14 QinQ Termination Supports the 802.1p Remark and DSCP Remark

As shown in **Figure 5-18**, according to RFC 2724, six bits of the Type of Service (ToS) field in an IPv4 packet header serve as the DiffServ Code Point (DSCP), which provides reference for

differentiated services (DiffServ) and is used to ensure the Quality of Service (QoS) on the IP network. The operation of the traffic controller on the gateway depends on the DSCP field.

Figure 5-18 Structure of the DSCP signaling

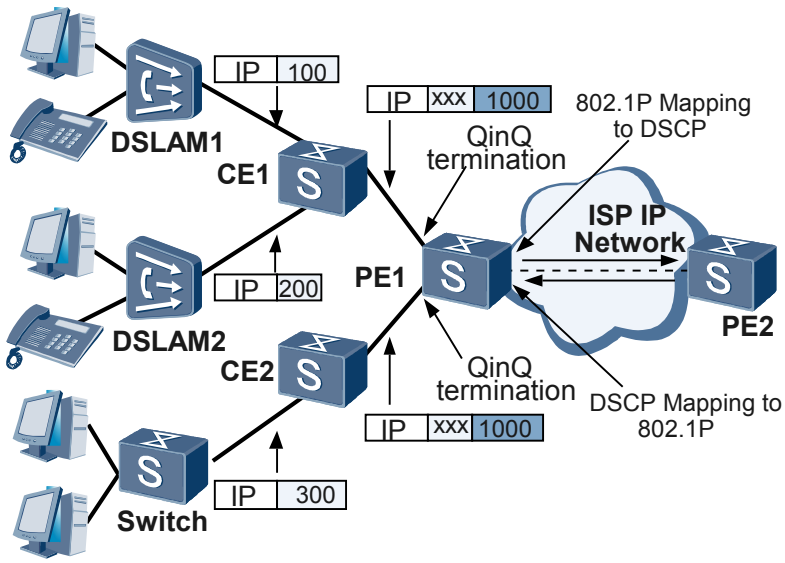


As shown in **Figure 5-19**, after being terminated on the PE, the packet is sent to the ISP network. To ensure the completeness of the QoS information in the packet, the mapping relationship between the 802.1p values in outer and inner tags and the DSCP field needs to be configured.

The following lists the mapping modes:

- The 802.1p value in the inner VLAN tag is mapped to the DSCP field.
- The 802.1p value in the outer VLAN tag is mapped to the DSCP field.
- A vale ranging from 0 to 7 is selected and mapped to the DSCP field.

Figure 5-19 Typical networking of the 802.1p remark and DSCP remark supported by QinQ termination



5.4.15 QinQ Termination Supports the 802.1p Remark and EXP (MPLS) Remark

As shown in **Figure 5-20**, the EXP field in an MPLS packet is used for Class of Service (CoS). The operation of the traffic controller on the gateway depends on the field.

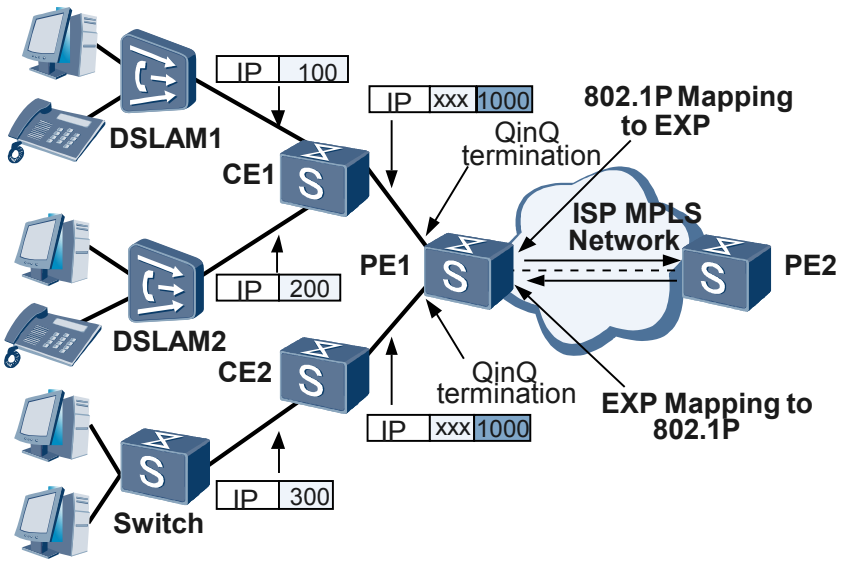
Figure 5-20 Structure of an MPLS packet

As shown in **Figure 5-21**, after a user packet is terminated, it is sent to the ISP MPLS network. To ensure the completeness of the QoS information in the packet, the mapping relationship between the 802.1p values in outer and inner tags and the EXP field needs to be configured.

The following lists the mapping modes:

- The 802.1p value in the inner VLAN tag is mapped to the EXP field.
- The 802.1p value in the outer VLAN tag is mapped to the EXP field.
- A vale ranging from 0 to 7 is selected and mapped to the EXP field.

Figure 5-21 Typical networking of the 802.1p remark and EXP (MPLS) remark supported by QinQ termination



5.4.16 Summary of QinQ

The development of the QinQ technology is as follows:

QinQ Layer 2 tunnel (port-based QinQ)-> Flexible QinQ (selective QinQ/VLAN stacking)-> Enhanced QinQ (QinQ termination&Dot1q termination&QinQ stacking)-> Dynamic QinQ applied based on specified application scenarios

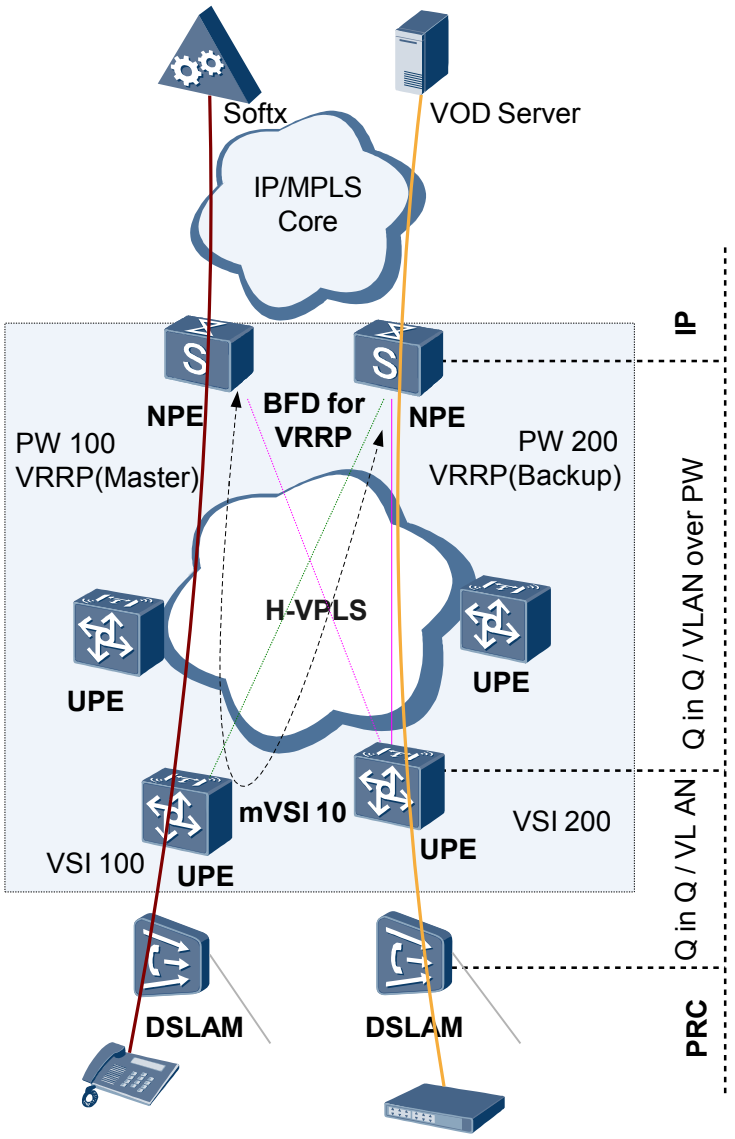
QinQ is used to expand the number of VLANs and implement traffic distribution over user services. The outer tag represents DSLAM and service types of individual services, different VPN sites of enterprise services, or different ISPs of batch services. As required by the planning of network services, QinQ can be deployed at different network layers, such as the access layer, convergence layer, bearer layer, and core layer.

QinQ can be used with other technologies to help operators implement refined management over individual services, enterprise services, and batch services.

- Individual services include unicast services such as HSI services, VoIP/(Video On Demand) VOD unicast services, and BTV services.
- Enterprise services include services of accessing a public network, L3VPN services, and L2VPN services.
- Batch services include line-based batch services.

Figure 5-22 shows the hierarchical deployment of QinQ.

Figure 5-22 Individual Services - VoIP/VOD Unicast Services



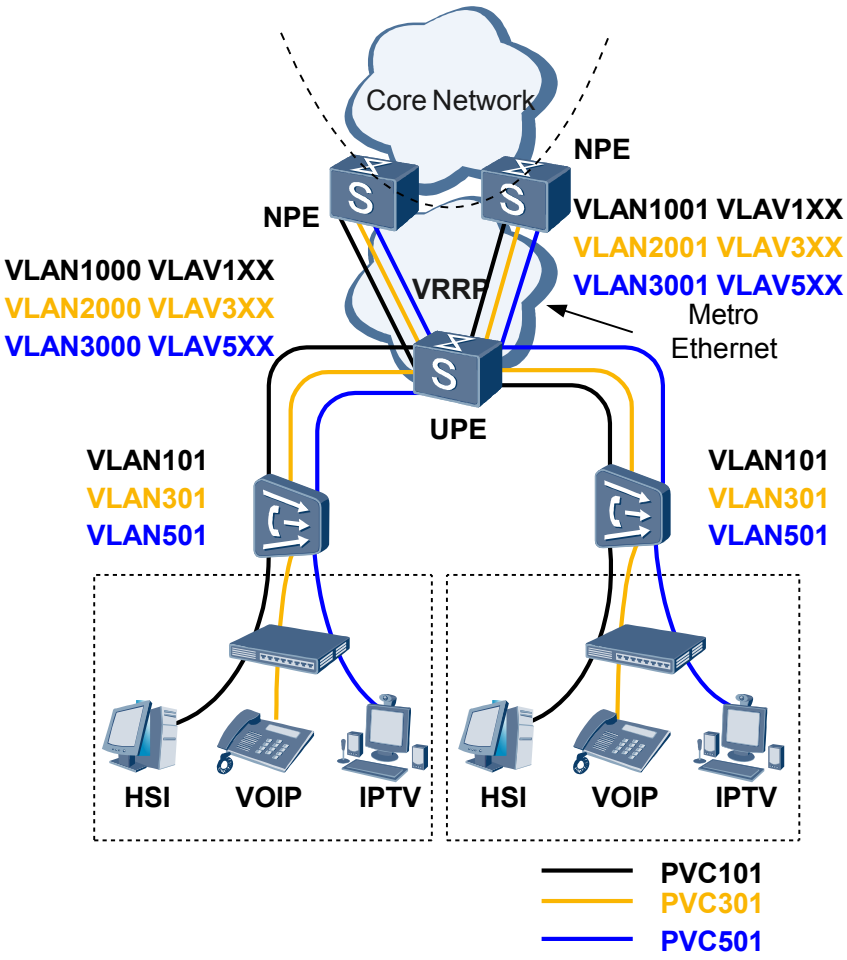
- The DSLAMs perform QinQ encapsulation over VoIP and VOD services of users. The outer tag represents the DSLAM number and the service type (VoIP or VOD). The QinQ encapsulation can be performed on a UPE through selective QinQ.
- Services of the same type enter the same VSI on a UPE according to their outer tags. The UPE transparently transmits QinQ packets to the NPE through a PW. On the NPE, the PW is terminated and services are sent to L3 for process according to VLAN/QinQ information.
- The NPE works in load balance and active/standby modes. It carries out load balance according to the outer tag and determines the active or standby status through BFD for VRRP.
- DHCP authentication packets are broadcast to two NPEs. The active NPE processes the packets through DHCP Relay and checks the binding; The DSLAM and UPE enable DHCP Snooping (insert option 82 field) and check the binding.

Differences between QinQ layer 2 tunnel, flexible QinQ, enhanced QinQ (QinQ termination&Dot1q termination&QinQ stacking), and dynamic QinQ are not mentioned here.

5.5 Application

5.5.1 Public User Services On the ME Network

Figure 5-23 Typical networking of applying QinQ on the ME network



As shown in **Figure 5-23**, the DSLAM supports multiple PVC access. In this case, the same user can use multiple services, such as HSI, IPTV and VoIP.

As shown in **Figure 5-23**, the operator defines the mapping relationship between PVCs and services and between services and range of VLAN IDs.

Table 5-4 Mapping relationship between services and VLAN IDs

Service Name	Full Spelling	Range of VLAN IDs
HSI	High Speed Internet	101 to 300
VoIP	Voice Over IP	301 to 500
IPTV	Internet Protocol Television	501 to 700

Presume that a user adopts the VoIP service. Data reaches the DSLAM through a specified PVC and is marked with a tag indicating the range of VoIP VLAN ID, such as 301, according to the

mapping relationship between the PVC and the VLAN. When reaching the UPE, the VoIP packet is marked with difference outer VLAN IDs according to different ranges of VLAN IDs, such as 2000. The inner VLAN ID represents user information and the outer VLAN ID represents service information and the location of the DSLAM (Packets from different DSLAMs are marked with different outer VLAN tags.) When the packet reaches the NPE according to the outer VLAN tag, the VLAN tag is terminated on the QinQ termination sub-interface. According to the configuration of the core network, the packet is forwarded on the IP network or enters the relevant VPN.

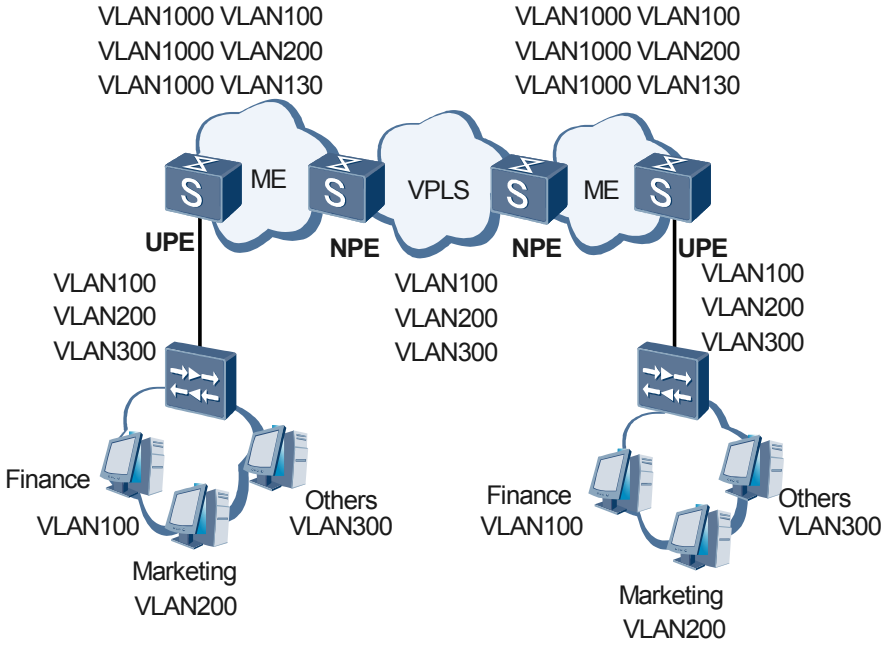
HSI and IPTV services are processed in the same way, and the difference is that QinQ termination of HSI services is implemented on the BRAS.

As required, the NPE can perform HQoS scheduling based on the two-layer tags and generate a DHCP binding table to avoid network attacks. In addition, the NPE can implement DHCP authentication based on the two-layer tags or other information and enable QinQ VRRP to ensure the reliable access of services.

5.5.2 Enterprise Users Are Connected through Private Line

As shown in **Figure 5-24**, an enterprise has two sites in different places. Each site has three networks of finance, sales and others. To ensure network security, it is required that users of different networks cannot communicate with each other.

Figure 5-24 Typical networking of the private line for the communication between enterprise users



The operator adopts VPLS on the MPLS/IP core network and QinQ on the ME network. Each site is configured with three VLANs that separately represent finance, sales and other departments, and their VLAN IDs are 100, 200 and 300. An outer VLAN 1000 is encapsulated on the UPE (Packets can be added with different VLAN tags on different UPEs.) The VSI on the NPE is in symmetry mode, thus only users of the same VLAN in different sites can communicate with each other.

5.6 Terms and Abbreviations

Terms

Terms	Description
QinQ port	The 802.1Q-in-802.1Q (QinQ) interface refers to an interface that can process VLAN frames with a single tag (dot1q termination or VLAN mapping) or VLAN frames with double tags (QinQ termination, or VLAN stacking).
QinQ termination sub-interface	A QinQ termination sub-interface can identify the one-layer or two-layer tags of QinQ packets and then strips the tags or sends the packets according to the subsequent forwarding.

Abbreviation

Abbreviation	Full Spelling
ARP	Address Resolution Protocol
DHCP	Dynamic Host Configuration Protocol
HSI	High Speed Internet
IPTV	Internet Protocol Television
PVC	Permanent Virtual Connection
PWE3	Pseudo Wire Emulation Edge-to-edge
QinQ	802.1Q in 802.1Q
QinQ Termination	QinQ Termination
Selective QinQ	Selective QinQ
VLAN	Virtual Local Area Netw
VLAN Stacking	VLAN Stacking
VLL	Virtual Leased Line
VOIP	Voice over IP
VPLS	Virture Private LAN Service
VRRP	Virtual Router Redundancy Protocol
VSI	Virtual Switch Instance

6 GVRP

About This Chapter

6.1 Introduction to GVRP

Definition

The Generic Attribute Registration Protocol (GARP) provides a mechanism to propagate attributes so that a protocol entity can register and deregister attributes. By filling different attributes into GARP packets, GARP supports different upper-layer applications.

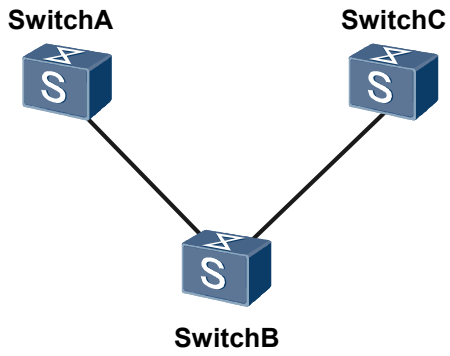
The GARP VLAN Registration Protocol (GVRP) is used to register and deregister VLAN attributes.

GARP identifies applications through destination MAC addresses. IEEE Std 802.1Q assigns 01-80-C2-00-00-21 to the VLAN application (GVRP).

Purpose

To deploy certain VLANs on all devices on a network, the network administrator needs to manually create these VLANs on each device. As shown in **Figure 6-1**, three switches are connected through trunk links. VLAN 2 is configured on SwitchA, and VLAN 1 is configured on SwitchB and SwitchC. To forward packets of VLAN 2 from SwitchA to SwitchC, the network administrator must manually create VLAN 2 on SwitchB and SwitchC.

Figure 6-1 Networking of GVRP application



When a network is complicated and the network administrator is unfamiliar with the network topology or when many VLANs are configured on the network, huge workload is required for manual configuration. In addition, configuration errors may occur. In this case, you can configure GVRP on the network to implement automatic registration of VLANs.

6.2 References

The following table lists the references of this document.

Document	Description	Remarks
IEEE Std 802.1D	Information technology-Telecommunications and information exchange between systems-Local and metropolitan area networks-Common specifications-Media Access Control (MAC) Bridges	-
IEEE Std 802.1Q	IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks	-

6.3 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

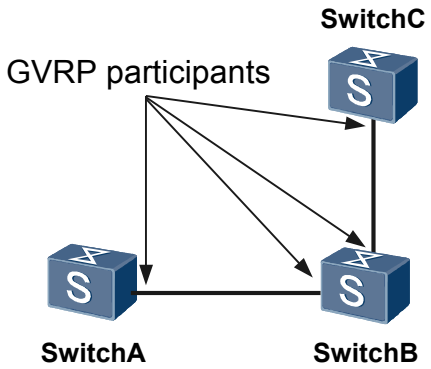
6.4 Principles

6.4.1 Basic Concepts

Participant

On a device, each port running a protocol is considered as a participant. On a device running GVRP, each GVRP-enabled port is considered as a GVRP participant, as shown in **Figure 6-2**.

Figure 6-2 GVRP participant



VLAN Registration and Deregistration

GVRP implements automatic registration and deregistration of VLAN attributes. The functions of VLAN registration and deregistration are:

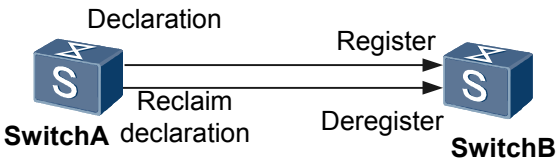
- VLAN registration: adds a port to a VLAN.
- VLAN deregistration: removes a port from a VLAN.

GVRP registers and deregisters VLAN attributes through attribute declarations and reclaim declarations as follows:

- When a port receives a VLAN attribute declaration, it registers the VLAN specified in the declaration. That is, the port is added to the VLAN.
- When a port receives a VLAN attribute reclaim declaration, it deregisters the VLAN specified in the declaration. That is, the port is removed from the VLAN.

A port registers or deregisters VLANs only when it receives GVRP messages.

Figure 6-3 VLAN registration and deregistration



GARP Messages

GARP participants exchange VLAN information through GARP messages. Major GARP messages are Join messages, Leave messages, and LeaveAll messages.

- Join message

When a GARP participant expects other devices to register its attributes, it sends Join messages to other devices. When the GARP participant receives a Join message from another participant or is configured with attributes statically, it also sends Join messages to other devices for the devices to register the new attributes.

Join messages are classified into JoinEmpty messages and JoinIn messages. The difference between the two types of messages is:

- JoinEmpty: declares an unregistered attribute.
- JoinIn: declares a registered attribute.

- Leave message

  When a GARP participant expects other devices to deregister its attributes, it sends Leave messages to other devices. When the GARP participant receives a Leave message from another participant or some of its attributes are deregistered statically, it also sends Leave messages to other devices.

  Leave messages are classified into LeaveEmpty messages and LeaveIn messages. The difference between the two types of messages is:

  - LeaveEmpty: deregisters an unregistered attribute.
  - LeaveIn: deregisters a registered attribute.

- LeaveAll message

  When a participant starts, it starts the LeaveAll timer. When the LeaveAll timer expires, the participant sends LeaveAll messages to other devices.

  A participant sends LeaveAll messages to deregister all attributes so that other participants can re-register attributes of the local participant. LeaveAll messages are used to periodically delete useless attributes on the network. For example, an attribute of a participant is deleted but the participant does not send Leave messages to request other participants to deregister the attribute because of a sudden power failure. Then this attribute becomes useless.

GARP Timers

The GARP protocol defines four timers, which are described as follows:

- Join timer

  The Join timer controls sending of Join messages including JoinIn messages and JoinEmpty messages.

  After sending the first Join message, a participant starts the Join timer. If the participant receives a JoinIn message before the Join timer expires, it does not send the second Join message. If the participant does not receive any JoinIn message, it sends the second Join message when the Join timer expires. This ensures that the Join message can be sent to other participants. Each port maintains an independent Join timer.

- Hold timer

  The Hold timer controls sending of Join messages (JoinIn messages and JoinEmpty messages) and Leave messages (LeaveIn messages and LeaveEmpty messages).

  After a participant is configured with an attribute or receives a message, it does not send the message to other participants before the Hold timer expires. The participant encapsulates messages received within the hold time into a minimum number of packets, reducing the packets sent to other participants. If the participant does not use the Hold timer but forwards a message immediately after receiving one, a large number of packets are transmitted on the network. This makes the network unstable and wastes data fields of packets.

Each port maintains an independent Hold timer. The Hold timer value must be equal to or smaller than half of the Join timer value.

- Leave timer

  The Leave timer controls attribute deregistration.

  A participant starts the Leave timer after receiving a Leave or LeaveAll message. If the participant does not receive any Join message of the corresponding attribute before the Leave timer expires, the participant deregisters the attribute.

  A participant sends a Leave message if one of its attributes is deleted, but this attribute may still exist on other participants. Therefore, the participant receiving the Leave message cannot deregister the attribute immediately and needs to wait for messages from other participants.

  For example, an attribute has two sources on the network: participant A and participant B. Other participants register the attribute through GARP. If the attribute is deleted from participant A, participant A sends a Leave message to other participants. After receiving the Leave message, participant B sends a Join message to other participants because the attribute still exists on participant B. After receiving the Join message from participant B, other participants retain the attribute. Other participants deregister the attribute only if they do not receive any Join message of the attribute within a period longer than two times the Join timer value. Therefore, the Leave timer value must be greater than two times the Join timer value.

  Each port maintains an independent Leave timer.

- LeaveAll timer

  When a GARP participant starts, it starts the LeaveAll timer. When the LeaveAll timer expires, the participant sends a LeaveAll message and restarts the LeaveAll timer.

  After receiving a LeaveAll message, a participant restarts all GARP timers. The participant sends another LeaveAll message when its LeaveAll timer expires. This reduces LeaveAll messages sent in a period of time.

  If LeaveAll timers of multiple devices expire at the same time, they send LeaveAll messages at the same time, which causes unnecessary LeaveAll messages. To solve this problem, each device uses a random value between the LeaveAll timer value and 1.5 times the LeaveAll timer value as its LeaveAll timer value. When a LeaveAll event occurs, all attributes on the entire network are deregistered. The LeaveAll event affects the entire network; therefore, you need to set the LeaveAll timer to a proper value, at least greater than the Leave timer value.

  Each device maintains a global LeaveAll timer.

Registration Modes

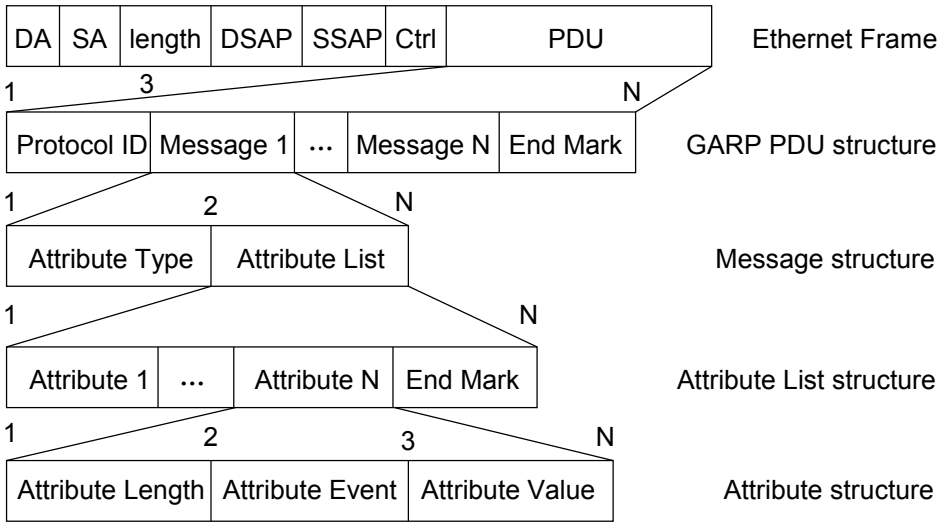
A manually configured VLAN is a static VLAN, and a VLAN created through GVRP is a dynamic VLAN. GVRP provides three registration modes. Static VLANs and dynamic VLANs are processed differently in each registration mode as follows:

- Normal mode: Dynamic VLANs can be registered on a port, and the port can send declarations of static VLANs and dynamic VLANs.
- Fixed mode: Dynamic VLANs cannot be registered on a port, and the port can send only declarations of static VLANs.
- Forbidden mode: Dynamic VLANs cannot be registered on a port. All VLANs except VLAN 1 are deleted from the port, and the port can send only the declaration of VLAN 1.

6.4.2 Packet Structure

GARP packets are encapsulated in the IEEE 802.3 Ethernet format, as shown in **Figure 6-4**.

Figure 6-4 GARP packet structure



The following table describes the fields in a GARP packet.

Field	Description	Value
Protocol ID	Indicates the protocol ID.	The value is 1.
Message	Indicates the messages in the packet. Each message consists of the Attribute Type and Attribute list fields.	-
Attribute Type	Indicates the type of an attribute, which is defined by the GARP application.	The value is 0x01 for GVRP, indicating that the attribute value is a VLAN ID
Attribute List	Indicates the attribute list of a message, which consists of multiple attributes.	-
Attribute	Indicates an attribute, which consists of the Attribute Length, Attribute Event, and Attribute Value fields.	-
Attribute Length	Indicates the length of an attribute.	The value ranges from 2 to 255, in bytes.

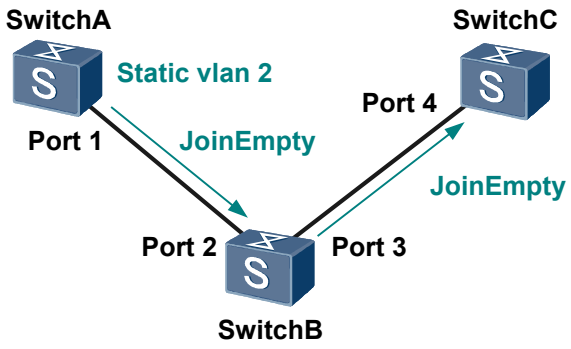
Field	Description	Value
Attribute Event	Indicates the event that an attribute describes.	The value can be: ● 0: LeaveAll Event ● 1: JoinEmpty Event ● 2: JoinIn Event ● 3: LeaveEmpty Event ● 4: LeaveIn Event ● 5: Empty Event
Attribute Value	Indicates the value of an attribute.	The value is a VLAN ID for GVRP. This field is invalid in a LeaveAll attribute.
End Mark	Indicates the end of a GARP PDU.	The value is 0x00.

6.4.3 Working Procedure

This section describes the working procedure of GVRP by using an example. This example illustrates how a VLAN attribute is registered and deregistered on a network in four phases.

One-Way Registration

Figure 6-5 One-Way registration of VLAN an attribute



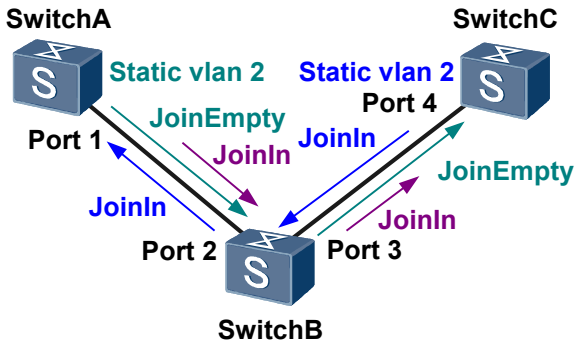
Static VLAN 2 is created on SwitchA. Ports on SwitchB and SwitchC can join VLAN 2 automatically through one-way registration. The process is as follows:

1. After VLAN 2 is created on SwitchA, Port 1 of SwitchA starts the Join timer and Hold timer. When the Hold timer expires, Port 1 sends the first JoinEmpty message to SwitchB. When the Join timer expires, Port 1 restarts the Hold timer. When the Hold timer expires again, Port 1 sends the second JoinEmpty message.

2. After Port 2 of SwitchB receives the first JoinEmpty message, SwitchB creates dynamic VLAN 2 and adds Port 2 to VLAN 2. In addition, SwitchB requests Port 3 to start the Join timer and Hold timer. When the Hold timer expires, Port 3 sends the first JoinEmpty message to SwitchC. When the Join timer expires, Port 3 restarts the Hold timer. When the Hold timer expires again, Port 3 sends the second JoinEmpty message. After Port 2 receives the second JoinEmpty message, SwitchB does not take any action because Port 2 has been added to VLAN 2.
3. When Port 4 of SwitchC receives the first JoinEmpty message, SwitchC creates dynamic VLAN 2 and adds Port 4 to VLAN 2. After Port 4 receives the second JoinEmpty message, SwitchC does not take any action because Port 4 has been added to VLAN 2.
4. Every time the LeaveAll timer expires or a LeaveAll message is received, each switch restarts the LeaveAll timer, Join timer, Hold timer, and Leave timer. Then Port 1 repeats step 1 to send JoinEmpty messages. Port 3 of SwitchB sends JoinEmpty messages to SwitchC in the same way.

Two-Way Registration

Figure 6-6 Two-way registration of VLAN an attribute



After one-way registration is complete, Port 1, Port 2, and Port 4 are added to VLAN 2 but Port 3 is not added to VLAN 2 because only ports receiving a JoinEmpty or JoinIn message can be added to dynamic VLANs. To transmit traffic of VLAN 2 in both directions, VLAN registration from SwitchC to SwitchA is required. The process is as follows:

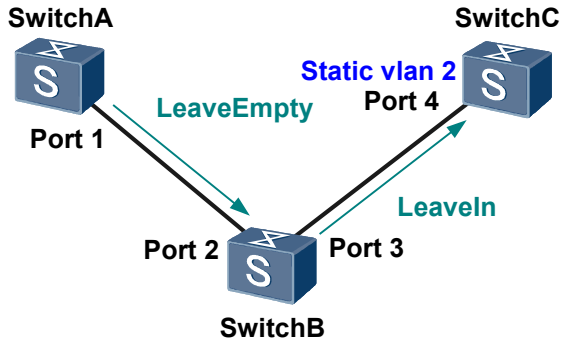
1. After one-way registration is complete, static VLAN 2 is created on SwitchC (the dynamic VLAN is replaced by the static VLAN). Port 4 of SwitchC starts the Join timer and Hold timer. When the Hold timer expires, Port 4 sends the first JoinIn message (because it has registered VLAN 2) to SwitchB. When the Join timer expires, Port 4 restarts the Hold timer. When the Hold timer expires, Port 4 sends the second JoinIn message.
2. After Port 3 of SwitchB receives the first JoinIn message, SwitchB adds Port 3 to VLAN 2 and requests Port 2 to start the Join timer and Hold timer. When the Hold timer expires, Port 2 sends the first JoinIn message to SwitchA. When the Join timer expires, Port 2 restarts the Hold timer. When the Hold timer expires again, Port 2 sends the second JoinIn message. After Port 3 receives the second JoinIn message, SwitchB does not take any action because Port 3 has been added to VLAN 2.
3. When SwitchA receives the JoinIn message, it stops sending JoinEmpty messages to SwitchB. Every time the LeaveAll timer expires or a LeaveAll message is received, each

switch restarts the LeaveAll timer, Join timer, Hold timer, and Leave timer. Port 1 of SwitchA sends a JoinIn message to SwitchB when the Hold timer expires.

4. SwitchB sends a JoinIn message to SwitchC.
5. After receiving the JoinIn message, SwitchC does not create dynamic VLAN 2 because static VLAN 2 has been created.

One-Way Deregistration

Figure 6-7 One-way deregistration of VLAN an attribute

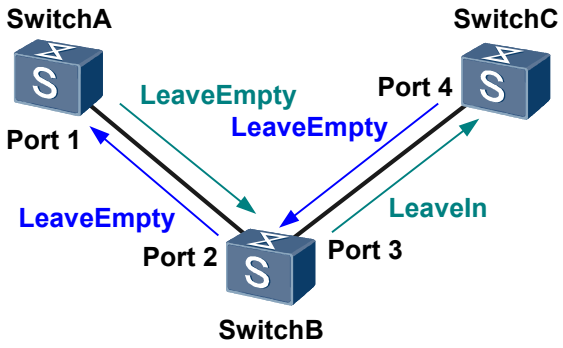


When VLAN 2 is not required on the switches, the switches can deregister VLAN 2. The process is as follows:

1. After static VLAN 2 is manually deleted from SwitchA, Port 1 of SwitchA starts the Hold timer. When the Hold timer expires, Port 1 sends a LeaveEmpty message to SwitchB. Port 1 needs to send only one LeaveEmpty message.
2. After Port 2 of SwitchB receives the LeaveEmpty message, it starts the Leave timer. When the Leave timer expires, Port 2 deregisters VLAN 2. Then Port 2 is deleted from VLAN 2, but VLAN 2 is not deleted from SwitchB because Port 3 is still in VLAN 2. At this time, SwitchB requests Port 3 to start the Hold timer and Leave timer. When the Hold timer expires, Port 3 sends a LeaveIn message to SwitchC. Static VLAN 2 is not deleted from SwitchC; therefore, Port 3 can receive the JoinIn message sent from Port 4 after the Leave timer expires. In this case, SwitchA and SwitchB can still learn dynamic VLAN 2.
3. After SwitchC receives the LeaveIn message, Port 4 is not deleted from VLAN 2 because VLAN 2 is a static VLAN on SwitchC.

