11 Layer 3 Forwarding

This topic describes the network layer (Layer 3) features implemented by the system.

11.1 IPv4

11.2 IPv6
Internet Protocol Version 6 (IPv6), also called IP Next Generation (IPng), is a second-generation protocol of the network layer protocol family.

11.3 ARP

11.4 ARP Proxy
ARP proxy is used to communication between users in different VLAN or in the same VLAN but isolated from each other by layer 3 forwarding.

11.5 DHCP Server

11.6 DHCPv6 Server

11.7 DHCP Relay
This section describes the implementation and configuration of Dynamic Host Configuration Protocol (DHCP) relay.

11.8 DHCP Proxy
This section describes the implementation and configuration of Dynamic Host Configuration Protocol (DHCP) proxy.

11.9 BFD

11.10 TWAMP
The Two-Way Active Measurement Protocol (TWAMP) is a technology that measures the round-trip performance of an IP network.

11.11 TWAMP Light
Two-Way Active Measurement Protocol (TWAMP) Light is a light version of TWAMP, which is defined in RFC 5357. TWAMP Light measures the round-trip performance of an IP network by using simplified control protocols to establish test sessions.

11.12 Routing
Routing is a common term used for describing the path through which the packets from a host in a network travel to a host in another network.

 Copyright © Huawei Technologies Co., Ltd.

11.1 IPv4

11.1.1 Introduction to IPv4

Definition

At the core of the TCP/IP protocol suite, Internet Protocol Version 4 (IPv4) works at the Internet layer in the TCP/IP model. This layer corresponds to the network layer in the OSI model. At the IP layer, information is divided into data units, and address and control information is added to allow datagrams to be routed.

IP provides unreliable and connectionless data transmission services. Unreliable transmission means that IP does not ensure that IP datagrams successfully arrive at the destination. IP only provides best effort delivery. Once an error occurs, for example, a router exhausts the buffer, IP discards the excess datagrams and sends ICMP messages to the source. The upper layer protocols, such as TCP, are responsible for resolving reliability issues.

Connectionless transmission means that IP does not maintain status information for subsequent datagrams. Every datagram is processed independently, meaning that IP datagrams may not be received in the same order they are sent. If a source sends two consecutive datagrams A and B in sequence to the same destination, each datagram is possibly routed over a different path to the destination, and therefore B may arrive ahead of A.

Purpose

IPv4 shields link layer protocol differences and provides a uniform standard for transmission at the network layer.

11.1.2 Principles to IPv4

11.1.2.1 ICMP

The Internet Control Message Protocol (ICMP) is an error-reporting mechanism and is used by IP or an upper-layer protocol (TCP or UDP). An ICMP message is encapsulated as a part of an IP datagram and transmitted through the Internet.

An IP datagram contains information about only the source and destination, not about all nodes along the entire path through which the IP datagram passes. The IP datagram can record information about all nodes along the path only when route record options are set in the IP datagram. Therefore, if a device detects an error, it reports the error to the source and not to intermediate devices.

When an error occurs during the IP datagram forwarding, ICMP reports the error to the source of the IP datagram, but does not rectify the error or notify the intermediate devices of the error. A majority of errors generally occur on the source. When an error occurs on an intermediate device, however, the source cannot locate the device on which the error occurs even after receiving the error report.

 Copyright © Huawei Technologies Co., Ltd.

Redirect Message

After a device receives an IP datagram that should have been sent to another device, the device sends an ICMP Redirect message to the sending device to redirect the IP datagram. Redirect messages help hosts that have a limited amount of routing information to gradually build up complete routing tables. A host may initially have only one default routing entry in the routing table. When an error occurs in the default routing entry, the default gateway notifies the host of generating another route to redirect the IP datagram and allows the host to modify the routing table.

ICMP Redirect messages can be generated and sent when all of the following conditions are met:

- The IP datagram is received and forwarded by the same interface. Note that both the physical interface and VLANIF interface are the same.
- The route used to send an IP datagram is not generated or modified through an ICMP Redirect message. That is, when a redirected route is used to send an IP datagram, no ICMP Redirect message is generated when the IP datagram is received.
- The route through which an IP datagram is transmitted cannot be the default route.
- The subnet mask of the source address in an IP datagram is the same as the subnet mask of the outgoing interface's address.

A large number of generated redirection ICMP packets will degrade the device performance. The device does not send redirection ICMP packets by default. The **ip redirect enable** command is used to enable the function of sending redirection ICMP packets.

TTL timeout message

During the process of forwarding or assembling an IP datagram, if the time-to-live (TTL) field in the IP datagram is zero, the receiving device sends a TTL timeout message to the source.

Port Unreachable Message

If a host or routing device receives a local UDP or TCP datagram but cannot find the corresponding process according to the destination port of the datagram, the host or routing device sends a Port Unreachable message to the source.

Destination Unreachable Message

If a network is unreachable, route selection fails. If a host is unreachable, message delivery fails. The source device can determine which address is unreachable by checking the IP header and the leftmost 64 bits in the original IP datagram (Internet Header + 64 bits of the Original Data Datagram field).

When a routing device forwards a message that meets the following conditions:

- No route is available for the destination address of the message.
- The message is not sent for the local host.

 Copyright © Huawei Technologies Co., Ltd.

The routing device will discard the message and return an ICMP Host Unreachable message to the source address to inform the source host to stop sending messages to this destination.

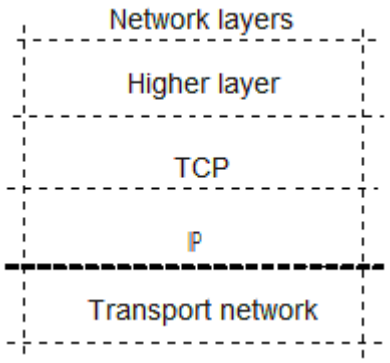
11.1.2.2 TCP

The Transmission Control Protocol (TCP) defined in RFC793 ensures high-reliability transmission between hosts. TCP provides reliable, connection-oriented, and full-duplex services for user processes. TCP transmits data through sequenced and nonstructural byte streams.

TCP is an end-to-end, connection-oriented, and reliable protocol. TCP supports multiple network applications. In addition, TCP assumes that the lower layer provides only unreliable datagram services, and it can run over a network of different hardware structures.

Figure 11-1 shows the position of TCP in a layered protocol architecture, where TCP is above IP. TCP can transmit variable-length data through IP encapsulation. IP then performs data fragmentation and assembly and transmits the data over multiple networks.

Figure 11-1 TCP in the layered protocol architecture



TCP works below applications and above IP. Its upper-layer interface consists of a series of calls similar to the interrupt call of an operating system.

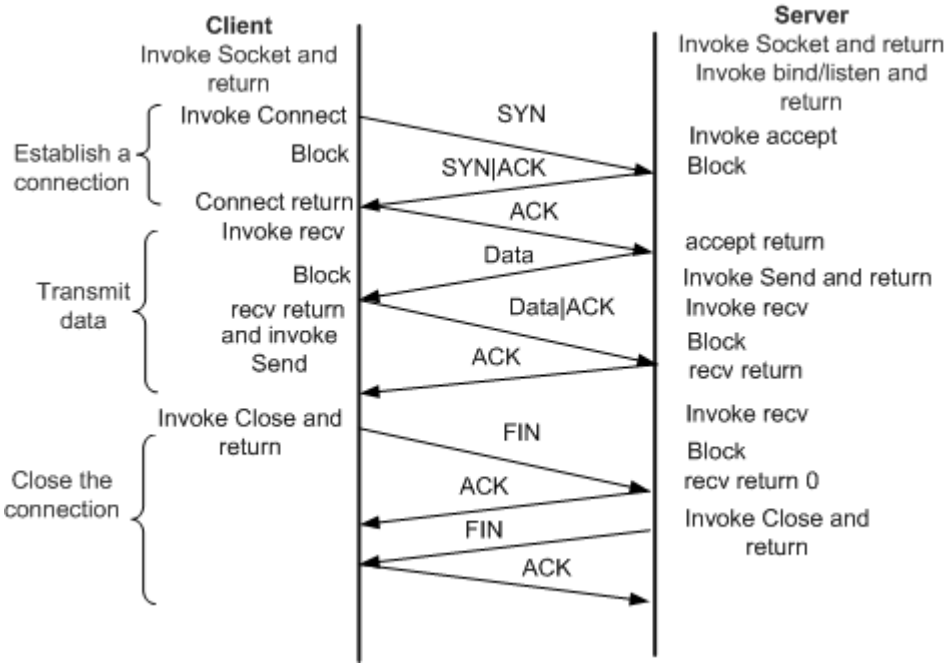
TCP can asynchronously transmit data of upper-layer applications. The lower-layer interfaces are assumed as IP interfaces. To implement connection-oriented and reliable data transmission over unreliable networks, TCP must provide the following functions:

- Ensures reliability and flow control.
- Provides multiple interfaces for upper-layer applications.
- Provides data for multiple applications.
- Ensures connections.
- Ensures communication security.

Figure 11-2 shows the process of setting up and tearing down a TCP connection.

 Copyright © Huawei Technologies Co., Ltd.

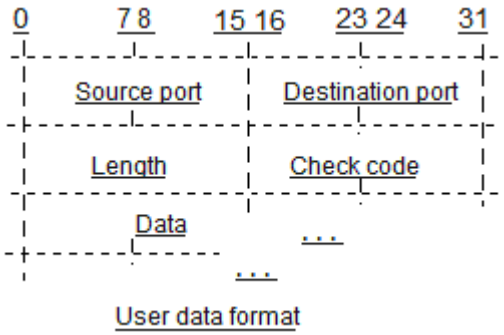
Figure 11-2 Setup and teardown of a TCP connection



11.1.2.3 UDP

The User Datagram Protocol (UDP) is a computer communication protocol that provides packet switching services on the Internet. By default, UDP uses IP as the lower-layer protocol. UDP provides the simplest protocol mechanism that sends information to a user application. UDP is transaction-oriented and does not support delivery or duplicate protection. TCP, however, is required by applications for reliable data transmission. **Figure 11-3** shows the format of a UDP datagram.

Figure 11-3 UDP datagram format



11.1.2.4 RawIP

RawIP only fills in certain fields of an IP header and allows an application to provide its own IP header. Similar to UDP, RawIP is unreliable. No control

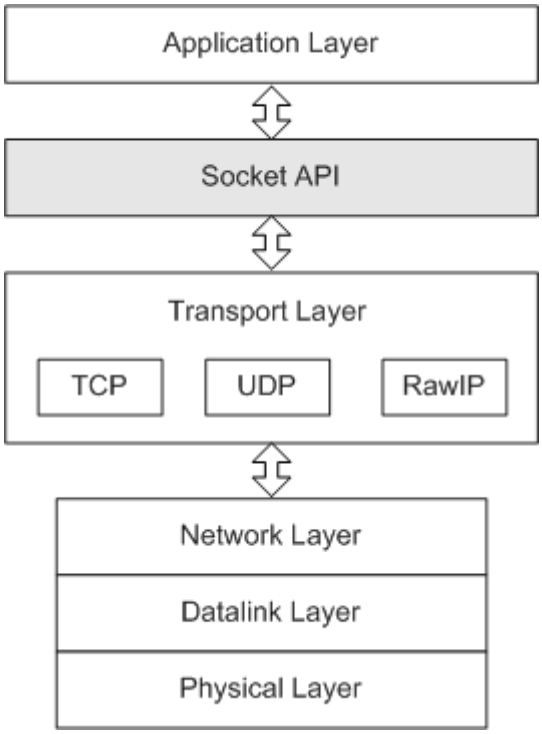
 Copyright © Huawei Technologies Co., Ltd.

mechanism is available to verify whether a RawIP datagram is received. RawIP is connectionless, and it transmits data between hosts without an electric circuit of any type. Unlike UDP, RawIP allows application data to be directly processed at the IP layer through a socket. This is helpful to the applications that need to directly communicate with the IP layer.

11.1.2.5 Socket

A socket consists of a set of application programming interfaces (APIs) working between the transport layer and application layer. The socket shields differences of transport layer protocols and provides the uniform programming interfaces for the application layer. In this manner, the application layer, being exempt from the detailed process of the TCP/IP protocol suite, can transmit data over IP networks by calling socket functions. **Figure 11-4** shows the position of the socket in the TCP/IP protocol stack.

Figure 11-4 Schematic diagram of the socket in the TCP/IP protocol stack



The following types of sockets are supported by different protocols at the transport layer:

- TCP-based socket: provides reliable byte-stream communication services for the application layer.
- UDP-based socket: supports connectionless and unreliable data transmission for the application layer and preserves datagram boundaries.
- RawIP socket: also called raw socket. Similar to the UDP-based socket, the RawIP socket supports connectionless and unreliable data transmission and preserves datagram boundaries. The RawIP socket is unique in that it can be used by applications to directly access the network layer.

 Copyright © Huawei Technologies Co., Ltd.

- Link layer-based socket: used by Intermediate System to Intermediate System (IS-IS) to directly access the link layer.

11.1.3 IPv4 Reference Standards and Protocols

The following table lists the references.

Document No.	Document Name
RFC 793	Transmission Control Protocol
RFC 768	User Datagram Protocol
RFC 1155	Structure and identification of management information for TCP/IP-based internets
RFC 1200	IAB official protocol standards
RFC 1219	On the assignment of subnet numbers
RFC 1878	Variable Length Subnet Table For IPv4
RFC 2581	TCP Congestion Control
RFC 2644	Changing the Default for Directed Broadcasts in Routers
RFC 5925	The TCP Authentication Option

11.2 IPv6

Internet Protocol Version 6 (IPv6), also called IP Next Generation (IPng), is a second-generation protocol of the network layer protocol family.

11.2.1 Why IPv6 is Required

Definition

Internet Protocol Version 6 (IPv6), complies with a set of specifications defined by the Internet Engineering Task Force (IETF), IPv6 is an upgrade of Internet Protocol Version 4 (IPv4). The most significant difference between IPv6 and IPv4 is that the length of IP addresses is extended from 32 bits to 128 bits in IPv6. Featuring a simplified header format, sufficient address space, layered address structure, flexible extension header, and enhanced neighbor discovery (ND) mechanism, IPv6 is more competitive in the future market.

Purpose

As the IPv4-based Internet achieves great success, the IP technology is widely applied. With the rapid development of the Internet, however, deficiencies of IPv4 are more strongly felt, especially in the following aspects:

 Copyright © Huawei Technologies Co., Ltd.

- The IPv4 address space is insufficient.

  An IPv4 address is identified by using 32 bits. In theory, a maximum of 4.3 billion addresses can be provided. In practice, less than 4.3 billion addresses are available due to address allocation reasons. In addition, IPv4 address resources are allocated unevenly. The USA uses almost half of the world's IP addresses, while Europe uses fewer IP addresses than the USA and the Asian-Pacific region uses even fewer. The development of mobile IP and broadband technologies requires more IP addresses. The shortage of IPv4 addresses directly restricts further development of the IP technology.

  There are several solutions to IPv4 address shortage. Classless Inter-domain Routing (CIDR) and Network Address Translation (NAT) are two major solutions. CIDR and NAT, however, have their disadvantages and outstanding problems. This also drives the need for and the development of IPv6.

- The backbone device needs to maintain a large number of routing entries.

  Due to allocation and planning problems in the early phase of IPv4 development, many discontinuous IPv4 addresses are allocated and as a result routes cannot be aggregated effectively. The increasingly large routing tables consume a lot of memory, leading to higher costs for equipment and lower forwarding efficiency. To tackle these issues, device manufacturers have to constantly upgrade their products in order to improve route addressing and forwarding performance.

- Address autoconfiguration and readdressing cannot be performed easily.

  An IPv4 address occupies only 32 bits and IP addresses are allocated unevenly. Consequently, IP addresses need to be reallocated during network expansion or network replanning. Address autoconfiguration and readdressing are required to simplify maintenance. With IPv4, however, such tasks cannot be performed easily.

- Security cannot be well guaranteed.

  With the development of the Internet, security problems become more serious. During the design of IPv4, security was not fully considered. Hence, the original framework cannot implement end-to-end security. IPv6 provides end-to-end security.

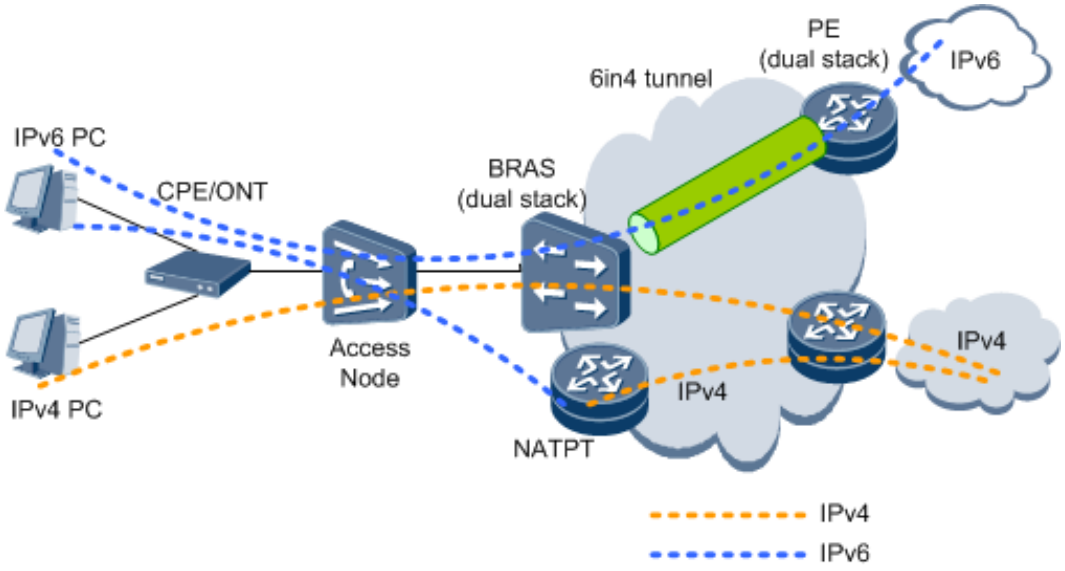
IPv6 solves the IP address shortage issue from the very outset. Moreover, IPv6 is easy to deploy, compatible with various applications, allows IPv4 networks to smoothly transit to IPv6 networks, and can coexist and interwork with IPv4. With so many advantages over IPv4, IPv6 is in rapid development.

11.2.2 IPv6 network deployment

Internet protocol version 6 (IPv6) is developed to tackle global IPv4 address shortage. At the early stage of IPv4-to-IPv6 transition, IPv4 networks have been in large-scale deployment whereas deployment of IPv6 networks is comparatively scarce, and both types of networks will coexist. **Figure 11-5** shows the IPv6 network deployment at the early stage of IPv4-to-IPv6 transition.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-5 Early-stage IPv6 network deployment



IPv4-based broadband remote access servers (BRASs) on the live network are progressively upgraded to support IPv4/IPv6 dual-stack. The dual-stack BRAS provides a 6in4 tunnel or a dedicated link to transmit IPv6 traffic to an IPv6 network.

The dual-stack BRAS provides the network address translation-protocol translation (NAT-PT) function to allow IPv6 users to access an IPv4 network.

The access network equipment is able to detect IPv6 packets, allocate IPv6 addresses (using DHCPv6), and support ACLv6.

11.2.3 IPv6 Principles

Basic functions of IPv6 include IPv6 neighbor discovery and IPv6 path maximum transmission unit (PMTU) discovery. Neighbor discovery and PMTU discovery are implemented through Internet Control Message Protocol for IPv6 (ICMPv6) messages.

11.2.3.1 IPv6 Highlights

- A 128-bit address structure, providing sufficient address space

  A major advantage of IPv6 is the almost infinite IP address space. IPv6 increases the size of an IP address from 32 bits to 128 bits, which is four times of that of IPv4. A 128-bit address structure is able to provide about $4{,}300{,}000{,}000^4$ addresses, meeting almost any address assignment requirements that can be predicted. (Theoretically, 4,300,000,000 IPv4 addresses can be provided; however, $4{,}300{,}000{,}000^4$IPv6 addresses can be provided.)

- Layered address structure

  The layered address structure realizes rapid route lookup, reduces the size of IPv6 routing tables with the aid of route aggregation, and thereby improves the forwarding efficiency of routers.

 Copyright © Huawei Technologies Co., Ltd.

- Address autoconfiguration

  IPv6 enables hosts to discover networks and obtain IPv6 addresses using address autoconfiguration, which greatly improves the network manageability. Using address autoconfiguration, user devices (such as mobile phones and wireless devices) support plug-and-play, without requiring manual configuration or using a private server (such as a DHCP server). IPv6 supports stateful address autoconfiguration and stateless address autoconfiguration.

  – In stateful address autoconfiguration, the host obtains the address and other configuration information from the server.

  – In stateless address autoconfiguration, the host automatically configures address information that contains the prefix and interface ID of the host as reported by the local router. If there is no router on the link, the host can automatically configure only a link-local address for interoperating with the local node.

- Source/Destination address selection

  To specify or plan source/destination addresses of the packets sent by the system, the network administrator can define a set of address selection rules. These rules form an address selection policy table. The policy table is similar to a routing table and employs the longest matching rule for prefix lookup. The address selection result is determined together by the source address and destination address.

  A source address is selected according to the following rules. Among the rules below, the rule with a smaller number has a higher priority. A candidate address is preferred if it:

  a. Is the same as the destination address.

  b. Has an appropriate effective scope.

  c. Is not a deprecated address.

  d. Is a home address.

  e. Is the address of an outgoing interface.

  f. Has the same *label* value as the destination address.

  g. Has the longest matching prefix.

  NOTE

  The candidate address can be a unicast address that is configured on the specified outgoing interface. If a source address that has the same *label* value as and is in the same address range with the destination address is not found on the outgoing interface, such a source address can be selected on another interface.

  A destination address is selected according to the following rules. Among the rules below, the rule with a smaller number has a higher priority. A candidate address is preferred if it:

  a. Is not an unusable address.

  b. Has an appropriate effective scope.

  c. Is not a deprecated address.

  d. Is a home address.

  e. Has the same *label* value as the source address.

 Copyright © Huawei Technologies Co., Ltd.

f. Has a larger *precedence* value.
g. Prefer native transport (6over4 or 6to4 tunnel is not required).
h. Has a smaller effective scope.
i. Has the longest matching prefix.
j. Otherwise, leave the order unchanged.

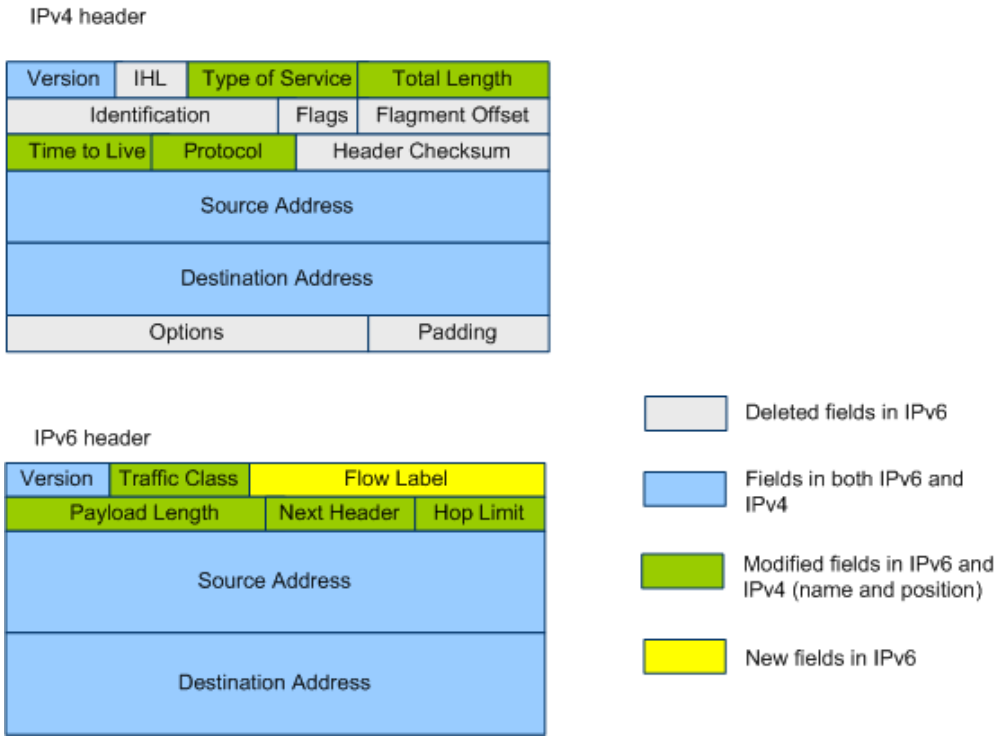
- Support for QoS

  New fields are added in the IPv6 header to define how to label and process flows. Flows are identified by the Flow Label field in the header. The Flow Label field allows routers to identify the packets of a certain flow and provide special processing for these packets.

- Flexible and simple extension headers

  Figure 11-6 shows the comparison between IPv6 and IPv4 header formats. Compared with an IPv4 header, an IPv6 header deletes the IHL, Identification, Flags, Fragment Offset, Header Checksum, Options, and Padding fields, and adds the Flow Label field, which helps improve the header processing efficiency. In addition, to better support different options, IPv6 introduces multiple extension headers. With these extension headers, it is unnecessary to modify the existing packet structure when new options are added, which greatly improves the flexibility of IPv6.

Figure 11-6 Comparison between IPv6 and IPv4 header formats



 Copyright © Huawei Technologies Co., Ltd.

11.2.3.2 IPv6 Addresses

Format of an IPv6 Address

A 128-bit IPv6 address has two formats:

- X:X:X:X:X:X:X:X
  - In this format, the 128 bits of an IPv6 address are divided into 8 groups. The 16 bits of each group are represented by 4 hexadecimal characters (0 to 9, and A to F). Groups are separated by colons. Every "X" represents the numerical value of a group of hexadecimal characters. The following is an example:

    2001:0DB8:130F:0000:0000:09C0:876A:130B

    For convenience, the zeros at the beginning of each group can be omitted. The preceding example, thus, can be written as 2001:DB8:130F:0:0:9C0:876A:130B.
  - Furthermore, the two or more consecutive zeros in the address can be replaced by "::", which reduces the written length of an IPv6 address. The preceding example can be further compressed as 2001:DB8:130F::9C0:876A:130B.

    An IPv6 address contains only one "::". Otherwise, a computer cannot determine the count of zeros when restoring the original 128-bit address from the compressed address.
- X:X:X:X:X:X:d.d.d.d

  Addresses in this format fall into the following two types:
  - IPv4-compatible IPv6 address: The format of an IPv4-compatible IPv6 address is 0:0:0:0:0:0:IPv4-address. The high-order 96 bits are all 0s, and the low-order 32 bits are an IPv4 address. This IPv4 address must be reachable in an IPv4 network, and cannot be a multicast, broadcast, loopback, or unspecified address (0.0.0.0).
  - IPv4-mapped IPv6 address: The format of an IPv4-mapped IPv6 address is 0:0:0:0:0:FFFF:IPv4-address. This type of IPv6 address is used to represent the addresses of IPv4 nodes.
  - An IPv4-compatible IPv6 address is used for the configuration of IPv6 over IPv4 tunnels.

    "X:X:X:X:X:X" represent the high-order 6 groups of numbers, and each "X" stands for 16 bits expressed in hexadecimal notation. "d.d.d.d" represent the low-order 4 groups of numbers, and each "d" stands for 8 bits expressed in decimal notation. "d.d.d.d" is a standard IPv4 address.

Structure of an IPv6 Address

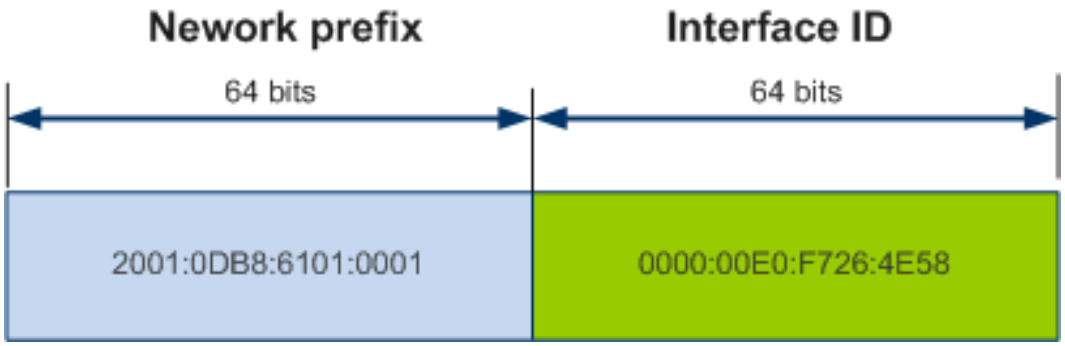
An IPv6 address can be divided into two parts:

- Network prefix: n bits, equaling the network ID in an IPv4 address.
- Interface identifier (ID): 128-n bits, equaling the host ID in an IPv4 address.

Figure 11-7 illustrates the structure of the address 2001:DB8:6101:1::E0:F726:4E58 /64.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-7 Structure of the address 2001:DB8:6101:1::E0:F726:4E58 /64



IPv6 Address Classification

IPv6 has the following types of addresses:

- Unicast address: uniquely identifies an interface and is similar to an IPv4 unicast address. The packets sent to a unicast address are transmitted to the unique interface identified by this address.

  Unicast addresses can be classified into the following types, as shown in **Table 11-1**.

  Table 11-1 Types of IPv6 unicast addresses

Address Type	Binary Prefix	IPv6 Prefix Identifier
Link-local unicast address	1111111010	FE80::/10
Loopback address	00...1 (128 bits)	::1/128
Unspecified address	00...0 (128 bits)	::/128
Global unicast address	Others	-

  The meanings of each type of address are described as follows:

  – Link-local IPv6 unicast address: It is used in the Neighbor Discovery (ND) protocol, and in the communication between nodes on the local link during the stateless address autoconfiguration. The packets carrying the link-local IPv6 unicast address as the source or destination address are forwarded only on the local link. A link-local IPv6 unicast address can be automatically configured on any interface by using the link-local prefix FE80::/10(1111 1110 10) and the interface identifier in the IEEE EUI-64 format (an EUI-64 identifier can be derived from an EUI-48 identifier).

  – Loopback address: It is 0:0:0:0:0:0:0:1 or ::1 and not assigned to any interface. Similar to the case of IPv4 loopback address 127.0.0.1, IPv6 packets carrying the IPv6 loopback address as the destination address are looped back to the sending node.

  – Unspecified address (::): It cannot be assigned to any node or function as a destination address. The unspecified address can be used in the Source Address field of the IPv6 packet sent by an initializing host that has not

 Copyright © Huawei Technologies Co., Ltd.

obtained its own address. During duplicate address detection (DAD), the Source Address field of a Neighbor Solicitation (NS) message is an unspecified address.

  - Global unicast address: It is equivalent to an IPv4 public network address. Global unicast addresses are used on the links that can be aggregated, and are provided to Internet service providers (ISPs). The structure of this type of address allows for route prefix aggregation to relieve the global routing entry resources limitation. A global unicast address consists of a 48-bit route prefix that is managed by the carrier, a 16-bit subnet ID that is managed by the local node, and a 64-bit interface ID. Unless otherwise specified, global unicast addresses include site-local unicast addresses.

- Anycast address: identifies a group of interfaces, which generally belong to different nodes. The packets carrying an anycast destination address are transmitted to the interface that is nearest to the source node in the interface group identified by the anycast address. The nearest interface refers to the interface with the smallest distance metric measured by the routing protocol.

  Application scenario: When a mobile host needs to communicate with the mobile agent on the home subnet, the mobile host uses the anycast address of the routing device in the subnet.

  Specifications of anycast addresses: Anycast addresses do not have independent address space. They can use the format of any unicast address. Thus, syntax is required to differentiate an anycast address from a unicast address.

- Multicast address: identifies a group of interfaces that belong to different nodes and is similar to an IPv4 multicast address. The packets carrying a multicast destination address are transmitted to all the interfaces identified by this multicast address.

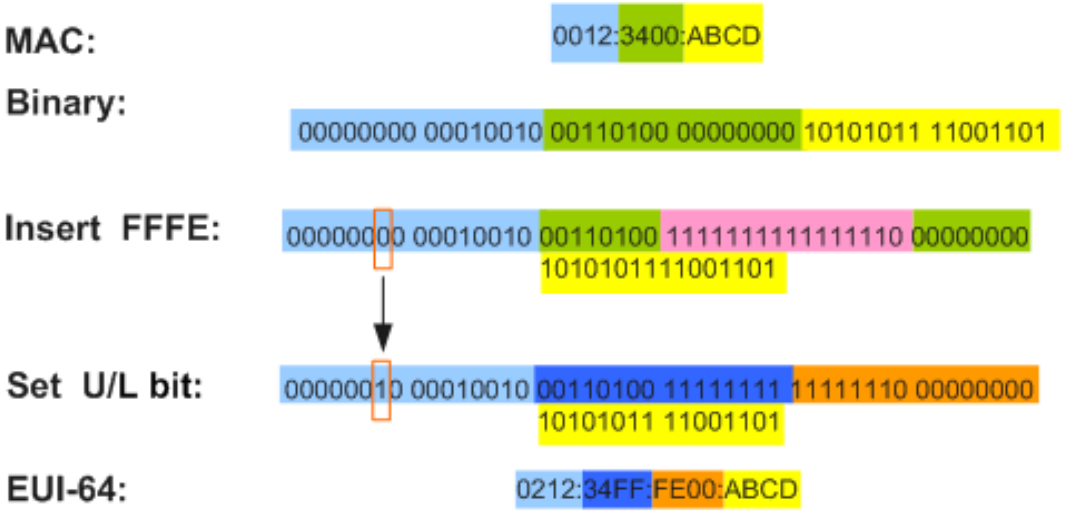
  IPv6 addresses do not include broadcast addresses. In IPv6, functions of broadcast addresses are provided by multicast addresses.

Interface ID in the IEEE EUI-64 Format

The 64-bit interface ID in an IPv6 address identifies a unique interface on a link. This address is derived from the link-layer address (such as a MAC address) of the interface. The interface ID of an IPv6 address contains 64 bits, while a MAC address contains 48 bits. By inserting a hexadecimal number FFFE (1111 1111 1111 1110) into the MAC address and then setting the U/L bit (the leftmost seventh bit) to 1, an interface ID in the EUI-64 format is generated. **Figure 11-8** shows conversion process.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-8 Converting a MAC address into an EUI-64 address

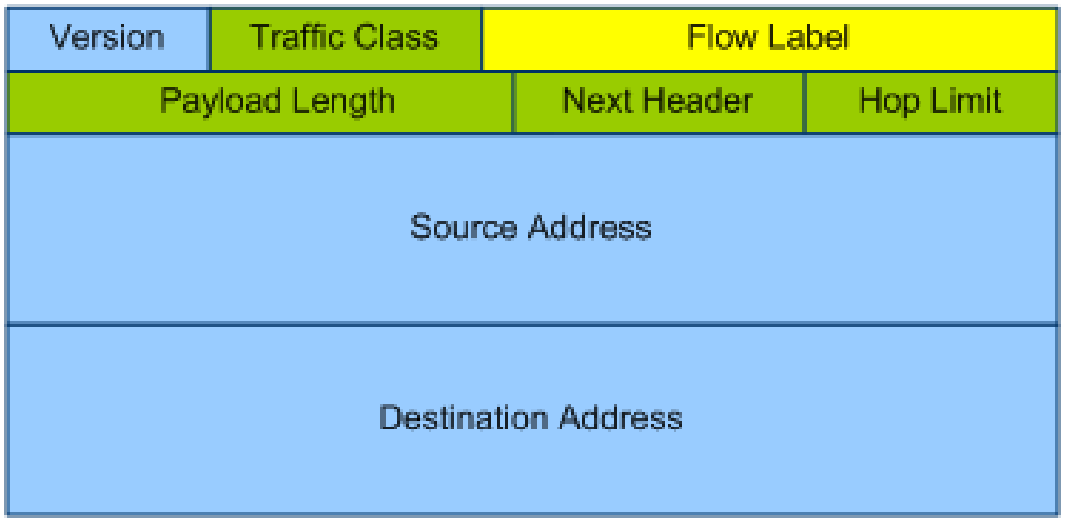


11.2.3.3 IPv6 Packet Format

Format of an IPv6 Header

Figure 11-9 shows the format of an IPv6 header.

Figure 11-9 Format of an IPv6 header



- Version:
  4 bits. The value of this field is 6, indicating an IPv6 packet.
- Traffic Class:
  8 bits. This field is similar to the TOS field in an IPv4 header.
- Flow Label:

 Copyright © Huawei Technologies Co., Ltd.

20 bits. This field is new in IPv6. Flow labels are used to label particular flows so as to differentiate packets at the network layer. Routers along a forwarding path differentiate and process packets according to the flow labels. Because the Flow Label field locates in the IPv6 header, forwarding routers and destination nodes do not need to differentiate packets according to the packet content.

- Payload Length:

  16 bits. This field indicates the length of the IPv6 payload in octets. The payload is the rest of the packet following the IPv6 header. (Note that any extension headers present are considered part of the payload.)

- Next Header:

  8 bits. This field identifies the type of header immediately following the current IPv6 header (which may be a header or an extension header). This field uses the same values as the IPv4 Protocol field. The Next Header field in the IPv6 header and the Next Header fields in the IPv6 extension headers form a chain. The chain mechanism helps improve the efficiency of extension header processing because the router processes only the option headers needed.

- Hop Limit:

  8 bits. This field is similar to the IPv4 TTL field. This field decreases by 1 by each node that forwards the packet. The packet is dropped if this field decreases to 0.

- Source Address:

  128 bits. This field indicates the source address of the packet.

- Destination Address:

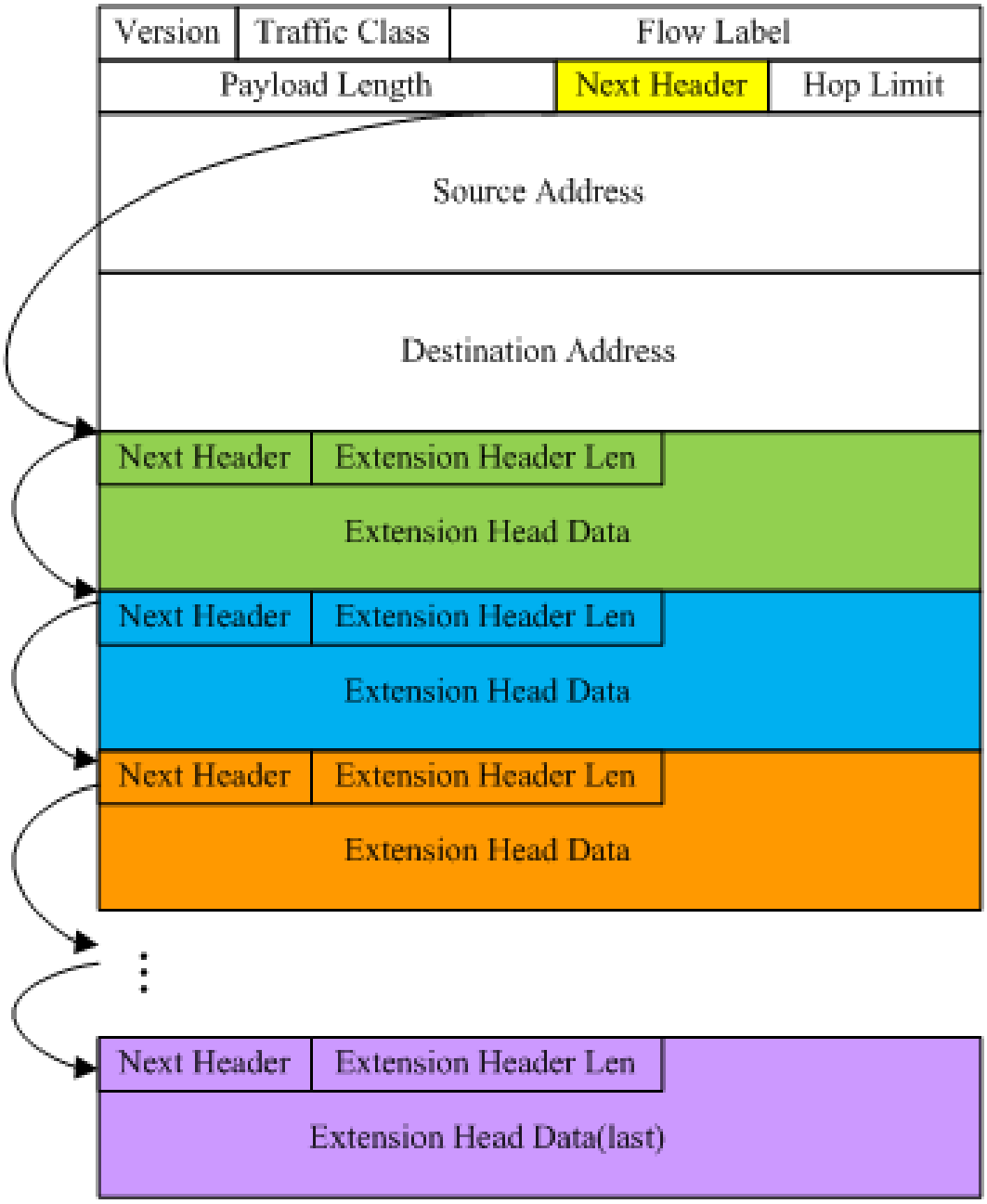
  128 bits. This field indicates the destination address of the packet.

Format of IPv6 Extension Headers

Figure 11-10 shows the format of IPv6 extension headers.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-10 Format of IPv6 extension headers



IPv6 option fields are supported through a chain of extension headers. An IPv6 packet can carry zero, one, or multiple extension headers.

IPv6 extension headers appear in the following order:

- Hop-by-Hop Options Header

  The value of this header is 0, which is defined in the IPv6 header. It is used for routing alarms and jumbo frames. This header is processed by every node along the packet forwarding path.
- Destination Options Header

 Copyright © Huawei Technologies Co., Ltd.

The value of this header is 60. This header may occur before the following two headers:

- Routing Header

  In such a case, the Destination Options header is processed by the destination node and the node specified in the Routing header.
- Upper-layer Header (located behind any ESP option)

  In such a case, the Destination Options header is processed only by the destination node.

- Routing Header

  The value of this header is 43. This header is used for source routing options and mobile IPv6.
- Fragment Header

  The value of this header is 44. This header is used for packet fragmentation when the packet sent by the source node is larger than the path maximum transmission unit (PMTU). PMTU is the MTU specified for the path from the source node to destination node.
- Authentication Header

  The value of this header is 51. This header is used for authentication and integrity checking of a packet. The definition of this header in IPv6 is the same as that in IPv4.
- ESP Header

  The value of this header is 50. This header is used for authentication, integrity checking, and encryption of a packet. The definition of this header in IPv6 is the same as that in IPv4.
- Upper-layer Header

  This header is an upper-layer protocol (such as TCP, UDP, or ICMP) header.

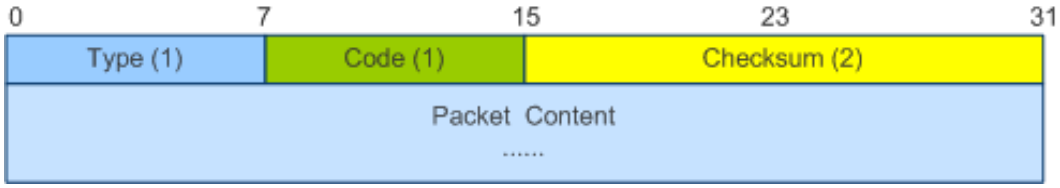
The Destination Options header occurs at most twice (once before the Routing header and once before the upper-layer header). The other extension headers occur at most once.

IPv6 nodes must be able to process the extension headers, regardless of the order and the number of occurrences they appear in the same packet, except for the Hop-by-Hop Options header that is restricted to appear immediately after an IPv6 header only. This requirement ensures interoperability between IPv6 nodes.

11.2.3.4 ICMPv6

As a basic protocol of IPv6, Internet Control Message Protocol for IPv6 (ICMPv6) is used by a device to generate two types of messages: error messages and informational messages. Using these messages, IPv6 nodes report errors and information generated during packet processing. **Figure 11-11** shows the format of an ICMPv6 message.

Figure 11-11 Format of an ICMPv6 message



 Copyright © Huawei Technologies Co., Ltd.

The meaning of each field in an ICMPv6 message is as follows:

- Type field: indicates the message type. The values from 0 to 127 indicate the error message type, and values from 128 to 255 indicate the informational message type.
- Code field: indicates the specific message type.
- Checksum field: indicates the checksum of an ICMPv6 message.

Classification of ICMPv6 Error Messages

- Destination Unreachable message

  When an IPv6 node is forwarding IPv6 packets and finds that the destination address of the packets is unreachable, it sends an ICMPv6 Destination Unreachable message to the source node of the packets. Specific causes for the error message are carried in the message. Destination Unreachable messages are classified into types that include the following:

  – No route to destination
  – Address unreachable
  – Port unreachable

- Packet Too Big message

  When an IPv6 node is forwarding IPv6 packets and finds that the size of the packets exceeds the path maximum transmission unit (PMTU) of the outbound interface, it sends an ICMPv6 Packet Too Big message to the source node of the packets. The PMTU of the outbound interface is carried in the message. PMTU discovery is implemented based on Datagram Too Big messages.

- Time Exceeded message

  During the reception and transmission of IPv6 packets, when a device receives a packet with the hop limit of 0 or when the device reduces the hop limit to 0, it sends an ICMPv6 Time Exceeded message to the source node of the packets. When reassembling fragmented packets, an ICMPv6 Time Exceeded message is also generated if the reassembly time exceeds the specified duration.

- Parameter Problem message

  When a destination node receives an IPv6 packet, it checks the validity of the packet. If the destination node detects any of the following errors, it sends an ICMPv6 Parameter Problem message to the source node of the packet:

  – A field in the IPv6 header or extension header is incorrect.
  – The Next Header field in the IPv6 header or extension header cannot be identified.
  – Unknown options exist in the extension header.

Classification of ICMPv6 Informational Messages

ICMPv6 informational messages are classified into Echo Request messages and Echo Reply messages. ICMPv6 informational messages can be used for network fault diagnosis, PMTU discovery, and neighbor discovery. During the interoperation

 Copyright © Huawei Technologies Co., Ltd.

check between two nodes, the node that receives an Echo Request message sends an Echo Reply message to the source node. In this manner, subsequent packets are received and transmitted between the two nodes.

NOTE

A large number of generated redirection ICMP packets will degrade the device performance. The device does not send redirection ICMP packets by default. The **icmpv6 redirect enable** command is used to enable the function of sending redirection ICMP packets.

11.2.3.5 PMTU

Problems Related to MTU

A path from a source address to a destination address may traverse interfaces that have different maximum transmission unit (MTU) values. The smallest MTU on this path is called the path MTU (PMTU).

- During transmission of IPv6 packets, the packets cannot be fragmented on the intermediate nodes. Therefore, it often happens that the packet length exceeds the PMTU. In such a case, the source node needs to retransmit the IPv6 packets, which reduces the transmission efficiency.
- If the minimum link IPv6 MTU (1280 bytes) is used as the maximum fragment length on the source node, in most cases, the IPv6 MTU of the path is greater than the IPv6 MTU of the minimum link, and the fragment sent by a node is far less than the IPv6 MTU of the path, which is a waste of network resources.

The PMTU discovery protocol is introduced to solve this problem.

Principle of PMTU Discovery

PMTU discovery describes a method of dynamically discovering the PMTU for a path. When an IPv6 node sends a large amount of data to another node, the data is transmitted by means of a sequence of IPv6 fragments. When these fragments are of the maximum length allowed in successful transmission between the source node and destination node, the fragment length is considered optimal and called PMTU.

A source node assumes that a PMTU of a path is the known IPv6 MTU of the first hop on the path. If the packet sent from this path is too large to be forwarded along the path, the intermediate node discards this packet and returns an ICMPv6 Packet Too Big message to the source node. The source node then sets the PMTU of the path according to the IPv6 MTU specified in the message.

When the PMTU learned by the node is smaller than or equal to the actual PMTU, the PMTU discovery process ends. Before the PMTU discovery process end, ICMPv6 Packet Too Big messages may be repeatedly sent and received because smaller IPv6 MTUs may be found further down the path.

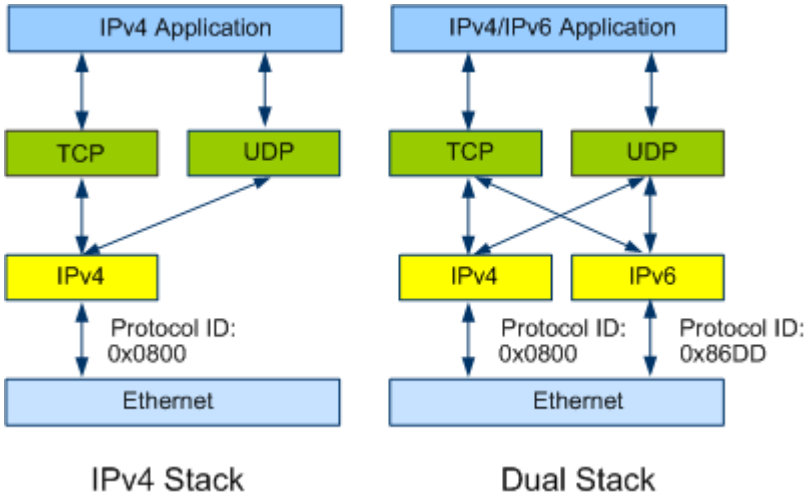
11.2.3.6 Dual Protocol Stacks

For an IPv6 node, the most effective way of being compatible with IPv4 is to retain a complete IPv4 protocol stack on the node. Such a node is called a dual-stack

 Copyright © Huawei Technologies Co., Ltd.

node. **Figure 11-12** shows the structures of a single protocol stack and dual protocol stacks.

Figure 11-12 Structures of a single protocol stack and dual protocol stacks in an Ethernet



Dual protocol stacks have the following advantages:

- Multiple link-layer protocols support dual protocol stacks.

  Multiple link-layer protocols, such as Ethernet, support dual protocol stacks. In **Figure 11-12**, the link-layer protocol is Ethernet. In an Ethernet frame, if the Protocol ID field is 0x0800, it indicates that the network layer receives IPv4 packets; if the field is 0x86DD, it indicates that the network layer receives IPv6 packets.
- Multiple applications support dual protocol stacks.

  The upper layer applications, such as the DNS, can use TCP or UDP as the transmission layer protocol, and prefers the IPv6 protocol stack rather than the IPv4 protocol stack as the network-layer protocol.

11.2.3.7 TCP6

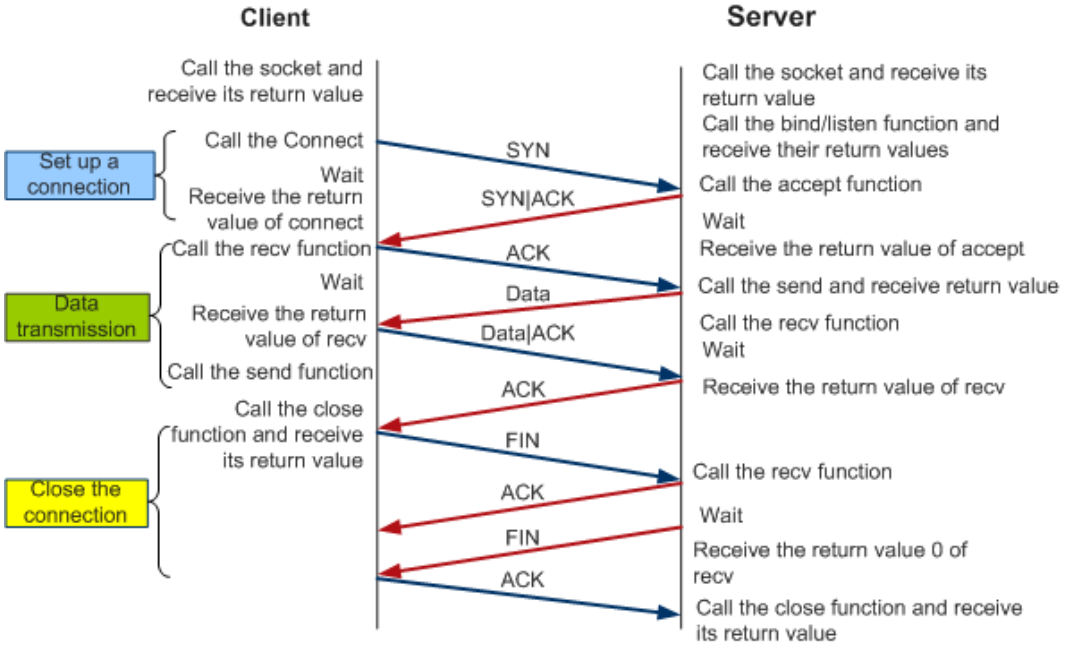
Transmission Control Protocol Version 6 (TCP6) provides a mechanism to establish virtual circuits between processes of two endpoints. A TCP6 virtual circuit is similar to the full-duplex circuit that transmits data between systems. TCP6 is called a reliable protocol because it provides reliable data transmission between processes. TCP6 also provides a mechanism to optimize the transmission performance according to the network status. When all the data can be received and acknowledged, the transmission rate increases gradually.

TCP6 is generally used in interactive applications, such as the web application. TCP6 establishes virtual circuits by using the three-way handshake mechanism, and all the virtual circuits are deleted using the four-way handshake mechanism. TCP6 connections provide a variety of checksums and reliability-ensuring functions, but increase the cost. As a result, TCP6 has lower efficiency than User Datagram Protocol Version 6 (UDP6).

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-13 shows the establishment and removal of a TCP6 connection.

Figure 11-13 Establishment and removal of a TCP6 connection



11.2.3.8 UDP6

User Datagram Protocol Version 6 (UDP6) is a computer communication protocol used to exchange packets on a network. UDP6 has the following characteristics:

- UDP uses only source and destination information and runs mainly in a simple request/response structure.
- UDP is unreliable. No control mechanism is provided to ascertain whether UDP6 datagrams have reached their destinations.
- UDP is connectionless. No virtual circuits are required for data transmission between hosts.

The connectionless feature of UDP6 enables UDP6 to send data to broadcast addresses. This is different from TCP6, which requires specific source and destination addresses.

11.2.3.9 RawIP6

RawIP6 is an implementation in which only a limited number of fields in the IPv6 header are filled with values, and RawIP6 allows applications to provide their own IPv6 headers.

RawIP6 is similar to UDP6 in the following aspects:

- RawIP6 is unreliable. No control mechanism is provided to ascertain whether RawIP6 datagrams have reached their destinations.
- RawIP6 is connectionless. No virtual circuits are required for data transmission between hosts.

 Copyright © Huawei Technologies Co., Ltd.

Unlike UDP6, RawIP6 allows applications to directly operate the IP layer through sockets. Therefore, RawIP6 is convenient for the applications that need to interact directly with the lower layer.

11.2.3.10 IPv6 Neighbor Discovery

The Neighbor Discovery (ND) protocol defines a set of messages and processes for determining the relationship between neighboring nodes. The IPv6 ND protocol supports the Address Resolution Protocol (ARP) messages, and the Router Discovery and Redirect messages of the Internet Control Message Protocol (ICMP) of IPv4, and also provides other functions.

After an IPv6 address is configured on a node, the node first verifies that the address is available and does not conflict with other addresses. If the node is a host, the router needs to notify the host of a better first-hop address to reach a particular destination. If the node is a router, this node needs to advertise its address, address prefix, and other configuration information, which provides guidance for the host to configure parameters. During IPv6 packet forwarding, a node needs to verify the link-layer address and reachability of its neighboring node. IPv6 ND defines five types of ICMPv6 packets:

- Router Solicitation (RS): Sent by an enabled host to a router. The router then responds with a Router Advertisement (RA).
- RA: Advertised by a router periodically, which contains information such as the prefix and some flags.
- Neighbor Solicitation (NS): Sent by an IPv6 node to determine the link-layer address of a neighbor, to check whether a neighbor is reachable, or for duplicate address detection (DAD).
- Neighbor Advertisement (NA): A response to an NS. An IPv6 node may also send unsolicited NAs to announce a link-layer address change.
- Redirect: Used by a router to inform hosts of a better first hop for a specific destination when the router finds that the incoming interface and outgoing interface of the packets are the same.

The ND protocol for IPv6 has the following functions.

DAD

In an IPv6 network, a link-local address (LLA) is used for the communication between neighboring nodes on the same link, for example, the communication between hosts on a single link where no routers exist. If an LLA is available, it is automatically used for neighbor discovery. Duplicate address detection (DAD) is a detection mechanism used for determining whether an LLA address is available. The process is as follows:

1. When an IPv6 address is configured on a node, the node sends an NS to its neighboring node to check whether the IP address is already used.
2. When receiving the NS, the neighboring node checks whether it has the same IPv6 address. If yes, the neighboring node responds with an NA carrying the IPv6 address information.
3. The node, upon receiving the NA, considers that the IPv6 address configured is already used by its neighboring node. If the node does not receive any

 Copyright © Huawei Technologies Co., Ltd.

response packet from its neighboring node, the IPv6 address configured is available.

DAD Proxy

A DAD proxy is used to prevent LLA conflicts on a Layer 3 interface. The DAD proxy feature resolves the issue of LLA conflicts between isolated ports on the same Layer 3 interface.

- When a user performs DAD, the MA5800 checks whether any LLA conflict occurs in the Layer 3 interface VLAN based on the information about the packet sent by the user.
  - If no conflict occurs, the MA5800 forwards the packet.
  - If a conflict occurs, the MA5800 responds to this conflict and drops the packet.
- The MA5800 obtains the LLA lease time from Router Advertisement (RA) and DHCPv6 packets, and the buffered LLA entries age when the lease time expires.

Neighbor Discovery

ND for IPv6, similar to ARP in IPv4, is used to parse the addresses of neighbors and detect whether neighbors are reachable using NSs and NAs.

To obtain the link-layer address of another node in the same local link, a node (source node) sends an NS with its ICMPv6 type as 135. This packet is similar to an ARP request packet in IPv4; but unlike the ARP request packet using a broadcast address, the NS uses a multicast address. A node with the last 24 bits of its address the same as the multicast address will receive this NS, which reduces the possibility of broadcast storm. The node receiving the NS (destination node) fills in its link-layer address in the response packet.

An NS can also be used to check whether a neighboring node is reachable when the link-layer address of the neighboring node is known. An IPv6 NA message is sent in response to an IPv6 NS message. The destination node, upon receiving an NS, responds with an NA with its ICMPv6 type as 136 over the local link. The source node then is able to communicate with the destination node after receiving the NA. A node may also send unsolicited NAs to announce a link-layer address change on the local link.

Router Discovery

Router Discovery is used to locate neighboring routers as well as learn prefixes and configuration parameters related to stateless address autoconfiguration (SLAAC). Router Discovery in IPv6 is implemented using the following two mechanisms:

- Router Solicitation (RS) message

  When unicast addresses are not configured on a host (for example, the system just starts up), the host sends an RS. An RS facilitates the host autoconfiguration, without having to wait for the RA sent by IPv6 routers. The IPv6 RS message is an ICMPv6 message of Type 133.
- Router Advertisement (RA) message

 Copyright © Huawei Technologies Co., Ltd.

Each RA-enabled IPv6 router periodically sends RAs. After receiving an RS from an IPv6 node on the local link, an IPv6 router also responds with an RA. An IPv6 router sends an RA to the multicast addresses (FF02::1) of all nodes or to the IPv6 unicast address of the node sending an RS. An RA is an ICMPv6 of type 134, including the following content:

- Whether to use address autoconfiguration
- Supported autoconfiguration type (stateless or stateful)
- One or more local link prefixes (nodes on the local link can use these prefixes to perform address autoconfiguration)
- Lifetime of local link prefixes advertised
- Whether the router sending an RA can serve as a default router. If yes, the information also contains the lifetime of the default router in units of seconds.
- Other host-related information, such as the hop limit and the MTU used for the host to send packets

An IPv6 node on the local link receives an RA and updates information such as the default router, prefix list, and other information from this RA.

Address Autoconfiguration

RAs and per-prefix flags enable routers to inform hosts how to perform address autoconfiguration. For example, routers can specify whether hosts use stateful (DHCPv6) or stateless (SLAAC) address configuration.

When using the SLAAC protocol, a host uses the prefix information and local interface ID obtained from the RA received to automatically generate an IPv6 address. Also, the host can set the default router according to the default router information in the RA.

Redirect

Redirect messages are sent by routers to inform a host of a better first-hop IPv6 address for a specific destination. Like IPv4, IPv6 Redirect messages are sent only for redirecting packets to a better router. Nodes receiving these Redirect messages will send packets to this better router. Routers send Redirect messages only for unicast flows, and Redirect messages are only sent to and processed by those nodes (hosts) triggering redirect.

Default Router Priority and Routing Information

An RA defines two fields: default router priority and routing information, which helps hosts select a better forwarding router for packets.

When the link where a host resides has multiple routers, the host needs to select the forwarding router according to the packet's destination address. In such cases, routers advertise default router priority and specific routing information to hosts, improving hosts' capability of selecting better forwarding routers according to different destination addresses.

A host, upon receiving an RA that contains the routing information, will update its routing list. Before sending packets to other devices, the host selects a better route according to its routing list.

 Copyright © Huawei Technologies Co., Ltd.

A host updates its default router list after receiving an RA that contains default router priority information. Before sending packets to other devices, the host queries its default router list and selects the highest-priority router for sending packets if no route is available. If the highest-priority router is faulty, the host will choose the second-highest-priority router and so on.

ND Proxy

NOTE

This function is supported in V100R020C00 and later versions.

User-side packets are sent to only the upper-layer device and they cannot be forwarded between user sides. After multicast domains are isolated, user networks are more secured. However, user terminals require interconnection, that is, different VLANs or some users isolated in the same VLAN need to communicate with each other.

Both IPv4 and IPv6 networks have such a requirement. It is similar to ARP proxy for IPv4 networks. ND proxy is an expansion of ARP proxy for IPv6 networks, which is used to resolve the IPv6 network interconnection issue. To configure it,

- Enable the ND proxy on a super VLAN interface, and then the super VLAN interface acts as the proxy of all the sub VLANs in the super VLAN.
- To enable the ND proxy for the NEs in a sub VLAN, the ND proxy of the sub VLAN must be enabled.

11.2.4 Configuring Basic IPv6 Information

This topic describes the IPv6 features supported by the MA5800. The basic IPv6 configuration includes configuration of the IPv6 address, IPv6 neighbor discovery, path maximum transmission unit (PTMU), and transmission control protocol 6 (TCP6).

Context

The following table lists the IPv6 features supported by the MA5800.

Table 11-2 IPv6 features supported by the MA5800

Feature	Sub-feature	Configuration Process or Command
IPv6 address management and assignment	Static configuration of IPv6 global unicast addresses and IPv6 link-local addresses	**Configuring an IPv6 Address for an Interface**
	Automatic configuration of IPv6 link-local addresses	**ipv6 address auto link-local**
	DHCPv6, DHCPv6 L2/L3 Relay	**11.7.4 Configuring DHCPv4 Relay**

 Copyright © Huawei Technologies Co., Ltd.

Feature	Sub-feature	Configuration Process or Command
	Management information base (MIB) for IPv6 address management	-
IPv6 stack and IPv6 host function	IPv6/IPv4 dual-stack to ensure compatibility of IPv6 and IPv4	-
	Basic IPv6 protocols, including ICMPv6, TCP6, UDP6, and RawIP6	-
	IPv6 Neighbor Discovery (ND) protocol and static configuration of IPv6 neighbors	**Configuring IPv6 Neighbor Discovery**
	IPv6 PMTU	**Configuring PMTU**
	IPv6 ping and tracert	• **ping ipv6** • **tracert ipv6**
	IPv6 statistics query and clearance	• **display ipv6 statistics** • **reset ipv6 statistics**
IPv6 route	IPv6 static routes	**11.12.2.5 Configuration Example of the IPv6 Static Route**
	BGP4+	**11.12.8.4 Configuration Example of BGP4+**
IPv6 QoS and security	IPv6 ACL	**Configuring Traffic Management Based on ACL Rules**
	Anti-MAC spoofing	**Configuring MAC Anti-Spoofing**
	Anti-IPv6 spoofing	**Configuring IPv6/ICMPv6 Address Attack Prevention**
	Anti-denial of service (DoS) attack	**Configuring Anti-DoS Protection**
DAD Proxy and Proxy advertisement for NS	DAD Proxy	**ipv6 dad proxy**
	Proxy advertisement for neighbor solicitation (NS) on the network side	-

 Copyright © Huawei Technologies Co., Ltd.

Feature	Sub-feature	Configuration Process or Command
IPv6 Layer 2 Transparent Transmission	Differentiation of service virtual ports based on the IPv6 over Ethernet (IPv6oE) type (0x86DD) and defining of VLANs for service virtual ports	**service-port**
	Transparent transmission of IPv6 over PPPoE packets	-
	VLAN-based transparent transmission of IPv6 packets	-

NOTE

In this manual:

- For the IPv6 features that are different from IPv4 features, configuration procedures and examples are provided for both IPv6 and IPv4 features.
- For the IPv6 features that are similar with IPv4 features, configuration procedures and examples are not provided for IPv6 features because they are the same as IPv4 features. To configure these IPv6 features, use IPv6 commands and follow the procedures of IPv4 features.

11.2.4.1 Configuring an IPv6 Address for an Interface

The MA5800 can communicate with other IPv6 equipment only after its interface is configured with an IPv6 address. Before an IPv6 global unicast address or IPv6 link-local address is configured on an interface, the IPv6 packet forwarding function must be enabled on the device.

Context

Each interface can be configured with a maximum of 10 IPv6 global unicast addresses but only one IPv6 link-local address.

- An IPv6 global unicast address is equivalent to an IPv4 public address. It is used for forwarding data across the public network and is necessary for the communication between users. An EUI-64 address has the same function as an IPv6 global unicast address. The difference is that only the network bits need to be specified for the EUI-64 address and the host bits are transformed from the media access control (MAC) addresses of the interface, while a complete 128-bit address needs to be specified for the IPv6 global unicast address.
- The IPv6 link-local address is used in neighbor discovery (ND), and for the communication between nodes on the local link in the stateless address autoconfiguration (SLAAC) process. The packets using the link-local address as the source or destination address are not forwarded to other links.

  The link-local address can be automatically generated or manually configured. It is recommended to automatically generate a link-local address because the

 Copyright © Huawei Technologies Co., Ltd.

link-local address is used to implement communication requirements of protocol and is not directly related to the communication between users.

NOTE

In the SLAAC process, a host uses the prefix information and local interface ID obtained from the received router advertisement (RA) to automatically generate an IPv6 address, rather than using the stateful address autoconfiguration mechanism (DHCPv6).

The MA5800 supports IPv6 address configuration on the VLANIF interface, METH interface, and loopback interface. This topic uses the VLANIF interface as an example.

Procedure

Step 1 Enable IPv6 packet forwarding capability.

Enabling the IPv6 function on the device and the interface is a prerequisite for configuring IPv6 features. To enable a device to forward IPv6 packets, you must enable the IPv6 capability in both the global config mode and the interface mode.

By default, the IPv6 function is disabled on the device and interface.

1. In global config mode, run the **interface vlanif** command to enter the VLANIF mode.
2. In VLANIF mode, run the **ipv6 enable** command to enable the IPv6 function on the interface.

NOTE

Before configuring other IPv6 features on an interface, you must enable the IPv6 function in interface mode.

Step 2 Configure an IPv6 global unicast address on the interface.

In VLANIF mode, run the **ipv6 address** or **ipv6 address eui-64** command to configure an IPv6 global unicast address on the interface.

To implement successful communication, you can configure both the EUI-64 address and the IPv6 global unicast address, or configure only one of them. The IPv6 addresses configured on one interface cannot be in the same network segment.

Step 3 Configure an IPv6 link-local address on the interface.

In VLANIF mode, use either of the following methods to configure the IPv6 link-local address:

- Run the **ipv6 address auto link-local** command to automatically generate a local-link address on the interface.

  After this command is executed, the deletion of the global unicast address does not affect local link communication. If the device only needs to communicate with another device that is directly connected to the device, using the link-local address saves IPv6 global unicast address resources.
- Run the **ipv6 address link-local** command to manually configure a link-local address on the interface. The prefix of the IPv6 address configured by running this command must be FE80::/10.

 Copyright © Huawei Technologies Co., Ltd.

If an automatically allocated link-local address exists on the interface, the link-local address will be overwritten by the new link-local address after this command is executed. If possible, avoid changing the link-local address.

If you have not run either of the preceding commands to configure a link-local address for an interface, a link-local address will be automatically allocated to the interface after an IPv6 global unicast address is configured on the interface.

Step 4 Query the address configuration information about an IPv6 interface.

- Run the **display ipv6 interface** command to query the IPv6 interface information.
- Run the **display ipv6 statistics** command to query the IPv6 packet statistics.

----End

Example

To create VLAN 10, set the IPv6 address of VLANIF interface 10 to 2001:db8:1::1/64, and set the generation mode of the link-local address to the automatic generation mode, run the following commands:

```
huawei(config)#vlan 10
huawei(config)#interface vlanif 10
huawei(config-if-vlanif10)#ipv6 enable
huawei(config-if-vlanif10)#ipv6 address 2001:db8:1::1/64
huawei(config-if-vlanif10)#ipv6 address auto link-local
```

11.2.4.2 Configuring PMTU

By setting the path maximum transmission unit (PMTU), the device can select a proper maximum transmission unit (MTU) for packet transmission. In this manner, packets do not have to be fragmented during transmission. Employing PMTU reduces the load on intermediate devices, improves network resources utilization, and achieves optimal throughput on the network.

Context

Dynamic PMTU is enabled on a device by default, ensuring that the smallest MTU value is used on all interfaces along the path from the source to the destination node. You can also configure static PMTU to specify the maximum length of a packet that can be forwarded from the source to the destination node. Configuring static PMTU protects devices on a network from jumbo packets. Static PMTU has a higher priority over dynamic PMTU. When both static PMTU and dynamic PMTU are configured, only static PMTU takes effect.

Procedure

Step 1 Configure static PMTU.

In global config mode, run the **ipv6 pathmtu** command to configure a static PMTU value for a path destined for a specified IPv6 address. By default, the PMTU of the path destined for an IPv6 address is 1500 bytes.

The static PMTU value should be smaller than or equal to the MTU value of every interface on the same path. If the static PMTU value is larger than the MTU value of the interfaces, the system segments packets according to the MTU value. By

 Copyright © Huawei Technologies Co., Ltd.

manually configuring the static PMTU value based on the smallest MTU value of the path over which packets are transmitted, you can achieve a higher packet transmission rate.

Step 2 Configure the aging time of dynamic PMTU entries.

Run the **ipv6 pathmtu age** command to configure the aging time of dynamic PMTU entries. By default, the aging time of dynamic PMTU entries is 10 minutes.

The **ipv6 pathmtu age** command is used to modify the lifetime of dynamic PMTU entries in the buffer. The aging time is invalid on static PMTU entries, because static PMTU entries do not age.

Step 3 Query the PMTU information.

- Run the **display ipv6 pathmtu** command to query the PMTU information.
- Run the **display ipv6 interface** command to query the current MTU value of an IPv6 interface.

----End

Example

To set the PMTU for the IPv6 address 2001:db8:: to 1300 bytes, and to set the aging time of the dynamic PMTU entries to 40 minutes, run the following commands.

```
huawei(config)#ipv6 pathmtu 2001:db8:: 1300
huawei(config)#ipv6 pathmtu age 40
```

11.2.4.3 Configuring IPv6 Neighbor Discovery

The IPv6 Neighbor Discovery (ND) protocol provides a set of packets and processes for establishing the relationship between neighboring nodes. The IPv6 ND protocol supports the Address Resolution Protocol (ARP) messages, and the Router Discovery and Redirect messages of the Internet Control Message Protocol (ICMP) of IPv4. In addition, ND supports neighbor reachability detection.

Prerequisite

The IPv6 address has been configured. For details on the configuration method, see **Configuring an IPv6 Address for an Interface**.

Context

Most of the ND configurations are implemented based on the interface. The MA5800 supports ND configurations on VLAN Layer 3 interfaces.

Procedure

- Configure the static IPv6 neighbor.

  By configuring a static neighbor, the device can obtain the mapping of the IPv6 address and MAC address of the neighbor. The statically configured neighbor entries will overwrite the dynamically learned neighbor entries and will not age.

 Copyright © Huawei Technologies Co., Ltd.

a. In global config mode, run the **interface vlanif** command to enter the VLANIF mode.

b. In VLANIF mode, run the **ipv6 neighbor** command to configure a static IPv6 neighbor.

- Configure the parameters of the Router Advertisement (RA) message.

  The device periodically sends RA messages that contain information such as address prefixes, hop limit value, neighbor reachable time, and message lifetime. The IPv6 node on the local link receives the RA messages and updates its information, such as the IPv6 prefix list and other configuration data, according to the RA messages.

  a. In global config mode, run the **interface vlanif** command to enter the VLANIF mode.

  b. Run the **undo ipv6 nd ra halt** command to enable RA message advertising.

     - When a device is connected to an IPv6 node, the RA message advertising function needs to be enabled so that the device periodically sends RA messages to the IPv6 node.
     - When a device is not connected to an IPv6 node, the RA message advertising function does not need to be enabled. By default, this function is disabled.

     By default, RA message advertising is disabled on the device.

  c. (Optional) Configure parameters carried in an RA message.

     Perform the following operations as needed:

     - Run the **ipv6 nd ra** command to configure the interval for advertising RA messages.

       By default, the maximum interval is 600 seconds and the minimum interval is 200 seconds. The maximum interval cannot be shorter than the minimum interval. When the maximum interval is shorter than 9 seconds, the minimum interval is adjusted to the same value as the maximum interval.

     - Run the **ipv6 nd ra router-lifetime** command to configure the RA message lifetime.

       By default, the lifetime is 1800 seconds. The lifetime of the message advertised by the device must be longer than or equal to the interval at which the device advertises RA messages.

     - Run the **ipv6 nd ra prefix** command to configure the address prefixes to be advertised in RA messages.

       By default, RA messages contain only the address prefixes specified through the **ipv6 address** command. Run the **ipv6 nd ra prefix** command when you need the device to advertise only the specified prefixes.

     - Set the stateful autoconfiguration flag bit carried in an RA message.

 Copyright © Huawei Technologies Co., Ltd.

- Run the **ipv6 nd autoconfig managed-address-flag** command to set the flag bit carried in an RA message for stateful address autoconfiguration.

  After this flag bit is set, the host uses a stateful address autoconfiguration protocol; otherwise, the host uses a stateless address autoconfiguration protocol.

- Run the **ipv6 nd autoconfig other-flag** command to set the flag bit for stateful autoconfiguration of other information.

  After this flag bit is set, the host uses a stateful autoconfiguration protocol for configuring information other than the address.

- Run the **ipv6 nd ns retrans-timer** command to configure the interval for detecting neighbor reachability, that is, the neighbor solicitation (NS) message retransmission timer.

  Frequently sending NS packets helps to determine the neighbor reachability but also affects the device performance. Therefore, you are not advised to set the interval to a small value. The default interval, 1000 milliseconds, is recommended.

- Run the **ipv6 nd nud reachable-time** command to configure the neighbor reachable time specified in an RA message.

  The device detects neighbor reachability by using the neighbor unreachability detection (NUD) mechanism. The neighbor reachable time configured through the **ipv6 nd nud reachable-time** command is the interval for running a NUD detection. A smaller neighbor reachable time set on a device means that the device can probe the neighbor reachability more quickly but, meanwhile, more network bandwidth and CPU resources will be consumed. Therefore, you are not advised to set the neighbor reachable time to a short interval. The default interval, 30000 milliseconds, is recommended.

- Configure the hop limit value for the router.

  Run the **quit** command to quit the VLANIF mode. In global config mode, run the **ipv6 nd hop-limit** command to configure the hop limit value for the router, that is, the maximum number of hops for the IPv6 unicast packets initiated by the router. The hop limit value for the router is the same as the hop limit value specified in the RA message. The default value is 64.

- Configure duplicate address detection (DAD).

  DAD is used to check whether an IPv6 address is available. When a node is configured with an IPv6 address, it immediately sends an NS message to check whether this address is already used by other neighboring nodes.

  a. In global config mode, run the **interface vlanif** command to enter the VLANIF mode.
  b. Run the **ipv6 nd dad attempts** command to configure the number of DAD attempts, that is, number of attempts to send NS messages. The default value is 1.

 Copyright © Huawei Technologies Co., Ltd.

c. Run the **ipv6 nd ns retrans-timer** command to configure the interval of DAD, that is, the timer for retransmitting NS messages. The default interval, 1000 milliseconds, is recommended.

- Query the IPv6 neighbor information.
  - Run the **display ipv6 neighbors** command to query the IPv6 neighbor information.
  - Run the **display ipv6 interface** command to query the IPv6 interface information.

----End

Example

To configure the function of automatically generating a link-local unicast address on VLANIF interface 10, set the prefix to be advertised by the local unicast address of site EUI-64 and by the RA message to 2001:db8:3::2/64, set both the valid lifetime and the preferred lifetime of the prefix to 1000s, and enable RA message advertising (so that the host can automatically obtain the address prefix in the RA message), run the following commands:

```
huawei(config)#vlan 10
huawei(config)#interface vlanif 10
huawei(config-if-vlanif10)#ipv6 enable
huawei(config-if-vlanif10)#ipv6 address auto link-local
huawei(config-if-vlanif10)#ipv6 address 2001:db8:3::2/64 eui-64
huawei(config-if-vlanif10)#ipv6 nd ra prefix 2001:db8:3::2/64 1000 1000  { <cr>|no-autoconfig<K>|off-link<K> }:  Command:  ipv6 nd ra prefix 2001:db8:3::2/64 1000 1000
huawei(config-if-vlanif10)#undo ipv6 nd ra halt
```

11.2.5 IPv6 Reference Standards and Protocols

The following table lists the reference standards and protocols of the IPv6 feature.

Standard/ Protocol	Description
RFC1887	An Architecture for IPv6 Unicast Address Allocation
RFC1981	Path MTU Discovery for IP version 6
RFC2375	IPv6 Multicast Address Assignments
RFC2460	Version 6 of the Internet Protocol (IPv6), also sometimes referred to as IP Next Generation or IPng.
RFC2461/ RFC4861	Neighbor Discovery for IP Version 6 (IPv6)
RFC2462/ RFC4862	IPv6 Stateless Address Auto configuration
RFC2463/ RFC4443	Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification
RFC2464	Transmission of IPv6 Packets over Ethernet Networks

 Copyright © Huawei Technologies Co., Ltd.

Standard/ Protocol	Description
RFC2466	Management Information Base for IP Version 6 ICMPv6 Group
RFC2711	IPv6 Router Alert Option
RFC2893	Transition Mechanisms for IPv6 Hosts and Routers
RFC3315	DHCPv6 snooping
RFC3484	Default Address Selection for Internet Protocol Version 6 (IPv6)Section 2.1
RFC3493	Basic Socket Interface Extensions for IPv6
RFC3633	IPv6 Prefix Options for Dynamic Host Configuration Protocol(DHCP) version 6
RFC3736	Stateless Dynamic Host Configuration Protocol (DHCP) Service for IPv6
RFC3849	IPv6 Address Prefix Reserved for Documentation
RFC4001	Textual Conventions for Internet Network Addresses
RFC4007	IPv6 Scoped Address Architecture
RFC4191	Default Router Preferences and More-Specific Routes
RFC4193	Unique Local IPv6 Unicast Addresses
RFC4213	Basic Transition Mechanisms for IPv6 Hosts and Routers
RFC4214	Intra-Site Automatic Tunnel Addressing Protocol(ISATAP)
RFC4429	Duplicate Address Detection
RFC4282	A Model of IPv6/IPv4 Dual Stack Internet Access Service
RFC2373/ RFC3513/ RFC4291	Internet Protocol Version 6 (IPv6) Addressing Architecture
RFC4862/ RFC5006	Router Advertisement (RA) filtering
RFC6221	DHCPv6 LDRA

11.3 ARP

 Copyright © Huawei Technologies Co., Ltd.

11.3.1 Introduction

Definition

The Address Resolution Protocol (ARP) is an Internet protocol used to map IP addresses to MAC addresses.

Purpose

To communicate with each other, the sending device must know the network-layer IP address of the receiving device. IP datagrams, however, must be encapsulated with MAC addresses before they can be transmitted over the physical network. Therefore, a resolution from the known IP address to a MAC address is required.

ARP was developed to enable communication between devices.

Function Overview

Table 11-3 lists ARP features.

Table 11-3 ARP features

Feature	Description	Usage Scenario
Dynamic ARP	Devices dynamically learn and update the mapping between IP addresses and MAC addresses by exchanging ARP messages.	Real-time communication is a priority, or network resources are insufficient.
Static ARP	The mapping between IP addresses and MAC addresses is manually created and cannot be dynamically modified.	Communication security is a priority, and network resources are sufficient.
Gratuitous ARP	A device broadcasts gratuitous ARP packets that carry the local IP address as both the source and destination IP addresses to notify the other devices on the same network segment of its address information.	Gratuitous ARP is used to check whether the local IP address conflicts with that of another device, to notify other devices on the same network segment of the new MAC address after the local network interface card is replaced, or to notify master/slave switchovers in a Virtual Router Redundancy Protocol (VRRP) backup group.

 Copyright © Huawei Technologies Co., Ltd.

Benefits

ARP ensures Ethernet communication by mapping IP addresses at the network layer to MAC addresses at the link layer on Ethernet networks.

11.3.2 Principles

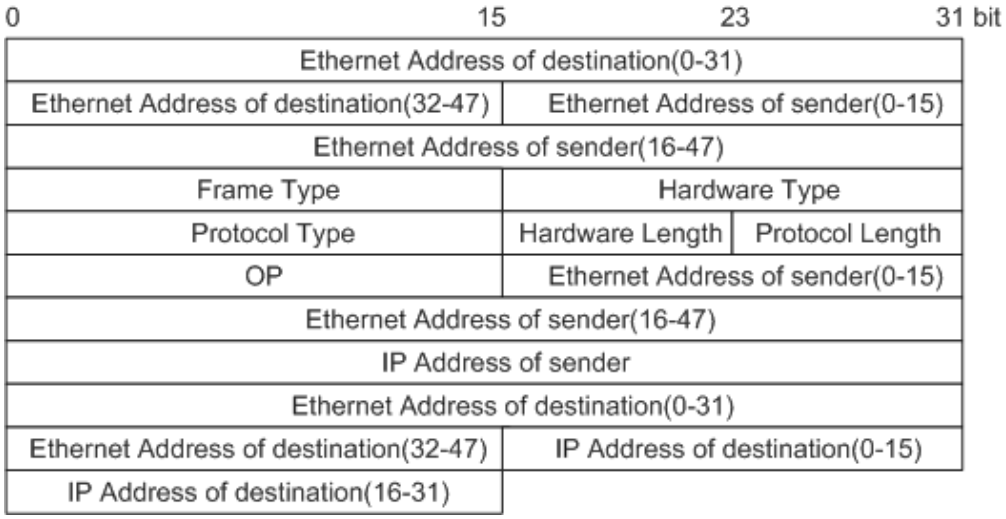
11.3.2.1 Basic Principles

Related Concepts

- **Address Resolution Protocol (ARP) messages**

  An ARP message can be an ARP request or reply message. **Figure 11-14** shows the ARP message format.

  Figure 11-14 ARP message format



  NOTE

  The **Ethernet Address of destination** field contains a total of 48 bits. **Ethernet Address of destination (0-31)** indicates the first 32 bits of the **Ethernet Address of destination** field, and **Ethernet Address of destination (32-47)** indicates the last 16 bits of the **Ethernet Address of destination** field.

  An ARP message consists of 42 bytes. The first 14 bytes indicate the Ethernet frame header, and the last 28 bytes are the ARP request or reply message content. **Table 11-4** describes the fields in an ARP message.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-4 Description of fields in an ARP message

Field	Length	Description
Ethernet Address of destination	48 bits	Ethernet destination MAC address in the Ethernet frame header. This field in an ARP Request message is the broadcast MAC address, with a value of 0xFF-FF-FF-FF-FF-FF.
Ethernet Address of sender	48 bits	Ethernet source MAC address in the Ethernet frame header.
Frame Type	16 bits	Data type. For an ARP request or reply message, the value of this field is 0x0806.
Hardware Type	16 bits	Hardware address type. For an Ethernet network, the value of this field is 1.
Protocol Type	16 bits	Type of the protocol address to be mapped by the sending device. For an IP address, the value of this field is 0x0800.
Hardware Length	8 bits	Hardware address length. For an ARP request or reply message, the value of this field is 6.
Protocol Length	8 bits	Protocol address length. For an ARP request or reply message, the value of this field is 4.
OP	16 bits	Operation type. The values are as follows: ● 1: ARP request ● 2: ARP reply ● 3: RARP request ● 4: RARP reply
Ethernet Address of sender	48 bits	Source MAC address. The value of this field is the same as the Ethernet source MAC address in the Ethernet frame header.
IP Address of sender	32 bits	Source IP address.
Ethernet Address of destination	48 bits	Destination MAC address. The value of this field in an ARP Request message is 0x00-00-00-00-00-00.
IP Address of destination	32 bits	Destination IP address.

- **ARP table**

  An ARP table contains the latest mapping between IP addresses and MAC addresses. If a host always broadcasts an ARP Request message for a MAC address before it sends an IP datagram, network communication traffic will greatly increase. Furthermore, all other hosts on the network have to receive and process the ARP Request messages, which lowers network efficiency. To

 Copyright © Huawei Technologies Co., Ltd.

solve this problem, an ARP table is maintained on each host to ensure efficient ARP operations. The mapping between an IP address and a MAC address is called an ARP entry.

ARP entries can be classified as dynamic or static.

- Dynamic ARP entries are automatically generated and maintained by using ARP messages. Dynamic ARP entries can be aged and overwritten by static ARP entries.
- Static ARP entries are manually configured and maintained by a network administrator. Static ARP entries can neither be aged nor be overwritten by dynamic ARP entries.

Before sending IP datagrams, a host searches the ARP table for the MAC address corresponding to the destination IP address.

- If the ARP table contains the corresponding MAC address, the host directly sends the IP datagrams to the MAC address instead of sending an ARP Request message.
- If the ARP table does not contain the corresponding MAC address, the host broadcasts an ARP Request message to request the MAC address of the destination host.

- **Reverse Address Resolution Protocol (RARP)**

  If only the MAC address of a host is available, the host can send and receive RARP messages to obtain its IP address.

  To do so, the network administrator must establish the mapping between MAC addresses and IP addresses on a gateway. When a new host must be configured, its RARP client requests the host's IP address from the RARP server on the gateway.

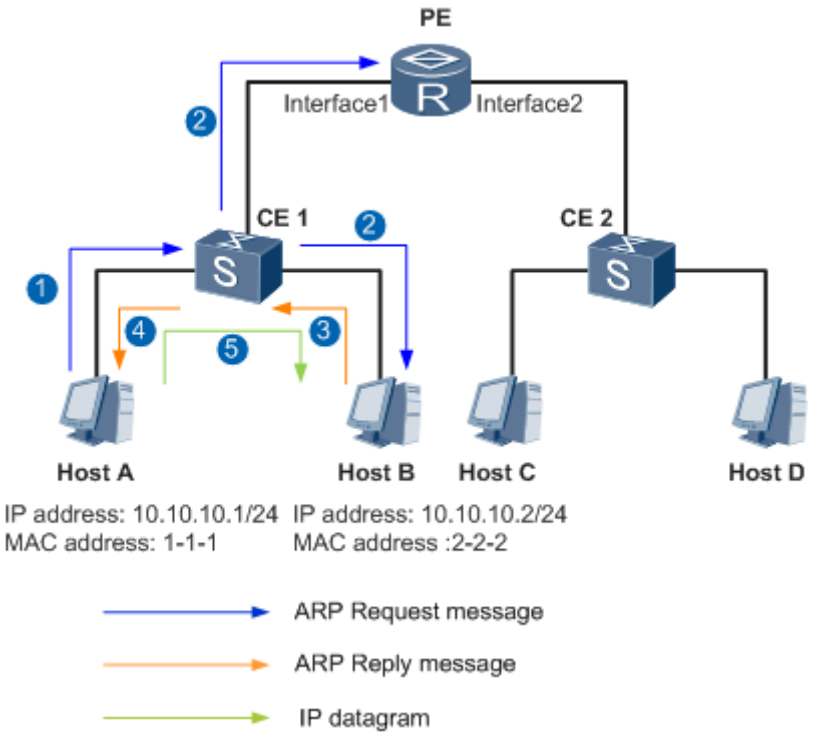
Implementation

- **ARP implementation within a network segment**

  Figure 11-15 illustrates how ARP is implemented within a network segment, by using IP datagram transmission from Host A to Host B as an example.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-15 ARP implementation between Host A and Host B on the same network segment



a. Host A searches its ARP table and does not find the mapping between the IP address and MAC address of Host B. Host A then sends an ARP Request message for the MAC address of Host B. In this ARP Request message, the source IP address and source MAC address are respectively the IP address and MAC address of Host A, the destination IP address and destination MAC address are respectively the IP address of Host B and 00-00-00-00-00-00, and the Ethernet source MAC address and Ethernet destination MAC address are respectively the MAC address of Host A and the broadcast MAC address.

b. After CE1 receives the ARP Request message, CE1 broadcasts it on the network segment.

c. After Host B receives the ARP Request message, Host B adds the MAC address of Host A to its ARP table and sends an ARP Reply message to Host A. In this ARP Reply message, the source IP and MAC addresses are respectively the IP and MAC addresses of Host B, the destination IP and MAC addresses are respectively the IP and MAC addresses of Host A, and the Ethernet source and destination MAC addresses are respectively the MAC addresses of Host B and Host A.

   NOTE

   The PE also receives the ARP Request message but discards it because the destination IP address in the ARP Request message is not its own IP address.

d. CE1 receives the ARP Reply message and forwards it to Host A.

e. After Host A receives the ARP Reply message, host A adds the MAC address of Host B to its ARP table and sends the IP datagrams to Host B.

 Copyright © Huawei Technologies Co., Ltd.

- **ARP implementation between different network segments**

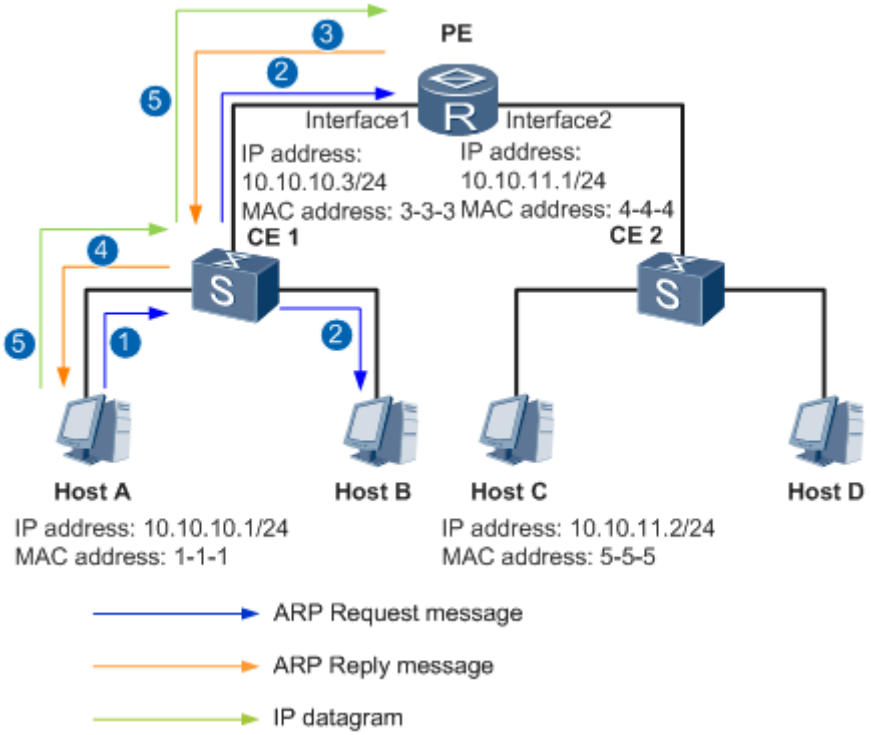
  NOTE

  ARP messages are Layer 2 messages. Therefore, ARP is applicable only to devices on the same network segment. If two hosts on different network segments need to communicate, the source host sends IP datagrams to the default gateway, which in turns forwards the IP datagrams to the destination host. ARP implementation between different network segments involves separate ARP implementation within network segments. In this manner, hosts on different network segments can communicate.

  The following examples show how ARP is implemented between different network segments, by using IP datagram transmission from Host A to Host C as an example.

  Figure 11-16 illustrates how ARP is implemented between Host A and the PE on the same network segment.

  Figure 11-16 ARP implementation between Host A and the PE



  a. Host A searches its ARP table and does not find the mapping between the IP address and MAC address of Interface 1 on the default gateway PE that connects to Host C. Host A then sends an ARP Request message for the MAC address of the PE's Interface. In this ARP Request message, the source IP and MAC addresses are respectively the IP and MAC addresses of Host A, the destination IP and MAC addresses are respectively the IP address of the PE's Interface 1 and 00-00-00-00-00-00, and the Ethernet source and destination MAC addresses are respectively the MAC address of Host A and the broadcast MAC address.

  b. After CE1 receives the ARP Request message, CE1 broadcasts it on the network segment.

 Copyright © Huawei Technologies Co., Ltd.

c. After the PE receives the ARP Request message, the PE adds the MAC address of Host A to its ARP table and sends an ARP Reply message to Host A. In this ARP Reply message, the source IP and MAC addresses are respectively the IP and MAC addresses of the PE's Interface 1, the destination IP and MAC addresses are respectively the IP and MAC addresses of Host A, and the Ethernet source and destination MAC addresses are respectively the MAC address of Interface 1 on PE and the MAC address of Host A.

NOTE

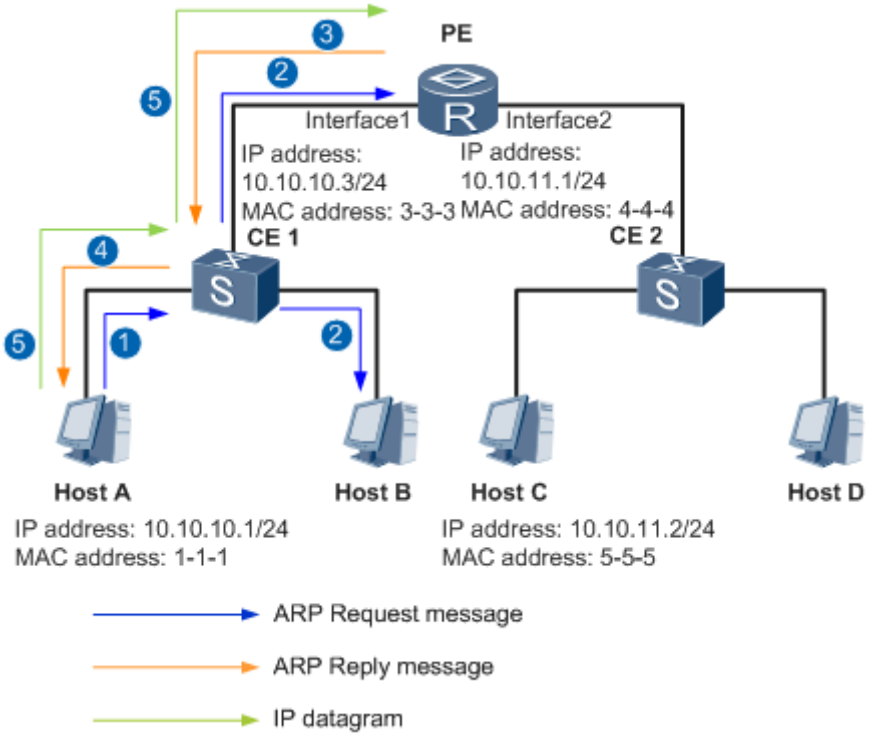
Host B also receives the ARP Request message but discards it because the destination IP address in the ARP Request message is not its own IP address.

d. CE1 receives the ARP Reply message and forwards it to Host A.

e. After Host A receives the ARP Reply message, host A adds the MAC address of the PE's Interface 1 to its ARP table and sends the IP datagrams to the PE.

Figure 11-17 illustrates ARP implementation between the PE and Host C on the same network segment.

Figure 11-17 ARP implementation between the PE and Host C



The PE searches its routing table and sends the IP datagrams from Interface 1 to Interface 2.

a. The PE searches its ARP table and does not find the mapping between the IP address and MAC address of Host C. Then, the PE sends an ARP Request message for the MAC address of Host C. In this ARP Request message, the source IP and MAC addresses are respectively the IP and

 Copyright © Huawei Technologies Co., Ltd.

MAC addresses of the PE's Interface 2, the destination IP and MAC addresses are respectively the Host C's IP address and 00-00-00-00-00-00, and the Ethernet source and destination MAC address are respectively the MAC address of Interface 2 on PE and the broadcast MAC address.

b. After CE2 receives the ARP Request message, CE2 broadcasts it on the network segment.

c. After Host C receives the ARP Request message, Host C adds the MAC address of the PE's Interface 2 to its ARP table and sends an ARP Reply message to the PE. In this ARP Reply message, the source IP and MAC addresses are respectively the IP and MAC addresses of Host C, the destination IP and MAC addresses are respectively the IP and MAC addresses of the PE's Interface 2, and the Ethernet source and destination MAC addresses are respectively the MAC address of Host C and the MAC address of Interface 2 on PE.

NOTE

Host D also receives the ARP Request message but discards it because the destination IP address in the ARP Request message is not the IP address of itself.

d. CE2 receives the ARP Reply message and forwards it to the PE.

e. After the PE receives the ARP Reply message, the PE adds the MAC address of Host C to its ARP table and sends the IP datagrams to Host C.

So far, the IP datagram transmission from Host A to Host C is complete.

NOTE

1. ARP Request messages are broadcast, whereas ARP Reply messages are unicast.
2. In ARP implementation, the switches CE1 and CE2 transparently forward IP datagrams and do not modify them.

11.3.2.2 Dynamic ARP

Definition

The Dynamic Address Resolution Protocol (ARP) indicates that devices dynamically learn and update the mapping between IP addresses and MAC addresses by using ARP messages. You do not need to manually configure the mapping.

Related Concepts

- Dynamic ARP involves the dynamic ARP aging mechanism.

  The dynamic ARP aging mechanism enables an ARP entry that is not used over a specified period to be automatically deleted. This mechanism helps reduce storage space of ARP tables and speed up ARP table queries.

  Table 11-5 describes concepts related to the dynamic ARP aging mechanism.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-5 Concepts related to the dynamic ARP aging mechanism

Concept	Description	Usage Scenario
Aging probe mode	Before a dynamic ARP entry on a device is aged, the device sends ARP aging probe messages to the other devices on the same network segment. An ARP aging probe message can be a unicast or broadcast message. By default, a device broadcasts ARP aging probe messages.	• If the IP address of the peer device remains unchanged but its MAC address changes frequently, it is recommended that you configure ARP aging probe messages to be broadcast. • If the MAC address of the peer device remains unchanged, network bandwidth resources are insufficient, and the aging time of ARP entries is set to a small value, it is recommended that you configure ARP aging probe messages to be unicast.
Aging time	A dynamic ARP entry has a life cycle. If a dynamic ARP entry is not updated before its life cycle ends, this dynamic ARP entry is deleted from the ARP table. The life cycle is called aging time.	Two interconnected devices can learn the mapping between their IP and MAC addresses using ARP and can save the mapping in their ARP tables. Then, the two devices can communicate by using the ARP entries. When the peer device becomes faulty, or the network adapter of the peer device is replaced but the local device does not receive any status change information about the peer device, the local device continues sending IP datagrams to the peer device. As a result, network traffic is interrupted because the ARP table of the local device is not promptly updated. To reduce the risk of network traffic interruption, an aging timer can be set for each ARP entry. After the aging timer of a dynamic ARP entry expires, the entry is automatically deleted.

 Copyright © Huawei Technologies Co., Ltd.

