Huawei AR1200 系列企业路由器

V200R002C00

特性描述-IP 路由

文档版本　01

发布日期　2011-12-30

华为技术有限公司

HUAWEI

商标声明

注意

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

前 言

读者对象

本文档针对 AR1200 的 IP 路由协议，介绍 AR1200 支持的各协议的基本概念、原理、特性和应用。

通过阅读本文档，可以更好地理解和掌握 AR1200IP 路由协议产生背景、基本概念、基本原理、参考标准等。

本文档主要适用于以下工程师：

- 网络规划工程师
- 调测工程师
- 数据配置工程师
- 系统维护工程师

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

符号	说明
危险	以本标志开始的文本表示有高度潜在危险，如果不能避免，会导致人员死亡或严重伤害。
警告	以本标志开始的文本表示有中度或低度潜在危险，如果不能避免，可能导致人员轻微或中等伤害。
注意	以本标志开始的文本表示有潜在风险，如果忽视这些文本，可能导致设备损坏、数据丢失、设备性能降低或不可预知的结果。
窍门	以本标志开始的文本能帮助您解决某个问题或节省您的时间。
说明	以本标志开始的文本是正文的附加信息，是对正文的强调和补充。

命令行格式约定

格式	意义
粗体	命令行关键字（命令中保持不变、必须照输的部分）采用**加粗**字体表示。
斜体	命令行参数（命令中必须由实际值进行替代的部分）采用*斜体*表示。
[ ]	表示用“[ ]”括起来的部分在命令配置时是可选的。
{ x | y | ... }	表示从两个或多个选项中选取一个。
[x | y | ...]	表示从两个或多个选项中选取一个或者不选。
{ x | y | ... } *	表示从两个或多个选项中选取多个，最少选取一个，最多选取所有选项。
[x | y | ...] *	表示从两个或多个选项中选取多个或者不选。
&<1-n>	表示符号&的参数可以重复 1 ～ n 次。
#	由“#”开始的行表示为注释行。

修订记录

修改记录累积了每次文档更新的说明。最新版本的文档包含以前所有文档版本的更新内容。

文档版本 01 (2011-12-30)

第一次正式发布。

目 录

1 IP 路由概述

关于本章

1.1 介绍

定义

路由是数据通信网络中最基本的要素。路由信息就是指导报文发送的路径信息，路由的过程就是报文中继转发的过程。

1.2 参考标准和协议

无

1.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R001C00

1.4 原理描述

1.4.1 路由器

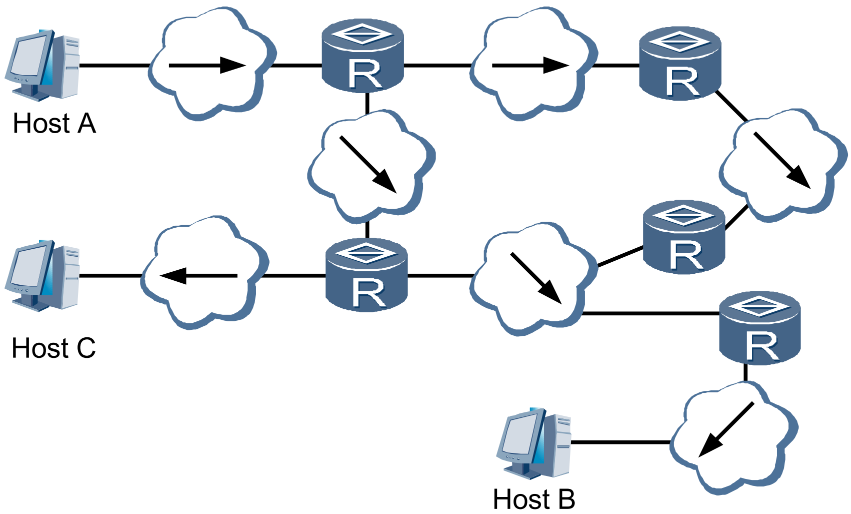
在因特网中，网络连接设备用来控制网络流量和保证网络数据传输质量。常见的网络连接设备有集线器（Hub）、网桥（Bridge）、交换机（Switch）和路由器（Router）。

路由器是一种典型的网络连接设备，用来进行路由选择和报文转发。路由器根据收到报文的目的地址选择一条合适的路径（包含一个或多个路由器的网络），将报文传送到下一个路由器，路径目的终端的路由器负责将报文送交目的主机。路由器可以为数据传输选择最佳路径。

路由器到与它直接相连网络的跳数为 0，通过一台路由器可达的网络的跳数为 1，其余以此类推。若一台路由器通过一个网络与另一台路由器相连接，则这两台路由器相隔一个网段，在因特网中认为这两台路由器相邻。至于每一个网段又由哪几条物理链路构成，路由器并不关心。

例如，在图 1-1 中，HostA 到 HostC 共经过了 3 个网络和 2 台路由器。在图中用箭头表示这些网段。

图 1-1 路由跳数和网段的概念



由于网络大小可能相差很大，而每个网段的实际长度并不相同，因此对不同的网络，可以将其网段乘以一个加权系数，用加权后的网段数来衡量通路的长短。

采用网段数最小的路由有时也并不一定是最理想的。例如，经过三个高速局域网段的路由可能比经过两个低速广域网段的路由快得多。

1.4.2 路由协议

上面提到路由器的主要功能是路由选择和报文转发，这种功能的实现需用到路由协议。路由协议是路由器之间维护路由表的规则，用于发现路由，生成路由表，并指导报文转发。路由协议可分为链路状态协议和距离矢量协议。路由协议决定 IP 路由表和转发表（Forwarding Information Base）中存放哪些路由信息。

1.4.3 路由表和 FIB 表

每个路由器都至少保存着一张路由表和一张 FIB 表。路由器通过路由表选择路由，通过 FIB（Forwarding Information Base）表指导报文转发。

- 路由表中保存了各种路由协议发现的路由，根据来源不同，路由表中的路由通常可分为以下三类：
  - 链路层协议发现的路由（也称为接口路由或直连路由）。
  - 由网络管理员手工配置的静态路由。
  - 动态路由协议发现的路由。
- FIB 表中每条转发项都指明到达某网段或某主机的报文应通过路由器的哪个物理接口或逻辑接口发送，然后就可到达该路径的下一个路由器，或者不再经过别的路由器而传送到直接相连的网络中的目的主机。

路由表

每台路由器中都保存着一张本地核心（管理）路由表，同时各个路由协议也维护着自己的路由表。

- 协议路由表

  协议路由表中存放着该协议发现的路由信息。

  路由协议可以引入并发布其他协议生成的路由。例如，在路由器上运行 OSPF（Open Shortest Path First）协议，需要使用 OSPF 协议通告直连路由、静态路由或者 IS-IS（Intermediate System-Intermediate System）路由时，要将这些路由引入到 OSPF 协议的路由表中。

- 本地核心路由表

  路由器使用本地核心路由表用来保存协议路由和决策优选路由，并负责把优选路由下发到 FIB，FIB 进行指导转发。这张路由表依据各种路由协议的优先级和度量值来选取路由。可以使用 **display ip routing-table** 命令查看。

  说明

  对于支持 L3VPN（Layer 3 Virtual Private Network）的路由器，每一个 VPN-Instance 拥有一个自己的管理路由表（本地核心路由表）。

路由表中的内容

在 AR1200 中，执行命令 **display ip routing-table** 时，可以查看到路由器的路由表简表，如下：

```
<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
GigabitEthernet2/0/0
     10.10.10.0/24  Direct 0    0           D    10.10.10.10
Virtual-Template1
    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
GigabitEthernet2/0/0
 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
```

路由表中包含了下列关键项：

- Destination：目的地址。用来标识 IP 包的目的地址或目的网络。
- Mask：网络掩码。与目的地址一起来标识目的主机或路由器所在的网段的地址。
  - 将目的地址和网络掩码“逻辑与”后可得到目的主机或路由器所在网段的地址。例如：目的地址为 1.1.1.1，掩码为 255.255.255.0 的主机或路由器所在网段的地址为 1.1.1.0。

  - 掩码由若干个连续“1”构成，既可以用点分十进制表示，也可以用掩码中连续“1”的个数来表示。例如掩码 255.255.255.0 长度为 24，即可以表示为 24。
- Proto：用来学习路由的协议。
- Pre：本条路由加入 IP 路由表的优先级。针对同一目的地，可能存在不同下一跳、出接口等的若干条路由，这些不同的路由可能是由不同的路由协议发现的，也可以是手工配置的静态路由。优先级高（数值小）者将成为当前的最优路由。各协议路由优先级请参见表 **1-1**。
- Cost：路由开销。当到达同一目的地的多条路由具有相同的优先级时，路由开销最小的将成为当前的最优路由。

  说明

  Preference 用于不同路由协议间路由优先级的比较，Cost 用于同一种路由协议内部不同路由优先级的比较。
- NextHop：下一跳 IP 地址。说明 IP 包所经由的下一个设备。
- Interface：输出接口。说明 IP 包将从该路由器哪个接口转发。

根据路由的目的地不同，可以划分为：

- 网段路由：目的地为网段
- 主机路由：目的地为主机

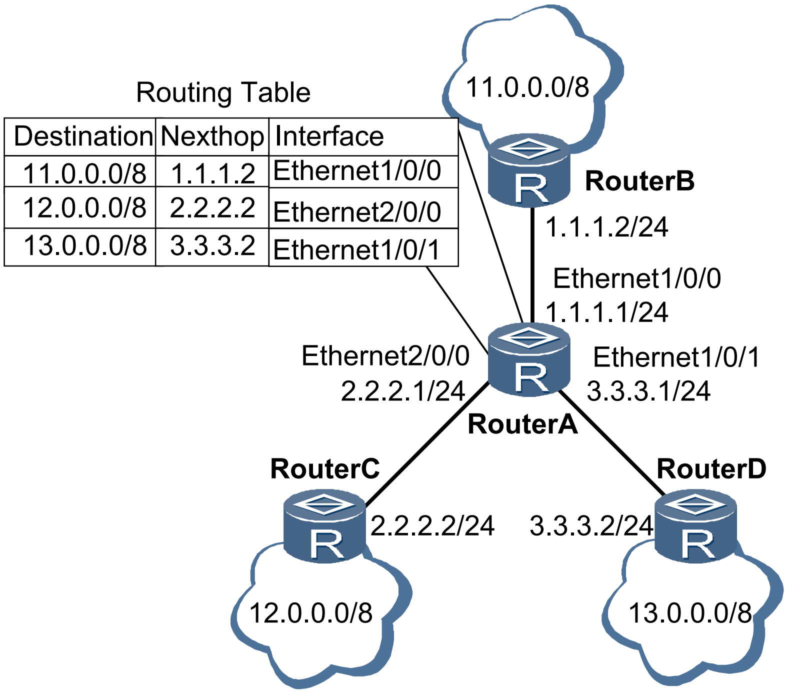
另外，根据目的地与该路由器是否直接相连，又可分为：

- 直接路由：目的地所在网络与路由器直接相连
- 间接路由：目的地所在网络与路由器不是直接相连

为了不使路由表过于庞大，可以设置一条缺省路由。凡遇到查找路由表失败后的数据包，就选择缺省路由转发。例如上面路由表中目的地址是 0.0.0.0/0 的路由就是缺省路由。

在图 **1-2** 所示的网络中，RouterA 与三个网络相连，因此有三个 IP 地址和三个出接口，其路由表如图所示。

图 1-2 路由表示意图



路由超限自动恢复

本地核心路由表里保存着各路由协议的路由，如果本地核心路由表里的路由数量达到系统上限，协议路由表将无法向本地核心路由表添加路由。本地核心路由表有以下几种路由限制：

- 整机路由限制
- 整机路由前缀限制
- 组播 IGP 路由限制
- 所有私网路由限制
- VPN 路由限制
- VPN 路由前缀限制

如果协议由于某种路由限制而向本地核心路由表添加路由失败，系统会记录下本次添加路由的协议和对应的路由表 TableID，表示该协议曾经向该路由表中添加路由失败。

当协议删除本地核心路由表里的路由释放了路由表的空间之后，路由超限解除，系统会通知所有向本地核心路由表添加路由失败的协议，重新向本地核心路由表添加路由，使得本地核心路由表中的路由能够得到最大程度的恢复。是否可以完全恢复，取决于释放的路由表空间的大小。

FIB 表的匹配

在路由表选择出路由后，路由表会将激活路由下发到 FIB 表中。当报文到达路由器时，会通过查找 FIB 表进行转发。

FIB 表的匹配遵循最长匹配原则。查找 FIB 表时，报文的目的地址和 FIB 中各表项的掩码进行按位“逻辑与”，得到的地址符合 FIB 表项中的网络地址则匹配。最终选择一个最长匹配的 FIB 表项转发报文。

说明

FIB 表的详细描述请参见《Huawei AR1200 系列 企业路由器 特性描述-IP 业务》。

例如，一台路由器上的路由表简表如下：

```
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        GigabitEthernet1/0/1
 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        GigabitEthernet2/0/1
```

说明

完整的路由表中包含激活路由和未激活路由，路由表简表中只显示激活路由。完整的路由表可以通过命令 **display ip routing-table verbose** 查看。

一个目的地址是 9.1.2.1 的报文进入路由器，查找对应的 FIB 表。

```
 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]        GigabitEthernet1/0/1  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]        GigabitEthernet2/0/1  0x0
```

首先，目的地址 9.1.2.1 与 FIB 表中各表项的掩码“0、8、16”作“逻辑与”运算，得到下面的网段地址：0.0.0.0/0、9.0.0.0/8、9.1.0.0/16。这三个结果可以匹配到 FIB 表中对应的三个表项：0.0.0.0/0 匹配长度是 0bit、9.0.0.0/8 匹配长度是 8bit、9.1.0.0/16 匹配长度是 16bit。

最终，AR1200 会选择最长匹配 9.1.0.0/16 表项，从接口 GigabitEthernet2/0/0 转发这条目的地址是 9.1.2.1 的报文。

1.4.4 静态路由与动态路由

AR1200 不仅支持静态路由，同时也支持 RIP（Routing Information Protocol）、OSPF、IS-IS 和 BGP（Border Gateway Protocol）等动态路由协议。

静态路由配置方便，对系统要求低，适用于拓扑结构简单并且稳定的小型网络。缺点是不能自动适应网络拓扑的变化，需要人工干预。

动态路由协议有自己的路由算法，能够自动适应网络拓扑的变化，适用于具有一定数量三层设备的网络。缺点是配置对用户要求比较高，对系统的要求高于静态路由，并将占用一定的网络资源。

1.4.5 动态路由协议的分类

对动态路由协议的分类可以采用以下不同标准：

作用范围

根据作用的范围，路由协议可分为：

- 内部网关协议（Interior Gateway Protocol，简称 IGP）：在一个自治系统内部运行，常见的 IGP 协议包括 RIP、OSPF 和 IS-IS。
- 外部网关协议（Exterior Gateway Protocol，简称 EGP）：运行于不同自治系统之间，BGP 是目前最常用的 EGP 协议。

使用的算法

根据使用的算法，路由协议可分为：

- 距离矢量协议（Distance-Vector）：包括 RIP 和 BGP。其中，BGP 也被称为路径矢量协议（Path-Vector）。
- 链路状态协议（Link-State）：包括 OSPF 和 IS-IS。

以上两种算法的主要区别在于发现路由和计算路由的方法。

目的地址类型

根据目的地址的类型，路由协议可分成：

- 单播路由协议（Unicast Routing Protocol）：包括 RIP、OSPF、BGP 和 IS-IS 等。
- 组播路由协议（Multicast Routing Protocol）：包括 PIM-SM（Protocol Independent Multicast-Sparse Mode）、PIM-DM（Protocol Independent Multicast-Dense Mode）等。

本部分手册主要介绍单播路由协议，组播路由协议请参见《Huawei AR1200 系列 企业路由器 特性描述-组播》。

静态路由和由路由协议发现的动态路由在路由器中是统一管理的，这些路由可通过相互引入等操作实现路由信息的重发布。

1.4.6 路由协议及路由优先级

路由优先级

对于相同的目的地，不同的路由协议（包括静态路由）可能会发现不同的路由，但这些路由并不都是最优的。事实上，在某一时刻，到某一目的地的当前路由仅能由唯一的路由协议来决定。为了判断最优路由，各路由协议（包括静态路由）都被赋予了一个优先级，当存在多个路由信息源时，具有较高优先级（取值较小）的路由协议发现的路由将成为最优路由。各种路由协议及其发现路由的缺省优先级如**表 1-1** 所示。

其中：0 表示直接连接的路由，255 表示任何来自不可信源端的路由；数值越小表明优先级越高。

表 1-1 路由协议及缺省时的路由优先级

路由协议或路由种类	相应路由的优先级
DIRECT	0
OSPF	10
IS-IS	15
STATIC	60

路由协议或路由种类	相应路由的优先级
RIP	100
OSPF ASE	150
OSPF NSSA	150
IBGP	255
EBGP	255

除直连路由（DIRECT）外，各种路由协议的优先级都可由用户手工进行配置。另外，每条静态路由的优先级都可以不相同。

AR1200 分别定义了外部优先级和内部优先级，外部优先级即前面提到的用户为各路由协议配置的优先级，缺省情况下如**表 1-1** 所示。

当不同的路由协议配置了相同的优先级后，系统会通过内部优先级决定哪个路由协议发现的路由将成为最优路由。路由协议的内部优先级如**表 1-2** 所示。

表 1-2 路由协议内部优先级

路由协议或路由种类	相应路由的优先级
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

例如，到达同一目的地 10.1.1.0/24 有两条路由可供选择，一条静态路由，另一条是 OSPF 路由，且这两条路由的协议优先级都被配置成 5。这时 AR1200 系统将根据**表 1-2** 所示的内部优先级进行判断。因为 OSPF 协议的内部优先级是 10，高于静态路由的内部优先级 60。所以系统选择 OSPF 协议发现的路由作为可用路由。

1.4.7 路由按优先级收敛

定义

按优先级收敛是提高网络可靠性的一项重要技术，可为关键业务提供更快的路由收敛速度。

按优先级收敛是指系统为路由设置不同的收敛优先级，分为 critical、high、medium、low 四种。系统根据这些路由的收敛优先级采用相对的优先收敛原则，即按照一定的调度比例进行路由收敛安装，指导业务的转发。

目的

随着网络的融合，区分服务的需求越来越强烈。某些路由可能指导关键业务的转发，如 VoIP，视频会议、组播等，对于运营商而言，这些关键的业务路由需要尽快收敛，而非关键路由可以相对慢一点收敛。因此，系统需要对不同路由按不同的收敛优先级处理，来提高网络可靠性。

原理描述

缺省情况下的公网路由收敛优先级如下表所示。路由协议优先计算并下发高收敛优先级的路由给系统。缺省情况下，系统按照优先级调度的权重为 critical: high: medium: low=8:4:2:1 对路由进行收敛，用户也可以根据实际组网重新配置收敛优先级的调度权重。

表 1-3 缺省时的公网路由收敛优先级

路由协议或路由种类	收敛优先级
DIRECT	high
STATIC	medium
OSPF 和 IS-IS 的 32 位主机路由	medium
OSPF（除 32 位主机路由外）	low
IS-IS（除 32 位主机路由外）	low
RIP	low
BGP	low

说明

对于私网路由，除了 OSPF 和 IS-IS 的 32 位主机路由标识为 medium，其余路由统一标识为 low。

1.4.8 负载分担与路由备份

负载分担

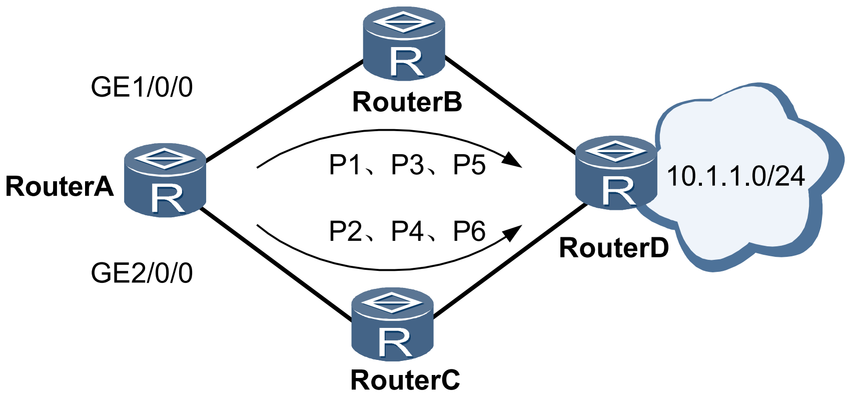
AR1200 支持多路由模式，即允许配置多条目的地相同且优先级也相同的路由。当到达同一目的地存在同一路由协议发现的多条路由时，且这几条路由的开销值也相同，那么

就满足负载分担的条件。通过在各协议视图下，配置 **maximum load-balancing** *number* 实现负载分担。负载分担分为两种方式：

- 逐包

  当配置负载分担的方式为逐包负载分担时，路由器在转发去往同一目的地的报文时，由 IP 层依次通过各条路径发送，也就是总是选择与发上一个包时不同的下一跳地址发送报文，即逐包的负载分担。如图 **1-3** 所示。

  图 1-3 逐包负载分担组网图

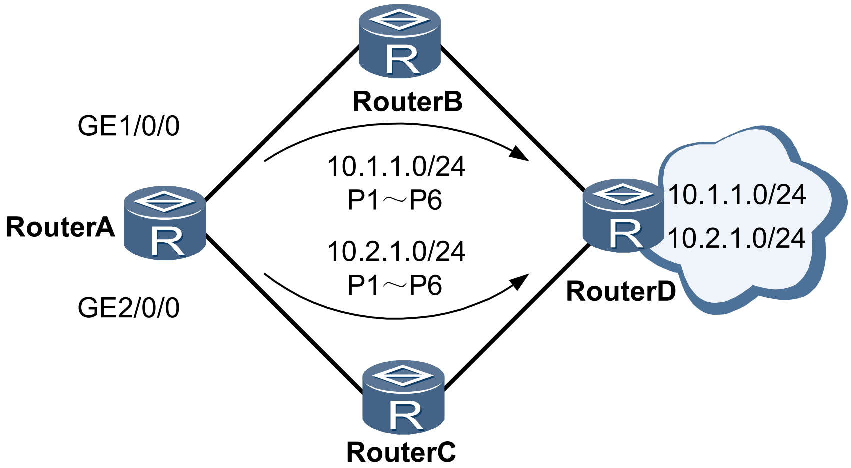
  RouterA 转发报文到目的地 10.1.1.0/24，待发送的报文为 P1、P2、P3、P4、P5、P6。为了进行负载分担，RouterA 的两个接口轮流发送到同一目的地址 10.1.1.0/24 的报文。报文发送的过程是：

  - 从接口 GigabitEthernet1/0/0 发送第 1 个报文 P1
  - 从接口 GigabitEthernet2/0/0 发送第 2 个报文 P2
  - 从接口 GigabitEthernet1/0/0 发送第 3 个报文 P3
  - 从接口 GigabitEthernet2/0/0 发送第 4 个报文 P4
  - 从接口 GigabitEthernet1/0/0 发送第 5 个报文 P5
  - 从接口 GigabitEthernet2/0/0 发送第 6 个报文 P6

- 逐流

  当配置负载分担的方式为逐流负载分担时，路由器根据五元组（源地址、目的地址、源端口、目的端口、协议）进行转发，当五元组相同时，路由器总是选择与上一次相同的下一跳地址发送报文。如图 **1-4** 所示。

图 1-4 逐流负载分担组网图



RouterA 分别转发报文到目的地址 10.1.1.0/24 和 10.2.1.0/24。逐流负载分担的选路原则是：同一流中的报文总是选择以前走过的路径。RouterA 转发报文的过程如下：

- 到达 10.1.1.0/24 的第 1 个报文 P1 是从接口 GigabitEthernet1/0/0 转发出去的，所以之后到达该目的地的报文都从 GigabitEthernet1/0/0 转发。
- 到达 10.2.1.0/24 的第 1 个报文 P1 是从接口 GigabitEthernet2/0/0 转发出去的，所以之后到达该目的地的报文都从 GigabitEthernet2/0/0 转发。

说明

AR1200 仅支持逐流负载分担方式。

目前的实现中，支持负载分担的路由协议为 RIP、OSPF、BGP 和 IS-IS，静态路由也支持负载分担。

说明

系统所允许的负载分担的具体路由条数，与实际产品型号相关。

1.4.9 路由信息的重发布

由于采用的算法不同，不同的路由协议可以发现不同的路由。当网络规模比较大，使用多种路由协议时，不同的路由协议间通常需要重发布各自发现的路由。

AR1200 支持将一种路由协议发现的路由引入到另一种路由协议中。每种路由协议都有相应的路由引入机制，具体内容请参见路由策略。

1.4.10 下一跳分离

定义

下一跳分离是一种提高路由收敛性能的技术，它把路由前缀与下一跳转发信息的直接关系变为间接关系。下一跳分离支持下一跳信息的单独刷新，而不必逐条刷新大量前缀，从而获得较快的收敛速度。

目的

在路由需要迭代的场景中，当 IGP 路由或隧道发生切换时，快速刷新 FIB 转发路径，实现流量的快速收敛，降低对业务的影响。

前缀与下一跳映射

下一跳分离的基础是路由前缀与下一跳的映射法则。为了满足不同场景下迭代路由和迭代隧道的需求，下一跳信息由地址族、原始下一跳地址、隧道策略等要素构成。系统给每个下一跳信息分配一个索引，进行迭代并将迭代结果通知路由协议和下发 FIB。

按需迭代

按需迭代是当依赖路由变化时，只对相关的下一跳进行重新迭代。如果路由的目的地址是某个下一跳信息的原始下一跳地址，或是其网段地址，则路由变化时会影响这个下一跳信息的迭代结果，否则路由的变化对下一跳没有影响。所以，当一条路由发生变化时，通过判断目的地址可以实现只对相关的下一跳进行重新迭代。

对于迭代隧道的情况，当一条隧道的状态（UP/DOWN）改变时，只需要对原始下一跳地址与隧道目的地址相同的下一跳信息进行重新迭代。

迭代策略

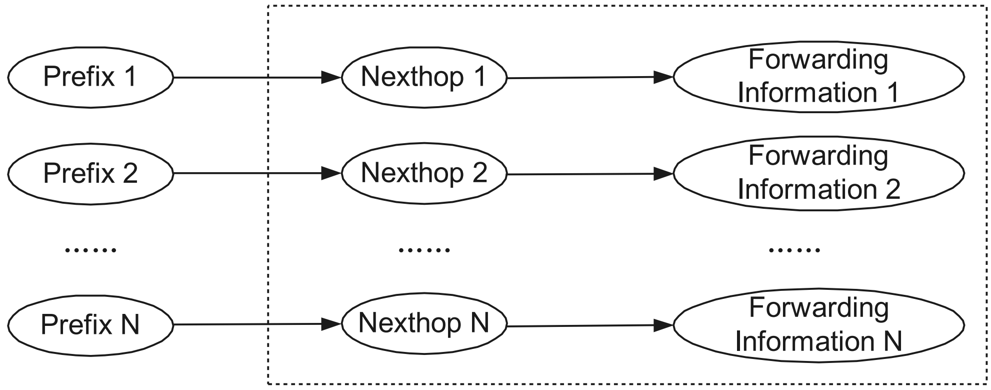
迭代策略用于控制下一跳的迭代结果，以满足不同应用场景下的需求。迭代路由时不需要通过策略进行控制，只遵循最长匹配原则。只有私网路由迭代隧道时需要应用隧道策略。

缺省情况下，系统为 VPN 选择 LSP 隧道且不进行负载分担。如果要进行负载分担或选择其它类型的隧道，需要配置隧道策略并应用该隧道策略。VPN 应用隧道策略后，对其下一跳进行迭代时，选择隧道策略中绑定的隧道，或根据隧道策略中为不同类型的隧道指定的优先级，选择使用的隧道。

下一跳分离下刷

从转发层面看，对于公网路由，只需要下一跳和出接口就能指导转发，对于私网路由，还需要指定公网隧道，在采用下一跳分离下刷之前，转发信息（下一跳、出接口、隧道）需要由前缀直接带到 FIB，导致总的收敛时间与前缀数量相关，采用下一跳分离下刷后，大量的前缀与同一个下一跳对应，只需要通过这个下一跳将转发信息带到 FIB 中，相关前缀的流量就能同时切换。

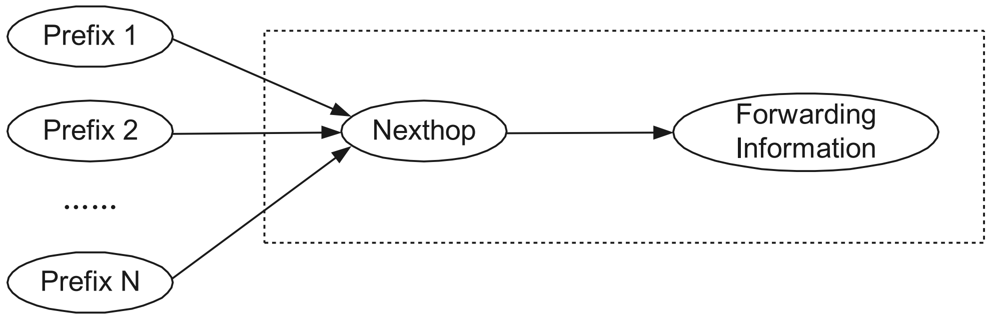
图 1-5 没有实现下一跳分离时路由的组织形式



如图 1-5 所示，没有实现下一跳分离时，前缀之间完全独立，分别对应自己的下一跳和迭代转发信息。当依赖路径变化时，分别对每个前缀对应的下一跳进行迭代，并以前缀为单位下刷，收敛速度与前缀数量相关。

实际上，如果前缀来自同一个 BGP 邻居，则它们的下一跳和迭代转发信息也一定相同，下刷的迭代转发信息是重复的。

图 1-6 实现下一跳分离时路由的组织形式



如图 1-6 所示，实现下一跳分离后，来自同一个邻居的前缀共享同一个下一跳，依赖路径变化时，只对这个共同的下一跳进行迭代，并以下一跳为单位刷新其对应的迭代转发信息。这样所有前缀的流量就可以同时收敛，收敛速度与前缀数量无关。

迭代路由与迭代隧道的比较

迭代路由与迭代隧道的不同如下表所示。

表 1-4 迭代路由与迭代隧道的比较

迭代类型	特点
迭代路由	● 对 BGP 公网路由进行迭代。 ● 由路由变化触发迭代。 ● 支持按策略迭代下一跳。
迭代隧道	● 对 BGP 私网路由进行迭代。 ● 由隧道或隧道策略变化触发迭代。 ● 可通过隧道策略对迭代行为进行控制，满足不同应用场景下的需求。

1.4.11 缺省路由

缺省路由是另外一种特殊的路由。通常情况下，管理员可以通过手工方式配置缺省静态路由；但有些时候，也可以使动态路由协议生成缺省路由，如 OSPF 和 IS-IS。

简单来说，缺省路由是没有在路由表中找到匹配的路由表项时才使用的路由。在路由表中，缺省路由以到网络 0.0.0.0（掩码也为 0.0.0.0）的路由形式出现。可通过命令 **display ip routing-table** 查看当前是否设置了缺省路由。

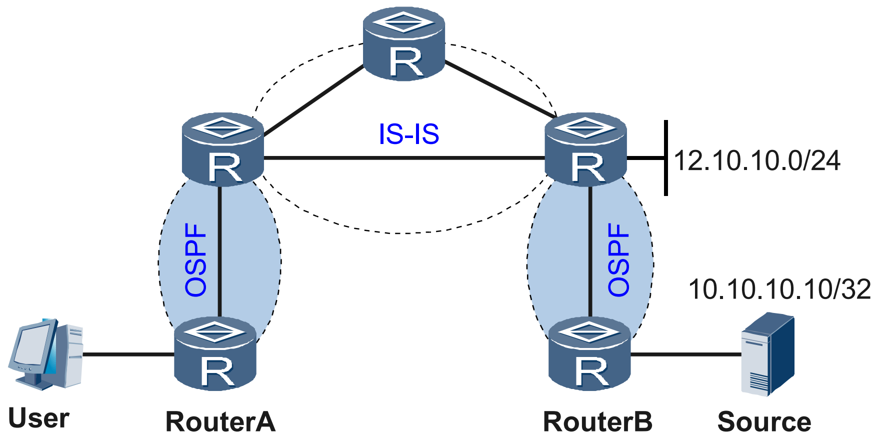
如果报文的目的地址不能与路由表的任何目的地址相匹配，那么该报文将选取缺省路由。如果没有缺省路由且报文的目的地址不在路由表中，那么该报文将被丢弃，并向源端返回一个 ICMP（Internet Control Message Protocol）报文，报告该目的地址或网络不可达。

1.5 应用

1.5.1 路由按优先级收敛典型应用

如图 **1-7** 所示，在图中的组播服务中，网络上运行 IGP 协议，接收者在 RouterA 端，组播源服务器 10.10.10.10/32 在 RouterB 端。其中要求到组播服务器的路由优先于其他路由（例如 12.10.10.0/24）收敛。这时可以配置路由 10.10.10.10/32 的收敛优先级高于路由 12.10.10.0/24 的收敛优先级，这样当网络路由重新收敛时，就能确保到组播源的路由 10.10.10.10/32 优先收敛，保证组播业务的转发。

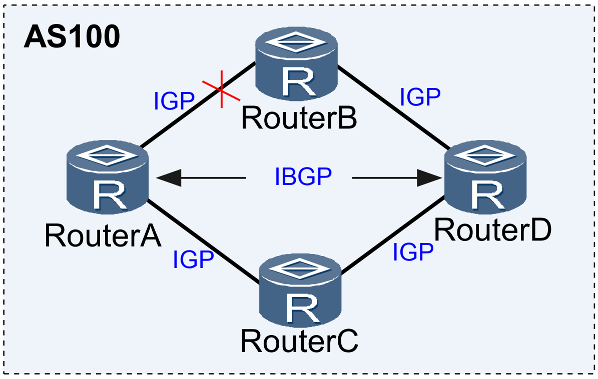
图 1-7 路由按优先级收敛应用组网图



1.5.2 下一跳分离的典型应用

IBGP 公网路由迭代

图 1-8 IBGP 路由迭代组网图



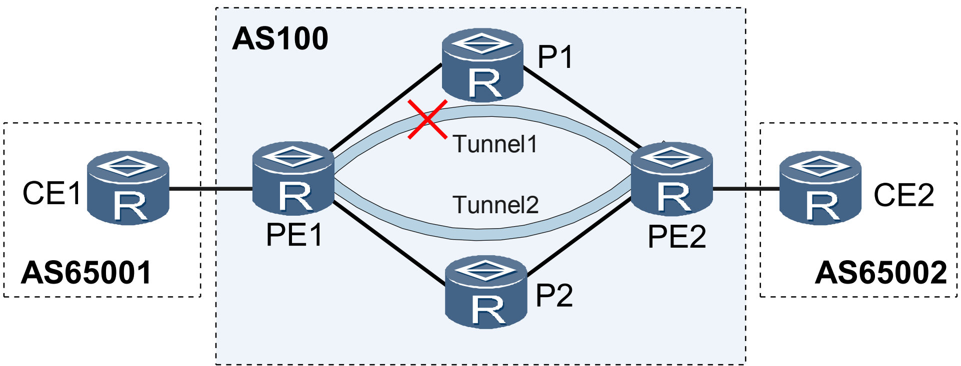
如图 1-8 所示，RouterA 和 RouterD 之间建立 IBGP 邻居。由于 IBGP 路由的下一跳是非直连可达的（一般 IBGP 邻居是通过两个设备的 Loopback 接口地址建立），不能直接用来指导转发，必须根据 IBGP 路由的原始下一跳信息找到真正的出接口和直连下一跳，才能下发 FIB，指导报文转发。

RouterD 从 RouterA 收到 4 千条路由，这些路由具有相同的 BGP 原始下一跳，迭代到相同的 IGP 路径（A-B-D）。当路径（A-B-D）发生故障时，这 4 千条 IBGP 路由不需要分别迭代并刷新 FIB，仅仅需要对共同的下一跳进行一次迭代和刷新，就可以在转发层面上实现 4 千条 IBGP 路由收敛到路径（A-C-D）。收敛时间仅与下一跳数量有关，实现了和前缀数量无关的亚秒级收敛。

如果 RouterA 和 RouterD 之间建立的是多跳 EBGP 邻居，收敛过程与上述过程相同，下一跳分离同样适用于多跳 EBGP 路由迭代的场景。

VPN 路由迭代隧道

图 1-9 VPN 路由迭代组网图



如图 1-9 所示，PE1 和 PE2 之间建立邻居，PE2 从 PE1 收到 4 千条 VPN 路由，这些路由具有相同的 BGP 原始下一跳，迭代到相同的公网隧道（Tunnel1），当 Tunnel1 发生故障时，这 4 千条路由不需要分别迭代并刷新 FIB，仅仅需要对共同的下一跳进行一次迭代和刷新，就可以在转发层面上实现 4 千条私网路由收敛到 Tunnel2。这样收敛时间仅与下一跳数量有关，实现了和前缀数量无关的亚秒级收敛。

1.6 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
ARP	Address Resolution Protocol	地址解析协议
BGP	Border Gateway Protocol	边界网关协议
CE	Customer Edge	用户边缘设备
FIB	Forwarding Information Base	转发信息库
IBGP	Internal Border Gateway Protocol	内部边界网关协议
IGP	Interior Gateway Protocol	内部网关协议
IS-IS	Intermedia System-Intermedia System	中间系统－中间系统
NDP	Neighbor Discover Protocol	邻居发现协议
OSPF	Open Shortest Path First	开放最短路径优先
PE	Provider Edge	服务商边缘设备
RIP	Routing Information Protocol	选路信息协议
RM	Route Management	路由管理
Vlink	Virual Link	虚拟链接
VoIP	Voice Over IP	IP 语音
VPN	Virtual Private Network	虚拟私有网络
VRP	Versatile Routing Platform	通用路由平台

2 静态路由

关于本章

2.1 介绍

定义

静态路由是一种需要管理员手工配置的特殊路由。

目的

当网络结构比较简单时，只需配置静态路由就可以使网络正常工作。仔细设置和使用静态路由可以改进网络的性能，并可为重要的应用保证带宽。

但是，当网络发生故障或者拓扑发生变化后，静态路由不会自动改变，必须有管理员的介入。

Huawei AR1200 系列支持普通静态路由，也支持与 VPN 实例关联的静态路由，后者主要用于 VPN 路由的管理。有关 VPN 实例请参见《Huawei AR1200 系列 企业路由器 特性描述-VPN》。

2.2 参考标准和协议

无

2.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R001C00

2.4 原理描述

2.4.1 静态路由的组成

在 AR1200 中，使用 **ip route-static** 命令配置静态路由，一条静态路由包含以下要素：

- 目的地址与掩码
- 出接口与下一跳地址

目的地址与掩码

在 **ip route-static** 命令中，IPv4 地址为点分十进制格式，掩码可以用点分十进制表示，也可用掩码长度（即掩码中连续‘1’的位数）表示。

出接口与下一跳地址

在配置静态路由时，可指定出接口 *interface-type interface-number*，也可指定下一跳地址 *nexthop-address*，还可以同时指定出接口和下一跳地址，具体要根据实际需要来定。

实际上，所有的路由项都必须明确下一跳地址。在发送报文时，首先根据报文的目的地址寻找路由表中与之匹配的路由（遵循最长匹配原则）。只有指定了下一跳地址，链路层才能找到对应的链路层地址，并转发报文。

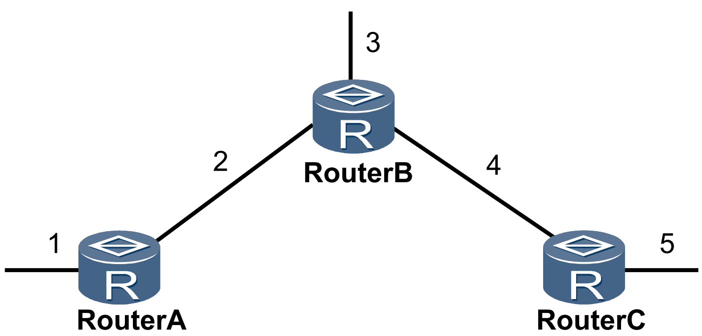
指定发送接口时需要注意：

- 对于点到点类型的接口，指定发送接口即隐含指定了下一跳地址，这时认为与该接口相连的对端接口地址就是路由的下一跳地址。
- 对于 NBMA（Non Broadcast Multiple Access）类型的接口（如 ATM 接口），它支持点到多点网络，这时除了配置 IP 路由外，还需在链路层建立 IP 地址到链路层地址的映射。这种情况下应配置下一跳 IP 地址。
- 在配置静态路由时，如果指定以广播口（如以太网接口）和 VT（Virtual-template）接口作为出接口，必须指定下一跳。因为以太网接口是广播类型的接口，而 VT 接口下可以关联多个虚拟访问接口（Virtual Access Interface），这都会导致出现多个下一跳，无法唯一确定下一跳。在应用中，如果必须指定广播接口（如以太网接口）或 VT 接口作为出接口，建议同时指定通过该接口发送时对应的下一跳地址。

2.4.2 静态路由的应用

如图 **2-1** 所示，该网络结构比较简单，可以使用静态路由实现网络互通。首先需要确定每个物理网络的地址，并为每个路由器标识出非直连的物理网络，最后为每个非直连的物理网络配置静态路由命令。

图 2-1 静态路由组网图



此例中需要在 RouterA 上配置到网络 3、4、5 的静态路由，在 RouterB 上配置到网络 1、5 的静态路由，在 RouterC 上配置到网络 1、2、3 的静态路由。

缺省静态路由

在使用 **ip route-static** 配置静态路由时，如果将目的地址与掩码配置为全零（0.0.0.0 0.0.0.0），则表示配置的是缺省路由。这样可以简化网络的配置。

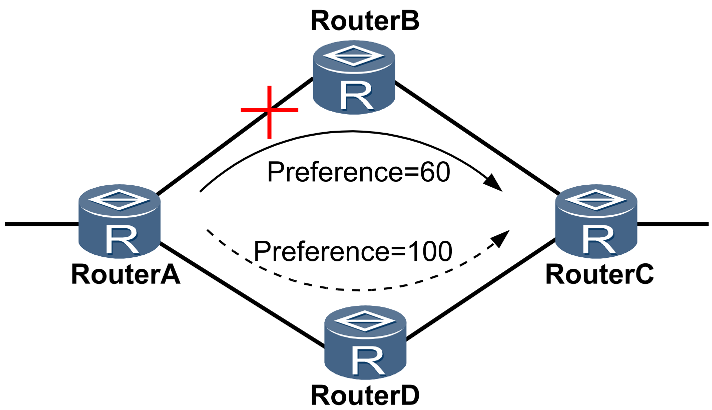
图 **2-1** 中，因为 RouterA 发往 3、4、5 网络的报文下一跳都是 RouterB，因此可在 RouterA 上配置一条缺省路由，代替上个例子中通往 3、4、5 网络的 3 条静态路由。同理，RouterC 也只需要配置一条到 RouterB 的缺省路由，代替上个例子中通往 1、2、3 网络的 3 条静态路由。

浮动静态路由

对于不同的静态路由，可以为它们配置不同的优先级 **preference**，从而更灵活地应用路由管理策略。配置到达相同目的地的多条路由，如果指定不同优先级，则可实现路由备份。图 **2-2** 是浮动静态路由组网图。

如图 **2-2**，从 RouterA 到 RouterC 有两条静态路由。在正常情况下，路由表上仅下一跳是 RouterB 的静态路由的状态为“Active”，因为这条路由具有更高的优先级。另一条下一跳是 RouterD 的静态路由则作为备份路由，只有在主链路上出现故障的时候，备份路由才会被激活，承担数据转发的业务。在主链路恢复正常后，路由表又恢复到原来的样子，即下一跳是 RouterB 的静态路由又成为活跃路由来承担数据转发，因此这条备份路由也叫做浮动静态路由。RouterB 和 RouterC 之间链路发生故障时，浮动静态路由就无能为力了。

图 2-2 浮动静态路由

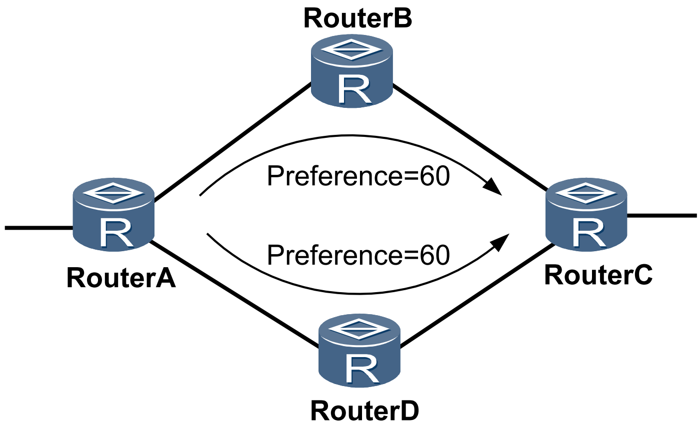


静态路由负载分担

配置到达相同目的地的多条路由，如果指定相同优先级，则可实现负载分担。

如图 **2-3**，从 RouterA 到 RouterC 有两条优先级相同的静态路由。两条路由都会出现在路由表上，同时进行数据的转发。

图 2-3 静态路由负载分担



2.4.3 静态路由特性

IPv4 静态路由

AR1200 支持普通静态路由，也支持与 VPN 实例关联的静态路由，后者主要用于 VPN 路由的管理。有关 VPN 实例请参见《Huawei AR1200 系列 企业路由器 特性描述-VPN》。

2.4.4 BFD for 静态路由

与动态路由协议不同，静态路由自身没有检测机制，当网络发生故障的时候，需要管理员介入。BFD for 静态路由特性可为静态路由绑定 BFD 会话，利用 BFD 会话来检测静态路由所在链路的状态。

BFD for 静态路由可为每条静态路由绑定一个 BFD 会话。

- 当某条静态路由上的 BFD 会话检测到故障（BFD 会话由 Up 转为 Down），BFD 会将故障上报系统，系统将这条路由从 IP 路由表中删除。
- 当某条静态路由上的 BFD 会话成功建立（BFD 会话由 Down 转为 Up），BFD 会上报系统，系统将这条路由加入 IP 路由表。

BFD for 静态路由有单跳检侧和多跳检测两种方式。

- 单跳检测

  对于非迭代的静态路由，所配置的出接口和下一跳就是直连下一跳信息。这样，BFD 会话的出接口即静态路由的出接口，对端地址即路由的下一跳。
- 多跳检测

  对于迭代的静态路由，仅配置了下一跳，需要迭代出直连下一跳和出接口。这样，BFD 会话的对端地址为路由的原始下一跳，出接口则不限。一般情况下，迭代的原始下一跳是多跳的，非直接可达，故支持迭代的静态路由进行多跳检测。

说明

如果路由的下一跳信息不是直接可达的，那么该路由就不能用来指导转发，系统会根据下一跳信息计算出一个实际的出接口和下一跳，这个过程就叫做路由迭代。另外，使用 **display ip routing-table** 命令查看路由时，如果 **Flags** 字段显示为 R，则表示路由为迭代路由，否则，为非迭代路由。

说明

BFD 的详细介绍请参见《Huawei AR1200 系列 企业路由器 特性描述-可靠性》。

2.4.5 NQA for 静态路由

静态路由本身并没有检测机制，如果链路发生了故障，静态路由不会自动改变（不会从 IP 路由表中自动删除），需要管理员介入，这就无法保证及时进行链路切换，可能造成较长时间的业务中断。

基于以上原因，需要有一种有效的方案来检测静态路由所在的链路。对于静态路由而言，现有的 BFD for 静态路由特性，由于受到互通设备两端都必须支持 BFD 的限制，在某些应用场景无法实施。而 NQA（Network Quality Analysis） for 静态路由则只要求互通设备的其中一端支持 NQA 即可，并不要求两端都支持，且不受二层设备的限制。

NQA for 静态路由特性即为静态路由绑定 NQA 测试例，利用 NQA 测试例来检测静态路由所在链路的状态，根据 NQA 的检测结果，决定静态路由是否活跃，达到避免通信的中断或服务质量降低的目的。NQA for 静态路由特性的功能如下：

- 如果 NQA 测试例检测到链路故障，路由器将这条静态路由设置为“非激活”状态（此条路由不可用，从 IP 路由表中删除）
- 如果 NQA 测试例检测到链路恢复正常，路由器将这条静态路由设置为“激活”状态（此条路由可用，添加到 IP 路由表）

说明

每条静态路由只可以绑定一个 NQA 测试例。

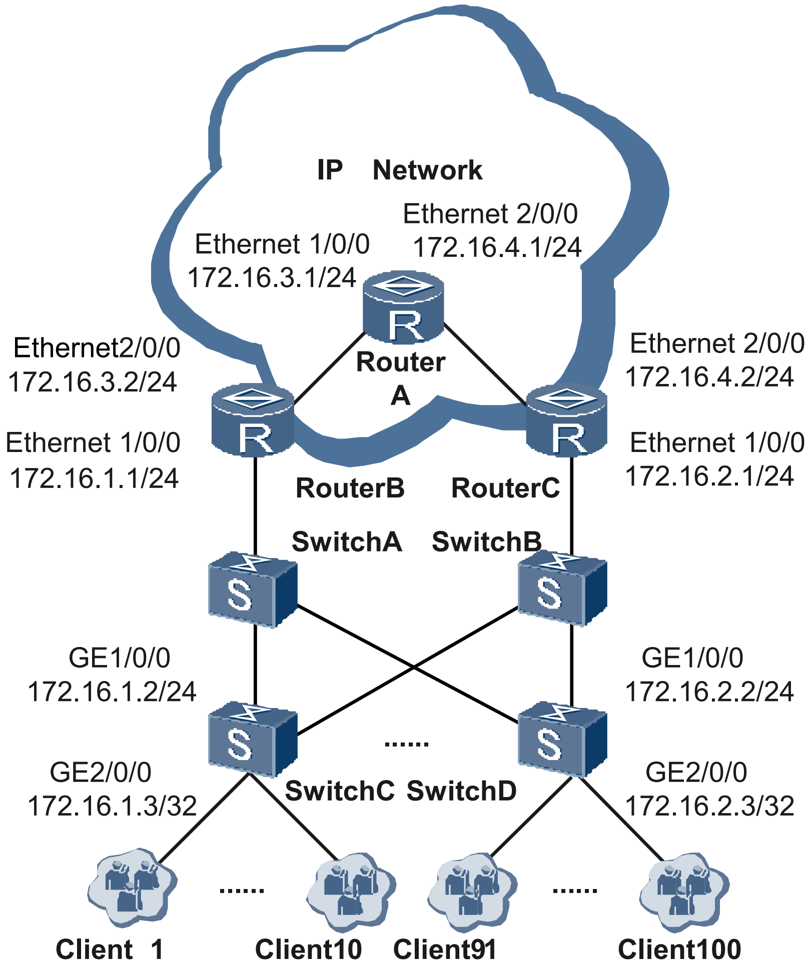
NQA 的详细介绍请参见《Huawei AR1200 系列 企业路由器 特性描述-系统管理》。

应用

如图 2-4 所示，每台接入交换机下连接 10 个用户，共 100 个用户。由于在 RouterB 和用户之间无法使用动态路由协议，所以在 RouterB 上配置到用户的静态路由。出于网络稳定性的考虑，在 RouterC 上进行同样的配置，作为冗余备份。RouterA、RouterB 和 RouterC 上运行动态路由协议，相互间可以学习路由。其中，RouterB 和 RouterC 配置动态路由协议引入静态路由，并且设置不同的花费值，这样 RouterA 也能通过动态路由协议从 RouterB 和 RouterC 分别学习到用户的路由，RouterA 根据两条链路的花费值不同选择一条主用链路，另一条链路做为备份。

在 RouterB 上配置 NQA for 静态路由特性，利用 NQA 测试例检测主用链路 RouterB→SwitchA→SwitchC（SwitchD）的状态，当主用链路发生故障时，撤销静态路由发布，使下行流量走无故障的链路 RouterC→SwitchB→SwitchC（SwitchD）。在两条链路都正常时，控制下行流量优先走主用链路。

