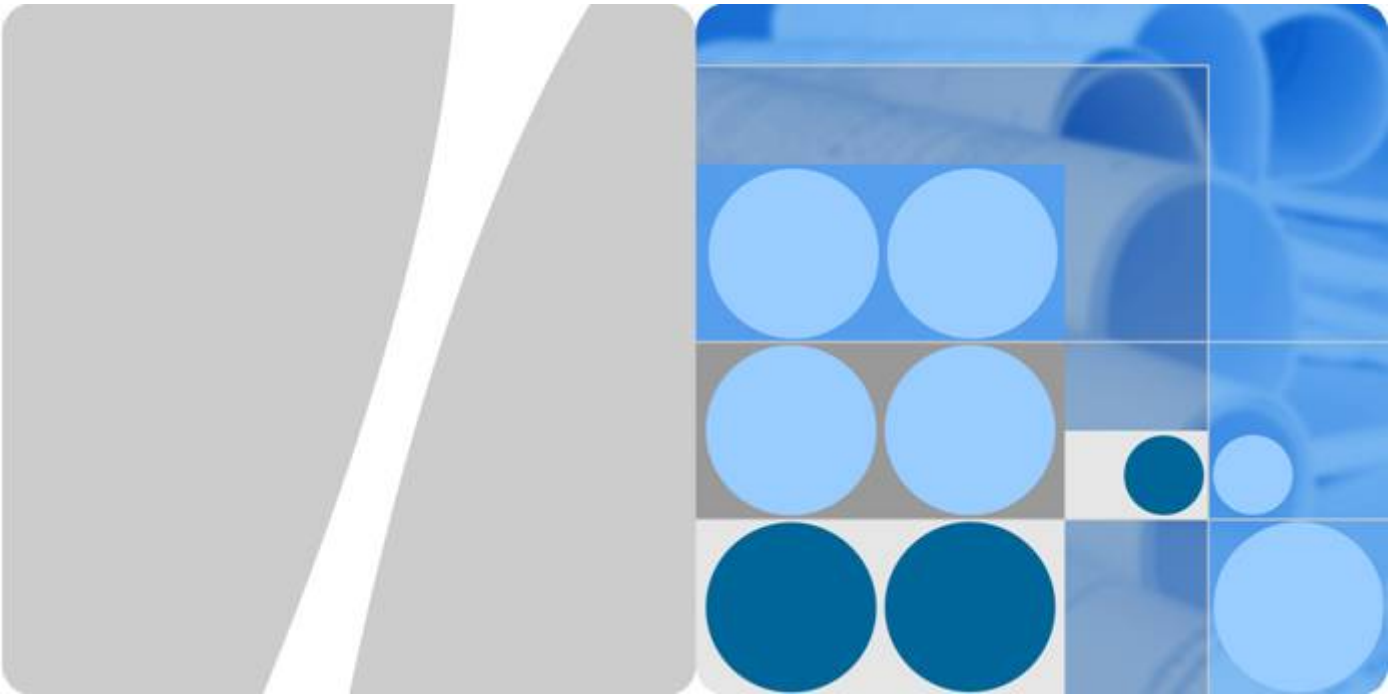


Quidway S6700 Series Ethernet Switches

V100R006C00

Feature Description - Reliability

Issue 01

Date 2011-07-15

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 the 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://www.huawei.com

Email: support@huawei.com

About This Document

Intended Audience

This document describes the Reliability 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 Reliability 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 01 (2011-07-15)

Initial commercial release.

Contents

1 DLDP

About This Chapter

1.1 DLDP Overview

Definition

The Device Link Detection Protocol (DLDP) monitors the link status of optical fibers or copper twisted-pair cables such as super Category 5 twisted pairs. If a unidirectional link is found on an interface, DLDP automatically shuts down or requests users to manually shut down the interface. This prevents network faults.

Purpose

As shown in **Figure 1-1** and **Figure 1-2**, unidirectional links occur on networks. That is, the local device can receive packets from the remote device through the link layer, but the remote device cannot receive packets from the local device. Unidirectional links result in problems such as loops on an STP-enabled network.

Take fiber links as an example. Unidirectional links are classified into the following types:

- Unidirectional link caused by crossed connections of fibers
- Unidirectional link caused by disconnection of fibers

Figure 1-1 Unidirectional link caused by crossed connections of fibers

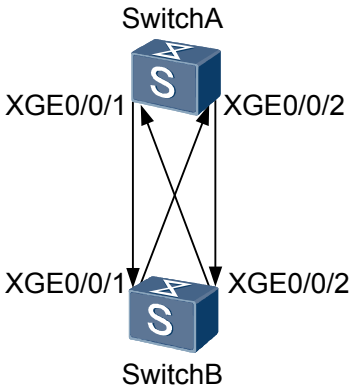
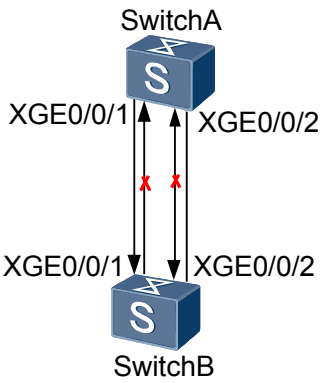


Figure 1-2 Unidirectional link caused by disconnection of a fiber



As a link layer protocol, DLDP works with physical layer protocols to detect the link status of devices. The auto negotiation mechanism at the physical layer detects physical signals and faults, and DLDP identifies the remote device and unidirectional link and disables unreachable interfaces. The auto negotiation mechanism and DLDP work together to detect and close unidirectional links at physical and logical layers.

If the interfaces on both ends of the link work normally at the physical layer, DLDP checks connections and packet exchange between the two interfaces at the link layer. This process cannot be implemented through the auto negotiation mechanism.

1.2 References

None.

1.3 DLDP Principle

DLDP Status

DLDP has the following statuses:

- Initial
- Inactive
- Active
- Advertisement
- Probe
- Disable
- DelayDown

Table 1-1 DLDP status

Status	Description
Initial	Indicates that DLDP is disabled.
Inactive	Indicates that DLDP is enabled but the link is Down.
Active	Indicates that DLDP is enabled and the link is Up, or entries of neighbors are cleared.
Advertisement	Indicates the status when all neighbors are bidirectionally reachable or have been in Active state for more than five seconds. This is a stable state where no unidirectional link has been detected.
Probe	Indicates the status when probe packets are sent to detect whether the link is unidirectional. When an interface enters this state, DLDP starts the probe timer and starts an echo timer for each neighbor to be detected.

Status	Description
Disable	Indicates the status when DLDP detects a unidirectional link or a neighbor disappears in enhanced mode. In this state, the DLDP-enabled interface does not receive or send DLDP packets.
Delaydown	Indicates the temporary status. When an interface in Active, Advertisement, or Probe state receives a Port-Down event, it enters this state instead of deleting neighbor entries immediately and transiting to the Inactive state. In this state, DLDP neighbor information is reserved and the system only responds to Port-Up events.

DLDP Timers

DLDP uses the following timers.

Table 1-2 DLDP timers

Timer	Description
Active timer	This timer determines the interval for sending Advertisement packets with RSY tags. The default interval is 1s. That is, a DLDP-enabled interface in Active state sends one Advertisement packet with RSY tags every second by default. Up to five Advertisement packets with RSY tags can be sent successively.
Advertisement timer	This timer determines the interval for sending Advertisement packets, which can be set through a command. By default, the interval for sending Advertisement packets is 10s.
Probe timer	This timer determines the interval for sending Probe packets (the default value is 1s). A DLDP-enabled interface in Probe state sends two Probe packets every second.
Echo timer	This timer is triggered when DLDP transits to the Probe state. The value is 10s. If the local interface in Probe state does not receive any Echo packet from a neighbor when the Echo timer expires, the interface is set to unidirectional and DLDP transits to the Disable state. In this case, the system displays logs and information about traced packets and sends Flush packets. In addition, the local interface is manually or automatically shut down according to the DLDP Down mode and the neighbor entry is deleted.

Timer	Description
Neighbor aging timer	When a new neighbor joins, a neighbor entry is created and the corresponding neighbor aging timer is triggered. When a DLDP packet is received, a DLDP-enabled interface updates the corresponding neighbor entry and resets the neighbor aging timer. ● In normal mode, if no DLDP packet is received from a neighbor when the corresponding neighbor aging timer expires, a DLDP-enabled device sends Advertisement packets with RSY tags and removes the neighbor entry. ● In enhanced mode, the enhanced timer is triggered if no DLDP packet is received from a neighbor when the neighbor aging timer expires. The value of the neighbor aging timer is three times the value of the Advertisement timer.
Enhanced timer	In enhanced mode, this timer is triggered if no packet is received from a neighbor when the neighbor aging timer expires. The value of the enhanced timer is 10s. After the enhanced timer is started, a DLDP-enabled device sends up to eight probe packets to the neighbor at a frequency of one probe packet per second.
DelayDown timer	If a DLDP-enabled device in Active, Advertisement, or Probe state detects a Port-Down event, it transits to the DelayDown state instead of deleting the neighbor entry and transiting to the Inactive state. At this time, the DLDP-enabled interface reserves DLDP neighbor information and only responds to Port-Up events. ● If the DLDP-enabled interface does not receive any Port-Up event when the DelayDown timer expires, it deletes the neighbor information and enters the Inactive state. ● If a DLDP-enabled device receives the Port-Up event before the DelayDown timer expires, it returns to the previous state.
RecoverProbe timer	The value of the RecoverProbe timer is 2s. That is, an interface in Disable state sends one RecoverProbe packet every two seconds to detect whether a unidirectional link is restored.

Operation Mode of DLDP

DLDP can work in normal mode or enhanced mode.

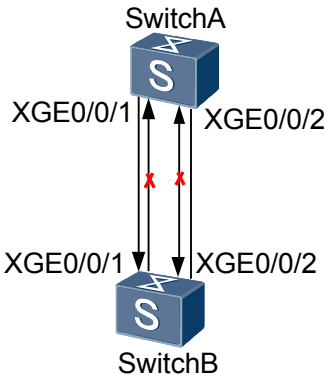
Table 1-3 DLDP timers

Operation Mode of DLDP	Detecting a Neighbor After an Neighbor Aging Timer Expires	Starting the Neighbor Aging Timer After an Neighbor Aging TimerExpires	Starting the Enhanced Timer After an Neighbor Aging Timer Expires
Normal mode	No	Yes (the neighbor entry is aged immediately after the neighbor aging timer expires)	No
Enhanced mode	Yes	Yes (the enhanced timer and Echo timer are started after the neighbor aging timer expires)	Yes (the status of the local link is set to unidirectional and the neighbor entry is deleted after the Echo timer expires)

If DLDP works in normal mode, the system can identify only unidirectional links caused by crossed connections of fibers. In this mode, the system cannot detect that a bidirectional link changes to a unidirectional link because the system does not start the enhanced timer.

If DLDP works in enhanced mode, the system can identify unidirectional links caused by crossed connections of fibers and disconnection of a fiber. To detect unidirectional links caused by disconnection of one fiber, you need to manually set the rate and full duplex mode of the interconnected interfaces; otherwise, DLDP does not take effect even if it is enabled. If a unidirectional link is caused by disconnection of fibers, the interface where the Rx fiber receives optical signals is in Disable state, and the interface where the Rx fiber receives no optical signals is in Inactive state.

Figure 1-3 Networking of DLDP in enhanced mode



Working Process of DLDP

If a link of the DLDP-enabled interface is in Up state, the DLDP-enabled interface sends DLDP packets to the peer device and processes the DLDP packets received from the peer device. DLDP packets sent vary with DLDP states.

Table 1-4 DLDP packet types

DLDP State	Types of DLDP Packets Sent
Active	Advertisement packets with RSY tags.
Advertisement	Common Advertisement packets.
Probe	Probe packets.
Disable	Disable packets. After one Disable packet is sent, a RecoverProbe packet is sent.

A received DLDP packet is processed as follows:

- In authentication modes, the DLDP packet is authenticated. The DLDP packet is discarded if it fails to pass the authentication.
- The DLDP packet is discarded if the interval for sending Advertisement packets in the DLDP packet is different from that on the local device.
- The DLDP packet is processed.

Table 1-5 Procedures for processing different types of DLDP packets

Packet Type	Procedure	
Advertisement packets	Retrieve the neighbor	Create a neighbor entry, trigger the neighbor aging timer, and transit to the Probe state if the neighbor entry does not exist.

Packet Type	Procedure			
with RSY tags	information.	Update the neighbor aging timer and transit to the Probe state if the neighbor entry exists.		
Common Advertisement packets	Retrieve the neighbor information.	Create a neighbor entry, trigger the neighbor aging timer, and transit to the Probe state if the neighbor entry does not exist.		
		Update the neighbor aging timer if the neighbor entry exists.		
Flush packets	Determine whether the local interface is in Disable state.	No action is required if the local interface is in Disable state.		
		Delete the neighbor information if the interface is not in Disable state and the neighbor information exists in the neighbor table.		
Probe packets	Retrieve the neighbor information.	Create a neighbor entry, transit to the Probe state, and return Echo packets to the peer end if the neighbor entry does not exist.		
		Update the neighbor aging timer and return Echo packets to the peer end if the neighbor entry does not exist.		
Echo packets	Retrieve the neighbor information.	Create a neighbor entry, trigger the neighbor aging timer, and transit to the Probe state if the neighbor entry does not exist.		
		Check whether the neighbor information in the DLDP packet is the same as that of the local device if the neighbor entry exists.	Discard the DLDP packet if the neighbor information in the DLDP packet is different from that of the local device.	

Packet Type	Procedure			
			Set the neighbor as bidirectionally connected if the neighbor information in the DLDP packet is the same as that of the local device. In addition, if all the neighbors are connected bidirectionally, DLDP transits from the Probe state to the Advertisement state and the Echo timer is disabled.	
Disable packets	Determine whether the local interface is in Disable state.	No action is required if the local interface is in Disable state.		
		The local interface enters the Disable state if the local interface is not in Disable state.		

Packet Type	Procedure	
RecoverProbe packets	Check whether the local interface is in Disable or Advertisement state.	No action is required if the local interface is not in Disable or Advertisement state.
		Return RecoverEcho packets if the local interface is in Disable or Advertisement state.
RecoverEcho packets	Determine whether the local interface is in Disable state.	No action is required if the local interface is not in Disable state.
		Check whether the neighbor information in the DLDP packet carries is the same as that on the local interface. If yes, the local interface transits to the Active state.
LinkDown packets	Check whether the local interface works in enhanced mode.	No action is required if the local interface does not work in enhanced mode.
		The local transits to the Disable state if the local interface works in enhanced mode and the local interface is not in Disable state.

If no echo packet is received from the neighbor, DLDP performs the following operations.

Table 1-6 Processing procedure when no echo packet is received from the neighbor

Mode	Procedure
In normal mode, no echo packet is received when the Echo timer expires.	The DLDP-enabled interface transits to the Disable state, and displays logs and information about traced packets and sends Flush packets. In addition, the local interface is manually or automatically shut down according to the DLDP Down mode and the neighbor entry is deleted.
In enhanced mode, no echo packet is received when the enhanced timer expires.	

Link Auto-recovery Mechanism

If the interface shutdown mode is set to **auto**, DLDP sets the interface where a unidirectional link is detected to DLDP Down automatically. A DLDP Down interface cannot forward service traffic or send/receive any protocol packets except DLDPDUs. A DLDP Down interface can be recovered before link recovery. The DLDP Down interface sends RecoverProbe packets

periodically. If correct RecoverEcho packets are received, it indicates that the unidirectional link changes to the bidirectional link and the DLDP Down interface becomes Up. The detailed process is described as follows:

The DLDP Down interface sends a RecoverProbe packet, which carries information about the local interface, every two seconds. When receiving the RecoverProbe packet, the peer end returns a RecoverEcho packet. When receiving the RecoverEcho packet, the local interface checks whether the neighbor information in the RecoverEcho packet is the same as that on the local interface. If they are the same, the link between the local interface and the neighbor is considered to have been restored. The local interface transits from Disable state to Active state and the neighbor relationship is set up. Only DLDP Down interfaces can send and process Recover packets, including RecoverProbe packets and RecoverEcho packets. The auto-recovery mechanism does not take effect on interfaces that are manually shut down.

2 Smart Link

About This Chapter

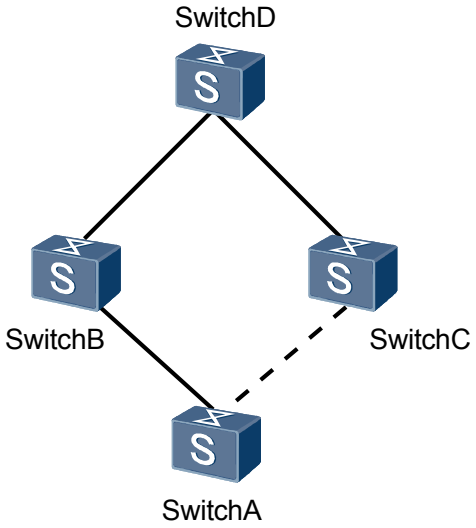
2.1 Introduction to Smart Link

Definition

Smart Link is introduced to provide efficient link backup and switchover for dual-homing networks. Compared with the Spanning Tree Protocol (STP), Smart Link provides faster convergence. Compared with the Rapid Ring Protection Protocol (RRPP), Smart Link provides simpler configuration methods.

Purpose

Figure 2-1 Ring network



As shown in **Figure 2-1**, link backup can be provided on the dual-uplink network, whereas a loop (SwitchA -> SwitchB ->SwitchD -> SwitchC -> SwitchA) result in broadcast storms. STP can be used to prevent loops, but the convergence speed is low. When the active link is faulty, traffic is switched to the standby link. During the switchover, a large amount of traffic is lost because the convergence takes several seconds. STP cannot be applied to the networks that require short convergence time. RRPP and SEP can improve the convergence performance, whereas they are applied to complicated ring networks and are difficult to configure.

To address the preceding problem, Huawei introduces Smart Link in dual-homing networks to implement link backup and fast transition. In this manner, the high performance is ensured and the configuration is simplified. In addition, Monitor Link is used to monitor the uplink interfaces, improving redundancy of Smart Link. Smart Link has the following advantages:

- Preventing broadcast storms caused by loops

  When two links are running properly on a dual-uplink network, only one link transmits traffic and the other link is blocked.
- Ensuring data forwarding

  When the active link is faulty, traffic is switched to the standby link in milliseconds.
- Simplifying configuration

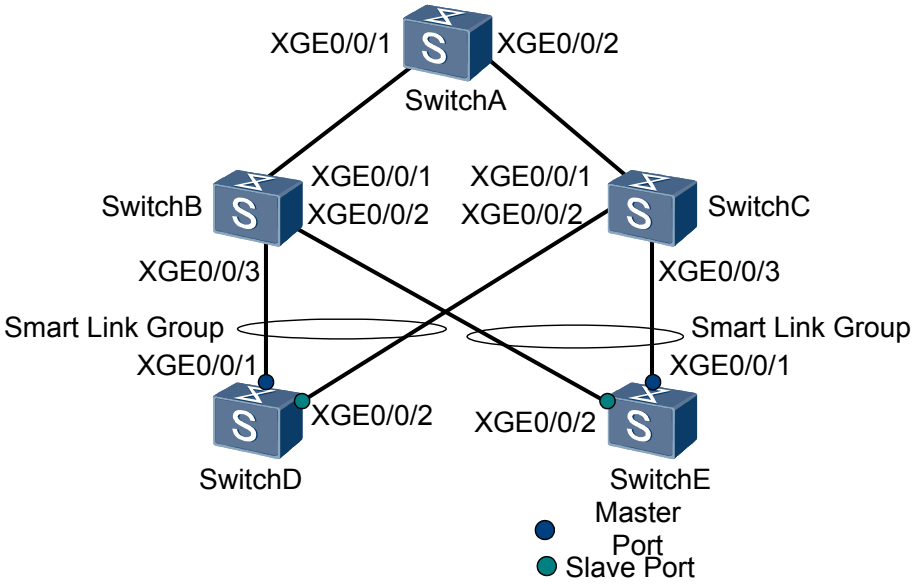
2.2 References

Smart Link is the private protocol of Huawei Technologies Co., Ltd.

2.3 Principles

2.3.1 Concepts of Smart Link

Figure 2-2 Networking diagram of Smart Link



Smart Link Instance

Similar to the MSTP instance, a Smart Link instance is bound to VLANs in a range; different instances are bound to VLANs of different ranges; the slave links of a Smart Link group are bound to different instances, implementing load balancing.

Smart Link Group

A Smart Link group consists of a maximum of two interfaces. One is the master interface, and the other is the slave interface. Normally, only one interface in the Smart Link group is active; the other interface is inactive. When the link of the master interface fails due to the physical fault, OAM connectivity fault, or unidirectional connectivity, the Smart Link group blocks this interface automatically and switches the originally inactive interface to active.

As shown in **Figure 2-2**, interfaces XGE 0/0/1 and XGE 0/0/2 on device SwitchD constitute a Smart Link group; interfaces XGE 0/0/1 and XGE 0/0/2 on device SwitchE constitute a Smart Link group.

Master Interface

Master interface is a role in a Smart Link group. If two interfaces in a Smart Link group are both inactive, the master interface turns active first. The master interface, however, is not always active. When the traffic is switched on the link, the slave interface becomes active and the master interface remains inactive even after the faulty link recovers. The master interface turns active until the next switching on the link. If revertive switching is configured, services are switched to the master link after the interval for revertive switching. In **Figure 2-2**, interface XGE 0/0/1 on device SwitchD is the master interface. Interface XGE 0/0/1 on device SwitchE is also the master interface even though it is blocked.

Slave Interface

Slave interface is a role in a Smart Link group. If two interfaces in a Smart Link group are both inactive, the slave interface remains inactive. The slave interface, however, is not always inactive. After the traffic is switched from the master link to the slave link, the slave interface becomes active. In **Figure 2-2**, interface XGE 0/0/2 on device SwitchD is the master interface. Interface XGE 0/0/2 on device SwitchE is also the master interface even though it is blocked.

Flush Packet

When switching is performed between the links of a Smart Link group, the original forwarding entries no longer apply to the new network topology. The MAC address entries and Address Resolution Protocol (ARP) entries on the whole network need to be updated. In this case, the Smart Link group informs other devices of link switching by sending Flush packets so that other devices can update the address table. The format of Flush packets is as follows:

NOTE

Flush packets are used to update MAC entries and ARP entries based on interfaces. Only the interfaces that are allowed to accept Flush packets can update MAC entries and ARP entries according to Flush packets.

Figure 2-3 Format of Flush packets

DMAC = 010F-E200-0004 (6 bytes)
Source MAC Address (6 bytes)
Length (1 byte)
DSAP = 0xaa (1 byte)
SSAP = 0xaa (1 byte)
Control Field = 0x03 (1 byte)
Organization Code = 0x000fe2 (3 bytes)
PID = 0x0127 (2 bytes)
Control Type = 0x01 (1 byte)
Control Version = 0x00 (1 byte)
Device ID (6 bytes)
Control VLAN ID (2 bytes)
Auth-mode (1 bytes)
Password (16 bytes)
VLAN Bitmap (512 bytes)
FCS (4 bytes)

IEEE802.3 encapsulation is used for Flush packets:

- DMAC indicates the unknown multicast address that can be used to differentiate protocols.
- Control Type indicates the control type. Currently, the field has only one value, 0x01, which means clearing MAC addresses.
- Control Version indicates the version number of Flush packets.
- Device ID indicates the bridge MAC address of the site.
- Control VLAN ID indicates the ID of the control VLAN.
- Auth-mode indicates the authentication mode, which is used with the password. It is used for future security extension.
- VLAN Bitmap indicates the VLAN bitmap, which carries the VLAN list in which addresses need to be updated.
- FCS indicates Frame Check Sequence, which is used to check the validity of packets.

Load Balancing

When the master and slave links of a Smart Link group work normally, Smart Link allows these two links to forward different data traffic. In load balancing mode, the two interfaces are active.

Note that the slave interface forwards load-balancing instance traffic and the master interface forwards non-load-balancing instance traffic. When a link fails due to physical fault, OAM connectivity fault, or unidirectional connectivity, the Smart Link group switches all the traffic to the other link automatically.

2.3.2 Working Mechanism of Smart Link

This section describes the working mechanism of Smart Link based on the network shown in **Figure 2-2**.

On SwitchD, XGE 0/0/1 functions as the master interface and XGE 0/0/2 functions as the slave interface. XGE 0/0/1 is active, whereas XGE 0/0/2 is inactive. When the link connected to XGE 0/0/1 fails, XGE 0/0/1 becomes inactive and XGE 0/0/2 becomes active.

When link switching occurs in a Smart Link group, MAC address entries and ARP entries stored in the devices on the network may be incorrect. A new mechanism is required to update MAC address entries and ARP entries.Devices enabled with Smart Link send Flush packets from a new link.

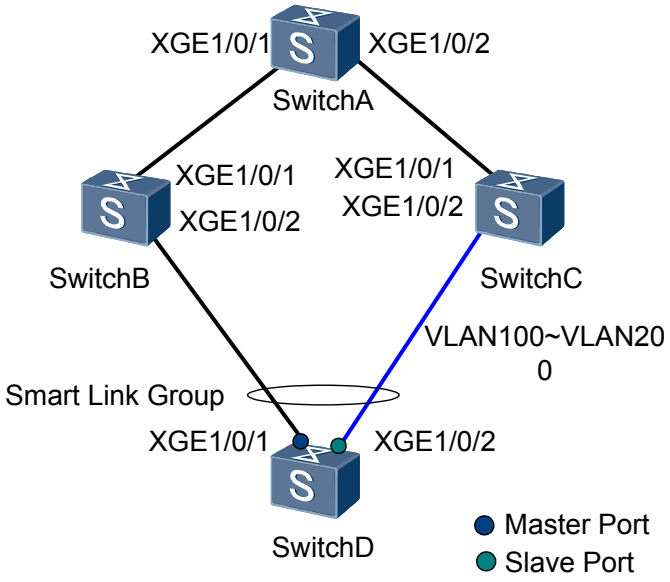
There are two ways to update MAC address entries and ARP entries:

- After a switchover, the two ends of a link sends traffic to each other. The device automatically updates the MAC address entries and ARP entries. This way is effective when the is connected to a non-Huawei device.
- The uplink device is capable of identifying the Flush packets of Smart Link. After a switchover, the device sends Flush packets to update the MAC address entries and ARP entries.

When the faulty active link recovers, it remains blocked, and thus traffic is not switched back to the recovered link to ensure stability.

Smart Link Load Balancing

Figure 2-4 Network diagram of Smart Link load balancing



In the implementation of the traditional Smart Link technology, the inactive link cannot carry any service data. Therefore, the bandwidth usage is low. Even if there are two uplink 10GE links, the actual bandwidth is only 10 Gbit/s. In addition, certain bandwidth is reserved during network planning. Assume that 50% bandwidth is reserved. The actual bandwidth is only 5 Gbit/s. Even if there are two uplink 10GE links, the actual bandwidth usage is only 25%.

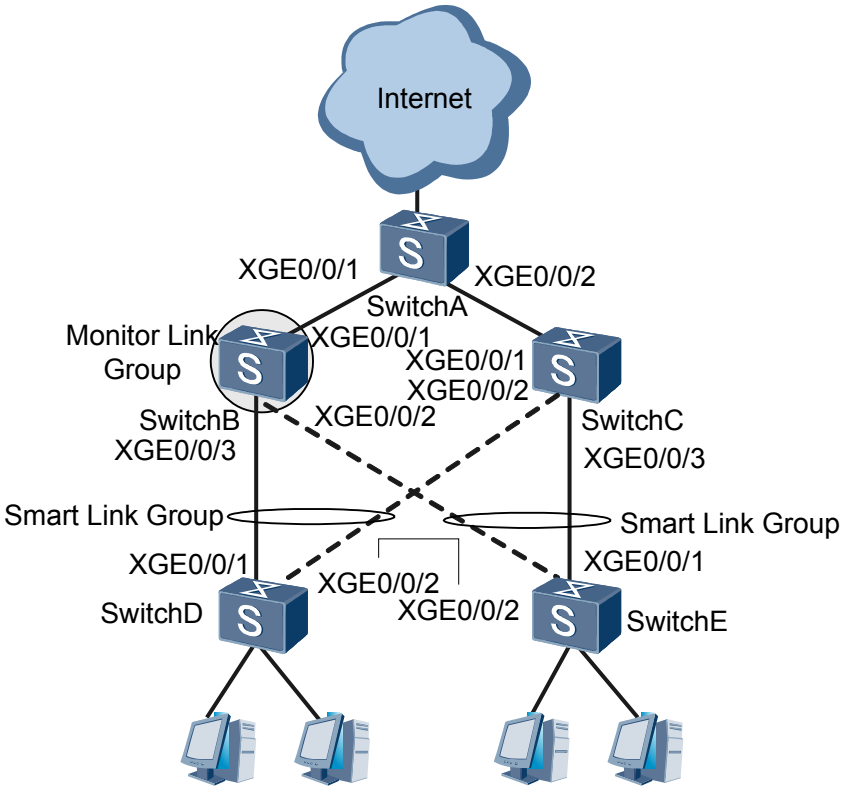
As shown in **Figure 2-4**, all the packets are transmitted on the active link through XGE 0/0/1 on SwitchD. To improve the bandwidth usage, you can configure Smart Link load balancing so that some VLAN data is transmitted through XGE 0/0/2 on SwitchD. You can configure VLAN 100 to VLAN 200 in an instance to load balance traffic of the instance between links in the Smart Link group.

