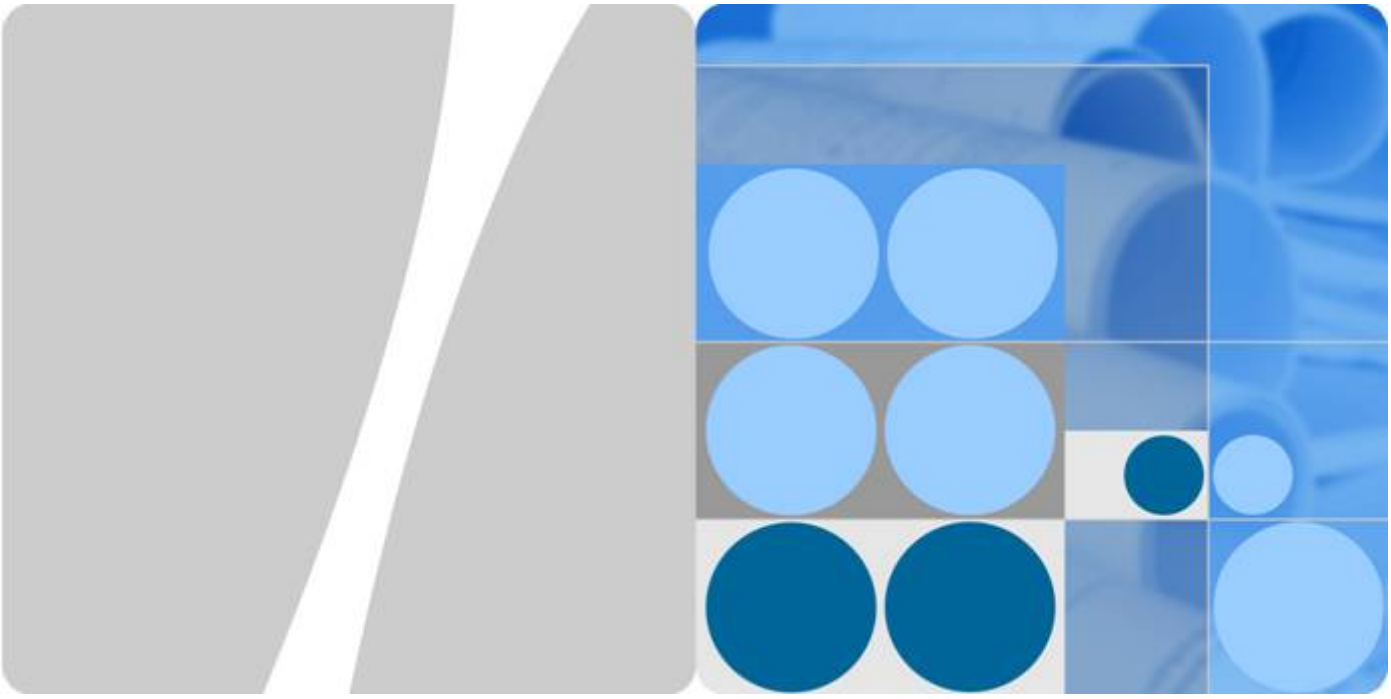


Enterprise Data Communication Products

Feature Description - IP Routing

Issue 01

Date 2012-09-30

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions


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

Notice

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

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

Huawei Technologies Co., Ltd.

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

Website: http://enterprise.huawei.com

About This Document

Intended Audience

This document describes the definition, purpose, and implementation of features on enterprise datacom products including the campus network switch, enterprise router, data center switch, and WLAN. For features supported by the device, see *Configuration Guide*.

This document is intended for:

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

Symbol Conventions

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

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

Command Conventions

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

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

Change History

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

Changes in Issue 01 (2012-09-30)

Initial commercial release.

Contents

1 IP Routing Overview

About This Chapter

1.1 Introduction to IP Routing

Routing is the basic element of data communication networks. Routing is the process of selecting paths on a network along which packets are sent from a source to a destination.

Routes are classified into the following types based on the destination address:

- Network segment route: The destination is a network segment. The subnet mask of an IPv4 destination address is less than 32 bits or the prefix length of an IPv6 destination address is less than 128 bits.
- Host route: The destination is a host. The subnet mask of an IPv4 destination address is 32 bits or the prefix length of an IPv6 destination address is 128 bits.

Routes are classified into the following types based on whether the destination is directly connected to a router:

- Direct route: The router is directly connected to the network where the destination is located.
- Indirect route: The router is indirectly connected to the network where the destination is located.

Routes are classified into the following types based on the destination address type:

- Unicast route: The destination address is a unicast address.
- Multicast route: The destination address is a multicast address.

This manual describes unicast routing. For details about mulitcast routing, see *Huawei AR150&AR200&AR1200&AR2200&AR3200 Series Enterprise Routers Feature Description - IP Multicast*.

1.2 Principles

1.2.1 Routers and Routing Principles

On the Internet, network connecting devices control traffic and ensure the quality of data transmission on the network. Common network connecting devices include hubs, bridges, switches, and routers. These network devices have similar basic principles. The following uses a router as an example to describe basic principles.

As a typical network connection device, a router is used to select routes and forward packets. According to the destination address in the received packet, a router selects a proper path, which has single-hop or multiple hops in it, to send the packet to the next router. The last router is responsible for sending the packet to the destination host.

A route is a path along which packets are sent from the source to the destination. When multiple routes are available to send packets from a router to the destination, the router can select the optimal route from an IP routing table to forward the packets. Optimal route selection depends on the routing protocol preferences and metrics of routes. When multiple routes have the same routing protocol preference and metric, load balancing can be implemented among these routes to relieve network pressure. When multiple routes have different routing protocol preferences and metrics, route backup can be implemented among these routes to improve network reliability.

1.2.2 Static Routes and Dynamic Routes

Routes support static routes and dynamic routes, including Routing Information Protocol (RIP) routes, Open Shortest Path First (OSPF) routes, Intermediate System-to-Intermediate System (IS-IS), and Border Gateway Protocol (BGP) routes.

Differences Between Static Routes and Dynamic Routes

Routing protocols are the rules used by routers to maintain routing tables, discover routes, generate routing tables, and guide packet forwarding. Routes are classified into the following types based on the origin:

- Direct route: is discovered by link layer protocols.
- Static route: is manually configured by network administrators.
- Dynamic route: is discovered by dynamic routing protocols.

Static routes are easy to configure, have low requirements on the system, and apply to simple, stable, and small networks. The disadvantage of static routes is that they cannot automatically adapt to network topology changes. Therefore, static routes require subsequent maintenance.

Dynamic routing protocols have their routing algorithms. Therefore, dynamic routes can automatically adapt to network topology changes and apply to the networks on which Layer 3 devices are deployed. The configurations of dynamic routes are complex. Dynamic routes have higher requirements on the system than static ones and consume network resources and system resources.

Classification of Dynamic Routing Protocols

Based on the application range, dynamic routing protocols are classified into the following types:

- Interior Gateway Protocol (IGP): runs inside an AS, such as RIP, OSPF, and IS-IS.
- Exterior Gateway Protocol (EGP): runs between different ASs, such as BGP.

Based on the type of algorithm they use, dynamic routing protocols are classified into the following types:

- Distance-vector routing protocol: includes RIP and BGP (BGP is also called Path-Vector).
- Link-state routing protocol: includes OSPF and IS-IS.

The preceding algorithms differ mainly in route discovery and calculation methods.

1.2.3 Routing Table and FIB Table

Routers forward packets based on routing tables and forwarding information base (FIB) tables. Each router maintains at least one routing table and one FIB table. Routers select routes based on routing tables and forward packets based on FIB tables.

Routing Table

Each router maintains a local core routing table, and each routing protocol maintains its routing table.

- Local core routing table

  A router uses the local core routing table to store protocol routes and preferred routes. The router then sends the preferred routes to the FIB table to guide packet forwarding. The

router selects routes according to the priorities of protocols and costs stored in the routing table.

NOTE

A router that supports Layer 3 Virtual Private Network (L3VPN) maintains a local core routing table for each VPN instance.

- 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.

Routing Table Contents

You can run the **display ip routing-table** command on a router to view brief information about a routing table of the router. The command output is as follows:

```
<Huawei> display ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
         Destinations : 14       Routes : 14

Destination/Mask    Proto   Pre  Cost      Flags NextHop         Interface

        0.0.0.0/0   Static  60   0           RD   10.137.216.1    GigabitEthernet
2/0/0
     10.10.10.0/24  Direct  0    0           D   10.10.10.10     GigabitEthernet
1/0/0
    10.10.10.10/32  Direct  0    0           D   127.0.0.1       InLoopBack0
   10.10.10.255/32  Direct  0    0           D   127.0.0.1       InLoopBack0
     10.10.11.0/24  Direct  0    0           D   10.10.11.1      LoopBack0
     10.10.11.1/32  Direct  0    0           D   127.0.0.1       InLoopBack0
   10.10.11.255/32  Direct  0    0           D   127.0.0.1       InLoopBack0
   10.137.216.0/23  Direct  0    0           D   10.137.217.208  GigabitEthernet
2/0/0
 10.137.217.208/32  Direct  0    0           D   127.0.0.1       InLoopBack0
 10.137.217.255/32  Direct  0    0           D   127.0.0.1       InLoopBack0
      127.0.0.0/8   Direct  0    0           D   127.0.0.1       InLoopBack0
      127.0.0.1/32  Direct  0    0           D   127.0.0.1       InLoopBack0
127.255.255.255/32  Direct  0    0           D   127.0.0.1       InLoopBack0
255.255.255.255/32  Direct  0    0           D   127.0.0.1       InLoopBack0
```

A routing table contains the following key data for each IP packet:

- Destination: is used to identify the destination IP address or the destination network address of an IP packet.
- Mask: is combined with the destination address to identify the address of the network segment where the destination host or router resides.

  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 1.1.1.1 and the mask is 255.255.255.0, the address of the network where the host or router resides is 1.1.1.0.

  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.

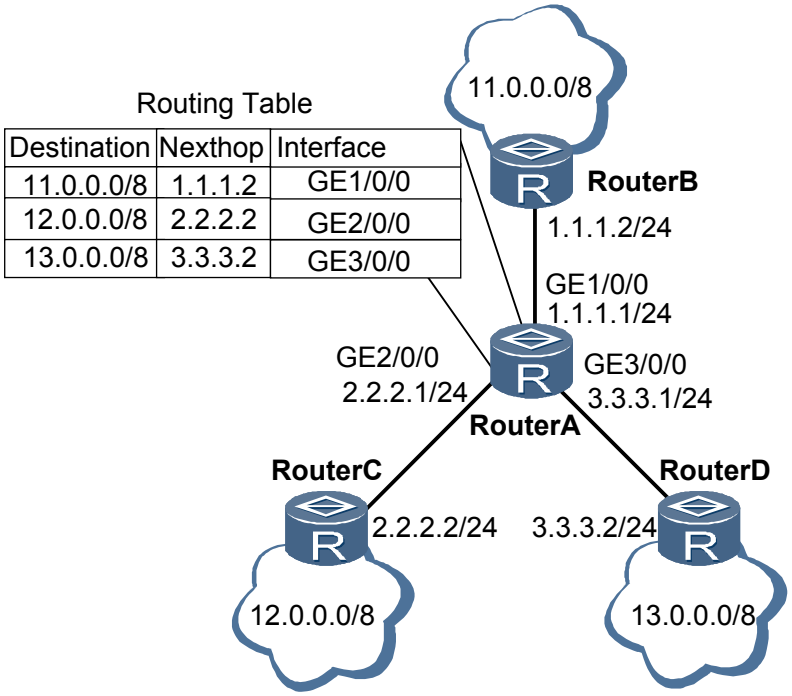
- Proto: indicates the protocol through which routes are learned.
- Pre: indicates the preference added to the IP routing table for a route. To the same destination, multiple routes with different next hops and outgoing interfaces exist. The routes in the table are those discovered by different routing 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 the priorities of routing protocols, see **1.2.4 Routing Protocol Preference**.
- 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.

As shown in **Figure 1-1**, Router A is connected with three networks, so it has three IP addresses and three physical interfaces. **Figure 1-1** also shows the routing table of Router A.

Figure 1-1 Schematic diagram of routing table



Automatic Restoration After the Number of Routes Exceeds the Upper Limit

A local core routing table stores the routes of different routing protocols. If the number of routes in the local core routing table reaches the upper limit, no route can be added to the table. The local core routing table has the following types of route limitations:

- System route limit: specifies the maximum number of routes supported by the system.

- System route prefix limit: specifies the range of prefixes for all the routes supported by the system.
- Multicast IGP route limit: specifies the maximum number of multicast IGP routes.
- Multi-topology route limit: specifies the maximum number of multi-topology routes.
- Private network route limit: specifies the maximum number of private network routes supported by the system.
- VPN route limit: specifies the maximum number of VPN routes supported by the system.
- VPN route prefix limit: specifies the range of prefixes for all the VPN routes supported by the system.

If a protocol fails to add routes to the local core routing table due to a specific route limitation, the system records the failure with the protocol name and routing table ID.

After routes of protocols are deleted from the local core routing table, and the number of routes falls below the upper limit, the system prompts all the protocols that failed to add routes to the local core routing table to re-add the routes to the local core routing table. This process restores most of the routes in the local core routing table. The mound of released table space determines whether all routes in the local core routing table can be restored.

Matching with 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.

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.

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      GigabitEthernet1/0/0
 8.0.0.0/8    RIP      100  3         D   120.0.0.2      GigabitEthernet1/0/0
 9.0.0.0/8    OSPF     10   50        D   20.0.0.2       GigabitEthernet3/0/0
 9.1.0.0/16   RIP      100  4         D   120.0.0.2      GigabitEthernet2/0/0
 20.0.0.0/8   Direct   0    0         D   20.0.0.1       GigabitEthernet4/0/0
```

After receiving a packet that carries the destination address 9.1.2.1, 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]         GigabitEthernet1/0/0  0x0
8.0.0.0/8          120.0.0.2        DU   t[37]         GigabitEthernet1/0/0  0x0
9.0.0.0/8          20.0.0.2         DU   t[9992]       GigabitEthernet3/0/0  0x0
9.1.0.0/16         120.0.0.2        DU   t[9992]       GigabitEthernet2/0/0  0x0
20.0.0.0/8         20.0.0.1         U    t[9992]       GigabitEthernet4/0/0  0x0
```

Then the router performs the "AND" operation on the destination address 9.1.2.1 and the masks 0, 8, 16 to obtain the network segment addresses: 0.0.0.0/0, 9.0.0.0/8, and 9.1.0.0/16. The three

addresses match three entries in the table. The router chooses the 9.1.0.0/16 entry because it is the longest match. The router then forwards the packet through GigabitEthernet 2/0/0 for the 9.1.0.0/16 entry.

1.2.4 Routing Protocol Preference

Routing protocols (including 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, each routing protocols (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 (a smaller value indicates a higher highest preference) and added to the local routing table.

Routers define external preference and internal preference. External preference is the manually configured for each routing protocol. **Table 1-1** lists the default preferences of routing protocols.

Table 1-1 Routing protocols and their default preferences

Routing Protocol or Route Type	Default External Preference
Direct	0
OSPF	10
IS-IS	15
Static	60
RIP	100
OSPF ASE	150
OSPF NSSA	150
IBGP	255
EBGP	255

NOTE

In **Table 1-1**, the value 0 indicates direct routes and the value 255 indicates routes learned from unreliable sources. A smaller value indicates a higher preference.

Except for direct routes, you can manually configure a routing protocol's preference. In addition, the preference for each static route varies.

Internal preferences of routing protocols cannot be manually configured. **Table 1-2** lists the internal preferences of routing protocols.

Table 1-2 Internal preferences of routing protocols

Routing Protocol or Route Type	Internal Preference
Direct	0
OSPF	10
IS-IS Level-1	15
IS-IS Level-2	18
Static	60
RIP	100
OSPF ASE	150
OSPF NSSA	150
IBGP	200
EBGP	20

During route selection, a router first compares the external preferences of routes. When the same external preference is set for different routing protocols, the router selects the optimal route based on the internal preference. Assume that there are two routes to 10.1.1.0/24: a static route and an OSPF route. Both routes have the same external preference that is set to 5. In this case, the router determines the optimal route based on the internal preference listed in **Table 1-2**. An OSPF route has an internal preference 10, and a static route has an internal preference 60. This indicates that the OSPF route has a higher preference than the static route. Therefore, the router selects the OSPF route as the optimal route.

1.2.5 Route Metric

A route metric specifies the cost of a route to a specified destination address. The following factors often affect the route metric:

- Path length

  The path length is the most common factors affecting the route metric. Link-state routing protocols allow you to assign a link cost for each link to identify the path length of a link. In this case, the path length is the sum of link costs of all the links that packets pass through. Distance-vector routing protocols use the hop count to identify the path length. The hop count is the number of devices that packets pass through from the source to the destination. For example, the hop count from a router to its directly connected network is 0, and the hop count from a router to a network that can be reached through another router is 1. The rest can be deduced in the same manner.

- Network bandwidth

  The network bandwidth is the transmission capability of a link. For example, a 10-Gigabit link has a higher transmission capability than a 1-Gigabit link. Although bandwidth defines the maximum transmission rate of a link, routes over high-bandwidth links are not

necessarily better than routes over low-bandwidth links. For example, when a high-bandwidth link is congested, forwarding packets over this link will require more time.

- Load

  The load is the degree to which a network resource is busy. You can calculate the load by calculating the CPU usage and packets processed per second. Monitoring the CPU usage and packets processed per second continually helps learn about network usage.

- Communication cost

  The communication cost measures the operating cost of a route over a link. The communication cost is another important indicator, especially if you do not care about network performance but the operating expenditure.

1.2.6 Load Balancing and Route Backup

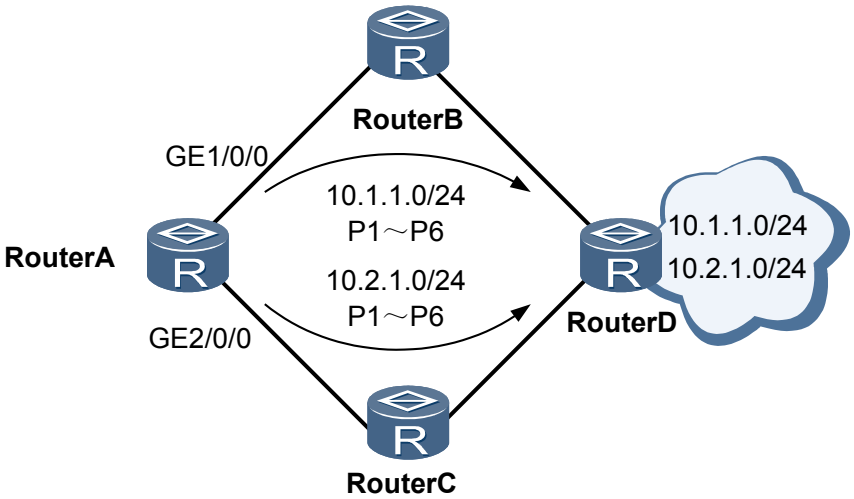
When multiple routes have the same routing protocol preference and metric, these routes are called equal-cost routes, among which load balancing can be implemented. When multiple routes have different routing protocol preferences and metrics, route backup can be implemented among these routes.

Load Balancing

Routers support the multi-route mode, allowing you to configure multiple routes with the same destination and preference. If the destinations and costs of multiple routes discovered by the same routing protocol are the same, load balancing can be performed among the routes.

During load balancing, a router forwards packets based on the 5-tuple (source IP address, destination IP address, source port, destination port, and transport protocol) in the packets. When the 5-tuple information is the same, the router always chooses the next-hop address that is the same as the last one to send packets. When the 5-tuple information is different, the router forwards packet over idle paths.

Figure 1-2 Networking diagram of load balancing



As shown in **Figure 1-2**, RouterA forwards the first packet P1 to 10.1.1.0/24 through GE1/0/0 and needs to forward subsequent packets to 10.1.1.0/24 and 10.2.1.0/24 respectively. The forwarding process is as follows:

- When forwarding the second packet P2 to 10.1.1.0/24, RouterA forwards P2 and subsequent packets destined for 10.1.1.0/24 through GE1/0/0 if it finds that the 5-tuple information of P2 is the same as that of P1 destined for 10.1.1.0/24.
- When forwarding the first packet P1 to 10.2.1.0/24, RouterA forwards this packet and subsequent packets destined for 10.2.1.0/24 through GE2/0/0 if it finds that the 5-tuple information of P1 destined for 10.2.1.0/24 is different from that of P1 destined for 10.1.1.0/24.

NOTE

The number of equal-cost routes for load balancing varies with products.

Route Backup

Route backup can improve network reliability. You can configure multiple routes to the same destination as required. The route with the highest preference functions as the primary route, and the other routes with lower preferences function as backup routes.

A router generally uses the primary route to forward data. When the primary link fails, the primary route becomes inactive. The router selects a backup route with the highest preference to forward data. In this manner, data is switched from the primary route to a backup route. When the primary link recovers, the router selects the primary route to forward data again because the primary route has the highest preference. Data is then switched back from the backup route to the primary route.

1.2.7 IP FRR

Definition

FRR refers to the mechanism that a fault detected at the physical layer or data link layer is reported to the upper-layer routing system, and a backup link is immediately used to forward packets. IP FRR is a method that implements fast route backup.

Purpose

On traditional IP networks, when a fault occurs at the lower layer of the forwarding link, the visible evidence is that the physical interface on the router becomes Down. After the router detects the fault, it informs the upper layer routing system to recalculate routes and then update routing information. Usually, it takes the routing system several seconds to re-select an available route.

Second-level convergence is intolerable to the services that are quite sensitive to delay and packet loss because it may lead to service interruption. For example, Voice over Internet Protocol (VoIP) services are only tolerant of millisecond-level interruption.

IP FRR ensures that the forwarding system swiftly detects such a fault and then takes measures to restore services as soon as possible.

IP FRR Classification and Implementation

IP FRR, which is designed for routes on IP networks, is classified into IP FRR on the public network and IP FRR on the private network.

- IP FRR on the public network: protects routers of the public network.

- IP FRR on the private network: protects Customer Edges (CEs).

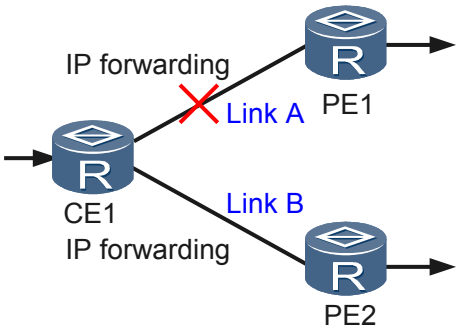
IP FRR is implemented as follows:

1. If the primary link is available, you can configure IP FRR by using a routing policy to provide the forwarding information of the backup route for the forwarding engine.
2. If the forwarding engine is notified of a link fault, the engine uses the backup link to forward traffic before the routes on the control plane converge.

IP FRR Typical Applications

As shown in **Figure 1-3**, IP FRR is configured to improve network reliability. CE1 is dual-homed to PE1 and PE2. CE1 is configured with two outbound interfaces and two next hops. That is, link B functions as the backup of link A. When link A fails, traffic can be rapidly switched to link B.

Figure 1-3 Configuring the IP FRR function



1.2.8 Route Convergence

Definition

Route convergence is the action of recalculating routes to replace existing routes in the case of network topology changes. The integration of network services urgently requires differentiated services. Routes for key services, such as Voice over IP (VoIP), video conferences, and multicast services, need to be converged rapidly, while routes for common services can be converged relatively slowly. In this case, the system needs to converge routes based on their convergence priorities to improve network reliability.

Priority-based convergence is a mechanism that allows the system to converge routes based on the convergence priority. You can set different convergence priorities for routes: critical, high, medium, and low, which are in descending order of priority. The system then converge routes according to the scheduling weight to guide service forwarding.

Principles

The routing protocols first compute and deliver routes of high convergence priorities to the system. You can re-configure the scheduling weight values as required. **Table 1-3** lists the default convergence priorities of public routes.

Table 1-3 Default convergence priorities of public routes

Routing Protocol or Route Type	Convergence Priority
Direct	high
Static	medium
32-bit host routes of OSPF and IS-IS	medium
OSPF routes (excluding 32-bit host routes)	low
IS-IS routes (excluding 32-bit host routes)	low
RIP	low
BGP	low

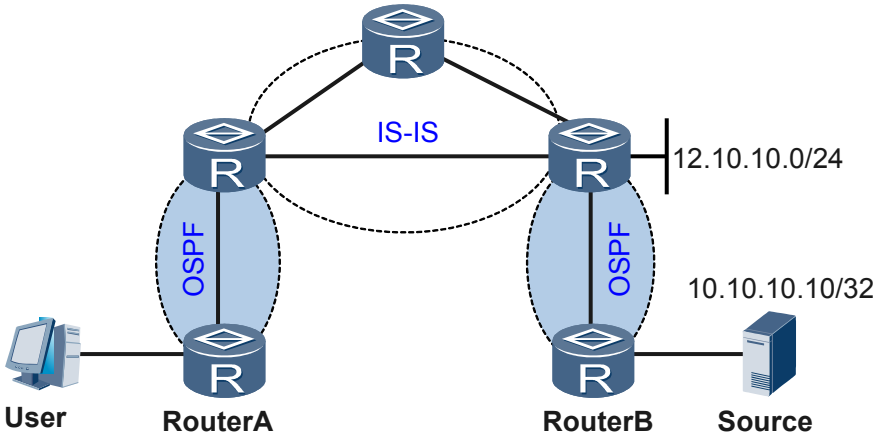
NOTE

For private routes, only 32-bit host routes of OSPF and IS-IS can be identified as medium and all other routes are identifies as low.

Priority-based Route Convergence

Figure 1-4 shows the networking for multicast services. OSPF and IS-IS run on the network; the receiver connects to RouterA; the multicast source server 10.10.10.10/32 connects to RouterB. The route to the multicast source server must be converged faster than other routes, such as 12.10.10.0/24. You can set the convergence priority of route 10.10.10.10/32 higher than that of route 12.10.10.0/24. When routes are converged on the network, the route to the multicast source server 10.10.10.10/32 is converged first. This ensures the transmission of multicast services.

Figure 1-4 Networking diagram of priority-based route convergence



1.2.9 Default Routes

Default routes are special routes. Default routes are used only when packets to be forwarded have no matching routing entry in a routing table. If the destination address of a packet does not

match any entry in the routing table, the packet is sent through 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.

In a routing table, a default route is the route to network 0.0.0.0 (with the mask 0.0.0.0). You can run the **display ip routing-table** command to check whether a default route is configured. Generally, administrators can manually configure default static routes. Default routes can also be generated through dynamic routing protocols such as OSPF and IS-IS.

1.2.10 Route Import

Different routing protocols may discover different routes because they use different algorithms. If multiple routing protocols run on a large network, the routing protocols need to re-advertise the routes they discover.

Each routing protocol can import the routes discovered by other routing protocols, direct routes, and static routes using its mechanism.

1.2.11 Autonomous System

An Autonomous System (AS) is a set of IP networks and routers that are controlled by one entity that presents a common routing policy to the Internet.

Each AS supports multiple internal routing protocols. All the networks in an AS are assigned the same AS number and managed by the same administration group. Two types of AS numbers are available: 2-byte AS number and 4-byte AS number. A 2-byte AS number ranges from 1 to 65535. Available AS numbers become almost exhausted. Therefore, 2-byte AS numbers need to be extended to 4-byte AS numbers that range from 1 to 4294967295. A 4-byte AS number is in the X.Y format, where X ranges from 1 to 65535 and Y ranges from 0 to 65535.

AS numbers are classified into two types based on the network where they are used. **Table 1-4** lists the two types of AS numbers and their ranges.

Table 1-4 AS number types and ranges

AS Number Type	2-Byte AS Number	4-Byte AS Number
Public AS number	1 to 64511	1 to 64511, 65536 to 4294967295
Private AS number	64512 to 65535	64512 to 65535

1.2.12 Indirect Next Hop

Definition

Indirect next hop can change the direct association between route prefixes and the next hop into an indirect association. Then, next hop information can be refreshed independently, the prefixes of the same next hop need not be refreshed one by one, and thus route convergence is speeded up.

Purpose

In the scenario in need of route iteration, when IGP routes or tunnels are switched, FIB entries are quickly refreshed. This implements traffic fast convergence and reduces the impact on services.

Mapping Between the Route Prefix and the Next Hop

The mapping between the route prefix and the next hop is the basis of indirect next hop. To meet the requirements of route iteration and tunnel iteration in different scenarios, next hop information involves the address family, the original next hop address, or the tunnel policy. The system assigns an index to information about each next hop, performs route iteration, and then notifies the iteration result to the route protocol and distributes FIB entries.

On-Demand Route Iteration

On-demand route iteration indicates that when a dependent route is changed, only the next hop related to the dependent route is re-iterated. If the destination address of a route is the original next hop address or network segment address of next hop information, route changes affect the iteration result of next hop information. Otherwise, route changes do not affect next hop information. Therefore, when a route changes, you can re-iterate only the related next hop by judging the destination address of the route.

With respect to tunnel iteration, when a tunnel alternates between up and down, you just need to re-iterate the next hop information whose next hop address is the same as the destination address of the tunnel.

Iteration Policy

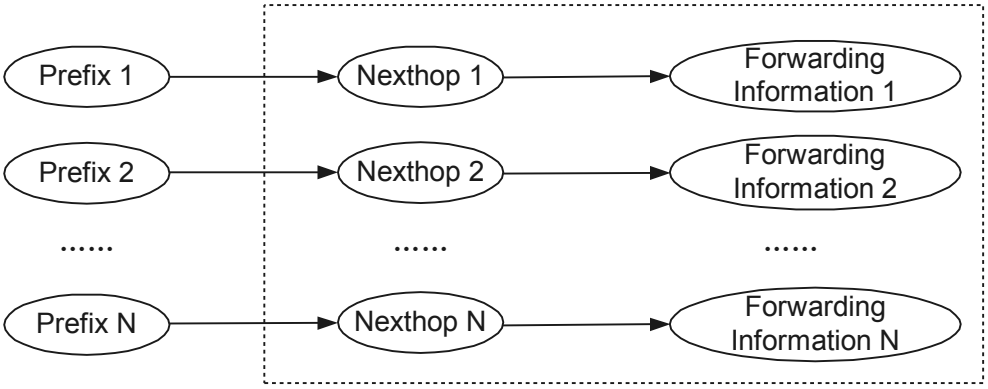
An iteration policy is used to control the iteration result of the next hop to meet the requirements of different application scenarios. In route iteration, iteration behaviors do not need to be controlled by the iteration policy. Instead, iteration behaviors only needs to comply with the longest matching rule. What is more, the iteration policy needs to be applied only when VPN routes iterate tunnels.

By default, the system selects LSPs for a VPN without performing load balancing. If load balancing or other types of tunnels are required, you need to configure a tunnel policy and bind the tunnel policy to a tunnel. After a tunnel policy is applied, the system adopts the tunnel bound in the tunnel policy or selects a tunnel according to the priorities of different types of tunnels.

Refreshment of Indirect Next Hop

On the forwarding plane, public network routes are forwarded based on the next hop and outbound interface while VPN routes are forwarded based on the public network tunnel in addition to the next hop and outbound interface. Before indirect next hop is adopted, forwarding information, including the next hop, outbound interface, and the tunnel token, needs to be added into the FIB entry by using the route prefix. In this manner, the route convergence speed is relevant to the number of route prefixes. After indirect next hop is adopted, many route prefixes corresponds to a shared next hop. Forwarding information is added into the FIB entry by using the next hop, and the traffic with the relevant route prefixes can be switched simultaneously. Therefore, the route convergence speed becomes faster.

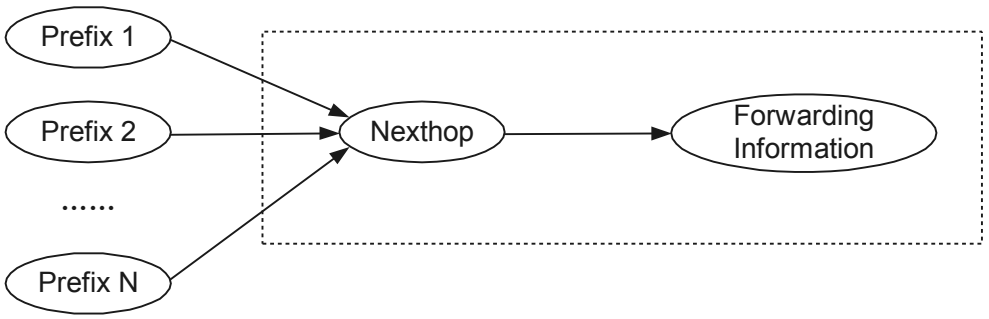
Figure 1-5 Schematic diagram before indirect next hop is adopted



As shown in **Figure 1-5**, before indirect next hop is adopted, prefixes are totally independent, each corresponding to its next hop and forwarding information. When a dependent route changes, the next hop corresponding to each prefix is iterated and forwarding information is updated based on the prefix. In this case, the convergence speed is related to the number of prefixes.

Actually, prefixes of a BGP neighbor have the same next hop, forwarding information, and refreshed forwarding information.

Figure 1-6 Schematic diagram after indirect next hop is adopted



As shown in **Figure 1-6**, after indirect next hop is adopted, prefixes of a BGP neighbor share a next hop. When a dependent route changes, only the shared next hop is iterated and forwarding information is updated based on the next hop. In this case, traffic of all prefixes can be converged at a time. The convergence speed is irrelevant to the number of prefixes.

Comparison Between Route Iteration and Tunnel Iteration

Comparison between route iteration and tunnel iteration is shown in the following table.

Table 1-5 Comparison between route iteration and tunnel iteration

Iteration Type	Description
Route iteration	● Iterating BGP public routes. ● It is triggered by route changes. ● It supports next-hop iteration based on the specified routing policy.
Tunnel iteration	● Iterating BGP VPN routes. ● It is triggered by tunnel changes or tunnel policy changes. ● Iteration behaviors can be controlled through the tunnel policy to meet the requirements of different application scenarios.

1.3 References

None

2 Static Route

About This Chapter

2.1 Introduction to Static Routes

Definition

Static routes are routes that are manually configured by the administrator.

Purpose

Static routes provide different functions on different networks.

- On a simple network, only static routes are required to ensure normal running of the network.
- On a complex network, static routes improve the network performance and ensure the required bandwidth for important applications.
- The static routes associated with VPN instances are used to manage VPN routes.

2.2 Principles

2.2.1 Basics of Static Routes

A router forwards data packets based on routing entries containing route information. The routing entries can be manually configured or calculated by dynamic routing protocols. Static routes refer to the routes that are manually added to the routing table.

Static routes use less bandwidth than dynamic routes. No CPU cycle is required for calculating or analyzing routing update. When a fault occurs on the network or the topology changes, static routes cannot automatically change and must be changed manually. The configuration of a static route includes destination IP address and mask, outbound interface and next-hop address, and preference.

Destination Address and Mask

The destination IPv4 address is expressed in dotted decimal notation. The mask can be expressed either in dotted decimal notation or by the mask length, that is, the number of consecutive 1s in the mask. For details about the destination IPv6 address and mask, see "IPv6 - Principles - IPv6 Addresses" in the *Feature Description - IP Service*. When the destination and mask are set to all 0s, the default static route is configured. For details about the default static route, see **2.3.2 Application of the Default Static Route**.

Outbound Interface and Next-Hop IP Address

When configuring a static route, you can specify the outbound interface and the next-hop IP address based on outbound interfaces types.

- Configure the outbound interface for point-to-point (P2P) interfaces. For a P2P interface, the next-hop address is specified after the outbound interface is specified. That is, the address of the remote interface (interface on the peer device) connected to this interface is the next-hop address. For example, the protocol used to encapsulate 10GE is the Point-to-

Point protocol (PPP). The remote IP address is obtained through PPP negotiation. You need specify only the outbound interface.

- Configure the next hop for Non Broadcast Multiple Access (NBMA) interfaces (for example, ATM interfaces). You need to configure the IP route and the mapping between IP addresses and link-layer addresses.
- Configure the next hop for broadcast interfaces (for example, Ethernet interfaces) and virtual template (VT) interfaces. The Ethernet interface is a broadcast interface, and the VT interface can be associated with several virtual access (VA) interfaces. If the Ethernet interface or the VT interface is specified as the outbound interface, multiple next hops exist and the system cannot decide which next hop is to be used. Therefore, this configuration is not recommended.

Static Route Preference

Different static routes can be configured with different preferences. A smaller preference value indicates a higher priority of static routes. If you specify the same preference for the static routes to the same destination, you can implement load balancing among these routes. If you specify different preferences for the static routes, you can implement route backup among the routes. For details, see **2.3.1 Load Balancing and Route Backup**.

2.2.2 BFD for Static Routes

Different from dynamic routing protocols, static routes do not have a detection mechanism. As a result, when a link fault occurs on the network, the administrator needs to handle it. Bidirectional Forwarding Detection (BFD) for static route is introduced to bind a static route to a BFD session so that the BFD session can detect the status of the link where the static route resides.

- When the BFD session that is bound to a static route detects a link fault, BFD reports the link fault to the Routing Management (RM) module. The RM module sets the route to inactive. The route is unavailable in the routing table.
- When the BFD session that is bound to a static route detects that the faulty link is re-established, BFD reports a message to the RM module. The RM module sets the route to active. The route is available in the IP routing table.

NOTE

For details about BFD, see "BFD" in the *Feature Description - Reliability*.

2.2.3 NQA for Static Routes

As mentioned previously, static routes do not have a dedicated detection mechanism. After a fault occurs, the corresponding static route is automatically deleted from the IP routing table. This condition delays the link switchover and can interrupt services for a comparatively long time. The network administrator must delete the corresponding static route to allow traffic to switch to an available path.

An effective method is required to detect faults in links related to static routes. BFD for static routes is applicable only to the scenario where both communicating devices support BFD. If either of the two communicating devices supports NQA, NQA for static routes can be used to detect faults in links where Layer 2 devices reside.

NQA for static routes refers to the association between a static route and an NQA test instance. The system can use the NQA test instance to check the link status. Then, according to the NQA

test result, the system can determine an optimal route in time to prevent communication interruption and ensure service quality. NQA for static routes functions as follows:

- If NQA detects a fault in the link, the system sets the static route to inactive. The route becomes unavailable and is deleted from the IP routing table.
- If NQA finds that the link recovers, the system sets the static route to active. The route becomes available and is added to the IP routing table.

NOTE

Each static route can be associated with only one NQA test instance.

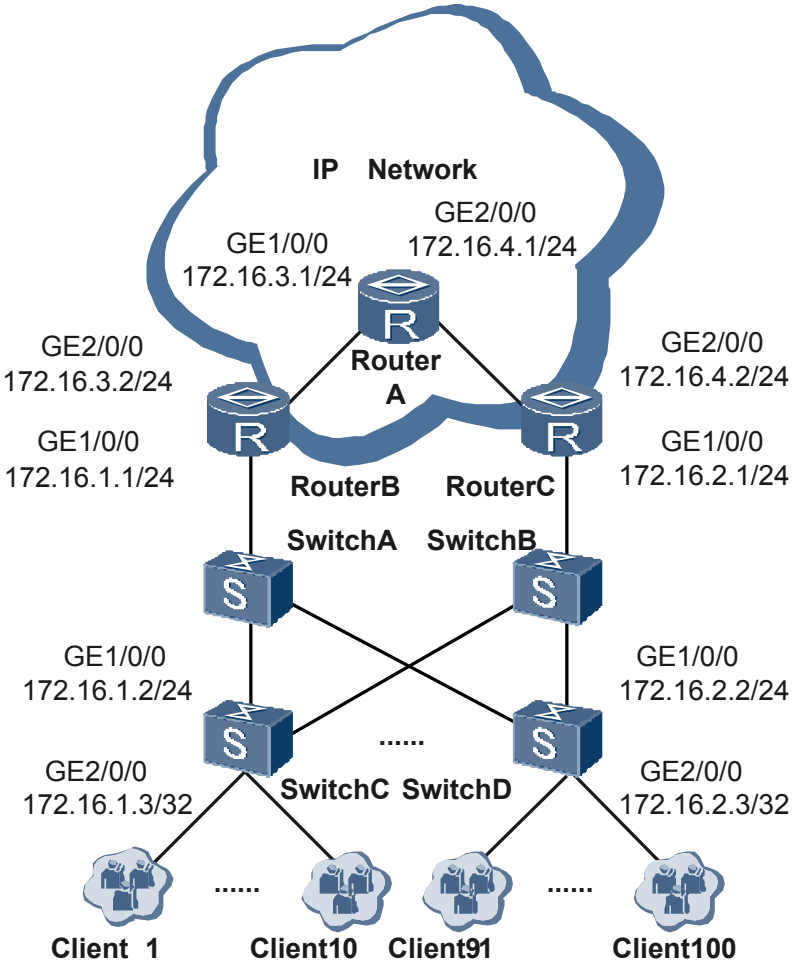
For details about NQA, see the *Huawei AR150&AR200&AR1200&AR2200&AR3200 Series Enterprise Routers Feature Description - Network Management*.

Applications

On the network shown in **Figure 2-1**, each access switch provides access services for 10 users, and a total of 100 users are connected to the network. Because dynamic routing protocols are unavailable for communication between Router B and users, static routes are configured on Router B. For network stability, Router C, functioning as the backup for Router B, is configured with static routes to the same destination. Router A, Router B, and Router C run a dynamic routing protocol to learn routes from each other. Router B and Router C import static routes using a dynamic routing protocol and have different costs for these static routes. After the configuration is complete, Router A can use the dynamic routing protocol to learn routes destined for users from Router B and Router C. Router A uses the link related to the static route with a lower cost as the active link and the other link as the standby link.

NQA for static routes is configured on Router B. NQA tests are performed to check the active link of Router B → Switch A → Switch C (Switch D). If the active link fails, the corresponding static route is deleted from the routing table, and traffic diverts to the standby link of Router C → Switch B → Switch C (Switch D). If both links work properly, traffic travels along the active link.

Figure 2-1 Networking to apply NQA for static routes



2.2.4 Permanent Advertisement of Static Routes

Link connectivity determines the stability and availability of a network. Therefore, link detection plays an important role in network maintenance. BFD, as a link detection mechanism, is inapplicable to certain scenarios. For example, a simpler and more natural method is required for link detection between different ISPs.

Permanent advertisement of static routes provides a low-cost and simple link detection mechanism and improves compatibility between Huawei devices and non-Huawei devices. If service traffic needs to be forwarded along a specified path, you can ping the destination addresses of static routes to detect the link connectivity.

After permanent advertisement of static routes is configured, the static routes that cannot be advertised are still preferred and are added to the routing table in the following cases:

- If an outbound interface configured with an IP address is specified for a static route, the static route is always preferred and added to the routing table regardless of whether the outbound interface is Up or Down.

- If no outbound interface is specified for a static route, the static route is always preferred and added to the routing table regardless of whether the static route can be iterated to an outbound interface.

In this way, you can enable IP packets to be always forwarded through this static route. The permanent advertisement mechanism provides a way for you to monitor services and detect link connectivity.

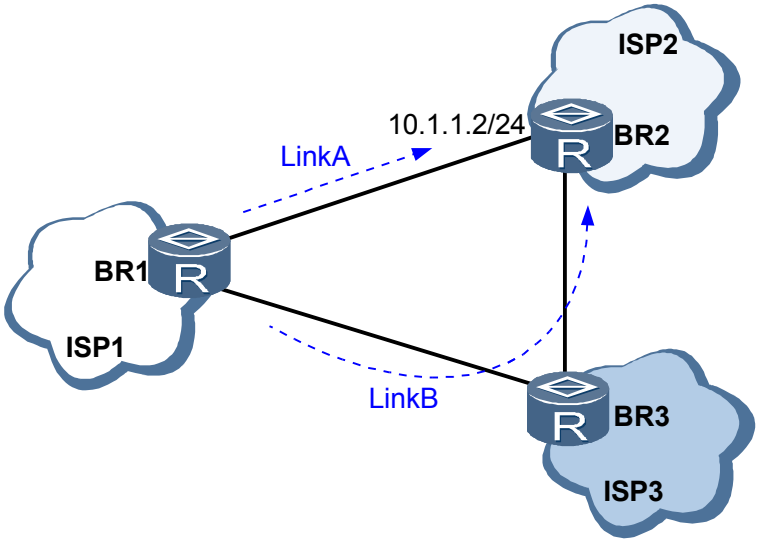
NOTE

A device enabled with this feature always stores static routes in its IP routing table, regardless of whether the static routes are reachable. If a path is unreachable, the corresponding static route may become a blackhole route.

Applications

In **Figure 2-2**, BR1, BR2, and BR3 belong to ISP1, ISP2, and ISP3 respectively. Between BR1 and BR2 are two links, Link A and Link B. ISP1, however, requires that service traffic be forwarded to ISP2 over Link A without traveling through ISP3.

Figure 2-2 Networking for applying permanent advertisement of static routes



The External Border Gateway Protocol (EBGP) peer relationship is established between BR1 and BR2. For service monitoring, a static route destined for the BGP peer (BR2) at 10.1.1.2/24 is configured on BR1, and permanent advertisement of static routes is enabled. The interface that connects BR1 to BR2 is specified as the outbound interface of the static route. Then, the network monitoring system periodically pings 10.1.1.2 to determine the status of Link A.

If Link A works properly, ping packets are forwarded over Link A. If Link A becomes faulty, although service traffic can reach BR2 over Link B, the static route is still preferred because its preference is higher. Therefore, ping packets are still forwarded over Link A, but packet forwarding fails. This scenario is also applicable to BGP packets. That is, a link fault causes the BGP peer relationship to be interrupted. The monitoring system detects service faults as returned in the ping result and prompts maintenance engineers to rectify the faults before services are affected.

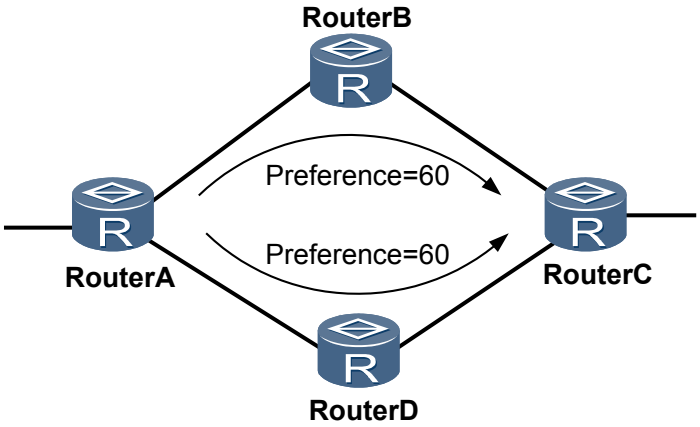
2.3 Applications

2.3.1 Load Balancing and Route Backup

Load Balancing Among Static Routes

To implement load balancing, specify the same preference for multiple routes to the same destination.