Two-Way Deregistration

Figure 6-8 Two-way deregistration of VLAN an attribute



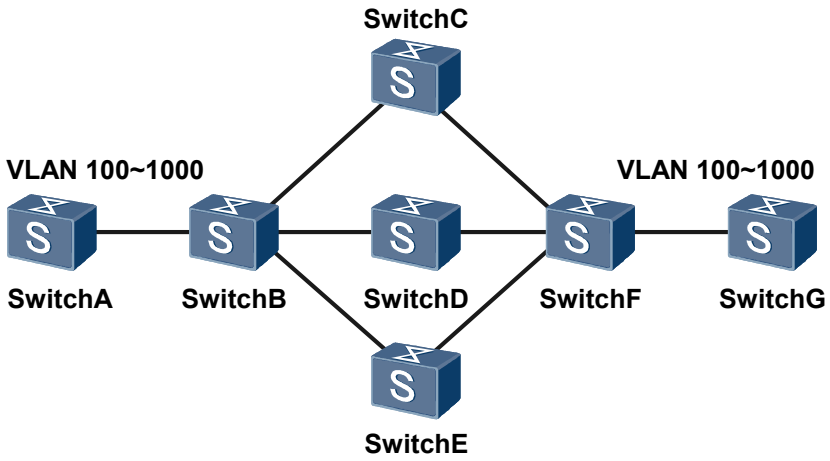
To delete VLAN 2 from all the switches, two-way deregistration is required. The process is as follows:

1. After static VLAN 2 is manually deleted from SwitchC, Port 4 of SwitchC starts the Hold timer. When the Hold timer expires, Port 4 sends a LeaveEmpty message to SwitchB.
2. After Port 3 of SwitchB receives the LeaveEmpty message, it starts the Leave timer. When the Leave timer expires, Port 3 deregisters VLAN 2. Then Port 3 is deleted from VLAN 2, and VLAN 2 is deleted from SwitchB. At this time, SwitchB requests Port 2 to start the Hold timer. When the Hold timer expires, Port 2 sends a LeaveEmpty message to SwitchA.
3. After Port 1 of SwitchA receives the LeaveEmpty message, it starts the Leave timer. When the Leave timer expires, Port 1 deregisters VLAN 2. Then Port 1 is deleted from VLAN 2, and VLAN 2 is deleted from SwitchA.

6.5 Applications

GVRP enables switches on a network to dynamically maintain and update VLAN information. With GVRP, you can adjust the VLAN deployment on the entire network by configuring only a few devices. You do not need to analyze the topology and manage configurations. As shown in **Figure 6-9**, GVRP is enabled on all switches. Switches are interconnected through trunk ports and each trunk port allows packets of all VLANs to pass. You simply need to configure static VLANs 100 to 1000 on SwitchA and SwitchG. Then the other switches can learn VLANs 100 to 1000 through GVRP.

Figure 6-9 Typical application of GVRP



6.6 Terms and Abbreviations

Abbreviations

Abbreviation	Full Spelling
GARP	Generic Attribute Registration Protocol
GVRP	GARP VLAN Registration Protocol

7 MAC

About This Chapter

7.1 Introduction to MAC

Definition

A MAC address consists of 48 bits and is usually displayed as a 12-digit hexadecimal number in dotted notation. The MAC addresses are unique, managed and distributed by IEEE. Each MAC address consists of a vendor code and a sequence number. The first 24 bits indicate the vendor code, and the last 24 bits are defined by the manufacturer.

The Ethernet identifies a network element by MAC address. On a LAN, data is encapsulated in Ethernet frames. After receiving an Ethernet frame, a device checks whether the MAC address of the frame is its own MAC address. If so, the device sends the frame to the upper layer software. This process does not apply to broadcast and multicast frames.

MAC addresses are classified into the following types:

- Physical MAC address: identifies a device on a LAN. Each physical MAC address is globally unique.
- Broadcast MAC address: indicates all devices on a LAN. The broadcast address is all 1s (FF-FF-FF-FF-FF-FF).
- Multicast MAC address: indicates a group of stations on a LAN. All the MAC addresses with the eighth bit as 1 are the multicast MAC address (xxxxxxx1-xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx-xxxxxxxx), excluding the broadcast MAC address.

The S7700 maintains a MAC address table. The MAC address table records MAC addresses of all the devices connected to ports of the S7700. When forwarding a data frame, the S7700 searches the MAC address table for the outbound interface according to the destination MAC address of the frame. This helps to decrease frame broadcasting.

Purpose

- Manually creating or deleting a MAC address entry

  By manually adding a MAC address entry and configuring a static entry, you can bind an authorized user to a specified interface. The static entry does not age and ensures that traffic is sent to the user correctly.

  By configuring blackhole MAC address entries, you can enable a switch to drop packets with a specified source MAC address or destination MAC address.

  You can delete a static or blackhole MAC entry or delete dynamic MAC addresses before the aging time expires. By properly configuring a MAC address table on a device, you can improve the forwarding efficiency of the device.

- Setting the aging time of MAC addresses

  An interface of a switch learns a dynamic MAC address entry based on the source MAC address of packets. If no packet with this MAC address as the destination or source MAC address reaches the S7700 within the specified time, the switch deletes the dynamic MAC address entry. The total number of MAC address entries in the MAC address table is limited. Therefore, periodically deleting inactive MAC address entries ensures that the MAC addresses of other online users can be learned.

  By configuring the aging time of MAC addresses on a switch, you can change the minimum holdtime of dynamic MAC address entries in the MAC address table. If the network topology changes frequently, you can set shorter aging time of MAC addresses to improve

the frequency of updating MAC address entries. If the network is stable, you can set longer aging time of MAC addresses to reduce unknown unicast packets to a certain extent.

- Disabling MAC address learning on an interface or a VLAN

  After MAC address learning is disabled on an interface or a VLAN, no MAC address entry can be learned on the interface or VLAN. This prevents a switch from learning invalid MAC addresses when being attacked.

- Limiting the number of MAC addresses on an interface or a VLAN

  You can limit the maximum number of dynamic MAC addresses that can be learned on an interface or a VLAN. This prevents a switch from learning invalid MAC addresses when being attacked.

- Port Security

  The port security function changes MAC addresses learned by an interface to secure dynamic MAC addresses or sticky MAC addresses. It prevents devices with untrusted MAC addresses from accessing an interface and improves device security.

7.2 Reference

The following table lists the references of this document.

Document	Description	Remarks
IEEE 802.1D	Standard for Information technology--Telecommunications and information exchange between systems--IEEE standard for local and metropolitan area networks--Common specifications--Media access control (MAC) Bridges	-
IEEE 802.1Q	IEEE standard for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks	-

7.3 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

7.4 Principles

7.4.1 MAC Address Table

Categories of MAC Address Entries

The S7700 holds one MAC address table (MAC table for short).

The MAC address entry can be classified into the dynamic entry, the static entry and the blackhole entry.

- The dynamic entry is created by learning the source MAC address. It has aging time.
- The static entry is set by users and is delivered to each SIC. It does not age.
- The blackhole entry is used to discard the frame with the specified source MAC address or destination MAC address. Users manually set the blackhole entries and send them to each SIC. Blackhole entries have no aging time.

The dynamic entry will be lost after the system is reset or the interface board is hot swapped or reset. The static entry and the blackhole entry, however, will not be lost.

Automatically Generated MAC Address Entries

Generally, the S7700 learns the source MAC addresses and then creates MAC address entries. When a device connected to a port of the S7700 sends a packet to the S7700, the S7700 obtains the source MAC address in the frame, and adds the source MAC address and the port to the MAC address table. Since then, when receiving data packets destined for that device, the S7700 can find the outbound port by checking the MAC address table.

The S7700 updates the MAC table at intervals to adapt to the changes of network. The entries in the MAC table will not be valid all the time. Each entry has its own lifetime. If the entry has not been refreshed at the expiration of its lifetime, the S7700 will delete that entry from the MAC table. That lifetime is called aging time. If the entry is refreshed before its lifetime expires, the S7700 resets the aging time for it.

Manually Configured MAC Address Entries

When creating MAC address entries by itself, the S7700 cannot identify whether the packets are from the legal users or the hackers. This threatens the network safety.

Hackers can fake the source MAC address in attack packets. The packet with a forged address enters the S7700 from the other port. Then the S7700 learns a fault MAC table entry. That is why the packets sent to the legal users are forwarded to the hackers.

For security, the network administrator can add static entries to the MAC table manually to bind the user's device and the port of the S7700. In this way, the S7700 can stop the illegal users from stealing data.

By configuring blackhole MAC address entries, you can configure the specified user traffic not to pass through a switch to prevent attacks from unauthorized users.

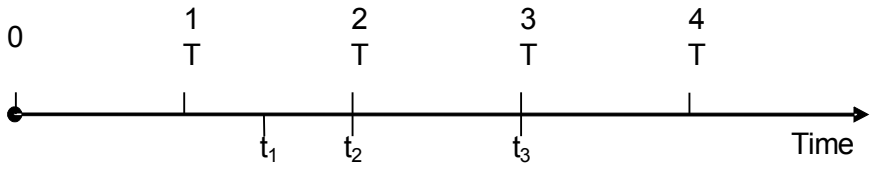
The priority of MAC entries set up by users is higher than that generated by the S7700 itself.

Aging Time of MAC Addresses

To adapt to the changes of networks, the MAC table needs to be updated constantly. The dynamic entries automatically in a MAC address table are not always valid. Each entry has a life cycle. The entry that has never been updated till its life cycle ends will be deleted. This life cycle is called aging time. If the entry is updated before its life cycle ends, the aging time of the entry is recalculated.

Dynamic learned MAC address entries age, whereas static MAC address entries do not age.

Figure 7-1 Aging of MAC addresses



As shown in the preceding figure, the aging time of MAC addresses is set to T. At t_1, packets with the source MAC address 00e0-fc00-0001 and VLAN ID 1 reach an interface. Assume that the interface is added to VLAN 1. If no entry with the MAC address as 00e0-fc00-0001 and the VLAN ID as 1 exists in the MAC address table, the MAC address is added to the MAC address table as a dynamic MAC address entry and the flag of the matching entry is set to 1.

The switch checks all learned dynamic MAC address entries at an interval of T. For example, at t_2, if the switch discovers that the flag of the matching dynamic MAC address entry with the MAC address as 00e0-fc00-0001 and the VLAN ID as 1 is 1, the flag of the matching MAC address entry is set to 0 and the MAC address entry is not deleted. If packets with the source MAC address as 00e0-fc00-0001 and the VLAN ID as 1 enter the switch at t_2 and t_3, the flag of the matching MAC address entry is set to 1 again. If no packet with the source MAC address as 00e0-fc00-0001 and the VLAN ID as 1 enters the switch between t_2 and t_3, the flag of the matching MAC address entry is always 0. At t_3, after discovering that the flag of the matching MAC address entry is 0, the switch assumes that the aging time of the MAC address entry expires and deletes the MAC address entry.

As stated above, the minimum holdtime of a dynamic MAC address entry in the MAC address table ranges from the aging time T to 2 T configured on the switch through automatic aging.

The aging time of MAC addresses is configurable. By setting the aging time of MAC addresses, you can flexibly control the holdtime of learned dynamic MAC address entries in the MAC address table.

7.4.2 Port Security

The port security function changes MAC addresses learned by an interface to secure dynamic MAC addresses or sticky MAC addresses. It prevents devices with untrusted MAC addresses from accessing an interface and improves device security.

Differences between secure dynamic MAC addresses and sticky MAC addresses are:

- Secure dynamic MAC addresses are learned after port security is enabled and will not be aged out by default. Secure dynamic MAC addresses will be lost after the device restarts and the device needs to learn the MAC addresses again.

- Sticky MAC addresses are learned after the sticky MAC function is enabled. Sticky MAC addresses will not be aged out and will exist after the S7700 restarts.

7.4.3 Disabling MAC Address Learning and Limiting the Number of MAC Addresses

The capacity of a MAC address table is limited. Therefore, when hackers forge a large quantity of packets with different source MAC addresses and send the packets to a switch, the MAC address table of the switch may reach its full capacity. After the MAC address table is full, the switch cannot learn the source MAC addresses from received valid packets. A switch limits the number of learned MAC addresses in one of the following modes:

- Disabling MAC address learning on an interface or a VLAN
- Limiting the number of MAC addresses on an interface or a VLAN

After MAC address learning is disabled on an interface or a VLAN, no MAC address entry can be learned on the interface or VLAN. The system deletes the previously learned dynamic MAC entries after the aging time expires. You can also manually delete these entries.

You can limit the maximum number of dynamic MAC address entries on a specified VLAN or interface. After the number of MAC address entries learned by the VLAN or interface reaches the limit, no MAC address entry can be learned on the VLAN or interface until the previously learned MAC address entries age out.

In most cases, attack packets sent by a hacker enter a switch through the same interface. Therefore, you can set the limit on the number of MAC address entries or disable MAC address learning on an interface to prevent attack packets from exhausting the MAC address table.

7.4.4 MAC Address Anti-flapping

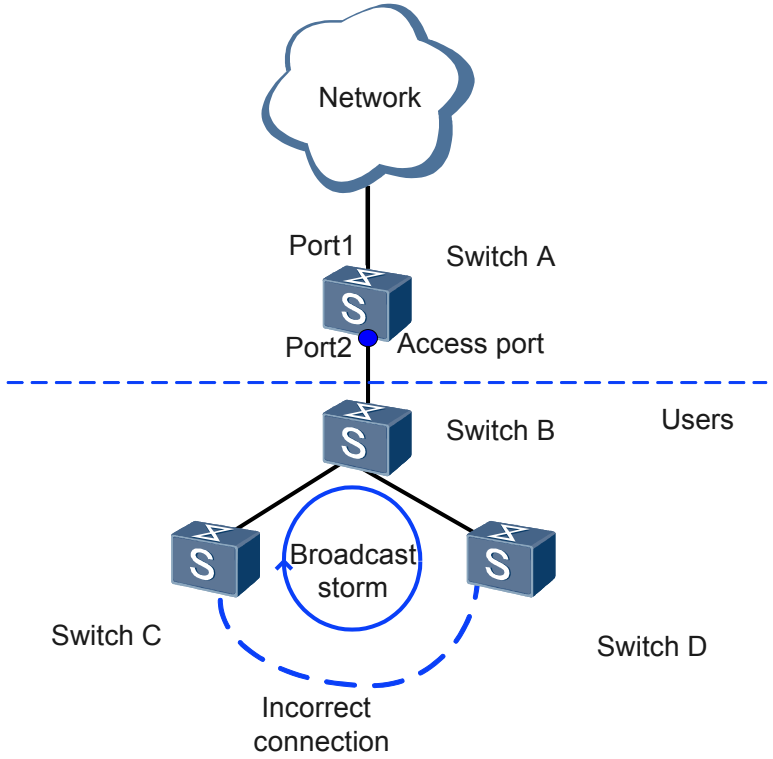
MAC address flapping occurs on a network when the network has a loop or is attacked. To prevent MAC address flapping, you can set MAC address learning priorities for interfaces so that MAC addresses can be learned by correct interfaces. When the same MAC address is learned by interfaces with different priorities, the MAC address entry learned by the interface with the highest priority overrides the MAC address entries learned by other interfaces. You can also determine whether to allow MAC address flapping between interfaces with the same priority.

7.4.5 MAC Address Flapping Detection

The switch can detect MAC address flapping. MAC address flapping occurs when a MAC address is learned on different interfaces. MAC address flapping is usually caused by loops. When a loop occurs and causes a broadcast storm, all the switches affected by the broadcast storm encounter MAC address flapping. Therefore, MAC address flapping detection can be used to check for loops on a network.

When MAC address flapping occurs, the switch can provide diagnosis information, including the flapping MAC address, interfaces between which the MAC address flaps, and VLAN that the interfaces belong to. A loop may exist on the interfaces between which the MAC address flaps. You will know how the loop is generated by checking interfaces where MAC addresses are flapping.

Figure 7-2 MAC address flapping detection



As shown in **Figure 7-2**, SwitchC should not be connected to SwitchD. When the two switches are connected, SwitchB, SwitchC, and SwitchD form a loop. When Port1 of SwitchA receives a broadcast packet, SwitchA forwards the packet to SwitchB. The packet is then sent to Port2 of SwitchA. SwitchA detects that the source MAC address of the packet flaps from Port1 to Port2. If the MAC address flaps between the two ports frequently, SwitchA considers that MAC address flapping occurs.

Because SwitchB, SwitchC, and SwitchD form a loop, broadcast packets sent from other switches can cause frequent MAC address flapping on the three switches. The number of MAC address flapping events on SwitchA is smaller than those on SwitchB, SwitchC, and SwitchD.

NOTE

- By default, MAC address flapping detection is enabled on a switch to help locate loops.
- MAC address flapping detection allows a switch to detect changes in traffic based on learned MAC addresses, but the switch cannot obtain the entire network topology. It is recommended that this function be used on an interface when the interface connects to a user network where loops may occur.

As shown in **Figure 7-2**, the upper-layer network connected to SwitchA is properly planned and supports loop prevention protocols. Therefore, loops seldom occur on the upper-layer network. The loops may occur on the Layer 2 user network because of incorrect device connections. If some devices on the user network do not support any loop prevention protocol, enable MAC address flapping detection on the devices and configure a loop prevention action (error-down or quit-VLAN).

- MAC address flapping detection and the error-down action

You can enable MAC address flapping on SwitchA and set the loop prevention action on the user-side interface (Port2) to error-down. When a loop occurs on the user network, SwitchA detects MAC address flapping on Port2 and shuts down Port2 to reduce impact of the loop on

the entire network. If the error-down action is configured on the network-side interface (Port1), SwitchA shuts down Port1 when MAC address flapping occurs. As a result, network maintenance personnel cannot log in to the switches remotely. Therefore, do not configure the error-down action on network-side interfaces.

- MAC address flapping detection and the quit-VLAN action

You can enable MAC address flapping on SwitchA and set the loop prevention action on the user-side interface (Port2) to quit-VLAN. When a loop occurs on the user network, SwitchA detects MAC address flapping on Port2 and removes Port2 from the VLAN where MAC address flapping occurs. This reduces impact of the loop on the entire network. Similarly, do not configure the quit-VLAN action on network-side interfaces.

NOTE

Do not use the quit-VLAN action together with other dynamic VLAN functions such as GVRP, HVRP, guest VLAN, and voice VLAN.

7.5 Terms and Abbreviations

Terms

Term	Description
MAC address	A hardware address that is used to identify a network node.
VLAN	A Virtual Local Area Network (VLAN) is a switched network and is an end-to-end logical network that is constructed by using the network management software across different network segments and networks. A VLAN forms a logical subnet, that is, a logical broadcast domain. One VLAN can include multiple network devices.
Frame	A bit group that consists of data, one or more addresses, and other protocol control messages. In general, a frame is a protocol data unit at the link layer, which is the second layer in the Open Systems Interconnection (OSI) reference model.

Abbreviations

Abbreviation	Full Spelling
MAC	Media Access Control
VLAN	Virtual Local Area Network

8 STP/RSTP/MSTP

About This Chapter

8.1 Introduction

Definition

Generally, redundant links are used on an Ethernet switching network to provide link backup and enhance network reliability. The use of redundant links, however, may produce loops, causing broadcast storms and rendering the MAC address table unstable. As a result, the communication quality deteriorates, and the communication service may even be interrupted. The Spanning Tree Protocol (STP) is introduced to solve this problem.

STP has a narrow sense and broad sense:

- STP, in the narrow sense, refers to only the STP protocol defined in IEEE 802.1D.
- STP, in the broad sense, refers to the STP protocol defined in IEEE 802.1D, the Rapid Spanning Tree Protocol (RSTP) defined in IEEE 802.1W, and the Multiple Spanning Tree Protocol (MSTP) defined in IEEE 802.1S.

Currently, the following spanning tree protocols are defined:

- STP

  IEEE 802.1D, issued in 1998, defines STP.

  STP, a management protocol at the data link layer, is used to detect and prevent loops on a Layer 2 network. STP blocks redundant links on a Layer 2 network and trims a network into a loop-free tree topology.

  The STP topology, however, converges at a slow speed. Even an edge port cannot be changed to the Forwarding state until twice the amount of time specified by the Forward Delay timer elapses. The default time specified by the forward delay timer is 15 seconds.

- RSTP

  IEEE 802.1W, issued in 2001, defines RSTP.

  RSTP, as an enhancement of STP, achieves fast convergence of the network topology.

  Both RSTP and STP have one defect: All the Virtual Local Area Networks (VLANs) in a LAN share the same spanning tree. As a result, data traffic from different VLANs cannot be evenly balanced. Even worse, packets in some VLANs cannot be forwarded.

  RSTP is backward compatible with STP and can be used together with STP on a network.

- MSTP

  IEEE 802.1S, issued in 2002, defines MSTP.

  MSTP defines a VLAN mapping table in which VLANs are associated with multiple spanning tree instances (MSTIs). In addition, MSTP divides a switching network into multiple regions, each of which has multiple independent MSTIs. In this manner, the entire network is trimmed into a loop-free tree topology, and replication and circular propagation of packets and broadcast storms are prevented on the network. In addition, MSTP provides multiple redundant paths to load-balance VLAN traffic.

  MSTP is compatible with STP and RSTP. **Table 8-1** shows the comparison between STP, RSTP, and MSTP.

Table 8-1 Comparison between STP, RSTP, and MSTP

Spanning Tree Protocol	Characteristics	Usage Scenario
STP	In an STP region, a loop-free tree is generated. Thus, broadcast storms are prevented and redundancy is achieved.	STP or RSTP is used in a scenario where all VLANs share one spanning tree. In this situation, users or services do not need to be differentiated.
RSTP	● In an RSTP region, a loop-free tree is generated. Thus, broadcast storms are prevented and redundancy is achieved. ● RSTP allows fast convergence of the network topology.	
MSTP	● In an MSTP region, a loop-free tree is generated. Thus, broadcast storms are prevented and redundancy is achieved. ● MSTP achieves fast convergence of the network topology. ● MSTP implements load balancing among VLANs. Traffic in different VLANs is transmitted along different paths.	MSTP is used in a scenario where traffic in different VLANs is forwarded through different spanning trees that are independent of each other to implement load balancing. In this situation, users or services are distinguished by using VLANs.

Purpose

After a spanning tree protocol is configured on an Ethernet switching network, it calculates the network topology and implements the following functions to remove network loops:

- Loop cut-off: The potential loops on the network are cut off by blocking redundant links.
- Link redundancy: When an active path becomes faulty, a redundant link can be activated to ensure network connectivity.

8.2 References

The following table lists the references of MSTP.

Document	Description	Remarks
IEEE 802.1D	IEEE Standard for: Local and metropolitan area networks Virtual Bridged Local Area Networks	-
IEEE 802.1S	IEEE Standard for: Local and metropolitan area networks Virtual Bridged Local Area Networks	-
IEEE 802.1W	IEEE Standard for: Local and metropolitan area networks Common specifications	-

8.3 Availability

Involved Network Element

It is unnecessary to cooperate with other network elements.

License Support

This feature can be used without a license.

Version Support

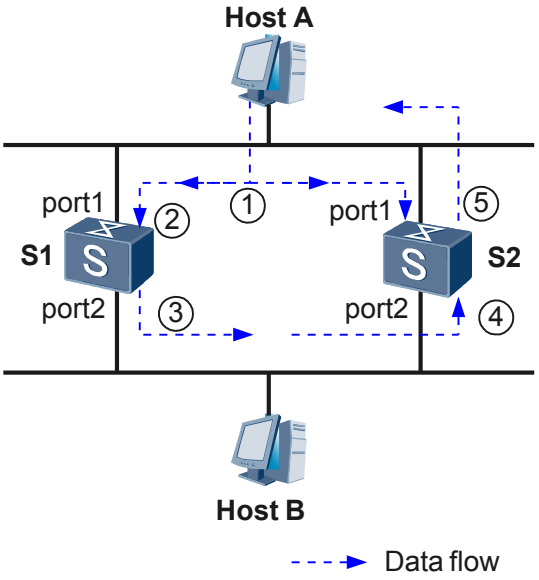
Product	Version
S7700	V100R003, V100R006, V200R001

8.4 Principles of STP/RSTP

8.4.1 Background

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.

Figure 8-1 Networking diagram for a typical LAN



On the network shown in **Figure 8-1**, the following situations may occur:

- Broadcast storms render the network unavailable.

  It is known that loops lead to broadcast storms. In **Figure 8-1**, assume that STP is not enabled on the switching devices. If Host A broadcasts a request, the request is received by port 1 and forwarded by port 2 and port3 on S1. Then, again on S1 and S2, port 2 and port 3 receive the request broadcast by the other and port 1 forwards the request. As such transmission repeats, resources on the entire network are exhausted, causing the network unable to work.

- Flapping of MAC address tables damages MAC address entries.

  As shown in **Figure 8-1**, even update of MAC address entries upon the receipt of unicast packets damages the MAC address table.

  Assume that no broadcast storm occurs on the network. Host A unicasts a packet to Host B. If Host B is temporarily removed from the network at this time, the MAC address entries of Host B on S1 and S2 are deleted. The packet unicast by Host A to Host B is received by port 1 on S1. S1, however, does not have associated MAC address entries. Therefore, the unicast packet is forwarded to port 2 and port 3. Then, port 2 on S2 receives the unicast packet from port 2 on S1 and sends it out through port 3. As such transmission repeats, port 2 and port 3 on S1 and S2 continuously receive unicast packets from Host A. Therefore, S1 and S2 modify the MAC address entries continuously, causing the MAC address table to flap. As a result, MAC address entries are damaged.

8.4.2 Basic Concepts

Basic Design

STP runs at the data link layer. The devices running STP discover loops on the network by exchanging information with each other and trim the ring topology into a loop-free tree topology by blocking a certain interface. In this manner, replication and circular propagation of packets are prevented on the network. In addition, STP prevents the processing performance of network devices from deteriorating.

The devices running STP usually communicate with each other by exchanging configuration Bridge Protocol Data Units (configuration BPDUs). BPDUs are classified into two types:

- Configuration BPDU: used to calculate a spanning tree and maintain the spanning tree topology.
- Topology Change Notification BPDU (TCN BPDU): used to inform upstream devices of a topology change by downstream device.

NOTE

Configuration BPDUs contain sufficient information for devices to calculate the spanning tree. They contain the following information:

- Root bridge ID: is composed of a root bridge priority and the root bridge's MAC address. Each STP network has only one root bridge.
- Cost of the root path: indicates the cost of the shortest path to the root bridge.
- ID of a designated bridge: is composed of a bridge priority and a MAC address.
- ID of a designated port: is composed of a port priority and a port name.
- Message Age: sets the lifetime of a BPDU on the network.
- Max Age: sets the maximum time a BPDU is saved.
- Hello Time: sets the interval at which BPDUs are sent.
- Forward Delay: indicates the time interface status transition takes.

One Root Bridge

A tree topology must have a root. Therefore, the root bridge is introduced by STP.

There is only one root bridge on the entire STP-capable network. The root bridge is the logical center of but is not necessarily at the physical center of the entire network. The root bridge changes dynamically with the network topology.

After the network converges, the root bridge generates and sends out configuration BPDUs at specific intervals. The other devices forward only the configuration BPDUs to advertise the changes in the topology to ensure a stable network.

Two Types of Measurements

The spanning tree is calculated based on two types of measurements: ID and path cost.

- ID

  IDs are classified into Bridge IDs (BIDs) and Port IDs (PIDs).

  - BID

    IEEE 802.1D defines that a BID is composed of a 16-bit bridge priority and a bridge MAC address. The bridge priority occupies the leftmost 16 bits and the MAC address occupies the rightmost 48 bits.

    On an STP-capable network, the device with the smallest BID is selected to be the root bridge.

  - PID

    The PID is composed of a 4-bit port priority and a 12-bit port number. The port priority occupies the left most 4 bits and the port number occupies remaining bits on the right.

    The PID is used to select the designated port.

NOTE

The port priority affects the role of a port in a specified spanning tree instance. For details, see **8.4.4 STP Topology Calculation**.

- Path cost

  The path cost is a port variable and is used to select a link. STP calculates the path cost to select a robust link and blocks redundant links to trim the network into a loop-free tree topology.

  On an STP-capable network, the accumulative cost of the path from a certain port to the root bridge is the sum of the costs of all the segment paths into which the path is separated by the ports on the transit bridges.

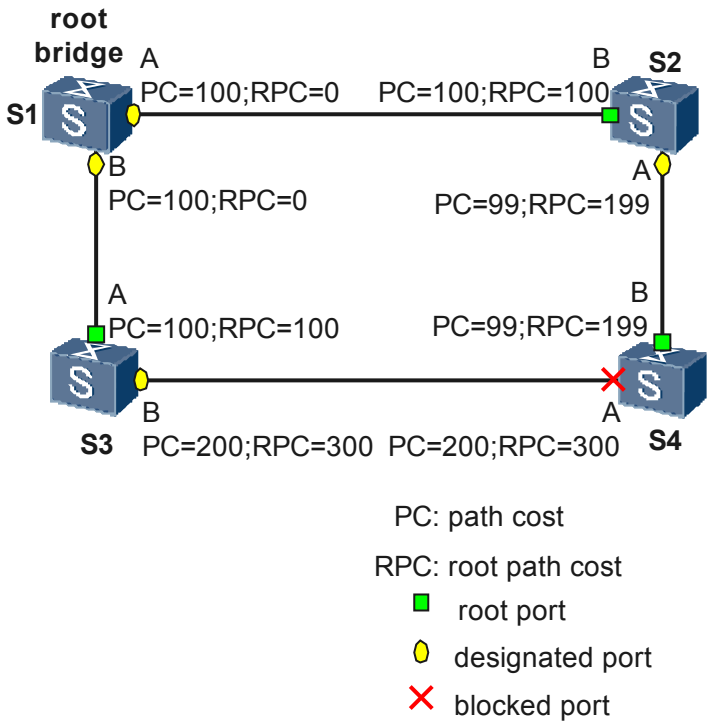
  NOTE

  The rate of an aggregated link is the sum of the rates of all Up member links in the aggregated group.

Three Elements

There are generally three elements used when a ring topology is to be trimmed into a tree topology: root bridge, root port, and designated port. **Figure 8-2** shows the three elements.

Figure 8-2 STP network architecture



- Root bridge

  The root bridge is the bridge with the smallest BID. The smallest BID is discovered by exchanging configuration BPDUs.

- Root port

  The root port is the port with the smallest root path to the root bridge. The root port is determined based on the path cost. Among all STP-capable ports on a network bridge, the

port with the smallest root path cost is the root port. There is only one root port on an STP-capable device, but there is no root port on the root bridge.

- Designated port

  For description of the designated bridge and designated port, see **Table 8-2**.

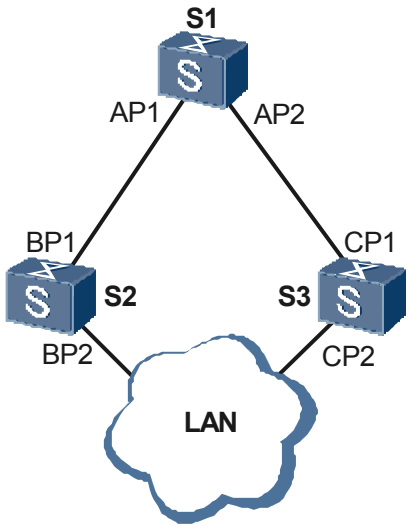
Table 8-2 Description of the designated bridge and designated port

Object	Designated Bridge	Designated Port
Device	Device that forwards configuration BPDUs to a directly connected device	Designated bridge port that forwards configuration BPDUs to a device
LAN	Device that forwards configuration BPDUs to a network segment	Designated bridge port that forwards configuration BPDUs to a network segment.

As shown in **Figure 8-3**, AP1 and AP2 reside on S1; BP1 and BP2 reside on S2; CP1 and CP2 reside on S3.

- S1 sends configuration BPDUs to S2 through AP1. S1 is the designated bridge of S2, and AP1 on S1 is the designated port.
- Two devices, S2 and S3, are connected to the LAN. If S2 is responsible for forwarding configuration BPDUs to the LAN, S2 is the designated bridge of the LAN and BP2 on S2 is the designated port.

Figure 8-3 Networking diagram of the designated bridge and designated port



After the root bridge, root port, and designated port are selected successfully, the entire tree topology is set up. When the topology is stable, only the root port and the designated port forward traffic. All the other ports are in the Blocking state and receive only STP protocol packets instead of forwarding user traffic.

Four Comparison Principles

STP has four comparison principles that form a BPDU priority vector { root BID, total path costs, sender BID, port ID }.

Table 8-3 shows the port information that is carried in the configuration BPDUs.

Table 8-3 Four important fields

Field	Brief Description
Root BID	Each STP-capable network has only one root bridge.
Root path cost	Cost of the path from the port sending configuration BPDUs to the root bridge.
Sender BID	BID of the device sending configuration BPDUs.
Port ID	PID of the port sending configuration BPDUs.

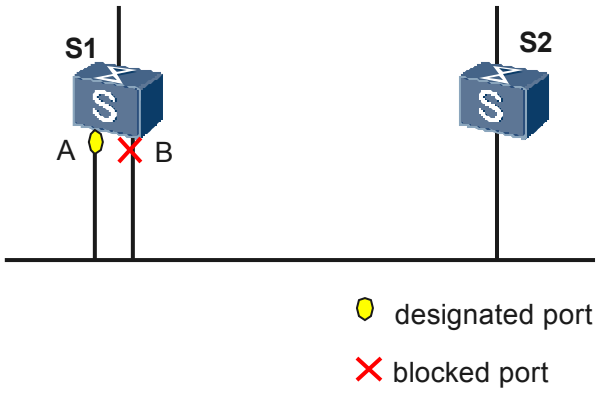
After a device on the STP-capable network receives configuration BPDUs, it compares the fields shown in **Table 8-3** with that of the configuration BPDUs on itself. The four comparison principles are as follows:

NOTE

During the STP calculation, the smaller the value, the higher the priority.

- Smallest BID: used to select the root bridge. Devices running STP select the smallest BID as the root BID shown in **Table 8-3**.
- Smallest root path cost: used to select the root port on a non-root bridge. On the root bridge, the path cost of each port is 0.
- Smallest sender BID: used to select the root port when a device running STP selects the root port between two ports that have the same path cost. The port with a smaller BID is selected as the root port in STP calculation. Assume that the BID of S2 is smaller than that of S3 in **Figure 8-2**. If the path costs in the BPDUs received by port A and port B on S4 are the same, port B becomes the root port.
- Smallest PID: used to block the port with a greater PID but not the port with a smaller PID when the ports have the same path cost. The PIDs are compared in the scenario shown in **Figure 8-4**. The PID of port A on S1 is smaller than that of port B. In the BPDUs that are received on port A and port B, the path costs and BIDs of the sending devices are the same. Therefore, port B with a greater PID is blocked to cut off loops.

Figure 8-4 Topology to which PID comparison is applied



Five Port States

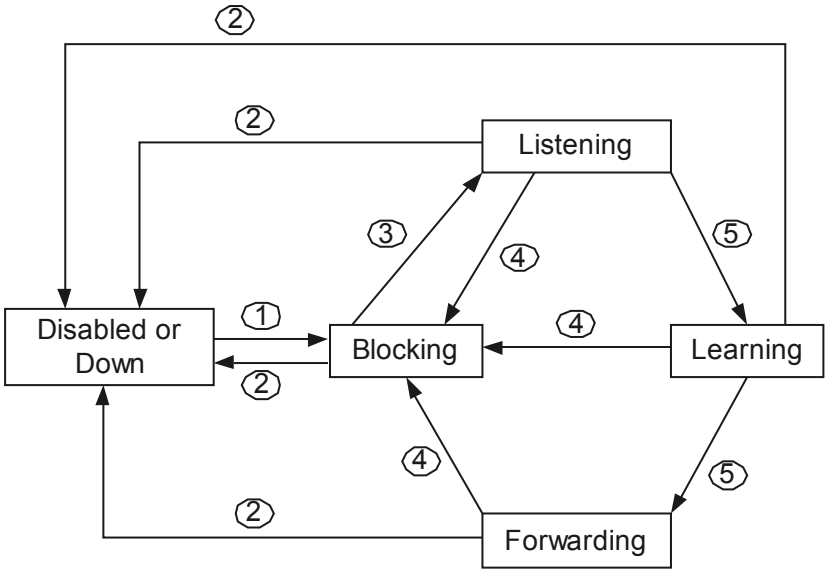
Table 8-4 shows the port status of an STP-capable device.

Table 8-4 Port states

Port State	Purpose	Description
Forwarding	A port in the Forwarding state forwards user traffic and BPDUs.	Only the root port and designated port can enter the Forwarding state.
Learning	When a device has a port in the Learning state, the device creates a MAC address table based on the received user traffic but does not forward user traffic.	This is a transitional state, which is designed to prevent temporary loops.
Listening	All ports are in the Listening state when STP calculation is being implemented to determine port roles.	This is a transitional state.
Blocking	A port in the Blocking state receives and forwards only BPDUs, not user traffic.	This is the final state of a blocked port.
Disabled	A port in the Disabled state does not forward BPDUs or user traffic.	The port is Down.

Figure 8-5 shows the process of the state transition of a port.

Figure 8-5 State transition of a port



1. The port is initialized or enabled.
2. The port is blocked or the link fails.
3. The port is selected as the root or designated port.
4. The port is no longer the root or designated port.
5. The forwarding delay timer expires.

CAUTION

After a device transitions from the MSTP mode to the STP mode, its STP-capable port supports the same port states as those supported by an MSTP-capable port, including the Forwarding, Learning, and Discarding states. For details, see **Table 8-5**.

Table 8-5 Port status

Port Status	Description
Forwarding	A port in the Forwarding state can send and receive BPDUs as well as forward user traffic.
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 not forward user traffic.
Discarding	A port in the Discarding state can only receive BPDUs.

The following parameters affect the STP-capable port states and convergence.

- Hello time

  The Hello timer specifies the interval at which an STP-capable device sends configuration BPDUs to detect link faults.