2.3.3 Concepts of Monitor Link

As an interface association mechanism introduced to complement Smart Link, Monitor Link perfects link backup of Smart Link. By synchronizing the downlink with the uplink, Monitor Link quickly notifies the downstream device of uplink failures and thus triggers the switchover between the active link and the standby link, preventing packet loss caused by unannounced uplink failures.

As shown in **Figure 2-5**, the S6700 enabled with Smart Link can rapidly switch the traffic to the slave interface when the master interface fails. The duration of service interruption is thus shortened. If the uplink connected to the master interface fails, services are also interrupted. Therefore, the uplink needs to be monitored so that the downlink can detect the fault and switch the traffic.

Figure 2-5 Application scenario of Monitor Link



As shown in **Figure 2-5**, the uplink of SwitchB is faulty. Although Smart Link is enabled on SwitchD, an switchover is not performed because SwitchD does not detect the fault on the uplink. In this case, services are thus interrupted. To enable the Smart Link group to respond to the faults of the uplink quickly, you need to configure the Monitor Link function on SwitchB connected to the active link to monitor the status of the uplink. When a fault occurs on an uplink, the active link of the Smart Link group is rapidly blocked. Thus, the Smart Link group can detect the fault and switch the traffic to the standby link to shorten the service interruption duration.

Monitor Link Group

A Monitor Link group consists of uplink and downlink interfaces. In a Monitor Link group, there is one uplink interface and several downlink interfaces. A member of the Monitor Link group can be a single interface, an interface in a static aggregation group, an interface in a manual aggregation group, or an interface in a Smart Link group. Only the interface in a Smart Link group can function as the uplink interface. The status of the downlink interfaces varies with the status of the uplink interface.

As shown in **Figure 2-5**, XGE 0/0/1 on SwitchB, XGE 0/0/2, and XGE 0/0/3 constitute a Monitor Link group.

Uplink Interface

An uplink interface is monitored by the downlink interfaces in a Monitor Link group. If the uplink interface fails, the Monitor Link group is faulty and all the downlink interfaces in the Monitor Link group are forcibly shut down.

When the uplink interface is a Smart Link group, the uplink interface is considered as faulty only if the maser and slave interfaces of the Smart Link group are in Inactive or Down state.

As shown in **Figure 2-5**, XGE 0/0/1 on SwitchB is the uplink interface.

Downlink Interface

Downlink interfaces monitor the uplink interface in a Monitor Link group. The fault of a downlink interface does not affect the uplink interface or the other downlink interfaces.

As shown in **Figure 2-5**, XGE 0/0/2 on SwitchB and XGE 0/0/3 are downlink interfaces.

2.3.4 Operation Mechanism of Monitor Link

After a Monitor Link group is configured, the uplink interface is monitored in real time. When the uplink interface fails such as the fault on a link, unidirectional OAM connectivity, or failure to establish OAM connections, all the downlink interfaces in Up state in the Monitor Link group are forcibly shut down. When the uplink interface recovers, the downlink interfaces recover.

If the uplink interface is the interface in a Smart Link group, the uplink interface is considered as faulty only when the two interfaces of the Smart Link group are in Inactive state or Down state.

When downlink interfaces are the interfaces in an aggregation group, all the interfaces in the aggregation group are forcibly shut down if the uplink interface fails. When the uplink interface recovers, all the downlink interfaces in the aggregation group recover.

2.4 Applications

Figure 2-6 Networking diagram of Smart Link

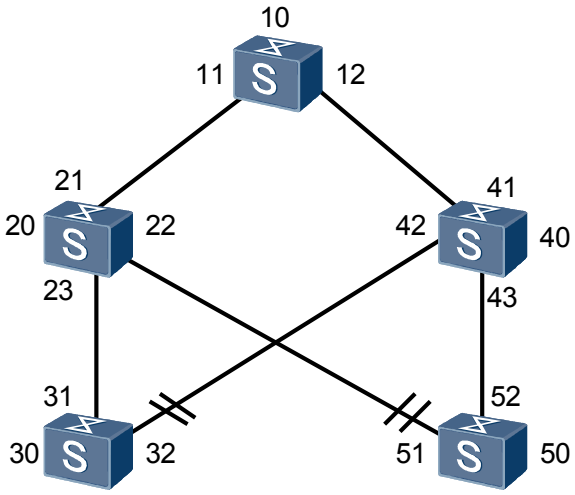
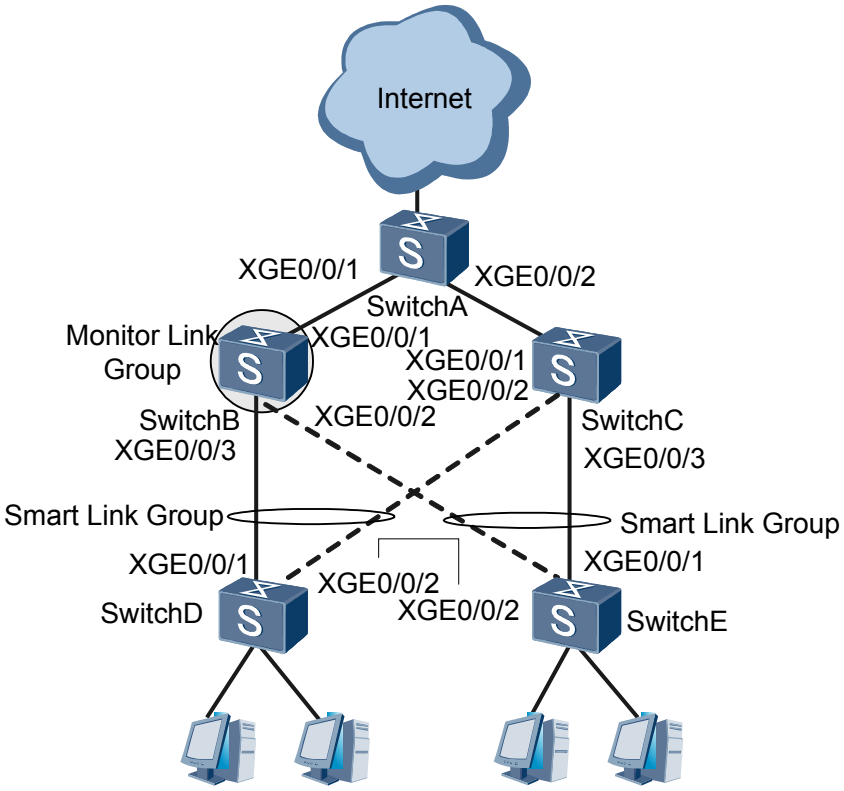


Figure 2-6 shows a typical dual-homing networking where Smart Link and Monitor Link are used. In the networking, Smart Link is configured on devices 30 and 50. One of the dual uplinks is blocked, and the other is in active state. When the active link fails, Smart Link can detect the fault rapidly and <keyword conref="..\commonterms\commonterms.xml#commonterms/productgroup"></keyword> the traffic. The fault of the link near device 10 cannot be directly detected by devices 30 and 50 because the diameter of the network is great. Thus, Monitor Link needs to be configured on devices 20 and 40. Interfaces 21 and 41 function as uplink interfaces, interfaces 22 and 23 and interfaces 42 and 43 function as downlink interfaces. When the fault of the uplink interface is detected, the downlink interfaces are automatically shut down. When the uplink interface recovers, the downlink interfaces recover accordingly. In this manner, devices 30 and 50 can detect the link change near device 10 rapidly.

2.4.1 Combination of Smart Link and Monitor Link

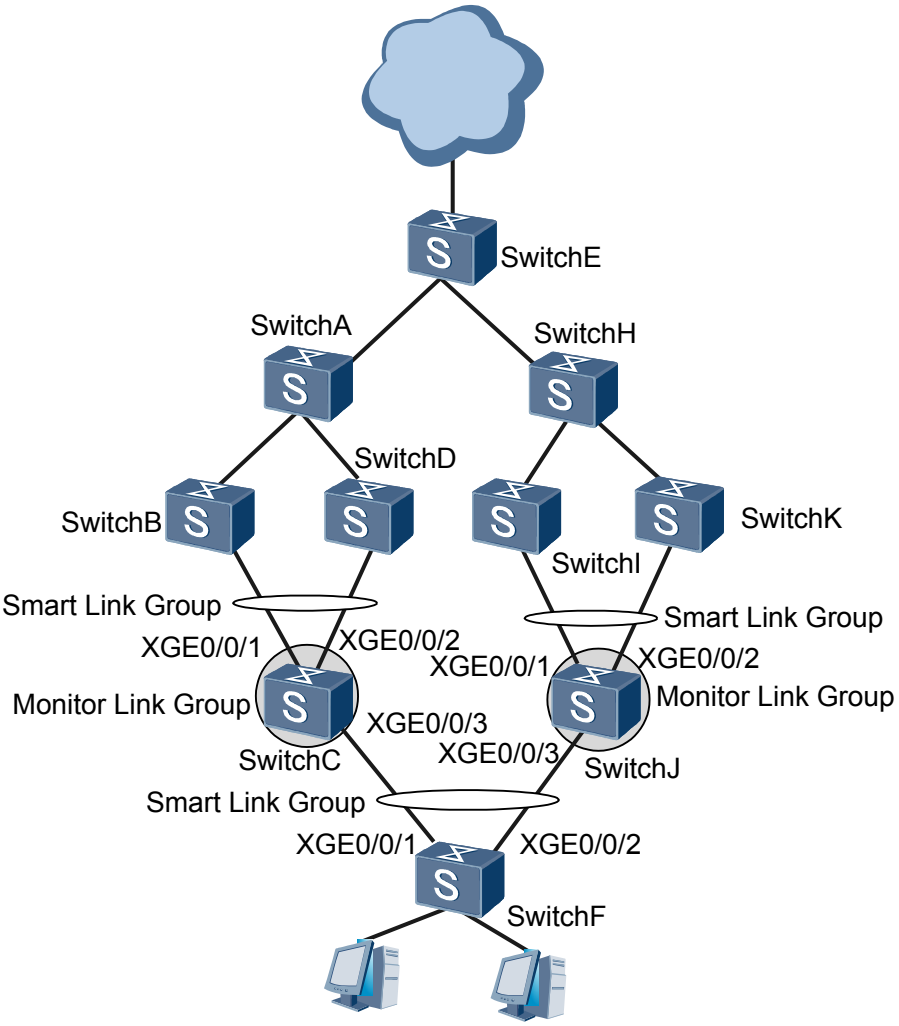
Figure 2-7 Network where Smart Link and Monitor Link are used together



As shown in **Figure 2-7**, Smart Link is configured on SwitchD and SwitchE. When the link between SwitchB and SwitchD or between SwitchC and SwitchE is faulty, the Smart Link group can detect the fault and switch services to the standby link quickly. To enable SwitchD or SwitchE to detect the fault on the link between SwitchA and SwitchB or between SwitchA and SwitchC, you need to configure a Monitor Link group on SwitchB or SwitchC. XGE 0/0/1 functions as the uplink interface, and XGE 0/0/2 and XGE 0/0/3 function as downlink interfaces. When the Monitor Link group detects a fault on the uplink, downlink interfaces are shut down forcibly. An active/standby switchover is then triggered in the Smart Link groups on SwitchD and SwitchE. When the fault on the uplink interface is rectified, downlink interfaces are enabled automatically. In this way, SwitchD or SwitchE can detect the uplink status change.

2.4.2 Cascading of Smart Link and Monitor Link

Figure 2-8 Cascading of Smart Link and Monitor Link



Cascading of Smart Link and Monitor Link improves link reliability. Smart Link groups can function as uplink member interfaces of a Monitor Link group. As shown in **Figure 2-8**, the uplink interfaces on SwitchC and SwitchJ are added to Smart Link groups to guarantee reliability. Monitor Link groups are configured on SwitchC and SwitchJ and the Smart Link groups on the Switches added to the Monitor Link groups. An active/standby switchover is performed between the downlink interfaces only when both the two uplink interfaces in a Smart Link group become Down.

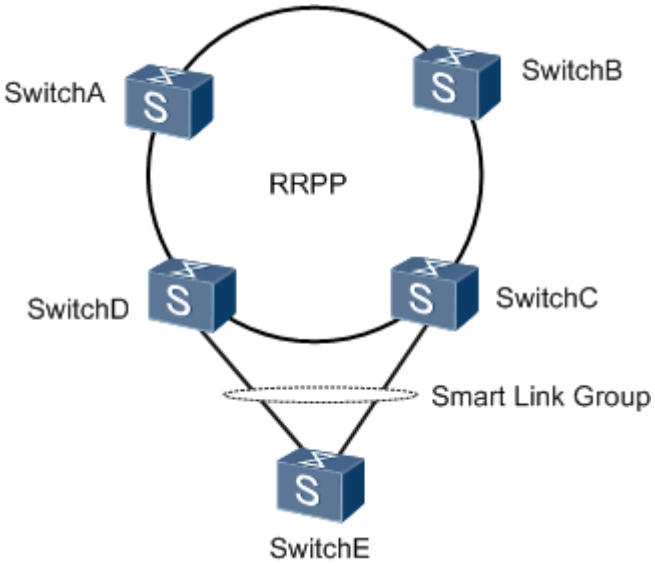
Table 2-1 lists the configuration of Smart Link and Monitor Link.

Table 2-1 Configuration of Smart Link and Monitor Link

Switch	Smart Link Group 1		Monitor Link Group 1	
	Master Interface	Slave Interface	Uplink Interface	Downlink Interface
SwitchC	XGE 0/0/1	XGE 0/0/2	Smart Link Group 1 XGE 0/0/1 and XGE 0/0/2	XGE 0/0/3
SwitchJ	XGE 0/0/1	XGE 0/0/2	Smart Link Group 1 XGE 0/0/1 and XGE 0/0/2	XGE 0/0/3
SwitchF	XGE 0/0/1	XGE 0/0/2	The Monitor Link group is not created.	

2.4.3 Combination of Smart Link and RRPP

Figure 2-9 Network where Smart Link and RRPP are used together



As shown in **Figure 2-9**, RRPP is configured on SwitchA, SwitchB, SwitchC, and SwitchD to provide link backup. To use STP to implement link backup, you must enable STP on connected interfaces between SwitchC, SwitchD, and SwitchE. RRPP is enabled on the two interfaces connecting SwitchC and SwitchD; therefore, STP cannot be enabled. In this case, you can configure a Smart Link group on SwitchE to implement link backup. Such configuration is easier than the configuration of RRPP subrings on SwitchC, SwitchD, and SwitchE. This method can also be used when SwitchE does not support RRPP.

2.5 Terms and Abbreviations

Abbreviation

Abbreviation	Full Spelling
STP	Spanning Tree Protocol
RRPP	Rapid Ring Protection Protocol
SMLK	Smart Link
MTLK	Monitor Link

3 RRPP

About This Chapter

3.1 Introduction

Definition

The Rapid Ring Protection Protocol (RRPP) is a link layer protocol specially used to prevent loops on an Ethernet ring network. Devices running RRPP discover loops on the network by exchanging RRPP packets with one another, and eliminate loops by blocking certain interfaces.

Purpose

For most MANs and LANs, the ring network is adopted to provide high reliability. The fault of any node on the ring, however, affects the service.

When a device or link is faulty, it takes a period of time for data switches to a backup device or link. To reduce convergence time and remove the impact of network scale on convergence time, Huawei develops RRPP. Compared with other Ethernet ring technologies, RRPP boasts following features as shown in **Table 3-1**:

- Convergence time is irrelevant to the number of nodes on the ring network. Thus, RRPP can be applied to a network with a great diameter.
- RRPP can prevent broadcast storm caused by loops when an Ethernet ring network is complete.
- On an Ethernet ring network, when a link is torn down, a backup link immediately starts to resume the normal communication between nodes.
- The cost is low.

Table 3-1 Comparison of ring network protocols

Ring Network Rrotocol	Description
Token Ring	Based on the MAC layer protocol, the token ring is the first ring technology that is introduced to the data communication field. The token ring is a single-direction ring of low speed and is used in Local Area Networks (LANs). The token ring does not have self-healing capability.
FDDI	As an enhancement of the token ring, the Fiber Distributed Digital Interface (FDDI) uses a token to transmit the right to control a ring network. Different from the token ring, FDDI adopts the double-ring structure. In addition, FDDI uses fibers as the transmission media, which greatly improves the performance and efficiency compared with the token ring. Same as the token ring, FDDI does not have self-healing capability. The bandwidth of an FDDI ring network cannot be efficiently utilized because of the adoption of source address stripping.

Ring Network Rrotocol	Description
SDH/SONET	Synchronous Digital Hierarchy/Synchronous Optical Network (SDH/SONET) is a widely applied ring technology that supports both single ring and double rings. SDH/SONET features high reliability and provides Automatic Protection Switching (APS), which is a mechanism of automatic fault recovery. On an SDH/SONET, the bandwidth between two nodes is fixed, which is determined by the point-to-point (P2P) structure and the circuit switching design. The bandwidth cannot adapt itself to the ever changing situation, which impedes the efficient utilization of bandwidth. Thus, the SDH/SONET technology cannot fully meet the bandwidth requirements of burst IP traffic. In addition, broadcast packets and multicast packets on the SDH/SONET are transmitted as unicast packets. The bandwidth is thus severely wasted. The APS feature requires a maximum of 50 % redundancy bandwidth, which makes a flexible selection mechanism impossible.
RPR	Designed and standardized by IEEE 802.17 and RPR confederation, Resilient Packet Ring (RPR) is a MAC layer protocol running on ring networks. RPR designs a logical P2P closed topology based on the MAC layer. For the physical layer, an RPR network is a ring network that consists of P2P links; for the data link layer, an RPR network is a broadcast network similar to an Ethernet network. RPR is realized through special hardware. The fair algorithm of RPR is complex.
Spanning Tree Protocol (STP)/ Rapid Spanning Tree Protocol (RSTP)/Multiple Spanning Tree Protocol (MSTP)	STP (Spanning Tree Protocol)/RSTP (Rapid Spanning Tree Protocol)/ MSTP (Multi-Spanning Tree Protocol) On an MSTP network, a loop-free tree is formed. Thus, broadcast storm is eliminated and backup is implemented. Multiple spanning trees carry out load balancing among VLANs. In this case, traffic of different VLANs is transmitted along different paths. STP/RSTP/MSTP is a protocol with the automatic calculation function and supports any topology. The convergence time is affected by the network topology.
RRPP	RRPP(Rapid Ring Protection Protocol) RRPP features fast convergence. Convergence time is irrelevant to the number of nodes on a ring network. RRPP multi-instance supports load balancing of different kinds of service traffic. RRPP is a Huawei private protocol and only takes effect on Huawei devices. Therefore, RRPP can be enabled on a network deployed with pure Huawei devices.

3.2 References

None

3.3 Principle Description

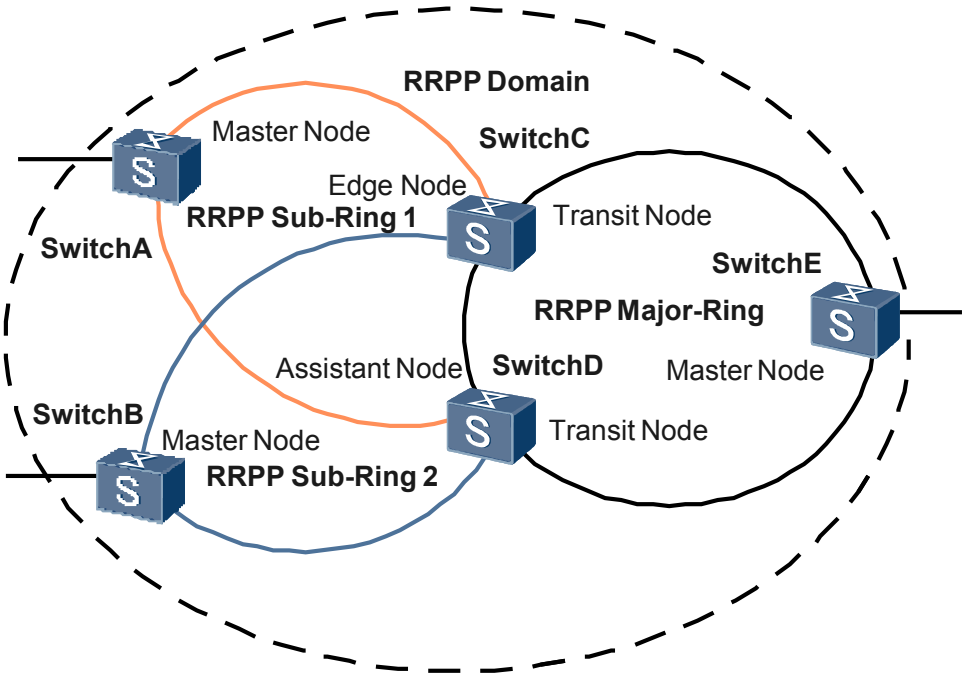
3.3.1 Basic Concepts of RRPP

Devices on the Ethernet can be configured to serve as different nodes on the RRPP ring. Nodes on the RRPP ring exchange and process RRPP packets to detect the status of the ring network and transmit the topology change of the ring network. The master node on the ring blocks or unblocks the secondary port according to the status of the ring network. In this case, when a device or link on the ring network is faulty, the backup link can be immediately started to eliminate loops.

Network architecture of RRPP

As shown in **Figure 3-1**, an RRPP domain consists of the following elements.

Figure 3-1 Schematic diagram of an RRPP domain



- RRPP domain

  An RRPP domain is identified uniquely with the ID that is an integer.

  An RRPP domain consists of a group of switches that are mutually connected and configured with the same domain ID and control VLAN.

Network elements that form an RRPP domain are as follows:

- RRPP major ring
- RRPP sub-ring
- Control VLAN
- Master node
- Transit node
- Edge node
- Assistant edge node
- Common port
- Edge port
- Primary port
- Secondary port

- RRPP ring

  Physically, an RRPP ring corresponds to an Ethernet ring topology. Each RRPP ring is a part of its RRPP domain.

  An RRPP domain can be composed of a single RRPP ring or the combination of a major ring and multiple sub-rings.

  Protocol packets of sub-rings are transmitted in the major ring as data packets; protocol packets of the major ring are transmitted only in the major ring.

  NOTE

  An RRPP domain can have only one RRPP major ring.

- Control VLAN

  The control VLAN is a concept relative to the data VLAN. In an RRPP domain, the control VLAN is used to transmit only RRPP packets.

  Different from the control VLAN, the data VLAN is used to transmit data packets. The data VLAN can contain both the RRPP port and non-RRPP port.

  Each RRPP domain is configured with two control VLANs, namely, the major control VLAN and sub-control VLAN. During the configuration, only the major control VLAN needs to be specified. The VLAN whose ID is greater than the ID of the major control VLAN by 1 is the sub-control VLAN.

  Protocol packets of the major ring are transmitted in the major control VLAN; protocol packets of the sub-ring are transmitted in the sub-control VLAN. Interfaces of both the major control VLAN and the sub-control VLAN cannot be configured with VLANIF interfaces.

  On a switch, the port connected to the RRPP ring network belongs to the control VLAN.

- Node type

  Each device on the Ethernet ring is a node. Nodes on the RRPP ring are classified into following types:

  - Master node

    The master node determines how to handle topology changes. Each RRPP ring must have and only have one master node.

    Any device on the Ethernet ring can serve as the master node.

    The status of the master node can be either Complete or Failed.

When all links on the ring network are in the Up state and the master node can receive Hello packets sent by itself from the secondary port, the master node is in the Complete state.

The status of the master node is the status of the RRPP ring. Thus, the RRPP ring is also in the Complete state. In this case, the master node blocks the secondary port to prevent data packets from forming broadcast loops on the ring topology. After being blocked, the secondary can only receive RRPP protocol packets rather than transmit data packets.

When a link on the ring network is in the Down state, the master node is in the Failed state. In this situation, the master node unblocks the secondary port to ensure the uninterrupted communication of nodes on the ring network.

- Transit node

  On an RRPP ring, all nodes except the master node are transit nodes. The transit node monitors the status of its directly connected RRPP link and notifies the link status change to the master node for processing.

  The status of the transit node can be Link-Up, Link-Down, or Preforwarding.

  When both the primary port and secondary port of a transit node are in the Up state, the transit node is in the Link-Up state.

  When the primary port or secondary port of a transit node is in the Down state, the transit node is in the Link-Down state.

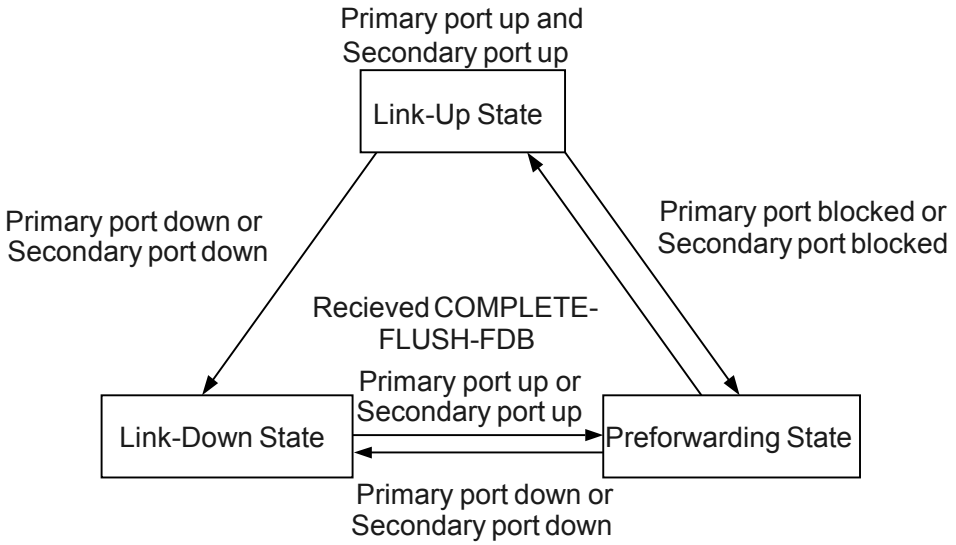
  When the primary port or secondary port of a transit node is in the blocked state, the transit node is in the Preforwarding state.

  As shown in **Figure 3-2**, when the transit node in the Link-Up state detects that the link of the primary port or secondary port turns Down, the transit node switches to the Link-Down state and sends a Link-Down packet to notify the master node.

  The transit node never directly switches back to the Link-Up state from the Link-Down state. When the link on a port of the transit node in the Link-Down state turns Up and the primary port and secondary port return to the Up state, the transit node switches to the Preforwarding state and blocks the recovered port. When the primary and secondary ports go Up, the master node does not immediately detect the change and the secondary port thus remains unblocked. If the transit node immediately switches back to the Link-Up state, broadcast loops formed by data packets occur on the ring network. Therefore, the transit node first enters Preforwarding from the Link-Down state.

  When an interface on the transit node in the Preforwarding state goes Down, the transit node enters the Link-Down state. When an interface on the transit node in the Preforwarding state goes Up and the transit node receives a COMPLETE-FLUSH-FDB packet from the master node, the transit node enters the Link-Up state. If the COMPLETE-FLUSH-FDB packet is lost during the transmission, RRPP provides a backup mechanism to unblock temporarily blocked ports and trigger the state transition. That is, the transit node automatically goes Link-Up and the temporarily blocked port is unblocked.

Figure 3-2 Diagram of state transition of transit node



- Edge node and assistant edge node

  A switch serves as an edge node and an assistant edge node on the sub-ring, and serves as a transit node on the major ring.

  - Edge node

    On an RRPP sub-ring, either of the two nodes crossed with the major ring can be specified as the edge node.