Figure 2-3 Load Balancing Among Static Routes

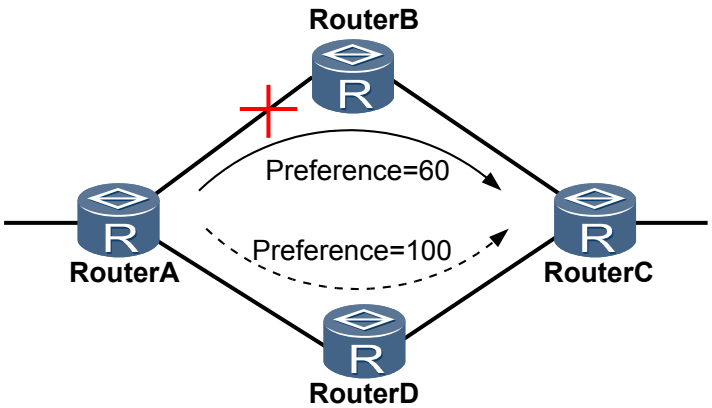


As shown in **Figure 2-3**, there are two static routes with the same preference from RouterA to RouterC. The two routes are stored in the routing table and used to forward data.

Route Backup

To implement route backup, specify different preferences for multiple routes to the same destination.

Figure 2-4 Route backup



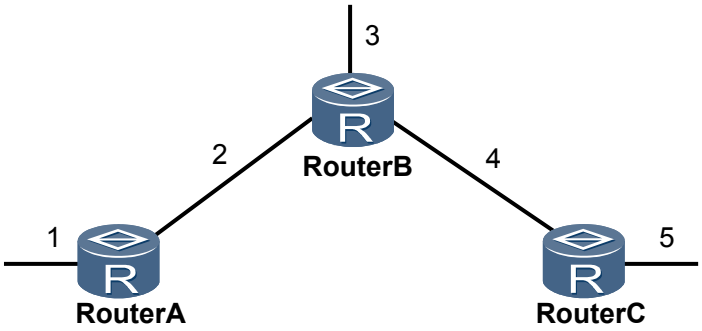
As shown in **Figure 2-4**, there are two static routes with different preferences from RouterA to RouterC. Static route B with next hop RouterB has a higher preference. The link that static route B belongs to functions as the active link. Static route D with next hop RouterD has a higher preference. The link that static route D belongs to functions as the standby link.

- In normal situations, static route B is activated, and the active link forwards data. Static route D is not shown in the routing table.
- If a fault occurs on the active link, static route B is deleted from the routing table. Static route D is activated, and the standby link forwards data.
- When the active link restores, static route B is activated again, and the active link forwards data. Static route D is deleted from the routing table and functions as the backup route. Static route D is also called a floating static route.

2.3.2 Application of the Default Static Route

When the destination IP address is set to all 0s, a default route is configured. The default route can be automatically generated by a routing protocol or manually configured. The default route manually configured simplifies network configuration. If the destination address of a packet fails to match any entry in the routing table, the router selects the default route to forward the packet.

Figure 2-5 Networking diagram of static routes



As shown in **Figure 2-5**, if no default static route is configured, you need to configure static routes destined for networks 3, 4, and 5 on RouterA, configure static routes destined for networks 1 and 5 on RouterB, and configure static routes destined for networks 1, 2, and 3 on RouterC. In this way, RouterA, RouterB, and RouterC can communicate with each other.

The next hop of the packets sent by RouterA to networks 3, 4, and 5 is RouterB. Therefore, a default route configured on RouterA can replace the three static routes destined for networks 3, 4, and 5 in the preceding example. Similarly, only a default route from RouterC to RouterB can replace the three static routes destined for networks 1, 2, and 3 in the preceding example.

2.4 References

None.

3 RIP

About This Chapter

3.1 Introduction to RIP

Definition

Routing Information Protocol (RIP) is a simple Interior Gateway Protocol (IGP). RIP is a Distance-Vector protocol that uses hot count to measure the distance between the local device and the destination. RIP exchanges routing information using UDP packets on UDP port 520.

Two versions are available for RIP: RIP-1 and RIP-2. RIP-2 is an extension to RIP-1.

Purpose

RIP is easy to implement, and is easier to configure and manage than OSPF and IS-IS. Therefore, RIP is applicable to small-sized networks, such as campus networks and simple LANs. It is not suitable for complex environments or large-sized networks.

3.2 Principles

3.2.1 Principles

RIP is based on the Distance-Vector (DV) algorithm. RIP uses hop count (HC) to measure the distance to the destination. The distance is called the metric value. In RIP, the default HC from a router to its directly connected network is 0, and the HC from a router to a reachable network through another router is 1, and so on. That is to say, the HC equals the number of routers passed from the local network to the destination network. To speed up network convergence, RIP defines the HC as an integer that ranges from 0 to 15. An HC 16 or greater is defined as infinity, that is, the destination network or the host is unreachable. For this reason, RIP is not applied to large-scale networks.

RIP Routing Table

When RIP starts on a router, the RIP routing table contains only the routes to the directly connected interfaces. After neighboring routers on different network segments learn the routing entries from each other, they can communicate with each other.

Figure 3-1 RIP routing table generation

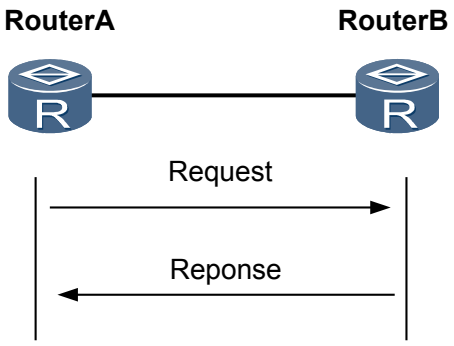


Figure 3-1 shows the process of RIP routing table generation.

- RIP starts, and then router A broadcasts Request packets to neighboring routers.
- When receiving the Request packet, router B encapsulates its own routing table into the Response packet and broadcasts the Response packet to the network segment connected to the interface receiving the Request packet.
- Router A generates a routing table based on the Response packet sent from router B.

RIP Update and Maintenance

RIP uses four timers to update and maintain routing information:

- Update timer: When this timer expires, a router immediately sends an Update packet.
- Age timer: If a RIP device does not receive an Update packet from a neighbor within the aging time, the RIP device considers the route to this neighbor unreachable.
- Garbage-collect timer: If a RIP device does not receive an Update packet of an unreachable route within the timeout interval, the device deletes the routing entry from the routing table.
- Suppress timer: When a RIP device receives an Update packet with the Cost field being 16 from a neighbor, the route is suppressed and the suppress timer starts. To avoid route flapping, the RIP device does not accept any Update packet before the suppress timer expires even if the Cost field in an Update packet is smaller than 16. After the suppress timer expires, the RIP device accepts new Update packets.

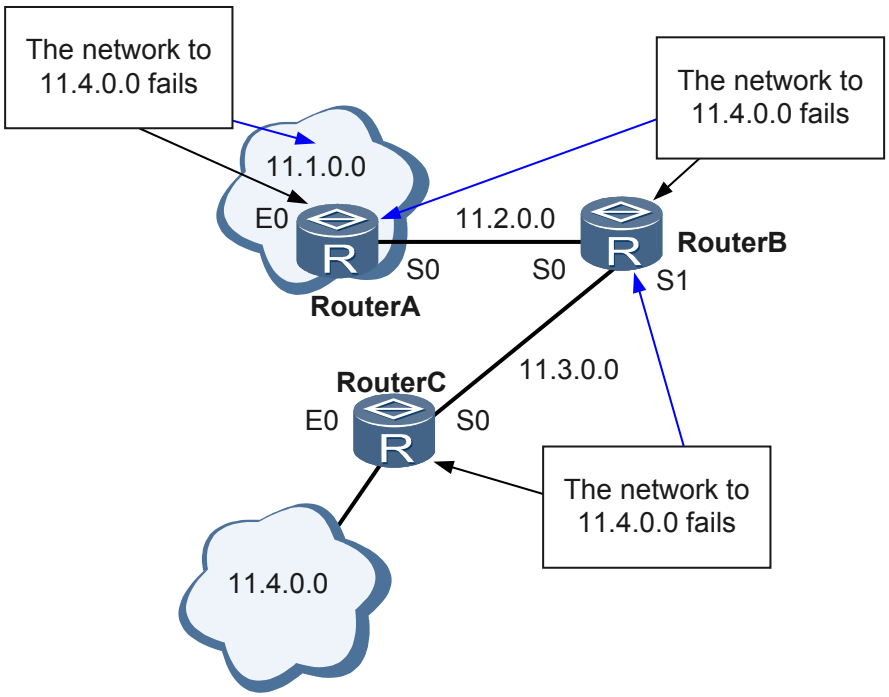
Relationships between RIP routes and timers:

- The interval for sending Update packets is determined by the Update timer, which is 30 seconds by default.
- Each routing entry has two timers: age timer and Garbage-collect timer. When a RIP device adds a learned route to the local routing table, the age timer starts for the routing entry. If the RIP device does not receive an Update packet from the neighbor within the age time, the RIP device sets the Cost value of the route to 16 (unreachable) and starts the Garbage-collect timer. If the RIP device still does not receive an Update packet within the Garbage-collect timer, the RIP device deletes the routing entry from the routing table.

Triggered Update

When routing information changes, a device immediately sends an Update packet to its neighbors, without waiting for Update timer expiration. This function avoids loops.

Figure 3-2 Triggered update



As shown in **Figure 3-2**, router C first learns that network 11.4.0.0 is unreachable.

- If router C does not support triggered update when detecting a link fault, it has to wait until the Update timer expires. If router C receives an Update packet from router B before its Update timer expires, router C learns a wrong route to network 11.4.0.0. In this case, the next hops of the routes from router B or router C to network 11.4.0.0 are router C and router B respectively. A routing loop is generated.
- If router C supports triggered update when detecting a link fault, router C immediately sends an Update packet to router B so that a routing loop is prevented.

3.2.2 RIP-2 Enhanced Features

Two versions are available for RIP: RIP-1 and RIP-2. RIP-2 is an extension to RIP-1.

Comparison Between RIP-1 and RIP-2

RIP version 1 (RIP-1) is a classful (as opposed to classless) routing protocol. It supports the advertisement of protocol packets only in broadcast mode. **Figure 3-3** shows the packet format. The RIP-1 protocol packet does not carry any mask, so it can identify only the routes of the natural network segment such as Class A, Class B, and Class C, and does not support route aggregation or discontinuous subnet.

RIP version 2 (RIP-2), is a classless routing protocol. **Figure 3-4** shows the packet format.

Figure 3-3 RIP-1 packet format

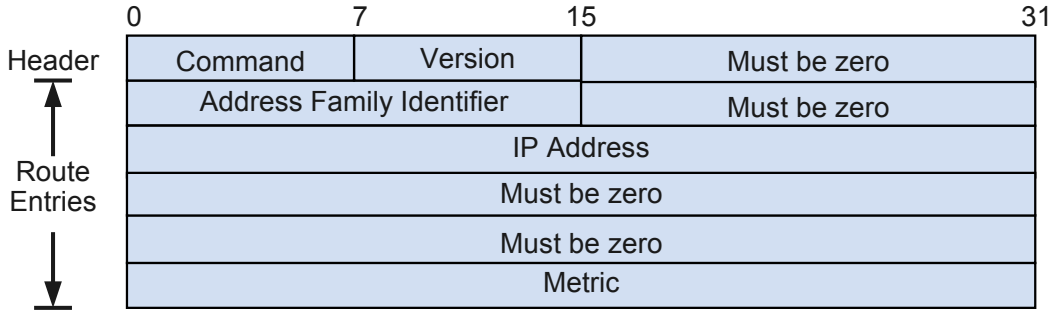
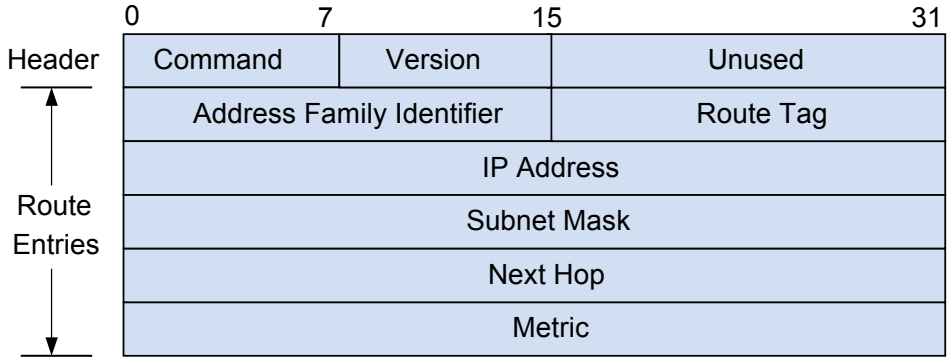


Figure 3-4 RIP-2 packet format



Compared with RIP-1, RIP-2 has the following advantages:

- Supports route tag and can flexibly control routes on the basis of the tag in the routing policy.
- Has packets that contain mask information and support route summarization and Classless Inter-domain Routing (CIDR).
- Supports the next hop address and can select the optimal next hop address in the broadcast network.
- Supports sending update packets in multicast mode. Only RIP-2 routers can receive protocol packets. This reduces resource consumption.
- Provides two authentication modes to enhance security: plain-text authentication and MD5 authentication.

RIP-2 Route Summarization

When different subnet routes in the same natural network segment are transmitted to other network segments, these routes are summarized into one route of the same segment. This process is called route summarization.

RIP-1 packets do not carry mask information, so RIP-1 can advertise only the routes with natural masks. Because RIP-2 packets carry mask information, RIP-2 supports subnetting. RIP-2 route summarization improves extensibility and efficiency and minimizes the routing table size of a large-sized network.

Route summarization is classified into two types:

- RIP process-based classful summarization

  Summarized routes are advertised using nature masks. For example, route 10.1.1.0/24 (metric=2) and route 10.1.2.0/24 (metric=3) are summarized as a route 10.0.0.0/8 (metric=2) in the natural network segment. RIP-2 supports classful summarization to obtain the optimal metric.

- Interface-based summarization

  A user can specify a summarized address. For example, a route 10.1.0.0/16 (metric=2) can be configured on the interface as a summarized route of route 10.1.1.0/24 (metric=2) and route 10.1.2.0/24 (metric=3).

3.2.3 RIPng

In addition to IPv4 networks, RIP is also applicable to IPv6 networks to provide accurate route information for IPv6 packets. IETF has defined RIP next generation (RIPng) based on RIP for IPv6 networks. RIPng is an important protocol for IPv6 networks.

Comparison Between RIPng and RIP

RIPng made the following modifications to RIP:

- RIPng uses UDP port 521 to send and receive routing information.
- RIPng uses the destination addresses with 128-bit prefixes (mask length).
- RIPng uses 128-bit IPv6 addresses as next hop addresses.
- RIPng uses the local link address FE80::/10 as the source address to send RIPng Update packets.
- RIPng periodically sends routing information in multicast mode and uses FF02::9 as multicast address.
- A RIPng packet consists of a header and multiple route table entries (RTEs). In a RIPng packet, the maximum number of RTEs depends on the MTU on the interface.

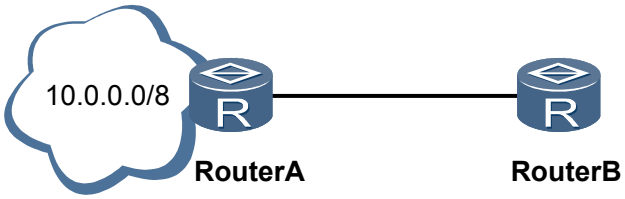
3.2.4 Split Horizon and Poison Reverse

Split Horizon

Split horizon ensures that a route learned by RIP on an interface is not sent to neighbors from the interface. This feature reduces bandwidth consumption and avoids routing loops.

Split horizon provides two models for different networks: interface-based split horizon and neighbor-based split horizon. Broadcast, P2P, and P2MP networks use interface-based split horizon, as shown in **Figure 3-5**.

Figure 3-5 Interface-based split horizon

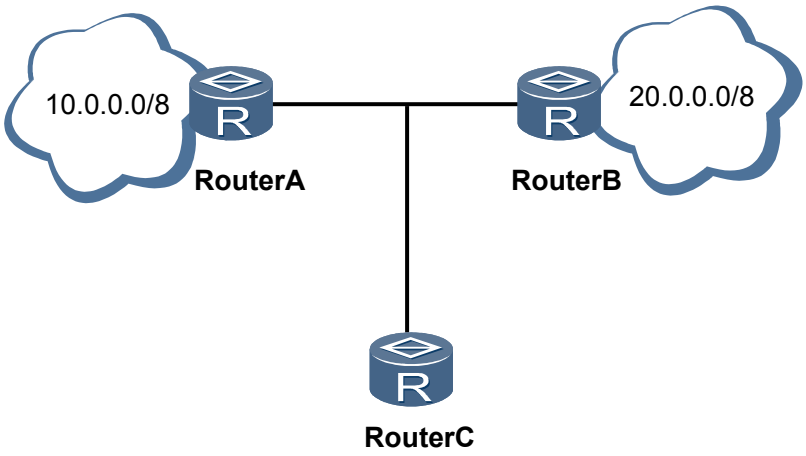


Router A sends routing information destined for 10.0.0.0/8 to router B. If split horizon is not configured, router B sends the route learned from router A back to router A. Thus router A can learn two routes destined for 10.0.0.0/8: a direct route with hop count 0 and a route with the next hop router B and hop count 2.

However, only the direct route in the RIP routing table on router A is active. When the route from router A to network 10.0.0.0 is unreachable, router B does not receive the unreachable message immediately and still notifies router A that network 10.0.0.0/8 is reachable. Therefore, router A receives incorrect routing information that network 10.0.0.0/8 is reachable through router B, and router B considers that network 10.0.0.0/8 is reachable through router A. A routing loop is thus generated. With the split horizon feature, router B does not send the route destined for 10.0.0.0/8 back to router A. Routing loops are avoided.

On a Non-Broadcast Multiple Access (NBMA) network, an interface connects to multiple neighbors; therefore, split horizon is performed based on neighbors. Routes are advertised in unicast mode. The routes received by an interface are differentiated by neighbors. The route learned from a neighbor is not sent back through the same interface.

Figure 3-6 Neighbor-based split horizon



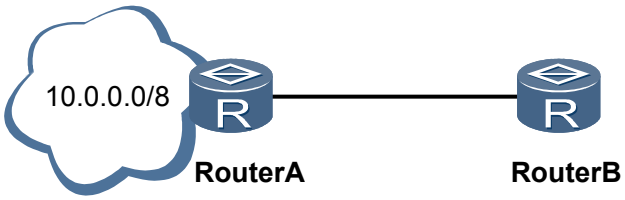
As shown in **Figure 3-6**, after split horizon is configured on an NBMA network, router A sends route 10.0.0.0/8 learned from router B to router C, but does not send it to router B.

Poison Reverse

Poison reverse ensures that RIP sets the cost of the route learned from an interface of a neighbor to 16 (unreachable) and then sends the route from the same interface back to the neighbor. This feature deletes useless routes from the routing table and avoids routing loops.

Poison reverse prevents loops.

Figure 3-7 Poison reverse



As shown in **Figure 3-7**, after receiving a route from router A, router B sends an unreachable message (with the route Cost being 16) to router A. Router A then does not learn the route from router B. A routing loop is avoided.

3.2.5 Multi-process and Multi-instance

The multi-process feature associates a RIP process with multiple interfaces, ensuring that the specific process performs all the protocol-related operations only on these interfaces. With the multi-process feature, multiple RIP processes can run on a device independently. Route exchange between RIP processes is similar to route exchange between routing protocols.

RIP multi-instance associates a VPN instance with a RIP process so that the VPN instance can be associated with all interfaces on this process.

3.2.6 RIP and BFD Association

A link fault or topology change causes routers to recalculate routes. Therefore, route convergence must be quick enough to ensure network performance. A solution to speed up route convergence is to quickly detect faults and notify routing protocols of the faults.

Bidirectional Forwarding Detection (BFD) detects faults on links between neighboring routers. Associated with a routing protocol, BFD can rapidly detect link faults and report the faults to the protocol so that the protocol quickly triggers route convergence. Traffic loss caused by topology changes is minimized. After RIP is associated with BFD, BFD rapidly detects link faults and reports the faults to RIP so that RIP quickly responds to network topology changes.

Table 3-1 lists the link fault detection mechanisms and convergence speed before and after BFD is associated with RIP.

Table 3-1 BFD speeds up convergence

RIP and BFD Association Feature	Link Fault Detection Mechanism	Convergence Speed
Disabled	The RIP age timer expires. By default, the timeout interval is 180 seconds.	Second-level (> 180 seconds)
Enabled	The BFD session goes Down.	Second-level (< 30 seconds)

Principle

BFD is classified into static BFD and dynamic BFD:

- Static BFD

  In static BFD, BFD session parameters (including local and remote discriminators) are set manually using commands, and BFD session setup requests are manually delivered.

- Dynamic BFD

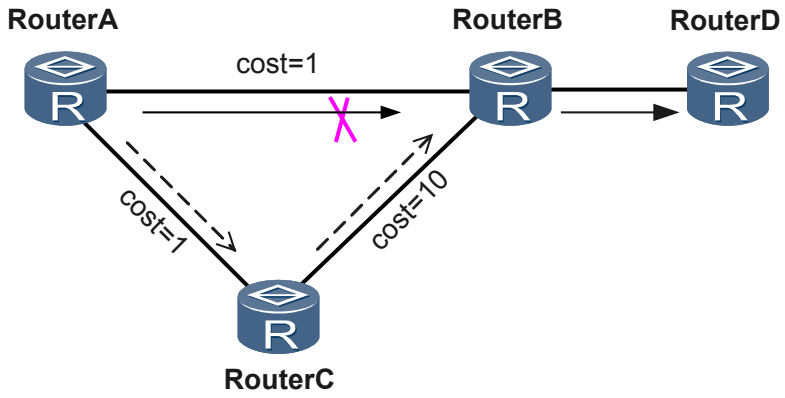
  In dynamic BFD, BFD session setup is triggered by routing protocols. The local discriminator is dynamically allocated and remote discriminator is obtained from the peer. A routing protocol notifies BFD of the neighbor parameters (including destination and source addresses), and then BFD sets up a session based on the received parameters. When a link fault occurs, the protocol associated with BFD quickly detects that the BFD session is Down, and switches traffic to the backup link. This feature minimizes data loss.

A device can implement static BFD even if the peer device does not support BFD. Dynamic BFD is more flexible than static BFD.

Application

After RIP is associated with BFD, BFD reports link faults to RIP within several milliseconds. The RIP router then deletes the faulty links from the local routing table and starts the backup link. This feature increases route convergence speed.

Figure 3-8 RIP and BFD association network



Implementation of RIP and BFD association:

- As shown in **Figure 3-8**, router A, router B, router C, and router D set up RIP neighbor relationships. Router B is the next hop on the route from router A to router D. RIP and BFD association is configured on router A and router B.
- When the link between router A and router B is faulty, BFD quickly detects the fault and notify router A of the fault. Router A deletes the route with router B as the next hop, and then recalculates a route. The new route passes router C and router B and reaches router D.
- When the link between router A and router B recovers, a session is set up again. Router A receives routing information from router B and selects the optimal route.

3.2.7 Hot Standby

Devices with distributed architecture support the RIP hot standby feature.

During hot standby, a device backs up RIP data from the active main board (AMB) to the standby main board (SMB). When the AMB becomes faulty, the SMB becomes active and takes over the AMB's tasks. This prevents RIP from being affected and ensures normal data forwarding.

3.3 References

The following table lists the references that apply in this chapter.

Document	Description	Remarks
RFC1058	Describes RIP protocol, describes the elements, characteristic, limitation of RIP version 1.	-
RFC2453	Specifies an extension of the Routing Information Protocol (RIP), as defined in [1], to expand the amount of useful information carried in RIP messages and to add a measure of security.	-
RFC 2080	This document specifies a routing protocol for an IPv6 Internet. It is based on protocols and algorithms currently in wide use on the IPv4 Internet.	-

4 OSPF

About This Chapter

4.1 Introduction to OSPF

Definition

The Open Shortest Path First (OSPF) protocol, developed by the Internet Engineering Task Force (IETF), is a link-state Interior Gateway Protocol (IGP).

At present, OSPF Version 2, defined in RFC 2328, is intended for IPv4, and OSPF Version 3, defined in RFC 2740, is intended for IPv6. Unless otherwise stated, OSPF stated in this document refers to OSPF Version 2.

Purpose

Before the emergence of OSPF, the Routing Information Protocol (RIP) is widely used on networks as an IGP.

RIP is a routing protocol based on the distance vector algorithm. Due to its slow convergence, routing loops, and poor scalability, RIP is gradually replaced by OSPF.

As a link-state protocol, OSPF can solve many problems encountered by RIP. Additionally, OSPF features the following advantages:

- Receives or sends packets in multicast mode to reduce load on the router that does not run OSPF.
- Supports Classless Interdomain Routing (CIDR).
- Supports load balancing among equal-cost routes.
- Supports packet encryption.

With the preceding advantages, OSPF is widely accepted and used as an IGP.

4.2 Principle

4.2.1 Fundamentals of OSPF

OSPF has the following functions:

- Divides an Autonomous System (AS) into one or multiple logical areas.
- Advertises routes by sending Link State Advertisements (LSAs).
- Exchanges OSPF packets between devices in an OSPF area to synchronize routing information.
- Encapsulates OSPF packets into IP packets and sends the packets in unicast or multicast mode.

Packet Type

Table 4-1 packet type

Packet Type	Function
Hello packet	Sent periodically to discover and maintain OSPF neighbor relationships.
Database Description (DD) packet	Contains brief information about the local link-state database (LSDB) and synchronizes the LSDBs on two devices.
Link State Request (LSR) packet	Requests the required LSAs from neighbors. LSR packets are sent only after DD packets are exchanged successfully.
Link State Update (LSU) packet	Sends the required LSAs to neighbors.
Link State Acknowledgement (LSAck) packet	Acknowledges the receipt of an LSA.

LSA Type

Table 4-2 LSA type

LSA Type	Function
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.
Network-summary-LSA (Type 3)	Describes routes to a specific network segment in an area. Network-summary-LSAs are generated by an Area Border Router (ABR) and advertised in all areas except totally stub areas and Not-So-Stubby Areas (NSSA Areas).
ASBR-summary-LSA (Type 4)	Describes routes to an Autonomous System Boundary Router (ASBR). ASBR-summary-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 NSSA areas.
NSSA-LSA (Type7)	Describes routes to a destination outside the AS. Generated by an ASBR and advertised in NSSAs only.

LSA Type	Function
Opaque-LSA (Type 9/Type 10/Type 11)	Provides a universal mechanism for OSPF extension. ● Type 9 LSAs are advertised only on the network segment where the interface originating Type 9 LSAs resides. Grace LSAs used to support GR are a type of Type 9 LSAs. ● Type 10 LSAs are advertised inside an OSPF area. LSAs used to support TE are a type of Type 10 LSAs. ● Type 11 LSAs are advertised within an AS. At present, there are no applications of Type 11 LSAs.

router Type

Figure 4-1 lists common router types used in OSPF.

Figure 4-1 router type

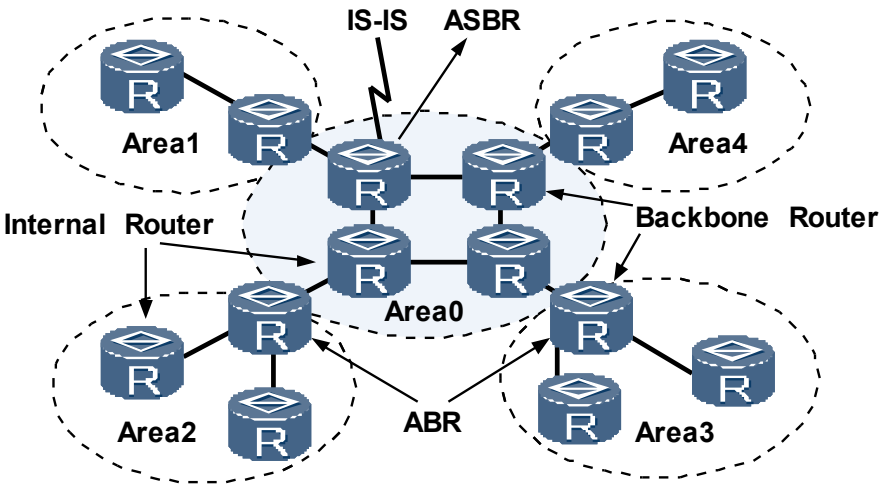


Table 4-3 router type

router Type	Description
Internal router	All interfaces on an internal router belong to the same OSPF area.
Area Border Router (ABR)	An ABR belongs to two or more than two 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.

router Type	Description
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.
ASBR (AS Boundary Router)	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 OSPF device that has imported external routing information will become an ASBR.

Route Type

Inter-area and intra-area routes in an AS describe the AS's network structure. AS external routes describe the routes to destinations outside an AS. OSPF classifies the imported AS external routes into Type 1 and Type 2 external routes.

Table 4-4 lists route types in descending priority order.

Table 4-4 route type

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 OSPF considers that the cost of the route from an ASBR to the destination of a Type 2 external route is much greater than the cost of any internal route to the ASBR. Cost of a Type 2 external route = Cost of the route from the ASBR to the destination of the Type 2 external route

Area Type

Table 4-5 area type

Area Type	Function
Common area	OSPF areas are common areas by default. Common areas include standard areas and backbone areas. ● A standard area is the most common area and transmits intra-area routes, inter-area routes, and external routes. ● A backbone area connects all the other OSPF areas. It is usually named Area 0.
Stub area	A stub area does not advertise AS external routes, but only intra-area and inter-area routes. Compared with a non-stub area, the router in a stub area maintains fewer routing entries and transmits less routing information. To ensure the reachability of AS external routes, the ABR in a stub area advertises Type 3 default routes to the entire stub area. All AS external routes must be advertised by the ABR.
Totally stub area	A totally stub area does not advertise AS external routes or inter-area routes, but only intra-area routes. Compared with a non-stub area, the router in a totally stub area maintains fewer routing entries and transmits less routing information. To ensure the reachability of AS external routes, the ABR in a totally stub area advertises Type 3 default routes to the entire totally stub area. All AS external routes must be advertised by the ABR.
NSSA area	An NSSA area can import AS external routes. An ASBR uses Type 7 LSAs to advertise the imported AS external routes to the entire NSSA area. These Type 7 LSAs are translated into Type 5 LSAs on an ABR, and are then flooded in the entire OSPF AS. An NSSA area has the characteristics of the stub areas in an AS. An ABR in an NSSA area advertises Type 3 default routes to the entire NSSA area. All inter-area routes must be advertised by the ABR.
Totally NSSA area	A totally NSSA area can import AS external routes. An ASBR uses Type 7 LSAs to advertise the imported AS external routes to the entire NSSA area. These Type 7 LSAs are translated into Type 5 LSAs on an ABR, and are then flooded in the entire OSPF AS. A totally NSSA area has the characteristics of the totally stub areas in an AS. An ABR in a totally NSSA area advertises Type 3 default routes to the entire totally NSSA area. All inter-area routes must be advertised by the ABR.

OSPF Network Type

Table 4-6 lists four OSPF network types that are classified based on link layer protocols.

Table 4-6 OSPF network type

Network Type	Description
Broadcast	A network with the link layer protocol of Ethernet or Fiber Distributed Data Interface (FDDI) is a broadcast network by default. On a broadcast network: ● Hello packets, LSU packets, and LSAck packets are usually transmitted in multicast mode. 224.0.0.5 is an IP multicast address reserved for an OSPF device. 224.0.0.6 is an IP multicast address reserved for an OSPF DR or backup designated router (BDR). ● DD and LSR packets are transmitted in unicast mode.
Non-Broadcast Multi-Access (NBMA)	A network with the link layer protocol of frame relay (FR), X.25 is an NBMA network 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.
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 network can be changed to a P2MP network. The common practice is to change a non-fully meshed NBMA network to a P2MP network. On a P2MP network: ● Hello packets are transmitted in multicast mode using the multicast address 224.0.0.5. ● 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)	By default, a network where the link layer protocol is PPP, HDLC, or LAPB is a P2P network. 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 224.0.0.5.

Stub Area

Stub areas are specific areas where ABRs do not flood the received AS external routes. In stub areas, routers maintain fewer routing entries and less routing information.

Configuring a stub area is optional. Not every area can be configured as a stub area. 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.
- Before configuring an area as a stub area, you must configure stub area attributes on all routers in the area.
- There should be no ASBR in a stub area, meaning that AS external routes cannot be transmitted in the stub area.
- Virtual connections cannot cross a stub area.

NSSA Area

NSSA areas are a special type of OSPF areas. There are many similarities between an NSSA area and a stub area. Both of them do not advertise the external routes received from the other OSPF areas. The difference is that a stub area cannot import AS external routes, whereas an NSSA area can import AS external routes and advertise the imported routes to the entire AS.

After an area is configured as an NSSA area, an ABR in the NSSA area generates a default route and advertises the route to the other routers in the NSSA area. This is to ensure the reachability of the routes to the destinations outside an AS.

Note the following points when configuring an NSSA area:

- The backbone area cannot be configured as an NSSA area.
- Before configuring an area as an NSSA area, you must configure NSSA area attributes on all routers in the area.
- Virtual connections cannot cross an NSSA area.

Neighbor State Machine

OSPF has eight state machines: Down, Attempt, Init, 2-way, Exstart, Exchange, Loading, and Full.

- Down: It is in the initial stage of setting up sessions between neighbors. The state machine is Down when a router fails to receive Hello packets from its neighbor before the dead interval expires.
- Attempt: It occurs only on an NBMA network. The state machine is Attempt when a neighbor does not reply with Hello packets after the dead interval has expired. The local router, however, keeps sending Hello packets to the neighbor at every poll interval.
- Init: The state machine is Init after a router receives Hello packets.
- 2-way: The state machine is 2-way when the Hello packets received by a router contain its own router ID. The state machine will remain in the 2-way state if no neighbor relationship is established, and will become Exstart if a neighbor relationship is established.
- Exstart: The state machine changes from Init to Exstart when the neighbor relationship is established. The two neighbors then start to negotiate the master/slave status and determine the sequence numbers of DD packets.
- Exchange: The state machine is Exchange when a router starts to exchange DD packets with its neighbor after the master/slave status negotiation is completed.

- Loading: The state machine is Loading after a router has finished exchanging DD packets with its neighbor.
- Full: The state machine is Full when the LSA retransmission list is empty.

OSPF Packet Authentication

OSPF supports packet authentication. Only the OSPF packets that have been authenticated can be received. If OSPF packets are not authenticated, a neighbor relationship cannot be established.

The AR150&AR200&AR1200&AR2200&AR3200 supports two authentication methods:

- Area-based authentication
- Interface-based authentication

The authentication modes supported by the AR150&AR200&AR1200&AR2200&AR3200 can be classified into null, simple, and MD5 based on encryption algorithms.

When both area-based and interface-based authentication methods are configured, interface-based authentication takes effect.

OSPF Route Summarization

Route summarization means that an ABR in an area summarizes the routes with the same prefix into one route and advertises the summarized route to the other areas.

Route summarization between areas reduces the amount of routing information to be transmitted, reducing the size of routing tables and improving device performance.

Route summarization can be carried out by an ABR or an ASBR:

- Route summarization on an ABR:

  When an ABR in an area advertises routing information to other areas, it generates Type 3 LSAs by network segment. If this area contains consecutive network segments, you can run a command to summarize these network segments into one network segment. The ABR only needs to send one summarized LSA, and will not send the LSAs that belong to the summarized network segment specified in the command.
- Route summarization on an ASBR:

  If the local device is an ASBR and route summarization is configured, the ASBR will summarize the imported Type 5 LSAs within the aggregated address range. After an NSSA area is configured, the ASBR needs to summarize the imported Type 7 LSAs within the aggregated address range.

  If the local device is an ASBR and ABR, the device will summarize the Type 5 LSAs that are translated from Type 7 LSAs.

OSPF Default Route

A default route is a route of which the destination address and mask are all 0s. If a router cannot find a route in its routing table for forwarding packets, it can forward packets using a default route. Due to hierarchical management of OSPF routes, the priority of default Type 3 routes is higher than the priority of default Type 5 or Type 7 routes.

OSPF default routes are usually used in the following cases:

- An ABR advertises default Type 3 Summary LSAs to instruct routers within an area to forward packets between areas.
- An ASBR advertises default Type 5 ASE LSAs or default Type 7 NSSA area LSAs to instruct routers in an AS to forward packets to other ASs.

Principles for advertising OSPF default routes are described below:

- An OSPF router can advertise LSAs carrying default route information only when it has an interface connected to an upper-layer network.
- If an OSPF router has advertised an LSA carrying information about a type of default route, the OSPF router does not learn this type of default routes advertised by other routers. This means that the OSPF router no longer calculates routes based on the LSAs carrying information about the same type of the default routes advertised by other routers, but stores these LSAs in its LSDB.
- The route on which default external route advertisement depends cannot be a route in the local OSPF AS. This means that the route cannot be the one learned by the local OSPF process. This is because default external routes are used to guide packet forwarding outside an AS, whereas the routes within an AS have the next hop pointing to the devices within the AS.

Table 4-7 lists principles for advertising default routes in different areas.

Table 4-7 Principles for advertising OSPF default routes

Area Type	Function
Common area	By default, devices in a common OSPF area do not automatically generate default routes, even if the common OSPF area has default routes. When a default route on the network is generated by another routing process (not OSPF process), the device that generates the default route must advertise the default route in the entire OSPF AS. (Run a command on an ASBR to configure the ASBR to generate a default route. After the configuration, the ASBR generates a default Type 5 ASE LSA and advertises the LSA to the entire OSPF AS.)
STUB area	A stub area does not allow AS external routes (Type 5 LSAs) to be transmitted within the area. All routers within the stub area must learn AS external routes from the ABR. The ABR automatically generates a default Summary LSA (Type 3 LSA) and advertises it to the entire stub area. Then all routes to destinations outside an AS can be learned from the ABR.
Totally STUB area	A totally stub area does not allow AS external routes (Type 5 LSAs) or inter-area routes (Type 3 LSAs) to be transmitted within the area. All routers within the totally stub area must learn AS external routes and other areas' routes from the ABR. The ABR automatically generates a default Summary LSA (Type 3 LSA) and advertises it to the entire totally stub area. Then, all routes to destinations outside an AS and to destinations in other areas can be learned from the ABR.

Area Type	Function
NSSA area	An NSSA area allows its ASBRs to import a small number of AS external routes, but does not advertise ASE LSAs received from other areas within the NSSA area. This means that AS external routes can be learned only from ASBRs in the NSSA area. Devices in an NSSA area do not automatically generate default routes. Use either of the following methods as required: ● To advertise some external routes using the ASBR in the NSSA area and advertise other external routes through other areas, configure a default Type 7 LSA on the ABR and advertise this LSA in the entire NSSA area. ● To advertise all the external routes using the ASBR in the NSSA area, configure a default Type 7 LSA on the ASBR and advertise this LSA in the entire NSSA area. The difference between these two configurations is described below: ● An ABR will generate a default Type 7 LSA regardless of whether the routing table contains the default route 0.0.0.0. ● An ASBR will generate a default Type 7 LSA only when the routing table contains the default route 0.0.0.0. A default route is flooded only in the local NSSA area and is not flooded in the entire OSPF AS. If routers in the local NSSA area cannot find routes to the outside of the AS, the routers can forward packets to the outside of the AS through an ASBR. Packets of other OSPF areas, however, cannot be sent to the outside of the AS through this ASBR. Default Type 7 LSAs will not be translated into default Type 5 LSAs and flooded in the entire OSPF AS.
Totally NSSA area	A totally NSSA area does not allow AS external routes (Type 5 LSAs) or inter-area routes (Type 3 LSAs) to be transmitted within the area. All routers within the totally NSSA area must learn AS external routes from the ABR. The ABR automatically generates a default Summary LSAs and advertises it to the entire totally NSSA area. Then all external routes received from other areas and inter-area routes can be advertised within the totally NSSA area.

OSPF Route Filtering

OSPF supports route filtering using routing policies. By default, OSPF does not filter routes.

Routing policies used by OSPF include the route-policy, access-list, and prefix-list. For details on routing policy description, see the *Huawei AR150&AR200&AR1200&AR2200&AR3200 Series Enterprise Routers Feature Description - Routing Policies*

OSPF route filtering can be used for:

- Importing routes

OSPF can import routes learned by other routing protocols. You can configure routing policies to filter the imported routes to allow OSPF to import only the routes that match specific conditions.

- Advertising imported routes

  OSPF advertises the imported routes to its neighbors.

  You can configure filtering rules to filter the routes to be advertised. The filtering rules can be configured only on ASBRs.

- Learning routes

  Filtering rules can be configured to allow OSPF to filter the received intra-area, inter-area, and AS external routes.

  After receiving routes, an OSPF device adds only the routes that match the filtering rules to the local routing table, but can still advertise all routes from the OSPF routing table.

- Learning inter-area LSAs

  You can run a command to configure an ABR to filter the incoming Summary LSAs. This configuration takes effect only on ABRs because only ABRs can advertise Summary LSAs.

  Table 4-8 Differences between inter-area LSA learning and route learning

Inter-area LSA Learning	Route Learning
Directly filters the incoming LSAs.	Filters the routes that are calculated based on LSAs, but does not filter LSAs. This means that all incoming LSAs are learned.

- Advertising inter-area LSAs

  You can run a command to configure an ABR to filter the outgoing Summary LSAs. This configuration takes effect only on ABRs.

OSPF Multi-Process

OSPF supports multi-process. Multiple OSPF processes can run on the same router, and they are independent of each other. Route exchanges between different OSPF processes are similar to route exchanges between different routing protocols.

Each interface on the router belongs to only one OSPF process.

A typical application of OSPF multi-process is that OSPF runs between PEs and CEs in a VPN, whereas OSPF is used as an IGP on the backbone of the VPN. Two OSPF processes on the same PE are independent of each other.

OSPF RFC 1583 Compatibility

RFC 1583 is an earlier version of OSPFv2.

When OSPF calculates external routes, routing loops may occur because RFC 2328 and RFC 1583 define different route selection rules. To prevent routing loops, both communication ends must use the same route selection rules.

- After RFC 1583 compatibility is enabled, OSPF use the route selection rules defined in RFC 1583.

- When RFC 1583 compatibility is disabled, OSPF uses the route selection rules defined in RFC 2328.

OSPF calculates external routes based on Type 5 LSAs. If the router enabled with RFC 1583 compatibility receives a Type 5 LSA:

- The router selects a route to the ASBR that originates the LSA, or to the forwarding address (FA) described in the LSA.
- The router selects external routes to the same destination.

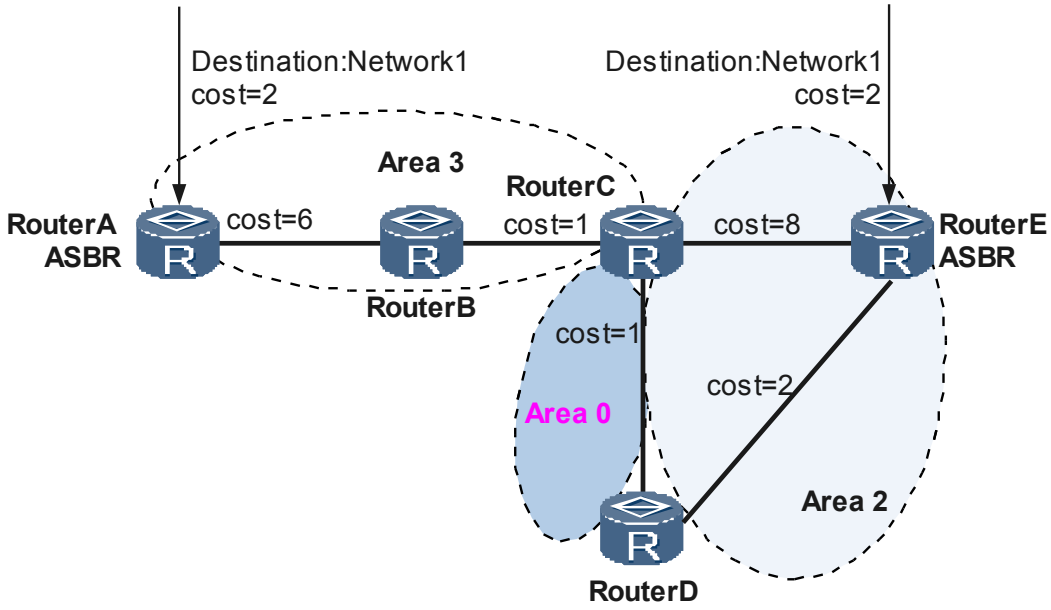
By default, OSPF uses the route selection rules defined in RFC 1583.

4.2.2 OSPF TE

OSPF Traffic Engineering (TE) is a new feature developed on the basis of OSPF to support MPLS TE and establish and maintain the Label Switch Path (LSP) of TE. In the MPLS TE architecture described in "MPLS Feature Description", OSPF functions as the information advertising component, responsible for collecting and advertising MPLS TE information.

In addition to the network topology, TE also needs to know network constraints, such as the bandwidth, TE metric, administrative group, and affinity attribute. Current OSPF functions, however, cannot meet these requirements. Therefore, OSPF needs to be extended by introducing a new type of LSAs to advertise network constraints. Based on the network constraints, the Constraint Shortest Path First (CSPF) algorithm can calculate the path that satisfies certain constraints.

Figure 4-2 Function of OSPF in the MPLS TE architecture



Function of OSPF in the MPLS TE Architecture

In the MPLS TE architecture, OSPF functions as the information advertising component:

- Collects related information about TE.
- Floods TE information to devices in the same area.
- Uses the collected TE information to form the TE database (TEDB) and provides it for CSPF to calculate routes.

OSPF does concern with what the specific information is or how MPLS uses the information.

TE-LSA

OSPF uses a new type of LSAs, namely, Type 10 opaque LSAs, to collect and advertise TE information. This type of LSAs contain the link status information required by TE, including the maximum link bandwidth, maximum reservable bandwidth, current reserved bandwidth, and link color. Type 10 opaque LSAs synchronize link status information among devices in an area through the OSPF flooding mechanism. By so doing, a uniform TEDB is formed for route calculation.

Interaction Between OSPF TE and CSPF

OSPF collects TE information in an area by using Type 10 LSAs, including the bandwidth, priority, and link metric. After processing the collected TE information, OSPF provides it for CSPF to calculate routes.

IGP Shortcut and Forwarding Adjacency

OSPF supports IGP shortcut and forwarding adjacency. The two features allow OSPF to use a tunnel interface as an outgoing interface to reach a destination.

Differences between IGP shortcut and forwarding adjacency are as follows:

- A device enabled with IGP shortcut uses a tunnel interface as an outgoing interface, but it does not advertise the link of the tunnel interface to neighbors. Therefore, other devices cannot use this tunnel.
- A device enabled with forwarding adjacency uses a tunnel interface as an outgoing interface, and advertises the tunnel interface to neighbors. Therefore, other devices can use this tunnel.
- IGP shortcut is unidirectional and needs to be configured only on the device that uses IGP shortcut.

