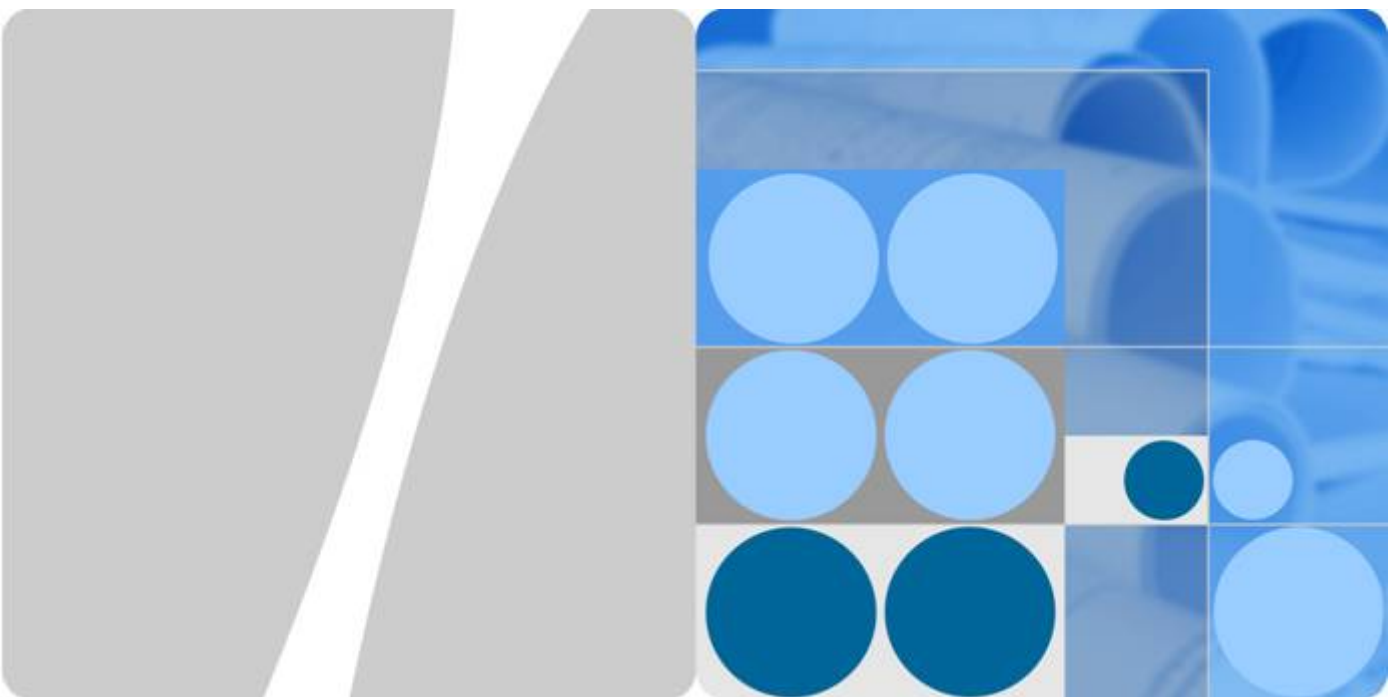




## 华为 Secoway USG2000&5000 系列统一安全网关

### 特性描述

文档版本 01

发布日期 2012-07-06

版权所有 © 华为技术有限公司 2012。保留一切权利。

非经本公司书面许可，任何单位和个人不得擅自摘抄、复制本文档内容的部分或全部，并不得以任何形式传播。

## 商标声明



HUAWEI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://enterprise.huawei.com>

客户服务邮箱：[ChinaEnterprise\\_TAC@huawei.com](mailto:ChinaEnterprise_TAC@huawei.com)

客户服务电话：4008229999

# 前言

## 产品版本

与本文档相对应的产品版本如下所示。

| 产品名称                                                                    | 产品版本     |
|-------------------------------------------------------------------------|----------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | V300R001 |

## 读者对象

本文档介绍了 Secoway USG2110-X/2100/2200/5100/5500、USG2100/2200/5100 BSR/HSR（以下简称 USG）的功能特性。

本文档提供了 USG 的特性介绍、原理描述和相关参考文档。

本文档主要适用于以下工程师：

- 网络规划工程师
- 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 | 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 | 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。 |

| 符号                                                                                   | 说明                                             |
|--------------------------------------------------------------------------------------|------------------------------------------------|
|  注意 | 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 | 表示能帮助您解决某个问题或节省您的时间。                           |
|  说明 | 表示是正文的附加信息，是对正文的强调和补充。                         |

## 命令行格式约定

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

## 修订记录

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

### 文档版本 01 (2012-07-06)

第一次正式发布。



# 目 录

|                    |    |
|--------------------|----|
| 前 言.....           | ii |
| 1 系统管理.....        | 1  |
| 1.1 FTP.....       | 3  |
| 1.1.1 介绍.....      | 3  |
| 1.1.2 可获得性.....    | 3  |
| 1.1.3 原理描述.....    | 3  |
| 1.2 TFTP.....      | 4  |
| 1.2.1 介绍.....      | 4  |
| 1.2.2 参考标准和协议..... | 4  |
| 1.2.3 可获得性.....    | 4  |
| 1.2.4 原理描述.....    | 5  |
| 1.3 Telnet.....    | 7  |
| 1.3.1 介绍.....      | 7  |
| 1.3.2 可获得性.....    | 8  |
| 1.3.3 原理描述.....    | 8  |
| 1.4 SSH.....       | 11 |
| 1.4.1 介绍.....      | 11 |
| 1.4.2 可获得性.....    | 12 |
| 1.4.3 原理描述.....    | 12 |
| 1.5 信息中心.....      | 15 |
| 1.5.1 简介.....      | 15 |
| 1.5.2 参考标准和协议..... | 16 |
| 1.5.3 可获得性.....    | 16 |
| 1.5.4 原理描述.....    | 16 |
| 1.6 SNMP.....      | 21 |
| 1.6.1 介绍.....      | 21 |
| 1.6.2 参考标准和协议..... | 22 |
| 1.6.3 可获得性.....    | 23 |
| 1.6.4 应用场景.....    | 23 |
| 1.6.5 原理描述.....    | 24 |
| 1.7 LLDP.....      | 31 |
| 1.7.1 介绍.....      | 31 |

|                          |           |
|--------------------------|-----------|
| 1.7.2 参考标准和协议.....       | 31        |
| 1.7.3 可获得性.....          | 31        |
| 1.7.4 原理描述.....          | 32        |
| 1.7.4.1 LLDP 基本概念.....   | 32        |
| 1.7.4.2 LLDP 工作机制.....   | 35        |
| 1.8 CWMP.....            | 36        |
| 1.8.1 介绍.....            | 36        |
| 1.8.2 可获得性.....          | 37        |
| 1.8.3 原理描述.....          | 38        |
| 1.9 NetStream.....       | 39        |
| 1.9.1 介绍.....            | 39        |
| 1.9.2 可获得性.....          | 40        |
| 1.9.3 原理描述.....          | 41        |
| 1.10 NTP.....            | 44        |
| 1.10.1 介绍.....           | 44        |
| 1.10.2 参考标准和协议.....      | 44        |
| 1.10.3 可获得性.....         | 44        |
| 1.10.4 原理描述.....         | 44        |
| 1.11 NQA.....            | 48        |
| 1.11.1 介绍.....           | 48        |
| 1.11.2 参考标准和协议.....      | 50        |
| 1.11.3 可获得性.....         | 50        |
| 1.11.4 原理描述.....         | 51        |
| 1.12 Ping 和 Tracert..... | 56        |
| 1.12.1 介绍.....           | 56        |
| 1.12.2 参考标准和协议.....      | 56        |
| 1.12.3 可获得性.....         | 56        |
| 1.12.4 原理描述.....         | 57        |
| 1.13 探测 PMTU.....        | 58        |
| 1.13.1 简介.....           | 58        |
| 1.13.2 参考标准和协议.....      | 58        |
| 1.13.3 可获得性.....         | 58        |
| 1.13.4 应用场景.....         | 59        |
| 1.13.5 原理描述.....         | 59        |
| <b>2 可靠性.....</b>        | <b>61</b> |
| 2.1 双机热备份.....           | 62        |
| 2.1.1 介绍.....            | 62        |
| 2.1.2 参考标准和协议.....       | 63        |
| 2.1.3 可获得性.....          | 63        |
| 2.1.4 应用场景.....          | 64        |
| 2.1.4.1 主备备份.....        | 64        |
| 2.1.4.2 负载分担.....        | 68        |

|                                     |            |
|-------------------------------------|------------|
| 2.1.5 协议原理描述.....                   | 71         |
| 2.1.5.1 协议引入.....                   | 71         |
| 2.1.5.2 VRRP.....                   | 73         |
| 2.1.5.3 VGMP.....                   | 74         |
| 2.1.5.4 HRP.....                    | 78         |
| 2.1.6 典型组网分析.....                   | 81         |
| 2.1.6.1 组网一：业务接口工作在三层，上下行连接交换机..... | 82         |
| 2.1.6.2 组网二：业务接口工作在三层，上下行连接路由器..... | 83         |
| 2.1.6.3 组网三：业务接口工作在二层，上下行连接路由器..... | 84         |
| 2.1.6.4 组网四：业务接口工作在二层，上下行连接交换机..... | 86         |
| 2.2 IP-link.....                    | 87         |
| 2.2.1 介绍.....                       | 87         |
| 2.2.2 可获得性.....                     | 88         |
| 2.2.3 原理描述.....                     | 88         |
| 2.3 Link-group.....                 | 90         |
| 2.3.1 介绍.....                       | 90         |
| 2.3.2 可获得性.....                     | 90         |
| 2.3.3 原理描述.....                     | 91         |
| 2.4 BFD.....                        | 91         |
| 2.4.1 介绍.....                       | 91         |
| 2.4.2 参考标准和协议.....                  | 92         |
| 2.4.3 可获得性.....                     | 92         |
| 2.4.4 应用场景.....                     | 92         |
| 2.4.4.1 BFD 与 OSPF 联动.....          | 92         |
| 2.4.4.2 BFD 与静态路由联动.....            | 93         |
| 2.4.4.3 BFD 与 FRR 联动.....           | 95         |
| 2.4.4.4 BFD 与 DHCP 联动.....          | 96         |
| 2.4.4.5 BFD 与策略路由联动.....            | 97         |
| 2.4.5 原理描述.....                     | 98         |
| 2.4.5.1 BFD 报文.....                 | 98         |
| 2.4.5.2 BFD 检测机制.....               | 101        |
| 2.4.5.3 BFD 会话管理.....               | 103        |
| <b>3 虚拟防火墙.....</b>                 | <b>106</b> |
| 3.1 介绍.....                         | 107        |
| 3.2 可获得性.....                       | 107        |
| 3.3 原理描述.....                       | 107        |
| <b>4 接入.....</b>                    | <b>109</b> |
| 4.1 链路聚合.....                       | 110        |
| 4.1.1 介绍.....                       | 110        |
| 4.1.2 参考标准和协议.....                  | 110        |
| 4.1.3 可获得性.....                     | 110        |

|                              |     |
|------------------------------|-----|
| 4.1.4 应用场景.....              | 111 |
| 4.1.5 原理描述.....              | 111 |
| 4.1.5.1 链路聚合协议（LACP） .....   | 111 |
| 4.1.5.2 链路聚合实现原理.....        | 112 |
| 4.2 PPP 和 MP.....            | 113 |
| 4.2.1 介绍.....                | 114 |
| 4.2.2 可获得性.....              | 114 |
| 4.2.3 原理描述.....              | 114 |
| 4.2.3.1 PPP 的基本原理.....       | 114 |
| 4.2.3.2 PPP 报文格式.....        | 115 |
| 4.2.3.3 PPP 的运行过程.....       | 118 |
| 4.2.3.4 PPP 的 PAP 验证协议.....  | 121 |
| 4.2.3.5 PPP 的 CHAP 验证协议..... | 125 |
| 4.2.3.6 PPP 的报文压缩.....       | 129 |
| 4.2.3.7 MP 的实现方式.....        | 129 |
| 4.2.3.8 MP 的协商过程.....        | 130 |
| 4.3 HDLC.....                | 130 |
| 4.3.1 介绍.....                | 130 |
| 4.3.2 参考标准和协议.....           | 131 |
| 4.3.3 可获得性.....              | 131 |
| 4.3.4 原理描述.....              | 131 |
| 4.3.4.1 HDLC 的基本原理.....      | 131 |
| 4.3.4.2 HDLC 的操作方式.....      | 132 |
| 4.3.4.3 HDLC 的应用特点.....      | 133 |
| 4.4 VLAN.....                | 134 |
| 4.4.1 介绍.....                | 134 |
| 4.4.2 参考标准和协议.....           | 135 |
| 4.4.3 可获得性.....              | 135 |
| 4.4.4 应用场景.....              | 136 |
| 4.4.4.1 VLAN 内通信.....        | 136 |
| 4.4.4.2 VLAN 间通信.....        | 136 |
| 4.4.5 原理描述.....              | 138 |
| 4.4.5.1 VLAN 基本概念.....       | 138 |
| 4.4.5.2 VLAN 通信原理.....       | 140 |
| 4.5 MSTP.....                | 143 |
| 4.5.1 介绍.....                | 143 |
| 4.5.2 参考标准和协议.....           | 144 |
| 4.5.3 可获得性.....              | 144 |
| 4.5.4 原理描述.....              | 144 |
| 4.5.4.1 MSTP 的基本原理.....      | 145 |
| 4.5.4.2 MSTP 报文格式.....       | 145 |
| 4.5.4.3 MSTP 的基本概念.....      | 148 |

|                               |            |
|-------------------------------|------------|
| 4.5.4.4 MSTP 的快速收敛机制.....     | 154        |
| 4.5.4.5 MSTP 的保护功能.....       | 156        |
| 4.6 PPPoE.....                | 157        |
| 4.6.1 介绍.....                 | 157        |
| 4.6.2 参考标准和协议.....            | 157        |
| 4.6.3 可获得性.....               | 157        |
| 4.6.4 原理描述.....               | 158        |
| 4.6.4.1 PPPoE 基本原理.....       | 158        |
| 4.6.4.2 Discovery 阶段.....     | 159        |
| 4.6.4.3 PPPoE Session 阶段..... | 163        |
| 4.7 WLAN.....                 | 163        |
| 4.7.1 介绍.....                 | 163        |
| 4.7.2 可获得性.....               | 164        |
| 4.7.3 原理描述.....               | 164        |
| 4.7.3.1 WLAN 的基本原理.....       | 164        |
| 4.7.3.2 802.11 协议族.....       | 165        |
| 4.7.3.3 常见术语.....             | 165        |
| <b>5 IP 业务.....</b>           | <b>167</b> |
| 5.1 IP 地址.....                | 169        |
| 5.1.1 介绍.....                 | 169        |
| 5.1.2 参考标准和协议.....            | 169        |
| 5.1.3 可获得性.....               | 169        |
| 5.1.4 原理描述.....               | 170        |
| 5.1.4.1 IP 地址分类.....          | 170        |
| 5.1.4.2 子网划分.....             | 172        |
| 5.1.4.3 IP 地址获取方式.....        | 173        |
| 5.1.4.4 IP 地址解析.....          | 174        |
| 5.2 ARP.....                  | 174        |
| 5.2.1 介绍.....                 | 174        |
| 5.2.2 参考标准和协议.....            | 175        |
| 5.2.3 可获得性.....               | 175        |
| 5.2.4 原理描述.....               | 176        |
| 5.2.4.1 ARP 报文.....           | 176        |
| 5.2.4.2 ARP 原理.....           | 177        |
| 5.2.4.3 静态 ARP.....           | 179        |
| 5.2.4.4 动态 ARP.....           | 180        |
| 5.2.4.5 代理 ARP.....           | 180        |
| 5.2.4.6 免费 ARP.....           | 180        |
| 5.2.4.7 授权 ARP.....           | 181        |
| 5.3 DNS.....                  | 181        |
| 5.3.1 介绍.....                 | 182        |
| 5.3.2 参考标准和协议.....            | 182        |

|                                                |     |
|------------------------------------------------|-----|
| 5.3.3 可获得性.....                                | 182 |
| 5.3.4 原理描述.....                                | 182 |
| 5.3.4.1 Internet 域名体系.....                     | 182 |
| 5.3.4.2 DNS 报文.....                            | 183 |
| 5.3.4.3 静态域名解析.....                            | 187 |
| 5.3.4.4 动态域名解析.....                            | 188 |
| 5.3.4.5 DNS Proxy.....                         | 190 |
| 5.3.4.6 DDNS.....                              | 190 |
| 5.4 DHCP.....                                  | 191 |
| 5.4.1 介绍.....                                  | 191 |
| 5.4.2 参考标准和协议.....                             | 192 |
| 5.4.3 可获得性.....                                | 193 |
| 5.4.4 应用场景.....                                | 193 |
| 5.4.4.1 DHCP 服务器.....                          | 193 |
| 5.4.4.2 DHCP 中继.....                           | 195 |
| 5.4.4.3 DHCP 客户端.....                          | 195 |
| 5.4.5 原理描述.....                                | 196 |
| 5.4.5.1 DHCP 报文.....                           | 196 |
| 5.4.5.2 DHCP 服务器和客户端的交互过程.....                 | 201 |
| 5.4.5.3 DHCP 客户端与 DHCP 服务器通过 DHCP 中继的交互过程..... | 203 |
| 5.4.5.4 地址池.....                               | 206 |
| 5.4.5.5 租期与定时器.....                            | 207 |
| 5.4.5.6 地址冲突探测.....                            | 208 |
| 5.5 DHCP Snooping.....                         | 208 |
| 5.5.1 介绍.....                                  | 208 |
| 5.5.2 参考标准和协议.....                             | 210 |
| 5.5.3 可获得性.....                                | 210 |
| 5.5.4 原理描述.....                                | 210 |
| 5.5.4.1 DHCP Server 仿冒者攻击.....                 | 210 |
| 5.5.4.2 中间人攻击与 IP/MAC Spoofing 攻击.....         | 211 |
| 5.5.4.3 改变 CHADDR 值的 DoS 攻击.....               | 213 |
| 5.5.4.4 Option82.....                          | 214 |
| 5.6 IP 性能.....                                 | 218 |
| 5.6.1 介绍.....                                  | 218 |
| 5.6.2 参考标准和协议.....                             | 218 |
| 5.6.3 可获得性.....                                | 218 |
| 5.6.4 原理描述.....                                | 218 |
| 5.6.4.1 IP 报文.....                             | 219 |
| 5.6.4.2 ICMP 报文.....                           | 221 |
| 5.6.4.3 差错和重定向 ICMP 报文处理.....                  | 225 |
| 5.6.4.4 Ping 和 Tracert.....                    | 227 |
| 5.6.4.5 报文保序.....                              | 227 |

|                             |            |
|-----------------------------|------------|
| 5.7 IPv6 基础.....            | 228        |
| 5.7.1 介绍.....               | 228        |
| 5.7.2 参考标准和协议.....          | 229        |
| 5.7.3 可获得性.....             | 229        |
| 5.7.4 原理描述.....             | 229        |
| 5.7.4.1 IPv6 的特点.....       | 229        |
| 5.7.4.2 IPv6 地址.....        | 230        |
| 5.7.4.3 IPv6 报文.....        | 233        |
| 5.7.4.4 ICMPv6 报文.....      | 237        |
| 5.8 IPv6 邻居发现.....          | 238        |
| 5.8.1 介绍.....               | 239        |
| 5.8.2 参考标准和协议.....          | 239        |
| 5.8.3 可获得性.....             | 240        |
| 5.8.4 原理描述.....             | 240        |
| 5.8.4.1 IPv6 邻居发现.....      | 240        |
| 5.8.4.2 IPv6 安全邻居发现.....    | 243        |
| 5.9 IPv6 性能.....            | 244        |
| 5.9.1 介绍.....               | 244        |
| 5.9.2 参考标准和协议.....          | 244        |
| 5.9.3 可获得性.....             | 245        |
| 5.9.4 原理描述.....             | 245        |
| 5.10 DHCPv6.....            | 246        |
| 5.10.1 介绍.....              | 246        |
| 5.10.2 参考标准和协议.....         | 247        |
| 5.10.3 可获得性.....            | 247        |
| 5.10.4 原理描述.....            | 247        |
| 5.10.4.1 DHCPv6 概述.....     | 247        |
| 5.10.4.2 DHCPv6 报文.....     | 249        |
| 5.10.4.3 DHCPv6 工作原理.....   | 253        |
| 5.11 IPv6 over IPv4 隧道..... | 256        |
| 5.11.1 介绍.....              | 256        |
| 5.11.2 参考标准和协议.....         | 256        |
| 5.11.3 可获得性.....            | 257        |
| 5.11.4 原理描述.....            | 257        |
| 5.12 IPv4 over IPv6 隧道..... | 258        |
| 5.12.1 介绍.....              | 258        |
| 5.12.2 参考标准和协议.....         | 258        |
| 5.12.3 可获得性.....            | 258        |
| 5.12.4 原理描述.....            | 259        |
| <b>6 IP 路由.....</b>         | <b>260</b> |
| 6.1 静态路由.....               | 262        |
| 6.1.1 介绍.....               | 262        |

|                              |     |
|------------------------------|-----|
| 6.1.2 可获得性.....              | 262 |
| 6.1.3 原理描述.....              | 262 |
| 6.2 RIP.....                 | 263 |
| 6.2.1 介绍.....                | 263 |
| 6.2.2 参考标准和协议.....           | 264 |
| 6.2.3 可获得性.....              | 264 |
| 6.2.4 原理描述.....              | 265 |
| 6.2.4.1 RIP-1 报文格式.....      | 265 |
| 6.2.4.2 RIP-2 报文格式.....      | 266 |
| 6.2.4.3 RIP-2 的验证报文格式.....   | 267 |
| 6.2.4.4 RIP 的路由数据库.....      | 267 |
| 6.2.4.5 RIP 定时器.....         | 268 |
| 6.2.4.6 RIP 运行和路由计算的过程.....  | 268 |
| 6.2.4.7 RIP 发送请求报文和响应报文..... | 270 |
| 6.2.4.8 RIP 报文的认证.....       | 270 |
| 6.2.4.9 水平分割.....            | 270 |
| 6.2.4.10 毒性逆转.....           | 271 |
| 6.2.4.11 触发更新.....           | 271 |
| 6.2.4.12 路由聚合.....           | 272 |
| 6.2.4.13 多进程和多实例.....        | 273 |
| 6.3 RIPvng.....              | 273 |
| 6.3.1 介绍.....                | 273 |
| 6.3.2 参考标准和协议.....           | 274 |
| 6.3.3 可获得性.....              | 274 |
| 6.3.4 原理描述.....              | 274 |
| 6.3.4.1 RIPvng 报文格式.....     | 274 |
| 6.3.4.2 定时器.....             | 277 |
| 6.3.4.3 水平分割.....            | 277 |
| 6.3.4.4 毒性逆转.....            | 277 |
| 6.3.4.5 触发更新.....            | 278 |
| 6.3.4.6 路由聚合.....            | 279 |
| 6.4 OSPF.....                | 279 |
| 6.4.1 介绍.....                | 279 |
| 6.4.2 参考标准和协议.....           | 280 |
| 6.4.3 可获得性.....              | 280 |
| 6.4.4 原理描述.....              | 280 |
| 6.4.4.1 OSPF 基础概念.....       | 280 |
| 6.4.4.2 OSPF 报文格式.....       | 288 |
| 6.4.4.3 OSPF 的 LSA 格式.....   | 294 |
| 6.4.4.4 OSPF 区域特性.....       | 300 |
| 6.4.4.5 邻居和邻接.....           | 303 |
| 6.4.4.6 网络类型和 DR 选举.....     | 308 |



|                                 |     |
|---------------------------------|-----|
| 6.4.4.7 OSPF 缺省路由和路由聚合.....     | 310 |
| 6.4.4.8 OSPF 报文验证.....          | 312 |
| 6.4.4.9 多进程和多实例.....            | 312 |
| 6.5 OSPFv3.....                 | 313 |
| 6.5.1 介绍.....                   | 313 |
| 6.5.2 参考标准和协议.....              | 313 |
| 6.5.3 可获得性.....                 | 313 |
| 6.5.4 原理描述.....                 | 314 |
| 6.5.4.1 OSPFv3 报文格式.....        | 314 |
| 6.5.4.2 OSPFv3 的 LSA 类型.....    | 314 |
| 6.5.4.3 OSPFv3 的网络类型和区域类型.....  | 315 |
| 6.5.4.4 OSPFv3 和 OSPF 协议比较..... | 316 |
| 6.6 IS-IS.....                  | 318 |
| 6.6.1 介绍.....                   | 318 |
| 6.6.2 参考标准和协议.....              | 319 |
| 6.6.3 可获得性.....                 | 319 |
| 6.6.4 原理描述.....                 | 319 |
| 6.6.4.1 IS-IS 的发展.....          | 319 |
| 6.6.4.2 IS-IS 协议地址结构.....       | 321 |
| 6.6.4.3 IS-IS 的报文结构.....        | 322 |
| 6.6.4.4 IS-IS 区域.....           | 329 |
| 6.6.4.5 IS-IS 网络类型.....         | 333 |
| 6.6.4.6 IS-IS 邻居建立过程.....       | 334 |
| 6.6.4.7 IS-IS 的 LSP 交互过程.....   | 336 |
| 6.6.4.8 IS-IS 对 IPv6 的支持.....   | 340 |
| 6.6.4.9 多进程和多实例.....            | 341 |
| 6.7 BGP.....                    | 341 |
| 6.7.1 介绍.....                   | 341 |
| 6.7.2 参考标准和协议.....              | 343 |
| 6.7.3 可获得性.....                 | 344 |
| 6.7.4 原理描述.....                 | 344 |
| 6.7.4.1 BGP 协议基本原理.....         | 344 |
| 6.7.4.2 BGP 报文格式.....           | 350 |
| 6.7.4.3 BGP 属性分类.....           | 356 |
| 6.7.4.4 BGP 路由聚合.....           | 365 |
| 6.7.4.5 BGP 对等体组.....           | 366 |
| 6.7.4.6 BGP 团体.....             | 366 |
| 6.7.4.7 BGP 路由反射器.....          | 366 |
| 6.7.4.8 BGP 联盟.....             | 370 |
| 6.7.4.9 BGP 衰减.....             | 372 |
| 6.7.4.10 BGP 安全性.....           | 373 |
| 6.7.4.11 BGP4+特性.....           | 373 |

|                                     |            |
|-------------------------------------|------------|
| 6.8 IP 单播策略路由.....                  | 374        |
| 6.8.1 介绍.....                       | 374        |
| 6.8.2 参考标准和协议.....                  | 374        |
| 6.8.3 可获得性.....                     | 375        |
| 6.8.4 特性增强.....                     | 375        |
| 6.8.5 应用场景.....                     | 375        |
| 6.8.6 原理描述.....                     | 376        |
| 6.8.6.1 策略路由.....                   | 376        |
| 6.8.6.2 FIB.....                    | 379        |
| 6.9 IPv6 单播策略路由.....                | 381        |
| 6.9.1 介绍.....                       | 381        |
| 6.9.2 参考标准和协议.....                  | 382        |
| 6.9.3 可获得性.....                     | 382        |
| 6.9.4 原理描述.....                     | 382        |
| 6.9.4.1 IPv6 FIB.....               | 382        |
| 6.9.4.2 策略路由工作原理.....               | 383        |
| 6.9.4.3 策略路由的应用.....                | 383        |
| 6.10 路由策略.....                      | 384        |
| 6.10.1 介绍.....                      | 384        |
| 6.10.2 可获得性.....                    | 384        |
| 6.10.3 原理描述.....                    | 385        |
| 6.10.3.1 路由策略的基本原理.....             | 385        |
| 6.10.3.2 路由策略与策略路由的比较.....          | 386        |
| 6.10.3.3 过滤器的比较.....                | 387        |
| 6.10.3.4 地址前缀列表的组成和匹配规则.....        | 387        |
| 6.10.3.5 地址前缀列表的应用.....             | 389        |
| 6.10.3.6 Route-Policy 的组成和匹配规则..... | 390        |
| 6.10.3.7 Route-Policy 应用简介.....     | 393        |
| <b>7 防火墙.....</b>                   | <b>396</b> |
| 7.1 安全区域.....                       | 398        |
| 7.1.1 介绍.....                       | 398        |
| 7.1.2 可获得性.....                     | 398        |
| 7.1.3 原理描述.....                     | 398        |
| 7.2 ACL.....                        | 402        |
| 7.2.1 介绍.....                       | 402        |
| 7.2.2 可获得性.....                     | 403        |
| 7.2.3 原理描述.....                     | 404        |
| 7.3 ACL6.....                       | 405        |
| 7.3.1 介绍.....                       | 405        |
| 7.3.2 可获得性.....                     | 405        |
| 7.3.3 IPv6 ACL.....                 | 406        |
| 7.4 安全策略.....                       | 407        |

|                                    |     |
|------------------------------------|-----|
| 7.4.1 介绍.....                      | 407 |
| 7.4.2 可获得性.....                    | 407 |
| 7.4.3 特性增强.....                    | 408 |
| 7.4.4 应用场景：通过域间安全策略进行安全检查.....     | 408 |
| 7.4.4.1 控制流量转发.....                | 408 |
| 7.4.4.2 控制设备访问.....                | 409 |
| 7.4.5 应用场景：通过域内安全策略进行安全检查.....     | 411 |
| 7.4.6 应用场景：通过接口上的包过滤规则控制报文的收发..... | 411 |
| 7.4.6.1 控制接口收发 IP 报文.....          | 411 |
| 7.4.6.2 控制接口接收的以太网帧.....           | 412 |
| 7.4.6.3 控制二层接口卡接收报文.....           | 413 |
| 7.4.7 原理描述.....                    | 414 |
| 7.4.7.1 安全策略的实现.....               | 414 |
| 7.4.7.2 安全策略与状态检测机制的关系.....        | 418 |
| 7.5 IPv6 包过滤.....                  | 420 |
| 7.5.1 介绍.....                      | 420 |
| 7.5.2 原理描述.....                    | 420 |
| 7.6 限流策略.....                      | 420 |
| 7.6.1 介绍.....                      | 420 |
| 7.6.2 可获得性.....                    | 421 |
| 7.6.3 应用场景.....                    | 421 |
| 7.6.3.1 每 IP 限流.....               | 421 |
| 7.6.3.2 整体限流.....                  | 422 |
| 7.6.3.3 保证带宽.....                  | 423 |
| 7.6.4 原理描述.....                    | 424 |
| 7.7 NAT.....                       | 425 |
| 7.7.1 介绍.....                      | 425 |
| 7.7.2 参考标准和协议.....                 | 426 |
| 7.7.3 可获得性.....                    | 426 |
| 7.7.4 原理描述.....                    | 426 |
| 7.7.4.1 NAT 的实现.....               | 426 |
| 7.7.4.2 NAT ALG 的实现.....           | 433 |
| 7.7.4.3 NAT 组网应用.....              | 436 |
| 7.8 NAT64.....                     | 439 |
| 7.8.1 介绍.....                      | 439 |
| 7.8.2 参考标准和协议.....                 | 439 |
| 7.8.3 可获得性.....                    | 440 |
| 7.8.4 原理描述.....                    | 440 |
| 7.8.4.1 NAT64 概述.....              | 440 |
| 7.8.4.2 NAT64 地址转换原理.....          | 441 |
| 7.8.4.3 NAT64 工作原理.....            | 442 |
| 7.9 负载均衡.....                      | 443 |

|                              |     |
|------------------------------|-----|
| 7.9.1 介绍.....                | 443 |
| 7.9.2 可获得性.....              | 444 |
| 7.9.3 原理描述.....              | 444 |
| 7.9.3.1 虚服务技术.....           | 444 |
| 7.9.3.2 服务器健康性检测.....        | 445 |
| 7.9.3.3 基于流的转发.....          | 445 |
| 7.10 攻击防范.....               | 445 |
| 7.10.1 介绍.....               | 445 |
| 7.10.2 可获得性.....             | 445 |
| 7.10.3 原理描述.....             | 446 |
| 7.11 IP 与 MAC 地址绑定.....      | 455 |
| 7.11.1 介绍.....               | 455 |
| 7.11.2 可获得性.....             | 456 |
| 7.11.3 原理描述.....             | 456 |
| 7.12 黑名单.....                | 461 |
| 7.12.1 介绍.....               | 461 |
| 7.12.2 可获得性.....             | 461 |
| 7.12.3 原理描述.....             | 462 |
| 7.13 会话表与长连接.....            | 464 |
| 7.13.1 介绍.....               | 464 |
| 7.13.2 可获得性.....             | 465 |
| 7.13.3 原理描述.....             | 465 |
| 7.14 Server-map 表与 ASPF..... | 469 |
| 7.14.1 介绍.....               | 469 |
| 7.14.2 可获得性.....             | 469 |
| 7.14.3 原理描述.....             | 469 |
| 7.15 端口映射.....               | 473 |
| 7.15.1 介绍.....               | 473 |
| 7.15.2 可获得性.....             | 474 |
| 7.15.3 原理描述.....             | 474 |
| 7.16 分片报文处理.....             | 475 |
| 7.16.1 介绍.....               | 475 |
| 7.16.2 参考标准和协议.....          | 475 |
| 7.16.3 可获得性.....             | 475 |
| 7.16.4 原理描述.....             | 476 |
| 7.17 GTP.....                | 478 |
| 7.17.1 介绍.....               | 478 |
| 7.17.2 参考标准和协议.....          | 481 |
| 7.17.3 可获得性.....             | 481 |
| 7.17.4 原理描述.....             | 481 |
| 7.17.4.1 GTP 策略.....         | 481 |
| 7.17.4.2 GTP 计费溢出攻击防范.....   | 484 |

|                                                           |            |
|-----------------------------------------------------------|------------|
| 7.17.4.3 GTP 全局参数.....                                    | 486        |
| 7.17.4.4 GTP 策略在核心网中的应用.....                              | 487        |
| 7.18 IDS 联动.....                                          | 487        |
| 7.18.1 介绍.....                                            | 488        |
| 7.18.2 可获得性.....                                          | 488        |
| 7.18.3 原理描述.....                                          | 488        |
| 7.18.3.1 IDS 联动的特点.....                                   | 488        |
| 7.18.3.2 IDS 联动的典型应用.....                                 | 489        |
| 7.19 URPF.....                                            | 489        |
| 7.19.1 介绍.....                                            | 489        |
| 7.19.2 可获得性.....                                          | 491        |
| 7.19.3 原理描述.....                                          | 491        |
| 7.20 终端安全管理（TSM 联动）.....                                  | 494        |
| 7.20.1 介绍.....                                            | 494        |
| 7.20.2 可获得性.....                                          | 494        |
| 7.20.3 原理描述.....                                          | 494        |
| 7.20.3.1 终端安全理念与 TSM 简介.....                              | 494        |
| 7.20.3.2 SACG 工作原理简介.....                                 | 496        |
| 7.20.3.3 SACG 的部署方式.....                                  | 497        |
| <b>8 UTM（USG2200/5100/5500、USG2200/5100 BSR/HSR）.....</b> | <b>502</b> |
| 8.1 反病毒（AV）.....                                          | 503        |
| 8.1.1 介绍.....                                             | 503        |
| 8.1.2 可获得性.....                                           | 503        |
| 8.1.3 原理描述.....                                           | 503        |
| 8.2 入侵防御（IPS）.....                                        | 504        |
| 8.2.1 介绍.....                                             | 504        |
| 8.2.2 可获得性.....                                           | 504        |
| 8.2.3 原理描述.....                                           | 505        |
| 8.3 邮件过滤.....                                             | 507        |
| 8.3.1 介绍.....                                             | 507        |
| 8.3.2 参考标准和协议.....                                        | 507        |
| 8.3.3 可获得性.....                                           | 508        |
| 8.3.4 应用场景.....                                           | 508        |
| 8.3.4.1 RBL 过滤.....                                       | 508        |
| 8.3.4.2 Webmail 邮件过滤.....                                 | 509        |
| 8.3.4.3 SMTP/POP3 邮件过滤.....                               | 510        |
| 8.3.5 原理描述.....                                           | 511        |
| 8.3.5.1 电子邮件介绍.....                                       | 511        |
| 8.3.5.2 RBL 过滤.....                                       | 512        |
| 8.3.5.3 邮件内容过滤.....                                       | 514        |
| 8.4 Web 过滤.....                                           | 515        |
| 8.4.1 介绍.....                                             | 515        |

|                                                   |            |
|---------------------------------------------------|------------|
| 8.4.2 参考标准和协议.....                                | 515        |
| 8.4.3 可获得性.....                                   | 515        |
| 8.4.4 应用场景.....                                   | 516        |
| 8.4.5 原理描述.....                                   | 516        |
| 8.4.5.1 URL 过滤.....                               | 516        |
| 8.4.5.2 搜索关键字过滤.....                              | 518        |
| 8.4.5.3 Web 内容过滤.....                             | 519        |
| 8.5 FTP 过滤.....                                   | 519        |
| 8.5.1 介绍.....                                     | 520        |
| 8.5.2 参考标准和协议.....                                | 520        |
| 8.5.3 可获得性.....                                   | 520        |
| 8.5.4 应用场景.....                                   | 520        |
| 8.6 应用控制（DPI） .....                               | 521        |
| 8.6.1 介绍.....                                     | 521        |
| 8.6.2 可获得性.....                                   | 521        |
| 8.6.3 特性增强.....                                   | 522        |
| 8.6.4 应用场景.....                                   | 522        |
| 8.6.5 原理描述.....                                   | 522        |
| 8.6.5.1 DPI 工作原理.....                             | 522        |
| 8.6.5.2 关联协议识别.....                               | 523        |
| 8.6.5.3 全包检测.....                                 | 524        |
| <b>9 UTM（USG2110-X/2100、USG2100BSR/HSR） .....</b> | <b>525</b> |
| 9.1 反病毒（AV） .....                                 | 526        |
| 9.1.1 介绍.....                                     | 526        |
| 9.1.2 可获得性.....                                   | 526        |
| 9.1.3 原理描述.....                                   | 526        |
| 9.2 协议识别.....                                     | 528        |
| 9.2.1 介绍.....                                     | 528        |
| 9.2.2 可获得性.....                                   | 528        |
| 9.2.3 原理描述.....                                   | 529        |
| 9.3 入侵防御（IPS） .....                               | 529        |
| 9.3.1 介绍.....                                     | 529        |
| 9.3.2 可获得性.....                                   | 530        |
| 9.3.3 原理描述.....                                   | 530        |
| 9.4 协议检测.....                                     | 531        |
| 9.4.1 介绍.....                                     | 532        |
| 9.4.2 参考协议和标准.....                                | 532        |
| 9.4.3 可获得性.....                                   | 532        |
| 9.4.4 原理描述.....                                   | 533        |
| 9.4.4.1 检测类型和响应策略.....                            | 533        |
| 9.4.4.2 HTTP 协议检测.....                            | 533        |
| 9.4.4.3 FTP 协议检测.....                             | 535        |

|                             |            |
|-----------------------------|------------|
| 9.4.4.4 SMTP 协议检测.....      | 536        |
| 9.4.4.5 IMAP/POP3 协议检测..... | 537        |
| 9.4.4.6 DNS 协议检测.....       | 538        |
| 9.5 反垃圾邮件（RBL 过滤） .....     | 539        |
| 9.5.1 介绍.....               | 539        |
| 9.5.2 可获得性.....             | 539        |
| 9.5.3 原理描述.....             | 539        |
| 9.6 URL 过滤.....             | 541        |
| 9.6.1 介绍.....               | 541        |
| 9.6.2 可获得性.....             | 541        |
| 9.6.3 原理描述.....             | 542        |
| 9.6.3.1 URL 过滤概述.....       | 542        |
| 9.6.3.2 URL 格式.....         | 542        |
| 9.6.3.3 URL 匹配方式.....       | 543        |
| 9.6.3.4 URL 过滤方式.....       | 544        |
| 9.6.3.5 URL 过滤工作原理.....     | 545        |
| 9.7 应用控制（DPI） .....         | 547        |
| 9.7.1 介绍.....               | 547        |
| 9.7.2 可获得性.....             | 547        |
| 9.7.3 特性增强.....             | 548        |
| 9.7.4 应用场景.....             | 548        |
| 9.7.5 原理描述.....             | 548        |
| 9.7.5.1 DPI 工作原理.....       | 548        |
| 9.7.5.2 关联协议识别.....         | 549        |
| 9.7.5.3 全包检测.....           | 550        |
| <b>10 MPLS 与 VPN.....</b>   | <b>551</b> |
| 10.1 VPN.....               | 552        |
| 10.1.1 介绍.....              | 552        |
| 10.1.2 原理描述.....            | 552        |
| 10.1.2.1 VPN 的原理.....       | 552        |
| 10.1.2.2 VPN 的特征.....       | 553        |
| 10.1.2.3 VPN 的优势.....       | 553        |
| 10.1.2.4 VPN 的分类.....       | 554        |
| 10.1.2.5 VPN 基本组网应用.....    | 555        |
| 10.2 IPSec.....             | 555        |
| 10.2.1 介绍.....              | 555        |
| 10.2.2 参考标准和协议.....         | 556        |
| 10.2.3 可获得性.....            | 557        |
| 10.2.4 应用场景.....            | 558        |
| 10.2.4.1 局域网通过 VPN 互通.....  | 558        |
| 10.2.4.2 移动用户远程接入 VPN.....  | 563        |
| 10.2.4.3 IPSec 冗余设计.....    | 564        |

|                                    |     |
|------------------------------------|-----|
| 10.2.4.4 IPSec 多实例应用.....          | 567 |
| 10.2.5 IPSec 协议框架.....             | 568 |
| 10.2.5.1 协议框架简介.....               | 568 |
| 10.2.5.2 安全协议.....                 | 569 |
| 10.2.5.3 封装模式.....                 | 571 |
| 10.2.5.4 加密.....                   | 574 |
| 10.2.5.5 验证.....                   | 575 |
| 10.2.5.6 密钥交换.....                 | 576 |
| 10.2.6 IPSec 工作原理.....             | 577 |
| 10.2.7 IKEv1 协商安全联盟的过程.....        | 579 |
| 10.2.7.1 IKEv1 协商阶段 1.....         | 579 |
| 10.2.7.2 IKEv1 协商阶段 2.....         | 582 |
| 10.2.8 IKEv2 协商安全联盟的过程.....        | 583 |
| 10.2.9 EAP 认证.....                 | 585 |
| 10.3 L2TP.....                     | 586 |
| 10.3.1 介绍.....                     | 586 |
| 10.3.2 参考标准和协议.....                | 587 |
| 10.3.3 可获得性.....                   | 587 |
| 10.3.4 原理描述.....                   | 587 |
| 10.3.4.1 VPDN 简介.....              | 587 |
| 10.3.4.2 L2TP 协议简介.....            | 588 |
| 10.3.4.3 L2TP 典型应用.....            | 592 |
| 10.4 GRE.....                      | 593 |
| 10.4.1 介绍.....                     | 593 |
| 10.4.2 参考标准和协议.....                | 593 |
| 10.4.3 可获得性.....                   | 593 |
| 10.4.4 原理描述.....                   | 594 |
| 10.4.4.1 GRE 的实现.....              | 594 |
| 10.4.4.2 GRE 典型应用.....             | 595 |
| 10.5 MPLS.....                     | 597 |
| 10.5.1 介绍.....                     | 597 |
| 10.5.2 参考标准和协议.....                | 597 |
| 10.5.3 可获得性.....                   | 598 |
| 10.5.4 原理描述.....                   | 598 |
| 10.5.4.1 MPLS 产生背景.....            | 598 |
| 10.5.4.2 MPLS 基本概念.....            | 599 |
| 10.5.4.3 MPLS 体系结构.....            | 602 |
| 10.5.4.4 MPLS 网络结构.....            | 603 |
| 10.5.4.5 MPLS Ping/Traceroute..... | 604 |
| 10.5.4.6 MPLS LDP 的基本概念.....       | 605 |
| 10.5.4.7 LDP 会话基本概念.....           | 607 |
| 10.5.4.8 LDP 的发现机制.....            | 608 |



|                                     |            |
|-------------------------------------|------------|
| 10.5.4.9 LDP 会话的建立.....             | 609        |
| 10.5.4.10 LDP 会话的维护.....            | 610        |
| 10.6 BGP MPLS IP VPN.....           | 610        |
| 10.6.1 介绍.....                      | 610        |
| 10.6.2 参考标准和协议.....                 | 612        |
| 10.6.3 可获得性.....                    | 612        |
| 10.6.4 原理描述.....                    | 612        |
| 10.6.4.1 BGP/MPLS IP VPN 基本概念.....  | 612        |
| 10.6.4.2 Hub and Spoke.....         | 615        |
| 10.6.4.3 跨域 VPN.....                | 617        |
| 10.6.4.4 HoVPN.....                 | 619        |
| 10.6.4.5 OSPF Sham link.....        | 621        |
| 10.6.4.6 Multi-VPN-Instance CE..... | 623        |
| 10.6.4.7 VPN 与 Internet 互联.....     | 624        |
| 10.7 SSL VPN.....                   | 625        |
| 10.7.1 介绍.....                      | 625        |
| 10.7.2 参考标准和协议.....                 | 626        |
| 10.7.3 可获得性.....                    | 626        |
| 10.7.4 原理描述.....                    | 627        |
| 10.7.4.1 SSL 协议简介.....              | 627        |
| 10.7.4.2 虚拟网关.....                  | 628        |
| 10.7.4.3 SSL VPN 业务.....            | 630        |
| 10.7.4.4 认证与授权.....                 | 632        |
| 10.7.4.5 策略.....                    | 633        |
| 10.7.4.6 DNS.....                   | 634        |
| <b>11 用户与认证.....</b>                | <b>635</b> |
| 11.1 上网用户管理.....                    | 636        |
| 11.1.1 介绍.....                      | 636        |
| 11.1.2 参考标准和协议.....                 | 636        |
| 11.1.3 可获得性.....                    | 637        |
| 11.1.4 应用场景.....                    | 637        |
| 11.1.4.1 免认证.....                   | 637        |
| 11.1.4.2 密码认证.....                  | 638        |
| 11.1.4.3 单点登录.....                  | 639        |
| 11.1.5 原理描述.....                    | 640        |
| 11.1.5.1 组/用户.....                  | 640        |
| 11.1.5.2 认证原理.....                  | 644        |
| 11.2 接入用户管理.....                    | 653        |
| 11.2.1 介绍.....                      | 653        |
| 11.2.2 参考标准和协议.....                 | 653        |
| 11.2.3 可获得性.....                    | 653        |
| 11.2.4 原理描述.....                    | 654        |

|                                  |            |
|----------------------------------|------------|
| 11.2.4.1 AAA 协议.....             | 654        |
| 11.2.4.2 RADIUS 协议.....          | 655        |
| 11.2.4.3 HWTACACS 协议.....        | 657        |
| 11.2.4.4 基于域的用户管理.....           | 658        |
| 11.2.5 应用.....                   | 660        |
| 11.3 证书.....                     | 661        |
| 11.3.1 介绍.....                   | 661        |
| 11.3.2 参考标准和协议.....              | 663        |
| 11.3.3 可获得性.....                 | 663        |
| 11.3.4 原理描述.....                 | 664        |
| 11.3.4.1 PKI 体系.....             | 664        |
| 11.3.4.2 证书申请.....               | 665        |
| 11.3.4.3 证书获取.....               | 666        |
| 11.3.4.4 证书吊销列表.....             | 667        |
| 11.3.4.5 OCSP.....               | 667        |
| 11.3.5 证书应用.....                 | 667        |
| 11.3.5.1 证书在 IPSec VPN 中的应用..... | 667        |
| 11.3.5.2 基于证书属性的 VPN 访问控制.....   | 668        |
| 11.4 802.1X.....                 | 669        |
| 11.4.1 介绍.....                   | 669        |
| 11.4.2 参考标准和协议.....              | 670        |
| 11.4.3 可获得性.....                 | 670        |
| 11.4.4 原理描述.....                 | 670        |
| 11.4.4.1 802.1X 的基本原理.....       | 670        |
| 11.4.4.2 802.1X 的认证过程.....       | 672        |
| <b>12 IP 组播.....</b>             | <b>675</b> |
| 12.1 组播概述.....                   | 676        |
| 12.1.1 介绍.....                   | 676        |
| 12.1.2 原理描述.....                 | 676        |
| 12.1.2.1 组播特点和应用.....            | 676        |
| 12.1.2.2 组播基本概念.....             | 679        |
| 12.1.2.3 组播协议.....               | 681        |
| 12.1.2.4 组播地址.....               | 682        |
| 12.1.2.5 组播模型.....               | 684        |
| 12.1.2.6 组播报文转发.....             | 685        |
| 12.2 IGMP.....                   | 685        |
| 12.2.1 介绍.....                   | 685        |
| 12.2.2 参考标准和协议.....              | 686        |
| 12.2.3 可获得性.....                 | 686        |
| 12.2.4 原理描述.....                 | 686        |
| 12.2.4.1 IGMPv1.....             | 686        |
| 12.2.4.2 IGMPv2.....             | 688        |

|                                     |     |
|-------------------------------------|-----|
| 12.2.4.3 IGMPv3.....                | 688 |
| 12.2.4.4 SSM Mapping.....           | 690 |
| 12.2.4.5 IGMP 消息.....               | 691 |
| 12.2.4.6 IGMP 典型应用.....             | 693 |
| 12.3 IGMP Snooping.....             | 697 |
| 12.3.1 介绍.....                      | 697 |
| 12.3.2 参考标准和协议.....                 | 698 |
| 12.3.3 可获得性.....                    | 698 |
| 12.3.4 应用场景.....                    | 698 |
| 12.3.5 原理描述.....                    | 699 |
| 12.3.5.1 IGMP Snooping 基本原理和概念..... | 699 |
| 12.3.5.2 IGMP Snooping 实现机制.....    | 702 |
| 12.3.5.3 IGMP Proxy.....            | 705 |
| 12.4 PIM.....                       | 709 |
| 12.4.1 介绍.....                      | 709 |
| 12.4.2 参考标准和协议.....                 | 710 |
| 12.4.3 可获得性.....                    | 710 |
| 12.4.4 原理描述.....                    | 710 |
| 12.4.4.1 基本概念.....                  | 710 |
| 12.4.4.2 PIM-DM.....                | 712 |
| 12.4.4.3 PIM-SM.....                | 714 |
| 12.4.4.4 PIM-SM 的 BSR 管理域.....      | 720 |
| 12.4.4.5 PIM-SSM.....               | 721 |
| 12.4.4.6 PIM 协议比较.....              | 723 |
| 12.4.4.7 PIM 消息.....                | 724 |
| 12.4.4.8 PIM 安全性.....               | 726 |
| 12.4.4.9 PIM 典型应用.....              | 727 |
| 12.5 MSDP.....                      | 730 |
| 12.5.1 介绍.....                      | 730 |
| 12.5.2 参考标准和协议.....                 | 731 |
| 12.5.3 可获得性.....                    | 731 |
| 12.5.4 原理描述.....                    | 731 |
| 12.5.4.1 MSDP 对等体.....              | 731 |
| 12.5.4.2 MSDP 实现 PIM-SM 域间组播.....   | 733 |
| 12.5.4.3 MSDP 实现 Anycast RP.....    | 734 |
| 12.5.4.4 Mesh Group.....            | 735 |
| 12.5.4.5 SA 消息的 RPF 检查.....         | 736 |
| 12.5.4.6 MSDP 典型应用.....             | 737 |
| 12.6 组播路由与转发.....                   | 739 |
| 12.6.1 介绍.....                      | 739 |
| 12.6.2 参考标准和协议.....                 | 740 |
| 12.6.3 可获得性.....                    | 741 |

|                               |            |
|-------------------------------|------------|
| 12.6.4 原理描述.....              | 741        |
| 12.6.4.1 控制平面.....            | 741        |
| 12.6.4.2 转发平面.....            | 742        |
| 12.6.4.3 组播转发流程.....          | 743        |
| 12.6.4.4 RPF 检查.....          | 745        |
| 12.6.4.5 组播静态路由.....          | 747        |
| 12.6.4.6 组播路由匹配策略.....        | 749        |
| 12.6.4.7 控制组播转发.....          | 750        |
| 12.6.4.8 GRE 隧道在组播转发中的应用..... | 750        |
| <b>13 QoS.....</b>            | <b>752</b> |
| 13.1 流量监管、流量整形和接口限速.....      | 753        |
| 13.1.1 介绍.....                | 753        |
| 13.1.2 参考标准和协议.....           | 753        |
| 13.1.3 可获得性.....              | 753        |
| 13.1.4 原理描述.....              | 754        |
| 13.1.4.1 令牌桶.....             | 754        |
| 13.1.4.2 流量监管.....            | 755        |
| 13.1.4.3 流量整形.....            | 755        |
| 13.1.4.4 接口限速.....            | 756        |
| 13.2 拥塞管理.....                | 757        |
| 13.2.1 介绍.....                | 757        |
| 13.2.2 可获得性.....              | 758        |
| 13.2.3 原理描述.....              | 758        |
| 13.2.3.1 FIFO 队列.....         | 758        |
| 13.2.3.2 CQ 队列.....           | 759        |
| 13.2.3.3 PQ 队列.....           | 759        |
| 13.2.3.4 WFQ 队列.....          | 760        |
| 13.2.3.5 CBWFQ 队列.....        | 761        |
| 13.2.3.6 RTP 队列.....          | 763        |
| 13.3 拥塞避免.....                | 763        |
| 13.3.1 介绍.....                | 763        |
| 13.3.2 可获得性.....              | 764        |
| 13.3.3 原理描述.....              | 764        |
| 13.4 基于类的 QoS.....            | 765        |
| 13.4.1 介绍.....                | 765        |
| 13.4.2 可获得性.....              | 765        |
| 13.4.3 原理描述.....              | 766        |
| 13.4.3.1 流分类.....             | 766        |
| 13.4.3.2 标记.....              | 766        |
| 13.4.3.3 DSCP.....            | 767        |
| 13.4.3.4 标准的 PHB.....         | 767        |
| 13.4.3.5 基于类的 CBWFQ.....      | 768        |

|                         |     |
|-------------------------|-----|
| 13.4.3.6 策略的实现.....     | 768 |
| 13.5 链路效率机制.....        | 769 |
| 13.5.1 介绍.....          | 769 |
| 13.5.2 参考标准和协议.....     | 770 |
| 13.5.3 可获得性.....        | 770 |
| 13.5.4 原理描述.....        | 770 |
| 13.6 VPN QoS.....       | 771 |
| 13.6.1 介绍.....          | 771 |
| 13.6.2 参考标准和协议.....     | 772 |
| 13.6.3 可获得性.....        | 772 |
| 13.6.4 应用场景.....        | 772 |
| 13.6.5 原理描述.....        | 774 |
| 13.7 HQoS.....          | 775 |
| 13.7.1 介绍.....          | 775 |
| 13.7.2 可获得性.....        | 776 |
| 13.7.3 应用场景.....        | 776 |
| 13.7.4 原理描述.....        | 776 |
| 13.8 二层 QoS.....        | 777 |
| 13.8.1 简介.....          | 778 |
| 13.8.2 可获得性.....        | 778 |
| 13.8.3 原理描述.....        | 779 |
| 13.8.3.1 优先级映射.....     | 779 |
| 13.8.3.2 SP 队列.....     | 780 |
| 13.8.3.3 WRR 队列.....    | 780 |
| 13.8.3.4 SP+WRR 队列..... | 780 |

# 1 系统管理

## 关于本章

### 1.1 FTP

FTP 协议主要向用户提供与远程主机之间的文件传输，系统提供的 FTP 服务包括 FTP Server 服务、FTP Client 服务。

### 1.2 TFTP

TFTP 协议是一种简单文件传输协议，适用于客户机和服务器之间交互不复杂的环境。TFTP 协议可以传输文件包括:二进制模式、ASCII 码模式。

### 1.3 Telnet

Telnet 协议在 TCP/IP 协议族中属于应用层协议，通过网络提供远程登录和虚拟终端功能。

### 1.4 SSH

SSH（Secure Shell）特性可以提供安全的信息保障和强大的认证功能，以保护 USG 不受诸如 IP 地址欺诈、明文密码截取等攻击。

### 1.5 信息中心

信息中心接管大多数的信息输出，并且能够进行细致的分类，从而能够有效地进行信息筛选。它通过与 Debug 程序的结合，为网络管理员和开发人员监控网络运行情况和诊断网络故障提供了强有力的支持。

### 1.6 SNMP

SNMP（Simple Network Management Protocol）为网管和被管理设备的通信提供一套标准协议，保证网管能够正常的管理设备和接收设备的告警。

### 1.7 LLDP

LLDP 功能用于获取邻居信息和拓扑发现。

### 1.8 CWMP

CWMP 是一种广域网管理协议，可实现 ACS 对 CPE 的集中管理，主要功能为配置管理、版本管理、远程监控和诊断。

### 1.9 NetStream

NetStream 对网络流量进行统计，并定期向 NSC 发送统计数据。这些数据可用于计费、网络管理和规划等。

#### 1.10 NTP

NTP 基于 UDP 传输，其目的是对网络内所有时钟设备进行时钟同步。运行 NTP 的系统可以主动同步其它系统，也可以接受其它系统的同步。

#### 1.11 NQA

NQA 功能可以测量网络上运行的各种协议的性能，使运营商能够实时采集到各种网络运行指标。

#### 1.12 Ping 和 Tracert

当正常网络通信受到影响时，可以首先使用 ping 与 tracert 命令检测网络是否正常工作。

#### 1.13 探测 PMTU

探测 PMTU 的目的就是要找到从源端到目的端路径上使报文不分片的最大 MTU 值。

## 1.1 FTP

FTP 协议主要向用户提供与远程主机之间的文件传输，系统提供的 FTP 服务包括 FTP Server 服务、FTP Client 服务。

### 1.1.1 介绍

#### 定义

FTP（File Transfer Protocol）在 TCP/IP 协议族中属于应用层协议，是文件传输的国际标准。

#### 目的

FTP 主要功能是向用户提供本地和远程主机之间的文件传输，在进行版本升级、日志下载、文件传输和配置保存等业务操作时，FTP 功能被广泛地使用。

FTP 除了完成文件传输基本功能外，同时还提供了交互存取、格式规范和验证控制等超出传输功能的配置功能。FTP 实现主机间文件的传输，并提供常用的文件操作命令，供用户进行文件系统的简单管理。客户可以利用 USG 外部的 FTP 客户端程序与 USG 进行文件的上传、下载和目录访问等操作；还可以使用 USG 内部的 FTP 客户端程序与其他 USG 或其他设备的 FTP 服务器端的程序进行文件传输。

### 1.1.2 可获得性

#### License 支持

不需要 License 支持。

#### 版本支持

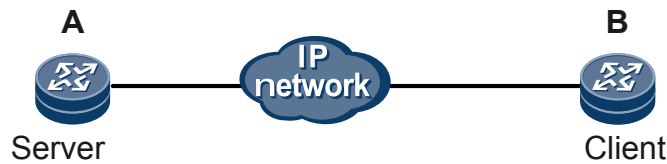
| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 1.1.3 原理描述

FTP 采用 C/S（Client/Server）结构，常见组网如图 1-1 所示。



图 1-1 FTP 组网图



说明

需要保证 FTP Client 和 FTP Server 之间路由可达。

- **FTP Server:** 提供 FTP 服务的设备。通常是网络中的文件服务器，为 FTP Client 提供文件访问服务。
- **FTP Client:** 访问 FTP 服务器的设备，是 FTP 连接的发起者。当 FTP 通过 TCP 建立连接后，FTP Client 和 FTP Server 之间即可进行文件传输。

FTP 属于多通道协议，会建立数据通道和控制通道 2 个 TCP 连接。当设备既开启 NAT 功能，又需要转发多通道协议报文（例如 FTP 等）时，就必须开启相应的 NAT ALG 功能。

## 1.2 TFTP

TFTP 协议是一种简单文件传输协议，适用于客户机和服务器之间交互不复杂的环境。TFTP 协议可以传输文件包括：二进制模式、ASCII 码模式。

### 1.2.1 介绍

#### 定义

TFTP（Trivial File Transfer Protocol，简单文件传送协议）是 TCP/IP 协议族中用来在客户机与服务器之间进行简单文件传输的协议，提供不复杂、开销不大的文件传输服务。

#### 目的

TFTP 是一个传输简单文件的协议，主要基于 UDP 协议而实现。TFTP 是基于小文件的传输而设计的，所以它缺少了 FTP 协议的一些功能。例如，TFTP 协议只能从服务器上获得或写入文件，不能列出目录，不进行认证。

### 1.2.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 1350: The TFTP Protocol (Revision 2)

### 1.2.3 可获得性

## License 支持

不需要 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 1.2.4 原理描述

### 基本概念

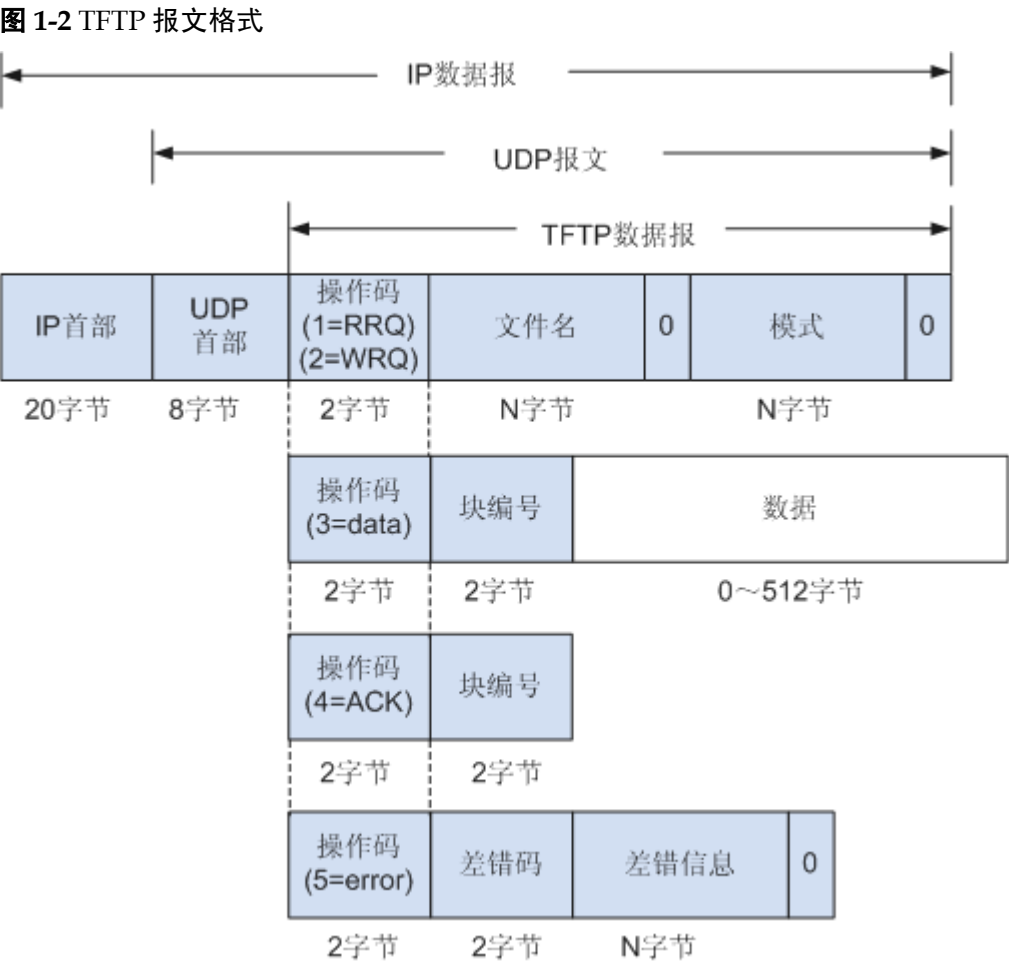
- 操作码  
TFTP 报文的头 2 个字节表示操作码，其取值和含义如下：
  - 数值为 1：表示读请求（RRQ）。
  - 数值为 2：表示写请求（WRQ）。
  - 数值为 3：表示数据（data）。
  - 数值为 4：表示肯定应答（ACK）。
  - 数值为 5：表示差错（error）。
- 文件模式  
TFTP 传输文件有两种模式：
  - 二进制模式：用于传输程序文件。
  - ASCII 码模式：用于传输文本文件。

### 基本原理

- TFTP 不提供用户名和口令  
因为 TFTP 初始设计是用于系统引导进程的，所以它不提供用户名和口令。
- TFTP 协议传输  
TFTP 协议传输由客户端发起
  - 当需要下载文件时，由客户端向 TFTP 服务器发送读请求包，然后从服务器接收数据包，并向服务器发送确认。
  - 当需要上传文件时，由客户端向 TFTP 服务器发送写请求包，然后向服务器发送数据包，并接收服务器的确认。

### 报文格式

TFTP 共有 5 种报文格式，如图 1-2 所示。



以下内容分别描述 TFTP 的 5 种报文。

- 读/写请求报文  
即操作码为 1 和 2 的报文，这 2 种报文使用相同的格式。  
对于读请求（RRQ）报文和写请求（WRQ）报文来说，文件名字段标识客户端要读/写的位于服务器上的文件；文件名字段以字节“0”作为结束标识。  
模式字段是一个 ASCII 码串“netascii”或者“octet”（可大小写任意组合），以字节“0”结束。其取值和含义如表 1-1 所示。

表 1-1 TFTP 报文中文件模式字段的取值及含义

| 取值       | 含义                                               |
|----------|--------------------------------------------------|
| netascii | 表示数据是以成行的 ASCII 码字符组成，以 2 个字节（回车符号后跟换行字符）作为行结束符。 |
| octet    | 将数据看作 8 比特一组的字节流，不作任何解释。                         |

- 数据分组报文  
即操作码为 3 的报文，每个数据分组包含一个块编号字段，用于确认分组。

- ACK 报文  
即操作码为 4 的报文，ACK 报文是通信过程中的一种确认、证实的报文。
- 差错报文  
即操作码为 5 的报文，当服务器不能处理读/写请求时，则发送差错报文给客户端。  
文件传输过程中的读/写差错也会导致差错报文。差错码字段给出一个数字标识的差错码；差错信息字段是由 2 字节 ASCII 码表示的差错信息，可能包含额外的操作系统说明信息。

## 1.3 Telnet

Telnet 协议在 TCP/IP 协议族中属于应用层协议，通过网络提供远程登录和虚拟终端功能。

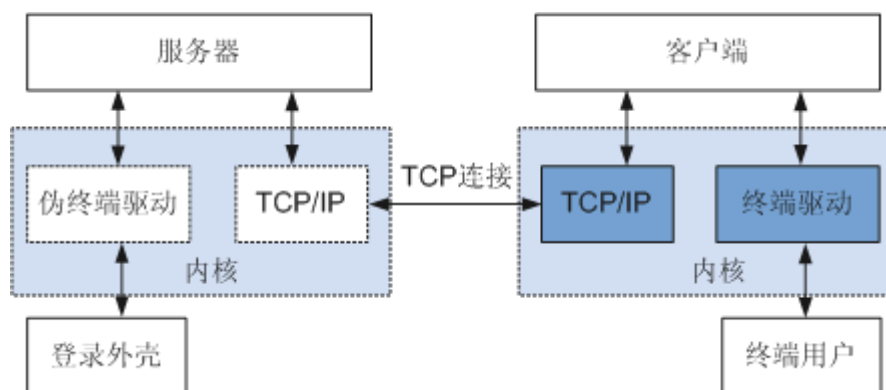
### 1.3.1 介绍

#### 定义

Telnet（Telecommunication Network Protocol）起源于 1969 年的 ARPANET，最早应用于 Internet。

Telnet 采用客户端/服务器模式。Telnet 客户端和服务器的连接如图 1-3 所示。

图 1-3 Telnet 客户端和服务器的连接



通过该示意图，可知：

- Telnet 使用的传输协议为 TCP 协议。
- Telnet 连接的任何回显信息，最终都会输出到终端上。
- 服务器进程直接与“伪终端设备”交互。
- 客户端和服务器通过 TCP 连接传输命令和数据。
- 客户端登录到服务器上。

目的

Telnet 提供了一种通过终端远程登录到服务器的方式，呈现一个交互式操作界面。用户可以先登录到一台主机，然后再通过 Telnet 的方式远程登录到网络上的其他主机，并不需要为每一台主机都连接一个硬件终端，然后对设备进行配置和管理。

1.3.2 可获得性

License 支持

不需要 License 支持。

版本支持

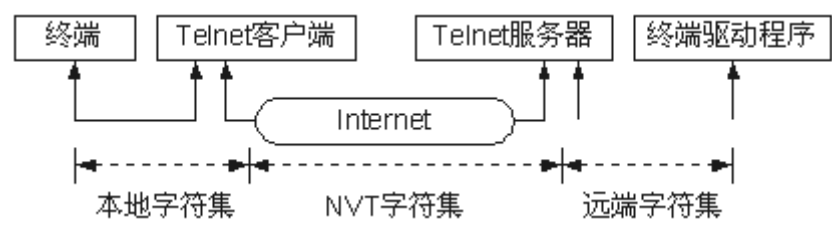
| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

1.3.3 原理描述

基本概念

- NVT  
即网络虚拟终端（Network Virtual Terminal），是一种双向的虚拟设备，连接的双方都必须把它们各自的物理终端和 NVT 进行转换。由于使用了统一的 NVT，Telnet 协议可以工作在任何主机（例如，任何操作系统）或任何终端之间。  
NVT 是虚拟设备，对于连接的双方（即客户机和服务器），都必须把它们的物理终端和 NVT 进行相互转换。也就是说，无论客户进程终端是什么类型，操作系统都必须把它转换为 NVT 格式。同时，不管服务器进程的终端是什么类型，操作系统都能把 NVT 格式转换为终端所能支持的格式。  
物理终端与 NVT 的转换模型如图 1-4 所示。

图 1-4 物理终端与 NVT 的转换模型



- NVT ASCII

即 7 比特的 ASCII 字符集。发送时，在每个 7 比特的字符的最高比特位加 0，以 8 比特的格式发送。网间网协议族使用的字符集是 NVT ASCII，例如 FTP、SMTP 等。

- IAC  
Telnet 通信的两个方向都采用带内信令的方式。字节 0xFF 称作 IAC（Interpret As Command），即作为命令来解释。该字节后面的一个字节代表命令。  
以下列出 USG 用到的命令及其含义：
  - SE：子选项结束
  - SB：子选项开始
  - WILL：选项协商
  - WONT：选项协商
  - DO：选项协商
  - DONT：选项协商
  - IAC：其后面的一个字节作为命令来解释