    On one sub-ring, there must be only one edge node.

  - Assistant edge node

    On an RRPP sub-ring, if one of the two nodes crossed with the major ring is specified as the edge node, the other node is the assistant edge node.

    On one sub-ring, there must be only one assistant edge node.

  Edge nodes and assistant edge nodes are special transit nodes. Thus, they have the same three states as transit nodes, but with different meanings.

  If an edge port is in the Up state, it indicates that the edge node or assistant edge node is in the Link-Up state.

  If an edge port is in the Down state, it indicates that the edge node or assistant edge node is in the Link-Down state.

  If an edge port is blocked, it indicates that the edge node or assistant edge node is in the Pre-forwarding state.

  The state transition of an edge node or assistant edge node is similar to that of a transit node. The difference is that if the state transition is caused by the changes of the link status on the port, only the status of an edge port changes.

- Port role

  - Primary port and secondary port

    On both the master node and transit node, one of the two ports connected to the Ethernet ring is the primary port, and the other is the secondary port. You can specify the role of the ports.

    The primary port and secondary port of the master node provide different functions.

The master node sends a Hello packet from the primary port. If the secondary port can receive this packet, it indicates that the RRPP ring of the node is complete. So, the master node needs to block the secondary port to prevent the data loop.

On the contrary, if the packet is not received within the specified period, it indicates that the RRPP ring is faulty. In this case, the master node needs to unblock the secondary port to guarantee normal communication between nodes on the ring.

If the secondary port on the master node of the major ring is blocked, not only the data packets but also the protocol packets of the sub-rings are prevented from passing through the port. When the secondary port is unblocked, both the data packets and the protocol packets of the sub-rings are permitted to pass through the port. That is, in the major ring, protocol packets of the sub-rings are processed as data packets.

The primary port and secondary port of the transit node provide the same function.

- Common port and edge port

  On an edge node or an assistant edge node, the port shared by the sub-ring and major ring is called the common port. Only the port on the sub-ring is called the edge port.

  A common port is regarded as the port on the major ring and belongs to both the major control VLAN and the sub-control VLAN. The RRPP port on the sub-ring only belongs to the sub-control VLAN. The major ring is regarded as a logical node of the sub-ring; thus the packets of the sub-ring are transparently transmitted through the major ring. The packets of the major ring, however, are transmitted only in the major ring.

RRPP Packets

Table 3-2 lists different types of RRPP packets.

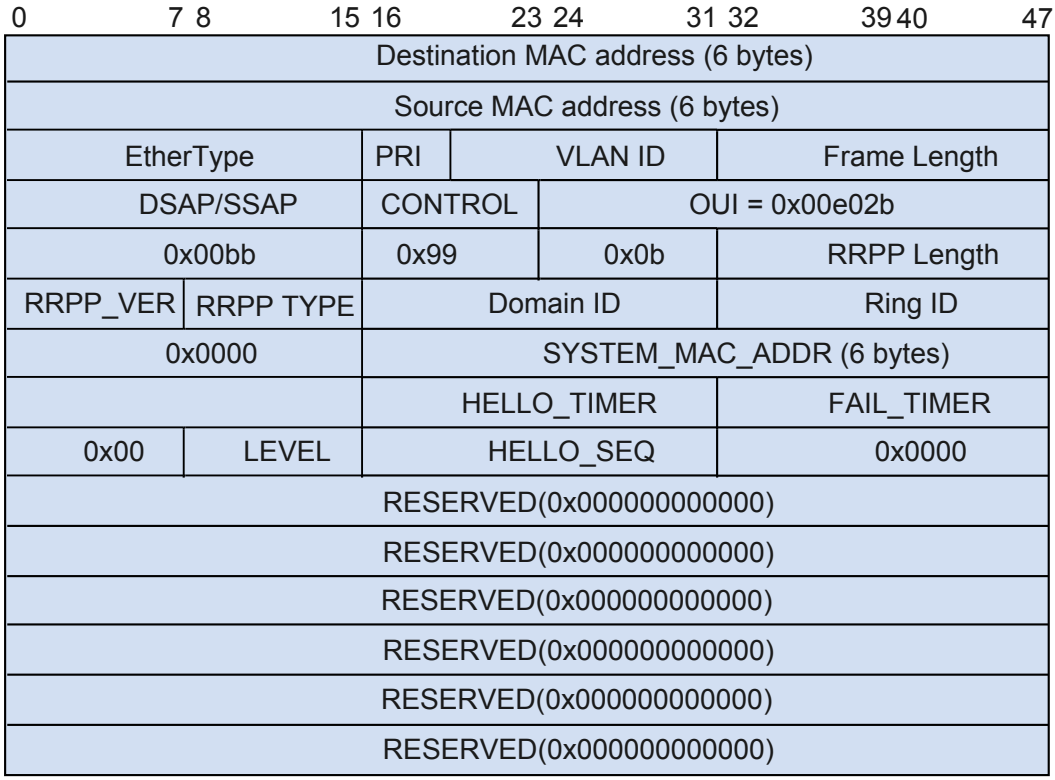
Table 3-2 Types of RRPP packets

Packet Type	Description
HEALTH (HELLO)	A packet sent from the master node to detect whether a loop exists on a network.
LINK-DOWN	A packet sent from a transit node, edge node, or assistant edge node to notify the master node that a port goes Down and the loop disappears.
COMMON-FLUSH-FDB	A packet sent from the master node to notify the transit nodes to refresh their MAC address forwarding tables and ARP entries.
COMPLETE-FLUSH-FDB	A packet sent from the master node to notify the transit node, edge node, or assistant edge node to update its MAC address forwarding table and ARP entries. In addition, it can notify the transit node to unblock the temporarily blocked ports.
EDGE-HELLO	A packet sent from an edge port of a sub-ring and received by an assistant edge port on the same sub-ring. The packet is used to check the completeness of the major ring in the domain where the sub-ring is located.

Packet Type	Description
MAJOR-FAULT	A packet sent from an assistant edge node to notify the edge node that the major ring in the domain fails if the assistant edge node does not receive the Edge-Hello packet from the edge port within a specified period.

Figure 3-3 illustrates the format of an RRPP packet.

Figure 3-3 Format of an RRPP packet



The following describes the fields in the packet:

- Destination MAC Address: indicates the destination MAC address of the packet. The field occupies 48 bits.
- Source MAC Address: indicates the source MAC address of the packet. The fixed value is the MAC address of the device.
- EtherType: indicates the encapsulation type. The fixed value is 0x8100, which indicates the tagged encapsulation. The field occupies 16 bits.
- PRI: indicates the priority of Class of Service (COS). The fixed value is 0xe. The field occupies 4 bits.
- VLAN ID: indicates the ID of the VLAN to which the packet belongs. The field occupies 12 bits.

- Frame Length: indicates the length of the Ethernet frame. The fixed value is 0x0048. The field occupies 16 bits.
- DSAP/SSAP: indicates the destination service access point/source service access point. The fixed value is 0xaaaa. The field occupies 16 bits.
- CONTROL: The field occupies 8 bits but has no actual significance. The fixed value is 0x03.
- OUI: The field occupies 24 bits but has no actual significance. The fixed value is 0x00e02b.
- RRPP_LENGTH: indicates the length of the RRPP data unit. The fixed value is 0x0040. The field occupies 16 bits.
- RRPP_VERS: indicates the version of the RRPP packet. The current version is 0x01. The field occupies 8 bits.
- RRPP TYPE: indicates the type of the RRPP packet. The field occupies 8 bits.
  - HEALTH = 0x05
  - COMPLETE-FLUSH-FDB = 0x06
  - COMMON-FLUSH-FDB = 0x07
  - LINK-DOWN = 0x08
  - EDGE-HELLO = 0x0a
  - MAJOR-FAULT= 0x0b
- DOMAIN_ID: indicates the ID of the RRPP domain to which the packet belongs. The field occupies 16 bits.
- RING_ID: indicates the ID of the RRPP ring to which the packet belongs. The field occupies 16 bits.
- SYSTEM_MAC_ADDR: indicates the bridge MAC address from where the packet is sent. The field occupies 48 bits.
- HELLO_TIMER: indicates the timeout period of the Hello timer on the node that sends the packet, in seconds. The field occupies 16 bits.
- FAIL_TIMER: indicates the timeout period of the Fail timer on the node that sends the packet, in seconds. The field occupies 16 bits.
- LEVEL: indicates the level of the RRPP ring to which the packet belongs. The field occupies 8 bits.
- HELLO_SEQ: indicates the sequence number of the Hello packet. The field occupies 16 bits.

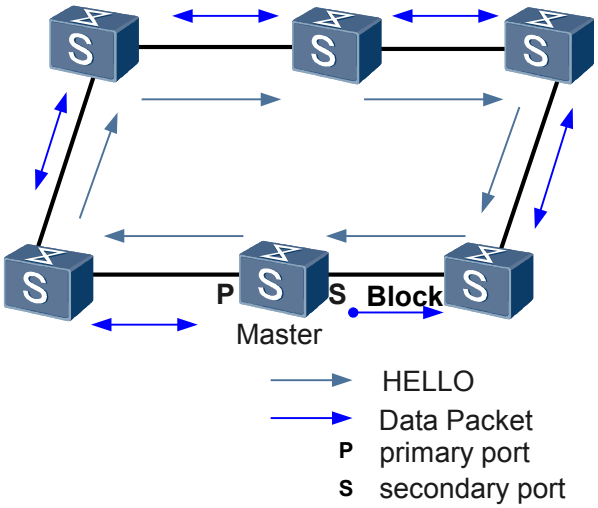
3.3.2 RRPP Operation Principles

Operation Principles of a Single RRPP Ring

The following takes a single RRPP ring as an example to describe the RRPP operation and topology convergence when the ring status changes from healthy to faulty and then to healthy.

1. The ring is in the healthy state.

Figure 3-4 RRPP ring in the Complete state

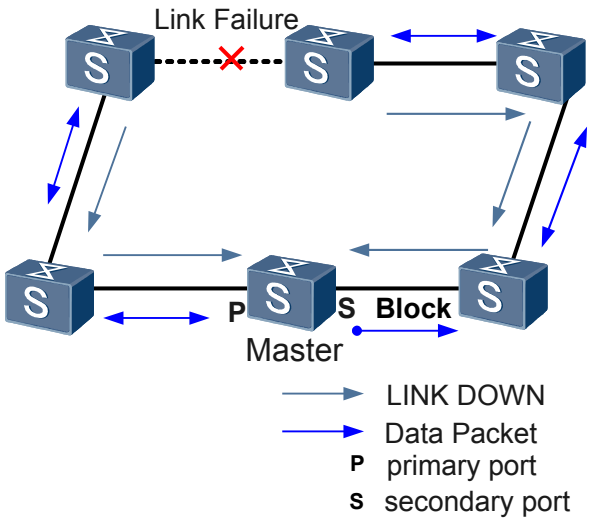


As shown in **Figure 3-4**,

- if all links on the ring are in the Up state, the RRPP ring is in the healthy state. The status of the master node reflects the health status of the ring.
- When the ring is in the healthy state, the master node blocks its secondary port to prevent the broadcast loop formed by data packets.
- The master node periodically sends the Hello packet from the primary port. The Hello packet is transmitted through all transit nodes and reaches the secondary port of the master node.

2. The link is faulty.

Figure 3-5 Schematic diagram of reporting link fault from the transit node

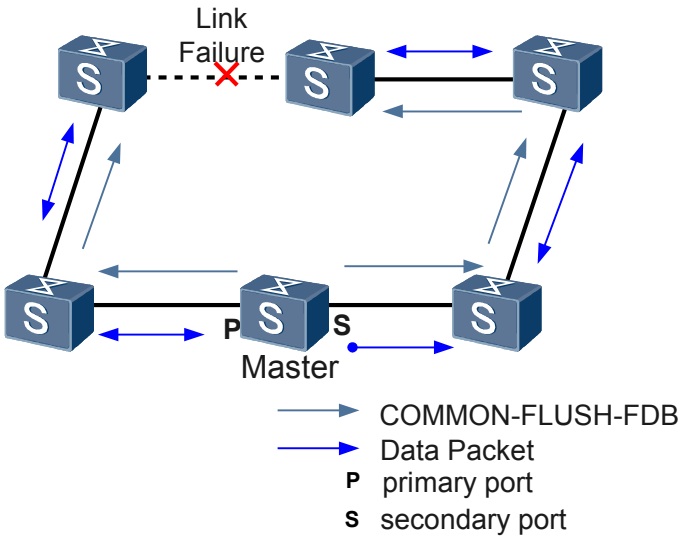


As shown in **Figure 3-5**,

- when the link on the RRPP port of the transit node is faulty, the node sends the Link-Down packet to inform the master node.
- When receiving the Link-Down packet, the master node changes its status from Complete to Failed and unblocks the secondary port.

  If the Link-Down packet is lost during transmission, the Polling mechanism of the master node is needed. If the secondary port of the master node does not receive the Hello packet sent from the primary port within the period specified by the Fail timer, the master node also assesses that the ring is faulty and then unblocks the secondary port.
- When the network topology is changed, the master node needs to refresh the FDB to ensure that packet can be sent to the correct destination. In addition, the master node sends the COMMON-FLUSH-FDB packet from the primary port and secondary port to notify all transit nodes to refresh FDBs. As shown in **Figure 3-6**.

Figure 3-6 Schematic diagram of the master node changing to the Failed state

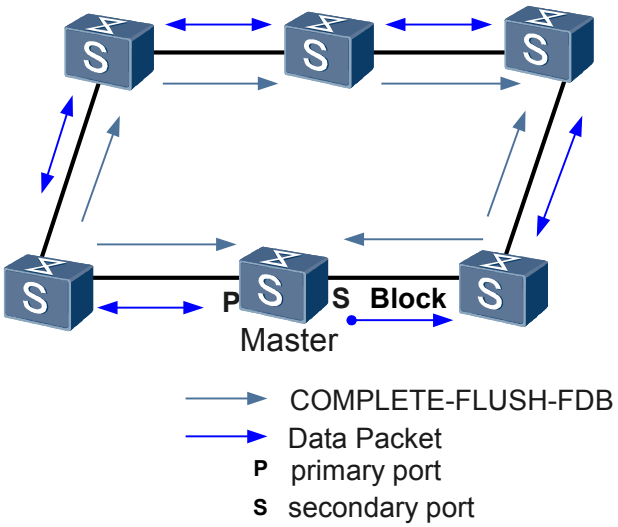


3. The fault is rectified.
   - When the RRPP port of the transit node is recovered, the transit node enters the Preforwarding state and blocks the recovered port.
   - The master node periodically sends the Hello packet through the primary port. When all the faulty links on the ring are recovered, the secondary port of the master node can receive the Hello packet.
   - When receiving the Hello packet, the master node changes to the Complete state and blocks the secondary port.
   - The master node sends the COMPLETE_FLUSH_FDB packet from the primary port to notify all transit nodes to refresh FDBs.

   If the packet of refreshing the FDB is lost during transmission, a standby mechanism is available to unblock the temporarily blocked ports of transit nodes. If the transit node is in the Preforwarding state, it unblocks the temporarily blocked port in case no FDB refreshing packet is received from the master node within the period specified by the Fail timer.

After receiving the COMPLETE_FLUSH_FDB packet, the transit node changes to the Link-Up state, unblocks the temporarily blocked ports, and refreshes FDBs, as shown in **Figure 3-7**.

Figure 3-7 Schematic diagram of ring recovery



Operation Principles of Multiple RRPP Rings

The operation process in the case of multiple rings is similar to that of the single-ring. The difference is that the mechanism of checking the channel status of sub-ring protocol packets on the major ring is used in the case of multiple rings. For details, see "Mechanism of Checking the Channel Status of Sub-ring Protocol Packets on the Major Ring".

Another difference is that, in the case of multiple rings, when receiving COMMON-FLUSH-FDB or COMPLETE-FLUSH-FDB packets from the sub-ring, nodes on the sub-ring refresh FDBs to relearn the address entries. Then, re-routing is performed for the data traffic.

The transit node on the major ring unblocks the temporarily blocked port only when it receives the COMPLETE-FLUSH-FDB packet sent from the major ring rather than the sub-ring.

3.3.3 RRPP Implementation Mechanism

Polling Mechanism

- Hello timer and fail timer

  When RRPP detects the link status of the Ethernet ring through the Polling mechanism, the master node sends Hello packets according to the Hello timer and checks whether the secondary port receives Hello packets within a set period according to the Fail timer. Then, the master node determines whether to unblock the secondary port.

  - The value of the Hello timer specifies the interval at which the master node sends Hello packets from the primary port. The value of the Hello timer ranges from 1 to 10, in seconds.

- The value of the Fail timer specifies the maximum delay during which the primary port sends the Hello packet and the secondary port receives the Hello packet. The value of the Fail timer ranges from 3 to 1200, in seconds.

When the link is faulty, RRPP fast convergence enables the transit node in the Link-Up state to immediately notify the master node to unblock the secondary port through the Link-Down packet. When the link is recovered, the master node notifies the transit node to unblock the temporarily blocked port through the COMPLETE-FLUSH-FDB packet. This has no relation with the value range of the Hello and Fail timers.

The Fail timer on the transit node is used to set the time to unblock the temporarily blocked port.

- Process of the polling mechanism

  Polling mechanism is used by the master node on an RRPP ring to detect the network status. The process of the polling mechanism is as follows:

  1. The value of the Hello timer specifies the interval at which the master node periodically sends Hello packets from the primary port.
  2. Hello packets are transmitted through all transit nodes on the ring.
     - If the secondary port of the master node receives the Hello packet before the Fail timer times out, the master node believes that the ring is in the Complete state.
     - If the secondary port of the master node does not receive the Hello packet after the Fail timer times out, the master node believes that the ring is in the Failed state.

     When the secondary port on the master node in the Failed state receives the Hello packet, the master node performs the following operations.
     - Goes to the Complete state.
     - Blocks the secondary port.
     - Refreshes the FDB.
     - Sends packets from the primary port to notify all transit nodes to unblock temporarily blocked ports and refresh FDBs.

Link Fault Notification Mechanism

Devices serving as master nodes on the RRPP ring need to fast sense the status change of links on the ring. The master node can only fast sense faults on its directly-connected links. For faults on non-directly connected links, the master node asks the link fault notification mechanism for help. When links are recovered, a link recovery mechanism is used to notify the master node.

- Link fault notification mechanism

  The process of the link fault notification mechanism is as follows:

  1. If a link on the ring is faulty, the port directly connected to the link becomes Down.
  2. The transit node immediately sends a Link-Down packet to the master node to report the change of the link status.
  3. When receiving the Link-Down packet, the master node assesses that the ring fails. Then, it immediately unblocks the secondary port and sends the packet to notify other transit nodes to refresh FDBs.
  4. After other transit nodes refresh their FDBs, the data stream is switched to the normal link.

As shown in **Figure 3-8**, when a link on the ring goes Down, the two directly-connected devices immediately send LINK-DOWN packets from the other ports to the master node, which, after receiving packets, unblock the blocked secondary port.

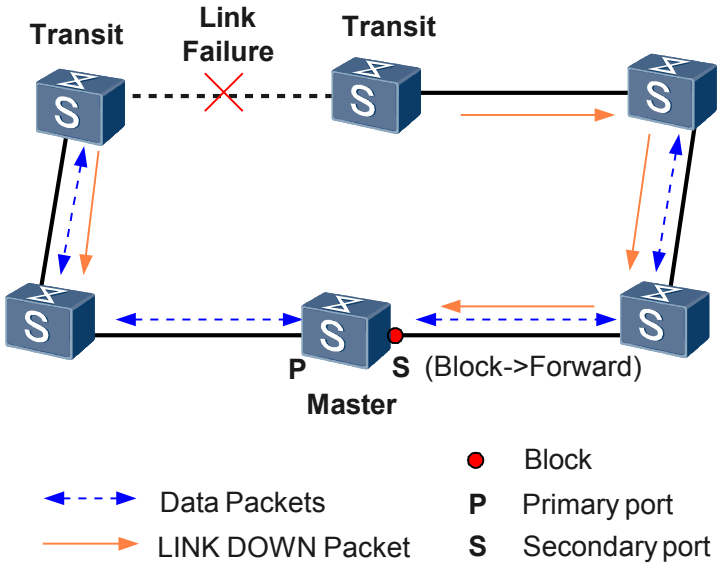
- Link recovery mechanism

  The process of the link recovery mechanism is as follows:

  1. If the faulty link is recovered, the port of the transit node changes to the Up state.
  2. In this case, the transit node temporarily blocks the recovered port. The Hello packet sent from the master node, however, can pass through the temporarily blocked port.
  3. Once the secondary port on the master node receives the Hello packet, the master node considers that the ring recovers to the healthy status.
  4. Then, the master node blocks the secondary port and sends packets to notify all transit nodes to unblock temporarily blocked ports and refresh FDBs.

  As shown in **Figure 3-8**, when the faulty link is recovered, transits nodes at the two ends temporarily block the recovered port. The Hello packet sent from the master node, however, can pass through the temporarily blocked port. Once the secondary port on the master node receives the Hello packet, the master node considers that the ring recovers to the healthy status. Then, the master node blocks the secondary port and sends the COMPLETE-FLUSH-FDB packet to notify transit nodes to unblock the temporarily blocked port and refresh FDBs. In this case, data stream can be switched back to the normal path from the backup link and the ring restores to the normal state.

Figure 3-8 Schematic diagram of link recovery



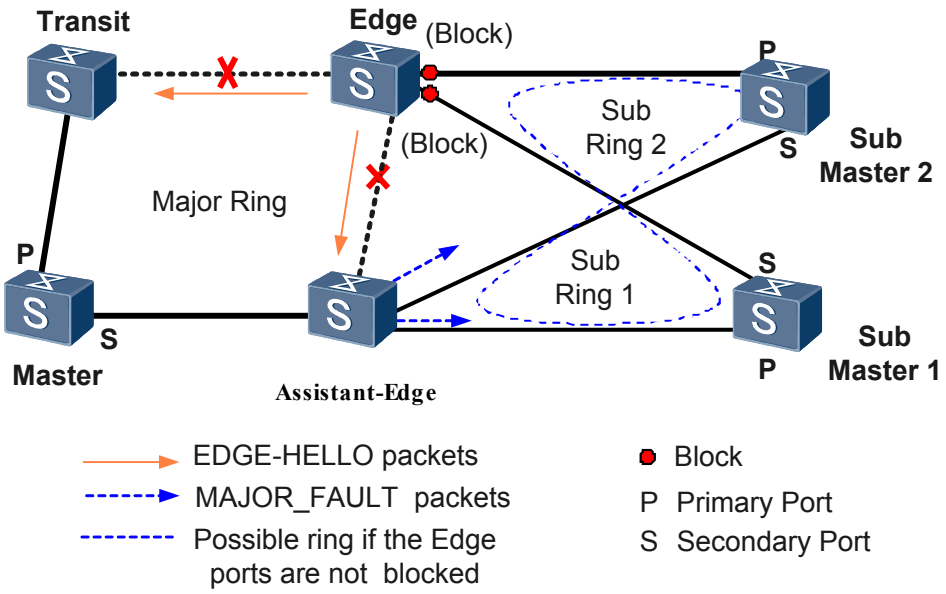
Mechanism of Checking the Channel Status of the Sub-Ring Protocol Packets on the Major Ring

This mechanism is applied on the network where multiple sub-rings are crossed with the master ring to prevent loops among sub-rings after secondary ports are unblocked by master nodes on sub-rings.

As shown in **Figure 3-9**, when the common link between the major ring and sub-ring is faulty and at lease one non-common link is faulty, the master node of each sub-ring blocks its secondary

port ("S" in the preceding figure) because the secondary port no longer receives the Hello packet. In this case, broadcast loops (blue dashed lines in the preceding figure) may occur between sub-rings. To prevent this situation, the mechanism of checking the channel status of the sub-ring protocol packets on the major ring is used.

Figure 3-9 Schematic diagram of loop formation between sub-rings



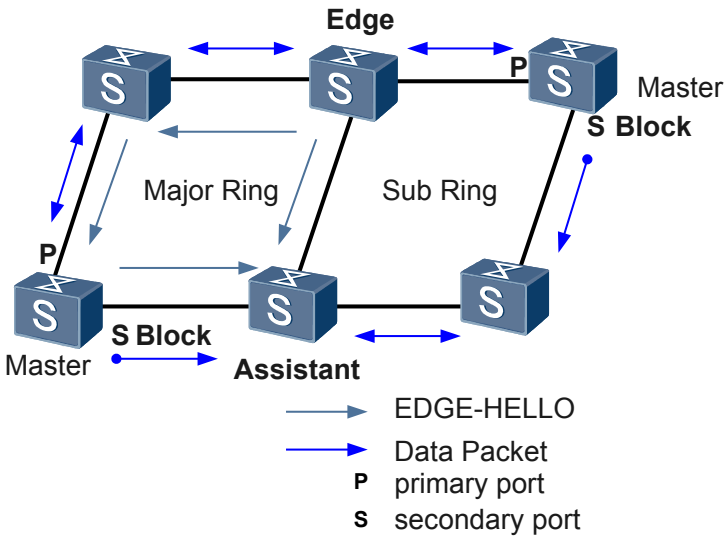
The process of the mechanism is described as follows:

1. Check the channel status of sub-ring protocol packets on the major ring.

   The edge nodes on a sub-ring periodically send Edge-Hello packets to the major ring through two RRPP ports on the major ring. The Edge-Hello packets pass through all the nodes on the ring before reaching the assistant edge node. After receiving the Edge-Hello packets, the assistant edge node does not forward the packets.

   As shown in **Figure 3-10**, edge nodes at the two ends of the major ring sends Edge-Hello packets to the major ring through two RRPP ports.

Figure 3-10 Schematic diagram of sending Edge-Hello packets from edge nodes



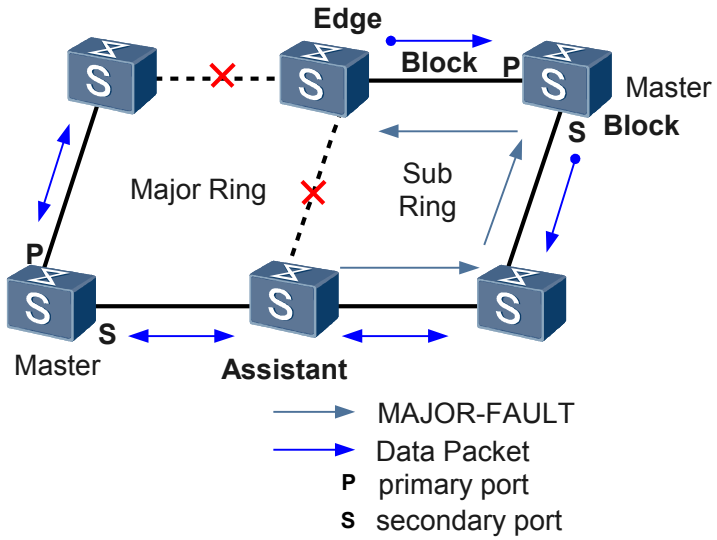
If the assistant edge node receives the Edge-Hello packets within the specified period, it indicates that the packet channel is normal. Otherwise, it indicates that the channel is faulty.

2. The channel breaks off and the edge node blocks the edge ports.

   After the assistant edge node detects that the channel for sub-ring protocol packets breaks off, the edge port immediately sends the Major-Fault packet to the edge node through the sub-ring. After receiving the Major-Fault packet, the edge node blocks its edge port.

   As shown in **Figure 3-11**, the assistant node sends the Major-Fault packet to the edge node through the sub-ring.

Figure 3-11 Schematic diagram of blocking edge ports in response to the Major-Fault packet received on edge nodes

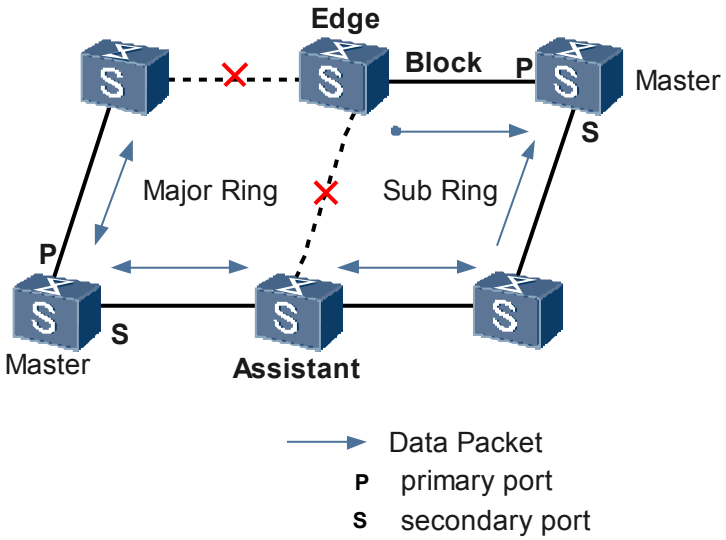


3. The master node of the sub-ring unblock the secondary port even after the Hello timer expires.

After the edge node blocks its edge port, the channel for sub-ring protocol packets breaks off because of the failure in the major ring. Thus, the master node of the sub-ring cannot receive the Hello packet sent within the specified period. The master node, therefore, turns Failed and unblocks the secondary port.