图 2-4 NQA for 静态路由应用组网图



2.4.6 静态路由永久发布

链路有效性直接影响网络的稳定性和可用性，因此链路状态的检测对网络维护具有重要意义。BFD 作为一种常用方案，并不适合所有的场景。例如，在不同的 ISP 之间，客户更希望采用更简单、更自然的方式来达到这一目的。

静态路由永久发布可以为客户提供一种低成本、部署简单的链路检测机制，并提高与其它厂商设备的兼容性。在客户希望确定业务流量的转发路径，不希望流量从其它路径穿越时，静态路由永久发布可以通过 Ping 静态路由目的地址的方式来检测链路的有效性而达到业务监控的目的。

配置永久发布属性后，之前无法发布的静态路由仍然被优选并添加到路由表中。具体可以分为以下两种情况：

- 静态路由配置出接口且出接口的 IP 地址存在时，无论接口状态是 Up 或 Down，只要配置了永久发布属性，静态路由都会被优选并添加到路由表。
- 静态路由没有配置出接口时，无论静态路由是否能迭代到出接口，只要配置了永久发布属性，路由都会被优选并添加到路由表中。

这样，通过控制静态路由的优先级和前缀长度，使 Ping 数据包始终通过静态路由转发，就可以检测出链路的有效性。

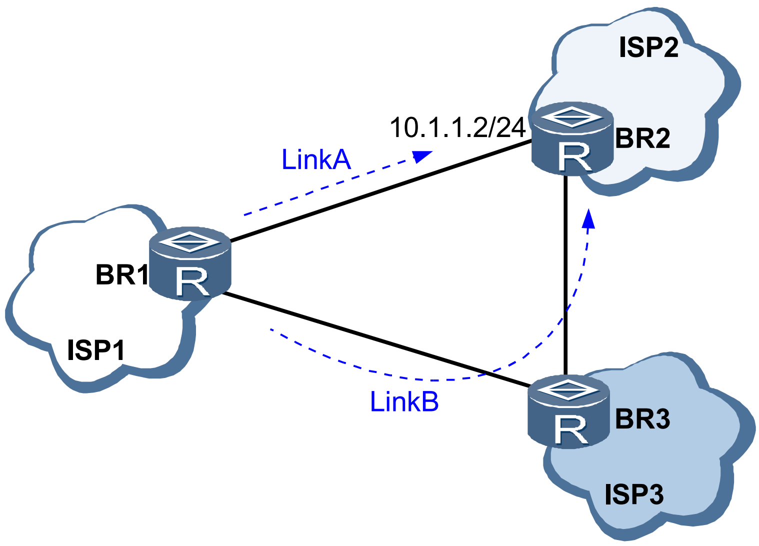
说明

该特性不判断路由是否可达而一直将静态路由保留在 IP 路由表中，如果实际路径不可达，静态路由可能形成黑洞路由。

应用

如图 2-5 所示，BR1、BR2 和 BR3 分别属于 ISP1、ISP2 和 ISP3。从 BR1 到 BR2 有两条链路（LinkA 和 LinkB）可达，但 ISP1 希望业务流量都通过 LinkA 直接转发到 ISP2，而不从 ISP3 穿越。

图 2-5 静态路由永久发布应用组网图



在 BR1 和 BR2 之间建立直连单跳 EBGP 邻居，同时为了进行业务状态监控，在 BR1 上配置到对端（BR2）BGP 邻居地址（10.1.1.2/24）的静态路由（出接口为与 BR2 直连的本地接口），并使能路由永久发布。网络监控系统周期性的 Ping 10.1.1.2，可通过 Ping 结果来判断 LinkA 的状态，进而间接的监控 BGP 业务状态。

当 LinkA 正常时，Ping 数据包都是通过 LinkA 进行转发。如果 LinkA 发生故障，即使能通过 LinkB 到达 BR2，但由于静态路由优先级较高，却仍然有效，Ping 数据包还是通过 LinkA 进行转发，但此时转发不通。对于 BGP 数据包也是相同的情况，故障会导致 BGP 邻居断开，监控系统可以通过 Ping 结果间接的检测到业务问题，并通知维护人员及时响应。

2.5 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
BFD	Bidirectional Forwarding Detection	双向转发检测
FIB	Forwarding Information Base	转发信息库
VRP	Versatile Routing Platform	通用路由平台
RM	Route Management	路由管理

3 RIP

关于本章

3.1 介绍

定义

RIP 是 Routing Information Protocol（路由信息协议）的简称。它是一种较为简单的内部网关协议 IGP（Interior Gateway Protocol），主要应用于规模较小的网络中，例如校园网以及结构较简单的地区性网络。对于更为复杂的环境和大型网络，一般不使用 RIP 协议。

RIP 是一种基于距离矢量（Distance-Vector）算法的协议，它通过 UDP 报文进行路由信息的交换，使用的端口号为 520。

RIP 使用跳数（Hop Count）来衡量到达目的地址的距离，称为度量值。在 RIP 中，缺省情况下，设备到与它直接相连网络的跳数为 0，通过一个设备可达的网络的跳数为 1，其余依此类推。也就是说，度量值等于从本网络到达目的网络间的设备数量。为限制收敛时间，RIP 规定度量值取 0 ～ 15 之间的整数，大于或等于 16 的跳数被定义为无穷大，即目的网络或主机不可达。由于这个限制，使得 RIP 不可能在大型网络中得到应用。

为提高性能，防止产生路由循环，RIP 支持水平分割（Split Horizon）和毒性逆转（Poison Reverse）功能。

目的

RIP 协议是最早的内部网关协议之一，RIP 协议被设计用于使用同种技术的中小型网络。由于 RIP 的实现较为简单，在配置和维护管理方面也远比 OSPF 和 IS-IS 容易，因此在实际组网中仍有广泛的应用。

3.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC1058	This document describes RIP protocol, describes the elements, characteristic, limitation of rip version 1.	-
RFC2453	This document 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.	-

3.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R001C00

3.4 原理描述

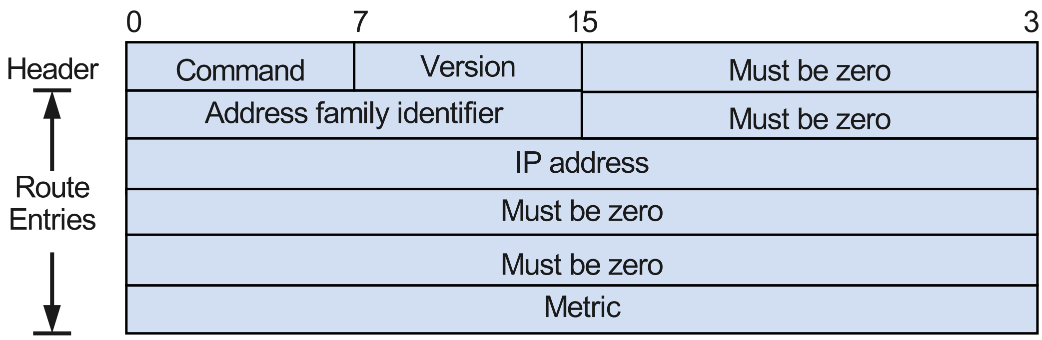
RIP 协议是一种距离矢量路由协议，报文转发基于 UDP 协议，RIP 协议使用三个定时器保证路由信息的发布，更新和老化。由于协议本身的设计缺陷可能会产生环路，所以 RIP 增加了水平分割、毒性逆转和触发更新的特性，最大限度避免环路的产生。

另外，由于 RIP 定时向外通告自己的路由表，所以增加 RIP 路由聚合特性来缩减路由表的规模。

3.4.1 RIP-1

RIP-1（即 RIP version1）是有类别路由协议（Classful Routing Protocol），它只支持以广播方式发布协议报文，报文格式如图 3-1 所示。在一个 RIP 报文中，最多可以有 25 个路由表项。RIP 是一个基于 UDP 协议的，并且 RIP-1 的数据包不能超过 512 字节。RIP-1 的协议报文中没有携带掩码信息，它只能识别 A、B、C 类这样的自然网段的路由，因此 RIP-1 无法支持路由聚合，也不支持不连续子网（Discontiguous Subnet）。

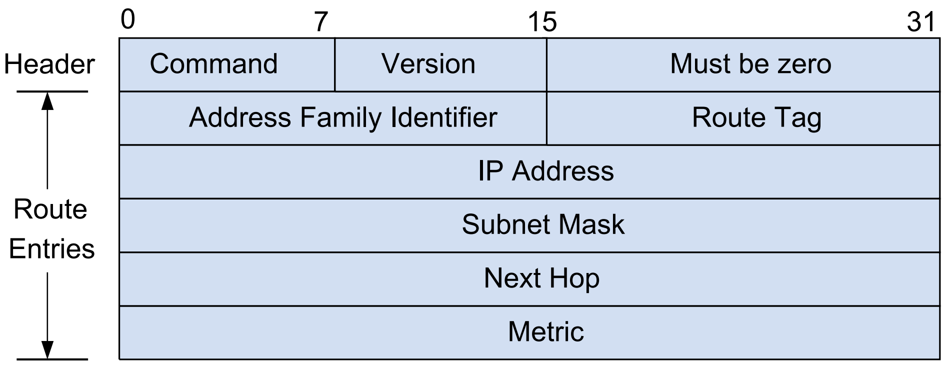
图 3-1 RIP-1 **的报文格式**



3.4.2 RIP-2

RIP-2（即 RIP version2）是一种无分类路由协议（Classless Routing Protocol），报文格式如图 3-2 所示。

图 3-2 RIP-2 的报文格式



与 RIP-1 相比，RIP-2 有以下优势：

- 支持外部路由标记（Route Tag），可以在路由策略中根据 Tag 对路由进行灵活的控制。
- 报文中携带掩码信息，支持路由聚合和 CIDR（Classless Inter-Domain Routing）。
- 支持指定下一跳，在广播网上可以选择到最优下一跳地址。
- 支持组播路由发送更新报文，只有支持 RIP-2 的设备才能收到协议报文，减少资源消耗。
- 支持对协议报文进行验证，并提供明文验证和 MD5 验证两种方式，增强安全性。

3.4.3 定时器

RIP 主要使用三个定时器：

- 更新定时器（Update timer）：它定时触发更新报文的发送，更新周期默认为 30 秒。
- 老化定时器（Age timer）：RIP 设备如果在老化时间内没有收到邻居发来的路由更新报文，则认为该路由不可达。
- 垃圾收集定时器：如果在垃圾收集时间内不可达路由没有收到来自同一邻居的更新，则该路由将被从路由表中彻底删除。

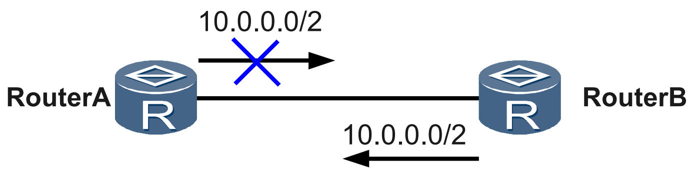
三个定时器之间的关系：

RIP 的更新信息发布是由 Update 定时器控制的，默认为每 30 秒发送一次。每一条路由表项对应两个定时器：老化定时器和垃圾收集定时器。当学到一条路由并添加到路由表中时，老化定时器启动。如果在默认 180 秒后没有收到邻居发来的更新报文，则把该路由的度量值置为 16（表示路由不可达），并启动垃圾收集定时器，如果在默认 120 秒内仍然没有收到更新报文，垃圾收集定时器超时后在路由中删除该表项。

3.4.4 水平分割

水平分割（Split Horizon）指的是 RIP 从某个接口学到的路由，不会从该接口再发回给邻居设备。这样不但减少了带宽消耗，还可以防止路由环路。

图 3-3 水平分割原理图



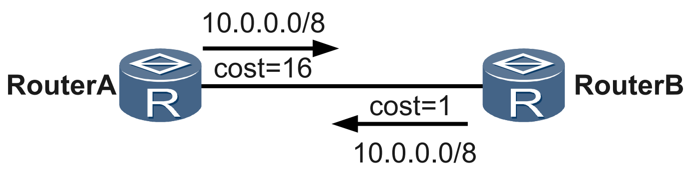
如图 3-3 所示，RouterB 目的地址是 10.0.0.0 的路由信息通告给 RouterA 后，RouterA 不会再把到网络 10.0.0.0 的路由发回给 RouterB。

3.4.5 毒性逆转

毒性逆转（Poison Reverse）指的是 RIP 从某个接口学到路由后，将该路由的开销设置为 16（即指明该路由不可达），并从原接口发回邻居设备。利用这种方式，可以清除对方路由表中的无用路由。

RIP 毒性逆转可以防止产生路由环路。

图 3-4 毒性逆转原理图



如图 3-4 所示，在不配置水平分割的情况下，RouterB 会向 RouterA 发送从 RouterA 学到的路由，并且 RouterA 到网络 10.0.0.0 的路由开销值为 1。如果 RouterA 到网络 10.0.0.0 的路由变成不可达，同时 RouterB 没有收到 RouterA 的更新报文，而继续向 RouterA 发送到达网络 10.0.0.0 路由信息，则会导致路由环路。

如果 RouterA 在接收到从 RouterB 发来的路由后，向 RouterB 发送一个这条路由不可达的消息，这样 RouterB 就不会再从 RouterA 学到这条可达路由，因此就可以避免上述环路的发生。

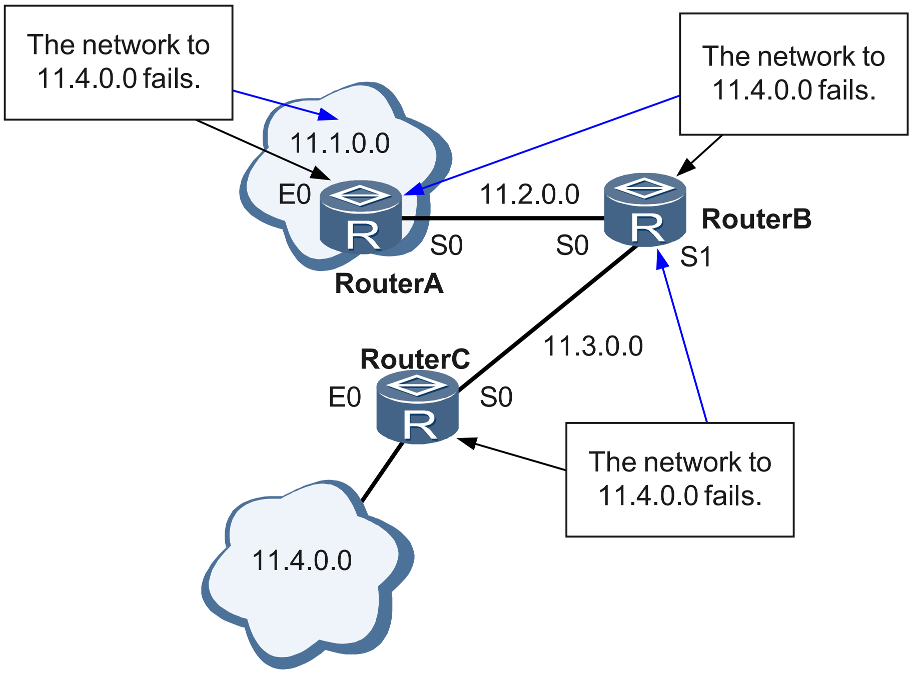
如果毒性逆转和水平分割都配置了，简单的水平分割行为（从某接口学到的路由再从这个接口发布时将被抑制）会被毒性逆转行为代替。

3.4.6 触发更新

触发更新的原理是，路由信息发生变化时，立即向邻居设备发送触发更新报文，通知变化的路由信息。

触发更新缩短了收敛时间，触发更新可以缩短网络收敛时间，在路由表项变化时立即向其他设备广播该信息，而不必等待定时更新。

图 3-5 触发更新原理图



如图 3-5 所示，网络 11.4.0.0 不可达时，RouterC 最先得到这一信息。通常，更新路由信息会定时发送给相邻 Router（RIP 协议每隔 30 秒发送一次）。但如果在 RouterC 等待更新周期到来的时候，RouterB 的更新报文传到了 RouterC，RouterC 就会学到 RouterB 的去往网络 11.4.0.0 的错误路由。这样 RouterB 和 RouterC 上去往网络 11.4.0.0 的路由都指向对方从而形成路由环路。如果 RouterC 发现网络故障之后，不再等待更新周期到来，就立即发送路由更新信息给路由器 B，使路由器 B 的路由表及时更新，则可以避免产生上述问题。

触发更新还存在另外一种方式：当下一跳不可用之后（如因为链路故障）需要及时通告给其它设备，此时要把该路由的 cost 设置为 16 然后发布出去，此更新也叫做路由毒杀。

3.4.7 路由聚合

路由聚合的原理是，同一个自然网段内的不同子网的路由在向外（其它网段）发送时聚合成一个网段的路由发送。RIP-1 的协议报文中没有携带掩码信息，故 RIP-1 发布的就是自然掩码的路由。RIP-2 支持路由聚合，因为 RIP-2 报文携带掩码位，所以支持子网划分。

在 RIP-2 中进行路由聚合可提高大型网络的可扩展性和效率，缩减路由表。

路由聚合有两种方式：

- 基于 RIP 进程的有类聚合：

  聚合后的路由使用自然掩码的路由形式发布，但在配置水平分割或毒性逆转的情况下，有类聚合将失效，这是因为水平分割或毒性逆转将抑制一些路由的发布，配置了有类聚合时一条聚合路由可能是聚合了从不同的接口上学到的路由，这样在向外发布时就会产生冲突。

比如，对于 10.1.1.0/24（metric=2）和 10.1.2.0/24（metric=3）这两条路由，会聚合成自然网段路由 10.0.0.0/8（metric=2）。RIP Version2 聚合是按类聚和的，聚合得到最优的 metric 值。

- 基于接口的聚合：

  用户可以指定聚合地址。

  比如，对于 10.1.1.0/24（metric=2）和 10.1.2.0/24（metric=3）这两条路由，可以在此接口上配置聚合路由 10.1.0.0/16（metric=2）。

3.4.8 多进程和多实例

为了方便管理，提高控制效率，RIP 支持多进程和多实例特性。多进程允许为一个指定的 RIP 进程关联一组接口，从而保证该进程进行的所有协议操作都仅限于这一组接口。这样，就可以实现一台设备有多个 RIP 协议进程，每个进程负责唯一的一组接口。而且每个 RIP 进程的路由数据也是相互独立的，但进程之间可以相互引入路由。

对于支持 VPN 的设备，每个 RIP 进程都与一个指定的 VPN 实例相关联。这样，所有附加到该进程的接口都应与该进程相关联的 VPN 实例相关联。

3.4.9 BFD for RIP

定义

双向转发检测 BFD（Bidirectional Forwarding Detection）是一种用于检测邻居路由器之间链路故障的检测机制，它通常与路由协议联动，通过快速感知链路故障并通告使得路由协议能够快速地重新收敛，从而减少由于拓扑变化导致的流量丢失。

在 BFD for RIP 中，BFD 可以快速检测到链路故障并通知 RIP 协议，从而加快 RIP 协议对于网络拓扑变化的响应。

BFD 包括静态 BFD 和动态 BFD。

- 静态 BFD

  静态 BFD 是指通过命令行手工配置 BFD 会话参数，包括了配置本地标识符和远端标识符等，手工下发 BFD 会话建立请求。

  这种方式可以不受对端设备的限制，在对端设备不支持 BFD 功能的情况下，本端通过静态 BFD 实现单臂 BFD 检测功能；也可以用来配置在指定链路上，实现普通 BFD 检测功能。

- 动态 BFD

  动态 BFD 是指由路由协议动态触发 BFD 会话建立。动态 BFD 中，本地标识符是动态分配的，远端标识符从对端的 BFD 报文中获取。

  路由协议在建立了新的邻居关系时，将对应的参数及检测参数（包括目的地址、源地址等）通告给 BFD，BFD 根据收到的参数建立起会话。当发生链路故障是，联动了 BFD 的路由协议可以快速感知到 BFD 会话状态变为 Down，从而实现将流量快速切换到备份路径，避免了数据大量丢失。

  动态 BFD 比静态 BFD 更具有灵活性。

目的

网络上的链路故障会导致路由器重新计算路由，因此缩短路由协议的收敛时间对于提高网络性能是非常重要的。加快故障感知速度并快速通告给路由协议是一种可行的方案。

BFD 和 RIP 相关联后，一旦链路发生故障，BFD 在毫秒级时间内感知该故障并通知 RIP 协议，然后路由器在路由表中删除掉故障链路的路由并快速启用备份路径，提高了路由协议的收敛速度。

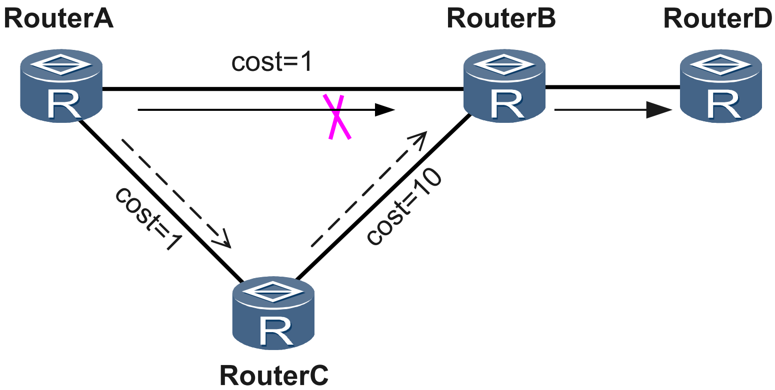
RIP 协议在使用 BFD 前后的链路故障检测机制及收敛速度如表 3-1 所示：

表 3-1 是否使用 BFD 检测的区别

是否配置 BFD 检测	链路故障检测机制	收敛速度
否	RIP 老化定时器超时（默认配置 180s）	秒级(>180s)
是	BFD 会话状态为 Down	秒级(<30s)

原理

图 3-6 BFD for RIP 组网图



- 动态 BFD for RIP 的原理：
  1. RouterA 、RouterB 及 RouterC 三台设备两两之间以及 RouterB 和 RouterD 之间已经分别建立 RIP 邻居。
  2. 在 RouterA 及 RouterB 上使能动态 RIP BFD 检测机制。
  3. 经过路由计算，RouterA 到达 RouterD 的路由下一跳为 RouterB。
  4. 当 RouterA 和 RouterB 之间的链路出现故障时，BFD 快速感知并通知给 RouterA，RouterA 删除掉下一跳为 RouterB 的路由。
  5. RouterA 重新进行路由计算并选取新的路径，经过 RouterC、RouterB 到达 RouterD。
  6. 当 RouterA 与 RouterB 之间的链路恢复之后，二者之间的会话重新建立，RouterA 收到 RouterB 的路由信息，重新选择最优路径进行报文转发。
- 静态 BFD for RIP 的原理：
  1. RouterA、RouterB 及 RouterC 三台设备两两之间以及 RouterB 和 RouterD 之间已经分别建立 RIP 邻居。

2. 在 RouterA 与 RouterB 相连的接口上使能静态 BFD 检测机制。
3. 当 RouterA 和 RouterB 之间的链路出现故障时，BFD 快速感知并通知给 RouterA，RouterA 删除掉下一跳为 RouterB 的路由。
4. 当 RouterA 与 RouterB 之间的链路恢复之后，二者之间的会话重新建立，RouterA 收到 RouterB 的路由信息，重新选择最优路径进行报文转发。

3.5 术语与缩略语

术语

术语	解释
毒性逆转	RIP 从某个接口学到路由后，将该路由的开销设置为 16（不可达），并从原接口发回邻居设备。
水平分割	RIP 从某个接口学到的路由，不会从该接口再发回给该邻居设备。

缩略语

缩略语	英文全称	中文全称
RIP	Routing Information Protocol	路由信息协议

4 RIPng

关于本章

4.1 介绍

定义

RIPng（RIP next generation，下一代 RIP 协议）是对原来的 IPv4 网络中 RIP version 2 协议在 IPv6 网络上的扩展，大多数 RIP 的概念都可以应用于 RIPng。

RIPng 协议是基于 D-V（Distance Vector，距离矢量）算法的路由协议，用跳数来衡量到达目的主机的距离（也称为度量值或开销）。在 RIPng 协议中，从一个路由器到其直连网络的跳数为 0，到通过另一台路由器可达的网络的跳数为 1，如此类推，当跳数大于或等于 16 时，目的网络或主机就被定义为不可达。

为了在 IPv6 网络中应用，RIPng 对原有的 RIP 协议进行了修改：

- UDP 端口号：使用 UDP 的 521 端口（RIP 使用 520 端口）发送和接收路由信息。
- 组播地址：使用 FF02::9 作为链路本地范围内的 RIPng 路由器组播地址。
- 前缀长度：目的地址使用 128 比特的前缀长度（掩码长度）。
- 下一跳地址：使用 128 比特的 IPv6 地址。
- 源地址：使用链路本地地址 FE80::/10 作为源地址发送 RIPng 路由信息更新报文。

目的

RIPng 是为了支持 IPv6 而对 RIP 协议进行的扩展。

4.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC2080	This document specifies a routing protocol for an IPv6 internet.It is based on protocols and algorithms currently in wide use in the IPv4 Internet.	

4.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R002C00

4.4 原理描述

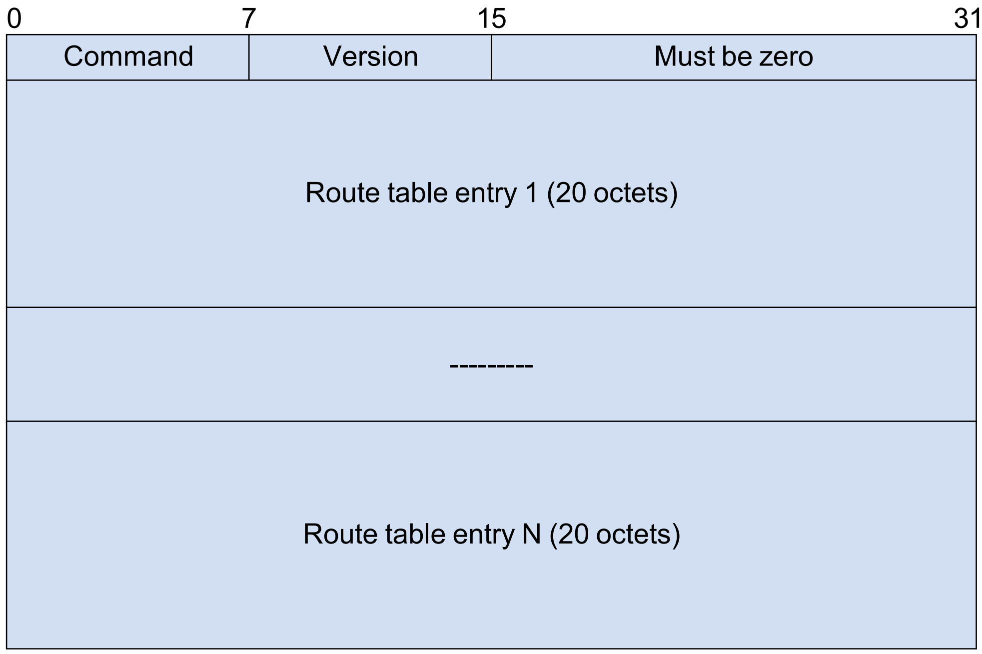
RIPng 协议是对 RIP Version 2 协议在 IPv6 网络上的扩展，RIPng 使用与 RIP Version 2 相同的定时器。RIPng 支持水平分割、毒性逆转和触发更新，用以避免路由环路。

4.4.1 RIPng 报文格式

RIPng 报文由头部（Header）和多个路由表项 RTEs（Route Table Entry）组成。在同一个 RIPng 报文中，RTE 的最大数目根据接口的 MTU 值来确定。

RIPng 报文基本格式如图 **4-1** 所示。

图 4-1 RIPng 报文格式

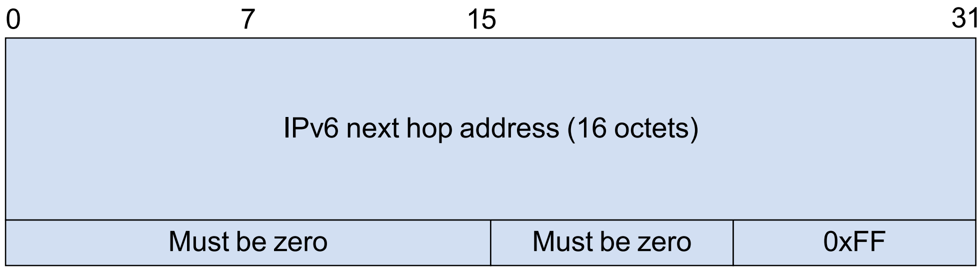


在 RIPng 里有两类 RTE，分别是：

- 下一跳 RTE：位于一组具有相同下一跳的“IPv6 前缀 RTE”的最前面，它定义了下一跳的 IPv6 地址。
- IPv6 前缀 RTE：位于某个“下一跳 RTE”的后面，同一个“下一跳 RTE”的后面可以有多个不同的“IPv6 前缀 RTE”。它描述了 RIPng 路由表中的目的 IPv6 地址及开销。

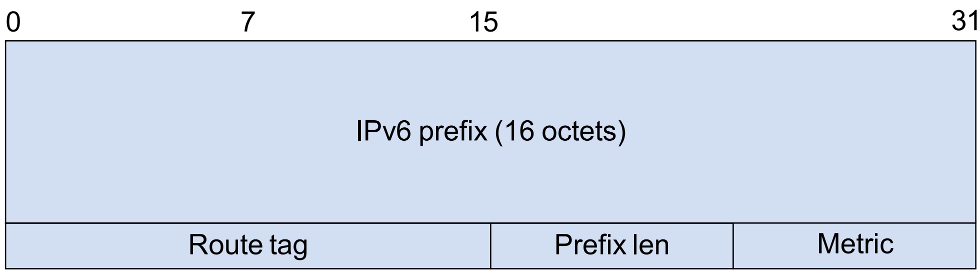
下一跳 RTE 的格式如图 **4-2** 所示。

图 4-2 下一跳 RTE 格式



IPv6 前缀 RTE 的格式如图 **4-3** 所示。

图 4-3 IPv6 前缀 RTE 格式



4.4.2 定时器

RIPng 主要使用三个定时器，分别是更新定时器、老化定时器和垃圾超时定时器，它们的具体功能和实现原理如下：

- 更新定时器（即 Update timer），它定时触发更新报文的发送，更新周期默认为 30 秒，这个定时器是为了同步网络中设备的 RIPng 路由。
- 老化定时器（即 Age timer），RIPng 设备如果在老化时间内没有收到邻居发来的路由更新报文，则认为该路由不可达。
- 垃圾收集定时器（即 Garbage-Collect timer）,在垃圾收集时间内不可达路由没有收到来自同一邻居的更新，则该路由被从路由表中彻底删除。

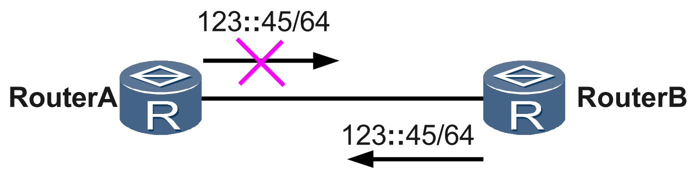
三个定时器之间的关系：

RIP 的更新信息发布是由 Update 定时器控制的，默认为每 30 秒发送一次。每一条路由表项对应两个定时器：老化定时器和垃圾收集定时器。当学到一条路由并安装到路由表中时，老化定时器启动。如果在默认 180 秒后没有收到邻居发来的更新报文，则把该路由的度量值置为 16；并启动垃圾收集定时器，如果在默认 120 秒内仍然没有收到更新报文，定时器超时后在路由中删除该表项。

4.4.3 水平分割（Split Horizon）

水平分割的原理是，RIPng 从某个接口学到的路由，不会从该接口再发回给邻居设备。这样不但减少了带宽消耗，还可以防止路由循环。

图 4-4 水平分割原理图



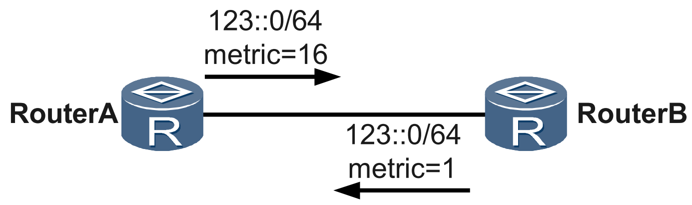
如图 4-4 所示 RouterB 发过来的到网络 123::45 的路由到 RouterA 上后，RouterA 不会再把到网络 123::45 的路由发回给 RouterB。

4.4.4 毒性逆转（Poison Reverse）

毒性逆转的原理是，RIPng 从某个接口学到的路由，将该路由的开销设置为 16（即指明该路由不可达），并从原接口发回邻居设备。通过这种方式，可以清除对方路由表中的无用路由。

RIPng 毒性逆转也是为了防止产生路由环路。

图 4-5 毒性逆转原理图



如图 4-5 所示，在不配置水平分割的情况下，RouterB 会向 RouterA 发送从 RouterA 学到的路由。RouterA 到网络 123::0/64 的路由开销为 1，如果 RouterA 到网络 123::0/64 的路由变成不可达，同时 RouterB 没有收到 RouterA 的更新报文，而继续向 RouterA 发送到网络 123::0/64 的路由信息，则会导致路由环路。

如果 RouterA 在接收到从 RouterB 发来的路由后，向 RouterB 发送一个这条路由不可达的消息，这样 RouterB 就不会再从 RouterA 学到这条可达路由，因此就可以避免上述环路的发生。

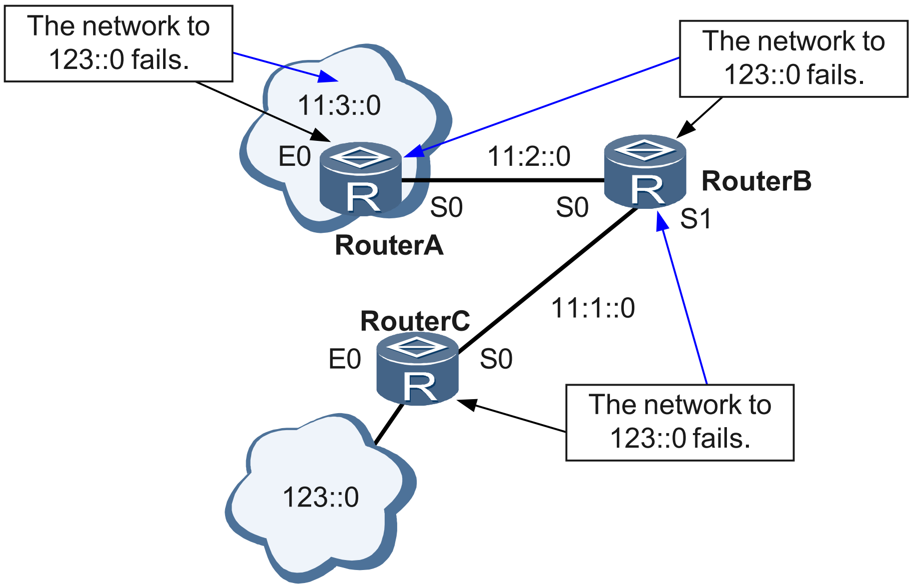
如果毒性逆转和水平分割都配置了，简单的水平分割行为（从某接口学到的路由再从这个接口发布时将被抑制）会被毒性逆转行为代替。

4.4.5 触发更新

触发更新是指路由信息发生变化时，立即向邻居设备发送触发更新报文，通知变化的路由信息。

触发更新缩短了收敛时间，触发更新可以缩短网络收敛时间，在路由表项发生变化时立即向其他设备广播该信息，而不必等待定时更新。

图 4-6 触发更新原理图



如图 4-6 所示，网络 123::0 不可达时，RouterC 最先得到这一信息。通常，更新路由信息会定时发送给相邻 Router。例如，RIPng 协议每隔 30 秒发送一次。但如果在 RouterC 等待更新周期到来的时候，RouterB 的更新报文传到了 RouterC，RouterC 就会学到 RouterB 的去往网络 123::0 的错误路由。这样 RouterB 和 C 上去往网络 123::0 的路由都指向对方从而形成路由环路。如果 RouterC 发现网络故障之后，不再等待更新周期到来，就立即发送路由更新信息给路由器 B，使路由器 B 的路由表及时更新，则可以避免产生上述问题。

触发更新还存在另外一种方式：当下一跳不可用之后（如因为链路故障）需要及时通告给其它 Router，此时要把该路由的 cost 设置为 16 然后发布出去，此更新也叫做路由毒杀。

4.4.6 路由聚合

RIPng 的路由聚合是在接口上实现的，在接口上配置路由聚合，此时可以将 RIPng 要在这个接口上发布出去的路由按最长匹配原则聚合后发布出去。

RIPng 路由聚合可提高大型网络的可扩展性和效率，缩减路由表。

路由聚合的实现原理：

例如，RIPng 可以从这个接口发布出去的路由有两条：

11:11:11::24 Metric=2 和 11:11:12::34 Metric=3，在此接口上配置的聚合路由为 11::0/16，则最终发布出去的路由为 11::0/16 Metric=2。

4.4.7 多进程

为了方便管理，提高控制效率，RIPng 支持多进程和多实例特性。多进程允许为一个指定的 RIPng 进程关联一组接口，从而保证该进程进行的所有协议操作都仅限于这一组接口。这样，就可以实现一台设备有多个 RIPng 协议进程，每个进程负责唯一的一组接口。而且每个 RIPng 进程的路由数据也是相互独立的，但进程之间可以相互引入路由。

4.5 术语与缩略语

术语

术语	解释
毒性逆转	RIPng 从某个接口学到路由后，将该路由的开销设置为 16（不可达），并从原接口发回邻居路由器。
水平分割	RIPng 从某个接口学到的路由，不会从该接口再发回给该邻居路由器。

缩略语

缩略语	英文全称	中文全称
RIPng	RIP next generation	下一代 RIP 协议

5 IS-IS

关于本章

5.1 介绍

定义

IS-IS（Intermediate System to Intermediate System，中间系统到中间系统）最初是国际标准化组织 ISO（the International Organization for Standardization）为它的无连接网络协议 CLNP（ConnectionLess Network Protocol）设计的一种动态路由协议。

随着 TCP/IP 协议的流行，为了提供对 IP 路由的支持，IETF 在 RFC1195 中对 IS-IS 进行了扩充和修改，使它能够同时应用在 TCP/IP 和 OSI 环境中，称为集成 IS-IS（Integrated IS-IS 或 Dual IS-IS）。

本文所指的 IS-IS，如不加特殊说明，均指集成 IS-IS。

目的

IS-IS 属于内部网关协议 IGP（Interior Gateway Protocol），用于自治系统内部。IS-IS 是一种链路状态协议，使用最短路径优先 SPF（Shortest Path First）算法进行路由计算。

5.2 参考标准和协议

表 5-1 本特性的参考资料清单如下：

文档	描述	备注
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	不支持配置多个认证密码
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	-

文档	描述	备注
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	-
draft-ietf-IS-IS-ipv6-05	Routing IPv6 with IS-IS	-
draft-ietf-isis-admin-tags-02(Admin Tag)	Admin Tag	-

5.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R001C00

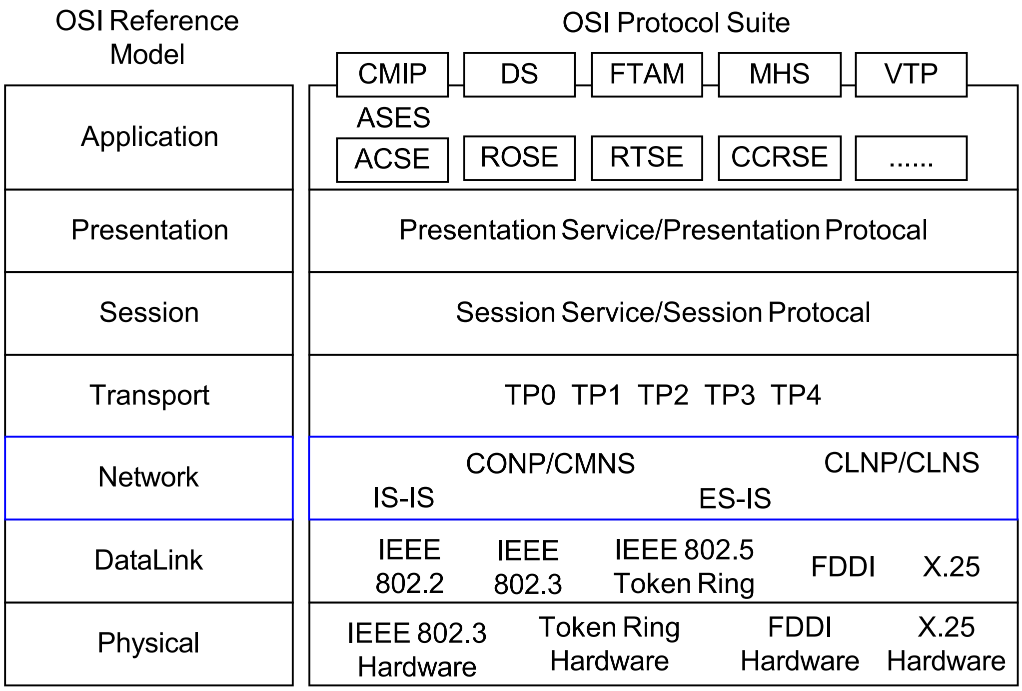
5.4 原理描述

5.4.1 IS-IS 基本概念

IS-IS 的发展

CLNS（Connectionless Network Service）是国际标准化组织 ISO 提出的 OSI 协议栈中的第三层协议。IS-IS 最早由 ISO 设计，用于实现基于 CLNP 寻址的路由协议。

图 5-1 OSI 结构模型



OSI 协议采用体系化（或层次化 Hierarchical）编址，通过 NSAP（Network Service Access Point）来寻址 OSI 网络中处于传输层的各种服务。

OSI 协议的几个常用术语：

- CLNS（Connectionless Network Service）：无连接网络服务
- CLNP（Connectionless Network Protocol）：无连接网络协议
- CMNS（Connection-Mode Network Service）：连接模式网络服务
- CONP（Connection-Oriented Network Protocol）：面向连接网络协议

OSI 通过 CLNP 实现 CLNS，通过 CONP 实现 CMNS。

CLNS 由以下三个协议构成：

- CLNP：类似于 TCP/IP 中的 IP 协议；
- IS-IS：中间系统间的路由协议；
- ES-IS：主机系统与中间系统间的协议，相当于 IP 中的 ARP，ICMP 等。

表 5-2 OSI 与 IP 相对应的概念

缩略语	OSI 中的概念	IP 中对应的概念
IS	Intermediate System 中间系统	路由器
ES	End System 端系统	主机
DIS	Designated Intermediate System 选举中间系统	OSPF 选举路由器
SysID	System ID 系统 ID	OSPF 中的 Router ID
PDU	Protocol Data Unit 协议报文数据单元	IP 报文
LSP	Link state Protocol Data Unit 链路状态协议数据单元	OSPF 中的 LSA
NSAP	Network Service Access Point 网络服务访问点（网络层地址）	IP 地址

随着 TCP/IP 协议的流行，为了提供对 IP 路由的支持，IETF 在 RFC1195 中对 IS-IS 进行了扩充和修改，使它能够同时应用在 TCP/IP 和 OSI 环境中，称为集成化 IS-IS（Integrated IS-IS 或 Dual IS-IS）。

IS-IS 的地址结构

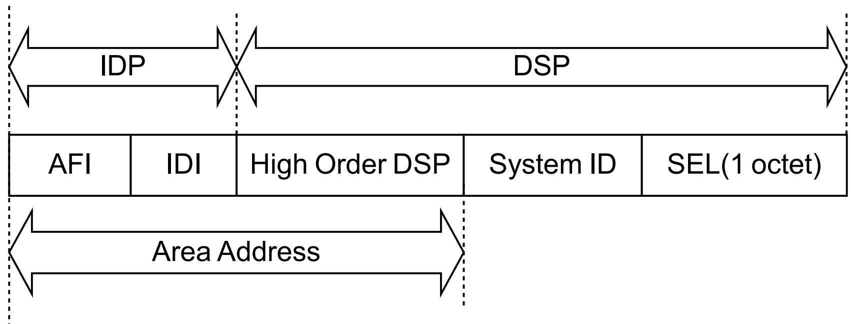
NSAP 是 OSI 协议中用于定位资源的地址。ISO 采用如图 5-2 所示的地址结构，即 NSAP，它由 IDP（Initial Domain Part）和 DSP（Domain Specific Part）组成。IDP 相当于 IP 地址中的主网络号，DSP 相当于 IP 地址中的子网号和主机地址。

IDP 部分是 ISO 规定的，它由 AFI（Authority and Format Identifier）与 IDI（Initial Domain Identifier）组成，AFI 表示地址分配机构和地址格式，IDI 用来标识域。

DSP 由 HODSP、System ID 和 SEL 三个部分组成。HODSP 用来分割区域，System ID 用来区分主机，SEL 指示服务类型。

IDP 和 DSP 的长度都是可变的，NSAP 总长最多是 20 个字节，最少 8 个字节。

图 5-2 IS-IS 协议的地址结构示意图



- 区域地址

  IDP 和 DSP 中的 HODSP（High Order DSP）一起，既能够标识路由域，也能够标识路由域中的区域，因此，它们一起被称为区域地址（Area Address），相当于

OSPF 中的区域编号。同一个路由域中不允许有相同的区域地址。同一区域中路由器的 Level-1区域地址必须相同。

一般情况下，一个路由器只需要配置一个区域地址，且同一区域中所有节点的区域地址都要相同。为了支持区域的平滑合并、分割及转换，在设备的实现中，一个 IS-IS 进程下最多可配置 3 个区域地址。

- System ID

  System ID 用来在区域内唯一标识主机或路由器。在设备的实现中，它的长度固定为 48bit（6 字节）。

  在实际应用中，一般使用 Router ID 与 System ID 进行对应。假设一台路由器使用接口 Loopback0 的 IP 地址 168.10.1.1 作为 Router ID，则它在 IS-IS 使用的 System ID 可通过如下方法转换得到：

  - 将 IP 地址 168.10.1.1 的每个十进制数都扩展为 3 位，不足 3 位的在前面补 0；
  - 将扩展后的地址 168.010.001.001 分为 3 部分，每部分由 4 位数字组成；
  - 重新组合的 1680.1000.1001 就是 System ID。

  实际 System ID 的指定可以有不同的方法，但要保证能够唯一标识主机或路由器。

- SEL

  SEL（NSAP Selector，有时也写成 N-SEL）的作用类似 IP 中的“协议标识符”，不同的传输协议对应不同的 SEL。在 IP 上 SEL 均为 00。

- NET

  网络实体名称 NET（Network Entity Title）指的是 IS 本身的网络层信息，可以看作是一类特殊的 NSAP（SEL ＝ 0）。NET 的长度与 NSAP 的相同，最多为 20 个字节，最少为 8 个字节。在路由器上配置 IS-IS 时，只需要考虑 NET 即可，NSAP 可不必去关注。

  通常情况下，一个 IS-IS 进程下配置一个 NET 即可，当区域需要重新划分时，例如将多个区域合并，或者将一个区域划分为多个区域，这种情况下配置多个 NET 可以在重新配置时仍然能够保证路由的正确性。

  由于一个 IS-IS 进程中区域地址最多可配置 3 个，所以 NET 最多也只能配 3 个。在配置多个 NET 时，必须保证它们的 System ID 都相同。

  例如有 NET 为：ab.cdef.1234.5678.9abc.00，则其中 Area 为 ab.cdef，System ID 为 1234.5678.9abc，SEL 为 00。

  说明

  位于同一区域内的路由器的区域地址必须相同。

IS-IS PDU 格式

IS-IS PDU 有以下类型：HELLO、LSP、CSNP 和 PSNP。

表 5-3 PDU 类型对应关系表

类型值	PDU 类型	简称
15	Level-1 LAN IS-IS Hello PDU	L1 LAN IIH
16	Level-2 LAN IS-IS Hello PDU	L2 LAN IIH
17	Point-to-Point IS-IS Hello PDU	P2P IIH
18	Level-1 Link State PDU	L1 LSP

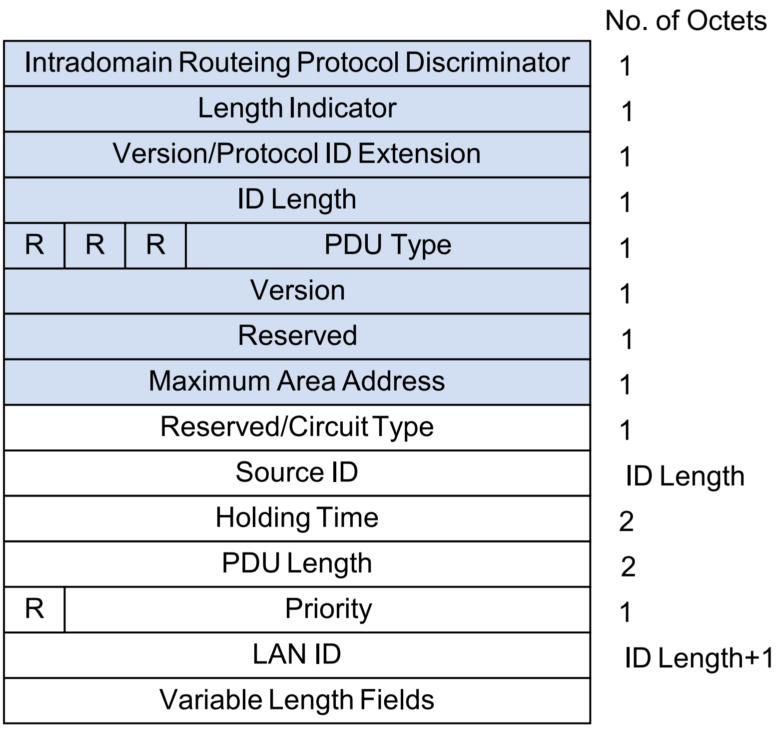
类型值	PDU 类型	简称
20	Level-2 Link State PDU	L2 LSP
24	Level-1 Complete Sequence Numbers PDU	L1 CSNP
25	Level-2 Complete Sequence Numbers PDU	L2 CSNP
26	Level-1 Partial Sequence Numbers PDU	L1 PSNP
27	Level-2 Partial Sequence Numbers PDU	L2 PSNP

- Hello 报文格式

  Hello 报文用于建立和维持邻居关系，也称为 IIH（IS-to-IS Hello PDUs）。其中，广播网中的 Level-1 IS-IS 使用 Level-1 LAN IIH；广播网中的 Level-2 IS-IS 使用 Level-2 LAN IIH；非广播网络中则使用 P2P IIH。它们的报文格式有所不同。

  广播网中的 Hello 报文格式如图 5-3 所示（蓝色部分是通用报文头）。

图 5-3 Level-1/Level-2 LAN IIH 格式



P2P 网络中的 Hello 报文格式如图 5-4 所示。

图 5-4 P2P IIH 格式

	No. of Octets
Intradomain Routeing Protocol Discriminator	1
Length Indicator	1
Version/Protocol ID Extension	1
ID Length	1
R | R | R | PDU Type	1
Version	1
Reserved	1
Maximum Area Address	1
Reserved/Circuit Type	1
Source ID	ID Length
Holding Time	2
PDU Length	2
Local Circuit ID	1
Variable Length Fields	

从图中可以看出，P2P IIH 中的多数字段与 LAN IIH 相同。不同的是没有 Priority 和 LAN ID 字段，而多了一个 Local Circuit ID 字段，表示本地链路 ID。

- LSP 报文格式

  链路状态报文 LSP（Link State PDUs）用于交换链路状态信息。LSP 分为两种：Level-1 LSP 和 Level-2 LSP。Level-1 LSP 由 Level-1 IS-IS 传送，Level-2 LSP 由 Level-2 IS-IS 传送，Level-1-2 IS-IS 则可传送以上两种 LSP。