Concept	Description	Usage Scenario
Number of aging probe attempts	Before a dynamic ARP entry is aged, the local device sends ARP aging probe messages to the peer device. If the local device does not receive an ARP Reply message after the number of aging probe attempts reaches a specified number, the dynamic ARP entry is deleted.	The ARP aging timer can help reduce the risk of network traffic interruptions that occur because an ARP table is not updated quickly enough, but cannot eliminate problems due to delays. Specifically, if the dynamic ARP entry aging time is *N* seconds, the local device can detect the status change of the peer device after *N* seconds. During the *N* seconds, the ARP table of the local device is not updated. If the number of aging probe attempts is specified, the local device can obtain the status change information about the peer device and update its ARP table.

Implementation

Dynamic ARP entries can be created, updated, and aged.

- Creating and updating dynamic ARP entries

  If a device receives an ARP message that meets either of the following conditions, the device automatically creates or updates an ARP entry:

  - The source IP address of the ARP message is on the same network segment as the IP address of the inbound interface. The destination IP address of the ARP message is the IP address of the inbound interface.
  - The source IP address of the ARP message is on the same network segment as the IP address of the inbound interface. The destination IP address of the ARP message is the virtual IP address of the Virtual Router Redundancy Protocol (VRRP) backup group configured on the interface on the device.

- Aging dynamic ARP entries

  After the aging timer of a dynamic ARP entry on a device expires, the device sends ARP aging probe messages to the peer device. If the device does not receive an ARP Reply message after the number of aging probe attempts reaches a specified number, the dynamic ARP entry is aged.

NOTE

The MA5800 supports semi-static ARP for a specified IP address or interface. Semi-static ARP entries are dynamically learned and can be updated but cannot be aged. After the semi-static ARP attribute is deleted, the generated ARP entry can be aged.

Usage Scenario

Dynamic ARP is applicable to a network with a complex topology, insufficient bandwidth resources, and a high requirement for real-time communication.

 Copyright © Huawei Technologies Co., Ltd.

Benefits

Dynamic ARP entries are dynamically created and updated using ARP messages. They do not need to be manually maintained, greatly reducing maintenance workload.

11.3.2.3 Static ARP

Definition

The Static Address Resolution Protocol (ARP) indicates that the mapping between IP addresses and MAC addresses is manually created by a network administrator.

Principles

The main difference between static ARP and dynamic ARP is how ARP entries are generated and maintained. Dynamic ARP entries are automatically generated and maintained using ARP packets. Static ARP entries are manually configured and maintained by network administrators. The advantages and disadvantages of dynamic ARP and static ARP are as follows:

- Dynamic ARP
  - Advantages: Dynamic ARP entries free network administrators from manual configuration and maintenance. Especially when the network device is faulty or the network adapter of the host is frequently changed, the ARP entries can be dynamically updated in real time, which greatly reduces the maintenance workload of the network administrator.
  - Disadvantages:
    - Dynamic ARP entries can be aged or overwritten by new dynamic ARP entries, which cannot ensure the stability and security of network communication.
    - The execution of dynamic ARP consumes certain network resources. Therefore, dynamic ARP does not apply to networks with insufficient bandwidth resources and may affect user services.
- Static ARP
  - Advantages:
    - Static ARP entries are not aged or overwritten by dynamic ARP entries, ensuring network stability.
    - With static ARP, IP addresses and MAC addresses are bound to prevent network attackers from tampering with ARP entries using ARP packets, ensuring network communication security.
    - Static ARP eliminates the need of dynamic ARP execution and reduces network resource consumption.
  - Disadvantages: Static ARP entries need to be manually configured by network administrators, causing heavy maintenance workload when it comes to frequent changes of the network structure.

To ensure the stability and security of network communication, deploy static ARP based on actual requirements and network resources.

 Copyright © Huawei Technologies Co., Ltd.

Static ARP can implement the following functions:

- Binds IP addresses to the MAC address of a specified gateway so that IP datagrams destined to these IP addresses must be forwarded by this gateway.
- Binds the destination IP addresses of IP datagrams sent by a specified host to a nonexistent MAC address, helping filter out unwanted IP datagrams.

Related Concepts

Static ARP entries are classified as short or long entries.

- Short static ARP entries

  Short static ARP entries contain only IP and MAC addresses. A device still has to send ARP Request messages. If the source IP and MAC addresses of the received reply messages are the same as the configured IP and MAC addresses, the device adds the interface that receives the ARP Reply messages to the static ARP entries. The device can use this interface to forward subsequent messages directly. Short static ARP entries cannot be directly used to forward messages

  Configure short static ARP entries to enable the device and a specified host to communicate using fixed IP and MAC addresses.

  NOTE

  In Network Load Balancing (NLB) scenarios, you must configure both MAC entries with multiple outbound interfaces and short static ARP entries for the gateway. These MAC entries and short static ARP entries must have the same MAC address. In this scenario, short static ARP entries are also called ARP entries with multiple outbound interfaces, and cannot be updated manually.

- Long static ARP entries

  Long static ARP entries contain IP and MAC addresses as well as the VLAN and outbound interface through which devices send packets. Long static ARP entries are directly used to forward messages.

  Configure long static ARP entries to further specify the interface through which only the host can communicate with the device in a VLAN.

Usage Scenario

Static ARP applies to the following scenarios:

- Networks with a simple topology and high stability
- Networks on which information security is of high priority, such as a government or military network
- Short static ARP entries mainly apply to scenarios in which network administrators want to bind hosts' IP and MAC addresses but hosts' access interfaces can change.

Benefits

Static ARP ensures communication security. If a static ARP entry is configured on a device, the device can communicate with the peer device using only the specified MAC address. Network attackers cannot modify the mapping between the IP address and MAC address by using ARP messages, ensuring normal communication between the two devices.

 Copyright © Huawei Technologies Co., Ltd.

11.3.2.4 Gratuitous ARP

Principles

To ensure the stability and reliability of network communication, a device can broadcast gratuitous Address Resolution Protocol (ARP) messages to notify the other devices on the same network segment of its address information in the following scenarios:

- You need to check whether the IP address of a device conflicts with the IP address of another device on the same network segment. The IP address of each device must be unique to ensure the stability of network communication.
- After the MAC address of a host changes after its network adapter is replaced, the host must quickly notify other devices on the same network segment of the MAC address change before the ARP entry is aged. This ensures the reliability of network communication.
- When a master/slave switchover occurs in a Virtual Router Redundancy Protocol (VRRP) backup group, the new master router needs to notify other devices on the same network segment of its status change.

Related Concepts

Gratuitous ARP involves gratuitous ARP messages.

A gratuitous ARP message is a special ARP message that carries the sending device's IP address as both the source and destination IP addresses.

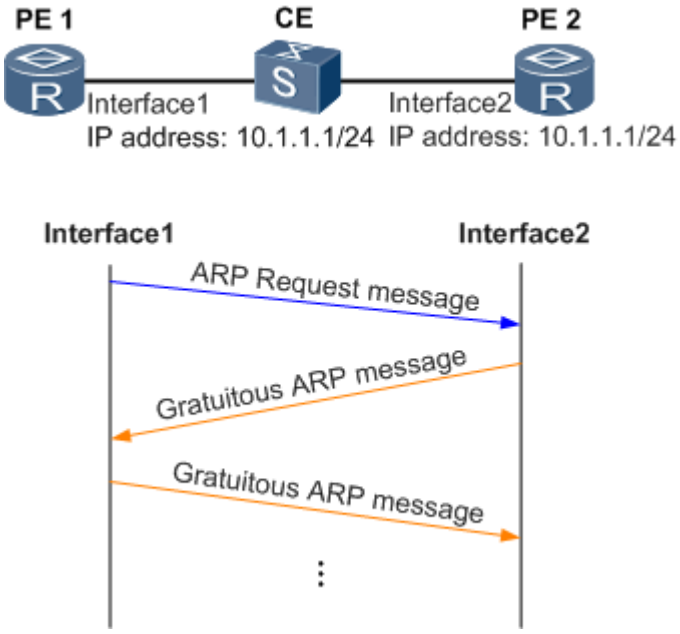
Implementation

- If a device finds that the source IP address in a received gratuitous ARP message is the same as its own IP address, the device sends a gratuitous ARP message to notify the sending device of the address conflict.
- If a device finds that the source IP address in a received gratuitous ARP message is different from its own IP address, the device updates the corresponding ARP entry with the sending device's IP address and MAC address carried in the gratuitous ARP message.

Figure 11-18 illustrates how gratuitous ARP is implemented.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-18 Gratuitous ARP implementation



As shown in **Figure 11-18**, the IP address of Interface 1 on PE1 is 10.1.1.1, and the IP address of Interface 2 on PE2 is 10.1.1.1.

1. Interface 1 broadcasts an ARP Request message. Interface 2 receives the ARP Request message and finds that the source IP address in the message conflicts with its own IP address. Then, Interface 2 performs the following operations:
   a. Interface 2 sends a gratuitous ARP message to notify Interface 1 of its IP address.
   b. A conflict node is generated on the conflict link of Interface 2. Interface 2 then sends gratuitous ARP messages to Interface 1 at an interval of 5 seconds.
2. Interface 1 receives the gratuitous ARP messages from Interface 2 and finds that the source IP address in the message conflicts with its own IP address. Interface 1 then performs the following operations:
   a. Interface 1 sends a gratuitous ARP message to notify Interface 2 of its IP address.
   b. A conflict node is generated on the conflict link of Interface 1. Interface 1 then sends gratuitous ARP messages to Interface 2 at an interval of 5 seconds.

Interface 1 and Interface 2 send gratuitous ARP messages to each other at an interval of 5 seconds until the address conflict is rectified.

If one interface does not receive a gratuitous ARP message from the other interface within 8 seconds, the interface considers the address conflict rectified. The interface deletes the conflict node on its conflict link and stops sending gratuitous ARP messages to the other interface.

 Copyright © Huawei Technologies Co., Ltd.

Functions

Gratuitous ARP implements the following functions:

- To check for IP address conflict on the network, send a gratuitous ARP message from a device. If the device receives a gratuitous ARP message from another device, the IP addresses of the two devices conflict.
- When the MAC address of a host changes after its network adapter is replaced, the host sends a gratuitous ARP message to notify other devices of the MAC address change before the ARP entry is aged. This ensures the reliability of network communication. After receiving the gratuitous ARP message, other devices maintain the corresponding ARP entry in their ARP tables based on the address information carried in the message.
- When a master/slave switchover occurs in the VRRP backup group, the new master router sends a gratuitous ARP message to notify other devices on the network of its status change.

Benefits

Gratuitous ARP reveals address conflict on a network so that ARP tables of devices can be quickly updated. This feature ensures the stability and reliability of network communication.

11.3.2.5 IP Address Conflict Detection

Background

The occurrence of an IP address conflict causes route flapping and traffic interruptions, affecting user services. IP address conflicts are often caused by incorrect networking or configurations. Users expect that devices can automatically detect IP address conflicts on a network and immediately notify users of conflict reasons, so that they can rapidly resolve such conflicts and minimize impact on services.

IP address conflict detection can help users quickly locate and modify the conflicted IP addresses and instruct users to properly configure and manage the IP addresses of devices on a network.

Implementation

IP address conflict detection can be classified into active and passive detection, and their differences are as follows:

- Active detection

  When the physical status of an interface on a device changes to Up, the device actively sends gratuitous ARP packets to detect possible IP address conflicts. For the detailed detection procedure, see **11.3.2.4 Gratuitous ARP**.

- Passive detection

  When a device receives ARP packets that are not gratuitous ARP packets, it checks the IP addresses carried by the ARP packets. The device concludes that IP address conflicts exist on the network if any of the following conditions are met:

 Copyright © Huawei Technologies Co., Ltd.

- The source IP address in an ARP packet is the same as the IP address of the inbound interface that receives the ARP packet, but the source MAC address in the ARP packet is different than the MAC address of the inbound interface.
- The source IP address in an ARP packet is the same as the IP address in an existing ARP entry, but the source MAC address is different than the MAC address in the ARP entry.
- The source IP address in an ARP packet is different than the CE IP address configured on the inbound interface that connects to the CE, or the source MAC address is different than the CE MAC address configured on the inbound interface that connects to the CE.

Usage Scenario

IP address conflict detection is applicable to Ethernet LANs.

Benefits

IP address conflict detection helps users quickly locate and modify IP address conflicts to ensure stability and security of user services.

11.3.3 Configuring ARP

This topic describes how to configure dynamic ARP, static ARP, and gratuitous ARP.

Context

Dynamic ARP, static ARP, and gratuitous ARP are applicable for different usage scenario. Therefore, configure them according to actual requirements.

Procedure

- Configuring dynamic ARP parameters

  Proper dynamic ARP parameter adjustment improves network reliability. The following lists optional steps.

  a. Run the **arp aging-mode** command to set the aging mode of the dynamic ARP entries.
  b. Run the **arp aging-time** command to set the aging time of the dynamic ARP entries.
  c. Run the **arp learning uni enable** command to enable user network ARP learning function for VLAN interface. By default, the system enables user network ARP learning function for VLAN interface.
  d. Run the **arp learning strict enable** command to enable dynamic ARP strict learning. By default, the system enables dynamic ARP strict learning.

- Configuring static ARP

  Static ARP applies to networks with a simple topology and high stability or networks on which information security is of high priority.

 Copyright © Huawei Technologies Co., Ltd.

a. Run the **arp** command to configure the static ARP mapping entries to specify a static mapping between an IP address and a MAC address.

   For boards on which service flows in Ethernet access mode are created by running the **service-port** command, run the **arp** *ip_addr mac_addr vlanid* **service-port** *service-portid* command to configure static ARP. If you configure static ARP by port, a configuration failure message is displayed.

- Configuring gratuitous ARP

  Gratuitous ARP reveals address conflict on a network so that ARP tables of devices can be quickly updated. This feature ensures the stability and reliability of network communication.

  a. Run the **arp gratuitous-arp period-send-mode** command to enable VLAN interfaces to periodically send gratuitous ARP packets. By default, VLAN interfaces are enabled to periodically send gratuitous ARP packets.
  b. Run the **arp conflict detect enable** command to enable the ARP conflict detection. By default, the system disables the ARP conflict detection.

----End

Result

Run the **display arp config** command to query ARP configurations.

Example

- Assume that:
- IP address of the LAN: 10.10.10.1
- Mapped MAC address: 00e0-fc01-0000
- Index of the service port: 0 (belongs to VLAN 10)

To configure static ARP by service flow, do as follows:

```
huawei(config)#arp 10.10.10.1 00e0-fc01-0000 10 service-port 0
```

- Assume that:
- IP address of the LAN: 10.10.20.1
- Mapped MAC address: 00e0-fc01-0022
- Port ID: 0/1/0 (belongs to VLAN 100)

To configure static ARP by port, do as follows:

```
huawei(config)#arp 10.10.20.1 00e0-fc01-0022 100 0/1/0
```

11.3.4 ARP Reference Standards and Protocols

The following table lists the references.

Document No.	Document Name
RFC 826	Ethernet Address Resolution Protocol
RFC 903	Reverse Address Resolution Protocol

 Copyright © Huawei Technologies Co., Ltd.

Document No.	Document Name
RFC 1027	Using ARP to Implement Transparent Subnet Gateways
RFC 1042	Standard for the Transmission of IP Datagrams over IEEE 802 Networks

11.4 ARP Proxy

ARP proxy is used to communication between users in different VLAN or in the same VLAN but isolated from each other by layer 3 forwarding.

11.4.1 Introduction to ARP proxy

Definition

When a host sends an ARP request to another host, the request is processed by the access device connected to the two hosts. This process is called ARP proxy.

Purpose

On the MA5800, ARP proxy is often used for interconnection between sub VLANs in a super VLAN.

11.4.2 Proxy ARP

Principles

The Address Resolution Protocol (ARP) is applicable only to devices on the same physical network. When a device on a physical network needs to send IP datagrams to another physical network, the gateway is used to query the routing table to implement communication between the two networks. However, routing table query consumes system resources and can affect other services. To resolve this problem, deploy proxy ARP on an intermediate device. The proxy ARP feature helps reduce system resource consumption caused by routing table queries and improves the efficiency of system processing.

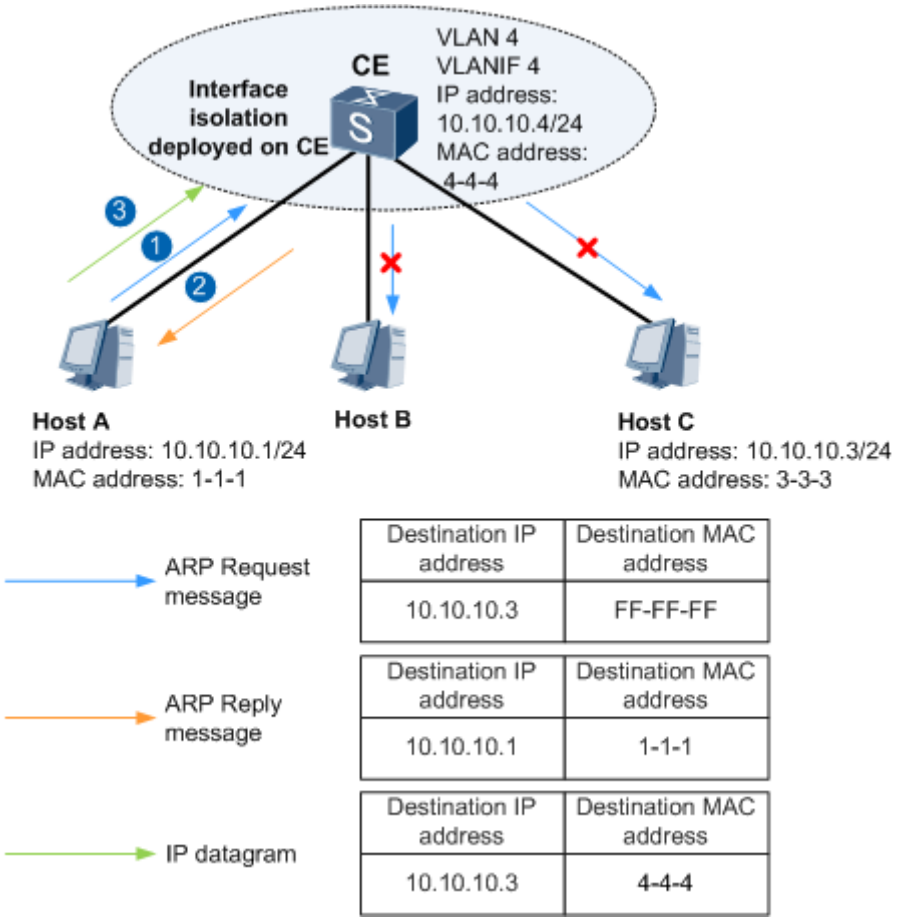
Implementation

- **Proxy ARP within a VLAN**

  Figure 11-19 illustrates how proxy ARP is implemented within a VLAN by using the communication between Host A and Host C as an example.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-19 Proxy ARP within a VLAN implementation



Host A, Host B, and Host C belong to the same VLAN. Port isolation is configured on the CE. Therefore, Host A and Host C cannot communicate at Layer 2. You can configure a VLANIF interface on the CE and enable proxy ARP within a VLAN to allow Host A and Host C to communicate.

a. Host A sends an ARP Request message for the MAC address of Host C.
b. After the CE receives the ARP Request message, the CE checks the destination IP address of the message and finds that it is not its own IP address and determines that the requested MAC address is not the MAC address of its VLANIF 4. The CE then searches its ARP table for the ARP entry indicating the mapping between the IP address and MAC address of Host C.

   - If the CE finds this ARP entry in its ARP table, the CE checks whether proxy ARP within a VLAN is enabled on it.
     - If proxy ARP within a VLAN is enabled on the CE, the CE sends the MAC address of its VLANIF 4 to Host A.
     - If proxy ARP within a VLAN is not enabled on the CE, the CE discards the ARP Request message sent by Host A.
   - If the CE does not find this ARP entry in its ARP table, the CE discards the ARP Request message sent by Host A and checks whether proxy ARP within a VLAN is enabled on it.

 Copyright © Huawei Technologies Co., Ltd.

- If proxy ARP within a VLAN is enabled on the CE, the CE broadcasts the ARP Request message with the IP address of Host C as the destination IP address within VLAN 4. After the CE receives an ARP Reply message from Host C, the CE generates an ARP entry indicating the mapping between the IP address and MAC address of Host C.
- If proxy ARP within a VLAN is not enabled on the CE, the CE does not perform any operations.

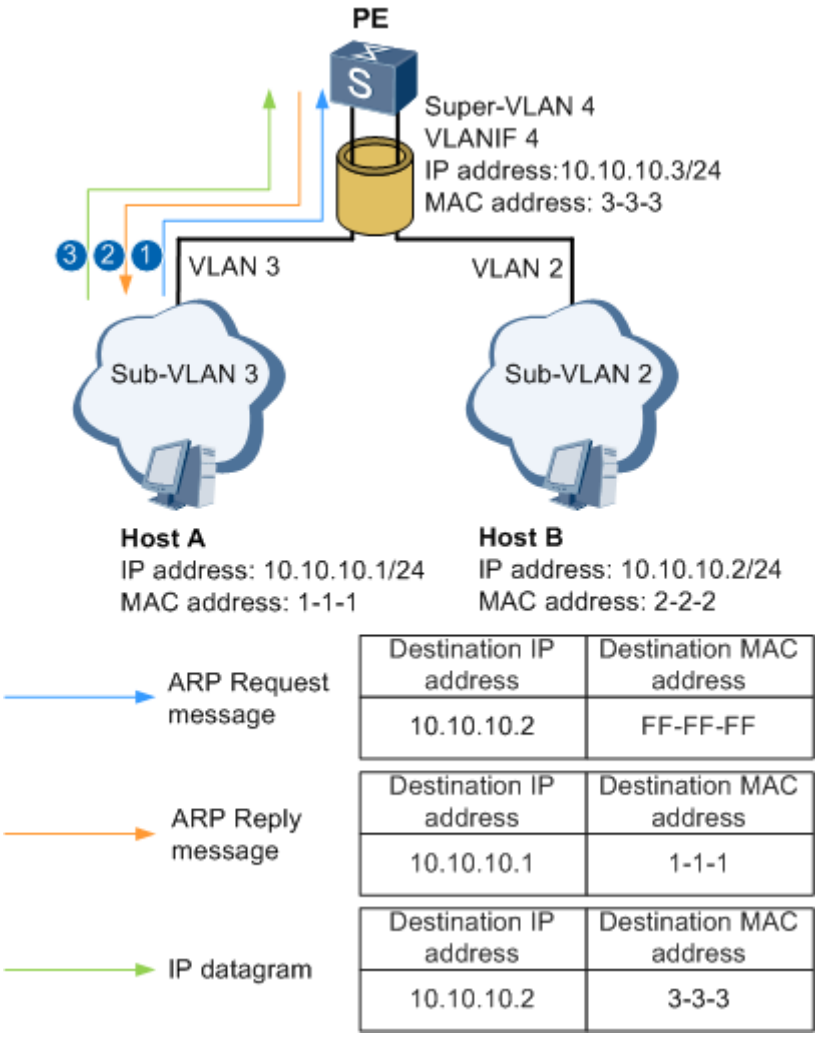
c. After Host A learns the MAC address of VLANIF 4, Host A sends IP datagrams to the CE based on this MAC address.

The CE receives the datagrams and forwards them to Host C.

- **Proxy ARP between VLANs**

  Figure 11-20 illustrates how proxy ARP is implemented between VLANs by using the communication between Host A and Host B as an example.

  Figure 11-20 Proxy ARP between VLANs implementation



Host A belongs to VLAN 3, whereas Host B belongs to VLAN 2. Therefore, Host A cannot communicate with Host B. You can configure a VLANIF

 Copyright © Huawei Technologies Co., Ltd.

interface on the PE and enable proxy ARP between VLANs on super-VLAN 4 to allow Host A and Host B to communicate.

a. Host A sends an ARP Request message for the MAC address of Host B.
b. After the PE receives the ARP Request message, the PE checks the destination IP address of the message and finds that it is not its own IP address and determines that the requested MAC address is not the MAC address of its VLANIF 4. The PE then searches its ARP table for the ARP entry indicating the mapping between the IP address and MAC address of Host B.
   - If the PE finds this ARP entry in its ARP table, the PE checks whether proxy ARP between VLANs is enabled on it.
     - If proxy ARP between VLANs is enabled on the PE, the PE sends the MAC address of its VLANIF 4 to Host A.
     - If proxy ARP between VLANs is not enabled on the PE, the PE discards the ARP Request message sent by Host A.
   - If the PE does not find this ARP entry in its ARP table, the PE discards the ARP Request message sent by Host A and checks whether proxy ARP between VLANs is enabled on it.
     - If proxy ARP within a VLAN is enabled on the CE, the CE broadcasts the ARP Request message with the IP address of Host B as the destination IP address within VLAN 2. After the PE receives an ARP Reply message from Host B, the PE generates an ARP entry indicating the mapping between the IP address and MAC address of Host B.
     - If proxy ARP between VLANs is not enabled on the PE, the PE does not perform any operations.
c. After Host A learns the MAC address of VLANIF 4, Host A sends IP datagrams to the PE based on this MAC address.

The PE receives the IP datagrams and forwards them to Host B.

Usage Scenarios

Table 11-6 Proxy ARP usage scenarios

Proxy ARP Type	Usage Scenario
Proxy ARP within a VLAN	Two hosts that need to communicate belong to the same VLAN in which user isolation is configured and the same network segment.
Proxy ARP between VLANs	Two hosts that need to communicate belong to the same network segment but different VLANs. **NOTE** In VLAN aggregation scenarios, proxy ARP between VLANs can be enabled on the VLANIF interface corresponding to the super-VLAN to implement communication between sub-VLANs.

 Copyright © Huawei Technologies Co., Ltd.

Benefits

- Proxy ARP enables a host on a network to mistakenly consider that the destination host is on the same network segment. In this manner, the details of the physical network are unavailable, and the division of the network into subnets is transparent to hosts.
- All processing related to proxy ARP is performed on a gateway, with no configuration needed on the hosts connected to it. In addition, proxy ARP affects only the ARP tables on hosts and does not affect the ARP table and routing table on a gateway.
- Proxy ARP can be used when no default gateway is configured for a host or a host cannot route messages.

11.4.3 Configuring ARP Proxy for Interworking

ARP proxy allows isolated users in a VLAN or different sub VLANs to communicate with each other. In this way, all PCs in a network segment or all voice service users connected to the AGs in the same network segment can communicate with each other. ARP proxy specially applies to triple play and enterprise users.

Context

Two users in the same network segment communicate with each other at Layer 2. To ensure user-side information security, all users connected to an ONU are isolated at Layer 2. Access services, such as the Internet access and voice services, are isolated in the same network segment. This rule also applies to FTTx networks. Specifically, the access services on the same ONU are isolated. In this situation, the users, especially enterprise users, in a LAN cannot share resources or call each other.

ARP proxy is implemented as follows: A host sends an ARP request to another one locating in the same network segment but not in the same physical network. Then, the ARP proxy-enabled device connected to the two hosts replies to this request. This allows the hosts in separate physical networks to communicate with each other.

By default, ARP proxy is disabled globally.

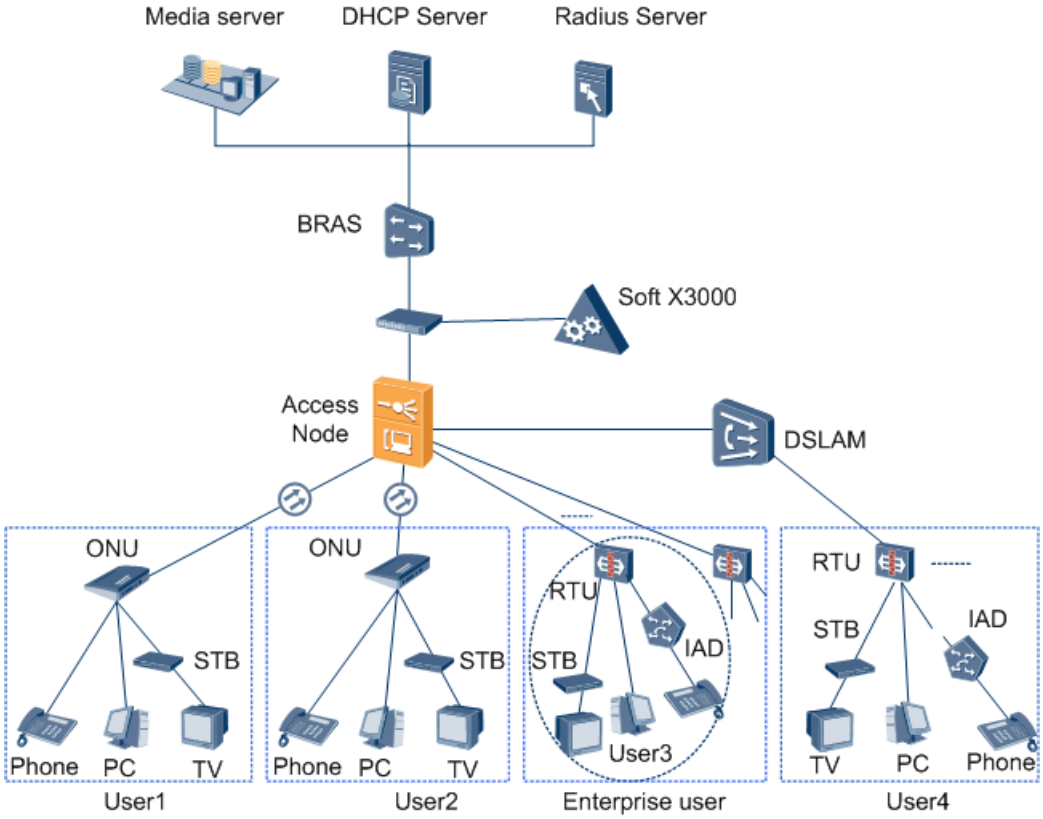
Application Context

As shown in **Figure 11-21**, ARP proxy applies to the following scenarios:

- Triple play users, users 1 and 2 in **Figure 11-21**, in the same network segment are required to communicate with each other.
- Enterprise users in the same network segment or users of different branches (enterprise branches in **Figure 11-21**) in the same enterprise are required to communicate with each other.
- Cascaded users, users 3 and 4 in **Figure 11-21**, in the same network segment are required to communicate with each other.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-21 Application scenarios of ARP proxy interworking



Prerequisite

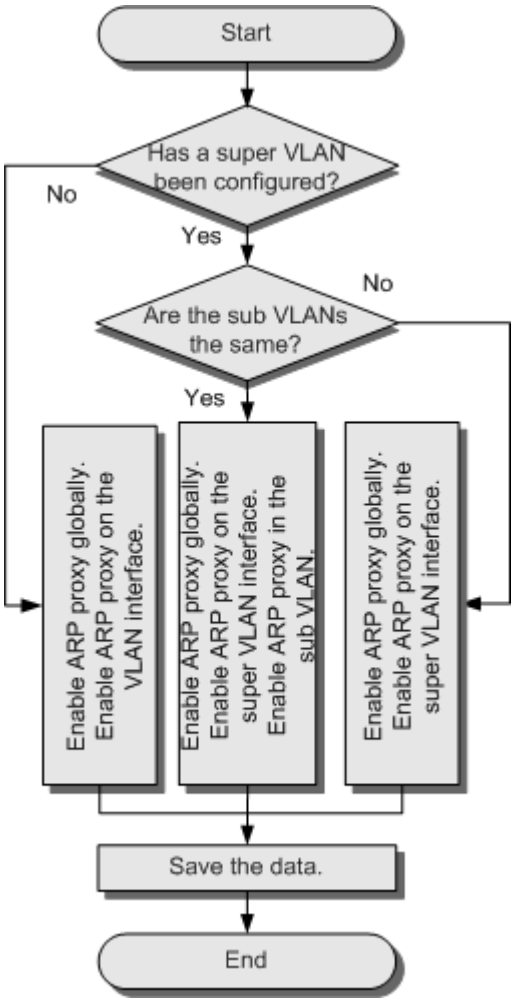
- All users to communicate with each other locate in the same VLAN or super VLAN.
- Services have been configured on the users, and the users can access the Internet and use the voice service properly.
- The IP addresses of PCs or VoIP phone sets of triple play users are in the same network segment.
- The IP addresses of PCs or IADs of enterprise users are in the same network segment.

Configuration Flow

The configuration flow varies depending on VLAN planning, as shown in **Figure 11-22**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-22 ARP proxy configuration flow



Data Plan

Table 11-7 lists the data plan for implementing the ARP proxy interworking shown in Figure 11-21.

Table 11-7 Data plan for configuring ARP proxy

Item	Data
User 1	Super VLAN: 100 Sub VLAN: smart VLAN 10
User 2	Super VLAN: 100 Sub VLAN: smart VLAN 10
Enterprise branches	Branch 1 (Smart VLAN: 50) Branch 2 (Smart VLAN: 50)
User 3	Super VLAN: 100 Sub VLAN: smart VLAN 30

 Copyright © Huawei Technologies Co., Ltd.

Item	Data
User 4	Super VLAN: 100 Sub VLAN: smart VLAN 40

Procedure

- Enable users 1 and 2 to communicate with each other.

  Both users 1 and 2 belong to super VLAN 100 and sub VLAN 10. When packet transmission in the same VLAN is isolated at Layer 2, the transmission is also isolated at Layer 3. To enable users 1 and 2 to communicate with each other, enable ARP proxy at global, super VLAN, and sub VLAN levels.

  NOTE

  - Before ARP proxy takes effect, user services must have been configured, and the users can access the Internet and use the VoIP service properly. This section does not describe service configurations. For details, see the sections for configuring the Internet access service, the VoIP service, and the triple play service.
  - The IP address of the Layer 3 interface of the super VLAN must be in the same network segment as the IP addresses of user terminals. To set the IP address of the Layer 3 interface, run the **ip address** command in the VLAN interface mode. After user terminals request for IP addresses, the DHCP server allocates the IP addresses in the same network segment to them.

  a. Configure the super VLAN and sub VLAN.

  ```
  huawei(config)#vlan 100 super
  huawei(config)#vlan 10 smart
  huawei(config)#supervlan 100 subvlan 10
  ```

  b. Enable ARP proxy globally.

  ```
  huawei(config)#arp proxy enable
  ```

  c. Enable ARP proxy on the super VLAN interface.

  ```
  huawei(config)#interface vlanif 100
  huawei(config-if-vlanif100)#arp proxy enable
  ```

  d. Enable ARP proxy in the subVLAN.

  ```
  huawei(config-if-vlanif100)#arp proxy enable subvlan 10
  huawei(config-if-vlanif100)#quit
  ```

  e. Save the data.

  ```
  huawei(config)#save
  ```

- Enable branch users of an enterprise to communicate with each other.

  No super VLAN is configured for enterprise branches, but they belong to smart VLAN 30. Therefore, the packet transmission in the same VLAN for branch users is isolated at Layer 2. To enable branch users of an enterprise to communicate with each other, enable ARP proxy at global and VLAN levels.

  a. Configure the VLAN.

  ```
  huawei(config)#vlan 30 smart
  ```

  b. Enable ARP proxy globally.

  ```
  huawei(config)#arp proxy enable
  ```

  c. Enable ARP proxy on the VLAN interface.

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config)#interface vlanif 30
huawei(config-if-vlanif30)#arp proxy enable
huawei(config-if-vlanif30)#quit
```

d. Save the data.

```
huawei(config)#save
```

- Enable cascaded users 3 and 4 to communicate with each other.

  Cascaded users 3 and 4 belong to super VLAN 100 but different sub VLANs, sub VLANs 30 and 40, respectively. To enable cascaded users 3 and 4 to communicate with each other, enable ARP proxy globally and on the super VLAN interface.

  a. Configure the super VLAN and sub VLAN for user 3 on the access node and add cascaded port 0/3/0 to sub VLAN 40.

  NOTE

  After a cascaded port is added to a sub VLAN, this port is automatically added to the super VLAN.

  ```
  huawei(config)#vlan 100 super
  huawei(config)#vlan 30,40 smart
  huawei(config)#supervlan 100 subvlan 30
  huawei(config)#supervlan 100 subvlan 40
  huawei(config)#port vlan 40 0/3 0
  ```

  b. Enable ARP proxy globally and on the super VLAN interface on the access node.

  ```
  huawei(config)#arp proxy enable
  huawei(config)#interface vlanif 100
  huawei(config-if-vlanif100)#arp proxy enable
  ```

  c. Configure the super VLAN and sub VLAN for user 4 on the cascaded device. Then, add upstream port 0/19/0 to the super VLAN and sub VLAN, respectively.

  ```
  Cascade_dev(config)#vlan 100 super
  Cascade_dev(config)#vlan 40 smart
  Cascade_dev(config)#supervlan 100 subvlan 40
  Cascade_dev(config)#port vlan 100,40 0/19 0
  ```

----End

Result

Enabled ARP proxy supports the following functions based on service requirements:

- Triple play users, users 1 and 2 in **Figure 11-21**, in the same network segment can communicate with each other.
- Enterprise users in the same network segment or users of different branches (enterprise branches in **Figure 11-21**) in the same enterprise can communicate with each other.
- Cascaded users, users 3 and 4 in **Figure 11-21**, in the same network segment can communicate with each other.

11.4.4 ARP Proxy Reference Standards and Protocols

The following lists the reference documents of ARP proxy:

- IETF RFC1027: Using ARP to Implement Transparent Subnet Gateways

 Copyright © Huawei Technologies Co., Ltd.

11.5 DHCP Server

11.5.1 Feature Change History

Product Version	Change Description
V100R022C00	First version to support this feature.

11.5.2 Introduction to the DHCP Server

Why Is the DHCP Server Required?

A DHCP server is used to assign IP addresses in the following scenarios:

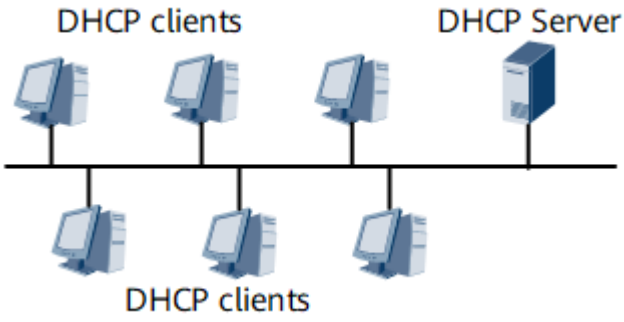
- On a large network, manual configurations take more time and efforts and bring difficulties to the centralized management over the entire network.
- The number of hosts on a network is greater than that of the available IP addresses. Thus, not every host can be assigned a fixed IP address. For example, if Internet service providers (ISPs) limit the number of concurrent network access users, many hosts must dynamically obtain IP addresses.
- Only a few hosts on the network require fixed IP addresses.

What Is a DHCP Server?

A DHCP server assigns IP addresses to clients. A client sends a message to a server to request configuration parameters, such as the IP address, subnet mask, and default gateway address. The server responds with a message carrying the requested configuration parameters. Both the request and reply messages are encapsulated in UDP packets.

Networking Scenario

As shown in Figure, a typical DHCP network consists of a DHCP server and multiple DHCP clients, such as PCs and portable computers. DHCP uses the client/server model. DHCP clients apply for configuration information from the DHCP server. The DHCP server provides configuration information including IP addresses to the DHCP clients.



 Copyright © Huawei Technologies Co., Ltd.

11.5.3 Basic Principles

A DHCP server assigns IP addresses to clients. A DHCP client sends a message to a DHCP server to request configuration parameters, such as the IP address, subnet mask, and default gateway address. The DHCP server responds with a message carrying the requested configuration parameters. Both the request and reply messages are encapsulated in UDP packets.

Modes for Interaction Between the DHCP Client and Server

To obtain a valid dynamic IP address, a DHCP client exchanges different information with a server at different stages. Generally, the DHCP client and server interact in the following modes (defined in standard protocols):

- A DHCP client accesses a network for the first time.

  When a DHCP client accesses a network for the first time, the DHCP client goes through the following stages to set up a connection with a DHCP server:

  - Discovery stage: The DHCP client searches for a DHCP server. The DHCP client broadcasts a DHCPDISCOVER message and only DHCP servers respond to the message.
  - Offer stage: Each DHCP server offers an IP address to the DHCP client. After receiving the DHCPDISCOVER message from the DHCP client, each DHCP server selects an unassigned IP address from the IP address pool, and sends a DHCPOFFER message with the leased IP address and other configurations to the DHCP client.
  - Request stage: The DHCP client selects an IP address. If multiple DHCP servers send DHCPOFFER messages to the DHCP client, the DHCP client accepts the first DHCPOFFER message it receives, and broadcasts to each DHCP server a DHCPREQUEST message carrying information about the selected IP address.
  - Acknowledgment stage: The DHCP server acknowledges the IP address that is offered. After receiving the DHCPREQUEST message, the DHCP server sends a DHCPACK message to the client. The DHCPACK message contains the offered IP address and other settings. The DHCP client then binds its TCP/IP protocol suite to the network interface card.

  Except the IP address offered by the DHCP server selected by the DHCP client, the unassigned IP addresses offered by other DHCP servers are available for other clients.

- A DHCP client extends the IP address lease.

  An IP address dynamically assigned to a DHCP client usually has a validity period. The DHCP server withdraws the IP address after the validity period expires. To continue using the IP address, the DHCP client must renew the IP address lease.

  In actual application, a DHCP client automatically sends a DHCPREQUEST message to the DHCP server to renew the IP address lease when the DHCP client is started or half the duration of the lease is remaining. If the IP address is valid, the DHCP server replies with a DHCPACK message to inform the DHCP client of the new lease.

 Copyright © Huawei Technologies Co., Ltd.

IP Address Assignment Sequence

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

- IP address that is in the database of the DHCP server and is statically bound to the MAC address of the client
- IP address that is first found when the DHCP server searches the DHCP address pool for available IP addresses
- If the DHCP address pool has no available IP address, the DHCP server searches the expired IP addresses and conflicting IP addresses, and then assigns a valid IP address to the client.

Method of Preventing Repeated IP Address Assignment

To avoid address conflicts, the DHCP server pings the IP address before assigning it to a client.

The ping command checks whether a response to the ping packet is received within the specified period. If no response to the ping packet is received, the DHCP server continues to send ping packets to the IP address until the number of sent ping packets reaches the maximum limit. If there is still no response, this IP address is not in use, and the DHCP server assigns the IP address to a client. (This method is implemented based on standard protocols.)

IP Address Reservation

DHCP supports IP address reservation for clients. The reserved IP addresses must belong to the address pool. If an address in the address pool is reserved, it is no longer assignable. Addresses are usually reserved for specific clients, such as DNS and WWW servers.

11.5.4 Configuring the DHCP Server

Context

A DHCP server that uses an address pool is configured to dynamically assign IP addresses to computers that are indirectly connected to the DHCP server.

To configure the DHCP server, you need to configure the IP address of the interface, address segment, and address lease time.

Procedure

1. In the global config mode, run the **interface vlanif** command to enter the VLANIF mode.
2. In the VLANIF mode, run the **ip address** command to configure the IP address and subnet mask of the VLANIF interface.
3. In the VLANIF mode, run the **dhcp select interface** command to enable the DHCP server function.
4. In VLANIF mode, run the **dhcp server ip-range** command to configure the address pool range of the DHCP server.
5. In VLANIF mode, run the **dhcp server mask** command to configure the subnet mask of the IP address assigned by the DHCP server to a DHCP client.

 Copyright © Huawei Technologies Co., Ltd.

If the **dhcp server mask** command is used to configure a client mask, the maximum address pool range is determined by the configured mask length. Otherwise, the larger value between the 18-bit mask length and the mask length of the Layer 3 interface is used as the maximum address pool range.

6. (Optional) To prevent conflicts between the fixed IP address of the device and IP addresses in the address pool on the DHCP server, run the **dhcp server excluded-ip-address** command to configure the range of IP addresses that cannot be automatically assigned to clients from the address pool on the DHCP server. Ensure that the IP addresses that cannot be automatically assigned are not statically bound to the MAC addresses of DHCP clients using the **dhcp server static-bind** command.
7. In the VLANIF mode, run the **dhcp server lease** command to configure the lease time of the IP addresses in the address pool on the DHCP server.
8. In global config mode, run the **security anti-ipattack vlan** *vlanid* **disable** command to disable the VLAN-based anti-IP attack function.

   NOTE

   For V100R021C10, run the **security anti-ipattack disable** command to disable anti-IP attack. For V100R022C00 and later versions, run the **security anti-ipattack vlan** *vlanid* **disable** command to disable VLAN-based anti-IP attack.
9. (Optional) In VLANIF interface mode, run the **dhcp server dns-list** command to configure the IP address of the domain name system (DNS) server in the address pool of the DHCP server.
10. (Optional) In VLANIF interface mode, run the **dhcp server gateway-list** command to configure the gateway IP address assigned by the DHCP server to a DHCP client.
11. (Optional) In VLANIF interface mode, run the **dhcp server option** command to configure the user-defined options of the address pool of the DHCP server.
12. (Optional) To use a fixed IP address to ensure device stability, run the **dhcp server static-bind** command to statically bind an IP address in the address pool of the DHCP server to the MAC address of a DHCP client.

    NOTE

    If a DHCP user is online and does not age or a static user is bound by running the **dhcp server static-bind** command, you can disable the DHCP server function only after the user ages and the bound static user is deleted by running the **undo dhcp server static-bind** command. You can run the **display dhcp server interface vlanif** command to query the configurations of the DHCP server on a specified VLANIF interface.

Example

To enable the DHCP server function on VLANIF interface 2, set the address pool range of the DHCP server to 10.10.100.10 to 10.10.100.20, set the subnet mask of the IP addresses to 24, and set the IP address lease time to 2 minutes, do as follows:

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.10.100.1 255.255.255.0
huawei(config-if-vlanif2)#dhcp select interface
huawei(config-if-vlanif2)#dhcp server ip-range 10.10.100.10 10.10.100.20
huawei(config-if-vlanif2)#dhcp server mask 24
huawei(config-if-vlanif2)#dhcp server lease day 0 hour 0 minute 2
```

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config-if-vlanif2)quit
huawei(config)security anti-ipattack vlan 2 disable
```

11.6 DHCPv6 Server

11.6.1 DHCPv6 Server Feature Changes

Product Version	Change Description
V100R022C10	This feature is supported for the first time.

11.6.2 Introduction to DHCPv6

Why Is a DHCPv6 Server Required?

A DHCPv6 server is used to assign IPv6 addresses in the following scenarios:

- On a large network, manual configurations take more time and efforts and bring difficulties to centralized management of the entire network.
- The number of hosts on a network is greater than that of supported IPv6 addresses. Thus, not every host can be assigned a fixed IPv6 address. For example, if Internet service providers (ISPs) limit the number of concurrent network access users, many hosts must dynamically obtain IPv6 addresses.
- Only a few hosts on the network require fixed IPv6 addresses.

What Is a DHCPv6 Server?

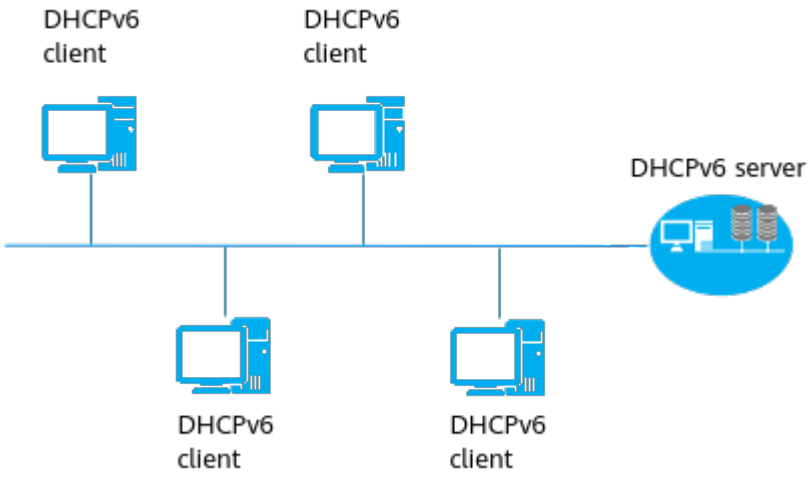
A DHCPv6 server processes address assignment, lease renewal, and address release requests from clients or relay agents, and assigns IPv6 addresses/prefixes and other network configuration parameters to clients.

Networking Scenario

As shown in **Figure 11-23**, the basic DHCPv6 architecture generally contains a DHCPv6 server and multiple DHCPv6 clients (such as PCs and laptops). DHCPv6 uses the client/server model. Clients apply for configuration information from the server. The server provides configuration information including IPv6 addresses to the clients.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-23 DHCPv6 server networking



Basic Concepts

1. Multicast address

   In DHCP, a client sends broadcast packets to locate the server. To prevent broadcast storms, IPv6 uses multicast packets instead of broadcast packets. DHCPv6 uses the following two multicast addresses:

   - FF02::1:2: multicast address of all DHCP relay agents and servers. The address is a link-local multicast address and is used for communication between a DHCPv6 client and its neighboring servers or between a DHCPv6 client and DHCPv6 relay agents. All DHCPv6 servers and relay agents are members of the multicast group corresponding to this multicast address.
   - FF05::1:3: multicast address of all DHCP servers. The address is a site-local address and is used for communication between DHCPv6 relay agents and servers within a site. All DHCPv6 servers within a site are members of this multicast group.

2. UDP port number: DHCPv6 packets are carried over UDPv6.
3. DHCP unique identifier
   - A DHCP unique identifier (DUID) identifies a DHCP device. Each server or client has a unique DUID. Servers use DUIDs to identify clients and clients use DUIDs to identify servers.
   - The DUIDs of a client and server are carried in the Client Identifier and Server Identifier options in DHCPv6 packets, respectively. The two options have the same format and are distinguished by the option-code field value. When the value of the option-code field is 1, the option is the Client Identifier option. When the value of the option-code field is 2, the option is the Server Identifier option.

11.6.3 DHCPv6 Server Principles

A DHCPv6 server assigns addresses and prefixes to clients in the following two modes:

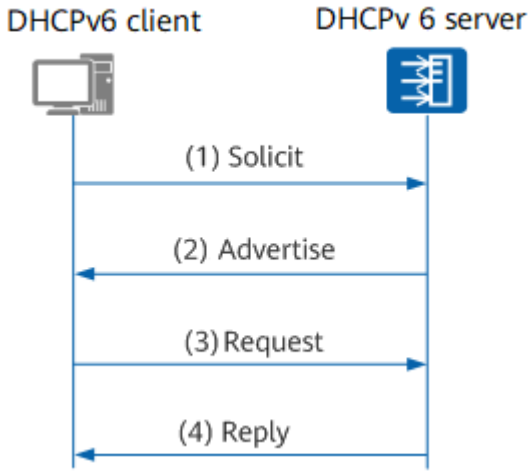
- DHCPv6 assignment with four-step interactions applies to the scenario when multiple DHCPv6 servers are deployed on a network.

 Copyright © Huawei Technologies Co., Ltd.

- DHCPv6 rapid assignment with two-step interactions applies to the scenario when only one DHCPv6 server is deployed on a network.

The process of DHCPv6 address assignment using four-step interactions is as follows:

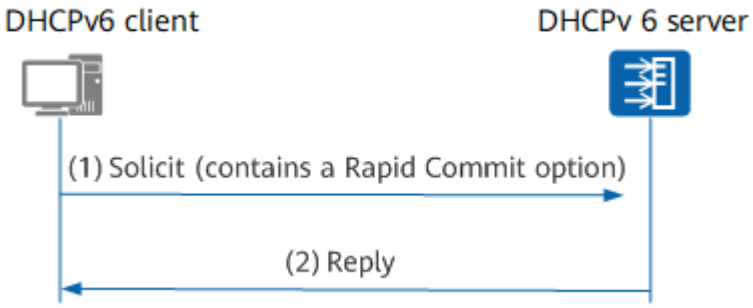
1. A DHCPv6 client sends a Solicit packet to request a DHCPv6 server to assign an IPv6 address and network configuration parameters.
2. If the Solicit packet does not carry the Rapid Commit option, or the Solicit packet carries the Rapid Commit option but the DHCPv6 server does not support the rapid assignment mode, the DHCPv6 server replies with the Advertise packet to notify the client of the address, prefix, and other network configuration parameters that can be assigned to the client.
3. If the DHCPv6 client receives Advertise packets from multiple DHCPv6 servers, the DHCPv6 client selects the DHCPv6 server with the highest priority based on the Preference option in the Advertise packets and sends a Request multicast packet to all DHCPv6 servers. The Request multicast packet carries the DUID of the selected DHCPv6 server.
4. The DHCPv6 server responds with a Reply packet, which contains the assigned address and network configuration parameters.



The process of DHCPv6 address assignment using two-step interactions is as follows:

1. A DHCPv6 client sends a Solicit packet that carries the Rapid Commit option, indicating that the client expects the server to rapidly assign an address and network configuration parameters.
2. After receiving the Solicit packet, the DHCPv6 server performs the following operations:
   - If the DHCPv6 server supports two-step rapid address assignment, it directly returns a Reply packet to assign an IPv6 address and other network configuration parameters to the client. The Reply packet also carries the Rapid Commit option.
   - If the DHCPv6 server does not support rapid address assignment, it assigns an IPv6 address/prefix and other network configuration parameters to the client through four-step interactions.

 Copyright © Huawei Technologies Co., Ltd.



11.6.4 Configuring the DHCPv6 Server

Context

If a computer on the network cannot be directly connected to a server that assigns IPv6 addresses, the computer needs to pass through other network devices. In this case, to enable the computer to dynamically obtain an IPv6 address, you need to configure a DHCPv6 server with an address pool.

To configure the DHCPv6 server, you need to configure the IPv6 address of the interface, address segment, and address lease time.

Procedure

1. In global config mode, run the **interface vlanif** command to enter VLANIF mode.
2. In VLANIF mode, run the **ipv6 enable** command to enable the IPv6 function on the interface.
3. In VLANIF mode, run the **ipv6 address** command to configure the IPv6 address of the VLAN interface.
4. In VLANIF mode, run the **dhcpv6 select interface** command to enable the DHCPv6 server function.
5. In VLANIF mode, run the **dhcpv6 server** [**allow-hint** | **preference** *preference-value* | **rapid-commit** | **unicast**] command to configure the DHCPv6 server attributes.
6. In VLANIF mode, run the **dhcpv6 server address prefix** *ipv6-prefix/ipv6-prefix-length* [**life-time** { *valid-lifetime* | **infinite** } { *preferred-lifetime* | **infinite** }] command to configure the network prefix and lifetime.
7. In VLANIF mode, run the **dhcpv6 server renew-time-percent rebind-time-percent** command to configure the renewal time and rebinding time of the IPv6 address pool.
8. In VLANIF mode, run the **dhcpv6 server excluded-address** *start-ipv6-address* [**to** *end-ipv6-address*] command to configure the range of the IPv6 addresses that cannot be automatically assigned to clients from the IPv6 address pool.

Example

In VLANIF interface 2, to set the IPv6 address 2001:db8:1::1/64 for the interface, enable DHCPv6 server address assignment, set the network prefix of the DHCPv6

 Copyright © Huawei Technologies Co., Ltd.

server to 2001:db8::/64, set the lifetime to infinite, set the percentages of the lease renewal time and rebinding time of the DHCPv6 server to the preferred lifetime of the DHCPv6 server's network prefix to 60% and 80% respectively, and exclude IPv6 addresses in the range of 2001:db8::1 to 2001:db8::10 from automatic assignment from the DHCPv6 server address pool, do as follows:

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1::1 64
huawei(config-if-vlanif2)#dhcpv6 select interface
huawei(config-if-vlanif2)#dhcpv6 server allow-hint
huawei(config-if-vlanif2)#dhcpv6 server address prefix 2001:db8::/64 life-time infinite infinite
huawei(config-if-vlanif2)#dhcpv6 server renew-time-percent 60 rebind-time-percent 80
huawei(config-if-vlanif2)#dhcpv6 server excluded-address 2001:db8::1 to 2001:db8::10
```

11.7 DHCP Relay

This section describes the implementation and configuration of Dynamic Host Configuration Protocol (DHCP) relay.

11.7.1 DHCP Relay Overview

What is DHCP?

Dynamic Host Configuration Protocol (DHCP) assigns IP addresses in a dynamic and centralized manner.

DHCP works in client-server (C/S) mode, including the following two processes:

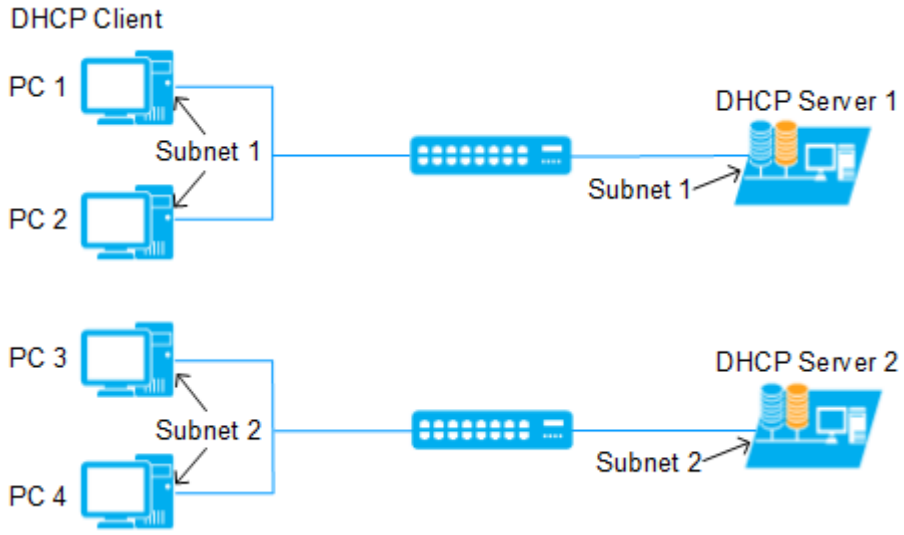
- A DHCP client sends a Request message to the DHCP server to request configuration parameters.
- The DHCP server sends a Reply message that contains information such as IP address.

Why Is DHCP Relay Needed?

As shown in **Figure 11-24**, without DHCP relay, DHCP works only when the DHCP clients and server are in the same subnet. When DHCP clients locate in different subnets, you need to deploy a DHCP server in each subnet. This is an inefficient way of using DHCP servers.

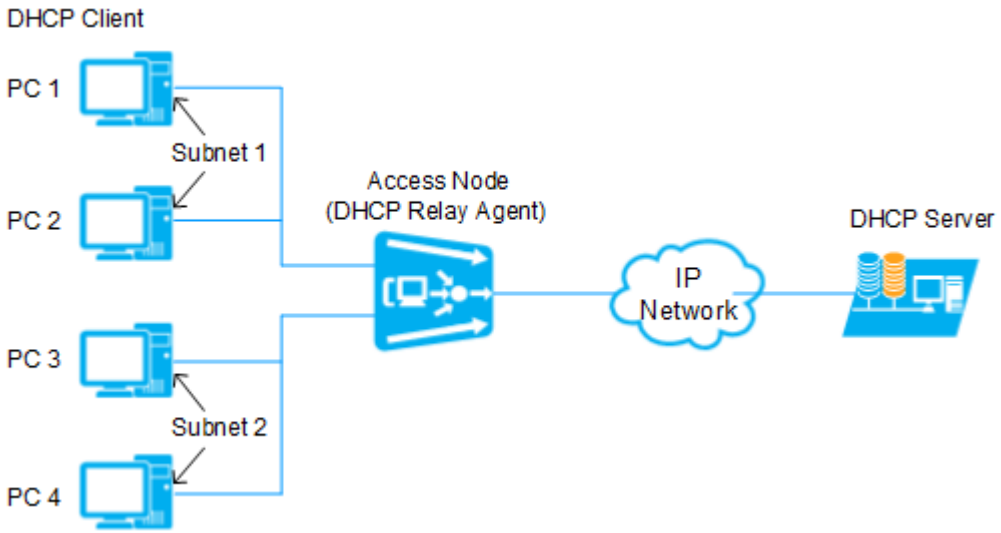
 Copyright © Huawei Technologies Co., Ltd.

Figure 11-24 Without DHCP relay



As shown in **Figure 11-25**, DHCP relay solves this problem. DHCP relay can forward DHCP packets between the DHCP server and DHCP clients across subnets. This reduces hardware costs and achieves centralized client management.

Figure 11-25 DHCP relay



- A DHCP client sends a Request message to the DHCP server to request configuration parameters.
- The DHCP server sends the client a Reply message that contains information such as IP address.

DHCP Relay Networking

As shown in **Figure 11-25**, a DHCP relay network consists of the following roles:

- DHCP client, which exchanges packets with the DHCP server to obtain an IP address.

 Copyright © Huawei Technologies Co., Ltd.

- DHCP relay agent, which forwards packets between DHCP clients and the DHCP server if they are in different subnets.
- DHCP server: assigns IP address and other network parameters to DHCP clients.

DHCP Relay Versions

DHCP relay has two versions, DHCPv4 relay and DHCPv6 relay, which apply to IPv4 and IPv6 networks, respectively.

The content in this section applies to both DHCPv4 relay and DHCPv6 relay.

11.7.2 DHCPv4 Relay Principles

DHCPv4 Layer 2 Relay Principle

If the access device between a DHCP client and the DHCP server is only a bridge device on which the routing function is unavailable or disabled, the DHCPv4 Layer 2 relay function can transparently transmit DHCP packets.

DHCPv4 Layer 3 Relay Principle

The DHCPv4 Layer 3 relay function can forward DHCP packets across subnets so that DHCP clients can obtain dynamic IP addresses from the DHCP server on another subnet. Figure 1 shows the process of the DHCPv4 L3 relay.

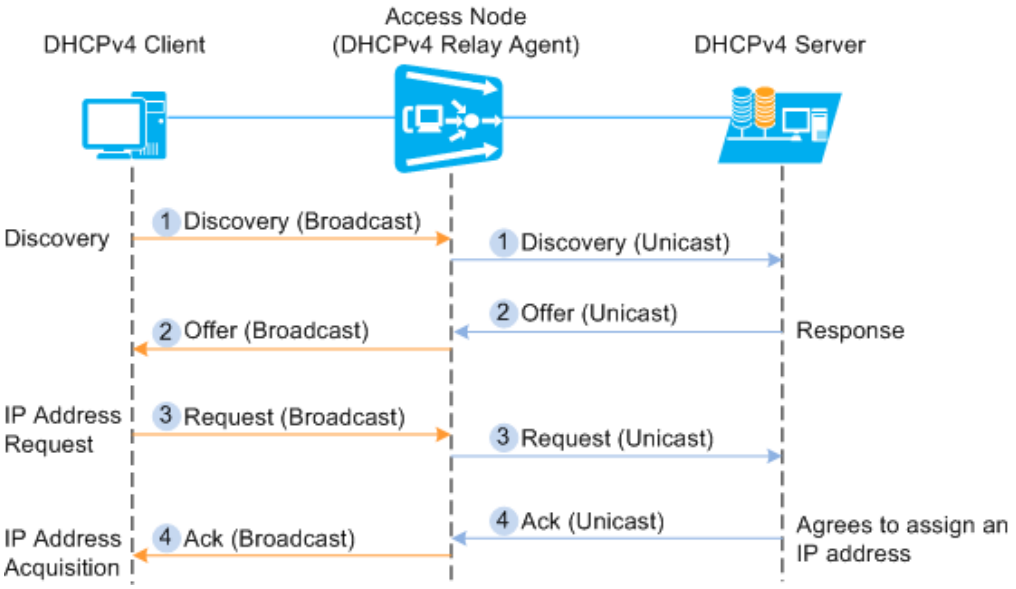
1. When DHCPv4 Relay is enabled on a DHCPv4 client, the client broadcasts a discovery message on the local network to search for the DHCPv4 server.
   - If a DHCPv4 server is available on the local network, the client can obtain an IP address from the server.
   - If no DHCPv4 server exists on the local network, the access device connected to the local network receives the Discovery message and forwards it to the DHCPv4 server on another network.
2. If the DHCPv4 server determines that the DHCPv4 client can apply for an IP address, the server returns a unicast Offer message to the access device. Upon receiving the offer message, the access device broadcasts the Offer message to the DHCPv4 clients.
3. The DHCPv4 client broadcasts a Request message to request an IP address. Upon receiving the Request message, the access device sends the Offer message to the DHCPv4 server.
4. The DHCPv4 server agrees to assign an IP address and sends an ACK message to the access device. Upon receiving the ACK message, the access device broadcasts the ACK message to configure the dynamic IP address of the DHCPv4 client.

 Copyright © Huawei Technologies Co., Ltd.

NOTE

After DHCPv4 relay is enabled, the access device modifies the XID (a field in DHCP packet header, which is equivalent to the sequence number of a DHCP packet) of DHCP packets sent by users. As a result, the XIDs in the DHCP packets sent by DHCPv4 clients are different from those in the DHCP packet received by the DHCPv4 server. Generally, the DHCPv4 server does not verify the XID. Therefore, services are not affected. If an operator adds information to the XID field of the packets sent by DHCPv4 clients for verification on the DHCPv4 server, the verification may fail because the added information is not defined in the protocol. In this case, services may be interrupted.

Figure 11-26 DHCPv4 Layer 3 relay working process



Concerning the DHCPv4 Layer 3 relay principle, you need to understand the following two questions.

1. How does a DHCPv4 relay agent convert broadcast packets into unicast packets?

   The DHCPv4 relay agent knows the IP address of the DHCPv4 server. Therefore, it changes the broadcast IP addresses in Discovery and Request messages received from the clients to the destination IP address of the DHCPv4 server.

2. How does a DHCPv4 server know which address pool is used to allocate IP addresses to clients?

   Upon receiving the Discovery messages from the clients, the DHCPv4 relay agent adds the gateway address, for example, 10.1.1.1, to the messages. After receiving the Discovery messages, the DHCPv4 server determines that the requests are sent from network 10.1.1.0/24, and then assigns IP addresses from the range of 10.1.1.2–10.1.1.254.

DHCPv4 Relay and DHCP Option 82

The DHCPv4 relay function has no authentication or security mechanisms. In actual applications, there are many security problems, such as excessive broadcast packets and IP exhaustion attacks. The DHCPv4 relay agent inserts Option 82 into

 Copyright © Huawei Technologies Co., Ltd.

DHCPv4 packets. The DHCPv4 server can determine the validity of the DHCPv4 clients based on the Option 82 information, ensuring user access security. When the DHCP Option 82 function is enabled, the DHCPv4 relay agent adds or removes Option 82 in the received DHCPv4 packets regardless of DHCPv4 L2 Relay or DHCPv4 L3 Relay. For details, see **DHCP Option 82**.

11.7.3 DHCPv6 Relay Principle

Principle of DHCPv6 L2 Relay

By default, the system checks the IPv6 addresses of all DHCPv6 packets according to the RFC6221 standard. If an IPv6 address meets the requirement, the corresponding packet is relayed. If an IPv6 address does not meet the requirement, the corresponding packet is directly forwarded and is not relayed. The check criteria are as follows:

- DHCPv6 request packet: The destination address is All_DHCP_Relay_Agents_and_Servers (ff02::1:2).
- DHCPv6 response packet: The source and destination addresses are both the link-local address (LLA).

The DHCPv4 Layer 2 relay working process is as follows:

1. When a DHCPv6 relay agent receives a message that needs to be processed using L2 relay, the relay agent constructs a new Relay-forward message, regardless of the original message type. The relay agent copies the IP address of the original message to the peer-address field of the new Relay-forward message, and copies the entire original DHCP message (excluding the IP header and UDP header) to the Relay Message Option (Option 9) of the new Relay-forward message.
2. If the original message originates from a DHCPv6 client, the DHCPv6 relay agent initializes the hop-count field of the Relay-forward message to 0. If the original message is a Relay-forward message that originates from another DHCPv6 relay agent, the relay agent first checks the hop-count field of the message. If the value of the field is greater than or equal to 32, the relay agent discards the message; if the value is smaller than 32, the relay agent adds 1 to the value and uses the new value as the value of the hop-count field of the new Relay-forward message.
3. The DHCPv6 relay agent invariably sets the link-address Field of the Relay-forward message to an unspecified address (::), and invariably includes the interface-id field (Option 18) and remote-id field (Option 37) in the Relay-forward message.
4. In the downstream direction, the Relay Message Option (Option 9) of the original Relay-reply message is extracted, and the content of the Relay Message Option (Option 9) is forwarded as a new downstream message to the DHCPv6 client.

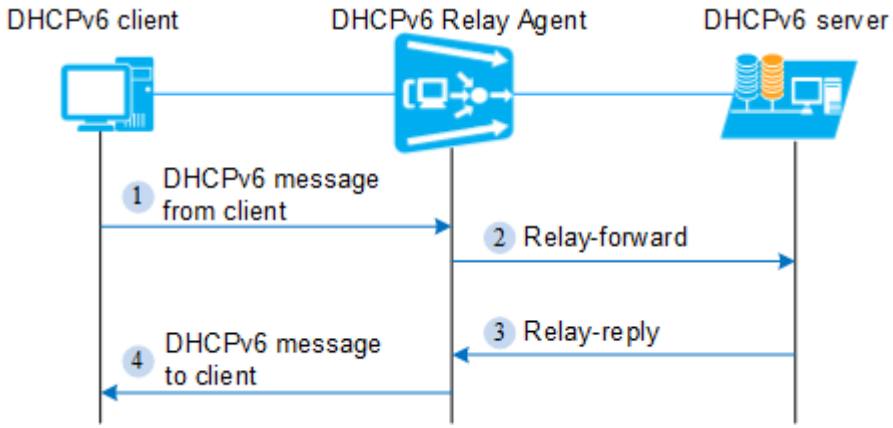
Principle of DHCPv6 L3 Relay

The processing flows between the DHCPv6 client and server are the same when the IPv6 address, IPv6 prefix and other network configuration parameters are dynamically obtained using or not using the DHCPv6 relay agent. **Figure 11-27** illustrates only the forwarding process when DHCPv6 relay is used.

 Copyright © Huawei Technologies Co., Ltd.

The MA5800 supports DHCPv6 option 53. DHCPv6 option 53 is Relay-ID option which carries the DUID information of DHCPv6 L3 relay agent.

Figure 11-27 Working process with DHCPv6 relay



1. The DHCPv6 client sends a request message to the multicast address FF02::1:2 of all DHCPv6 servers and relay agents.
2. After receiving the request message, the DHCPv6 relay agent encapsulates the message into the Relay Message Option of the Relay-forward message and sends the Relay-forward message to the DHCPv6 server.
3. The DHCPv6 server parses the Relay-forward message and obtains the client's request message. Then, the server selects an IPv6 address and other parameters for the client, constructs a reply message, encapsulates the reply message into the Relay Message Option of the Relay-reply message, and sends the Relay-reply message to the DHCPv6 relay agent.
4. The DHCPv6 relay agent parses the Relay-reply message, obtains the server's reply message, and then forwards the message to the DHCPv6 client.
5. The DHCPv6 client implements network configuration according to the IPv6 address, IPv6 prefix, and other parameters contained in the reply message.

NOTE

A DHCPv6 server allocates both IPv6 addresses and IPv6 prefixes to DHCPv6 clients. After obtaining an IPv6 prefix allocated by the DHCPv6 server, a DHCPv6 client sends a Router Advertisement (RA) message containing the IPv6 prefix to the network in which it is located. Using the prefix, hosts in the network can automatically configure their IPv6 addresses.

11.7.4 Configuring DHCPv4 Relay

The MA5800 can work in DHCPv4 Layer 2 relay or DHCPv4 Layer 3 relay mode to forward packets between DHCPv4 clients and the server across networks.

- By default, the MA5800 works in the Layer 2 DHCP relay mode. In this mode, the MA5800 transparently transmits the DHCPv4 packets from users.
- When the MA5800 works in Layer 3 mode, DHCPv4 server groups need to be configured. A DHCPv4 server group contains one or more DHCPv4 servers. A DHCPv4 server can be added to multiple DHCPv4 server groups at the same time.

 Copyright © Huawei Technologies Co., Ltd.

DHCPv4 Layer 3 relay mode can be configured on the MA5800 to select a DHCPv4 server group. The following DHCPv4 relay modes are supported:

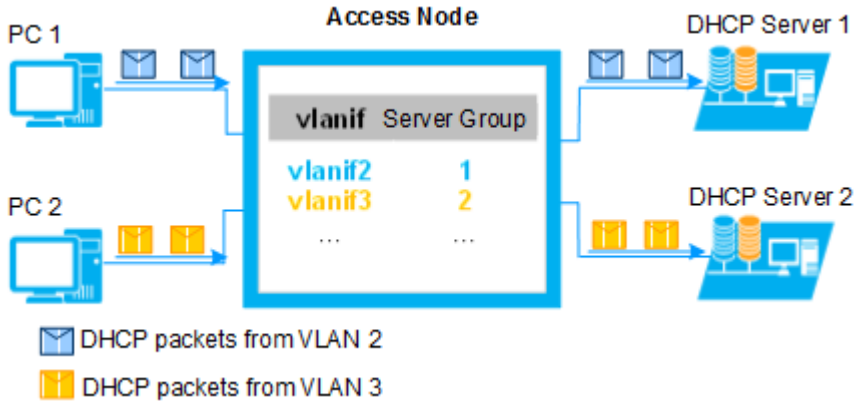
Standard Mode

Figure 11-28 shows the standard mode, which is the default mode and is applicable when you need to specify a DHCPv4 server group for users in different VLANs. In the preceding figure, the two servers belong to two server groups. In actual deployment, the server groups can be mapped to one or multiple servers. The DHCPv4 relay agent selects a DHCPv4 server group based on the addresses of the primary Layer 3 VLAN interfaces, which are the gateway addresses of DHCPv4 relay agents. This mode is the most common and simplest DHCPv4 relay mode. The disadvantage of this mode is that different services in the same VLAN cannot be distinguished.

To use this mode, you must bind Layer 3 VLAN interfaces to DHCP server groups. The configuration procedure is as follows:

Set the working mode of DHCPv4 relay to standard, configure DHCPv4 server groups, and bind VLANs to DHCPv4 server groups.

Figure 11-28 Standard mode of DHCPv4 relay



Option 60 Mode

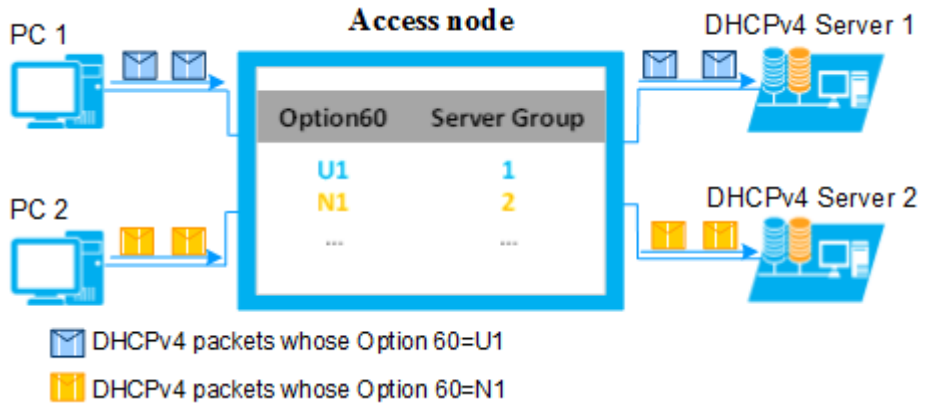
As shown in Figure 2 DHCPv4 Relay Option 60 Mode, this mode applies when you need to specify DHCPv4 server groups for users in different domains indicated by Option 60 strings. The DHCPv4 relay agent selects a DHCPv4 server group based on Option 60 strings (domain names) in the received DHCPv4 packets. This mode is also commonly used to distinguish services in the same VLAN through the Option 60 strings.

To use this mode, you must configure domain names through the Option 60 strings and bind the domain names to DHCPv4 server groups. The configuration procedure is as follows:

Set the DHCPv4 relay working mode to Option 60, configure DHCPv4 server groups, create domains using the Option 60 strings, and bind the Option 60 strings to DHCPv4 server groups.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-29 DHCPv4 relay Option 60 mode



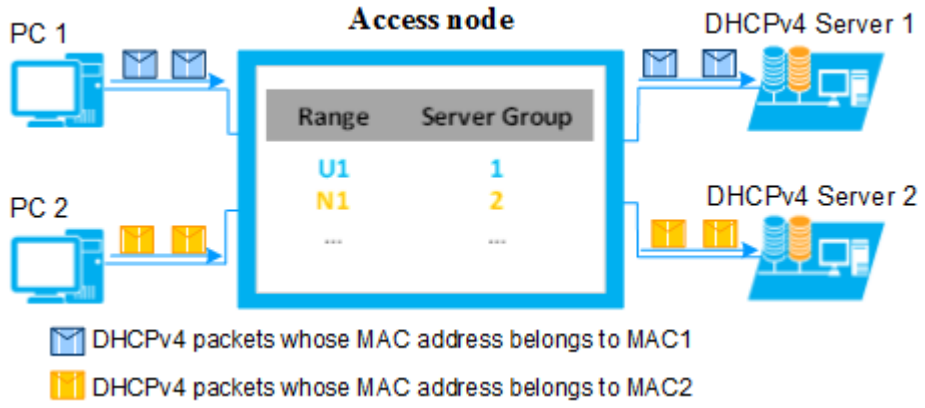
MAC Address Range Mode

As shown in **Figure 11-30**, this mode applies when you need to specify DHCPv4 server groups for users in different MAC address ranges. The DHCPv4 relay agent selects a DHCPv4 server group based on the source MAC address range of the received DHCPv4 packets. In this mode, services in the same VLAN are distinguished based on the MAC address range.

To use this mode, you must configure MAC address ranges and bind them to DHCPv4 server groups. The configuration procedure is as follows:

Set the DHCPv4 relay working mode to mac-range, configure DHCPv4 server groups, define MAC address ranges, and bind the MAC address ranges to DHCPv4 server groups.

Figure 11-30 DHCPv4 relay MAC address range mode



11.7.4.1 Configuring the Standard Mode

The standard mode is used to specify a DHCP server group for users of a VLAN that contains service ports created on the MA5800.

Prerequisites

The user VLAN and network VLAN have been configured. For details about how to configure a VLAN, see **Configuring a VLAN**.

 Copyright © Huawei Technologies Co., Ltd.

Procedure

Step 1 Configure the DHCP forwarding mode.

The DHCP forwarding mode can be configured in global config mode or VLAN service profile mode.

- In global config mode, run the **dhcp mode layer-3 standard** command to configure the DHCP Layer 3 relay mode to **standard**. If you have selected the **vlan** keyword and specified a VLAN ID, the standard DHCP Layer 3 relay mode takes effect only in the specified VLAN.
- In VLAN service profile mode, perform the following operations to configure the DHCP forwarding mode in the user VLAN:
  a. Run the **vlan service-profile** command to create a VLAN service profile.
  b. Run the **dhcp mode layer-3 standard** command to configure the DHCP Layer 3 relay mode to **standard**.
  c. Run the **commit** command to make the profile configuration take effect.
  d. Run the **quit** command to quit the VLAN service profile mode.
  e. Run the **vlan bind service-profile** command to bind the VLAN service profile created in **Step 1.a** to the VLAN.

Step 2 Configure a DHCP server group.

1. In global config mode, run the **dhcp-server** command to create a DHCP server group.

   igroup-number and *ip-addr* are keywords in the **dhcp-server** command.

   – *igroup-number*: number that uniquely identifies a DHCP server group.

     Before specifying a number for a DHCP server group, check whether the number is unique among the DHCP server group numbers that have been configured on the MA5800. To do so, run the **display dhcp-server all-group** command on the MA5800.

   – *ip-addr*: IP address of a DHCP server in a DHCP server group. A maximum of four IP addresses can be specified for this parameter.

     NOTE

     Ensure that the IP address configured in this parameter is the same as the IP address of the network-side DHCP server.

2. (Optional) Run the **dhcp server mode** command to configure the working mode of the DHCP servers in the DHCP server group.

   The DHCP servers in a DHCP server group can work in load sharing or active/standby mode. By default, the DHCP servers work in load sharing mode.

Step 3 Bind the VLAN to the DHCP server.

1. In global config mode, run the **interface vlanif** command to create a VLAN Layer 3 interface.

   Ensure that the value of the **VLANID** keyword in the **interface vlanif** command is the same as the ID of the VLAN that has been created.

2. In VLAN interface (VLANIF) mode, run the **ip address** command to configure the IP address of the VLAN Layer 3 interface.

 Copyright © Huawei Technologies Co., Ltd.

After the configuration, all IP packets in the VLAN use this IP address as the source IP address for Layer 3 forwarding.

NOTE

- If all the devices between the MA5800 and the DHCP server work at Layer 2, the IP address of the VLAN Layer 3 interface must be in the same network segment as the IP address of the DHCP server.
- If a device between the MA5800 and the DHCP server works at Layer 3, the IP address of the VLAN Layer 3 interface can be in a network segment different from that of the DHCP server. In this case, ensure that the DHCP server is reachable from the VLAN Layer 3 interface.

3. Versions earlier than V100R020C00, in VLANIF mode, run the **dhcp-server** command to bind the VLAN to the DHCP server.

   In V100R020C00 and later versions, in VLANIF-UAM mode, run the **dhcp-server** command to bind the VLAN to the DHCP server.

   NOTE

   In V100R020C00 and later versions, the DHCP relay service port is disabled by default. Run the **sysman service dhcp-relay enable** command to enable the DHCP relay service.

   This command requires the *group-number*. Ensure that the *group-number* value is the same as the number of the DHCP server group.

----End

Example

The following is an example of the configurations used to enable DHCP relay on the MA5800 so that DHCP clients in VLAN 2 can obtain IP addresses from DHCP server group 1:

- The DHCP clients in VLAN 2 are in a network segment different from DHCP server group 1.
- The IP address of the Layer 3 interface in VLAN 2 is 10.1.1.101/24.
- DHCP server group 1 contains two DHCP servers working in active/standby mode.
  - The IP address of the active server is 10.1.1.9.
  - The IP address of the standby server is 10.1.1.10.
- The maximum response time of the DHCP servers is 20s.
- The maximum number of DHCP server response timeout times is 10.

```
Versions earlier than V100R020C00
huawei(config)#dhcp mode layer-3 standard
huawei(config)#dhcp server mode backup 20 10
huawei(config)#dhcp-server 1 ip 10.1.1.9 10.1.1.10
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.1.101 24
huawei(config-if-vlanif2)#dhcp-server 1
In V100R020C00 and later versions
huawei(config)#dhcp mode layer-3 standard
huawei(config)#dhcp server mode backup 20 10
huawei(config)#dhcp-server 1 ip 10.1.1.9 10.1.1.10
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.1.101 24
```

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config-if-vlanif2)#quit
huawei(config)sysman service dhcp-relay enable
huawei(config)#interface vlanif-uam 2
huawei(config-if-vlanif-uam-2)#dhcp-server 1
```

11.7.4.2 Configuring the DHCP Option 60 Mode

The DHCP Option 60 mode is used to specify a DHCP server group for DHCP Option 60 users.

Prerequisites

- The user VLAN and network VLAN have been configured. For details about how to configure a VLAN, see **Configuring a VLAN**.
- Before configuring the DHCP Option 60 mode, ensure that the Option 60 domain name is available.

Context

If the MA5800 provisions multiple services, such as the multicast video and VoIP services, the services may be provided by different service providers. These service providers use different DHCP servers or different relay IP addresses of the same DHCP server to allocate IP addresses to DHCP users. Therefore, the DHCP Option 60 mode needs to be configured for the users in different DHCP Option 60 domains to apply for IP addresses.

Procedure

Step 1 Configure the DHCP forwarding mode.

The DHCP forwarding mode can be configured in global config mode or VLAN service profile mode.

- In global config mode, run the **dhcp mode layer-3 option60** command to configure the DHCP Layer 3 relay mode to **option60**. If you have selected the **vlan** keyword and specified a VLAN ID, the Option 60 DHCP Layer 3 relay mode takes effect only in the specified VLAN.
- In VLAN service profile mode, perform the following operations to configure the DHCP forwarding mode in the user VLAN:
  a. Run the **vlan service-profile** command to create a VLAN service profile.
  b. Run the **dhcp mode layer-3 option60** command to configure the DHCP Layer 3 relay mode to **option60**.
  c. Run the **commit** command to make the profile configuration take effect.
  d. Run the **quit** command to quit the VLAN service profile mode.
  e. Run the **vlan bind service-profile** command to bind the VLAN service profile created in **Step 1.a** to the VLAN.

Step 2 Configure a DHCP server group.

1. In global config mode, run the **dhcp-server** command to create a DHCP server group.

   igroup-number and *ip-addr* are keywords in the **dhcp-server** command.

 Copyright © Huawei Technologies Co., Ltd.

- *igroup-number*: number that uniquely identifies a DHCP server group.

  Before specifying a number for a DHCP server group, check whether the number is unique among the DHCP server group numbers that have been configured on the MA5800. To do so, run the **display dhcp-server all-group** command on the MA5800.
- *ip-addr*: IP address of a DHCP server in a DHCP server group. A maximum of four IP addresses can be specified for this parameter.

  NOTE

  Ensure that the IP address configured in this parameter is the same as the IP address of the network-side DHCP server.

2. (Optional) Run the **dhcp server mode** command to configure the working mode of the DHCP servers in the DHCP server group.

   The DHCP servers in a DHCP server group can work in load sharing or active/standby mode. By default, the DHCP servers work in load sharing mode.

Step 3 In global config mode, run the **dhcp domain** command to create a DHCP Option 60 domain.

Configure the domain name based on the type of the terminal connected to the MA5800. For example, if the terminal connected to the MA5800 is a DHCP client running the Windows 98, Windows 2000, Windows XP, or Windows NT OS, the domain name must be **msft**.

Step 4 In DHCP Option 60 mode, run the **dhcp-server** command to bind the DHCP Option 60 domain to the DHCP server group.

After the configuration, the DHCP server group serves all DHCP clients in the DHCP Option 60 domain.

NOTE

In V100R020C00 and later versions, the DHCP relay service port is disabled by default. Run the **sysman service dhcp-relay enable** command to enable the DHCP relay service.

Step 5 Configure the gateway address of the DHCP Option 60 domain.

1. In global config mode, run the **interface vlanif** command to create a VLAN Layer 3 interface.

   Ensure that the value of the **VLANID** keyword in the **interface vlanif** command is the same as the ID of the VLAN that has been created.
2. In VLANIF mode, run the **ip address** command to configure the IP address of the VLAN Layer 3 interface.

   After the configuration, all IP packets in the VLAN use this IP address as the source IP address for Layer 3 forwarding.

   NOTE

   - If all the devices between the MA5800 and the DHCP server work at Layer 2, the IP address of the VLAN Layer 3 interface must be in the same network segment as the IP address of the DHCP server.
   - If a device between the MA5800 and the DHCP server works at Layer 3, the IP address of the VLAN Layer 3 interface can be in a network segment different from that of the DHCP server. In this case, ensure that the DHCP server is reachable from the VLAN Layer 3 interface.

 Copyright © Huawei Technologies Co., Ltd.

3. Versions earlier than V100R020C00, in VLANIF mode, run the **dhcp domain gateway** command to configure the gateway address of the DHCP Option 60 domain.

   In V100R020C00 and later versions, in VLANIF-UAM mode, run the **dhcp domain gateway** command to configure the gateway address of the DHCP Option 60 domain.

   Ensure that this gateway address is one of the IP addresses of the VLAN Layer 3 interface. Different gateways can be configured for different DHCP Option 60 domains under the same VLAN Layer 3 interface. Therefore, the DHCP Option 60 users connected to different DHCP servers can be differentiated based on their DHCP Option 60 domain name.

----End

Example

The following is an example of the configurations used to enable DHCP relay on the MA5800 so that DHCP clients in VLAN 2 can obtain IP addresses from DHCP server group 1.

- The DHCP clients run the Windows 98, Windows 2000, Windows XP, or Windows NT OS.
- The DHCP clients in VLAN 2 are in a network segment different from DHCP server group 1.
- The IP address of the Layer 3 interface in VLAN 2 is 10.1.2.1/24.
- DHCP server group 1 contains two DHCP servers working in load sharing mode.
  - The IP address of the active server is 10.10.10.10.
  - The IP address of the standby server is 10.10.10.11.

```
Versions earlier than V100R020C00
huawei(config)#dhcp mode layer-3 Option60
huawei(config)#dhcp-server 1 ip 10.10.10.10 10.10.10.11
huawei(config)#dhcp domain msft
huawei(config-dhcp-domain-msft)#dhcp-server 1
huawei(config-dhcp-domain-msft)#quit
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.2.1 24
huawei(config-if-vlanif2)#dhcp domain msft gateway 10.1.2.1
In V100R020C00 and later versions
huawei(config)#dhcp mode layer-3 Option60
huawei(config)#dhcp-server 1 ip 10.10.10.10 10.10.10.11
huawei(config)#dhcp domain msft
huawei(config-dhcp-domain-msft)#dhcp-server 1
huawei(config-dhcp-domain-msft)#quit
huawei(config)sysman service dhcp-relay enable
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.2.1 24
huawei(config-if-vlanif2)#quit
huawei(config)#interface vlanif-uam 2
huawei(config-if-vlanif-uam-2)#dhcp domain msft gateway 10.1.2.1
```

11.7.4.3 Configuring the DHCP MAC Address Segment Mode

The DHCP MAC address segment mode is used to specify a DHCP server group for the users in a MAC address segment.

 Copyright © Huawei Technologies Co., Ltd.

Prerequisites

The user VLAN and network VLAN have been configured. For details about how to configure a VLAN, see **Configuring a VLAN**.

Context

The devices on a network may be from different vendors. Each vendor has a consistent MAC address segment and all devices from a vendor use the vendor's MAC address segment. Configure the DHCP MAC address segment mode for the devices to obtain IP addresses from the DHCP server.

The MA5800 selects a DHCP server group based on a MAC address segment. After a MAC address segment is configured, all DHCP clients in the MAC address segment can obtain IP addresses from the DHCP server group.

Procedure

Step 1 Configure the DHCP forwarding mode.

The DHCP forwarding mode can be configured in global config mode or VLAN service profile mode.

- In global config mode, run the **dhcp mode layer–3 mac-range** command to configure the DHCP Layer 3 relay mode to **mac-range**. If you have selected the **vlan** keyword and specified a VLAN ID, the DHCP MAC address segment mode takes effect only in the specified VLAN.
- In VLAN service profile mode, perform the following operations to configure the DHCP forwarding mode in the user VLAN:
  a. Run the **vlan service-profile** command to create a VLAN service profile.
  b. Run the **dhcp mode layer–3 mac-range** command to configure the DHCP Layer 3 relay mode to **mac-range**.
  c. Run the **commit** command to make the profile configuration take effect.
  d. Run the **quit** command to quit the VLAN service profile mode.
  e. Run the **vlan bind service-profile** command to bind the VLAN service profile created in **Step 1.a** to the VLAN.

Step 2 Configure a DHCP server group.

1. In global config mode, run the **dhcp-server** command to create a DHCP server group.

   igroup-number and *ip-addr* are keywords in the **dhcp-server** command.

   – *igroup-number*: number that uniquely identifies a DHCP server group.

     Before specifying a number for a DHCP server group, check whether the number is unique among the DHCP server group numbers that have been configured on the MA5800. To do so, run the **display dhcp-server all-group** command on the MA5800.

   – *ip-addr*: IP address of a DHCP server in a DHCP server group. A maximum of four IP addresses can be specified for this parameter.

 Copyright © Huawei Technologies Co., Ltd.

NOTE

Ensure that the IP address configured in this parameter is the same as the IP address of the network-side DHCP server.

2. (Optional) Run the **dhcp server mode** command to configure the working mode of the DHCP servers in the DHCP server group.

   The DHCP servers in a DHCP server group can work in load sharing or active/standby mode. By default, the DHCP servers work in load sharing mode.

Step 3 Specify a MAC address segment.

1. In global config mode, run the **dhcp mac-range** command to create a MAC address segment.

   range-name is the MAC address segment name, which is used only for commenting on the MAC address segment.

2. In MAC address segment mode, run the **mac-range mac-address-start to mac-address-end** command to specify the range of the MAC address segment.

Step 4 In MAC address segment mode, run the **dhcp-server** command to bind the MAC address segment to the DHCP server.

NOTE

In V100R020C00 and later versions, the DHCP relay service port is disabled by default. Run the **sysman service dhcp-relay enable** command to enable the DHCP relay service.

Step 5 Configure the gateway address of the MAC address segment.

1. In global config mode, run the **interface vlanif** command to create a VLAN Layer 3 interface.

   Ensure that the value of the **VLANID** keyword in the **interface vlanif** command is the same as the ID of the VLAN that has been created.

2. In VLANIF mode, run the **ip address** command to configure the IP address of the VLAN Layer 3 interface.

   After the configuration, all IP packets in the VLAN use this IP address as the source IP address for Layer 3 forwarding.

   NOTE

   - If all the devices between the MA5800 and the DHCP server work at Layer 2, the IP address of the VLAN Layer 3 interface must be in the same network segment as the IP address of the DHCP server.
   - If a device between the MA5800 and the DHCP server works at Layer 3, the IP address of the VLAN Layer 3 interface can be in a network segment different from that of the DHCP server. In this case, ensure that the DHCP server is reachable from the VLAN Layer 3 interface.

3. Versions earlier than V100R020C00, in VLANIF mode, run the **dhcp mac-range gateway** command to configure the gateway address of the MAC address segment.

   In V100R020C00 and later versions, in VLANIF-UAM mode, run the **dhcp mac-range gateway** command to configure the gateway address of the MAC address segment.

 Copyright © Huawei Technologies Co., Ltd.

Ensure that this gateway address is one of the IP addresses of the VLAN Layer 3 interface. Different gateways can be configured for different MAC address segments under the same VLAN Layer 3 interface. Therefore, the users connected to different DHCP servers can be differentiated based on their DHCP MAC address segment.

----End

Example

The following is an example of the configurations used to enable DHCP relay on the MA5800 so that users in the MAC address segment ranging from 00e0-fc00-0001 to 00e0-fc00-0100 in VLAN 2 can obtain IP addresses from DHCP server group 1.

- The DHCP clients in VLAN 2 are in a network segment different from DHCP server group 1.
- The IP address of the Layer 3 interface in VLAN 2 is 10.1.2.1/24.
- DHCP server group 1 contains two DHCP servers working in load sharing mode.
  - The IP address of the active server is 10.10.10.10.
  - The IP address of the standby server is 10.10.10.11.