OSPF DS-TE

DiffSer Aware Traffic Engineering (DS-TE) controls and forwards flows differently based on Class of Service (CoS). DS-TE combines the advantages of MPLS TE and Differentiated Services (DiffServ) and controls flow paths precisely. By so doing, DS-TE effectively uses network resources and reserves required resources for different service flows. For details, refer to the chapter "MPLS" in this manual.

To support DS-TE in MPLS, OSPF supports the local overbooking multiplier TLV and bandwidth constraint (BC) TLV in the TE-LSA, which are used to advertise and collect the reservable bandwidths of class types (CTs) with different priorities on the link (A CT refers to a collection of bandwidths of an LSP or a group of LSPs with the same CoS.)

OSPF SRLG

OSPF supports the applications of the Shared Risk Link Group (SRLG) in MPLS by obtaining information about the SRLG that floods TE information to devices in an area. For details, refer to the chapter "MPLS" in this manual.

4.2.3 BFD for OSPF

Definition

Bidirectional Forwarding Detection (BFD) is a mechanism to detect communication faults between forwarding engines.

To be specific, BFD detects connectivity of a data protocol on a path between two systems. The path can be a physical link, a logical link, or a tunnel.

In BFD for OSPF, a BFD session is associated with OSPF. The BFD session quickly detects a link fault and then notifies OSPF of the fault. This speeds up OSPF's response to the change of the network topology.

Purpose

The link fault or the topology change may cause devices to re-calculate routes. Therefore, the convergence of routing protocols must be as quick as possible to improve the network performance.

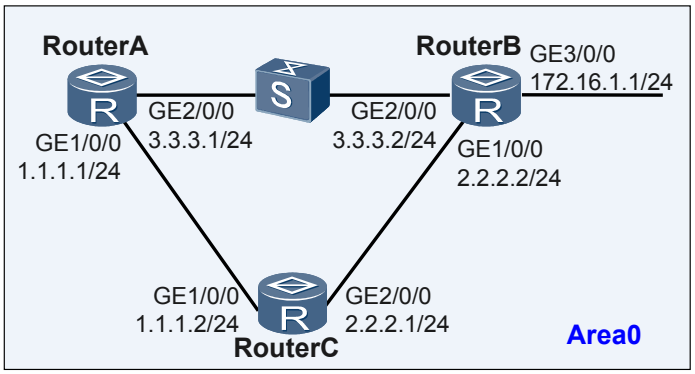
Link faults are unavoidable. Therefore, a feasible solution is required to detect faults faster and notify the faults to routing protocols immediately. If BFD is associated with OSPF, once a fault occurs on a link between neighbors, BFD can speed up the OSPF convergence.

Table 4-9 Comparison before and after BFD for OSPF is enabled

Associated with BFD or Not	Link Fault Detection Mechanism	Convergence Speed
Not associated with BFD	An OSPF Dead timer expires. By default, the timeout period of the timer is 40s.	At the second level
Associated with BFD	A BFD session goes Down.	At the millisecond level

Principle

Figure 4-3 BFD for OSPF



The principle of BFD for OSPF is shown in **Figure 4-3**.

1. OSPF neighbor relationships are established between these three routers.
2. After a neighbor relationship becomes Full, this triggers BFD to establish a BFD session.
3. The outbound interface on Router A connected to Router B is GE 2/0/0. If the link fails, BFD detects the fault and then notifies Router A of the fault.
4. Router A processes the event that a neighbor relationship becomes Down and re-calculates routes. After calculation, the outbound interface is GE1 /0/0 passes through Router C and then reaches Router B.

4.2.4 OSPF GTSM

Definition

GTSM is short for Generalized TTL Security Mechanism, a mechanism that protects the services over the IP layer by checking whether the TTL value in the IP packet header is within a pre-defined range.

Purpose

On the network, an attacker may simulate valid OSPF packets and keeps sending them to a device. After receiving these packets, the device identifies the destination of the packets. The forwarding plane of the device then directly sends the packets to the control plane for processing without checking the validity of the packets. As a result, the device is busy processing these "valid" packets, resulting in high CPU usage.

In applications, the GTSM is mainly used to protect the TCP/IP-based control plane from CPU-utilization based attacks, for example, attacks that cause CPU overload.

Principle

Devices enabled with GTSM check the TTL values in all the received packets according to the configured policies. The packets that fail to pass the policies are discarded or sent to the control

plane. This prevents devices from possible CPU-utilization based attacks. A GTSM policy involves the following items:

- Source address of the IP packet sent to the device
- VPN instance to which the packet belongs
- Protocol number of the IP packet (89 for OSPF, and 6 for BGP)
- Source interface number and destination interface number of protocols above TCP/UDP
- Valid TTL range

The method of implementing GTSM is as follows:

- For the directly connected OSPF neighbors, the TTL value of the unicast protocol packets to be sent is set to 255.
- For multi-hop neighbors, a reasonable TTL range is defined.

The applicability of GTSM is as follows:

- GTSM is effective with unicast packets rather than multicast packets. This is because the TTL file of multicast packets can only be 255, and therefore GTSM is not needed to protect against multicast packets.
- GTSM does not support tunnel-based neighbors.

4.2.5 OSPF Smart-discover

Definition

Generally, routers periodically send Hello packets through OSPF interfaces. That is, a router sends Hello packets at the Hello interval set by a Hello timer. Because Hello packets are sent at a fixed interval, the speed at which OSPF neighbor relationship is established is lowered.

Enabling Smart-discover can speed up the establishment of OSPF neighbor relationships in specific scenarios.

Table 4-10 OSPF Smart-discover

Smart-discover Configured or Not	Processing
Smart-discover is not configured	● Hello packets are sent only when the Hello timer expires. ● The gap between the sending of two Hello packets is the Hello interval. ● Neighbors keep waiting to receive Hello packets within the Hello interval.
Smart-discover is configured	● Hello packets are sent directly regardless of whether the Hello timer expires. ● Neighbors can receive packets rapidly and perform status transition immediately.

Principle

In the following scenarios, the interface enabled with Smart-discover can send Hello packets to neighbors without having to wait for the Hello timer to expire:

- The neighbor status becomes 2-way for the first time.
- The neighbor status changes from 2-way or a higher state to Init.

4.2.6 OSPF VPN

Definition

As an extension of OSPF, OSPF VPN multi-instance enables Provider Edges (PEs) and Customer Edges (CEs) in VPNs to run OSPF for interworking and use OSPF to learn and advertise routes.

Purpose

As a widely used IGP, in most cases, OSPF runs in VPNs. If OSPF runs between PEs and CEs, and PEs advertise VPN routes to CEs using OSPF, CEs do not need to support other routing protocols for interworking with PEs. This simplifies management and configuration of CEs.

Running OSPF Between PEs and CEs

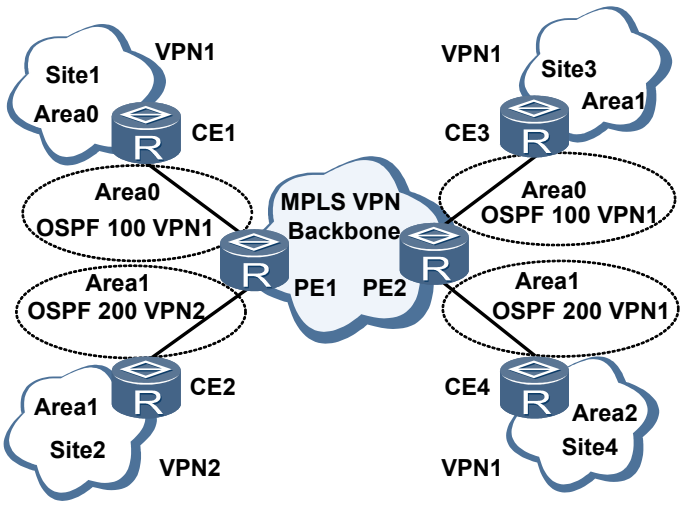
In BGP/MPLS VPN, routing information is transmitted between PEs using Multi-Protocol BGP (MP-BGP), whereas routes are learned and advertised between PEs and CEs using OSPF.

Running OSPF between PEs and CEs has the following benefits:

- OSPF is used in a site to learn routes. Running OSPF between PEs and CEs can reduce the protocol types that CEs must support, reducing the requirements for CEs.
- Similarly, running OSPF both in a site and between PEs and CEs simplifies the workload of network administrators. In this manner, network administrators do not have to be familiar with multiple protocols.
- When a network using OSPF but not VPN on the backbone network begins to use BGP/MPLS VPN, running OSPF between PEs and CEs facilitates the transition.

As shown in **Figure 4-4**, CE1, CE3, and CE4 belong to VPN 1, and the numbers following OSPF refer to the process IDs of multiple OSPF instances running on PEs.

Figure 4-4 Running OSPF between PEs and CEs



The process of advertising routes of CE1 to CE3 and CE4 is as follows:

1. PE1 imports OSPF routes of CE1 into BGP and forms BGP VPNv4 routes.
2. PE1 advertises BGP VPNv4 routes to PE2 using MP-BGP.
3. PE2 imports BGP VPNv4 routes into OSPF, and then advertises these routes to CE3 and CE4.

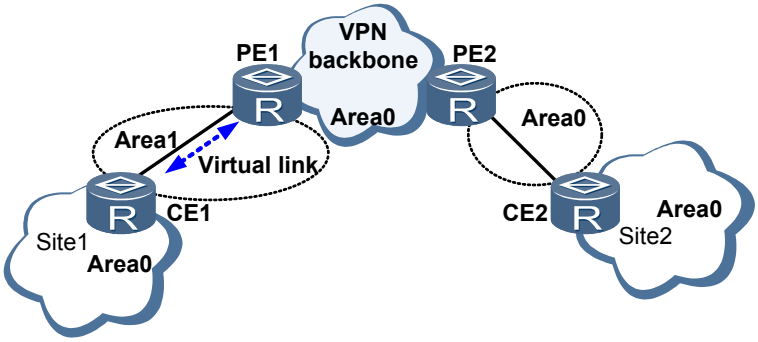
The process of advertising routes of CE4 or CE3 to CE1 is the same as the preceding process.

Configuring OSPF Areas Between PEs and CEs

OSPF areas between PEs and CEs can be either non-backbone areas or backbone areas (Area 0). A PE can only be an area border router (ABR).

In the extended application of OSPF VPN, the MPLS VPN backbone network serves as Area 0. OSPF requires that Area 0 be contiguous. Therefore, Area 0 of all VPN sites must be connected to the MPLS VPN backbone network. If a VPN site has OSPF Area 0, the PEs that CEs access must be connected to the backbone area of this VPN site through Area 0. If no physical link is available to directly connect PEs to the backbone area, a virtual link can be used to implement logical connection between the PEs and the backbone area, as shown in **Figure 4-5**.

Figure 4-5 Configuring OSPF areas between PEs and CEs



A non-backbone area (Area 1) is configured between PE1 and CE1, and a backbone area (Area 0) is configured in Site 1. As a result, the backbone area in Site 1 is separated from the VPN backbone area. Therefore, a virtual link is configured between PE1 and CE1 to ensure that the backbone area is contiguous.

OSPF Domain ID

If inter-area routes are advertised between local and remote OSPF areas, these areas are considered to be in the same OSPF domain.

- Domain IDs identify and differentiate different domains.
- Each OSPF domain has one or more domain IDs, one of which is a primary ID with the others being secondary IDs.
- If an OSPF instance does not have a specific domain ID, its ID is considered as null.

Before advertising the remote routes sent by BGP to CEs, PEs need to determine the type of OSPF routes (Type 3, Type 5 or Type 7) to be advertised to CEs according to domain IDs.

- If local domain IDs are the same as or compatible with remote domain IDs in BGP routes, PEs advertise Type 3 routes.
- Otherwise, PEs advertise Type 5 or Type 7 routes.

Table 4-11 Domain ID

Comparison Between Local and Remote Domain IDs	Local and Remote Domain IDs the Same Or Not	Route Type
Both the local and remote domain IDs are null.	The same	Inter-area route
The remote domain ID is the same as the local primary domain ID or one of the local secondary domain IDs.	The same	Inter-area route
The remote domain ID is different from the local primary domain ID or any of the local secondary domain IDs.	Not the same	If the local area is a non-NSSA, external routes are generated. If the local area is an NSSA, NSSA routes are generated.

Disabling Routing Loop Prevention

CAUTION

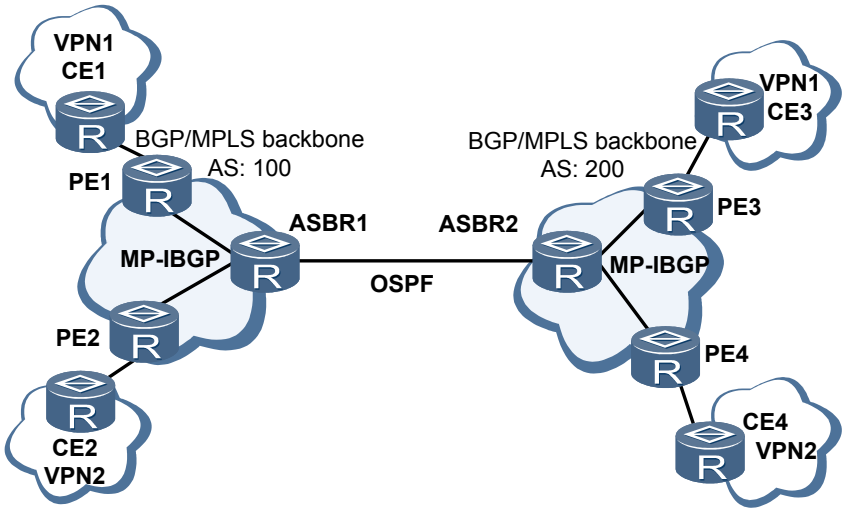
Disabling routing loop prevention may cause routing loops. Exercise caution when performing this operation.

During BGP or OSPF route exchanges, routing loop prevention prevents OSPF routing loops in VPN sites.

In the inter-AS VPN Option A scenario, if OSPF is running between ASBRs to transmit VPN routes, the remote ASBR may be unable to learn the OSPF routes sent by the local ASBR due to the routing loop prevention mechanism.

As shown in **Figure 4-6**, inter-AS VPN Option A is deployed. OSPF is running between PE1 and CE1. CE1 sends VPN routes to CE2.

Figure 4-6 Networking diagram for inter-AS VPN Option A



1. PE1 learns routes to CE1 using the OSPF process in a VPN instance, and imports these routes into MP-BGP, and sends the MP-BGP routes to ASBR1.
2. After having received the MP-BGP routes, ASBR1 imports the routes into the OSPF process in a VPN instance and generates Type 3, Type 5, or Type 7 LSAs in which the DN bit is set to 1.
3. ASBR2 learns these LSAs using OSPF and checks the DN bit of each LSA. After learning that the DN bit in each LSA is set to 1, ASBR2 does not add the routing information carried in these LSAs to its routing table.

Due to the routing loop prevention mechanism, ASBR2 cannot learn the OSPF routes sent from ASBR1, causing CE1 to be unable to communicate with CE3.

To address the preceding problem, use either of the following methods:

- A device does not set the DN bit to 1 in the LSAs when importing BGP routes into OSPF. For example, ASBR1 does not set the DN bit to 1 when importing MP-BGP routes into OSPF. After ASBR2 receives these routes and checks that the DN bit in the LSAs carrying these routes is 0, ASBR2 adds the routes to its routing table.
- A device does not check the DN bit after having received LSAs. For example, ASBR1 sets the DN bit to 1 in LSAs when importing MP-BGP routes into OSPF. ASBR2, however, does not check the DN bit after having received these LSAs.

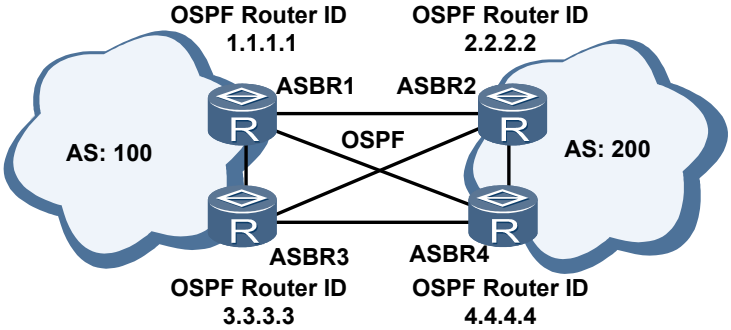
The preceding methods can be used more flexibly based on specific types of LSAs. For Type 3 LSAs, you can configure a sender to determine whether to set the DN bit to 1 or configure a

receiver to determine whether to check the DN bit in the Type 3 LSAs based on the router ID of the device that generates the Type 3 LSAs.

In the inter-AS VPN Option A scenario shown in **Figure 4-7**, the four ASBRs are fully meshed and run OSPF. ASBR2 may receive the Type 3, Type 5, or Type 7 LSAs generated on ASBR4. If ASBR2 is not configured to check the DN bit in the LSAs, ASBR2 will accept the Type 3 LSAs, and routing loops will occur, as described in **Figure 4-7**. ASBR2 will deny the Type 5 or Type 7 LSAs, because the VPN route tags carried in the LSAs are the same as the default VPN route tag of the OSPF process on ASBR2.

To address the routing loop problem caused by Type 3 LSAs, configure ASBR2 not to check the DN bit in the Type 3 LSAs that are generated by devices with the router ID 1.1.1.1 and the router ID 3.3.3.3. After the configuration is complete, if ASBR2 receives Type 3 LSAs sent by ASBR4 with the router ID 4.4.4.4, ASBR2 will check the DN bit and deny these Type 3 LSAs because the DN bit is set to 1.

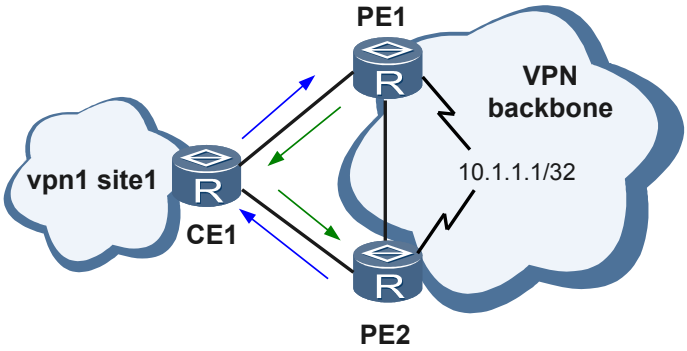
Figure 4-7 Networking diagram for full-mesh ASBRs in the inter-AS VPN Option A scenario



Routing Loop Prevention

Between PEs and CEs, routing loops may occur when OSPF and BGP learn routes from each other.

Figure 4-8 OSPF VPN routing loops



As shown in **Figure 4-8**, on PE1, OSPF imports a BGP route whose destination address is 10.1.1.1/32, and then generates and advertises a Type 5 or Type 7 LSA to CE1. Then, CE1 learns

an OSPF route with the destination address and next hop being 10.1.1.1/32 and PE1 respectively, and advertises the route to PE2. In this manner, PE2 learns an OSPF route with the destination address and next hop being 10.1.1.1/32 and CE1 respectively.

Similarly, CE1 also learns an OSPF route with the destination address and next hop being 10.1.1.1/32 and PE2 respectively. PE1 learns an OSPF route with the destination address and next hop being 10.1.1.1/32 and CE1 respectively.

As a result, CE1 has two equal-cost routes with next hops being PE1 and PE2 respectively, and the next hops of the routes from PE1 and PE2 to 10.1.1.1/32 are CE1. Thus, a routing loop occurs.

In addition, the preference of an OSPF route is higher than that of a BGP route. Therefore, on PE1 and PE2, BGP routes to 10.1.1.1/32 are replaced by the OSPF route. That is, the OSPF route with the destination address and next hop being 10.1.1.1/32 and CE1 respectively is active in the routing tables of PE1 and PE2.

The BGP route then becomes inactive, and thus the LSA generated when this route is imported by OSPF is deleted. This causes the OSPF route to be withdrawn. As a result, there is no OSPF route in the routing table, and the BGP route becomes active again. This cycle causes route flapping.

OSPF VPN provides a solution to this problem, as shown in **Table 4-12**.

Table 4-12 Routing loop prevention

Feature	Definition	Function
DN-bit	To prevent routing loops, an OSPF multi-instance process uses one bit as a flag bit, which is called the DN-bit.	When advertising the generated Type 3, Type 5, or Type 7 LSAs to CEs, PEs set the DN-bit of these LSAs to 1 and the DN-bit of other LSAs to 0. When calculating routes, the OSPF multi-instance process of a PE ignores the LSAs with the DN-bit being 1. This avoids routing loops that occur when PEs learn the self-originated LSAs from CEs.
VPN Route Tag	The VPN route tag is carried in Type 5 or Type 7 LSAs generated by PEs according to the received BGP private route. Not transmitted in BGP extended community attributes, the VPN route tag is valid only on the PEs that receive BGP routes and generate OSPF LSAs.	When a PE detects that the VPN route tag in the incoming LSA is the same as that in the local LSA, the PE ignores this LSA. Consequently, routing loops are avoided.

Feature	Definition	Function
Default Route	A route with the destination address and mask being all 0s is a default route.	PEs do not calculate default routes. Default routes are used to forward the traffic from CEs or the sites where CEs reside to the VPN backbone network.

Multi-VPN-Instance CE

OSPF multi-instance generally runs on PEs. The devices that run OSPF multi-instance within the LANs of users are called Multi-VPN-Instance CEs (MCEs), that is, multi-instance CEs.

Compared with OSPF multi-instance running on PEs, MCEs have the following characteristics:

- MCEs do not need to support OSPF-BGP synchronization.
- MCEs establish different OSPF instances for different services. Different virtual CEs transmit different services. This solves the security issue of the LAN at a low cost.
- MCEs implement different OSPF multi-instances on a CE. The key to implementing MCEs is to disable loop detection and calculate routes directly. MCEs also need to use the received LSAs with the ND-bit for route calculation.

4.2.7 OSPF NSSA

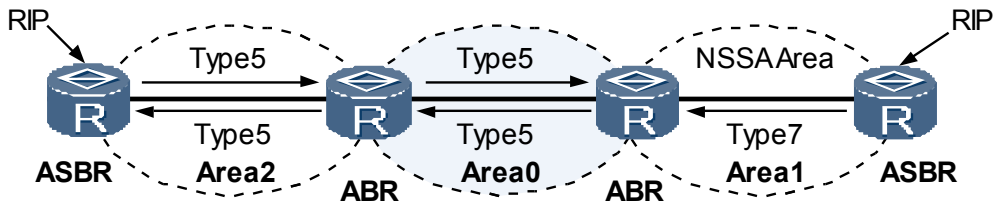
Definition

As defined in OSPF, stub areas cannot import external routes. This prevents a large number of external routes from consuming bandwidth and storage resources of the routers in stub areas. To import external routes and to prevent external routes from consuming resources, NSSAs are used, because stub areas cannot meet requirements.

NSSAs are a new type of OSPF areas.

There are many similarities between NSSAs and stub areas. The difference between NSSAs and stub areas is that NSSAs can import AS external routes into the entire OSPF AS and advertise the imported routes in the OSPF AS, but do not learn external routes from other areas on the OSPF network.

Figure 4-9 NSSA



N-bit

All routers in an area must be configured with the same area type. In OSPF, the N-bit is carried in a Hello packet and is used to identify the area type supported by the router. OSPF neighbor relationships cannot be established between routers configured with different area types.

Some manufacturers do not comply with the standard and set the N-bit in both OSPF Hello and DD packets. To allow Huawei devices to interwork with these manufacturers' devices, set the N-bit in OSPF DD packets on Huawei devices.

Type 7 LSA

- Type 7 LSAs are a new type of LSAs that can only be used in NSSAs and describe the imported external routes.
- Type 7 LSAs are generated by ASBRs in an NSSA and flooded only in the NSSA where the ASBRs reside.
- When the ABRs in the NSSA receive these Type 7 LSAs, they translate some of the Type 7 LSAs into Type 5 LSAs to advertise AS external routes to the other areas on the OSPF network.

Translating Type 7 LSAs Into Type 5 LSAs

To advertise the external routes imported by an NSSA to other areas, Type 7 LSAs need to be translated into Type 5 LSAs so that the external routes can be advertised on the entire OSPF network.

- The Propagate bit (P-bit) in a Type 7 LSA is used to instruct the router whether to translate Type 7 LSAs into Type 5 LSAs.
- By default, the ABR with the largest router ID in an NSSA is responsible for translating Type 7 LSAs into Type 5 LSAs.
- Only the Type 7 LSAs in which the P-bit is set to 1 and the FA is not 0 can be translated into Type 5 LSAs. The FA indicates that the packet to a specific destination address will be forwarded to the address specified by the FA.
- The P-bit in the Type 7 LSAs generated by ABRs is not set to 1.

Preventing Loops Caused by Default Routes

There may be multiple ABRs in an NSSA. To prevent routing loops, these ABRs not to calculate default routes advertised by each other.

4.2.8 OSPF Fast Convergence

OSPF fast convergence is an extended feature of OSPF to speed up route convergence. The characteristics of OSPF fast convergence are as follows:

- **4.2.10 Priority-based OSPF Convergence**
- When certain routes on the network change, only the changed routes are recalculated. This is called Partial Route Calculation (PRC).
- An intelligent timer is used to implement LSA management (the generating and receiving of LSAs). With the intelligent timer, infrequent changes are responded to quickly, whereas frequent changes are suppressed as desired.

To avoid excessive consumption of device resources by network connections or due to frequent route flapping, RFC 2328 maintains that:

- After an LSA is generated, it cannot be generated again in five seconds. That is, the interval for updating LSAs is one second.
- The interval for receiving LSAs is one second.

On a stable network where routes need to be fast converged, you can use the intelligent timer to set the interval for receiving LSAs to 0 seconds. This ensures that topology or route changes can be advertised to the network or be immediately sensed, thus speeding up route convergence on the network.

- Route calculation is controlled through the intelligent timer.

  When the network topology changes, devices need to recalculate routes according to OSPF. This means that frequent changes in the network topology affect the performance of devices. To address issue, RFC 2328 requires the use of a delay timer in route calculation so that route calculation is performed only after the specified delay. But the delay suggested by RFC is a fixed value, and cannot ensure both fast response to topology changes and effective suppression of flapping.

  By means of the intelligent timer, the delay in route calculation can be flexibly set as desired. As a result, infrequent changes are responded to quickly, whereas frequent changes are suppressed as desired.

- **4.2.5 OSPF Smart-discover**

4.2.9 OSPF NSR

Non-Stop Routing (NSR) is a routing technique that prevents a neighbor from sensing the fault on the control plane of a device that provides a slave control plane. With NSR, when the control plane of the device becomes faulty, the neighbor relationship set up through specific routing protocols, MPLS, and other protocols that carry services are not interrupted.

As networks develop at a fast pace, operators are having increasingly higher requirements for reliability of IP networks. NSR, as a high availability (HA) solution, is introduced to ensure that services transmitted by a device are not affected when a hardware or software failure occurs on the device.

OSPF NSR synchronizes the protocol data on the master MPU/SRU to the slave MPU/SRU in real time. When the master MPU/SRU becomes faulty or needs to be upgraded, the slave MPU/SRU rapidly takes over services from the master MPU/SRU without being sensed by the neighbor. OSPF NSR synchronizes the real-time data between the master and slave MPUs/SRUs in the following manners:

- 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), and Traffic Engineering DataBases (TEDBs). All these can be restored through the source data by using the database backup process.
- When the master-slave switchover occurs, the new master MPU/SRU restores the operation data and takes over services from the former master MPU/SRU without being sensed by the neighbor.

4.2.10 Priority-based OSPF Convergence

Priority-based OSPF convergence ensures that specific routes converge first when a great number of routes need to converge. Different routes can be set with different convergence

priorities. This allows important routes to converge first and therefore improves network reliability.

By using priority-based OSPF convergence, you can assign a higher convergence priority to routes for key services so that those routes can converge fast. By so doing, the impact on key services is reduced.

4.2.11 OSPF IP FRR

OSPF IP Fast Reroute (FRR) is dynamic IP FRR in which a backup link is pre-computed by an OSPF based on the LSDBs on the entire network. The backup link is stored in the forwarding table to protect traffic in the case of failures. In this manner, the failure recovery time can be reduced to less than 50 ms.

OSPF IP FRR complies with RFC 5286, that is, Basic Specification for IP Fast Reroute Loop-Free Alternates, which protects traffic when links or nodes become faulty.

Background

With the development of networks, Voice over IP (VoIP) and online video services require high-quality real-time transmission. Nevertheless, if an OSPF fault occurs, multiple processes, including fault detection, LSP update, LSP flooding, route calculation, and FIB entry delivery, must be performed to switch traffic to a new link. As a result, the fault recovery time is much greater than 50 ms, the time for users to sense traffic interruption, which cannot meet the requirement for real-time services.

Implementation Principle

OSPF IP FRR pre-computes a backup link by using the Loop-Free Alternate (LFA) algorithm, and then adds the backup link and the primary link to the forwarding table. In the case of failures, OSPF IP FRR can fast switch traffic to the backup link before routes on the control plane converge. This prevents traffic interruption and thus protects traffic and improves reliability of an OSPF network. The AR150&AR200&AR1200&AR2200&AR3200 supports IPv4 OSPF IP FRR.

In the LFA algorithm, considering a neighbor that can provide a backup link as the root node, the neighbor computes the shortest path from itself to the destination of the primary link by using the SPF algorithm. The neighbor then computes a loop-free backup link with the smallest cost by using the inequality defined in RFC 5286.

OSPF IP FRR can filter backup routes that need to be added to the IP routing table. Only the backup routes that are filtered through the filtering policy are added to the IP routing table. In this manner, users can flexibly manage the addition of OSPF backup routes to the IP routing table.

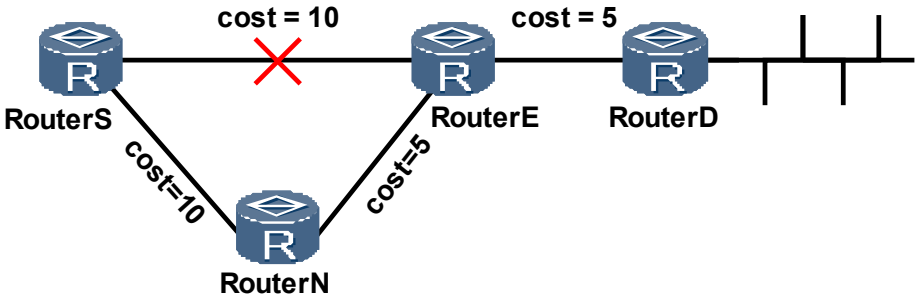
Application Environment

OSPF IP FRR is classified into link protection and link-node dual protection. Distance_opt(X,Y) indicates the shortest path between node X and node Y.

Link protection: indicates that the object to be protected is the traffic passing through an OSPF IP FRR-enabled link. The link cost must satisfy the inequality Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D). S indicates the source node of traffic; N indicates the node on the backup link; D indicates the destination node of traffic.

As shown in **Figure 4-10**, traffic is transmitted from Router S to Router D. The link cost satisfies the link protection inequality. When the primary link fails, Router S switches the traffic to the backup link Router S -> Router N so that the traffic can be further transmitted along downstream paths. This ensures that traffic interruption is less than 50 ms.

Figure 4-10 OSPF IP FRR link protection



Link-node dual protection: **Figure 4-11** shows link-node dual protection of OSPF IP FRR. Node protection takes precedence over link protection.

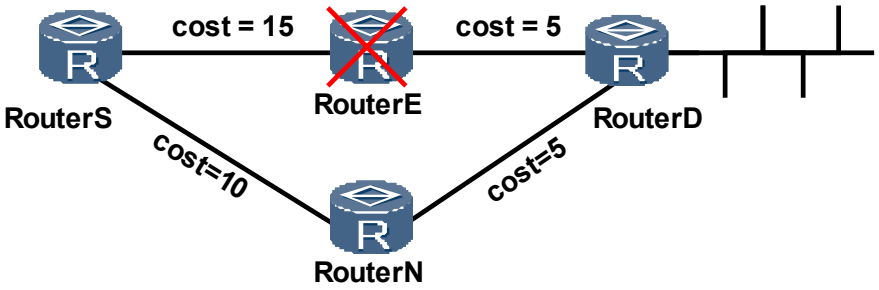
Link-node dual protection must satisfy the following situations:

The link cost must satisfy the inequality Distance_opt(N, D) < Distance_opt(N, S) + Distance_opt(S, D).

The interface cost of the router must satisfy the inequality Distance_opt(N, D) < Distance_opt (N, E) + Distance_opt(E, D).

S indicates the source node of traffic; E indicates the faulty node; N indicates the node on the backup link; D indicates the destination node of traffic.

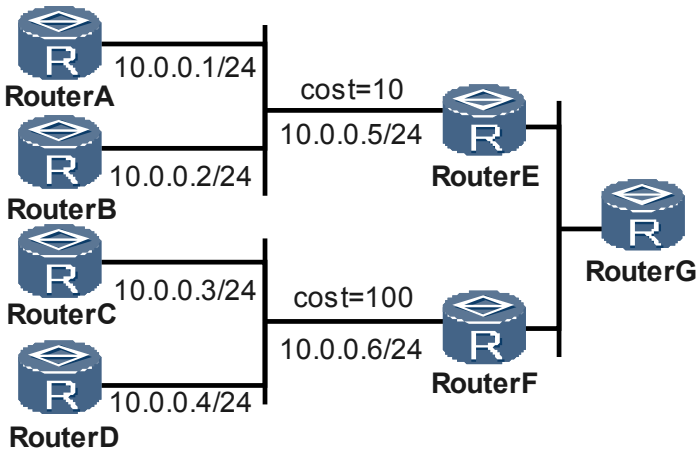
Figure 4-11 OSPF IP FRR link-node dual protection



4.2.12 Advertising Host Routes

Unlike in a data communication network, in an optical network, IP addresses of the same network can be configured as belonging to different physical networks. As a result of this kind of IP address planning, some host addresses may become unreachable. By configuring OSPF to advertise the corresponding host routes (routes whose destinations are hosts) of interfaces in addition to advertising network segment routes (routes whose destinations are network segments), you can ensure that these host addresses are reachable.

Figure 4-12 IP address planning on a typical optical network



As shown in **Figure 4-12**, if the function of advertising host routes is not enabled, all routes are advertised on the network in the form of 10.0.0.0/24. The next hop of the route from Router G to 10.0.0.0/24 is RouterE. As a result, 10.0.0.3, 10.0.0.4, and 10.0.0.6 become unreachable. To solve this problem, you can configure OSPF to advertise host routes so that routes are advertised as host addresses (such as 10.0.0.1/32).

4.2.13 OSPF-BGP Association

Definition

When a new device is deployed in the network or a device is restarted, network traffic may be lost during BGP convergence. This is because IGP convergence is faster than BGP convergence.

This problem can be solved through the synchronization between OSPF and BGP.

Purpose

If a backup link exists, during traffic switchback, BGP traffic is lost because BGP route convergence is slower than OSPF route convergence.

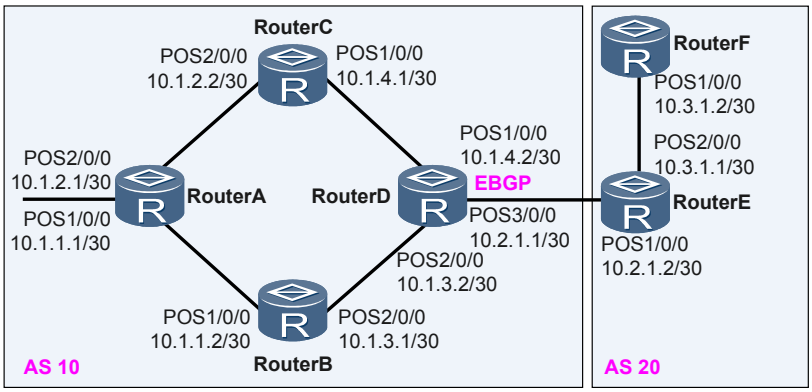
As shown in **Figure 4-13**, Router A, Router B, Router C, and Router D run OSPF and establish IBGP connections. Router C functions as the backup of Router B. When the network is stable, BGP and OSPF routes converge completely on the router.

Normally, traffic from Router A to 10.3.1.0/30 passes through Router B. When Router B becomes faulty, traffic is switched to Router C. After Router B recovers, traffic is switched back to Router B. During this process, packet loss occurs.

This is because when traffic is switched back to Router B, IGP route convergence is faster than BGP route convergence. Consequently, convergence of OSPF routes is already complete when BGP route convergence is still going on. As a result, Router B does not know the route to 10.3.1.0/30.

Therefore, when packets from Router A to 10.3.1.0/30 arrive at Router B, they are discarded because Router B does not have the route to 10.3.1.0/30.

Figure 4-13 OSPF-BGP synchronization



Principle

The device enabled with OSPF-BGP synchronization remains as a stub router within the set synchronization period. That is, the link metric in the LSA advertised by the device is the maximum value 65535. Therefore, the device instructs other OSPF routers not to use it for data forwarding.

As shown in **Figure 4-13**, OSPF-BGP synchronization is enabled on Router B. In this situation, before BGP route convergence is complete, Router A continues to use the backup link Router C rather than forward traffic to Router B until BGP route convergence on Router B is complete.

4.2.14 OSPF Local MT

Definition and Purpose

When multicast and an MPLS TE tunnel are deployed in a network, multicast may be affected by the TE tunnel, which causes multicast services to become unavailable.

To solve this problem, you can enable local multicast-topology (MT) for multicast packet forwarding.

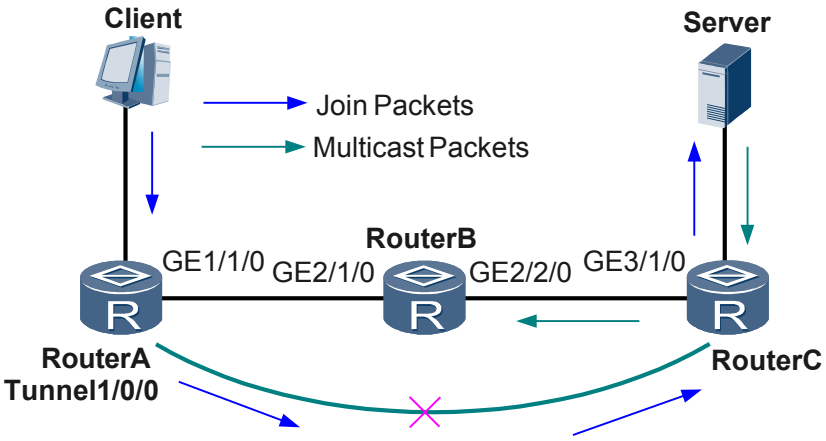
Local MT

After IGP Shortcut is configured on a TE tunnel, the outbound interface of the route calculated by an IGP may not be the actual physical interface but a TE tunnel interface.

According to the unicast route to the multicast source address, a router sends a Join message through a TE tunnel interface. In this situation, routers spanned by the TE tunnel cannot detect the Join message, so they do not create any multicast forwarding entry.

As shown in **Figure 4-14**, Router B spanned by the TE tunnel does not create any multicast forwarding entry.

Figure 4-14 OSPF Local MT



A TE tunnel is unidirectional, so multicast data packets sent by the multicast source are sent to the routers spanned by the tunnel through physical interfaces. These routers discard the multicast data packets, because they do not have any multicast forwarding entry. As a result, services become unavailable.

After local MT is enabled, if the outbound interface of the calculated route is a TE tunnel interface of IGP Shortcut type, the route management (RM) module creates a separate Multicast IGP (MIGP) routing table for the multicast protocol, calculates the actual physical outbound interface for the route, and then adds the route to the MIGP routing table. Multicast then uses routes in the MIGP routing table to forward packets.

In **Figure 4-14**, the packets requesting to join a multicast group is sent from Router A to Router B through GE 1/1/0. Router B then can create the multicast forwarding table correctly.

4.2.15 OSPF GR

Routers generally operate with separation of the control plane and forwarding plane. When the network topology remains stable, a restart of the control plane does not affect the forwarding plane, and the forwarding plane can still forward data properly. This separation ensures non-stop service forwarding.

In graceful restart (GR) mode, the forwarding plane continues to direct data forwarding after a restart occurs. The actions on the control plane, such as re-establishment of neighbor relationships and route calculation, do not affect the forwarding plane. Network reliability is improved because service interruption caused by route flapping is prevented.

Basic Concepts of OSPF GR

As mentioned in chapter 4, Graceful Restart (GR) is a technology used to ensure normal traffic forwarding and non-stop forwarding of key services during the restart of routing protocols.

Unless otherwise stated, GR described in this section refers to the GR technology defined in RFC 3623.

GR is one of the high availability (HA) technologies, which comprise a set of comprehensive technologies, such as fault-tolerant redundancy, link protection, faulty node recovery, and traffic

engineering. As a fault-tolerant redundancy technology, GR is widely used to ensure non-stop forwarding of key services during master/slave switchover and system upgrade.

The following concepts are involved in GR:

- Grace-LSA

  OSPF supports GR by flooding grace LSAs. Grace LSAs are used to inform the neighbor of the GR time, cause, and interface address when the GR starts and ends.
- Role of a router during GR
  - Restarter: is the router that restarts. The Restarter can be configured to support totally GR or partly GR.
  - Helper: is the router that helps the Restarter. The Helper can be configured to support planned GR or unplanned GR or to selectively support GR through the configured policies.
- Conditions that cause GR
  - Unknown: indicates that GR is triggered for an unknown reason.
  - Software restart: indicates that GR is triggered by commands.
  - Software reload/upgrade: indicates that GR is triggered by software restart or upgrade.
  - Switch to redundant control processor: indicates that GR is triggered by the abnormal master/slave switchover.
- GR period

  The GR period cannot exceed 1800 seconds. OSPF routers can exit from GR regardless of whether GR succeeds or fails, without waiting for GR to expire.

Classification of OSPF GR

- Totally GR: indicates that when a neighbor of a router does not support GR, the router exits from GR.
- Partly GR: indicates that when a neighbor does not support GR, only the interface associated with this neighbor exits from GR, whereas the other interfaces perform GR normally.
- Planned GR: indicates that a router restarts or performs the master/slave switchover using a command. The Restarter sends a grace LSA before restart or master/slave switchover.
- Unplanned GR: indicates that a router restarts or performs the master/slave switchover because of faults. A router performs the master/slave switchover, without sending a grace LSA, and then enters GR after the slave board goes Up. The process of unplanned GR is the same as that of planned GR.

GR Process

- A router starts GR.

  In planned GR mode, after master/slave switchover is triggered through a command, the Restarter sends a grace LSA to all neighbors to notify them of the start, period, and cause of GR, and then performs the master/slave switchover.

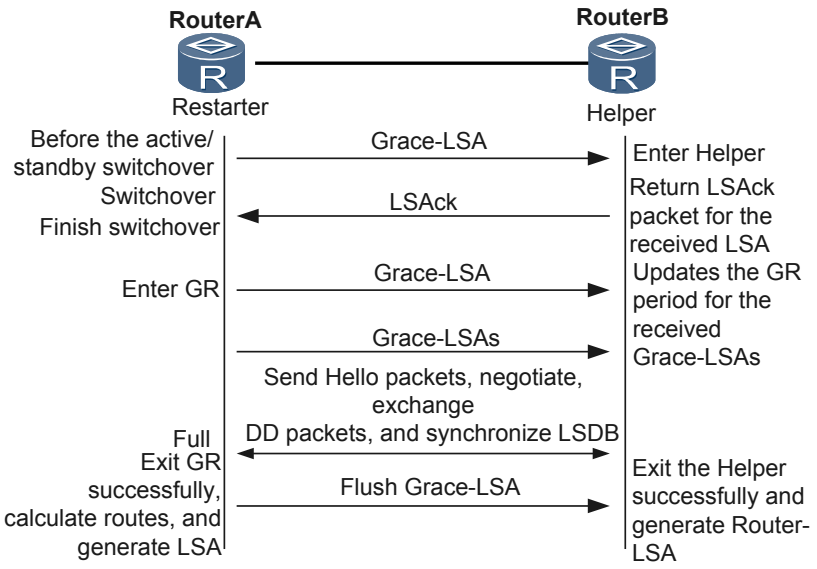
  In unplanned GR, the Restarter does not send the grace LSA.

  In unplanned GR mode, the Restarter sends a grace LSA immediately after the slave board goes Up, informing neighbors of the start, period, and cause of GR. The Restarter then sends a grace LSA to each neighbor five times consecutively. This ensures that neighbors receive the grace LSA. This operation is proposed by manufacturers but not defined by the OSPF protocol.

The Restarter sends a grace LSA to notify neighbors that it enters GR. During GR, neighbors keep neighbor relationships with the Restarter so that other routers cannot detect the switchover of the Restarter.

- The GR process runs, as shown in Figure 1

Figure 4-15 OSPF GR process



- The router exits from GR.

Table 4-13 Reasons that a router exits GR

Execution of GR	Restarter	Helper
GR succeeds.	Before GR expires, the Restarter re-establishes neighbor relationships with all neighbors before master/slave switchover.	After the Helper receives the grace LSA with the Age being 3600s from the Restarter, the neighbor relationship between the Helper and Restarter enters the Full state.

Execution of GR	Restarter	Helper
GR fails.	● GR expires, and neighbor relationships do not recover completely. ● Router LSA or network LSA sent by the Helper causes Restarter to fail to perform bidirectional check. ● Status of the interface that functions as the Restarter changes. ● Restarter receives the one-way Hello packet from the Helper. ● The Restarter receives the grace LSA that is generated by another router on the same network segment. Only one router can perform GR on the same network segment. ● On the same network segment, neighbors of the Restarter have different DRs or BDRs because of the topology changes.	● Helper does not receive the grace LSA from Restarter before the neighbor relationship expires. ● Status of the interface that functions as the Helper changes. ● Helper receives the LSA that is inconsistent with the LSA in the local LSDB from another router. This situation can be excluded after the Helper is configured not to perform strict LSA check. ● Helper receives grace LSAs from two routers on the same network segment at the same time. ● Neighbor relationships between Helper and other neighbors change.

Comparison Between GR Mode and Non-GR Mode

Table 4-14 Comparison of master/slave switchover in the GR mode and non-GR mode

Switchover in Non-GR Mode	Switchover in GR Mode
● OSPF neighbor relationships are re-established. ● Routes are recalculated. ● Forwarding table changes. ● Entire network detects route changes, and route flapping occurs for a short period of time. ● Packets are lost during forwarding, and services are interrupted.	● OSPF neighbor relationships are re-established. ● Routes are recalculated. ● Forwarding table remains unchanged. ● Except for neighbors of the device where master/slave switchover occurs, other routers do not detect route changes. ● No packets are lost during forwarding, and services are not affected.

4.2.16 OSPF-LDP Association

Definition

In the networking that uses primary and backup links, when the faulty primary link recovers, traffic is switched from the backup link back to the primary link.

