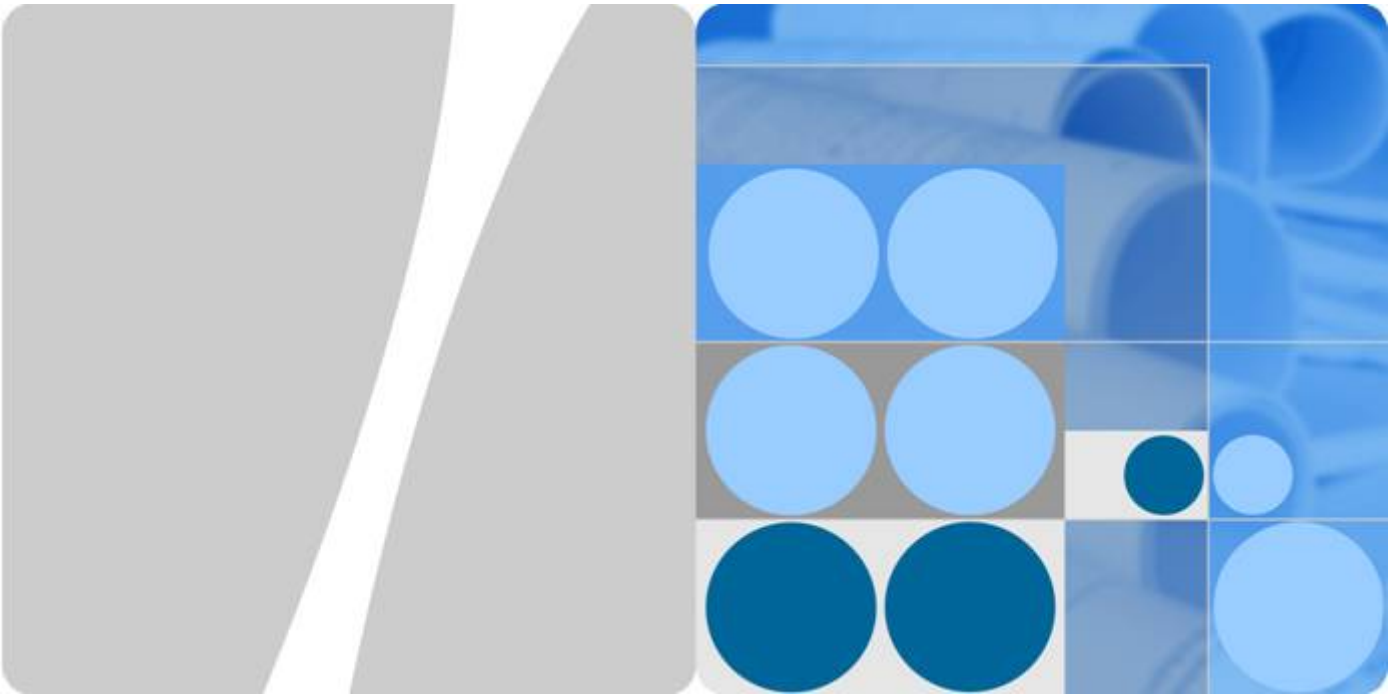


HUAWEI NetEngine20E-X6 高端业务路由器

V600R003C00

特性描述-IP 组播

文档版本　01

发布日期　2011-05-15

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

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： 0755-28560000 4008302118

客户服务传真： 0755-28560111

前 言

读者对象

本文档针对组播，从简介、原理描述和应用三个方面介绍了组播特性。

本文档与其它类型手册相结合，便于读者深入掌握特性的实现原理。

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。

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 (2011-05-15)

第一次正式归档。

目 录

插图目录

表格目录

1 IP 组播基础

关于本章

1.1 介绍

随着 Internet 网络的不断发展，网络中交互的各种数据、语音和视频信息越来越多，同时新兴的电子商务、网上会议、网上拍卖、视频点播、远程教学等服务也在逐渐兴起。这些服务大多符合点对多点的模式，对信息安全性、有偿性、网络带宽提出了较高的要求。

IP 数据传输

IP 数据传输基础是 IP 地址，Internet 使用 IP 地址标识并区分连接在网络上的各种设备。

- IP 报文使用 IP 地址标识发送对象，也就是报文目的地址。
- 用户主机根据接口能够识别的 IP 地址，来接收 IP 报文。
- 路由器根据 IP 报文的目的地址找出下一跳，执行转发。

IP 数据传输过程

IP 数据传输的大致过程如下：

1. 信源发送 IP 报文，目的地址字段使用目的主机能够识别的 IP 地址。
2. 路由器执行转发，将报文送达目的主机所在的网段。该网段可能同时连接多台用户主机。
3. 每一台用户主机检查网段内所有报文的目的地址，只接收自己能够识别的 IP 报文。如果同一网段内的各台用户主机能够识别的 IP 地址不同，则接收到的 IP 报文就不同。

用户主机能够识别的 IP 地址

用户主机能够识别的 IP 地址分为三类：

- 单播 IP 地址

  一个单播 IP 地址只能标识一台用户主机，一台用户主机只能识别一个单播 IP 地址。一份使用单播 IP 地址为目的地址的 IP 报文，只能被一台用户主机接收。
- 广播 IP 地址

  一个广播 IP 地址能够标识某确定网段内的所有用户主机，一台用户主机只能识别一个广播 IP 地址。一份使用广播 IP 地址为目的地址的 IP 报文，能够被该网段内的所有用户主机接收。IP 广播报文不能跨网段传播。
- 组播 IP 地址

  一个组播 IP 地址能够标识网络不同位置的多个用户主机，一台用户主机可以同时识别多个组播 IP 地址。一份使用组播 IP 地址为目的地址的 IP 报文，能够被网络不同位置的多个用户主机接收。

IP 传输三种方式

IP 传输分为三种方式，分别使用以上三类 IP 地址。

- IP 单播（Unicast），简称为单播。
- IP 广播（Broadcast），简称为广播。

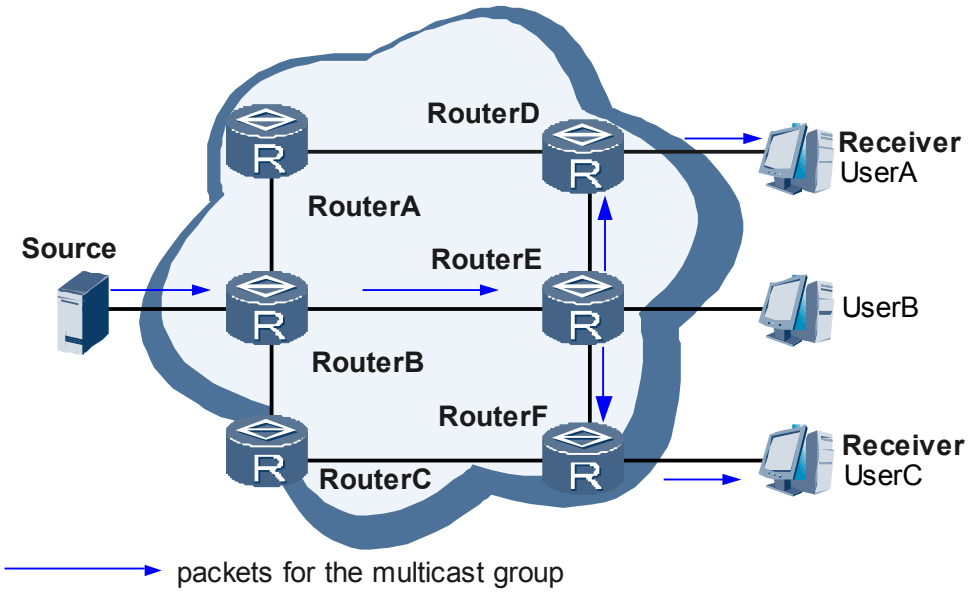
- IP 组播（Multicast），简称为组播。

应用以上三种传输方式，分别进行点对多点的数据传输。经过比较可以发现，组播在此方面更具优势。

- 单播方式
  - 单播的特点
    - 一份单播报文，使用一个单播地址作为目的地址。Source 向每个 Receiver 发送一份独立的单播报文。如果网络中存在 N 个接收者，则 Source 需要发送 N 份单播报文。
    - 网络为每份单播报文执行独立的数据转发，形成一条独立的数据传送通路。N 份单播报文形成 N 条相互独立的传输路径。
  - 单播的缺陷
    - 单播方式下，网络中传输的信息量和需求该信息的用户量成正比，当需求该信息的用户量较大时，网络中将出现多份相同信息流，不仅占用处理器资源而且浪费带宽。
    - 单播方式较适合用户稀少的网络，当用户量较大时很难保证网络传输质量。
- 广播方式
  - 广播的特点
    - 一份广播报文，使用一个广播地址作为目的地址。Source 向本网段对应的广播地址发送且仅发送一份报文。
    - 不管是否有需求，保证报文被网段中的所有用户主机接收。
  - 广播的缺陷
    - 广播方式下，信息发送者与用户主机被限制在一个共享网段中，且该网段所有用户主机都能接收到该信息。
    - 广播方式只适合共享网段，且信息安全性和有偿服务得不到保障。
- 组播方式

  如图 1-1 所示，网络中存在信息发送者 Source，UserA 和 UserC 提出信息需求，网络采用组播方式传输信息。

图 1-1 组播方式传输信息



- 组播的特点
  - 一份组播报文，使用一个组播地址作为目的地址。Source（组播源）向一个组播地址发送且仅发送一份报文。如图 1-1 所示：packets for the multicast group。
  - 网络中部署的组播协议为此组播报文建立一棵树型路由，根连接 Source，分支连接所有组播组成员。如图 1-1 所示：Source→ RouterB → RouterE [→ RouterD → UserA | → RouterF → UserC] 。
- 组播的优势
  - 组播方式下，单一的信息流沿组播分发树被同时发送给一组用户，相同的组播数据流在每一条链路上最多仅有一份。相比单播来说，使用组播方式传递信息，用户的增加不会显著增加网络的负载，减轻了服务器和 CPU 的负荷。
  - 组播报文可以跨网段传输，不需要此报文的用户不能收到此报文。相比广播来说，使用组播方式可以远距离传输信息，且只将信息传输到有接收者的地方，保障了信息的安全性。
  - 组播技术有效地解决了单点发送多点接收的问题，实现了 IP 网络中点到多点的高效数据传送。
- 组播的应用

  组播适用于任何“点到多点”的数据发布，主要包含以下几方面：
  - 多媒体、流媒体的应用
  - 培训、联合作业场合的通信
  - 数据仓库、金融应用（股票）

  如今 ISP 提供的互联网信息服务中，已经应用了 IP 组播技术。例如：在线直播、网络电视、远程教育、远程医疗、网络电台和实时视/音频会议等。

1.2 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
RFC1112	Host Extensions for IP Multicasting	-
RFC2236	Internet Group Management Protocol, Version 2	-
RFC3376	Internet Group Management Protocol, Version 3	-
RFC3810	Multicast Listener Discovery Version 2 (MLDv2) for IPv6	-
RFC4610	Protocol Independent Multicast - Sparse Mode (PIM-SM)	-
RFC3618	Multicast Source Discovery Protocol (MSDP)	-
RFC3973	Protocol Independent Multicast - Dense Mode (PIM-DM)	-

1.3 原理描述

1.3.1 基本概念

组播组

组播组使用一个 IP 组播地址标识。任何用户主机（或其他接收设备），加入一个组播组，就成为了该组成员，可以识别并接收以该 IP 组播地址为目的地址的 IP 报文。

组播源

以组播组地址为目的地址，发送 IP 报文的信源称为组播源。

- 一个组播源可以同时向多个组播组发送数据。
- 多个组播源可以同时向一个组播组发送报文。

组播组成员

组播组中的成员是动态的，网络中的用户主机可以在任何时刻加入和离开组播组。组成员可能广泛分布在网络中的任何地方。

组播源通常不会同时是数据的接收者，不属于组播组成员。

说明

下文以收看某电视频道的节目为例，可以帮助理解 IP 组播中的概念。

- 组播组是发送者和接收者之间的一个约定，如同电视频道。
- 电视台是组播源，它向某频道内发送数据。
- 电视机是接收者主机，观众打开电视机选择收看某频道的节目，表示主机加入某组播组；然后电视机播放该频道电视节目，表示主机接收到发送给这个组的数据。
- 观众可以随时控制电视机的开关和频道间的切换，表示主机动态的加入或退出某组播组。

组播路由器

网络中支持组播功能的路由器称为组播路由器。

组播路由器的功能：

- 在与用户主机连接的末梢网段，提供组播组成员管理功能。
- 实现组播路由，指导组播报文的转发。

组播分发树

根据组播组成员的分布情况，组播路由协议为多目的端的数据包转发建立树型路由。报文在距离组播源尽可能远的分叉路口才开始复制和分发，最终传送到组播组成员。

1.3.2 基本构架

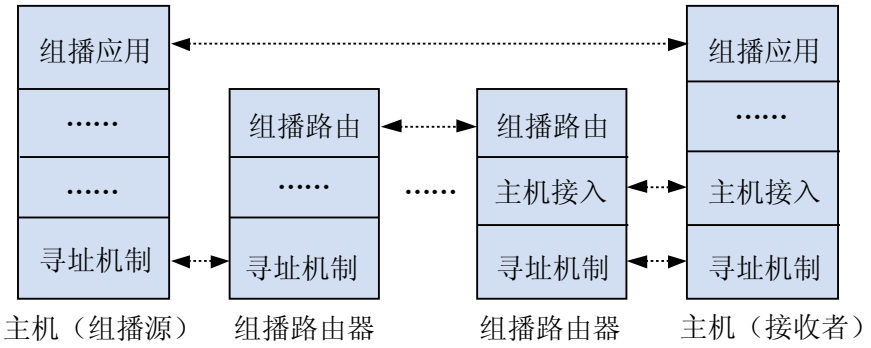
组播模型复杂多样，其目的都是以组播方式将信息从组播源传输到接收者手中，同时满足接收者对信息的各种需求。对于组播，需要关注的事项如表 1-1 所示。

表 1-1 组播技术关注事项

关注事项	组播技术
哪里有组播接收者	主机接入
从哪里可以获得组播数据	组播源发现技术
将组播数据传输到哪里	组播寻址机制
如何传输组播信息	组播路由

组播属于一种端到端服务，按照协议层从下往上划分，IP 组播基本构架包括寻址机制、主机接入、组播路由、组播应用四个部分，如图 1-2 所示。

图 1-2 IP 组播基本构架



- 寻址机制：使用组播地址，将一份数据报文发送给一组接收者。
- 主机接入：基于组播协议实现，允许用户主机动态加入和离开某组播组，实现组播成员管理。
- 组播路由：基于组播协议实现，构建报文分发树进行组播路由，从组播源传输报文到接收者。
- 组播应用：组播源与接收者必须支持视频会议等组播应用软件，TCP/IP 协议栈必须支持组播信息的发送和接收。

1.3.3 组播地址

如果采用组播方式传输信息，信息源该将信息发往何处，组播报文目的地址如何选取？这些问题简而言之就是组播寻址。

- 为了使信息源和组播组成员跨越互联网进行通讯，需要提供网络层组播，使用 IP 组播地址。
- 为了在本地物理网络上实现组播信息的正确传输，需要提供链路层组播，即硬件组播。当链路层应用以太网时，硬件组播使用组播 MAC 地址。
- 同时必须存在一种技术将 IP 组播地址映射为组播 MAC 地址。

IPv4 组播地址

IPv4 地址空间分为五类，即 A 类、B 类、C 类、D 类和 E 类。D 类地址为 IPv4 组播地址，用于标识组播组，使用在 IPv4 组播报文的目的地址字段。

IPv4 组播报文的源地址字段为 IPv4 单播地址，可使用 A、B 或 C 类地址，不能出现 D 类地址。E 类地址保留。

在网络层上，加入同一组播组的所有用户主机能够识别同一个 IPv4 组播组地址。一旦网络中某用户加入该组播组，则此用户就能接收以该组地址为目的地址的 IP 报文。

D 类组播地址范围是从 224.0.0.0 到 239.255.255.255，范围及含义见表 1-2。

表 1-2 D 类地址的范围及含义

D 类地址范围	含义
224.0.0.0 ～ 224.0.0.255	为路由协议预留的永久组地址。

D 类地址范围	含义
224.0.1.0 ～ 231.255.255.255 233.0.0.0 ～ 238.255.255.255	用户可用的 ASM 临时组地址，全网范围内有效。
232.0.0.0 ～ 232.255.255.255	用户可用的 SSM 临时组地址，全网范围内有效。
239.0.0.0 ～ 239.255.255.255	用户可用的 ASM 临时组地址，仅在特定的本地管理域内有效，称为本地管理组播地址。本地管理组播地址属于私有地址，在不同的管理域内使用相同的本地管理组播地址不会导致冲突。

- 永久组地址：IANA 为路由协议预留的组播地址（也称为保留组地址），用于标识一组特定的网络设备。具体请参见表 1-3。永久组地址保持不变，组成员的数量可以是任意的，甚至可以为零。
- 临时组地址：为用户组播组临时分配的 IPv4 地址（也称为普通组地址），组成员的数量一旦为零，即取消。

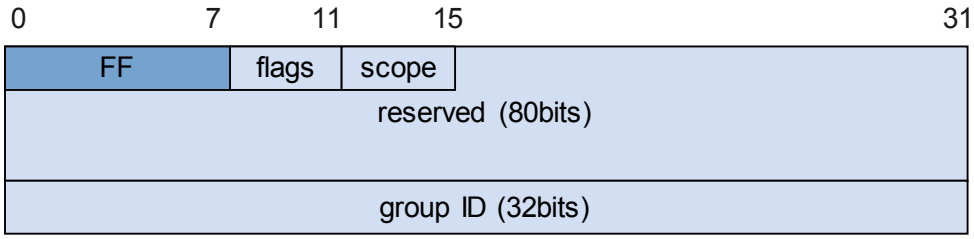
表 1-3 常见的永久组地址列表

永久组地址	含义
224.0.0.0	不分配
224.0.0.1	网段内所有主机和路由器（等效于广播地址）
224.0.0.2	所有组播路由器的地址
224.0.0.3	不分配
224.0.0.4	DVMRP 路由器
224.0.0.5	OSPF 路由器
224.0.0.6	OSPF DR
224.0.0.7	ST 路由器
224.0.0.8	ST 主机
224.0.0.9	RIP-2 路由器
224.0.0.11	移动代理
224.0.0.12	DHCP 服务器/中继代理
224.0.0.13	所有 PIM 路由器
224.0.0.14	RSVP 封装
224.0.0.15	所有 CBT 路由器
224.0.0.16	指定 SBM
224.0.0.17	所有 SBMS

永久组地址	含义
224.0.0.18	VRRP
224.0.0.19 ～ 224.0.0.21	未指定
224.0.0.22	所有使能 IGMPv3 的路由器
224.0.0.23 ～ 224.0.0.255	未指定

IPv6 组播地址

图 1-3 IPv6 组播地址的格式



IPv6 组播地址的格式如图 1-3 所示。

- IPv6 组播地址以 FF 开头。
- 标识字段（4 位），其含义如下：
  - 0：表示是 Internet 地址分配机构制定的熟知的多播地址
  - 1：表示是 ASM 范围的组播地址
  - 2：表示是 ASM 范围的组播地址
  - 3：表示是 SSM 范围的组播地址
  - 其他：未分配
- 范围字段（4 位）：用于指示组播组是只包含同一本地网络、同一站点、同一机构中的节点，还是包含全球地址空间内的任何节点。其含义如下：
  - 0：保留
  - 1：节点（或接口）本地范围（node/interface-local scope）
  - 2：链路本地范围（link-local scope）
  - 3：保留
  - 4：管理本地范围（admin-local scope）
  - 5：站点本地范围（site-local scope）
  - 8：机构本地范围（organization-local scope）
  - E：全球范围（globel scope）
  - F：保留
  - 其他：未分配

固定的 IPv6 组播地址的范围及含义如表 1-4。

表 1-4 IPv6 组播地址的范围及含义

范围	含义
FF0x::/32	Internet 地址分配机构制定的熟知的多播地址，具体请参见表 1-5
FF1x::/32（x 不能是 1 或者 2） FF2x::/32（x 不能是 1 或者 2）	任意源组播地址。全网范围内有效。
FF3x::/32（x 不能是 1 或者 2）	指定源组播地址。缺省的 SSM 组地址范围，全网范围内有效。

表 1-5 IPv6 常用多播地址范围及含义

范围	IPv6 组播地址	含义
节点（或接口）本地范围	FF01:0:0:0:0:0:0:1	所有节点（接口）地址
	FF01:0:0:0:0:0:0:2	所有路由器地址
链路本地范围	FF02:0:0:0:0:0:0:1	所有节点地址
	FF02:0:0:0:0:0:0:2	所有路由器地址
	FF02:0:0:0:0:0:0:3	未定义的地址
	FF02:0:0:0:0:0:0:4	DVMRP（Distance Vector Multicast Routing Protocol）路由器
	FF02:0:0:0:0:0:0:5	OSPF IGP Routers
	FF02:0:0:0:0:0:0:6	OSPF IGP Designated Routers
	FF02:0:0:0:0:0:0:7	ST 路由器
	FF02:0:0:0:0:0:0:8	ST 主机
	FF02:0:0:0:0:0:0:9	RIP 路由器
	FF02:0:0:0:0:0:0:A	EIGRP 路由器
	FF02:0:0:0:0:0:0:B	移动代理（Mobile-Agents）
	FF02:0:0:0:0:0:0:D	所有 PIM 路由器
	FF02:0:0:0:0:0:0:E	RSVP-ENCAPSULATION
	FF02:0:0:0:0:0:1:1	Link Name
	FF02:0:0:0:0:0:1:2	所有 DHCP 代理
	FF02:0:0:0:0:1:FFXX:XXXX	Solicited-Node 地址，XX:XXXX 表示节点 IPv6 地址的后 24 位
站点本地范围	FF05:0:0:0:0:0:0:2	所有路由器地址

范围	IPv6 组播地址	含义
	FF05:0:0:0:0:0:1:3	所有 DHCP 服务器
	FF05:0:0:0:0:0:1:4	所有 DHCP 中继（DHCP relay）
	FF05:0:0:0:0:0:1:1000 ～ FF05:0:0:0:0:0:1:13FF	服务位置（Service Location）

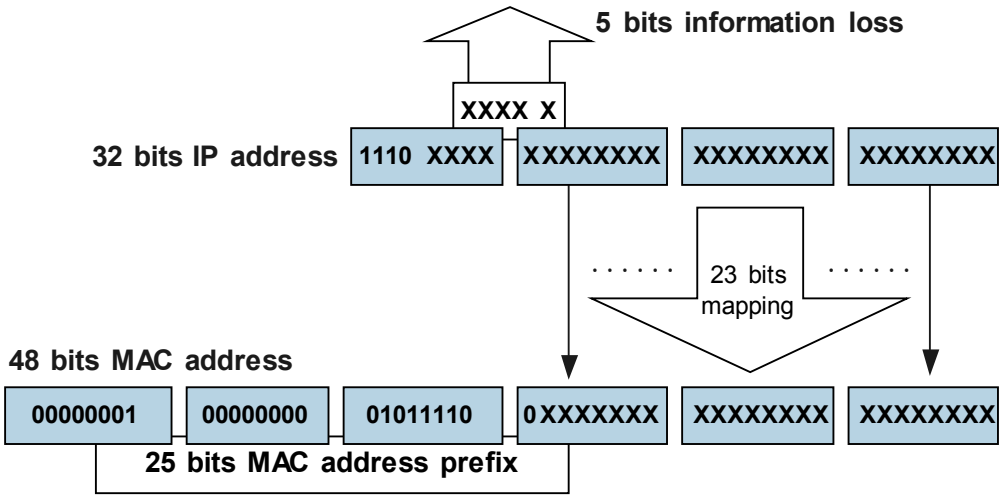
以太网组播 MAC 地址

组播 MAC 地址用于在链路层上标识属于同一组播组的接收者。

网络设备上的以太网接口板可以识别组播 MAC 地址。通过在驱动程序中配置某组播 MAC 地址，设备就可以在以太网上接收和转发该组播组的数据。

IANA 规定，组播 MAC 地址的高 25bit 为 0x01005e，第 25bit 为 0，低 23bit 为组播 IP 地址的低 23bit，映射关系如图 1-4 所示。

图 1-4 组播 IP 地址与组播 MAC 地址的映射关系



IP 组播地址的前 4bit 是固定的 1110，对应组播 MAC 地址的高 25bit。IP 组播地址的后 28bit 中只有 23bit 被映射到 MAC 地址，因此丢失了 5bit 的地址信息，直接结果是有 32 个 IP 组播地址映射到同一 MAC 地址上。

说明

本手册重点介绍 IP 组播技术及设备操作。如果不加特别说明，本手册中出现的组播均指 IP 组播。

1.3.4 组播协议

实现一套完整的组播服务，需要在网络各个位置部署多种组播协议相互配合，共同运作，如图 1-5 和图 1-6 所示。

图 1-5 IPv4 组播网络

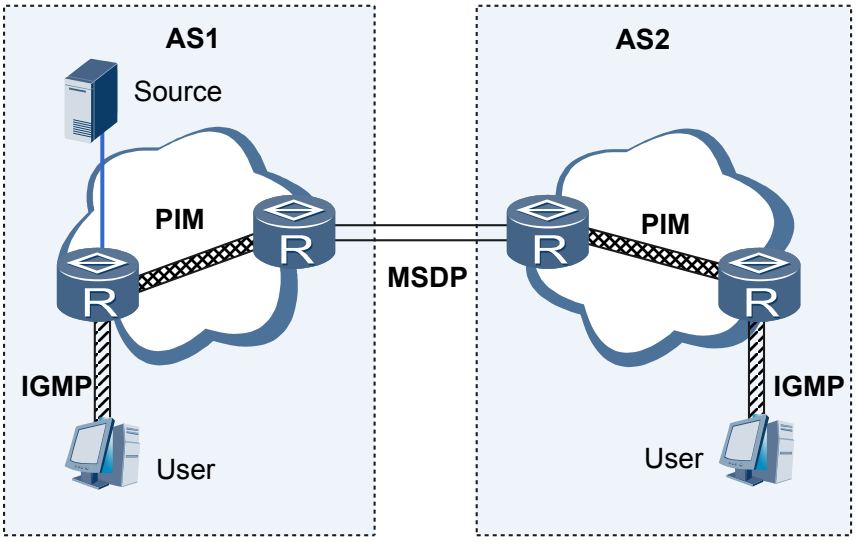
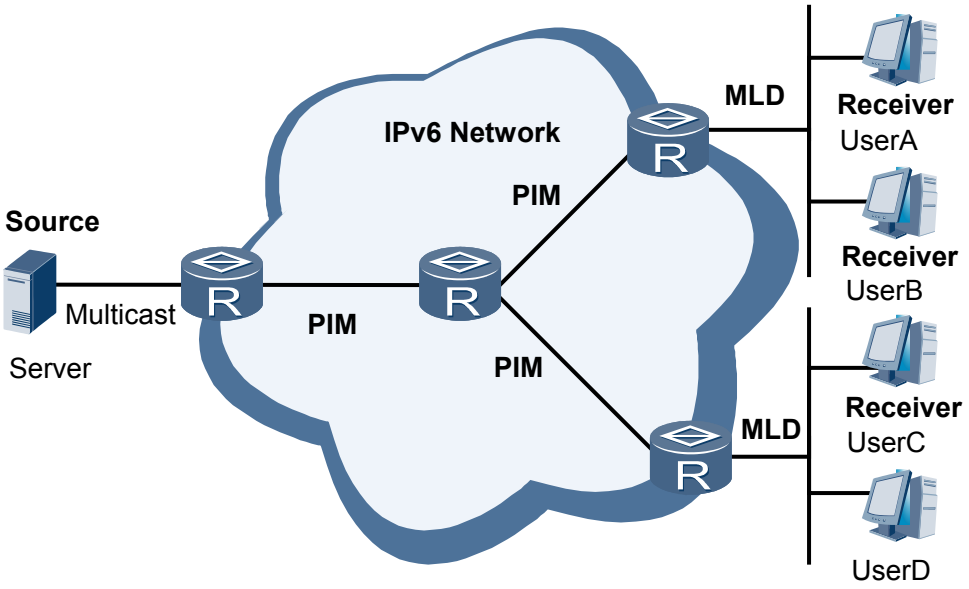


图 1-6 IPv6 组播网络



针对不同的应用位置与目的，NE20E-X6 开发出功能各异的多种组播路由协议，如表 1-6 所示。

表 1-6 组播协议

应用位置	目的	组播协议
用户主机与组播路由器之间	将用户主机接入组播网络： ● 在主机侧实现组播组成员动态加入与离开。 ● 在路由器侧实现组成员关系的维护与管理，同时支持与上层组播路由协议的信息交互。	组成员关系管理协议有： ● IGMP（Internet Group Management Protocol）因特网组管理协议，用于IPv4 网络。 ● MLD（Multicast Listener Discovery）组播监听器发现协议，用于 IPv6 网络。
域内组播路由器之间	组播路由与转发： ● 按需创建组播路由。 ● 动态响应网络拓扑变化，维护组播路由表。 ● 按照路由表项执行转发。	PIM（Protocol Independent Multicast）协议无关组播。
域间组播路由器之间	域间组播源信息共享： ● 源所在域内的路由器将本地源信息传播给其他域内的路由器。 ● 不同域的路由器之间传递源信息。	MSDP（Multicast Source Discovery Protocol）组播源发现协议，目前仅用于 IPv4 网络。

从表 1-6 中观察可知，组播协议的功能主要分为以下两类：

组播组成员关系管理

组播组成员关系管理是指在主机与路由器之间建立和维护组成员关系。

IGMP 是用于 IPv4 网络的组播组成员关系管理协议，有以下特点：

- 包含三个版本，分别是 IGMPv1、IGMPv2 和 IGMPv3。新版本完全兼容旧版本。目前应用最广泛的是 IGMPv2。
- 三个版本都支持 ASM 模型；IGMPv3 可以直接支持 SSM 模型，而 IGMPv1 和 IGMPv2 需要结合 SSM-Mapping 技术才能支持 SSM 模型。

MLD 是用于 IPv6 网络的组播组成员关系管理协议，有以下特点：

- 包含两个版本，分别是 MLDv1 和 MLDv2。
- MLDv1 的功能与 IGMPv2 相似。
- MLDv2 的功能与 IGMPv3 相似。
- 两个版本都支持 ASM 模型；MLDv2 可以直接支持 SSM 模型，而 MLDv1 需要结合 SSM-Mapping 技术才能支持 SSM 模型。

建立并维护组播路由

组播路由也称为组播分发树，指从一个组播源到所有组成员的数据传输路径。组播路由单向、无环且路径最短。通过在路由器之间建立和维护组播路由，网络才能够正确、高效地转发组播数据包。

- 域内组播路由协议：用来在自治系统 AS（Autonomous System）内发现组播源并构建组播分发树，将信息传递到接收者。PIM 是典型的域内组播路由协议，有两套独立的模式：
  - DM（Dense Mode）：适用于小规模、接收者分布较为密集的情况，支持 ASM 模型。
  - SM（Sparse Mode）：适用于大规模、接收者分布较为稀疏的情况，同时支持 ASM 模型和 SSM 模型。
- 域间组播路由协议：用来在 AS 之间传递组播源信息，从而跨域建立组播路由，实现域间组播资源共享。MSDP 是典型的域间组播路由协议，通常与 MBGP 协同工作。MSDP 适用于各域内运行 PIM-SM 的情况。

对于 SSM 模型来说，没有域内和域间的划分。由于接收者预先知道组播源的具体位置，因此可以借助 PIM-SM 的部分功能直接创建组播传输路径。

1.3.5 组播模型分类

根据对组播源的控制程度的不同，IP 组播分为三种模型，分别为：

- ASM 模型
- SFM 模型
- SSM 模型

ASM 模型

ASM 全称为 Any-Source Multicast，译为任意源组播。在 ASM 模型中，任意发送者都可以成为组播源，向某组播组地址发送信息。接收者加入该组播组后，能够接收到发往该组播组的所有信息。

在 ASM 模型中，接收者无法预先知道组播源的位置，接收者可以在任意时间加入或离开该组播组。

SFM 模型

SFM 全称为 Source-Filtered Multicast，译为过滤源组播。SFM 模型继承了 ASM 模型，从发送者角度来看，组播组成员关系相同。

同时，SFM 在功能上对 ASM 进行了扩展：上层软件对接收到的组播报文的源地址进行检查，允许或禁止来自某些组播源的报文通过。最终，接收者只能接收到来自部分组播源的数据。从接收者角度来看，只有部分组播源是有效的，组播源经过了筛选。

说明

SFM 在 ASM 的基础上添加了组播源过滤策略，此外基本原理和配置方法相同。本手册中将 SFM 与 ASM 统称为 ASM。

SSM 模型

SSM 全称为 Source-Specific Multicast，译为指定源组播。在现实生活中，用户可能仅对某些源发送的组播信息感兴趣，而不愿接收其它源发送的信息。SSM 模型为用户提供了一种能够在客户端指定信源的传输服务。

SSM 模型和 ASM 模型的根本区别是接收者已经通过其他手段预先知道了组播源的具体位置。SSM 和 ASM 使用不同的组播地址范围，直接在接收者和组播源之间建立组播转发树。

1.3.6 组播报文转发

组播报文转发和单播报文转发相互隔离，互不影响。

在组播模型中，IP 报文的目的地址字段为组播组地址，组播源向以此目的地址所标识的主机群组传送信息。因此，转发路径上的路由器为了将组播报文传送到各个方位的接收站点，往往需要将从一个入接口接收到的组播报文，从多个出接口转发出去。与单播模型相比，组播模型的复杂性就在于此。

- 由组播路由表来指导组播报文转发。
- 由 RPF（Reverse Path Forwarding）机制保证组播路由是一棵最短路径树。RPF 机制是大部分组播路由协议创建组播路由表项、进行组播转发的基础。

组播报文转发的过程为：

- 对于 ASM，设备接收到组播数据时，先查找 MFIB，如果 MFIB 中存在此表项，根据此转发表项进行转发。 如果 MFIB 中没有此表项，则转发平面通知 PIM 协议创建组播路由表，由 PIM-SM 协议根据用户的加入信息创建组播路由表下发到 MFIB，然后指导组播数据的转发。
- 对于 SSM，由 PIM-SSM 协议根据用户的加入信息创建组播路由表直接下发到 MFIB。当接收到组播数据时，查找 MFIB，如果 MFIB 中存在此表项，根据此转发表项进行转发；否则直接丢弃。

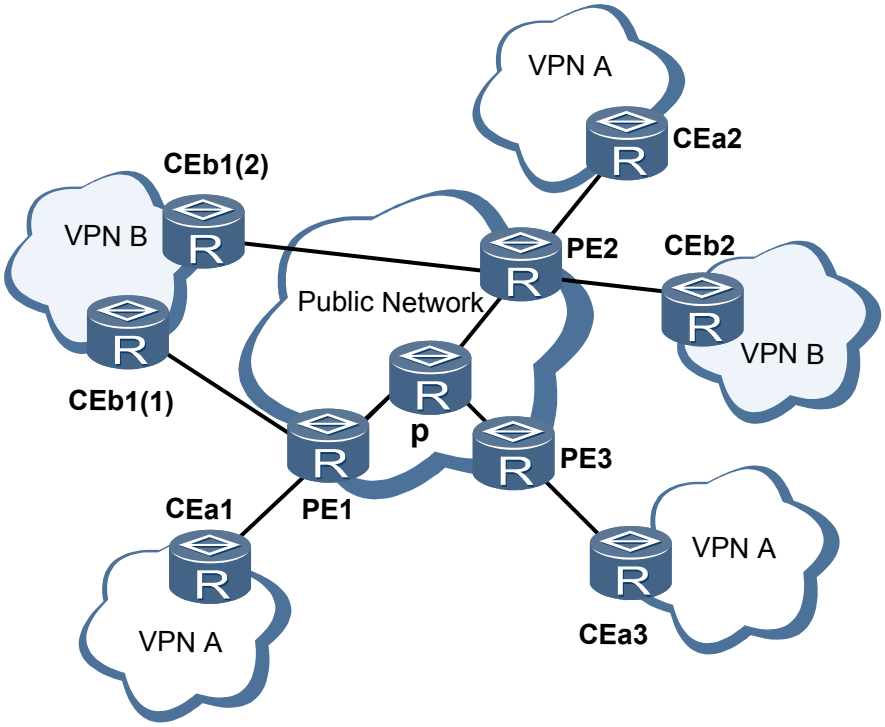
1.4 应用

多实例组播

多实例组播是指在 IPv4 VPN（Virtual Private Network）中应用的组播。

各 VPN 网络、VPN 网络和公共网络之间要求信息隔离。如图 1-7 所示，VPN A、VPN B 通过 PE 路由器接入公共网络。

图 1-7 VPN 典型组网



- P 专属于公网，各 CE 专属于某一 VPN。每个路由器只为其专属网络服务，仅维护一套转发机制。
- PE 同时接入公网和 VPN 网络，同时为多个网络服务。在路由器上必须严格区分各个网络的信息并为各个网络独立维护一套转发机制。这时，PE 上为同一网络服务的一套软硬件设施统称为一个实例。PE 上同时存在多实例，同一实例分布在多个 PE 上。

说明

关于多实例的更多内容请参见《HUAWEI NetEngine20E-X6 高端业务路由器 特性描述 VPN》。

多实例在组播中的应用

HUAWEI NetEngine20E-X6 支持多实例组播。在 PE 上应用多实例技术后，具备以下功能：

- 为每个实例独立维护一套组播转发机制：支持各种组播协议，拥有自己的 PIM 邻居列表、组播路由表等信息。每个实例转发组播数据时只查找本实例的转发表或路由表。
- 保证各个实例之间相互隔离。
- 实现公网实例和 VPN 实例之间的信息交流和数据转换。

多实例组播是实现跨越 VPN 传播组播数据的基础。以此为前提，VRP 开发出组播 VPN 技术。以图 1-7 中的 VPN A 实例为例，组播 VPN 是指：

- 组播源属于 VPN A，向某组播组 G 发送组播报数据。
- 网络中所有可能的数据接收者中，只有属于 VPN A 的组成员才能接收到 S 发送的组播数据。

- 组播数据在 VPN A 中以组播方式传输，在公网中也以组播方式传输。

1.5 术语与缩略语

术语

术语	解释
IGMP	Internet Group Management Protocol，称为因特网组管理协议，是 IP 组播在末端网络上使用的主机对路由器的信令机制。 IGMP 在主机侧实现组播组成员动态加入与离开，在路由器侧实现组成员关系的维护与管理，同时支持与上层组播路由协议的信息交互。
MLD	Multicast Listener Discovery，称为组播监听者发现协议，用于 IPv6 网络。 在 IPv6 网络中，通过在接收者主机和与其直连的组播路由器上配置 MLD，可以实现主机动态加入和组播路由器对本地网络组成员信息的管理。
PIM	Protocol Independent Multicast，称为协议无关组播，属于组播路由协议。 网络中单播路由畅通是 PIM 转发的基础。PIM 利用现有的单播路由信息，对组播报文执行 RPF 检查，从而创建组播路由表项，构建组播分发树。
MSDP	Multicast Source Discovery Protocol，称为组播源发现协议。只适用于 PIM-SM 域，仅对 ASM（Any-Source Multicast）模型有意义。 通过在不同 PIM-SM 域的 RP 之间建立 MSDP 对等体关系，在域间共享组播源信息，实现跨域组播。 通过在同一 PIM-SM 域的多个 RP 之间建立 MSDP 对等体关系，在域内共享组播源信息，实现 Anycast RP。

缩略语

缩略语	英文全称	中文全称
ASM	Any-Source Multicast	任意源组播
SFM	Source-Filtered Multicast	过滤源组播
SSM	Source-Specific Multicast	指定源组播

2 PIM

关于本章

2.1 介绍

定义

PIM（Protocol Independent Multicast）称为协议无关组播，作为一种组播路由解决方案，主要用于将网络中的组播数据流引入到有组播数据请求的组成员，从而实现组播数据流的转发。

目前在实际网络中应用较为广泛的实现方式主要有以下三种，均可应用于 IPv4 和 IPv6 网络。

- PIM-DM（Protocol Independent Multicast Dense Mode）：协议无关组播－密集模式。
- PIM-SM（Protocol Independent Multicast Sparse Mode）：协议无关组播－稀疏模式。
- PIM-SSM（Protocol Independent Multicast Source-Specific Multicast）：协议无关组播－指定源组播。

目的

组播源向组播组地址发出组播报文，经过中间网络到达组播组所有成员。为使中间网络能够实现组播报文的复制和转发，必须为网络中的路由器配置组播路由协议。PIM 协议是网络中实现组播报文的复制和转发的一个重要协议。

2.2 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
RFC4601	Protocol Independent Multicast - Sparse Mode （PIM-SM）	-
RFC5059	Bootstrap Router （BSR） Mechanism for PIM	-
RFC3973	Protocol Independent Multicast - Dense Mode protocol	-
RFC4607	Source-Specific Multicast for IP	-
RFC4610	Anycast Rendezvous Point （Anycast RP）	

2.3 原理描述

2.3.1 基本概念

PIM 路由器

支持 PIM 协议的组播路由器称为 PIM 路由器。使能了 PIM 协议的接口称为 PIM 接口。

PIM 域

由 PIM 路由器所组成的网络称为 PIM 网络。

通过在组播设备接口上设置“边界”，可以将一个大的 PIM 网络划分多个 PIM 域。“边界”可以拒绝特定组播报文通过，或者限制 PIM 控制消息的传输。

组播分发树

在 PIM 组播域中，以组播组为单位建立一点到多点的组播转发路径。由于组播转发路径呈现树型结构，也称为组播分发树（MDT，Multicast Distribution Tree）。

- 以组播源为根，组播组成员为叶子的组播分发树称为 SPT（Shortest Path Tree）。SPT 同时适用于 PIM-DM 和 PIM-SM。
- 以 RP（Rendezvous Point）为根，组播组成员为叶子的组播分发树称为 RPT（RP Tree）。RPT 仅适用于 PIM-SM。

组播分发树的特点：

- 无论网络中的组成员有多少，每条链路上相同的组播数据最多只有一份。
- 被传递的组播数据在距离组播源尽可能远的分叉路口才开始复制和分发。

叶子路由器

与用户主机相连的 PIM 路由器称为叶子路由器。

组播源 DR

与组播源直接相连且负责向 RP 发送注册报文的 PIM 路由器。

接收者 DR

与组播组成员（通常为接收者主机）直接相连且负责向该组成员转发组播数据的 PIM 路由器。

中间路由器

组播转发路径上，第一跳路由器与最后一跳路由器之间的 PIM 路由器。

2.3.2 PIM-SM

PIM-SM（Protocol Independent Multicast-Sparse Mode）称为协议无关组播-稀疏模式，主要采用接收者主动加入的方式建立组播转发树，适用于网络中的组成员相对比较稀疏，分布广泛的大型网络。