As shown in **Figure 3-12**, the edge node blocks its edge port. The master node of the sub-ring unblocks the secondary port that is blocked in **Figure 3-12**.

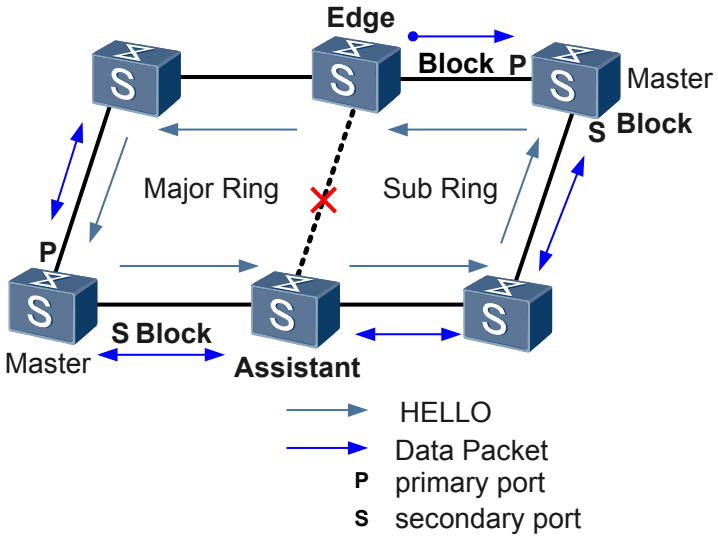
Figure 3-12 Schematic diagram of the failed sub-ring caused by the blocked channel on the major ring



4. The channel for the sub-ring protocol packets is recovered.

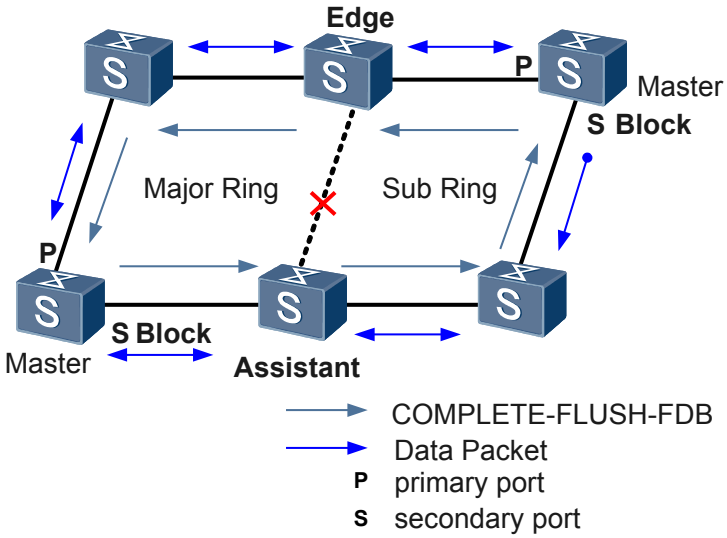
   As shown in **Figure 3-13**, after the links on the major ring are restored, the communication between the edge node and assistant edge node recovers. Then, the channel for the sub-ring protocol packets is recovered. The secondary port on the sub-ring can receive the Hello packets sent from the master node. Then, the master node goes Complete and blocks the secondary port.

Figure 3-13 Schematic diagram of the recovery of the channel of the sub-ring protocol packets



As shown in **Figure 3-14**, the master node on the sub-ring sends the COMPLETE-FLUSH-FDB packets. After receiving the packets, the edge node unblock the edge port.

Figure 3-14 Schematic diagram of unblocking the edge ports on the edge node of the sub-ring



3.3.4 RRPP Features Supported by the S6700

Port Types Supported by RRPP

Multiple physical interfaces can be bound to a logical interface, namely, a trunk, to increase the bandwidth, enhance reliability, and implement load balancing. RRPP supports Eth-Trunk, which can improve the performance and reliability of the RRPP ring.

Table 3-3 Port types supported by RRPP

Supported Interface Type	Description
XGE interface	10GE interface
Eth-Trunk interface	XGE trunk interface

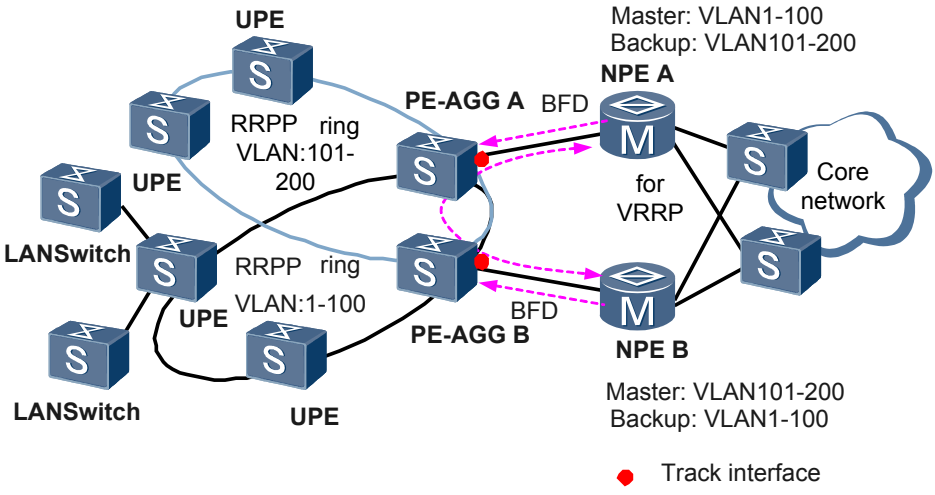
Association of RRPP and BFD

RRPP solution for the ME makes the fast switchover of Layer 2 services possible on the active and standby NPEs.

As shown in **Figure 3-15**, NPE A and NPE B provide active/standby protection and load balancing for the Layer 2 services of VLANs 1 to 200.

- For the services of VLANs 1 to 100, NPE A is in the Active state; for the services of VLANs 101 to 200, NPE A is in the Standby state. When all devices and connections are normal, NPE A directly discards all the packets of VLANs 101 to 200.
- For the services of VLANs 101 to 200, NPE B is in the Active state; for the services of VLANs 1 to 100, NPE B is in the Standby state. When all devices and connections are normal, NPE B directly discards all the packets of VLANs 1 to 100.

Figure 3-15 Networking diagram of Metro Ethernet RRPP



DSLAM: Digital Subscriber Line Access Multiplexer

PE-AGG: PE-Aggregation

UPE: Underlayer Provider Edge

NPE: Network Provider Edge

BFD: Bidirectional Forwarding Detection

VRRP: Virtual Router Redundancy Protocol

When a node on NPE A or the connection between NPE A and PE-AGG A is faulty, NPE B can fast detect the fault based on BFD for VRRP and then automatically upgrade to the Active state to process data of VLANs 1 to 100.

After the active/standby switchover, PE-AGGs and UPEs send packets to incorrect destinations according to the MAC addresses that are originally learned. For example, after NPE B is switched

to the Active state, PE-AGG B transmits packets of VLANs 1 to 100 to PE-AGG A according to the original MAC address table.

To solve this problem, PE-AGG A needs to fast sense its connection status with NPE A and clear the local MAC address table when the connection status changes. At the same time, all nodes on the RRPP ring of VLANs 1 to 100 must clear the learnt MAC address table. When the connection status restores, active/standby switchover is performed over the NPE. Therefore, when the link goes Up from Down, PE-AGG A notifies other nodes on the RRPP ring to clear the MAC address table.

After monitoring interfaces are configured on PE-AGG nodes, RRPP rings can monitor the status of the connections between PE-AGG nodes and NPEs. The status of the monitoring interface changes in the following situations:

- When a node on NPEs or the connection between the PE-AGG node and NPE is faulty, the status of the monitoring interfaces goes Down.
- When a node on NPEs or the connection between the PE-AGG node and NPE is recovered, the status of the monitoring interfaces goes Up.

When the status of the monitoring interface or the status of the BFD session on the interface changes, the node where the monitoring interface resides clears the dynamic MAC address entry and at the same time sends a COMMON-FLUSH-FDB packet to notify other nodes on the RRPP ring to clear their dynamic MAC address entries.

After clearing the original dynamic MAC address entries, nodes on the RRPP ring, when receiving data packets of VLANs 1 to 100, broadcast these packets throughout the VLAN. NPE-B is switched to the Active state and processes data packets of VLANs 1 to 100. When the reverse traffic is forwarded through NPE-B, all PE-AGGs and UPEs can correctly learn the MAC addresses. In this case, the incoming data packets can be normally forwarded to NPE-B.

The monitoring interface supports hot swap. Each time the monitoring interface is pulled out or reinserted, dynamic MAC address entries are cleared on the RRPP ring. When the monitoring interface is pull out but another interface is inserted, configuration of the monitoring interface is cleared.

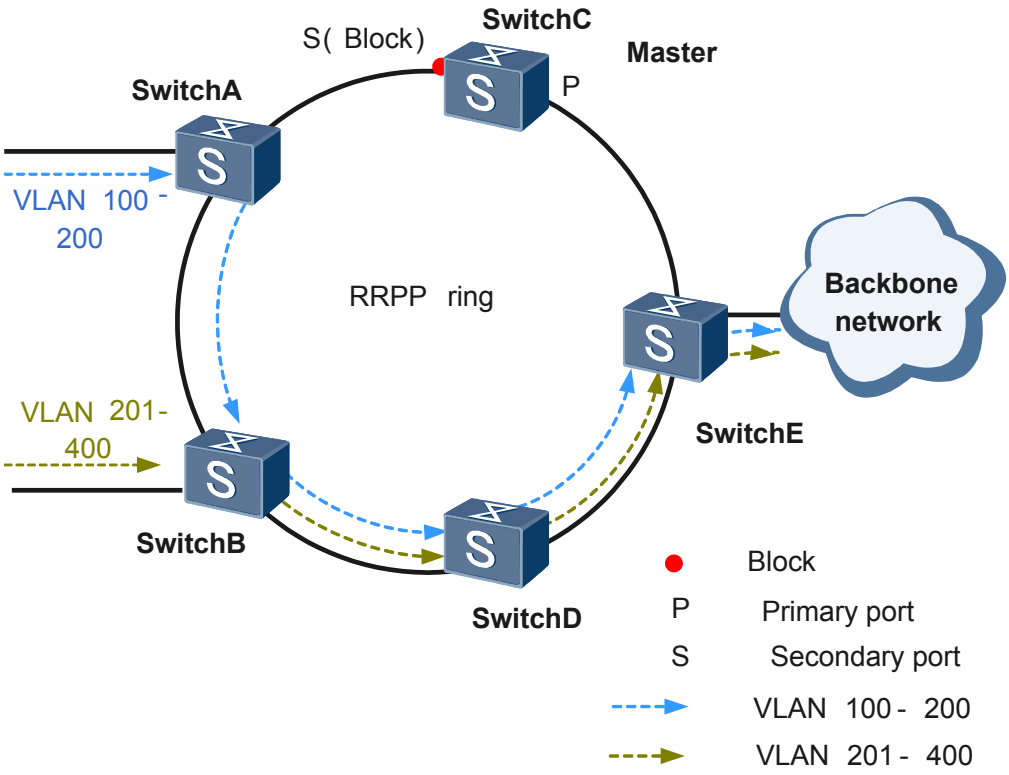
RRPP Multi-Instance

- Multi-Instance

  On a common RRPP network, a physical ring can be configured with only one RRPP domain and a physical ring can have only one master node.

  As shown in **Figure 3-16**, when the master node is in the Complete state, the blocked secondary port prevents all user packets from passing. In this case, all user packets are transmitted on the RRPP ring through one path. As a result, the link at the secondary port side of the master node becomes idle, which leads to a waste of bandwidth.

Figure 3-16 Networking diagram of RRPP



RRPP multi-instance is implemented on the basis of domains. One physical ring can be configured with multiple RRPP domains. In a domain, all ports, nodes, and topologies must comply with basic RRPP principles. Correspondingly, a physical ring can have multiple master nodes. Each master node independently detects the completeness of the physical ring and then blocks or unblocks its secondary port.

A domain can contain one or more instances, each of which specifies a VLAN range. The VLANs contained in the domain are called the protected VLANs of the RRPP domain. Protected VLANs include data VLANs, major ring control VLANs, and sub-ring control VLANs. The topology calculated by the ring protection protocol in each domain is valid only for the protected VLAN of its domain.

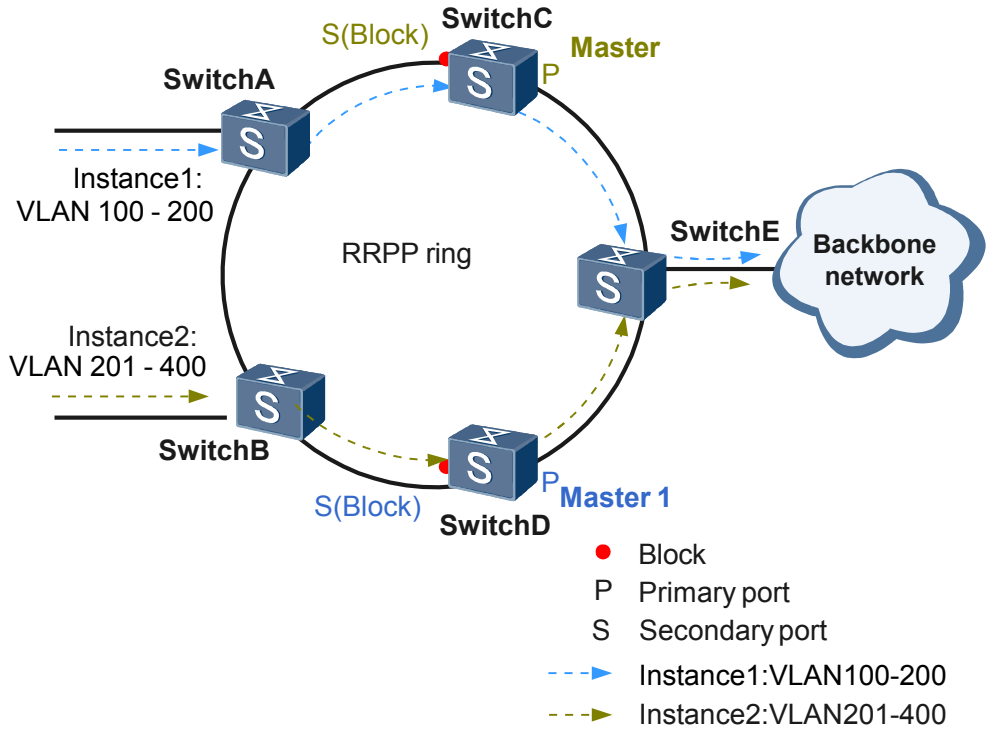
Domains are configured with different protected VLANs and data belonging to different protected VLANs is transmitted through different paths. In this manner, when the master node in a domain blocks its secondary port, data of other protected VLANs are not affected. This implements link backup and traffic load balancing.

As shown in **Figure 3-17**, two domains exist on the RRPP ring of multi-instance that consists of Switch A, Switch B, Switch C, Switch D, and Switch E. Switch C is the master node of Domain 2, and Switch D is the master node of Domain 1.

- Instance 1 is created in Domain 1, and data of VLANs 100 to 200 is mapped to instance 1 and transmitted along the path Switch A->Switch C->Switch E. Master 2 (Switch C) serves as the master node of Domain 2. The secondary port of Master 2 is Blocked. Only data of VLANs 201 to 400 is blocked and data of VLANs 100 to 200 can pass through.
- Instance 2 is created in Domain 2, and data of VLANs 201 to 400 is mapped to instance 2 and transmitted along the path Switch B->Switch D->Switch E. Master 1 (Switch D)

serves as the master node of Domain 1. The secondary port of Master 1 is Blocked. Only data of VLANs 100 to 200 is blocked and data of VLANs 201 to 400 can pass through.

Figure 3-17 Networking diagram of RRPP multi-instance

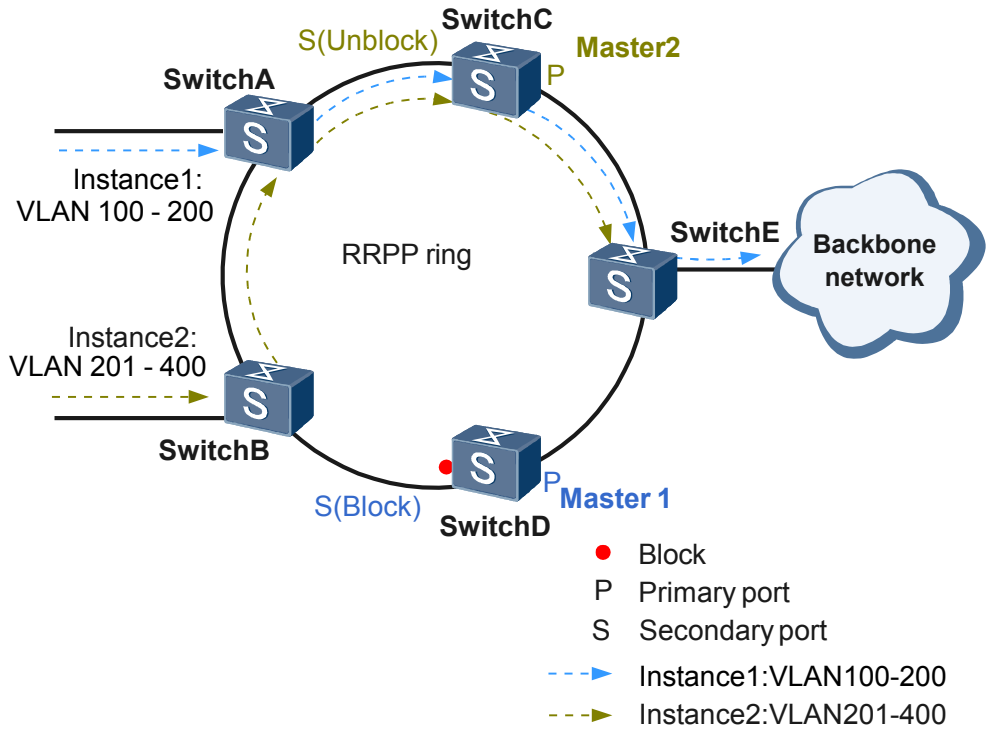


When a node or a link fails, each RRPP domain calculates the topology independently and updates forwarding database (FDB) entries on each node.

As shown in **Figure 3-18**, a fault occurs on the link between Switch D and Switch E. This fault does not affect the transmission path of the packets of VLANs 100 to 200 in domain 1, but the fault blocks the transmission path of the packets of VLANs 201 to 400 in domain 2.

The master node Switch C in domain 2 cannot receive Hello packets on the secondary port. Thus, Switch C unblocks the secondary port and notifies each node in domain 2 to refresh FDB entries. After the topology in domain 2 re-converges, the transmission path of the packets of VLANs 201 to 400 changes to Switch B->Switch A->Switch C->Switch E.

Figure 3-18 Networking diagram of RRPP multi-instance (link fault)



After the link between Switch D and Switch E recovers, Switch C receives Hello packets on the secondary port. Thus, Switch C unblocks the secondary port and notifies each node in domain 2 to refresh FDB entries. After the topology in domain 2 re-converges, the packets of VLANs 201 to 400 are switched back to the previous path Switch B->Switch D->Switch E.

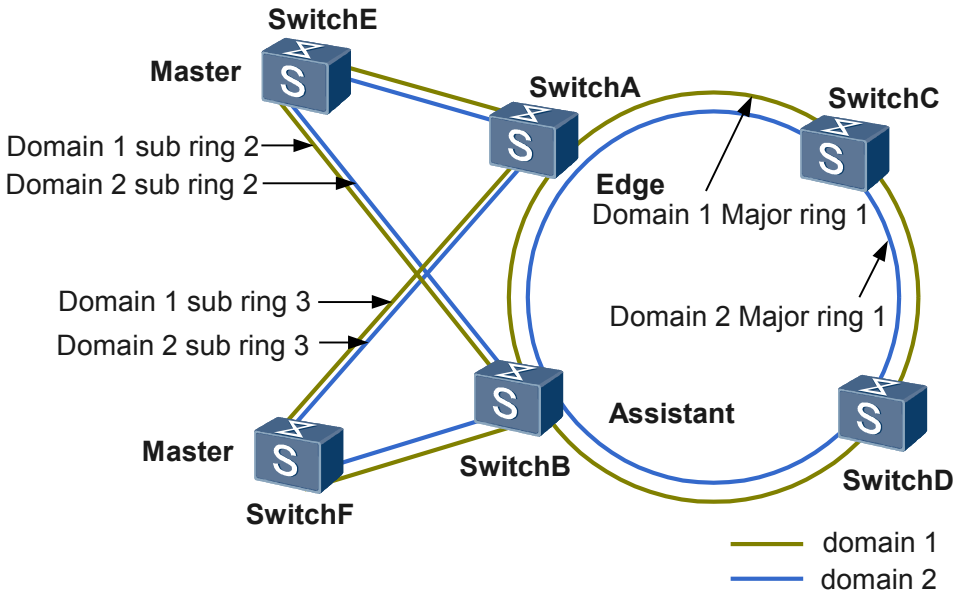
- Ring Group

  In RRPP multi-instance, sub-rings are group to reduce the number of received and sent Edge-Hello packets and improve system performance.

  In the mechanism of checking channel status of the sub-ring protocol packets on the major ring, the edge node on the sub-ring periodically sends EDGE-Hello packets to the two RRPP ports on the major ring to detect the completeness of the channel for sub-ring protocol packets.

  As shown in **Figure 3-19**, the edge nodes of multiple sub-rings (sub-ring 2 and sub-ring 3 in Domain 1; sub-ring 2 and sub-ring 3 in Domain 2) are the same device, and the assistant edge nodes of the sub-rings are the same device. In addition, edge nodes and assistant edge nodes have the same link of the major ring. That is, the Edge-Hello packets of edge nodes on the sub-rings arrive at assistant edge nodes along the same path. In this case, the sub-rings with the same edge nodes and assistant edge nodes can be added into a ring group. A sub-ring in the ring group is selected to send Edge-Hello packets to detect the channel for sub-ring protocol packets on the major ring. This can reduce the number of received and sent Edge-Hello packets and improve system performance.

Figure 3-19 Networking diagram of ring group in RRPP multi-instance



The process of selecting a sub-ring in the ring group to send the Edge-Hello packet is as follows:

1. The sub-rings with the smallest domain ID are selected from all the activated rings in the ring group of the edge node. In **Figure 3-19**, the sub-rings with the smallest domain ID are ring 2 in Domain 1 and ring 3 in Domain 1.
2. The smallest ring ID is selected from the sub-rings with the smallest domain ID, and then the node on the sub-ring with the smallest ring ID sends the Edge-Hello packet. In **Figure 3-19**, the sub-rings with the smallest ring ID is ring 2 in Domain 1. Thus, in the ring group formed by ring 2 in Domain 1, ring 3 in Domain 1, ring 2 in Domain 2, and ring 3 in Domain 2, the edge node on ring 2 in Domain 1 sends the Edge-Hello packet.
3. On all the activated rings in the ring group where assistant edge nodes reside, when any sub-ring receives an Edge-Hello packet, it notifies other sub-rings of the packet.

- LinkUp Delay Timer

  The LinkUp delay timer is used to set a delay time for the link on the RRPP ring to become Up from Down.

  If the link status changes frequently on a ring, transmission paths of the traffic are switched frequently. As a result, loop flapping occurs. Starting the LinkUp delay timer on the master node can reduce the impact on the system caused by loop flapping.

  The value of the LinkUp delay timer ranges from 0 to 1000, in seconds. Its value must be smaller than the value of the Fail timer by at least 2.

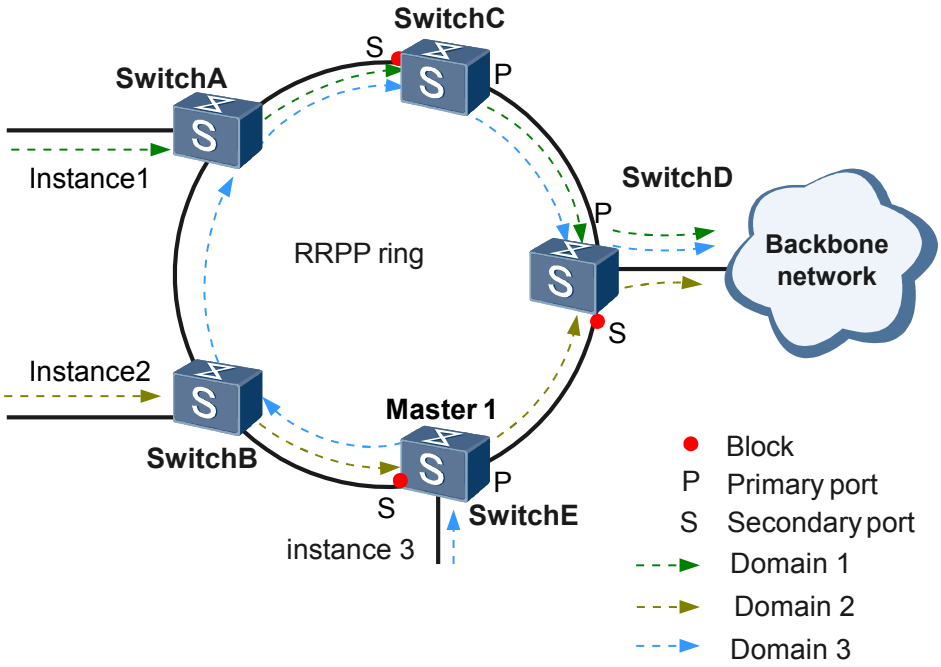
  As shown in **Figure 3-20**, three domains exist on the RRPP multi-instance ring formed by Switch A, Switch B, Switch C, Switch D, and Switch E.

  The protected VLANs of Domain 1 are the VLANs mapped to instance 1, the master node is Switch E, and the transmission path of packets is Switch A->Switch C->Switch D.

  The protected VLANs of Domain 2 are the VLANs mapped to instance 2, the master node is Switch C, and the transmission path of packets is Switch B->Switch E->Switch D.

The protected VLANs of Domain 3 are the VLANs mapped to instance 3, the master node is Switch D, and the transmission path of packets is Switch B->Switch A->Switch C->Switch D.

Figure 3-20 LinkUp delay timer when a ring is in the Complete state



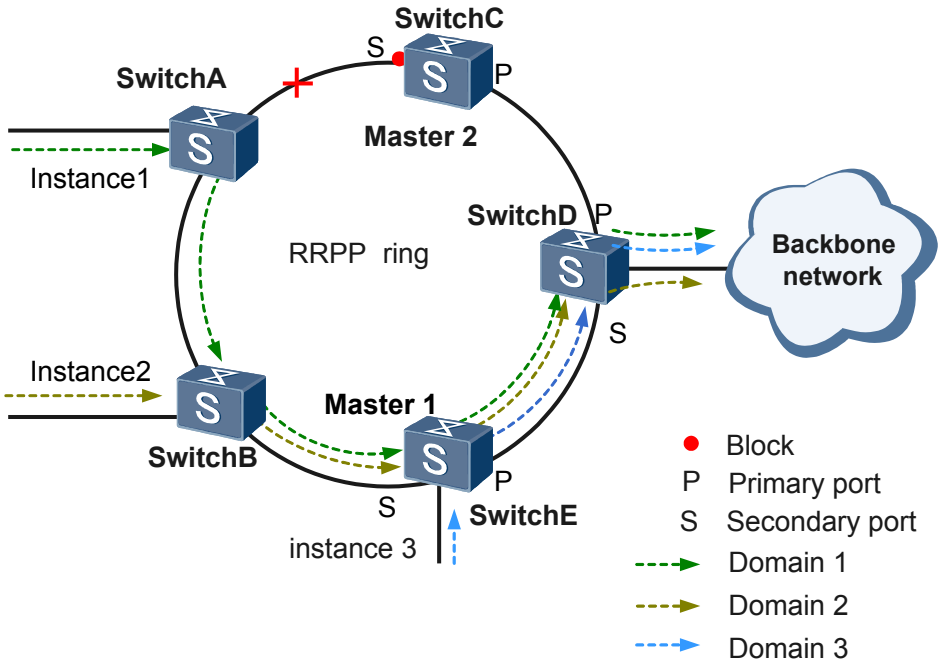
If a fault occurs on the link between Switch A and Switch C, packets in Domain 1 and Domain 3 cannot be transmitted. After the topology reconverges, the transmission paths of the packets in Domain 1 and Domain 3 change. Conversely, the transmission path of the packets in Domain 2 remains unchanged.