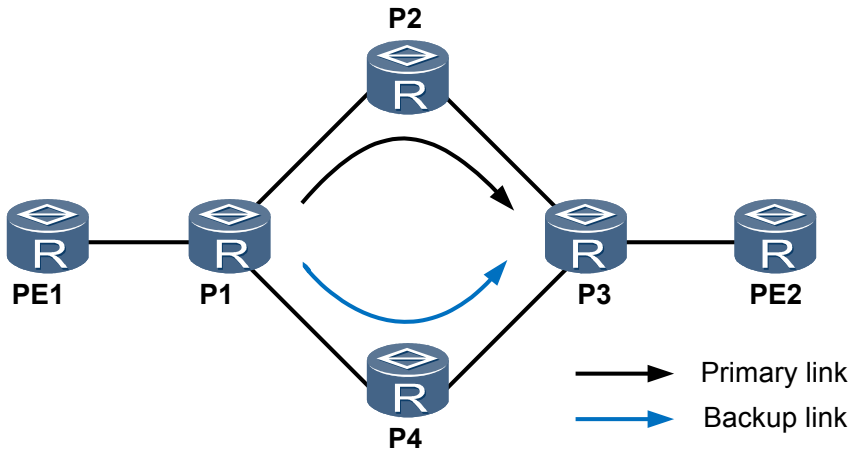
IGP route convergence completes before an LDP session is established. Consequently, the old LSP is deleted before the new LSP is established and LSP traffic is interrupted.

Purpose

As shown in **Figure 4-16**, the primary link adopts the path PE1→P1→P2→P3→PE2, and the backup link adopts the path PE1→P1→P4→P3→PE2.

When the primary link is faulty, traffic is switched to the backup link. After the primary link recovers, traffic is switched back to the primary link. During this process, traffic is interrupted for a long period of time.

Figure 4-16 OSPF-LDP association



Synchronizing LDP and IGP on P1 and P2 can shorten traffic interruption caused by traffic switchover from the backup link to the primary link.

Table 4-15 OSPF-LDP association

Enabling Status of OSPF-LDP Association	Traffic Interruption Time
Not enabled.	Seconds level
Enabled.	Milliseconds level

Principle

The principle of LDP-IGP synchronization is to delay route switchback by suppressing the establishment of IGP neighbor relationships until LDP convergence is complete. That is, before

an LSP on the primary link is established, the backup link continues to forward traffic. Then the link is deleted after the LSP is established.

Synchronization of LDP and IGP involves three timers:

- Hold-down
- Hold-max-cost
- Delay

After the primary link recovers, a router responds as follows:

1. Starts the hold-down timer. The IGP interface does not establish IGP neighbors but waits for establishment of an LDP session. The Hold-down timer specifies the period that the IGP interface waits.
2. Starts the hold-max-cost timer after the hold-down timer expires. The hold-max-cost timer specifies the interval for advertising the maximum link metric of the interface in the Link State Advertisement (LSA) to the primary link.
3. Starts the Delay timer to allow time for establishment of an LSP after an LDP session is re-established for the faulty link.

   After the Delay timer expires, LDP notifies IGP that synchronization is complete regardless of the status of IGP.

4.2.17 OSPF Database Overflow

Definition

OSPF requires that routers in the same area have the same Link State Database (LSDB).

With the continuous increase in routes on the network, some routers fail to carry the additional routing information because of limited system resources. This situation is called *OSPF database overflow*.

Purpose

You can configure stub areas or NSSAs to solve the problem of the continuous increase in routing information that causes the exhaustion of system resources of routers. However, configuring stub areas or NSSAs cannot solve the problem when the unexpected increase in dynamic routes causes database overflow. Setting the maximum number of external LSAs in the LSDB can dynamically limit the LSDB capacity, to avoid the problems caused by database overflow.

Principle

To prevent database overflow, you can set the maximum number of non-default external routes on a router.

All routers on the OSPF network must be set with the same upper limit. If the number of external routes on a router reaches the upper limit, the router enters the Overflow state and starts an overflow timer. The router automatically exits from the overflow state after the timer expires, By default, it is 5 seconds.

Table 4-16 OSPF database overflow

Overflow Phase	OSPF Processing
Entering overflow state	A router deletes all non-default external routes that is generated.
Staying in overflow state	● Router does not generate non-default external routes. ● Router discards the newly received, non-default external routes, and does not reply with an LSAck packet. ● When the overflow timer expires, the router checks whether the number of external routes still exceeds the upper limit. – If so, the router restarts the timer. – If not, the router exits from overflow state.
Exiting from the overflow state	● Router deletes the overflow timer. ● Router generates non-default routes. ● Router learns the newly received non-default routes, and replies with an LSAck packet. ● Router prepares to enter Overflow state for the next time it occurs.

4.2.18 OSPF Mesh-Group

Definition

In the scenario where there are multiple concurrent links, you can deploy OSPF mesh-group to classify links into a mesh group. Then, OSPF floods LSAs to only a link selected from the mesh group. Using OSPF mesh-group prevents unnecessary burden on the system caused by repetitive flooding.

The mesh-group feature is disabled by default.

Purpose

After receiving or generating an LSA, an OSPF process floods the LSA. When there are multiple concurrent links, OSPF floods the LSA to each link and sends Update messages.

In this scenario, if there are 2000 concurrent links, OSPF floods each LSA 2000 times. Only one flooding, however, is valid. The other 1999 times are useless repetition.

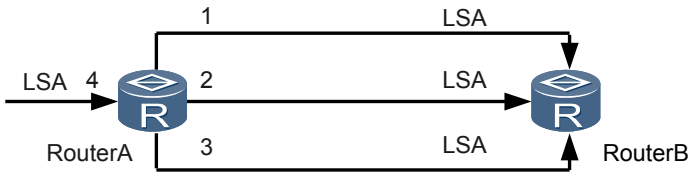
To prevent burden on the system caused by repetitive flooding, you can enable mesh-group to classify multiple concurrent links between a router and its neighbor into a group and then select a primary link to use for flooding.

Principles

As shown in **Figure 4-17**, Router A and Router B, which are connected through three links, establish an OSPF neighbor relationship. After receiving a new LSA from interface 4, Router A floods the LSA to Router B through interfaces 1, 2, and 3.

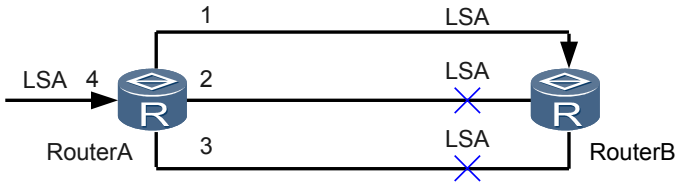
This flooding causes a heavy load on the concurrent links. For the neighbor with concurrent links, only a primary link is selected to flood the LSA.

Figure 4-17 LSAflooding with OSPF mesh-group disabled



When multiple concurrent links exist between a device enabled with OSPF mesh-group and its neighbor, the device selects? to flood the received LSAs, as shown in **Figure 4-18**.

Figure 4-18 LSA flooding with OSPF mesh-group enabled

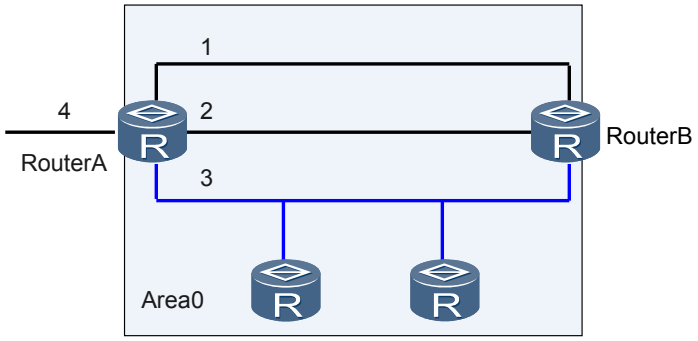


As defined in OSPF, LSAs can be flooded to a link only when the neighbor status is not lower than Exchange. In this case, when the status of the interface on the primary link is lower than Exchange, OSPF reselects a primary link from the concurrent links and then floods the LSA. After receiving the LSA flooded by Router A from link 1, Router B no longer floods the LSA to Router A through interfaces 2 and 3.

As defined by the mesh-group feature, the Router ID of a neighbor uniquely identifies the mesh group. Interfaces connected to the same neighbor that have a status greater than Exchange belong to the same mesh group.

In **Figure 4-19**, a mesh group of Router A resides in Area 0, which contains the links of interface 1 and interface 2. More than one neighbor of interface 3 resides on the broadcast link. Therefore, interface 3 cannot be defined as part of the mesh group.

Figure 4-19 Interface not added to mesh group



NOTE

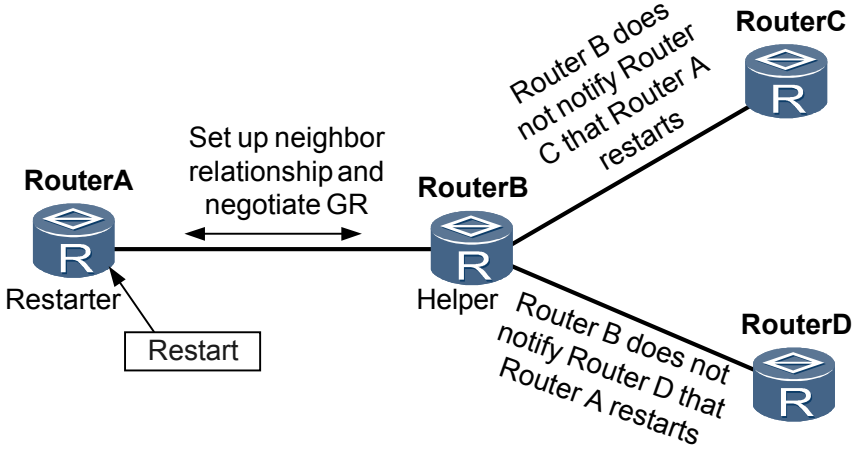
After a router is enabled with mesh-group, if the Router IDs of the router and its directly connected neighbor are the same, LSDBs cannot be synchronized and routes cannot be calculated correctly. In this case, you need to reconfigure the Router ID of the neighbor.

4.3 OSPF Applications

4.3.1 OSPF GR

In **Figure 4-20**, Router A, Router B, Router C, and Router D run OSPF for interworking, and Router A and Router B are enabled with GR. When Router A restarts, Router B helps Router A perform GR, without notifying other neighbors of Router A. OSPF GR ensures non-interrupted network traffic.

Figure 4-20 OSPF GR

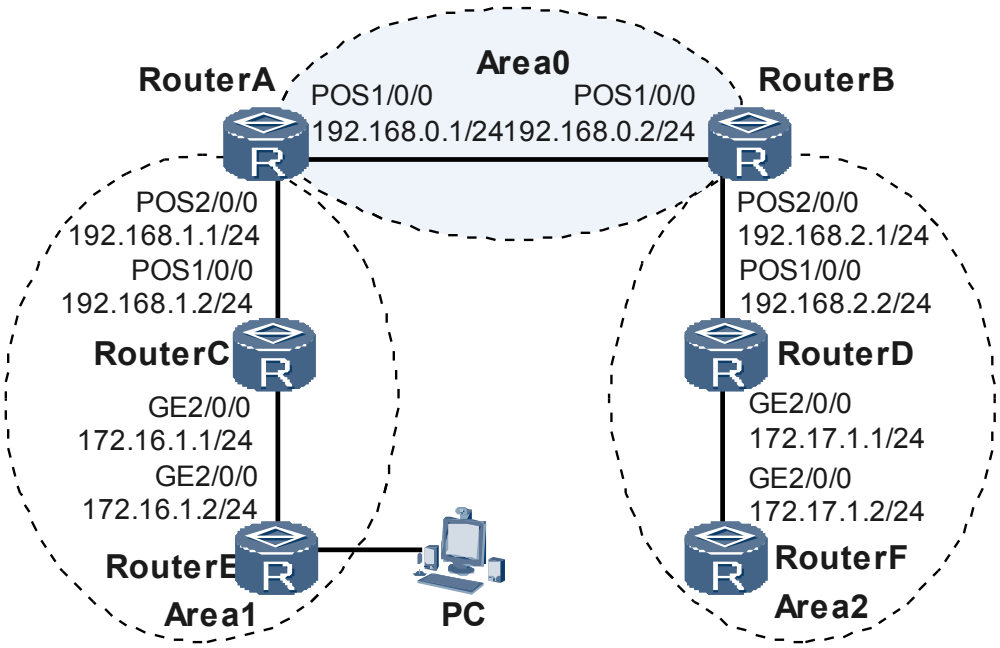


4.3.2 OSPF GTSM

As shown in **Figure 4-21**, OSPF runs between routers, and GTSM is enabled on Router C. The following are the valid TTL ranges of the packets that are sent from routers to Router C:

- Router A and Router E are the neighbors of Router C, and their valid TTL range of packets is [255 - hops + 1, 255].
- The valid TTL ranges of the packets sent from Router B, Router D, and Router F to Router C are respectively [254, 255], [253, 255], and [252, 255].

Figure 4-21 OSPF GTSM



NOTE

For detailed description of OSPF GTSM, refer to the *Feature Description - Security*.

4.4 References

The following table lists the references that apply in this chapter.

Document	Description	Remarks
RFC 1587	This document describes a new optional type of OSPF area, referred to humorously as a "not-so-stubby" area (or NSSA). NSSAs are similar to the existing OSPF stub area configuration option, but have the additional capability to import AS external routes on a limited basis.	-

Document	Description	Remarks
RFC 1765	Proper operation of the OSPF protocol requires that all OSPF routers maintain an identical copy of the OSPF link-state database. However, when the size of the link-state database becomes very large, some routers might be unable to store the entire database due to resource shortages. This condition is called *"database overflow"*.	This RFC is experimental and non-standard.
RFC 2328	This memo documents version 2 of the OSPF protocol.	-
RFC 2370	This memo defines enhancements to the OSPF protocol to support a new class of link-state advertisements (LSA) called Opaque LSAs. Opaque LSAs provide a generalized mechanism to allow for future extensibility of OSPF.	-
RFC 3137	This memo describes a backward-compatible technique that can be used by OSPF (Open Shortest Path First) implementations to advertise unavailability to forward transit traffic or to lower the preference level for the paths through such a router.	This RFC is informational and non-standard.
RFC 3623	This memo documents an enhancement to the OSPF routing protocol, whereby an OSPF device can stay on the forwarding path even as its OSPF software is restarted.	-
RFC 3630	This document describes extensions to the OSPF protocol version 2 to support intra-area Traffic Engineering (TE), using Opaque Link State Advertisements.	-
RFC 3682	This document describes the use of a packets Time to Live (TTL) (IPv4) or Hop Limit (IPv6) to protect a protocol stack from CPU-utilization based attacks, which has been proposed in many settings.	This RFC is experimental and non-standard.
RFC 3906	This document describes how conventional hop-by-hop link-state routing protocols interact with new Traffic Engineering capabilities to create Interior Gateway Protocol (IGP) shortcuts.	-
RFC 4576	This document specifies the procedure, using one of the options bits in the LSA (Link State Advertisements) to indicate that an LSA has already been forwarded by a PE and should be ignored by any other PEs.	-

Document	Description	Remarks
RFC 4577	This document extends the RFC 4576 specification by allowing the routing protocol on the PE/CE interface to be Open Shortest Path First (OSPF) protocol.	-
RFC 4750	This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in TCP/IP-based Internet networks. In particular, it defines objects to manage version 2 of the Open Shortest Path First Routing Protocol. Version 2 of the OSPF protocol is specific to the IPv4 address family.	-

5 OSPFv3

About This Chapter

5.1 Introduction to OSPFv3

Definition

The Open Shortest Path First (OSPF) protocol, developed by the Internet Engineering Task Force (IETF), is an interior gateway protocol based on the link status.

At present, OSPF Version 2 is used for IPv4 and OSPF Version 3 is used for IPv6.

- OSPFv3 is short for OSPF Version 3.
- As defined in RFC 2740, OSPFv3 is a routing protocol over IPv6.
- OSPFv3 is an independent routing protocol whose functions are modified on the basis of OSPFv2.

Purpose

The primary purpose of OSPFv3 is to develop a routing protocol independent of any specific network layer. The internal routing information of OSPFv3 is redesigned to serve this purpose.

The differences between OSPFv3 and OSPFv2 are as follows:

- OSPFv3 does not insert IP-based data in the header of the packet and Link State Advertisement (LSA).
- OSPFv3 executes some crucial tasks that originally require the data in the IP packet header by making use of the information independent of any network protocol. For example, OSPFv3 can identify the LSA that advertises the routing data.

5.2 Principle

5.2.1 Principle of OSPFv3

Running on IPv6, OSPFv3 (defined in RFC 2740) is an independent routing protocol whose functions are enhanced on the basis of OSPFv2.

- OSPFv3 and OSPFv2 are the same in respect of the working principles of the Hello message, state machine, link-state database (LSDB), flooding, and route calculation.
- OSPFv3 divides an Autonomous System (AS) into one or more logical areas and advertises routes through LSAs.
- OSPFv3 achieves unity of routing information by exchanging OSPFv3 packets between routers within an OSPFv3 area.
- OSPFv3 packets are encapsulated into IPv6 packets, which can be transmitted in unicast or multicast mode.

Formats of OSPFv3 Packets

Packet Type	Description
Hello message	Hello messages are sent regularly to discover and maintain OSPFv3 neighbor relationships.
Database Description (DD) packet	A DD packet contains the summary of the local LSDB. It is exchanged between two OSPFv3 routers to update the LSDBs.
Link State Request (LSR) packet	LSR packets are sent to the neighbor to request the required LSAs. An OSPFv3 router sends LSR packets to its neighbor only after they exchange DD packets.
Link State Update (LSU) packet	The LSU packet is used to transmit required LSAs to the neighbor.
Link State Acknowledgment (LSAck) packet	The LSAck packet is used to acknowledge the received LSA packets.

LSA Type

LSA Type	Description
Router-LSA (Type1)	Generated by a router for each area to which an OSPFv3 interface belongs, the router LSA describes the status and costs of links of the router and is advertised in the area where the OSPFv3 interface belongs.
Network-LSA (Type2)	Generated by a designated router (DR), the network LSA describes the link status and is broadcast in the area that the DR belongs to.
Inter-Area-Prefix-LSA (Type3)	Generated on the area border router (ABR), an inter-area prefix LSA describes the route of a certain network segment within the local area and is used to inform other areas of the route.
Inter-Area-Router-LSA (Type4)	Generated on the ABR, an inter-area router LSA describes the route to the autonomous system boundary router (ASBR) and is advertised to all related areas except the area that the ASBR belongs to.
AS-external-LSA (Type5)	Generated on the ASBR, the AS-external LSA describes the route to a destination outside the AS and is advertised to all areas except the stub area and NSSA area.
NSSA-LSA (Type7)	Describes routes to a destination outside the AS. It is generated by an ASBR and advertised in NSSAs only.

LSA Type	Description
Link-LSA (Type8)	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.
Intra-Area-Prefix-LSA (Type9)	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. ● An LSA generated on a DR describes the IPv6 address prefix associated with the network LSA.

Router Type

Figure 5-1 Router type

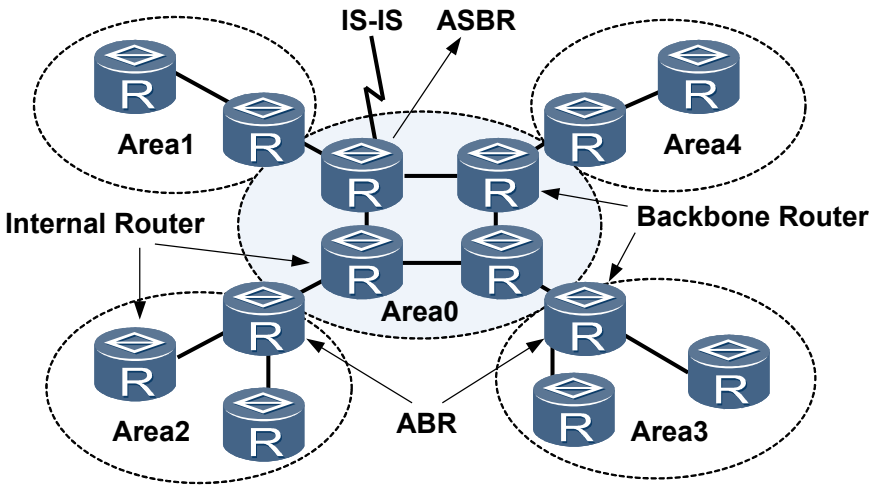


Table 5-1 Router types and descriptions

Router Type	Description
Internal router	All interfaces on an internal router belong to the same OSPFv3 area.
Area border router (ABR)	An ABR can belong to two or more areas, but one of the areas must be a backbone area. An ABR is used to connect the backbone area and the non-backbone areas. It can be physically or logically connected to the backbone area.

Router Type	Description
Backbone router	At least one interface on a backbone router belongs to the backbone area. All ABRs and internal routers in Area 0, therefore, are backbone routers.
AS boundary router (ASBR)	A router that exchanges routing information with other ASs is called an ASBR. An ASBR may not locate on the boundary of an AS. It can be an internal router or an ABR.

OSPFv3 Route Type

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 5-2 lists route types in a descending order of priority.

Table 5-2 Types of OSPFv3 routes

Route Type	Description
Intra Area	Intra-area routes
Inter Area	Inter-area routes
Type1 external routes	Because of the high reliability of Type 1 external routes, the calculated cost of external routes is equal to that of AS internal routes, and can be compared with the cost of OSPFv3 routes. That is, the cost of a Type1 external route equals the cost of the route from the router to the corresponding ASBR plus the cost of the route from the ASBR to the destination address.
Type2 external routes	Because of the low reliability of Type2 external routes, the cost of the route from the ASBR to a destination outside the AS is considered far greater than the cost of any internal path to an ASBR. Therefore, OSPFv3 only takes the cost of the route from the ASBR to a destination outside the AS into account when calculating route costs. That is, the cost of a Type2 external route equals the cost of the route from the ASBR to the destination of the route.

Area Type

Table 5-3 Types of OSPFv3 areas

Area Type	Description
Totally stub area	A totally stub area allows the Type3 default routes advertised by the ABR, and disallows the routes outside the AS and inter-area routes.
Stub area	A stub area allows inter-area routes, which is different from a totally stub area.
NSSA	Imports routes outside an AS, which is different from a stub area. An ASBR advertises Type7 LSAs in the local area. These Type 7 LSAs are translated into Type 5 LSAs on an ABR, and are then flooded in the entire OSPFv3 AS.

Network Types Supported by OSPFv3

OSPFv3 classifies networks into the following types according to link layer protocols.

Table 5-4 Types of OSPFv3 networks

Network Type	Description
Broadcast	If the link layer protocol is Ethernet or FDDI, OSPFv3 defaults the network type to broadcast. In this type of networks, the following situations occur: ● Hello messages, LSU packets, and LSAck packets are transmitted in multicast mode (FF02::5 is the reserved IPv6 multicast address of the OSPFv3 router; FF02::6 is the reserved IPv6 multicast address of the OSPFv3 DR or BDR). ● DD packets and LSR packets are transmitted in unicast mode.
Non-broadcast Multiple Access (NBMA)	If the link layer protocol is frame relay, ATM, or X.25, OSPFv3 defaults the network type to NBMA. In this type of networks, protocol packets such as Hello messages, DD packets, LSR packets, LSU packets, and LSAck packets, are transmitted in unicast mode.
Point-to-Multipoint (P2MP)	Regardless of the link layer protocol, OSPFv3 does not default the network type to P2MP. A P2MP network must be forcibly changed from other network types. The common practice is to change a non-fully connected NBMA to a P2MP network. In this type of networks, the following situations occur: ● Hello messages are transmitted in multicast mode with the multicast address as FF02::5. ● Other protocol packets, including DD packets, LSR packets, LSU packets, and LSAck packets, are transmitted in unicast mode.

Network Type	Description
Point-to-point (P2P)	If the link layer protocol is PPP, HDLC, or LAPB, OSPFv3 defaults the network type to P2P. In this type of network, the protocol packets, including Hello messages, DD packets, LSR packets, LSU packets, and LSAck packets, are transmitted to the multicast address FF02::5.

Stub Area

A stub area is a special area where the ABRs do not flood the received external routes. In stub areas, the size of the routing table of the routers and the routing information in transmission are reduced.

Configuring a stub area is optional. Not all areas can be configured as stub areas. Usually, a stub area is a non-backbone area with only one ABR and is located at the AS boundary.

To ensure the reachability of a destination outside the AS, the ABR in the stub area generates a default route and advertises it to the non-ABR routers in the stub area.

Note the following when configuring a stub area:

- The backbone area cannot be configured as a stub area.
- If an area needs to be configured as a stub area, all the routers in this area must be configured with the **stub** command.
- An ASBR cannot exist in a stub area. That is, external routes are not flooded in the stub area.
- A virtual link cannot pass through the stub area.

OSPFv3 Route Summarization

Routing information can be decreased after route aggregation so that the size of routing tables is reduced, which improves the performance of routers.

The procedure for OSPFv3 route aggregation is as follows:

- Route summarization on an ABR

  An ABR can summarize routes with the same prefix into one route and advertise the summarized route in other areas.

  When sending routing information to other areas, an ABR generates Type 3 LSAs based on IPv6 prefixes. If consecutive IPv6 prefixes exist in an area and route summarization is enabled on the ABR of the area, the IPv6 prefixes can be summarized into one prefix. If there are multiple LSAs that have the same prefix, the ABR summarizes these LSAs and advertises only one summarized LSA. The ABR does not advertise any specific LSAs.
- Route summarization on an ASBR

  An ASBR can summarize imported routes with the same prefix into one route and then advertise the summarized route to other areas.

  After being enabled with route summarization, an ASBR summarizes imported Type 5 LSAs within the summarized address range. After route summarization, the ASBR does not generate a separate Type 5 LSA for each specific prefix within the configured range. Instead, the ASBR generates a Type 5 LSA for only the summarized prefix. In an NSSA,

an ASBR summarizes multiple imported Type 7 LSAs within the summarized address range into one Type 7 LSA.

OSPFv3 Virtual Link

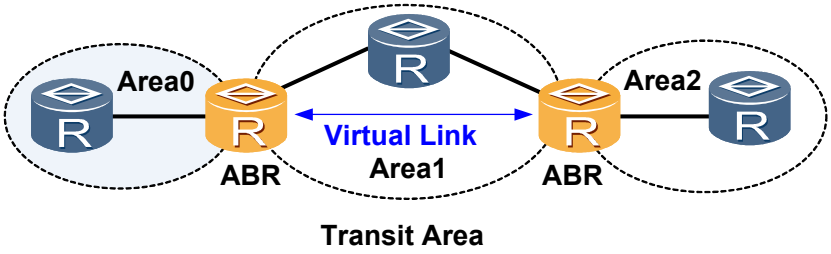
A virtual link refers to a logical channel established between two ABRs through a non-backbone area.

- A virtual link must be set up on both ends of the link; otherwise, it does not take effect.
- The transmit area refers to the area that provides an internal route of a non-backbone area for both the ends of the virtual link.

In actual applications, the physical connectivity between non-backbone areas and the backbone area cannot be ensured owing to various limitations. To solve this problem, you can configure OSPFv3 virtual links.

The virtual link is similar to a point-to-point connection between two ABRs. Similar to physical interfaces, the interfaces on the virtual link can be configured with parameters such as the hello interval.

Figure 5-2 OSPFv3 virtual link



As shown in **Figure 5-2**, OSPFv3 packets transmitted between two ABRs are only forwarded by the OSPFv3 devices that reside between the two ABRs. The OSPFv3 devices detect that they are not the destinations of the packets, so they forward the packets as common IP packets.

OSPFv3 Multi-process

OSPFv3 supports multi-process. More than one OSPFv3 process can run on the same router because processes are independent of each other. Route interaction between different OSPFv3 processes is similar to the route interaction between different routing protocols.

An interface of a router belongs to only a certain OSPFv3 process.

5.2.2 OSPFv3 GR

Graceful restart (GR) is a technology used to ensure normal traffic forwarding when a routing protocol restarts and guarantee that key services are not affected in the process.

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 redundancy technology, GR is widely used to ensure uninterrupted forwarding of key data in active/standby switchover and system upgrade.

If GR is not enabled, the active/standby switchover occurring owing to various causes leads to transient interruption of data forwarding, and as a result, route flapping occurs on the whole network. Such route flapping and service interruption are unacceptable on a large-scale network, especially on a carrier network.

In GR mode, the forwarding plane continues to direct data forwarding once a restart occurs, and the actions on the control plane, such as reestablishment of neighbor relationships and route calculation, do not affect the forwarding plane. In this manner, service interruption caused by route flapping is prevented so that the network reliability is improved.

Basic Concepts

- Grace LSA
  - OSPFv3 supports GR by flooding grace LSAs on the link.
  - Grace LSAs are used to inform the neighbor of the GR time, cause, and interface instance ID when GR starts and ends.
- Router function
  - A router can function as a GR restarter.
  - A router can function as a GR helper.
- GR implementation
  - Planned-GR: This refers to the smooth restart of OSPFv3 through the **reset ospfv3 graceful-restart** command. In this mode, a grace LSA is sent to the neighbor before the restart.
  - Unplanned-GR: This refers to the active/standby switchover triggered by router faults like power down, dead loop, exception or reset in master.

    Unlike planned-GR, no grace LSA is sent before the active/standby switchover in unplanned GR mode. Instead, the switchover is directly performed. When the standby board becomes Up, a grace LSA is sent and the GR process starts. The following procedure is the same as that of planned GR.

GR Process

Figure 5-3 OSPFv3 planned-GR process (reset ospfv3 graceful-restart)

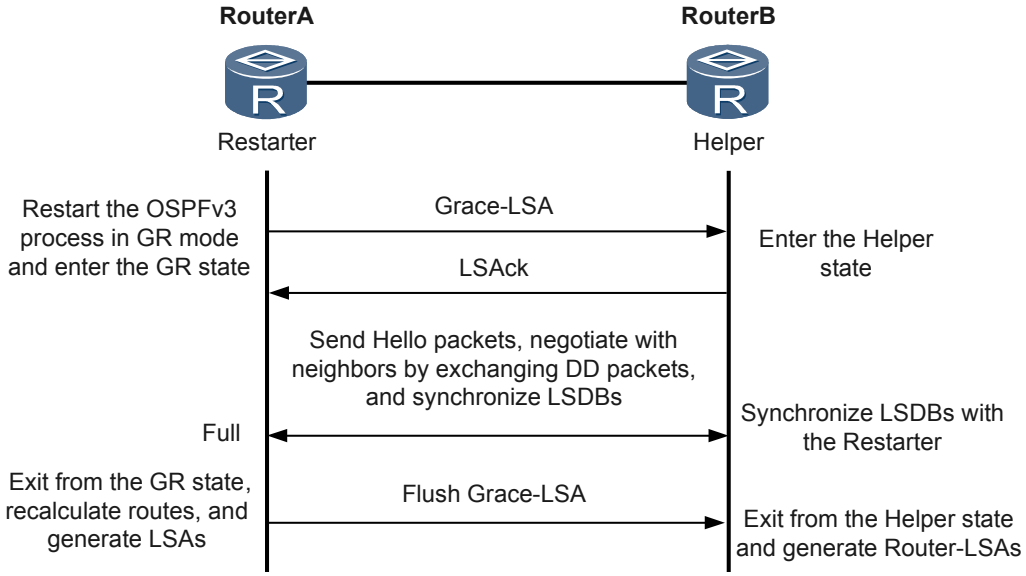
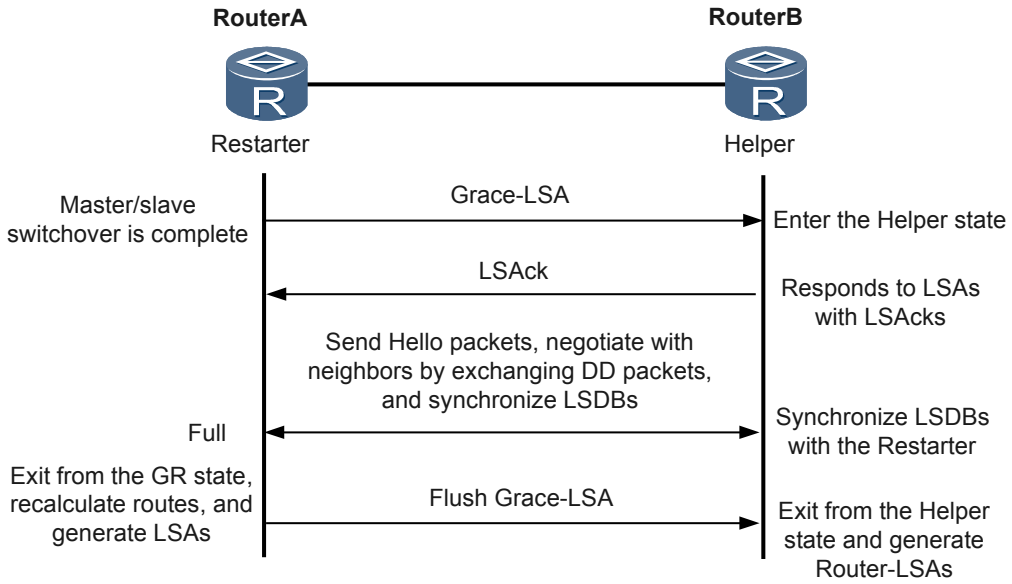


Figure 5-4 OSPFv3 unplanned-GR process (active/standby switchover)



- On the GR restarter:

1. In planned-GR mode, the GR restarter sends a grace LSA to all neighbors to inform them of the start of a GR process and the period and cause of this process.

   In unplanned GR mode, a grace LSA is sent to each neighbor immediately after the standby board is Up to inform the neighbors of the start of a GR process and the period and cause of the process.
2. The GR restarter performs negotiation with neighbors again to set up new neighbor relationships.
3. When all the neighbor relationships between the GR restarter and the original neighbors enter the Full state:
   - The GR restarter exits from the GR process and OSPFv3 recalculates routes.
   - The GR restarter updates the routing table on the main control board and the FIBs on interface boards and deletes invalid routing entries.
   - The GR restarter sends a grace LSA whose aging time is 3600 seconds to instruct the GR helper to exit from the GR process.

   Now, the GR process is complete.
4. If errors occur, the GR timer expires, or the neighbor relationship fails to enter the Full state during a GR process, the GR restarter exits from the process and OSPFv3 is restarted in non-GR mode. In this case, packets are lost.

- On the GR helper:

1. If a router is configured to support the GR process on its neighbor, the router enters the helper mode after receiving a grace LSA.
2. The GR helper maintains its neighbor relationship with the GR restarter, and the status of the neighbor relationship does not change.

3. If the GR helper continues to receive grace LSAs whose GR period is different from that on the GR helper, the GR helper updates its GR period.
4. Being informed of the successful GR process through a grace LSA whose aging time is 3600 seconds from the GR restarter, the GR helper exits from the GR process.
5. If errors occur during a GR process, the GR helper exits from the helper state and deletes invalid routes after route calculation.

Comparison between the GR Mode and the Non-GR Mode

Table 5-5 Comparison between the OSPFv3 GR mode and the OSPFv3 non-GR mode

Active/Standby Switchover in Non-GR Mode	Active/Standby Switchover in GR Mode
● OSPFv3 neighbor relationships are reestablished. ● Routes are recalculated. ● The forwarding table changes. ● Route changes are sensed on the network and route flapping occurs over 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 the same. ● Except the neighbor of the device where the active/standby switchover occurs, other routers do not sense the route changes. ● No packets are lost during forwarding, and services are not affected.

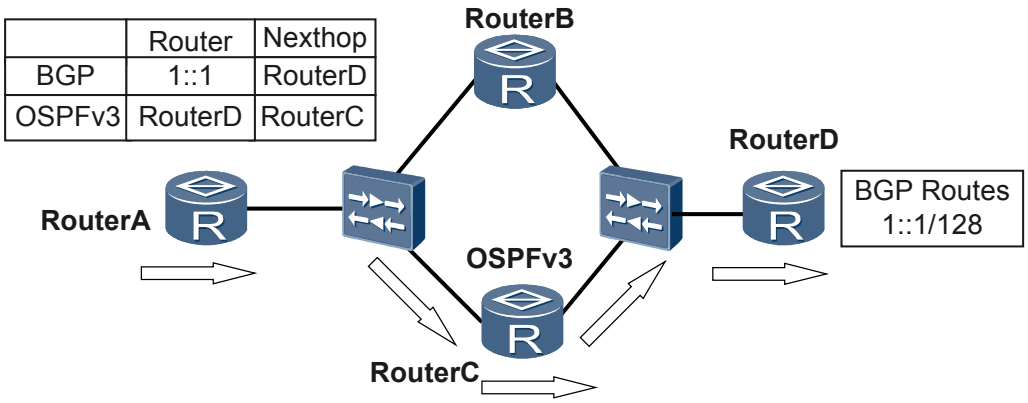
5.2.3 Association between OSPFv3 and BGP

When a new router is deployed in the network or a router is restarted, the network traffic may be lost during BGP convergence. This is because IGP convergence is quicker than BGP convergence. This problem can be solved through the association between OSPFv3 and BGP.

If a router on a BGP network recovers from a fault, BGP convergence is performed again and certain packets may be lost during the convergence.

As shown in **Figure 5-5**, traffic from Router A to Router D traverses a BGP network.

Figure 5-5 Traffic traversing a BGP network

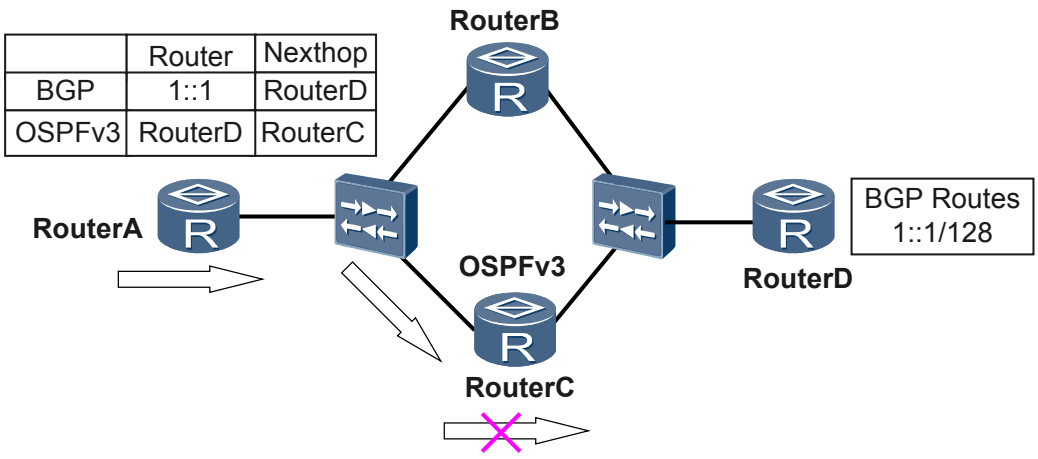


If a fault occurs on Router C, traffic is redirected to Router B after rerouting. Packets are lost when Router C is restored to the normal status.

Because OSPFv3 convergence is quicker than BGP convergence, OSPFv3 convergence is complete when Router C recovers. The next hop of the route from Router A to Router D is Router C, which, however, does not know the route to Router D since BGP convergence on Router C is not complete.

Thus, when the packets destined for Router D are transmitted from Router A to Router C, they are discarded by Router C because Router C has no route to Router D, as shown in **Figure 5-6**.

Figure 5-6 Packet loss during the restart of the device not enabled with association between OSPFv3 and BGP



Process of Association between OSPFv3 and BGP

When a router enabled with association between OSPFv3 and BGP restarts, the router advertises a message in the local OSPFv3 area to instruct other routers not to use it as a transit router.

At the same time, the router sets the largest weight value of 65535 in its LSAs to ensure that it is not used by other routers as the transit router. The BGP route, however, can still reach the router.

5.2.4 Comparison between OSPFv3 and OSPFv2

OSPFv3 and OSPFv2 are the same in the following aspects:

- Network type and interface type
- Interface state machine and neighbor state machine
- LSDB
- Flooding mechanism
- Five types of packets, including Hello, DD, LSR, LSU, and LSAck packets
- Route calculation

OSPFv3 and OSPFv2 are different in the following aspects:

- OSPFv3 is based on links rather than network segments.

  OSPFv3 runs on IPv6, which is based on links rather than network segments.

  Therefore, you need not to configure OSPFv3 on the interfaces in the same network segment. It is only required that the interfaces enabled with OSPFv3 are on the same link. In addition, the interfaces can set up OSPFv3 sessions without IPv6 global addresses.

- OSPFv3 does not depend on IP addresses.

  This is to separate topology calculation from IP addresses. That is, OSPFv3 can calculate the OSPFv3 topology without knowing the IPv6 global address, which only applies to virtual link interfaces for packet forwarding.

- OSPFv3 packets and LSA format change.
  - OSPFv3 packets do not contain IP addresses.
  - OSPFv3 router LSAs and network LSAs do not contain IP addresses, which are advertised by 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 in broadcast, NBMA, or P2MP networks.

- Information about the flooding scope is added in LSAs of OSPFv3.

  Information about the flooding scope is added in the LSA Type field of LSAs of OSPFv3. Thus, OSPFv3 routers can process LSAs of unidentified types, which makes the processing more flexible.
  - OSPFv3 can store or flood unidentified packets, whereas OSPFv2 just discards unidentified packets.
  - OSPFv3 floods packets in an OSPF area or on a link. It sets the U flag bit of packets (the flooding area is based on the link local) so that unidentified packets are stored or forwarded to the stub area.

  For example, Router A and Router B can identify LSAs of a certain type. They are connected through Router C, which, however, cannot identify this type of LSAs. When Router A floods an LSA of this type, Router C can still flood the received LSA to Router B although it does not identify this LSA. Router B then processes the LSA.

  If OSPFv2 is run, Router C discards the unidentified LSA so that the LSA cannot reach Router B.

- OSPFv3 supports multi-process on a link.

  Only one OSPF process can be configured on a physical interface.

  In OSPFv3, one physical interface can be configured with multiple processes that are identified by different instance IDs. That is, multiple OSPFv3 instances can run on one physical link. They establish neighbor relationships with the other end of the link and transmit packets to the other end without interfering with each other.

  Thus, the resources of a link can be shared among OSPFv3 instances that simulate multiple OSPFv3 routers, which improves the utilization of limited router resources.

- 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. 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 that of the next hop to transmit OSPFv3 packets.

The advantages are as follows:

- The OSPFv3 can calculate the topology without knowing the global IPv6 addresses so that topology calculation is not based on IP addresses.
- The packets flooded on a link are not transmitted to other links, which prevents unnecessary flooding and saves bandwidth.

- OSPFv3 packets do not contain authentication fields.

  OSPFv3 directly adopts IPv6 authentication and security measures. Thus, OSPFv3 does not need to perform authentication. It only focuses on the processing of packets.

- OSPFv3 supports two new LSAs.
  - Link LSA: A router 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 router advertises an intra-area prefix LSA in the local OSPF area to inform the other routers in the area or the network, which can be a broadcast network or a 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. Thus, even if global IPv6 addresses are not configured or they are configured in different network segments, OSPFv3 can still establish and maintain neighbor relationships so that topology calculation is not based on IP addresses.

5.3 References

The following table lists the references of this document.

Document	Description	Remarks
RFC 2740	This document describes the modifications to OSPF to support version 6 of the Internet Protocol (IPv6).	-
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.	-

6 IS-IS

About This Chapter

6.1 Introduction to IS-IS

Definition

Intermediate System-to-Intermediate System (IS-IS) is an Interior Gateway Protocol (IGP) that runs within an autonomous system (AS). IS-IS is also a link-state routing protocol, using the shortest path first (SPF) algorithm to calculate routes.

Purpose

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) extended and modified IS-IS in RFC 1195. This modification enables IS-IS to apply to TCP/IP and OSI environments. This type of IS-IS is called Integrated IS-IS or Dual IS-IS.

NOTE

IS-IS stated in this document refers to Integrated IS-IS, unless otherwise stated.

In addition to IPv4 networks, IS-IS also applies to IPv6 networks to provide accurate routing information for IPv6 packets. IS-IS has good scalability, supports IPv6 network layer protocols, and is capable of discovering, generating, and forwarding IPv6 routes.

6.2 Principles

6.2.1 IS-IS Basic Concepts

IS-IS Topology Structure

Overall IS-IS Topology

IS-IS uses a two-level hierarchy (backbone area and non-backbone area) to support large-scale routing networks. Generally, Level-1 routers are deployed in non-backbone areas, whereas Level-2 and Level-1-2 routers are deployed in backbone areas. Each non-backbone area connects to the backbone area through a Level-1-2 router.

Figure 6-1 shows a network that runs IS-IS. The network is similar to an OSPF network typology with multiple areas. The backbone area contains all the routers in Area 1 and Level-1-2 routers in other areas.

Figure 6-1 IS-IS topology I

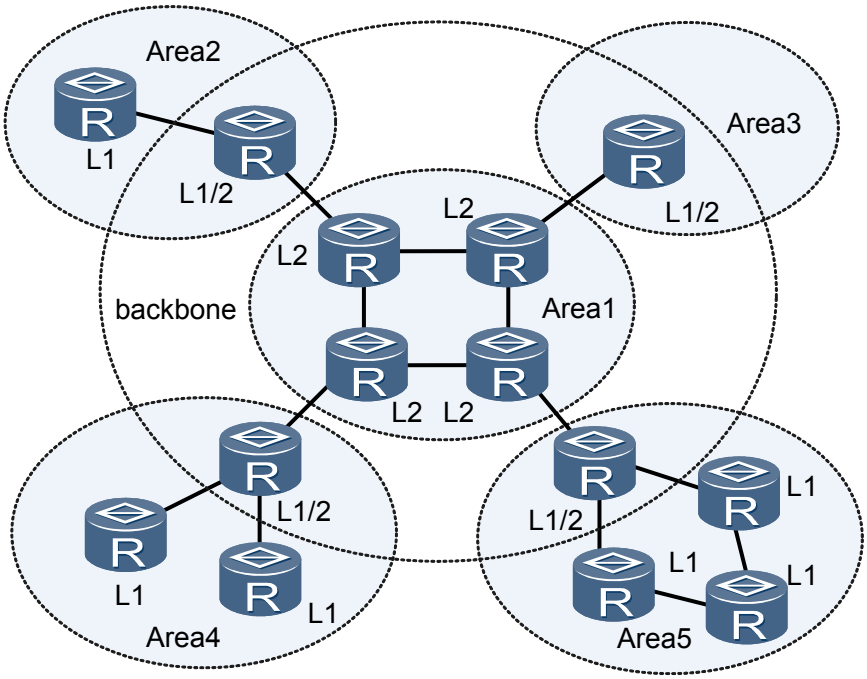
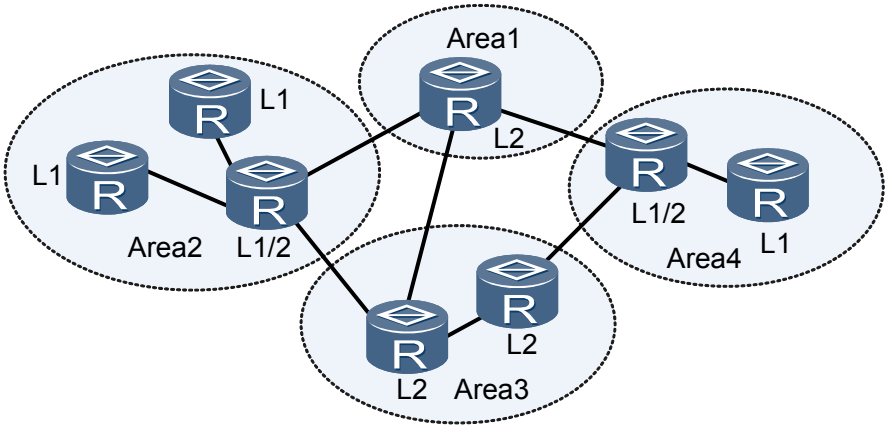


Figure 6-2 shows another type of IS-IS topology. In this topology, Level-2 routers belong to different areas. All the physically contiguous Level-1-2 and Level-2 routers form the backbone area of IS-IS.