基本原理

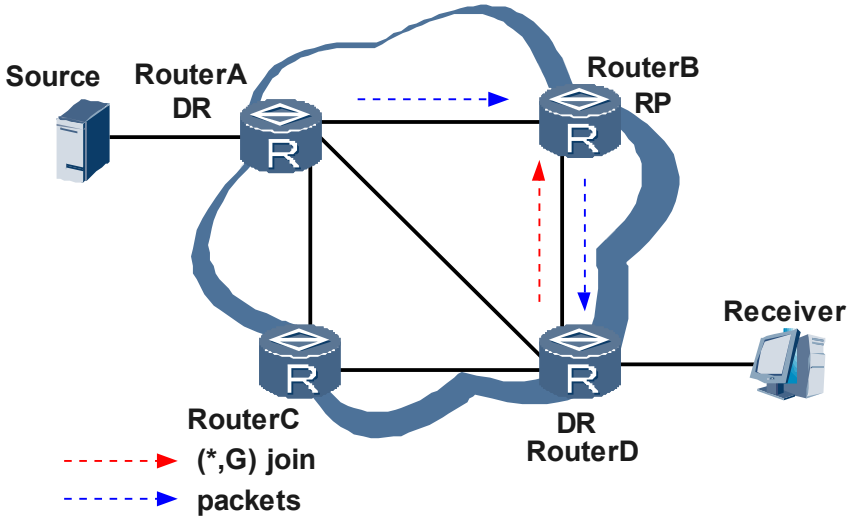
PIM-SM 转发组播数据的关键是建立 RPT（Rendezvous Point Tree，汇聚点树也称共享树）和 SPT（Shortest Path Tree，最短路径树）。

RPT 建立原理

PIM-SM RPT 是一棵以 RP 为根，以存在组成员关系的路由器为叶子的组播分发树。

汇聚点 RP 为网络中一台重要的 PIM 路由器，用于处理组播源 DR 注册信息及组成员加入请求，网络中的所有 PIM 路由器都知道 RP 的位置，RP 类似于一个供求信息汇聚中心。

图 2-1 RPT 建立原理图



建立 RPT 的过程即建立组播数据转发路径的过程。如图 2-1 所示，RPT 的建立及数据转发过程如下：

- 当网络中出现活跃的组播源（组播源向某组播组 G 发送第一个组播数据）时，组播源端 DR 将组播数据封装在 Register 消息中单播发往 RP，在 RP 上创建（S，G）表项，注册源信息。
- 当网络中出现组成员（用户主机通过 IGMP 加入某组播组 G）时，接收者 DR 向 RP 发送 Join 消息，在通向 RP 的路径上逐跳创建（*，G）表项，生成以一棵以 RP 为根的 RPT。
- 当网络中同时出现组成员和向该组发送数据的组播源时，以 RP 为中转站，组播数据先被封装在 Register 消息中单播发往 RP，再沿 RPT 到达组成员。

RPT 实现了组播数据按需转发的目的，减少无需求数据对网络带宽的占用。

DR（Designated Router）的分类：

- 在连接组播源的共享网段，由 DR 负责向 RP 发送 Register 注册消息。与组播源相连的 DR 称为组播源端 DR。
- 在连接组成员的共享网段，由 DR 负责向 RP 发送 Join 加入消息。与组成员相连的 DR 称为接收者 DR。

说明

为了减轻 RPT 的转发负担、提高组播数据转发效率，PIM-SM 允许进行 SPT 切换。即建立一条从数据源直接到接收者的转发链路，组播源可以沿 SPT 将组播数据转发到接收者。

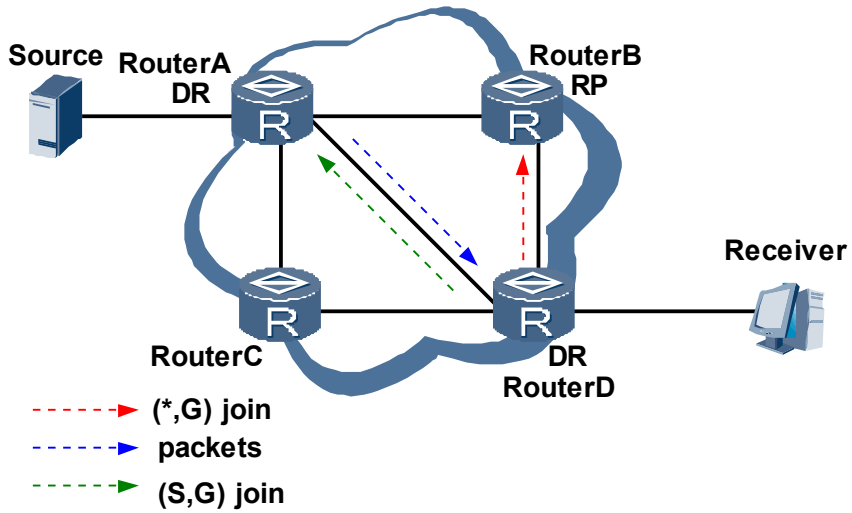
SPT 切换原理

PIM-SM SPT 是以组播源为根，以组播组成员为叶子的组播分发树。

在 PIM-SM 网络中，一个组播组只对应一个 RP，只构建一棵 RPT。在未进行 SPT 切换的情况下，所有发往该组的组播报文都必须先封装在注册消息中发往 RP，RP 解封装后，再沿 RPT 分发。

由于所有通过 RPT 转发的组播数据报文必须经过 RP 中转，所以当组播数据报文逐渐增多时，会对 RP 形成巨大的负担。为了解决此问题，PIM-SM 允许在组播报文速率增大到某个阈值时，由 RP 或接收者 DR 触发 SPT 切换，建立一条从数据源直接到接收者的转发链路。以下组播源简称为 S。

图 2-2 接收者 DR 进行 SPT 切换的原理图



SPT 切换的两种方式:

- RP 触发 SPT 切换

  RP 收到源端 DR 的注册报文后，将封装在注册报文中的组播数据沿 RPT 转发给组成员，同时 RP 会向源端 DR 发送 SPT 加入报文，建立 RP 到源的 SPT 树。

  SPT 树建立成功后，RP 停止使用注册消息，使源端 DR 和 RP 免除了频繁的封装/解封装。组播数据从与组播源直接相连的路由器，通过 SPT 树转发到 RP，再沿 RPT 转发给组成员。
- 组成员端 DR 触发 SPT 切换。

如图 2-2 所示：组成员端 DR 周期性检测组播报文的转发速率。一旦发现（S，G）报文的转发速率超过阈值，则触发 SPT 切换。

建立从源到组成员的 SPT 后，后续报文可能不再流经 RP。由于 RPT 不一定是路径最短的树，进行 SPT 切换后，减少了组播数据在网络中的传输延迟。

网络中可能存在一个组播源向多个组播组发送组播报文。若指定了针对某个组播组范围的 SPT 切换策略，SPT 切换前，这些报文都沿 RPT 到达组成员端 DR；完成 SPT 切换后，只有组播源发往属于 SPT 切换策略中组播组范围内的组播组的报文沿 SPT 转发，而组播源发往其他组播组的报文，仍会沿 RPT 转发。

说明

缺省情况下，RP 在收到第一个组播注册报文后立即进行 SPT 切换，组成员端 DR 收到第一个组播数据报文后立即进行 SPT 切换。

邻居发现

PIM 路由器在每个使能了 PIM 的接口上，对外发送 Hello 消息。封装 Hello 消息的组播报文的目的地址是 224.0.0.13（表示同一网段中所有 PIM 路由器）、源地址为接口的 IP 地址、TTL 数值为 1。

Hello 消息的作用：发现邻居、协调各项协议参数、维持邻居关系。

- 发现 PIM 邻居

  同一网段中的 PIM 路由器都必须接收目的地址为 224.0.0.13 的组播报文。这样在收到 Hello 报文以后，直接相连的组播路由器之间，就可以彼此知道自己的邻居信息。

  只有在路由器接收到来自邻居的 Hello 消息后，才会接收其他的 PIM 控制消息或组播报文，从而创建组播路由表项，维护组播分发树。

- 协调各项协议参数

  Hello 消息中携带多项协议参数，介绍如下：

  - DR_Priority：表示各路由器接口竞选 DR 的优先级，优先级越高越容易获胜。适用于 PIM-SM。
  - Holdtime：表示保持邻居为可达状态的超时时间。
  - LAN_Delay：表示共享网段内传输 Prune 消息的延迟时间。
  - Neighbor-Tracking：表示邻居跟踪功能。
  - Override-Interval：表示 Hello 消息中携带的否决剪枝的时间间隔。

- 维持邻居关系

  PIM 路由器之间周期性地发送 Hello 消息。如果 Holdtime 超时还没有收到该 PIM 邻居发出的新的 Hello 报文，则认为该邻居不可达，将其从邻居列表中清除。

  PIM 邻居的变化将导致网络中组播拓扑的变化。如果组播分发树上的某上游邻居或下游邻居不可达，将导致组播路由重新收敛，组播分发树迁移。

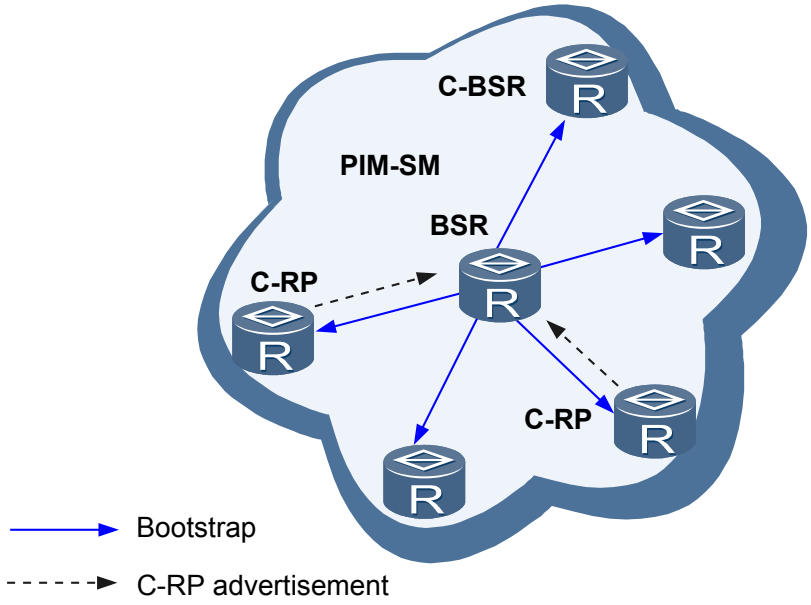
RP 发现机制

- RP 分类

  一个 RP 可以同时为多个组播组服务，但一个组播组只能对应一个 RP。RP 是 PIM-SM 网络的核心，网络中的路由器必须知道 RP 的地址，目前可以通过以下方式获取 RP 的地址：

  - 静态 RP：用户通过配置命令在网络中的所有路由器上配置相同的 RP 地址；
  - 动态 RP：在 PIM 域内选择几台 PIM 路由器，配置 C-RP（Candidate-RP）。从 C-RP 中竞选产生 RP。

图 2-3 动态 RP 竞选机制原理图



如图 2-3 所示：

1. 使用动态 RP，必须同时配置 C-BSR（Candidate-BootStrap Router），由 C-BSR 竞选产生 BSR。网络中的所有路由器都知道 BSR 的地址。
2. C-RP 向 BSR 发送 Advertisement 消息，消息中携带 C-RP 地址、服务的组范围和 C-RP 优先级。
3. BSR 将这些信息汇总为 RP-Set，封装在 Bootstrap 消息中，发布给全网的每一台 PIM-SM 路由器。
4. 各路由器根据 RP-Set，使用相同的规则进行计算和比较，从多个针对特定组的 C-RP 中竞选出该组 RP。规则如下：
   - C-RP 接口地址掩码最长者获胜。
   - C-RP 优先级较高者获胜（优先级数值越大优先级越低）。
   - 如果优先级相同，则执行 Hash 函数，计算结果较大者获胜。
   - 如果以上都相同，则 C-RP 地址较大者获胜。
5. 由于所有路由器使用相同的 RP-Set 和竞选规则，所以得到的“组播组－RP”对应关系也相同。路由器将“组播组—RP”对应关系保存下来，指导后续的组播操作。

说明

为了和使用 Auto-RP 的厂商进行互通，支持作为其 Auto-RP 的 Listening 端，并支持 IPv6 的 Embedded RP。

- Anycast RP

  在传统的 PIM-SM 域中，每个组播组只能映射到一个 RP。但当网络负载较大或者流量过于集中时，可能导致 RP 的压力过大、RP 失效后路由收敛较慢、组播转发路径非最优等问题。

  针对上述问题实现 Anycast RP，目前有两种配置方式：

- 采用 MSDP 协议：在同一 PIM-SM 域内设置多个具有相同地址的 RP，并在这些 RP 之间通过建立 MSDP（Multicast Source Discovery Protocol）对等体的方式共享组播数据源信息。

  采用 MSDP 协议实现 Anycast RP 支持 IPv4 网络。
- 采用 PIM 协议：在同一个 PIM-SM 域内设置多个具有相同地址的 RP，同时在这些 RP 所在的设备上配置全网唯一标识该 RP 的本地地址，用于这些设备之间相互建立无连接的对等体，对等体之间以注册报文的方式共享组播源信息。

  采用 PIM 协议实现 Anycast RP 支持 IPv4 和 IPv6 网络。

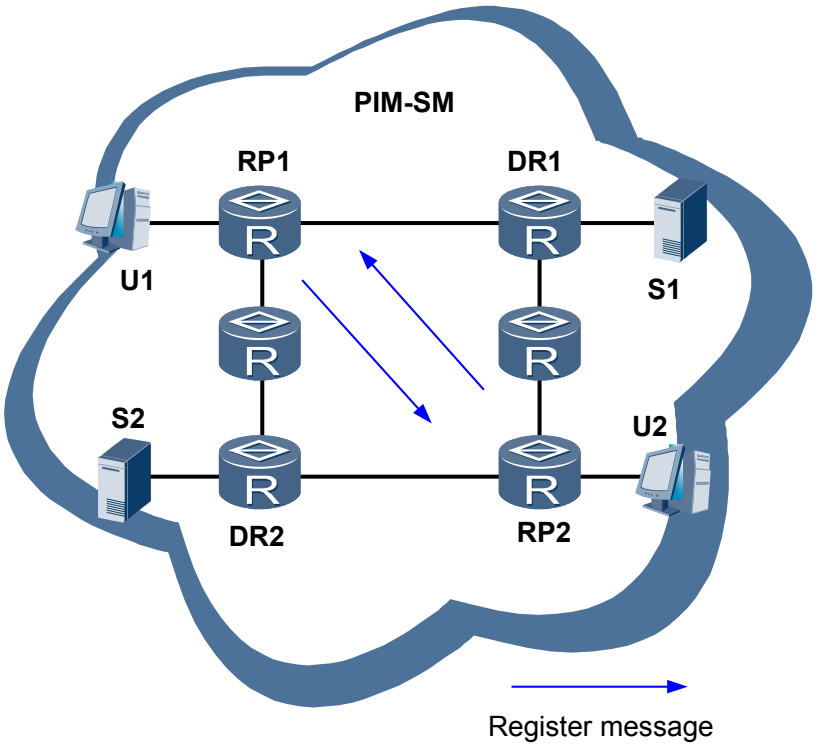
说明

在同一 PIM-SM 域内，不支持同时使用 MSDP 和 PIM 协议两种方式配置同一个 Anycast RP；采用 PIM 协议实现 Anycast RP 支持通过命令行设置有选择地将从 MSDP 获得的域外数据源信息，通知给域内的其它对等体（可配置）。

通过以上两种方式之一实现 Anycast RP，接收者和组播源分别选择距离自己最近的 RP 进行 RPT 的创建，当接收者 DR 接收到组播数据后自行决定是否发起 SPT 切换。从而实现 RP 路径最优及负荷分担。

对于采用 MSDP 实现 Anycast RP 的实现原理，请参见 **5.3.2 MSDP 实现 Anycast RP**，下面重点介绍采用 PIM 协议实现 Anycast RP 的原理。

图 2-4 采用 PIM 协议实现 Anycast RP 典型组网图



如图 2-4 所示，在 PIM-SM 域内，组播源 S1 和 S2 向组播组 G 发送组播数据，U1 和 U2 是组播组 G 的成员。在 PIM-SM 域内应用 PIM 协议实现 Anycast RP 的配置方法如下：

- 配置 RP1 和 RP2，使用相同的 IP 地址（使用 Loopback 接口，假设为 10.10.10.10）。

- 在 RP1 和 RP2 两个设备间配置无连接的对等体关系（使用网络中唯一标识的 IP 地址，假设 RP1 的 IP 地址为 1.1.1.1，RP2 的 IP 地址为 2.2.2.2）。

采用 PIM 协议实现 Anycast RP 的实现过程如下：

1. 接收者选择距离最近的 RP 发送加入消息构建 RPT 树。
   - U1 加入以 RP1 为根的 RPT，在 RP1 上创建（*，G）。
   - U2 加入以 RP2 为根的 RPT，在 RP2 上创建（*，G）。
2. 组播源选择距离最近的 RP 进行注册。
   - DR1 向 RP1 发送注册消息，在 RP1 上创建（S1，G）。从 S1 发来的组播数据沿 RPT 到达 U1。
   - DR2 向 RP2 发送注册消息，在 RP2 上创建（S2，G）。从 S2 发来的组播数据沿 RPT 到达 U2。
3. RP 收到源 DR 发过来的注册报文，重新封装成注册报文转给自己的对等体，共享组播源信息。
   - RP1 收到源 DR1 发过来的（S1，G）注册报文后，将该注册报文的源地址和目的地址替换为 1.1.1.1 和 2.2.2.2，重新封装后发送到 RP2。RP2 收到该注册报文后检查发现该注册报文来自对等体 1.1.1.1，不再转发给其它对等体，只是处理该注册报文。
   - RP2 收到源 DR2 发过来的（S2，G）注册报文后，将该注册报文的源地址和目的地址替换为 2.2.2.2 和 1.1.1.1，重新封装后发送到 RP1。RP1 收到该注册报文后检查发现该注册报文来自对等体 2.2.2.2，不再转发给其它对等体，只是处理该注册报文。
4. RP 加入以源端 DR 为根的 SPT，将组播数据引下来。
   - RP1 向 S2 发送加入消息。从 S2 发来的组播数据先沿 SPT 到达 RP1，再沿 RPT 到达 U1。
   - RP2 向 S1 发送加入消息。从 S1 发来的组播数据先沿 SPT 到达 RP2，再沿 RPT 到达 U2。
5. 接收者 DR 接收到组播数据后，自行决定是否发起 SPT 切换。

BootStrap Router 机制

- BSR 竞选机制

  BSR 称为自举路由器（Bootstrap Router），负责收集并发布网络中的 C-RP 信息，确保网络中的所有路由器都知道 RP 的位置。

  BSR 是从众多 C-BSR 中竞争产生的。最初，每个 C-BSR 都认为自己是 BSR，向全网发送 Bootstrap 消息。Bootstrap 消息中携带 C-BSR 地址、C-BSR 的优先级。每一台路由器都收到所有 C-BSR 发出的 Bootstrap 消息，通过比较这些 C-BSR 信息，竞选产生 BSR。竞选规则如下：

  - 优先级较高者获胜（优先级数值越大优先级越高）。
  - 如果优先级相同，IP 地址较大者获胜。

  由于所有路由器使用相同的竞选规则，所以得到的当选 BSR 也相同。

  设备发送 BSR 报文时需要携带网络中所有的 C-RP 信息。当网络中存在大量 C-RP，BSR 报文携带这些 C-RP 信息时，会导致报文长度过大，超过接口 MTU 值，造成设备无法正确处理 BSR 报文，从而无法选举出 RP 信息，组播业务无法正常传输。此时可以使用 BSR 报文分片功能对 BSR 报文进行分片处理。

推荐使用 BSR 报文分片功能，可以解决 IP 分片时，分片信息丢失而导致所有分片不可用的问题。

- BSR 管理域

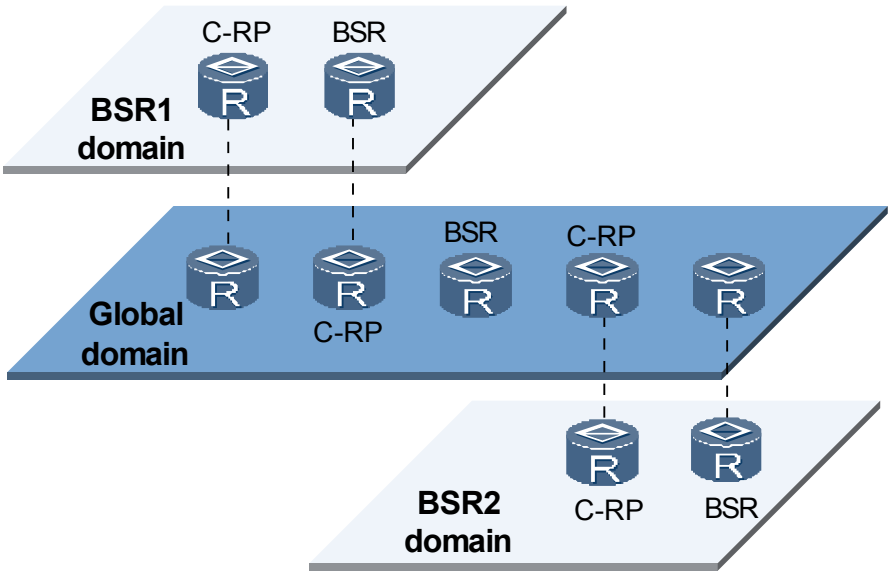
  为了实现网络管理精细化，可以选择将一个 PIM-SM 网络划分为多个 BSR 管理域和一个 Global 域。这样一方面可以有效地分担单一 BSR 的管理压力，另一方面可以使用私有组地址为特定区域的用户提供专门服务。

  每个 BSR 管理域中维护一个 BSR，为某一特定地址范围的组播组服务。Global 域中维护一个 BSR，为所有剩余的组播组服务。

  下文将从地域空间、组地址范围、组播功能三个角度分析 BSR 管理域和 Global 域的关系。

  - 地域空间

    图 2-5 BSR 管理域_地域空间

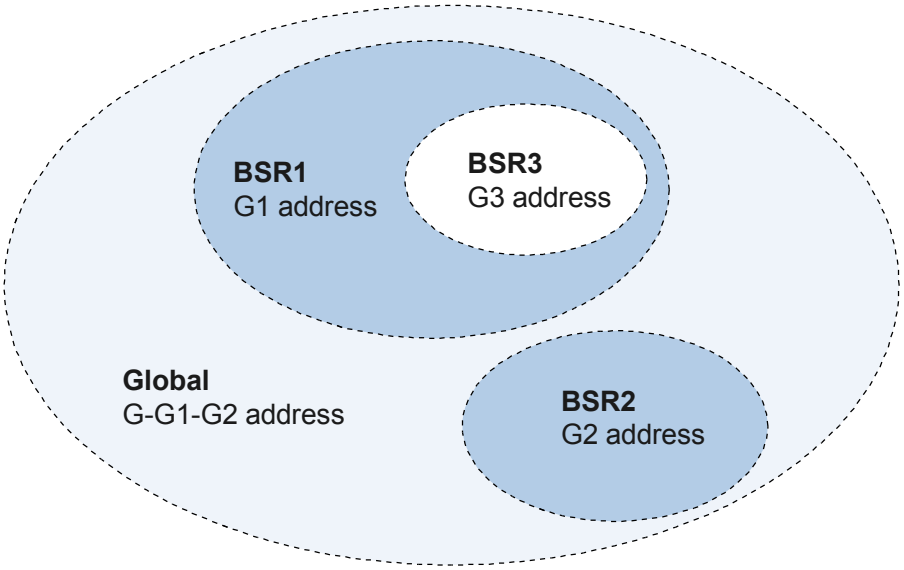
    

    如图 2-5 所示，对于有相同组地址的不同管理域，各 BSR 管理域所包含的路由器互不相同，同一路由器不能从属于多个 BSR 管理域。各 BSR 管理域在地域上相互独立，且相互隔离。BSR 管理域是针对特定地址范围的组播组的管理区域，属于此范围的组播报文只能在本管理域内传播，无法通过 BSR 管理域边界。

    Global 域包含 PIM-SM 网络内的全部路由器。不属于任意 BSR 管理域的组播报文，可以在整个 PIM 网络范围内传播。

  - 组地址范围

图 2-6 BSR 管理域_地址范围



每个 BSR 管理域为特定地址范围的组播组提供服务，不同的 BSR 管理域服务的组播组范围可以重叠。该组播地址只在本 BSR 管理域内有效，相当于私有组地址。如图 2-6 所示，BSR1 域和 BSR3 域对应的组地址范围出现重叠。

不属于任何 BSR 管理域的组播组，一律属于 Global 域的服务范围。即 Global 域组地址范围是 G-G1-G2。

- 组播功能

  如图 2-5 所示，Global 域和每个 BSR 管理域都包含针对自己域的 C-RP 和 BSR 设备，这些设备在行使相应功能时，仅在本域内有效。即 BSR 机制和 RP 竞选在各管理域之间是隔离的。

  每个 BSR 管理域都有自己的边界，该管理域的组播信息（C-RP 宣告消息、BSR 自举消息等）不能跨越域传播。同时 Global 域的组播信息可以在整个 Global 域内传递，可以穿越任意 BSR 管理域。

Assert 的基本原理

当满足如下条件时，说明网段上还存在着其他的组播转发者。路由器执行 Assert：

- 该组播报文不能通过 RPF 检查。
- 接收到组播报文的接口是本路由器上（S，G）表项中的一个下游接口。

路由器从该下游接口发送 Assert 消息。同时，该下游接口也接收到了来自该网段上其他组播转发者的 Assert 消息。PIM Assert 消息的目的地址为 224.0.0.13，源地址为下游接口地址，TTL 为 1。Assert 消息中携带：该 PIM 路由器到组播源或 RP 的开销、所采用的单播路由协议的优先级、组播组地址 G。

路由器将自身条件与对方报文中携带的信息进行比较，称为 Assert 竞选。规则如下：

- 单播路由协议优先级较高者获胜。
- 如果优先级相同，则到组播源的开销较小者获胜。
- 如果以上都相同，则下游接口 IP 地址最大者获胜。

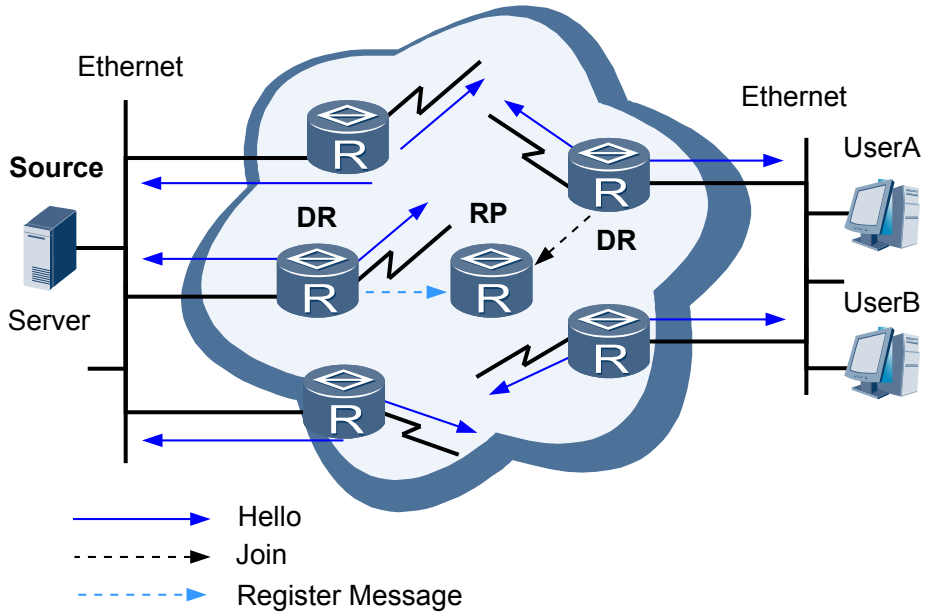
根据 Assert 竞选结果，路由器将执行不同的操作：

- 如果获胜，则该下游接口保持转发状态，路由器负责后续在该网段上的（S，G）转发，该下游接口称为 Assert winner。
- 如果落败，则禁止该下游接口转发组播报文，将其从（S，G）表项下游接口列表中删除。该下游接口称为 Assert loser。

Assert 竞选结束后，该网段上只存在一个有下游接口的上游路由器，只传输一份组播报文。Assert winner 周期性发送 Assert 消息，维持 Assert loser 的状态。若定时器超时后，Assert loser 仍没有收到 Assert winner 的 Assert 消息，则重新添加下游接口转发组播数据。

PIM DR 竞选的基本原理

图 2-7 DR 竞选示意图



如图 2-7 所示，DR（Designated Router）应用在 PIM-SM 网络中的如下两个位置：

- 在连接组播源的共享网段，由 DR 负责向 RP 发送 Register 注册消息。与组播源相连的 DR 称为源端 DR。
- 在连接组成员的共享网段，由 DR 负责向 RP 发送 Join 加入消息。与组成员相连的 DR 称为接收者 DR。

在组播源或组成员所在的网段，通常同时连接着多台 PIM 路由器。这些 PIM 路由器之间通过交互 Hello 消息成为 PIM 邻居，Hello 消息中携带 DR 优先级和该网段接口地址。路由器将自身条件与对方报文中携带的信息进行比较，称为 DR 竞选。规则如下：

- DR 优先级较高者获胜（网段中所有路由器都支持 DR 优先级）。
- 如果 DR 优先级相同或该网段存在至少一台路由器不支持在 Hello 报文中携带 DR 优先级，则 IP 地址较大者获胜。

如果当前 DR 出现故障，将导致 PIM 邻居关系超时，其他 PIM 邻居之间会触发新一轮的 DR 竞选过程。

PIM DR 切换延迟的基本原理

多台路由器处于同一共享网段时，通过一定的 DR 选举策略（通常在不配置 DR 优先级的情况下，共享网段内 IP 地址高的路由器将被选举为 DR）使其中一台路由器被选举为接收者 DR，在接收端负责共享网段内的组播数据转发。

缺省情况下，接口由 DR 变为非 DR 时，路由器会立即停止使用此接口转发数据。如果此时新 DR 的组播数据还未到达，那么将会出现短暂的组播数据断流。

在一个接口配置 PIM DR 切换延迟后，当某个使能 PIM-SM 的接口由于收到一个新邻居的 Hello 报文导致该接口由 DR 变为非 DR 时，该接口在延迟时间超时前仍然具有部分 DR 功能，并继续转发组播数据。

如果在 DR 延迟期间收到新 DR 转发过来的数据，处于 DR 延迟的路由器将会立刻停止转发数据，从而保证不会出现重复数据流的现象。此时在共享网段上若收到新的 IGMP 加入，处于 DR 切换延迟状态的旧 DR 将不会为其向上游发送 PIM 加入报文，而由新 DR 处理。

说明

在 DR 切换延迟期间，如果新 DR 收到原 DR 发送的组播数据，则会触发 Assert 竞选。

BFD for PIM 的基本原理

为了减小设备故障对业务的影响，提高网络的可靠性，网络设备需要快速检测到与相邻设备间的通信故障，以便及时采取措施，保证业务继续进行。

现有的故障检测方法主要包括：

- 硬件检测：例如通过 SDH（Synchronous Digital Hierarchy，同步数字体系）告警检测链路故障。硬件检测的优点是可以很快发现链路故障，但此检测方法不适用于所有介质。
- 慢 Hello 机制：通常是指路由协议的 Hello 机制。这种机制检测到故障所需时间为秒级。对于高速数据传输，例如吉比特速率级，超过 1 秒的检测时间将导致大量数据丢失；对于时延敏感的业务，例如语音业务，超过 1 秒的延迟也是不能接受的。
- 其他检测机制：不同的协议或设备制造商有时会提供专用的检测机制，但在系统间互联互通时，这样的专用检测机制通常难以部署。

BFD（Bidirectional Forwarding Detection）检测机制可提供毫秒级的快速检测，并采用单一机制对所有类型的介质、协议层进行检测，实现全网统一的检测机制。其检测原理是在两个系统间建立 BFD 会话，并沿它们之间的路径周期性发送 BFD 检测报文，如果一方在检测周期内没有收到 BFD 检测报文，则认为该路径发生了故障。

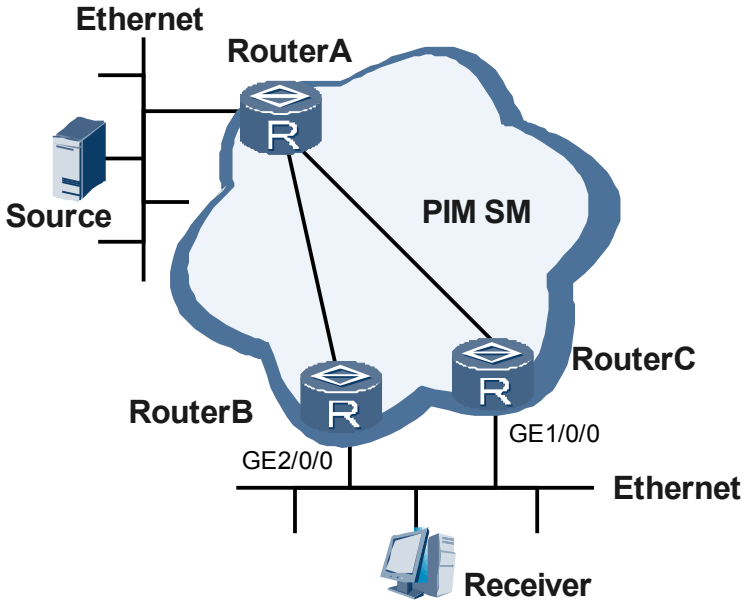
在组播的应用中，如果共享网段上的当前 DR 或 Assert winner 发生故障，其他 PIM 邻居会等到邻居关系超时或 Assert timer 超时才触发新一轮的 DR 竞选或 Assert 竞选过程，导致组播数据传输中断，中断的时间将不小于邻居关系的超时时间或 Assert timer 超时时间，通常是秒级。

BFD for PIM 能够在毫秒级内检测共享网段内的链路状态，快速响应 PIM 邻居故障。如果配置了 BFD for PIM 功能的接口在检测周期内没有收到当前 DR 或 Assert winner 发送的 BFD 检测报文，则认为当前 DR 或 Assert winner 发生故障，BFD 快速把会话状态通告给 RM，再由 RM 通告给 PIM。PIM 模块触发新一轮的 DR 竞选或 Assert 竞选过程，而不是等到邻居关系超时或 Assert timer 超时，从而缩小组播数据传输的中断时间，提高组播数据传输的可靠性。

说明

目前，BFD for PIM 功能支持 IPv4 和 IPv6 PIM SM/SSM 网络。

图 2-8 BFD for PIM 原理图



如图 2-8 所示，在与用户主机相连的共享网段上，RouterB 的下游接口 GE2/0/0 和 RouterC 的下游接口 GE1/0/0 之间建立 PIM BFD session，通过在链路两端发送 BFD 检测报文检测链路状态。

RouterB 的下游接口 GE2/0/0 作为当前 DR，负责接收端组播数据的转发。若接口 GE2/0/0 发生故障，BFD 快速把会话状态通告给 RM，再由 RM 通告给 PIM。PIM 模块触发新一轮的 DR 竞选,RouterC 的下游接口 GE1/0/0 作为新当选的 DR,在短时间内向接收端转发组播数据，从而缩小组播数据传输的中断时间。

PIM Silent 的基本原理

若路由器直连用户主机的接口上使能了 PIM 协议，就可以在该接口上建立 PIM 邻居，处理各类 PIM 协议报文。但此配置同时存在着安全隐患：当恶意主机模拟发送 PIM Hello 报文时，有可能导致路由器瘫痪。

为了避免上述情况，可以在路由器直连用户主机的接口上配置 PIM Silent，用来禁止该接口接收和转发任何 PIM 协议报文。同时，此接口上的 IGMP 功能不受影响。

2.3.3 PIM-SSM

PIM 支持 ASM（Any-Source Multicast）模型和 SSM（Source-Specific Multicast）两种模型，本节介绍 SSM 模型。

SSM 模型是借助 PIM-SM 的部分技术和 IGMPv3/MLDv2 来实现的，其建立组播转发树的过程与 PIM-SM 创建 SPT 树的过程相似，即接收者 DR 在知道组播数据源的具体位置后，直接向组播数据源发送 Join 消息，将组播数据流发送到接收者。

缺省情况下，SSM 组播组地址的范围为 232.0.0.0 ～ 232.255.255.255。当用户加入的组播组属于 SSM 组地址范围内，适用于 SSM 模型；当用户加入的组播组不属于 SSM 组地址范围，则适用 ASM 模型，ASM 模型原理即 PIM-SM 原理。

SSM 的特点是网络用户能够预先知道组播源的具体位置。因此用户在加入组播组时，可以明确指定从哪些源接收信息。组成员端 DR 了解到用户的需求后，直接向组播源的方向发送 Join消息。Join 消息逐跳向上传输，在源与组成员之间建立 SPT。

SSM 只使用了 PIM-SM 的部分技术：无需维护 RP、无需构建 RPT、无需注册组播源，可以直接在源与组成员之间建立 SPT。

说明

可以通过静态组加入和 SSM-Mapping 建立组播分发树，构建 SPT。

在 SSM 中，DR 仅在与组成员相连的共享网段上有效。由 DR 向组播源的方向发送加入消息，逐跳创建（S，G）表项，构建 SPT。

PIM-SSM 支持 PIM DR 切换延迟、PIM Silent、BFD for PIM 特性。

2.3.4 PIM-DM

应用场景

PIM-DM（Protocol Independent Multicast Dense Mode）协议无关组播－密集模式，主要采用扩散-剪枝的方式转发组播数据流，对于组播组成员稀疏的网络会产生大量剪枝报文，而对规模较大的网络扩散-剪枝周期会很长。所以适合规模较小、组播组成员相对比较密集的网络。

基本原理

PIM-DM 假设网络中的组成员分布非常稠密，每个网段都可能存在组成员。其设计思路是：首先将数据报文扩散到各个网段，然后再裁剪掉不存在组成员的网段。通过周期性的“扩散—剪枝”，构建并维护一棵连接组播源和组成员的单向无环 SPT（Source Specific Shortest Path Tree）。

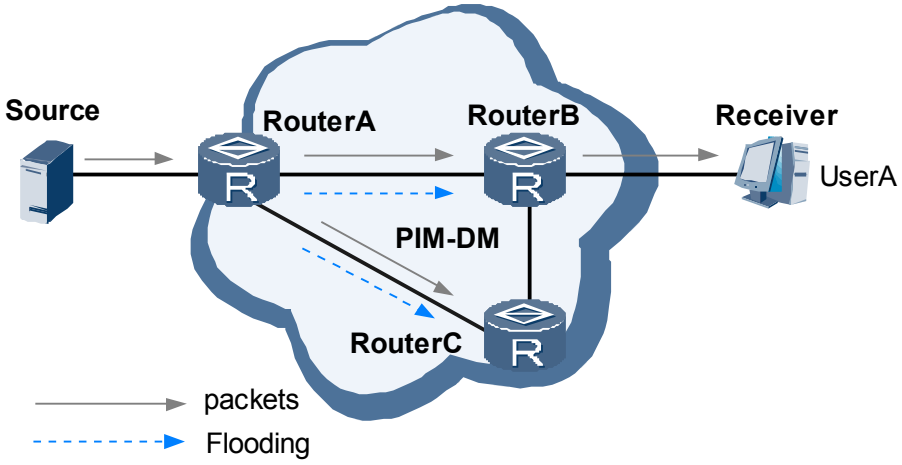
邻居发现

与 PIM SM 协议相同，请参见 **2.3.2 PIM-SM**。

扩散 Flooding

如**图 2-9** 所示，组播数据源（Source）把数据发送到 RouterA，RouterA 把数据报文发送给它所有的邻居（除了给它发送数据的邻居，如 RouterB 和 RouterC 不会把数据发给 RouterA,）。这时 RouterB 与 RouterC 也会相互转发数据报文，但 DM 协议采用 RPF 检查机制可以保证数据只从一个方向接收。最后数据被扩散到连接接收者的 RouterB，由 RouterB 把数据发送给它的接收者 UserA。

图 2-9 PIM-DM 扩散示意图

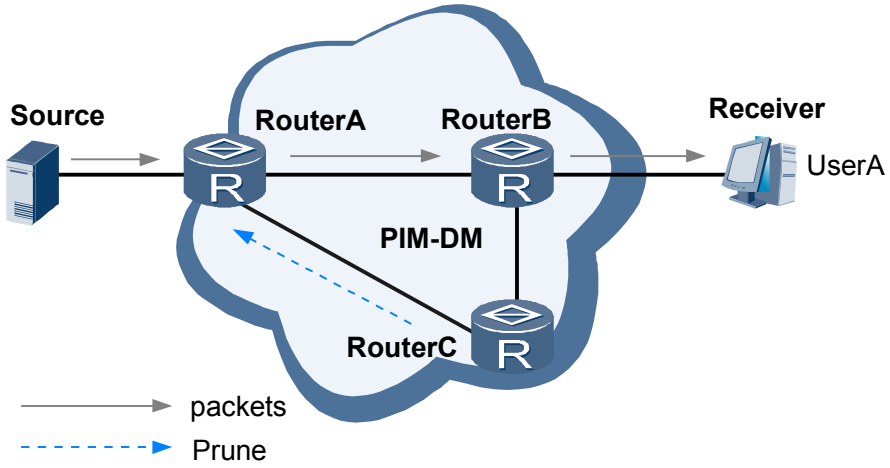


剪枝 Prune

如图 2-10 所示，由于 RouterC 没有接收者，不需要数据，则向上游 RouterA 发送 Prune 消息，通知 RouterA 不必再向该下游网段转发数据。

RouterA 收到 Prune 消息后，停止该下游接口转发，由于 RouterA 上还存在其他处于转发状态的下游接口，剪枝过程停止。后续到达的报文只向 RouterB 转发。从而实现了一棵连接组播源和组成员 UserA 的单向无环最短路径树。