工作原理

Telnet 协议可以工作在任何主机或者任何终端之间。因为，无论客户终端是什么类型，操作系统都会把它转换成 NVT 格式；同时，操作系统也会将 NVT 格式转换成服务器终端所支持的类型。那么，可以屏蔽具体的客户端和终端类型，简单地认为 Telnet 双方都连接在 NVT 上。

说明

Telnet 采用对称性模型，因此理论上一个 Telnet 连接的每一端都必须有一个 NVT。

Telnet 连接的两端，通过“WILL、WONT、DO、DONT”请求来进行选项协商，从而确定 Telnet 服务的具体内容。这些选项包括回显、改变命令字符集或行方式等。

本节从以下几个方面介绍 Telnet 的工作原理：

- Telnet 中的请求  
Telnet 连接的任何一方都可以主动发起请求。请求的含义和用法如表 1-2 所示。

表 1-2 Telnet 中请求的使用

| 发送方发出请求 | 含义           | 接受方应答 |                        |       |                        |
|---------|--------------|-------|------------------------|-------|------------------------|
|         |              | WILL  | WONT                   | DO    | DONT                   |
| WILL    | 发送方想激活选项     | —     | —                      | 接收方同意 | 接收方不同意                 |
| WONT    | 发送方想禁止选项     | —     | —                      | —     | 接收方必须同意 <sup>(1)</sup> |
| DO      | 发送方想让接收方激活选项 | 接收方同意 | 接收方不同意                 | —     | —                      |
| DONT    | 发送方想让接收方禁止选项 | —     | 接收方必须同意 <sup>(1)</sup> | —     | —                      |

 说明

发起方发送一些“选项有效”的请求，接收方可以接受或者拒绝这些请求：

- 如果接受请求，则选项立即生效；
- 如果拒绝请求，则选项不生效，而发送方仍然能保留 NVT 的特性。

<sup>(1)</sup>发起方发送“选项失效”请求（WONT 和 DONT）时，接收方必须同意；

- 选项协商

选项协商由 3 个字节组成，如下：

< IAC, WILL/DO/WONT/DONT 之一,选项代码 >

下面举例说明 Telnet 中如何进行选项协商。

例如：服务器想跟客户端请求激活“远程流量控制”（选项标识是 33），客户端表示同意激活该选项。两者交互的命令如下：

- 服务器：< IAC,WILL,33 >
- 客户端：< IAC,DO,33 >

- 子选项协商

在主机之间传递选项时，除了一个选项代码外，可能还需要其他更多的信息。例如，要求对方指定终端类型，则对方必须回应一个用 ASCII 字符串表示的终端类型。

子选项协商的格式如下：

< IAC,SB,选项代码,子选项内容信息,IAC,SE>

一次完整的子选项协商过程如下：

- 发送方发送一个带有选项代码的 DO/WILL 命令来请求激活该选项。
- 接收方发送一个带有选项代码的 WILL/DO 命令来同意激活该选项。

至此，双方都同意使用该选项。

其中一方通过 SB 命令后跟子选项代码，并且以 SE 结尾来开始子选项协商信息。

- 对方同样以 SB 命令后跟子选项代码和相关协商信息，并且以 SE 结尾来回应子选项协商信息。
- 另一方回应 DO/WILL 命令来同意该子选项。

如果没有其他子选项需要协商，则本次协商结束。

 说明

上述过程假设接收方同意发送方的请求。实际应用中，接收方根据需要，在任何时候都可以拒绝发送方的请求。

下面举一个终端类型协商的示例。

例如：客户端要激活本地的“终端类型”（选项标识是 24）选项；服务器同意；服务器会发出询问客户端的终端类型的请求；于是客户端向服务器发送其终端类型为“DELL PC 机”。两者交互的命令如下：

- 客户端：< IAC, WILL, 24 >
- 服务器：< IAC, DO, 24 >
- 服务器：< IAC, SB, 24, 1, IAC, SE >
- 客户端：< IAC, SB, 24, 0, 'D', 'E', 'L', 'L', 'P', 'C', IAC, SE >



说明

- 只有 DO 类型的发送端可以发送请求。
  - 只有 WILL 类型的发送端可以发送实际的类型信息。
- 终端类型信息不能以自动方式发送，而只能以<请求-响应>的方式。
- 终端类型信息是 NVT ASCII String 字符串类型。该类型编码不区分大小写。
- 操作方式
- Telnet 协议规定有 4 种操作方式。
- 半双工
  - 一次一个字符
  - 一次一行
  - 行方式

## 1.4 SSH

SSH（Secure Shell）特性可以提供安全的信息保障和强大的认证功能，以保护 USG 不受诸如 IP 地址欺诈、明文密码截取等攻击。

### 1.4.1 介绍

#### 定义

SSH 是 Secure Shell（安全外壳）的简称，标准协议端口号是 22。

SSH 通过 TCP 进行数据交互，它在 TCP 之上构建了一个安全的通道。SSH 服务除了支持标准端口 22 外，还支持其他服务端口，以防止受到非法攻击。

SSH 支持 Password 认证和 RSA 认证，对数据进行 DES、3DES、AES 等加密，有效防止了对密码的窃取，保护了数据的完整性和可靠性，以及数据的安全传输。特别是对于 RSA 认证的支持，对称加密和非对称加密的混合应用，密钥的安全交换，最终实现了安全的会话过程。

SSH 协议有两个版本：SSH1（SSH1.5）协议和 SSH2（SSH2.0）协议，两者是不同的协议，不兼容。SSH2.0 在安全、功能和性能上均优于 SSH1.5。

USG 支持 STelnet 以及 SFTP 的客户端和服务端，均支持 SSH1（SSH1.5）协议和 SSH2（SSH2.0）协议。

STelnet 是 Secure Telnet 的简称，使得用户可以从远端安全登录到设备。提供交互式配置界面，所有交互数据均经过加密，实现安全的会话。

SFTP 是 Secure FTP 的简称，使得用户可以从远端安全地登录到设备上文件管理，这样远程系统升级等维护所需要文件可以安全的传送，增加了数据传输的安全性。同时，由于提供了客户端功能，可以在本设备上通过 SFTP 安全的登录到远程设备，进行文件的安全传输。

#### 目的

Telnet 缺少安全的认证方式，而且传输过程采用 TCP 进行明文传输，存在很大的安全隐患。单纯提供 Telnet 服务容易导致 DoS（Deny of Service）、主机 IP 地址欺骗、路由欺骗等恶意攻击。



随着人们对网络安全的重视，传统的 Telnet 和 FTP 通过明文传送密码和数据的方式，已经逐渐不被人接受。SSH（Secure Shell）是一个网络安全协议，通过对网络数据的加密，解决了这个问题。它在一个不安全的网络环境中，提供了安全的远程登录和其他安全网络服务。

由于 SSH 数据加密传输，认证机制更加安全，SSH 已经越来越被广泛使用，成为了当前最重要的网络协议之一。

## 1.4.2 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 1.4.3 原理描述

### SSH 的基本概念

- SFTP  
SFTP（SSH File Transfer Protocol）。在一个传统不安全的网络环境中，服务器通过对客户端的认证及双向的数据加密，为网络文件传输提供了安全的服务。
- STelnet  
一种安全的 Telnet 服务。在一个传统不安全的网络环境中，服务器通过对客户端的认证及双向的数据加密，为网络终端访问提供了安全的服务。
- RSA 认证  
RSA（Rivest-Shamir-Adleman Algorithm）RSA 身份认证是基于客户端私钥的一种认证方式。它是一种公开密钥加密体系，是一种非对称加密算法，其原理是基于大整数因子分解这一著名的数学难题，主要用来传递对称加密算法所使用的密钥，通过这种方法可以有效地提高加密的效率并能简化对密钥的管理。  
服务器必须检查用户是否是合法的 SSH 用户，检查公钥对于该用户是否合法，用户数字签名是否合法。若三者同时满足，则身份认证通过；若其中任何一项不能验证通过均告验证失败，拒绝该用户的登录请求。
- Password 认证  
Password 身份认证是基于“用户名+口令”的一种认证方式。  
在服务器端由 AAA 为每一个合法用户分配一个用于登录时进行身份验证的口令，即在服务器端存在“用户名+口令”的一一对应的关系。当某个用户请求登录时，

服务器需要对用户名以及其口令分别进行鉴别，其中任何一项不能验证通过均告验证失败，拒绝该用户的登录请求。

- RSA-Password 认证

SSH 服务器可以要求客户端在进行身份认证的过程中同时进行 **Publickey** 身份认证和 **Password** 身份认证，只有在两者同时满足的情况下，才认为客户端身份认证通过。

- ALL 认证

服务器可以要求客户端在进行身份认证的过程中进行公钥认证或密码认证，只要满足其中一个认证，就认为客户端身份认证通过。

## SSH 协议的基本原理

SSH 采用了传统客户端/服务器应用模型，其安全特性通过以下方式保证。

**数据加密：**通过客户端/服务器协商交换生成的 **Encryption Key**，实现对数据报文的对称加密，确保数据在传输过程中的机密性。

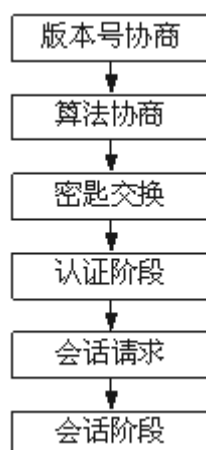
**数据完整性：**通过客户端/服务器协商交换生成的 **Integrity Key** 唯一标识的一条会话链路，所有会话交互报文被 **Integrity Key** 标识。一旦数据被第三方修改，接收方就能够检查出来，并丢弃报文，确保数据在传输过程中的完整性。

**权限认证：**通过提供多种认证方式，确保只有认证通过的合法用户才能和服务器进行会话，提高系统安全，同时保障合法用户的权益。

## SSH 处理流程

SSH 连接在整个通讯过程中，经过如图 1-5 所示的六个阶段，协商实现 SSH 认证安全连接，下面介绍一下 SSH 协商阶段的整个过程。

图 1-5 SSH 连接协商过程



1. 版本协商阶段

SSH 客户端向服务器发起 TCP 连接请求。TCP 连接建立后，SSH 服务器和客户端协商版本号。协商出匹配的版本协议，对不同版本协议走不同的状态机流程。如果版本匹配，进入密钥协商阶段；否则 SSH 服务器断开 TCP 连接。

2. 密钥算法协商

当进入算法协商过程后，客户端向服务器发送算法协商消息包，客户端同时将随机 Cookie、密钥交换的算法、主机密钥算法、支持的 MAC（Message Authentication Code）认证方法以及支持的语言等作为消息的参数发送给服务器。

服务器在等待接收到客户端的算法协商消息后，将客户端的算法列表集合与本地算法列表集合相比较。若密钥交换算法、公钥加密算法和 MAC 算法当中有一项没有找到，则断开与客户端的连接，算法协商失败。

### 3. 密钥交换阶段

当算法协商通过后，进入密钥交换阶段。双方开始计算会话 ID。客户端随机生成一个 32 字节的会话密钥（session key），用该密钥的前 16 字节异或（XOR）会话 ID 的 16 字节，后 16 字节不变，所得结果按 MSB 排列成一个 MP 型整数。用主机公钥和服务器公钥中模数较小的公钥进行加密，把结果按 MSB first 顺序排列成一个 MP 型整数，用模数较大的公钥进行加密，并把所得结果和客户端选择的加密算法、服务器传来的 8 字节 Cookie、自身的协议标志一起发给服务器端。

会话过程中，大数据量传输必须使用处理速度较快的对称加密算法。对称加解密需要共享密钥，密钥交换流程的目的是在不安全的通道中安全传送会话密钥。

服务器处于等待状态，当收到客户端发送的密钥生成消息包后，服务器回送一个密钥生成消息给客户端，表示密钥交换完成，采用新的密钥进行通讯。如果失败，则返回密钥交换失败消息，并断开连接。

### 4. 认证阶段

在计算出会话密钥后，SSH 服务器对客户端进行用户身份验证。SSH 客户端向服务器发送用户身份信息。如果 SSH 服务器上配置该用户无需验证，则直接进入请求会话阶段；如果在 SSH 服务器上配置对该用户进行验证，客户端将采用配置的验证方法向服务器提出验证请求，直到验证通过或连接超时断开。

SSH 服务器提供四种验证方法：RSA、Password、RSA-Password 和 ALL。

- 在 RSA 认证方式下，用户客户端程序生成一个 RSA 密钥对，并将公钥部分发送给服务器端。用户发起认证请求时，客户端程序随机生成一段有私钥加密的密文并发送给服务器端，服务器端利用公钥对其进行解密，解密成功就认为用户是可信的，随机对用户授予相应的访问权限。否则，中断连接。
- Password 认证依靠 AAA 实现，与 Telnet 和 FTP 类似，支持本地数据库和远程 RADIUS 服务器验证，SSH 服务器对来自客户端的用户名/口令和预先配置的用户名/口令进行比较，如果完全匹配则验证通过。

### 5. 会话请求阶段

认证完成后，客户端向服务器提交会话请求。会话请求包含 shell 和命令。服务器则进行等待，处理客户端的请求。在这个阶段，无论什么请求，只要成功处理了，服务器都向客户端回应认证成功消息；否则回应认证失败消息，这表示认证失败。

认证失败的原因如下：

- 服务器处理请求失败
- 服务器不能识别请求

### 6. 交互会话阶段

会话申请成功后，连接进入交互会话模式。在这个模式下，数据在两个方向上双向传送。

- a. 客户端将要执行的命令加密后打包传给服务器。
- b. 服务器接收到报文，解密后执行该命令，并将执行的结果加密、打包、发还给客户端。
- c. 客户端将接收到的结果解密后显示到终端上。

## 1.5 信息中心

信息中心接管大多数的信息输出，并且能够进行细致的分类，从而能够有效地进行信息筛选。它通过与 Debug 程序的结合，为网络管理员和开发人员监控网络运行情况和诊断网络故障提供了强有力的支持。

### 1.5.1 简介

#### 定义

信息中心是 USG 中不可或缺的一部分，它是系统软件模块的信息枢纽。信息中心管理大多数的信息输出，通过对系统输出信息进行细致的分类，可以有效地对信息进行筛选。通过与调试程序（**debugging** 命令）和 SNMP 模块的结合，信息中心为网络管理员监控 USG 的运行情况和诊断网络故障提供了强有力的支持。

信息中心的工作原理如下：

概括来说，信息中心的主要工作就是将 3 种信息，按照 8 个严重等级，分配到 10 个信息通道中，再输出到多个方向，具体如下：

1. 首先信息中心接收各模块输出的各种级别的日志信息（log）、告警信息（trap）和调试信息（debug）。



各模块将日志信息、告警信息、调试信息分别存放在信息中心对应的日志、告警和调试队列中。

2. 根据用户的设置，将不同类型、不同重要程度的信息分别输出到不同的信息通道。
3. 根据信息通道和输出方向的关联关系，将信息输出到各个方向。

信息中心主要包含如下特性：

表 1-3 信息中心特性列表

| 特性   | 说明                                                            |
|------|---------------------------------------------------------------|
| 信息分类 | 信息中心规定信息分为日志信息、告警信息、调试信息。                                     |
| 信息分级 | 信息中心规定信息根据严重程度分为 8 个等级。信息越严重，其严重等级越小。                         |
| 信息输出 | 信息中心可以向日志文件、控制台、VTY/TTY 终端、日志主机、SNMP agent、日志缓冲区、告警缓冲区分别输出信息。 |
| 信息屏蔽 | 通过命令可以配置屏蔽输出信息的等级、模块。                                         |

#### 目的

信息中心实现信息按照统一的格式向多个方向输出，增加了日志的可读性、可维护性、灵活性。

1. 控制信息输出的方向。目前支持的输出方向有日志文件、控制台、VTY/TTY 终端、日志主机、SNMP agent、日志缓冲区、告警缓冲区。
2. 过滤信息内容。目前支持的过滤条件有信息来源、信息级别、信息类别、信息输出方向。
3. 提供系统级的信息输出平台。
4. 提供系统级的调试开关。

## 1.5.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 3164: The BSD syslog

## 1.5.3 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 1.5.4 原理描述

为了让信息更加清晰，满足各个输出方向对不同信息的要求，信息中心将信息规定为三类：日志信息、告警信息、调试信息。

日志信息主要记录用户的操作和一些诊断信息。用户信息供用户查看，诊断信息供研发人员定位使用。

告警信息主要记录故障。信息中心接收告警并发送给网管协议模块 SNMP agent，再由 SNMP agent 发送给网管。

调试信息主要用于跟踪 USG 内部运行的轨迹。

### 日志信息

- 日志概述  
日志范围比较广，按照 ITU-T 定义，凡是管理对象事件和异常活动都可以以日志形式记录下来。因此一般认为日志模块具有跟踪用户活动、管理系统安全的功能，同时也能为系统进行诊断和维护提供依据，是操作维护、定位问题的重要手段。
- 日志信息在设备上的实现

缺省情况下信息中心是开启的，可以向控制台、日志缓存区、SNMP agent 和日志文件等方向输出日志信息。

在配置日志主机后，可以把日志信息发往日志主机。设备目前最多可以配置 4 个日志主机，这样日志信息就可以同时发往不同的日志主机，实现对日志信息的备份。

缺省情况下，可以向控制台和日志缓冲区中发送日志信息，日志缓冲区和告警缓冲区，可以存储 1024 条日志信息，可以在 1 ~ 1024 的范围内配置日志缓冲区的大小。当进入日志缓冲区的日志信息数目已经达到最大的日志缓冲区的尺寸时，设备就会对进入日志缓冲区中的时间最早的日志进行覆盖，直到满足新日志的存放为止。

● 诊断日志信息

在现有的系统日志中，有些日志信息是进行问题定位使用的，对于用户没有实际的指导意义，可以不通知用户。因此对现有的系统日志信息拆分为用户日志信息和诊断日志信息。

信息中心沿用原有的用户日志管理系统，增加了对诊断日志信息的独立处理。使得用户只可以看到用户日志，而对诊断日志进行的配置对用户不可见，用户可以看到生成的诊断日志文件，但由于诊断日志文件经过加密处理，用户无法查看到诊断日志文件中的诊断日志信息。

缺省情况下，诊断日志向诊断日志文件方向输出。

● 日志信息的输出格式

Syslog 是信息中心（info-center）的一个子功能。Syslog 使用 UDP 进行传输，使用端口号 514 将日志信息输出到日志主机中。

日志格式如图 1-6 所示：

图 1-6 日志输出格式

<Int\_16>TIMESTAMP HOSTNAME %%ddAAA/B/CCC(I):-Slot=k-XXX;YYYY

各字段的详细说明见表 1-4。

表 1-4 日志记录格式说明

| 字段        | 字段含义        | 说明                                                                                                                                                                                                                                                                                         |
|-----------|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <Int_16>  | 前导符         | 在向日志主机发送的时候添加前导符，在设备本机保存的日志不保存前导符。                                                                                                                                                                                                                                                         |
| TIMESTAMP | 时间戳，信息输出的时间 | <p>时间戳有 5 种格式可供选择。</p> <ul style="list-style-type: none"><li>● boot 型：相对时间类型。调试信息缺省采用 boot 型时间戳。</li><li>● date 型：系统时间类型。告警信息和日志信息缺省采用 date 型时间戳。</li><li>● short-date 型：系统时间类型。不含有年份信息。</li><li>● format-date 型：另一种系统时间形式。</li><li>● none 型：信息中不包含时间戳。</li></ul> <p>时间戳与主机名之间用一个空格隔开。</p> |

| 字段          | 字段含义   | 说明                                                                                                                             |
|-------------|--------|--------------------------------------------------------------------------------------------------------------------------------|
| HOSTNAME    | 本地的系统名 | 缺省是“USG”。<br>主机名与模块名之间用一个空格隔开。                                                                                                 |
| %%          | 厂商标识   | 用来标识该日志是由华为公司的产品输出的。                                                                                                           |
| dd          | 版本号    | 用来标识该日志格式的的版本。                                                                                                                 |
| AAA         | 模块名    | 向信息中心输出信息的模块名称。                                                                                                                |
| B           | 日志的级别  | 表示日志信息的严重级别。                                                                                                                   |
| CCC         | 简要描述   | 用以进一步说明信息的类型。                                                                                                                  |
| (l)         | 信息的类别  | <ul style="list-style-type: none"> <li>● l: 日志信息</li> <li>● T: 告警信息</li> <li>● d: debugging 信息</li> <li>● D: 诊断日志信息</li> </ul> |
| -Slot=k-XXX | 定位信息   | Slot: 表示发送定位信息的槽位号。此部分信息前后各有一个空格。根据日志产生模块的不同, 日志信息中有可能不包括此部分信息。                                                                |
| YYYY        | 描述符    | 各个模块向信息中心输出的信息的具体内容。由各个模块在每次输出时填充, 详细描述该日志的具体内容。                                                                               |

## 告警信息

### ● 告警概述

告警是系统检测到故障而产生的通知, 告警中携带对应的故障信息。这类信息不同于日志类信息的最大特点是需要及时通知、提醒管理用户, 对时间敏感, 因此管理中心对此类信息处理的方式也不同于其它类信息 (该类信息为设备发往网管站的 Trap 信息)。

Trap 信息是 USG 发送到网管设备的信息。在 USG 上使能了 SNMP agent, 并使能了相应模块的 Trap 功能, 配置了 trap 发往的网管主机后, 当某一特定事件发生时 (比如接口 “down”), USG 会生成 trap 信息并发往指定的目的地址。如果 USG 与网管主机之间路由可达, 网管软件就能够接收到由 USG 发出的 trap 信息。

另外在 USG 中有一个告警缓冲区, 用于储存 USG 产生的 Trap 信息, 如果在信息中心中对该缓冲区的信息源进行了配置, 该缓冲区就能够存储 USG 产生的 Trap 信息 (即使没有设置网管的目的主机)。

### ● 告警相关概念

- 事件: 是指被管对象发生的任何情况的通称。例如对象的增加、删除、修改、状态改变等。
- 故障: 对系统正常运行状态的偏离, 可能导致运作能力或冗余能力的丢失。
- 告警: 系统检测到故障而产生的通知。

### ● 告警信息的输出格式

图 1-7 告警输出格式

TimeStamp HostName ModuleName/Severity/Brief:Description

各字段的详细说明见表 1-5。

表 1-5 告警记录格式说明

| 字段          | 字段含义        | 说明                                                                                                                                                                                                                                                                                                                   |
|-------------|-------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| TimeStamp   | 时间戳，信息输出的时间 | <p>时间戳有 5 种格式可供选择。</p> <ul style="list-style-type: none"> <li>● boot 型：相对时间类型。调试信息缺省采用 boot 型时间戳。</li> <li>● date 型：系统时间类型。告警信息和日志信息缺省采用 date 型时间戳。</li> <li>● short-date 型：与 date 型的唯一区别是，short-date 型时间戳不含年份。</li> <li>● format-date 型：另一种系统时间形式。</li> <li>● none 型：信息中不包含时间戳。</li> </ul> <p>时间戳与主机名之间用一个空格隔开。</p> |
| HostName    | 本地的系统名      | <p>缺省是“USG”。</p> <p>主机名与模块名之间用一个空格隔开。</p>                                                                                                                                                                                                                                                                            |
| ModuleName  | 模块名         | 用来表示产生告警的模块名。                                                                                                                                                                                                                                                                                                        |
| Severity    | 严重级别        | <p>告警的严重级别。</p> <ul style="list-style-type: none"> <li>● Critical（紧急）</li> <li>● Major（重要）</li> <li>● Minor（次要）</li> <li>● Warning（提示）</li> <li>● indeterminate（不确定）</li> </ul>                                                                                                                                      |
| Brief       | 简要描述        | 告警信息的简要描述。                                                                                                                                                                                                                                                                                                           |
| Description | 描述信息        | 告警信息的描述信息。                                                                                                                                                                                                                                                                                                           |

## 调试信息

调试信息是系统对 USG 内部运行的跟踪信息的输出。只有在用户视图下打开相应模块的调试开关，USG 才能够产生调试信息。调试信息显示被调试模块接受或者发送数据报的信息内容。打开调试开关只能产生调试信息，如果需要显示调试信息还需要另做配置。与 log 信息和 trap 信息不同的是，debugging 信息没有 debugging 缓冲区的概念。Debugging 信息可以输出到控制台，也可以通过配置发送到日志主机。



用户可以通过 console 口的方式或者 Telnet 的方式进行配置，通过 console 口的方式称为控制台，而通过 Telnet 方式登录到 USG 的方式叫做监视终端。当用户需要在控制台或者监视终端对 USG 进行调试时，可以控制调试信息输出的内容。

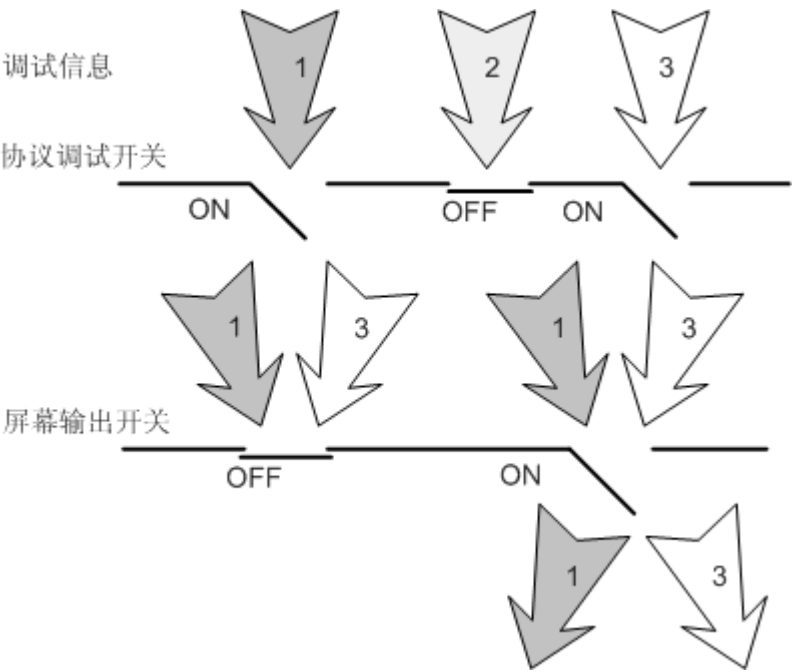
系统的命令行接口提供了种类丰富的调试功能，对于 USG 所支持的各种协议和功能，基本上都提供了相应的调试功能，可以帮助用户对错误进行诊断和定位。

调试信息的输出可以由两个开关控制：

- 协议调试开关，控制是否输出某协议的调试信息。
- 屏幕输出开关，控制是否在某个用户屏幕上输出调试信息。

二者关系如图 1-8 所示。当协议 1、3 的调试开关打开时，有调试信息输出。由于屏幕输出开关也打开，所以协议 1、3 的调试信息输出到屏幕上。协议 2 由于调试信息开关没有打开，所以没有调试信息输出。

图 1-8 调试信息输出示意图



信息严重等级

根据信息的严重等级或紧急程度，信息分为 8 个等级，信息越严重，其严重等级阈值越小。详细信息如表 1-6 所示。

表 1-6 信息严重等级的定义

| 显示值 | 严重等级      | 描述                                                  |
|-----|-----------|-----------------------------------------------------|
| 0   | Emergency | 设备致命的异常，系统已经无法恢复正常，必须重启设备。如程序异常导致设备重启，内存的使用被检测出错误等。 |

| 显示值 | 严重等级          | 描述                                                                        |
|-----|---------------|---------------------------------------------------------------------------|
| 1   | Alert         | 设备重大的异常，需要立即采取措施。如设备内存占用率达到极限等。                                           |
| 2   | Critical      | 设备重大的异常，需要采取措施进行处理或原因分析。如设备内存占用率超过低界线，温度超过低温告警线，检测出错误的信息（信息是由本设备内部生成）等。   |
| 3   | Error         | 错误的操作或设备的异常流程，不会影响后续业务，但是需要关注和原因分析。如用户的错误指令，用户密码错误，检测出错误协议报文（报文是由其他设备获得）。 |
| 4   | Warning       | 设备异常运转的异常点，可能引起业务故障的流程，需要引起注意。如用户对关闭路由进程，检测出错误协议报文等。                      |
| 5   | Notice        | 用于设备正常运转的关键操作信息。如邻居发现，协议状态机的正常跳转等。                                        |
| 6   | Informational | 用于设备正常运转的一般性操作信息。如用户使用 <b>display</b> 命令等。                                |
| 7   | Debugging     | 设备正常运转的一般性信息，用户无需关注。                                                      |

输出信息的严重等级是可配置的，根据配置的严重等级过滤信息时，仅输出严重等级阈值小于或等于所配置的严重等级阈值的信息，即输出所配置级别和比所配置级别更严重的信息。

例如，当配置严重等级阈值为 6 时，仅输出严重等级阈值为 0 ~ 6 的信息。

## 1.6 SNMP

SNMP（Simple Network Management Protocol）为网管和被管理设备的通信提供一套标准协议，保证网管能够正常的管理设备和接收设备的告警。

### 1.6.1 介绍

#### 定义

SNMP 是广泛用于 TCP/IP 网络的网络管理标准协议。SNMP 提供了一种通过运行网络管理软件的中心计算机（即网络管理工作站）来管理网元（如防火墙、路由设备、交换机等）的方法。

目前，设备支持 SNMPv3 版本，并兼容 SNMPv1 版本和 SNMPv2c 版本。

- SNMPv1 采用团体名（Community Name）认证。在每个 SNMPv1 报文中都包括 community 字段，在该字段中填入团体名，团体名起密码的作用。SNMPv1 假设如果网管侧使用正确的团体名进行认证，就认为该网管通过认证，是可靠的。

- SNMPv2c 也采用团体名认证。它在兼容 SNMPv1 的同时又扩充了 SNMP v1 的功能：它提供了更多的操作类型（GetBulk 和 InformRequest）；它支持更多的数据类型（Counter64 等）；它提供了更丰富的错误代码，能够更细致地区分错误。
- SNMPv3 提供了基于用户的安全模型（USM，User-Based Security Model）的认证机制。用户可以设置认证和加密功能，认证用于验证报文发送方的合法性，避免非法用户的访问；加密则是对网管站和 Agent 之间的传输报文进行加密，以免被窃听。

## 目的

随着网络业务的日益发展，现有的网络中，设备数量日益庞大，且这些设备与网络管理员所在的中心机房距离较远。当这些设备发生故障时，由于设备无法主动上报故障，导致网络管理员无法及时感知、及时定位和排除故障，从而导致网络的维护效率降低，维护工作量大大增加。

SNMP 解决了这个问题，通过 SNMP，网管可以远程询问设备的状态，设备能够在特定类型的事件发生时向网络管理工作站发出告警。

## 1.6.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 1212: Concise MIB definitions
- RFC 1155: Structure and identification of management information for TCP/IP-based internets
- RFC 1157: Simple Network Management Protocol (SNMP)
- RFC 1901: Introduction to Community-based SNMPv2
- RFC 1905: Protocol Operations for Version 2 of the Simple Network Management Protocol (SNMPv2)
- RFC 2271: An Architecture for Describing SNMP Management Frameworks
- RFC 2570: Introduction to Version 3 of the Internet-standard Network Management Framework (Status=3DINFORMATIONAL)
- RFC 2571: An Architecture for Describing SNMP Management Frameworks
- RFC 2572: Message Processing and Dispatching for the Simple Network Management Protocol (SNMP)
- RFC 2573: SNMP Applications
- RFC 2574: User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3)
- RFC 2575: View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP)
- RFC 2578: Structure of Management Information Version 2 (SMIv2)
- RFC 2579: Textual Conventions for SMIv2
- RFC 2580: Conformance Statements for SMIv2
- RFC 3410: An Architecture for Describing Simple Network Management Protocol (SNMP) Management Frameworks
- RFC 3411: An Architecture for Describing Simple Network Management Protocol (SNMP) Management frameworks

- RFC 3412: Message Processing and Dispatching for the Simple Network Management Protocol SNMP
- RFC 3413: Simple Network Management Protocol (SNMP) Applications
- RFC 3414: User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3)
- RFC 3415: View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP)
- RFC 3416: Version 2 of the Protocol Operations for the Simple Network Management Protocol (SNMP)
- RFC 3418: Management Information Base (MIB) for the Simple Network Management Protocol (SNMP)
- RFC 3512: Configuring Networks and Devices with Simple Network Management Protocol (SNMP).

### 1.6.3 可获得性

#### License 支持

不需要 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 1.6.4 应用场景

#### 应用环境

如果用户是统一规划建设新网络，建议根据应用场景选择 SNMP 协议的版本。如果用户是网络扩建或升级，建议用户根据网管使用的版本选择在设备上配置对应的 SNMP 协议版本，保证设备和网管的通信。

表 1-7 SNMP 各版本的应用环境

| SNMP 版本 | 应用环境                                                      |
|---------|-----------------------------------------------------------|
| SNMPv1  | 适用于小型网络，组网简单，对网络安全性要求不高或者网络环境本身比较安全，且比较稳定的网络，比如校园网，小型企业网。 |

| SNMP 版本 | 应用环境                                                                                                        |
|---------|-------------------------------------------------------------------------------------------------------------|
| SNMPv2c | 适用于大中型网络，对网络安全性要求不高或者网络环境本身比较安全（比如 VPN 网络），但业务比较繁忙，有可能发生流量拥塞的网络。<br><br>通过配置 Infrom 告警可以确保网管能够收到被管理设备发送的告警。 |
| SNMPv3  | 适用于各种规模的网络，尤其是对网络的安全性要求较高，确保合法的管理员才能对网络设备进行管理的网络。比如网管和被管理设备间的通信数据需要在公网上进行传输。                                |

典型应用

如图 1-9 所示，网管通过 SNMP 协议对设备进行管理。网管通过查询和接收设备发出的告警来了解设备的运行信息，在权限允许的情况下还可以设置设备的参数，对设备进行管理。

图 1-9 SNMP 典型应用



1.6.5 原理描述

SNMP 的发展历程

简单网络管理协议是目前 TCP/IP 网络中应用最为广泛的网络管理协议。1990 年 5 月，RFC 1157 定义了 SNMP 的第一个版本 SNMPv1。RFC 1157 和另一个关于管理信息的文件 RFC 1155 一起，提供了一种监控和管理计算机网络的系统方法。因此，SNMP 得到了广泛应用，并成为网络管理的事实上的标准。

SNMP 在 90 年代初得到了迅猛发展，同时也暴露出了明显的不足，如：难以实现大量的数据传输，缺少身份验证（Authentication）和加密（Privacy）机制。因此，1993 年发布了 SNMPv2，具有以下特点：

- 支持分布式网络管理
- 扩展了数据类型
- 可以实现大量数据的同时传输，提高了效率和性能
- 丰富了故障处理能力
- 增加了集合处理功能
- 加强了数据定义语言

但是 SNMPv2 并没有完全实现预期的目标，尤其是安全性能没有得到提高，如：身份验证（如用户初始接入时的身份验证、信息完整性的分析、重复操作的预防）、加密、授

权和访问控制、适当的远程安全配置和管理能力等都没有实现。1996 年发布的 SNMPv2c 是 SNMPv2 的修改版本，功能增强了，但是安全性能仍没有得到改善，继续使用 SNMPv1 的基于明文密钥的身份验证方式。

IETF SNMPv3 工作组于 1998 年元月提出了互联网建议 RFC 2271-2275，正式形成 SNMPv3。这一系列文件定义了包含 SNMPv1、SNMPv2 所有功能在内的体系框架和包含验证服务和加密服务在内的全新的安全机制，同时还规定了一套专门的网络安全和访问控制规则。可以说，SNMPv3 是在 SNMPv2 基础之上增加了安全和管理机制。

SNMPv3 的体系结构体现了模块化的设计思想，可以简单地实现协议功能的增加和修改。其特点：

- 适应性强：适用于多种操作环境，既可以管理最简单的网络，又能够满足复杂网络的管理需求。
- 扩充性好：可以根据需要增加模块。
- 安全性好：具有多种安全处理模块。

SNMPv3 主要有四个模块：

- 信息处理和控制模块：信息处理和控制模块（Message Processing And Control Model）在 RFC 2272 中定义，它负责 SNMP 报文的产生和分析，并判断信息在传输过程中是否要经过代理服务器等。在信息产生过程中，该模块接收来自调度器（Dispatcher）的 PDU，然后由用户安全模块在信息头中加入安全参数。在分析接收的信息时，先由用户安全模块处理信息头中的安全参数，然后再将解包后的 PDU 送给调度器处理。
- 本地处理模块：本地处理模块（Local Processing Model）的功能主要是进行访问控制、数据组包和中断。访问控制是指通过设置代理的有关信息使不同的管理站的管理进程在访问代理时具有不同的权限，通过 PDU（Protocol Date Unit）完成。常用的控制策略有两种：限定管理站可以向代理发出的命令或确定管理站可以访问代理的 MIB 中的具体内容。访问控制的策略必须预先设定。SNMPv3 通过使用带有不同参数的原语来灵活地确定访问控制方式。
- 用户安全模块：用户安全模块（User Security Model）则提供身份验证和数据加密服务。实现这个功能要求管理站和代理必须共享同一密钥。
  - 身份验证：身份验证是指代理（管理站）接到信息时首先必须确认信息是否来自有权的管理站（代理）并且信息在传输过程中未被改变。RFC2104 中定义了 HMAC，这是一种使用安全哈希函数和密钥来产生信息验证码的有效工具，在互联网中得到了广泛的应用。SNMP 使用的 HMAC 可以分为两种：HMAC-MD5-96 和 HMAC-SHA-96。前者的哈希函数是 MD5，使用 128 位 authKey 作为输入。后者的哈希函数是 SHA-1，使用 160 位 authKey 作为输入。
  - 加密：采用数据加密标准（DES）的密码组链接（CBC）码，使用 128 位的 privKey 作为输入。管理站使用密钥计算验证码，然后将其加入信息中，而代理则使用同一密钥从接收的信息中提取出验证码，从而得到信息。加密的过程与身份验证类似，也需要管理站和代理共享同一密钥来实现信息的加密和解密。
- 基于视图的访问控制模块：基于视图的访问控制模块（View-based Access Control Model）在 RFC 2515 中定义，它主要用于对用户组或者团体名实现基于视图的访问控制。用户必须首先配置一个视图，并指明权限。用户可以在配置用户或者用户组或者团体名的时候，加载这个视图达到限制读操作或写操作或告警操作的目的。

## SNMP 的工作机制

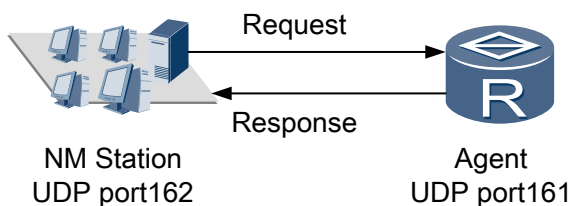
SNMP 是建立在 TCP/IP 基础上的应用层协议。

SNMP 定义了三类角色：

- 网管站：向被管理设备发送各种查询报文，以及接收被管理设备发送的告警。
- 被管理设备：接受网管的管理，产生和主动上报告警。
- Agent：驻留在被管理设备上的一个进程。Agent 的作用如下：
  - 接收、解析来自网管站的查询报文。
  - 根据报文类型对管理变量进行 Read 或 Write 操作，并生成响应报文，返回给网管站。
  - 根据各协议模块对告警触发条件的定义，在达到触发条件后，如进入、退出系统视图或设备重新启动等，相应的模块通过 Agent 主动向网管站发送告警，报告所发生的事件。

网管站与 Agent 的关系如图 1-10 所示。

图 1-10 SNMP 结构示意图



网管站通过在被管理设备中运行的 Agent 客户端上执行 GET-SET 操作来管理设备上的节点，设备上的节点由 MIB (Management Information Base) 来唯一标识。

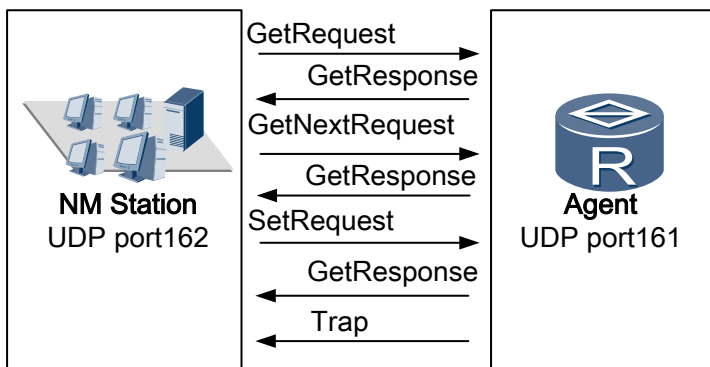
SNMP 有以下几种操作方式：

- 网管站通过 Get、Get-Next、Get-Bulk 操作来获取被管理设备相关信息。
- 网管站通过 Set 操作来设置被管理设备。
- Agent 主动上报 Trap 报文给网管站，可使网管站及时获取被管理设备的工作状态，从而网络管理员能够及时采取响应措施。

## SNMP 操作

SNMP 以 GET-SET 操作方式替代复杂的命令集，利用如图 1-11 的操作实现全部功能。

图 1-11 SNMP 操作示意图



说明

在 Agent 端是用知名端口 161 来接收 get 或 set 报文，而在网管端是用知名端口 162 来接收 Trap 报文。

SNMP 的操作如表 1-8 所示。

表 1-8 SNMP 操作说明

| 操作             | 功能                                                   |
|----------------|------------------------------------------------------|
| GetRequest     | 从某变量中取值，获取设备某功能节点状态，是网管向设备发出的请求。                     |
| GetNextRequest | 在 MIB 表项中取下一项值，获取设备某功能的另一节点状态，是网管向设备发出的请求。           |
| GetResponse    | 对 GetRequest、GetNextRequest、SetRequest 的响应，设备向网管的回应。 |
| GetBulk        | 该操作相当于连续执行多次 GetNextRequest 操作，是网管向设备发出的请求。          |
| SetRequest     | 设置具体变量的值，对设备功能节点状态进行调整，是网管向设备发出的指令。                  |
| Trap           | 报告事件信息，是设备主动向网管报告事件。                                 |

管理模型

SNMP 的管理体系，在网管站和 Agent 两侧进行信令交互。

- 网管站作为管理者，向 Agent 发送 SNMP 请求报文。
- Agent 通过查询设备端的 MIB 得到所要查询的信息，向网管站发送 SNMP 响应报文。
- 设备端的模块由于达到模块定义的告警触发条件，通过 Agent 向网管端工作站的网管站发送 Trap 消息，告知设备侧的出现的状况，这样便于网络管理人员及时的对于网络中出现的状况进行处理。

网络管理模型如图 1-12 所示。

图 1-12 SNMP 管理模型

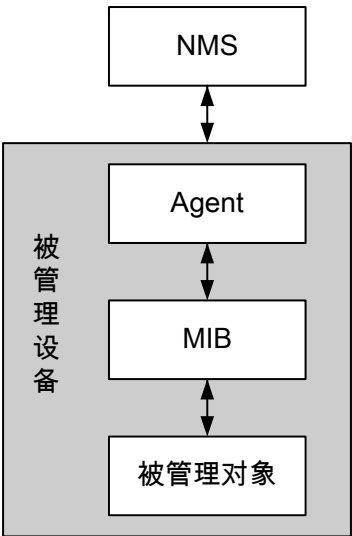
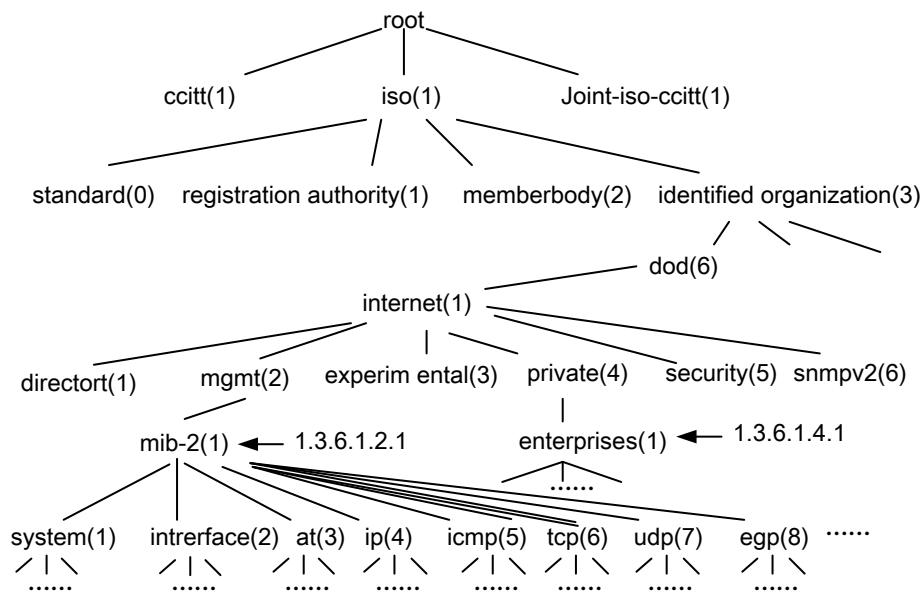






图 1-14 MIB 树结构



对象命名树的顶级对象有三个，即 ISO、ITU-T（即 ccitt）和这两个组织的联合体。在 ISO 的下面有 4 个结点，其中的一个（标号 3）是标识组织 organization。在其下面有一个 DoD（Department of Defense）的子树（标号是 6），再下面就是 Internet（标号是 1）。缺省能访问的 Viewdefault 视图就是这个 MIB 节点 Internet: {1.3.6.1}。

在 Internet 结点下面的第二个结点是 mgmt（管理），标号是 2。再下面是管理信息库，原先的结点名是 MIB。1991 年定义了新的版本 MIB-II，故结点名现改为 mib-2，其标识为 {1.3.6.1.2.1}，或 {Internet(1).2.1}。这种标识为对象标识符。

MIB 的定义与具体的网络管理协议无关。设备制造商可按照标准定义新的 MIB 节点。

## SMI

SMI（Structure of Management Information）为命名和定义 MIB 指定了一套规则，可定义 MIB 的对象标识、对象类型、访问级别和状态。目前 SMI 有两个版本，SMIv1 和 SMIv2。

以下是 SMI 中定义的标准数据类型：

- INTEGER
- OCTET STRING
- DisplayString
- OBJECT IDENTIFIER
- NULL
- IpAddress
- PhysAddress
- Counter
- Gauge

- TimeTicks
- SEQUENCE
- SEQUENCEEOF

## SNMP 各版本差异

表 1-9 SNMP 各版本特性差异

| 特性列表      | SNMPv1      | SNMPv2c     | SNMPv3                                        |
|-----------|-------------|-------------|-----------------------------------------------|
| 访问控制      | 基于团体名进行访问控制 | 基于团体名进行访问控制 | 基于用户和用户组进行访问控制                                |
| 认证加密      | 不支持         | 不支持         | 支持认证和加密。<br>认证：<br>● MD5<br>● SHA<br>加密：DES56 |
| 错误码       | 支持 6 个错误码   | 支持 16 个错误码  | 支持 16 个错误码                                    |
| Trap 告警   | 支持          | 支持          | 支持                                            |
| Inform 告警 | 不支持         | 支持          | 不支持                                           |
| GetBulk   | 不支持         | 支持          | 支持                                            |

各特性说明如下：

- 访问控制主要用来限制管理设备的用户的权限。通过该功能可以限制指定的用户管理设备上的指定节点，从而提升精细化管理。
- 认证加密主要是通过对网管和被管理设备传送的报文进行认证和加密，避免数据报文被窃取或篡改，从而提升数据传输的安全性。
- 错误码用来标识特定的故障现象，有助于管理员快速定位和解决故障，因此错误码越丰富越有利于管理员对设备进行管理。
- Trap 告警是被管理设备主动向网管发送告警。以便管理员能够及时发现设备的异常。被管理设备发送 Trap 告警后，不需要网管进行接收确认。
- Inform 告警也是被管理设备向网管主动发送告警。被管理设备发送 Inform 告警后，需要网管进行接收确认，被管理设备没有收到确认信息时会重复发送该告警，并且生成相应的告警日志，方便网管重启后，能够同步重启时间内发送的告警。系统没有收到 inform 告警的确认信息时，将告警暂时保存在内存中，因此使用 Inform 告警可能会占用较多的系统资源。

- GetBulk 主要方便管理员进行批量的 Get-next 操作。网络规模较大时，以节省管理员的工作量，提升管理效率。

## 1.7 LLDP

LLDP 功能用于获取邻居信息和拓扑发现。

### 1.7.1 介绍

#### 定义

LLDP（Link Layer Discovery Protocol，链路层发现协议）是 IEEE 802.1AB 中定义的邻居发现协议。通过采用 LLDP 技术，在网络规模迅速扩大时，网管系统可以快速掌握网络的拓扑信息和拓扑变化信息。

#### 目的

LLDP 是一种标准的链路发现协议，可以将本端设备的设备标识、主要能力等信息发布给与自己直连的邻居，邻居收到这些信息后将其以标准 MIB（Management Information Base，管理信息库）的形式保存起来，以供网络管理系统查询和判断链路的状况。

### 1.7.2 参考标准和协议

本特性的参考资料清单如下：

- IEEE 802.1AB-2005: Station and Media Access Control Connectivity Discovery
- ANSI/TIA-1057: Link Layer Discovery Protocol for Media Endpoint Devices
- IEEE Std 802.3-2005: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications

### 1.7.3 可获得性

#### License 支持

本特性无须 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                               |
|-------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V300R001 |

## 1.7.4 原理描述

### 1.7.4.1 LLDP 基本概念

#### LLDP 报文

LLDP 报文指封装有 LLDPDU（Link Layer Discovery Protocol Data Unit，链路层发现协议数据单元）的报文，它包含两种封装格式：Ethernet II 和 SNAP（Subnetwork Access Protocol，子网访问协议）。

1. Ethernet II 格式封装的 LLDP 报文

Ethernet II 格式封装的 LLDP 报文格式如[图 1-15](#) 所示。

**图 1-15** Ethernet II 格式的 LLDP 报文

|                               |  |
|-------------------------------|--|
| Destination MAC address       |  |
| Source MAC address            |  |
| Type                          |  |
| Data = LLDPDU<br>(1500 bytes) |  |
| FCS                           |  |

Ethernet II 格式封装的 LLDP 报文各字段的含义如[表 1-10](#) 所示。

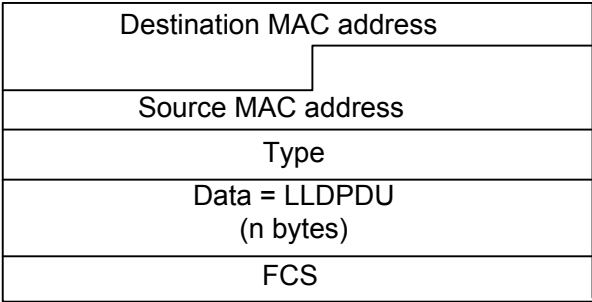
**表 1-10** Ethernet II 格式的 LLDP 报文各字段

| 字段                      | 含义                                                                |
|-------------------------|-------------------------------------------------------------------|
| Destination MAC address | 目的 MAC 地址，为固定的组播 MAC 地址 0x0180-C200-000E。                         |
| Source MAC address      | 源 MAC 地址，为接口 MAC 地址或设备 MAC 地址（如果有接口地址则使用接口 MAC 地址，否则使用设备 MAC 地址）。 |
| Type                    | 报文类型，为 0x88CC。                                                    |
| Data                    | 数据，为 LLDPDU。                                                      |
| FCS                     | 帧检验序列。                                                            |

2. SNAP 格式封装的 LLDP 报文

SNAP 格式封装的 LLDP 报文格式如[图 1-16](#) 所示。

图 1-16 SNAP 格式的 LLDP 报文



SNAP 格式封装的 LLDP 报文各字段的含义如表 1-11 所示。

表 1-11 SNAP 格式的 LLDP 报文各字段

| 字段                      | 含义                                                                |
|-------------------------|-------------------------------------------------------------------|
| Destination MAC address | 目的 MAC 地址，为固定的组播 MAC 地址 0x0180-C200-000E。                         |
| Source MAC address      | 源 MAC 地址，为接口 MAC 地址或设备 MAC 地址（如果有接口地址则使用接口 MAC 地址，否则使用设备 MAC 地址）。 |
| Type                    | 报文类型，为 0xAAAA-0300-0000-88CC。                                     |
| Data                    | 数据，为 LLDPDU。                                                      |
| FCS                     | 帧检验序列。                                                            |

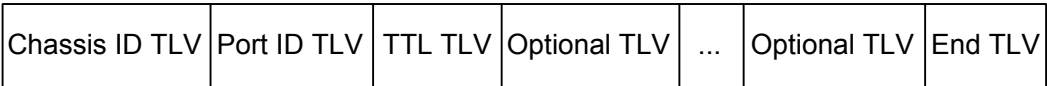
LLDPDU

LLDPDU 指封装在 LLDP 报文数据部分的数据单元。

设备首先将本地信息封装成 TLV 格式，再由若干个 TLV 组合成一个 LLDPDU 封装在 LLDP 报文的数据部分进行传送。

LLDPDU 报文格式如图 1-17 所示。

图 1-17 LLDPDU



如图 1-17 所示，每个 LLDPDU 均可携带多种 TLV。其中 Chassis ID TLV、Port ID TLV、TTL TLV 和 End TLV 这四种是必须携带的，其余的 TLV 则为可选携带。

## TLV

TLV 是组成 LLDPDU 的单元，每个 TLV 都代表一个信息。LLDP 可以封装的 TLV 包括基本 TLV、IEEE 802.1 组织定义 TLV、IEEE 802.3 组织定义 TLV。

基本 TLV 是网络设备管理中的基础 TLV。IEEE 802.1 组织定义 TLV 和 IEEE 802.3 组织定义 TLV 则是由标准组织定义的 TLV，用于增强对网络设备的管理，可根据实际需要选择是否在 LLDPDU 中发送。

### 1. 基本 TLV

在基本 TLV 中，包含几种实现 LLDP 功能时必须选的 TLV，即必须在 LLDPDU 中发布，如表 1-12 所示。

表 1-12 基本 TLV

| TLV 名称              | 含义                                          | 是否必须发布 |
|---------------------|---------------------------------------------|--------|
| Chassis ID          | 发送设备的 MAC 地址。                               | 是      |
| Port ID             | LLDPDU 发送端的接口。                              | 是      |
| Time To Live        | 本设备信息在邻居设备上的存活时间。                           | 是      |
| Port Description    | 以太网接口的描述字符串。                                | 否      |
| System Name         | 设备的名称。                                      | 否      |
| System Description  | 系统描述。                                       | 否      |
| System Capabilities | 系统的主要功能以及已启用的功能项。                           | 否      |
| Management Address  | 管理地址，以及对应的接口号和 OID（Object Identifier，对象标识）。 | 否      |
| End TLV             | 标识 LLDPDU 结束。                               | 是      |

### 2. IEEE 802.1 组织定义 TLV

IEEE 802.1 组织定义 TLV 的内容如表 1-13 所示。

表 1-13 IEEE 802.1 组织定义的 TLV

| TLV 名称       | 含义           |
|--------------|--------------|
| Port VLAN ID | 接口的 VLAN ID。 |

### 3. IEEE 802.3 组织定义 TLV

IEEE 802.3 组织定义 TLV 的内容如表 1-14 所示。

表 1-14 IEEE 802.3 组织定义的 TLV

| TLV 名称                       | 含义                                               |
|------------------------------|--------------------------------------------------|
| MAC/PHY Configuration/Status | 接口的速率和双工状态、是否支持接口速率自动协商、是否已启用自动协商功能以及当前的速率和双工状态。 |

## 管理地址

管理地址是供网络管理系统标识网络设备并进行管理的地址。

管理地址明确地标识一台设备，有利于网络拓扑的绘制，便于网络管理。管理地址被封装在 LLDP 报文的 Management Address TLV 中向外发布。

如果设备配置了全局管理地址，则系统将选取全局管理地址作为设备的管理地址。

如果设备没有配置全局管理地址，则对于二层接口，管理地址为该接口允许通过的、配置有 IP 地址的最小 VLAN 对应的 Vlanif 接口的主 IP 地址，但是如果所有 Vlanif 接口都没有配置 IP 地址，则不发布管理地址；对于三层接口，LLDP 管理地址为该接口配置的主 IP 地址，如果没有配置 IP 地址，则不发布管理地址。

### 1.7.4.2 LLDP 工作机制

#### LLDP 的工作模式

LLDP 包括以下四种工作模式：

- TxRx：既发送也接收 LLDP 报文。
- Tx：只发送不接收 LLDP 报文。
- Rx：只接收不发送 LLDP 报文。
- Disable：既不发送也不接收 LLDP 报文。

当接口上 LLDP 的工作模式发生变化的时候，接口将对协议状态机进行初始化操作。通过配置接口初始化的延迟时间，可以避免由于工作模式频繁改变而导致接口不断进行初始化。

#### LLDP 报文的发送机制

当接口工作在 TxRx 或 Tx 模式时，设备会周期性地向邻居设备发送 LLDP 报文。

当设备的本地配置发生变化时，设备会立即发送 LLDP 报文，以确保本地信息的变化情况可以被尽快通知给邻居设备。同时，为了防止由于本地信息的频繁变化而引起的 LLDP 报文的大量发送，设备在先后发送的 LLDP 报文之间都需要有一段时间的延迟。

当设备的工作模式由 Disable/Rx 切换为 TxRx/Tx，或者发现了新的邻居设备（即收到一个新的 LLDP 报文且本地尚未保存发送该报文设备的信息）时，设备将自动启用快速发送机制，即将 LLDP 报文的发送周期缩短为 1 秒，并连续发送指定数量的 LLDP 报文后再恢复为正常的发送周期。

#### LLDP 报文的接收机制

当接口工作在 TxRx 或 Rx 模式时，设备会对收到的 LLDP 报文及其携带的 TLV 进行有效性检查，判断 LLDPDU 报文中的前三个 TLV 和最后一个 TLV 是否合法。如果有任



何一个 TLV 不合法，则设备会进入 LLDP 报文丢弃流程。检查通过后，设备会将邻居信息保存到本地，并根据 TTL TLV 中 TTL 的值来设置邻居信息在本地设备上的老化时间，若该值为零，则立刻老化该邻居信息。

设备对收到的 LLDP 报文及其携带的 TLV 进行有效性检查的具体流程如下所示：

1. 第一个 TLV 检查：
  - a. 第一个 TLV 是 Chassis ID TLV，TLV 的类型为 1。
  - b. Chassis ID TLV 的字符串长度在 2-256 之间。
2. 第二个 TLV 检查：
  - a. 第二个 TLV 是 Port ID TLV，TLV 的类型为 2。
  - b. Port ID TLV 的信息字符串长度在 2-256 之间。
3. 第三个 TLV 检查：
  - a. 第三个 TLV 是 TTL，TLV 的类型为 3。
  - b. TTL 的长度等于 2。
  - c. TLV information string 的前两个字节是 rxTTL
    - a. 如果 rxTTL 为 0，表明是 shutdown 报文。应该直接被 MIB 处理，进一步的 LLDP 处理需要被终止
    - b. 如果 rxTTL 不为 0，LLDPDU 确认过程继续。
4. 最后一个 TLV 检查：
  - a. 最后一个 TLV 必须是 END TLV。

## 1.8 CWMP

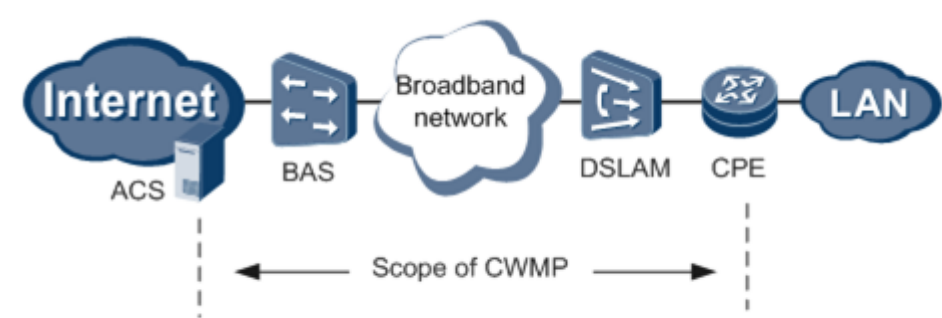
CWMP 是一种广域网管理协议，可实现 ACS 对 CPE 的集中管理，主要功能为配置管理、版本管理、远程监控和诊断。

### 1.8.1 介绍

#### 定义

CWMP（CPE WAN Management Protocol，CPE 广域网管理协议）是由 DSL（Digital Subscriber's Line，数字用户线路）论坛发起开发的技术规范之一，编号为 TR069，所以又被称为 TR069 协议。它提供了对下一代网络中家庭网络设备进行管理配置的通用框架、消息规范、管理方法和数据模型。CWMP 主要应用于 DSL 接入网络环境。在 DSL 接入网络中，由于用户设备数量繁多、部署分散，通常位于用户侧，不易进行设备的管理和维护，CWMP 提出通过 ACS（Auto-Configuration Server，自动配置服务器）对 CPE（Customer Premises Equipment，用户侧设备）进行远程集中管理，解决 CPE 设备的管理困难，节约维护成本，提高问题解决效率。

图 1-18 CWMP 典型组网图



通过 CWMP，可以实现以下功能：

- 配置管理：可以实现对 CPE 的“零配置安装”或进行远程控制配置参数的修改。
- 版本管理：可以提供对 CPE 的版本软件和固件的管理，如版本软件下载（远程升级支持断点续传）、配置文件备份和恢复等。
- 远程监控：CWMP 定义了 ACS 对 CPE 的状态和性能进行监测的方法，可以实时查询 CPE 的状态。
- 诊断功能：CWMP 提供诊断和报告功能。通过设置诊断参数触发诊断处理，获取 CPE 的诊断结果。如果发现 CPE 出现故障，维护人员在远端就可以对故障进行定位。

目的

通常，用户侧设备（例如网关、机顶盒）种类繁多，部署分散。当需要配置变更或者故障处理时，通常只有运营商的维护人员上门处理，管理非常不便。CWMP 的出现解决这样一个难题。它提供了对用户终端设备进行管理配置的通用框架和协议，实现了从网络侧对用户侧设备进行远程集中管理。通过部署 CWMP，用户侧相关设备的配置、诊断、升级等工作，均可以由 ACS 来完成，大大节省了维护成本。

1.8.2 可获得性

License 支持

不需要 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                           |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00 |

## 1.8.3 原理描述

### CWMP 方法

ACS 对 CPE 的管理和监控是通过一系列的操作来实现的，这些操作在 CWMP 协议里称为 RPC 方法。

主要方法如下：

- **Get:** ACS 使用该方法可以获取 CPE 上参数的值。
- **Set:** ACS 使用该方法可以设置 CPE 上参数的值。
- **Inform:** CPE 与 ACS 建立连接前，发送的通告信息。CWMP 中，该连接通常是 CPE 发起，CPE 可使用 Inform RPC 方法建立一条到 ACS 的连接。同时，CPE 也支持 ACS 发起的连接，ACS 可向 CPE 发出有关配置修改的通知。
- **Download:** 为了保证 CPE 端硬件的升级以及厂商配置文件的自动下载，ACS 使用该方法可以要求 CPE 到指定的 URL 下载指定的文件来更新 CPE 的本地文件。
- **Upload:** 为了方便 ACS 对 CPE 端的管理，ACS 使用该方法可以要求 CPE 将指定的文件上传到 ACS 指定的位置。

### CWMP 交互过程

CPE 与 ACS 之间的通信大致可分为：ACS 发现阶段和连接建立阶段。

ACS 发现阶段：

- CPE 必须知道 ACS 的 URL，该信息可以直接配置在 CPE 上。
- 当 CPE 知道 ACS 的 URL 后，就可以随时向 ACS 发起连接。

连接建立阶段：

CPE 作为 HTTP 客户端，其 SOAP 请求通过 HTTP Post 方式发给 ACS；而 ACS 作为 HTTP 服务端，通过 HTTP Response 方式回应 CPE 的请求。

下面以 CPE 重启后配置恢复为例，其通信过程如下：

- CPE 向发现的 ACS 发起建立 TCP 连接。
- CPE 通过向 ACS 发送 Inform 请求，上报设备信息，请求建立 CWMP 连接。
- 若 CPE 通过认证，ACS 回应 Inform Response，CWMP 连接建立成功。
- ACS 通过向 CPE 调用 Download 方法，告知 CPE 将要下载的文件名、下载地址等信息。
- CPE 通过向 ACS 发送 Download Response，告知下载状态（下载还未完成）。
- CPE 与指定的文件服务器建立连接。
- CPE 从服务器下载配置文件。
- 文件下载完成后，CPE 向 ACS 发送 TransferComplete 消息，告知 ACS 下载文件完成。
- ACS 发送 TransferCompleteResponse 回应 CPE。
- CPE 拆除 TCP 连接，完成本次连接。

## 1.9 NetStream

NetStream 对网络流量进行统计，并定期向 NSC 发送统计数据。这些数据可用于计费、网络管理和规划等。

### 1.9.1 介绍

#### 定义

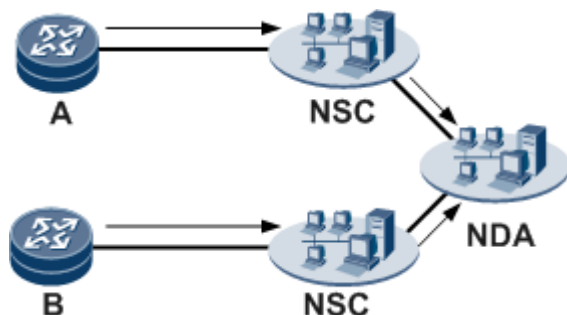
NetStream 主要是基于“流”的概念，通过提供基本而详细的流统计信息，为精细计费提供支持。在网络层的常规交换中，每个进来的信息包都是逐个单独处理的，有一系列的功能来进行访问列表检查、收集记帐数据并交换信息包。

NetStream 的流定义为在源、目的端点间的一系列单方向的数据包，端点由 IP 地址和传输层的端口号决定。此外，NetStream 还使用了 IP 协议类型、ToS 和输入接口来唯一地标识一个流，即如下的七元组：

- 源 IP 地址
- 目的 IP 地址
- 源端口号
- 目的端口号
- 协议类型
- 服务类型（报文的 ToS 值）
- 输入/输出接口

图 1-19 是数据采集和分析的一个过程，首先 NDE 负责对流量进行采集，图中的设备 A 和设备 B 是 NDE 设备，之后 NDE 把采集到的关于流的详细信息输出给 NSC，NSC 收集和存储 NDE 发来的流量统计数据信息，并进行过滤和聚合后，输出给 NDA，NDA 对数据进一步的聚合、排序并将各种数据以图形的形式显示出来。通过对 NDA 输出的结果进行分析，为网络计费，网络规划，网络监控，应用监控、分析和故障定位等提供依据。

图 1-19 NetStream 数据采集和分析图



#### 目的

Internet 的高速发展，为用户提供了更高的带宽和可预测的 QoS，随着网络支持的业务和应用日渐增多，运营商需要对网络进行更加细致的管理和计费，这样就对流量统计分

析提出了更高的要求。传统的流量统计存在着如下表所示的现状，这样就急切的需要一种新技术来更好的支持网络业务的发展。

表 1-15 传统的流量统计的实现方法和特点

| 名称       | 实现方法                                         | 特点                                                  |
|----------|----------------------------------------------|-----------------------------------------------------|
| 基于 IP 计数 | 在路由表中存放计数器索引，对通过路由器的字节和包分别计数。                | 统计的信息简单，无法针对多种信息进行统计。                               |
| 使用 ACL   | 通过 ACL 精确的匹配流，匹配后进行计数。                       | 要求 ACL 的容量很大，对于 ACL 规则以外的流没有办法统计。                   |
| SNMP 协议  | 使用网管协议，统计路由器上简单的统计功能，比如端口计数、IP 计数、ACL 匹配计数等。 | 功能不强。要不断的通过轮询向网管查询，对 CPU 和网络的压力都很大。                 |
| 端口镜像     | 通过端口镜像，把流量复制一份，给专用的 Server 进行统计。             | 成本高，需要购买专用的 Server 进行统计，同时消耗路由器的一个端口，对于无法镜像的端口无能为力。 |
| 物理层复制    | 在物理层通过分光器或者其他设备复制流量，同样要送到专用的 Server 进行统计。    | 成本高，需要购买专用的 Server 进行统计，同时还需要购买专用的硬件设备。             |

NetStream 是华为公司的专利技术，是一种基于网络流信息的统计与发布技术。NetStream 可以对网络中的通信量和资源使用情况进行分类和统计，实现对各种业务和不同的 QoS 进行管理和计费。

1.9.2 可获得性

License 支持

不需要 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 1.9.3 原理描述

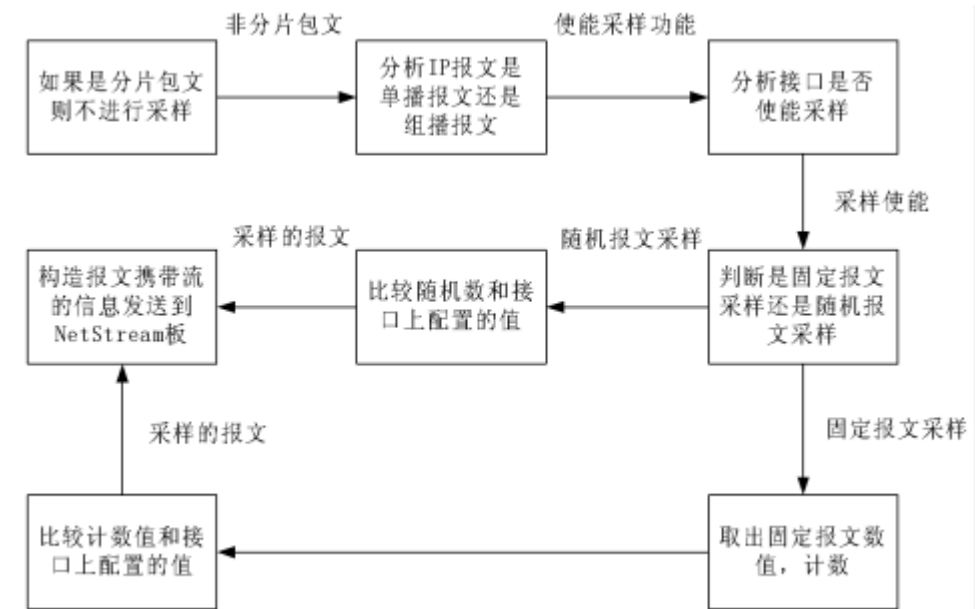
#### 采样的概念

流的提取在使能 NetStream 的接口上进行，流的建立、维护、老化、聚合、生成 export data 都通过 CPU 处理完成。进行流的提取之前，首先要在接口上使能 NetStream。如果接口使能 NetStream，系统会在相应的接口信息表中存放 NetStream 的相关信息。如果每一个报文都参与 NetStream 流的建立和计数，则对性能的影响比较巨大，尤其对于速率比较高的接口。因此，经过接口的报文，需要经过采样，采样选中的报文的相关信息被送到 NetStream 板进行处理，采样的比率越低对设备性能的影响越小。