```
Versions earlier than V100R020C00
huawei(config)#dhcp mode layer-3 mac-range
huawei(config)#dhcp-server 1 ip 10.10.10.10 10.10.10.11
huawei(config)#dhcp mac-range huawei
huawei(config-mac-range-huawei)#mac-range 00e0-fc00-0001 to 00e0-fc00-0100
huawei(config-mac-range-huawei)#dhcp-server 1
huawei(config-mac-range-huawei)#quit
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.2.1 24
huawei(config-if-vlanif2)#dhcp mac-range huawei gateway 10.1.2.1
In V100R020C00 and later versions
huawei(config)#dhcp mode layer-3 mac-range
huawei(config)#dhcp-server 1 ip 10.10.10.10 10.10.10.11
huawei(config)#dhcp mac-range huawei
huawei(config-mac-range-huawei)#mac-range 00e0-fc00-0001 to 00e0-fc00-0100
huawei(config-mac-range-huawei)#dhcp-server 1
huawei(config-mac-range-huawei)#quit
huawei(config)sysman service dhcp-relay enable
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.2.1 24
huawei(config-if-vlanif2)#quit
huawei(config)#interface vlanif-uam 2
huawei(config-if-vlanif-uam-2)#dhcp mac-range huawei gateway 10.1.2.1
```

11.7.5 Differences Between DHCPv4 and DHCPv6 Configurations

Dynamic Host Control Protocol version 6 (DHCPv6) is a DHCP protocol for IPv6. This topic describes the differences between DHCPv4 and DHCPv6 regarding function specifications and commands. Before configuring DHCPv6 services, it is recommended that you familiarize yourself with the procedures and principles of configuring DHCPv4 services and be aware of the differences between DHCPv4 and DHCPv6 configurations.

 Copyright © Huawei Technologies Co., Ltd.

Context

The differences between DHCPv4 and DHCPv6 configurations are as follows:

- When working in Layer 3 forwarding mode, DHCPv6 supports both the standard mode and option 16 mode. You can run the **dhcpv6 mode { layer-2 | layer-3 [option16] }** command to configure the DHCPv6 working mode. DHCPv4 supports the standard, media access control (MAC) address segment, and DHCP option 60 modes. You can run the **dhcp mode { layer-2 | layer-3 { mac-range | option60 | standard } }** command to configure the DHCPv4 working mode.
- DHCPv6 does not support DHCP proxy.
- Regarding commands, DHCPv4 uses **dhcp** while DHCPv6 uses **dhcpv6** as the command word. For example, the **dhcp-server** command is used to configure DHCPv4 server groups while the **dhcpv6-server** command is used to configure DHCPv6 server groups.
- In V100R020C00 and later versions, the DHCPv6 relay service port is disabled by default. Run the **sysman service dhcpv6-relay enable** command to enable the DHCPv6 relay service if necessary.

For the differences of other configuration commands, see "DHCPv6 Configuration" in the *Command Reference*. Some commands need to be executed in diagnose mode. For details about these commands, see "Diagnose Mode Command" in the *Command Reference*.

11.7.6 DHCP Relay Standards and Protocols Compliance

The DHCPv4 relay feature complies with the following standard and protocol:

- RFC 2131: Dynamic Host Configuration Protocol
- DHCPv4 Option 82: RFC 3046

The following lists the reference standards and protocols of the DHCPv6 relay feature:

- DHCPv6 relay: RFC 3315 (DHCPv6 protocol created by IETF)
- DHCPv6 Layer 2 relay: draft-ietf-dhc-dhcpv6-ldra-02 (Draft of a DHCPv6 Layer 2 relay protocol created by IETF), RFC6221
- DHCPv6 option 37: RFC 4649

DHCPv6 option 53: RFC 5460

11.8 DHCP Proxy

This section describes the implementation and configuration of Dynamic Host Configuration Protocol (DHCP) proxy.

11.8.1 What Is DHCP Proxy

Definition

DHCP proxy enables the MA5800 to modify the DHCP packets exchanged between a DHCP server and DHCP clients based on site requirements.

 Copyright © Huawei Technologies Co., Ltd.

DHCP proxy is composed of the server ID proxy and lease proxy functions.

- Server ID proxy

  The Option 54 field in a DHCP packet specifies the IP address of a DHCP server. Server ID proxy enables the MA5800 to change the value of the Option 54 field so that the actual IP address of the DHCP server is hidden from DHCP clients. This reduces the risk of DHCP server attacks initiated by DHCP clients.

- Lease proxy

  The IP address lease that a DHCP client applies for is determined based on Options 51, 58, and 59 in a DHCP packet initiated by the DHCP client. Lease proxy enables the MA5800 to change the values of these options so that the MA5800 provides a shorter lease than that provided by the DHCP server for the DHCP client. This facilitates lease management.

Purpose

The DHCP proxy functions offer solutions for different problems.

- Server ID proxy

  The actual IP address of the DHCP server can be hidden from DHCP clients, which reduces the risk of DHCP server attacks initiated by DHCP clients.

- Lease proxy

  The duration of the lease configured by the DHCP server is so long that the MA5800 cannot obtain the DHCP client status in a timely manner. This obstructs service provisioning.

The lease-time proxy, however, enables a DHCP client to obtain a shorter lease time for an IP address. The MA5800 with the DHCP proxy function enabled is capable of quickly perceiving whether a user is online. Meanwhile, the request packets from the DHCP client for re-leasing an IP address during a short lease time are processed by the MA5800 and are no longer forwarded to the DHCP server. This decreases the load of the DHCP server for frequently processing the request packets during the short lease time.

The MA5800 with DHCP proxy enabled changes the XID values of the DHCP packets sent from DHCP clients to values different from the XID values of the DHCP packets received by the DHCP server. The DHCP server generally does not check XID values. Therefore, the XID value change generally does not affect services. However, if carriers add data to the XID field for data checks on the DHCP server, a data check may fail, which would affect services.

NOTE

Exchange identification (XID) is a field in a DHCP packet that uniquely identifies the DHCP packet.

11.8.2 DHCP Proxy Principles

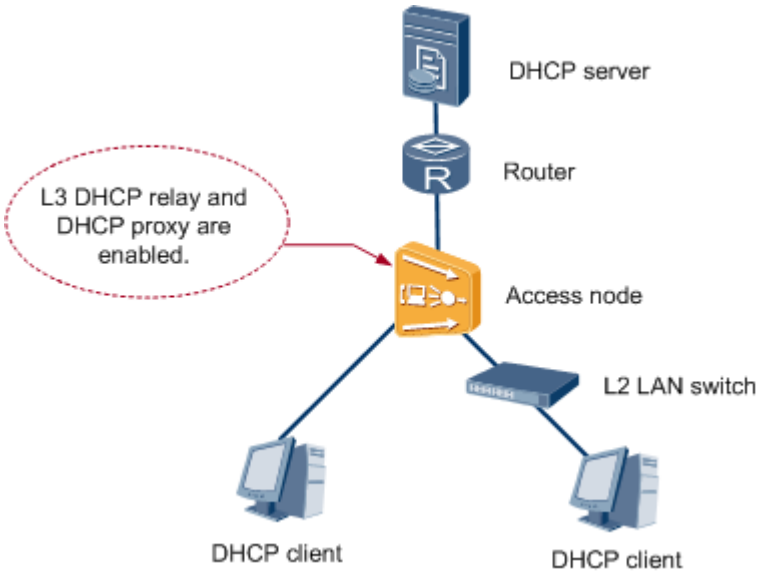
Application Scenario

The MA5800 supports DHCP proxy only when Layer 3 DHCP relay is enabled on it. Both user ports and cascading ports on the MA5800 support DHCP proxy. The MA5800 with DHCP proxy enabled can monitor all DHCP packets exchanged

 Copyright © Huawei Technologies Co., Ltd.

between DHCP clients and a DHCP server. **Figure 11-31** shows an application scenario of DHCP proxy.

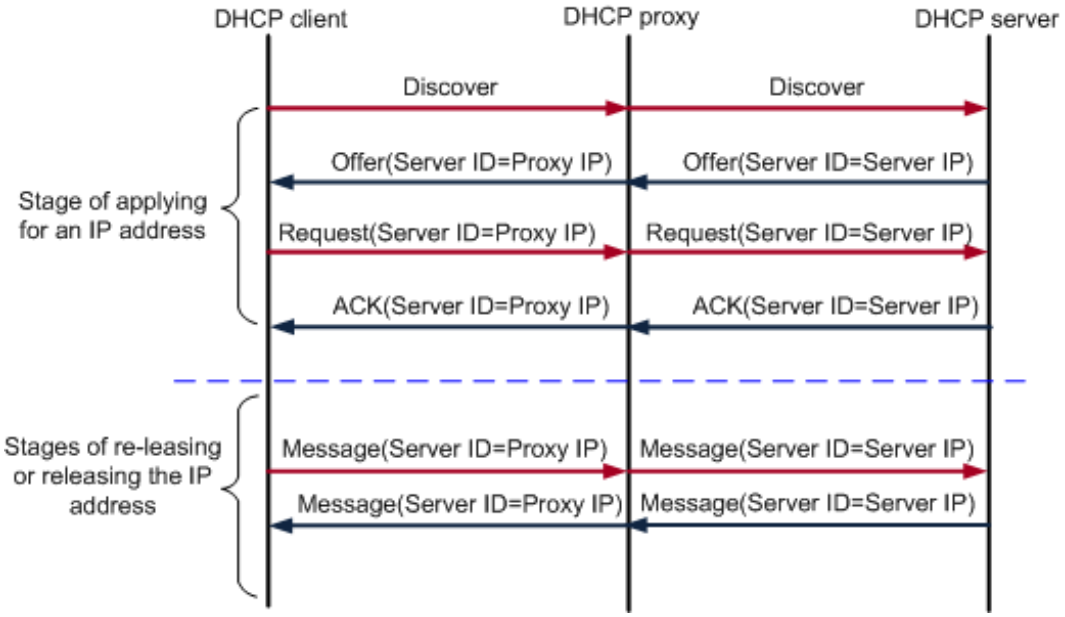
Figure 11-31 Application scenario of DHCP proxy



Server ID Proxy

With DHCP proxy enabled, the server ID proxy function alters the exchange of packets as follows:

Figure 11-32 Exchange of packets between a DHCP client and DHCP server (server ID proxy)



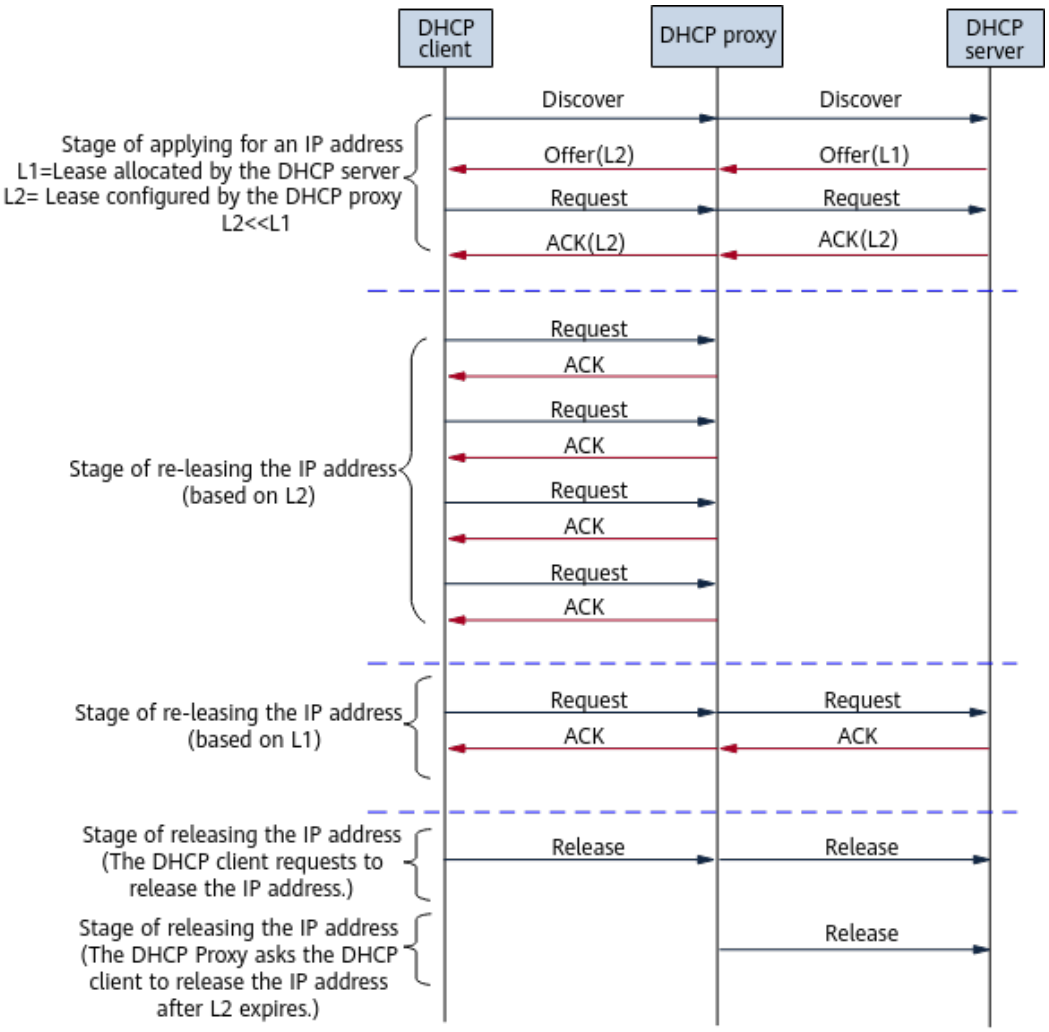
 Copyright © Huawei Technologies Co., Ltd.

- In the downstream direction, the MA5800 changes the value of the Option 54 field in the response packets, such as the Offer and ACK packets, sent by the DHCP server to the IP address of the MA5800. After the DHCP client receives the packets, Option 54 in these packets contains the IP address of the MA5800, and the related field in the DHCP packets is always the IP address of the MA5800 hereafter.
- In the upstream direction, the MA5800 recovers the value of Option 54 in the DHCP packets sent from the DHCP client to the IP address of the actual DHCP server.

Lease Proxy

With DHCP proxy enabled, exchange of packets between a DHCP client and DHCP server (lease proxy) as follows:

Figure 11-33 Exchange of packets between a DHCP client and DHCP server (lease proxy)



Stage of applying for an IP address:

 Copyright © Huawei Technologies Co., Ltd.

1. The DHCP client sends a request packet to the DHCP server for an IP address. The DHCP server then sends a response packet and allocates an IP address with lease L1 to the DHCP client.
2. The MA5800 captures the response packet from the DHCP server, changes the L1 value in the packet to a smaller value L2 (configurable on the MA5800), and sends the Offer packet containing lease L2 to the DHCP client. Then, the lease of the IP address allocated to the DHCP client is changed to L2.

Stage of re-leasing the IP address:

1. The DHCP client sends a request packet to the DHCP server to re-lease the IP address when half of the L2 lease duration has elapsed.
2. The MA5800 captures the request packet and determines whether to send the request packet to the DHCP server based on L1.
   - If the lease time does not reach half of the L1, the MA5800 does not need to send a lease renewal request to the DHCP server, but directly responds to the lease renewal request from the DHCP client, updates the lease renewal time to L2, and allows the client to renew the IP address lease.
   - If the elapsed lease is greater than or equal to half of L1, the MA5800 forwards the request packet to the DHCP server.
3. After receiving the request packet, the DHCP server performs the following operations:
   - Responds to the request packet if it allows the re-leasing. Then, the MA5800 forwards the response packet to the DHCP client.
   - Sends a NACK packet to the DHCP client if it does not allow the re-leasing. Then, the MA5800 notifies the DHCP client that this IP address will be released.

Stage of releasing the IP address:

- If the DHCP client sends a request packet to release the IP address, the MA5800 forwards the request packet to the DHCP server.
- If the MA5800 detects that lease L2 of the DHCP client has expired and the MA5800 has not received any request to re-lease the IP address from the DHCP client, the MA5800 sends a request to release the IP address to the DHCP server.

11.8.3 DHCP Proxy Standards and Protocols Compliance

Reference documents:

- dsl2006[1].127.00: Proposals of DHCP relay improvements

11.9 BFD

11.9.1 Introduction to BFD

The link detection mechanisms of access devices include: slow Hello mechanism, dedicated detection mechanism, and bidirectional forwarding detection (BFD) mechanism.

 Copyright © Huawei Technologies Co., Ltd.

- Slow Hello mechanism: It usually refers to the Hello mechanism offered by a routing protocol. This mechanism can detect a fault in seconds. In high-speed data transmission, for example, at gigabit rates, fault detection longer than 1s will bring the loss of a large amount of data. For delay-sensitive services such as voice services, the delay longer than 1s is also unacceptable.
- Dedicated detection mechanism: Different protocols or manufacturers may provide proprietary detection mechanisms; however, it is difficult to deploy the proprietary detection mechanisms when systems are interconnected.
- BFD detection mechanism: provides uniform detection for all media and protocol layers. It provides light-load and millisecond-level fast fault detection for channels between adjacent forwarding devices. The detected faults may occur on interfaces, data links, or forwarding devices.

In view of the pros and cons of the preceding detection mechanisms, the BFD detection mechanism is widely used in access devices.

11.9.2 Principles to BFD

BFD is used to detect communication link faults between forwarding devices. Specifically, it is used to detect the connectivity of a data protocol on a path between devices. The path can be a physical link, logical link, or tunnel.

- Upper-layer applications provide BFD with parameters, such as the detection address and the detection time.
- BFD creates, deletes, or modifies a BFD session based on these parameters and notifies the upper-layer applications of the session status.

The following describes the basic principles of BFD in terms of the BFD detection mechanism, session establishment mode, and session management.

BFD Detection Mechanism

In the BFD detection mechanism, two systems establish 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 has occurred on the path, selects another path for communication based on the existing routes, and reports a link session fault alarm.

BFD control packets are encapsulated in UDP packets. In the initial phase of a BFD session, both systems negotiate with each other using parameters in BFD control packets, such as discriminators, expected minimum intervals for sending and receiving BFD control packets, and local BFD session status. After the negotiation succeeds, the two systems transmit BFD control packets to each other at the negotiated interval.

BFD provides the following detection modes:

- Asynchronous mode: most commonly-used BFD detection mode. In this mode, two systems periodically send BFD control packets to each other. If a system fails to receive any packets within a specified period, the BFD session is considered Down.
- Demand mode: an alternative BFD detection mode. A system with a large number of BFD sessions can use the demand mode to prevent overheads

 Copyright © Huawei Technologies Co., Ltd.

generated by periodic BFD control packet transmission from affecting system operation. In this mode, once a BFD session is established, both systems stop sending BFD control packets to each other. Instead, the systems use another mechanism, such as the Hello mechanism of a routing protocol, to check link connectivity. In this manner, the systems are less affected by overheads resulting from BFD sessions.

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.

Table 11-8 BFD session establishment modes

BFD Session Establishment Mode	Description
Static mode	BFD session parameters, such as the local and remote discriminators, are manually configured and delivered to establish BFD sessions. **NOTE** In static mode, configure unique local and remote discriminators for each BFD session. This mode prevents incorrect discriminators from affecting BFD sessions that have correct discriminators and prevents BFD session flapping.
Dynamic mode	When a BFD session is dynamically established, the system processes the local and remote discriminators as follows: • Dynamically allocates the local discriminator. When a system triggers the dynamic establishment of a BFD session, the system allocates a dynamic discriminator as the local discriminator of the BFD session. Then, the system sends a BFD control packet with Your Discriminator set to 0 to the peer for session negotiation. • Automatically learns the remote discriminator. The local end of a BFD session sends a BFD control packet with **Your Discriminator** set to **0** to the remote end. After the remote end receives the packet, it checks whether the value of **Your Discriminator** in this packet matches the value of its **My Discriminator**. If they are matched, the remote end learns the value of **My Discriminator** of the local end and obtains its **Your Discriminator**.

 Copyright © Huawei Technologies Co., Ltd.

BFD Session Management

A BFD session has the following states:

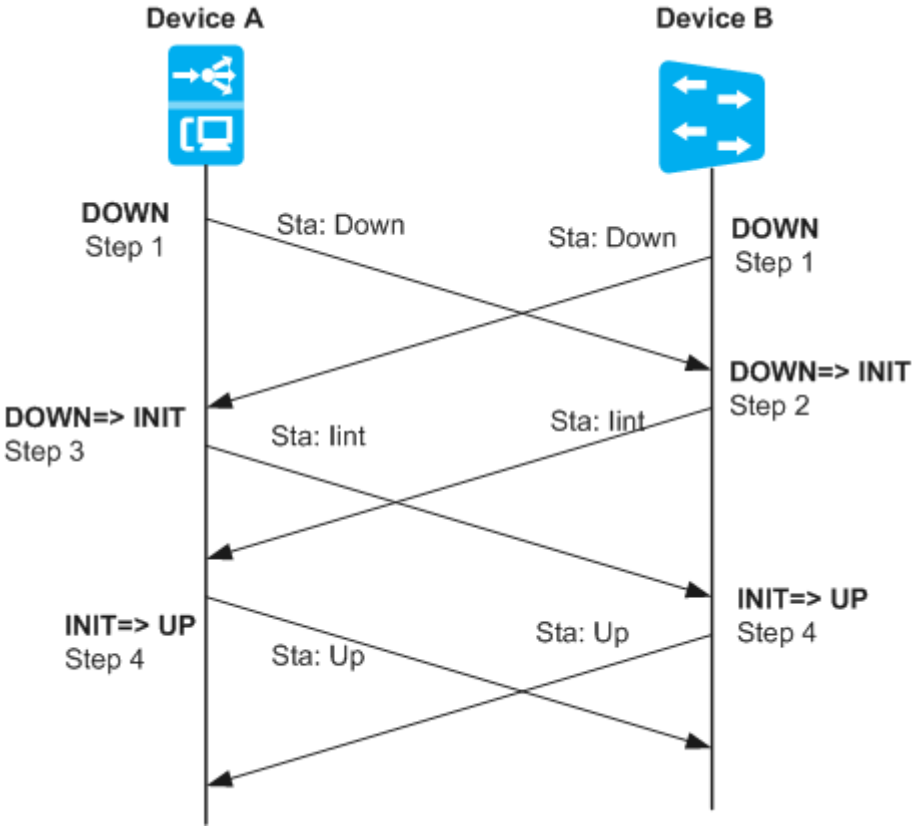
- **Down**: indicates that the BFD session is in the **Down** state or has been set up.
- **Init**: indicates that the local system can communicate with the remote system, and the local system expects a BFD session to go **Up**.
- **Up**: indicates that the BFD session is set up successfully.
- **AdminDown**: indicates that the BFD session is in the **AdminDown** state. After entering the **AdminDown** state, BFD sessions are not deleted but they do not monitor links.

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 both two systems are aware of the status change.

As shown in **Figure 11-34**, a BFD session is established as an example to describe the state machine migration process.

Figure 11-34 BFD session establishment



1. Devices A and B enable their BFD session state machines. Because the initial status of the state machines is **Down**, devices A and B send BFD control packets with the **State** field set to **Down**. For a statically established BFD

 Copyright © Huawei Technologies Co., Ltd.

session, the value of **Your Discriminator** in a BFD control packet is specified manually. For a dynamically established BFD session, the value of **Your Discriminator** is **0**.

2. After receiving a BFD control packet with the **State** field set to **Down**, device B changes the status to **Init** and sends BFD control packets with the **State** field set to **Init**.

   NOTE

   After the local BFD session status changes to **Init**, device B no longer processes the received BFD control packets with the **State** field set to **Down**.

3. The BFD status change of device A is the same as that of device B.
4. After receiving a BFD control packet with the **State** field set to **Init**, device B changes the local BFD session status to **Up**.
5. The BFD status change of device A is the same as that of device B.

11.9.3 BFD Echo

BFD echo is a rapid fault detection mechanism in which the local system sends BFD echo packets and the remote system loops back the packets. BFD echo is classified into passive BFD echo and one-arm BFD echo modes. These two BFD echo modes have the same detection mechanism but different application scenarios.

Passive BFD Echo

The device supports passive BFD echo for interworking with other vendors' devices.

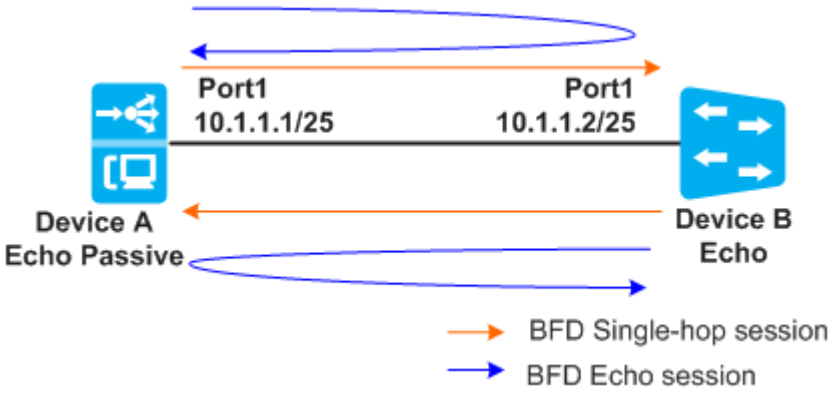
Passive BFD echo applies only to single-hop IP link scenarios and works with asynchronous BFD. When a BFD session works in asynchronous echo mode, the two endpoints of the BFD session perform both slow detection in asynchronous mode and quick detection in echo mode.

As shown in **Figure 11-35**, Device A is directly connected to Device B, and asynchronous BFD sessions are established between the two devices. After active BFD echo is enabled on Device B and passive BFD echo is enabled on Device A, the two devices work in asynchronous echo mode and send single-hop and echo packets to each other.

If Device A has a higher BFD performance than Device B, for example, the minimum intervals between receiving BFD packets supported by Device A and Device B are 3 ms and 100 ms respectively, then BFD sessions in asynchronous mode will adopt the larger interval (100 ms). If BFD echo is enabled, Device A can use echo packets to implement faster link failure detection. If BFD echo is disabled, Device A and Device B can still use asynchronous BFD packets to detect link failures. However, the minimum interval between receiving BFD packets is the larger interval value (100 ms in this example).

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-35 Passive BFD echo networking



The process of establishing a passive BFD echo session as shown in **Figure 11-35** is as follows:

1. Device B functions as a BFD session initiator and sends an asynchronous BFD packet to Device A. The Required Min Echo RX Interval field carried in the packet is a nonzero value, which specifies that Device A must support BFD echo.
2. After receiving the packet, Device A finds that the value of the Required Min Echo RX Interval field carried in the packet is a nonzero value. If Device A has passive BFD echo enabled, it checks whether any ACL that restricts passive BFD echo is referenced. If an ACL is referenced, only BFD sessions that match specific ACL rules can enter the asynchronous echo mode. If no ACL is referenced, BFD sessions immediately enter the asynchronous echo mode.
3. Device B periodically sends BFD echo packets, and Device A sends BFD echo packets (the source and destination IP addresses are the local IP address, and the destination physical address is Device B's physical address) at the interval specified by the Required Min RX Interval field. Both Device A and Device B start a receive timer, with a receive interval that is the same as the interval at which they each send BFD echo packets.
4. After Device A and Device B receive BFD echo packets from each other, they immediately loop back the packets at the forwarding layer. Device A and Device B also send asynchronous BFD packets to each other at an interval that is much less than that for sending echo packets.

One-Arm BFD Echo

One-arm BFD echo applies only to single-hop IP link scenarios. Generally, one-arm BFD echo is used when two devices are directly connected and only one of them supports BFD. Therefore, one-arm BFD echo does not require both ends to negotiate echo capabilities. A one-arm BFD echo session can be established on a device that supports BFD. After receiving a one-arm BFD echo session packet, devices that do not support BFD immediately loop back the packet, implementing quick link failure detection.

The local device that has one-arm BFD echo enabled sends a special BFD packet (both the source and destination IP addresses in the IP header are the local IP address, and the MD and YD in the BFD payload are the same). After receiving the packet, the remote device immediately loops the packet back to the local device

 Copyright © Huawei Technologies Co., Ltd.

to determine link reachability. One-arm BFD echo can be used on low-end devices that do not support BFD.

Similarities and Differences Between Passive BFD Echo and One-Arm BFD Echo

To ensure that passive BFD echo or one-arm BFD echo can take effect, disable strict URPF on devices that send BFD echo packets.

Strict URPF prevents attacks that use spoofed source IP addresses. If strict URPF is enabled on a device, the device obtains the source IP address and inbound interface of a packet and searches the forwarding table for an entry with the destination IP address set to the source IP address of the packet. The device then checks whether the outbound interface for the entry matches the inbound interface. If they do not match, the device considers the source IP address invalid and discards the packet. After a device enabled with strict URPF receives a BFD echo packet that is looped back, it checks the source IP address of the packet. As the source IP address of the echo packet is a local IP address of the device, the packet is sent to the platform without being forwarded at the lower layer. As a result, the device considers the packet invalid and discards it.

Table 11-9 Differences between BFD echo sessions and common static single-hop sessions

BFD Session	Supported IP Type	Session Type	Descriptor	Negotiation Prerequisite	IP Header
Common static single-hop session	IPv4 and IPv6	Static single-hop session	MD and YD must be configured.	A matching session must be established on the peer.	The source and destination IP addresses are different.
Passive BFD echo session	IPv4 and IPv6	Dynamic single-hop session	No MD or YD needs to be configured.	A matching session must be established and echo must be enabled on the peer.	Both the source and destination IP addresses are a local IP address of the device.
One-arm BFD echo session	IPv4	Static single-hop session	Only MD needs to be configured (MD and YD are the same).	A matching session does not need to be established on the peer.	Both the source and destination IP addresses are a local IP address of the device.

11.9.4 Application Environment

 Copyright © Huawei Technologies Co., Ltd.

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

Example 1

Figure 11-36 shows a single-hop BFD session detecting a path between directly-connected devices. The BFD session is bound to the outgoing interface.

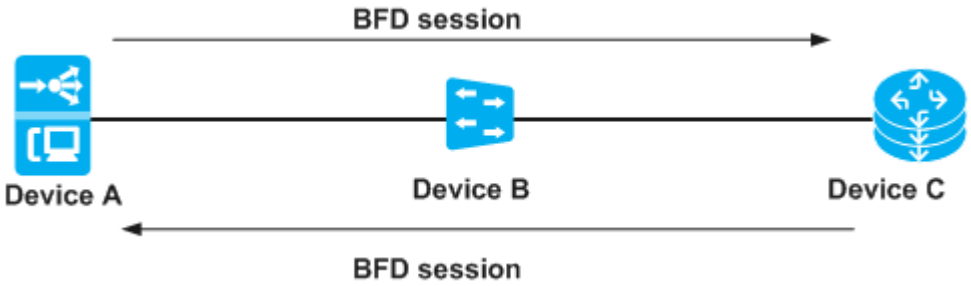
Figure 11-36 Single-hop BFD for IP



Example 2

Figure 11-37 shows a multi-hop BFD session detecting a path between Device A and Device C. The BFD session is bound to the peer IP address but not the outgoing interface.

Figure 11-37 Multi-hop BFD for IP



11.9.4.2 BFD for OSPF

A link fault or change in topology may lead to rerouting in a network. Quick convergence of a routing protocol is important for improving network availability.

A feasible solution is to fast detect the fault and immediately notify the routing protocol of the fault.

 Copyright © Huawei Technologies Co., Ltd.

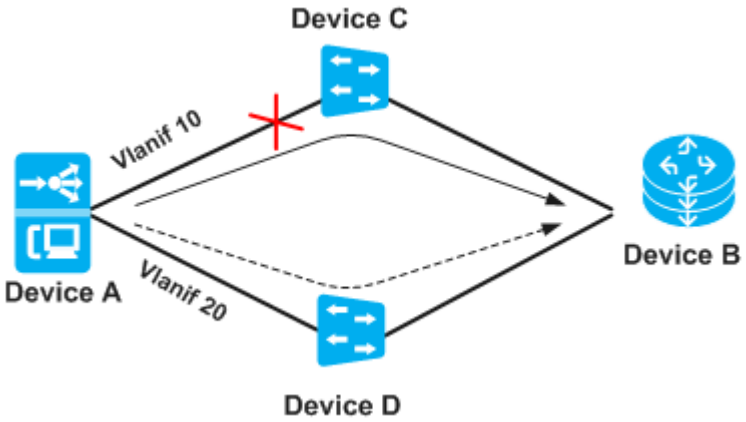
In BFD for OSPF, OSPF is associated with a BFD session. The BFD session fast detects a link fault and notifies OSPF of the fault. In this manner, OSPF speeds up responses to changes in network topology.

Table 11-10 shows convergence speed statistics when OSPF is and is not associated with a BFD session.

Table 11-10 OSPF convergence speed statistics

Associated with BFD	Link Fault Detection Mechanism	Convergence Speed
No	OSPF Hello keepalive timer timeout	Within seconds
Yes	BFD session in the Down state	Within milliseconds

Figure 11-38 BFD for OSPF networking diagram



As shown in Figure 11-38, Device A sets up OSPF neighbor relationships with Device C and Device D. The outbound interface VLANIF 10 on Device A is connected to Device B through Device 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 Device A and Device C, the BFD session detects the fault and notifies Device A.
2. Device A processes the neighbor-Down event and recalculates routes. Then, the outbound interface changes to VLANIF 20 on Device A, which is connected Device B through Device D.

11.9.4.3 BFD for IS-IS

Generally, the interval at which the Intermediate System to Intermediate System (IS-IS) protocol sends Hello messages is 10 seconds. If a device does not receive any Hello message from its neighbor within three Hello intervals, the device deletes the neighbor. Therefore, it takes a device a number of seconds to detect that a neighbor is Down. This leads to the loss of a large number of packets in a high-speed network.

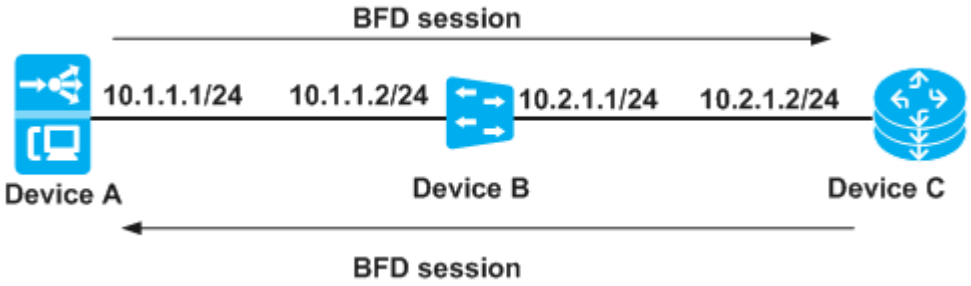
 Copyright © Huawei Technologies Co., Ltd.

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 Routing Management Module (RM). IS-IS processes the neighbor-Down event and quickly sends the link state PDU (LSP), and performs the partial route calculation (PRC). In this manner, IS-IS routes fast converge.

The BFD fault detection interval is at the millisecond level. Instead of replacing the IS-IS Hello mechanism, BFD works with IS-IS to detect the adjacency fault more quickly. In addition, BFD instructs IS-IS to recalculate routes, ensuring correct packet forwarding.

The RM allows IS-IS and BFD to interact with each other. Through the RM, IS-IS instructs BFD to dynamically set up or delete BFD sessions. The BFD event messages are also delivered to IS-IS through the RM.

Figure 11-39 BFD for IS-IS networking diagram



After BFD is enabled on Device A, Device B, and Device C, the BFD session can quickly detect faults on the link between Device A and Device B, 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 Device C (Device B's neighbor) can receive the updated LSPs from Device B in time. This implements fast network topology convergence.

11.9.4.4 BFD for BGP

The Border Gateway Protocol (BGP) periodically sends Keepalive messages to its peer to monitor the neighbor status. This detection process lasts more than 1 second. When the data is transmitted at gigabit rates, a large amount of data will be discarded, which cannot meet the requirement for carrier-class reliability.

BFD for BGP was developed to compensate for this shortcoming. The BFD session can fast detect a fault on a link between BGP peers and notify BGP, ensuring fast convergence.

NOTE

By default, a multi-hop BGP session is established between Huawei devices that set up an IBGP peer relationship. A BFD for IGP session and A BFD for IBGP session cannot be both set up between a Huawei device and a non-Huawei device that sets up a single-hop BGP session with its peer by default. In such a situation, setting up only A BFD for IGP session or A BFD for IBGP session between the Huawei and non-Huawei devices is recommended.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-40 BFD for BGP networking



As shown in **Figure 11-40**, Device A and Device B are directly connected through the External Border Gateway Protocol (EGBP). A BFD session is established to detect the BGP neighbor relationship between Device A and Device B. When the link between Device A and Device B is faulty, the BFD session can quickly detect the fault and notify BGP.

11.9.4.5 BFD for RSVP

BFD monitors RSVP neighbor relationships. When a Layer 2 device (a hub for example) exists between RSVP neighboring nodes, the two nodes can detect a link fault only using the Hello mechanism in seconds. This process results in the loss of lots of data. BFD for RSVP rapidly detects faults in a link between RSVP neighboring nodes within milliseconds. BFD for RSVP applies to TE FRR networks, on which Layer 2 devices exist on a primary CR-LSP between the PLR and its RSVP neighboring node.

BFD for RSVP is an IP-layer detection. Only a single-hop BFD session can be set up between RSVP neighboring nodes.

Figure 11-41 BFD for RSVP



A BFD session for RSVP is set up to monitor the link between RSVP neighbors. The RSVP module can rapidly detect a link failure.

BFD for RSVP can share BFD sessions with BFD for Open Shortest Path First (OSPF), BFD for Intermediate System to Intermediate System (IS-IS), or BFD for Border Gateway Protocol (BGP). The local node selects the smallest values of parameters between the two ends of the shared BFD session as local BFD parameters. The parameters include the interval at which BFD packets are sent, interval at which BFD packets are received, and local detection multiplier.

 Copyright © Huawei Technologies Co., Ltd.

11.9.4.6 BFD for LAG

Link aggregation group (LAG) is a feature to achieve load balancing and protection of Ethernet links. By bundling multiple Ethernet links between 2 Ethernet switching devices into a LAG, the Ethernet links can form a high-bandwidth Ethernet link. Ethernet traffic between the 2 Ethernet switching devices is automatically distributed on member links to implement load balancing. When a member link becomes faulty, services are automatically distributed to other normal links to implement protection switching.

The LAG feature detects faults on member links by using either of the following mechanisms:

- Port status detection. A port in the **Down** state is determined faulty. The detection is fast, but only physical-layer faults can be detected. In some Ethernet working modes, the faults of only the receive links can be detected.
- LACP protocol detection. Ethernet switching devices at both ends of a LAG periodically send LACP frames to each other. By detecting LACP frames and parsing the information carried in the LACP frames, the Ethernet switching devices detect whether a link is faulty. The detection is comprehensive, but the LACP protocol is a slow protocol and the sending interval is long (in seconds). Therefore, the detection is slow.

Figure 11-42 BFD for LAG



Based on the exchange of BFD packets, the BFD for LAG feature can quickly detect various forwarding faults of LAG member links and distribute the services on the faulty link to other normal links in a timely manner. This overcomes the limitations of the 2 detection mechanisms of the LAG feature. The BFD for LAG feature has the following advantages:

- BFD for LAG has a faster detection speed than LACP and can substitute for LACP.
- BFD for LAG can detect Layer 3 packet forwarding capabilities. In contrast, LACP is a Layer 2 protocol and can detect only Layer 2 forwarding capabilities.

11.9.4.7 BFD for PW

The bidirectional forwarding detection (BFD) for pseudo wire (PW) is used to quickly check the connectivity of the PWs in virtual private wire service (VPWS) and virtual private LAN service (VPLS) applications, providing a basis for fast switching for PW redundance. BFD is a fast detection mechanism, which consumes a lot of resources. Generally, BFD is used only for key PWs.

 Copyright © Huawei Technologies Co., Ltd.

Implementation Principle

In BFD for PW detection, BFD packets are encapsulated in a PW and sent periodically to check the connectivity of the PW.

This version supports the following types of BFD for PW:

- Static BFD for VPWS
- Dynamic BFD for VPWS
- Static BFD for VPLS

NOTE

Static sessions are implemented by Huawei. To interconnect with devices of other vendors, you can use dynamic BFD sessions.

In the PW standard, VCCV is used to define the connectivity check channel. That is, continuity check (CC) and CV are used to define the channel type and check type for connectivity check.

- **CC type:**
  - Control word mode: This field is encapsulated to indicate that a packet is a BFD detection packet.
  - Router Alert Label mode: This field is encapsulated to indicate that a packet is a BFD detection packet.
  - TTL mode: The TTL expiration mode is used to indicate that a packet is a BFD detection packet.

  This version supports only the control word mode.
- **CV type:**

Bit	**Value**	**MPLS PW CV Type**
Bit 0	**0x01**	**ICMP Ping**
Bit 1	**0x02**	**LSP Ping**
Bit 2	**0x04**	Indicates IP or UDP BFD.
Bit 3	**0x08**	Indicates IP/UDP BFD +status notification.
Bit 4	**0x10**	Indicates raw BFD.
Bit 5	**0x20**	Indicates raw BFD +status notification.
Bit 6	**0x40**	Reserved.
Bit 7	**0x80**	Reserved.

The BFD feature of this version supports the following CV types:

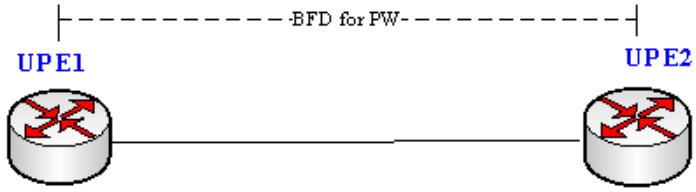
- All supported BFD for PW session types support 0x04.

 Copyright © Huawei Technologies Co., Ltd.

– Dynamic BFD for VPWS also supports 0x10.

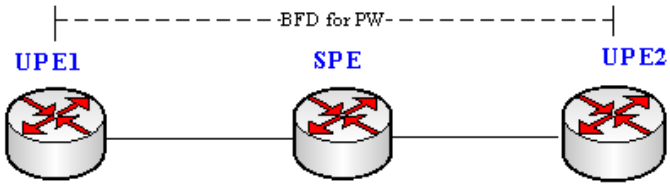
VPWS Scenario

Figure 11-43 Checking the connectivity between PEs for a single-hop PW



In this scenario, the PW connectivity between UPE1 and UPE2 is checked.

Figure 11-44 Checking the connectivity between PEs for a multi-hop PW.



In this VPWS scenario, the superstratum provider edge (SPE) in the middle switches PW labels. This version does not support the PW connectivity check between UPE1 and the SPE.

NOTE

In this scenario, the MA5800 cannot function as an SPE.

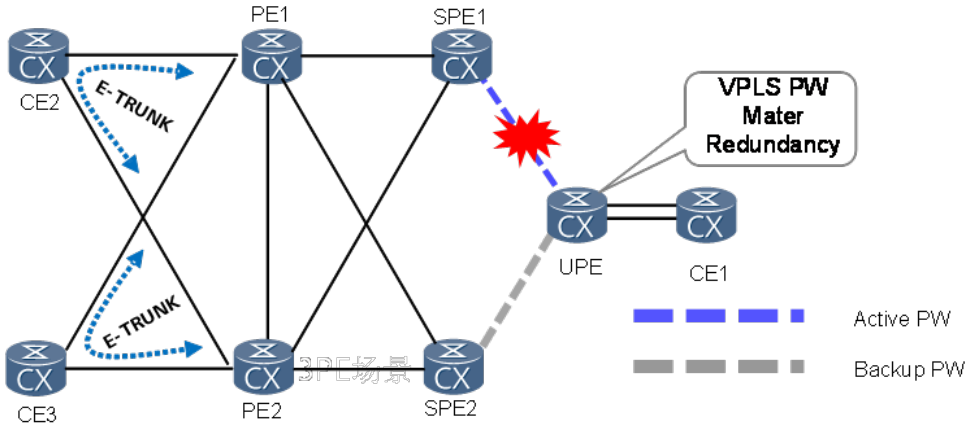
VPLS Scenario

In VPLS scenarios, BFD for PW cannot traverse SPEs on the hierarchical virtual private LAN service (HVPLS) network. That is, BFD for PW can only detect the segment from a UPE to an SPE on the VPLS network.

 Copyright © Huawei Technologies Co., Ltd.

As shown in the figure, the customer edge (CE) is dual-homed to the VPLS network, and Eth-Trunk is deployed on the CE and PE to ensure link reliability. The CE communicates with remote devices on the VPLS network through PE1.

Figure 11-45 Failure of the primary PW between the UPE and SPE1



VPLS PW redundancy is configured between SPEs and UPEs, and BFD is configured for the primary and secondary PWs.

Failover: Upon detecting that the primary PW is down, the UPE switches traffic to the secondary PW and sends MAC Withdraw messages in which the PE ID field carries the SPE1 LSR ID to SPE2. SPE2 transparently transmits the MAC Withdraw messages to PE1 and PE2. SPE2, PE1, and PE2 clear the MAC addresses learned from SPE1.

Switchback: After the primary PW recovers, the BFD negotiation is up, and the UPE instructs SPE2 to change its PW forwarding status to standby and SPE1 to change its PW forwarding status to active. The UPE sends MAC Withdraw messages in which the PE ID field carries the SPE2 LSR ID to SPE1. SPE1 transparently transmits the MAC Withdraw messages to PE1 and PE2. SPE1, PE1, and PE2 clear the MAC addresses learned from SPE2. PE1 and PE2 then relearn MAC addresses by broadcasting upon receiving multicast packets from the primary PW.

11.9.5 Configuring the BFD

This topic describes how to configure the BFD on the MA5800.

Configuration Scenario Description

- When you need to configure BFD in IP static route, OSPF, IS-IS, BGP, RSVP, and LAG scenarios, refer to this section to simplify the configuration process.
- When you need to configure links not belonging to the preceding scenarios, configure the links one by one by referring to "Configuring BFD Sessions" in this chapter. If BFD echo is required, refer to **11.9.5.2 Configuring One-Arm BFD Echo Sessions** to make the configurations.

11.9.5.1 Configuring BFD Sessions

A bidirectional forwarding detection (BFD) session rapidly detects faults in links over a network.

 Copyright © Huawei Technologies Co., Ltd.

Context

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 uses the local and remote discriminators to differentiate multiple BFD sessions between the same pair of systems. Based on the differences in methods of creating the local and the remote discriminators, MA5800 supports the following types of BFD sessions:

- Static BFD sessions with manually-specified discriminators

  The local and remote discriminators must be set manually. The discriminators on the remote end must also be manually specified.
- Static BFD sessions with automatically-negotiated discriminators

  If a dynamic BFD session is used by a remote device, a static BFD session with automatically negotiated discriminators must be created on a local device to interwork with the remote device and support the BFD for static routes. The discriminators on the remote end can be automatically negotiated or a dynamic BFD session can be established on the remote end.
- For BFD sessions dynamically triggered by protocols, no local or remote discriminator needs to be set. The implementation method is as follows:
  - BFD sessions with dynamically-allocated local discriminators.
  - BFD sessions with self-learned remote discriminators.

Procedure

Step 1 Enable BFD globally.

1. Run the **bfd** command to enable BFD globally and enter the BFD mode.

   BFD must be enabled globally before configurations relevant to BFD are performed. By default, BFD is disabled globally.
2. Run the **quit** command to quit the BFD mode.

Step 2 Create a BFD session. Select one of the following steps depending on the type of link to be checked by BFD.

- For an IPv4 link

  Run the **bfd bind peer-ip** command to create a BFD session.

  If the **bfd bind peer-ip source-ip auto** command is run, a BFD session is set up through automatic negotiations over discriminators. The device on which such a BFD session is created can interoperate with another device on which a dynamic BFD is set up. This command is used to configure BFD sessions for IPv4 static routes.
- For an IPv6 link

  Run the **bfd bind peer-ipv6** command to create a BFD session.

  If the **bfd bind peer-ipv6 source-ipv6 auto** command is run, a BFD session is set up through automatic negotiations over discriminators. The device on which such a BFD session is created can interoperate with another device on which a dynamic BFD is set up. This command is used to configure BFD sessions for IPv6 static routes.

 Copyright © Huawei Technologies Co., Ltd.

Pay attention to the following points:

- If a single-hop BFD session is to be set up on an interface for the first time, the interface and its peer address must be bound to the BFD session. The bindings cannot be modified after the BFD session is successfully created.
- If a multi-hop BFD session is to be set up on an interface for the first time, the peer address must be bound to the BFD session. The bindings cannot be modified after the BFD session is successfully created.
- During BFD configuration items are being created, the system checks only the format, not the correctness, of an IP address. Either an incorrect peer or source IP address leads to a failure in creating a BFD session.

Step 3 Configure the discriminators.

1. Run the **discriminator local** *discr-value* command to configure a local discriminator.
2. Run the **discriminator remote** *discr-value* command to configure a remote discriminator.

The local discriminator set on a device is equal to the remote discriminator set on a remote device, and the remote discriminator set on the local device is equal to the local discriminator set on the remote device. If the discriminators on the device and the remote device do not match, the session cannot be created. After the local and remote discriminators are set, they cannot be changed.

Step 4 (Optional) Configure the BFD parameters.

Select the following desired operations:

- Modify the detection time.
  - Run the **min-tx-interval** command to configure the interval for sending BFD packets.
  - Run the **min-rx-interval** command to configure the interval for receiving BFD packets.
- Run the **detect-multiplier** command to configure the local detection multiplier.
- Run the **description** command to add the description of a BFD session. Descriptions of BFD sessions help you distinguish between various BFD sessions.

  The **description** command takes effect only on the statically configured BFD sessions, rather than the BFD sessions that are dynamically configured or the BFD sessions that are set up through automatic negotiations over discriminators.
- Run the **tos-exp** command to configure the priority of the BFD packet. By default, the highest priority 7 is adopted. When the system is congested, the BFD packet with higher priority can be sent first.

  The command is valid in static BFD mode. The command is invalid in dynamic BFD mode.
- Run the **wtr** command to configure the time of waiting for recovery of the BFD session. By default, the value is 0, indicating no waiting.

  The BFD session is unidirectional. The detection is performed by BFD parameters configured on both ends respectively. If wait-to-recovery (WTR) is

 Copyright © Huawei Technologies Co., Ltd.

needed, configure it on two ends manually. Or, when the status of the session on one end changes, the applications on both ends can find that the states of the BFD sessions are inconsistent.

Step 5 Query the BFD session information and BFD session statistics.

- Run the **display bfd configuration** command to query the BFD configuration.
- Run the **display bfd interface** command to query the BFD configuration on an interface.
- Run the **display bfd session** command to query the BFD session information.
- Run the **display bfd statistics** command to query the BFD global statistics.
- Run the **display bfd statistics session** command to query the BFD session statistics.

----End

Example

Example: VLAN 10 has been created, and the IP address of VLANIF 10 has been configured. Single-hop BFD needs to be configured on VLANIF 10. Assume that the peer IP address is 10.1.1.1/24, BFD session name is test, the local discriminator is 100, the remote discriminator is 200, the minimum transmit interval and minimum receive interval of BFD control packets are both 10 milliseconds, the local detection multiplier is 3 (default value). The OLT is configured with xPON dual-homing protection.

```
Configure a BFD session as follows:
huawei(config)#bfd
huawei(config-bfd)#quit
huawei(config)#bfd test bind peer-ip 10.1.1.1 interface vlanif 10
huawei(config-bfd-session-test)#discriminator local 100   //The discriminator corresponds to the peer end.
huawei(config-bfd-session-test)#discriminator remote 200   //The discriminator corresponds to the peer end.
huawei(config-bfd-session-test)#min-tx-interval 10
huawei(config-bfd-session-test)#min-rx-interval 10
Add a protection group to the OLT as follows:
huawei(config)#protect-group protect-target gpon-uni-ont workmode dual-parenting
Bind the OLT protection group to the uplink interface, and create an association between the protection group and the BFD session as follows:
huawei(protect-group-0)#uplink-monitor port bfd test   //test is the name configured for the BFD.
Add a member to the OLT protection group and configure a work-side ONT as follows:
huawei(protect-group-0)#protect-group member port 0/1/0 ont 0 role work
```

11.9.5.2 Configuring One-Arm BFD Echo Sessions

A one-arm BFD echo session can be established on a device that supports BFD. After receiving a one-arm BFD echo session packet, devices that do not support BFD immediately loop back the packet, implementing quick link failure detection.

Procedure

Step 1 Enable BFD globally.

1. Run the **bfd** command to enable BFD globally and enter the BFD mode.

   BFD must be enabled globally before configurations relevant to BFD are performed. By default, BFD is disabled globally.

2. Run the **quit** command to quit the BFD mode.

 Copyright © Huawei Technologies Co., Ltd.

Step 2 Run the **bfd one-arm-echo** command to create a one-arm BFD echo session.

Pay attention to the following points:

- If a single-hop BFD session is to be set up on an interface for the first time, the interface and its peer address must be bound to the BFD session. The bindings cannot be modified after the BFD session is successfully created.
- During BFD configuration items are being created, the system checks only the format, not the correctness, of an IP address. Either an incorrect peer or source IP address leads to a failure in creating a BFD session.

Step 3 Run the **discriminator local** *discr-value* command to configure a local discriminator.

•When creating a one-arm BFD echo session, you need to specify only the **local** *discr-value* parameter.

Step 4 (Optional) Configure the BFD parameters.

Select the following desired operations:

- Run the **min-echo-rx-interval** command to configure the interval for receiving BFD packets.
- Run the **detect-multiplier** command to configure the local detection multiplier.
- Run the **description** command to add the description of a BFD session. Descriptions of BFD sessions help you distinguish between various BFD sessions.

  The **description** command takes effect only on the statically configured BFD sessions, rather than the BFD sessions that are dynamically configured or the BFD sessions that are set up through automatic negotiations over discriminators.
- Run the **tos-exp** command to configure the priority of the BFD packet. By default, the highest priority 7 is adopted. When the system is congested, the BFD packet with higher priority can be sent first.
- Run the **wtr** command to configure the time of waiting for recovery of the BFD session. By default, the value is 0, indicating no waiting.

Step 5 Query the BFD session information and BFD session statistics.

- Run the **display bfd configuration** command to query the BFD configuration.
- Run the **display bfd interface** command to query the BFD configuration on an interface.
- Run the **display bfd session** command to query the BFD session information.
- Run the **display bfd statistics** command to query the BFD global statistics.
- Run the **display bfd statistics session** command to query the BFD session statistics.

----End

Example

Assume that the peer IP address is 10.1.1.1/24, BFD session name is test, the local discriminator is 100, the minimum transmit interval and minimum receive interval

 Copyright © Huawei Technologies Co., Ltd.

of BFD control packets are both 10 milliseconds, the local detection multiplier is 3 (default value), VLAN 10 is created, and the IP address of VLAN interface 10 is configured. To configure one-arm BFD echo single-hop detection on VLAN interface 10, run the following commands:

```
huawei(config)#bfd
huawei(config-bfd)#quit
huawei(config)#bfd test bind peer-ip 10.1.1.1 interface vlanif 10 one-arm-echo
huawei(config-bfd-session-test)#discriminator local 100
huawei(config-bfd-session-test)#min-echo-rx-interval 10
```

11.9.5.3 Configuring BFD for IPv4 Static Routes

BFD for IPv4 static routes enables devices to fast detect link changes, improving network reliability.

Prerequisites

Configure parameters of the link layer protocol and IP addresses for interfaces and ensure that the link layer protocol on the interfaces is Up.

Context

Optimal IPv4 static routes are delivered to the forwarding table for packet forwarding. However, IPv4 static routes cannot detect the status of the link to the next hop. You can bind IPv4 static routes to BFD sessions. A BFD session can fast detect changes over a link and inform the routing management system of the changes. The routing management system immediately deletes the IPv4 static route that is bound to the BFD session from the forwarding table and recalculates another active route. In this manner, fast route convergence is implemented.

Procedure

Step 1 Enable BFD globally.

1. Run the **bfd** command to enable BFD globally and enter the BFD mode.
2. Run the **quit** command to quit the BFD mode.

Step 2 Create a static BFD session.

1. Run the **bfd bind peer-ip** command to create a BFD session.
2. Configure the discriminators.

   The local discriminator set on a device is equal to the remote discriminator set on a remote device, and the remote discriminator set on the local device is equal to the local discriminator set on the remote device. If the discriminators on the device and the remote device do not match, the session cannot be created.

   a. Run the **discriminator local** *discr-value* command to configure a local discriminator.
   b. Run the **discriminator remote** *discr-value* command to configure a remote discriminator.
3. Run the **quit** command to quit the BFD session mode.
4. Run the **ip route-static** *ip-address* { *mask* | *mask-length* } *interface-type interface-number* [*nexthop-address*] [**preference** *preference*] **track bfd-**

 Copyright © Huawei Technologies Co., Ltd.

session *cfg-name* [**description** *text*] command to bind a BFD session to an IPv4 static route on the public network.

Step 3 Create a dynamic BFD session.

1. Run the **ip route-static** *ip-address* { *mask* | *mask-length* } *interface-type interface-number* [*nexthop-address*] [**preference** *preference*] **bfd enable** [**description** *text*] command to bind a BFD session to an IPv4 static route on the public network.
2. (Optional) Run the **ip route-static default-bfd** command to configure global BFD parameters for IPv4 static routes.
3. Run the **ip route-static bfd** command to configure BFD parameters for an IPv4 static route.

Step 4 Query the BFD session information.

Run the **display bfd session** command to query the BFD session.

----End

11.9.5.4 Configuring BFD for OSPF

Configuring BFD for OSPF helps the system to rapidly detect the link status and speeds up OSPF convergence in the case of a link failure.

Prerequisites

Basic OSPF functions are configured.

Context

After BFD for OSPF is configured, when detecting a link fault, BFD rapidly notifies the routers on both ends of the link of the fault, triggering rapid OSPF convergence. When the OSPF neighbor relationship goes Down, the BFD session will be dynamically deleted.

Procedure

Step 1 Enable BFD globally.

1. Run the **bfd** command to enable BFD globally and enter the BFD mode.
2. Run the **quit** command to quit the BFD mode.

Step 2 Create a BFD session.

- To configure BFD for all interfaces in a specified OSPF process, perform the following steps:

1. Run the **ospf** command to enter the OSPF mode.
2. Run the **bfd all-interfaces enable** command to enable BFD for all the interfaces in a specific OSPF process. The default parameter values are used.
3. (Optional) Run the **bfd all-interfaces** { **min-rx-interval** *receive-interval* | **min-tx-interval** *transmit-interval* | **detect-multiplier** *multiplier-value* } * command to modify the parameters of BFD session.
4. Run the **quit** command to quit the OSPF mode.

 Copyright © Huawei Technologies Co., Ltd.

5. (Optional) Prevent an interface from dynamically creating a BFD session.

   After BFD for OSPF is configured, all interfaces on which neighbor relationships are Full in the OSPF process will create BFD sessions. To prevent specific interfaces from being enabled with BFD, disable these interfaces from dynamically creating BFD sessions.

   a. Run the **interface vlanif** command to enter the VLANIF interface mode.
   b. Run the **ospf bfd block** command to prevent the interface from dynamically creating a BFD session.
   c. Run the **quit** command to quit the VLANIF interface mode.

- To configure BFD for specified OSPF interfaces, perform the following steps:

1. Run the **interface vlanif** command to enter the VLANIF interface mode.
2. Run the **ospf bfd enable** command to enable BFD for an OSPF interface.
3. (Optional) Run the **ospf bfd** { **min-rx-interval** *receive-interval* | **min-tx-interval** *transmit-interval* | **detect-multiplier** *multiplier-value* } * command to modify the parameters of BFD session.
4. Run the **quit** command to quit the VLANIF interface mode.

Step 3 Query the BFD session information.

Run the **display ospf bfd session** command to query the BFD session.

----End

11.9.5.5 Configuring BFD for IS-IS

Configuring BFD for IS-IS helps the system to rapidly detect the link status and speeds up IS-IS convergence in the case of a link failure.

Prerequisites

Basic IS-IS functions are configured.

Context

To accelerate IS-IS convergence speed when the link status changes, you can configure BFD on the IS-IS link. The access device supports configuration of static and dynamic BFD for IS-IS. When BFD sessions are configured in both methods, the static BFD session takes precedence over the dynamic BFD session.

- Static BFD refers to configuring BFD session parameters manually including local and remote identifiers and delivering BFD session setup requests manually.
- Dynamic BFD refers to that routing protocols dynamically trigger the establishment of BFD sessions. When setting up new neighbor relationship, routing protocols send parameters of neighbors and detection parameters (including source and destination IP addresses) to the BFD module. BFD then sets up sessions according to the received parameters between neighbors. Dynamic BFD is more flexible than static BFD.

The procedures and methods for planning and configuring IPv6 services are similar to those for configuring IPv4 services. Only some commands are different, which are described in the following configuration procedure.

 Copyright © Huawei Technologies Co., Ltd.

Procedure

Step 1 Enable BFD globally.

1. Run the **bfd** command to enable BFD globally and enter the BFD mode.
2. Run the **quit** command to quit the BFD mode.

Step 2 Create static BFD for IS-IS.

1. Run the **interface vlanif** command to enter the VLANIF interface mode.
2. Run the **isis bfd static** command to enable static BFD function on the interface.
3. Run the **quit** command to quit the VLANIF interface mode.
4. Run the **bfd bind peer-ip** command to create a BFD session.
5. Configure the discriminators.

   The local discriminator set on a device is equal to the remote discriminator set on a remote device, and the remote discriminator set on the local device is equal to the local discriminator set on the remote device. If the discriminators on the device and the remote device do not match, the session cannot be created.

   a. Run the **discriminator local** *discr-value* command to configure a local discriminator.
   b. Run the **discriminator remote** *discr-value* command to configure a remote discriminator.
6. Run the **quit** command to quit the BFD session mode.

Step 3 Create dynamic BFD for IS-IS.

- To configure BFD for all interfaces in a specified IS-IS process, perform the following steps:

1. Run the **isis** command to enter the IS-IS mode.
2. Run the **bfd all-interfaces enable** command to enable BFD for all the interfaces in a specific IS-IS process. The default parameter values are used.
3. (Optional) Run the **bfd all-interfaces** { **min-rx-interval** *receive-interval* | **min-tx-interval** *transmit-interval* | **detect-multiplier** *multiplier-value* } * command to modify the parameters of BFD session.
4. Run the **quit** command to quit the IS-IS mode.
5. (Optional) Prevent an interface from dynamically creating a BFD session.

   After BFD for IS-IS is configured, all interfaces on which neighbor relationships are Full in the IS-IS process will create BFD sessions. To prevent specific interfaces from being enabled with BFD, disable these interfaces from dynamically creating BFD sessions.

   a. Run the **interface vlanif** command to enter the VLANIF interface mode.
   b. Run the **isis bfd block** command to prevent the interface from dynamically creating a BFD session.
   c. Run the **quit** command to quit the VLANIF interface mode.

- To configure BFD for specified IS-IS interfaces, perform the following steps:

1. Run the **interface vlanif** command to enter the VLANIF interface mode.

 Copyright © Huawei Technologies Co., Ltd.

2. Run the **isis bfd enable** command to enable BFD for an IS-IS interface.
3. (Optional) Run the **isis bfd** { **min-rx-interval** *receive-interval* | **min-tx-interval** *transmit-interval* | **detect-multiplier** *multiplier-value* } * command to modify the parameters of BFD session.
4. Run the **quit** command to quit the VLANIF interface mode.

Step 4 Query the BFD session information.

Run the **display isis bfd interface** command to query the BFD configuration information of an IS-IS interface.

Run the **display isis bfd session** command to query the BFD session information.

----End

11.9.5.6 Configuring BFD for BGP or BGP4+

BFD for BGP or BGP4+ speeds up fault detection and therefore increases the route convergence speed.

Prerequisites

Basic BGP or BGP4+ functions are configured.

Context

BGP periodically sends Keepalive packets to its peers to detect the status of its peers. The detection mechanism, however, takes more than one second. When the data transmission rate reaches the level of Gbit/s, such slow detection will cause a large amount of data to be lost. As a result, the requirement for high reliability of carrier-class networks cannot be met.

BFD for BGP or BGP4+ detects faults on links between BGP peers within 50 milliseconds. The fast detection speed ensures fast BGP route convergence and minimizes traffic loss.

NOTE

By default, a multi-hop IBGP session is established between Huawei devices that set up an IBGP peer relationship. A BFD for IGP session and a BFD for IBGP session cannot be both set up between a Huawei device and a non-Huawei device that sets up a single-hop BGP session with its peer by default. In such a situation, setting up only a BFD for IGP session or a BFD for IBGP session between the Huawei and non-Huawei devices is recommended.

Procedure

Step 1 Enable BFD globally.

1. Run the **bfd** command to enable BFD globally and enter the BFD mode.

   BFD must be enabled globally before configurations relevant to BFD are performed. By default, BFD is disabled globally.
2. Run the **quit** command to quit the BFD mode.

Step 2 Create a BFD session.

 Copyright © Huawei Technologies Co., Ltd.

1. Run the **bgp** command to enter the BGP mode.
2. (Optional) Configure BFD for BGP or BGP4+ in the VPN. To configure BFD for BGP or BGP4+ for the public network, skip this step.
   - To configure BFD for BGP, run the **ipv4-family vpn-instance** *vpn-instance-name* command to enter the BGP-VPN instance IPv4 address family mode.
   - To configure BFD for BGP4+, run the **ipv6-family vpn-instance** *vpn-instance-name* command to enter the BGP-VPN instance IPv6 address family mode.
3. Run the **peer bfd enable** command to enable BFD for the peer or peer group, and a BFD session is established.

   A BFD session is set up only when the BGP session is in the Established state.

   After BFD is enabled for a peer group, BFD sessions will be created on the peers that belong to this peer group and are not configured with the **peer bfd block** command.

Step 3 (Optional) Run the **peer bfd** command to configure the BFD parameters.

The BFD parameters of peers take precedence over those of peer groups. If BFD parameters are configured on peers, they will be used in BFD session establishment.

The default intervals for transmitting BFD packets and the default detection multiplier are recommended. When changing the default values, pay attention to the network status and the network reliability requirement. A short interval for transmitting BFD packets can be configured for a link that has a higher reliability requirement. A long interval for transmitting BFD packets can be configured for a link that has a lower reliability requirement. There are three formulas:

- Actual interval for the local device to send BFD packets = max {Locally configured interval for transmitting BFD packets, Remotely configured interval for receiving BFD packets}
- Actual interval for the local device to receive BFD packets = max {Remotely configured interval for transmitting BFD packets, Locally configured interval for receiving BFD packets}
- Local detection period = Actual interval for receiving BFD packets x Remotely configured BFD detection multiplier

For example:

- On the local device, the configured interval for transmitting BFD packets is 200 ms, the interval for receiving BFD packets is 300 ms, and the detection multiplier is 4.
- On the peer device, the configured interval for transmitting BFD packets is 100 ms, the interval for receiving BFD packets is 600 ms, and the detection multiplier is 5.

Then:

- On the local device, the actual interval for transmitting BFD packets is 600 ms calculated by using the formula max {200 ms, 600 ms}; the interval for receiving BFD packets is 300 ms calculated by using the formula max {100

 Copyright © Huawei Technologies Co., Ltd.

ms, 300 ms}; the detection period is 1500 ms calculated by multiplying 300 ms by 5.

- On the peer device, the actual interval for transmitting BFD packets is 300 ms calculated by using the formula max {100 ms, 300 ms}; the interval for receiving BFD packets is 600 ms calculated by using the formula max {200 ms, 600 ms}; the detection period is 2400 ms calculated by multiplying 600 ms by 4.

wtr *wtr-value* can be specified in the command to suppress frequent BFD and BGP session flapping caused by link flapping. If a BFD session over a link goes down, it does not go Up immediately after the link recovers. Instead, the BFD session waits for the WTR timer to expire before going Up. If the link fails again before the WTR timer expires, BFD does not send a link fault message to BGP, and the BGP session status is stabilized.

Step 4 (Optional) Run the **peer bfd block** command to prevent a peer from inheriting the BFD function of the peer group to which it belongs.

If a peer joins a peer group enabled with BFD, the peer inherits the BFD configuration of the group and creates a BFD session. To prevent the peer from inheriting the BFD function of the peer group, perform this step.

The **peer bfd block** command and the **peer bfd enable** command are mutually exclusive. After the **peer bfd block** command is run, the BFD session is automatically deleted.

Step 5 Query the BFD session information.

- Run the **display bgp bfd session** command to query the BFD session between BGP peers.
- Run the **display bgp ipv6 bfd session** command to query the BFD session between BGP4+ peers.

----End

11.9.5.7 Configuring BFD for RSVP

BFD for RSVP is applied to a scenario where TE FRR is used and a Layer 2 device exists on the primary LSP between a PLR and its downstream neighbors. On a network where GR is enabled on the PLR and MP, BFD for RSVP is also recommended.

Prerequisites

The RSVP-TE tunnel is configured.

Context

By default, the interval at which RSVP Hello messages are sent is 3 seconds. The interval at which a neighbor is declared down is three times the interval at which RSVP Hello messages are sent. This allows devices to detect a fault in an RSVP neighbor at seconds level. If a Layer 2 device exists on a link between RSVP neighboring nodes, the neighboring node cannot rapidly detect the fault after the link fails, resulting in a great loss of data.

 Copyright © Huawei Technologies Co., Ltd.

BFD detects faults at millisecond level in protected links or nodes. BFD for RSVP rapidly detects faults in an RSVP neighbor, allowing packets to switch to a backup LSP rapidly.

NOTE

BFD for LSP can function properly though the forward path is an LSP and the backward path is an IP link. The forward path and the backward path must be established over the same link; otherwise, if a fault occurs, BFD cannot identify the faulty path. Before deploying BFD, ensure that the forward and backward paths are over the same link so that BFD can correctly identify the faulty path.

Perform the following steps on the two RSVP neighboring nodes between which a Layer 2 device resides:

Procedure

Step 1 Enable BFD globally.

1. Run the **bfd** command to enable BFD globally and enter the BFD mode.

   BFD must be enabled globally before configurations relevant to BFD are performed. By default, BFD is disabled globally.
2. Run the **quit** command to quit the BFD mode.

Step 2 Create a BFD session.

- If most RSVP interfaces on a node need BFD for RSVP, enable BFD for RSVP globally.

1. Run the **mpls** command to enter the MPLS mode.
2. Run the **mpls rsvp-te bfd all-interfaces enable** command to enable BFD for RSVP globally.

   After this command is run in the MPLS mode, BFD for RSVP is enabled on all RSVP interfaces except the interfaces with BFD for RSVP that are blocked.
3. (Optional) Block BFD for RSVP on the RSVP interfaces that need not BFD for RSVP.

   Run the **interface vlanif** command to enter the VLANIF interface mode, and run the **mpls rsvp-te bfd block** command to block BFD for RSVP.

- If certain RSVP interfaces on a node need BFD for RSVP, enable BFD for RSVP on the RSVP interfaces.

1. Run the **interface vlanif** command to enter the VLANIF interface mode.
2. Run the **mpls rsvp-te bfd enable** command to enable BFD for RSVP on an RSVP interface.

Step 3 (Optional) Configure the BFD parameters.

- If most RSVP interfaces on a node use the same BFD parameters, configure global BFD parameters.

1. Run the **mpls** command to enter the MPLS mode.
2. Run the **mpls rsvp-te bfd all-interfaces** command to configure global BFD parameters.

- If certain RSVP interfaces require BFD parameters different from global BFD parameters, configure BFD parameters on the RSVP interfaces.

 Copyright © Huawei Technologies Co., Ltd.

1. Run the **interface vlanif** command to enter the VLANIF interface mode.
2. Run the **mpls rsvp-te bfd** command to configure BFD parameters on an RSVP interface.

Step 4 Run the **display mpls rsvp-te bfd session** command to query the BFD session information.

----End

11.9.5.8 Configuring BFD for LAG

This section describes how to configure the BFD for LAG feature to quickly detect the status of a link aggregation group (LAG).

Prerequisites

Basic LAG functions have been configured.

Context

The BFD session for detecting an Ethernet LAG needs to be configured at both ends of the Ethernet LAG and cannot communicate with other types of BFD sessions.

After a BFD session for detecting an Ethernet LAG is created, the system creates a BFD sub-session for each member link. The discriminators of the BFD sub-sessions are determined through auto-negotiation. The discriminators of the BFD sessions are also generated by the local system. You do not need to configure the local and remote discriminators for the BFD sessions of the Ethernet LAG. The status of a BFD session in an Ethernet LAG becomes **Down** only when all BFD sub-sessions are **Down**.

Procedure

Step 1 Enable BFD globally.

1. Run the **bfd** command to enable BFD globally and enter BFD mode.
2. Run the **quit** command to quit the BFD mode.

Step 2 Create a BFD session.

1. Run the **bfd bind link-aggregation peer-ip** command to create a BFD session.
2. Run the **quit** command to quit the BFD mode.

Step 3 View the information about the BFD session.

Run the **display bfd session link-aggregation** command to view the BFD session information in the Ethernet LAG.

----End

11.9.6 BFD Reference Standards and Protocols

The following table lists the references.

 Copyright © Huawei Technologies Co., Ltd.

Document No.	Document Name	Protocol Compliance
RFC 5880	Bidirectional Forwarding Detection	• Compliant only with the synchronous mode • Compliant except for authentication According to RFC5880, the authentication field in a BFD control packet is optional. The authentication implementation of the device is in compliance with the RFC recommendation. The identifier of a BFD session is Your Discriminator carried in the packet. In a scenario where no authentication mechanism is available, if attackers detect that the IP address of the transmitter of a BFD session is the IP address of the local device and Your Discriminator of the BFD session is the same as that of the BFD session allocated by the local device, attackers can launch attacks, which will lead to abnormalities in the BFD session status and even in applications that detect the BFD session status. For a "BFD for IP" single-hop session, the TTL field carried in the IP packet is used to prevent remote attacks. If the TTL value equals 255, this packet is discarded. This method is not applicable to a "BFD for IP" multi-hop session.
RFC 5881	BFD for IPv4 and IPv6 (Single Hop)	Fully compliant
RFC 5882	Generic Application of BFD	• Compliant except for OSPF virtual connection
RFC 5883	BFD for Multihop Paths	• Compliant except for discriminator learning in outband mode • Compliant except for bidirectional connection
RFC 5884	BFD for mpls	Compliant except for the security mechanism
RFC 5885	BFD for PW	Compliant only with IP/UDP encapsulation

11.10 TWAMP

The Two-Way Active Measurement Protocol (TWAMP) is a technology that measures the round-trip performance of an IP network.

 Copyright © Huawei Technologies Co., Ltd.

11.10.1 TWAMP Introduction

This topic describes the definition, usage scenario, and benefits of Two-Way Active Measurement Protocol (TWAMP).

Definition

TWAMP is a technology that measures the round-trip performance of an IP network. It defines a method for measuring round-trip IP network performance between two TWAMP-capable devices. TWAMP measures the network delay, jitter, and packet loss rate (PLR) through UDP packets. Using which, the overall IP transmission performance can be measured by cooperation between deployed network devices.

TWAMP uses the client/server mode.

- The client establishes, starts, and stops a TWAMP session, and generates and maintains statistics about IP network performance.
- The server responds to the client's request to establish, start, and stop a TWAMP session.

Usage Scenario

As networks continue to rapidly develop and applications widely apply, more and more new services are being deployed, posing increasingly high requirements on network performance. A tool that rapidly provides statistics about IP network performance is vital.

Traditionally, NEs themselves generate and maintain statistics about IP network performance. To display statistics about entire network performance, a network management system (NMS) needs to manage multiple NEs and collect their statistics. However, in scenarios where no NMS is deployed or the NMS is incapable of collecting statistics, carriers cannot rapidly obtain network performance statistics.

To solve this problem, TWAMP is introduced. NEs themselves no longer need to generate or maintain statistics. The performance management system can easily obtain statistics about entire network performance by managing TWAMP clients only.

Benefits

Compared with the traditional tools used to collect statistics about IP network performance, TWAMP has the following advantages:

- Has low requirements on NMS capabilities. Compared with network quality analysis (NQA), TWAMP has a unified detection model and unified packet format. Its configuration method is simple and no test case needs to be designed.
- Does not require clock synchronization. Compared with IP flow performance measurement (IP FPM), TWAMP is easy to obtain and deploy.

Therefore, TWAMP is applicable to scenarios where statistics about IP network performance must be rapidly obtained but do not need to be highly accurate.

 Copyright © Huawei Technologies Co., Ltd.

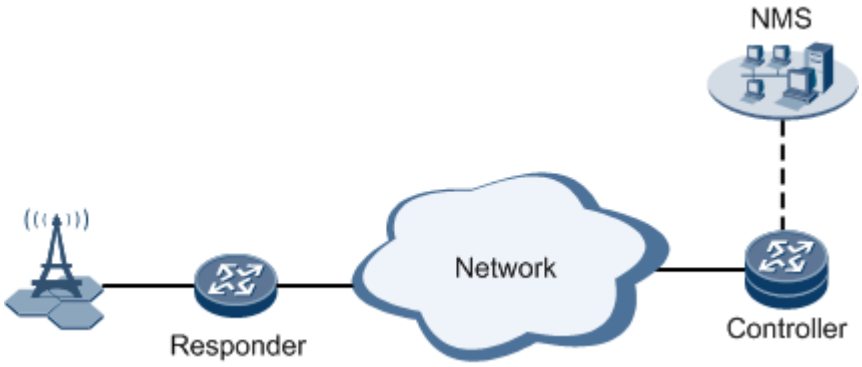
11.10.2 TWAMP Basic Concepts

This topic describes the basic concepts of the device roles and communication models of Two-Way Active Measurement Protocol (TWAMP).

Device Roles

TWAMP involves two physical devices: a Controller and a Responder. As shown in **Figure 11-46**, a device or tester works as the Controller (initiator), and the OLT or the ONT connected to an OLT works as the Responder (reflector).

Figure 11-46 Networking of TWAMP measurement



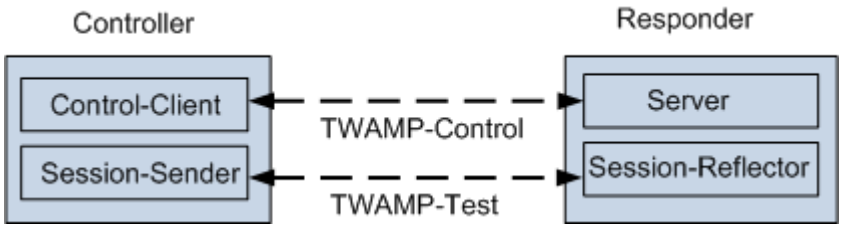
The Controller and Responder function as follows:

- Controller: Client of a TWAMP test. Serving as the initiator, it sends and receives packets over a test session, collects and calculates performance statistics, and reports these statistics to the NMS.
- Responder: Server of a TWAMP test. Serving as the receiver, it responds to packets received over a test session. It does not calculate or report performance statistics.

Communication Model

TWAMP uses client/server mode. In TWAMP Standard architecture, mapping relationships between physical devices and logical entities are shown in **Figure 11-47**. Specifically, Controller contains Control-Client and Session-Sender; Responder contains Server and Session-Reflector. The Control-Client and Server establish and start a test session. Once a test session starts, the Control-Client and Server notify the Session-Sender and Session-Reflector respectively of the session information and allow the Session-Sender to send probes and the Session-Reflector to respond to the probes.

Figure 11-47 TWAMP Standard Architecture



 Copyright © Huawei Technologies Co., Ltd.

Definition of logical roles:

- Control-Client: Establishes, starts, and stops a TWAMP session and collects statistics.
- Session-Sender: Sends probes for performance measurement after being notified by the Control-Client.
- Server: Responds to the Control-Client's request for establishing, starting, or stopping a TWAMP session.
- Session-Reflector: Responds to the probes sent by the Session-Sender after being notified by the Server.

Relationship of logical roles:

- Control-Client and Server exchange control packets through a TCP connection.
- Session-Sender and Session-Reflector exchange test packets through a UDP connection.

Protocol Packets

The Controller calculates the performance information according to fields carried in test packets. The TWAMP test packet is of the following format:

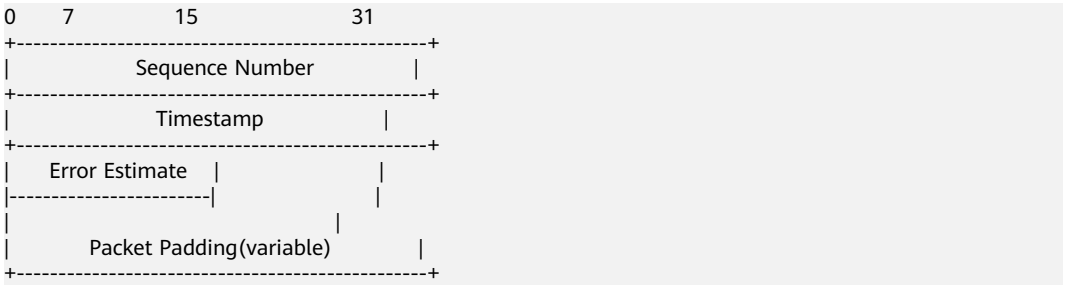


Table 11-11 Parameters of a TWAMP Test Packet

Field	Length	Description
Sequence Number	4 bytes	Sequence number (SN) of packets that is generated based on sequence of packets transmitted. It is numbered from 0 and increases with packet count. Every packet has an SN. SNs generated by the Session-reflector are irrelevant to those of packets received.
Timestamp	4 bytes	Transmission timestamp tagged on tested packets of the Session-Reflector.
Error Estimate	2 bytes	Error checks of the Session-Reflector.
Packet Padding	variable	Field for padding a packet. This field cannot be encrypted. Data fields of TWAMP test packets have the minimum 41 bytes in unauthentication mode and 104 bytes in authentication mode or encryption mode.

 Copyright © Huawei Technologies Co., Ltd.

Performance Measurement Data

TWAMP test data includes:

- Latency: Time from the point when the first bit of the packet enters the router to the point when the last bit of the packet outputs from the router.
- Jitter: Latency change. Due to network congestion, incorrect queue, or incorrect configuration, the interval at which the data packets (of even packets carried on consecutive data streams) are received on the receive is different. The change of the interval is called jitter.
- Packet loss rate (PLR): Ratio of the number of lost packets to the number of Tx packets.

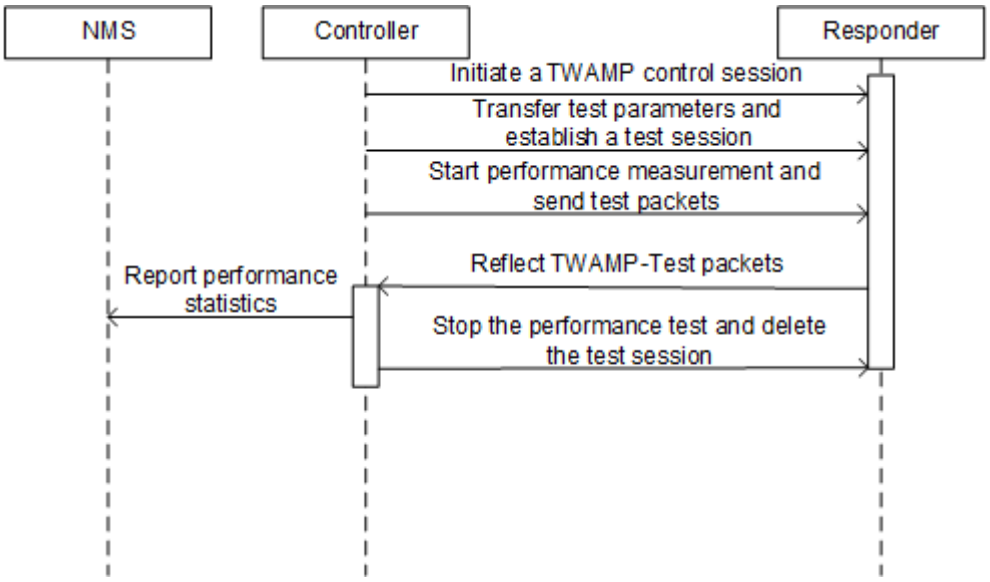
11.10.3 TWAMP Principles

This topic describes the principles and implementation process of Two-Way Active Measurement Protocol (TWAMP).

A TWAMP test covers how a control session is established, and how a test session is established, started, and stopped.

The Controller and Responder interact with each in the following way:

Figure 11-48 TWAMP test interaction



1. Control session establishment: The maintenance personnel create a control session on the configuration page of the Controller. Then, the Controller initiates a TWAMP control session.
2. Test session establishment: The Controller transmits test parameters to the Responder and the Responder initiates a test session.
3. Test session startup: The Controller starts a performance measurement test. In this test, the Controller sends the TWAMP-Test packet to the Responder and the Responder reflects the TWAMP-Test packet.
4. After receiving the reflected TWAMP-Test packet, the Controller calculates the round-trip performance information, including the packet loss rate (PLR),

 Copyright © Huawei Technologies Co., Ltd.

delay, and jitter according to the information carried in the packets, such as SN and timestamp. Then, the Controller reports the calculated performance information to the NMS.

5. Test session stop: The Controller stops the performance measurement test and deletes the Responder test session through the TWAMP control session.

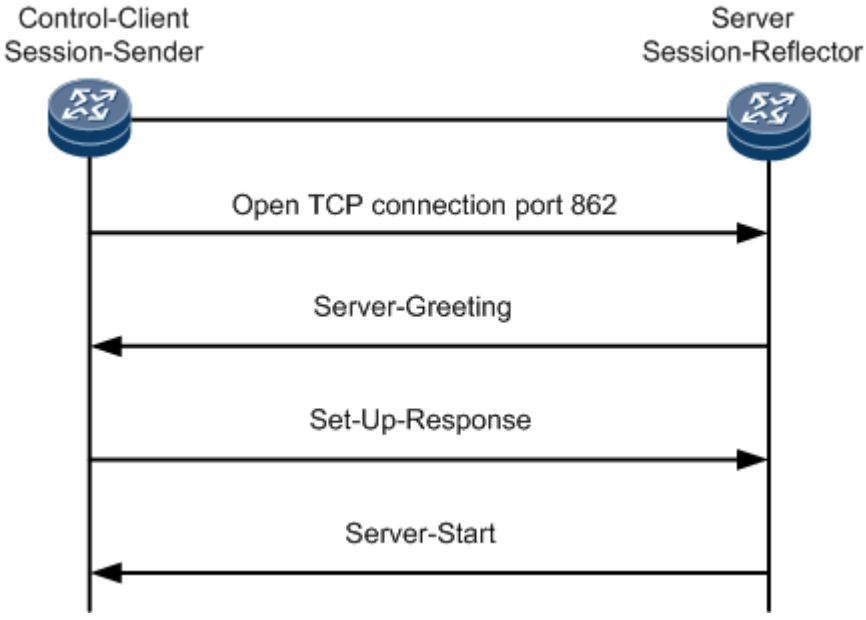
Establishing a Control Session

A control session is established based on a TCP connection for checking control layer validity and verifying parameters. A control session provides a basis for the establishment of a test session. **Figure 11-49** shows how a control session is established.

1. The Server specifies a TCP port ID (the default port ID is 862), and the Control-Client initiates a TCP connection.
2. The Server replies with a Server-Greeting message to notify the Control-Client of the server configurations.
3. After receiving the Server-Greeting message, the Control-Client sends a Set-Up-Response message to the Server to establish a control session.
4. The Server verifies the Set-Up-Response message and replies with a Server-Start message.

A control session is established between the Control-Client and the Server.

Figure 11-49 Establishing a control session



Establishing a Test Session

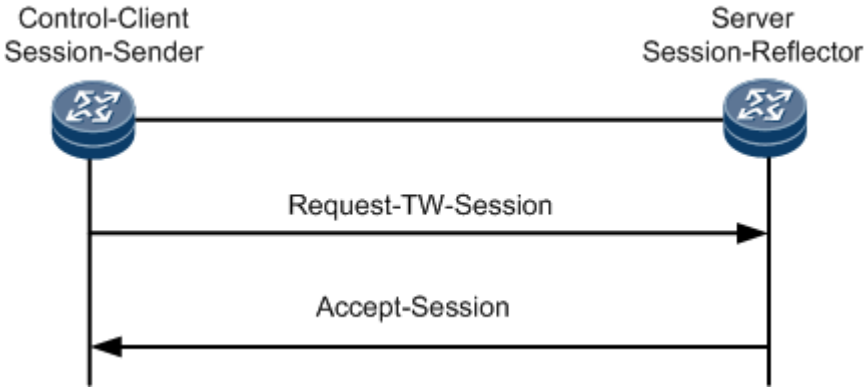
After a control session is established, users can specify an IP address and a UDP port ID on the Control-Client to establish a test session. **Figure 11-50** shows how a test session is established.

 Copyright © Huawei Technologies Co., Ltd.

1. The Control-Client sends a Request-TW-Session message carrying an IP address and a UDP port ID to the Server through the TCP connection.
2. After receiving the Request-TW-Session message, the Server establishes a test session based on the IP address and UDP port ID in the Request-TW-Session message and replies to the Control-Client with an Accept-Session message.

A test session is established between the Control-Client and the Server.

Figure 11-50 Establishing a test session



Starting a Test Session

A test session is started based on a control session. After a test session is established, the Control-Client can send a Start-Session message to the Server. After receiving the Start-Session message, the Server instructs the control session to start all test sessions that are established based on the control session. **Figure 11-51** shows how a test session is started.

1. The Control-Client sends a Start-Session message to the Server.
2. After receiving the Start-Session message, the Server notifies the Session-Reflector of the test session information, so the Session-Reflector can respond to probes.
3. The Session-Reflector replies to the Control-Client with a Start-ACK message to start the test session.
4. After receiving the Start-ACK message, the Control-Client notifies the Session-Sender of the test session information, so the Session-Sender can send probes.

The test session is started, and the Session-Reflector starts to respond to probes.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-51 Starting a test session

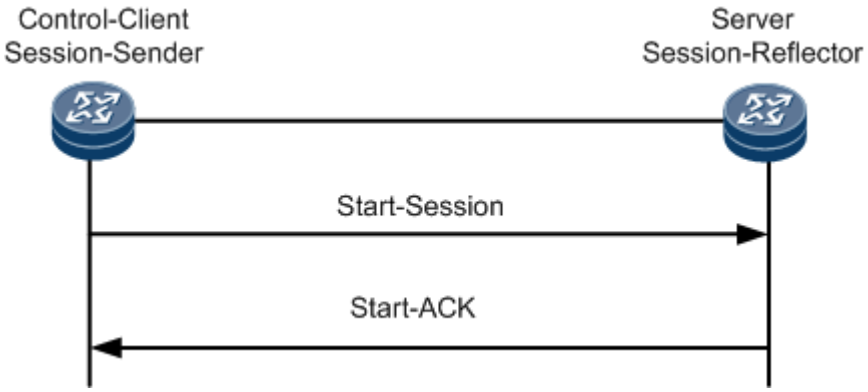
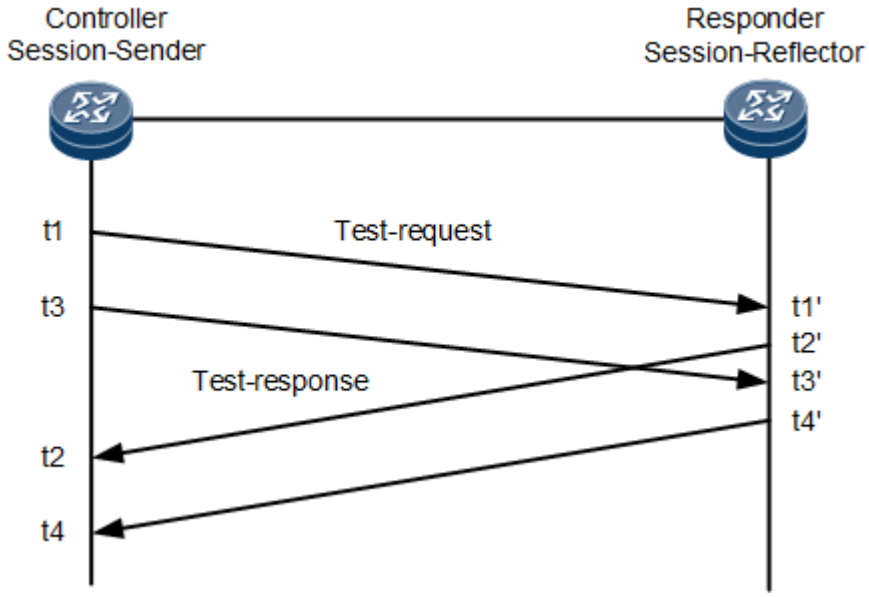


Figure 11-52 TWAMP Light performance measurement



As shown in **Figure 11-52**, after the statistical task is established, the TWAMP-Test protocol functions as the probe for sending and receiving performance measurement, and uses the preset IP address and UDP port ID of the statistical session. The Controller sends TWAMP-Test packets. After receiving the TWAMP-Test packets, the Responder reflects the packets to the Controller. The Controller collects TWAMP measurement statistics. The detailed performance measurement process is as follows:

1. After receiving the TWAMP-Test packet, the Controller calculates the round-trip packet loss rate, delay, and jitter based on the SN and timestamp in the packet.

   Delay

   The delay is calculated based on timestamps. The Controller sends a probe carrying a Tx timestamp t1, and the Responder replies with a response probe carrying an Rx timestamp t1' and a response timestamp t2'. After receiving

 Copyright © Huawei Technologies Co., Ltd.

the response probe, the Controller records the Rx timestamp t2. The delay during a single period is calculated based on 4 timestamps.

Delay1 = t2 - t1- (t2' - t1')

Jitter

The value of this parameter varies depending on the absolute value of the delay data in the adjacent period.

According to the previous delay formula, the delay value of the adjacent period is Delay2 = t4 - t3 - (t4'- t3').

Jitter = | Delay2 — Delay1 |

PLR

The controller sends a probe packet carrying the Tx SN. The Responder does not generate an independent SN. Instead, the Responder copies the Controller SN as the response SN. The final PLR is calculated based on the Rx and Tx packet data.

PLR = |Number of Tx packets - Number of Rx packets|/Number of Tx packets

2. The Controller reports the calculated performance statistics to the NMS. Then, customers can read related statistics on the NMS.

   Statistics calculated in different ways are reported to the NMS in different ways:

   - Continuous measurement: The test result is reported to the NMS through the performance monitoring (PM).
   - On-demand measurement: The test result is reported to the NMS through the MIB.

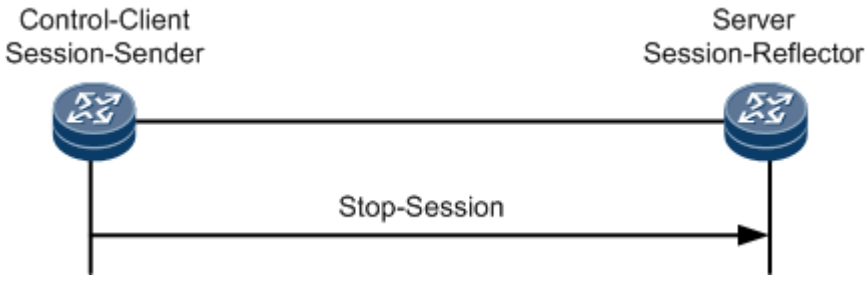
Stopping a Test Session

After statistics are collected, users can stop a test session. **Figure 11-53** shows how a test session is stopped.

1. The Control-Client sends a Stop-Session message to instruct the Server to stop collecting statistics.
2. After receiving the Stop-Session message, the Server disables the Session-Reflector from responding to probes in a test session.

The test session is stopped.

Figure 11-53 Stopping a test session



 Copyright © Huawei Technologies Co., Ltd.

11.10.4 TWAMP for ONT

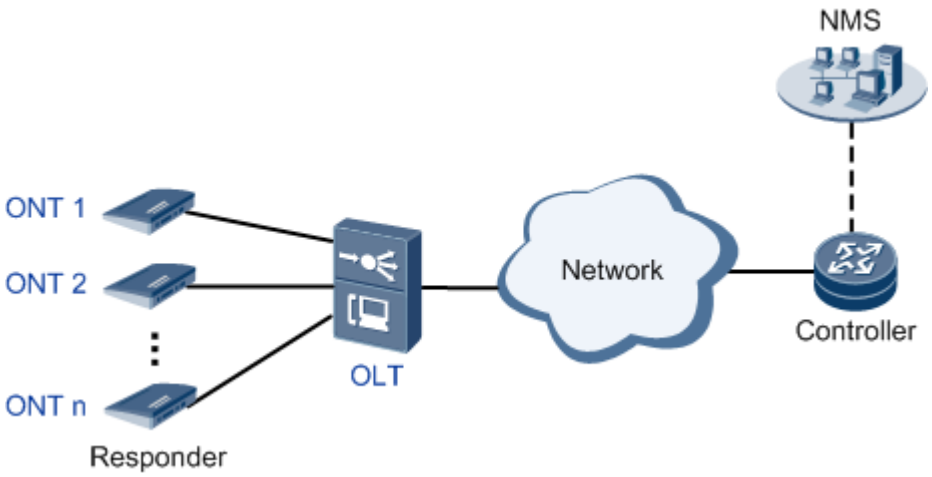
When an ONT connected to an OLT works as a responder, this ONT needs to be managed through the OLT.

Application Scenario

As shown in **Figure 11-54**, the OLT and ONT complete the Two-Way Active Measurement Protocol (TWAMP) test process together. The ONT connected to OLT works as a responder. The OLT implements the configuration functions. Through the configuration of OLT, frequent operations on a large number of ONTs can be avoided.

The TWAMP test service can be quickly restored in xPON line fault scenarios such as intermittent line disconnection, ONT reset, type B single-homing switching, and type B dual-homing switching.

Figure 11-54 TWAMP for ONT network



Logic Model

The mapping relationships between the physical devices and logical entities are shown in **Figure 11-55**

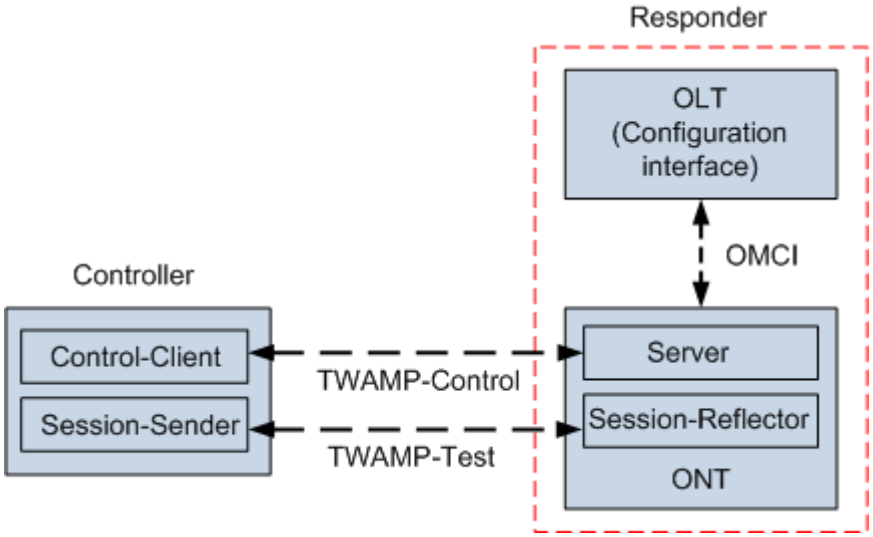
The OLT implements configuration and issues the configured parameters and running status to the ONT through OMCI.

The ONT implements control protocol and test protocol, that is, the ONT processes TWAMP control packets and test packets and interacts with the Controller logically.

Data issued by the OLT to the ONT is static data, and dynamic data is automatically generated and maintained through TWAMP control protocol interaction by the ONT.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-55 TWAMP for ONT logical model



11.10.5 Configuring TWAMP

In the TWAMP standard mode, the test parameters are flexibly negotiated by the control message between the responder and the controller. The test task is dynamically initiated.

11.10.5.1 Configuring TWAMP Responder

The control session parameters for the server role and the test session parameters for the reflector role are configured on the responder.

Prerequisites

Ensure that the Controller and Responder are routable and IP links between them work properly.

The controller creates the statistic session, initiates the TWAMP control session, and starts the test session of responder.

Procedure

Step 1 Run the **sysman server source twamp** command to bind a specified source port to the TWAMP service protocol.

NOTE

In versions earlier than V100R020C00, this step is not required. In V100R020C00 and later versions, this step must be performed to make the TWAMP function take effect.

Step 2 Run the **nqa twamp** command to create a TWAMP service, and enter TWAMP mode.

Step 3 Configure the control session parameters of TWAMP server.

1. Run the **server** command to enable the server function and enter TWAMP server mode.

 Copyright © Huawei Technologies Co., Ltd.

2. (Optional) To ensure compatibility of non-standard client devices, run the **tcp port** *port-number* [**all** | **vpn-instance** *vpn-instance-name*] command to configure a TCP dectecting port ID on the TWAMP server. The default TCP detecting port ID is 862.
3. (Optional) Run the **control-session inactive** *inactive-time* command to configure the inactive timeout duration of a TWAMP control session on the TWAMP server. If a TWAMP control session is inactive within the specified inactive timeout duration, the control session is deleted to avoid occupying system resources.
4. Run the **quit** command to return TWAMP mode.

Step 4 Configure the test session parameters of TWAMP reflector.

1. Run the **reflector** command to enable the TWAMP reflector function and enter TWAMP reflector mode.
2. (Optional) Run the **test-session inactive** *inactive-time* command to configure the inactive timeout duration of TWAMP test sessions on the TWAMP reflector. If a TWAMP test session is inactive within the specified inactive timeout duration, the test session is stopped.
3. Run the **quit** command to return TWAMP mode.

Step 5 Query the real-time statistics about TWAMP control session and TWAMP test session on a responder.

- Run the **display twamp control-session** command to query the information about control sessions on the TWAMP server.
- Run the **display twamp test-session** command to query the information about test sessions on the TWAMP reflector.
- Run the **display twamp global-info** command to query the global information about TWAMP server and reflector.

----End

Example

VLAN 10 and VLANIF 10 have been created. To set the TCP detecting port ID to 2000, the inactive timeout duration of a TWAMP control session to 600s, and the inactive timeout duration of a TWAMP test session to 600s, do as follows:

```
huawei(config)#sysman server source twamp vlanif 10    //V100R020C00 and later versions
huawei(config)#nqa twamp
huawei(config-twamp)#server
huawei(config-twamp-srv)#tcp port 2000
huawei(config-twamp-srv)#control-session inactive 600
huawei(config-twamp-srv)#quit
huawei(config-twamp)#reflector
huawei(config-twamp-ref)#test-session inactive 600
huawei(config-twamp-ref)#quit
```

11.10.5.2 Configuring TWAMP for ONT

Configuring the parameters through the MA5800 makes the ONT reflect the TWAMP test packet as the responder side.

 Copyright © Huawei Technologies Co., Ltd.

Prerequisites

Ensure that the Controller and Responder are routable and IP links between them work properly.

The controller creates the statistic session, initiates the TWAMP control session, and starts the test session of responder.

Procedure

Step 1 Run the **sysman server source** command to bind a specified source port to the TWAMP service protocol.

NOTE

In versions earlier than V100R020C00, this step is not required. In V100R020C00 and later versions, this step must be performed to make the TWAMP function take effect.

Step 2 Run the **nqa twamp ont** command to enable the TWAMP function of an ONT and enter TWAMP ONT mode.

Step 3 Run the **ont** [**eth** *ont-portid*] { **untag** | **vlan** *vlanid* [*priority1*] | **outer-vlan** *outer-vlanid* [*priority2*] **inner-vlan** *inner-vlanid* [*priority3*] } **ipaddress** *ipaddr* { *mask-length* | *mask-ipaddr* } **gateway** *gateway* command to configure the TWAMP parameters of an ONT.

Before performing a TWAMP test on an ONT, run the **ont** command to configure the TWAMP parameters of the ONT. Otherwise, the test cannot be performed.

Step 4 Configure the control session parameters of TWAMP Server on an ONT

1. Run the **server** command to enable the TWAMP server function on an ONT and enter TWAMP ONT server mode.
2. Run the **client permit source** command to configure the range of IP addresses of clients that are allowed to access the TWAMP server of an ONT. By default, the range of IP addresses of clients that are allowed to access the TWAMP server is not configured. In this case, clients with any IP addresses are not allowed to access the TWAMP server.
3. (Optional) To ensure compatibility of non-standard client devices, run the **tcp port** *port-number* [**all** | **vpn-instance** *vpn-instance-name*] command to configure a TCP detecting port ID on the TWAMP server on an ONT. The default TCP detecting port ID is 862.
4. (Optional) Run the **control-session inactive** *inactive-time* command to configure the inactive timeout duration of a TWAMP control session on the TWAMP server on an ONT. If a TWAMP control session is inactive within the specified inactive timeout duration, the control session is deleted to avoid occupying system resources.
5. Run the **quit** command to return TWAMP ONT mode.

Step 5 Configure the test session parameters of TWAMP reflector on an ONT.

1. Run the **reflector** command to enable the TWAMP reflector function on an ONT and enter TWAMP ONT reflector mode.
2. (Optional) Run the **test-session inactive** *inactive-time* command to configure the inactive timeout duration of TWAMP test sessions on the TWAMP reflector on an ONT. If a TWAMP test session is inactive within the specified inactive timeout duration, the test session is stopped.

 Copyright © Huawei Technologies Co., Ltd.

3. Run the **quit** command to return TWAMP ONT mode.

Step 6 Query the real-time statistics about TWAMP control session and TWAMP test session on an ONT responder.

- Run the **display twamp ont control-session** command to query the information about control sessions on the TWAMP server on an ONT.
- Run the **display twamp ont test-session** command to query the information about test sessions on the TWAMP reflector on an ONT.
- Run the **display twamp ont global-info** command to query the global information about TWAMP server and reflector on an ONT.

----End

Example

VLAN 10 and VLANIF 10 have been created. Enable the TWAMP function for ONT 1 connected to port 0/3/0. Configure the ONT to use the NNI port for the test. The packets are encapsulated with one VLAN tag, the VLAN ID is 10, the IP address is 192.168.1.2, the subnet mask length is 24 bits, and the gateway address is 192.168.1.1. The IP address of the client that is allowed to access is 10.10.10.2.

- On the TWAMP Responder of the ONT, set the TCP detecting port ID to 2000, the inactive timeout duration of a TWAMP control session to 600s.
- On the TWAMP Responder of the ONT, set the inactive timeout duration of the TWAMP test session to 600s.