  When the network topology becomes stable, the change made on the interval takes effect only after a new root bridge takes over. The new root bridge adds certain fields in BPDUs to inform non-root bridges of the change in the interval. After a topology changes, TCN BPDUs will be sent. This interval is irrelevant to the transmission of TCN BPDUs.

- Forward Delay time

  The Forward Delay timer specifies the delay for interface status transition. When a link fault occurs, STP recalculation is performed, causing the structure of the spanning tree to change. The configuration BPDUs generated during STP recalculation cannot be immediately transmitted over the entire network. If the root port and designated port forward data immediately after being selected, transient loops may occur. Therefore, an interface status transition mechanism is introduced by STP. The newly selected root port and designated port do not forward data until an amount of time equal to twice the forward delay has past. In this manner, the newly generated BPDUs can be transmitted over the network before the newly selected root port and designated port forward data, which prevents transient loops.

  NOTE

  The Forward Delay timer specifies the duration of a port spent in both the Listening and Learning states. The default value is 15 seconds. This means that the port stays in the Listening state for 15 seconds and then stays in the Learning state for another 15 seconds. The port in the Listening or Learning state is blocked, which is key to preventing transient loops.

- Max Age time

  The Max Age time specifies the aging time of BPDUs. The Max Age time can be manually configured on the root bridge.

  Configuration BPDUs are transmitted over the entire network, ensuring a unique Max Age value. After a non-root bridge running STP receives a configuration BPDU, the non-root bridge compares the Message Age value with the Max Age value in the received configuration BPDU.

  - If the Message Age value is smaller than or equal to the Max Age value, the non-root bridge forwards the configuration BPDU.
  - If the Message Age value is larger than the Max Age value, the configuration BPDU ages and the non-root bridge directly discards it. In this case, the network size is considered too large and the non-root bridge disconnects from the root bridge.

  NOTE

  If the configuration BPDU is sent from the root bridge, the value of Message Age is 0. Otherwise, the value of Message Age indicates the total time during which a BPDU is sent from the root bridge to the local bridge, including the delay in transmission. In real world situations, each time a configuration BPDU passes through a bridge, the value of Message Age increases by 1.

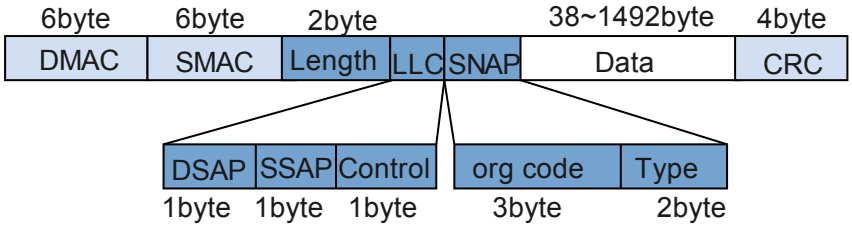
8.4.3 BPDU Format

The BID, path cost, and PID that are described in the previous sections are all carried in BPDUs.

- Configuration BPDUs are heartbeat packets. STP-enabled designated ports send BPDUs at intervals specified by the Hello timer.
- TCN BPDUs are sent only after the device detects network topology changes.

A BPDU is encapsulated into an Ethernet frame. In an Ethernet frame, the destination MAC address is the multicast MAC address 01-80-C2-00-00-00; the value of the Length/Type field is the length of MAC data; in the LLC header, as defined in the IEEE standard, the values of DSAP and SSAP are 0x42 and the value of Control is 0x03; the BPDU header follows the LLC header. **Figure 8-6** shows the format of an Ethernet frame.

Figure 8-6 Format of an Ethernet frame



Configuration BPDU

Configuration BPDUs are most commonly used.

During initialization, each bridge actively sends configuration BPDUs. After the network topology becomes stable, only the root bridge actively sends configuration BPDUs. Other bridges send configuration BPDUs only after receiving configuration BPDUs from upstream devices. A configuration BPDU is at least 35 bytes long, including the parameters such as the BID, path cost, and PID. A BPDU is discarded if both the sender BID and Port ID field values are the same as those of the local port. Otherwise, the BPDU is processed. In this manner, BPDUs containing the same information as that of the local port are not processed.

Table 8-6 shows the format of a BPDU.

Table 8-6 BPDU format

Field	Byte	Description
Protocol Identifier	2	Always 0
Protocol Version Identifier	1	Always 0
BPDU Type	1	Indicates the type of a BPDU. The value is one of the following: ● 0x00: configuration BPDU ● 0x80: TCN BPDU

Field	Byte	Description
Flags	1	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.
Root Identifier	8	Indicates the BID of the current root bridge.
Root Path Cost	4	Indicates the cumulative cost of all links to the root bridge.
Bridge Identifier	8	Indicates the BID of the bridge sending a BPDU.
Port Identifier	2	Indicates the ID of the port sending a BPDU.
Message Age	2	Records the time since the root bridge originally generated the information that a BPDU is derived from. If the configuration BPDU is sent from the root bridge, the value of Message Age is 0. Otherwise, the value of Message Age indicates the total time during which a BPDU is sent from the root bridge to the local bridge, including the delay in transmission. In real world situations, each time a configuration BPDU passes through a bridge, the value of Message Age increases by 1.
Max Age	2	Indicates the maximum time that a BPDU is saved.
Hello Time	2	Indicates the interval at which BPDUs are sent.
Forward Delay	2	Indicates the time spent in the Listening and Learning states.

Figure 8-7 shows the Flags field. Only the leftmost and rightmost bits are used in STP.

Figure 8-7 Format of the Flags field

A configuration BPDU is generated in one of the following scenarios:

- Once the ports are enabled with STP, the designated ports send configuration BPDUs at intervals specified by the Hello timer.
- When a root port receives configuration BPDUs, the device where the root port resides sends a copy of the configuration BPDUs to the specified ports on itself.
- When receiving a configuration BPDU with a lower priority, a designated port immediately sends its own configuration BPDUs to the downstream device.

TCN BPDU

The contents of TCN BPDUs are quite simple, including only three fields: Protocol ID, Version, and Type, as shown in **Table 8-6**. The value of the Type field is 0x80, four bytes in length.

TCN BPDUs are transmitted by each device to its upstream device to notify the upstream device of changes in the downstream topology, until they reach the root bridge. A TCN BPDU is generated in one of the following scenarios:

- Where the port is in the Forwarding state and at least one designated port resides on the device
- Where a designated port receives TCN BPDUs and sends a copy to the root bridge

8.4.4 STP Topology Calculation

Initialization of the Spanning Tree

After all devices on the network are enabled with STP, each device considers itself the root bridge. Each device only transmits and receives BPDUs but does not forward user traffic. All ports are in the Listening state. After exchanging configuration BPDUs, all devices participate in the selection of the root bridge, root port, and designated port.

1. Root bridge selection

   As shown in **Figure 8-8**, the quadruple marked with {} indicates a set of ordered vectors: root BID (S1_MAC and S2_MAC indicates the BIDs of two devices), total path costs, sender BID, and Port ID. Configuration BPDUs are sent at intervals set by the Hello timer. By default, the interval is 2 seconds.

   NOTE

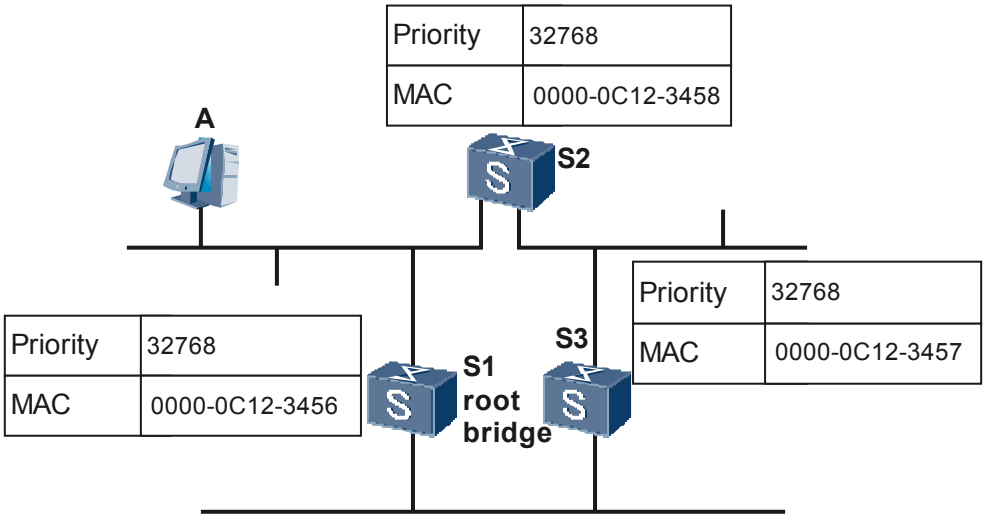
   As each bridge considers itself the root bridge, the value of the root BID field in the BPDU sent by each port is recorded as its BID; the value of the Root Path Cost field is the cumulative cost of all links to the root bridge; the sender BID is the ID of the local bridge; the Port ID is the PID of the local bridge port that sends the BPDU.

Figure 8-8 Exchange of initialization messages

Once a port receives a BPDU with a priority higher than that of itself, the port extracts certain information from the BPDU and synchronizes its own information with the obtained information. The port stops sending the BPDU immediately after saving the updated BPDU.

When sending a BPDU, each device fills in the Sender BID field with its own BID. When a device considers itself the root bridge, the device fills in the Root BID field with its own BID. As shown in **Figure 8-8**, Port B on S2 receives a BPDU with a higher priority from S1, and therefore considers S1 the root bridge. When another port on S2 sends a BPDU, the port fills in its Root BID field with S1_BID. The preceding intercommunication is repeatedly performed between two devices until all devices consider the same device as the root bridge. This indicates that the root bridge is selected. **Figure 8-9** shows the root bridge selection.

Figure 8-9 Diagram of root bridge selection



2. Root port selection

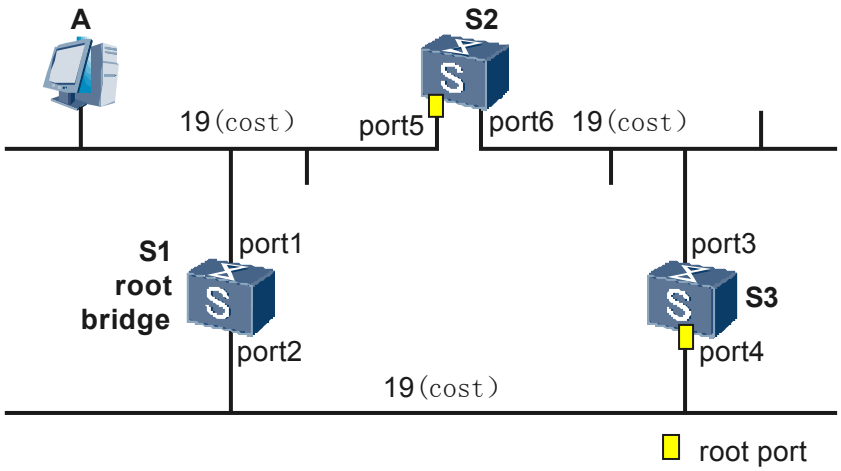
   Each non-root bridge must and can only select one root port.

   After the root bridge has been selected, each bridge determines the cost of each possible path from itself to the root bridge. From these paths, it picks one with the smallest cost (a least-cost path). The port connecting to that path becomes the root port of the bridge. **Figure 8-10** shows the root port selection.

   NOTE

   In the Root Path Cost algorithm, after a port receives a BPDU, the port extracts the value of the Root Path Cost field, and adds the obtained value and the path cost on the itself to obtain the root path cost. The path cost on the port covers only directly-connected path costs. The cost can be manually configured on a port. If the root path costs on two or more ports are the same, the port that sends a BPDU with the smallest sender BID value is selected as the root port.

Figure 8-10 Diagram of root port selection



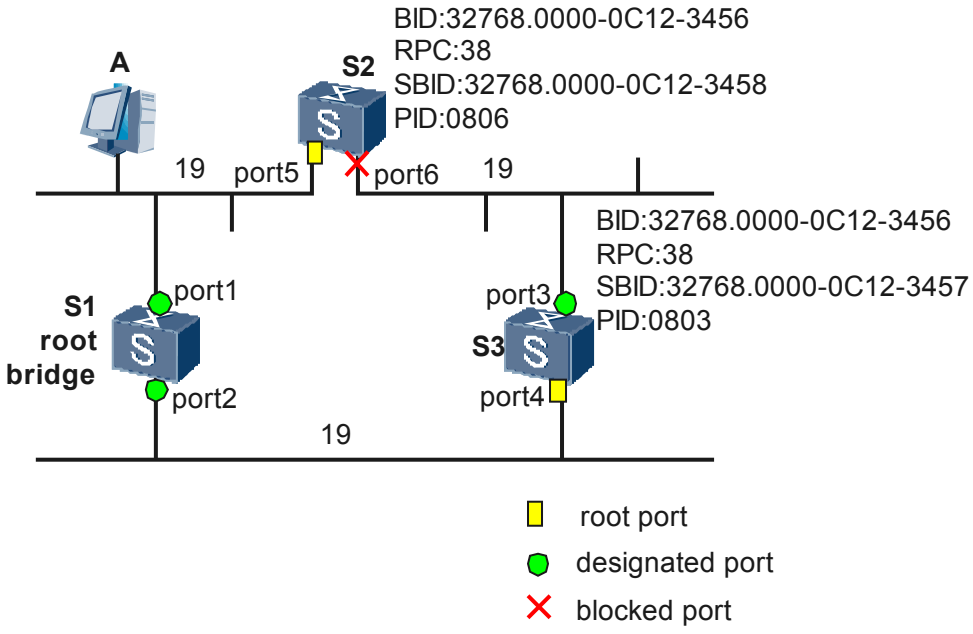
3. Selection of a designated port

   A port that discards lower-priority BPDUs received from other ports, whether on the local device or other devices on the network segment, is called a designated port on the network segment. As shown in **Figure 8-8**, assume that the MAC address of S1 is smaller than that of S2. Port A on S1 is selected as a designated port. The device where a designated port resides is called a designated bridge on the network segment. In **Figure 8-8**, S1 is a designated bridge on the network segment.

   After the network convergence is implemented, only the designated port and root port are in the Forwarding state. The other ports are in the Blocking state. They do not forward user traffic.

   Ports on the root bridge are all designated ports unless loops occur on the root bridge. **Figure 8-11** shows the designated port selection.

Figure 8-11 Diagram of designated port selection



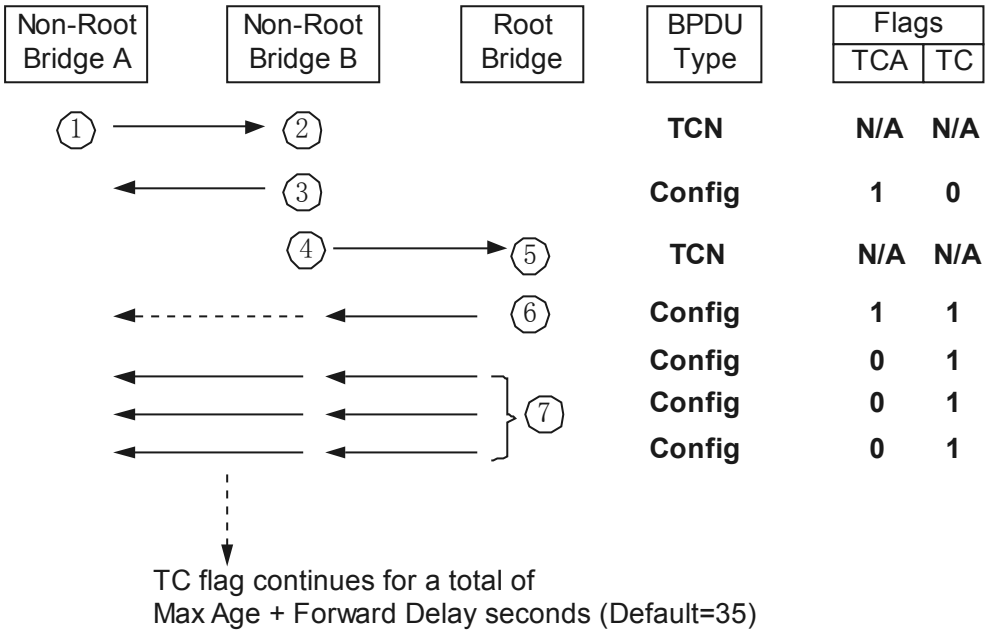
After the Topology Becomes Stable

After the topology becomes stable, the root bridge still sends configuration BPDUs at intervals set by the Hello timer. Each non-root bridge forwards the received configuration BPDUs by using its designated port. If the priority of the received BPDU is higher than that on the non-root bridge, the non-root bridge updates its own BPDU based on the information carried in the received BPDU.

STP Topology Changes

Figure 8-12 shows the packet transmission process after the STP topology changes.

Figure 8-12 Diagram of packet transmission after the topology changes



1. After the network topology changes, a downstream device continuously sends TCN BPDUs to an upstream device.
2. After the upstream device receives TCN BPDUs from the downstream device, only the designated port processes them. The other ports may receive TCN BPDUs but do not process them.
3. The upstream device sets the TCA bit of the Flags field in the configuration BPDUs to 1 and returns the configuration BPDUs to instruct the downstream device to stop sending TCN BPDUs.
4. The upstream device sends a copy of the TCN BPDUs to the root bridge.
5. Steps 1, 2, 3 and 4 are repeated until the root bridge receives the TCN BPDUs.
6. The root bridge sets the TC bit of the Flags field in the configuration BPDUs to 1 to instruct the downstream device to delete MAC address entries.

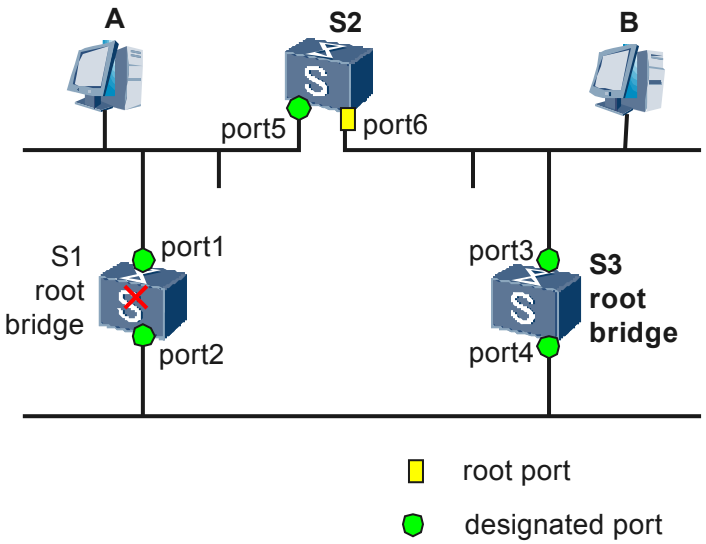
NOTE

- TCN BPDUs are used to inform the upstream device and root bridge of topology changes.
- Configuration BPDUs with the TCA bit being set to 1 are used by the upstream device to inform the downstream device that the topology changes are known and instruct the downstream device to stop sending TCN BPDUs.
- Configuration BPDUs with the TC bit being set to 1 are used by the upstream device to inform the downstream device of topology changes and instruct the downstream device to delete MAC address entries. In this manner, fast network convergence is achieved.

Figure 8-11 is used as an example to show how the network topology converges when the root bridge or designated port of the root bridge becomes faulty.

- The root bridge becomes faulty.

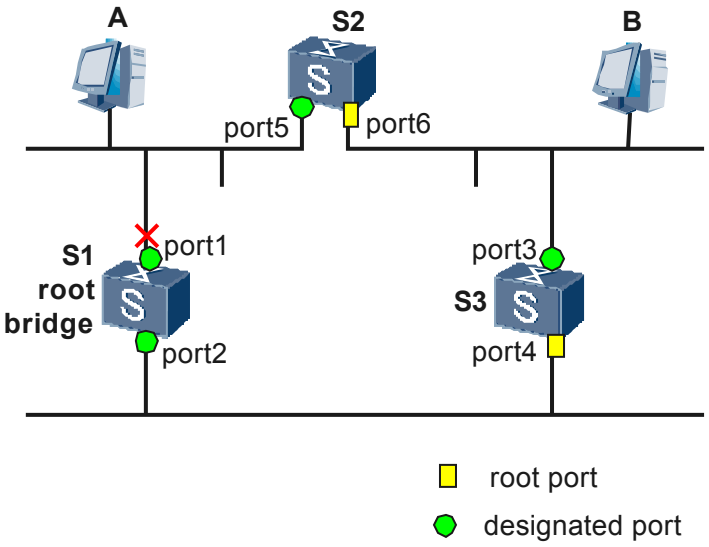
  Figure 8-13 Diagram of topology changes in the case of a faulty root bridge



  As shown in **Figure 8-13**, the root bridge becomes faulty, S2 and S3 will reselect the root bridge. S2 and S3 exchange configuration BPDUs to select the root bridge.

- The designated port of the root bridge becomes faulty.

Figure 8-14 Diagram of topology changes in the case of a faulty designated port on the root bridge



As shown in **Figure 8-14**, the designated port of the root bridge, port 1, becomes faulty. the port6 is selected as the root port through exchanging configuration BPDUs of S2 and S3.

In addition, port6 sends TCN BPDUs after entering the forwarding state. Once the root bridge receives the TCN BPDUs, it will send TC BPDUs to instruct the downstream device to delete MAC address entries.

8.4.5 Evolution from STP to RSTP

In 2001, IEEE 802.1w was published to introduce an extension of the Spanning Tree Protocol (STP), namely, Rapid Spanning Tree Protocol (RSTP). RSTP is developed based on STP but outperforms STP.

Disadvantages of STP

STP ensures a loop-free network but has a slow network topology convergence speed, leading to service deterioration. If the network topology changes frequently, the connections on the STP-capable network are frequently torn down, causing frequent service interruption. Users can hardly tolerate such a situation.

Disadvantages of STP are as follows:

- Port states or port roles are not subtly distinguished, which is not conducive to the learning and deployment for beginners.

  A network protocol that subtly defines and distinguishes different situations is likely to outperform the others.

  - Ports in the Listening, Learning, and Blocking states do not forward user traffic and are not even slightly different to users.
  - The differences between ports in essence never lie in the port states but the port roles from the perspective of use and configuration.

It is possible that the root port and designated port are both in the Listening state or Forwarding state.

- The STP algorithm determines topology changes after the time set by the timer expires, which slows down network convergence.
- The STP algorithm requires a stable network topology. After the root bridge sends configuration BPDUs, other devices forward them until all bridges on the network receive the configuration BPDUs.

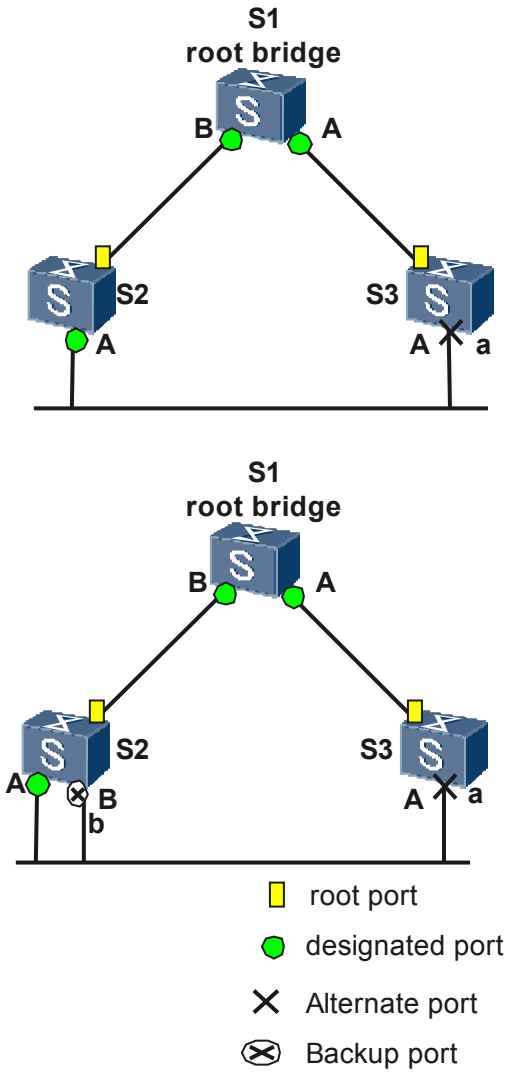
  This also slows down topology convergence.

Advantages of RSTP over STP

To make up for STP disadvantages, RSTP deletes three port states, introduces two port roles, and distinguishes port attributes based on port states and roles to provide more accurate port description. This offers beginners an easy access to protocols and speeds up topology convergence.

- More port roles are defined to simplify the knowledge and deployment of STP.

Figure 8-15 Diagram of port roles



As shown in **Figure 8-15**, RSTP defines four port roles: root port, designated port, alternate port, and backup port.

The functions of the root port and designated port are the same as those defined in STP. The alternate port and backup port are described as follows:

- From the perspective of configuration BPDU transmission:
  - An alternate port is blocked after learning the configuration BPDUs sent by other bridges.
  - A backup port is blocked after learning the configuration BPDUs sent by itself.
- From the perspective of user traffic
  - An alternate port backs up the root port and provides an alternate path from the designated bridge to the root bridge.
  - A backup port backs up the designated port and provides an alternate path from the root bridge to the related network segment.

  After all RSTP-capable ports are assigned roles, topology convergence is completed.

- Port states are redefined in RSTP.

  Port states are simplified from five types to three types. Based on whether a port forwards user traffic and learns MAC addresses, the port is in one of the following states:

  - If a port neither forwards user traffic nor learns MAC addresses, the port is in the Discarding state.
  - If a port does not forward user traffic but learns MAC addresses, the port is in the Learning state.
  - If a port forwards user traffic and learns MAC addresses, the port is in the Forwarding state.

  Table 8-7 shows the comparison between port states in STP and RSTP.

  NOTE

  Port states and port roles are not necessarily related. **Table 8-7** lists states of ports with different roles.

Table 8-7 Comparison between states of STP ports and RSTP ports with different roles

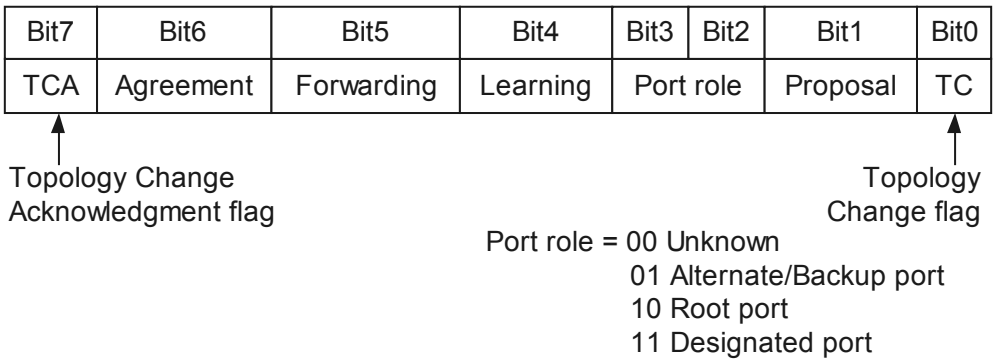
STP Port State	RSTP Port State	Port Role
Forwarding	Forwarding	Root port or designated port
Learning	Learning	Root port or designated port
Listening	Discarding	Root port or designated port
Blocking	Discarding	Alternate port or backup port
Disabled	Discarding	Disabled port

- Configuration BPDUs in RSTP are differently defined. Port roles are described based on the Flags field defined in STP.

  Compared with STP, RSTP slightly redefined the format of configuration BPDUs.

- The value of the Type field is no longer set to 0 but 2. Therefore, the RSTP-capable device always discards the configuration BPDUs sent by an STP-capable device.
- The 6 bits in the middle of the original Flags field are reserved. Such a configuration BPDU is called an RST BPDU, as shown in **Figure 8-16**.

Figure 8-16 Format of the Flags field in an RST BPDU



- Configuration BPDUs are processed in a different manner.
  - Transmission of configuration BPDUs

    In STP, after the topology becomes stable, the root bridge sends configuration BPDUs at an interval set by the Hello timer. A non-root bridge does not send configuration BPDUs until it receives configuration BPDUs sent from the upstream device. This renders the STP calculation complicated and time-consuming. In RSTP, after the topology becomes stable, a non-root bridge sends configuration BPDUs at Hello intervals, regardless of whether it has received the configuration BPDUs sent from the root bridge. Such operations are implemented on each device independently.
  - BPDU timeout period

    In STP, a device has to wait a Max Age period before determining a negotiation failure. In RSTP, if a port does not receive configuration BPDUs sent from the upstream device for three consecutive Hello intervals, the negotiation between the local device and its peer fails.
  - Processing of inferior BPDUs

    In RSTP, when a port receives an RST BPDU from the upstream designated bridge, the port compares the received RST BPDU with its own RST BPDU.

    If its own RST BPDU is superior to the received one, the port discards the received RST BPDU and immediately responds to the upstream device with its own RST BPDU. After receiving the RST BPDU, the upstream device updates its own RST BPDU based on the corresponding fields in the received RST BPDU.

    In this manner, RSTP processes inferior BPDUs more rapidly, independent of any timer that is used in STP.
- Rapid convergence
  - Proposal/agreement mechanism

    When a port is selected as a designated port, in STP, the port does not enter the Forwarding state until a Forward Delay period expires; in RSTP, the port enters the Discarding state, and then the proposal/agreement mechanism allows the port to immediately enter the Forwarding state. The proposal/agreement mechanism must be applied on the P2P links in full duplex mode.

For details, see **8.4.6 Details About RSTP**.

- Fast switchover of the root port

  If the root port fails, the most superior alternate port on the network becomes the root port and enters the Forwarding state. This is because there must be a path from the root bridge to a designated port on the network segment connecting to the alternate port.

  When the port role changes, the network topology accordingly changes. For details, see **8.4.6 Details About RSTP**.

- Edge ports

  In RSTP, a designated port on the network edge is called an edge port. An edge port directly connects to a terminal and does not connect to any other switching devices.

  An edge port does not receive configuration BPDUs, and thus does not participate in the RSTP calculation. It can directly change from the Disabled state to the Forwarding state without any delay, just like an STP-incapable port. If an edge port receives bogus configuration BPDUs from attackers, it is deprived of the edge port attributes and becomes a common STP port. The STP calculation is implemented again, causing network flapping.

- Protection functions

  Table 8-8 shows protection functions provided by RSTP.

Table 8-8 Protection functions

Protection Function	Scenario	Principle
BPDU protection	On a switching device, ports that are directly connected to a user terminal such as a PC or file server are configured as edge ports. Usually, no RST BPDU will be sent to edge ports. If a switching device receives bogus RST BPDUs on an edge port, the switching device automatically sets the edge port to a non-edge port, and performs STP calculation again. This causes network flapping.	After BPDU protection is enabled on a switching device, if an edge port receives an RST BPDU, the switching device shuts down the edge port without depriving of its attributes, and notifies the NMS of the shutdown event. The edge port can be started only by the network administrator. To allow an edge port to automatically start after being shut down, you can configure the auto recovery function and set the delay on the port. In this manner, an edge port starts automatically after the set delay. If the edge port receives RST BPDUs again, the edge port will again be shut down. **NOTE** The smaller the delay is set, the sooner the edge port becomes Up, and the more frequently the edge port alternates between Up and Down. The larger the delay is set, the later the edge port becomes Up, and the longer the service interruption lasts.

Protection Function	Scenario	Principle
Root protection	Due to incorrect configurations or malicious attacks on the network, the root bridge may receive RST BPDUs with a higher priority. Consequently, the valid root bridge is no longer able to serve as the root bridge, and the network topology incorrectly changes. This also causes the traffic that should be transmitted over high-speed links to be transmitted over low-speed links, leading to network congestion.	If a designated port is enabled with the root protection function, the port role cannot be changed. Once a designated port that is enabled with root protection receives RST BPDUs with a higher priority, the port enters the Discarding state and does not forward packets. If the port does not receive any RST BPDUs with a higher priority before a period (generally two Forward Delay periods) expires, the port automatically enters the Forwarding state. **NOTE** Root protection can take effect on only designated ports.
Loop protection	On an RSTP-capable network, the switching device maintains the status of the root port and blocked ports by continually receiving BPDUs from the upstream switching device. If ports cannot receive BPDUs from the upstream switching device due to link congestion or unidirectional link failures, the switching device re-selects a root port. Then, the previous root port becomes a designated port and the blocked ports change to the Forwarding state. As a result, loops may occur on the network.	After loop protection is configured, if the root port or alternate port does not receive RST BPDUs from the upstream switching device for a long time, the switching device notifies the NMS that the port enters the Discarding state. The blocked port remains in the Blocked state and does not forward packets. This prevents loops on the network. The root port or alternate port restores the Forwarding state after receiving new RST BPDUs. **NOTE** Loop protection can take effect on only the root port and alternate ports.

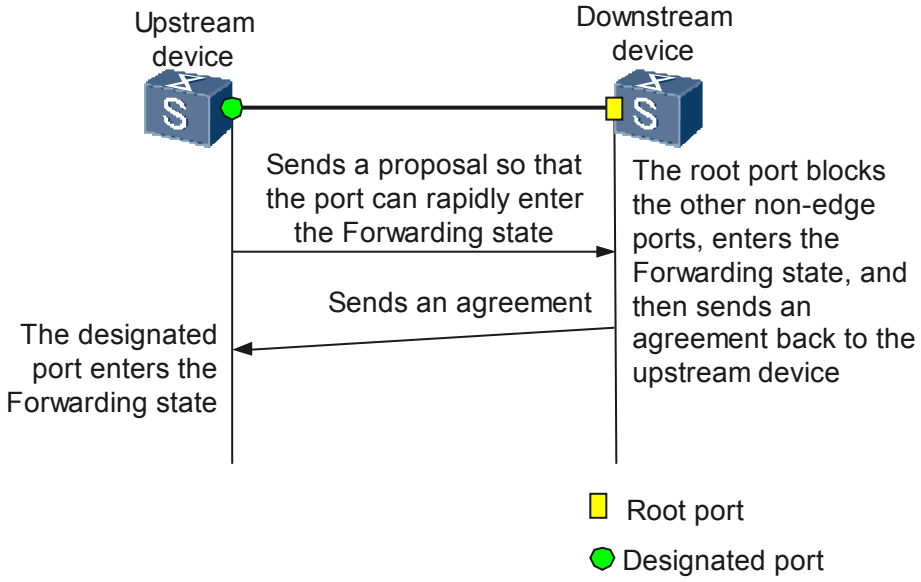
Protection Function	Scenario	Principle
TC BPDU attack defense	After receiving TC BPDUs, a switching device will delete its MAC entries and ARP entries. In the event of a malicious attack by sending bogus TC BPDUs, a switching device receives a large number of TC BPDUs within a short period, and busies itself deleting its MAC entries and ARP entries. As a result, the switching device is heavily burdened, rendering the network rather unstable.	After the TC BPDU attack defense is enabled, the number of times that TC BPDUs are processed by the switching device within a given time period is configurable. If the number of TC BPDUs that the switching device receives within the given time exceeds the specified threshold, the switching device processes TC BPDUs only for the specified number of times. Excess TC BPDUs are processed by the switching device as a whole for once after the specified period expires. In this manner, the switching device is prevented from frequently deleting its MAC entries and ARP entries, and thus is protected against overburden.

8.4.6 Details About RSTP

P/A Mechanism

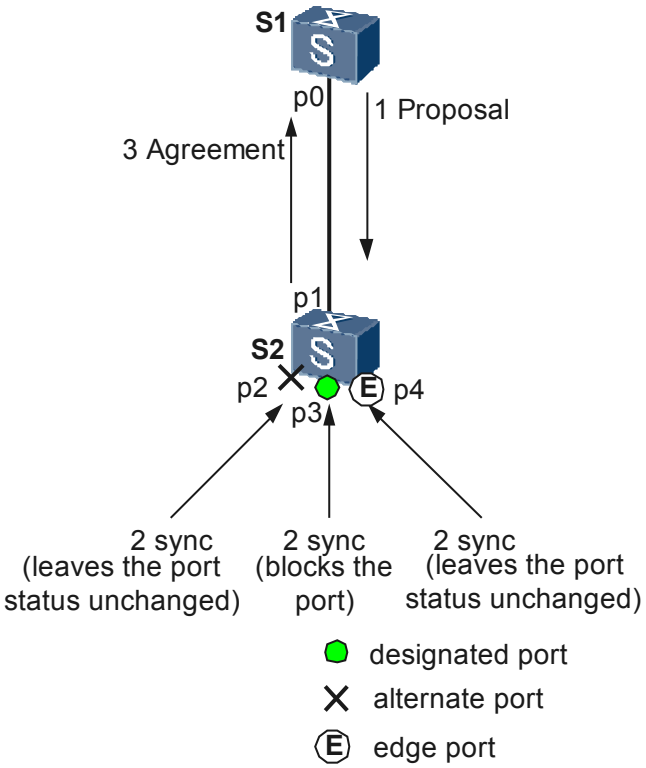
The Proposal/Agreement (P/A) mechanism helps a designated port to enter the Forwarding state as soon as possible. As shown in **Figure 8-17**, the P/A negotiation is performed based on the following port variables:

Figure 8-17 BPDU exchange during the P/A negotiation



1. proposing: When a port is in the Discarding or Learning state, this variable is set to 1. Additionally, an RST BPDU with the Proposal field being 1 is sent to the downstream switching device.
2. proposed: After a port receives an RST BPDU with the Proposal field being 1 from the designated port on the peer device, this variable is set to 1, urging the designated port on this network segment to enter the Forwarding state.
3. sync: After the proposed variable is set to 1, the root port receiving the proposal sets the sync variable to 1 for the other ports on the same device; a non-edge port receiving the proposal enters the Discarding state.
4. synced: After a port enters the Discarding state, it sets its synced variable to 1 in the following manner: If this port is the alternate, backup, or edge port, it will immediately set its synced variable to 1. If this port is the root port, it will monitor the synced variables of the other ports. After the synced variables of all the other ports are set to 1, the root port sets its synced variable to 1, and sends an RST BPDU with the Agreement field being 1.
5. agreed: After the designated port receives an RST BPDU with the Agreement field being 1 and the port role field indicating the root port, this variable is set to 1. Once the agreed variable is set to 1, this designated port immediately enters the Forwarding state.