图 2-10 PIM-DM 剪枝示意图

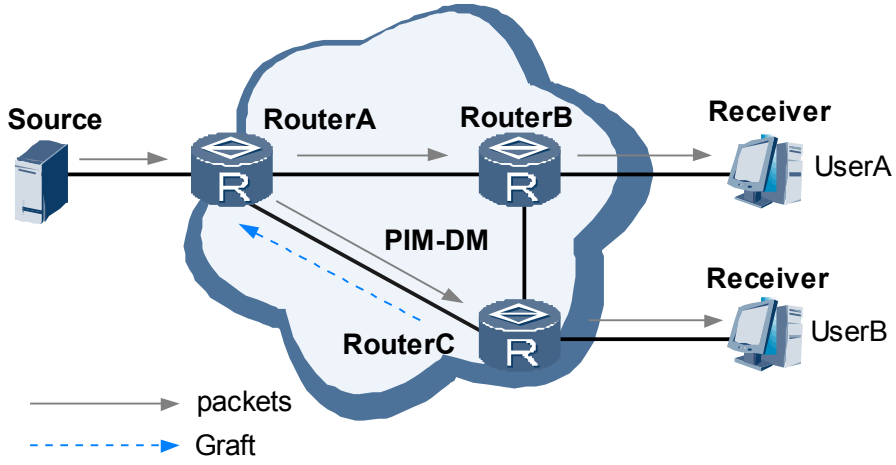


嫁接 Graft

如图 2-11 所示，如果 RouterC 收到接收者 UserB 的 IGMP Report 报文，请求转发组播源数据，即 RouterC 具有转发数据需求。为了避免周期性扩散-剪枝的时间延迟，PIM DM 用嫁接 Graft 方式实现数据的快速转发。

RouterC 发送 Graft 嫁接消息，请求上游 RouterA 恢复对应出接口的转发。RouterA 收到 Graft 消息后，将连接 RouterC 的出接口恢复转发，组播报文由该下游接口到达 RouterC。

图 2-11 PIM-DM 嫁接示意图

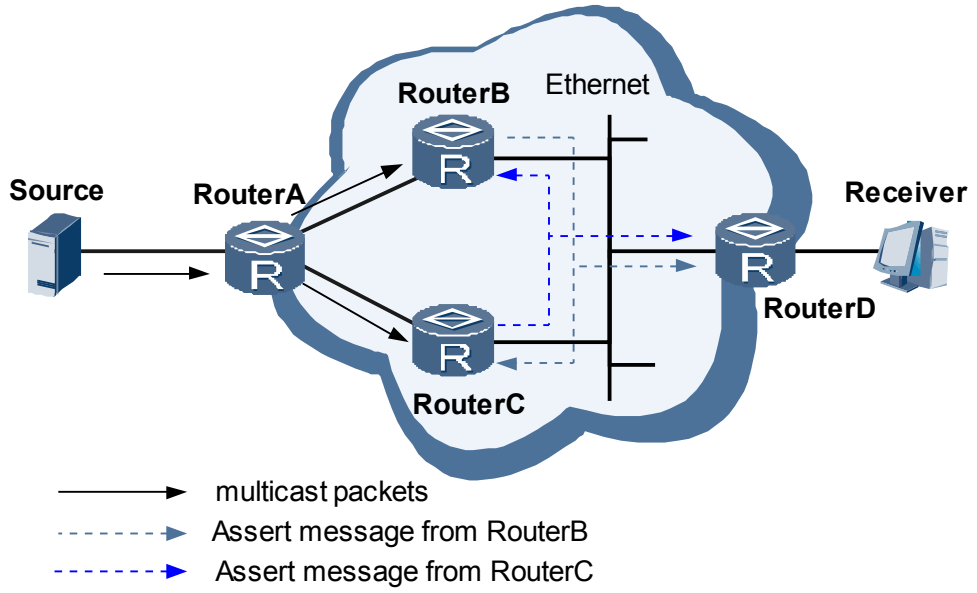


断言 Assert

如图 2-12 所示，如果 RouterB 和 RouterC 都能够接收到组播源 S 发出的组播报文，并且均能通过 RPF 检查，创建（S，G）表项。RouterB、RouterC 的下游接口连接在同一网段，那么 RouterB 和 RouterC 就会同时向该网段发送组播数据。Assert 可以保证一个网段只能存在一个组播数据转发者。断言过程如下：

1. RouterB 从下游接口接收到 RouterC 发来的组播报文，RPF 检查失败，报文被丢弃。同时，向该网段发送 Assert 消息。
2. RouterC 将自身的路由信息与对方报文中携带的路由信息进行比较，由于自身到组播源的开销较大而落败。于是禁止该下游接口转发组播报文，将其从（S，G）表项的下游接口列表中删除。
3. RouterC 从该网段接收到 RouterB 发来的组播报文，RPF 检查失败，报文被丢弃。Assert 过程结束。

图 2-12 PIM-DM 断言示意图



状态刷新

如图 2-12 所示，若 RouterA 对 RouterC 所在网段处于剪枝状态，那么 RouterA 对 RouterC 的接口会维护一个“剪枝定时器”，当剪枝定时器超时，RouterA 就会恢复对不需要数据的 RouterC 的数据转发，这样会导致不必要的网络资源浪费。

PIM DM 协议采用状态刷新特性解决此问题：离组播源最近的第一跳 RouterA 周期性触发 State Refresh 消息。State Refresh 消息在全网扩散，刷新所有设备上的剪枝定时器状态。

PIM Silent

与 PIM SM 协议相同，请参见 2.3.2 PIM-SM。

2.3.5 协议比较

表 2-1 协议比较

协议	特点
PIM-SM	协议无关组播－稀疏模式。采用接收者主动加入的方式建立组播转发树，适合网络中的组成员相对比较稀疏，分布广泛的大型网络。 需维护 RP、构建 RPT、注册组播源。
PIM-SSM	与 PIM-SM 类似，但无需维护 RP、无需构建 RPT、无需注册组播源。只要存在到数据源的路由，即可以直接在源与组成员之间建立 SPT。在跨域的组播数据流转发方面有很大的优势。

协议	特点
PIM-DM	协议无关组播－密集模式。采用扩散-剪枝的方式转发组播数据流，适合规模较小、组播组成员相对比较密集的局域网。 没有 RP 的概念，通过周期性“扩散-剪枝”维护一棵连接组播源和组成员的单向无环 SPT。

2.3.6 PIM GR

平滑重启（Graceful Restart，简称 GR），是一种控制平面主备倒换的协议。该功能可以使设备进行主备倒换时保持用户组播流量的不间断转发。目前 PIM GR 仅支持 PIM-SM 与 PIM-SSM，不支持 PIM-DM。

基本原理

组播 GR 依赖单播 GR。运行 PIM-SM/SSM 协议具有双主控的设备，在主备板倒换期间，保持接口板软件及硬件的不间断转发能力（Multicast Non-Stop Forwarding）。新主控板的 PIM 协议通过从下游邻居重新学习 PIM 加入状态及从 IGMP 主机学习加入的组成员完成如下动作：

- 重新计算 PIM 组播路由表项。
- 维持上游邻居的加入状态。
- 更新转发平面的组播路由表项。

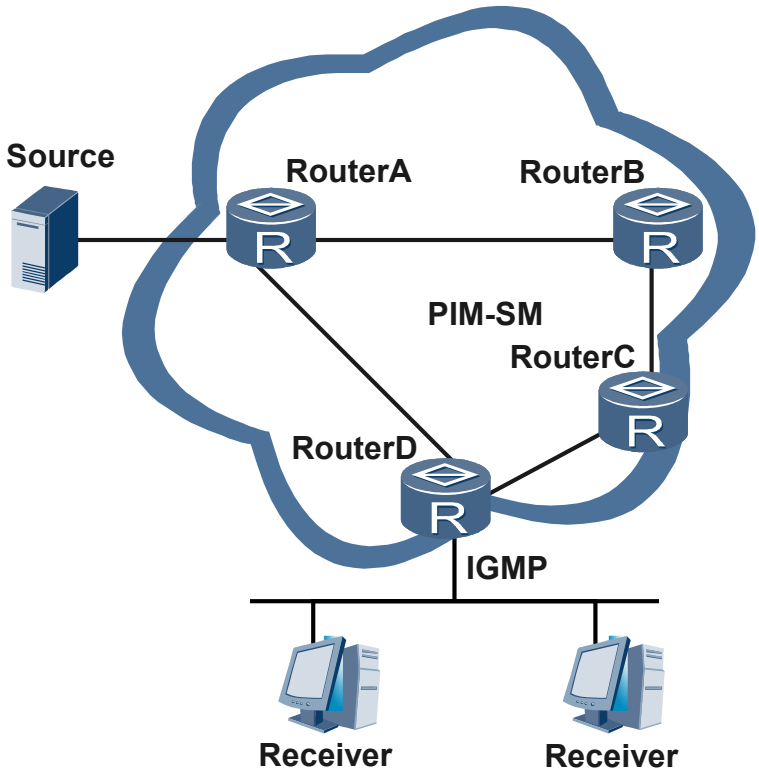
达到在倒换后快速恢复主板的 PIM 路由表项及刷新组播转发表项的目的，最大限度减少由于主备倒换对用户组播流量中断的影响。

PIM GR 适用于 PIM-SM/SSM 网络，同时也适用于 VPN 场景。在 PIM-SM 网络中的组播设备上部署 PIM GR 功能，可以使设备进行主备倒换时保持用户组播流量的不间断转发。PIM GR 还可用于 ISSU(In Service Software Upgrade)软件升级场景，使得在主控板及接口板 Image 升级过程中，保证组播流量的正常转发。

PIM GR 支持 VPN 场景。

如图 2-13 所示，介绍 RouterA 进行 PIM GR 的过程。

图 2-13 典型组网图



PIM GR 是建立在单播 GR 的基础上的，分为三个阶段，包括开始阶段（GR_START）、同步阶段（GR_SYNC）和完成阶段（GR_END）。

GR_START

1. RouterA 发生主备倒换，PIM 协议启动 GR 定时器，PIM GR 进入开始阶段。同时单播开始进行 GR。
2. PIM 协议向所有使能 PIM-SM 的接口发送携带新的 Generation ID 的 Hello 报文。
3. RouterA 的下游邻居 RouterD 发现 RPF 邻居的 Generation ID 改变，向 RouterA 重新发送 Join/Prune 报文。
4. 若网络中使用动态 RP，当网络中的邻居收到 Generation ID 改变的 Hello 报文后，向 RouterA 单播发送 BSM 报文，恢复 RouterA 的 BSR 及 RP 信息。
5. RouterA 通过接收下游 RouterD 发送的 Join/Prune 报文，在空的入接口表中创建 PIM 路由表项，记录下游的加入信息。
6. 在此期间，转发模块转发表项保持不变，维持组播业务数据的转发。

GR_SYNC

单播 GR 结束，PIM GR 进入同步阶段，根据单播路由信息建立组播分发树，恢复 PIM 路由表项的入接口，更新到源或到 RP 的加入队列，并通知组播转发模块更新转发表。

GR_END

GR 定时器超时，PIM 协议完成 GR，并通知组播转发模块。组播转发模块老化 GR 期间未更新的转发表项。

2.3.7 PIM 安全性

源地址过滤

适用于 PIM-DM 和 PIM-SM。

此功能用来实现路由器对接收的组播数据报文根据源或源组进行过滤。通过配置 ACL，路由器可以选择只转发源地址属于过滤规则范围内的组播报文，或转发源地址和组地址都属于过滤规则范围内的组播报文。

配置合法的 BSR 地址范围

适用于 PIM-SM。

此功能用来限定合法 BSR 地址范围，使路由器丢弃来自该地址范围之外的 BSR 报文，从而防止 BSR 欺骗。

配置合法的 C-RP 地址范围

适用于 PIM-SM。

此功能用来限定合法的 C-RP 地址范围及其服务的组播组地址范围，使 BSR 丢弃来自该地址范围之外的 C-RP 消息，从而防止 C-RP 欺骗。

Register 报文过滤

适用于 PIM-SM。

此功能用来实现 RP 过滤由组播源端 DR 发送的注册报文，根据报文过滤规则接受或拒绝和规则匹配的注册报文，从而防止非法注册报文攻击。

PIM 邻居过滤

适用于 PIM-DM 和 PIM-SM。

为了防止与其它未知设备建立 PIM 邻居，阻止未知设备成为 DR，需要对不期望的邻居进行过滤。配置此功能后，接口只与符合过滤规则的地址建立邻居关系，删除不符合过滤规则的邻居。

Join 信息过滤

适用于 PIM-SM。

接口上接收的 Join/Prune 消息中包含 Join 信息和 Prune 信息。此功能可过滤 Join 信息，路由器根据符合过滤规则的 Join 信息建立 PIM 表项，从而防止非法用户加入。

PIM 邻居检查

适用于 PIM-SM。

默认情况下，接收或发送 Join/Prune 消息和 Assert 消息时，不检查该消息是否来自 PIM 邻居或发送给 PIM 邻居。

如果需要配置 PIM 邻居检查功能，建议在与用户相连的设备上配置，在网络内部设备上不推荐使用此功能。接收或发送 Join/Prune 消息和 Assert 消息时，检查该消息是否来自 PIM 邻居或发送给 PIM 邻居，如果是则处理，否则丢弃。

PIM Silent

若路由器直连用户主机的接口上使能了 PIM 协议，就可以在该接口上建立 PIM 邻居，处理各类 PIM 协议报文。但此配置同时存在着安全隐患：当恶意主机模拟发送 PIM Hello 报文时，有可能导致路由器瘫痪。

为了避免上述情况，可以在路由器直连用户主机的接口上配置 PIM Silent，用来禁止该接口接收和转发任何 PIM 协议报文。同时，此接口上的 IGMP 功能不受影响。

IPv6 PIM IPSec

“Internet 协议安全性(IPSec)”是一种开放标准的框架结构，通过使用加密的安全服务以确保在 Internet 网络上进行保密而安全的通讯。IPSec 是安全联网的长期方向，通过端到端的安全性来提供主动的保护以防止专用网络与 Internet 网络之间的攻击。

IPSec 协议不是一个单独的协议，它给出了应用于 IP 层上网络数据安全的一整套体系结构，包括网络认证协议 Authentication Header（AH）、封装安全载荷协议 Encapsulating Security Payload（ESP）、密钥管理协议 Internet Key Exchange（IKE）和用于网络认证及加密的一些算法等。IPSec 提供了一套安全协议和安全算法，并定义了密钥交换机制，实现访问控制、数据源认证、数据加密等网络安全服务。

IPv6 PIM IPSec 利用 IPSec 提供的一整套安全保护机制对 IPv6 PIM 协议报文的发送和接收进行认证处理，防止伪造的 IPv6 PIM 协议报文对设备进行非法攻击。

本特性适用于 PIM-DM 和 PIM-SM。

目前，除 Register 和 Register-stop 报文外，其他 IPv6 PIM 协议报文均支持使用 IPSec 进行安全认证。

P2P 接口、Broadcast 和 NBMA 接口均支持 IPv6 PIM IPSec 特性。下面分别介绍其实现原理：

- P2P 接口支持 IPv6 PIM IPSec

  图 2-14 P2P 接口支持 IPv6 PIM IPSec 组网图

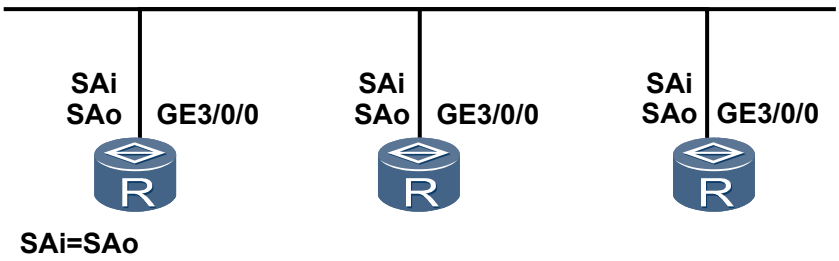
  

  如图 2-14 所示，对 P2P 接口采用 IPv6 PIM IPSec 特性实现对发送和接收的 IPv6 PIM 协议报文进行认证时，有如下要求：

  - 本端接口配置的 IPSec SAo=对端接口配置的 IPSec SAi
  - 本端接口配置的 IPSec SAi=对端接口配置的 IPSec SAo

- Broadcast 和 NBMA 接口支持 IPv6 PIM IPSec

图 2-15 Broadcast 和 NBMA 接口支持 IPv6 PIM IPSec 组网图



如图 2-15 所示，对 Broadcast 或 NBMA 接口采用 IPv6 PIM IPSec 特性实现对发送和接收的 IPv6 PIM 协议报文进行认证时，各设备上连接共享网段的各接口使用相同的 SA，即 SAi =SAo。

2.4 术语与缩略语

术语

术语	解释
（S，G）	属于组播路由表项，S 表示组播源，G 表示组播组。 源地址为 S、组地址为 G 的组播报文，到达组播设备后，从（S，G）表项中的下游接口转发出去。 通常，将源地址为 S，组地址为 G 的组播报文表示为（S，G）报文。
Assert	断言。同时在 PIM-DM 和 PIM-SM 中使用。 如果组播设备从下游接口收到组播报文，且 RPF 检查失败，则说明该网段存在其他的组播转发者。组播设备从该下游接口发出 Assert 消息，参与 Assert 竞选，如果落败则将该下游接口从下游接口列表中删除。 Assert 保证了一个网段上最多只存在一个组播转发者，传输一份组播报文。
Flooding	扩散。只在 PIM-DM 中使用。 PIM-DM 假设网络中的组成员分布非常稠密，每个网段都可能存在组成员。基于这一假设，PIM-DM 的设计思路是：首先将数据报文扩散到各个网段，然后再裁剪掉不存在组成员的网段。 PIM-DM 通过周期性的“扩散—剪枝”，构建并维护一棵连接组播源和组成员的单向无环 SPT。
Graft	嫁接。只在 PIM-DM 中使用。 原本下游接口列表为 NULL 的组播设备，添加第一个下游接口时，从上游接口发出嫁接消息。 如果上游组播设备上收到嫁接消息的接口处于剪枝状态，则立即恢复转发，添加到下游接口列表中。

术语	解释
PIM	Protocol Independent Multicast，称为协议无关组播，属于组播路由协议。 网络中单播路由畅通是 PIM 转发的基础。PIM 利用现有的单播路由信息，对组播报文执行 RPF 检查，从而创建组播路由表项，构建组播分发树。
Prune	剪枝。同时在 PIM-DM 和 PIM-SM 中使用。 当路由表项下游接口列表为 NULL 时，向上游发出剪枝消息，通知其停止向该下游接口转发组播报文。 上游组播设备将收到剪枝消息的接口从下游接口列表中删除。

缩略语

缩略语	英文全称	中文全称
RP	Rendezvous Point	汇聚点
PIM-SM	Protocol Independent Multicast Sparse Mode	协议无关组播－稀疏模式
SSM	Source-Specific Multicast	指定源组播
PIM-DM	Protocol Independent Multicast Dense Mode	协议无关组播－密集模式
PIM	Protocol Independent Multicast	协议无关组播

3 IGMP

关于本章

3.1 介绍

定义

IGMP（Internet Group Management Protocol）因特网组管理协议，是 TCP/IP 协议族中负责 IPv4 组播成员管理的协议，用来在 IP 主机和与其直接相邻的组播路由器之间建立、维护组播组成员关系。

通过在接收者主机和与其直连的组播路由器上配置 IGMP，可以实现主机动态加入组播组和组播路由器对本地网络组成员信息的管理。

到目前为止，IGMP 有三个版本：IGMPv1 版本（RFC1112）、IGMPv2 版本（RFC2236）和 IGMPv3 版本（RFC3376）。所有 IGMP 版本都支持 ASM（Any-Source Multicast）模型。IGMPv3 可以直接应用于 SSM（Source-Specific Multicast）模型，而 IGMPv1 和 IGMPv2 则需要通过 SSM-Mapping 技术来支持 SSM 模型。

目的

要使组播报文最终能够到达接收者，需要将接收者主机接入 IP 组播网络，并加入到相应的组播组中。IGMP 通过在主机侧和路由器侧交互 IGMP 报文实现组成员管理功能。IGMP 协议可以记录接口下主机的加入和离开等信息，以确保组播数据能够正确地转发到该接口。

3.2 参考标准和协议

本特性的参考资料清单如下：

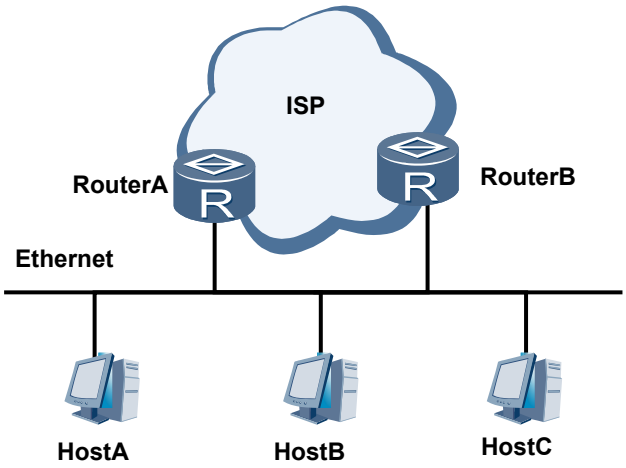
文档	描述	备注
RFC1112	Host Extensions for IP Multicasting	-
RFC2236	Internet Group Management Protocol, Version 2	-
RFC3376	Internet Group Management Protocol, Version 3	-
RFC3569	An Overview of Source-Specific Multicast (SSM)	-
RFC4601	Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)	-

3.3 原理描述

3.3.1 IGMPv1&v2&v3

IGMP 协议

图 3-1 IGMP 基本组网图



通过发送查询报文并接收主机反馈的加入报文和离开报文，路由器可以了解与该接口连接的网段上存在哪些组播组的接收者（即组成员）。如果出现组成员，组播路由器应将对应组的组播数据报文转发到这个网段；如果没有组成员则不转发。主机可以自主决定加入或退出某个组播组。

如**图 3-1** 所示，使能了 IGMP 协议的路由器 RouterA 会自动变为查询器并定时发出 IGMP 查询报文，与 RouterA 在同一网段的所有主机（HostA、HostB、HostC）都能收到它发出的查询报文。

- 主机收到查询报文的处理：
  - 如果主机之前已经加入了组 G，就会在路由器允许的响应时间内随机发送一个组 G 的 IGMP 加入报文。

    RouterA 收到组 G 的 IGMP 加入报文会记录组 G 的相关信息，同时对该组启动一个定时器（如果已经启动则刷新定时器），以便长时间没有主机响应时切断

该组的组播流量。RouterA 把组 G 的组播数据转发到 HostA 和 RouterA 相连的接口所在的网段上。

- 如果主机不是任何组播组成员，主机在收到 IGMP 查询报文时不作任何响应。

- 主机加入组播组：

  一个新加入组播组 G 的主机会主动发送一个该组的 IGMP 加入报文，通知路由器更新组播组信息，后续的加入报文则由路由器的查询报文来驱动。

- 主机离开组播组：

  如果一个主机决定离开某个组播组 G，会主动发送组 G 的 IGMP 离开报文。路由器收到后会触发一个指定组 G 的查询，确定该组在当前网段接收者的存在情况。如果在查询结束后仍然没有收到主机针对该组的 IGMP 加入报文，则删除已记录的组信息，停止转发该组数据到对应接口所在的网段上。

IGMPv1 处理 IGMP 报文

IGMPv1 协议主要基于查询和响应机制完成组播组管理，支持查询和加入报文，处理过程与 IGMPv2 相同。IGMPv1 与 IGMPv2 的不同之处是：主机离开组播组时不主动发送离开报文，收到查询消息后不反馈 Report 消息，待维护组成员关系的定时器超时后，路由器删除组记录。

IGMPv2 处理 IGMP 报文

运行 IGMPv2 的主机发送的 IGMP 报告中仅携带组信息。当主机发送一个组的 IGMP 加入报文给路由器后，路由器会通知组播转发模块，以便这个组的组播数据到来时能够正确转发给该主机。

IGMPv2 协议具有报告抑制机制，可以减少网络中的 IGMP 的重复报告。

当一个主机 HostA 加入了某个组播组 G，在收到路由器的查询报文后，HostA 会在 0 ～最大响应时间（查询报文中已经指定）之间选择一个随机值作为定时器的超时时间，并在该定时器超时后，向路由器发送组 G 的加入报文。如果在超时时间内，HostA 收到了加入同一个组的主机 HostB 发送的加入报文，则 HostA 不再向路由器发送组 G 的加入报文。

当主机退出某个组 G 时，会向路由器发送一个指定组 G 的 IGMP 离开报文。由于 IGMPv2 报告抑制机制，路由器无法确定是否还有其他主机加入了组 G。这时路由器会触发一个指定组 G 的查询，如果其他主机加入了组 G，就会发送针对组 G 的 IGMP 加入报文。

如果路由器发送了若干次数指定组 G 的查询之后，仍然没有收到主机针对组 G 的 IGMP 加入报文，那么路由器就不再记录组 G 的信息，停止转发该组数据到对应接口所在的网段。

说明

IGMP 的查询器和非查询器都会处理 IGMP 组加入信息，但是只有查询器负责发送查询报文。IGMP 非查询器不处理 IGMPv2 离开报文。

IGMPv3 处理 IGMP 报文

IGMPv2 报文中只能携带组播组的信息，不能携带组播源的信息。这样运行 IGMPv2 的主机就只能选择加入某个组，而不能选择加入某个组播源/组。IGMPv3 解决了该问题。运行 IGMPv3 的主机不仅能够选择组，还能根据需要选择组播源/组。主机发送的 IGMPv3 报文中可以包含多个组记录，每个组记录中可以包含多个组播源。

在路由器侧，查询器发送查询报文并接收主机反馈的加入报文和离开报文，以此来了解与该接口连接的网段上有哪些组播组存在接收者，并将组播数据转发到相应的网段。IGMPv3 的组记录有include 和 exclude 二种组过滤模式。

- 在 include 模式下
  - 处于激活状态的组播源表示需要路由器转发这个源的数据。
  - 不活动的源会被路由器删除并停止转发这个源的数据。
- 在 exclude 模式下
  - 处于激活状态的组播源表示处于冲突域中。也就是说，与该路由器接口同一网段的主机中，有的主机需要该源的数据，有的主机不需要该源的数据，在这种情况下该源的数据仍然需要转发。
  - 不活动的组播源表示不需要转发该源的数据。
  - 组中没有记录的组播源的数据全部都要转发。

在 IGMPv3 中，实现了对采用 Include 模式加入特定源组的 IGMPv3 主机成员信息的跟踪功能。

相对于 IGMPv2，IGMPv3 没有报文抑制机制，所有加入组播组的主机在收到查询时都会响应 IGMP 加入报文。由于有了对组播源的选择，IGMPv3 路由器在通用查询和组查询的基础上增加了指定源组查询，用以在收到特定组播源的数据时确定是否存在该数据的接收者。

3.3.2 IGMP 组兼容

IGMP 组兼容模式是指支持高 IGMP 版本的组播设备可以兼容低版本的主机。例如 v2 版本的组播设备可以正确处理 v1 主机的加入，v3 版本的组播设备可以正确处理 v1 和 v2 版本的主机加入。当组播设备工作在兼容模式时，收到低版本的主机的 IGMP 加入报文后会自动降低组的兼容版本到该主机对应的版本，并工作在该版本上。

工作在 v2 或 v3 版本的组播设备收到 IGMPv1 主机发送的 Report 报文时，会自动把该组的兼容模式设定为 v1 模式。在这种情况下，设备会忽略针对该组的 IGMPv2 Leave 报文。

工作在 v3 版本的组播设备收到 v2 版本的 Report 报文时，会自动把该组的兼容模式设定为 v2 模式。在这种情况下，设备会忽略 IGMPv3 的 BLOCK 报文、IGMPv3 的 TO_IN 报文以及 IGMPv3 的 TO_EX 报文的源列表，即抑制了 IGMPv3 对组播源的选择功能。

通过手工配置把组播设备从低版本升到高版本时，如果有组存在，则这些组继续工作在低版本的兼容模式，直到低版本的主机退出该组播组。

说明

缺省情况下，IGMP 的版本是 IGMPv2。

3.3.3 IGMP 查询器选举

使能了 IGMP 协议的组播设备在网段中的角色有两个：

- 查询器

  负责发送查询报文，并接收主机反馈的加入报文和离开报文，以此来了解与该接口连接的网段上有哪些组播组存在接收者（即组成员）。
- 非查询器

只接收主机反馈的加入报文，了解与该接口连接的网段上有哪些组播组存在接收者，并根据网段中查询器的动作确定当前网段中有哪些组播组成员离开。

通常情况下一个网段只有一个查询器，因此组播设备之间需要用某些方式来选出查询器。查询器选举时采用以下的原则：

- 组播设备 A 使能 IGMP 协议后，在 IGMP 协议启动阶段会默认自己为当前网段中的查询器，向网段中发送查询报文。如果收到 IP 地址比自己小的组播设备 B 发来的查询报文，则 A 由查询器转为非查询器，并启动其他查询器存在定时器，记录 B 为当前网段的查询器。
- 如果组播设备 A 在非查询器状态时，收到查询器组播设备 B 发送的查询报文，则更新其他查询器存在定时器。如果此时收到的查询报文不是先前记录的查询器 B 发来的，而是新的组播设备 C 发来的，且 C 的 IP 地址比 B 的小，则更新查询器为 C，同时更新其他查询器存在定时器。
- 如果组播设备 A 在非查询器状态时，其他查询器存在定时器超时，则由非查询器转为查询器状态，承担起查询器的职责。

说明

协议规定 IGMPv1 不支持查询器选举，IGMPv1 查询器由上层协议（如 PIM）指定。当前仅支持同网段上同版本的组播设备之间进行查询器选举。为了保证正常工作，需要在同网段所有组播设备上配置相同版本的 IGMP。

3.3.4 IGMP 支持 Router-Alert

通常情况下，只有目的地址属于本设备的接口地址时，报文才会上送给路由协议层处理。Router-Alert 是一种标示协议报文的特殊机制。如果一个报文中带有 Router-alert 选项，则表示该报文需要被上送到路由协议层去处理。在实际使用中，有些协议报文的目的地址是组播地址或其他特殊地址，如果报文中没有带有 Router-Alert 选项，可能不会被上送。

IGMP 报文的目的地址一般为组地址而非组播设备的接口地址，这样就导致报文可能不会被上送到路由协议层处理，Route-Alert 选项可以解决此类问题。

当组播设备收到 IGMP 报文时：

- 缺省情况下，不检查 Router-alert 选项。无论 IGMP 报文有没有 Route-Alert 选项，都会上送到路由协议层去处理。
- 在配置了检查 Router-Alert 选项的情况下，只有带有 Route-Alert 选项的 IGMP 报文才会被上送到路由协议层处理。

3.3.5 IGMP Only-Link

IGMP Only-Link 是指在组播设备与主机相连的接口上只使能 IGMP 协议，而不使能 PIM 等上层协议，只由 IGMP 协议来指导该网段的数据转发的一种机制。

相比由 PIM 协议指导某一网段数据转发的机制，应用该特性可以减少组播设备对 PIM 邻居、PIM 接口状态机等信息的维护。

应用 IGMP Only-Link 机制时，查询器有以下功能：

- 发送查询报文并接收主机反馈的加入报文和离开报文，以此来了解与该接口连接的网段上有哪些组播组存在接收者。

- 维护 IGMP 组播组加入/离开状态，并根据 IGMP 组播组的加入/离开状态来指导该网段的数据转发。

非查询器上则只维护 IGMP 组播组加入/离开状态。

说明

接口上使能 PIM 协议时由 DR 指导数据转发，详细介绍请参见 2.3.2 PIM-SM 中的“PIM DR 竞选的基本原理”。

3.3.6 IGMP On-Demand

组播设备通过发送查询报文并接收主机反馈的加入报文和离开报文来了解与该接口连接的网段上有哪些组播组存在接收者。和组播设备相连的可能不是主机，而是一个使能了 IGMP 代理的接入设备。

为了减少组播设备和接入设备间的报文交互，可以做这样的优化：接入设备汇聚其接收的 IGMP 组播组加入/离开状态，只在 IGMP 组播组的状态发生改变时才将该加入/离开状态上报给组播设备。即，接入设备在第一个用户加入组播组时发送加入报文，最后一个用户离开组播组时发送离开报文。在这样的应用场景下，组播设备侧的 IGMP On-Demand 特性应运而生。

使能 IGMP On-Demand 特性的组播设备不主动发送查询来确定某一个组播组在当前网段上是否有接收者，而是通过与该接口相连的 IGMP 代理设备将其汇聚的组播组加入/离开状态主动上报给组播设备的方式来实现 IGMP 组播组维护。

IGMP On-Demand 只适用于 IGMPv2 和 IGMPv3。组播设备使能了 IGMP On-Demand 特性后，与 IGMP 标准协议行为有 3 点不同：

- 不发送查询报文。
- 收到 IGMP 加入报文后创建组播组和源信息，且表项永不超时。
- 只有在收到 IGMP 离开报文后，组播设备才会删除对应表项。

3.3.7 IGMP Prompt-Leave

当主机退出某个组 G 时，会向组播设备发送一个指定组 G 的 IGMP 离开报文。由于 IGMPv2 报告抑制机制，组播设备无法确定是否还有其他主机加入了组 G。这时设备会触发一个指定组 G 的查询，如果其他主机加入了组 G，就会发送针对组 G 的 IGMP 加入报文。如果设备发送了若干次数指定组 G 的查询之后，仍然没有收到主机针对组 G 的 IGMP 加入报文，那么就不再记录组 G 的信息，停止转发该组数据到对应接口所在的网段。

如果组播设备只和一个使能了 IGMP 代理的接入设备相连，那么当该接入设备离开某个组播组 G 并向组播设备发送针对该组 G 的 IGMP 离开报文时，组播设备无需触发指定组 G 的查询报文来确定当前网段上该组 G 是否还有其他接收者，可以直接将该组 G 的组记录删除，停止转发该组数据到对应接口所在的网段。IGMP Prompt-Leave（快速离开）特性解决了该问题。

组播设备使能了 IGMP Prompt-Leave 特性后，当其接收到 IGMP 离开报文后，不会触发针对该组的查询报文，而是直接将该组的组记录删除，停止转发该组的数据到对应接口所在的网段。应用该特性可以提升组播设备响应组播组离开的速度。

说明

IGMP Prompt-Leave 特性只在 IGMPv2 版本上支持，其他版本不支持该特性。IGMP On-Demand 特性已经包含 IGMP Prompt-Leave 特性。

3.3.8 IGMP 策略控制

可控组播是对 IGMP 的行为进行限制或者扩充的附加特殊功能，对 IGMP 协议本身实现没有影响。具体包括 3 个功能：IGMP-Limit、静态组加入和 Group-Policy。

- IGMP-Limit

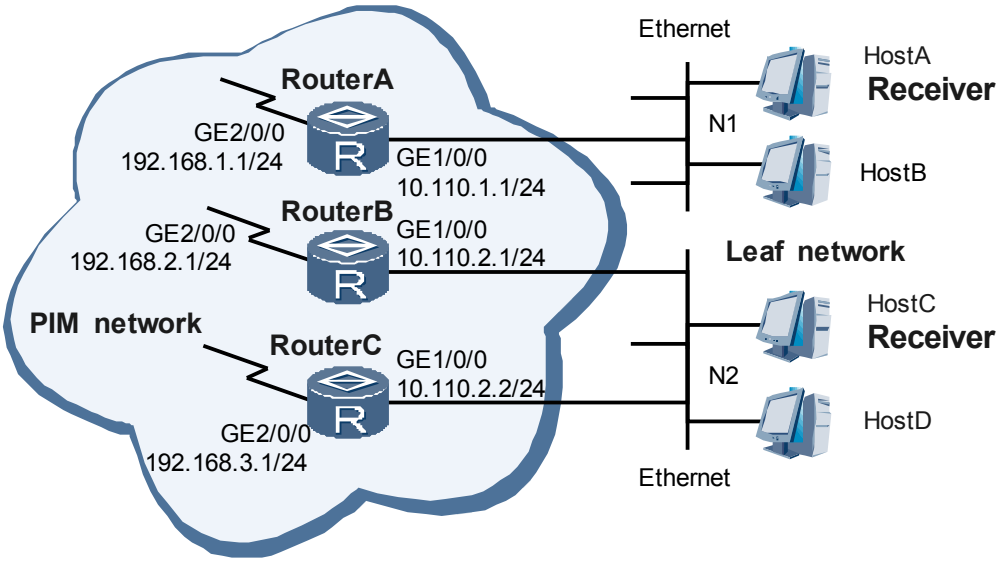
  通过配置接口表项限制、单实例和所有实例表项限制，限制组播组或源组数量。
- 静态组加入

  通过在接口上配置静态组播组加入，可以实现快速响应用户请求，将该组播数据转发给接收者，减少用户的频道切换时间。
- Group-Policy

  是管理员配置在路由器接口上的一种过滤组策略。配置 Group-Policy 之后，路由器可以对某些特定的组进行限制，不建立对应的表项。

IGMP-Limit

图 3-2 配置 IGMP-Limit 组网图



在大量用户同时收看多套节目时，需要占用路由器的大量带宽，会造成路由器性能下降。为了避免这种情况的发生，需要限制 IGMP 接口和全局下允许加入的组播组个数，使加入的组播组个数在给定的限制之内。这样能够使加入组播组的用户收看更加清晰稳定的节目。

IGMP-Limit 是指通过在路由器上的接口、单实例和所有实例下配置 IGMP 组播组限制个数。当收到 IGMP 加入报文时，首先判断是否超过配置的个数限制，如果没有超过，则建立组成员关系，给用户转发该组的数据流。

- IGMP 接口表项限制
  - 支持接口配置 IGMP 加入的接口表项限制。当接口收到 IGMP 协议加入报文时，可以对本接口的表项数量进行限制。
  - 支持通过配置 IGMP 表项限制，使某些组范围或源组范围不受接口加入限制，不参与接口 IGMP 加入的计数。

- IGMP 组播 VPN 单实例表项限制

  支持对组播 VPN 单实例的 IGMP 加入进行限制。单实例的 IGMP 加入限制即对当前实例下各个接口的 IGMP 加入的接口表项总和进行限制。

  - 接口收到 IGMP 协议加入报文时，判断是否超过本实例的 IGMP 加入表项限制。
  - 删除接口(*,G)/(S,G)成员关系时，减少本实例的 IGMP 接口加入表项的计数。

- IGMP 整路由器表项限制

  支持对整路由器的 IGMP 加入进行限制。整路由器的 IGMP 加入限制即对所有实例接口的 IGMP 加入的接口表项总和进行限制。

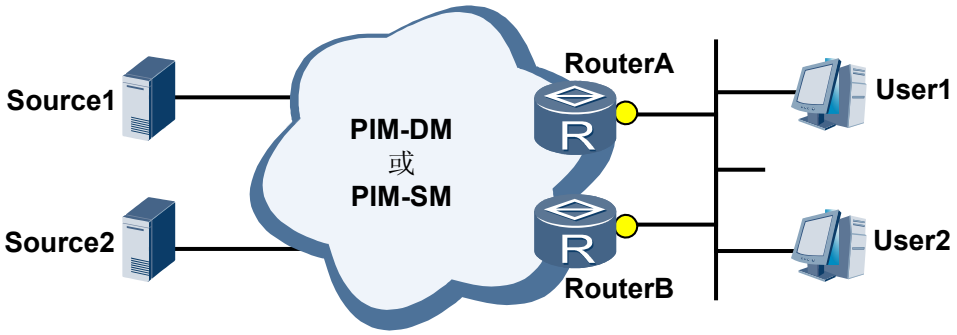
  - 收到 IGMP 协议加入报文时，判断是否超过整路由器的 IGMP 加入表项限制。
  - 删除接口(*,G)/(S,G)成员关系时，要减少整路由器的 IGMP 加入表项的计数。

以上三种表项限制策略均遵循以下计数规则：

- 每个(*,G)组成员关系和每个(S,G)源组成员关系各计为一个接口表项。
- 用于 SSM-Mapping 的每个(*,G)组成员关系计为一个接口表项，按照映射生成的(S,G)表项不进行计数。

静态组加入

图 3-3 配置静态组加入组网图



在路由器接口上配置静态加入组，相当于该网段上存在稳定的组成员

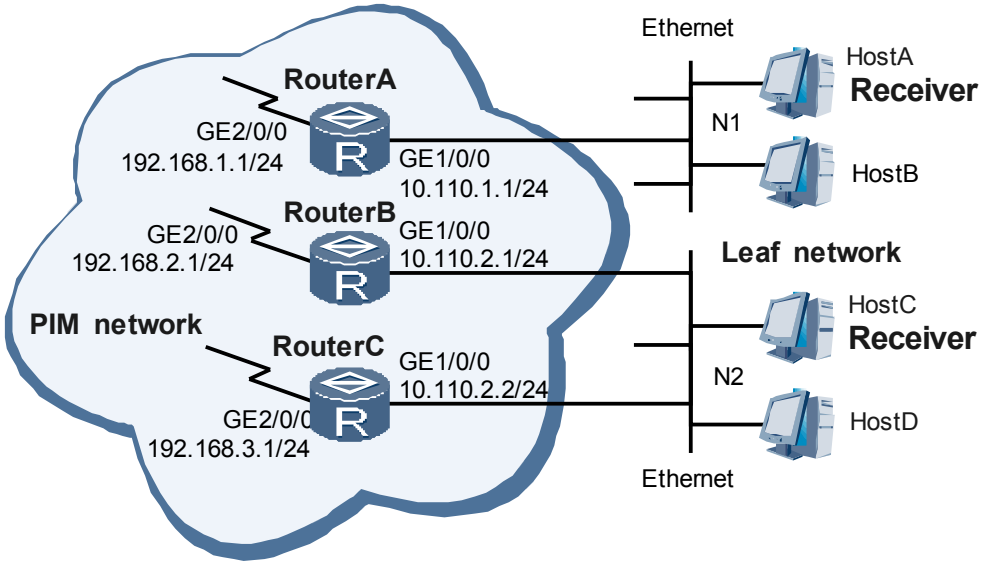
静态组播是通过在接口上配置静态组加入来实现的。配置静态组加入后，路由器创建的表项没有定时器，永远不会超时，因此路由器会持续地给接收者转发数据。接收者不再需要路由器转发该组播组数据时，不能通过表项超时自动删除，只能通过手动删除静态组配置来实现。

在实际应用中，静态组加入配置在接收端路由器与主机相连的接口上，以便将组播数据引到该路由器上。当与该路由器直连的主机或路由器中有接收者想接收对应的组播组数据时，该路由器就能做到快速响应，将相应的组播数据转发给接收者，这样可以减少用户的频道切换时间。

组播静态加入同时支持单条配置和批量配置，即一条静态加入配置支持一个组播组(源组)加入和多个组播组(源组)加入。

IGMP Group-Policy

图 3-4 配置 Group-Policy 组网图



Group-Policy 是管理员配置在路由器接口上的一种过滤策略，配置 Group-Policy 之后，路由器可以对某些特定的组进行限制，不建立对应的表项。

在大量的用户同时收看多套节目组时，会占用路由器大量的带宽，同时也会降低路由器的性能。这时可以使用 Group-Policy 对某些组进行限制，使组播组数量控制在一定的范围内。另外，出于安全考虑或者管理的需要，路由器可能不希望接收某些组的加入报文，不希望转发该组播组的数据，也可以通过 Group-Policy 加以限制。