#### 采样处理流程

处理的流程如图 1-20 所示。

图 1-20 NetStream 采样流程



从上面图中可以看出，在接口上配置的和 NetStream 相关的信息如下表所示。

| 名称       | 宽度     | 作用                                                      |
|----------|--------|---------------------------------------------------------|
| 单播报文使能标志 | 1bit   | 表示对目的 IP 地址是单播的报文是否使能                                   |
| 组播报文使能标志 | 1bit   | 表示对目的 IP 地址是组播的报文是否使能                                   |
| 采样使能标志   | 1bit   | 表示该接口是否使能采样                                             |
| 采样掩码     | 15bits | 随机采样和根据报文个数采样时的比较值                                      |
| 接口索引     | 16bits | 包括 4bits 的槽位信息和 12bits 的接口索引信息，这个字段作为 NetStream 流的入接口信息 |

上述信息包括上行和下行两个独立的部分，在接口上对报文采样支持入接口和出接口两种方式。

## 流的老化

由于网络上的流是短时间阵发的，在几秒时间就会产生数万条流，而 NDE 的内存的容量是一定的，这样就需要把当前的部分流删除，为后面到来的流提供内存空间，这个过程称为老化。

流老化的分类：

- 定时老化  
对不活跃时间（从最后一个报文到达当前的时间）超过 **inactive timeout**，或活跃时间（从第一个报文流过到当前的时间）超过 **active timeout** 的流进行老化。在流缓存区中的流超过 **active timeout** 时间，不会立刻进行老化，直到一个新的流到来时，系统对流进行老化；在超过 **inactive timeout** 时间后，系统立即对流进行老化，而不管此时是否仍然处于活跃老化时间内。活跃老化时间和不活跃老化时间可以设定，也可以使用默认值，默认的活跃老化时间是 30 分钟，不活跃老化时间是 30 秒。
- 由 TCP 连接的 FIN 和 RST 报文触发老化  
对于 TCP 连接，当有标志为 FIN 或 RST 的报文发送时，表示一次会话结束。因此当一条已经存在的 TCP 协议 NetStream 流中流过一条标志为 FIN 或 RST 的报文时，可以立即把相应的 NetStream 流老化掉。但是假如一条流的第一个报文就是 TCP 的 FIN 或 RST 报文，应该按正常的流程创建一条新流，然后继续执行，不进行老化。
- 统计字节超过限制时老化  
流缓存区中的流需要记录流过的报文字节数，当字节数量超过定义的变量上限时，缓存区就会溢出。所以系统在检测到某条流的字节统计超过限制（例如 3.9G 字节）时，就立即把该流老化掉。
- 命令行强制老化  
用户可以通过执行命令 **reset ip netstream statistics** 将当前流缓存区中的流全部老化掉。

## 流的输出

原始流的输出方式：老化后的 NetStream 流中信息被系统统计后，组成 UDP 报文发送出去，直接送入到 NSC，NSC 可以得到流的详细信息，可以对这些流记录进行更为灵活的后续处理，但这样增加了网络带宽和设备的 CPU 占有率，而且为了存储这些信息，需要占用大量的存储容量，增加了设备的压力。

聚合流的输出方式：老化后的 NetStream 流中信息被系统统计后，原始信息按照一定的规则进行分类、合并后生成聚合的信息，在系统中的定时器到时后，聚合流通过 UDP 报文发送出去。通过对原始流的聚合，可以明显减少网络带宽、CPU 占用率和存储介质空间。目前支持如表 1-16 所示的聚合方式。

表 1-16 聚合方式

| 聚合方式                   | 说明                                                                                                             |
|------------------------|----------------------------------------------------------------------------------------------------------------|
| as                     | 自治系统聚合，根据 NetStream 流的源、目的自治系统号，输入接口索引，输出接口索引，4 个关键值进行分类，具有相同关键值的流合并成一条聚合的流，并对应一条聚合信息。                         |
| as-tos                 | 自治系统-TOS 聚合，根据 NetStream 流的源、目的自治系统号，输入接口索引，输出接口索引，ToS，5 个关键值进行分类，具有相同关键值的流合并成一条聚合的流，并对应一条聚合信息。                |
| protocol-port          | 协议-端口聚合，根据 NetStream 流的协议号，源端口，目的端口，3 个关键值进行分类，具有相同关键值的流合并成一条聚合流，并对应一条聚合记录。                                    |
| protocol-port-tos      | 协议-端口-ToS 聚合，根据 NetStream 流的协议号，源端口，目的端口，ToS，输入接口索引，输出接口索引，6 个关键值进行分类，具有相同关键值的流合并成一条聚合流，并对应一条聚合记录。             |
| source-prefix          | 源前缀聚合，根据 NetStream 流的源自治系统号，源掩码长度，源前缀，输入接口索引，4 个关键值进行分类，具有相同的关键值的流合并成一条聚合的流，并对应一条聚合记录。                         |
| source-prefix-tos      | 源前缀-ToS 聚合，根据 NetStream 流的源自治系统号，源掩码长度，源前缀，ToS，输入接口索引，5 个关键值进行分类，具有相同的关键值的流合并成一条聚合的流，并对应一条聚合记录。                |
| destination-prefix     | 目的前缀聚合，根据 NetStream 流的目的自治系统号，目的掩码长度，目的前缀，输出接口索引，4 个关键值进行分类，具有相同的关键值的流合并成一条聚合的流，并对应一条聚合记录。                     |
| destination-prefix-tos | 目的前缀-Tos 聚合，根据 NetStream 流的目的自治系统号，目的掩码长度，目的前缀，ToS，输出接口索引，5 个关键值进行分类，具有相同的关键值的流合并成一条聚合的流，并对应一条聚合记录。            |
| prefix                 | 前缀聚合，根据 NetStream 流的源、目的自治系统号，源、目的掩码长度，源、目的前缀，输入接口索引，输出接口索引，8 个关键值进行分类，具有相同的关键值的流合并成一条聚合的流，并对应一条聚合记录。          |
| prefix-tos             | 前缀-ToS 聚合，根据 NetStream 流的源、目的自治系统号，源、目的掩码长度，源、目的前缀，ToS，输入接口索引，输出接口索引，9 个关键值进行分类，具有相同的关键值的流合并成一条聚合的流，并对应一条聚合记录。 |
| mpls-label             | MPLS 标签聚合，根据 MPLS label 对流进行分类，具有相同标签的流合并成一条聚合流，并对应一条聚合记录。                                                     |



## 1.10 NTP

NTP 基于 UDP 传输，其目的是对网络内所有时钟设备进行时钟同步。运行 NTP 的系统可以主动同步其它系统，也可以接受其它系统的同步。

### 1.10.1 介绍

#### 定义

NTP（Network Time Protocol，网络时间协议）是用于互联网中时钟同步的应用层协议，其用途是在分布式时间服务器和客户端之间进行时钟同步，把主机的时钟同步到某些时间标准。

#### 目的

NTP 的目的是对网络内所有具有时钟的设备进行时间同步，使网络内所有设备的时间基本保持一致，从而使设备能够提供基于统一时间的多种应用。如：交换机、PC、路由设备等，均可用于时间同步。

### 1.10.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 1305: NTP 模块需求规格提出的基础

### 1.10.3 可获得性

#### License 支持

不需要 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 1.10.4 原理描述

NTP 主要用于在分布式时间服务器和客户端之间进行时间同步，把主机的时间同步到某些时间标准。服务器和客户端的概念是相对而言的，提供时间标准的设备称为时间服务器，接收时间服务的设备称为客户端。对于运行 NTP 的本地系统，既可以接受来自其

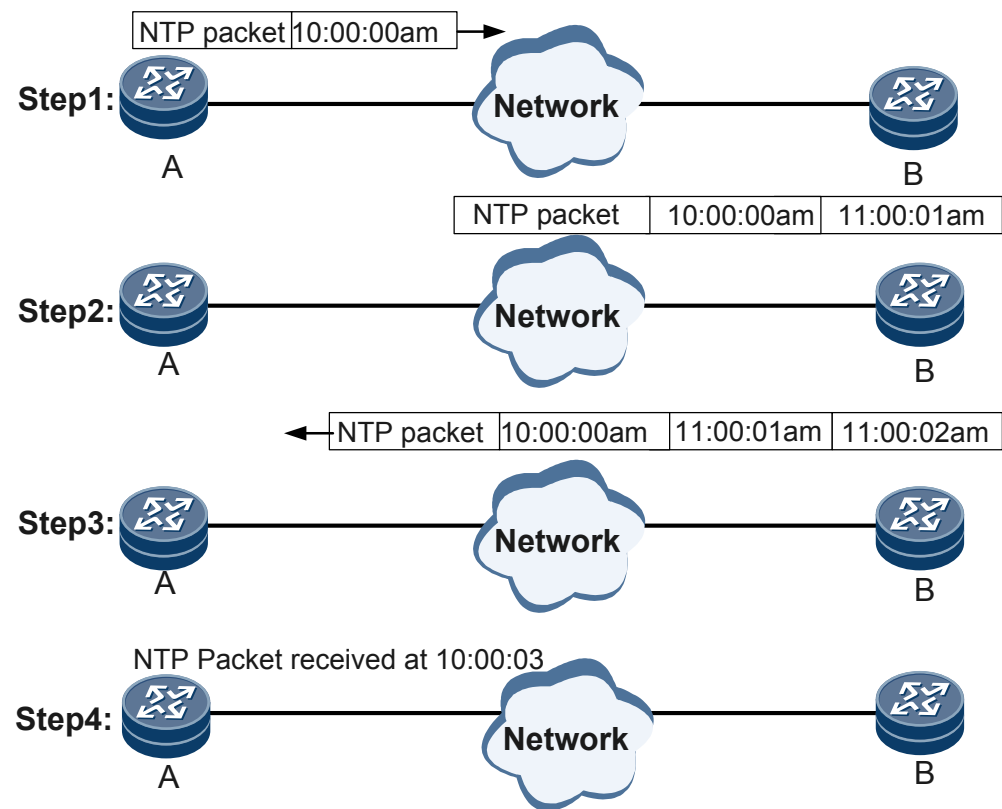
他时钟源的同步，也可以作为时钟源去同步别的时钟，并且可以通过交换 NTP 报文互相同步。NTP 基于 UDP 传输，使用端口号 123。

实现过程

NTP 的实现过程如图 1-21 所示：

设备 A 和设备 B 通过广域网相连，它们都有自己独立的系统时钟，通过 NTP 实现系统时钟自动同步。

图 1-21 NTP 实现过程图



作如下假设：

- 在设备 A 和设备 B 的系统时钟同步之前，设备 A 的时钟设定为 10:00:00am，设备 B 的时钟设定为 11:00:00am。
- 设备 B 作为 NTP 时间服务器，设备 A 的时钟与设备 B 的时钟同步。
- 数据包在设备 A 和设备 B 之间单向传输需要 1 秒。
- 设备 A 和设备 B 处理 NTP 数据包的时间都是 1 秒。

系统时钟同步的工作过程如下：

- 设备 A 发送一个 NTP 报文给设备 B，该报文中带有它离开设备 A 时的时间戳 10:00:00am（T1）。
- 此 NTP 报文到达设备 B 时，设备 B 加上到达时间戳 11:00:01am（T2）。

- 此 NTP 报文离开设备 B 时，设备 B 再加上离开时间戳 11:00:02am (T3)。
- 设备 A 接收到该响应报文时，加上新的时间戳 10:00:03am (T4)。

至此，设备 A 拥有足够信息来计算以下两个重要参数：

- NTP 消息来回一个周期的时延： $\text{Delay} = (T4 - T1) - (T3 - T2)$ 。
- 设备 A 相对设备 B 的时间差： $\text{Offset} = ((T2 - T1) + (T3 - T4)) / 2$ 。

设备 A 根据这些信息来设定自己的时钟，实现与设备 B 的时钟同步。

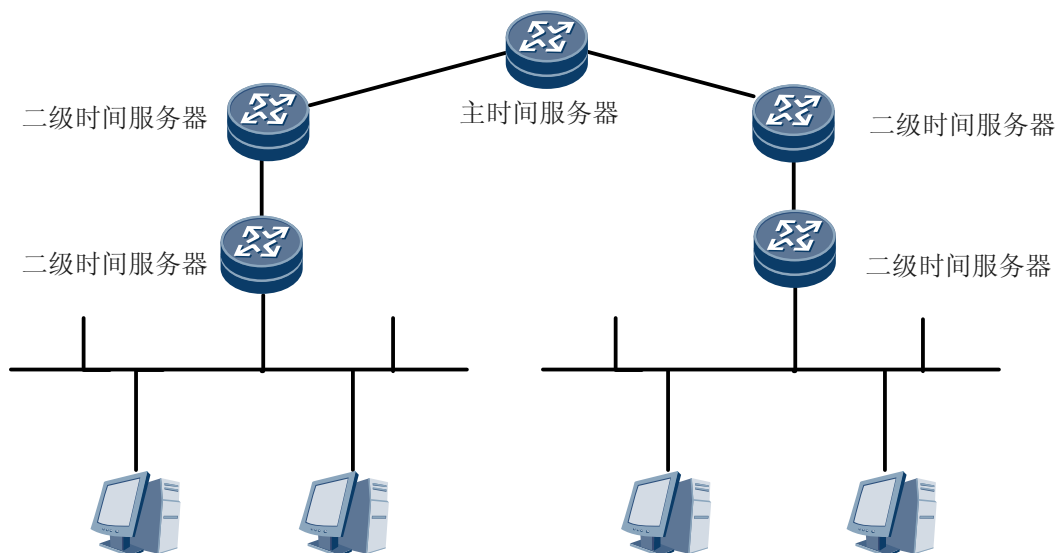
#### 说明

以上是对 NTP 工作原理的简要描述，NTP 使用标准的 RFC1305 中的算法来确保时钟同步的精确性。

## 网络结构

具体的组网方式是由主时间服务器、二级时间服务器、客户端和它们之间互连的传输路径组成，如图 1-22 所示。

图 1-22 NTP 网络结构



- 主时间服务器直接同步到主参考时钟，主参考时钟通常是 Radio Clock 或卫星定位系统等。
- 二级时间服务器通过网络中的主时间服务器或者其它二级服务器取得同步。二级时间服务器通过 NTP 将时间信息传送到网络内部的其它主机。

在正常情况下，同步子网中的主服务器和次级服务器呈现出一种分层主从结构，在这种分层结构中，主服务器位于根部，次级服务器向叶子节点靠近，层数递增，准确性递减。

## 工作模式

客户/服务器模式：

- **客户模式：**运行在客户模式的主机定期向服务器端发送报文，不管服务器端是否可达及服务器端的层数。运行在这种模式的主机，通常是网络内部的工作站，它可以依照对方的时钟进行同步，但不会修改对方的时钟。
- **服务器模式：**运行在服务器模式的主机接收并回应报文。运行在服务器模式的主机，通常是网络内部的时间服务器，它可以向客户端提供同步信息，但不会修改自己的时钟。

运行在客户模式的主机在重新启动时和重新启动后定期向运行在服务器模式的主机发送 NTP 报文。服务器收到客户端的报文后，首先将报文的 **目的 IP 地址** 和 **目的端口号** 分别与其 **源 IP 地址** 和 **源端口号** 相交换，再填写所需的信息，然后把报文发送给客户端。服务器在客户端发送请求之间无需保留任何状态信息，客户端根据本地情况自由管理发送报文的时间间隔。

对等体模式：

对等体模式下，主动对等体和被动对等体可以互相同步，等级低（层数大）的对等体向等级高（层数小）的对等体同步。

- **主动对等体：**运行在这一模式下的主机定期发送报文，不考虑它的对等体是否可达及对等体的层数。运行在这一模式下的主机可以向对方提供同步信息，但可以依照对方的时间信息同步本地时钟。
- **被动对等体：**运行在这一模式的主机接收并回应报文。运行在被动对等体模式的主机可以向对方提供同步信息，但可以依照对方的时间信息同步本地时钟。
- **运行被动对等体模式的必备条件：**本机接收的报文来自一个运行在主动对等体模式下的对等体，且该对等体的层数等于或低于本机并路由可达。

#### 说明

被动对等体模式运行在同步子网中层次较低层上时。这种模式下，不需要预先知道对等体的特性，因为只有当本机收到 NTP 报文时才建立连接及相关的状态变量。

广播模式：

- 运行在广播模式下，周期性向广播地址 255.255.255.255 发送时钟同步报文，不管它的对等体是否可达或层数为多少。运行在广播模式的主机通常是网络内运行高速广播介质的时间服务器，向所有对等体提供同步信息，但不会修改自己的时钟。
- 客户端侦听来自服务器的广播消息包。当接收到第一个广播消息包后，为估计网络延迟，客户端先启用一个短暂的服务器/客户端模式与远程服务器交换消息，之后恢复广播模式，继续侦听广播消息包的到来，根据到来的广播消息包对本地时钟再次进行同步。

广播模式应用在有多台工作站、不需要很高的准确度的高速网络。典型的情况是网络中的一台或多台时间服务器定期向工作站发送广播报文，广播报文在毫秒级的延迟基础上确定时间。

组播模式：

- 服务器端周期性向组播地址发送时钟同步报文。运行在组播模式的主机通常是网络内运行高速广播介质的时间服务器，向所有对等体提供同步信息，但不会修改自己的时钟。
- 客户端侦听来自服务器的组播消息包。当接收到第一个组播消息包后，为估计网络延迟，客户端先启用一个短暂的服务器/客户端模式与远程服务器交换消息，之后，客户端恢复组播模式，继续侦听组播消息包的到来，根据到来的组播消息包对本地时钟进行同步。

## 安全机制

当同步子网中的一台时间服务器发生意外或恶意的数据篡改或破坏时，通常不应该导致子网中其它时间服务器的计时错误。因此，NTP 还提供了两种安全机制：访问权限和 NTP 验证功能。这样就对网络的安全性提供了保障。

访问权限：

通过设置访问权限保护本地 NTP 服务，USG 提供了一种比较简单的安全措施。

设备提供 4 个等级的访问限制，当 1 个 NTP 访问请求到达本地时，按照最小访问限制到最大访问限制依次匹配，以第 1 个匹配的为准，匹配顺序如下：

- **peer:**（最小访问限制）可以对本地 NTP 服务进行时间请求和控制查询，本地时钟也可以同步到远程服务器。
- **server:** 可以对本地 NTP 服务进行时间请求和控制查询，但本地时钟不会同步到远程服务器。
- **synchronization:** 只允许对本地 NTP 服务进行时间请求。
- **query:**（最大访问限制）只允许对本地 NTP 服务进行控制查询。

验证功能：

在安全性要求较高的网络中，可以启用 NTP 验证功能。配置 NTP 验证功能分为两部分：配置客户端、配置服务器端。

在配置 NTP 验证功能时，应注意以下原则：

- 客户端和服务端均需要完整配置，NTP 验证才能生效。如果使能了 NTP 验证功能，应同时配置密钥，并声明可信的密钥。
- 服务器端和客户端应配置相同的密钥。

## 1.11 NQA

NQA 功能可以测量网络上运行的各种协议的性能，使运营商能够实时采集到各种网络运行指标。

### 1.11.1 介绍

#### 定义

NQA（Network Quality Analysis），即网络质量分析，是 USG 提供的一个特性，位于链路层之上，覆盖网络层、传输层和应用层，独立于底层硬件。

#### 目的

提供 NQA 主要目的是帮助运营商实时监视网络 QoS，同时在网络发生故障时进行有效的故障诊断和定位。

为了使网络服务质量可见，使用户能够自行检查网络服务质量是否达到要求，运营商需要采取以下措施：

- 在设备上提供能够说明网络服务质量的数据。

由于 IP 网络统计复用、流量突发的特点，NQA 只能以一种统计的方法描述。运营商需要在设备侧提供时延、抖动、丢包率等相关统计参数。

- 在网络中部署探针设备能对网络服务质量进行监控。  
随着网络规模的不断增大，如果使用专用的探针设备(如第三方的探测设备：Brix)，则所需设备数量将不断增加，运营商的成本就会相应的增加。

USG 提供的 NQA 功能，在系统上集成网络质量测试功能，不仅可以实现对网络运行状况的准确测试，输出统计信息。而且，由于不用部署专门的探针设备，还有效的节约了运营商的成本。

NQA 可以监测网络上运行的多种协议的性能，使运营商能够实时采集到各种网络运行指标，例如：HTTP 的总时延、TCP 连接时延、DNS 解析时延、文件传输速率、FTP 连接时延、DNS 解析错误率等。

使用传统的 **Ping** 命令也可以监测网络质量，但是相比 NQA 来说，**Ping** 获取的信息非常有限。下文将从功能，配置和调度三方面来比较 NQA 和 **Ping** 的差异。

表 1-17 NQA 与 Ping 的比较

|      | NQA                                                                                                                                                                                                                                                                                                                                                            | 命令行 Ping                                                                |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|
| 功能差异 | NQA 不仅可以支持 ICMP 探测，还可以探测 TCP、UDP、DHCP、FTP、HTTP、SNMP、DNS 等服务是否启动，以及测试各种服务的响应时间。NQA 还可以通过 Jitter 测试来计算网络抖动情况。                                                                                                                                                                                                                                                    | Ping 功能只能使用 ICMP 协议来测试数据包在本端和指定的目的端之间的往返时间及目的端是否可达。<br>一个用户只能发起一个 Ping。 |
| 配置差异 | NQA 功能可以在客户端执行命令查看 NQA 操作的统计结果。注意事项如下： <ul style="list-style-type: none"><li>● NQA 功能可以通过网管来设置各项操作的参数，并启动测试操作，通过测试结果和历史表中的数据取得统计信息。</li><li>● 在大多数的测试项中，只需要配置 NQA Client 端就可以了。但在进行 TCP、UDP 和 Jitter 类型的测试时，还必须配置 NQA 服务器。</li><li>● NQA 服务器通过监听功能实现对客户端发起的测试的响应。只有在 NQA 服务器上配置了相应的目的地址和端口号，客户端发起的请求才能得到服务器的响应，而且服务器监听服务中指定的 IP 地址、端口号要和客户端配置的参数相一致。</li></ul> | Ping 功能在控制台终端执行 <b>Ping</b> 命令，测试指定的 IP 地址是否可达，是否会实时显示每个包的往返时间或超时。      |

|      | NQA                                                                                                                                                                                                                          | 命令行 Ping  |
|------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 调度方式 | <p>实现对测试例的调度，降低由于测试的并发执行，导致设备负担过重。</p> <p>对单个测试例，支持多种启动时间、结束时间的设置。</p> <ul style="list-style-type: none"><li>● 支持立即启动、延迟启动、定时启动。</li><li>● 支持报文发送完后自动结束、立即结束、延迟结束、定时结束、生命周期结束。</li></ul> <p>多个任务同时启动时，设备主动合理分布启动时间和测试间隔。</p> | 只支持命令行下发。 |

## 1.11.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 1889: RTP: A Transport Protocol for Real-Time Applications
- RFC 2925: Definitions of Managed Objects for Remote Ping, Traceroute, and Lookup Operations
- RFC 2131: Dynamic Host Configuration Protocol
- RFC 1035: DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION
- RFC 414: FILE TRANSFER PROTOCOL (FTP) STATUS AND FURTHER COMMENTS
- RFC 1945: Hypertext Transfer Protocol - HTTP/1.0
- RFC 2616: Hypertext Transfer Protocol - HTTP/1.1
- RFC 792: INTERNET CONTROL MESSAGE PROTOCOL
- IEEE 802.1AG DRAFT6.1: IEEE 802.1AG DRAFT6.1
- RFC 792: INTERNET CONTROL MESSAGE PROTOCOL
- RFC 1157: A Simple Network Management Protocol (SNMP)
- RFC 1905: Protocol Operations for Version 2 of the Simple Network Management Protocol (SNMPv2)
- RFC 3414: User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3)
- RFC793\RFC862: Echo Protocol
- RFC 1393: Traceroute Using an IP Option

## 1.11.3 可获得性

### License 支持

不需要 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                      |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br><br>USG5500 V200R001C00 |

### 1.11.4 原理描述

NQA 把测试两端称为客户端和服务端，NQA 的测试是由客户端发起。在用户端通过命令行配置测试例或由网管端发送相应配置操作后，NQA 把相应的操作类型放入到测试例队列中进行调度。

启动 NQA 测试例，可以选择立即启动、延迟启动、定时启动。在定时器的时间到达后，则根据测试例的测试类型，构造符合相应协议的报文。但配置的测试报文的大小如果无法满足发送本协议报文的最小尺寸，则按照本协议规定的最小报文尺寸来构造报文发送。

测试例启动后，根据返回的报文，可以对相关协议的运行状态提供数据信息。发送报文时的系统的时间作为测试报文的发送时间，给报文打上时间戳，再发送给目的端。目的端接收报文后，返回给源端相应的回应信息，源端在接收到报文时，再一次读取系统的时间，给报文打上时间戳。根据报文的发送和接收时间，计算出报文的往返时间。

#### 说明

对于 JITTER 测试例，不仅源端需要在源端给报文打时间戳，而且目的端在接收到报文和发送报文时，也要读取自己的本地系统时间，再打上时间戳，从而能够计算出抖动时间。

这样用户就可以通过查看测试数据信息了解到网络的运行情况。

## HTTP 测试

如图 1-23 所示，NQA 的 HTTP 测试提供三个阶段的响应速度：

- DNS 解析时间：客户端发送 DNS 报文给名字解析器，将 HTTP 服务器名字解析为 IP 地址，DNS 解析报文返回的时间。
- TCP 建立连接时间：客户端与 HTTP 服务器通过 TCP “三次握手” 建立连接所用的时间。
- 交易时间：客户端发送 Get 或 Post 报文给 HTTP 服务器，响应报文到达 HTTP 服务器的时间。

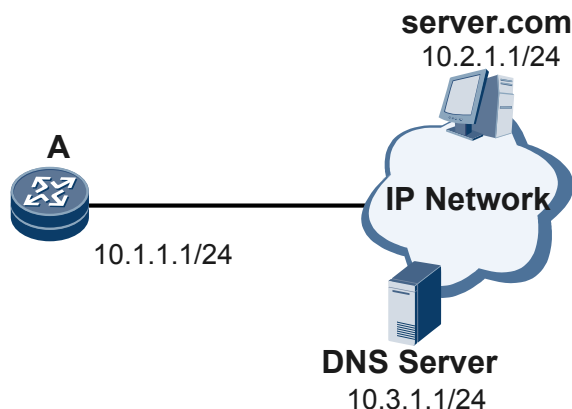
通过 HTTP 测试，从源端接收到的信息中可以计算出：

- 最小 DNS 查询时间、最大 DNS 查询时间及 DNS 查询时间总和。
- 最小 TCP 连接建立时间、最大 TCP 连接建立时间及 TCP 连接建立时间总和。
- 最小 HTTP 交易时间、最大 HTTP 交易时间及 HTTP 交易时间总和。

从而清晰的反映出网络 HTTP 协议的性能状况。



图 1-23 HTTP 应用场景



## DNS 测试

NQA 的 DNS 测试用于检测将给定的 DNS 名称解析成 IP 地址的速度，以 UDP 报文为承载。如图 1-23 所示，DNS 测试的过程如下：

1. 源端向 DNS Server 发送要求解析给定的 DNS 名称的 Query 报文。
2. DNS Server 收到报文后，通过解析构造 Response 报文，然后再把这个数据包发回到源端。
3. 源端收到数据包后通过计算源端接收报文的时间和源端发送报文的时间的差，计算出 DNS 域名解析时间。从而清晰的反映出网络 DNS 协议的性能状况。

## FTP 测试

NQA 的 FTP 测试以 TCP 报文为承载，用于检测从 FTP 服务器下载指定文件或向 FTP 服务器上载指定文件的速度。如图 1-24 所示，FTP 测试提供两个阶段的响应速度：

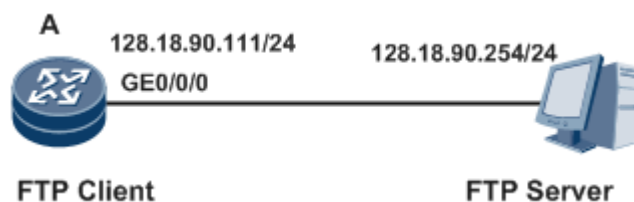
- 控制连接时间：源端与 FTP 服务器通过 TCP “三次握手” 建立控制连接的时间以及通过控制连接交互信令的时间。
- 数据连接时间：源端通过数据连接从 FTP 服务器下载指定文件或向 FTP 服务器上载指定文件的时间。

通过 FTP 测试，从源端接收到的信息中可以计算出：

- 最小控制连接时间、最大控制连接时间及平均控制连接时间。
- 最小数据连接时间、最大数据连接时间及平均数据连接时间。

从而清晰的反映出网络 FTP 协议的性能状况。

图 1-24 FTP 应用场景



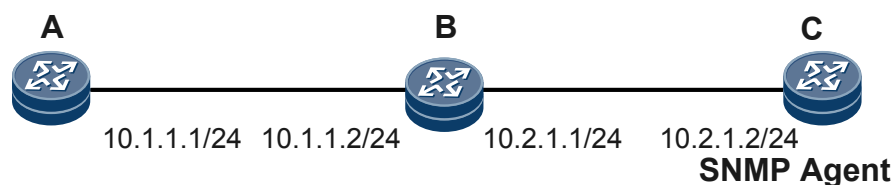
## SNMP 测试

NQA 的 SNMP 测试用于检测主机与 SNMP Agent 之间通信的速度，以 UDP 报文为承载。如图 1-25 所示，SNMP 测试的过程如下：

1. 源端向 SNMP Agent 发送要求获取系统时间的请求报文。
2. SNMP Agent 收到报文，查询系统时间并构造回应报文，然后再把这个数据包发回到源端。

源端收到数据包后通过计算源端接收报文的时间和源端发送报文的时间的差，计算出源端与 SNMP Agent 之间通信的时间。从而清晰的反映出网络 SNMP 协议的性能状况。

图 1-25 SNMP 应用场景



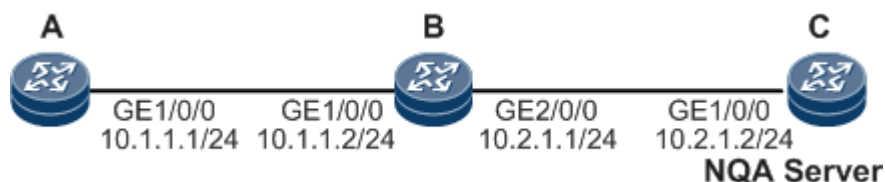
## TCP 测试

NQA 的 TCP 测试用于检测主机与 TCP Server 之间经过三次握手建立 TCP 连接的速度。如图 1-26 所示，TCP 测试的过程如下：

1. 设备 A 向设备 B(TCP server)发送要求建立连接的 TCP SYN 报文。
2. TCP server 收到报文，接受请求并向源端回应 TCP SYN ACK 报文。
3. 源端收到数据包后，向 TCP server 回应 ACK 报文，连接建立。

此后，源端通过接收报文和发送报文的时间的差，计算出与 TCP server 之间三次握手建立 TCP 连接的时间。从而清晰的反映出网络 TCP 协议的性能状况。

图 1-26 TCP 应用场景



## UDP 测试

NQA 的 UDP 测试用于检测主机与 UDP Server 之间通信的速度。如图 1-26 所示，UDP 测试的过程如下：

1. 设备 A 向设备 B(UDP server)发送构造的 UDP 报文。

2. UDP server 收到报文，直接将报文再回送给设备 A。

源端收到数据包后通过计算源端接收报文的时间和源端发送报文的时间的差，计算出源端与 UDP server 之间通信的时间。从而清晰的反映出网络 UDP 协议的性能状况。

## ICMP 测试

NQA 的 ICMP 测试例用于检测 NQA 客户端到目的端的路由是否可达。ICMP 测试提供类似于普通命令行下的 Ping 命令功能，但输出信息更为丰富：

- 默认情况下能够保存最近 5 次的测试结果。
- 结果中能够显示平均时延，丢报率，最后一个报文正确接收的时间等信息。

如图 1-27 所示，ICMP 测试的过程如下：

1. 客户端(A)端向目的端发送构造的 icmp echo request 报文。
2. 目的端(B)收到报文后，直接回应 icmp echo reply 报文给客户端。

客户端收到报文后，通过计算源端接收时间和源端发送时间之差，计算出源端到目的端的通信时间，从而清晰的反应出网络通路畅通情况。

图 1-27 ICMP 应用场景



## Trace 测试

NQA 的 trace 测试例用于检测 NQA 客户端到目的端的转发路径，并沿该路径记录源设备到中间各个设备的时延等信息。TRACE 测试例类似于普通命令行下的 tracert 命令功能，但输出信息更为丰富：每一跳信息中能够显示平均时延，丢包，最后一个包接收时间等信息。如图 1-28 所示，TRACE 测试的过程如下：

1. 客户端 A 端向目的端 B 发送构造的 UDP 报文，报文中的 TTL 为 1。
2. 第一跳设备 C 收到该报文后，判断 TTL 后丢弃该报文，返回一份超时 ICMP 报文。
3. 客户端 A 端收到该 ICMP 报文后，记录下第一跳的 IP 地址，并重新构造一份 UDP 报文，报文中的 TTL 为 2。
4. 报文到达第二跳设备 D 后，判断 TTL 后丢弃该报文，返回一份超时 ICMP 报文。
5. 以此类推，最终报文到达最后一跳，返回一份端口不可达的 ICMP 报文给客户端。

客户端收到每跳返回的 ICMP 报文后，统计并打印出从源端到目的端的转发路径和该路径上各路由设备的有关信息。从而清晰的反映出从源到目的主机的转发路径，以及该路径上各设备的有关的统计信息。

图 1-28 Trace 应用场景



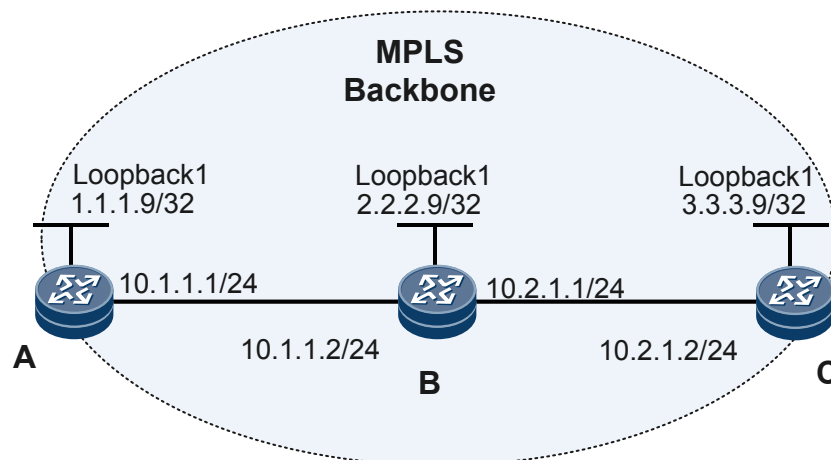
## LSP Ping 测试

NQA 的 LSP Ping 测试例用于检测两种类型（LDP，TE）的 LSP 路径是否可达。如图 1-29 所示，LSP Ping 测试的过程如下：

1. LSP Ping 测试首先构造 UDP 的 MPLS Echo Request 报文，在 IP 头填入 127.0.0.0/8 网段的地址作为 IP 的目的地，根据配置对端 LSR ID 查找相应的 LSP（对于 TE 的 LSP，可以指定从 Tunnel 接口发送，从而找到相应的 CR-LSP），按指定的 LSP 进行 MPLS 域内的转发。
2. 接收端 Egress 监听 3503 端口发送 MPLS Echo Reply 响应报文。

发送端通过接收到的响应报文，统计出测试结果，通过计算源端接收时间和源端发送时间之差，计算出源端到目的端的通信时间，从而清晰的反应出 MPLS 网络通路畅通情况。

图 1-29 LSP PING 应用场景



## 1.12 Ping 和 Tracert

当正常网络通信受到影响时，可以首先使用 ping 与 tracert 命令检测网络是否正常工作。

### 1.12.1 介绍

#### 定义

Ping（Packet Internet Groper）命令是最常见的用于检测网络设备可访问性的调试工具，它使用 ICMP（IPv6 中为 ICMP6）的 echo 信息来决定：

- 远程设备是否可用。
- 与远程主机通信的来回旅程（round-trip）的延迟（delay）。
- 包（packet）的丢失情况。

Tracert 命令用于测试数据包从发送主机到目的地所经过的网关。

tracert 的执行过程如下：

- 首先发送一个 TTL（IPv6 中为 hop limit）为 1 的数据包。
- 到达第一跳时 TTL 超时，第一跳路由器发回一个 ICMP 错误消息，指明此数据包不能被发送。
- 发送主机将 TTL 加 1，重新发送此数据包。
- 第二跳路由器返回 TTL 超时。

以上步骤循环进行，直到到达目的主机，由于目的主机的端口号无效，目的主机发送端口不到达报文，测试结束。这样，发送主机就能够记录每一个 ICMP TTL 超时消息的源地址，根据得到的回应 IP 数据包分析出到达目的地所经历的路径。

#### 目的

当设备出现故障时，可以首先使用 Ping 与 Tracert 命令测试网络连接是否正常工作。

- Ping 命令主要用于检查网络连接及主机是否可达。
- Tracert 主要用于检查网络连接是否可达以及分析网络什么地方发生了故障。

### 1.12.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 791: Internet Protocol
- RFC 1256: ICMP Discovery Messages
- RFC 1393: Traceroute Using an IP Option

### 1.12.3 可获得性

## License 支持

不需要 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 1.12.4 原理描述

### Ping 的工作过程

Ping 会发送 ICMP（IPv6 中为 ICMP6）echo request packet 到某个地址，然后等待应答（reply），当 echo request 到达目标地址以后，在一个有效的时间内（timeout 之前）返回 echo reply packet 给源地址，则说明目的地可达。如在有效时间内，没有收到回应，则在发送端显示超时。

Ping 命令是把 ICMP 报文中的标识符置为发送该 ICMP 报文的进程，这样在对端可以区分出从本端运行的多个 Ping 实例。

Ping 命令每发送一个 ICMP 回显请求报文，顺序号就加 1，顺序号从 1 开始，不同的系统发送回显请求的数量不同，默认情况发送 5 个回显请求报文，也可以通过命令行参数设置发送回显请求报文的个数，如果对端可达，则在对端会相应回应 5 个和请求端同样序号的 ICMP 回应报文。

报文在转发过程中，如果 TTL（IPv6 中为 hop limit）字段的值减为 0，报文到达的路由设备就会向源端发送 ICMP 超时报文，表明远程设备不可达。

### Tracert 的工作过程

源端首先发送 3 个 TTL（IPv6 中为 hop limit）字段的值都为 1 的 UDP 数据报给远程设备，使用任何大于 32768 的随机端口地址作为目标设备的接收报文端口，TTL 为 1 的数据报到达某个路由设备以后随即超时，路由设备响应源设备一个 ICMP 的超时报文，之后源端再发送 3 个 UDP 数据报，这次更改 TTL 值为 2，即经过 2 个路由设备以后，响应源端 ICMP 超时报文，依次类推，直到这些 UDP 报文到达了目标设备。

由于发送的报文中的目的端口，目标设备接收到 ICMP 报文后，由于报文的端口是一个在目标设备没有应用的端口，目标设备就会响应 ICMP port unreachable 信息给源端，表示目标端口不可达，同时说明 Tracert 执行完毕。从而可以从源端显示的结果中，看到到目标设备所经过的路径。

Tracert 发送数据报的 TTL 值最大可以到 30，每一次发送如果在指定的时间内没有回应报文，在发送端就会显示超时，如果发送 30 跳的值后，仍然显示为超时，表明无法达到目标设备，测试失败。没有发送报文的超时时间默认为 5 秒，可以在 0ms ~ 65535ms 之间进行设置。

## 1.13 探测 PMTU

探测 PMTU 的目的就是要找到从源端到目的端路径上使报文不分片的最大 MTU 值。

### 1.13.1 简介

#### MTU 定义

MTU (Maximum Transfer Unit) 指的是在网络中能够传输的最大数据报文的大小，以字节为单位。当设备要转发的报文大于接口 MTU 值时，该报文会被分成多个小于或等于 MTU 的分片报文，在到达目的地址后将分片的报文重组为原始报文。如果报文本身设置了不分片的标志，则经过 MTU 较小的接口时会被丢弃。

#### PMTU 定义

在 IP 网络中，从源地址到目的地址可能会经过具有不同 MTU 值的接口，其中最小的 MTU 值即为该路径的 PMTU。该定义意味着，小于 PMTU 值的报文在该路径上传输时，不会被分片。因为来回路径可能不一致，所以互相通信的两端的 PMTU 在两个方向上不一定相同。

#### 目的

由于分片报文的特殊性，分片报文经常会被利用来做网络攻击，同时分片报文的传输也经常因为某个分片的丢失而导致整个报文重传。所以在实际应用中，应尽量避免分片报文的出现。

通过探测获取指定目的 IPv4 地址的 PMTU 值，然后在网络中使用一个合适的 MTU 值来发送报文，使得报文在整个传输过程中不需要分片，减轻中间设备的工作压力，以便有效地利用网络资源并得到最佳的吞吐量。

 说明

目前设备不支持对 IPv6 地址的 PMTU 探测。

### 1.13.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 1191: Path MTU Discovery

### 1.13.3 可获得性

#### License 支持

不需要 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                      |
|-------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2200/5100 V300R001、<br>USG2200/5100 BSR/HSR V300R001<br>USG5500 V300R001 |

1.13.4 应用场景

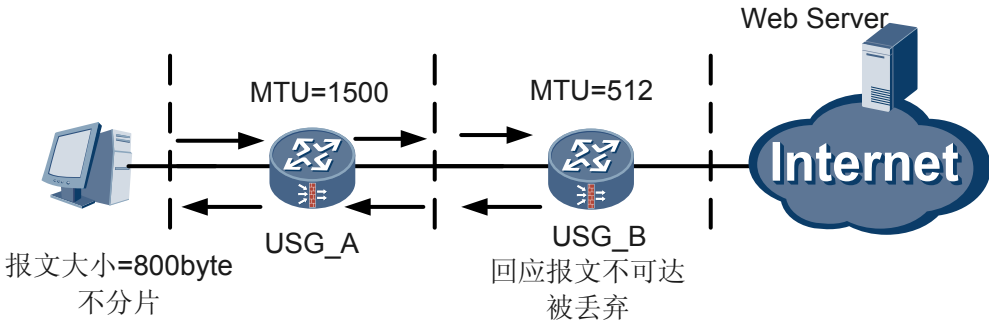
探测 PMTU 应用场景

如图 1-30，在 PC 要访问服务器（Web Server）时，PC 先与服务器建立 TCP 连接，SYN 报文中包含了 MSS（MTU 头），当 SYN 报文到达服务器时，报文与服务器协商，报文 MSS 值会被修改为一个合适的值，最后不会分片；但有时中间设备没有做出正确处理，导致最后协商的 MTU 比实际的大。当 PC 发送大于实际链路 MTU 的不分片报文时，到达 USG\_B 的时候就被丢了，造成访问故障。这时，可以通过调整 USG\_A 的 MTU 值，使 PC 与设备重新协商发送合适大小 MSS 值的报文，解决故障。

例如，PC 先与服务器连接协商，发送大小为 800 字节报文（设置不分片），USG\_A 的 MTU 为 1500，报文能顺利通过；但 USG\_B 的 MTU 为 512，报文大于 MTU 值，但又无法分片，被丢弃。

所以，通过在 USG\_A 上探测 PMTU，可以获取到指定目的地址，使报文转发不分片的所有接口的最小 MTU 值（512）。获取到的 PMTU 值之后，再改变 USG\_A 相应的配置（比如通过修改设备接口 MTU 值为 512），PC 与设备重新协商发送 MTU 较小的报文，就可以使报文在网络中不再需要分片，从而解决问题。

图 1-30 探测 PMTU



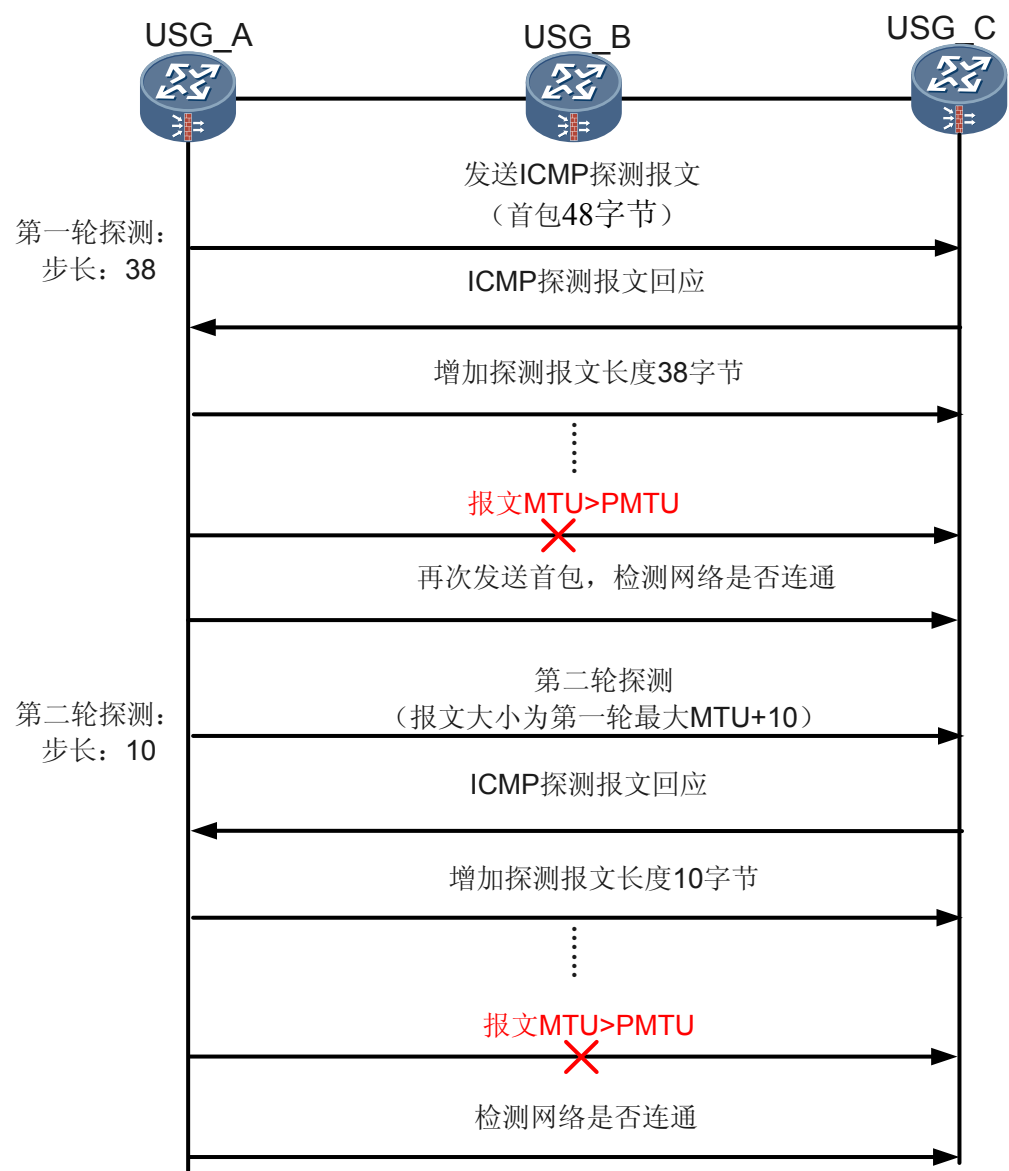
1.13.5 原理描述

PMTU 探测原理

PMTU 探测通过从源向目的端发起 ICMP 探测报文，从初始报文长度开始，逐步增加步长，最终确定路径上的 PMTU 值。具体过程如下：



图 1-31 PMTU 探测原理



1. USG\_A 向 USG\_C 发起第一轮 ICMP 探测（设置不分片标志），报文大小为探测区间的最小值 48 字节（固定不可配置）。
2. 测试连通后，USG\_A 向 USG\_C 继续发送 ICMP 报文，逐步增加第一轮探测递增步长（不可配置，固定为 38 字节），直至连续三个报文发送超时（不再增加步长），说明发送报文的 MTU 已经大于路径上的最小 MTU。
3. 此时 USG\_A 向 USG\_C 再发送一个检测报文（大小为第一轮探测区间的默认最小值 48），探测网络是否连通，如果是，则发起第二轮探测，进行下一步。
4. 减小步长（可配置，第二轮探测递增步长默认为 10），以上一次能到达 USG\_C 的最大报文长度为基础（第一轮探测的 MTU），发起第二轮探测，USG\_A 向 USG\_C 继续发送 ICMP 报文，逐步增加步长(10 字节)，直至连续三个报文发送超时（不再增加步长），说明发送报文的大小已经超过 PMTU。
5. 此时 USG\_A 向 USG\_C 再发送一个检测报文（大小为第一轮探测得到的 PMTU），探测网络是否连通，如果是，则第二轮探测得到的 MTU 为 PMTU。探测步长越小，PMTU 值越精准，但时间、内存等资源消耗越多。

# 2 可靠性

---

## 关于本章

### 2.1 双机热备份

双机热备份需要 VRRP、VGMP、HRP 三个协议的支持，主用设备的配置命令和状态信息能够立即备份到备用设备，保证业务正常进行。

### 2.2 IP-link

IP-link 是指发送 ICMP 或 ARP 请求自动侦测与设备不直接相连的链路的状态，主要应用在双机热备、静态路由和策略路由的环境。

### 2.3 Link-group

Link-group 是指将多个物理接口绑定在一个逻辑组中，保证组内物理接口的状态一致性。

### 2.4 BFD

BFD (Bidirectional Forwarding Detection) 是一种独立的 Hello 协议，可以实现轻负荷的快速故障探测功能。通过与上层协议的联动，可以让上层协议快速发现故障并进行恢复。

## 2.1 双机热备份

双机热备份需要 VRRP、VGMP、HRP 三个协议的支持，主用设备的配置命令和状态信息能够立即备份到备用设备，保证业务正常进行。

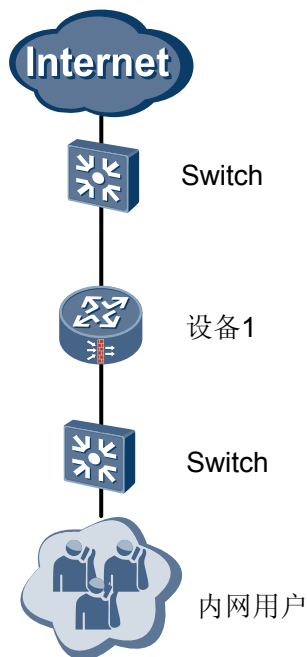
### 2.1.1 介绍

介绍了双机热备的产生背景以及基本的功能。

随着网络的快速普及和应用的日益深入，各种业务在网络上得到了广泛部署，网络带宽也以指数级增长，网络短时间的中断就可能影响大量业务，造成重大损失。高可靠性成为网络建设的关键因素。

如图 2-1 所示，设备 1 部署在网络节点处，内网用户的业务都通过设备 1 进行转发。如果设备 1 出现故障，内外网之间的网络业务将会全部中断。

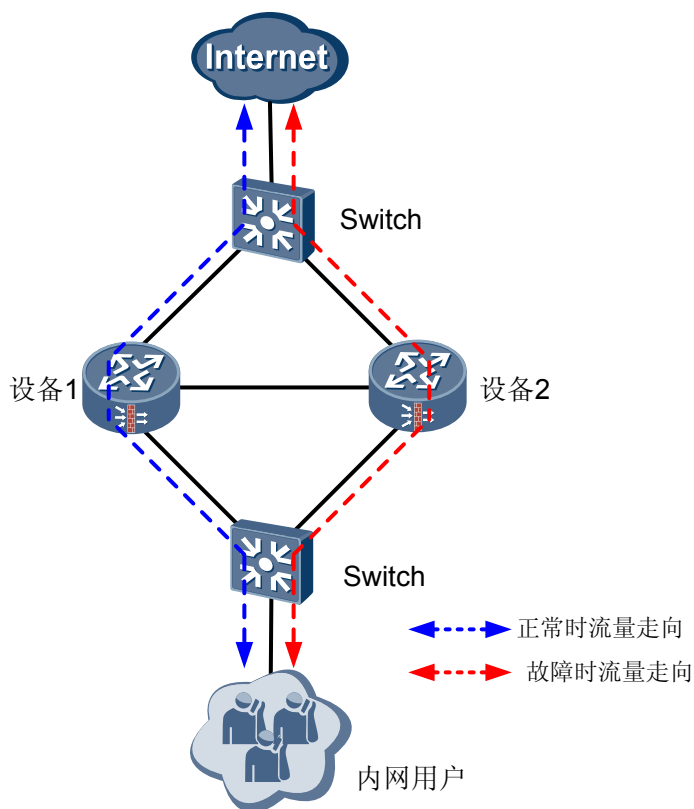
图 2-1 单条链路网络基本组网图



为了规避网关单点故障的风险，可以在网络节点处同时部署两台设备，形成双机热备组网。当其中一台设备出现故障时，业务流量能平滑地切换到备用设备上，保证业务不中断，使内外网用户交互的时候感知不到曾经出现过网络故障。

如图 2-2 所示，网络正常时业务通过设备 1 转发。当设备 1 发生故障时，流量切换到设备 2，业务通过设备 2 转发，从而保证了业务的正常转发，增强了网络的可靠性。

图 2-2 双机热备基本组网图



在双机热备组网中，主备设备通过 VRRP、VGMP、HRP 三个协议共同完成对接口、链路、设备状态的实时监控以及状态信息的实时交互，能够保证业务流量不会因主备倒换而引通信中断。由于该特性能够为用户提供网络级的可靠性保障，故常用于运营商网络，大型垂直行业网网关、中型企业网网关等重要位置。

## 2.1.2 参考标准和协议

介绍双机热备相关的参考标准和协议。

参考标准如下：

- RFC2338: Virtual Router Redundancy Protocol (version number One 1998)
- RFC3768: Virtual Router Redundancy Protocol (version number Two 2004)

相关协议如下：

- VRRP: Virtual Router Redundancy Protocol
- VGMP: VRRP Group Management Protocol
- HRP: Huawei Redundancy Protocol

## 2.1.3 可获得性

### License 支持

本特性无需 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                       |
|-------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100 BSR/<br>HSR V100R005C00<br><br>USG5500 V200R001C00 |

## 2.1.4 应用场景

介绍双机热备的典型应用场景以及典型组网。

### 2.1.4.1 主备备份

当您需要所有业务流量通过一台设备转发，而另一台设备作为备份保证业务不中断时，建议您使用主备备份的双机热备。

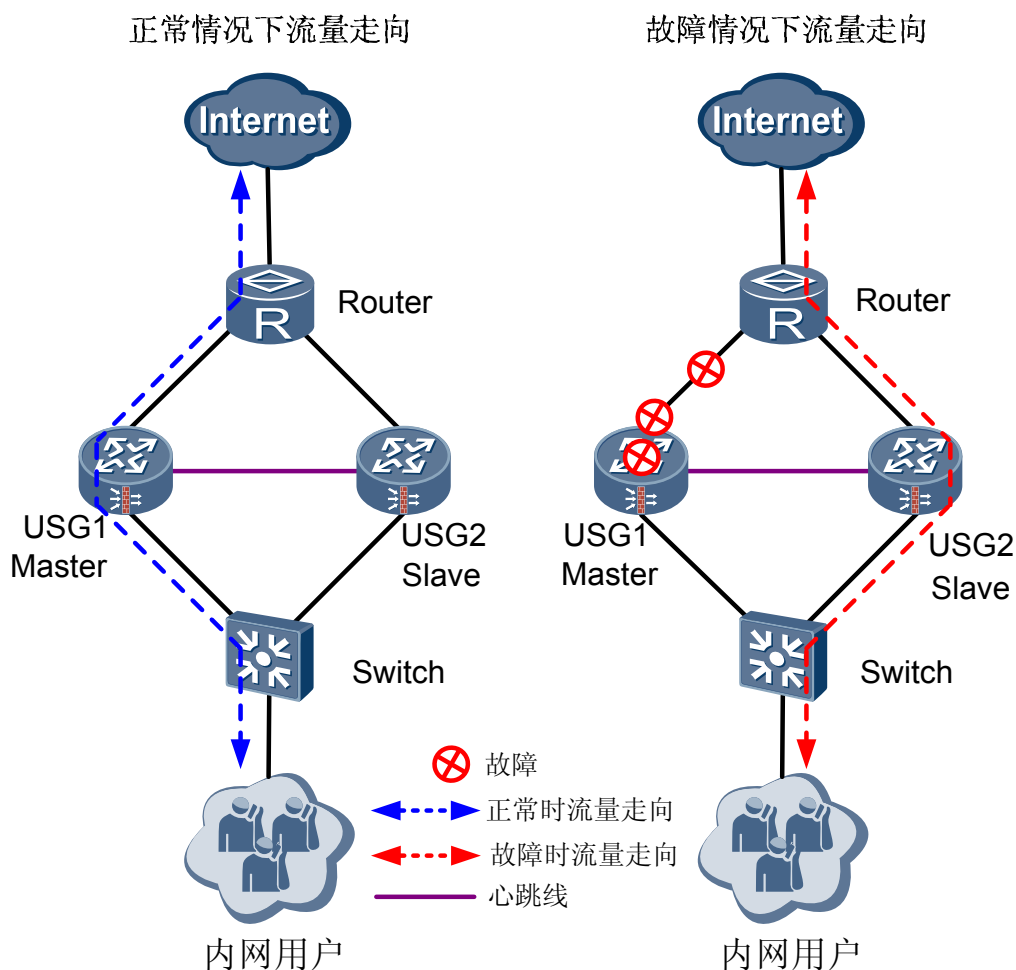
#### 定义

主备备份指正常情况下仅由主用设备处理业务，备用设备空闲；当主用设备接口、链路或整机故障时，备用设备切换为主用设备，接替主用设备处理业务。

主备备份可以有效防止一台设备出现故障导致的网络中断，大大提高了网络的可靠性，常用于重点业务的入口或接入点上，像企业的 Internet 接入点，银行的数据库服务器等。

如图 2-3 所示，在网络入口处同时部署两台 USG 设备，USG1 为主用设备，处理业务流量；USG2 为备用设备，不处理业务流量，只通过心跳线同步 USG1 的配置命令和状态信息。当 USG1 的接口、链路或整机发生故障时，USG2 立即切换为主用设备接替 USG1 处理业务流量，从而保证业务的不中断。

图 2-3 主备备份应用场景



## 典型组网

根据主备设备的上下行业务接口类型，以及上下行设备的类型，主备备份的双机热备场景可以分成以下四种组网。

- 业务接口工作在三层，上下行连接交换机

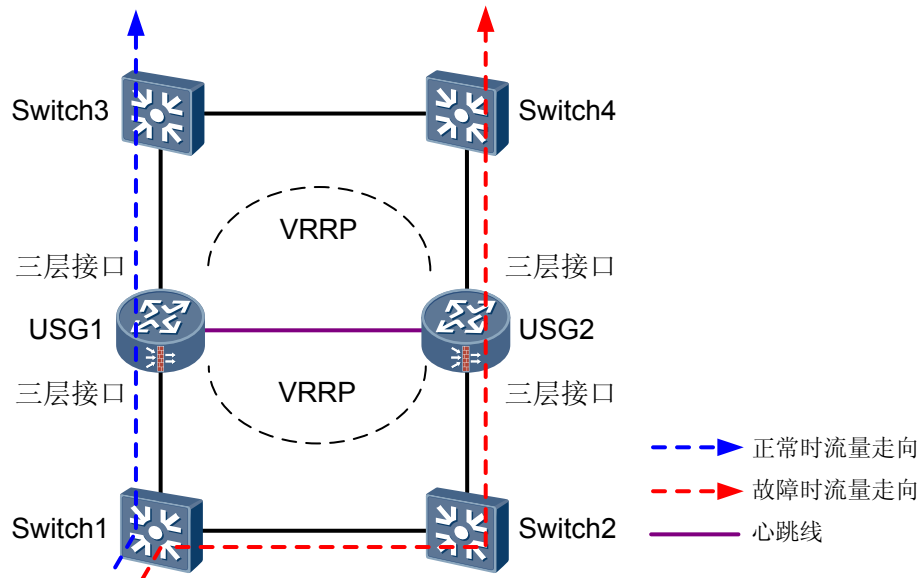
如图 2-4 所示，USG 的上、下行业务接口工作在三层，与二层交换机直连。

此组网是 USG 推荐的典型双机热备组网，是已经应用的很成熟的组网方式。由于上下行设备都是交换机，并且配置的是静态路由，所以当主设备发生故障时，流量能快速切换到备设备。

说明

与图 2-3 相比，图 2-4 中上下行各部署两台设备，可以进一步提升网络可靠性。

图 2-4 业务接口工作在三层，上下行连接交换机的组网

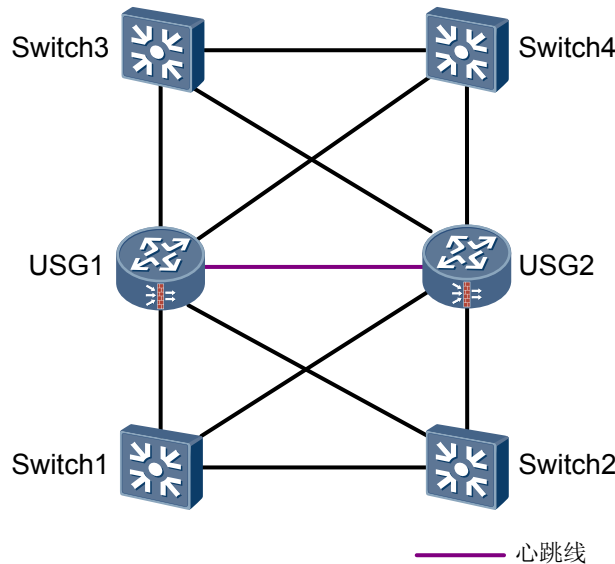


在图 2-4 的基础上，将 USG1 的上行接口与 Switch4 相连，下行接口与 Switch2 相连；将 USG2 的上行接口与 Switch3 相连，下行接口与 Switch1 相连。

这样就组成了双机热备的全冗余组网，如图 2-5 所示。

双机热备的全冗余组网能够进一步提升网络可靠性，避免多条链路故障时业务中断。

图 2-5 双机热备全冗余组网



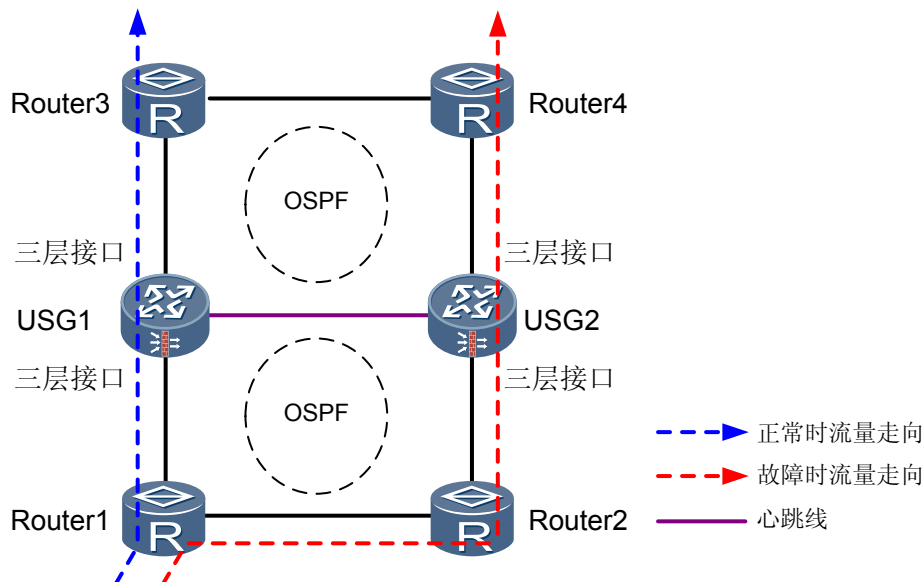
● 业务接口工作在三层，上下行连接路由器

如图 2-6 所示，USG 的上、下行业务接口工作在三层，与路由器直连。USG 与上、下行路由器之间运行 OSPF 协议。

此组网网络拓扑简单，是比较常用的一种组网方式。对已经存在的都是路由器的网络，如果需要增加 USG 作为防火墙可以采用这种组网方案。

此组网可以与业务接口工作在三层，上下行连接交换机的组网结合使用，即组成上行连接路由器，下行连接交换机的组网。

图 2-6 业务接口工作在三层，上下行连接路由器的组网

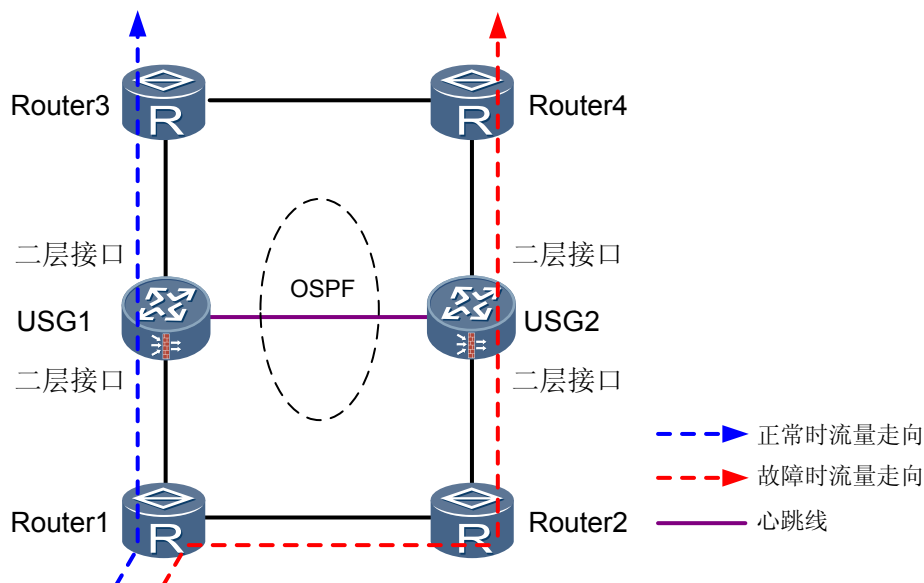


- 业务接口工作在二层，上下行连接路由器

如图 2-7 所示，USG 的上、下行业务接口工作在二层，与路由器直连。上、下行路由器之间运行 OSPF 动态路由协议。每台 USG 的上下行业务接口加入到同一个 VLAN。

在此组网中，USG 透明接入到原有网络，不改变网络拓扑。在接入原有网络时，如果没有额外的 IP 地址分配给 USG，可以采用此种组网。

图 2-7 业务接口工作在二层，上下行连接路由器的组网



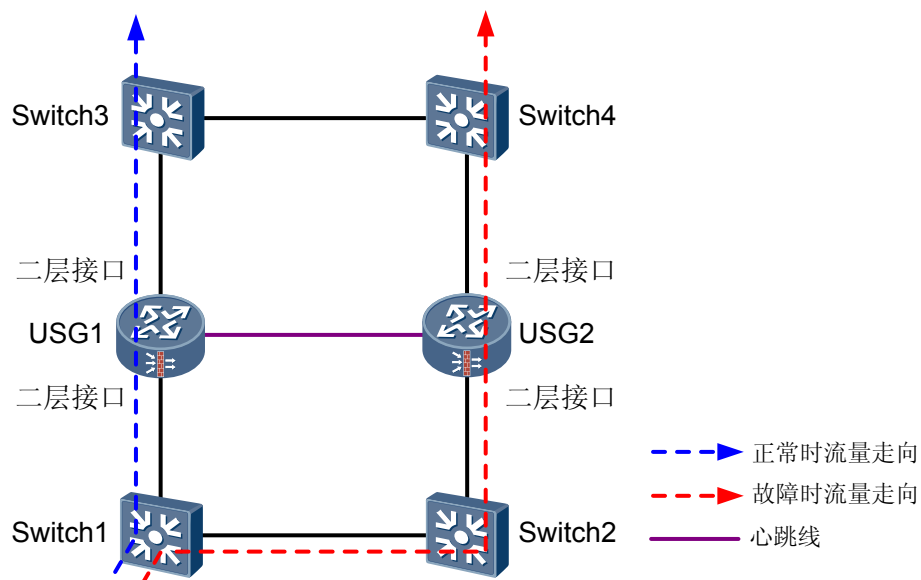


- 业务接口工作在二层，上下行连接交换机

如图 2-8 所示，USG 的上、下行业务接口工作在二层，与二层交换机直连。每台 USG 的上下行业务接口加入到同一个 VLAN。

对已经存在的都是交换机的网络，如果需要增加 USG 作为防火墙且不希望改变网络拓扑时，可以采用这种组网方案。

图 2-8 业务接口工作在二层，上下行连接交换机的组网



### 2.1.4.2 负载分担

当业务流量较大，需要两台设备共同处理并且互为备份时，建议您使用负载分担的双机热备。

#### 定义

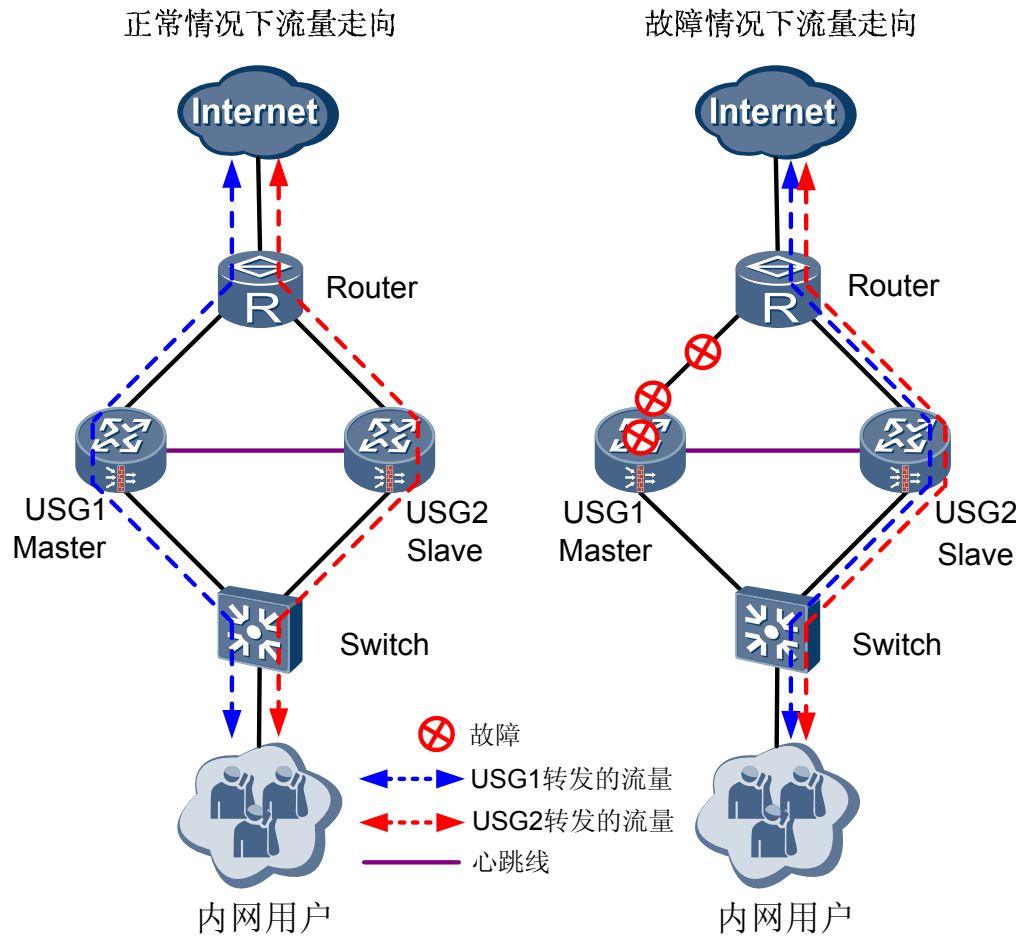
所谓负载分担，也可以称为“互为主备”，即两台设备同时处理业务。当其中一台设备发生故障时，另外一台设备会立即承担其业务，保证原来需要通过这台设备转发的业务不中断。在网络规划时，请保证经过两台设备的总流量不超过单台设备的处理能力。