Figure 6-2 IS-IS topology II



The two types of topologies show the differences between IS-IS and OSPF:

- In IS-IS, each router belongs to only one area. In OSPF, different interfaces of a router may belong to different areas.
- In IS-IS, no area is defined as the backbone area. In OSPF, Area 0 is defined as the backbone area.

- In IS-IS, Level-1 and Level-2 routes are calculated using the SPF algorithm to generate the shortest path tree (SPT). In OSPF, the SPF algorithm is used only in the same area, and inter-area routes are forwarded by the backbone area.

IS-IS Router Types

- Level-1 router

  A Level-1 router manages intra-area routing. It establishes neighbor relationships with only the Level-1 and Level-1-2 routers in the same area and maintains a Level-1 link state database (LSDB). The LSDB contains intra-area routing information. A packet to a destination outside this area is forwarded to the nearest Level-1-2 router.

- Level-2 router

  A Level-2 router manages inter-area routing. It can establish neighbor relationships with Level-2 or Level-1-2 routers in different areas and maintains a Level-2 LSDB. The LSDB contains inter-area routing information.

  All Level-2 routers form the backbone network of the routing domain. They establish Level-2 neighbor relationships and are responsible for inter-area communication. Level-2 routers in the routing domain must be physically contiguous to ensure the continuity of the backbone network. Only Level-2 routers can exchange data packets or routing information with routers outside the routing domain.

- Level-1-2 router

  A router that belongs to both a Level-1 area and a Level-2 area is called a Level-1-2 router. It can establish Level-1 neighbor relationships with Level-1 and Level-1-2 routers in the same area. It can also establish Level-2 neighbor relationships with Level-2 and Level-1-2 routers in different areas. A Level-1 router must be connected to other areas through a Level-1-2 router.

  A Level-1-2 router maintains two LSDBs: a Level-1 LSDB and a Level-2 LSDB. The Level-1 LSDB saves for intra-area routing and the Level-2 LSDB saves for inter-area routing.

IS-IS Network Types

IS-IS supports only two types of networks. In terms of physical links, IS-IS networks can be classified into the following link types:

- Broadcast: such as Ethernet and Token-Ring
- Point-to-point: such as PPP and HDLC

NOTE

For a Non-Broadcast Multi-Access (NBMA) network such as the ATM, you should configure its sub-interfaces as P2P interfaces.

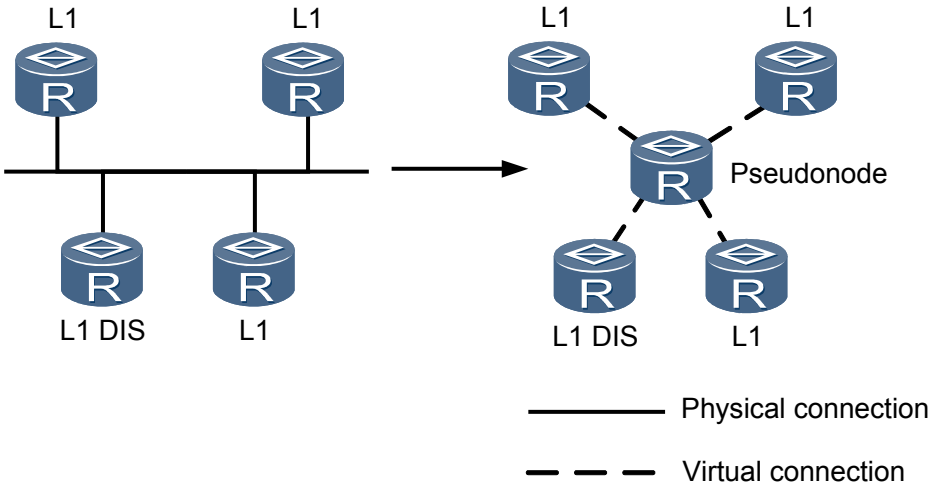
IS-IS cannot run on Point to MultiPoint (P2MP) networks.

DIS and Pseudonode

In a broadcast network, IS-IS needs to elect a Designated Intermediate System (DIS) from all the routers. DISs are used to create and update pseudonodes and generate link state protocol data units (LSPs) of pseudonodes to describe available network devices.

The pseudonode is used to simulate the virtual node in the broadcast network and is not an actual router. In IS-IS, a pseudonode is identified by the system ID of the DIS and the 1-byte Circuit ID (its value is not 0).

Figure 6-3 Pseudonode



The use of pseudonodes simplifies the network topology and shortens LSPs. When the network changes, the number of generated LSPs is reduced, and the SPF consumes fewer resources.

You can configure different priorities for DISs of different levels. The router with the highest priority is elected as the DIS. If there are multiple routers with the same highest priority on a broadcast network, the one with the highest MAC address is chosen. The DISs of different levels can be the same router or different routers.

DIS election in IS-IS differs from designated router (DR) election in OSPF:

- On an IS-IS broadcast network, the router with priority 0 also takes part in DIS election. In OSPF, the router with priority 0 does not take part in DR election.
- In IS-IS, when a new router that meets the requirements of being a DIS connects to a broadcast network, the router is elected as the new DIS, and the previous pseudonode is deleted. This causes a new flooding of LSPs. In OSPF, when a new router connects to a network, it is not immediately elected as the DR even if it has the highest DR priority.
- On an IS-IS broadcast network, routers (including non-DIS routers) of the same level on a network segment set up adjacencies. In OSPF, routers set up adjacencies with only the DR and backup designated router (BDR).

NOTE

On an IS-IS broadcast network, although all the routers set up adjacencies with each other, the LSDBs are synchronized by the DISs.

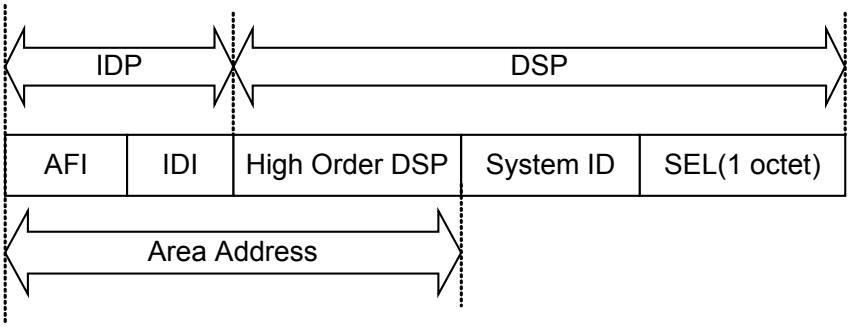
IS-IS Address Structure

The network service access point (NSAP) is an address defined by the OSI to locate resources. **Figure 6-4** shows the NSAP address structure. The NSAP is composed of the initial domain part (IDP) and the domain specific part (DSP). The lengths of the IDP and the DSP are variable. The maximum length of the NSAP is 20 bytes and its minimum length is 8 bytes.

- The IDP is similar to the network ID in an IP address. It is defined by the ISO and consists of the authority and format identifier (AFI) and the initial domain identifier (IDI). The AFI indicates the address allocation authority and address format, and the IDI identifies a domain.

- The DSP is similar to the subnet ID and host address in an IP address. 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, and the SEL indicates the service type.

Figure 6-4 IS-IS address structure



- Area Address

  The IDP and the HODSP of the DSP identify a routing domain and the areas in a routing domain. Therefore, the combination of the IDP and HODSP is called an area address, which is similar to an area number in OSPF. 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 fixed length of the system ID is 48 bits (6 bytes).

  In actual applications, a router ID corresponds to a system ID. If a router takes the IP address 168.10.1.1 of Loopback 0 as its router ID, its system ID used in IS-IS can be obtained in the following way:

  - Extend each part of IP address 168.10.1.1 to 3 bits and add 0 to the front of any part that is shorter than 3 bits. Then the IP address is extended as 168.010.001.001.
  - Divide the extended address 168.010.001.001 into three parts, each of which consists of four decimal digits. Then system ID 1680.1000.1001 is obtained.

  You can specify a system ID in many ways. You need to ensure that the system ID uniquely identifies a host or a router.

- SEL

  The role of an SEL is similar to that of the "protocol identifier" of IP. A transport protocol matches an SEL. The SEL is always "00" in IP.

A network entity title (NET) indicates network layer information about an IS. A NET can be regarded as a special NSAP. The NET length is the same as the NSAP length. Its maximum length is 20 bytes and minimum length is 8 bytes. When configuring IS-IS on a router, you only need to configure a NET but not an NSAP.

Assume that there is a NET: ab.cdef.1234.5678.9abc.00. In the NET, the area address is ab.cdef, the system ID is 1234.5678.9abc, and the SEL is 00.

IS-IS PDU Types

IS-IS PDUs include Hello PDUs, link state PDUs (LSPs), and sequence number PDUs (SNPs).

- Hello PDU

  Hello packets, also called IS-IS Hello PDUs (IIH), are used to set up and maintain neighbor relationships. Among them, Level-1 LAN IIHs apply to the Level-1 routers on broadcast LANs; Level-2 LAN IIHs apply to the Level-2 routers on broadcast LANs; and P2P IIHs apply to non-broadcast networks. Hello packets on different networks have different formats. Compared to a LAN IIH, a P2P IIH does not have the Priority and LAN ID fields, but has a Local Circuit ID field. The Priority field indicates the DIS priority on a broadcast network, the LAN ID field indicates the system ID of the DIS and pseudonode, and the Local Circuit ID indicates the local link ID.

- LSP

  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; and Level-1-2 IS-IS can transmit both Level-1 and Level-2 LSPs.

  The meanings of major fields in an LSP are as follows:

  - ATT field: When a Level-1-2 IS-IS transmits Level-1 LSPs in a Level-1 area, Level-1 IS-IS in the area can communicate with devices in other areas through the Level-1-2 IS-IS if the ATT bit is set in the Level-1 LSPs.
  - OL field: indicates the LSDB overload.

    LSPs with the overload bit are still flooded on the network, but these LSPs are ignored during the calculation of the routes that pass through a router in overload state. After the overload bit is set on a router, other routers ignore the router when performing SPF calculation and consider only the direct routes of the router. For details, see "IS-IS Overload" in Principles.
  - IS Type field: indicates the type of IS-IS that generates the LSP. The value 01 indicates Level-1, and the value 11 indicates Level-2.

- SNP

  SNPs describe the LSPs in all or some databases to help synchronize and maintain all LSDBs.

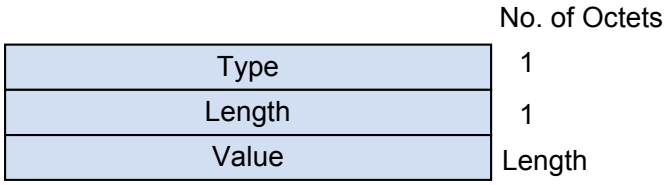
  SNPs include complete SNPs (CSNPs) and partial SNPs (PSNPs). They are further classified into Level-1 CSNPs, Level-2 CSNPs, Level-1 PSNPs, and Level-2 PSNPs.

  A CSNP contains the summary of all LSPs in an LSDB. This maintains LSDB synchronization between neighboring routers. On a broadcast network, the DIS periodically sends CSNPs. The default interval for sending CSNPs is 10 seconds. On a point-to-point link, CSNPs are sent only when the neighbor relationship is established for the first time.

  A PSNP lists only the sequence number of recently received LSPs. A PSNP can acknowledge multiple LSPs at one time. If an LSDB is not updated, the PSNP is also used to request a neighbor to send a new LSP.

The variable length fields in an IS-IS PDU are multiple type-length-values (TLVs). **Figure 6-5** shows the TLV format. A TLV is also called a code-length-value (CLV).

Figure 6-5 TLV format



TLVs vary according to PDU types, as shown in **Table 6-1**.

Table 6-1 PDU types and TLV names

TLV Type	Name	Applied 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, SNP
128	IP Internal Reachability Information	LSP
129	Protocols Supported	IIH, LSP
130	IP External Reachability Information	L2 LSP
131	Inter-Domain Routing Protocol Information	L2 LSP
132	IP Interface Address	IIH, LSP

TLVs with the type value ranging from 1 to 10 are defined in ISO 10589, and the other TLVs are defined in RFC 1195.

6.2.2 IS-IS Basic Principles

IS-IS is a link-state routing protocol. Each router generates an LSP that contains link state information about all the IS-IS interfaces on the router. The router can establish IS-IS neighbor relationships with neighboring devices and update its LSDB to synchronize the local LSDB with the LSDBs of all the other devices on the IS-IS network. Based on the local LSDB, the router uses the SPF algorithm to calculate IS-IS routes. If the router finds that an IS-IS route is the optimal route to a destination, the router adds the route to the local IP routing table to guide packet forwarding.

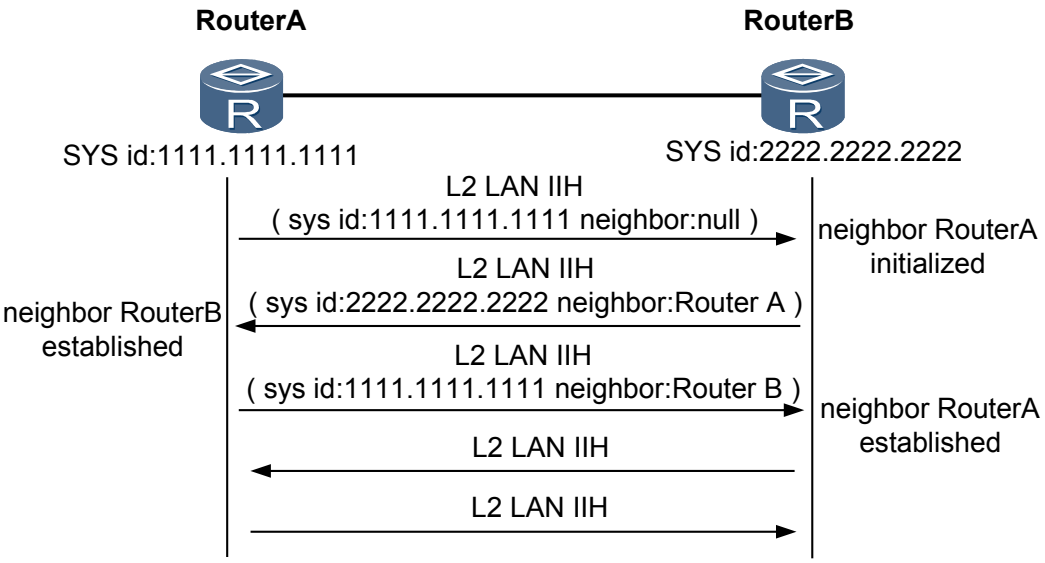
Establishment of IS-IS Neighbor Relationship

Two IS-IS routers need to establish a neighbor relationship before exchanging packets to implement routing. On different networks, the modes for establishing IS-IS neighbors are different.

- Establishment of a neighbor relationship on a broadcast link

  Figure 6-6 uses Level-2 routers as an example to describe the process of establishing a neighbor relationship on a broadcast link. The process of establishing a neighbor relationship between Level-1 routers is the same as the process of establishing a neighbor relationship between Level-2 routers.

  Figure 6-6 Process of establishing a neighbor relationship on a broadcast link



  1. RouterA broadcasts a Level-2 LAN IS-IS Hello PDU (IIH) with no neighbor ID specified.
  2. RouterB receives this packet and sets the status of the neighbor relationship with RouterA to Initial. RouterB then responds to RouterA with a Level-2 LAN IIH, indicating that RouterA is a neighbor of RouterB.
  3. RouterA receives this packet and sets the status of the neighbor relationship with RouterB to Up. RouterA then sends RouterB a Level-2 LAN IIH indicating that RouterB is a neighbor of RouterA.
  4. RouterB receives this packet and sets the status of the neighbor relationship with RouterA to Up. RouterA and RouterB establish a neighbor relationship successfully.

  The network is a broadcast network, so a DIS needs to be elected. After the neighbor relationship is established, routers wait for two intervals before sending Hello packets to elect the DIS. The IIH packets exchanged by the routers contain the Priority field. The router with the highest priority is elected as the DIS. If the routers have the same priority, the router with the largest interface MAC address is elected as the DIS.

- Establishment of a neighbor relationship on a P2P link

Unlike the establishment of a neighbor relationship on a broadcast link, the establishment of a neighbor relationship on a P2P link is classified into two modes: two-way mode and three-way mode.

- Two-way mode

  Upon receiving a P2P IIH from a neighbor, a router considers the neighbor Up and establishes a neighbor relationship with the neighbor.

- Three-way mode

  A neighbor relationship is established after P2P IIHs are sent for three times. The establishment of a neighbor relationship on a P2P link is similar to that on a broadcast link.

Two-way mode has distinct disadvantages. For example, when two or more links exist between two routers, the two routers can still establish a neighbor relationship if one link is Down and the other is Up in the same direction. The parameters of the link in Up state are used in SPF calculation. As a result, the router that does not detect the fault of the link in Down state still tries to forward packets over this link. Three-way mode addresses such problems on unreliable P2P links. In three-way mode, a router considers the neighbor Up only after confirming that the neighbor receives the packet sent by itself, and then establishes a neighbor relationship with the neighbor.

Basic rules for establishing an IS-IS neighbor relationship are as follows:

- Only neighboring routers of the same level can set up the neighbor relationship with each other.
- For Level-1 routers, their area IDs must be the same
- Network types of IS-IS interfaces on both ends of a link must be consistent.

  NOTE

  Ethernet interfaces can be simulated as P2P interfaces to establish a neighbor relationship on a P2P link.

- IP addresses of IS-IS interfaces on both ends of a link must be on the same network segment.

  IS-IS runs on the data-link layer and was initially designed for CLNP. Therefore, the establishment of an IS-IS neighbor relationship is not related to IP addresses. In the implementation of a device, IS-IS runs only over IP. Therefore, IS-IS needs to check the IP address of its neighbor. If secondary IP addresses are assigned to the interfaces, the routers can still set up the IS-IS neighbor relationship, but only when either the primary IP addresses or secondary IP addresses are on the same network segment.

  NOTE

  When IP addresses of IS-IS interfaces on both ends of a link are on different network segments, a neighbor relationship can still be established on the two interfaces if the interfaces are configured not to check the IP addresses in received Hello packets. You can configure P2P interfaces not to check the IP addresses in received Hello packets. Before configuring Ethernet interfaces not to check the IP addresses, simulate Ethernet interfaces as P2P interfaces.

Process of Exchanging IS-IS LSPs

Causes of LSP generation

All routers in the IS-IS routing domain can generate LSPs. The following events trigger the generation of a new LSP:

- Neighbor is Up or Down.
- Related interface goes Up or Down.

- Imported IP routes change.
- Inter-area IP routes change.
- Interface is assigned a new metric value.
- Periodic updates occur.

Processing of a new LSP received from a neighbor

1. The router installs the LSP to its LSDB and marks it for flooding.
2. The router sends the LSP to all interfaces except the interface that initially received the LSP.
3. The neighbors flood the LSP to their neighbors.

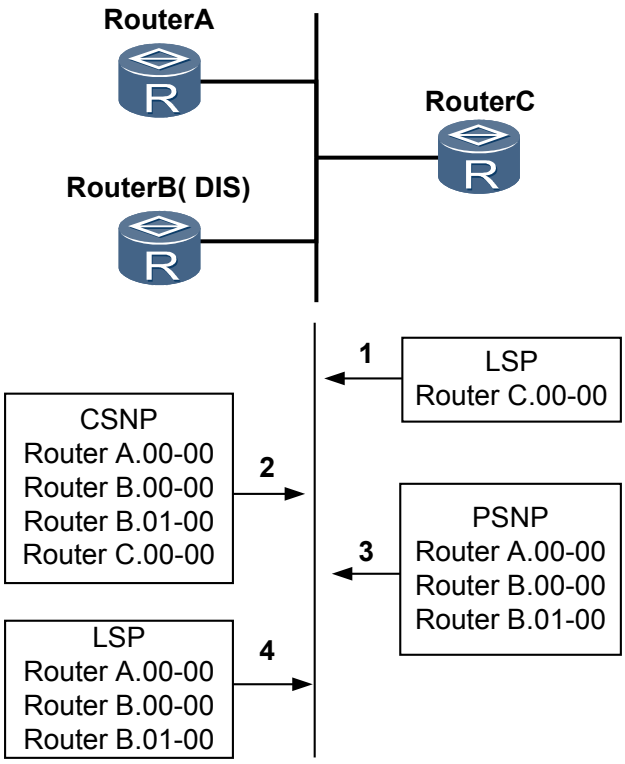
LSP flooding

In LSP flooding, a router sends an LSP to its neighbors and then the neighbors send the received LSP to their respective neighbors except the router that first sends the LSP. In this manner, the LSP is flooded among the routers of the same level. LSP flooding allows each router of the same level to have the same LSP information and synchronize its LSDB with each other.

Each LSP has a 4-byte sequence number. When a router is started, the sequence number of the first LSP sent by the router is 1. When a new LSP is generated, the sequence number of the LSP is equal to the sequence number of the previous LSP plus 1. The greater the sequence number, the newer the LSP.

Process of synchronizing LSDBs between a newly added router and DIS on a broadcast link

Figure 6-7 Process of updating LSDBs on a broadcast link



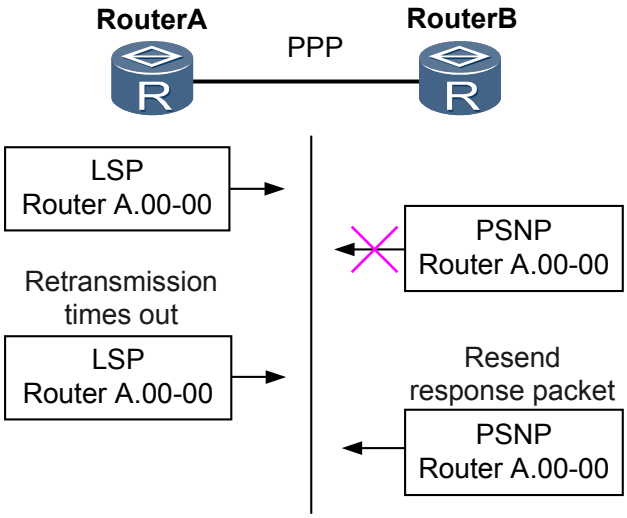
1. A new router (RouterC) sends a Hello packet to establish neighbor relationships with the other routers in the broadcast domain.
2. RouterC establishes neighbor relationships with RouterA and RouterB, waits for the timeout of the LSP refresh timer, and then sends its LSP to a multicast address (01-80-C2-00-00-1 in a Level-1 area and 01-80-C2-00-00-15 in a Level-2 area). All neighbors on the network can receive the LSP.
3. The DIS on the network segment adds the received LSP to its LSDB. After the CSNP timer expires, the DIS sends CSNPs to synchronize the LSDBs on the network.
4. RouterC receives the CSNPs from the DIS, checks its LSDB, and sends a PSNP to request the LSPs it does not have.
5. The DIS receives the PSNP and sends RouterC the required LSPs for LSDB synchronization.

The process of updating the LSDB of the DIS is as follows:

1. When the DIS receives an LSP, it searches the LSDB to check whether the same LSP exists. If the DIS does not find the same LSP in its LSDB, the DIS adds the LSP to its LSDB and broadcasts the content of the new LSDB.
2. If the sequence number of the received LSP is greater than that of the corresponding LSP in the LSDB, the DIS replaces the existing LSP with the received LSP and broadcasts the contents of the new LSDB. If the sequence number of the received LSP is smaller than that of the corresponding LSP in the LSDB, the DIS sends its LSP in the LSDB through the inbound interface of the received LSP.
3. If the sequence number of the received LSP is the same as that of the corresponding LSP in the LSDB, the DIS compares the remaining lifetime of the two LSPs. If the remaining lifetime of the received LSP is smaller than that of the corresponding LSP in the LSDB, the DIS replaces the existing LSP with the received LSP and broadcasts the contents of the new LSDB. If the remaining lifetime of the received LSP is greater than that of the corresponding LSP, the DIS sends its LSP in the LSDB through the inbound interface of the received LSP.
4. If the sequence number and remaining lifetime of the received LSP are the same as those of the corresponding LSP in the LSDB, the DIS compares the checksum of the two LSPs. If the checksum of the received LSP is greater than that of the corresponding LSP in the LSDB, the DIS replaces the existing LSP with the received LSP and broadcasts the content of the new LSDB. If the checksum of the received LSP is smaller than that of the corresponding LSP, the DIS sends its LSP in the LSDB through the inbound interface of the received LSP.
5. If the sequence number, remaining lifetime, and checksum of the received LSP are the same as those of the corresponding LSP in the LSDB, the DIS does not forward the received LSP.

Process of synchronizing the LSDB on a P2P link

Figure 6-8 Process of updating LSDBs on a P2P link



1. RouterA establishes a neighbor relationship with RouterB.
2. RouterA and RouterB send a CSNP to each other. If the LSDB of the neighbor and the received CSNP are not synchronized, the neighbor sends a PSNP to request the required LSP.
3. **Figure 6-8** assumes that RouterB requests the required LSP from RouterA. RouterA sends the required LSP to RouterB, starts the LSP retransmission timer, and waits for a PSNP from RouterB as an acknowledgement for the received LSP.
4. If RouterA does not receive a PSNP from RouterB after the LSP retransmission timer expires, RouterA resends the LSP until it receives a PSNP from RouterB.

NOTE

A PSNP on a P2P link is used as follows:

- An ACK packet to acknowledge the received LSP.
- A request packet to acquire LSPs.

The process of updating LSDBs on a P2P link is as follows:

1. If the sequence number of the received LSP is smaller than that of the corresponding LSP in the LSDB, the router directly sends its LSP to the neighbor and waits for a PSNP from the neighbor. If the sequence number of the received LSP is greater than that of the corresponding LSP in the LSDB, the router adds the received LSP to its LSDB, sends a PSNP to acknowledge the received LSP, and then sends the received LSP to all its neighbors except the neighbor that sends the LSP.
2. If the sequence number of the received LSP is the same as that of the corresponding LSP in the LSDB, the router compares the remaining lifetime of the two LSPs. If the received LSP has a smaller remaining lifetime than that of the corresponding LSP in the LSDB, the router adds the received LSP to its LSDB, sends a PSNP to acknowledge the received LSP, and then sends the received LSP to all its neighbors except the neighbor that sends the LSP. If the received LSP has a greater remaining lifetime than that of the corresponding LSP in the LSDB, the router directly sends its LSP to the neighbor and waits for a PSNP from the neighbor.
3. If the sequence number and remaining lifetime of the received LSP are the same as those of the corresponding LSP in the LSDB, the router compares the checksum of the two LSPs.

If the received LSP has a greater checksum than that of the corresponding LSP in the LSDB, the router adds the received LSP to its LSDB, sends a PSNP to acknowledge the received LSP, and then sends the received LSP to all its neighbors except the neighbor that sends the LSP. If the received LSP has a smaller checksum than that of the corresponding LSP in the LSDB, the router directly sends its LSP to the neighbor and waits for a PSNP from the neighbor.

4. If the sequence number, remaining lifetime, and checksum of the received LSP and the corresponding LSP in the LSDB are the same, the router does not forward the received LSP.

6.2.3 IS-IS Authentication

To ensure network security, IS-IS authentication encrypts IS-IS packets by adding the authentication field to packets. When a local router receives IS-IS packets from a remote router, the local router discards the packets if the authentication passwords do not match. This protects the local router.

Authentication Types

Based on the types of packets, the authentication is classified as follows:

- Interface authentication: authenticates Level-1 and Level-2 Hello packets sent and received on IS-IS interfaces using the specified authentication mode and password.

  NOTE

  You can configure a router to perform interface authentication in the following ways:

  - A router sends authentication packets carrying the authentication TLV and verifies the authentication information about the received packets.
  - A router sends authentication packets carrying the authentication TLV but does not verify the authentication information about the received packets.

- Area authentication: authenticates Level-1 LSPs and Level-1 SNPs transmitted in an IS-IS area using the specified authentication mode and password.
- Routing domain authentication: authenticates Level-2 LSPs and Level-2 SNPs transmitted in an IS-IS routing domain using the specified authentication mode and password.

  NOTE

  In area authentication and routing domain authentication, you can configure a router to authenticate LSPs and SNPs separately in the following ways:

  - A router sends LSPs and SNPs carrying the authentication TLV and verifies the authentication information about the received LSPs and SNPs.
  - A router sends LSPs carrying the authentication TLV and verifies the authentication information about the received LSPs. The router sends SNPs carrying the authentication TLV but does not verify the authentication information about the received SNPs.
  - A router sends LSPs carrying the authentication TLV and verifies the authentication information about the received LSPs. The router sends SNPs without the authentication TLV and does not verify the authentication information about the received SNPs.
  - A router sends LSPs and SNPs carrying the authentication TLV but does not verify the authentication information about the received LSPs and SNPs.

Based on the authentication modes of packets, authentication is classified into the following types:

- Plain text authentication: is a simple authentication mode in which passwords are directly added to packets. This authentication is insecure.

- MD5 authentication: uses the MD5 algorithm to encrypt passwords before they are added to packets, which improves password security.
- Keychain authentication: further improves network security with configurable key chain that changes with time.

Mode in Which Authentication Information Is Carried

IS-IS provides a TLV to carry authentication information, with the type of the TLV specified as 10.

- Type: is defined by the ISO as 0, with a length of 1 byte.
- Length: indicates the length of the authentication TLV, which is 1 byte.
- Value: indicates the authentication contents of 1 to 254 bytes, including the authentication type and password.

  The authentication type is 1 byte:

  - Type 0 is reserved.
  - Type 1 indicates plain text authentication.
  - Type 54 indicates MD5 authentication.
  - Type 255 indicates routing domain private authentication methods.

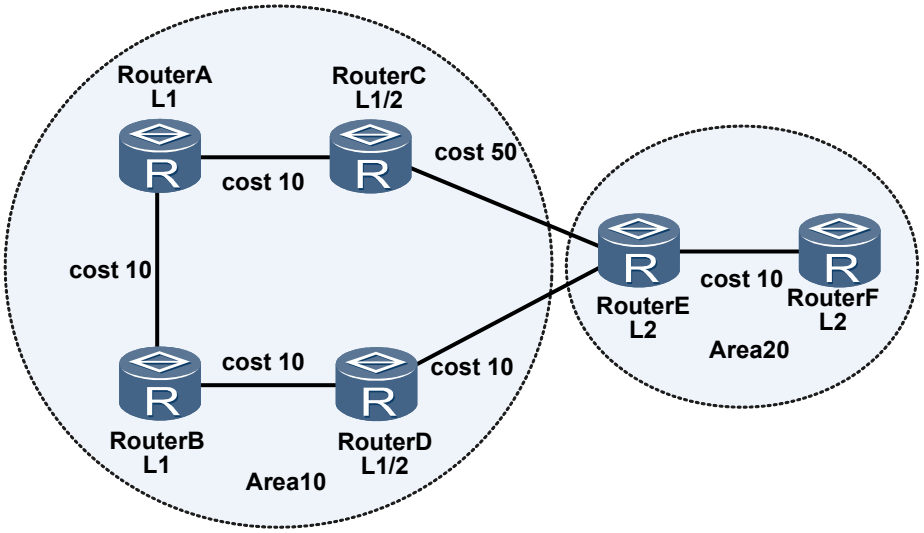
6.2.4 IS-IS Route Leaking

Normally, Level-1 routers manage routes in Level-1 areas. All Level-2 and Level-1-2 routers form a contiguous backbone area. Level-1 areas can only connect to the backbone area, but cannot connect to each other.

A Level-1-2 router encapsulates learned Level-1 routing information into a Level-2 LSP and floods the Level-2 LSP to other Level-2 and Level-1-2 routers. Then Level-1-2 and Level-2 routers know routing information about the entire IS-IS routing domain. To reduce the size of routing tables, a Level-1-2 router, by default, does not advertise the learned routing information of other Level-1 areas and the backbone area to its Level-1 area. In this case, Level-1 routers cannot know routing information outside the local area. As a result, Level-1 routers cannot select the optimal route to the destination outside the local area.

IS-IS route leaking can solve this problem. You can configure access control lists (ACLs) and routing policies and mark routes with tags on Level-1-2 routers to select eligible routes. Then a Level-1-2 router can advertise routing information of other Level-1 areas and backbone area to its Level-1 area.

Figure 6-9 IS-IS route leaking



In **Figure 6-9**, RouterA sends a packet to RouterF. The selected optimal route should be RouterA->RouterB->RouterD->RouterE->RouterF. This is because the cost of this route is 40, which is smaller than the cost (70) of the other route (RouterA->RouterC->RouterE->RouterF). However, when you check the route on RouterA to view the path of the packets sent to RouterF, the selected route is RouterA->RouterC->RouterE->RouterF but not the optimal route from RouterA to RouterF.

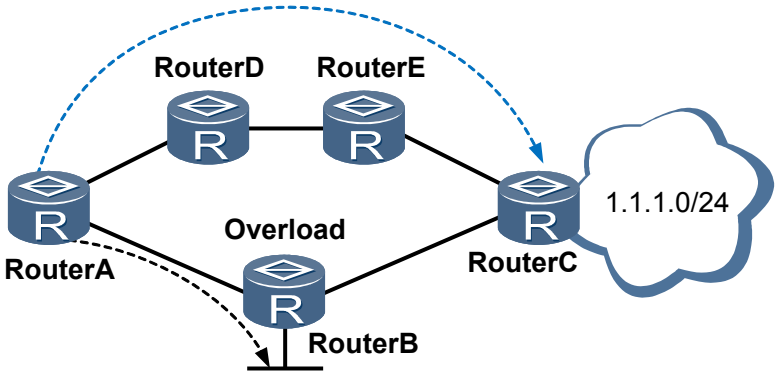
RouterA (Level-1 router) does not know routes outside its area, so it sends packets outside its area through the default route generated by the nearest Level-1-2 router. Therefore, the optimal route is not used to forward the packets.

If route leaking is enabled on Level-1-2 routers (RouterC and RouterD), Level-1 routers in Area 10 can know routes outside Area 10 and passing through the two Level-1-2 routers. After route calculation, the forwarding path becomes RouterA->RouterB->RouterD->RouterE->RouterF, which is the optimal route from RouterA to RouterF.

6.2.5 IS-IS Overload

IS-IS Overload allows a device to use the IS-IS overload bit to identify the overload state. The IS-IS overload bit is the OL field in an IS-IS LSP. After the overload bit is set on a device, other devices ignore this device when performing SPF calculation and consider only the direct routes of the device.

Figure 6-10 IS-IS Overload



As shown in **Figure 6-10**, RouterB forwards the packets sent from RouterA to network segment 1.1.1.0/24. If the overload bit in the LSP sent from RouterB is set to 1, RouterA considers the LSDB of RouterB incomplete and sends packets to 1.1.1.0/24 through RouterD and RouterE. This process does not affect the packets sent to the directly connected network segment of RouterB.

If a device cannot store new LSPs and fails to synchronize the LSDB, the routes calculated by this device are incorrect. In this situation, the device enters the overload state and does not calculate the routes passing through this device; however, the direct routes of the device are still valid.

A device may enter the overload state because of device abnormalities or is manually configured to enter the overload state. When an IS-IS device on the network needs to be upgraded or maintained, isolate this device from the network temporarily and set the overload bit on the device to prevent other devices from using this device to forward traffic.

NOTE

- If the system enters the overload state because of an abnormality, the system deletes all the imported or leaked routes.
- If the system is configured to enter the overload state, the system determines whether to delete all the imported or leaked routes based on the configuration.

6.2.6 IS-IS Network Convergence

Fast convergence and priority-based convergence can improve IS-IS network convergence. Fast convergence speeds up network convergence by fast calculating routes, while priority-based convergence sets different convergence priorities for routes to improve network convergence.

Fast Convergence

IS-IS fast convergence is an extended feature of IS-IS that is implemented to speed up the convergence of routes. Fast convergence includes the following:

- 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 the calculation of routes.

  In ISO 10589, the SPF algorithm is used to calculate routes. When a node changes on the network, this algorithm is used to recalculate all routes. The calculation takes a long time and consumes too many CPU resources, which affects the convergence speed.

I-SPF improves this algorithm. Except for the first time, only changed nodes instead of all nodes are involved in calculation. The shortest path tree (SPT) generated is the same as that generated by the previous algorithm. This decreases CPU usage and speeds up network convergence.

- Partial Route Calculation (PRC): calculates only the changed routes when the routes on the network change.

  Similar to I-SPF, PRC calculates only the changed routes, but it 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. If the SPT changes after I-SPF calculation, PRC processes all the leaves only on the changed node. If the SPT remains unchanged, PRC processes only the changed leaves. For example, if IS-IS is enabled on an interface of a node, the SPT calculated by I-SPF remains unchanged. PRC updates only the routes of this interface, consuming less CPU resources.

  PRC working with I-SPF further improves the convergence performance of the network. It is an improvement of the original SPF algorithm.

- Intelligent timer: applies to LSP generation and SPF calculation. The first timeout period of the intelligent timer is fixed. Before the intelligent timer expires, if an event that triggers the timer occurs, the next timeout period of the intelligent timer increases.

  Although the route calculation algorithm is improved, the long interval for triggering route calculation affects the convergence speed. Frequent network changes also consume too many CPU resources. The SPF intelligent timer addresses both of these problems. In general, an IS-IS network is stable under normal conditions. The probability of the occurrence of many network changes is very minimal, and IS-IS does not calculate routes frequently. The period for triggering the route calculation is very short (milliseconds). If the topology of the network changes very often, the intelligent timer increases the interval for the calculation times to avoid too much CPU consumption. The original mechanism uses a timer with uniform intervals, which makes fast convergence and low CPU consumption impossible to achieve.

  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 LSP generation timer is designed as an intelligent timer to respond to emergencies (such as the interface is Up or Down) quickly and speed up the network convergence.

- LSP fast flooding: speeds up the flooding of LSPs.

  In most cases, when an IS-IS router receives new LSPs from other routers, it updates the LSPs in its LSDB and periodically floods the updated LSPs according to a timer.

  LSP fast flooding speeds up LSDB synchronization because it allows a device to flood fewer LSPs than the specified number before route calculation when the device receives one or more new LSPs. This mechanism also speeds up network convergence.

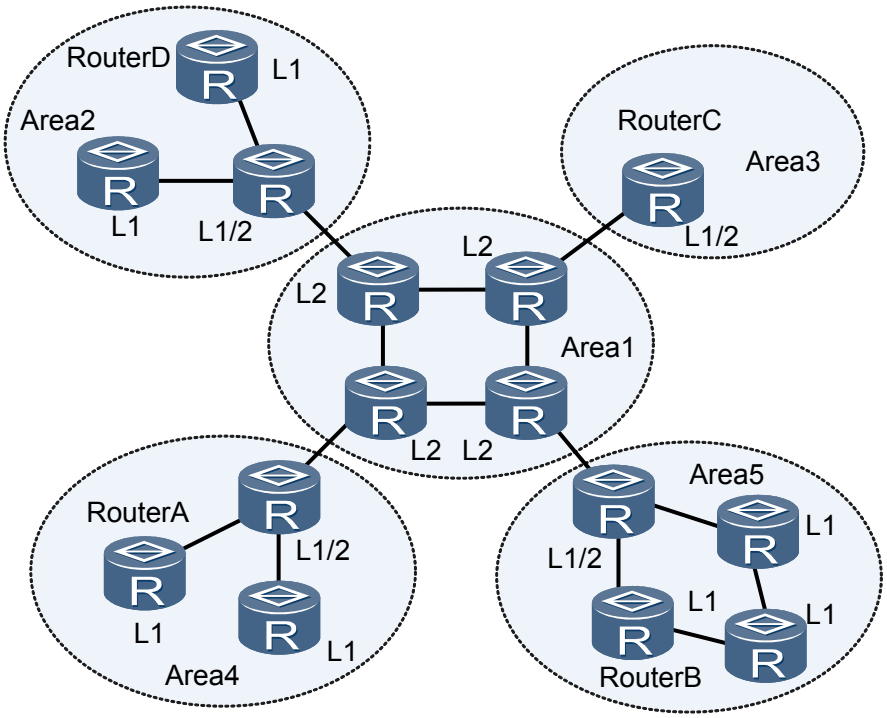
Priority-based Convergence

Priority-based IS-IS convergence ensures that specific routes are converged first when a great number of routes need to be converged. You can assign a high convergence priority to routes for key services so that these routes are converged quickly. This reduces the impact of route convergence on key services. Different routes can be set with different convergence priorities so that important routes can be converged first. This improves network reliability.

6.2.7 IS-IS Administrative Tag

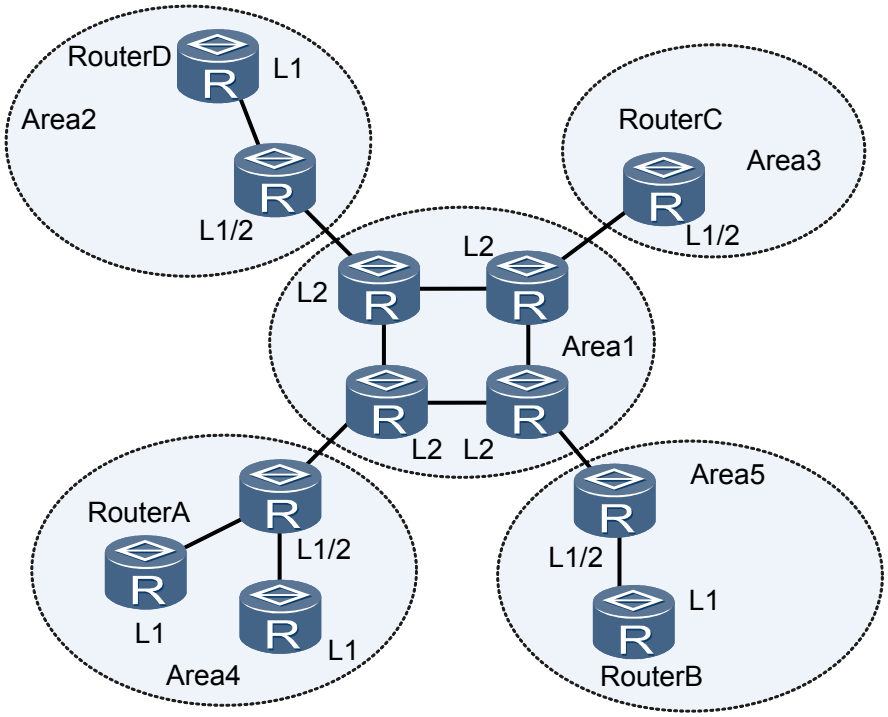
Administrative tags control the advertisement of IP prefixes in an IS-IS routing domain to simplify route management. You can use administrative tags to control the import of routes of different levels and different areas and control IS-IS multi-instances running on the same router.

Figure 6-11 IS-IS networking



In **Figure 6-11**, RouterA in Area 4 needs to communicate with RouterB in Area 5, RouterC in Area 3, and RouterD in Area 2. To ensure information security, it is required that other routers in Level-1 areas (Areas 2, 3, and 5) should not receive the packets sent from RouterA. To meet this requirement, configure the same administrative tag for IS-IS interfaces on RouterB, RouterC, and RouterD and configure the Level-1-2 router in Area 4 to leak only the routes matching the configured administrative tag from Level-2 to Level-1 areas. This allows RouterA to communicate with only RouterB, RouterC, and RouterD. **Figure 6-12** shows the topology formed on RouterA.

Figure 6-12 IS-IS administrative tag application



The value of an administrative tag is associated with certain attributes. If the cost style is wide, wide-compatible or compatible, when IS-IS advertises an IP address prefix with these attributes, IS-IS adds the administrative tag to the TLV in the prefix. The tag is flooded along with the prefix throughout the routing domain.

6.2.8 IS-IS Wide Metric

In ISO 10589, the maximum IS-IS interface metric value can only be 63 and the IS-IS cost style is narrow. A small range of metrics cannot meet the requirements on large-scale networks. Therefore, in RFC 3784, the maximum IS-IS interface metric value can reach 16777215, and the maximum IS-IS route metric value can reach 4261412864; in this case, the IS-IS cost style is wide.

- The following lists the TLVs used in narrow mode:
  - TLV 128 (IP Internal Reachability TLV): carries IS-IS routes in a routing domain.
  - TLV 130 (IP External Reachability TLV): carries IS-IS routes outside a routing domain.
  - TLV 2 (IS Neighbors TLV): carries neighbor information.
- The following lists the TLVs used in wide mode:
  - TLV 135 (Extended IP Reachability TLV): replaces the earlier IP reachability TLV and carries IS-IS routing information. This TLV expands the route metric and carries sub-TLVs.
  - TLV 22 (IS Extended Neighbors TLV): carries neighbor information.

Table 6-2 lists the cost styles of received and sent IS-IS routing information. The cost styles of received and sent IS-IS routing information vary according to the cost style configured on a device.

Table 6-2 Cost styles of received and sent IS-IS routing information

Cost Style Configured on a Device	Cost Style for Received IS-IS Routing Information	Cost Style for Sent IS-IS Routing Information
narrow	narrow	narrow
narrow-compatible	narrow&wide	narrow
compatible	narrow&wide	narrow&wide
wide-compatible	narrow&wide	wide
wide	wide	wide

NOTE

When the cost-style is set to compatible, IS-IS sends the information in narrow mode and then in wide mode.

IS-IS in wide mode and IS-IS in narrow mode cannot communicate. If IS-IS in wide mode and IS-IS in narrow mode need to communicate, you must change the mode to enable all routers on the network to receive packets sent by other routers.

6.2.9 IS-IS LSP Fragment Extension

When an IS-IS router needs to advertise the LSPs that contain much information, the IS-IS router generates multiple LSP fragments to carry more IS-IS information.

IS-IS LSP fragments are identified by the LSP Number field in their LSP IDs. This field is of 1 byte. An IS-IS process can generate a maximum of 256 LSP fragments; therefore, only a limited number of routes can be carried.