Group-Policy 的过滤规则统一通过 ACL 配置。

3.3.9 SSM Mapping

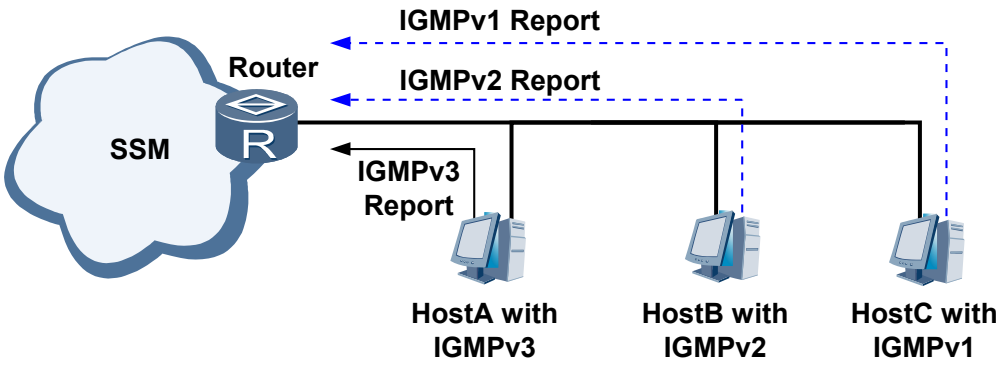
SSM Mapping（Source Specific Multicast Mapping）机制可以更好的兼容运行 IGMPv3 之前的版本的主机，使其也能够使用组播 SSM 范围的服务。SSM Mapping 机制是：将处于 SSM 范围的 IGMPv1/v2 的(*,G)加入按照配置的转化规则转化成对应的一组(S,G)加入。这样应用低版本 IGMP 的用户也可以获得 SSM 范围的组播服务。

同时，应用 SSM Mapping 机制能够很好的保护组播源服务器，减少其受到攻击的可能。

说明

对于 SSM 范围的组，组播设备不接收整个组的需求，只处理源组的需求。SSM 的详细介绍请参见 **PIM-SSM**。

图 3-5 SSM Mapping 应用组网图



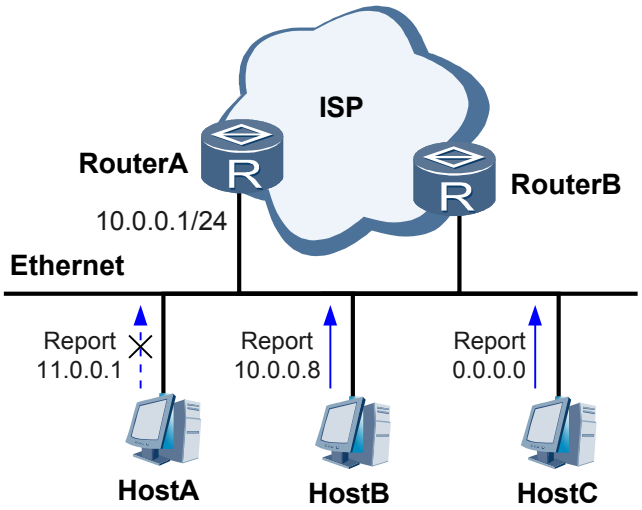
如图 3-5 所示，SSM 网络的用户网段中 HostA 运行 IGMPv3、HostB 运行 IGMPv2、HostC 运行 IGMPv1。在不允许将 HostB 和 HostC 升级为 IGMPv3 的情况下，为该网段中的所有主机提供 SSM 组播服务，需要在组播设备上支持 SSM Mapping。

如果组播设备支持 SSM Mapping，并且配置了相应的转化规则，设备在接收到 HostC 或者 HostB 发出的 IGMP(*,G)加入报文的时候，会分情况进行处理：

- 报文中的组播组为 ASM 范围，则按照 3.3.1 IGMPv1&v2&v3 章节描述的方式进行处理。
- 报文中的组播组为 SSM 范围，则通过查找配置的转化规则将报文中指定的(*,G)加入，转化成对应的一组(S,G)加入。

3.3.10 IGMP 主机地址过滤

图 3-6 IGMP 主机地址过滤应用组网图



为了保证组播流量发送的准确性，允许用户在组播设备的接口配置 IGMP 主机地址过滤策略：

- 如果 IGMP 报文主机地址和接收接口地址在同一网段或者主机地址是网络地址(0.0.0.0)，那么此报文的主机地址过滤检查通过；
- 如果报文主机地址和接收接口地址不在同一网段，那么此报文的主机地址过滤检查不通过，丢弃此报文。

如图 3-6 所示，RouterA 与 Host 相连的接口地址为(10.0.0.1/24)，HostA 发送的 IGMP Report 报文的主机地址为 11.0.0.1，HostB 发送的 IGMP Report 报文的主机地址为 10.0.0.8，HostC 发送的 IGMP Report 报文的主机地址为 0.0.0.0，那么 RouterA 处理 HostB 和 HostC 的报文，丢弃 HostA 的报文。

3.3.11 IGMP 支持多实例

IGMP 支持多实例时，组播设备根据接口所属的实例来处理协议报文的收发。当组播设备从网络上收到 IGMP 报文时，需要区分接收该报文的接口所属的实例，并在该实例范围内对其进行处理。当 IGMP 需要和其它组播协议交互信息时，也只会通知本实例内的其它组播协议。

具体 IGMP 报文的处理请详见文档 3.3.1 IGMPv1&v2&v3 章节。

3.3.12 协议的比较

IGMPv1 和 IGMPv2 协议的比较：

IGMPv1	IGMPv2	IGMPv2 较 IGMPv1 的优势
报文类型不包含成员离开报文	报文类型包括成员离开报文	可以更及时有效管理组播组成员
只支持普遍组查询	除了支持普遍组查询，还支持特定组查询	可以直接对特定的组播组进行选择，增加了选择的粒度

IGMPv2 和 IGMPv3 协议的比较：

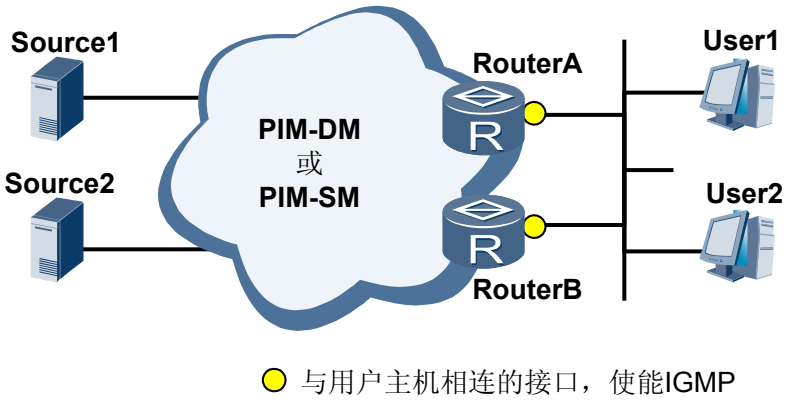
IGMPv2	IGMPv3	IGMPv3 较 IGMPv2 的优势
报文中不能携带组播源信息，只能指定组播组信息	报文中除了携带组播组信息，还能携带组播源信息	可以直接对特定的组播源进行选择，增加了选择的粒度
一个报文中只能携带一个组记录	一个报文中可以携带多个组记录	减少了网段中的 IGMP 报文数量
指定组查询报文无重传机制	指定组、指定源组查询报文有重传机制	非查询器和查询器维护的组播组信息能够更好地保持一致

3.4 应用

3.4.1 IGMP 典型应用

3.4.1 IGMP 典型应用

图 3-7 IGMP 应用典型组网图



IGMP 是处理主机加入路由网络的协议。因此，该协议应用于路由边界与主机相连的区域。该协议可以处理多主机和多个组播设备分别使用不同版本时的情况。

IGMP On-Demand 特性和 **IGMP Prompt-Leave** 特性只适用于共享网段上只有一个组播设备和单一接入设备的情况。

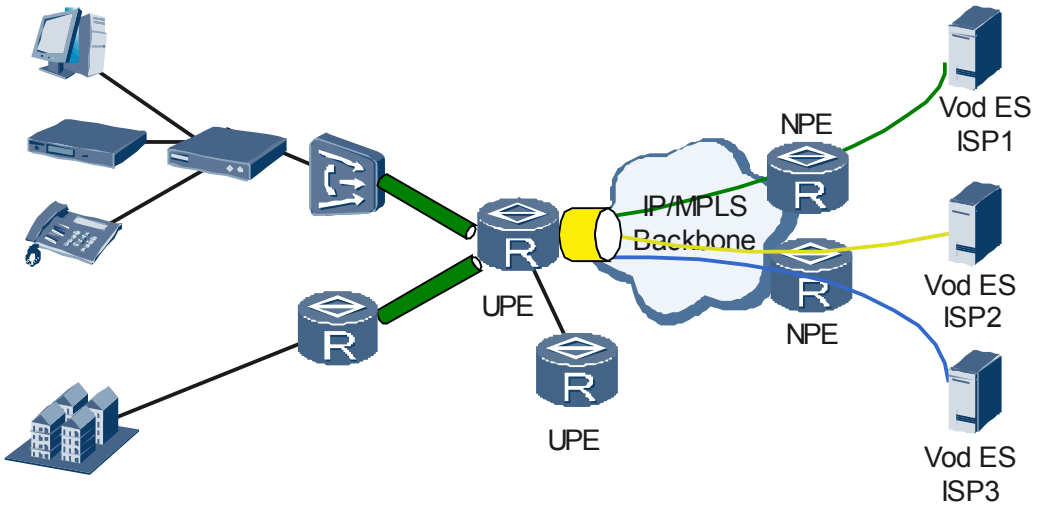
3.4.2 IGMP 表项限制应用

如图 3-8 所示，在 UPE 路由器的接口上可以对 IGMP 加入的数量进行限制。包括：

- IGMPv1/v2 report 加入的组记录数量
- IGMPv3 report 加入的源组记录数量和 exclude 模式组记录数量

在 UPE 上，不仅可以对每个接口进行 IGMP 加入的限制，也可以对全局的 IGMP 加入进行限制（包括所有实例和单实例）。

图 3-8 IGMP 表项限制组网图



3.4.3 BAS 用户组播

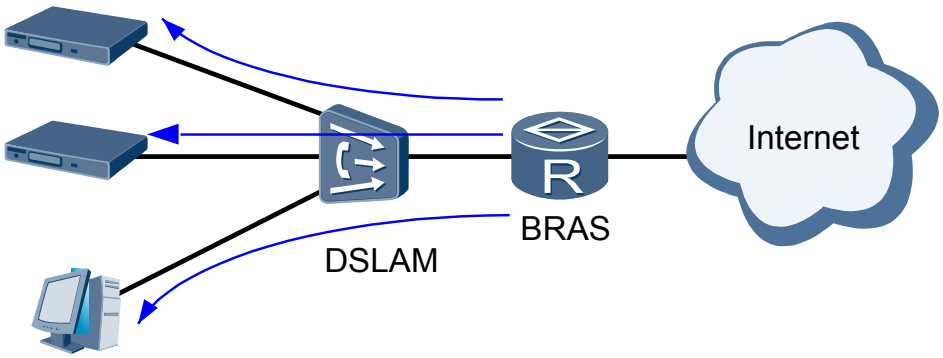
当 IGMP 特性应用在 Bas 接口上时，Bas 接口下的接入用户可以实时加入某组播组。

普通端口的组播数据按照端口复制，而由于 Bas 接口下可能存在多个接入用户，所以 Bas 接口下组播数据的复制方式存在以下几种。包括：

- 按用户进行复制
- 按 MVLAN（组播 VLAN）进行复制
- 按端口进行复制

当下游的二层设备不具备组播复制能力时，需要 Bras 设备启用按用户方式的组播复制，下游的二层设备只负责转发组播流量，由 Bras 设备完成组播复制，如图 3-9 所示。

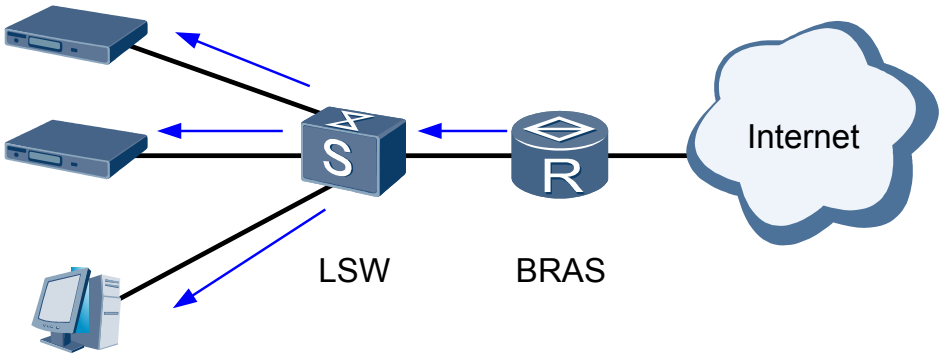
图 3-9 BRAS 设备组播复制



当下游的二层设备支持组播复制时，Bras 设备可以启用按 MVLAN 组播复制，由二层设备完成组播复制，如图 3-10 所示。

按端口复制是按 MVLAN 复制的一种特例，即 MVLAN 的值为 0 即为按端口复制。

图 3-10 二层设备组播复制



3.5 术语与缩略语

术语

术语	解释
IGMP	Internet Group Management Protocol，称为因特网组管理协议，是 IP 组播在末端网络上使用的主机对组播设备的信令机制。 主机通过 IGMP 加入或者离开组播组；组播设备通过 IGMP 了解下游网段是否存在组播组成员。
(S,G)	属于组播路由表项，S 表示组播源，G 表示组播组。 源地址为 S、组地址为 G 的组播报文，到达组播设备后，从(S,G)表项中的下游接口转发出去。 通常，将源地址为 S，组地址为 G 的组播报文表示为(S,G)报文。
(*,G)	属于 PIM 路由表项，*表示任意源，G 表示组播组。 (*,G)表项适用于所有组地址为 G 的组播报文。不论是哪个组播源发出的，只要是发往组播组 G 的组播报文，都应该从(*,G)表项中的下游接口转发出去。

缩略语

缩略语	英文全称	中文全称
ASM	Any-Source Multicast	任意源组播
IGMP	Internet Group Management Protocol	因特网组管理协议
SSM	Source-Specific Multicast	指定源组播

4 二层组播

关于本章

4.1 介绍

定义

二层组播是指提供链路层组播，以实现组播信息在物理网络上正确传输。当前利用 IGMP Snooping 对上游设备和主机之间交互的 IGMP 协议报文进行侦听，建立二层组播转发表，实现组播数据报文在数据链路层的按需分发。

二层组播的基本原理是使二层设备可以识别组播组地址，将组播组地址与对应的端口记录在自己内部的一个转发表中，根据这个转发表进行组播数据的转发。组播组地址可以为组播 IP 地址，也可以为映射后的组播 MAC 地址。二层设备可以根据组播地址查找转发表项里的出端口，进行组播数据的转发。

目的

在网络运行环境中，当上游设备将组播报文转发下来以后，处于接入边缘的路由器负责将组播报文转发给组播用户。如果路由器上没有二层组播功能，当其收到组播数据报文时，由于不知道哪些端口下存在接收者，所以以广播方式在报文所属的广播域内发送该组播报文，此广播域内的组播成员和非组播成员都能收到组播数据报文。这样不但浪费了网络带宽，而且影响了组网安全。

功能

二层组播主要包括以下功能：

- IGMP Snooping：当运行 IGMP Snooping 的路由器收到主机和上游设备之间传递的 IGMP 消息时，路由器的 IGMP Snooping 模块分析消息携带的信息，根据这些信息建立和维护二层组播转发表，从而管理和控制组播数据报文的转发，实现组播数据报文在数据链路层的按需分发。
- IGMP Proxy：通过配置 IGMP Proxy 可以实现路由器的如下功能：对于网络侧相当于主机，响应路上游设备的查询报文，并将用户主机加入，离开组播组的信息汇总处理后通告路由器；对于主机侧相当于上游设备，负责向用户发送 IGMP 查询报文，并处理用户发来的 IGMP 响应报文。这样可以减少两侧网络带宽的浪费。
- 组播 VLAN 复制：配置了组播 VLAN 复制功能后，上游设备只需把组播数据传送给组播 VLAN 即可，由组播 VLAN 在路由器上实现组播流的复制，减少网络带宽的浪费。
- 组播 VLAN 1+1 保护：组播 VLAN 1 ＋ 1 保护功能通过工作 VLAN 和保护 VLAN 实现了组播流的 1 ＋ 1 备份，提高了组播的可靠性。
- 静态二层组播：静态二层组播功能通过手工配置二层组播转发表，将端口与组播地址表项进行静态绑定，把组播数据报文转发给长期需要接收该数据的主机。
- 二层组播实例：二层组播实例可以实现相同的组播数据只用一份在 VLAN 或者 VPLS 域内传递，然后向不同域的用户复制，节约了大量带宽。
- 备用设备快速转发组播流：在 VPLS 组网情况下，为避免主用链路或设备发生故障时，经过主备切换后的备用设备无法立即转发组播数据流，可以在备用设备上配置在备用设备上快速转发组播流功能。

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
RFC4541	Considerations for IGMP and MLD Snooping Switches	-
RFC1112	Host Extensions for IP Multicasting	-
RFC2236	Internet Group Management Protocol， Version 2	-
RFC3376	Internet Group Management Protocol， Version 3	-

4.3 原理描述

4.3.1 二层组播的原理

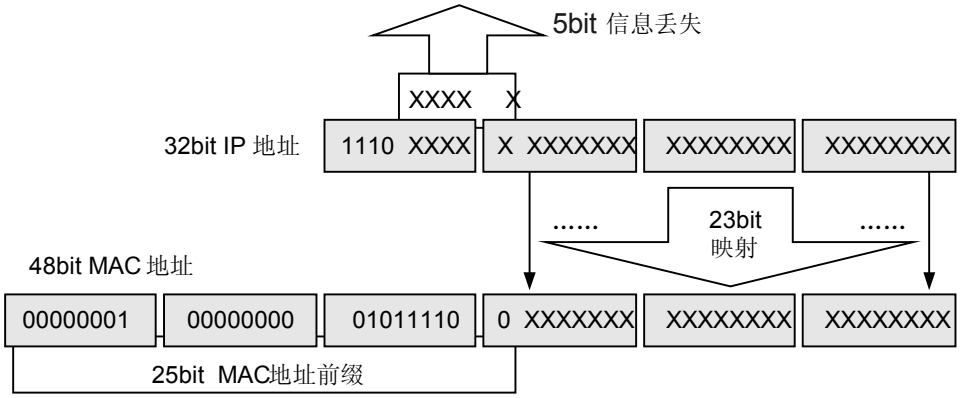
二层组播的基本原理是使交换机可以识别组播组地址，将组播组地址与对应的端口记录在自己内部的一个转发表中，根据这个转发表进行组播数据的转发。

组播组地址可以为组播 IP 地址，也可以为映射后的组播 MAC 地址。交换机根据组播地址查找转发表项里的出端口，进行组播数据的转发。

4.3.2 组播 MAC 地址

IANA 规定，组播 MAC 地址的高 24bit 为 0x01005e，第 25bit 为 0，低 23bit 为组播 IP 地址的低 23bit，映射关系如**图 4-1** 所示。

图 4-1 组播 IP 地址与组播 MAC 地址的映射关系



例如某组播组的组播 IP 地址为 224.0.1.1，则该组播组的组播 MAC 地址为 01-00-5e-00-01-01。

由于 IP 组播地址的前 4bit 是 1110，代表组播标识，而后 28bit 中只有 23bit 被映射到 MAC 地址，这样 IP 地址中就有 5bit 信息丢失，直接的结果是出现了 32 个 IP 组播地址映射到同一 MAC 地址上，当按 MAC 地址转发时，如果发生地址冲突，请将配置修改成按 IP 地址转发组播数据。例如组播 IP 地址为 224.0.1.1、224.128.1.1、225.0.1.1、239.128.1.1 等组播组的组播 MAC 地址都为 01-00-5e-00-01-01。

4.3.3 基本概念

组播转发表项

组播转发表项有两种：静态表项和动态表项。

- 静态表项
- 静态表项由用户手工配置，不会老化
- 动态表项
- 动态表项是通过运行在链路层设备上的协议分析主机和组播设备之间传递的 IGMP 消息来维护的。

动态表项具备老化功能，到达老化时间而未被更新的动态转发表项将被删除。

有了组播转发表，链路层设备就可以根据组播转发表项，将来自上游的组播数据报文转发给接收者主机。

运行 IGMP Snooping 功能的组播设备依据出端口信息，转发组播报文。

出端口信息

出端口信息相当于组播转发表项。每条“出端口信息”都包括如下内容：

- VLAN 的编号
- 组播组地址

- 路由器端口（相当于上游端口）
- 组播组成员端口（相当于下游端口）

当组播设备接收到组播报文时，依据报文所属 VLAN 和报文的目的地址（即组播组地址）查找是否存在对应的“出端口信息”。

- 如果存在，则将报文发送到所有组播组成员端口。只要能够维持组播组成员端口就可以保证组播报文按需转发。
- 如果不存在，则丢弃该报文或将报文在 VLAN 内广播。

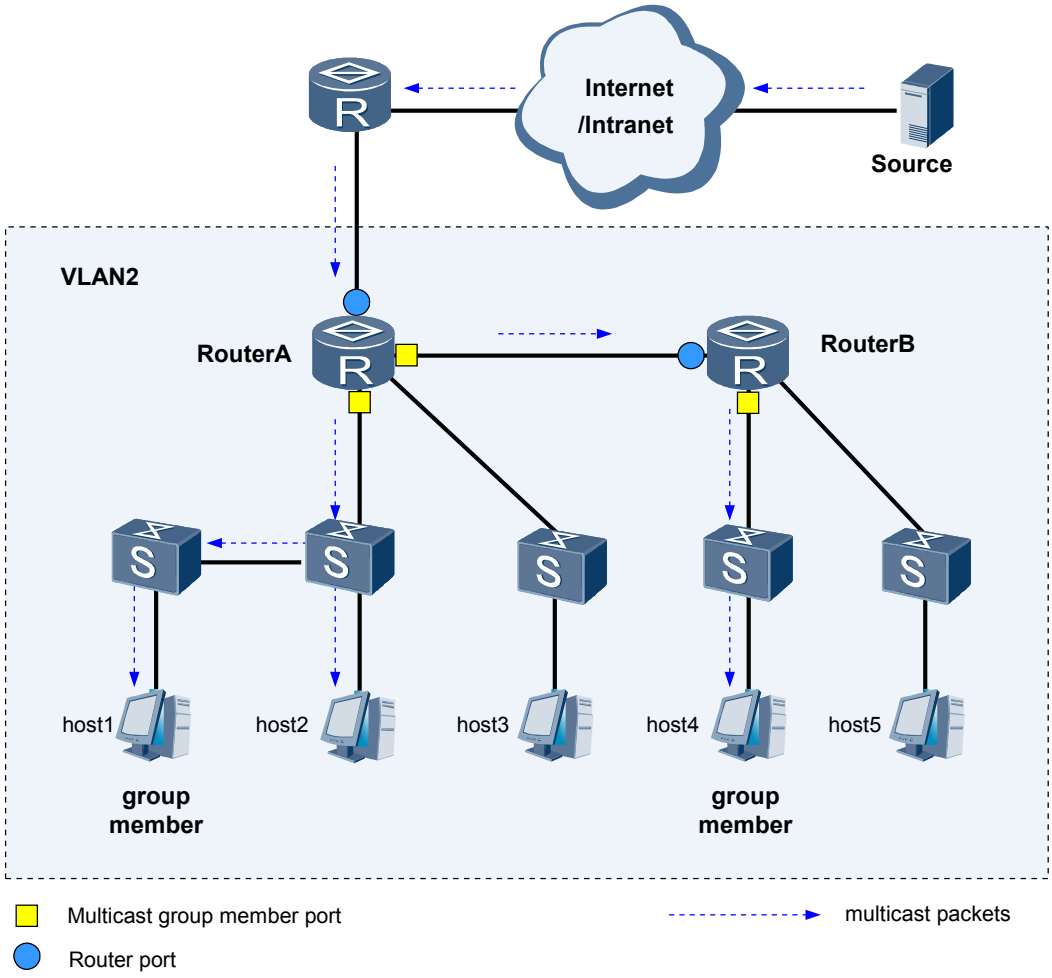
出端口类型

说明

本节结合具体示例介绍出端口类型。

- 在如图 4-2 所示，RouterA 和 RouterB 上运行 IGMP Snooping。

图 4-2 出端口类型



路由器端口

路由器端口是指朝向组播路由器的端口，二层组播设备从该端口接收组播报文。路由器端口分为如下两类。

- 动态路由器端口（图 4-2 中蓝色圆形）：能够接收到源地址不为 0.0.0.0 的 IGMP Query 报文或者 PIM Hello 报文的端口。动态路由器端口依赖于组播设备与主机之间交互的协议报文，动态维护。每个动态路由器端口启动一个“路由器端口老化定时器”，定时器超时则该路由器端口失效。
- 静态路由器端口：用户使用配置命令指定的，不会老化。（图 4-2 中未体现）

组播组成员端口

组播组成员端口是指朝向组成员主机的端口，二层组播设备从该端口发出组播报文。组播组成员端口简称为成员端口，分为如下两类。

- 动态成员端口（图 4-2 中黄色方块）：能够接收到 IGMP Report 报文的端口。动态成员端口依赖于组播设备与主机之间交互的协议报文，动态维护。每个动态成员端口启动一个“成员端口老化定时器”，定时器超时则该成员端口失效。
- 静态成员端口：用户使用配置命令指定的，不会老化。（图 4-2 中未体现）

4.3.4 IGMP Snooping

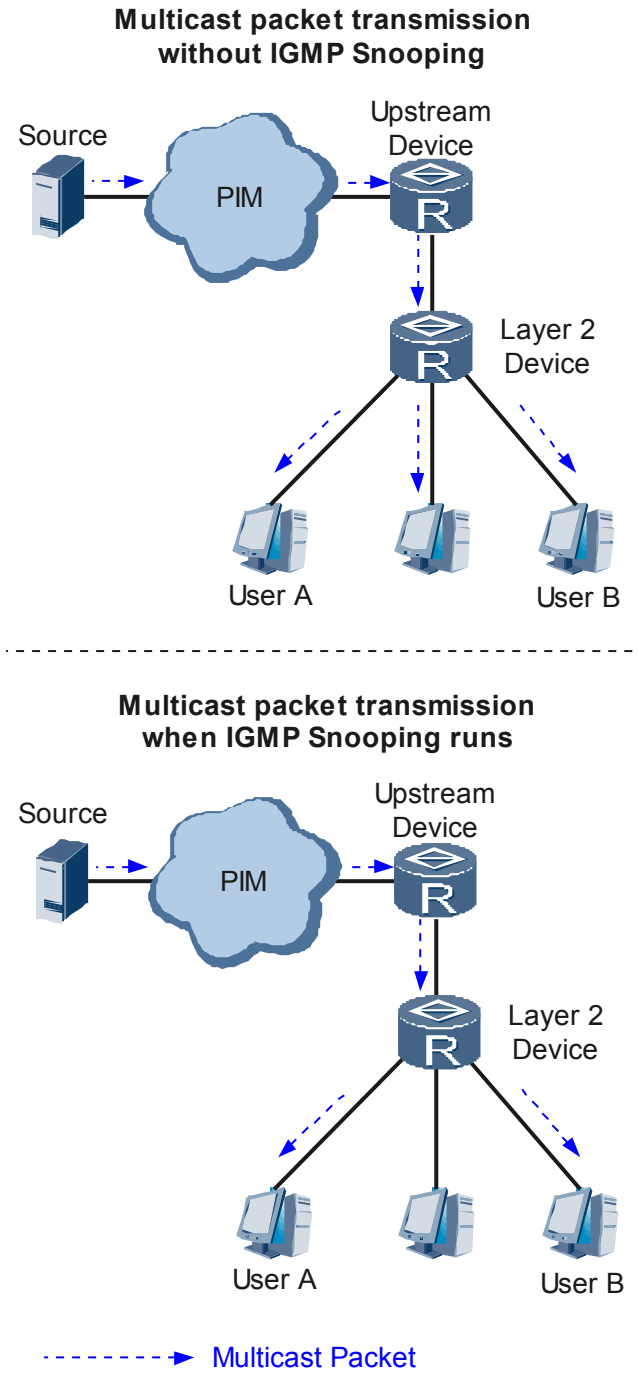
基本原理

IGMP Snooping 是 Internet Group Management Protocol Snooping（互联网组管理协议窥探）的简称，它是运行在二层设备上的组播约束机制，用于管理和控制组播组。

根据 IGMP 协议，主机加入组播组时，需要向上游设备发送 IGMP 报文，向其报告加入组播组消息，这样上游设备才可以将组播报文发送给主机。由于 IGMP 报文是封装在 IP 报文内，属于三层协议报文，而二层设备不处理报文的三层信息，所以这个过程它并不知道，而且通过对数据链路层数据帧的源 MAC 地址的学习也学不到组播 MAC 地址（数据帧的源 MAC 地址不会是组播 MAC 地址）。这样当二层设备在接收到一个目的 MAC 地址为组播 MAC 地址的数据帧时，在以前学习的 MAC 地址表中就不会找到对应的表项。那么这时候，它就会采用广播方式发送接收到的组播报文，这样一来不但会造成带宽的极大浪费，而且影响网络安全。

IGMP Snooping 是解决在路由器上实现数据链路层组播的一种方案。当运行 IGMP Snooping 的路由器收到主机和上游设备之间传递的 IGMP 消息时，路由器的 IGMP Snooping 模块分析消息携带的信息，根据这些信息建立和维护二层组播转发表，转发表中包括 VLAN 的编号，组播组地址，路由器端口（相当于上游端口），组播组成员端口（相当于下游端口）。当路由器接收到组播报文时，依据报文所属 VLAN 和报文的目的地址（即组播组地址）查找是否存在匹配的转发表项。如果存在，则将报文发送到所有组播组成员端口。如果不存在，则丢弃该报文或将报文在 VLAN 内广播。对于不存在匹配的转发表项的组播报文，路由器是丢弃还是在 VLAN 中广播，由产品具体决定。

图 4-3 二层设备运行 IGMP Snooping 前后的对比



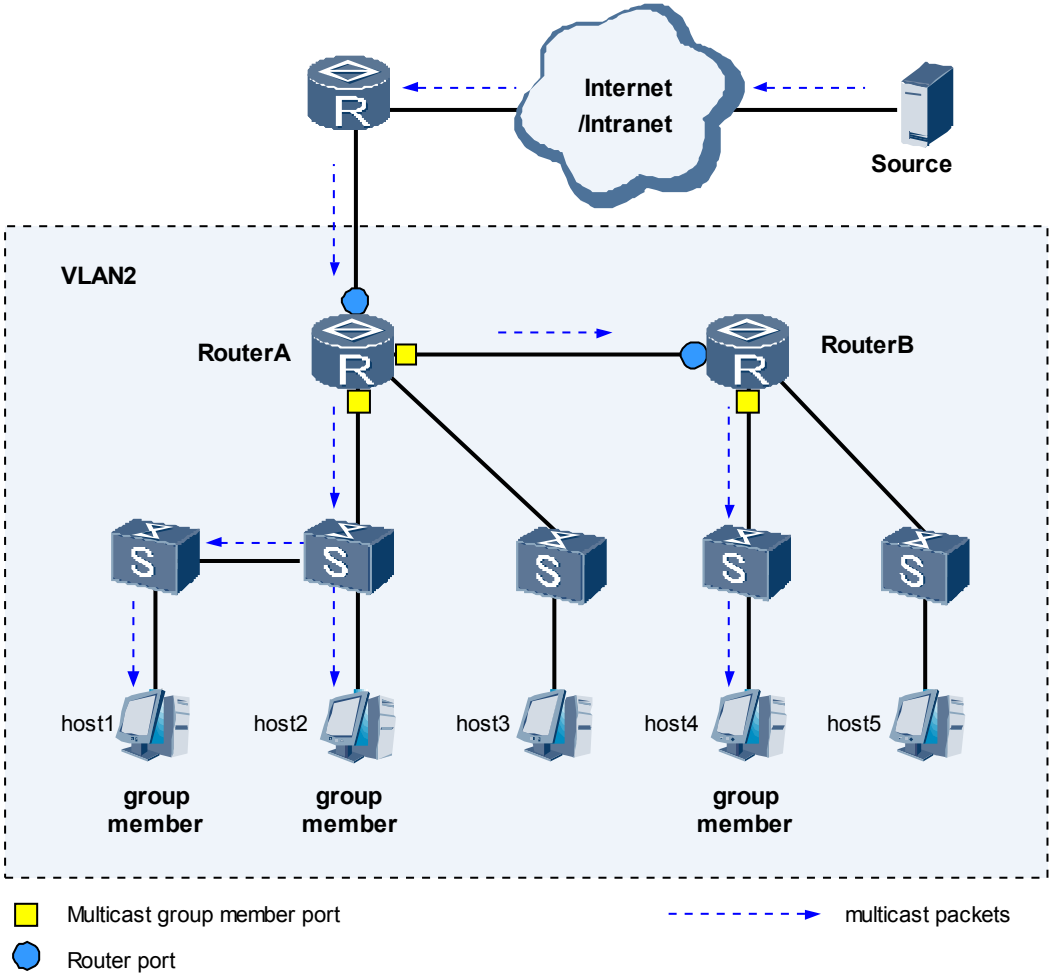
如图 4-3 所示，当路由器作为二层设备时，

- 如果路由器没有运行 IGMP Snooping，组播数据在数据链路层被广播。
- 如果路由器运行了 IGMP Snooping，已知组播组的组播数据不会在数据链路层被广播，而会发给指定的接收者。

相关概念

如图 4-4 所示，RouterA 和 RouterB 上运行 IGMP Snooping。

图 4-4 运行 IGMP Snooping 时组播数据报文的转发



结合图 4-4，介绍一下运行 IGMP Snooping 的路由器中的相关概念：

- 路由器端口

  路由器端口是指朝向组播路由器的端口，数据链路层组播设备从该端口接收组播报文。路由器端口分为如下两类。

  - 动态路由器端口（图 4-4 中圆圈）：能够接收到源地址不为 0.0.0.0 的 IGMP Query 报文或者 PIM Hello 报文的端口。动态路由器端口依赖于组播设备与主机之间交互的协议报文，动态维护。每个动态路由器端口启动一个路由器端口老化定时器，定时器超时则该路由器端口失效。
  - 静态路由端口：用户使用配置命令指定的，不会老化。（图 4-4 中未体现）

- 组播组成员端口

  组播组成员端口是指朝向组成员主机的端口，数据链路层组播设备从该端口发出组播报文。组播组成员端口简称为成员端口，分为如下两类。

- 动态成员端口（图 4-4 中方块）：能够接收到 IGMP Report 报文的端口。动态成员端口依赖于组播设备与主机之间交互的协议报文，动态维护。每个动态成员端口启动一个成员端口老化定时器，定时器超时则该成员端口失效。
- 静态成员端口：用户使用配置命令指定的，不会老化。（图 4-4 中未体现）

- 组播转发表项

  组播转发表项有两种：静态表项和动态表项。

  - 静态表项，静态表项由用户手工配置，不会老化
  - 动态表项，动态表项是通过 IGMP Snooping 协议分析主机和网络层设备之间传递的 IGMP 消息来生成的

  动态表项具备老化功能，到达老化时间而未被更新的动态转发表项将被删除。

  有了组播转发表，链路层设备就可以根据组播转发表项，将来自上游的组播数据报文转发给接收者主机。

  运行 IGMP Snooping 功能的组播设备依据出端口信息，转发组播报文。

- 出端口信息

  出端口信息相当于组播转发表项。每条“出端口信息”都包括如下内容：

  - VLAN 的编号
  - 组播组地址
  - 路由器端口（相当于上游端口）
  - 组播组成员端口（相当于下游端口）

  当组播设备接收到组播报文时，依据报文所属 VLAN 和报文的目的地址（即组播组地址）查找是否存在对应的出端口信息。

  - 如果存在，则将报文发送到所有组播组成员端口。
  - 如果不存在，则丢弃该报文或将报文在 VLAN 内广播。

其他功能

路由器在运行 IGMP Snooping 的基础上支持成员端口快速离开和组播组策略功能。

- 成员端口快速离开

  成员端口快速离开是指当路由器某端口收到主机发送的离开某指定组播组的 IGMP Leave 消息时，不发送 IGMP 查询报文，立刻将该端口从指定组播组的出端口信息中删除。而且只有当 VLAN/VSI 内的每个组播成员端口都只连接一台接收者主机时，才能在该 VLAN/VSI 内配置允许成员端口快速离开。否则，当端口下有多个接收者主机时，一个主机离开，可能会造成同一组播组中的其它接收主机组播中断。

  在组播应用于 IPTV 的场景下，需要配置成员端口快速离开机制，在这种场景中路由器端口一般只连接一个用户主机，快速离开可以保证用户切换频道的速度。

  IGMP Snooping prompt-leave 有以下优点：

  - 减小响应延迟
  - 节省因各种消息而占用的网络带宽

- 组播组策略

  IGMP Snooping 支持用户通过命令配置组播组策略，使 VLAN/VSI 内的端口只能加入符合 ACL 规则的组播组。配置策略后可以指定一组组播地址为不生成转发表的地址，这样即使路由器收到这些地址的 Report 报文，也不会生成转发表。通过组播组策略，可以限制用户加入组播组范围，提高了组播的可控性和安全性。

为了与组播路由器的 SSM 模式匹配，允许在路由器上配置 SSM-Policy，指定属于 SSM 范围的组播组。SSM 模式要求 IGMP 报文携带组播源信息，IGMP v2 报文不携带组播源信息，在该范围内的 IGMP v2 加入报文不会生成表项。

支持接口

IGMP Snooping 运行在二层设备上，目前 IGMP Snooping 支持以太网的大部分物理接口和逻辑接口：

- 支持的以太网物理接口类型有：GE 接口、Ethernet 接口
- 支持 QINQ 终结子接口和 Dot1q 终结子接口
- 支持的以太网逻辑接口有：ETH-Trunk 接口、GE 子接口、Ethernet 子接口、ETH-Trunk 子接口
- 也支持在 VSI 的 pw 上学习和配置二层组播转发表项

IGMP Snooping 不支持的以太网接口类型有：Virtual-Ethernet、QinQ 类型的 VLAN Stacking 接口。另外 IGMP Snooping 不支持其他不属于以太网络的接口，如 POS 接口、ATM 接口等。

使用价值

在用户主机网段的路由器上应用 IGMP Snooping，有益于以下三个方面：

- 节约网络带宽，方便对主机单独计费。
- 各个 VLAN 独立转发，提高信息安全性。
- 快速响应链路故障，增强可靠性。

4.3.5 静态二层组播

在二层组播中，除了通过二层组播协议动态建立组播转发表外，还可以通过手工配置二层组播转发表，将端口与组播地址表项进行静态绑定，即静态二层组播。

二层静态组播有以下优点：

- 避免协议报文的攻击。
- 采用直接查找组播报文转发表转发报文，减少网络延时。
- 避免未注册的用户收到组播流，提高了信息安全性，实现服务的有偿提供机制。

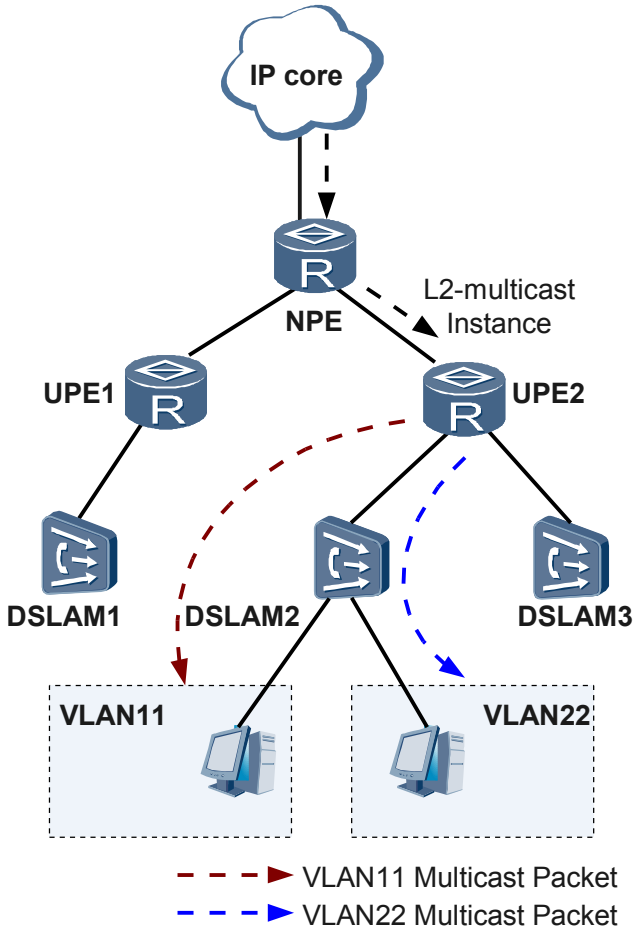
4.3.6 二层组播实例

传统的组播点播方式下，属于不同 VLAN 域或者 VPLS 域的用户通过同一台设备进行同一组播源点播时，需要为每个 VLAN 或 VSI 都复制一份组播数据，这样既造成了带宽浪费，同时也给上游设备增加了额外的负担。

为了解决这个问题，可以在整个二层网络中部署一个或几个二层组播实例，当每个用户在 VLAN/VSI 内都接收相同组播流的时候，上游设备只需要发送一份组播流然后向不同的用户复制，大量节约了带宽。

在如图 4-5 所示的组网中，如果在不同 VLAN 内的用户都需要相同的组播数据流，可以在 UPE2 设备上部署二层组播实例，实现从 NPE 向 UPE2 设备下发一份组播流后向各个不同 VLAN 的用户复制，从而节约了 NPE 与 UPE2 之间的带宽。

图 4-5 组播实例的应用



4.3.7 IGMP Proxy

IGMP Proxy 是靠拦截用户和上游设备之间的 IGMP 报文建立组播转发表，使能 IGMP Proxy 的路由器有两种身份：对于主机它是一个查询器，发送 Query 消息。对于上游设备，它是一个主机，发送 Report 和 Leave 消息。

IGMP Proxy 的工作过程和原理如下所述：

1. 使能 IGMP Proxy 的路由器代替上游设备发送 IGMP Query 消息，形成组成员和端口的对应关系，这样可以节约上游设备和路由器之间的带宽。
2. 当上游设备发送 IGMP 查询报文后，同一组播组内无论 VLAN 下多少主机组加入，路由器都只需要发送一份 IGMP Report 报文给上游设备，这样可以减少网络侧带宽。