The transmission path of the packets in Domain 1 is Switch A->Switch B->Switch E->Switch D.

The transmission path of the packets in Domain 3 is Switch E->Switch D.

As shown in **Figure 3-21**, if the LinkUp delay timer is not started, the topologies of Domain 1 and Domain 3 reconverge after the link between Switch A and Switch C goes Up. In this case, the transmission paths of the packets in Domain 1 and Domain 3 change again.

Figure 3-21 LinkUp delay timer when a ring is in the Failed state



Starting the LinkUp delay timer on the master node can prevent transmission paths from changing frequently and avoid loop flapping.

Before expiring, the timer is closed if the link status changes because the master node receives a Link-Down packet or the link goes Down. The master node does not process the Hello packets received by the secondary port, and the topology of the RRPP ring remains the re-converged topology after a fault occurs on the link.

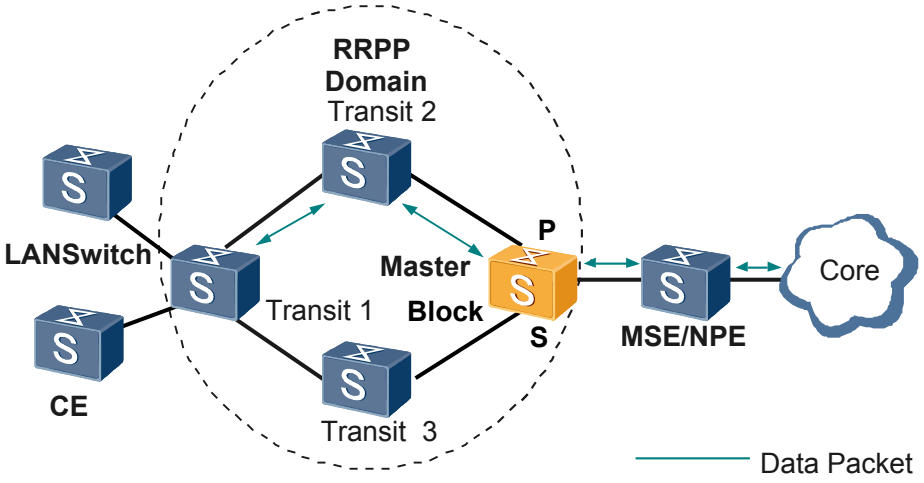
After the timer expires, the master node processes the Hello packets received by the secondary port. That is, the master node blocks the secondary port and notifies other nodes to refresh FDBs. In this case, the RRPP ring re-converges.

3.4 Application

A Single RRPP Ring

Figure 3-22 is the networking of a single RRPP ring. Normally, data flow is transmitted along the path of Transit 1->Transit 2->Master. If the link between Transit 1 and Transit 2 fails, the path of the data flow on the RRPP ring changes.

Figure 3-22 Networking diagram of a single RRPP ring



DSLAM: Digital Subscriber Line Access Multiplexer	UMG: Universal Media Gateway
CE	NPE: Network Provider Edge
MSE: Multi Service Edge	-

- Fast switchover of Layer 2 services

  After being notified of faults on the link between Transit 1 and Transit 2, the master node immediately unblocks the secondary port.

  At this time, the network topology changes, the original MAC address table of each node cannot correctly guide the Layer 2 forwarding. Thus, Layer 2 traffic is interrupted. After unblocking the secondary port, the master node immediately requires other nodes on the ring to re-learn MAC address entries. The Layer 2 traffic on the RRPP ring is switched onto the path of Transit 1->Transit 3->Master.

- Fast switchover of Layer 3 services

  After being notified of faults on the link between Transit 1 and Transit 2, the master node immediately unblocks the secondary port.

  At this time, the network topology changes, the original ARP and FIB of each node cannot correctly guide the Layer 3 forwarding. After unblocking the secondary port, the master node immediately requires other nodes on the ring to relearn MAC address entries. The Layer 2 traffic on the RRPP ring is switched onto the path of Transit 1->Transit 3->Master.

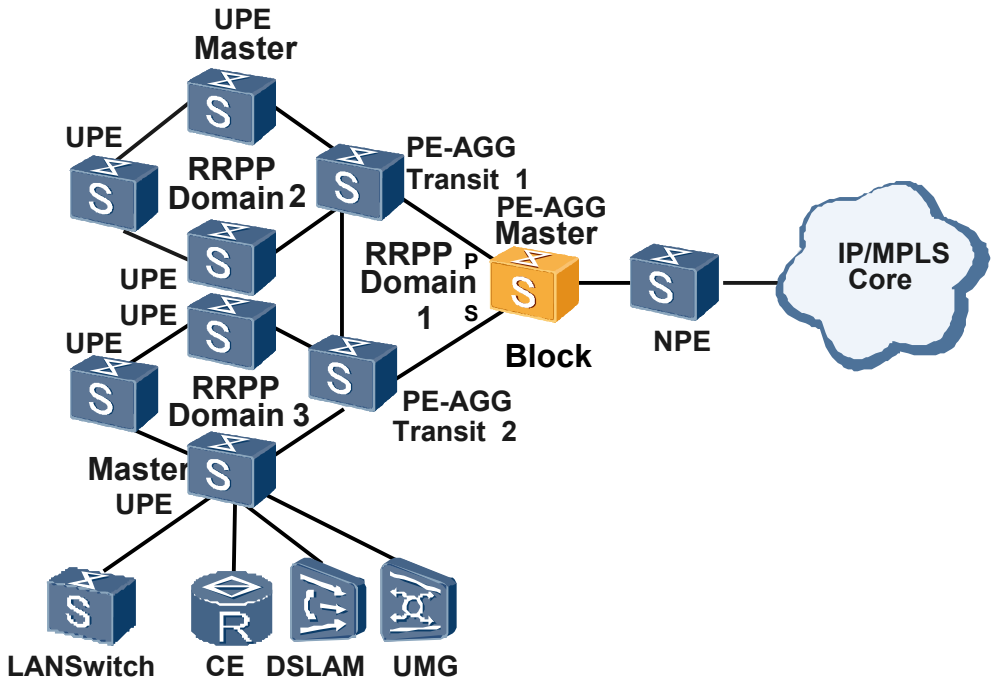
Tangent RRPP Rings

Generally, the metro Ethernet uses two-layer rings:

- One layer is the convergence layer between the convergence devices PE-AGGs, for example, RRPP Domain 1 in **Figure 3-23**.
- The other layer is the access layer between PE-AGGs and UPEs, such as RRPP Domain 2 and RRPP Domain 3 in **Figure 3-23**.

As shown in **Figure 3-23**, in this networking, tangent RRPP rings can be adopted. That is, the convergence layer is the RRPP major ring; the access layer is the RRPP sub-ring.

Figure 3-23 Application of the tangent RRPP rings



DSLAM: Digital Subscriber Line Access Multiplexer	UMG: Universal Media Gateway
UPE: Underlayer Provider Edge	PE
PE-AGG: PE-Aggregation	-

Two tangent rings cannot belong to the same RRPP domain. The tangency points are configured to two domains. The master node on a ring can be the tangency point.

For multiple tangent RRPP rings, the failure in one ring does not affect other domains. The convergence process of RRPP rings in a domain is the same as that of a single ring.

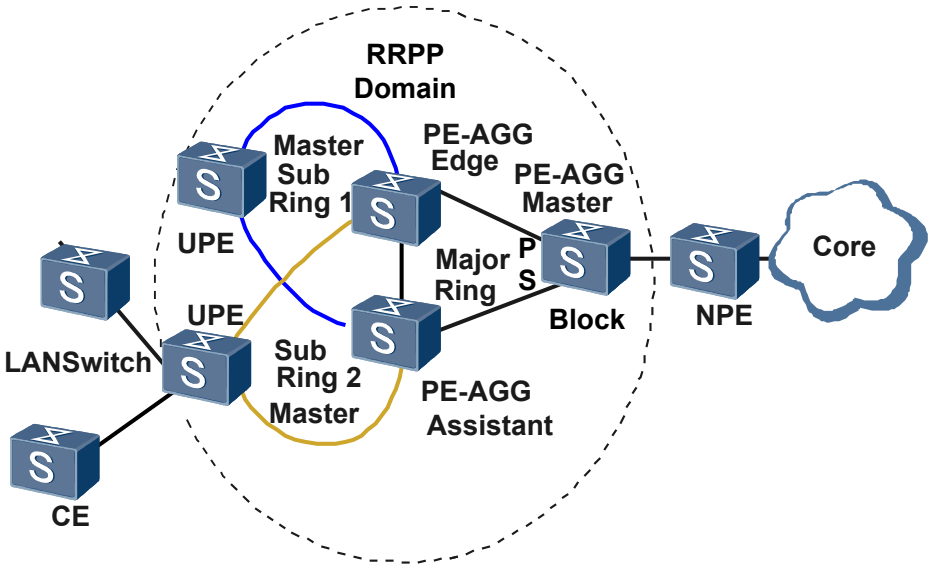
Intersected RRPP Rings

Generally, the metro Ethernet uses two-layer rings:

- One layer is the convergence layer between the convergence devices PE-AGGs.
- The other layer is the access layer between PE-AGGs and UPEs.

As shown in **Figure 3-24**, in this networking, intersected RRPP rings can be adopted. That is, the convergence layer is the RRPP major ring; the access layer is the RRPP sub-ring.

Figure 3-24 Application of intersected RRPP rings



DSLAM: Digital Subscriber Line Access Multiplexer	UMG: Universal Media Gateway
UPE: Underlayer Provider Edge	NPE: Network Provider Edge
PE-AGG: PE-Aggregation	-

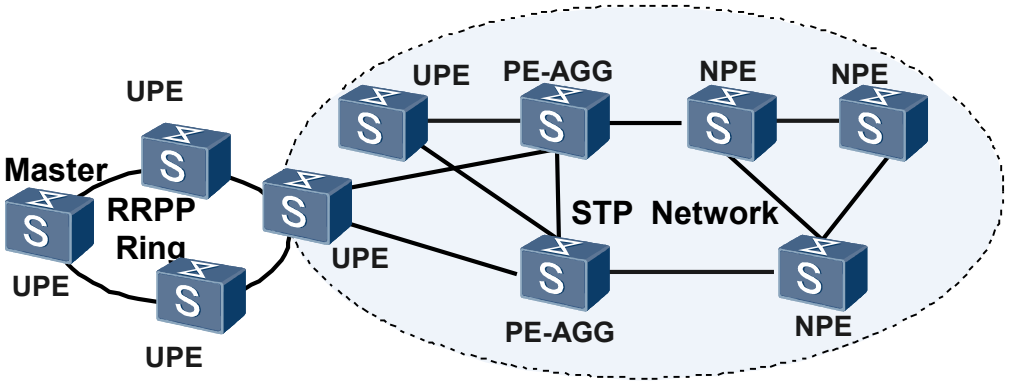
The RRPP major ring on the convergence layer and the RRPP sub-rings belong to the same RRPP domain.

The major ring and sub-ring have two intersection points. A node cannot exist on the intersected segments. The two intersection points can be configured only as transit nodes. On a sub-ring, if one of the two transit nodes is configured as an edge node, the other node must be configured as an assistant edge node.

Association of RRPP and STP

As shown in **Figure 3-25**, RRPP can be deployed on an Ethernet network enabled with STP/RSTP/MSTP only in tangent mode. On an Ethernet network enabled with STP/RSTP/MSTP, any RRPP node can be selected as the master node of the RRPP ring.

Figure 3-25 Networking diagram of the association of RRPP and STP



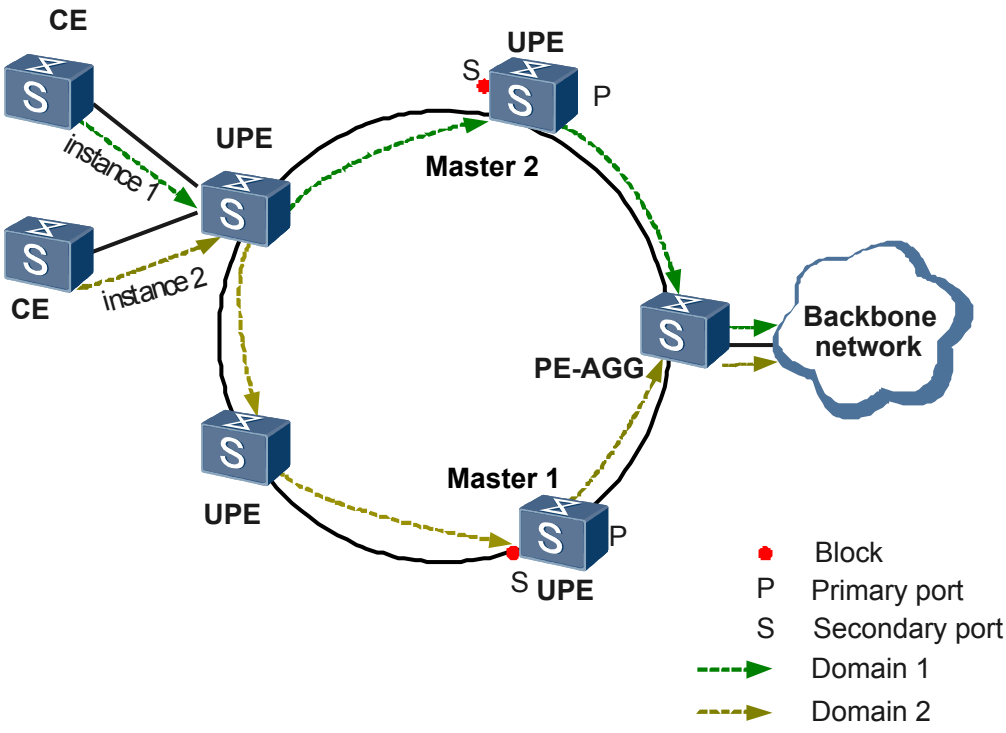
UPE: Underlayer Provider Edge NPE: Network Provider Edge

PE-AGG: PE-Aggregation -

Application Scenario for Single-homed Access to an RRPP Convergence Ring of Multi-instance

As shown in **Figure 3-26**, CEs access the RRPP ring through a UPE. Then, the traffic that flows into the RRPP ring through the UPE flows into the backbone network through a PE-AGG.

Figure 3-26 Networking diagram of single-homed access to an convergence RRPP ring of multi-instance



UPE: Underlayer Provider Edge NPE: Network Provider Edge

PE-AGG: PE-Aggregation	-

Four UPEs and a PE-AGG construct a ring in two domains, that is, ring 1 in Domain 1 and ring 1 in Domain 2. Domain 1 processes the packets from VLANs 101 to 200; Domain 2 processes the packets from VLANs 1 to 100.

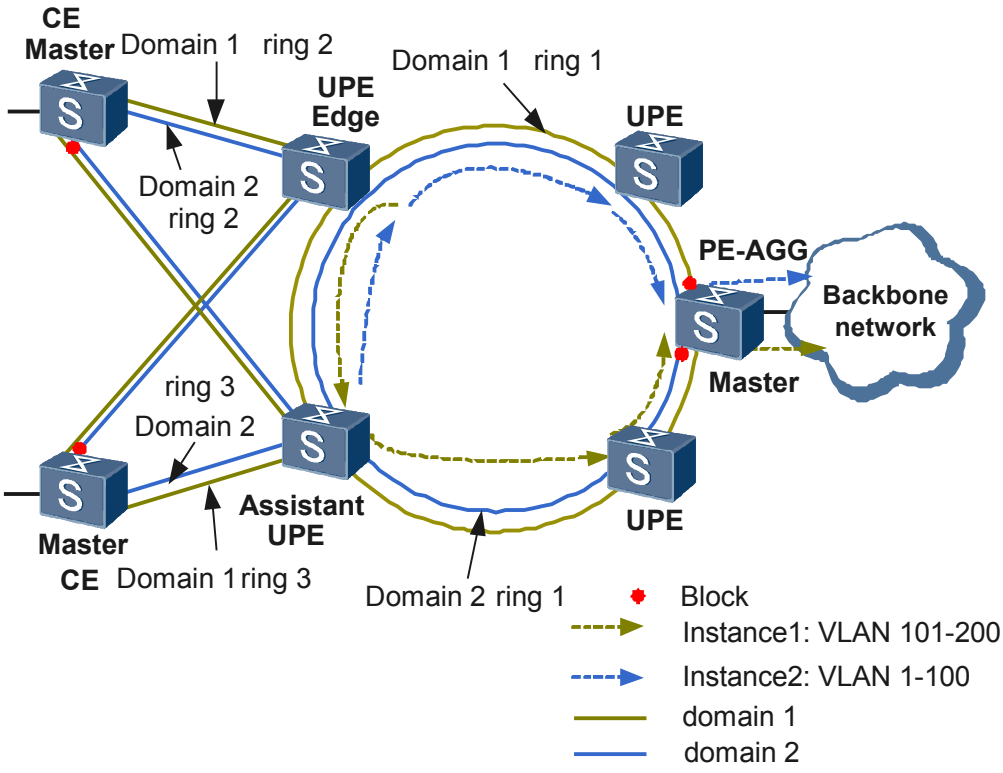
Domain 1 corresponds to instance 1; Domain 2 corresponds to instance 2. Service VLANs accessed from CEs range from 1 to 200.

All accessed service VLANs are covered and service VLANs processed in the two RRPP domains cannot overlap. Load balancing for the traffic of Domain 1 and Domain 2 is carried out on the RRPP ring.

Application Scenario for Intersected RRPP Rings of Multi-instance in MAN

As shown in **Figure 3-27**, CEs are dual-homed to UPEs and therefore two RRPP rings are formed.

Figure 3-27 Networking diagram of intersected RRPP rings of multi-instance in MAN (CEs support RRPP multi-instance)



UPE: Underlayer Provider Edge	NPE: Network Provider Edge
PE-AGG: PE-Aggregation	-

Four UPEs and a PE-AGG construct a ring. Enable RRPP multi-instance on the ring. Traffic flows into the backbone network through PE-AGGs.

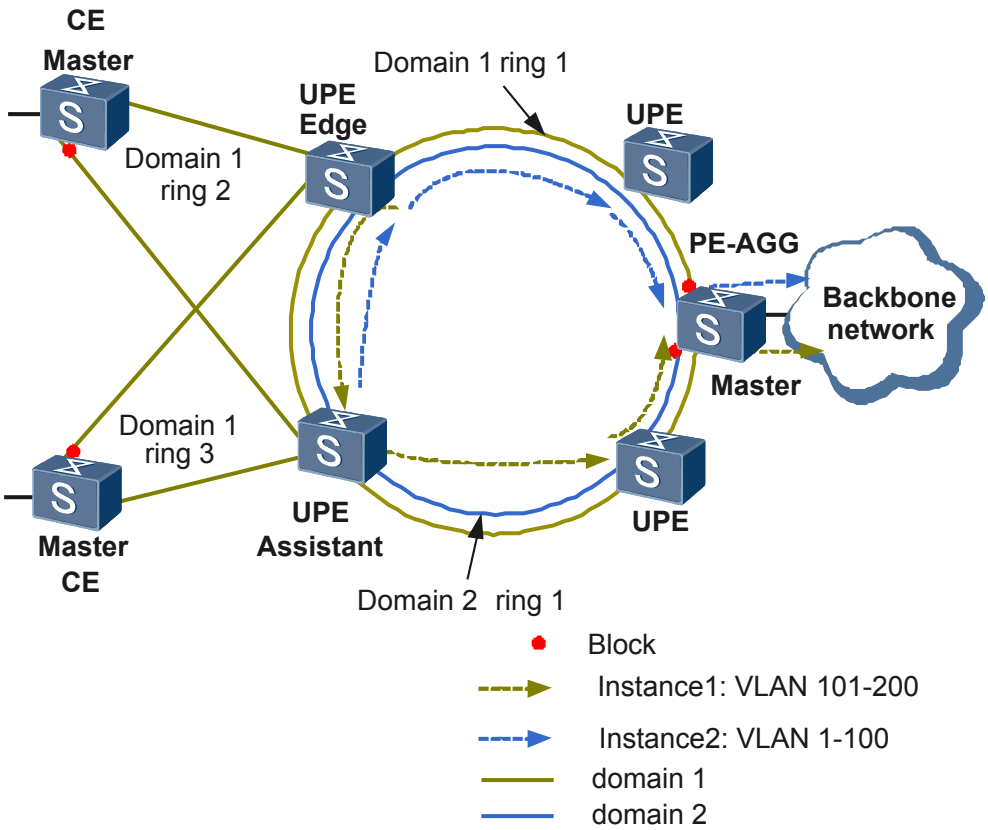
Two RRPP rings are configured on the four UPEs and a PE-AGG, which are ring 1 in Domain 1 and ring 1 in Domain 2. Domain 1 processes the packets from VLANs 101 to 200; Domain 2 processes the packets from VLANs 1 to 100.

Two CEs and two UPEs construct 4 RRPP sub-rings, which are ring 2 in Domain 1, ring 2 in Domain 2, ring 3 in Domain 1, and ring 3 in Domain 2.

Various services are accessed to sub-rings. RRPP rings provide master/slave protection and load balancing for the Layer 2 services of VLANs 1 to 200. When all the nodes and connection on rings are normal, the traffic accessed to sub-rings is transmitted along different paths according to different service VLANs. Load balancing is implemented.

As shown in **Figure 3-28**, CEs may not support RRPP multi-instance. The major ring constructed by four UPEs and a PE-AGG has multiple domains; however, the sub-rings constructed by CEs and UPEs have only one domain. In that case, load balancing is not implemented on sub-rings and paths of all service VLANs on sub-rings are the same. After entering the major ring, the traffic accessed to sub-rings is transmitted along different paths according to different service VLANs. Load balancing is implemented.

Figure 3-28 Networking diagram of intersected RRPP rings of multi-instance in MAN (CEs do not support multi-instance)



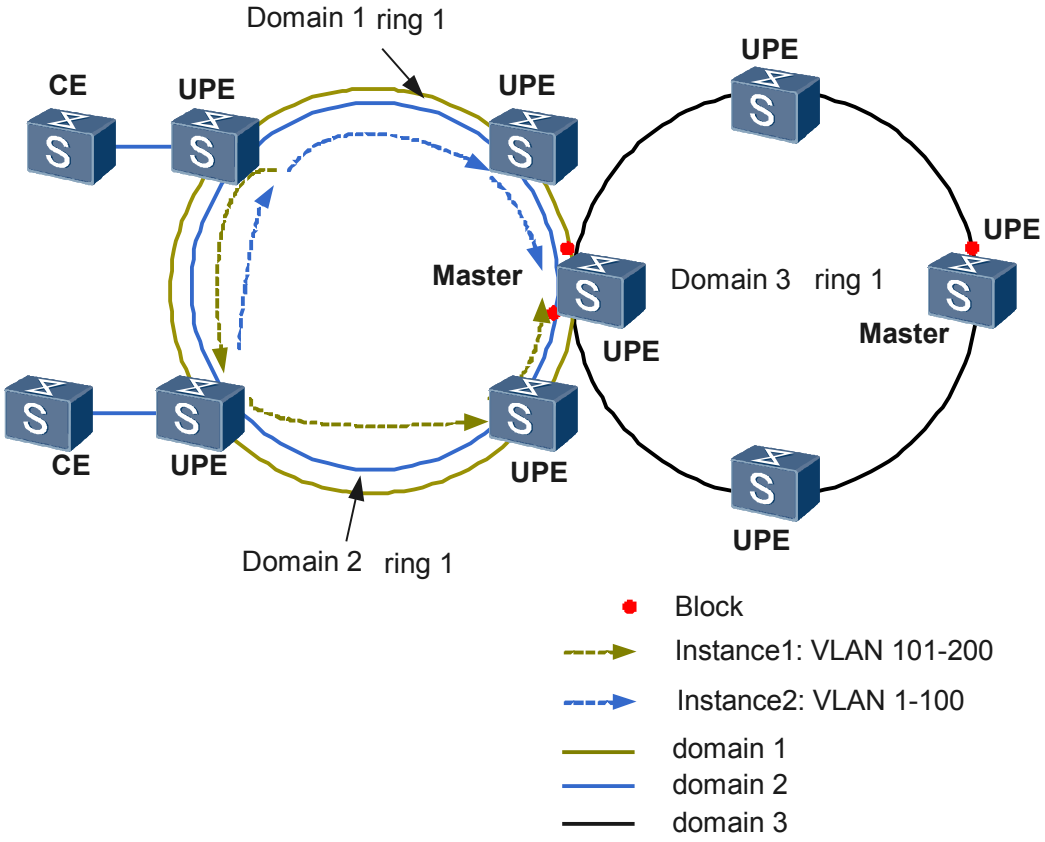
UPE: Underlayer Provider Edge	NPE: Network Provider Edge
PE-AGG: PE-Aggregation	-

Whatever CEs support multi-instance, service VLANs on RRPP rings must within the processing scope of the RRPP domain. Otherwise, loops of data packets occur.

Application Scenario for Tangent RRPP Rings of Multi-instance in MAN

As shown in **Figure 3-29**, two RRPP rings namely, ring 1 in Domain 1 and ring 1 in Domain 2, are configured on the five UPEs on the left. An RRPP ring, namely, ring 1 in Domain 3, is configured on the four UPEs on the right side.

Figure 3-29 Networking diagram of tangent RRPP rings of multi-instance in MAN



UPE: Underlayer Provider Edge NPE: Network Provider Edge

Domain 1 processes the packets from VLANs 101 to 200; Domain 2 processes the packets from VLANs 1 to 100; Domain 3 processes the packets from VLANs 1 to 200.

RRPP rings on the left side provide master/slave protection and load balancing for the Layer 2 services of VLANs 1 to 200. When all the nodes and connection on RRPP rings are normal, the traffic flowed into rings from CEs is transmitted along different paths according to different service VLANs. Load balancing is implemented.

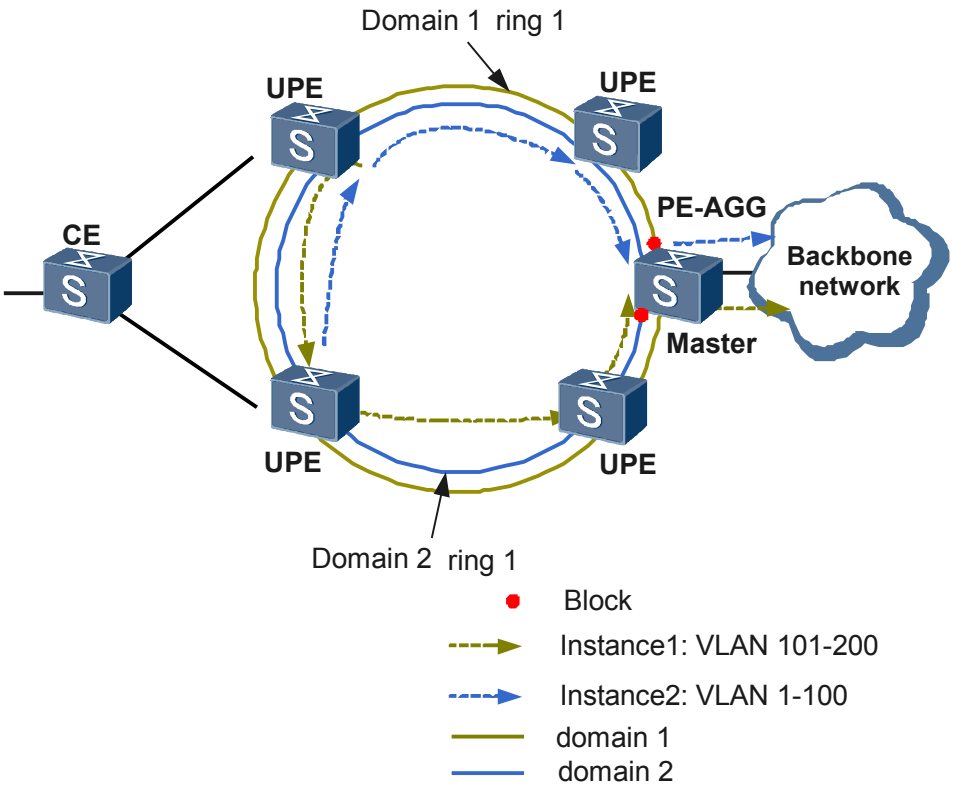
The traffic of VLANs 1 to 200 is injected into the RRPP ring on the right side from the tangent node.