Figure 8-18 Schematic diagram for the P/A negotiation



As shown in **Figure 8-18**, a new link is established between the root bridges S1 and S2. On S2, p2 is an alternate port; p3 is a designated port in the Forwarding state; p4 is an edge port. The P/A mechanism works in the following process:

1. p0 and p1 become designated ports and send RST BPDUs.

2. After receiving an RST BPDU with a higher priority, p1 realizes that it will become a root port but not a designated port, and thus it stops sending RST BPDUs.
3. p0 enters the Discarding state, and sends RST BPDUs with the Proposal field being 1.
4. After receiving an RST BPDU with the Proposal field being 1, S2 sets the sync variable to 1 for all its ports.
5. As p2 has been blocked, its status keeps unchanged; p4 is an edge port, and thus it does not participate in calculation. Therefore, only the non-edge designated port p3 needs to be blocked.
6. After p2, p3, and p4 enter the Discarding state, their synced variables are set to 1. The synced variable of the root port p1 is then set to 1, and p1 sends an RST BPDU with the Agreement field being 1 to S1. Except for the Agreement field, which is set to 1, and the Proposal field, which is set to 0, the RST BPDU is the same as that was received.
7. After receiving this RST BPDU, S1 identifies it as a reply to the proposal that it just sent, and thus p0 immediately enters the Forwarding state.

This P/A negotiation process finishes, and S2 continues to perform the P/A negotiation with its downstream device.

Theoretically, STP can quickly select a designated port. To prevent loops, STP has to wait for a period of time long enough to determine the status of all ports on the network. All ports can enter the Forwarding state at least one forward delay later. RSTP is developed to eliminate this bottleneck by blocking non-root ports to prevent loops. By using the P/A mechanism, the upstream port can rapidly enter the Forwarding state.

NOTE

To use the P/A mechanism, ensure that the link between the two devices is a P2P link in full-duplex mode. Once the P/A negotiation fails, a designated port can be selected by performing the STP negotiation after the forwarding delay timer expires twice.

RSTP Topology Change

In RSTP, if a non-edge port changes to the Forwarding state, the topology changes.

After a switching device detects the topology change (TC), it performs the following procedures:

- Start a TC While Timer for every non-edge port. The TC While Timer value doubles the Hello Timer value.

  All MAC addresses learned by the ports whose status changes are cleared before the timer expires.

  These ports send RST BPDUs with the TC field being 1. Once the TC While Timer expires, they stop sending the RST BPDUs.
- After another switching device receives the RST BPDU, it clears the MAC addresses learned by all ports excluding the one that receives the RST BPDU. The device then starts a TC While Timer for all non-edge ports and the root port, the same as the preceding process.

In this manner, RST BPDUs flood the network.

Interoperability Between RSTP and STP

When RSTP switches to STP, RSTP loses its advantages such as fast convergence.

On a network where both STP-capable and RSTP-capable devices are deployed, STP-capable devices ignore RST BPDUs; if a port on an RSTP-capable device receives a configuration BPDU

from an STP-capable device, the port switches to the STP mode after two Hello intervals and starts to send configuration BPDUs. In this manner, RSTP and STP are interoperable.

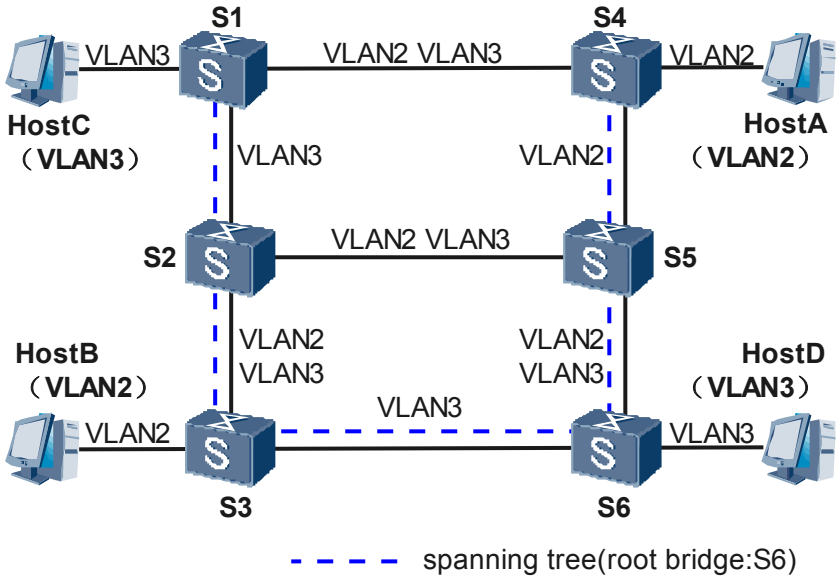
After STP-capable devices are removed, Huawei RSTP-capable datacom devices can switch back to the RSTP mode.

8.5 MSTP Principles

8.5.1 MSTP Background

RSTP, an enhancement to STP, implements fast convergence of the network topology. There is a defect for both RSTP and STP: All VLANs on a LAN use one spanning tree, and thus VLAN-based load balancing cannot be performed. Once a link is blocked, it will no longer transmit traffic, wasting bandwidth and causing the failure in forwarding certain VLAN packets.

Figure 8-19 STP/RSTP defect



On the network shown in **Figure 8-19**, STP or RSTP is enabled. The broken line shows the spanning tree. S6 is the root switching device. The links between S1 and S4 and between S2 and S5 are blocked. VLAN packets are transmitted by using the corresponding links marked with "VLAN2" or "VLAN3."

Host A and Host B belong to VLAN 2 but they cannot communicate with each other because the link between S2 and S5 is blocked and the link between S3 and S6 denies packets from VLAN 2.

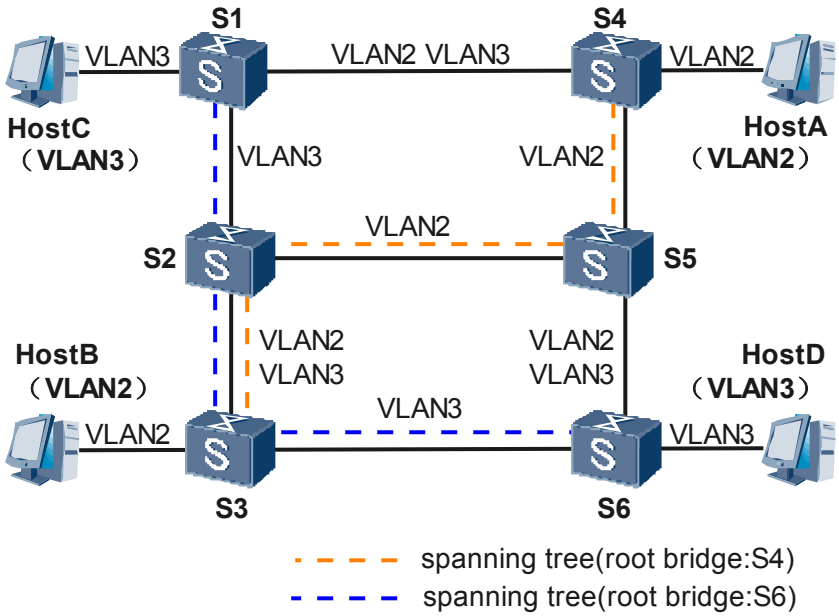
To fix the defect of STP and RSTP, the IEEE released 802.1s in 2002, defining the Multiple Spanning Tree Protocol (MSTP). MSTP implements fast convergence and provides multiple paths to load balance VLAN traffic.

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

NOTE

An instance is a collection of VLANs. Binding multiple VLANs to an instance saves communication costs and reduces resource usage. The topology of each MSTI is calculated independent of one another, and traffic can be balanced among MSTIs. Multiple VLANs that have the same topology can be mapped to one instance. The forwarding status of the VLANs for a port is determined by the port status in the MSTI.

Figure 8-20 Multiple spanning trees in an MST region



As shown in **Figure 8-20**, MSTP maps VLANs to MSTIs in the VLAN mapping table. Each VLAN can be mapped to only one MSTI. This means that traffic of a VLAN can be transmitted in only one MSTI. An MSTI, however, can correspond to multiple VLANs.

Two spanning trees are calculated:

- MSTI 1 uses S4 as the root switching device to forward packets of VLAN 2.
- MSTI 2 uses S6 as the root switching device to forward packets of VLAN 3.

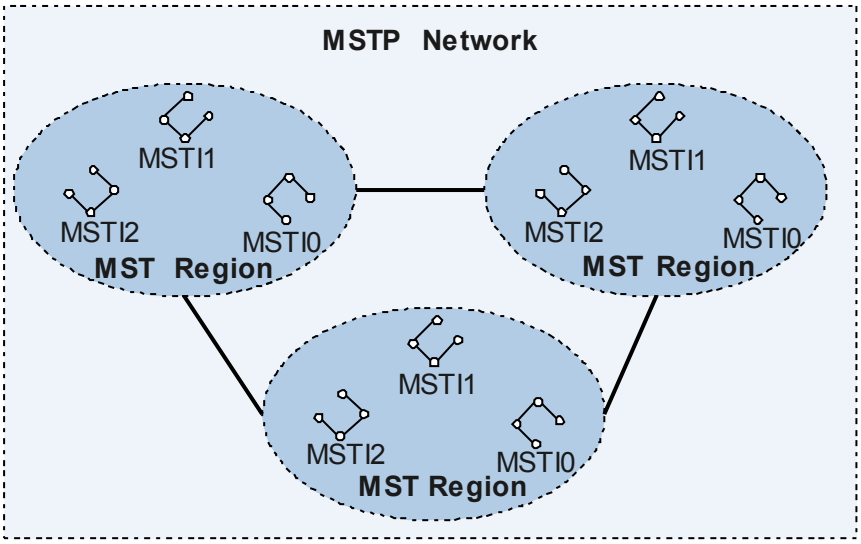
In this manner, devices within the same VLAN can communicate with each other; packets of different VLANs are load balanced along different paths.

8.5.2 Basic MSTP Concepts

MSTP Network Hierarchy

As shown in **Figure 8-21**, the MSTP network consists of one or more MST regions. Each MST region contains one or more MSTIs. An MSTI is a tree network consisting of switching devices running STP, RSTP, or MSTP.

Figure 8-21 MSTP network hierarchy



MST Region

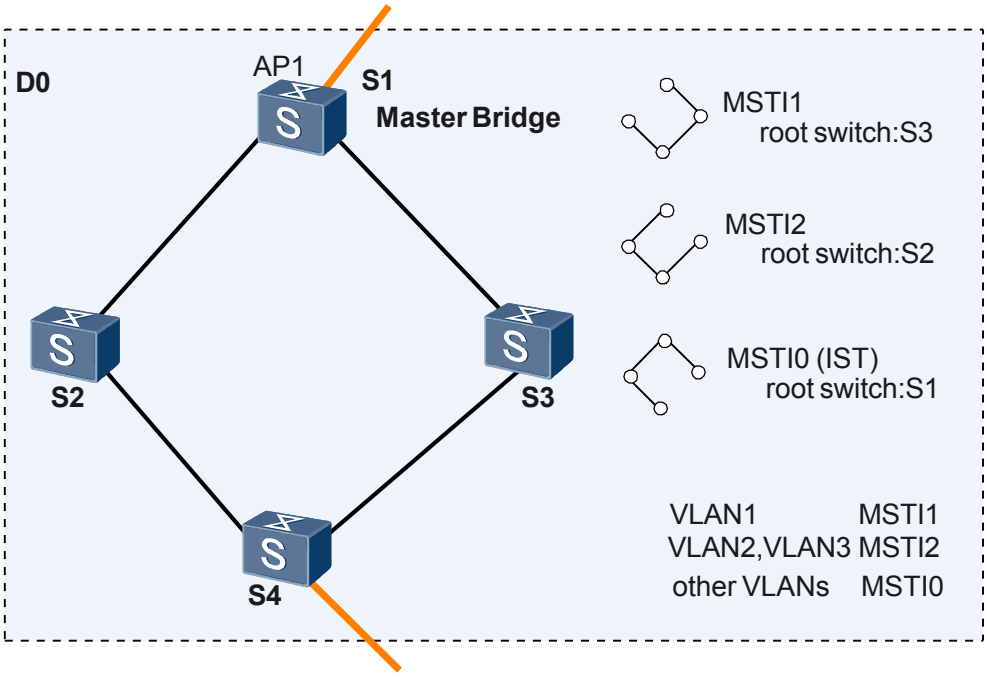
An MST region contains multiple switching devices and network segments between them. The switching devices of one MST region have the following characteristics:

- MSTP-enabled
- Same region name
- Same VLAN-MSTI mappings
- Same MSTP revision level

A LAN can comprise several MST regions that are directly or indirectly connected. Multiple switching devices can be grouped into an MST region by using MSTP configuration commands.

As shown in **Figure 8-22**, the MST region D0 contains the switching devices S1, S2, S3, and S4, and has three MSTIs.

Figure 8-22 MST region



VLAN Mapping Table

The VLAN mapping table is an attribute of the MST region. It describes mappings between VLANs and MSTIs.

As shown in **Figure 8-22**, the mappings in the VLAN mapping table of the MST region D0 are as follows:

- VLAN 1 is mapped to MSTI 1.
- VLAN 2 and VLAN 3 are mapped to MSTI 2.
- Other VLANs are mapped to MSTI 0.

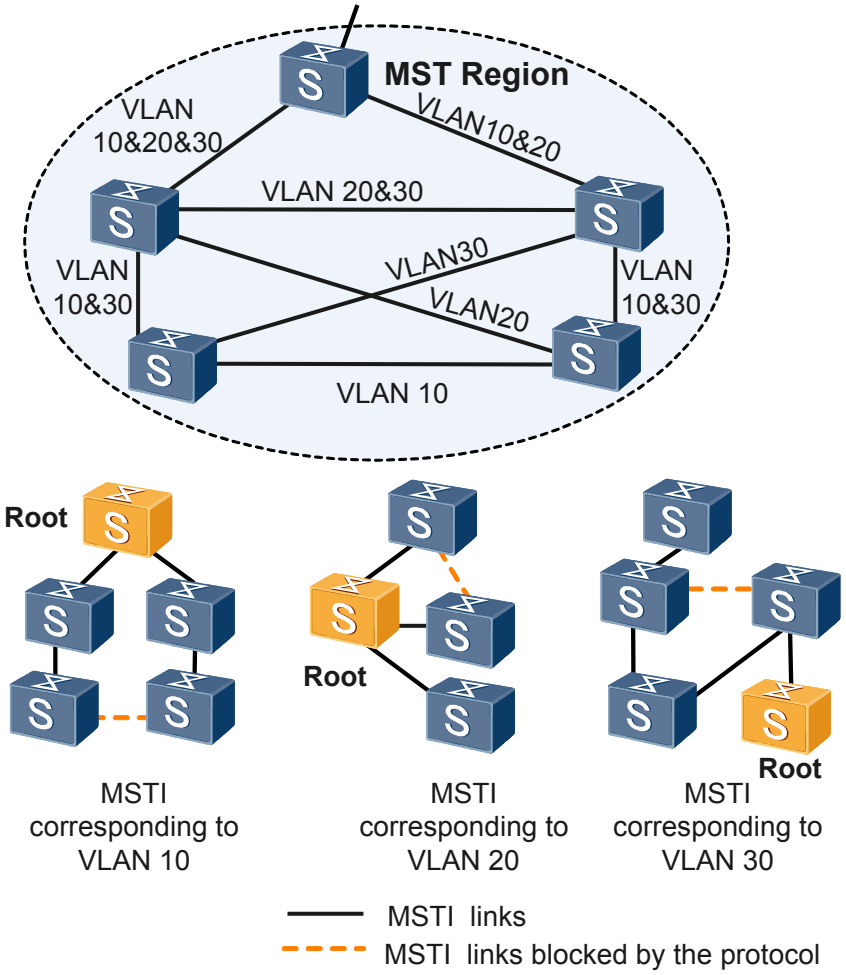
Regional Root

Regional roots are classified into Internal Spanning Tree (IST) and MSTI regional roots.

In the region B0, C0, and D0 on the network shown in **Figure 8-24**, the switching devices closest to the Common and Internal Spanning Tree (CIST) root are IST regional roots.

An MST region can contain multiple spanning trees, each called an MSTI. An MSTI regional root is the root of the MSTI. On the network shown in **Figure 8-23**, each MSTI has its own regional root.

Figure 8-23 MSTI



MSTIs are independent of each other. an MSTI can correspond to one or more VLANs, but a VLAN can be mapped to only one MSTI.

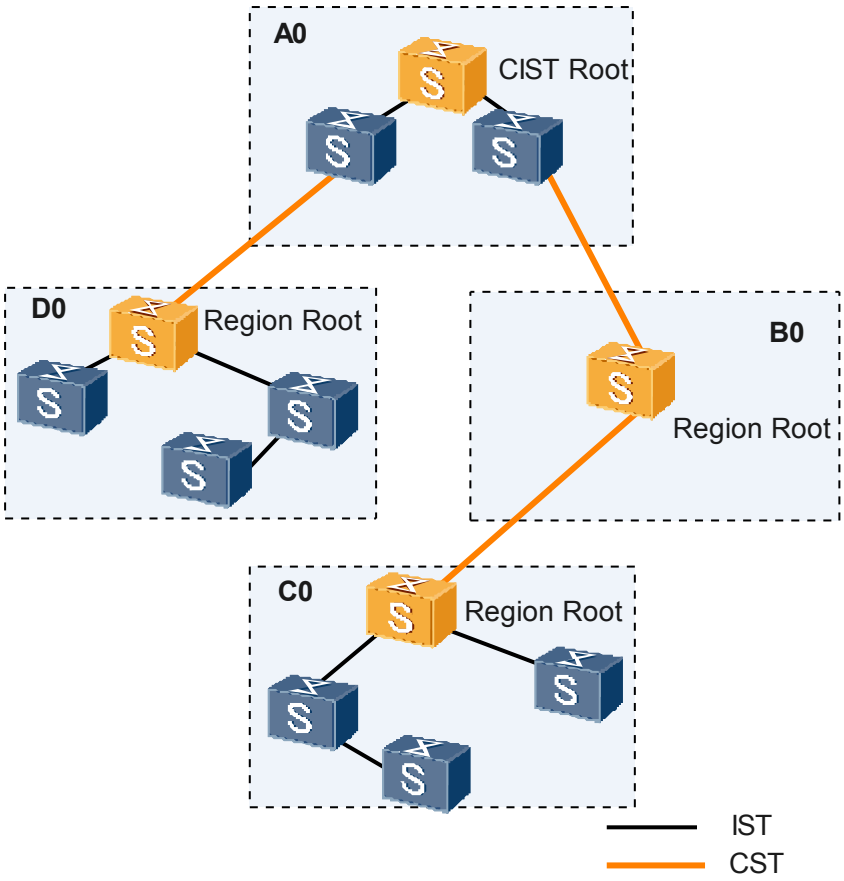
Master Bridge

The master bridge is the IST master, which is the switching device closest to the CIST root in a region, for example, S1 shown in **Figure 8-22**.

If the CIST root is in an MST region, the CIST root is the master bridge of the region.

CIST Root

Figure 8-24 MSTP network



On the network shown in **Figure 8-24**, the CIST root is the root bridge of the CIST. The CIST root is a device in A0.

CST

A Common Spanning Tree (CST) connects all the MST regions on a switching network.

If each MST region is considered a node, the CST is calculated by using STP or RSTP based on all the nodes.

As shown in **Figure 8-24**, the MST regions are connected to form a CST.

IST

An IST resides within an MST region.

An IST is a special MSTI with the MSTI ID being 0, called MSTI 0.

An IST is a segment of the CIST in an MST region.

As shown in **Figure 8-24**, the switching devices in an MST region are connected to form an IST.

CIST

A CIST, calculated by using STP or RSTP, connects all the switching devices on a switching network.

As shown in **Figure 8-24**, the ISTs and the CST form a complete spanning tree, the CIST.

SST

A Single Spanning Tree (SST) is formed in either of the following situations:

- A switching device running STP or RSTP belongs to only one spanning tree.
- An MST region has only one switching device.

As shown in **Figure 8-24**, the switching device in B0 forms an SST.

Port Role

Based on RSTP, MSTP has two additional port types. MSTP ports can be root ports, designated ports, alternate ports, backup ports, edge ports, master ports, and regional edge port.

The functions of root ports, designated ports, alternate ports, and backup ports have been defined in RSTP. **Table 8-9** lists all port roles in MSTP.

NOTE

Except edge ports, all ports participate in MSTP calculation.

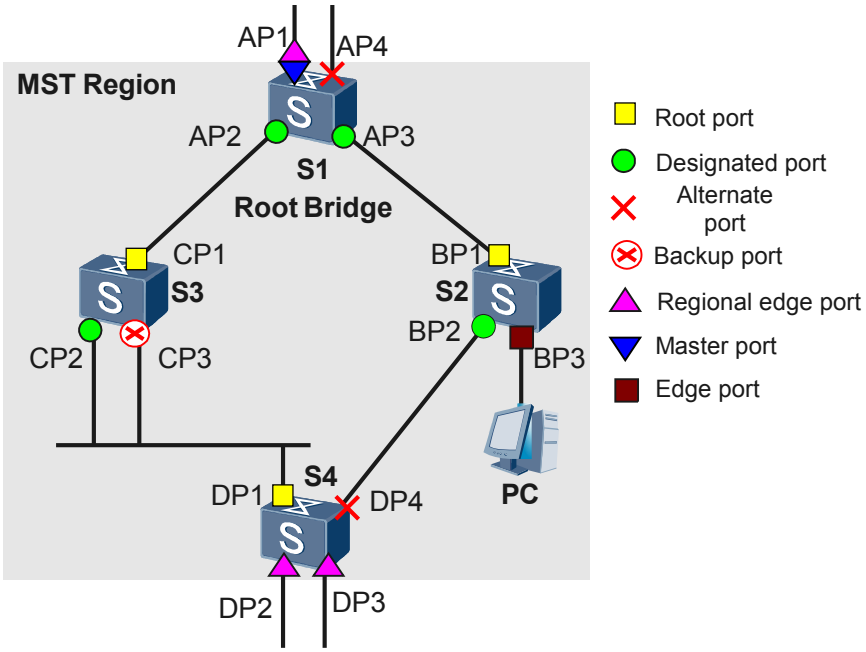
A port can play different roles in different spanning tree instances.

Table 8-9 Port roles

Port Role	Description
Root port	A root port is the non-root bridge port closest to the root bridge. Root bridges do not have root ports. Root ports are responsible for sending data to root bridges. As shown in **Figure 8-25**, S1 is the root; CP1 is the root port on S3; BP1 is the root port on S2; DP1 is the root port on S4.
Designated port	The designated port on a switching device forwards bridge protocol data units (BPDUs) to the downstream switching device. As shown in **Figure 8-25**, AP2 and AP3 are designated ports on S1; BP2 is a designated port on S2; CP2 is a designated port on S3.
Alternate port	● An alternate port is blocked after it receives a BPDU sent by another switching devices. ● An alternate port provides an alternate path to the root bridge. This path is different than using the root port. As shown in **Figure 8-25**, DP4 and AP4 are alternate ports.

Port Role	Description
Backup port	● A backup port is blocked after it receives a BPDU sent by itself. ● A backup port provides a redundant path to a segment and is the backup for the root port. As shown in **Figure 8-25**, CP3 is a backup port.
Master port	A master port is on the shortest path connecting MST regions to the CIST root. BPDUs of an MST region are sent to the CIST root through the master port. Master ports are special regional edge ports, functioning as root ports on ISTs or CISTs and master ports in instances. As shown in **Figure 8-25**, S1, S2, S3, and S4 form an MST region. AP1 on S1, being the nearest port in the region to the CIST root, is the master port.
Regional edge port	A regional edge port is located at the edge of an MST region and connects to another MST region or an SST. During MSTP calculation, the roles of a regional edge port in the MSTI and the CIST instance are the same. If the regional edge port is the master port in the CIST instance, it is the master port in all the MSTIs in the region. As shown in **Figure 8-25**, AP1, DP2, and DP3 in an MST region are directly connected to other regions, and therefore they are all regional edge ports of the MST region. As shown in **Figure 8-25**, AP1 is a regional edge port and also a master port in the CIST. Therefore, AP1 is the master port in every MSTI in the MST region.
Edge port	An edge port is located at the edge of an MST region and does not connect to any switching device. Generally, edge ports are directly connected to terminals. After MSTP is enabled on a port, edge-port detecting is started automatically. If the port fails to receive BPDU packets within (2 x Forward Delay - 2) seconds, the port is set to an edge port. Otherwise, the port is set to a non-edge port. As shown in **Figure 8-25**, BP3 is an edge port.

Figure 8-25 Port roles



MSTP Port Status

Table 8-10 lists the MSTP port status, which is the same as the RSTP port status.

Table 8-10 Port status

Port Status	Description
Forwarding	A port in the Forwarding state can send and receive BPDUs as well as forward user traffic.
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 not forward user traffic.
Discarding	A port in the Discarding state can only receive BPDUs.

There is no necessary link between the port status and the port role. Table 8-11 lists the relationships between port roles and port status.

Table 8-11 Relationships between port roles and port status

Port Status	Root Port/ Master Port	Designated Port	Regional Edge Port	Alternate Port	Backup Port
Forwarding	Yes	Yes	Yes	No	No
Learning	Yes	Yes	Yes	No	No
Discarding	Yes	Yes	Yes	Yes	Yes

Yes: The port supports this status.

No: The port does not support this status.

8.5.3 MST BPDUs

MSTP calculates spanning trees on the basis of Multiple Spanning Tree Bridge Protocol Data Units (MST BPDUs). By transmitting MST BPDUs, spanning tree topologies are computed, network topologies are maintained, and topology changes are conveyed.

Table 8-12 shows differences between TCN BPDUs, configuration BPDUs defined by STP, RST BPDUs defined by RSTP, and MST BPDUs defined by MSTP.

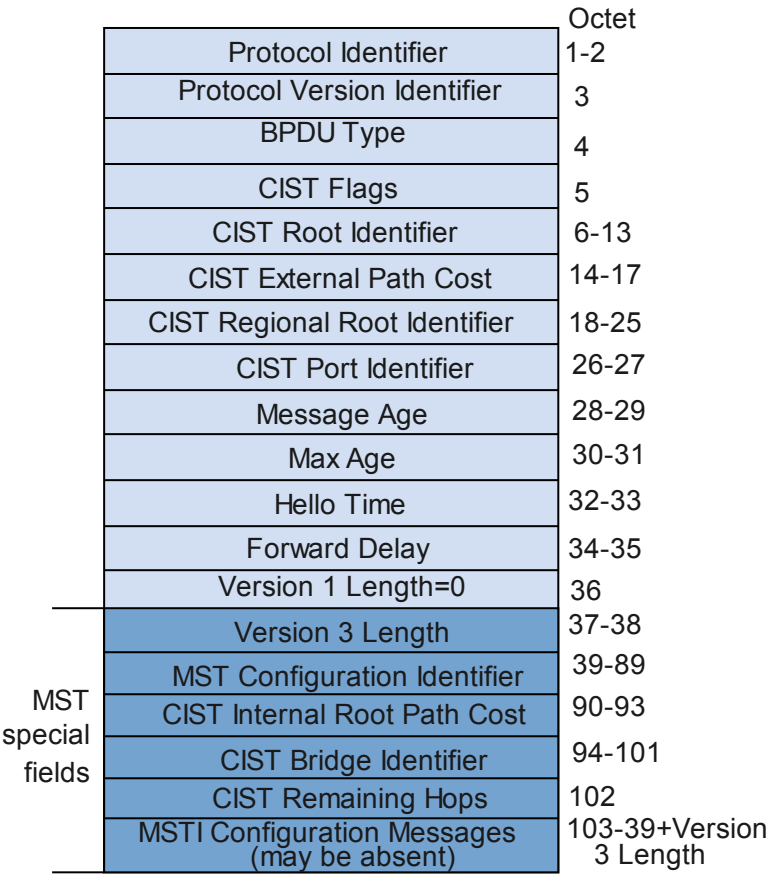
Table 8-12 Differences between BPDUs

Version	Type	Name
0	0x00	Configuration BPDU
0	0x80	TCN BPDU
2	0x02	RST BPDU
3	0x02	MST BPDU

MST BPDU Format

Figure 8-26 shows the MST BPDU format.

Figure 8-26 MST BPDU format



The first 36 bytes of an intra-region or inter-region MST BPDU are the same as those of an RST BPDU.

Fields from the 37th byte of an MST BPDU are MSTP-specific. The field **MSTI Configuration Messages** consists of configuration messages of multiple MSTIs.

Table 8-13 lists the major information carried in an MST BPDU.

Table 8-13 Major information carried in an MST BPDU

Field	Byte	Description
Protocol Identifier	2	Indicates the protocol identifier.
Protocol Version Identifier	1	Indicates the protocol version identifier. 0 indicates STP; 2 indicates RSTP; 3 indicates MSTP.

Field	Byte	Description
BPDU Type	1	Indicates the BPDU type: ● 0x00: Configuration BPDU for STP ● 0x80: TCN BPDU for STP ● 0x02: RST BPDU or MST BPDU
CIST Flags	1	Indicates the CIST flags.
CIST Root Identifier	8	Indicates the CIST root switching device ID.
CIST External Path Cost	4	Indicates the total path costs from the MST region where the switching device resides to the MST region where the CIST root switching device resides. This value is calculated based on link bandwidth.
CIST Regional Root Identifier	8	Indicates the ID of the regional root switching device on the CIST, that is, the IST master ID. If the root is in this region, the CIST Regional Root Identifier is the same as the CIST Root Identifier.
CIST Port Identifier	2	Indicates the ID of the designated port in the IST.
Message Age	2	Indicates the lifecycle of the BPDU.
Max Age	2	Indicates the maximum lifecycle of the BPDU. If the Max Age timer expires, it is considered that the link to the root fails.
Hello Time	2	Indicates the Hello timer value. The default value is 2 seconds.
Forward Delay	2	Indicates the forwarding delay timer. The default value is 15 seconds.
Version 1 Length	1	Indicates the BPDUv1 length, which is fixed to 0.
Version 3 Length	2	Indicates the BPDUv3 length.
MST Configuration Identifier	51	Indicates the MST regional label information, which includes four fields shown in **Figure 8-27**. Interconnected switching devices that are configured with the same MST configuration identifier belong to one region. For details about these four fields, see **Table 8-14**.
CIST Internal Root Path Cost	4	Indicates the total path costs from the local port to the IST master. This value is calculated based on link bandwidth.
CIST Bridge Identifier	8	Indicates the ID of the designated switching device on the CIST.

Field	Byte	Description
CIST Remaining Hops	1	Indicates the remaining hops of the BPDU in the CIST.
MSTI Configuration Messages(may be absent)	16	Indicates the MSTI configuration information. Each MSTI configuration message uses 16 bytes, and thus this field has N x 16 bytes in the case of N MSTIs. **Figure 8-28** shows the structure of a single MSTI configuration message. **Table 8-14** describes every sub-field.

Figure 8-27 MST Configuration Identifier

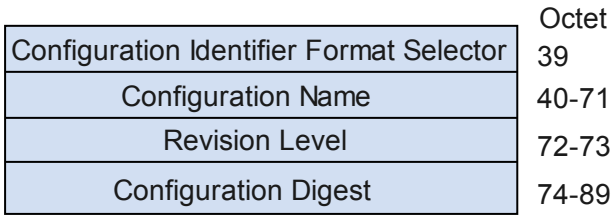


Table 8-14 Description of sub-fields in the MST Configuration Identifier field

Sub-field	Byte	Description
Configuration Identifier Format Selector	1	The value is 0.
Configuration Name	32	Indicates the regional name. The value is a 32-byte string.
Revision Level	2	The value is a 2-byte non-negative integer.
Configuration Digest	16	Indicates a 16-byte digest obtained by encrypting the mappings between VLANs and instances in the region based on the HMAC-MD5 algorithm.

Figure 8-28 MSTI Configuration Messages

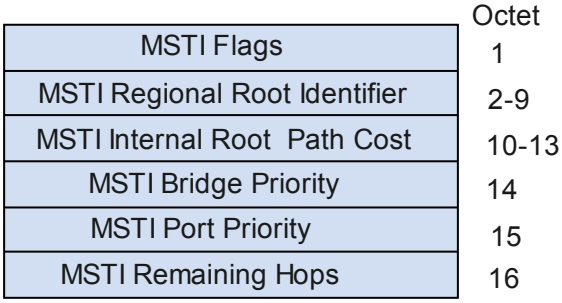


Table 8-15 Description of sub-fields in the MSTI Configuration Messages field

Sub-field	Byte	Description
MSTI Flags	1	Indicates the MSTI flags.
MSTI Regional Root Identifier	8	Indicates the MSTI regional root switching device ID.
MSTI Internal Root Path Cost	4	Indicates the total path costs from the local port to the MSTI regional root switching device. This value is calculated based on link bandwidth.
MSTI Bridge Priority	1	Indicates the priority value of the designated switching device in the MSTI.
MSTI Port Priority	1	Indicates the priority value of the designated port in the MSTI.
MSTI Remaining Hops	1	Indicates the remaining hops of the BPDU in the MSTI.

Configurable MST BPDU Format

Currently, there are two MST BPDU formats:

- dot1s: BPDU format defined in IEEE 802.1s.
- legacy: private BPDU format.

If a port transmits either dot1s or legacy BPDUs by default, the user needs to identify the format of BPDUs sent by the peer, and then runs a command to configure the port to support the peer BPDU format. Once the configuration is incorrect, a loop probably occurs due to incorrect MSTP calculation.

By using the **stp compliance** command, you can configure a port on a Huawei datacom device to automatically adjust the MST BPDU format. With this function, the port automatically adopts the peer BPDU format. The following MST BPDU formats are supported by Huawei datacom devices:

- auto
- dot1s
- legacy

In addition to dot1s and legacy formats, the auto mode allows a port to automatically switch to the BPDU format used by the peer based on BPDUs received from the peer. In this manner, the two ports use the same BPDU format. In auto mode, a port uses the dot1s BPDU format by default, and keeps pace with the peer after receiving BPDUs from the peer.

Configurable Maximum Number of BPDUs Sent by a Port at a Hello Interval

BPDUs are sent at Hello intervals to maintain the spanning tree. If a switching device does not receive any BPDU during a certain period of time, the spanning tree will be re-calculated.

After a switching device becomes the root, it sends BPDUs at Hello intervals. Non-root switching devices adopt the Hello Time value set for the root.

Huawei datacom devices allow the maximum number of BPDUs sent by a port at a Hello interval to be configured as needed.

The greater the Hello Time value, the more BPDUs sent at a Hello interval. Setting the Hello Time to a proper value limits the number of BPDUs sent by a port at a Hello interval. This helps prevent network topology flapping and avoid excessive use of bandwidth resources by BPDUs.

8.5.4 MSTP Topology Calculation

MSTP Principle

In MSTP, the entire Layer 2 network is divided into multiple MST regions, which are interconnected by a single CST. In an MST region, multiple spanning trees are calculated, each of which is called an MSTI. Among these MSTIs, MSTI 0 is also known as the internal spanning tree (IST). Like STP, MSTP uses configuration messages to calculate spanning trees, but the configuration messages are MSTP-specific.

Vectors

Both MSTIs and the CIST are calculated based on vectors, which are carried in MST BPDUs. Therefore, switching devices exchange MST BPDUs to calculate MSTIs and the CIST.

- Vectors are described as follows:
  - The following vectors participate in the CIST calculation:

    { root ID, external root path cost, region root ID, internal root path cost, designated switching device ID, designated port ID, receiving port ID }
  - The following vectors participate in the MSTI calculation:

    { regional root ID, internal root path cost, designated switching device ID, designated port ID, receiving port ID }

  The priorities of vectors in braces are in descending order from left to right.

  Table 8-16 describes the vectors.

Table 8-16 Vector description

Vector Name	Description
Root ID	Identifies the root switching device for the CIST. The root identifier consists of the priority value (16 bits) and MAC address (48 bits).
External root path cost (ERPC)	Indicates the path cost from a CIST regional root to the root. ERPCs saved on all switching devices in an MST region are the same. If the CIST root is in an MST region, ERPCs saved on all switching devices in the MST region are 0s.

Vector Name	Description
Regional root ID	Identifies the MSTI regional root. The regional root ID consists of the priority value (16 bits) and MAC address (48 bits).
Internal root path cost (IRPC)	Indicates the path cost from the local bridge to the regional root. The IRPC saved on a regional edge port is greater than the IRPC saved on a non-regional edge port.
Designated switching device ID	Identifies the nearest upstream bridge on the path from the local bridge to the regional root. If the local bridge is the root or the regional root, this ID is the local bridge ID.
Designated port ID	Identifies the port on the designated switching device connected to the root port on the local bridge. The port ID consists of the priority value (4 bits) and port number (12 bits). The priority value must be a multiple of 16.
Receiving port ID	Identifies the port receiving the BPDU. The port ID consists of the priority value (4 bits) and port number (12 bits). The priority value must be a multiple of 16.