4.3.8 组播 VLAN 复制

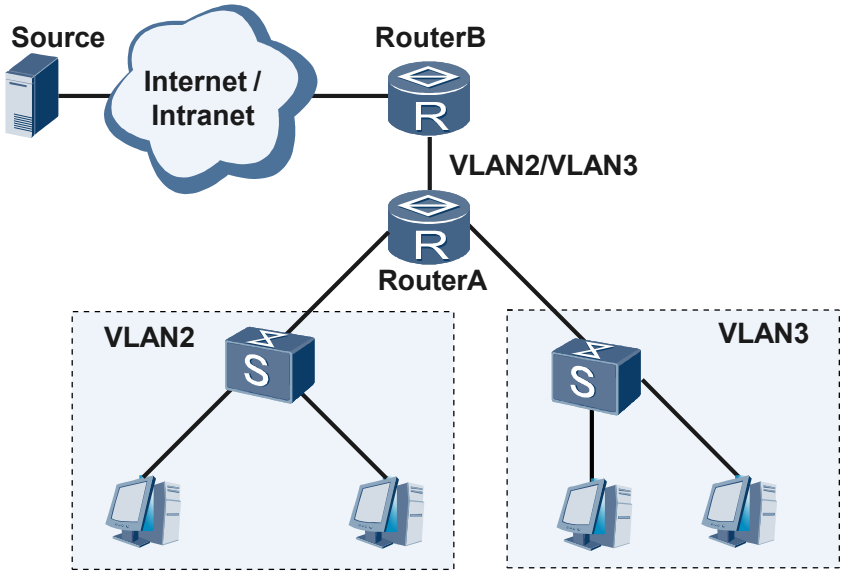
如果多个不同 VLAN 内的用户需要接收相同的组播流，可以通过配置组播 VLAN 复制功能来实现。通过实现组播 VLAN 复制功能，可以对组播源和组播组成员进行管理和控制，同时也可以减少带宽浪费。

组播 VLAN 复制功能中的 VLAN 分为组播 VLAN 和用户 VLAN。

- 组播 VLAN 是与组播源相连的接口所属的 VLAN，用于实现组播流的汇聚。
- 用户 VLAN 是组播组成员主机所属的 VLAN，用于接收组播 VLAN 的数据流。
- 一个组播 VLAN 下可以绑定多个用户 VLAN。

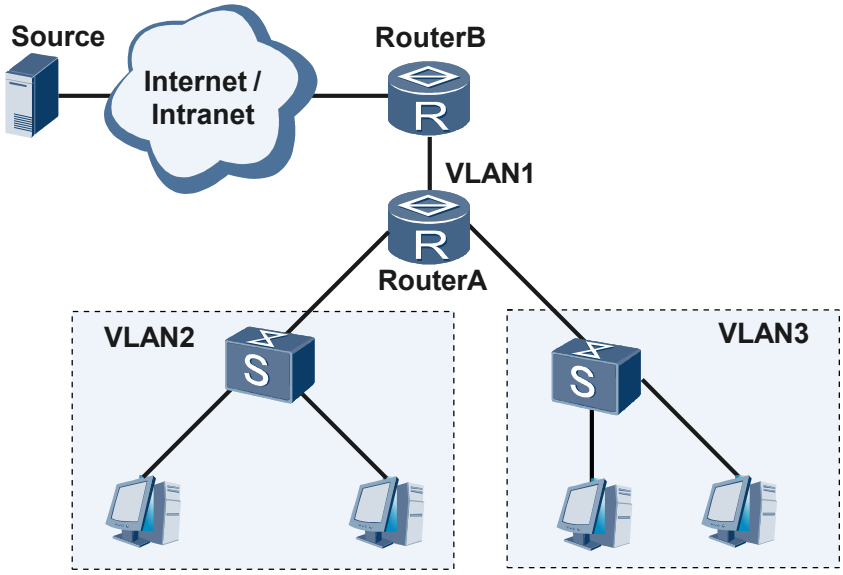
如图 4-6 所示，传统的组播点播方式下，属于不同 VLAN 的用户通过同一台设备 RouterA 进行同一组播源点播时，RouterB 需要为每个 VLAN 都复制一份组播数据，这样既造成了带宽浪费，同时也给上游设备增加了额外的负担。

图 4-6 传统组播点播组网图



如图 4-7 所示，不同 VLAN 的用户分别进行同一组播源点播时，可以在路由器 A 上配置组播 VLAN 复制，并将用户 VLAN 加入组播 VLAN，以实现组播数据在不同的 VLAN 内传送，便于对组播源和组播组成员的管理和控制，同时也可以减少带宽浪费。

图 4-7 组播 VLAN 复制组网图



路由器 B 与组播源侧相连，路由器 A 上运行组播 VLAN 复制功能，VLAN1 为组播 VLAN，VLAN2 和 VLAN3 为用户 VLAN。这样，在路由器 B 收到组播数据报文时只需把组播数据传送给该组播 VLAN 即可，而不必再为每个用户 VLAN 都复制一份。当路由器 A从上游接收到组播数据报文时，依据报文所属组播 VLAN 和报文的目的地址（即组播组地址）查询组播转发表。如果查询到对应的转发表项，则可以找到出端口和出端口 VLAN ID，然后将数据报文在每个出端口复制一份发送到用户 VLAN 中去；如果查询不到对应的转发表项，则将数据报文在本组播 VLAN 内广播。

4.3.9 组播 VLAN 1 ＋ 1 保护

为了提高组播业务的可靠性，在 VRP 组播 VLAN 复制功能的基础上实现了组播 VLAN 1 ＋ 1 保护功能。

组播 VLAN 1 ＋ 1 保护功能通过工作 VLAN 和保护 VLAN 实现组播流的 1 ＋ 1 备份。工作 VLAN 和保护 VLAN 接收到两份完全相同的组播数据，路由器选择工作 VLAN 还是保护 VLAN 的组播流是在组播流的入端口通过以太网 OAM 故障检测机制决定的。将 CCM 检测报文和两个 VLAN 绑定在一起，CCM 报文正常就说明 VLAN 中的组播流正常，否则组播流异常。

正常情况下，工作 VLAN 的组播流是有效的，保护 VLAN 的组播流被直接丢弃。当工作 VLAN 出现故障时，以太网 OAM 可以自动检测到链路变化，进行保护组的倒换，以保证组播数据流的正常转发。

组播 VLAN 1 ＋ 1 保护的对象是组播源，工作 VLAN 和保护 VLAN 应该接受到两份完全相同的组播数据。这是通过组网来保证的。可以在上游路由器上配置跨 VLAN 复制来保证在下游收到两分相同的组播数据流。保护组的倒换条件可以是手工倒换命令或者是以太网 OAM 检测到链路变化后自动倒换。

4.4 应用

4.5 术语与缩略语

术语/缩略语

缩略语	英文全称	中文全称
VLAN	Virtual Local Area Network	虚拟局域网
IGMP	Internet Group Management Protocol	Internet 组管理协议
PIM	Protocol Independent Multicast	协议无关组播

5 MSDP

关于本章

5.1 介绍

定义

MSDP（Multicast Source Discovery Protocol，组播源发现协议）是基于多个 PIM-SM（Protocol Independent Multicast Sparse Mode）域互连而开发的一种域间组播解决方案，目前只支持 IPv4。

目的

多个 PIM-SM 路由器相连组成的网络称为 PIM-SM 网络。一个大的 PIM-SM 网络可以由多个 ISP（Internet Service Provider）联合维护。

PIM-SM 域间 RP 信息隔离，组播源只向本域内的 RP 注册，用户主机只向本域内的 RP 发起加入。由于不同 PIM-SM 域的 RP 之间无法通信，所以 RP 知道且仅知道本域内的组播源，只能将本域内的组播源发出的数据分发给本地用户。

PIM-SM 网络依靠 RP（Rendezvous Point）实现组播转发。将一个大的 PIM-SM 网络划分为多个区域，每个区域维护一个 RP，可以实现 RP 负荷分担、增强网络的稳定性，益于管理。每一个这样的区域，称为一个 PIM-SM 域。

因此，将一个大的 PIM-SM 网络划分为多个 PIM-SM 域后，针对如何实现 PIM-SM 域间组播，使本 PIM-SM 域内的用户主机能够接收到其它域内组播源发出的组播数据，产生了 MSDP，使不同 PIM-SM 域的 RP 之间能够互相通信，共享组播源信息。

说明

本节中描述的 PIM-SM 域指某个 RP 的服务范围，可以是通过 BSR 边界划分出来的域，也可以是通过在不同路由器上配置不同的静态 RP 形成的域。

5.2 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
RFC3618	Multicast Source Discovery Protocol	-
RFC3446	Anycast Rendevous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)	-

5.3 原理描述

5.3.1 MSDP 实现域间组播

MSDP 对等体

MSDP 实现域间组播解决了不同 PIM-SM 域之间 RP 信息隔离、需要共享不同域内组播源信息的问题，实现了 PIM-SM 域间的 RP 之间的通信，共享组播源信息，保证域间组播业务正常运行。

按如下两种关系配置对等体关系：

- 在属于同一 AS，但属于不同 PIM-SM 域的 RP 之间建立 MSDP 对等体。
- 在跨 AS 的 RP 之间建立 MSDP 对等体。并采用 MBGP，将 MBGP 对等体和 MSDP 对等体建立在相同的接口上。

  说明

  有关 MBGP 的更多介绍，请参见《HUAWEI NetEngine20E-X6 高端业务路由器 配置指南-IP 组播》中的“MBGP 配置”。

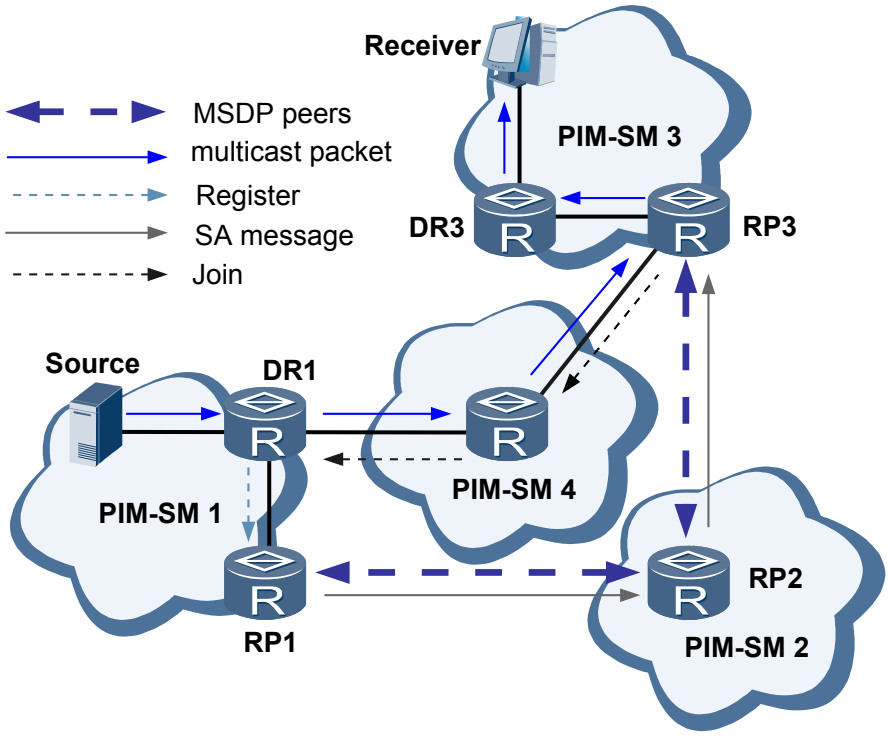
基本原理

通过配置 MSDP Peer 使各个 PIM-SM 域的 RP 之间建立 MSDP 对等体关系，各 MSDP Peer 之间彼此首尾相连，形成一张“MSDP 连通图”，连接各 PIM-SM 域的 RP。

MSDP 对等体之间交互 SA（Source Active）消息，SA 消息中携带组播源 DR 在 RP 上注册时的（S，G）信息。通过这些 MSDP 对等体之间的信息传递，任意一个 RP 发出的 SA 消息能够发送到其他所有的 RP。

如图 5-1 所示，PIM-SM 网络被划分为 4 个 PIM-SM 域。PIM-SM1 域内的组播源 Source 向组 G 发送数据。PIM-SM3 域内的 Receiver 为组 G 成员，RP3 和 Receiver 之间维护了一棵关于组 G 的 RPT（RP-rooted Shared Tree）。

图 5-1 MSDP 实现域间组播



如图 5-1 所示，通过在 RP1、RP2 和 RP3 之间建立 MSDP 对等体关系，可以使 Receiver 接收到 Source 发出的组播数据，具体过程如下：

1. Source 向组 G 发送组播数据。DR1（Designated Router）将组播数据封装在 Register 消息中，发给 RP1。RP1 作为源端 RP，创建 SA 消息，携带 Source 的 IP 地址、组 G 地址和 RP1 地址，发送给对等体 RP2。
2. RP2 接收到该 SA 消息后，执行 RPF（Reverse Path Forwarding）检查。检查通过，向 RP3 转发。
3. RP3 接收到该 SA 消息后，执行 RPF 检查，检查通过。由于 RP3 上存在（*，G）表项，表示本域内存在组 G 成员。
4. RP3 创建（S，G）表项，向 Source 逐跳发送（S，G）加入消息，创建一条从 Source 到 RP3 的组播路径（源树）。组播数据沿源树到达 RP3 后，再沿 RPT 向接收者转发。
5. 接收者接收到组播数据后，自行决定是否发起 SPT 切换。

5.3.2 MSDP 实现 Anycast RP

应用场景

在传统的 PIM-SM 域中，每个组播组只能映射到一个 RP。当网络负载较大或者流量过于集中时，可能导致 RP 路由器的压力过大、RP 失效后收敛较慢、组播转发路径非最优等问题。

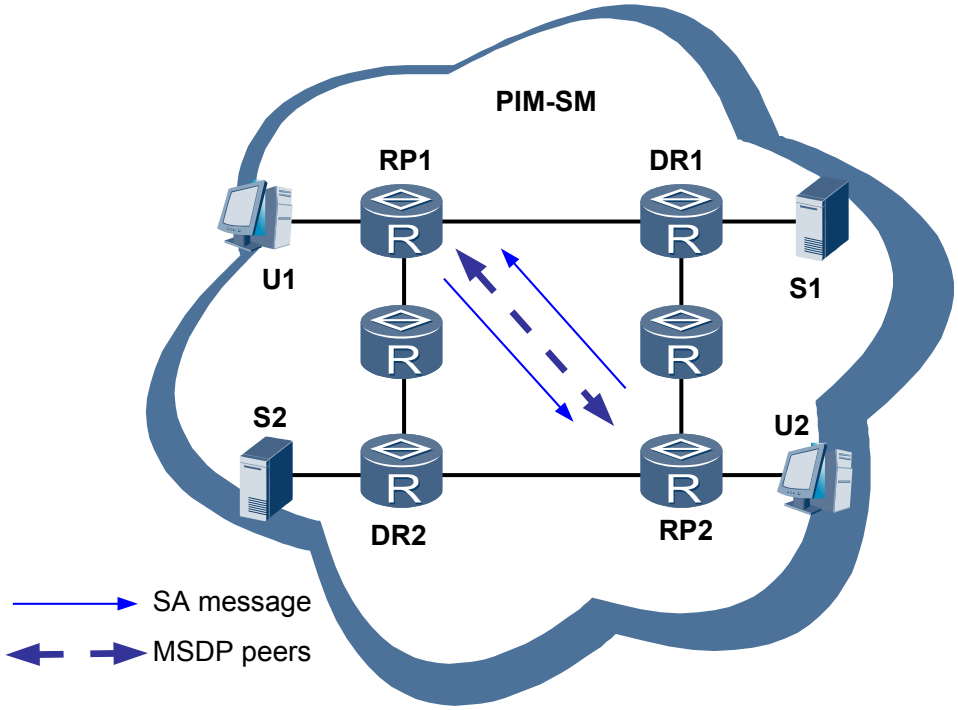
MSDP 实现 Anycast RP 是指在同一 PIM-SM 域内配置多个具有相同 IP 地址的 RP，这些相同的 IP 地址都配置在 loopback 接口上，且在这些 RP 之间建立 MSDP 对等体关系，从而实现 RP 路径最优及负荷分担。

在 PIM-SM 域内应用 Anycast RP 既可以解决组播源信息和组播组加入信息都需要向同一 RP 汇聚，导致单 RP 负荷重的问题。同时，接收者和组播源分别选择最近的 RP 发送加入和进行注册，保证了 RP 路径最优。

实现原理

如图 5-2 所示，在 PIM-SM 域内，组播源 S1 和 S2 向组播组 G 发送组播数据，U1 和 U2 是组播组 G 的成员。

图 5-2 Anycast RP 典型组网图



在 PIM-SM 域内应用 Anycast RP 的实现过程如下：

1. 在 RP1 和 RP2 两个路由器之间建立 MSDP 对等体关系，通过 MSDP 对等体进行域内组播。
2. 接收者选择距离最近的 RP 发送加入消息以构建 RPT 树。组播源选择距离最近的 RP 进行注册，RP 之间通过 MSDP 交互 SA 消息，共享组播源信息。
3. RP 加入以源端 DR 为根的 SPT，接收组播数据并转发，接收者接收到组播数据后，自行决定是否发起 SPT 切换。

5.3.3 多实例的 MSDP

VPN 实例支持 MSDP，属于同一实例（包括公网实例和 VPN 实例）的组播路由器接口之间可以建立 MSDP Peer。通过在 MSDP Peer 之间交互 SA 信息，实现跨域 VPN 组播。

应用多实例的组播路由器，为其支持的每一个实例独立维护一套 MSDP 机制，包括：SA 缓存、Peer 连接、定时器、发送缓存和 PIM 交互的缓冲区。同时，保证不同实例之间信息隔离。因此，只有属于同一实例的 MSDP 和 PIM-SM 信息才可以交互。

5.3.4 MSDP 支持 MD5/Key-chain 认证

MSDP 支持 MD5 或 Key-chain 认证，用于提高 MSDP 报文转发的安全性和可靠性，其应用场景同 MSDP 基本应用场景。当前支持 MD5 和 Key-chain 两种加密方式，MSDP peer 之间同时只能选择 MD5 和 Key-chain 两种加密方式之一，二者在功能上互斥。

Key-chain 能够集中的为所有应用提供加密认证功能，支持多种加密算法，且支持动态更新加密算法的 key 值。关于 Key-chain 的详细介绍，请参考 Key-chain 的特性描述文档。

5.3.5 SA 消息的 RPF 检查规则

为了防止 SA 消息在 MSDP Peer 之间被循环转发，MSDP 对接收到的 SA 消息执行 RPF 检查，在消息传递的入方向上进行严格的控制。不符合 RPF 规则的 SA 消息，将被丢弃。

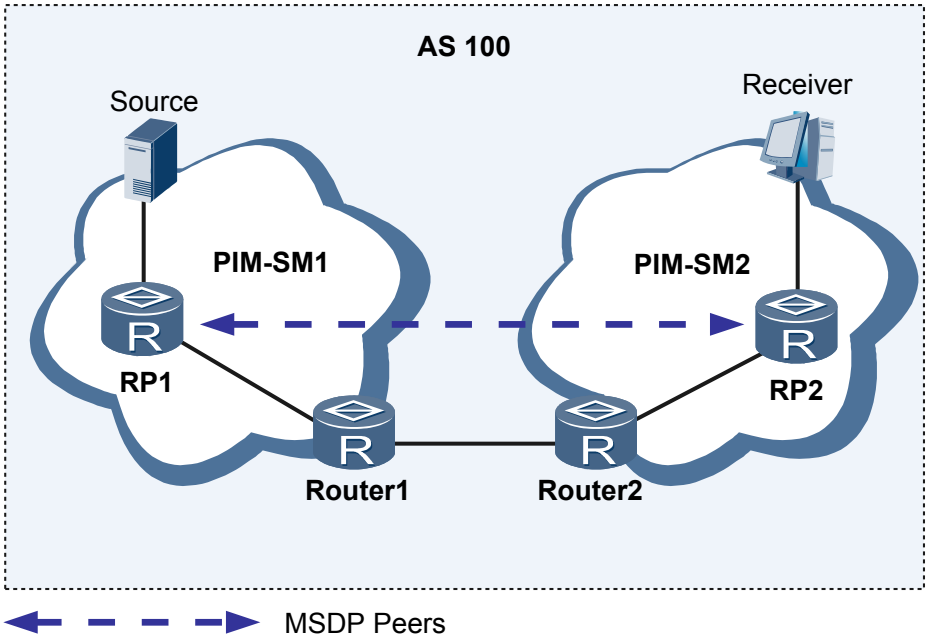
SA 消息的 RPF 规则主要有以下 6 点：

- 规则 1：发出 SA 消息的 Peer 就是源 RP（即创建该 SA 消息的 RP），则接受该 SA 消息并向其他对等体转发。
- 规则 2：接收从静态 RPF 对等体到来的 SA 消息。一台设备可以同时与多个设备建立 MSDP 对等体关系。用户可以从这些远端对等体中选取一个或多个，配置为静态 RPF 对等体。
- 规则 3：如果一台设备只拥有一个远端 MSDP 对等体，则该远端对等体自动成为 RPF 对等体，设备接受从该远端对等体发来的 SA 消息。如果 PIM-SM 域只存在一个域外远端 MSDP 对等体时，该域被称为 STUB 域。
- 规则 4：发出 SA 消息的 Peer 与本地设备属于同一 Mesh Group，则接受该 SA 消息。来自 Mesh group 的 SA 消息不再向属于该 Mesh group 的成员转发，但向该 mesh group 之外的所有对等体转发。
- 规则 5：发出 SA 消息的 Peer 是到源 RP 的“路由”下一跳或“路由”转发者，则接受该 SA 消息并向其他对等体转发。“路由”包括：MBGP、组播静态路由、单播路由（包括 BGP、IGP）。
- 规则 6：到达源 RP 的路由需要跨越多个 AS 时，接收从 AS-path（以 AS 为单位）中的 Peer 发出的 SA 消息。

5.4 应用

域间组播

图 5-3 AS 内 PIM-SM 域间组播

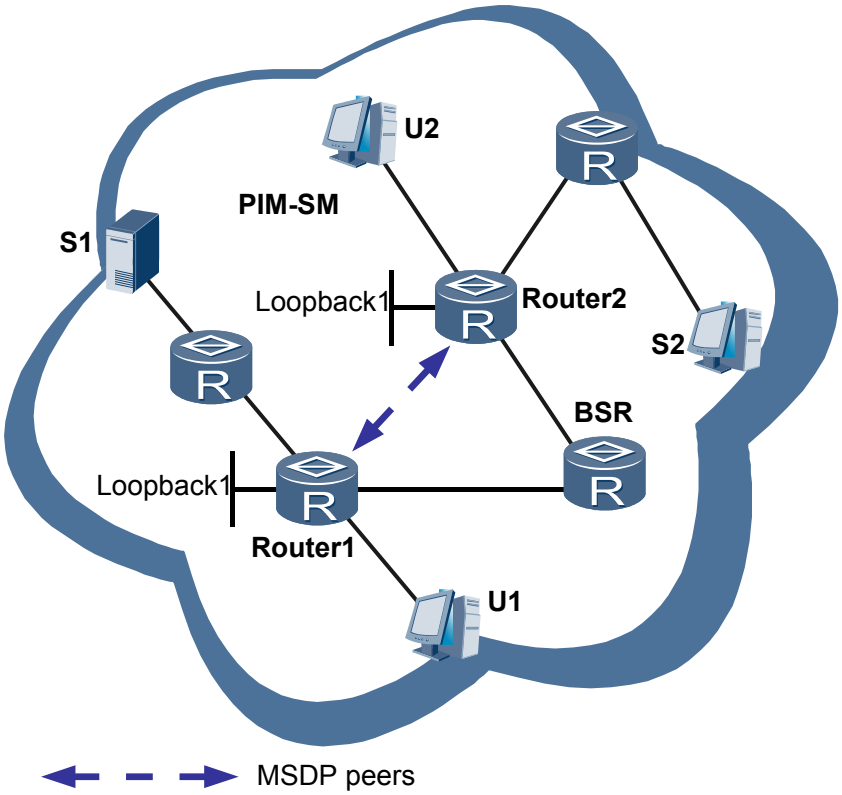


如图 5-3 所示：

- 在两个 PIM-SM 域的 RP 之间配置 MSDP 对等体,共享两个域内的组播源信息。
- 组播数据到达组播源侧 RP1，通过 SA 消息将对应的组播源信息发送给 RP2。
- RP2 将该组播数据转发给本域内的接收者。
- 接收者接收到组播数据后，自行决定是否发起 SPT 切换。

Anycast RP

图 5-4 Anycast RP 应用



如图 5-4 所示:

- Router1 和 Router2 作为 RP，两者之间建立 MSDP 对等体关系。
- 借助 MSDP 对等体进行域内组播，接收者选择距离最近的 RP 发送加入消息以构建 RPT 树。
- 组播源选择距离最近的 RP 进行注册，RP 之间交互 SA 消息，共享源信息。
- RP 加入以源端 DR 为根的 SPT，引入组播数据。
- 接收者接收到组播数据后，自行决定是否发起 SPT 切换。

5.5 术语与缩略语

术语

术语	解释
MSDP	Multicast Source Discovery Protocol，称为组播源发现协议。只适用于 PIM-SM 域，仅对 ASM（Any-Source Multicast）模型有意义。 通过在不同 PIM-SM 域的 RP 之间建立 MSDP 对等体关系，在域间共享组播源信息，实现跨域组播。 通过在同一 PIM-SM 域的多个 RP 之间建立 MSDP 对等体关系，在域内共享组播源信息，实现 Anycast RP。
PIM	Protocol Independent Multicast，称为协议无关组播，属于组播路由协议。网络中单播路由畅通是 PIM 转发的基础。PIM 利用现有的单播路由信息，对组播报文执行 RPF 检查，从而创建组播路由表项，构建组播分发树。
SA	Source Active，MSDP 消息类型。SA 消息中包含多组（S，G）信息，或封装一个 Register 消息。MSDP 对等体之间通过交互 SA 消息，共享组播源信息。
SPT	Shortest Path Tree，称为最短路径树。以组播源为根，组播组成员为叶子的组播分发树称为 SPT。SPT 同时适用于 PIM-DM、PIM-SM 和 PIM SSM。
BSR	BootStrap Router，称为自举路由器。是 PIM-SM 网络的管理核心。BSR 负责收集网络中的 C-RP 信息，汇集成为 RP-Set，封装在 Bootstrap 消息中，发布给全网的每一台 PIM-SM 设备。各设备根据 RP-Set 计算出特定组播组对应的 RP。

缩略语

缩略语	英文全称	中文全称
AS	Autonomous System	自治系统
BGP	Border Gateway Protocol	边界网关协议
BSR	BootStrap Router	自举路由器
MSDP	Multicast Source Discovery Protocol	组播源发现协议
PIM-SM	Protocol Independent Multicast Sparse Mode	协议无关组播－稀疏模式
RP	Rendezvous Point	汇聚点

6 组播管理

关于本章

6.1 介绍

定义

随着 Internet 网络的不断发展，网络中交互的各种数据、语音和视频信息越来越多，组播业务随之快速发展起来，组播管理是对组播业务探测、故障诊断等工具的管理，特性列表如下：

- 组播 Ping（Multicast Ping，以下简称 MPing）是一种组播业务的探测工具，通过发送 ICMP Echo Request 报文促进组播转发树的建立和检测网络中的保留组成员。

  说明

  保留组：保留的本地组播组地址 224.0.0.0-224.0.0.255 网段。例如：224.0.0.5 是 OSPF 协议组地址，224.0.0.13 是 PIMv2 协议组地址。

- 组播 Tracert（Multicast trace route，以下简称 MTrace），是一种组播路径的追踪工具，可以追踪某一接收者沿着组播转发树到组播源的路径。

目的

在组播应用日益增加的今天，缺乏组播 Ping 和 Tracert，已经不能满足组播业务的维护和故障定位的需要。在选择支持组播的网络设备时，用户不仅仅要求设备支持组播转发和组播路由协议，还要求支持组播故障诊断工具。伴随着组播业务的开展，组播维护和故障定位自然成为必要的需求。

MPing 主要有以下几种用途：

- 发起普通组播组的 MPing
- 通过查看组播设备上的组播路由表信息，检查协议运行状态是否正常，确认组播分发树是否正确建立。
- 通过对目的主机反馈的 ICMP Echo Reply 报文进行统计处理，计算从 MPing 发起者到组播组成员的 TTL、响应时间等。
- 按照一定时间间隔连续执行多次 MPing，计算网络时延和路由抖动。
- 发起保留组播组的 MPing
- 检测网络中的保留组成员。

MTrace 主要有以下几种用途：

- 在组播故障处理和日常维护中使用 MTrace，有助于定位故障结点和检测配置错误。
- 追踪报文实际转发路径，在追踪过程中收集流量信息。循环执行追踪过程，可以统计组播流速率。
- 网管通过分析 MTrace 输出的异常节点信息，产生告警信息。

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
draft-sarac-mPing-00.txt	This document describes a mechanism for discovering multicast reachability between end systems within/between multicast enabled networks. It uses request/response messages to verify multicast reachability between the local site and a remote site. With this utility, multicast users can test if they can successfully join a multicast group of a remote source and receiver its data.	-
draft-fenner-traceroute-ipm-01	This draft describes the IGMP multicast traceroute facility. Unlike unicast traceroute, multicast traceroute requires a special packet type and implementation on the part of routers. This speci-fication describes the required functionality in multicast routers, as well as how management applications can use the new router func-tionality.	-

6.3 原理描述

6.3.1 MPing

MPing 使用标准的 ICMP 消息，查询设备（用户发起检测命令的设备）构造 ICMP Echo Request 报文，其内部封装的 IP 报文的目的地址为组播地址，包括保留组地址和普通组地址。

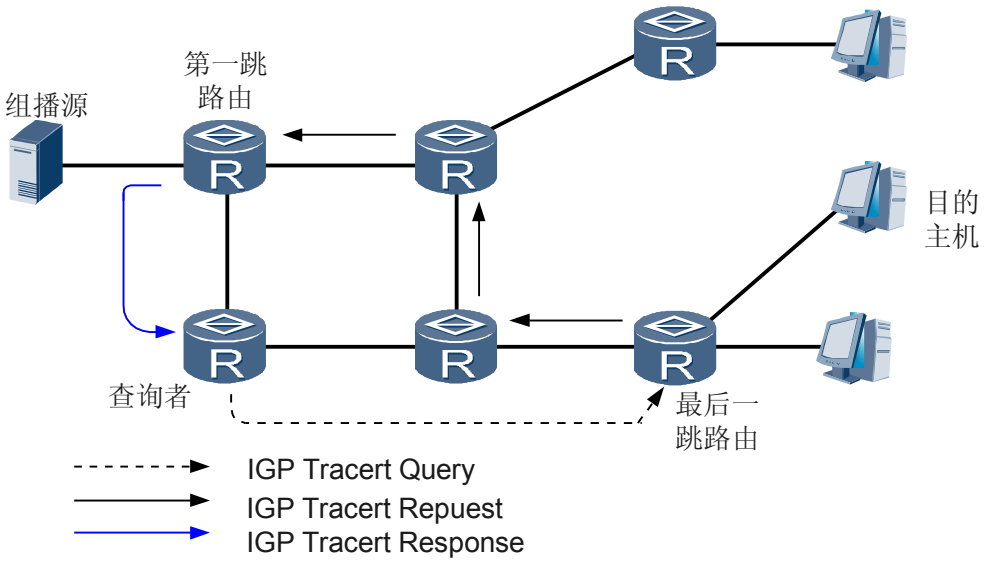
- 目的地址为保留组地址时，查询设备必须指定 ICMP Echo Request 报文的出接口，保留组成员（组播设备）收到目的组地址为保留组地址的 ICMP Echo Request 消息后，回应 ICMP Echo Reply 消息。因此，可以用来检测网络中的保留组成员。
- 目的地址为普通组地址时，查询设备不能指定 ICMP Echo Request 报文的出接口。ICMP Echo Request 报文作为有限的组播流量，在组播网络中正常转发，触发组播路由的建立。通过 NQA 调度多次 MPing 组播组，可以得到时延抖动数据以实现组播业务的正常维护和故障定位。

6.3.2 MTrace

MTrace 遵循 IETF 的协议标准 draft-fenner-traceroute-ipm-01.txt。

该标准描述了一种实现跟踪组播数据从组播源到特定目的接收者所经过的网络路径的机制。

图 6-1 MTrace 组网图



MTrace 基于的网络必须使能了组播协议，例如 PIM 协议（PIM-DM、PIM-SM），并建立了组播转发树。MTrace 通过发送查询报文探测组播路径。报文分为三类：IGMP Tracert Query、IGMP Tracert Request 和 IGMP Tracert Response。

- IGMP Tracert Request 消息完全继承了 IGMP Tracert Query 消息，并在消息尾部添加了 Response 数据块。
- IGMP Tracert Response 消息完全继承了 IGMP Tracert Request 消息，只修改了消息类型。

实现原理如下：

1. 在查询设备上（如图查询者）输入 MTrace 命令，指定源地址、目的主机地址、组播组。
2. 查询设备向目的主机所连接的最后一跳设备发送 IGMP Tracert Query 报文。
3. 最后一跳设备收到该报文后，增加包含本设备接口地址信息的响应数据块，然后沿着到组播源的逆向组播路径向上一跳设备发送 IGMP Tracert Request 报文。
4. 每一跳设备都会增加本跳的响应数据块并向组播源方向转发 IGMP Tracert Request 报文。
5. 当连接组播源的第一跳组播设备收到 IGMP Tracert Request 报文时，添加本设备响应数据块，然后向查询设备发送 IGMP Tracert Response 报文。
6. 在查询设备上通过解析 Response 报文获取转发的路径信息，显示从组播源到目的主机的组播路径信息。
7. 如果因为某种错误原因，请求消息在转发时无法到达第一跳组播设备，则直接向查询设备返回 Response 报文，根据解析数据块信息，了解故障节点，从而实现故障点监测的目的。

   MTrace 发起方式：

   MTrace 支持 4 种封装方式（也称为 MTrace 的发起方式），适用于不同的网络环境。

- all-router：当前组播设备与目的主机直接相连（但不是最后一跳设备），使用224.0.0.2 为报文目的地址，目的主机网段接口地址为报文源地址。该报文能够被连接在目的主机网段上的所有组播设备（包括最后一跳设备）接收。
- last-hop：使用最后一跳组播设备地址为报文目的地址。这种方法要求用户输入最后一跳设备地址。
- destination：使用目的主机地址为报文的目的地址。当与用户主机相连的组播设备接收到该报文时，判断自己是不是最后一跳设备，如果不是，则使用 all-router 方式重新封装 IGMP Tracert Query 消息。
- multicast-tree：查询设备正好位于从组播源到目的主机的组播路径上（比如第一跳组播设备），使用被追踪的组地址为报文的目的地址，组播源地址为报文的源地址。该报文沿组播路径下发，到达最后一跳组播设备。

6.4 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
MPing	Multicast Ping	组播 Ping
MTrace	Multicast trace route	组播 trace route
NQA	Network Quality Analysis	网络质量分析

7 组播路由管理

关于本章

7.1 介绍

定义

组播路由管理（Multicast Route Management）用于管理组播路由表，能够控制组播路由创建或改变组播路由。

组播路由管理包括以下 6 个功能：

- RPF 单播逆向路由检查
- 组播负载分担
- 按照最长匹配选择路由
- 指定组播转发边界(multicast boundary)
- 组播多拓扑
- 组播 NSR

目的

- RPF 单播逆向路由检查功能

  此功能用于查找到组播源的最优单播路由，创建组播转发树。单播路由的出接口作为转发表项中的数据入口。当转发模块收到组播数据时，在匹配转发表项的同时，还会匹配数据的入接口是否正确。如果组播数据报文实际到达接口和单播路由的出接口相同，则 RPF 检查通过；否则 RPF 检查失败，丢弃该报文。RPF 单播逆向路由检查功能能够防止组播数据在转发过程中出现流量环路。

- 组播负载分担

  在组播选路时，使用组播负载分担策略可对不同转发表项从多条等价路由中选取不同的等价路由作为 RPF 路由，指导数据转发。由于转发表项依赖的 RPF 路由能够分布在多条等价路由上，因此此功能可以达到组播数据分流的效果。

- 按照最长匹配选择路由功能

  在组播选路时，能够优先选取掩码长度最长的路由，以实现路由的精确匹配。

- 指定组播转发边界功能

  在接口上配置组播转发边界可阻塞相应的组播数据，使组播数据流无法从该接口转发。

- 组播多拓扑

  组播多拓扑是为了解决组播业务严重依赖单播，以及与 MPLS TE Tunnel 共同部署的问题，在物理网络上为组播业务单独规划出一张拓扑。组播进行 RPF 检查时，只在指定拓扑下查找路由，从而在该拓扑中建立组播转发树转发数据，实现网络资源的隔离。

- 组播 NSR

  使用组播 NSR 特性可使设备在进行主备倒换后，周边设备不感知，组播路由不间断，不会引起组播转发树的变化，以及触发周边设备的响应处理。

7.2 参考标准和协议

本特性的参考资料清单如下：

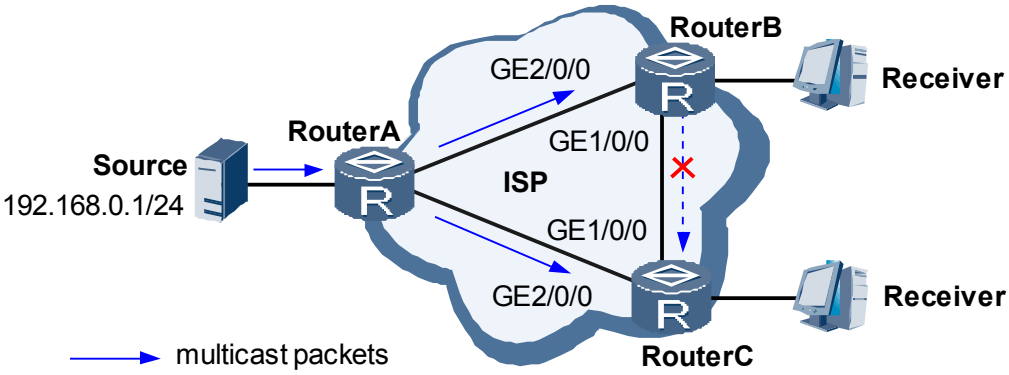
文档	描述	备注
RPF4601	Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)	-

7.3 原理描述

7.3.1 RPF 单播逆向路由检查

RPF 检查的规则是：依据“报文源”，查找单播路由表、MBGP 路由表、MIGP 路由表和组播静态路由表，从这些路由表中选出一条最优路由，作为 RPF 路由。如果报文实际到达接口与 RPF 接口相同，则 RPF 检查通过；否则 RPF 检查失败。

图 7-1 RPF 检查过程



IP Routing Table on RouterC	
Destination/Mask	Interface
192.168.0.0/24	GE2/0/0

如图 7-1 所示，组播报文从 GE1/0/0 到达 RouterC，RouterC 对数据报文进行 RPF 检查，发现数据到达接口与转发表项入接口不符，则 RPF 检查失败。如图 7-1 中的路由表，发现到达 Source 的最短路径出接口是 GE2/0/0，与（S，G）表项的入接口相同。于是判定当前（S，G）表项正确，该报文从错误的路径而来，丢弃该报文。

7.3.2 组播负载分担

组播负载分担共支持 5 种策略：

- 基于组播组的负载分担

  基于组播组的负载分担策略主要应用在网络中存在大量不同的组播组的场景。
- 基于组播源的负载分担

  基于组播源的负载分担策略主要应用在网络中存在大量不同的组播源的场景。
- 基于组播源组的负载分担

  基于组播源组的负载分担策略应用在网络中既存在大量不同的组播组，又存在大量不同的组播源的场景。
- 稳定优先负载分担

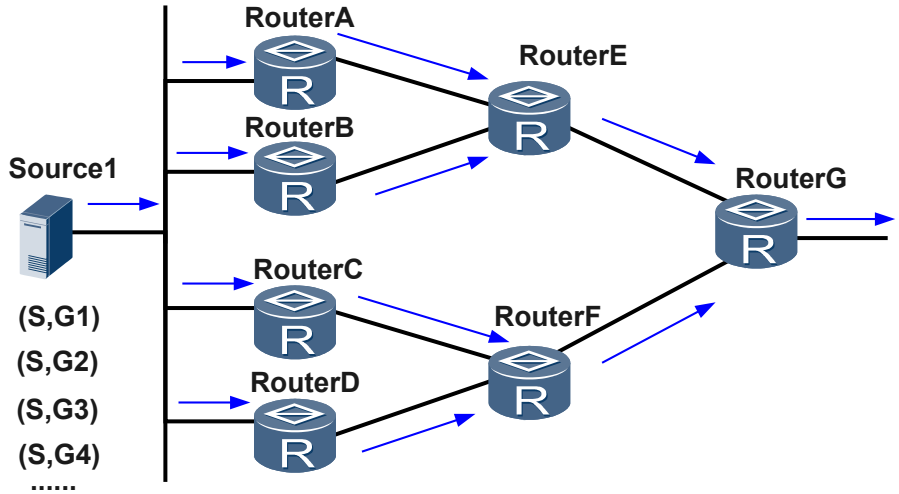
  稳定优先负载分担策略可以应用于上述三种负载分担场景，还可以应用于共享网段的应用场景。
- 均衡优先负载分担

  均衡优先负载分担的应用场景同稳定优先。

基于组播组的负载分担

如图 7-2 所示：

图 7-2 基于组播组的负载分担

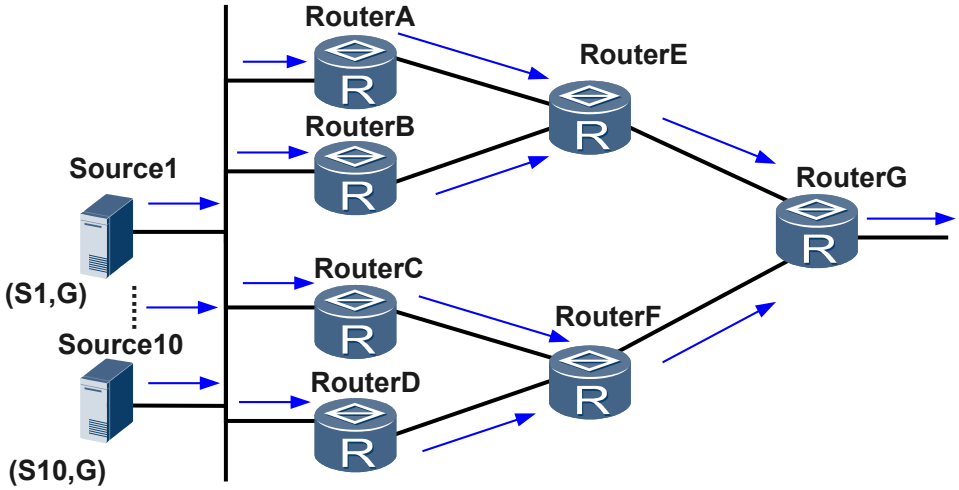


组播路由器根据组播组 G 的不同，经过一系列算法，为每个组播组 G 从多条等价路由中选取一条合适的路由，作为该组播组 G 的转发路由，最终达到不同的转发路径上的流量属于不同的组播组集合。