Service VLANs on RRPP rings must within the processing scope of the RRPP domain. Otherwise, loops of data packets occur.

Application of the Hybrid Networking of RRPP Multi-instance Rings and Smartlinks

As shown in **Figure 3-30**, CEs are dual-homed to UPEs through a smartlink.

Figure 3-30 Networking diagram of the association of RRPP multi-instance rings and smartlinks



UPE: Underlayer Provider Edge	NPE: Network Provider Edge
PE-AGG: PE-Aggregation	-

Four UPEs and a PE-AGG construct a ring. RRPP multi-instance is enabled on the ring.

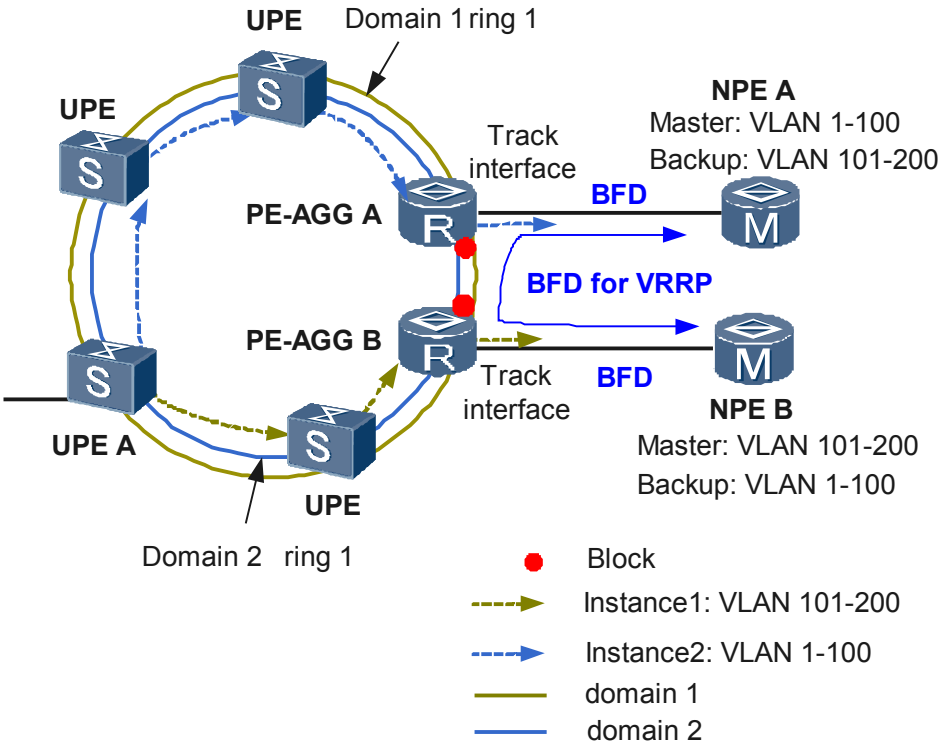
Traffic flows into the backbone network through the PE-AGG.

Nodes on the RRPP ring and the PE-AGG must identify and process the Flush packets of the smartlink.

Association of RRPP Multi-instance Rings and BFD

As shown in **Figure 3-31**, two RRPP rings are configured on the four UPEs and two PE-AGGs, which are ring 1 in Domain 1 and ring 1 in Domain 2. Domain 1 processes the packets from VLANs 101 to 200; Domain 2 processes the packets from VLANs 1 to 100.

Figure 3-31 Association of RRPP multi-instance rings and BFD



UPE: Underlayer Provider Edge	NPE: Network Provider Edge
PE-AGG: PE-Aggregation	-

Configure BFD for VRRP on at the network side. Configure Track Interface on the two RRPP rings of PE-AGG A and PE-AGG B to check the BFD status of interfaces at the network side.

NPE A and NPE B provide master/slave protection and load balancing for the Layer 2 services of VLAN 1 to VLAN 200.

- For the services of VLANs 1 to 100, NPE A is in the Active state; for the services of VLANs 101 to 200, NPE A is in the Standby state. When all the nodes and the connection of the devices are normal, the packets from VLANs 1 to 200 are directly discarded by NPE B.
- For the services of VLANs 101 to 200, NPE B is in the Active state; for the services of VLANs 1 to 100, NPE B is in the Standby state. When all the nodes and the connection of the devices are normal, the packets from VLANs 1 to 100 are directly discarded by NPE B.

RRPP rings provide active/standby protection and load balancing for the Layer 2 services of VLANs 1 to 200. When all the nodes and connection on rings are normal, the traffic accessed to rings from UPE A is transmitted along different paths according to different service VLANs. Load balancing is implemented.

3.5 Terms and Abbreviations

Term

Term	Description
RRPP	Rapid Ring Protection Protocol, a link layer protocol specially used to prevent loops on an Ethernet ring network. Devices running RRPP discover loops on the network by exchanging information with each other, and block certain interfaces to eliminate loops.
MSTP	Multi-Spanning Tree Protocol, a new spanning tree protocol defined in IEEE 802.1s. It introduces concepts of region and instance. Based on different requirements, MSTP divides a big 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) are transmitted between network bridges.
VPLS	Virtual Private LAN Service, a type of point-to-multipoint service provided in the public network. VPLS ensure that isolated user sites can be connected through MAN/WAN and two sites can communicate as if they were in a LAN.
FDB	Forwarding DataBase, including entries for guiding multicast data forwarding. There are layer 2 FDB and layer 3 FDB. The layer 2 FDB refers to the MAC table, which provides information about the MAC address and outbound interface and guides the layer 2 forwarding. The layer 3 FDB refers to the ARP table, which provides information about the IP address and outbound interface and guides the layer 3 forwarding.

Abbreviation

Abbreviation	Full Spelling
STP	Spanning Tree Protocol
VLAN	Virtual Local Area Network
VRRP	Virtual Router Redundancy Protocol
BFD	Bidirectional Forwarding Detection

4 Ethernet OAM

About This Chapter

4.1 Introduction to Ethernet OAM

Definition

Ethernet OAM is short for Ethernet Operation, Administration and Maintenance.

Ethernet OAM improves management and maintenance capabilities on the Ethernet and guarantees a stable network.

Purpose

Ethernet has developed as the major Local Area Network (LAN) technology because its protocol is easy to implement and features low-cost network implementations..

The original Ethernet was developed for LANs. LANs do not have high requirements for reliability and stability. Thus, the Ethernet lacks an OAM mechanism, which hinders the Ethernet from being an network. Therefore, Ethernet OAM is the trend.

4.2 References

The following table lists the references of this document.

Document	Description	Remarks
IEEE Std 802.3ah-2004	Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment: Media Access Control Parameters, Physical Layers, and Management Parameters for Subscriber Access Networks	-

4.3 Principles

Link-level Ethernet OAM

Link-level Ethernet OAM technologies, such as Ethernet in the First Mile OAM (EFM OAM) defined in IEEE 802.3ah, provide functions including link connectivity detection, link fault monitoring, remote fault notification, and remote loopback for two directly-connected devices..

A CE is an edge device of a customer network. It connects the customer network to the provider network. Different from a CE, a PE is an edge device of a provider network. It connects the provider network to the customer network.

Functioning as a link-level Ethernet OAM, EFM OAM provides point-to-point fault detection to perform OAM detection on the link between two directly-connected devices.

OAM Fault Association

OAM fault association is used to transmit fault information between detection protocols, such as between EFM OAM and Ethernet CFM or between Ethernet OAM and BFD. Along with wider applications of detection protocols such as Ethernet OAM, BFD, and MPLS OAM, the OAM fault association module will be applied in more scenarios.

4.3.1 EFM OAM

EFM OAM has the following functions: peer discovery, link monitoring, fault notification, and remote loopback.

Peer Discovery

The working mode of EFM OAM is an attribute of the interface enabled with EFM OAM. EFM OAM has two working modes: active mode and passive mode. The active mode is adopted by an interface by default.

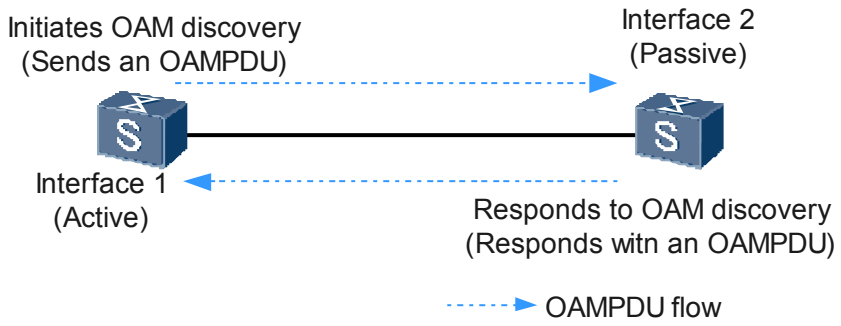
Before configuring EFM OAM on an interface, you need to configure its working mode:

- If the active mode is adopted, the interface initiates the peer discovery process.
- If the passive mode is adopted, the interface cannot initiate the peer discovery process.

  This prevents two interfaces in passive mode from negotiating for setting up sessions. In addition, interfaces in passive mode cannot initiate requests for remote loopback or variables.

When an interface is enabled with EFM OAM and works in active mode, the interface initiates the peer discovery process. The interface and the peer interface then enter the EFM OAM discovery state.

Figure 4-1 Schematic diagram of peer discovery



As shown in **Figure 4-1**, assume that the EFM OAM working mode of Interface 1 is active and that of Interface 2 is passive. After Interface 1 is enabled with EFM OAM, the peer discovery process is as follows:

1. Interface 1 sends an OAM Protocol Data Unit (OAM PDU) to Interface 2. The OAM PDU carries the EFM OAM configuration of Interface 1.
2. After receiving the OAM PDU, Interface 2 compares the EFM OAM configuration information of Interface 1 with that of itself, and then responds with an OAM PDU. The

OAM PDU sent from Interface 2 to Interface 1 carries not only the EFM OAM configuration information of both Interface 1 and Interface 2 but also the Flags field indicating whether Interface 2 is satisfied with the EFM OAM configuration of Interface 1.

For the format of an OAM PDU, see the following figure.

Figure 4-2 Format of an EFM OAM PDU

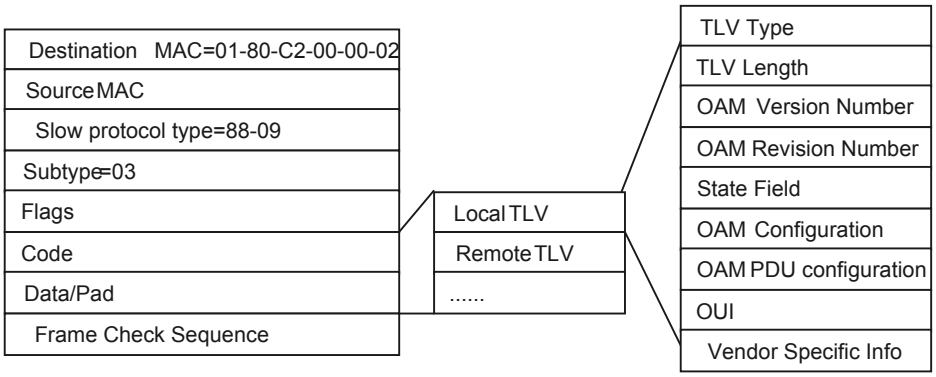


Figure 4-3 Description of the OAM configuration information

Name	Description
OAM Configuration	7:5 reserved, TLV is set to 0 in local information
	4 variable reachability 1=DTE supports that OAM PDUs are in response to sent variables 0=DTE does not support that OAM PDUs are in response to sent variables
	3 link event 1=DTE supports to parse link events 0=DTE does not support to parse link events
	2 OAM remote loopback 1=DTE can be configured with OAM remote loopback 0=DTE cannot be configured with OAM remote loopback
	1 Sending OAM PDUs 1=when receiving link works, DTE can send OAM PDUs 0=when receiving link works, DTE cannot send OAM PDUs
	0 OAM mode 1=DTE is configured to work in active mode 0=DTE is configured to work in passive mode

3. After receiving the OAM PDU sent from Interface 2, Interface 1 compares the EFM OAM configuration information of Interface 2 to check whether the configurations match.

After the preceding process, if the EFM OAM configurations of Interface 1 and Interface 2 match, the two interfaces enter the EFM OAM detect state. In the detect state, the two interfaces

send OAM PDUs regularly to maintain adjacencies. If the EFM OAM configurations of Interface 1 and Interface 2 do not match, the two interfaces remain in the discovery state and keep sending OAM PDUs for negotiation till the negotiation succeeds or EFM is disabled on Interface 1 (or Interface 2). OAM PDUs are sent at a fixed interval of 1s.

Link Monitoring

After the EFM OAM link monitoring function is configured, the statistics about the interface management module at the physical layer are queried and the communication quality of the link connecting the current interface is detected. In the set observation period, if the number of error frames, error codes, or error frame seconds reaches or exceeds the threshold that is set for the interface, it indicates that the link is faulty. As a result, an alarm is generated and reported to the NMS. An error frame second is a one-second interval during which at least one error frame is detected.

Fault Notification

The following faults are notified: timeout of protocol packets, faults on physical links, transmission faults of the OAM module.

- When a protocol packet times out or a physical link becomes faulty, the fault event is logged and reported to the NMS.
- When a transmission fault occurs in the OAM module, the fault event is recorded to the log and reported to the NMS.

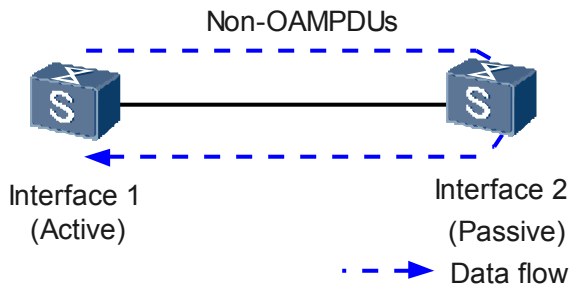
  If the reverse link is reachable, an OAM PDU is sent to notify the peer of the fault. Upon receiving the notification message, the peer logs the event carried in the message and reports it to the NMS.
- If the EFM OAM module is associated with other modules such as BFD, Ethernet CFM, and MPLS OAM, the OAM fault association module can notify the associated modules of the fault.

Remote Loopback

As shown in **Figure 4-4**, when the local interface sends to the peer, the peer loops back non-OAM PDUs to the local interface, instead of forwarding non-OAM PDUs based on their destination addresses. This is called remote loopback.

Remote loopback can be used to locate faults and test link performance. In remote loopback mode, the local interface sends testing packets to the peer. The local device computes communication quality parameters such as the packet loss ratio of the current link according to the number of sent packets and the number of received packets.

Figure 4-4 Schematic diagram of remote loopback



Only the interface in active mode can initiate remote loopback. When a local interface is in active mode and the local interface and the remote interface are in the EFM OAM detect state, the following can be achieved after remote loopback is enabled on the local interface:

1. The interface sends a loopback request to the remote interface and waits for response.
2. After receiving the request, the remote interface replies the local interface with a message and enters the remote loopback state.
3. If the local interface receives the message in 2s, it enters the remote loopback state. Otherwise, the local interface re-sends a loopback request to the remote interface. An interface can re-send a loopback request for up to three times.

When the local interface intends to end the remote loopback test, it sends a message to disable remote loopback. After receiving the message, the remote interface exits the remote loopback state.

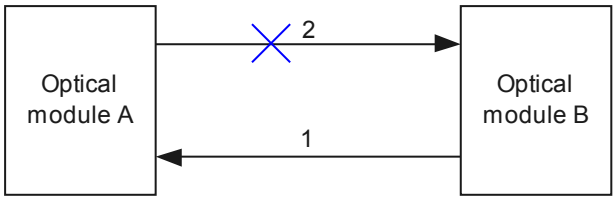
The customer may forget to disable remote loopback, which causes the link to fail to forward service data for a long time. To avoid such a problem, EFM OAM remote loopback can be automatically disabled after a timeout period. The timeout period for remote loopback is configurable. After a timeout period, the local interface automatically sends a message to the remote interface to disable remote loopback.

Single-fiber Fault Detection

Currently, optical interfaces work in full-duplex mode. An optical interface, which can receive packets, considers itself to be physically Up. In the following condition, the working status and the physical status of the interface, however, are inconsistent.

As shown in **Figure 4-5**, optical interface A and optical interface B are directly connected. If Line 2 is faulty, optical interface B cannot receive packets and thus changes to Down. Optical interface A can still receive packets from optical interface B through Line 1, and thus optical interface A remains in the Up state. Actually, optical interface A now cannot forward packets, but the device where optical interface A resides does not sense the actual working status of optical interface A. As a result, service transmission is affected.

Figure 4-5 Schematic diagram of EFM OAM single-fiber detection



EFM OAM single-fiber detection can solve the preceding problem.

When EFM detects a fault, if EFM is associated with interface status, EFM on the interface goes Down. For Layer 2 and Layer 3 services, if EFM on the interface goes Down, the interface is physically Down. Thus, service modules can sense the actual working status of the interface. If the fault is rectified and EFM OAM negotiation succeeds, the interface becomes Up and thus services can be normally transmitted.

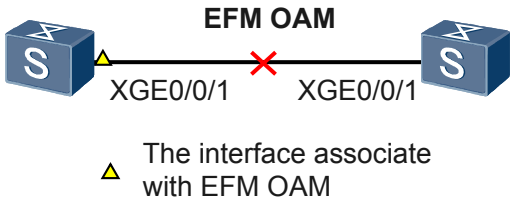
This mechanism effectively solves the problem that the interface status is inconsistent with its actual working status due to a single-fiber fault, thus ensuring the accuracy and reliability of interface status that is sensed by each service module.

4.3.2 OAM Fault Association

Association between EFM OAM and an Interface

As shown in **Figure 4-6**, when an interface running EFM OAM detects a connectivity fault, no packet except EFM protocol packets can be forwarded on the interface. As a result, Layer 2 and Layer 3 services are blocked. Therefore, the association between EFM OAM and the current interface may greatly affect services. When the current interface detects link fault recovery through EFM OAM, all packets can be forwarded on the interface and Layer 2 and Layer 3 services are unblocked.

Figure 4-6 Schematic diagram of the association between EFM OAM and an interface



4.4 Terms and Abbreviations

Terms

NA.

Abbreviations

Abbreviation	Full Spelling
EFM	Ethernet in the First Mile
OAM	Operation Administration & Maintenance

5 MAC SWAP Loopback

About This Chapter

The MAC swap loopback function modifies the header of a received Ethernet frame by swapping the source and destination MAC addresses so that the Ethernet frame can be sent back to the sender. This function checks Ethernet connectivity and network performance.

5.1 MAC SWAP Loopback Overview

Definition

The MAC swap loopback function modifies the header of a received Ethernet frame by swapping the source and destination MAC addresses so that the Ethernet frame can be sent back to the sender. This function checks Ethernet connectivity and network performance.

Purpose

The MAC swap loopback technique can work with a network performance tester to test the connectivity, throughput, and QoS of an Ethernet network.

A MAC swap loopback test can be performed only for Ethernet frames with IP payload.

5.2 Principle of MAC Swap Loopback

The MAC swap loopback function modifies the header of a received Ethernet frame by swapping the source and destination MAC addresses so that the Ethernet frame can be sent back to the sender. This function checks Ethernet connectivity and network performance.

The MAC swap loopback technique can work with a network performance tester to check the connectivity, throughput, and QoS of an Ethernet network.

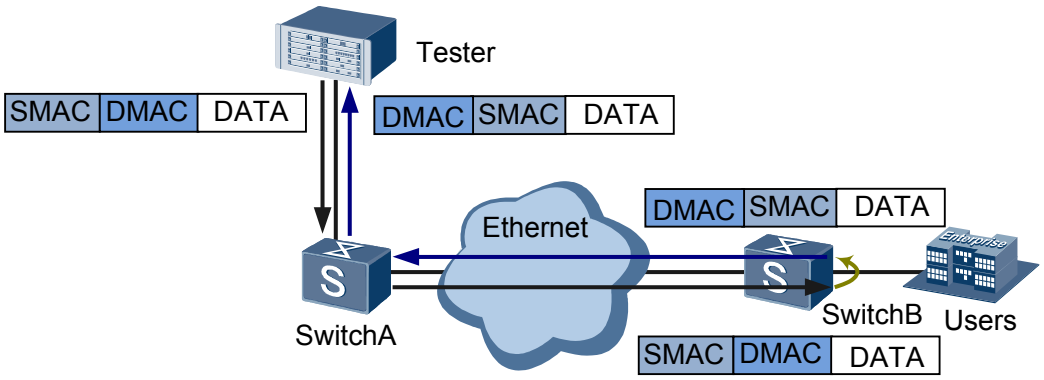
A MAC swap loopback test can be performed only for Ethernet frames with IP payload.

Local MAC Swap Loopback Test

A local MAC swap loopback test checks connectivity and performance of the network between a tester and a tested switch, and performance of the tested switch. As shown in **Figure 5-1**, Ethernet frames sent by a tester traverse an Ethernet network and reach the tested switch (SwitchB). After the tested switch receives the Ethernet frames, it forwards them to the interface where local MAC swap loopback is configured. The tested switch swaps source and destination MAC addresses of all Ethernet frames that match the local MAC swap loopback configuration and sends the Ethernet frames back to the tester through the specified outbound interface. The tester then analyzes network performance based on Ethernet frames sent from the tested switch.

When a local MAC swap test is performed on an interface, services on the interface are interrupted, but services on other interfaces are not affected. If a large number of test packets are sent, test packets occupy bandwidth of other services on the interface that sends test packets back to the tester.

Figure 5-1 Local MAC swap loopback test

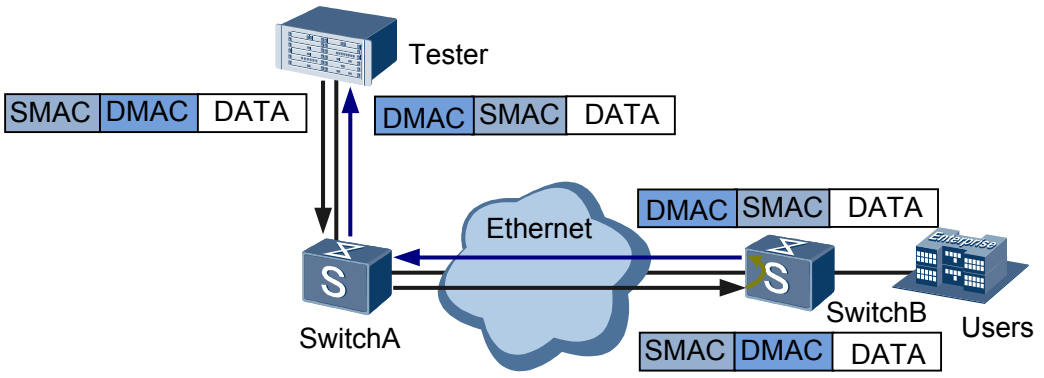


Remote MAC Swap Loopback Test

A remote MAC swap loopback test checks connectivity and performance of the network between a tester and a tested switch but does not check performance of the tested switch. As shown in **Figure 5-2**, Ethernet frames sent by a tester traverse an Ethernet network and reach the tested switch (SwitchB). After the uplink interface of the tested switch receives the Ethernet frames, the switch swaps source and destination MAC addresses of all Ethernet frames that match the remote MAC swap loopback configuration and sends the Ethernet frames back to the tester through this interface. The tester then analyzes network performance based on Ethernet frames sent from the tested switch.

A remote MAC swap loopback test modifies only Ethernet frames matching the MAC swap loopback configuration on an interface, and other Ethernet frames can still be forwarded on the interface. However, if a large number of test packets are sent, test packets occupy bandwidth of other services on the interface.

Figure 5-2 Remote MAC swap loopback test



6 BFD

About This Chapter

6.1 Introduction to BFD

Definition

Bidirectional forwarding detection (BFD) can fast detect communications faults between systems and notify the upper-layer application of the faults.

Purpose

To minimize the impact of a fault on services and improve network availability, a network device is required to fast detect a communications fault between adjacent devices and the upper layer protocol can rectify the fault to ensure normal services.

Currently, the existing detection mechanisms are as follows:

- Hardware detection: For example, the Synchronous Digital Hierarchy (SDH) alarms are used to detect a fault on a link. The hardware detection can fast detect a fault; however, not all media can provide the hardware detection mechanism.
- Slow Hello mechanism: It usually refers to the Hello mechanism offered by a routing protocol. The slow Hello mechanism can detect a fault in seconds. In high-speed data transmission, for example, at gigabit rates, the detection time of more than one second causes loss of a large amount of data. In delay-sensitive services such as the voice service, the delay of more than one second is unacceptable.
- Other detection mechanisms: Different protocols or manufacturers may provide private detection mechanisms; however, it is difficult to deploy the private detection mechanisms when systems are interconnected.

BFD is developed to supplement existing detection mechanisms.

BFD provides the following functions:

- Provides low-cost fast fault detection for channels between adjacent forwarding engines. The detected faults may occur on interfaces, data links, or forwarding engines.
- Provides a single mechanism to detect any media and protocol layers in real time. In addition, the detection period and cost range are variable.

6.2 References

The references of this feature are as follows:

Document	Description	Remarks
RFC5880	Bidirectional Forwarding Detection	-
RFC5882	Generic Application of BFD	-
RFC5883	BFD for Multihop Paths	-
RFC5881	BFD for IPv4 and IPv6 (Single Hop)	-

Document	Description	Remarks
RFC5884	BFD for mpls	-

6.3 Principle of BFD

BFD detects communications faults between forwarding engines. To be specific, BFD detects connectivity of a data protocol on the same path between systems. The path can be a physical link, a logical link, or a tunnel.

BFD can be regarded as a type of service that the system provides.

- The upper layer applications provide BFD with parameters, such as the detection address and the detection time.
- BFD creates, deletes, or modifies the BFD session according to the information and informs the upper layer applications of the session status.

BFD offers the following functions:

- Provides low-cost and short-duration detection for path faults between adjacent forwarding engines.
- Provides a single mechanism that can be used for detection over any media, at any protocol layer, and thus implements the unified detection mechanism in a network.

The following sections describe the basic principle of BFD, including the BFD detection mechanism, detected link types, session establishment modes, and session management.

BFD Detection Mechanism

In the BFD detection mechanism, two systems set up a BFD session, and periodically send BFD control packets along the path between them. If one system does not receive BFD control packets within a specified period, the system considers that a fault occurs on the path.

BFD control packets are encapsulated in the UDP packets. In the initial phase of a BFD session, both systems negotiate with each other over parameters, such as discriminators, expected minimum intervals for sending and receiving BFD control packets, and local BFD session status, carried in BFD control packets. When negotiations are successful, both systems send BFD control packets at the negotiated intervals on the path between them.

To meet the requirement for fast detection, the BFD draft defines that the intervals for sending and receiving BFD control packets are expressed in microseconds. Limited by the current processing capability, BFD-enabled devices of most manufacturers process BFD control packets at millisecond level. Their speed is transformed to microsecond during actual applications. The minimum detection time that the S6700 supports is 30 milliseconds.

BFD provides the following detection modes:

- Asynchronous mode: The main mode is asynchronous mode. In asynchronous mode, two systems periodically send BFD control packets to each other. If one system receives no packets consecutively, the BFD session is considered as Down.
- Query mode: The second mode is the query mode. If multiple BFD sessions exist in a system, periodically sending the costs of BFD control packets affects the running of the system. To prevent this, you can use the query mode. In query mode, after a BFD session

is set up, the system does not periodically send BFD control packets, but detects the connectivity through another mechanism (such as the Hello mechanism of a routing protocol or the hardware detection mechanism) to reduce the cost of the BFD session.

An auxiliary function of the two modes is the Echo function. When the Echo function is activated, a BFD control packet is sent as follows: The local system sends a BFD control packet and the remote system sends the BFD control packet back through the forwarding channel. If consecutive Echo packets are not received, the BFD session is declared Down. The Echo function can work with the asynchronous mode or demand mode.

At present, only the passive Echo function is supported.