```
huawei(config)#sysman server source twamp vlanif 10    //V100R020C00 and later versions
huawei(config)#nqa twamp ont 0/3/0 1
huawei(config-twamp-ont-0/3/0/1)#ont vlan 10 ipaddress 192.168.1.2 24 gateway 192.168.1.1
huawei(config-twamp-ont-0/3/0/1)#server
huawei(config-twamp-ont-0/3/0/1-srv)#tcp port 2000
huawei(config-twamp-ont-0/3/0/1-srv)#control-session inactive 600
huawei(config-twamp-ont-0/3/0/1-srv)#quit
huawei(config-twamp-ont-0/3/0/1)#reflector
huawei(config-twamp-ont-0/3/0/1-ref)#test-session inactive 600
```

11.10.6 TWAMP Reference Standards and Protocols

The following table lists the references.

Document	Document Name	Protocol Compliance
RFC 5357	Two-Way Active Measurement Protocol (TWAMP)	Partially compliant: • Supports the TWAMP Responder only. • Supports unauthenticated TWAMP test packets only. Does not support authentication and encryption TWAMP test packets.

11.11 TWAMP Light

Two-Way Active Measurement Protocol (TWAMP) Light is a light version of TWAMP, which is defined in RFC 5357. TWAMP Light measures the round-trip

 Copyright © Huawei Technologies Co., Ltd.

performance of an IP network by using simplified control protocols to establish test sessions.

11.11.1 TWAMP Light Introduction

This topic describes the definition and usage scenario of Two-Way Active Measurement Protocol (TWAMP) Light.

Definition

TWAMP Light is a light version of TWAMP, which is defined in RFC 5357. TWAMP Light measures the round-trip performance of an IP network by using simplified control protocols to establish test sessions.

Purpose

On conventional IP networks, a universal tool that can rapidly provide statistics about IP network performance is vital for carriers to perform operation, administration, and maintenance (OAM). Currently, Network Quality Analysis (NQA) and IP Flow Performance Measurement (IP FPM) are the tools most often used for this purpose. However, NQA requires complex deployment and does not support intercommunication between Huawei and non-Huawei devices, and IP FPM has high requirements on network devices and is applicable only to a few scenarios. To resolve this problem, the Internet Engineering Task Force IP performance monitoring (IETF IPPM) group has defined a set of protocols, including TWAMP.

TWAMP measures the round-trip performance of an IP network. It includes two versions: TWAMP Standard and TWAMP Light.

- TWAMP Standard: Involves control packet exchange and test packet exchange. In this version, two devices can flexibly negotiate test parameters through control packets and dynamically initiate test tasks. However, protocol implementation is complex.
- TWAMP Light: Involves only test packet exchange (TWAMP Light does not involve control packet exchange). In this version, test parameters are manually configured on two devices and test tasks are manually issued. However, protocol implementation is simple. TWAMP Light simplifies the working process of protocols and is easier to deploy in actual application.

11.11.2 TWAMP Light Basic Concepts

This topic describes the basic concepts of device roles and communication models of TWAMP, and the differences between TWAMP Standard and TWAMP Light.

Device Roles

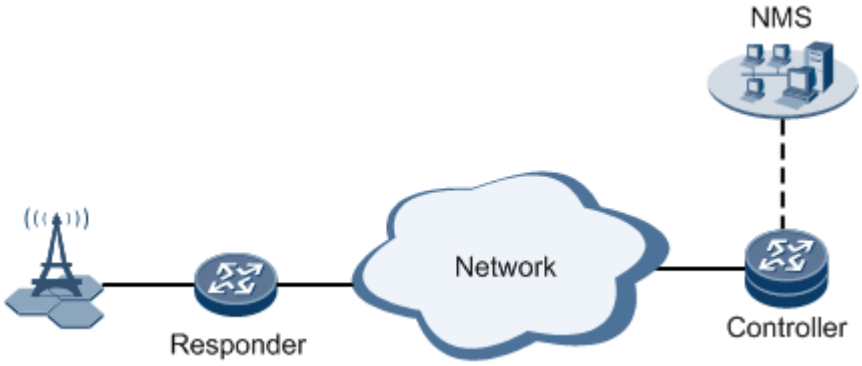
TWAMP involves two physical devices: a Controller and a Responder. As shown in **Figure 11-56**, a device or tester works as the Controller (initiator), and the MA5800 works as the Responder (sender).

 Copyright © Huawei Technologies Co., Ltd.

NOTE

The MA5800 works as the Responder only in TWAMP Light.

Figure 11-56 Networking of TWAMP measurement



The Controller and Responder function as follows:

- Controller: Serves as the initiator of a TWAMP test session, which sends and receives packets over a test session, collects and calculates performance statistics, and reports these statistics to the NMS.
- Responder: Serves as the receiver of a TWAMP test session, which responds to packets received over a test session. It does not calculate or report performance statistics.

Communication Models

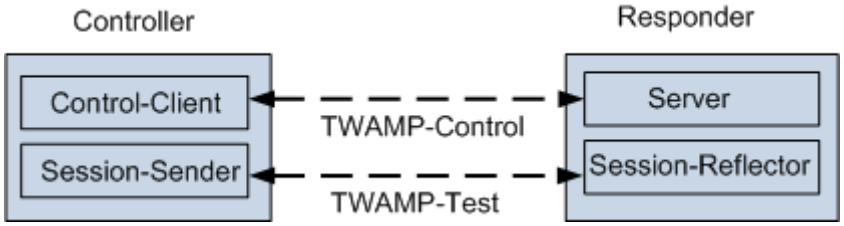
For details on the communication models and logical entities, see **Communication Model** in "TWAMP Basic Concepts."

In the TWAMP Standard and TWAMP Light architectures, physical devices and logical entities have different mapping relationships.

- TWAMP Standard

  Mapping relationships between physical devices and logical entities in TWAMP Standard architecture are shown in **Figure 11-57**. Specifically, Controller contains Control-Client and Session-Sender; Responder contains Server and Session-Reflector.

  Figure 11-57 TWAMP Standard architecture

- TWAMP Light

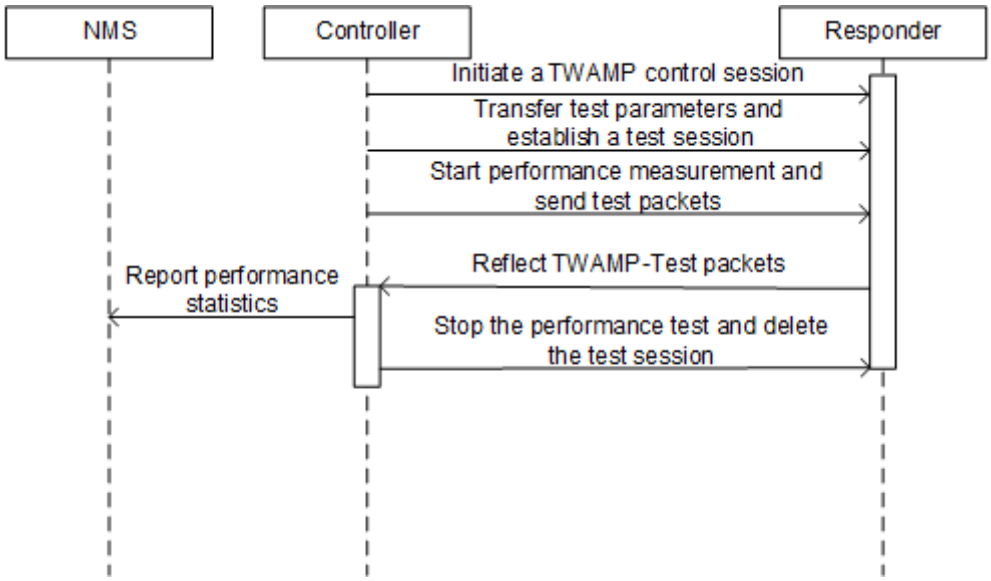
  Mapping relationships between physical devices and logical entities in TWAMP Light architecture are shown in **Figure 11-58**. Specifically, Controller

 Copyright © Huawei Technologies Co., Ltd.

contains Control-Client, Session-Sender, and Server; Responder contains only Session-Reflector.

Unlike TWAMP Standard, TWAMP Light has parameters statically configured for a test session. You can configure the IP address and UDP port ID on the Responder by running commands. After a test session is created, the performance statistics, such as the packet loss rate, delay, and jitter can be calculated based on the exchange of TWAMP-Test packets. No control session (required in TWAMP Standard) needs to be set up during performance detection. TWAMP Light does not need any control protocol for parameter negotiation. It simplifies the communication model of TWAMP and greatly reduces requirements on the Responder performance, allowing the Responder to be rapidly deployed. In addition, TWAMP Light supports plug-and-play.

Figure 11-58 TWAMP Light architecture



11.11.3 TWAMP Light Principles

This topic describes the principles and implementation process of Two-Way Active Measurement Protocol (TWAMP) Light.

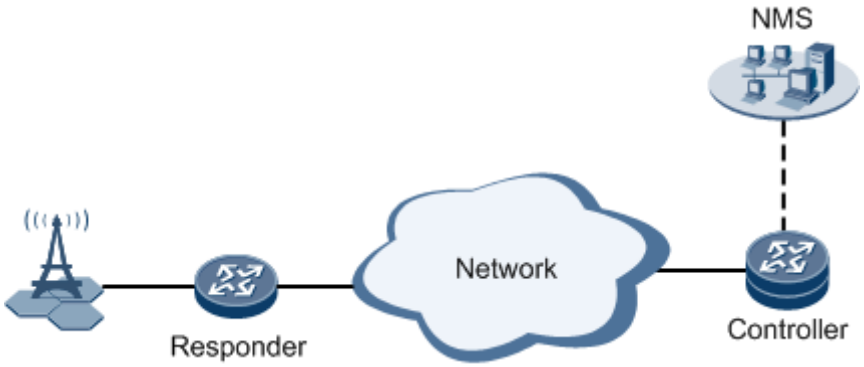
TWAMP Light Principles

A TWAMP Light service must be established before TWAMP Light is implemented.

1. **Establishing a TWAMP Light service**

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-59 Establishing a TWAMP Light service



The Controller and Responder are deployed as shown in **Figure 11-59**.

a. The Control-Client on the Controller creates a TWAMP Light test session.
b. The Responder is specified as the Session-Reflector.
c. The Session-Sender on the Controller starts the test session, and the Controller sends TWAMP-Test packets to the Responder based on the configured packet sending rate and packet profile.
d. The Responder responds to the TWAMP-Test packets.

2. **Measuring performance**

   TWAMP Light supports both on-demand measurement and proactive measurement.

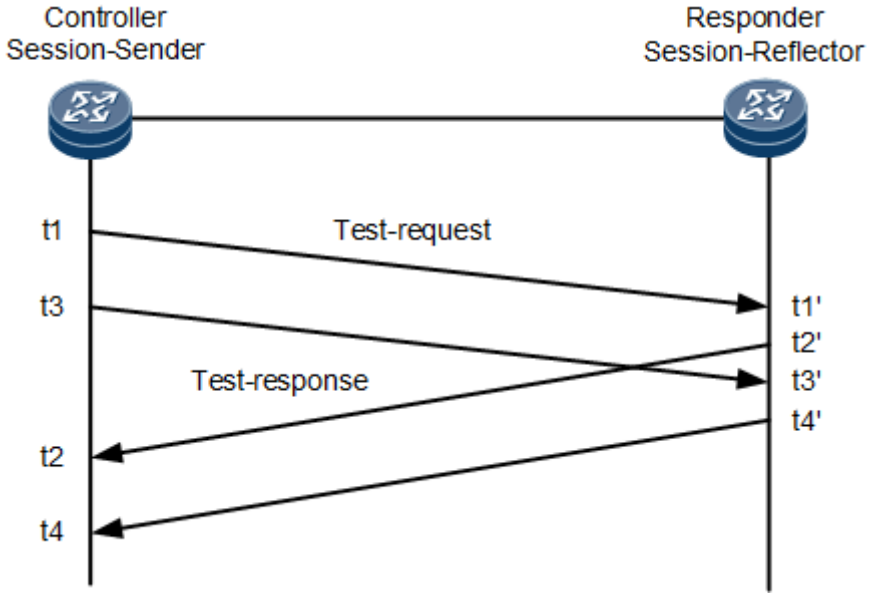
   – On-demand measurement works within a specified period after being started. It can be performed once or periodically within the specified period.
   – Proactive measurement works continuously after being started to collect statistics.

   TWAMP Light defines two types of TWAMP-Test packets: Test-request packets and Test-response packets.

   – Test-request packets are sent from the Controller to the Responder.
   – Test-response packets are sent as replies from the Responder to the Controller.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-60 TWAMP Light implementation



In **Figure 11-60**, TWAMP-Test packets function as probes and carry the IP address and UDP port ID that are predefined for the test session between the Controller and Responder. The Controller sends a TWAMP-Test packet to the Responder, and the Responder replies to it. The Controller collects TWAMP statistics as follows:

a. The Controller collects statistics about the round—trip delay, jitter, and packet loss rate based on the sequence numbers and timestamps carried in TWAMP-Test packets.

   Delay

   The delay is calculated based on timestamps. The Controller sends a probe carrying a sending timestamp t1, and the Responder replies with a response probe carrying a receiving timestamp t1' and a responding timestamp t2'. After receiving the response probe, the Controller records the receiving timestamp t2. The delay during a single period is calculated based on the four timestamps.

   Delay1 = t2 - t1- (t2' - t1')

   Jitter

   The jitter is calculated based on two consecutive delays.

   Based on the preceding delay formula, the delay of the neighboring measurement interval can be calculated as follows: Delay2 = t4 - t3 - (t4' - t3')

   Jitter = | Delay2 — Delay1 |

   Packet loss rate

   The packet loss rate is calculated based on the sequence numbers carried in probes. The Controller sends a probe with a sequence number, and the Responder replies with a response probe with the same sequence number. The packet loss rate is calculated as follows:

   Packet loss rate = |Number of sent packets - Number of received packets|/Number of sent packets

 Copyright © Huawei Technologies Co., Ltd.

b. The Controller reports the calculated performance statistics to the NMS. Then, customers can read related statistics on the NMS.

Statistics calculated in different ways are reported to the NMS in different ways:

- Continuous measurement: The test result is reported to the NMS through the performance monitoring (PM).
- On-demand measurement: The test result is reported to the NMS through the MIB.

11.11.4 Configuring TWAMP Light Responder

Users can create a test session on the Two-Way Active Measurement Protocol (TWAMP) Light Responder for responding to TWAMP-Test packets.

Prerequisites

Ensure that the controller and responder are routable and IP links between them work properly.

Procedure

Step 1 Run the **sysman server source** command to bind a specified source port to the TWAMP service protocol.

NOTE

In versions earlier than V100R020C00, this step is not required. In V100R020C00 and later versions, this step must be performed to make the TWAMP function take effect.

Step 2 Run the **nqa twamp-light** command to create a TWAMP Light service, and enter TWAMP Light mode.

Step 3 Run the **responder** command to enable the TWAMP Light responder function, and enter TWAMP Light responder mode.

Step 4 Run the **test-session** *session-id* **local-ip** *local-ip-address* **remote-ip** *remote-ip-address* **local-port** *local-port-number* **remote-port** *remote-port-number* [**vpn-instance** *vpn-instance-name*] [**description** *description-value*] command to create a session on the responder of a TWAMP Light service.

Step 5 Run the **display twamp-light responder test-session** command to query the real-time session statistics of a TWAMP Light responder.

----End

Example

VLAN 10 and VLANIF 10 have been created. Assume that the session ID is 1, IP address of the responder is 10.1.1.1, IP address of the sender is 10.2.2.2, UDP port IDs of the responder and sender are 3000, and the VPN instance name is vpn1. To create a session on a responder of the TWAMP Light service with the preceding parameters, do as follows:

```
huawei(config)#sysman server source twamp vlanif 10    //V100R020C00 and later versions
huawei(config)#nqa twamp-light
```

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config-twamp-light)#responder
huawei(config-twamp-light-responder)#test-session 1 local-ip 10.1.1.1 remote-ip 10.2.2.2 local-port
3000 remote-port 3000 vpn-instance vpn1
huawei(config-twamp-light-responder)#quit
huawei(config-twamp-light)#quit
huawei(config)#display twamp-light responder test-session
  Total number : 1
  ------------------------------------------------------------------------
  ID     Local-IP        Local-Port     Remote-IP       Remote-Port
  ------------------------------------------------------------------------
   1     10.1.1.1           3000        10.2.2.2           3000
  ------------------------------------------------------------------------
```

11.11.5 TWAMP Light Reference Standards and Protocols

The following table lists the references.

Document	Document Name	Protocol Compliance
RFC 5357	Two-Way Active Measurement Protocol (TWAMP)	Partially compliant: • Supports the TWAMP Light Responder only. • Supports unauthenticated TWAMP test packets only. Does not support authentication and encryption TWAMP test packets.

11.12 Routing

Routing is a common term used for describing the path through which the packets from a host in a network travel to a host in another network.

11.12.1 IP Routing Overview

11.12.1.1 Introduction to Routing

Definition

Routing is a common term used for describing the path through which the packets from a host in a network travel to a host in another network.

Routers send packets on the Internet. A router selects a suitable path in a network according to the destination address included in a received packet, and sends the packet to the next router on the path. In this way, the packet travels over the Internet Until it reaches the destination host.

Purpose

The access equipment, serving as a basic element in the entire telecom network, must support the functions of remote operation, management and maintenance on the equipment itself.

 Copyright © Huawei Technologies Co., Ltd.

With the development of small-size access equipment that can be managed remotely, the access equipment needs to feature the functions of a BRAS, such as allocation of network addresses and user management. In this way, the access equipment must support the routing feature.

A MA5800 can also serve as a router.

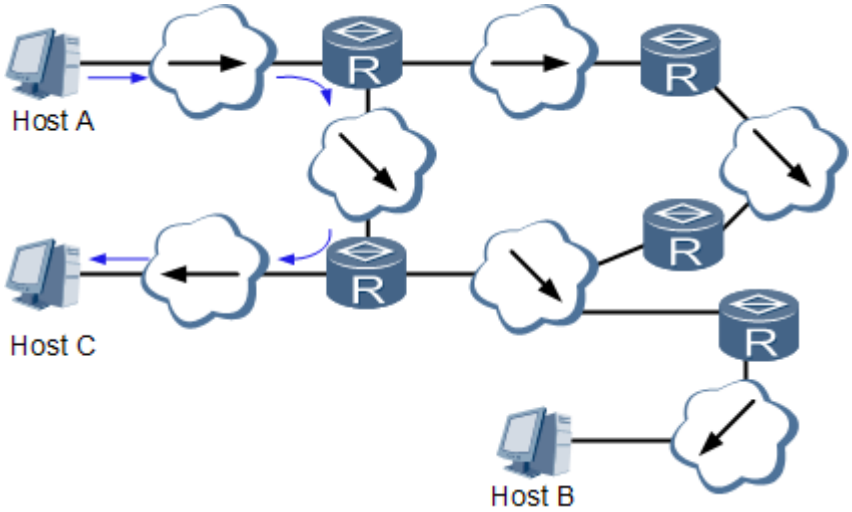
11.12.1.2 Routers

On the Internet, network connection devices control network traffic and ensure data transmission quality on networks. Common network connection devices include hubs, bridges, switches, and routers.

As a standard network connection device, a router is used to select routes and forward packets. Based on the destination address in the received packet, a router selects a path to send the packet to the next router. The last router is responsible for sending the packet to the destination host. In addition, a router can select an optimal path for data transmission.

For example, in **Figure 11-61**, traffic from Host A to Host C needs to pass through three networks and two routers. The hop count from a router to its directly connected network is zero. The hop count from a router to a network that the router can reach through another router is one. The rest can be deduced by analogy. If a router is connected to another router through a network, a network segment exists between the two routers, and they are considered adjacent on the Internet. In the figure, the arrows indicate network segments. The routers do not need to know about the physical link composition of each network segment.

Figure 11-61 Network segment and hop count



Network sizes may vary greatly, and the actual lengths of network segments vary as well. Therefore, you can set a weighted coefficient for the network segments of each network and then measure the cost of a route based on the number of network segments.

A route with the minimal network segments is not necessarily optimal. For example, a route passing through three high-speed Local Area Network (LAN) network segments may be a better choice than one passing through two low-speed Wide Area Network (WAN) network segments.

 Copyright © Huawei Technologies Co., Ltd.

11.12.1.3 Routing Table and FIB Table

Each router maintains one routing table and one FIB table at least. A router uses the routing table to select routes, and uses the FIB table to guide packet forwarding. In Equal and Weighted Cost Multi-Path (ECMP), according to different states of the network, the traffic to the same destination network can be distributed among multiple equal-cost paths to reduce the network load or to implement the link backup function.

- Routes discovered by the various routing protocols are stored in the routing table. The routes in the routing table are divided, according to their sources, into the following types:
  - Directly connected route or interface route: is the route discovered by the link layer protocols.
  - Static route: is the route manually configured by the network administrator.
  - Dynamic route: is the route discovered by dynamic routing protocols.
- Each entry in the FIB table contains the physical or logical interface through which a packet is sent to a network segment or host to reach the next router. An entry also indicates whether the packet can be sent directly to a destination host in a directly connected network.

Routing table

Each routing entry in the routing table specifies the physical port through which a packet is sent to a subnet or a host, to the next router on the path, or to the destination host on the directly connected network without passing through another router.

Each router maintains the protocol routing table for each type of protocol and a local core routing table (or routing management table).

- Protocol routing table

  A protocol routing table stores the routing information discovered by the protocol. A routing protocol can import and advertise the routes that are discovered by other protocols. For example, if a router that runs the Open Shortest Path First (OSPF) protocol needs to use OSPF to advertise direct routes, static routes, or Intermediate System-Intermediate System (IS-IS) routes, the router must import the routes into the OSPF routing table.
- Local core routing table

  A router uses the local core routing table to store protocol routes and preferred routes. The router then delivers the preferred routes to the FIB table for forwarding packets. The router selects routes according to the priorities of protocols and costs stored in the routing table. You can run the **display ip routing-table** command to view the local core routing table of a router.

  A router maintains a local core routing table for each Virtual route forwarding instance (VRF).

The key entries of the routing table are shown in **Table 11-12**.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-12 Key entries of the routing table

Entry	Description
Destination	The destination address is a 32-bit character that labels the destination IP address or destination network of an IP packet.
Mask	The mask is used with the destination address to identify the subnet address of the destination host or router. The network mask is composed of several consecutive 1s. These 1s can be expressed in either the dotted decimal notation or the number of consecutive 1s in the mask. For example, the network mask can be expressed either as 255.255.255.0 or 24. The network address of the destination host or router is obtained through the "AND" operation on the destination address and network mask. For example, if the destination address is 10.1.1.1 and the mask is 255.255.255.0, the address of the network where the host or router resides is 10.1.1.0.
Proto	Indicates the protocol through which routes are learned.
Pre	Indicates the preference added to the IP routing table for a route. To the same destination, multiple routes with different next hops and outgoing interfaces exist. The routes in the table are those discovered by different routing protocols or tare the manually configured static routes. The router selects the route with the highest preference (the smallest value) as the optimal route. For more information on the preference of each protocol, see **Table 11-15** in the Route Protocols.
Cost	Indicates the route cost. When multiple routes to the same destination have the same preference, the route with the lowest cost is selected as the optimal route. **NOTE** The Preference value is used to compare the preferences of various routing protocols, while the Cost value is used to compare the preferences of different routes of the same routing protocol.
NextHop	Indicates the IP address of the next device that an IP packet passes through.
Interface	Indicates the outgoing interface through which an IP packet is forwarded.

Based on the destination, routes can be classified as:

- Subnet route: Its destination is a subnet.
- Host route: Its destination is a host.

Based on the connection between the destination and the router, routes can be classified as:

- Direct route: Its destination network is directly connected to the router.

 Copyright © Huawei Technologies Co., Ltd.

- Indirect route: Its destination network is not directly connected to the router.

To avoid large routing tables, a default route can be assigned. Once a packet fails to find a dedicated route in the routing table, the default route is selected for forwarding the packet.

In the networks shown in **Figure 11-62**, the digits in each network represent the IP address of the network. Router 8 is connected to three networks. Therefore, it has three IP addresses and three physical ports. **Figure 11-62** is its routing table.

Figure 11-62 Interconnected networks

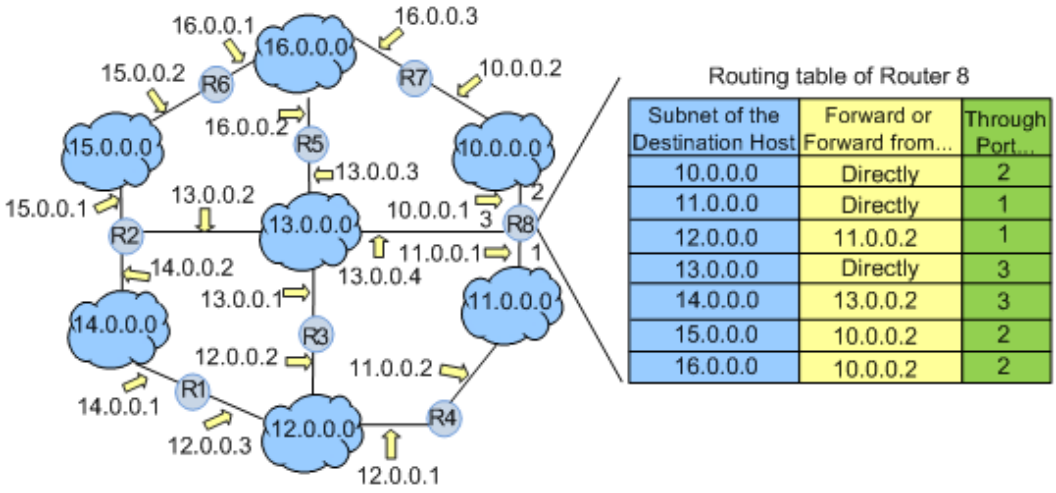


Table 11-13 shows the routing table of router 8.

Table 11-13 Routing table of Router 8

Subnet of the Destination Host	Forward or Forward from...	Through Port...
10.0.0.0	Directly	2
11.0.0.0	Directly	1
12.0.0.0	11.0.0.2	1
13.0.0.0	Directly	3
14.0.0.0	13.0.0.2	3
15.0.0.0	10.0.0.2	2
16.0.0.0	10.0.0.2	2

FIB Table

After route selection, routers send the active routes in the routing table to the FIB table. When a router receives a packet, the router searches the FIB table for the optimal route to forward the packet.

 Copyright © Huawei Technologies Co., Ltd.

The router performs the "AND" operation on the destination address in the packet and the network mask of each entry in the FIB table. The router then compares the result of the "AND" operation with the entries in the FIB table to find a match. The router chooses the optimal route to forward packets according to the best or "longest" match.

As an example, a certain router has the following brief routing table:

```
Routing Tables:
Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface
        0.0.0.0/0   Static  60   0           D   120.0.0.2       Vlanif10
        8.0.0.0/8   RIP     100  3           D   120.0.0.2       Vlanif10
       10.0.0.0/8   OSPF    10   50          D   20.0.0.2        Vlanif10
      10.1.0.0/16   RIP     100  4           D   120.0.0.2       Vlanif20
      20.0.0.0/8    Direct  0    0           D   20.0.0.1        Vlanif20
```

After receiving a packet that carries the destination address 10.1.2.1, the router searches the following table:

```
 FIB Table:
 Total number of Routes : 5
Destination/Mask    Nexthop         Flag TimeStamp     Interface       TunnelID
0.0.0.0/0           120.0.0.2       SU   t[37]         Vlanif10        0x0
8.0.0.0/8           120.0.0.2       DU   t[37]         Vlanif10        0x0
10.0.0.0/8          20.0.0.2        DU   t[9992]       Vlanif10        0x0
10.1.0.0/16         120.0.0.2       DU   t[9992]       Vlanif20        0x0
20.0.0.0/8          20.0.0.1        U    t[9992]       Vlanif20        0x0
```

First, AND operation is performed on the destination address 10.1.2.1 and the masks 0, 8, and 16 to obtain the network segment addresses: 0.0.0.0/0, 10.0.0.0/8, and 10.1.0.0/16. The three addresses match the three entries in the FIB table, that is, 0.0.0.0/0 (0 bits matching), 10.0.0.0/8 (8 bits matching), and 10.1.0.0/16 (16 bits matching). According to the longest match, the VRP selects the entry 10.1.0.0/16 and forwards the packet with its destination address as 10.1.2.1 through VLANIF 20.

11.12.1.4 Routing Protocols

The MA5800 supports the configuration of static routes and the dynamic routing protocols such as RIP, OSPF, IS-IS, and BGP. The MA5800 manages the static and dynamic routes in a unified manner. The static routes and the routes discovered by the routing protocols can be shared.

- Static routes can be easily configured on a system and have lower system requirements. Static routes are applicable to simple, stable, and small-scale networks. Static routes, however, cannot automatically adapt to changes in the network topology, so they must be manually configured.
- On the other hand, dynamic routing protocols use routing algorithms to automatically adapt to changes in network topology. Dynamic routes are applicable to the network that is equipped with Layer 3 devices. The dynamic route configuration, however, has a higher requirement (such as large memory capacity) for system performance and occupies more network resources.

Classification of Dynamic Routing Protocols

Dynamic routing protocols are classified according to the following factors:

Routing protocols are classified according to the application range:

 Copyright © Huawei Technologies Co., Ltd.

- Interior Gateway Protocol (IGP): runs inside an AS, such as RIP, OSPF, and IS-IS.

  Table 11-14 Differences among the three typical IGPs

Item	RIP	OSPF	IS-IS
Protocol type	IP layer protocol	IP layer protocol	Link layer protocol
Application scope	Applies to small networks with simple architectures, such as campus networks.	Applies to medium-sized networks, such as enterprise networks.	Applies to large networks, such as Internet service provider (ISP) networks.
Routing algorithm	Uses a distance-vector algorithm and exchanges routing information over the User Datagram Protocol (UDP).	Uses the shortest path first (SPF) algorithm to generate a shortest path tree (SPT) based on the network topology, calculates shortest paths to all destinations, and exchanges routing information over IP.	Uses the SPF algorithm to generate an SPT based on the network topology, calculates shortest paths to all destinations, and exchanges routing information over IP. In IS-IS, the SPF algorithm runs separately in Level-1 and Level-2 databases.
Route convergence speed	Slow	Less than 1 second	Less than 1 second
Scalability	Not supported	Supported by partitioning a network into areas	Supported by defining levels

- Exterior Gateway Protocol (EGP): runs between different ASs, such as BGP.

Routing protocols are classified according to the type of algorithm they use:

- Distance-Vector Routing Protocol: includes RIP and BGP (BGP is also called Path-Vector).
- Link-State Routing Protocol: includes OSPF and IS-IS.

Routing Protocols and Routing Preference

Routing protocols (including the static route) can learn different routes to the same destination, but not all routes are optimal. Only one routing protocol at one time determines the optimal route to a destination. To select the optimal route,

 Copyright © Huawei Technologies Co., Ltd.

each routing protocol (including the static route) is configured with a preference (the smaller the value, the higher the preference). When multiple routing information sources coexist, the route with the highest preference is selected as the optimal route (the smaller the value is, the higher the preference is).

Table 11-15 lists various routing protocols and the default priorities of the routes discovered by them. "0" indicates the direct route, and "255" indicates any route from an untrusted source.

Table 11-15 Default priorities of routing protocols

Routing Protocol or Route Type	Default Routing Preference
DIRECT	0
OSPF	10
IS-IS	15
INTERNAL EIGRP	50
STATIC	60
RIP	100
OSPF ASE (AS-External)	150
OSPF NSSA (Not-So-Stubby Area)	150
EXTERNAL EIGRP	160
IBGP	255
EBGP	255
UNKNOWN	255

You can define the priorities for all dynamic routing protocols except the direct route (DIRECT) and the BGP (IBGP, EBGP). In addition, the priorities of any two static routes can be different.

If different routing protocols are configured with the same preference, the system determines which routes discovered by these routing protocols become the preferred routes through an internal preference. Table 11-16 shows the internal preferences of routing protocols.

Table 11-16 Internal preferences of routing protocols

Routing Protocol or Route Type	Internal Routing Preference
DIRECT	0
OSPF	10
IS-IS Level-1	15

 Copyright © Huawei Technologies Co., Ltd.

Routing Protocol or Route Type	Internal Routing Preference
IS-IS Level-2	18
STATIC	60
RIP	100
OSPF ASE (AS-External)	150
OSPF NSSA (Not-So-Stubby Area)	150
IBGP	200
EBGP	20

For example, two routes, an OSPF route and a static route, can reach the destination 10.1.1.0/24, and the preferences of both routes are set to 5. In this case, the VRP determines the optimal route according to the internal preferences listed in **Table 11-15**. The internal preference value 10 of OSPF is higher than the internal preference value 60 of the static route. Therefore, the system selects the route discovered by OSPF as the optimal route.

Route Sharing through Route Policy

Different routing protocols can find different routes as they use different algorithms. A routing protocol might need to import routes discovered by other protocols to diversify its own routes. The MA5800 supports importing the routes discovered by one protocol to another protocol. Each protocol has its own route importing mechanism.

However, a protocol only needs to import qualified routes by setting attributes of the routes to be imported. To realize a route policy, you must define the attributes of the routes to which the route policy is to be applied, such as the destination address, and the address of the router distributing routes. You can define the matching rules in advance so that they can be applied in a route policy for route distribution, reception and importing.

The two applications of the routing policy are as follows:

- When importing routes discovered by other protocols, a routing protocol can apply this filter to obtain the required routes.
- When transmitting or receiving routes, a routing protocol can apply the filter so that only the required routes are transmitted or received.

11.12.1.5 Priority-based Route Convergence

Definition

Priority-based route convergence is an important technology that improves network reliability. It provides faster route convergence for key services. For example, to minimize the interruption of key services in case of network faults, real-time multicast services require that the routes to the multicast source quickly

 Copyright © Huawei Technologies Co., Ltd.

converge, and the Multiprotocol Label Switching (MPLS) VPN bearer network requires that routes between PEs also quickly converge.

Convergence priorities provide references for the system to converge routes for service forwarding. Different routes can be set with different convergence priorities, which can be identified as critical, high, medium, and low listed in descending order.

Purpose

With the integration of network services, requirements on service differentiation increase. Carriers require that the routes for key services, such as Voice over IP (VoIP) and video conferencing services converge faster than those for common services. Therefore, routes need to converge based on their convergence priorities to improve network reliability.

Route Convergence Priority

Table 11-17 lists the default convergence priorities of public network routes. You can set convergence priorities for routes based on the requirements of a live network.

Table 11-17 Default convergence priorities of public network routes

Routing Protocol or Route Type	Convergence Priority
DIRECT	critical
STATIC	medium
32-bit host routes of OSPF and IS-IS	medium
OSPF route (except 32-bit host routes)	low
IS-IS route (except 32-bit host routes)	low
RIP	low
BGP	low

NOTE

For VPN route priorities, only 32-bit host routes of OSPF and IS-IS are identified as medium, and the other routes are identified as low.

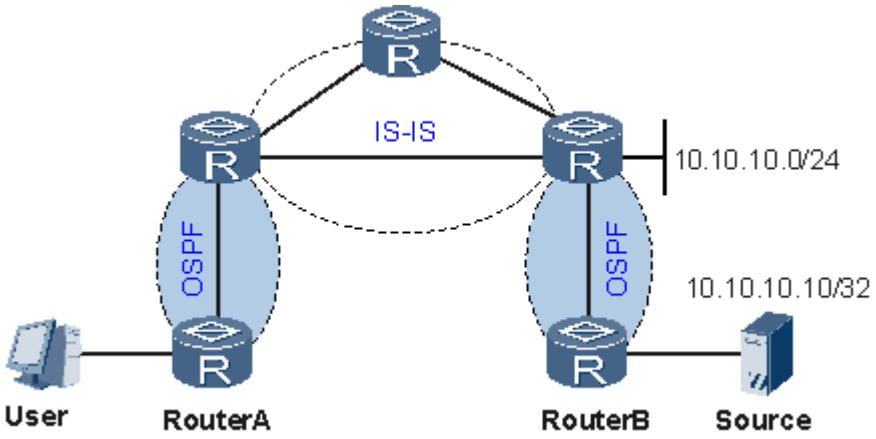
Applications

Figure 11-63 shows networking for multicast services. An IGP runs on the network; Router A is the receiver, and Router B is the multicast source server with IP address 10.10.10.10/32. The route to the multicast source server is required to

 Copyright © Huawei Technologies Co., Ltd.

converge faster than other routes, such as 10.10.10.0/24. In this case, you can set a higher convergence priority for 10.10.10.10/32 than that of 10.10.10.0/24. Then, when routes converge on the network, the route to the multicast source server 10.10.10.10/32 converges first, ensuring the transmission of multicast services.

Figure 11-63 Networking for priority-based route convergence



11.12.1.6 Load Balancing and Route Backup

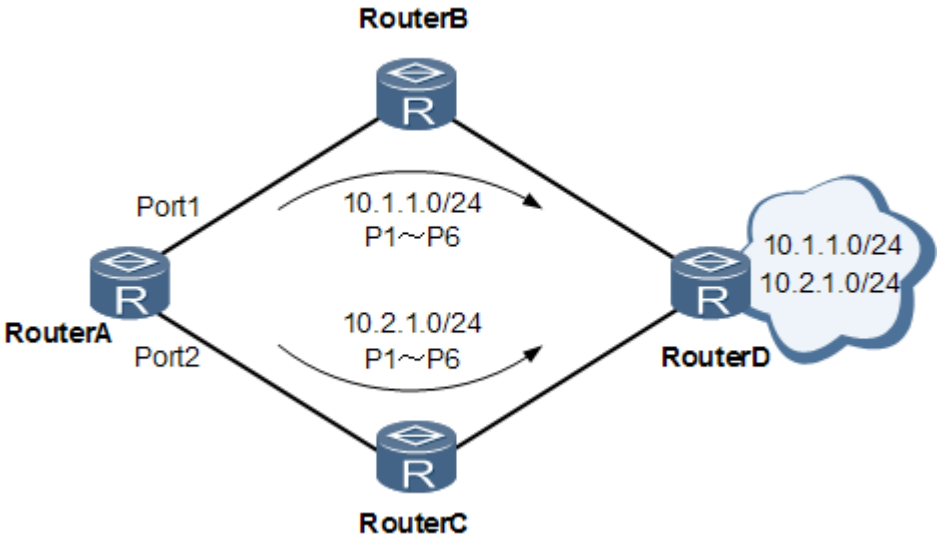
Load Balancing

The MA5800 supports the multi-route model (multiple routes with the same destination and priority). Routes discovered by one routing protocol with the same destination and cost can load-balance traffic. In each routing protocol view, you can run the **maximum load-balancing** *number* command to perform load balancing. The MA5800 adopts per-destination load balancing.

After per-destination load balancing is configured, the Router forwards packets based on the quintuple (the source address, destination address, source port, destination port, and protocol in the packets). When the quintuple is the same, the Router always chooses the next hop address that is the same as the last one to send packets. **Figure 11-64** per-destination load balancing.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-64 Networking for per-destination load balancing



Router A needs to forward packets to 10.1.1.0/24 and 10.2.1.0/24. Based on per-destination load balancing, packets of the same flow are transmitted along the same path. The processes for forwarding packets on Router A are as follows:

- The first packet P1 to 10.1.1.0/24 is forwarded through Port 1, and all subsequent packets to 10.1.1.0/24 are forwarded through Port 1.
- The first packet P1 to 10.2.1.0/24 is forwarded through Port 2, and all subsequent packets to 10.2.1.0/24 are forwarded through Port 2.

Currently, RIP, OSPF, BGP, and IS-IS support load balancing, and static routes also support load balancing.

NOTE

The number of equal-cost routes for load balancing varies with the router model.

Route Backup

The MA5800 supports route backup to improve network reliability. You can configure multiple routes to the same destination as required. The route with the highest priority functions as the primary route, and the other routes with lower priorities function as backup routes.

In most cases, the router uses the primary route to forward packets. If the link fails, the primary route becomes inactive, and the router selects the backup route with the highest priority to forward data. In this way, traffic is switched to the backup route. When the original primary route recovers, the router restores and re-selects the optimal route. Because the original primary route has the highest priority, the router selects this route to send packets. In this way, the backup route is switched back to the primary route.

11.12.1.7 Re-advertisement of Routing Information

Different routing protocols may discover different routes because they adopt different routing algorithms. When the scale of a network is large and multiple

 Copyright © Huawei Technologies Co., Ltd.

routing protocols run on the network, these protocols need to re-advertise their discovered routes.

On the MA5800, the routes discovered by a routing protocol can be imported into the routing table of another routing protocol. Each protocol has its mechanism to import routes. For details, see "Routing Policy."

11.12.1.8 Default Route

Default routes are special routes. In most cases, they are configured by administrators. Default routes can also be generated by dynamic routing protocols, such as OSPF and IS-IS.

Default routes are used only when no matching routing entry is available for packet forwarding in the routing table. A default route in the routing table is the route to the network 0.0.0.0 (with mask 0.0.0.0). You can check whether the default route is configured using the **display ip routing-table** command.

If the destination address of a packet does not match any entry in the routing table, the packet is sent along a default route. If no default route exists and the destination address of the packet does not match any entry in the routing table, the packet is discarded. An Internet Control Message Protocol (ICMP) packet is then sent, informing the originating host that the destination host or network is unreachable.

11.12.1.9 IP FRR

Definition

IP fast reroute (IP FRR) provides fast traffic recovery upon link-layer or physical-layer faults. When a fault is detected and reported to the upper-layer routing system, a backup link is used to forward the impacted traffic.

IP FRR recovers the traffic impacted by link faults.

Purpose

If a forwarding link fails on a traditional IP network, the device port goes down. After detecting such a fault, the device notifies the upper-layer routing system so that it updates the link status and recalculates a route. The process takes several seconds.

The convergence time cannot meet the requirements of delay-sensitive services, such as voice over IP (VoIP) and online video services. In contrast, IP FRR can recover impacted traffic in the level of 50 ms.

Principle

IP FRR supports static routes and dynamic routes, such as OSPF, OSPFv3, IS-IS, IS-IS IPv6, BGP, and BGP4+ routes.

IP FRR is implemented as follows:

1. Run commands to configure and enable IP FRR.

 Copyright © Huawei Technologies Co., Ltd.

2. After IP FRR is enabled, the MA5800-X17 selects primary routes using routing protocols, calculates the backup links in advance, and saves them in the forwarding table.

3. If a primary link is unavailable, the MA5800-X17 forwards traffic through the backup link.

NOTE

For detailed configuration of IP FRR for each routing protocol, see the corresponding protocol description.

IP FRR Triggering Mode

IP FRR can be triggered by a BFD session or port status.

IP FRR triggered by a BFD session

If a Bidirectional Forwarding Detection (BFD) session is configured, it automatically associates with IP FRR. If a BFD session detects a link failure, the BFD session goes down and triggers IP FRR.

If multiple BFD sessions are configured and they have identical outgoing interfaces and destination IP addresses, IP FRR can be triggered if any of the BFD sessions goes down.

IP FRR triggered by port status

NOTE

This function is supported in V100R019C1X and later versions.

You can run the **port frr-binding** command to enable the function of triggering IP FRR by port status. When a port goes down, IP FRR is triggered.

The following restrictions apply when enabling the function of triggering IP FRR by port status:

- Virtual service ports do not support IP FRR. Only standard VLAN ports support IP FRR.
- To enable the function of triggering IP FRR by port status in a VLAN, the VLAN can have only one standard port, which can be a common Ethernet port or a port aggregation group.
- Protection groups do not support IP FRR.

11.12.1.10 ECMP

Equal and Weighted Cost Multi-Path (ECMP) is a technique in which if two or more equal cost shortest paths exist between two nodes, the traffic between the nodes is distributed among the multiple equal-cost paths.

11.12.1.10.1 Introduction to ECMP

Definition

Equal and Weighted Cost Multi-Path (ECMP) is a technique in which if two or more equal cost shortest paths exist between two nodes, the traffic between the

 Copyright © Huawei Technologies Co., Ltd.

nodes is distributed among the multiple equal-cost paths. That is, in packet transmissions, if different routes with the same destination network exist in the system, the packets can be transmitted to the destination network through multiple next hops.

Purpose

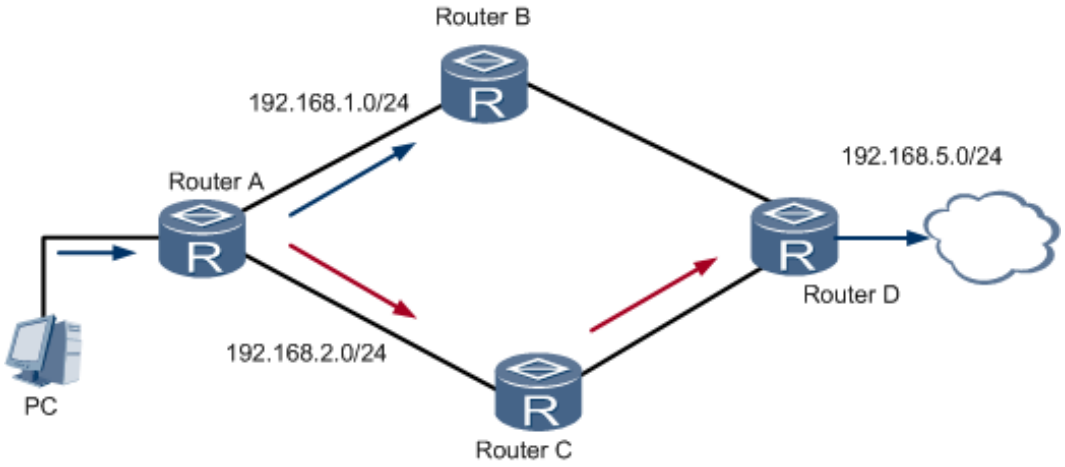
In ECMP, the traffic to the same destination network can be distributed among multiple equal-cost paths to reduce the network load, and the links in the network can back up each other. That is, when a link in the network fails, the packets on this link can be forwarded to the destination network through other links that are in the normal state.

11.12.1.10.2 ECMP Principle

In ECMP, according to different states of the network, the traffic to the same destination network can be distributed among multiple equal-cost paths to reduce the network load or to implement the link backup function.

As shown in **Figure 11-65**, assume that a packet is transmitted to the destination network (192.168.5.0) through Routers A-D, and two routes to the destination network exist in Router A. When receiving the packet from a user, Router A can select Router B or Router C as the next hop according to Hash algorithm to forward the packet to the destination network in load balancing mode.

Figure 11-65 ECMP diagram



The device supports two route load balancing modes. The **route load-balancing mode** command is used to configure the route load balancing mode.

- Based on the Hash algorithm of source and destination IP addresses, which is the default mode.
- Based on the Hash algorithm of IP address and UDP/TCP port.

11.12.1.10.3 ECMP Fast Switching

 Copyright © Huawei Technologies Co., Ltd.

Definition

The equal and weighted cost multi-path (ECMP) function transmits packets to the destination network through multiple paths of an equal cost. If one or multiple paths fail, the MA5800-X17 quickly switches traffic to normal paths. This process is known as ECMP fast switching.

Purpose

If one of the equal-cost paths fails, ECMP fast switching quickly switches the impacted traffic to another path to minimize traffic interruption duration.

Implementation

ECMP fast switching supports static routes and dynamic routes, such as OSPF, OSPFv3, IS-IS, IS-IS IPv6, BGP, and BGP4+ routes. ECMP fast switching can be triggered by a BFD session or port status change.

NOTE

If one of the equal-cost paths fails, the MA5800-X17 switches the impacted traffic to the adjacent path instead of evenly distributing the traffic among remaining available paths.

ECMP fast switching triggered by a BFD session

When a BFD session is configured, the session is automatically associated with ECMP fast switching. If a BFD session detects a link failure, the BFD session goes down and triggers ECMP fast switching.

ECMP fast switching triggered by port status

NOTE

This function is supported in V100R019C1X and later versions.

You can run the **port frr-binding** command to enable the function of triggering ECMP fast switching by port status. When a port failure is detected, ECMP fast switching is triggered.

The following restrictions apply when enabling the function of triggering ECMP fast switching by port status:

- Virtual service ports do not support ECMP fast switching. Only standard VLAN ports support ECMP fast switching.
- A VLAN with the function of triggering ECMP fast switching by port status enabled can have only one standard port, which can be a common Ethernet port or a port aggregation group.
- Protection groups do not support ECMP fast switching.

11.12.2 Static Route

 Copyright © Huawei Technologies Co., Ltd.

11.12.2.1 Introduction

Definition

Static routes are special routes that are configured by network administrators.

The MA5800 supports common static routes and the static routes associated with VPN instances. The static routes associated with VPN instances are used to manage VPN routes.

The MA5800 supports IPv4 static routes and IPv6 static routes.

Similar to IPv4 static routes, IPv6 static routes are configured by the administrator and are applicable to simple IPv6 networks. The major difference between IPv6 static routes and IPv4 static routes lies in their destination addresses and next hop addresses.

An IPv6 static route with destination address ::/0 (mask length 0) is a default IPv6 route. If the destination address of an IPv6 packet fails to match any entry in the routing table, a router selects the default IPv6 route to forward the IPv6 packet.

Purpose

On a simple network, only static routes can ensure that the network runs properly. Properly configuring and using static routes can improve network performance and guarantee the required bandwidth for important applications.

When a network fault occurs or the network topology changes, however, static routes must be changed manually by the administrator.

11.12.2.2 Components

On the MA5800, you can run the **ip route-static** command to configure a static route, which consists of the following components:

- **Destination address and mask**
- **Outbound interface and next hop address**

Destination Address and Mask

In the **ip route-static** command, the IPv4 address is expressed in dotted decimal notation, and the mask is expressed in dotted decimal notation or represented by the mask length (the number of consecutive 1s in the mask).

Outbound Interface and Next Hop Address

An outbound interface, a next hop address, or both of them can be configured for a static route.

Actually, each routing entry requires a next hop address. Before sending a packet, a device needs to search its routing table for the route matching the destination address in the packet based on the longest match rule. The device can find the associated link layer address to forward the packet only when the next hop address of the packet is available.

When specifying an outbound interface, note the following:

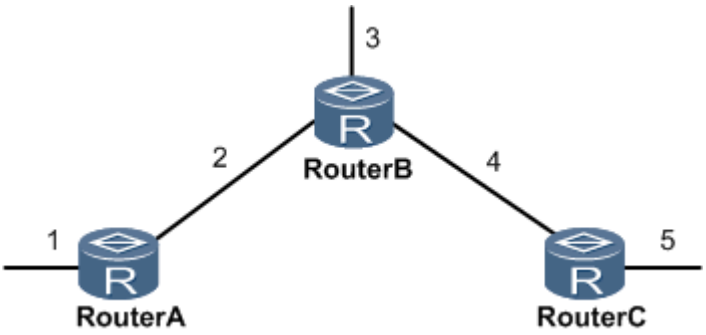
 Copyright © Huawei Technologies Co., Ltd.

- For a Point-to-Point (P2P) interface, if the outbound interface is specified, the next hop address is the address of the remote interface connected to the outbound interface. For example, when a POS interface is encapsulated with Point-to-Point Protocol (PPP) and obtains the remote IP address through PPP negotiation, you need to specify only the outbound interface rather than the next hop address.
- Non-Broadcast Multiple-Access (NBMA) interfaces are applicable to Point-to-Multipoint networks. Therefore, IP routes and the mappings between IP addresses and link layer addresses are required. In this case, you need to configure next hop addresses.
- An Ethernet interface is a broadcast interface and a virtual-template (VT) interface can be associated with multiple virtual access (VA) interfaces. If the Ethernet interface or the VT interface is specified as the outbound interface of a static route, the next hop cannot be determined because multiple next hops exist. Therefore, do not specify an Ethernet interface or a VT interface as the outbound interface unless necessary. If you need to specify a broadcast interface (such as an Ethernet interface) or a VT interface as the outbound interface, specify the associated next hop address.

11.12.2.3 Applications

As shown in **Figure 11-66**, the network topology is simple, and network communication can be implemented through static routes. You need to specify an address for each physical network, identify indirectly connected physical networks for each router, and configure static routes for indirectly connected physical networks.

Figure 11-66 Networking for static routes



In **Figure 11-66**, static routes to networks 3, 4, and 5 need to be configured on Router A; static routes to networks 1 and 5 need to be configured on Router B; static routes to networks 1, 2, and 3 need to be configured on Router C.

Default Static Route

The static route with the destination address and mask 0s (0.0.0.0 0.0.0.0) configured using the **ip route-static** command is a default route intended to simplify network configuration.

In **Figure 11-66**, because the next hop of the packets from Router A to networks 3, 4, and 5 is Router B, a default route can be configured on Router A to replace the three static routes destined for networks 3, 4, and 5. Similarly, only a default

 Copyright © Huawei Technologies Co., Ltd.

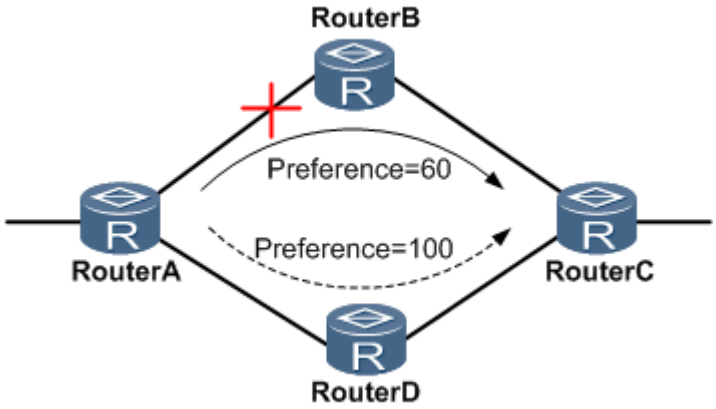
route from Router C to Router B needs to be configured to replace the three static routes destined for networks 1, 2, and 3.

Floating Static Routes

Different static routes can be configured with different priorities so that routing management policies can be flexibly applied. Route backup can be implemented by specifying different priorities for multiple routes to the same destination.

As shown in **Figure 11-67**, there are two static routes from Router A to Router C. In most cases, the only Active route is the static route with Router B as the next hop in the routing table because it has a higher priority. The other static route with Router D as the next hop functions as a backup route. The backup route is only activated to forward traffic when the primary link fails. After the primary link recovers, the static route with Router B as the next hop becomes Active to take over the traffic. Therefore, the backup route is also called a floating static route. The floating static route becomes ineffective if a fault occurs on the link between Router B and Router C.

Figure 11-67 Networking for a floating static route



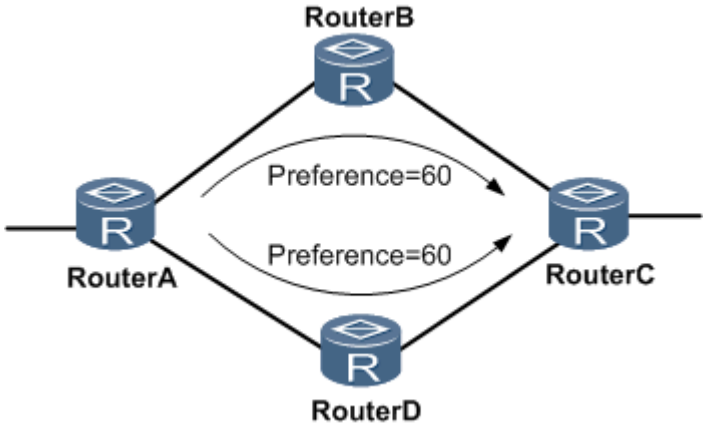
Load Balancing Among Static Routes

Routes to the same destination with the same priority can be used to load-balance traffic.

As shown in **Figure 11-68**, there are two static routes with the same priority from Router A to Router C. The two routes both exist in the routing table and forward traffic at the same time.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-68 Load balancing among static routes



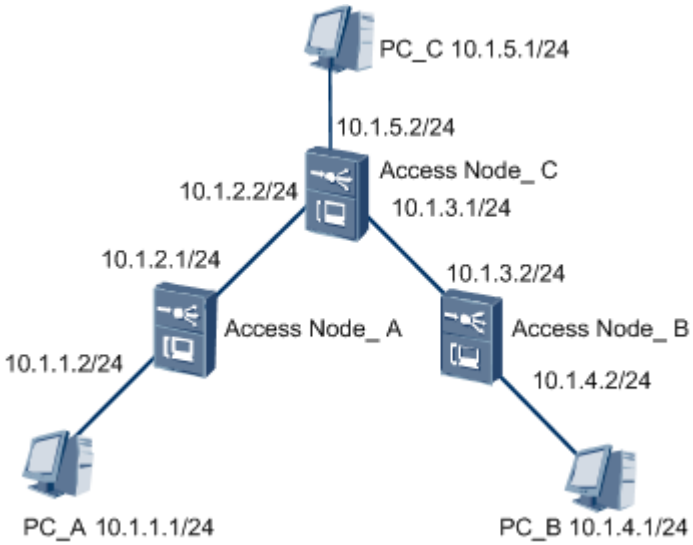
11.12.2.4 Configuration Example of the IPv4 Static Route

This topic describes how to manually add the IPv4 static route to implement the interconnection between MA5800.

Service Requirements

In this example network, MA5800_A, MA5800_B, and MA5800_C have the routing function. It is expected that after the IPv4 static route is configured, any two PCs can communicate with each other.

Figure 11-69 Example network for configuring the IPv4 static route



Procedure

Step 1 Configure the IP address of the Layer 3 interface.

The configurations for the three MA5800 devices are the same. The configuration of the MA5800 is considered as an example.

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.1.2 24
huawei(config-if-vlanif2)#ip address 10.1.2.1 24 sub
huawei(config-if-vlanif2)#quit
```

Step 2 Configure IPv4 static routes.

1. Configure an IPv4 static route for MA5800_A.

   ```
   huawei(config)#ip route-static 10.1.5.0 255.255.255.0 10.1.2.2
   huawei(config)#ip route-static 10.1.4.0 255.255.255.0 10.1.2.2
   ```

2. Configure an IPv4 static route for MA5800_B.

   ```
   huawei(config)#ip route-static 10.1.5.0 255.255.255.0 10.1.3.1
   huawei(config)#ip route-static 10.1.1.0 255.255.255.0 10.1.3.1
   ```

3. Configure IPv4 static routes for MA5800_C.

   ```
   huawei(config)#ip route-static 10.1.1.0 255.255.255.0 10.1.2.1
   huawei(config)#ip route-static 10.1.4.0 255.255.255.0 10.1.3.2
   ```

Step 3 Configure the host gateways.

1. Configure the default gateway of Host A to 10.1.1.2.
2. Configure the default gateway of Host B to 10.1.4.2.
3. Configure the default gateway of Host C to 10.1.5.2.

Step 4 Save the data.

```
huawei(config)#save
```

----End

Result

After the configuration, an interconnection can be set up between all the hosts and between all the MA5800 devices. Run the **ping** and **tracert** command to check the network connectivity.

Run the **display ip routing-table** command to query the IPv4 routing table which contains the static routing information that is configured.

Configuration File

Configuration example of MA5800_A.

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ip address 10.1.1.2 24
ip address 10.1.2.1 24 sub
quit
ip route-static 10.1.5.0 255.255.255.0 10.1.2.2
ip route-static 10.1.4.0 255.255.255.0 10.1.2.2
```

11.12.2.5 Configuration Example of the IPv6 Static Route

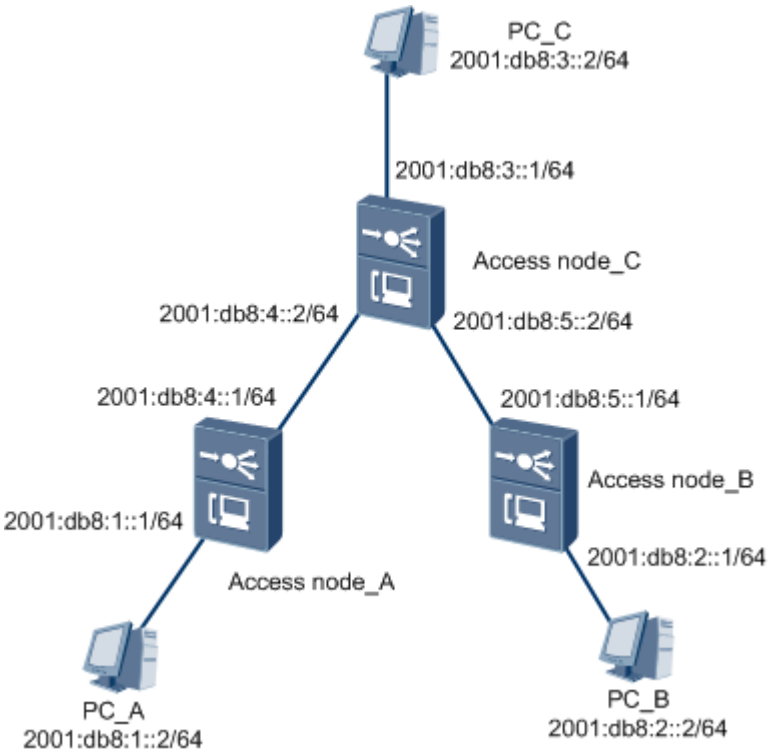
This topic describes how to manually add the IPv6 static route to implement the interconnection between access devices.

 Copyright © Huawei Technologies Co., Ltd.

Service Requirements

In this example network, three access nodes have the routing function. It is expected that after the IPv6 static route is configured, any two PCs can communicate with each other.

Figure 11-70 Example network for configuring the IPv6 static route



Procedure

Step 1 Configure the IPv6 address of the Layer 3 interface.

The configurations for the three access devices are the same. The configuration of the Access node_A is considered as an example.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1::1 64
huawei(config-if-vlanif2)#quit
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 1
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ipv6 enable
huawei(config-if-vlanif3)#ipv6 address 2001:db8:4::1 64
huawei(config-if-vlanif3)#quit
```

Step 2 Configure IPv6 static routes.

1. Configure IPv6 static route for Access node_A.

   ```
   huawei(config)#ipv6 route-static :: 0 2001:db8:4::2
   ```

 Copyright © Huawei Technologies Co., Ltd.

2. Configure IPv6 static route for Access node_B.

```
huawei(config)#ipv6 route-static :: 0 2001:db8:5::2
```

3. Configure IPv6 static route for Access node_C.

```
huawei(config)#ipv6 route-static 2001:db8:1::1 64 2001:db8:4::1
huawei(config)#ipv6 route-static 2001:db8:2::1 64 2001:db8:5::1
```

Step 3 Configure the host gateways.

1. Configure the default gateway of Host A to 2001:db8:1::1.
2. Configure the default gateway of Host B to 2001:db8:2::1.
3. Configure the default gateway of Host C to 2001:db8:3::1.

Step 4 Save the data.

```
huawei#save
```

----End

Result

After the configuration, an interconnection can be set up between all the hosts and between all the access devices. Run the **ping ipv6** and **tracert ipv6** commands to query the network connectivity.

Run the **display ipv6 routing-table** command to query the IPv6 routing table which contains the static routing information that is configured.

Configuration File

Configuration example of Access node_A.

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ipv6 enable
ipv6 address 2001:db8:1::1/64
quit
vlan 3 smart
port vlan 3 0/9 1
interface vlanif 3
ipv6 enable
ipv6 address 2001:db8:4::1/64
quit
ipv6 route-static :: 0 2001:db8:4::2
```

11.12.2.6 Example for Configuring IP FRR for an IPv4 Static Route

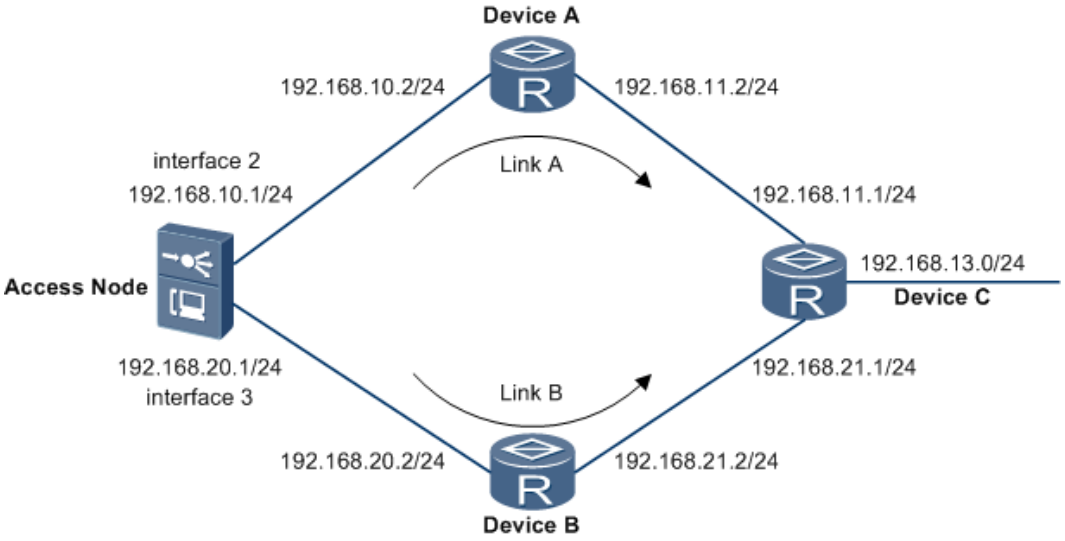
IP fast reroute (IP FRR) is applicable to services that are sensitive to packet loss and latency. Configuring IP FRR for IPv4 static routes implements link fast switching and relieves impacts on carried services brought by link faults.

Service Requirements

As shown in **Figure 11-71**, 2 IPv4 static routes having next hops device A and device B respectively are configured on the MA5800 so that link B is the backup for link A. In this case, when link A fails, traffic can be fast switched to link B.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-71 Networking of configuring IP FRR for an IPv4 static route



Data Plan

Table 11-18 describes data plan for configuring IP FRR for an IPv4 static route.

Table 11-18 Data plan for configuring IP FRR for an IPv4 static route

Item	Data
MA5800 port 2	IP address of the port: 192.168.10.1/24
	VLAN ID: 2
	Priority: 10
MA5800 port 3	IP address of the port: 192.168.20.1/24
	VLAN ID: 3
	Priority: default value

Prerequisite

Before configuring IP FRR for IPv4 static routes, complete the following task:

- You are advised to configure dynamic BFD for IPv4 static routes or static BFD for IPv4 static routes to quickly detect faults.

Procedure

Step 1 Perform configuration on the MA5800.

1. Configure the IP address of the Layer 3 interface.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
```

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 192.168.10.1 24
huawei(config-if-vlanif2)#quit
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 1
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ip address 192.168.20.1 24
huawei(config-if-vlanif3)#quit
```

2. Configure an IPv4 static route and set a higher priority for link A so that the system preferentially selects link A.

```
huawei(config)#ip route-static 192.168.13.0 24 192.168.10.2 preference 10
huawei(config)#ip route-static 192.168.13.0 24 192.168.20.2
huawei(config)#ip route-static 192.168.11.0 24 192.168.10.2
huawei(config)#ip route-static 192.168.21.0 24 192.168.20.2
```

3. Enable IP FRR for an IPv4 static route.

```
huawei(config)#ip route-static frr
```

4. Save the data.

```
huawei(config)#save
```

----End

Result

Run the **display ip routing-table verbose** command to query the information about primary/secondary routes configured.

Configuration File

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ip address 192.168.10.1 24
quit
vlan 3 smart
port vlan 3 0/9 1
interface vlanif 3
ip address 192.168.20.1 24
quit
ip route-static 192.168.13.0 24 192.168.10.2 preference 10
ip route-static 192.168.13.0 24 192.168.20.2
ip route-static 192.168.11.0 24 192.168.10.2
ip route-static 192.168.21.0 24 192.168.20.2
ip route-static frr
save
```

11.12.2.7 Example for Configuring IP FRR for an IPv6 Static Route

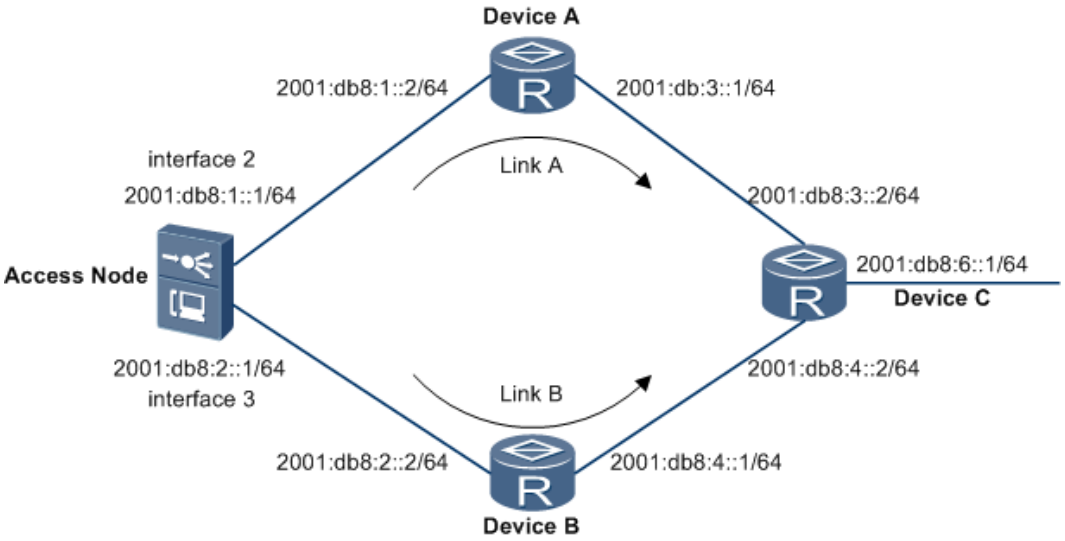
IP fast reroute (IP FRR) is applicable to services that are sensitive to packet loss and latency. Configuring IP FRR for IPv6 static routes implements link fast switching and relieves impacts on carried services brought by link faults.

Service Requirements

As shown in **Figure 11-72**, 2 IPv6 static routes having next hops device A and device B respectively are configured on the MA5800 so that link B is the backup for link A. In this case, when link A fails, traffic can be fast switched to link B.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-72 Networking of configuring IP FRR for an IPv6 static route



Data Plan

Table 11-19 describes data plan for configuring IP FRR for an IPv6 static route.

Table 11-19 Data plan for configuring IP FRR for an IPv6 static route

Item	Data
MA5800 port 2	IPv6 address of the port: 2001:db8:1::164
	VLAN ID: 2
	Priority: 10
MA5800 port 3	IPv6 address of the port: 2001:db8:2::164
	VLAN ID: 3
	Priority: default value

Prerequisite

Before configuring IP FRR for IPv6 static routes, complete the following task:

- You are advised to configure dynamic BFD for IPv6 static routes or static BFD for IPv6 static routes to quickly detect faults.

Procedure

Step 1 Perform configuration on the MA5800.

 Copyright © Huawei Technologies Co., Ltd.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1::1 64
huawei(config-if-vlanif2)#quit
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 1
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ipv6 enable
huawei(config-if-vlanif3)#ipv6 address 2001:db8:2::1 64
huawei(config-if-vlanif3)#quit
```

2. Configure an IPv6 static route and set a higher priority for link A so that the system preferentially selects link A.

```
huawei(config)#ipv6 route-static 2001:db8:6::1 64 2001:db8:1::2 preference 10
huawei(config)#ipv6 route-static 2001:db8:6::1 64 2001:db8:2::2
huawei(config)#ipv6 route-static 2001:db8:3:: 64 2001:db8:1::2
huawei(config)#ipv6 route-static 2001:db8:4:: 64 2001:db8:2::2
```

3. Enable IPv6 FRR for an IPv4 static route.

```
huawei(config)#ipv6 route-static frr
```

4. Save the data.

```
huawei(config)#save
```

----End

Result

Run the **display ipv6 routing-table verbose** command to query the information about primary/secondary routes configured.

Configuration File

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ipv6 enable
ipv6 address 2001:db8:1::1 64
quit
vlan 3 smart
port vlan 3 0/9 1
interface vlanif 3
ipv6 enable
ipv6 address 2001:db8:2::1 64
quit
ipv6 route-static 2001:db8:6:: 64 2001:db8:1::2 preference 10
ipv6 route-static 2001:db8:6:: 64 2001:db8:2::2
ipv6 route-static 2001:db8:3:: 64 2001:db8:1::2
ipv6 route-static 2001:db8:4:: 64 2001:db8:2::2
ipv6 route-static frr
save
```

11.12.3 RIP

 Copyright © Huawei Technologies Co., Ltd.

11.12.3.1 Introduction

Definition

Routing Information Protocol (RIP) is a simple Interior Gateway Protocol (IGP). RIP is used in small-scale networks, such as campus networks and simple regional networks.

As a distance-vector routing protocol, RIP exchanges routing information through User Datagram Protocol (UDP) packets with port number 520.

RIP employs the hop count as the metric to measure the distance to the destination. In RIP, by default, the number of hops from the Router to its directly connected network is 0; the number of hops from the Router to a network that is reachable through another Router is 1, and so on. The hop count (the metric) equals the number of Routers along the path from the local network to the destination network. To speed up route convergence, RIP defines the hop count as an integer that ranges from 0 to 15. A hop count that is equal to or greater than 16 is classified as infinite, indicating that the destination network or host is unreachable. Due to the hop limit, RIP is not applicable to large-scale networks.

RIP supports split horizon, poison reverse, and triggered update, which improves the performance and prevents routing loops.

Purpose

As the earliest IGP, RIP is used in small and medium-sized networks. Its implementation is simple, and the configuration and maintenance of RIP are easier than those of Open Shortest Path First (OSPF) and Intermediate System-to-Intermediate System (IS-IS). Therefore, RIP is widely used on live networks.

11.12.3.2 Principles

RIP is a distance-vector routing protocol. It forwards packets through UDP and uses timers to control the advertisement, update, and aging of routing information. However, design defects in RIP may cause routing loops. Therefore, split horizon, poison reverse, and triggered update were introduced into RIP to prevent routing loops.

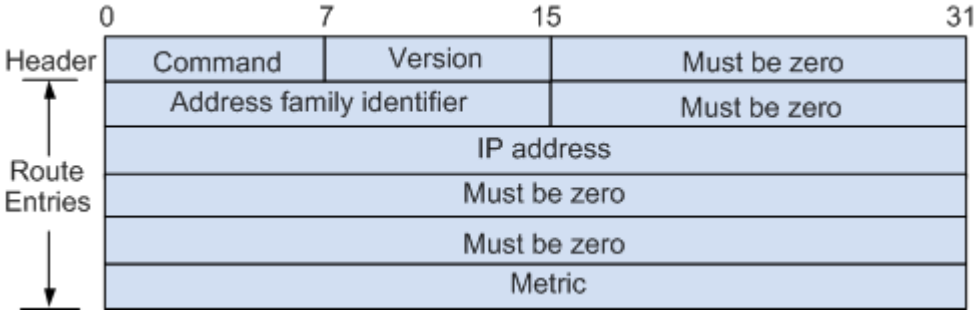
In addition, RIP periodically advertises its routing table to neighbors, and route summarization was introduced to reduce the size of the routing table.

11.12.3.2.1 RIP-1

RIP Version 1 (RIP-1) is a classful routing protocol, which supports only the broadcast of protocol packets. **Figure 11-73** shows the format of a RIP-1 packet. A RIP packet can carry a maximum of 25 routing entries. RIP is based on UDP, and a RIP-1 packet cannot be longer than 512 bytes. RIP-1 packets do not carry any mask information, and RIP-1 can identify only the routes to natural network segments, such as Class A, Class B, and Class C. Therefore, RIP-1 does not support route aggregation or discontinuous subnets.

 Copyright © Huawei Technologies Co., Ltd.

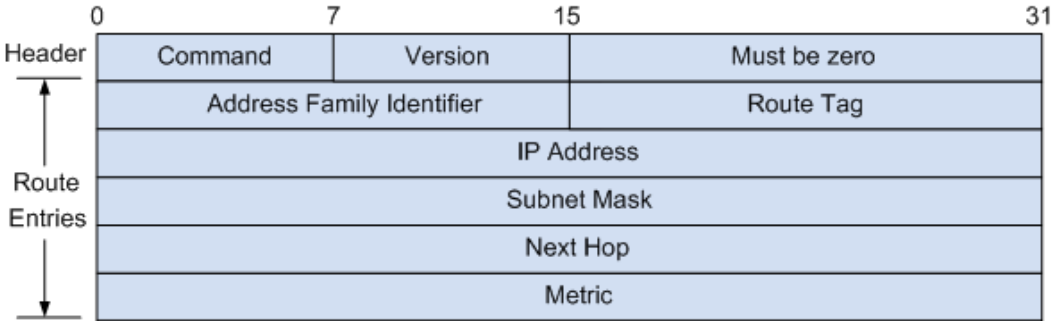
Figure 11-73 RIP-1 packet format



11.12.3.2.2 RIP-2

RIP Version 2 (RIP-2) is a classless routing protocol. **Figure 11-74** shows the format of a RIP-2 packet.

Figure 11-74 RIP-2 packet format



Compared with RIP-1, RIP-2 has the following advantages:

- Supports external route tags and flexibly controls routes based on the tag using a routing policy.
- Supports route aggregation and Classless Inter-domain Routing (CIDR) by adding mask information to RIP-2 packets.
- Supports next hop specification so that the optimal next hop address can be specified on the broadcast network.
- Supports Update packets transmission along multicast routes. Only the Routers that support RIP-2 can receive RIP-2 packets, which reduces resource consumption.
- Provides three packet authentication modes: simple authentication, Message Digest 5 (MD5) authentication and HMAC-MD5 authentication.

NOTICE

MD5 is an insecure encryption algorithm. Therefore, using HMAC-SHA256 is recommended.

11.12.3.2.3 Timers

RIP uses the following timers:

 Copyright © Huawei Technologies Co., Ltd.

- Update timer: The Update timer periodically triggers Update packet transmission. By default, the interval at which Update packets are sent is 30s.
- Age timer: If a RIP device does not receive any packets from its neighbor to update a route before the route expires, the RIP device considers the route unreachable. By default, the age timer interval is 180s.
- Garbage-collect timer: If a route becomes invalid after the age timer expires or a route unreachable message is received, the route is placed into a garbage queue instead of being immediately deleted from the RIP routing table. The garbage-collect timer monitors the garbage queue and deletes expired routes. If an Update packet of a route is received before the garbage-collect timer expires, the route is placed back into the age queue. The garbage-collect timer is set to avoid route flapping. By default, the garbage collect timer interval is 120s.
- Hold-down timer: If a RIP device receives an updated route with cost 16 from a neighbor, the route enters the holddown state, and the hold-down timer is started. To avoid route flapping, the RIP device does not accept any updated routes until the hold-down timer expires, even if the cost is less than 16 except in the following scenarios:
  a. The cost carried in the update packet is less than or equal to that carried in the last update packet.
  b. The hold-down timer expires, and the corresponding route enters the Garbage state.

The relationship between RIP routes and the four timers is as follows:

- The advertisement of RIP routing updates is triggered by the update timer with a default value 30 seconds.
- Each routing entry is associated with two timers: the age timer and garbage-collect timer.
  a. Each time a route is learned and added to the routing table, the age timer is started.
  b. If no Update packet is received from the neighbor within 180 seconds after the age timer is started, the metric of the corresponding route is set to 16, and the garbage-collect timer is started.
- If no Update packet is received within 120 seconds after the garbage-collect timer is started, the corresponding routing entry is deleted from the routing table after the garbage-collect timer expires.
- By default, the hold-down timer is disabled. If you configure a hold-down timer, it starts after the system receives a route with a cost greater than 16 from its neighbor.

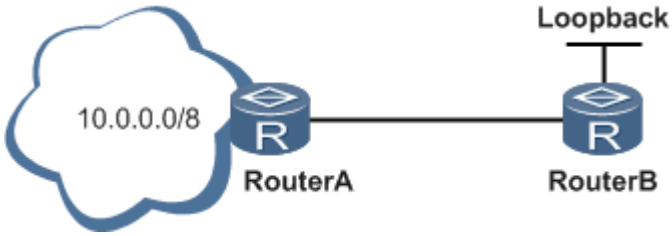
11.12.3.2.4 Split Horizon

Split Horizon on Broadcast, P2PMP, and P2P Networks

Split horizon prevents a RIP-enabled interface from sending back the routes it learns, which reduces bandwidth consumption and prevents routing loops.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-75 Networking for interface-based split horizon



In **Figure 11-75**, Router A sends Router B a route to 10.0.0.0/8. If split horizon is not configured, Router B will send this route back to Router A after learning it from Router A. As a result, Router A learns the following routes to 10.0.0.0/8:

- A direct route with zero hops
- A route with Router B as the next hop and total two hops

Only direct routes, however, are active in the RIP routing table of Router A.

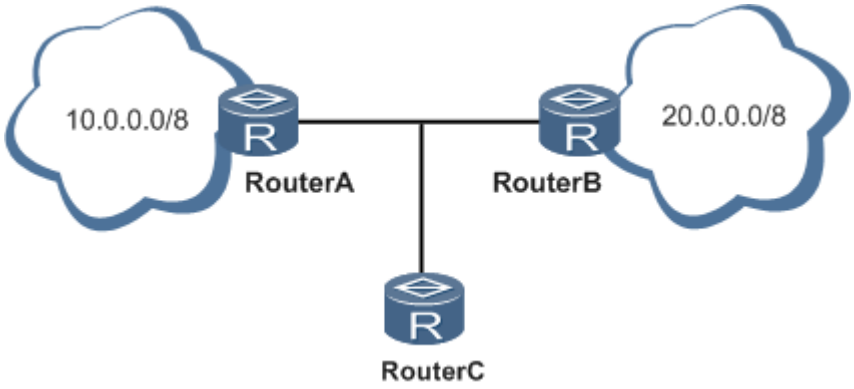
If the route from Router A to 10.0.0.0/8 becomes unreachable and Router B is not notified, Router B still considers the route to 10.0.0.0/8 reachable and continues sending this route to Router A. Then, Router A receives incorrect routing information and regards the route to 10.0.0.0/8 reachable through Router B; Router B considers the route to 10.0.0.0/8 reachable through Router A. As a result, a loop occurs on the network.

After split horizon is configured, Router B no longer sends the route back after learning the route, which prevents such a loop.

Split Horizon on NBMA Networks

On a Non-Broadcast Multi-Access (NBMA) network where an interface is connected to multiple neighbors, RIP supports neighbor-based split horizon. On NBMA networks, routes are sent in unicast mode, and an interface can differentiate which neighbor each route was learned from, and the interface will not send the routes back to the neighbor it learned them from.

Figure 11-76 Networking for neighbor-based split horizon on an NBMA network



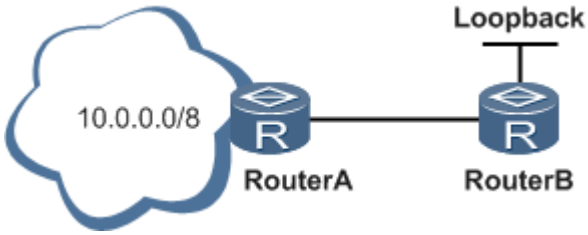
As shown in **Figure 11-76**, Router A sends the route to 10.0.0.0/8 that it learns from Router B only to Router C.

 Copyright © Huawei Technologies Co., Ltd.

11.12.3.2.5 Poison Reverse

Poison reverse allows a RIP-enabled interface to set the cost of the route that it learns from a neighbor to 16 (indicating that the route is unreachable) and then send the route back. After receiving this route, the neighbor deletes the useless route from its routing table, which prevents loops.

Figure 11-77 Networking for poison reverse



As shown in **Figure 11-77**, Router A sends Router B a route to 10.0.0.0/8. If poison reverse is not configured, Router B will send this route back to Router A after learning it from Router A. As a result, Router A learns the following routes to 10.0.0.0/8:

- A direct route with zero hops
- A route with Router B as the next hop and total two hops

Only direct routes, however, are active in the RIP routing table of Router A.

If the route from Router A to 10.0.0.0 becomes unreachable and Router B is not notified, Router B still considers the route to 10.0.0.0/8 reachable and continues sending this route to Router A. Then, Router A receives incorrect routing information and considers the route to 10.0.0.0/8 reachable through Router B; Router B considers the route to 10.0.0.0/8 reachable through Router A. As a result, a loop occurs on the network.

With poison reverse, after Router B receives the route from Router A, Router B sends a route unreachable message to Router A with cost 16. Router A then no longer learns the reachable route from Router B, which prevents routing loops.

If both split horizon and poison reverse are configured, only poison reverse takes effect.

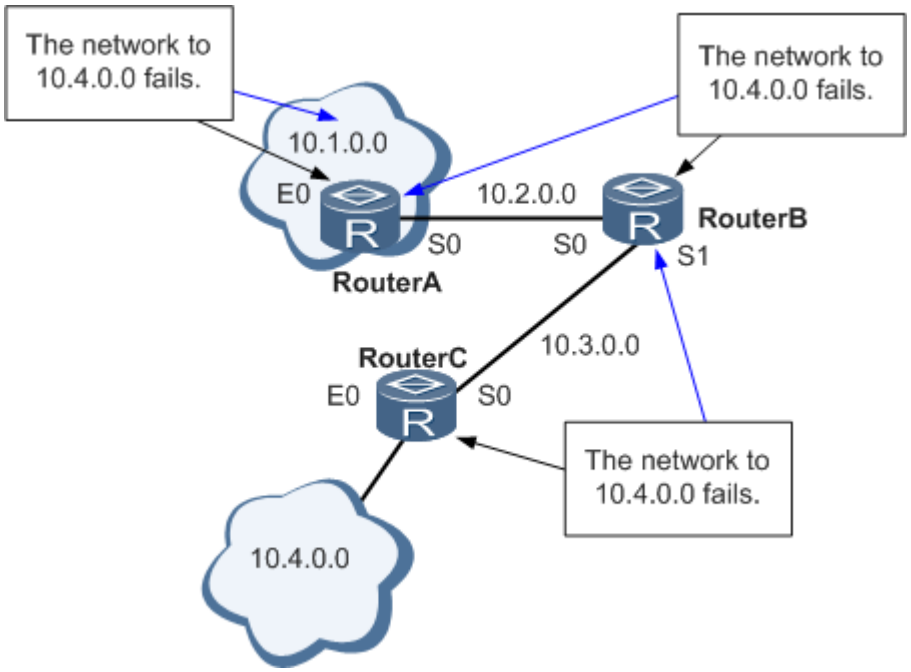
11.12.3.2.6 Triggered Update

Triggered update allows a device to advertise routing information changes to neighbor routers immediately.

Triggered update shortens the network convergence time. When local routing information changes, the local router immediately notifies its neighbor routers of the changes rather than waiting for periodical update.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-78 Networking for triggered update



In the networking shown in **Figure 11-78**, if the route to 10.4.0.0 becomes unreachable, Router C learns the information first. By default, a RIP-enabled device sends routing updates to its neighbors every 30s. If Router C receives an Update packet from Router B within 30s while Router C is still waiting to send Update packets, Router C learns the incorrect route to 10.4.0.0. In this case, the next hops of the routes from Router B or Router C to network 10.4.0.0 are Router C and Router B respectively, which results in routing loops. If Router C sends an Update packet to Router B immediately after it detects a network, Router B can rapidly update its routing table, which prevents routing loops.

In addition, if the next hop of a route becomes unavailable due to a link failure, the local Router sets the cost of the route to 16 and then advertises the route immediately to its neighbors. This process is called route poisoning.

11.12.3.2.7 Route Summarization

Route summarization allows routes to the same natural network segment but different subnets to be summarized into a single route belonging to the same network segment before it is transmitted to other network segments. RIP-1 packets do not carry mask information, and therefore RIP-1 can advertise only routes with natural masks. RIP-2 supports route summarization because RIP-2 packets carry mask information. Therefore, RIP-2 supports subnetting.

In RIP-2, route summarization can reduce the size of the routing table and improve the extensibility and efficiency of a large-scale network.

Route summarization has two modes:

- Process-based classful summarization

  Summarized routes are advertised with natural masks. If split horizon or poison reverse is configured, classful summarization becomes invalid because split horizon or poison reverse suppresses some routes from being advertised.

 Copyright © Huawei Technologies Co., Ltd.

In addition, when classful summarization is configured, routes learned from different interfaces may be summarized into a single route. As a result, a conflict occurs in the advertisement of the summarized route.

For example, a RIP process summarizes the route 10.1.1.0 /24 with metric 2 and route 10.2.2.0/24 with metric 3 into the route 10.0.0.0/8 with metric 2.

- Interface-based summarization

  Users can specify a summary address.

  For example, users can configure a RIP-enabled interface to summarize the route 10.1.1.0/24 with metric 2 and route 10.2.2.0/24 with metric 3 into the route 10.1.0.0/16 with metric 2.

11.12.3.2.8 Multi-Process

RIP supports multi-process and multi-instance to simplify network management and improve service control efficiency. Multi-process allows a set of interfaces to be associated with a specific RIP process, which ensures that the specific RIP process performs all the protocol operations only on this set of interfaces. Therefore, multiple RIP processes can run on one Router, and each process manages a unique set of interfaces. In addition, the routing data of each RIP process is independent; however, processes can import routes from each other.

11.12.3.2.9 RIP NSR

Devices with a distributed architecture support RIP Non-stop Routing (NSR). RIP backs up all route data from the Active Main Board (AMB) to the Standby Main Board (SMB). Whenever the AMB fails, the SMB becomes active and takes over traffic. RIP NSR ensures that routes are synchronous between the AMB and SMB. Therefore, during the AMB/SMB switchover, the neighbor will not detect the fault on the local device.

11.12.3.2.10 RIP Authentication

As networks develop, there has been considerable growth in all types of data, voice, and video information exchanged on networks. In addition, new services, such as E-commerce, online conferencing and auctions, video on demand (VoD), and e-learning have sprung up increasingly, requiring higher information security than before. Carriers must protect data packets from being intercepted or modified by attackers and prohibit unauthorized users from accessing network resources. RIP packet authentication effectively meets these security requirements.

RIP authentication falls into the following modes:

- Simple authentication: The authenticated party adds the configured password directly to packets for authentication. This authentication mode provides the lowest password security.
- MD5 authentication: The authenticated party uses the Message Digest 5 (MD5) algorithm to generate a ciphertext password and adds it to packets for authentication. This authentication mode improves password security.

NOTICE

MD5 is an insecure encryption algorithm. Therefore, using HMAC-SHA256 is recommended.

 Copyright © Huawei Technologies Co., Ltd.

- HMAC-SHA256 authentication: The authenticated party uses the HMAC-SHA256 algorithm to generate a ciphertext password and adds it to packets for authentication.

RIP authentication ensures network security by adding an authentication field used to encrypt a packet before sending the packet to ensure network security. After receiving a RIP packet from a remote router, the local router discards the packet if the authentication password in the packet does not match the local authentication password. This authentication mode protects the local router.

On IP networks of carriers, RIP authentication ensures the secure transmission of packets, improves the system security, and provides secure network services for carriers.

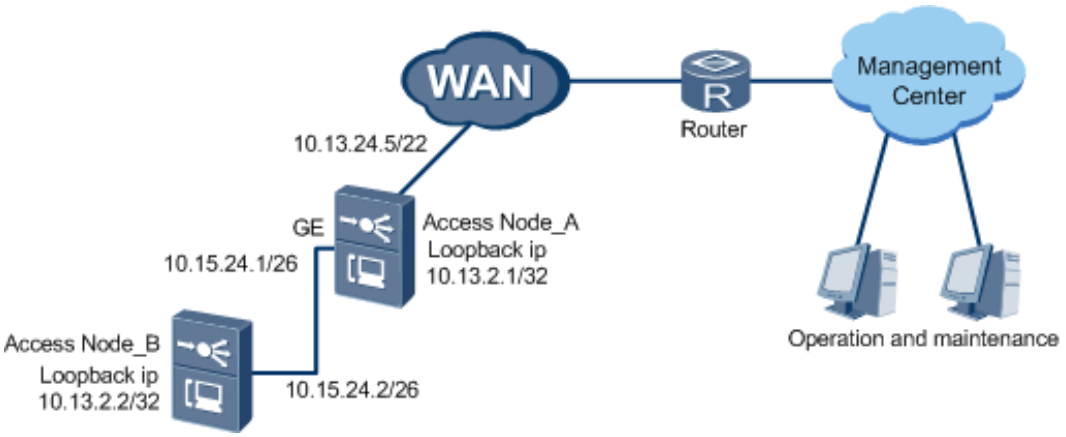
11.12.3.3 Configuration Example of RIP

This topic provides an example for configuring RIP on the MA5800.

Service Requirements

- MA5800_A is cascaded with MA5800_B through port 0/9/1, and uses port 0/9/0 to transmit services in the upstream. Besides, it connects to the management center network through the WAN.
- RIP is enabled on MA5800_A and MA5800_B so that the administrator can access MA5800_A and MA5800_B through the RIP route. Then, you can operate and maintain MA5800_A and MA5800_B.

Figure 11-79 Example network for configuring RIP



Data Plan

Table 11-20 provides the data plan for configuring RIP.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-20 Data plan for configuring RIP

Item	Data
MA5800_A	Upstream port: 0/9/0 Administration VLAN: smart VLAN 100 IP address of the Layer 3 interface in the administration VLAN: 10.13.24.5/22 Loopback interface address: 10.13.2.1/32
	RIP version: V2 RIP route filtering policy: filtering routes based on the IP address prefix list "abc". Only the routes with the IP addresses 10.13.2.1 and 10.13.2.2 can be advertised through the Layer 3 interface of VLAN 100.
	Cascading port: 0/9/0 Cascading administration VLAN: smart VLAN 10 IP address of the Layer 3 interface in the cascading administration VLAN: 10.15.24.1/26
MA5800_B	Cascading port: 0/9/1 Administration VLAN: smart VLAN 10 IP address of the Layer 3 interface in the administration VLAN: 10.15.24.2/26 Loopback interface address: 10.13.2.2/32
	RIP version: V2 RIP route filtering policy: filtering routes based on the IP address prefix list "abc". Only the route with the IP address 10.13.2.2 can be advertised through the Layer 3 interface of VLAN 10.

Procedure

- Configure MA5800_A.
  a. Configure the RIP-supported Layer 3 interface.

```
huawei(config)#vlan 100 smart
huawei(config)#port vlan 100 0/9 0
huawei(config)#interface vlanif 100
huawei(config-if-vlanif100)#ip address 10.13.24.5 22
huawei(config-if-vlanif100)#quit
huawei(config)#interface loopBack 0
huawei(config-if-loopback0)#ip address 10.13.2.1 32
huawei(config-if-loopback0)#quit
```

  b. Enable RIP.

```
huawei(config)#rip 1
huawei(config-rip-1)#network 10.0.0.0
huawei(config-rip-1)#version 2
huawei(config-rip-1)#quit
```

  c. Configure the route filtering policy.

```
huawei(config)#ip ip-prefix abc permit 10.13.2.1 32
huawei(config)#ip ip-prefix abc permit 10.13.2.2 32
```

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config)#rip 1
huawei(config-rip-1)#filter-policy ip-prefix abc export vlanif 100
huawei(config-rip-1)#quit
```

d. Configure the cascading port.

```
huawei(config)#vlan 10 smart
huawei(config)#port vlan 10 0/9 1
huawei(config)#interface mpu 0/9
huawei(config-if-mpu-0/9)#network-role 1 cascade
huawei(config-if-mpu-0/9)#quit
huawei(config)#interface vlanif 10
huawei(config-if-vlanif10)#ip address 10.15.24.1 26
huawei(config-if-vlanif10)#quit
```

e. Enable RIP on the cascading port.

```
huawei(config)#rip 1
huawei(config-rip-1)#network 10.0.0.0
huawei(config-rip-1)#quit
```

f. Save the data.

```
huawei(config)#save
```

- Configure MA5800_B.

a. Configure the RIP-supported Layer 3 interface.

```
huawei(config)#vlan 10 smart
huawei(config)#port vlan 10 0/9 0
huawei(config)#interface vlanif 10
huawei(config-if-vlanif10)#ip address 10.15.24.2 26
huawei(config-if-vlanif10)#quit
huawei(config)#interface loopBack 0
huawei(config-if-loopback0)#ip address 10.13.2.2 32
huawei(config-if-loopback0)#quit
```

b. Enable RIP.

```
huawei(config)#rip 1
huawei(config-rip-1)#network 10.0.0.0
huawei(config-rip-1)#version 2
huawei(config-rip-1)#quit
```

c. Configure the route filtering policy.

```
huawei(config)#ip ip-prefix abc permit 10.13.2.2 32
huawei(config)#rip 1
huawei(config-rip-1)#filter-policy ip-prefix abc export vlanif 10
huawei(config-rip-1)#quit
```

d. Save the data.

```
huawei(config)#save
```

----End

Result

The maintenance terminal of the administration center can access MA5800_A and MA5800_B, and operate and maintain the two devices.

Configuration File

Configuration on MA5800_A

```
vlan 100 smart
port vlan 100 0/9 0
interface vlanif 100
ip address 10.13.24.5 22
quit
interface loopBack 0
ip address 10.13.2.1 24
```

 Copyright © Huawei Technologies Co., Ltd.

```
quit
rip 1
network 10.0.0.0
version 2
quit
ip ip-prefix abc permit 10.13.2.1 32
ip ip-prefix abc permit 10.13.2.2 32
rip 1
filter-policy ip-prefix abc export vlanif 100
quit
vlan 10 smart
port vlan 10 0/9 1
interface mpu 0/9
network-role 1 cascade
quit
interface vlanif 10
ip address 10.15.24.1 26
quit
rip 1
network 10.0.0.0
quit
save
```

Configuration on MA5800_B

```
vlan 10 smart
port vlan 10 0/9 0
interface vlanif 10
ip address 10.15.24.2 26
quit
interface loopBack 0
ip address 10.13.2.2 32
quit
rip 1
network 10.0.0.0
version 2
quit
ip ip-prefix abc permit 10.13.2.2 32
rip 1
filter-policy ip-prefix abc export vlanif 10
quit
save
```

11.12.3.4 RIP Reference Standards and Protocols

The following table lists the references.

Document NO.	Document Name
RFC 1058	Routing Information Protocol
RFC 1721	RIP Version 2 Protocol Analysis
RFC 1722	RIP Version 2 Protocol Applicability Statement
RFC 1723	RIP Version 2 - Carrying Additional Information
RFC 2082	RIP-2 MD5 Authentication
RFC 2453	RIP Version 2

11.12.4 RIPng

 Copyright © Huawei Technologies Co., Ltd.

11.12.4.1 Introduction

Definition

RIP next generation (RIPng) is an extension of RIPv2 in IPv4 networks to support of IPv6, the next generation Internet Protocol. Most RIP concepts apply to RIPng.

RIPng, based on the Distance Vector (D-V) algorithm, is a routing protocol that measures the distance (metrics or cost) to the host by Hop Count (HC). RIPng defines that the HC from a device to its directly connected network is 0, and the HC from a device to a network that is reachable through another device is 1, and so on. When the HC reaches 16, the destination network or host is defined as unreachable.

To be applied on IPv6 networks, RIPng makes the following changes to RIP:

- RIPng uses UDP port 521 to send and receive routing information. RIPv2 uses UDP port 520.
- Multicast address: RIPng uses FF02::9 as the link-local multicast address of a RIPng device.
- Prefix length: RIPng uses a 128-bit (the mask length) prefix in the destination address.
- Next hop address: RIPng uses a 128-bit IPv6 address.
- Source address: RIPng uses link-local address FE80::/10 as the source address to send RIPng Update packets.