As defined in RFC 3786, virtual system IDs can be configured and virtual LSPs that carry routing information can be generated for IS-IS.

Concepts

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

  NOTE

  Like a normal system ID, an additional system ID must be unique in a routing domain.
- TLV 24 (IS Alias ID TLV): describes the relationship between the originating system and virtual system.

Principles

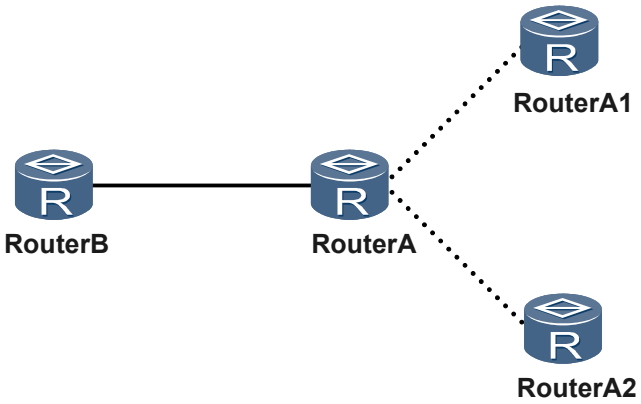
In IS-IS, each system ID identifies a system, which can generate a maximum of 256 LSP fragments. With more additional system IDs (up to 50 virtual systems can be configured), an IS-IS process can generate a maximum of 13,056 LSP fragments.

After LSP fragment extension is configured, the system prompts you to restart the IS-IS process if information is lost because LSPs overflow. After being restarted, the originating system loads as much routing information to LSPs, adds the overloaded information to the LSPs of the virtual system for transmission, and uses TLV 24 to notify other routers of its relationship with the virtual system.

Operating Modes

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

Figure 6-13 IS-IS LSP fragment extension



Operating Mode	Usage Scenario	Principles	Example	Precautions
Mode-1	Some routers on the network do not support LSP fragment extension.	Virtual systems participate in SPF calculation. The originating system advertises LSPs containing information about links to each virtual system. Similarly, each virtual system advertises LSPs containing information about links to the originating system. Virtual systems look like the physical routers that connect to the originating system. Mode-1 is a transitional mode for the earlier versions that do not support LSP fragment extension. In earlier versions, IS-IS cannot identify the IS Alias ID TLV and processes the received LSP that is advertised by a virtual system as an LSP advertised by an IS-IS process.	In **Figure 6-13**, RouterB does not support LSP fragment extension, and RouterA is configured to support LSP fragment extension in mode-1. RouterA1 and RouterA2 are virtual systems of RouterA and send LSPs carrying some routing information of RouterA. After receiving LSPs from RouterA, RouterA1, and RouterA2, RouterB considers that there are three individual routers at the remote end and calculates routes. Because the cost of the route from RouterA to RouterA1 and the cost of the route from RouterA to RouterA2 are both 0, the cost of the route from RouterB to RouterA is the same as the cost of the route from RouterB to RouterA1.	The LSP sent by a virtual system contains the same area address and overload bit as those in a common LSP. If the LSPs sent by a virtual system contain TLVs specified in other features, these TLVs must be the same as those in common LSPs. The virtual system carries neighbor information indicating that the neighbor is the originating system, with the metric equal to the maximum value minus 1. The originating system carries neighbor information indicating that the neighbor is the virtual system, with the metric 0. This ensures that the virtual system is the downstream node of the originating system when other routers calculate routes.

Operating Mode	Usage Scenario	Principles	Example	Precautions
Mode-2	All the routers on the network support LSP fragment extension.	Virtual systems do not participate in SPF calculation. All the routers on the network know that the LSPs generated by virtual systems actually belong to the originating system. An IS-IS router working in mode-2 can identify the IS Alias ID TLV, which is used as a reference for calculating the SPT and routes.	In **Figure 6-13**, RouterB supports LSP fragment extension, and RouterA is configured to support LSP fragment extension in mode-2. RouterA1 and RouterA2 are virtual systems of RouterA and send LSPs carrying some routing information of RouterA. When receiving LSPs from RouterA1 and RouterA2, RouterB obtains the IS Alias ID TLV and knows that the originating system of RouterA1 and RouterA2 is RouterA. RouterB then considers that information advertised by RouterA1 and RouterA2 belongs to RouterA.	-

NOTE

When the originating system and virtual system send the LSPs with fragment number 0, the LSPs must carry the IS Alias ID TLV to indicate the originating system regardless of the operation mode (mode-1 or mode-2).

6.2.10 IS-IS Host Name Mapping

The IS-IS host name mapping mechanism maps host names to system IDs for IS-IS devices, including dynamic host name mapping and static host name mapping. Dynamic host name mapping takes precedence over static host name mapping. When both a dynamic host name and a static host name are configured, the dynamic host name takes effect.

On an IS-IS router where host name exchange is disabled, information about IS-IS neighbors and LSDBs shows that each device in an IS-IS routing domain is identified by a system ID with 12-digit hexadecimal number, for example, aaaa.eeee.1234. This device identification method is complex and not easy to use. The host name exchange mechanism facilitates IS-IS network management and maintenance.

The system ID is replaced by a host name in the following situations:

- When an IS-IS neighbor is displayed, the system ID of the IS-IS neighbor is replaced by its host name. When the neighbor is the DIS, the system ID of the DIS is also replaced by its host name.
- When an LSP in the IS-IS LSDB is displayed, the system ID in the LSP ID is replaced by the host name of the IS-IS device that advertises the LSP.
- When details about the IS-IS LSDB are displayed, the Host Name field is added to the LSP generated by the device where dynamic host name exchange is enabled, and the system ID in the Host Name field is replaced by the dynamic host name of the device that generates the LSP.

Dynamic Host Name Mapping

On a device where dynamic host name mapping is enabled, dynamic host name information is advertised as TLV 137 (Dynamic Hostname TLV) in LSPs. When you run IS-IS commands on other devices to view IS-IS information, the system ID of the local device is replaced by the configured host name. The host name is easier to identify and memorize than the system ID.

The Dynamic Hostname TLV is optional and can be inserted anywhere in an LSP. The value of this TLV cannot be empty. A device can determine whether to send LSPs carrying TLV 137, while the device that receives LSPs can determine whether to ignore TLV 137 or whether to obtain TLV 137 for its mapping table.

Static Host Name Mapping

Static host name mapping allows you to configure the mapping between host names and system IDs of other IS-IS devices on a device. Static host name mapping takes effect only on the local device and is not advertised using LSPs.

6.2.11 IS-IS Reliability

As networks develop, services have higher network requirements. IS-IS provides high reliability to ensure uninterrupted service forwarding when a network fault occurs or when network devices need maintenance.

IS-IS reliability includes hot standby, non-stop routing (NSR), batch backup, and real-time backup, **IS-IS GR**, **BFD for IS-IS**, and **IS-IS Auto FRR**.

In hot standby, IS-IS backs up data from the Active Main Board (AMB) to the Standby Main Board (SMB). Whenever the AMB fails, the SMB becomes active and takes over the tasks of the AMB to ensure normal IS-IS running. This improves IS-IS reliability.

IS-IS information backup includes data backup and command line backup:

- Data backup: The system backs up data of processes and interfaces.

  Data backup ensures the same IS-IS data on the AMB and SMB. When an AMB/SMB switchover occurs, neighbors do not detect the switchover.
- Command line backup: The system backs up the command lines that are successfully executed on the AMB to the SMB.

  Whether to send command lines to the SMB for processing is determined by the the execution results of command lines on the AMB. If command lines are successfully executed on the AMB, the command lines are sent to the SMB for processing. Otherwise, the command lines are not sent to the SMB and the command line execution failure is logged. If the command lines fail to be executed on the SMB, this failure is logged.

The AMB sends only the successfully executed command lines to the SMB for processing. If a fault occurs on the AMB, IS-IS neighbor relationships on the device need to be established again after the AMB/SMB switchover is performed.

Hot Standby

Devices with distributed architecture support IS-IS hot standby.

In IS-IS hot standby, IS-IS configurations on the AMB and SMB are consistent. When an AMB/SMB switchover occurs, the new AMB performs GR and resends a request for establishing neighbor relationships to neighbors to synchronize its LSDB. This prevents traffic transmission from being affected.

NSR

NSR ensures continuous service forwarding on a device when a hardware or software failure occurs on the device. NSR uses data backup to ensure that a neighbor of a device does not detect the fault on the AMB of the device that provides the SMB. NSR ensures that the neighbor relationships established using routing protocols, MPLS, and other protocols that transmit services are not interrupted when a device fault occurs.

IS-IS NSR ensures that data is synchronized in real time between the AMB and SMB. When the AMB/SMB switchover occurs, the SMB can rapidly take over services on the AMB. This ensures that neighbors do not detect device faults.

Batch Backup

- Batch data backup

  When the SMB is installed, all data of the AMB is backed up to the SMB at a time. No configuration can be changed during batch backup.
- Batch command line backup

  When the SMB is installed, all configurations of the AMB are backed up to the SMB at a time. No configuration can be changed during batch backup.

Real-time Backup

- Real-time data backup

  Changed data of processes and interfaces are backed up in real time to the SMB.
- Real-time command line backup

  The command lines that are executed successfully on the AMB are backed up to the SMB.

6.2.12 IS-IS GR

IS-IS graceful restart (GR) is a high availability technology that implements non-stop data forwarding.

After the master/slave switchover, no neighbor information is stored on the restarted router. The first Hello packets sent by the router after restart do not contain the neighbor list. After receiving the Hello packets, the neighbor checks the two-way neighbor relationship and detects that it is not in the neighbor list of the Hello packets sent by the router. The neighbor relationship is interrupted. The neighbor then generates new LSPs and floods the topology changes to all other routers in the area. Routers in the area calculate routes based on the new LSDBs, which leads to route interruption or routing loops.

The IETF defined the GR standard, RFC 3847, for IS-IS. The restart of the protocol is processed for both the reserved FIB tables and unreserved FIB tables. Therefore, the route flapping and interruption of the traffic forwarding caused by the restart can be avoided.

Concepts

IS-IS GR involves two roles, namely, GR restarter and GR helper.

- GR restarter: is a device that has the GR capability and restarts in GR mode.
- GR helper: is a device that has the GR capability and helps the GR restarter complete the GR process. The GR restarter must have the GR helper capability.

To implement GR, IS-IS uses TLV 211 (restart TLV) and three timers, T1, T2, and T3.

Restart TLV

The restart TLV is an extended part of an IS-to-IS Hello (IIH) PDU. All IIH packets of the router that supports IS-IS GR contain the restart TLV. The restart TLV carries the parameters for the protocol restart. **Figure 6-14** shows the format of the restart TLV.

Figure 6-14 Restart TLV

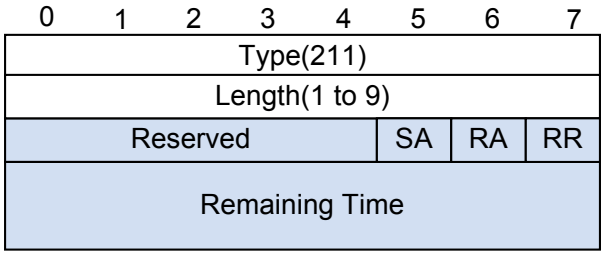


Table 6-3 describes the fields of the restart TLV.

Table 6-3 Restart TLV fields

Field	Length	Description
Type	1 byte	TLV type. Type value 211 indicates the restart TLV.
Length	1 byte	Length of value in the TLV.
RR	1 bit	Restart request bit. A router sends an RR packet to notify the neighbors of its restarting or starting and to require the neighbors to retain the current IS-IS adjacency and return CSNPs.
RA	1 bit	Restart acknowledgement bit. A router sends an RA packet to respond to the RR packet.
SA	1 bit	Suppress adjacency advertisement bit. The starting router uses an SA packet to require its neighbors to suppress the broadcast of their neighbor relationships to prevent routing loops.

Field	Length	Description
Remaining Time	2 bytes	Time during which the neighbor does not reset the adjacency. The length of the field is 2 bytes. The time is measured in seconds. When RA is reset, the value is mandatory.

Timers

Three timers are introduced to enhance IS-IS GR: T1, T2, and T3.

- T1: If the GR restarter has already sent an IIH packet with RR being set but does not receive any IIH packet that carries the restart TLV and the RA set from the GR helper even after the T1 timer expires, the GR restarter resets the T1 timer and continues to send the restart TLV. If the ACK packet is received or the T1 timer expires three times, the T1 timer is deleted. The default value of a T1 timer is 3 seconds.

  Any interface enabled with IS-IS GR maintains a T1 timer. On a Level-1-2 router, broadcast interfaces maintain a T1 timer for Level-1 and Level-2 neighbor relationships.
- T2: is the time from when the GR restarter restarts until the LSDBs of all devices of the same level are synchronized. T2 is the maximum time that the system waits for synchronization of all LSDBs. T2 is generally 60 seconds.

  Level-1 and Level-2 LSDBs maintain their respective T2 timers.
- T3: is the maximum time during which the GR restarter performs GR. The T3 initial value is 65535 seconds. After the IIH packets that carry the RA are received from neighbors, the T3 value becomes the smallest value among the Remaining Time fields of the IIH packets. If the T3 timer expires, GR fails.

  The entire system maintains a T3 timer.

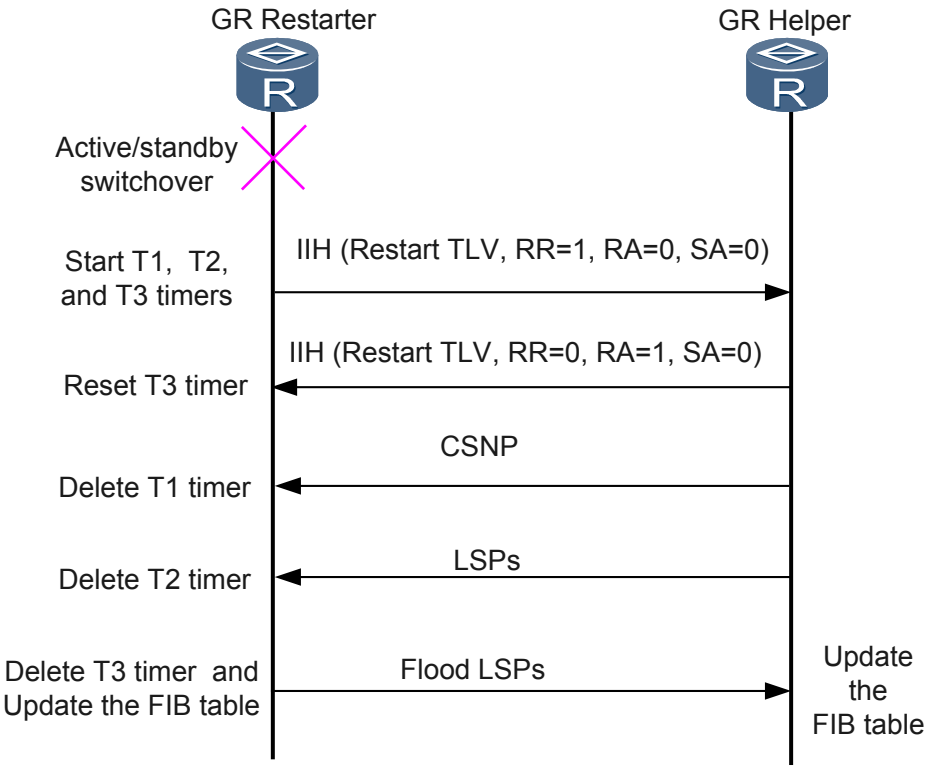
Session Mechanism

For differentiation, GR triggered by the master/slave switchover or the restart of an IS-IS process is referred to as restarting. In restarting, the FIB table remains unchanged. GR triggered by router restart is referred to as starting. In starting, the FIB table is updated.

The following describes the process of IS-IS GR in restarting and starting modes:

- **Figure 6-15** shows the process of IS-IS restarting.

Figure 6-15 IS-IS restarting



1. After performing the protocol restart, the GR restarter performs the following actions:
   - Starts T1, T2, and T3 timers.
   - Sends IIH packets that contain the restart TLV from all interfaces. In such a packet, RR is set to 1, and RA and SA are set to 0.
2. After receiving an IIH packet, the GR helper performs the following actions:
   - Maintains the neighbor relationship and refreshes the current Holdtime.
   - Replies with an IIH packet containing the restart TLV. In the packet, RR is set to 0; RA is set to 1, and the value of the Remaining Time field indicates the period from the current moment to the timeout of the Holdtime.
   - Sends CSNPs and all LSPs to the GR restarter.

   NOTE

   On a P2P link, a neighbor must send CSNPs.

   On a LAN link, only the neighbor of the DIS sends CSNPs. If the DIS is restarted, a temporary DIS is elected from the other routers on the LAN.

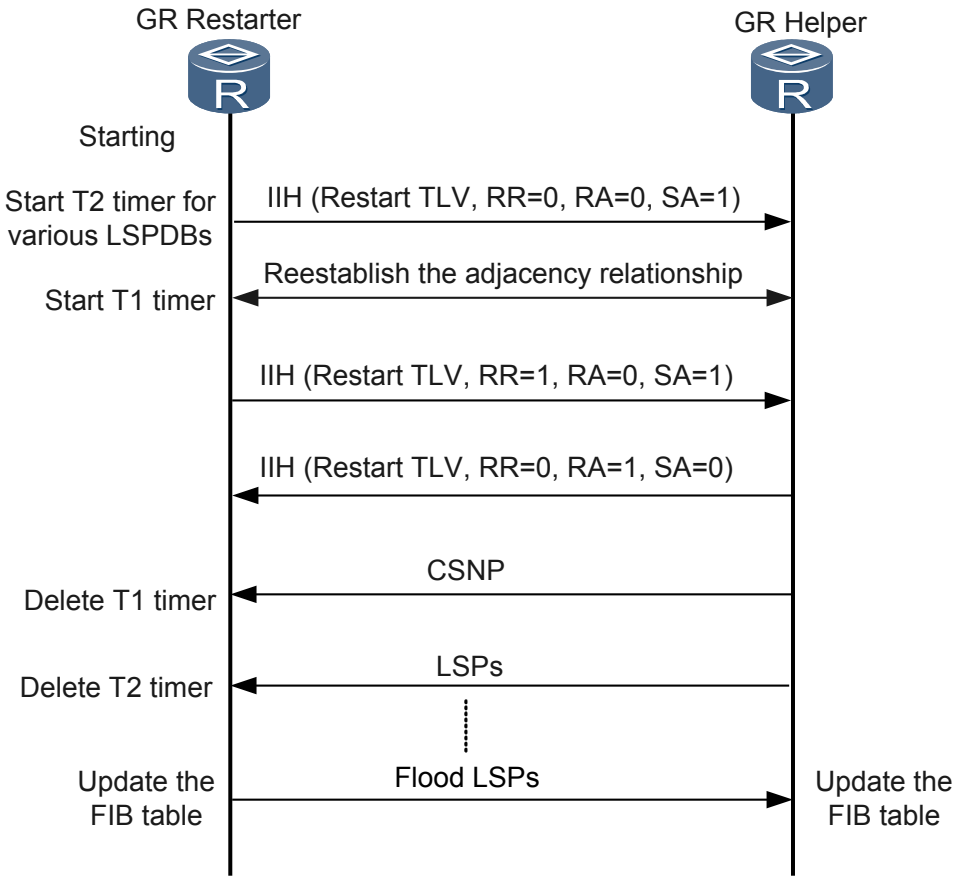
   If the neighbor does not have the GR helper capability, it ignores the restart TLV and resets the adjacency with the GR restarter according to normal IS-IS processing.
3. After the GR restarter receives the IIH response packet, in which RR is set to 0 and RA is set to 1, from the neighbor, it performs the following actions:
   - Compares the current value of the T3 timer with the value of the Remaining Time field in the packet. The smaller value is taken as the value of the T3 timer.

- Deletes the T1 timer maintained by the interface that receives the ACK packet and CSNPs.
- If the interface does not receive the ACK packet or CSNPs, the GR restarter constantly resets the T1 timer and resends the IIH packet that contains the restart TLV. If the number of timeouts of the T1 timer exceeds the threshold value, the GR restarter forcibly deletes the T1 timer and turns to the normal IS-IS processing to complete LSDB synchronization.

4. After the GR restarter deletes the T1 timers on all interfaces, the synchronization with all neighbors is complete when the CSNP list is cleared and all LSPs are collected. The T2 timer is then deleted.
5. After the T2 timer is deleted, the LSDB of the level is synchronized.
   - In the case of a Level-1 or Level-2 router, SPF calculation is triggered.
   - In the case of a Level-1-2 router, determine whether the T2 timer on the router of the other level is also deleted. If both T2 timers are deleted, SPF calculation is triggered. Otherwise, the router waits for the T2 timer of the other level to expire.
6. After all T2 timers are deleted, the GR restarter deletes the T3 timer and updates the FIB table. The GR restarter re-generates the LSPs of each level and floods them. During LSDB synchronization, the GR restarter deletes the LSPs generated before restarting.
7. At this point, the IS-IS restarting of the GR restarter is complete.

- The starting device does not retain the FIB table. The starting device depends on the neighbors, whose adjacency with itself is Up before it starts, to reset their adjacency and suppress the neighbors from advertising their adjacency. The IS-IS starting process is different from the IS-IS restarting process, as shown in **Figure 6-16**.

Figure 6-16 IS-IS starting



1. After the GR restarter is started, it performs the following actions:
   - Starts the T2 timer for the synchronization of LSDBs of each level.
   - Sends IIH packets that contain the restart TLV from all interfaces.

     If RR in the packet is set to 0, a router is started.

     If SA in the packet is set to 1, the router requests its neighbor to suppress the advertisement of their adjacency before the neighbor receives the IIH packet in which SA is set to 0.
2. After the neighbor receives the IIH packet that carries the restart TLV, it performs the following actions depending on whether GR is supported:
   - GR is supported.

     Re-initiates the adjacency.

     Deletes the description of the adjacency with the GR restarter from the sent LSP. The neighbor also ignores the link connected to the GR restarter when performing SPF calculation until it receives an IIH packet in which SA is set to 0.
   - GR is not supported.

     Ignores the restart TLV and resets the adjacency with the GR restarter.

     Replies with an IIH packet that does not contain the restart TLV. The neighbor then returns to normal IS-IS processing. In this case, the neighbor does not suppress

the advertisement of the adjacency with the GR restarter. On a P2P link, the neighbor also sends a CSNP.

3. After the adjacency is re-initiated, the GR restarter re-establishes the adjacency with the neighbors on each interface. When an adjacency set on an interface is in the Up state, the GR restarter starts the T1 timer for the interface.
4. After the T1 timer expires, the GR restarter sends an IIH packet in which both RR and SA are set to 1.
5. After the neighbor receives the IIH packet, it replies with an IIH packet, in which RR is set to 0 and RA is set to 1, and sends a CSNP.
6. After the GR restarter receives the IIH ACK packet and CSNP from the neighbor, it deletes the T1 timer.

   If the GR restarter does not receive the IIH packet or CSNP, it constantly resets the T1 timer and resends the IIH packet in which RR and SA are set to 1. If the number of the timeouts of the T1 timer exceeds the threshold value, the GR restarter forcibly deletes the T1 timer and turns to the normal IS-IS processing to complete LSDB synchronization.
7. After receiving the CSNP from the helper, the GR restarter synchronizes the LSDB.
8. After the LSDB of this level is synchronized, the T2 timer is deleted.
9. After all T2 timers are deleted, the SPF calculation is started and LSPs are regenerated and flooded.
10. At this point, the IS-IS starting of the GR restarter is complete.

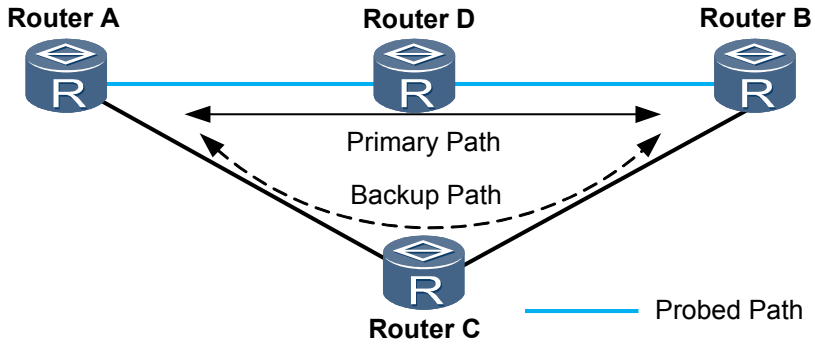
6.2.13 BFD for IS-IS

In IS-IS, the interval for sending Hello packets is 10s, and the holddown time for keeping the neighbor relationship is three times the interval for sending Hello packets. If a router does not receive a Hello packet from its neighbor within the holddown time, the router deletes the corresponding neighbor relationship. This indicates that the router detects neighbor faults in seconds. Second-level fault detection, however, may result in heavy packet loss on high-speed networks.

Bidirectional forwarding detection (BFD) provides light-load and millisecond-level link fault detection to prevent heavy packet loss. BFD is not used to substitute the Hello mechanism of IS-IS but helps IS-IS rapidly detect the faults on neighbors or links and instructs IS-IS to recalculate routes for packet forwarding.

In **Figure 6-17**, basic IS-IS functions are configured on every router, and BFD for IS-IS is enabled on RouterA and RouterB.

Figure 6-17 BFD for IS-IS



When a fault occurs on the primary link (RouterA->RouterD->RouterB), BFD fast detects the fault and reports it to IS-IS. IS-IS sets the neighbors of the interface on the faulty link to Down, which triggers topology calculation, and updates LSPs so that neighbors such as RouterC can receive the updated LSPs from RouterB. This process implements fast network convergence.

Classification of BFD for IS-IS

BFD for IS-IS includes static BFD for IS-IS and dynamic BFD for IS-IS.

Table 6-4 Two implementation modes for BFD for IS-IS

Implementation Mode	Principles	Differences
Static BFD for IS-IS	BFD session parameters, including local and remote discriminators, are manually configured using commands, and the requests for establishing BFD sessions are manually delivered.	● Static BFD can be manually controlled and is easy to deploy. To save memory and ensure reliability of key links, deploy BFD on specified links. ● Establishing and deleting BFD sessions need to be manually triggered and lack flexibility. Configuration errors may occur. For example, if an incorrect local or remote discriminator is configured, a BFD session cannot work properly.
Dynamic BFD for IS-IS	BFD sessions are dynamically created but not manually configured. When detecting faults, BFD informs IS-IS of the faults through the routing management (RM) module. IS-IS then turns the neighbors Down, rapidly advertises the changed LSPs, and performs incremental SPF. This implements fast route convergence.	Dynamic BFD is more flexible than static BFD. In dynamic BFD, routing protocols trigger the setup of BFD sessions, preventing the configuration errors caused by manual configuration. Dynamic BFD is easy to configure and applies to the scenarios where BFD needs to be configured on the entire network.

NOTE

BFD uses local and remote discriminators to differentiate multiple BFD sessions between the same pair of systems.

Because IS-IS establishes only single-hop neighbors, BFD for IS-IS detects only single-hop links between IS-IS neighbors.

Establishment and Deletion of BFD Sessions

The RM module provides related services for association with the BFD module for IS-IS. Through RM, IS-IS prompts BFD to set up or tear down BFD sessions by sending notification messages. In addition, BFD events are transmitted to IS-IS through RM.

Conditions for setting up a BFD session

- Basic IS-IS functions are configured on each router and IS-IS is enabled on the interfaces of the routers.
- BFD is globally enabled on each router, and BFD is enabled on a specified interface or process.
- BFD is enabled on interfaces or processes, and the neighbors are Up. A DIS needs to be elected on a broadcast network.

Process of setting up a BFD session

- P2P network

  After the conditions for setting up a BFD session are satisfied, IS-IS instructs BFD through RM to directly set up a BFD session between neighbors.

- Broadcast network

  After the conditions for establishing BFD sessions are met, and the DIS is elected, IS-IS instructs BFD through RM to establish a BFD session between the DIS and each router. No BFD session is established between non-DISs.

  NOTE

  On a broadcast network, routers (including non-DIS routers) of the same level on a network segment can establish neighbor relationships. In the implementation of BFD for IS-IS, however, BFD sessions are established only between a DIS and a non-DIS. On a P2P network, BFD sessions are directly established between neighbors.

  If a Level-1-2 neighbor relationship is set up between two routers on a link, IS-IS sets up two BFD sessions for the Level-1 and Level-2 neighbors on a broadcast network, but sets up only one BFD session on a P2P network.

Conditions for tearing down a BFD session

- P2P network

  When a neighbor relationship that was set up on P2P interfaces by IS-IS is down (that is, the neighbor relationship is not in the Up state) or when the IP protocol type of a neighbor is deleted, IS-IS tears down the BFD session.

- Broadcast network

  When a neighbor relationship that was set up on P2P interfaces by IS-IS is torn down (that is, the neighbor relationship is not in the Up state), when the IP protocol type of a neighbor is deleted, or when the DIS is re-elected, IS-IS tears down the BFD session.

NOTE

After dynamic BFD is globally disabled in an IS-IS process, the BFD sessions on all the interfaces in this IS-IS process are deleted.

IS-IS Responding to BFD Session Down Event

When detecting a link failure, BFD generates a Down event, and then notifies RM of the event. RM then instructs IS-IS to deletes the neighbor relationship. IS-IS recalculates routes to speed up route convergence on the entire network.

When both the local router and its neighbor are Level-1-2 routers, they establish two neighbors of different levels. Then IS-IS establishes two BFD sessions for the Level-1 neighbor and Level-2

neighbor respectively. When BFD detects a link failure, it generates a Down event and informs the RM module of the event. The RM module then instructs IS-IS to delete the neighbor relationship of a specific level.

6.2.14 IS-IS Auto FRR

With the development of networks, the services such as Voice over IP (VoIP) and online video services require high-quality real-time transmission. Nevertheless, if an IS-IS link fault occurs, traffic can be switched to a new link only after the processes, including fault detection, LSP update, LSP flooding, route calculation, and FIB entry delivery, are complete. As a result, it takes much more than 50 ms to rectify the fault, which cannot meet the requirement for real-time transmission services on the network.

Complying with RFC 5286 (Basic Specification for IP Fast Reroute Loop-Free Alternates), IS-IS Auto FRR protects traffic when links or nodes become faulty. IS-IS Auto FRR allows the forwarding system to rapidly detect such faults and take measures to restore services as soon as possible.

In most cases, you can bind BFD to IS-IS Auto FRR to ensure that the fault recovery time is within 50 ms. When BFD detects a link fault on an interface, the BFD session goes Down, triggering FRR on the interface. Subsequently, traffic is switched from the faulty link to the backup link, which protects services.

Principles

IS-IS Auto FRR pre-computes a backup link by using the Loop-Free Alternate (LFA) algorithm, and then adds the backup link and the primary link to the forwarding table. In the case of an IS-IS network failure, IS-IS Auto FRR can fast switch traffic to the backup link before routes on the control plane converge. This ensures normal transmission of traffic and improves the reliability of the IS-IS network.

The backup link is calculated through the LFA algorithm. With the neighbor that can provide the backup link being the root, the shortest path to the destination node is calculated by a device through the SPF algorithm. Then, the loop-free backup link is calculated according to the inequality defined in RFC 5286.

IS-IS Auto FRR can filter backup routes that need to be added to the IP routing table. Only the backup routes matching the filtering policy are added to the IP routing table. In this manner, users can flexibly control the addition of IS-IS backup routes to the IP routing table.

Applications

IS-IS Auto FRR support traffic engineering (TE) links, including the following types:

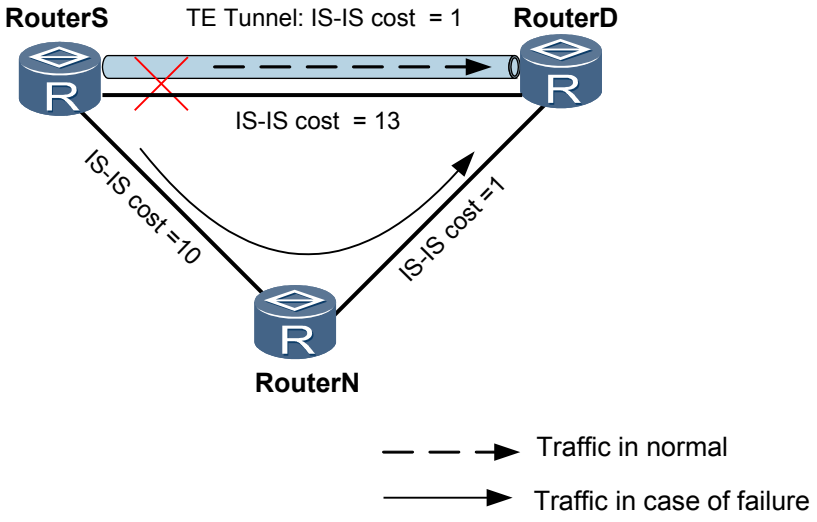
- IP protecting TE

  As shown in **Figure 6-18**, the TE tunnel has the smallest IS-IS cost among the paths from Router S to Router D. Therefore, Router S selects the TE tunnel as the primary path to Router D. The path Router S->Router N->Router D has the second smallest cost. According to the LFA algorithm, Router S selects the path Router S->Router N->Router D as the backup path. The outbound interface of the backup path is the interface that connects Router S to Router N.

  NOTE

  If the outbound interface of the backup link is the actual outbound interface of the TE tunnel, IP protecting TE fails.

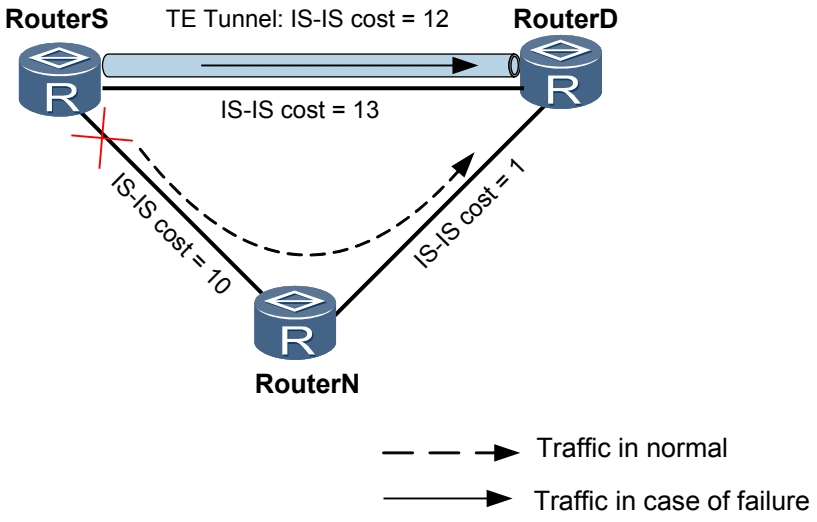
Figure 6-18 IP protecting TE



- TE protecting IP

  As shown in **Figure 6-19**, the physical path Router S-->Router N-->Router D has the smallest IS-IS metric among the paths from Router S to Router D. Therefore, Router S prefers the path Router S-->Router N-->Router D as the primary path from Router S to Router D. The IS-IS cost of the TE tunnel is 12, and the explicit path of the TE tunnel is the direct link from Router S to Router D. The IS-IS metric of the direct link from Router S to Router D is 13, which is greater than the IS-IS metric of the TE tunnel. Therefore, IS-IS selects the TE tunnel as the backup path. TE protecting IP is implemented.

Figure 6-19 TE protecting IP



IS-IS Auto FRR traffic protection is classified into link protection and link-node dual protection.

Figure 6-20 IS-IS Auto FRR link protection

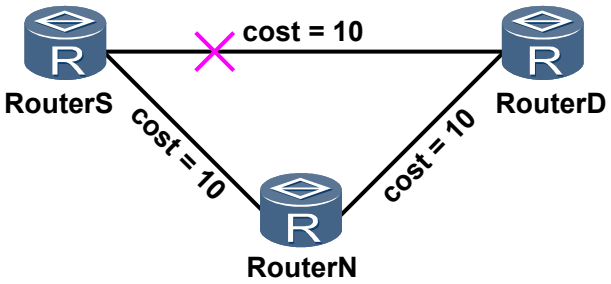


Figure 6-21 IS-IS Auto FRR link-node dual protection

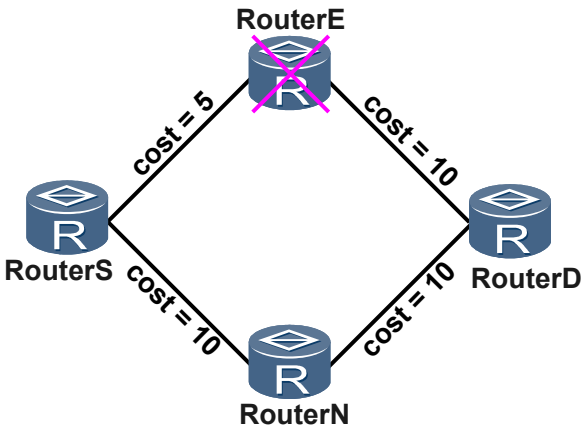


Table 6-5 IS-IS Auto FRR traffic protection

Traffic Protection Type	Object Protected	Condition	Application Example
Link protection	Traffic passing through a specific link	The link cost must satisfy the following inequality: Distance_opt(N,D) < Distance_opt(N,S) + Distance_opt(S,D)	In **Figure 6-20**, traffic is transmitted from RouterS to RouterD. The link cost satisfies the link protection inequality. When the primary link fails, RouterS switches the traffic to the backup link RouterS->RouterN so that the traffic can be further transmitted along downstream paths. This ensures that the traffic interruption time is within 50 ms.
Link-node dual protection	Next-hop node or link from the local node to the next-hop node. Node protection takes precedence over link protection.	Link-node dual protection must satisfy the following conditions: ● The link cost must satisfy the following inequality: Distance_opt(N,D) < Distance_opt (N,S) + Distance_opt(S,D) ● The interface cost of the router must satisfy the following inequality: Distance_opt(N,D) < Distance_opt (N,E) + Distance_opt(E,D)	In **Figure 6-21**, traffic is transmitted along the path RouterS->RouterE->RouterD. The link cost satisfies the link protection inequality. When RouterE or the link between RouterS and RouterE fails, RouterS switches the traffic to the backup link RouterS->RouterN so that the traffic can be further transmitted along downstream paths. This ensures that the traffic interruption time is within 50 ms.

NOTE

In **Table 6-5**, Distance_opt(X,Y) indicates the cost of the optimal path between node X and node Y. S indicates the source node of traffic; E indicates the faulty node; N indicates the node on the backup link; D indicates the destination node of traffic.

6.2.15 IS-IS TE

Traditional routers select the shortest path as the master route regardless of other factors, such as bandwidth. In this manner, the traffic is not switched to other paths even if a path is congested.

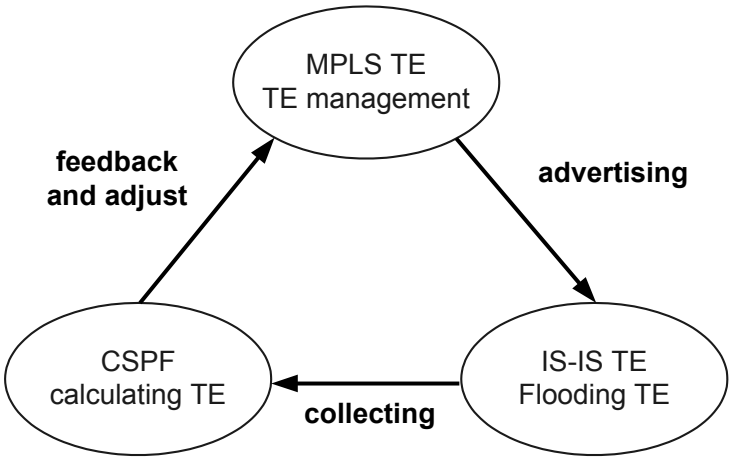
MPLS traffic engineering (TE) has advantages in solving the problem of network congestion. With MPLS TE, you can precisely control the traffic path and prevent traffic from passing through congested nodes. Meanwhile, MPLS TE can reserve resources to ensure the quality of services during the establishment of LSPs.

To ensure the continuity of services, MPLS TE introduces the LSP backup and fast reroute (FRR) mechanisms. When faults occur on the link, the traffic can be switched immediately. Through MPLS TE, service providers (SPs) can fully utilize the current network resources to provide diversified services, optimize network resources, and scientifically manage the network.

To achieve the preceding purpose, MPLS needs to learn TE information of all routers in this network. MPLS TE lacks such a mechanism through which each router floods its TE information in the entire network to implement the synchronization of TE information. This mechanism is provided by the IS-IS protocol. Therefore, MPLS TE can advertise and synchronize TE information with the help of the IS-IS protocol.

IS-IS TE is an extension of IS-IS to support MPLS TE and complies with RFC 5305 and RFC 4205. IS-IS TE defines new TLVs in IS-IS LSPs to carry TE information and floods LSPs to flood and synchronize TE information. It extracts TE information from all LSPs and then transmits the TE information to the Constraint Shortest Path First (CSPF) module of MPLS for tunnel path calculation. IS-IS TE plays the role of a porter in MPLS TE. **Figure 6-22** shows the relationships between IS-IS TE, MPLS TE, and CSPF.

Figure 6-22 Relationships between MPLS TE, CSPF, and IS-IS TE



New TLVs in IS-IS TE

To carry TE information in LSPs, IS-IS TE defines the following TLVs in RFC 5305:

- Extended IS reachability TLV

  This TLV takes the place of IS reachability TLV and extends the TLV formats with sub-TLVs. Sub-TLVs are implemented in TLVs in the same manner as TLVs are implemented in LSPs. Sub-TLVs are used to carry TE information configured on physical interfaces.

  NOTE

  Currently, all sub-TLVs defined in RFC 5305 and sub-TLV type 22 defined in RFC 4124 are supported.

Table 6-6 Sub-TLVs defined in Extended IS reachability TLV

Name	Type	Length (Byte)	Value
Administrative Group	3	4	Indicates the administrative group.
IPv4 Interface Address	6	4	Indicates the IPv4 address of a local interface.
IPv4 Neighbour Address	8	4	Indicates the IPv4 address of a neighbor interface.
Maximum Link Bandwidth	9	4	Indicates the maximum bandwidth of a link.
Maximum Reserved Link Bandwidth	10	4	Indicates the maximum reserved bandwidth of a link.
Unreserved Bandwidth	11	32	Indicates the unreserved bandwidth.
Traffic Engineering Default Metric	18	3	Indicates the default metric of TE.
Bandwidth Constraints sub-TLV	22	36	Indicates the TLV of the bandwidth constraint.

- Traffic Engineering router ID TLV

  It is of TLV type 134, with a 4-byte Router ID. It is used as the MPLS LSR ID. In MPLS TE, a Router ID uniquely identifies a router. Each router has a Router ID.

- Extended IP reachability TLV

  This TLV takes the place of IP reachability TLV and carries routing information. It extends the length of the route cost field and carries sub-TLVs.

- Shared Risk Link Group TLV

  It is of TLV type 138 and used to carry information about the shared risk link group. This TLV can carry information about multiple shared links, each of which is a 4-byte positive integer.

IS-IS TE Implementation

IS-IS TE is implemented in two processes.

- Process of responding to MPLS TE configurations.

  IS-IS TE functions only after MPLS TE is enabled.

  IS-IS TE updates the TE information in IS-IS LSPs based on MPLS TE configurations.

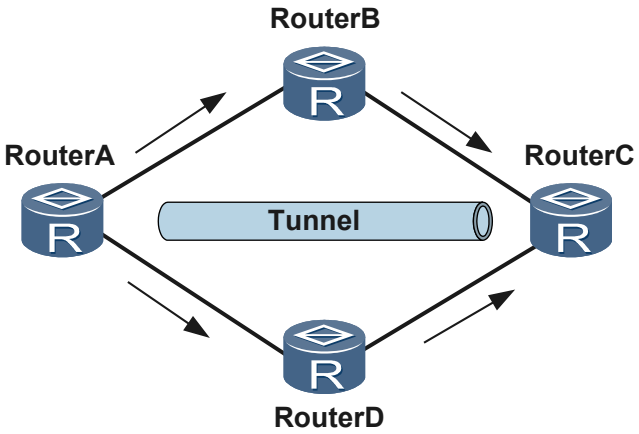
  IS-IS TE transmits MPLS TE configurations to the CSPF module.

- Process of handling TE information in LSPs.

  IS-IS TE extracts TE information from IS-IS LSPs and transmits the TE information to the CSPF module.

In typical applications, IS-IS TE helps MPLS TE set up TE tunnels. As shown in **Figure 6-23**, a TE tunnel is set up between RouterA and RouterD.

Figure 6-23 IS-IS TE networking



The networking configuration is as follows:

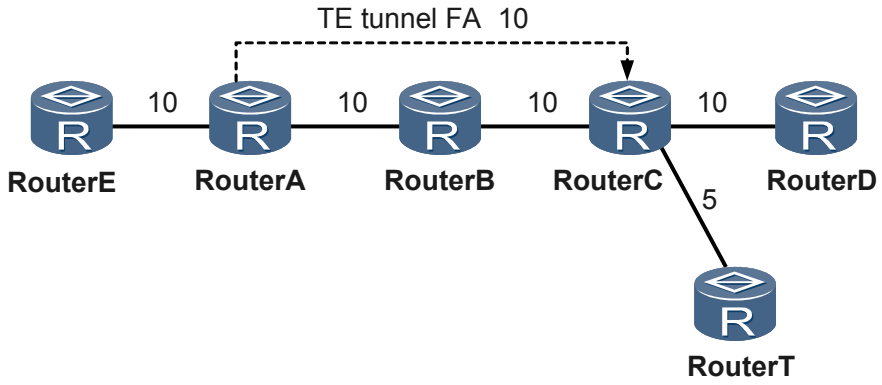
- Enable MPLS TE on RouterA, RouterB, RouterC, and RouterD and enable MPLS TE CSPF on RouterA to calculate the tunnel path.
- Run IS-IS and enable IS-IS TE on RouterA, RouterB, RouterC, and RouterD to implement communication between the four routers.

After the preceding configuration is complete, IS-IS on RouterA, RouterB, RouterC, and RouterD sends LSPs carrying TE information configured on each router. RouterA then obtains the TE information of RouterB, RouterC, and RouterD from the received LSPs. The CSPF module can calculate the path required by the TE tunnel based on the TE information on the entire network.

Route Calculation on TE Tunnel Interfaces

IS-IS Shortcut (AA) and IS-IS Advertise (FA) calculate routes through TE tunnel interfaces. For the traffic transmitted through a specific route, MPLS guarantees the forwarding comparing with IP, which is unreliable. When IS-IS Shortcut (AA) and IS-IS Advertise (FA) are configured, MPLS forwarding is achieved with TE tunnel interfaces involving in route calculation and being the outbound interfaces of specific routes.