Types of Links Detected by BFD

- IP links
- Eth-Trunk
  - Layer 2 Eth-Trunk links
  - Layer 2 Eth-Trunk member links

  BFD sessions used to detect the trunk member interfaces and the trunk interface are independent from each other and can detect links at the same time.
- VLANIF
  - VLAN Ethernet member links
  - VLANIF interfaces

  The BFD session of the VLANIF and the BFD sessions of the vlan member interfaces are independent from each other and can detect these interfaces at the same time.

BFD Session Establishment Modes

A BFD session can be set up in the following modes:

BFD differentiates sessions by My Discriminator and Your Discriminator in the control packets. The main difference in establishment of static and dynamic BFD sessions is that My Discriminator and Your Discriminator are set differently.

- Statically configuring a BFD session

  Statically configuring a BFD session means setting BFD session parameters, including the local discriminator and the remote discriminator, through command lines. Then, a request of BFD session establishment is distributed manually.
- Dynamically establishing a BFD session

  When a BFD session is set up dynamically, the system processes the local discriminator and the remote discriminator as follows:
  - Dynamically allocating the local discriminator

    When the application triggers the dynamic setup of a BFD session, the system allocates the value that belongs to the dynamic session discriminator area as the local discriminator of the BFD session. Then, the local system sends a BFD control packet with Your Discriminator being 0 to the remote system to negotiate over the BFD session.
  - Self learning the remote discriminator

    When one end of a BFD session receives a BFD control packet with Your Discriminator being 0, the BFD control packet is checked. If the packet matches the local BFD session,

the end learns the value of My Discriminator in the received BFD control packet to obtain the remote discriminator.

Managing the BFD Session

The BFD session has the following status:

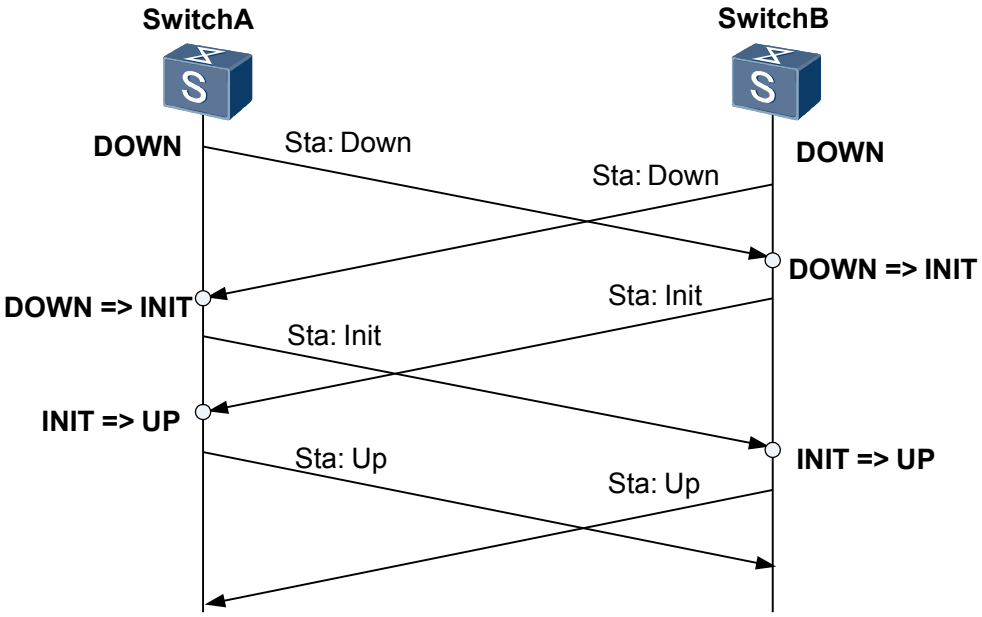
- Down: indicates that a BFD session is in the Down state or has just been set up.
- Init: indicates that the local system can communicate with the remote system, and the local system expects a BFD session to go Up.
- Up: indicates that a BFD session is set up successfully.
- AdminDown: indicates that a BFD session is in the administratively Down state.

The session status is conveyed in the State field of a BFD control packet. The system changes the session status based on the local session status and the received session status of the peer.

When a BFD session is to be set up or deleted, the BFD state machine implements a three-way handshake to ensure that the two systems are aware of the status change.

Figure 6-1 shows the transition process of the state machine in establishment of a BFD session.

Figure 6-1 Establishment of a BFD session



1. Switch A and Switch B enable BFD state machines respectively. The initial status of BFD state machines is Down. Switch A and Switch B send BFD control packets with the State field being Down. In the static configuration of a BFD session, Your Discriminator in the BFD control packet is specified manually. In dynamic establishment of a BFD session, Your Discriminator is 0.
2. After receiving the BFD packet with the State field being Down, Switch B switches the session status to Init and sends the BFD packet with State field being Init.

3. After the local BFD session status of Switch B changes to Init, Switch B no longer processes the received BFD packets with the State field being Down.
4. The status change of the BFD session on Switch A is the same as the status change of the BFD session on Switch B.
5. After receiving the BFD packet with the State field being Init, Switch B changes the local session status to Up.
6. The status change of the BFD session on Switch A is the same as the status change of the BFD session on Switch B.

6.3.1 BFD for IP

A BFD session is established on an IP link to fast detect faults.

BFD can detect single-hop and multi-hop IP links.

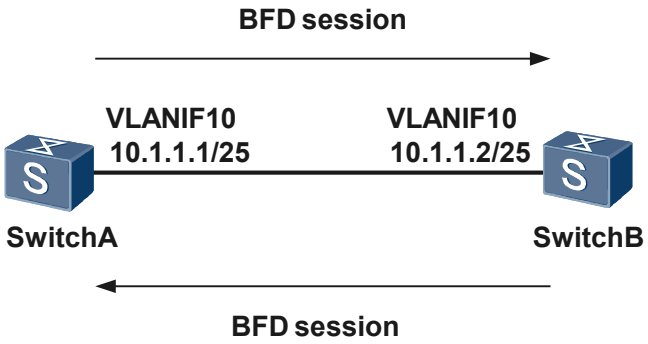
- Single-hop BFD is a mechanism that detects IP route connectivity between directly-connected systems. The single hop refers to an IP hop. Between these two systems detected by the single-hop BFD session, only one BFD session can be set up on a specified interface enabled with a specified data protocol.
- Multi-hop BFD is a mechanism that detects any paths between systems. A path may span multiple hops or paths may partially overlap.

Application Environment

Typical Application 1

Figure 6-2 shows that a BFD session detects a single-hop path between devices and the BFD session is bound to the outgoing interface.

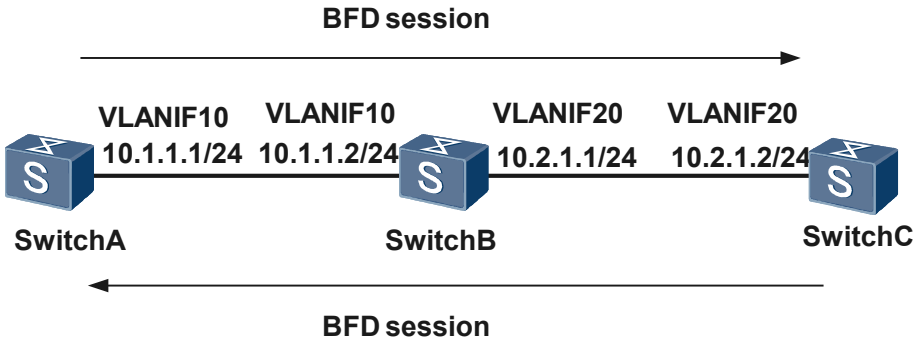
Figure 6-2 Networking diagram of single-hop BFD for IP



Typical Application 2

Figure 6-3 shows that a BFD session detects a multi-hop path between Switch A and Switch C and the BFD session is bound to the peer IP address but not the outgoing interface.

Figure 6-3 Networking diagram of multi-hop BFD for IP



6.3.2 BFD for USR

BFD for Unicast Static Route (USR) is used to detect IPv4 USRs. After a BFD session is bound to an IPv4 USR, faster detection of links can be performed.

Unlike dynamic routing protocols, USRs do not have a detection mechanism. If a fault occurs on a network, an administrator needs to handle it. In BFD for USR, BFD sessions are bound to IPv4 USRs in a public network and are used to detect the link status of the IPv4 USR.

One BFD session is bound to one IPv4 USR. When a BFD session detects a fault (for example, the link changes from Up to Down) on a link of the USR, BFD reports the fault to the routing management module. Then, the RM sets the USR as "inactive" (indicating that the route is unavailable and deleted from the IP routing table.)

When the BFD session bound to the USR is successfully set up or the link of the USR recovers from the fault (that is, the link changes from Down to Up), BFD reports the event to the RM and the RM sets the USR as "active" (indicating that the route is available and added to the IP routing table).

6.3.3 BFD for OSPF

A link fault or the change of topology may lead to rerouting in a network. The short-duration convergence of a routing protocol is important for the improvement of availability of the network. A feasible solution is to fast detect the fault and notify the fault to the routing protocol immediately.

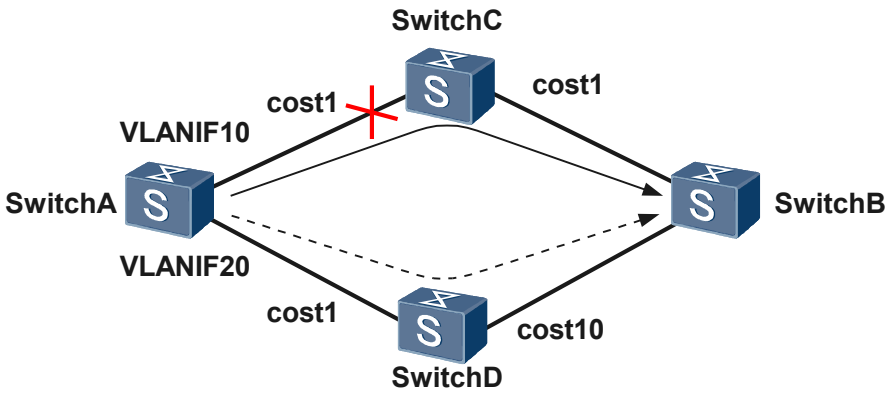
In BFD for OSPF, OSPF is associated with a BFD session. The BFD session fast detects a link fault and notifies OSPF of the fault. In this manner, OSPF speeds up the response to the change of the network topology.

Table 6-1 shows statistics of convergence speeds when OSPF is and is not associated with a BFD session.

Table 6-1 Statistics of OSPF convergence speeds

Associated with BFD or Not	Link Fault Detection Mechanism	Convergence Speed
Not associated with BFD	Timeout of the OSPF Hello keepalive timer	At the second level
Associated with BFD	BFD session in the Down state	At the millisecond level

Figure 6-4 Networking diagram of BFD for OSPF



As shown in **Figure 6-4**, Switch A sets up OSPF neighbor relationships with Switch C and Switch D. The outgoing interface on Switch A is connected to Switch B through Switch C. When the neighbor state is Full, BFD is notified of the status and starts to set up a BFD session.

1. When a fault occurs on the link between Switch A and Switch C, the BFD session detects the fault and then notifies Switch A.
2. Switch A processes the event that the neighbor goes Down and then recalculates routes. Then, the outgoing interface changes to be on Switch A that is connected Switch B through Switch D.

6.3.4 BFD for IS-IS

Generally, the interval for the Intermediate System to Intermediate System (IS-IS) protocol to send Hello messages is 10 seconds. The interval for notifying that a neighbor is Down when the neighbor fails, is three times the interval for sending Hello messages. If a device does not receive Hello messages from its neighbor before the neighboring device fails, the device deletes the neighbor. That is, the device detects neighbor faults in seconds. This leads to the loss of a large number of packets in a high-speed network.

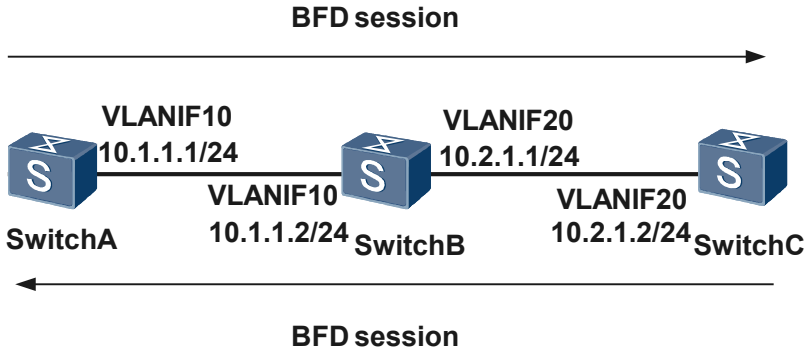
In BFD for IS-IS, the establishment of a BFD session is dynamically triggered by IS-IS but not configured manually. When detecting a fault, the BFD session notifies IS-IS of the fault through the RM. Then, IS-IS processes the event that the neighbor goes Down, fast sends the link state PDU (LSP), and performs the partial route calculation (PRC). In this manner, IS-IS routes fast converge.

The interval for BFD to detecting faults can be at the millisecond level. Instead of replacing the Hello mechanism of IS-IS, BFD works with IS-IS to detect the adjacency fault faster. In addition, BFD notifies IS-IS to recalculate routes. This ensures correct packet forwarding.

The RM achieves the interaction between the ISIS protocol and the BFD module. Through the RM, IS-IS notifies BFD of dynamically setting up or deleting BFD sessions. In addition, the BFD event message is delivered to IS-IS through the RM.

Application Environment

Figure 6-5 Networking diagram of BFD for IS-IS



After BFD is enabled on Switch A, Switch B, and Switch C, when a fault occurs on the link between Switch A and Switch B, the BFD session can fast detect the fault and notify IS-IS through the RM. Then, IS-IS sets the neighbor status to Down to trigger the IS-IS topology calculation. In addition, IS-IS updates LSPs to ensure that Switch C, that is, Switch B's neighbor, can receive the updated LSPs from Switch B in time. In this manner, the network topology fast converges.

6.3.5 BFD for VRRP

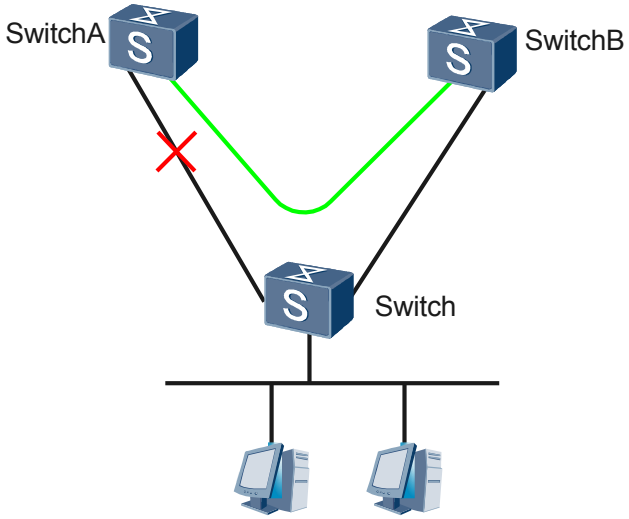
BFD can notify the interface board of detected faults to speed up the Virtual Router Redundancy Protocol (VRRP) master/backup switchover in the following cases:

- Faults occur on the interfaces on which VRRP backup groups are created.
- The master device and backup device are not directly connected.
- The master device and the backup device are directly connected, but a transmission device exists on the link between the master device and the backup device. The configurations are different according to whether the transmission device supports BFD:
  - If the transmission device is a switch or a device that supports BFD, you can configure two types of BFD sessions to detect links. One type is Peer BFD, which detects the link between the master device and the backup device; the other type is Link BFD, which detects the link between the transmission device and the master device (or the backup device).
  - If the transmission device does not support BFD, either of the following BFD sessions for VRRP can be created to implement fast swithover:
    - A peer BFD session can be created on the master and backup devices in the VRRP backup group to detect the link between the two devices. After the peer BFD session

goes Down, the backup device preempts to be the master device, in which case two master devices may coexist. These two master devices send gratuitous ARP packets based on which the system determines which link fails.

The BFD session detects connectivity of actual IP addresses between the backup device and the master device. If the communication is abnormal, the backup device considers that the master device is Down and a backup device becomes the master device. VRRP tracks the status of a BFD session to fast perform the master/backup switchover within 50 milliseconds.

Figure 6-6 Networking diagram of VRRP tracking the status of a BFD session



As shown in **Figure 6-6**, VRRP is enabled on Switch A and Switch B. Switch A functions as the master device and Switch B functions as the backup device. The user traffic is transmitted through Switch A. A BFD session is established between Switch A and Switch B. VRRP backup group tracks the status of the BFD session. When the status of the BFD session changes, the priority of the backup device is changed and then VRRP master/backup switchover is performed.

When a BFD session detects a link fault between Switch A and Switch B, a Down event is notified to VRRP. Then, the priority of the VRRP backup group on Switch B is increased to be higher than the priority of the VRRP backup group on Switch A. As a result, the VRRP master/backup switchover is performed. Switch B becomes the master device and the subsequent user traffic is forwarded through Switch B.

6.3.6 BFD for PIM

Normally, if the current Designate Router (DR) is faulty in the shared network segment, other Protocol Independent Multicast (PIM) neighbors participate in DR election after the neighbor relationship times out. The period of interruption of multicast data transmission, usually in seconds, is not shorter than the timeout period of the neighbor relationship.

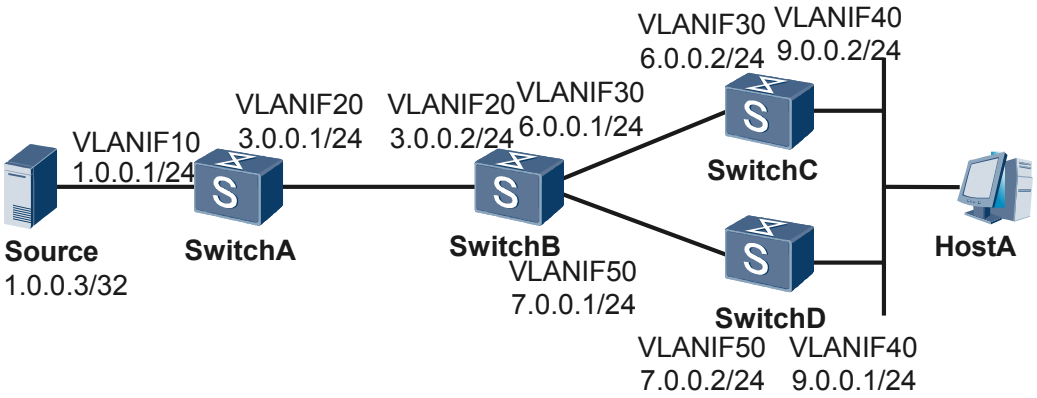
The characteristic of BFD for PIM is that fast fault detection can be performed, and the fault is notified to the PIM module within milliseconds, but not after the neighbor relationship times out, to trigger DR election. This greatly shortens the interruption of multicast data transmission and improves reliability of multicast networks.

BFD for PIM can fast detect faults on interfaces of Assert Winner and thus is also applicable to the Assert election in a shared network segment

Application Environment

BFD for PIM can fast detect link faults in a shared network segment consisting of non-broadcast multi-access (NBMA) interfaces and broadcast interfaces.

Figure 6-7 Networking diagram of BFD for PIM



As shown in **Figure 6-7**, Host A needs to receive data flows from the source. After the multicast PIM is correctly configured, data flows are transmitted as follows:

- In the case that Switch C and Switch D are not configured with DR priorities, Switch C functions as a DR at the Receiver side. Then, the data flows are sent by the source, forwarded through Switch A, Switch B, and Switch C, and received by Host A.
- PIM BFD is enabled on VLANIF30 interface of Switch C and VLANIF50 interface of Switch D. When a fault occurs on VLANIF30 of Switch C, Switch D immediately detects the fault on the other end of the link. Then, a new round of DR election is triggered. As a result, data flows from the source are switched to Switch A, Switch B, and Switch D, and then received by Host A. This greatly shortens the interruption of multicast data transmission. Note: By default, no DR priority is set. In this case, the switch with a higher IP address functions as the DR.

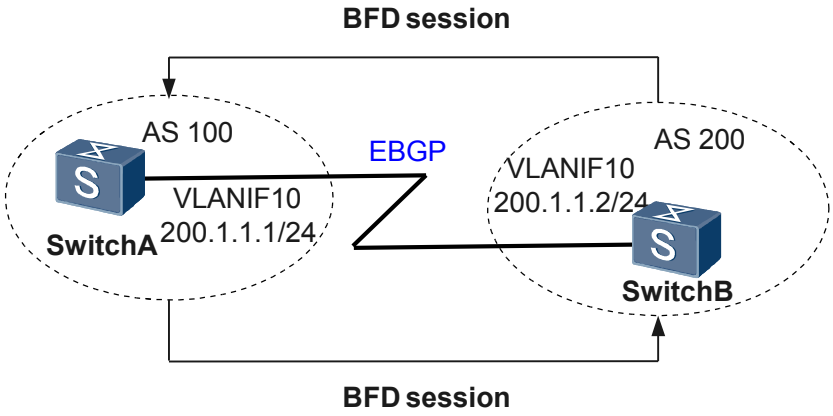
6.3.7 BFD for BGP

In the fault detection mechanism of the Border Gateway Protocol (BGP), BGP periodically sends Keepalive messages to the peer to detect the status of the neighbor. In BFD for BGP, the detection lasts more than one second. Therefore, when the data is transmitted at gigabit rates, a large amount of data is discarded, which cannot meet the requirement for high reliability of carrier-class networks.

Therefore, BFD for BGP is developed. The BFD session can fast detect a fault on a link between BGP peers and notify BGP. In this manner, BGP routes fast converge.

Application Environment

Figure 6-8 Networking diagram of BFD for BGP



As shown in **Figure 6-8**, Switch A belongs to AS 100 and Switch B belongs to AS 200. Switch A and Switch B are directly connected through the External Border Gateway Protocol (EGBP). A BFD session is established to detect the BGP neighbor relationship between Switch A and Switch B. When the link between Switch A and Switch B is faulty, the BFD session can fast detect the default and notify BGP.

6.3.8 Multicast BFD

Multicast BFD fast detects link faults by detecting connectivity of links between Layer 3 interfaces that do not have IP addresses.

Multicast BFD is enabled on switchs between which the link is to be detected, and then sends multicast control packets through the network layer. The local switch sends a multicast packet. If the link is reachable, the interface of the peer switch can receive the multicast packet and forward the packet to the BFD module. Then, the BFD session detects that the link is normal. On Layer 2 trunk links, the multicast packets can be directly forwarded to the link layer, without IP forwarding, to fast detect link connectivity. The IP address used by the BFD module is the default known multicast IP address. Any interface that receives packets containing the IP address forwards the packets to the BFD module, and the IP forwarding is complete.

Application Environment

Figure 6-9 Networking diagram of multicast BFD

As shown in **Figure 6-9**, multicast BFD can fast detect connectivity of a link between interfaces. A BFD session is established on Switch A and Switch B. The BFD session sends a packet with the destination address being the default multicast IP address to XGE 0/0/1 to detect the single-hop link. In this manner, BFD can fast detect connectivity of the link between interfaces.

6.3.9 BFD for PIS

BFD for process interface status (PIS) is a simple mechanism, in which the BFD session is associated with the interface status. This improves the sensitivity of interfaces in detecting a link fault and minimizes the impact of faults on non-directly-connected links.

In BFD for PIS, after detecting a link fault, a BFD session immediately sends a Down message to the corresponding interface. Then, the interface enters the BFD Down status, which is equal to the Down state of the link protocol. In the BFD Down state, only BFD packets can be normally processed and thus interface can fast detect the link fault.

Each BFD session that needs to be associated with an interface is configured as multicast BFD session and the interface type is set. In this manner, the forwarding of BFD packets are independent of the IP attribute on interfaces.

Application Environment

Figure 6-10 Networking diagram of BFD for PIS

As shown in **Figure 6-10**, a BFD session is established on Switch A and Switch B. The BFD session sends a packet with the source address being the default multicast IP address to XGE 0/0/1 to detect the single-hop link. After BFD for PIS is enabled, when BFD detects a link fault, the BFD session sends a Down message to the corresponding interface and then the interface enters the BFD Down state.

6.4 Terms and Abbreviations

Abbreviation

Abbreviation	Full Spelling
ISIS	Intermediate System-Intermediate System

Abbreviation	Full Spelling
BFD	Bidirectional Forwarding Detection
VC	Virtual Circuit
VLL	Virtual Leased Line
AC	Attachment Circuit
PE	Provider Edge Router
CE	Customer Edge Router
OSPF	Open Shortest Path First
TE	Traffic Engineer
CSPF	Constraint Shortest Path First
VRRP	Virtual Router Redundancy Protocol
L2VPN	Layer 2 virtual private network
PW	Pseudo Wire
MPLS	Multi Protocol Label Switching

7 VRRP

About This Chapter

7.1 Introduction to VRRP

Definition

The Virtual Router Redundancy Protocol (VRRP) is a fault tolerant protocol, and it groups several switches into a virtual router. If the next hop switch of a host is defective, VRRP uses a mechanism to switch traffic to another switch. This ensures continuity and reliability of communications.

The basic concepts related to VRRP are as follows:

- VRRP Router: It is the router running VRRP and it may join one or multiple virtual routers.
- Virtual router: It is an abstract device managed by VRRP, also called a VRRP backup group. A virtual router functions as a default gateway on a shared local area network (LAN). A virtual router is identified by a virtual router identifier and has a set of virtual IP addresses.
- Virtual IP address: It is the IP address of a virtual router. A virtual router is manually assigned one or multiple virtual IP addresses.
- IP address owner: It is a VRRP that uses an IP address of a virtual router as an actual interface address. When working normally, the VRRP router responds to packets destined for the virtual IP address, such as ping packets and TCP packets.
- Virtual MAC address: It is a MAC address that is generated according to a virtual router ID. A VRRP virtual router has a virtual MAC address in the format of 00-00-5E-00-01-{VRID},and a VRRP6 virtual router has a virtual MAC address in the format of 00-00-5E-00-02-{VRID}. A virtual router responds to an Address Resolution Protocol (ARP) request using the virtual MAC address rather than interface's actual MAC address.
- Primary IP address: It is selected from one of physical interfaces' IP addresses. It is usually the first configured IP address. The primary IP address functions as the source IP address in VRRP broadcast packets.
- Master Router (virtual router master): It is a VRRP router that forwards packets to the virtual IP address, or responds to ARP requests. When an IP address owner is available, it usually functions as the master router.
- Backup Router (virtual router backup): It is a set of VRRP routers that do not forward packets. If the master router is defective, the backup routers will compete to be the new master router.
- Preemption mode: It means that the backup router automatically becomes the master router if the priority of a backup router is higher than the priority of the current master router.

Purpose

With the growth of Internet, networks are required for higher reliability. For LAN users, it is important to be in contact with external networks at any time.

Generally, all hosts within an internal network are configured with one default route destined for an egress gateway to communicate with external networks. If the egress gateway is defective, communication between these hosts and external networks will be interrupted.

A common method to improve reliability of the system is to configure multiple egress gateways. The solution to route selection among these gateways must be figured out in the case that hosts in a LAN do not support dynamic routing protocols.

Therefore, the Internet Engineering Task Force (IETF) puts forward VRRP. VRRP provides reliability for hosts on a LAN to access external networks. VRRP provides the following functions:

- Master/backup mode: VRRP provides the IP address backup function in master/backup mode. A virtual router must be set up, consisting of a master router and multiple backup routers. The master router and backup routers form a backup group. Normally, the master router transmits all services. When the master router fails, a backup router takes over the services.
- VRRP load balancing: In VRRP local balancing, multiple virtual routers transmit service at the same time. Load balancing can be performed only on two or more backup groups on multiple routers, rather than on one VRRP backup group. In load balancing, the master status is load balanced in VRRP backup groups and each device transmits a part of services.
- VRRP tracking interface status: Each VRRP backup group can track the status of all interfaces that are bound to a VRRP backup group. If an interface fails, the router with the highest priority will be re-selected as the master router.
- Pinging the virtual IP address: In VRRP, virtual IP addresses can be pinged through command lines.
- VRRP security: Different authentication modes and authentication keys can be set in VRRP packet headers in networks at different security levels.
- VRRP fast switchover: VRRP tracks the status of a BFD session to perform a fast switchover in milliseconds.

7.2 References

The following table lists the references of this document.