Purpose

RIPng is an extension to RIP for support of IPv6.

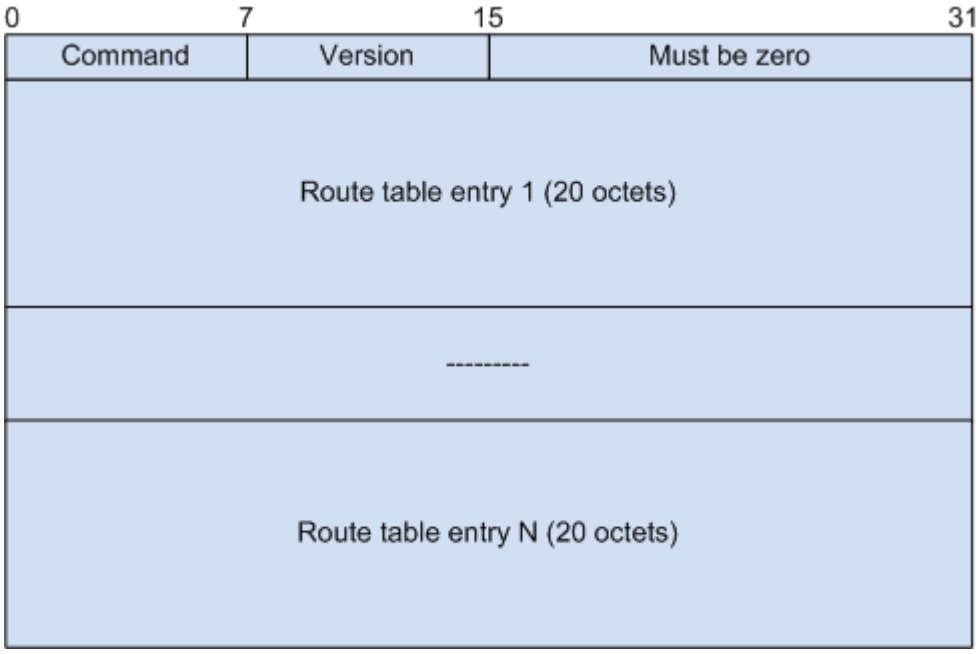
11.12.4.2 RIPng Packet Format

A RIPng packet is composed of a header and multiple route table entries (RTEs). In a RIPng packet, the maximum number of RTEs is determined by the maximum transmission unit (MTU) of an interface.

Figure 11-80 shows the basic format of a RIPng packet.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-80 RIPng packet format



A RIPng packet contains two types of RTEs:

- Next hop RTE: It defines the IPv6 address of the next hop and is located before a group of IPv6-prefix RTEs that have the same next hop. The Metric field of a next hop RTE is always 0xFF.
- IPv6-prefix RTE: It describes the destination IPv6 address and the cost in the RIPng routing table and is located after a next hop RTE. A next hop RTE can be followed by multiple different IPv6-prefix RTEs.

Figure 11-81 shows the format of a next hop RTE.

Figure 11-81 Format of the next hop RTE

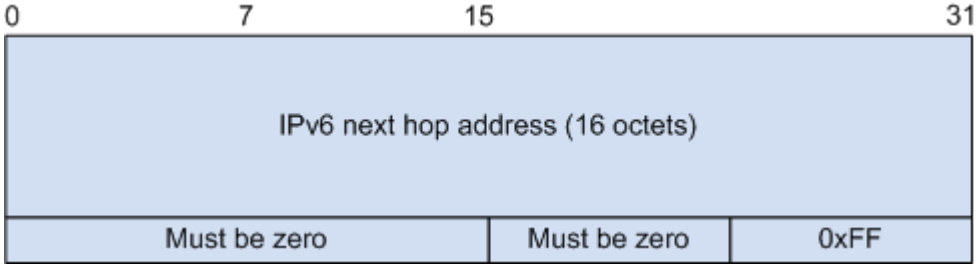
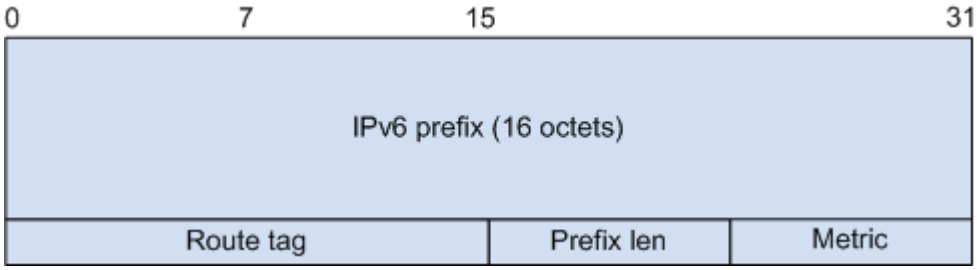


Figure 11-82 shows the format of an IPv6-prefix RTE.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-82 Format of the IPv6-prefix RTE



11.12.4.3 Timers

RIPng uses the following timers:

- Update timer: This timer periodically triggers Update packet transmission. By default, the interval at which Update packets are sent is 30s. This timer is used to synchronize RIPng routes on the network.
- Age timer: If a RIPng device does not receive any Update packet from its neighbor before a route expires, the RIPng device considers the route to its neighbor unreachable.
- Garbage-collect timer: If no packet is received to update an unreachable route after the Age timer expires, this route is deleted from the RIPng routing table.
- Hold-down timer: If a RIP device receives an updated route with cost 16 from a neighbor, the route enters the holddown state, and the hold-down timer is started.

The relationship between these timers is as follows:

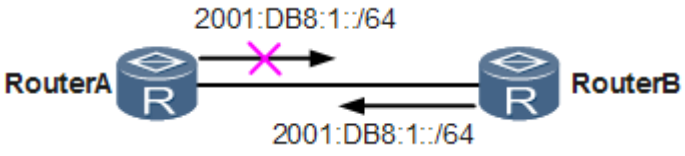
The advertisement of RIPng routing updates is periodically triggered by the update timer with default value 30 seconds. Each routing entry is associated with the age timer and garbage-collect timer. Each time a route is learned and added to the routing table, the age timer is started. If no update packet is received from the neighbor within 180 seconds, the metric of the route is set to 16, and the garbage-collect timer is started. If no update packet is received within 120 seconds, the route is deleted after the garbage-collect timer expires.

By default, hold-down timer is disabled. If you configure a hold-down timer, it starts after the system receives a route with a cost greater than 16 from its neighbor.

11.12.4.4 Split Horizon

Split horizon prevents a RIPng-enabled interface from sending back the routes it learns, which reduces bandwidth consumption and prevents routing loops.

Figure 11-83 Networking for split horizon



 Copyright © Huawei Technologies Co., Ltd.

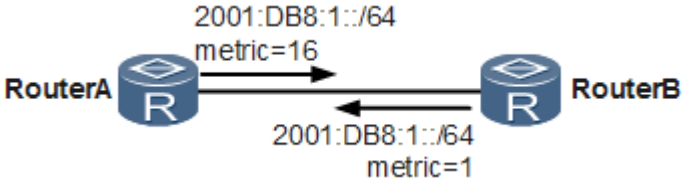
On the network shown in **Figure 11-83**, after Router B sends a route to network 2001:DB8:1::/64 to Router A, Router A does not send the route back to Router B.

11.12.4.5 Poison Reverse

Poison reverse allows a RIPng-enabled interface to set the cost of the route that it learns from a neighbor to 16 (indicating that the route is unreachable) and then send the route back. In this way, RIPng can delete useless routes from the routing table of the neighbor.

Poison reverse of RIPng can also avoid route loops.

Figure 11-84 Networking for poison reverse



On the network shown in **Figure 11-84**, if poison reverse is not configured, Router B sends Router A a route learned from Router A. The metric of the route from Router A to network 2001:DB8:1::/64 is 1. If the route from Router A to network 2001:DB8:1::/64 becomes unreachable and Router B does not receive an Update packet from Router A and keeps sending Router A the route from Router A to network 2001:DB8:1::/64, a routing loop occurs.

With poison reverse, after Router A receives the route from Router B, Router A sends a route unreachable message to Router B with metric 16. Router B then no longer learns the reachable route from Router A, which prevents routing loops.

If both poison reverse and split horizon are configured, only poison reverse takes effect.

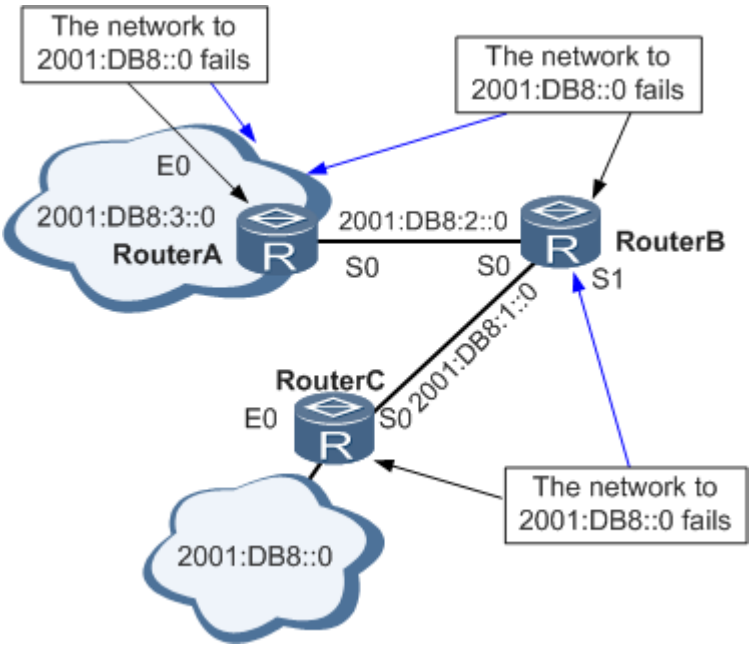
11.12.4.6 Triggered Update

Triggered update allows a device to advertise the routing information changes immediately, which speeds up network convergence.

Triggered update shortens the network convergence time. When the local routing information changes, the local router immediately notifies its neighbors of the changes in routing information rather than wait for a periodic update.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-85 Networking for triggered update



As shown in **Figure 11-85**, when network 2001:DB8::0 is unreachable, Router C learns the information first. Usually, the route update message is periodically sent to neighbor routers. For example, RIPng sends the route update message every 30s. If Router C receives an Update packet from Router B within 30s while Router C is still waiting to send Update packets, Router C learns the incorrect route to 2001:DB8::0 from Router B. In this case, the next hops of the routes from Routers B and C to 2001:DB8::0 point to each other and form a loop. If Router C sends an Update packet to Router B immediately after it detects a network fault, Router B can rapidly update its routing table, which prevents routing loops.

In addition, if the next hop of a route becomes unavailable due to a link failure, the local Router sets the metric of the route to 16 and then advertises the route immediately to its neighbors. This process is called route poisoning.

11.12.4.7 Route Summarization

Background

On large networks, the Routing Information Protocol next generation (RIPng) routing table of each device contains a large number of routes, which consumes lots of system resources. In addition, if a specific link connected to a device within an IP address range frequently alternates between Up and Down, route flapping occurs.

To address these problems, RIPng route summarization was introduced. With RIPng route summarization, a device summarizes routes destined for different subnets of a network segment into one route destined for one network segment and then advertises the summarized route to other network segments. RIPng route summarization reduces the number of routes in the routing table, minimizes system resource consumption, and prevents route flapping.

 Copyright © Huawei Technologies Co., Ltd.

Implementation

RIPng route summarization is interface-based. After RIPng route summarization is enabled on an interface, the interface summarizes routes based on the longest matching rule and then advertises the summarized route. The smallest metric among the specific routes for the summarization is used as the metric of the summarized route.

For example, an interface has two routes: 2001:db8:11::24 with metric 2 and 2001:db8:12::34 with metric 3. After RIPng route summarization is enabled on the interface, the interface summarizes the two routes into the route 2001:db8::0/32 with metric 2 and then advertises it.

11.12.4.8 Multi-Process and Multi-Instance

RIPng supports multi-process and multi-instance, which simplifies network management and improves service control efficiency. Multi-process allows a set of interfaces to be associated with a specific RIPng process, which ensures that the specific RIPng process performs all the protocol operations only on this set of interfaces. Therefore, multiple RIPng processes can run on one router, and each process manages a unique set of interfaces. In addition, the routing data of each RIPng process is independent; however, processes can import routes from each other.

On routers that support VPN, each RIPng process is associated with a specific VPN instance. Therefore, all the interfaces associated with the RIPng process need to be associated with the RIPng process-related VPN instance.

11.12.4.9 Hot Standby

Devices of a distributed architecture support RIPng Hot Standby (HSB). In the RIPng HSB process, RIPng backs up RIPng configuration from the Active Main Board (AMB) to the Standby Main Board (SMB). Whenever the AMB fails, the SMB becomes active to take over the work. RIPng re-sends synchronization requests to ensure that RIPng works properly.

11.12.4.10 IPSec Authentication

Background

As networks develop, network security has become an increasing concern. Internet Protocol Security (IPsec) authentication is used to authenticate Routing Information Protocol next generation (RIPng) packets for the network security concern. The packets that fail to be authenticated are discarded, which prevents data transmitted based on TCP/IP from being intercepted, tampered with, or attacked.

Implementation

IPSec has an open standard architecture and ensures secure packet transmission on the Internet by encrypting packets. RIPng IPSec provides a complete set of security protection mechanisms to authenticate RIPng packets, which prevents devices from being attacked by forged RIPng packets.

 Copyright © Huawei Technologies Co., Ltd.

IPSec includes a set of protocols that are used at the network layer to ensure data security, such as Authentication Header (AH), Encapsulating Security Payload (ESP), and Internet Key Exchange (IKE). The three protocols are described as follows:

- AH: A protocol that provides data origin authentication, data integrity check, and anti-replay protection. AH does not encrypt packets to be protected.
- ESP: A protocol that provides IP packet encryption and authentication mechanisms besides the functions provided by AH. The encryption and authentication mechanisms can be used together or independently.

  NOTE

  AH and ESP can be used together or independently.

Benefits

RIPng IPSec offers the following benefits:

- Improves carriers' reputation and competitiveness by preventing services from being tampered with or attacked by unauthorized users.
- Ensures confidentiality and integrity of user packets.

11.12.4.11 RIPng NSR

Background

As networks develop, the demand for data, audio, and video services is growing, which impose increasing requirements on IP network reliability. If an AMB/SMB switchover is performed on a device during a maintenance operation or a single point of failure occurs, the device may fail to converge routing information, which may result in traffic loss or even a network breakdown. Non-stop routing (NSR) can address this problem and ensure uninterrupted forwarding of key services.

Related Concepts

- High availability (HA): supports data backup between the AMB and SMB.
- Non-stop forwarding (NSF): enables a node to use the GR mechanism to ensure uninterrupted transmission during an AMB/SMB switchover.
- NSR: allows a standby control plane to take over traffic from an active control plane if the active control plane fails, preventing the control planes of neighbors from detecting the fault.
- AMB and SMB: run control plane processes.

Implementation

With Routing Information Protocol next generation (RIPng) NSR, RIPng real-time data is synchronized between the AMB and SMB. After an AMB/SMB switchover is performed on a device, the SMB takes over services from the AMB, and neighbors are unaware of the local fault. After the switchover, the new AMB restores IS-IS immediately based on the synchronized IS-IS real-time data. Therefore, neighbors are unaware of the switchover as well. RIPng NSR requires synchronization of the following data:

 Copyright © Huawei Technologies Co., Ltd.

- All configuration data, such as information about neighbors, timer parameters, and process configurations.
- Dynamic data, such as the interface parameters and state, and information about neighbors and the link state database (LSDB).

Usage Scenario

NSR minimizes the impact of control plane faults and prevents route flapping on networks that require high reliability.

Benefits

NSR improves network reliability and ensures uninterrupted traffic forwarding.

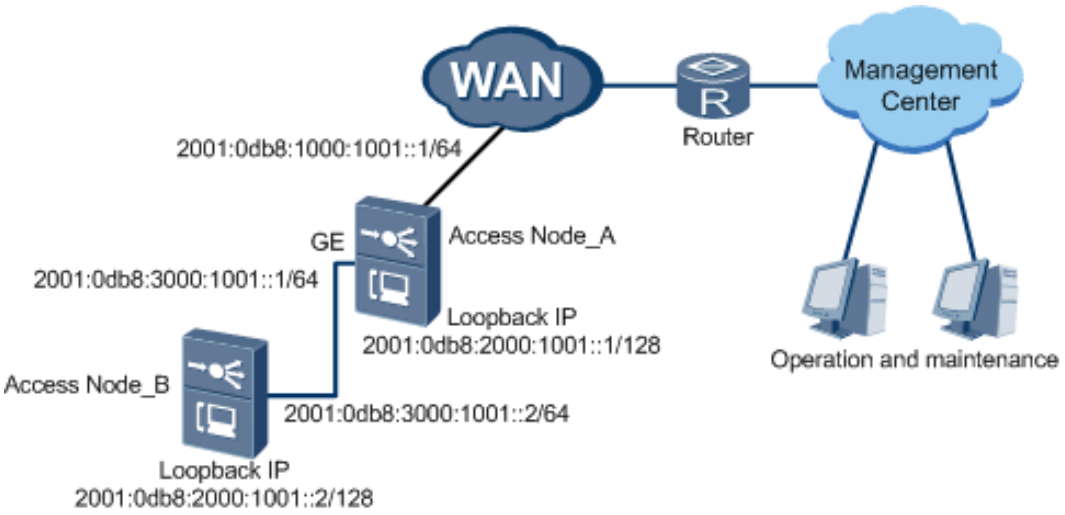
11.12.4.12 Configuration Example of RIPng

This topic provides an example for configuring RIPng on the MA5800.

Service Requirements

- Access Node_A is cascaded with Access node_B through port 0/9/1, and uses port 0/9/0 to transmit services in the upstream. Besides, it connects to the management center network through the WAN.
- RIPng is enabled on Access Node_A and Access Node_B so that the administrator can access Access Node_A and Access Node_B through the RIPng route. Then, you can operate and maintain Access Node_A and Access Node_B.

Figure 11-86 Example network for configuring RIPng



Data Plan

Table 11-21 provides the data plan for configuring RIPng.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-21 Data plan for configuring RIPng

Item	Data
Access node_A	Upstream port: 0/9/0 Administration VLAN: smart VLAN 100 IPv6 address of the Layer 3 interface in the administration VLAN: 2001:0db8:1000:1001::1/64 Loopback interface address: 2001:0db8:2000:1001::1/128
	RIPng route filtering policy: filtering routes based on the IPv6 address prefix list "abc". Only the routes with the IP addresses 2001:0db8:2000:1001::1 and 2001:0db8:2000:1001::2 can be advertised through the Layer 3 interface of VLAN 100.
	Cascading port: 0/9/0 Cascading administration VLAN: smart VLAN 10 IPv6 address of the Layer 3 interface in the cascading administration VLAN: 2001:0db8:3000:1001::1/64
Access node_B	Cascading port: 0/9/1 Administration VLAN: smart VLAN 10 IPv6 address of the Layer 3 interface in the administration VLAN: 2001:0db8:3000:1001::2/64 Loopback interface address: 2001:0db8:2000:1001::2/128
	RIPng route filtering policy: filtering routes based on the IPv6 address prefix list "abc". Only the route with the IP address 2001:0db8:2000:1001::2 can be advertised through the Layer 3 interface of VLAN 10.

Procedure

- Configure Access Node_A.
  a. Configure the RIPng-supported Layer 3 interface.

```
huawei(config)#vlan 100 smart
huawei(config)#port vlan 100 0/9 0
huawei(config)#interface vlanif 100
huawei(config-if-vlanif100)#ipv6 enable
huawei(config-if-vlanif100)#ipv6 address 2001:0db8:1000:1001::1 64
huawei(config-if-vlanif100)#quit
huawei(config)#interface loopBack 0
huawei(config-if-loopback0)#ipv6 enable
huawei(config-if-loopback0)#ipv6 address 2001:0db8:2000:1001::1/128
huawei(config-if-loopback0)#quit
```

  b. Enable RIPng.

```
huawei(config)#ripng 1
huawei(config-ripng-1)#quit
huawei(config)#interface vlanif 100
huawei(config-if-vlanif100)#ripng 1 enable
```

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config-if-vlanif100)#quit
huawei(config)#interface loopBack 0
huawei(config-if-loopback0)#ripng 1 enable
huawei(config-if-loopback0)#quit
```

c. Configure the route filtering policy.

```
huawei(config)#ip ipv6-prefix abc permit 2001:0db8:2000:1001::1 128
huawei(config)#ip ipv6-prefix abc permit 2001:0db8:2000:1001::2 128
huawei(config)#ripng 1
huawei(config-ripng-1)#filter-policy ipv6-prefix abc export
huawei(config-ripng-1)#quit
```

d. Configure the cascading port.

```
huawei(config)#vlan 10 smart
huawei(config)#port vlan 10 0/9 1
huawei(config)#interface mpu 0/9
huawei(config-if-mpu-0/9)#network-role 1 cascade
huawei(config-if-mpu-0/9)#quit
huawei(config)#interface vlanif 10
huawei(config-if-vlanif10)#ipv6 enable
huawei(config-if-vlanif10)#ipv6 address 2001:0db8:3000:1001::1 64
huawei(config-if-vlanif10)#quit
```

e. Enable RIPng on the cascading port.

```
huawei(config)#interface vlanif 10
huawei(config-if-vlanif10)#ripng 1 enable
huawei(config-if-vlanif10)#quit
```

f. Save the data.

```
huawei(config)#save
```

- Configure Access Node_B.

a. Configure the RIPng-supported Layer 3 interface.

```
huawei(config)#vlan 10 smart
huawei(config)#port vlan 10 0/9 0
huawei(config)#interface vlanif 10
huawei(config-if-vlanif10)#ipv6 enable
huawei(config-if-vlanif10)#ipv6 address 2001:0db8:3000:1001::2 64
huawei(config-if-vlanif10)#quit
huawei(config)#interface loopback 0
huawei(config-if-loopback0)#ipv6 enable
huawei(config-if-loopback0)#ipv6 address 2001:0db8:2000:1001::2 128
huawei(config-if-loopback0)#quit
```

b. Enable RIPng.

```
huawei(config)#ripng 1
huawei(config-ripng-1)#quit
huawei(config)#interface vlanif 10
huawei(config-if-vlanif10)#ripng 1 enable
huawei(config-if-vlanif10)#quit
huawei(config)#interface loopback 0
huawei(config-if-loopback0)#ripng 1 enable
huawei(config-if-loopback0)#quit
```

c. Configure the route filtering policy.

```
huawei(config)#ip ipv6-prefix abc permit 2001:0db8:2000:1001::2 128
huawei(config)#ripng 1
huawei(config-ripng-1)#filter-policy ipv6-prefix abc export
huawei(config-ripng-1)#quit
```

d. Save the data.

```
huawei(config)#save
```

----End

 Copyright © Huawei Technologies Co., Ltd.

Result

The maintenance terminal of the administration center can access Access Node_A and Access Node_B, and operate and maintain the two devices.

Configuration File

Configuration on Access Node_A

```
vlan 100 smart
port vlan 100 0/9 0
interface vlanif 100
ipv6 enable
ipv6 address 2001:0db8:1000:1001::1 64
quit
interface loopBack 0
ipv6 enable
ipv6 address 2001:0db8:2000:1001::1/128
quit
ripng 1
quit
interface vlanif 100
ripng 1 enable
quit
interface loopBack 0
ripng 1 enable
quit
ip ipv6-prefix abc permit 2001:0db8:2000:1001::1 128
ip ipv6-prefix abc permit 2001:0db8:2000:1001::2 128
ripng 1
filter-policy ipv6-prefix abc export
quit
vlan 10 smart
port vlan 10 0/9 1
interface mpu 0/9
network-role 1 cascade
quit
interface vlanif 10
ipv6 enable
ipv6 address 2001:0db8:3000:1001::1 64
ripng 1 enable
quit
save
```

Configuration on Access node_B

```
vlan 10 smart
port vlan 10 0/9 0
interface vlanif 10
interface vlanif 10
ipv6 enable
ipv6 address 2001:0db8:3000:1001::2 64
quit
interface loopback 0
ipv6 enable
ipv6 address 2001:0db8:2000:1001::2 128
quit
ripng 1
quit
interface vlanif 10
ripng 1 enable
quit
interface loopback 0
ripng 1 enable
quit
ip ipv6-prefix abc permit 2001:0db8:2000:1001::2 128
ripng 1
filter-policy ipv6-prefix abc export
```

 Copyright © Huawei Technologies Co., Ltd.

```
quit
save
```

11.12.4.13 RIPng Reference Standards and Protocols

The following table lists the references.

Document No.	Document Name	Remarks
RFC 2080	RIPng for IPv6	-

Document No.	Document Name	Protocol Compliance
RFC 2080	RIPng for IPv6	Fully compliant.

11.12.5 IS-IS

11.12.5.1 Introduction

Definition

Intermediate System to Intermediate System (IS-IS) is a dynamic routing protocol initially designed by the International Organization for Standardization (ISO) for its Connectionless Network Protocol (CLNP).

To support IP routing, the Internet Engineering Task Force (IETF) extends and modifies IS-IS in RFC 1195, which enables IS-IS to be applied to both TCP/IP and Open System Interconnection (OSI) environments. This type of IS-IS is called Integrated IS-IS or Dual IS-IS.

In this document, IS-IS refers to Integrated IS-IS, unless otherwise stated.

Purpose

As an Interior Gateway Protocol (IGP), IS-IS is used in Autonomous Systems (ASs). IS-IS is a link state protocol, and it uses the Shortest Path First (SPF) algorithm to calculate routes.

11.12.5.2 Principles

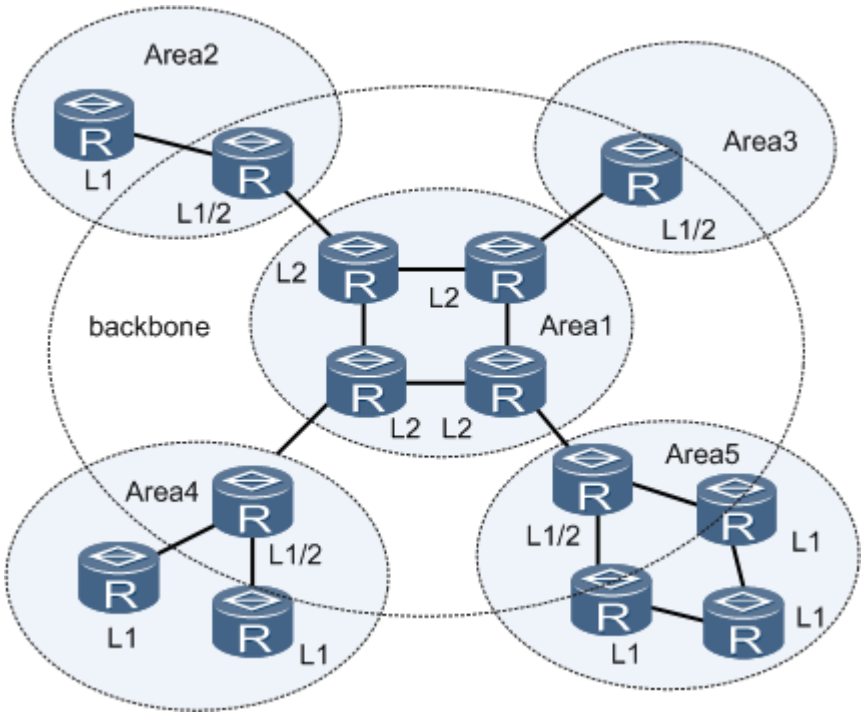
 Copyright © Huawei Technologies Co., Ltd.

11.12.5.2.1 Basic Concepts of IS-IS

IS-IS Areas

To support large-scale routing networks, IS-IS adopts a two-level structure in a routing domain. A large domain is divided into areas. **Figure 11-87** shows an IS-IS network. The entire backbone area covers all Level-2 routers in area 1 and Level-1-2 routers in other areas. Three types of routers on the IS-IS network are described as follows:

Figure 11-87 IS-IS topology



- Level-1 router

  A Level-1 device manages intra-area routing. It establishes neighbor relationships with only the Level-1 and Level-1-2 devices in the same area and maintains a Level-1 LSDB. The LSDB contains routing information in the local area. A packet to a destination beyond this area is forwarded to the nearest Level-1-2 device.

- Level-2 router

  A Level-2 device manages inter-area routing. It can establish neighbor relationships with all Level-2 devices and Level-1-2 devices, and maintains a Level-2 LSDB which contains inter-area routing information.

  All Level-2 devices form the backbone network of the routing domain. Level-2 neighbor relationships are set up between them. They are responsible for communications between areas. The Level-2 devices in the routing domain must be in succession to ensure the continuity of the backbone network. Only Level-2 devices can exchange data packets or routing information with the devices beyond the routing domain.

- Level-1-2 router

 Copyright © Huawei Technologies Co., Ltd.

A device, which can establish neighbor relationships with both Level-1 devices and Level-2 devices, is called a Level-1-2 device. A Level-1-2 device can establish Level-1 neighbor relationships with Level-1 devices and Level-1-2 devices in the same area. It can also establish Level-2 neighbor relationships with Level-2 devices and Level-1-2 devices in other areas. Level-1 devices can be connected to other areas only through Level-1-2 devices.

A Level-1-2 device maintains two LSDBs: a Level-1 LSDB and a Level-2 LSDB. The Level-1 LSDB is used for intra-area routing, while the Level-2 LSDB is used for inter-area routing.

NOTE

Level-1 routers in different areas cannot establish neighbor relationships. Level-2 routers can establish neighbor relationships with each other, regardless of the areas to which the Level-2 routers belong.

In general, Level-1 devices are located within an area, Level-2 devices are located between areas, and Level-1-2 devices are located between Level-1 devices and Level-2 devices.

Interface level

A Level-1-2 device may need to establish only a Level-1 adjacency with a neighbor and establish only a Level-2 adjacency with another neighbor. In this case, you can set the level of an interface to control the setting of adjacencies on the interface. Specifically, only Level-1 adjacencies can be established on a Level-1 interface, and only Level-2 adjacencies can be established on a Level-2 interface.

Address Structure of IS-IS

In OSI, the NSAP is used to locate resources. The ISO adopts the address structure shown in **Figure 11-88**. An NSAP is composed of the Initial Domain Part (IDP) and the Domain Specific Part (DSP). IDP is the counterpart of network ID in an IP address, and DSP is the counterpart of the subnet number and host address in an IP address.

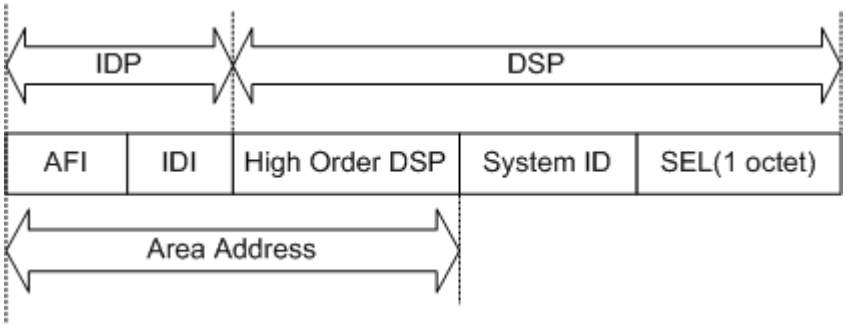
As defined by the ISO, the IDP consists of the Authority and Format Identifier (AFI) and Initial Domain Identifier (IDI). AFI specifies the address assignment mechanism and the address format; the IDI identifies a domain.

The DSP consists of the High Order DSP (HODSP), system ID, and NSAP Selector (SEL). The HODSP is used to divide areas; the system ID identifies a host; the SEL indicates the service type.

The lengths of the IDP and DSP are variable. The length of the NSAP varies from 8 bytes to 20 bytes.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-88 IS-IS address structure



- Area address

  An IDP and HODSP of the DSP can identify a routing domain and the areas in a routing domain; therefore, the combination of the IDP and HODSP is referred to as an area address, equal to an area ID in OSPF. An area address is used to uniquely identify an area in a routing domain. The area addresses of routers in the same Level-1 area must be the same, while the area addresses of routers in the Level-2 area can be different.

  In general, a router can be configured with only one area address. The area address of all nodes in an area must be the same. In the implementation of a device, an IS-IS process can be configured with a maximum of three area addresses to support seamless combination, division, and transformation of areas.

- System ID

  A system ID uniquely identifies a host or a router in an area. In the device, the length of the system ID is 48 bits (6 bytes).

  In actual applications, a router ID corresponds to a system ID. If a device uses the IP address of Loopback 0 (192.168.1.1) as its router ID, its system ID used in IS-IS can be obtained in the following steps:

  – Extend each part of the IP address 192.168.1.1 to 3 digits and add 0 or 0s to the front of the part that is shorter than 3 digits.
  – Divide the extended address 192.168.001.001 into three parts, with each part consisting of 4 decimal digits.
  – The reconstructed 1921.6800.1001 is the system ID.

  There are many ways to specify a system ID. You need to ensure that the system ID uniquely identifies a host or a device.

- SEL

  The role of an SEL (also referred to as NSAP Selector or N-SEL) is similar to that of the "protocol identifier" of IP. A transport protocol matches an SEL. The SEL is "00" in IP.

- NET

  A Network Entity Title (NET) indicates the network layer information of an IS itself. It does not contain the transport layer information (SEL = 0). A NET can be regarded as a special NSAP. The length of the NET field is the same as that of an NSAP, varying from 8 bytes to 20 bytes. For example, in NET ab.cdef.1234.5678.9abc.00, the area is ab.cdef, the system ID is 1234.5678.9abc, and the SEL is 00.

 Copyright © Huawei Technologies Co., Ltd.

In general, an IS-IS process is configured with only one NET. When areas need to be redefined, for example, areas need to be combined or an area needs to be divided into sub-areas, you can configure multiple NETs.

NOTE

A maximum of three area addresses can be configured in an IS-IS process, and therefore, you can configure only a maximum of three NETs. When you configure multiple NETs, ensure that their system IDs are the same.

The routers in an area must have the same area address.

IS-IS Network Types

IS-IS supports the following types of networks:

- Broadcast network
- Point-to-point (P2P) network

11.12.5.2.2 Basic Protocols of IS-IS

Related Concepts

DIS and Pseudo Node

A Designated Intermediate System (DIS) is an intermediate router elected in IS-IS communication. A pseudo node simulates a virtual node on a broadcast network and is not a real router. In IS-IS, a pseudo node is identified by the system ID and 1-byte circuit ID (a non-zero value) of a DIS.

The DIS is used to create and update pseudo nodes and generate the **link state protocol data units (LSPs)** of pseudo nodes. The routers advertise a single link to a pseudo node and obtain routing information about the entire network through the pseudo node. The router does not need to exchange packets with all the other routers on the network. Using the DIS and pseudo nodes simplifies network topology and reduces the length of LSPs generated by routers. When the network changes, fewer LSPs are generated. As a result, the SPF consumes fewer resources.

SPF Algorithm

The SPF algorithm, also named Dijkstra's algorithm, is used in a link-state routing protocol to calculate the shortest paths to other nodes on a network. In the SPF algorithm, a local router takes itself as the root and generates a shortest path tree (SPT) based on the network topology to calculate the shortest path to every destination node on a network. In IS-IS, the SPF algorithm runs separately in Level-1 and Level-2 databases.

Implementation

All routers on the IS-IS network communicate through the following steps:

- **Establishment of IS-IS Neighbor Relationships**
- **LSDB Synchronization**
- **Route Calculation**

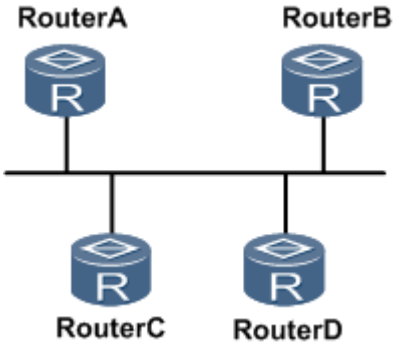
Establishment of IS-IS Neighbor Relationships

 Copyright © Huawei Technologies Co., Ltd.

On different types of networks, the modes for establishing IS-IS neighbor relationships are different.

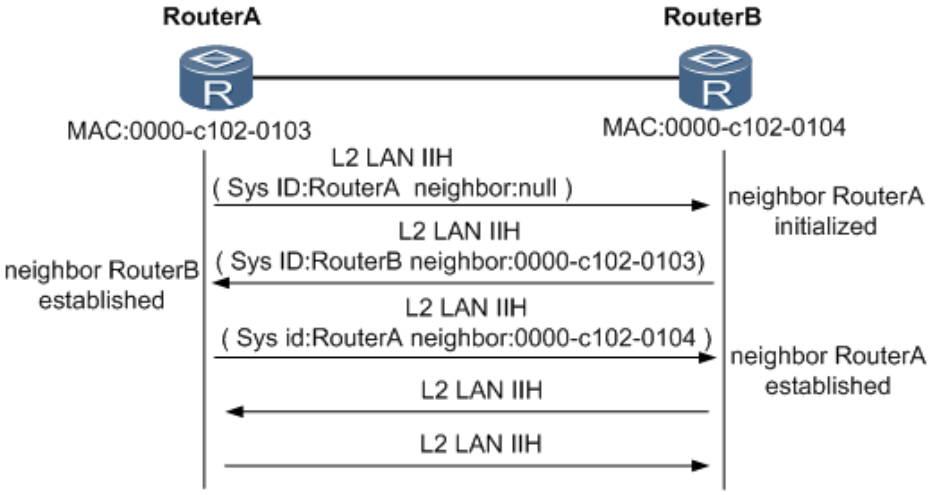
- Establishment of a neighbor relationship on a broadcast link

Figure 11-89 Networking for a broadcast link



Router A, Router B, Router C, and Router D are Level-2 routers. Router A is newly added to the broadcast network. **Figure 11-90** demonstrates the process of establishing the neighbor relationship between Router A and Router B, the process of establishing the neighbor relationship between Router A and Router C or Router D is similar to that between Router A and Router B.

Figure 11-90 Establishing a neighbor relationship on a broadcast link



As shown in **Figure 11-90**, the process for establishing a neighbor relationship on a broadcast link consists of the following phases:

- Router A broadcasts a Level-2 local area network (LAN) **IS-to-IS Hello PDU (IIH)**. After Router B receives the IIH, Router B detects that the neighbor field in the IIH does not contain its media access control (MAC) address, and sets its neighbor status with Router A to **Initial**.
- Router B replies a Level-2 LAN IIH to Router A. After Router A receives the IIH, Router A detects that the neighbor field in the IIH contains its MAC address, and sets its neighbor status with Router B to **Up**.

 Copyright © Huawei Technologies Co., Ltd.

– Router A sends a Level-2 LAN IIH to Router B. After Router B receives the IIH, Router B detects that the neighbor field in the IIH contains its MAC address, and sets its neighbor status with Router A to **Up**.

DIS Election

On a broadcast network, any two routers exchange information. If n routers are available on the network, n x (n - 1)/2 adjacencies must be established. Each status change of a router is transmitted to other routers, which wastes bandwidth resources. IS-IS resolves this problem by introducing the DIS. All routers send information to the DIS, which then broadcasts the network link status. Using the DIS and pseudo nodes simplifies network topology and reduces the length of LSPs generated by routers. When the network changes, fewer LSPs are generated. As a result, the SPF consumes fewer resources.

A DIS is elected after a neighbor relationship is established. Level-1 and Level-2 DISs are elected separately. You can configure different priorities for DISs at different levels. In DIS election, a Level-1 priority and a Level-2 priority are specified for every interface on every router. A router uses every interface to send IIHs and advertises its priorities in the IIHs to neighboring routers. The higher the priority, the higher the probability of being elected as the DIS. If there are multiple routers with the same highest priority on a broadcast network, the one with the largest MAC address is elected. The DISs at different levels can be the same router or different routers.

In the DIS election procedure, IS-IS is different from Open Shortest Path First (OSPF) in the following aspects:

– The router with the priority of 0 also takes part in the DIS election.

– When a new router that meets the requirements of being a DIS is added to the broadcast network, the router is selected as the new DIS, which triggers a new round of LSP flooding.

- Establishment of a neighbor relationship on a P2P link

  The establishment of a neighbor relationship on a P2P link is different from that on a broadcast link. A neighbor relationship on a P2P link can be established in 2-way or 3-way mode, as shown in **Table 11-22**. By default, the 3-way handshake mechanism is used to establish a neighbor relationship on a P2P link.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-22 Comparison between 2-way mode and 3-way mode

Mode	Description	Advantages and Disadvantages	Reliability
2-way mode	When a router receives an IIH, it unidirectionally sets up a neighbor relationship.	Disadvantages: • The unstable link status causes the loss of **complete sequence numbers protocol data units (CSNPs)** that are sent once an adjacency is set up. As a result, the link state databases (LSDBs) of two neighboring routers are not synchronized during the LSP update period. • If two or more links exist between two routers, an adjacency can still be set up when one link is Down and another is Up in the same direction. A router that fails to detect the faulty link may also forward packets over this link.	Low

 Copyright © Huawei Technologies Co., Ltd.

Mode	Description	Advantages and Disadvantages	Reliability
3-way mode	A neighbor relationship is established after IIHs are sent three times.	Advantages: A neighbor relationship is established only when both ends are Up. This mechanism ensures that packets are transmitted securely.	High

IS-IS Strict Neighbor Check

In IS-IS, IPv4 and IPv6 share a standard topology. When both IPv4 and IPv6 are deployed in a standard topology, a problem can occur: if only IPv4 goes Up, IPv6 traffic is interrupted; if only IPv6 goes Up, IPv4 traffic is interrupted. Another problem occurs when a primary link recovers after a switchover occurs between primary and backup links. In this scenario, IPv4 traffic is switched to the primary link first because IPv4 goes Up faster than IPv6. IPv6 is also switched but discarded because IPv6 links are not Up. IS-IS strict neighbor check can resolve these problems.

IS-IS must check the following items when establishing a neighbor relationship:

- Whether the address family configured on the local router is the same as that carried in a received IIH. The address family status of neighboring routers is set to Up only when both are the same.
- Whether the address family configured in the standard topology is Up in the case of establishing a standard topology neighbor relationship identified by MT0, which is carried in an IIH. The address family configured in the standard topology goes Up only when all address families go Up.

LSDB Synchronization

IS-IS is a link-state protocol. An IS-IS router obtains first-hand information from other routers running link-state protocols. Every router generates information about itself, directly connected networks, and links between itself and directly connected networks. The router then sends the generated information to other routers through adjacent routers. Every router saves link state information without modifying it. Finally, every router has the same network interworking information, and LSDB synchronization is complete. The process of synchronizing LSDBs is called LSP flooding. In LSP flooding, a router sends an LSP to its neighbors and the neighbors send the received LSP to their neighbors except the router that first sends the LSP. The LSP is flooded among the routers at the same level. This implementation allows each router at the same level to have the same LSP information and keep a synchronized LSDB.

All routers in the IS-IS routing domain can generate LSPs. The following events trigger the generation of a new LSP:

 Copyright © Huawei Technologies Co., Ltd.

- Neighbor goes Up or Down.
- related interface goes Up or Down.
- Imported IP routes change.
- Inter-area IP routes change.
- A new metric value is configured for an interface.
- Periodic updates occur.

A router processes a received LSP as follows:

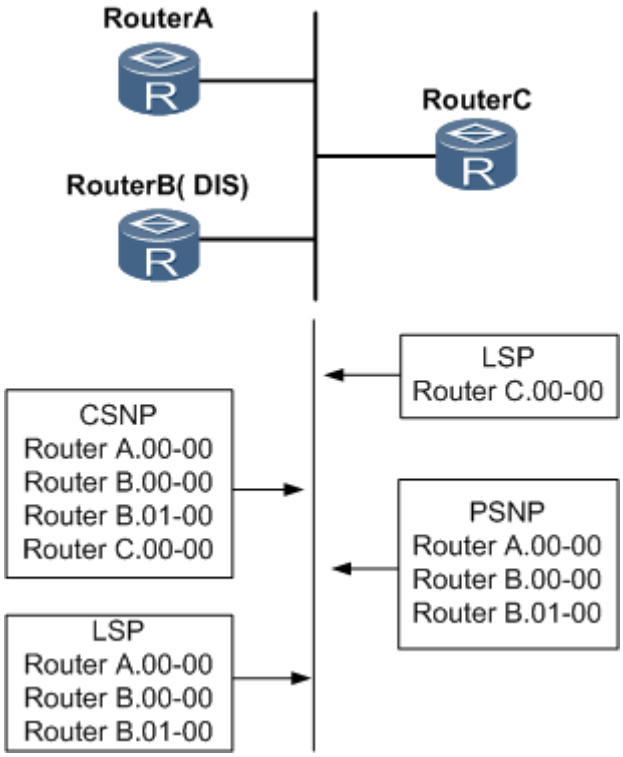
- Updating the LSDB on a broadcast link

  The DIS updates the LSDB to synchronize LSDBs on a broadcast network. **Figure 11-91** shows the process of synchronizing LSDBs on a broadcast network.

  a. When the DIS receives an LSP, it searches the LSDB for the related records. If the DIS does not find the LSP in its LSDB, it adds the LSP to its LSDB and broadcasts the new LSDB.
  b. If the sequence number of the received LSP is greater than that of the local LSP, the DIS replaces the local LSP with the received LSP in the LSDB and broadcasts the new LSDB.
  c. If the sequence number of the received LSP is less than that of the local LSP, the DIS sends the local LSP in the LSDB to the inbound interface.
  d. If the sequence number of the received LSP is equal to that of the local LSP, the DIS compares the Remaining Lifetime of the two LSPs. If Remaining Lifetime of the received LSP is 0, the DIS replaces the LSP with the received LSP, and broadcasts the new LSDB. If the Remaining Lifetime of local LSP is 0, the DIS sends the LSP to the inbound interface.
  e. If the sequence number of the received LSP and the local LSP in the LSDB are the same and neither Remaining Lifetime is 0, the DIS compares the checksum of the two LSPs. If the received LSP has a greater checksum than that of the local LSP in the LSDB, the DIS replaces the local LSP in the LSDB with the received LSP and advertises the new LSDB. If the received LSP has a smaller checksum than that of the local LSP in the LSDB, the DIS sends the local LSP in the LSDB to the inbound interface.
  f. If the checksums of the received LSP and the local LSP are the same, the LSP is not forwarded.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-91 Process of updating the LSDB on a broadcast link



- Updating the LSDB on a P2P link
  a. If the sequence number of the received LSP is greater than that of the local LSP in the LSDB, the router adds the received LSP to its LSDB. The router then sends a **PSNP packet** to acknowledge the received LSP and sends the LSP to all its neighbors except the neighbor that sends the LSP.
  b. If the sequence number of the received LSP is less than that of the local LSP, the router directly sends its LSP to the neighbor and waits for a PSNP from the neighbor as an acknowledgement.
  c. If the sequence number of the received LSP is the same as that of the local LSP in the LSDB, the router compares the Remaining Lifetimes of the two LSPs. If Remaining Lifetime of the received LSP is 0, the router adds the LSP to its LSDB. The router then sends a PSNP to acknowledge the received LSP. If Remaining Lifetime of the local LSP is 0, the router directly sends the local LSP to the neighbor and waits for a PSNP from the neighbor.
  d. If the sequence number of the received LSP and the local LSP in the LSDB are the same, and neither Remaining Lifetime is 0, the router compares the checksum of the two LSPs. If the received LSP has a greater checksum than that of the local LSP, the router adds the received LSP to its LSDB. The router then sends a PSNP to acknowledge the received LSP. If the received LSP has a smaller checksum than that of the local LSP, the router directly sends the local LSP to the neighbor and waits for a PSNP from the neighbor. At last, the router sends the LSP to all its neighbors except the neighbor that sends the LSP.
  e. If the checksums of the received LSP and the local LSP are the same, the LSP is not forwarded.

Route Calculation

 Copyright © Huawei Technologies Co., Ltd.

When LSDB synchronization is complete and network convergence is implemented, IS-IS performs SPF calculation by using LSDB information to obtain the SPT. IS-IS uses the SPT to create a forwarding database (a routing table).

In IS-IS, link costs are used to calculate shortest paths. The default cost for an interface on a Huawei router is 10. The cost is configurable. The cost of a route is the sum of the cost of every outbound interface along the route. There may be multiple routes to a destination, among which the route with the smallest cost is the optimal route.

Level-1 routers can also calculate the shortest path to Level-2 routers to implement inter-area route selection. When a Level-1-2 router is connected to other areas, the router sets the value of the attachment (ATT) bit in its LSP to 1 and send the LSP to neighboring routers. In the route calculation process, a Level-1 router selects the nearest Level-1-2 router as an intermediate router between the Level-1 and Level-2 areas.

11.12.5.2.3 IS-IS Routing Information Control

IS-IS routes calculated using the shortest path first (SPF) algorithm may bring about some problems. For example, too many routing entries slow down route lookup, or link usage is unbalanced. As a result, IS-IS routing cannot meet carriers' network planning and traffic management requirements.

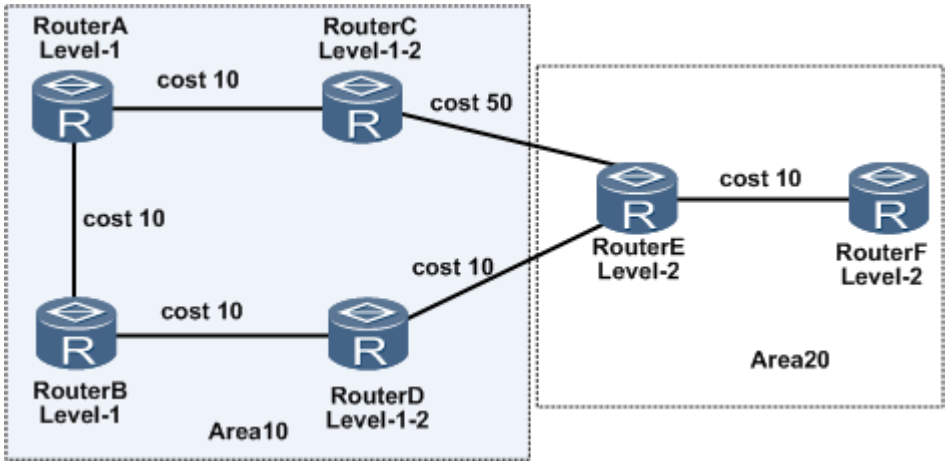
IS-IS routing information control can refine control over route selection using the following methods.

Route Leaking

When Level-1 and Level-2 areas both exist on an IS-IS network, Level-2 routers do not advertise the learned routing information about a Level-1 area and the backbone area to any other Level-1 area by default. Therefore, Level-1 routers do not know the routing information beyond the local area. As a result, the Level-1 routers cannot select the optimal routes to the destination beyond the local area.

With route leaking, Level-1-2 routers can select routes using ACLs, routing policies, or tags and advertise the selected routes of other Level-1 areas and the backbone area to the Level-1 area. **Figure 11-92** shows the typical networking for route leaking.

Figure 11-92 Typical networking for route leaking



 Copyright © Huawei Technologies Co., Ltd.

- Router A, Router B, Router C, and Router D belong to area 10. Router A and Router B are Level-1 routers. Router C and Router D are Level-1-2 routers.
- Router E and Router F belong to area 20 and are Level-2 routers.

If Router A sends a packet to Router F, the selected optimal route should be Router A -> Router B -> Router D -> Router E -> Router F because its cost is 40 (10 + 10 + 10 + 10 = 40) which is less than that of Router A -> Router C -> Router E -> Router F (10 + 50 + 10 = 70). However, if you check routes on Router A, you can find that the selected route is Router A -> Router C -> Router E -> Router F, which is not the optimal route from Router A to Router F.

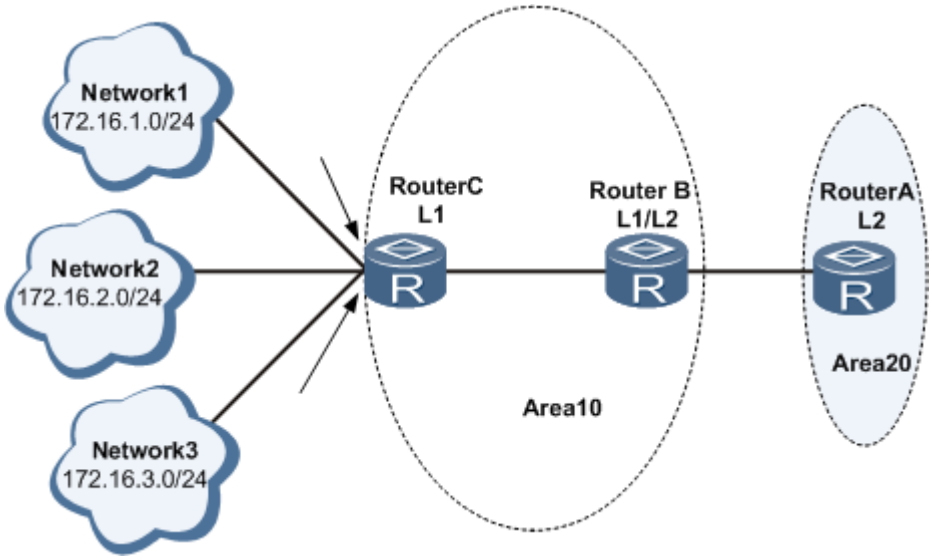
This is because Router A does not know the routes beyond the local area, and therefore, the packets sent by Router A to other network segments are sent through the default route generated by the nearest Level-1-2 device.

In this case, you can enable route leaking on the Level-1-2 devices (Router C and Router D). Then, check the route and you can find that the selected route is Router A -> Router B -> Router D -> Router E -> Router F.

Route Summarization

On a large-scale IS-IS network, links connected to devices within an IP address range may alternate between Up and Down. With route summarization, multiple routes with the same IP prefix are summarized into one route, which prevents route flapping, reduces routing entries and system resource consumption, and facilitates route management. **Figure 11-93** shows the typical networking for route summarization.

Figure 11-93 Typical networking for route summarization



- Router A, Router B, and Router C use IS-IS to communicate with each other.
- Router A belongs to area 20, and Router B and Router C belong to area 10.
- Router A is a Level-2 router. Router B is a Level-1-2 router. Router C is a Level-1 router.

 Copyright © Huawei Technologies Co., Ltd.

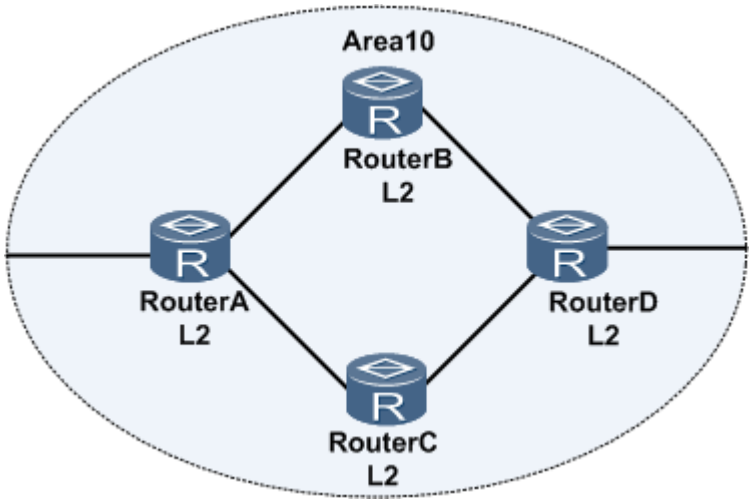
- Router B maintains Level-1 and Level-2 LSDBs and leaks the routes to three network segments (172.16.1.0/24, 172.16.2.0/24, and 172.16.3.0/24) from the Level-1 area to the Level-2 area. If a link fault causes the Router C interface with IP address 172.16.1.1/24 to frequently alternate between Up and Down, the status change is advertised to the Level-2 area, triggering frequent LSP flooding and SPF calculation on Router A. As a result, the CPU usage on Router A increases, and even network flapping occurs.

  On Router B, routes to the three network segments in the Level-1 area are summarized to one route to 172.16.0.0/16, which reduces the number of routing entries on Router B and minimizes the impact of route flapping in the Level-1 area on route convergence in the Level-2 area.

Load Balancing

If multiple equal-cost routes are available on a network, they can load-balance traffic, which improves link usage and prevents network congestion caused by link overload. **Figure 11-94** shows the typical networking for load balancing.

Figure 11-94 Typical networking for load balancing



- Router A, Router B, Router C, and Router D communicate with each other on an IP network using IS-IS.
- Router A, Router B, Router C, and Router D belong to area 10 and are Level-2 routers.
- If load balancing is not enabled, traffic on Router A is transmitted along the optimal route obtained using the SPF calculation. Consequently, traffic on different links is unbalanced. Enabling load balancing on Router A sends traffic to RouterRouter D through RouterRouter B and Router C. This transmission mode relieves the load on the optimal route.

Load balancing can work per-destination or per-packet. For details, see the section "Overview" in *MA5800 Feature Description - IP Routing*.

Administrative Tag

Administrative tags carry administrative information about IP address prefixes. When the cost type is wide, wide-compatible, or compatible and the prefix of the

 Copyright © Huawei Technologies Co., Ltd.

reachable IP address to be advertised by IS-IS has this cost type, IS-IS adds the administrative tag to the reachability type-length-value (TLV) in the prefix. In this manner, the administrative tag is advertised throughout the entire IS-IS area so that routes can be imported or filtered based on the administrative tag.

IS-IS Mesh Group

As defined in IS-IS, a router must flood the received LSP to all neighbors. On a network with multiple connections and point-to-point (P2P) links, this flooding method causes repeated LSP flooding and wastes bandwidth resources.

To address this issue, you can add certain interfaces to a mesh group. These interfaces flood the LSPs received from a group only to interfaces of other groups or interfaces no which no mesh groups are configured. All the interfaces that join a mesh group ensure the synchronization of the LSDBs in the entire network segment using the CSNP and PSNP mechanisms.

11.12.5.2.4 IS-IS Overload

The overload (OL) field of LSPs configured on a device prevents other devices from calculating the routes passing through this device.

If a system fails to store new LSPs for LSDB synchronization, the routes calculated by the system are incorrect. In that case, the system enters the Overload state. The user can configure the device to enter the Overload state when the system lacks sufficient memory. At present, users can set the Overload timer when IS-IS is started and configure whether to delete the leaked routes and whether to advertise the imported routes. A device enters the Overload state after an exception occurs on the device or when it is configured to enter the state.

- If IS-IS enters the Overload state after an exception occurs on the device, the system deletes all imported or leaked routes.
- If IS-IS enters the Overload state based on a user configuration, the system only deletes all imported or leaked routes if configured to do so.

Although LSPs with overload fields are flooded throughout the network, they are ignored in the calculation of the routes passing through the device in the Overload state. Specifically, after the overload field of LSPs is configured on a device, other devices do not count the routes that pass through the device when performing SPF calculation, but the direct routes between the device and other devices are still calculated.

If a device in an IS-IS domain is faulty, routes may be incorrectly calculated across the entire domain. The overload field can be configured for the device to isolate it from the IS-IS network temporarily, which facilitates fault isolation.

11.12.5.2.5 IS-IS Fast Convergence

IS-IS fast convergence is an extended feature of IS-IS implemented to speed up route convergence. It includes the following concepts:

- I-SPF

  Incremental SPF (I-SPF) recalculates only the routes of the changed nodes rather than the routes of all nodes when the network topology changes, which speeds up the calculation of routes.

 Copyright © Huawei Technologies Co., Ltd.

- PRC

  Partial Route Calculation (PRC) calculates only those routes which have changed when the network topology changes.
- LSP fast flooding

  LSP fast flooding speeds up LSP flooding.
- Intelligent timer

  The first timeout period of the timer is fixed. If an event that triggers the timer occurs before the set timer expires, the next timeout period of the timer increases.

  The intelligent timer is applicable to LSP generation and SPF calculation.

I-SPF

In ISO 10589, the Dijkstra algorithm was adopted to calculate routes. When a node changes on the network, the algorithm recalculates all routes. The calculation requires a long time to complete and consumes a significant amount of CPU resources, reducing convergence speed.

I-SPF improves the algorithm. Except for the first time the algorithm is run, only the nodes that have changed rather than all nodes in the network are used in the calculation. The SPT generated using I-SPF is the same as that generated using the previous algorithm. This significantly decreases CPU usage and speeds up network convergence.

PRC

Similar to I-SPF, PRC calculates only routes that have changed. PRC, however, does not calculate the shortest path. It updates routes based on the SPT calculated by I-SPF.

In route calculation, a leaf represents a route, and a node represents a device. If the SPT changes after I-SPF calculation, PRC calculates all the leaves only on the changed node. If the SPT remains unchanged, PRC calculates only the changed leaves.

For example, if IS-IS is enabled on an interface of a node, the SPT calculated by I-SPF remains unchanged. In this case, PRC updates only the routes of this interface, which consumes less CPU resources.

PRC working with I-SPF further improves network convergence performance and replaces the original SPF algorithm.

NOTE

On the MA5800, only I-SPF and PRC are used to calculate IS-IS routes.

LSP Fast Flooding

When an IS-IS device receives new LSPs from other devices, it updates the LSPs in the LSDB and periodically floods the updated LSPs based on a timer. Therefore, the synchronization of all LSDBs is slow.

With lSP fast flooding, when the Router receives LSPs that can trigger route calculation or route update, it floods these LSPs before route calculation occurs,

 Copyright © Huawei Technologies Co., Ltd.

which speeds up network convergence and LSDB synchronization throughout the entire network.

Intelligent Timer

Although the route calculation algorithm is improved, the long interval for triggering route calculation also affects the convergence speed. A millisecond-level timer can shorten the interval. Frequent network changes, however, also consume too much CPU resources. The SPF intelligent timer addresses these problems.

In most cases, an IS-IS network running normally is stable. The frequent changes on a network are rather rare, and IS-IS does not calculate routes frequently. Therefore, a short period (within milliseconds) can be configured as the first interval for route calculation. If the network topology changes frequently, the interval set by the intelligent timer increases with the calculation times to reduce CPU consumption.

The delay for the first SPF calculation is *init-interval*; the delay for the second SPF calculation is *incr-interval*. From the third time on, the delay doubles each time until the delay reaches *max-interval*. After the delay remains *max-interval* three times or the IS-IS process is restarted, the delay decreases to *init-interval*.

The LSP generation intelligent timer is similar to the SPF intelligent timer. When the LSP generation intelligent timer expires, the system generates a new LSP based on the current topology. The original mechanism uses a timer with fixed intervals, which results in slow convergence and high CPU consumption. Therefore, the LSP generation timer is designed as an intelligent timer to respond to emergencies (for example, the interface goes Up or Down) quickly and speed up network convergence. In addition, when the network changes frequently, the interval for the intelligent timer becomes longer to reduce CPU consumption.

11.12.5.2.6 IS-IS LSP Fragment Extension

When the LSPs to be advertised by IS-IS contain a large amount of information, they are advertised in multiple LSP fragments belonging to the same system.

As defined in RFC 3786, virtual system IDs can be configured, and virtual LSPs that carry routing information can be generated for IS-IS.

IS-IS LSP fragment extension allows an IS-IS device to generate more LSP fragments and carry more IS-IS information.

Terms

- Originating system

  The originating system is a device that runs the IS-IS protocol. A single IS-IS process advertises LSPs as virtual devices do, except that the originating system refers to a real IS-IS process.
- Normal system ID

  The normal system ID is the system ID of the originating system.
- Additional system ID

  The additional system ID, assigned by the network administrator, is used to generate additional or extended LSP fragments. A maximum of 256 additional or extended LSP fragments can be generated. Like a normal system ID, an additional system ID must be unique in a routing domain.

 Copyright © Huawei Technologies Co., Ltd.

- Virtual system

  The virtual system, identified by an additional system ID, is used to generate extended LSP fragments. These fragments carry additional system IDs in their LSP IDs.

Principles

IS-IS LSP fragments are identified by the LSP Number field in their LSP IDs. The LSP Number field is 1 byte. Therefore, an IS-ISprocess can generate a maximum of 256 fragments. A 1497-byte LSPcan carry about 30,000 routes. With fragment extension, more information can be carried.

Each system ID represents a virtual system, and each virtual system can generate 256 LSP fragments. In addition, a maximum of another 99 virtual systems can be configured. Therefore, an IS-IS process can generate a maximum of 25,600 LSP fragments.

After a virtual system and fragment extension are configured, an IS-IS device adds the contents that cannot be contained in its LSPs to the LSPs of the virtual system and notifies other devices of the relationship between the virtual system and itself through a special TLV in the LSPs.

IS Alias ID TLV

RFC 3786 defines a special TLV: IS Alias ID TLV.

Table 11-23 IS Alias ID TLV

Field	Length	Description
Type	1 byte	TLV type. If the value is 24, it indicates the IS Alias ID TLV.
Length	1 byte	TLV length.
System ID	6 bytes	System ID.
Pseudonode number	1 byte	Pseudonode number.
sub-TLVs length	1 byte	Length of sub-TLVs.
sub-TLVs	0 to 247 bytes	Sub-TLVs.

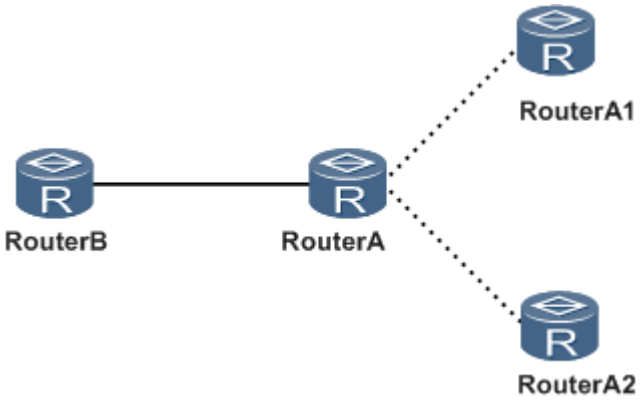
LSPs with fragment number 0 sent by the originating system and virtual system carry IS Alias ID TLVs to indicate the originating system.

Operation Modes

IS-IS devices can use the LSP fragment extension feature in the following modes:

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-95 Networking for IS-IS LSP fragment extension



- Mode-1

  Mode-1 is used when some devices on the network do not support LSP fragment extension.

  In this mode, virtual systems participate in SPF calculation. The originating system advertises LSPs containing information about links to each virtual system and each virtual system advertises LSPs containing information about links to the originating system. In this manner, the virtual systems function the same as the actual devices connected to the originating system on the network.

  Mode-1 is a transitional mode for earlier versions that do not support LSP fragment extension. In the earlier versions, IS-IS cannot identify Alias ID TLVs. Therefore, the LSP sent by a virtual system must look like a common IS-IS LSP.

  The LSP sent by a virtual system contains the same area address and overload bit as those in the common LSP. If the LSPs sent by a virtual system contain TLVs specified in other features, the TLVs must be the same as those in common LSPs.

  LSPs sent by a virtual system carry information of the neighbor (the originating system), and the carried metric is the maximum value minus 1. LSPs sent by the originating system carry information of the neighbor (the virtual system), and the carried metric is 0. This mechanism ensures that the virtual system is a node downstream of the originating system when other devices calculate routes.

  As shown in **Figure 11-95**, Router B does not support LSP fragment extension; Router A supports LSP fragment extension in mode-1; Router A1 and Router A2 are virtual systems of Router A. Router A1 and Router A2 send LSPs carrying partial routing information of Router A. After receiving LSPs from Router A, Router A1, and Router A2, Router B considers there to be three devices at the peer end and calculates routes normally. Because the cost of the route from Router A to Router A1 or Router A2 is 0, the cost of the route from Router B to Router A is equal to that from Router B to Router A1.

- Mode-2

  Mode-2 is used when all the devices on the network support LSP fragment extension. In this mode, virtual systems do not participate in SPF calculation. All the devices on the network know that the LSPs generated by the virtual systems actually belong to the originating system.

 Copyright © Huawei Technologies Co., Ltd.

IS-IS working in mode-2 identifies IS Alias ID TLVs, which are used to calculate the SPT and routes.

As shown in **Figure 11-95**, Router B supports LSP fragment extension, and Router A supports LSP fragment extension in mode-2; Router A1 and Router A2 send LSPs carrying some routing information of Router A. After receiving LSPs from Router A1 and Router A2, Router B obtains IS Alias ID TLV and learns that the originating system of Router A1 and Router A2 is Router A. Router B then considers information advertised by Router A1 and Router A2 to be about Router A.

Whatever the LSP fragment extension mode, LSPs can be resolved. However, if LSP fragment extension is not supported, only LSPs in mode-1 can be resolved.

Table 11-24 Comparison between mode-1 and mode-2

LSP Field	Carried in Mode-1	Carried in Mode-2
IS Alias ID	Yes	Yes
Area	Yes	No
Overload bit	Yes	Yes
IS NBR/IS EXTENDED NBR	Yes	No
Routing	Yes	Yes
ATT bit	Yes, with value 0	Yes, with value 0
P bit	Yes, with value 0	Yes, with value 0

Process

After LSP fragment extension is configured, if information is lost because LSPs overflow, the system restarts the IS-IS process. After being restarted, the originating system loads as much routing information as possible. Any excessive information beyond the forwarding capability of the system is added to the LSPs of the virtual systems for transmission. In addition, if a virtual system with routing information is deleted, the system automatically restarts the IS-IS process.

Usage Scenario

NOTE

If there are non-Huawei devices on the network, LSP fragment extension must be set to mode-1. Otherwise, these devices cannot identify LSPs.

Configuring LSP fragment extension and virtual systems before setting up IS-IS neighbors or importing routes is recommended. If IS-IS neighbors are set up or routes are imported first and the information to be carried exceeds the forwarding capability of 256 fragments before LSP fragment extension and virtual systems are configured, you have to restart the IS-IS process for the configurations to take effect.

 Copyright © Huawei Technologies Co., Ltd.

11.12.5.2.7 IS-IS 3-Way Handshake

IS-IS introduces the 3-way handshake mechanism on P2P links to ensure a reliable data link layer.

Based on ISO 10589, the IS-IS 2-way handshake mechanism uses Hello packets to set up P2P adjacencies between neighboring devices. When a device receives a Hello packet from the other end, it regards the other end as Up and sets up an adjacency with it. However, this mechanism has some serious shortcomings.

When two or more links exist between two devices, an adjacency can still be set up where one link is Down and the other is Up in the same direction. The parameters of the other link are used in SPF calculation. As a result, a device that does not detect any fault along the faulty link will continue trying to forward packets over the link.

The 3-way handshake mechanism resolves these problems on P2P links. In 3-way handshake mode, a device regards a neighbor Up and sets up an adjacency with it only after confirming that the neighbor has received the packet that the device sends.

In addition, the 3-way handshake mechanism uses the 32-bit Extended Local Circuit ID field, which extends the original 8-bit Extended Local Circuit ID field and the limit of only 255 P2P links.

NOTE

By default, the IS-IS 3-way handshake mechanism is implemented on P2P links, as defined in RFC 3373.

11.12.5.2.8 IS-IS NSR

Background

As networks develop, the demand for data, audio, and video services is growing, which imposes increasing requirements on IP network reliability. If an AMB/SMB switchover is performed on a device due to a maintenance operation or a single point of failure, routes may fail to converge, which may result in traffic loss or even a network breakdown. Non-stop routing (NSR) can address this problem and ensure uninterrupted forwarding of key services.

Related Concepts

- High availability (HA): supports data backup between the AMB and SMB.
- Non-stop forwarding (NSF): enables a node to use the GR mechanism to ensure uninterrupted transmission during an AMB/SMB switchover.
- NSR: allows a standby control plane to take over traffic from an active control plane if the active control plane fails, preventing the control planes of neighbors from detecting the fault.
- AMB and SMB: run control plane processes.

Implementation

With IS-IS NSR, IS-IS real-time data is synchronized between the AMB and SMB. After an AMB/SMB switchover is performed on a device, the SMB takes over

 Copyright © Huawei Technologies Co., Ltd.

services from the AMB, and neighbors are unaware of the local fault. After the switchover, the new AMB recovers IS-IS immediately based on the synchronized IS-IS real-time data. Therefore, neighbors are unaware of the switchover as well. IS-IS NSR requires synchronization of the following data:

- All configuration data, such as information about neighbors, timer parameters, and process configurations.
- Dynamic data, such as the interface parameters and state, and information about neighbors and the link state database (LSDB).

Usage Scenario

NSR minimizes the impact of control plane faults and prevents route flapping on networks that require high reliability.

Benefits

NSR improves network reliability and ensures uninterrupted traffic forwarding.

11.12.5.2.9 IS-IS Wide Metric

In the earlier ISO 10589, the largest metric of an interface is 63. TLV type 128 and TLV type 130 contain information about routes, and TLV type 2 contains information about IS-IS neighbors. However, on large-scale networks, the metric range cannot meet the requirements.

As defined in RFC 3784, with IS-IS wide metric, the largest metric of an interface is extended to 16777215, and the largest metric of a route is 4261412864.

After IS-IS wide metric is enabled, TLV type 135 contains information about routes; TLV type 22 contains information about IS-IS neighbors.

- The following lists the TLVs used in narrow mode:
  - IP Internal Reachability TLV: carries routes within an area.
  - IP External Reachability TLV: carries routes outside an area.
  - IS Neighbors TLV: carries information about neighbors.
- The following lists the TLVs used in wide mode:
  - Extended IP Reachability TLV: replaces the earlier IP Reachability TLV and carries information about routes. This TLV expands the range of the route cost to 4 bytes and carries sub-TLVs.
  - IS Extended Neighbors TLV: carries information about neighbors.

NOTE

The metric style can be set to narrow, narrow-compatible, compatible, wide-compatible, or wide mode. **Table 11-25** shows which metric styles are carried in received and sent packets. A device can calculate routes only when it can receive, send, and process corresponding TLVs. Therefore, to ensure correct data forwarding on a network, the proper metric style must be configured for each device on the network.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-25 Metric style carried in received and sent under different metric style configurations

Configured Metric Style	Metric Style Carried in Received Packets	Metric Style Carried in Sent Packets
Narrow	Narrow	Narrow
Narrow-compatible	Narrow and wide	Narrow
Compatible	Narrow and wide	Narrow and wide
Wide-compatible	Narrow and wide	Wide
Wide	Wide	Wide

When the metric style is set to compatible, IS-IS sends the information both in narrow and wide modes.

Process

NOTICE

Once the metric style is changed, the IS-IS process restarts.

- If the metric style carried in sent packets is changed from narrow to wide:

  The information previously carried by TLV type 128, TLV type 130, and TLV type 2 is now carried by TLV type 135 and TLV type 22.
- If the metric style carried in sent packets is changed from wide to narrow:

  The information previously carried by TLV type 135 and TLV type 22 is now carried by TLV type 128, TLV type 130, and TLV type 2.
- If the metric style carried in sent packets is changed from narrow or wide to narrow and wide:

  The information previously carried in narrow or wide mode is now carried by TLV type 128, TLV type 130, TLV type 2, TLV type 135, and TLV type 22.

11.12.5.2.10 IS-IS Authentication

Background

As the Internet develops, more data, voice, and video information are exchanged over the Internet. New services, such as e-commerce, online conferencing and auctions, video on demand, and distance learning, emerge gradually. The new services have high requirements for network security. Carriers need to prevent data packets from being intercepted or modified by attackers or unauthorized users. IS-IS authentication applies to the area or interface where packets need to be protected. Using IS-IS authentication enhances system security and helps carriers provide safe network services.

 Copyright © Huawei Technologies Co., Ltd.

Related Concepts

Authentication Classification

Based on packet types, the authentication is classified as follows:

- Interface authentication: is configured in the interface view to authenticate Level-1 and Level-2 IS-to-IS Hello PDUs (IIHs).
- Area authentication: is configured in the IS-IS process view to authenticate Level-1 CSNPs, PSNPs, and LSPs.
- Routing domain authentication: is configured in the IS-IS process view to authenticate Level-2 CSNPS, PSNPs, and LSPs.

Based on the authentication modes of packets, the authentication is classified into the following types:

- Simple authentication: The authenticated party directly adds the configured password to packets for authentication. This authentication mode provides the lowest password security.
- MD5 authentication: uses the MD5 algorithm to encrypt a password before adding the password to the packet, which improves password security.

NOTICE

MD5 is an insecure encryption algorithm. Therefore, using HMAC-SHA256 or Keychain is recommended.

- Keychain authentication: further improves network security with configurable key chain that changes with time.
- HMAC-SHA256 authentication: uses the HMAC-SHA256 algorithm to encrypt a password before adding the password to the packet, which improves password security.

Implementation

IS-IS authentication encrypts IS-IS packets by adding the authentication field to packets to ensure network security. After receiving IS-IS packets from a remote router, a local router discards the packets if the authentication passwords in the packets are different from the locally configured one. This mechanism protects the local router.

IS-IS provides a type-length-value (TLV) to carry authentication information. The TLV components are as follows:

- Type: indicates the type of a packet, which is 1 byte. The value defined by ISO is 10, while the value defined by IP is 133.
- Length: indicates the length of the authentication TLV, which is 1 byte.
- Value: indicates the authentication information, including authentication type and authenticated password, which ranges from 1 to 254 bytes. The authentication type is 1 byte:
  - 0: reserved

 Copyright © Huawei Technologies Co., Ltd.

- 1: simple authentication
- 3: general authentication, and only HMAC-SHA256 authentication currently
- 54: MD5 authentication
- 255: private authentication

Interface Authentication

Authentication passwords for IIHs are saved on interfaces. The interfaces send authentication packets with the authentication TLV. Interconnected router interfaces must be configured with the same password.

Area Authentication

Every router in an IS-IS area must use the same authentication mode and have the same key chain.

Routing Domain Authentication

Every Level-2 or Level-1-2 router in an IS-IS area must use the same authentication mode and have the same key chain.

For area authentication and routing domain authentication, you can set a router to authenticate SNPs and LSPs separately in the following ways:

- A router sends LSPs and SNPs that carry the authentication TLV and verifies the authentication information of the LSPs and SNPs it receives.
- A router sends LSPs that carry the authentication TLV and verifies the authentication information of the LSPs it receives. The router sends SNPs that carry the authentication TLV and does not verify the authentication information of the SNPs it receives.
- A router sends LSPs that carry the authentication TLV and verifies the authentication information of the LSPs it receives. The router sends SNPs without the authentication TLV and does not verify the authentication information of the SNPs it receives.
- A router sends LSPs and SNPs that carry the authentication TLV but does not verify the authentication information of the LSPs and SNPs it receives.

11.12.5.2.11 IS-IS Control Messages

IS-IS routers implement routing by exchanging control messages. This section describes IS-IS control messages.

IS-IS PDU Formats

Nine types of IS-IS protocol data units (PDUs) are available for processing control information. Each PDU is identified by a 5-digit type code. IS-IS has three major types of PDUs: Hello PDUs, Link State PDUs (LSPs), and Sequence Number PDUs (SNPs). **Table 11-26** shows the mapping between PDUs and type values.

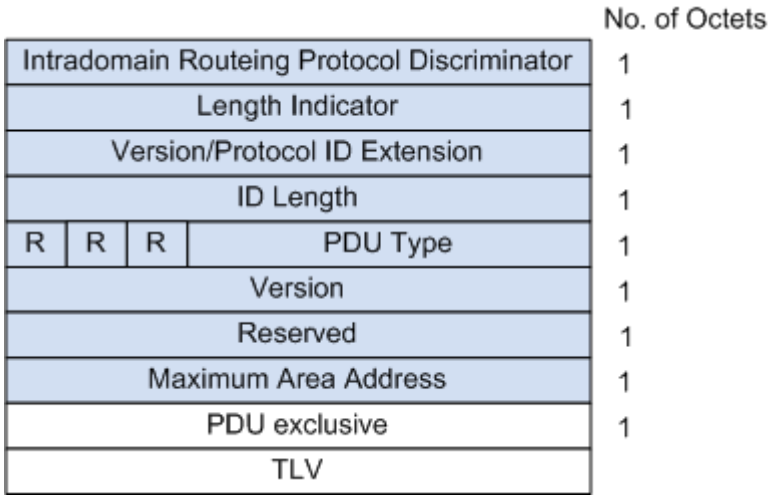
 Copyright © Huawei Technologies Co., Ltd.

Table 11-26 Mapping between PDUs and type values

PDU Type	Acronym	Type Value
Level-1 LAN IS-IS Hello PDU	L1 LAN IIH	15
Level-2 LAN IS-IS Hello PDU	L2 LAN IIH	16
Point-to-Point IS-IS Hello PDU	P2P IIH	17
Level-1 Link State PDU	L1 LSP	18
Level-2 Link State PDU	L2 LSP	20
Level-1 Complete Sequence Numbers PDU	L1 CSNP	24
Level-2 Complete Sequence Numbers PDU	L2 CSNP	25
Level-1 Partial Sequence Numbers PDU	L1 PSNP	26
Level-2 Partial Sequence Numbers PDU	L2 PSNP	27

The first eight bytes in all IS-IS PDUs are public. **Figure 11-96** shows the IS-IS PDU format.

Figure 11-96 IS-IS PDU format



The main fields are as follows:

- Intradomain Routing Protocol Discriminator: network layer protocol identifier assigned to IS-IS, which is 0x83.
- Length Indicator: length of the fixed header, in bytes.
- ID Length: length of the system ID of network service access point (NSAP) addresses or NETs in this routing domain.
- PDU Type: type of a PDU. For details, see **Table 11-26**.

 Copyright © Huawei Technologies Co., Ltd.

- Maximum Area Address: maximum number of area addresses supported by an IS-IS area. The value **0** indicates that a maximum of three area addresses are supported by this IS-IS area.
- Type/Length/Value (TLV): encoding type that features high efficiency and expansibility. Each type of PDU contains a different TLV. **Table 11-27** shows the mapping between TLV codes and PDU types.

Table 11-27 Mapping between TLV codes and PDU types

TLV Code	TLV Code Name	PDU Type
1	Area Addresses	IIH, LSP
2	IS Neighbors (LSP)	LSP
4	Partition Designated Level2 IS	L2 LSP
6	IS Neighbors (MAC Address)	LAN IIH
7	IS Neighbors (SNPA Address)	LAN IIH
8	Padding	IIH
9	LSP Entries	SNP
10	Authentication Information	IIH, LSP, or SNP
128	IP Internal Reachability Information	LSP
129	Protocols Supported	IIH or LSP
130	IP External Reachability Information	L2 LSP
131	Inter-Domain Routing Protocol Information	L2 LSP
132	IP Interface Address	IIH or LSP

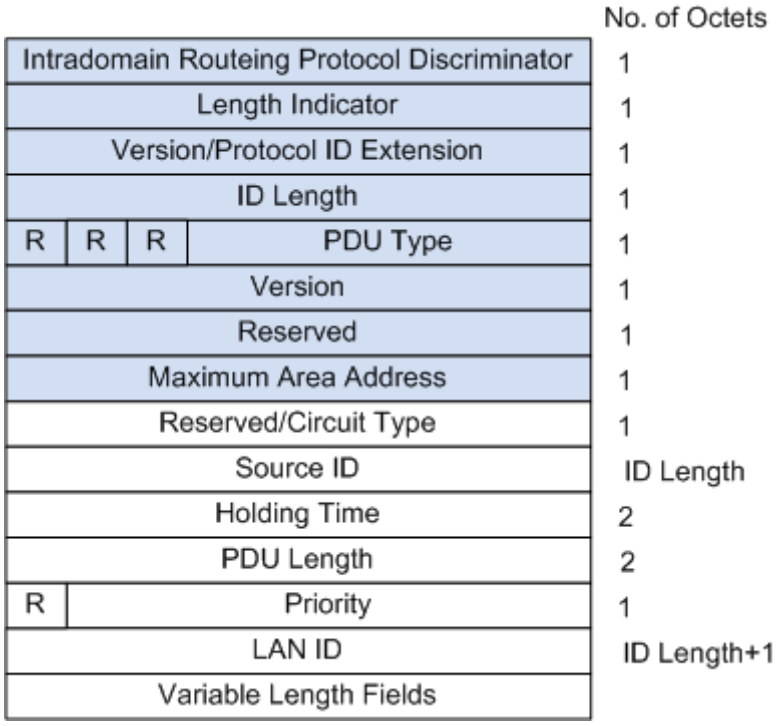
Hello Packet Format

Hello packets, also called the IS-to-IS Hello PDUs (IIHs), are used to set up and maintain neighbor relationships. Level-1 LAN IIHs are applied to the Level-1 routers on broadcast LANs. Level-2 LAN IIHs are applied to the Level-2 routers on broadcast LANs. P2P IIHs are applied to non-broadcast networks. IIHs in different networks have different formats.

- LAN IIHs: **Figure 11-97** shows the format of IIHs on a broadcast network.

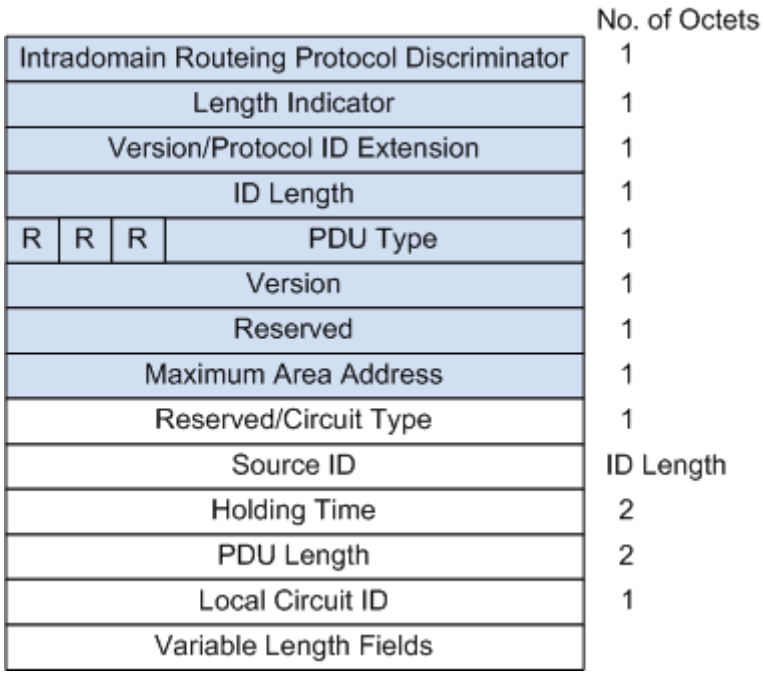
 Copyright © Huawei Technologies Co., Ltd.

Figure 11-97 Level-1/Level-2 LAN IIH format



- P2P IIHs: **Figure 11-98** shows the format of IIHs on a P2P network.

Figure 11-98 P2P IIH format



As shown in **Figure 11-98**, most fields in a P2P IIH are the same as those in a LAN IIH. The P2P IIH does not have the priority and LAN ID fields but has a local circuit ID field. The local circuit ID indicates the local link ID.

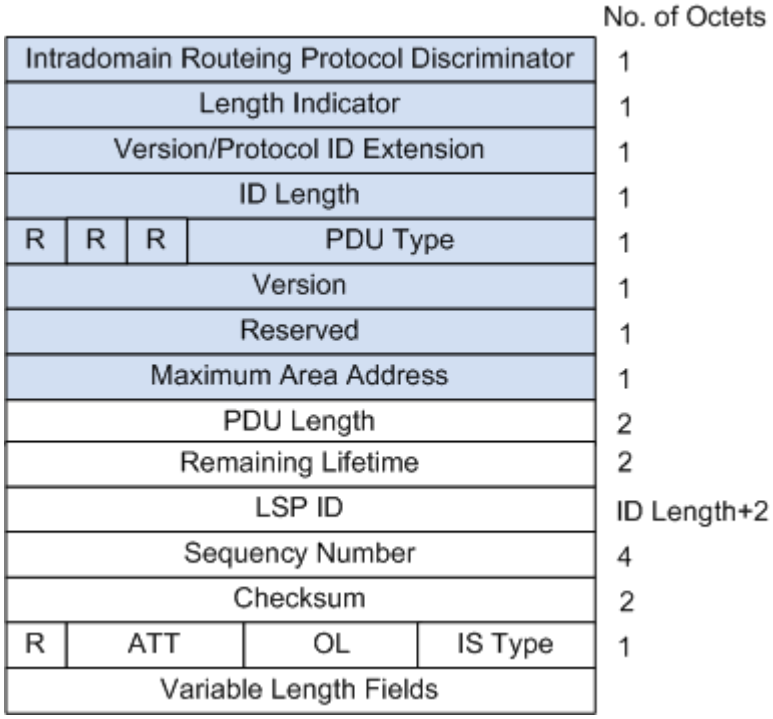
 Copyright © Huawei Technologies Co., Ltd.

LSP Format

LSPs are used to exchange link-state information. There are two types of LSPs: Level-1 and Level-2. Level-1 IS-IS transmits Level-1 LSPs. Level-2 IS-IS transmits Level-2 LSPs. Level-1-2 IS-IS can transmit both Level-1 and Level-2 LSPs.

Level-1 and Level-2 LSPs have the same format, as shown in **Figure 11-99**.

Figure 11-99 Level-1 or Level-2 LSP



The main fields are as follows:

- ATT: Attached bit

  ATT is generated by a Level-1-2 router to identify whether the originating router is connected to other areas. When a Level-1 router receives a Level-1 LSP with ATT as 1 from a Level-1-2 router, the Level-1 router generates a default route destined for the Level-1-2 router so that data can be transmitted to other areas.

  Although ATT is defined in both the Level-1 LSP and Level-2 LSP, it is set only in the Level-1 LSP only by the Level-1-2 router.

- OL: LSDB overload

  LSPs with the overload bit are still flooded on networks, but the LSPs are not used when routes that pass through a device configured with the overload bit are calculated. That is, after a device is configured with the overload bit, other devices ignore the device when performing the SPF calculation except for the direct routes of the device.

- IS Type: type of the IS-IS generating the LSP

  IS Type is used to specify whether the IS-IS type is Level-1 or Level-2 IS-IS. The value **01** indicates Level-1; the value **11** indicates Level-2.

 Copyright © Huawei Technologies Co., Ltd.

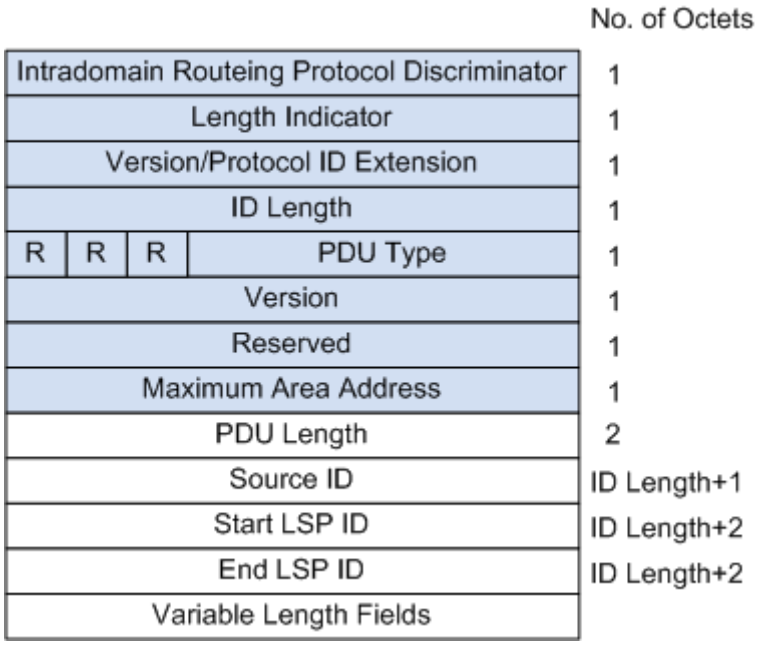
SNP Format

SNPs describe the LSPs in all or some of the databases and are used to synchronize and maintain all LSDBs. SNPs consist of complete SNPs (CSNPs) and partial SNPs (PSNPs).

- CSNPs carry summaries of all LSPs in LSDBs, which ensures LSDB synchronization between neighboring routers. On a broadcast network, the designated intermediate system (DIS) sends CSNPs at an interval. The default interval is 10 seconds. On a P2P link, neighboring devices send CSNPs only when a neighbor relationship is established for the first time.

  Figure 11-100 shows the CSNP format.

  Figure 11-100 Level-1/Level-2 CSNP format

  The main fields are as follows:

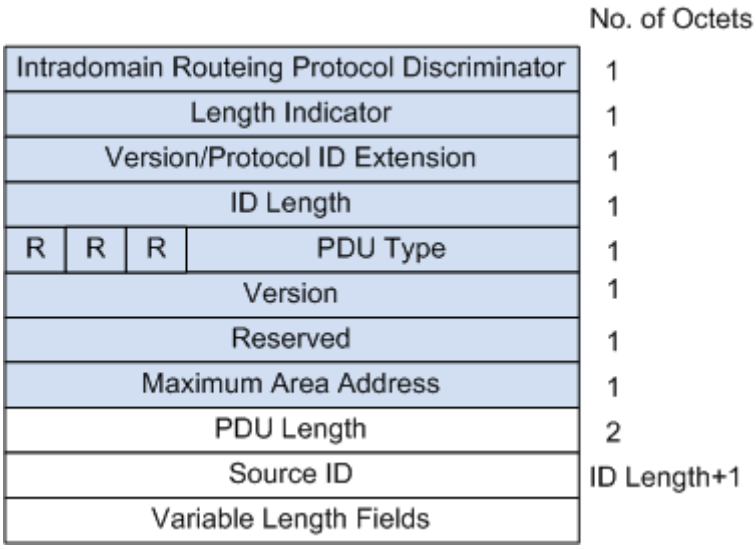
  - Source ID: system ID of the router that sends SNPs
  - Start LSP ID: ID of the first LSP in a CSNP
  - End LSP ID: ID of the last LSP in a CSNP

- PSNPs list only the sequence numbers of recently received LSPs. A PSNP can acknowledge multiple LSPs at a time. If an LSDB is not updated, PSNPs are also used to request a new LSP from a neighbor.

  Figure 11-101 shows the PSNP format.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-101 Level-1/Level-2 PSNP format



11.12.5.2.12 IS-IS for IPv6

RFC 5308 (Routing IPv6 with IS-IS) released by the IETF defines two new TLVs that can support IPv6 routes and a new Network Layer Protocol Identifier (NLPID), which ensures that IS-IS can process and calculate IPv6 routes.

The two new TLVs are as follows:

- IPv6 Reachability

  The IPv6 Reachability TLV indicates the reachability of a network by specifying the route prefix and metric. The type value is 236 (0xEC).

- IPv6 Interface Address

  The IPv6 Interface Address TLV is similar to the IP interface address TLV of IPv4 in function, except that it changes the original 32-bit IPv4 address to a 128-bit IPv6 address. The type value is 232 (0xE8).

The NLPID is an 8-bit field that identifies network layer protocol packets. The NLPID of IPv6 is 142 (0x8E). If an IS-IS router supports IPv6, it advertises routing information through the NLPID value.

11.12.5.2.13 IS-IS Auto FRR

Background

Some services, such as voice over IP (VoIP) and online video services, are delay-sensitive. Nevertheless, if a primary link fails, IS-IS must perform multiple operations, including detecting the fault, updating the link-state advertisement (LSA), flooding the LSA, calculating routes, and delivering forward information base (FIB) entries before switching traffic to a new link. This process takes a much longer time than the delay-sensitive services can tolerate. To solve this problem, IS—IS Auto FRR is developed.

IS—IS auto fast reroute (FRR) is a kind of dynamic IP FRR. To implement this feature, Interior Gateway Protocol (IGP) calculates the backup paths beforehand by using the link status database of the entire network and then stores the backup

 Copyright © Huawei Technologies Co., Ltd.

paths in the forwarding table to minimize service interruption duration in case of a link fault.

LFA

A device uses each neighbor that can be used as a backup next hop as a root node, uses the shortest path first (SPF) algorithm to calculate the shortest path from each root node to the destination, and then uses the inequalities defined in RFC 5286 to calculate the backup link that is loop-free and has the lowest cost.

The link costs meet the inequality: Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D). Distance_opt(X, Y) indicates the shortest link from X to Y. S stands for a source node, E for the faulty node, N for a node along a backup link, and D for a destination node.

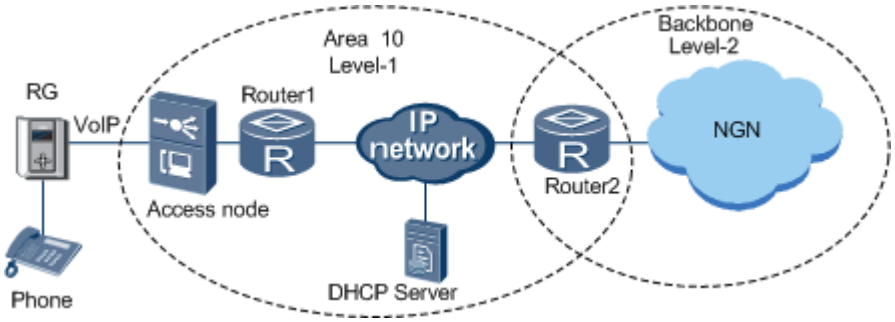
11.12.5.3 Configuration Example of IS-IS

This operation enables the corresponding device configured data to run the IS-IS protocol on the access device.

Service Requirements

- The access device forwards the access VoIP service through the Layer 3 interface to the NGN network.
- The access device obtains the routes of the NGN networking through the IS-IS protocol. The area ID of the Level-2 router differs from the area ID of the Level-1-2 router to which the Level-2 router connects.

Figure 11-102 Example network for configuring IS-IS



Data Plan

Table 11-28 provides the data plan for configuring IS-IS.

Table 11-28 Data plan for configuring IS-IS

Item	Data
Access node	IS-IS process ID: 1

 Copyright © Huawei Technologies Co., Ltd.

Item	Data
	NET (Network entity title): 10.0000.0000.0001.00, where: ● Area ID: 10 ● System ID: 0000.0000.0001 ● Level: Level-1 ● Host name: MA5800
	IS-IS interface: ● Port number: 0/9/0 ● VLAN ID: 20 ● IP address: 10.5.5.5/16
Router1	IS-IS process ID: 1
	NET (Network entity title): 10.0000.0000.0002.00, where: ● Area ID: 10 ● System ID: 0000.0000.0002 ● Level: Level-1 ● Host name: Router1
	IS-IS interface: 1/0/0 IP address: 10.8.8.8/16
Router2	IS-IS process ID: 1
	NET (Network entity title): 10.0000.0000.0005.00, where: ● Area ID: 10 ● System ID: 0000.0000.0005 ● Level: Level-1-2 ● Host name: Router2
	IS-IS interface: 1/0/0 IP address: 10.9.9.9/16

Procedure

- Configure IS-IS on the access node.
  a. Configure the Layer 3 interface.

```
huawei(config)#vlan 20 standard
huawei(config)#port vlan 20 0/9 0
huawei(config)#interface vlanif 20
huawei(config-if-vlanif20)#ip address 10.5.5.5 16
huawei(config-if-vlanif20)#quit
```

  b. Start the IS-IS process.

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config)#isis 1
huawei(config-isis-1)#
```

c. Configure the NET.

```
huawei(config-isis-1)#network-entity 10.0000.0000.0001.00
```

d. Configure the router level.

```
huawei(config-isis-1)#is-level level-1
```

e. Configure the local host name.

```
huawei(config-isis-1)#is-name MA5800
huawei(config-isis-1)#quit
```

f. Enable the IS-IS function on an interface.

```
huawei(config)#interface vlanif 20
huawei(config-if-vlanif20)#isis enable 1
```

- Configure IS-IS on Router1.

  The process of configuring IS-IS on Router1 is similar to that of configuring IS-IS on the access node. The details are not provided in this chapter.

- Configure IS-IS on Router2.

  The process of configuring IS-IS on Router2 is similar to that of configuring IS-IS on the access node. The details are not provided in this chapter.

----End

Result

- Run the **display isis lsdb** command and you can query the IS-IS LSDB.
- Run the **display isis route** command and you can query the IS-IS route. The routing table of the Level-1 router should have a default route, and the next hop should be the Level-1-2 router. The Level-2 router should have the routes to all the Level-1 routers and the Level-2 routers.