- The vector comparison principle is as follows:

  For a vector, the smaller the priority value, the higher the priority.

  Vectors are compared based on the following rules:

  1. Compare the IDs of the roots.
  2. If the IDs of the roots are the same, compare ERPCs.
  3. If ERPCs are the same, compare the IDs of regional roots.
  4. If the IDs of regional roots are the same, compare IRPCs.
  5. If IRPCs are the same, compare the IDs of designated switching devices.
  6. If the IDs of designated switching devices are the same, compare the IDs of designated ports.
  7. If the IDs of designated ports are the same, compare the IDs of receiving ports.

  If the priority of a vector carried in the configuration message of a BPDU received by a port is higher than the priority of the vector in the configuration message saved on the port, the port replaces the saved configuration message with the received one. In addition, the port updates the global configuration message saved on the device. If the priority of a vector carried in the configuration message of a BPDU received on a port is equal to or lower than the priority of the vector in the configuration message saved on the port, the port discards the BPDU.

CIST Calculation

After completing the configuration message comparison, the switching device with the highest priority on the entire network is selected as the CIST root. MSTP calculates an IST for each MST region, and computes a CST to interconnect MST regions. On the CST, each MST region is considered a switching device. The CST and ISTs constitute a CIST for the entire network.

MSTI Calculation

In an MST region, MSTP calculates an MSTI for each VLAN based on mappings between VLANs and MSTIs. Each MSTI is calculated independently. The calculation process is similar to the process for STP to calculate a spanning tree. For details, see **8.4.4 STP Topology Calculation**.

MSTIs have the following characteristics:

- The spanning tree is calculated independently for each MSTI, and spanning trees of MSTIs are independent of each other.
- MSTP calculates the spanning tree for an MSTI in the manner similar to STP.
- Spanning trees of MSTIs can have different roots and topologies.
- Each MSTI sends BPDUs in its spanning tree.
- The topology of each MSTI is configured by using commands.
- A port can be configured with different parameters for different MSTIs.
- A port can play different roles or have different status in different MSTIs.

On an MSTP-aware network, a VLAN packet is forwarded along the following paths:

- MSTI in an MST region
- CST among MST regions

MSTP Responding to Topology Changes

MSTP topology changes are processed in the manner similar to that in RSTP. For details about how RSTP processes topology changes, see **8.4.6 Details About RSTP**.

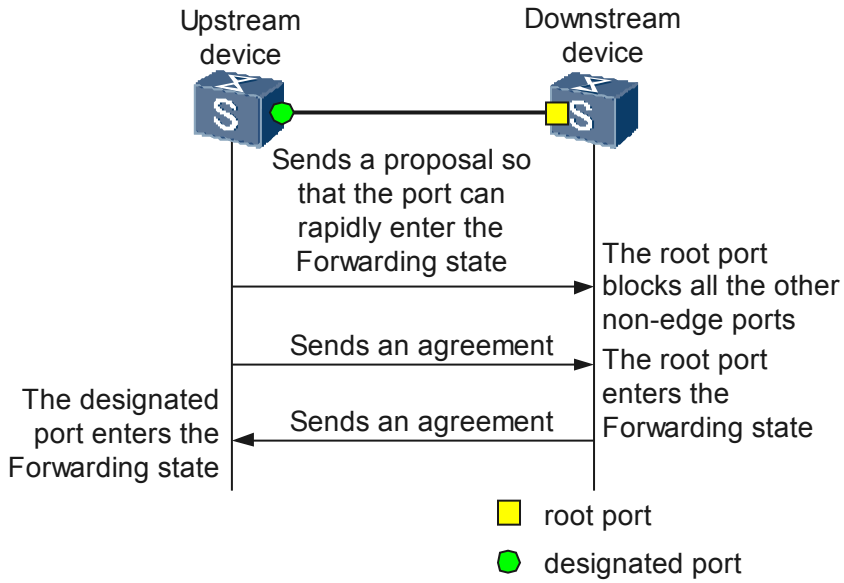
8.5.5 MSTP Fast Convergence

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

- Ordinary P/A

  The ordinary P/A mechanism supported by MSTP is implemented in the same manner as that supported by RSTP. For details about the P/A mechanism supported by RSTP, see **8.4.6 Details About RSTP**.
- Enhanced P/A

Figure 8-29 Enhanced P/A mechanism



As shown in **Figure 8-29**, in MSTP, the P/A mechanism works as follows:

1. The upstream device sends a proposal to the downstream device, indicating that the port connecting to the downstream device wants to enter the Forwarding state as soon as possible. After receiving this BPDU, the downstream device sets its port connecting to the upstream device to the root port, and blocks all non-edge ports.
2. The upstream device continues to send an agreement. After receiving this BPDU, the root port enters the Forwarding state.
3. The downstream device replies with an agreement. After receiving this BPDU, the upstream device sets its port connecting to the downstream device to the designated port, and the port enters the Forwarding state.

By default, Huawei datacom devices use the enhanced P/A mechanism. If a Huawei datacom device needs to communicate with a non-Huawei device that uses the ordinary P/A mechanism, run the **stp no-agreement-check** command to configure the Huawei device to use the ordinary P/A mechanism. In this manner, these two devices can communicate with each other.

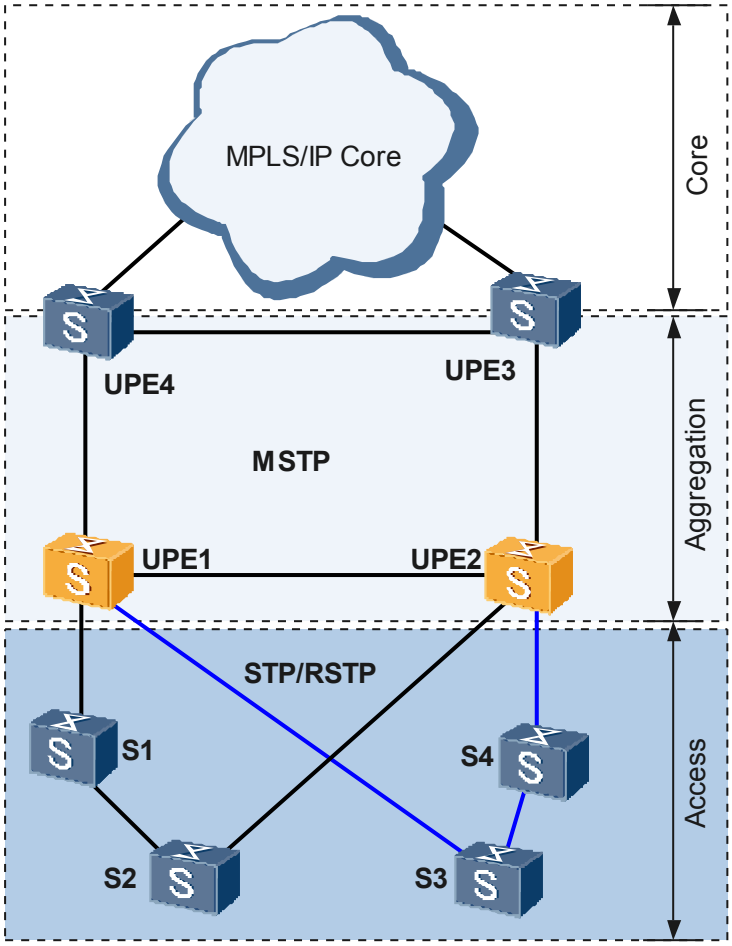
8.5.6 MSTP Multi-Process

Background

On the network shown in **Figure 8-30**:

- UPEs are deployed at the aggregation layer, running MSTP.
- UPE1 and UPE2 are connected by a Layer 2 link.
- Multiple rings are connected to UPE1 and UPE2 through different ports.
- Switching devices on the rings reside at the access layer, running STP or RSTP. In addition, UPE1 and UPE2 work for different carriers, and thus they need to reside on different spanning trees whose topology changes do not affect each other.

Figure 8-30 Application with both MSTP and STP/RSTP



On the network shown in **Figure 8-30**, switching devices and UPEs construct multiple Layer 2 rings. STP must be enabled on these rings to prevent loops. UPE1 and UPE2 are connected to multiple access rings that are independent of each other. The spanning tree protocol cannot calculate a single spanning tree for all switching devices. Instead, the spanning tree protocol must be enabled on each ring to calculate a separate spanning tree.

MSTP supports MSTIs, but these MSTIs must belong to one MST region and devices in the region must have the same configurations. If the devices belong to different regions, MSTP calculates the spanning tree based on only one instance. Assume that devices on the network belong to different regions, and only one spanning tree is calculated in one instance. In this case, the status change of any device on the network affects the stability of the entire network. On the network shown in **Figure 8-30**, the switching devices connected to UPEs support only STP or RSTP but not MSTP. When MSTP-enabled UPEs receive RST BPDUs from the switching devices, the UPEs consider that they and switching devices belong to different regions. As a result, only one spanning tree is calculated for the rings composed of UPEs and switching devices, and the rings affect each other.

To prevent this problem, MSTP multi-process is introduced. MSTP multi-process is an enhancement to MSTP. The MSTP multi-process mechanism allows ports on switching devices to be bound to different processes. MSTP calculation is performed based on processes. In this

manner, only ports that are bound to a process participate in the MSTP calculation for this process. With the MSTP multi-process mechanism, spanning trees of different processes are calculated independently and do not affect each other. The network shown in **Figure 8-30** can be divided into multiple MSTP processes by using MSTP multi-process. Each process takes charge of a ring composed of switching devices. The MSTP processes have the same functions and support MSTIs. The MSTP calculation for one process does not affect the MSTP calculation for another process.

NOTE

MSTP multi-process is applicable to MSTP as well as RSTP and STP.

Purpose

On the network shown in **Figure 8-30**, MSTP multi-process is configured to implement the following:

- Greatly improves applicability of STP to different networking conditions.

  To help a network running different spanning tree protocols run properly, you can bind the devices running different spanning tree protocols to different processes. In this manner, every process calculates a separate spanning tree.
- Improves the networking reliability. For a network composed of many Layer 2 access devices, using MSTP multi-process reduces the adverse effect of a single node failure on the entire network.

  The topology is calculated for each process. If a device fails, only the topology corresponding to the process to which the device belongs changes.
- Reduces the network administrator workload during network expansion, facilitating operation and maintenance.

  To expand a network, you only need to configure new processes, connect the processes to the existing network, and keep the existing MSTP processes unchanged. If device expansion is performed in a process, only this process needs to be modified.
- Implements separate Layer 2 port management

  An MSTP process manages parts of ports on a device. Layer 2 ports on a device are separately managed by multiple MSTP processes.

Principle

- Public link status

  As shown in **Figure 8-30**, the public link between UPE1 and UPE2 is a Layer 2 link running MSTP. The public link between UPE1 and UPE2 is different from the links connecting switching devices to UPEs. The ports on the public link need to participate in the calculation for multiple access rings and MSTP processes. Therefore, the UPEs must identify the process from which MST BPDUs are sent.

  In addition, a port on the public link participates in the calculation for multiple MSTP processes, and obtains different status. As a result, the port cannot determine its status.

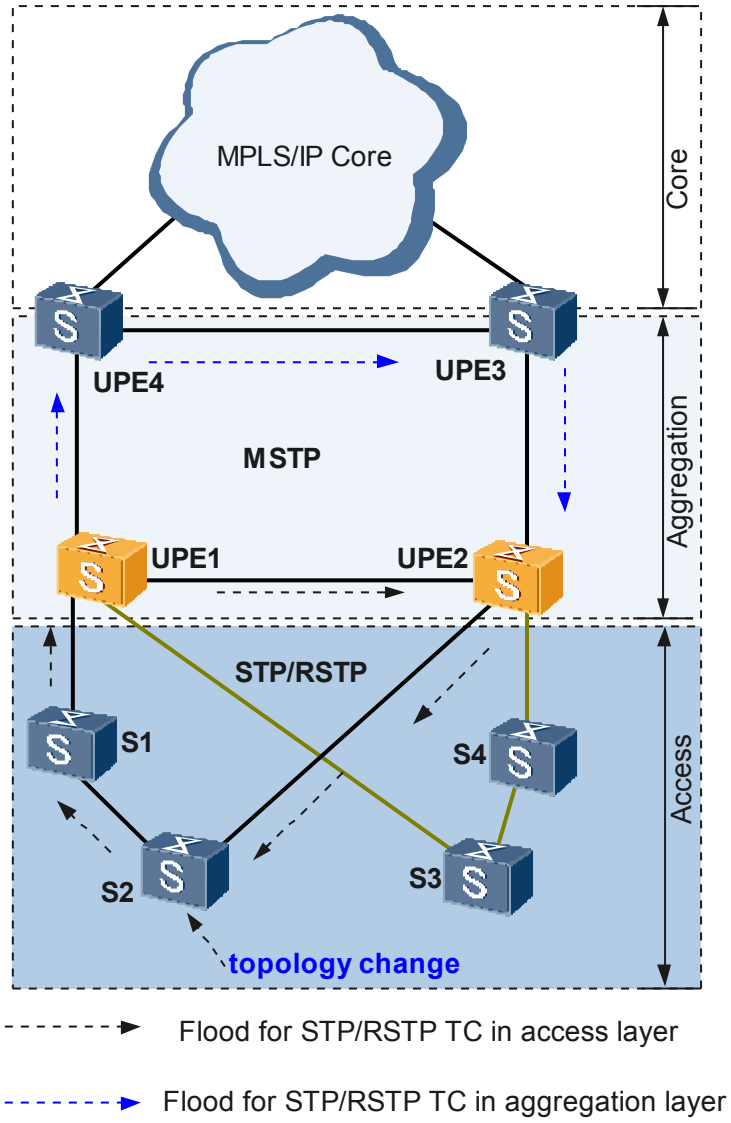
  To prevent this situation, it is defined that a port on a public link always adopts its status in MSTP process 0 when participating in the calculation for multiple MSTP processes.

  NOTE

  After a device normally starts, MSTP process 0 exists by default, and MSTP configurations in the system view and interface view belong to this process.
- Reliability

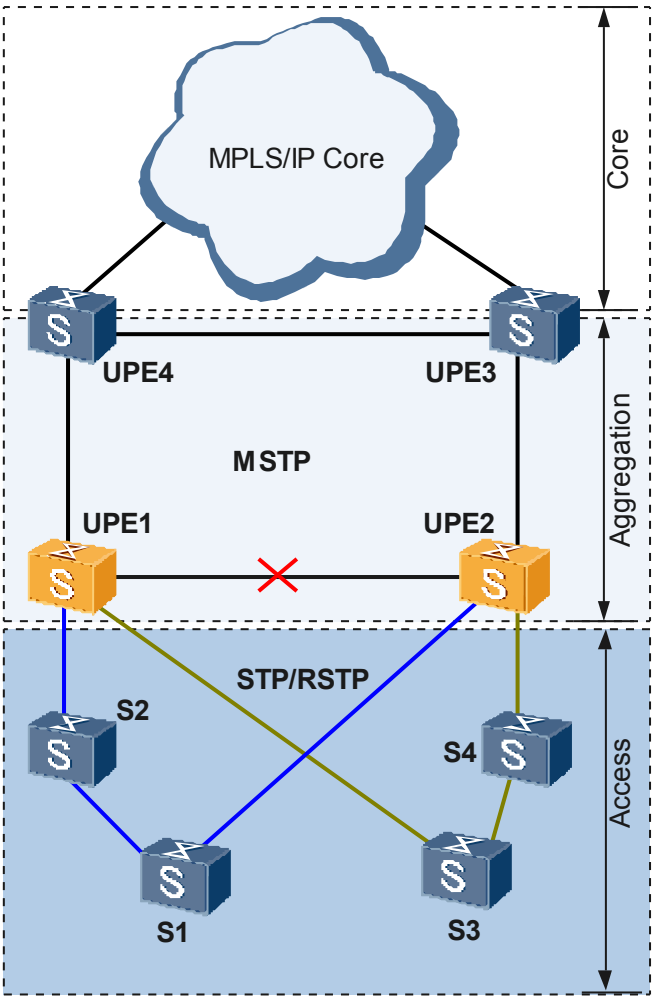
On the network shown in **Figure 8-31**, after the topology of a ring changes, the MSTP multi-process mechanism helps UPEs flood a TC packet to all devices on the ring and prevent the TC packet from being flooded to devices on the other ring. UPE1 and UPE2 update MAC and ARP entries on the ports corresponding to the changed spanning tree.

Figure 8-31 MSTP multi-process topology change



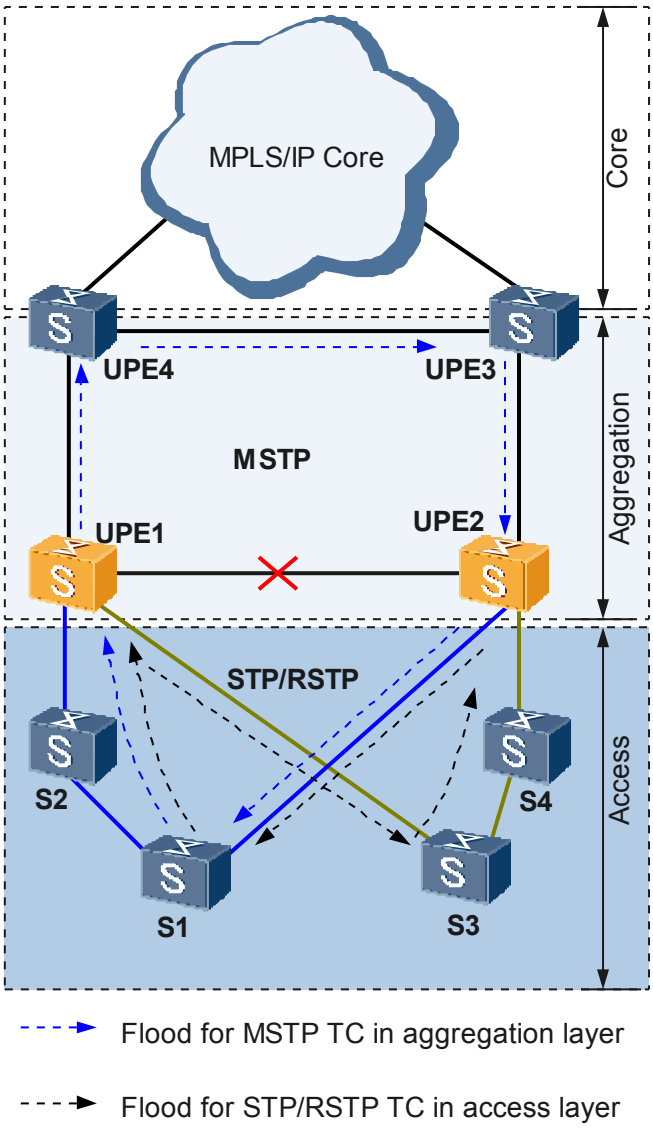
On the network shown in **Figure 8-32**, if the public link between UPE1 and UPE2 fails, multiple switching devices that are connected to the UPEs will unblock their blocked ports.

Figure 8-32 Public link fault



Assume that UPE1 is configured with the highest priority, UPE2 with the second highest priority, and switching devices with default or lower priorities. After the link between UPE1 and UPE2 fails, the blocked ports (replacing the root ports) on switching devices no longer receive packets with higher priorities and re-performs state machine calculation. If the calculation changes the blocked ports to designated ports, a permanent loop occurs, as shown in **Figure 8-33**.

Figure 8-33 Loop between access rings



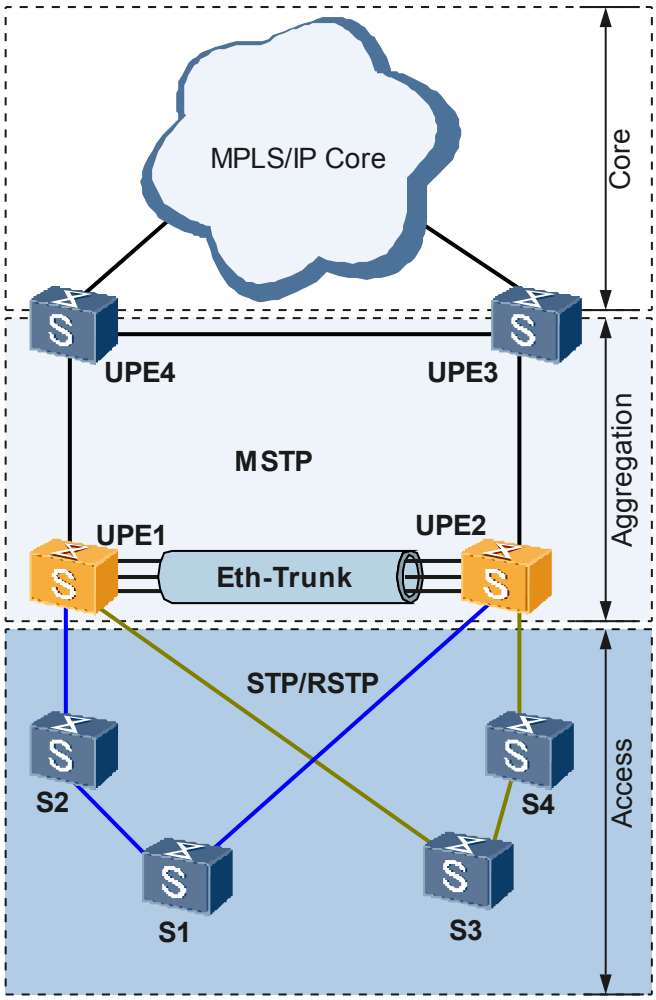
- Solutions

  To prevent a loop between access rings, use either of the following solutions:

  - Configure an inter-board Eth-Trunk link between UPE1 and UPE2.

    An inter-board Eth-Trunk link is used as the public link between UPE1 and UPE2 to improve link reliability, as shown in **Figure 8-34**.

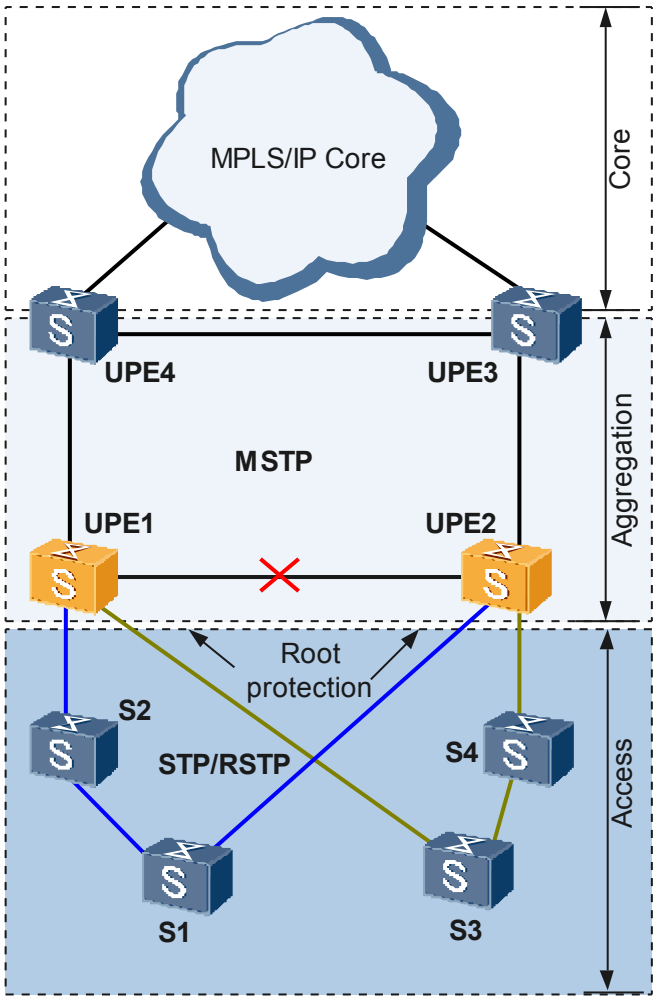
Figure 8-34 Inter-board Eth-Trunk link



- Configure root protection between UPE1 and UPE2.

  If all physical links between UPE1 and UPE2 fail, configuring an inter-board Eth-Trunk link cannot prevent the loop. Root protection can be configured to prevent the loop shown in **Figure 8-33**.

Figure 8-35 MSTP multi-process with root protection



Use the blue ring shown in **Figure 8-35** as an example. UPE1 is configured with the highest priority, UPE2 with the second highest priority, and switching devices on the blue ring with default or lower priorities. In addition, root protection is enabled on UPE2.

Assume that a port on S1 is blocked. When the public link between UPE1 and UPE2 fails, the blocked port on S1 begins to calculate the state machine because it no longer receives BPDUs of higher priorities. After the calculation, the blocked port becomes the designated port and performs P/A negotiation with the downstream device.

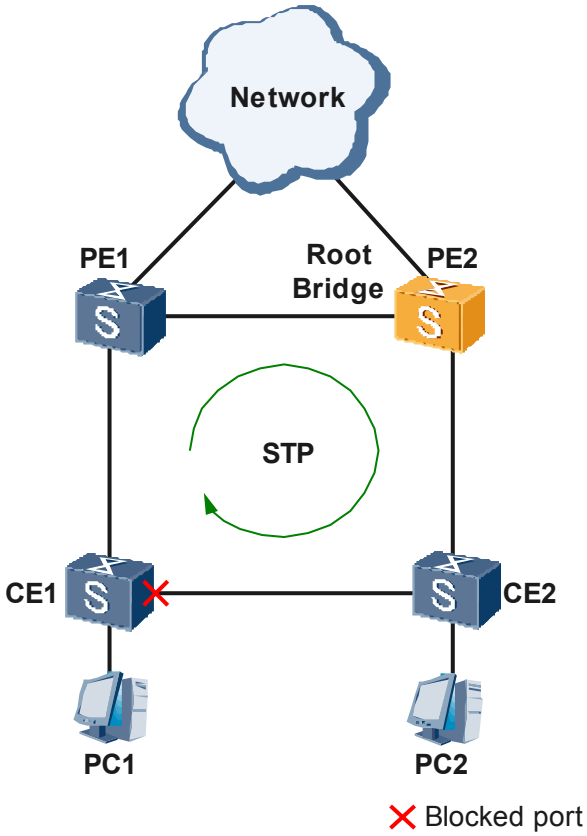
After S1, which is directly connected to UPE2, sends BPDUs of higher priorities to the UPE2 port enabled with root protection, the port is blocked. From then on, the port remains blocked because it continues receiving BPDUs of higher priorities. In this manner, no loop will occur.

8.6 Applications

Application of STP

On a complex network, loops are inevitable. With the requirement for network redundancy backup, network designers tend to deploy multiple physical links between two devices, one of which is the master and the others are the backup. Loops are likely or bound to occur in such a situation.

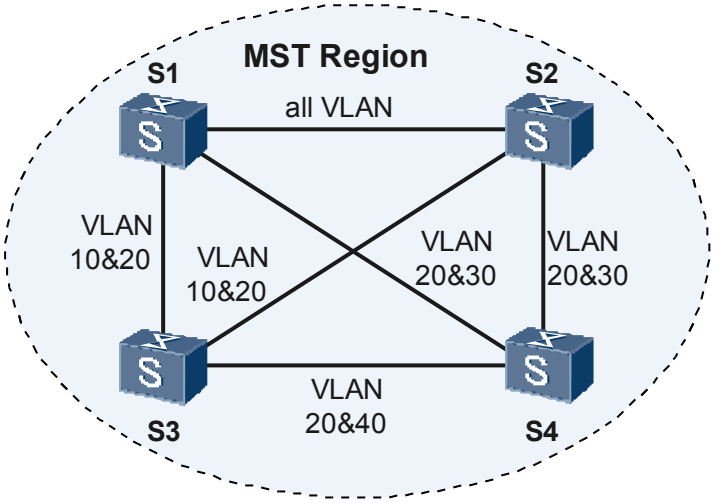
Figure 8-36 Networking diagram for a typical STP application



On the network shown in **Figure 8-36**, after CE and PE running STP discover loops on the network by exchanging information with each other, they trim the ring topology into a loop-free tree topology by blocking a certain port. In this manner, replication and circular propagation of packets are prevented on the network and the switching devices are released from processing duplicated packets, thereby improving their processing performance.

Application of MSTP

Figure 8-37 Networking diagram for a typical MSTP application



MSTP allows packets in different VLANs to be forwarded by using different spanning tree instances, as shown in **Figure 8-37**. The configurations are as follows:

- All switches on the network belong to the same MST region.
- VLAN 10 packets are forwarded within MSTI 1; VLAN 30 packets are forwarded within MSTI 3; VLAN 40 packets are forwarded within MSTI 4; VLAN 20 packets are forwarded within MSTI 0.

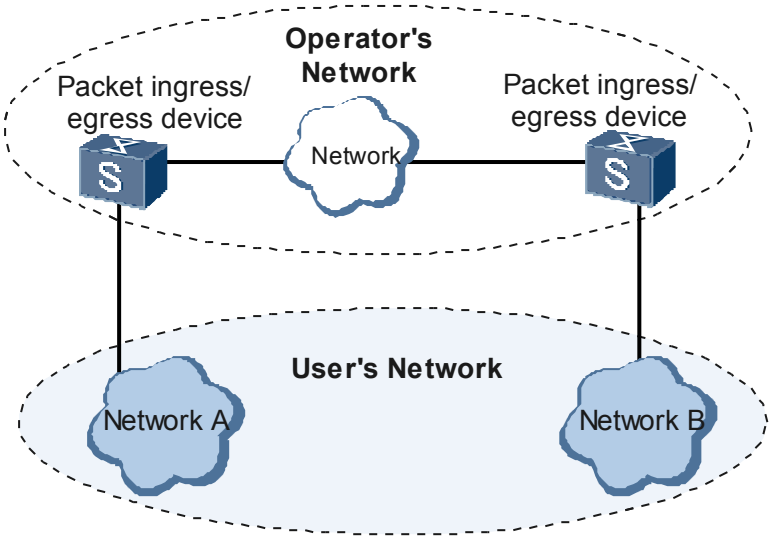
BPDU Tunneling

The BPDU tunneling technology allows a user's networks located in different areas to transparently transmit BPDUs on a specified VLAN VPN within an operator's network. In this manner, all devices on the user's networks can calculate the spanning tree. The user's networks and the operator's networks have their own independent spanning trees.

As shown in **Figure 8-38**, the upper part is an operator's network; the lower part is a user's network. The operator's networks hold ingress/egress devices; the user's networks consist of user's network A and user's network B.

You can configure the packet ingress device to replace the original destination MAC address of a BPDU with a MAC address in a special format and the packet egress device to replace the MAC address in a special format with the original MAC address. In this manner, the BPDU is transparently transmitted.

Figure 8-38 Networking diagram for BPDU transparent transmission

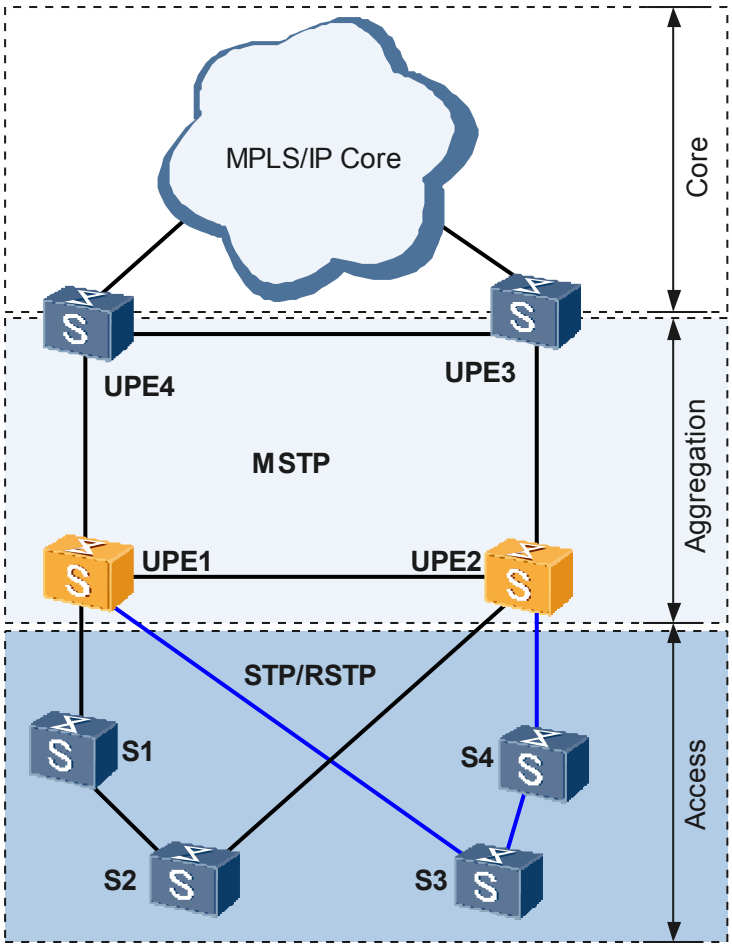


Application of MSTP Multi-process

As shown in **Figure 8-39**, the UPEs are connected to each other through Layer 2 links and enabled with MSTP. The rings connected to the UPEs must be independent of each other. The devices on the rings connected to the UPEs support only RSTP, not MSTP.

After MSTP multi-process is enabled, each MSTP process corresponds to a ring connected to the UPE. The spanning tree protocol on each ring calculates a tree independently.

Figure 8-39 Application with both MSTP and STP/RSTP



8.7 Terms and Abbreviations

Terms

Term	Explanation
STP	Spanning Tree Protocol, a protocol used in a local area network (LAN) to eliminate loops. STP-capable devices exchange protocol packets to discover loops in the network and block redundant interfaces to eliminate loops.
RSTP	Rapid Spanning Tree Protocol, a protocol defined in the IEEE 802.1w. RSTP is a supplement to STP and implements faster convergence than STP.

Term	Explanation
MSTP	Multi-Spanning Tree Protocol, a spanning tree protocol defined in IEEE 802.1s, introducing the concepts of region and instance. MSTP divides a large network into regions and creates multiple spanning tree instances (MSTIs), which are mapped to virtual LANs (VLANs). Network bridges exchange bridge protocol data units (BPDUs) carrying information about regions and instances to know which MSTIs they belong to. Multi-instance RSTP run within regions, whereas RSTP-compatible protocols run between regions.
VLAN	A Virtual Local Area Network (VLAN) is a logical switched network that is constructed across different network segments by using network management software. A VLAN forms a logical subnet (broadcast domain). One VLAN can include multiple network devices.

Abbreviations

Abbreviation	Full Name
STP	Spanning Tree Protocol
RSTP	Rapid Spanning Tree Protocol
MSTP	Multiple Spanning Tree Protocol
BPDU	Bridge Protocol Data Unit
CIST	Common and Internal Spanning Tree
CST	Common Spanning Tree
IST	Internal Spanning Tree
SST	Single Spanning Tree
MST	Multiple Spanning Tree
MSTI	Multiple Spanning Tree Instance
TCN	Topology Change Notification
VLAN	Virtual Local Area Network

9 SEP

About This Chapter

9.1 Introduction

Definition

The Smart Ethernet Protection (SEP) protocol is a link layer protocol dedicated to Ethernet rings. SEP takes an SEP segment as a basic unit. An SEP segment is composed of multiple interconnected Layer 2 switching devices configured with the same SEP segment ID and the same control VLAN ID.

Purpose

Generally, redundant links are used on an Ethernet switching network to provide link backup and enhance network reliability. The use of redundant links, however, may produce loops, causing broadcast storms and rendering the MAC address table unstable. As a result, the communication quality deteriorates, and communication services may even be interrupted. To solve the loop problem, Huawei datacom devices support the following ring network protocols:

- STP/RSTP/MSTP

  STP, RSTP, and MSTP are standard protocols for breaking loops on Ethernet networks. They are mature and widely applied. Huawei devices running one of them can communicate with non-Huawei devices. The convergence time is at the second level on a network running STP, RSTP, or MSTP, which cannot meet the requirements of some real-time services. The convergence time is affected by the network topology.

- RRPP

  RRPP is a propriety protocol of Huawei. It features short convergence time (less than 50 ms) and supports load balancing for different types of traffic. A Huawei device running RRPP cannot communicate with any non-Huawei device. RRPP has a high requirement on network topologies. Logical topologies need to be configured for a physical topology, and primary rings and sub-rings need to be defined for these logical topologies. Therefore, RRPP is not applicable to complex networks.

Therefore, Huawei develops SEP. Like RRPP, SEP boasts short convergence time (less than 50 ms). Compared with RRPP, SEP has the following advantages:

- SEP supports various types of networking modes. For example, a network running SEP can communicate with a network running STP, RSTP, MSTP, or RRPP. SEP supports all topologies and the display of network topologies.

  The blocked interface, therefore, can be quickly located. When a fault occurs, SEP can quickly locate the fault, improving network maintainability.

- SEP supports various policies for specifying an interface to block. This allows the implementation of traffic load balancing.
- Link switchback is not performed after fault recovery, improving network stability.