  两类 LSP 有相同的报文格式，如图 **5-5** 所示。

图 5-5 Level-1/Level-2 LSP 格式

				No. of Octets
IntradomainRouteingProtocolDiscriminator				1
Length Indicator				1
Version/Protocol IDExtension				1
ID Length				1
R	R	R	PDU Type	1
Version				1
Reserved				1
Maximum AreaAddress				1
PDULength				2
RemainingLifetime				ID Length+2
SequencyNumber				4
Checksum				2
R	ATT	OL	IS Type	1
Variable Length Fields				

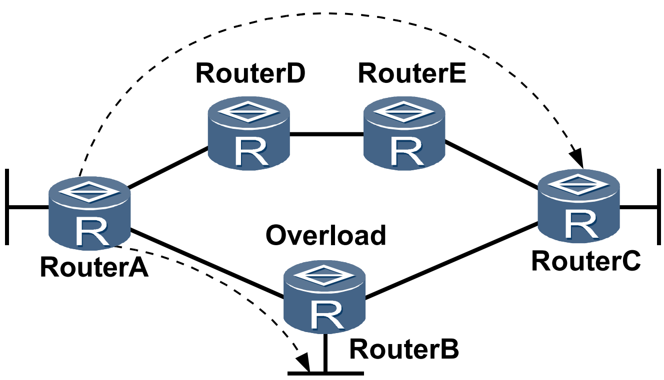
主要字段的解释如下：

- OL（LSDB Overload）：过载标志位。

  设置了过载标志位的 LSP 虽然还会在网络中扩散，但是在计算通过过载路由器的路由时不会被采用。即对路由器设置过载位后，其它路由器在进行 SPF 计算时不会使用这台路由器做转发，只计算该节点上的直连路由。

  如图 5-6 所示，RouterA 到 RouterC 的报文由 RouterB 转发，但如果 RouterB 的 OL 位置 1，则 RouterA 认为 RouterB 的 LSDB 不完整，将转发流量通过 RouterD、RouterE 转发给 RouterC，但到 RouterB 直连地址的流量不受影响。

图 5-6 LSDB Overload 示意图



- IS Type：生成 LSP 的 IS-IS 类型。

  用来指明是 Level-1 还是 Level-2 IS-IS（01 表示 Level-1，11 表示 Level-2）。

- SNP 格式

  SNP（Sequence Number PDUs）通过描述全部或部分数据库中的 LSP 来同步各 LSDB（Link-State DataBase），从而维护 LSDB 的完整与同步。

  SNP 包括 CSNP（Complete SNP，全序列号报文）和 PSNP（Partial SNP，部分序列号报文），进一步又可分为 Level-1 CSNP、Level-2 CSNP、Level-1 PSNP 和 Level-2 PSNP。

  CSNP 包括 LSDB 中所有 LSP 的摘要信息，从而可以在相邻路由器间保持 LSDB 的同步。在广播网络上，CSNP 由 DIS 定期发送（缺省的发送周期为 10 秒）；在点到点链路上，CSNP 只在第一次建立邻接关系时发送。

  CSNP 的报文格式如图 5-7 所示。

图 5-7 Level-1/Level-2 CSNP 格式

	No. of Octets
Intradomain Routeing Protocol Discriminator	1
Length Indicator	1
Version/Protocol ID Extension	1
ID Length	1
R | R | R | PDU Type	1
Version	1
Reserved	1
Maximum Area Address	1
PDU Length	2
Source ID	ID Length+1
Start LSP ID	ID Length+2
End LSP ID	ID Length+2
Variable Length Fields	

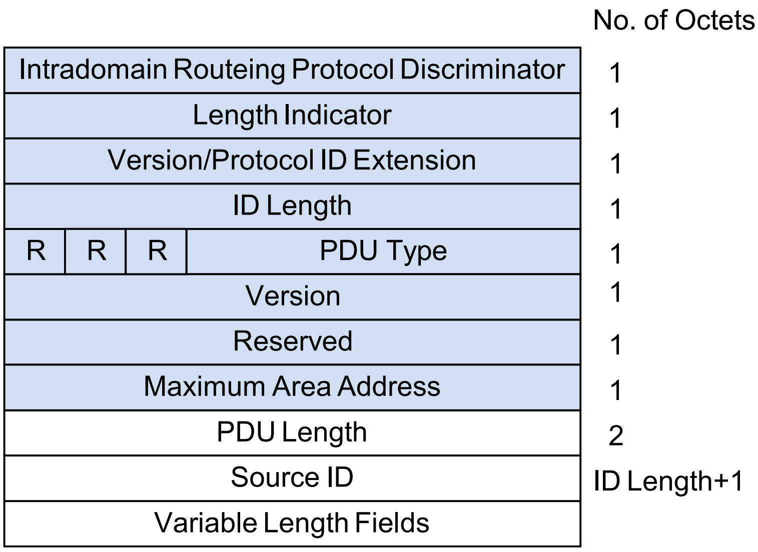
  主要字段的解释如下：

  - Source ID：发出 SNP 报文的设备的 System ID。
  - Start LSP ID：CSNP 报文中第一个 LSP 的 ID 值。
  - End LSP ID：CSNP 报文中最后一个 LSP 的 ID 值。

  PSNP 只列举最近收到的一个或多个 LSP 的序号，它能够一次对多个 LSP 进行确认，当发现 LSDB 不同步时，也用 PSNP 来请求邻居发送新的 LSP。

  PSNP 的报文格式如图 5-8 所示。

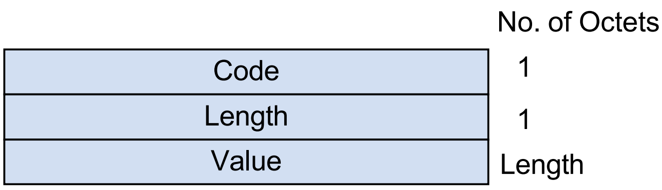
图 5-8 Level-1/Level-2 PSNP 格式



- CLV

  PDU 中的变长字段部分是多个 CLV（Code-Length-Value）三元组。其格式如图 **5-9** 所示。CLV 也称为 TLV（Type-Length-Value）。

图 5-9 CLV 格式



不同 PDU 类型所包含的 CLV 是不同的。如表 **5-4** 所示。

表 5-4 PDU 类型和包含的 CLV 名称

CLV Code	名称	所应用的 PDU 类型
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

CLV Code	名称	所应用的 PDU 类型
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

其中，Code 值从 1 到 10 的 CLV 在 ISO10589 中定义（有 2 类未在上表中列出），其他几种 CLV 在 RFC1195 中定义。

IS-IS 区域

- 两级结构

  为了支持大规模的路由网络，IS-IS 在路由域内采用两级的分层结构。一个大的 Domain（域）可以被分为多个 Areas（区域）。一般来说，将 Level-1 路由器部署在区域内，Level-2 路由器部署在区域间，Level-1-2 路由器部署在 Level-1 和 Level-2 路由器的中间。

- Level-1 路由器

  Level-1 路由器负责区域内的路由，它只与属于同一区域的 Level-1 和 Level-1-2 路由器形成邻居关系。一个 Level-1 路由器只负责维护 Level-1 的 LSDB，该 LSDB 包含本区域的路由信息，到本区域外的报文转发给最近的 Level-1-2 路由器。

- Level-2 路由器

  Level-2 路由器负责区域间的路由，可以与 Level-2 或其它区域的 Level-1-2 路由器形成邻居关系，维护一个 Level-2 的 LSDB，该 LSDB 包含区域间的路由信息。

  所有 Level-2 级别（即形成 Level-2 邻居关系）的路由器组成路由域的骨干网，负责在不同区域间通信，路由域中 Level-2 级别的路由器必须是连续的，以保证骨干网的连续性。只有 Level-2 级别的路由器才能直接与区域外的路由器交换数据报文或路由信息。

- Level-1-2 路由器

  同时属于 Level-1 和 Level-2 的路由器称为 Level-1-2 路由器，可以与同一区域的 Level-1 和 Level-1-2 路由器形成 Level-1 邻居关系，也可以与其他区域的 Level-2 和 Level-1-2 路由器形成 Level-2 的邻居关系。Level-1 路由器必须通过 Level-1-2 路由器才能连接至其他区域。

  Level-1-2 路由器维护两个 LSDB，Level-1 的 LSDB 用于区域内路由，Level-2 的 LSDB 用于区域间路由。

  说明

  属于不同区域的 Level-1 路由器不能形成邻居关系。Level-2 路由器之间可以直接形成邻居，与所在区域无关。

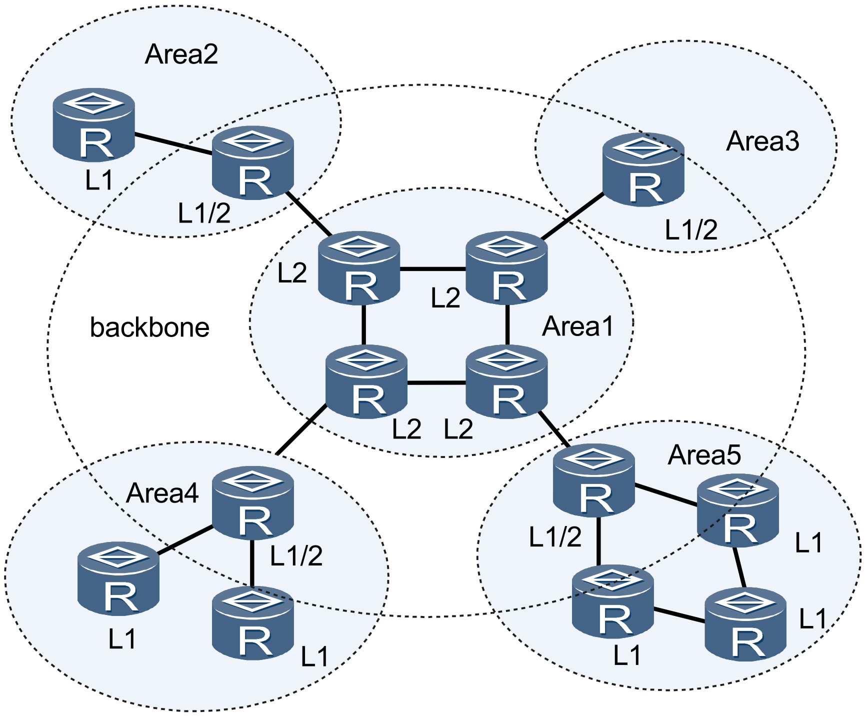
- 接口的级别

  对于 Level-1-2 路由器，可能需要与某个对端只建立 Level-1 的邻接关系，与另一个对端只建立 Level-2 的邻接关系。可以通过设置相应接口的级别来限制接口上所能

建立的邻接关系，如 Level-1 的接口只能建立 Level-1 的邻接关系，Level-2 的接口只能建立 Level-2 的邻接关系。

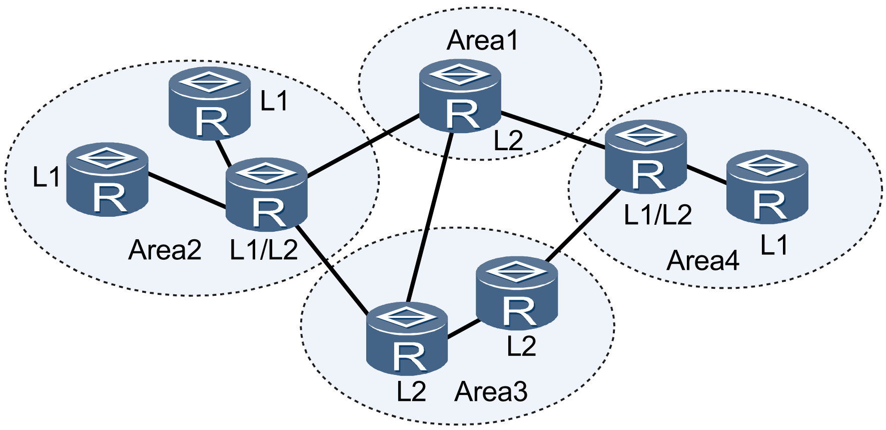
如图 5-10 所示为一个运行 IS-IS 协议的网络，它与 OSPF 的多区域网络拓扑结构非常相似。整个 backbone 区域不仅包括 Area1 中的所有路由器，还包括其它区域的 Level-1-2 路由器。

图 5-10 IS-IS 拓扑结构图一



如图 5-11 所示是 IS-IS 的另外一种拓扑结构图。所有连续的 Level-1-2 和 Level-2 路由器构成了 IS-IS 的骨干区域。在这个拓扑中，Level-2 / Level-1-2 级别的路由器分别属于不同的区域，并没有规定哪个区域是骨干区域。

图 5-11 IS-IS 拓扑结构图二



说明

IS-IS 的骨干网（Backbone）指的不是一个特定的区域。

这种组网方案也体现出 IS-IS 与 OSPF 的不同点。在 OSPF 中，区域之间的路由需要通过骨干区域转发，只有在同一个区域内才使用 SPF 算法。而 IS-IS 不论是 Level-1 还是 Level-2 路由，都采用 SPF 算法，分别生成最短路径树 SPT（Shortest Path Tree）。

IS-IS 的网络类型

IS-IS 只支持两种类型的网络，根据物理链路不同可分为：

- 广播链路：如 Ethernet、Token-Ring 等。
- 点到点链路：如 PPP、HDLC 等。

对于 NBMA（Non-Broadcast Multi-Access）网络，如 ATM，需对其配置子接口，并注意子接口类型应配置为 P2P。IS-IS 不能在点到多点链路 P2MP（Point to MultiPoint）上运行。

DIS 和伪节点

在广播网络中，IS-IS 需要在所有的路由器中选举一个路由器作为 DIS（Designated Intermediate System）。

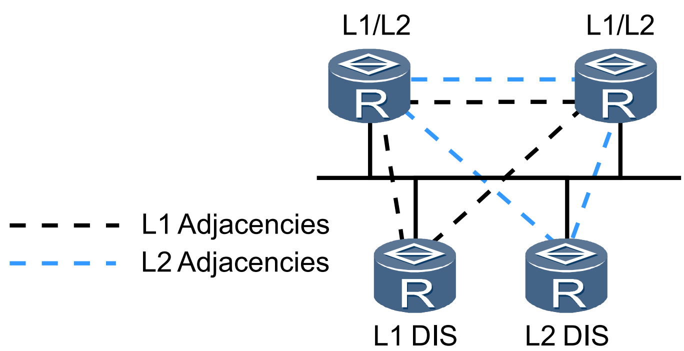
Level-1 和 Level-2 的 DIS 是分别选举的，用户可以为不同级别的 DIS 选举设置不同的优先级。DIS 优先级数值最大的被选为 DIS。如果优先级数值最大的路由器有多台，则其中 MAC 地址最大的路由器会被选中。不同级别的 DIS 可以是同一台路由器，也可以是不同的路由器。

与 OSPF 的不同点：

- 优先级为 0 的路由器也参与 DIS 的选举；
- 当有新的路由器加入，并符合成为 DIS 的条件时，这个路由器会被选中成为新的 DIS，原有的伪节点被删除。此更改会引起一组新的 LSP 泛洪。

在 IS-IS 广播网中，同一网段上的同一级别的路由器之间都会形成邻接关系，包括所有的非 DIS 路由器之间也会形成邻接关系，这一点与 OSPF 是不同的。如图 5-12 所示。

图 5-12 IS-IS 广播网的 DIS 和邻接关系



DIS 用来创建和更新伪节点（Pseudonodes），并负责生成伪节点的 LSP，用来描述这个网络上有哪些路由器。

伪节点是用来模拟广播网络的一个虚拟节点，并非真实的路由器。在 IS-IS 中，伪节点用 DIS 的 System ID 和一个字节的 Circuit ID（非 0 值）标识。

使用伪节点可以简化网络拓扑，使路由器产生的 LSP 长度较小。另外，当网络发生变化时，需要产生的 LSP 数量也会较少，减少 SPF 的资源消耗。

说明

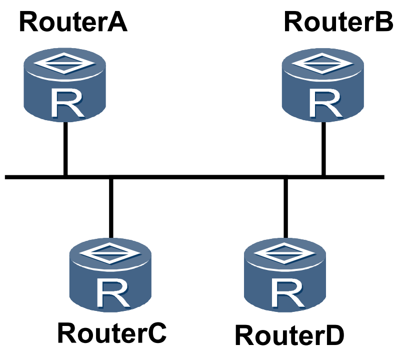
IS-IS 广播网上所有的路由器之间都形成邻接关系，但 LSDB 的同步仍然依靠 DIS 来保证。

IS-IS 邻居关系的建立

两台运行 IS-IS 的路由器在交互协议报文实现路由功能之前必须首先建立邻居关系。在不同类型的网络上，IS-IS 的邻居建立方式并不相同。

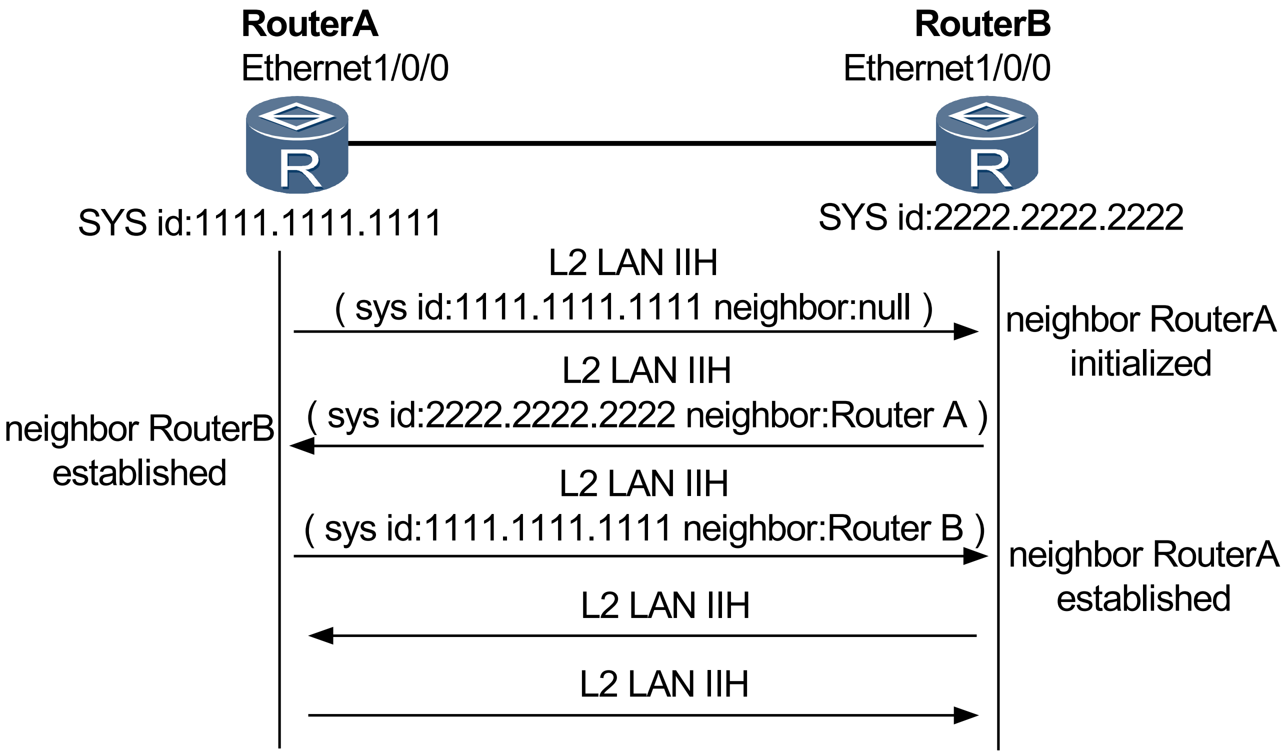
- 广播链路邻居关系的建立

图 5-13 广播链路组网图



RouterA、RouterB、RouterC 和 RouterD 都是 Level-2 路由器。RouterA 新加入到此广播网络中。图 5-14 只列出 RouterA 和 RouterB 建立邻居的过程，RouterA 与 RouterC 和 RouterD 建立邻居的过程与此相同。

图 5-14 广播链路邻居关系建立过程



RouterA 广播发送 Level-2 LAN IS-IS Hello PDU，RouterB 收到此报文后，将自己和 RouterA 的邻居状态标识为 Initial；然后，RouterB 再回复 L2 LAN IIH 报文，RouterA 收到这个带有 RouterA 为 RouterB 邻居信息的 IIH 报文后，RouterA 再将自己与 RouterB 的邻居状态标识为 Up。

因为是广播网络，需要选举 DIS，所以在邻居关系建立过程后，路由器会等待两个 Hello 报文间隔，再进行 DIS 的选举。IIH 报文中包含 Priority 字段，Priority 值最大的将被选举为该广播网的 DIS。若优先级相同，接口 MAC 地址较大的被选举为 DIS。

- P2P 链路邻居关系的建立

  在 P2P 链路上，邻居关系的建立不同于广播链路。分为 2-way 和 3-way 方式。

  - 2-way 方式

    即只要设备收到 IS-IS Hello 报文，就会单方向建立起邻居关系。

  - 3-way 方式

    此方式通过三次发送 P2P 的 IS-IS Hello PDU 最终建立起邻居关系，类似广播邻居关系的建立。

    说明

    对 IS-IS 三次握手机制特性会在 IS-IS 三次握手机制章节进行更详细的讨论。

IS-IS 按如下原则建立邻居关系：

- 只有同一层次的相邻路由器才有可能成为邻居。
- 对于 Level-1 路由器来说要求区域号一致。
- 在同一网段。

链路两端的 IS-IS 接口的网络类型必须一致，否则双方不可以建立起邻居关系。可以通过将以太网接口模拟成 P2P 接口，建立 P2P 链路邻居关系。

由于 IS-IS 是直接运行在数据链路层上的协议，并且最早设计是给 CLNP 使用的，IS-IS 邻居关系的形成与 IP 地址无关。但在实际的实现中，由于只在 IP 上运行 IS-IS，所以是要检查对方的 IP 地址的。如果接口配置了从 IP，那么只要双方有某个 IP（主 IP 或者从 IP）在同一网段，就能建立邻居，不一定要主 IP 相同。

在没有配置 IP 地址借用的情况下，如果对方的 IP 地址不和自己收到报文的接口 IP 地址在同一网段上，将不形成邻居关系，这样可以避免 IP 的不可达性。如果配置接口对接收的 Hello 报文不作 IP 地址检查，就可以建立邻居关系。

- 对于 P2P 接口，可以配置接口忽略 IP 地址检查。
- 对于以太网接口，需要将以太网接口模拟成 P2P 接口，然后才可以配置接口忽略 IP 地址检查。

IS-IS 的 LSP 交互过程

- LSP 的“泛洪”（flooding）

  LSP 报文的“泛洪”指当一个路由器向相邻路由器报告自己的 LSP 后，相邻路由器再将同样的 LSP 报文传送到除发送该 LSP 的路由器外的其它邻居，并这样逐级将 LSP 传送到整个层次内的一种方式。通过这种“泛洪”，整个层次内的每一个路由器就都可以拥有相同的 LSP 信息，并保持 LSDB 的同步。

  每一个 LSP 都拥有其自己的一个 4 字节的序列号。在路由器启动时所发送的第一个 LSP 报文中的序列号为 1，以后当需要生成新的 LSP 时，新 LSP 的序列号在前一个 LSP 序列号的基础上加 1。更高的序列号意味着更新的 LSP。

- LSP 产生的原因

  IS-IS 路由域内的所有路由器都会产生 LSP，以下事件会触发一个新的 LSP：

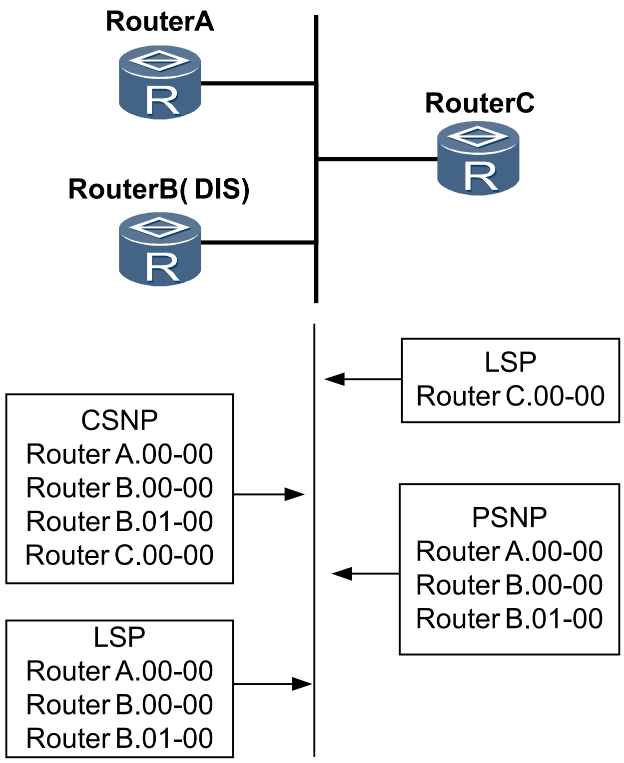
  - 邻居 Up 或 Down
  - IS-IS 相关接口 Up 或 Down
  - 引入的 IP 路由发生变化
  - 区域间的 IP 路由发生变化
  - 接口被赋了新的 metric 值
  - 周期性更新

- 收到邻居新的 LSP 的处理过程

  1. 将新的 LSP 安装到自己的 LSDB 数据库中标记为 flooding。
  2. 发送新的 LSP 到除了收到该 LSP 的接口之外的接口。
  3. 邻居再扩散到其他邻居。

- 新加入路由器与 DIS 同步 LSDB 数据库过程

图 5-15 广播链路数据库更新过程



- 新加入的路由器 RouterC 首先发送 Hello 报文，与该广播域中的路由器建立邻居关系。（请参见“广播链路邻居关系的建立”）
- 邻居关系建立起来后，RouterC 等待 LSP 定时器超时，然后将自己的 LSP 发送往组播地址：

  Level-1：01-80-C2-00-00-14

  Level-2：01-80-C2-00-00-15

  网络上所有的邻居都将收到该 LSP。
- 该网段中的 DIS 会把收到 RouterC 的 LSP 加入到 LSDB 中，并等待 CSNP 报文定时器超时并发送 CSNP 报文，进行该网络内的 LSDB 同步。CSNP 报文的发送间隔缺省值为 10 秒。
- RouterC 收到 DIS 发来的 CSNP 报文，对比自己的 LSDB 数据库，发送 PSNP 报文请求自己没有的 LSP。
- DIS 收到该 PSNP 报文请求后发送对应的 LSP 进行 LSDB 的同步。

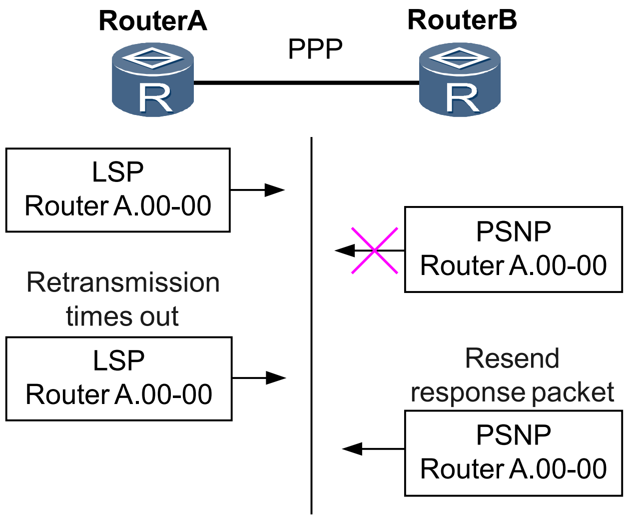
● DIS 的 LSDB 更新过程
- DIS 接收到 LSP，在数据库中搜索对应的记录。若没有该 LSP，则将其加入数据库，并广播新数据库内容。
- 若收到的 LSP 序列号大于本地 LSP 的序列号，就替换为新报文，并广播新数据库内容。
- 若收到的 LSP 序列号小本地 LSP 的序列号，就向入端接口发送本地 LSP 报文。
- 若两个序列号相等，则比较 Remaining Lifetime。若收到的 LSP 的 Remaining Lifetime 小于本地 LSP 的 Remaining Lifetime，就替换为新报文，并广播新数据

库内容；若收到的 LSP 的 Remaining Lifetime 大于本地 LSP 的 Remaining Lifetime，就向入端接口发送本地 LSP 报文。

- 若两个序列号和 Remaining Lifetime 都相等，则比较 Checksum。若收到的 LSP 的 Checksum 大于本地 LSP 的 Checksum，就替换为新报文，并广播新数据库内容；若收到的 LSP 的 Checksum 小于本地 LSP 的 Checksum，就向入端接口发送本地 LSP 报文。
- 若两个序列号、Remaining Lifetime 和 Checksum 都相等，则不转发该报文。

- P2P 链路上数据库的同步过程

图 5-16 点到点链路数据库更新过程



1. 邻居关系建立请参见“点到点链路邻居关系的建立”。
2. 第一次建立起邻居时，路由器会先发送 CSNP 给对端。如果对端的 LSDB 与 CSNP 没有同步，则发送 PSNP 请求索取相应的 LSP。
3. 本地将邻居请求的 LSP 发给邻居同时启动 LSP 重传定时器，并等待邻居发送的 PSNP 作为收到 LSP 的确认。
4. 如果在接口 LSP 重传定时器超时后还没有收到对端发送的 PSNP 报文作为应答，则重新发送该 LSP。

说明

在 P2P 链路上 PSNP 有两种作用：

- 作为 Ack 应答以确认收到的 LSP。
- 用来请求所需的 LSP。

- P2P 的 LSDB 更新过程
  - 若收到的 LSP 比本地的序列号更大，则将这个新的 LSP 存入自己的 LSDB，再通过一个 PSNP 报文来确认收到此 LSP，最后再将这个新 LSP 发送给除了发送该 LSP 的邻居以外的邻居。
  - 若收到的 LSP 比本地的序列号更小，则直接给对方发送本地的 LSP，然后等待对方给自己一个 PSNP 报文作为确认。
  - 若收到的 LSP 序列号和本地相同，则比较 Remaining Lifetime，若收到 LSP 的 Remaining Lifetime 小于本地 LSP 的 Remaining Lifetime，则将收到的 LSP 存入

LSDB 中并发送 PSNP 报文来确认收到此 LSP，然后将该 LSP 发送给除了发送该 LSP 的邻居以外的邻居；若收到 LSP 的 Remaining Lifetime 大于本地 LSP 的 Remaining Lifetime，则直接给对方发送本地的 LSP，然后等待对方给自己一个 PSNP 报文作为确认。

- 若收到的 LSP 和本地 LSP 的序列号和 Remaining Lifetime 都相同，则比较 Checksum，若收到 LSP 的 Checksum 大于本地 LSP 的 Remaining Lifetime，则将收到的 LSP 存入 LSDB 中并发送 PSNP 报文来确认收到此 LSP，然后将该 LSP 发送给除了发送该 LSP 的邻居以外的邻居；若收到 LSP 的 Checksum 小于本地 LSP 的 Remaining Lifetime，则直接给对方发送本地的 LSP，然后等待对方给自己一个 PSNP 报文作为确认。
- 若收到的 LSP 和本地 LSP 的序列号、Remaining Lifetime 和 Checksum 都相同，则不转发该报文。

5.4.2 IS-IS 多实例和多进程

对于支持 VPN 的路由器，可以将每个 IS-IS 进程都与一个指定的 VPN 实例相关联。因此可以配置多个 IS-IS 进程分别绑定多个 VPN 实例。

- IS-IS 多实例指在同一台路由器上，可以配置多个 IS-IS 实例。
- IS-IS 多进程指在同一个 VPN 下（或者同在公网下）创建多个 IS-IS 进程。
  - 多进程允许为一个指定的 IS-IS 进程关联一组接口，从而保证该进程进行的所有协议操作都仅限于这一组接口。这样，就可以实现一台路由器有多个 IS-IS 协议进程，每个进程负责唯一的一组接口。
  - IS-IS 多进程共用同一个 RM 路由表。IS-IS 多实例使用 VPN 中的 RM 路由表。每个 VPN 有自己单独的 RM 路由表。
  - IS-IS 进程在创建时可以选择绑定 VPN，绑定 VPN 后，IS-IS 进程就从属于这个 VPN，只接受和处理此 VPN 内的事件，VPN 删除时，IS-IS 进程也跟着删除。

为了方便管理，提高控制效率，IS-IS 支持多进程和多实例特性。

例如：为私网用户提供 IS-IS 协议功能。配置了 VPN 后，VPN 所绑定的接口，以及产生的路由都与其他 VPN 以及公网数据完全相隔离，因此若要在 VPN 中使用 IS-IS 进行部署，就可以使用 IS-IS 多实例。

对于支持 VPN 的路由器，每个 IS-IS 进程都与一个指定的 VPN 实例相关联。这样，所有附加到该进程的接口都应与该进程相关联的 VPN 实例相关联。

目前实例本身由 VPN 模块维护，所以 IS-IS 的实现就是在创建进程时绑定对应的 VPN，以此实现 IS-IS 的多实例。

配置 IS-IS 多实例和多进程时，有以下注意事项：

- 创建 IS-IS 多实例时，必须在创建 IS-IS 进程时绑定 VPN。如果在创建时没有进行绑定，后面无法通过配置将一个已存在的进程绑定到一个 VPN 上。
- 对一个已绑定了 VPN 的 IS-IS 进程，无法通过配置绑定到另一个 VPN 上。
- 一个 IS-IS 进程只能绑定同一个协议（IPv4 或 IPv6）的一个 VPN，但可以同时绑定一个 IPv4 和一个 IPv6 的 VPN。
- 多个 IS-IS 进程可以绑定到同一个 VPN 上，这就是 IS-IS 多进程。
- 需要使能 IS-IS 多实例的接口必须和 IS-IS 绑定相同的 VPN。
- 绑定了 VPN 的 IS-IS 进程从属于 VPN，所以当 VPN 删除时，IS-IS 进程也跟着删除。

- 不同 VPN 的路由不能相互引入。

5.4.3 IS-IS 路由渗透

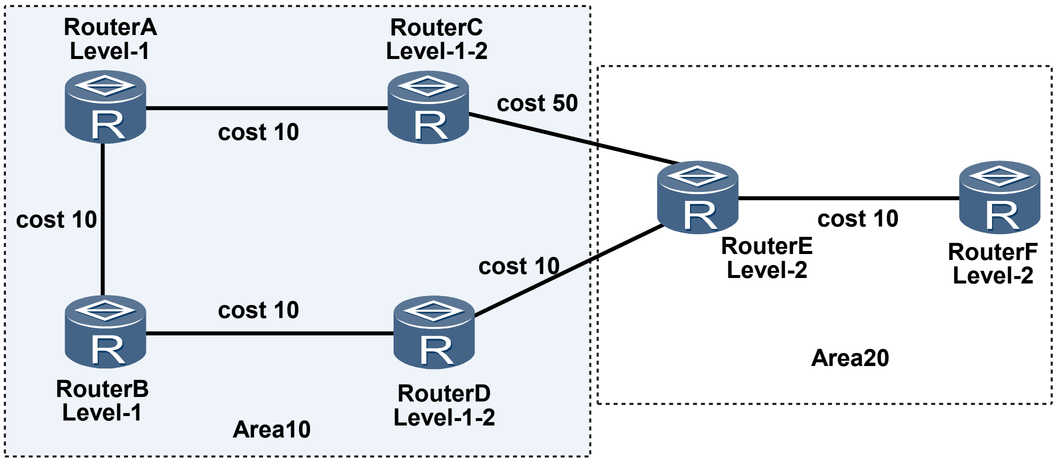
路由渗透特性是指 Level-1-2 IS-IS 将已知的其他 Level-1 区域以及 Level-2 区域的路由信息通报给指定的 Level-1 区域。

通常情况下，区域内的路由通过 Level-1 的路由器进行管理。所有的 Level-2 和 Level-1-2 路由器构成一个连续的骨干区域。Level-1 区域必须且只能与骨干区域相连，不同的 Level-1 区域之间并不相连。

Level-1 区域内的路由信息通过 Level-1-2 路由器通报给 Level-2 区域的，即 Level-1-2 路由器将学习到的 Level-1 路由信息装进 Level-2 LSP，再泛洪 LSP 给其他 Level-2 和 Level-1-2 路由器。因此，Level-1-2 和 Level-2 路由器知道整个 IS-IS 路由域的路由信息。但是，为了有效减小路由表的规模，在缺省情况下，Level-2 路由器并不将自己知道的 Level-1 区域以及骨干区域的路由信息通报给 Level-1 区域。这样，Level-1 路由器将不了解本区域以外的路由信息，可能导致对本区域之外的目的地址无法选择最佳的路由。

为解决上述问题，IS-IS 提供了路由渗透功能。通过在 Level-1-2 路由器上定义 ACL（Access Control List）、路由策略、Tag 标记等方式，将符合条件的路由筛选出来，实现将其他 Level-1 区域和骨干区域的部分路由信息通报给自己所在的 Level-1 区域。

图 5-17 路由渗透示例



- RouterA、RouterB、RouterC 和 RouterD 同属于 Area10 区域，RouterA 和 RouterB 为 Level-1 路由器，RouterC 和 RouterD 为 Level-1-2 路由器。
- RouterE、RouterF 同属于 Area20 区域，为 Level-2 路由器。

RouterA 发送报文给 RouterF，选择的最佳路径应该是 RouterA->RouterB->RouterD->RouterE->RouterF。因为这条链路上的 cost 值为 10+10+10+10=40，但在 RouterA 上查看发送到 RouterF 的报文选择的路径是：RouterA->RouterC->RouterE->RouterF，其 cost 值为 10+50+10=70，不是 RouterA 到 RouterF 的最优路由。

这是由于 RouterA 并不知道本区域外部的路由，所以发往非本区域网段内的报文都是通过由最近的 Level-1-2 路由器产生的缺省路由发送出去的。

此时分别在 Level-1-2 路由器 RouterC 和 RouterD 上使能路由渗透。再查看报文选择的路径，发现路径是 RouterA->RouterB->RouterD->RouterE->RouterF，为 RouterA 到 RouterF 的最优路由。

5.4.4 IS-IS 快速收敛

IS-IS 快速收敛是为了提高路由的收敛速度而做的扩展特性。包括：

- I-SPF（Incremental SPF）

  增量最短路径优先算法，是指当网络拓扑改变的时候，只对受影响的节点进行路由计算，而不是对全部节点重新进行路由计算，从而加快了路由的计算。

- PRC（Partial Route Calculation）

  部分路由计算，是指当网络上路由发生变化的时候，只对发生变化的路由进行重新计算。

- LSP 快速扩散

  可以加快 LSP 的扩散速度。

- 智能定时器

  定时器第一次超期时间是一个固定的时间。如果在定时器被设置但是还未超期的时候，又有触发定时器的事件发生，则该定时器下一次超期的时间会增加。

  在产生 LSP 和进行 SPF 计算的时候都用到这种定时器。

I-SPF

在 ISO10589 中定义使用 Dijkstra 算法进行路由计算。当网络拓扑中有一个节点发生变化时，这种算法需要重新计算网络中的所有节点，计算时间长，占用过多的 CPU 资源，影响整个网络的收敛速度。

I-SPF 改进了这个算法，除了第一次计算时需要计算全部节点外，每次只计算受到影响的节点，而最后生成的最短路径树 SPT 与原来的算法所计算的结果相同，大大降低了 CPU 的占用率，提高了网络收敛速度。

PRC

PRC 的原理与 I-SPF 相同，都是只对发生变化的路由进行重新计算。不同的是，PRC 不需要计算节点路径，而是根据 I-SPF 算出来的 SPT 来更新路由。

在路由计算中，叶子代表路由，节点则代表路由器。如果 I-SPF 计算后的 SPT 改变，PRC 会只处理那个变化的节点上的所有叶子；如果经过 I-SPF 计算后的 SPT 并没有变化，则 PRC 只处理变化的叶子信息。

比如一个节点使能一个 IS-IS 接口，则整个网络拓扑的 SPT 是不变的，这时 PRC 只更新这个节点的接口路由，从而节省 CPU 占用率。

PRC 和 I-SPF 配合使用可以将网络的收敛性能进一步提高，它是原始 SPF 算法的改进，所以已经代替了原有的算法。

说明

在设备的实现中，使用 I-SPF 和 PRC 作为 IS-IS 路由计算的唯一算法。

LSP 快速扩散

当 IS-IS 收到其它路由器发来的 LSP 时，如果此 LSP 比本地 LSDB 中相应的 LSP 要新，则更新 LSDB 中的 LSP，并用一个定时器定期将 LSDB 内已更新的 LSP 扩散出去。

LSP 快速扩散特性改进了这种方式，配置此特性的设备收到一个或多个比较新的 LSP 时，在路由计算之前，先将小于指定数目的 LSP 扩散出去，加快 LSDB 的同步过程。这种方式在很大程度上可以提高整个网络的收敛速度。

智能定时器

改进了路由算法后，如果触发路由计算的时间间隔较长，同样会影响网路的收敛速度。使用毫秒级定时器可以缩短这个间隔时间，但如果网络变化比较频繁，又会造成过度占用 CPU 资源。SPF 智能定时器既可以对少量的外界突发事件进行快速响应，又可以避免过度的占用 CPU。

通常情况下，一个正常运行的 IS-IS 网络是稳定的，发生大量的网络变动的几率很小，IS-IS 不会频繁的进行路由计算，所以第一次触发的时间可以设置的非常短（毫秒级）。如果拓扑变化比较频繁，智能定时器会随着计算次数的增加，间隔时间也会逐渐延长，避免占用大量的 CPU 资源。

与 SPF 智能定时器类似的还有 LSP 生成智能定时器。在 IS-IS 协议中，当 LSP 生成定时器到期时，系统会根据当前拓扑重新生成一个自己的 LSP。原有的实现机制是采用间隔时间定长的定时器，不能同时满足快速收敛和低 CPU 占用率的需要。为此将 LSP 生成定时器也设计成智能定时器，使其可以对于突发事件（如接口 Up/Down）快速响应，加快网络的收敛速度。同时，当网络变化频繁时，智能定时器的间隔时间会自动延长，避免过度占用 CPU 资源。

5.4.5 IS-IS 按优先级收敛

IS-IS 按优先级收敛是指在大量路由情况下，能够让某些特定的路由优先收敛的一种技术。通过对不同的路由配置不同的收敛优先级，达到重要的路由先收敛的目的，提高网络的可靠性。

IS-IS 按优先级收敛能够让某些特定的路由（例如匹配指定 IP 前缀的路由）优先收敛，因此用户可以把和关键业务相关的路由配置成相对较高的优先级，使这些路由更快的收敛，从而使关键的业务收到的影响减小。

5.4.6 IS-IS LSP 分片扩展

当 IS-IS 要发布的链路状态协议数据报文 PDU（Protocol Data Unit）中的信息量变大时，以同一系统的多个 LSP 分片的形式发布。

在 RFC3786 中规定，IS-IS 配置虚拟的 SystemID，并生成虚拟 IS-IS 的 LSP 报文，在这些 LSP 报文中携带路由等信息。

IS-IS LSP 分片扩展特性可使 IS-IS 路由器生成更多的 LSP 分片，用来携带更多的 IS-IS 信息。

术语

- 初始系统（Originating System）

  初始系统是实际运行 IS-IS 协议的路由器。允许一个单独的 IS-IS 进程像多个虚拟路由器一样发布 LSP，而“Originating System”指的是那个“真正”的 IS-IS 进程。

- 系统 ID（Normal System-ID）

  初始系统的系统 ID。

- 附加系统 ID（Additional System-ID）

附加系统 ID 由网络管理器分配。每个附加系统 ID 都允许生成 256 个额外的或扩展的 LSP 分片。附加系统 ID 和普通系统 ID 一样，在整个路由域中必须唯一。

- 虚拟系统（Virtual System）

  由附加系统 ID 标识的系统，用来生成扩展 LSP 分片。这些分片在其 LSP ID 中携带附加系统 ID。

原理

IS-IS LSP 分片由 LSP ID 的 LSP Number 字段进行标识，这个字段的长度是 1 字节，因此，一个 IS-IS 进程最多可产生 256 个分片，携带的路由数量有限（长度 1497 时只能带三万左右的路由）。通过分片扩展可以达到携带更多信息的目的。

每个系统 ID 代表一个虚拟系统，每个虚拟系统都可生成 256 个 LSP 分片。通过增加附加的系统 ID（最多可配置 50 个虚拟系统），IS-IS 进程可最多可生成 13056 个 LSP 分片。

配置虚拟系统和分片扩展后，IS-IS 会将初始系统发布的 LSP 报文无法装下的内容，放入到虚拟系统的 LSP 中发出，并通过特定的 TLV 来告知别的路由器虚拟系统和自己的关系。

IS Alias ID TLV

RFC3786 中规定了一种特殊的 TLV：IS Alias ID TLV。

表 5-5 IS Alias ID TLV

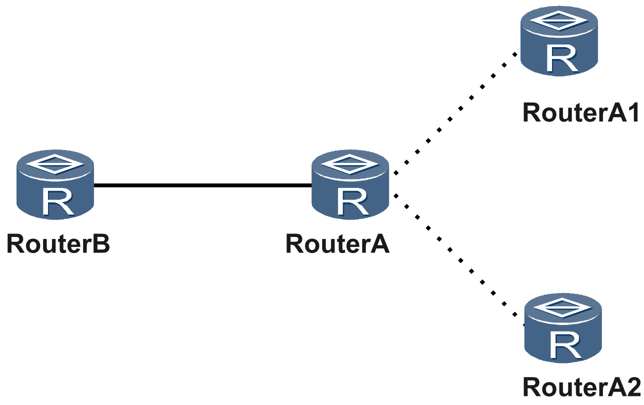
字段名	长度	含义
Type	1 字节	TLV 的类型。值为 24 表示 IS Alias ID TLV。
Length	1 字节	TLV 值的长度。
System ID	6 字节	System ID。
Pseudonode number	1 字节	pseudonode number。
sub-TLVs length	1 字节	sub-TLVs length。
sub-TLVs	0 ～ 247 字节	sub-TLVs

无论在哪种方式下，初始系统和虚拟系统的 LSP 零分片中，都必须包含 IS Alias ID TLV 来表示初始系统是谁。

操作模式

IS-IS 路由器可以在两种模式下运行 LSP 分片扩展特性：

图 5-18 IS-IS LSP 分片扩展



- Mode-1 方式：

  用于网络中的部分路由器不支持 LSP 分片扩展特性的情况。

  在该模式下，虚拟系统参与路由 SPF 计算，初始系统发布的 LSP 中携带了到每个虚拟系统的链路信息。类似地，虚拟系统发布的 LSP 也包含到初始系统的链路信息。这样，在网络中虚拟系统看起来与初始系统相连的真实路由器是一样的。

  这种方式是为了兼容不支持分片扩展的老版本所做的一个过渡模式。在老版本中，IS-IS 无法识别 Alias ID TLV，所以虚拟系统的 LSP 必须表现的像一个普通 IS-IS 发出的报文。

  虚拟系统的 LSP 中包含和原 LSP 中相同的区域地址和 overload bit。如果还有其它特性的 TLV，也必须保持一致。

  虚拟系统所携带的邻居信息指向初始系统，metric 为最大值减 1；初始系统所携带的邻居信息指向虚拟系统，metric 必须为 0。这样就保证了其它路由器在进行路由计算的时候，虚拟系统一定会成为初始系统的下游节点。

  如图 5-18 所示，RouterB 是不支持分片扩展的路由器，RouterA 设置为 mode-1 的分片扩展，RouterA1 和 RouterA2 是 RouterA 的虚拟系统，RouterA 将一部分路由信息放入 RouterA1 和 RouterA2 的 LSP 报文中向外发送。RouterB 收到 RouterA，RouterA1 和 RouterA2 的报文时，认为对端有三台独立的路由器，并进行正常的路由计算。同时 RouterA 到 RouterA1 和 RouterA2 的开销都是 0，所以，RouterB 到 RouterA 的路由开销值与 RouterB 到 RouterA1 路由开销值都相等。

- Mode-2 方式

  用于网络中所有路由器都支持 LSP 分片扩展特性的情况。在该模式下，虚拟系统不参与路由 SPF 计算，网络中所有路由器都知道虚拟系统生成的 LSP 实际属于初始系统。

  在 Mode-2 方式下工作的 IS-IS，可以识别 IS Alias ID TLV 的内容，并作为计算树和路由的依据。

  如图 5-18 所示，RouterB 支持分片扩展，RouterA 设置为 Mode-2 的分片扩展，RouterA 将一部分路由信息放入到 RouterA1 和 RouterA2 的 LSP 报文中向外发送。当 RouterB 收到 RouterA1 和 RouterA2 的 LSP 时，通过 IS Alias ID TLV 知道他们的初始系统是 RouterA，则把 RouterA1，RouterA2 所发布的信息都视为 RouterA 的信息。

无论是配置了哪种 Mode，都可以解析出任何一种 Mode，但对于不支持分片扩展的路由器，只有 Mode-1 的报文能正常解析。

表 5-6 Mode-1 和 Mode-2 的比较

发送报文内容\配置的模式	Mode-1	Mode-2
IS Alias ID	Yes	Yes
area	Yes	No
overload bit	Yes	Yes
IS NBR/IS EXTENDED NBR	Yes	No
路由	Yes	Yes
ATT bits	must 0	must 0
P bit	must 0	must 0

基本流程

配置分片扩展后，如果存在由于报文装满而丢失的信息，系统会提醒重启 IS-IS。重启之后，初始系统会尽最大能力装载路由信息，装不下的信息将放入虚拟系统的 LSP 中发送出去。

组网应用

说明

如果网络上还有其他厂商的设备，配置分片扩展必须配置成 Mode-1，否则其他设备无法识别。

建议先配置分片扩展和虚拟系统，然后将 IS-IS 建立邻居或者引入路由。如果先让 IS-IS 携带大量信息，256 个分片无法装下，再配置分片扩展和虚拟系统，需要重启 IS-IS 才可以让配置生效，所以要谨慎。

5.4.7 IS-IS 管理标记

管理标记特性允许在 IS-IS 域中通过管理标记对 IP 地址前缀进行控制。

管理标记用来携带关于 IP 地址前缀的管理信息，可以达到简化管理。其用途包括控制不同级别和不同区域间的路由引入，各种路由协议，以及同一路由器上运行的 IS-IS 多实例。

管理标记值与某些属性相关联。当 cost-sytle 为 wide、wide-compatible 或 compatible 时，如果发布可达的 IP 地址前缀具有该属性，IS-IS 会将管理标记加入到该前缀的 IP 可达信息 TLV 中。这样，管理标记就会随着前缀发布到整个路由域。

5.4.8 IS-IS 动态主机名交换

动态主机名交换机制（Dynamic Hostname Exchange Mechanism）为运行 IS-IS 协议的路由器提供了一种从主机名到 System ID 映射的服务。

IS-IS 最早是 ISO 为 CLNS（Connectionless Network Service）而设计的动态路由协议，因而保留了独特的地址编码方式。

在没有实现/使能主机名交换特性的运行 IS-IS 协议的路由器上，查看 IS-IS 邻居和链路状态数据库等信息时，IS-IS 域中的各路由器（IS）都是用由 12 位十六进制数组成的 System ID 来表示的，例如：aaaa.eeee.1234。这种表示方法比较繁琐，而且易用性不好。

动态主机名交换机制就是为了方便对 IS-IS 网络的维护和管理而引入的。

IS-IS 动态主机名的信息在 LSP 中以一个动态主机名 TLV（137 号）的形式发布。这个机制同时还提供将主机名与广播网中的 DIS 相关联的服务，并将此信息通过 LSP 以动态主机名 TLV 的形式发布出去。

在 AR1200 的实现中，使能了 IS-IS 动态主机名映射的路由器，在其生成的 LSP 中添加 Dynamic Hostname TLV （TLV type 137），记录本地主机名并发布出去。