基于组播源的负载分担

如图 7-3 所示：

图 7-3 基于组播源的负载分担

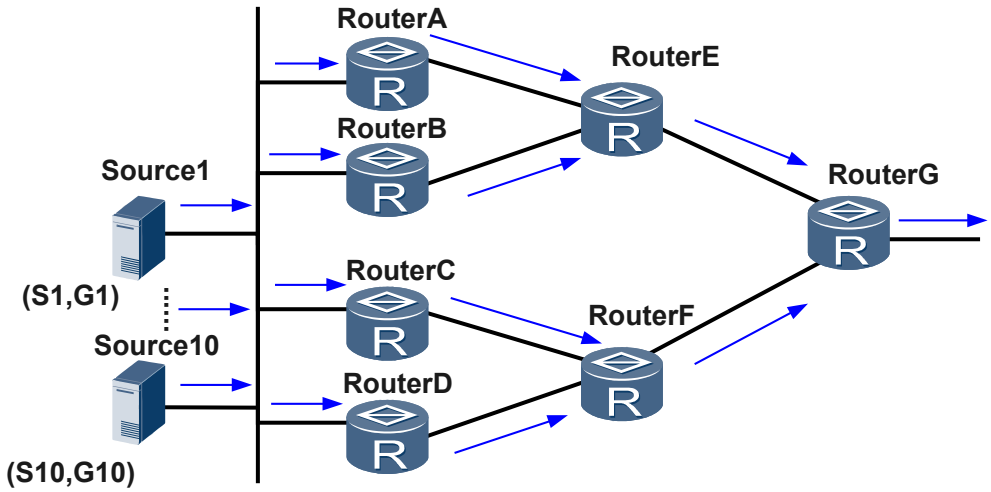


组播路由器根据组播源 S 的不同，经过一系列算法，为每个组播源 S 从多条等价路由中选取一条合适的路由，作为该组播源 S 的转发路由，最终达到不同的转发路径上的流量分属于不同的组播源集合。

基于组播源组的负载分担

如图 7-4 所示：

图 7-4 基于组播源组的负载分担



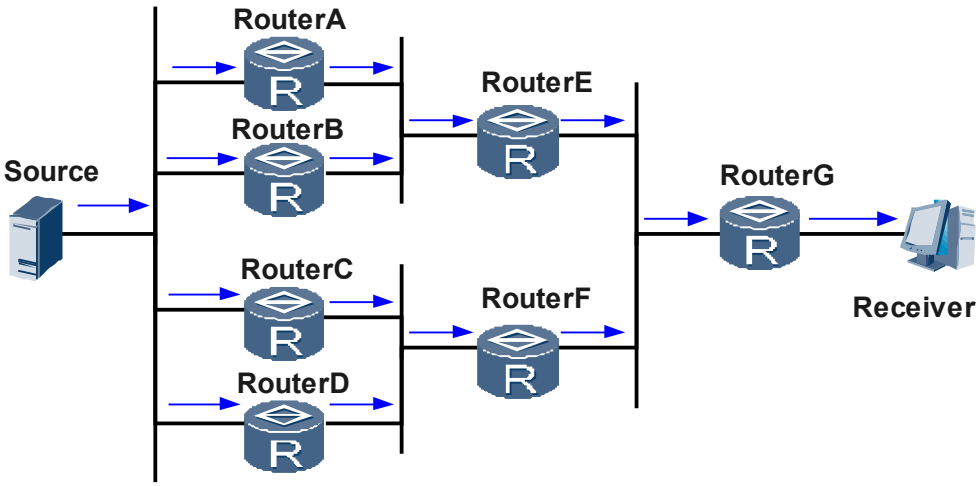
组播路由器根据组播组 G 和组播源 S 的不同，为每一个组播源组(S,G)经过一系列算法，从多条等价路由中选取一条合适的路由，作为该组播源组(S,G)的转发路由，最终达到分配到不同的转发路径上的流量分属于不同的组播源组集合。

稳定优先负载分担

- 应用场景

  该负载分担策略可以应用于上述三种负载分担场景，如图 7-2、图 7-3 和图 7-4 所示。负载分担还可以应用于共享网段的应用场景，如图 7-5 所示。

  图 7-5 稳定优先负载分担



- 实现原理

  在配置稳定优先负载分担的路由器上，对于新的表项加入，会选择最合适的路由，即当前依赖表项最少的路由。在网络拓扑稳定同时表项稳定的情况下，依赖同一个网段源的所有表项会均衡分布在各个等价路由之上。

  如果表项退出或者路由权值变化，导致负载不均衡，在配置稳定优先负载分担策略的情况下，会通过对后续新的加入选择最合适路由来慢慢“愈合”这种不均衡。

  在稳定优先负载分担场景中，若出现表项分布不均衡，经过一定的延迟时间，设备会对所有表项进行均衡调整，均衡调整的延迟时间用来防止频繁变化对设备的冲击。

  目前，可通过配置组播负载分担均衡调整定时器控制系统开始调整表项分布不均衡状态的延迟时间。

均衡优先负载分担

在配置均衡优先负载分担的路由器上，对于新的表项加入，会选择最合适的路由，即当前依赖表项最少的路由。均衡优先策略即不论在何种情况下，都要求依赖同一网段源的表项最终均衡的分布在其等价路由上，包括表项退出、权值变化和等价路由发生变化。在发生不均衡以后，均衡调整的动作将会有一定的延迟时间，用来防止频繁变化，同时还可以在延迟时间内，通过新加入表项来“愈合”不均衡。

目前，可通过配置组播负载分担均衡调整定时器控制系统开始调整表项分布不均衡状态的延迟时间。

不均衡负载分担

- 应用场景

  不均衡负载分担是对均衡优先负载分担和稳定优先负载分担的功能的一个补充，不会改变这两种策略的基本行为，只是会让分布在等价路由上的表项数呈现一种比例

关系。它的应用场景图同上述两种负载分担（除图 7-5 所示的共享网段场景）。两种典型应用场景为：

- 当几条等价路由之间转发能力存在较大差异，或者流量拥塞程度存在较大差异时，可以通过配置路由负载分担权值来调节等价路由上分配的表项，权值大的路由将会分配更多的表项。稳定优先负载分担策略，只会对后来的加入表项起到作用；均衡优先负载分担策略，则会对已经存在的表项按照权值的比例进行调整。
- 当路由器的某条等价路由路径上，有路由器需要进行版本升级时，可以通过在接口配置权值为 0，让流量从这条等价路由上转移到其它等价路由上（对均衡优先负载分担策略起作用）。

- 实现原理

  单播各条等价路由的转发能力、实际网络负载情况和链路的用途均可能存在差异，因此在特定场景下对组播表项还继续进行均衡负载，很难满足需求。不均衡负载分担是一种针对上述情况的解决策略，允许用户在接口上配置权值，权值越大的接口所在路由可以依赖的表项越多。

7.3.3 按照最长匹配选择路由

选路时，分别从域内单播路由表、域间单播路由表和组播静态路由表中各选出一条最优路由，并从中选择一条作为组播数据转发路径。

按照最长匹配选择路由时，组播路由器会优先选取目的地址与报文源地址匹配“掩码”最长的路由。若存在多条路由掩码匹配长度相同，则按照组播静态路由、域间单播路由、域内单播路由的顺序选择一条路由作为组播数据的转发路径。

例如：组播源地址为 10.1.1.1，地址为 192.168.1.1 的主机需要接收组播源的组播数据。查找组播静态路由表和域内单播路由表有两条可到达组播源的路由。到达的网段分别为：10.1.1.0/16 和 10.1.1.0/24。按照最长匹配选择路由，选择到达网段为 10.1.1.0/24 的路由作为组播数据的转发路径。若掩码匹配长度相同，则按照路由的优先级顺序选择一条最优路由作为组播数据的转发路径。

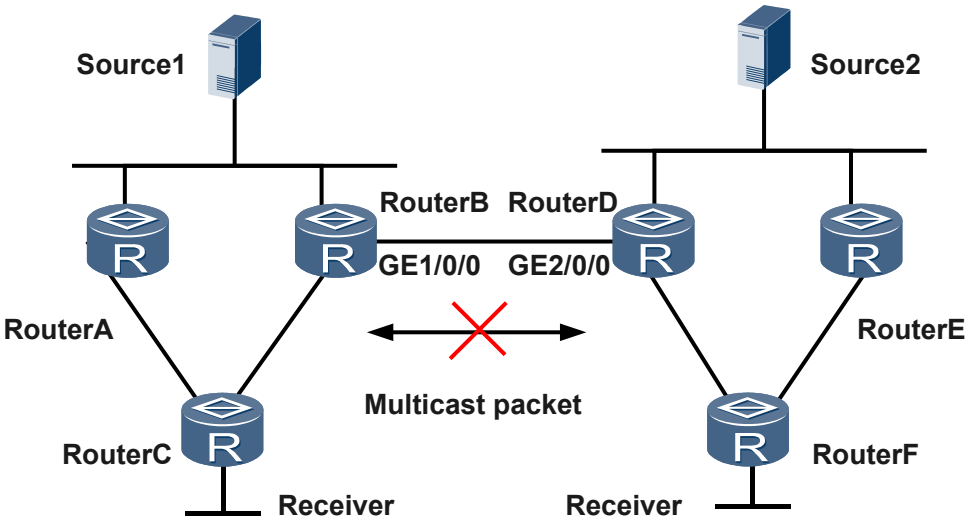
7.3.4 指定组播转发边界

应用场景

应用组播转发边界（multicast boundary）功能可以实现每个组播组对应的组播信息在一个确定的范围内传递。在接口上配置组播边界，以形成一个封闭的组播转发区域。当组播设备接口配置了针对某组播组的转发边界以后，该接口将不能再发出或接收对应的组播组的报文。

实现原理

图 7-6 指定组播转发边界组网图



如图 7-6 所示，RouterA，RouterB，RouterC 构成组播域 1，RouterD，RouterE，RouterF 构成组播域 2，两个组播域通过 RouterB 和 RouterD 进行互通。

如果要两个组播域将组播组 G 的数据隔离开，则只需在接口 GE1/0/0 或接口 GE2/0/0 上配置该组 G 为组播转发边界，对应的接口在配置组播边界后将不再转发和接收对应组播组 G 的报文。

7.4 术语与缩略语

术语

术语	解释
组播负载分担	组播负载分担不等同于负载均衡，负载分担是指组播表项能够分布在各条等价路由上，且每条等价路由上的组播表项数目可以不同。
MBGP 路由表	Multicast BGP，组播边界网关协议路由表。 MBGP 是 MP-BGP 在组播上的应用。MBGP 路由表是在 BGP 协议中定义一个地址族，用于发布组播使用的单播路由。
MIGP 路由表	Multicast IGP，组播内部网关协议路由表。 MIGP 路由表是对单播路由表中出接口为 Shortcut Tunnel 的路由，计算出不使用 shortcut tunnel 作为出接口的路由表。 若组播在单播路由表中选路时选中单向 TE Tunnel 接口，会在 MIGP 路由表中重新选取一条最优路由作为组播数据转发的路径。

术语	解释
组播多拓扑	组播多拓扑是为了解决组播业务严重依赖单播，以及与 MPLS TE Tunnel 共同部署的冲突问题，在物理网络上为组播业务单独规划出一张拓扑。组播进行 RPF 检查时，只在指定拓扑下查找路由，从而在该拓扑中建立组播分发树转发数据，实现与单播业务或 MPLS TE Tunnel 业务网络资源的隔离。

缩略语

缩略语	英文全称	中文全称
RPF	Reverse Path Forwarding	反向转发路径
NSR	Non-Stop Routing	不间断路由

8 组播 VPN

关于本章

8.1 介绍

定义

MVPN（Multicast VPN，组播 VPN）解决方案基于 draft-rosen-vpn-mcast 草案制定的 MD（Multicast Domains）方案，是为了在现有 MPLS/BGP VPN 上开通组播业务。

目的

实现 MVPN 的目的是：为了在现有 MPLS/BGP VPN 上开通组播业务，将私网 PIM 实例中的组播数据和控制报文透过公网传递到 VPN 的远端站点。

公网 PIM 实例不需要了解私网中传递的组播数据，私网 PIM 实例也不需要了解公网实例中的组播路由信息，各个私网 PIM 实例之间相互隔离。

8.2 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
draft-rosen-vpn-mcast-08	Multicast in MPLS/BGP IP VPNs	● Share-Group 不支持 SSM 范围。 ● 不同的 VPN 的 Share-Group 不能使用相同的组地址。
draft-ietf-l3vpn-2547bis-mcast-08	Multicast in MPLS/BGP IP VPNs	
draft-ietf-l3vpn-2547bis-mcast-07	Multicast in MPLS/BGP IP VPNs	
draft-rosen-vpn-mcast-12	Multicast in MPLS/BGP IP VPNs	

8.3 原理描述

8.3.1 MVPN 术语介绍

- MD

  Multicast Domains，表示各个 PE 上，能够互相发送和接收组播报文的所有 VPN 实例组成的集合。

- Share-Group

  根据 Multicast Domains 的原理，所有属于同一 MD 的 PE 上的 VPN 实例都要加入一个公共的组，称之为 Shared-Group。

  目前组播 VPN 的实现中，一个 VPN 实例只能配置一个 Share-Group，即一个 VPN 实例只能加入一个 MD。

- Share-Multicast Distribution Tree

  Share-MDT 是 Share-Multicast Distribution Tree 的简称。实际是由 PE 上的私网 PIM 实例加入到根据 Share-Group 建立的组播分发树，用于将 VPN 内的 PIM 协议报文和数据报文分发给其他同属于一个 VPN 的 PE，这个组播分发树称之为 Share-MDT。通常被称为 Multicast Tunnel，即 MD 的组播隧道。

- MTI

  Multicast Tunnel Interface，称为组播隧道接口。MTI 是 MT 的入/出口，相当于 MD 的入/出口。本地 PE 将私网数据从 MTI 发出，远端 PE 从 MTI 接收私网数据。

  PE 定义 MTI 调用整个 MT 传输过程。MTI 实际上是 PE 上公网实例和 VPN 实例进行交互的“通道”。PE 使用 MTI 连接到 MT 上，相当于被连接到了共享网段上。各个 PE 上属于该 MD 的 VPN 实例在 MTI 上建立 PIM 邻居关系。

- Switch-Group

  在建立起 Share-MDT 后，所有有私网接收者的 PE 为建立 Switch-MDT 而加入的组。

- Switch-MDT

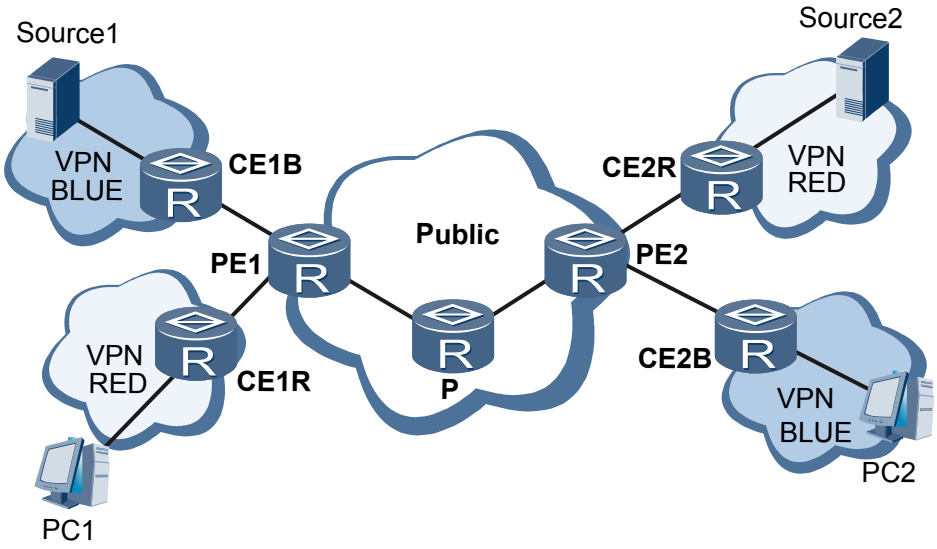
  Switch-MDT 是 Switch-Multicast Distribution Tree 的简称。为了避免数据流向不必要的 PE，在建立起 Share-MDT 后，所有有私网接收者的 PE 加入到一个用 Switch-Group 组建立起来的按需发送的组播分发树，用于将 VPN 的高速数据报文分发给其他同属于一个 VPN 的 PE。

8.3.2 MVPN 实现域间组播

应用 MVPN 方案需要在运营商（Service Provider）骨干网（核心网络或公网）中支持组播功能。

- PE 在 VPN 实例中运行的 PIM 实例称为私网 PIM 实例。
- PE 的公网部分运行的 PIM 实例称为公网 PIM 实例。

图 8-1 MVPN 应用组网图



应用 MVPN 方案实现 PE 上的私网 PIM 实例互相访问的过程是：

1. 在 PE 私网 PIM 实例之间建立一个虚拟的组播隧道 MT（Multicast Tunnel）。
2. 私网 PIM 实例创建一个组播隧道接口 MTI（Multicast Tunnel Interface）与该组播隧道相连。
3. 各 VPN 私网实例根据自己所配置的 Share-Group 加入各自的组播隧道。

这样，配置了相同的 Share-Group 组地址的私网实例就形成了一个 MD。

图 8-1 中，分别与 PE1、PE2 相连的两个私网实例 VPN BLUE、VPN RED 通过对应的 MD BLUE、MD RED 实现了各自实例的互通，如图 8-2 和图 8-3 所示。

图 8-2 基于 MD 的 VPN BLUE 组网

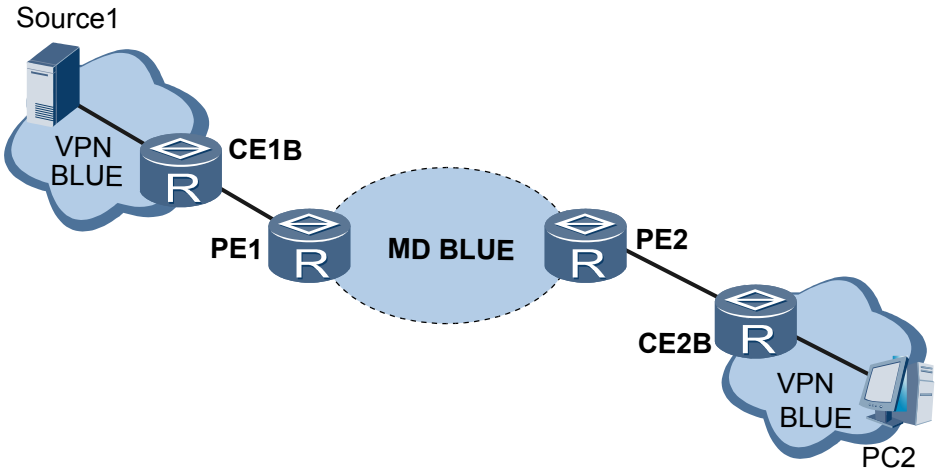
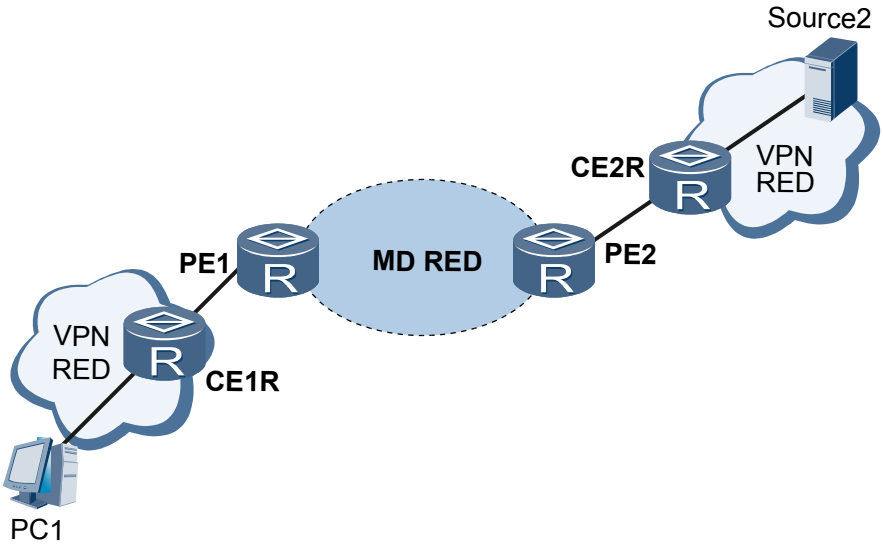


图 8-3 基于 MD 的 VPN RED 组网



PE 上的私网 PIM 实例将组播隧道接口当作一个 LAN 接口，私网 PIM 实例在组播隧道接口上与远端的私网 PIM 实例建立 PIM 邻居，进行 DR 选举，发送 Join/Prune 报文，并从该接口转发和接收组播数据。

私网 PIM 实例向组播隧道接口上发送 PIM 协议报文或者组播数据报文时，将对报文进行封装。封装后的报文是公网组播数据报文，交给公网 PIM 实例在公网中进行转发。可以看出，组播隧道实际上是公网中的组播分发树。

- 不同的 VPN 使用不用的组播隧道，不同的组播隧道使用不同的封装。这样就使得不同 VPN 之间的组播数据相互隔离。
- 同一 VPN 中的 PE 上的私网 PIM 实例使用相同的组播隧道，并通过该组播隧道相互连接。

说明

一个 VPN 唯一确定一个 MD；一个 MD 只能为一个 VPN 服务。这种关系称为一一对应。VPN、MD、MTI、Share-Group 地址和 Switch-group-pool 两两之间是一一对应的关系。

8.3.3 CE、PE 和 P 之间的 PIM 邻居关系

PIM 邻居关系建立在直接相连并且属于同一网段的两台或多台组播设备之间，在 MD VPN 中存在三种 PIM 邻居关系：PE-CE 邻居关系、PE-P 邻居关系、PE-PE 邻居关系。

在图 8-4 中，各个 PE 上的 VPNA 实例与属于 VPNA 的 site 共同实现 VPNA 组播，PIM 邻居关系如图 8-5 所示。

图 8-4 VPNA 组播

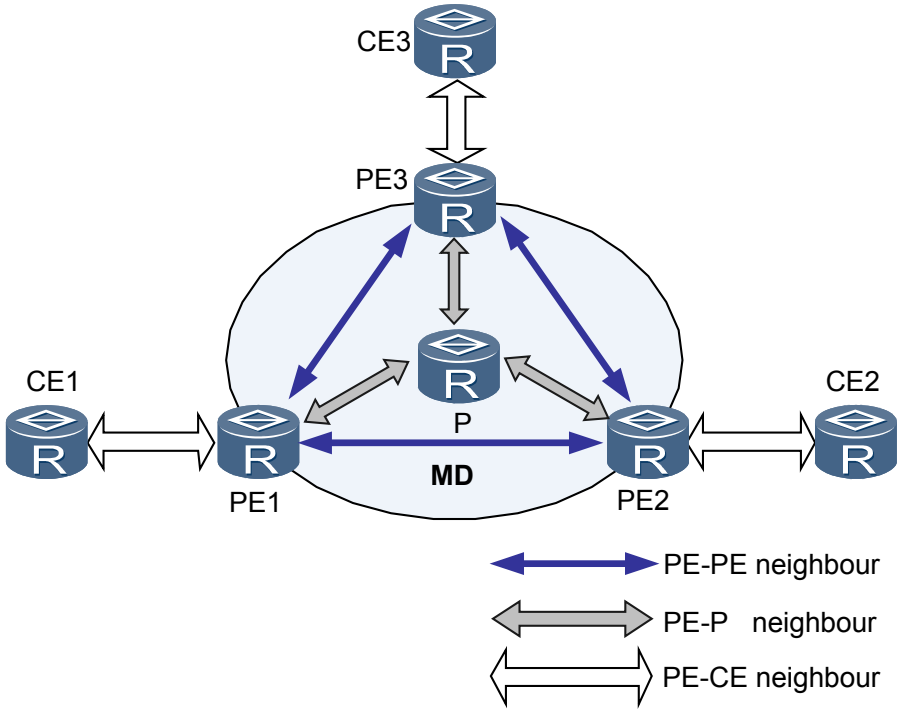
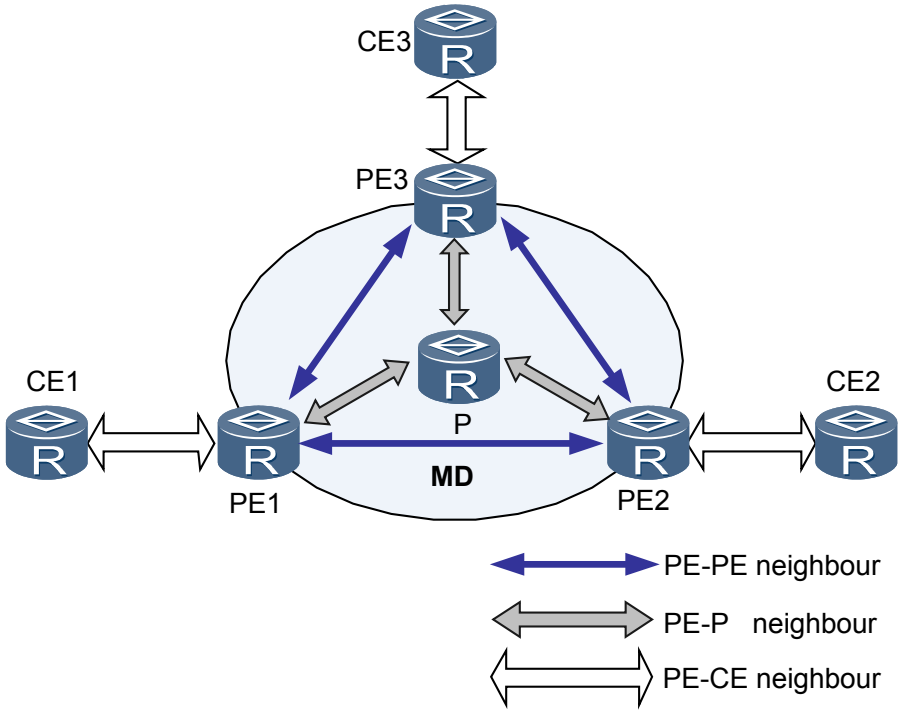


图 8-5 MD 方案中 CE、PE 和 P 的邻居关系



- PE-CE 邻居关系

  PE 上绑定 VPN 实例的接口与链路对端 CE 上的接口之间建立的 PIM 邻居关系。
- PE-P 邻居关系

  PE 上公网实例接口与链路对端 P 上的接口之间建立的 PIM 邻居关系。
- PE-PE 邻居关系

  PE 上的 VPN 实例通过 MTI 接收到从远端 PE 上的 VPN 实例发送来的 Hello 报文，建立邻居关系。

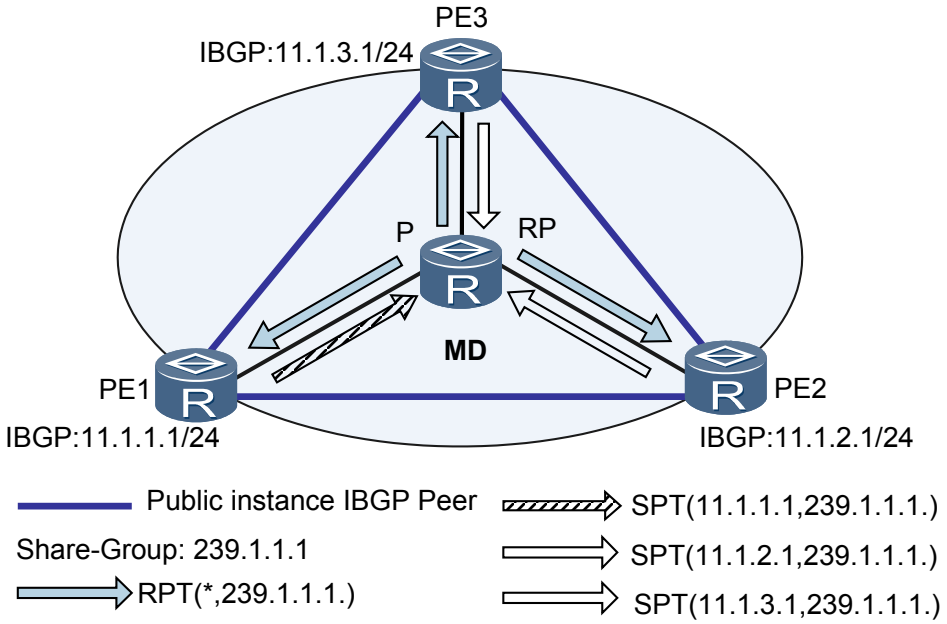
8.3.4 Share-MDT 建立过程

以 Share-Group 为组地址的 MDT（Multicast Distribution Tree），称为 Share-MDT。VPN 使用 Share-Group，唯一标识一棵 Share-MDT。

公网组播可以运行 PIM-SM 或 PIM-DM。这两种情况下，构建 Share-MDT 的过程是有区别的。

在 PIM-SM 网络中创建 Share-MDT

图 8-6 在 PIM-SM 网络中创建 Share-MDT



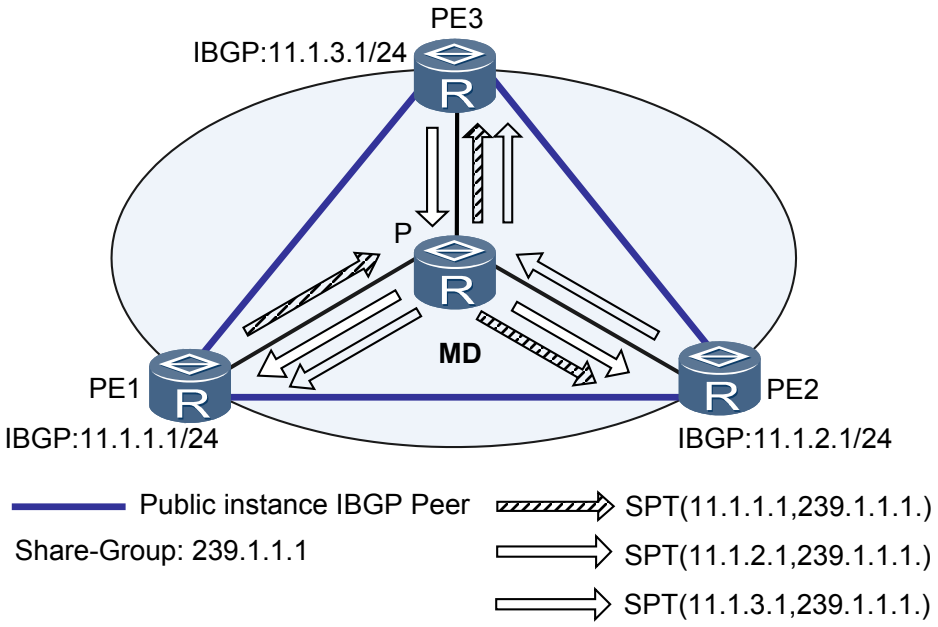
如图 8-6 所示，公网运行 PIM-SM，Share-MDT 的创建过程如下：

1. PE1 的公网实例向公网 RP 发起加入，以 Share-Group 地址为组播组地址，沿途在公网 PE 上创建（*，239.1.1.1）项。同时，PE2 和 PE3 也向公网 RP 发起加入，在 MD 中形成一棵以公网 RP 为根，PE1、PE2 和 PE3 为叶的 RPT。
2. PE1 的公网实例向公网 RP 发起注册，以 MTI（Multicast Tunnel Interface）地址为组播源地址，Share-Group 地址为组播组地址，公网 RP 上创建（11.1.1.1，239.1.1.1）项。同时，PE2 和 PE3 也向公网 RP 发起注册，在 MD 中形成三棵连接 PE 与 RP、相互独立的 RP-源树。

PIM-SM 网络中，由一棵 RPT（*，239.1.1.1）和三棵相互独立的 RP-源树共同组成了一棵 Share-MDT。

在 PIM-DM 网络中创建 Share-MDT

图 8-7 在 PIM-DM 网络中创建 Share-MDT



如图 8-7 所示，公网运行 PIM-DM，Share-MDT 创建过程如下：

以 PE1 上的公网实例为组播源、Share-Group 地址为组播组地址、其他支持 VPN A 的 PE 作为组成员，在整个公网的范围内发起扩散-剪枝过程。沿途在公网 PE 上创建（11.1.1.1，239.1.1.1）项，形成一棵以 PE1 为根，PE2 和 PE3 为叶的 SPT。同时，PE2 和 PE3 也在公网内发起扩散-剪枝，再分别形成两棵 SPT。

PIM-DM 网络中，这三棵相互独立的 SPT 共同组成了一棵 Share-MDT。

8.3.5 基于 Share-MDT 的 MT 传输过程

在 Share-MDT 构建完成后，就可以进行 MT 传输过程。

基于 Share-MDT 的 MT 传输过程简述

1. PE 上的 VPN 实例向 MTI 发出私网组播报文。
2. PE 不区分私网组播报文是协议报文或是数据报文，统一使用 MTI 地址作为源地址、Share-Group 地址作为组地址进行封装，转换为公网组播数据报文。
3. PE 将封装好的公网组播数据报文交给公网实例，公网实例将报文发出。
4. 公网组播数据报文沿 Share-MDT 转发，到达远端 PE 上的公网实例。
5. 远端 PE 对公网组播数据报文进行解封装，还原为私网组播报文，交给 VPN 实例。

MT 传输过程的主要任务

- 在 MTI 之间交互 Hello 报文，在各 PE 的 VPN 实例之间建立私网的 PIM 邻居。
- 在 MTI 之间交互其他协议报文，构建私网组播分发树。
- 传输私网组播数据。

说明

- 属于同一 VPN 的所有接口，包括 PE 上绑定 VPN 实例的接口和 MTI，必须运行统一的 PIM 模式。
- VPN 实例与公网实例相互独立，可以分别运行不同的 PIM 模式。

组播协议报文的传递过程

当 VPN 中运行 PIM-DM，协议报文的职责：

- MTI 之间交互 Hello 报文，建立 PIM 邻居。
- 跨越公网发起扩散-剪枝，创建 SPT。

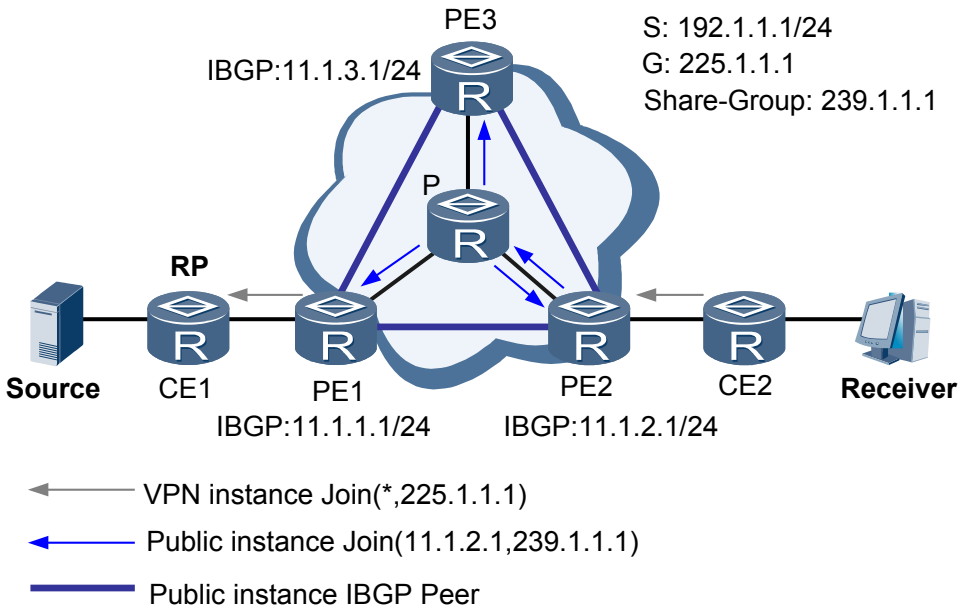
当 VPN 中运行 PIM-SM，协议报文的职责：

- MTI 之间交互 Hello 报文，建立 PIM 邻居。
- 如果接收者与私网 RP 属于不同 site，需要跨越公网发起加入，创建共享树。
- 如果组播源与私网 RP 属于不同 site，需要跨越公网发起注册，创建源树。

下面以公网运行 PIM-SM、VPN 运行 PIM-SM、私网接收者跨越公网发起加入为例，介绍基于 Share-MDT 的组播协议报文的传递过程。

如图 8-8 所示，VPN A 中的接收者 Receiver 属于 site2，与 CE2 相连。CE1 为私网组播组 G（225.1.1.1）的 RP，属于 site1。

图 8-8 组播协议报文的传递过程



组播协议报文的交互过程如下：

1. Receiver 通过 IGMP 协议通知 CE2 接收并转发组播组 G 的数据。CE2 在本地创建（*，225.1.1.1）项，同时向私网 RP（CE1）发起加入过程。
2. PE2 上的 VPN 实例接收到 CE2 发送的 Join 消息，本地创建（*，225.1.1.1）项，指定 MTI 为上游接口。然后将 Join 消息交由 P 做进一步处理。这时，PE2 上的 VPN 实例认为 Join 消息已从 MTI 发出。
3. PE2 对 Join 消息进行 GRE 封装，以 PE2 的 IBGP 接口地址为组播源地址，Share-Group 地址为组播组地址，转换成普通的公网组播数据报文（11.1.2.1，239.1.1.1）。然后交由 PE2 上的公网实例向公网转发。
4. 组播数据报文（11.1.2.1，239.1.1.1）沿 Share-MDT 传输到各 PE 上的公网实例。各 PE 对报文进行解封装，还原为发往私网 RP 的 Join 消息。然后，各 PE 检查该 Join 消息，如果发现私网 RP（CE1）在其直连 site 中，则交由其上的 VPN 实例处理，否则丢弃该 Join 消息。
5. PE1 上的 VPN 实例接收到 Join 消息，认为是从 MTI 获得的。本地创建（*，225.1.1.1）项，指定下游接口为 MTI、上游接口为朝向 CE1 的接口。同时向私网 RP 发送 Join 消息。
6. CE1 收到 PE1 上的 VPN 实例发送的 Join 消息，本地更新或创建（*，225.1.1.1）项，至此跨越 VPN 的组播共享树构建完成。

组播数据报文的传递过程

当 VPN 中运行 PIM-DM，沿私网 SPT 跨越公网传输私网组播数据。

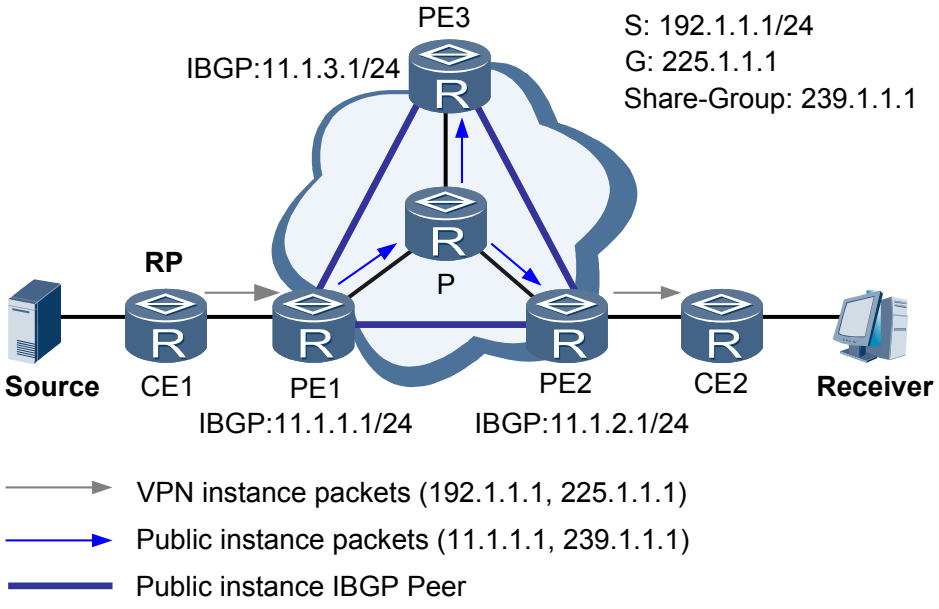
当 VPN 中运行 PIM-SM：

- 如果接收者与私网 RP 属于不同 site，沿私网 RPT 跨越公网传输私网组播数据。
- 如果组播源与接收者属于不同 site，沿私网源树跨越公网传输私网组播数据。

下面以公网运行 PIM-DM、VPN 运行 PIM-DM、沿 SPT 跨越公网传输私网组播数据为例，介绍基于 Share-MDT 的组播数据报文的传递过程。

如图 8-9 所示，VPN A 中组播源 Source 向组播组 G（225.1.1.1）发送组播数据。接收者 Receiver 属于 site2，与 CE2 相连。

图 8-9 组播数据报文的传递过程



私网组播数据跨越公网传输过程如下：

1. 组播源 Source 发送私网组播数据（192.1.1.1，225.1.1.1）到 CE1。
2. CE1 沿 SPT 将私网组播数据转发到 PE1，PE1 上的 VPN 实例查找转发项。如果对应的转发项出接口包含 MTI，则将该私网组播数据交由 P 做进一步处理。这时，PE1 上的 VPN 实例认为私网组播数据已从 MTI 发出。
3. PE1 对该私网组播数据进行 GRE 封装，以 PE1 的 IBGP 接口地址为组播源地址，Share-Group 地址为组播组地址，转换成普通的公网组播数据报文（11.1.1.1，239.1.1.1）。然后交由 PE1 上的公网实例向公网转发。
4. 组播数据报文（11.1.1.1，239.1.1.1）沿 Share-MDT 传输到各 PE 上的公网实例。各 PE 对报文进行解封装，还原为私网组播数据，然后交由对应的 VPN 实例处理。如果该 PE 上存在 SPT 下游接口，则沿 SPT 转发该私网组播数据，否则丢弃。
5. PE2 上的 VPN 实例搜索转发表项，将私网组播数据送达接收者 Receiver。至此，私网组播数据跨越公网传输完成。

8.3.6 Switch-MDT 切换

从上述 Share-MDT 的建立过程中，可以看到与 PE3 相连接的私网实例中没有接收者，但(192.1.1.1, 225.1.1.1)的私网组播数据还会到达 PE3，这是 MD 方案的一个缺点，即：所有属于同一 MD 的 PE 无论其是否有下游接收者，都会接收到组播数据报文。这样造成了带宽浪费，也增加了 PE 的处理负担。

在组播 VPN 的实现中，提出了一种优化的按需发送方法－Switch-MDT。下面介绍其实现过程，假设已按照上述步骤成功建立 Share-MDT。

1. 在 PE1 上配置 Switch-MDT 的切换组地址范围为 238.1.1.0 ～ 238.1.1.255，并且配置切换到 Switch-MDT 的数据转发阈值。
2. 当 CE1 上所连接的源发送数据速率超过 PE1 上所配置的转发阈值，PE1 在所配置的 Switch-MDT 地址范围中选择一个组地址 238.1.1.0，并通过 Share-MDT 向其他 PE 周期性发送一个切换到 Switch-MDT 的信令。
3. PE2 有下游接收者，接收到该信令报文后，加入到组 238.1.1.0 中，建立 Switch-MDT。建立 Swith-MDT 的过程与建立 Share-MDT 的过程的类似，。PE3 接收到切换信令后，没有下游接收者，不加入到 Switch-MDT。此后，私网数据报文(192.1.1.1, 225.1.1.1)只有 PE2 接收到。PIM 协议的控制报文仍然通过 Share-MDT 进行分发。