Figure 6-24 Principle of IS-IS Shortcut (AA) and Advertise (FA)



IS-IS Shortcut (AA) and IS-IS Advertise (FA) have the following differences:

- IS-IS Advertise (FA) advertises TE tunnel information to other ISs, whereas IS-IS Shortcut (AA) does not.

  As shown in **Figure 6-24**, if the TE tunnel is enabled with IS-IS Advertise (FA), RouterA advertises information indicating that RouterC is its neighbor. The neighbor information is carried in TLV type 22 with no sub-TLVs. That is, no TE information is carried. If the TE tunnel is enabled with IS-IS Shortcut (AA), RouterA does not advertise such information.

- IS-IS Advertise (FA) affects the SPF tree of other routers, whereas IS-IS Shortcut (AA) does not.

  IS-IS Shortcut (AA) does not affect the original structure of the IS-IS SPF tree, irrespective of whether a TE tunnel exists or not. Apart from the link from RouterA to RouterB, and that from RouterB to RouterC, a link marked with an Shortcut from RouterA to RouterC is added. The link marked with an Shortcut participates in route calculation.

  If the TE tunnel is enabled with IS-IS Advertise (FA), RouterA advertises the message that "RouterC is a neighbor of RouterA" to other routers on the network. Other routers then consider RouterC a neighbor of RouterA and add RouterC to the SPF tree without marking it with an Shortcut.

- IS-IS Advertise (FA) does not support a relative metric, whereas IS-IS Shortcut (AA) supports.

  IS-IS Shortcut (AA) supports an absolute metric and a relative metric.

  If you use an absolute metric, the metric value of TE tunnels in IS-IS is fixed. If you use a relative metric, the metric value of TE tunnels in IS-IS is the sum of the physical link cost and relative metric. As shown in **Figure 6-24**, if the relative metric is set to 1, the cost of the path from SwitchA to SwitchC through the TE tunnel is 21 (10+10+1). If the relative metric is set to 0, the TE tunnel and physical link are of equal-cost on the outbound interface. If the relative metric is less than 0, the TE tunnel interface is preferred as the outbound interface.

- IS-IS Advertise (FA) requires bidirectional TE tunnels, whereas IS-IS Shortcut (AA) requires only unidirectional tunnels.

6.2.16 IS-IS Local MT

IS-IS local multicast-topology (MT) creates a separate multicast topology on the local device, without affecting the protocol packets exchanged between devices, to allow both TE tunnels and multicast to be configured on the backbone network.

NOTE

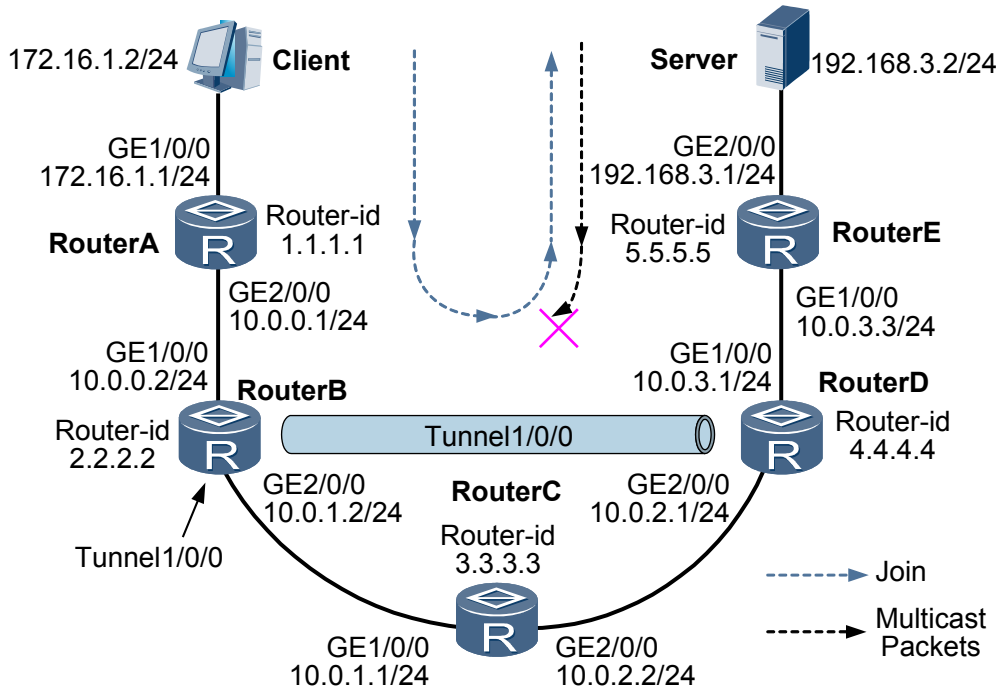
The mentioned TE tunnel specifies the TE tunnel enabled with IGP Shortcut (AA).

Background

When multicast and an MPLS TE tunnel are deployed in a network simultaneously, the multicast function may be affected by the TE tunnel.

This is because after the TE tunnel is enabled with IS-IS Shortcut (AA), the outbound interface of a route calculated by an IS-IS is not the actual physical interface but a TE tunnel interface. According to the unicast route to the multicast source address, a router sends a Report message through a TE tunnel interface. Routers spanned by the TE tunnel cannot sense the Report message, so multicast forwarding entries cannot be created. The TE tunnel is unidirectional, so multicast data packets sent by the multicast source are sent to the routers spanned by the tunnel through the related physical interfaces. The routers do not have any multicast forwarding entry. Therefore, the multicast data packets are discarded.

Figure 6-25 TE tunnel scenario



As shown in **Figure 6-25**, RouterA, RouterB, RouterC, RouterD, and RouterE are Level-2 routers. The routers run IS-IS to implement interconnection. The multicast services are normal. A unidirectional MPLS TE tunnel is set up between RouterB and RouterD. The MPLS TE tunnel is enabled with IS-IS Shortcut (AA). When you view the multicast routing table on RouterC spanned by the TE tunnel, you cannot find any multicast forwarding entry. Therefore, the multicast services are interrupted.

The process of transmitting multicast packets between the client and the multicast server is as follows:

1. To join a multicast group, the client sends a Report message to SwitchA. SwitchA then sends a Join message to SwitchB.
2. When the Join message reaches SwitchB, SwitchB uses Tunnel 1/0/0 as the Reverse Path Forwarding (RPF) interface and forwards the message to SwitchC through GE 2/0/0 by using the MPLS label.
3. The Join message is forwarded with the MPLS label, so SwitchC just forwards the message and does not create a multicast routing entry. In the topology shown in **Figure 6-25**, SwitchC is the penultimate hop of the MPLS forwarding. SwitchC pops out the MPLS label, and then forwards the Join message to SwitchD through GE 2/0/0.
4. After receiving the Join message, SwitchD creates a multicast forwarding entry. The inbound interface is GE 2/0/0 and the outbound interface is GE 1/0/0. SwitchD then forwards the message to SwitchE. The SPT is set up.
5. When the multicast source sends the traffic to SwitchD, SwitchD forwards the traffic to SwitchC. SwitchC does not create any forwarding entry in advance. Therefore, the traffic is discarded and the multicast service is interrupted.

As described in the preceding process of transmitting multicast packets, the forwarding of multicast packets relies on the unicast routing table and the TE tunnel is unidirectional. Therefore, the multicast packets are discarded. This problem can be avoided by using the following methods:

- Manually configuring static multicast routes to guide the forwarding of multicast packets.
- Configuring a bidirectional TE tunnel. In this case, the returned multicast packets can be sent by using the same tunnel. Routers spanned by the TE tunnel use the tunnel to transmit multicast packets.
- Configuring the Multicast Border Gateway Protocol (MBGP) to separate the unicast topology from the multicast topology. MBGP provides the topology that does not contain the TE tunnel for multicast separately. Multicast is used to perform RPF check on MBGP routes.
- Configuring local MT

The preceding methods are used to prevent the interruption of multicast services. The disadvantage of the first three methods is that a lot of manual configurations need to be done. As a result, if the network is complex, the planning, configuration, and maintenance tasks become heavier. Therefore, in the preceding network environment, local MT needs to be configured.

Principles

Local MT creates a separate multicast topology on the local device, without affecting the protocol packets exchanged between devices. Devices support local MT. This ensures that multicast services are still available when both multicast and the MPLS TE tunnel enabled with IGP Shortcut are deployed.

After local MT is enabled, the router at the ingress of a TE tunnel creates a separate multicast IGP (MIGP) routing table to store the physical interfaces to which the TE tunnel corresponds. This ensures that multicast protocol packets are correctly forwarded. The correct routing entries are created in the multicast routing table (MRT).

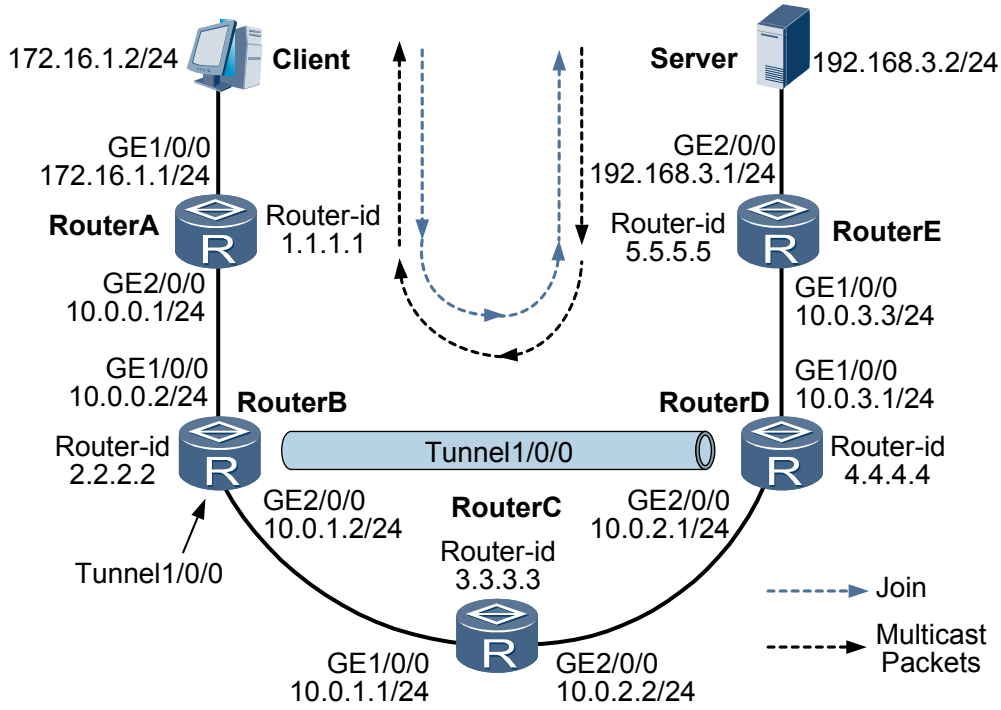
- Create an MIGP routing table.

Multicast protocol packets are forwarded according to the unicast routing table. After local MT is enabled on SwitchB, RM creates separate MIGP routing tables for multicast protocols. When the outbound interface of a route is a TE tunnel interface, an IGP calculates out the actually physical outbound interface for the route and adds the outbound interface to the MIGP routing table.

- Guide the forwarding of multicast protocol packets.

  Before forwarding a multicast protocol packet, a router needs to search the unicast routing table. If the router finds that the next hop is the TE tunnel, the router continues to search the MIGP routing table for the related physical outbound interface to guide the forwarding of the multicast protocol packet.

Figure 6-26 Local MT Topology



As shown in **Figure 6-26**, if the outbound interface of multicast source 192.168.3.2/24 is TE tunnel 1/0/0, the physical outbound interface of the route calculated by IS-IS is GE 2/0/0. IS-IS installs the route to the MIGP routing table. The multicast services are not affected by the TE tunnel. Multicast packets are forwarded through the physical outbound interfaces according to the MIGP routing table for the general IP forwarding. The related routing entries are created in the MRT. Multicast data packets are then correctly forwarded.

6.2.17 IS-IS Multi-Instance and Multi-Process

On a VPN-supporting device, you can associate multiple VPN instances with multiple IS-IS processes to implement IS-IS multi-instance. IS-IS multi-process allows you to create multiple IS-IS processes in the same VPN (or on the public network). These IS-IS processes are independent of each other. Route exchange between IS-IS processes is similar to route exchange between routing protocols.

Each IS-IS process can be bound to a specified VPN instance. A typical application is as follows: In a VPN, IS-IS runs between PEs and CEs and also runs on the VPN backbone network. On the PEs, the two IS-IS processes are independent of each other.

IS-IS multi-instance and multi-process have the following characteristics:

- IS-IS multi-processes share an RM routing table. IS-IS multi-instances use the RM routing tables in VPNs, and each VPN has its own RM routing table.
- IS-IS multi-process allows a set of interfaces to be associated with a specified IS-IS process. This ensures that the specified IS-IS process performs all the protocol operations only on this set of interfaces. In this manner, multiple IS-IS processes can work on a single router and each process is responsible for managing a unique set of interfaces.
- When creating an IS-IS process, you can bind it to a VPN instance to associate the IS-IS process with the VPN instance. The IS-IS process accepts and processes only the events related to the VPN instance. When the bound VPN instance is deleted, the IS-IS process is also deleted.

6.2.18 IS-IS IPv6

IS-IS is a link-state dynamic routing protocol initially designed by the OSI. To support IPv4 routing, IS-IS is applied to IPv4 networks and called as Integrated IS-IS.

As IPv6 networks are built, IS-IS also needs to provide accurate routing information for IPv6 packet forwarding. IS-IS has good scalability, supports IPv6 network layer protocols, and is capable of discovering, generating, and forwarding IPv6 routes.

Extended IS-IS for IPv6 is defined in the draft-ietf-isis-ipv6-05.txt of the IETF. To process and calculate IPv6 routes, IS-IS uses two new TLVs and one network layer protocol identifier (NLPID).

The two TLVs are as follows:

- TLV 236 (IPv6 Reachability): describes network reachability by defining the route prefix and metric.
- TLV 232 (IPv6 Interface Address): is similar to the IP Interface Address TLV of IPv4, except that it changes a 32-bit IPv4 address to a 128-bit IPv6 address.

The NLPID is an 8-bit field that identifies the protocol packets of the network layer. The NLPID of IPv6 is 142 (0x8E). If IS-IS supports IPv6, it advertises routing information through the NLPID value.

6.2.19 IS-IS MT

During the transition from IPv4 networks to IPv6 networks, IPv4 topologies and IPv6 topologies must coexist for a long time. The IPv4/IPv6 dual stack is a widely used technology that is applicable to IPv4 networks and IPv6 networks. The function is that a router that supports only IPv4 or IPv6 can communicate with a router that supports both IPv4 and IPv6.

Background

IS-IS implements IPv6 by extending TLV and complies with the rules for establishing and maintaining neighbor databases and topology databases as defined in ISO 10589 and RFC 1195. As a result, IPv4 networks and IPv6 networks have the same topology. The mixed topology of IPv4 and IPv6 is considered as an integrated topology, which utilizes the SPT to perform the SPF calculation. This requires that IPv6 and IPv4 topology information should be consistent.

In actual applications, the deployment of IPv4 and IPv6 may be different on the network; therefore, information about IPv4 topologies may be different from information about IPv6 topologies. Some routers and links in a mixed topology do not support IPv6. However, routers that support the IPv4/IPv6 dual stack in the mixed topology cannot sense the routers or links, and still forward IPv6 packets to them. As a result, the IPv6 packets are discarded. Similarly, when routers and links that do not support IPv4 exist in the topology, IPv4 packets cannot be forwarded.

IS-IS multi-topology (MT) can be used to solve the preceding problems. IS-IS MT is an extension of IS-IS to support multiple topologies, complying with draft-ietf-IS-IS-wg-multi-topology. IS-IS MT defines new TLVs in IS-IS packets, transmits MT information, and performs separate SPF calculation in different topologies.

Principles

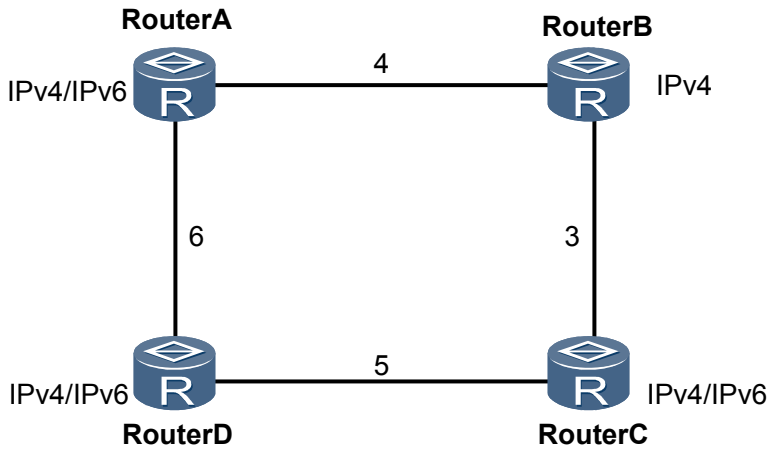
IS-IS MT refers to multiple separate IP topologies that are run in an IS-IS AS, such as IPv4 topology and IPv6 topology. The separate IP topologies are not considered as an integrated and single topology. This is helpful for calculating IS-IS routes of separate IPv4 networks and IPv6 networks. Based on the IP protocols supported by links, separate SPF calculation is performed in different topologies to shield networks from each other.

Figure 6-27 shows the IS-IS MT. Values in **Figure 6-27** indicate link costs. RouterA, RouterC, and RouterD support the IPv4/IPv6 dual stack. RouterB supports only IPv4 and cannot forward IPv6 packets.

If RouterA does not support IS-IS MT, only the single topology is considered during SPF calculation. The shortest path from RouterA to RouterC is RouterA->RouterB->RouterC. However, RouterB does not support IPv6. IPv6 packets sent from RouterA cannot be forwarded by RouterB to RouterC.

If IS-IS MT is enabled on RouterA, RouterA performs SPF calculation in different topologies. When RouterA needs to send IPv6 packets to RouterC, RouterA chooses only IPv6 links to forward IPv6 packets. The shortest path from RouterA to RouterC changes to RouterA->RouterD->RouterC. IPv6 packets are then forwarded.

Figure 6-27 IS-IS MT networking



IS-IS MT is implemented as follows:

1. Setting up topologies: Neighbors are set up by exchanging various packets for setting up MTs.
2. Performing the SPF calculation: The SPF calculation is performed for different MTs.

6.3 References

Table 6-7 The following table lists the references of this document.

Document	Description	Remarks
ISO 10589	ISO IS-IS Routing Protocol	-
ISO 8348/Ad2	Network Services Access Points	-
RFC 1195	Use of OSI IS-IS for Routing in TCP/IP and Dual Environments	Multiple authentication passwords are not supported.
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 3373	Three-Way Handshake for IS-IS Point-to-Point Adjacencies	-
RFC 3567	Intermediate System to Intermediate System (IS-IS) Cryptographic Authentication	-
RFC 3719	Recommendations for Interoperable Networks using IS-IS	-
RFC 3784	IS-IS extensions for Traffic Engineering	-
RFC 3786	Extending the Number of IS-IS LSP Fragments Beyond the 256 Limit	-
RFC 3787	Recommendations for Interoperable IP Networks using IS-IS	-
RFC 3847	Restart signaling for IS-IS	-
RFC 3906	Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels	-
RFC 4444	Management Information Base for IS-IS	-

Document	Description	Remarks
RFC 5120	Multi Topology (MT) Routing in IS-IS	-
draft-ietf-IS-IS-ipv6-05	Routing IPv6 with IS-IS	-
draft-ietf-IS-IS-wg-multi-topology-11	M-IS-IS: Multi Topology (MT) Routing in IS-IS	-
draft-ietf-isis-admin-tags-02(Admin Tag)	Admin Tag	-

7 BGP

About This Chapter

7.1 Introduction to BGP

Definition

The Border Gateway Protocol (BGP) is a path vector protocol that allows devices between Autonomous Systems (ASs) to communicate and selects optimal routes. BGP-1 (defined in RFC 1105), BGP-2 (defined in RFC 1163), and BGP-3 (defined in RFC 1267) are three earlier versions of BGP. BGP-4 (defined in RFC 1771) has been used since 1994. Since 2006, unicast IPv4 networks have been using BGP-4 defined in RFC 4271, and other networks have been using **MP-BGP** defined in RFC 4760.

Purpose

A network is divided into different ASs to facilitate the management over the network. In 1982, the Exterior Gateway Protocol (EGP) was used to dynamically exchange routing information between ASs. EGP advertises only reachable routes but not select optimal routes or prevent routing loops. Therefore, EGP cannot meet network management requirements.

BGP was designed to replace EGP. Different from EGP, BGP can select optimal routes, prevent routing loops, transmit routing information efficiently, and maintain a large number of routes.

Although BGP is used to transmit routing information between ASs, BGP is not the best choice in some scenarios. For example, on the egress connecting a data center to the Internet, static routes instead of BGP are used to prevent a huge number of Internet routes from affecting the data center internal network.

Benefits

BGP ensures high network security, flexibility, stability, reliability, and efficiency:

- BGP uses authentication and Generalized TTL Security Mechanism (GTSM) to ensure **network security**.
- BGP provides routing policies to allow for flexible **route selection** and **routing policy-based route advertisement**.
- BGP provides **7.2.8 Route Summarization** and **7.2.9 Route Dampening** to prevent route flapping and improve network stability.
- BGP uses the Transport Control Protocol (TCP) with listening port number 179 as the transport layer protocol and supports **7.2.10 Association Between BGP and BFD**, **7.2.11 BGP Tracking**, **7.2.12 BGP Auto FRR**, and **7.2.13 BGP GR and NSR** to improve network reliability.
- BGP uses the **7.2.15 Dynamic Update Peer-Groups** technology to send packets in groups when a large number of peers and routes exist and most peers share the same outbound policies, improving BGP forwarding performance.

7.2 Principles

7.2.1 BGP Concepts

This section describes BGP concepts to help you better understand BGP functions.

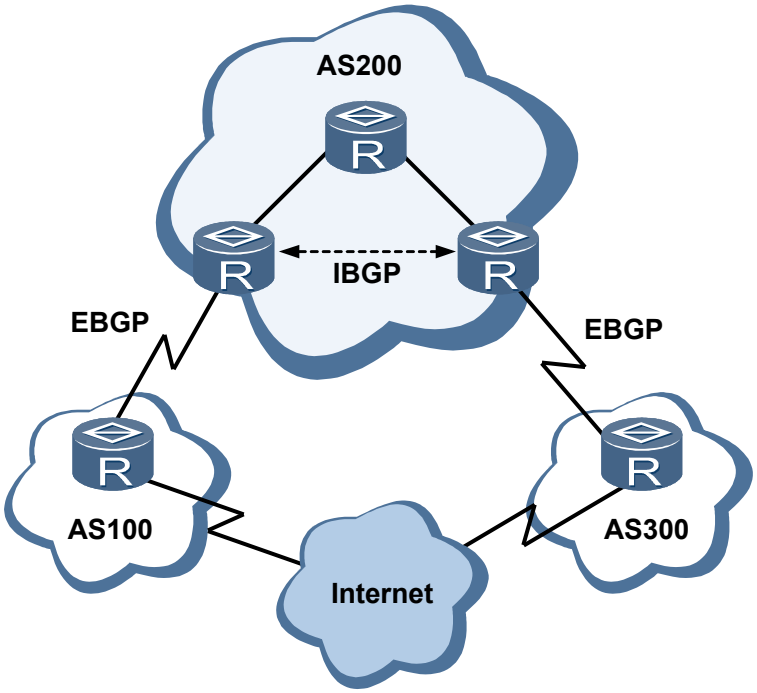
Autonomous System

An Autonomous System (AS) is a group of Internet Protocol (IP) networks that are controlled by one entity, typically an Internet service provider (ISP), and that have the same routing policy. Each AS is assigned a unique AS number, which identifies an AS on a BGP network. Two types of AS numbers are available: 2-byte AS numbers and 4-byte AS numbers. A 2-byte AS number ranges from 1 to 65535, and a 4-byte AS number ranges from 1 to 4294967295. Devices supporting 4-byte AS numbers are compatible with devices supporting 2-byte AS numbers.

BGP Classification

As shown in **Figure 7-1**, BGP is classified into two types according to where it runs: Internal BGP (IBGP) and External BGP (EBGP). When BGP runs between two peers in the same AS, BGP is called IBGP. When BGP runs between ASs, BGP is called EBGP.

Figure 7-1 BGP operating mode



- EBGP: runs between ASs. To prevent routing loops between ASs, a BGP device discards the routes with the local AS number when receiving the routes from EBGP peers.
- IBGP: runs within an AS. To prevent routing loops within an AS, a BGP device does not advertise the routes learned from an IBGP peer to the other IBGP peers and establishes full-mesh connections with all the IBGP peers. To address the problem of too many IBGP connections between IBGP peers, BGP uses **7.2.6 Route Reflector** and **7.2.7 BGP Confederation**.

  NOTE

  If a BGP device needs to advertise the route received from an EBGP peer outside an AS through another BGP device, IBGP is recommended.

Device Roles in BGP Message Exchange

There are two device roles in BGP message exchange:

- Speaker: The device that sends BGP messages is called a BGP speaker. The speaker receives and generates new routes, and advertises the routes to other BGP speakers.
- Peer: The speakers that exchange messages with each other are called BGP peers. A group of peers sharing the same policies can form a peer group.

BGP Router ID

The BGP router ID is a 32-bit value that is often represented by an IPv4 address to identify a BGP device. It is carried in the Open message sent during the establishment of a BGP session. When two BGP peers need to establish a BGP session, they each require a unique router ID. Otherwise, the two peers cannot establish a BGP session.

The BGP router ID of a device must be unique on a BGP network. It can be manually configured or selected from IPv4 addresses on the device. By default, an IPv4 address of a loopback interface on a device is used as the BGP router ID. If no loopback interface is configured on the device, the system selects the largest IPv4 address from all IPv4 addresses of interfaces as the BGP router ID. Once the BGP router ID is selected, the system retains this router ID even if a larger IPv4 address is configured on the device later. The system changes the BGP router ID only when the corresponding IPv4 address is deleted.

7.2.2 BGP Working Principles

BGP peer establishment, update, and deletion involve five types of messages, six state machine states, and five route exchange rules.

BGP Messages

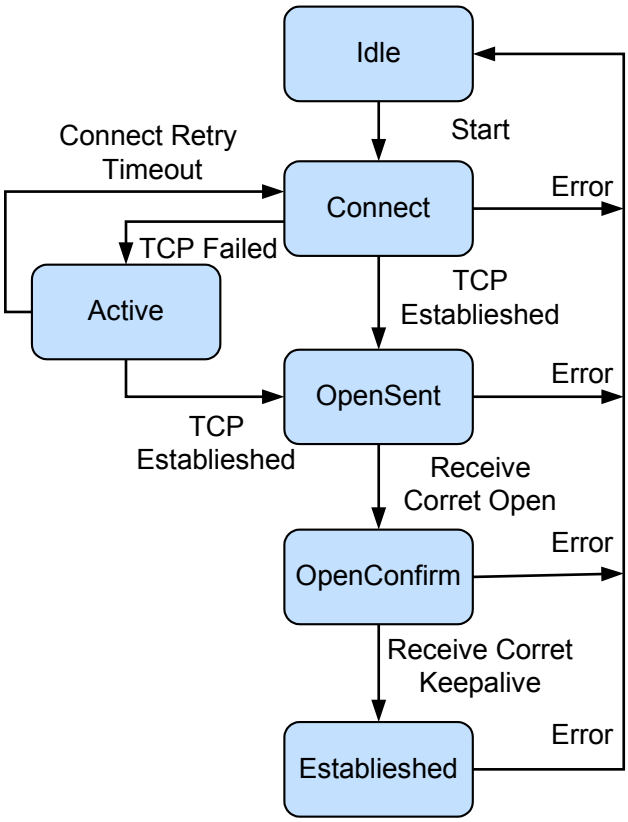
BGP peers exchange the following messages, among which Keepalive messages are periodically sent and other messages are triggered by events.

- Open message: is used to establish BGP peer relationships.
- Update message: is used to exchange routes between BGP peers.
- Notification message: is used to terminate BGP connections.
- Keepalive message: is used to maintain BGP connections.
- Route-refresh message: is used to request the peer to resend routes if routing policies are changed. Only the BGP devices supporting route-refresh can send and respond to Route-refresh messages.

BGP State Machine

As shown in **Figure 7-2**, a BGP device uses a finite state machine (FSM) to determine its operations with peers. The FSM has six states: Idle, Connect, Active, OpenSent, OpenConfirm, and Established. Three common states are involved in BGP peer establishment: Idle, Active, and Established.

Figure 7-2 BGP state machine



1. The Idle state is the initial BGP state. In Idle state, the BGP device refuses all connection requests from neighbors. The BGP device initiates a TCP connection with its BGP peer and changes its state to Connect only after receiving a Start event from the system.

   NOTE

   - The Start event occurs when an operator configures a BGP process or resets an existing BGP process or when the router software resets a BGP process.
   - If an error occurs at any state of the FSM, for example, the BGP device receives a Notification packet or TCP connection termination notification, the BGP device returns to the Idle state.

2. In Connect state, the BGP device starts the ConnectRetry timer and waits to establish a TCP connection.
   - If the TCP connection is established, the BGP device sends an Open message to the peer and changes to the OpenSent state.
   - If the TCP connection fails to be established, the BGP device moves to the Active state.
   - If the BGP device does not receive a response from the peer before the ConnectRetry timer expires, the BGP device attempts to establish a TCP connection with another peer and stays in Connect state.
3. In Active state, the BGP device keeps trying to establish a TCP connection with the peer.
   - If the TCP connection is established, the BGP device sends an Open message to the peer, closes the ConnectRetry timer, and changes to the OpenSent state.
   - If the TCP connection fails to be established, the BGP device stays in the Active state.

- If the BGP device does not receive a response from the peer before the ConnectRetry timer expires, the BGP device returns to the Connect state.

4. In OpenSent state, the BGP device waits an Open message from the peer and then checks the validity of the received Open message, including the AS number, version, and authentication password.
   - If the received Open message is valid, the BGP device sends a Keepalive message and changes to the OpenConfirm state.
   - If the received Open message is invalid, the BGP device sends a Notification message to the peer and returns to the Idle state.
5. In OpenConfirm state, the BGP device waits for a Keepalive or Notification message from the peer. If the BGP device receives a Keepalive message, it transitions to the Established state. If it receives a Notification message, it returns to the Idle state.
6. In Established state, the BGP device exchanges Update, Keepalive, Route-refresh, and Notification messages with the peer.
   - If the BGP device receives a valid Update or Keepalive message, it considers that the peer is working properly and maintains the BGP connection with the peer.
   - If the BGP device receives a valid Update or Keepalive message, it sends a Notification message to the peer and returns to the Idle state.
   - If the BGP device receives a Route-refresh message, it does not change its status.
   - If the BGP device receives a Notification message, it returns to the Idle state.
   - If the BGP device receives a TCP connection termination notification, it terminates the TCP connection with the peer and returns to the Idle state.

Route Exchange Rules

A BGP device adds optimal routes to the BGP routing table to generate BGP routes. After establishing a BGP peer relationship with a neighbor, the BGP device follows the following rules to exchange routes with the peer:

- Advertises the BGP routes received from IBGP peers only to its EBGP peers.
- Advertises the BGP routes received from EBGP peers to its EBGP peers and IBGP peers.
- Advertises the optimal route to its peers when there are multiple valid routes to the same destination.
- Sends only updated BGP routes when BGP routes change.
- Accepts all the routes sent from its peers.

7.2.3 Interaction Between BGP and an IGP

BGP and IGPs use different routing tables. To enable different ASs to communicate, you need to configure interaction between BGP and IGPs so that BGP routes can be imported into IGP routing tables and IGP routes can also be imported to BGP routing tables.

Importing IGP Routes to BGP Routing Tables

BGP does not discover routes and so needs to import the routes discovered by IGPs to BGP routing tables so that different ASs can communicate. When an AS needs to advertise routes to another AS, an Autonomous System Boundary Router (ASBR) imports IGP routes to its BGP routing table. To better plan the network, you can use routing policies to filter routes and set

route attributes when BGP imports IGP routes. Alternatively, you can set the multi-exit discriminator (MED) to help EBGP peers select the best path for traffic entering an AS.

BGP imports routes in either import or network mode:

- In import mode, BGP imports IGP routes, including RIP, OSPF, and IS-IS routes, into BGP routing tables based on protocol type. To ensure the validity of imported IGP routes, BGP can also import static routes and direct routes in import mode.
- In network mode, BGP imports the routes in the IP routing table one by one to BGP routing tables. The network mode is more accurate than the import mode.

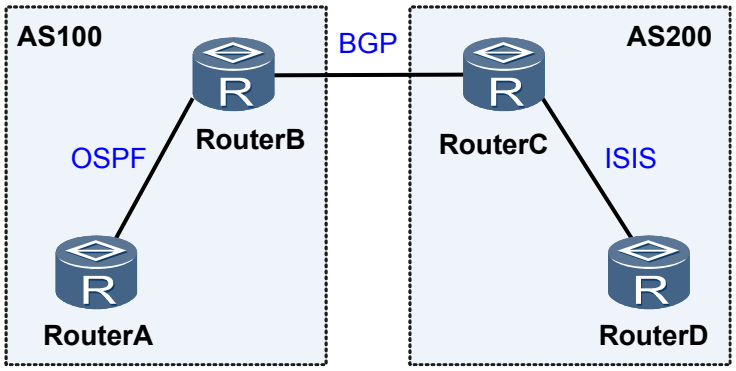
Importing BGP Routes to IGP Routing Tables

When an AS needs to import routes from another AS, an ASBR imports BGP routes to its IGP routing table. To prevent a large number of BGP routes from affecting devices within the AS, IGPs can use routing policies to filter routes and set route attributes when importing BGP routes.

Applications

As shown in **Figure 7-3**, an OSPF network is deployed in AS 100 where the Overseas Market Department of a company resides, and an IS-IS network is deployed in AS 200 where the Domestic R&D Department of the company resides. AS 100 and AS 200 communicate using BGP. The company requires that the Overseas Market Department can send files to the Domestic R&D Department but the Domestic R&D Department cannot send files to the Overseas Market Department.

Figure 7-3 IGPs importing BGP routes



According to the preceding requirement of the company, devices in AS 100 must know routes of AS 200, but devices in AS 200 do not know routes of AS 100. To meet this requirement, configure BGP to import IS-IS routes on RouterC. Then RouterC has routes of AS 200 in the BGP routing table and advertises these routes to RouterB. In addition, configure OSPF to import BGP routes on RouterB. Devices in AS 100 can know routes of AS 200, but devices in AS 200 do not know routes of AS 100.

7.2.4 BGP Security

BGP uses authentication and Generalized TTL Security Mechanism (GTSM) to ensure exchange security between BGP peers.

BGP Authentication

BGP authentication includes Message Digest 5 (MD5) authentication and keychain authentication, which improves communication security between BGP peers. In MD5 authentication, you can only set the authentication password for a TCP connection. In keychain authentication, you can set the authentication password for a TCP connection and authenticate BGP messages.

BGP GTSM

BGP GTSM checks whether the time to live (TTL) value in the IP packet header is within a predefined range and permits or discards the packets of which the TTL values are out of the predefined range to protect services above the IP layer. BGP GTSM enhances system security.

Assume that the TTL value range of packets from BGP peers is set to 254-255. When an attacker forges valid BGP packets and keeps sending these packets to attack a device, the TTL values of these packets are smaller than 254. If BGP GTSM is not enabled on the device, the device finds that these packets are destined for itself and sends the packets to the control plane for processing. Then the control layer needs to process a large number of such attack packets, causing high CPU usage. If BGP GTSM is enabled on the device, the system checks the TTL values in all BGP packets and discards the attack packets of which the TTL values are smaller than 254. This prevents network attack packets from consuming CPU resources.

7.2.5 BGP Route Selection Rules and Load Balancing

There may be multiple routes to the same destination in a BGP routing table. BGP will select one route as the optimal route and advertise it to peers. To select the optimal route among these routes, BGP compares the BGP attributes of the routes in sequence based on route selection rules.

BGP Attributes

Route attributes describe routes. BGP route attributes are classified into the following types. **Table 7-1** lists common BGP attributes.

- Well-known mandatory attribute

  All BGP devices can identify this type of attributes, which must be carried in Update messages. Without this type of attributes, errors occur in routing information.

- Well-known discretionary attribute

  All BGP devices can identify this type of attributes, which are optional in Update messages. Without this type of attributes, errors do not occur in routing information.

- Optional transitive attribute

  BGP devices may not identify this type of attributes but still accepts them and advertises them to peers.

- Optional non-transitive attribute

  BGP devices may not identify this type of attributes. If a BGP device does not identify this type of attributes, it ignores them and does not advertise them to peers.

Table 7-1 Common BGP attributes

Attribute	Type
Origin	Well-known mandatory
AS_Path	Well-known mandatory
Next_Hop	Well-known mandatory
Local_Pref	Well-known discretionary
Community	Optional transitive
MED	Optional non-transitive
Originator_ID	Optional non-transitive
Cluster_List	Optional non-transitive

The following describes common BGP route attributes:

- **Origin**

  The Origin attribute defines the origin of a route and marks the path of a BGP route. The Origin attribute is classified into three types:

  - IGP

    A route with IGP as the Origin attribute is of the highest priority. The Origin attribute of the routes imported into a BGP routing table using the **network** command is IGP.

  - EGP

    A route with EGP as the Origin attribute is of the secondary highest priority. The Origin attribute of the routes obtained through EGP is EGP.

  - Incomplete

    A route with Incomplete as the Origin attribute is of the lowest priority. The Origin attribute of the routes learned by other means is Incomplete. For example, the Origin attribute of the routes imported by BGP using the **import-route** command is Incomplete.

- **AS_Path**

  The AS_Path attribute records all the ASs that a route passes through from the source to the destination in the vector order. To prevent inter-AS routing loops, a BGP device does not receive the routes of which the AS_Path list contains the local AS number.

  When a BGP speaker advertises an imported route:

  - If the route is advertised to EBGP peers, the BGP speaker creates an AS_Path list containing the local AS number in an Update message.
  - If the route is advertised to IBGP peers, the BGP speaker creates an empty AS_Path list in an Update message.

  When a BGP speaker advertises a route learned in the Update message sent by another BGP speaker:

  - If the route is advertised to EBGP peers, the BGP speaker adds the local AS number to the leftmost of the AS_Path list. According to the AS_Path list, the BGP speaker that receives the route can learn about the ASs through which the route passes to reach the

destination. The number of the AS that is nearest to the local AS is placed on the top of the AS_Path list. The other AS numbers are listed according to the sequence in which the route passes through ASs.

- If the route is advertised to IBGP peers, the BGP speaker does not change the AS_Path attribute of the route.

- **Next_Hop**

  The Next_Hop attribute records the next hop that a route passes through. The Next_Hop attribute of BGP is different from that of an IGP because it may not be the neighbor IP address. A BGP speaker processes the Next_Hop attribute based on the following rules:

  - When advertising a route to an EBGP peer, a BGP speaker sets the Next_Hop attribute of the route to the address of the local interface through which the BGP peer relationship is established with the peer.
  - When advertising a locally originated route to an IBGP peer, the BGP speaker sets the Next_Hop attribute of the route to the address of the local interface through which the BGP peer relationship is established with the peer.
  - When advertising a route learned from an EBGP peer to an IBGP peer, the BGP speaker does not change the Next_Hop attribute of the route.

- **Local_Pref**

  The Local_Pref attribute indicates the BGP preference of a device and helps determine the optimal route when traffic leaves an AS. When a BGP device obtains multiple routes to the same destination address but with different next hops from different IBGP peers, the BGP device prefers the route with the highest Local_Pref. The Local_Pref attribute is exchanged only between IBGP peers and is not advertised to other ASs. The Local_Pref attribute can be manually configured. If no Local_Pref attribute is configured for a route, the Local_Pref attribute of the route uses the default value 100.

- **MED**

  The multi-exit discriminator (MED) attribute helps determine the optimal route when traffic enters an AS. When a BGP device obtains multiple routes to the same destination address but with different next hops from EBGP peers, the BGP device selects the route with the smallest MED value as the optimal route.

  The MED attribute is exchanged only between two neighboring ASs. The AS that receives the MED attribute does not advertise it to any other ASs. The MED attribute can be manually configured. If no MED attribute is configured for a route, the MED attribute of the route uses the default value 0.

- **Community**

  The Community attribute identifies the BGP routes with the same characteristics, simplifies the applications of routing policies, and facilitates route maintenance and management.

  The Community attribute includes self-defined community attributes and well-known community attributes. **Table 7-2** lists well-known community attributes.

Table 7-2 Well-known community attributes

Community Attribute	Value	Description
Internet	0 (0x00000000)	A BGP device can advertise the received route with the Internet attribute to all peers.

Community Attribute	Value	Description
No_Advertise	4294967042 (0xFFFFFF02)	A BGP device does not advertise the received route with the No_Advertise attribute to any peer.
No_Export	4294967041 (0xFFFFFF01)	A BGP device does not advertise the received route with the No_Export attribute to devices outside the local AS.
No_Export_Sub-confed	4294967043 (0xFFFFFF03)	A BGP device does not advertise the received route with the No_Export_Subconfed attribute to devices outside the local AS or to devices outside the local sub-AS.

- **Originator_ID and Cluster_List**

  The Originator_ID attribute and Cluster_List attribute help eliminate loops in route reflector scenarios. For details, see **7.2.6 Route Reflector**.

BGP Route Selection Policies

When there are multiple routes to the same destination, BGP compares the following attributes in sequence to select the optimal route:

1. Prefers the route with the largest PrefVal value.

   The PrefVal attribute is a Huawei proprietary attribute and is valid only on the device where it is configured.
2. Prefers the route with the highest Local_Pref.

   If a route does not have the Local_Pref attribute, the Local_Pref attribute of the route uses the default value 100.
3. Prefers the manually summarized route, automatically summarized route, route imported using the **network** command, route imported using the **import-route** command, and route learned from peers. These routes are in descending order of priority.
4. Prefers the route with the shortest AS_Path.
5. Prefers the route with the lowest origin type. IGP is lower than EGP, and EGP is lower than Incomplete.
6. Prefers the route with the lowest MED if routes are received from the same AS.
7. Prefers EBGP routes, IBGP routes, LocalCross routes, and RemoteCross routes, which are listed in descending order of priority.

   LocalCross allows a PE to add the VPNv4 route of a VPN instance to the routing table of the VPN instance if the export RT of the VPNv4 route matches the import RT of another VPN instance on the PE. RemoteCross allows a local PE to add the VPNv4 route learned from a remote PE to the routing table of a VPN instance on this local PE if the export RT of the VPNv4 route matches the import RT of the VPN instance.
8. Prefers the route with the lowest IGP metric to the BGP next hop.

   NOTE

   If there are multiple routes to the same destination, an IGP calculates the route metric using its routing algorithm.

9. Prefers the route with the shortest Cluster_List.
10. Prefers the route advertised by the device with the smallest router ID.

    NOTE

    If a route carries the Originator_ID attribute, BGP prefers the route with the smallest Originator_ID without comparing the router ID.
11. Prefers the route learned from the peer with the lowest IP address.

BGP Load Balancing

When there are multiple equal-cost routes to the same destination, you can perform load balancing among these routes to load balance traffic. Equal-cost BGP routes can be generated for traffic load balancing only when the first eight route attributes described in "BGP Route Selection Policies" are the same.

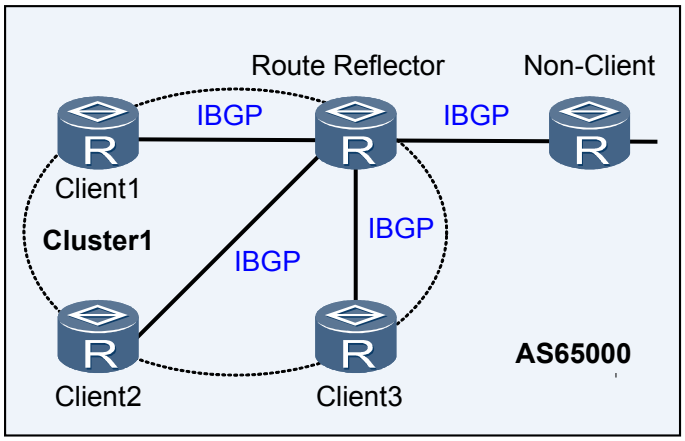
7.2.6 Route Reflector

To ensure connectivity between IBGP peers, you need to establish full-mesh connections between IBGP peers. If there are n devices in an AS, n(n-1)/2 IBGP connections need to be established. When there are a large number of devices, many network resources and CPU resources are consumed. A route reflector (RR) can be used between IBGP peers to solve this problem.

Roles in RR

As shown in **Figure 7-4**, the following roles are involved in RR scenarios in an AS.

Figure 7-4 Networking diagram of the RR



- Route reflector (RR): a BGP device that can reflect the routes learned from an IBGP peer to other IBGP peers. An RR is similar to a designated router (DR) on an OSPF network.
- Client: an IBGP device of which routes are reflected by the RR to other IBGP devices. In an AS, clients only need to directly connect to the RR.
- Non-client: an IBGP device that is neither an RR nor a client. In an AS, a non-client must establish full-mesh connections with the RR and all the other non-clients.

- Originator: is a device that originates routes in an AS. The Originator_ID attribute helps eliminate routing loops in a cluster.
- Cluster: is a set of the RR and clients. The Cluster_List attribute helps eliminate routing loops between clusters.

RR Principles

Clients in a cluster only need to exchange routing information with the RR in the same cluster. Therefore, clients only need to establish IBGP connections with the RR. This reduces the number of IBGP connections in the cluster. As shown in **Figure 7-4**, in AS 65000, Cluster1 is comprised of an RR and three clients. The number of IBGP connections in AS 65000 is then reduced from 10 to 4, which simplifies the device configuration and reduces the loads on the network and CPU.

The RR allows a BGP device to advertise the BGP routes learned from an IBGP peer to other IBGP peers, and uses the Cluster_List and Originator_ID attributes to eliminate routing loops. The RR advertises routes to IBGP peers based on the following rules:

- The RR advertises the routes learned from a non-client to all the clients.
- The RR advertises the routes learned from a client to all the other clients and all the non-clients.
- The RR advertises the routes learned from an EBGP peer to all the clients and non-clients.

Cluster_List Attribute

An RR and its clients form a cluster, which is identified by a unique cluster ID in an AS. To prevent routing loops between clusters, an RR uses the Cluster_List attribute to record the cluster IDs of all the clusters that a route passes through.

- When a route is reflected by an RR for the first time, the RR adds the local cluster ID to the top of the cluster list. If there is no cluster list, the RR creates a Cluster_List attribute.
- When receiving an updated route, the RR checks the cluster list of the route. If the cluster list contains the local cluster ID, the RR discards the route. If the cluster list does not contain the local cluster ID, the RR adds the local cluster ID to the cluster list and then reflects the route.