动态主机名交换机制 TLV（137 号）包含的内容如下：

- Type：动态主机名交换机制。
- Length：Value 字段的总长度。
- Value：1 ～ 255 字节的字符串。

动态主机名的 TLV 是可选的，它可以存在于 LSP 中的任何位置。主机名的值不能为空。路由器在发送 LSP 的时候可以决定是否携带该 TLV，接收端的路由器可以决定是否忽略该 TLV，或者提取该 TLV 的内容放在自己的映射表中。

基本实现

- 匹配原则

  动态主机名采用最长匹配原则，即先匹配 System ID+NSEL，若不能匹配则匹配 System ID。
- 动态主机名的传输

  动态主机名只能存在于 Original LSP 中。
- DIS 动态主机名的传输

  DIS 动态主机名在 DIS 产生的 LSP 中传输。
- 动态主机名的优先级

  动态主机名的优先级高于静态主机名。当两者都配置时，由动态主机名代替静态主机名。
- 动态主机名的配置与解析

  动态主机名最长支持配置 64 字节长，可以解析最大 255 字节长度的内容。

组网应用

在维护和管理中，使用主机名比使用 System ID 会更直观，也更容易记忆。配置此功能后，当在路由器上查看 IS-IS 相关信息时，看到的是路由器的主机名，而不再是 System ID。

在 AR1200 的实现中，主机名交换特性包括动态主机名映射和静态主机名映射两个主要功能。在下列三种情况下会将 System ID 替换为主机名显示：

- 显示 IS-IS 邻居时，将 IS-IS 邻居的 System ID 替换为其动态主机名。如果该邻居为 DIS，则 DIS 的 System ID 也替换为该邻居的动态主机名。
- 显示 IS-IS 链路状态数据库中的 LSP 时，将 LSP ID 中的 System ID 替换为发布该 LSP 的路由器的动态主机名。
- 显示 IS-IS 链路状态数据库的详细信息时，对于使能了动态主机名交换的 IS 的 LSP，会增加显示 Host Name 字段，而 IS 字段显示内容中的 System ID 也将替换为该的 IS-IS 邻居的动态主机名。

5.4.9 IS-IS 三次握手机制（3-Way HandShake）

IS-IS 协议在点到点链路上，增加 3 次握手机制，提升链路层的可靠性。

ISO 10589 中的 IS-IS 的 2 次握手机制（2-Way Handshake）使用 Hello 报文来建立相邻路由器间点到点链路的邻接关系。只要路由器收到对端发来的 Hello 报文，就宣布邻居为 Up 状态，建立邻接关系。这种机制存在明显缺陷。

当路由器间存在两条及以上的链路时，如果某条链路上到达对端的单向状态为 Down，而另一条链路同方向的状态为 Up，路由器之间还是能建立起邻接关系。SPF 在计算时会使用另一条链路上的参数，这就导致没有检测到故障的路由器在转发报文时仍然试图通过状态为 Down 的链路。

三次握手机制解决了上述不可靠点到点链路中存在的问题。这种方式下，路由器只有在知道邻居路由器也接收到它的报文时，才宣布邻居路由器处于 Up 状态，从而建立邻接关系。

同时，三次握手机制中使用 32 比特的扩展 Circuit ID，打破了目前由本地 8 比特 Circuit ID 字段限制的 255 个点到点链路。

说明

缺省情况下，IS-IS 在点到点链路上执行三次握手特性。（RFC3373）

5.4.10 IS-IS GR

IS-IS GR（Graceful Restart）是指为了实现不间断转发，通过对 IS-IS 做扩展，以支持 GR 能力的高可靠性技术（HA，High Availability）。RFC3847 制定了 IS-IS GR 规范。

IS-IS 是一种链路状态路由协议，需要同一个区域内的每一台路由器都保持完全一致的网络拓扑信息，即完全一致的链路状态数据库（LSDB）。

路由器发生主备倒换后，由于没有保存任何重启前的邻居信息，因此一开始发送的 Hello 报文中不包含邻居列表。此时邻居路由器收到后，执行 2-way 邻居关系检查，发现在重启路由器的 Hello 报文的邻居列表中没有自己，这样邻居关系将会断掉。

同时，邻居路由器通过生成新的 LSP 报文，将拓扑变化的信息泛洪给区域内的其它路由器。区域内的其他路由器会基于新的链路状态数据库进行路由计算，从而造成路由中断或者路由环路。

由于没有保存重启前的任何链路状态信息（LSDB），重启路由器在主备倒换后，需要快速和邻居间同步链路状态信息。因此，IS-IS 协议若不以 GR（Graceful Restart）方式重启，则会重置 IS-IS 邻居关系，重新生成 LSP 和泛洪 LSP，进而在整个区域引发 SPF 计算，引起整个区域的路由震荡和转发中断。

IETF 针对这种情况为 IS-IS 制定了 GR 规范（RFC3847），对保留 FIB 表和不保留 FIB 表的协议重启都进行了处理，避免协议重启带来的路由震荡和流量转发中断的现象。

路由器故障后，其路由协议层面的邻居会检测到它们之间的邻居关系 Down 掉，过一段时间再次 Up，这个过程被称之为邻居关系震荡。邻居关系的震荡将最终导致路由震荡，使得重启路由器在一段时间内出现路由黑洞，或者导致邻居将数据业务从重启路由器处环路，从而导致网络的可靠性大大降低。GR 的目标就是为了解决上述路由震荡的问题。

IS-IS GR 基本概念

IS-IS GR 过程由 GR-Restarter 和 GR-Helper 配合完成。

- GR-Restarter

  具备 GR 能力、并将要进行 GR 的路由器。

- GR-Helper

  用于辅助 GR 路由器完成 GR 功能的另外一个具备 GR 能力的路由器。GR-Restarter 一定具有 GR-Helper 的能力。

  说明

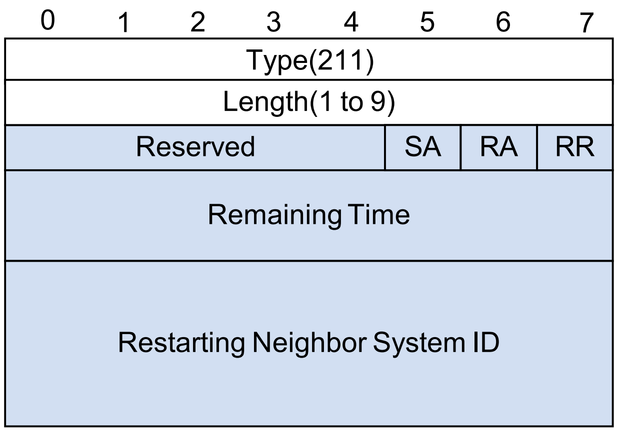
  AR1200 只能作为 Helper 路由器，不能作为 Restarter 路由器。

为了实现 GR，IS-IS 引入 Restart TLV（Type-Length-Value）和 T1、T2、T3 定时器。

Restart TLV

Restart TLV 是包含在 IIH（IS-to-IS Hello PDUs）报文中的扩展部分。支持 IS-IS GR 能力的路由器的所有 IIH 报文都包含 Restart TLV。Restart TLV 中携带了协议重启的一些参数。其报文格式如图 5-19 所示。

图 5-19 Restart TLV 格式



Restart TLV 各字段的含义如表 5-7 所示。

表 5-7 Restart TLV 报文字段含义

字段名	长度	含义
Type	1 字节	TLV 的类型。值为 211 表示是 Restart TLV。
Length	1 字节	TLV 值的长度。
RR	1 比特	重启请求位（Restart Request）。路由器发送的 RR 置位的 Hello 报文用于通告邻居自己发生 Restarting/Starting，请求邻居保留当前的 IS-IS 邻接关系并返回 CSNP 报文。
RA	1 比特	重启应答位（Restart Acknowledgement）。路由器发送的 RA 置位的 Hello 报文用于通告邻居确认收到了 RR 置位的报文。

字段名	长度	含义
SA	1 比特	抑制发布邻接关系位（Suppress adjacency advertisement）。用于发生 Starting 的设备请求邻居抑制与自己相关的邻居关系的广播，以避免路由黑洞。
Remaining Time	2 字节	邻居保持邻接关系不重置的时间。长度是 2 字节，单位是秒。当 RA 置位时，这个值是必需的。
Restarting Neighbor System ID	6 字节	回应重启应答报文的邻居路由器的 System ID。

定时器

IS-IS 的 GR 能力扩展中，引入了三个定时器，分别是 T1、T2 和 T3。

- T1

  使能了 IS-IS GR 特性的进程，在每个接口都会维护一个 T1 定时器。在 Level-1-2 路由器上，广播网接口为每个 Level 维护一个 T1 定时器。

  如果 GR Restarter 已发送 RR 置位的 IIH 报文，但直到 T1 定时器超时还没有收到 GR Helper 的包含 Restart TLV 且 RA 置位的 IIH 报文的确认消息时，会重置 T1 定时器并继续发送包含 Restart TLV 的 IIH 报文。

  当收到确认报文或者 T1 定时器已超时 3 次时，取消 T1 定时器。T1 定时器缺省设置为 3 秒。

- T2

  Level-1 和 Level-2 的 LSDB 各维护一个 T2 定时器。

  T2 是系统等待各层 LSDB 同步的最长时间，一般情况下为 60 秒。

- T3

  整个系统维护一个 T3 定时器。

  T3 定时器可理解为成功完成 GR 所允许的最大时间。

  T3 定时器超时表示 GR 失败。

  T3 定时器的初始值为 65535 秒，但在收到邻居回应的 RA 置位的 IIH 报文后，取值会变为各个 IIH 报文的 Remaining time 字段值中的最小者。

  T3 定时器只用于 Restarting 设备。

IS-IS GR 的会话机制

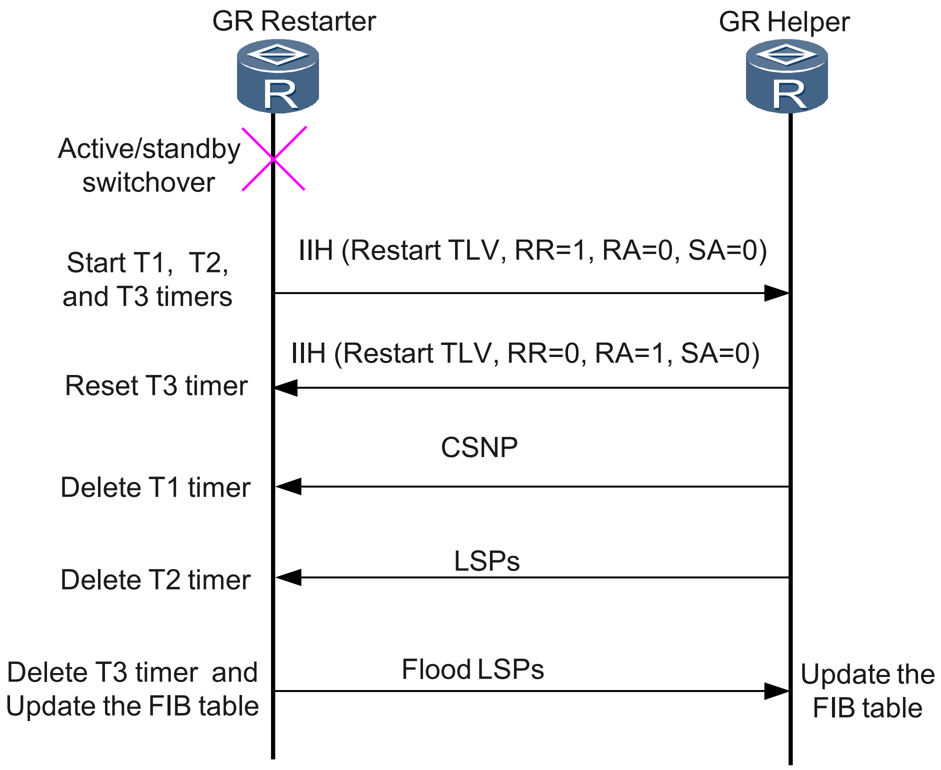
为了以示区别，主备倒换和重启 IS-IS 进程触发的 GR 过程称为 Restarting，FIB 表保持不变。路由器重启触发的 GR 过程称为 Starting，进行 FIB 表更新。

下面分 Restarting 和 Starting 两种情况说明 IS-IS GR 的详细过程。

IS-IS Restarting

IS-IS Restarting 的过程如**图 5-20** 所示。

图 5-20 IS-IS Restarting 过程



1. GR Restarter 进行协议重启后，GR Restarter 进行如下操作：
   - 启动 T1、T2 和 T3 定时器。
   - 从所有接口发送包含 Restart TLV 的 IIH 报文，其中 RR 置位，RA 和 SA 位清除。
2. GR Helper 收到 IIH 报文以后，进行如下操作：
   - GR Helper 维持邻居关系，刷新当前的 Holdtime。
   - 回送一个包含 Restart TLV 的 IIH 报文（RR 清除，RA 置位，Remaining time 是从现在到 Holdtime 超时的时间间隔）。
   - 发送 CSNP 报文和所有 LSP 报文给 GR Restarter。

   说明

   - *在点到点链路上，邻居必须发送 CSNP。*
   - *在 LAN 链路上，是 DIS 的邻居才发送 CSNP 报文，如果重启的是 DIS，则在 LAN 中的其它路由器中选举一个临时的 DIS。*

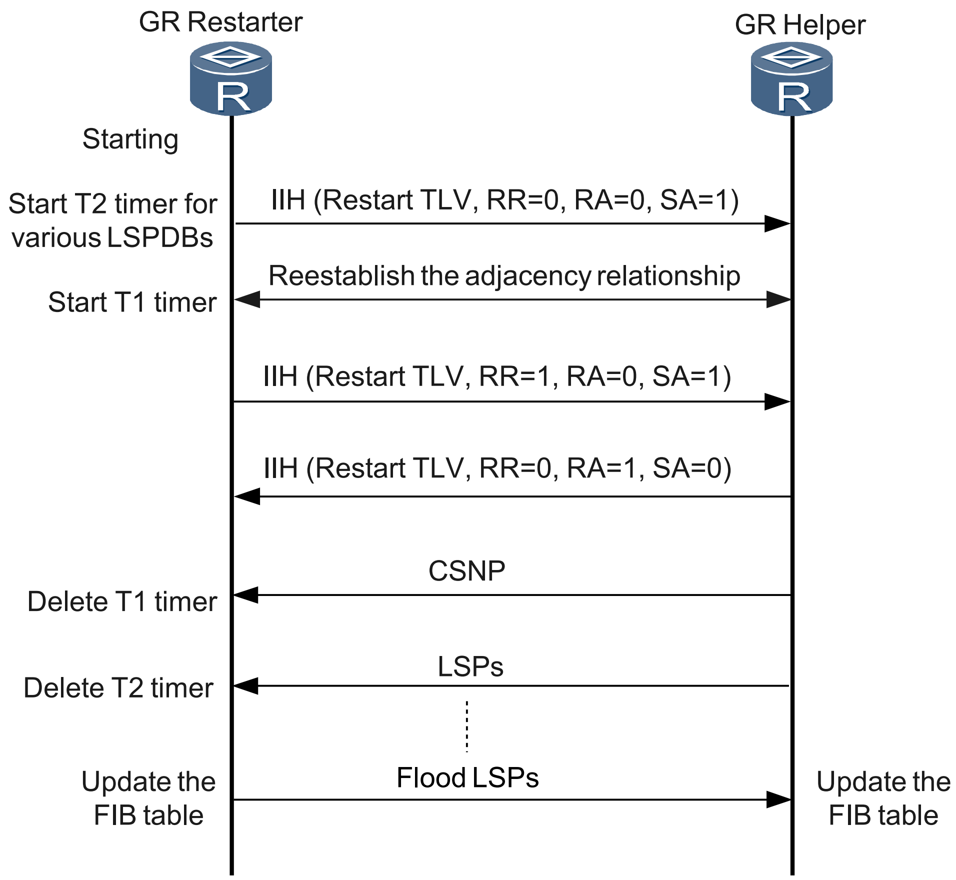
   如果 GR Helper 不支持 GR，就忽略 Restart TLV，按正常的 IS-IS 过程处理，重置和 GR Restarter 的邻接关系。
3. GR Restarter 接收到邻居的 IIH 回应报文（RR 清除、RA 置位），做如下处理：
   - 把 T3 的当前值和报文中 Remaining time 比较，取其中较小者作为 T3 的值。
   - 在接口收到确认报文和 CSNP 报文之后，取消该接口的 T1 定时器。
   - 如果该接口没有收到确认报文和 CSNP 报文，T1 会不停地重置，重发含 Restart TLV 的 IIH 报文。如果 T1 超时次数超过阈值，GR Restarter 强制取消 T1 定时器，启动正常的 IS-IS 处理流程。

4. 当 GR Restarter 所有接口上的 T1 定时器都取消，CSNP 列表清空并且收集全所有的 LSP 报文后，可以认为和所有的邻居都完成了同步，取消 T2 定时器。
5. T2 定时器被取消，表示本 Level 的 LSDB 已经同步。
   - 如果是单 Level 系统，则直接触发 SPF 计算。
   - 如果是 Level-1-2 系统，此时判断另一个 Level 的 T2 定时器是否也取消。如果两个 Level 的 T2 定时器都被取消，那么触发 SPF 计算，否则等待另一个 Level 的 T2 定时器超时。
6. 各层的 T2 定时器都取消后，GR Restarter 取消 T3 定时器，更新 FIB 表。GR Restarter 可以重新生成各层的 LSP 并泛洪，在同步过程中收到的自己重启前生成的 LSP 此时也可以被删除。
7. 至此，GR Restarter 的 IS-IS Restarting 过程结束。

IS-IS Starting

对于 Starting 设备，因为没有保留 FIB 表项，所以一方面希望在 Starting 之前和自己的邻接关系为“Up”的邻居重置和自己的邻接关系，同时希望邻居能在一段时间内抑制和自己的邻接关系的发布。其处理过程和 Restarting 不同，具体如图 5-21 所示。

图 5-21 IS-IS Starting 过程



1. GR Restarter Starting 后，进行如下操作：
   - 为每层 LSDB 的同步启动 T2 定时器。

- 从各个接口发送携带 Restart TLV 的 IIH 报文，其中 RR 位清除，SA 位置位。

  RR 位清除表示是 Starting 完成。

  SA 位置位则表示希望邻居在收到 SA 位清除的 IIH 报文之前，一直抑制和自己的邻接关系的发布。

2. 邻居收到携带 Restart TLV 的 IIH 报文，根据路由器是否支持 GR，进行如下处理。
   - 支持 GR

     重新初始化邻接关系。

     在发送的 LSP 中取消和 GR Restarter 邻接关系的描述，进行 SPF 计算时也不考虑和 GR Restarter 相连的链路，直到收到 SA 位清除的 IIH 为止。
   - 不支持 GR

     邻居忽略 Restart TLV，重置和 GR Restarter 之间的邻接关系。

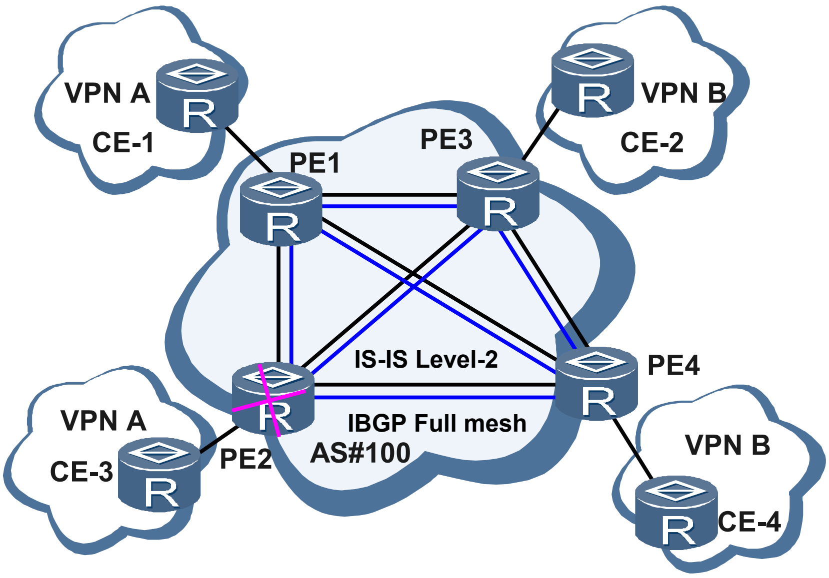
     回应一个不含 Restart TLV 的 IIH 报文，转入正常的 IS-IS 处理流程。这时不会抑制和 GR Restarter 的邻接关系的发布。在点到点链路上，还会发送一个 CSNP 报文。
3. 邻接关系重新初始化之后，在每个接口上 GR Restarter 都和邻居重建邻接关系。当有一个邻接关系到达 Up 状态后，GR Restarter 为该接口启动 T1 定时器。
4. 在 T1 定时器超时之后，GR Restarter 发送 RR 置位、SA 置位的 IIH 报文。
5. 邻居收到 RR 置位和 SA 置位的 IIH 报文后，发送一个 RR 清除、RA 置位的 IIH 报文作为确认报文，并发送 CSNP 报文。
6. GR Restarter 收到邻居的 IIH 确认报文和 CSNP 报文以后，取消 T1 定时器。

   如果没有收到 IIH 报文或者 CSNP 报文，就不停重置 T1 定时器，重发 RR 置位、SA 置位的 IIH 报文。如果 T1 超时次数超过阈值，GR Restarter 强制取消 T1 定时器，进入正常的 IS-IS 处理流程完成 LSDB 同步。
7. GR Restarter 收到 Helper 端的 CSNP 以后，开始同步 LSDB。
8. 本 Level 的 LSDB 同步完成后，GR Restarter 取消 T2 定时器。
9. 所有的 T2 定时器都取消以后，启动 SPF 计算，重新生成 LSP，并泛洪。
10. 至此，GR Restarter 的 IS-IS Starting 过程完成。

组网应用

在运营商网络的边缘，即 PE（Provider Edge）是典型的 GR（不间断转发）应用场所，特别是用户单点（Single point）连入运营商网络的情况。当单点 PE 出现故障，或者出于维护目的（比如升级软件版本）导致 PE 主备倒换发生，如果部署了 GR，则能够给用户的关键业务提供不间断转发的高可靠性保障。具体如**图 5-22** 所示。

图 5-22 GR 在运营商网络中的应用



说明

在 PE2 上部署 NSF 防止单点故障，需同时使能 IS-IS GR、BGP GR 和 LDP GR。

在 PE 上，应用 IS-IS、BGP 和 LDP GR 等协议，同时，在 P 设备上，应用 IS-IS、LDP GR 等协议。PE、P 设备均具备双主控冗余备份能力。

5.4.11 IS-IS for IPv6

IETF 的 draft-ietf-isis-ipv6.txt 中规定了 IS-IS 为支持 IPv6 所新增的内容。支持 IPv6 路由的处理和计算。主要是新添加的支持 IPv6 路由信息的两个 TLVs（Type-Length-Values）和一个新的 NLPID（Network Layer Protocol Identifier）。

新增的两个 TLV 分别是：

- IPv6 Reachability

  类型值为 236（0xEC），通过定义路由信息前缀、度量值等信息来说明网络的可达性。

- IPv6 Interface Address

  类型值为 232（0xE8），它相当于 IPv4 中的“IP Interface Address”TLV，只不过把原来的 32 比特的 IPv4 地址改为 128 比特的 IPv6 地址。

NLPID 是标识网络层协议报文的一个 8 比特字段，IPv6 的 NLPID 值为 142（0x8E）。如果 IS-IS 支持 IPv6，那么向外发布 IPv6 路由时必须携带 NLPID 值。

5.4.12 IS-IS Wide Metric

在 RFC3784 中规定，扩展后的接口 Metirc 可以配置到 16777215，路由的 Metric 可达 4261412864。

在大型网络设计中，较小的 Metric 范围不能满足需求。同时，为了支持 IS-IS TE 特性，所以采用了 Wide-Metric。

在早期的 ISO10589 中，接口下最大只能配置值为 63 的 Metric。使用 128 号和 130 号 TLV 作为携带路由的 TLV，使用 2 号 TLV 作为携带邻居信息的 TLV。

在使能 IS-IS Wide-Metric 后，使用 135 号 TLV 作为携带路由的 TLV，并使用 22 号 TLV 作为携带邻居信息的 TLV。

- narrow 模式下使用的 TLV：
  - IP Internal Reachability TLV：用来携带域内的路由。
  - IP External Reachability TLV：用来携带域外的路由信息。
  - IS Neighbors TLV：用来携带邻居信息。
- wide 模式下使用的 TLV：
  - Extended IP Reachability TLV：用来替换原有的 IP reachability TLV，携带路由信息，它扩展了路由开销值的范围（4 字节），并可以携带 sub TLV。
  - IS Extended Neighbors TLV：用来携带邻居信息。

说明

wide 模式的 IS-IS 和 narrow 模式下的 IS-IS 不可实现互通。如果需要互通，就必须修改模式，让网络上所有路由器都可以接收其他路由器发的所有报文。

表 5-8 接收和发送的模式详细列表

模式\收发	接收	发送
narrow	narrow	narrow
narrow-compatible	narrow&wide	narrow
compatible	narrow&wide	narrow&wide
wide-compatible	narrow&wide	wide
wide	Wide	wide

当配置模式为 compatible 的时候，会按照 narrow 模式和 wide 模式分别发送一份信息。

流程

注意

当修改 cost-style 后，会导致 IS-IS 进程的重启，所以一定要谨慎。

- 如果发送模式由 narrow 变成 wide

  原来由 128，130 和 2 号 TLV 携带的信息，变化成 135 和 22 号 TLV 携带。
- 如果发送模式由 wide 变成 narrow

  原来由 135 和 22 号 TLV 携带的信息，变化成 128，130 和 2 号 TLV 携带。

- 如果发送模式由 narrow/wide 变成 narrow&wide

  由原来的信息，变化成 128，130，2 号 TLV 和 135 和 22 号 TLV 同时携带。

组网应用

- 当使用 IS-IS TE 的时候，必须先使能 IS-IS Wide Metric。
- 当需要应用管理标记特性时，必须先使能 IS-IS Wide Metric。（参见 IS-IS 管理标记特性）

5.4.13 IS-IS LDP 联动

在存在主备链路的网络中，当主链路故障恢复后，流量会从备份链路切换到主链路。

由于 IGP 的收敛在 LDP 会话建立之前完成，导致旧的 LSP 已经删除，新的 LSP 还没有建立，因此 LSP 流量中断。

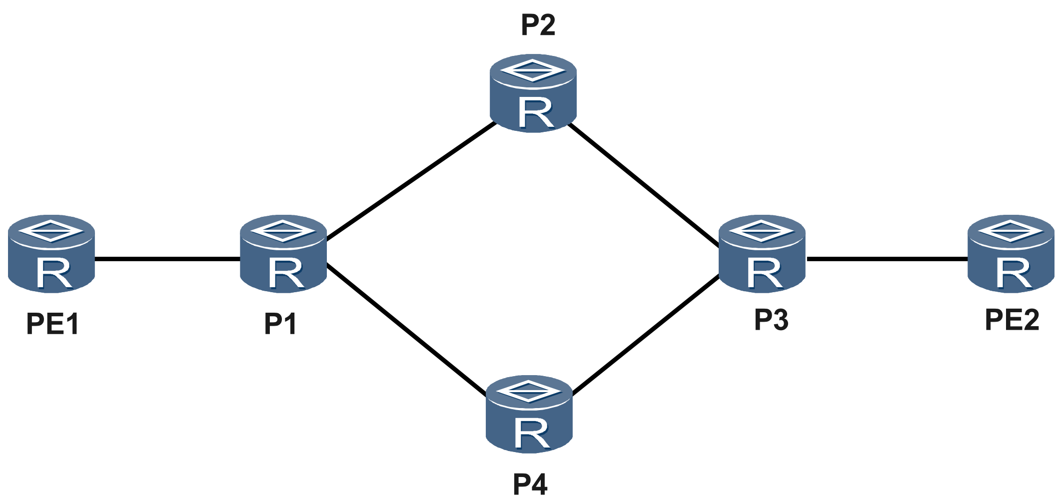
如图 5-23 所示，PE1-P1-P2-P3-PE2 为主链路，PE1-P1-P4-P3-PE2 为备份链路。

主链路发生故障，流量从主链路切换到备份链路。主链路故障恢复，流量从备份链路回切到主链路，此时流量会有较长时间的中断。

说明

此特性中提到的 LSP 是 Lable Switch Path 的缩写。

图 5-23 IS-IS LDP 联动



通过在 P1 和 P2 上配置 LDP 和 IGP 同步功能，能够缩短流量从备份链路切换到主链路时的中断时间。

解决 LDP 回切丢包问题的一个方法是 LDP-IGP 联动，即 IGP 推迟路由的回切，直至 LDP 完成收敛。即在新的 LSP 没有收敛之前，保持老的 LSP，让流量继续从老 LSP 路径转发，直至新的 LSP 建立成功，再删除老的 LSP。

图 5-24 LDP-IGP 联动状态机

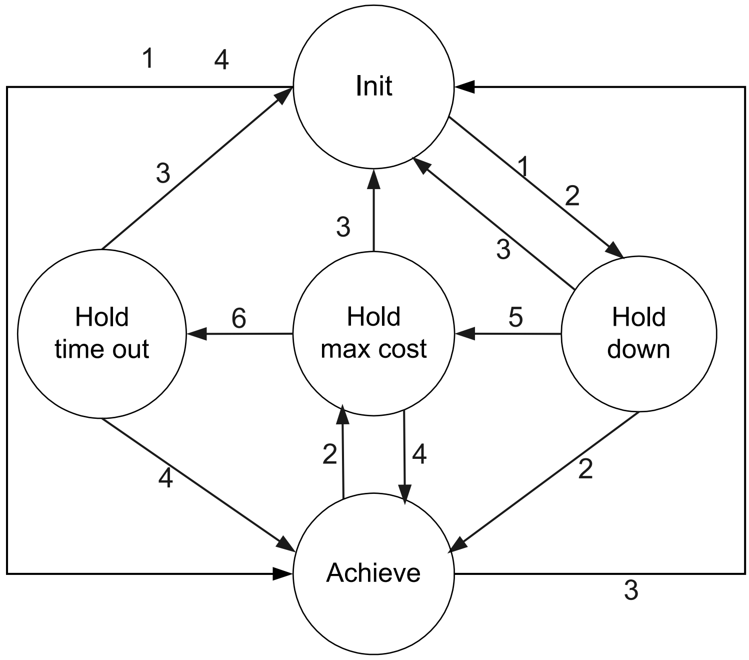


图 5-24 中各个数字含义如下：

- 1 表示接口状态为 Up。
- 2 表示 LDP 会话状态为 Down。
- 3 表示接口状态为 Down。
- 4 表示 LDP 会话状态为 Up。
- 5 表示 LDP 路由不可达或 Hold Down Timer 超时。
- 6 表示 Hold max cost Timer 超时。
- 状态描述
  - Init 状态：是 LDP-IGP 联动的初始状态。
  - Holdtimeout 状态：是指接口超时状态。
  - Holddown 状态：是指接口抑制状态。当接口处于 Holddown 状态时，会抑制 Hello 报文的接收和发送。
  - HoldMaxCost 状态：是指接口发布最大开销的状态。
  - Achieve 状态：是 LDP-IGP 联动同步状态。
- 状态迁移描述
  - 当接口处于 Init 状态时，且 LDP 会话状态为 Down，如果收到接口 Up 消息，则接口状态会迁移到 HoldDown 状态。
  - 当接口处于 Init 状态时，且 LDP 会话状态为 Up，如果收到接口 Up 消息，则接口状态会迁移到 Achieve 状态。
  - 当接口处于 Holdtimeout 状态时，如果收到接口 Down 消息，则接口状态会迁移到 Init 状态。
  - 当接口处于 Holdtimeout 状态时，如果收到 LDP Session Up 消息，则接口状态会迁移到 Achieve 状态。

- 当接口处于 HoldMaxCost 状态时，如果收到 LDP Session Up 消息，则接口状态迁移到 Achieve 状态。
- 当接口处于 HoldMaxCost 状态时，如果收到接口 Down 消息，则接口状态迁移到 Init 状态。
- 当接口处于 HoldMaxCost timer 超时状态时，且没有收到 LDP Session UP 消息，则接口状态迁移到 Holdtimeout 状态。
- 当接口处于 HoldDown 状态时，如果收到 LDP Session Down 消息，则接口状态迁移到 Achieve 状态。
- 当接口处于 HoldDown 状态时，如果收到接口 Down 消息，则接口状态迁移到 Init 状态。
- 当接口处于 HoldDown 状态时，如果收到 Hold Down Timer 超时消息，则接口状态迁移到 HoldMaxCost 状态。
- 当接口处于 Achieve 状态时，如果收到 LDP Session Down 消息，则接口状态迁移到 HoldMaxCost 状态。
- 当接口处于 Achieve 状态时，如果收到接口 Down 消息，则接口状态迁移到 Init 状态。

组网应用

在图 **5-23** 的典型组网中，当链路回切时，为避免 LDP 回切流量时丢包，可以配置 IS-IS LDP 联动来实现。

5.4.14 BFD for IS-IS

双向转发检测 BFD（Bidirectional Forwarding Detection）是一个简单的“Hello”协议。在很多方面，它与路由协议的邻居检测部分相似。

一对系统在它们之间的所建立会话的通道上周期性的发送检测报文，如果某个系统在检测时间内没有收到对端的检测报文，则认为在这条到相邻系统的双向通道的某个部分发生了故障。在某些条件下，为了减少负荷，系统之间的发送和接收速率需要协商。

BFD 包括静态 BFD 和动态 BFD。

说明

BFD 使用本地标识符（Local Discriminator）和远端标识符（Remote Discriminator）区分同一对系统之间的多个 BFD 会话。

- 静态 BFD

  静态 BFD 是指通过命令行手工配置 BFD 会话参数，包括了配置本地标识符和远端标识符等，然后手工下发 BFD 会话建立请求。

- 动态 BFD（包括 BFD for IPv4）

  动态 BFD 是指由路由协议动态触发 BFD 会话建立。动态 BFD 中，本地标识符是动态分配的，远端标识符是通过路由协议自学习得到。

IS-IS 动态 BFD 是指 BFD 的会话由 IS-IS 动态创建，不再依靠手工配置。当 BFD 检测到故障的时候，通过路由管理通知 IS-IS。IS-IS 进行相应邻居 Down 处理，快速发布变化的 LSP 信息和进行增量路由计算，从而实现路由的快速收敛。

通常情况下，IS-IS 设定发送 Hello 报文的时间间隔为 10 秒钟，一般将宣告邻居 Down 掉的时间（即邻居的保持时间）配置为 Hello 报文间隔的 3 倍。若在相邻路由器失效时间内没有收到邻居发来的 Hello 报文，将会删除邻居。

路由器能感知到邻居故障的时间最小为秒级。由此可能会出现高速的网络环境中大量报文丢失的问题。

双向转发检测 BFD 就是为解决现有检测机制的不足而产生的，能够提供轻负荷、快速（毫秒级）的通道故障检测。

通过配置 BFD 可以设置毫秒级的时间检测间隔。使用 BFD 并不是代替 IS-IS 协议本身的 Hello 机制，而是配合 IS-IS 协议更快的发现邻接方面出现的故障，并及时通知 IS-IS 重新计算相关路由以便正确指导报文的转发。

静态 BFD

静态 BFD 是指通过命令行手工配置 BFD 会话参数，包括了配置本地标识符和远端标识符等，然后手工下发 BFD 会话建立请求。

这种方式的缺点是建立和删除 BFD 会话时都需要手工触发，缺乏灵活性，而且有可能造成人为的配置错误，比如配置了错误的本地标识符或者远端标识符时，BFD 会话将不能正常工作。

动态 BFD

动态建立 BFD 会话指的是由路由协议动态触发 BFD 会话建立。BFD for IPv4 会话是 IS-IS 在 IPv4 邻居建立的时候触发建立的。

路由协议在建立了新的邻居关系时，根据邻居的 IP 协议类型，将对应的参数及检测参数（包括目的地址、源地址等）通告给 BFD，BFD 根据收到的参数建立起会话。动态 BFD 比静态 BFD 更具有灵活性。

路由管理模块 RM（Routing Management Module）为 IS-IS 提供与 BFD 模块交互的相关服务。IS-IS 通过 RM 通知 BFD 来动态创建或删除 BFD 会话，同时 BFD 的事件消息也通过 RM 传递给 IS-IS。

BFD 会话的建立与删除

- 创建 BFD 会话的条件
  - 各路由器配置了 IS-IS 基本功能并且在接口下使能了 IS-IS。
  - 各路由器配置了全局 BFD 功能并且使能了接口或者进程的 BFD for IPv4 特性。
  - 使能了接口或者进程的 BFD for IPv4 特性，且相邻路由器的邻居状态为 Up（广播网中须等到 DIS 选举出来）。
- 创建 BFD 会话的过程
  - P2P 网络

    满足创建 BFD 会话的条件后，IS-IS 将通过 RM 模块通知 BFD 模块直接在邻居间创建 BFD 会话。
  - 广播网络

    满足创建 BFD 会话的条件且 DIS 已经选举出来后，IS-IS 将通过 RM 模块通知 BFD 模块，DIS 与每台路由器之间都自动创建 BFD 会话。都不是 DIS 的两台路由器之间不建立 BFD 会话。

  广播网与 P2P 网络不同的是：

  虽然广播网中 IS-IS 同一网段上的同一级别的路由器之间都会形成邻接关系，即包括所有的非 DIS 路由器之间也会形成邻接关系，但在 IS-IS BFD 实现上，只在 DIS 和非 DIS 之间建立 BFD 会话，非 DIS 之间不启动 BFD 会话，而 P2P 网络直接在邻居间创建会话。

如果同一链路上的同一对路由器形成的是 Level-1-2 的类型的邻居，在广播网中 IS-IS 会针对这两个 Level 分别创建两个 BFD 会话，但在 P2P 网络中 IS-IS 只会创建一个 BFD 会话。

- 删除 BFD 会话的条件
  - P2P 网络

    当 IS-IS 在 P2P 网络接口类型上建立的邻接关系断开时（非 Up 状态），或者邻居对应的 IP 协议类型删除时，删除对应的 BFD 会话。
  - 广播网络

    当 IS-IS 在广播网络接口类型上建立的邻接关系断开（非 Up 状态），邻居对应的 IP 协议类型删除时，或者广播网络 DIS 发生变化时，删除对应的 BFD 会话。

  在接口上删除动态创建 BFD 会话的配置或者禁用了 IS-IS BFD 功能后，该接口相关的所有 Up 或 DIS Up 的邻接关系对应的 BFD 会话都被删除。

  在 IS-IS 进程下去使能全局动态 BFD 后，该进程下的所有接口的 BFD 会话都被删除。

  说明

  由于 IS-IS 只能建立单跳邻居，IS-IS BFD 只对 IS-IS 邻居间的单跳链路进行检测。

- 响应 BFD 会话 Down 事件

  当 BFD 检测到链路发生故障并产生 Down 事件时，会通知 RM。RM 通知 IS-IS 删除此邻接。IS-IS 响应这个事件并重新进行路由计算，实现网络迅速收敛。BFD for IPv4 通知 IS-IS 链路故障后，IS-IS 只改变其 IPv4 路由。

  当本地路由器与邻居路由器均为 Level-1-2 时，二者之间会针对不同的 Level 分别创建两个邻居，此时 IS-IS 也会创建两个不同 Level 的会话，在这种情况下，RM 会删除根据相应 Level 的邻接关系。

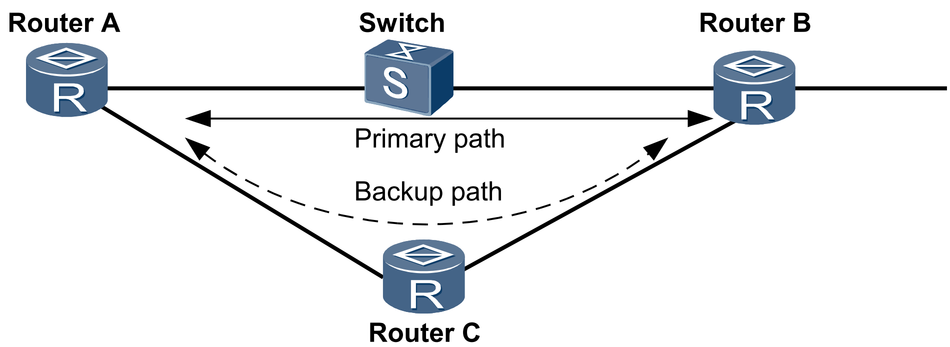
组网应用

注意

请根据网络环境配置 BFD，如果时间参数设置不当将会导致网络震荡。

BFD for IS-IS 可以快速感知链路变化实现路由收敛。

图 5-25 IS-IS BFD 组网示意图



配置要求：

- 如图 5-25 所示在各路由器上使能 IS-IS 基本功能。
- 使能全局 BFD 特性。
- 在 RouterA 和 RouterB 上使能 IS-IS BFD 检测机制。

这样，当 RouterA 和 RouterB 之间的链路故障时，BFD 能够快速检测到故障并通告给 IS-IS 协议，IS-IS Down 掉故障链路的接口邻居并删除邻接对应的 IP 协议类型，从而触发拓扑计算，同时更新 LSP 使得其他邻居（如 RouterB 的邻居 RouterC）及时收到 RouterB 的更新 LSP，实现了网络拓扑的快速收敛。

5.4.15 IS-IS 认证

IS-IS 认证是基于网络安全性的要求而实现的一种加密手段，通过在 IS-IS 报文中增加认证字段对报文进行加密。当本地路由器接收到远端路由器发送过来的 IS-IS 报文，如果发现认证密码不匹配，则将收到的报文进行丢弃，达到自我保护的目的。

根据报文的种类，认证可以分为以下三类：

- 区域认证

  在 IS-IS 进程视图下配置，对 Level-1 的 CSNP、PSNP 和 LSP 报文进行认证。
- 路由域认证

  在 IS-IS 进程视图下配置，对 Level-2 的 CSNP、PSNP 和 LSP 报文进行认证。
- 接口认证

  在接口视图下配置，对 Level-1 和 Level-2 的 Hello 报文进行认证。

根据报文的认证方式，可以分为以下三类：

- 明文认证

  这是一种简单的加密方式，将配置的密码直接加入报文中，这种加密方式安全性不够，从而产生了下面的认证方式。
- MD5 认证

  通过将配置的密码进行 MD5 算法之后再加入报文中，这样提高了密码的安全性。
- Keychian 认证

  通过配置随时间变化的密码链表来进一步提升网络的安全性。

IS-IS 通过 TLV 的形式携带认证信息，认证 TLV 的类型为 10：

- Type

  ISO 定义认证报文的类型值为 10，1 字节。
- Length

  认证 TLV 值的长度，1 字节。
- Value

  认证的具体内容，其中包括了认证的类型和认证的密码，1 ～ 254 字节。

  在 Value 中，认证的类型为 1 字节，具体定义如下：

  - 0：保留的类型
  - 1：明文认证
  - 54：MD5 认证

－ 255：路由域私有认证方式

认证密码的保存情况如下：

- 对于 IIH 报文，使用的认证密码保存在接口下，即前面提到的接口认证。
- 对于 Level-1 LSP 和 SNP 报文，使用的认证密码保存在 IS-IS 进程下，即前面提到的区域认证。
- 对于 Level-2 LSP 和 SNP 报文，使用的认证密码保存在 IS-IS 进程下，即前面提到的路由域认证。

对于接口认证，有以下两种设置：

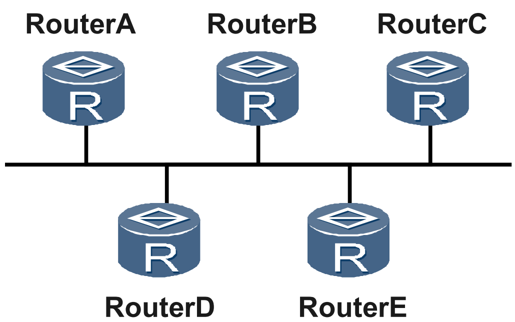
- 发送带认证 TLV 的认证报文，本地对收到的报文也进行认证检查。
- 发送带认证 TLV 的认证报文，但是本地对收到的报文不进行认证检查。

对于区域和路由域认证，可以设置为 SNP 和 LSP 分开认证。

- 本地发送的 LSP 报文和 SNP 报文都携带认证 TLV，对收到的 LSP 报文和 SNP 报文都进行认证检查。
- 本地发送的 LSP 报文携带认证 TLV，对收到的 LSP 报文进行认证检查；发送的 SNP 报文携带认证 TLV，但不对收到的 SNP 报文进行检查。
- 本地发送的 LSP 报文携带认证 TLV，对收到的 LSP 报文进行认证检查；发送的 SNP 报文不携带认证 TLV，也不对收到的 SNP 报文进行认证检查。
- 本地发送的 LSP 报文和 SNP 报文都携带认证 TLV，对收到的 LSP 报文和 SNP 报文都不进行认证检查。

组网应用

图 5-26 广播网中的 IS-IS 认证



配置要求：

- 在同一网络中的多台路由器，当配置的接口认证完全相同，才能建立 IS-IS 邻居。
- 如果多台路由器在同一个区域中，那么为了保证它们的 Level-1 LSDB 能够完全同步，必须将区域认证配置成完全相同。
- 如果多台路由器建立的是 Level-2 邻居，那么为了保证它们的 Level-2 LSDB 能够完全同步，必须将路由域认证配置成完全相同。

5.5 术语与缩略语

术语

术语	解释
TLV	TLV（type-length-value）。TLV 编码方式是一种高效率，扩展性好的协议报文编码方式。也称为 CLV 编码（code-length-value） T-Type：采用不同的值定义不同类型 L-Length：Value 域的整个长度 V-Value：本 TLV 的实际内容，最重要的部分 TLV 编码的优点：可扩展性好，如果想增加对于新特性的支持，只需增加新的 TLV 类型，它采取这样一种形式更加灵活的方式来描述报文中需要加载的信息。
LSP	链路状态协议数据单元（Link State Protocol Data Unit），用来在区域中传播链路状态记录，包含了一个路由器的所有信息：IS 邻居、所连接的 IP 前缀、连接的 ES、区域地址等。LSP 分为两种：Level-1 LSP 和 Level-2 LSP，一个路由器对应不同的 Level 产生且只产生一个 LSP（包含分片）。
CSNP	全序列号协议数据单元（Complete Sequence Numbers Protocol Data Unit），包括数据库的摘要信息，用于邻居间同步数据库。分 Level 发送，分 Level 解析。
DIS	指定中间系统（Designated Intermediate System）
Pseudonodes	伪结点（Pseudonodes）是个虚拟的结点，而非真实的路由器，将广播网络模拟成伪结点。由 DIS 产生，和广播网络中的所有路由器相互宣称建立邻接关系。
PE	提供商边缘（Provider Edge Router）
CE	用户边缘（Customer edge router）
NSR	无间断路由（Non-Stop Routing）

缩略语

缩略语	英文全称	中文全称
IS-IS	Intermediate System-Intermediate System	中间系统到中间系统
IGP	Interior Gateway Protocol	内部网关协议
LSP	Link State Protocol Data Unit	链路状态协议数据单元
CSNP	Complete Sequence Numbers Protocol Data Unit	全序列号协议数据单元

缩略语	英文全称	中文全称
SNP	Sequence Number PDU	序列号报文
DIS	Designated Intermediate System	指定中间系统
TLV	Type-Length-Value	代码类型-长度-值的三元组
SPF	Shortest Path First	最短路由优先算法
MI	Multiple Instance	多实例
MT	Multi-topology	多拓扑
Local-MT	Local Multicast-Topology	本地组播拓扑
URT	Unicast Routing Table	单播路由表
MIGP	IGP Routing Table for Multicast	组播内部网关路由协议路由表（仅供组播使用）
Shortcut（AA）	IGP-Shortcut (Auto Announcement)	自动路由通告类型的 TE-Tunnel
Advertise（FA）	IGP-Advertise (Forwarding Adjacency)	转发邻接体类型的 TE-Tunnel
GR	Graceful Restart	优雅重启
BGP	Border Gateway Protocol	边界网关协议
RM	Routing Management	路由管理
VPN	Virtual Private Networks	虚拟专用网
BFD	Bidirectional Forwarding Detection	双向转发检测
MPLS	Multiprotocol Label Switching	多协议标签交换
CSPF	Constraint-based Shortest Path First	基于约束的最短路由优先
TE	Traffic Engineering	流量工程
LDP	Lable Distribution Protocol	标签分发协议
LSP	Lable Switched Path	标签交换路径
SNMP	Simple Network Management Protocol	简单网络管理协议
MIB	Management Information Base	管理信息库
PE	Provider Edge	运营商边界路由器
CE	Customers Edge	用户边界路由器
RIB	Routing Information Base	路由信息库

6 OSPF

关于本章

6.1 介绍

定义

OSPF（Open Shortest Path First）是 IETF 组织开发的一个基于链路状态的内部网关协议（Interior Gateway Protocol）。

目前针对 IPv4 协议使用的是 OSPF Version 2（RFC2328）；针对 IPv6 协议使用 OSPF Version 3（RFC2740）。本文中所指的 OSPF 如不特殊说明均为 OSPF Version 2。

目的

在 OSPF 出现前，网络上广泛使用 RIP（Routing Information Protocol）作为内部网关协议。

由于 RIP 是基于距离矢量算法的路由协议，存在着收敛慢、路由环路、可扩展性差等问题，所以逐渐被 OSPF 取代。

OSPF 作为基于链路状态的协议，能够解决 RIP 所面临的诸多问题。此外，OSPF 还有以下优点：

- OSPF 采用多播形式收发报文，这样就可以减少其它不运行 OSPF 设备的负担。
- OSPF 支持无类型域间选路（CIDR）。
- OSPF 支持对等价路由进行负载分担。
- OSPF 支持报文加密。

由于 OSPF 具有以上优势，使得 OSPF 作为优秀的内部网关协议被快速接受并广泛使用。

6.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC1587	This document describes a new optional type of OSPF area, somewhat humorously referred to as a "not-so-stubby" area (or NSSA). NSSAs are similar to the existing OSPF stub area configuration option but have the additional capability of importing AS external routes in a limited fashion.	-
RFC1765	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 may be unable to keep the entire database due to resource shortages; we term this "database overflow".	该 RFC 为 Experimental，非 Standard。

文档	描述	备注
RFC2328	This memo documents version 2 of the OSPF protocol. OSPF is a link-state routing protocol.	-
RFC2370	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 the future extensibility of OSPF.	-
RFC3137	This memo describes a backward-compatible technique that may 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.	该 RFC 为 Informational，非 Standard。
RFC3623	This memo documents an enhancement to the OSPF routing protocol, whereby an OSPF router can stay on the forwarding path even as its OSPF software is restarted.	-
RFC3630	This document describes extensions to the OSPF protocol version 2 to support intra-area Traffic Engineering (TE), using Opaque Link State Advertisements.	-
RFC3682	The use of a packet's Time to Live (TTL) (IPv4) or Hop Limit (IPv6) to protect a protocol stack from CPU-utilization based attacks has been proposed in many settings.	该 RFC 为 Experimental，非 Standard。
RFC3906	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.	-
RFC4576	This document specifies the necessary 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 that see it.	-
RFC4577	This document extends that specification by allowing the routing protocol on the PE/CE interface to be the Open Shortest Path First (OSPF) protocol.	-

文档	描述	备注
RFC4750	This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in TCP/IP-based internets. In particular, it defines objects for managing version 2 of the Open Shortest Path First Routing Protocol. Version 2 of the OSPF protocol is specific to the IPv4 address family.	-

6.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R001C00

6.4 原理描述

6.4.1 OSPF 基础

OSPF 协议具有以下特点：

- OSPF 把自治系统划分成逻辑意义上的一个或多个区域。
- OSPF 通过 LSA（Link State Advertisement）的形式发布路由。
- OSPF 依靠在 OSPF 区域内各路由器间交互 OSPF 报文来达到路由信息的统一。
- OSPF 报文封装在 IP 报文内，可以采用单播或组播的形式发送。