   发起 Switch-MDT 切换，必须满足以下两点要求：

   - 私网组播数据报文的源地址和组地址符合 ACL 过滤规则指定的源地址和组地址范围，不符合此范围的组播数据报文仍沿 Share-MDT 转发。
   - 私网组播数据报文的转发速率超过了切换阈值，且维持一定的时间。
4. 在某些情况下，私网组播数据的转发速率会在切换阈值上下振荡。为了避免组播数据流在 Share-MDT 与 Switch-MDT 之间频繁切换，当系统经过计算发现转发速率高于阈值后，并不立即执行切换，而是启动 switch-delay 定时器。在 switch-MDT 建立期间，仍然使用 share-MDT 转发组播数据。即 switch-delay 定时器可以保证组播数据在 Share-MDT 向 Switch-MDT 切换过程中不断流。在建立在 switch-delay 定时器超时前，系统将继续检测数据转发速率。如果速率始终高于切换阈值则切换至 Switch-MDT，否则继续使用 Share-MDT 进行报文转发。

支持 Switch-MDT 回切

当私网组播数据切换到 Switch-MDT 以后，由于情况变化导致切换条件变得不满足，PE1 就会把此私网组播数据从 Switch-MDT 反向切换回 Share-MDT，反向切换条件包括：

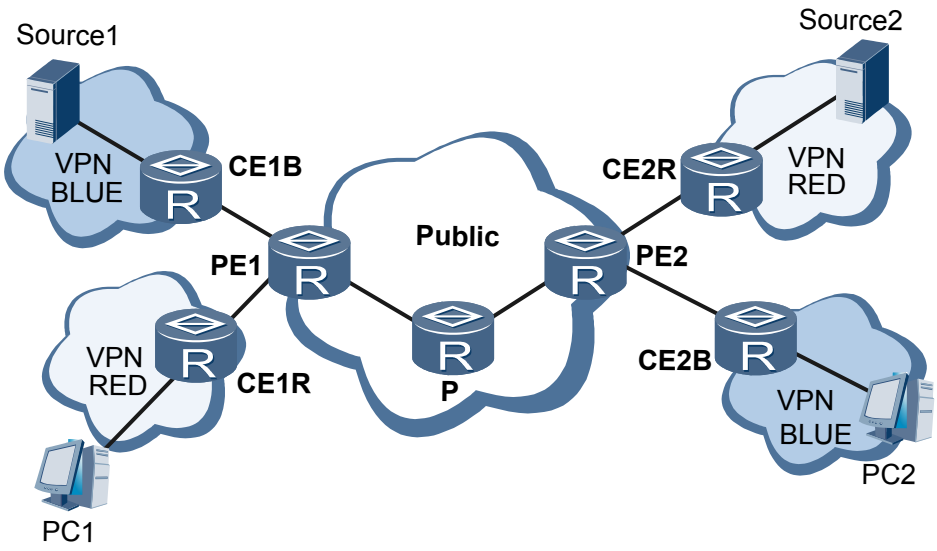
- 私网组播数据转发速率低于指定阈值，且维持 Switch-Holddown 的时间。
- 在某些情况下，私网组播数据的转发速率会在切换阈值上下振荡。为了避免组播数据流在 Switch-MDT 与 Share-MDT 之间频繁切换，当系统经过计算发现转发速率低于阈值后，并不立即执行切换，而是启动 holddown 定时器，超时时间为对应命令的配置值。在 holddown 定时器超时前，系统将继续检测数据转发速率。如果速率始终低于切换阈值则切换回 Share-MDT，否则继续使用 Switch-MDT 进行报文转发。
- 当更改切换组地址池时，用于私网组播数据封装的 Switch-Group 地址不在地址池中。
- 当控制私网组播数据切换到 Switch-MDT 的高级 ACL 规则发生变化，导致私网组播数据不能通过新 ACL 规则过滤。

8.4 应用

8.4.1 单自治域 MD VPN

应用场景：主要应用于同一个域内不同实例的组播业务的隔离。

图 8-10 单自治域 MD VPN



如图 8-10 所示，单 AS 中正常运行 MPLS/BGP VPN。在 PE1 和 PE2 上都配置两个 VPN 实例 VPN BLUE 和 VPN RED，并且保证在两台 PE 上相同的 VPN 实例中配置相同的

Share-Group 组地址。这样就使得配置相同 Share-Group 的 VPN 实例加入到同一个 MD 中。创建了对应的 Share-MDT 后，相应私网内的协议报文、低速数据就可以通过各自的 MT进行传输。

下面以 VPN BLUE 为例，描述私网组播业务互通的过程：

1. 在 PE1 和 PE2 上都配置实例 VPN BLUE，并且使用相同的 Share-Group。创建了对应的 Share-MDT 后，分别与 CE1B 和 CE2B 相连的两个私网 VPN BLUE 就可以通过对应的 MT 隧道相互发送组播协议报文。
2. 分别处在与 CE1B 和 CE2B 相连的两个私网中的组播设备建立邻居关系，相互发送组播加入、剪枝、BSR 等报文。这些私网协议报文只有在 PE 上才进行 MT 封装和解封装等特殊处理。私网中的设备并不知道自己处于私网中，像公网设备一样处理组播协议报文、转发组播数据，从而实现了同一实例内组播业务的互通、不同实例内组播业务的隔离。

8.4.2 跨自治域 MD VPN

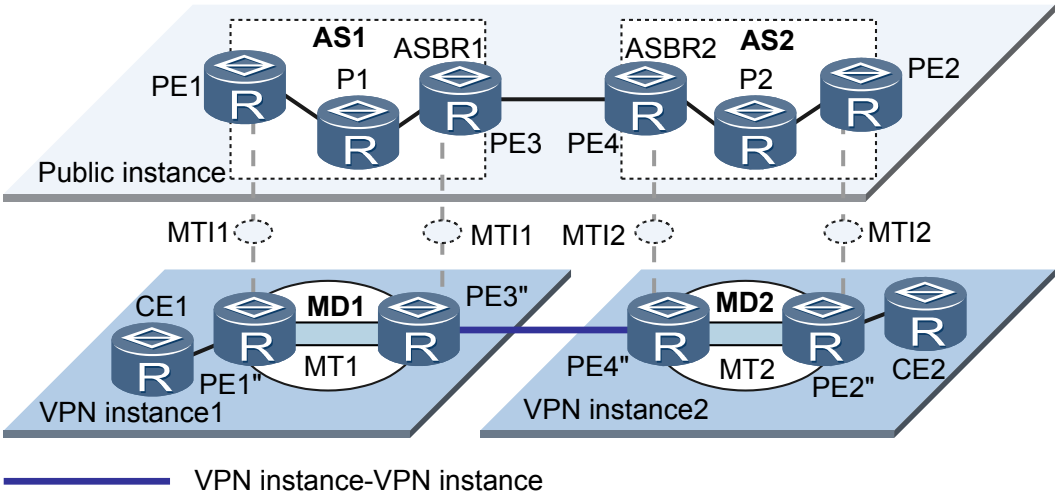
应用场景：主要应用于跨自治域的组播业务互通，并将不同的组播业务进行隔离。

跨自治域 MD VPN 有以下两种实现方式。

VPN 实例-VPN 实例方式

VPN 跨越多个 AS，各个 AS 之间通过 VPN 实例相连。

图 8-11 VPN 实例-VPN 实例跨自治域 MD VPN



如图 8-11 所示，VPN 跨越 AS1 和 AS2 两个自治域。PE3 和 PE4 分别是 AS1、AS2 的 ASBR（AS 域边界路由器）。PE3 和 PE4 通过各自的 VPN 实例相连，互相把对方当作“CE”设备来看待。

采用“VPN 实例-VPN 实例”方式，需要在每个自治域中分别建立一个独立的 MD，在两个 MD 之间实现私网组播数据跨自治域传输。

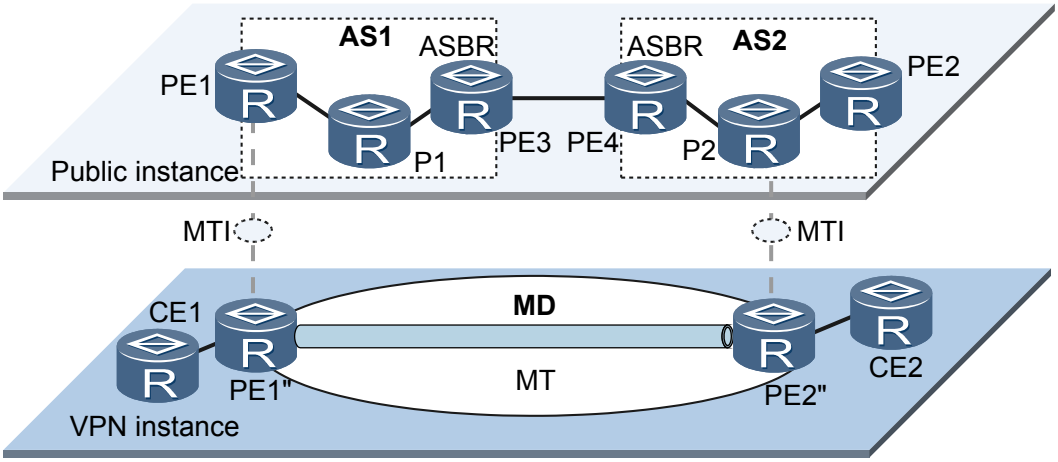
采用“VPN 实例-VPN 实例”方式实现跨自治域传输私网组播数据的过程为：

1. 在 AS1 中，CE1 与 PE3 相连的私网 CE 之间可以互通组播业务，其中与 PE3 相连的私网 CE 实际上就是 PE4。虽然 PE4 在 AS2 中被当作 PE 部署，但是对于 PE3 来说，PE4 只是与其私网接口相连的一个 CE；同样，对于 PE4 来说，PE3 也只是与其私网接口相连的一个 CE。
2. CE1 的私网 VPN1 组播协议报文和数据到达 PE4 后，PE4 会认为是与自己私网接口相连的私网实例 VPN2 的组播协议报文，并将其封装后通过 MD2 转发到与 PE2 相连的私网 CE2 上。与此类似，CE2 的私网协议报文也可以反向到达 CE1。这样，就实现了 CE1 与 CE2 上的组播业务跨自治域互通。

Multi-hop EBGP 方式

VPN 跨越多个 AS，各个 AS 之间通过公网 EBGP 相连。

图 8-12 Multihop EBGP 跨自治域 MD VPN



如图 8-12 所示，VPN 跨越 AS1 和 AS2 两个自治域，PE3 和 PE4 分别是 AS1、AS2 两个自治域的 ASBR。两个 ASBR 通过各自的公网 EBGP 相连，互相把对方当作“P”设备来看待。

采用“Multi-hop EBGP”方式，只需要在两个自治域 AS1 和 AS2 中建立一个 MD，在该 MD 内实现公网组播数据跨自治域的传输。

采用 Multi-hop EBGP”方式实现跨自治域传输私网组播数据的过程为：

1. 来自 CE1 的私网协议报文和数据通过 PE1 上的 MTI 封装后在 MT 隧道中进行传输。这样，被封装后的私网协议报文在公网就变成了普通的组播数据流，通过公网的 Share-Group 表项进行转发。
2. 在 PE3 和 PE4 两台域边界路由器配置公网 EBGP 连接之后，AS1 和 AS2 就实现了公网的互通。同样，公网的组播业务也实现了互通，被封装成公网普通组播数据的私网协议报文就可以顺利到达 PE2。对于 PE1 和 PE2，不需要关心被封装的私网报文是通过何种方式到达的，就像处在同一个自治域内一样。这样，就实现了跨自治域的私网 VPN 互通。

8.5 术语与缩略语

术语

术语	解释
PIM	Protocol Independent Multicast，协议无关组播，属于组播路由协议。网络中单播路由畅通是 PIM 转发的基础。PIM 利用现有的单播路由信息，对组播报文执行 RPF 检查，从而创建组播路由表项，构建组播分发树。
SPT	Shortest Path Tree，最短路径树。以组播源为根，组播组成员为叶子的组播分发树称为 SPT。SPT 同时适用于 PIM-DM、PIM-SM 和 PIM-SSM。
Share-Group	根据 Multicast Domains 的原理，所有属于同一 MD 的 PE 上的 VPN 实例都要加入一个公共的组，将其称之为 Shared-Group。 目前组播 VPN 的实现中，一个 VPN 实例只能配置一个 Share-Group，即一个 VPN 实例只能加入到一个 MD 中。
Share-MDT	Share-Multicast Distribution Tree，即共享组播分发树，实际是由 PE 上的公网 PIM 实例加入到根据 Share-Group 建立的组播分发树，用于将 VPN 内的 PIM 协议报文和低速数据报文分发给其他同属于一个 VPN 的 PE，这个组播分发树称之为 Share-MDT。
MTI	Multicast Tunnel Interface，组播隧道接口。MTI 是 MT 的入/出口，相当于 MD 的入/出口。本地 PE 将私网数据从 MTI 发出，远端 PE 从 MTI 接收私网数据。 PE 定义 MTI 调用整个 MT 传输过程。MTI 实际上是 PE 上公网实例和 VPN 实例进行交互的“通道”。PE 使用 MTI 连接到 MT 上，相当于被连接到了共享网段上。各个 PE 上属于该 MD 的 VPN 实例在 MTI 上建立 PIM 邻居关系。
Switch-Group	在建立起 Share-MT 后，所有有私网接收者的 PE 为建立 Switch-MT 而加入的组。
Switch-MDT	Switch-Multicast Distribution Tree，为了避免数据流向不必要的 PE，在建立起 Share-MDT 后，所有有私网接收者的 PE 加入到一个用 Switch-Group 组而建立起来的按需发送的组播分发树，用于将 VPN 的高速数据报文分发给其他同属于一个 VPN 的 PE。
组播 VPN Extranet	组播 VPN Extranet 用来实现不同企业组织之间组播业务的分发，以及服务或内容提供商对多个不同的企业组织用户分发组播业务的需要。
自动发现	采用 BGP 协议发现属于同一组播 VPN 的 PE 邻居地址。在 BGP 携带的消息中定义新的地址族，当在 PE 上配置组播 VPN 的时候，通过 BGP PEER，向所有 PEER 发布组播 VPN 的配置信息，包括 RD 以及 share-group 地址。对端 PE 收到 BGP 发布的 SAFI 消息以后，与自己配置的 share-group 地址进行匹配比较，确认加入相同的 VPN 后，利用该信息建立公网组播 PIM-SSM 转发树，承载私网组播业务。

缩略语

缩略语	英文全称	中文全称
AS	Autonomous System	自治系统
ASBR	Autonomous System Boundary Router	自治系统边界路由器
BGP	Border Gateway Protocol	边界网关协议
PIM-SM	Protocol Independent Multicast Sparse Mode	协议独立组播－稀疏模式
RP	Rendezvous Point	汇聚点
A-D	Auto-Discovery	自动发现

9 MLD

关于本章

9.1 介绍

定义

MLD（Multicast Listener Discovery）组播监听者发现协议，是负责 IPv6 组播成员管理的协议，用来在 IPv6 主机和与其直连的组播路由器之间建立、维护组播组成员关系。

通过在接收者主机和与其直连的组播路由器上配置 MLD，可以实现主机动态加入和组播路由器对本地网络组成员信息的管理。

到目前为止，MLD 有两个版本：MLDv1 版本（RFC2710）、MLDv2 版本（RFC3810）。所有 MLD 版本都支持 ASM（Any-Source Multicast）模型。MLDv2 版本可以直接应用于 SSM（Source-Specific Multicast）模型，而 MLDv1 则需要通过使用 SSM Mapping 机制来支持 SSM 模型。

MLD 可以理解为 IGMP 的 IPv6 版本。两者的实现方式具有类比性，如 MLDv1 可以类比 IGMPv2；MLDv2 可以类比 IGMPv3。

鉴于 MLD 和 IGMP 在一些特性的实现上两者没有区别，在以下的篇幅中主要介绍 MLD 特有的特性，对于两者实现相同的特性不再赘述。这些特性相同的包括：

- MLD Router-Alert
- MLD Only-Link
- MLD On-Demand
- MLD Prompt-Leave
- MLD static group
- MLD Group-Policy
- MLD SSM Mapping
- MLD Limit

MLD 特有的特性为：MLDv1/v2 原理描述、MLD 查询器选举、MLD 组兼容。

目的

在 IPv6 网络中，通过在接收者主机和与其直连的组播路由器上配置 MLD，可以实现主机动态加入和组播路由器对本地网络组成员信息的管理。

9.2 参考标准和协议

本特性的参考资料清单如下：

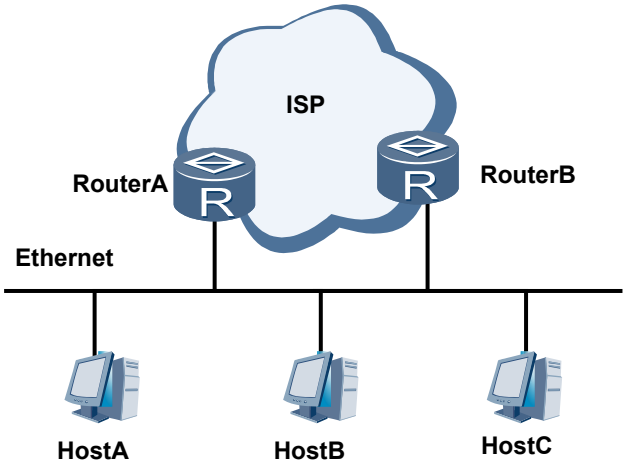
文档	描述	备注
RFC2710	Multicast Listener Discovery (MLDv1) for IPv6	-
RFC3810	Multicast Listener Discovery Version 2 (MLDv2) for IPv6	-
RFC3569	An Overview of Source-Specific Multicast (SSM)	-

文档	描述	备注
RFC4601	Protocol Independent Multicast - Sparse Mode (PIM-SM):Protocol Specification (Revised)	-

9.3 原理描述

9.3.1 MLDv1&v2

图 9-1 MLD 基本组网图



路由器通过发送查询报文并接收主机反馈的加入报文和离开报文来了解与该接口连接的网段上有哪些组播组存在接收者（即组成员）。如果出现组成员，组播路由器应将对应组的组播数据报文转发到这个网段；如果没有组成员则不转发。主机可以自主决定加入或退出某个组播组。

如图 9-1 所示，一个使能了 MLD 协议的路由器 RouterA 会自动变为查询器并定时发出 MLD 查询报文，与 RouterA 在同一网段的所有主机（HostA、HostB、HostC）都能收到它发出的查询报文。

- 主机收到查询报文的处理：

  如果 HostA 之前已经加入了组 G，就会在路由器允许的响应时间内随机地发送一个报告组 G 的 MLD 加入报文。

路由器收到组 G 的 MLD 加入报文会记录组 G 的相关信息，同时对该组启动一个定时器（如果已经启动则刷新），以便如果长时间没有主机响应时切断该组流量，并负责把组 G 的组播数据转发到 HostA 和 RouterA 相连的接口所在的网段上。

如果主机不是任何的组播组成员，在收到 MLD 查询报文时则不做任何响应。

- 主机加入组播组：

  一个新加入组播组 G 的主机会主动发送一个该组的 MLD 加入报文，通知路由器更新组播组信息，后续的加入报文则由路由器的查询报文来驱动。

- 主机离开组播组：

  如果一个主机决定离开某个组播组 G，会主动发送组 G 的 MLD 离开报文，路由器收到后会触发一个指定组 G 的查询，确定该组在当前网段接收者的存在情况。如果在查询结束后仍然没有收到主机针对该组的 MLD 加入报文，则删除已记录的组信息，停止转发该组数据到对应接口所在的网段上。

MLDv1 报文处理

运行 MLDv1 的主机发送的 MLD 报告中仅携带组信息，当主机发送一个组的 MLD 加入报文给路由器后，路由器就会通知组播转发模块，以便这个组的组播数据到来时能够正确转发给该主机。

MLDv1 协议具有报告抑制机制，可以减少网络中的 MLD 的重复报告。

当一个主机 HostA 加入了某个组播组 G，在收到路由器的查询报文后，HostA 会在 0 ～最大响应时间（查询报文中已经指定）之间选择一个随机值作为定时器的超时时间，并在该定时器超时后，向路由器发送该组 G 的加入报文。如果在超时时间内 HostA 收到了其他加入同一个组的主机 HostB 发送的加入报文，则 HostA 不再向路由器发送该组 G 的加入报文。

当主机退出某个组 G 的时候，会向路由器报告一个指定组 G 的 MLD 离开报文。由于 MLDv1 报告抑制机制，路由器无法确定是否还有其他主机加入了组 G。这时路由器会触发一个指定组 G 的查询，如果还有其他路由器加入了组 G 就会报告针对组 G 的 MLD 加入报文。

如果路由器发送了若干次数指定组 G 的查询之后仍然没有收到主机针对该组 G 的 MLD 加入报文，那么路由器就不再记录这个组，停止转发该组数据到对应接口所在的网段。

说明

MLD 的查询器和非查询器都会处理 MLD 组加入信息，但是只有查询器负责发送查询报文。MLD 非查询器不处理 MLDv1 离开报文。

MLDv2 报文处理

MLDv1 协议报文中只能携带组播组的信息，不能携带组播源的信息，这样运行 MLDv1 的主机就只能选择加入某个组，而不能选择加入某个组播源/组。MLDv2 协议解决了该问题。运行 MLDv2 的主机不仅能够选择组，还能够根据需要选择相应的组播源/组。同时主机发送的 MLDv2 报文中还可以包含多个组记录，每个组记录中可以包含多个组播源。

MLDv2 主机侧的报文类型共有 6 种：

- MODE_IS_INCLUDE，表示组播组与源列表之间的对应方式为 INCLUDE，即接收从指定源列表发往该组播组的数据。

- MODE_IS_EXCLUDE，表示组播组与源列表之间的对应方式为 EXCLUDE，即接收从指定源列表以外的组播源发往该组播组的数据。当 EXCLUDE 的列表为空等价于 MLDv1 的 report 报文。
- CHANGE_TO_INCLUDE_MODE，表示组播组与源列表之间的对应方式由 EXCLUDE 转换到 INCLUDE。如果这时指定源列表为空，则表示离开该组播组。
- CHANGE_TO_EXCLUDE_MODE，表示组播组与源列表之间的对应方式由 INCLUDE 转换到 EXCLUDE。
- ALLOW_NEW_SOURCES，表示在现有的基础上，还希望从某些组播源接收组播数据。如果当前对应关系为 INCLUDE，则向现有源列表中添加某些组播源；如果当前对应关系为 EXCLUDE，则从现有源列表中删除某些组播源。
- BLOCK_OLD_SOURCES，表示在现有的基础上，不再希望从这些组播源接收组播数据。如果当前对应关系为 INCLUDE，则从现有源列表中删除这些组播源；如果当前对应关系为 EXCLUDE，则向现有源列表中添加这些组播源。

在路由器侧，查询器通过发送查询报文并接收主机反馈的加入报文了解与该接口连接的网段上有哪些组播组存在接收者，并将对应的组播数据转发到相应的网段。MLDv2 的组记录有 include 和 exclude 二种组过滤模式。

- 在 include 模式下
  - 处于激活状态的组播源表示需要路由器转发这个源的数据。
  - 不活动的源会被路由器删除并停止转发这个源的数据。
- 在 exclude 模式下
  - 处于激活状态的组播源表示处于冲突域中。也就是说，与该路由器接口同一网段的主机中，有的主机需要该源的数据，有的主机不需要该源的数据，在这种情况下该源的数据仍然需要转发。
  - 不活动的组播源表示不需要转发该源的数据。
  - 组中没有记录的组播源的数据全部都要转发。

在 MLDv2 中，实现了对采用 Include 模式加入特定源组的 MLDv2 主机成员信息的跟踪功能。

MLDv2 协议相对于 MLDv1 协议，没有报文抑制机制，所有加入组播组的主机在收到查询时都会响应 MLD 加入报文。由于有了对组播源的选择，MLDv2 路由器在通用查询和指定组查询的基础上增加了指定源/组查询，用于在收到特定组播源的数据时确定是否存在该数据的接收者。

9.3.2 MLD 组兼容

MLD 组兼容模式是指支持高 MLD 版本的组播设备可以兼容低版本的主机。例如 v2 版本的设备可以正确处理 v1 主机的加入。当组播设备工作在兼容模式时，收到低版本的主机的 MLD 加入报文后会自动降低组的兼容版本到该主机对应的版本，并工作在该版本上。

工作在 MLDv2 版本的组播设备收到 MLDv1 主机发送的加入报文，会自动把该组的兼容模式设定为 v1 模式。在这种情况下，设备忽略 MLDv2 的 BLOCK 报文以及 MLDv2 的 TO_EX 报文的源列表，即抑制了 MLDv2 对组播源的选择功能。

通过手工配置把组播设备的 MLD 版本从低版本升到高版本时，如果有组存在，则会继续使组工作在低版本的兼容模式，直到低版本的主机退出该组播组。

说明

缺省情况下，MLD 的版本是 MLDv2。

9.3.3 MLD 查询器选举

使能了 MLD 协议的组播设备在网段中的角色有两个：

- 查询器

  负责发送查询报文并接收主机反馈的加入报文和离开报文，以此来了解与该接口连接的网段上有哪些组播组存在接收者（即组成员）。

- 非查询器

  只接收主机反馈的加入报文，了解与该接口连接的网段上有哪些组播组存在接收者，并根据网段中查询器的动作确定当前网段中有哪些组播组成员离开。

通常情况下一个网段只有一个查询器，因此组播设备之间需要用某些方式来选出查询器。查询器选举时采用以下的原则：

- RouterA 使能 MLD 协议后，在 MLD 协议启动阶段会默认自己为当前网段中的查询器，向网段中发送查询报文。如果收到 IP 地址比自己小的 RouterB 发来的查询报文，则 RouterA 由查询器转为非查询器，并启动“其他查询器存在定时器”，记录 RouterB 为当前网段的查询器。
- 如果 RouterA 在非查询器状态时收到查询器 RouterB 发送的查询报文，则更新“其他查询器存在定时器”；如果此时收到的查询报文不是先前记录的查询器 RouterB 发来的，而是新的 RouterC 发来的，且 RouterC 的 IP 地址比 RouterB 的小，则更新查询器为 RouterC，同时更新“其他查询器存在定时器”。
- 如果 RouterA 在非查询器状态时，“其他查询器存在定时器”超时，则由非查询器转为查询器状态，承担起查询器的职责。

说明

当前仅支持同网段上同版本的组播设备之间进行查询器选举。为了保证正常工作，需要在同网段所有设备上配置相同版本的 MLD。

9.3.4 协议的比较

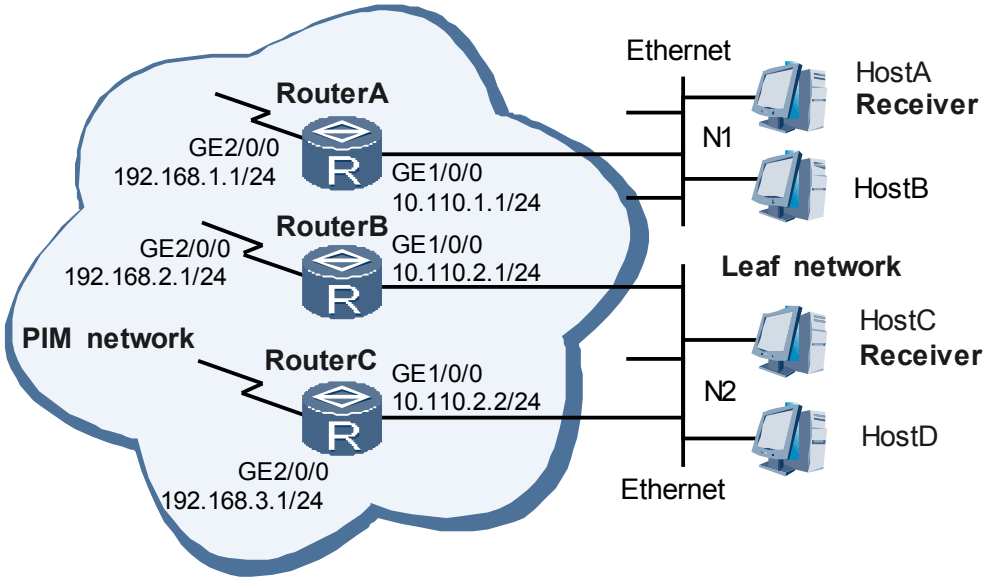
MLDv1 和 MLDv2 协议的比较：

MLDv1	MLDv2	MLDv2 较 MLDv1 的优势
报文中不能携带组播源信息，只能指定组播组信息	报文中除了携带组播组信息，还能携带组播源信息	可以直接对特定的组播源进行选择，增加了选择的粒度
一个报文中只能携带一个组记录	一个报文中可以携带多个组记录	减少了网段中的 MLD 报文数量
指定组查询报文无重传机制	指定组、指定源组查询报文有重传机制	非查询器和查询器维护的组播组信息能够更好地保持一致

9.4 应用

如图 9-2 所示，主机以组播方式接收视频点播信息（Video information On Demand）。隶属于不同组织的接收者组成了叶子网络，每个叶子网络中至少包含一个主机接收者。

图 9-2 MLD 应用组网图



主机 HostA 和 HostC 分别是叶子网络 N1、N2 网络的接收者。与主机直连的 RouterA 在接口 GE1/0/0 上配置 MLDv1，即叶子网络 N1 运行 MLDv1；与主机直连的 RouterB、RouterC 在各自的 GE1/0/0 上均配置 MLDv2，即叶子网络 N2 运行 MLDv2。同一网段的设备上要配置相同的 MLD 版本。

9.5 术语与缩略语

术语

术语	解释
MLD	Multicast Listener Discovery，称为组播监听者发现协议，用于 IPv6 组播设备发现其直连网段上组播监听者（Multicast Listener）、建立、维护组成员关系。 在 IPv6 网络中，通过在接收者主机和与其直连的组播路由器上配置 MLD，可以实现主机动态加入和组播路由器对本地网络组成员信息的管理。
(S,G)	属于组播路由表项，S 表示组播源，G 表示组播组。 源地址为 S、组地址为 G 的组播报文，到达组播设备后，从(S,G)表项中的下游接口转发出去。 通常，将源地址为 S，组地址为 G 的组播报文表示为(S,G)报文。
(*,G)	属于 PIM 路由表项，*表示任意源，G 表示组播组。 (*,G)表项适用于所有组地址为 G 的组播报文。不论是哪个组播源发出的，只要是发往组播组 G 的组播报文，都应该从(*,G)表项中的下游接口转发出去。

缩略语

缩略语	英文全称	中文全称
MLD	Multicast Listener Discovery	组播监听者发现协议

10 三层组播 CAC

关于本章

10.1 介绍

定义

组播 CAC（Call Admission Control，接入管理控制）技术主要实现了按照节目组对组播表项进行管理，并根据节目组所配置的表项个数和带宽限制，以及组播路由器接入接口的最大表项个数和带宽来限制组播路由器可创建的最大组播表项个数。

运用三层组播 CAC 技术可以使运营商在 IP 核心网络对用户接入的数量及可服务的 IPTV ICP（Internet Content Provider，网络信息提供者）进行限制，为运营商运营 IPTV 业务提供必要的控制手段和增值业务。

节目组：由多个成员组播组构成，这些成员组播组有自己的带宽属性，比如 CCTV 节目组，其成员 CCTV-1 带宽为 4M，CCTV-5 带宽为 18M。

目的

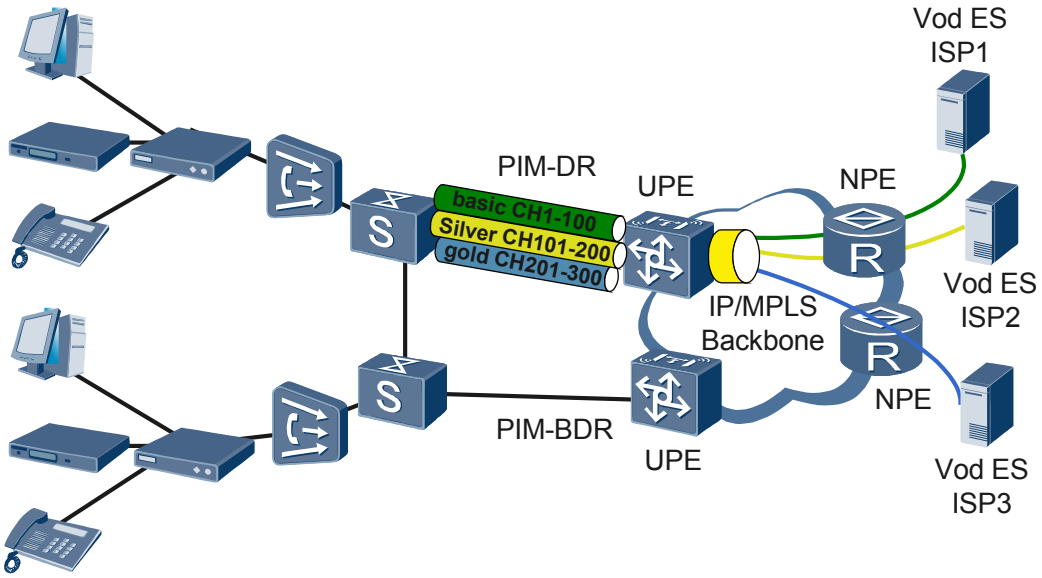
三层组播 CAC 主要实现以下两个功能：

- 通过限制 PIM 路由表项的个数来控制可以服务的组播组数量，避免引入超过转发能力的组播数据流。
- 通过为节目组或接口预留带宽的方式来规划组播网络，当节目组或接口的带宽资源不足时不允许新增组播组，以保障服务质量。

如图 10-1 所示：

- 通过组播 CAC 在 IP/MPLS BackBone 入口 NPE 和出口 UPE 上限制组播控制层面建立组播分发树，在 IP/MPLS BackBone 不引入超过转发能力的冗余流量。
- 对 IP/MPLS BackBone 内已经建立的组播分发树提供带宽保障。

图 10-1 组播 CAC 典型组网图



10.2 参考标准和协议

无

10.3 原理描述

10.3.1 组播 CAC 实现策略

组播 CAC 主要实现了表 1 中对组播表项和带宽限制的基本策略：

表 10-1 组播 CAC 实现策略

组播 CAC 策略	具体策略		描述
组播 CAC	组播 CAC 节目组管理		实现了全局按节目组管理组范围或源组范围，并设定相应的带宽。
	组播 CAC 全局限制	组播 CAC 未指定节目组加入控制	支持配置不接受未落在任何全局配置的节目组范围的组或源组加入。如果没有配置该功能，则默认为接受不属于任何全局配置的节目组范围的组或源组加入。
		组播 CAC 全局（单实例下）总体表项限制	当全局（单实例下）配置了总体的表项限制，收到 IGMP 或 PIM 的新加入消息时发现全局总体（单实例下）的表项限制超限，则 PIM 不创建表项。
		组播 CAC 全局（单实例下）节目组表项限制	当全局（单实例下）配置了节目组的表项限制，收到指定节目组范围的 IGMP 或 PIM 的新加入消息时发现全局（单实例下）基于当前节目组的表项限制超限，则 PIM 不创建表项。
	组播 CAC 出接口限制	组播 CAC 出接口总体限制	配置组播出接口上的 PIM 表项个数限制和带宽限制，当接口的可用带宽小于组播组的预留带宽或 PIM 表项个数达到最大值时，该接口拒绝新的加入报文。
		组播 CAC 出接口节目组限制	配置组播出接口上指定节目组的最大 PIM 表项数和带宽限制，当接口的可用带宽小于组播组的预留带宽或 PIM 路由表项个数达到最大值时，该接口拒绝新的加入报文。

10.3.2 组播 CAC

组播 CAC 通过在城域网入口 NPE 和出口 UPE 上限制组播控制层面建立组播分发树，在城域网不引入超过转发能力的冗余流量。同时，组播 CAC 能够对已经建立的组播分发树提供带宽保障。

为便于用户管理组播组，按节目划分不同的节目组，为每个节目组配置允许的组播表项个数和带宽限制。根据组播的 ASM/SSM 模型，表项个数计算方法如下：

- 对于 ASM 模型，同一 G 下无论有无(*,G)，有多少(S1,G)，(S2,G)．．．，都计为一个表项。
- 对于 SSM 模型，一个(S,G)计为一个表项。

组播 CAC 节目组管理

为便于提供 IPTV 业务的运营商管理 IP 核心网设备上的组播表项，组播 CAC 采取以易于管理的节目组名字的方式，来管理组播组或组播源组。

组播 CAC 提供了一套完整的节目组管理方法，使运营商对现有网络中的组播数据流按照节目组的方式进行管理，并根据节目组制定统一的组播 CAC 策略。IPTV 运营商可能会将相同带宽的组播表项划分到同一节目组，或者将来自同一 ISP 的组或者源组划分到同一节目组。

目前组播业务主要采用 ASM 模型和 SSM 模型提供组播路由和转发能力：

- 对于 ASM 模型，用户侧设备只需要知道要加入的组播组，就可以接收到该组所有源的组播数据。
- 对于 SSM 模型，用户侧设备需要知道要加入的组播组和组播源，才能接收到对应的组播数据。

组播 CAC 节目组管理支持以下配置策略：

- 支持配置属于某节目组表项的组或源组范围，且指定该范围组播表项的带宽。

  当配置节目组表项范围时，已存在表项对应的节目组可能已经发生变化，且已经存在表项的默认带宽也可能发生改变，那么需更新已存在表项基于节目组的全局和出接口的统计计数以及总体的统计计数。
- 支持改变节目组下的组或源组范围的带宽时，更新已存在表项的出接口带宽计数。
- 支持对未落在全局配置的节目组范围内的组或源组加入进行控制。

组播 CAC 全局限制

当运营商部署 IPTV 业务时，为防止进入接入网络的组播数据流超过组播设备的处理能力或超过接入网络的带宽限制，在 NPE 上配置可以建立的最大组播表项个数。

组播 CAC 支持配置全局的表项限制。在配置组播 CAC 全局表项限制时，对已经存在的表项不进行限制，仅更新 PIM 已经存在(*,G)/(S,G)表项的计数。组播设备收到 IGMP 或 PIM 的新加入，PIM 新创建协议表项时，若全局表项限制超限，则不会创建 PIM 表项。

组播 CAC 全局限制包括以下三种限制策略：

- 组播 CAC 未指定节目组加入控制

  如果没有配置该功能，则默认为接受不属于任何全局配置的节目组范围的组或源组加入。配置此功能可实现组播设备不接受未落在任何全局配置的节目组范围的组或源组加入。

- 组播 CAC 全局（单实例下）总体表项限制

  根据 ASM 和 SSM 模型下表项的计算方法，对本实例的所有表项（无论是否属于节目组范围）进行个数和带宽统计，从而对本实例用户的表项个数和带宽进行限制。

- 组播 CAC 全局（单实例下）节目组表项限制

  根据 ASM 和 SSM 模型下表项的计算方法，对本实例属于此节目组表项进行个数和带宽统计，从而对本实例本节目组用户的表项个数和带宽进行限制。

说明

组播 CAC 支持全局可容纳表项限制增大后，收到由于全局表项限制没有创建表项成功的加入消息时，会重新创建 PIM 表项。

组播 CAC 出接口限制

当运营商部署 IPTV 业务时，为防止进入接入网络的组播数据流超过组播设备的处理能力或超过接入网络的带宽限制，可以在 UPE 的接入接口上配置组播 CAC 表项和带宽限制。UPE 的接入接口即组播出接口，用于进行组播流量复制。

组播 CAC 出接口限制通过限制接入接口上允许的用户加入，从而达到限制此接口流量的复制。组播 CAC 出接口限制包括表项个数和带宽限制两方面。

当接口配置组播 CAC 出接口限制时，更新 PIM 已经存在表项的出接口计数。当收到 IGMP/PIM 加入，增加 PIM 表项时，进行组播 CAC 出接口限制检查，若没有超限，则根据 ASM 和 SSM 模型下表项的计算方法计入表项个数和带宽统计。

当配置的出接口限制表项计数小于已存在的表项个数时，不删除已经建立的表项。但该接口拒绝新的加入报文。

组播 CAC 出接口表项和带宽限制包括以下两种限制策略，分别对出接口上的表项个数和带宽进行统计，从而限制用户加入：

- 组播 CAC 出接口总体限制

  根据 ASM 和 SSM 模型下表项的计算方法，对所有表项（无论是否属于节目组范围）在此接口上进行个数和带宽统计，从而对指定接入接口上所有用户的表项个数和带宽进行限制。

- 组播 CAC 出接口节目组限制

  根据 ASM 和 SSM 模型下表项的计算方法，对属于此节目组的表项在此接口上进行个数和带宽统计，从而对指定接入接口本节目组的用户进行表项个数和带宽限制。

说明

组播 CAC 支持出接口可容纳表项个数增大或带宽限制增大时，收到由于出接口表项和带宽限制没有创建表项成功的加入消息时，会重新创建 PIM 表项。

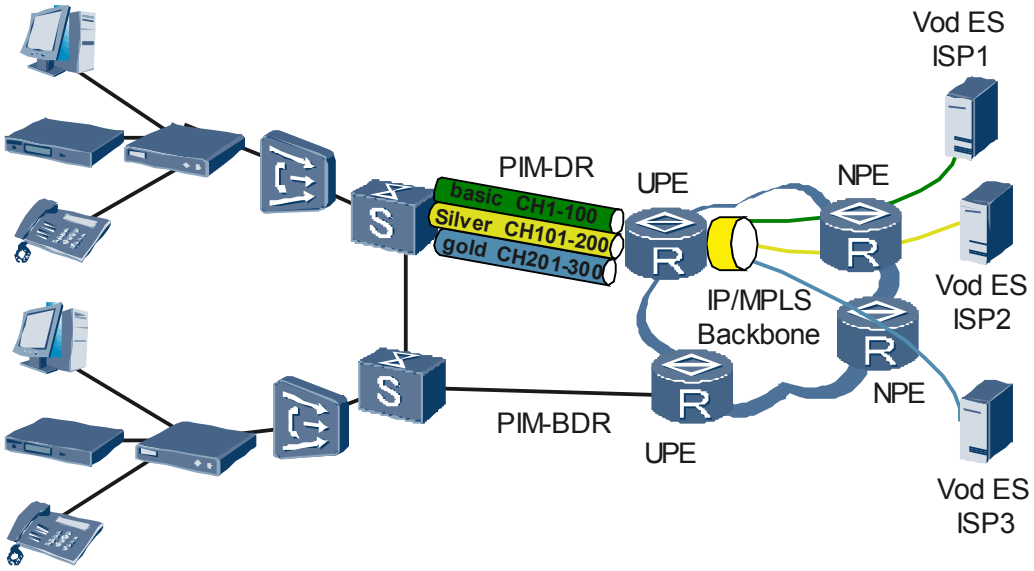
10.4 应用

10.4.1 组播 CAC 应用

10.4.1 组播 CAC 应用

组播 CAC 节目组管理应用

图 10-2 组播 CAC 节目组管理应用组网图



如图 10-2 所示，在 UPE 上可以配置节目组组范围或源组范围及允许带宽。

在 IPTV 业务网络中，运营商的 IP 核心网络可能连接多个 ISP，每个 ISP 利用预先分配的组播组（Multicast Group，简称 G）或者源组（简称(S,G)）承载各自的 IPTV 节目。每个 G 或者(S,G)会消耗不同的带宽，为了便于 IP 核心网对 G 或(S,G)对应的组播表项进行管理，可以为这些组播组分配一个易于管理的节目组名。