9.2 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

9.3 Principles

9.3.1 Principles of SEP

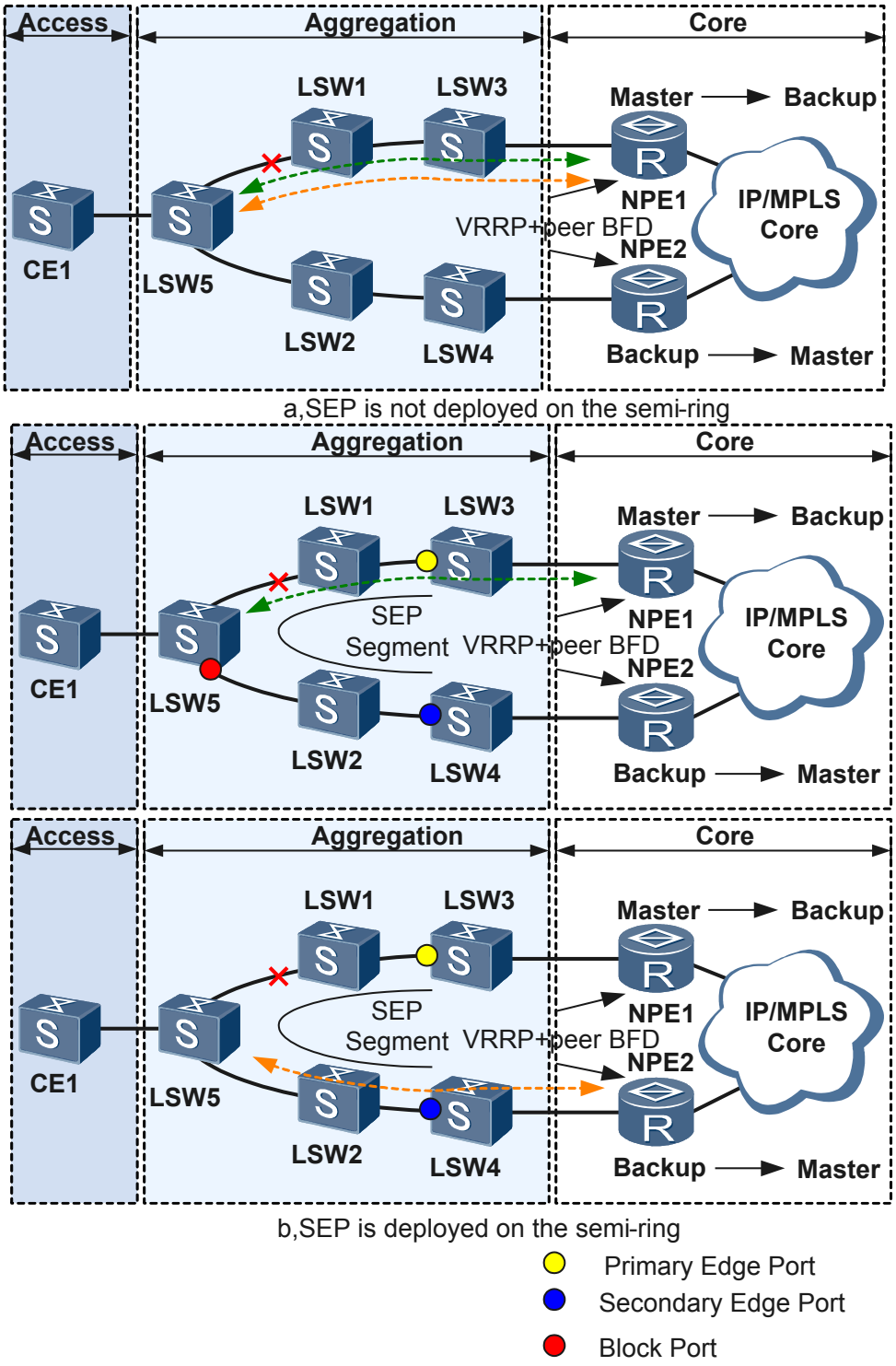
The SEP protocol is a dedicated link layer protocol for use on Ethernet ring networks. A SEP segment is the basic unit of the protocol. Only two interfaces on a Layer 2 switching device can be added to the same SEP segment.

In a SEP segment, loops can be prevented by starting a protection mechanism to selectively block certain interfaces and eliminate Ethernet redundant links. When a fault occurs on a ring network, a device running SEP can quickly unblock the blocked interface to perform link switching. This maintains normal communication between nodes on the ring network.

Figure 9-1 shows a typical SEP application. CE1 is connected to NPEs through a semi-ring formed by switches. A VRRP backup group is deployed on the NPEs. Initially, the status of NPE1 is master and the status of NPE2 is backup. When the link between NPE1 and LSW5 or a node on the link becomes faulty (it is assumed that the link between LSW1 and LSW5 becomes faulty), the status of NPE1 changes from master to backup and the status of NPE2 changes from backup to master, and the following situations occur:

- If SEP is not deployed on the semi-ring, CE1 still forwards traffic along the original path. NPE1 that becomes backup does not forward traffic, causing traffic interruption.
- If SEP is deployed on the semi-ring, the blocked interface on LSW5 becomes unblocked and enters the forwarding state. In addition, it sends Link Status Advertisements (LSAs) to instruct other nodes on the SEP segment to refresh their LSA databases. CE1 sends traffic along the backup link LSW5->LSW2->LSW4->NPE2. This ensures proper traffic transmission.

Figure 9-1 Schematic diagram for SEP



In ordinary SEP networking, a physical ring can be configured with only one SEP segment in which only one interface can be blocked. If an interface in the SEP segment in the complete state is blocked, all user package is transmitted only along the path where the primary edge interface resides. The path where the secondary edge interface resides is idle, wasting bandwidths.

SEP multi-instance developed by Huawei is a solution to limit bandwidth waste and to implement traffic load balancing. SEP multi-instance allows two SEP segments to be configured on a

physical ring network. All devices, interface roles, and control VLANs in each SEP segment must be configured by conforming to basic SEP configuration principles. Each of the two SEP segments has a blocked interface. The blocked interfaces detect whether the physical ring network is complete. The blocked interfaces in the two SEP segments are independent of each other.

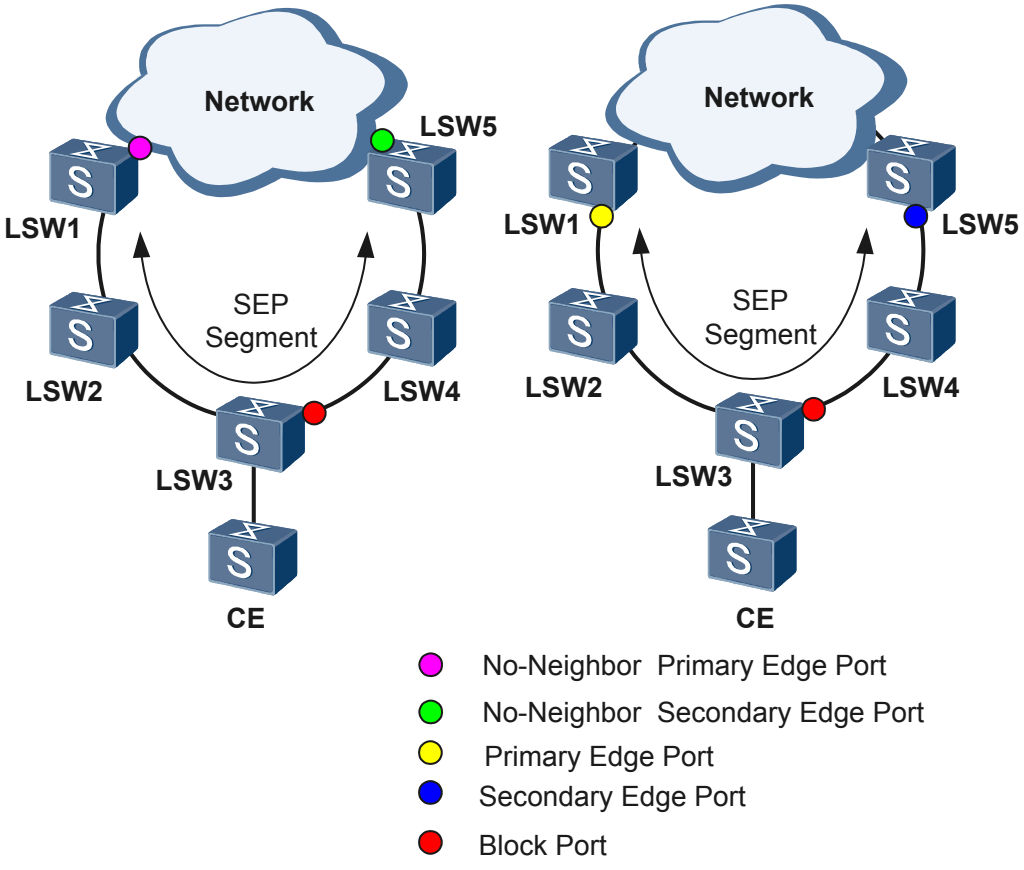
For details about SEP multi-instance, see **9.3.3 SEP Implementation Mechanisms**.

9.3.2 Basic Concepts of SEP

Network Architecture of SEP

As shown in **Figure 9-2**, LSW1, LSW2, LSW3, LSW4, and LSW5 are connected to access a Layer 2 network. On the Layer 2 network, two edge devices LSW1 and LSW5 are indirectly connected. This networking is called open-ring networking. Such networking mode results in a new loop on the entire network. To eliminate redundant loops on the network and ensure the connectivity of links, a protection mechanism is required. **Figure 9-2** shows the typical networking of an open ring running SEP. The following describes the basic concepts of SEP.

Figure 9-2 Networking diagram of an open ring running SEP



- SEP segment

  SEP takes an SEP segment as a basic unit. An SEP segment is composed of multiple interconnected Layer 2 switching devices that are configured with the same SEP segment ID and control VLAN ID.

An SEP segment physically corresponds to a ring-shaped or line-shaped Ethernet topology. Each SEP segment has a control VLAN, edge interfaces, and common interfaces.

- Control VLAN

  In an SEP segment, the control VLAN is used to transmit only SEP packets.

  Each SEP segment must be configured with a control VLAN. After an interface is added to an SEP segment that is configured with a control VLAN, the interface is added to the control VLAN automatically.

  Different SEP segments can use the same control VLAN ID.

  Different from a control VLAN, a data VLAN is used to transmit data packets.

- Node

  A node refers to a Layer 2 switching device that is added to an SEP segment. A maximum of two interfaces on a node can be added to the same SEP segment.

- Interface role

  As defined by SEP, there are two interface roles: common interfaces and edge interfaces.

  As shown in **Table 9-1**, edge interfaces are further classified into primary edge interfaces, secondary edge interfaces, no-neighbor primary edge interfaces, and no-neighbor secondary edge interfaces.

  NOTE

  Normally, edge interfaces and no-neighbor edge interfaces do not reside in the same SEP segment.

  The interfaces connected to a primary edge interface and a secondary edge interface must be SEP interfaces. The interfaces connected to a no-neighbor primary edge interface and a no-neighbor secondary edge interface can not be SEP interfaces.

Table 9-1 Interface roles

Interface Role	Sub-role	Description
Common interface	-	In an SEP segment, all interfaces except edge interfaces are common interfaces. A common interface monitors the status of its directly connected SEP link and sends a message about link status changes to a neighboring interface in time. The neighboring interface constantly floods the message to other interfaces in the SEP segment until the message reaches the primary edge interface. The primary edge interface then processes the message.
Edge interface	Primary edge interface	There is only one primary edge interface in an SEP segment. The primary edge interface is elected after being configured. The primary edge interface initiates blocked-interface preemption, terminates packets, and sends messages about topology changes to other networks.

Interface Role	Sub-role	Description
	Secondary edge interface	There is only one secondary edge interface in an SEP segment. The secondary edge interface is elected after being configured. The secondary edge interface terminates packets and sends messages about topology changes to other networks.
	No-neighbor primary edge interface	The interface at the edge of the SEP segment is a no-neighbor edge interface. The no-neighbor edge interface is configured by a user. The no-neighbor primary edge interface terminates packets and sends messages about topology changes to other networks. No-neighbor primary edge interfaces are used to interconnect Huawei devices and non-Huawei devices or interconnect Huawei devices and devices that do not support SEP. **NOTE** Whether the device where a no-neighbor edge interface resides sends preemption packets is determined by whether the brother interface of the no-neighbor edge interface is blocked. ● If the brother interface of the no-neighbor edge interface is blocked, the device does not need to send preemption packets. ● If the brother interface of the no-neighbor edge interface is unblocked, the brother interface sends preemption packets.
	No-neighbor secondary edge interface	There is only one no-neighbor secondary edge interface in an SEP segment. The no-neighbor secondary edge interface is elected after being configured. The no-neighbor secondary edge interface terminates packets and sends messages about topology changes to other networks. No-neighbor secondary edge interfaces are used to interconnect Huawei devices and non-Huawei devices or interconnect Huawei devices and devices that do not support SEP.

- Blocked interface

  A blocked interface refers to the interface that is blocked to prevent loops in an SEP segment.

In an SEP segment, no interface is specified as a blocked interface. Each interface in the SEP segment may become as a blocked interface. When an SEP segment works normally, there is only one blocked interface in the SEP segment.

- Status of interfaces enabled with SEP

  In an SEP segment, the status of interfaces enabled with SEP is classified into two types, as shown in **Table 9-2**.

  Table 9-2 Interface status

Interface Status	Description
Forwarding	In the Forwarding state, an interface can forward user traffic, and receive and send SEP packets.
Discarding	In the Discarding state, an interface just receives and sends SEP packets.

Interface status and interface roles are not necessarily related. Interfaces playing different roles support Forwarding and Discarding states.

SEP Packet

Table 9-3 shows the types of SEP packets.

Table 9-3 Types of SEP packets

Packet Type	Packet Subtype	Description
Hello packet	-	After an interface is added to an SEP segment, the neighbor negotiation mechanism is started on the interface. By exchanging Hello packets, the interface and its neighboring interface establish the neighbor relationship. After neighbor negotiation succeeds, the interfaces continue to exchange Hello packets to detect the neighbor status.
LSA	LSA Request packet	After an interface is enabled with SEP, the interface periodically sends Link Status Advertisements (LSA) to its neighboring interface. After the state machine of the neighboring interface is Up, the two interfaces update their link status databases, that is, all topology information.
	LSA ACK packet	
TC packet	-	When the topology of an SEP segment changes, a Topology Change (TC) packet is sent to notify the upper-layer network. Then, all nodes on the upper-layer network need to update their MAC address forwarding tables and ARP tables. The TC packet is sent by the device where the SEP segment and the upper-layer network are intersected.

Packet Type	Packet Subtype	Description
GR packet	-	If a device sends an SEP Graceful Restart (GR) packet, it indicates that active/standby switchover occurs on the device. A GR packet is sent by a device to instruct other nodes to prolong the aging time of the LSA received from the device. After active/standby switchover is complete, the device needs to send another GR packet to instruct other nodes to restore the aging time of the LSA received from the device to the previous value.
Primary edge interface-election packet	-	After an interface is enabled with SEP, the interface sets its interface role as the primary edge interface if it has the right to participate in the election of the primary edge interface. Then, the interface periodically sends primary edge interface-election packets without waiting for the success of neighbor negotiation. A primary edge interface-election packet contains the interface role (primary edge interface, secondary edge interface, or common interface), bridge MAC address of the interface, interface ID, and status of the topology database.
Preemption packet	Preemption Request packet	A preemption packet is used to block a specified interface. Preemption packets are sent by the selected primary edge interface or the brother interface of a no-neighbor primary edge interface.
	Preemption ACK packet	

9.3.3 SEP Implementation Mechanisms

Neighbor Negotiation Mechanism

After an interface is added to a SEP segment, the interface begins neighbor negotiations. The newly added interface and its neighboring interfaces establish neighbor relationships by exchanging Hello packets. After neighbor negotiations succeed, the interfaces continue to exchange Hello packets to detect the neighbor status.

The neighbor negotiation mechanism can prevent unidirectional links. The neighbor negotiation mechanism is bidirectional. The interfaces at both ends of a link need to send Hello packets to each other. If one of the two interfaces does not receive a Hello packet from the other interface before a timeout occurs, this interface sets the neighbor status to Down.

The neighbor negotiation mechanism provides information required to display the topology of a SEP segment. After the neighbor negotiation mechanism is used to establish neighbor relationships between interfaces, links can be connected to form a complete SEP segment. This helps to display the complete topology of the SEP segment.

Synchronization of SEP Link Status Databases and Display of the Topology

- Synchronization of SEP link status databases

  After neighbor negotiations are complete, nodes in the SEP segment enter the phase of link status database synchronization. Each node periodically sends LSAs. After receiving LSAs from other nodes, a node updates its neighbor status database. This ensures that the link status databases of all nodes in the SEP segment are consistent.

  If a device does not receive any LSAs from its peer device or other devices in the SEP segment within three LSA transmission intervals, the device will age the database that save the other devices LSAs in the SEP segment.

  When a faulty node in a SEP segment recovers, the node needs to obtain topology information from the other nodes in the SEP segment. After receiving LSA request packets from the node, neighboring interfaces reply with LSA ACK packets containing the latest link status information.

- Display of SEP segment topologies

  The function for displaying SEP segment topologies allows you to view the topology of normal SEP segments on each device. Link status synchronization ensures that topology information on every device in a SEP segment is consistent.

  Table 9-4 shows the types of SEP segment topologies.

Table 9-4 Types of SEP segment topologies

Topology Type	Description	Constraint
Ring-shaped topology	The neighbor status of each interface in a SEP segment is Up, and each interface has a neighbor interface and a brother interface. That is, only two interfaces on a node segment can be added to the same SEP segment.	● If the primary edge interface is elected on a ring network, topology information is displayed beginning from the primary edge interface. ● If there is no primary edge interface on a ring network and the secondary edge interface is elected, topology information is displayed beginning from the secondary edge interface.

Topology Type	Description	Constraint
Line-shaped topology	All topologies except ring-shaped topologies are line-shaped topologies.	For interfaces at both ends of a link: ● If one interface functions as the primary edge interface, topology information is displayed beginning from the primary edge interface. ● If there is no primary edge interface but there is a secondary edge interface, topology information is displayed beginning from the secondary edge interface.

NOTE

The conditions detailed in **Table 9-4** ensure that topology information displayed on every node in a ring-shaped or a line-shaped topology is consistent.

Election of a Primary Edge Interface

Only interfaces that are configured as no-neighbor edge interfaces, primary edge interfaces and secondary edge interfaces have rights to participate in the election of a primary edge interface.

NOTE

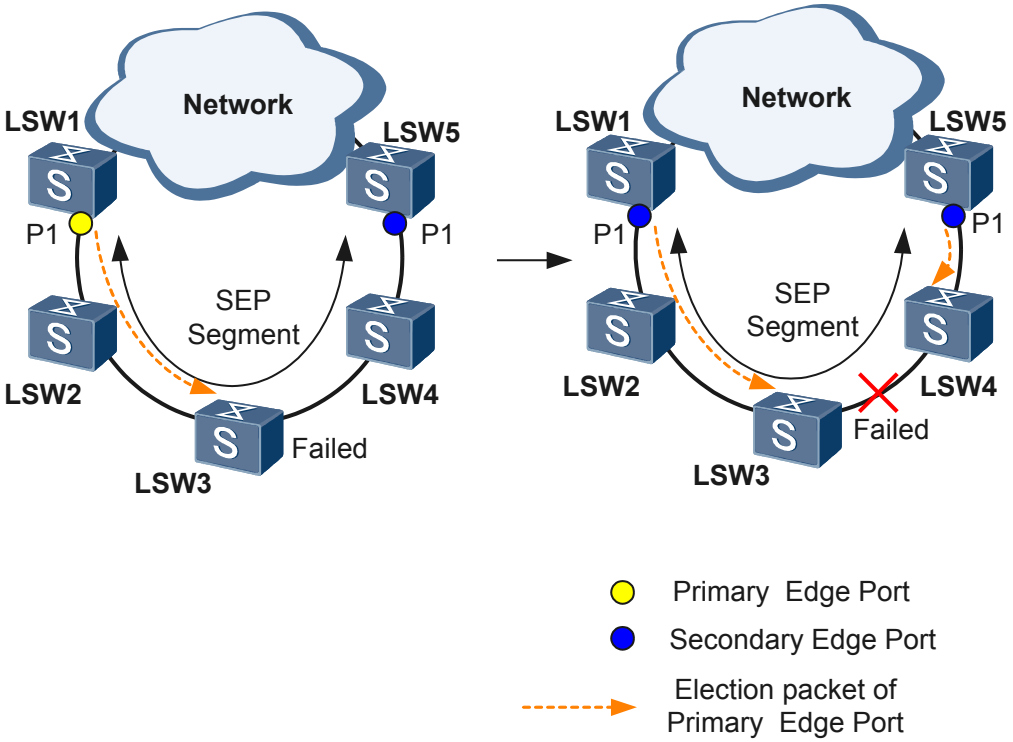
If only one interface on a node is enabled with SEP, the corresponding command must be run to set the role of the interface to **edge** so that it can function as an edge interface.

As shown in **Figure 9-3**, if there is no faulty link on the network and SEP is enabled on interfaces, the following situations occur:

- Common interfaces do not participate in the election of the primary edge interface. Only P1 on LSW1 and P1 on LSW5 participate in the election of the primary edge interface.
- If P1 on LSW1 and P1 on LSW5 have the same role, the interface that has a higher MAC address is elected as the primary edge interface.

After the primary edge interface is selected, it begins sending primary edge interface-election packets periodically without waiting for the success of neighbor negotiations. A primary edge interface-election packet contains information about the interface role (primary edge interface, secondary edge interface, or common interface), bridge MAC address of the interface, interface ID, and status of the topology database.

Figure 9-3 Networking diagram for electing a primary edge interface



As shown in **Figure 9-3**, if a link fault occurs in the SEP segment, P1 on LSW1 and P1 on LSW5 receive fault notification packets, or P1 on LSW5 does not receive any primary edge interface-election packets before a timeout occurs, P1 on LSW1 becomes the secondary edge interface. Therefore, two secondary edge interfaces exist on the SEP segment and send edge interface-election packets periodically.

After the last link fault in the SEP segment is rectified, both secondary edge interfaces can receive edge interface-election packets from each other and a new primary edge interface is elected.

Specifying an Interface to Block

In general, a blocked interface is one of the last two interfaces that complete neighbor negotiation. In some cases, however, the negotiated blocked interface may not be the one a user expects to be blocked. A user can specify an interface to block as needed. The designated blocking does not, however, become effective immediately. A preemption mechanism allows an interface designated by a user to be blocked instead of a previously blocked interface.

- Interface blocking mode

  A user can configure an interface blocking mode in order to specify the location of a blocked interface. **Table 9-5** lists interface blocking modes.

Table 9-5 Interface blocking mode

Interface Blocking Mode	Description
Specifying the interface with the highest priority as the blocked interface	The rules for comparing the priorities of interfaces are as follows: 1. The interface with the highest priority is designated as the blocked interface. The priorities of interfaces can be set. The greater the priority value, the higher the priority. 2. If the interfaces have the same priority, their bridge MAC addresses are compared. The interface with the lowest bridge MAC address is more likely to be designated as a blocked interface. 3. If the interfaces have both the same priority and bridge MAC address, their interface numbers are compared. The interface with the smallest interface number is more likely to be designated as a blocked interface.
Specifying the interface in the middle of a SEP segment as the blocked interface	-
Specifying a blocked interface based on the hop count set by users	The hop count of the primary edge interface is 1, and the hop count from the primary edge interface to the neighboring interface is 2. Continuing to move downstream, each downstream neighbor of the primary edge interface increases the hop count by one.
Specifying a blocked interface based on the device name and interface name	After SEP is configured, a device name and an interface name are used to designate the interface to be blocked. Before specifying an interface to block, run a **display** command to obtain information about all interfaces, and then specify the device name and interface name. If there are several devices with the same device name and interface name on a ring network, preemption packets search from the device where the primary edge interface resides and then block the first searched interface with the specified device name and interface name. **NOTE** If the device name and interface name are used to specify an interface to block, changing the device name or interface name will render the preemption mechanism ineffective.

- Preemption

  After the interface blocking mode is specified, whether the specified interface will be blocked is determined by the preemption mode. **Table 9-6** lists the preemption modes.

Table 9-6 Preemption mode

Preemption Mode	Description
Non-preemption mode	When the last faulty link recovers or last two interfaces complete neighbor negotiation, the interface to be blocked is determined by exchanging packets containing the blocked status of interfaces. The other interfaces enter the forwarding state.
Preemption Mode **NOTE** Preemption can be implemented only on the device where the primary edge interface resides or the no-neighbor primary edge interface resides.	The preemption mode is classified into delayed preemption and manual preemption. ● Delayed preemption When the last faulty edge interface recovers, the edge interface no longer receives fault advertisement packets. If the primary edge interface receives no fault advertisement packet within 3 seconds, it starts a delay timer. After the delay timer expires, nodes in the SEP segment preempt blocked interfaces. ● Manual preemption When the manual preemption mode is configured by using commands and the link status databases of the primary edge interface and the secondary edge interface are complete, the elected primary edge interface or the brother interface of no-neighbor primary edge interface will send preemption packets to block a specified interface. The interface sends a packet to advertise its status immediately after being blocked. The interface that is blocked before preemption enters the forwarding state. Then, manual preemption is complete. **NOTE** Only two interfaces on a device can be added to the same SEP segment. If one interface is the no-neighbor primary edge interface, the other interface is the brother interface of the no-neighbor primary edge interface.

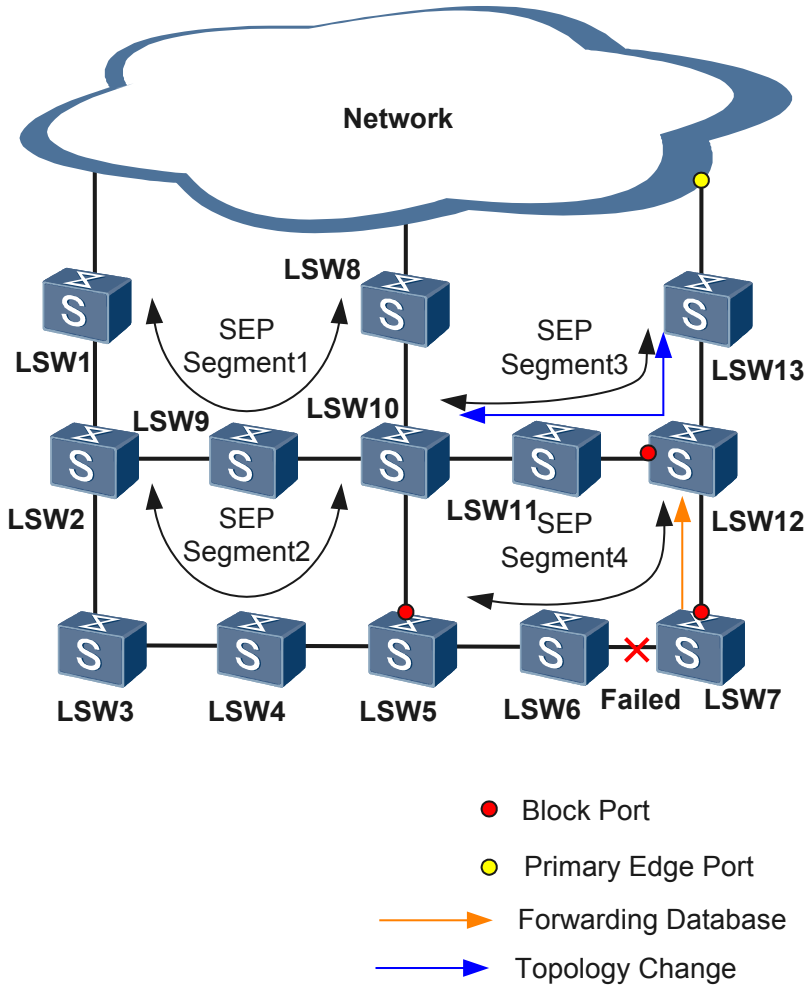
SEP Topology Change Notification

Table 9-7 lists the situations in which the topology of a SEP segment changes.

Table 9-7 SEP topology changes

SEP Topology Change	Description
Topology change caused by an interface fault	Figure 9-4 shows an interface fault in a SEP segment. An interface fault can be a link fault or a neighboring interface fault. If a device that has an interface in the forwarding state in the SEP segment receives a fault advertisement packet, the device needs to send a Flush-Forwarding Database (Flush-FDB) packet through the interface to notify other nodes in the SEP segment that the topology has changed.
Topology change caused by a fault being rectified and the preemption function taking effect	One or more faults occur in the SEP segment. When the last faulty interface recovers and the blocked interface is preempted, the topology is considered changed. Preemption is triggered by the primary edge interface. When an interface in a SEP segment receives a preemption packet from the primary edge interface, the interface needs to send Flush-FDB packets to notify other nodes in the SEP segment that the topology has changed.

Figure 9-4 Networking diagram for SEP topology change notification



NOTE

The topology change notification function is configured on devices that connect an upper-layer network and a lower-layer network. If the topology of either of the networks changes, these devices inform the other network of the change.

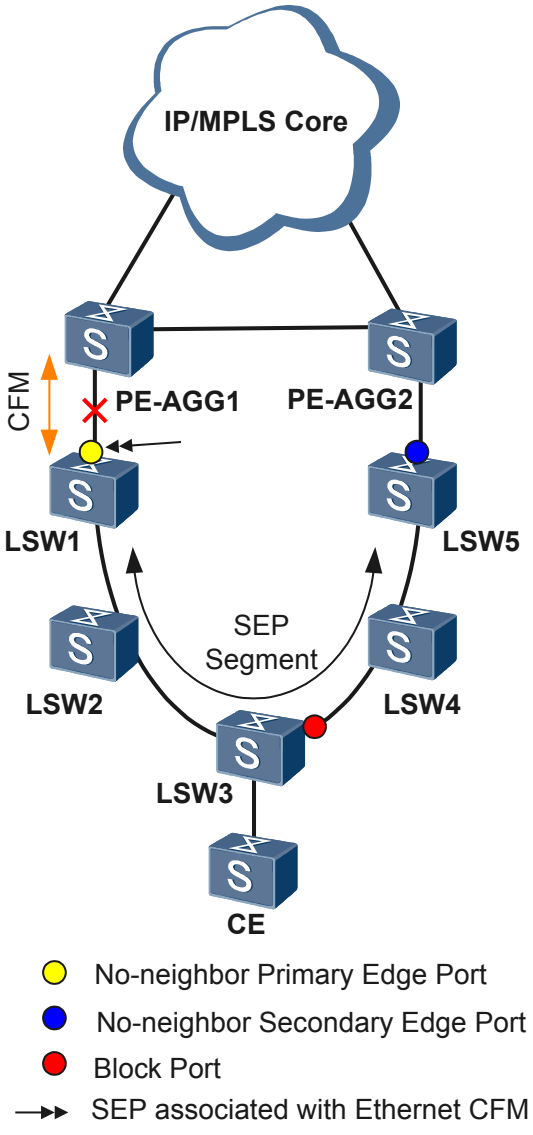
Table 9-8 lists the scenarios in which topology changes are reported.

Table 9-8 SEP topology change notification

SEP Topology Change Notification	Scenario	Description	Solution
Topology change notification from a lower-layer network to an upper-layer network	Networking where a SEP network is connected to an upper-layer network running other features such as SEP, STP, RRPP	● If the blocked interface in a lower-layer SEP network is manually changed, the topology of the SEP segment changes. Because the upper-layer network cannot detect this topology change, traffic is interrupted. ● If an interface in a lower-layer SEP network becomes faulty, the topology of the SEP segment changes but the upper-layer network cannot detect the change. As a result, traffic is interrupted.	Configure the SEP topology change notification function.
	Networking scenario where a host is connected to a SEP network by using a SmartLink group	During an active/standby switchover of member interfaces in the SmartLink group, the host sends a SmartLink Flush packet to notify the connected devices in the SEP segment of the switchover. If the connected devices in the SEP segment cannot identify the SmartLink Flush packet (that is, if these connected devices in the SEP segment cannot detect any topology change of the lower-layer network), traffic will be interrupted.	Enable the edge devices in the SEP segment to process SmartLink Flush packets.

SEP Topology Change Notification	Scenario	Description	Solution
Topology change notification from an upper-layer network to a lower-layer network	Networking scenario where a SEP network is connected to an upper-layer network configured with CFM.	If a fault occurs on the upper-layer network, the topology of that network changes but the lower-layer network cannot detect the change. As a result, traffic is interrupted.	Configure association between SEP and CFM As shown in **Figure 9-5**, association between SEP and CFM is configured on LSW1.

Figure 9-5 Networking diagram of association between SEP and CFM



As shown in **Figure 9-5**, association between SEP and CFM is configured on LSW1 in the SEP segment. When CFM detects a fault on the network at the convergence layer, LSW1 uses a CCM to notify the Operation, Administration, and Maintenance (OAM) module of the fault. Then, the SEP status of the interface associated with CFM goes Down.

The interface associated with CFM is in the SEP segment. If this interface goes Down, LSW2 needs to send a Flush-FDB packet to notify the other nodes in the SEP segment that the topology has changed. After LSW3 receives the Flush-FDB packet, the blocked interface on LSW3 is unblocked and enters the Forwarding state. This interface then sends a Flush-FDB packet to instruct the other nodes in the SEP segment to refresh their MAC address forwarding tables and ARP tables. The lower-layer network can detect the fault of the upper-layer network, and therefore reliable service transmission is guaranteed.

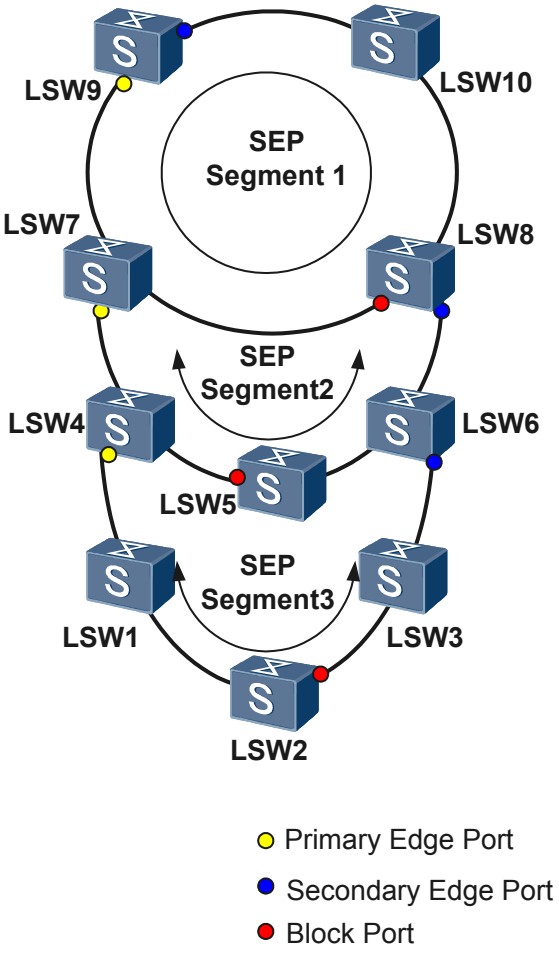
Suppression of SEP TC Notification Packets

Topology changes of a SEP segment are advertised to other SEP segments or upper-layer networks. A large number of topology change (TC) notification packets are generated in the following cases:

- A link becomes disconnected transiently.
- A SEP segment is attacked by invalid TC notification packets.
- A networking scenario has multiple SEP ring networks.

  Figure 9-6 shows a networking scenario with three SEP ring networks. If the topology of SEP segment 3 changes, the number of TC notification packets doubles and SEP segment 2 is flooded with these packets. Each time TC notification packets pass through a SEP segment, the number of these TC notification packets doubles.

Figure 9-6 Networking diagram for multiple SEP ring networks



The sending of many TC notification packets reduces the CPU's capability in processing other types of packets. In addition, devices in the SEP segments are caused to frequently refresh MAC address entries and this consumes bandwidth resources. To solve such problems, the following measures can be taken to suppress TC notification packets:

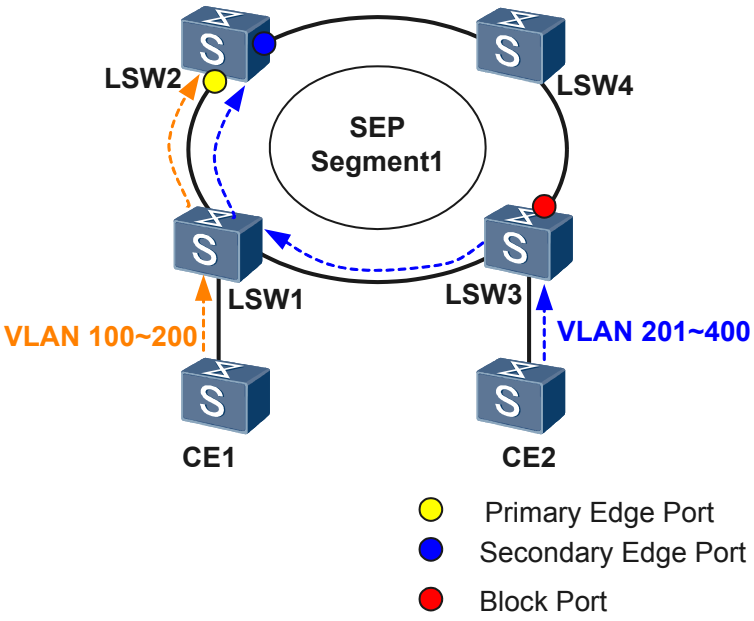
- Suppress TC notification packets based on their source addresses by configuring a device to process only one of the TC notification packets carrying the same source address.
- Configure a device to process a specified number of TC notification packets within a specified time period. By default, three TC notification packets with different source addresses are processed in 2s.
- Avoid the networking scenario that has more than three SEP ring networks.