OSPF 报文类型

表 6-1 OSPF 报文类型

报文类型	报文作用
Hello 报文	周期性发送，用来发现和维持 OSPF 邻居关系。
DD 报文（Database Description packet）	描述本地 LSDB 的摘要信息，用于两台路由器进行数据库同步。
LSR 报文（Link State Request packet）	用于向对方请求所需的 LSA。 路由器只有在 OSPF 邻居双方成功交换 DD 报文后才会向对方发出 LSR 报文。
LSU 报文（Link State Update packet）	用于向对方发送其所需要的 LSA。
LSAck 报文（Link State Acknowledgment packet）	用来对收到的 LSA 进行确认。

LSA 类型

表 6-2 OSPF LSA 类型

LSA 类型	LSA 作用
Router-LSA（Type1）	每个路由器都会产生，描述了路由器的链路状态和开销，在发布路由器所属的区域内传播。
Network-LSA（Type2）	由 DR 产生，描述本网段的链路状态，在 DR 所属的区域内传播。
Network-summary-LSA（Type3）	由 ABR 产生，描述区域内某个网段的路由，并通告给发布或接收此 LSA 的非 Totally STUB 或 NSSA 区域。
ASBR-summary-LSA（Type4）	由 ABR 产生，描述到 ASBR 的路由，通告给除 ASBR 所在区域的其他区域。
AS-external-LSA（Type5）	由 ASBR 产生，描述到 AS 外部的路由，通告到所有的区域（除了 STUB 区域和 NSSA 区域）。
NSSA LSA（Type7）	由 ASBR 产生，描述到 AS 外部的路由，仅在 NSSA 区域内传播。
Opaque LSA（Type9/Type10/Type11）	Opaque LSA 提供用于 OSPF 的扩展的通用机制。其中： ● Type9 LSA 仅在接口所在网段范围内传播。用于支持 GR 的 Grace LSA 就是 Type9 LSA 的一种。 ● Type10 LSA 在区域内传播。用于支持 TE 的 LSA 就是 Type10 LSA 的一种。 ● Type11 LSA 在自治域内传播，目前还没有实际应用的例子。

路由器类型

OSPF 协议中常用到的路由器类型如图 6-1 所示。

图 6-1 路由器类型

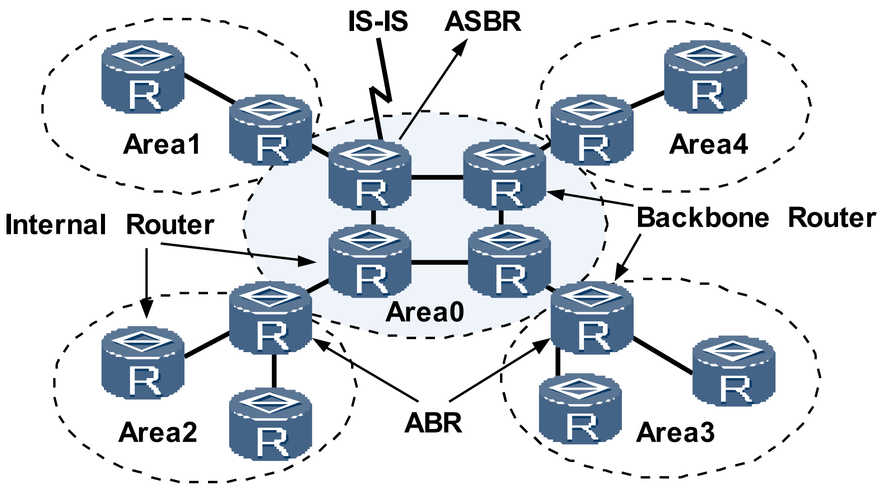


表 6-3 OSPF 路由器类型

路由器类型	含义
区域内路由器（Internal Router）	该类路由器的所有接口都属于同一个 OSPF 区域。
区域边界路由器 ABR（Area Border Router）	该类路由器可以同时属于两个以上的区域，但其中一个必须是骨干区域。 ABR 用来连接骨干区域和非骨干区域，它与骨干区域之间既可以是物理连接，也可以是逻辑上的连接。
骨干路由器（Backbone Router）	该类路由器至少有一个接口属于骨干区域。 所有的 ABR 和位于 Area0 的内部路由器都是骨干路由器。
自治系统边界路由器 ASBR（AS Boundary Router）	与其他 AS 交换路由信息的路由器称为 ASBR。 ASBR 并不一定位于 AS 的边界，它可能是区域内路由器，也可能是 ABR。只要一台 OSPF 路由器引入了外部路由的信息，它就成为 ASBR。

OSPF 路由类型

AS 区域内和区域间路由描述的是 AS 内部的网络结构，AS 外部路由则描述了应该如何选择到 AS 以外目的地址的路由。OSPF 将引入的 AS 外部路由分为 Type1 和 Type2 两类。

表 6-4 中按优先级从高到低顺序列出了路由类型。

表 6-4 OSPF 路由类型

路由类型	含义
Intra Area	区域内路由。
Inter Area	区域间路由。
第一类外部路由（Type1 External）	这类路由的可信程度高一些，所以计算出的外部路由的开销与自治系统内部的路由开销是相当的，并且和 OSPF 自身路由的开销具有可比性。 到第一类外部路由的开销=本路由器到相应的 ASBR 的开销+ASBR 到该路由目的地址的开销。
第二类外部路由（Type2 External）	这类路由的可信度比较低，所以 OSPF 协议认为从 ASBR 到自治系统之外的开销远远大于在自治系统之内到达 ASBR 的开销。 所以，OSPF 计算路由开销时只考虑 ASBR 到自治系统之外的开销，即到第二类外部路由的开销=ASBR 到该路由目的地址的开销。

区域类型

表 6-5 OSPF 区域类型

区域类型	作用
Totally Stub Area	允许 ABR 发布的 Type3 缺省路由，不允许自治系统外部路由和区域间的路由。
Stub Area	和 Totally Stub 区域的不同在于该区域允许区域间路由。
NSSA Area	和 Stub 区域的不同在于该区域允许自治系统外部路由的引入，由 ASBR 发布 Type 7 LSA 通告给本区域。
Totally NSSA Area	和 NSSA 区域的不同在于该区域不允许区域间路由。

OSPF 支持的网络类型

OSPF 根据链路层协议类型，将网络分为如表 6-6 所列四种类型。

表 6-6 OSPF 网络类型

网络类型	含义
广播类型（Broadcast）	当链路层协议是 Ethernet、FDDI 时，缺省情况下，OSPF 认为网络类型是 Broadcast。 在该类型的网络中： ● 通常以组播形式发送 Hello 报文、LSU 报文和 LSAck 报文。其中，224.0.0.5 的组播地址为 OSPF 路由器的预留 IP 组播地址；224.0.0.6 的组播地址为 OSPF DR 的预留 IP 组播地址。 ● 以单播形式发送 DD 报文和 LSR 报文。
NBMA 类型（Non-broadcast multiple access）	当链路层协议是 ATM 时，缺省情况下，OSPF 认为网络类型是 NBMA。 在该类型的网络中，以单播形式发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。
点到多点 P2M 类型（Point-to-Multipoint）	没有一种链路层协议会被缺省的认为是 Point-to-Multipoint 类型。点到多点必须是由其他的网络类型强制更改的。常用做法是将非全连通的 NBMA 改为点到多点的网络。 在该类型的网络中： ● 以组播形式（224.0.0.5）发送 Hello 报文； ● 以单播形式发送其他协议报文（DD 报文、LSR 报文、LSU 报文、LSAck 报文）。
点到点 P2P 类型（point-to-point）	当链路层协议是 PPP、HDLC 和 LAPB 时，缺省情况下，OSPF 认为网络类型是 P2P。 在该类型的网络中，以组播形式（224.0.0.5）发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。

Stub 区域

Stub 区域是一些特定的区域，Stub 区域的 ABR 不传播它们接收到的自治系统外部路由，在这些区域中路由器的路由表规模以及路由信息传递的数量都会大大减少。

Stub 区域是一种可选的配置属性，但并不是每个区域都符合配置的条件。通常来说，Stub 区域位于自治系统的边界，是那些只有一个 ABR 的非骨干区域。

为保证到自治系统外的路由依旧可达，该区域的 ABR 将生成一条缺省路由，并发布给 Stub 区域中的其他非 ABR 路由器。

配置 Stub 区域时需要注意下列几点：

- 骨干区域不能配置成 Stub 区域。
- 如果要将一个区域配置成 Stub 区域，则该区域中的所有路由器都要配置 STUB 区域属性。
- Stub 区域内不能存在 ASBR，即自治系统外部的路由不能在本区域内传播。
- 虚连接不能穿过 Stub 区域。

OSPF 报文认证

OSPF 支持报文验证功能，只有通过验证的 OSPF 报文才能接收，否则将不能正常建立邻居。

AR1200 支持两种验证方式：

- 区域验证方式
- 接口验证方式

AR1200 支持的验证模式按加密算法不同分为 null、simple、MD5 以及 HMAC-MD5。

当两种验证方式都存在时，优先使用接口验证方式。

OSPF 路由聚合

路由聚合是指将具有相同前缀的路由信息聚合在一起，只发布一条路由到其它区域。

通过路由聚合，可以减少路由信息，从而减小路由表的规模，提高路由器的性能。

OSPF 有两种路由聚合方式：

- ABR 聚合

  ABR 向其它区域发送路由信息时，以网段为单位生成 Type3 LSA。如果该区域中存在一些连续的网段，则可以通过命令将这些连续的网段聚合成一个网段。这样 ABR 只发送一条聚合后的 LSA，所有属于命令指定的聚合网段范围的 LSA 将不会再被单独发送出去。

- ASBR 聚合

  配置引入路由聚合后，如果本地路由器是自治系统边界路由器 ASBR，将对引入的聚合地址范围内的 Type5 LSA 进行聚合。当配置了 NSSA 区域时，还要对引入的聚合地址范围内的 Type7 LSA 进行聚合。

  如果本地路由器既是 ASBR 又是 ABR，则对由 Type7 LSA 转化成的 Type5 LSA 进行聚合处理。

OSPF 缺省路由

缺省路由是指目的地址和掩码都是 0 的路由。当路由器无精确匹配的路由时，就可以通过缺省路由进行报文转发。

OSPF 缺省路由通常应用于下面两种情况：

- 由区域边界路由器（ABR）发布 Type3 缺省 Summary LSA，用来指导区域内路由器进行区域之间报文的转发。
- 由自治系统边界路由器（ASBR）发布 Type5 外部缺省 ASE LSA，或者 Type7 外部缺省 NSSA LSA，用来指导自治系统（AS）内路由器进行自治系统外报文的转发。

当路由器无精确匹配的路由时，就可以通过缺省路由进行报文转发。由于 OSPF 路由的分级管理，Type3 缺省路由的优先级高于 Type5/7 路由。

OSPF 缺省路由的发布原则如下：

- OSPF 路由器只有具有对外的出口时，才能够发布缺省路由 LSA。
- 如果 OSPF 路由器已经发布了缺省路由 LSA，那么不再学习其它路由器发布的相同类型缺省路由。即路由计算时不再计算其它路由器发布的相同类型的缺省路由 LSA，但数据库中存有对应 LSA。

- 外部缺省路由的发布如果要依赖于其它路由，那么被依赖的路由不能是本 OSPF 路由域内的路由，即不是本进程 OSPF 学习到的路由。因为外部缺省路由的作用是用于指导报文的域外转发，而本 OSPF 路由域的路由的下一跳都指向了域内，不能满足指导报文域外转发的要求。

不同区域缺省路由发布原则如表 6-7 所示。

表 6-7 不同区域的缺省路由发布原则

区域类型	缺省路由发布原则
普通区域	缺省情况下，普通 OSPF 区域内的 OSPF 路由器是不会产生缺省路由的，即使它有缺省路由。 当网络中缺省路由通过其他路由进程产生时，路由器必须将缺省路由通告到整个 OSPF 自治域中。实现方法是在 ASBR 上手动通过命令进行配置，产生缺省路由。配置完成后，路由器会产生一个缺省 ASE LSA（Type5 LSA），并且通告到整个 OSPF 自治域中。 如果 ASBR 上没有缺省路由，则路由器不会通告缺省路由。
Stub Area	Stub 区域不允许自治系统外部的路由（Type5 LSA）在区域内传播。 区域内的路由器必须通过 ABR 学到自治系统外部的路由。实现方法是 ABR 会自动产生一条缺省的 Summary LSA（Type3 LSA）通告到整个 Stub 区域内。这样，到达自治系统的外部路由就可以通过 ABR 到达。
Totally Stub Area	Totally Stub 区域既不允许自治系统外部的路由（Type5 LSA）在区域内传播，也不允许区域间路由（Type3 LSA）在区域内传播。 区域内的路由器必须通过 ABR 学到自治系统外部和其他区域的路由。实现方法是配置 Totally Stub 区域后，ABR 会自动产生一条缺省的 Summary LSA（Type3 LSA）通告到整个 Stub 区域内。这样，到达自治系统外部的路由和其他区域间的路由都可以通过 ABR 到达。

区域类型	缺省路由发布原则
NSSA Area	NSSA 区域允许引入少量通过本区域的 ASBR 到达的外部路由，但不允许其他区域的外部路由 ASE LSA（Type5 LSA）在区域内传播。即到达自治系统外部的路由只能通过本区域的 ASBR 到达。 只配置了 NSSA 区域是不会自动产生缺省路由的。 此时，有两种选择： ● 如果希望到达自治系统外部的路由通过该区域的 ASBR 到达，而其它外部路由通过其它区域出去。则必须在 ABR 上手动通过命令进行配置，使 ABR 产生一条缺省的 NSSA LSA（Type7 LSA），通告到整个 NSSA 区域内。这样，除了某少部分路由通过 NSSA 的 ASBR 到达，其它路由都可以通过 NSSA 的 ABR 到达其它区域的 ASBR 出去。 ● 如果希望所有的外部路由只通过本区域 NSSA 的 ASBR 到达。则必须在 ASBR 上手动通过命令进行配置，使 ASBR 产生一条缺省的 NSSA LSA（Type7 LSA），通告到整个 NSSA 区域内。这样，所有的外部路由就只能通过本区域 NSSA 的 ASBR 到达。 上面两种情况使用相同的命令在不同的视图下进行配置，区别是在 ABR 上无论路由表中是否存在路由 0.0.0.0，都会产生 Type7 LSA 缺省路由，而在 ASBR 上只有当路由表中存在路由 0.0.0.0 时，才会产生 Type7 LSA 缺省路由。 因为缺省路由只是在本 NSSA 区域内泛洪，并没有泛洪到整个 OSPF 域中，所以本 NSSA 区域内的路由器在找不到路由之后可以从该 NSSA 的 ASBR 出去，但不能实现其他 OSPF 域的路由从这个出口出去。Type7 LSA 缺省路由不会在 ABR 上转换成 Type5 LSA 缺省路由泛洪到整个 OSPF 域。
Totally NSSA Area	Totally NSSA 区域既不允许其他区域的外部路由 ASE LSA（Type5 LSA）在区域内传播，也不允许区域间路由（Type3 LSA）在区域内传播。 区域内的路由器必须通过 ABR 学到其他区域的路由。实现方法是配置 Totally NSSA 区域后，ABR 会自动产生一条缺省的 Type3 LSA 通告到整个 NSSA 区域内。这样，其他区域的外部路由和区域间路由都可以通过 ABR 在区域内传播。

OSPF 路由过滤

OSPF 支持使用路由策略对路由信息进行过滤。缺省情况下，OSPF 不进行路由过滤。

OSPF 可以使用的路由策略包括 route-policy，访问控制列表（access-list），地址前缀列表（prefix-list）。具体策略的描述可以参看 RM 特性描述部分。

OSPF 路由过滤可以应用于以下几个方面：

- 路由引入

  OSPF 可以引入其它路由协议学习到的路由。在引入时可以通过配置路由策略来过滤路由，只引入满足条件的路由。

- 引入路由发布

  OSPF 引入了路由后会向其它邻居发布引入的路由信息。

  可以通过配置过滤规则来过滤向邻居发布的路由信息。该过滤规则只在 ASBR 上配置才有效（只有 ASBR 才能引入路由）。

- 路由学习

  通过配置过滤规则，可以设置 OSPF 对接收到的区域内、区域间和自制系统外部的路由进行过滤。

  该过滤只作用于路由表项的添加与否，即只有通过过滤的路由才被添加到本地路由表中，但所有的路由仍可以在 OSPF 路由表中被发布出去。

- 区域间 LSA 学习

  通过命令可以在 ABR 上配置对进入本区域的 Summary LSA 进行过滤。该配置只在 ABR 上有效（只有 ABR 才能发布 Summary LSA）。

表 6-8 区域间 LSA 学习与路由学习的差异

区域间 LSA 学习	路由学习
直接对进入区域的 LSA 进行过滤。	路由学习中的过滤不对 LSA 进行过滤，只针对 LSA 计算出来的路由是否添加本地路由表进行过滤。学习到的 LSA 是完整的。

- 区域间 LSA 发布

  通过命令可以在 ABR 上配置对本区域出方向的 Summary LSA 进行过滤。该配置只在 ABR 上配置有效。

OSPF 虚连接

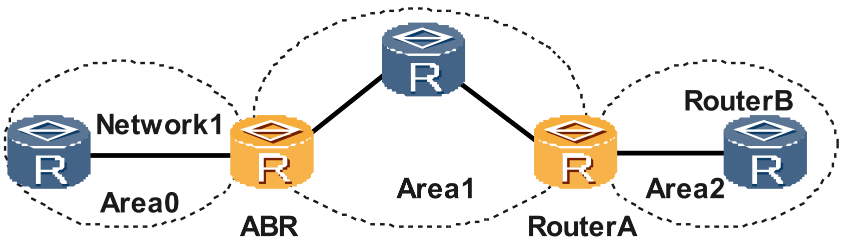
虚连接（Virtual link）是指在两台 ABR 之间通过一个非骨干区域建立的一条逻辑上的连接通道。

- 虚连接必须在两端同时配置方可生效。
- 为虚连接两端提供一条非骨干区域内部路由的区域称为传输区域（Transit Area）。

按照 RFC2328 的建议，在部署 OSPF 时，要求所有的非骨干区域与骨干区域相连。否则会出现有的区域不可达的问题。

如图 6-2 中所示，Area2 没有连接到骨干区 Area0，RouterA 不是 ABR，因此不会向 Area2 生成 Area0 中 Network1 的路由信息，所以 RouterB 上没有到达 Network1 的路由。

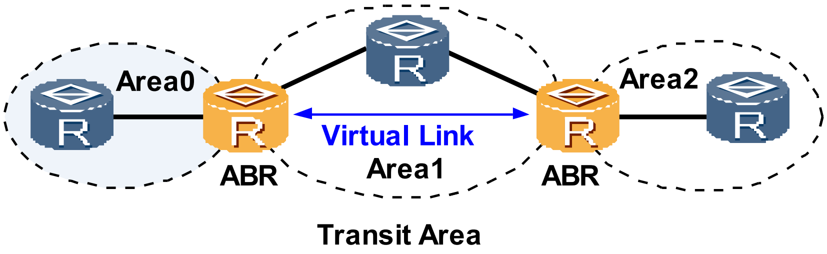
图 6-2 OSPF 非骨干区没有连接骨干区



在实际应用中，可能会因为各方面条件的限制，无法满足所有非骨干区域与骨干区域保持连通的要求。这时可以通过配置 OSPF 虚连接予以解决。

虚连接相当于在两个 ABR 之间形成了一个点到点的连接，因此，虚连接的两端和物理接口一样可以配置接口的各参数，如发送 Hello 报文间隔等。

图 6-3 OSPF 虚连接



如图 6-3 所示，通过虚连接，两台 ABR 之间直接传递 OSPF 报文信息，他们之间的 OSPF 路由器只是起到一个转发报文的作用。由于 OSPF 协议报文的目的地址不是这些路由器，所以这些报文对于他们而言是透明的，只是当作普通的 IP 报文来转发。

OSPF 多进程

OSPF 支持多进程，在同一台路由器上可以运行多个不同的 OSPF 进程，它们之间互不影响，彼此独立。不同 OSPF 进程之间的路由交互相当于不同路由协议之间的路由交互。

路由器的一个接口只能属于某一个 OSPF 进程。

OSPF 多进程的一个典型应用就是在 VPN 场景中 PE 和 CE 之间运行 OSPF 协议，同时 VPN 骨干网上的 IGP 也采用 OSPF。在 PE 上，这两个 OSPF 进程互不影响。

6.4.2 OSPF GR

随着路由设备普遍采用了控制和转发分离的技术，在网络拓扑保持稳定的情况下，控制层面的重启并不会影响转发层面，转发层面仍然可以很好地完成数据转发任务，从而保证业务不受影响。

GR 技术保证了在重启过程中转发层面能够继续指导数据的转发，同时控制层面邻居关系的重建以及路由计算等动作不会影响转发层面的功能，从而避免了路由震荡引发的业务中断，提高了整网的可靠性。

基本概念

GR 是 Graceful Restart 的简称，又被称为平滑重启，是一种用于保证当路由协议重启时数据正常转发并且不影响关键业务的技术。

如果没有特殊说明，以下所说 GR 均表示 RFC3623 所规定的 GR 技术。

GR 技术是属于高可靠性（HA, High Availability）技术的一种。HA 是一整套综合技术，主要包括冗余容错、链路保证、节点故障修复及流量工程。GR 是一种冗余容错技术，目前已经被广泛的使用在主备切换和系统升级方面，以保证关键业务的不间断转发。

和 GR 相关的概念如下：

- Grace-LSA

  OSPF 通过新增 Grace-LSA 来支持 GR 功能。这种 LSA 用于在开始 GR 和退出 GR 时向邻居通告 GR 的时间、原因以及接口地址等内容。

- 路由器在 GR 中的角色
  - Restarter：重启路由器。可以通过配置支持完全 GR 或者部分 GR。
  - Helper：协助重启路由器。可以通过配置支持有计划 GR、无计划 GR 或者通过策略有选择支持 GR。

  说明

  AR1200 只能作为 Helper 路由器，不能作为 Restarter 路由器。

- GR 的原因
  - Unknown：未知原因导致的 GR 操作。
  - Software restart：通过命令行主动触发的 GR 操作。
  - Software reload/upgrade：软件重启或升级导致的 GR 操作。
  - Switch to redundant control processor：异常主备倒换导致的 GR 操作。
- GR 的持续时间

  GR 持续时间最长不超过 1800 秒。GR 成功或失败都可以提前退出，不必等到超时才退出。

GR 的分类

- 完全 GR（Totally GR）：指当有一个邻居不支持 GR 功能时，整个路由器退出 GR 状态。
- 部分 GR（Partly GR）：指当有一个邻居不支持 GR 时，仅该邻居所关联的接口退出 GR，其它接口正常进行 GR 过程。
- 有计划 GR（Planned-GR）：指手动通过命令使路由器执行重启或主备倒换。在进行重启或主备倒换前 Restarter 会先发送 Grace-LSA。
- 非计划 GR（Unplanned-GR）：与 Planned-GR 的区别在于，路由器是由于故障等原因进行重启或主备倒换，并且在主备倒换前不会事先发送 Grace-LSA，而是直接开始主备倒换，在备板正常 Up 后才进入 GR 过程。以下的步骤同 Planned-GR。

GR 的过程

- GR 开始

  对于 Planned-GR，主备倒换命令执行后，Restarter 会首先向每个邻居发送一个 Grace-LSA，通知邻居 GR 的开始以及 GR 的周期、原因等，然后进行主备倒换。

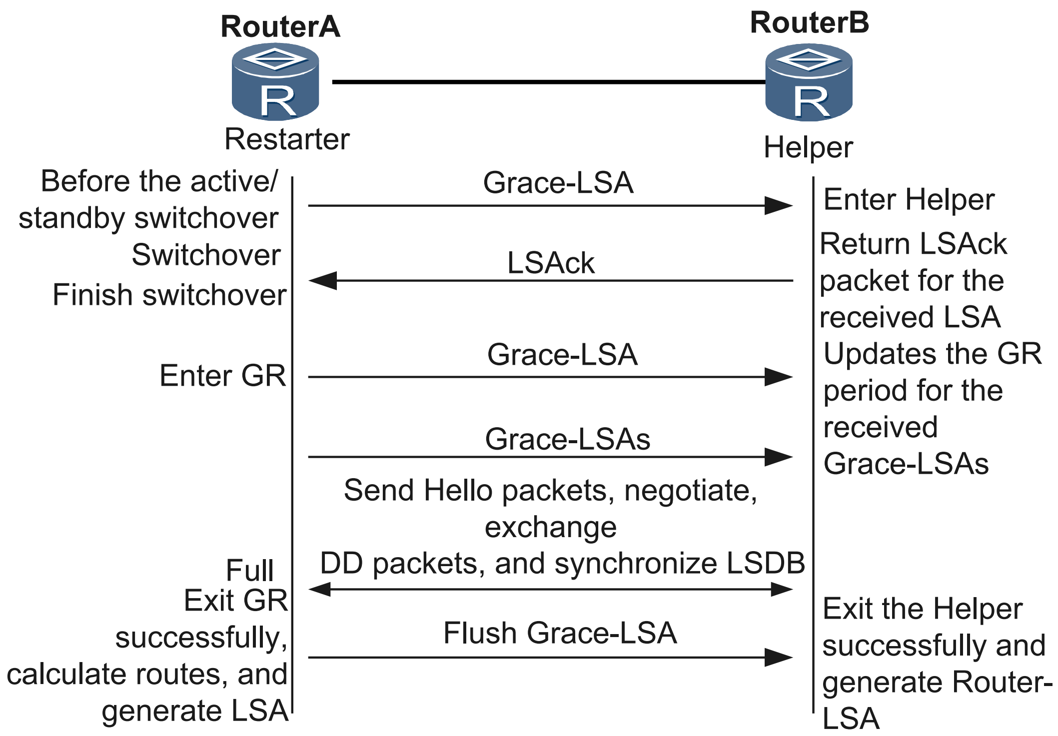
  对于 Unplanned-GR，则不发送这个 Grace-LSA。

  当备板正常 Up 后，立即发送一个 Grace-LSA，通知邻居自己进入 GR，包括 GR 的周期、原因等。然后会再向每个邻居连续发送 5 个 Grace-LSA。（连续发送 5 个是为了确保邻居收到该 Grace-LSA。此为各厂商实现方案，非协议规定）。

  此时发送的 Grace-LSA 是为了告知邻居自己进入 GR 状态，邻居会在 GR 期间保持与 Restarter 的邻居关系，让其它路由器感知不到 Restarter 的倒换。

- GR 过程

图 6-4 OSPF GR 过程



- GR 退出

表 6-9 GR 退出原因

GR 执行情况	Restarter	Helper
GR 成功	Restarter 在 GR 超时前与主备倒换前的所有邻居都重新建立好邻居关系。	收到 Restarter 发送的 Age 为 3600 秒的 Grace-LSA 时与 Restarter 的邻居关系为 Full 状态。

GR 执行情况	Restarter	Helper
GR 失败	● GR 超时并且邻居关系尚未完全恢复。 ● Helper 发送的 Router-LSA 或 Network-LSA 导致 Restarter 端进行双向检查时失败。 ● Restarter 接口状态变化。 ● Restarter 收到 Helper 发送的 1-way Hello 报文。 ● Restarter 收到同一网段上另一台路由器产生的 Grace-LSA。同一网段同一时间只能有一台路由器做 GR。 ● Restarter 同一个网段的邻居之间存在 DR/BDR 不一致的情况（拓扑变化）。	● 在邻居关系超时前没有收到 Restarter 发送的 Grace-LSA。 ● Helper 接口状态发生变化。 ● 收到其它路由器发送的与 Helper 本地数据库不一致的 LSA。（可以通过配置不进行严格 LSA 检查排除这种情况。） ● 同一网段上同一时间收到两台路由器发送的 Grace-LSA。 ● 与其它路由器邻居关系变化。

有无 GR 技术的比较

表 6-10 有无 GR 技术的比较

无 GR 技术的主备倒换	有 GR 技术的主备倒换
● OSPF 邻居重建 ● 路由重新计算 ● 转发表变化 ● 整网感知路由变化，路由短时震荡 ● 转发流量丢失，业务中断	● OSPF 邻居重建 ● 路由重新计算 ● 转发表保持不变 ● 除主备倒换设备的邻居外的其他路由器感知不到路由变化 ● 转发流量零丢失，业务不受影响

6.4.3 OSPF VPN

定义

OSPF VPN 多实例特性是为了支持在 VPN 场景中 PE（Provider Edge routers）和 CE（Customer Edge devices）之间能够运行 OSPF 协议、使用 OSPF 进行路由的学习和发布而在 OSPF 基础协议上进行的扩展。

目的

OSPF 是一种应用广泛的 IGP 协议，很多情况下，VPN 用户内部网络运行 OSPF。如果能够在 PE-CE 之间使用 OSPF，PE 通过 OSPF 向 CE 发布 VPN 路由，则 CE 上就不需要为到 PE 的连接支持其它路由协议，从而简化 CE 的管理和配置。

PE-CE 间运行 OSPF

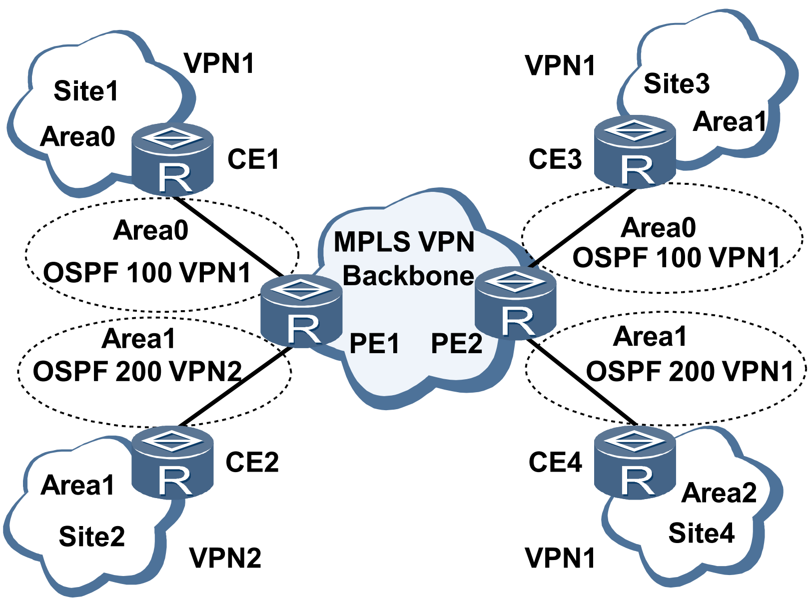
BGP/MPLS VPN 中，PE 之间使用 MP-BGP 传递路由信息，而 PE-CE 间则广泛使用 OSPF 进行路由学习和传递。

PE-CE 间使用 OSPF 有如下优势：

- 通常在一个 Site 内部使用 OSPF 学习路由。如果 PE-CE 间也使用 OSPF 则可以减少 CE 设备所支持的协议种类，降低对 CE 设备的要求。
- 同样，Site 内部和 PE-CE 间都使用 OSPF 可以降低网络管理人员的工作复杂度，不必要求管理人员对多种协议熟练掌握。
- 对于在骨干网上使用 OSPF 而不使用 VPN 的网络，将其转换为使用 BGP/MPLS VPN 时，由于 PE-CE 间继续使用 OSPF，从而降低了转换的难度。

如图 6-5 所示，CE1、CE3 和 CE4 都属于 VPN1，图中 OSPF 之后的数字表示 PE 设备上运行的 OSPF 多实例进程号。

图 6-5 PE-CE 间运行 OSPF



CE1 上的路由发布给 CE3 和 CE4 过程可以描述为：

1. PE1 将 CE1 上的 OSPF 路由引入到 BGP 中，形成 BGP VPNv4 路由。
2. PE1 通过 MP-BGP 将这些 BGP VPNv4 路由发布给 PE2。
3. PE2 将 BGP VPNv4 路由引入到 OSPF，再发布给 CE3 和 CE4。

同理，CE4 和 CE3 上的路由发布给 CE1 的过程类似。

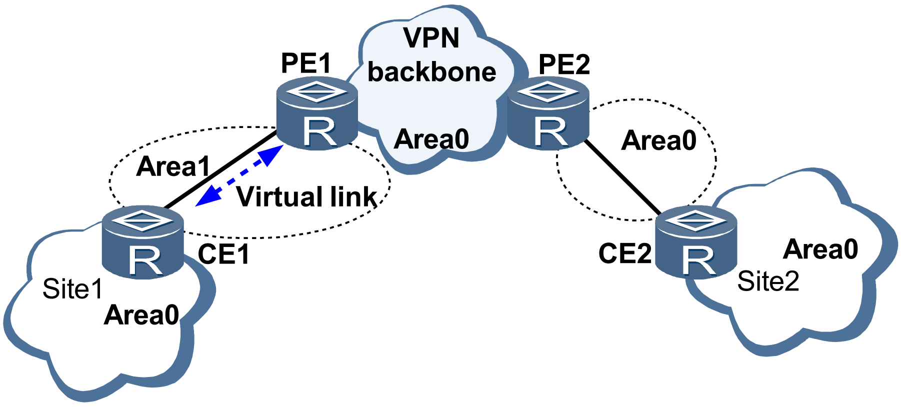
PE-CE 间 OSPF 区域配置

PE 与 CE 之间的 OSPF 区域可以是非骨干区域，也可以是骨干区域（区域 0），并且 PE 永远是 ABR（Area Border Router）。

在 OSPF VPN 扩展应用中，MPLS VPN 骨干网被看作是 Area0。由于 OSPF 要求 Area 0 连续，因此，所有 VPN Site 的 Area0 必须与 MPLS VPN 骨干网相连。如果 VPN Site 中

存在 OSPF Area0，则 CE 接入的 PE 必须通过 Area0 与这个 VPN Site 的骨干区域相连（可以通过 Virtual-link 实现逻辑连通），如图 6-6 所示。

图 6-6 PE-CE 间 OSPF 区域配置



PE-CE 间配置为非骨干区域 1，而 Site1 内配置了骨干区域 0，此时 Site1 的骨干区域就与 VPN 骨干区域分离了，所以在 CE1 与 PE1 间配置虚连接（Virtual link）来保持骨干区域连续。

OSPF Domain ID

本地 OSPF 区域和 VPN 远端的 OSPF 区域间如果相互发布区域间路由（Inter-area routes），则认为这些区域属于同一个 OSPF 域（OSPF Domain）。

- 域标识符（Domain ID）用来标识和区分不同的域。
- 每一个 OSPF 域都有一个或多个域标识符，其中有一个是主标识符，其它为从标识符。
- 如果 OSPF 实例没有明确域标识符，则认为它的标识符为 NULL。

PE 把 BGP 传来的远端路由向 CE 发布时，需要根据域标识符的情况选择向 CE 发布 3 类或 5 类的 OSPF 路由。

- 如果本地的域标识符与 BGP 路由信息中携带的远端域标识符相等或相互兼容，则发布 3 类路由；
- 否则，发布 5 类路由。

表 6-11 Domain ID

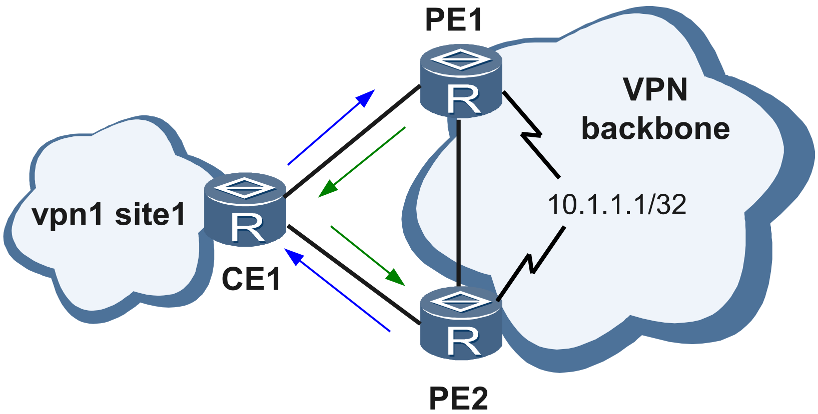
本地和远端域标识符比较情况	是否相等	路由类型
两者都为 NULL	相等	Inter-area 路由。
远端域标识符=本地主域标识符或者远端域标识符=本地从域标识符中的一个	相等	Inter-area 路由。

本地和远端域标识符比较情况	是否相等	路由类型
远端标识符≠本地主从标识符并且远端标识符≠本地从域标识符中的任何一个	不相等	如果本地是非 NSSA（Not So Stubby Area）区域，生成 External 路由 如果是 NSSA 区域，生成 NSSA 路由。

路由环路预防

PE 和 CE 之间，如果 OSPF 与 BGP 的路由相互学习，则有可能导致路由环路问题。

图 6-7 OSPF VPN 路由环路



如图 6-7 所示，PE1 上 OSPF 引入了目的地址为 10.1.1.1./32 的 BGP 路由，产生 5 类或 7 类 LSA 发布给 CE1，CE1 上学到一条目的地址为 10.1.1.1/32，下一跳为 PE1 的 OSPF 路由，并发布给 PE2，这样 PE2 上就学到一条目的地址为 10.1.1.1/32，下一跳为 CE1 的 OSPF 路由。

同理，CE1 上也会学到一条目的地址为 10.1.1.1/32，下一跳为 PE2 的 OSPF 路由，PE1 上学到一条目的地址为 10.1.1.1/32，下一跳为 CE1 的 OSPF 路由。

此时，CE1 上存在两条等价路由，分别指向 PE1 和 PE2，而 PE1 和 PE2 上到 10.1.1.1/32 的下一跳也都指向 CE1，环路就产生了。

同时，由于 OSPF 路由的优先级高于 BGP 路由，PE1 和 PE2 上到 10.1.1.1/32 的 BGP 路由被 OSPF 路由所替代，也就是说，PE1 和 PE2 的路由表中活跃的是到 10.1.1.1/32，下一跳为 CE1 的 OSPF 路由。

既然 BGP 路由转为不活跃状态，之前 OSPF 引入这条 BGP 路由时所产生的 LSA 就会被删除，而这样又会导致 OSPF 路由被撤消。路由表中没有了 OSPF 路由，BGP 路由又变为活跃状态，继续重复之前的循环，导致路由振荡。

OSPF VPN 特性专门针对这种情况提供了解决方案，如表 6-12 所列。

表 6-12 路由环路预防

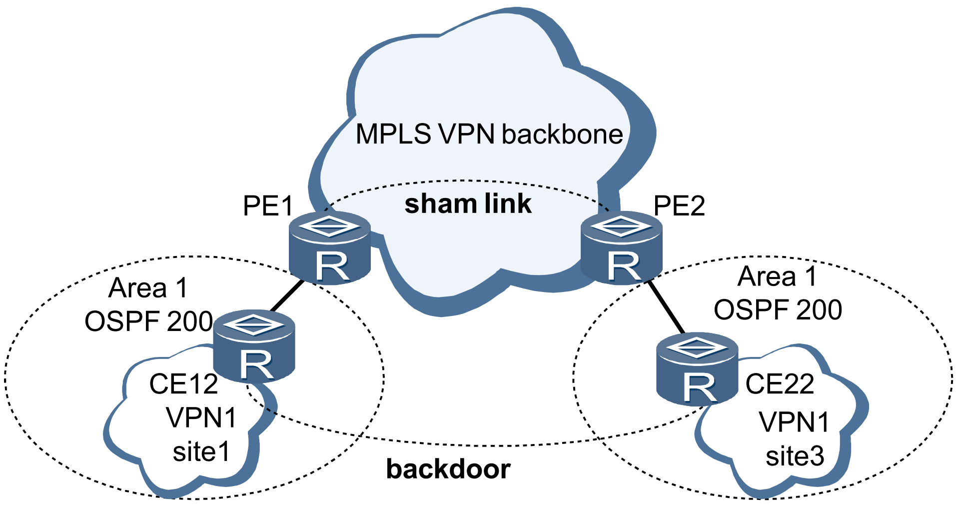
特性名	定义	作用
DN-bit	为了防止路由环路，OSPF 多实例进程使用一个 bit 位作为标志位，称为 DN 位。	PE 在生成 Type3、Type5 或 Type7 LSA 发布给 CE 时，都将 DN 位置位（值为 1），其他类型 LSA 的 DN 位不置位（值为 0）。 PE 路由器的 OSPF 多实例进程在进行计算时，忽略 DN 置位的 LSA。这样就防止了 PE 又从 CE 学到发出的 LSA 而引起的环路。
VPN Route Tag	VPN 路由标记（VPN Route Tag），PE 根据收到的 BGP 的私网路由产生的 5/7 类 LSA 中必须包含这个参数。 VPN 路由标记不在 BGP 的扩展团体属性中传递，只是本地概念，只在收到 BGP 路由并且产生 OSPF LSA 的 PE 路由器上有意义。	当 PE 发现 LSA 的 VPN 路由标记（LSA 的 Tag 值）和自己的一样，就会忽略这条 LSA，因此避免了环路。
缺省路由	目的地址和掩码全为 0 的路由。	PE 不计算缺省路由。 缺省路由用于转发源自 CE 和 CE 所在 Site 的流量到 VPN 骨干网。

伪连接 Sham link

OSPF 伪连接（Sham link）是 MPLS VPN 骨干网上两个 PE 路由器之间的点到点链路，这些链路使用借用（Unnumbered）的地址。

通常情况下，BGP 对等体之间通过 BGP 扩展团体属性在 MPLS VPN 骨干网上承载路由信息。另一端 PE 上运行的 OSPF 可利用这些信息来生成 PE 到 CE 的区域间路由。

图 6-8 OSPF Sham link



如图 6-8 所示，如果本地 CE 所在网段和远端 CE 所在网段间存在一条区域内 OSPF 链路，则称之为后门链路（Backdoor link）。

这种情况下经过后门链路的路由是区域内路由，其优先级要高于经过 MPLS VPN 骨干网的区域间路由，这将导致 VPN 流量总是通过后门路由转发，而不走骨干网。

为了避免这一问题，可以在 PE 路由器之间建立 OSPF 伪连接（Sham link），使经过 MPLS VPN 骨干网的路由也成为 OSPF 区域内路由，并且被优选。

- Sham link 被看成是两个 VPN 实例之间的链路，每个 VPN 实例中必须有一个 Sham link 的端点地址，它是 PE 路由器上 VPN 地址空间中的一个有 32 位掩码的 Loopback 接口地址。
- Sham link 在两个 PE 之间建立起来后，这两个 PE 将成为 Sham link 邻居，交互路由信息。
- Sham link 作为区域内的一条点到点链路，用户可以通过调整度量值在 Sham link 和 Backdoor 之间进行选路。

Multi-VPN-Instance CE

OSPF 多实例通常运行在 PE 路由器上，在用户局域网内部运行 OSPF 多实例的路由器称为 Multi-VPN-Instance CE（MCE），即多实例 CE。

与 PE 上的 OSPF 多实例相比：

- Multi-VPN-Instance CE 不需要支持 BGP/OSPF 互操作功能。
- Multi-VPN-Instance CE 通过为不同的业务建立各自的 OSPF 实例，相当于不同的业务使用不同的虚拟 CE 路由器，从而以较低的成本解决局域网的安全问题。
- Multi-VPN-Instance CE 在同一台 CE 路由器上实现不同的 OSPF 多实例。其实现的关键在于禁止路由环的检查，直接进行路由计算。也就是说，MCE 路由器收到了带有 DN-bit 的 LSA 也会用于路由计算。

6.4.4 OSPF NSSA

定义

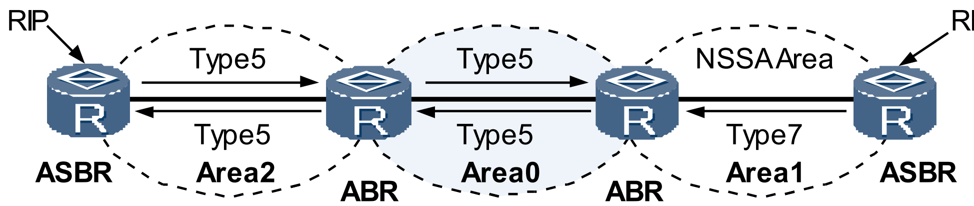
OSPF NSSA 区域（Not-So-Stubby Area）是 OSPF 新增的一类特殊的区域类型。

NSSA 区域其实是 Stub 区域的一个变形，它和 Stub 区域有许多相似的地方。两者的差别在于，NSSA 区域能够将自治域外部路由引入并传播到整个 OSPF 自治域中，同时又不会学习来自 OSPF 网络其它区域的外部路由。

目的

OSPF 规定 Stub 区域是不能引入外部路由的，这样可以避免大量外部路由对 Stub 区域路由器带宽和存储资源的消耗。对于既需要引入外部路由又要避免外部路由带来的资源消耗的场景，Stub 区域就不再满足需求了。因此 Stub 区域的变形——NSSA 区域就产生了。

图 6-9 NSSA 区域



Type-7 LSA

- Type-7 LSA 是为了支持 NSSA 区域而新增的一种 LSA 类型，用于描述引入的外部路由信息。
- Type-7 LSA 由 NSSA 区域的自治域边界路由器（ASBR）产生，其扩散范围仅限于边界路由器所在的 NSSA 区域。
- NSSA 区域的区域边界路由器（ABR）收到 Type-7 LSA 时，会有选择地将其转化为 Type-5 LSA，以便将外部路由信息通告到 OSPF 网络的其它区域。
- 缺省路由也可以通过 Type-7 LSA 来表示，用于指导流量流向其它自治域。

N-bit

一个区域内所有路由器上配置的区域类型必须保持一致。OSPF 在 Hello 报文中使用 N-bit 来标识路由器对 NSSA 区域的支持。区域类型选择不一致的路由器不能建立 OSPF 邻居关系。

虽然协议没有要求，但有些厂商实现时违背了，在 OSPF DD 报文中也置位了 N-bit。为了和这些厂商互通，AR1200 的实现方式是可以通过配置来兼容。

Type-7 LSA 转化为 Type-5 LSA

为了将 NSSA 区域引入的外部路由发布到其它区域，需要把 Type-7 LSA 转化为 Type-5 LSA 以便在整个 OSPF 网络中通告。

- Propagate bit（P-bit）用于告知转化路由器该条 Type-7 LSA 是否需要转化。
- 进行转化的是 NSSA 区域中 Router ID 最大的区域边界路由器（ABR）。
- 只有 Propagate bit（P-bit）置位并且 Forwarding Address 不为 0 的 Type-7 LSA 才能转化为 Type-5 LSA。Forwarding Address 用来表示发送的某个目的地址的报文将被转发到 Forwarding Address 所指定的地址。
- 满足以上条件的缺省 Type-7 LSA 也可以被转化。
- 区域边界路由器产生的 Type-7 LSA 不会置位 P-bit。

缺省路由环路预防

在 NSSA 区域中，可能同时存在多个边界路由器。为了防止路由环路产生，边界路由器之间不计算对方发布的缺省路由。

6.4.5 BFD for OSPF

定义

双向转发检测 BFD（Bidirectional Forwarding Detection）是一种用于检测转发引擎之间通信故障的检测机制。

BFD 对两个系统间的、同一路径上的同一种数据协议的连通性进行检测，这条路径可以是物理链路或逻辑链路，包括隧道。

BFD for OSPF 就是将 BFD 和 OSPF 协议关联起来，将 BFD 对链路故障的快速感应通知 OSPF 协议，从而加快 OSPF 协议对于网络拓扑变化的响应。

目的

网络上的链路故障或拓扑变化都会导致路由器重新进行路由计算，所以缩短路由协议的收敛时间对于提高网络的性能是非常重要的。

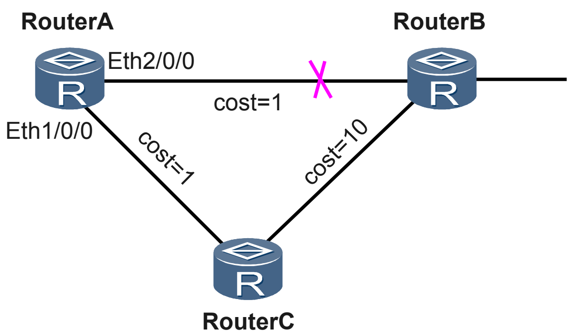
由于链路故障是无法完全避免的，因此，加快故障感知速度并将故障快速通告给路由协议是一种可行的方案。BFD 和路由协议相关联，一旦链路出现故障，BFD 的快速性能够加快路由协议的收敛速度。

表 6-13 BFD for OSPF

有无 BFD	链路故障检测机制	收敛速度
无 BFD	OSPF Dead 定时器超时（默认配置 40s）	秒级
有 BFD	BFD 会话状态为 Down	毫秒级

原理

图 6-10 BFD for OSPF



BFD for OSPF 的原理如图 6-10 所示：

1. 三台设备间建立 OSPF 邻居关系。
2. 邻居状态到达 Full 状态时通知 BFD 建立 BFD 会话。

3. RouterA 到 RouterB 的路由出接口为 GE2/0/0，当这两台设备间的链路出现故障后，BFD 首先感知到并通知 RouterA。
4. RouterA 处理邻居 Down 事件，重新进行路由计算，新的路由出接口为 GE1/0/0，经过 RouterC 到达 RouterB。

6.4.6 OSPF GTSM

定义

GTSM（Generalized TTL Security Mechanism），即通用 TTL 安全保护机制。GTSM 通过检查 IP 报文头中的 TTL 值是否在一个预先定义好的范围内，对 IP 层以上业务进行保护。

目的

网上存在一些“有效报文”攻击，可能通过对一台路由器不断发送报文，使路由器收到这些发送给本机的报文后，不辨别其“合法性”，直接由转发层面上送控制层面处理，导致路由器因为处理这些“合法”报文，系统异常繁忙，CPU 占用率高。

在实际应用中，GTSM 特性主要用于保护建立在 TCP/IP 基础上的控制层面（路由协议等）免受 CPU 利用（CPU-utilization）类型的攻击，如 CPU 过载（CPU overload）。

原理

使能了 GTSM 特性的设备会对收到的所有报文进行策略检查。对于没有通过策略的报文丢弃或者上送控制平面，从而达到防治攻击的目的。策略内容包括：

- 发送给本机 IP 报文的源地址。
- 报文所属的 VPN 实例。
- IP 报文的协议号（OSPF 是 89，BGP 是 6）。
- TCP/UDP 之上协议的协议源端口号、目的端口号。
- 有效 TTL 范围。

GTSM 的实现手段如下：

- 对于直连的协议邻居：将需要发出的协议报文的 TTL 值设定为 255。
- 对于多跳的邻居：可以定义一个合理的 TTL 范围。

GTSM 的应用范围是：

- GTSM 对单播报文有效，对组播报文无效。这是因为组播报文本身具有 TTL 值为 255 的限制，不需要使用 GTSM 进行保护。
- GTSM 不支持基于 Tunnel 的邻居。
- OSPF GTSM 主要应用于 NBMA 网络，Virtual Link 和 Sham link 场景。

6.4.7 OSPF Smart-discover

定义

通常情况下，路由器会周期性地从运行 OSPF 协议的接口上发送 Hello 报文。这个周期被称为 Hello Interval，通过一个 Hello Timer 定时器控制 Hello 报文的发送。这种按固定周期发送报文的方式减缓了 OSPF 邻居关系的建立。

通过使能 Smart-discover 特性，可以在特定场景下加快 OSPF 邻居的建立。

表 6-14 OSPF Smart-discover

接口是否配置 Smart-discover	处理
接口没有配置 Smart-discover	● 必须等待 Hello Timer 超时才能发送 Hello 报文； ● 两次报文发送间隔为 Hello Interval； ● 在这期间邻居一直在等待接收报文。
接口上配置 Smart-discover	● 直接发送 Hello 报文，不需要等待 Hello Timer 超时； ● 邻居可以很快收到报文迅速进行状态迁移。