Originator_ID Attribute

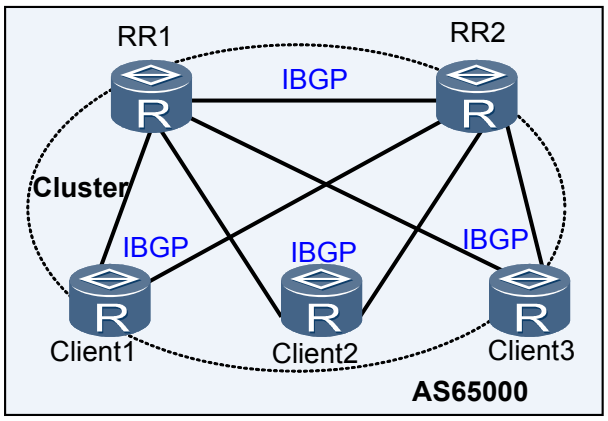
The originator ID identifies the originator of a route and is generated by an RR to prevent routing loops in a cluster. Its value is the same as the router ID.

- When a route is reflected by an RR for the first time, the RR adds the Originator_ID attribute to this route. The Originator_ID attribute identifies the originator of the route. If the route contains the Originator_ID attribute, the RR retains this Originator_ID attribute.
- When a device receives a route, the device compares the originator ID of the route with the local router ID. If they are the same, the device discards the route.

Backup RR

To ensure network reliability and prevent single points of failures, redundant RRs are required in a cluster. An RR allows a BGP device to advertise the routes received from an IBGP peer to other IBGP peers. Therefore, routing loops may occur between RRs in the same cluster. To solve this problem, all the RRs in the cluster must use the same cluster ID.

Figure 7-5 Backup RR



As shown in **Figure 7-5**, RR1 and RR2 reside in the same cluster and have the same cluster ID configured.

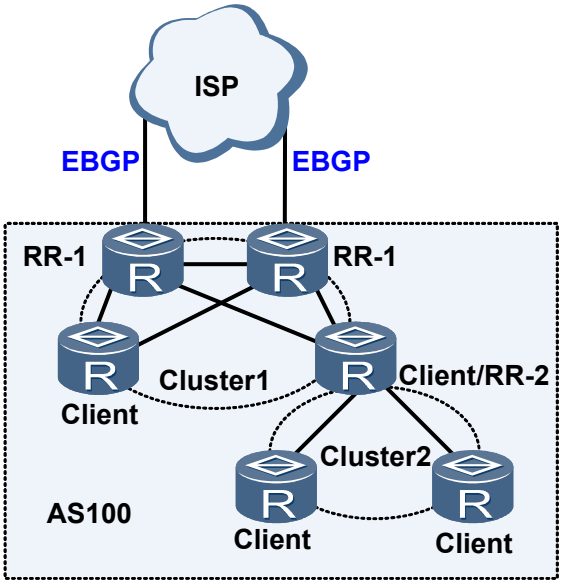
- When Client1 receives an updated route from an EBGP peer, Client1 advertises this route to RR1 and RR2 using IBGP.
- After RR1 and RR2 receive this route, they add the local cluster ID to the top of the cluster list of the route and then reflect the route to other clients (Client2 and Client3) and to each other.
- After RR1 and RR2 receive the reflected route from each other, they check the cluster list of the route, finding that the cluster list contains their local cluster IDs. RR1 and RR2 discard this route to prevent routing loops.

RRs of Multiple Clusters in an AS

There may be multiple clusters in an AS. RRs of the clusters establish IBGP peer relationships. When RRs reside at different network layers, an RR at the lower network layer can be configured as a client to implement hierarchical RR. When RRs reside at the same network layer, RRs of different clusters can establish full-mesh connections to implement flat RR.

Hierarchical RR

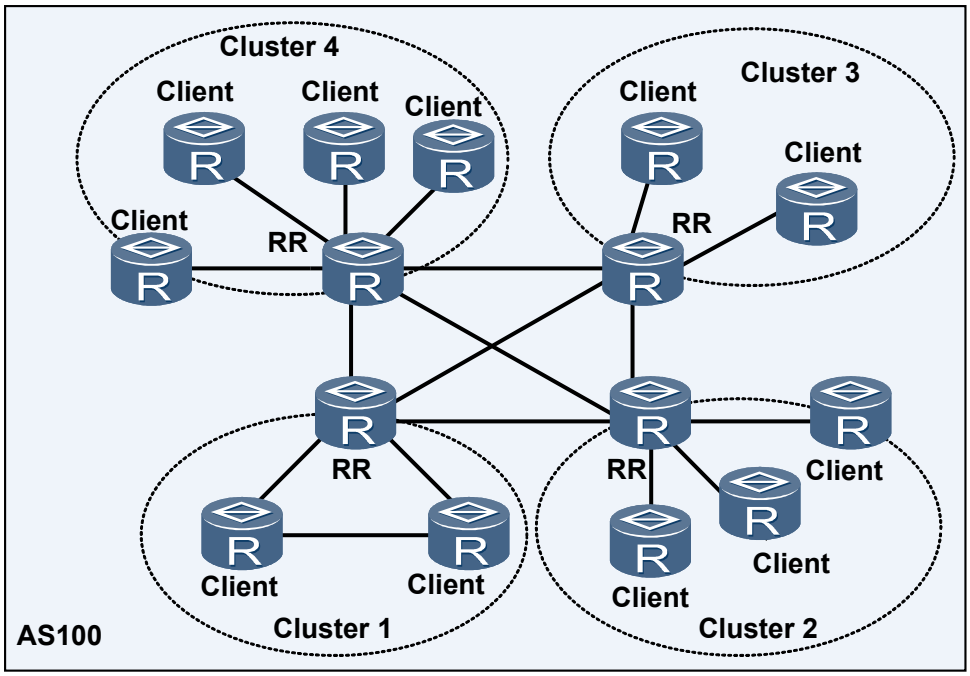
Figure 7-6 Hierarchical RR



In practice, hierarchical RR is often used. As shown in **Figure 7-6**, the ISP provides Internet routes to AS 100. AS 100 is divided into two clusters, Cluster1 and Cluster2. Four devices in Cluster1 are core routers and use a backup RR to ensure reliability.

Flat RR

Figure 7-7 Flat RR



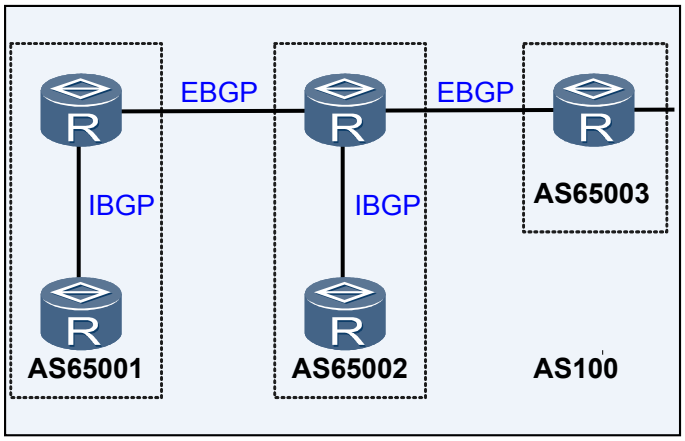
As shown in **Figure 7-7**, the backbone network is divided into multiple clusters. RRs of the clusters are non-clients and establish full-mesh connections with each other. Although each

client only establishes an IBGP connection with its RR, all the RRs and clients can receive all routing information.

7.2.7 BGP Confederation

In addition to a route reflector, the confederation is another method that reduces the number of IBGP connections in an AS. A confederation divides an AS into sub-ASs. Full-mesh IBGP connections are established in each sub-AS. EBGP connections are established between sub-ASs. ASs outside a confederation still consider the confederation as an AS. After a confederation divides an AS into sub-ASs, it assigns a confederation ID (the AS number) to each router within the AS. This brings two benefits. First, original IBGP attributes are retained, including the Local_Pref attribute, MED attribute, and Next_Hop attribute. Secondly, confederation-related attributes are automatically deleted when being advertised outside a confederation. Therefore, the administrator does not need to configure the rules for filtering information such as sub-AS numbers at the egress of a confederation.

Figure 7-8 Networking diagram of a confederation



As shown in **Figure 7-8**, AS 100 is divided into three sub-ASs after a confederation is configured: AS65001, AS65002, and AS65003. The AS number AS 100 is used as the confederation ID. The number of IBGP connections in AS 100 is then reduced from 10 to 4, which simplifies the device configuration and reduces the loads on the network and CPU. In addition, BGP devices outside AS 100 only know the existence of AS 100 but not the confederation within AS 100. Therefore, the confederation does not increase the CPU load.

Comparisons Between a Route Reflector and a Confederation

Table 7-3 compares a route reflector and a confederation in terms of the configuration, device connection, and applications.

Table 7-3 Comparisons between a route reflector and a confederation

Route Reflector	Confederation
Retains the existing network topology and ensures compatibility.	Requires the logical topology to be changed.

Route Reflector	Confederation
Requires only a route reflector to be configured because clients do not need to know that they are clients of a route reflector.	Requires all devices to be reconfigured.
Requires full-mesh connections between clusters.	Does not require full-mesh connections between sub-ASs of a confederation because the sub-ASs are special EBGP peers.
Applies to medium and large networks.	Applies to large networks.

7.2.8 Route Summarization

The BGP routing table of each device on a large network is large. This burdens devices, increases the route flapping probability, and affects network stability.

Route summarization is a mechanism that combines multiple routes into one route. This mechanism allows a BGP device to advertise only the summarized route but not all the specific routes to peers, therefore reducing the size of the BGP routing table. If the summarized route flaps, the network is not affected, so network stability is improved.

BGP supports automatic summarization and manual summarization on IPv4 networks, and supports only manual summarization on IPv6 networks.

- Automatic summarization: summarizes the routes imported by BGP. After automatic summarization is configured, BGP summarizes routes based on the natural network segment and advertises only the summarized route to peers. For example, BGP summarizes 10.1.1.1/24 and 10.2.1.1/24 (two Class A addresses with non-natural mask) into 10.0.0.0/8 (Class A address with natural mask).
- Manual summarization: summarizes routes in the local BGP routing table. Manual summarization can help control the attributes of the summarized route and determine whether to advertise specific routes.

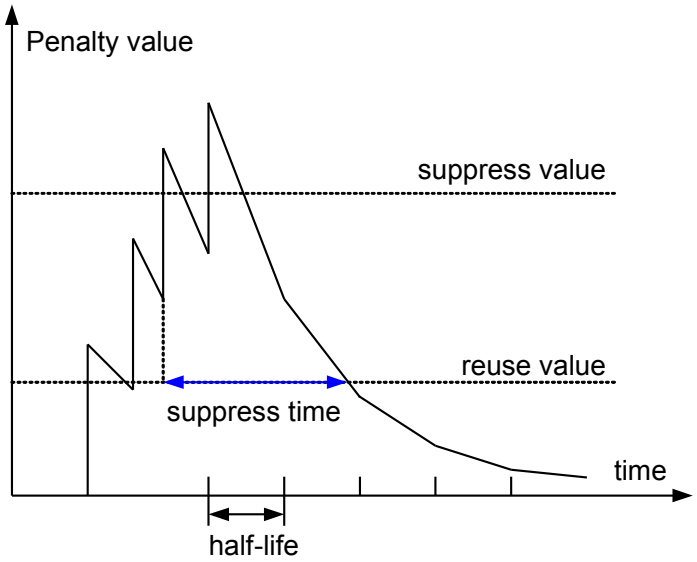
To prevent routing loops caused by route summarization, BGP uses the AS_Set attribute. The AS_Set attribute is an unordered set of all ASs that a route passes through. When the summarized route enters an AS in the AS_Set attribute again, BGP finds that the local AS number has been recorded in the AS_Set attribute of the route and discards this route to prevent a routing loop.

7.2.9 Route Dampening

When BGP is used on complex networks, route flapping occurs frequently. To prevent frequent route flapping, BGP uses route dampening to suppress unstable routes.

Route flapping is a process of adding a route to an IP routing table and then withdrawing this route. When route flapping occurs, a BGP device sends an Update message to its neighbors. The devices that receive the Update message need to recalculate routes and modify routing tables. Frequent route flapping consumes lots of bandwidths and CPU resources and even affects normal network operation.

Figure 7-9 Diagram of BGP route dampening



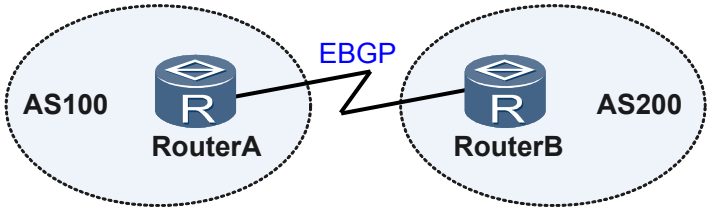
Route dampening measures the stability of a route using a penalty value. A larger penalty value indicates a less stable route. As shown in **Figure 7-9**, each time route flapping occurs, BGP increases the penalty of this route by a value of 1000. When the penalty value of a route exceeds the suppression threshold, BGP suppresses this route, and does not add it to the IP routing table or advertise any Update message to peers. After a route is suppressed for a period of time (half life), the penalty value is reduced by half. When the penalty value of a route decreases to the suppression threshold, the route is reusable and is added to the routing table. At the same time, BGP advertises an Update message to peers. The suppression time is the period from when a route is suppressed to when the route is reusable.

Route dampening applies only to EBGP routes but not IBGP routes. IBGP routes may include the routes of the local AS, and an IGP network requires that the routing tables of devices within an AS be the same. If IBGP routes were dampened, routing tables on devices are inconsistent when these devices have different dampening parameters. Therefore, route dampening does not apply to IBGP routes.

7.2.10 Association Between BGP and BFD

BGP periodically sends messages to peers to detect the status of the peers. It takes more than 1 minute for this detection mechanism to detect a fault. When data is transmitted at gigabit rates, long-time fault detection will cause packet loss. This cannot meet high reliability requirements of networks. Association between BGP and bidirectional forwarding detection (BFD) uses the millisecond-level fault detection of BFD to improve network reliability.

Figure 7-10 Networking diagram of association between BGP and BFD



As shown in **Figure 7-10**, RouterA belongs to AS 100 and RouterB belongs to AS 200. RouterA and RouterB are directly connected and establish the EBGP peer relationship. Association between BGP and BFD is configured on RouterA and RouterB. When a fault occurs on the link between RouterA and RouterB, BFD can rapidly detect that the BFD session changes from Up to Down and notify this fault to RouterA and RouterB. RouterA and RouterB process the neighbor Down event and select routes again using BGP.

7.2.11 BGP Tracking

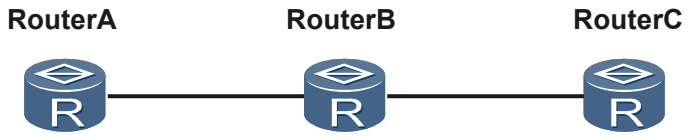
BGP tracking provides fast link fault detection to speed up network convergence. When a fault occurs on the link between BGP peers that have BGP tracking configured, BGP tracking can quickly detect peer unreachability and instruct the routing management module to notify BGP of the fault, implementing rapid network convergence.

Compared to BFD, BGP tracking is easy to configure because it needs to be configured only on the local device. BGP tracking is a fault detection mechanism at the routing layer, whereas BFD is a fault detection mechanism at the link layer. BGP route convergence on a network where BGP tracking is configured is slower than that on a network where BFD is configured. Therefore, BGP tracking cannot meet the requirements of voice services that require fast convergence.

Applications

As shown in **Figure 7-11**, RouterA and RouterB, and RouterB and RouterC establish IGP connections. RouterA and RouterC establish an IBGP peer relationship. BGP tracking is configured on RouterA. When a fault occurs on the link between RouterA and RouterB, IGP performs fast convergence. Subsequently, BGP tracking detects the unreachability of the route to RouterC and notifies the fault to BGP on Router A, which then interrupts the BGP connection with RouterC.

Figure 7-11 Networking diagram of BGP tracking



NOTE

If establishing an IBGP peer relationship requires IGP routes, the interval between peer unreachability discovery and connection interruption needs to be configured, and this interval must be longer than the IGP route convergence time. Otherwise, the BGP peer relationship may have been interrupted before IGP route flapping caused by transient interruption is suppressed, causing unnecessary BGP convergence.

7.2.12 BGP Auto FRR

BGP Auto Fast Reroute (FRR) is a protection measure against link failures. It applies to the network topology with primary and backup links and provides sub-second-level switching between two BGP peers or two next hops.

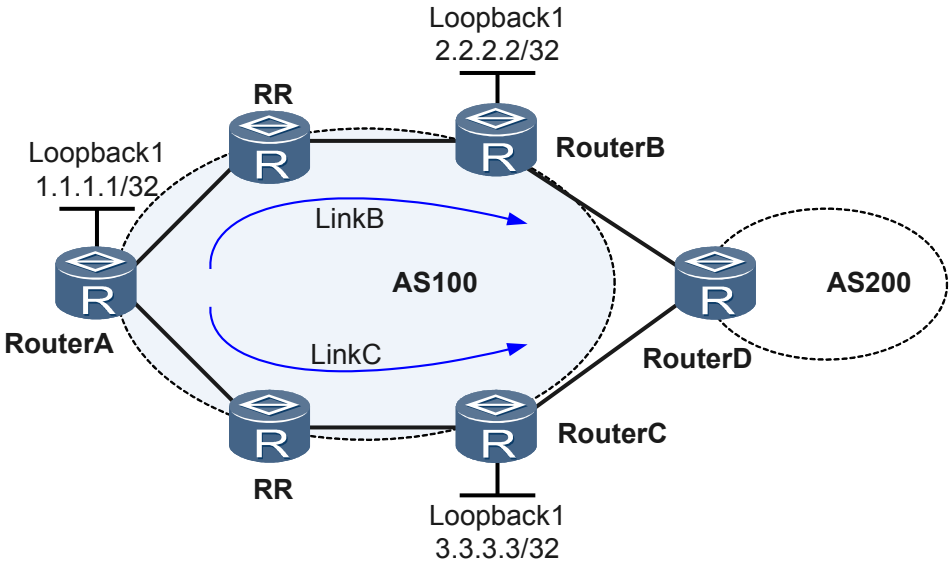
After BGP Auto FRR is enabled on a device, the device selects the optimal route from the routes that carry the same prefix and are learned from multiple peers as the primary link to forward

packets, and uses the second optimal route as the backup link. When the primary link becomes faulty, the system rapidly responds to the notification that the BGP route becomes unreachable, and then switches traffic from the primary link to the backup link. After BGP convergence is complete, BGP Auto FRR uses the optimal route selected by BGP to guide traffic forwarding. For details about Auto FRR, see "Auto FRR" in *Feature Description - IP Routing*.

Applications

As shown in **Figure 7-12**, RouterD advertises a learned BGP route to RouterB and RouterC in AS 100; RouterB and RouterC then advertise the BGP route to RouterA through a route reflector. RouterA receives two routes whose next hops are RouterB and RouterC respectively. Then RouterA selects a route according to the configured policy. Assume that the route sent from RouterB, namely LinkB, is preferred. The route sent from RouterC, namely LinkC, then functions as the backup link.

Figure 7-12 Networking diagram of BGP Auto FRR



When a router along LinkB fails or faults occur on LinkB, the next hop of the route from RouterA to RouterB becomes invalid. If BGP Auto FRR is enabled on RouterA, the forwarding plane quickly switches traffic sent from RouterA to RouterD to LinkC. This prevents traffic loss. In addition, RouterA reselects the route sent from RouterC and updates the FIB table.

7.2.13 BGP GR and NSR

BGP graceful restart (GR) and non-stop routing (NSR) are high availability solutions that minimize the impact of device failures on user services.

BGP GR

BGP GR ensures that the forwarding plane continues to guide data forwarding during a device restart or active/standby switchover. The operations on the control plane, such as reestablishing peer relationships and performing route calculation, do not affect the forwarding plane. This

mechanism prevents service interruptions caused by route flapping and improves network reliability.

GR concepts are as follows:

- GR restarter: is the device that is restarted by the administrator or triggered by failures to perform GR.
- GR helper: is the neighbor that helps the GR restarter to perform GR.
- GR time: is the time during which the GR helper retains forwarding information after detecting the restart or active/standby switchover of the GR restarter.

BGP GR process is as follows:

1. Using the BGP capability negotiation mechanism, the GR restarter and helper know each other's GR capability and establish a GR session.
2. When detecting the restart or active/standby switchover of the GR restarter, the GR helper does not delete the routing information and forwarding entries of the GR restarter or notify other neighbors of the restart or switchover, but waits to reestablish a BGP connection with the GR restarter.
3. The GR restarter reestablishes neighbor relationships with all GR helpers before the GR time expires.

BGP NSR

NSR is a reliability technique that prevents neighbors from detecting the control plane switchover. It applies to the devices that have the active and standby MPUs configured. Compared to GR, NSR does not require the help of neighbors and does not need to deal with interoperability issues. For details about NSR, see "NSR" in the *Feature Description - Reliability*.

Comparisons Between Active/Standby Switchovers with and Without GR and NSR

Table 7-4 Comparisons between active/standby switchovers with and without GR and NSR

Active/Standby Switchover Without GR and NSR	Active/Standby Switchover in GR Mode	Active/Standby Switchover in NSR Mode
The BGP peer relationship is reestablished.	The BGP peer relationship is reestablished.	The BGP peer relationship is reestablished.
Routes are recalculated.	Routes are recalculated.	Routes are recalculated.
The forwarding table changes.	The forwarding table remains unchanged.	The forwarding table remains unchanged.
Traffic is lost during forwarding, and services are interrupted.	No traffic is lost during forwarding, and services are not affected.	No traffic is lost during forwarding, and services are not affected.

Active/Standby Switchover Without GR and NSR	Active/Standby Switchover in GR Mode	Active/Standby Switchover in NSR Mode
The network detects route changes, and route flapping occurs for a short period of time.	Except the neighbors of the device where the active/standby switchover occurs, other routers do not detect route changes.	The network does not detect route changes.
-	The GR restarter requires neighbors to support the GR helper function. The GR helper function does not allow multiple neighbors to perform active/standby switchovers in GR mode simultaneously.	Neighbors do not need to support the NSR function.

7.2.14 BGP ORF

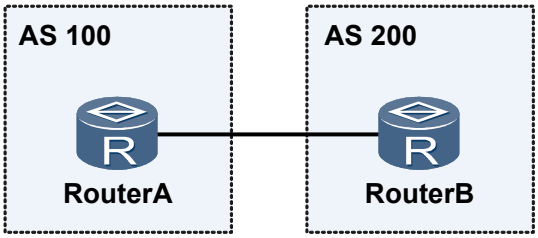
RFC 5291 and RFC 5292 define the prefix-based BGP outbound route filtering (ORF) capability to advertise required BGP routes. BGP ORF allows a device to send prefix-based import policies in a Route-refresh message to BGP peers. BGP peers construct export policies based on these import policies to filter routes before sending these routes, which has the following advantages:

- Prevents the local device from receiving a large number of unnecessary routes.
- Reduces CPU usage of the local device.
- Simplifies the configuration of BGP peers.
- Improves link bandwidth efficiency.

Applications

BGP ORF applies to the scenario when a device wants BGP peers to send only required routes, and BGP peers do not want to maintain different export policies for different devices.

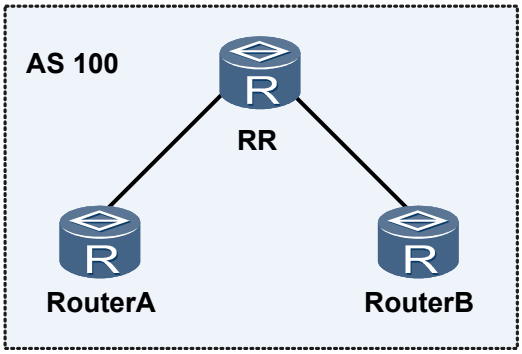
Figure 7-13 Inter-AS EBGP peers



As shown in **Figure 7-13**, after negotiating the prefix-based ORF capability with RouterB, RouterA adds the local prefix-based import policies to a Route-refresh message and sends the message to RouterB. RouterB constructs export policies based on the received Route-refresh

message and sends required routes to RouterA using a Route-refresh message. RouterA receives only required routes, and RouterB does not need to maintain routing policies. This reduces the configuration workload.

Figure 7-14 Intra-AS route reflector



As shown in **Figure 7-14**, there is a route reflector (RR) in AS 100. RouterA and RouterB are the clients of the RR. RouterA, Router B, and the RR negotiate the prefix-based ORF capability. RouterA and RouterB then add the local prefix-based import policies to Route-refresh messages and send the messages to the RR. The RR constructs export policies based on the received import policies and reflects required routes in Route-refresh messages to RouterA and RouterB. RouterA and RouterB receive only required routes, and the RR does not need to maintain routing policies. This reduces the configuration workload.

7.2.15 Dynamic Update Peer-Groups

Currently, the rapid growth in the size of the routing table and the complexity of the network topology require BGP to support more peers. Especially in the case of a large number of peers and routes, high-performance grouping and forwarding are required when a router needs to send routes to a large number of BGP peers, most of which share the same outbound policies.

The dynamic update peer-groups feature treats all the BGP peers with the same outbound policies as an update-group. In this case, routes are grouped uniformly and then sent separately. That is, each route to be sent is grouped once and then sent to all peers in the update-group, improving grouping efficiency exponentially. For example, a route reflector (RR) has 100 clients and needs to reflect 100,000 routes to these clients. If the RR sends the routes grouped per peer to 100 clients, the total number of times that all routes are grouped is 10,000,000 (100,000 x 100). After the dynamic update peer-groups feature is used, the total number of grouping times changes to 100,000 (100,000 x 1), improving grouping performance by a factor of 100.

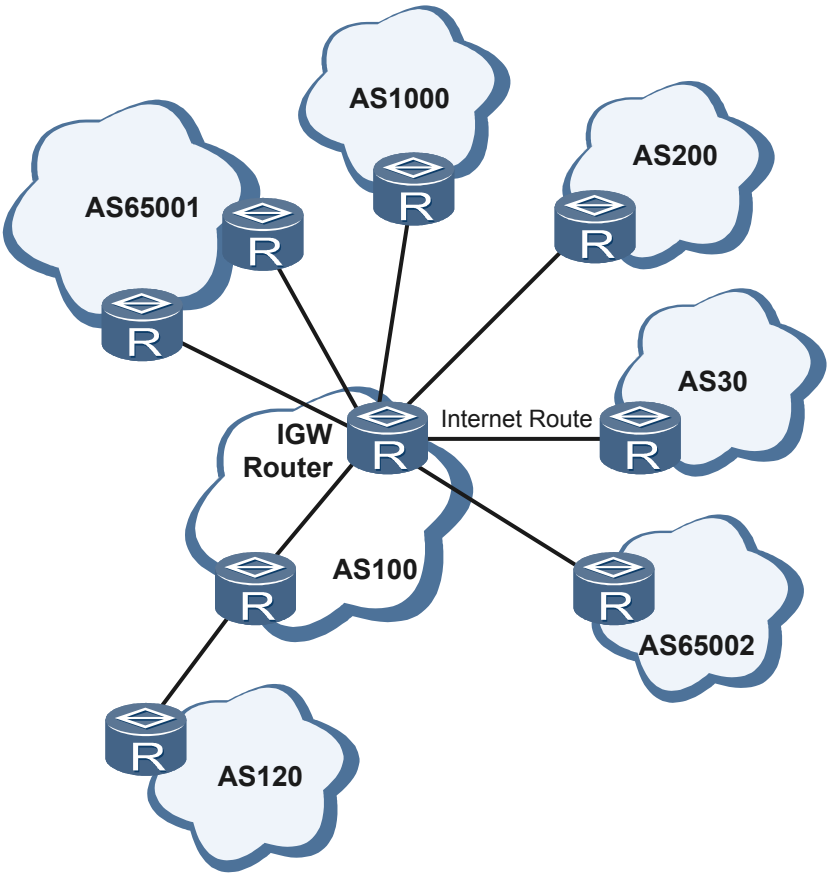
Applications

BGP uses the dynamic update peer-groups technology when a large number of peers and routes exist and most peers share the same outbound policies, improving BGP route grouping and forwarding performance. The dynamic update peer-groups feature applies to the following scenarios:

- International gateway

As shown in **Figure 7-15**, the Internet gateway (IGW) router sends routes to all neighboring ASs. If the IGW router supports the dynamic update peer-groups feature, its BGP route forwarding performance will be greatly improved.

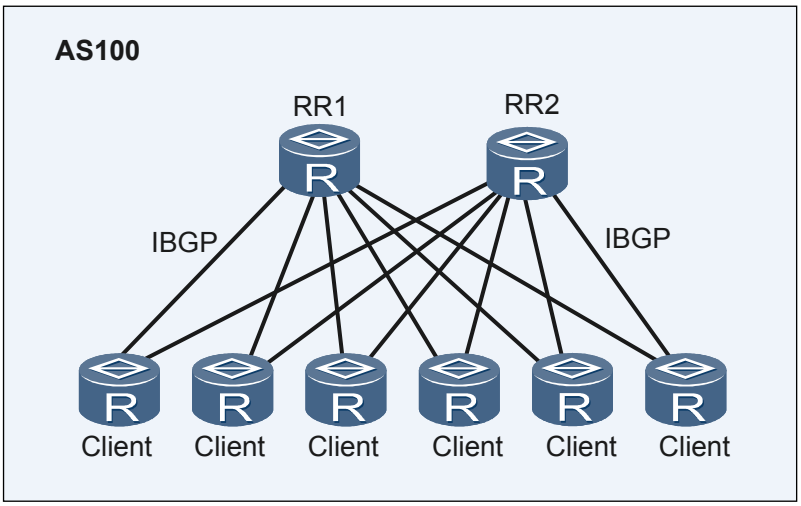
Figure 7-15 Networking diagram of the international gateway



- RR

  As shown in **Figure 7-16**, RRs send routes to all clients. If the RRs support the dynamic update peer-groups feature, their BGP route forwarding performance will be greatly improved.

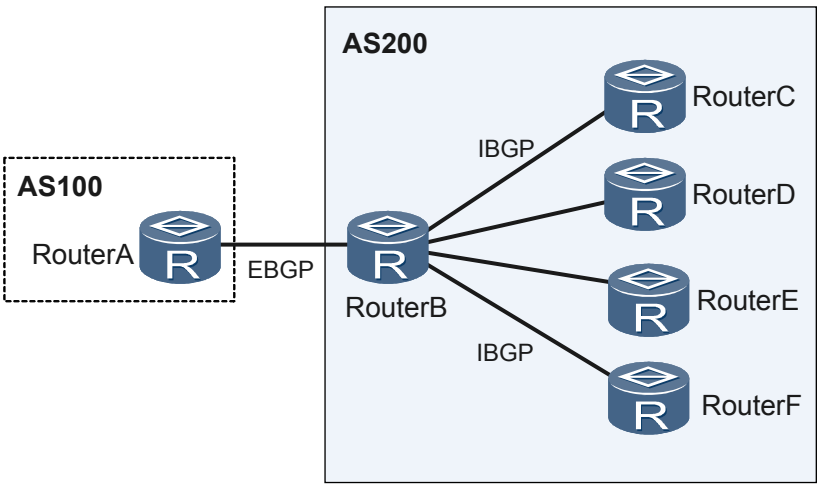
Figure 7-16 Networking diagram of RRs



- ASBR

  As shown in **Figure 7-17**, RouterB, as an Autonomous System Boundary Router (ASBR), sends all the routes received from an EBGP neighbor RouterA to all IBGP neighbors. If RouterB supports the dynamic update peer-groups feature, its BGP route forwarding performance will be greatly improved.

Figure 7-17 Networking diagram of a PE connecting to multiple IBGP neighbors



7.2.16 MP-BGP

Traditional BGP-4 manages only IPv4 routing information. Inter-AS transmission of other network layer protocol packets (such as IPv6 and multicast packets) is limited. To support multiple network layer protocols, Multiprotocol BGP (MP-BGP) is designed in RFC 4760 as an extension to BGP-4. MP-BGP uses extended attributes and address families to support IPv6, multicast, and VPN, without changing the existing BGP packet forwarding and routing mechanism.

MP-BGP is called BGP4+ on IPv6 unicast networks or called multicast BGP (MBGP) on IPv4 multicast networks. MP-BGP establishes separate topologies for IPv6 unicast networks and IPv4 multicast networks, and stores IPv6 unicast and IPv4 multicast routing information in different routing tables. This ensures that routing information of IPv6 unicast networks and IPv4 multicast networks is separated from each other, and allows routes of different networks to be maintained using different routing policies.

Extended Attributes

In BGP, an Update message carries three IPv4-related attributes: NLRI, Next_Hop, and Aggregator.

To support multiple network layer protocols, BGP requires NLRI and Next_Hop attributes to carry information about network layer protocols. Therefore, MP-BGP uses the following new optional non-transitive attributes:

- MP_REACH_NLRI: indicates the multiprotocol reachable NLRI. It is used to advertise reachable routes and next hop information.
- MP_UNREACH_NLRI: indicates the multiprotocol unreachable NLRI. It is used to withdraw unreachable routes.

Address Families

MP-BGP uses address families to differentiate network layer protocols. For the values of address families, see RFC 3232 (Assigned Numbers). Currently, devices support the following address family views:

- BGP-IPv4 unicast address family view
- BGP-IPv4 multicast address family view
- BGP-VPN instance IPv4 address family view
- BGP-VPNv4 address family view
- BGP-IPv6 unicast address family view
- BGP-IPv6 unicast address family view
- BGP-VPN instance IPv6 address family view
- BGP-VPNv6 address family view

For details about the BGP VPNv4 address family and BGP VPN instance IPv4 address family, see "BGP/MPLS IP VPN" in *Feature Description - VPN*.

7.3 References

Table 7-5 lists the references of this feature.

Table 7-5 References

Document	Description	Remarks
RFC 827	Exterior Gateway Protocol (EGP)	-
RFC 1997	BGP Communities Attribute	-

Document	Description	Remarks
RFC 2439	BGP Route Flap Damping	-
RFC 2918	Route Refresh Capability for BGP-4	-
RFC 3065	Autonomous System Confederations for BGP	-
RFC 3232	Assigned Numbers: RFC 1700 is Replaced by an On-line Database	-
RFC 3392	Capabilities Advertisement with BGP-4	-
RFC 3682	The Generalized TTL Security Mechanism (GTSM)	-
RFC 4271	A Border Gateway Protocol 4 (BGP-4)	-
RFC 4456	BGP Route Reflection	-
RFC 4486	Subcodes for BGP Cease Notification Message	-
RFC 4724	Graceful Restart Mechanism for BGP	-
RFC 4760	Multiprotocol Extensions for BGP-4	-
RFC 4893	BGP Support for Four-octet AS Number Space	-
draft-rijsman-bfd-down-subcode-00	BFD Down Subcode for BGP Cease Notification Message	-

8 Routing Policy

About This Chapter

8.1 Introduction to the Routing Policy

Definition

Routing policies are used to filter routes and set attributes for routes. By changing route attributes (including reachability), a route policy changes the path that network traffic passes through.

Purpose

When advertising, receiving, and importing routes, routing protocols implement certain policies based on actual networking requirements to filter routes and change the attributes of the routes. Routing policies serve the following purposes:

- To control route receiving and advertising

  Only the required and valid routes are received or advertised. This reduces the size of the routing table and improves network security.
- To control route importing

  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.
- To 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:

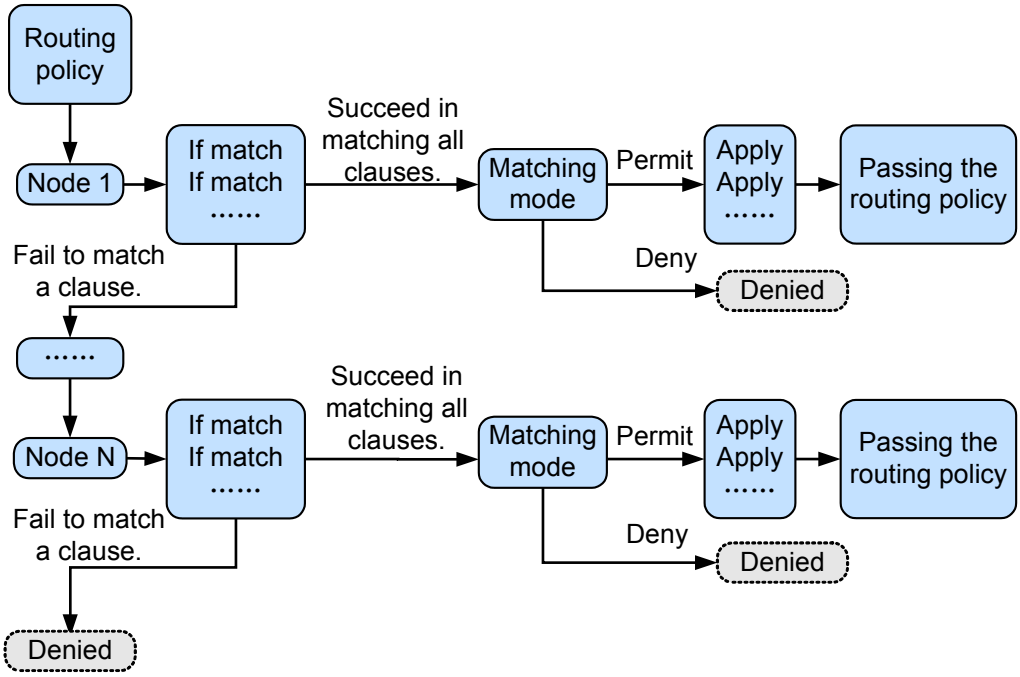
- Controls the size of the routing table, saving system resources.
- Controls route receiving, advertising and importing, improving network security.
- Modifies attributes of routes for proper traffic planning, improving network performance.

8.2 Principle

A routing policy uses different matching rules and modes to select routes and change route attributes. Six filters in the routing policy can be used independently to filter routes in special scenarios. If the device supports the BGP to IGP function, the private attributes of BGP can serve as matching rules when the IGP imports BGP routes.

Routing Policy Principle

Figure 8-1 Working mechanism of the routing policy



As shown in **Figure 8-1**, a routing policy consists of *N* nodes ($N \geqslant 1$). The system checks routes in the nodes of a routing policy with the node ID in ascending order. The **If-match** clauses define matching rules related to route attributes and six filters.

When a route matches all **If-match** clauses in a node, the route enters the matching mode without being checked in other nodes. The following two matching modes are supported:

- **permit**: A route is permitted, and actions defined by the **Apply** clauses are performed on the route to set its attributes.
- **deny**: A route is denied.

If a route does not match one **If-match** clause in a node, the route enters to the next node. If a route does not match any one of the nodes, the route is filtered out.

Filters

The six filters specified in **If-match** clauses in a routing policy are access control list (ACL), IP prefix list, AS_Path filter, community filter, extended community filter, and RD filter. The six filters have their own matching rules and modes. Therefore, they can be used independently to filter routes in some special situations.

ACL

ACLs check inbound interface, source or destination IP address, source or destination port number, and protocol of packets to filter routes. ACLs can be used independently when routing protocols advertise and receive routes. The **If-match** clauses in a routing policy support only basic ACLs.

ACLs can be used in not only a routing policy but other scenarios. For details, see the *Feature Description - Security - ACL*.

IP prefix list

IP prefix lists check IP prefixes of the source IP address, destination IP address, and next hop address to filter routes. They can be used independently when routing protocols advertise and receive routes.

Each IP prefix list consists of multiple indexes, and each index matches a node. An IP prefix list checks routes in all nodes with the indexes in ascending order. If a route matches one node, the route is no longer checked by other nodes. If a route does not match any one of the nodes, the route is filtered out.

The IP prefix list supports exact matching or matching within a specified mask length.

NOTE

When the IP address is 0.0.0.0, a wildcard address, all routes in the mask length range are permitted or denied.

AS_Path filter

The AS_Path filter uses the AS_Path attribute of BGP to filter routes. It can be used independently when BGP advertises and receives routes.

The AS_Path attribute records all ASs that a route passes through. For details about the AS_Path attribute, see "Introduction to BGP" in the *Feature Description - IP Routing - BGP*.

Community filter

The community filter uses the community attribute of BGP to filter routes. It can be used independently when BGP advertises and receives routes.

The community attribute identifies a group of routes with the same properties. For details about the community attribute, see "Introduction to BGP" in the *Feature Description - IP Routing - BGP*.

Extended community filter

The extended community filter uses the extended community attribute of BGP to filter routes. It can be used independently when VPN targets are used to identify routes in a VPN.

Currently, the extended community filter applies only to the VPN target attribute in a VPN. On a BGP/MPLS IP VPN, VPN targets are used to control the advertising and receiving of VPN routing information between sites. For details about the VPN target attribute, see "Introduction to BGP/MPLS IP VPN" in the *Feature Description - VPN - BGP/MPLS IP VPN*.

Route Distinguisher (RD) filter

The RD filter uses the RD attribute in a VPN to filter routes. It can be used independently when the RD attribute is used to identify routes in a VPN.

A VPN instance uses RDs to separate address spaces and distinguish the IP prefixes with the same address space. For details about the RD attribute, see "Introduction to BGP/MPLS IP VPN" in the *Feature Description - VPN - BGP/MPLS IP VPN*.

BGP to IGP function

The BGP to IGP function enables IGP to identify private attributes of BGP such as the community, extended community, and AS-Path attributes.

Routing policies can be used when an IGP imports BGP routes. BGP private attributes can be used as matching rules in routing policies only when the device supports the BGP to IGP function. When the device does not support the BGP to IGP function, the IGP cannot identify private attributes of BGP routes. Therefore, the matching rule does not take effect.

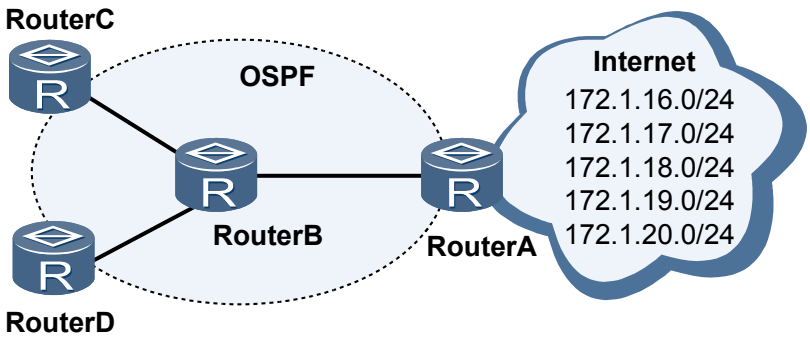
8.3 Applications

Specific Routes Filtering

On the OSPF-enabled network shown in **Figure 8-2**, Router A receives routes from the Internet and advertises some of the routes to Router B.

- Router A advertises only routes 172.1.17.0/24, 172.1.18.0/24, and 172.1.19.0/24 to Router B.
- Router C accepts only the route 172.1.18.0/24.
- Router D accepts all the routes advertised by Router B.

Figure 8-2 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:

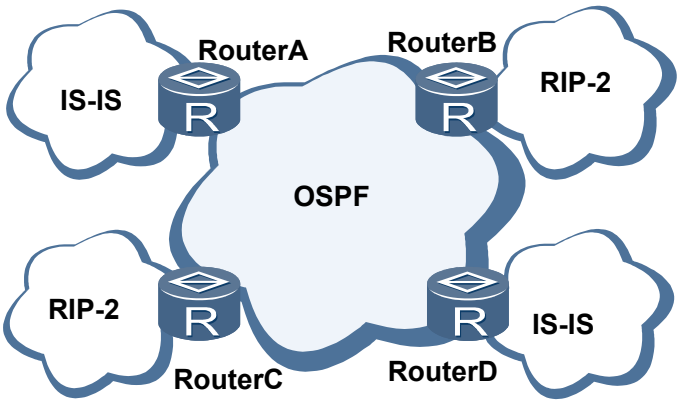
- Use IP prefix lists
  - Configure an IP prefix list on Router A and configure the IP prefix list as an export policy of Router A to be used by OSPF.
  - Configure another IP prefix list on Router C and configure the IP prefix list as an import policy of Router C to be used by OSPF.
- Use route-policies
  - Configure a Route-Policy (the matching rules can be the IP prefix list, cost, or route tag) on Router A and configure the Route-Policy as an export policy of Router A to be used by OSPF.
  - Configure another Route-Policy on 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 8-3**, 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.

Figure 8-3 Networking diagram for transparently transmitting routes of other protocols through an OSPF AS

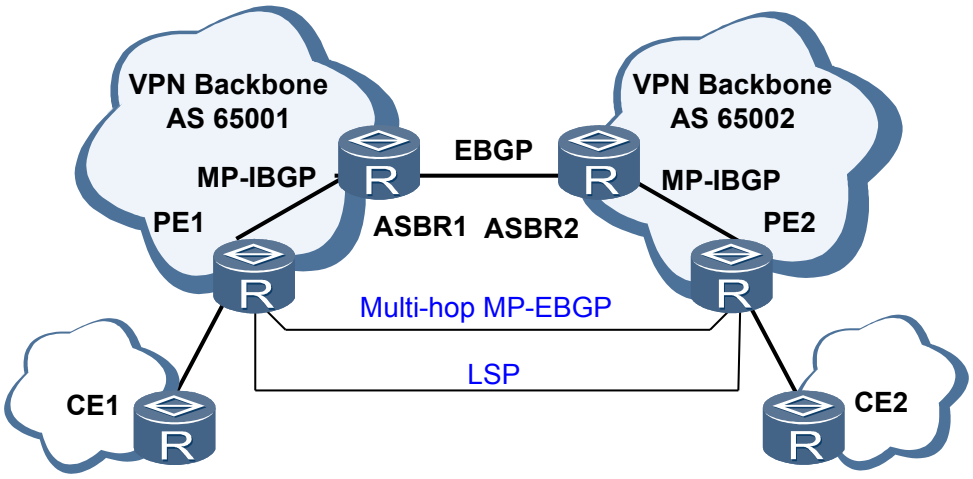


To meet the preceding requirements, configure a Route-Policy on Router A to set a tag for the imported IS-IS routes. Router D identifies the IS-IS routes from OSPF routes based on the tag.

Routing Policy Application in Inter-AS VPN Option C

On the network shown in Router, CE1 and CE2 communicate with each other through inter-AS VPN Option C.

Figure 8-4 Networking diagram for implementing route-policies in the inter-AS VPN Option C scenario



To establish an inter-AS LSP between PE1 and PE2, route-policies need to be configured on ASBRs.

- When an ASBR advertises the routes received from a PE in the same AS to the peer ASBR, the ASBR allocates MPLS labels to the routes using a Route-Policy.
- When an ASBR advertises labeled IPv4 routes to a PE in the same AS, the ASBR reallocates MPLS labels to the routes using another Route-Policy.

In addition, to control route transmission between different VPN instances on a PE, configure a Route-Policy on the PE and configure the Route-Policy as an import or export policy on the VPN instances.

8.4 References

None.