Configuration File

```
vlan 20 standard
port vlan 20 0/9 0
interface vlanif 20
ip address 10.5.5.5 16
quit
isis 1
network-entity 10.0000.0000.0001.00
is-level level-1
is-name MA5800
quit
interface vlanif 20
isis enable 1
```

11.12.5.4 Configuration Example of IS-IS for IPv6

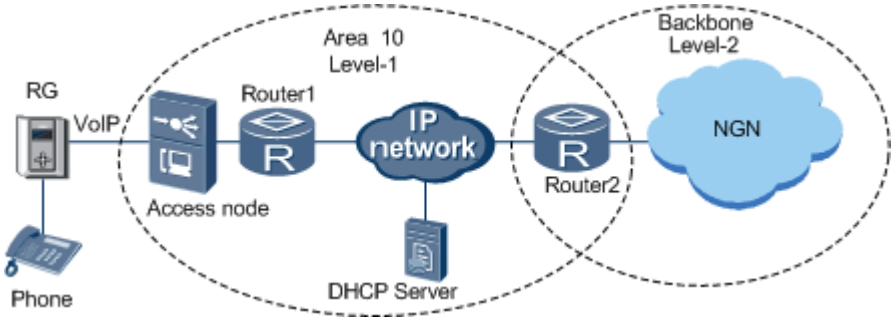
This operation enables the corresponding device configured data to run the IS-IS protocol on the access device in the IPv6 network.

Service Requirements

- The access device forwards the access VoIP service through the Layer 3 interface to the NGN network.
- The access device obtains the routes of the NGN networking through the IS-IS protocol.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-103 Example network for configuring IS-IS for IPv6



Data Plan

Table 11-29 provides the data plan for configuring IS-IS.

Table 11-29 Data plan for configuring IS-IS

Item	Data
Access node	IS-IS process ID: 1
	NET (Network entity title): 10.0000.0000.0001.00, where: • Area ID: 10 • System ID: 0000.0000.0001 • Level: Level-1 • Host name: MA5800
	IS-IS interface: • Port number: 0/9/0 • VLAN ID: 20 • IP address: 2001:db8:1::1/64

Procedure

- Configure IS-IS on the access node.
  a. Configure the Layer 3 interface.

```
huawei(config)#vlan 20 standard
huawei(config)#port vlan 20 0/9 0
huawei(config)#interface vlanif 20
huawei(config-if-vlanif20)#ipv6 enable
huawei(config-if-vlanif20)#ipv6 address 2001:db8:1::1 64
huawei(config-if-vlanif20)#quit
```

  b. Start the IS-IS process, and enable IPv6 for the process.

```
huawei(config)#isis 1
huawei(config-isis-1)#ipv6 enable
```

  c. Configure the NET.

```
huawei(config-isis-1)#network-entity 10.0000.0000.0001.00
```

 Copyright © Huawei Technologies Co., Ltd.

d. Configure the router level.

```
huawei(config-isis-1)#is-level level-1
```

e. Configure the local host name.

```
huawei(config-isis-1)#is-name MA5800
huawei(config-isis-1)#quit
```

f. Enable the IS-IS for IPv6 function on an interface.

```
huawei(config)#interface vlanif 20
huawei(config-if-vlanif20)#isis ipv6 enable 1
huawei(config-if-vlanif20)#quit
```

- Configure IS-IS for IPv6 on Router1.

  The process of configuring IS-IS on Router1 is similar to that of configuring IS-IS on the access node. The details are not provided in this chapter.

- Configure IS-IS for IPv6 on Router2.

  The process of configuring IS-IS on Router2 is similar to that of configuring IS-IS on the access node. The details are not provided in this chapter.

----End

Result

- Run the **display isis lsdb** command and you can query the IS-IS LSDB.
- Run the **display isis route** command and you can query the IS-IS route. The routing table of the Level-1 router should have a default route, and the next hop should be the Level-1-2 router. The Level-2 router should have the routes to all the Level-1 routers and the Level-2 routers.

Configuration File

```
vlan 20 standard
port vlan 20 0/9 0
interface vlanif 20
ipv6 enable
ipv6 address 2001:db8:1::1 64
quit
isis 1
ipv6 enable
network-entity 10.0000.0000.0001.00
is-level level-1
is-name MA5800
quit
interface vlanif 20
isis ipv6 enable 1
quit
```

11.12.5.5 Example for Configuring IS-IS Auto FRR

IS-IS auto fast reroute (FRR) is applicable to services that are sensitive to packet loss and latency. Configuring IS-IS auto FRR implements fast link switching, which greatly improves the reliability of an IS-IS network.

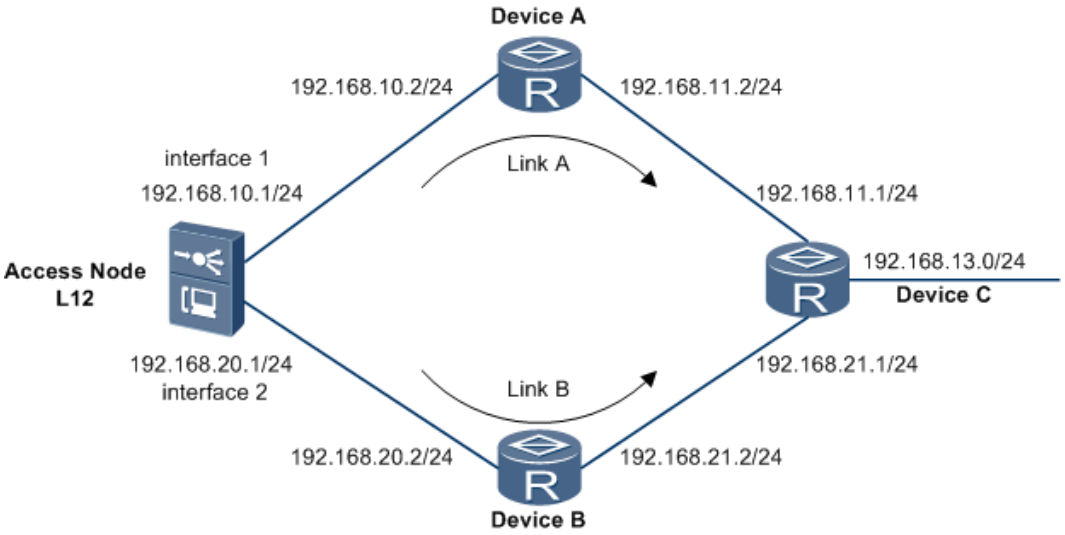
Service Requirements

When a fault occurs on the network, IS-IS auto FRR fast switches traffic to the backup link without waiting for route convergence, reducing the switching time.

As shown in **Figure 11-104**,

- The 4 routers run IS-IS to interconnect IP networks.
- The 4 routers are level-1-2 routers.
- When link A fails, the MA5800 fast switches traffic to backup link B.

Figure 11-104 Networking of configuring IS-IS auto FRR



Data Plan

Table 11-30 describes data plan for configuring IP IS-IS auto FRR.

Table 11-30 Data plan for configuring IS-IS auto FRR

Item	Data
MA5800 port 1	IP address of the port: 192.168.10.1 24
	VLAN ID: 2
	Cost: 10
MA5800 port 2	IP address of the port: 192.168.20.1 24
	VLAN ID: 3
	Cost: 30

Prerequisite

Before you configure IS-IS Auto FRR, complete the following tasks:

- Configuring link layer protocol parameters and IP addresses for ports to ensure that the link layer protocol on the ports is Up
- Configuring basic IP IS-IS functions

 Copyright © Huawei Technologies Co., Ltd.

Procedure

Step 1 Perform configuration on the MA5800.

1. Configure the IP address of the Layer 3 interface on the MA5800.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 192.168.10.1 24
huawei(config-if-vlanif2)#quit
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 1
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ip address 192.168.20.1 24
huawei(config-if-vlanif3)#quit
```

2. Configure basic IS-IS functions.

```
huawei(config)#isis 1
huawei(config-isis-1)#is-level level-1-2
huawei(config-isis-1)#network-entity 10.0000.0000.0001.00
huawei(config-isis-1)#quit
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#isis enable 1
huawei(config-if-vlanif2)#quit
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#isis enable 1
huawei(config-if-vlanif3)#quit
```

3. Set the port cost of the MA5800 to 30.

```
huawei(config)#interface vlan 3
huawei(config-if-vlanif3)#isis cost 30
huawei(config-if-vlanif3)#quit
```

4. Enable IS-IS auto FRR on the MA5800.

```
huawei(config)#isis
huawei(config-isis-1)#frr
huawei(config-isis-1-ipv6-frr)#loop-free-alternate
huawei(config-isis-1-ipv6-frr)#quit
huawei(config-isis-1)#quit
```

5. Save the data.

```
huawei(config)#save
```

----End

Result

Run the **display isis route** command to query the information about primary/secondary routes configured.

Configuration File

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ip address 192.168.10.1 24
quit
vlan 3 smart
port vlan 3 0/9 1
interface vlanif 3
ip address 192.168.20.1 24
quit
isis 1
is-level level-1-2
network-entity 10.0000.0000.0001.00
quit
```

 Copyright © Huawei Technologies Co., Ltd.

```
interface vlanif 2
isis enable 1
quit
interface vlanif 3
isis enable 1
quit
interface vlan 3
isis cost 30
quit
isis
frr
loop-free-alternate
quit
quit
save
```

11.12.5.6 Example for Configuring IPv6 IS-IS Auto FRR

IPv6 IS-IS auto fast reroute (FRR) is applicable to services that are sensitive to packet loss and latency. Configuring IPv6 IS-IS auto FRR implements fast link switching, which greatly improves the reliability of an IS-IS network.

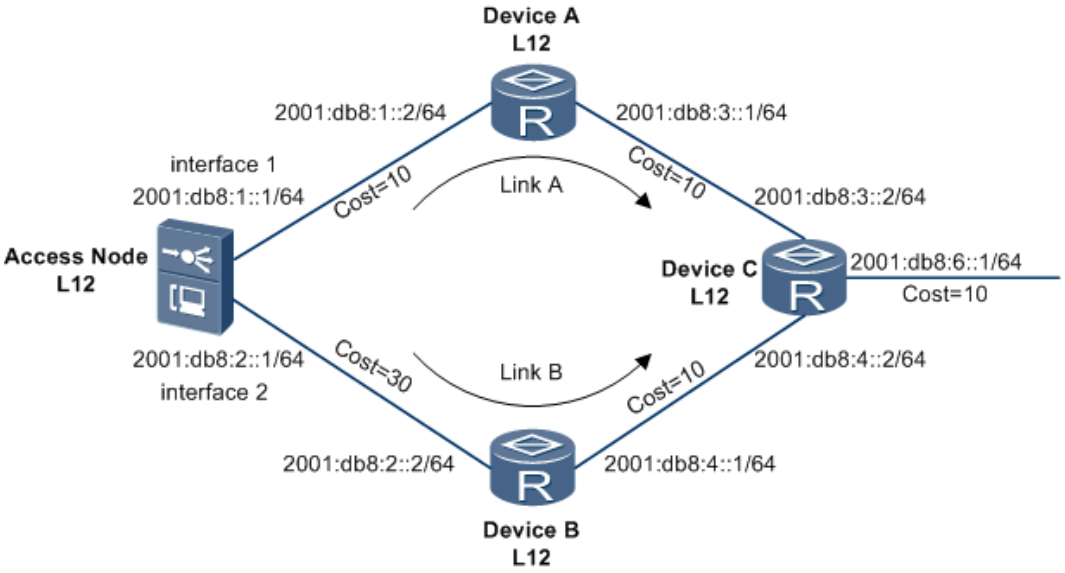
Service Requirements

When a fault occurs on the network, IPv6 IS-IS auto FRR fast switches traffic to the backup link without waiting for route convergence, reducing the switching time.

As shown in **Figure 11-105**,

- The 4 routers run IS-IS to interconnect IPv6 networks.
- The 4 routers are level-1-2 routers.
- When link A fails, the MA5800 fast switches traffic to backup link B.

Figure 11-105 Networking of configuring IPv6 IS-IS auto FRR



Data Plan

Table 11-31 describes data plan for configuring IPv6 IS-IS auto FRR.

Table 11-31 Data plan for configuring IPv6 IS-IS auto FRR

Item	Data
MA5800 port 1	IPv6 address of the port: 2001:db8:1::1 64
	VLAN ID: 2
	Cost: 10
MA5800 port 2	IPv6 address of the port: 2001:db8:2::1 64
	VLAN ID: 3
	Cost: 30

Prerequisite

Before you configure IPv6 IS-IS Auto FRR, complete the following tasks:

- Configuring link layer protocol parameters and IPv6 addresses for ports to ensure that the link layer protocol on the ports is Up
- Configuring basic IPv6 IS-IS functions

Procedure

Step 1 Perform configuration on the MA5800.

1. Configure the IPv6 address of the Layer 3 interface on the MA5800.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1::1 64
huawei(config-if-vlanif2)#quit
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 1
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ipv6 enable
huawei(config-if-vlanif3)#ipv6 address 2001:db8:2::1 64
huawei(config-if-vlanif3)#quit
```

2. Configure basic IS-IS IPv6 functions.

```
huawei(config)#isis 1
huawei(config-isis-1)#is-level level-1-2
huawei(config-isis-1)#network-entity 10.0000.0000.0001.00
huawei(config-isis-1)#quit
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#isis ipv6 enable 1
huawei(config-if-vlanif2)#quit
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#isis ipv6 enable 1
huawei(config-if-vlanif3)#quit
```

3. Set the port cost of the MA5800 to 30.

```
huawei(config)#interface vlan 3
huawei(config-if-vlanif3)#isis ipv6 cost 30
huawei(config-if-vlanif3)#quit
```

 Copyright © Huawei Technologies Co., Ltd.

4. Enable IPv6 IS-IS auto FRR on the MA5800.

```
huawei(config)#isis
huawei(config-isis-1)#ipv6 frr
huawei(config-isis-1-ipv6-frr)#loop-free-alternate
huawei(config-isis-1-ipv6-frr)#quit
huawei(config-isis-1)#quit
```

5. Save the data.

```
huawei(config)#save
```

----End

Result

Run the **display isis route** command to query the information about primary/secondary routes configured.

Configuration File

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ipv6 enable
ipv6 address 2001:db8:1::1 64
quit
vlan 3 smart
port vlan 3 0/9 1
interface vlanif 3
ipv6 enable
ipv6 address 2001:db8:2::1 64
quit
isis 1
is-level level-1-2
network-entity 10.0000.0000.0001.00
quit
interface vlanif 2
isis ipv6 enable 1
quit
interface vlanif 3
isis ipv6 enable 1
quit
interface vlan 3
isis ipv6 cost 30
quit
isis
ipv6 frr
loop-free-alternate
quit
quit
save
```

11.12.5.7 IS-IS Reference Standards and Protocols

Table 11-32 The following table lists the references.

Document NO.	Document Name	Remarks
ISO 10589	ISO IS-IS Routing Protocol	-
RFC 1142	OSI IS-IS Intra-domain Routing Protocol	-

 Copyright © Huawei Technologies Co., Ltd.

Document NO.	Document Name	Remarks
RFC 1195	Use of OSI IS-IS for Routing in TCP/IP and Dual Environments	Multiple authentication passwords cannot be configured.
RFC 2763	Dynamic Hostname Exchange Mechanism for IS-IS	-
RFC 2966	Domain-wide Prefix Distribution with Two-Level IS-IS	-
RFC 2973	IS-IS Mesh Groups	-
RFC 3277	IS-IS Transient Blackhole Avoidance	-
RFC 3359	Summary of all Reserved Type, Length and Value (TLV) Codepoints in IS-IS	-
RFC 3373	Three-Way Handshake for IS-IS Point-to-Point Adjacencies	-
RFC 3567	IS-IS Cryptographic Authentication	-
RFC 3719	Recommendations for Interoperable Networks using IS-IS	-
RFC 3786	Extension of the Number of IS-IS LSP Fragments Beyond the 256 Limit	-
RFC 3787	Recommendations for Interoperable IP Networks using IS-IS	-
RFC 3906	Calculation of Interior Gateway Protocol (IGP) Routes	-
RFC 4444	Management Information Base for IS-IS	Only some features are supported.
RFC 5130	A Policy Control Mechanism in IS-IS Using Administrative Tags	-
RFC 5286	Basic Specification for IP Fast Reroute: Loop-Free Alternates	-
RFC 5301	Dynamic Hostname Exchange Mechanism for IS-IS	Same as RFC 2763.
RFC 5302	Domain-Wide Prefix Distribution with Two-Level IS-IS	Same as RFC 2966.
RFC 5303	Three-Way Handshake for IS-IS Point-to-Point Adjacencies	Same as RFC 3373.
RFC 5304	IS-IS Cryptographic Authentication	Same as RFC 3567.

 Copyright © Huawei Technologies Co., Ltd.

Document NO.	Document Name	Remarks
RFC 5309	Point-to-Point Operation over LAN in Link State Routing Protocols	-
draft-ietf-isis-igp-p2p-over-lan-06	Point-to-point operations over LAN in link-state routing protocols	-
draft-ietf-isis-3way-03	Three-Way Handshake for IS-IS Point-to-Point Adjacencies	-
draft-ietf-rtgwg-ipfrr-framework-11	IP Fast Reroute Framework	-

11.12.6 OSPF

11.12.6.1 Introduction to OSPF

Definition

Open Shortest Path First (OSPF) is a link-state Interior Gateway Protocol (IGP) developed by the Internet Engineering Task Force (IETF).

OSPF version 2 (OSPFv2), which is defined in RFC 2328, is intended for IPv4. OSPF version 3 (OSPFv3), which is defined in RFC 2740, is intended for IPv6.

NOTE

In this document, OSPF refers to OSPFv2, unless otherwise stated.

Purpose

Before the emergence of OSPF, the Routing Information Protocol (RIP) was widely used as an IGP on networks. RIP is a distance-vector routing protocol, and therefore has slow convergence, routing loops, and poor scalability. OSPF was developed to replace RIP. Typical IGPs include RIP, OSPF, and Intermediate System to Intermediate System (IS-IS). **Table 11-33** describes differences among the three typical IGPs.

Table 11-33 Differences among the three typical IGPs

Item	RIP	OSPF	IS-IS
Protocol type	IP layer protocol	IP layer protocol	Link layer protocol

 Copyright © Huawei Technologies Co., Ltd.

Item	RIP	OSPF	IS-IS
Application scope	Applies to small networks with simple architectures, such as campus networks.	Applies to medium-sized networks with several hundred Routers supported, such as enterprise networks.	Applies to large networks, such as Internet service provider (ISP) networks.
Routing algorithm	Uses a distance-vector algorithm and exchanges routing information over the User Datagram Protocol (UDP).	Uses the shortest path first (SPF) algorithm to generate a shortest path tree (SPT) based on the network topology, calculates shortest paths to all destinations, and exchanges routing information over IP.	Uses the SPF algorithm to generate an SPT based on the network topology, calculates shortest paths to all destinations, and exchanges routing information over IP. The SPF algorithm runs separately in Level-1 and Level-2 databases.
Route convergence speed	Slow	Less than 1 second	Less than 1 second
Scalability	Not supported	Supported by partitioning a network into areas	Supported by defining Router levels

Benefits

OSPF offers the following benefits:

- Wide application scope: OSPF applies to medium-sized networks with several hundred Routers supported, such as enterprise networks.
- Network masks supported: OSPF packets can carry masks, and therefore the packet length is not limited by natural IP masks. OSPF can process variable length subnet masks (VLSMs).
- Fast convergence: When the network topology changes, OSPF immediately sends link state update (LSU) packets to synchronize the changes to the link state databases (LSDBs) of all routers in an autonomous system (AS).
- Loop-free: OSPF uses the SPF algorithm to calculate loop-free routes based on the collected link status.
- Area partition: OSPF allows an AS to be partitioned into areas to simplify management. Routing information transmitted between areas is further summarized, which reduces network bandwidth consumption.

 Copyright © Huawei Technologies Co., Ltd.

- Equal-cost route: OSPF allows multiple equal-cost routes to the same destination.
- Routing hierarchy: OSPF uses four route types: intra-area routes, inter-area routes, Type 1 external routes, and Type 2 external routes, which are listed in descending order of priority.
- Authentication: OSPF supports area- and interface-based packet authentication, which ensures packet exchange security.
- Multicast: OSPF uses multicast addresses to send packets on certain types of links, which minimizes the impact on other devices.

11.12.6.2 Principles

11.12.6.2.1 Basic Concepts of OSPF

This section describes the basic Open Shortest Path First (OSPF) concepts.

Router ID

A router ID is a 32-bit unsigned integer, which identifies a Router in an autonomous system (AS). A router ID must exist before a Router runs OSPF.

A router ID can be obtained in two modes:

- In manual mode, a router ID needs to be manually configured.
- Automatically set by the router.

  If no router ID has been configured, the router automatically obtains a router ID using the following methods in descending order of priority. The selection sequence is as follows:

  a. The router preferentially selects the largest IP address from its loopback interface addresses as the router ID.

  b. If no loopback interface has been configured, the router selects the largest IP address from its interface IP addresses as the router ID.

A router can obtain a router ID again only after a router ID is reconfigured for the router or an OSPF router ID is reconfigured and the OSPF process restarts.

Area

When a large number of Routers run OSPF, link state databases (LSDBs) become very large and require a large amount of storage space. Large LSDBs also complicate shortest path first (SPF) computation and are computationally intensive for the Routers. Network expansion causes the network topology to change, which results in route flapping and frequent OSPF packet transmission. When a large number of OSPF packets are transmitted on the network, bandwidth usage efficiency decreases. Each change in the network topology causes all Routers on the network to recalculate routes.

OSPF resolves this problem by partitioning an AS into different areas. An area is regarded as a logical group, and each group is identified by an area ID. A Router, not a link, resides at the border of an area. A network segment or link can belong only to one area. An area must be specified for each OSPF interface.

 Copyright © Huawei Technologies Co., Ltd.

OSPF areas include common areas, stub areas, and not-so-stubby areas (NSSAs), as described in **Table 11-34**.

Table 11-34 OSPF areas

Area Type	Function	Notes
Common area	By default, OSPF areas are defined as common areas. Common areas include: • Standard area: transmits intra-area, inter-area, and external routes. • Backbone area: connects to all other OSPF areas and transmits inter-area routes. The backbone area is represented by area 0. Routes between non-backbone areas must be forwarded through the backbone area.	• The backbone area must have all its devices connected. • All non-backbone areas must remain connected to the backbone area.
Stub area	A stub area is a non-backbone area with only one ABR and generally resides at the border of an AS. The area border router (ABR) in a stub area does not transmit received AS external routes, which significantly decreases the number of entries in the routing table on the ABR and the amount of routing information to be transmitted. To ensure the reachability of AS external routes, the ABR in the stub area generates a default route and advertises the route to non-ABRs in the stub area. A totally stub area allows only intra-area routes and ABR-advertised Type 3 link state advertisements (LSAs) carrying a default route to be advertised within the area.	• The backbone area cannot be configured as a stub area. • An autonomous system boundary router (ASBR) cannot exist in a stub area. Therefore, AS external routes cannot be advertised within the stub area. • A virtual link cannot pass through a stub area.
NSSA	An NSSA is similar to a stub area. An NSSA does not advertise **Type5 LSA**s, but can import AS external routes. ASBRs in an NSSA generate **Type7 LSA**s to carry the information about the AS external routes. The Type 7 LSAs are advertised only within the NSSA. When the Type 7 LSAs reach an ABR in the NSSA, the ABR translates the Type 7 LSAs into Type 5 LSAs and floods them to the entire AS. A totally NSSA area allows only intra-area routes to be advertised within the area.	• ABRs in an NSSA advertise Type 7 LSAs carrying a default route within the NSSA. All inter-area routes are advertised by ABRs. • A virtual link cannot pass through an NSSA.

 Copyright © Huawei Technologies Co., Ltd.

Router Type

Routers are classified as internal routers, ABRs, backbone routers, or ASBRs by location in an AS. **Figure 11-106** and **Table 11-35** shows the four router types.

Figure 11-106 Router type layout

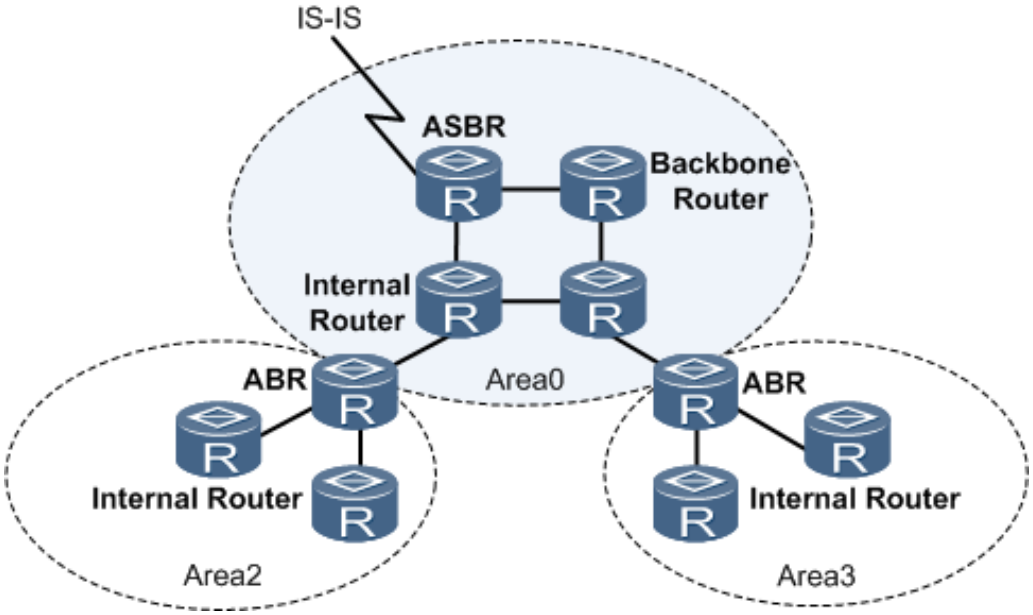


Table 11-35 Router type description

Router Type	Description
Internal router	All interfaces of an internal router belong to the same OSPF area.
ABR	An ABR can belong to two or more areas, one of which must be a **backbone area**. An ABR connects the backbone area and non-backbone areas. An ABR can be physically or logically connected to the backbone area.
Backbone router	At least one interface on a backbone router belongs to the backbone area. Internal routers in the backbone area and all ABRs are backbone routers.
ASBR	An ASBR exchanges routing information with other ASs. An ASBR may not reside at the boundary of an AS. An ASBR can be an internal router or an ABR.

 Copyright © Huawei Technologies Co., Ltd.

LSA

OSPF encapsulates routing information into LSAs for transmission. **Table 11-36** describes LSAs and their functions.

Table 11-36 LSAs and their functions

LSA Type	Function
Router-LSA (Type 1)	Describes the link status and cost of a Router. Router-LSAs are generated by a Router and advertised within the area to which the Router belongs.
Network-LSA (Type 2)	Describes the link status of all routers on the local network segment. Network-LSAs are generated by a designated router (DR) and advertised within the area to which the DR belongs.
Network-summary-LSA (Type 3)	Describes routes on a network segment. Network-summary-LSAs are generated by an **ABR** and are advertised within the non-totally stub area or NSSA.
ASBR-summary-LSA (Type 4)	Describes routes to an ASBR in an area. ASBR-summary-LSAs are generated by an **ABR** and are advertised to the areas except the area to which the ASBR belongs.
AS-external-LSA (Type 5)	Describes AS external routes, which are advertised to all areas except **stub areas** and **NSSAs**. AS-external-LSAs are generated by an **ASBR**.
NSSA-LSA (Type 7)	Describes AS external routes. NSSA-LSAs are generated by an **ASBR** and advertised only within NSSAs.
Opaque-LSA (Type 9/ Type 10/Type 11)	Provides a general mechanism for OSPF extension. Different types of LSAs are described as follows: • Type 9 LSAs are advertised only on the network segment where the interface advertising the LSAs resides. Grace LSAs used in graceful restart (GR) are Type 9 LSAs. • Type 10 LSAs are advertised within an OSPF area. LSAs that are used to support traffic engineering (TE) are Type 10 LSAs. • Type 11 LSAs are advertised within an AS but have not been used in practice.

Packet Type

OSPF packets are encapsulated into IP packets. The OSPF protocol number is 89. OSPF packets are classified as Hello, Database Description (DD), Link State

 Copyright © Huawei Technologies Co., Ltd.

Request (LSR), Link State Update (LSU), or Link State Acknowledgment (LSAck) packets, as described in **Table 11-37**.

Table 11-37 OSPF packets and their functions

Packet Type	Function
Hello packet	Hello packets are periodically sent to discover and maintain OSPF neighbor relationships.
DD packet	A DD packet contains the summaries of LSAs in the local LSDB. DD packets are used for LSDB synchronization between two Routers.
LSR packet	LSR packets are sent to OSPF neighbors to request required LSAs. A Router sends LSR packets to its OSPF neighbor only after DD packets have been successfully exchanged.
LSU packet	LSU packets are used to transmit required LSAs to OSPF neighbors.
LSAck packet	LSAck packets are used to acknowledge received LSAs.

Route Type

Intra-area and inter-area routes describe the network structure of an AS. AS external routes describe how to select routes to destinations outside an AS. AS external routes imported by OSPF are classified as Type 1 or Type 2 external routes.

Table 11-38 describes OSPF routes in descending order of priority.

Table 11-38 OSPF routes

Route Type	Description
Intra-area route	-
Inter-area route	-
Type 1 external route	Because a Type 1 external route has high reliability, its cost equals the cost of an AS internal route. Cost of a Type 1 external route = Cost of the route from a Router to an ASBR + Cost of the route from the ASBR to the destination

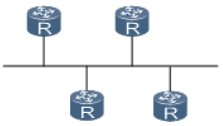
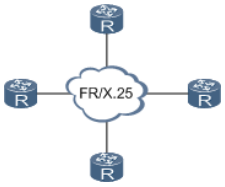
 Copyright © Huawei Technologies Co., Ltd.

Route Type	Description
Type 2 external route	Because a Type 2 external route has low reliability, its cost is considered to be much greater than the cost of any internal route to an ASBR. Cost of a Type 2 external route = Cost of the route from an ASBR to the destination

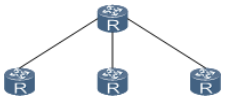
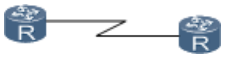
Network Type

Networks are classified as broadcast, non-broadcast multiple access (NBMA), point-to-multipoint (P2MP), or point-to-point (P2P) networks by link layer protocol, as described in **Table 11-39**.

Table 11-39 OSPF network classification

Network Type	Link Layer Protocol	Packet Transmission Mode	Graph
Broadcast	• Ethernet • FDDI	• **Hello** and **LSAck** packets are transmitted by an OSPF device using the address 224.0.0.5 in multicast mode. **LSU** packets are first transmitted in multicast mode and retransmitted in unicast mode. When an OSPF DR is used, the OSPF DR transmits packets using the address 224.0.0.6 in multicast mode. • **DD** and **LSR** packets are transmitted in unicast mode.	
NBMA	X.25	Hello, DD, LSR, LSU, and LSAck packets are transmitted in unicast mode.	

 Copyright © Huawei Technologies Co., Ltd.

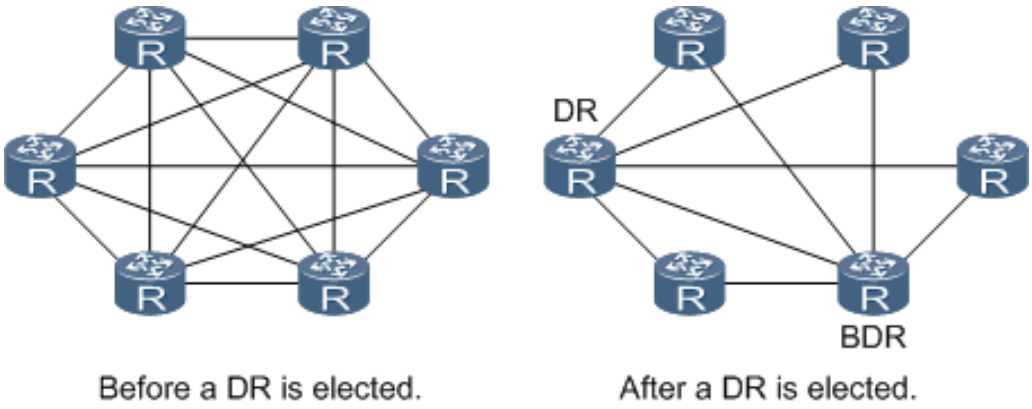
Network Type	Link Layer Protocol	Packet Transmission Mode	Graph
P2MP	Regardless of the link layer protocol, OSPF does not default the network type to P2MP. The network type must be manually changed to P2MP. The common practice is to change a non-fully connected NBMA network to a P2MP network.	• Hello packets are transmitted using the address 224.0.0.5 in multicast mode. • DD, LSR, LSU, and LSAck packets are transmitted in unicast mode.	
P2P	• PPP • HDLC • LAPB	• Hello, DD, LSR, LSU, and LSAck packets are transmitted using the address 224.0.0.5 in multicast mode. • LSU packets are retransmitted in multicast mode.	

 Copyright © Huawei Technologies Co., Ltd.

DR and BDR

On broadcast or NBMA networks, any two Routers need to exchange routing information. As shown in **Figure 11-107**, n routers are deployed on the network. n x (n - 1)/2 adjacencies must be established. Any route change on a Router is transmitted to other Routers, which wastes bandwidth resources. OSPF resolves this problem by defining a DR and a backup designated router (BDR). After a DR is elected, all Routers send routing information only to the DR. Then the DR broadcasts LSAs. Routers other than the DR and BDR are called DR others. The DR others establish only adjacencies with the DR and BDR and not with each other. This process reduces the number of adjacencies established between Routers on broadcast or NBMA networks.

Figure 11-107 Network topologies before and after a DR election



If the original DR fails, Routers must reelect a DR and the Routers except the new DR must synchronize routing information to the new DR. This process is lengthy, which may cause incorrect route calculations. A BDR is used to shorten the process. The BDR is a backup for a DR. A BDR is elected together with a DR. The BDR establishes adjacencies with all Routers on the network segment and exchanges routing information with them. When the DR fails, the BDR immediately becomes a new DR. The Routers need to reelect a new BDR, but this process does not affect route calculations.

The DR priority of a Router interface determines its qualification for DR and BDR elections. The Router interfaces with their DR priorities greater than 0 are eligible. Each Router adds the elected DR to a Hello packet and sends it to other Routers on the network segment. When both Router interfaces on the same network segment declare that they are DRs, the Router interface with a higher DR priority is elected as a DR. If the two Router interfaces have the same DR priority, the Router interface with a larger router ID is elected as a DR.

OSPF Multi-process

OSPF supports multi-process. Multiple OSPF processes can independently run on the same Router. Route exchange between different OSPF processes is similar to route exchange between different routing protocols. A Router's interface can belong only to one OSPF process.

 Copyright © Huawei Technologies Co., Ltd.

A typical application of OSPF multi-process is that OSPF runs between provider edges (PEs) and customer edges (CEs) on virtual private networks (VPNs) and OSPF is also used as an Interior Gateway Protocol (IGP) on the VPN backbone network. The two OSPF processes on the PEs are independent of each other.

OSPF Default Route

A default route is the route whose destination address and mask are both all 0s. When no exactly matching route is discovered, a Router uses a default route to forward packets.

A default route generally applies to the following scenarios:

- An ABR in an area advertises Type 3 LSAs carrying a default route within the area. The Routers in the area use the received default route to forward inter-area packets.
- An ASBR in an AS advertises Type 5 or Type 7 LSAs carrying a default route within the AS. The Routers in the AS use the received default route to forward AS external packets.

OSPF routes are hierarchically managed. The priority of the default route carried in Type 3 LSAs is higher than the priority of the default route carried in Type 5 or Type 7 LSAs.

The basic principles for advertising LSAs carrying a default route are as follows:

- A Router in an area can advertise LSAs carrying a default route only when the Router has an interface connected to a device outside the area.
- If a Router has advertised LSAs carrying a default route, the Router no longer learns the same type of LSA advertised by other Routers, which carry a default route. That is, the Router uses only the LSAs advertised by itself to calculate routes. The LSAs advertised by other Routers are still saved in the LSDB.
- If a Router must use a route to advertise LSAs carrying an external default route, the route cannot be a route learned by the local OSPF process. A Router in an area uses an external default route to forward packets outside the area. If the next hops of routes in the area are Routers in the area, packets cannot be forwarded outside the area.
- Before a Router advertises a default route, it checks whether a neighbor in the full state is present in area 0. The Router advertises a default route only when a neighbor in the full state is present in area 0. If no such a neighbor exists, the backbone area cannot forward packets and advertising a default route is meaningless.

Table 11-40 describes the principles for advertising default routes in different areas.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-40 Principles for advertising default routes in different areas

Area Type	Advertisement Principles
Common area	By default, a Router in a common area does not generate a default route. After being configured to do so, an ASBR generates a Type 5 LSA carrying a default route. The Router then advertises the default route in the entire AS. If no default route is generated on the ASBR, the Router does not advertise a default route.
Stub area	Type 5 LSAs cannot be advertised within a stub area. A Router in the stub area must learn AS external routes from an ABR. The ABR automatically generates a Type 3 LSA carrying a default route and advertises it within the entire stub area. Then the Router can learn AS external routes from the ABR.
Totally stub area	Neither Type 3 nor Type 5 LSAs can be advertised within a totally stub area. A Router in the totally stub area must learn AS external and inter-area routes from an ABR. After you configure a totally stub area, an ABR automatically generates a Type 3 LSA carrying a default route and advertises it within the entire totally stub area. Then the Router can learn AS external and inter-area routes from the ABR.
NSSA	A small number of AS external routes learned from the ASBR in an NSSA can be imported to the NSSA. Type 5 LSAs cannot be advertised within the NSSA. The ABR automatically generates a Type 7 LSA carrying a default route and advertises it within the entire NSSA. A small number of AS external routes can be learned from the ASBR in the NSSA, and other inter-area routes can be learned from the ABR in the NSSA. Manual configurations must be performed on the ASBR to enable the ASBR to generate a Type 7 LSA carrying a default route and advertise the LSA within the entire NSSA. An ABR does not translate Type 7 LSAs carrying a default route into Type 5 LSAs carrying a default route or flood them to the entire AS.
Totally NSSA	Neither Type 3 nor Type 5 LSAs can be advertised within a totally NSSA. A Router in the totally NSSA must learn AS external routes from an ABR. The ABR automatically generates Type 3 and Type7 LSAs carrying a default route and advertises them to the entire totally NSSA. Then AS external and inter-area routes can be advertised within the totally NSSA.

11.12.6.2.2 Basic Principles of OSPF

OSPF route calculation involves the following processes:

 Copyright © Huawei Technologies Co., Ltd.

1. The **adjacency establishment** process is as follows:
   a. The local device sends Hello packets through an interface to a remote device to establish a **neighbor relationship**.
   b. The local and remote devices negotiate a master/slave relationship and exchange Database Description (DD) packets.
   c. The local and remote devices exchange link state advertisements (LSAs) to synchronize their link state databases (LSDBs).

   The **adjacency** is established.
2. **Route calculation**

   OSPF uses the shortest path first (SPF) algorithm to calculate routes, resulting in fast route convergence.

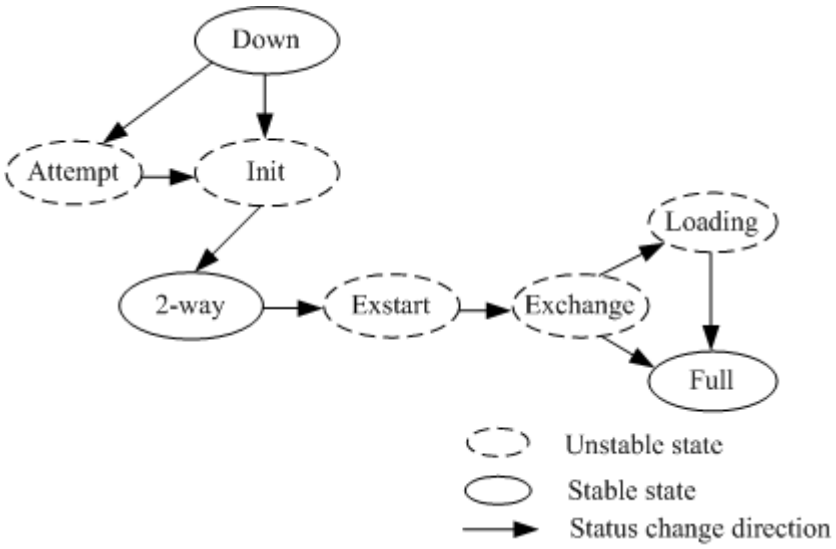
OSPF Neighbor States

To exchange routing information on an OSPF network, neighbor Routers must establish adjacencies. The differences between neighbor relationships and adjacencies are described as follows:

- Neighbor relationship: After the local Router starts, it uses an OSPF interface to send a Hello packet to the remote Router. After the remote Router receives the packet, it checks whether the parameters carried in the packet are consistent with its own parameters. If the parameters carried in the packet are consistent with its own parameters, the local and remote Routers establish a neighbor relationship.
- Adjacency: After the local and remote Routers establish a neighbor relationship, they exchange DD packets and LSAs to establish an adjacency.

OSPF has eight neighbor states: Down, Attempt, Init, 2-way, Exstart, Exchange, Loading, and Full, as shown in **Figure 11-108**. Down, 2-way, and Full are stable states. Attempt, Init, Exstart, Exchange, and Loading are unstable states, which last only several minutes.

Figure 11-108 OSPF neighbor states



 Copyright © Huawei Technologies Co., Ltd.

Table 11-41 OSPF neighbor states and their meanings

OSPF Neighbor State	Meaning
Down	This is the initial state of a neighbor conversation. This state indicates that a Router has not received any Hello packets from its neighbors within a dead interval.
Attempt	In the Attempt state, a Router periodically sends Hello packets to manually configured neighbors. **NOTE** This state applies only to non-broadcast multiple access (NBMA) interfaces.
Init	This state indicates that a Router has received Hello packets from its neighbors but the neighbors did not receive Hello packets from the Router.
2-way	This state indicates that a Router has received Hello packets from its neighbors and **neighbor relationships** have been established between the Routers. If no adjacency needs to be established, the neighbors remain in the 2-way state. If adjacencies need to be established, the neighbors enter the Exstart state.
Exstart	In the Exstart state, Routers establish a master/slave relationship to ensure that DD packets are sequentially exchanged.
Exchange	In the Exchange state, Routers exchange DD packets. A Router uses a DD packet to describe its own LSDB and sends the packet to its neighbors.
Loading	In the Loading state, a Router sends Link State Request (LSR) packets to its neighbors to request their LSAs for LSDB synchronization.
Full	In the Full state, a Router establishes **adjacencies** with its neighbors and all LSDBs have been synchronized.

NOTE

The neighbor state of the local Router may be different from that of the remote Router. For example, the neighbor state of the local Router is Full, but the neighbor state of the remote Router is Loading.

Adjacency Establishment

Adjacencies can be established in either of the following situations:

- Two Routers have established a neighbor relationship and communicate for the first time.
- The designated router (DR) or backup designated router (BDR) on a network segment changes.

 Copyright © Huawei Technologies Co., Ltd.

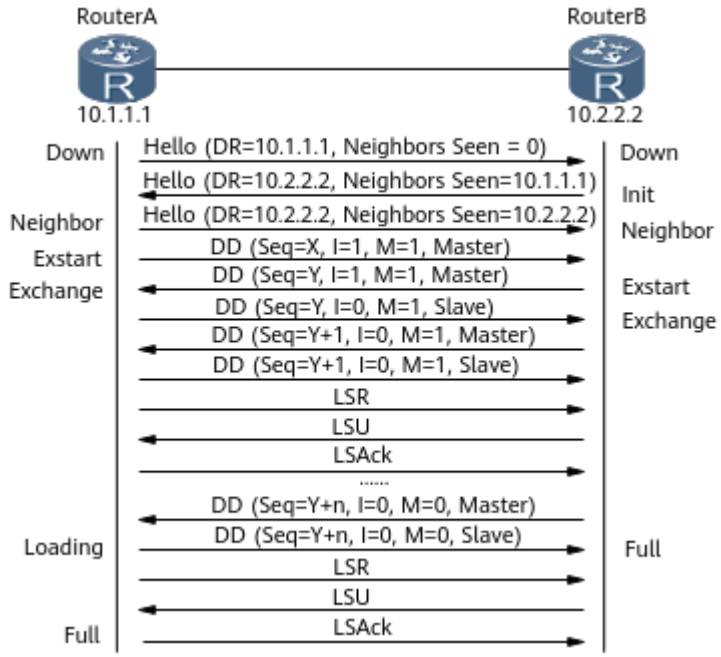
The adjacency establishment process is different on different networks.

Adjacency establishment on a broadcast network

On a broadcast network, the DR and BDR establish adjacencies with each Router on the same network segment, but DR others establish only neighbor relationships.

Figure 11-109 shows the adjacency establishment process on a broadcast network.

Figure 11-109 Adjacency establishment process on a broadcast network



The adjacency establishment process on a broadcast network is as follows:

1. Neighbor relationship establishment
   a. Router A uses the multicast address 224.0.0.5 to send a Hello packet to Router B through the OSPF interface connected to a broadcast network. The packet carries the DR field of 10.1.1.1 (ID of Router A) and the Neighbors Seen field of 0. A neighbor Router has not been discovered, and Router A regards itself as a DR.
   b. After Router B receives the packet, it returns a Hello packet to Router A. The returned packet carries the DR field of 10.2.2.2 (ID of Router B) and the Neighbors Seen field of 10.1.1.1 (Router A's router ID). Router A has been discovered but its router ID is less than that of Router B, and therefore Router B regards itself as a DR. Then Router B's status changes to Init.
   c. After Router A receives the packet, Router A's status changes to Exstart.

   NOTE

   The following procedures are not performed for DR others on a broadcast network.
2. Master/Slave negotiation and DD packet exchange

 Copyright © Huawei Technologies Co., Ltd.

a. Router A sends a DD packet and claims it as the master (MS = 1), and sets the sequence number (Seq) to x. I field: The value 1 indicates that the packet is the first DD packet, which is used to negotiate a master/slave relationship and does not carry LSA summaries. M field: The value 1 indicates that the packet is not the last DD packet.

   To improve transmission efficiency, Router A and Router B determine which LSAs in each other's LSDB need to be updated. If one party determines that an LSA of the other party is already in its own LSDB, it does not send an LSR packet for updating the LSA to the other party. To achieve the preceding purpose, Router A and Router B first send DD packets, which carry summaries of LSAs in their own LSDBs. Each summary identifies an LSA. To ensure packet transmission reliability, a master/slave relationship must be determined during DD packet exchange. One party serving as a master uses the Seq field to define a sequence number. The master increases the sequence number by one each time it sends a DD packet. When the other party serving as a slave sends a DD packet, it adds the sequence number carried in the last DD packet received from the master to the Seq field of the packet.

b. After Router B receives the DD packet, Router B's status changes to Exstart and Router B returns a DD packet to Router A. The returned packet does not carry LSA summaries. Because Router B's router ID is greater than Router A's router ID, Router B declares itself a master and sets the Seq field to *y*.

c. After Router A receives the DD packet, it agrees that Router B is a master and Router A's status changes to Exchange. Then Router A sends a DD packet to Router B to transmit LSA summaries. The packet carries the Seq field of *y* and the MS field of 0. The value 0 indicates that Router A declares itself a slave.

d. After Router B receives the packet, Router B's status changes to Exchange and Router B sends a new DD packet containing its own LSA summaries to Router A. The value of the Seq field carried in the new DD packet is changed to *y* + 1.

Router A uses the same sequence number as Router B to confirm that it has received DD packets from Router B. Router B uses the sequence number plus one to confirm that it has received DD packets from Router A. When Router B sends the last DD packet, it sets the M field of the packet to 0.

3. LSDB synchronization

   a. After Router A receives the last DD packet, it finds that many LSAs in Router B's LSDB do not exist in its own LSDB, so Router A's status changes to Loading. After Router B receives the last DD packet from Router A, Router B's status directly changes to Full, because Router B's LSDB already contains all LSAs of Router A.

   b. Router A sends an LSR packet for updating LSAs to Router B. Router B returns an LSU packet to Router A. After Router A receives the packet, it sends an LSAck packet for acknowledgement.

   The preceding procedures continue until the LSAs in Router A's LSDB are the same as those in Router B's LSDB. Router A's status changes to Full. After Router A and Router B exchange DD packets and update all LSAs, they establish an adjacency.

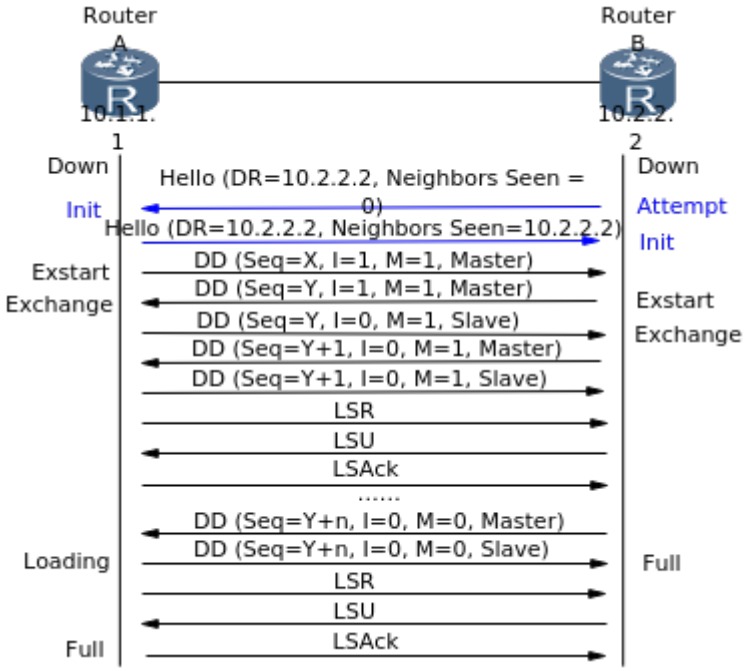
Adjacency establishment on an NBMA network

 Copyright © Huawei Technologies Co., Ltd.

The adjacency establishment process on an NBMA network is similar to that on a broadcast network. The blue part shown in **Figure 11-110** highlights the differences from a broadcast network.

On an NBMA network, all Routers establish adjacencies only with the DR and BDR.

Figure 11-110 Adjacency establishment process on an NBMA network



The adjacency establishment process on an NBMA network is as follows:

1. Neighbor relationship establishment
   a. After Router B sends a Hello packet to a Down interface of Router A, Router B's status changes to Attempt. The packet carries the DR field of 2.2.2.2 (ID of Router B) and the Neighbors Seen field of 0. A neighbor Router has not been discovered, and Router B regards itself as a DR.
   b. After Router A receives the packet, Router A's status changes to Init and Router A returns a Hello packet. The returned packet carries the DR and Neighbors Seen fields of 2.2.2.2. Router B has been discovered but its router ID is greater than that of Router A, and therefore Router A agrees that Router B is a DR.

   NOTE

   The following procedures are not performed for DR others on an NBMA network.
2. The procedures for negotiating a master/slave relationship and exchanging DD packets on an NBMA network are the same as those on a broadcast network.
3. The procedure for synchronizing LSDBs (LSA request, LSA transmission, and LSA response) on an NBMA network is the same as that on a broadcast network.

 Copyright © Huawei Technologies Co., Ltd.

Adjacency establishment on a point-to-point (P2P)/point-to-multipoint (P2MP) network

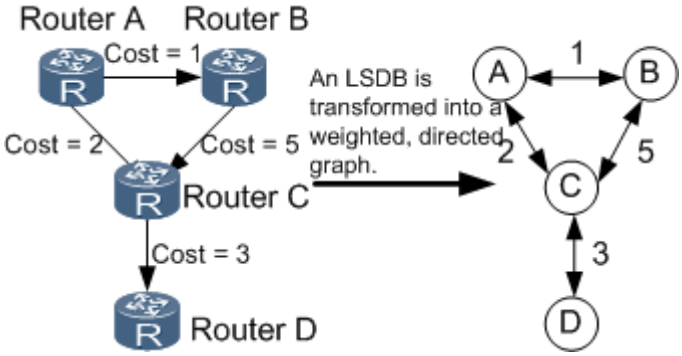
The adjacency establishment process on a P2P/P2MP network is similar to that on a broadcast network. On a P2P/P2MP network, however, no DR or BDR needs to be elected and DD packets are transmitted in multicast mode.

Route Calculation

OSPF uses the shortest path first (SPF) algorithm to calculate routes, resulting in fast route convergence.

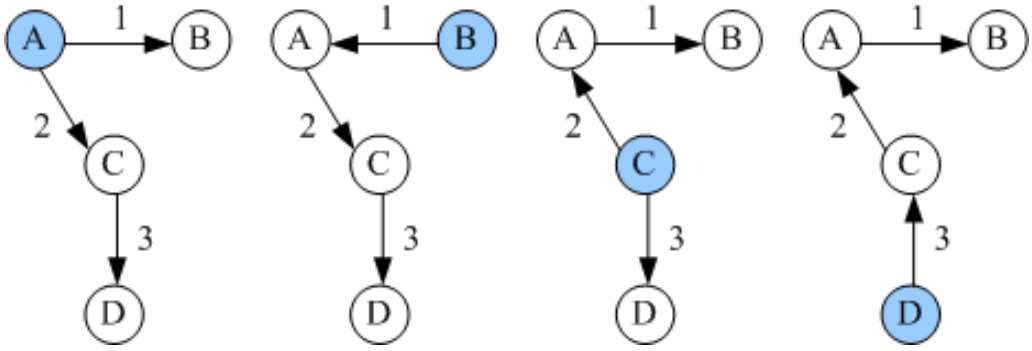
OSPF uses an LSA to describe the network topology. A Type 1 LSA describes the attributes of a link between Routers. A Router transforms its LSDB into a weighted, directed graph, which reflects the topology of the entire AS. All Routers in the same area have the same graph. **Figure 11-111** shows a weighted, directed graph.

Figure 11-111 Weighted, directed graph



Based on the graph, each Router uses an SPF algorithm to calculate an SPT with itself as the root. The SPT shows routes to nodes in the AS. **Figure 11-112** shows an SPT.

Figure 11-112 SPT



When a Router's LSDB changes, the Router recalculates a shortest path. Frequent SPF calculations consume a large amount of resources and affect Router efficiency. Changing the interval between SPF calculations can prevent resource consumption caused by frequent LSDB changes. The default interval between SPF calculations is 5 seconds.

 Copyright © Huawei Technologies Co., Ltd.

The route calculation process is as follows:

1. A Router calculates intra-area routes.

   The Router uses an SFP algorithm to calculate shortest paths to other Routers in an area. Type 1 and Type 2 LSAs accurately describe the network topology in an area. Based on the network topology described by a Type 1 LSA, the Router calculates paths to other Routers in the area.

   NOTE

   If multiple equal-cost routes are produced during route calculation, the SPF algorithm retains all these routes in the LSDB.

2. The Router calculates inter-area routes.

   The network segment of the routes in an adjacent area can be considered to be directly connected to the area border router (ABR). Because the shortest path to the ABR has been calculated in the preceding phase, the Router can directly check a Type 3 LSA to obtain the shortest path to the network segment. The autonomous system boundary router (ASBR) can also be considered to be connected to the ABR. Therefore, the shortest path to the ASBR can also be calculated in this phase.

   NOTE

   If the Router performing an SPF calculation is an ABR, the Router needs to check only Type 3 LSAs in the backbone area.

3. The Router calculates AS external routes.

   AS external routes can be considered to be directly connected to the ASBR. Because the shortest path to the ASBR has been calculated in the preceding phase, the Router can check Type 5 LSAs to obtain the shortest paths to other ASs.

11.12.6.2.3 OSPF Route Control

You can use the following features to control the advertising and receiving of Open Shortest Path First (OSPF) routing information to import routes from other protocols:

- **Route summarization**

  Route summarization enables a Router to summarize routes with the same prefix into a single route and to advertise only the summarized route to other areas. Route summarization reduces the size of a routing table and improves Router performance.

- **Route filtering**

  OSPF can use routing policies to filter routes. By default, OSPF does not filter routes.

- **Smart-discover**

  Smart-discover enables a Router to send Hello packets to its neighbors without waiting for the Hello timer to expire, which speeds up neighbor relationship establishment.

Route Summarization

When a large OSPF network is deployed, an OSPF routing table includes a large number of routing entries. To accelerate route searches and simplify management,

 Copyright © Huawei Technologies Co., Ltd.

configure route summarization to reduce the size of the OSPF routing table. If a link frequently alternates between Up and Down, the links not involved in the route summarization are not affected. This process prevents route flapping and improves network stability.

Route summarization can be carried out in the following manners:

- Area border router (ABR) summarization

  When an ABR transmits routing information to other areas, it generates Type 3 LSAs for each network segment. If consecutive network segments exist in this area, you can manually summarize these network segments into a single network segment. The ABR generates one LSA for the summarized network segment and advertises only that LSA.

- Autonomous system boundary router (ASBR) summarization

  If route summarization has been configured and the local Router is an ASBR, the local Router summarizes imported Type 5 LSAs within the summarized address range. If a not-so-stubby area (NSSA) has been configured, the local Router also summarizes imported Type 7 LSAs within the summarized address range.

  If the local Router is both an ASBR and an ABR, it summarizes Type 5 LSAs translated from Type 7 LSAs.

Route Filtering

OSPF routing policies include access control lists (ACLs), IP prefix lists, and route-policies. For details about these policies, see the section "Routing Policy" in the *Feature Description - IP Routing*.

OSPF route filtering applies in the following aspects:

- Route import

  OSPF can import the routes learned by other routing protocols. A Router uses a configured routing policy to filter routes and imports only the routes matching the routing policy. Only an ASBR can import routes, and therefore a routing policy for importing routes must be configured on the ASBR.

- Advertising of imported routes

  A Router advertises imported routes to its neighbors. Only an ASBR can import routes, and therefore a routing policy for the advertising of imported routes must be configured on the ASBR.

- Route learning

  A Router uses a routing policy to filter received intra-area, inter-area, and AS external routes. The Router adds only the routes matching the routing policy to its routing table. All routes can still be advertised from an OSPF routing table.

  The Router filters only routes calculated based on LSAs, and therefore learned LSAs are complete.

- Inter-area LSA learning

  An ABR in an area can be configured to filter Type 3 LSAs advertised to the area. The ABR can advertise only Type 3 LSAs, and therefore a routing policy for inter-area LSA learning must be configured on the ABR.

 Copyright © Huawei Technologies Co., Ltd.

During inter-area LSA learning, the ABR directly filters Type 3 LSAs advertised to the area.

- Inter-area LSA advertising

  An ABR in an area can be configured to filter Type 3 LSAs advertised to other areas. The ABR can advertise only Type 3 LSAs, and therefore a routing policy for inter-area LSA advertising must be configured on the ABR.

Smart-Discover

Generally, Routers periodically send Hello packets through OSPF interfaces. OSPF uses a Hello timer to control the interval at which Hello packets are sent. A Router can send Hello packets again only when the Hello timer expires. Neighbors keep waiting to receive Hello packets until the Hello timer expires. This process delays the establishment of OSPF neighbor relationships. You can enable smart-discover to resolve this issue.

After you enable smart-discover on a Router, the Router sends Hello packets to its neighbors without waiting for the Hello timer to expire. Smart-discover ensures that neighbors can quickly receive Hello packets and perform a fast state transition.

Enable smart-discover when either of the following cases is true:

- The neighbor status becomes 2-way for the first time.
- The neighbor status changes from 2-way or higher to Init.

11.12.6.2.4 OSPF NSSA

Background

As defined in the Open Shortest Path First (OSPF) protocol, stub areas cannot import external routes. This mechanism prevents external routes from consuming the bandwidth and storage resources of Routers in stub areas. If you need to both import external routes and prevent resource consumption caused by external routes, you can configure not-so-stubby areas (NSSAs).

Derived from stub areas, NSSAs resemble stub areas in many ways. Different from stub areas, NSSAs can import autonomous system (AS) external routes and advertise them within the entire AS, without learning external routes from other areas in the AS.

Related Concepts

- **N-bit**

  A Router uses the N-bit carried in a Hello packet to identify the area type that it supports. The same area type must be configured for all Routers in an area. If Routers have different area types, they cannot establish OSPF neighbor relationships. Some vendors' devices do not comply with RFC 1587, and the N-bit is also set in OSPF Database Description (DD) packets. You can manually set the N-bit on a Router to interwork with the vendors' devices.

- **Type 7 LSA**

  Type 7 LSAs, which describe imported external routes, are introduced to support NSSAs. Type 7 LSAs are generated by an autonomous system

 Copyright © Huawei Technologies Co., Ltd.

boundary router (ASBR) in an NSSA and advertised only within the NSSA. After an area border router (ABR) in an NSSA receives Type 7 LSAs, it selectively translates Type 7 LSAs into Type 5 LSAs to advertise external routes to other areas on an OSPF network.

Principles

To advertise external routes imported by an NSSA to other areas, a translator must translate Type 7 LSAs into Type 5 LSAs. Notes for an NSSA are as follows:

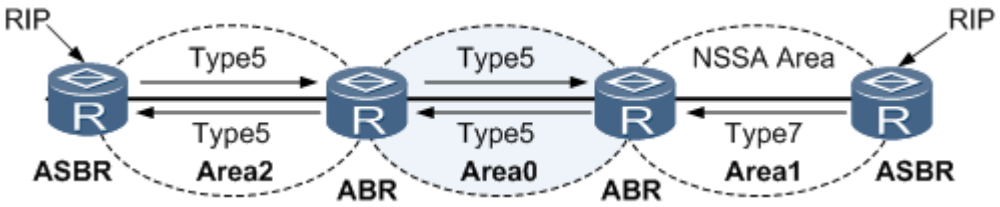
- By default, the translator is the ABR with the largest router ID in the NSSA.
- The propagate bit (P-bit) is used to notify a translator whether Type 7 LSAs need to be translated.
- Only Type 7 LSAs with the P-bit set and a non-zero forwarding address (FA) can be translated into Type 5 LSAs. An FA indicates that packets to a destination address will be forwarded to the address specified by the FA.

  NOTE

  FA indicates that the packet to a specific destination address is to be forwarded to the address specified by.

  The loopback interface address in an area is preferentially selected as the FA. If no loopback interface exists, the address of the interface that is Up and has the largest logical index in the area is selected as the FA.
- The P-bit is not set for Type 7 LSAs generated by an ABR.

Figure 11-113 NSSA



Advantages

Multiple ABRs may be deployed in an NSSA. To prevent routing loops caused by default routes, ABRs do not calculate the default routes advertised by each other.

11.12.6.2.5 OSPF Fast Convergence

OSPF fast convergence is an extended feature of OSPF implemented to speed up the convergence of routes. It includes the following components:

- Incremental SPF (I-SPF): recalculates only the routes of the changed nodes rather than all the nodes when the network topology changes. This speeds up route calculation.
- Partial Route Calculation (PRC): calculates only the changed routes when the routes on the network change.
- An OSPF intelligent timer: can dynamically adjust its value based on the user's configuration and the interval at which an event is triggered, such as the route calculation interval, which ensures rapid and stable network operation.

 Copyright © Huawei Technologies Co., Ltd.

OSPF intelligent timer applies the exponential backoff technology so that the value of the timer can reach the millisecond level.

I-SPF

In ISO 10589, the Dijkstra algorithm was adopted to calculate routes. When a node changes on the network, this algorithm is used to recalculate all routes. The calculation takes a long time and consumes too many CPU resources, which affects the convergence speed.

I-SPF improves the Dijkstra algorithm. Except for the first time, only changed nodes instead of all nodes are involved in calculation. The SPT generated at last is the same as that generated by the previous algorithm, but I-SPF decreases CPU usage and speeds up network convergence.

PRC

Similar to I-SPF, PRC calculates only the changed routes. However, PRC does not calculate the shortest path. It updates routes based on the SPT calculated by I-SPF.

In route calculation, a leaf represents a route, and a node represents a router. SPT and leaf changes both cause the change of routing information, but the SPT change is irrelevant to the leaf change. PRC processes routing information based on either SPT or leaf information.

- If the SPT changes, PRC processes the routing information of all leaves on a changed node.
- If the SPT does not change, PRC does not process the routing information on any node.
- If a leaf changes, RPC processes the routing information on the leaf only.
- If a leaf does not change, PRC does not process the routing information on any leaf.

For example, if OSPF is enabled on an interface of a node, the SPT calculated by I-SPF remains unchanged. PRC updates only the routes of this interface, consuming less CPU resources.

PRC improves on the SPF algorithm. Working with I-SPF, RPC further improves the convergence performance of the network.

NOTE

In the implementation of a device, only I-SPF and PRC are used to calculate OSPF routes.

OSPF Intelligent Timer

On an unstable network, routes are calculated frequently, which consumes a great number of CPU resources. In addition, LSPs that describe the unstable topology are generated and transmitted on the unstable network. Frequently processing such LSAs affects the rapid and stable operation of the entire network.

To speed up route convergence on the entire network, the OSPF intelligent timer controls route calculation, LSA generation, and LSA receiving.

OSPF intelligent timer speeds up route convergence in the following modes:

 Copyright © Huawei Technologies Co., Ltd.

- On a network where routes are calculated repeatedly, the OSPF intelligent timer dynamically adjusts the route calculation based on user's configuration and the exponential backoff technology. The number of route calculation times and the CPU resource consumption are decreased. Routes are calculated after the network topology becomes stable.
- On an unstable network, if a router generates or receives LSAs due to frequent topology changes, the OSPF intelligent timer can dynamically adjust its value. No LSA is generated or handled within an interval, which prevents invalid LSAs from being generated and advertised on the entire network.

The OSPF intelligent timer is started by default and uses the default value.

11.12.6.2.6 OSPF NSR

Non-Stop Routing (NSR) is a routing technique that prevents neighboring routers from detecting faults on the control plane of a device with a backup control plane. With NSR, when the control plane of the device fails, the neighbor relationship set up through specific routing protocols, MPLS, and other protocols that carry services continue uninterrupted.

As networks develop and expand, carriers place increasingly higher requirements on the reliability of IP networks. NSR was introduced as a high availability (HA) solution to ensure that services transmitted by a device were unaffected by hardware or software failures on the device.

OSPF NSR synchronizes the protocol data on the master main control board to the slave main control board in real time. When the master main control board fails or needs to be upgraded, the slave main control board immediately takes over services from the master main control board, and the neighbor remains unaware the fault occurred. Real-time data synchronization between the master and slave main control boards through OSPF NSR is implemented as follows:

- OSPF backs up configuration data and dynamic data, including information about interfaces, neighbors, and LSDBs.
- OSPF does not back up routes, shortest path trees (SPTs), or Traffic Engineering Databases (TEDBs). Such information can be restored using the source data in the database backup process.
- When the master/slave main control board switchover occurs, the new master main control board restores the operations data and takes over services from the former master main control board. Throughout this process, the neighbor remains unaware of the fault.

11.12.6.2.7 OSPF Authentication

OSPF authentication encrypts OSPF packets by adding the authentication field to packets to ensure network security. When a local device receives OSPF packets from a remote device, the local device discards the packets if the authentication passwords carried in these packets do not match the local one, which protects the local device from potential attacks.

In terms of the packet type, the authentication is classified as follows:

- Area authentication

  Area authentication is configured in the OSPF area view and applies to packets received by all interfaces in the OSPF area.

 Copyright © Huawei Technologies Co., Ltd.

- Interface authentication

  Interface authentication is configured in the interface view and applies to all packets received by the interface.

In terms of packet the authentication modes, the authentication is classified as follows:

- Non-authentication

  Authentication is not required.
- Simple authentication

  The authenticated party directly adds the configured password to packets for authentication. This authentication mode provides the lowest password security.
- MD5 authentication

  The authenticated party encrypts the configured password using a Message Digest 5 (MD5) algorithm and adds the ciphertext password to packets for authentication. This authentication mode improves password security. The supported MD5 algorithms include MD5 and HMAC-MD5.

  NOTICE

  MD5 and HMAC-MD5 are insecure encryption algorithms. Therefore, using HMAC-SHA256 is recommended.
- HMAC-SHA256 authentication

  A password is encrypted using the HMAC-SHA256 algorithm before it is added to the packet, which improves password security.

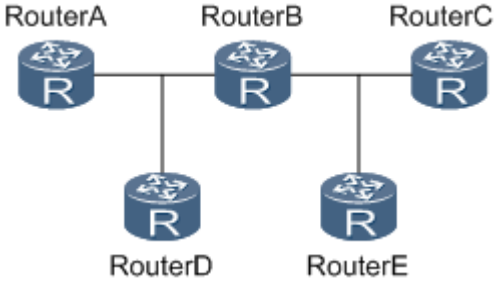
OSPF carries authentication types in packet headers and authentication information in packet tails.

The authentication types include:

- 0: Non-authentication
- 1: Simple authentication
- 2: Ciphertext authentication

Usage Scenario

Figure 11-114 Networking for OSPF authentication on a broadcast network



The configuration requirements are as follows:

 Copyright © Huawei Technologies Co., Ltd.

- The interface authentication configurations must be the same on all devices on the same network so that OSPF neighbor relationships can be established.
- The area authentication configurations must be the same on all devices in the same area.

11.12.6.2.8 OSPF Packet Format

Open Shortest Path First (OSPF) packets are encapsulated into IP packets. The OSPF protocol number is 89. OSPF packets are classified into the following types:

- **Hello packet**
- **Database Description (DD) packet**
- **Link State Request (LSR) packet**
- **Link State Update (LSU) packet**
- **Link State Acknowledgment (LSAck) packet**

Packet Header Format

The five types of OSPF packets have the same packet header format. The length of an OSPF packet header is 24 bytes. **Figure 11-115** shows an OSPF packet header.

Figure 11-115 OSPF packet header

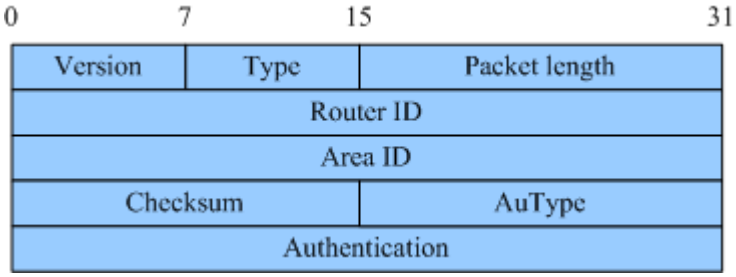


Table 11-42 Packet header fields

Field	Length	Description
Version	8 bits	OSPF version number. For OSPFv2, the value is 2.
Type	8 bits	OSPF packet type. The values are as follows: • 1: Hello packet • 2: DD packet • 3: LSR packet • 4: LSU packet • 5: LSAck packet
Packet length	16 bits	Length of the OSPF packet containing the packet header, in bytes.
Router ID	32 bits	ID of the Router that sends the OSPF packet.

 Copyright © Huawei Technologies Co., Ltd.

Field	Length	Description
Area ID	32 bits	ID of the area to which the Router that sends the OSPF packet belongs.
Checksum	16 bits	Checksum of the OSPF packet that does not contain the Authentication field.
AuType	16 bits	Authentication type. The values are as follows: • 0: non-authentication • 1: simple authentication • 2: message digest algorithm 5 (MD5) authentication **NOTICE** MD5 is an insecure encryption algorithm. Therefore, using HMAC-SHA256 is recommended.
Authentication	64 bits	This field has different meanings for different AuType values: • 0: This field is not defined. • 1: This field defines password information. • 2: This field contains the key ID, MD5 authentication data length, and sequence number.

NOTE

MD5 authentication data is added to an OSPF packet and is not included in the Authentication field.

Hello Packet

Hello packets are commonly used packets, which are periodically sent on OSPF interfaces to establish and maintain neighbor relationships. A Hello packet includes information about the designated router (DR), backup designated router (BDR), timers, and known neighbors. **Figure 11-116** shows the format of a Hello packet.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-116 Format of a Hello packet

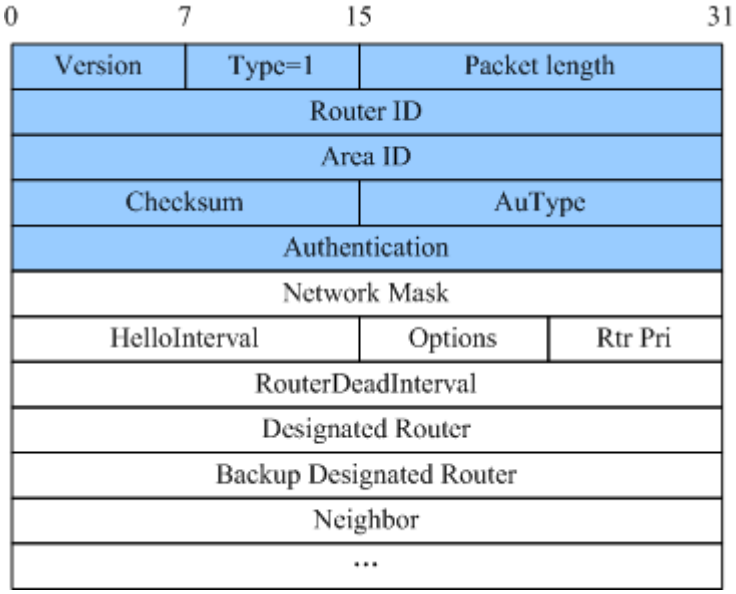


Table 11-43 Hello packet fields

Field	Length	Description
Network Mask	32 bits	Mask of the network on which the interface that sends the Hello packet resides.
HelloInterval	16 bits	Interval at which Hello packets are sent.
Options	8 bits	The values are as follows: • E: Type 5 link state advertisements (LSAs) are flooded. • MC: IP multicast packets are forwarded. • N/P: Type 7 LSAs are processed. • DC: On-demand links are processed.
Rtr Pri	8 bits	DR priority. The default value is 1. **NOTE** If the DR priority of a Router interface is set to 0, the interface cannot participate in a DR or BDR election.
RouterDeadInterval	32 bits	Dead interval. If a Router does not receive any Hello packets from its neighbors within a specified dead interval, the neighbors are considered Down.
Designated Router	32 bits	Interface address of the DR.
Backup Designated Router	32 bits	Interface address of the BDR.
Neighbor	32 bits	Router ID of the neighbor.

 Copyright © Huawei Technologies Co., Ltd.

Table 11-44 lists the address types, interval types, and default intervals used when Hello packets are transmitted on different networks.

Table 11-44 Hello packet characteristics for various network types

Network Type	Address Type	Interval Type	Default Interval
Broadcast	Multicast address	HelloInterval	10 seconds
Non-broadcast multiple access (NBMA)	Unicast address	• HelloInterval for the DR, BDR, and Router that can become a DR • PollInterval for the case when neighbors become Down and HelloInterval for other cases	30 seconds for HelloInterval 120 seconds for PollInterval
Point-to-point (P2P)	Multicast address	HelloInterval	10 seconds
Point-to-multipoint (P2MP)	Unicast address	HelloInterval	30 seconds

NOTE

To establish neighbor relationships between Routers on the same network segment, you must set the same HelloInterval, PollInterval, and RouterDeadInterval values for the Routers. PollInterval applies only to NBMA networks.

DD Packet

During an adjacency initialization, two Routers use DD packets to describe their own link state databases (LSDBs) for LSDB synchronization. A DD packet contains the header of each LSA in an LSDB. An LSA header uniquely identifies an LSA. The LSA header occupies only a small portion of the LSA, which reduces the amount of traffic transmitted between Routers. A neighbor can use the LSA header to check whether it already has the LSA. When two Routers exchange DD packets, one functions as the master and the other functions as the slave. The master defines a start sequence number. The master increases the sequence number by one each time it sends a DD packet. After the slave receives a DD packet, it uses the sequence number carried in the DD packet for acknowledgement.

Figure 11-117 shows the format of a DD packet.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-117 Format of a DD packet

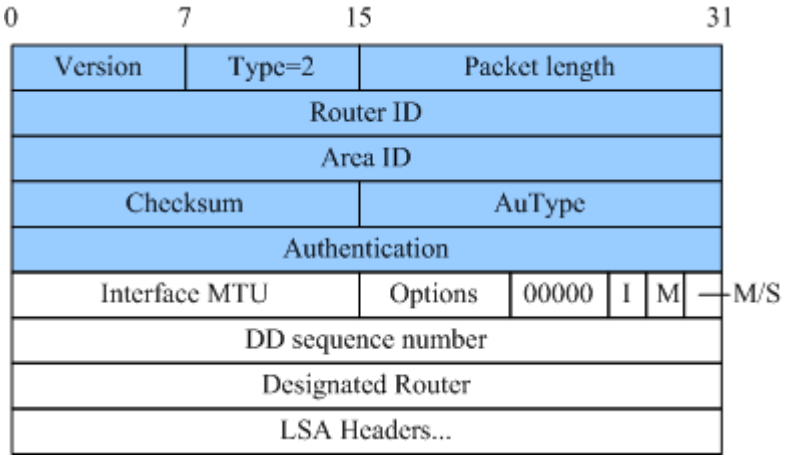


Table 11-45 DD packet fields

Field	Length	Description
Interface MTU	16 bits	Maximum length of the DD packet sent by the interface with packet fragmentation disabled.
Options	8 bits	The values are as follows: • E: Type 5 LSAs are flooded. • MC: IP multicast packets are forwarded. • N/P: Type 7 LSAs are processed. • DC: On-demand links are processed.
I	1 bit	If the DD packet is the first packet among multiple consecutive DD packets sent by a Router, this field is set to 1. In other cases, this field is set to 0.
M (More)	1 bit	If the DD packet is the last packet among multiple consecutive DD packets sent by a Router, this field is set to 0. In other cases, this field is set to 1.
M/S (Master/ Slave)	1 bit	When two Routers exchange DD packets, they negotiate a master/slave relationship. The Router with a larger router ID becomes the master. If this field is set to 1, the DD packet is sent by the master.
DD sequence number	32 bits	Sequence number of the DD packet. The master and slave use the sequence number to ensure that DD packets are correctly transmitted.
LSA Headers	-	LSA header information included in the DD packet.

 Copyright © Huawei Technologies Co., Ltd.

LSR Packet

After two Routers exchange DD packets, they send LSR packets to request each other's LSAs. The LSR packets contain the summaries of the requested LSAs. **Figure 11-118** shows the format of an LSR packet.

Figure 11-118 Format of an LSR packet

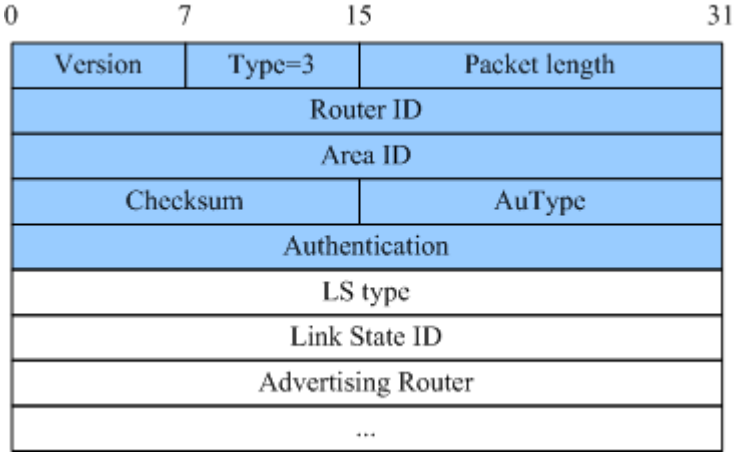


Table 11-46 LSR packet fields

Field	Length	Description
LS type	32 bits	Type of the LSA
Link State ID	32 bits	This field together with the LS type field describes an LSA in an AS.
Advertising Router	32 bits	Router ID of the Router that generates the LSA.

NOTE

The LS type, Link State ID, and Advertising Router fields can uniquely identify an LSA. If two LSAs have the same LS type, Link State ID, and Advertising Router fields, a Router uses the LS sequence number, LS checksum, and LS age fields to obtain a required LSA.

LSU Packet

A Router uses an LSU packet to transmit LSAs requested by its neighbors or to flood its own updated LSAs. The LSU packet contains a set of LSAs. For multicast and broadcast networks, LSU packets are multicast to flood LSAs. To ensure reliable LSA flooding, a Router uses an LSAck packet to acknowledge the LSAs contained in an LSU packet that is received from a neighbor. If an LSA fails to be acknowledged, the Router retransmits the LSA to the neighbor. **Figure 11-119** shows the format of an LSU packet.

Copyright © Huawei Technologies Co., Ltd.

Figure 11-119 Format of an LSU packet

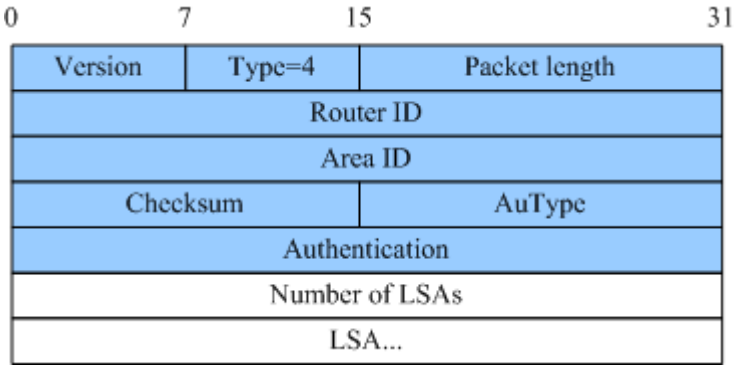


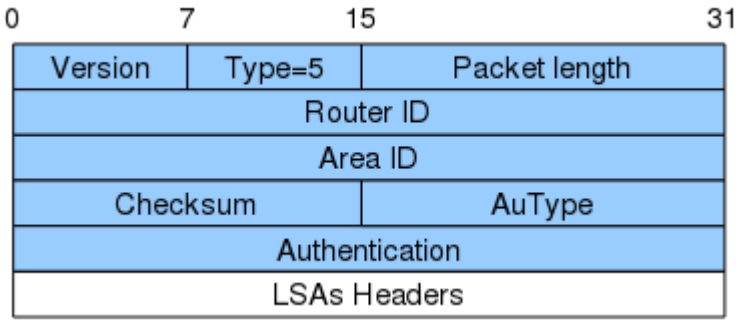
Table 11-47 LSU packet field

Field	Length	Description
Number of LSAs	32 bits	Number of LSAs contained in the LSU packet

LSAck Packet

A Router uses an LSAck packet to acknowledge the LSAs contained in a received LSU packet. The LSAs can be acknowledged using LSA headers. LSAck packets can be transmitted over different links in unicast or multicast mode. **Figure 11-120** shows the format of an LSAck packet.

Figure 11-120 Format of an LSAck packet



 Copyright © Huawei Technologies Co., Ltd.

Table 11-48 LSAck packet field

Field	Length	Description
LSAs Headers	Determined by the header length of the LSA to be acknowledged.	This field is used to acknowledge an LSA.

11.12.6.2.9 OSPF LSA Format

Each Router in an autonomous system (AS) generates one or more types of link state advertisements (LSAs), depending on the Router's type. Multiple LSAs form a link state database (LSDB). Open Shortest Path First (OSPF) encapsulates routing information into LSAs for transmission. Commonly used LSAs include:

- **Router-LSAs (Type 1)**
- **Network-LSAs (Type 2)**
- **Summary-LSAs**, including network-summary-LSAs (Type 3) and ASBR-summary-LSAs (Type 4)
- **AS-external-LSAs (Type 5)**

LSA Header Format

All LSAs have the same header. **Figure 11-121** shows an LSA header.

Figure 11-121 LSA header

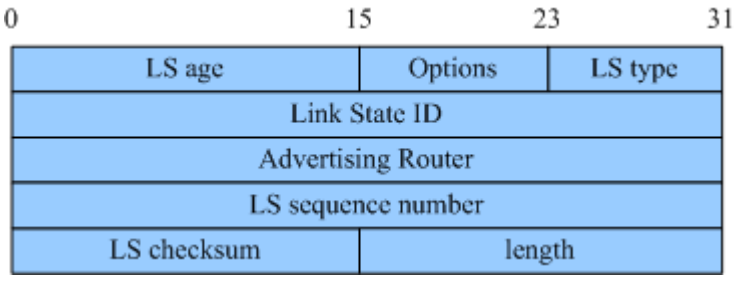


Table 11-49 LSA header fields

Field	Length	Description
LS age	16 bits	Time that elapses after the LSA is generated, in seconds. The value of this field continually increases regardless of whether the LSA is transmitted over a link or saved in an LSDB.

 Copyright © Huawei Technologies Co., Ltd.

Field	Length	Description
Options	8 bits	The values are as follows: • E: Type 5 LSAs are flooded. • MC: IP multicast packets are forwarded. • N/P: Type 7 LSAs are processed. • DC: On-demand links are processed.
LS type	8 bits	Type of the LSA. The values are as follows: • Type1: Router-LSA • Type2: Network-LSA • Type3: Network-summary-LSA • Type4: ASBR-summary-LSA • Type5: AS-external-LSA • Type7: NSSA-LSA
Link State ID	32 bits	This field together with the LS type field describes an LSA in an AS.
Advertising Router	32 bits	Router ID of the Router that generates the LSA.
LS sequence number	32 bits	Sequence number of the LSA. Neighbors can use this field to identify the latest LSA.
LS checksum	16 bits	Checksum of all fields except the LS age field.
length	16 bits	Length of the LSA including the LSA header, in bytes.

Router-LSA

A router-LSA describes the link status and cost of a Router. Router-LSAs are generated by a Router and advertised within the area to which the Router belongs. **Figure 11-122** shows the format of a router-LSA.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-122 Format of a router-LSA

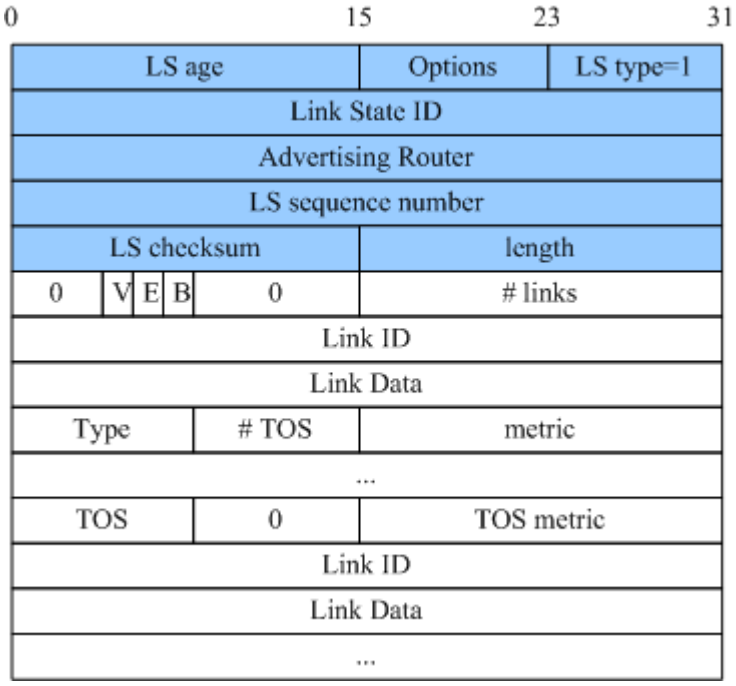


Table 11-50 Router-LSA fields

Field	Length	Description
Link State ID	32 bits	Router ID of the Router that generates the LSA.
V (Virtual Link)	1 bit	If the Router that generates the LSA is located at one end of a virtual link, this field is set to 1. In other cases, this field is set to 0.
E (External)	1 bit	If the Router that generates the LSA is an autonomous system boundary router (ASBR), this field is set to 1. In other cases, this field is set to 0.
B (Border)	1 bit	If the Router that generates the LSA is an area border router (ABR), this field is set to 1. In other cases, this field is set to 0.
# links	16 bits	Number of links and interfaces described in the LSA, including all links and interfaces in the area to which the Router belongs.
Link ID	32 bits	Object to which the Router is connected. This field has different meanings for different link types: ● 1: router ID ● 2: interface IP address of the designated router (DR) ● 3: network segment or subnet number ● 4: router ID of the neighbor on a virtual link

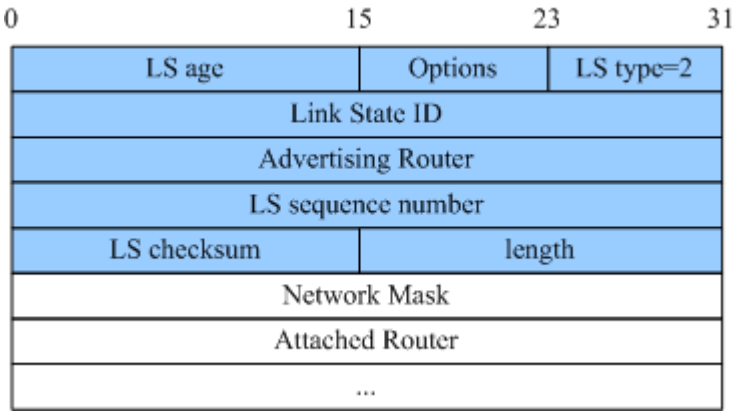
 Copyright © Huawei Technologies Co., Ltd.

Field	Length	Description
Link Data	32 bits	Link data. This field has different meanings for different link types: • 1: interface index • 3: subnet mask • 2 and 4: interface address of the Router
Type	8 bits	Type of the Router link. The values are as follows: • 1: The Router is connected to another Router in point-to-point (P2P) mode. • 2: The Router is connected to a transport network. • 3: The Router is connected to a stub network. • 4: The Router is connected to another Router over a virtual link.
# TOS	8 bits	Number of types of service (ToSs).
metric	16 bits	Cost of the link.
TOS	8 bits	Type of service.
TOS metric	16 bits	Metric for the specified ToS.

Network-LSA

A network-LSA describes the link status of all Routers on the local network segment. Network-LSAs are generated by a DR on a broadcast or non-broadcast multiple access (NBMA) network and advertised within the area to which the DR belongs. **Figure 11-123** shows the format of a network-LSA.

Figure 11-123 Format of a network-LSA



 Copyright © Huawei Technologies Co., Ltd.

Table 11-51 Network-LSA fields

Field	Length	Description
Link State ID	32 bits	Interface IP address of the DR
Network Mask	32 bits	Mask of the broadcast or NBMA network
Attached Router	32 bits	Router IDs of all Routers on the broadcast or NBMA network, including the router ID of the DR

Summary-LSA

A network-summary-LSA describes routes on a network segment in an area. The routes are advertised to other areas.

An ASBR-summary-LSA describes routes to the ASBR in an area. The routes are advertised to all areas except the area to which the ASBR belongs.

The two types of summary-LSAs have the same format and are generated by an ABR. **Figure 11-124** shows the format of a summary-LSA.

Figure 11-124 Format of a summary-LSA

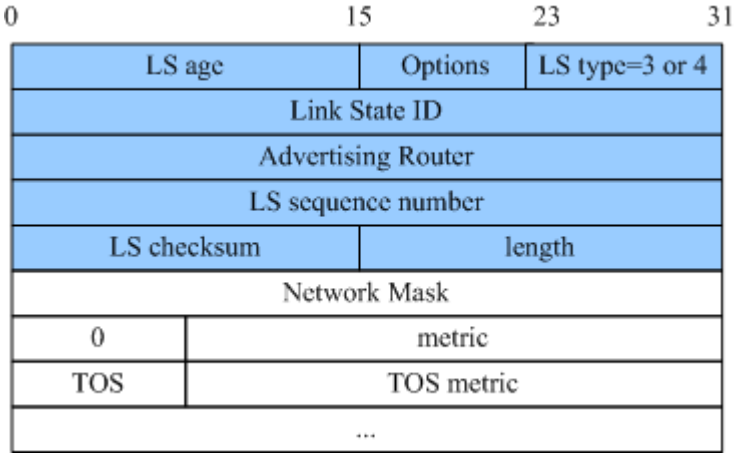


Table 11-52 Network-summary-LSA fields

Field	Length	Description
Link State ID	32 bits	Advertised network address
Network Mask	32 bits	Mask of the broadcast or NBMA network
metric	24 bits	Cost of the route to the destination address
TOS	8 bits	Type of service.

Copyright © Huawei Technologies Co., Ltd.

Field	Length	Description
TOS metric	24 bits	Metric for the specified ToS

NOTE

When a default route is advertised, both the Link State ID and Network Mask fields are set to 0.0.0.0.

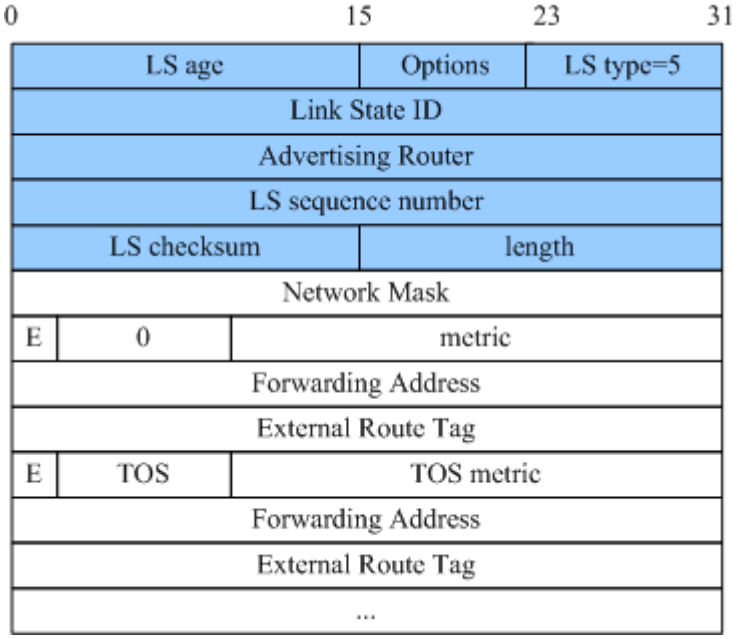
Table 11-53 ASBR-summary-LSA fields

Field	Length	Description
Link State ID	32 bits	Router ID of the ASBR
Network Mask	32 bits	Set to 0.0.0.0
metric	24 bits	Cost of the route to the destination address
TOS	8 bits	Type of service.
TOS metric	24 bits	Metric for the specified ToS

AS-External-LSA

An AS-external-LSA describes AS external routes. AS-external-LSAs are generated by an ASBR. Among the five types of LSAs, only AS-external-LSAs can be advertised to all areas except stub areas and not-so-stubby areas (NSSAs). **Figure 11-125** shows the format of an AS-external-LSA.

Figure 11-125 Format of an AS-external-LSA



 Copyright © Huawei Technologies Co., Ltd.

Table 11-54 AS-external-LSA fields

Field	Length	Description
Link State ID	32 bits	Advertised network address
Network Mask	32 bits	Mask of the advertised destination address.
E	1 bit	Type of the external route. The values are as follows: ● 0: Type 1 external route ● 1: Type 2 external route
metric	24 bits	Cost of the route to the destination address
Forwarding Address	32 bits	Packets destined for the advertised destination address are forwarded to the address specified by this field.
External Route Tag	32 bits	Tag added to the external route. This field can be used to manage external routes. OSPF itself does not use this field.
TOS	8 bits	Type of service.
TOS metric	24 bits	Metric for the specified ToS.

NOTE

When AS-external-LSAs are used to advertise default routes, both the Link State ID and Network Mask fields are set to 0.0.0.0.

11.12.6.2.10 OSPF IP FRR

OSPF IP fast reroute (FRR) is a kind of dynamic IP FRR. To implement this feature, Interior Gateway Protocol (IGP) calculates the backup paths beforehand by using the link status database of the entire network and then stores the backup paths in the forwarding table to minimize service interruption duration in case of a link fault.

Background

As networks develop, voice over IP (VoIP) and online video services pose higher requirements for real-time transmission. Nevertheless, if a primary link fails, OSPF-enabled devices need to perform multiple operations, including detecting the fault, updating the link-state advertisement (LSA), flooding the LSA, calculating routes, and delivering forward information base (FIB) entries before switching traffic to a new link. This process takes a time much longer than the minimum delay that can cause users to sense service interruption. As a result, the requirements for real-time transmission cannot be met. OSPF IP FRR can solve this problem. OSPF IP FRR conforms to dynamic IP FRR defined by RFC 5286. With OSPF IP FRR, devices can quickly switch traffic from a faulty primary link to a backup link, protecting against a link or node failure.

 Copyright © Huawei Technologies Co., Ltd.

Related Concepts

OSPF IP FRR

OSPF IP FRR refers to a mechanism in which a device uses the loop-free alternate (LFA) algorithm to compute the next hop of a backup link and stores the next hop together with the primary link in the forwarding table. If the primary link fails, the device switches the traffic to the backup link before routes are converged on the control plane. This mechanism reduces the traffic interruption duration and minimizes the impacts.

OSPF IP FRR Policy

An OSPF IP FRR policy can be configured to filter alternate next hops. Only the alternate next hops that match the filtering rules of the policy can be added to the IP routing table.

LFA algorithm

A device uses each neighbor that can be used as a backup next hop as a root node, uses the shortest path first (SPF) algorithm to calculate the shortest path from each root node to the destination, and then uses the inequalities defined in RFC 5286 to calculate the backup link that is loop-free and has the lowest cost.

Networking

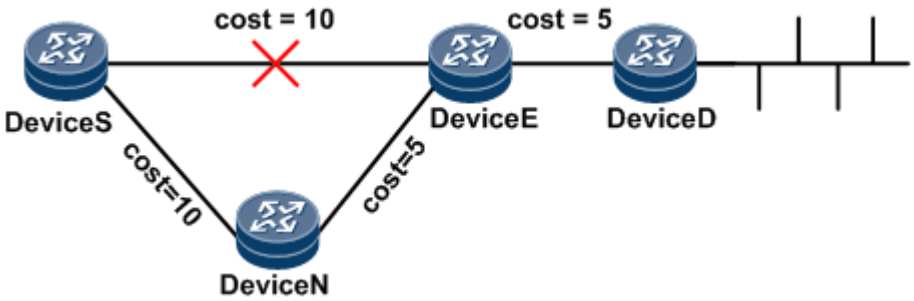
OSPF IP FRR protects traffic from a fault on a specified link.

In **Figure 11-126**, traffic flows from Device S to Device D. The primary link is Device S->Device E->Device D, and the backup link is Device S->Device N->Device E->Device D. If link costs meet the inequality: Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D) and OSPF IP FRR is enabled, Device S switches the traffic to the backup link if the primary link fails, reducing the traffic interruption duration.

NOTE

Distance_opt(X, Y) indicates the shortest link from X to Y. S stands for a source node, E for the faulty node, N for a node along a backup link, and D for a destination node.

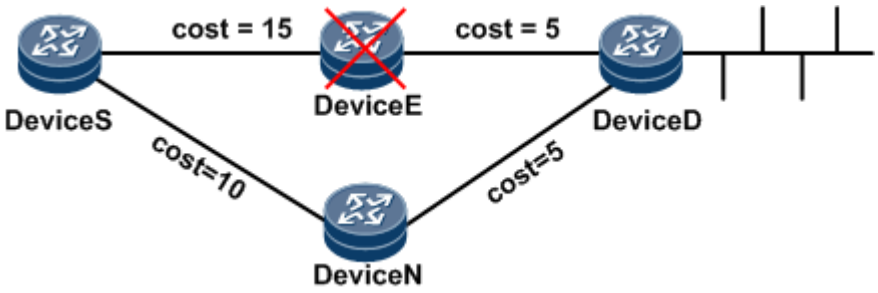
Figure 11-126 OSPF IP FRR link protection



The switching process in the following figure is similar. Therefore, the details are omitted.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-127 OSPF IP FRR node-and-link protection



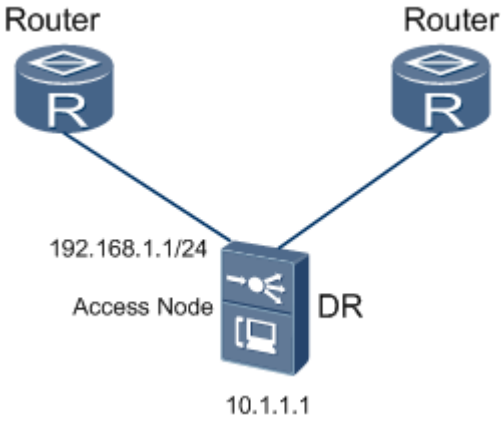
11.12.6.3 Configuration Example of OSPF

This topic provides an example for configuring OSPF on the MA5800.