原理

在以下场景中，使能了 Smart-discover 特性的接口不需要等待 Hello Timer 超时，可以主动向邻居发送 Hello 报文：

- 当邻居状态首次到达 2-way 状态。
- 当邻居状态从 2-way 或更高状态迁移到 Init 状态。

6.4.8 OSPF-BGP 联动

定义

当有新的路由器加入到网络中，或者路由器重启时，可能会出现在 BGP 收敛期间内网络流量丢失的现象。这是由于 IGP 收敛速度比 BGP 快而造成的。

通过使能 OSPF-BGP 联动特性可以解决这个问题。

目的

在存在备份链路的情况下，BGP 在链路回切时，由于路由收敛速度滞后于 OSPF 路由收敛速度，从而造成流量丢失。

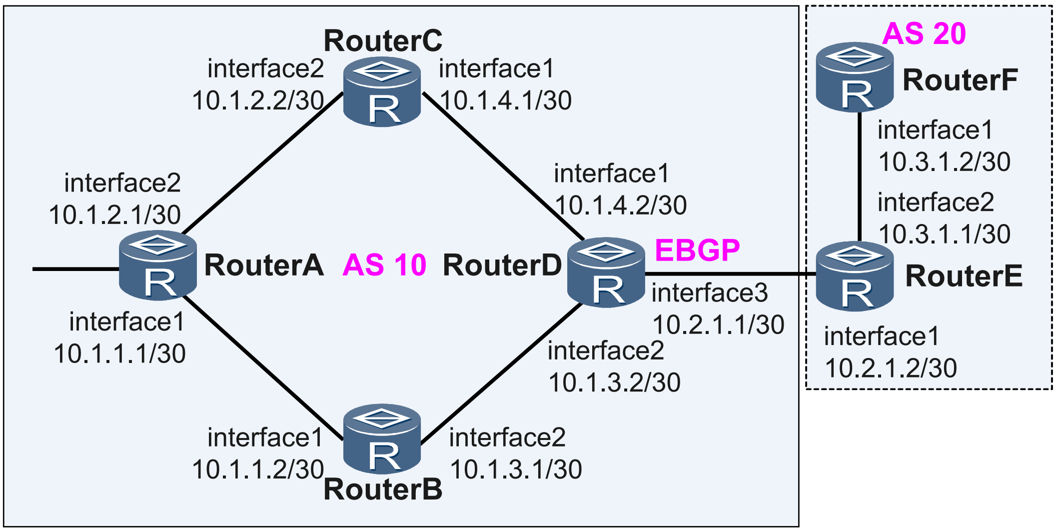
如**图 6-11** 所示，四台设备 RouterA、RouterB、RouterC、RouterD 之间运行 OSPF 协议，并建立 IBGP 连接。RouterC 为 RouterB 的备份设备。当网络环境稳定时，BGP 与 OSPF 在设备上是完全收敛的。

正常情况下，从 RouterA 到 10.3.1.0/30 的流量会途经 RouterB。当 RouterB 发生故障后，流量切换到 RouterC。RouterB 故障恢复以后，流量回切到 RouterB，此时会有流量丢失。

这是因为，在流量回切到 RouterB 的过程中，IGP 收敛速度比 BGP 快，因此 OSPF 先收敛，BGP 还没有完成收敛，导致 RouterB 不知如何到达 10.3.1.0/30。

这样，当从 RouterA 去往 10.3.1.0/30 的流量被发送给 RouterB 时，由于没有必要的路由选择信息，这些流量就会被丢弃。

图 6-11 OSPF-BGP 联动



原理

使能了 OSPF-BGP 联动特性的路由器会在设定的联动时间内保持为 Stub 路由器，也就是说，该路由器发布的 LSA 中的链路度量值为最大值（65535），从而告知其它 OSPF 路由器不要使用这个 Stub 路由器来转发数据，由此保证该路由器不会被用作穿越路由器。

图 **6-11** 中，在 RoutetB 上使能 BGP 联动，这样，在 BGP 收敛完成前，RouterA 不把流量转发到 RouterB 上，而是继续使用备份链路 RouterC，直到 RouterB 上的 BGP 路由完成收敛。

6.4.9 OSPF-LDP 联动

定义

在存在主备链路的网络中，当主链路故障恢复后，流量会从备份链路切换到主链路。

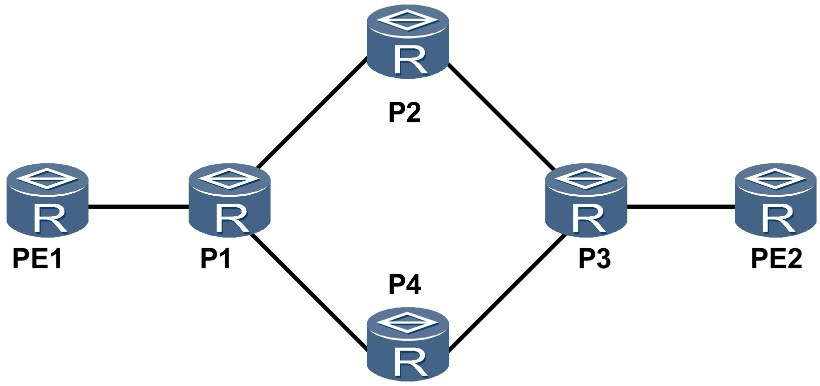
由于 IGP 的收敛在 LDP 会话建立之前完成，导致旧的 LSP 已经删除，新的 LSP 还没有建立，因此 LSP 流量中断。

目的

如图 **6-12** 所示，PE1-P1-P2-P3-PE2 为主链路，PE1-P1-P4-P3-PE2 为备份链路。

主链路发生故障，流量从主链路切换到备份链路。主链路故障恢复，流量从备份链路回切到主链路，此时流量会有较长时间的中断。

图 6-12 OSPF-LDP 联动



通过在 P1 和 P2 上配置 LDP 和 IGP 同步功能，能够缩短流量从备份链路切换到主链路时的中断时间。

表 6-15 OSPF-LDP 联动

是否使能 OSPF-LDP 联动特性	流量中断时间
不使能 OSPF-LDP 联动特性	秒级
使能 OSPF-LDP 联动特性	毫秒级

原理

LDP 和 IGP 同步的基本原理是：通过抑制 IGP 建立邻居关系来推迟路由的回切，直至 LDP 完成收敛。也就是在主链路的 LSP 建立之前，先保留备份链路，让流量继续从备份链路转发，直至主链路的 LSP 建立成功，再删除备份链路。

LDP 和 IGP 同步包括三个定时器：

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

当主链路故障恢复后，路由器进行以下操作：

1. 启动 Hold-down 定时器，IGP 接口先不建立 IGP 邻居，等待 LDP 会话的建立。
2. Hold-down 定时器超时后，启动 Hold-max-cost 定时器。IGP 在本地路由器的链路状态通告中，向主链路通告接口链路的最大 metric 值。
3. 故障链路的 LDP 会话重新建立以后，启动 Delay 定时器等待 LSP 的建立。
4. Delay 定时器超时以后，无论 IGP 的状态如何，LDP 都通知 IGP 同步流程结束。

6.4.10 OSPF Database Overflow

定义

OSPF 协议要求同一个区域中的路由器保存相同的链路状态数据库 LSDB（Link-State Database）。

随着网络上路由数量不断增加，一些路由器由于系统资源有限，不能再承载如此多的路由信息，这种状态就被称为数据库超限（OSPF Database Overflow）。

目的

对于路由信息不断增加导致路由器系统资源耗尽而失效的问题，可以通过配置 Stub 或 NSSA 区域来解决，但 Stub 或 NSSA 区域的方案不能解决动态路由增长导致的数据库超限问题。为了解决数据库超限引发的问题，通过设置 LSDB 中 External LSA 的最大条目数，可以动态限制链路数据库的规模。

原理

通过设置路由器上非缺省外部路由数量的上限，来避免数据库超限。

OSPF 网络中所有路由器都必须配置相同的上限值。这样，只要路由器上外部路由的数量达到该上限，路由器就进入 Overflow 状态，并同时启动超限状态定时器（默认超时时间为 5 秒），路由器在定时器超过 5 秒后自动退出超限状态。

表 6-16 OSPF Database Overflow

Overflow 状态阶段	OSPF 处理流程
进入 Overflow 状态时	路由器删除所有自己产生的非缺省外部路由。
处于 Overflow 状态中	● 不产生非缺省外部路由。 ● 丢弃新收到的非缺省外部路由，不回复确认报文。 ● 当超限状态定时器超时，检查外部路由数量是否仍然超过上限。 　– N=>退出超限状态。 　– Y=>重启定时器。
退出 Overflow 状态时	● 删除超限状态定时器。 ● 产生非缺省外部路由。 ● 接收新收到的非缺省外部路由，回复确认报文。 ● 准备下一次进入超限状态。

6.4.11 OSPF 快速收敛

OSPF 快速收敛是为了提高路由的收敛速度而做的扩展特性。包括：

- I-SPF（Incremental SPF）

增量最短路径优先算法，是指当网络拓扑改变的时候，只对受影响的节点进行路由计算，而不是对全部节点重新进行路由计算，从而加快了路由的计算。

- PRC（Partial Route Calculation）

  部分路由计算，是指当网络上路由发生变化的时候，只对发生变化的路由进行重新计算。

- 智能定时器

  OSPF 智能定时器可以根据用户的配置和触发事件（如路由计算）的频率动态调整时间间隔，使网络快速稳定。

  OSPF 智能定时器使用了指数衰减（Exponential Backoffs）技术，用户的配置可以精确到毫秒。

I-SPF

在 ISO10589 中定义使用 Dijkstra 算法进行路由计算。当网络拓扑中有一个节点发生变化时，这种算法需要重新计算网络中的所有节点，计算时间长，占用过多的 CPU 资源，影响整个网络的收敛速度。

I-SPF 改进了这个算法，除了第一次计算时需要计算全部节点外，每次只计算受到影响的节点，而最后生成的最短路径树 SPT 与原来的算法所计算的结果相同，大大降低了 CPU 的占用率，提高了网络收敛速度。

PRC

PRC 的原理与 I-SPF 相同，都是只对发生变化的路由进行重新计算。不同的是，PRC 不需要计算节点路径，而是根据 I-SPF 算出来的 SPT 来更新路由。

在路由计算中，叶子代表路由，节点则代表路由器。SPT 变化和叶子变化都会引起路由信息的变化，但两者不存在依赖关系，PRC 根据 SPT 或叶子的不同情况进行相应的处理：

- SPT 变化，PRC 处理变化节点上的所有叶子的路由信息。
- SPT 没有变化，PRC 不会处理节点的路由信息。
- 叶子变化，PRC 处理变化的叶子的路由信息。
- 叶子没有变化，PRC 不会处理叶子的路由信息。

比如一个节点使能一个 OSPF 接口，则整个网络拓扑的 SPT 是不变的，这时 PRC 只更新这个节点的接口路由，从而节省 CPU 占用率。

PRC 和 I-SPF 配合使用可以将网络的收敛性能进一步提高，它是原始 SPF 算法的改进，所以已经代替了原有的算法。

说明

在设备的实现中，使用 I-SPF 和 PRC 作为 OSPF 路由计算的唯一算法。

智能定时器

网络不稳定时，可能会频繁进行路由计算，造成系统 CPU 消耗过大。尤其是在不稳定网络中，经常会产生和传播描述不稳定拓扑的 LSA，频繁处理这样的 LSA，不利于整个网络的快速稳定。

OSPF 智能定时器分别对路由计算、LSA 的产生、LSA 的接收进行控制，加速网络收敛。

OSPF 智能定时器可以通过以下两种方式来加速网络收敛：

- 在频繁进行路由计算的网络中，OSPF 智能定时器根据用户的配置和指数衰减技术动态调整两次路由计算的时间间隔，减少路由计算的次数，从而减少 CPU 的消耗，待网络拓扑稳定后再进行路由计算。
- 在不稳定网络中，当路由器由于拓扑的频繁变化需要产生或接收 LSA 时，OSPF 智能定时器可以动态调整时间间隔，在时间间隔之内不产生 LSA 或对接受到的 LSA 不进行处理，从而减少整个网络无效 LSA 的产生和传播。

缺省情况下，OSPF 智能定时器已经打开并且配置了缺省值。

6.4.12 OSPF MIB

定义

MIB（Management Information Base，管理信息库）是一个信息存储库。网络管理员通过 Agent 来调用 MIB 数据对象，从而实现对网络设备控制、配置或监控。（请参看特性描述 SNMP 部分）

RFC4750 定义了 OSPF MIB，主要用来实现设置、修改、查看网络设备中 OSPF 协议的运行状况。

目的

网络管理员可以通过 MIB 查询被管理设备的运行状况，并通过对 MIB 的 SET 操作配置网络设备，从而更加快捷、有效的监控和管理网络。

OSPF 支持对 RFC4750 中全部 MIB 相关节点进行 GET 和 GET-NEXT 操作（不支持 SET 操作）。

为了增强和补充 RFC4750，OSPF 支持私有 MIB，并支持对 OSPF 私有 MIB 的 SET 操作。

原理

在将一个 OSPF 进程绑定到 MIB 后，网管可以对 OSPF MIB 进行 GET 和 GET-NEXT 操作，来获取被绑定进程中 OSPF LSDB、区域、接口、邻居等相关信息。

OSPF 支持 3 个私有 MIB 表的 SET 操作，分别是进程（Process）表，区域（Area）表，网络（Network）表。

- 通过对私有 MIB 进程表的 SET 操作可以创建或删除 OSPF 进程，以及配置或取消 OSPF 进程相关的参数。
- 通过对私有 MIB 区域表的 SET 操作可以在 OSPF 进程下创建或删除 OSPF 区域，以及配置或取消 OSPF 区域相关的参数。
- 通过对私有 MIB 网络表的 SET 操作可以为 OSPF 区域配置或取消具体的网段。

网络管理员可以通过对 3 个 OSPF 私有 MIB 表的 SET 操作进行基础的 OSPF 配置，并搭建基本的 OSPF 拓扑，方便的对网络进行管理和配置。

6.4.13 OSPF Mesh-Group

定义

OSPF Mesh-Group 是将并行链路场景中的链路分组，从而洪泛时从群组中选取代表链路进行洪泛，避免重复洪泛而造成不必要的系统压力。

缺省情况下，不使能 Mesh-Group 功能。

目的

当 OSPF 进程收到一个 LSA 或者新产生一个 LSA 时，会进行洪泛操作。并行链路场景下，OSPF 会对每一条链路洪泛 LSA，发送 Update 报文。

这样，如果有 2000 条并行链路，则每个 LSA 洪泛都要发送 2000 次，然而只有一次洪泛是有效的，其他 1999 次洪泛为重复洪泛。

为了避免这种重复洪泛而造成的系统压力，使能 Mesh Group 特性，可以将并行链路进行归组，选取代表链路进行洪泛。

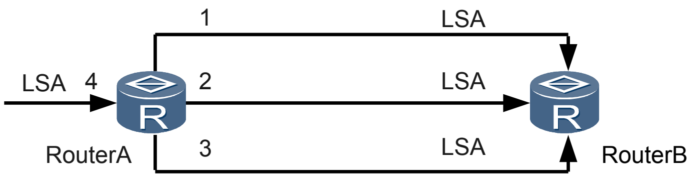
原理

当路由器和邻居存在多条并行链路时，通过 OSPF Mesh-Group 特性，可以明显减轻链路的压力。

如图 **6-13** 所示，RouterA 和 RouterB 建立 OSPF 邻居关系，通过 3 条链路相连。当 RouterA 从接口 4 接收到新的 LSA 后，会将该 LSA 通过 1、2、3 接口洪泛到 RouterB。

这种洪泛方式会造成并行链路的压力，因为对于存在多条并行链路的邻居来说，只需要选取一条主链路进行洪泛 LSA 即可。

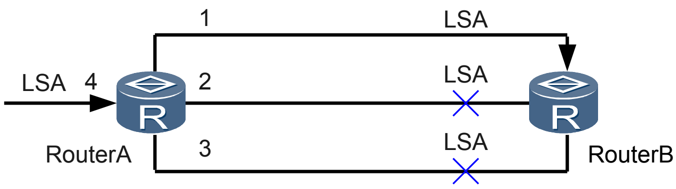
图 6-13 没有使能 OSPF Mesh-Group 特性时 LSA 的洪泛情况



使能了 OSPF Mesh-Group 特性的设备和邻居存在多条并行链路时，当其收到 LSA 后，会选取一条主链路进行泛洪，如图 **6-14** 所示。

当主链路上接口状态低于 Exchange 时，OSPF 会在并行链路中重新选取主链路，并继续洪泛 LSA，这是因为，OSPF 规定，只有当邻居状态达到 Exchange 时，才能洪泛 LSA。并且，当 RouterB 从链路 1 收到来自 RouterA 洪泛的 LSA 后，不会再将该 LSA 从链路 2、3 反向洪泛给 RouterA。

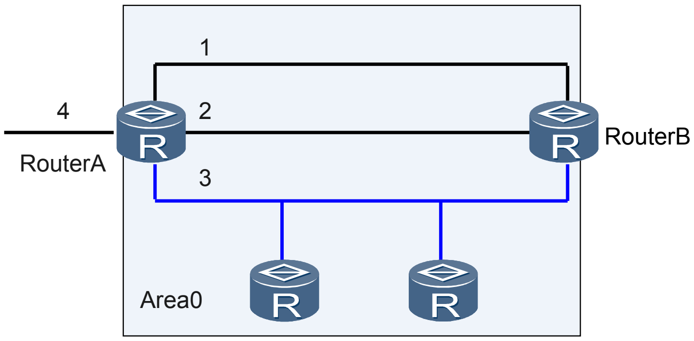
图 6-14 使能 OSPF Mesh-Group 特性时 LSA 的洪泛情况



Mesh-Group 以邻居的 Router ID 唯一标识一个群组，接口状态大于 Exchange 且与同一个邻居相连的接口属于同一个 Mesh-Group。

如图 6-15 所示，RouterA 在区域 0 中有一个群组，分别是接口 1 和接口 2 所在的链路。由于接口 3 所在的链路为广播链路，有超过一个邻居，所以不能加入到群组中。

图 6-15 接口不能加入到群组中的情况



说明

另外，路由器使能 Mesh-Group 后，若其直连的邻居路由器 Router ID 配置重复，会引起全网 LSDB 不同步、路由计算不正确的情况，需要重新配置邻居路由器的 Router ID（注：配置重复 Router ID 属于错误配置）。

6.4.14 按优先级收敛

OSPF 按优先级收敛是指在大量路由情况下，能够让某些特定的路由优先收敛的一种技术。通过对不同的路由配置不同的收敛优先级，达到重要的路由先收敛的目的，提高网络的可靠性。

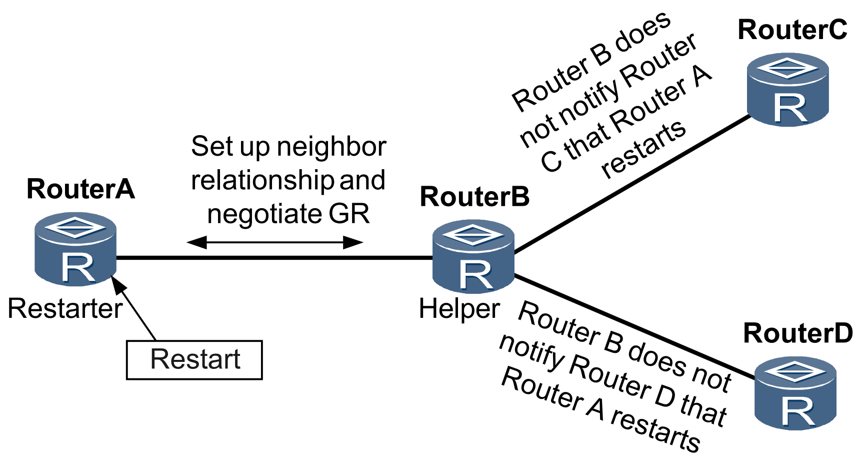
OSPF 按优先级收敛能够让某些特定的路由优先收敛，因此用户可以把和关键业务相关的路由配置成相对较高的优先级，使这些路由更快的收敛，从而使关键的业务受到的影响减小。

6.5 应用

6.5.1 OSPF GR

如图 6-16 所示，RouterA、RouterB、RouterC 和 RouterD 运行 OSPF 协议实现网络互通，RouterA 和 RouterB 使能了 GR 功能。当 RouterA 重启时，RouterB 协助 RouterA 完成平滑重启，但并不通告给其它邻居，网络流量并不中断。

图 6-16 OSPF GR

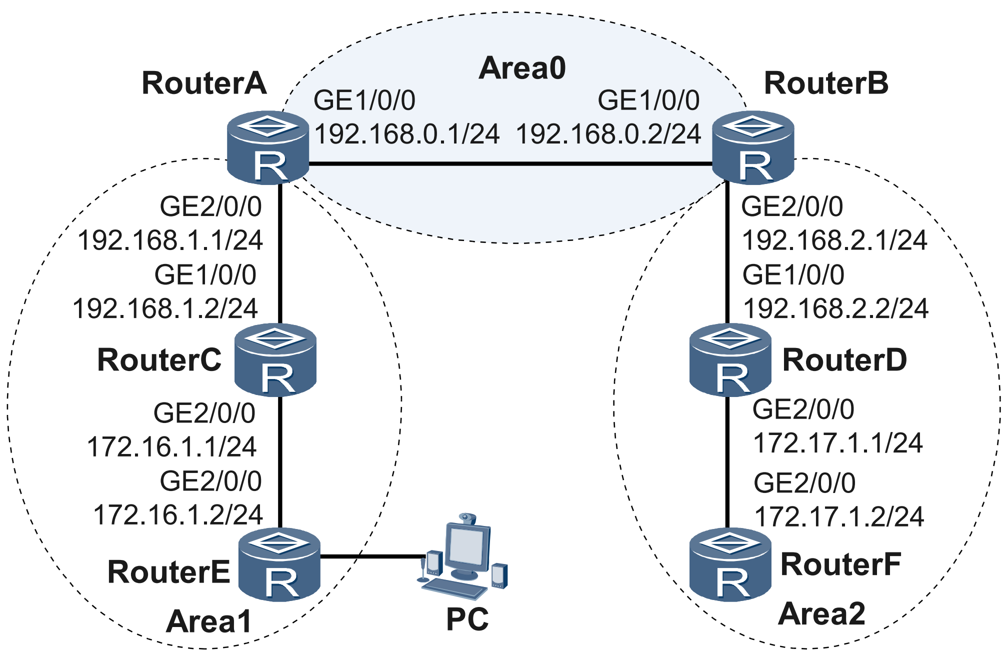


6.5.2 OSPF GTSM

如图 **6-17** 所示，各设备间运行 OSPF 协议，在 RouterC 上启用 GTSM 保护功能。各设备发往 RouterC 的报文有效 TTL 范围如下：

- RouterA 和 RouterE 是 RouterC 的相邻设备，报文有效 TTL 范围是 255 到 255。
- RouterB、RouterD、RouterF 发往 RouterC 的报文有效 TTL 范围分别是 254 到 255、253 到 255、252 到 255。

图 6-17 OSPF GTSM



6.6 术语与缩略语

术语

术语	解释
PE	Provider Edge Router，提供商边缘路由器。该路由器属于连接 CE 路由器的服务提供商网络的一部分。PE 路由器完成所有 VPN 处理。
CE	Customer Edge Router，用户边缘路由器。CE 路由器属于用户网络的一部分，并且和 PE 路由器接口。CE 路由器不能感知相连的 VPN。

缩略语

缩略语	英文全称	中文全称
OSPF	Open Shortest Path First	开放式最短路径优先协议
GR	Graceful Restart	平滑重启
LSA	Link State Advertisement	链路状态发布
TE	Traffic Engineer	流量工程
MPLS	Multiprotocol Label Switching	多协议标记交换
CSPF	Constraint Shortest Path First	约束最短路径优先

7 OSPFv3

关于本章

7.1 介绍

定义

OSPF（Open Shortest Path First）是 IETF 组织开发的一个基于链路状态的内部网关协议（Interior Gateway Protocol）。

目前针对 IPv4 协议使用的是 OSPF Version 2，针对 IPv6 协议使用 OSPF Version 3。

- OSPFv3 是 OSPF Version 3 的简称。
- OSPFv3 是运行于 IPv6 的 OSPF 路由协议（RFC2740）。
- OSPFv3 在 OSPFv2 基础上进行了增强，是一个独立的路由协议。

目的

OSPFv3 的主要目的是开发一种独立于任何具体网络层的路由协议。为实现这一目的，OSPFv3 的内部路由器信息被重新进行了设计。

OSPFv3 与 OSPFv2 的不同在于：

- OSPFv3 不在位于数据包和链路状态公告（LSA）起始位置的报文头部插入基于 IP 的数据。
- OSPFv3 利用独立于网络协议的信息，来执行过去需要 IP 报文头部数据的关键任务，如识别发布路由数据的 LSA。

7.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC2740	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.	

7.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R002C00

7.4 原理描述

7.4.1 OSPFv3 基本原理

OSPFv3 是运行于 IPv6 的 OSPF 路由协议（RFC2740），它在 OSPFv2 基础上进行了增强，是一个独立的路由协议。

- OSPFv3 在 Hello 报文、状态机、LSDB、洪泛机制和路由计算等方面的工作原理和 OSPFv2 保持一致。
- OSPFv3 协议把自治系统划分成逻辑意义上的一个或多个区域，通过 LSA（Link State Advertisement）的形式发布路由。
- OSPFv3 依靠在 OSPFv3 区域内各路由器间交互 OSPFv3 报文来达到路由信息的统一。
- OSPFv3 报文封装在 IPv6 报文内，可以采用单播和组播的形式发送。

OSPFv3 报文类型

报文类型	报文作用
Hello 报文	周期性发送，用来发现和维持 OSPFv3 邻居关系。
DD 报文（Database Description packet）	描述了本地 LSDB 的摘要信息，用于两台路由器进行数据库同步。
LSR 报文（Link State Request packet）	用于向对方请求所需的 LSA。 路由器只有在 OSPFv3 邻居双方成功交换 DD 报文后才会向对方发出 LSR 报文。
LSU 报文（Link State Update packet）	向对方发送其所需要的 LSA。
LSAck 报文（Link State Acknowledgment packet）	用来对收到的 LSA 进行确认。

LSA 类型

LSA 类型	LSA 作用
Router-LSA（Type1）	路由器会为每个运行 OSPFv3 接口所在的区域产生一个 LSA，描述了路由器的链路状态和开销，在所属的区域内传播。
Network-LSA（Type2）	由 DR 产生，描述本链路的链路状态，在所属的区域内传播。
Inter-Area-Prefix-LSA（Type3）	由 ABR 产生，描述区域内某个网段的路由，并通告给其他相关区域。
Inter-Area-Router-LSA（Type4）	由 ABR 产生，描述到 ASBR 的路由，通告给除 ASBR 所在区域的其他相关区域。
AS-external-LSA（Type5）	由 ASBR 产生，描述到 AS 外部的路由，通告到所有的区域（除了 Stub 区域和 NSSA 区域）。
NSSA LSA（Type7）	由 ASBR 产生，描述到 AS 外部的路由，仅在 NSSA 区域内传播。
Link-LSA（Type8）	每个路由器都会为每个链路产生一个 Link-LSA，描述到此 Link 上的 link-local 地址、IPv6 前缀地址，并提供将会在 Network-LSA 中设置的链路选项，它仅在此链路内传播。
Intra-Area-Prefix-LSA（Type9）	每个路由器及 DR 都会产生一个或多个此类 LSA，在所属的区域内传播。 ● 路由器产生的此类 LSA，描述与 Route-LSA 相关联的 IPv6 前缀地址。 ● DR 产生的此类 LSA，描述与 Network-LSA 相关联的 IPv6 前缀地址。

路由器类型

图 7-1 路由器类型

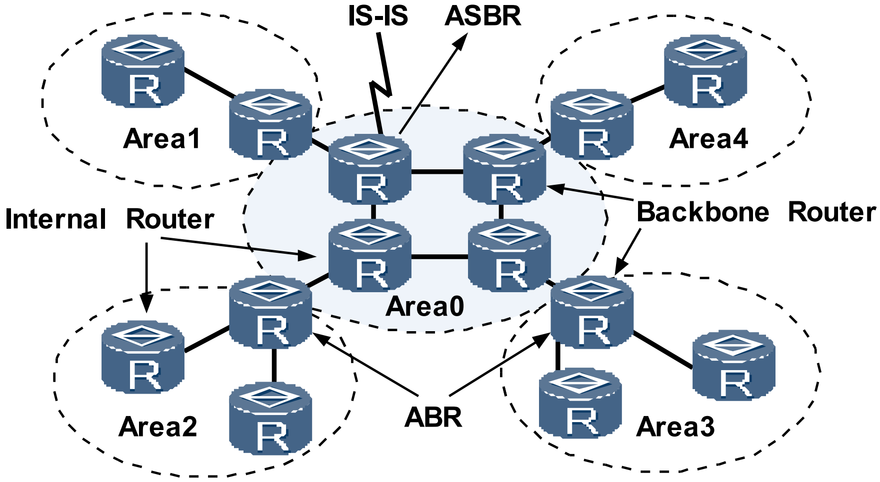


表 7-1 路由器的类型及含义

路由器类型	含义
区域内路由器（Internal Router）	该类路由器的所有接口都属于同一个 OSPFv3 区域。
区域边界路由器 ABR（Area Border Router）	该类路由器可以同时属于两个以上的区域，但其中一个必须是骨干区域。 ABR 用来连接骨干区域和非骨干区域，它与骨干区域之间既可以是物理连接，也可以是逻辑上的连接。
骨干路由器（Backbone Router）	该类路由器至少有一个接口属于骨干区域。 因此，所有的 ABR 和位于 Area0 的内部路由器都是骨干路由器。
自治系统边界路由器 ASBR（AS Boundary Router）	与其他 AS 交换路由信息的路由器称为 ASBR。 ASBR 并不一定位于 AS 的边界，它可能是区域内路由器，也可能是 ABR。只要一台 OSPFv3 路由器引入了外部路由的信息，它就成为 ASBR。

OSPFv3 路由类型

AS 区域内和区域间路由描述的是 AS 内部的网络结构，AS 外部路由则描述了应该如何选择到 AS 以外目的地址的路由。OSPFv3 将引入的 AS 外部路由分为 Type1 和 Type2 两类。

表 7-2 中按优先级从高到低顺序列出了路由类型。

表 7-2 OSPFv3 路由类型

路由类型	含义
Intra Area	区域内路由。
Inter Area	区域间路由。
第一类外部路由（Type1 External）	这类路由的可信程度高一些，所以计算出的外部路由的开销与自治系统内部的路由开销是相当的，并且和 OSPFv3 自身路由的开销具有可比性。 到第一类外部路由的开销=本路由器到相应的 ASBR 的开销+ASBR 到该路由目的地址的开销。
第二类外部路由（Type2 External）	这类路由的可信度比较低，所以 OSPFv3 协议认为从 ASBR 到自治系统之外的开销远远大于在自治系统之内到达 ASBR 的开销。 所以，OSPFv3 计算路由开销时只考虑 ASBR 到自治系统之外的开销，即到第二类外部路由的开销=ASBR 到该路由目的地址的开销。

区域类型

表 7-3 OSPFv3 区域类型

区域类型	作用
Totally Stub Area	允许 ABR 发布的 Type3 缺省路由，不允许自治系统外部路由和区域间的路由。
Stub Area	和 Totally Stub 区域的不同在于，该区域允许区域间路由。
NSSA Area	和 Stub 区域的不同在于该区域允许自治系统外部路由的引入，由 ASBR 发布 Type7 LSA 通告给本区域。

OSPFv3 支持的网络类型

OSPFv3 根据链路层协议类型，将网络分为如表 7-4 所列四种类型。

表 7-4 OSPFv3 网络类型

网络类型	含义
广播类型（Broadcast）	当链路层协议是 Ethernet、FDDI 时，缺省情况下，OSPFv3 认为网络类型是 Broadcast。 在该类型的网络中： ● 通常以组播形式发送 Hello 报文、LSU 报文和 LSAck 报文。其中，FF02::5 为 OSPFv3 路由器的预留 IPv6 组播地址；FF02::6 为 OSPFv3 DR/BDR 的预留 IPv6 组播地址。 ● 以单播形式发送 DD 报文和 LSR 报文。
NBMA 类型（Non-broadcast multiple access）	当链路层协议是帧中继、ATM 或 X.25 时，缺省情况下，OSPFv3 认为网络类型是 NBMA。 在该类型的网络中，以单播形式发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。
点到多点 P2M 类型（Point-to-Multipoint）	没有一种链路层协议会被缺省的认为是 Point-to-Multipoint 类型。点到多点必须是由其他的网络类型强制更改的。常用做法是将非全连通的 NBMA 改为点到多点的网络。 在该类型的网络中： ● 以组播形式（FF02::5）发送 Hello 报文； ● 以单播形式发送其他协议报文（DD 报文、LSR 报文、LSU 报文、LSAck 报文）。
点到点 P2P 类型（point-to-point）	当链路层协议是 PPP、HDLC 和 LAPB 时，缺省情况下，OSPFv3 认为网络类型是 P2P。 在该类型的网络中，以组播形式（FF02::5）发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。

Stub 区域

Stub 区域是一些特定的区域，Stub 区域的 ABR 不传播它们接收到的自治系统外部路由，在这些区域中路由器的路由表规模以及路由信息传递的数量都会大大减少。

Stub 区域是一种可选的配置属性，但并不是每个区域都符合配置的条件。通常来说，Stub 区域位于自治系统的边界，是那些只有一个 ABR 的非骨干区域。

为保证到自治系统外的路由依旧可达，该区域的 ABR 将生成一条缺省路由，并发布给 Stub 区域中的其他非 ABR 路由器。

配置 Stub 区域时需要注意下列几点：

- 骨干区域不能配置成 Stub 区域。
- 如果要将一个区域配置成 Stub 区域，则该区域中的所有路由器必须都要配置成 Stub 路由器。
- Stub 区域内不能存在 ASBR，即自治系统外部的路由不能在本区域内传播。
- 虚连接不能穿过 Stub 区域。

OSPFv3 路由聚合

通过路由聚合，可以减少路由信息，从而减小路由表的规模，提高路由器的性能。

OSPFv3 路由聚合过程如下：

- ABR 聚合

  ABR 路由聚合将具有相同前缀的路由信息聚合在一起，只发布一条路由到其它区域。

  ABR 在发送路由信息到别的区域的过程中，基于 IPv6 地址前缀生成 type3 LSA。如果一些连续的 IPv6 地址前缀在某个区域内存在并且 ABR 聚合功能是使能的，这些 IPv6 地址前缀聚合成一个地址前缀。 对于前缀相同的多条 LSA，区域边界路由器（ABR）只发送一条聚合的 LSA。属于命令指定网段的 LSA 不单独传输。
- ASBR 聚合

  ASBR 路由聚合将具有相同前缀的引入路由聚合在一起，只发布一条路由到其它区域。

  使能 ASBR 路由聚合功能后，自治系统边界路由器将处于聚合地址范围内的引入的 type5 LSA 进行聚合。聚合后，自治系统边界路由器不会为配置范围内所有的前缀生成单独的 type5 LSA，相反，自治系统边界路由器只为聚合的前缀生成单一的 type5 LSA。对于 NSSA 区域，自治系统边界路由器将处于聚合地址范围内的多条引入的 type7 LSA 聚合成一条 type7 LSA。

OSPFv3 虚连接

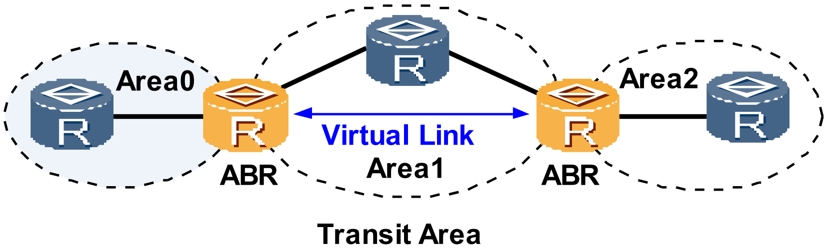
虚连接（Virtual link）是指在两台 ABR 之间通过一个非骨干区域建立的一条逻辑上的连接通道。

- 虚连接必须在两端同时配置方可生效。
- 为虚连接两端提供一条非骨干区域内部路由的区域称为传输区域（Transit Area）。

在实际应用中，可能会因为各方面条件的限制，无法满足所有非骨干区域与骨干区域保持连通的要求。这时可以通过配置 OSPFv3 虚连接予以解决。

虚连接相当于在两个 ABR 之间形成了一个点到点的连接，因此，虚连接的两端和物理接口一样可以配置接口的各参数，如发送 Hello 报文间隔等。

图 7-2 OSPFv3 虚连接



如图 7-2 所示，通过虚连接，两台 ABR 之间直接传递 OSPFv3 报文信息，他们之间的 OSPFv3 路由器只是起到一个转发报文的作用。由于 OSPFv3 协议报文的目的地址不是这些路由器，所以这些报文对于他们而言是透明的，只是当作普通的 IP 报文来转发。

OSPFv3 多进程

OSPFv3 支持多进程，在同一台路由器上可以运行多个不同的 OSPFv3 进程，它们之间互不影响，彼此独立。不同 OSPFv3 进程之间的路由交互相当于不同路由协议之间的路由交互。

路由器的一个接口只能属于某一个 OSPFv3 进程。

7.4.2 OSPFv3 GR

GR 是 Graceful Restart 的简称，又被称为平滑重启，是一种用于保证当路由协议重启时数据正常转发并且不影响关键业务的技术。

GR 技术属于高可靠性（HA, High Availability）技术的一种。HA 是一整套综合技术，主要包括冗余容错、链路保证、节点故障修复及流量工程。GR 是一种冗余容错技术，目前已经被广泛的使用在主备切换和系统升级方面，以保证关键业务的不间断转发。

- 在没有使用 GR 时，由于各种原因触发的主备切换，都会造成短时间的转发中断，并且在全网造成路由振荡。对于一个大型网络，尤其是运营商网络，这些路由振荡和业务中断是不可接受的。

GR 技术保证了在重启过程中转发层面能够继续指导数据的转发，同时控制层面邻居关系的重建以及路由计算等动作不会影响转发层面的功能，从而避免了路由震荡引发的业务中断，提高了整网的可靠性。

基本概念

- Grace-LSA
  - OSPFv3 通过在链路上泛洪一种 Grace-LSA 来支持 GR 功能。
  - Grace-LSA 用于在开始和退出 GR 时向邻居通告 GR 的时间、原因、接口实例 ID 等内容。
- 路由器在 GR 中的角色
  - Restarter：重启路由器；
  - Helper：协助重启路由器。

说明

AR1200 只能作为 Helper 路由器，不能作为 Restarter 路由器。

- GR 的实现方式
  - Planned-GR：指通过执行 **reset ospfv3 graceful-restart** 命令进行的协议平滑重启。这种方式在重启前，会给邻居先发送 Grace-LSA。
  - Unplanned-GR：通过命令引起的主备倒换，或路由器故障（非命令）引起的重启和主备倒换都被认为是 Unplanned GR。

    与 Planned-GR 的区别在于，Unplanned-GR 在主备倒换前不事先发送 Grace-LSA，而是直接开始主备倒换，并在备板正常 Up 后发送 Grace-LSA 并进入 GR 过程。以后的步骤同 Planned-GR。

GR 过程

图 7-3 OSPFv3 Planned-GR 过程（reset ospfv3 graceful-restart）

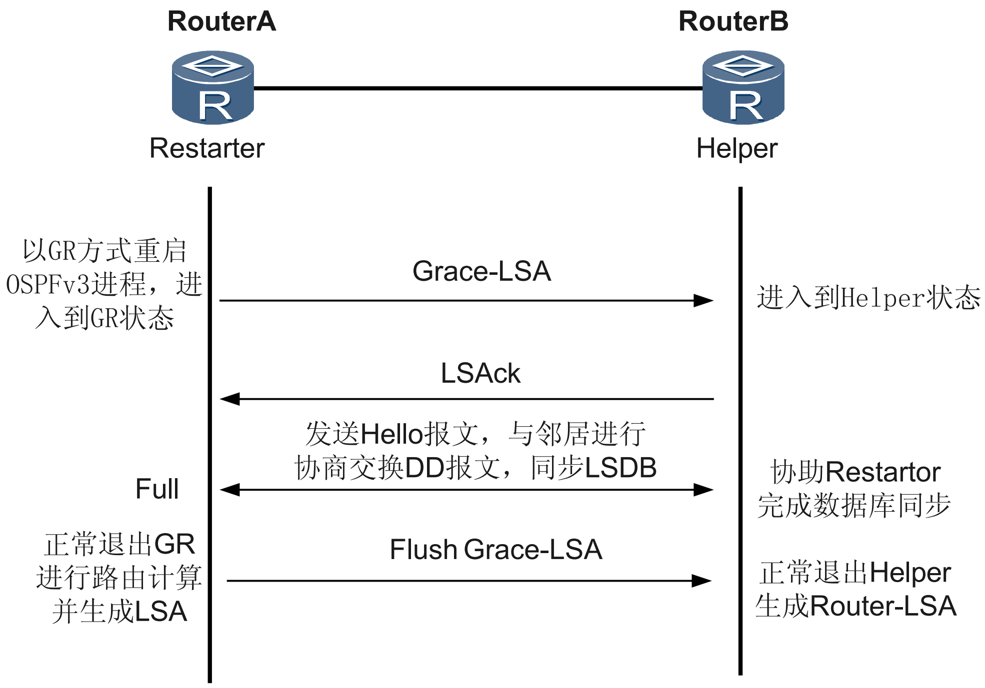
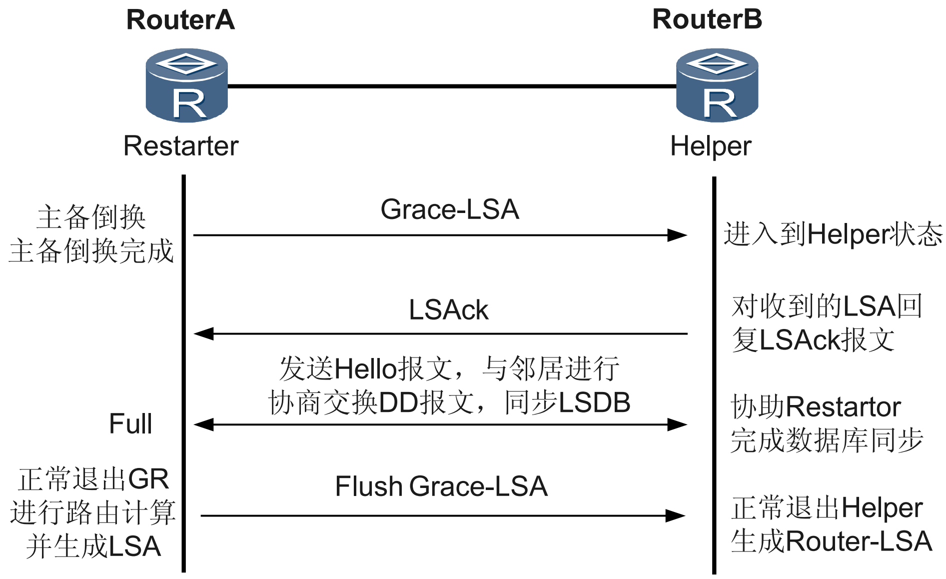


图 7-4 OSPFv3 Unplanned-GR 过程（主备倒换）



- Restarter 端：

1. 对于 Planned-GR（AR1200 为通过命令对协议进行重启），Restarter 会首先向每个邻居发送一个 Grace-LSA 通知邻居 GR 的开始以及 GR 的周期、原因等。

   对于 Unplanned-GR（AR1200 为主备倒换或非命令导致的重启），当备板正常 Up 后，马上发送一个 Grace-LSA，通知邻居自己进入 GR，包括 GR 的周期，原因等。
2. Restarter 与邻居重新开始协商建立邻接关系。
3. Restarter 与所有 GR 前邻居的邻接关系都达到 Full 状态后，
   - 正常退出 GR 并重新计算路由；
   - 更新主控板路由表和接口板 FIB 表，并删除失效的路由表项；
   - 向 Helper 发送 LSA 年龄为 3600 秒的 Grace-LSA 通知 Helper 退出 GR。

   此时 GR 为成功执行。
4. 如果在 GR 过程中出错，或 GR 定时器超时还有邻居没有达到 Full 状态，则 GR 失败退出，进行非 GR 的重启。这种情况下会导致报文丢失。

- Helper 端：

1. 路由器收到 Grace-LSA 后，如果配置了允许支持邻居执行 GR，则进入 Helper 模式。
2. Helper 与 Restarter 继续保持邻接关系，状态不发生改变。
3. Helper 如果继续收到包含不同 GR 周期的 Grace-LSA，则只更新平滑重启的周期。
4. 收到 Restarter 发送的 Age 为 3600 秒的表示 GR 成功的 Grace-LSA 后，正常退出 GR。
5. 如果 GR 过程出错，则退出 Helper 状态，重新进行路由计算，删除失效的路由。

有无 GR 技术的比较

表 7-5 有无 OSPFv3 GR 的比较

无 GR 技术的主备倒换	有 GR 技术的主备倒换
● OSPFv3 邻居重建 ● 路由重新计算 ● 转发表发生改变 ● 整网感知路由变化，路由短时震荡 ● 转发流量丢失，业务中断	● OSPFv3 邻居重建 ● 路由重新计算 ● 转发表保持不变 ● 除主备倒换设备的邻居外的其他路由器感知不到路由变化 ● 转发流量零丢失，业务不受影响

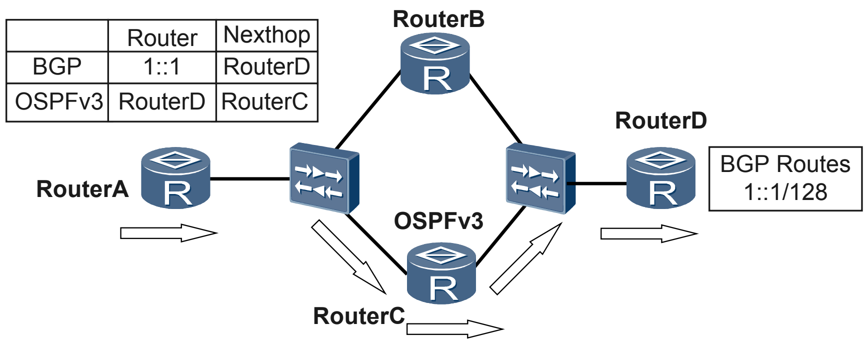
7.4.3 OSPFv3 与 BGP 联动

当有新的路由器加入到网络中，或者路由器重启时，可能会出现在 BGP 收敛期间内网络流量丢失的现象。这是由于 IGP 收敛速度比 BGP 快而造成的。通过使能 OSPFv3-BGP 联动特性可以解决这个问题。

在 BGP 网络中，如果一台路由器从故障中恢复正常，其 BGP 会重新收敛，这段时间内可能会有流量丢失。

如图 7-5 所示，从 RouterA 流到目的地 RouterD 的流量经过 RouterC，穿越了 BGP 网络。

图 7-5 流量穿越 BGP 网络

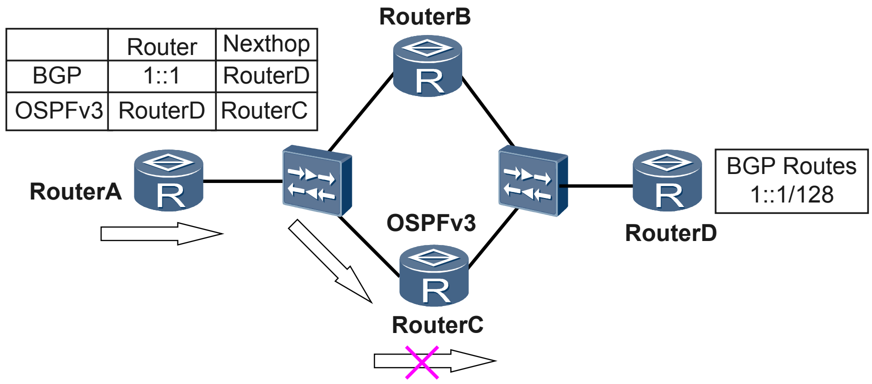


如果 RouterC 上发生故障，流量会由于路由重新选取而经过 RouterB。流量丢失的问题发生在 RouterC 恢复正常的时候。

由于 OSPFv3 收敛速度比 BGP 快，RouterC 恢复正常时，OSPFv3 先收敛。从而在 RouterA 上，到 RouterD 的路由下一跳是 RouterC，但由于 RouterC 上的 BGP 还没有重新收敛完成，这会导致 RouterC 不知道如何到达 RouterD。

这样，当有 RouterA 到 RouterD 的流量被发送给 RouterC 时，由于没有必要的路由选择信息，这些流量就会被丢弃，如图 7-6 所示。

图 7-6 没有使能 OSPFv3-BGP 联动特性的设备重启时导致流量丢失



OSPFv3 与 BGP 联动过程

使能了 OSPFv3-BGP 联动特性的路由器会在重新启动时，向 OSPFv3 域中通告一条信息，告知其它路由器不要将它用作穿越路由器。

同时，该 OSPFv3 路由器在它的路由器链路状态通告（link-state advertisement，LSA）中设置最大的度量值（65535），来确保自己不会被用作穿越路由器，但对于 BGP 会话依然可达。

7.4.4 OSPFv3 和 OSPFv2 协议比较

相同点：

- 网络类型和接口类型
- 接口状态机和邻居状态机
- 链路状态数据库（LSDB）
- 洪泛机制（Flooding mechanism）
- 相同类型的报文：Hello 报文、DD 报文、LSR 报文、LSU 报文和 LSAck 报文
- 路由计算基本相同

不同点：

- OSPFv3 基于链路，而不是网段

  OSPFv3 运行在 IPv6 协议上，IPv6 是基于链路而不是网段的。

  这样，在配置 OSPFv3 时，不需要考虑是否配置在同一网段，只要在同一链路，就可以不配置 IPv6 全局地址而直接建立联系。

- OSPFv3 上移除了 IP 地址的意义

  这样做的目的是为了使“拓扑与地址分离”。OSPFv3 可以不依赖 IPv6 全局地址的配置来计算出 OSPFv3 的拓扑结构。IPv6 全局地址仅用于 Vlink 接口及报文的转发。

- OSPFv3 的报文及 LSA 格式发生改变
  - OSPFv3 报文不包含 IP 地址。
  - OSPFv3 的 Router LSA 和 Network LSA 里不包含 IP 地址。IP 地址部分由新增的两类 LSA（Link LSA 和 Intra Area Prefix LSA）宣告。

- OSPFv3 的 Router ID、Area ID 和 LSA Link State ID 不再表示 IP 地址，但仍保留 IPv4 地址格式。
- 广播、NBMA 及 P2MP 网络中，邻居不再由 IP 地址标识，只由 Router ID 标识。

- OSPFv3 的 LSA 报文里添加 LSA 的洪泛范围

  OSPFv3 在 LSA 报文头的 LSA Type 里，添加 LSA 的洪泛范围，这使得 OSPFv3 的路由器更加灵活，可以处理不能识别类型的 LSA：

  - OSPFv3 可存储或洪泛不识别报文，而 OSPF 只简单丢弃掉不识别报文。
  - OSPFv3 允许洪泛范围为区域或链路本地（Link-local），并且设置 U 位（报文可按洪泛范围为链路本地来处理）的不识别报文存储或通过 Stub 区域。

  例如，RouterA 和 B 都可识别某类 LSA，它们之间通过<RouterC 连接，但 RouterC 不识别该类 LSA。这样，当 RouterA 洪泛此类 LSA 时，RouterC 虽然不识别，但还是可以洪泛给 RouterB，B 收到后继续处理。

  如果运行的是 OSPF 协议，只会丢弃不能识别的报文，RouterB 则不能收到此类 LSA。

- OSPFv3 支持一个链路上多个进程

  一个 OSPFv2 物理接口，只能和一个多实例绑定。但一个 OSPFv3 物理接口，可以和多个多实例绑定，并用不同的 Instance ID 区分。这些运行在同一条物理链路上的多个 OSPFv3 实例，分别与链路对端设备建立邻居及发送报文，且互不干扰。这样可以充分共享同一链路资源。