主备设备共同处理业务流量，可以提高网络的转发效率，降低主用设备发生故障的几率。

如图 2-9 所示，在网络入口处同时部署两台 USG 设备。正常情况下，业务流量分别送到两台 USG 上进行处理。每台 USG 既作为主用设备处理业务流量，也作为备用设备通过心跳线同步另外一台 USG 的配置及状态信息。

如果 USG1 的接口、链路或整机发生故障，USG2 会立即承担全部业务流量的转发。

图 2-9 负载分担应用场景



典型组网

根据主备设备的上下行业务接口类型，以及上下行设备的类型，负载分担的双机热备场景可以分成以下两种组网。

说明

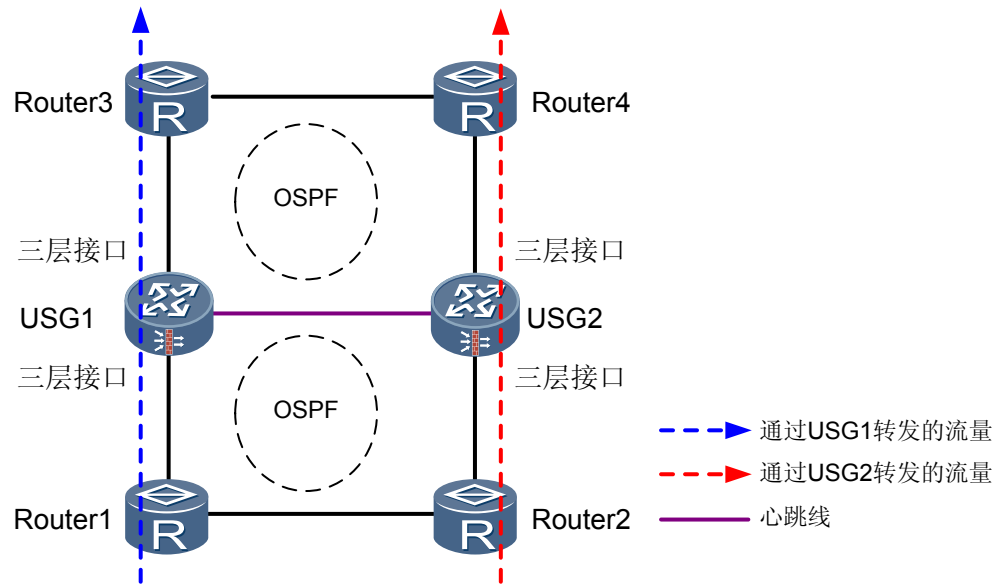
在负载分担场景中，我们推荐您的 USG 上下行连接路由器。

● 业务接口工作在三层，上下行连接路由器

如图 2-10 所示，USG 的上、下行业务接口工作在三层，与路由器直连。USG 与上、下行路由器之间运行 OSPF 协议，使业务流量能通过 USG1 和 USG2 共同转发。

此组网网络拓扑简单，是比较常用的一种组网方式。对已经存在的都是路由器的网络，如果需要增加 USG 作为防火墙且希望两台设备共同处理流量时，可以采用这种组网方案。

图 2-10 业务接口工作在三层，上下行连接路由器的组网

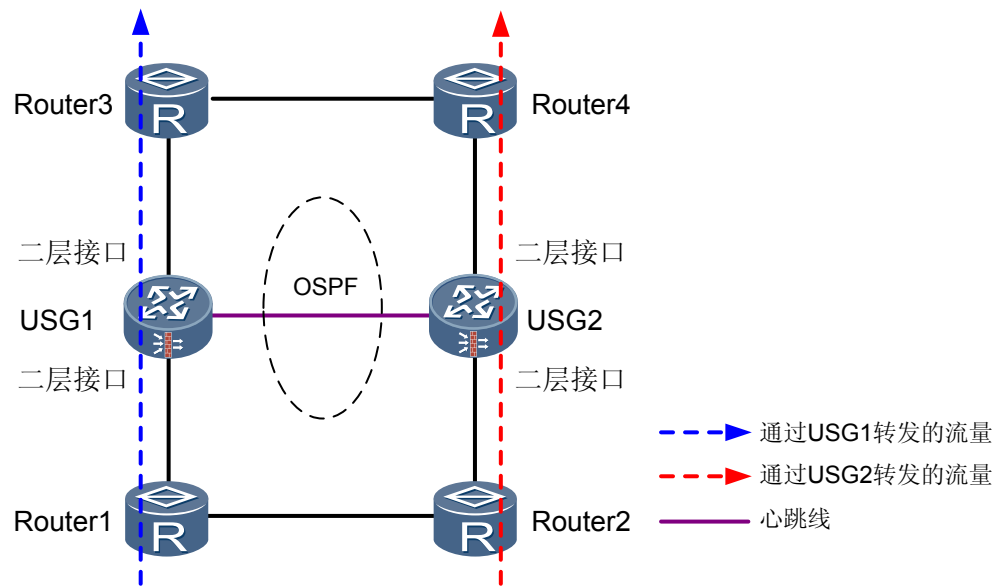


● 业务接口工作在二层，上下行连接路由器

如图 2-11 所示，USG 的上、下行业务接口工作在二层，与路由器直连。上、下行路由器之间运行 OSPF 动态路由协议，使业务流量能通过 USG1 和 USG2 共同转发。每台 USG 的上下行业务接口加入到同一个 VLAN。

在此组网中，USG 透明接入到原有网络，不改变网络拓扑。在接入原有网络时，如果没有额外的 IP 地址分配给 USG，且希望两台 USG 共同处理业务时，可以采用此种组网。

图 2-11 业务接口工作在二层，上下行连接路由器的组网



## 2.1.5 协议原理描述

介绍 VRRP、VGMP、HRP 协议的实现原理。

### 2.1.5.1 协议引入

通过对双机热备正常运行必须解决的四个问题以及业务流程的四个阶段的分析，引入了双机热备的基础协议。

双机热备组网正常运行需要解决以下四个问题：

- 主备设备如何监测到接口、链路或设备故障？
- 主备设备监测到故障后，如何触发主备设备的状态切换？
- 主备设备状态切换后，如何保证业务流量引导到备用设备上？
- 主备设备状态切换后，如何保证业务流量在备用设备上正常转发？

根据以上四个问题，可以将双机热备的业务流程分为四个阶段：

1. 故障监测

故障监测是双机热备的基础。为了保证链路可靠性，通过监控接口、链路和设备状态，尽快监测到各种故障，以触发后续的状态切换和流量引导。

2. 状态切换

状态切换是指双机热备组网中两台设备在故障出现时迅速切换状态，原主用设备切换成备用设备，原备用设备切换成主用设备。在双机热备组网中设备的主备状态由所属 VGMP 管理组的状态决定。

3. 流量引导

流量引导是指在双机热备组网中，当主用设备故障时，需要及时进行主备之间流量倒换，有效地将业务流量从一台设备引导到另一台设备。

4. 信息同步

在双机热备组网中，为保证报文正常转发，需要在主用设备和备用设备之间备份配置命令和状态信息。

如表 2-1 所示，双机热备在不同应用场景和组网下，每个阶段涉及的协议及实现的具体机制有所不同。

表 2-1 双机热备协议列表

| 应用场景 | 典型组网               | 故障监测             | 状态切换 | 流量引导     | 信息同步 |
|------|--------------------|------------------|------|----------|------|
| 主备备份 | 业务接口工作在三层，上下行连接交换机 | VRRP、IP-Link     | VGMP | 静态路由、ARP | HRP  |
|      | 业务接口工作在三层，上下行连接路由器 | VGMP (HRP Track) | VGMP | OSPF     | HRP  |

| 应用场景 | 典型组网               | 故障监测             | 状态切换 | 流量引导                 | 信息同步 |
|------|--------------------|------------------|------|----------------------|------|
|      | 业务接口工作在二层，上下行连接路由器 | VGMP (HRP Track) | VGMP | OSPF、VGMP、Link-group | HRP  |
|      | 业务接口工作在二层，上下行连接交换机 | VGMP (HRP Track) | VGMP | VGMP                 | HRP  |
| 负载分担 | 业务接口工作在三层，上下行连接路由器 | VGMP (HRP Track) | VGMP | OSPF                 | HRP  |
|      | 业务接口工作在二层，上下行连接路由器 | VGMP (HRP Track) | VGMP | OSPF、VGMP、Link-group | HRP  |

从以上列表可以看出，为了保证双机热备在不同场景下的灵活应用，主要引入了以下协议。

- **VRRP (Virtual Router Redundancy Protocol)**

是一种基本的容错协议，它将同一个广播域的一组路由器组织成一个虚拟路由器，称之为一个备份组。备份组中的所有路由器一起，对外共同提供一个虚拟 IP 地址。在双机热备特性中，VRRP 是实现接口、链路故障监测的一种手段。当需要监测远端链路（非直连链路）时，可通过将 IP-Link 功能与 VRRP 备份组绑定实现。

- **VGMP (VRRP Group Management Protocol)**

VGMP 是 VRRP 备份组管理协议，它可以将一组 VRRP 备份组或一组接口组成一个 VGMP 管理组。VGMP 管理组通过 VRRP 备份组监控接口、链路的状态，或者直接监控接口、链路的状态。利用 VGMP 管理组统一管理上下行接口或链路的状态，可以保证其成员的状态一致。VGMP 管理组的状态决定了设备的主备状态，主用设备的 VGMP 管理组状态为 Master，备用设备的 VGMP 管理组状态为 Slave。

- **HRP (Huawei Redundancy Protocol)**

信息备份协议，它可以在主用和备用设备之间实时备份关键配置命令和会话表状态信息，保证主用设备出现故障时，业务能够平滑切换到备用设备上。

- **Link-group**

Link-group 保证组内接口的状态一致性。如果将设备的上下行业务接口加入到一个 Link-group 中，当其中一个接口故障状态变为 Down 时，另一接口状态也立即变为 Down。

- **OSPF (Open Shortest Path First)**

OSPF 是基于链路状态的内部网关协议。在双机热备特性中，OSPF 是路由器环境下保证流量引导的重要手段。

- **静态路由**

静态路由可以在网络的路由选择中实现非常精确的控制。在双机热备特性中，静态路由主要用于业务接口工作在三层，上下行连接交换机的组网，保证流量被引导到正确的出接口上。

- **ARP (Address Resolution Protocol)**

ARP 实现三层 IP 地址与二层 MAC 地址之间的动态映射。在双机热备特性中，ARP 的主要作用是在主备状态切换时发送免费 ARP 报文，以刷新上下游交换机的 MAC 表。

## 2.1.5.2 VRRP

介绍 VRRP 协议的基本原理以及双机热备中的 VRRP 基本功能。

### 基本原理

VRRP 是一种基本的容错协议，它将同一个广播域的一组路由器组织成一个虚拟路由器，称之为一个备份组。在同一个备份组中，只有一台路由器处于工作状态，称为主用设备（状态为 Master），其余都处于备份状态，称为备用设备（状态为 Slave）。备份组中的所有路由器一起对外提供一个虚拟 IP 地址，只有主用设备能转发以虚拟 IP 地址作为下一跳的报文。

主用设备和备用设备的优先级不同。正常情况下，主用设备的优先级高于备用设备。当主用设备发生故障时，优先级会降低，如果低于备用设备的优先级，则备用设备会变成主用设备，接替主用设备的工作继续转发业务报文。

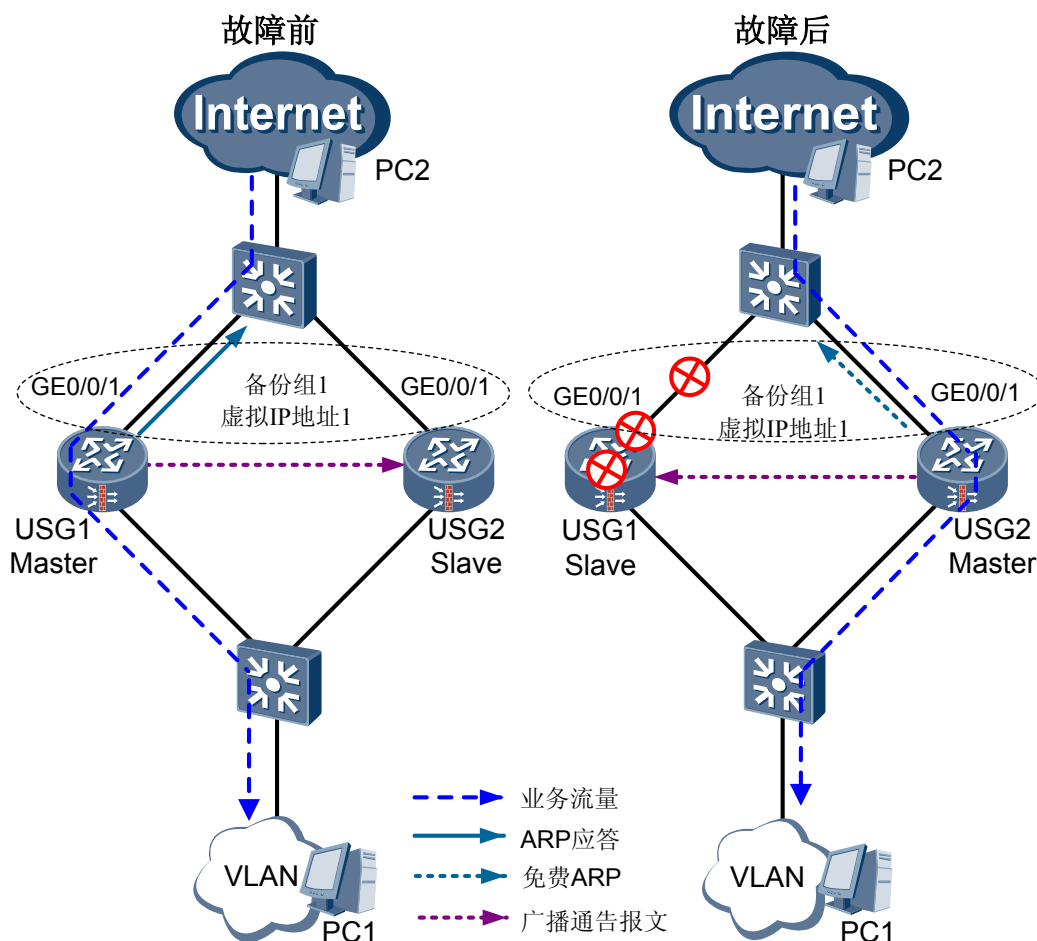
如图 2-12 所示，每个 VRRP 备份组都有一个虚拟 IP 地址，内网用户将网关地址设置为 VRRP 备份组的虚拟 IP 地址。

当链路正常时，虚拟 IP 地址在 USG1 的 GE0/0/1 接口（VRRP 备份组状态为 Master）上生效。PC2 请求网关地址（即 VRRP 备份组的虚拟 IP 地址）的地址解析（ARP）时，只有 USG1 会应答 ARP 报文，反馈自己的 MAC 地址。因此业务流量被交换机引导到 USG1 上进行转发。

当 USG1 出现故障时，VRRP 状态发生切换，流量从 USG1 倒换到 USG2。具体过程如下：

1. USG1 的接口、链路或整机故障导致接口 GE0/0/1 状态变为 Down，触发接口 VRRP 备份组优先级降低。
2. USG2 的接口上 VRRP 备份组优先级高于主用设备，VRRP 备份组状态切换到 Master，并发送免费 ARP。
3. 交换机收到免费 ARP 后刷新 MAC 表，将出接口指向 USG2。这样 PC 发出的业务流量被引导到 USG2 上进行转发。

图 2-12 链路正常时流量转发



## 双机热备中的 VRRP 基本功能

双机热备特性中，VRRP 的主要作用是通过 VRRP 备份组来实现接口及链路的故障监测。

- 接口故障监测

每个 VRRP 备份组可以监视所有与此 VRRP 备份组绑定的接口的状态。当接口 Down 时，会导致相应 VRRP 备份组优先级降低。

- 链路故障监测

- 当接口直连链路 Down 时会触发接口 Down，故也会导致相应 VRRP 备份组优先级降低。
- 当需要监测远端链路时，可以采用 VRRP 备份组绑定 IP-Link 功能的方式。当 IP-Link 监测到远端链路 Down 时，也会导致相应 VRRP 备份组优先级降低。

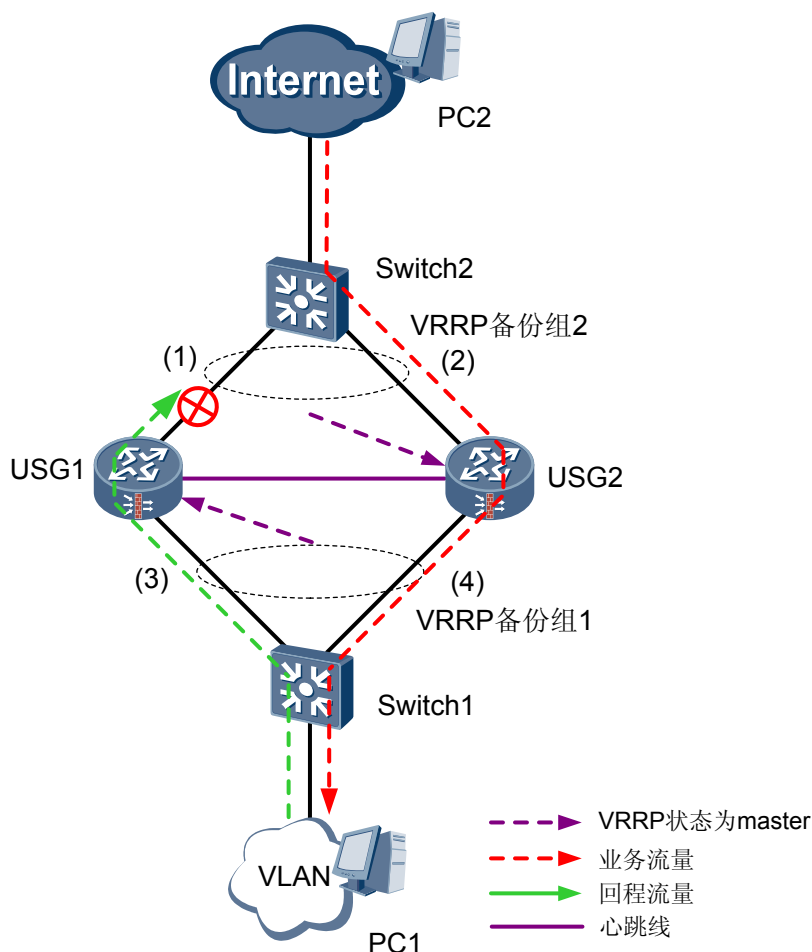
### 2.1.5.3 VGMP

介绍 VGMP 的引入和基本原理，以及双机热备中设备状态切换的过程。

## VGMP 的引入

当 USG 的多个接口同时加入到不同的 VRRP 备份组中，多个 VRRP 备份组将出现状态无法同步的现象。如图 2-13 所示，USG 的上下行业务接口分别加入到 VRRP 备份组 1、2 中。正常情况下，业务流量通过主用设备 USG1 进行转发。当 PC2 与 USG1 之间的链路（1）出现故障时，VRRP 备份组 2 的状态切换，交换机 Switch2 更新 MAC 表，流量倒换，PC2 到 PC1 的后续报文会通过路径（2）发送到 USG2；但由于 USG1 与 PC1 之间的链路（3）是正常的，无法感知上行链路的故障，因此 VRRP 备份组 1 的状态没有切换，从 PC1 发出的返回流量仍然会通过路径（3）到达 USG1，从而导致返回流量丢失。

图 2-13 多个 VRRP 备份组主备状态切换



为了解决上述问题，USG 引入 VGMP 实现 VRRP 备份组状态管理，以保证多个 VRRP 备份组状态的一致性。

## VGMP 基本原理

VGMP 是 VRRP 备份组管理协议，它可以将一组 VRRP 备份组或一组接口组成一个 VGMP 管理组。

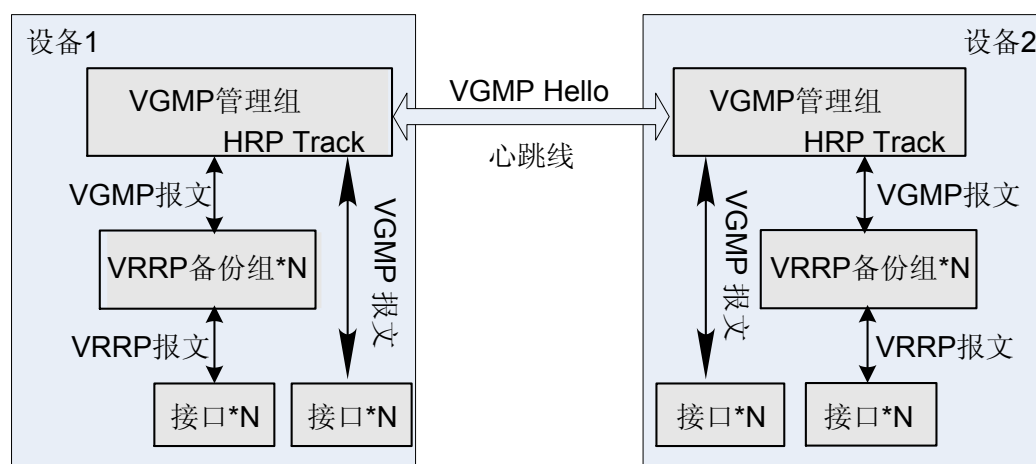


所谓 VGMP 管理组，是指 USG 上满足某种备份需求的若干 VRRP 备份组组成的逻辑集合。VGMP 管理组统一管理加入其中的 VRRP 备份组，同步调整组中所有 VRRP 备份组的状态，从而克服了独立工作的 VRRP 备份组信息无法互通的缺陷。

VGMP 管理组通过 VRRP 备份组监控接口状态。当接口或链路故障时，VRRP 备份组更新自己的状态，并向 VGMP 管理组汇报。此外，VGMP 管理组也可以直接监控上下行业务接口的状态，称为 HRP Track。

如图 2-14 所示，VGMP 管理组之间通过 VGMP Hello 报文通信，VGMP 管理组与 VRRP 备份组或接口之间使用 VGMP 报文通信，VRRP 备份组和接口之间使用 VRRP 报文通信。

图 2-14 VGMP 功能示意图



## VGMP 管理组基本功能

VGMP 管理组是双机热备的核心，主要功能有接口集中监控和设备状态统一管理。

- 接口集中监控
  - 将系统中的所有 VRRP 备份组都加入到 VGMP 管理组中，由 VGMP 管理组统一监控组内所有的 VRRP 备份组状态。这样 VGMP 管理组就可以通过 VRRP 备份组统一监控系统的接口状态。
  - 将系统中的业务接口加入到 VGMP 管理组中，由 VGMP 管理组统一监控组内所有接口状态（HRP Track）。

以上两种接口监控方式可以一起使用。

- 设备状态统一管理

USG 是通过 VGMP 管理组来进行设备状态管理的。VGMP 管理组的主备状态决定了双机热备组网中设备的主备状态，决定了 VGMP 管理组内的成员接口和 VRRP 备份组的状态。当一台设备的 VGMP 管理组的状态为 Master 时，组内所有成员的接口状态统一为 Master，该设备成为主用设备。另外一台设备的 VGMP 管理组状态为 Slave，则该设备为备用设备。

每台设备的 VGMP 管理组初始状态由用户指定（Master 或 Slave），Master 的优先级为 65001，Slave 的优先级为 65000。当 VGMP 管理组通过 VRRP 备份组或直接监测到接口 Down 时，会重新计算 VGMP 管理组优先级，计算公式如下：

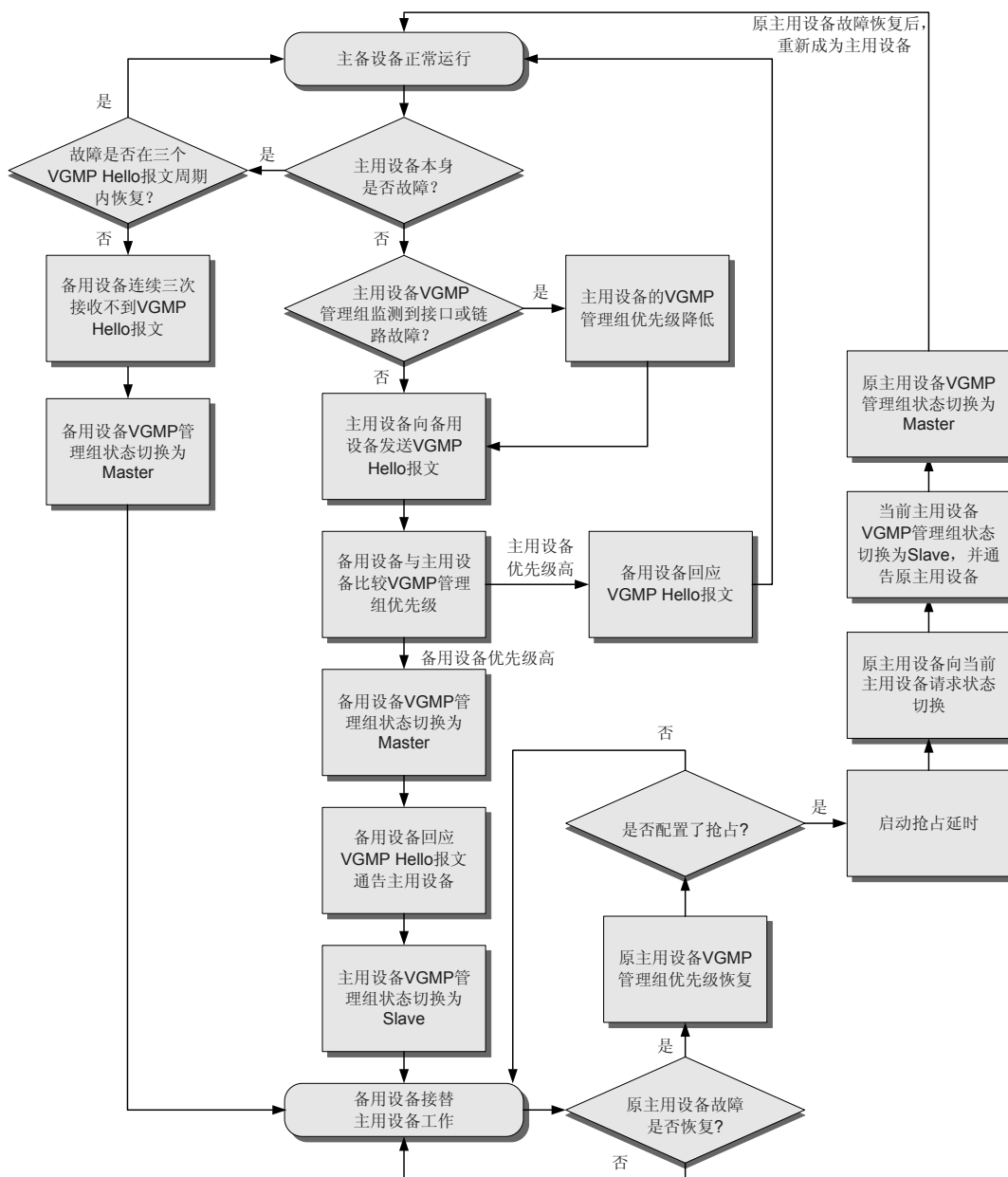
VGMP 管理组优先级=VGMP 管理组初始优先级-N\*2（N=状态变为 Down 的被监测接口的个数）

即每个被监测的接口 Down 时，VGMP 管理组优先级降低 2。

## 设备状态切换过程

双机热备组网中的两台设备通过 VGMP Hello 报文交互各自的状态信息并协商确定各自的主备状态。主用设备的 VGMP 管理组状态为 Master，备用设备的 VGMP 管理组状态为 Slave。具体协商过程如图 2-15 所示。

图 2-15 设备状态切换过程业务流程图



- 主用设备正常时

1. 主用设备会定时通过心跳线向备用设备发送 VGMP Hello 报文，通知对端自身的运行状态（包括 VGMP 管理组优先级、VRRP 成员状态等信息）。
  2. 备用设备收到 VGMP Hello 报文后，会回应一个 VGMP Hello 报文进行确认，该报文中也携带本端的 VGMP 管理组优先级、VRRP 成员状态等信息。
- 接口或链路发生故障时
    1. 当主用设备的链路或接口发生故障时，VGMP 管理组监测到故障，并降低自身的 VGMP 管理组优先级。
    2. 主用设备向备用设备发送 VGMP Hello 报文，请求将本端 VGMP 管理组状态从 Master 切换到 Slave。
    3. 备用设备将本端的 VGMP 管理组优先级与 VGMP Hello 报文中的对端优先级进行比较。比较后发现对端的 VGMP 管理组优先级低于本端，于是备用设备的 VGMP 管理组状态切换为 Master，组内成员状态也统一切换为 Master。同时备用设备会回应 VGMP Hello 报文通告主用设备。
    4. 主用设备收到 VGMP Hello 报文后，会将 VGMP 管理组状态切换为 Slave，组内成员状态也统一切换为 Slave。此时主备设备之间完成状态切换过程。
  - 主用设备故障时  
备用设备如果在三个周期内没有收到主用设备发送的 VGMP Hello 报文，则认为主用设备出现故障。这时，备用设备会将本端的 VGMP 管理组状态切换为 Master，从而成为主用设备。
  - 原主用设备故障恢复（抢占管理）  
如果配置了抢占模式，原主用设备故障恢复后会重新抢占为主用设备。抢占过程如下：  
 说明  
如果关闭抢占模式，即使原主用设备的优先级高于当前主用设备，也不会抢占成主用设备。
    1. 原主用设备故障恢复后，VGMP 管理组的优先级升高。  
 说明  
原主用设备当前为备用设备，VGMP 管理组的状态为 Slave。
    2. 原主用设备在收到当前主用设备的 VGMP Hello 报文后发现本端的 VGMP 管理组优先级高于对端，因此进入到抢占状态并启动抢占延时计时。
    3. 抢占延时结束后，原主用设备向当前主用设备发送 VGMP Hello 报文，请求将 VGMP 管理组的状态从 Slave 切换到 Master。
    4. 当前主用设备收到请求后，将 VGMP Hello 报文中的优先级与本端进行比较。如果本端的 VGMP 管理组优先级较低，则将 VGMP 管理组状态切换为 Slave，并向对端发送 VGMP Hello 报文确认。
    5. 原主用设备收到确认报文后，将 VGMP 管理组状态切换为 Master，重新成为主用设备。

#### 2.1.5.4 HRP

介绍 HRP 的信息备份内容和备份方式。

USG 通过执行命令行来实现用户所需的各种功能。如果备用设备切换为主用设备前，配置命令没有备份到备用设备，则备用设备无法实现主用设备的相关功能，从而导致业务中断。

USG 是状态防火墙，对于每一个动态生成的会话连接，都有一个会话表项与之对应。主用设备处理业务过程中创建了很多动态会话表项；而备用设备没有流量经过，因此没有

创建会话表。如果备用设备切换为主用设备前，会话表项没有备份到备用设备，则会导致先前经过主用设备的业务流量因为无法匹配会话表而中断，从而影响业务正常进行。

为了实现主用设备出现故障时备用设备能平滑地接替工作，需要在主用和备用设备之间备份关键配置命令和会话表状态信息。为此，USG 引入 HRP 备份功能。启动 HRP 备份功能后，关键配置命令和会话表状态信息会实时同步备份到备用设备。

在双机热备组网中，指定心跳线作为专门的备份通道，用于备份配置命令和状态信息。

心跳线是指主用设备和备用设备之间的直连网线，主备设备通过心跳线交互报文了解对方状态，以及实现配置命令和状态信息的备份。心跳线两端的设备上的接口称为心跳口。

备份方式

USG 支持三种不同类型的备份方式，如表 2-2 所示。

表 2-2 USG 支持的 HRP 备份方式

| 备份方式 | 配置命令                                                                                                                                                                                                            | 状态信息                                                                                        | 说明                                             |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------|------------------------------------------------|
| 自动备份 | 在主用设备、备用设备都正常工作的情况下，如果启用自动备份功能，则在主用设备上每执行一条可以备份的命令时，此配置命令就会被同步执行到备用设备。如果在主用设备上执行的命令不在备份范围内，则该命令仅在主用设备上执行。<br><br>对于备份范围内的配置命令，只能在主用设备上配置，备用设备不能配置。<br><br>对于备份范围外的配置命令，备用设备上可以配置。<br><br>当备用设备故障时，配置命令自动备份无法进行。 | 在主备运行过程中，当主用设备上产生了需要备份的状态信息时，主用设备自动将状态信息同步到备用设备进行备份，备用设备更新状态。<br><br>当备用设备故障时，状态信息自动备份无法进行。 | 此功能在 USG 上缺省为开启，即默认自动备份配置命令和状态信息。应用于各种双机热备组网中。 |

| 备份方式   | 配置命令                                                                                                             | 状态信息                                                                                                                                                                                                                                 | 说明                                                                                                  |
|--------|------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| 手工批量备份 | <p>当主用设备、备用设备都正常工作时，在主用设备上执行手工批量备份命令，则主用设备将备份范围内的配置命令批量发送到备用设备，备用设备同步执行这些配置命令。</p> <p>当备用设备故障时，配置命令无法手工批量备份。</p> | <p>当主用设备、备用设备都正常工作时，在主用设备上执行手工批量备份命令，则主用设备将备份范围内的状态信息批量发送到备用设备，备用设备更新状态。</p> <p>当备用设备故障时，状态信息无法手工批量备份。</p>                                                                                                                           | 主要用于主备设备之间配置不同步，需要批量备份的场景。                                                                          |
| 快速备份   | 不备份                                                                                                              | <p>当 USG 工作于负载分担组网环境下，如果报文的来回路径不一致，即来回两个方向的报文分别从不同的 USG 经过，如果主用设备的状态信息没有及时备份到备用设备，则备用设备会将到达的报文丢弃。</p> <p>为防止上述现象发生，可以通过快速备份会话，将主用设备相应的状态信息表项快速备份到备用设备，使返回报文在备用设备上能够查找到相应的状态信息表项，从而保证内外部用户的业务不中断。</p> <p>对于来回路径不一致的组网，需要开启快速备份功能。</p> | <p>为了保证状态信息的及时同步，快速备份功能只是备份状态信息，不备份配置命令。配置命令的备份由自动备份功能实现。</p> <p>通常用于负载分担组网环境下，主备设备需要同时处理业务的场景。</p> |

## 备份信息

USG 备份的内容包括以下特性的配置命令和状态信息。

- 配置命令
  - AAA 命令
  - ACL 命令

- 转发策略命令
- 攻击防范命令
- 地址绑定命令
- 虚拟防火墙命令
- GTP 配置命令
- 黑名单命令
  - 包括黑名单的启动命令、手工添加黑名单表项和删除黑名单表项命令。
- 日志命令
- NAT 命令
- 统计命令
- 区域命令
  - 包括创建安全区域，设置区域优先级，接口加入安全区域以及域间配置命令。
- ASPF 命令
- 清除会话表项的命令
- 清除配置的命令
- IPS 命令
- URL 过滤命令
- AV 命令（不包括 **av scan-level**）
- Web 过滤命令
- 邮件过滤命令
- FTP 过滤命令
- IPSec VPN 命令
- SSL VPN 命令
- 升级服务命令
  - 不包括 **update local { ips | av } file filename**、**update restore factory-default { ips | av }**、**update install factory-default**。
- 状态信息
  - USG 生成的会话表（UTM 的会话不备份）
  - 虚拟防火墙生成的会话表
  - SeverMap 表
  - IP 监控表
  - 分片缓存表
  - GTP 表
  - 黑名单
  - 白名单
  - PAT 方式端口映射表
  - NO-PAT 方式地址映射表

## 2.1.6 典型组网分析

介绍双机热备的四种典型组网，并对他们的实现原理进行了分析。

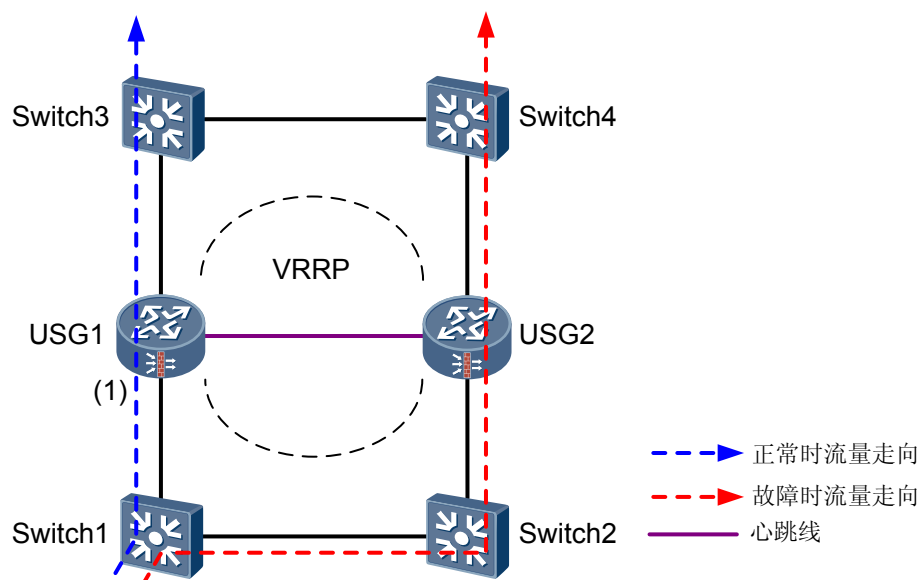
### 2.1.6.1 组网一：业务接口工作在三层，上下行连接交换机

介绍 USG 业务接口工作在三层，上下行直连设备为交换机的组网。

如图 2-16 所示，USG 的上、下行业务接口工作在三层，分别与二层交换机直连。

此组网一般采用主备备份方式的双机热备。

图 2-16 USG 业务接口工作在三层，上下行连接交换机组网图



## 网络部署分析

- 由于 USG 的业务接口工作在三层，能够配置 IP 地址，且上下行连接的交换机能够透传 VRRP 报文，因此采用在 USG 的上、下行业务接口配置 VRRP 备份组的方式实现故障监测。
- 配置 VRRP 备份组时会将 VRRP 备份组加入 VGMP 管理组，并设置 VGMP 管理组的状态。当链路发生故障时，VGMP 管理组会保证 VRRP 备份组状态的统一切换，实现设备状态的切换。
- 由于 USG 的上下行设备为交换机，不能配置路由，因此需要配置 VRRP 备份组的虚拟 IP 地址，并在交换机的下行或上行设备上配置静态路由，将下一跳指向 VRRP 备份组的虚拟 IP 地址，实现流量引导。
- HRP 协议可以在主用和备用设备之间实时备份关键配置命令和会话表状态信息，实现信息同步。因此需要在 USG 上指定心跳口，并启用 HRP 功能。

## 网络运行分析

如图 2-16 所示，链路正常的时候，流量走向为图中蓝色的路径；链路故障的时候，流量走向为图中红色的路径。

- 正常情况下，USG1 为主用设备（VGMP 管理组状态为 Master，优先级为 65001），USG2 为备用设备（VGMP 管理组状态为 Slave，优先级为 65000）。由于 USG1 的上下行接口的 VRRP 备份组都加入了 USG1 的 VGMP 管理组，所以他们的 VRRP 备份组状态也为 Master。

由于虚拟 IP 地址在 VRRP 备份组状态为 Master 的接口上生效（即上下行交换机的 MAC 转发表中记录虚拟 IP 地址与 USG1 的接口 MAC 地址的对映表项），因此以 VRRP 虚拟 IP 地址为下一跳的流量都被引导到 USG1。

- 当 USG1 和 Switch1 之间的链路发生故障时，（1）接口的 VRRP 备份组状态变成初始状态，导致 USG1 的 VGMP 管理组优先级降低 2。这时 USG1 的 VGMP 管理组优先级低于 USG2，两台 USG 主备状态发生倒换，USG2 变成主用设备（VGMP 管理组状态为 Master，上下行接口的 VRRP 备份组状态变为 Master），USG1 变成备用设备（VGMP 管理组状态为 Slave，上下行接口的 VRRP 备份组状态变为 Slave）。

同时，USG2 会对外发布 VRRP 备份组虚拟 IP 地址的免费 ARP，更新上、下行交换机的 MAC 转发表，这样以 VRRP 虚拟 IP 地址为下一跳的流量都能被引导到状态为主 USG2。

- 当 USG1 整机发生故障或者重启时，USG2 连续三次收不到 USG1 发送的 VGMP Hello 报文，则认为 USG1 已经发生故障，立即切换为主用设备；同时，USG2 对外发布 VRRP 备份组虚拟 IP 地址的免费 ARP，更新上、下行交换机的 MAC 转发表，使上、下行设备的报文流量都被引导到 USG2。

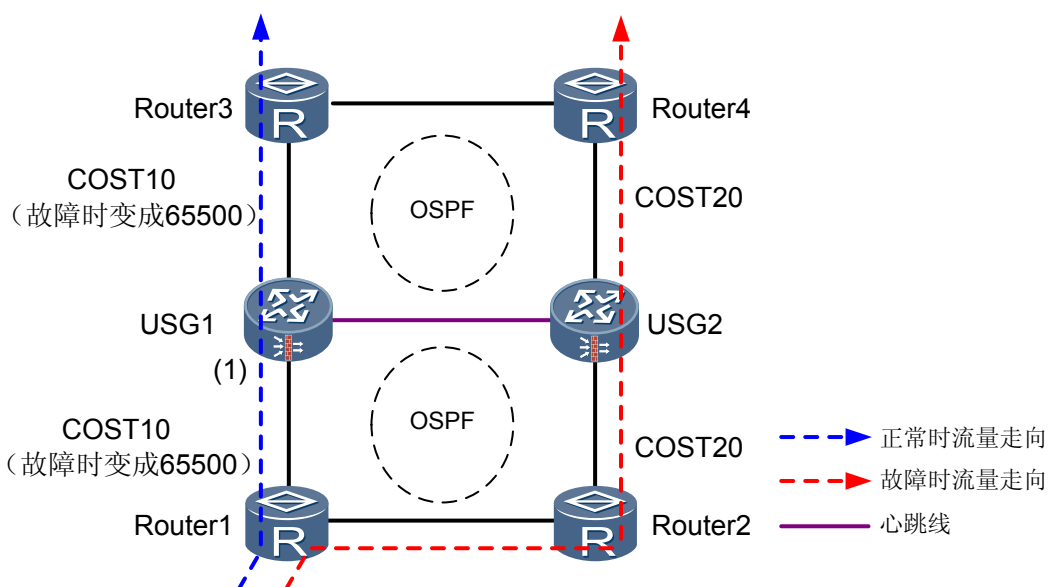
### 2.1.6.2 组网二：业务接口工作在三层，上下行连接路由器

介绍 USG 业务接口工作在三层，上下行直连设备是路由器的组网。

如图 2-17 所示，USG 上、下行业务接口工作在三层，与路由器直连。USG 与上、下行路由器之间运行 OSPF 协议。这是比较常用的一种组网方式。

此组网既可以用于主备备份方式，又可以用于负载分担方式的双机热备。

图 2-17 USG 业务接口工作在三层，上下行连接路由器组网图





## 网络部署分析

- 由于 USG 的上下行设备为路由器，而路由器不能透传 VRRP 报文，所以 USG 的业务接口上不能配置 VRRP 备份组。业务接口工作三层，能够配置 IP 地址，因此采用 VGMP 管理组直接监控接口状态的方式实现故障监测。
- VGMP 管理组监控的接口故障时，VGMP 管理组优先级降低。当主用设备的 VGMP 管理组优先级低于备用设备的 VGMP 管理组优先级，会发生主备设备的状态切换。
- 由于 USG 的上下行设备为路由器，且业务接口工作三层，因此可以在 USG 与上下行设备之间运行 OSPF 协议。如果需要通过主备备份，则需要分别在两台 USG 的直连路由器上配置不同的 COST 值；如果需要通过负载分担，则需要分别在两台 USG 的直连路由器上配置相同的 COST 值。
- HRP 协议可以在主用和备用设备之间实时备份关键配置命令和会话表状态信息，实现信息同步。因此需要在 USG 上指定心跳口，并启用 HRP 功能。如果是负载分担组网，为了保证流量来回路径的一致，还需要启用快速会话备份功能。

## 网络运行分析

下面以主备备份方式的组网为例分析网络的运行情况。

如图 2-17 所示，链路正常的时候，流量走向为图中蓝色的路径；链路故障的时候，流量走向为图中红色的路径。

- 正常情况下，USG1 为主用设备（VGMP 管理组状态为 Master，优先级为 65001），USG2 为备用设备（VGMP 管理组状态为 Slave，优先级为 65000）。由于 Router1 连接 USG1 的接口 COST 值小于 Router2 连接 USG2 的接口 COST 值，且 Router3 连接 USG1 的接口 COST 值小于 Router4 连接 USG2 的接口 COST 值，所以流量都被引导到 USG1。
- 当 USG1 和 Router1 之间的链路发生故障时，（1）接口状态变为 Down，导致 USG1 的 VGMP 管理组优先级降低 2。这时 USG1 的 VGMP 管理组优先级低于 USG2，两台 USG 主备状态发生倒换，USG2 变成主用设备，USG1 变成备用设备。同时，USG1 和 USG2 都会更新自身的路由并对外发布，此时的 USG2 为主用设备，对外直接发布自身的路由，而 USG1 变成了备用设备，对外发布的路由 COST 值为 65500。这样，当整个网络的路由收敛时，业务流量都被引导到 USG2。
- 当 USG1 发生故障或者重启时，USG2 连续 3 次收不到 USG1 发送的 VGMP Hello 报文，则认为 USG1 已经发生故障，立即切换为主用设备。同时链路的路由 COST 值发生变化，上、下行路由器重新计算路由，路由收敛后所有的业务流量都引导到 USG2，从而保证业务的正常转发。

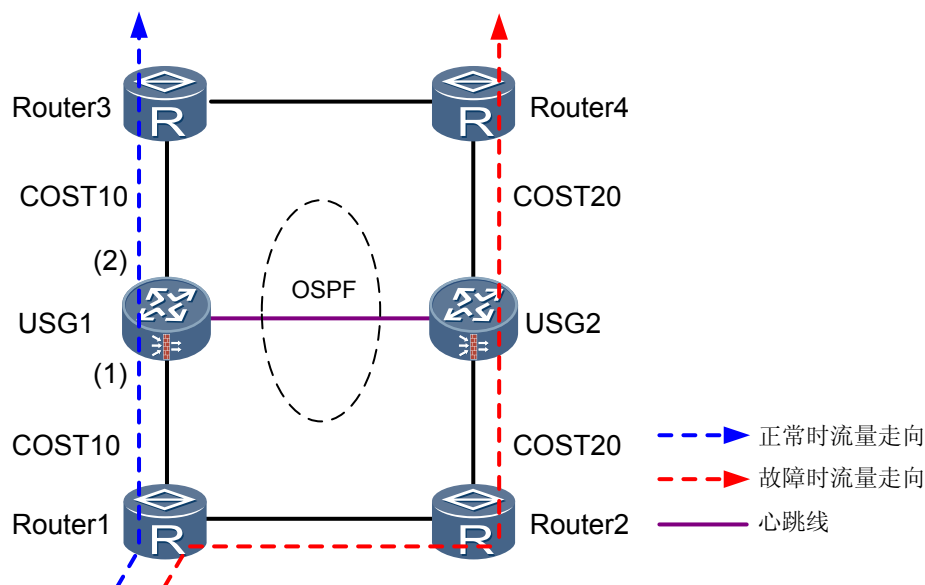
### 2.1.6.3 组网三：业务接口工作在二层，上下行连接路由器

介绍 USG 业务接口工作在二层，上下行直连设备为路由器的组网。

如图 2-18 所示，USG 上、下行业务接口工作在二层，与路由器直连。上、下行路由器之间运行 OSPF 动态路由协议。每台 USG 的上下行业务接口加入到同一个 VLAN。

此组网既可以用于主备备份方式，又可以用于负载分担方式的双机热备。

图 2-18 USG 业务接口工作在二层，上下行连接路由器组网图



## 网络部署分析

- USG 的业务接口工作在二层，不能配置 IP 地址，所以上、下行业务接口不能配置 VRRP 备份组。在这种情况下，需要将 USG 的上下行接口加入到同一个 VLAN，并将 VGMP 管理组与 VLAN 绑定，由 VGMP 管理组直接监控接口状态以实现故障监测。当 VGMP 管理组状态为 Master 时，VLAN 能转发流量；VGMP 管理组状态为 Slave 时，VLAN 被禁用，不能转发流量。
- 当 VLAN 内的业务接口故障时，VGMP 管理组优先级降低。当主用设备的 VGMP 管理组优先级低于备用设备的 VGMP 管理组优先级，会发生主备设备的状态切换。
- 由于 USG 的业务接口工作在二层，且上下行连接路由器，所以 USG 自身不运行 OSPF 路由协议，只是依靠上下行路由器运行的 OSPF 来实现流量引导。如果需要实现主备备份，则需要在两台 USG 的直连路由器上配置不同的 COST 值；如果需要实现负载分担，则需要在两台 USG 的直连路由器上配置相同的 COST 值。

为保证链路故障时路由的快速收敛，建议将同台 USG 设备的上下行接口加入到同一 Link-group 组。

- HRP 协议可以在主用和备用设备之间实时备份关键配置命令和会话表状态信息，实现信息同步。因此需要在 USG 上指定心跳口，并启用 HRP 功能。如果是负载分担组网，为了保证流量来回路径的一致，还需要启用快速会话备份功能。

## 网络运行分析

下面以主备备份方式的组网为例分析网络的运行情况。

如图 2-18 所示，链路正常的时候，流量走向为图中蓝色的路径；链路故障的时候，流量走向为图中红色的路径。

- 正常情况下，USG1 为主用设备（VGMP 管理组状态为 Master，优先级为 65001），USG2 为备用设备（VGMP 管理组状态为 Slave，优先级为 65000）。由于 Router1 连接 USG1 的接口 COST 值小于 Router2 连接 USG2 的接口 COST 值，且 Router3

连接 USG1 的接口 COST 值小于 Router4 连接 USG2 的接口 COST 值，所以流量都被引导到 USG1。

- 当 USG1 和 Router1 之间的链路发生故障时，（1）接口状态变为 Down，导致 USG1 的 VGMP 管理组优先级降低 2。这时 USG1 的 VGMP 管理组优先级低于 USG2，两台 USG 主备状态发生倒换，USG2 变成主用设备，USG1 变成备用设备。USG1 切换成备用设备（VGMP 管理组状态切换到 Slave）时，VGMP 管理组监控的 VLAN 内所有接口都会 Down 然后 Up 一次。这会导致链路的 COST 值发生变化，上、下行路由器重新计算路由，路由收敛后所有的业务流量都引导到 USG2。如果在 USG 上配置了 Link-group，那么在（1）接口状态变为 Down 时，同一 Link-group 组内的（2）接口也会变为 Down，OSPF 路由将迅速收敛（不必等待 VGMP 管理组的状态切换），将业务流量引导到 USG2。
- 当 USG1 发生故障或者重启时，USG2 连续 3 次收不到 USG1 发送的 VGMP Hello 报文，则认为 USG1 已经发生故障，立即切换为主用设备。同时链路的 COST 值发生变化，上、下行路由器重新计算路由，路由收敛后所有的业务流量都引导到 USG2，从而保证业务的正常转发。

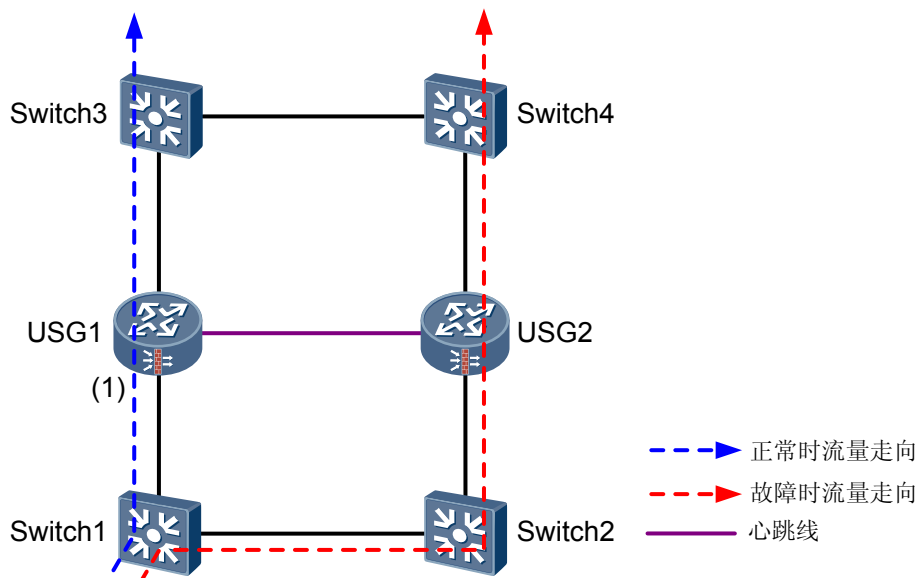
#### 2.1.6.4 组网四：业务接口工作在二层，上下行连接交换机

介绍 USG 业务接口工作在二层，上下行直连设备为交换机的组网。

如图 2-19 所示，USG 的上、下行业务接口工作在二层，分别与二层交换机直连。每台 USG 的上下行业务接口加入到同一个 VLAN。

此组网一般采用主备备份方式的双机热备。

图 2-19 USG 业务接口工作在二层，上下行连接交换机组网图



#### 网络部署分析

- USG 的业务接口工作在二层，不能配置 IP 地址，所以上、下行业务接口不能配置 VRRP 备份组。在这种情况下，需要将 USG 的上下行接口加入到同一个 VLAN，并将 VGMP 管理组与 VLAN 绑定，由 VGMP 管理组直接监控接口状态以实现故障

监测。当 VGMP 管理组状态为 Master 时，VLAN 能转发流量；VGMP 管理组状态为 Slave 时，VLAN 被禁用，不能转发流量。

- 当 VLAN 内的业务接口故障时，VGMP 管理组优先级降低。当主用设备的 VGMP 管理组优先级低于备用设备的 VGMP 管理组优先级，会发生主备设备的状态切换。
- 由于 USG 的业务接口工作在二层，且上下行连接交换机，网络不能运行路由协议，所以只能通过 VGMP 管理组控制 VLAN 来实现流量引导。
- HRP 协议可以在主用和备用设备之间实时备份关键配置命令和会话表状态信息，实现信息同步。因此需要在 USG 上指定心跳口，并启用 HRP 功能。

## 网络运行分析

如图 2-19 所示，链路正常的时候，流量走向为图中蓝色的路径；链路故障的时候，流量走向为图中红色的路径。

- 正常情况下，USG1 为主用设备，VGMP 管理组状态为 Master，VLAN 能转发流量；USG2 为备用设备，VGMP 管理组状态为 Slave，VLAN 不能转发流量。由于 USG1 的 VLAN 能转发流量，USG2 的 VLAN 不能转发流量，所以流量都被引导到 USG1。
- 当 USG1 和 Switch1 之间的链路发生故障时，（1）接口状态变为 Down，导致 USG1 的 VGMP 管理组优先级降低 2。这时 USG1 的 VGMP 管理组优先级低于 USG2，两台 USG 主备状态发生倒换，USG2 变成主用设备，USG1 变成备用设备。USG1 切换成备用设备（VGMP 管理组状态切换到 Slave）时，VGMP 管理组监控的 VLAN 内所有接口都会 Down 然后 Up 一次。这会导致上下行交换机会更新自身 MAC 转发表，重新学习 MAC 地址，将流量引导到 USG2 上。接口（1）重新 Up 后，由于 VGMP 管理组状态为 Slave，所以 VLAN 被禁用，USG1 将无法转换该 VLAN 的流量。  
USG2 切换成主用设备（VGMP 管理组状态切换到 Master）时，VGMP 管理组监控的 VLAN 被启用，能转发流量。
- 当 USG1 发生故障或者重启时，USG2 连续 3 次收不到 USG1 发送的 VGMP Hello 报文，则认为 USG1 已经发生故障，立即切换为主用设备。USG 切换成主用设备，VGMP 管理组监控的 VLAN 被启用，能转发流量，所以流量被引导到 USG2。

## 2.2 IP-link

IP-link 是指发送 ICMP 或 ARP 请求自动侦测与设备不直接相连的链路的状态，主要应用在双机热备、静态路由和策略路由的环境。

### 2.2.1 介绍

#### 定义

IP-Link，即链路可达性检查，通过 USG 定时地向指定的目的 IP 进行 ICMP 回显请求或 ARP 请求，并等待应答。在设定的时限内（默认 3 秒）未收到回应报文时，则认为当前链路发生故障，并进行与链路相关的后续操作。当原来认为故障的链路，在之后设定的时限内，有连续的 3 个回应报文收到，则认为链路故障已经消除，此时进行链路恢复的后续操作。

## 目的

IP-Link 主要用于业务链路正常与否的自动侦测，可以检测到与 USG 不直接相连的链路状态，保证业务持续通畅。

## 2.2.2 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                   |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100 BSR/<br>HSR V100R005C00<br>USG5500 V200R001C00 |

## 2.2.3 原理描述

### 实现原理

通过 USG 定时地向指定的目的 IP 进行 ICMP 回显请求或 ARP 请求，并等待应答。在设定的时限内（默认 3 秒）未收到回应报文时，则认为当前链路发生故障，并进行与链路相关的后续操作。当原来认为故障的链路，在之后设定的时限内，有连续的 3 个回应报文收到，则认为链路故障已经恢复，此时进行链路恢复的后续操作。

### 组网应用

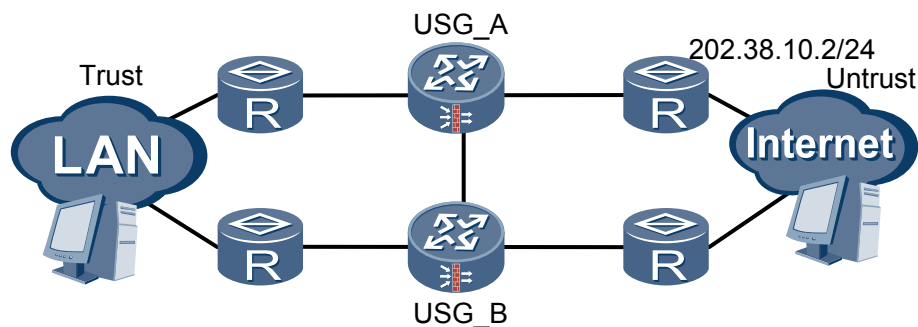
IP-Link 主要应用在以下环境：

- 双机热备环境

当 USG 工作于双机热备份环境时，IP-Link 自动侦测后发现链路故障影响主备业务，通过配置 VGMP 管理组绑定 IP-Link，USG 会对 VGMP 管理组的优先级进行相关调整，触发主备 USG 切换，从而保证业务能够持续流通。

配置 VGMP 管理组监控 IP-Link 后，可以检测到与 USG 不直接相连的接口或链路状态。如图 2-20 所示，当位于 Untrust 区域路由器接口（IP 地址为 202.38.10.2/24）发生故障，启用 IP-link 链路可达性检查功能后，系统将会触发主备切换，保证业务正常进行。

图 2-20 IP-Link 链路可达性检查组网图

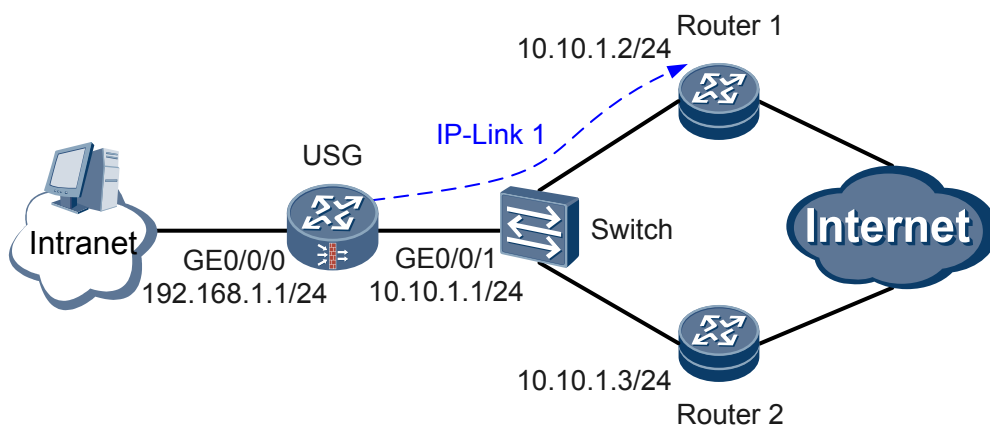


- 静态路由环境

当 IP-Link 自动侦测发现链路故障时，USG 会对自身的静态路由进行相应的调整，保证每次用到的链路是最高优先级和链路可达的，以保持业务的持续流通。

如图 2-21 所示，内部网络用户访问 Internet 的时候有两条静态路由可供选择，其中一条静态路由绑定了 IP-Link 进行链路可达性检查，当该链路不通的时候流量切换至另一条路由，以保证业务的畅通。

图 2-21 静态路由环境下的 IP-Link 链路可达性检查



- 策略路由环境

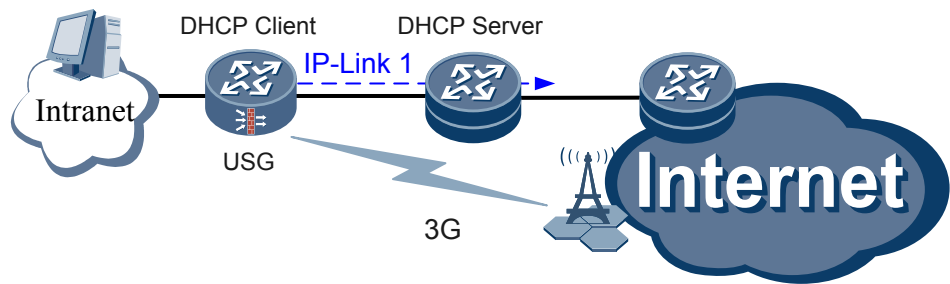
当 IP-Link 自动侦测发现链路故障时，系统可以触发链路绑定的策略路由失效，这样 USG 在查找路由时将查找备份的路由，以保持业务的持续流通。

- DHCP Client 环境

如图 2-22 所示，USG 作为出口网关，采用双上行链路。主链路是 USG 作为 DHCP Client 获取 IP 地址，备链路是 3G 链路。当 IP-Link 检查 DHCP Server 后的链路时，USG 会获取网关地址作为下一跳进行链路检测。如果发现 DHCP Server 后的链路故障，则将 USG 切换到备份链路。



图 2-22 DHCP Client 环境下的 IP-Link 链路可达性检查



2.3 Link-group

Link-group 是指将多个物理接口绑定在一个逻辑组中，保证组内物理接口的状态一致性。

2.3.1 介绍

定义

Link-group 功能是将多个接口的状态相互绑定，组成一个逻辑组。当组内任一接口出现故障，系统将组内其它接口状态设置为 Down。当组内所有接口恢复正常后，整个组内的接口状态才重新被设置为 Up。

目的

Link-group 管理组保证了组内物理接口的状态一致性，同时在链路故障时能够加快路由收敛速度。

2.3.2 可获得性

License 支持

本特性无需 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                   |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100 BSR/<br>HSR V100R005C00<br>USG5500 V200R001C00 |

## 2.3.3 原理描述

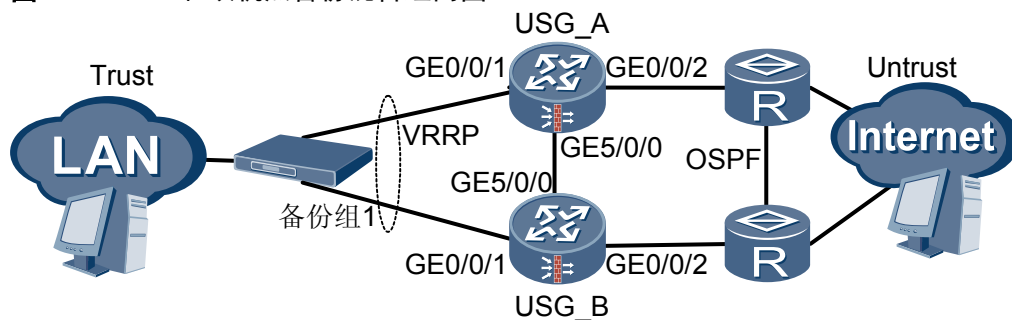
### 组网应用

如图 2-23 所示，在 OSPF（Open Shortest Path First）和双机热备份混合组网中，USG\_A 和 USG\_B 组成双机热备份组网环境，与路由器之间运行 OSPF 协议，与交换机之间运行 VRRP 协议。并且 USG 使用 VGMP 管理组监测与路由器相连的接口 GigabitEthernet 0/0/2。

当 USG\_A 与交换机相连的接口 GigabitEthernet 0/0/1 出现故障时，触发主备倒换，主用设备由 USG\_A 转换成 USG\_B。然而由于 USG 和路由器组成的 OSPF 区域无法感知到接口 GigabitEthernet 0/0/1 故障，路由不会发生切换。这样发送到路由器的报文仍将发送到 USG\_A，造成业务中断。

将接口 GigabitEthernet 0/0/1 与接口 GigabitEthernet 0/0/2 加入同一 Link-group 管理组后。如果接口 GigabitEthernet 0/0/1 因故障而状态变为 Down，将会触发接口 GigabitEthernet 0/0/2 状态变为 Down，从而保证两接口的状态一致，实现了业务正常运行。

图 2-23 OSPF 和双机热备份混合组网图



## 2.4 BFD

BFD（Bidirectional Forwarding Detection）是一种独立的 Hello 协议，可以实现轻负荷的快速故障探测功能。通过与上层协议的联动，可以让上层协议快速发现故障并进行恢复。

### 2.4.1 介绍

#### 定义

双向转发检测 BFD（Bidirectional Forwarding Detection）用于快速检测系统之间的通信故障，并在出现故障时通知上层协议。

#### 目的

为了减小设备故障对业务的影响，提高网络的可用性，网络设备需要能够尽快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故障，但并不是所有介质都能提供硬件检测。
- 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 就是为解决现有检测机制的不足而产生的。

BFD 的目标如下：

- 对相邻转发引擎之间的链路提供轻负荷、快速故障检测。这些故障包括接口，数据链路，甚至有可能是转发引擎本身。
- 提供一个单一的机制，能够用来对任何媒介、任何协议层进行实时地检测，并且检测的时间与开销范围比较宽。

## 2.4.2 参考标准和协议

与 BFD 相关的参考标准与协议如下：

- RFC 5880: Bidirectional Forwarding Detection (BFD)
- RFC 5881: Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)
- RFC 5882: Generic Application of Bidirectional Forwarding Detection (BFD)
- RFC 5883: Bidirectional Forwarding Detection (BFD) for Multihop Paths

## 2.4.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                      |
|-------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2200/5100 V300R001、<br>USG2200/5100 BSR/HSR V300R001<br>USG5500 V300R001 |

## 2.4.4 应用场景

### 2.4.4.1 BFD 与 OSPF 联动

## 应用环境

网络上的链路故障或拓扑变化都会导致路由器重新进行路由计算，要提高网络的可用性，缩短路由协议的收敛时间非常重要。由于链路故障无法完全避免，因此，加快故障感知速度并将故障快速通告给路由协议是一种可行的方案。

BFD 与 OSPF 联动就是将 BFD 和 OSPF 协议关联起来，通过 BFD 对链路故障的快速感知进而通知 OSPF 协议，从而加快 OSPF 协议对于网络拓扑变化的响应。