Service Requirements

- OSPF is enabled on the MA5800.

Figure 11-128 Example network for configuring OSPF



Data Plan

Table 11-55 provides the data plan for configuring OSPF.

Table 11-55 Data plan for configuring OSPF

Item	Data
MA5800	IP address of the Layer 3 interface: 192.168.1.1/24
	Priority: 100
	VLAN ID: 2
	Router ID: 10.1.1.1

 Copyright © Huawei Technologies Co., Ltd.

Prerequisite

- The native VLAN of each interface of the MA5800 must be configured to ensure a normal communication.
- The OSPF area IDs of the MA5800 device and the routers must be consistent.

Procedure

Step 1 Configure MA5800.

1. Configure the IP address of the Layer 3 interface.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 192.168.1.1 24
huawei(config-if-vlanif2)#quit
```

2. Configure the OSPF Router ID.

```
huawei(config)#router id 10.1.1.1
```

3. Enable OSPF.

```
huawei(config)#ospf
huawei(config-ospf-1)#area 0
huawei(config-ospf-1-area-0.0.0.0)#network 192.168.1.0 0.0.0.255
huawei(config-ospf-1-area-0.0.0.0)#network 10.1.1.1 0.0.0.0
huawei(config-ospf-1-area-0.0.0.0)#quit
huawei(config-ospf-1)#quit
```

4. Configure the OSPF priority.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ospf dr-priority 100
huawei(config-if-vlanif2)#quit
```

5. Save the data.

```
huawei(config)#save
```

----End

Result

Run the **display ip routing-table** command and you can find the learned route table. Hosts can communicate with each other.

Configuration File

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ip address 192.168.1.1 24
quit
router id 10.1.1.1
ospf
area 0
network 192.168.1.0 0.0.0.255
network 10.1.1.1 0.0.0.0
quit
quit
interface vlanif 2
ospf dr-priority 100
quit
save
```

 Copyright © Huawei Technologies Co., Ltd.

11.12.6.4 Example for Configuring OSPF IP FRR

OSPF IP FRR can generate a loop-free backup link for fast link switching in case of a link failure.

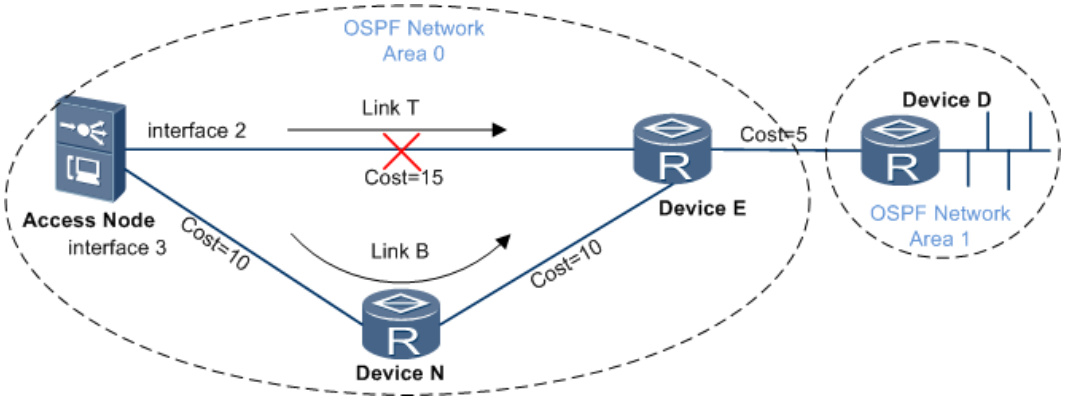
Service Requirements

Configuring OSPF IP FRR quickly switches traffic to the backup link without waiting for route convergence.

As shown in **Figure 11-129**,

- OSPF is running on all devices.
- The link cost meets the OSPF IP FRR traffic protection inequality.
- If the primary link T fails, traffic forwarded by the MA5800 is quickly switched to the backup link B.

Figure 11-129 Networking of configuring OSPF IP FRR



Data Plan

Table 11-56 describes data plan for configuring OSPF IP FRR.

Table 11-56 Data plan for configuring OSPF IP FRR

Item	Data
MA5800 port 2	IP address of the port: 10.1.2.1/24
	VLAN ID:3
	Cost:15
MA5800 port 3	IP address of the port: 10.1.3.1/24
	VLAN ID:4
	Cost:10

 Copyright © Huawei Technologies Co., Ltd.

Prerequisite

Before you configure OSPF IP FRR, complete the following tasks:

- Configuring link layer protocol parameters and IP addresses for ports to ensure that the link layer protocol on the ports is Up
- Enabling basic OSPF functions on routers
- Configuring basic BFD for OSPF functions on all devices in area 0 to quickly detect faults

Procedure

Step 1 Perform configuration on the MA5800.

1. Configure the IP address of the Layer 3 interface on the MA5800.

```
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 1
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ip address 10.1.2.1 24
huawei(config-if-vlanif3)#quit
huawei(config)#vlan 4 smart
huawei(config)#port vlan 4 0/9 2
huawei(config)#interface vlanif 4
huawei(config-if-vlanif4)#ip address 10.1.3.1 24
huawei(config-if-vlanif4)#quit
```

2. Configure the cost so that the traffic preferentially goes through link T.

```
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ospf cost 15
huawei(config-if-vlanif3)#quit
huawei(config)#interface vlanif 4
huawei(config-if-vlanif4)#ospf cost 10
huawei(config-if-vlanif4)#quit
```

3. Enable OSPF IP FRR on the MA5800.

```
huawei(config)#ospf
huawei(config-ospf-1)#frr
huawei(config-ospf-1-frr)#loop-free-alternate
huawei(config-ospf-1-frr)#quit
huawei(config-ospf-1)#quit
```

4. Save the data.

```
huawei(config)#save
```

----End

Result

Run the **display ospf routing** command to query the information about primary/secondary routes configured.

Configuration File

```
vlan 3 smart
port vlan 3 0/9 1
interface vlanif 3
ip address 10.1.2.1 24
quit
vlan 4 smart
port vlan 4 0/9 2
interface vlanif 4
```

 Copyright © Huawei Technologies Co., Ltd.

```
ip address 10.1.3.1 24
quit
interface vlanif 3
ospf cost 15
quit
interface vlanif 4
ospf cost 10
quit
ospf
frr
loop-free-alternate
quit
quit
save
```

11.12.6.5 OSPF Reference Standards and Protocols

Reference documents for OSPF are as follows:

Document	Description	Remarks
RFC 1583	OSPF Version 2	–
RFC 1765	OSPF Database Overflow	–
RFC 2328	This memo documents OSPFv2. Implementations of this memo are interoperable with RFC 1583.	–
RFC 2370	The OSPF Opaque LSA Option	–
RFC 3137	OSPF Stub Router Advertisement	This RFC is informational and non-standard.
RFC 3623	Graceful OSPF Restart	Partially supported by Huawei devices.
RFC 3630	Traffic Engineering (TE) Extensions to OSPF Version 2	–
RFC 3682	The Generalized TTL Security Mechanism (GTSM)	This RFC is experimental and non-standard.
RFC 3906	Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels	-
RFC 4576	Using a Link State Advertisement (LSA) Options Bit to Prevent Looping in BGP/MPLS IP Virtual Private Networks (VPNs)	-
RFC 4577	PDF OSPF as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)	-

 Copyright © Huawei Technologies Co., Ltd.

Document	Description	Remarks
RFC 4750	OSPF Version 2 Management Information Base	-

11.12.7 OSPFv3

11.12.7.1 Introduction

Definition

Open Shortest Path First (OSPF), developed by the Internet Engineering Task Force (IETF), is a link-state Interior Gateway Protocol (IGP).

At present, OSPF Version 2 (OSPFv2) is used for IPv4, while OSPF Version 3 (OSPFv3), developed on the basis of OSPFv2, is used for IPv6.

Purpose

The primary purpose of OSPFv3 is to develop a routing protocol independent of any specific network layer. The internal OSPFv3 router information is redesigned to achieve this purpose.

11.12.7.2 Principles

11.12.7.2.1 OSPFv3 Fundamentals

Running on IPv6, OSPFv3 (defined in RFC 2740) is an independent routing protocol that is developed on the basis of OSPFv2.

- OSPFv3 and OSPFv2 are the same in respect of the working principles of the Hello packet, state machine, link-state database (LSDB), flooding, and route calculation.
- OSPFv3 packets are encapsulated into IPv6 packets and can be transmitted in unicast or multicast mode.

OSPFv3 Packet Types

Packet Type	Function
Hello packet	Hello packets are sent periodically to discover and maintain OSPFv3 neighbor relationships.
Database Description (DD) packet	DD packets contain the summary of the local LSDB and are exchanged between two OSPFv3 routers to update the LSDBs.

 Copyright © Huawei Technologies Co., Ltd.

Packet Type	Function
Link State Request (LSR) packet	LSR packets are sent to the neighbor to request the required LSAs. An OSPFv3 device sends LSR packets to its neighbor only after they exchange DD packets.
Link State Update (LSU) packet	LSU packets carry the LSAs required by neighbors.
Link State Acknowledgment (LSAck) packet	LSAck packets acknowledge the receipt of an LSA.

LSA Types

LSA Type	Description
Router-LSA (Type 1)	Describes the link status and link cost of a router. It is generated by every router and advertised in the area to which the router belongs.
Network-LSA (Type 2)	Describes the link status of all routers on the local network segment. Network-LSAs are generated by a designated router (DR) and advertised in the area to which the DR belongs.
Inter-Area-Prefix-LSA (Type 3)	Describes routes to a specific network segment in an area. Inter-Area-Prefix-LSAs are generated on the Area Border Router (ABR) and sent to related areas.
Inter-Area-Router-LSA (Type 4)	Describes routes to an Autonomous System Boundary Router (ASBR). Inter-Area-Router-LSAs are generated by an ABR and advertised to all related areas except the area to which the ASBR belongs.
AS-external-LSA (Type 5)	Describes routes to a destination outside the AS. AS-external-LSAs are generated by an ASBR and advertised to all areas except stub areas and NSSAs.
Link-LSA (Type 8)	Describes the link-local address and IPv6 address prefix associated with the link and the link option set in the network LSA. Link LSAs are transmitted only on the link.

 Copyright © Huawei Technologies Co., Ltd.

LSA Type	Description
Intra-Area-Prefix-LSA (Type 9)	Each device or DR generates one or more intra-area prefix LSAs and transmits it in the local area. • An intra-area prefix LSA generated by a device describes the IPv6 address prefix associated with the router LSA. • An intra-area prefix LSA generated by a DR describes the IPv6 address prefix associated with the network LSA.

Router Types

Figure 11-130 Router types

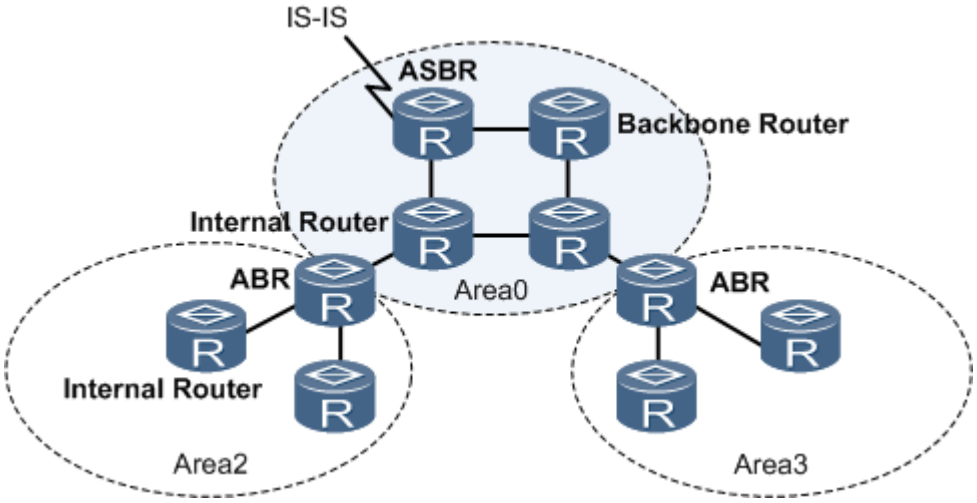


Table 11-57 Router types and descriptions

Router Type	Description
Internal router	All interfaces on an internal router belong to the same OSPFv3 area.
Area border router (ABR)	An ABR belongs to two or more areas, one of which must be the backbone area. An ABR is used to connect the backbone area and non-backbone areas. It can be physically or logically connected to the backbone area.
Backbone router	At least one interface on a backbone router belongs to the backbone area. Internal routers in Area 0 and all ABRs are backbone routers.

 Copyright © Huawei Technologies Co., Ltd.

Router Type	Description
AS boundary router (ASBR)	An ASBR exchanges routing information with other ASs. An ASBR does not necessarily reside on the border of an AS. It can be an internal router or an ABR. An OSPFv3 device that has imported external routing information will become an ASBR.

OSPFv3 Route Types

Inter-area routes and intra-area routes describe the network structure of an AS. External routes describe how to select a route to the destination outside an AS. OSPFv3 classifies the imported AS external routes into Type 1 routes and Type 2 routes.

Table 11-58 lists route types in descending order of priority.

Table 11-58 Types of OSPFv3 routes

Route Type	Description
Intra area route	Indicates routes within an area.
Inter area route	Indicates routes between areas.
Type 1 external route	Type 1 external routes have high reliability. Cost of a Type 1 external route = Cost of the route from a local router to an ASBR + Cost of the route from the ASBR to the destination of the Type 1 external route
Type 2 external route	Type 2 external routes have low reliability, and therefore OSPFv3 considers the cost of the route from an ASBR to the destination of a Type 2 external route to be much greater than the cost of any internal route to the ASBR. Cost of a Type 2 external route = Cost of the route from the ASBR to the destination of the Type 2 external route

Area

When a large number of Routers run OSPFv3, link state databases (LSDBs) become very large and require a large amount of storage space. Large LSDBs also complicate shortest path first (SPF) computation and are computationally intensive for the Routers. Network expansion causes the network topology to change, which results in route flapping and frequent OSPFv3 packet transmission. When a large number of OSPFv3 packets are transmitted on the network,

 Copyright © Huawei Technologies Co., Ltd.

bandwidth usage efficiency decreases. Each change in the network topology causes all Routers on the network to recalculate routes.

OSPFv3 resolves this problem by partitioning an AS into different areas. An area is regarded as a logical group, and each group is identified by an area ID. A Router, not a link, resides at the border of an area. A network segment or link can belong only to one area. An area must be specified for each OSPFv3 interface.

OSPFv3 areas include common areas, stub areas, and not-so-stubby areas (NSSAs), as described in **Table 11-59**.

Table 11-59 OSPFv3 areas

Area Type	Function	Notes
Common area	By default, OSPFv3 areas are defined as common areas. Common areas include: • Standard area: transmits intra-area, inter-area, and external routes. • Backbone area: connects to all other OSPFv3 areas and transmits inter-area routes. The backbone area is represented by area 0. Routes between non-backbone areas must be forwarded through the backbone area.	• The backbone area must have all its devices connected. • All non-backbone areas must remain connected to the backbone area.
Stub area	A stub area is a non-backbone area with only one ABR and generally resides at the border of an AS. The area border router (ABR) in a stub area does not transmit received AS external routes, which significantly decreases the number of entries in the routing table on the ABR and the amount of routing information to be transmitted. To ensure the reachability of AS external routes, the ABR in the stub area generates a default route and advertises the route to non-ABRs in the stub area. A totally stub area allows only intra-area routes and ABR-advertised Type 3 link state advertisements (LSAs) carrying a default route to be advertised within the area.	• The backbone area cannot be configured as a stub area. • An autonomous system boundary router (ASBR) cannot exist in a stub area. Therefore, AS external routes cannot be advertised within the stub area. • A virtual link cannot pass through a stub area.

 Copyright © Huawei Technologies Co., Ltd.

Area Type	Function	Notes
NSSA	An NSSA is similar to a stub area. An NSSA does not advertise Type 5 LSAs, but can import AS external routes. ASBRs in an NSSA generate Type 7 LSAs to carry the information about the AS external routes. The Type 7 LSAs are advertised only within the NSSA. When the Type 7 LSAs reach an ABR in the NSSA, the ABR translates the Type 7 LSAs into Type 5 LSAs and floods them to the entire AS. A totally NSSA area allows only intra-area routes to be advertised within the area.	• ABRs in an NSSA advertise Type 3 LSAs carrying a default route within the NSSA. All inter-area routes are advertised by ABRs. • A virtual link cannot pass through an NSSA.

Network Types Supported by OSPFv3

OSPFv3 classifies networks into the following types based on link layer protocols.

Table 11-60 Types of OSPFv3 networks

Network Type	Description
Broadcast	OSPFv3 considers networks with Ethernet or Fiber Distributed Data Interface (FDDI) as the link layer protocol as broadcast networks by default. On a broadcast network: • Hello packets, LSU packets, and LSAck packets are usually transmitted in multicast mode. FF02::5 is an IPv6 multicast address reserved for an OSPFv3 device. FF02::6 is an IPv6 multicast address reserved for an OSPFv3 DR or backup designated router (BDR). • DD and LSR packets are transmitted in unicast mode.
Non-broadcast Multiple Access (NBMA)	OSPFv3 considers networks with X.25 as the link layer protocol as NBMA networks by default. On an NBMA network, protocol packets, such as Hello packets, DD packets, LSR packets, LSU packets, and LSAck packets are sent in unicast mode.

 Copyright © Huawei Technologies Co., Ltd.

Network Type	Description
Point-to-Multipoint (P2MP)	No network is a P2MP network by default, no matter what type of link layer protocol is used on the network. A non-fully meshed NBMA network can be changed to a P2MP network. On a P2MP network: • Hello packets are transmitted in multicast mode using the multicast address FF02::5. • Other types of protocol packets, such as DD packets, LSR packets, LSU packets, and LSAck packets are sent in unicast mode.
Point-to-point (P2P)	OSPFv3 considers networks with PPP, HDLC, or LAPB as the link layer protocol to be P2P networks by default. On a P2P network, protocol packets, such as Hello packets, DD packets, LSR packets, LSU packets, and LSAck packets are sent in multicast mode using the multicast address FF02::5.

Stub Area

Stub areas are specific areas where ABRs do not flood received AS external routes. In stub areas, routers maintain fewer routing entries and less routing information than those in other areas.

Configuring a stub area is optional. Not every area can be configured as a stub area, because a stub area is usually a non-backbone area with only one ABR and is located at the AS border.

To ensure the reachability of the routes to destinations outside an AS, the ABR in the stub area generates a default route and advertises the route to the non-ABRs in the same stub area.

Note the following points when configuring a stub area:

- The backbone area cannot be configured as a stub area.
- Configure stub area attributes on all routers in the area to be configured as a stub area.
- No ASBRs are allowed in the area to be configured as a stub area because AS external routes cannot be transmitted in the stub area.

OSPFv3 Multi-process

OSPFv3 supports multi-process. Multiple OSPFv3 processes can independently run on the same router. Route exchange between different OSPFv3 processes is similar to that between different routing protocols.

11.12.7.2.2 Comparison Between OSPFv3 and OSPFv2

OSPFv3 and OSPFv2 are the same in the following aspects:

 Copyright © Huawei Technologies Co., Ltd.

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

OSPFv3 and OSPFv2 are different in the following aspects:

- In OSPFv3, only LSUs contain IP addresses.
- OSPFv3 runs on IPv6 which is based on links rather than network segments.

  Therefore, the interfaces on which OSPFv3 is to be configured should be on the same link rather than in the same network segment. In addition, the interfaces can establish OSPFv3 sessions without IPv6 global addresses.
- OSPFv3 does not depend on IP addresses.

  OSPFv3 separates topology calculation from IP addresses. Specifically, OSPFv3 can calculate the OSPFv3 topology without IPv6 global addresses which only apply to virtual link interfaces and packet forwarding.
- OSPFv3 packets and the LSA format change.
  - OSPFv3 packets do not contain IP addresses.
  - OSPFv3 router LSAs and network LSAs do not contain IP addresses, which are advertised through link LSAs and intra-area prefix LSAs.
  - In OSPFv3, Router IDs, area IDs, and LSA link state IDs no longer indicate IP addresses, but the IPv4 address format is still reserved.
  - Neighbors are identified by Router IDs instead of IP addresses on broadcast, NBMA, or P2MP networks.
- Information about the flooding scope is added to OSPFv3 LSAs.

  Information about the flooding scope is added to the LSA Type field of OSPFv3 LSAs. Therefore, OSPFv3 routers can process LSAs of unidentified types more flexibly.
  - OSPFv3 can store or flood unidentified packets, while OSPFv2 discards unidentified packets.
  - In OSPFv3, unknown LSAs with 1 as the U flag bit can be flooded, and the flooding scope of such LSAs is specified by the LSAs.

  For example, Router A and Router B can identify LSAs of a certain type. Router A and Router B are connected through Router C which, however, cannot identify these LSAs. If Router A floods such LSA to Router C, Router C can still flood the received LSAs to Router B although Router C does not identify these LSAs. Router B then processes these LSAs.

  If OSPFv2 is run, Router C discards the unidentified LSAs. As a result, these LSAs cannot reach Router B.
- OSPFv3 supports multi-process on a link.

  In OSPFv2, one physical interface can be bound to only one multi-instance. In OSPFv3, one physical interface can be bound to multiple multi-instances that are identified by different instance IDs. In these OSPFv3 multi-instances running on one physical interface, neighbor relationships are established separately, sharing resources on the same link.

 Copyright © Huawei Technologies Co., Ltd.

- OSPFv3 uses IPv6 link-local addresses.

  IPv6 implements neighbor discovery and automatic configuration based on link-local addresses. Routers running IPv6 do not forward IPv6 packets whose destination address is a link-local address, and those packets can only be exchanged on the same link. The unicast link-local address starts from FE80/10.

  As a routing protocol running on IPv6, OSPFv3 also uses link-local addresses to maintain neighbor relationships and update LSDBs. Except Vlink interfaces, all OSPFv3 interfaces use link-local addresses as the source address and the next hop to transmit OSPFv3 packets.

  The advantages are as follows:

  – OSPFv3 can calculate the topology without global IPv6 addresses.

  – The packets flooded on a link are not transmitted to other links, which prevents unnecessary flooding and saves bandwidth.

- OSPFv3 supports two new LSAs.

  – Link LSA: A device floods a link LSA on the link where it resides to advertise its link-local address and the configured global IPv6 address.

  – Intra-area prefix LSA: A device advertises an intra-area prefix LSA in the local OSPF area to inform the other routers in the area or the network (either a broadcast network or an NBMA network) of its IPv6 global address.

- OSPFv3 identifies neighbors based on Router IDs only.

  On broadcast, NBMA, and P2MP networks, OSPFv2 identifies neighbors based on IPv4 addresses of interfaces.

  OSPFv3 identifies neighbors based on Router IDs only.

11.12.7.2.3 OSPFv3 GR

NOTE

The can be configured as a GR helper rather than a GR restarter.

Graceful restart (GR) is a technology used to ensure proper traffic forwarding, especially the forwarding of key services, during the restart of routing protocols.

Without GR, the master/slave main control board switchover due to various reasons leads to transient service interruption, and as a result, route flapping occurs on the whole network. Such route flapping and service interruption are unacceptable on large-scale networks, especially carrier networks.

GR is one of the high availability (HA) technologies which comprise a series of comprehensive technologies, such as fault-tolerant redundancy, link protection, faulty node recovery, and traffic engineering technologies. As a fault-tolerant redundancy technology, GR is widely used to ensure non-stop forwarding of key data during the master/slave main control board switchovers and system upgrade.

In GR mode, the forwarding plane continues data forwarding during a restart, and operations on the control plane, such as re-establishment of neighbor relationships and route calculation, do not affect the forwarding plane, preventing service interruptions caused by route flapping and improving network reliability.

 Copyright © Huawei Technologies Co., Ltd.

Comparison Between Master/Slave Main Control Board Switchovers with and Without GR

Table 11-61 Comparison between master/slave main control board switchovers with and without GR

Master/Slave Main Control Board Switchovers Without GR	Master/Slave Main Control Board Switchovers with GR
• OSPFv3 neighbor relationships are reestablished. • Routes are recalculated. • The forwarding table changes. • The entire network detects route changes, and route flapping occurs for a short period of time. • Packets are lost during forwarding, and services are interrupted.	• OSPFv3 neighbor relationships are reestablished. • Routes are recalculated. • The forwarding table remains unchanged. • Except the neighbors of the router on which a master/slave main control board switchover occurs, other routers do not detect route changes. • No packets are lost during forwarding, and services are not affected.

11.12.7.2.4 OSPFv3 Authentication

In Open Shortest Path First version 3 (OSPFv3) authentication, an authentication field is added to each OSPFv3 packet for encryption. When a local device receives an OSPFv3 packet from a remote device, the local device discards the packet if the authentication password carried in the packet is different from the local one, which protects the local device against potential attacks. Therefore, OSPFv3 authentication improves network security.

Based on the applicable scope, OSPFv3 authentication is classified as follows:

- Area authentication

  This authentication is configured in the OSPFv3 area view and applies to packets received by all interfaces in an OSPF area.
- Process authentication

  This authentication is configured in the OSPFv3 view and applies to all packets in an OSPF process.
- Interface authentication

  This authentication is configured in the interface view and applies to all packets received by the interface.

OSPFv3 uses HMAC-SHA256 authentication to authenticate packets. In HMAC-SHA256 authentication, a password is encrypted using the HMAC-SHA256 algorithm before being added to a packet, which improves password security.

Each OSPFv3 packet carries an authentication type in the header and authentication information in the tail.

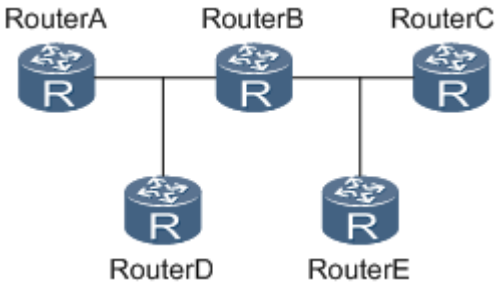
The authentication types are as follows:

 Copyright © Huawei Technologies Co., Ltd.

- 1: simple authentication
- 2: ciphertext authentication

Networking Application

Figure 11-131 OSPFv3 authentication on a broadcast network



The configuration requirements are as follows:

- Interface authentication configurations must be the same on all devices of the same network so that OSPFv3 neighbor relationships can be established.
- Area authentication configurations must be the same on all devices in the same area.

11.12.7.2.5 OSPFv3 Packet Format

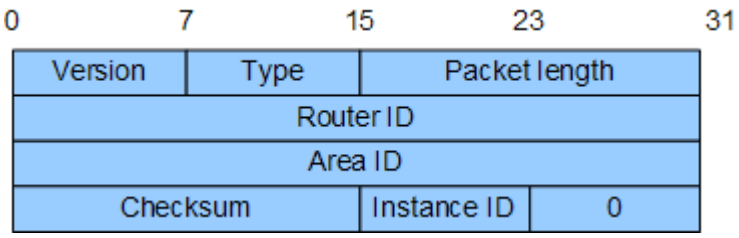
Open Shortest Path First (OSPF) for IPv6 packets are encapsulated into IPv6 packets. The OSPFv3 protocol number is 89. OSPFv3 packets are classified into the following types:

- **Hello packet**
- **Database Description (DD) packet**
- **Link State Request (LSR) packet**
- **Link State Update (LSU) packet**
- **Link State Acknowledgment (LSAck) packet**

Packet Header Format

The five types of OSPFv3 packets have the same packet header format. The length of an OSPFv3 packet header is 24 bytes. **Figure 11-132** shows an OSPFv3 packet header.

Figure 11-132 OSPFv3 packet header



 Copyright © Huawei Technologies Co., Ltd.

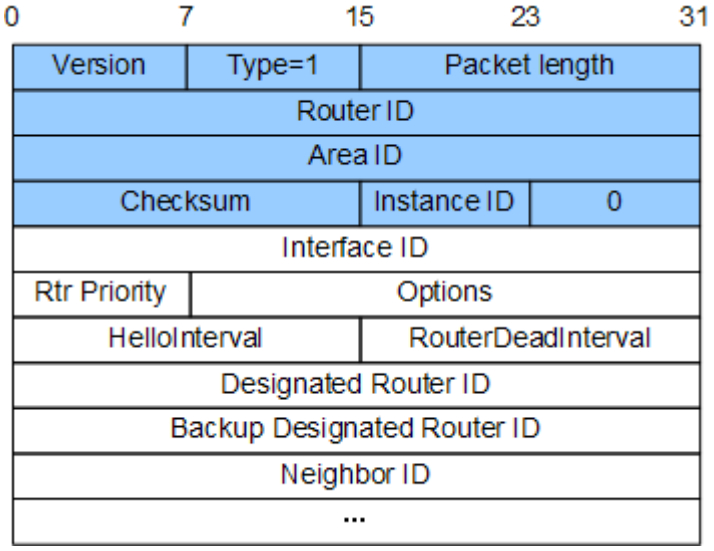
Table 11-62 Packet header fields

Field	Length	Description
Version	8 bits	OSPF version number. For OSPFv3, the value is 3.
Type	8 bits	OSPFv3 packet type. The values are as follows: • 1: Hello packet • 2: DD packet • 3: LSR packet • 4: LSU packet • 5: LSAck packet
Packet length	16 bits	Length of the OSPFv3 packet containing the packet header, in bytes.
Router ID	32 bits	ID of the Router that sends the OSPFv3 packet.
Area ID	32 bits	ID of the area to which the Router that sends the OSPFv3 packet belongs.
Checksum	16 bits	Checksum of the OSPFv3 packet that does not contain the Authentication field.
Instance ID	8 bits	ID of an OSPFv3 instance.
0	8 bits	Reserved fields.

Hello Packet

Hello packets are commonly used packets, which are periodically sent on OSPFv3 interfaces to establish and maintain neighbor relationships. A Hello packet includes information about the designated router (DR), backup designated router (BDR), timers, and known neighbors. **Figure 11-133** shows the format of a Hello packet.

Figure 11-133 Format of a Hello packet



 Copyright © Huawei Technologies Co., Ltd.

Table 11-63 Hello packet fields

Field	Length	Description
Interface ID	32 bits	ID of the interface that sends the Hello packets.
Rtr Priority	8 bits	DR priority. The default value is 1. **NOTE** If the DR priority of a Router interface is set to 0, the interface cannot participate in a DR or BDR election.
Options	24 bits	The values are as follows: • E: Type 5 link state advertisements (LSAs) are flooded. • MC: IP multicast packets are forwarded. • N/P: Type 7 LSAs are processed. • DC: On-demand links are processed.
HelloInterval	16 bits	Interval at which Hello packets are sent.
RouterDeadInterval	16 bits	Dead interval. If a Router does not receive any Hello packets from its neighbors within a specified dead interval, the neighbors are considered Down.
Designated Router ID	32 bits	Interface address of the DR.
Backup Designated Router ID	32 bits	Interface address of the BDR.
Neighbor ID	32 bits	Router ID of the neighbor.

Table 11-64 lists the address types, interval types, and default intervals used when Hello packets are transmitted on different networks.

Table 11-64 Hello packet characteristics for various network types

Network Type	Address Type	Interval Type	Default Interval
Broadcast	Multicast address	HelloInterval	10 seconds

 Copyright © Huawei Technologies Co., Ltd.

Network Type	Address Type	Interval Type	Default Interval
Non-broadcast multiple access (NBMA)	Unicast address	• HelloInterval for the DR, BDR, and Router that can become a DR • PollInterval for the case when neighbors become Down and HelloInterval for other cases	30 seconds for HelloInterval 120 seconds for PollInterval
Point-to-point (P2P)	Multicast address	HelloInterval	10 seconds
Point-to-multipoint (P2MP)	Unicast address	HelloInterval	30 seconds

NOTE

To establish neighbor relationships between Routers on the same network segment, you must set the same HelloInterval, PollInterval, and RouterDeadInterval values for the Routers. PollInterval applies only to NBMA networks.

DD Packet

During an adjacency initialization, two Routers use DD packets to describe their own link state databases (LSDBs) for LSDB synchronization. A DD packet contains the header of each LSA in an LSDB. An LSA header uniquely identifies an LSA. The LSA header occupies only a small portion of the LSA, which reduces the amount of traffic transmitted between Routers. A neighbor can use the LSA header to check whether it already has the LSA. When two Routers exchange DD packets, one functions as the master and the other functions as the slave. The master defines a start sequence number. The master increases the sequence number by one each time it sends a DD packet. After the slave receives a DD packet, it uses the sequence number carried in the DD packet for acknowledgement.

Figure 11-134 shows the format of a DD packet.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-134 Format of a DD packet

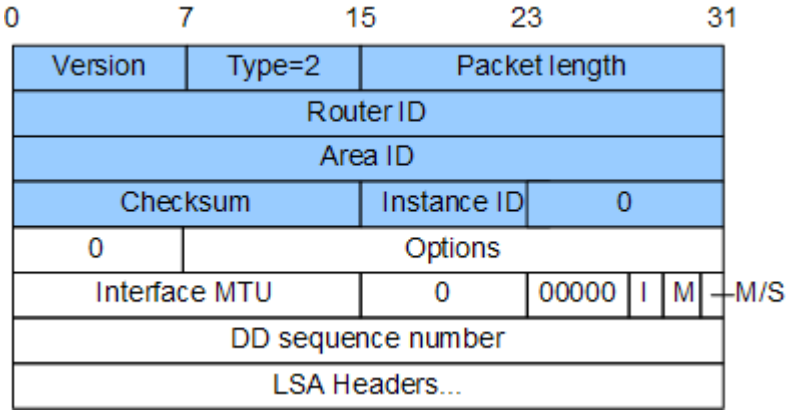


Table 11-65 DD packet fields

Field	Length	Description
Options	24 bits	The values are as follows: • E: Type 5 LSAs are flooded. • MC: IP multicast packets are forwarded. • N/P: Type 7 LSAs are processed. • DC: On-demand links are processed.
Interface MTU	16 bits	Maximum length of the DD packet sent by the interface with packet fragmentation disabled.
I	1 bit	If the DD packet is the first packet among multiple consecutive DD packets sent by a Router, this field is set to 1. In other cases, this field is set to 0.
M (More)	1 bit	If the DD packet is the last packet among multiple consecutive DD packets sent by a Router, this field is set to 0. In other cases, this field is set to 1.
M/S (Master/ Slave)	1 bit	When two Routers exchange DD packets, they negotiate a master/slave relationship. The Router with a larger router ID becomes the master. If this field is set to 1, the DD packet is sent by the master.
DD sequence number	32 bits	Sequence number of the DD packet. The master and slave use the sequence number to ensure that DD packets are correctly transmitted.
LSA Headers	-	LSA header information included in the DD packet.

LSR Packet

After two Routers exchange DD packets, they send LSR packets to request each other's LSAs. The LSR packets contain the summaries of the requested LSAs. **Figure 11-135** shows the format of an LSR packet.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-135 Format of an LSR packet

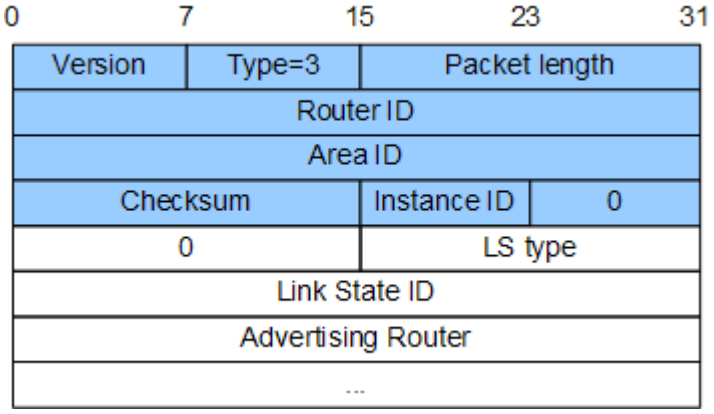


Table 11-66 LSR packet fields

Field	Length	Description
LS type	16 bits	Type of the LSA
Link State ID	32 bits	This field together with the LS type field describes an LSA in an AS.
Advertising Router	32 bits	Router ID of the Router that generates the LSA.

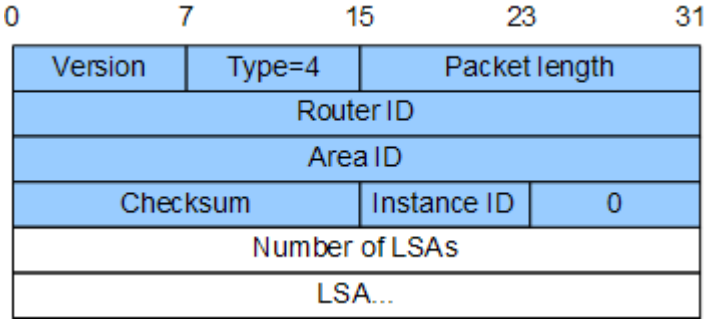
NOTE

The LS type, Link State ID, and Advertising Router fields can uniquely identify an LSA. If two LSAs have the same LS type, Link State ID, and Advertising Router fields, a Router uses the LS sequence number, LS checksum, and LS age fields to obtain a required LSA.

LSU Packet

A Router uses an LSU packet to transmit LSAs requested by its neighbors or to flood its own updated LSAs. The LSU packet contains a set of LSAs. For multicast and broadcast networks, LSU packets are multicast to flood LSAs. To ensure reliable LSA flooding, a Router uses an LSAck packet to acknowledge the LSAs contained in an LSU packet that is received from a neighbor. If an LSA fails to be acknowledged, the Router retransmits the LSA to the neighbor. **Figure 11-136** shows the format of an LSU packet.

Figure 11-136 Format of an LSU packet



 Copyright © Huawei Technologies Co., Ltd.

Table 11-67 LSU packet field

Field	Length	Description
Number of LSAs	32 bits	Number of LSAs contained in the LSU packet

LSAck Packet

A Router uses an LSAck packet to acknowledge the LSAs contained in a received LSU packet. The LSAs can be acknowledged using LSA headers. LSAck packets can be transmitted over different links in unicast or multicast mode. **Figure 11-137** shows the format of an LSAck packet.

Figure 11-137 Format of an LSAck packet

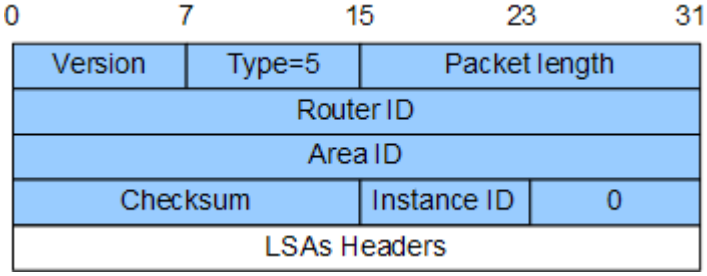


Table 11-68 LSAck packet field

Field	Length	Description
LSAs Headers	Determined by the header length of the LSA to be acknowledged.	This field is used to acknowledge an LSA.

11.12.7.2.6 OSPFv3 LSA Format

Each Router in an autonomous system (AS) generates one or more types of link state advertisements (LSAs), depending on the Router's type. Multiple LSAs form a link state database (LSDB). Open Shortest Path First (OSPF) for IPv6 encapsulates routing information into LSAs for transmission. Commonly used LSAs include:

- **Router-LSA (Type1)**
- **Network-LSA (Type2)**
- **Inter-Area-Prefix-LSA (Type3)**
- **Inter-Area-Router-LSA (Type4)**
- **AS-external-LSA (Type5)**
- **NSSA LSA (Type7)**

 Copyright © Huawei Technologies Co., Ltd.

- **Link-LSA (Type8)**
- **Intra-Area-Prefix-LSA (Type9)**

LSA Header Format

All LSAs have the same header. **Figure 11-138** shows an LSA header.

Figure 11-138 LSA header

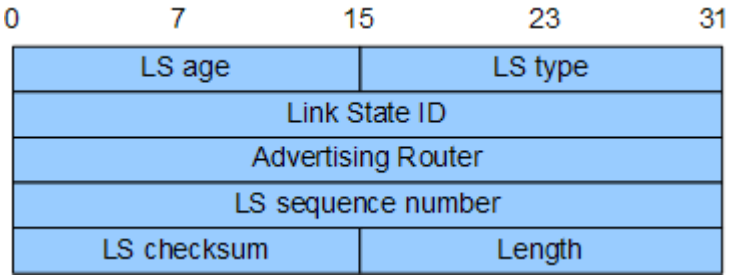


Table 11-69 LSA header fields

Field	Length	Description
LS age	16 bits	Time that elapses after the LSA is generated, in seconds. The value of this field continually increases regardless of whether the LSA is transmitted over a link or saved in an LSDB.
LS type	16 bits	Type of the LSA. The values are as follows: • Type1: Router-LSA. • Type2: Network-LSA. • Type3: Inter-Area-Prefix-LSA. • Type4: Inter-Area-Router-LSA. • Type5: AS-external-LSA. • Type7: NSSA-LSA. • Type8: Link-LSA. • Type9: Intra-Area-Prefix-LSA.
Link State ID	32 bits	This field together with the LS type field describes an LSA in an AS.
Advertising Router	32 bits	Router ID of the Router that generates the LSA.
LS sequence number	32 bits	Sequence number of the LSA. Neighbors can use this field to identify the latest LSA.
LS checksum	16 bits	Checksum of all fields except the LS age field.
Length	16 bits	Length of the LSA including the LSA header, in bytes.

 Copyright © Huawei Technologies Co., Ltd.

Router-LSA

A router-LSA describes the link status and cost of a Router. Router-LSAs are generated by a Router and advertised within the area to which the Router belongs. **Figure 11-139** shows the format of a router-LSA.

Figure 11-139 Format of a router-LSA

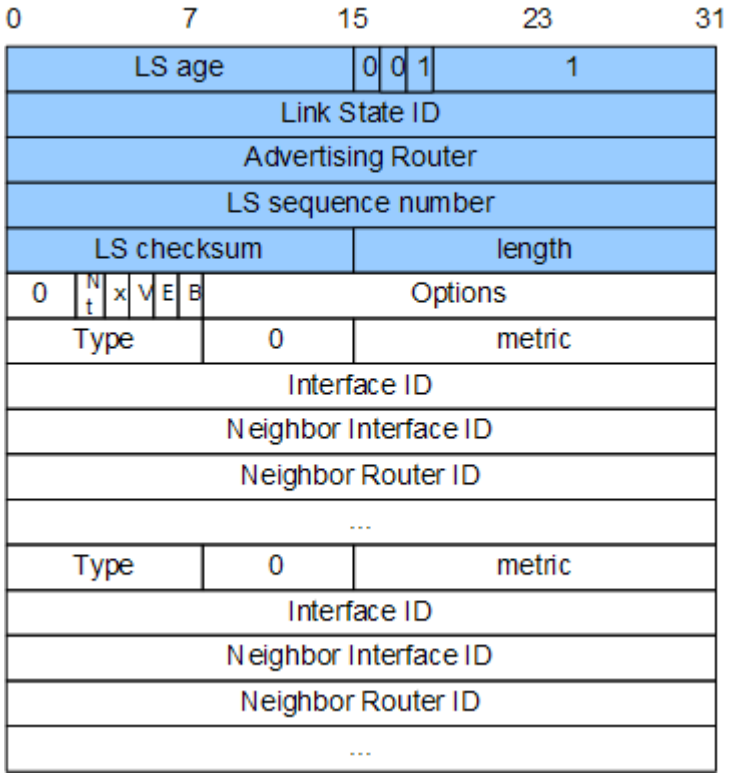


Table 11-70 Router-LSA fields

Field	Length	Description
Nt (NSSA translation)	1 bit	If the Router that generates the LSA is an NSSA border router, this field is set to 1. In other cases, this field is set to 0. When this field is set to 1, the router is unconditionally translating NSSA-LSAs into AS-external-LSAs.
x	1 bit	This field is deprecated.
V (Virtual Link)	1 bit	If the Router that generates the LSA is located at one end of a virtual link, this field is set to 1. In other cases, this field is set to 0.
E (External)	1 bit	If the Router that generates the LSA is an autonomous system boundary router (ASBR), this field is set to 1. In other cases, this field is set to 0.

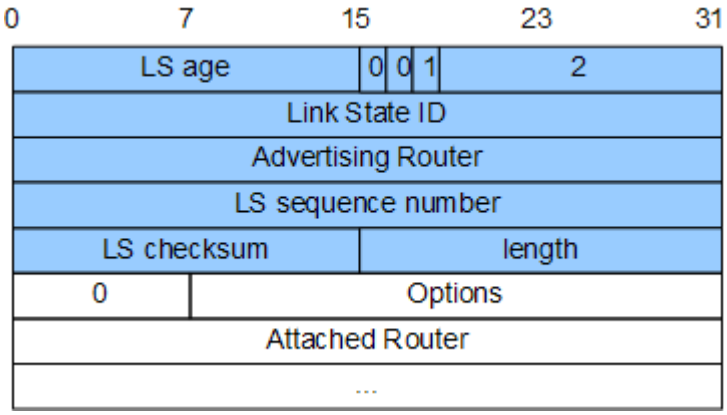
Copyright © Huawei Technologies Co., Ltd.

Field	Length	Description
B (Border)	1 bit	If the Router that generates the LSA is an area border router (ABR), this field is set to 1. In other cases, this field is set to 0.
Options	24 bits	The optional capabilities supported by the Router.
Type	8 bits	Type of the Router link. The values are as follows: • 1: The Router is connected to another Router in point-to-point (P2P) mode. • 2: The Router is connected to a transport network. • 3: Reserved. • 4: The Router is connected to another Router over a virtual link.
metric	16 bits	Cost of the link.
Interface ID	32 bits	The Interface ID assigned to the interface.
Neighbor Interface ID	32 bits	The Interface ID the neighbor Router has associated with the link.
Neighbor Router ID	32 bits	The Router ID of the neighbor Router.

Network-LSA

A network-LSA describes the link status of all Routers on the local network segment. Network-LSAs are generated by a DR on a broadcast or non-broadcast multiple access (NBMA) network and advertised within the area to which the DR belongs. **Figure 11-140** shows the format of a network-LSA.

Figure 11-140 Format of a network-LSA



 Copyright © Huawei Technologies Co., Ltd.

Table 11-71 Network-LSA fields

Field	Length	Description
Options	24 bits	The optional capabilities supported by the Router.
Attached Router	32 bits	Router IDs of all Routers on the broadcast or NBMA network, including the router ID of the DR

Inter-Area-Prefix-LSA

An inter-area-prefix-LSA describes routes on a network segment in an area. It is generated by the ABR. The routes are advertised to other areas.

Figure 11-141 shows the format of an inter-area-prefix-LSA.

Figure 11-141 Format of An inter-area-prefix-LSA

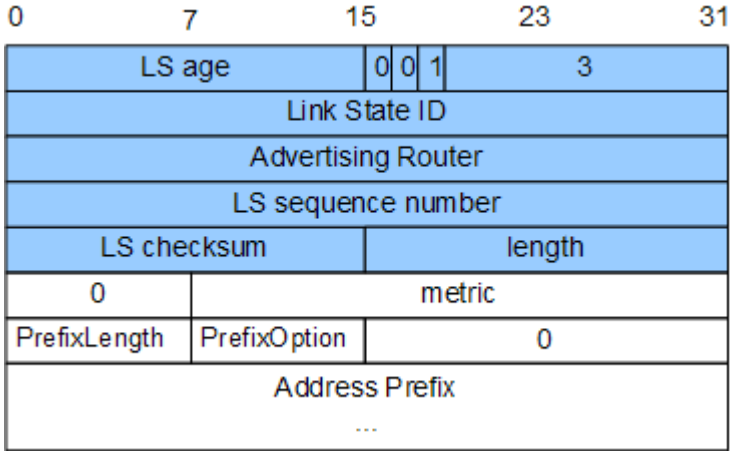


Table 11-72 Network-summary-LSA fields

Field	Length	Description
PrefixLength	8 bits	Length of the prefix.
PrefixOption	8 bits	The capabilities associated with the prefix.
Address Prefix	32 bits	An encoding of the prefix itself.

Inter-Area-Router-LSA

An inter-area-router-LSA describes routes to ASBR in other areas. It is generated by the ABR. The routes are advertised to all related areas except the area that the ASBR belongs to.

Figure 11-142 shows the format of an inter-area-prefix-LSA.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-142 Format of an inter-area-router-LSA

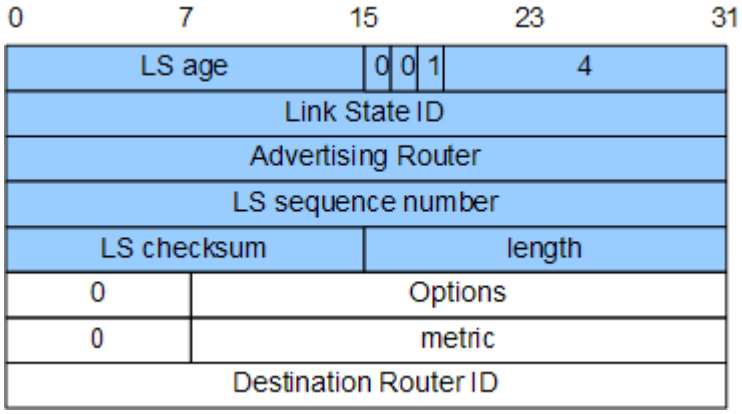


Table 11-73 Inter-area-router-LSA fields

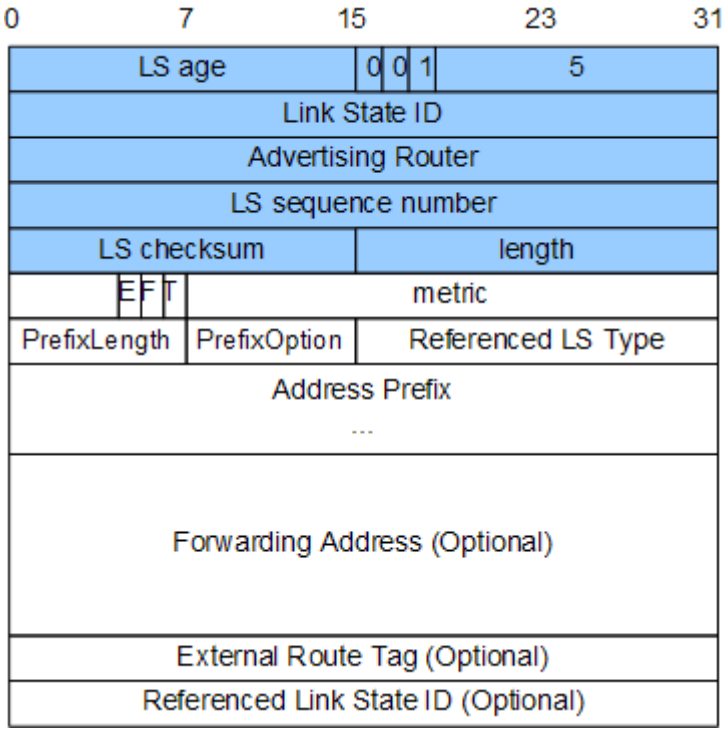
Field	Length	Description
Destination Router ID	32 bits	The Router ID of the router being described by the LSA.

AS-External-LSAs

An as-external-LSA describes destinations outside the AS, it is originated by ASBR.

Figure 11-143 shows the format of an as-external-LSA.

Figure 11-143 Format of an as-external-LSA



 Copyright © Huawei Technologies Co., Ltd.

Table 11-74 As-external-LSA fields

Field	Length	Description
E	1 bit	The type of external metric. • If this field is 1, the specified metric is a Type 2 external metric. • If this field is 0, the specified metric is a Type 1 external metric.
F	1 bit	Whether a Forwarding Address has been included in the LSA. • If this field is 1, a Forwarding Address has been included in the LSA. • If this field is 0, no Forwarding Address is included in the LSA.
T	1 bit	Whether an External Route Tag has been included in the LSA. • If this field is 1, an External Route Tag has been included in the LSA. • If this field is 0, no External Route Tag is included in the LSA.
Referenced LS Type	16 bits	Indicates the referenced LS type. If non-zero, an LSA with this LS type is to be associated with this LSA (see Referenced Link State ID below).
Forwarding Address	128 bits	A fully qualified global IPv6 address.
External Route Tag	32 bits	Indicates the external route tag, which can be used to communicate additional information between ASBRs.
Referenced Link State ID	32 bits	Indicates the referenced link state ID.

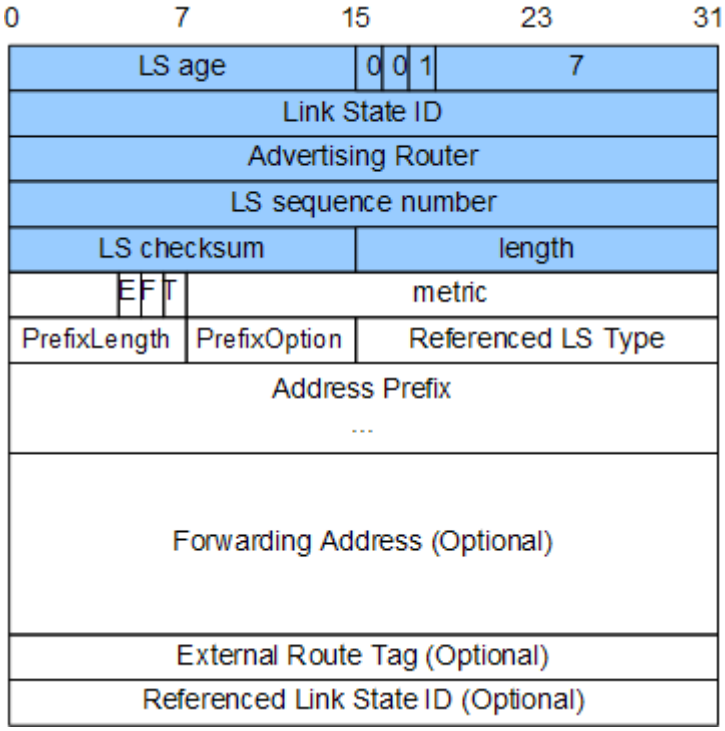
NSSA-LSAs

An NSSA-LSA describes destinations outside the AS, it is originated by ASBR.

Figure 11-144 shows the format of an NSSA-LSA.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-144 Format of an NSSA-LSA



Link-LSAs

Each Router generates a link LSA for each link. A link LSA describes the link-local address and IPv6 address prefix associated with the link and the link option set in the network LSA. It is transmitted only on the link.

Figure 11-145 shows the format of a Link-LSA.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-145 Format of a Link-LSA

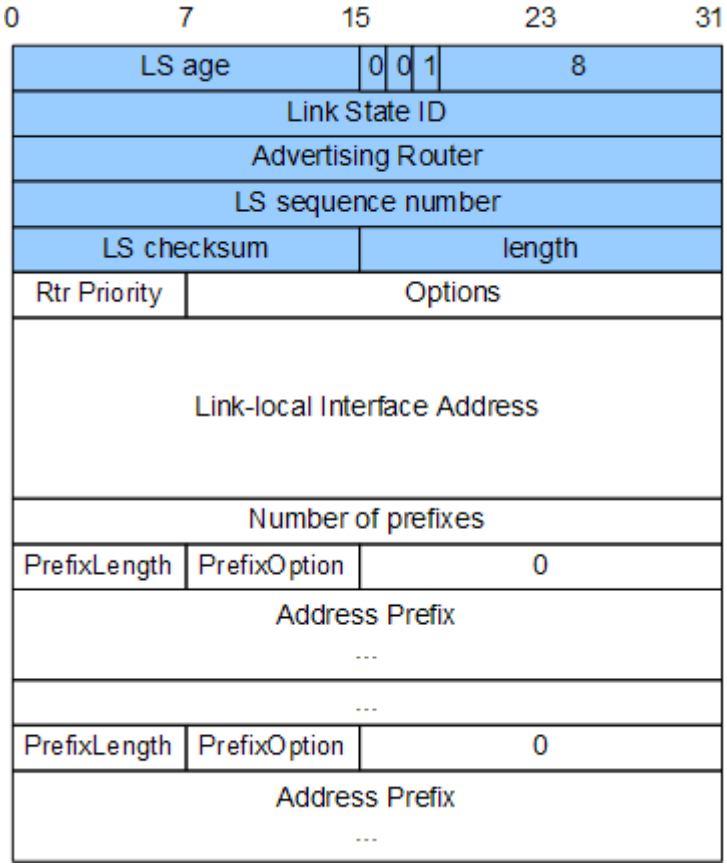


Table 11-75 Link-LSA fields

Field	Length	Description
Rtr Priority	8 bits	The Router Priority of the interface.
Options	24 bits	The set of Options bits that the Router would like set in the network-LSA that will be originated by the DR on broadcast or NBMA links.
Link-local Interface Address	128 bits	The originating Router's link-local interface address on the link.
Number of prefixes	32 bits	The number of IPv6 address prefixes contained in the LSA.

Intra-Area-Prefix-LSAs

Each Router or DR generates one or more intra-area prefix LSAs and transmits it in the local area.

- An LSA generated on a router describes the IPv6 address prefix associated with the router LSA.

 Copyright © Huawei Technologies Co., Ltd.

- An LSA generated on a DR describes the IPv6 address prefix associated with the network LSA.

Figure 11-146 shows the format of an intra-area-prefix-LSA.

Figure 11-146 Format of an intra-area-prefix-LSA

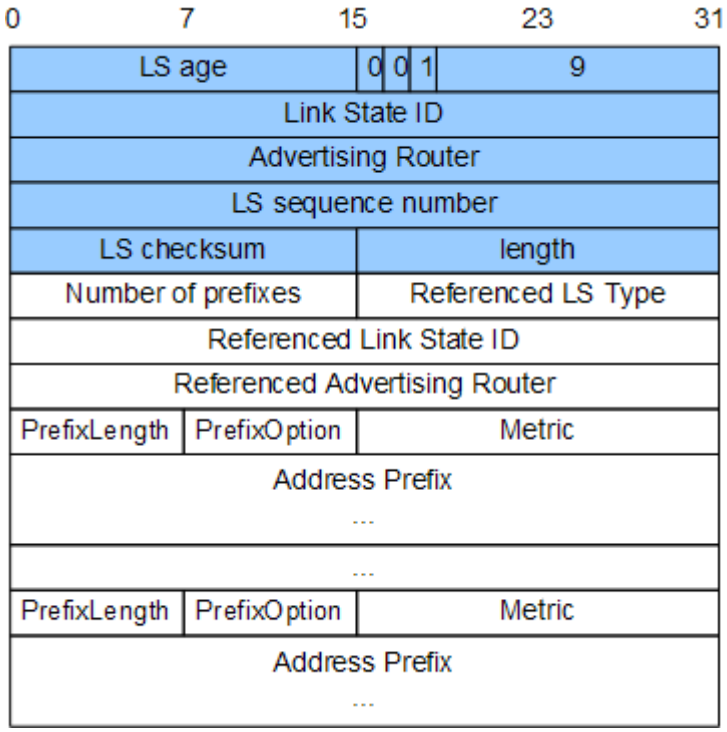


Table 11-76 Intra-area-prefix-LSA fields

Field	Length	Description
Referenced LS Type	16 bits	This field identifies the router-LSA or network-LSA with which the IPv6 address prefixes should be associated. • If Referenced LS Type is 0x2001, the prefixes are associated with a router-LSA. • If Referenced LS Type is 0x2002, the prefixes are associated with a network-LSA.
Referenced Link State ID	32 bits	Indicates the referenced link state ID. • If Referenced LS Type is 0x2001, Referenced Link State ID should be 0. • If Referenced LS Type is 0x2002, Referenced Link State ID should be the Interface ID of the link's DR.

 Copyright © Huawei Technologies Co., Ltd.

Field	Length	Description
Referenced Advertising Router	32 bits	Indicates the ID of the referenced advertising Router. • If Referenced LS Type is 0x2001, Referenced Advertising Router should be the originating router's Router ID. • If Referenced LS Type is 0x2002, Referenced Advertising Router should be the Designated Router's Router ID.

11.12.7.2.7 OSPFv3 Auto FRR

OSPFv3 auto fast reroute (FRR) is a kind of dynamic IP FRR. To implement this feature, Interior Gateway Protocol (IGP) calculates the backup paths beforehand by using the link status database of the entire network and then stores the backup paths in the forwarding table to minimize service interruption duration in case of a link fault.

Background

Some services, such as voice over IP (VoIP) and online video services, are delay-sensitive. Nevertheless, if a primary link fails, OSPFv3 must perform multiple operations, including detecting the fault, updating the link-state advertisement (LSA), flooding the LSA, calculating routes, and delivering forward information base (FIB) entries before switching traffic to a new link. This process takes a much longer time than the delay-sensitive services can tolerate. To solve this problem, OSPFv3 Auto FRR is developed.

Related Concepts

OSPFv3 Auto FRR

OSPFv3 Auto FRR refers to a mechanism in which a device uses the loop-free alternate (LFA) algorithm to compute a backup link and stores it with the primary link in the forwarding table. If the primary link fails, the device switches the traffic to the backup link before routes are converged on the control plane to minimize the impact of the link failure.

NOTE

If a Bidirectional Forwarding Detection (BFD) session is configured, it automatically associates with OSPFv3 auto FRR. If a BFD session detects a link failure, the BFD session goes down and triggers FRR.

OSPFv3 Auto FRR Policy

An OSPFv3 Auto FRR policy can be configured to filter backup routes. Only the backup routes that match the filtering rules of the policy can be added to the IP routing table.

LFA algorithm

A device uses each neighbor that can be used as a backup next hop as a root node, uses the shortest path first (SPF) algorithm to calculate the shortest path

 Copyright © Huawei Technologies Co., Ltd.

from each root node to the destination, and then uses the inequalities defined in RFC 5286 to calculate the backup link that is loop-free and has the lowest cost.

Networking

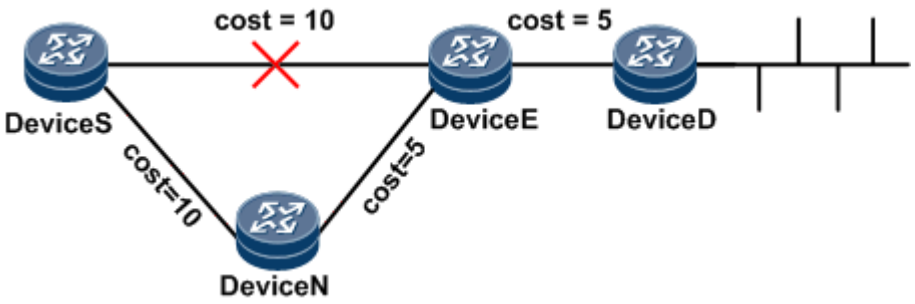
OSPFv3 auto FRR protects traffic from a fault on a specified link.

In **Figure 11-147**, traffic flows from Device S to Device D. The primary link is Device S->Device E->Device D, and the backup link is Device S->Device N->Device E->Device D. The link costs meet the inequality: Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D). With OSPFv3 Auto FRR, Device S switches the traffic to the backup link if the primary link fails, minimizing the traffic interruption duration.

NOTE

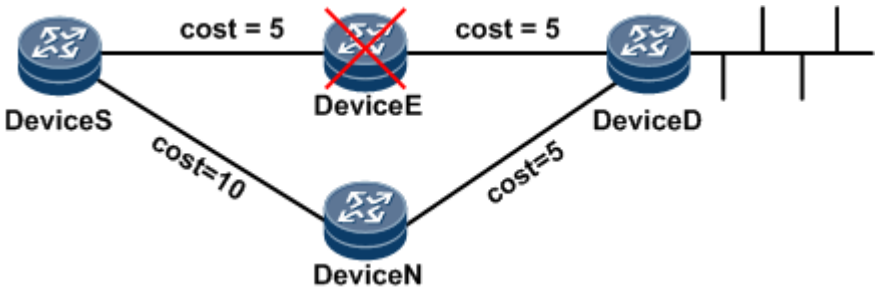
Distance_opt(X, Y) indicates the shortest path from X to Y. S stands for the source node, N for a node along the backup link, and D for the destination node.

Figure 11-147 OSPFv3 Auto FRR link protection



The switching process in the following figure is similar. Therefore, the details are omitted.

Figure 11-148 OSPFv3 Auto FRR node-and-link protection



11.12.7.3 Configuration Example of OSPFv3

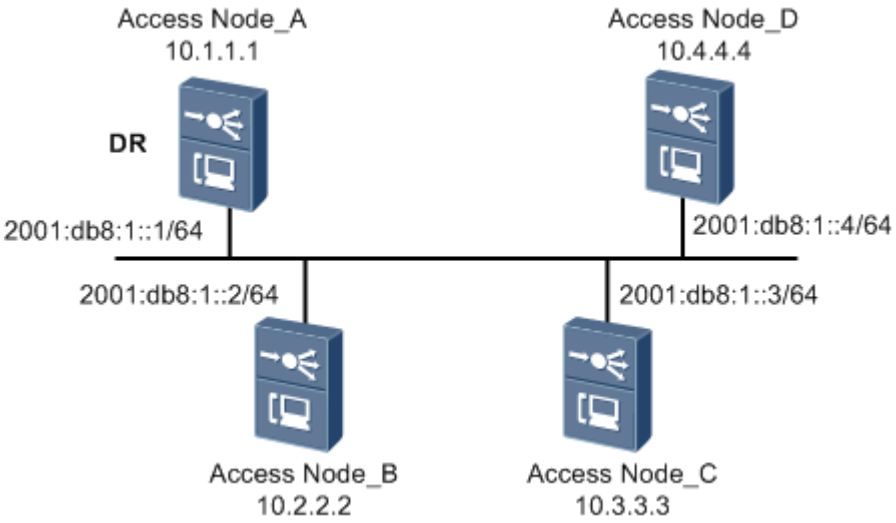
This topic provides an example for configuring OSPFv3 on the MA5800.

 Copyright © Huawei Technologies Co., Ltd.

Service Requirements

- OSPFv3 is enabled on the four MA5800s.
- Access Node_A is configured with the highest designated router (DR) priority, Access Node_C is configured with the second highest DR priority, and Access Node_A implements the broadcast of network link status for the DR.

Figure 11-149 Example network for configuring OSPFv3



Data Plan

Table 11-77 provides the data plan for configuring OSPFv3.

Table 11-77 Data plan for configuring OSPFv3

Item	Data	Remarks
Access Node_A	IPv6 address of the Layer 3 interface: 2001:db8:1::1/64	-
	Priority: 100	Ensure one of the access nodes is configured with the highest designated router (DR) priority. Take Access Node_A for example.
	VLAN ID: 2	-
	Router ID: 10.1.1.1	-
Access Node_B	IPv6 address of the Layer 3 interface: 2001:db8:1::2/64	-
	Priority: 80	-

 Copyright © Huawei Technologies Co., Ltd.

Item	Data	Remarks
	VLAN ID: 2	-
	Router ID: 10.2.2.2	-
Access Node_C	IPv6 address of the Layer 3 interface: 2001:db8:1::3/64	-
	Priority: 90	-
	VLAN ID: 2	-
	Router ID: 10.3.3.3	-
Access Node_D	IPv6 address of the Layer 3 interface: 2001:db8:1::4/64	-
	Priority: not configured	Default: 1
	VLAN ID: 2	-
	Router ID: 10.4.4.4	-

Prerequisite

The OSPFv3 area IDs of the MA5800 devices must be consistent. Take area 0 for example.

Procedure

Step 1 Configure Access Node_A.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#vlan 2
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1::1 64
huawei(config-if-vlanif2)#quit
```

2. Enable OSPFv3, and configure the OSPFv3 Router ID.

```
huawei(config)#ospfv3
huawei(config-ospfv3-1)#router-id 10.1.1.1
huawei(config-ospfv3-1)#quit
```

3. Enable OSPFv3 on the vlanif interface, and configure the OSPFv3 priority.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ospfv3 1 area 0
huawei(config-if-vlanif2)#ospfv3 dr-priority 100
huawei(config-if-vlanif2)#quit
```

4. Save the data.

```
huawei(config)#save
```

Step 2 Configure Access Node_B.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#vlan 2
huawei(config)#port vlan 2 0/9 0
```

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1::2 64
huawei(config-if-vlanif2)#quit
```

2. Enable OSPFv3, and configure the OSPFv3 Router ID.

```
huawei(config)#ospfv3
huawei(config-ospfv3-1)#router-id 10.2.2.2
huawei(config-ospfv3-1)#quit
```

3. Enable OSPFv3 on the vlanif interface, and configure the OSPFv3 priority.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ospfv3 1 area 0
huawei(config-if-vlanif2)#ospfv3 dr-priority 80
huawei(config-if-vlanif2)#quit
```

4. Save the data.

```
huawei(config)#save
```

Step 3 Configure Access Node_C.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#vlan 2
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1::3 64
huawei(config-if-vlanif2)#quit
```

2. Enable OSPFv3, and configure the OSPFv3 Router ID.

```
huawei(config)#ospfv3
huawei(config-ospfv3-1)#router-id 10.3.3.3
huawei(config-ospfv3-1)#quit
```

3. Enable OSPFv3 on the vlanif interface, and configure the OSPFv3 priority.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ospfv3 1 area 0
huawei(config-if-vlanif2)#ospfv3 dr-priority 90
huawei(config-if-vlanif2)#quit
```

4. Save the data.

```
huawei(config)#save
```

Step 4 Configure Access Node_D.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#vlan 2
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1::4 64
huawei(config-if-vlanif2)#quit
```

2. Enable OSPFv3, and configure the OSPFv3 Router ID.

```
huawei(config)#ospfv3
huawei(config-ospfv3-1)#router-id 10.4.4.4
uawei(config-ospfv3-1)#quit
```

3. Enable OSPFv3 on the vlanif interface.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ospfv3 1 area 0
huawei(config-if-vlanif2)#quit
```

4. Save the data.

```
huawei(config)#save
```

----End

 Copyright © Huawei Technologies Co., Ltd.

Result

Run the **display ipv6 routing-table** command and you can find the learnt route table. Hosts can communicate with each other.

Configuration File

Configuration on each MA5800 is similar. Take Access Node_A for example.

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ipv6 enable
ipv6 address 2001:db8:1::1 64
quit
ospfv3
router-id 10.1.1.1
quit
interface vlanif 2
ospfv3 1 area 0
ospfv3 dr-priority 100
quit
save
```

11.12.7.4 Example for Configuring OSPFv3 Auto FRR

OSPFv3 auto FRR can generate a loop-free backup link for fast link switching in case of a link failure.

Service Requirements

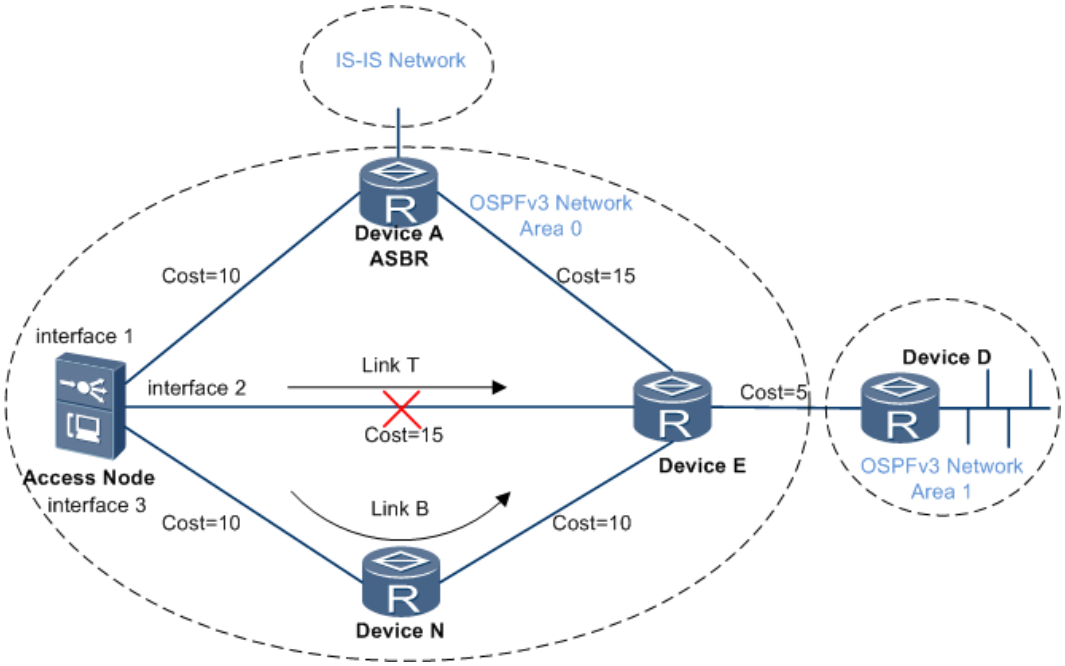
OSPFv3 auto FRR quickly switches traffic to the backup link without waiting for route convergence. Moreover, you can exclude a specified link from backup link options.

As shown in **Figure 11-150**,

- OSPFv3 is running on all devices.
- The link costs meet the OSPFv3 auto FRR traffic protection inequality.
- If the primary link T fails, traffic forwarded by the MA5800 is quickly switched to the backup link B.
- In network planning, you can exclude a link from backup link options.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-150 Networking for configuring OSPFv3 auto FRR



Data Plan

Table 11-78 describes data plan for configuring OSPFv3 auto FRR.

Table 11-78 Data plan for configuring OSPFv3 auto FRR

Item	Data
MA5800 port 1	IPv6 address of the interface: 2001:db8:1000::1 64
	VLAN ID: 2
	Cost: 10
MA5800 port 2	IPv6 address of the interface: 2001:db8:1001::1 64
	VLAN ID: 3
	Cost: 15
MA5800 port 3	IPv6 address of the interface: 2001:db8:1002::1 64
	VLAN ID: 4
	Cost: 10

 Copyright © Huawei Technologies Co., Ltd.

Prerequisite

Before you configure OSPFv3 auto FRR, complete the following tasks:

- Configuring link layer protocol parameters and interface IP addresses to ensure that the state of the link layer protocol on the interfaces is Up
- Enabling basic OSPFv3 functions on routers
- Configuring basic BFD for OSPFv3 functions on all devices in area 0 to quickly detect faults

Procedure

Step 1 Configure the MA5800.

1. Configure the IPv6 addresses of the Layer 3 interfaces on the MA5800.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:1000::1 64
huawei(config-if-vlanif2)#quit
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 1
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ipv6 enable
huawei(config-if-vlanif3)#ipv6 address 2001:db8:1001::1 64
huawei(config-if-vlanif3)#quit
huawei(config)#vlan 4 smart
huawei(config)#port vlan 4 0/9 2
huawei(config)#interface vlanif 4
huawei(config-if-vlanif4)#ipv6 enable
huawei(config-if-vlanif4)#ipv6 address 2001:db8:1002::1 64
huawei(config-if-vlanif4)#quit
```

2. Configure the costs so that the traffic preferentially goes through link T.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ospfv3 cost 10
huawei(config-if-vlanif2)#quit
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ospfv3 cost 15
huawei(config-if-vlanif3)#quit
huawei(config)#interface vlanif 4
huawei(config-if-vlanif4)#ospfv3 cost 10
huawei(config-if-vlanif4)#quit
```

3. Block IP FRR on the specified interface on the MA5800.

```
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ospfv3 frr block
huawei(config-if-vlanif2)#quit
```

4. Enable OSPFv3 auto FRR on the MA5800.

```
huawei(config)#ospfv3
huawei(config-ospfv3-1)#frr
huawei(config-ospfv3-1-frr)#loop-free-alternate
huawei(config-ospfv3-1-frr)#quit
huawei(config-ospfv3-1)#quit
```

5. Save the data.

```
huawei(config)#save
```

----End

 Copyright © Huawei Technologies Co., Ltd.

Result

Run the **display ospfv3 routing** command to display the information about primary/secondary routes configured.

Configuration File

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ipv6 enable
ipv6 address 2001:db8:1000::1 64
quit
vlan 3 smart
port vlan 3 0/9 1
interface vlanif 3
ipv6 enable
ipv6 address 2001:db8:1001::1 64
quit
vlan 4 smart
port vlan 4 0/9 2
interface vlanif 4
ipv6 enable
ipv6 address 2001:db8:1002::1 64
quit
interface vlanif 2
ospfv3 cost 10
quit
interface vlanif 3
ospfv3 cost 15
quit
interface vlanif 4
ospfv3 cost 10
quit
interface vlanif 2
ospfv3 frr block
quit
ospfv3
frr
loop-free-alternate
quit
quit
save
```

11.12.7.5 OSPFv3 Reference Standards and Protocols

The following table lists the references.

Document NO.	Description
RFC 2740	This document describes the modifications to OSPF to support IPv6.
RFC 5340	This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

 Copyright © Huawei Technologies Co., Ltd.

Document NO.	Description
draft-ietf-ospf-ospfv3-graceful-restart	This document describes the OSPFv3 graceful restart. The OSPFv3 graceful restart is identical to OSPFv2 except for the differences described in this document. These differences include the format of the grace Link State Advertisements (LSA) and other considerations.
draft-ietf-ospf-ospfv3-mib-11	This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in IPv6-based internets. In particular, it defines objects for managing the Open Shortest Path First Routing Protocol for IPv6.

11.12.8 BGP

11.12.8.1 Introduction

Definition

Border Gateway Protocol (BGP) is a dynamic routing protocol used between Autonomous Systems (ASs). BGP is widely used by Internet Service Providers (ISPs).

As three earlier-released versions of BGP, BGP-1 (defined in RFC 1105), BGP-2 (defined in RFC 1163), and BGP-3 (defined in RFC 1267) are used to exchange reachable inter-AS routes, establish inter-AS paths, avoid routing loops, and apply routing policies between ASs.

Currently, BGP-4 defined in RFC 4271 is used.

As an exterior routing protocol on the Internet, BGP is widely used by Internet service providers (ISPs).

BGP has the following characteristics:

- Unlike an Interior Gateway Protocol (IGP), such as Open Shortest Path First (OSPF) and Routing Information Protocol (RIP), BGP is an Exterior Gateway Protocol (EGP) which controls route advertisement and selects optimal routes between ASs rather than discovering or calculating routes.
- BGP uses Transport Control Protocol (TCP) with port number 179 as the transport layer protocol, which enhances BGP reliability.
  - BGP selects inter-AS routes, which poses high requirements on stability. Therefore, using TCP enhances BGP's stability.
  - BGP peers must be logically connected through TCP. The destination port number is 179 and the local port number is a random value.
- BGP supports Classless Inter-Domain Routing (CIDR).
- When routes are updated, BGP transmits only the updated routes, which reduces bandwidth consumption during BGP route distribution. Therefore,

 Copyright © Huawei Technologies Co., Ltd.

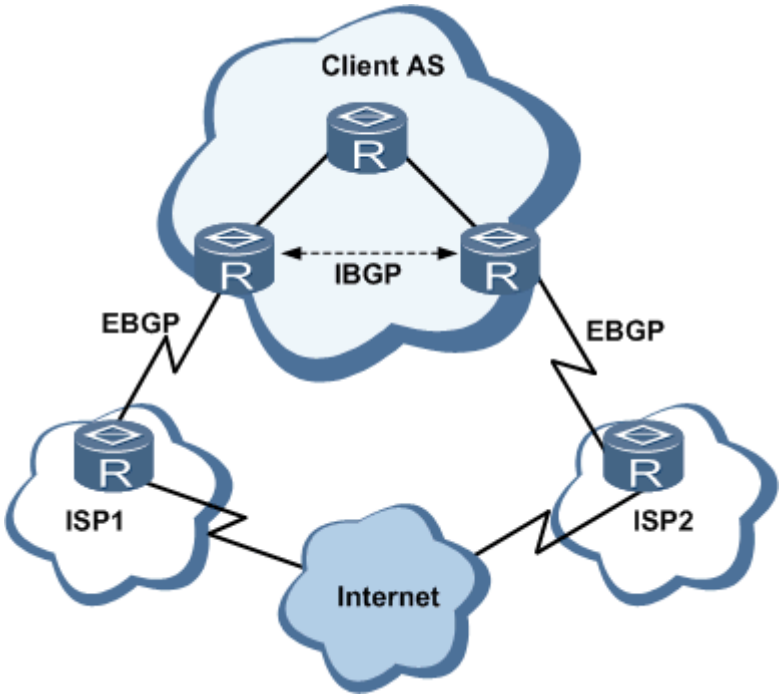
BGP is applicable to the Internet where a large number of routes are transmitted.

- BGP is a distance-vector routing protocol.
- BGP is designed to prevent loops.
  - Between ASs: BGP routes carry information about the ASs along the path. The routes that carry the local AS number are discarded to prevent inter-AS loops.
  - Within an AS: BGP does not advertise routes learned in an AS to BGP peers in the AS to prevent intra-AS loops.
- BGP provides many routing policies to flexibly select and filter routes.
- BGP provides a mechanism that prevents route flapping, which effectively enhances Internet stability.
- BGP can be easily extended.

Purpose

BGP transmits route information between ASs. It, however, is not required in all scenarios.

Figure 11-151 BGP networking



BGP is required in the following scenarios:

- On the network shown in **Figure 11-151**, users need to be connected to two or more ISPs. The ISPs need to provide all or part of the Internet routes for the users. Routers, therefore, need to select the optimal route through the AS of an ISP to the destination based on the attributes carried in BGP routes.

 Copyright © Huawei Technologies Co., Ltd.

- The AS_Path attribute needs to be transmitted between users in different organizations.
- Users need to transmit VPN routes through a Layer 3 VPN.

BGP is not required in the following scenarios:

- Users are connected to only one ISP.
- The ISP does not need to provide Internet routes for users.
- ASs are connected through default routes.

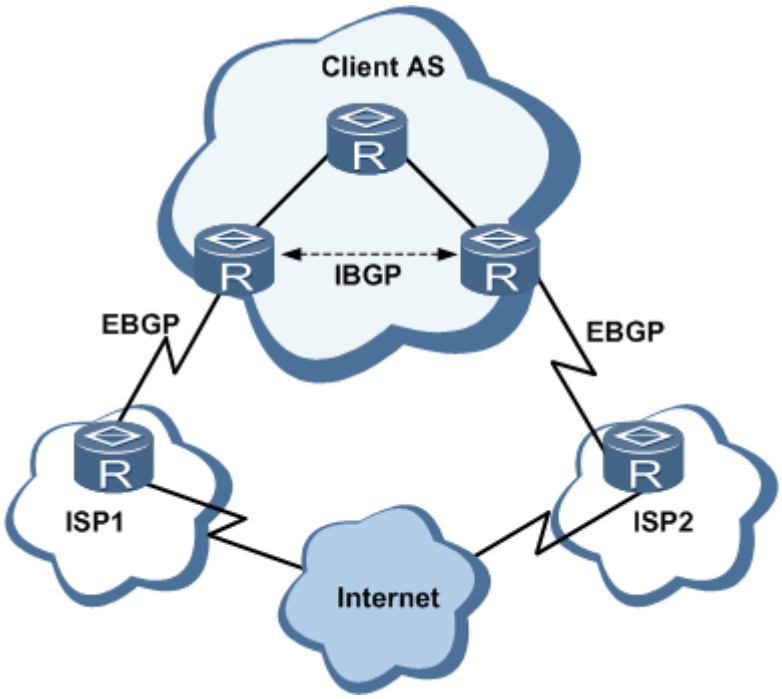
11.12.8.2 Principles

11.12.8.2.1 Basic Principle

BGP Operating Modes

BGP is called Internal BGP (IBGP) when it runs within an AS; it is called External BGP (EBGP) when it runs between ASs, as shown in **Figure 11-152**.

Figure 11-152 BGP operating modes



Roles in Transmitting BGP Messages

- Speaker: Any Router that sends BGP messages is called a BGP speaker. The speaker receives or generates new routing information and then advertises the routing information to other BGP speakers. After receiving a route from another AS, a BGP speaker compares the route with its local routes. If the route is better than its local routes, or the route is new, the speaker advertises this route to all other BGP speakers.

 Copyright © Huawei Technologies Co., Ltd.

- Peer: BGP speakers that exchange messages with each other are called peers.

BGP Messages

BGP runs by sending five types of messages: Open, Update, Notification, Keepalive, and Route-refresh.

- Open: The first message sent after a TCP connection is set up is an Open message, which is used to set up BGP peer relationships. After a peer receives an Open message and the peer negotiation is successful, the peer sends a Keepalive message to confirm and maintain the peer relationship. Then, peers can exchange Update, Notification, Keepalive, and Route-refresh messages.
- Update: This type of message is used to exchange routes between BGP peers.
  - An Update message can advertise multiple reachable routes with the same attributes. These route attributes are applicable to all destination addresses (expressed by IP prefixes) in the Network Layer Reachability Information (NLRI) field of the Update message.
  - An Update message can be used to delete multiple unreachable routes. Each route is identified by its destination address (using the IP prefix), which identifies the routes previously advertised between BGP speakers.
  - An Update message can be used only to delete routes. In this case, it does not need to carry the route attributes or NLRI. In addition, an Update message can be used only to advertise reachable routes. In this case, it does not need to carry information about the deleted routes.
- Notification: When BGP detects an error, it sends a Notification message to its peer. The BGP connection is then torn down immediately.
- Keepalive: BGP periodically sends Keepalive messages to peers to maintain peer relationships.
- Route-refresh: This type of message is used to request that the peer resend all reachable routes.

  If all BGP routers are enabled with the Route-refresh capability and the import policy of BGP changes, the local BGP router sends a Route-refresh message to its peers. After receiving the Route-refresh message, the peers resend their routing information to the local BGP router. In this manner, BGP routing tables are dynamically refreshed and new routing policies are used without tearing down BGP connections.

BGP Finite State Machine

The BGP Finite State Machine (FSM) has six states: Idle, Connect, Active, OpenSent, OpenConfirm, and Established.

Three common states during the establishment of BGP peer relationships are Idle, Active, and Established.

- In the Idle state, BGP denies all connection requests. This is the initial status of BGP.
- In the Connect state, BGP decides subsequent operations after a TCP connection is established.
- In the Active state, BGP attempts to establish a TCP connection.

 Copyright © Huawei Technologies Co., Ltd.

- In the OpenSent state, BGP is waiting for an Open message from the peer.
- In the OpenConfirm state, BGP is waiting for a Notification or Keepalive message.
- In the Established state, BGP peers can exchange Update, Route-refresh, Keepalive, and Notification messages.

The BGP peer relationship can be established only when both BGP peers are in the Established state. Both peers send Update messages to exchange routes.

BGP Processing

- BGP adopts TCP as its transport layer protocol. Therefore, a TCP connection must be available between the peers. BGP peers negotiate parameters by exchanging Open messages to establish a BGP peer relationship.
- After the peer relationship is established, BGP peers exchange BGP routing tables. BGP does not periodically update a routing table. When BGP routes change, BGP updates the changed BGP routes in the BGP routing table by sending Update messages.
- BGP sends Keepalive messages to maintain the BGP connection between peers.
- After detecting an error on a network, BGP sends a Notification message to report the error and the BGP connection is torn down.

BGP Attributes

BGP route attributes are a set of parameters that describe specific BGP routes. With BGP route attributes, BGP can filter and select routes. BGP route attributes are classified into the following types:

- Well-known mandatory: This type of attribute can be identified by all BGP routers and must be carried in Update messages. Without this attribute, errors occur in the routing information.
- Well-known discretionary: This type of attribute can be identified by all BGP routers. This type of attribute is optional and, therefore, is not necessarily carried in Update messages.
- Optional transitive: This indicates the transitive attribute between ASs. A BGP router may not recognize this attribute, but the router still receives it and advertises it to other peers.
- Optional non-transitive: If a BGP router does not recognize this type of attribute, the router does not advertise it to other peers.

The most common BGP route attributes are as follows:

- Origin

  The Origin attribute defines the origin of a route. The Origin attribute is classified into the following types:

  – Interior Gateway Protocol (IGP): This attribute type has the highest priority. IGP is the Origin attribute for routes obtained through an IGP of the AS from which the routes originate. For example, the Origin attribute of the routes imported to the BGP routing table using the **network** command is IGP.

 Copyright © Huawei Technologies Co., Ltd.

  - Exterior Gateway Protocol (EGP): This attribute type has the second highest priority. The Origin attribute of the routes obtained through EGP is EGP.
  - Incomplete: This attribute type has the lowest priority. Incomplete is the Origin attribute type of all routes that do not have the IGP or EGP Origin attribute. For example, for the routes imported using the **import-route** command by BGP, the Origin attribute is Incomplete.

- AS_Path

  The AS-Path attribute records all ASs through which a route passes from the local end to the destination in distance-vector (DV) order.

  When a BGP speaker advertises a local route:

  - When advertising the route beyond the local AS, the BGP speaker adds the local AS number to the AS_Path list and then advertises it to the neighboring Routers through Update messages.
  - When advertising the route within the local AS, the BGP speaker creates an empty AS_Path list in an Update message.

  When a BGP speaker advertises a route learned from the Update messages of another BGP speaker:

  - When advertising the route beyond the local AS, the BGP speaker adds the local AS number to the left of the AS_Path list. From the AS_Path attribute, the BGP router that receives the route learns the ASs through which the route passes to the destination. The number of the AS that is nearest to the local AS is placed on the left of the list, while other AS numbers are listed in sequence.
  - When advertising the route within the local AS, the BGP speaker does not change the AS_Path attribute.

  The AS_Path attribute has four types:

  - AS_Sequence: a sequenced set of numbers of the ASs that a route passes through from a local end to the destination
  - AS_Set: an unsequenced set of numbers of the ASs that a route passes through from a local end to the destination The AS_Set attribute is used in the route aggregation scenario. After route aggregation, the device cannot sequence the numbers of ASs that specific routes pass through, so the AS_Set attribute is used to record the unsequenced AS numbers. No matter how many AS numbers an AS_Set contains, BGP regards the AS_Set as one AS number to calculate routes.
  - AS_Confed_Sequence: a sequenced set of sub-AS numbers in a confederation
  - AS_Confed_Set: an unsequenced set of sub-AS numbers in a confederation. The AS_Confed_Set attribute is used in the route aggregation scenario in a confederation.

  The AS_Confed_Sequence and AS_Confed_Set attributes are used to prevent route loops and to select routes among the various sub-ASs in a confederation.

- Next_Hop

  The Next_Hop attribute in BGP is different from that in IGP. It is not necessarily the IP address of a neighboring Router. Generally, the Next_Hop attribute complies with the following rules:

 Copyright © Huawei Technologies Co., Ltd.

  - When advertising a route to an EBGP peer, a BGP speaker sets the next hop of the route to the address of the local interface through which the BGP peer relationship is established.
  - When advertising a local route to an IBGP peer, a BGP speaker sets the next hop of the route to the address of the local interface through which the BGP peer relationship is established.
  - When advertising a route learned from an EBGP peer to an IBGP peer, the BGP speaker does not change the next hop of the route.
- MED

  The Multi-Exit-Discriminator (MED) is transmitted only between two neighboring ASs. The AS that receives the MED does not advertise it to a third AS.

  Similar to the metric used by an IGP, the MED is used to determine the optimal route when traffic enters an AS. When a BGP peer obtains multiple routes to the same destination address but with different next hops through EBGP peers, the route with the smallest MED value is selected as the optimal route if all other attributes are the same.
- Local_Pref

  The Local_Pref attribute indicates the BGP priority of a route. It is available only to IBGP peers and is not advertised to other ASs.

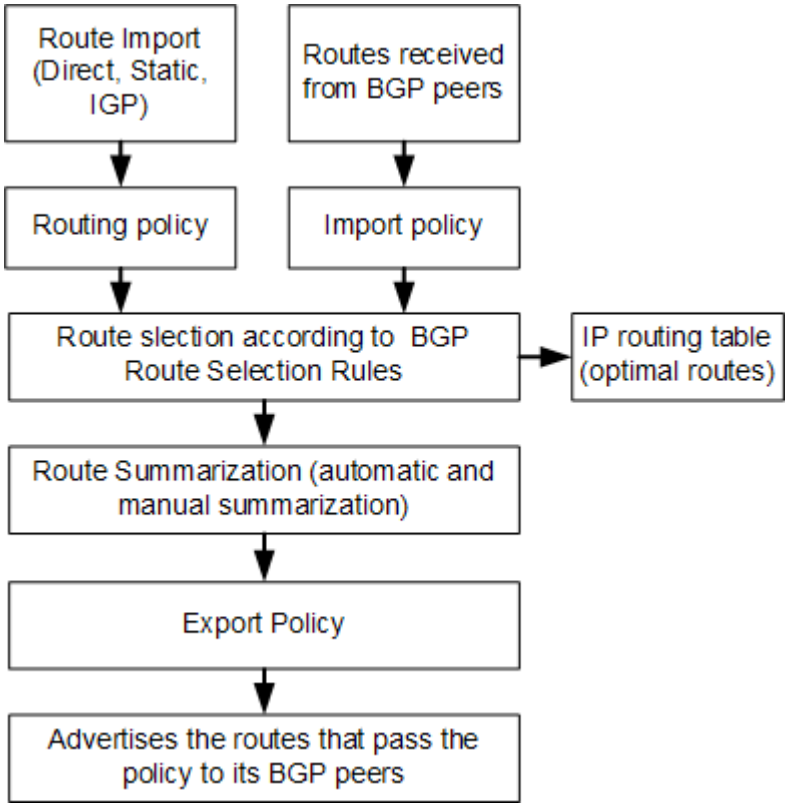
  The Local_Pref attribute is used to determine the optimal route when traffic leaves an AS. When a BGP Router obtains multiple routes to the same destination address but with different next hops through IBGP peers, the route with the largest Local_Pref value is selected.

11.12.8.2.2 BGP Route Processing

Figure 11-153 shows how BGP processes routes. BGP routes can be imported from other protocols or learned from peers. To reduce the routing size, you can configure route summarization after BGP selects routes. In addition, you can configure route-policies and apply them to route import, receipt, or advertisement to filter routes or modify route attributes.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-153 BGP route processing



Route Import

BGP itself cannot discover routes. Therefore, it needs to import other protocol routes, such as IGP routes or static routes, to the BGP routing table. Imported routes can be transmitted within an AS or between ASs.

BGP routes are imported in either of the following modes:

- The **import** command imports routes based on protocol types, such as RIP routes, OSPF routes, Intermediate System to Intermediate System (IS-IS) routes, static routes, or direct routes.
- The **network** command imports a route with the specified prefix and mask to the BGP routing table, which is more precise than the previous mode.

BGP Route Selection

When multiple routes to the same destination are available, BGP selects routes based on the following rules:

1. Prefers routes in descending order of Valid, Not Found, and Invalid after BGP origin AS validation states are applied to route selection in a scenario where the device is connected to an RPKI server.
2. Prefers the route with the largest PreVal value.

   PrefVal is Huawei-specific. It is valid only on the device where it is configured.
3. Prefers the route with the highest Local_Pref.

   If a route does not carry Local_Pref, the default value 100 takes effect. To change the value, run the **default local-preference** command.

 Copyright © Huawei Technologies Co., Ltd.

4. Prefers a locally originated route to a route learned from a peer.

   Locally originated routes include routes imported using the **network** or **import-route** command, as well as manually and automatically summarized routes.

   a. Prefers a summarized route over a non-summarized route.

   b. Prefers a route obtained using the **aggregate** command over a route obtained using the **summary automatic** command.

   c. Prefers a route imported using the **network** command over a route imported using the **import-route** command.

5. Prefers the route with the shortest AS_Path length.

   - The AS_CONFED_SEQUENCE and AS_CONFED_SET are not included in the AS_Path length.
   - During route selection, a router assumes that an AS_SET carries only one AS number regardless of the actual number of ASs it is carrying.
   - If the **bestroute as-path-ignore** command is run, BGP no longer compares the AS_Path attribute.

6. Prefers the route with the Origin type as IGP, EGP, and Incomplete in descending order.

7. Prefers the route with the smallest MED value.

   - BGP compares the MEDs of only routes from the same AS (excluding confederation sub-ASs). MEDs of two routes are compared only when the first AS number in the AS_Sequence (excluding AS_Confed_Sequence) of one route is the same as its counterpart in the other route.
   - If a route does not carry MED, BGP considers its MED as the default value (0) during route selection. If the **bestroute med-none-as-maximum** command is run, BGP considers its MED as the largest MED value (4294967295).
   - If the **compare-different-as-med** command is run, BGP compares MEDs of routes even when the routes are received from peers in different ASs. Do not run this command unless the ASs use the same IGP and route selection mode. Otherwise, a loop may occur.
   - If the **deterministic-med** command is run, routes are no longer selected in the sequence in which they are received.

8. Prefers EBGP routes to IBGP routes.

   EBGP, IBGP, LocalCross, and RemoteCross routes are preferred in descending order.

   If the ERT of a VPNv4 route in the routing table of a VPN instance on a PE matches the IRT of another VPN instance on the PE, the VPNv4 route is added to the routing table of the second VPN instance. This route is called a LocalCross route. If the ERT of a VPNv4 route learned from a remote PE matches the IRT of a VPN instance on the local PE, the VPNv4 route is added to the routing table of that VPN instance. This route is called a RemoteCross route.

9. Prefers the route that is iterated to an IGP route with the smallest cost.

   If the **bestroute igp-metric-ignore** command is run, BGP no longer compares the IGP cost.

 Copyright © Huawei Technologies Co., Ltd.

NOTE

With load balancing, if the preceding conditions are equal and multiple external routes with the same AS_Path are available, load balancing is performed among them. The number of routes load-balancing traffic must be less than or equal to the configured number. After the **load-balancing as-path-ignore** command is run, the routes with different As_Path values can load-balance traffic.

10. Prefers the route with the shortest Cluster_List length.
11. Prefers the route advertised by the Router with the smallest router ID.

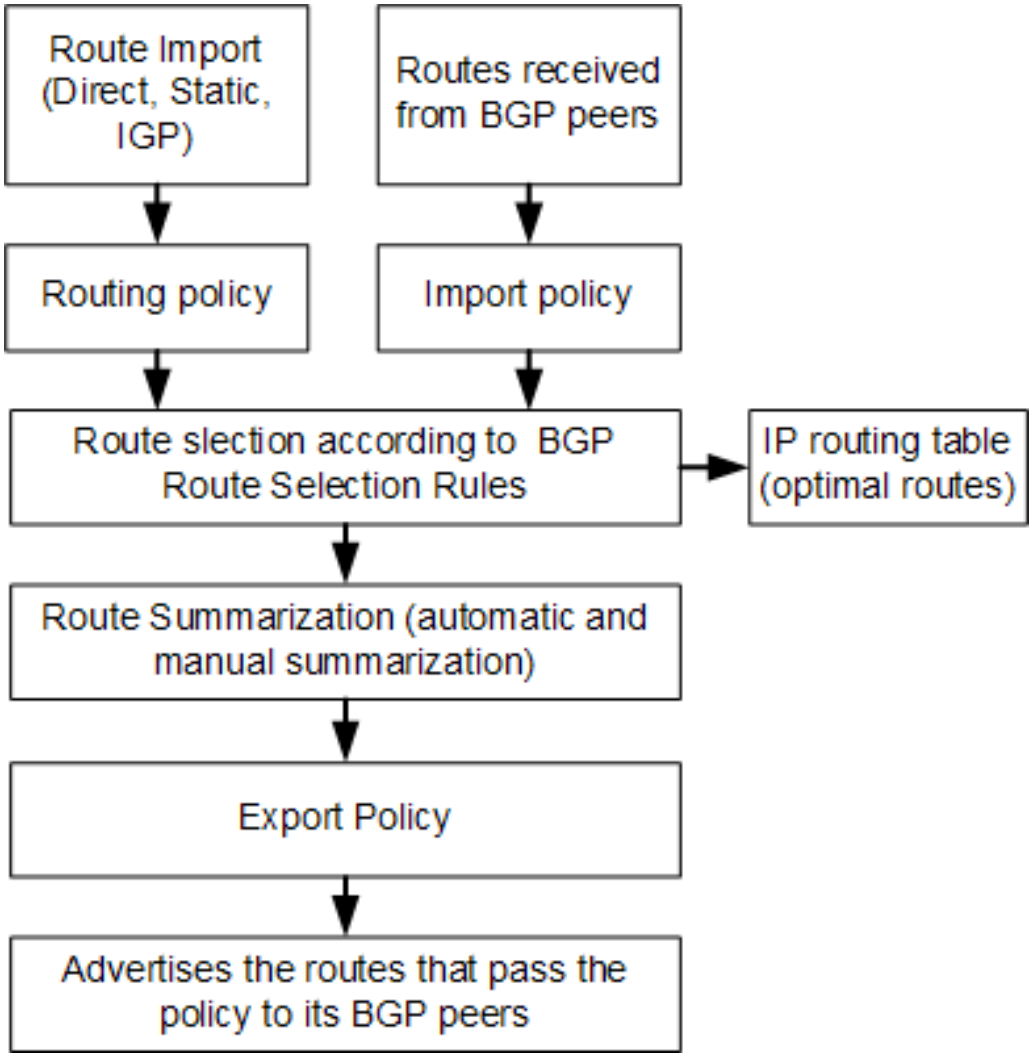
    If the **bestroute router-id-ignore** command is run, router IDs do not determine which route is selected for BGP.