- OSPFv3 利用 IPv6 链路本地地址

  IPv6 使用链路本地（Link-local）地址在同一链路上发现邻居及自动配置等。运行 IPv6 的路由器不转发目的地址为链路本地地址的 IPv6 报文，此类报文只在同一链路有效。链路本地单播地址从 FE80/10 开始。

  OSPFv3 是运行在 IPv6 上的路由协议，同样使用链路本地地址来维持邻居，同步 LSA 数据库。除 Vlink 外的所有 OSPFv3 接口都使用链路本地地址作为源地址及下一跳来发送 OSPFv3 报文。

  这样的好处是：

  - 不需要配置 IPv6 全局地址，就可以得到 OSPFv3 拓扑，实现拓扑与地址分离。
  - 通过在链路上泛洪的报文不会传到其他链路上，来减少报文不必要的泛洪来节省带宽。

- OSPFv3 移除所有认证字段

  OSPFv3 的认证直接使用 IPv6 的认证及安全处理，不再需要其自身来完成认证，使用协议时只需关注协议本身即可。

- 新增两种 LSA

  - Link LSA：用于路由器宣告各个链路上对应的链路本地地址及其所配置的 IPv6 全局地址，仅在链路内洪泛。
  - Intra Area Prefix LSA：用于向其他路由器宣告本路由器或本网络（广播网及 NBMA）的 IPv6 全局地址信息，在区域内洪泛。

- OSPFv3 只通过 Router ID 来标识邻居

  OSPF 在广播网，NBMA 及 P2MP 网络中是通过 IPv4 接口地址来标识的。

  OSPFv3 只通过 Router ID 来标识邻居，这样即使没有配置 IPv6 全局地址，或是 IPv6 全局地址配置都不在同一网段，OSPFv3 的邻居还是可以建立并维护的，以达到“拓扑与地址分离”的目的。

7.5 术语与缩略语

缩略语

缩略语	英文全称	中文全称
OSPF	Open Shortest Path First	开放式最短路径优先协议
OSPFv3	OSPF Version 3	运行于 IPv6 的 OSPF 路由协议
GR	Graceful Restart	平滑重启
LSA	Link State Advertisement	链路状态发布

8 BGP

关于本章

8.1 介绍

定义

BGP（Border Gateway Protocol）是一种用于自治系统 AS（Autonomous System）之间的动态路由协议。

早期发布的三个版本分别是 BGP-1（RFC1105）、BGP-2（RFC1163）和 BGP-3（RFC1267），主要用于交换 AS 之间的可达路由信息，构建 AS 域间的传播路径，防止路由环路的产生，并在 AS 级别应用一些路由策略。

当前使用的版本是 BGP-4（RFC4271）。

BGP 作为事实上的 Internet 外部路由协议标准，被广泛应用于 ISP（Internet Service Provider）之间。

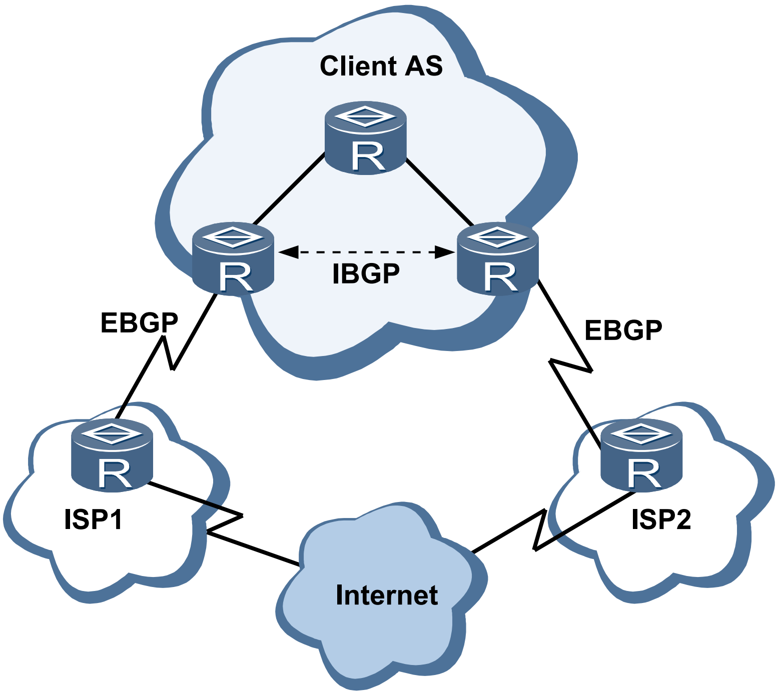
BGP 协议具有如下特点：

- BGP 是一种外部网关协议（EGP），与 OSPF、RIP 等内部网关协议（IGP）不同，其着眼点不在于自动发现网络拓扑，而在于在 AS 之间选择最佳路由和控制路由的传播。
- BGP 使用 TCP 作为其传输层协议（监听端口号为 179），提高了协议的可靠性。
  - BGP 进行域间的路由选择，对协议的稳定性要求非常高。因此用 TCP 协议的高可靠性来保证 BGP 协议的稳定性。
  - BGP 的对等体之间必须在逻辑上连通，并进行 TCP 连接。目的端口号为 179，本地端口号任意。
- BGP 支持无类别域间路由 CIDR（Classless Inter-Domain Routing）。
- 路由更新时，BGP 只发送更新的路由，大大减少了 BGP 传播路由所占用的带宽，适用于在 Internet 上传播大量的路由信息。
- BGP 是一种距离矢量（Distance-Vector）路由协议。
- BGP 从设计上避免了环路的发生。
  - AS 之间：BGP 通过携带 AS 路径信息来标记途经的 AS，带有本地 AS 号的路由将被丢弃，从而避免了域间产生环路。
  - AS 内部：BGP 在 AS 内学到的路由不再通告给 AS 内的 BGP 邻居，避免了 AS 内产生环路。
- BGP 提供了丰富的路由策略，能够对路由实现灵活的过滤和选择。
- BGP 提供了防止路由振荡的机制，有效提高了 Internet 网络的稳定性。
- BGP 易于扩展，能够适应网络新的发展。

目的

BGP 用于在 AS 之间传递路由信息，并不是所有情况都需要运行 BGP。

图 8-1 BGP 的应用场景



以下情况中需要使用 BGP 协议：

- 如图 8-1，用户需要同时与两个或者多个 ISP 相连，ISP 需要向用户提供部分或完全的 Internet 路由。这时可以通过 BGP 路由携带的 AS 信息来决定到达目的地，走哪一个 ISP 的 AS 更为经济。
- 不同组织下的用户之间需要传递 AS 路径信息。
- 用户需要传播组播路由构造组播拓扑，请参见《Huawei AR1200 系列 特性描述-IP 组播》。

以下情况不需要使用 BGP 协议：

- 用户只与一个 ISP 相连。
- ISP 不需要向用户提供 Internet 路由。
- AS 间使用了缺省路由进行连接。

8.2 参考标准和协议

本特性的参考资料清单如下所示：

表 8-1 参考标准和协议

文档	描述	备注
RFC4271	A Border Gateway Protocol 4 (BGP-4)	-

文档	描述	备注
RFC4760	Multiprotocol Extensions for BGP-4	-
RFC3392	Capabilities Advertisement with BGP-4	-
RFC2918	Route Refresh Capability for BGP-4	-
RFC2439	BGP Route Flap Damping	-
RFC1997	BGP Communities Attribute	-
RFC4456	BGP Route Reflection	-
RFC3065	Autonomous System Confederations for BGP	-
RFC3232	Assigned Numbers: RFC 1700 is Replaced by an On-line Database	-
RFC827	Exterior Gateway Protocol (EGP)	-
RFC3682	The Generalized TTL Security Mechanism (GTSM)	-
RFC 4724	Graceful Restart Mechanism for BGP	-
draft-rijsman-bfd-down-subcode-00	BFD Down Subcode for BGP Cease Notification Message	-
RFC 4486	Subcodes for BGP Cease Notification Message	-

8.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R001C00

8.4 原理描述

本章主要介绍 BGP 的特性如下所示：

表 8-2 BGP 主要特性列表

特性中文名称	特性英文名称	参考文档
基本原理	BGP Basic	RFC 4271
路由引入	Route Redistribution	
路由聚合	Route Aggregation	RFC 2519
路由衰减	BGP Route Flap Damping	RFC 2439
团体属性	BGP Communities Attribute	RFC 1997
路由反射器	BGP Route Reflection	RFC 4456
BGP 联盟	Autonomous System Confederations for BGP	RFC 3065
MP-BGP	Multiprotocol Extensions for BGP-4	RFC 4760
BGP4+	Multiprotocol Extensions for BGP-4	RFC 4760
BGP 扩展团体属性	BGP Extended Communities Attribute	RFC 4360
BGP GR	Graceful Restart Mechanism for BGP	RFC 4724
BGP 安全性	Security Requirements for Keys used with the TCP MD5 Signature Option	RFC 3562
BFD for BGP	Bidirectional Forwarding Detection for BGP	draft-ietf-bfd-base-06

8.4.1 协议基本原理

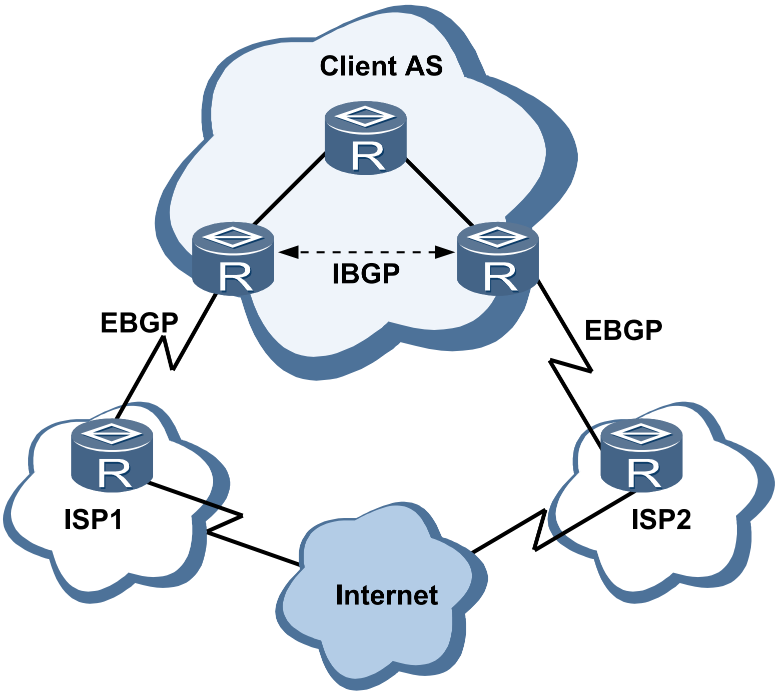
BGP 运行方式

BGP 在路由器上以下列两种方式运行，如图 8-2 所示：

- IBGP（Internal BGP）
- EBGP（External BGP）

当 BGP 运行于同一 AS 内部时，被称为 IBGP；当 BGP 运行于不同 AS 之间时，称为 EBGP。

图 8-2 BGP 的运行方式



BGP 消息中的角色

- Speaker：发送 BGP 消息的路由器称为 BGP 发言者（Speaker），它接收或产生新的路由信息，并发布（Advertise）给其它 BGP Speaker。当 BGP Speaker 收到来自其它 AS 的新路由时，如果该路由比当前已知路由更优、或者当前还没有该路由，它就把这条路由发布给所有其他 BGP Speaker（发送这条路由的 BGP Speaker 除外）。
- Peer：相互交换消息的 BGP Speaker 之间互称对等体（Peer），若干相关的对等体可以构成对等体组（Peer Group）。

BGP 的消息

BGP 的运行是通过消息驱动的，共有 Open、Update、Notification、Keepalive 和 Route-Refresh 等 5 种消息类型。

- Open 消息：是 TCP 连接建立后发送的第一个消息，用于建立 BGP 对等体之间的连接关系。对等体在接收到 Open 消息并协商成功后，将发送 Keepalive 消息确认并保持连接的有效性。确认后，对等体间可以进行 Update、Notification、Keepalive 和 Route-Refresh 消息的交换。
- Update 消息：用于在对等体之间交换路由信息。Update 消息可以发布多条属性相同的可达路由信息，也可以撤销多条不可达路由信息。
  - 一条 Update 消息可以发布多条具有相同路由属性的可达路由，这些路由可共享一组路由属性。所有包含在一个给定的 Update 消息里的路由属性适用于该 Update 消息中的 NLRI（Network Layer Reachability Information）字段里的所有目的地（用 IP 前缀表示）。

- 一条 Update 消息可以撤销多条不可达路由。每一个路由通过目的地（用 IP 前缀表示），清楚的定义了 BGP Speaker 之间先前通告过的路由。
- 一条 Update 消息可以只用于撤销路由，这样就不需要包括路径属性或者 NLRI。相反，也可以只用于通告可达路由，就不需要携带撤销路由信息了。

- Notification 消息：当 BGP 检测到错误状态时，就向对等体发出 Notification 消息，之后 BGP 连接会立即中断。
- Keepalive 消息：BGP 会周期性的向对等体发出 Keepalive 消息，用来保持连接的有效性。
- Route-Refresh 消息：Route-Refresh 消息用来通知对等体自己支持路由刷新能力（Route-Refresh capability）。

  在所有 BGP 路由器使能 Route-Refresh 能力的情况下，如果 BGP 的入口路由策略发生了变化，本地 BGP 路由器会向对等体发布 Route-Refresh 消息，收到此消息的对等体会将其路由信息重新发给本地 BGP 路由器。这样，可以在不中断 BGP 连接的情况下，对 BGP 路由表进行动态刷新，并应用新的路由策略。

BGP 有限状态机

BGP 有限状态机共有六种状态，分别是 Idle、Connect、Active、OpenSent、OpenConfirm 和 Established。

- Idle 状态下，BGP 拒绝任何进入的连接请求，是 BGP 初始状态。
- Connect 状态下，BGP 等待 TCP 连接的建立完成后再决定后续操作。
- Active 状态下，BGP 将尝试进行 TCP 连接的建立，是 BGP 的中间状态。
- OpenSent 状态下，BGP 等待对等体的 Open 消息。
- OpenConfirm 状态下，BGP 等待一个 Notification 报文或 Keepalive 报文。
- Established 状态下，BGP 对等体间可以交换 Update 报文、Route-Refresh 报文、Keepalive 报文和 Notification 报文。

在 BGP 对等体建立的过程中，通常可见的三个状态是：Idle、Active、Established。

BGP 对等体双方的状态必须都为 Established，BGP 邻居关系才能成立，双方通过 Update 报文交换路由信息。

BGP 处理过程

- 因为 BGP 的传输层协议是 TCP 协议，所以在 BGP 对等体建立之前，对等体之间首先进行 TCP 连接。BGP 邻居间会通过 Open 消息协商相关参数，建立起 BGP 对等体关系。
- 建立连接后，BGP 邻居之间交换整个 BGP 路由表。BGP 协议不会定期更新路由表，但当 BGP 路由发生变化时，会通过 Update 消息增量地更新路由表。
- BGP 会发送 Keepalive 消息来维持邻居间的 BGP 连接。当 BGP 检测到网络中的错误状态时（例如：收到不支持的协商能力或者收到错误报文时），BGP 会发送 Notification 消息进行报错，BGP 连接会随即中断。

BGP 属性

BGP 路由属性是一套参数，它对特定的路由进一步的描述，使得 BGP 能够对路由进行过滤和选择。事实上，所有的 BGP 路由属性都可以分为以下 4 类：

- 公认必须遵循的（Well-known mandatory）：所有 BGP 路由器都可以识别，且必须存在于 Update 消息中。如果缺少这种属性，路由信息就会出错。
- 公认任意（Well-known discretionary）：所有 BGP 路由器都可以识别，但不要求必须存在于 Update 消息中，可以根据具体情况来选择。
- 可选过渡（Optional transitive）：在 AS 之间具有可传递性的属性。BGP 路由器可以不支持此属性，但它仍然会接收这类属性，并传递给其他对等体。
- 可选非过渡（Optional non-transitive）：如果 BGP 路由器不支持此属性，则相应的这类属性会被忽略，且不会传递给其他对等体。

下面介绍几种常用的 BGP 路由属性：

- Origin 属性

  Origin 属性用来定义路径信息的来源，标记一条路由是怎么成为 BGP 路由的。它有以下 3 种类型：

  - IGP：具有最高的优先级。通过路由始发 AS 的 IGP 得到的路由信息，比如通过 **network** 命令注入到 BGP 路由表的路由，其 Origin 属性为 IGP。
  - EGP：优先级次之。通过 EGP 得到的路由信息，其 Origin 属性为 EGP。
  - Incomplete：优先级最低。通过其他方式学习到的路由信息。比如 BGP 通过 **import-route** 命令引入的路由，其 Origin 属性为 Incomplete。

- AS_Path 属性

  AS_Path 属性按矢量顺序记录了某条路由从本地到目的地址所要经过的所有 AS 编号。

  当 BGP Speaker 本地通告一条路由时：

  - 当 BGP Speaker 将这条路由通告到其他 AS 时，便会将本地 AS 号添加在 AS_Path 列表中，并通过 Update 消息通告给邻居路由器。
  - 当 BGP Speaker 将这条路由通告到本地 AS 时，便会在 Update 消息中创建一个空的 AS_Path 列表。

  当 BGP Speaker 传播从其他 BGP Speaker 的 Update 消息中学习到的路由时：

  - 当 BGP Speaker 将这条路由通告到其他 AS 时，便会把本地 AS 编号添加在 AS_Path 列表的最前面（最左面）。收到此路由的 BGP 路由器根据 AS_Path 属性就可以知道去目的地址所要经过的 AS。离本地 AS 最近的相邻 AS 号排在前面，其他 AS 号按顺序依次排列。
  - 当 BGP Speaker 将这条路由通告到本地 AS 时，不会改变这条路由相关的 AS_Path 属性。

- Next_Hop 属性

  BGP 的下一跳属性和 IGP 的有所不同，不一定就是邻居路由器的 IP 地址。通常情况下，Next_Hop 属性遵循下面的规则：

  - BGP Speaker 在向 EBGP 对等体发布某条路由时，会把该路由信息的下一跳属性设置为本地与对端建立 BGP 邻居关系的接口地址。
  - BGP Speaker 将本地始发路由发布给 IBGP 对等体时，会把该路由信息的下一跳属性设置为本地与对端建立 BGP 邻居关系的接口地址。
  - BGP Speaker 在向 IBGP 对等体发布从 EBGP 对等体学来的路由时，并不改变该路由信息的下一跳属性。

- MED

  MED（Multi-Exit-Discriminator）属性仅在相邻两个 AS 之间传递，收到此属性的 AS 一方不会再将其通告给任何其他第三方 AS。

MED 属性相当于 IGP 使用的度量值（Metrics），它用于判断流量进入 AS 时的最佳路由。当一个运行 BGP 的路由器通过不同的 EBGP 对等体得到目的地址相同但下一跳不同的多条路由时，在其它条件相同的情况下，将优先选择 MED 值较小者作为最佳路由。

- Local_Pref 属性

  Local_Pref 属性仅在 IBGP 对等体之间有效，不通告给其他 AS。它表明路由器的 BGP 优先级。

  Local_Pref 属性用于判断流量离开 AS 时的最佳路由。当 BGP 路由器通过不同的 IBGP 对等体得到目的地址相同但下一跳不同的多条路由时，将优先选择 Local_Pref 属性值较高的路由。

BGP 选择路由的策略

当到达同一目的地存在多条路由时，BGP 采取如下策略进行路由选择：

1. 优选协议首选值（PrefVal）最高的路由。

   协议首选值（PrefVal）是华为设备的特有属性，该属性仅在本地有效。
2. 优选本地优先级（Local_Pref）最高的路由。

   如果路由没有本地优先级，BGP 选路时将该路由按缺省的本地优先级 100 来处理。通过执行 **default local-preference** 命令可以修改 BGP 路由的缺省本地优先级。
3. 优选本地生成的路由（本地生成的路由优先级高于从邻居学来的路由）。

   本地生成的路由包括通过 **network** 命令或 **import-route** 命令引入的路由、手动聚合路由和自动聚合路由。

   a. 优选聚合路由（聚合路由优先级高于非聚合路由）。
   b. 通过 **aggregate** 命令生成的手动聚合路由的优先级高于通过 **summary automatic** 命令生成的自动聚合路由。
   c. 通过 **network** 命令引入的路由的优先级高于通过 **import-route** 命令引入的路由。
4. 优选 AS 路径（AS_Path）最短的路由。
   - AS_Path 的长度不包括 AS_CONFED_SEQUENCE 和 AS_CONFED_SET。
   - AS_SET 的长度为 1，无论 AS_SET 中包括多少 AS 号。
   - 执行 **bestroute as-path-ignore** 命令后，BGP 选路时，忽略 AS_Path 的比较。
5. 比较 Origin 属性，依次优选 Origin 类型为 IGP、EGP、Incomplete 的路由。
6. 优选 MED（Multi Exit Discriminator）值最低的路由。
   - BGP 只比较来自同一个 AS（不包括联盟的子 AS）的路由的 MED 值。即，只有两条路由的 AS_SEQUENCE（不包括 AS_CONFED_SEQUENCE）属性的第一个 AS 号相同时，BGP 才会比较二者的 MED 值。
   - 如果路由没有 MED 属性，BGP 选路时将该路由的 MED 值按缺省值 0 来处理；执行 **bestroute med-none-as-maximum** 命令后，BGP 选路时将该路由的 MED 值按最大值 4294967295 来处理。
   - 执行 **compare-different-as-med** 命令后，BGP 将强制比较来自不同自治系统中的邻居的路由的 MED 值。除非能够确认不同的自治系统采用了同样的 IGP 和路由选择方式，否则不要使用 **compare-different-as-med** 命令（可能产生环路）。
   - 执行 **bestroute med-confederation** 命令后，只有当 AS_Path 中不包含外部 AS 号（不属于联盟的子 AS），且 AS_CONFED_SEQUENCE 的第一个 AS 号相同时，才能比较 MED 值的大小。

- 执行 **deterministic-med** 命令后，将消除路由接收顺序对选路结果的影响。

7. 优选从 EBGP 邻居学来的路由（EBGP 路由优先级高于 IBGP 路由）。

   依次优选 EBGP 路由、IBGP 路由、LocalCross 路由、RemoteCross 路由。

   PE 上某个 VPN 实例的 VPNv4 路由的 ERT 匹配其他 VPN 实例的 IRT 后复制到该 VPN 实例，称为 LocalCross；从远端 PE 学习到的 VPNv4 路由的 ERT 匹配某个 VPN 实例的 IRT 后复制到该 VPN 实例，称为 RemoteCross。

8. 优选到 BGP 下一跳 IGP Metric 较小的路由。

   说明

   如果配置了负载分担，当上述所有规则相同，且存在多条 As_Path 完全相同的外部路由，则根据配置的路由条数选择多条路由进行负载分担。

9. 优选 Cluster_List 最短的路由。

10. 优选 Router ID 最小的路由器发布的路由。

    说明

    如果路由携带 Originator_ID 属性，选路过程中将比较 Originator_ID 的大小（不再比较 Router ID），并优选 Originator_ID 最小的路由。

11. 比较对等体的 IP Address，优选从具有较小 IP Address 的对等体学来的路由。

BGP 等价负载分担

当到达同一目的地址存在多条等价路由时，可以通过 BGP 等价负载分担实现均衡流量的目的。

形成 BGP 等价负载分担的条件是："BGP 选择路由的策略"的 1 至 8 条规则中需要比较的属性完全相同。

BGP 发布路由的策略

BGP 发布路由时采用如下策略：

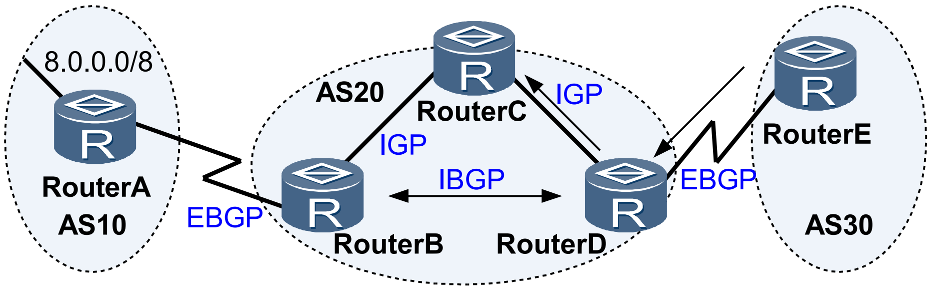
- 存在多条有效路由时，BGP Speaker 只将最优路由发布给对等体。
- BGP Speaker 从 EBGP 获得的路由会向它所有 BGP 对等体发布（包括 EBGP 对等体和 IBGP 对等体）。
- BGP Speaker 从 IBGP 获得的路由不向它的 IBGP 对等体发布。
- BGP Speaker 从 IBGP 获得的路由发布给它的 EBGP 对等体。
- 连接一旦建立，BGP Speaker 将把自己所有 BGP 路由发布给新对等体。

IBGP 和 IGP 同步

同步是指 IBGP 和 IGP 之间的同步，其目的是避免误导外部 AS 的路由器。

如果一个 AS 中有非 BGP 路由器提供转发服务，经该 AS 转发的 IP 报文将可能因为目的地址不可达而被丢弃。如图 8-3 所示，RouterE 通过 BGP 从 RouterD 可以学到 RouterA 的一条路由 8.0.0.0/8，于是将到这个目的地址的报文转发给 RouterD，RouterD 查询路由表，发现下一跳是 RouterB。由于 RouterD 从 IGP 学到了到 RouterB 的路由，所以通过路由迭代，RouterD 将报文转发给 RouterC。但 RouterC 并不知道去 8.0.0.0/8 的路由，于是将报文丢弃。

图 8-3 IBGP 和 IGP 同步



如果设置了同步特性，在 IBGP 路由加入路由表并发布给 EBGP 对等体之前，会先检查 IGP 路由表。只有在 IGP 也知道这条 IBGP 路由时，它才会被加入到路由表，并发布给 EBGP 对等体。

在下面的情况中，可以安全地关闭同步特性。

- 本 AS 不是过渡 AS（图 8-3 中的 AS20 就属于一个过渡 AS）
- 本 AS 内所有路由器建立 IBGP 全连接

8.4.2 路由引入

BGP 协议自身不能发现路由，所以需要引入其他协议的路由（如 IGP 或者静态路由等）注入到 BGP 路由表中，从而将这些路由在 AS 之内和 AS 之间传播。

BGP 引入路由时支持 Import 和 Network 两种方式：

- Import 方式是按协议类型，将 RIP 路由、OSPF 路由、ISIS 路由、静态路由和直连路由等某一协议的路由注入到 BGP 路由表中。
- Network 方式比 Import 方式更精确，将指定前缀和掩码的一条路由注入到 BGP 路由表中。

8.4.3 路由聚合

在大规模的网络中，BGP 路由表十分庞大，使用路由聚合（Routes Aggregation）可以大大减小路由表的规模。

路由聚合实际上是将多条路由合并的过程。这样 BGP 在向对等体通告路由时，可以只通告聚合后的路由，而不是通告所有的具体路由。

BGP 路由聚合支持两种方式：

- 自动聚合：对 BGP 引入的路由进行聚合。配置自动聚合后，对参加聚合的具体路由进行抑制，BGP 将按照自然网段聚合路由（如 10.1.1.1/24 和 10.2.1.1/24 将聚合为 A 类地址 10.0.0.0/8），并且 BGP 向对等体只发送聚合后的路由。
- 手动聚合：对 BGP 本地路由进行聚合。手动聚合可以控制聚合路由的属性，以及决定是否发布具体路由。

IPv4 支持自动聚合和手动聚合两种方式，而 IPv6 仅支持手动聚合。

8.4.4 路由衰减

路由不稳定的主要表现形式是路由振荡（Route Flapping），即路由表中的某条路由反复消失和重现。

说明

路由表中添加一条路由后，该路由又被撤销，这个过程称为一次路由振荡。

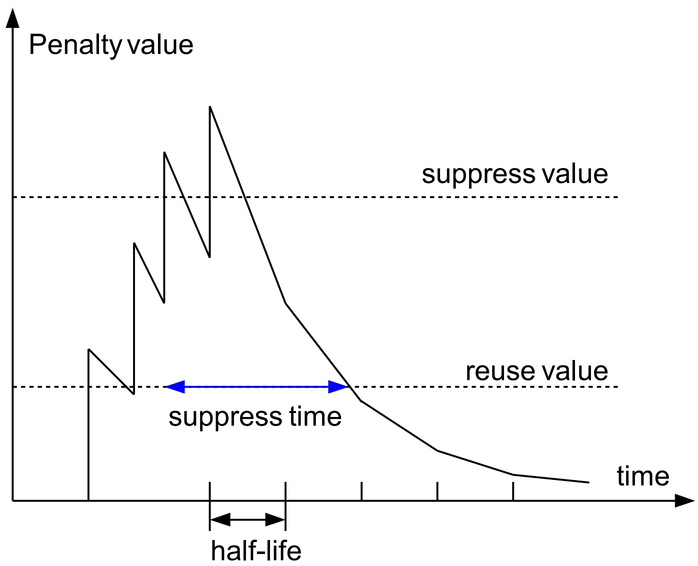
发生路由振荡时，设备就会向邻居发布路由更新，收到更新报文的设备需要重新计算路由并修改路由表。所以频繁的路由振荡会消耗大量的带宽资源和 CPU 资源，严重时会影响到网络的正常工作。

路由衰减（Route Dampening）用来解决路由不稳定的问题。多数情况下，BGP 协议都应用于复杂的网络环境中，路由变化十分频繁。为了防止持续的路由振荡带来的不利影响，BGP 使用路由衰减来抑制不稳定的路由。

BGP 衰减使用惩罚值（Penalty Value）来衡量一条路由的稳定性，惩罚值越高则说明路由越不稳定。路由每发生一次振荡（路由从激活状态变为未激活状态，称为一次路由振荡），BGP 便会给此路由增加一定的惩罚值（1000）。当惩罚值超过抑制阈值（Suppress Value）时，此路由被抑制，不加入到路由表中，也不再向其他 BGP 对等体发布更新报文。

被抑制的路由每经过一段时间，惩罚值便会减少一半，这个时间称为半衰期（Half-life）。当惩罚值降到再使用阈值（Reuse Value）时，此路由变为可用并被加入到路由表中，同时向其他 BGP 对等体发布更新报文。上文提到的惩罚值、抑制阈值和半衰期都可以手动配置。BGP 衰减的处理过程如图 8-4 所示。

图 8-4 BGP 衰减示意图



路由衰减只适用于 EBGP 路由。对于从 IBGP 收来的路由不能进行衰减，因为 IBGP 路由经常含有本 AS 的路由，内部网络路由要求转发表尽可能一致，IGP 快速收敛就是为了达到信息同步，转发一致。如果衰减对 IBGP 路由起作用，不同设备的衰减参数不一致时，会导致转发表不一致。

8.4.5 团体属性

团体属性（Community）是一组有相同特征的目的地址的集合。团体属性用一组以 4 字节为单位的列表来表示，设备中团体属性的格式是 aa:nn 或团体号。

- aa:nn：aa 和 nn 的取值范围都是 0 ～ 65535，管理员可根据实际情况设置具体数值。通常 aa 表示自治系统 AS 编号，nn 是管理员定义的团体属性标识。例如，来自 AS100 的一条路由，管理员定义的团体属性标识是 1，则该路由的团体属性格式是 100:1。
- 团体号：团体号是 0 ～ 4294967295 的整数。RFC1997 中定义，0（0x00000000）～ 65535（0x0000FFFF）和 4294901760（0xFFFF0000）～ 4294967295（0xFFFFFFFF）是预留的。

团体属性用来简化路由策略的应用和降低维护管理的难度，利用团体可以使多个 AS 中的一组 BGP 设备共享相同的策略。团体是一个路由属性，在 BGP 对等体之间传播，且不受 AS 的限制。BGP 设备在将带有团体属性的路由发布给其它对等体之前，可以先改变此路由原有的团体属性。

对等体组可以使一组对等体共享相同的策略，而团体可以使一组路由共享相同的策略。

除了使用公认的团体属性外，用户还可以使用团体属性过滤器过滤自定义扩展团体属性，以便更为灵活的控制路由策略。

公认的团体属性

BGP 公认的团体属性见表 **8-3**。

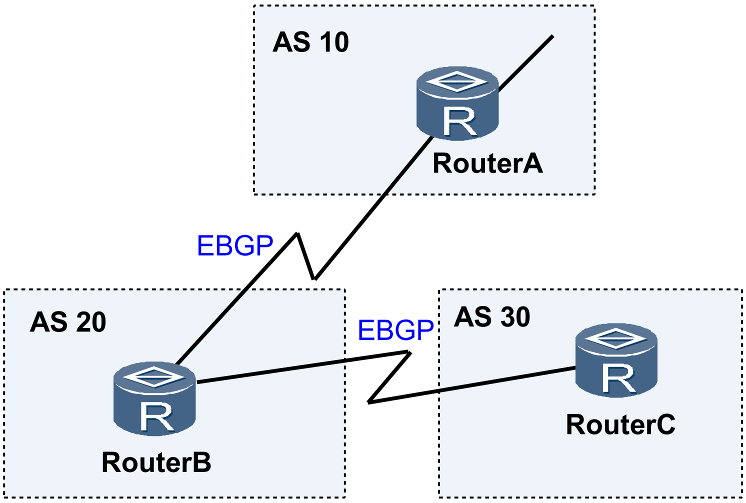
表 8-3 BGP 公认团体属性

团体名称	团体标识	说明
Internet	0（0x00000000）	缺省情况下，所有的路由都属于 Internet 团体。具有此属性的路由可以被通告给所有的 BGP 对等体。
No_Export	4294967041（0xFFFFFF01）	具有此属性的路由在收到后，不能被发布到本地 AS 之外。如果使用了联盟，则不能被发布到联盟之外，但可以发布给联盟中的其他子 AS。
No_Advertise	4294967042（0xFFFFFF02）	具有此属性的路由在收到后，不能被通告给任何其他的 BGP 对等体。
No_Export_Subconfed	4294967043（0xFFFFFF03）	具有此属性的路由在收到后，不能被发布到本地 AS 之外，也不能发布到联盟中的其他子 AS。

组网应用

如图 **8-5** 所示，RouterB 分别与 RouterA、RouterC 之间建立 EBGP 连接。通过在 RouterA 上配置 No_Export 团体属性，使得 AS10 发布到 AS20 中的路由，不再被 AS20 向其他 AS 发布。

图 8-5 配置 BGP 团体组网图



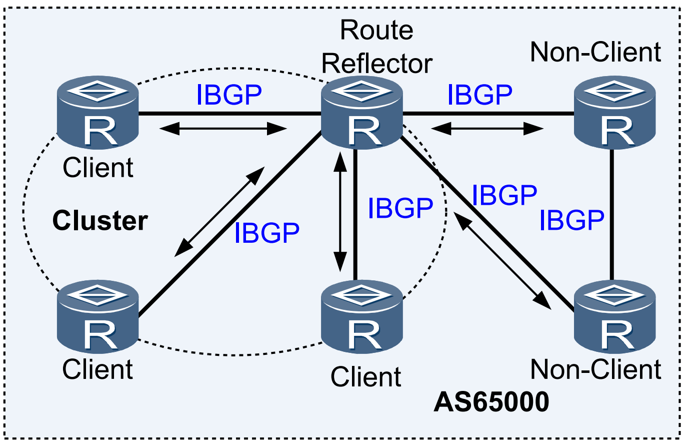
8.4.6 路由反射器

为保证 IBGP 对等体之间的连通性，需要在 IBGP 对等体之间建立全连接（Full-mesh）关系。假设在一个 AS 内部有 n 台路由器，那么应该建立的 IBGP 连接数就为 n(n-1)/2。当 IBGP 对等体数目很多时，对网络资源和 CPU 资源的消耗都很大。利用路由反射可以解决这一问题。

在一个 AS 内，其中一台路由器作为路由反射器 RR（Route Reflector），其它路由器作为客户机（Client）。客户机与路由反射器之间建立 IBGP 连接。路由反射器和它的客户机组成一个集群（Cluster）。路由反射器在客户机之间反射路由信息，客户机之间不需要建立 BGP 连接。

既不是反射器也不是客户机的 BGP 设备被称为非客户机（Non-Client）。非客户机与路由反射器之间，以及所有的非客户机之间仍然必须建立全连接关系。如图 **8-6** 所示。

图 8-6 路由反射器示意图



应用

当 RR 收到对等体发来的路由，首先使用 BGP 选路策略来选择最佳路由。在向 IBGP 邻居发布学习到的路由信息时，RR 按照 RFC2796 中的规则发布路由。

- 从非客户机 IBGP 对等体学到的路由，发布给此 RR 的所有客户机。
- 从客户机学到的路由，发布给此 RR 的所有非客户机和客户机（发起此路由的客户机除外）。
- 从 EBGP 对等体学到的路由，发布给所有的非客户机和客户机。

RR 的配置方便，只需要对作为反射器的路由器进行配置，客户机并不需要知道自己是客户机。

在某些网络中，路由反射器的客户机之间已经建立了全连接，它们可以直接交换路由信息，此时客户机到客户机之间的路由反射是没有必要的，而且还占用带宽资源。AR1200 支持配置命令 **undo reflect between-clients** 来禁止客户机之间的路由反射，但客户机到非客户机之间的路由仍然可以被反射。缺省情况下，允许客户机之间的路由反射。

Originator_ID 属性

RFC2796 定义了 Originator_ID 属性和 Cluster_List 属性，用于检测和防止路由环路。

Originator_ID 属性长 4 字节，由路由反射器（RR）产生，携带了本地 AS 内部路由发起者的 Router ID。

- 当一条路由第一次被 RR 反射的时候，RR 将 Originator_ID 属性加入这条路由，标识这条路由的发起路由器。如果一条路由中已经存在了 Originator_ID 属性，则 RR 将不会创建新的 Originator_ID。
- 当其他 BGP Speaker 接收到这条路由的时候，将比较收到的 Originator_ID 和本地的 Router ID，如果两个 ID 相同，BGP Speaker 会忽略掉这条路由，不做处理。

Cluster_List 属性

对于 AS 之间，BGP 用于防止环路的主要措施是通过 AS_Path 属性记录途经的 AS 路径，带有本地 AS 号的路由将被路由器丢弃；对于 AS 之内，BGP 防止路由环路的方法是禁止 IBGP 对等体发布从 AS 内部学来的路由。

RR 的实现是基于放宽对“BGP 在 AS 内学到的路由不会在 AS 中转发”的要求，即允许 IBGP 对等体之间发布从 AS 内部学来的路由。在这种情况下，Cluster_List 属性被引入，用于防止 AS 内部的环路。

路由反射器和它的客户机组成一个集群（Cluster）。在一个 AS 内，每个路由反射器使用唯一的 CLUSTER_ID 作为标识。

为防止产生路由环路，路由反射器使用 CLUSTER_LIST，记录反射路由经过的所有 CLUSTER_ID。

Cluster_List 由一系列的 Cluster_ID 组成，描述了一条路由所经过的反射器路径，这和描述路由经过的 AS 路径的 AS_Path 属性有相似之处。Cluster_List 由路由反射器产生。

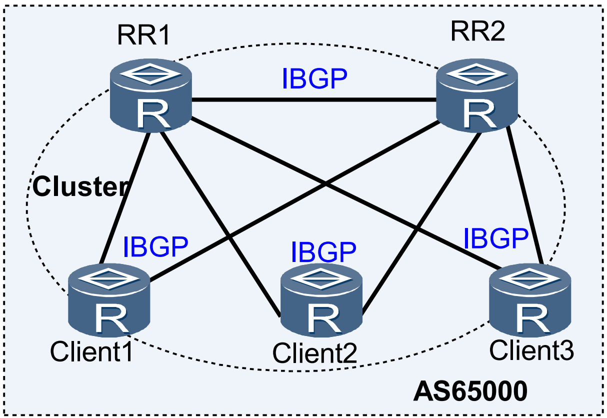
- 当 RR 在它的客户机之间或客户机与非客户机之间反射路由时，RR 会把本地 Cluster_ID 添加到 Cluster_List 的前面。如果 Cluster_List 为空，RR 就创建一个。
- 当 RR 接收到一条更新路由时，RR 会检查 Cluster_List。如果 Cluster_List 中已经有本地 Cluster_ID，丢弃该路由；如果没有本地 Cluster_ID，将其加入 Cluster_List，然后反射该更新路由。

备份 RR

为增加网络的可靠性，防止单点故障，有时需要在一个集群中配置一个以上的路由反射器。这时，相同集群中的路由反射器要共享相同的 Cluster_ID，以避免路由环路。AR1200 中需要使用命令 **reflector cluster-id** 给所有位于同一个集群内的路由反射器配置相同的 Cluster_ID。

在冗余的环境里，客户机会收到不同反射器发来的到达同一目的地的多条路由，这时客户机应用 BGP 选择路由的策略来选择最佳路由。

图 8-7 备份路由反射器



如图 **8-7**，路由反射器 RR1 和 RR2 在同一个 Cluster 内。RR1 和 RR2 之间配置 IBGP 连接，即两个反射器互为非客户机。

- 当客户机 Client1 从外部对等体接收到一条更新路由后，它通过 IBGP 向 RR1 和 RR2 通告这条路由。
- RR1 接收到该更新路由后，它向其他的客户机（Client2、Client3）和非客户机（RR2）反射，同时将本地 Cluster_ID 添加到 Cluster_List 前面。
- RR2 接收到该反射路由后，检查 Cluster_List，发现自己的 Cluster_ID 已经包含在 Cluster_List 中。因此，它丢弃该更新路由，不再向自己的客户机反射。

说明

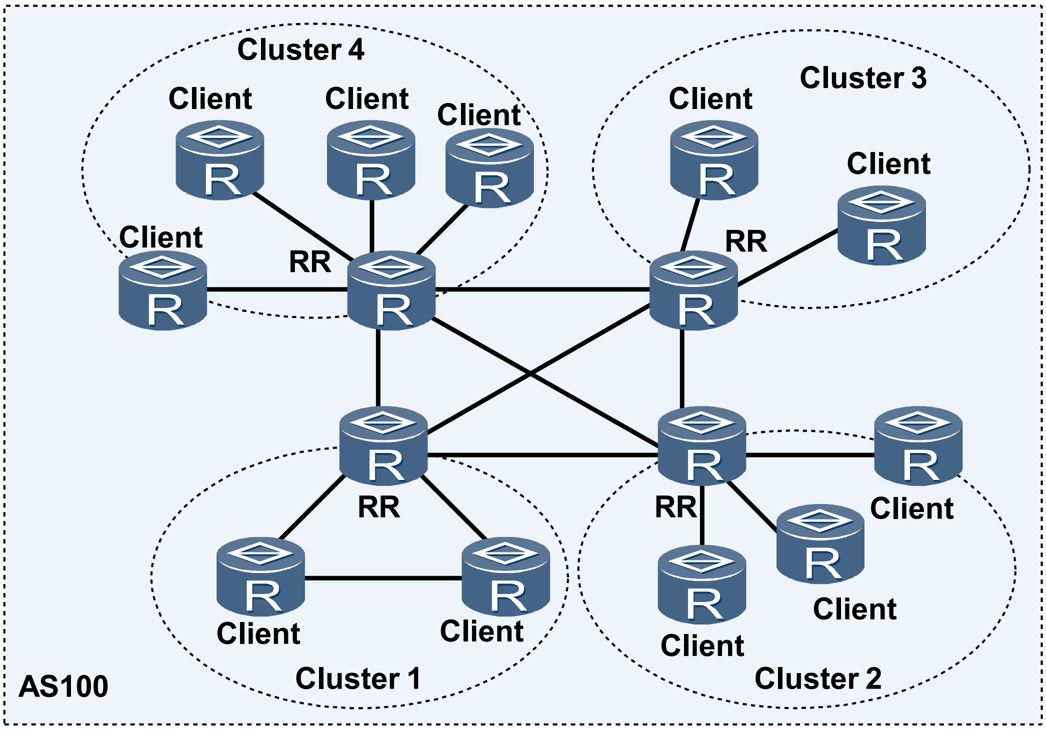
Cluster_List 的应用保证了同一 AS 内的不同 RR 之间不出现路由循环。

AS 内多个集群

一个 AS 中可能存在多个集群（Cluster）。各个 RR 之间是 IBGP 对等体的关系，一个 RR 可以把另一个 RR 配置成自己的客户机或非客户机。因此可以灵活的配置 AS 内部集群与集群之间的关系。

例如，一个骨干网被分成多个反射集群，每个 RR 将其它的 RR 配置成非客户机，各 RR 之间建立全连接。每个客户机只与所在集群的 RR 建立 IBGP 连接。这样该自治系统内的所有 BGP 路由器都会收到反射路由信息。如图 **8-8** 所示。

图 8-8 AS 内多个集群

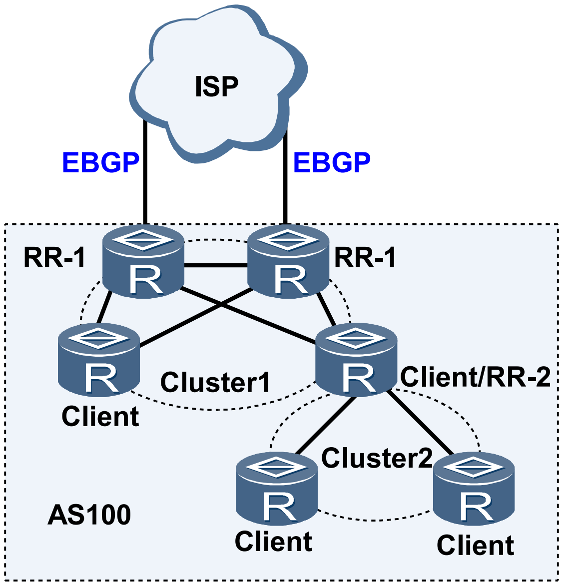


分级反射器

在实际的反射器部署中，常用的是分级反射器的场景。如图 **8-9**，ISP 为 AS100 提供 Internet 路由，ISP 与 AS100 内建立双出口 EBGP 连接。AS100 内部分为两个集群。Cluster1 内的四台路由器是核心路由器。

- Cluster1 中部署了两个一级 RR（RR-1），这种冗余结构保证了 AS100 内部网络核心层的可靠性。核心层其余两台路由器作为 RR-1 的客户机。
- Cluster2 中部署了一个二级 RR（RR-2），这个 RR-2 同时也是 RR-1 的客户机。

图 8-9 分级反射器



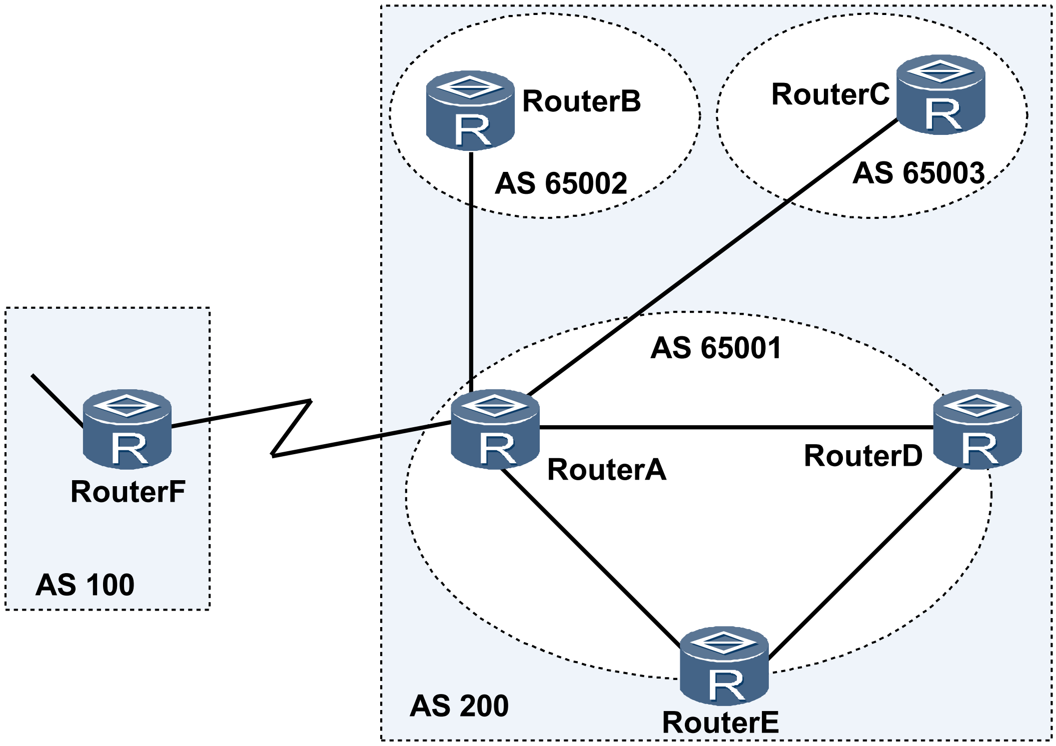
窍门

反射器场景下，如果 BGP 的优选路由不需要指导转发，通过配置 BGP-RIB-ONLY 特性，使所有 BGP 的优选路由都不加入 IP 路由表，也不进入转发层，从而提高转发效率和提升系统容量。

8.4.7 BGP 联盟

联盟（Confederation）是处理 AS 内部的 IBGP 网络连接激增的另一种方法，它将一个 AS 划分为若干个子自治系统（Sub AS），每个子 AS 内部建立 IBGP 全连接关系，子 AS 之间建立 EBGP 连接关系。如图 **8-10** 所示。

图 8-10 联盟示意图



在不属于联盟的 BGP Speaker（如 AS100 中的设备）看来，属于同一个联盟的多个子 AS（AS65001、AS65002、AS65003）是一个整体，外界不需要了解内部的子 AS 情况，联盟 ID 就是标识联盟这一整体的自治系统号，如图 **8-10** 中的 AS200 就是联盟 ID。

如图 **8-10** 所示，AS200 中有多台 BGP 设备，为了减少 IBGP 的连接数，现将他们划分为 3 个子自治系统：AS65001、AS65002 和 AS65003。其中 AS65001 内的三台设备建立 IBGP 全连接。

应用和限制

联盟需要在每个设备上进行配置，且要求加入联盟的设备具有联盟功能。

联盟的缺陷是：从非联盟向联盟方案转变时，要求设备重新进行配置，逻辑拓扑也要改变。

在大型 BGP 网络中，路由反射器和联盟可以被同时使用。

说明

同一联盟内不能同时配置 2 字节 AS 号的 Old Speaker 和 4 字节 AS 号的 New Speaker。因为 AS4_Path 不支持联盟，这种配置可能会引起环路。

8.4.8 MP-BGP

传统的 BGP-4 只能管理 IPv4 单播路由信息，对于使用其它网络层协议（如 IPv6、组播等）的应用，在跨 AS 传播时就受到一定限制。

为了提供对多种网络层协议的支持，IETF 对 BGP-4 进行了扩展，形成 MP-BGP，目前的 MP-BGP 标准是 RFC4760（Multiprotocol Extensions for BGP-4，BGP-4 的多协议扩

展）。MP-BGP 向前兼容，即支持 BGP 扩展的路由器与不支持 BGP 扩展的路由器可以互通。

MP-BGP 在现有 BGP-4 协议的基础上增强功能，使 BGP 能够为多种路由协议提供路由信息，包括 IPv6（即 BGP4+）和组播。

- MP-BGP 可以同时为单播和组播维护路由信息，将它们储存在不同的路由表中，保持单播和组播之间路由信息相互隔离。
- MP-BGP 可以同时支持单播和组播模式，为两种模式构建不同的网络拓扑结构。
- 原 BGP-4 支持的单播路由策略和配置方法大部分都可应用于组播模式，从而根据路由策略为单播和组播维护不同的路由。

扩展属性

BGP-4 使用的报文中，与 IPv4 相关的三处信息都由 Update 报文携带，这三处信息分别是：NLRI 字段、Next_Hop 属性、Aggregator 属性（该属性中包含形成聚合路由的 BGP Speaker 的 IP 地址）。