OSPF 协议在有、无 BFD 协议下收敛速度的数据如表 2-3 所示。

表 2-3 OSPF 协议收敛速度的数据

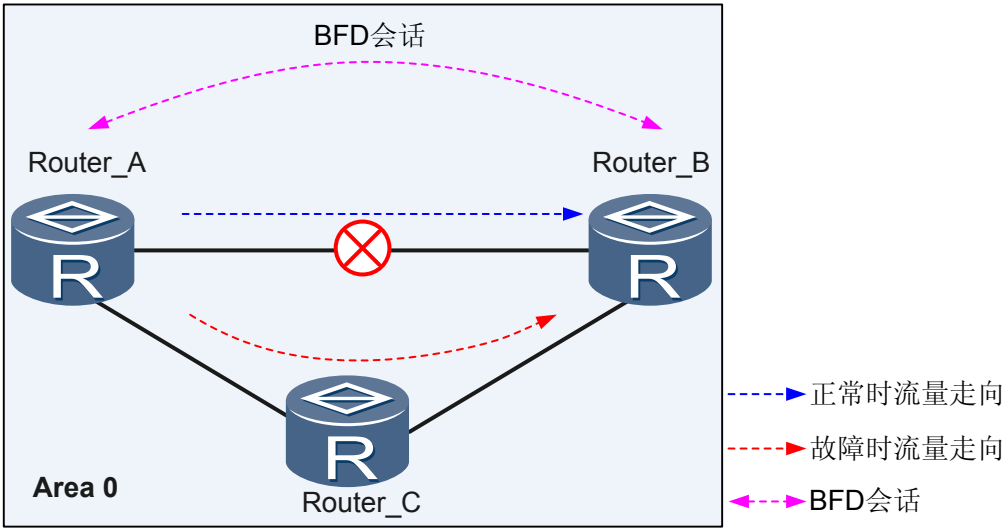
| 有无 BFD | 链路故障检测机制                   | 收敛速度 |
|--------|----------------------------|------|
| 无 BFD  | OSPF Hello Keepalive 定时器超时 | 秒级   |
| 有 BFD  | BFD 会话 Down                | 毫秒级  |

## 典型应用

如图 2-24 所示，Router\_A、Router\_B 和 Router\_C 三台设备之间运行 OSPF，互为邻居。Router\_A->Router\_B 为主链路，Router\_A->Router\_C->Router\_B 为备份链路。

为提高链路状态变化时 OSPF 的收敛速度，在 Router\_A 和 Router\_B 之间的链路上创建 BFD 会话，当 Router\_A 和 Router\_B 之间的链路出现故障时，BFD 能够快速检测到故障并通告给 OSPF 协议，使业务流量通过备份链路传送。

图 2-24 BFD 与 OSPF 联动典型组网应用



### 2.4.4.2 BFD 与静态路由联动

## 应用环境

静态路由是一种需要管理员对某一已知路径手工配置的特殊路由。与动态路由不同，静态路由本身没有检测机制，当网络故障时需要管理员介入。

BFD 与静态路由联动就是将静态路由绑定静态 BFD 会话，可以使静态路由的状态随着 BFD 会话的状态更新。

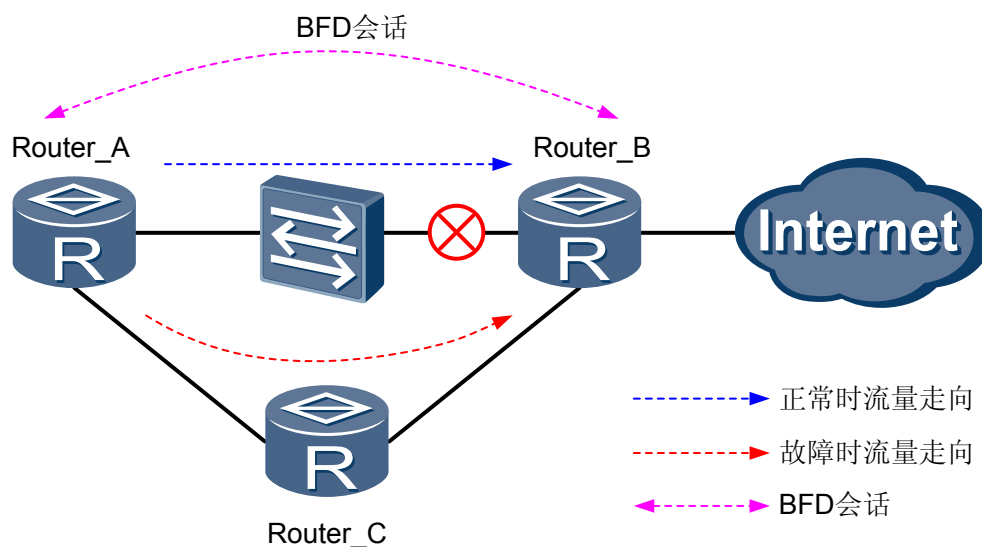
## 典型应用

如图 2-25 所示，Router\_A 通过二层交换机和 Router\_B 相连，通过配置静态路由与外部网络进行正常通信。其中，Router\_A->Router\_B 为主链路，Router\_A->Router\_C->Router\_B 为备份链路。

为增强网络可靠性，提高故障发生时的路由收敛速度，在 Router\_A 和 Router\_B 之间建立 BFD 会话，利用 BFD 检测 Router\_A 和 Router\_B 之间的链路状态：

- 当静态路由上的 BFD 会话检测到故障（由 Up 转为 Down），BFD 会将故障上报系统，系统将这条路由从 IP 路由表中删除，业务流量切换到备份链路上。
- 当静态路由上的 BFD 会话成功建立（由 Down 转为 Up），BFD 会上报系统，系统将这条路由加入 IP 路由表，业务流量恢复到主链路上。

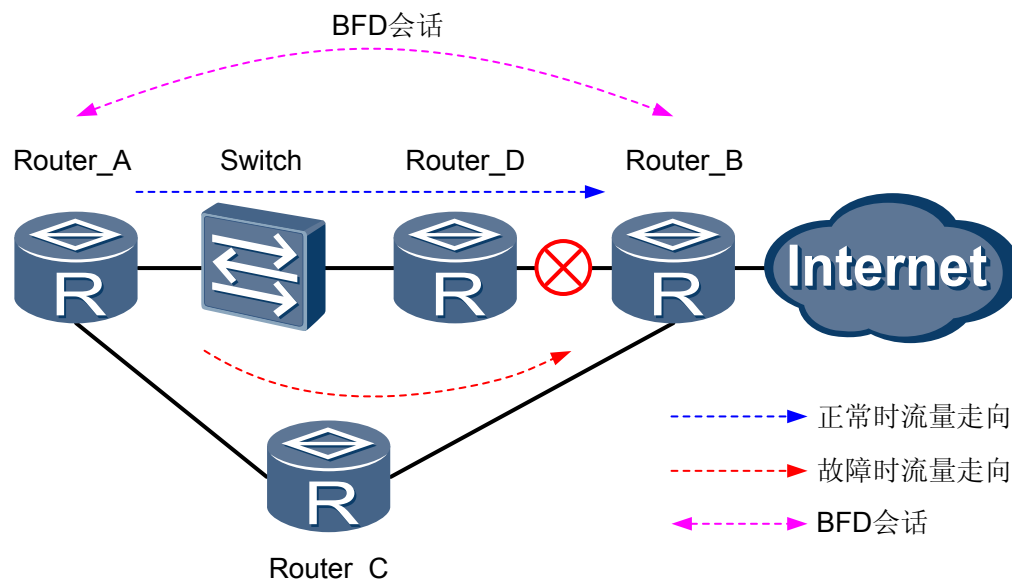
图 2-25 BFD 与静态路由联动典型组网应用（单跳检测）



BFD 与静态路由联动时，支持两种检测方式：

- 单跳检测  
建立 BFD 会话的两端直连或者通过二层交换机相连，BFD 会话的出接口即静态路由的出接口，对端 IP 地址即路由的下一跳。图 2-25 所示即为典型的单跳检测组网应用。
- 多跳检测  
建立 BFD 会话的两端之间非直接相连，而是存在多跳路由通道，如图 2-26 所示。此时，BFD 会话绑定对端 IP 地址但不绑定静态路由的出接口。

图 2-26 BFD 与静态路由联动典型组网应用（多跳检测）



### 2.4.4.3 BFD 与 FRR 联动

#### 应用环境

FRR（Fast ReRoute）是指当物理层或链路层检测到故障时将此消息上报上层路由系统，同时立即开始采取措施，使用一条备份的链路将报文转发出去，从而将链路故障对承载业务的影响降低到最小限度。

在传统的 IP 网络上，转发链路出现底层故障后，最为直观的表现是路由器上的物理接口状态变为 Down 状态。路由器检测到这种故障后，会通知上层路由系统进行相应更新，并重新计算路由。通常从链路故障发生到路由系统完成路由收敛（重新选择了一条可用的路由），要经历几秒钟的时间。

对于网络上某些对延时、丢包等非常敏感的业务来说，秒级的收敛时间是不能忍受的，因为这将导致当前业务的中断。比如 VoIP（Voice Over IP）业务所能容忍的网络中断时间为毫秒级。IP FRR 特性能够保证转发系统快速地对这种故障进行检测并采取措施，尽快让业务流恢复正常。

但是 IP FRR 需要故障检测机制（如 BFD）触发才能生效。

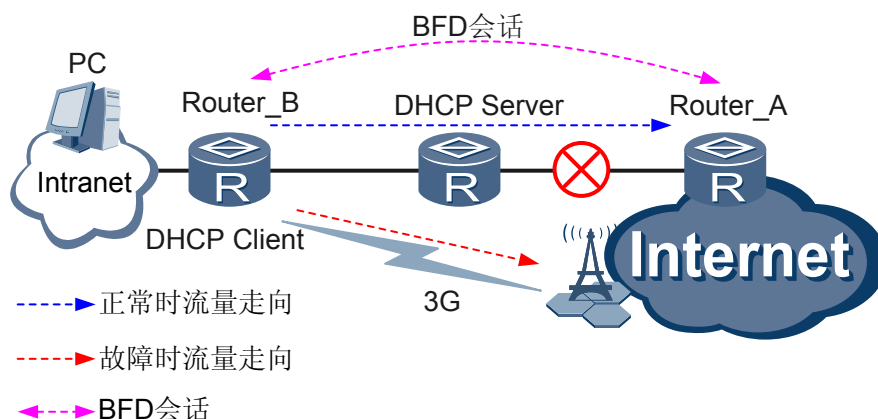
#### 典型应用

如图 2-27 所示，Router\_A 和 Router\_B 之间存在两条链路，主链路为 Router\_A->Router\_C->Router\_B，备份链路为 Router\_A->Router\_D->Router\_B。

在 Router\_A 和 Router\_B 之间创建 BFD 会话，当主链路发生故障时，BFD 将此消息上报给 FRR，FRR 会快速将流量切换到备份链路上。



图 2-28 BFD 与 DHCP 联动典型组网应用



为确保主链路出现故障时，DHCP 客户端能快速感知并迅速采取链路切换的措施，在 Router\_A 和 Router\_B 之间创建静态 BFD 会话，并在 Router\_B 上将 DHCP 与 BFD 绑定。

Router\_B 通过 DHCP 与 BFD 联动，可实现以下功能：

- 当 BFD 检测到主链路故障时，系统会将 DHCP 链路失效，业务流量切换到备份链路上。
- 当 BFD 检测到主链路恢复故障时，业务流量将切换回主链路上。

#### 2.4.4.5 BFD 与策略路由联动

##### 应用环境

策略路由与单纯依照 IP 报文的目的地址查找 FIB 表进行转发不同，是一种依据用户制定的策略进行路由选择的机制，可应用于安全、负载分担等目的。

策略路由支持基于到达报文的源地址、报文类型等信息进行路由选择。对于满足一定条件（ACL 规则）的报文，将执行一定的操作（依据转发报文的出接口和下一跳、报文的缺省出接口和下一跳等进行转发），以指导报文的转发。

由于策略路由无法感知所在链路的可达性，当链路不可达时，对报文执行设置转发操作，可能会导致报文转发失败。

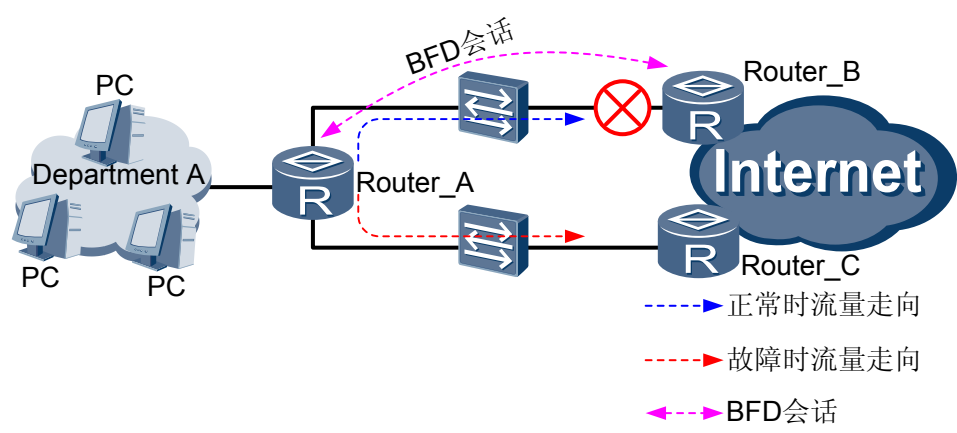
策略路由通过与 BFD 联动，可以解决上述问题，增强了策略路由应用的灵活性，以及策略路由对网络环境的动态感知能力。通过将策略路由的执行动作与静态 BFD 会话关联，由 BFD 快速监视策略路由指定的下一跳或出接口的链路的可达性，通过 BFD 会话的状态来动态决定策略路由的可用性。

##### 典型应用

如图 2-29 所示，Router\_A 是某公司的出口网关，存在两条接入 Internet 的业务链路。正常情况下，期望该公司部门 A 发起的业务流量通过 Router\_A 流向 Router\_B。异常情况下，期望由另外一条链路承担业务流量。

为快速动态感知策略路由的链路可达性，在 Router\_A 和 Router\_B 上建立 BFD 会话，当 Router\_B 和二层交换机之间的链路发生故障后，BFD 能够快速检测并通告 Router\_A，使得 BFD 会话绑定的策略路由失效。这样 Router\_A 在查找路由时将查找备份的路由，以保持业务的持续流通。

图 2-29 BFD 与策略路由联动的典型组网应用



2.4.5 原理描述

2.4.5.1 BFD 报文

BFD 报文有两种类型，分别为 BFD 控制报文和 BFD 回声报文。

BFD 控制报文

BFD 控制报文封装在 UDP 报文中传送，其 UDP 目的端口号为 3784。  
BFD 控制报文包括一个必选部分和一个可选的验证部分，报文格式如图 2-30 所示。

图 2-30 BFD 控制报文格式

|                               |      |                        |   |   |   |                                      |   |   |             |    |  |  |        |    |  |
|-------------------------------|------|------------------------|---|---|---|--------------------------------------|---|---|-------------|----|--|--|--------|----|--|
| 0                             |      | 7                      |   |   |   | 16                                   |   |   |             | 23 |  |  |        | 31 |  |
| Vers                          | Diag | Sta                    | P | F | C | A                                    | D | R | Detect Mult |    |  |  | Length |    |  |
| My Discriminator              |      |                        |   |   |   |                                      |   |   |             |    |  |  |        |    |  |
| Your Discriminator            |      |                        |   |   |   |                                      |   |   |             |    |  |  |        |    |  |
| Required Min TX Interval      |      |                        |   |   |   |                                      |   |   |             |    |  |  |        |    |  |
| Required Min RX Interval      |      |                        |   |   |   |                                      |   |   |             |    |  |  |        |    |  |
| Required Min Echo RX Interval |      |                        |   |   |   |                                      |   |   |             |    |  |  |        |    |  |
| Auth Type<br>(Optional)       |      | Auth Len<br>(Optional) |   |   |   | Authentication Data...<br>(Optional) |   |   |             |    |  |  |        |    |  |



说明

USG 不支持 BFD 验证功能。

报文中各字段的含义解释如表 2-4 所示。

表 2-4 BFD 控制报文字段含义

| 字段                | 长度   | 含义                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-------------------|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Vers (Version)    | 3 比特 | 协议的版本号，目前版本号为 1。                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| Diag (Diagnostic) | 5 比特 | 本地系统最后一次从 Up 状态转换到其他状态的原因。不同值表示的原因如下： <ul style="list-style-type: none"> <li>● 0: No Diagnostic (不诊断)</li> <li>● 1: Control Detection Time Expired (控制检测超时)</li> <li>● 2: Echo Function Failed (回声功能失败)</li> <li>● 3: Neighbor Signaled Session Down (邻居通告会话 Down)</li> <li>● 4: Forwarding Plane Reset (转发面复位)</li> <li>● 5: Path Down (通道失效)</li> <li>● 6: Concatenated Path Down (链接通道失效)</li> <li>● 7: Administratively Down (管理 Down)</li> <li>● 8: Reverse Concatenated Path Down (反向链路 Down)</li> <li>● 9 ~ 31: Reserved for future use (保留)</li> </ul> |
| Sta (State)       | 2 比特 | 当前 BFD 会话的状态。不同值表示的状态如下： <ul style="list-style-type: none"> <li>● 0: AdminDown, 会话处于管理性 Down 状态。</li> <li>● 1: Down, 会话处于 Down 状态或刚刚创建。</li> <li>● 2: Init, 已经能够与对端系统通信，本端希望使会话进入 Up 状态。</li> <li>● 3: Up, 会话已经建立成功。</li> </ul>                                                                                                                                                                                                                                                                                                                                |
| P (Poll)          | 1 比特 | 连接请求确认位。不同值表示的含义如下： <ul style="list-style-type: none"> <li>● 1: 表示发送系统请求进行连接确认，或者发送请求参数改变的确认。</li> <li>● 0: 表示发送系统不请求确认。</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                            |
| F (Final)         | 1 比特 | 是否对收到的 P 比特为 1 的 BFD 控制报文进行响应。不同值表示的含义如下： <ul style="list-style-type: none"> <li>● 1: 表示发送系统响应一个接收到 P 比特为 1 的 BFD 控制报文。</li> <li>● 0: 表示发送系统不响应一个 P 比特为 1 的 BFD 控制报文。</li> </ul>                                                                                                                                                                                                                                                                                                                                                                               |



| 字段                                   | 长度   | 含义                                                                                                                                                                                   |
|--------------------------------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| C (Control Plane Independent)        | 1 比特 | BFD 控制报文是否在控制平面传输。不同值表示的含义如下：<br><ul style="list-style-type: none"> <li>● 1：表示发送系统的 BFD 实现不依赖于它的控制平面。即，BFD 报文在转发平面传输，即使控制平面失效，BFD 仍然能够起作用。</li> <li>● 0：表示 BFD 报文在控制平面传输。</li> </ul> |
| A (Authentication Present)           | 1 比特 | BFD 控制报文中是否包含验证字段。不同值表示的含义如下：<br><ul style="list-style-type: none"> <li>● 1：包含验证字段，会话需要被验证。</li> <li>● 0：不包含验证字段，会话不需要验证。</li> </ul> <b>说明</b><br>USG 目前不提供 BFD 验证功能，A 比特始终置 0。      |
| D (Demand)                           | 1 比特 | 查询模式操作位。不同值表示的含义如下：<br><ul style="list-style-type: none"> <li>● 1：表示发送系统希望工作在查询模式。</li> <li>● 0：表示发送系统不希望、或不能工作在查询模式。</li> </ul>                                                     |
| R (Reserved)                         | 1 比特 | 该字段在发送时设置为 0，在接收时忽略。                                                                                                                                                                 |
| Detect Mult (Detect time multiplier) | 1 字节 | 检测时间倍数，即接收方允许发送方发送报文的最大连续丢包数，用来检测链路是否正常。<br><ul style="list-style-type: none"> <li>● 查询模式：采用本地检测倍数。</li> <li>● 异步模式：采用对端检测倍数。</li> </ul>                                             |
| Length                               | 1 字节 | BFD 控制报文的长度，单位为字节。                                                                                                                                                                   |
| My Discriminator                     | 4 字节 | 发送系统产生的一个唯一的、非 0 鉴别值，用来区分一个系统的多个 BFD 会话。                                                                                                                                             |
| Your Discriminator                   | 4 字节 | 接收到的远端系统的“My Discriminator”，如果没有收到远端的“My Discriminator”，该字段填 0。                                                                                                                      |
| Desired Min Tx Interval              | 4 字节 | 本地系统发送 BFD 控制报文时想要采用的最小时间间隔，单位为微秒。                                                                                                                                                   |
| Required Min Rx Interval             | 4 字节 | 本地系统能够支持的接收两个 BFD 控制报文之间的间隔，单位为微秒。                                                                                                                                                   |
| Required Min Echo Rx Interval        | 4 字节 | 本地系统能够支持的接收两个 BFD 回声报文之间的间隔，单位为微秒。如果这个值设置为 0，则发送系统不支持接收 BFD 回声报文。                                                                                                                    |

| 字段                  | 长度   | 含义                                                                                                                                                                                                                                                                                      |
|---------------------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Auth Type           | 1 字节 | BFD 控制报文使用的认证类型。不同值表示的认证类型如下： <ul style="list-style-type: none"><li>● 0: Reserved</li><li>● 1: Simple Password</li><li>● 2: Keyed MD5</li><li>● 3: Meticulous Keyed MD5</li><li>● 4: Keyed SHA1</li><li>● 5: Meticulous Keyed SHA1</li><li>● 6 ~ 255: Reserved for future use</li></ul> |
| Auth Len            | 1 字节 | 认证字段的长度，包括认证类型与认证长度字段，单位为字节。                                                                                                                                                                                                                                                            |
| Authentication Data | 2 字节 | 认证数据区。                                                                                                                                                                                                                                                                                  |

## BFD 回声报文

BFD 回声报文（Echo）提供了一种不依赖于 BFD 控制报文的故障检测方法。本端发送本端接收，远端不对报文进行处理，而只是将此报文在反向通道上返回。因此 BFD 协议并没有对 BFD 回声报文的格式进行定义，唯一的要求是发送方能够通过报文内容区分会话。

BFD 回声报文采用 UDP 封装，目的端口号为 3784，目的 IP 地址为发送接口的地址，源 IP 地址由配置产生。

### 2.4.5.2 BFD 检测机制

BFD 的检测机制是两个系统建立 BFD 会话，并沿它们之间的路径周期性发送 BFD 控制报文，如果一方在既定的时间内没有收到 BFD 控制报文，则认为路径上发生了故障。

BFD 控制报文封装在 UDP 报文中传送。会话开始阶段，双方系统通过控制报文中携带的参数（会话标识符、期望的收发报文最小时间间隔、本端 BFD 会话状态等）进行协商。协商成功后，以协商的报文收发时间在彼此之间的路径上定时发送 BFD 控制报文。

为满足快速检测的需求，BFD 协议规定发送间隔和接收间隔单位是微秒。但限于目前的设备处理能力，大部分厂商的设备配置 BFD 时只能达到毫秒级，在进行内部处理时再转换到微秒。

## 检测模式

BFD 提供以下检测模式：

- 异步模式

在异步模式下，系统之间会按照协商好的周期发送 BFD 控制报文，如果某个系统在检测时间内没有收到对端发来的 BFD 控制报文，就宣告 BFD 会话的状态为 Down。异步模式是最常用的 BFD 检测模式。

- 查询模式

在查询模式下，一旦 BFD 会话建立，系统就不再周期性发送 BFD 控制报文，而是通过其他与 BFD 无关的机制检测连通性（比如路由协议的 Hello 机制、硬件检测机制等），从而减少 BFD 会话带来的开销。在查询模式下，系统存在一个查询定时器，当查询定时器到期后，系统会发送一个短序列的查询报文检查链路，如果没有收到返回的报文就宣布会话的状态为 Down。

两种模式的一个辅助功能是回声功能。当回声功能激活时，一个 BFD 控制报文按照如下方式发送：本地发送一个 BFD 控制报文，远端系统通过它的转发通道将它们环回回来。如果连续几个回声包都没有接收到，会话状态就被宣布为 Down。回声功能可以与异步模式或者查询模式一起使用。

目前系统只支持异步模式下的单跳会话的被动回声功能。当网络中有支持回声功能的设备时，需要配置 BFD 被动回声功能，以和其他设备兼容。当该设备进入被动回声模式后，普通的 BFD 控制报文周期变长，两端互发源地址和目的地址是本端出接口地址的 BFD 回声报文，通过 ICMP 重定向到达本端，来检测链路状态。

## 检测时间

BFD 的检测时间由三个值决定：

- DMTI (Desired Min Tx Interval)：本端想要采用的最短 BFD 控制报文发送周期
- RMRI (Required Min Rx Interval)：本端能够支持的最短 BFD 控制报文接收周期
- Detect Mult (Detect time multiplier)：检测时间倍数

首先，一个系统接收到对端发来的 BFD 控制报文后，将该报文携带的 RMRI 与自己本地的 DMTI 进行比较，取二者中的较大值作为自己发送 BFD 控制报文的周期。即，速度较慢的系统决定 BFD 控制报文的发送速率。

Detect Mult 不需要协商，两端系统各自配置。

异步模式的检测时间=接收到的远端 Detect Multi  $\times$  max（本地的 RMRI，接收到的 DMTI）

查询模式的检测时间=本地的 Detect Multi  $\times$  max（本地的 DMTI，接收到的 RMRI）

例如，本地的 RMRI 为 400 毫秒，本地的 DMTI 为 300 秒，接收到的 DMTI 为 300 毫秒，接收到的 RMRI 为 400 秒，接收到的 Detect Multi 为 4，本地的 Detect Multi 为 5。

则：异步模式的检测时间=4 $\times$ max（400 毫秒，300 毫秒）=1600 毫秒；查询模式的检测时间=5 $\times$ max（300 毫秒，400 毫秒）=2000 毫秒。

DMTI、RMRI、Detect Multi 都是可以独立配置的，因此，有可能存在两个系统收发 BFD 控制报文速率不同的情况。

对于系统间使用同一种介质的情况，建议两端配置参数相同。

## 检测参数协商过程

BFD 会话建立后，可以动态改变相关检测参数，而不影响会话的当前状态。设备对于各个检测参数改变后的处理过程如下：

- 改变 DMTI

1. 本端立即在发送周期中发送一个 P 比特置 1 的 BFD 控制报文，携带新的 DMTI。

2. 本端重新计算发送周期，和当前的发送周期进行比较。

如果发送周期需要从大改小，则：

- 本端立即重启发送定时器，以新的发送周期发送 P 比特置 1 的 BFD 控制报文。
- 对端收到 P 比特置 1 的控制报文后，响应一个 F 比特置 1 的 BFD 控制报文；重新计算检测时间，并立即重启检测定时器，按照新的检测时间检测。
- 本端收到对端回应的 F 比特置 1 的 BFD 控制报文后，停止发送 P 比特置 1 的 BFD 控制报文。

如果发送周期从小改大，则：

- 本端按原来的发送间隔发送 P 比特置 1 的 BFD 控制报文，并携带新的 DMTI。
- 对端收到 P 比特置 1 的控制报文后，响应一个 F 比特置 1 的 BFD 控制报文；重新计算检测时间，并立即重启检测定时器，按照新的检测时间检测。
- 本端收到对端回应的 F 比特置 1 的 BFD 控制报文后，停止发送 P 比特置 1 的 BFD 控制报文，并且重启发送定时器，以新的发送周期发送 BFD 报文。

如果重新计算得到的发送周期和当前的发送周期相等，则本端不改变发送周期。

● 改变 RMRI

1. 本端立即在发送周期中发送一个 P 比特置 1 的 BFD 控制报文，携带新的 RMRI。
2. 本端重新计算检测时间，与当前的检测时间进行比较。

如果检测周期需要从小变大，则：

- 本端立即重启检测时间定时器，并以新的检测时间进行检测；继续发送 P 比特置 1、携带新 RMRI 的 BFD 控制报文。
- 对端收到 P 比特置 1 的 BFD 控制报文后，立即响应一个 F 比特置 1 的 BFD 控制报文；重新计算发送周期，并立即重启发送定时器。
- 本端收到对端响应的 F 比特置 1 的 BFD 控制报文，停止发送 P 比特置 1 的 BFD 控制报文。

如果检测周期需要从大变小，则：

- 本端按照以前的发送间隔发送 P 比特置 1、携带新 RMRI 的 BFD 控制报文。
- 对端收到 P 比特置 1 的 BFD 控制报文后，立即响应一个 F 比特置 1 的 BFD 控制报文；重新计算发送周期，并立即重启发送定时器。
- 本端收到对端响应的 F 比特置 1 的 BFD 控制报文后，停止发送 P 比特置 1 的 BFD 报文，并更新检测时间、重启检测定时器。

如果重新计算得到的检测时间和当前检测时间相等，则本端不改变检测周期。

3. 改变 Detect Multi

- a. 本端立即在发送周期中发送一个 P 比特置 1 的 BFD 控制报文，携带改变后的检测倍数，并且在以后的报文中都携带此检测倍数。
- b. 对端收到 BFD 控制报文后，重新计算检测时间并使用新的检测时间检测。

### 2.4.5.3 BFD 会话管理

## 会话建立方式

BFD 通过控制报文中的 My Discriminator（本地标识符）和 Your Discriminator（远端标识符）区分不同的会话。按照本地标识符和远端标识符创建方式的差异区分，USG 支持以下 BFD 会话类型：

- 手工指定标识符的静态 BFD 会话

手工指定标识符的静态 BFD 会话是指手工配置 BFD 会话参数，包括配置本地标识符和远端标识符，然后手工下发 BFD 会话建立请求。

这种方式的缺点是会带来人为的配置错误，比如配置了错误的本地标识符或者远端标识符时，BFD 会话将不能正常工作。并且 BFD 会话的建立和删除需要手工触发，缺乏灵活性。

在策略路由、DHCP、FRR 与 BFD 联动应用中，必须使用手工指定标识符的静态 BFD 会话。而针对静态路由与 BFD 联动的应用，可以根据网络情况选择使用手工指定标识符的静态 BFD 会话或者标识符自协商的静态 BFD 会话。

- 标识符自协商的静态 BFD 会话

标识符自协商的静态 BFD 会话是指手工创建 BFD 会话，但不需要配置本地标识符和远端标识符，本地标识符和远端标识符通过会话协商获得。

在静态路由与 BFD 联动应用中，如果对端设备不支持静态 BFD 会话，而采用动态 BFD 会话，此时本端设备既要与之互通，又要能够实现静态路由与 BFD 联动时，必须使用静态标识符自协商 BFD 会话。

- 协议触发的动态 BFD 会话

协议触发的动态 BFD 会话是指由路由协议动态触发而建立 BFD 会话。

动态建立 BFD 会话时，系统对本地标识符和远端标识符的处理方式如下：

- 动态分配本地标识符

当应用程序触发动态创建 BFD 会话时，系统分配属于动态会话标识符区域的值作为 BFD 会话的本地标识符。然后向对端发送 Your Discriminator 的值为 0（My Discriminator 的值为分配到的本地标识符，State 的值为 Down）的 BFD 控制报文，进行会话协商。



说明

系统通过划分标识符区域的方式来区分静态 BFD 会话和动态 BFD 会话，静态配置 BFD 会话的本地标识符取值范围是 1 ~ 8191，动态创建 BFD 会话的本地标识符取值范围是 8192 ~ 16383。

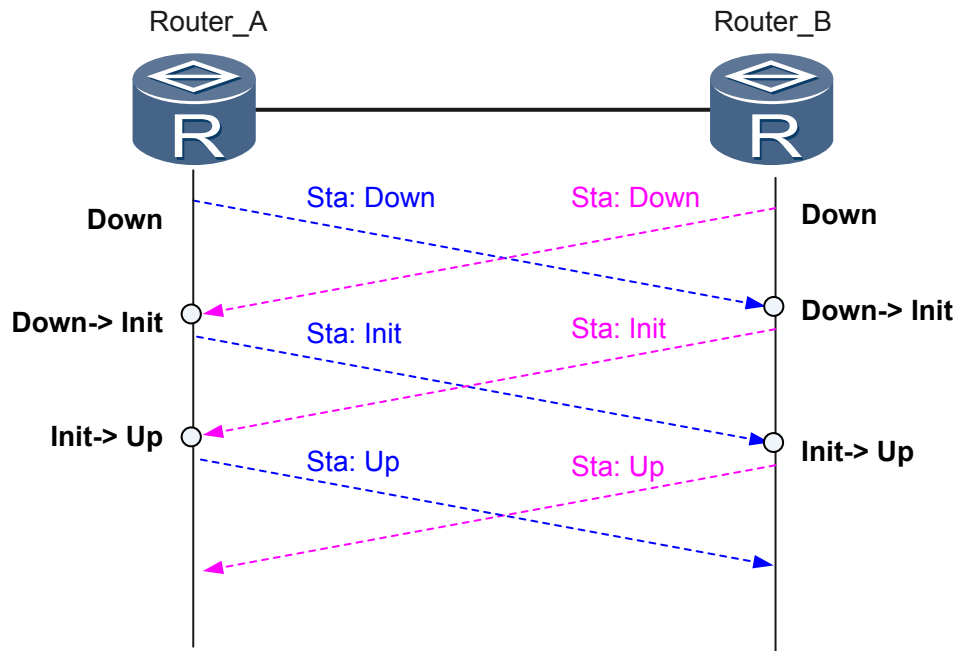
- 自学习远端标识符

当 BFD 会话的一端收到 Your Discriminator 的值为 0 的 BFD 控制报文时，根据四元组（源地址、目的地址、出接口、VPN 索引）判断该报文是否与本地 BFD 会话匹配，如果匹配，则学习接收到的 BFD 控制报文中 My Discriminator 的值，获取远端标识符。

## 会话建立过程

BFD 使用三路握手的机制来建立会话，发送方在发送 BFD 控制报文时会在 Sta 字段填入本地当前的会话状态，接收方根据收到的 BFD 控制报文的 Sta 字段以及本端当前会话状态来进行状态机的迁移，建立会话。图 2-31 以 BFD 会话建立为例，简单介绍状态机的迁移过程。

图 2-31 BFD 会话连接建立



1. Router\_A 和 Router\_B 的 BFD 收到上层应用的通知后，发送状态为 Down 的 BFD 控制报文。对于手工指定标识符的静态 BFD 会话，报文中的 Your Discriminator 的值是用户指定的；对于标识符自协商的静态 BFD 会话，报文中的 Your Discriminator 的值由双方协商而定；对于动态创建 BFD 会话，Your Discriminator 的值是 0。
2. Router\_B 收到状态为 Down 的 BFD 控制报文后，本地状态切换至 Init，并发送状态为 Init 的 BFD 控制报文。Router\_A 的 BFD 状态变化同 Router\_B。
3. Router\_B 收到状态为 Init 的 BFD 控制报文后，本地状态切换至 Up，并发送状态为 Up 的 BFD 控制报文。Router\_A 的 BFD 状态变化同 Router\_B。
4. Router\_A 和 Router\_B 双方状态都为 Up，会话成功建立并开始检测链路状态。

Router\_A 和 Router\_B 发生“DOWN=>INIT”的状态迁移后，会启动一个超时定时器。如果定时器超时仍未收到状态为 Init 或 Up 的 BFD 控制报文，则本地状态自动切换回 Down。

# 3 虚拟防火墙

---

## 关于本章

[3.1 介绍](#)

[3.2 可获得性](#)

[3.3 原理描述](#)

介绍 VPN 实例的功能和 USG 上的实现。

## 3.1 介绍

### 定义

通过在防火墙上创建虚拟防火墙 VFW（Virtual-firewall），能够提供防火墙的出租业务。将一台防火墙从逻辑上划分为多个 VPN（Virtual Private Network），向多个小型私有网络提供相对独立的安全服务。

### 目的

针对此类用户的需求，通过在防火墙上创建虚拟防火墙 VFW（Virtual-firewall），能够提供防火墙的出租业务。将一台防火墙从逻辑上划分为多个 VPN（Virtual Private Network），向多个小型私有网络提供相对独立的安全服务。每个虚拟防火墙都是 VPN 实例（VPN-Instance）、安全实例和配置实例的综合体，能够为虚拟防火墙用户提供私有的路由转发平面、安全服务和配置管理平面。

VPN 实例为虚拟防火墙提供相互隔离的 VPN 路由，与虚拟防火墙相关的信息主要包括：VPN 路由以及与 VPN 实例绑定的接口。VPN 路由将为转发来自与 VPN 实例绑定的接口的报文提供路由支持。

## 3.2 可获得性

### License 支持

USG2110-X/2100 虚拟防火墙特性不受 License 控制，最多支持 10 个虚拟防火墙。  
USG2200/5100/5500 虚拟防火墙特性受 License 控制，无 License 情况下，缺省开启 10 个虚拟防火墙。USG2100/2200/5100 BSR/HSR 不支持虚拟防火墙。

### 版本支持

| 产品                                                                      | 最低支持版本                                                         |
|-------------------------------------------------------------------------|----------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00<br>USG5500 V200R001C00 |

## 3.3 原理描述

介绍 VPN 实例的功能和 USG 上的实现。

近年来小型私有网络不断增加，这些网络一般对应小型企业。此类用户有如下特点：

- 有较强的安全防范需求。
- 经济上无法负担一台专有安全设备。



USG 提出多实例解决方案，即将一台 USG 从逻辑上划分为多台虚拟防火墙，分别为多个小型私有网络提供独立的安全保障。对于网络运营商，应用这种技术即可向外出租网络安全保障服务。

每台虚拟防火墙都是 VPN（Virtual Private Network）实例（VPN-Instance）、安全实例和配置实例的综合体。它能够为用户提供私有的路由转发平面、安全服务和配置管理平面。

## VPN 实例

VPN 实例为虚拟防火墙用户提供相互隔离的 VPN 路由，与虚拟防火墙一一对应。这些 VPN 路由将为其对应的虚拟防火墙接收的报文提供路由支持。

## 安全实例

安全实例为虚拟防火墙用户提供相互隔离的安全服务，与虚拟防火墙一一对应。这些安全实例具备私有的接口、安全区域、安全域间、ACL（Access Control List）和 NAT（Network Address Translation）地址池，并能为虚拟防火墙用户提供地址绑定、黑名单、地址转换、包过滤、统计、攻击防范、ASPF 和 NAT 等私有的安全服务。

## 配置实例

配置实例为虚拟防火墙用户提供相互隔离的配置管理平面，与虚拟防火墙一一对应。这些配置实例使虚拟防火墙用户能够登录到各自的虚拟防火墙，并管理和维护上述私有 VPN 实例和安全实例。

# 4 接入

## 关于本章

### 4.1 链路聚合

链路聚合（Link Aggregation）是将一组物理接口捆绑在一起作为一个逻辑接口来增加带宽的一种方法。

### 4.2 PPP 和 MP

PPP 运行在点对点链路上，支持对链路两端设备的认证。MP 捆绑多个 PPP 链路，提高了传输带宽。

### 4.3 HDLC

HDLC 运行在点对点链路，配置简单。

### 4.4 VLAN

VLAN 可以把连接到同一个物理以太网的用户区分到逻辑 LAN 中，提高网络的管理能力。

### 4.5 MSTP

在交换式局域网络中开启 MSTP 功能，可有效的避免广播风暴。

### 4.6 PPPoE

PPPoE 利用以太网将大量主机组成网络，通过一个远端接入设备连入因特网，并对接入的每个主机实现控制、计费功能。

### 4.7 WLAN

无线局域网是指以无线信道作传输媒介的计算机局域网，是有线联网方式的重要补充和延伸。

## 4.1 链路聚合

链路聚合（Link Aggregation）是将一组物理接口捆绑在一起作为一个逻辑接口来增加带宽的一种方法。

### 4.1.1 介绍

#### 定义

链路聚合接口把多个物理接口捆绑为一个逻辑接口（Eth-Trunk 接口），提高了传输带宽。捆绑在一起的每个物理接口称为成员接口。

链路聚合根据是否启用链路聚合控制协议分为以下两种类型：

- 手工链路聚合

手工模式是一种最基本的链路聚合方式，在该模式下 Eth-Trunk 接口的建立，成员接口的加入，以及哪些接口作为活动接口完全由手工来配置，没有链路聚合控制协议的参与。

- 静态 LACP 模式链路聚合

静态 LACP 模式下，Eth-Trunk 接口的建立，成员接口的加入，都是由手工配置完成的。但与手工负载分担模式链路聚合不同的是，该模式下 LACP 协议报文负责活动接口的选择。

 说明

链路聚合存在活动接口和非活动接口两种。转发数据的接口称为活动接口，而不转发数据的接口称为非活动接口。

#### 目的

链路聚合技术可以增加带宽和提高可靠性。

在没有使用链路聚合前，百兆接口互连的网络若想达到更高的数据传输速率，则需要更换接口硬件（比如使用千兆接口）。这样的解决方案成本昂贵，不适合中小型企业应用。采用链路聚合就可以把多个接口捆绑在一起，则可以用较低的成本满足增加接口带宽的需求。

同时，一条聚合链路中可以有多个成员接口。某个成员接口 Down 后，不会导致整条链路 Down，提高了链路的可靠性。

### 4.1.2 参考标准和协议

与本特性相关的参考标准与协议如下：

- IEEE 802.3AD: IEEE Std 802.3ad - 2005 IEEE Standard for Link Aggregation operation, Link Aggregation Control, Link Aggregation Control Protocol, Marker protocol and Configuration capabilities and restrictions

### 4.1.3 可获得性

## License 支持

本特性无需 License 支持。

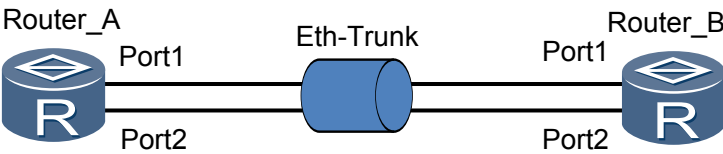
## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 4.1.4 应用场景

如图 4-1 所示，Router\_A 与 Router\_B 之间创建 Eth-Trunk 接口，将两个 GigabitEthernet 接口（Port1 和 Port2）捆绑成一个 Eth-Trunk。Router\_A 与 Router\_B 之间的聚合链路最大带宽达到原接口的两倍。

图 4-1 链路聚合组网图



Eth-Trunk 接口将 Router\_A 和 Router\_B 之间的流量分担到两条链路上，避免所有流量都走同一条链路而导致网络拥塞。当其中一条链路故障，流量切换到另一条链路，提高链路可靠性。

## 4.1.5 原理描述

### 4.1.5.1 链路聚合协议（LACP）

#### 链路聚合协议的引入

作为链路聚合技术，手工链路聚合可以完成多个物理端口聚合成一个 Eth-Trunk 接口来提高带宽，同时只能够检测到成员链路有断路等有限故障，无法检测链路错连等故障。LACP 的技术出现后，提高了链路聚合的容错性，并且能提供 M:N 备份功能，保证成员链路的高可靠性。

例如 Router\_A 与 Router\_B 之间创建手工链路聚合，将四个以太网接口捆绑成一条聚合链路。由于错将 Router\_A 上的一个以太网接口与 Router\_C 相连，这将会导致 Router\_A 向 Router\_B 传输数据时可能会将本应该发到 Router\_B 的数据发送到 Router\_C 上。而手工链路聚合不能及时的检测到故障。如果在路由器上都启用 LACP 协议，通过 LACP 协议协商可以保证 Router\_A 发送的数据能够正确到达 Router\_B。

## LACP 基本概念

LACP 为交换数据的设备提供一种标准的协商方式，以供系统根据自身配置自动形成聚合链路并启动聚合链路收发数据。聚合链路形成以后，负责维护链路状态。在聚合条件发生变化时，自动调整或解散链路聚合。

- 系统 LACP 优先级

系统 LACP 优先级是为了区分两端设备优先级的高低而配置的参数。静态 LACP 模式下，要求两端设备所选择的接口一致，否则链路聚合组就无法建立。所以需要指定其中一端具有高优先级（称为主动端），另一端（称为被动端）根据高优先级的一端来选择活动接口。

系统 LACP 优先级值越小优先级越高。在两端设备中选择系统 LACP 优先级数值较小一端作为主动端，如果系统 LACP 优先级相同则选择 MAC 地址较小的一端作为主动端。

- 接口 LACP 优先级

接口 LACP 优先级是为了区别不同接口被选为活动接口的优先程度。接口 LACP 优先级值越小，优先级越高。

- 活动接口和非活动接口

链路聚合存在活动接口和非活动接口两种。转发数据的接口称为活动接口，而不转发数据的接口称为非活动接口。

只有当前活动接口出现故障时，备份的接口才可以由非活动接口转变为活动接口。

- M:N 备份

静态 LACP 模式链路聚合是一种利用 LACP 协议进行参数协商选取活动链路的聚合模式。该模式由 LACP 协议确定活动和非活动链路，又称为 M:N 模式，即 M 条活动链路与 N 条备份链路的模式。这种模式提供了更高的链路可靠性，并且可以在 M 条链路中实现不同方式的负载均衡。

当 M 条链路中有一条链路故障时，LACP 会从 N 条备份链路中找出一条正常链路替换有故障的链路。此时链路的实际带宽还是 M 条链路的总和。

- LACP 优先级抢占

LACP 优先级抢占功能可以保持接口 LACP 优先级最高的接口为活动接口。当一条高优先级的接口因故障切换为非活动状态而后再恢复时，如果使能了抢占，则高优先级的接口将在延时一定时间后重新成为活动接口；如果未使能抢占，该接口不能重新成为活动接口。

### 4.1.5.2 链路聚合实现原理

#### 手工链路聚合

手工模式链路聚合是应用比较广泛的一种链路聚合，允许在聚合组中手工加入多个成员接口，所有的接口均处于转发状态，分担负载的流量。当需要在两个直连设备间提供一个较大的链路带宽而对端设备又不支持 LACP 协议时，可以使用手工负载分担模式。

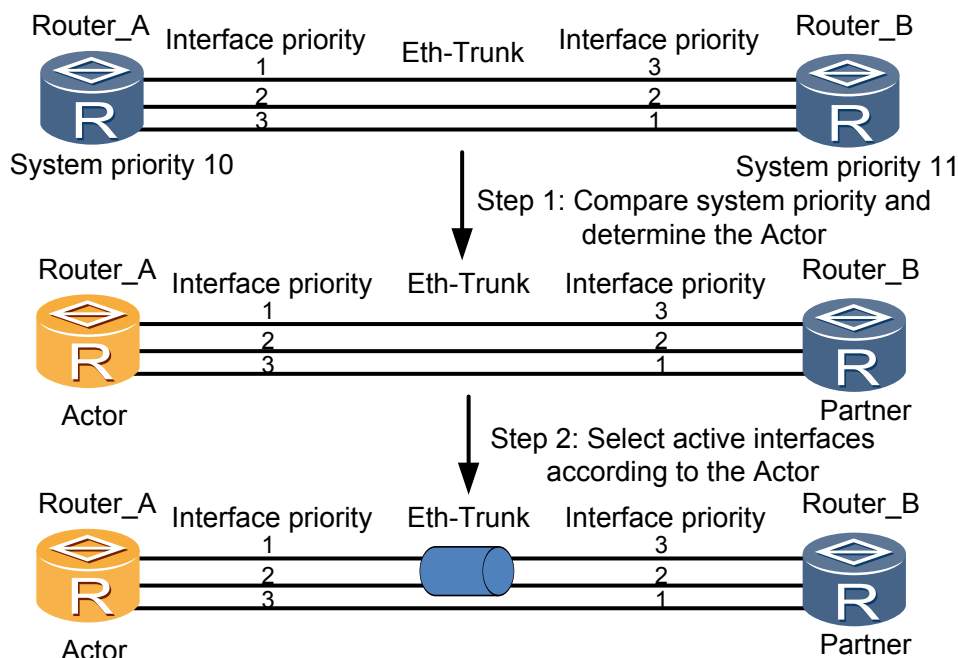
#### 静态 LACP 链路聚合

基于 IEEE802.3ad 标准的 LACP（链路汇聚控制协议）是一种实现链路动态聚合与解聚合的协议。LACP 协议通过 LACPDU（Link Aggregation Control Protocol Data Unit）与对端交互信息。

在静态 LACP 模式的 Eth-Trunk 中加入成员接口后，这些接口将通过发送 LACPDU 向对端通告自己的系统优先级、系统 MAC、接口优先级、接口号和操作 Key 等信息。对端

接收到这些信息后，将这些信息与自身接口所保存的信息比较以选择能够聚合的接口，双方对哪些接口能够成为活动接口达成一致，确定活动链路。活动链路建立的过程如图 4-2 所示。

图 4-2 活动链路建立过程示意图



1. 两端互相发送 LACPDU 报文，两端设备根据系统 LACP 优先级和 MAC 地址确定主动端。  
系统 LACP 优先级值越小优先级越高。在两端设备中选择系统 LACP 优先级数值较小一端作为主动端（Actor），如果系统 LACP 优先级相同则选择 MAC 地址较小的一端作为主动端。
2. 两端设备根据主动端接口 LACP 优先级和接口 ID 确定活动接口。  
两端设备选择了一致的活动接口，数据便可正常转发。

静态 LACP 链路聚合检测到以下事件，会触发的链路切换：

- 链路 Down。
- LACP 协议发现链路故障。
- 在使能了 LACP 抢占前提下，更改备份接口的优先级高于当前活动接口的优先级后，会发生切换的过程。

切换时，系统会使用优先级最高的备份链路转为活动状态并转发数据。

## 4.2 PPP 和 MP

PPP 运行在点对点链路上，支持对链路两端设备的认证。MP 捆绑多个 PPP 链路，提高了传输带宽。

## 4.2.1 介绍

### 定义

PPP（Point-to-Point Protocol）协议是一种在点到点链路上传输、封装网络层数据包的数据链路层协议。

MP（MultiLink PPP）是将多个 PPP 链路捆绑使用的技术。

### 目的

PPP 协议处于 OSI（Open Systems Interconnection）参考模型的第二层，主要用在支持全双工的同异步链路上，进行点到点之间的数据传输。由于它能够为用户提供用户验证，易于扩充，并且支持同异步通信，因而获得广泛应用。

MP 通过将多个 PPP 链路进行捆绑，增加链路带宽。

## 4.2.2 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 4.2.3 原理描述

### 4.2.3.1 PPP 的基本原理

#### PPP 简介

PPP 协议是一种在点到点链路上传输、封装网络层数据包的数据链路层协议。

点到点的直接连接是广域网连接的一种比较简单的形式，点到点连接的线路上链路层封装的协议主要有 PPP 和 HDLC。其中，HDLC 协议只支持同步方式，而 PPP 协议支持同、异步两种传输方式。因此，PPP 协议得到广泛的应用。

从 1994 年至今，PPP 协议本身并没有太大的改变，但由于 PPP 协议所具有的其他链路协议所无法比拟的特性，它得到了越来越广泛的应用，其扩展支持协议也层出不穷，随之而来的是 PPP 协议功能的逐步强大。

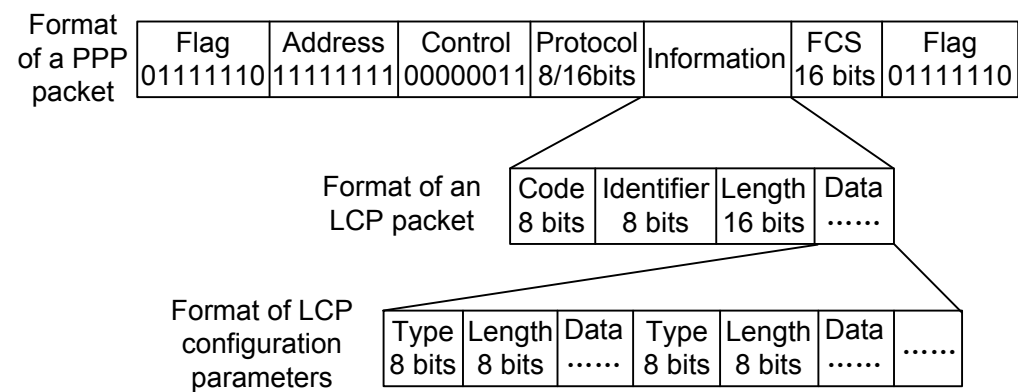
PPP 的基本构架

- PPP 主要由三类协议组成：
- 链路控制协议族（Link Control Protocol）：主要用来建立、拆除和监控 PPP 数据链路。此外，还进行链路层参数的协商，如 MRU、验证方式。
  - 网络层控制协议族（Network Control Protocol）：主要用来协商在数据链路上传输的数据包的格式与类型。
  - PPP 扩展协议族（如 PPPoE）：主要用于提供对 PPP 功能的进一步支持。
- 同时，PPP 还提供了用于网络安全方面的验证协议族（PAP 和 CHAP）。

4.2.3.2 PPP 报文格式

PPP 报文格式如 图 4-3 所示。

图 4-3 PPP 报文格式



PPP 报文封装的帧格式

- PPP 报文封装格式请参见 图 4-3。
- PPP 报文的内容是指 Address、Control、Protocol 和 Information 四个域的内容。各字段的含义如下。
- Flag 域  
Flag 域标识了一个物理帧的起始和结束，该字节为 0x7E。
  - Address 域  
PPP 协议是被运用在点对点的链路上，它可以唯一标识对方。因此使用 PPP 协议互连的两个通信设备无需知道对方的数据链路层地址。所以该字节已无任何意义，按照协议的规定将该字节填充为全 1 的广播地址。
  - Control 域  
同 Address 域一样，PPP 数据帧的 Control 域也没有实际意义，按照协议的规定通信双方将该字节的内容填充为 0x03。  
Address 和 Control 域一起表示了此报文为 PPP 报文，即 PPP 报文头为 FF03。
  - Protocol 域



协议域可用来区分 PPP 数据帧中信息域所承载的数据报文的内容。

协议域的内容必须依据 ISO 3309 的地址扩展机制所给出的规定。该机制规定协议域所填充的内容必须为奇数，也就是要求低字节的最低位为“1”，高字节的最低位为“0”。

如果当发送端发送的 PPP 数据帧的协议域字段不符合上述规定，接收端则会认为此数据帧是不可识别的。接收端向发送端发送一个 Protocol-Reject 报文，在该报文尾部将填充被拒绝报文的协议号。

表 4-1 常见的协议代码

| 协议代码 | 协议类型                                        |
|------|---------------------------------------------|
| 0021 | Internet Protocol                           |
| 002b | Novell IPX                                  |
| 002d | Van Jacobson Compressed TCP/IP              |
| 002f | Van Jacobson Uncompressed TCP/IP            |
| 8021 | Internet Protocol Control Protocol          |
| 802b | Novell IPX Control Protocol                 |
| 8031 | Bridging NC                                 |
| C021 | Link Control Protocol                       |
| C023 | Password Authentication Protocol            |
| C223 | Challenge Handshake Authentication Protocol |

● Information 域

信息域最大长度是 1500 字节，其中包括填充域的内容。信息域的最大长度等于 PPP 协议中 MRU（Maximum Receive Unit）的缺省值。在实际应用当中可根据实际需要进行信息域最大封装长度选项的协商。

如果信息域长度不足 1500 字节，可被填充，但不是必须的。如果填充则需通信双方的两端能辨认出有用与无用的信息方可正常通信。

● FCS 域

校验域的功能主要对 PPP 数据帧传输的正确性进行检测。

在数据帧中引入了一些传输的保证机制，会引入更多的开销，这样可能会增加应用层交互的延迟。

LCP 报文封装的帧格式

LCP 报文封装格式请参见图 4-3。

在链路建立阶段，PPP 协议通过 LCP 报文进行链路的建立和协商过程。此时 LCP 报文作为 PPP 的净载荷被封装在 PPP 数据帧的信息域中，PPP 数据帧的协议域的值固定填充 0xC021。

在链路建立阶段的整个过程中信息域的内容是变化的，它包括很多种类型的报文，所以这些报文也要通过相应的字段来区分。

● Code 域

代码域的长度为一个字节，主要是用来标识 LCP 数据报文的类型。

在链路建立阶段，接收方接收到 LCP 数据报文。当其代码域的值无效时，就会向对端发送一个 LCP 的代码拒绝报文（Code-Reject 报文）。

如果是 IP 报文，则不存在此域，取而代之的是 IP 报文内容。

表 4-2 常见 code 值

| code 值 | 报文类型              |
|--------|-------------------|
| 0x01   | Configure-Request |
| 0x02   | Configure-Ack     |
| 0x03   | Configure-Nak     |
| 0x04   | Configure-Reject  |
| 0x05   | Terminate-Request |
| 0x06   | Terminate-Ack     |
| 0x07   | Code-Reject       |
| 0x08   | Protocol-Reject   |
| 0x09   | Echo-Request      |
| 0x0A   | Echo-Reply        |
| 0x0B   | Discard-Request   |
| 0x0C   | Reserved          |

● Identifier 域

标识域的值表示进行协商报文的匹配关系。

标识域也是一个字节，其目的是用来匹配请求和响应报文。

一般而言，在进入链路建立阶段时，通信双方任何一端都会连续发送几个配置请求报文（Configure-Request 报文）。这几个请求报文的数据域的值可能是完全一样的，只是它们的标志域不同。

通常一个配置请求报文的 ID 是从 0x01 开始逐步加 1 的。

当对端接收到该配置请求报文后，无论使用何种报文回应对方，但必须要求回应报文中的 ID 要与接收报文中的 ID 一致。

当通信设备收到回应后就可以将该回应与发送时的进行比较来决定下一步的操作。

● Length 域

长度域表示此协商报文长度，它包含 Code 域及 Identifier 域的长度。

长度域的值就是该 LCP 报文的总字节数据。它是代码域、标志域、长度域和数据域四个域长度的总和。

长度域所指示字节数之外的字节将被当作填充字节而忽略掉，而且该域的内容不能超过 MRU 的值。

● Data 域

数据域所包含的是协商报文的内容。

- Type 为协商选项类型。
- Length 为协商选项长度，它是指 Data 域的总长度，也就是包含 Type、Length 和 Data。
- Data 为协商的选项具体内容。

表 4-3 常见 Type 中的协商类型值

| 协商类型值 | 协商报文类型                                |
|-------|---------------------------------------|
| 0x01  | Maximum-Receive-Unit                  |
| 0x02  | Async-Control-Character-Map           |
| 0x03  | Authentication-Protocol               |
| 0x04  | Quality-Protocol                      |
| 0x05  | Magic-Number                          |
| 0x06  | RESERVED                              |
| 0x07  | Protocol-Field-Compression            |
| 0x08  | Address-and-Control-Field-Compression |

### 4.2.3.3 PPP 的运行过程

PPP 链路的建立是通过一系列的协商完成的。

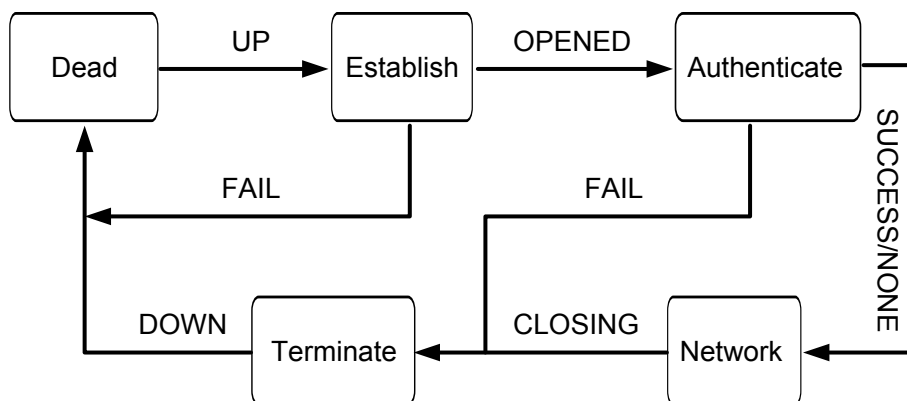
数据通信设备的两端如果希望通过 PPP 协议建立点对点的通信，无论哪一端的设备都需发送 LCP 数据报文来建立链路。LCP 的配置参数选项协商完成后，通信的双方就会根据 LCP 配置请求报文中所协商的认证配置参数选项，决定链路两端设备所采用的认证方式。

协议缺省情况下双方是不进行认证的，直接进入到了 NCP 配置参数选项的协商。直至所经历的几个配置过程全部完成后，点对点的双方就可以开始通过已建立好的链路进行网络层数据报文的传送了，整个链路就处于可用状态。

任何一端收到 LCP 或 NCP 的链路关闭报文，物理层无法检测到载波或管理人员对该链路进行关闭操作，都会将该条链路断开，从而终止 PPP 会话。一般而言，协议是不要求 NCP 有关闭链路的能力的，因此通常情况下关闭链路的数据报文是在 LCP 协商阶段或应用程序会话阶段发出的。

下图是 PPP 协议整个链路过程需经历阶段的状态转移图：

图 4-4 PPP 运行流程图



PPP 运行的过程简单描述如下。

1. 开始建立 PPP 链路时，先进入到 Establish 阶段。
2. 在 Establish 阶段，PPP 链路进行 LCP 协商。协商内容包括工作方式是 SP（Single-link PPP）还是 MP（Multilink PPP）、最大接收单元 MRU、验证方式、魔术字（magic number）和异步字符映射等选项。LCP 协商成功后进入 Opened 状态，表示底层链路已经建立。
3. 如果配置了验证，将进入 Authenticate 阶段，开始 CHAP 或 PAP 验证。如果没有配置验证，则直接进入 NCP 协商阶段。
4. 对于 Authenticate 阶段，如果验证失败，进入 Terminate 阶段，拆除链路，LCP 状态转为 Down。如果验证成功，进入 NCP 协商阶段，此时 LCP 状态仍为 Opened，而 NCP 状态从 Initial 转到 Starting。
5. NCP 协商支持 IPCP、MPLSCP、OSCI CP 等协商。IPCP 协商主要包括双方的 IP 地址。通过 NCP 协商来选择和配置一个网络层协议。只有相应的网络层协议协商成功后，该网络层协议才可以通过这条 PPP 链路发送报文。
6. PPP 链路将一直保持通信，直至有明确的 LCP 或 NCP 帧关闭这条链路，或发生了某些外部事件，例如用户干预。

在点对点链路的配置、维护和终止过程中，PPP 需经历以下几个阶段：

## 链路不可用阶段（Dead）

它有时也称为物理层不可用阶段。PPP 链路都需从这个阶段开始和结束。

当通信双方的两端检测到物理线路激活（通常是检测到链路上有载波信号）时，就会从当前这个阶段跃迁至下一个阶段，即链路建立阶段。

在链路建立阶段主要是通过 LCP 协议进行链路参数的配置，LCP 在此阶段的状态机也会根据不同的事件发生变化。当处于在链路不可用阶段时，LCP 的状态机是处于初始化 Initial 状态或准备启动 Starting 状态，一旦检测到物理线路可用，则 LCP 的状态机就要发生改变。

当然链路被断开后也同样会返回到链路不可用阶段。

## 链路建立阶段（Establish）

它是 PPP 协议最关键和最复杂的阶段。

该阶段主要是发送一些配置报文来配置数据链路，这些配置的参数不包括网络层协议所需的参数。当完成数据报文的交换后，则会继续向下一个阶段跃迁。

下一个阶段既可以是验证阶段，也可以是网络层协议阶段。下一阶段的选择是依据链路两端的设备配置的。通常是由用户来配置。

在链路建立阶段，LCP 的状态机会发生两次改变。

- 当链路处于不可用阶段时，此时 LCP 的状态机处于 Initial 或 Starting。当检测到链路可用时，则物理层会向链路层发送一个 Up 事件。链路层收到该事件后，会将 LCP 的状态机从当前状态改变为 Request-Sent（请求发送状态），根据此时的状态机 LCP 会进行相应的动作，也就是开始发送 Configure-Request 报文来配置数据链路。
- 无论哪一端接收到了 Configure-Ack 报文时，LCP 的状态机又要发生改变，从当前状态改变为 opened 状态。进入 Opened 状态后收到 Configure-Ack 报文的一方则完成了当前阶段，应该向下一个阶段跃迁。

同理可知，另一端也是一样的，但须注意的是在链路配置阶段双方的链路配置操作过程是相互独立的。如果在该阶段收到了非 LCP 数据报文，则会将这些报文丢弃。

## 验证阶段（Authenticate）

多数情况下的链路两端设备是需要经过认证后才进入到网络层协议阶段。

PPP 链路缺省情况下，不进行验证。如果要求验证，在链路建立阶段必须指定验证协议。

PPP 验证有两种用途：

- 主要是用于主机和设备之间，通过 PPP 网络服务器交换电路或拨号接入连接的链路。
- 偶尔也用于专用线路。

PPP 提供两种验证方式。

- PAP: Password Authentication Protocol，密码验证协议
- CHAP: Challenge-Handshake Authentication Protocol，挑战握手验证协议

验证方式的选择是依据在链路建立阶段双方进行协商的结果。然而，链路质量的检测也会在这个阶段同时发生，但协议规定不会让链路质量的检测无限制的延迟验证过程。

在这个阶段仅支持链路控制协议、验证协议和质量检测数据报文，其它的数据报文都会被丢弃。如果在这个阶段再次收到了 Configure-Request 报文，则又会返回到链路建立阶段。

## 网络层协议阶段（Network）

一旦 PPP 完成了前面几个阶段，每种网络层协议（IP、IPX 和 AppleTalk）会通过各自相应的网络控制协议进行配置，每个 NCP 协议可在任何时间打开和关闭。当一个 NCP 的状态机变成 Opened 状态时，则 PPP 就可以开始在链路上承载网络层的数据报文了。

如果在个阶段收到了 Configure-Request 报文，则又会返回到链路建立阶段。

## 网络终止阶段（Terminate）

PPP 能在任何时候终止链路。当载波丢失、认证失败、链路质量检测失败或管理员人为关闭链路等情况均会导致链路终止。

链路建立阶段可能通过交换 LCP 的链路终止报文来关闭链路，当链路关闭时，链路层会通知网络层做相应的操作，而且也会通过物理层强制关断链路。对于 NCP 协议，它是不能也没有必要去关闭 PPP 链路的。

#### 4.2.3.4 PPP 的 PAP 验证协议

##### PAP 验证过程概述

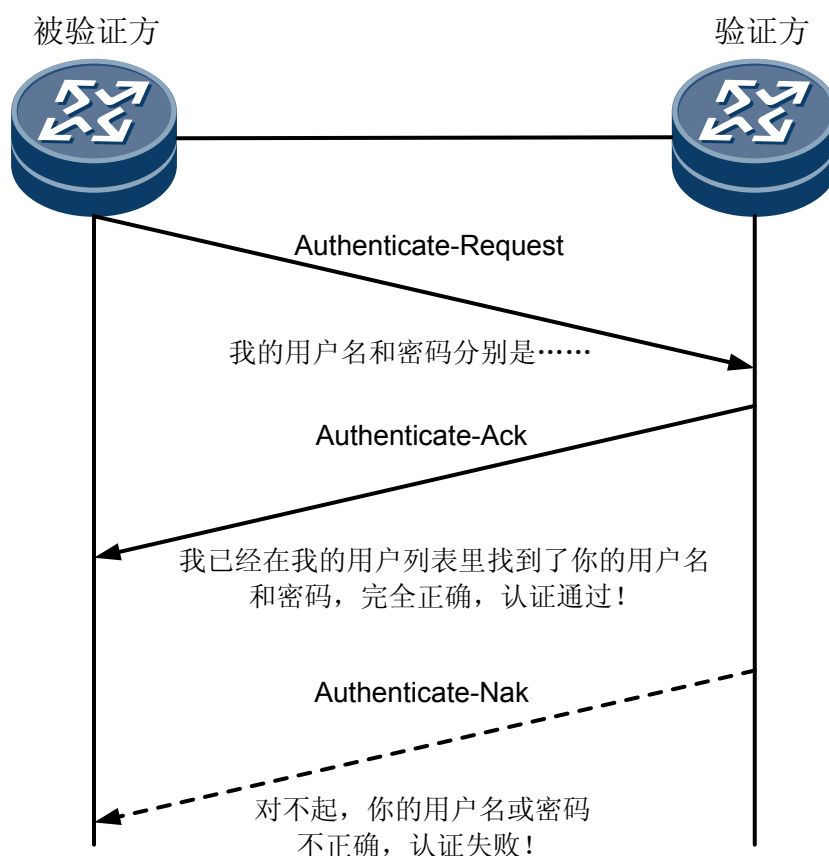
PAP 验证协议为两次握手验证，密码为明文。验证过程仅在链路初始建立阶段进行。

当链路建立阶段结束后，用户名和密码将由被验证方重复地在链路上发送给验证方，直到验证被通过或者链路连接终止。

如果必须在远端主机上使用明文密码进行模拟登录，这种验证方式是最合适的。

PAP 验证的过程如下图所示。

图 4-5 PAP 认证过程



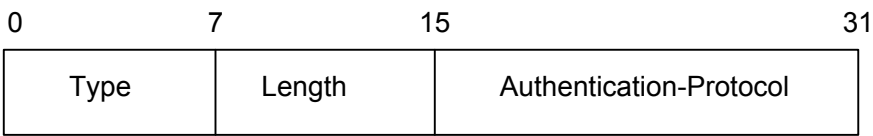
1. 被验证方把本地用户名和密码发送到验证方。
2. 验证方根据本地用户表查看是否有被验证方的用户名以及密码是否正确，然后返回不同的响应（接受或拒绝）。

PAP 不是一种安全的验证协议。当验证时，密码以明文方式在链路上发送，并且由于完成 PPP 链路建立后，被验证方会不停地在链路上反复发送用户名和密码，直到身份验证过程结束，所以不能防止攻击。

PAP 验证报文帧格式

- PAP 的配置参数选项帧格式  
协商 PAP 验证协议的配置参数选项帧格式如下图所示。

图 4-6 PAP 的配置参数选项帧格式



各字段的含义如下表所示。

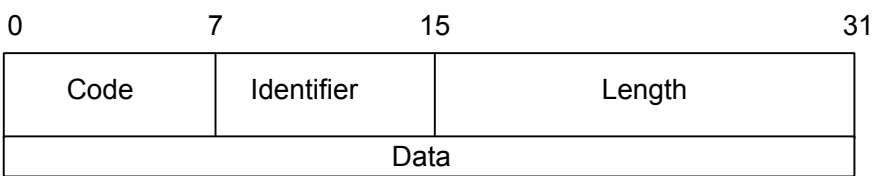
表 4-4 PAP 的配置参数选项各字段解释表

| 字段                      | 长度<br>(字节) | 含义                                                                                     |
|-------------------------|------------|----------------------------------------------------------------------------------------|
| Type                    | 1          | 当值是 0x03 时表示是验证协议。常用的协商类型请参见表 4-3。                                                     |
| Length                  | 1          | 此时固定值是 4，表示这个配置参数选项帧格式总长度是 4 个字节。                                                      |
| Authentication-Protocol | 2          | 当值是 0xC023 表示是 PAP 协议。常用的协议代码请参见表 4-1。在验证协议中，Data 字段的内容就是 Authentication-Protocol 的内容。 |