SEP Multi-Instance

In regular SEP networking shown in **Figure 9-7**, a physical ring network can be configured with only one SEP segment in which only one interface can be blocked.

If an interface in the SEP segment in the complete state is blocked, all user package is transmitted only along the path where the primary edge interface resides. The path where the secondary edge interface resides is idle, wasting bandwidths.

Figure 9-7 Networking diagram for SEP

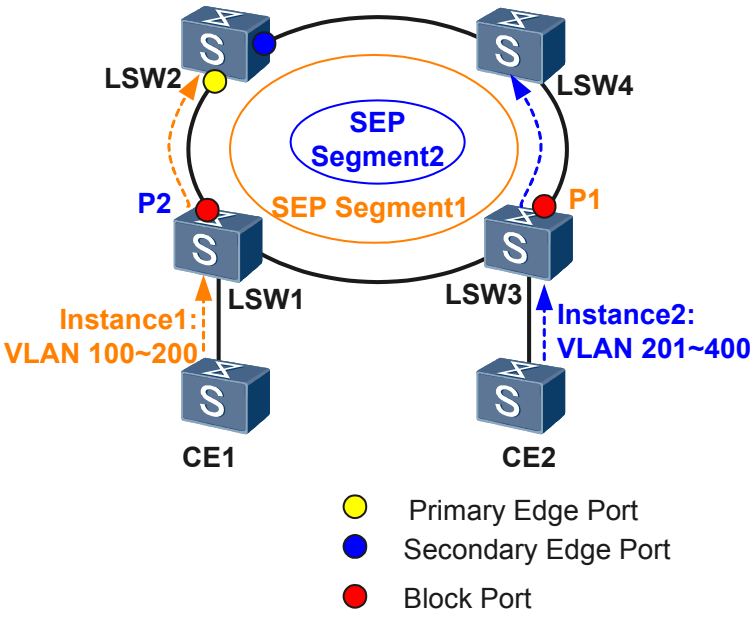


SEP multi-instance allows two SEP segments to be configured on one physical ring network. All devices, interface roles, and control VLANs in each SEP segment must be configured by conforming to basic SEP configurations principles. Each SEP segment has one blocked interface. Each blocked interface detects whether the physical ring network is complete. The blocked interfaces in the two SEP segments are independent of each other.

Each SEP segment needs to be configured with a protected instance and each protected instance represents a VLAN range. The topology calculated by a SEP segment is valid only for that SEP segment.

After different protected instances are configured for SEP segments and the mapping between protected instances and VLANs is set, a blocked interface is valid only for the VLANs protected by the SEP segment where the blocked interface resides. Data traffic of different VLANs can be transmitted along different paths. This implements traffic load balancing and link backup.

Figure 9-8 Networking diagram for SEP multi-instance



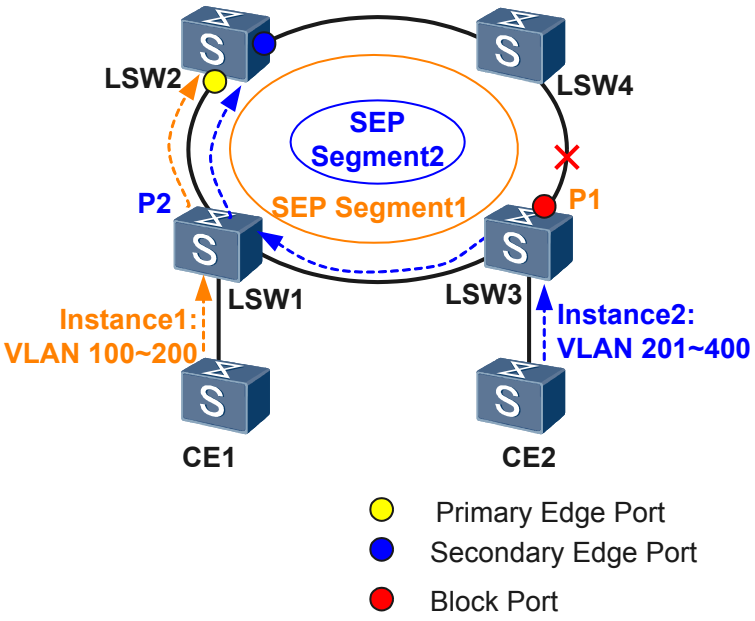
As shown in **Figure 9-8**, the SEP multi-instance ring network that consists of LSW1 to LSW4 has two SEP segments. P1 is the blocked interface in SEP segment 1, and P2 is the blocked interface in SEP segment 2.

- Protected instance 1 is configured in SEP segment 1 to protect the data of VLAN 100 to VLAN 200. The data is transmitted along path LSW1->LSW2. As the blocked interface in SEP segment 2, P2 blocks only the data of VLAN 201 to VLAN 400.
- Protected instance 2 is configured in SEP segment 2 to protect the data of VLAN 201 to VLAN 400. The data is transmitted along path LSW3->LSW4. As the blocked interface in SEP segment 1, P1 blocks only the data of VLAN 100 to VLAN 200.

In the case of a node or a link failure, each SEP segment calculates its own topology independently, and the nodes in each SEP segment update their LSA databases.

As shown in **Figure 9-9**, a fault occurs on the link between LSW3 and LSW4. The link fault does not affect the transmission path for the data of VLAN 100 to VLAN 200 in SEP segment 1, but blocks the transmission path for the data of VLAN 201 to VLAN 400 in SEP segment 2.

Figure 9-9 Networking diagram for a fault in a link on a SEP muti-instance network



After the link between LSW3 and LSW4 becomes faulty, LSW3 starts to send LSAs to instruct the other devices in SEP segment 2 to refresh their LSA databases, and the blocked interface enters the forwarding state. After the topology of SEP segment 2 is recalculated, the data of VLAN 201 to VLAN 400 is transmitted along path LSW3->LSW1->LSW2.

After the link between LSW3 and LSW4 recovers, the devices in SEP segment 2 performs delayed preemption. After the preemption delay expires, P1 becomes the blocked interface again, and sends LSAs to instruct the other devices in SEP segment 2 to refresh their LSA databases. After the topology of SEP segment 2 is recalculated, the data of VLAN 201 to VLAN 400 is transmitted along path LSW3->LSW4.

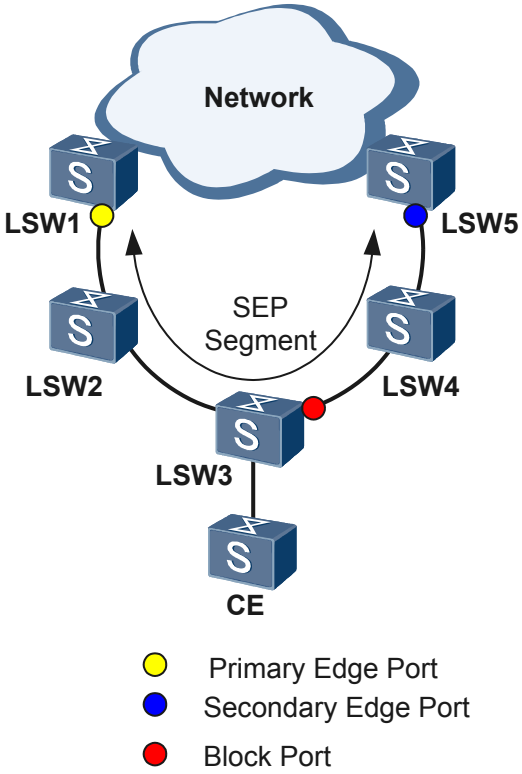
9.4 Applications

9.4.1 Open-Ring Networking

As shown in **Figure 9-10**, LSW1 to LSW5 are connected to form an open ring to access a Layer 2 network. The two edge devices on the Layer 2 network, that is, LSW1 and LSW5, are not directly connected. This networking is called open-ring networking. The open-ring networking is at the access layer and is used to transparently transmit Layer 2 unicast and multicast services. After SEP is run at the access layer, redundancy protection switching can be implemented at the access layer and topology of the SEP segment can be displayed.

On a closed ring network, an edge interface is deployed on each of the two edge devices.

Figure 9-10 Networking diagram of an open ring running SEP

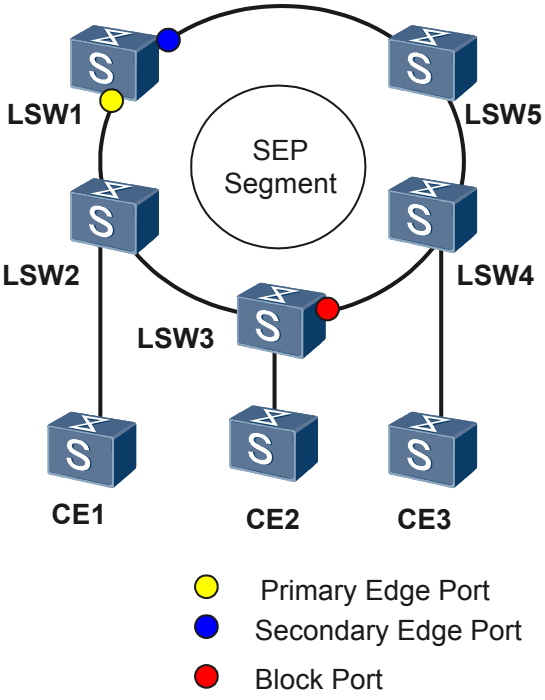


9.4.2 Closed-ring Networking

As shown in **Figure 9-11**, LSW1 to LSW5 are connected to form a dual-homed link to access a Layer 2 network. LSW1 and LSW5 at the edge of the Layer 2 network are directly connected. This networking is called closed-ring networking. The networking is at the convergence layer and is used to converge Layer 2 unicast and multicast services. After SEP is run at the convergence layer, redundancy protection switching can be implemented at the convergence layer and the topology of the SEP segment can be displayed.

On a closed ring network, two edge interfaces are deployed on one edge device.

Figure 9-11 Networking diagram of a closed ring running SEP



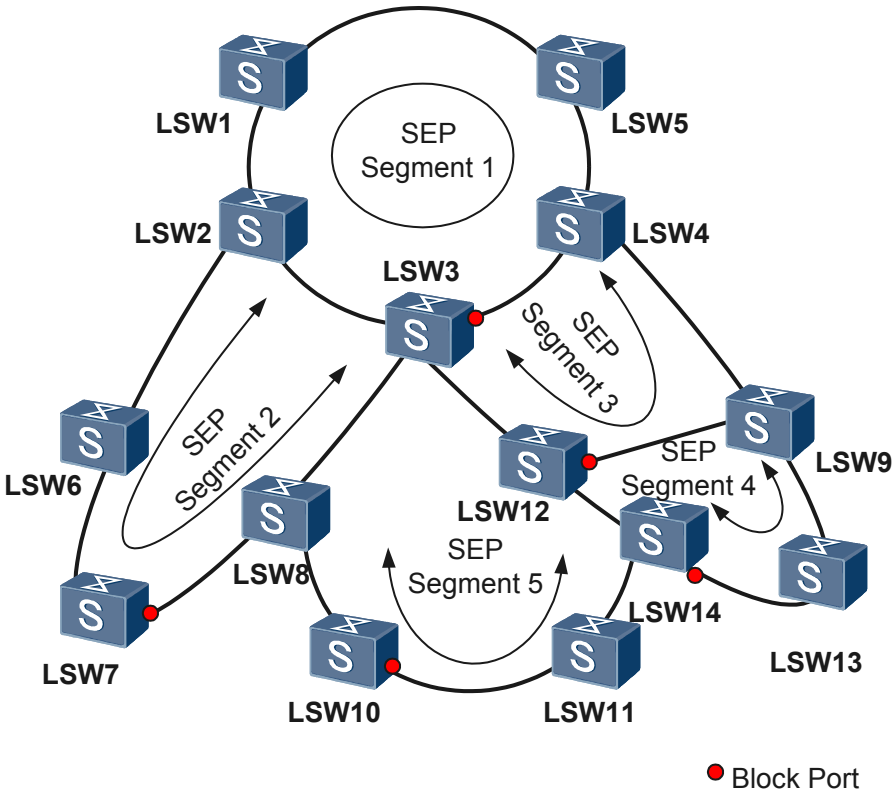
9.4.3 Multiple-Ring Networking

As shown in **Figure 9-12**, the networking composed of LSW1 to LSW14 is called multiple-ring networking. LSW1 to LSW5 are at the convergence layer, and LSW6 to LSW14 are at the access layer. Layer 2 services are transparently transmitted at the access layer and the convergence layer. After SEP is run at the access layer and the convergence layer, redundancy protection switching can be implemented at the access layer and the convergence layer and the topology of the SEP segment can be displayed.

If the topology of the access layer changes, a node in the SEP segment sends a Flush-FDB packet to instruct other nodes in the SEP segment to refresh the MAC address forwarding table and the ARP table. Edge devices in the SEP segment send TC packets to notify the upper-layer network that the topology of the SEP segment changes.

In multi-ring networking, topology change notification among ring networks needs to be configured.

Figure 9-12 Networking diagram of multiple rings running SEP



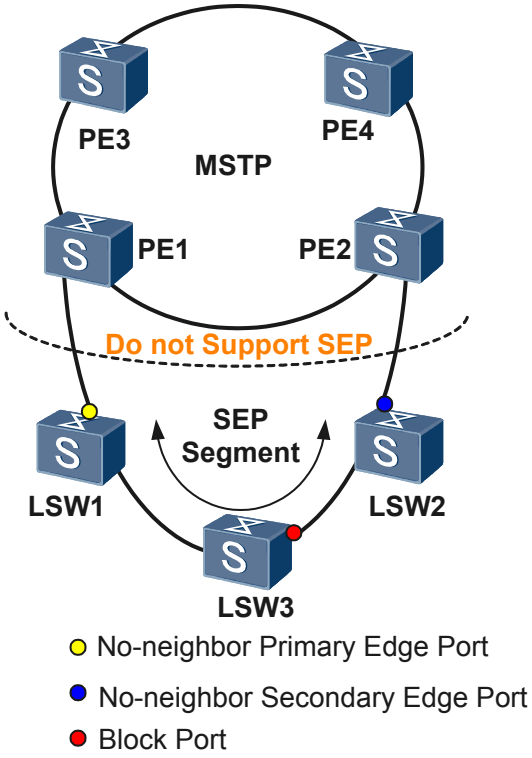
9.4.4 Hybrid SEP+MSTP Ring Networking

As shown in **Figure 9-13**, LSW1 to LSW3 form an SEP segment to access the MSTP ring. The networking is called hybrid SEP+MSTP ring networking. LSW1 to LSW3 are at the access layer and transparently transmit Layer 2 unicast and multicast services. After SEP is run at the access layer, redundancy protection switching can be implemented at the access layer.

If the topology of the access layer changes, a node in the SEP segment sends a Flush-FDB packet to instruct the other nodes in the SEP segment to refresh the MAC address forwarding table and the ARP table. LSW1 and LSW2 at the edge the SEP segment send a TC packet to notify the convergence layer that the topology of the SEP segment changes.

In hybrid-ring networking, no-neighbor edge interfaces need to be deployed on the edge devices of SEP networks, and the SEP networks need to report topology changes to STP networks.

Figure 9-13 Networking diagram of hybrid rings running SEP+MSTP



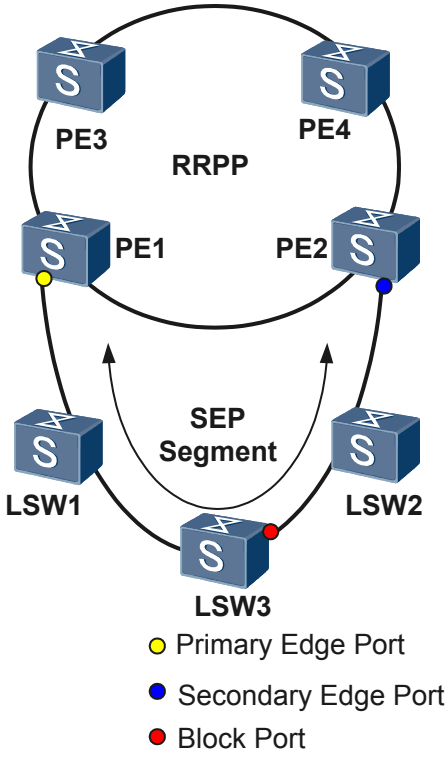
9.4.5 Hybrid SEP+RRPP Ring Networking

As shown in **Figure 9-14**, PE1, PE2 and LSW1 to LSW3 form an SEP segment to access the RRPP ring. The networking is called hybrid SEP + RRPP ring networking. PE1, PE2 and LSW1 to LSW3 are at the access layer and transparently transmit Layer 2 unicast and multicast services. After SEP is run at the access layer, redundancy protection switching can be implemented at the access layer.

If the topology of the access layer changes, a node in the SEP segment sends a Flush-FDB packet to instruct the other nodes in the SEP segment to refresh the MAC address forwarding table and the ARP table. PE1 and PE2 at the edge the SEP segment send a TC packet to notify the convergence layer that the topology of the SEP segment changes.

In hybrid SEP+RRPP ring networking, the SEP networks need to report topology changes to RRPP networks on the edge devices of SEP networks.

Figure 9-14 Networking diagram of hybrid rings running SEP+RRPP



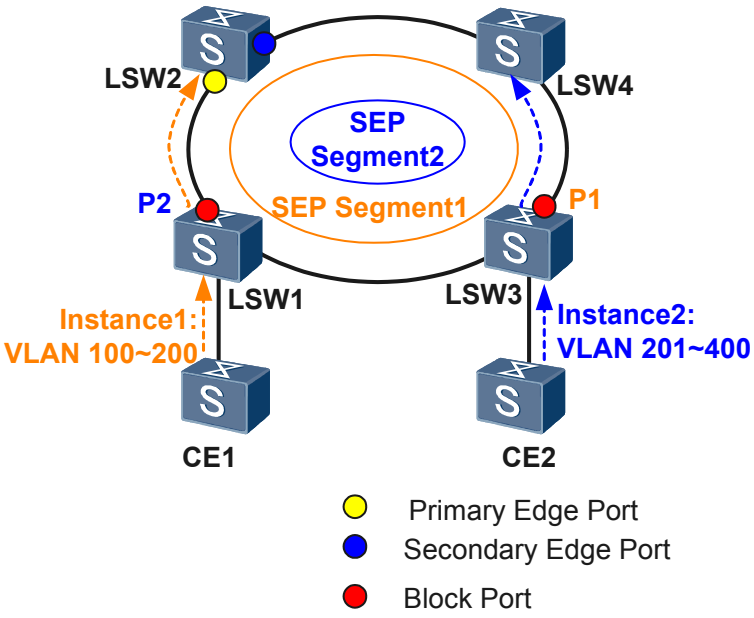
9.4.6 SEP Multi-Instance

As shown in **Figure 9-15**, SEP multi-instance allows two SEP segments to be configured on one physical ring network. All devices, interface roles, and control VLANs in each SEP segment must be configured by conforming to basic SEP configurations principles. Each SEP segment has one blocked interface. Each blocked interface detects whether the physical ring network is complete. The blocked interfaces in the two SEP segments are independent of each other.

Each SEP segment needs to be configured with a protected instance and each protected instance represents a VLAN range. The topology calculated by a SEP segment is valid only for that SEP segment.

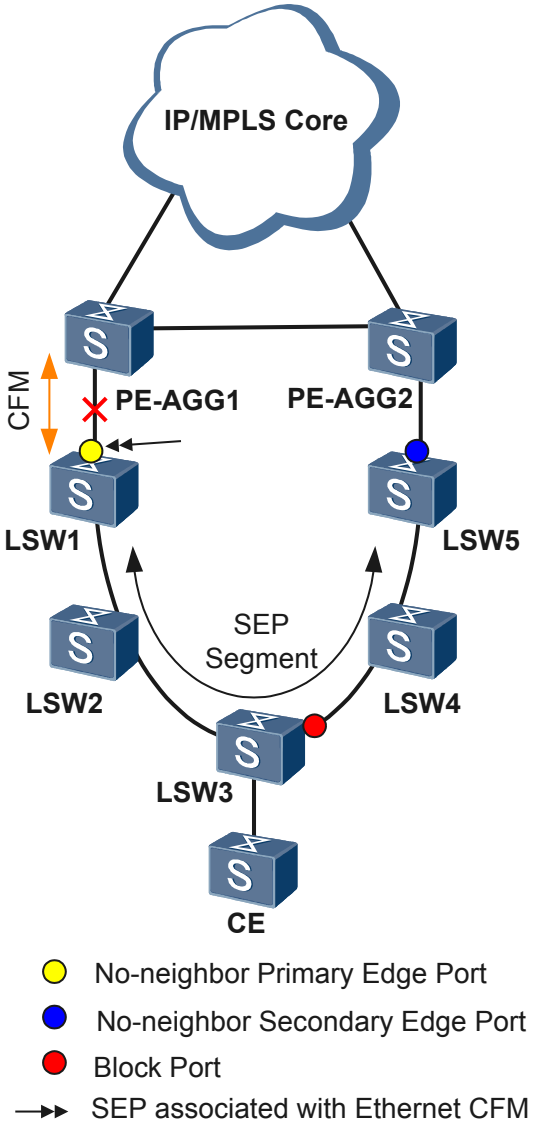
After different protected instances are configured for SEP segments and the mapping between protected instances and VLANs is set, a blocked interface is valid only for the VLANs protected by the SEP segment where the blocked interface resides. Data traffic of different VLANs can be transmitted along different paths. This implements traffic load balancing and link backup.

Figure 9-15 Networking diagram for SEP multi-instance



9.4.7 Association Between SEP and CFM

Figure 9-16 Networking diagram of association between SEP and CFM



As shown in **Figure 9-16**, LSW1 to LSW5 run SEP to implement redundancy protection switching at the access layer and display the topology. Association between SEP and CFM is configured on LSW1 in the SEP segment. When CFM detects a fault on the network at the convergence layer, LSW1 notifies the fault to the Operation, Administration, and Maintenance (OAM) module through a CCM. Then, the SEP status of the interface associated with CFM becomes Down.

The interface associated with CFM is in the SEP segment. Therefore, when the SEP status of the interface associated with CFM goes Down, LSW2 needs to send a Flush-FDB packet to notify the other nodes in the SEP segment that the topology changes. After LSW3 receives the Flush-FDB packet, the blocked interface on LSW3 is unblocked and enters the Forwarding state. Then, the interface sends a Flush-FDB packet to instruct the other nodes in the SEP segment to

refresh the MAC address forwarding table and the ARP table. Therefore, the lower-layer network can detect the fault of the upper-layer network, and reliable service transmission is guaranteed.

9.5 Terms and Abbreviations

Terms

Term	Description
FDB	Forwarding Database, including entries for guiding data forwarding. There are Layer 2 and Layer 3 FDBs. The Layer 2 FDB refers to the MAC address table, which provides information about MAC addresses and outbound interfaces and guides Layer 2 information forwarding. The Layer 3 FDB refers to the ARP table, which provides information about IP addresses and outbound interfaces and guides Layer 3 information forwarding.
MSTP	Multi-Spanning Tree Protocol, a new spanning tree protocol defined in IEEE 802.1s. MSTP introduces the concepts of region and instance. To meet different requirements, MSTP divides a large network into regions where multiple spanning tree instances (MSTIs) are created. These MSTIs are mapped to virtual LANs (VLANs), and bridge protocol data units (BPDUs) carrying information about regions and instances are transmitted between network bridges. Therefore, a network bridge can know which region itself belongs to according to the BPDU information. Multi-instance RSTP is run within regions, whereas RSTP-compatible protocols are run between regions.
RRPP	Rapid Ring Protection Protocol, a link layer protocol exclusively used to prevent loops in Ethernet ring networks. Devices running RRPP detect loops on the network by exchanging information with each other, and block certain interfaces to eliminate loops.
SEP	Smart Ethernet Protection, a link layer protocol exclusively used in Ethernet ring networks. By blocking an interface, SEP can eliminate loops. A network running SEP can interwork with an upper-layer network running STP, RSTP, MSTP, or RRPP. The topology of a network running SEP can be displayed on any device on the network.

Abbreviations

Abbreviation	Full Spelling
EPA	Edge Port Advertisement
LSA	Link Status Advertisement
TC	Topology Change
GR	Graceful Restart

10 Transparent Transmission of Layer 2 Protocol Packets

About This Chapter

10.1 Introduction to Transparent Transmission of Layer 2 Protocol Packets

Definition

Transparent transmission of Layer 2 protocol packets indicates that the packets of standard protocols such as Spanning Tree Protocol (STP), Link Aggregation Control Protocol (LACP), HUAWEI Group Management Protocol (HGMP), and user-defined protocols are transparently transmitted on a Layer 2 network through Layer 2 tunneling technologies.

Purpose

Transparent transmission of Layer 2 protocol packets is a technology used to transparently transmit the protocol packets of users over the ISP network. On the ingress of the ISP network, protocol packets sent by users are forwarded to the ISP network after their multicast destination MAC addresses are changed or modified; on the egress of the ISP network, the multicast destination MAC addresses of the protocol packets are restored to the original ones.

10.2 References

The following table lists the references of this document.

Document	Description	Remarks
IEEE802.1Q	IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks	-
IEEE 802.1ad/ D6.0	Virtual Bridged Local Area Networks-Amendment 4: Provider Bridges	-

10.3 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

10.4 Principles

10.4.1 Basic Concepts of Transparent Transmission of Layer 2 Protocol Packets

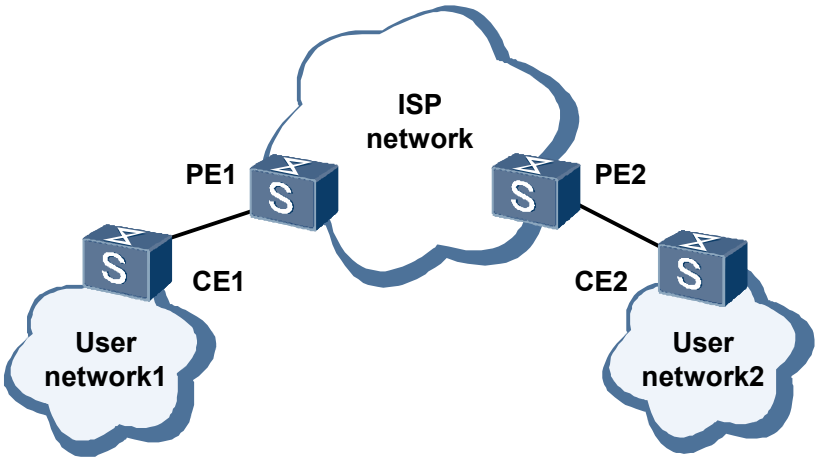
Background of Transparent Transmission of Layer 2 Protocol Packets

Some Layer 2 protocols, such as MSTP, HGMP, and LACP running between user networks, need to traverse the Internet Service Provider (ISP) network to perform Layer 2 protocol calculation.

As shown in **Figure 10-1**, a certain Layer 2 protocol such as MSTP is running in user network1 and user network2. The Layer 2 protocol packets in user network1 must traverse the ISP network to reach user network2 to perform Spanning Tree Protocol (STP) calculation. Generally, the destination MAC addresses of Layer 2 protocol packets are the same. For example, the MSTP packets are BPDUs, of which the destination MAC address is 0180-C200-0000. Therefore, when a Layer 2 protocol packet reaches a PE on the ISP network, the PE sends the protocol packet to the CPU to perform STP calculation, without identifying whether the protocol packet comes from a user network or the ISP network.

In this case, devices in user network1 perform STP calculation together with PE1 rather than devices in user network2. As a result, the Layer 2 protocol packets in user network1 cannot traverse the ISP network to reach user network2.

Figure 10-1 Transparent transmission of Layer 2 protocol packets in the ISP network



To address the preceding problem, you can configure transparent transmission of Layer 2 protocol packets. Currently, the Huawei devices support the transparent transmission of packets of the following Layer 2 protocols:

- Spanning Tree Protocol (STP)
- Link Aggregation Control Protocol (LACP)
- Ethernet Operation, Administration, and Maintenance 802.3ah (EOAM3ah)
- Link Layer Discovery Protocol (LLDP)
- Generic VLAN Registration Protocol (GVRP)
- Generic Multicast Registration Protocol (GMRP)
- HUAWEI Group Management Protocol (HGMP)
- VLAN Trunking Protocol (VTP)
- Unidirectional Link Detection (UDLD)
- Port Aggregation Protocol (PAGP)
- Cisco Discovery Protocol (CDP)
- Per VLAN Spanning Tree Plus (PVST+)
- Dynamic Trunking Protocol (DTP)
- User-defined protocols

If Layer 2 protocol packets need to be transparently transmitted on the ISP network, the following conditions must be met during the transmission process:

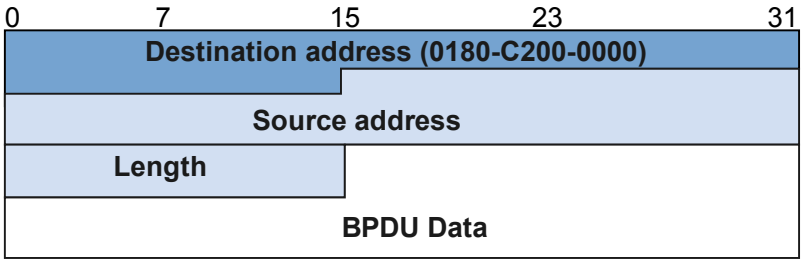
- Each site of a user network can receive the Layer 2 protocol packets from other sites.
- The Layer 2 protocol packets of a user network cannot be processed by the CPUs of the devices on the ISP network.
- Layer 2 protocol packets of different user networks must be isolated and do not affect each other.

Transparent transmission of Layer 2 protocol packets can prevent the Layer 2 protocol packets of different user networks from interfering in each other, which cannot be achieved by the previous technologies.

BPDU

BPDUs are common Layer 2 protocol packets. For example, STP and HGMP use BPDUs as protocol packets. The BPDUs are special protocol packets that are multicast between Layer 2 switches. The encapsulation of BPDUs conforms to IEEE 802.3 and the encapsulation format is shown in **Figure 10-2**. BPDUs of various protocols are multicast with different destination MAC addresses.

Figure 10-2 Format of a BPDU



A BPDU consists of the following fields:

- Destination Address: is of 6 bytes and indicates the destination MAC address.
- Source Address: is of 6 bytes and indicates the source MAC address.
- Length: is of 2 bytes and indicates the length of the BPDU.
- BPDU Data: indicates the contents of the BPDU.

Transparent transmission of Layer 2 protocol packets provides a BPDU tunnel for BPDUs. BPDU tunneling is a Layer 2 tunneling technology that enables the provider network to transparently transmit BPDUs from customer networks at different locations. In this manner, mutual interference between the customer networks and the provider network is prevented.

10.4.2 Principles of Transparent Transmission of Layer 2 Protocol Packets

Layer 2 protocol packets are transparently transmitted based on the following principles:

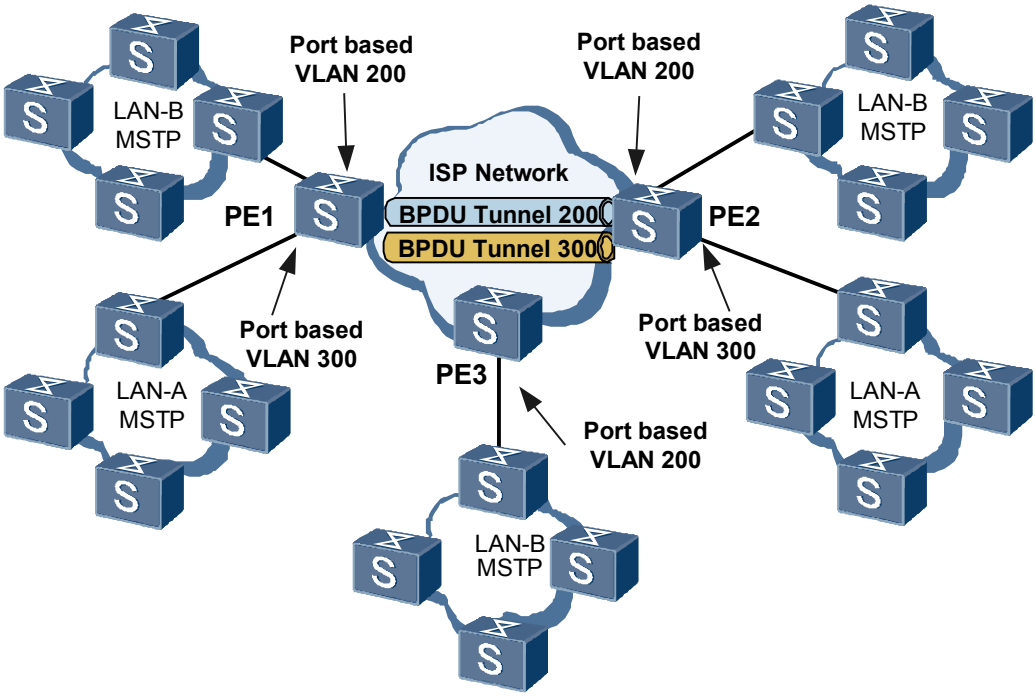
- On the ingress PE of the ISP network, the destination multicast MAC address of a Layer 2 protocol packet is replaced with a specified multicast MAC address.
- The devices on the ISP network determine whether to add an outer VLAN tag to the protocol packet according to the configured transparent transmission mode.
- When the Layer 2 protocol packet reaches the egress, the destination multicast MAC address of the Layer 2 protocol packet is restored to the standard destination multicast MAC address according to the mapping between the specific destination multicast MAC address configured on the device and the Layer 2 protocol. In addition, the egress determines whether to remove the outer VLAN tag according to the configured transparent transmission mode, and then forwards the protocol packet to the UPE.

The Huawei devices support the following transparent transmission modes of Layer 2 protocol packets in different application scenarios:

- Interface-based transparent transmission of Layer 2 protocol packets
- VLAN-based transparent transmission of Layer 2 protocol packets
- QinQ-based transparent transmission of Layer 2 protocol packets

Interface-based Transparent Transmission of Layer 2 Protocol Packets

Figure 10-3 Interface-based transparent transmission of Layer 2 protocol packets



As shown in **Figure 10-3**, each interface on a PE connects to one user network. The user networks belong to different LANs, that is, LAN-A and LAN-B. BPDUs sent from user networks to the PE are untagged. The PE, however, needs to identify that LAN from which the BPDUs come. BPDUs of a user network in LAN-A must be sent to other user networks in LAN-A rather than the user networks in LAN-B. In addition, BPDUs must not be processed by PEs.

In this application scenario, the following processing methods are available:

- Change the default multicast MAC address of the Layer 2 BPDU that can be identified by the devices on the ISP network into another multicast MAC address.
  1. Set the roles of all devices in the ISP network to provider. Thus, the destination MAC addresses of the BPDUs sent by the devices on the ISP network are changed to 01-80-C2-00-00-08 instead of the original 01-80-C2-00-00-00.
  2. Set the roles of all devices in a user network to customer. Thus, the destination MAC addresses of the BPDUs sent by the user network are still 01-80-C2-00-00-00.
  3. On the device of the ISP network, add the interfaces that connect to the same user network to the same VLAN. After receiving the Layer 2 protocol packet from the user network, the device on the ISP network adds the default VLAN ID of the interface to the packet.
  4. The devices (of the provider type) on the ISP network do not take the BPDU as the Layer 2 BPDU and do not send the BPDU to the CPU for processing. Instead, the devices select a corresponding Layer 2 tunnel according to the default VLAN ID of the interface to forward the BPDU.

5. The BPDU is normally forwarded by the devices on the ISP network and normally traverses the ISP network.
6. When reaching the egress on the ISP network, the Layer 2 BPDU is forwarded to the UPE without being changed.

NOTE

This method applies to only the STP, RSTP, or MSTP protocol, and the associated configuration command is **bpdu-tunnel stp bridge role provider**.

- Replace the original multicast MAC address of the Layer 2 BPDU with a specified multicast MAC address.

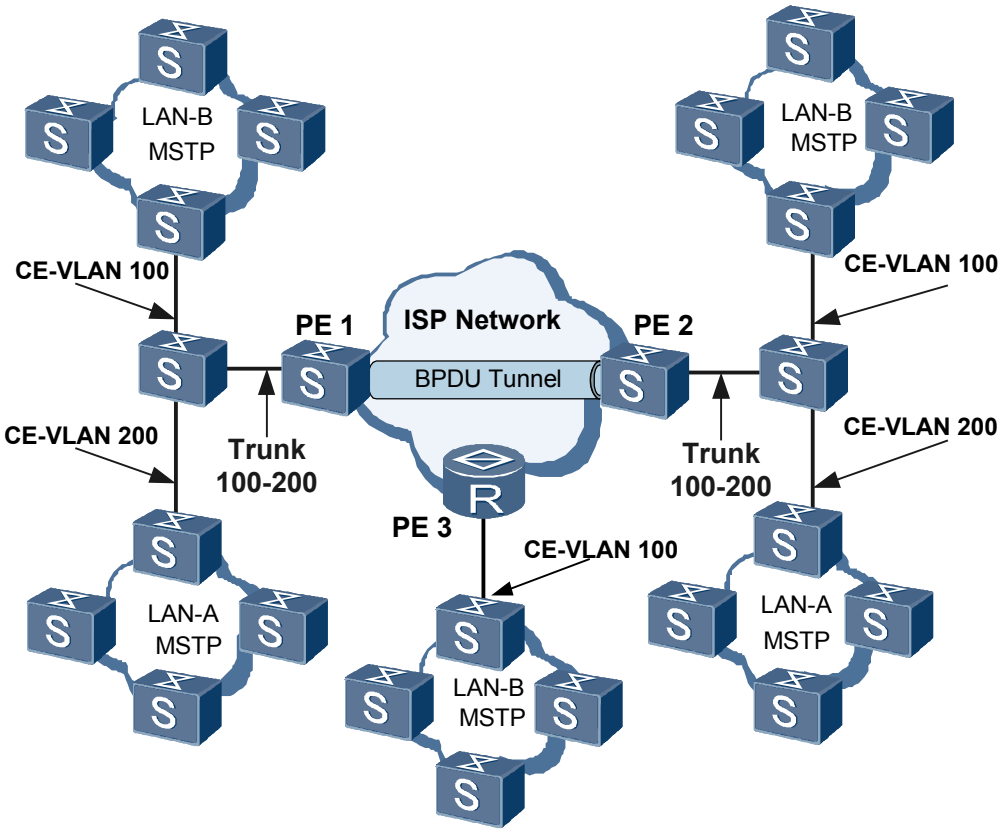
NOTE

This method applies to all types of transparent transmission of Layer 2 protocol packets.