为实现对多种网络层协议的支持，BGP-4 需要将网络层协议的信息反映到 NLRI 及 Next_Hop。MP-BGP 中引入了两个新的路径属性：

- MP_REACH_NLRI：Multiprotocol Reachable NLRI，多协议可达 NLRI。用于发布可达路由及下一跳信息。
- MP_UNREACH_NLRI：Multiprotocol Unreachable NLRI，多协议不可达 NLRI。用于撤销不可达路由。

这两种属性都是可选非过渡（Optional non-transitive）的，因此，不提供多协议能力的 BGP Speaker 将忽略这两个属性的信息，不把它们传递给其它邻居。

地址族

BGP 采用地址族（Address Family）来区分不同的网络层协议，关于地址族的一些取值可以参考 RFC3232（Assigned Numbers）。AR1200 实现多种 MP-BGP 扩展应用，包括对 VPN 的扩展、对 IPv6 的扩展等，不同的扩展应在各自的地址族视图下配置。

MP-BGP 在组播中的详细介绍请参见《Huawei AR1200 系列 特性描述-组播》。

8.4.9 BGP GR

当 BGP 协议重启时会导致对等体关系重新建立和转发中断，使能平滑重启 GR（Graceful Restart）功能后可以避免流量中断。

在系统进行 GR 时有两种角色：

- GR Restarter：以 GR 方式重启的设备，指由管理员触发或故障触发重启的设备，必须是有 GR 能力的设备，即路由协议使能并协商了 GR 能力。
- GR Helper：GR Restarter 的邻居，本身必须具备了 GR 能力，才能协助 GR Restarter 进行 GR。

说明

AR1200 只能作为 Helper 路由器，不能作为 Restarter 路由器。

系统进行 GR 时有如下会话和定时器的概念：

- GR Session：GR 会话，是 GR Restarter 和 GR Helper 之间的协议关系。通过控制协议的会话协商机制，GR Restarter 和 GR Helper 可以了解彼此的 GR 能力，建立有 GR 能力的会话。
- GR Time：是 GR Helper 发现 GR Restarter Down 后，保持转发信息不删除的时间。当 GR Helper 发现对端的 GR Restarter 处于 Down 状态时，在 GR Time 时间内仍保留从 GR Restarter 得到的拓扑信息或路由，不删除这些信息。

BGP GR 的基本原理是：

- 利用 BGP 的能力协商机制，在 Restarting 前建立进行 GR 能力协商，建立有 GR 能力的 BGP 会话。
- 当邻居检查到 GR Restarter 发生重启时，不删除和 GR Restarter 相关的路由和转发表项，而是等待重建 BGP 连接。
- GR Restarter 和邻居在新连接上完成 BGP 路由更新。

这样既可以保证转发不中断，也可以让 BGP 协议的震荡仅限于和 Restarting 设备相连邻居，不会扩散到整个路由域，这对 BGP 这种路由数据大的协议来说，尤其有意义。

8.4.10 BGP 安全性

BGP 验证

BGP 使用 TCP 作为传输层协议，为提高 BGP 的安全性，可以在建立 TCP 连接时进行 MD5 认证。但 BGP 的 MD5 认证并不能对 BGP 报文认证，它只是为 TCP 连接设置 MD5 认证密码，由 TCP 完成认证。如果认证失败，则不建立 TCP 连接。

BGP 的 GTSM

通用 TTL 安全保护机制 GTSM（Generalized TTL Security Mechanism）机制通过对 TTL（Time-to-Live：生存时间字段，设置数据报文可以经过的路由器的最多数量）的检测来达到防止攻击的目的，如果攻击者模拟真实的 BGP 协议报文，对一台路由器不断的发送报文，路由器收到这些报文后，发现是发送给本机的报文，则直接上送控制层面的 BGP 协议处理，而不加辨别其“合法性”，这样导致路由器控制层面因为处理这些“合法”报文，系统异常繁忙，CPU 占用率高。

GTSM 通过检测 IP 报文头中的 TTL 值是否在一个预先定义好的特定范围内，对 IP 层以上业务进行保护，增强系统的安全性。

使能 BGP 的 GTSM 策略后，接口板对所有 BGP 报文的 TTL 值进行检查。根据实际组网的需要，对于不符合 TTL 值范围的报文，GTSM 可以设置为通过或丢弃。配置 GTSM 缺省动作为丢弃时，可以根据网络拓扑选择合适的 TTL 有限值范围，不符合 TTL 值范围的报文会被接口板直接丢弃，这样就避免了网络攻击者模拟的“合法”BGP 报文占用 CPU。

对于丢弃的报文，可以通过 LOG 信息开关，控制是否对报文被丢弃的情况记录日志，以方便故障的定位。

BGP 协议业务安全

- 限制从对等体接收的路由数量，防止资源耗尽性攻击。
- AS_Path 长度保护。通过在入口和出口两个方向对 AS_Path 的长度进行限定，直接丢弃 AS_Path 超限的报文。

8.4.11 BFD for BGP

BFD（Bidirectional Forwarding Detection）可为 BGP 协议提供更快速的链路故障检测。BGP 协议通过周期性的向对等体发送报文来实现邻居检测机制。但这种机制检测到故障所需时间比较长，超过 1 秒钟。当数据达到 Gbit/s 的速率等级时，这种机制的检测时间将导致大量数据丢失，无法满足电信级网络高可靠性的需求。

因此，BGP 协议通过引入 BFD for BGP 特性，利用 BFD 的快速检测机制（检测到故障的时间可以达到毫秒级）即迅速发现 BGP 对等体间链路的故障，并报告给 BGP 协议，从而实现 BGP 路由的快速收敛。

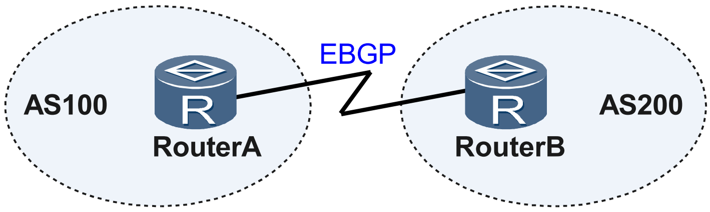
BFD for BGP 适用于 IPv4 网络。

组网

如图 **8-11** 所示，RouterA 和 RouterB 分别属于 AS100 和 AS200，两台路由器之间建立 EBGP 连接。

使用 BFD 检测 RouterA 和 RouterB 之间的 BGP 邻居关系，当 RouterA 和 RouterB 之间的链路发生故障时，BFD 能够快速检测到故障并通告给 BGP 协议。

图 8-11 BFD for BGP **组网图**



8.4.12 BGP Tracking

BFD 可为 BGP 协议提供更快速的链路故障检测，但是 BFD 需要全网部署，比较复杂，此时可以在本地配置 BGP Tracking 功能，同样达到快速检测链路故障的效果。

BGP 对等体使能 BGP Tracking 功能后，当对等体之间的链路发生故障时，BGP 对等体可快速感知邻居不可达，并中断与邻居之间的连接，同时删除从该邻居收到的路由，从而实现快速收敛。

BGP 发现邻居不可达到中断连接的时间间隔（*delay-time*）需配置合适，才可以保证网络的稳定性。

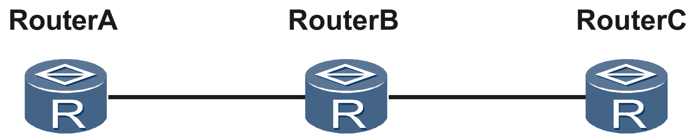
- 时间间隔为 0 时，BGP 发现邻居不可达后立即断开连接。
- 网络中的闪断会导致 IGP 路由震荡，如果 IBGP 邻居配置时间间隔为 0，则会导致邻居关系震荡。因此根据实际组网，IBGP 邻居配置的 *delay-time* 需要大于 IGP 路由收敛时间。
- BGP 邻居 GR 协商成功的情况下，BGP 邻居主备倒换，需配置 *delay-time* 大于 GR 收敛时间。如果时间间隔小于 GR 收敛时间，BGP 邻居会中断连接，导致 GR 失效。

BGP Tracking 可以加速网络的收敛，布局也比较方便。但是收敛速度比 BFD 慢，不适用于对收敛时间要求较高的语音业务。

组网

如图 **8-12** 所示，RouterA 和 RouterC 之间建立 IBGP 邻居。在 RouterA 上配置 BGP Tracking 功能，当 RouterA 和 RouterB 之间的链路发生故障时，IGP 快速收敛后会通知 RouterA 邻居 RouterC 不可达，RouterA 中断 BGP 连接。

图 8-12 BGP Tracking 组网图



8.4.13 BGP ORF

ORF（Outbound Route Filtering）是通过将本地的路由策略应用到邻居的出口，使邻居在发布路由时过滤掉无用路由。

用户希望运营商只发送自己需要的路由，而运营商又不想针对每个用户维护不同的出口策略。因此，运营商需要一个解决方案，在无需维护单个客户需求的情况下，能够满足用户对路由的过滤要求。在这种背景下，ORF 特性的产生能够很好的满足客户和运营商的需求。ORF 支持路由按需发布，在降低对带宽占用的同时，可以有效减少配置工作。

AR1200 支持以下两种 ORF：

- 基于前缀的 ORF

  RFC5291、RFC5292 规定了基于前缀的 ORF 能力，能将用户配置的基于前缀的入口策略通过 RR 报文发送给运营商，运营商将这些策略应用到出口，在路由发送时对路由进行过滤，避免用户接收大量无用路由，节省资源。

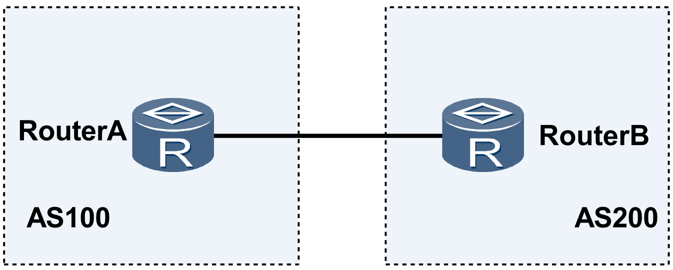
- VPN ORF

  RFC4684 规定了 VPN ORF 能力，使 PE 只接收本地期望的路由，减少接收压力，从而减少域内 RR、域间 ASBR 路由容量压力。

应用

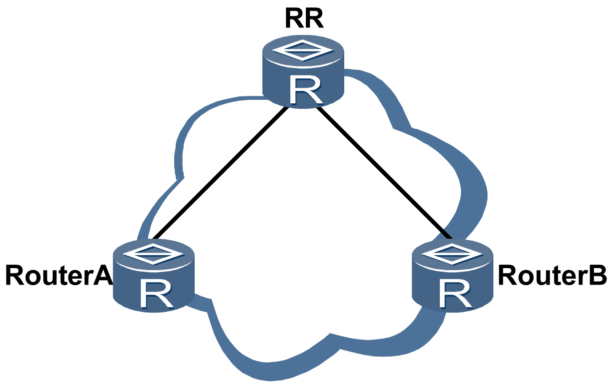
如图 **8-13** 所示，基本 BGP 邻居中，RouterA、RouterB 协商基于前缀的 ORF 能力后，RouterA 将本地配置的基于前缀的入口策略打包到 RR 报文中发送给 RouterB。RouterB 根据接收到的 RR 报文构造出口策略，指导发送路由给 RouterA。

图 8-13 基本 BGP 邻居



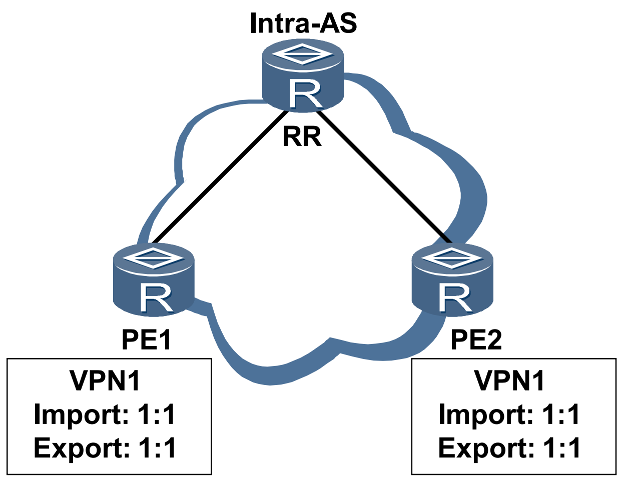
如图 **8-14** 所示，域内带 RR 的 BGP 邻居，RouterA、RouterB 为 RR 的客户端，RouterA 与 RR、RouterB 与 RR，分别协商基于前缀的 ORF 能力，RouterA、RouterB 将本地配置的基于前缀的入口策略打包到 RR 报文中发送给 RR。在 RR 上，根据接收到的前缀信息，构造出口策略，指导路由反射给 RouterA、RouterB。

图 8-14 域内 RR 场景



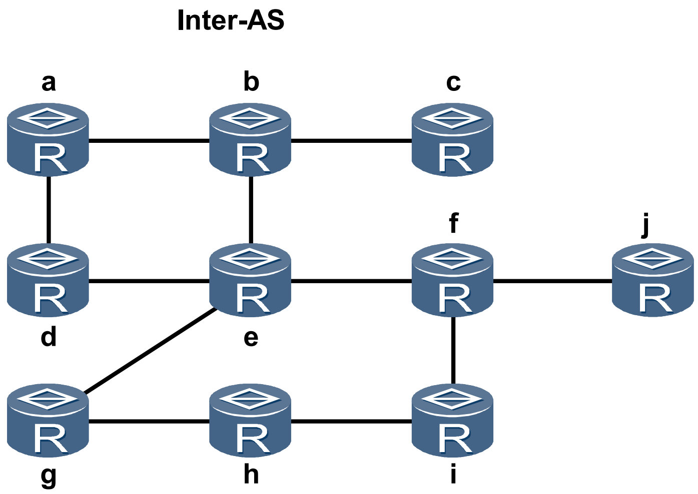
如图 **8-15** 所示，假设 PE1、PE2 上承载了相同的 VPN 业务 VPN1，并为这个 VPN 配置相同的 IRT（Import Route Target）。PE1、PE2 分别向 RR 发布这条 RT 路由，假设在 RR 上优选了 PE1 发布的，RR 不仅仅要向 PE1 发布 VPN 路由，也要向 PE2 发布 ERT（Export Route Target）为 1:1 的 VPN 路由。

图 8-15 域内 VPN 场景



如图 **8-16** 所示，a,b,c,d,e,f,g,h,i,j 是不同的 AS。假设从 i 起源一条 RT 路由，在节点 e 上将会有 2 条到达节点 i 的路径：(i,f,e), (i,h,g,e)。在 e 上优选(i,f,e)这条路径（因为 AS-PATH 较短），并将此路径前发给 b,d，由 b,d 发布给 a。假设 a 上优选了路径(e,b,a)，则 a 上 VPN 路由从 a 向 i 发布的路径为(a,b,e,f,i)，而次优路径(e,d,a), (i,h,g,e)上的节点 d,g,h 不会收到此 VPN 路由。

图 8-16 域间 VPN 场景



8.4.14 Active-Route-Advertise

目前，设备只向邻居发布 BGP 优选的路由，之前版本的发布规则为路由管理层优选才向邻居发布，为了前向兼容，设计实现 Active-Route-Advertise。

默认情况下路由只需在 BGP 中优选即可向邻居发布。配置了此特性之后，路由必须同时满足在 BGP 协议层面优选与在路由管理层面活跃两个条件，才能向邻居发布。

说明

引入的路由本身就是 IP 路由表中的活跃路由，不受命令 **active-route-advertise** 的影响。

8.4.15 BGP 按组打包

目前现网路由表的快速增长，以及网络拓扑的复杂性导致 BGP 需要支持更多的邻居。特别是一些邻居数目多且路由量大的场景下，针对设备需要给大量的 BGP 邻居发送路由，且大部分邻居具有相同出口策略的特点，要求较高的打包发包性能。

按组打包技术将所有拥有共同出口策略的 BGP 邻居当作是一个打包组。

在支持按组打包特性之前每条待发送路由需要针对每个邻居单独打包。按组打包实现了统一打包和分别发送，即每条待发送路由只被打包一次，然后发给组内的所有邻居，使得打包发包效率指数级提升。在邻居数目多且路由量大的场景下，按组打包极大的提高了 BGP 打包发包性能。

典型应用

按组打包的典型应用场景主要有以下三种情况，分别是：

- 国际关口局
- 反射器
- 从 EBGP 邻居收来路由向所有 IBGP 邻居发送

下面结合图形按序描述这三种场景。

图 8-17 国际关口局典型组网图

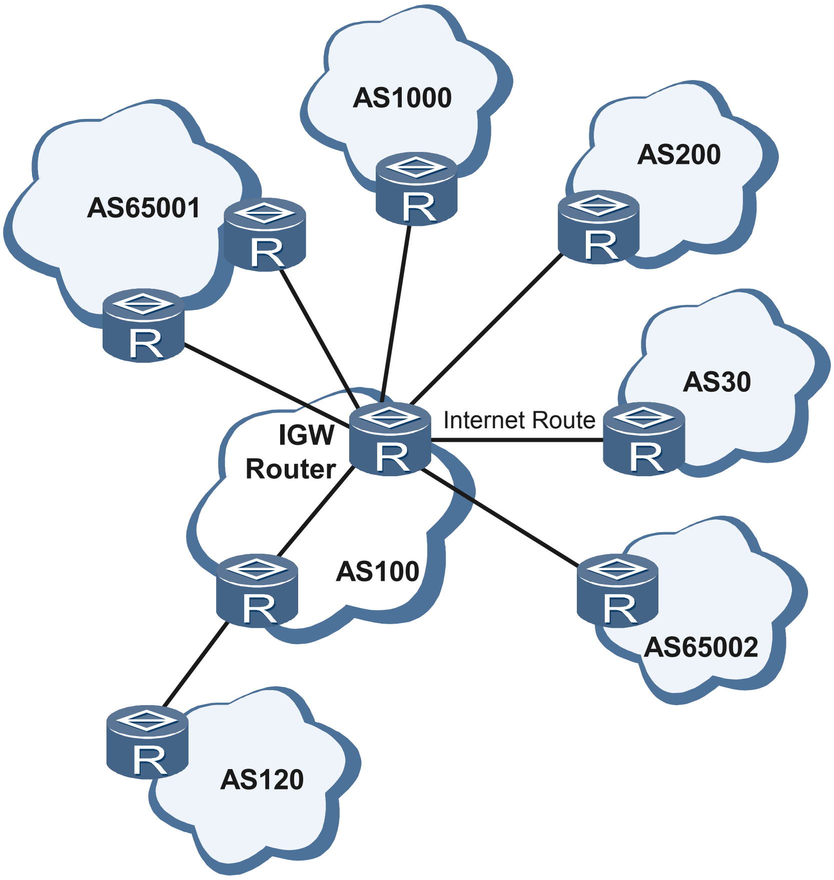


图 8-18 多个客户机的反射器典型组网图

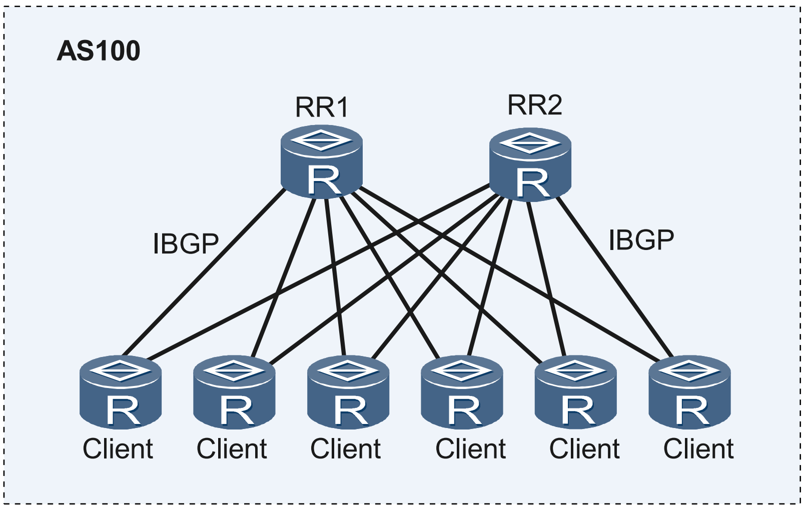
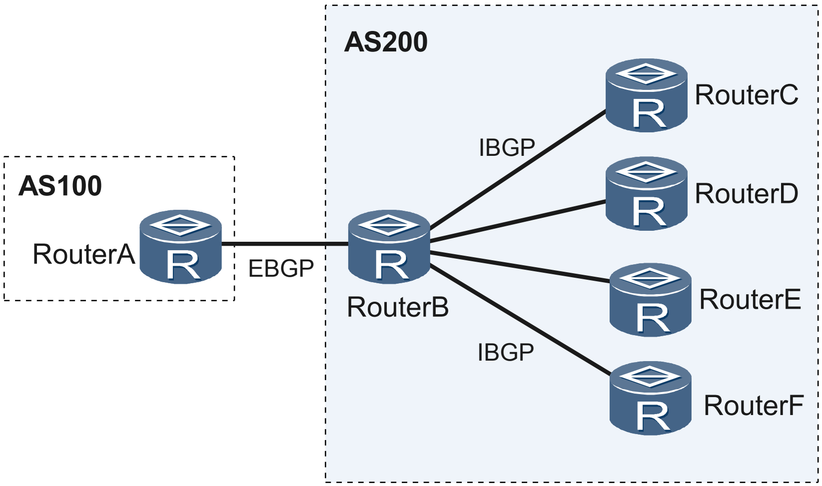


图 8-19 PE 与多个 IBGP 邻居连接典型组网图



上述三种场景都具有一个共性：一个设备要给大量的 BGP 邻居发送路由，同时大部分邻居拥有相同的出口策略，其中图 8-18 最典型。而且在邻居数目多且路由量大的场景下，它们的发包效率都是性能瓶颈。

应用按组打包技术后，每条待发送路由只被打包一次，然后发给组内的所有邻居，使得打包效率指数级提升。例如，一个反射器有 100 个客户机，有 100,000 条路由需要反射。如果按照每个邻居分别打包的方式，反射器 RR 在向 100 个客户机发送路由时候，所有路由被打包的总次数是 100,000×100。而按组打包技术将这个过程变为 100,000×1，性能相当于提升了 100 倍。

8.4.16 4 字节 AS 号

目前网络上使用的 AS 号范围为 0 至 65535（2 字节），随着时间推进，可分配的 AS 号已经濒临枯竭，需要将 AS 号范围扩展为 4 字节，且能够与仅支持 2 字节 AS 号的 Old speaker 兼容。

4 字节 AS 号特性是将 AS 号的编码范围由 2 字节扩大为 4 字节，并通过定义新的能力码和新的可选过渡属性来协商 4 字节 AS 号能力和传递 4 字节 AS 号信息，使支持 4 字节能力的 New speaker 之间、New speaker 和只支持 2 字节 AS 号的 Old speaker 之间能够进行通信。

- New Speaker：支持 4 字节 AS 号扩展能力的对等体。
- Old Speaker：不支持 4 字节 AS 号扩展能力的对等体。
- New Session：New Speaker 之间建立的 BGP 连接。
- Old Session：New Speaker 和 Old Speaker 之间或者 Old Speaker 之间建立的 BGP 连接。

协议扩展

为了支持 4 字节 AS 号，定义了一种新的 Open 能力码（0x41，代表本端支持 4 字节能力扩展）用于进行 BGP 连接的能力协商。

定义了 2 种新的可选过渡属性 AS4_Path（属性码为 0x11）和 AS4_Aggregator 属性（属性码为 0x12）用于在 Old Session 上传递 4 字节 AS 信息。

如果 New Speaker 和 Old Speaker 建立连接且 New Speaker 的 AS 号大于 65535，则在 Old Speaker 端需指定对端 AS 号为 AS_TRANS。其中 AS_TRANS 是保留 AS 号，值为 23456。

基本原理

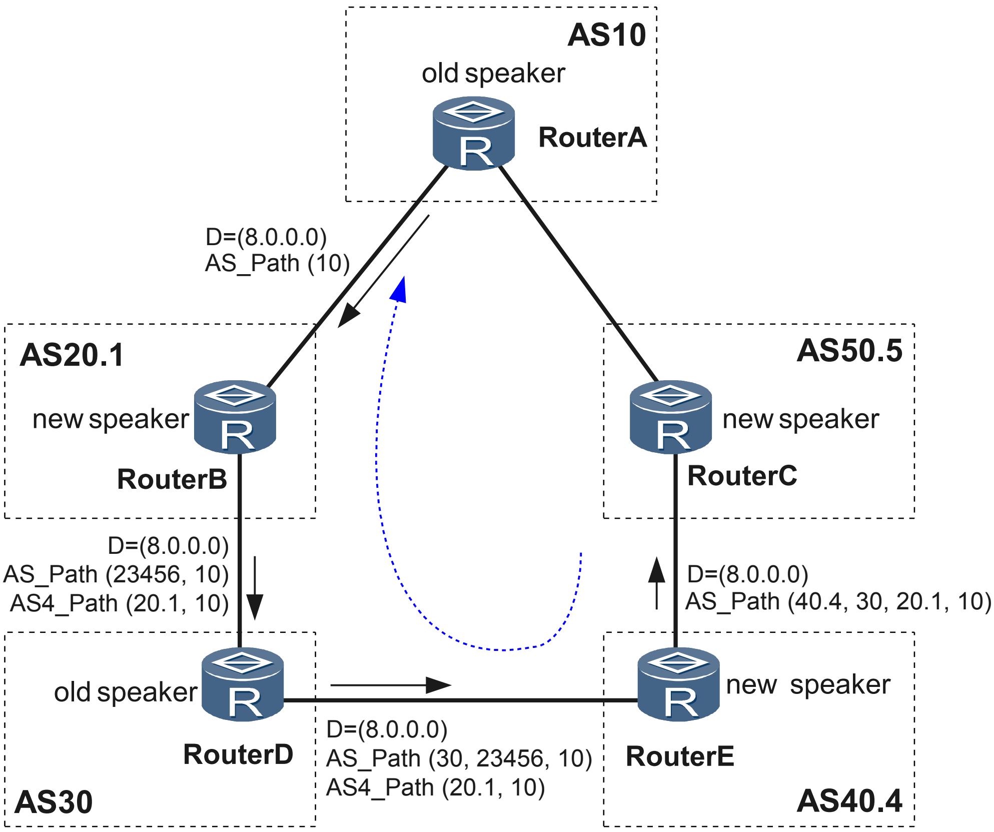
在邻居建立连接阶段，邻居间通过 Open 报文的可选能力字段获知对方是否支持 4 字节 AS 能力。

- New Speaker 之间建立 New Session，Update 报文的 AS_Path 属性和 Aggregator 属性中的 AS 号按照 4 字节进行编码。
- New Speaker 和 Old Speaker 之间建立 Old Session，由于 Old Speaker 中 AS_Path 属性和 Aggregator 属性中 AS 号都是按照 2 字节编码，
  - 当 New Speaker 向 Old Speaker 发送 Update 报文时，如果存在大于 65535 的 AS 号信息，会使用 AS4_PATH 属性和 AS4_Aggregator 属性辅助 AS_Path 属性和 AS_Aggregator 属性传递 4 字节 AS 号信息，AS4_PATH 属性和 AS4_Aggregator 属性对 Old Speaker 来说是完全透明的，
  - 当 New Speaker 从 Old Speaker 收到带有 AS_Path 属性、AS4_PATH 属性、AS_Aggregator 属性、AS4_Aggregator 属性的报文时，会根据重构算法重构出真正的 AS_Path 属性和 AS_Aggregator 属性。

组网

如图 8-20 所示，拓扑中既有仅支持 2 字节 AS 的 Old Speaker，又有支持 4 字节 AS 的 New Speaker。4 字节 AS 特性通过 AS4_PATH 属性的辅助，顺利的完成路由在 Old Speaker 和 New Speaker 之间的传递。

图 8-20 4 字节 AS 号典型组网图



BGP 要将自治系统 AS10 的路由 D=8.0.0.0 通告到其他 AS 时：

1. 首先把 AS10 的 AS 编号 10 放到 AS_Path 列表中（10）。
2. 在经过 AS20.1 后，为了能够让 RouterD（Old speaker）传递携带 4 字节 AS 号的 AS 路径信息，该路径开始携带 AS4_Path 属性（20.1，10），并又将自己的 AS 号 20.1 放到 AS_Path 列表的最左边（23456，10），其中 23456 就是使用 AS_TRANS 替代 20.1 的结果。
3. 在经过 AS30 后，RouterD 作为 Old Speaker，并不知道 AS4_Path 的存在，只是把收到的 AS4_Path（20.1，10）透传给 RouterE，并把自己的 AS 号 30 添加到 AS_Path 列表的最左边（30，23456，10）。
4. 在经过 AS40.4 后，通过对 AS_Path 和 AS4_Path 的重构处理，AS40.4 再将自己的 AS 编号 40.4 放到 AS_Path 列表的最左边（40.4，30，20.1，10）。

   以此类推，当 AS50.5 中的设备收到这条路由后，就可以根据 AS_Path 列表知道要去 AS10 的路径选择。

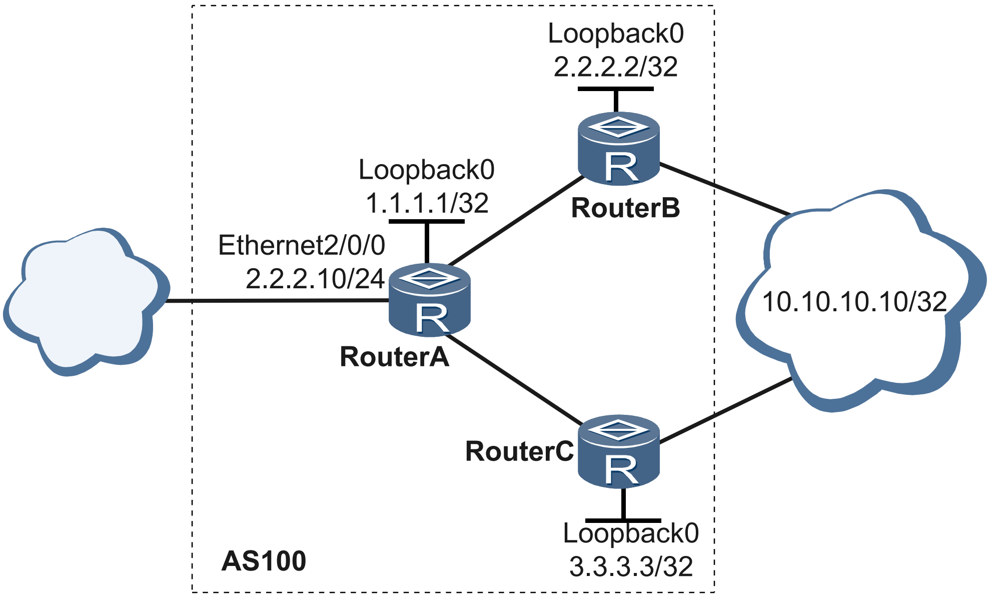
8.4.17 按策略进行下一跳迭代

BGP 需要对非直连的下一跳进行路由迭代，但是如果不对迭代到的路由进行过滤的话，可能会迭代到一个错误的转发路径上。按策略进行下一跳迭代就是通过配置路由策略来限制迭代到的路由。如果路由不能通过路由策略，则该路由迭代失败。

应用

如图 **8-21** 所示，RouterA 和 RouterB、RouterC 之间通过 Loopback 口建立 IBGP 邻居。RouterA 从 RouterB、RouterC 分别收到了前缀为 10.10.10.10/32 的 BGP 路由。其中从 RouterB 收到的 BGP 路由的原始下一跳为 2.2.2.2。另外，RouterA 上 GigabitEthernet2/0/0 的接口地址为 2.2.2.10/24。

图 8-21 按策略进行下一跳迭代组网图



当 RouterB 正常运行时，RouterA 收到从 RouterB 发来的前缀为 10.10.10.10/32 的路由会迭代到 IGP 路由 2.2.2.2/32。但是当 RouterB 发生故障时，IGP 路由 2.2.2.2/32 被撤销，这样就导致下一跳重新迭代。在 RouterA 上会用原始下一跳 2.2.2.2 在 IP 路由表中进行最长匹配迭代，结果会迭代到 2.2.2.0/24 的路由上。但此时用户期望的是，当到 2.2.2.2 的路由不可达时，可以重新选路优选到 3.3.3.3 的路由。

配置下一跳迭代策略，可以通过到 BGP 路由原始下一跳所依赖路由的掩码长度来过滤迭代路由。如**图 8-21** 中，可以通过配置下一跳迭代策略，使到原始下一跳 2.2.2.2 只能依赖于 2.2.2.2/32 的 IGP 路由。

8.5 术语与缩略语

术语

术语	解释
BGP	BGP 是一种用于自治系统 AS 之间的动态路由协议，与 OSPF、RIP 等内部网关协议（IGP）不同，其着眼点不在于发现和计算路由，而在于控制路由的传播和选择最佳路由。
BFD	Bidirectional Forwarding Detection－－双向转发检测，是一种通用的快速 Hello 报文检测机制，它的主要功能是快速感知链路状态的变化,使协议能快速的知道链路状态,是建立 Peer 还是断开它的 Peer。
Fake-AS	Fake-Autonomous System－－虚拟自治系统号。BGP Fake-AS 能够为 EBGP 对等体（组）配置一个虚拟的本地自治系统号，将本地真实的自治系统号隐藏起来，位于其他 AS 内的 EBGP 对等体只能看到这个虚拟 AS 号。通过 BGP Fake-AS 功能，邻居自治系统无需修改 BGP 配置就可以完成网络迁移。这可以保证现有的 BGP 业务尽量不中断。

缩略语

缩略语	英文全称	中文全称
BGP	Border Gateway Protocol	边界网关协议
VPN	Virtual Private Network	虚拟专用网
RM	Routing Management	路由管理
AS	Autonomous System	自治系统
ISP	Internet Service Provider	互联网服务提供商
EGP	Exterior Gateway Protocol	外部网关协议
IGP	Interior Gateway Protocol	内部网关协议
IBGP	Internal BGP	内部 BGP 关系
EBGP	External BGP	外部 BGP 关系
CE	Customer Edge	用户网络边缘设备

缩略语	英文全称	中文全称
PE	Provider Edge	服务提供商边缘设备
P	Provider	服务提供商网络中的骨干设备
MPLS	MultiProtocol Label Switch	多协议标签交换
LSP	Label Switched Path	标签交换路径
NLRI	Network Layer Reachability Information	网络层可达信息
CIDR	Classless Inter-Domain Routing	无分类域间路由
RR	Route Reflector	路由反射器
RIB	Route Information Base	BGP 路由表
MP-BGP	Multiprotocol Extensions for BGP	BGP 多协议扩展
GR	Graceful Restart	平滑重启
GTSM	Generalized TTL Security Mechanism	通用 TTL 安全保护机制
TTL	Time-To-Live	生存时间
6PE	IPv6 Provider Edge	IPv6 运营商边缘（设备）

9 路由策略

关于本章

9.1 介绍

定义

路由策略（Routing Policy）主要实现了对路由的过滤、设置等控制功能，通过改变路由属性（包括可达性）改变网络流量所经过的途径。

目的

路由器在发布、接收和引入路由信息时，根据实际组网需要实施一些策略，以便对路由信息进行过滤和改变路由信息的属性，如：

- 控制路由的发布

  只发布满足条件的路由信息。

- 控制路由的接收

  只接收必要、合法的路由信息，以控制路由表的容量，提高网络的安全性。

- 过滤和控制引入的路由

  一种路由协议在引入其它路由协议发现的路由信息丰富自己的路由知识时，只引入一部分满足条件的路由信息，并对所引入的路由信息的某些属性进行设置，以使其满足本协议的要求。

- 设置特定路由的属性

  为通过路由策略过滤的路由设置相应的属性。

9.2 参考标准和协议

无

9.3 可获得性

涉及网元

无需其他网元的配合。

License 支持

无需获得 License 许可，即可获得该特性的服务。

版本支持

AR1200 支持该特性的版本如下：

产品	最低支持版本
AR1200	V200R001C00

9.4 原理描述

9.4.1 路由策略的基本原理

路由策略的实现分为两个步骤：

1. 定义规则。首先要定义将要实施路由策略的路由信息的特征，即定义一组匹配规则，可以以路由信息中的不同属性作为匹配依据进行设置，如目的地址、AS 号等。
2. 应用规则。根据设置的匹配规则，再将它们应用于路由的发布、接收和引入等过程的路由策略中。

目前提供了如下几种过滤器供路由协议引用：

- 访问控制列表
- 地址前缀列表
- AS 路径过滤器
- 团体属性过滤器
- 扩展团体属性过滤器
- 路由标识属性过滤器

访问控制列表

访问控制列表包括针对 IPv4 报文的 ACL（Access Control List）。用户在定义 ACL 时可以指定 IP 地址和子网范围用于匹配路由信息的目的网段地址或下一跳地址。

地址前缀列表

地址前缀列表包括 IPv4 和 IPv6 地址前缀列表。

一个地址前缀列表由前缀列表名标识。每个前缀列表可以包含多个表项，每个表项可以独立指定一个网络前缀形式的匹配范围，并用一个索引号（index-number）来标识，索引号指明了进行匹配检查的顺序。

在匹配的过程中，设备按升序依次检查由索引号标识的各个表项。只要有某一表项满足条件，就意味着本次匹配过程结束，而不再进行下一个表项的匹配。

AS 路径过滤器

BGP 的路由信息中，包含一个自治系统路径域。As-path-filter 就是针对自治系统路径域指定匹配条件。AS 路径过滤器仅应用于 BGP 协议。

AS 路径属性参考文献：RFC1965。

团体属性过滤器

团体属性过滤器 Community-filter 仅用于 BGP。BGP 的路由信息中，包含一个 community 属性域，用来标识一个团体。Community-filter 就是针对团体属性域指定匹配条件。

团体属性的参考文献：RFC1997。

扩展团体属性过滤器

扩展团体属性过滤器 Extcommunity-filter 仅用于 BGP。目前我们只支持对 VPN 的 RT（Route-Target）扩展团体属性的过滤。

路由标识属性过滤器

路由标识属性列表 Rd-filter 仅用于 BGP。路由标识属性过滤器就是针对 VPN 的 RD（Route Distinguisher）属性指定匹配条件。

路由策略

路由策略最核心的内容就是它的匹配规则。

Route-policy 可以选择使用上述 6 种过滤器定义自己的匹配规则。一个 Route-policy 可以由多个节点（node）构成，不同节点之间是“或”的关系。系统按节点序号依次检查各个节点，如果通过了其中一节点，就意味着通过该策略，不再对其它节点进行匹配。

每个节点可以由一组 **If-match** 和 **Apply** 子句组成。**If-match** 子句定义匹配规则，匹配对象是路由信息的一些属性。同一节点中的不同 **If-match** 子句是“与”的关系，只有满足节点内所有 **If-match** 子句指定的匹配条件，才能通过该节点的匹配。**Apply** 子句指定动作，也就是在通过节点的匹配后，对路由信息的一些属性进行设置。

节点的匹配模式有两种：

- 允许模式（Permit）：当路由项满足该节点的所有 If-match 子句时被允许通过该节点的过滤，并执行该节点的 Apply 子句，如路由项不满足该节点的 If-match 子句，将继续匹配下一个节点。
- 拒绝模式（Deny）：当路由项满足该节点的所有 If-match 子句时，被拒绝通过并且不再匹配其他节点。

9.4.2 路由策略与策略路由的比较

表 9-1 路由策略与策略路由的比较

特性	特点
路由策略	● 基于目的地址按路由表转发； ● 基于控制平面，为路由协议和路由表服务； ● 与路由协议结合完成策略； ● 应用命令 **route-policy**。
策略路由	● 基于策略的转发，失败后再查找路由表转发； ● 基于转发平面，为转发策略服务； ● 需要手工逐跳配置，以保证报文按策略转发； ● 应用命令 **policy-based-route**。

9.4.3 组网应用

路由策略应用较灵活，无具体组网，主要有两种应用方式：

- **filter-policy** { **import** | **export** }

  import 策略定义了可接受的路由，即本端设备从它的对端设备接收什么样的路由。

  export 策略定义了可被路由协议对外发布的路由，即本端设备向它的对端设备发送什么样的路由。

- **import-route**（又称 Redistribute）

  import-route 策略定义了各个协议之间的路由交换，即不同协议之间路由信息的交换。缺省状态下，一个路由协议只对外发送由该协议找到的路由信息。**import-route** 策略使各个协议能进行路由交换，使其他协议的路由能被本协议发布。

9.4.4 地址前缀列表

功能

地址前缀列表即 IP-Prefix List。可以通过地址前缀列表，将与所定义的前缀过滤列表相匹配的路由，根据定义的匹配模式进行过滤，满足使用者的需要。

IP-Prefix 作为 IP 地址前缀过滤列表，可以通过在系统视图下进行 **ip ip-prefix** 命令行配置，可以通过配置关键字 **permit** 或者 **deny** 决定 IP 前缀过滤列表的匹配模式。前缀过滤列表由 IP 地址和掩码组成，IP 地址可以是网段地址或者主机地址，掩码长度的配置范围为 0 ～ 32。

基本实现

- IP-Prefix List 中的每一条 IP-Prefix 都有一个序列号 **index**，匹配的时候根据序列号从小到大进行匹配。
- 如果不配置 IP-Prefix 的 **index**，那么对应的 index 在上次配置的同名 IP-Prefix 的 index 的基础上，步长 10 进行增长。
- 如果未配置 IP-Prefix 的 **index**，则当以步长为 10 增长到大于等于 65525 时，再进行配置一条没有 **index** 设置的 IP-Prefix List 时，由于 **index** 不能超过 65535，所以这时采用上一次配置的 index 加 1 作为该次配置的 IP-Prefix 的 **index**。
- 如果未配置 IP-Prefix 的 **index**，则当以步长为 1 增长到大于等于 65535 时，再进行配置一条没有 **index** 设置的 IP-Prefix List 时，由于 **index** 不能超过 65535，所以 **index** 从 1 开始，以步长为 1 增长。
- IP-Prefix List 的数目受 LICENSE 限制，默认数目为 65535。
- 如果配置的 IP-Prefix 的名字与 **index** 都和已经配置了的一项 IP-Prefix List 的相同，仅仅只是匹配的内容不同，则该 IP-Prefix List 将覆盖原有的 IP-Prefix List。
- 当所有前缀过滤列表均未匹配时，缺省情况下，存在最后一条默认匹配模式为 **deny**。
- 当引用的前缀过滤列表不存在时，默认匹配模式为 **permit**。

前缀掩码长度范围

前缀过滤列表可以进行精确匹配或者在一定掩码长度范围内匹配，可以通过配置关键字 **greater-equal** 和 **less-equal** 指定待匹配的前缀掩码长度范围。如果没有配置关键字 **greater-equal** 或 **less-equal**，前缀过滤列表进行精确匹配，即只匹配掩码长度为前缀过滤列表掩码长度的相同 IP 地址路由；如果只配置了关键字 **greater-equal**，则待匹配的掩码长度范围为从 **greater-equal** 指定值到 32 位长度；如果只匹配了关键字 **less-equal**，则待匹配的掩码长度范围为从指定的掩码到关键字 **less-equal** 指定值。

greater-equal-value 与 *less-equal-value* 的取值限制：*mask-length*<=*greater-equal-value*<=*less-equal-value*<=32。

通配地址

0.0.0.0 为通配地址。当前缀为 0.0.0.0 时，可以在其后指定掩码以及掩码范围，不论掩码指定为多少，都表示掩码长度范围内的所有路由全部被 Permit 或 Deny。

关于通配地址的典型应用，请参见下文的“通配地址匹配”。

应用

假设有如下五条路由：1.1.1.1/24、1.1.1.1/32、1.1.1.1/26、2.2.2.2/24 和 1.1.1.2/16。

- 单节点匹配
  - Case1：

    ```
    ip ip-prefix aa index 10 permit 1.1.1.1 24
    ```

    匹配结果：路由 1.1.1.1/24 被 Permit，其他都被 Deny。

    说明：这种情况属于单节点的精确匹配，只有目的地址，掩码完全相同的路由才会匹配成功，而且节点的匹配模式为 Permit，所以路由 1.1.1.1/24 被 Permit，属于匹配成功并被 Permit，其他路由由于未匹配成功被 Deny。
  - Case2：

    ```
    ip ip-prefix aa index 10 deny 1.1.1.1 24
    ```

    匹配结果：路由全部被 Deny。

    说明：这种情况依然属于单节点的精确匹配，但节点的匹配模式为 **deny**，所以路由 1.1.1.1/24 还是被 Deny，属于匹配成功但被 Deny，其他路由则属于未匹配成功被默认 Deny。
- 多节点匹配
  - Case1：

    ```
    ip ip-prefix aa index 10 deny 1.1.1.1 24
    ip ip-prefix aa index 20 permit 1.1.1.1 32
    ```

    匹配结果：路由 1.1.1.1/24 被 Deny，路由 1.1.1.1/32 被 Permit，其他路由都被 Deny。

    说明：这种情况属于多节点的精确匹配：
    - 路由 1.1.1.1/24 在匹配 **index 10** 时，满足匹配条件，但匹配模式是 **deny**，属于匹配成功但被 Deny。
    - 路由 1.1.1.1/32 在匹配 **index 10** 时，不满足匹配条件，则继续匹配 **index 20**，此时匹配成功，且 **index 20** 的匹配模式是 **permit**，属于匹配成功并被 Permit。
    - 其他路由由于都不符合 **index 10** 和 **20** 的条件，属于未匹配成功被默认 Deny。
  - Case2：

    ```
    ip ip-prefix aa index 10 permit 1.1.1.1 24
    ```

    配置结果：没有配置关键字 **greater-equal** 或 *less-equal*（相当于 *greater-equal-value*=0，*less-equal-value*=0），前缀过滤列表进行精确匹配，即只匹配掩码长度为指定的掩码长度（此例为 24）的相同 IP 地址路由。

    匹配结果：路由 1.1.1.1/24 被 Permit，其他路由都被 Deny。
  - Case3：

    ```
    ip ip-prefix aa index 10 permit 1.1.1.1 24 less-equal 32
    ```

配置结果：此时 *greater-equal-value*=24，*less-equal-value*=32。

匹配结果：路由 1.1.1.1/24，1.1.1.1/32，1.1.1.1/26 被 Permit，其他路由被 Deny。

对 **greater-equal** 以及 *less-equal* 的说明：配置时，需满足 *mask-length*<=*greater-equal-value*<=*less-equal-value*，否则配置不成功。

- Case4：

```
ip ip-prefix aa index 10 permit 1.1.1.1 24 greater-equal 26
```

配置结果：此时 *greater-equal-value*=26，*less-equal-value*=32。

匹配结果：路由 1.1.1.1/32，1.1.1.1/26 被 Permit，其他路由被 Deny。

- Case5：

```
ip ip-prefix aa index 10 permit 1.1.1.1 24 greater-equal 26 less-equal 32
```

配置结果：此时 *greater-equal-value*=26，*less-equal-value*=32。

匹配结果：路由 1.1.1.1/32，1.1.1.1/26 被 Permit，其他路由被 Deny。

- 通配地址匹配
  - Case1：

```
ip ip-prefix aa index 10 permit 0.0.0.0 8 less-equal 32
```

配置结果：此时 *greater-equal-value*=8，*less-equal-value*=32，由于 0.0.0.0 为通配地址，路由全部被 Permit，即匹配所有掩码长度在 8 到 32 的路由。

匹配结果：所有掩码长度在 8 到 32 的路由都被 Permit。

  - Case2：

```
ip ip-prefix aa index 10 deny 0.0.0.0 24 less-equal 32
ip ip-prefix aa index 20 permit 0.0.0.0 0 less-equal 32
```

配置结果：对于 **index 10**，*greater-equal-value*=24，*less-equal-value*=32，由于 0.0.0.0 为通配地址，所有掩码长度在 24 到 32 的路由全部被 Deny；对于 **index 20**，*greater-equal-value*=0，*less-equal-value*=32，由于 0.0.0.0 为通配地址，所有其他路由全部被 Permit。

匹配结果：路由 1.1.1.2/16 被 Permit，其他路由被 Deny。

  - Case3：

```
ip ip-prefix aa index 10 deny 2.2.2.2 24
ip ip-prefix aa index 20 permit 0.0.0.0 0 less-equal 32
```

配置结果：对于 **index 10**，符合条件的路由 2.2.2.2/24 被 Deny，对于 **index 20**，其他路由都被 Permit。

匹配结果：除路由 2.2.2.2/24 外的其他路由被 Permit。

9.4.5 BGP to IGP

BGP to IGP 是路由策略的一个增强特性：当 IGP 通过路由策略引入 BGP 路由时，根据 BGP 路由的 Community、Extcommunity、AS-Path 等私有属性，来设置对应的路由属性值。

未实现该特性前，IGP 通过策略引入 BGP 路由时，因 IGP 无法识别 BGP 路由的 Community 等私有属性，则在过滤策略时，始终认为不匹配过滤条件而被拒绝，导致设置路由属性的 **apply** 子句不生效。一些特定场景下，如图 **9-1**，IGP 希望能根据 BGP 路由的 Community 等私有属性，来设置其 cost 值。由此产生了 BGP to IGP 这一特性。

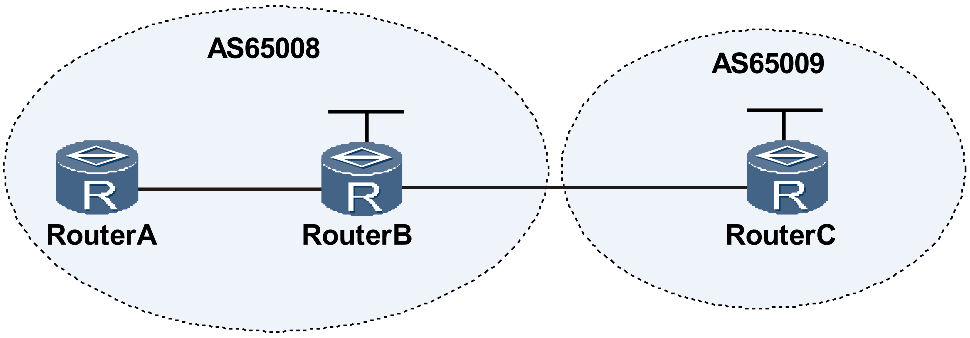
- 当 IGP 通过路由策略引入 BGP 路由时，根据 BGP 路由的 Community 等私有属性设置路由的 Cost 值。

- 若 BGP 路由携带 Community 等私有属性，系统会获取这些私有属性来进行策略过滤：若符合匹配条件，允许该路由通过策略，则 **apply** 子句生效。
- 若 BGP 路由未携带 Community 等私有属性，系统就会认为不符合匹配条件，拒绝该路由通过策略，则 **apply** 子句不生效。

BGP to IGP 应用

如图 **9-1** 所示，在 RouterA 和 RouterB 之间建立 IS-IS 邻居，同属一个自治系统。RouterB 和 RouterC 之间建立 EBGP 连接。RouterA 为 AS 内部的一台非 BGP 设备。当 IS-IS 协议引入 BGP 路由，并应用路由策略时，能通过匹配 BGP 路由的 Community 等私有属性来改变路由的 cost。

图 9-1 BGP to IGP 组网图



9.5 术语与缩略语

术语

术语	解释
PBR	Policy Based Routing——策略路由，是在路由表查找之前的 IP 转发流程。PBR 支持基于到达报文的源地址、报文长度等信息，依据用户制定的策略进行路由选择，可应用于安全、负载分担等目的

缩略语

缩略语	英文全称	中文全称
FIB	Forwarding Information Base	转发基本信息
IBGP	Internal Border Gateway Protocol	内部边界网关协议
EBGP	External Border Gateway Protocol	外部边界网关协议
VRP	Versatile Routing Platform	通用路由平台
ACL	Access Control List	访问控制列表

缩略语	英文全称	中文全称
USR	Unicast Static Route	单播静态路由
RM	Route Management	路由管理