    NOTE

    If each route carries an Originator_ID, the originator IDs rather than router IDs are compared during route selection. The route with the smallest Originator_ID is preferred.
12. Prefers the route learned from the peer with the smallest IP address.

For details about the BGP route selection process, see **Figure 11-154**.

Figure 11-154 BGP route selection process



 Copyright © Huawei Technologies Co., Ltd.

Route Summarization

On a large-scale network, the BGP routing table can be very large. Route summarization can reduce the size of the routing table.

Route summarization is the process of summarizing specific routes with the same IP prefix into a summarized route. After route summarization, BGP advertises only the summarized route rather than all specific routes to BGP peers.

BGP supports automatic and manual route summarization.

- Automatic route summarization: takes effect on the routes imported by BGP. With automatic route summarization, the specific routes for the summarization are suppressed, and BGP summarizes routes based on the natural network segment and sends only the summarized route to BGP peers. For example, 10.1.1.1/24 and 10.2.1.1/24 are summarized into 10.0.0.0/8, which is a Class A address.
- Manual route summarization: takes effect on routes in the local BGP routing table. With manual route summarization, users can control the attributes of the summarized route and determine whether to advertise the specific routes.

IPv4 supports both automatic and manual route summarization, while IPv6 supports only manual route summarization.

BGP Route Advertisement

BGP adopts the following policies to advertise routes:

- When there are multiple valid routes, a BGP speaker advertises only the optimal route to its peers.
- A BGP speaker advertises the routes learned from EBGP peers to all BGP peers, including EBGP peers and IBGP peers.
- A BGP speaker does not advertise the routes learned from an IBGP peer to other IBGP peers.
- A BGP speaker advertises the routes learned from IBGP peers to its EBGP peers.
- A BGP speaker advertises all BGP optimal routes to new peers after peer relationships are established.

11.12.8.2.3 Peer Group

A peer group is a set of peers with the same policies. When a peer joins a peer group, it inherits the configurations of the peer group. If the configurations of the peer group change, the configurations of all the peers in the group changes accordingly.

On a large-scale BGP network, there are many peers and most of them need the same policies. Therefore, some commands need to be run repeatedly on each peer. Configuring a peer group can simplify the configuration.

Each peer in a peer group can be configured with unique policies to advertise and receive routes.

 Copyright © Huawei Technologies Co., Ltd.

11.12.8.2.4 Route Dampening

Route instability is reflected by route flapping. When a route flaps, it repeatedly disappears from the routing table and then reappears.

NOTE

A route is added to the routing table, and then is withdrawn. This process is called one route flapping.

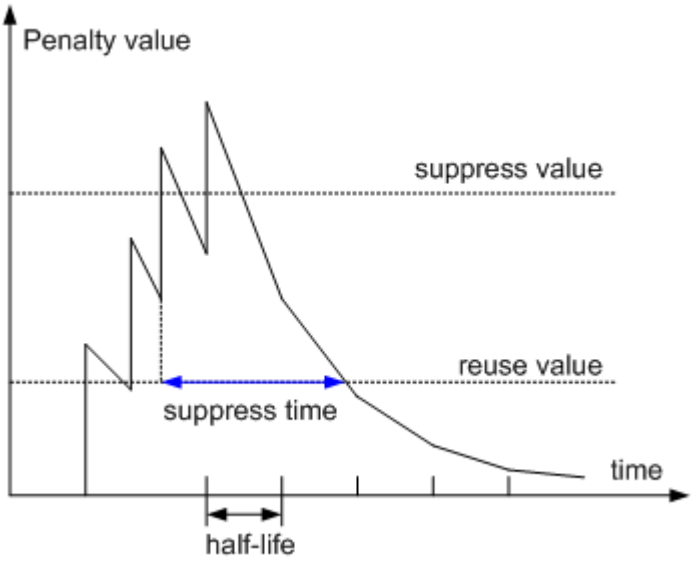
If route flapping occurs, a Router sends an Update packet to its peers. After the peers receive the Update packet, they recalculate routes and update their routing tables. Frequent route flapping consumes lots of bandwidth and CPU resources and can even affect network operations.

Route dampening can address this problem. In most cases, BGP is deployed on complex networks where routes change frequently. To reduce the impact of frequent route flapping, BGP adopts route dampening to suppress unstable routes.

BGP dampening measures route stability using a penalty value. The greater the penalty value, the more unstable a route. Each time route flapping occurs, BGP adds a penalty value to this route. If a route changes from active to inactive, the penalty value increases by 1000. If a route is updated when it is active, the penalty value increases by 500. When the penalty value of a route exceeds the suppress value, the route is suppressed. As a result, BGP does not add the route to the routing table or advertise any Update message to BGP peers.

The penalty value of a suppressed route reduces by half after a half-life period. When the penalty value decreases to the reuse value, the route is reusable and is added to the routing table. At the same time, BGP advertises an Update message to BGP peers. The penalty value, suppression threshold, and half-life are configurable. **Figure 11-155** shows the process of BGP route dampening.

Figure 11-155 BGP route dampening



Route dampening applies only to EBGP routes. IBGP routes cannot be dampened because IBGP routing tables contain the routes from the local AS, which require

 Copyright © Huawei Technologies Co., Ltd.

that the forwarding entries be the same on IBGP peers in the AS. If IBGP routes are dampened, the forwarding entries may be inconsistent because dampening parameters may vary among these IBGP peers.

11.12.8.2.5 Community Attribute

A community is a set of destination addresses with the same characteristics. It is four bytes long. On the , the community attribute is expressed in the format of aa:nn or as a community number.

- aa:nn: aa indicates an AS number and nn indicates the community identifier defined by an administrator. The value of aa or nn ranges from 0 to 65535, which is configurable. For example, if a route is from AS 100 and the community identifier defined by the administrator is 1, the community is 100:1.
- Community number: It is an integer ranging from 0 to 4294967295. As defined in RFC 1997, numbers from 0 (0x00000000) to 65535 (0x0000FFFF) and from 4294901760 (0xFFFF0000) to 4294967295 (0xFFFFFFFF) are reserved.

The community attribute is used to simplify the application, maintenance, and management of routing policies. With the community attribute, a group of BGP peers in multiple ASs can share the same routing policy. The community attribute is a route attribute. It is transmitted between BGP peers and is not restricted by the AS. Before advertising a route with the community attribute to peers, a BGP peer can change the original community attribute of this route.

The peers in a peer group share the same policy, while the routes with the same community attribute share the same policy.

The well-known communities are described in the next section. Users can also create their own communities to filter routes.

Well-known Community

Table 11-79 lists well-known communities of BGP routes.

Table 11-79 Well-known communities of BGP routes

Community Name	Community Identifier	Description
Internet	0 (0x00000000)	By default, all routes belong to the Internet community. A route with this attribute can be advertised to all BGP peers.
No_Export	4294967041 (0xFFFFFF01)	A route with this attribute cannot be advertised beyond the local AS.
No_Advertise	4294967042 (0xFFFFFF02)	A route with this attribute cannot be advertised to any other BGP peers.

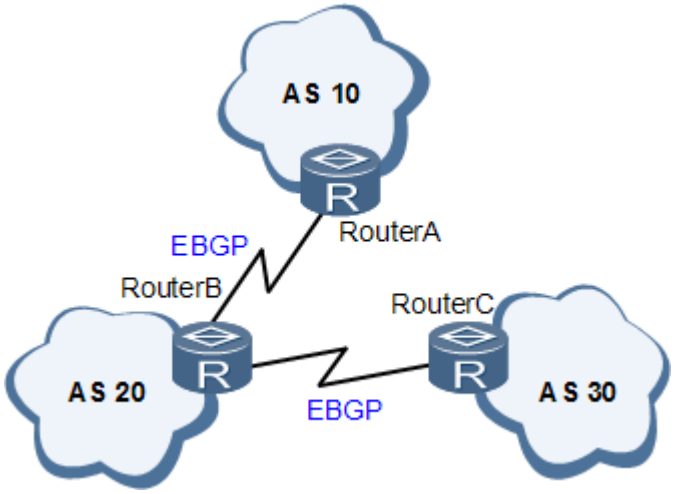
 Copyright © Huawei Technologies Co., Ltd.

Community Name	Community Identifier	Description
No_Export_Subconfed	4294967043 (0xFFFFFF03)	A route with this attribute cannot be advertised beyond the local AS or to other sub-ASs.

Usage Scenario

On the network shown in **Figure 11-156**, EBGP connections are established between Router A and Router B, and between Router B and Router C. If the community attribute of No_Export is configured on Router A and Router A sends a route with the community attribute to Router B, Router B does not advertise the route to other ASs after receiving it.

Figure 11-156 Networking for BGP communities



11.12.8.2.6 Route Reflector

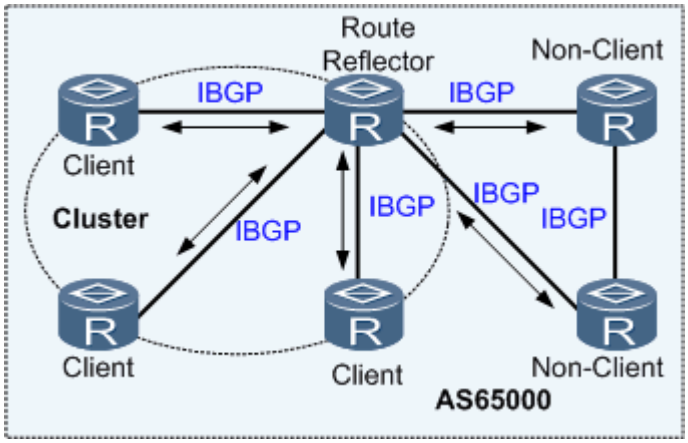
Fully meshed connections need to be established between IBGP peers to ensure the connectivity between IBGP peers. If there are n Routers in an AS, n x (n-1)/2 IBGP connections need to be established. When there are a lot of IBGP peers, network resources and CPU resources are greatly consumed. Route reflection can solve the problem.

In an AS, one Router functions as a Route Reflector (RR) and the other Routers as clients. The clients establish IBGP connections with the RR. The RR and its clients form a cluster. The RR reflects routes among clients, and BGP connections do not need to be established between the clients.

A BGP peer that functions as neither an RR nor a client is called a non-client. A non-client must establish full meshed connections with the RR and all the other non-clients, as shown in **Figure 11-157**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-157 Networking with an RR



Applications

After an RR receives routes from its peers, it selects the optimal route based on BGP route selection policies and performs one of the following operations:

- If the optimal route is from a non-client IBGP peer, the RR advertises the route to all clients.
- If the optimal route is from a client, the RR advertises the route to all non-clients and clients.
- If the optimal route is from an EBGP peer, the RR advertises the route to all clients and non-clients.

An RR is easy to configure because it only needs to be configured on the RR itself, and clients do not need to know whether they are clients.

On some networks, if fully meshed connections have already been established among clients of an RR, they can exchange routing information directly. In this case, route reflection among the clients is unnecessary and occupies bandwidth. For example, on the , route reflection can be disabled, but the routes between clients and non-clients are still exchanged. By default, route reflection between clients is enabled.

On the , an RR can change various attributes of BGP routes, such as the AS_Path, MED, Local_Pref, and community attributes.

Originator_ID

Originator_ID and Cluster_List, defined in RFC 2796, are used to detect and prevent routing loops.

The Originator_ID attribute is four bytes long and is generated by an RR. It carries the router ID of the originator of the route in the local AS.

- When a route is reflected by an RR for the first time, the RR adds the Originator_ID to this route. If a route already carries an Originator_ID, the RR does not add a new one.

 Copyright © Huawei Technologies Co., Ltd.

- After receiving the route, a BGP speaker checks whether the Originator_ID is the same as its router ID. If Originator_ID is the same as its router ID, the BGP speaker discards this route.

Cluster_List

To prevent routing loops between ASs, a BGP Router uses the AS_Path attribute to record the ASs through which a route passes. Routes with the local AS number are discarded by the Router. To prevent routing loops within an AS, IBGP peers do not advertise routes learned from the local AS.

With RR, IBGP peers can advertise routes learned from the local AS to each other. However, the Cluster_List attribute must be deployed to prevent routing loops within the AS.

An RR and its clients form a cluster. In an AS, each RR is uniquely identified by a Cluster_ID.

Similar to an AS_Path, a Cluster_List is composed of a series of Cluster_IDs and is generated by an RR. It records all the RRs through which a route passes.

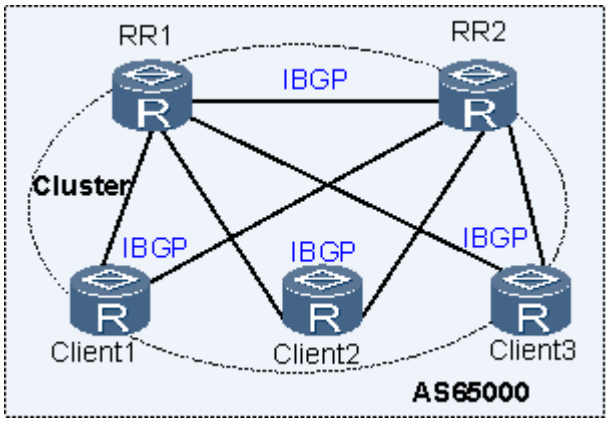
- Before an RR reflects a route between its clients or between its clients and non-clients, the RR adds the local Cluster_ID to the head of the Cluster_List. If a route does not carry any Cluster_List, the RR creates one for the route.
- After the RR receives an updated route, it checks the Cluster_List of the route. If the RR finds that its cluster ID is included in the Cluster_List, the RR discards the route. If its cluster ID is not included in the Cluster_List, the RR adds its cluster ID to the Cluster_List and then reflects the route.

Backup RR

To enhance network reliability and prevent single points of failure, more than one route reflector needs to be configured in a cluster. The route reflectors in the same cluster must share the same Cluster_ID to prevent routing loops.

With backup RRs, clients can receive multiple routes to the same destination from different RRs. The clients then apply route selection policies to choose the optimal route.

Figure 11-158 Backup RR



 Copyright © Huawei Technologies Co., Ltd.

On the network shown in **Figure 11-158**, RR1 and RR2 are in the same cluster. RR1 and RR2 establish an IBGP connection so that each RR is a non-client of the other RR.

- If Client 1 receives an updated route from an external peer, Client 1 advertises the route to RR1 and RR2 through IBGP.
- After receiving the updated route, RR1 reflects the route to other clients (Client 2 and Client 3) and the non-client (RR2) and adds the local Cluster_ID to the head of the Cluster_List.
- After receiving the reflected route, RR2 checks the Cluster_List. RR2 finds that its Cluster_ID is contained in the Cluster_List; therefore, it discards the updated route.

If RR1 and RR2 are configured with different Cluster_IDs, each RR receives both the route from Client 1 and the updated route reflected from the other RR. Therefore, configuring the same Cluster_ID for RR1 and RR2 reduces the number of routes that each RR receives and memory consumption.

NOTE

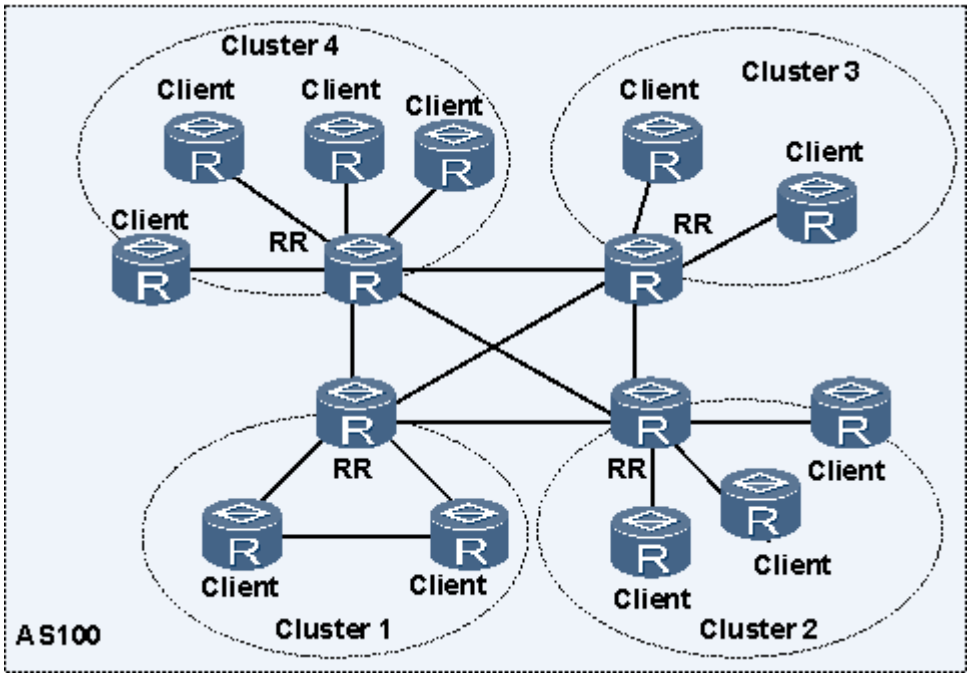
The application of Cluster_List prevents routing loops among RRs in the same AS.

Multiple Clusters in an AS

Multiple clusters may exist in an AS. RRs are IBGP peers of each other. An RR can be configured as a client or non-client of another RR.

For example, the backbone network shown in **Figure 11-159** is divided into multiple clusters. Each RR is configured as a non-client of the other RRs, and these RRs are fully meshed. Each client establishes IBGP connections with only the RR in the same cluster. In this manner, all BGP peers in the AS can receive reflected routes.

Figure 11-159 Multiple clusters in an AS



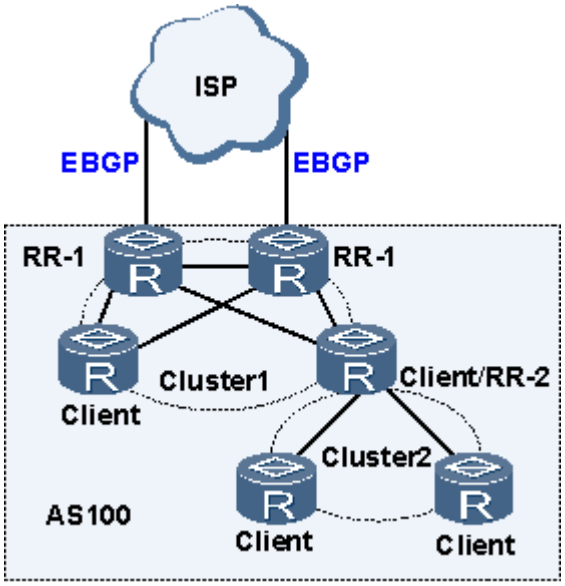
 Copyright © Huawei Technologies Co., Ltd.

Hierarchical Reflector

Hierarchical reflectors are deployed on live networks. On the network shown in **Figure 11-160**, the ISP provides Internet routes for AS 100. Two EBGP connections are established between the ISP and AS 100. AS 100 is divided into two clusters. The four Routers in Cluster 1 are core routers.

- Two Level-1 RRs (RR-1s) are deployed in Cluster 1, which ensures the reliability of the core layer of AS 100. The other two Routers in the core layer are clients of RR-1s.
- One Level-2 RR (RR-2) is deployed in Cluster 2. RR-2 is a client of RR-1.

Figure 11-160 Hierarchical reflector

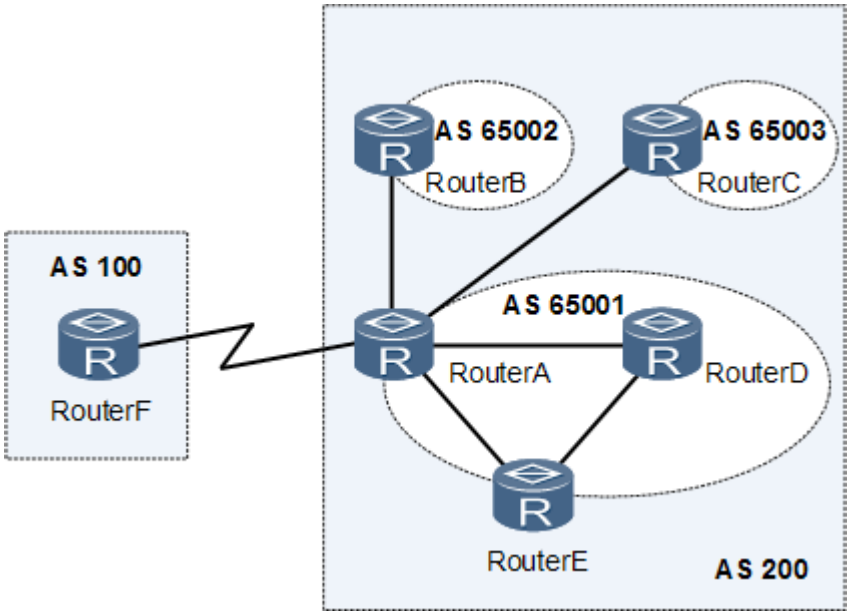


11.12.8.2.7 BGP Confederation

Besides RR, BGP confederation can also reduce IBGP connections in an AS. It divides an AS into several sub-ASs. Fully meshed IBGP connections are established in each sub-AS, and fully meshed EBGP connections are established between sub-ASs. As shown in **Figure 11-161**.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-161 BGP confederation



As shown in **Figure 11-161**, there are multiple BGP routers in AS 200. To reduce the number of IBGP connections, AS 200 is divided into three sub-ASs: AS 65001, AS 65002, and AS 65003. In AS 65001, fully meshed IBGP connections are established between the three routers.

BGP speakers outside a confederation such as Router F in AS 100, do not know the existence of the sub-ASs (AS 65001, AS 65002, and AS 65003) in the confederation. The confederation ID is the AS number that is used to identify the entire confederation. For example, AS 200 in **Figure 11-161** is the confederation ID.

Applications and Limitations

The confederation needs to be configured on each Router, and the Router that joins the confederation must support the confederation function.

BGP speakers need to be reconfigured when a network in non-confederation mode switches to confederation mode. As a result, the logical topology changes accordingly.

On large-scale BGP networks, the RR and confederation can both be used.

11.12.8.2.8 MP-BGP

Conventional BGP-4 manages only IPv4 unicast routing information, and inter-AS transmission of packets of other network layer protocols, such as IPv6 and multicast, is limited.

To support multiple network layer protocols, the Internet Engineering Task Force (IETF) extends BGP-4 to Multiprotocol Extensions for BGP-4 (MP-BGP). RFC 4760 defines the MP-BGP standard. MP-BGP is forward compatible. Specifically, Routers supporting MP-BGP can communicate with the Routers that do not support MP-BGP.

 Copyright © Huawei Technologies Co., Ltd.

Extended Attributes

BGP-4 Update packets carry three IPv4-related attributes: NLRI (Network Layer Reachable Information), Next_Hop, and Aggregator. Aggregator contains the IP address of the BGP speaker that performs route aggregation.

To carry information about multiple network layer protocols in NLRI and Next_Hop, MP-BGP introduces the following route attributes:

- MP_REACH_NLRI: indicates the multiprotocol reachable NLRI. It is used to advertise a reachable route and its next hop.
- MP_UNREACH_NLRI: indicates the multiprotocol unreachable NLRI. It is used to delete an unreachable route.

The preceding two attributes are optional non-transitive. Therefore, the BGP speakers that do not support MP-BGP will ignore the information carried in the two attributes and do not advertise the information to other peers.

Address Family

The Address Family Information field consists of a 2-byte Address Family Identifier (AFI) and a 1-byte Subsequent Address Family Identifier (SAFI).

BGP uses address families to distinguish different network layer protocols. For the values of address families, see RFC 3232 (Assigned Numbers). The supports multiple MP-BGP extension applications, such as VPN extension and IPv6 extension, which are configured in their respective address family views.

11.12.8.2.9 BGP GR

Graceful restart (GR) is one of the high availability (HA) technologies, which comprise a series of comprehensive technologies such as fault-tolerant redundancy, link protection, faulty node recovery, and traffic engineering. As a fault-tolerant redundancy technology, GR ensures normal forwarding of data during the restart of routing protocols to prevent interruption of key services. Currently, GR has been widely applied to the master/slave switchover and system upgrade.

GR is usually used when the active route processor (RP) fails because of a software or hardware error, or used by an administrator to perform the master/slave switchover.

Prerequisite for Implementation

On a traditional routing device, a processor implements both control and forwarding. The processor finds routes based on routing protocols, and maintains the routing table and forwarding table of the device. Mid-range and high-end devices generally adopt the multi-RP structure to improve forwarding performance and reliability. The processor in charge of routing protocols is located on the main control board, whereas the processor responsible for data forwarding is located on the interface board. The design helps to ensure the continuity of packet forwarding on the interface board during the restart of the main processor. The technology that separates control from forwarding satisfies the prerequisite for GR implementation.

 Copyright © Huawei Technologies Co., Ltd.

At present, a GR-capable device must have two main control boards. In addition, the interface board must have an independent processor and memory.

Related Concepts

The concepts related to GR are as follows:

- GR Restarter: indicates a device that performs master/slave switchover triggered by the administrator or a failure. A GR Restarter must support GR.
- GR helper: indicates the neighbor of a GR Restarter. A GR helper must support GR.
- GR session: indicates a session, through which a GR Restarter and a GR helper can negotiate GR capabilities.
- GR time: indicates the time when the GR helper finds that the GR Restarter is Down but keeps the topology information or routes obtained from the GR Restarter.
- End-of-RIB (EOR): indicates a BGP information, notifying a peer BGP that the first route upgrade is finished after the negotiation.
- EOR timer: indicates a maximum time of a local device waiting for the EOR information sent from the peer. If the local device does not receive the EOR information from the peer within the EOR timer, the local device will select an optimal route from the current routes.

Principles

Principles of BGP GR are as follows:

1. During BGP peer relationship establishment, devices negotiate GR capabilities by sending supported GR capabilities to each other.
2. When detecting the master/slave switchover of the GR Restarter, a GR helper does not delete the routing information and forwarding entries related to the GR Restarter within the GR time, but waits to re-establish a BGP connection with the GR Restarter.
3. After the master/slave switchover, the GR Restarter receives routes from all the negotiated peers with GR capabilities before the switchover, and starts the EOR timer. The GR Restarter selects a route when either of the following conditions is met:
   - The GR Restarter receives the EOR information of all peers and the EOR timer is deleted.
   - The EOR timer times out but the GR Restarter receives no EOR information from all peers.
4. The GR Restarter sends the optimal route to the GR Helper and the GR Helper starts the EOR timer. The GR Helper quits GR when either of the following conditions is met:
   - The GR Helper receives the EOR information from the GR Restarter and the EOR timer is deleted.
   - The EOR timer times out and the GR Helper receives no EOR information from the GR Restarter.

   NOTE

   Currently, the system supports only the GR helper.

 Copyright © Huawei Technologies Co., Ltd.

GR Reset

Currently, BGP does not support dynamic capability negotiation. Therefore, each time a new BGP capability (such as the IPv4, IPv6, VPNv4, and VPNv6 capabilities) is enabled on a BGP speaker, the BGP speaker tears down existing sessions with its peer and renegotiates BGP capabilities. This process will interrupt ongoing services.

To prevent the service interruptions, the provides the GR reset function that enables the to reset a BGP session in GR mode. With the GR reset function configured, when you enable a new BGP capability on the BGP speaker, the BGP speaker enters the GR state, resets the BGP session, and renegotiates BGP capabilities with the peer. In the whole process, the BGP speaker re-establishes the existing sessions but does not delete the routing entries for the existing sessions, so that the existing services are not interrupted.

Benefits

Through BGP GR, the forwarding is not interrupted. In addition, the flapping of BGP occurs only on the neighbors of the GR Restarter, and does not occur in the entire routing domain. This is important for BGP that needs to process a large number of routes.

11.12.8.2.10 BGP Dynamic Update Peer-Groups

As the routing table increases in size and the network topology increases in complexity, BGP needs to be able to support more peers. When the Router needs to send a large number of routes to many BGP peers and most of the peers share the same outbound policy, if the Router groups each route and then send the route to each peer, the efficiency is low.

The dynamic update peer-groups feature can address this problem. It considers all the BGP peers with the same outbound policy as an update-group. With the dynamic update peer-groups feature, each route to be sent is grouped once for all and then sent to all peers in the update-group, improving grouping efficiency and forwarding performance exponentially.

Usage Scenario

The BGP dynamic update peer-groups feature is applicable to the following scenarios:

- Scenario with an international gateway
- Scenario with an RR
- Scenario where routes received from EBGP neighbors need to be sent to all IBGP peers

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-162 Networking for the international gateway

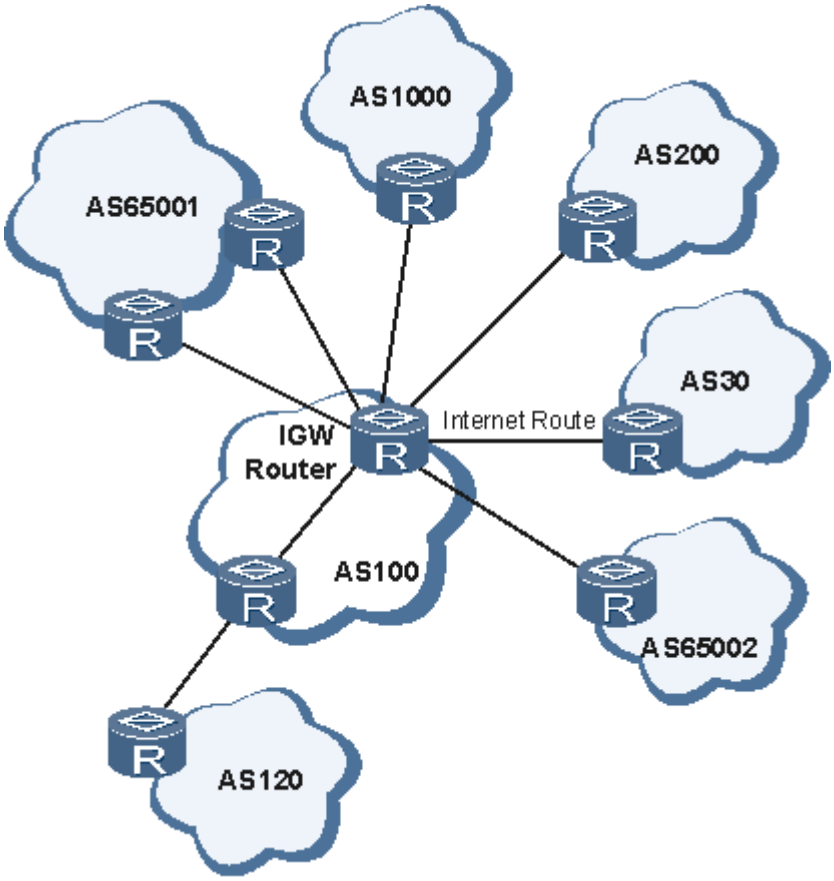
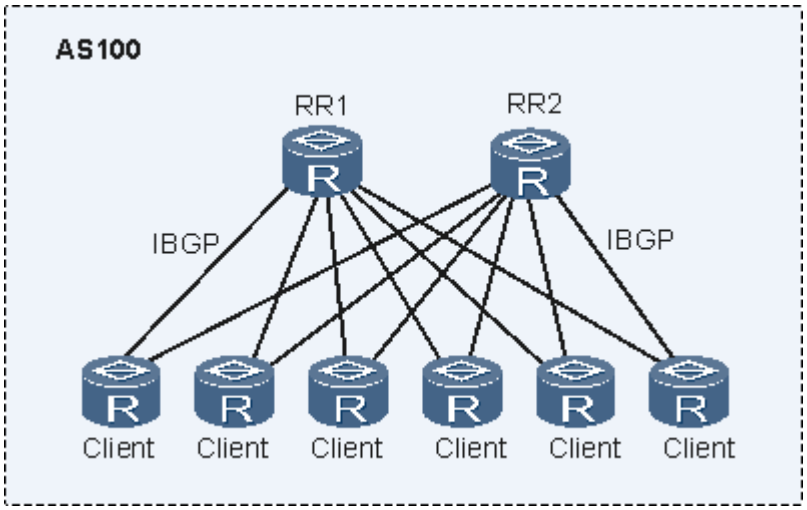
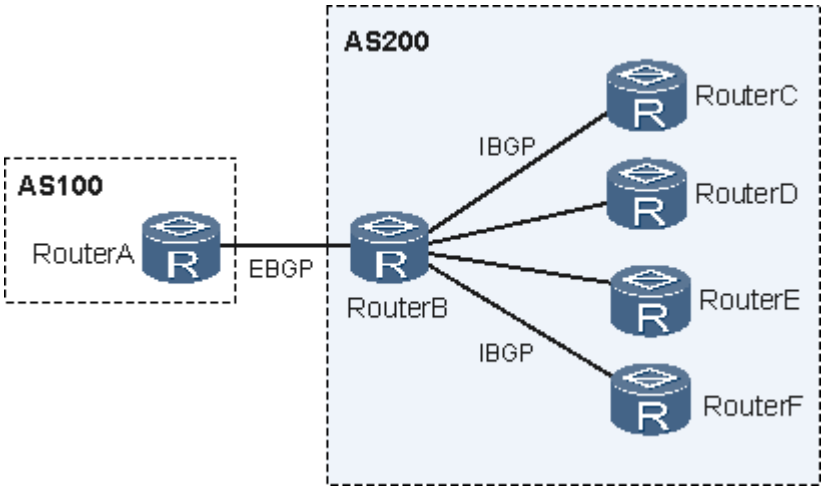


Figure 11-163 Networking for RRs with many clients



 Copyright © Huawei Technologies Co., Ltd.

Figure 11-164 Networking for a PE connected to multiple IBGP neighbors



The preceding scenarios have in common that a router needs to send routes to a large number of BGP peers, most of which share the same outbound policy. This situation is most evident in the networking shown in **Figure 11-163**.

For example, an RR has 100 clients and needs to reflect 100,000 routes to them. If the RR groups the routes for each peer before sending the routes to 100 clients, the total number of times that all routes are grouped is 100,000 x 100. After the dynamic update peer-groups feature is applied, the total number of times that all routes are grouped changes to 100,000 x 1. The efficiency is 100 times higher than before.

11.12.8.2.11 BGP NSR

The Border Gateway Protocol (BGP) non-stop routing (NSR) technique ensures the control plane connection to peers and uninterrupted traffic transmission on the forwarding plane if the BGP control plane fails because of active main board (AMB) failures.

Background

Carriers have increasing demands for IP network reliability. Conventional non-stop forwarding (NSF) and graceful restart (GR) techniques cannot prevent traffic interruptions if a peer does not support GR or multiple peers fail simultaneously during a GR process. Traffic is interrupted temporarily before the GR process is complete because a GR-enabled router cannot obtain routing information from or establish control plane connections to its peers during the GR process. To tackle this problem, NSR is introduced. NSR is an innovation, compared with NSF. NSR can be used to ensure uninterrupted traffic transmission and retain control plane connections if a software or hardware fault occurs on the control plane of a router. In addition, the fault is transparent to the control planes of its peers.

Related Concepts

- High availability (HA): supports data backup between the AMB and standby main board (SMB).

 Copyright © Huawei Technologies Co., Ltd.

- NSR: allows a standby control plane to take over traffic from an active control plane if the active control plane fails and prevents the control planes of peers from detecting the fault.
- NSF: enables a node to use the GR mechanism to ensure uninterrupted transmission during an AMB/SMB switchover.
- AMB and SMB: run control plane processes.

Implementation

BGP NSR implements synchronization of the following BGP data between the AMB and SMB to support NSR.

- BGP forwarding entries
- BGP control blocks

NOTE

The AMB and SMB must be installed on an NSR-enabled node to support NSR.

The NSR process is as follows:

1. Batch backup

   The AMB backs up BGP data in batches to the SMB immediately after the SMB starts.
2. Real-time backup

   The SMB backs up BGP data in real time and receives packets with the AMB simultaneously.
3. Switchover

   If the AMB fails, the SMB takes over services. The SMB retains uninterrupted operation of both the control and forwarding planes because data is synchronous between the AMB and SMB.

Other Usages

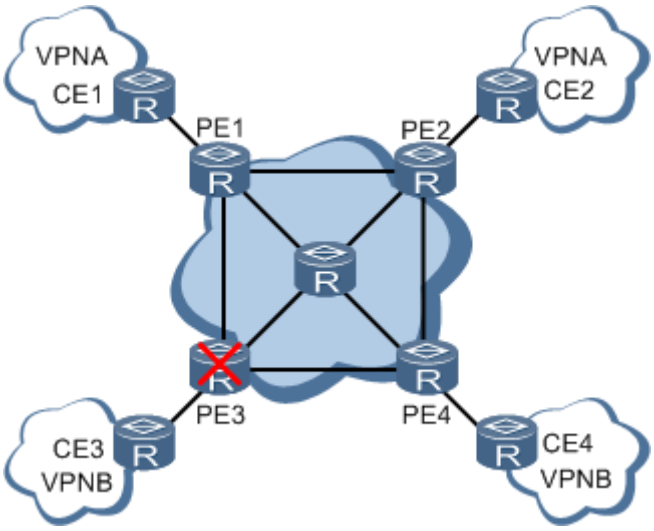
An NSR device can function as a GR Helper. The GR Helper communicates with NSR-disabled devices and responds to peers' GR requests during an AMB/SMB switchover.

Usage Scenario

- NSR is enabled on a node (for example, PE3 shown in **Figure 11-165**) with multiple links among which traffic is load-balanced. NSR helps to prevent traffic interruptions caused by a single point of failure. **Figure 11-165** shows a typical networking for NSR application.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-165 Typical networking for BGP NSR application



- NSR minimizes the impact of control plane faults and prevents route flapping on heavily loaded networks.

Benefits

- NSR ensures the uninterrupted operation of the forwarding plane and allows a standby control plane to retain connections between a local router and its peers if the local BGP control plane fails.
- With NSR, BGP routers can work independently without the assistance of peers. Even though control planes of multiple BGP routers fail, NSR can still ensure the AMB/SMB switchover on each router.

11.12.8.2.12 4-Byte AS Number

Purpose

2-byte autonomous system (AS) numbers used on networks range from 1 to 65535, and the available AS numbers are close to exhaustion as networks expand. Therefore, the AS number range needs to be extended. 4-byte AS numbers ranging from 1 to 4294967295 can address this problem. New speakers that support 4-byte AS numbers can co-exist with old speakers that support only 2-byte AS numbers.

Definition

4-byte AS numbers are extended from 2-byte AS numbers. Border Gateway Protocol (BGP) peers use a new capability code and optional transitive attributes to negotiate the 4-byte AS number capability and transmit 4-byte AS numbers. This mechanism enables communication between new speakers and between old speakers and new speakers.

Open capability code (0x41), defined by RFC 3392, indicates that the local end supports 4-byte capability extension.

The following new optional transitive attributes are defined by RFC 4893 and used to transmit 4-byte AS numbers in old sessions:

 Copyright © Huawei Technologies Co., Ltd.

- **AS4_Path** coded 0x11
- **AS4_Aggregator** coded 0x12

If a new speaker with an AS number greater than 65535 communicates with an old speaker, the old speaker needs to set the peer AS number to **AS_TRANS**. The value of **AS_TRANS** is 23456 and reserved.

Related Concepts

- New speaker: a peer that supports 4-byte AS numbers
- Old speaker: a peer that does not support 4-byte AS numbers
- New session: a BGP connection established between new speakers
- Old session: a BGP connection established between a new speaker and an old speaker, or between old speakers

Principles

BGP speakers negotiate capabilities by exchanging Open messages. **Figure 11-166** shows the format of Open messages exchanged between new speakers. The header of a BGP Open message is fixed, in which **My AS Number** is supposed to be the local AS number. However, **My AS Number** can carry only 2-byte AS numbers. Therefore, a new speaker adds 23456 to **My AS Number** and its local AS number to **Optional parameters** before it sends an Open message to a peer. After the peer receives the message, it can determine whether the new speaker supports 4-byte AS numbers by checking **Optional parameters** in the message.

Figure 11-166 Format of Open messages sent by new speakers

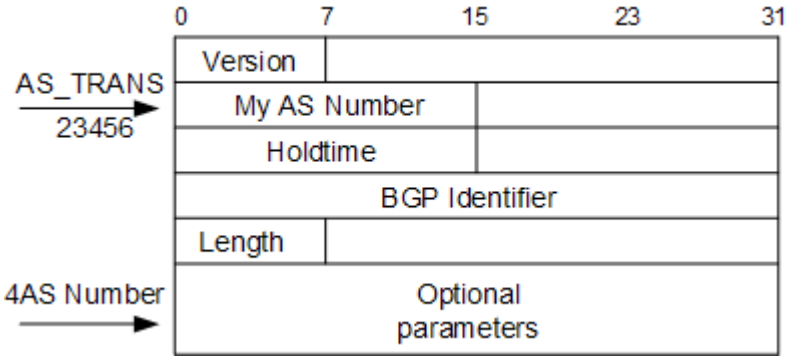
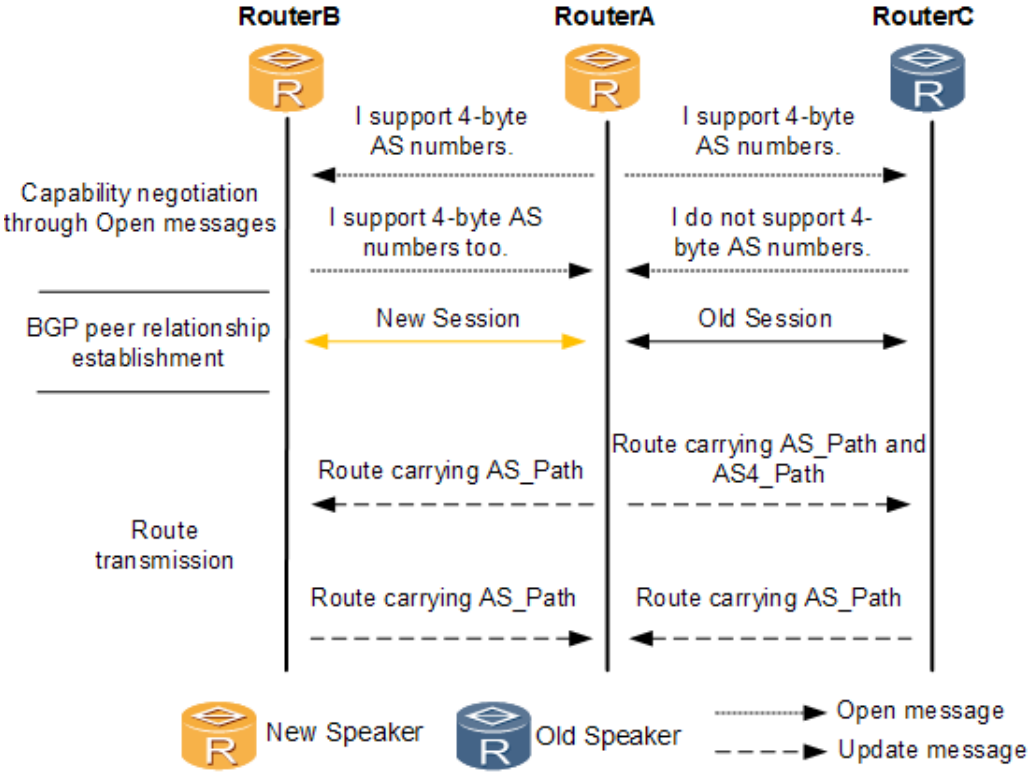


Figure 11-167 shows how peer relationships are established between new speakers, and between an old speaker and a new speaker. BGP speakers notify each other of whether they support 4-byte AS numbers by exchanging Open messages. After the capability negotiation, new sessions are established between new speakers, and old sessions are established between a new speaker and an old speaker.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-167 Process of establishing a BGP peer relationship

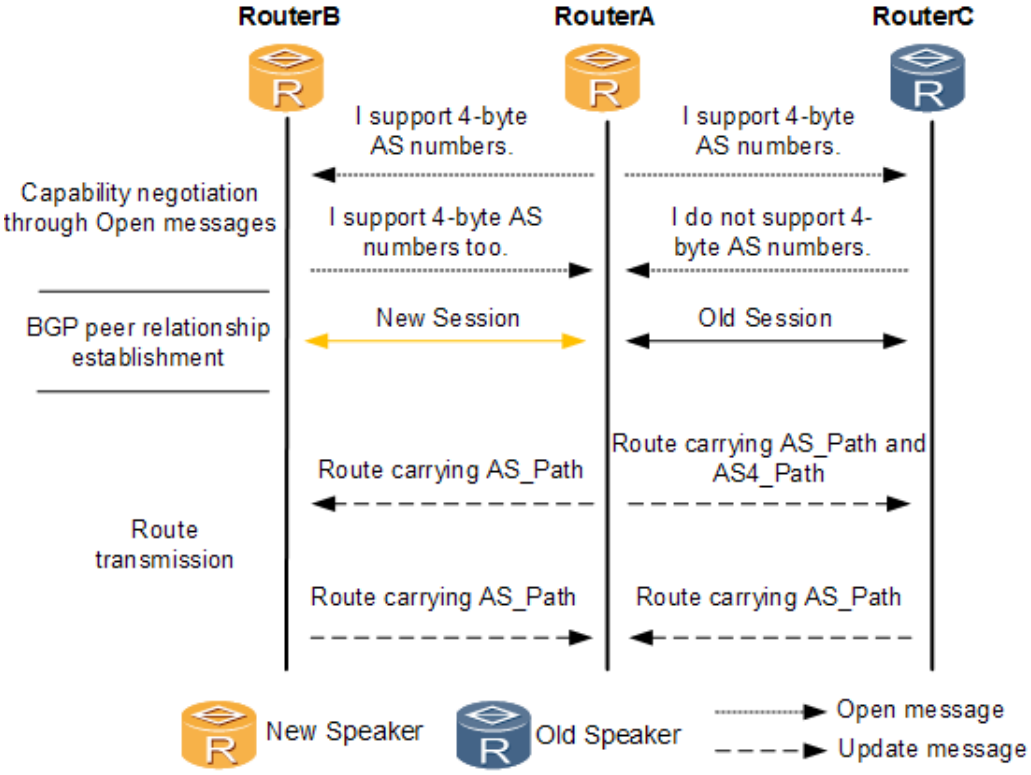


AS_Path and **Aggregator** in Update messages exchanged between new speakers carry 4-byte AS numbers, while **AS_Path** and **Aggregator** in Update messages sent by an old speaker carry 2-byte AS numbers.

- When a new speaker sends an Update message carrying an AS number greater than 65535 to an old speaker, the new speaker uses **AS4_Path** and **AS4_Aggregator** to assist **AS_Path** and **AS_Aggregator** in transferring 4-byte AS numbers. **AS4_Path** and **AS4_Aggregator** are transparent to the old speaker. In the networking shown in **Figure 11-168**, before the new speaker in AS 2.2 sends an Update message to the old speaker in AS 65002, the new speaker replaces each 4-byte AS number (1.1 and 2.2) with 23456 in **AS_Path**. Therefore, the **AS_Path** carried in the Update message is (23456, 23456, 65001), and the carried **AS4_Path** is (1.1, 2.2). After the old speaker in AS 65002 receives the Update message, it transparently transmits the message to other ASs.
- When the new speaker receives an Update message carrying **AS_Path**, **AS4_Path**, **AS_Aggregator**, and **AS4_Aggregator** from the old speaker, the new speaker uses the reconstruction algorithm to reconstruct the actual **AS_Path** and **AS_Aggregator**. In the networking shown in **Figure 11-168**, after the new speaker in AS 65003 receives an Update message carrying **AS_Path** (65002, 23456, 23456, 65001) and **AS4_Path** (1.1, 2.2) from the old speaker in AS 65002, the new speaker reconstructs the actual **AS_Path** (65002, 2.2, 1.1, 65001).

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-168 Process of transmitting a BGP Update message



Format of 4-byte AS numbers

A 4-byte AS number can be an integer or in dotted notation. The system stores 4-byte AS numbers as unsigned integers, regardless of their formats. 4-byte AS numbers in dotted notation are in the format of A.B. The formula of the conversion between 4-byte AS numbers for the two formats is as follows: Integer 4-byte AS number = A x 65536 + B. For example, the 4-byte AS number in dotted notation **2.3** can be converted to the integer 4-byte AS number 131075 (2 x 65536 + 3).

The supports 4-byte AS numbers of both formats. The 4-byte AS numbers displayed in the configuration files are in the format configured by users.

By default, the 4-byte AS numbers displayed in the **display** and **debugging** command outputs are in dotted notation, regardless of the configured format. If users change the default display format of 4-byte AS numbers from dotted notation to an integer, the displayed 4-byte AS numbers are integers.

NOTICE

Adjusting the display format of 4-byte AS numbers affects the matching results of **AS_Path** regular expressions and **extcommunity** filters. If you adjust the display format of 4-byte AS numbers on a system that uses an **AS_Path** regular expression or **extcommunity** filter as the export or import policy, reconfigure the **AS_Path** regular expression or **extcommunity** filter. If you do not reconfigure the **AS_Path** regular expression or **extcommunity** filter, routes cannot match the export or import policy, and a network error may occur.

 Copyright © Huawei Technologies Co., Ltd.

Benefits

4-byte AS numbers alleviate AS number exhaustion and therefore are beneficial to carriers who need to expand the network scale.

11.12.8.2.13 BGP Best-external

Background

If multiple routes to the same destination are available, a Border Gateway Protocol (BGP) device selects one optimal route based on BGP route selection policies and advertises the route to its BGP peers.

However, in scenarios with master and backup provider edges (PEs) or route reflectors (RRs), if routes are selected based on the preceding policies and the primary link fails, the BGP route convergence takes a long time because no backup route is available. To address this problem, the BGP Best-external feature was introduced.

Related Concepts

BGP Best-external: A mechanism that enables a backup device to select a sub-optimal route and send the route to its BGP peers if the route preferentially selected based on BGP route selection policies is an Internal Border Gateway Protocol (IBGP) route advertised by the master device. Therefore, BGP Best-external speeds up BGP route convergence if the primary link fails.

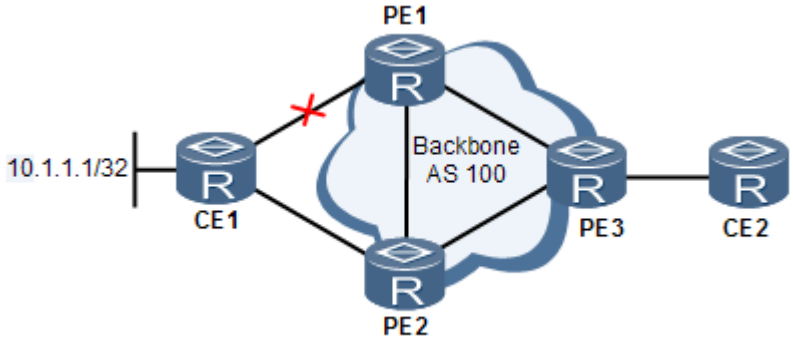
Best-external route: The sub-optimal route selected after BGP Best-external is enabled.

Networking with Master and Backup PEs

In the networking shown in **Figure 11-169**, CE1 is dual-homed to PE1 and PE2. PE1 has a greater Local_Pref value than PE2, and therefore PE1 is the master device while PE2 is the backup device. EBGP peer relationships are established between CE1 and PE1, and between CE1 and PE1. In addition, IBGP peer relationships are established among PE1, PE2, and PE3. PE1 and PE2 receive the same route to 10.1.1.1/32 from CE1. After receiving this route, PE1 advertises it to PE2 and PE3. Therefore, PE2 has two routes to 10.1.1.1/32. Of the two routes, PE2 preferentially selects the route from PE1 because PE1 has a larger Local_Pref value. PE2 does not advertise the selected route to PE3. Therefore, PE3 has only one route to 10.1.1.1/32, which is advertised by PE1. If the link between CE1 and PE1 fails, a new route must be selected to take over traffic after routes are converged, which takes a long time.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-169 Networking with master and backup PEs



BGP Best-external can be enabled on PE2 to address this problem. With BGP Best-external, PE2 selects the EBGP route from CE1 and advertises it to PE3. Therefore, PE3 has two routes to 10.1.1.1/32, in which the route CE1 -> PE2 -> PE3 backs up CE1 -> PE1 -> PE3. **Table 11-80** lists the differences with and without BGP Best-external.

Table 11-80 Differences with and without BGP Best-external

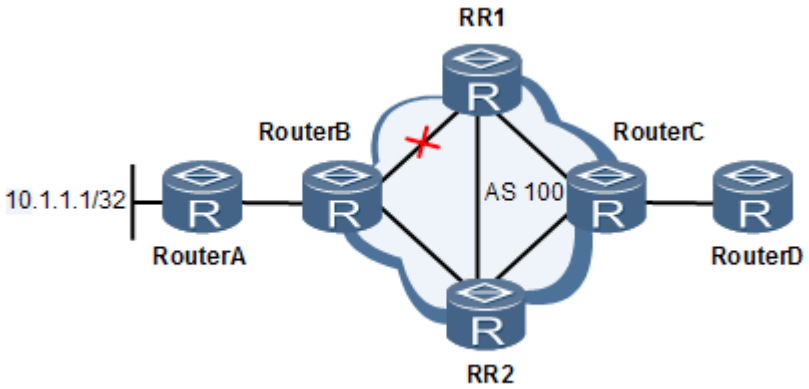
BGP Best-external	Route Available on PE3	Optimal Route	If the Link Between CE1 and PE1 Fails
Not enabled	CE1 -> PE1 -> PE3	CE1 -> PE1 -> PE3	A new route must be selected to take over traffic after routes are converged.
Enabled	CE1 -> PE1 -> PE3 CE1 -> PE2 -> PE3	CE1 -> PE1 -> PE3	Traffic is switched to CE1 -> PE2 -> PE3 immediately.

Networking with Master and Backup RRs

In the networking shown in **Figure 11-170**, an EBGP peer relationship is established between Router A and Router B. An IBGP peer relationship is established between every two devices among RR1, RR2, Router B, and Router C except between Router B and Router C. Router B is a client of RR1 and RR2. RR1 has a greater Local_Pref value than RR2, and therefore RR1 is the master device while RR2 is the backup device. RR1 and RR2 receive the same route to 10.1.1.1/32 from Router B. After receiving this route, RR1 advertises it to RR2 and Router C. Therefore, RR2 has two routes to 10.1.1.1/32. Of the two routes, RR2 preferentially selects the route from RR1 because RR1 has a greater Local_Pref value. RR2 does not advertise the selected route to Router C. Therefore, Router C has only one route to 10.1.1.1/32, which is advertised by RR1. If the link between Router B and RR1 fails, a new route must be selected to take over traffic after routes are converged, which takes a long time.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-170 Networking with master and backup RRs



BGP Best-external can be enabled on RR2 to address this problem. With BGP Best-external, RR2 selects the EBGP route from Router B and advertises it to Router C. Therefore, Router C has two routes to 10.1.1.1/32, in which the route Router A -> Router B -> RR2 -> Router C backs up Router A -> Router B -> RR1 -> Router C. **Table 11-81** lists the differences with and without BGP Best-external.

Table 11-81 Differences with and without BGP Best-external

BGP Best-external	Route Available on Router C	Optimal Route	If the Link Between Router B and RR1 Fails
Not enabled	Router A -> Router B -> RR1 -> Router C	Router A -> Router B -> RR1 -> Router C	A new route must be selected to take over traffic after routes are converged.
Enabled	Router A -> Router B -> RR1 -> Router C Router A -> Router B -> RR2 -> Router C	Router A -> Router B -> RR1 -> Router C	Traffic is switched to Router A -> Router B -> RR2 -> Router C immediately.

Usage Scenario

The BGP Best-external feature applies to scenarios in which master and backup PEs or RRs are deployed and the backup PE or RR needs to advertise the sub-optimal route (Best-external route) to its BGP peers to speed up BGP route convergence.

Advantages

As networks develop, services, such as voice over IP (VoIP), online video, and financial services, pose higher requirements for real-time transmission. With BGP Best-external, the backup device selects the sub-optimal route and advertises the

 Copyright © Huawei Technologies Co., Ltd.

route to its BGP peers, which speeds up BGP route convergence and minimizes service interruptions.

11.12.8.2.14 BGP ADD-PATH

Background

In a scenario with a route reflector (RR) and clients, if the RR has multiple routes to the same destination (with the same prefix), the RR selects an optimal route from these routes and then sends only the optimal route to its clients. Therefore, the clients have only one route to the destination. If a link along this route fails, route convergence takes a long time, which cannot meet the requirements on high reliability.

To address this issue, deploy the BGP ADD-PATH feature on the RR. With BGP ADD-PATH, the RR can send two or more routes with the same prefix to its clients. After reaching the clients, these routes can back up each other or load-balance traffic, which ensures high reliability in data transmission.

NOTE

- BGP ADD-PATH is deployed on RRs in most cases although it can be configured on any router.
- With BGP ADD-PATH, you can configure the maximum number of routes with the same prefix that an RR can send to its clients. The actual number of routes with the same prefix that an RR can send to its clients is the smaller value between the configured maximum number and the number of available routes with the same prefix.

Related Concepts

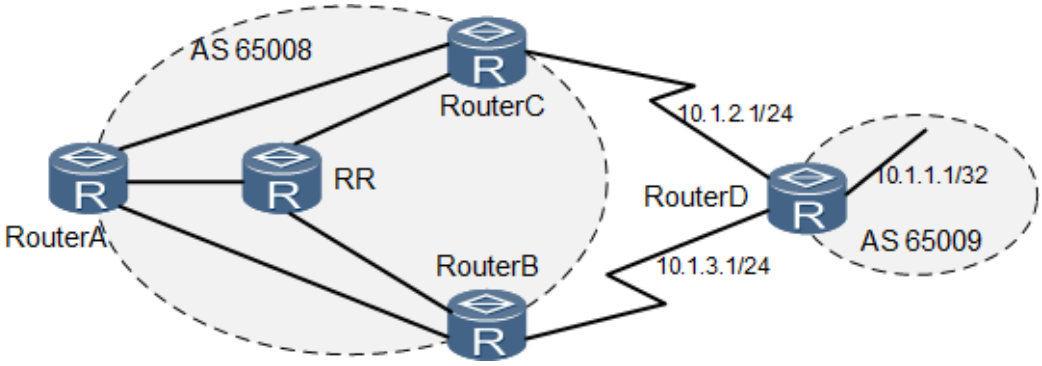
ADD-PATH route: The routes selected by BGP after BGP ADD-PATH is configured.

Typical Networking

On the network shown in **Figure 11-171**, Router A, Router B, and Router C are clients of the RR, and Router D is an EBGP peer of Router B and Router C.

Each of Router B and Router C receives a route to 10.1.1.1/32 from Router D, with 10.1.2.1/24 and 10.1.3.1/24 as the next hops, respectively. Then, each of Router B and Router C sends the received route to the RR. After receiving the two routes, the RR selects an optimal route from them and sends it to Router A. Therefore, Router A has only one route to 10.1.1.1/32.

Figure 11-171 Networking with BGP ADD-PATH



 Copyright © Huawei Technologies Co., Ltd.

BGP ADD-PATH can be configured to allow the RR to send more than one route with the same prefix to Router A. Suppose that the configured maximum number of routes with the same prefix that the RR can send to Router A is 2 and that the optimal route selected by the RR is the one from Router B. **Table 11-82** lists the differences with and without BGP ADD-PATH.

Table 11-82 Differences with and without BGP ADD-PATH

BGP ADD-PATH	Route Available on Router A	If the Link Between Router D and Router B Fails
Not enabled	Only one route to 10.1.1.1/32, with 10.1.2.1/24 as the next hop	A new route must be selected to take over traffic after route convergence.
Enabled	Two routes to 10.1.1.1/32, with 10.1.2.1/24 and 10.1.3.1/24 as the next hops	If the two routes back up each other, the link between Router D and Router C takes over traffic. If the two routes load-balance traffic, the link between Router D and Router C takes over all the traffic that ran along the two routes.

Usage Scenario

The BGP ADD-PATH feature applies to scenarios in which an RR and clients are deployed and the RR needs to send more than one route with the same prefix to its clients to ensure high reliability in data transmission.

Benefits

Deploying BGP ADD-PATH can improve network reliability.

11.12.8.2.15 IP FRR

Definition

IP fast reroute (IP FRR) provides fast traffic recovery upon link-layer or physical-layer faults. When a fault is detected and reported to the upper-layer routing system, a backup link is used to forward the impacted traffic.

IP FRR recovers the traffic impacted by link faults.

 Copyright © Huawei Technologies Co., Ltd.

Purpose

If a forwarding link fails on a traditional IP network, the device port goes down. After detecting such a fault, the device notifies the upper-layer routing system so that it updates the link status and recalculates a route. The process takes several seconds.

The convergence time cannot meet the requirements of delay-sensitive services, such as voice over IP (VoIP) and online video services. In contrast, IP FRR can recover impacted traffic in the level of 50 ms.

Principle

IP FRR supports static routes and dynamic routes, such as OSPF, OSPFv3, IS-IS, IS-IS IPv6, BGP, and BGP4+ routes.

IP FRR is implemented as follows:

1. Run commands to configure and enable IP FRR.

2. After IP FRR is enabled, the MA5800-X17 selects primary routes using routing protocols, calculates the backup links in advance, and saves them in the forwarding table.

3. If a primary link is unavailable, the MA5800-X17 forwards traffic through the backup link.

NOTE

For detailed configuration of IP FRR for each routing protocol, see the corresponding protocol description.

IP FRR Triggering Mode

IP FRR can be triggered by a BFD session or port status.

IP FRR triggered by a BFD session

If a Bidirectional Forwarding Detection (BFD) session is configured, it automatically associates with IP FRR. If a BFD session detects a link failure, the BFD session goes down and triggers IP FRR.

If multiple BFD sessions are configured and they have identical outgoing interfaces and destination IP addresses, IP FRR can be triggered if any of the BFD sessions goes down.

IP FRR triggered by port status

NOTE

This function is supported in V100R019C1X and later versions.

You can run the **port frr-binding** command to enable the function of triggering IP FRR by port status. When a port goes down, IP FRR is triggered.

The following restrictions apply when enabling the function of triggering IP FRR by port status:

- Virtual service ports do not support IP FRR. Only standard VLAN ports support IP FRR.

 Copyright © Huawei Technologies Co., Ltd.

- To enable the function of triggering IP FRR by port status in a VLAN, the VLAN can have only one standard port, which can be a common Ethernet port or a port aggregation group.
- Protection groups do not support IP FRR.

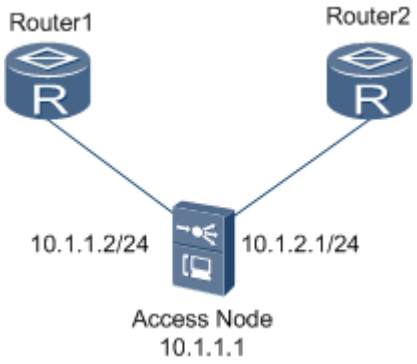
11.12.8.3 Configuration Example of BGP

This topic provides an example for configuring the BGP on the device.

Service Requirements

In this example network, an IBGP connection is set up between access node and routers.

Figure 11-172 Example network for configuring the BGP



Data Plan

Table 11-83 provides the data plan for configuring the BGP.

Table 11-83 Data plan for configuring the BGP

Item	Data	Remarks
Access node	IP address of VLAN interface 6: 10.1.1.2/24	It is used for the IBGP connection to the Router1.
	IP address of VLAN interface 2: 10.1.2.1/24	It is used for the IBGP connection to the Router12.
	Router ID: 10.1.1.1	-
	AS number: 2000	-

Procedure

Step 1 Configure access node.

 Copyright © Huawei Technologies Co., Ltd.

1. Configure the IP address of the Layer 3 interface.

```
huawei(config)#vlan 6 smart
huawei(config)#port vlan 6 0/9 0
huawei(config)#interface vlanif 6
huawei(config-if-vlanif6)#ip address 10.1.1.2 24
huawei(config-if-vlanif6)#quit
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 1
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.2.1 24
huawei(config-if-vlanif2)#quit
```

2. Enable the BGP function.

```
huawei(config)#bgp 2000
huawei(config-bgp)#router-id 10.1.1.1
huawei(config-bgp)#peer 10.1.3.3 as-number 2000
huawei(config-bgp)#peer 10.1.2.2 as-number 2000
huawei(config-bgp)#quit
```

3. Save the data.

```
huawei(config)#save
```

----End

Result

- Run the **display bgp peer** command, and you can query the configuration.

Configuration File

```
vlan 6 smart
port vlan 6 0/9 0
interface vlanif 6
ip address 10.1.1.2 24
quit
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ip address 10.1.2.1 24
quit
bgp 2000
router-id 10.1.1.1
peer 10.1.3.3 as-number 2000
peer 10.1.2.2 as-number 2000
quit
save
```

11.12.8.4 Configuration Example of BGP4+

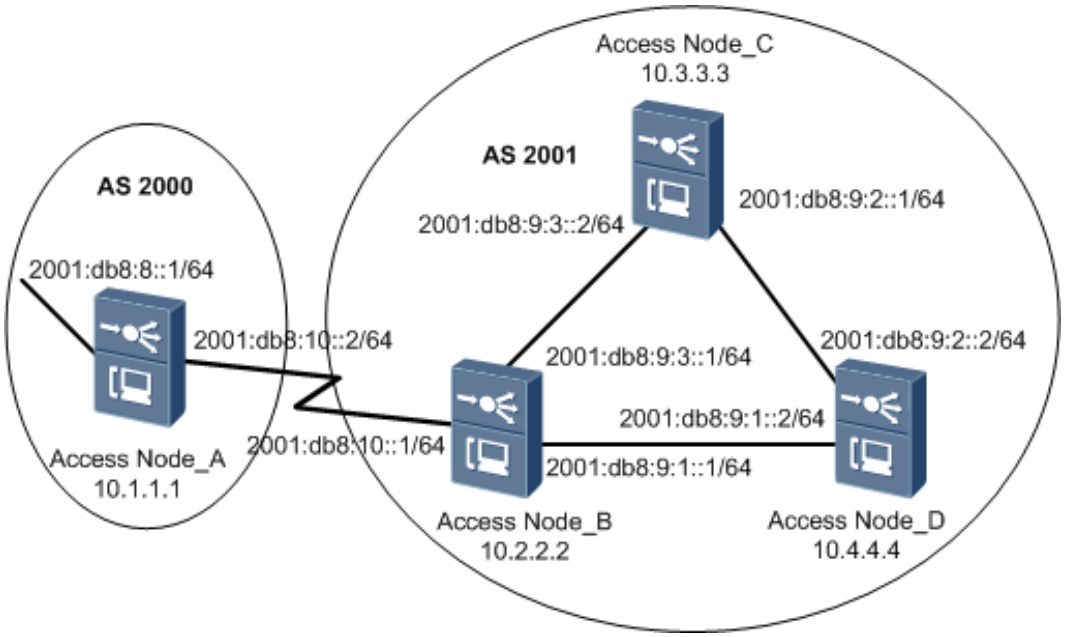
This topic provides an example for configuring the BGP4+ on the device.

Service Requirements

In this example network, an external Border Gateway Protocol (EBGP) connection is set up between Access node_A and Access node_B, and an Interior Border Gateway Protocol (IBGP) connection is set up among Access node_B, Access node_C, and Access node_D.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-173 Example network for configuring the BGP4+



Data Plan

Table 11-84 provides the data plan for configuring the BGP4+.

Table 11-84 Data plan for configuring the BGP4+

Item	Data	Remarks
Access node_A	IPv6 address of virtual local area network (VLAN) interface 6: 2001:db8:10::2/64	It is used for the EBGP connection to Autonomous System (AS) 2001.
	IPv6 address of VLAN interface 2: 2001:db8:8::1/64	-
	Router ID: 10.1.1.1	-
	AS number: 2000	-
Access node_B	IPv6 address of VLAN interface 6: 2001:db8:10::1/64	It is used for the EBGP connection to AS 2000.
	IPv6 address of VLAN interface 3: 2001:db8:9:3::1/64	It is used for the IBGP connection to the Access node_C.
	IPv6 address of VLAN interface 4: 2001:db8:9:1::1/64	It is used for the IBGP connection to the Access node_D.
	Router ID: 10.2.2.2	-

 Copyright © Huawei Technologies Co., Ltd.

Item	Data	Remarks
	AS number: 2001	-
Access node_C	IPv6 address of VLAN interface 3: 2001:db8:9:3::2/64	It is used for the IBGP connection to the Access node_B.
	IPv6 address of VLAN interface 4: 2001:db8:9:2::1/64	It is used for the IBGP connection to the Access node_D.
	Router ID: 10.3.3.3	-
	AS number: 2001	-
Access node_D	IPv6 address of VLAN interface 5: 2001:db8:9:2::2/64	It is used for the IBGP connection to the Access node_C.
	IPv6 address of VLAN interface 4: 2001:db8:9:1::2/64	It is used for the IBGP connection to the Access node_B.
	Router ID: 10.4.4.4	-
	AS number: 2001	-

Procedure

Step 1 Configure Access node_A.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#ipv6
huawei(config)#vlan 6 smart
huawei(config)#port vlan 6 0/9 0
huawei(config)#interface vlanif 6
huawei(config-if-vlanif6)#ipv6 enable
huawei(config-if-vlanif6)#ipv6 address 2001:db8:10::2 64
huawei(config-if-vlanif6)#quit
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 1
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ipv6 enable
huawei(config-if-vlanif2)#ipv6 address 2001:db8:8::1/64
huawei(config-if-vlanif2)#quit
```

2. Enable the Border Gateway Protocol (BGP) function and configure the EBGP neighbor between Access node_B.

```
huawei(config)#bgp 2000
huawei(config-bgp)#router-id 10.1.1.1
huawei(config-bgp)#peer 2001:db8:10::1 as-number 2001
huawei(config-bgp)#ipv6-family unicast
huawei(config-bgp-af-ipv6)#peer 2001:db8:10::1 enable
huawei(config-bgp-af-ipv6)#network 2001:db8:10:: 64
huawei(config-bgp-af-ipv6)#network 2001:db8:8:: 64
huawei(config-bgp-af-ipv6)#quit
huawei(config-bgp)#quit
```

3. Save the data.

```
huawei(config)#save
```

 Copyright © Huawei Technologies Co., Ltd.

Step 2 Configure Access node_B.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#ipv6
huawei(config)#vlan 6 smart
huawei(config)#port vlan 6 0/9 0
huawei(config)#interface vlanif 6
huawei(config-if-vlanif6)#ipv6 enable
huawei(config-if-vlanif6)#ipv6 address 2001:db8:10::1 64
huawei(config-if-vlanif6)#quit
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 0
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ipv6 enable
huawei(config-if-vlanif3)#ipv6 address 2001:db8:9:3::1 64
huawei(config-if-vlanif3)#quit
huawei(config)#vlan 4 smart
huawei(config)#port vlan 4 0/9 0
huawei(config)#interface vlanif 4
huawei(config-if-vlanif4)#ipv6 enable
huawei(config-if-vlanif4)#ipv6 address 2001:db8:9:1::1 64
huawei(config-if-vlanif4)#quit
```

2. Enable the BGP function. Configure the EBGP neighbor between Access node_B and Access node_A, and the IBGP neighbor between Access node_B, Access node_C, and Access node_D.

```
huawei(config)#bgp 2001
huawei(config-bgp)#router-id 10.2.2.2
huawei(config-bgp)#peer 2001:db8:10::2 as-number 2000
huawei(config-bgp)#peer 2001:db8:9:3::2 as-number 2001
huawei(config-bgp)#peer 2001:db8:9:1::2 as-number 2001
huawei(config-bgp)#ipv6-family unicast
huawei(config-bgp-af-ipv6)#peer 2001:db8:10::2 enable
huawei(config-bgp-af-ipv6)#peer 2001:db8:9:3::2 enable
huawei(config-bgp-af-ipv6)#peer 2001:db8:9:1::2 enable
huawei(config-bgp-af-ipv6)#import-route direct
huawei(config-bgp-af-ipv6)#quit
huawei(config-bgp)#quit
```

3. Save the data.

```
huawei(config)#save
```

Step 3 Configure Access node_C.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#ipv6
huawei(config)#vlan 3 smart
huawei(config)#port vlan 3 0/9 0
huawei(config)#interface vlanif 3
huawei(config-if-vlanif3)#ipv6 enable
huawei(config-if-vlanif3)#ipv6 address 2001:db8:9:3::2 64
huawei(config-if-vlanif3)#quit
huawei(config)#vlan 5 smart
huawei(config)#port vlan 5 0/9 0
huawei(config)#interface vlanif 5
huawei(config-if-vlanif5)#ipv6 enable
huawei(config-if-vlanif5)#ipv6 address 2001:db8:9:2::1 64
huawei(config-if-vlanif5)#quit
```

2. Enable the BGP function. Configure the IBGP neighbor between Access node_B and Access node_D.

```
huawei(config)#bgp 2001
huawei(config-bgp)#router-id 10.3.3.3
huawei(config-bgp)#peer 2001:db8:9:3::1 as-number 2001
huawei(config-bgp)#peer 2001:db8:9:2::2 as-number 2001
huawei(config-bgp)#ipv6-family unicast
huawei(config-bgp-af-ipv6)#peer 2001:db8:9:3::1 enable
huawei(config-bgp-af-ipv6)#peer 2001:db8:9:2::2 enable
```

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config-bgp-af-ipv6)#import-route direct
huawei(config-bgp-af-ipv6)#quit
huawei(config-bgp)#quit
```

3. Save the data.

```
huawei(config)#save
```

Step 4 Configure Access node_D.

1. Configure the IPv6 address of the Layer 3 interface.

```
huawei(config)#ipv6
huawei(config)#vlan 4 smart
huawei(config)#port vlan 4 0/9 0
huawei(config)#interface vlanif 4
huawei(config-if-vlanif4)#ipv6 enable
huawei(config-if-vlanif4)#ipv6 address 2001:db8:9:1::2 64
huawei(config-if-vlanif4)#quit
huawei(config)#vlan 5 smart
huawei(config)#port vlan 5 0/9 0
huawei(config)#interface vlanif 5
huawei(config-if-vlanif5)#ipv6 enable
huawei(config-if-vlanif5)#ipv6 address 2001:db8:9:2::2 64
huawei(config-if-vlanif5)#quit
```

2. Enable the BGP function. Configure the IBGP neighbor between Access node_B and Access node_C.

```
huawei(config)#bgp 2001
huawei(config-bgp)#router-id 10.4.4.4
huawei(config-bgp)#peer 2001:db8:9:1::2 as-number 2001
huawei(config-bgp)#peer 2001:db8:9:2::1 as-number 2001
huawei(config-bgp)#ipv6-family unicast
huawei(config-bgp-af-ipv6)#peer 2001:db8:9:1::2 enable
huawei(config-bgp-af-ipv6)#peer 2001:db8:9:2::1 enable
huawei(config-bgp-af-ipv6)#import-route direct
huawei(config-bgp-af-ipv6)#quit
huawei(config-bgp)#quit
```

3. Save the data.

```
huawei(config)#save
```

----End

Result

- Run the **display bgp peer** command, and you can see that:
  - The EBGP connection is set up between Access node_A and Access node_B.
  - The IBGP connections are set up among Access node_B, Access node_C, and Access node_D.
  - The route with the destination subnet 2001:db8:8::/64 exists on Access node_C and Access node_D, and the next hop of the route is the interface address of Access node_A
- Run the **ping ipv6** command on Access node_C and Access node_D to ping the Layer 3 interface (2001:db8:8::1/64) on Access node_A. The **ping ipv6** command is executed successfully.

Configuration File

Configuration on each Access node is similar. Take Access Node_A for example.

```
ipv6
vlan 6 smart
```

 Copyright © Huawei Technologies Co., Ltd.

```
port vlan 6 0/9 0
interface vlanif 6
ipv6 enable
ipv6 address 2001:db8:10::2 64
quit
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ipv6 enable
ipv6 address 2001:db8:8::1 64
quit
bgp 2000
router-id 10.1.1.1
peer 2001:db8:10::1 as-number 2001
ipv6-family unicast
peer 2001:db8:10::1 enable
network 2001:db8:10:: 64
network 2001:db8:8:: 64
quit
quit
```

11.12.8.5 BGP Reference Standards and Protocols

The following table lists the references.

Table 11-85 References

Document No.	Document Name
RFC 827	Exterior Gateway Protocol (EGP)
RFC 1213	Management Information Base for Network Management of TCP/IP-based internets: MIB-II
RFC 1397	Default Route Advertisement in BGP2 and BGP3 Version of the Border Gateway Protocol
RFC 1403	BGP OSPF Interaction
RFC 1519	Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy
RFC 1654	A Border Gateway Protocol 4 (BGP-4)
RFC 1655	Application of the Border Gateway Protocol in the Internet
RFC 1656	BGP-4 Protocol Document Roadmap and Implementation Experience
RFC 1657	basic BGP4 MIB
RFC 1771	A Border Gateway Protocol 4 (BGP-4)
RFC 1772	BGP basic functions support
RFC 1773	obsoletes RFC 1656
RFC 1774	BGP-4 Protocol Analysis
RFC 1930	Guidelines for creation, selection, and registration of an Autonomous System (AS)

 Copyright © Huawei Technologies Co., Ltd.

Document No.	Document Name
RFC 1965	Autonomous System Confederations for BGP
RFC 1966	BGP Route-Reflection
RFC 1997	BGP Communities Attribute
RFC 1998	An Application of the BGP Community Attribute
RFC 2270	Using a Dedicated AS for Sites Homed to a Single Provider
RFC 2283	Multiprotocol Extensions for BGP-4
RFC 2385	TCP MD5
RFC 2439	BGP Route Flap Damping
RFC 2519	A Framework for Inter-Domain Route Aggregation
RFC 2545	BGP supporting IPV6
RFC 2796	BGP Route Reflection
RFC 2842	Capabilities Advertisement with BGP-4
RFC 2858	Multiprotocol Extensions for BGP-4
RFC 2918	Route Refresh Capability for BGP-4
RFC 3065	Autonomous System Confederations for BGP
RFC 3107	Support BGP carry Label for MPLS
RFC 3232	Assigned Numbers: RFC 1700 is Replaced by an On-line Database
RFC 3392	Capabilities Advertisement with BGP-4
RFC 3562	Key Management Considerations for the TCP MD5 Signature Option
RFC 4271	A Border Gateway Protocol 4 (BGP-4)
RFC 4272	BGP Security Vulnerabilities Analysis
RFC 4273	Definitions of Managed Objects for the Fourth Version of Border Gateway
RFC 4274	BGP-4 Protocol Analysis
RFC 4276	BGP 4 Implementation Report
RFC 4277	Experience with the BGP-4 Protocol
RFC 4360	BGP Extended Communities Attribute
RFC 4456	BGP Route Reflection
RFC 4486	Subcodes for BGP Cease Notification Message

 Copyright © Huawei Technologies Co., Ltd.

Document No.	Document Name
RFC 4632	Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan
RFC 4724	Graceful Restart Mechanism for BGP
RFC 4760	Multiprotocol Extensions for BGP-4
RFC 4781	Graceful Restart Mechanism for BGP with MPLS
RFC 4893	BGP Support for Four-octet AS Number Space
RFC 5004	Avoid BGP Best Path Transitions from One External to Another
RFC 5065	Autonomous System Confederations for BGP
RFC 5291	Outbound Route Filtering Capability for BGP-4
RFC 5292	Address-Prefix-Based Outbound Route Filter for BGP-4
RFC 5396	Textual Representation of Autonomous System (AS) Numbers
RFC 5398	Autonomous System (AS) Number Reservation for Documentation Use
RFC 5492	Capabilities Advertisement with BGP-4
draft-ietf-idr-as4bytes-13	BGP Support for Four-octet AS Number Space
draft-ietf-idr-bgp-ext-communities-05	Extended Community Attribute
draft-ietf-idr-as-pathlimit-03	The AS_PATHLIMIT Path Attribute
draft-ietf-idr-aspath-orf-09	Aspath Based Outbound Route Filter for BGP-4
Draft-marques-l3vpn-ibgp-01	Internal BGP as PE-CE protocol draft-marques-l3vpn-ibgp-01
draft-ramachandra-bgp-ext-communities-04	Address Prefix Based Outbound Route Filter for BGP-4
draft-kato-bgp-ipv6-link-local-00	BGP4+ Peering Using IPv6 Link-local Address
draft-ietf-idr-cap-neg-01	Capabilities Negotiation with BGP4
draft-ietf-idr-bgp-prefix-orf-04	Address Prefix Based Outbound Route Filter for BGP-4

 Copyright © Huawei Technologies Co., Ltd.

Document No.	Document Name
draft-ietf-idr-rfc3065bis-06	Autonomous System Confederations for BGP

11.12.9 Routing Policy

11.12.9.1 Introduction to Routing Policies

Definition

Routing policies are used to filter routes and set attributes for routes. Changing route attributes (including reachability) changes the path that network traffic passes through.

Purpose

When advertising, receiving, and importing routes, the Router implements certain policies based on actual networking requirements to filter routes and change the attributes of the routes. Routing policies serve the following purposes:

- Control route advertising

  Only routes that match the rules specified in a policy are advertised.
- Control route receiving

  Only the required and valid routes are received. This reduces the size of the routing table and improves network security.
- Filter and control imported routes

  A routing protocol may import routes discovered by other routing protocols. Only routes that satisfy certain conditions are imported to meet the requirements of the protocol.
- Modify attributes of specified routes

  Attributes of the routes that are filtered by a routing policy are modified to meet the requirements of the local device.

Benefits

This feature brings the following benefits:

- Controls the size of the routing table, saving system resources.
- Controls route receiving and advertising, improving network security.
- Modifies attributes of routes for proper traffic planning, improving network performance.

 Copyright © Huawei Technologies Co., Ltd.

11.12.9.2 Principles

Implementation

Routing policies are implemented using the following procedures:

- Define rules: Define features of routes to which routing policies are applied. Users define a set of matching rules based on different attributes of routes, such as the destination address and the address of the router that advertises the routes.
- Implement the rules: Apply the matching rules to routing policies for advertising, receiving, and importing routes.

Filter

Filters are the core of a routing policy and is defined using a set of matching rules. The provides several filters for routing policies.

Table 11-86 lists applicable scopes and matching rules of different filters.

Table 11-86 Comparisons between filters

Filter	Applicable Scope	Matching Rules
Access control list (ACL)	Dynamic routing protocols	Inbound interface, source or destination IP address, protocol type, and source or destination port number
IP prefix list	Dynamic routing protocols	Source and destination IP addresses and next hop address
Route-Policy	Dynamic routing protocols	Destination IP address, next hop address, cost, interface information, route type, ACL, IP prefix list, AS_Path filter, community filter, extcommunity filter, and RD filter.

The ACL and IP prefix list can be used only to filter routes but not modify attributes of the filtered routes. A Route-Policy is a comprehensive filter, and it can use the matching rules of the ACL and IP prefix list to filter routes. In addition, attributes of the filtered routes can be modified using the Route-Policy. The following section describes the filters in more detail.

ACL

An ACL is a set of sequential filtering rules. Users can define rules based on packet information, such as inbound interfaces, source or destination IP addresses, protocol types, or source or destination port numbers and specify an action to deny or permit packets. After an ACL is configured, the system classifies received packets based on the rules defined in the ACL and denies or permits the packets accordingly.

 Copyright © Huawei Technologies Co., Ltd.

An ACL only classifies packets based on defined rules and can be used to filter packets only when it is applied to a routing policy.

ACLs can be configured for both IPv4 packets and IPv6 packets. Based on the usage, ACLs are classified into three types: interface-based ACLs, basic ACLs, and advanced ACLs. Users can specify the IP address and subnet address range in an ACL to match the source IP address, destination network segment address, or the next hop address of a route.

ACLs can be configured on access or core devices to:

- Protect the devices against IP, TCP, and Internet Control Message Protocol (ICMP) packet attacks.
- Control network access. For example, ACLs can be used to control the access of enterprise network users to external networks, the specific network resources that users can access, and the period for which users can access networks.
- Limit network traffic and improve network performance. For example, ACLs can be used to limit bandwidth for upstream and downstream traffic, charge for the bandwidth that users have applied for, and fully use high-bandwidth network resources.

IP Prefix List

An IP prefix list contains a group of route filtering rules. Users can specify the prefix and mask length range to match the destination network segment address or the next hop address of a route. An IP prefix list is used to filter routes that are advertised and received by various dynamic routing protocols.

An IP prefix list is easier and more flexible than an ACL. However, if a large number of routes with different prefixes need to be filtered, configuring an IP prefix list to filter the routes is complex.

IP prefix lists can be configured for both IPv4 routes and IPv6 routes, and they share the same implementation process. An IP prefix list filters routes based on the mask length or mask length range.

- Mask length: An IP prefix list filters routes based on IP address prefixes. An IP address prefix is defined by an IP address and the mask length. For example, in a route to 10.1.1.1/16, the mask length is 16 bits, and the valid prefix is 16 bits (10.1.0.0).
- Mask length range: Routes with the IP address prefix and mask length within the range defined in the IP prefix list meet the matching rules.

NOTE

0.0.0.0 is a wildcard address. If the IP prefix is 0.0.0.0, users must specify either a mask or a mask length range, with the following results:

- If a mask is specified, all routes with the mask are permitted or denied as required.
- If a mask length range is specified, all routes with the mask length in the range are permitted or denied as required.

 Copyright © Huawei Technologies Co., Ltd.

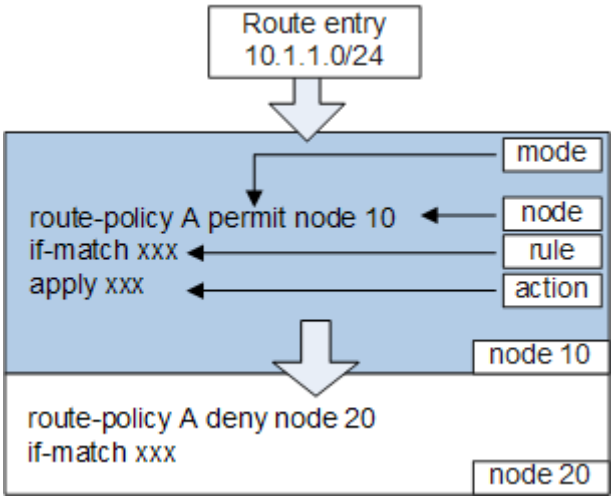
Route-Policy

A Route-Policy is a complex filter. It is used to match attributes of specified routes and change route attributes when specific conditions are met. A Route-Policy can use the preceding six filters to define its matching rules.

- **Composition of a Route-Policy**

  As shown in **Figure 11-174**, a Route-Policy consists of node IDs, matching mode, **if-match** clauses, and **apply** clauses.

  Figure 11-174 Composition of a Route-Policy



  - Node ID

    A Route-Policy consists of one or more nodes. Node IDs are specified as indexes in the IP prefix list. In a Route-Policy, routes are filtered based on the following rules:

    - Sequential matching: The system checks entries based on node IDs in ascending order. Therefore, specifying the node IDs in the required sequence is recommended.
    - One-time matching: The relationship between the nodes of a Route-Policy is "OR". If a route matches one node, the route matches the Route-Policy and will not be matched against the next node.

  - Matching mode

    Either of the following matching modes can be used:

    - **permit**: specifies the **permit** mode of a node. If a route matches the **if-match** clauses of a node, all the actions defined by **apply** clauses are performed, and the matching is complete. If a route does not match the **if-match** clauses of the node, the route continues to match the next node.
    - **deny**: specifies the **deny** mode of a node. In the **deny** mode, the **apply** clauses are not used. If a route matches all the **if-match** clauses of the node, the route is denied by the node and the next

 Copyright © Huawei Technologies Co., Ltd.

node is not matched. If the entry does not match all the **if-match** clauses, the next node is matched.

NOTE

To allow other routes to pass through, a Route-Policy that contains no **if-match** or **apply** clause in the **permit** mode needs to be configured for a node next to multiple nodes that are in the **deny** mode.

- **if-match** clause

  The **if-match** clause defines the matching rules.

  Each node of a Route-Policy can comprise multiple **if-match** clauses or no **if-match** clause at all. If no **if-match** clause is configured for a node in the **permit** mode, all routes match the node.

- **apply** clause

  The **apply** clauses specify actions. When a route matches a Route-Policy, the system sets some attributes for the route based on the **apply** clause.

  Each node of a Route-Policy can comprise multiple **apply** clauses or no **apply** clause at all. The **apply** clause is not used when routes need to be filtered but attributes of the routes do not need to be set.

- **Matching results of a Route-Policy**

  The matching results of a Route-Policy are obtained based on the following aspects:

  - Matching mode of the node, either **permit** or **deny**
  - Matching rules (either **permit** or **deny**) contained in the **if-match** clause (such as ACLs or IP prefix lists)

  The matching results are listed in **Table 11-87**.

Table 11-87 Matching results of a Route-Policy

Rule (Matching Rule Contained in if-match Clauses)	Mode (Matching Mode of a Node)	Matching Result
permit	**permit**	• Routes matching the **if-match** clauses of the node match the Route-Policy, and the matching is complete. • Routes not matching the **if-match** clauses of the node continue to match the next node of the Route-Policy.

 Copyright © Huawei Technologies Co., Ltd.

Rule (Matching Rule Contained in if-match Clauses)	Mode (Matching Mode of a Node)	Matching Result
	deny	• Routes matching the **if-match** clauses of the node are denied by the Route-Policy, and the matching is complete. • Routes not matching the **if-match** clauses of the node continue to match the next node of the Route-Policy.
deny	**permit**	• Routes matching the **if-match** clauses of the node are denied by the Route-Policy and continue to match the next node. • Routes not matching the **if-match** clauses of the node continue to match the next node of the Route-Policy.
	deny	• Routes matching the **if-match** clauses of the node are denied by the Route-Policy and continue to match the next node. • Routes not matching the **if-match** clauses of the node continue to match the next node of the Route-Policy. **NOTE** If all **if-match** clauses and nodes of the Route-Policy are in the **deny** mode, all the routes to be filtered are denied by the Route-Policy.

NOTE

All unmatched routes are denied by the Route-Policy by default. If more than one node is defined in a Route-Policy, at least one of them must be in the **permit** mode. The reason is as follows:

- If a route fails to match any of the nodes, the route is denied by the Route-Policy.
- If all the nodes in the Route-Policy are set in the **deny** mode, all the routes to be filtered are denied by the Route-Policy.

 Copyright © Huawei Technologies Co., Ltd.

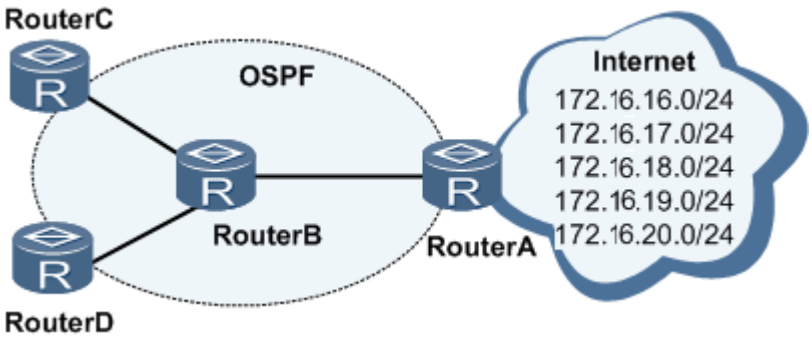
11.12.9.3 Applications

Specific Routes Filtering

On the OSPF-enabled network shown in **Figure 11-175**, Router A receives routes from the Internet and advertises some of the routes to Router B.

- Router A advertises only routes 172.16.17.0/24, 172.16.18.0/24, and 172.16.19.0/24 to Router B.
- Router C accepts only the route 172.16.18.0/24.
- Router D accepts all the routes advertised by Router B.

Figure 11-175 Networking diagram for filtering received and advertised routes



There are multiple approaches to meet the preceding requirements, and the following two approaches are used in this example:

- Use IP prefix lists.
  - Configure an IP prefix list for Router A and configure the IP prefix list as an export policy of Router A to be used by OSPF.
  - Configure another IP prefix list for Router C and configure the IP prefix list as an import policy of Router C to be used by OSPF.
- Use route-policies
  - Configure a Route-Policy (the matching rules can be the IP prefix list, cost, or route tag) for Router A and configure the Route-Policy as an export policy of Router A to be used by OSPF.
  - Configure another Route-Policy for Router C and configure the Route-Policy as an import policy of Router C to be used by OSPF.

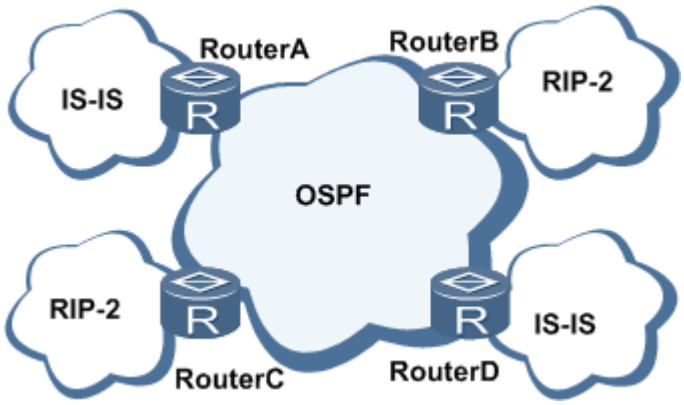
  Compared with an IP prefix list, a Route-Policy allows route attributes to be modified and can be used to control routes more flexibly, but it is more complex to configure.

Transparent Transmission of Routes of Other Protocols Through an OSPF AS

On the network shown in **Figure 11-176**, an AS runs OSPF and functions as a transit AS for other areas. Routes from the IS-IS area connected to Router A need to be transparently transmitted through the OSPF AS to the IS-IS area connected to Router D. Routes from the RIP-2 area connected to Router B need to be transparently transmitted through the OSPF AS to the RIP-2 area connected to Router C.

 Copyright © Huawei Technologies Co., Ltd.

Figure 11-176 Networking diagram for transparently transmitting routes of other protocols through an OSPF AS



To meet the preceding requirements, configure a Route-Policy for Router A to set a tag for the imported IS-IS routes. Router D identifies the IS-IS routes from OSPF routes based on the tag.

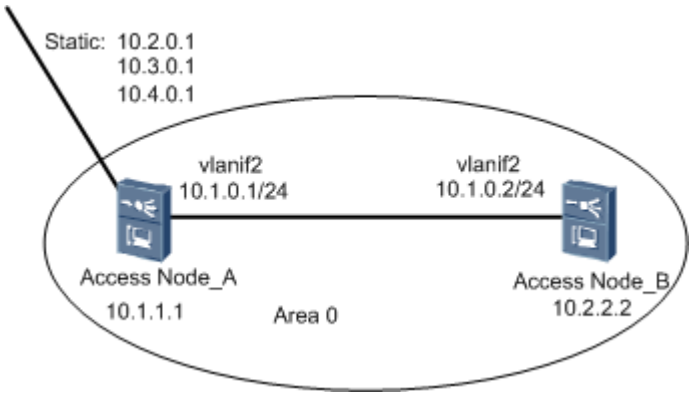
11.12.9.4 Configuration Example of the Routing Policy

This topic provides an example for configuring a routing policy for imported routes.

Service Requirements

- Consider two MA5800s with routing function enabled, namely MA5800_A and MA5800_B. Both of them are running the OSPF routing protocol, and within area 0.
- MA5800_A imports static routes, and MA5800_B is configured with the routing filtering policy.

Figure 11-177 Example network for configuring the routing policy



Procedure

Step 1 Configuring MA5800_A.

1. Configure the IP address of the Layer 3 interface on MA5800_A.

 Copyright © Huawei Technologies Co., Ltd.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.0.1 24
huawei(config-if-vlanif2)#quit
```

2. Enable OSPF on MA5800_A and specify the area ID to which the interface belongs.

```
huawei(config)#ospf
huawei(config-ospf-1)#area 0
huawei(config-ospf-1-area-0.0.0.0)#network 10.1.0.0 0.0.0.255
huawei(config-ospf-1-area-0.0.0.0)#quit
huawei(config-ospf-1)#quit
```

3. Configure the OSPF router ID of MA5800_A.

```
huawei(config)#router id 10.1.1.1
```

4. Configure three static routes.

```
huawei(config)#ip route-static 10.2.0.1 32 NULL 0
huawei(config)#ip route-static 10.3.0.1 32 NULL 0
huawei(config)#ip route-static 10.4.0.1 32 NULL 0
```

5. Import static routes into the OSPF routing table to improve its capability of obtaining routes.

```
huawei(config)#ospf
huawei(config-ospf-1)#import-route static
huawei(config-ospf-1)#quit
```

6. Save the data.

```
huawei(config)#save
```

Step 2 Configuring MA5800_B.

1. Configure the IP address of the Layer 3 interface on MA5800_B.

```
huawei(config)#vlan 2 smart
huawei(config)#port vlan 2 0/9 0
huawei(config)#interface vlanif 2
huawei(config-if-vlanif2)#ip address 10.1.0.2 24
huawei(config-if-vlanif2)#quit
```

2. Configure the ACL.

```
huawei(config)#acl 2000
huawei(config-acl-basic-2000)#rule deny source 10.3.0.0 255.255.255.0
huawei(config-acl-basic-2000)#rule permit source any
huawei(config-acl-basic-2000)#quit
```

3. Enable OSPF on MA5800_B and specify the area ID to which the interface belongs.

```
huawei(config)#ospf
huawei(config-ospf-1)#area 0
huawei(config-ospf-1-area-0.0.0.0)#network 10.1.0.0 0.0.0.255
huawei(config-ospf-1-area-0.0.0.0)#quit
huawei(config-ospf-1)#quit
```

4. Configure the OSPF router ID of MA5800_B.

```
huawei(config)#router id 10.2.2.2
```

5. Configure OSPF to filter the received external routes.

```
huawei(config)#ospf
uawei(config-ospf-1)#filter-policy 2000 import
huawei(config-ospf-1)#quit
```

6. Save the data.

```
huawei(config)#save
```

----End

 Copyright © Huawei Technologies Co., Ltd.

Result

1. MA5800_A and MA5800_B run OSPF successfully, and they can communicate well with each other.
2. After a filter is configured on MA5800_B, parts of the three imported static routes are available while part of them is screened on MA5800_B. That is, routes from segments 10.2.0.0 and 10.4.0.0 are available, while the route from segment 10.3.0.0 is screened.

Configuration File

Configuration on MA5800_A.

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ip address 10.1.0.1 24
quit
ospf
area 0
network 10.1.0.0 0.0.0.255
quit
quit
router id 10.1.1.1
ip route-static 10.2.0.1 32 NULL 0
ip route-static 10.3.0.1 32 NULL 0
ip route-static 10.4.0.1 32 NULL 0
ospf
import-route static
quit
save
```

Configuration on MA5800_B.

```
vlan 2 smart
port vlan 2 0/9 0
interface vlanif 2
ip address 10.1.0.2 24
acl 2000
rule deny source 10.3.0.0 255.255.255.0
rule permit source any
quit
ospf
area 0
network 10.1.0.0 0.0.0.255
quit
quit
router id 10.2.2.2
ospf
filter-policy 2000 import
quit
save
```

11.12.9.5 Routing Policy Reference Standards and Protocols

None

 Copyright © Huawei Technologies Co., Ltd.