Document	Description	Remarks
RFC 2281	Hot Standby Router Protocol (HSRP)	-
RFC 2338	Virtual Router Redundancy Protocol (version number One 1998)	-
RFC 2787	Definitions of Managed Objects for the Virtual Router Redundancy Protocol	-
RFC 3768	Virtual Router Redundancy Protocol (version number Two 2004)	-
RFC 5798	Virtual Router Redundancy Protocol Version 3 for IPv4 and IPv6	-

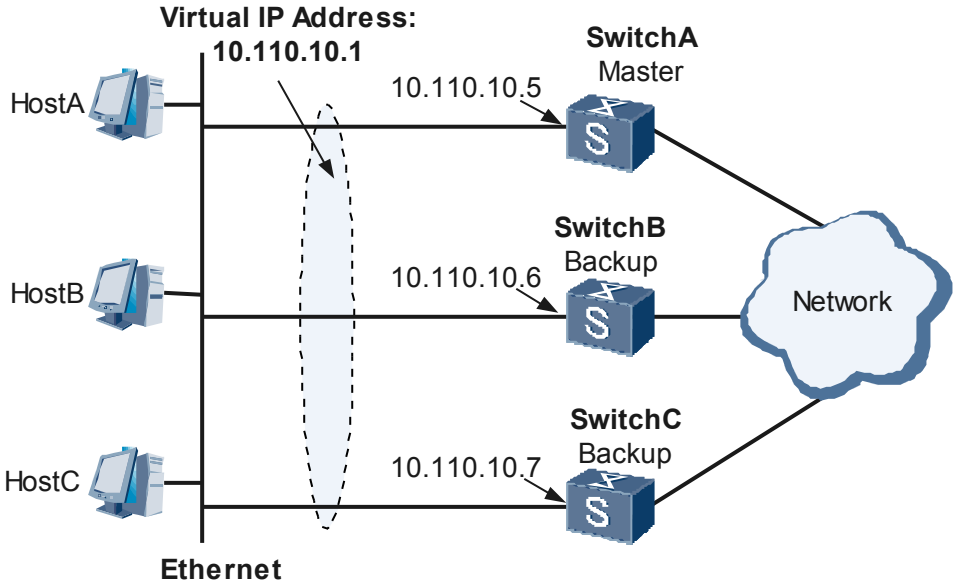
7.3 Principles

VRRP combines a group of routing devices in a LAN into a backup group that functions as a virtual router. In the LAN, hosts need to obtain only the IP address of the virtual router rather than the IP address of a specific device in the backup group. The hosts set the IP address of the virtual router as the default gateway. Then, the hosts can communicate with an external network through the virtual gateway.

VRRP dynamically associates the virtual router with a physical device that transmits services. When the device fails, another device is selected to transmit services. The switchover is

transparent to users and thus the internal network and the external network can communicate without interruption.

Figure 7-1 Schematic diagram of a virtual router



As shown in **Figure 7-1**, the implementation of the virtual router is as follows:

- Switch A, Switch B, and Switch C form a VRRP backup group that functions as a virtual router. The IP address of the virtual router is 10.110.10.1. The virtual IP address can be specified or borrowed from an interface of a device in this VRRP backup group.
- The actual IP addresses of Switch A, Switch B, and Switch C are 10.110.10.5, 10.110.10.6, and 10.110.10.7, respectively.
- Hosts of a LAN need only to set the default route to 10.111.10.1 rather than a physical interface address of a specific device.

Hosts communicate with external networks through this virtual gateway. The working mechanism of the virtual router is as follows:

- The master device is selected according to device priorities:
  - The device with a higher priority is elected as the master device.
  - If two devices have the same priority and one of them is the master device, the backup device will remain in the backup state. If the two devices with the same priority compete for becoming the master device, the device with a greater interface IP address will be selected as the master device.
- Other devices function as backup devices and track the status of the master device at any time.
  - When working normally, the master device sends a VRRP multicast packet at intervals of Advertisement_Interval to notify backup devices in the backup group that the master device works normally.
  - In a VRRP group consisting of a master device and a backup device, when the backup device does not receive packets from the master device within the period of

Master_Down_Interval, the backup device transits itself to be the master device. In a VRRP group consisting of a master device and multiple backup devices, when the backup devices do not receive packets from the master device within the period of Master_Down_Interval, multiple backup devices may transit themselves to be master devices in a short period. The devices then compare the priorities in the received VRRP packets with local priorities, and the device with a higher priority is selected to be the master device. After a backup device is switched to be the master device, it sends gratuitous ARP packets to refresh MAC entries on the switches. In this manner, user traffic is switched to the master device. The entire process is transparent to users.

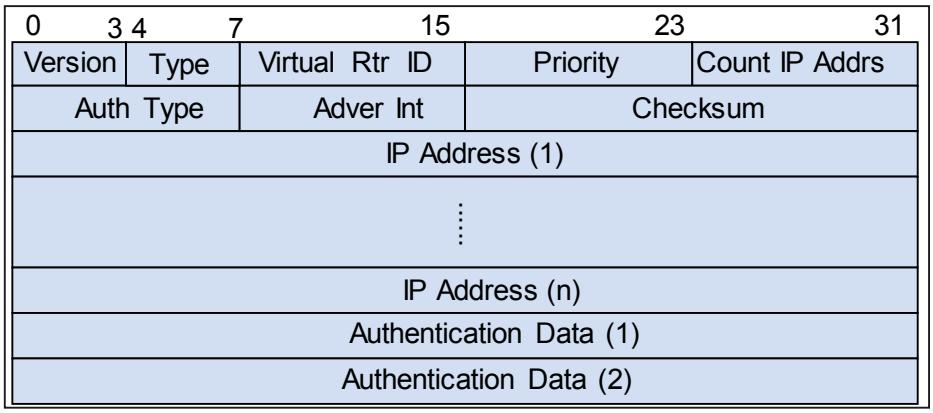
The preceding analysis shows that in VRRP, the hosts do not need additional operations and can communicate with external networks normally even when a device fails.

VRRP Packet Format

VRRP packets are used to notify the priority and the status of the master device to all VRRP routers associated with the same virtual router ID.

VRRP packets are encapsulated in IP packets and sent to the IPv4 multicast address assigned to VRRP. In the IP packet header, the source address is the primary IP address of the interface sending the packet, but not the virtual address or secondary address. The destination multicast address is 224.0.0.18. The TTL value is 255, and the protocol number is 112. **Figure 7-2** shows the VRRP packet format.

Figure 7-2 VRRP packet format



The descriptions of each field are as follows:

- Version: indicates the version number of the protocol. The VRRP protocol number is 2.
- Type: indicates the type of VRRP Advertisement packets. The value is fixed at 1.
- Virtual Rtr ID: indicates the virtual router identifier. This field identifies the virtual router. The value ranges from 1 to 255.
- Priority specifies the priority of the sending VRRP router for the virtual router. The priority value ranges from 0 to 255, and the valid range for users is 1 to 254. The priority value 0 indicates that the switch stops joining VRRP. This is used to trigger backup switchs to quickly switch to the master without waiting for timer expiration. The priority value 255 is reserved for the IP address owner. By default, the priority value is 100.

- Count IP Addrs: indicates the number of virtual IP addresses contained in this VRRP advertisement packet.
- Authentication Type: indicates the authentication type that is in use. The authentication types defined in the protocol are as follows:
  - 0: Non Authentication
  - 1: Simple Text Password
  - 2: IP Authentication Header

  NOTE

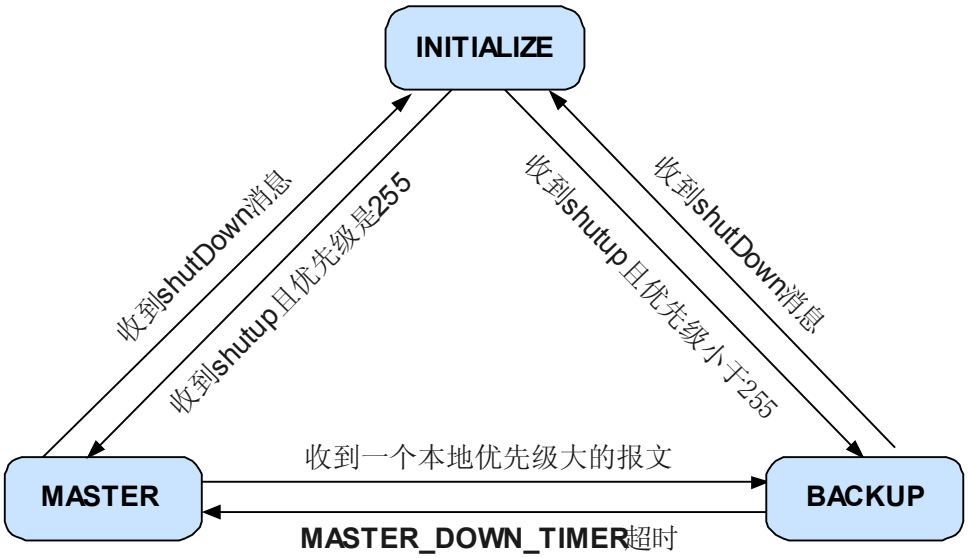
  Currently, the S6700 supports the following authentication modes:
  - Simple Text Password: Plain text authentication
  - IP Authentication Header: MD5 authentication
- Advertisement interval: indicates the interval for sending the advertisement packet. The default value is 1 second.
- Checksum: indicates the packet checksum.
- IP Address: indicates the virtual addresses of the virtual router. The number of addresses is specified in the Count IP Addrs field.
- Authentication data indicates the authentication key. Currently, this field is used only in plain text authentication mode and MD5 authentication mode. In other authentication modes, this field is filled with 0.

State Machine

VRRP defines three types of state machines: Initialize, Master, and Backup. Only the switch in the Master state can forward packets that are sent to the virtual IP address.

Figure 7-3 shows the transition of the state machines.

Figure 7-3 Transition of the VRRP state machine



Initialize: The switch is in the Initialize state when being started. If a Startup message is received, the switch changes to the Backup state or the Master state. The switch, which is the IP address

owner, changes to the Master state directly. In this state, the switch does not process the VRRP packets.

Master: In the Master state, the switch must do the following:

- Send the VRRP packets periodically.
- Respond to ARP requests for the virtual IP address with the virtual MAC address.
- Forward IP packets with a destination MAC address as the virtual MAC address.
- Accept the IP packet with the destination IP address as the virtual IP address if the switch is the virtual IP address owner. Otherwise, the IP packet is discarded.
- Change to the Backup state if the priority in the received packet is greater than the local priority.
- Change to the Initialize state when the interface is shut down.

Backup: In the Backup state, the switch must do the following:

- Accept the VRRP packets sent by the master switch and judge whether the master is normal.
- Do not respond to the ARP requests for the virtual IP address.
- Discard IP packets with the destination MAC address as the virtual MAC address.
- Discard IP packets with the destination IP address as the virtual IP address.
- Discard the packets with lower priority. The timer is not reset. If the packets with equal priority are received, the switch resets the timer and does not compare the IP addresses.
- Change to the master device when receiving the event that MASTER_DOWN_TIMER times out.
- Change to the Initialize state when receiving the Shutdown event from the interface.

7.3.1 Master/Backup Mode

VRRP provides the IP address backup function in master/backup mode. That is, a virtual router must be set up, consisting of a master device and multiple backup devices. The master device and backup devices form a backup group.

- Normally, the master device transmits all services.
- When the master device fails, a backup device takes over the services.

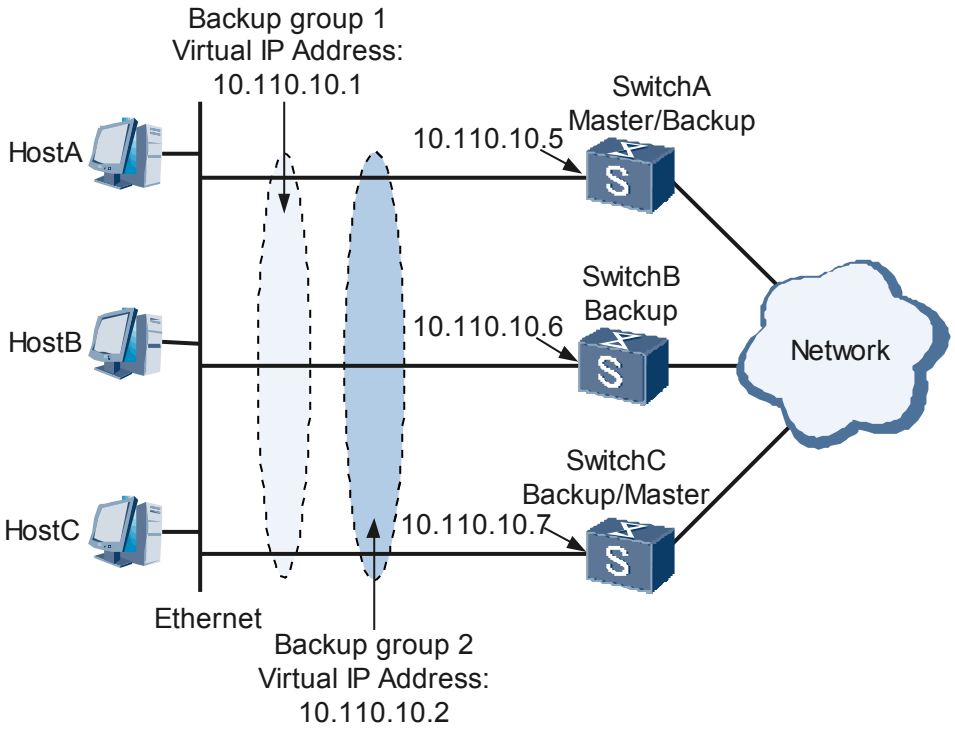
7.3.2 VRRP Load Balancing

A device can function as a backup device in multiple VRRP backup groups. Load balancing is performed over configured multiple virtual routers. In load balancing mode, multiple virtual routers transmit services at the same time; therefore, two or more backup groups must be set up.

The load balancing mode has the following characteristics:

- Each backup group consists of a master device and multiple backup devices.
- The master devices of backup groups can be different.
- A device can join multiple backup groups and obtain different priorities.

Figure 7-4 VRRP in load balancing mode



As shown in **Figure 7-4**, two backup groups are configured, that is, Backup group 1 and Backup group 2.

- Switch A is the master in Backup group 1 and the backup in Backup group 2.
- Switch B functions as the Backup in both Backup group 1 and Backup group 2.
- Switch C is the master in Backup group 2 and the backup in Backup group 1.
- Backup groups 1 and 2 are gateways for different hosts.

In this manner, load balancing of data traffic and the mutual backup can be implemented.

7.3.3 VRRP Tracking Interface Status

VRRP can track the status of all interfaces. When the interface that is tracked by VRRP goes Up or Down, the priority of the device automatically changes by a certain value. The order of device priorities in the backup group thus changes, and then the VRRP devices re-compete with each other to be the master device.

A VRRP backup group tracks a maximum of eight interfaces in Increase mode or Reduce mode.

- In Increase mode, when a tracked interface goes Down, the priority of the VRRP backup group increases by a set value.
- In Reduce mode, when a tracked interface goes Down, the priority of the VRRP backup group decreases by a set value.

  The Reduce mode takes effect on both master and backup devices.

For information about the typical application environment, see the section "VRRP Tracking Interface Status."

7.3.4 VRRP Fast Switchover

The bidirectional forwarding detection (BFD) mechanism is implemented to rapidly detect and monitor connectivity of network links or IP routes. VRRP tracks the status of a BFD session to perform fast master/backup switchover within one second.

BFD can notify the interface board of detected faults to speed up VRRP master/backup switchover in the following cases:

- Faults occur on the interfaces where VRRP backup groups are created.
- The master device and the backup device are not directly connected.
- The master device and the backup device are directly connected; however, other transmission devices exist on the link between the master device and the backup device.

The BFD session detects connectivity between the backup device and the master device according to the actual IP address. If the communications are abnormal, the backup device considers that the master device is Down and becomes the master device. A backup device becomes the master device in the following cases:

- When the back-to-back connection of two devices is disconnected, the backup device becomes the master device to transmit the upstream traffic.
- When the connection of two devices is terminated, the backup device becomes the master device and then transmits the upstream traffic in the following cases:
  - The master device is restarted.
  - The link between the master device and the switch is disconnected.
  - The switch that is connected to the master device is restarted.

The VRRP fast switchover requires the following:

- On the backup devices, the interfaces detected by BFD sessions must be connected to the master device.
- When the master device is faulty, the priority of a backup device is increased and greater than the priority of the faulty master device. Then, the backup device can fast transit to be the master device.

7.3.5 Pinging the Virtual IP Address

The RFC 3768 does not define whether virtual IP addresses should be successfully pinged. Pinging the virtual IP addresses of VRRP backup groups can facilitate the monitoring of virtual routers. Virtual routers, however, are prone to be attacked by Internet Control Message Protocol (ICMP) packets. A command is provided for you to ping a virtual IP address.

7.3.6 VRRP Security

Different authentication modes and authentication keys can be set in VRRP packet headers in networks at different security levels.

In a secure network, the default setting can be adopted. That is, the device does not authenticate the sent or received VRRP packets. All received VRRP packets are considered as valid. In this case, no authentication key needs to be set.

VRRP provides simple text authentication and MD5 authentication for networks that are vulnerable to attacks. In simple text authentication mode, a string of 1 to 8 characters can be configured as the authentication key. In MD5 authentication mode, a string of 1 to 8 characters

in plain text or a string of 24 characters in encrypted text can be configured as the authentication key.

7.3.7 VRRP Smooth Switching

After the AMB/SMB switchover is performed on the master device, the new AMB can work normally only after a period varying with the device and configuration. During this period, the master device may not process VRRP packets normally; the backup devices may not receive VRRP broadcast packets, and thus a backup device preempts to be the master device. Then, the new master device sends a gratuitous ARP packet to the virtual IP address of each virtual router to notify the related bound modules of the status change. In preemption mode, if the original master device has a higher priority, it can preempt to be the master device again after the AMB/SMB switchover. This causes twice jitters of the VRRP status, which affects service traffic.

In this case, the CEs enabled with VRRP need to support VRRP smooth switching to avoid the service traffic from being affected by the AMB/SMB switchover.

When the AMB and SMB on a device work normally, the master device in a VRRP backup group sends VRRP broadcast packets at intervals of Advertisement_Interval; the backup device determines whether the master device works normally by detecting the received broadcast packets.

During VRRP smooth switching, the master device cooperates with backup devices to ensure smooth transmission of services.

- To perform VRRP smooth switching, you must enable the function of learning the interval for sending VRRP packets on the master device and backup devices.
  - After being enabled with the learning function, the master device neither learns the interval for sending VRRP packets nor checks consistency of the intervals.
  - After receiving a packet from the master device, the non-master device checks the interval in the VRRP packets. If the received interval is different from the interval of the non-master device, the non-master device learns the interval and adjusts its own interval to be the same as the learned interval.
- Switch A is configured with VRRP smooth switching. After the AMB/SMB switchover is performed and the new AMB starts, VRRP saves the currently configured interval, adjusts the interval of the master VRRP backup group, and sends the VRRP smooth switching packet carrying the new interval to Switch B at the currently set intervals.
- After receiving the VRRP packet, Switch B finds that the interval carried in the received VRRP packet is inconsistent with the current interval. Switch B then adjusts the current interval to make it consistent with the interval carried in the received VRRP packet.
- After the AMB/SMB switchover, Switch A sends a VRRP Recovery packet carrying the interval set before the AMB/SMB switchover. Switch B then learns the interval again.

When perform VRRP smooth switching, notes the following:

- During VRRP smooth switching, the learning function takes precedence over the preemption function. That is, when the interval carried in the received packet is inconsistent with the current interval and the priority carried in the received packet is lower than the current priority, VRRP prefers the learning function and resets the timeout timer.
- VRRP smooth switching also depends on the system. If the system is quite busy since the AMB/SMB switchover and cannot schedule the operation of the VRRP module, VRRP smooth switching cannot take effect.

7.3.8 mVRRP

Naturally, the Management Virtual Router Redundancy Protocol (mVRRP) is an ordinary VRRP backup group with all features supported by the current device, whereas its unique character is as follows:

An mVRRP backup group can be bound to other service VRRP backup groups and determine the status of related service VRRP backup groups.

After ordinary VRRP backup groups are added to an mVRRP backup group, they do not need to send VRRP packets to determine the status. The mVRRP backup group sends VRRP packets to determine its status and the status of all its bound service VRRP backup groups. This reduces the bandwidth that VRRP packets occupy.

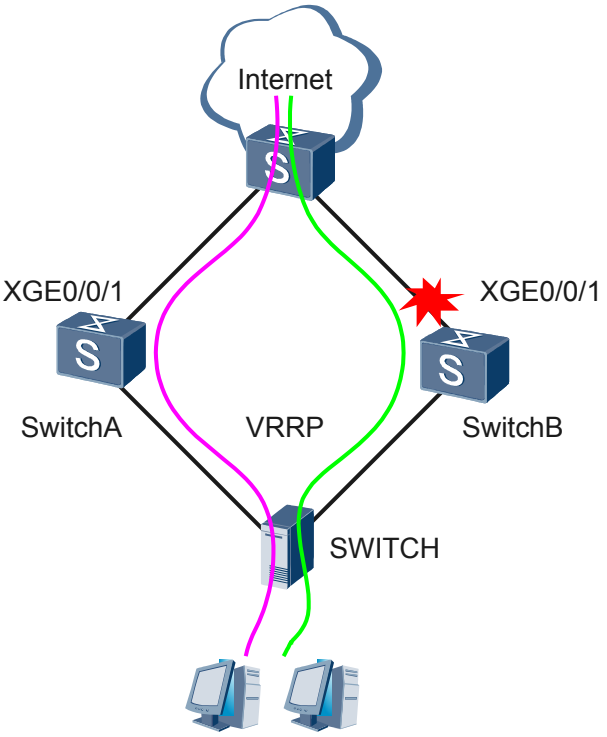
Although an mVRRP backup group can be bound to multiple service VRRP backup groups, the mVRRP backup group, functioning as a service VRRP backup group, cannot be bound to any other mVRRP backup group.

In a VPLS network, after PWs or service interfaces are bound to the mVRRP backup group, the mVRRP backup group can be associated with an mVSI. For information about the application environments, see the sections "mVRRP" and "VRRP in the ME Solution."

7.4 Application Environment

7.4.1 VRRP Tracking Interface Status

Figure 7-5 Networking diagram of VRRP tracking interface status



Solved problem: VRRP cannot detect status changes on interfaces that are not enabled with VRRP. In this case, when the outbound interface is faulty, VRRP cannot detect the fault, which causes service interruption.

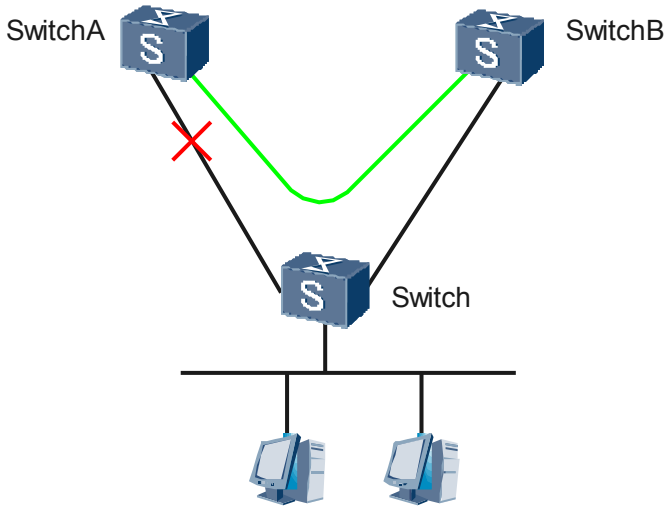
The configuration is as follows:

- VRRP is enabled to track specified interfaces.
- A VRRP backup group tracks a maximum of eight interfaces in Increase mode or Reduce mode.
- When the status of the interface tracked by VRRP changes, the VRRP backup group is notified of the change and then increases or decreases the VRRP priority to determine VRRP switchover.

As shown in **Figure 7-5**, Switch A and Switch B are enabled with VRRP. In addition, the priority of the VRRP backup group on Switch B is higher than the priority of the VRRP group on Switch A. Switch B tracks interface in Reduce mode. Switch B functions as the master device and the user traffic is sent by the master Switch B, as shown in dotted lines in **Figure 7-5**. Now, interface on Switch B connected to the Internet is faulty. The VRRP backup group that tracks XGE 0/0/1 in Reduce mode decreases the priority. Then, Switch A preempts to be the master device and receives user traffic and sends the traffic to the Internet.

7.4.2 VRRP Fast Switchover

Figure 7-6 Networking diagram of VRRP fast switchover



Solved problem: Traffic loss lasts a long time after a VRRP backup group detects a link fault.

The configuration is as follows:

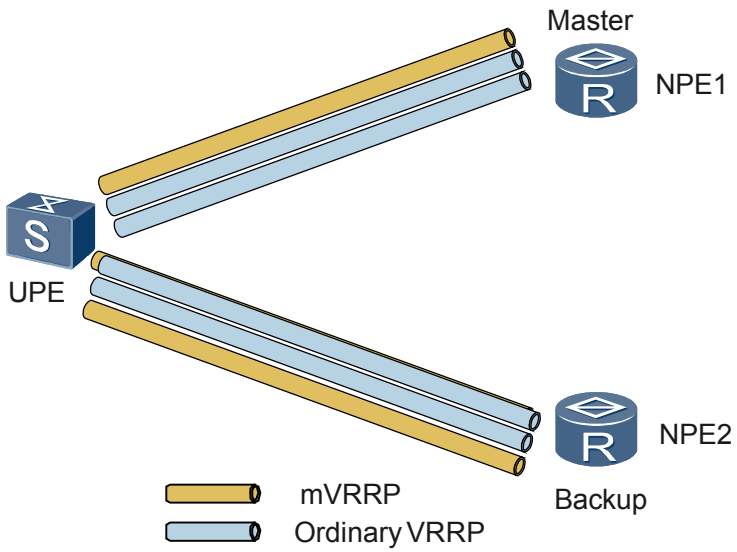
- After BFD that can detect faults in milliseconds is enabled, in the case that a fault occurs on the link between Switch A and Switch B or a remote host, the BFD session can fast detect the fault.
- After the VRRP backup group is configured to track the status of the BFD session, the BFD session can detect a link fault and notify VRRP of the link fault.

- Then, the VRRP backup group adjusts priorities according to the status notified by BFD, or performs fast switchover to start the preemption process.
- A VRRP backup group can track a maximum of eight BFD sessions.
- The VRRP master/backup switchover can be performed within 200 ms through the tracking of BFD sessions.

As shown in **Figure 7-6**, VRRP is enabled on Switch A and Switch B. In the VRRP backup group, Switch A functions as the master device and forwards user traffic and Switch B functions as the backup device. A BFD session is established between Switch A and Switch B. The VRRP backup group tracks the status of the BFD session. When the status of the BFD session changes, the priorities of the backup group are changed and then the master/backup switchover is performed. When a BFD session detects a link fault between Switch A and Switch, a Down event is notified to VRRP. Then, the priority of Switch B is increased to be higher than the priority of Switch A. As a result, the VRRP master/backup switchover is performed. Switch B becomes the master device and the subsequent user traffic is forwarded through Switch B.

7.4.3 mVRRP

Figure 7-7 Typical networking diagram of mVRRP



Solved problem: A great number of VRRP packets waste bandwidths and CPU resources.

The configuration is as follows:

- The mVRRP backup group and ordinary VRRP backup groups are set up on NPE 1 and NPE 2. The ordinary VRRP backup groups are bound to the mVRRP backup group and thus called a service VRRP backup group.
- The UPE does not sense the mVRRP backup group and service VRRP backup groups.

As shown in **Figure 7-7**, when an mVRRP backup group on NPE 1 changes from the Master state to the Backup state or the Initialize state, the mVRRP backup group notifies all its bound service VRRP backup groups to change their status to Backup. In this case, the mVRRP backup group on NPE 2 changes from the Backup state to the Master state, and it also notifies all its service VRRP backup groups to change their status to Master. When the mVRRP backup group

and the service backup groups change to the Master state, gratuitous ARP packets are broadcast and then the user traffic is switched to the new master backup group.

7.5 Terms and Abbreviations

Abbreviations

Abbreviation	Full Spelling
VRRP	Virtual Router Redundancy Protocol
ARP	Address Resolution Protocol
BFD	Bidirectional Forwarding Detection
L2VPN	Layer 2 virtual private network
PW	Pseudo Wire
VSI	Virtual Switching Instance
QinQ	802.1Q in 802.1Q
ME	Metro Ethernet
mVRRP	Management Virtual Router Redundancy Protocol
mVPLS	Management Virtual Private LAN Service
mVSI	Management Virtual Switching Instance