IPTV 运营商可能会将相同带宽的 G 或者(S,G)划分到同一节目组，或者会将来自同一 ISP 的 G 或者(S,G)划分到同一节目组。并根据节目组制定统一的组播 CAC 策略。

组播分发树有两种模式：

- ASM 模型

  用户只需要知道要加入的组播组，就可以接收到组播节目，目前用于建立 ASM 模型组播分发树的组播路由协议是 PIM-SM 协议。

- SSM 模型

  用户需要预先知道要加入的组播组和组播源，才能接收到组播节目，用于建立 SSM 模型组播分发树的组播路由协议是 PIM-SSM 协议。

在进行节目组编排时，用户必须指定节目组所建立的组播分发树模型。

为确保指定组播表项属于单一的节目组，对节目组作如下约束：

- 任何一个节目组必须提供一个可管理的节目组名。
- ASM 模型和 SSM 模型下节目组的配置规则。
  - 对于 ASM 模型的节目组，用户只能配置 G/Mask 的规则。

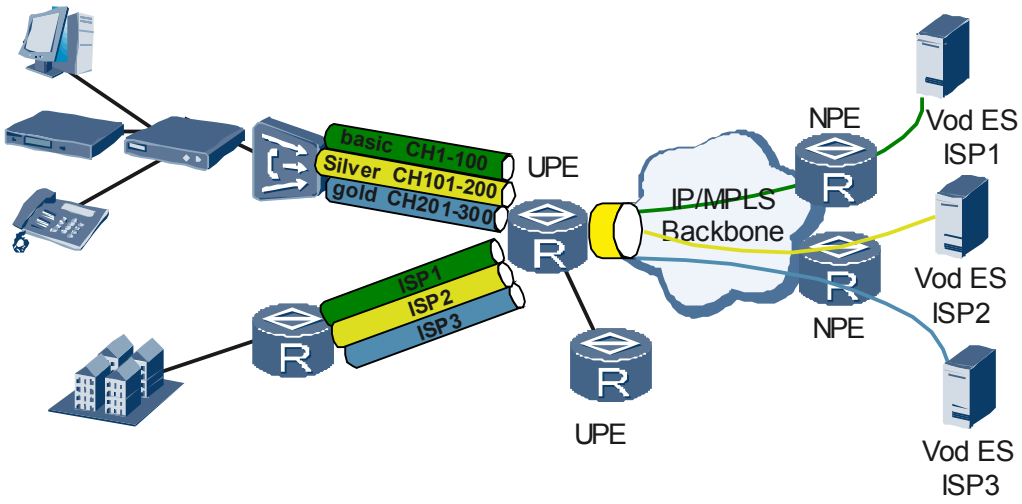
    当一个 ASM 模型的节目组配置了一条 G1/Mask，本节目组及其它节目组不允许再配置(G1/Mask，S/Mask)，也不允许配置与 G1/Mask 重叠的 G2/Mask 或者（G2/Mask, S/Mask）。G2/Mask 与 G1/Mask 重叠意味着 G2/Mask 包含该 G1/Mask 或被该 G1/Mask 包含。

- 对于 SSM 模型的节目组，用户只能配置 G/Mask, S/Mask 的规则。

  如果一个 SSM 模型的节目组配置了一条（G1/Mask, S1/Mask），本节目组及其它节目组不允许再配置 G1/Mask，也不允许再配置与 G1/Mask 重叠的 G2/Mask，或者(G2/Mask, S1/Mask)。但只要 S2/Mask 与 S1/Mask 不重叠，本节目组或其它节目组可以配置(G1/Mask, S2/Mask)。

组播 CAC 全局限制应用

图 10-3 组播 CAC 全局限制组网图



如图 10-3 所示，组播 CAC 全局限制用于 NPE 上，从组播流量入口处限制进入 IP/MPLS Backbone 的组播流量。当运营商部署 IPTV 业务时，为防止进入接入网络的组播数据流超过组播设备的处理能力或超过接入网络的带宽限制，在 NPE 上配置全局可以建立的最大组播表项个数和节目组的最大组播表项个数。

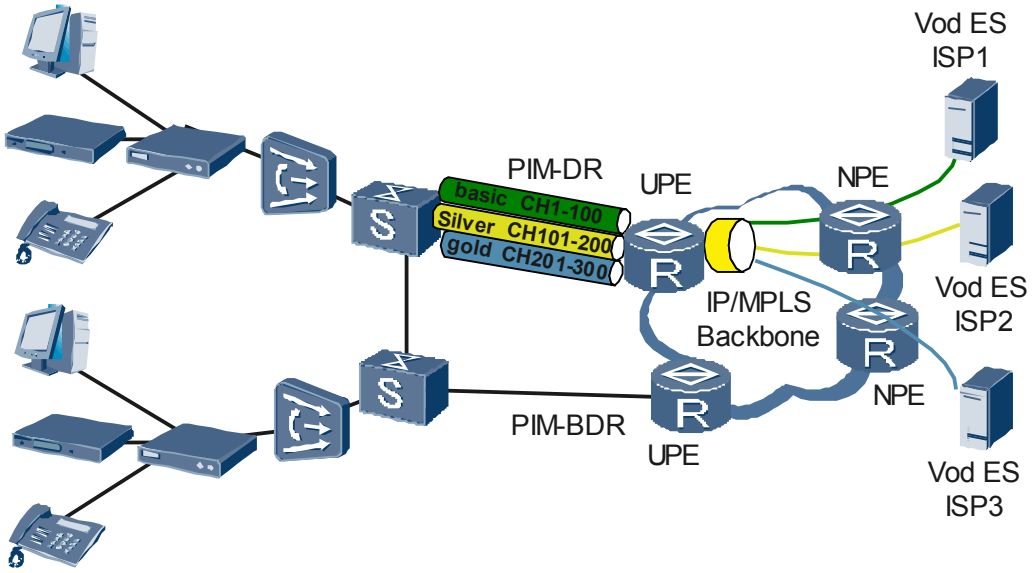
通过组播 CAC 全局基于节目组和总体的表项限制，可以实现运营商对 IP 核心网络连接的多个 ISP 提供的节目组频道和所有频道按照不同的控制策略进行的统一管理。

说明

由于组播 CAC 是对后续加入的组播表项进行限制，因此当全局配置表项限制或基于节目组的表项限制时，对已经建立的表项不进行删除操作。NPE 全局限制对配置静态组加入，或者私网加入而触发创建的表项只进行个数统计，不对这些表项进行 CAC 限制。

组播 CAC 出接口限制应用

图 10-4 组播 CAC 出接口限制组网图



如图 10-4 所示，组播 CAC 出接口限制用于 UPE，从组播流量出口处限制可以申请进入 IP/MPLS Backbone 的组播流量。

当运营商部署 IPTV 业务时，为防止进入接入网络的组播数据流超过组播设备的处理能力或超过接入网络的带宽限制，可以在 UPE 的接入接口上配置出接口总体/节目组表项个数和带宽限制。

UPE 的接入接口收到 IGMP 加入主要有以下几种场景：

- 运营商采用 IGMPv2+PIM-SM 的方式提供 IPTV 业务。

  用户设备运行 IGMPv2，接入接口运行 PIM-SM，并且在接入接口上配置组播 CAC 出接口总体/节目组表项个数和带宽限制。
- 运营商采用 IGMPv3+PIM-SM/SSM 的方式提供 IPTV 业务。

  用户设备运行 IGMPv3，接入接口运行 PIM-SM/SSM，并且在接入接口上配置组播 CAC 出接口总体/节目组表项个数和带宽限制。
- 运营商采用 IGMPv2+SSM-Mapping+PIM-SSM 的方式提供 IPTV 业务。

  用户设备运行 IGMPv2，接入接口运行 PIM-SM/SSM，UPE 上配置 SSM-Mapping 映射关系，并且在接入接口上配置组播 CAC 出接口总体/节目组表项个数和带宽限制。

10.5 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
ASM	Any Source Multicast	任意源组播
CAC	Call Admission Control	接入管理控制
ICP	Internet Content Provider	网络信息提供者
ISP	Internet Service Provider	网络服务提供者
PIM	Protocol Independent Multicast	协议无关组播
PIM-SM	Protocol Independent Multicast - Sparse Mode	协议无关组播－稀疏模式
PIM-SSM	Protocol Independent Multicast - Source-Specific Multicast	协议无关组播－指定源组播

11 二层组播 CAC

关于本章

11.1 介绍

定义

CAC（Call Admission Control）称为接入管理控制。二层组播 CAC 是 IPTV 组播方案的组成部分，主要在二层组播场景下控制 IPTV 的节目数量和带宽，避免出现带宽需求超出接入汇聚网带宽的情况，保证大多数用户的服务质量。

目的

随着 IPTV 的发展，节目频道数量快速增加，当有用户点播频道数量增加时，会出现带宽需求超出接入汇聚网带宽的问题，导致汇聚层设备负载过重而使整体用户满意度下降。倘若存在使用组播进行的网络攻击，可能会致使设备忙于处理大量攻击报文，无法有效地响应正常的网络要求。

因此，在运营商提供 IPTV 业务时，需要考虑节目频道数量很多的情况下，网络能否承担过度分散的频道的问题。在网络带宽不足时，需要拒绝用户加入新的频道的请求，这样，在牺牲少量用户的满意度的同时，能够保证绝大多数用户的服务质量。

二层组播 CAC 在部署了二层组播的场景下可以根据不同的维度进行接入限制，实现其组播业务的精确控制，避免出现带宽需求超出汇聚网带宽的情况，保证大多数用户的服务质量；同时在一定程度上也能减少组播攻击带来的危害。

11.2 参考标准和协议

与二层组播特性相关的参考标准与协议如下：

文档	描述	备注
RFC1112	Host Extensions for IP Multicasting	-
RFC2236	Internet Group Management Protocol, Version 2	-
RFC4541	Considerations for IGMP and MLD Snooping Switches	-

11.3 原理描述

11.3.1 基本概念

二层组播：在二层网络（VLAN/VPLS）中，通过 IGMP Snooping 对 IGMP 协议报文侦听学习，建立二层组播转发表，指导组播数据流转发。

CAC：Call Admission Control，接入控制管理。提供一系列规则来控制组播组学习，包括对组播组数量限制，组播组带宽限制，同时可以对节目组进行限制。

节目组：包含多个成员组播组构成，这些成员组播组有自己的带宽属性，比如 CCTV 节目组，其成员 CCTV-1 带宽为 4M，CCTV-5 带宽为 18M。

11.3.2 组播 CAC 基本原理

组播 CAC 是在二层组播的基础上通过控制二层组播转发表的建立来实现的。而二层转发表的建立和删除则是根据 IGMP Report 报文、IGMP leave 报文生成。在使用 IGMP、IGMP Snooping 协议开展组播业务时，都可以进行组播 CAC 的控制。CAC 控制二层组播转发表的生成，在超出设置的限制时，就不允许生成组播转发表项，从而保证了设备处理能力、链路带宽的控制。

组播 CAC 功能根据控制的内容的不同，可以分为：

- 限制组播组数量
- 限制组播组带宽
- 限制节目组内组播组数量
- 限制节目组内组播组带宽

根据控制的维度，组播 CAC 又可以分为：

- 全局组播 CAC 限制
- VLAN 组播 CAC 限制
- 端口组播 CAC 限制
- 端口+VLAN 组播 CAC 限制
- VSI 组播 CAC 限制
- 子接口组播 CAC 限制
- PW 组播 CAC 限制

各种不同网络类型中的组播 CAC 功能相互并列，用户可以根据需要选择相应配置。

配置组播组成员数量限制，即在建立组播成员转发表项时对组播组成员数量进行控制。这样，可以保证组播组的成员数量不超过设定值。如果某些组播组成员不需要限制，则可以配置 ACL 将这些组播组成员排除在外。同样也可以通过配置带宽限制，将某些组播组成员排除在外。

同时组播 CAC 还可以为特定的业务提供商定义各自的规则（配置节目组的数量、带宽限制），满足其业务需求。

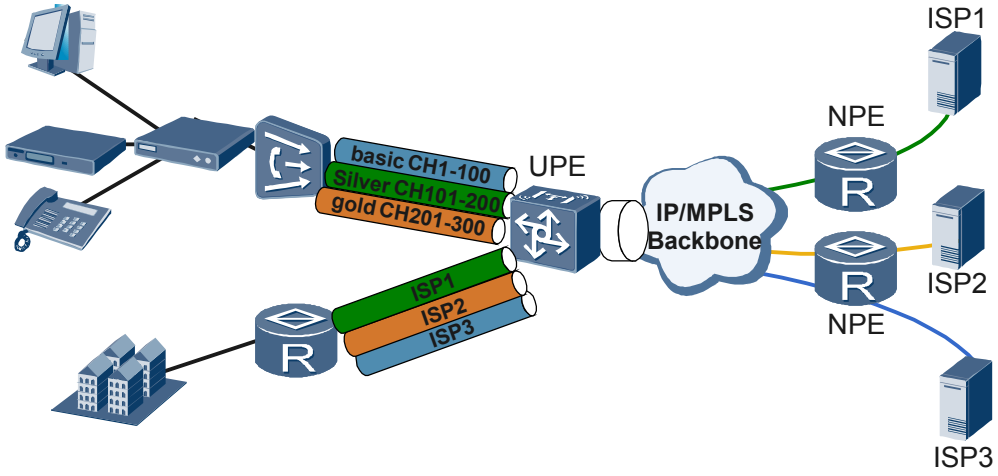
具体的应用见下面应用场景分析，此处不再赘述。

11.4 应用

11.4.1 IPTV 典型组网

IPTV 组网

图 11-1 IPTV 典型的组网



组播 CAC 是 IPTV 解决方案的一部分，主要用于控制 IPTV 的节目数量及其带宽，避免出现带宽需求超出接入汇聚网带宽的情况，保证大多数用户的服务质量。

运营商通过 ME 网络提供 IPTV 业务，根据汇聚网的不同方案，在 UPE 上部署 IGMP 或 IGMP Snooping， 建立组播转发表，指导组播数据流。同时在运营商充分评估其网络负载能力及需求后，可以根据自身需要和组网在 UPE 上部署基于 VLAN 的组播 CAC 功能或者基于 VSI 的组播 CAC 功能。

UPE 基于 VLAN 的组播 CAC 的应用

- 组播组数量限制

  在端口、端口+VLAN 配置组播组数量限制，在建立表项时进行数量控制。这样，可以保证组播组的数量不超过设定值。如果某些组不需要限制，则可以通过 ACL 策略将这些组排除在外。

  在处理 IGMP Report 报文，进行表项添加时，每添加一条表项则 CAC 各个维度下统计计数随之增加。如果 CAC 计数统计值已经达到 CAC 限制，则不建立表项。

  在处理 IGMP Leave 报文或者表项老化时，表项删除，更新 CAC 统计计数。

  如果各端口或端口+VLAN 的组播组数量没有超出限制，但各端口的组播组都不相同，则设备总的组播组可能很多，因此，有必要进行 VLAN、全局的组播组数量限制。

- 组播组带宽限制

  在各个组播组的带宽确定且差异不大的情况下，组播流量的带宽占用也就基本确定。例如：各组播组的带宽均为 4Kbit/s，则 20 个组播组，带宽就为 80Kbit/s。但是如果部分组播组的带宽为 4Kbit/s，部分组播组带宽为 18Kbits/s，那么 20 个组播组，总的带宽变动范围则不确定。这种情况下，单纯按照组播组数量控制，实际上就起不到控制带宽的作用了，所以有必要进行带宽的限制。

  在端口或端口+VLAN 上配置组播带宽限制。

在处理 IGMP Report 报文，进行表项添加时，每添加一条表项 CAC 各个维度下带宽统计随之增加。如果统计带宽超过限制值，则不生成表项。如果带宽统计值未超过限制，则生成表项。

在处理 IGMP Leave 报文或者表项老化时，更新 CAC 带宽计数。

同数量限制一样，也有必要支持全局的组播组带宽限制。

- 节目组内的组播组限制

  如果运营商网络是给多个节目组（既不同内容提供商）服务的，则需要针对不同的节目组进行带宽或组播组数量限制。

  在 IGMP Snooping 组播表项生成时，需要检查组播组属于哪个节目组的地址范围，再看该地址范围是否有 CAC 限制，如果超过限制，则不生成表项。检查所有匹配的地址范围段后，如果都没有超过限制，才允许生成表项。

UPE 基于 VSI 的组播 CAC 的应用

VSI 下组播 CAC 的应用和 VLAN 下应用是类似的。应用场景从 VLAN 变成了 VSI，增加了子接口、PW 等应用。

- 组播组数量限制

  在 VSI、子接口或者 PW 配置组播组数量，在建立表项时进行数量控制。这样，在组播组数量达到限制值后，新的组播组添加不生成表项。如果某些组不需要限制，则可以配置 ACL 规则将这些组排出在外。

- 组播组带宽限制

  在 VSI 或子接口、PW 上配置组播带宽限制。

  具体处理和 VLAN 方式类似。不再重复。

- 节目组内的组播组限制

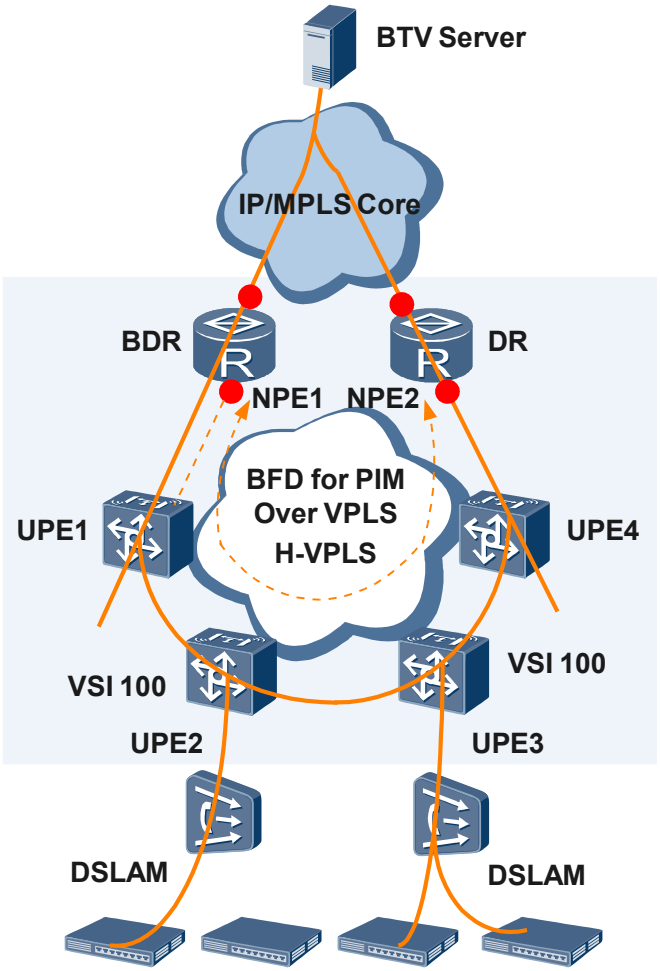
  如果运营商网络是给多个节目组服务的，则需要针对不同的节目组进行带宽或组播组数量限制。

  和 VLAN 方式类似，不再重复。

11.4.2 H-VPLS 典型组网

H-VPLS 组网

图 11-2 组播 H-VPLS 组网



在 UPE 间物理拓扑是环网的情况下，使用 H-VPLS 组网（称为“菊花链”）可以减少 UPE 之间组播复制的流量。

在 H-VPLS 组网中，来自下游方向 UPE 的 IGMP Report 报文是通过 PW 转发过来的，组播转发表中，出接口也包括 PW。因此有必要对 PW 进行组播 CAC 的限制。

基于 PW 组播 CAC 的应用

H-VPLS 组网下组播 CAC 支持 PW 上的限制，包括：

- 支持各 PW 上组播组数量限制。
- 支持各 PW 上组播组带宽限制。
- 支持各 PW 上区分内容提供上的组播组带宽限制。

在 PW 上实现组播 CAC 控制的功能和上面应用场景中 UPE 基于 VSI 的组播 CAC 在 PW 下应用是相同的，此处不再重复描述。

同时，如果各 PW 组播组统计值没有超出限制值，但各 PW 的组播组都不相同，则设备总的组播组可能很多，因此，有必要进行 VSI、全局的组播组限制。

11.5 术语与缩略语

术语

术语	解释
CAC	Call Admission Control，接入管理控制。是指在连接接收者的最后一跳组播路由器、中间组播路由器及连接源的第一跳组播路由器上配置策略限制组播路由器允许创建的组播表项，使运营商在 IP 核心网可以对用户接入的组播组数量进行控制。

缩略语

缩略语	英文全称	中文全称
CAC	Call Admission Control	接入管理控制
IGMP	Internet Group Management Protocol	因特网组管理协议
VLAN	Virtual Local Area Network	虚拟局域网
VSI	Virtual Switch Instance	虚拟交换实例
PW	Pseudo Wire	伪线

12 组播 trunk 负载分担

关于本章

12.1 介绍

定义

二层组播 Trunk 负载分担，在 VLAN/VSI 下不配置 IGMP Snooping 的情况下，支持组播流量在 Trunk 链路各个成员接口之间，基于组播组做负载分担。

目的

使能 IGMP Snooping 时，设备支持不同组播流量在 Trunk 各个成员端口之间做负载分担，但使能 IGMP Snooping 需要监听协议报文，需要软件对协议报文进行处理，加重设备的负担。部分运营商在部署网络时为了不影响设备性能，不在关键节点配置 IGMP Snooping。

不使能 IGMP Snooping 时，组播流量在 VLAN/VSI 内以广播方式转发。组播流量只会在 Trunk 的一个成员端口转发，不能充分利用带宽。在使能二层组播 Trunk 负载分担功能后，组播流量在 Trunk 链路各个成员接口上负载分担。

受益

运营商受益

组播 Trunk 负载分担给运营商带来的收益：

- 提升网络带宽利用率，充分利用 Trunk 链路的作用，运营商部署网络更灵活。
- 避免处理大量协议报文，提升设备性能。

用户受益

不涉及。

12.2 参考标准和协议

无

12.3 原理描述

12.3.1 实现原理

在部署组播业务的二层网络中，组播流量在 VLAN/VSI 内以广播方式转发，不按照组播组区分，广播报文只会在 Trunk 的一个成员口转发。

本特性采用命令行配置，命令行使能后，上层软件按照转发引擎的哈希规则预先下发转发表项，转发引擎对于不同组播组的报文，按照报文的目的 MAC 进行哈希运算，根据结果选择不同的表项进行转发。如果出接口是非 Trunk 接口，则组播报文的转发结果和不做哈希是一样的，只在此接口转发；如果出接口是 Trunk 接口，经过哈希运算后，不同组播组的报文会选择不同的 Trunk 成员接口进行转发。从而实现组播报文在 Trunk 链路上的负载分担，并且不影响未知单播、保留组播组报文的处理。

该特性需要转发引擎在转发组播报文时根据目的 MAC 做哈希，和配置 IGMP Snooping 的转发行为不同，所以和 IGMP Snooping 功能互斥，但不同的 VLAN/VSI 间该特性可以和 IGMP Snooping 同时使用。

12.3.2 协议流程

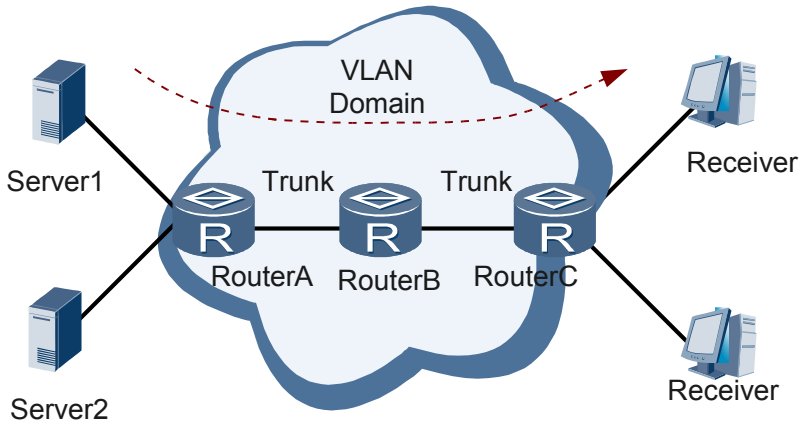
无

12.3.3 组网应用

组播 Trunk 负载分担的组网应用有 VLAN、VPLS、H-VPLS 等方式，支持 VPLS 组网应用支持 QinQ 接入，同时支持和 IGMP Snooping 基于不同的 VLAN、VSI 同时使用。

在 VLAN 中应用

图 12-1 组播 Trunk 负载分担 VLAN 组网



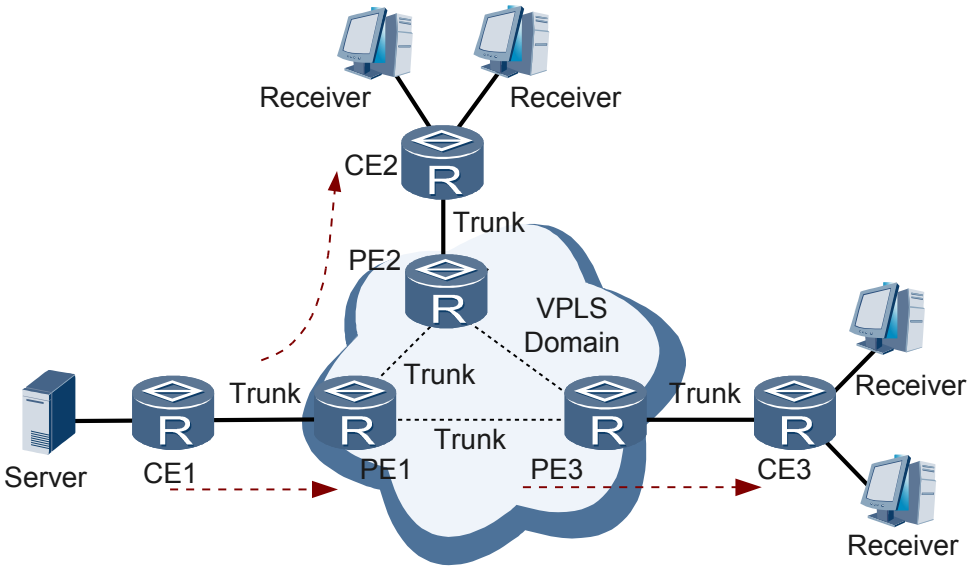
如图 12-1 所示：

- 在 VLAN 域中部署组播业务，路由器之间使用 Trunk 链路提升带宽和可靠性。
- 在 RouterA、RouterB 上使能组播 Trunk 负载分担功能（RouterC 连用户的端口不是 Trunk 链路，不需要使能组播 Trunk 负载分担功能）。
- 从组播服务器发往接收者的不同组播组的组播报文在 VLAN 域内的 Trunk 链路各成员口之间做负载分担，RouterA、RouterB 上 Trunk 每个成员口都有组播出流量。

在 VPLS 中应用

- 组播 Trunk 负载分担支持在 VPLS 组网中的应用，在 PE 节点的 Trunk 链路上负载分担，同时支持 AC 侧和 PW 侧都是 Trunk 的场景。

图 12-2 组播 Trunk 负载分担 VPLS 组网



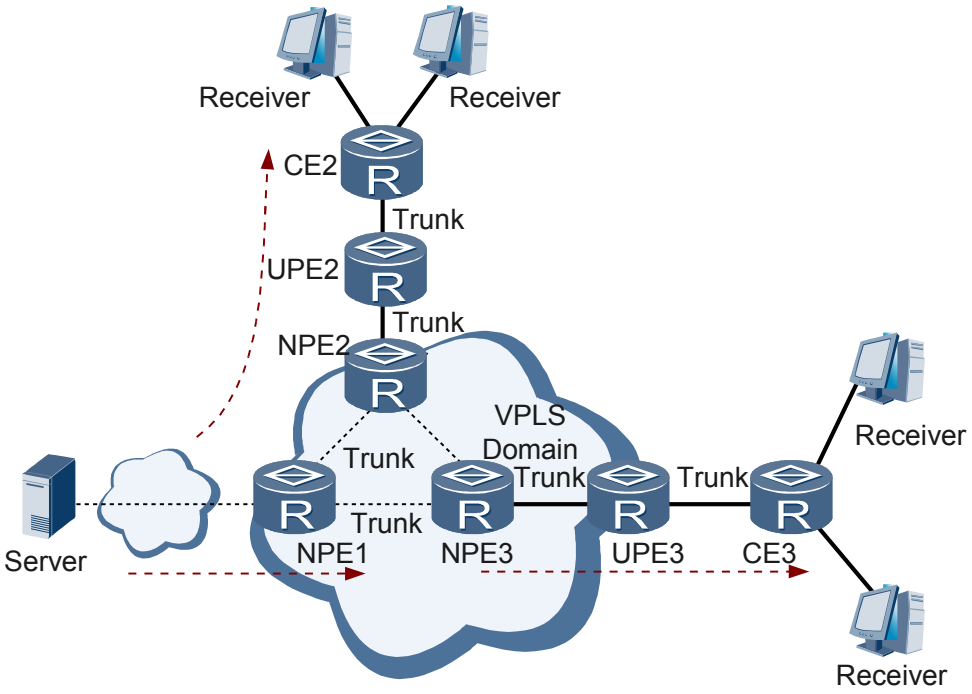
如图 12-2 所示：

- 在 VPLS 域中部署组播业务，路由器之间使用 Trunk 链路提升带宽和可靠性。
- CE1 连组播服务器，接入 VPLS，CE2 和 CE3 最为组播接收者。
- 同时支持 AC 侧和 PW 侧都是 Trunk 链路。
- 连接组播源的 CE1 上使能组播 Trunk 负载分担特性，组播源发往 PE1 的不同组播组的报文在 CE1-PE1 之间的 Trunk 链路上做负载分担；
- PE1 上使能组播 Trunk 负载分担特性后，发往其他 PE 的不同组播组的报文在 PW 侧的 Trunk 链路上做负载分担；
- PE2 上使能组播 Trunk 负载分担特性后，发往 CE2 的不同组播组的报文在 PE2-CE2 之间的 Trunk 链路上做负载分担。

在 H-VPLS 中应用

- 组播 Trunk 负载分担支持在 H-VPLS 组网中的应用，支持 UPE 节点、NPE 节点的 Trunk 链路的负载分担，同时支持 AC 侧、PW 侧及 UPE 和 NPE 之间都使用 Trunk 链路的场景。

图 12-3 组播 Trunk 负载分担 H-VPLS 组网



如图 12-3 所示：

- 在 H-VPLS 域中部署组播业务，路由器之间使用 Trunk 链路提升带宽和可靠性。
- 在 NPE1 上使能组播 Trunk 负载分担特性，从组播源发往其他 NPE 的不同组播组的流量在 PW 侧的 Trunk 链路上做负载分担。
- 在 NPE2 上使能组播 Trunk 负载分担特性，从 NPE2 发往 UPE2 的不同组播组的流量在 NPE2-UPE2 之间的 Trunk 链路上做负载分担。
- 在 UPE2 上使能组播 Trunk 负载分担特性，从 UPE2 发往本 Site 接收者的不同组播组的流量在 UPE2-CE2 之间的 Trunk 链路上做负载分担。

12.3.4 计费和话单

无

12.3.5 性能统计

无

12.4 应用

无

12.5 术语与缩略语

术语

无

缩略语

缩略语	英文全称	中文全称
VLAN	Virtual Local Area Network	虚拟局域网
VSI	Virtual Switch Instance	虚拟交换实例
IGMP	Internet Group Management Protocol	Internet 组管理协议

13 组播安全

关于本章

13.1 介绍

随着 Internet 的不断发展，网络中交互的各种数据、语音和视频信息越来越多。同时，新兴的电子商务、网上会议、网上拍卖、视频点播、远程教学等服务也在逐渐兴起。这些服务大多符合点对多点的模式，对信息安全性、有偿性、网络带宽提出了较高的要求。

为解决 IP 网络中点对多点的数据传输问题，除了实现各组播协议的基本功能外，还实现了组播成员管理、组播报文转发、域内组播路由和域间组播路由等功能。为了保证组播业务的安全性，又并设计了组播安全的管理功能，提供了限制、过滤、认证等手段，确保组播业务正常运行。

13.2 参考标准和协议

无

13.3 原理描述

13.3.1 组播表项总数限制

组播表项总数限制指对 IGMP、MLD、MSDP、PIM 协议表项的限制，禁止设备创建过多的组播表项，消耗设备资源。

IGMP/MLD 接口表项限制

支持接口配置 IGMP/MLD 加入的表项限制。

有关“IGMP/MLD 接口表项限制”的详细内容，请参见 **IGMP 策略控制**中的“IGMP-Limit”。

IGMP/MLD 单实例表项限制

支持对当前实例所有接口可创建的 IGMP/MLD 表项数量进行限制。

有关“IGMP/MLD 单实例表项限制”的详细内容，请参见 **IGMP 策略控制**中的“IGMP-Limit”。

IGMP/MLD 全局表项限制

支持对所有实例的所有接口可创建的 IGMP/MLD 表项数量进行限制。

有关“IGMP/MLD 全局表项限制”的详细内容，请参见 **IGMP 策略控制**中的“IGMP-Limit”。

PIM 表项限制

支持对设备可创建的 PIM 表项数量进行限制。

MSDP SA 消息缓存限制

MSDP SA 消息缓存用于存储 MSDP 对等体转发的 SA 消息中的(S, G, RP)信息。

MSDP SA 消息缓存的（S,G）个数限制，分为三种限制方式：

- 单实例下缓存的（S,G）表项个数限制
- 所有配置 MSDP 的设备缓存的（S,G）表项总数限制
- 每个 MSDP 对等体缓存的（S,G）表项个数限制

13.3.2 组播表项出接口限制

支持转发平面限制每个组播表项的最大出接口数量，从而控制组播数据的复制。

13.3.3 组播协议状态限制

组播协议状态限制主要用来限制各种组播协议的状态维护数量，防止非法报文对设备的资源消耗。主要包括 PIM 邻居数量、组播 VPN 数量、BSR 数量以及 C-RP 数量的控制。

PIM 邻居个数限制

PIM 协议报文的发送和接收都是建立在 PIM 邻居的基础上的。

PIM 邻居个数限制是指设备某接口上能够建立的 PIM 邻居个数的最大限制，防止接口建立太多 PIM 邻居导致设备无法正常运行。

C-RP 个数限制

支持限制当选 BSR 允许记录的 C-RP 的最大个数。

管理域 BSR 个数限制

管理域 BSR 个数限制包括单台设备支持的管理域 BSR 数量限制和单实例下每个管理域存储的 RP 数量限制。

组播 VPN 个数限制

设备通过配置 VPN 来实现私网的划分。各个 VPN 之间独立运行组播功能，将私网组播数据经公网透传到对端。对设备运行的组播 VPN 数量进行限制，可防止设备超载。

13.3.4 组播 CAC

组播 CAC 包括对带宽和 PIM 表项的全局限制和出接口限制。

组播 CAC 可通过在设备上限制组播协议层建立组播分发树，不引入超过转发能力的冗余流量。

为便于用户管理组播组，按节目划分不同的节目组，为每个节目组配置可以允许的组播表项个数和带宽。

13.3.5 组播过滤策略

在设备上配置各种策略，对协议报文进行过滤，可达到拒绝接收非法报文以及维护表项状态的目的。

IGMP/MLD group policy

支持在接口上配置 IGMP/MLD 组播组的过滤器，限制主机能够加入的组播组范围。

有关“IGMP/MLD group policy”的详细内容，请参见 **IGMP 策略控制**中的“IGMP Group-Policy”。

Source policy

用来对设备接收的组播数据报文根据源或源/组进行过滤。

SSM policy

用来改变某实例下的 SSM 组地址范围。

通过配置 SSM policy 可以指定 PIM SSM 组地址范围，所有使能 PIM-SM 协议的接口将会认为属于该范围内的组播组采用了 PIM-SSM 模式。

有关“SSM policy”的详细内容，请参见 **SSM Mapping**。

BSR policy

用来限定合法 BSR 地址范围，使设备丢弃来自该地址范围之外的 BSR 报文，从而防止 BSR 欺骗。

C-RP policy

用来限定合法的 C-RP 地址范围及其服务的组播组地址范围，使 BSR 丢弃来自该地址范围之外的 C-RP 消息，从而防止 C-RP 欺骗。

Register policy

通过配置 Register 报文过滤规则，接受或拒绝和规则匹配的 Register 报文，防止非法 Register 报文的攻击。

MSDP SA policy

用来配置接收或转发 SA 消息的过滤列表。

在接收来自指定 MSDP 对等体的 SA 消息或向指定 MSDP 对等体转发 SA 消息时，对其通告的（S，G）转发项进行源或源/组过滤，从而实现在接收和转发 SA 消息时对源或源/组消息传播的控制。

MSDP SA Request policy

用来配置设备响应由指定 MSDP 对等体发出的 SA 请求的策略。一旦 SA 请求通过过滤，立即回复 SA 消息。

MSDP SA Import source policy

用来配置在 MSDP 创建 SA 消息时，限制本域内被通告的活动源信息。

在创建 SA 消息时，对设备通告的（S，G）转发表项进行源或源/组过滤，从而实现在创建 SA 消息时对源或源/组消息传播的控制。

Multicast Boundary

网络中每个组播组对应的组播信息都需要在一个确定的范围内传递。支持在接口上配置 Multicast Boundary（组播边界）定义某个组播组数据的转发范围，以形成一个封闭的组播转发区域。当设备接口配置了针对某组播组的转发边界以后，将不能再发出或接收该组播组的报文。

BSR Boundary

为了达到对网络的精细化管理，通过在边界设备上配置 BSR Boundary 可划分 PIM 域的范围。

IGMP Host Address policy

为了提高接收端信息的可靠性，允许用户在设备接口配置 IGMP 主机地址过滤策略，用来过滤 IGMP Report 报文的主机地址。

有关“IGMP Host Address policy”的详细内容，请参见 **3.3.10 IGMP 主机地址过滤**。

PIM Neighbor policy

为了防止某些未知设备参与 PIM 协议，阻止本设备成为 DR，需要过滤 PIM 邻居。配置此功能后，接口只与符合过滤规则的地址建立邻居关系，删除不符合过滤规则的邻居。

PIM Join policy

接口上接收的 Join/Prune 消息中包含 Join 信息和 Prune 信息。配置此功能后，设备根据符合过滤规则的 Join 信息建立 PIM 表项，从而防止非法用户加入。

PIM Silent

为了避免恶意主机模拟 PIM Hello 报文攻击设备，可以在直连用户的接口上配置 PIM Silent，将接口设置为 PIM 消极模式。接口进入消极状态后，禁止接收和转发任何 PIM 协议报文，删除该接口上的所有 PIM 邻居及 PIM 状态机，并自动成为 DR。同时，该接口上的 IGMP 功能不受影响。

有关“PIM Silent”的详细内容，请参见 **2.3.2 PIM-SM** 中的“PIM Silent 的基本原理”。

13.3.6 组播协议报文防攻击

在某些应用中，设备只需要使用其需要的部分协议，对于不需要的上层协议，其协议报文是不需要上送到 CPU 处理的，转发平面感知上层开启的协议，对于开启服务的协议上送其协议报文，否则丢弃。可降低协议层受到攻击的风险，提高系统的安全性。

目前组播各协议都支持报文防攻击特性，包括 IGMP、MLD、PIM 和 MSDP 协议。

13.3.7 组播安全认证

组播安全认证主要包括 MSDP 支持 MD5/Key-chain 认证以及 IPv6 PIM IPSec。

MSDP 支持 MD5/Key-chain 认证

- MSDP MD5 认证：通过在 MSDP 对等体上配置 MD5 认证功能，只有对等体两端都配置了相同的 MD5 密码认证，才能建立 MSDP 对等体关系及报文交互。
- MSDP Key-chain 认证：支持用户为每条 TCP 链接配置一组密码，每个密码可以设置不同加密算法和有效期限，且密码可以随时更换。防止与非法用户建立 TCP 链接及接收非法报文。

有关“MSDP 支持 MD5/Key-chain 认证”的详细内容，请参见 **MSDP 支持 MD5/Key-chain 认证**。

IPv6 PIM IPSec

IPv6 PIM IPSec 利用 IPSec 提供的一整套安全保护机制对 IPv6 PIM 协议报文的发送和接收进行认证处理，防止伪造的 IPv6 PIM 协议报文对设备进行非法攻击。

有关“IPv6 PIM IPSec”的详细内容，请参见 **PIM 安全性**中的“IPv6 PIM IPSec”。

13.4 应用

13.4.1 网络安全保障措施

对于网络安全，组播目前提供的机制包括：

- reject-data-inbound：阻止组播流量从非法接口进入网络
- PIM Silent Interface：阻止在接入侧建立 PIM 协议状态
- Boundary：阻止组播数据和组播报文从组播边界发送和接收

13.4.2 协议层安全保障措施

组播目前提供的机制如下：

- MSDP MD5、keychain 认证：阻止伪造的 MSDP 对等体与设备建立 MSDP 邻居关系
- MSDP SA policy、MSDP SA Request policy：阻止伪造的 MSDP SA 消息

- PIM Neighbor Policy：阻止非法 PIM 邻居建立
- C-RP、BSR Policy：阻止学习非法的 RP 和 BSR 地址
- Register、Source、Join/Prune Policy、 IGMP group policy：阻止学习非法的组播表项
- 组播协议报文防攻击：在底层丢弃上层协议不需要处理的协议报文，或优先上送某些协议报文

13.4.3 设备安全保障措施

组播目前提供的机制如下：

- IGMP、PIM 表项限制、MSDP 对等体限制、MSDP SA 消息缓存限制：防止内存被无限占用
- 组播 CAC：防止链路带宽被过度占用影响业务正常进行
- CPCAR：控制协议报文上送的内部带宽占用
- 报文限速：控制处理协议报文的 CPU 占用率

13.5 术语与缩略语

无