它和 PPP 报文、LCP 报文的关系如图 4-3 所示。

- PAP 数据报帧格式  
一个 PAP 数据报是封装在协议域为 C023 的 PPP 数据链路层帧的信息域中的。  
PPP 数据报文格式和 PAP 数据报文格式的关系请参见图 4-5。  
PAP 数据报的帧格式如下图所示。

图 4-7 PAP 数据报的帧格式



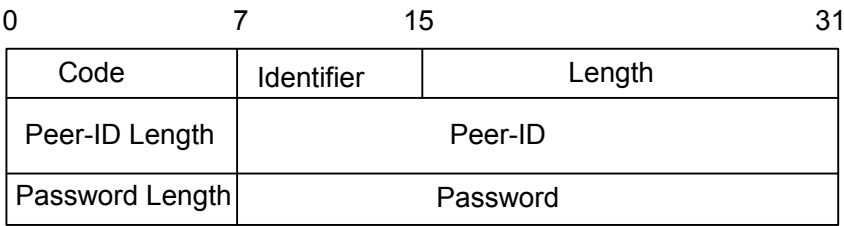
各字段的含义如下表所示。

表 4-5 PAP 数据报的帧格式各字段解释表

| 字段         | 长度（字节）  | 含义                                                                                                              |
|------------|---------|-----------------------------------------------------------------------------------------------------------------|
| Code       | 1       | 标识 PAP 数据报的类型。<br>● 1 表示是 Authenticate-Request 报文<br>● 2 表示是 Authenticate-Ack 报文<br>● 3 表示是 Authenticate-Nak 报文 |
| Identifier | 1       | 标识请求报文和应答报文的匹配。                                                                                                 |
| Length     | 2       | 表示包括 Code、Identifier、Length 和 Data 域在内的 PAP 报文长度。超出此长度的报文将被认为是填充字节并被丢弃。                                         |
| Data       | 0 或多个字节 | Data 域的帧由 Code 域来决定。                                                                                            |

- Authenticate-Request 报文  
验证-请求报文用于表示 PAP 验证的开始。链路的被验证方在验证阶段必须传输 code 值为 0x01 的 PAP 验证报文。验证报文必须被重复发送，直到收到了有效的回复报文，或计数器的值已满，此时应该终止链路连接。  
验证方只能等待被验证方发送验证-请求报文。当收到验证-请求报文后，必须根据实际情况回复不同的应答报文。  
验证-请求报文的帧格式如下图所示。

图 4-8 Authenticate-Request 报文格式



各字段的解释如下表所示。

表 4-6 Authenticate-Request 报文帧格式各字段解释表

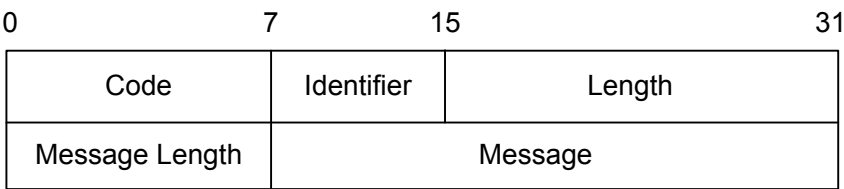
| 字段   | 长度（字节） | 含义                                                     |
|------|--------|--------------------------------------------------------|
| Code | 1      | 0x01 表示验证-请求报文。具体的 code 值含义请参见 <a href="#">表 4-5</a> 。 |



| 字段              | 长度（字节） | 含义                                    |
|-----------------|--------|---------------------------------------|
| Identifier      | 1      | 标识请求报文和应答报文的匹配。对于每个请求报文的应答报文，该域都必须不同。 |
| Length          | 2      | 表示该报文的总长度。                            |
| Peer-ID Length  | 1      | 标识 Peer-ID 域的长度。                      |
| Peer-ID         | 0 或多个  | 标识被验证方的名字。                            |
| Password Length | 1      | 标识 Password 域的长度。                     |
| Password        | 0 或多个  | 标识被验证的密码。                             |

- **Authenticate-Ack 和 Authenticate-Nak 报文帧格式**  
如果在验证-请求报文中的用户名和密码都能被验证方验证通过，验证方必须返回 Code 值是 2 的 Authenticate-Ack 报文，表示验证通过。  
如果验证-请求报文中用户名或密码有一项没有通过验证，验证方必须返回 Code 值是 3 的 Authenticate-Nak 报文，表示验证失败。  
Authenticate-Ack 和 Authenticate-Nak 报文的帧格式如下图所示。

图 4-9 Authenticate-Ack 和 Authenticate-Nak 报文的帧格式



各字段的含义如下表所示。

表 4-7 Authenticate-Ack 和 Authenticate-Nak 报文帧格式各字段解释表

| 字段             | 长度（字节） | 含义                                                                            |
|----------------|--------|-------------------------------------------------------------------------------|
| Code           | 1      | 0x02 表示 Authenticate-Ack 报文，0x03 表示 Authenticate-Nak 报文。具体的 code 值含义请参见表 4-5。 |
| Identifier     | 1      | 标识请求报文和应答报文的匹配。该域的值和引起该应答报文的 Authenticate-Request 报文一样。                       |
| Length         | 2      | 表示该报文的总长度。                                                                    |
| Message Length | 1      | 标识 Message 域的长度。                                                              |

| 字段      | 长度（字节） | 含义        |
|---------|--------|-----------|
| Message | 0 或多个  | 由报文的内容决定。 |

4.2.3.5 PPP 的 CHAP 验证协议

CHAP 验证过程概述

CHAP 验证协议为三次握手验证协议。它只在网络上传输用户名，而并不传输用户密码，因此安全性要比 PAP 高。

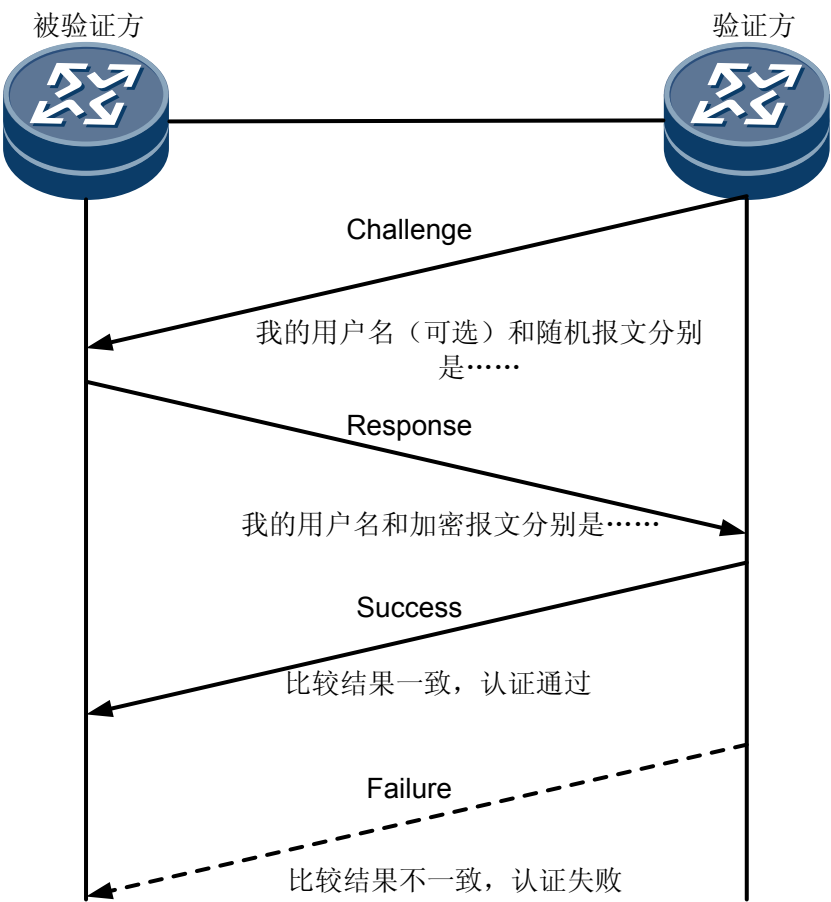
CHAP 协议是在链路建立的开始就完成的。在链路建立完成后的任何时间都可以重复发送进行再验证。

当链路建立阶段完成后，验证方发送一个“challenge”报文给被验证方。被验证方经过一次哈希算法后，给验证方返回一个值。

验证方把自己经过哈希算法生成的值和被验证方返回的值进行比较。如果两者匹配，那么验证通过。否则验证不通过，连接应该被终止。

CHAP 的验证过程如下图所示。

图 4-10 CHAP 的验证过程



CHAP 单向验证是指一端作为验证方，另一端作为被验证方。双向验证是单向验证的简单叠加，即两端都是既作为验证方又作为被验证方。在实际应用中一般只采用单向验证。

CHAP 单向验证过程分为两种情况：验证方配置了用户名和验证方没有配置用户名。推荐使用验证方配置用户名的方式，这样可以对验证方的用户名进行确认。

- 验证方配置了用户名的验证过程  
验证方配置了用户名的验证过程如下：
  - 验证方把随机产生的“质询（Challenge）”报文和本地主机名一起发送给被验证方。
  - 被验证方接收到报文后，根据验证方的用户名在本地用户列表中查找本地口令。根据查找到的口令和质询报文，通过 MD5 算法进行计算得出一个数值，并将计算得出的数值和自己的主机名发回验证方（Response）。
  - 验证方接收到 Response 后，根据其中携带的被验证方主机名，在本地用户表中查找被验证方口令字，找到匹配项后，利用质询报文和被验证方口令字，通过 MD5 算法进行计算得出一个数值，根据此数值与收到的 Response 的结果进行比较，然后返回不同的响应（接受或拒绝）。
- 验证方没有配置用户名  
验证方没有配置用户名时，验证方只把“质询”报文发送到被验证方。被验证方直接根据本地接口设置的口令和质询报文通过 MD5 算法计算得出一个数值，并将计算得出的数值和自己的主机名发回验证方。其他过程和验证方配置了用户名时相同。

CHAP 验证报文帧格式

- 配置参数选项帧格式  
协商 CHAP 协议的配置参数选项帧格式如下图所示。

图 4-11 CHAP 的配置参数选项帧格式

|           |        |                         |    |
|-----------|--------|-------------------------|----|
| 0         | 7      | 15                      | 31 |
| Type      | Length | Authentication-Protocol |    |
| Algorithm |        |                         |    |

各字段含义如下表所示。

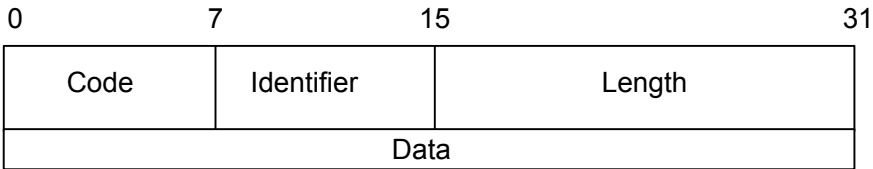
表 4-8 CHAP 的配置参数选项各字段解释表

| 字段     | 长度（字节） | 含义                            |
|--------|--------|-------------------------------|
| Type   | 1      | 0x03 表示是验证协议报文。具体值含义请参见表 4-3。 |
| Length | 1      | 此时固定是 5，表示该报文的总长度是 5 个字节。     |

| 字段                      | 长度（字节） | 含义                                                                                              |
|-------------------------|--------|-------------------------------------------------------------------------------------------------|
| Authentication-Protocol | 2      | 当值是 0xC223 时表示是 CHAP 验证协议。具体值的含义请参见表 4-1。                                                       |
| Algorithm               | 1      | 表示使用的一次哈希方法。 <ul style="list-style-type: none"><li>● 0 ~ 4: 不用，保留</li><li>● 5: MD5 算法</li></ul> |

- CHAP 数据报的帧格式  
确切来讲，CHAP 报文是封装在 PPP 数据链路层的 Information 域的。在 Information 域中 protocol 域的值固定是 0xC223，表示 CHAP 验证协议。  
CHAP 协议数据报格式如下图所示。

图 4-12 CHAP 数据报的帧格式



各字段含义如下表所示。

表 4-9 CHAP 数据报文帧格式各字段解释表

| 字段         | 长度（字节） | 含义                                                                                                                                                                  |
|------------|--------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Code       | 1      | 表示 CHAP 数据报文的类型。 <ul style="list-style-type: none"><li>● 1 表示 Challenge 报文</li><li>● 2 表示 Response 报文</li><li>● 3 表示 Success 报文</li><li>● 4 表示 Failure 报文</li></ul> |
| Identifier | 1      | 表示挑战报文、应答报文等之间的对应。                                                                                                                                                  |
| Length     | 2      | 表示包括 Code、Identifier、Length 和 Data 域在内的 CHAP 数据报文的长度。超出该长度值的字节应该被认为是数据链路层的填充字节，在接收时应该被忽略。                                                                           |
| Data       | 0 或多个  | 帧格式由 Code 域值决定。                                                                                                                                                     |

- Challenge 报文和 Response 报文  
Challenge 报文用来发起 CHAP 验证。验证方必须发送一个 Code 域值是 1 的 CHAP 数据报文，来表示是 Challenge 报文。其他附加挑战报文必须在收到有效的回应报文或计数器满后才可以发送。

挑战报文也可以在网络层协议阶段发送，以确认连接是否完整。

在验证阶段和网络层协议阶段，被验证方要等待验证方发送 Challenge 报文。只要接收到 Challenge 报文，被验证方必须返回一个 Code 域值是 2 的 CHAP 报文，表示是回应报文。

只要接收到回应报文，验证方就会把自己计算的值和返回值进行比较。根据比较的结果，验证方返回不同的回应报文。

挑战和回应报文的帧格式如下图所示。

图 4-13 Challenge 报文和 Response 报文帧格式

|            |            |        |    |
|------------|------------|--------|----|
| 0          | 7          | 15     | 31 |
| Code       | Identifier | Length |    |
| Value-size | Value      | Name   |    |

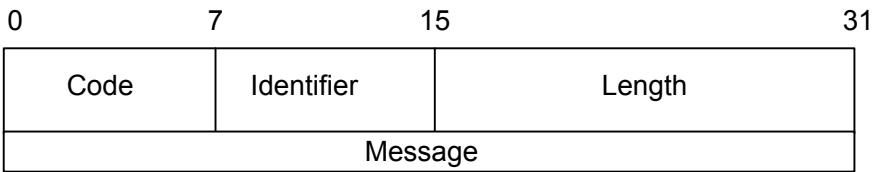
各字段的含义如下表所示。

表 4-10 Challenge 报文和 Response 报文的帧格式各字段解释表

| 字段         | 长度（字节） | 含义                                                 |
|------------|--------|----------------------------------------------------|
| Code       | 1      | ● 1 表示是 Challenge 报文<br>● 2 表示是回应报文                |
| Identifier | 1      | 它标识挑战报文和回应报文的对应关系。                                 |
| Length     | 2      | 表示该报文的总长度。                                         |
| Value-size | 1      | 表示 Value 域的长度。                                     |
| Value      | 1 或多个  | 挑战报文中，此域的内容是一些字节流。回应报文中此域的内容是挑战报文字节流经过一次哈希算法后得到的值。 |
| Name       | 1 或多个  | 该域的大小由 Length 域决定。                                 |

- Success 报文和 Failure 报文  
如果验证方接收到的值和自己计算的值相同，验证方必须要返回一个 Code 域值是 3 的 CHAP 报文，表示验证通过。  
如果验证方接收到的值和自己计算出的值不同，验证方必须返回一个 Code 域值是 4 的 CHAP 报文，表示验证失败。并且应该终止链路连接。  
验证通过或验证失败时的报文格式如下图所示。

图 4-14 Success 报文和 Failure 报文帧格式



各字段的含义如下表所示。

表 4-11 Success 报文和 Failure 报文帧格式各字段解释表

| 字段         | 长度（字节） | 含义                                                                            |
|------------|--------|-------------------------------------------------------------------------------|
| Code       | 1      | <ul style="list-style-type: none"><li>● 3 表示验证通过</li><li>● 4 表示验证失败</li></ul> |
| Identifier | 1      | 表示请求报文和该请求报文的响应报文的对应关系。                                                       |
| Length     | 2      | 表示该报文的总长度。                                                                    |
| Message    | 0 或多个  | 该域的长度由 Length 域的值决定。                                                          |

### 4.2.3.6 PPP 的报文压缩

USG 实现 PPP 协议上的对 IP、UDP、RTP 和 TCP 报文头的压缩，压缩方式有如下两种：

- TCP/IP 报文头压缩：需要 IPCP 协商，只压缩报文头部分。
- PPP 头压缩：只压缩 PPP 报文头部分。

### 4.2.3.7 MP 的实现方式

#### MP-Group

采用 MP-Group 方式配置 MP 时，直接将物理接口加入到 MP-Group 中即可。

#### 虚拟接口模板 VT（Virtual-Template）

采用虚拟接口模板配置 MP 时，又可以分为两种情况：

- 直接绑定
  - 直接将链路绑定到指定的虚拟接口模板，可以配置验证，也可以配置不验证。
    - 配置验证：接口通过验证后，绑定才能生效。
    - 配置不验证：当物理接口的 LCP 状态为 Up 后，绑定才能生效。
- 需要验证的绑定

系统可以根据两个参数进行验证：

- 用户名：PPP 链路进行 PAP 或 CHAP 验证时接收到的对端用户名。
- 终端标识符（Endpoint Discriminator）：唯一标识一台设备，是进行 LCP 协商时接收到的对端终端标识符。

可以只根据用户名或终端标识符验证，也可以同时根据用户名和终端标识符验证。拥有相同用户名或终端标识符的接口被捆绑到同一个虚拟接口模板。

如果本地按用户名来绑定 MP，需要配置本端验证对端（CHAP 或 PAP 方式）。

对于需要绑定在一起的接口，其验证方式的配置也必须相同。

实际使用中，也可以配置单向验证，即一端直接绑定到虚拟接口模板，另一端则通过用户名绑定到相应的虚拟接口模板。

#### 4.2.3.8 MP 的协商过程

MP 的协商较为特殊。MP 一些选项的协商是在 LCP 协商过程中完成的，如 MRRU、SSNHF、Endpoint Discriminator(终端描述符)等。

而决定不同通道是否需进行多链路捆绑有两个条件：只有两个链路的终端描述符和验证方式、用户完全相符时，才能对两个链路进行捆绑。这就意味着只有当验证完成后，才能真正完成 MP 的协商过程。MP 不会导致链路的拆除。

如果配置了 MP，两个链路不符合 MP 条件，则会建立一条新的 MP 通道，这同时也表明允许 MP 为单链路。MP 的捆绑是完全依照用户进行的，只有相同用户才能进行捆绑。如一端配置了 MP，另一端不支持或未配 MP，则建立起来的链路为非 MP 链路。

MP 的协商包括 LCP 协商和 NCP 协商两个过程：

- LCP 协商：两端首先进行 LCP 协商，除了协商一般的 LCP 参数外，还要验证对端接口是否也工作在 MP 方式下。如果两端工作方式不同，LCP 协商不成功。
- NCP 协商：根据 MP-Group 接口或指定虚拟接口模板的各项 NCP 参数（如 IP 地址等）进行 NCP 协商，物理接口配置的 NCP 参数不起作用。NCP 协商通过后，即可建立 MP 链路。

## 4.3 HDLC

HDLC 运行在点对点链路，配置简单。

### 4.3.1 介绍

#### 定义

HDLC（High-level Data Link Control procedure）是国际标准组织定义的多用途的数据链路控制协议，用于点到点和多点数据链路。HDLC 协议只支持同步方式。

#### 目的

当需要建立一种面向比特的链路层协议进行同步传输时，可以采用 HDLC 协议。

HDLC 最大的特点是不需要规定数据必须是字符集，对任何一种比特流，均可以实现透明的传输。

标准 HDLC 协议族中的协议运行于同步串行线路之上，如 DDN（Digital Data Network）。

HDLC 的地址字段和控制字段均是 8 比特，用来实现 HDLC 协议的各种控制信息，并标识是否是数据。

### 4.3.2 参考标准和协议

本特性的参考资料清单如下：

- RFC1549: PPP in HDLC Framing

### 4.3.3 可获得性

#### License 支持

不需要 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                           |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00 |

### 4.3.4 原理描述

#### 4.3.4.1 HDLC 的基本原理

数据链路控制协议一般可分为同步协议和异步协议两大类。其中同步协议以帧作为传输单位。采用帧作为传输单位，同步协议能更好地利用信道，也便于实现差错控制和流量控制等功能。

同步协议又可分为面向字符的同步协议、面向比特的同步协议及面向字节计数的同步协议。

作为面向比特的同步数据控制协议的典型，HDLC 具有以下几个特点。

- 全双工通信，不必等待确认可连续发送数据，有较高的数据链路传输效率。
- 所有帧均采用 CRC 校验，对信息帧进行顺序编号，可防止漏收或重收，传输可靠性高。
- 传输控制功能与处理功能分离，具有较大的灵活性和较完善的控制功能。
- 协议不依赖于任何一种字符编码集，数据报文可透明传输。
- 用于透明传输的“0 比特插入法”易于硬件实现。

数据链路控制协议着重对分段成物理块或包的数据进行逻辑传输。块或包也称为帧，由起始标志符引导并由终止标志符结束。



帧主要用于传输控制信息和响应信息。在 HDLC 中，所有面向比特的数据链路控制协议均采用统一的帧格式，不论是数据还是单独的控制信息均以帧为单位传输。

HDLC 协议的每个帧前后均有一标识码 01111110，用作帧的起始符、终止符或指示帧的同步。标识码不允许在帧的内部出现，以免引起歧义。

为保证标识码的唯一性但又兼顾帧内数据的透明性，可以采用“0 比特插入法”来解决。该法在发送端监视除标识码以外的所有字段，当发现有连续的 5 个“1”出现时，便在其后添加一个“0”，然后继续发送后继的比特流。在接收端，同样监视除标识码以外的所有字段。当连续发现 5 个“1”出现后，若其后一个比特为“0”，则自动删除它，以恢复原来的比特流；若发现连续 6 个“1”，则可能是插入的“0”发生错误，也可能是收到了终止标识码。

由于以上特点，目前网络设计及整机内部通信设计普遍使用 HDLC 数据链路控制协议。

#### 4.3.4.2 HDLC 的操作方式

##### HDLC 操作方式简介

HDLC 是通用的数据链路控制协议。当开始建立数据链路时，允许选用特定的操作方式。

所谓链路操作方式，通俗地讲就是以主节点方式操作，还是以从节点方式操作，或者是二者兼备。

在链路上起控制作用的节点称为主节点，其他受主节点控制的节点称为从节点。二者兼备的节点称为组合节点。

主节点负责对数据流进行组织，并且对数据上的差错进行恢复。由主节点发往从节点的帧称作命令帧，而由从节点返回主节点的帧称作响应帧。

连接多个节点的链路通常使用轮询技术，使其他节点轮询作为主节点，而在点到点链路中每个节点均可为主节点。在一个节点连接多条链路的情况下，该节点对于一些链路而言可能是主节点，而对另外一些链路而言有可能是从节点。

##### HDLC 常用的操作方式

HDLC 中常用的操作方式有以下 3 种。

- **正常响应方式**

正常响应方式 NRM (Normal Response Mode) 是一种非平衡数据链路操作方式，有时也称为非平衡正常响应方式。

该操作方式使用面向终端的点到点或一点到多点的链路。

在这种操作方式下，传输过程由主节点启动，从节点只有收到主节点某个命令帧后，才能作为响应向主节点传输信息。

响应信息可以由一个或多个帧组成，若信息由多个帧组成，则应指出哪一帧是最后一帧。

主节点负责管理整个链路，且具有轮询、选择从节点及向从节点发送命令的权利，同时也负责对超时、重发及各种恢复操作的控制。

- **异步响应方式**

异步响应方式 ARM (Asynchronous Response Mode) 也是一种非平衡数据链路操作方式。

与 NRM 不同的是，ARM 的传输过程由从节点启动。从节点主动发送给主节点的一个或一组帧。

在这种操作方式下，由从节点来控制超时和重发。

该方式对采用轮询方式的多节点链路来说是必不可少的。

- **异步平衡方式**

异步平衡方式 ABM (Asynchronous Balanced Mode) 是一种允许任何节点来启动传输的操作方式。

为了提高链路传输效率，节点之间在两个方向上都需要有较高的信息传输量。在这种操作方式下，任何时候任何节点都能启动传输操作，每个节点既可以作为主节点又可作为从节点。各个节点都有相同的一组协议，任何节点都可以发送或接受命令，也可以给出应答，并且各节点对差错恢复过程都负有相同的责任。

### 4.3.4.3 HDLC 的应用特点

#### 应用场合

- 就系统结构而言，HDLC 适用于点到点或点到多点式的结构；
- 就工作方式而言，HDLC 适用于半双工或全双工；
- 就传输方式而言，HDLC 只用于同步传输；
- 在传输速率方面考虑，HDLC 常用于中高速传输。

#### 传输效率

HDLC 开始发送一帧后，就要连续不断地发完该帧。HDLC 可以同时确认几个帧。

HDLC 中的每个帧含有地址字段。在多点的结构中，每个从节点只接收含有本节点地址的帧。

因此主节点在选中一个从节点并与之通信的同时，不用拆链，便可以选择其他的节点通信。即，可以同时与多个节点建立链路。

由于以上特点，HDLC 具有较高的传输效率。

#### 传输可靠性

HDLC 中所有的帧（包括响应帧）都有 FCS，I 帧按窗口序号顺序编号，传输可靠性比异步通信高。

#### 数据透明性

HDLC 采用“0 比特插入法”对数据进行透明传输，传输信息的比特组合模式无任何限制，处理简单。

#### 信息传输格式

HDLC 采用统一的帧格式来实现数据、命令和响应的传输，方便实现。

#### 链路控制

HDLC 利用改变一帧中的控制字段的编码格式来完成各种规定的链路操作功能，提供的是面向比特的传输功能。

## 4.4 VLAN

VLAN 可以把连接到同一个物理以太网的用户区分到逻辑 LAN 中，提高网络的管理能力。

### 4.4.1 介绍

#### 定义

VLAN（Virtual Local Area Network）即虚拟局域网，是指将一个物理的 LAN 在逻辑上划分成多个广播域的通信技术。

#### 目的

早期的局域网 LAN 技术基于总线型结构，它存在以下几个问题：

- 若某时刻有多个节点同时试图发送消息，则它们将产生冲突。
- 从任意节点发出的消息都会被发送到其他所有节点，形成广播。
- 所有主机共享一条传输通道，无法控制网络中的信息安全。

这种网络构成了一个冲突域，网络中计算机数量越多冲突越严重，网络效率越低。同时，该网络也是一个广播域，当网络中发送信息的计算机数量越多时，广播流量将会耗费大量带宽。

因此，传统网络不仅面临冲突域和广播域两大难题，而且无法保障传输信息的安全。

为了扩展传统 LAN，以接入更多计算机，同时避免冲突的恶化，出现了网桥和二层交换机，它们能有效隔离冲突域。

网桥和交换机采用交换方式将来自入端口的信息转发到出端口上，克服了共享介质上的访问冲突问题，从而将冲突域缩小到端口级。采用交换机进行组网，通过二层快速交换解决了冲突域问题，但是广播域和信息安全问题依旧存在。

#### 说明

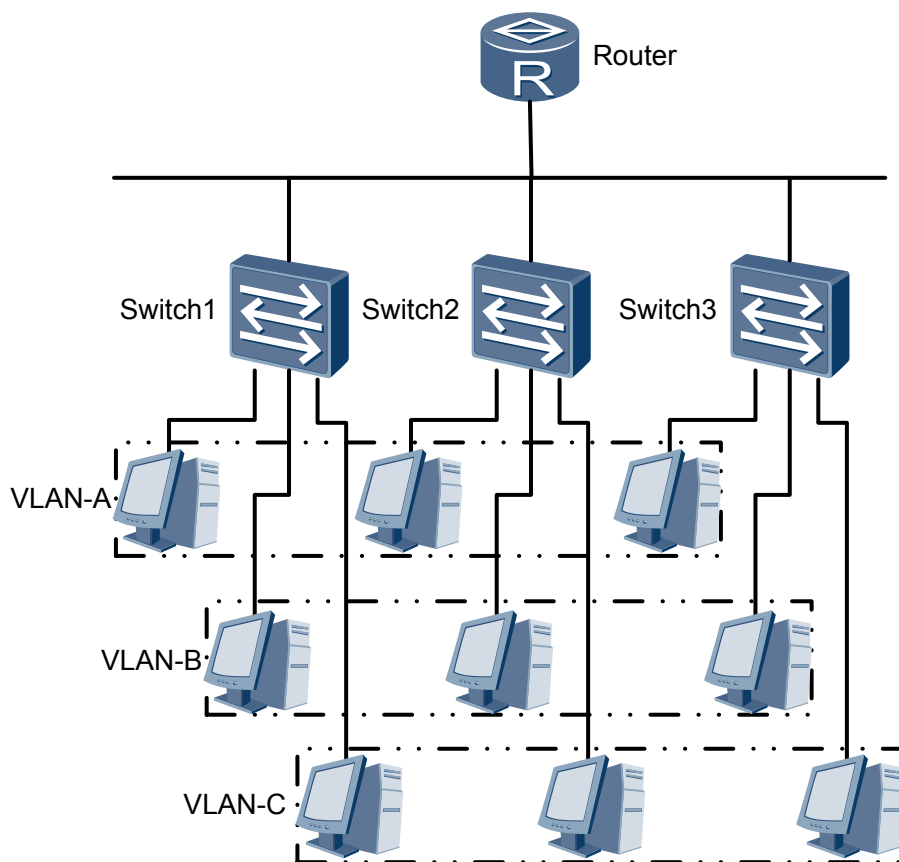
本手册中将二层局域网交换机简称为交换机。

为减少广播，需要在没有互访需求的主机之间进行隔离。路由器是基于 IP 地址信息来选择路由的，其连接两个网段时可以有效抑制广播报文的转发，但成本较高。因此人们设想在物理局域网上构建多个逻辑局域网，即 VLAN。

VLAN 技术将广播报文限制在一个 VLAN 内，使得同一 VLAN 内的主机之间可以直接相互通信，而不同 VLAN 间的主机则不能直接通信，提高了网络的安全性。

**图 4-15** 是一个典型的 VLAN 应用组网图。

图 4-15 VLAN 的典型应用示意图



三台交换机放置在写字楼的不同楼层。每台交换机分别连接 3 台计算机，他们分别属于 3 个不同的企业客户。若建立各自独立的 LAN，企业的网络投资成本将很高；若共用写字楼已有的 LAN，又会导致企业信息安全无法保证。此时采用 VLAN 技术，就可以实现各企业客户共享 LAN 设施，同时保证各自的网络信息安全。在图中，一个虚线框表示一个 VLAN。

## 4.4.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 3069: VLAN Aggregation for Efficient IP Address Allocation
- IEEE 802.1Q: IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks
- IEEE 802.1ad: IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks— Amendment 4
- IEEE 802.10: IEEE Standards for Local and Metropolitan Area Networks: Standard for Interoperable LAN/MAN Security
- YD/T 1260-2003: Technical and Testing Specification of Virtual LAN Based on Port

## 4.4.3 可获得性

License 支持

不需要 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

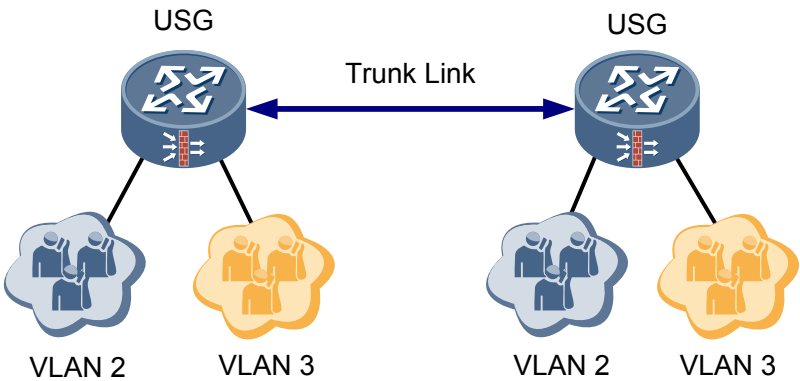
4.4.4 应用场景

4.4.4.1 VLAN 内通信

VLAN 内的主机可以直接通信，当相同 VLAN 内的主机不在同一地域（经过多个设备）时，可以通过 Trunk 链路实现。

如图 4-16 所示，公司业务发展存在多个部门，部门之间需要跨越不同的 USG；同时公司每个部门属于不同的 VLAN。此时，可以通过 Trunk 链路，实现不同部门的业务数据隔离，同一部门内业务数据的互通。

图 4-16 VLAN Trunk 组网图



4.4.4.2 VLAN 间通信

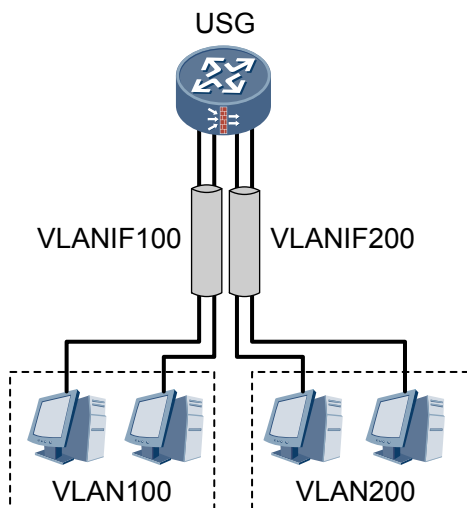
不同 VLAN 间的通信，可以通过 VLAN 接口或者以太网子接口实现。

## 通过 VLAN 接口实现 VLAN 间通信

VLAN 接口在功能上与普通三层物理接口基本相同，可实现配置 IP 地址等多种三层特性，实现不同 VLAN 间的数据通信。

如图 4-17 所示，USG 下面有 2 个部门（分别属于 VLAN100 和 200）。当不同部门需要互访时，可以在 USG 上为每个 VLAN 配置一个 VLAN 接口，实现 VLAN100、VLAN200 之间的互访。

图 4-17 VLAN 接口组网图



注意事项：

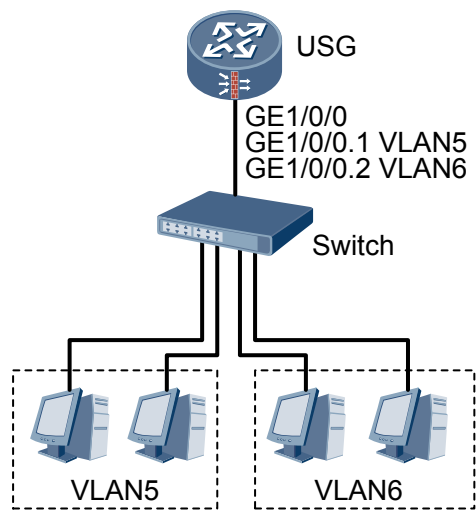
- USG 连接 PC 的接口为二层以太网接口，同时该接口需要加入到对应的 VLAN。
- 该组网方式会占用 USG 上多个接口。

## 通过以太网子接口实现 VLAN 间通信

通过 VLAN 接口可以实现不同 VLAN 间的通信，但是会占用多个物理接口。为解决这个问题，可以使用以太网子接口。

如图 4-18 所示，USG 下面有 2 个部门（分别属于 VLAN5 和 6）。如果不同部门需要互访时，可以在一个物理接口上配置多个子接口，这些子接口分别对应不同 VLAN，这样只需连接一个物理接口就可实现不同 VLAN 之间的数据通信。

图 4-18 以太网子接口组网图



- 注意事项：
- 为实现 VLAN5 和 VLAN6 间的通信，需要在 USG 与交换机相连的以太网接口上创建 2 个子接口与 VLAN5 和 VLAN6 分别对应。
  - 在子接口上配置 802.1Q 封装和 IP 地址。
  - 将交换机与 USG 相连的以太网口类型改为 Trunk 或 Hybrid，允许 VLAN5 和 VLAN6 的报文通过。

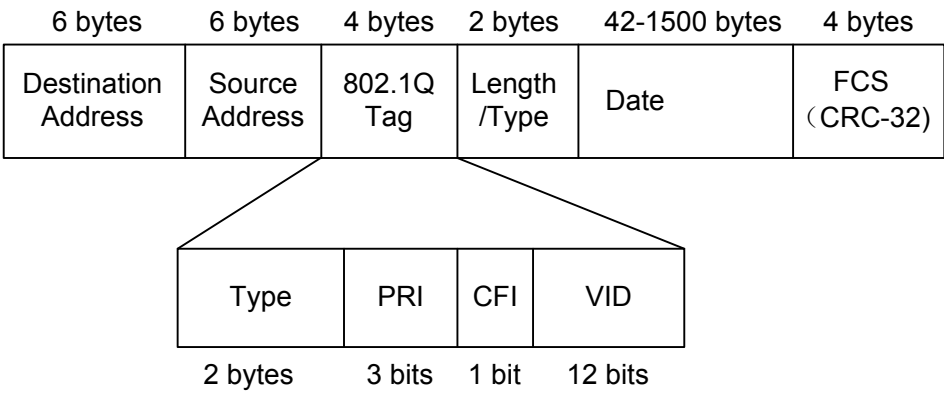
4.4.5 原理描述

4.4.5.1 VLAN 基本概念

VLAN 的帧格式

IEEE 802.1Q 标准对传统 Ethernet 帧格式进行了修改，在源 MAC 地址字段和协议类型字段之间加入 4 字节的 802.1Q Tag，形成了 VLAN 的帧格式。具体如图 4-19 所示。

图 4-19 基于 802.1Q 的 VLAN 帧格式



802.1Q Tag 包含 4 个字段，其含义如下：

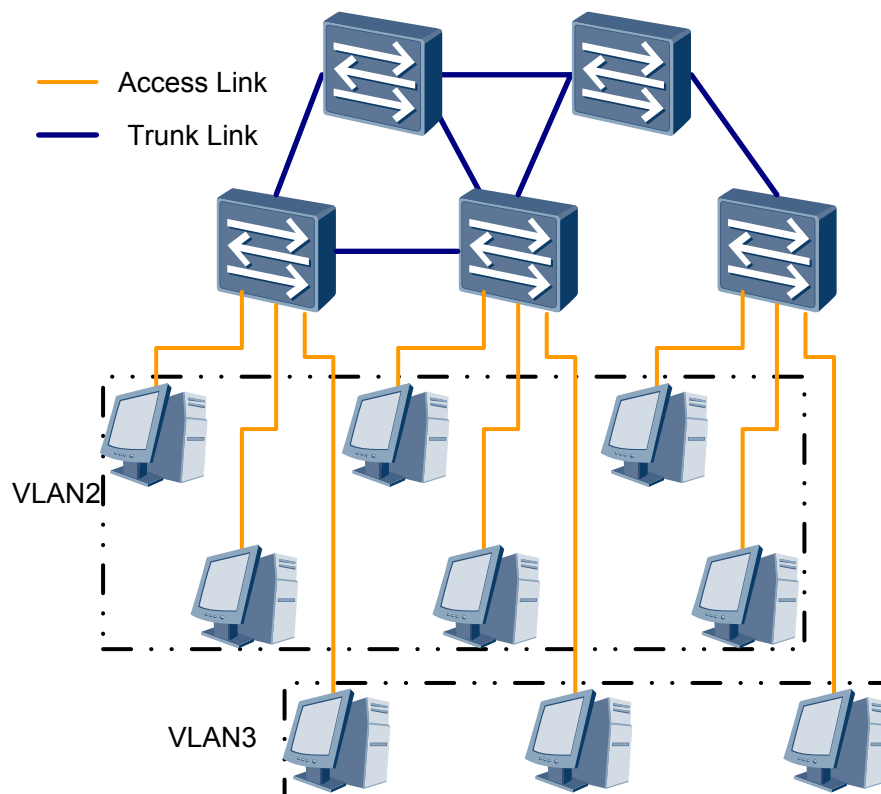
- Type  
长度为 2 字节，表示帧类型。取值为 0x8100 时表示 802.1Q Tag 帧。如果不支持 802.1Q 的设备收到这样的帧，会将其丢弃。
- PRI  
Priority，长度为 3 比特，表示帧的优先级，取值范围为 0 ~ 7，值越大优先级越高。用于当设备阻塞时，优先发送优先级高的数据帧。
- CFI  
Canonical Format Indicator，长度为 1 比特，表示 MAC 地址是否是经典格式。
- VID  
VLAN ID，长度为 12 比特，表示该帧所属的 VLAN。

## 链路类型

VLAN 内的链路包括：

- 接入链路（Access Link）：连接用户主机和设备的链路为接入链路。如图 4-20 所示，图中 PC 机和设备之间的链路都是接入链路。接入链路上通过的帧为不带 Tag 的以太网帧。
- 干道链路（Trunk Link）：连接设备和设备的链路称为干道链路。如图 4-20 所示，图中设备之间的链路都是干道链路。干道链路上通过的帧一般为带 Tag 的 VLAN 帧。

图 4-20 链路类型示意图





## 端口类型

在 802.1Q 中定义 VLAN 帧后，设备的有些端口可以识别 VLAN 帧，有些端口则不能识别 VLAN 帧。根据对 VLAN 帧的识别情况，将端口分以下几类：

- Access 端口

Access 端口是设备上用来连接用户主机的端口，它只能连接接入链路。有如下特点：

- 只允许 VLAN ID 与端口的 PVID（端口缺省的 VLAN ID）相同的 VLAN 通过该端口。
- 如果该端口收到的对端设备发送的帧是 untagged（不带 VLAN 标签），设备将强制加上该端口的 PVID。
- Access 端口发往对端设备的以太网帧永远是 untagged 的帧。

- Trunk 端口

Trunk 端口是设备上用来和其他设备连接的端口，它只能连接干道链路。有如下特点：

- 允许多个 VLAN 的帧（带 Tag）通过。
- 如果从 Trunk 端口发送的帧带 Tag，且 Tag 与 PVID 相同，则设备会剥掉该帧中的 Tag。仅在这种情况下，Trunk 端口发送的帧不带 Tag。

- Hybrid 端口

Hybrid 端口是设备上既可以连接用户主机，又可以连接其他设备的端口。Hybrid 端口既可以连接接入链路又可以连接干道链路。Hybrid 端口允许多个 VLAN 的帧通过，并可以在出端口方向将某些 VLAN 帧的 Tag 剥掉。

## VLAN 的划分

VLAN 的划分主要分为以下几种方式：

- 基于端口

根据设备的端口号来划分 VLAN。用户主机所属的 VLAN 由用户主机所连的网络设备端口所属的 VLAN 决定。其优点是定义成员简单，缺点是成员移动需重新配置 VLAN。

- 基于 MAC 地址

根据用户主机网卡的 MAC 地址来划分 VLAN。其优点是成员移动方便，缺点是需要预先定义所有成员。

- 基于网络层协议

根据网络层协议来划分 VLAN。例如将运行 IP 的用户主机划分为一个 VLAN，将运行 IPX 的用户主机划分为另一个 VLAN。

目前，最常用的 VLAN 划分方式为基于端口划分的方式。USG 只支持基于端口的 VLAN 划分。

### 4.4.5.2 VLAN 通信原理

## VLAN 基本通信原理

为了提高设备对数据帧的处理效率，设备内部的数据帧一律都带有 VLAN Tag，以便设备对这些数据帧以统一的方式处理。

根据端口类型的不同，设备对数据帧的处理过程也不同。具体如表 4-12 所示。

表 4-12 不同类型端口处理帧的过程

| 端口类型      | 接收帧处理过程                                                                                                                                                                                                                                                                                   | 发送帧处理过程                                                                                                                                                                 |
|-----------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Access 端口 | <ol style="list-style-type: none"><li>判断数据帧的 VLAN Tag:<ul style="list-style-type: none"><li>无 Tag，则添加本端口 PVID Tag，进行下一步处理。</li><li>有 Tag，若 Tag 与 PVID Tag 相同，则允许该 VLAN 帧进入，否则丢弃。</li></ul></li><li>二层设备根据数据帧的目的 MAC 地址和 VLAN ID 查找 VLAN 配置信息，决定从哪个端口把帧发送出去。</li></ol>                     | 先剥离帧的 PVID Tag，然后再发送。                                                                                                                                                   |
| Trunk 端口  | <ol style="list-style-type: none"><li>判断数据帧 VLAN Tag:<ul style="list-style-type: none"><li>无 Tag，且端口允许带 PVID Tag 的帧通过，则添加本端口 PVID Tag，进行下一步处理，否则丢弃。</li><li>有 Tag，判断该端口是否允许该 VLAN 帧进入，允许则进行下一步处理，否则丢弃。</li></ul></li><li>二层设备根据帧的目的 MAC 地址和 VLAN ID 查找 VLAN 配置信息，决定从哪个端口把帧发送出去。</li></ol> | <p>判断 VLAN 在本端口的属性：</p> <ul style="list-style-type: none"><li>如果是端口 PVID Tag，先剥离帧的 PVID Tag，再发送。</li><li>如果不是端口 PVID Tag，且是端口允许通过的 VLAN ID，则直接发送。否则丢弃。</li></ul>        |
| Hybrid 端口 | <ol style="list-style-type: none"><li>判断数据帧 VLAN Tag:<ul style="list-style-type: none"><li>无 Tag，且端口允许带 PVID Tag 的帧通过，则添加本端口 PVID Tag，进行下一步处理，否则丢弃。</li><li>有 Tag，则判断端口是否允许该 VLAN 帧进入，允许则进行下一步处理，否则丢弃。</li></ul></li><li>二层设备根据帧的目的 MAC 地址和 VLAN ID 查找 VLAN 配置信息，决定从哪个端口把帧发送出去。</li></ol> | <p>判断 VLAN 在本端口的属性：</p> <p>判断端口是否允许该 VLAN 帧通过</p> <ul style="list-style-type: none"><li>允许，则判断是否为允许不带 Tag 的帧，如果是，则先剥离帧的 VLAN Tag，再发送；否则直接发送。</li><li>不允许，则丢弃。</li></ul> |

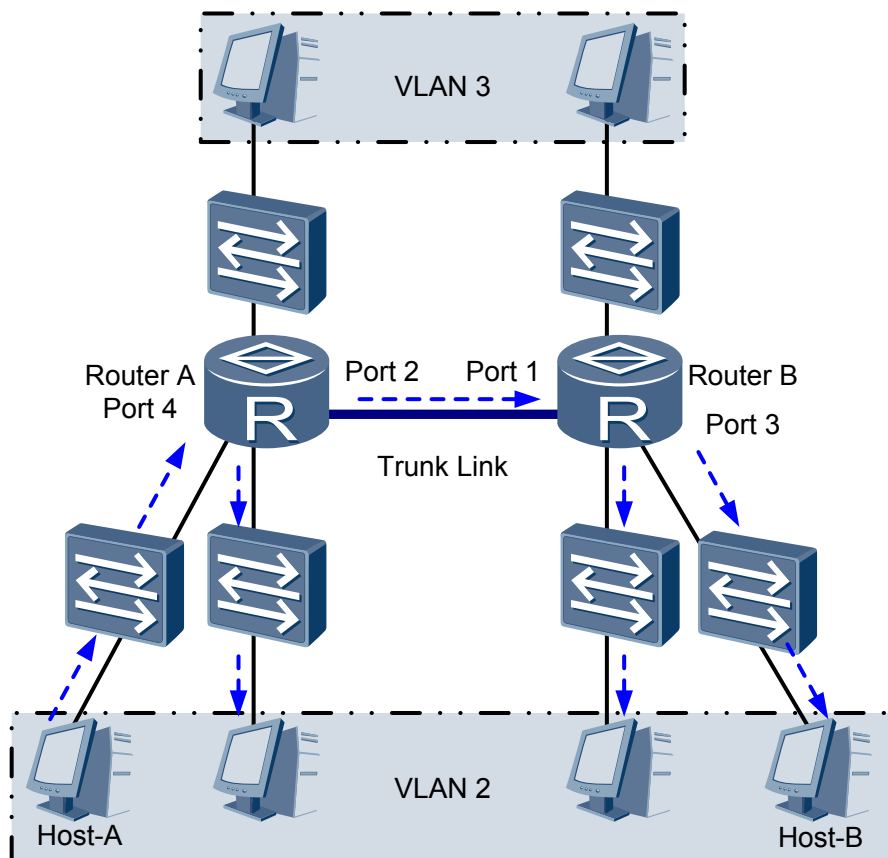
## 相同 VLAN 内主机跨越设备进行通信

有时属于同一个 VLAN 的用户主机被连接在不同的设备上。当 VLAN 跨越设备时，就需要设备间的端口能够同时识别和发送跨越设备的 VLAN 报文。这时，需要用到 Trunk Link 技术。

Trunk Link 有两个作用：

- 中继作用  
把 VLAN 报文透传到互联的设备。
- 干线作用  
一条 Trunk Link 上传输多个 VLAN 的报文。

图 4-21 Trunk Link 通信方式示意图



如图 4-21 所示的网络中，为了让 Router A 和 Router B 之间的链路既支持 VLAN2 内的用户通信又支持 VLAN3 内的用户通信，需要配置连接端口同时属于两个 VLAN。即应配置 Router A 的以太网端口 Port 2 和 Router B 的以太网端口 Port 1 既属于 VLAN2 也属于 VLAN3。

当用户主机 Host-A 发送数据给用户主机 Host-B 时，数据帧的发送过程如下：

1. 数据帧首先到达 Router A 的端口 Port 4。
2. 端口 Port 4 给数据帧加上 Tag，Tag 的 VID 字段填入该端口所属的 VLAN 的编号 2。
3. 对于广播或未知单播报文：Router A 将帧发送到本设备上除 Port 4 外的所有属于 VLAN2 的端口；对于已知单播报文，则根据 MAC 地址表转发。
4. 端口 Port 2 将帧转发到 Router B 上。
5. 对于广播或未知单播报文：Router B 收到帧后，会根据帧中的 Tag 识别出该帧属于 VLAN2，于是将该帧发给本设备上除 Port 1 外所有属于 VLAN2 的端口；对于已知单播报文，则根据 MAC 地址表转发。

6. 端口 Port 3 将数据帧发送给主机 Host-B。

VLAN3 内的主机通信过程同理，此处略。

4.5 MSTP

在交换式局域网络中开启 MSTP 功能，可有效的避免广播风暴。

4.5.1 介绍

定义

IEEE 于 2002 年发布的 802.1S 标准定义了多生成树协议 MSTP（Multiple Spanning Tree Protocol），该协议兼容 STP（Spanning Tree Protocol）和 RSTP（Rapid Spanning Tree Protocol）。

目的

在二层交换网络中，一旦网络存在环路就会造成报文在环路内不断复制和循环，产生广播风暴。在广播风暴的情况下，所有的有效带宽都被广播风暴占用，网络将变得不可用。

在这种环境下，IEEE 802.1D 标准中提出的 STP 协议应运而生。STP 通过选择性地阻塞网络冗余链路，将网络修剪成树状，达到消除环路的目的，同时具备链路备份功能。

但是 STP 拓扑收敛速度慢，即使是边缘端口也必须等待 30 秒的时间延迟才能迁移到转发状态。IEEE 802.1W 描述的 RSTP 协议在 STP 基础上进行了改进，实现了网络拓扑快速收敛。

但 RSTP 和 STP 还存在同一个缺陷：由于局域网内所有的 VLAN 共享一棵生成树，因此无法在 VLAN 间实现数据流量的负载均衡，还有可能造成部分 VLAN 的报文无法转发。

MSTP 兼容 STP 和 RSTP，并且可以弥补 STP 和 RSTP 的缺陷。MSTP 既可以快速收敛，同时还提供了数据转发的多个冗余路径，在数据转发过程中实现 VLAN 数据的负载均衡。

这样所有 VLAN 内部可以互通，同时不同 VLAN 的报文沿不同的路径转发，实现了负载分担。

如表 4-13 所示，三种生成树协议的比较。

表 4-13 三种生成树协议的比较

| 生成树协议 | 特点                                 |
|-------|------------------------------------|
| STP   | 形成一棵无环路的树：解决广播风暴并实现冗余备份。           |
| RSTP  | 形成一棵无环路的树：解决广播风暴并实现冗余备份。<br>收敛速度快。 |

| 生成树协议 | 特点                                                                              |
|-------|---------------------------------------------------------------------------------|
| MSTP  | 形成一棵无环路的树：解决广播风暴并实现冗余备份。<br>收敛速度快。<br>多颗生成树在 VLAN 间实现负载均衡，不同 VLAN 的流量按照不同的路径转发。 |

## 4.5.2 参考标准和协议

本特性的参考资料清单如下：

- IEEE 802.1D:  
IEEE Standard for:
  - Local and metropolitan area networks
  - Virtual Bridged Local Area Networks
- IEEE 802.1S:  
IEEE Standard for:
  - Local and metropolitan area networks
  - Virtual Bridged Local Area Networks
- IEEE 802.1W:  
IEEE Standard for:
  - Local and metropolitan area networks
  - Common specifications

## 4.5.3 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R002C00 |

## 4.5.4 原理描述

4.5.4.1 MSTP 的基本原理

MSTP 通过设置 VLAN 映射表（即 VLAN 和生成树实例的对应关系表），把 VLAN 和生成树实例联系起来。同时它把一个交换网络划分成多个域，每个域内形成多棵生成树实例，生成树实例之间彼此独立。MSTP 将环路网络修剪成为一个无环的树型网络，避免报文在环路网络中的增生和无限循环，同时还提供了数据转发的多个冗余路径，在数据转发过程中实现 VLAN 数据的负载均衡。

4.5.4.2 MSTP 报文格式

MSTP 使用多生成树桥协议数据单元 MST BPDU（Multiple Spanning Tree Bridge Protocol Data Unit）作为生成树计算的依据。

MST BPDU 报文用来计算生成树的拓扑、维护网络拓扑以及传达拓扑变化记录。MST BPDU 报文结构如图 4-22 所示。

图 4-22 MST BPDU 报文结构

|                                                |                              | Octet                      |
|------------------------------------------------|------------------------------|----------------------------|
| Protocol Identifier                            |                              | 1-2                        |
| Protocol Version Identifier                    |                              | 3                          |
| BPDU Type                                      |                              | 4                          |
| CIST Flags                                     |                              | 5                          |
| CIST Root Identifier                           |                              | 6-13                       |
| CIST External Path Cost                        |                              | 14-17                      |
| CIST Regional Root Identifier                  |                              | 18-25                      |
| CIST Port Identifier                           |                              | 26-27                      |
| Message Age                                    |                              | 28-29                      |
| Max Age                                        |                              | 30-31                      |
| Hello Time                                     |                              | 32-33                      |
| Forward Delay                                  |                              | 34-35                      |
| Version 1 Length=0                             |                              | 36                         |
| MST<br>special<br>fields                       | Version 3 Length             | 37-38                      |
|                                                | MST Configuration Identifier | 39-89                      |
|                                                | CIST Internal Root Path Cost | 90-93                      |
|                                                | CIST Bridge Identifier       | 94-101                     |
|                                                | CIST Remaining Hops          | 102                        |
| MSTI Configuration Messages<br>(may be absent) |                              | 103-39+Version<br>3 Length |

无论是域内的 MST BPDU 还是域间的，前 35 个字节和 RST BPDU 相同。

从第 36 个字节开始是 MSTP 专有字段。最后的 MSTI 配置信息字段由若干 MSTI 配置信息组连缀而成。

MST BPDU 中的主要信息如图 4-22 所示。

表 4-14 MST BPDU 中主要信息说明

| 字段内容                          | 说明                                                                                                                                                                                                                                                                            |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Protocol Identifier           | 协议标识符。                                                                                                                                                                                                                                                                        |
| Protocol Version Identifier   | 协议版本标识符，STP 为 0，RSTP 为 2，MSTP 为 3。                                                                                                                                                                                                                                            |
| BPDU Type                     | BPDU 类型，MSTP 为 0x02。<br><ul style="list-style-type: none"> <li>● 0x00: STP 的 Configuration BPDU</li> <li>● 0x80: STP 的 TCN BPDU (Topology Change Notification BPDU)</li> <li>● 0x02: RST BPDU (Rapid Spanning-Tree BPDU) 或者 MST BPDU (Multiple Spanning-Tree BPDU)</li> </ul> |
| CIST Flags                    | CIST 标志字段。                                                                                                                                                                                                                                                                    |
| CIST Root Identifier          | CIST 的总根交换机 ID。                                                                                                                                                                                                                                                               |
| CIST External Path Cost       | CIST 外部路径开销指从本交换机所属的 MST 域到 CIST 根交换机的累计路径开销。CIST 外部路径开销根据链路带宽计算。                                                                                                                                                                                                             |
| CIST Regional Root Identifier | CIST 的域根交换机 ID，即 IST Master 的 ID。如果总根在这个域内，那么域根交换机 ID 就是总根交换机 ID。                                                                                                                                                                                                             |
| CIST Port Identifier          | 本端口在 IST 中的指定端口 ID。                                                                                                                                                                                                                                                           |
| Message Age                   | BPDU 报文的生存期。                                                                                                                                                                                                                                                                  |
| Max Age                       | BPDU 报文的最大生存期，超时则认为到根交换机的链路故障。                                                                                                                                                                                                                                                |
| Hello Time                    | Hello 定时器，缺省为 2 秒。                                                                                                                                                                                                                                                            |
| Forward Delay                 | Forward Delay 定时器，缺省为 15 秒。                                                                                                                                                                                                                                                   |
| Version 1 Length              | Version1 BPDU 的长度，值固定为 0。                                                                                                                                                                                                                                                     |
| Version 3 Length              | Version3 BPDU 的长度。                                                                                                                                                                                                                                                            |
| MST Configuration Identifier  | MST 配置标识，表示 MST 域的标签信息，包含 4 个字段，如图 4-23 所示。只有 MST Configuration Identifier 中的四个字段完全相同的，并且互联的交换机，才属于同一个域。字段说明如表 4-15 所示。                                                                                                                                                       |
| CIST Internal Root Path Cost  | CIST 内部路径开销指从本端口到 IST Master 交换机的累计路径开销。CIST 内部路径开销根据链路带宽计算。                                                                                                                                                                                                                  |
| CIST Bridge Identifier        | CIST 的指定交换机 ID。                                                                                                                                                                                                                                                               |
| CIST Remaining Hops           | BPDU 报文在 CIST 中的剩余跳数。                                                                                                                                                                                                                                                         |

| 字段内容                                       | 说明                                                                                                                        |
|--------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|
| MSTI Configuration Messages(may be absent) | MSTI 配置信息。每个 MSTI 的配置信息占 16 bytes，如果有 n 个 MSTI 就占用 n×16bytes。单个 MSTI Configuration Messages 的结构如图 4-24 所示，字段说明如表 4-16 所示。 |

图 4-23 MST Configuration Identifier

|                                          |             |
|------------------------------------------|-------------|
| Configuration Identifier Format Selector | Octet<br>39 |
| Configuration Name                       | 40-71       |
| Revision Level                           | 72-73       |
| Configuration Digest                     | 74-89       |

表 4-15 MST Configuration Identifier 字段说明

| 字段内容                                     | 说明                                           |
|------------------------------------------|----------------------------------------------|
| Configuration Identifier Format Selector | 固定为 0。                                       |
| Configuration Name                       | “域名”，32 字节长字符串。                              |
| Revision Level                           | 2 字节非负整数。                                    |
| Configuration Digest                     | 利用 HMAC-MD5 算法将域中 VLAN 和实例的映射关系加密成 16 字节的摘要。 |

图 4-24 MSTI Configuration Messages

|                               |            |
|-------------------------------|------------|
| MSTI Flags                    | Octet<br>1 |
| MSTI Regional Root Identifier | 2-9        |
| MSTI Internal Root Path Cost  | 10-13      |
| MSTI Bridge Priority          | 14         |
| MSTI Port Priority            | 15         |
| MSTI Remaining Hops           | 16         |



表 4-16 MSTI Configuration Messages 字段说明

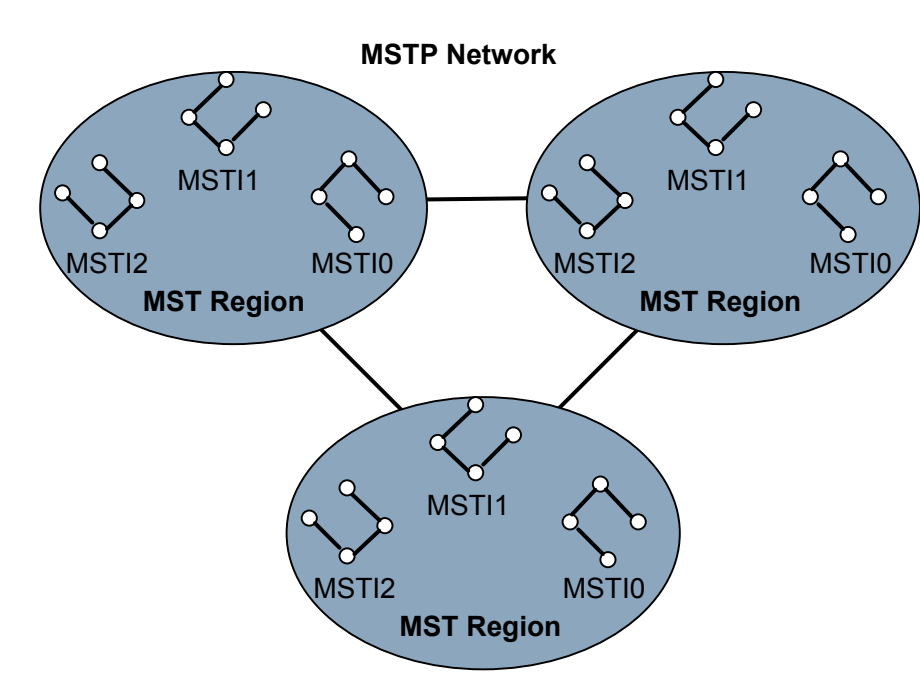
| 字段内容                          | 说明                                                       |
|-------------------------------|----------------------------------------------------------|
| MSTI Flags                    | MSTI 标志。                                                 |
| MSTI Regional Root Identifier | MSTI 域根交换机 ID。                                           |
| MSTI Internal Root Path Cost  | MSTI 内部路径开销指从本端口到 MSTI 域根交换机的累计路径开销。MSTI 内部路径开销根据链路带宽计算。 |
| MSTI Bridge Priority          | 本交换机在 MSTI 中的指定交换机的优先级。                                  |
| MSTI Port Priority            | 本交换机在 MSTI 中的指定端口的优先级。                                   |
| MSTI Remaining Hops           | BPDU 报文在 MSTI 中的剩余跳数。                                    |

4.5.4.3 MSTP 的基本概念

MSTP 的网络层次

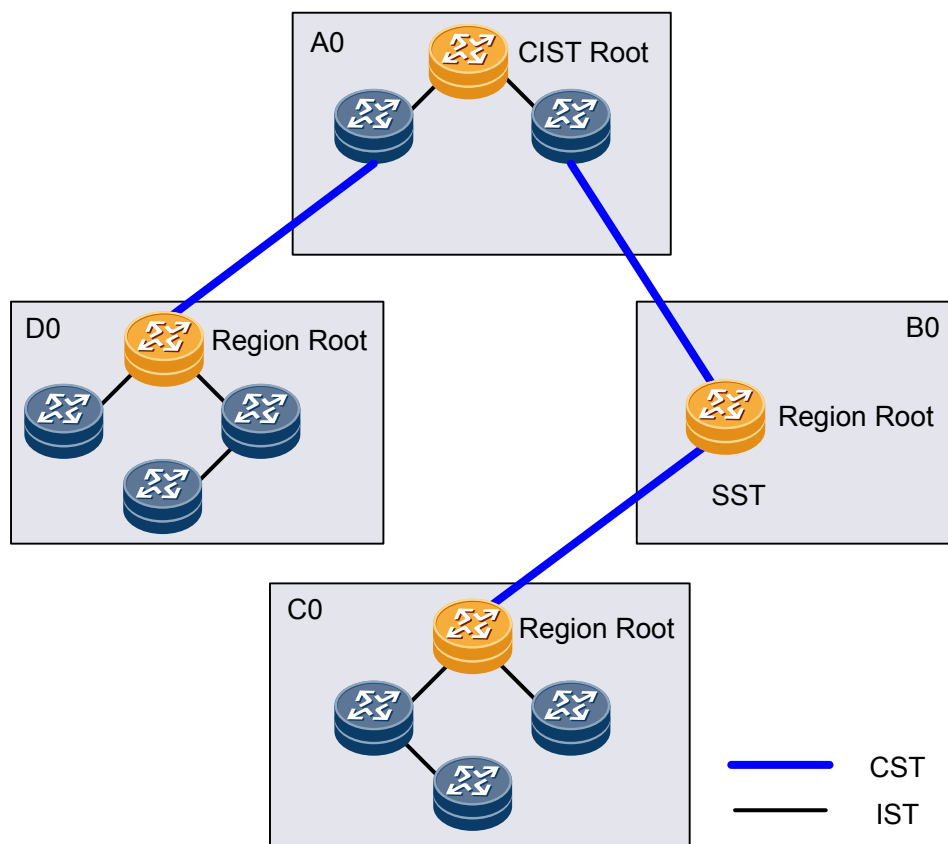
- MSTP 网络可分为以下层次：
- MSTP 网络
  - 多生成树域 MST Region（Multiple Spanning Tree Region）
  - 多生成树实例 MSTI（Multiple Spanning Tree Instance）
- 一个 MSTP 网络中可能包含一个或多个 MST Region。如图 4-25 所示，MSTP 网络中包含 3 个 MST Region，每个 MST Region 中包含 3 个 MSTI。MSTI 由运行 STP/RSTP/MSTP 的交换机组成，MSTI 是所有运行 STP/RSTP/MSTP 的交换机经 MSTP 协议计算后形成的树状网络。

图 4-25 MSTP 的网络层次示意图



## MSTP 网络

图 4-26 MSTP 网络基本概念示意图



如图 4-26 所示，MSTP 网络层包含以下基本概念。

- 总根  
总根是 CIST（Common and Internal Spanning Tree）的根交换机（CIST Root）。如图 4-26，总根 CIST Root 在 A0 中。
- CST  
公共生成树 CST（Common Spanning Tree）是连接交换网络内所有 MST 域的一棵生成树。  
如果把每个 MST 域看作是一个节点，CST 就是这些节点通过 STP 或 RSTP 协议计算生成的一棵生成树。  
如图 4-26 中较粗的蓝色线条连接各个域构成 CST。
- IST  
内部生成树 IST（Internal Spanning Tree）是各 MST 域内的一棵生成树。  
IST 是一个特殊的 MSTI，MSTI 的 ID 为 0 即表示 IST，通常称为 MSTI0。  
IST 是 CIST 在 MST 域中的一个片段。  
如图 4-26 中较细的线条在域中连接该域的所有交换机构成 IST。
- CIST  
公共和内部生成树 CIST 是通过 STP 或 RSTP 协议计算生成的，连接一个交换网络内所有交换机的单生成树。

如图 4-26 所示，所有 MST 域的 IST 加上 CST 就构成一棵完整的生成树，即 CIST。

- SST  
构成单生成树 SST（Single Spanning Tree）有两种情况：
  - 运行 STP 或 RSTP 的交换机只能属于一个生成树。
  - MST 域中只有一个交换机，这个交换机构成单生成树。如图 4-26 所示，B0 中的交换机就是一棵单生成树。

MST 域

图 4-27 MST 域的基本概念示意图

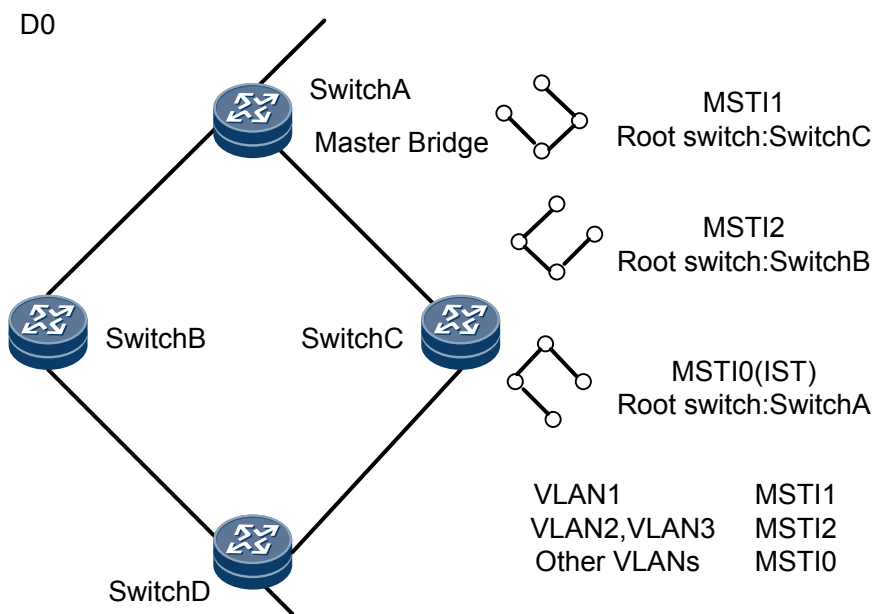


图 4-27 为图 4-26 中截取的一个 MST 域 D0，MST 域中包含以下基本概念。

- MST 域  
MST 域是多生成树域，由局域网中的多台交换机以及它们之间的网段构成。一个局域网可以存在多个 MST 域，各 MST 域之间在物理上直接或间接相连。用户可以通过 MSTP 配置命令把多台交换机划分在同一个 MST 域内。  
如图 4-27 所示的 MST 域 D0 中由交换机 SwitchA、SwitchB、SwitchC 和 SwitchD 构成，域中有 3 个 MSTI。
- VLAN 映射表  
VLAN 映射表是 MST 域的属性，它描述了 VLAN 和 MSTI 之间的映射关系。  
如图 4-27 中，MST 域 D0 的 VLAN 映射表是：
  - VLAN 1 映射到 MSTI1
  - VLAN2 和 VLAN3 映射到 MSTI2
  - 其余 VLAN 映射到 MSTI0
- 域根  
域根（Regional Root）分为 IST 域根和 MSTI 域根。