1. After receiving and identifying the Layer 2 protocol packet (such as a BPDU of the STP protocol) from the user network, the device on the ISP network adds the default VLAN ID of the interface to the Layer 2 protocol packet.
2. According to the mapping between the special destination multicast MAC addresses and Layer 2 protocols, the device on the ISP network changes the standard destination multicast MAC address of the Layer 2 BPDU into the specified destination multicast MAC address.
3. The Layer 2 BPDU is normally forwarded by the devices on the ISP network, thus successfully traversing the ISP network.
4. When the Layer 2 BPDU reaches the egress, the egress restores the destination multicast MAC address to the standard destination multicast MAC address of the Layer 2 BPDU according to the mapping between the special destination multicast MAC addresses and Layer 2 protocols, and then forwards the BPDU to the UPE.

VLAN-based Transparent Transmission of Layer 2 Protocol Packets

Figure 10-4 VLAN-based transparent transmission of Layer 2 protocol packets



In most cases, a PE serves as a convergence device. As shown in **Figure 10-4**, the convergence interface on PE1 receives Layer 2 protocol packets from LAN-A and LAN-B. To differentiate BPDUs from two LANs, BPDUs sent from the CE to the PE must be tagged. The VLAN ID of a BPDU from LAN-A is 200 and the VLAN ID of a BPDU from LAN-B is 100.

Currently, some Layer 2 protocol packets, such as protocol packets of STP, RSTP, or MSTP, do not carry VLAN tags. When receiving Layer 2 protocol packets with VLAN tags, a device on the ISP network considers them as invalid protocol packets and discards them. To avoid this problem, you can configure VLAN-based transparent transmission of Layer 2 protocol packets on the devices on the ISP network. In this manner, the Layer 2 protocol packets can traverse the ISP network through Layer 2 tunnels.

Similar to the interface-based transparent transmission of Layer 2 protocol packets, there are two processing methods in this application scenario:

- Change the default multicast MAC address of the Layer 2 protocol packet that can be identified by the device on the ISP network into another multicast MAC address.

  1. Set the roles of all devices in the ISP network to provider. Thus, the destination MAC addresses of the BPDUs sent by the devices in the ISP network are changed to 01-80-C2-00-00-08 instead of the original 01-80-C2-00-00-00.

2. Set the roles of all devices in a user network to customer. Thus, the destination MAC addresses of the BPDUs sent by the user network are still 01-80-C2-00-00-00.
3. Set specific VLAN IDs for the Layer 2 protocol packets that are sent from user networks to the ISP network.
4. Configure the devices in the ISP network to identify the Layer 2 protocol packets with VLAN IDs and allow the packets to pass through.
5. The devices (of the provider type) on the ISP network do not take the packet as the BPDU and do not send the packet to the CPU for processing. Instead, the devices select a corresponding Layer 2 tunnel to forward the packet according to the VLAN IDs with which the packets are allowed to pass through.
6. The Layer 2 protocol packet is transmitted as an ordinary Layer 2 packet by the devices on the ISP network, thus successfully traversing the ISP network.
7. When reaching the egress on the ISP network, the Layer 2 protocol packet is forwarded to the CE without being changed.

NOTE

This method applies to only the STP, RSTP, or MSTP protocol, and the related configuration command is **bpdu-tunnel stp bridge role provider**.

- Replace the original multicast MAC address of the Layer 2 protocol packet with a specified multicast MAC address.

NOTE

This method applies to transparent transmission of all types of Layer 2 protocol packets.

1. Set specific VLAN IDs for the Layer 2 protocol packets that are sent from user networks to the ISP network.
2. Configure the devices on the ISP network to identify the Layer 2 protocol packets with VLAN IDs and allow the packets to pass through.
3. According to the mapping between the specified destination multicast MAC address and the Layer 2 protocol, the device on the ISP network changes the standard destination multicast MAC address of the Layer 2 protocol packet into the specified destination multicast MAC address.
4. After the MAC address is changed, the Layer 2 protocol packet is transmitted as an ordinary Layer 2 packet by the devices on the ISP network, thus successfully traversing the ISP network.
5. When the Layer 2 protocol packet reaches the egress, the egress restores the destination multicast MAC address to the standard destination multicast MAC address according to the mapping between the specified destination multicast MAC addresses and Layer 2 protocols, and then forwards the Layer 2 protocol packet to the CE.

QinQ-based Transparent Transmission of Layer 2 Protocol Packets

- QinQ overview

  The QinQ protocol is a Layer 2 tunneling protocol based on the IEEE 802.1Q technology. The QinQ technology improves the utilization of VLANs by adding another 802.1Q tag. In this manner, services in the private VLAN can be transparently transmitted on the public network. The packet transmitted on the ISP network carries double 802.1Q tags (a public VLAN tag and a private VLAN tag), that is, 802.1Q-in-802.1Q. It is also called the QinQ protocol.

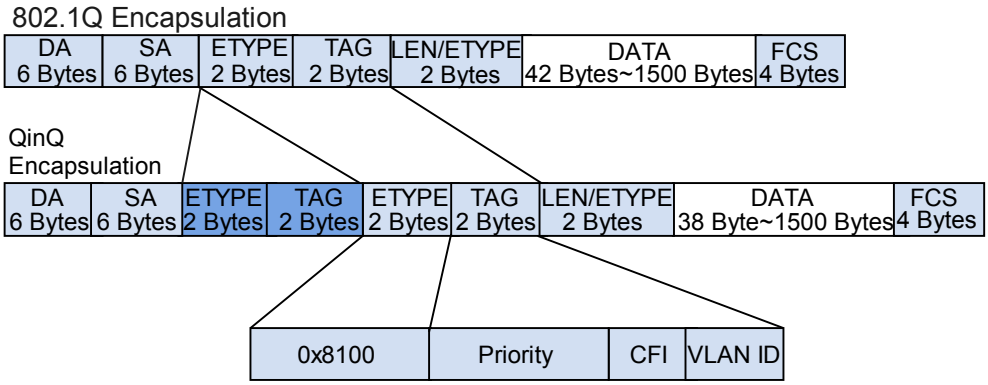
  Figure 10-5 shows the format of a QinQ packet. Compared with the 802.1Q packet, the QinQ packet has a tag suffixed to the source address (SA). This tag is known as the outer

tag or public tag, used for carrying the VLAN ID of a public network. The inner tag is usually known as the private tag, used for carrying the VLAN ID of a private network.

NOTE

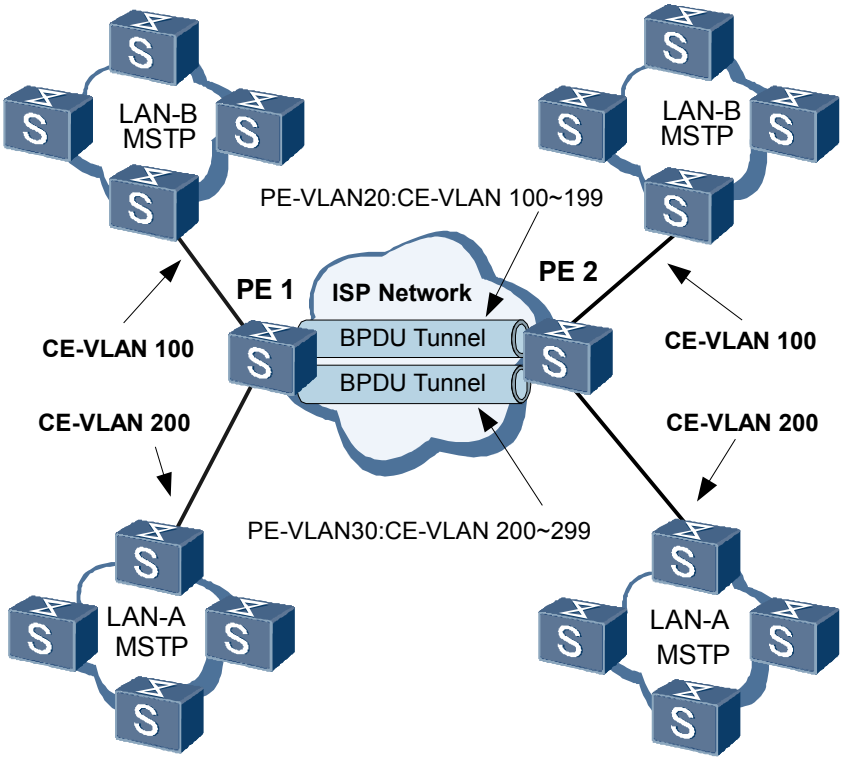
The QinQ function configured on a Layer 2 interface is also called VLAN stacking.

Figure 10-5 802.1Q Encapsulation and QinQ Encapsulation



- QinQ-based transparent transmission of Layer 2 protocol packets

Figure 10-6 QinQ-based transparent transmission of Layer 2 protocol packets



If Layer 2 protocol packets are still transmitted transparently in VLAN-based mode when many user networks are connected to the ISP network, a large number of VLAN IDs of the

ISP network are required. This may result in insufficient VLAN ID resources. In this case, you can configure the QinQ function to forward Layer 2 protocol packets.

As shown in **Figure 10-6**, the convergence interfaces on the PEs are configured with the function of QinQ-based transparent transmission of Layer 2 protocol packets. Then, the PEs add different outer tags to the packets from different user networks.

1. Set specific VLAN IDs for the Layer 2 protocol packets that are sent from user networks to the ISP network.
2. Configure transparent transmission of Layer 2 protocol packets and the QinQ function on the interfaces of the ingress on the ISP network.
3. According to the user VLAN IDs, the ingress on the ISP network allocates different outer tags, that is, the public VLAN IDs, to the Layer 2 protocol packets.
4. The ingress on the ISP network selects different Layer 2 tunnels according to different outer tags. Then, the layer 2 protocol packets are transmitted as ordinary Layer 2 packets by the devices on the ISP network.
5. Configure transparent transmission of Layer 2 protocol packets and the QinQ function on the interfaces of the egress on the ISP network.
6. The egress removes the outer tags and forwards the Layer 2 protocol packets to the corresponding user networks according to the inner tags.

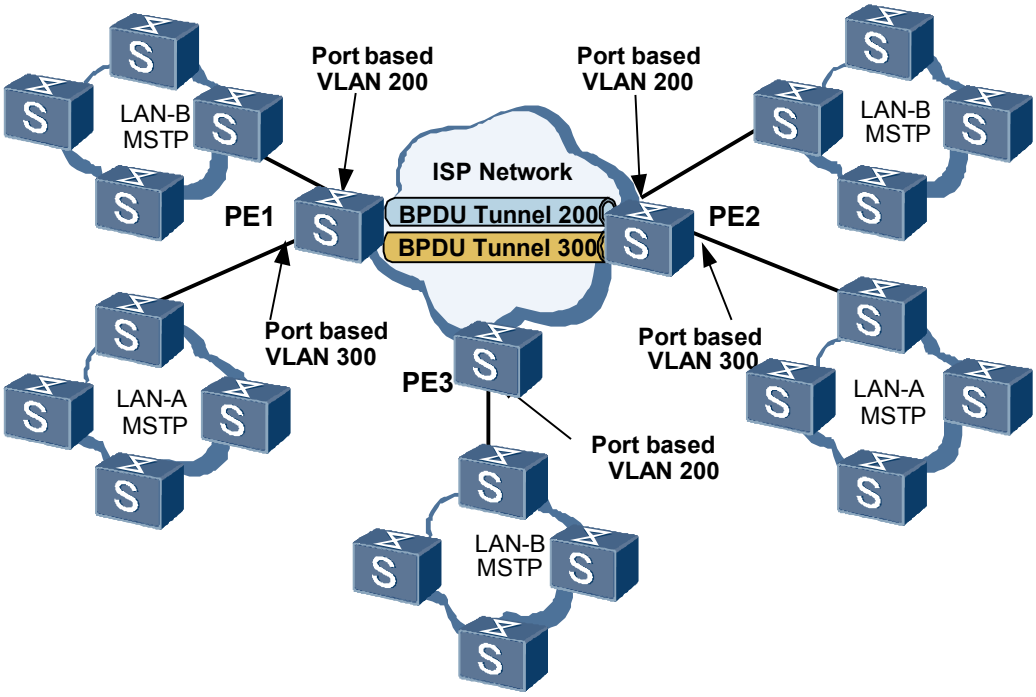
As shown in **Figure 10-6**, after receiving a Layer 2 protocol packet from a VLAN with the ID ranging from 100 to 199, PE1 adds VLAN ID 20 as an outer VLAN ID to the packet, and forwards the packet on the ISP network through a Layer 2 tunnel. After receiving a Layer 2 protocol packet from a VLAN with the ID ranging from 200 to 299, PE1 adds VLAN ID 30 as an outer VLAN ID to the packet, and forwards the packet on the ISP network through a Layer 2 tunnel. In this manner, Layer 2 protocol packets from different user networks can be transparently transmitted on the ISP network, and VLAN ID resources of the operator can be saved.

10.5 Applications

10.5.1 Interface-based Transparent Transmission of Layer 2 Protocol Packets

As shown in **Figure 10-7**, PEs on the Layer 2 switching network can transparently transmit Layer 2 Control Protocol packets from access users.

Figure 10-7 Interface-based transparent transmission of Layer 2 control protocol packets on a Layer 2 network



PE1, PE2, and PE3 are connected to construct a Layer 2 switching network, and access LAN-A and LAN-B through different interfaces. Each LAN runs Layer 2 control protocol packets. Here, MSTP is taken as an example.

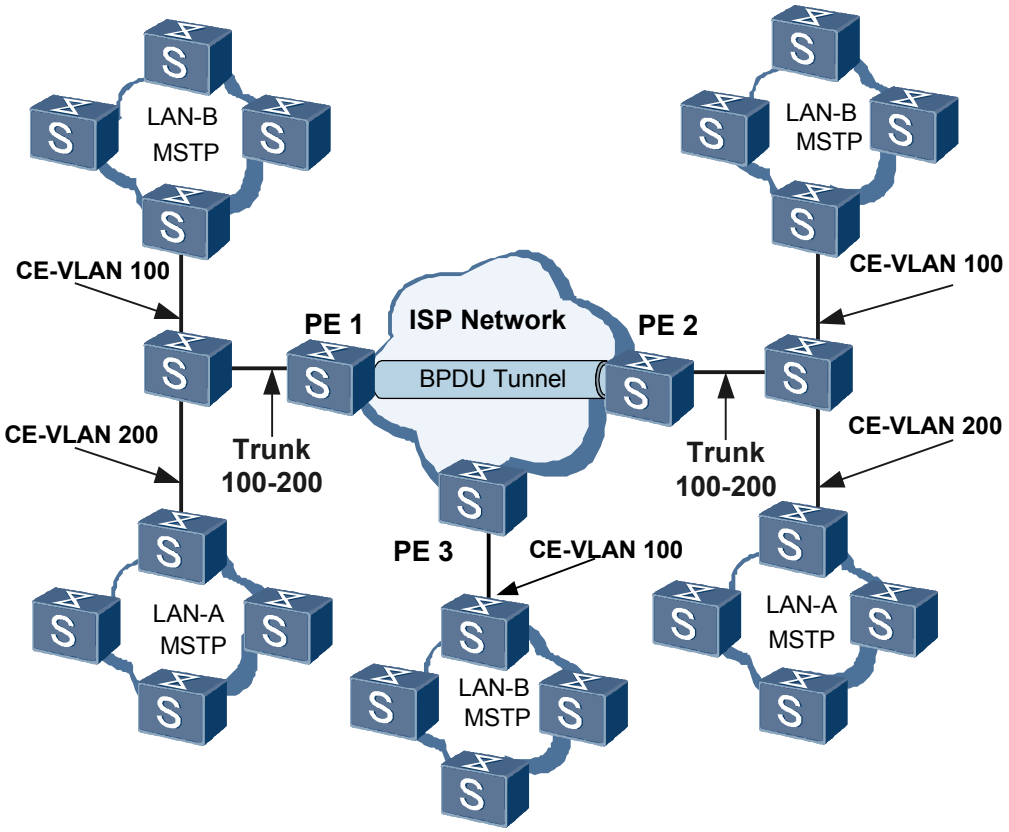
The process of transparently transmitting Layer 2 control protocol packets is as follows:

- The type of the Layer 2 control protocol packets that need to be transparently transmitted is set on the interfaces that connect PE1, PE2, and PE3 to CEs, and the original multicast MAC address of Layer 2 protocol packets from user networks is replaced with a specified multicast MAC address.
- After identifying that the packets received from CEs are Layer 2 control protocol packets, PE1 replaces the original multicast MAC address of the packets with the specified multicast MAC address according to the configured mapping, and then forwards the packets. The packets whose multicast MAC address is replaced with the specified multicast MAC address are forwarded as common Layer 2 packets on the ISP network.
- When the packets reach PE2, PE2 restores the multicast MAC address of the packets to the standard multicast MAC address according to the configured mapping between multicast MAC addresses and Layer 2 control protocol packets, and then forwards the packets to the corresponding CE, completing transparent transmission of Layer 2 protocol packets.

10.5.2 VLAN-based Transparent Transmission of Layer 2 Protocol Packets

As shown in **Figure 10-8**, Layer 2 control protocol packets with a VLAN tag need to be transparently transmitted. Therefore, the devices in VLAN 100 and VLAN 200 are required to transparently transmit the Layer 2 protocol packets.

Figure 10-8 VLAN-based transparent transmission of Layer 2 control protocol packets on a Layer 2 network



PE1, PE2, and PE3 are connected to construct a Layer 2 ISP network. CEs add one tag to Layer 2 control protocol packets from user networks and then send them to the PEs. The packets received by PEs have only one tag.

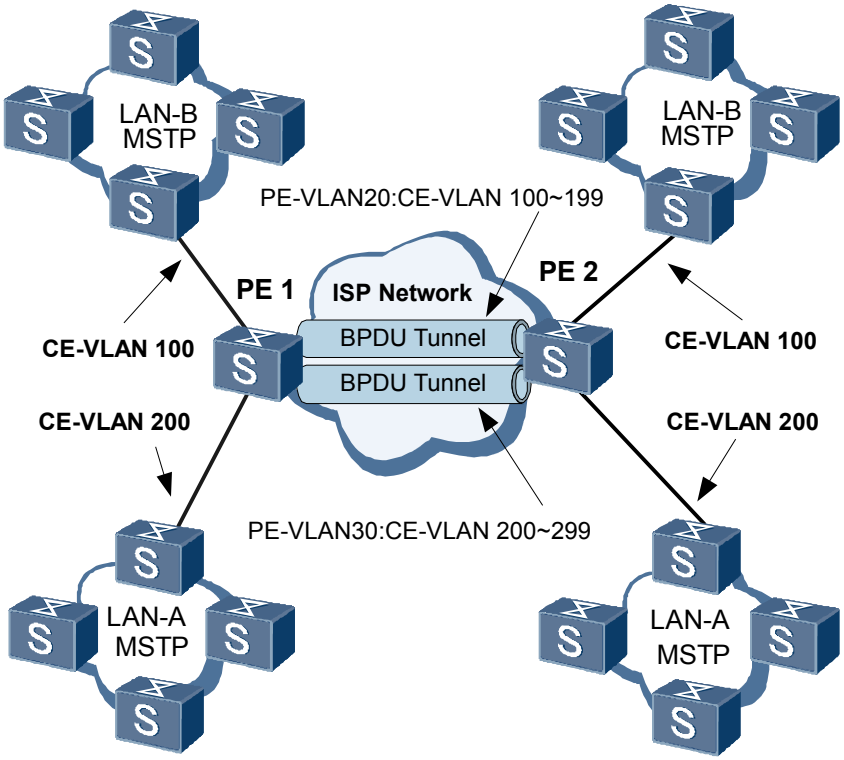
The process of transparently transmitting Layer 2 control protocol packets is as follows:

- VLAN-based transparent transmission of Layer 2 control protocol packets is configured on the interfaces that connect PE1, PE2, and PE3 to CEs.
- After identifying that the packets received from CEs are Layer 2 control protocol packets, PE1 replaces the original multicast MAC address of the packets with a specified multicast MAC address according to the configured mapping, and then forwards the packets. The packets whose multicast MAC address is replaced with the specified multicast MAC address are forwarded as common Layer 2 VLAN packets on the ISP network.
- When the packets reach PE2, PE2 restores the multicast MAC address of the packets to the standard multicast MAC address according to the configured mapping between the multicast MAC addresses and Layer 2 control protocol packets, and then forwards the packets to the corresponding CE, completing transparent transmission of Layer 2 protocol packets.

10.5.3 QinQ-based Transparent Transmission of Layer 2 Protocol Packets

As shown in **Figure 10-9**, when the edge devices on the ISP are connected to a large number of VLAN users, you can configure QinQ-based transparent transmission of Layer 2 control protocol packets on the devices to save VLAN resources.

Figure 10-9 QinQ-based transparent transmission of Layer 2 control protocol packets



PE1 and PE2 are connected to construct a Layer 2 switching network. VLAN 20 and VLAN 30 are configured on the PEs. CEs send tagged Layer 2 control protocol packets (VLAN ID being 100 or 200) to the PEs. QinQ is configured on the interfaces that connect PE1 and PE2 to CEs.

The process of transparently transmitting Layer 2 control protocol packets is as follows:

- Set specific VLAN IDs for the Layer 2 protocol packets that are sent from user networks to the ISP network.
- Configure transparent transmission of Layer 2 protocol packets and the QinQ function on the interfaces of the ingress on the ISP network.
- According to the user VLAN IDs, the ingress on the ISP network allocates different outer tags, that is, the public VLAN IDs, to the Layer 2 protocol packets.
- The ingress on the ISP network selects different Layer 2 tunnels according to different outer tags. Then, the layer 2 protocol packets are transmitted as ordinary Layer 2 packets by the devices on the ISP network.
- Configure transparent transmission of Layer 2 protocol packets and the QinQ function on the interfaces of the egress in the ISP network.

- The egress removes the outer tags and forwards the Layer 2 protocol packets to the corresponding user networks according to the inner tags.

10.6 Terms and Abbreviations

Acronyms and Abbreviations

Acronym and Abbreviation	Full Name
BPDU	Bridge Protocol Data Unit
STP	Spanning Tree Protocol
LACP	Link Aggregation Control Protocol
LLDP	Link Layer Discovery Protocol
GMRP	Generic Multicast Registration Protocol
GVRP	Generic VLAN Registration Protocol
HGMP	HUAWEI Group Management Protocol

11 HVRP

About This Chapter

11.1 Introduction to HVRP

Definition

Hierarchy VLAN Register Protocol (HVRP) can dynamically register and age the VLANs on the interfaces that do not forward packets. This saves MAC addresses.

Purpose

When constructing a metropolitan area network (MAN), carriers usually adopt the ring topology or tree topology. Regardless of the topology, devices on the convergence layer must support a large number of MAC address entries to meet the requirements of users. The number of users on the network increases quickly, and the MAC addresses supported by a switch may be insufficient for the users connected to the switch. As a result, the switch cannot learn the MAC addresses of some users. In this case, packets are broadcast in the VLAN, which wastes network bandwidth and degrades the network performance.

The HVRP protocol can be used when the number of MAC addresses supported by a switch is smaller than the total number of users connected to the switch. HVRP can identify user VLANs (that is, local VLANs) and non-user VLANs. In special networking, HVRP can save MAC addresses and increase the number of users that the switch supports.

11.2 References

The following table lists the references of this document.

Document	Description	Remarks
IEEE Std 802.1D	Information technology-Telecommunications and information exchange between systems-Local and metropolitan area networks-Common specifications-Media Access Control (MAC) Bridges	-
IEEE Std 802.1Q	IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks	-

11.3 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V100R003, V100R006, V200R001

11.4 Principles

11.4.1 Basic Concepts

Terms of HVRP

- HVRP interface

  An HVRP interface is an interface that is configured with HVRP attributes and can send, receive, and process HVRP packets.
- HVRP root interface

  An HVRP root interface is an HVRP interface that functions as the root interface in an STP region.
- HVRP designated interface

  An HVRP designated interface is an HVRP interface that functions as the designated interface in an STP region.
- Local VLAN

  A local VLAN is a VLAN that does not contain any HVRP interface.
- VLAN registration

  VLAN registration is a process of adding HVRP interfaces to VLANs meeting certain conditions in tagged mode.
- VLAN aging

  VLAN aging is a process of deleting a VLAN from an HVRP interface.
- Permanent VLAN

  A permanent VLAN is a VLAN that are never aged by an HVRP interface.
- Sending local VLAN information

  After STP and HVRP are enabled, the HVRP root interface sends HVRP packets containing the local VLAN information.
- VLAN registration timer

  The VLAN registration timer specifies the interval for the HVRP root interface to send HVRP VLAN registration packets.
- Aging timer of registered VLANs

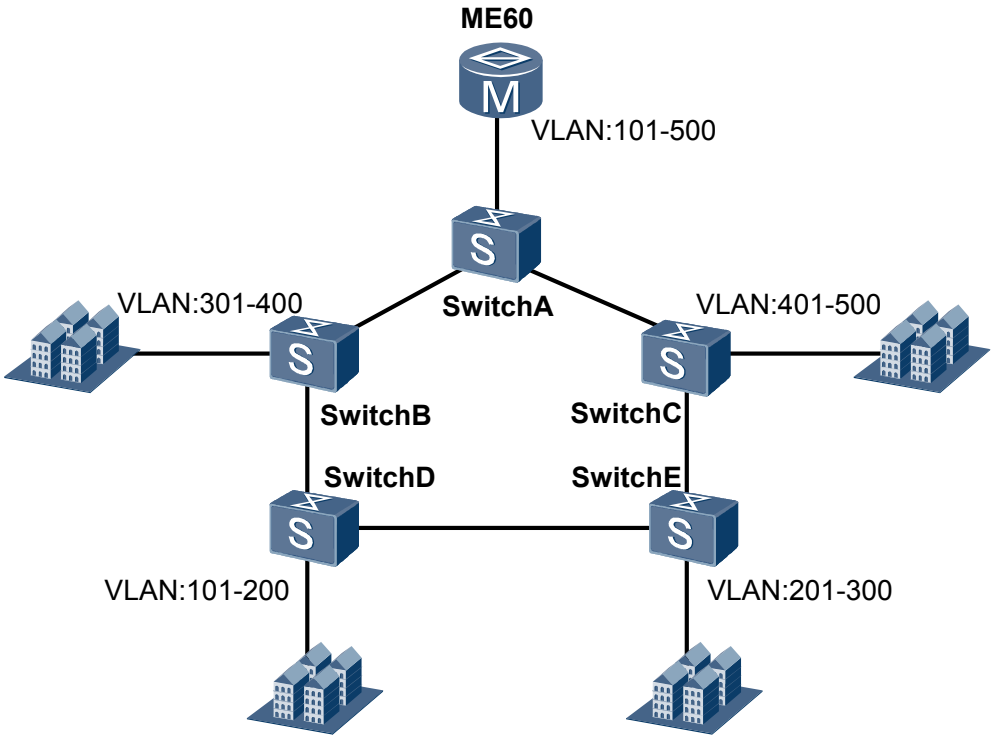
  The aging timer of registered VLAN specifies the aging time of registered VLANs. If the HVRP designated interface does not receive the registration packet of a VLAN within the aging time, the VLAN is aged on the HVRP designated interface.

11.4.2 Working Procedure

Figure 11-1 shows the networking of HVRP. The working mechanism of HVRP is described based on this networking.

- STP is enabled on the entire network, and the HVRP root interface and HVRP designated interfaces are calculated through STP.
- The Switches are connected through trunk interfaces. The trunk interfaces are all enabled with HVRP and can forward packets of VLAN 101 to VLAN 500.
- HVRP is disabled on the interfaces outside the STP network, that is, edge interfaces.

Figure 11-1 Networking diagram of HVRP



The HVRP application involves the following operations:

1. Registering VLANs
   - Each Switch periodically sends the local VLAN information through the HVRP root interface.
   - Each Switch forwards the received local VLAN information through the HVRP root interface. In addition, each Switch registers local VLANs on the HVRP designated interface according to the local VLAN information received from the HVRP designated interface.
   - VLAN registration and aging can be performed only on HVRP designated interfaces.
   - A VLAN can be registered on an interface only after the interface is added to the VLAN statically. For example, if an HVRP designated interface does not belong to VLAN 999, VLAN 999 cannot be registered on the HVRP designated interface even if the interface receives an HVRP packet with local VLAN 999.
2. Aging VLANs

If an HVRP designated interface does not receive any VLAN registration packet within the aging time, the VLANs on the HVRP designated interface are aged.

By default, only local VLANs are aged. You can configure the S7700 to age all the VLANs.

3. Sending and maintaining local VLAN information

   The HVRP root interface periodically sends local VLAN registration packets according to the VLAN registration timer.

   When the role of a local VLAN changes, for example, the VLAN is not a local VLAN any more because the configuration is changed, the Switch sends the local VLAN information through the HVRP root interface immediately.

4. Re-registering VLANs when the status of an HVRP interface changes to Up or Down

   When the status of an HVRP interface changes to Up or Down, the aged VLANs may interrupt forwarding of Layer 2 packets on the entire network. Therefore, when a Switch detects that the status of an HVRP interface changes, the Switch immediately sends a restore packet to notify all the other Switches on the network. Then the Switches re-register the aged VLANs on the original interfaces.

5. Re-registering VLANs when the STP role of an HVRP interface changes

   When the STP role of an HVRP interface changes, the aged VLANs on this interface are re-registered on the interface.

6. Counting interfaces in a VLAN

   - A Switch updates the number of interfaces in a VLAN every time an interface is added to or deleted from the VLAN, the VLAN is registered, or the VLAN is aged.
   - A trunk interface is counted as one interface.

7. MAC address learning mode in a VLAN

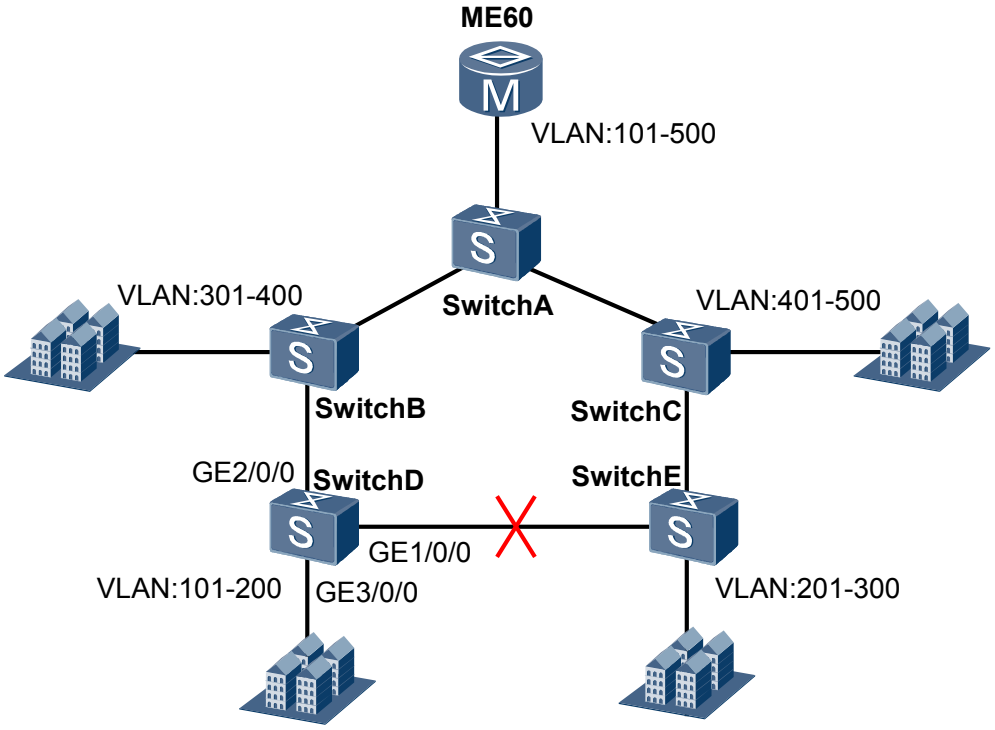
   - When a VLAN contains more than two non-aged interfaces, the interfaces learn MAC addresses.
   - When a VLAN contains two or less than two non-aged interfaces, the interfaces do not learn MAC addresses. In addition, the dynamic MAC addresses learned before are deleted.

11.5 Applications

A switch on a Layer 2 network needs to learn a large number of MAC addresses. To reduce the MAC addresses that the switch needs to learn, you can enable HVRP on the switch. As shown in **Figure 11-2**, HVRP needs to be configured on a single-ring network.

Through the dynamic VLAN registration and aging mechanism, HVRP ages the VLANs on the interfaces that do not forward packets and saves only necessary VLANs. When a VLAN contains two or less than two interfaces, the interfaces do not need to learn MAC addresses. Instead, the interfaces broadcast data packets in the VLAN without affecting the bandwidth.

Figure 11-2 Networking diagram of HVRP application



11.6 Terms and Abbreviations

Abbreviation

Abbreviation	Full Spelling
HVRP	Hierarchy VLAN Register Protocol

12 Loopback Detection

About This Chapter

12.1 Loopback Detection Overview

Definition

The loopback detection function detects loops on the network connected to an interface by checking loopbacks on the interface.

Purpose

A loopback occurs when packets sent from an interface are sent back to the interface. Loopbacks are usually caused by loops on the network connected to the interface, and loops cause broadcast storms.

Figure 12-1 Schematic diagram of loopback detection

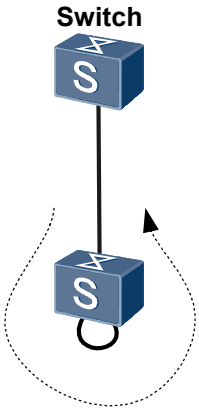
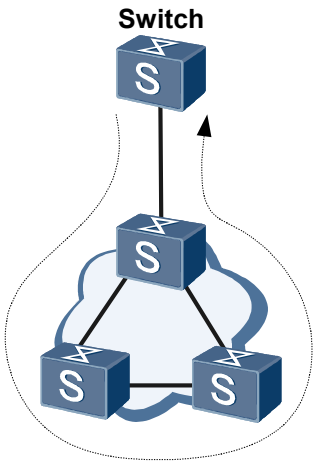


Figure 12-2 Schematic diagram of loopback detection



After loopback detection is enabled on the S7700, the S7700 periodically detects loopbacks on each Ethernet interface.

When a loopback is detected on an interface, the S7700 switches the interface to the loopback detection state. In addition, the S7700 sends a trap and blocks or shuts down the interface to minimize impact on the system and the entire network caused by the loop.

12.2 Availability

Involved Network Element

None.

License Support

This feature can be used without a license.

Version Support

Product	Version
S7700	V200R001

12.3 Principles

How Loopback Detection Works

After loopback detection is enabled on an interface, the interface sends a loopback detection packet every 5s by default. If a loopback detection packet is sent back to the interface, there is a loopback on the interface, indicating that a loop occurs on the network connected to the interface. Then the interface switches to the loopback detection state. The interface automatically restores to the previous status three detection intervals after the loop is removed.

You can configure the action performed on an interface when a loopback is detected, as shown in the following table.

Action	Description
Trap	Sends a trap.
Block	Sends a trap and blocks the interface. The interface can be unblocked automatically.
Shutdown	Sends a trap and shuts down the interface. The interface needs to be started manually.
Nolearning	Sends a trap and disables MAC address learning on the interface. MAC address learning can be re-enabled automatically.

After loopback detection is enabled on an interface, the interface sends untagged loopback detection packets by default. You can configure the interface to send loopback detection packets of a specified VLAN.

Various interfaces send loopback detection packets as follows:

- Access interfaces and dot1q tunnel interfaces can send only untagged loopback detection packets.
- Trunk interfaces and hybrid interfaces can send both untagged and tagged loopback detection packets. If a VLAN is specified on a trunk or hybrid interface, the interface sends loopback detection packets with the specified VLAN tag. A maximum of eight VLANs can be specified on an interface. If an interface does not belong to the specified VLAN, the interface does not send loopback detection packets with the specified VLAN tag. If no VLAN is specified on an interface, the interface sends untagged loopback detection packets.

NOTE

Loopback detection cannot be configured on an Eth-Trunk or its member interfaces.

12.4 Terms and Abbreviations

None