IST 域根如图 4-26 所示，在 B0、C0 和 D0 中，IST 生成树中距离总根最近的交换机是 IST 域根。

MSTI 域根是每个多生成树实例的树根。如图 4-27 所示，域中不同的 MSTI 有各自的域根。

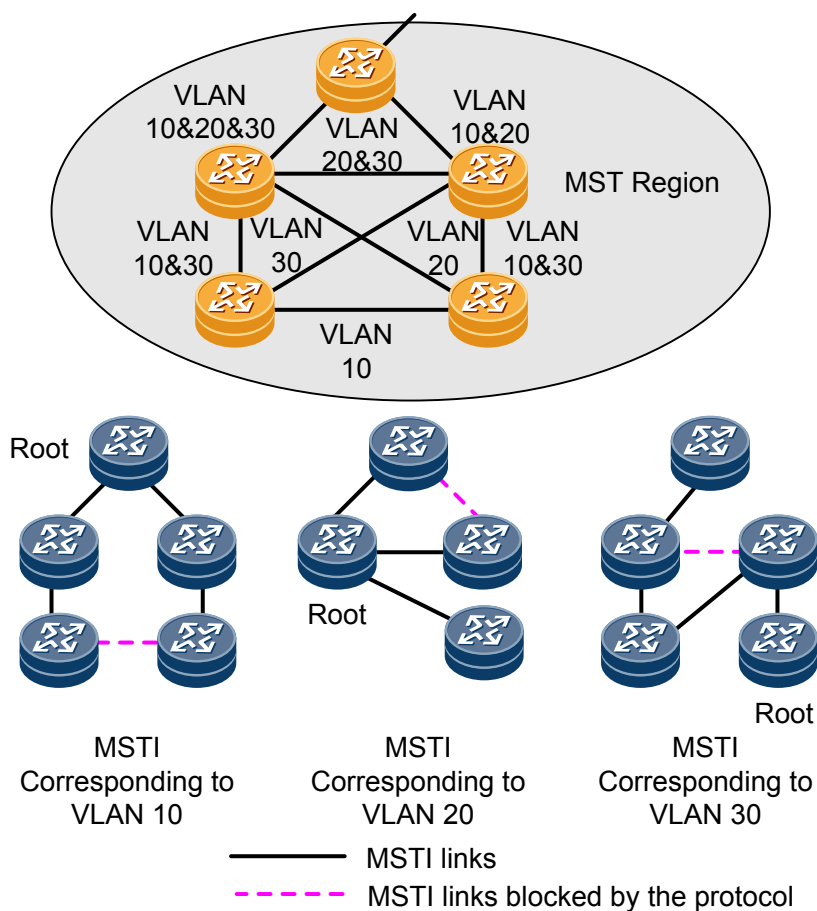
- 主桥

主桥（Master Bridge）也就是 IST Master，它是域内距离总根最近的交换机。如图 4-27 中的 SwitchA。

如果总根在 MST 域中，则总根为该域的主桥。

## MSTI

图 4-28 MSTI 的基本概念示意图



一个 MST 域内可以生成多棵生成树，每棵生成树都称为一个 MSTI。

MSTI 之间彼此独立，MSTI 可以与一个或者多个 VLAN 对应。但一个 VLAN 只能与一个 MSTI 对应。

如图 4-28 中，MST 域 VLAN10、VLAN20 和 VLAN30 分别对应一个 MSTI。

## 端口角色

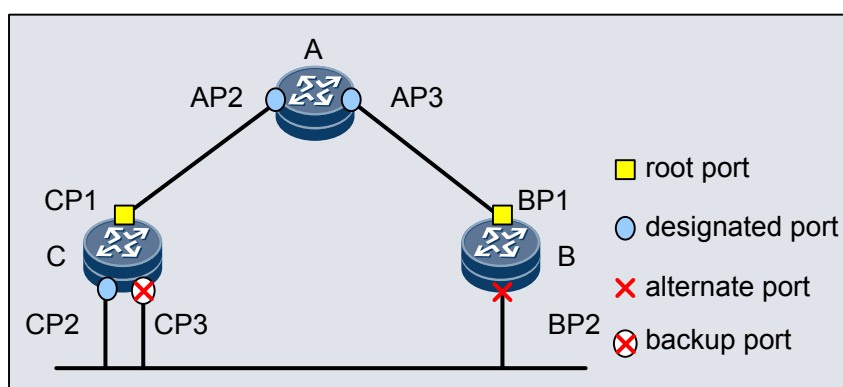
MSTP 中的端口角色主要有：

- 根端口（Root port）
- 指定端口（Designated port）
- Alternate 端口（Alternate port）
- Backup 端口（Backup port）
- Master 端口（Master port）
- 域边缘端口（Region edge port）
- 边缘端口（Edge port）

除边缘端口外，其他端口角色都参与 MSTP 的计算过程。

同一端口在不同的生成树实例中可以担任不同的角色。

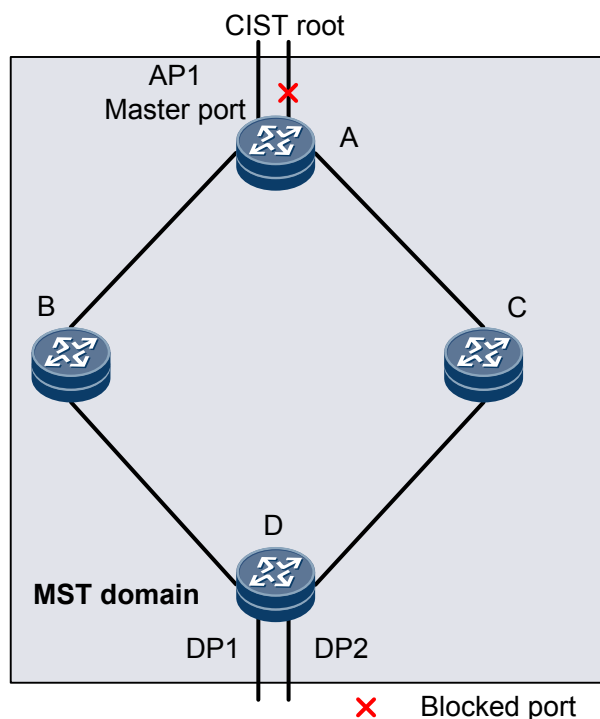
图 4-29 根端口、指定端口、Alternate 端口和 Backup 端口示意图



- 根端口  
在非根交换机上，距离根交换机最近的端口称为根端口。根交换机没有根端口。  
根端口负责向树根方向转发数据。  
图 4-29 中，A 为根交换机，CP1 为 C 的根端口，BP1 为 B 的根端口。
- 指定端口  
对一台交换机而言，它的指定端口是向下游交换机转发 BPDU 报文的端口。  
图 4-29 中 AP2 和 AP3 为 A 的指定端口，CP2 为 C 的指定端口。
- 边缘端口  
如果指定端口位于整个域的边缘，不再与任何交换机连接，这种端口叫做边缘端口。  
边缘端口一般与用户终端设备直接连接。
- Alternate 端口  
从学习 BPDU 的过程来看，Alternate 端口就是由于学习到由其它交换机发送的 BPDU 而被阻塞的端口。从转发用户流量来看，Alternate 端口提供了从指定交换机到根交换机的一条备份路径。  
Alternate 端口是根端口的备份端口，如果根端口被阻塞后，Alternate 端口将成为新的根端口。  
图 4-29 中 BP2 为 Alternate 端口。
- Backup 端口  
当同一台交换机的两个端口互相连接时就存在一个环路，此时交换机会将其中一个端口阻塞，Backup 端口就是被阻塞的那个端口。图 4-29 中 CP3 为 Backup 端口。

从发送 BPDU 来看，Backup 端口就是由于学习到自己发送的 BPDU 而被阻塞的端口。从转发用户流量来看，Backup 端口，作为指定端口的备份，提供了一条从根交换机到叶节点的备份通路。

图 4-30 Master 端口和域边缘端口示意图



- Master 端口

Master 端口是 MST 域和总根相连的所有路径中最短路径上的端口，它是交换机上连接 MST 域到总根的端口。

Master 端口是域中的报文去往总根的必经之路。

Master 端口是特殊域边缘端口，Master 端口在 IST/CIST 上的角色是 Root Port，在其它各实例上的角色都是 Master。

如图 4-30 所示，交换机 A、B、C、D 和它们之间的链路构成一个 MST 域，A 交换机的端口 AP1 在域内的所有端口中到总根的路径开销最小，所以 AP1 为 Master 端口。

- 域边缘端口

域边缘端口是指位于 MST 域的边缘并连接其它 MST 域或 SST 的端口。

进行 MSTP 计算时，域边缘端口在 MSTI 上的角色和 CIST 实例的角色保持一致。即如果边缘端口在 CIST 实例上的角色是 Master 端口（连接域到总根的端口），则它在域内所有 MSTI 上的角色也是 Master 端口。

例如如图 4-30 中，MST 域内的 AP1、DP1 和 DP2 都和其它域直接相连，它们都是本 MST 域的边缘端口。

域边缘端口在生成树实例上的角色与在 CIST 的角色保持一致。如图 4-30 所示，AP1 是域边缘端口，它在 CIST 上的角色是 Master 端口，则 AP1 在 MST 域内所有生成树实例上的角色都是 Master 端口。

## MSTP 的端口状态

MSTP 的端口状态分为三种，如表 4-17 所示。

表 4-17 MSTP 的端口状态

| 状态         | 说明                                                                                                          |
|------------|-------------------------------------------------------------------------------------------------------------|
| Forwarding | 在这种状态下，端口既转发用户流量又接收/发送 BPDU 报文。                                                                             |
| Learning   | 这是一种过渡状态。在 Learning 下，交换机会根据收到的用户流量,构建 MAC 地址表，但不转发用户流量，所以叫做学习“状态”。<br>Learning 状态的端口接收/发送 BPDU 报文，不转发用户流量。 |
| Discarding | Discarding 状态的端口只接收 BPDU 报文。                                                                                |

表 4-18 显示了各种端口角色能够具有的端口状态。

表 4-18 各种端口角色具有的端口状态

| 端口状态       | 根端口/<br>Master 端口 | 指定端口 | 域边缘端口 | Alternate<br>端口 | Backup 端口 |
|------------|-------------------|------|-------|-----------------|-----------|
| Forwarding | Yes               | Yes  | Yes   | No              | No        |
| Learning   | Yes               | Yes  | Yes   | No              | No        |
| Discarding | Yes               | Yes  | Yes   | Yes             | Yes       |

Yes: 表示端口支持的状态。

No: 表示端口不支持的状态。

### 4.5.4.4 MSTP 的快速收敛机制

#### Proposal/Agreement 机制

端口的快速迁移机制即为 Proposal/Agreement 机制，使用两种协议报文：

- proposal 报文：指定端口请求快速迁移的报文。
- agreement 报文：同意对端进行快速迁移的报文。

Proposal/Agreement 机制包括以下两种方式：

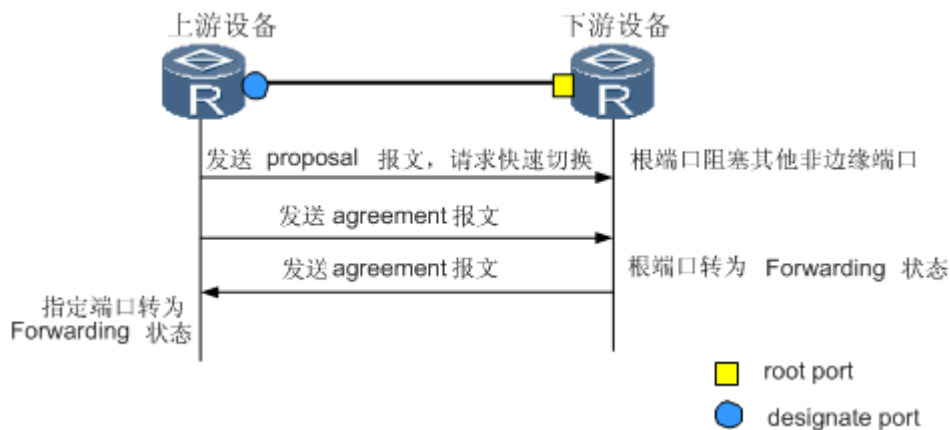
- 增强方式

- 普通方式

USG 支持增强方式和普通方式。

## 增强方式

图 4-31 MSTP 指定端口快速迁移机制（增强方式）

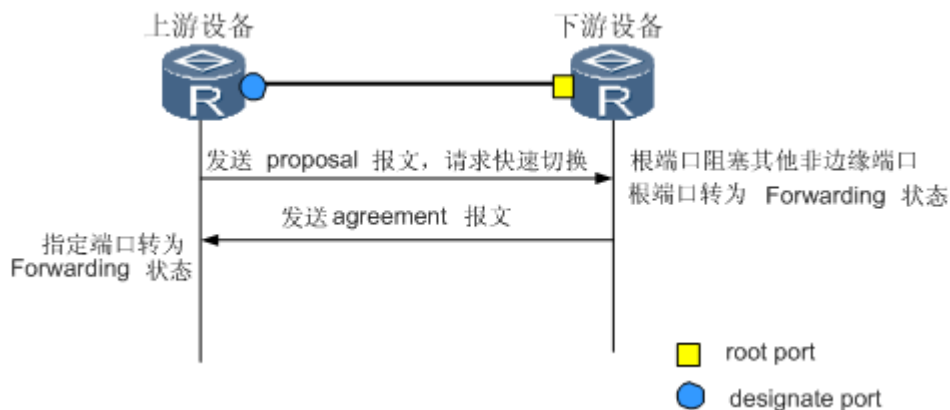


如图 4-31 所示，在 MSTP 中，Proposal/Agreement 机制工作过程如下。

1. 上游设备发送 Proposal 报文，请求进行快速迁移。下游设备接收到后，把与上游设备相连的端口设置为根端口，并阻塞所有非边缘端口。
2. 上游设备继续发送 Agreement 报文。下游设备接收到后，根端口转为 Forwarding 状态。
3. 下游设备回应 Agreement 报文。上游设备接收到后，把与下游设备相连的端口设置为指定端口，指定端口进入 Forwarding 状态。

## 普通方式

图 4-32 MSTP 指定端口快速迁移机制（普通方式）



如图 4-32 所示，在 MSTP 中，Proposal/Agreement 机制工作过程如下。



1. 上游设备发送 Proposal 报文，请求进行快速迁移。下游设备接收到后，把与上游设备相连的端口设置根端口，并阻塞所有非边缘端口，根端口转为 Forwarding 状态。
2. 下游设备回应 Agreement 报文，上游设备接收到后。把与下游设备相连的端口设置为指定端口，指定端口进入 Forwarding 状态。

## 根端口快速切换机制

如果交换机的根端口失效，该交换机的 Alternate 端口立即替代失效的根端口，成为交换机新的根端口，进入 Forwarding 状态。

## 边缘端口

由于边缘端口不参与 MSTP 运算，所以可以直接转到 Forwarding 状态，并不经历时延，缩短了网络收敛时间。

在没有配置 BPDU 保护的情况下，边缘端口一旦收到 BPDU，就丧失了边缘端口属性，成为了普通的 STP 端口。

### 4.5.4.5 MSTP 的保护功能

## BPDU 保护

在交换机上，通常将直接与用户终端（如 PC 机）或文件服务器等非交换机设备相连的端口配置为边缘端口。

正常情况下，这些端口不会收到 BPDU。如果有人伪造 BPDU 恶意攻击交换机，当这些端口接收到 BPDU 时，交换机会自动将这些端口设置为非边缘端口，并重新进行生成树计算，从而引起网络震荡。

MSTP 提供 BPDU 保护功能来防止这种攻击。交换机上启动 BPDU 保护功能后，如果边缘端口收到了 BPDU，交换机将关闭这些端口，同时通知网管系统。被关闭的端口只能由网络管理人员手动恢复。

## 根保护

由于维护人员的错误配置或网络中的恶意攻击，网络中的合法根交换机有可能会收到优先级更高的 BPDU，使得合法根交换机失去根交换机的地位，引起网络拓扑结构的错误变动。

这种不合法的变动，会导致原来应该通过高速链路的流量被牵引到低速链路上，造成网络拥塞。

为了防止以上情况发生，交换机提供 Root 保护功能。Root 保护功能通过维持指定端口的角色来保护根交换机的地位。配置了 Root 保护功能的端口，在所有实例上的端口角色都保持为指定端口。

当端口收到优先级更高的 BPDU 时，端口的角色不会变为非指定端口，而是进入侦听状态（Listening），不再转发报文。经过足够长的时间，如果端口一直没有再收到优先级较高的 BPDU，端口会恢复到原来的正常状态。

## 环路保护

在交换机上，根端口和其他阻塞端口状态是依靠不断接收来自上游交换机的 BPDU 来维持的。

当由于链路拥塞或者单向链路故障导致这些端口收不到来自上游交换机的 BPDU 时，此时交换机会重新选择根端口。原先的根端口会转变为指定端口，而原先的阻塞端口会迁移到转发状态，从而造成交换网络中可能产生环路。

环路保护功能会抑制这种环路的产生。在启动了环路保护功能后，如果根端口收不到来自上游的 BPDU 时，根端口会被设置进入阻塞状态；而阻塞端口则会一直保持在阻塞状态，不转发报文，从而不会在网络中形成环路。

## 4.6 PPPoE

PPPoE 利用以太网将大量主机组成网络，通过一个远端接入设备连入因特网，并对接入的每个主机实现控制、计费功能。

### 4.6.1 介绍

#### 定义

PPPoE 是 Point-to-Point Protocol over Ethernet 的简称，它利用以太网将多个主机组成网络，通过一个远端接入设备接入因特网，并对接入的每个主机实现控制、计费功能。极高的性价比使 PPPoE 被广泛于包括小区组网等一系列应用中。

#### 目的

当多个用户同时通过接入服务器获取服务时：

- 用户希望接入成本低，不要或者很少改变配置即可接入成功。以太网无疑是最好的组网方式。
- 服务提供商希望通过同一个接入服务器连接到远程站点上的多个主机，同时要求服务器能提供与使用 PPP 拨号上网类似的访问控制功能和支付功能。

PPP（Point-to-Point）应用虽然很广泛，但是不能应用于以太网，因此提出了 PPPoE 技术。PPPoE 是对 PPP 的扩展，它可以使 PPP 协议应用于以太网。

PPPoE 提供通过简单桥接的接入服务器把一个网络的多个主机连接到远程接入服务器的功能。

### 4.6.2 参考标准和协议

本特性的参考资料清单如下：

- RFC2516: A Method for Transmitting PPP Over Ethernet (PPPoE)
- RFC1661: The Point-to-Point Protocol (PPP)

### 4.6.3 可获得性

## License 支持

不需要 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 4.6.4 原理描述

### 4.6.4.1 PPPoE 基本原理

PPPoE 提供了在以太网上建立 PPPoE 会话以及封装 PPP 数据报的方法。PPPoE 要求在通信双方之间建立的是点到点关系，而不是在以太网和其他多路访问环境中所出现的多点关系。

在 PPPoE 通信过程中，每一个主机使用自己的 PPP 协议栈，但用户使用的还是熟悉的用户接口。

PPPoE 支持的访问控制、计费支付以及服务类型（Type of Service）等功能都是基于每一个用户的。

PPPoE 分为以下两个阶段：

- Discovery 阶段：地址发现阶段
- PPPoE Session 阶段：PPPoE 会话阶段

为了在以太网上建立点到点连接，每一个 PPPoE 会话必须知道通信对方的 MAC 地址，并建立一个唯一的会话标识符。PPPoE 通过地址发现协议查找对方的 MAC 地址。

当某个主机希望发起一个 PPPoE 会话时，它首先通过地址发现协议来确定对方的 MAC 地址并建立起一个 PPPoE 会话标识符 Session\_ID。

虽然 PPP 定义的是端到端的对等关系，地址发现却是一种客户端-服务器关系。在地址发现的过程中，主机作为客户端，发现某个作为服务器的接入访问集中器 AC（Access Concentrator）的以太网地址。

根据网络的拓扑结构，可能主机与不只一个访问集中器通信。Discovery 阶段允许主机发现所有的访问集中器，并从中选择一个进行通信。

当 Discovery 阶段成功完成之后，主机和访问集中器两者都具备了在以太网上建立点到点连接所需的所有信息。

在开始建立一个 PPPoE 会话之前，Discovery 阶段一直保持无状态（stateless）。

一旦开始建立 PPPoE 会话，主机和作为接入服务器的访问集中器都必须为一个 PPP 虚拟接口分配资源。

PPPoE 会话建立成功后，主机和接入服务器便可以通信了。

#### 4.6.4.2 Discovery 阶段

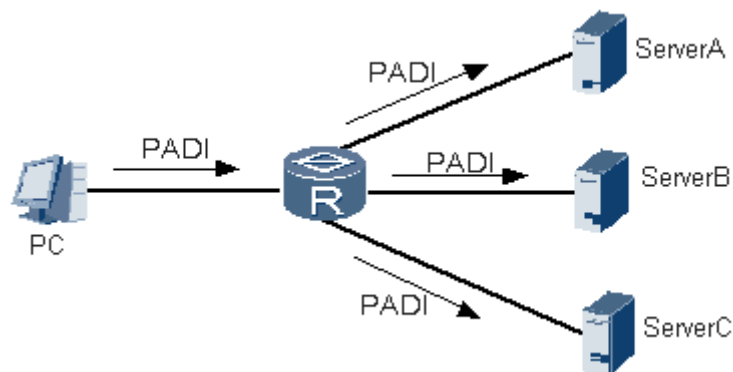
当主机开始通过 PPPoE 接入服务器时，它必须先识别接入端的以太网 MAC 地址，建立 PPPoE 的 Session\_ID。这就是 Discovery 阶段的目的。

Discovery 阶段由四个过程组成。完成之后通信双方都会知道 PPPoE 的 Session\_ID 以及对方以太网地址，它们共同确定了唯一的 PPPoE 会话。

Discovery 阶段的四个过程如下。

1. 主机在本以太网内广播一个 PADI（PPPoE Active Discovery Initial）报文，在此报文中包含主机想要得到的服务类型信息。

图 4-33 主机以广播形式发送 PADI 报文

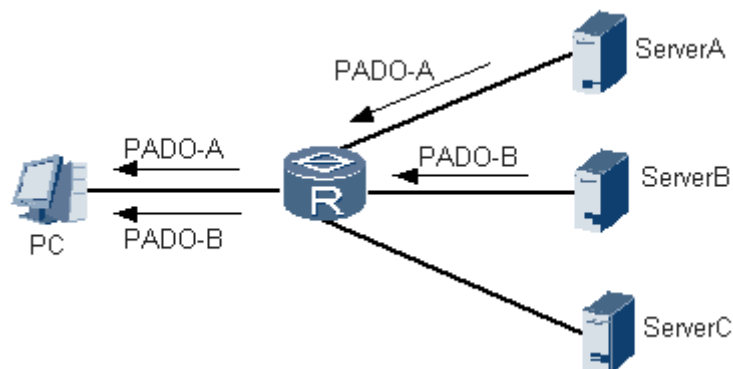


##### 说明

- 如果在 PPPoE 的服务器端配置 service-name，client 将发送 Discovery 阶段的 PADI 报文给服务器端请求建立连接。
  - 如果该 PADI 报文中包含有不为空的服务名称时，服务器端将用配置的 service-name 和该报文中的 service-name 进行完全匹配性检测。如果两者完全相同，服务器端提供后续服务，否则，服务器端不提供服务。
  - 以上是两者的 service-name 都不为空时的情况。但如果两者中有一个 service-name 为空，就不进行此项检测，直接按照原来的程序执行。
2. 以太网内的所有服务器收到这个 PADI 报文后，将其中请求的服务与自己能提供的服务进行比较，可以提供此服务的服务器发回 PADO（PPPoE Active Discovery Offer）报文。

如图 4-34 中，ServerA 和 ServerB 都可以提供服务，所以都会向主机发回 PADO 报文。

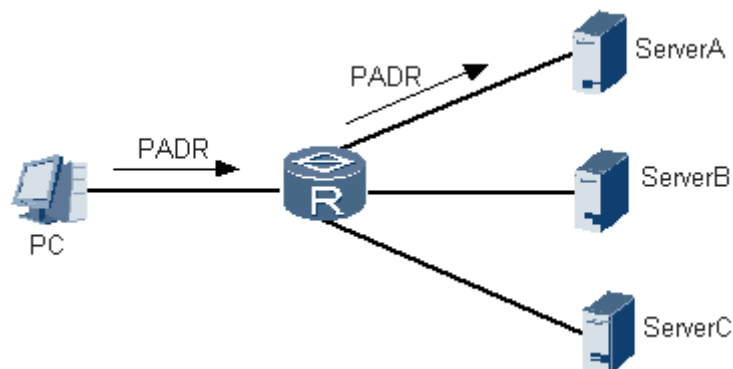
图 4-34 服务器发回 PADO 报文



3. 主机可能收到多个服务器的 PADO 报文，主机将依据 PADO 的内容，从多个服务器中选择一个，并向它发回一个会话请求报文 PADR（PPPoE Active Discovery Request）。

例如在图 4-35 中，主机选择 ServerA，并发回 PADR 报文。

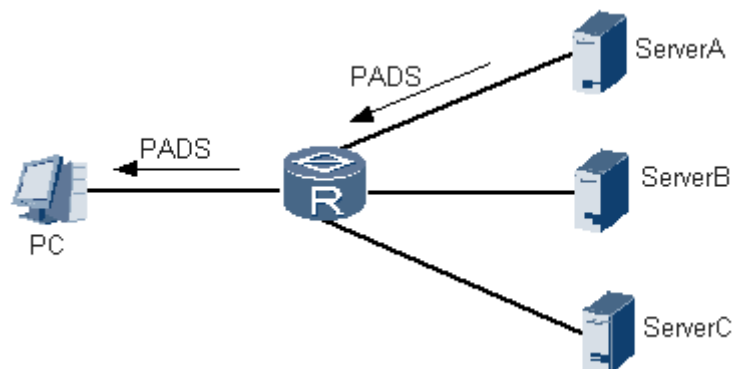
图 4-35 主机选择一个服务器并发送 PADR 报文



4. 服务器产生一个唯一的会话标识，标识和主机的这段 PPPoE 会话。并把此会话标识通过会话确认报文 PADS（PPPoE Active Discovery Session-confirmation）发回给主机，如果没有错误，双方进入 PPPoE Session 阶段。

例如在图 4-36 中，ServerA 收到 PADR 报文后，会向主机发送 PADS 报文。

图 4-36 服务器向主机发回 PADS 报文



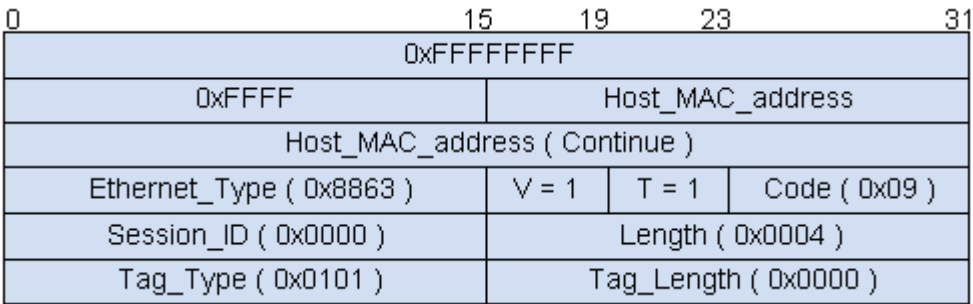
接入服务器发送确认数据包后，它就可以进入到 PPPoE 会话阶段。当主机接收到该确认数据包后，它就可以进入 PPPoE 会话阶段。

PADI 数据包

主机发送的 PADI 数据包中，Destination\_address 域为广播地址，Code 域设置为 0x09，Session\_ID 域设置为 0x0000。

PADI 数据包必须包含且仅包含一个 Tag\_Type 为 Service-Name 的 Tag，以表明主机请求的服务，以及任意数目的其它类型的 TAG。整个 PADI 数据包，包括 PPPoE 头部不允许超过 1484 个字节，以预留空间让中继代理向数据包中增加类型为 Relay-Session-Id 的 Tag。

图 4-37 PADI 报文结构示例图



PADO 数据包

如果接入服务器能够为收到的 PADI 请求提供服务，它将通过发送一个 PADO 数据包来做出应答。Destination\_address 是发送 PADI 报文的主机的单播地址，Code 域设置为 0x07，Session\_ID 域设置为 0x0000。

PADO 数据包必须包含一个类型为 AC-Name 的 Tag，AC 是接入服务器的名字。还必须包含与 PADI 中相同的 Service-Name，以及任意数目的类型为 Service-Name 的 Tag，表明接入服务器提供的其它服务。

如果接入服务器不能为 PADI 提供服务，则不允许用 PADO 作响应。

图 4-38 PADO 报文结构示例图

|                                              |      |                                 |       |               |
|----------------------------------------------|------|---------------------------------|-------|---------------|
| 0                                            | 15   | 19                              | 23    | 31            |
| Host_MAC_address                             |      |                                 |       |               |
| Host_MAC_address ( Continue )                |      | Access_Concentrator_MAC_address |       |               |
| Access_Concentrator_MAC_address ( Continue ) |      |                                 |       |               |
| Ethernet_Type ( 0x8863 )                     |      | V = 1                           | T = 1 | Code ( 0x07 ) |
| Session_ID ( 0x0000 )                        |      | Length ( 0x0020 )               |       |               |
| Tag_Type ( 0x0101 )                          |      | Tag_Length ( 0x0000 )           |       |               |
| Tag_Type ( 0x0102 )                          |      | Tag_Length ( 0x0018 )           |       |               |
| 0x47                                         | 0x6F | 0x20                            | 0x52  |               |
| 0x65                                         | 0x64 | 0x42                            | 0x61  |               |
| 0x63                                         | 0x6B | 0x20                            | 0x2D  |               |
| 0x20                                         | 0x65 | 0x73                            | 0x68  |               |
| 0x73                                         | 0x68 | 0x65                            | 0x73  |               |
| 0x68                                         | 0x6F | 0x6F                            | 0x74  |               |

PADR 数据包

由于 PADI 是广播的，主机可能收到不只一个 PADO 回应报文。它将审查接收到的所有 PADO 报文，并根据其中的 AC-Name 或 PADO 所提供的服务来选择一个做为自己的服务器。

主机向选中的接入服务器发送一个 PADR 数据包。其中，Destination\_address 域设置为发送 PADO 的接入服务器的单播地址，Code 域设置为 0x19，Session\_ID 设置为 0x0000。

PADR 必须包含且仅包含一个 Tag\_Type 为 Service-Name 的 TAG，表明主机请求的服务，以及任意数目其他类型的 Tag。

PADS 数据包

当接入服务器收到一个 PADR 数据包，它就准备开始一个 PPPoE 会话。接入服务器为 PPPoE 会话创建一个唯一的 Session\_ID 并用一个 PADS 数据包向主机响应。Destination\_address 域是发送 PADR 数据包的主机的单播以太网地址，Code 域设置为 0x65，Session\_ID 设置为创建好的 PPPoE 会话标识符。

PADS 数据包中包含且仅包含一个 Tag\_Type 为 Service-Name 的 Tag，表明接入服务器已经接受的该 PPPoE 会话的服务类型，以及任意数目的其他类型的 Tag。

如果接入服务器不接受 PADR 中的 Service-Name，则接入服务器也回应给主机一个 PADS 报文，但该 PADS 中带有类型为 Service-Name-Error 的 Tag 以及任意数目的其它 TAG 类型。这种情况下，Session\_ID 必须设置为 0x0000。

PADT 数据包

PADT（PPPoE Active Discovery Terminate）数据包可以在会话建立以后的任意时刻发送，表明 PPPoE 会话已经终止。

它可以由主机或接入服务器发送，Destination\_address 域为单播以太网地址，Code 域设置为 0xA7，Session\_ID 必须表明终止的会话，这种数据包不需要任何 Tag。

当收到 PADT 以后，就不允许再使用该会话发送 PPP 流量了。在发送或接收到 PADT 后，即使是常规的 PPP 结束数据包也不允许发送。

PPPoE 通信双方应该使用 PPP 协议自身来结束 PPPoE 会话，但在无法使用 PPP 时可以使用 PADT 结束通信。

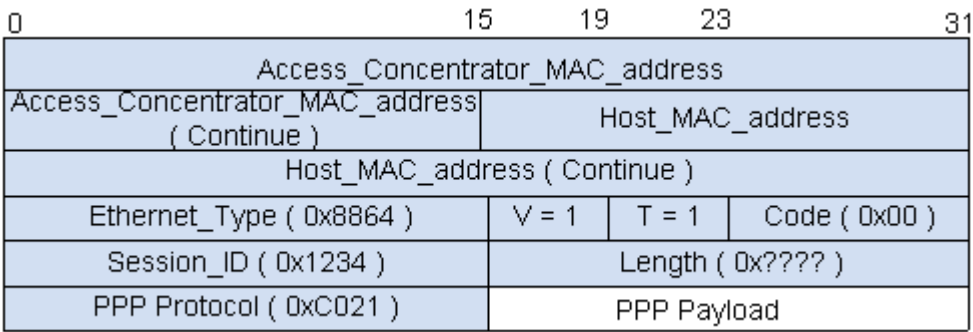
4.6.4.3 PPPoE Session 阶段

PPPoE 会话（PPPoE Session）开始后，PPP 报文作为 PPPoE 帧的净荷，封装在以太网帧内发送到对端。

这时所有的以太网数据包都是单播的。

- Ethernet\_Type 域设置为 0x8864。
- PPPoE 的 Code 必须设置为 0x00。
- PPPoE 会话的 Session\_ID 不允许发生改变，必须是 Discovery 阶段所指定的值。
- PPPoE 的 Payload 包含一个 PPP 帧。PPP 帧的开始字段是 PPP Protocol-ID。

图 4-39 从主机发送到接入服务器的 PPP LCP 数据包示例图



进入 PPPoE Session 阶段后，主机或服务器任何一方都可发 PADT 报文通知对方结束 PPPoE 会话。

4.7 WLAN

无线局域网是指以无线信道作传输媒介的计算机局域网，是有线联网方式的重要补充和延伸。

4.7.1 介绍



## 定义

无线局域网（Wireless Local Area Network）是指以无线信道作传输媒介的计算机局域网，是有线联网方式的重要补充和延伸，并逐渐成为计算机网络中一个至关重要的组成部分，广泛适用于需要可移动数据处理或无法进行物理传输介质布线的领域。

## 目的

传统有线网络在涉及到网络移动和重新布局时存在较大制约。

无线局域网使用户可以对有线网络进行更加灵活的扩展和延伸，在不进行传统布线的同时，提供有线局域网的所有功能。

## 4.7.2 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                           |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00 |

## 4.7.3 原理描述

### 4.7.3.1 WLAN 的基本原理

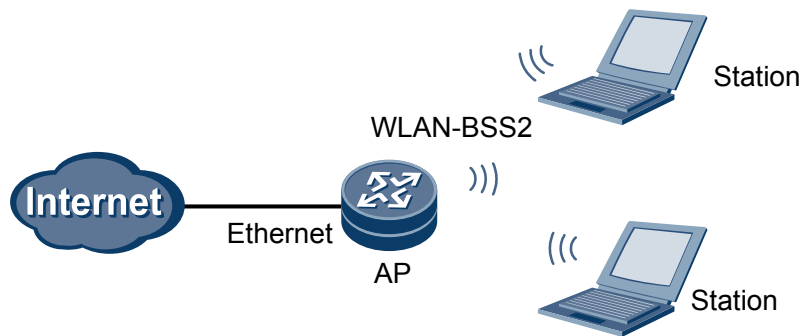
无线局域网（Wireless Local Area Network）是指以无线信道作传输媒介的计算机局域网，是有线联网方式的重要补充和延伸，并逐渐成为计算机网络中一个至关重要的组成部分，广泛适用于需要可移动数据处理或无法进行物理传输介质布线的领域。

作为新型的宽带接入方式，WLAN 越来越受到人们的宠爱，并得到迅猛的发展，由此涌现出多种 WLAN 的标准：

- 802.11
- Bluetooth
- HiperLAN2（High Performance Radio LAN 2）
- HomeRF（Home Radio Frequency）

其中，802.11 因技术简单、通信质量稳定、传输带宽较高而成为了构建 WLAN 的常用标准。WLAN 典型应用如图 4-40 所示。

图 4-40 WLAN 典型应用组网图



### 4.7.3.2 802.11 协议族

随着 IEEE802.11 无线网络标准的制定与发展，使无线网络技术更加成熟与完善。目前，常见的 IEEE802.11 标准有：

- IEEE802.11—IEEE 制定的第 1 个版本标准。其中定义了 MAC 层和物理层，传输速率为 2Mbit/s，工作在 2.4GHz。
- IEEE802.11a—主要对物理层做了补充，传输速率为 54Mbit/s，工作在 5GHz。
- IEEE802.11b—主要对物理层做了补充，传输速率为 11Mbit/s，工作在 2.4GHz。
- IEEE802.11e—对 MAC 层作了一些修正以改进服务质量（QoS），并解决了一些安全问题。
- IEEE802.11g—工作在 2.4GHz，兼容 802.11b，传输速率为 54Mbit/s。
- IEEE802.11i—主要补充了无线网络安全方面的内容。
- IEEE802.11n—传输最大值为 600Mbit/s，支持 MIMO（Multi-Input Multi-Output）技术。

### 4.7.3.3 常见术语

- AP（Access Point）：提供无线工作站到局域网的桥接功能，提供无线到有线和有线到无线的帧转换。
- BSS（Basic Service Set）：所有通信都必须通过接入点，服务区域相当于接入点的传输范围。
- SSID（Service Set Identifier）：每个 BSS 由 SSID 来区分。无线工作站必须配置与 AP 相同的 SSID 才能访问 AP。
- 虚拟 AP（Virtual AP）：一块无线接口卡上模拟出的多个逻辑 AP，实现同一物理接口卡上多个逻辑 AP 的功能。多个逻辑 AP 之间依靠不同的 SSID 加以区分。
- WEP（Wired Equivalent Privacy）：在链路层采用 RC4 对称加密技术。用户的加密密钥必须与 AP 的密钥相同时才能获准存取网络的资源。
- WPA（WiFi Protected Access）：是继承了 WEP 基本原理而又解决了部分 WEP 缺点的一种新技术。它改变了密钥的生成方式，加强了密钥的生成算法，采用更频繁地变换密钥方式来获得更高的安全，因此 WPA 加密就有效的解决了 WEP 加密应用中的不足。

- **WPA2:** WPA 的最新版本，提供比 WPA 更加安全、更适合应用在无线局域网环境下的加密协议—CCMP。
- **802.1X:** 802.1X 协议是一种基于端口的网络接入控制协议。连接在端口上的客户端必须通过认证，才能访问网络中的资源；如果不能通过认证，则无法访问网络。
- **WMM:** WMM 是一种无线 QoS 协议，用于保证高优先级的报文有优先的发送权利，从而保证语音、视频等应用在无线网络中有更好的质量。

# 5 IP 业务

## 关于本章

### 5.1 IP 地址

连接到 Internet 上的设备接口必须有一个全球唯一的 IP 地址。IP 地址可以通过配置主从 IP 地址、协商或者接口借用 IP 地址获得。

### 5.2 ARP

ARP 用于实现网络层的 IP 地址和数据链路层的 MAC 地址的映射。

### 5.3 DNS

DNS 域名解析系统建立域名与 IP 地址的对应关系，方便用户记忆。

### 5.4 DHCP

DHCP 采用客户端/服务器通信模式，由客户端向服务器提出配置申请（包括 IP 地址、默认网关、DNS 服务器、WINS 服务器等参数），服务器根据策略返回相应配置信息。

### 5.5 DHCP Snooping

介绍了 DHCP Snooping 安全的基本原理、基本概念和应用。

### 5.6 IP 性能

在一些特定的网络环境里，需要调整 IP 参数，以便网络性能达到最佳。

### 5.7 IPv6 基础

IPv6 是新版本的网络层协议。

### 5.8 IPv6 邻居发现

邻居发现 ND 是确定邻居节点之间关系的一组消息和进程，安全邻居发现 SEND 协议是 ND 的增强。

### 5.9 IPv6 性能

在一些特定的网络环境里，需要调整 IPv6 参数，以便网络性能达到最佳。

### 5.10 DHCPv6

在 IPv6 网络中，客户端可以通过 DHCPv6 技术动态获取 IPv6 地址等参数。

### 5.11 IPv6 over IPv4 隧道

IPv6 over IPv4 隧道是为了在 IPv4 网络中传递 IPv6 节点间信息而出现的隧道技术。

### 5.12 IPv4 over IPv6 隧道

IPv4 over IPv6 隧道是在 IPv4 向 IPv6 过渡的后期用于联系 IPv4 孤岛的一种隧道技术。

## 5.1 IP 地址

连接到 Internet 上的设备接口必须有一个全球唯一的 IP 地址。IP 地址可以通过配置主从 IP 地址、协商或者接口借用 IP 地址获得。

### 5.1.1 介绍

#### 定义

IP 地址就是在 Internet 上使用的 32 比特地址。IP 地址的层次是按网络结构进行划分的，一个 IP 地址是由网络号和主机号两部分组成的。

在 IP 网络上，需要为网络上的主机分配 IP 地址。如果用户要将一台计算机连接到 Internet 上，就需要向 ISP 申请一个 IP 地址。

IP 地址由 32 位二进制数值组成，但为了便于用户识别和记忆，采用了“点分十进制表示法”。采用了这种表示法的 IP 地址由 4 个由点分隔的十进制整数来表示，每个十进制整数对应一个字节。例如，A 主机的 IP 地址使用二进制的表示形式为 00001010 00000001 00000001 00000010，那么采用点分十进制表示法表示为 10.1.1.2。

IP 地址由如下两部分组成：

- 网络号码字段（net-id）：用于区分不同网络。网络号码字段的前几位称为类别字段（又称为类别比特），用来区分 IP 地址的类型。
- 主机号码字段（host-id）：用于区分一个网络内的不同主机。

#### 目的

IP 地址的网络号码字段用来标识一个网络，主机号码字段用来标识网络中网络设备的一个连接。如果有多台网络设备，无论它们分别处于任何物理位置，只要它们具有相同的网络号，那他们就处在同一网络中。也就是说，在公共网络内的多台网络设备是否处于相同网络与它们所处的物理位置无关。

### 5.1.2 参考标准和协议

与 IP 地址相关的参考标准与协议如下：

- RFC 1166: Internet Numbers
- RFC 1918: Address Allocation for Private Internets

### 5.1.3 可获得性

#### License 支持

本特性无需 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

5.1.4 原理描述

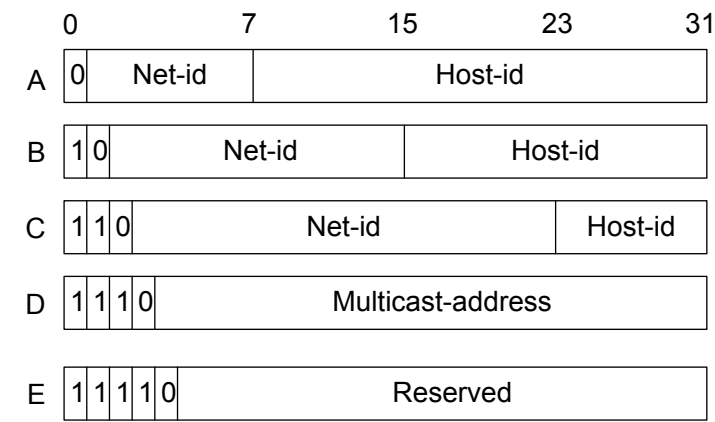
5.1.4.1 IP 地址分类

五类 IP 地址

为了方便 IP 地址的管理及组网，IP 地址分成五类，如图 5-1 所示。

通过网络号码字段的前几个比特就可以判断 IP 地址属于哪一类，这是区分各类地址最简单的方法。

图 5-1 五类 IP 地址



目前大量使用中的 IP 地址属于 A、B、C 三类 IP 地址中的一种。D 类地址是组播地址，E 类地址保留。在 IETF（Internet Engineering Task Force）发布的 RFC1166 Internet Numbers 中详细描述了各类 IP 地址。

在使用 IP 地址时要注意，一些 IP 地址是保留作为特殊用途的，一般的用户不能使用。表 5-1 列出各类 IP 地址的范围。

表 5-1 IP 地址分类及范围

| 网络类型 | 地址范围                        | 用户可用的 IP 网络范围             | 说明                                                                                                                                                                                                                                      |
|------|-----------------------------|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| A    | 0.0.0.0 ~ 127.255.255.255   | 1.0.0.0 ~ 126.255.255.0   | 全 0 的主机号码表示该 IP 地址就是网络的地址，用于网络路由；全 1 的主机号码表示广播地址，即对该网络上所有的主机进行广播；IP 地址 0.0.0.0 仅在采用 DHCP 方式的系统启动时允许本主机利用它进行临时的通信，并且永远不是有效目的地址；网络号码为 0 的 IP 地址表示当前网络，可以让机器引用自己的网络而不必知道其网络号；所有形如 127.X.Y.Z 的地址都保留作环回测试，发送到这个地址的分组不会输出到线路上，它们被内部处理并当作输入分组。 |
| B    | 128.0.0.0 ~ 191.255.255.255 | 128.0.0.0 ~ 191.255.255.0 | 全 0 的主机号码表示该 IP 地址就是网络的地址，用于网络路由；全 1 的主机号码表示广播地址，即对该网络上所有的主机进行广播。                                                                                                                                                                       |
| C    | 192.0.0.0 ~ 223.255.255.255 | 192.0.0.0 ~ 223.255.255.0 | 全 0 的主机号码表示该 IP 地址就是网络的地址，用于网络路由；全 1 的主机号码表示广播地址，即对该网络上所有的主机进行广播。                                                                                                                                                                       |
| D    | 224.0.0.0 ~ 239.255.255.255 | 无                         | D 类地址是一种组播地址。                                                                                                                                                                                                                           |
| E    | 240.0.0.0 ~ 255.255.255.255 | 无                         | 保留。255.255.255.255 用于局域网广播地址。                                                                                                                                                                                                           |

特殊的 IP 地址

在实际使用过程中，有一些特殊的 IP 地址，其范围和描述如表 5-2 所示。

表 5-2 特殊情况的 IP 地址

| IP 地址网络号 | IP 地址子网号 | IP 地址主机号 | 能否作为源端地址 | 能否作为目的端地址 | 描述                   |
|----------|----------|----------|----------|-----------|----------------------|
| 全 0      | -        | 全 0      | 可以       | 不可以       | 用于网络上的主机             |
| 全 0      | -        | 主机号      | 可以       | 不可以       | 用于网络上的特定主机           |
| 127      | -        | 任何值      | 可以       | 可以        | 用于环回地址               |
| 全 1      | -        | 全 1      | 不可以      | 可以        | 用于受限的广播（永远不被转发）      |
| net-id   | -        | 全 1      | 不可以      | 可以        | 用于向以 net-id 为目的的网络广播 |



| IP 地址<br>网络号 | IP 地址<br>子网号 | IP 地址<br>主机号 | 能否作为<br>源端地址 | 能否作为目<br>的端地址 | 描述                             |
|--------------|--------------|--------------|--------------|---------------|--------------------------------|
| net-id       | subnet-id    | 全 1          | 不可以          | 可以            | 用于向以 net-id，subnet-id 为目的的子网广播 |
| net-id       | 全 1          | 全 1          | 不可以          | 可以            | 用于向以 net-id 为目的的所有子网广播         |

 说明

net-id，subnet-id 分别表示不全为 0 和不全为 1 的对应字段。

私有 IP 地址

为了解决 IP 地址短缺的问题，提出了私有地址的概念。私有地址是指内部网络或主机地址，这些地址只能用于某个内部网络，不能用于公共网络。RFC1918 描述了为私有网络预留的 3 个 IP 地址段。

IP 地址分配组织规定将下列的 IP 地址保留用作私有地址，如表 5-3 所示。

表 5-3 私有 IP 地址

| 网络类型 | 地址范围                          |
|------|-------------------------------|
| A    | 10.0.0.0 ~ 10.255.255.255     |
| B    | 172.16.0.0 ~ 172.31.255.255   |
| C    | 192.168.0.0 ~ 192.168.255.255 |

5.1.4.2 子网划分

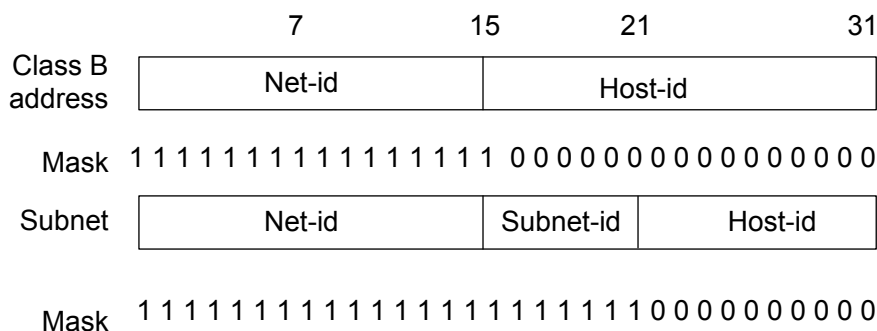
从 1985 年起，为了使 IP 地址的使用更加灵活，只分配 IP 地址的网络号码 net-id，而后面的主机号码 host-id 则是受本单位控制。即某个单位申请到 IP 地址时，实际上只是拿到了一个网络号码 net-id，具体的各个主机号码 host-id 则由该单位自行分配，只要做到在该单位管辖的范围内无重复的主机号码即可。

当一个单位的主机很多而且分布在很大的地理范围时，为了便于管理，可将单位内部的主机号码再进一步划分为多个子网。通过子网划分，整个网络地址可以划分成更多的小网络。

需要注意的是，子网的划分对于在本单位以外是看不见的。从外部看，这个单位只有一个网络号码。只有当外面的报文进入到本单位范围后，本单位的 USG 才根据子网号码再进行选路，找到目的主机。

如图 5-2 所示，为一个 B 类 IP 地址划分子网情况，其中子网掩码由一串连续的“1”和一串连续的“0”组成。“1”对应于网络号码和子网号码字段，而“0”对应于主机号码字段。

图 5-2 IP 地址子网划分



将 32 位的 IP 地址和子网掩码的对应位作与运算可以确定 IP 地址的网络号。例如，IP 地址为 10.1.1.2，子网掩码为 255.255.0.0，那么将 IP 地址与其相应掩码位执行与运算的结果就是网络地址 10.1.0.0。

多划分出一个子网号码字段是要付出代价的。举例来说，本来一个 B 类 IP 地址可以容纳 65534 个主机号码。但划分出 6bits 长的子网字段后，最多可有 64 个子网，每个子网有 10bit 的主机号码，即每个子网最多可有 1022 ( $2^{10}-2$ ，去掉全 1 和全 0 的主机号码) 个主机号码。因此主机号码的总数是 ( $64 \times 1022 = 65408$ ) 个，比不划分子网时要少 126 个。

若一个单位不进行子网的划分，则其子网掩码即为默认值，此时子网掩码中“1”的长度就是网络号码的长度。因此，对于 A、B 和 C 类的 IP 地址，其对应子网掩码的默认值分别为 255.0.0.0、255.255.0.0 和 255.255.255.0。

### 5.1.4.3 IP 地址获取方式

用户访问 Internet 必须要有合法的 IP 地址。目前有以下几种主要的 IP 地址获取方式。

#### 手工分配 IP 地址

可以直接在用户计算机上手工配置 IP 地址，这种方式一般用于固定用途的服务器或有特殊需要的用户，例如 Web 服务器、防火墙、路由器等。

#### 为 PPP 接入的用户分配 IP 地址

采用 PPP 方式接入的用户，可以利用 PPP 的地址协商功能，由接入服务器分配 IP 地址。有两种方法可以为 PPP 用户分配 IP 地址。

- 多用户情况下，配置 IP 地址池，在接口视图下指定该接口使用的地址池。
- 单用户情况下，不配置 IP 地址池，在接口上直接给用户配置指定的 IP 地址。

#### 使用 DHCP 服务分配 IP 地址

DHCP 采用客户/服务器通信模式。网络管理员在 DHCP 服务器上设定一个 IP 地址范围，客户端向服务器提出配置申请（包括分配的 IP 地址、子网掩码、缺省网关等参数），服务器根据策略返回相应的配置信息。

采用以太网接入的用户，可以通过 DHCP 获得 IP 地址。

## IP 地址借用

一个接口如果没有 IP 地址就无法生成路由，也就无法转发报文。IP 地址借用（IP Address Unnumbered）就是在本接口没有 IP 地址的情况下，可以使用其他接口的 IP 地址。所谓“借用 IP 地址”，其实质就是：一个接口上没有配置 IP 地址，但是还想使用该接口。就向其他有 IP 地址的接口借一个 IP 地址过来，以使该接口能够正常使用。

IP 地址借用的主要目的是节省 IP 地址资源。有时某个接口只是偶尔使用，这种情况也可配置该接口借用其他接口的 IP 地址，而不必让其一直占用一个单独的 IP 地址。

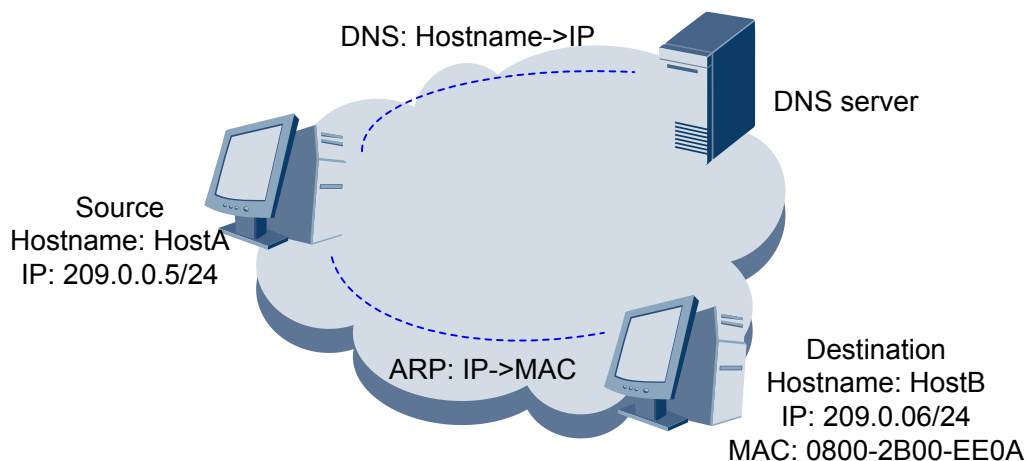
### 5.1.4.4 IP 地址解析

一台 USG 用来连接多个网络，具有多个网络的 IP 地址。上面讲的 IP 地址还不能直接用来进行通信。具体原因如下：

- IP 地址只是主机在网络层中的地址，若要将网络层中传送的数据报交给目的主机，必须知道该主机的物理地址。因此必须将 IP 地址解析为物理地址。
- 用户平时不愿意使用难于记忆的 IP 地址，而更愿意使用易于记忆的主机名，因此也需要将主机名解析为 IP 地址。

图 5-3 表示了主机名、IP 地址和物理地址之间的关系。在 Ethernet 上，主机的物理地址就是指 MAC 地址。将主机名解析为 IP 地址的操作是由 DNS 服务器来完成，而将 IP 地址解析为 MAC 地址的操作是由 ARP 来完成的。

图 5-3 主机名、IP 地址和物理地址之间的关系



## 5.2 ARP

ARP 用于实现网络层的 IP 地址和数据链路层的 MAC 地址的映射。

### 5.2.1 介绍

## 定义

ARP（Address Resolution Protocol）用于将 IP 地址解析为 MAC 地址。ARP 表项可以分为动态和静态两种类型。另外 ARP 还有扩展应用功能，包括代理 ARP、免费 ARP 等。

## 目的

局域网中每台主机或路由器都有一个 32 位的 IP 地址，这个地址用于该主机的所有通信。IP 地址的分配独立于机器的硬件地址。而在以太网中，主机或路由器是根据 48 位的 MAC 地址来发送、接收以太网数据帧的，这个 MAC 地址又称为物理地址或硬件地址，是制造设备时分配到以太网接口中的。因而，在实际的网络互联中，需要一种地址解析的机制来为这两种不同的地址形式提供映射。

ARP 协议主要是解决以上问题，此外 ARP 特性中还包括如下的应用特性：

- 动态 ARP：利用 ARP 报文，ARP 动态执行并自动进行 IP 地址到以太网 MAC 地址的解析，无需网络管理员手工处理。
- 静态 ARP：建立 IP 地址和 MAC 地址之间固定的映射关系，在主机和路由器上不能动态调整此映射关系。需要网络管理员手工添加。
- 代理 ARP：也叫路由式代理 ARP。当主机上没有配置缺省网关地址（即不知道如何到达本网络的中介系统），它可以发送一个 ARP 请求（请求目的主机的 MAC 地址），启动代理 ARP 功能的设备收到这样的请求后，会使用自己的 MAC 地址作为该 ARP 请求的回应，使得处于不同物理网络但网络号相同的内部主机之间可以正常的相互通信。
- 免费 ARP：用于检查重复的 IP 地址和通告新的 MAC 地址。
- 授权 ARP：只在启用 DHCP 服务器功能的设备上实现，且只适用于 DHCP 服务器与 DHCP 客户端在同一网段的组网环境，防止攻击者仿冒其他合法 DHCP 客户端的 IP 地址或 MAC 地址对网络进行攻击。

## 5.2.2 参考标准和协议

与 ARP 相关的参考标准与协议如下：

- RFC 826: Ethernet Address Resolution Protocol
- RFC 903: Reverse Address Resolution Protocol
- RFC 1027: Using ARP to Implement Transparent Subnet Gateways
- RFC 1042: Standard for the Transmission of IP Datagrams over IEEE 802 Networks

## 5.2.3 可获得性

### License 支持

本特性无需 License 支持。

版本支持

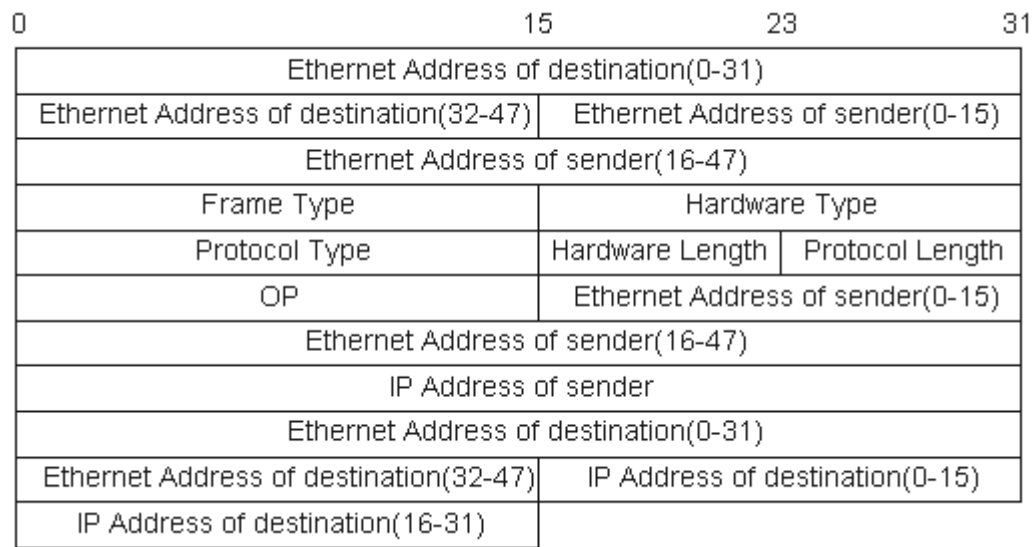
| 产品                                                                      | 最低支持版本                                                                                                      |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br><br>USG5500 V200R001C00 |

5.2.4 原理描述

5.2.4.1 ARP 报文

ARP 请求和应答的报文格式如图 5-4 所示。

图 5-4 ARP 请求和应答报文格式



ARP 报文的长度是 42 字节。最前面 14 字节的字段表示以太网首部，后面的 28 字节内容是 ARP 请求或响应的报文格式。

主要字段的解释如表 5-4 所示。

表 5-4 ARP 报文各字段的含义

| 字段                              | 长度    | 含义                                                  |
|---------------------------------|-------|-----------------------------------------------------|
| Ethernet Address of destination | 48 比特 | 目的以太网地址。发送 ARP 请求时，为广播的 MAC 地址，0xFF.FF.FF.FF.FF.FF。 |

| 字段                              | 长度    | 含义                                               |
|---------------------------------|-------|--------------------------------------------------|
| Ethernet Address of sender      | 48 比特 | 源以太网地址。                                          |
| Frame Type                      | 16 比特 | 表示后面数据的类型。对于 ARP 请求或应答来说，该字段的值为 0x0806。          |
| Hardware Type                   | 16 比特 | 表示硬件地址的类型。对于以太网，该类型的值为“1”。                       |
| Protocol Type                   | 16 比特 | 表示发送方要映射的协议地址类型。对于 IP 地址，该值为 0x0800。             |
| Hardware Length                 | 8 比特  | 表示硬件地址的长度，单位是字节。对于 ARP 请求或应答来说，该值为 6 和 4。        |
| Protocol Length                 | 8 比特  | 表示协议地址的长度，单位是字节。对于 ARP 请求或应答来说，该值为 4。            |
| OP                              | 16 比特 | 操作类型。OP 的值与操作类型的关系如表 5-5 所示。                     |
| Ethernet Address of sender      | 48 比特 | 发送方以太网地址。这个字段和 ARP 报文首部的源以太网地址字段是重复信息。           |
| IP Address of sender            | 32 比特 | 发送方的 IP 地址。                                      |
| Ethernet Address of destination | 48 比特 | 接收方的以太网地址。发送 ARP 请求时，该处填充值为 0x00.00.00.00.00.00。 |
| IP Address of destination       | 32 比特 | 接收方的 IP 地址。                                      |

表 5-5 OP 值与操作类型的对应关系

| 操作 | 操作类型    |
|----|---------|
| 1  | ARP 请求  |
| 2  | ARP 应答  |
| 3  | RARP 请求 |
| 4  | RARP 应答 |

 说明

对于一个 ARP 请求来说，除目的端硬件地址外的所有其他的字段都有填充值。

5.2.4.2 ARP 原理

## ARP 地址解析过程

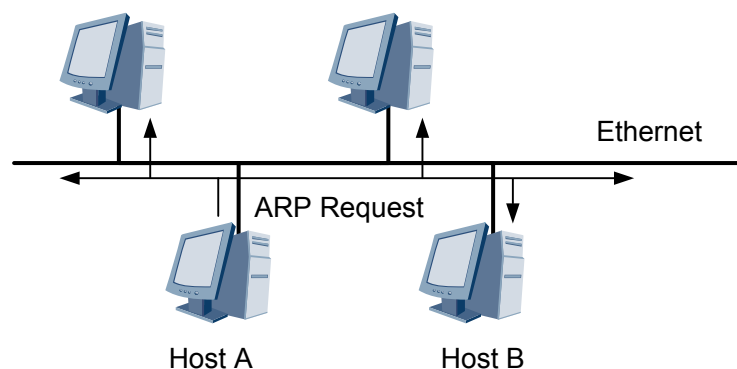
TCP/IP 协议的设计人员根据以太网这种具有广播特性的网络开发出的 ARP 地址解析协议。主机在仅知道同一物理网络上的目的端的 IP 地址情况下，通过 ARP 解析到目的端的 MAC 地址。即使网络上的主机发生变化，比如主机的增加或减少、主机更换计算机的网卡等，仍可以完成从 IP 地址到 MAC 地址的转换，并且这个转换关系可以动态更新。

ARP 是通过以下两个步骤完成地址解析过程的。

如图 5-5 所示，当主机 A 仅知道主机 B 的 IP 地址而不知道其 MAC 地址时，它广播一个 ARP 请求报文，请求得到主机 B 的 MAC 地址。

### ARP 请求

图 5-5 ARP 请求过程

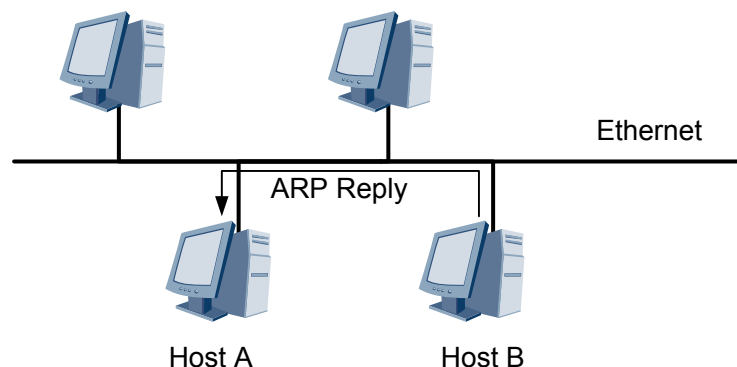


### ARP 响应

因为主机 A 发送的是一个广播报文，所以包括主机 B 在内的所有主机都会接收到这个请求。由于 ARP 请求报文的目的 IP 地址是主机 B 的 IP 地址，所以只有主机 B 会响应这个 ARP 请求。如图 5-6 所示，主机 B 向主机 A 发出一个包含其 MAC 地址的 ARP 响应报文。

当主机 A 接收到主机 B 的响应报文后，就用这个 MAC 地址和主机 B 通信。

图 5-6 ARP 响应过程



## ARP 老化机制

### 高速缓存

如果每次主机 A 向主机 B 发送一个分组前都要发送一个广播的 ARP 请求报文，则会增加很多网络的通信量。而且网络上的所有机器都必须接受和处理这个广播的 ARP 请求报文，这也极大的影响了网络运行效率。

为了解决以上问题，每台主机上都维护着一个高速缓存，这是 ARP 高效运行的一个关键。在这个高速缓存中，存放最近获得的 IP 地址到 MAC 地址的映射关系。

发送方在每次发送分组时，都先在缓存中查找目的 IP 地址所对应的 MAC 地址。如果 ARP 缓存中有对应的 MAC 地址，主机就不会再发送 ARP 请求报文，而是直接将分组发至这个 MAC 地址。如果 ARP 缓存中没有对应的 MAC 地址时，主机才会发送广播的 ARP 请求报文。

### 动态 ARP 表项的老化超时时间

如图 5-6 所示，当主机 B 回应了主机 A 的 ARP 请求后，在主机 A 的缓存中会形成主机 B 的 IP 地址和其 MAC 地址的映射关系。但是，如果主机 B 发生故障后或者更换了网卡时，主机 A 没有得到任何关于主机 B 的任何通告，于是主机 A 仍会继续将分组发送给主机 B。造成地址解析出现错误的原因就是主机 A 中的缓存表的信息没有得到及时的更新。

为了减少地址解析过程中所出现的错误，ARP 高速缓存中的表项一般都会设定一个定时器。当达到定时器的动态 ARP 表项的老化超时时间后，删除掉这个表项。

通过设置定时器，在地址解析过程中出现错误的现象得到了改善但并没有完全消除，其原因在于时延。如果定时器的动态 ARP 表项的老化超时时间是 N 秒，发送方只有等到 N 秒后才能检测到接收方出现了故障，在此期间发送方缓存表的信息还是没有得到及时的更新。

### 动态 ARP 表项的老化探测次数

除了设置定时器中动态 ARP 表项的老化超时时间，还可以通过设置动态的探测次数来减少地址的解析错误。在一个动态 ARP 表项老化之前，系统先进行探测，如果超过设置的探测次数后仍没有应答，则此 ARP 表项将被删除。

## 5.2.4.3 静态 ARP

静态 ARP 是指 IP 地址和 MAC 地址之间有固定的映射关系，在主机和路由设备上不能动态调整此映射关系。静态 ARP 需要网络管理员手动配置。

配置静态 ARP 表项主要应用于以下情况：

- 为了将目的 IP 地址不在本网段的报文，穿过本网段的某个网关，使得到该 IP 地址的报文能通过该网关进行转发。
- 当用户需要过滤掉一些非法的报文时，可以将这些非法报文的目的 IP 地址绑定到某个不存在的 MAC 地址。
- 将 IP 地址与 MAC 地址绑定，防止 ARP flood 等攻击。

静态 ARP 表项的优先级高于动态 ARP 表项。如果 ARP 表中存在某 IP 地址的动态 ARP 表项，当手动配置同一 IP 地址的静态 ARP 表项时，该表项将覆盖原有的动态 ARP 表项。



#### 5.2.4.4 动态 ARP

一般情况下，ARP 动态执行并自动进行 IP 地址到以太网 MAC 地址的解析，无需网络管理员的介入。

在 USG 的实现中，如果收到的 ARP 报文满足以下条件中的任何一条，系统将创建或更新 ARP 表项。

- ARP 报文的源 IP 地址与入接口 IP 地址在同一网段，且不是广播地址，目的 IP 地址是本接口 IP 地址。
- ARP 报文的源 IP 地址与入接口 IP 地址在同一网段，且不是广播地址，目的 IP 地址是本接口的 VRRP（Virtual Router Redundancy Protocol）虚拟 IP 地址。
- ARP 报文的源 IP 地址与入接口 IP 地址在同一网段，且不是广播地址，入接口是 IPoEoA 应用的 Virtual Ethernet 接口。
- ARP 报文的目的 IP 地址是入接口上配置的 NAT 地址池中的地址。

如果收到的 ARP 报文的源 IP 地址在入接口的 ARP 表中已经存在对应表项，也将对 ARP 表项进行更新。

#### 5.2.4.5 代理 ARP

代理 ARP 主要是通过代理的方式来解决网络互通问题的 ARP 实现功能。

代理 ARP 有以下特点：

- 所有处理在 ARP 子网网关（ARP Subnet Gateways）进行，所连网络中的主机不必做任何改动。
- 在主机端看不到子网，只是一个标准 IP 网络。
- 代理 ARP 只影响主机的 ARP 高速缓存，对网关的 ARP 高速缓存和路由表没有影响。
- 使用代理 ARP 后，主机应该减小 ARP 老化时间，以尽快使无效 ARP 项失效，减少发给路由器而路由器却不能转发的报文。

本设备支持路由式代理 ARP。路由式代理 ARP 就是使那些在同一网段却不在同一物理网络上的计算机或路由器能够相互通信的一种功能。在实际应用中，如果连接路由器的当前主机上没有配置缺省网关地址（即不知道如何到达本网络的中介系统），此时将无法进行数据转发。

路由式代理 ARP 可以解决这个问题，主机发送一个 ARP 请求（请求目的主机的 MAC 地址），使能代理 ARP 功能的路由器收到这样的请求后，会使用自己的 MAC 地址作为该 ARP 请求的回应，以此进行数据转发。

使能代理 ARP 功能的路由器还可隐藏物理网络的细节，使得处于不同物理网络但网络号相同的两个以太网的内部主机之间可以正常的相互通信。

#### 5.2.4.6 免费 ARP

设备主动使用自己的 IP 地址作为目标地址发送 ARP 请求，此种方式称免费 ARP。免费 ARP 有两个方面的作用：

- 用于检查重复的 IP 地址：正常情况下应当不能收到 ARP 回应，如果收到，则表明本网络中存在与自身 IP 地址重复的地址。

- 用于通告一个新的 MAC 地址：发送方更换了网卡，MAC 地址变了，为了能够在 ARP 表项老化前就通告所有主机，发送方可以发送一个免费 ARP。

### 5.2.4.7 授权 ARP

授权 ARP（Authorized Address Resolution Protocol）是指 DHCP 服务器在为客户端成功分配 IP 地址的同时，会自动在 ARP 映射表中添加一条包含客户端 MAC 地址、IP 地址的 ARP 表项，即授权 ARP 表项。

### 授权 ARP 表项

授权 ARP 表项是一类特殊的静态 ARP 表项，不会被老化，但当 DHCP 客户端下线后，该客户端对应的授权 ARP 表项将自动从 ARP 映射表中删除。

授权 ARP 表项具有较高的优先级，新添加的授权 ARP 表项可以覆盖相同的动态 ARP 表项而同时不被新添加的相同动态 ARP 表项所覆盖；但同时其优先级低于静态 ARP 表项，可以被相同的静态 ARP 表项所覆盖。

### 授权 ARP 的工作机制

授权 ARP 的工作机制是 ARP 的工作机制和 DHCP 的工作机制的结合。目前，授权 ARP 特性只在启用 DHCP 服务器功能的设备上实现，且只适用于 DHCP 服务器与 DHCP 客户端在同一网段的组网环境，不适用于 DHCP 中继组网环境。

授权 ARP 的工作机制如下：

1. DHCP 客户端以广播方式发送 DHCPDISCOVER 报文。DHCP 服务器接收到客户端的 DHCPDISCOVER 报文后，回应携带预分配的 IP 地址和其他网络配置参数的 DHCPOFFER 报文。
2. 如果有多台 DHCP 服务器向该客户端发来 DHCPOFFER 报文，DHCP 客户端只接受第一个收到的 DHCPOFFER 报文，然后以广播方式向各 DHCP 服务器回应 DHCPREQUEST 报文。DHCPREQUEST 报文中包含 DHCP 客户端的 MAC 地址以及向所选定的 DHCP 服务器请求 IP 地址的内容。
3. DHCP 服务器收到 DHCP 客户端回答的 DHCPREQUEST 报文后，便向客户端发送包含它所提供的 IP 地址和其他网络配置参数的 DHCPACK 确认报文。同时 DHCP 服务器会在自身的 ARP 表项中添加一条包含了 DHCP 客户端 MAC 地址、IP 地址的授权 ARP 表项。
4. 因为授权 ARP 表项的优先级高于动态 ARP 表项，所以授权 ARP 可以阻止 DHCP 服务器对非法 ARP 应答进行动态学习，即只有通过 DHCP 服务器分配了 IP 地址的客户端能够依据 DHCPACK 报文自动添加 ARP 表项。如果攻击者冒用合法 DHCP 客户端的 IP 地址或 MAC 地址发送 ARP 请求，因为 IP 地址或 MAC 地址与网关（DHCP 服务器与 DHCP 客户端在同一网段的组网环境下，DHCP 服务器地址通常为网关地址）记录的授权 ARP 表项不匹配，伪造的 ARP 请求将不能得到应答，从而限制冒用合法 IP 地址或 MAC 地址的攻击者通过网关访问网络，增加了网络的安全性。

## 5.3 DNS

DNS 域名解析系统建立域名与 IP 地址的对应关系，方便用户记忆。

## 5.3.1 介绍

### 定义

TCP/IP 提供了通过 IP 地址来连接到设备的功能，但对用户来讲，记住某台设备的 IP 地址是相当困难的，因此专门设计了一种字符串形式的主机命名机制，这些主机名与 IP 地址一一对应。在 IP 地址与主机名之间需要有一种转换和查询机制，提供这种机制的系统就是域名系统 DNS（Domain Name System）。

### 目的

DNS 使用一种有层次的命名方式，为网上的设备指定一个有意义的名字，并且在网络上设置域名解析服务器，建立域名与 IP 地址的对应关系。这样用户就可以使用便于记忆的、有意义的域名，而不必去记忆复杂的 IP 地址。

## 5.3.2 参考标准和协议

与 DNS 相关的参考标准与协议如下：

- RFC 1034: Domain Names - Concepts and Facilities
- RFC 1035: Domain Names - Implementation and Specification

## 5.3.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 5.3.4 原理描述

### 5.3.4.1 Internet 域名体系

最初所有机器使用的名字构成了一个非等级的名字空间，由字符序列组成所有的名字。网络信息中心 NIC（Network Information Center）管理这个名字空间，并负责接纳新名字。非等级名字空间由于自身的特点不能管理大量的机器名字，主要原因如下所示：

- 名字由字符序列组成，存在潜在的命名冲突。
- 名字空间的管理结构位于一个站点，机器数量的不断增多对管理工作带来很大负担。
- 名字到 IP 地址的映射经常发生变化，导致维护域名空间的工作量巨大。

因此，TCP/IP 把 DNS 的域名设计成了分级的结构。Internet 域名结构由 TCP/IP 协议栈中的 DNS 进行定义。DNS 把因特网划分成多个顶级域，如表 5-6 所示为每个顶级域的域名。顶级域采用组织模式和地理模式的划分模式。地理模式按照国家进行划分，每个申请加入 Internet 的国家都要向 NIC 注册一个顶级域名，如 cn 代表中国、us 代表美国等。

表 5-6 顶级 Internet 域名及其含义

| 顶级 Internet 域名 | 含义               |
|----------------|------------------|
| com            | 商业组织             |
| edu            | 教育机构             |
| gov            | 政府机构             |
| mil            | 军事部门             |
| net            | 主要网络支持中心         |
| int            | 国际组织             |
| org            | 其他组织             |
| 国家代码           | 其他各个国家（按照地理模式划分） |

 说明

前七个域对应组织模式，其余的域对应地理模式。

NIC 将顶级域的管理权分派给由其指定的管理机构，由这些管理机构再对被授权管理的域继续进行划分，从而形成了二级域。负责划分二级域的管理机构可以授权其下属的管理结构，由它们继续划分域。由此下去，便形成了层次型的 Internet 域名体系结构。

5.3.4.2 DNS 报文

DNS 查询和响应报文的格式

DNS 查询和响应报文的格式如图 5-7 所示。

图 5-7 DNS 查询和响应报文的格式

|                     |                      |    |
|---------------------|----------------------|----|
| 0                   | 15                   | 31 |
| Identification      | Flag                 |    |
| Number of Question  | Number of Answer     |    |
| Number of Authority | Number of Additional |    |
| Question<br>.....   |                      |    |
| Answer<br>.....     |                      |    |
| Authority<br>.....  |                      |    |
| Additional<br>..... |                      |    |

其中，16 比特的 Flag 字段的被划分为若干子字段，如图 5-8 所示；Flag 字段中各子字段的解释如表 5-8 所示。

图 5-8 DNS 报文中的标志字段

|    |        |   |    |    |    |    |        |       |
|----|--------|---|----|----|----|----|--------|-------|
| 0  | 1      | 5 | 6  | 7  | 8  | 9  | 12     | 15    |
| QR | opcode |   | AA | TC | RD | RA | (zero) | rcode |

这个报文由 12 字节长的首部和 4 个长度可变的字段组成。报文中问题字段是由客户填入的，由服务器返回问题的回答、授权和附加信息字段。

主要字段的解释如表 5-7 所示。

表 5-7 DNS 查询和相应的报文字段的含义

| 字段                   | 长度    | 含义                                   |
|----------------------|-------|--------------------------------------|
| Identification       | 16 比特 | 标识字段，客户通过标识字段来确定 DNS 响应是否与查询请求匹配。    |
| Flag                 | 16 比特 | 标志字段，标志字段被划分为若干子字段，各子字段的解释如表 5-8 所示。 |
| Number of Question   | 16 比特 | Question 变长字段的项目数。                   |
| Number of Answer     | 16 比特 | Answer 变长字段的项目数。                     |
| Number of Authority  | 16 比特 | Authority 变长字段的项目数。                  |
| Number of Additional | 16 比特 | Additional 变长字段的项目数。                 |

| 字段         | 长度 | 含义          |
|------------|----|-------------|
| Question   | 可变 | 查询报文中的问题部分。 |
| Answer     | 可变 | 回答字段。       |
| Authority  | 可变 | 授权字段。       |
| Additional | 可变 | 附加信息字段。     |

表 5-8 标志字段的详细格式

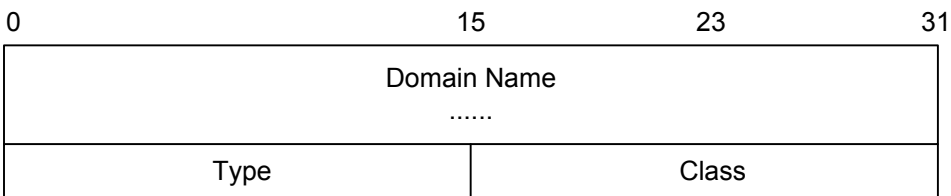
| 标志字段   | 含义                                                                                                                                                                                       |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| QR     | 操作类型：<br><ul style="list-style-type: none"> <li>● 0：查询报文</li> <li>● 1：响应报文</li> </ul>                                                                                                    |
| opcode | 查询类型：<br><ul style="list-style-type: none"> <li>● 0：标准查询</li> <li>● 1：反向查询</li> <li>● 2：服务器状态查询</li> <li>● 3 ~ 15：保留未用</li> </ul> <b>说明</b><br>反向查询是客户端请求服务器根据回答生成导致此回答的问题，这个查询类型的使用并不多。 |
| AA     | 如果将此位置为 1，则表示该域名解析服务器是授权回答该域的。                                                                                                                                                           |
| TC     | 如果将此位置为 1，则表示报文被截断。<br><b>说明</b><br>使用 UDP 传输时，应答的总长度超过 512 字节时，只返回报文的前 512 个字节内容。                                                                                                       |
| RD     | 客户端希望域名解析服务器采取的解析方式：<br><ul style="list-style-type: none"> <li>● 0：表示希望域名解析服务器采取迭代解析</li> <li>● 1：表示希望域名解析服务器采取递归解析</li> </ul>                                                           |
| RA     | 域名解析服务器采取的解析方式：<br><ul style="list-style-type: none"> <li>● 0：表示域名解析服务器采取迭代解析</li> <li>● 1：表示域名解析服务器采取递归解析</li> </ul>                                                                    |
| (zero) | 全部置 0，保留未用。                                                                                                                                                                              |

| 标志字段  | 含义                                                                                                                                                        |
|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| rcode | 响应类型： <ul style="list-style-type: none"><li>● 0：无差错</li><li>● 1：查询格式错</li><li>● 2：服务器失效</li><li>● 3：域名不存在</li><li>● 4：查询没有被执行</li><li>● 5：查询被拒绝</li></ul> |

DNS 查询报文中问题部分的格式

DNS 查询报文中问题部分的报文格式如图 5-9 所示。

图 5-9 查询报文中的问题部分格式



主要字段的解释如表 5-9 所示。

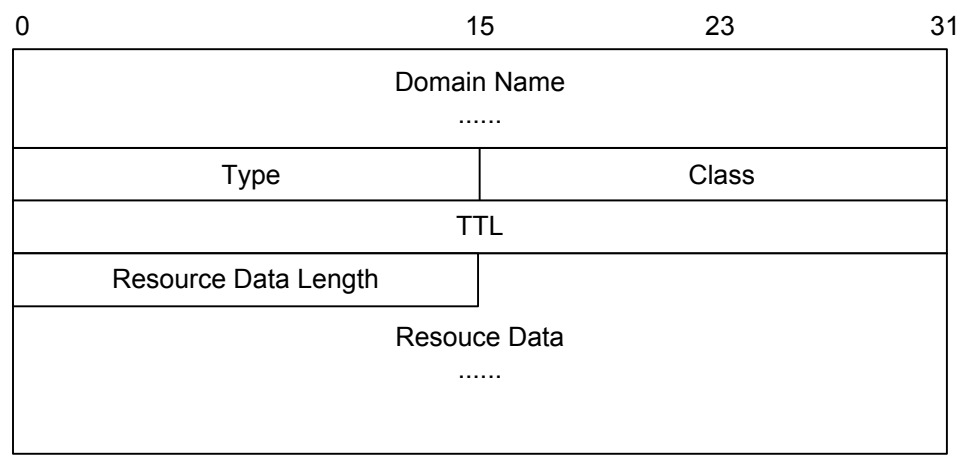
表 5-9 DNS 查询报文中问题部分的报文字段含义

| 字段          | 长度    | 含义                                |
|-------------|-------|-----------------------------------|
| Domain Name | 可变    | 客户要查询的域名，最大长度为 255 字节。            |
| Type        | 16 比特 | 查询类型。                             |
| Class       | 16 比特 | 查询等级。表示允许将域名用于任何对象。通常置 1，表示互联网地址。 |

DNS 响应报文中资源记录部分的报文格式

DNS 响应报文中最后三个字段，回答字段、授权字段和附加信息字段均采用了一种称为资源记录（Resource Record）的相同格式，如图 5-10 所示。

图 5-10 DNS 响应报文中资源记录部分的报文格式



主要字段的解释如表 5-10 所示。

表 5-10 DNS 响应报文中资源记录部分的报文字段含义

| 字段                   | 长度    | 含义                                                 |
|----------------------|-------|----------------------------------------------------|
| Domain Name          | 可变    | 资源域名，包含这个资源记录涉及的域名，与查询报文中客户要查询的域名相对应，最大长度为 255 字节。 |
| Type                 | 16 比特 | 查询类型，与查询报文中的查询类型相对应。                               |
| Class                | 16 比特 | 查询等级，与查询报文中的查询等级相对应。                               |
| TTL                  | 32 比特 | 生存时间，该值代表解析的域名可与 IP 地址绑定存在的有效时间值。                  |
| Resource Data Length | 16 比特 | 说明资源数据字段中的长度，单位为字节。                                |
| Resource Data        | 可变    | 资源数据，这部分的长度是可变的。该信息字段格式由 Type 和 Class 字段决定。        |

5.3.4.3 静态域名解析

域名解析分为动态域名解析和静态域名解析。在解析域名时，首先采用静态域名解析的方法，如果静态解析不成功，再采用动态域名解析的方法。

静态域名解析是通过静态域名解析表进行的，即手动建立域名和 IP 地址之间的对应关系表，该表的作用类似于 Windows 9X 操作系统下的 hosts 文件，可以将一些常用的域名放入表中。当客户端需要域名所对应的 IP 地址时，即到静态域名解析表中去查找指定的域名，从而获得所对应的 IP 地址，提高域名解析的效率。

说明

将一些常用的域名放入静态域名解析表中，可以提高域名解析效率。



#### 5.3.4.4 动态域名解析

### DNS 服务器

动态域名解析需要专用的域名解析服务器，该服务器运行域名解析服务器程序，提供从域名到 IP 地址的映射关系，负责接收客户提出的域名解析请求。

根域名服务器包含有关根和顶级域的信息。DNS 要求每个域名解析服务器至少要知道一个根域名解析服务器的地址。客户也必须知道如何与至少一个域名解析服务器取得联系。

域名解析服务器接收到客户端提出的域名解析请求后，首先判断请求的域名是否处于自己被授权管理的子域里。如果是，就查询数据库把域名转换为 IP 地址，并把转换结果发送给客户端。如果域名解析服务器不能解析出域名，它就根据客户在查询报文中所指明的解析方式来进行下一步操作。

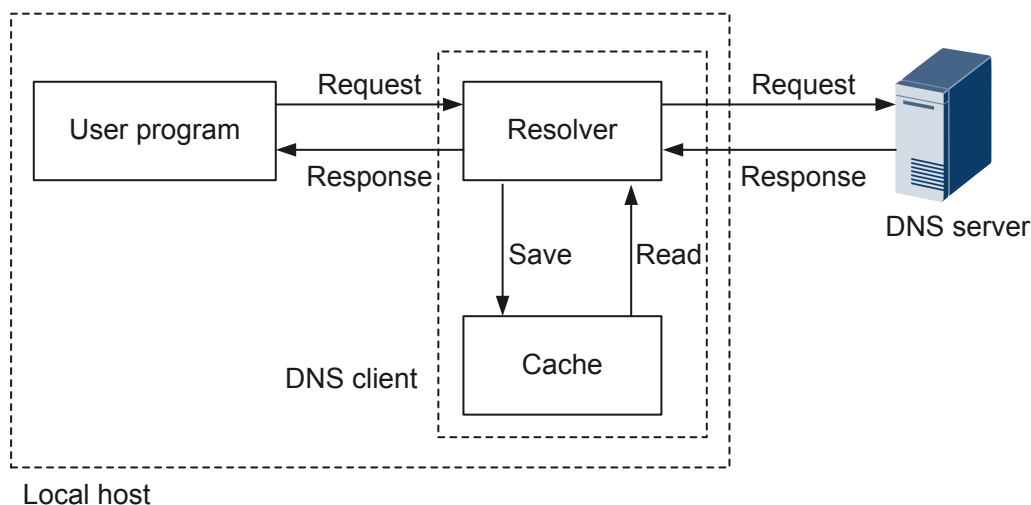
有以下两种域名解析方式：

- 递归解析  
域名解析服务器和能解析该域名的服务器联系，并将查询结果即域名所对应的 IP 地址返回给客户端。
- 迭代解析  
如果该域名解析服务器不能提供解析结果，会在给客户端的响应报文中指明客户端应联系的下一个域名解析服务器。客户端会向指明的下一个域名解析服务器再次发出查询请求。

用户程序对域名服务器（DNS server）的访问是通过 DNS 客户端（DNS client）的一个地址解析器（Resolver）来完成的。

用户程序、解析器和域名服务器以及解析器上的缓存区关系如图 5-11 所示，其中解析器和缓存区集成在一起构成 DNS Client，它的作用是接受用户程序的 DNS 咨询，并对其做出反应。一般来说，用户程序、缓存区和解析器是在同一台主机上，域名服务器和它们在不同的主机上。

图 5-11 动态域名解析



## 动态域名解析的工作过程

动态域名解析的工作过程如图 5-11 所示：

1. 用户程序首先向 DNS client 发出请求。
2. DNS client 收到请求后，首先查询本机数据库/缓存，如果没有发现所要查找的映射项，就向域名服务器发送查询报文。
3. 域名服务器收到查询报文后，首先判断请求的域名是否处于自己被授权管理的子域里，再根据不同的判断结果，向 DNS client 发送相应的响应报文。
4. DNS client 收到响应后，解析域名服务器发回来的响应报文，并根据响应报文的内容决定下一步的操作。

## 高速缓存

如果解析者把每个非本地名字的解析请求都送到根域名解析服务器，会造成很大的查询开销。为了降低非本地域名查询的开销，域名解析服务器使用高速缓存。对于每次动态解析成功的域名和 IP 地址的映射关系，存放在内存的动态域名缓存区中，下一次查询相同域名的时候，就可以直接从缓存区中读取，不用再向根域名解析服务器请求了。每个域名解析服务器都在本机的高速缓存里维护着最近用过的域名的记录，并且同时保存从哪个服务器获得的此域名映射。

但是当域名与 IP 地址的映射关系发生变化时，高速缓存中的信息就不再正确。为保证缓存信息的正确性，域名解析服务器在 DNS 响应报文中，会指定一个 TTL 值，该值代表解析的域名与 IP 地址绑定存在的有效时间值。客户端在缓存区中保存的映射关系在一段时间后被老化删除，保证能够及时从域名解析服务器得到最新的内容。对于安装不同操作系统的域名解析服务器，老化时间都有其相对应的值。客户端从 DNS 协议报文中获得老化时间。

同样，主机中也存在一个高速缓存，用来维护自己最近使用过的域名与 IP 地址的绑定。只有主机在未能从缓存中找到要解析的域名时才会使用域名解析服务器进行查询。

## 域名后缀

动态域名解析支持域名后缀列表功能。用户可以预先设置一些域名后缀，在域名解析的时候，用户只需要输入域名的部分字段，系统会自动将输入域名加上不同的后缀进行解析。

例如，用户想查询域名“test.com”，那么可以在后缀列表中配置 com，然后输入“test”，系统会自动将输入域名与后缀连接成“test.com”进行查询。使用域名后缀的时候，可能会出现以下几种情况：

- 如果用户输入的域名中没有“.”  
比如“test”，系统认为这是一个主机名，会首先加上域名后缀进行查询。如果所有域名都查询失败，最后用最初输入的域名进行查询。
- 如果用户输入的域名中间有“.”  
比如“www.test”，系统直接用它进行查询。如果查询失败，再依次加上各个后缀进行查询。
- 如果用户输入的域名最后有“.”  
比如“test.com.”，系统首先会将输入域名最后的“.”去掉，用剩下的部分查询，如果没有成功，会再去匹配域名列表进行查询。

### 5.3.4.5 DNS Proxy

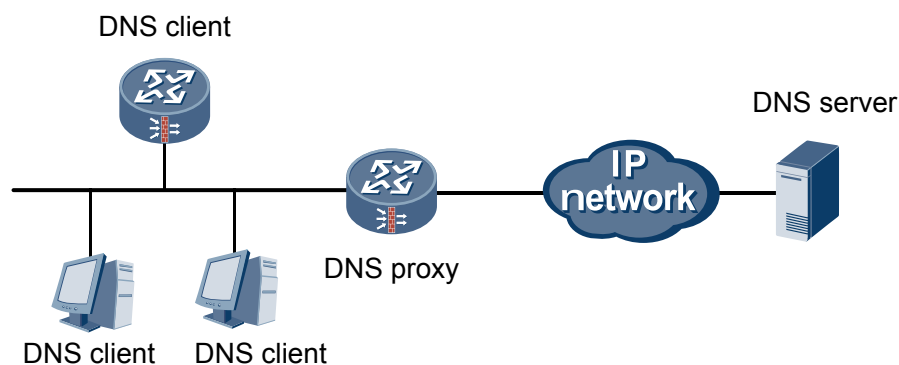
DNS 代理（DNS proxy）用来在 DNS client 和 DNS server 之间转发 DNS 请求和应答报文。

局域网内的 DNS client 把 DNS proxy 当作 DNS server，将 DNS 请求报文发送给 DNS proxy。DNS proxy 将该请求报文转发到真正的 DNS server，并将 DNS server 的应答报文返回给 DNS client，从而实现域名解析。

使用 DNS proxy 功能后，当 DNS server 的地址发生变化时，只需改变 DNS proxy 上的配置，无需改变局域网内每个 DNS client 的配置，从而简化了网络管理。

DNS proxy 的典型应用环境如图 5-12 所示。

图 5-12 DNS proxy 典型组网应用



DNS proxy 的工作过程如下：

1. DNS client 把 DNS proxy 当作 DNS server，将 DNS 请求报文发送给 DNS proxy，即请求报文的目地址为 DNS proxy 的 IP 地址。
2. DNS proxy 收到请求报文后，首先查找本地的静态域名解析表，如果静态域名解析表中存在请求的信息，则 DNS proxy 直接通过 DNS 应答报文，将域名解析结果返回给 DNS client。
3. 如果静态域名解析表中没有请求的信息，则 DNS proxy 将报文转发给 DNS server，通过 DNS server 进行域名解析。
4. DNS proxy 收到 DNS server 的应答报文后，将报文转发给 DNS client。DNS client 利用域名解析的结果进行相应的处理。

### 5.3.4.6 DDNS

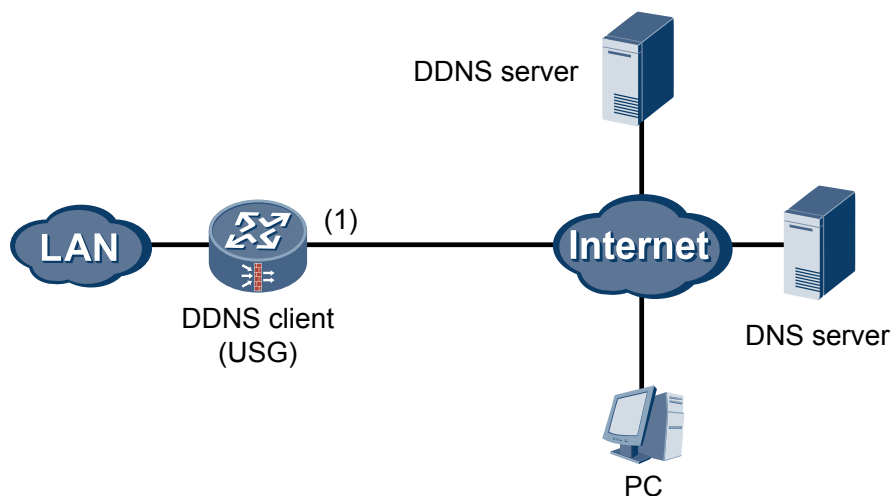
TCP/IP 提供了基于 IP 地址来寻址，并同时设计了一种字符串形式的主机命名机制，这就是所谓的域名系统 DNS（Domain Name System）。但是，DNS 仅提供了域名和 IP 地址之间的静态对应关系，当节点的 IP 地址发生变化时，DNS 无法动态地更新域名和 IP 地址的对应关系。此时，如果某台设备仍然使用域名访问该节点，只能获得错误的 IP 地址，导致访问失败。

由于 DNS 存在的上述局限性，产生了 DDNS（Dynamic Domain Name System）技术。

DDNS 通过动态更新 DNS 服务器上域名和 IP 地址之间的对应关系，保证通过域名解析后得到正确的 IP 地址。

DDNS 的典型组网应用如图 5-13 所示。

图 5-13 DDNS 典型组网应用



#### 说明

- 部署 DDNS 需要 DDNS 服务提供商 (DDNS server) 的支持。目前设备支持以下的 DDNS 服务提供商: [www.3322.org](http://www.3322.org)、[dyndns.org](http://dyndns.org)、[freedns.afraid.org](http://freedns.afraid.org)、[zoneedit.com](http://zoneedit.com) 和 [no-ip.com](http://no-ip.com)。
- 由于 DDNS server 通常部署在 Internet, 所以在应用 DDNS 时, 需要保证 DDNS client (即 USG) 能正常访问 Internet。

图 5-13 中, DDNS client (即 USG) 的接口 (1) 的 IP 地址从网络运营商动态获得, 每次获得的 IP 地址不同, 因此 PC 需要通过域名访问 USG。而传统的 DNS 无法动态更新域名和 IP 地址的对应关系, 这就会导致 PC 访问 USG 失败。此时, 可通过部署 DDNS 来解决该问题。

部署 DDNS 后, 当接口 (1) 的 IP 地址变化时, USG 会自动向 DDNS server 发送更新域名和 IP 地址对应关系的 DDNS 更新请求。DDNS server 处理该请求后, 会将更新后的域名和 IP 地址对应关系发送给 DNS server, 从而更新了 DNS server 上的信息。此时, 当 PC 再次通过域名访问 USG 时, 即可获取到 USG 正确的 IP 地址, 完成正常访问。

## 5.4 DHCP

DHCP 采用客户端/服务器通信模式, 由客户端向服务器提出配置申请 (包括 IP 地址、默认网关、DNS 服务器、WINS 服务器等参数), 服务器根据策略返回相应配置信息。

### 5.4.1 介绍

#### 定义

DHCP (Dynamic Host Configuration Protocol) 是一种用于集中对用户进行动态管理和配置的技术。

DHCP 采用客户端/服务器通信模式, 由客户端 (DHCP Client) 向服务器 (DHCP Server) 提出配置申请, 服务器返回为客户端分配的配置信息 (包括 IP 地址、缺省网

关、DNS 服务器、WINS 服务器等参数），可以实现 IP 地址动态分配，以及其他网络参数的集中配置管理。

## 目的

随着网络规模的不断扩大和网络复杂度的不断提高，计算机的数量经常超过可供分配的 IP 地址数量。同时随着便携机及无线网络的广泛使用，计算机的位置也经常变化，相应的 IP 地址也必须经常更新。为了动态合理地分配 IP 地址给主机使用，需要用到 DHCP。

DHCP 是在 BOOTP（BOOTstrap Protocol）基础上发展而来，但 BOOTP 运行在相对静态（每台主机都有固定的网络连接）的环境中，管理员为每台主机配置专门的 BOOTP 参数文件，该文件会在相当长的时间内保持不变。而 DHCP 从两方面对 BOOTP 进行了扩展：

- DHCP 加入了对重新使用的网络地址的动态分配和附加配置选项的功能，可使计算机仅用一个消息就获取它所需要的所有配置信息。
- DHCP 允许计算机快速、动态地获取 IP 地址，而不是静态为每台主机指定地址。

DHCP 技术保证了 IP 地址的合理分配问题，从而避免了 IP 地址的浪费，提高了整网的 IP 地址使用率。

## 受益

企业受益：DHCP 技术实现用户地址和配置信息的动态分配和集中管理，使企业可以快速、动态地为企业用户分配和管理地址，避免繁琐的手工配置，可以快速适应网络的变化。

## 5.4.2 参考标准和协议

与 DHCP 相关的参考标准与协议如下：

- RFC 1534: Interoperation Between DHCP and BOOTP
- RFC 1542: Clarifications and Extensions for the Bootstrap Protocol
- RFC 2131: Dynamic Host Configuration Protocol
- RFC 2132: DHCP Options and BOOTP Vendor Extensions
- RFC 2241: DHCP Options for Novell Directory Services
- RFC 2485: DHCP Option for The Open Group's User Authentication Protocol
- RFC 2563: DHCP Option to Disable Stateless Auto-Configuration in IPv4 Clients
- RFC 2610: DHCP Options for Service Location Protocol
- RFC 2937: The Name Service Search Option for DHCP
- RFC 2939: Procedures and IANA Guidelines for Definition of New DHCP Options and Message Types
- RFC 3004: The User Class Option for DHCP
- RFC 3011: The IPv4 Subnet Selection Option for DHCP
- RFC 3046: DHCP Relay Agent Information Option
- RFC 3361: Dynamic Host Configuration Protocol (DHCP-for-IPv4) Option for Session Initiation Protocol (SIP) Servers
- RFC 3397: Dynamic Host Configuration Protocol (DHCP) Domain Search Option

- RFC 3442: The Classless Static Route Option for Dynamic Host Configuration Protocol (DHCP) version 4
- RFC 3495: Dynamic Host Configuration Protocol (DHCP) Option for CableLabs Client Configuration

### 5.4.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 5.4.4 应用场景

#### 5.4.4.1 DHCP 服务器

#### 应用环境

在以下场合通常利用 DHCP 服务器来完成 IP 地址分配：

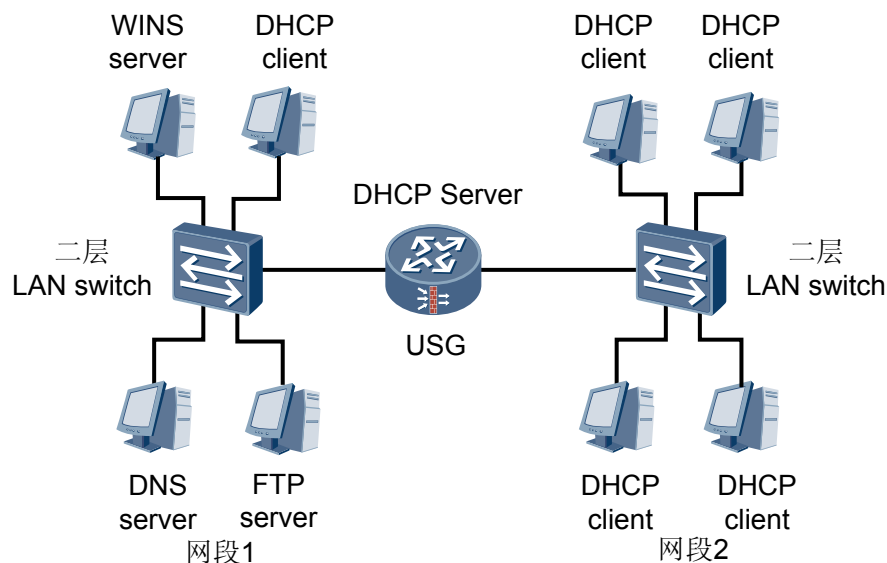
- 网络规模较大，手工配置需要很大的工作量。
- 网络中需要固定 IP 地址的主机比较少，大部分主机可以不使用固定 IP 地址。
- 网络中主机数目大于该网络可分配的 IP 地址数量，无法给每个主机分配一个固定 IP 地址，主机必须通过 DHCP 服务共享较少的 IP 地址资源。

#### 典型应用

DHCP 服务器的典型应用有以下两种：

- DHCP 客户端与 DHCP 服务器在同一网段  
如图 5-14 所示，USG 作为 DHCP 服务器通过二层交换机（或 HUB）连接客户端。

图 5-14 DHCP 服务器典型组网应用 1

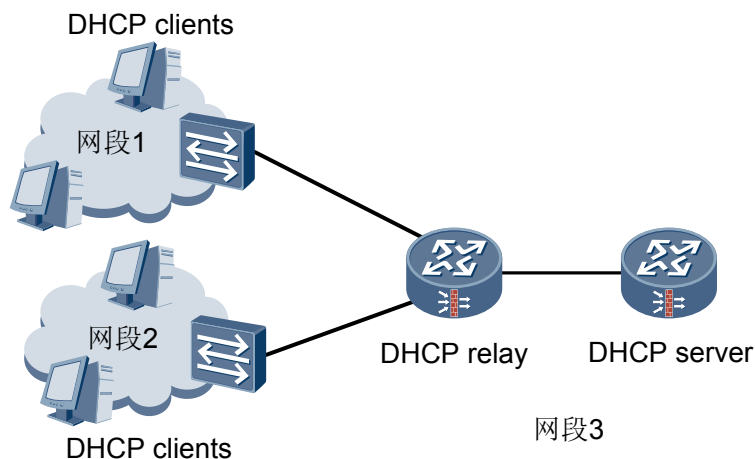


- DHCP 客户端与 DHCP 服务器不在同一网段

当 DHCP 客户端与 DHCP 服务器不在同一网段时，DHCP 服务器需要与 DHCP 中继配合来实现跨网段动态分配 IP 地址。

如图 5-15 所示，DHCP 客户端和 DHCP 服务器不在同一网段。在客户端所在的网段上部署 DHCP 中继，使得客户端通过 DHCP 中继和 DHCP 服务器通信。

图 5-15 DHCP 服务器典型组网应用 2



DHCP 服务器可以提供以下服务：

- 可以在 DHCP 服务器上配置地址池、租期，实现为 DHCP 客户端动态分配 IP 地址。
- 可以针对已经配置了固定 IP 地址的设备（例如 FTP 服务器），在 DHCP 服务器上定义保留 IP 地址。
- 可以实现静态分配 IP 地址，为服务器及特殊主机分配固定 IP 地址。



- 可以在 DHCP 服务器上配置 IP 地址探测，防止服务器重复分配 IP 地址。
- 可以在 DHCP 服务器上为客户端集中配置 DNS 服务器、缺省网关、WINS 服务器等网络参数。

#### 5.4.4.2 DHCP 中继

##### 应用环境

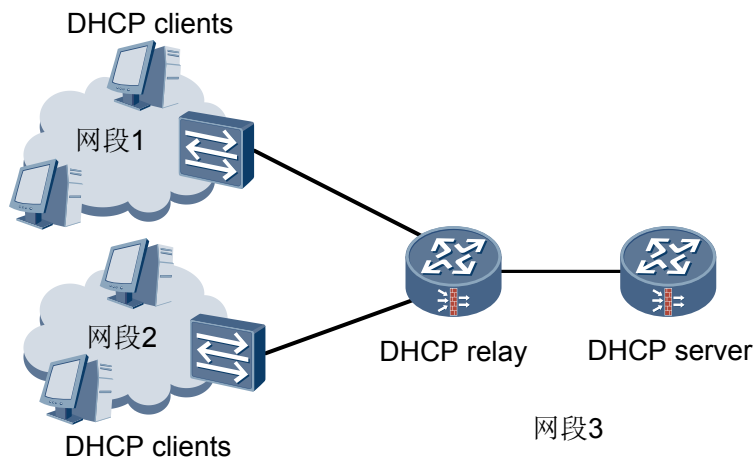
由于在 IP 地址动态获取过程中 DHCP 客户端采用广播方式发送请求报文，因此 DHCP 只适用于 DHCP 客户端和服务端处于同一个网段内的情况，不能跨网段。为进行动态主机配置，需要在每个网段设置一个 DHCP 服务器，这显然是很不经济的。

DHCP 中继（DHCP Relay）功能的引入解决了这一难题。客户端可以通过 DHCP 中继与其他网段的 DHCP 服务器通信，最终取得合法的 IP 地址。这样，多个网段的 DHCP 客户端可以使用同一个 DHCP 服务器，既节省了成本，又便于进行集中管理。

##### 典型应用

如图 5-16 所示，DHCP 客户端和 DHCP 服务器不在同一网段。在客户端所在的网段上部署 DHCP 中继，使得客户端通过 DHCP 中继和 DHCP 服务器通信，并从 DHCP 服务器获取 IP 地址等配置信息。

图 5-16 DHCP 中继典型应用



#### 5.4.4.3 DHCP 客户端

##### 应用环境

由于 IP 地址资源有限，边缘网络设备常常无法从运营商处获得固定 IP 地址，此时可以考虑将网络设备作为 DHCP 客户端，从 DHCP 服务器处动态获取 IP 地址。

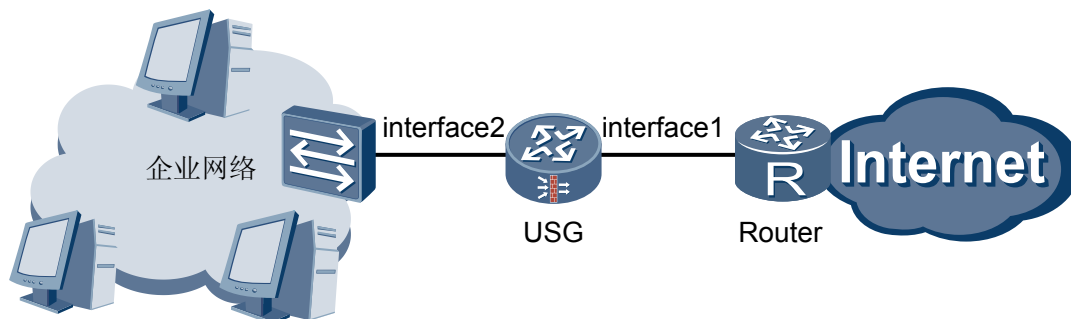
##### 典型应用

如图 5-17 所示，某写字楼通过 Router 接入 Internet，同时 Router 又作为 DHCP 服务器为写字楼下企业用户分配 IP 地址。USG 作为该写字楼下某小型企业的网关，在 USG



的接口 1（interface1）开启 DHCP 客户端功能后，可以从 DHCP 服务器动态获得 IP 地址等参数，为企业用户提供上网服务。

图 5-17 DHCP 客户端典型应用



当 USG 作为 DHCP 客户端时，可以提供以下服务：

- 动态获取 IP 地址、出口网关、DNS 服务器、域名后缀、静态路由等网络参数
- 主动刷新 IP 地址、出口网关、DNS 服务器、域名后缀、静态路由等网络参数

另外，DHCP 客户端和服务端可以在一个设备上同时开启，只要应用在不同接口上。例如图 5-17，在 USG 连接写字楼所在网络的接口（interface1）上开启 DHCP 客户端功能，使用 DHCP 协议从 DHCP Server 获取 IP 地址和配置信息。而在连接企业网络的接口（interface2）上配置 DHCP 服务器功能，为企业网内部 PC 分配 IP 地址。

## 5.4.5 原理描述

### 5.4.5.1 DHCP 报文

#### DHCP 报文类型

DHCP 报文分为 8 种类型，DHCP 服务器和客户端之间通过这 8 种类型的报文进行通信：

- DHCPDISCOVER：DHCP 客户端首次登录网络时进行 DHCP 过程的第一个报文，用来寻找 DHCP 服务器。
- DHCPOFFER：DHCP 服务器用来响应 DHCPDISCOVER 报文，此报文携带了各种配置信息。
- DHCPREQUEST：此报文用于以下三种用途。
  - 客户端初始化后，发送广播的 DHCPREQUEST 报文来回应服务器的 DHCPOFFER 报文。
  - 客户端重启初始化后，发送广播的 DHCPREQUEST 报文来确认先前被分配的 IP 地址等配置信息。
  - 当客户端已经和某个 IP 地址绑定后，发送单播的 DHCPREQUEST 报文来延长 IP 地址的租期。

- DHCPACK：服务器对客户端的 DHCPREQUEST 报文的确认响应报文，客户端收到此报文后，才真正获得了 IP 地址和相关的配置信息。
- DHCPNAK：服务器对客户端的 DHCPREQUEST 报文的拒绝响应报文，比如服务器对客户端分配的 IP 地址已超过使用租借期限或者客户端移到了另一个新的网络。
- DHCPDECLINE：当客户端发现服务器分配给它的 IP 地址发生冲突时会通过发送此报文来通知服务器，并且会重新向服务器申请地址。
- DHCPRELEASE：客户端可通过发送此报文主动释放服务器分配给它的 IP 地址，当服务器收到此报文后，可将这个 IP 地址分配给其他的客户端。
- DHCPINFORM：客户端已经获得了 IP 地址，发送此报文的目的是为了从服务器获得其他的一些网络配置信息，比如网关地址、DNS 服务器地址等。

DHCP 报文格式

以上 8 种类型的 DHCP 报文的格式相同，只是某些字段的取值不同。DHCP 继承了 BOOTP 的报文结构，采用 UDP 进行封装，具体报文格式如图 5-18 所示。

图 5-18 DHCP 的报文格式

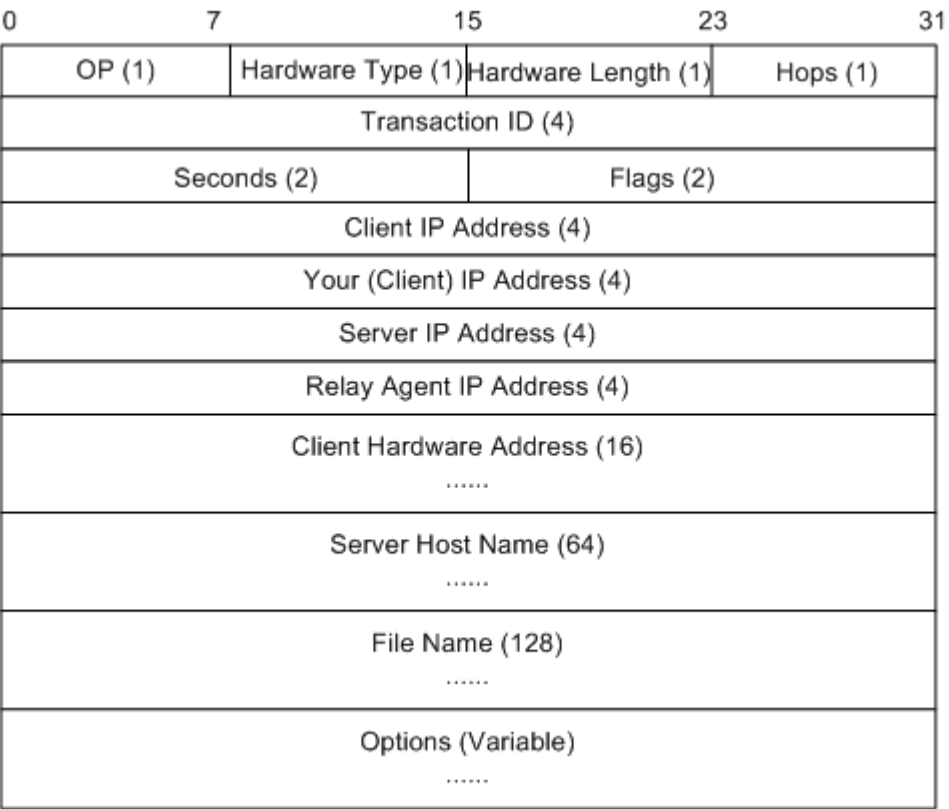


图 5-18 中括号里的数字表示字段的长度，单位是字节。DHCP 报文中各个字段的含义如表 5-11 所示。

表 5-11 DHCP 报文中各个字段的含义

| 字段                       | 长度   | 含义                                                                                                                                                                          |
|--------------------------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| OP                       | 1 字节 | 表示报文的操作类型：<br>● 1：客户端请求报文<br>● 2：服务器响应报文                                                                                                                                    |
| Hardware Type            | 1 字节 | 表示网络硬件地址的类型。系统目前只对以太网支持，该值为 1。                                                                                                                                              |
| Hardware Length          | 1 字节 | 表示网络硬件地址的长度，单位是字节。系统目前只对以太网支持，该值为 6。                                                                                                                                        |
| Hops                     | 1 字节 | 表示当前的 DHCP 报文经过的 DHCP 中继的数目。该字段由客户端设置为 0，每经过一个 DHCP 中继时，该字段加 1。此字段的作用是限制 DHCP 报文所经过的 DHCP 中继数目。<br><b>说明</b><br>服务器和客户端之间的 DHCP 中继不能超过 4 次，也就是 Hops 值不能大于 4，否则 DHCP 报文将被丢弃。 |
| Transaction ID           | 4 字节 | 表示客户端每次发送 DHCP 请求报文时选择的随机数，使服务器的回复与客户端的报文相关联。                                                                                                                               |
| Seconds                  | 2 字节 | 表示客户端在开始 DHCP 请求后逝去的秒数，单位是秒。                                                                                                                                                |
| Flags                    | 2 字节 | 表示标志字段。目前只有最高位才有意义，其余的位均被置为 0。最高位为广播响应标志位，用来表示 DHCP 服务器响应报文时采用单播还是广播方式发送，取值含义如下：<br>● 0：采用单播方式<br>● 1：采用广播方式                                                                |
| Client IP Address        | 4 字节 | 表示客户端的 IP 地址。可以是服务器分配给客户端的 IP 地址或者客户端已有的 IP 地址。客户端在初始化状态时没有 IP 地址，此字段为 0.0.0.0。<br><b>说明</b><br>IP 地址 0.0.0.0 仅在采用 DHCP 方式的系统启动时允许本主机利用它进行临时的通信，并且永远不是有效目的地址。               |
| Your (Client) IP Address | 4 字节 | 表示服务器分配给客户端的 IP 地址。当服务器响应客户端的 DHCP 请求时，将分配给客户端的 IP 地址填入此字段。                                                                                                                 |
| Server IP Address        | 4 字节 | 表示服务器 IP 地址。                                                                                                                                                                |

| 字段                      | 长度     | 含义                                                                                                                                                                                                                                                                               |
|-------------------------|--------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Relay Agent IP Address  | 4 字节   | 表示第一个 DHCP 中继的 IP 地址。当客户端发出 DHCP 请求时，如果服务器和客户端不在同一个网络中，那么第一个 DHCP 中继在转发这个 DHCP 请求报文时会把自己的 IP 地址填入此字段。服务器会根据此字段来判断出网段地址，从而选择为用户分配地址的地址池。服务器还会根据此地址将响应报文发送给此 DHCP 中继，再由 DHCP 中继将此报文转发给客户端。<br><b>说明</b><br>若在到达 DHCP 服务器前经过了不止一个 DHCP 中继，那么第一个 DHCP 中继后的中继不会改变此字段，只是把 Hops 的数目加 1。 |
| Client Hardware Address | 16 字节  | 表示客户端的 MAC 地址，此字段与前面的“Hardware Type”和“Hardware Length”保持一致。当客户端发出 DHCP 请求时，将自己的硬件地址填入此字段。对于以太网，当“Hardware Type”和“Hardware Length”分别为“1”和“6”时，此字段必须填入 6 字节的以太网 MAC 地址。                                                                                                            |
| Server Host Name        | 64 字节  | 表示客户端获取配置信息的服务器名字。此字段由 DHCP 服务器填写，是可选的。如果填写，必须是一个以 0 结尾的字符串。                                                                                                                                                                                                                     |
| File Name               | 128 字节 | 表示客户端的启动配置文件名。此字段由 DHCP 服务器填写，是可选的，如果填写，必须是一个以 0 结尾的字符串。                                                                                                                                                                                                                         |
| Options                 | 可变     | 表示 DHCP 的可选变长选项字段，至少为 312 字节。此字段包含了服务器分配给客户端的配置信息，如缺省网关 IP 地址、DNS 服务器的 IP 地址、WINS 服务器的 IP 地址、客户端可以使用 IP 地址的有效租期等信息。对 Options 字段的详细解释请参见“Options 字段的格式和作用”。                                                                                                                       |

Options 字段的格式和作用

Options 字段的格式

Options 字段的格式如图 5-19 所示。

图 5-19 Options 字段的格式

|      |        |       |
|------|--------|-------|
| 0    | 7      | 15    |
| Code | Length | Value |

Options 字段由 Code、Length 和 Value 三部分组成。这三部分的表示含义如表 5-12 所示。

**表 5-12 Options 各字段的含义**

| 字段     | 长度                | 含义                 |
|--------|-------------------|--------------------|
| Code   | 1 字节              | 表示标识号，唯一标识后面的信息内容。 |
| Length | 1 字节              | 表示后面信息内容的长度。       |
| Value  | 其长度为 Length 字段所指定 | 表示信息内容。            |

一些比较常用的 Options 如表 5-13 所示。

**表 5-13 Options 举例**

| Code 值 | Length 值           | Value 内容             | 说明                                                                                 |
|--------|--------------------|----------------------|------------------------------------------------------------------------------------|
| 1      | 4                  | 子网掩码                 | Option 1，子网掩码选项，用来指定为客户端分配的子网掩码。                                                   |
| 3      | IP 地址长度，必须为 4 的整数倍 | 客户端的默认网关 IP 地址       | Option 3，路由器选项，用来指定为客户端分配的默认网关地址，可以指定多个网关地址。                                       |
| 6      | IP 地址长度，必须为 4 的整数倍 | 客户端的 DNS 服务器的 IP 地址  | Option 6，DNS 服务器选项，用来指定为客户端分配的 DNS 服务器地址，可以指定多个 DNS 服务器地址。                         |
| 33     | 必须为 8 的整数倍         | 目的 IP 地址和下一跳地址       | Option 33，静态路由选项。该选项中包含一组有分类静态路由（即目的地址的掩码固定为自然掩码，不能划分子网），客户端收到该选项后，将在路由表中添加这些静态路由。 |
| 44     | IP 地址长度，必须为 4 的整数倍 | 客户端的 WINS 服务器的 IP 地址 | Option 44，WINS 服务器选项，用来指定为客户端分配的 WINS 服务器地址，可以指定多个 WINS 服务器地址。                     |
| 50     | 4                  | 客户端请求的 IP 地址         | Option 50，客户端请求的 IP 地址选项。客户端通过在 DHCPDISCOVER 报文中添加该选项来指明希望获取的 IP 地址。               |

| Code 值 | Length 值   | Value 内容                                      | 说明                                                                                                                                                                                                                                                                                                                                             |
|--------|------------|-----------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 53     | 1          | 1 ~ 8                                         | Option 53, DHCP 消息类型选项, 用来标识 DHCP 消息的类型。Vaule 内容与报文类型的对应关系如下: <ul style="list-style-type: none"><li>● 1: DHCPDISCOVER</li><li>● 2: DHCPOFFER</li><li>● 3: DHCPREQUEST</li><li>● 4: DHCPDECLINE</li><li>● 5: DHCPACK</li><li>● 6: DHCPNAK</li><li>● 7: DHCPRELEASE</li><li>● 8: DHCPINFORM</li></ul> 具体的报文类型说明请参见“ <a href="#">DHCP 报文类型</a> ”。 |
| 54     | 4          | DHCP 服务器的 IP 地址                               | Option 54, DHCP 服务器选项, 用来指定 DHCP 服务器的 IP 地址。                                                                                                                                                                                                                                                                                                   |
| 121    | 必须为 8 的整数倍 | 目的描述符和下一跳地址 (其中, 目的描述符由子网掩码长度和目的网络地址组成, 长度可变) | Option 121, 无分类路由选项。该选项中包含一组无分类静态路由 (即目的地址的掩码为任意值, 可以通过掩码来划分子网), 客户端收到该选项后, 将在路由表中添加这些静态路由。                                                                                                                                                                                                                                                    |

**Options 字段的作用**

DHCP 报文中的 Options 字段可以用来存放某些普通协议中没有定义的控制信息和参数。如果用户在 DHCP 服务器端配置了 Options, DHCP 客户端在申请 IP 地址的时候, 会通过服务器端回应的 DHCP 报文获得 Options 字段中的配置信息。

根据 Options 选项功能的不同, 此字段的作用对象也不同。比如 Option 77 用于 DHCP 客户端, 用于识别用户或应用程序所属的类型, 根据 Option 字段中所携带的用户类型 (User Class), DHCP 服务器选择适当的地址池为客户端分配 IP 地址以及相关配置参数。Option 77 一般在客户端由用户进行配置, 而不必用命令在服务器端配置。

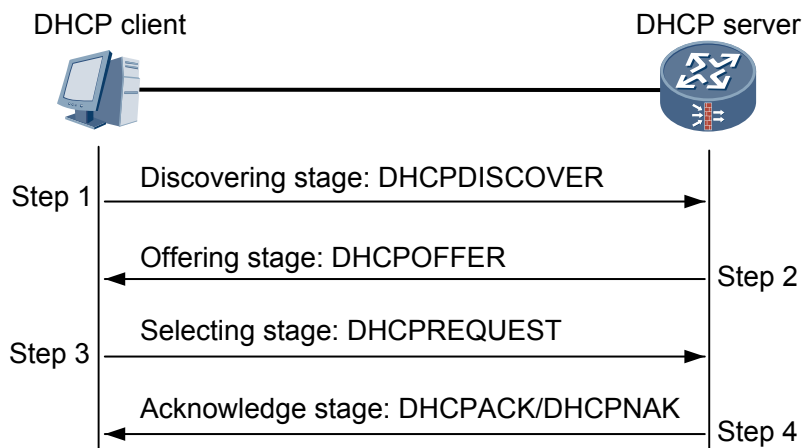
关于 DHCP Option Code 的含义和用法, 请参见 RFC 2132。

**5.4.5.2 DHCP 服务器和客户端的交互过程**

**DHCP 客户端首次登录网络**

DHCP 客户端首次登录网络时, 主要通过四个阶段与 DHCP 服务器建立联系。如[图 5-20](#)所示。

图 5-20 DHCP 客户端首次登录网络时获取 IP 地址的过程



1. 发现阶段

即 DHCP 客户端寻找 DHCP 服务器的阶段。客户端通过 UDP 知名端口 68 以广播方式发送 DHCPDISCOVER 报文，只有 DHCP 服务器才会进行响应。

2. 提供阶段

即 DHCP 服务器提供 IP 地址的阶段。DHCP 服务器接收到客户端的 DHCPDISCOVER 报文后，从 IP 地址池中挑选一个尚未分配的 IP 地址准备分配给客户端，通过 UDP 知名端口 67 向该客户端发送包含待分配的 IP 地址、租期和其他配置信息（如网关地址、DNS 服务器地址等）的 DHCPOFFER 报文。

3. 选择阶段

即 DHCP 客户端选择 IP 地址的阶段。如果有多台 DHCP 服务器向该客户端发来 DHCPOFFER 报文，客户端只接受第一个收到的 DHCPOFFER 报文，然后以广播方式向各 DHCP 服务器回应 DHCPREQUEST 报文，该 DHCPREQUEST 报文中包含向所选定的 DHCP 服务器请求 IP 地址的内容。

4. 确认阶段

即 DHCP 服务器确认所提供 IP 地址的阶段。当 DHCP 服务器收到 DHCP 客户端回应的 DHCPREQUEST 报文后：

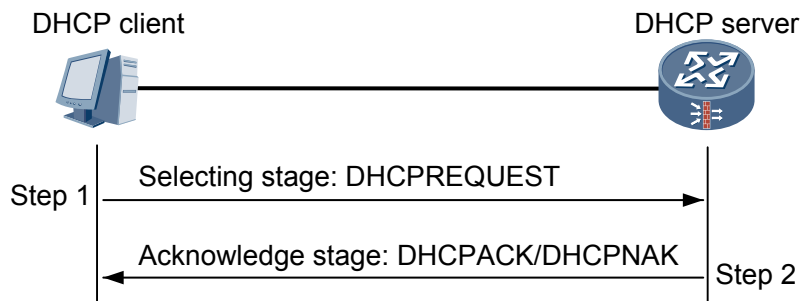
- 如果是客户端选定的服务器，则该服务器根据 DHCPREQUEST 报文中携带的 MAC 地址来查找有没有相应的租约记录。
  - 如果有，则向客户端发送包含它所提供的 IP 地址和其他配置信息的 DHCPACK 确认报文。客户端收到 DHCPACK 确认报文后，会以广播的方式发送免费 ARP 报文，探测是否有主机使用服务器分配的 IP 地址，如果在规定的时间内没有回应，客户端才使用此 IP 地址。否则，客户端将发送 DHCPDECLINE 报文给服务器，通知服务器该 IP 地址不可用，并重新申请 IP 地址。
  - 如果没有或由于某些原因无法正常分配 IP 地址，则向客户端发送 DHCPNAK 报文作为应答，通知客户端无法分配合适的 IP 地址。客户端需要重新发送 DHCPDISCOVER 报文来请求新的 IP 地址。
- 如果不是客户端选定的服务器，则这些服务器的本次未分配出的 IP 地址仍可分配给其他客户端。



## DHCP 客户端再次登录网络

DHCP 客户端首次正确登录网络后，再登录网络时不需要在发送 DHCPDISCOVER 报文，主要通过以下几个步骤与 DHCP 服务器建立联系，如图 5-21 所示。

图 5-21 DHCP 客户端再次登录网络时获取 IP 地址的过程



1. DHCP 客户端直接发送广播的 DHCPREQUEST 报文（在此 DHCPREQUEST 报文里包含上次分配到的 IP 地址等信息）。
2. DHCP 服务器收到 DHCPREQUEST 报文后，判断客户端是否可以使用申请的 IP 地址：
  - 如果 DHCP 客户端申请的地址没有被分配，则返回 DHCPACK 报文，通知该 DHCP 客户端继续使用原来的 IP 地址。客户端收到 DHCPACK 报文后进入到绑定状态。此时客户端可以继续使用原来的 IP 地址进行通信。
  - 如果申请的 IP 地址无法再分配给该 DHCP 客户端使用（例如已分配给其他客户端使用），DHCP 服务器将返回 DHCPNAK 报文。客户端收到 DHCPNAK 报文后，进入到初始化状态。此时客户端会重新发送 DHCPDISCOVER 报文请求新的 IP 地址。接下来 DHCP 客户端和 DHCP 服务器交互的过程和 **DHCP 客户端首次登录网络** 时一样。

## DHCP 客户端延长 IP 地址的租用有效期

DHCP 服务器分配给客户端的动态 IP 地址通常有一定的租借期限，期满后服务器会收回该 IP 地址。如果 DHCP 客户端希望继续使用该地址，需要更新 IP 租约（如延长 IP 地址租约）。

DHCP 客户端延长 IP 地址的租用有效期的具体过程请参见 **5.4.5.5 租期与定时器**。

## DHCP 客户端主动释放 IP 地址

DHCP 客户端不再使用分配的 IP 地址时，会主动向 DHCP 服务器发送 DHCPRELEASE 报文，通知 DHCP 服务器释放 IP 地址的租约。DHCP 服务器会保留这个 DHCP 客户端的配置信息，以便该客户端重新申请地址时，重用这些参数。

### 5.4.5.3 DHCP 客户端与 DHCP 服务器通过 DHCP 中继的交互过程

DHCP 客户端在发送交互消息时是广播报文，只适用于 DHCP 客户端和服务器处于同一个子网内的情况，不可以跨网段工作。当 DHCP 服务器和客户端处于不同网段时，DHCP 中继处于二者之间承担中继服务，将 DHCP 协议报文跨网段中继到目的 DHCP 服务器（或客户端）。这种交互过程分首次申请和延长地址租期两种情况。

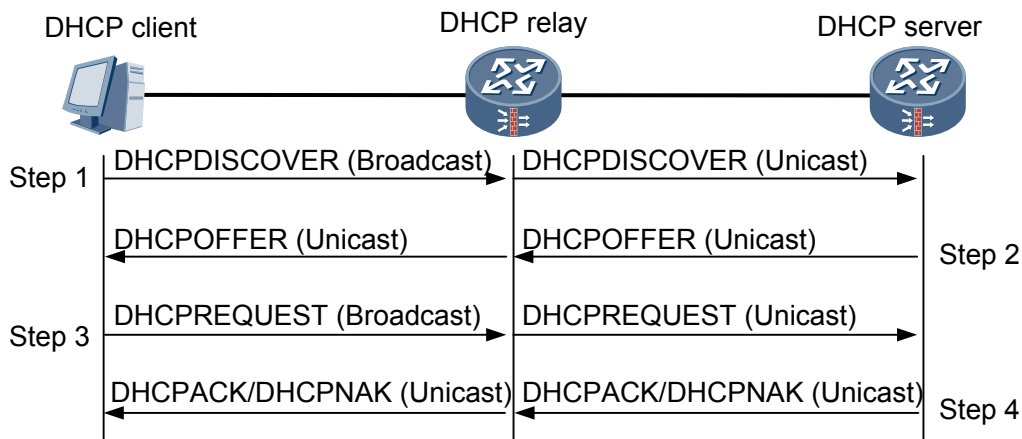


在实际网络环境中，DHCP 中继功能一般是在网络设备某个具体的接口上实现的。该接口应与 DHCP 客户端处于同一网段，能够在 DHCP 客户端和 DHCP 服务器之间透明转发 DHCP 报文。文中称这种接口为 DHCP 中继接口。

## DHCP 客户端通过 DHCP 中继首次申请地址

DHCP 客户端通过 DHCP 中继首次申请 IP 地址的过程如图 5-22 所示。

图 5-22 DHCP 客户端通过 DHCP 中继首次申请 IP 地址的过程



### 1. DHCP 客户端在本地网络广播 DHCPDISCOVER 报文。

如果本地网络没有 DHCP 服务器，则与本地网络相连的、带 DHCP 中继功能的网络设备收到该广播报文后，进行如下处理：

- 检查报文中的 Hops 字段，如果该值大于限定的最大跳数 4，则丢弃报文；否则继续下面的操作。
- 检查报文中的 Relay Agent IP Address 字段，如果该值为 0，则将该字段设置为接收报文的接口 IP 地址，即 DHCP 中继接口 IP 地址。以后，从该接口接收的所有 DHCP 报文都使用该 IP 地址。如果 Relay Agent IP Address 字段不是 0，则不修改该字段。
- 将报文中 Hops 字段值加 1，表明经过一次 DHCP 中继。
- 将报文的目的 IP 地址修改为 DHCP 服务器或下一个 DHCP 中继的 IP 地址。

经过上述处理后，DHCP 中继将报文单播转发给指定的、其他网络上的 DHCP 服务器或下一个 DHCP 中继。

#### 说明

设备作为 DHCP 中继时，发送给 DHCP 服务器或下一个 DHCP 中继的单播报文中的源 IP 地址为接收 DHCP 报文的接口 IP 地址，即 DHCP 中继接口的 IP 地址。

### 2. DHCP 服务器根据 Relay Agent IP Address 字段为客户端分配与 Relay Agent IP Address 处于同一网段的 IP 地址和其他配置信息等参数，并将 DHCPOFFER 报文单播发送给 Relay Agent IP Address 字段标识的 DHCP 中继。DHCP 中继收到该报文后，进行如下处理：

- 检查报文的 Relay Agent IP Address 字段，如果不是 DHCP 中继上的接口 IP 地址，则丢弃该报文，否则继续下面的操作。



说明

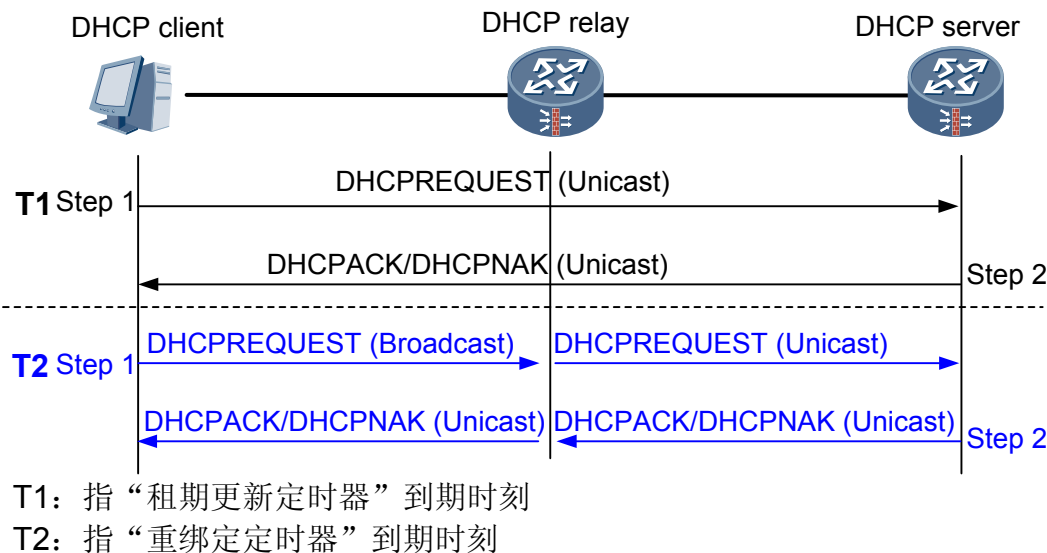
当网络中存在多个 DHCP 中继时，Relay Agent IP Address 字段标识的 DHCP 中继即为与 DHCP 客户端相连的第一个 DHCP 中继。其他的二级中继不会检查 Relay Agent IP Address 字段。

- b. 检查报文的广播标志位，如果广播标志位为 1，则将报文广播发送给 DHCP 客户端；否则将报文单播发送给 DHCP 客户端，其目的 IP 地址为报文中的 Your (Client) IP Address 字段值，MAC 地址为报文中的 Client Hardware Address 字段值。
3. DHCP 客户端向 DHCP 中继发送广播的 DHCPREQUEST 报文来回应服务器的 DHCPOFFER 报文。DHCP 中继收到该报文后，也进行第 1 步中的处理，然后将报文以单播的形式发送给 DHCP 服务器。
4. DHCP 服务器对客户端通过 DHCP 中继发送 DHCPACK 或者 DHCPNAK 报文。当 DHCP 中继收到该报文后，进行第 2 步中处理，然后将报文发送 DHCP 客户端。

## DHCP 客户端通过 DHCP 中继延长地址租期

DHCP 客户端通过 DHCP 中继延长地址租期的过程如图 5-23 所示。

图 5-23 DHCP 客户端通过 DHCP 中继延长地址租期的过程



- 当 DHCP 客户端的“租期更新定时器”到期时，DHCP 客户端通过以下步骤延长租期，报文不需要经过 DHCP 中继：
  1. DHCP 客户端向上次为它分配 IP 地址的 DHCP 服务器直接以单播的形式发送 DHCPREQUEST 报文。
  2. DHCP 服务器直接向客户端发送单播的 DHCPACK 或者 DHCPNAK 报文。如果 DHCP 客户端收到 DHCPACK 报文，则租期更新成功；如果收到 DHCPNAK 报文，则重新发起申请过程。
- 当 DHCP 客户端的“重绑定定时器”到期时，如果仍未收到 DHCP 服务器的应答，则 DHCP 客户端通过以下步骤延长租期，报文经过 DHCP 中继：



说明

DHCP 中继收到客户端或服务器发送的报文后，都会进行适当的处理，处理过程请参见 [DHCP 客户端通过 DHCP 中继首次申请地址](#)。

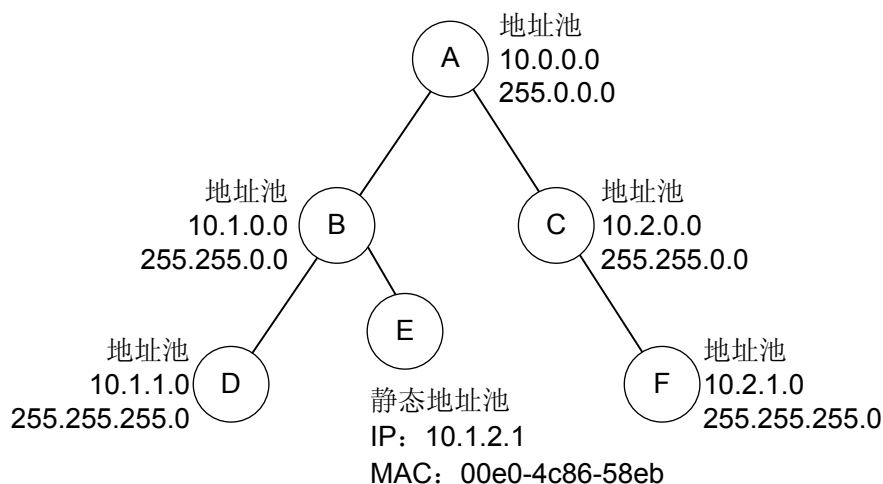
1. DHCP 客户端广播 DHCPREQUEST 报文。当 DHCP 中继收到此报文后，经过适当的处理，将此报文以单播的形式转发给 DHCP 服务器。
2. DHCP 服务器向客户端通过 DHCP 中继发送 DHCPACK 或者 DHCPNAK 报文。当 DHCP 中继收到该报文后，进行适当的处理，将此报文发送 DHCP 客户端。如果 DHCP 客户端收到 DHCPACK 报文，则租期更新成功；如果收到 DHCPNAK 报文，则重新发起申请过程。

#### 5.4.5.4 地址池

#### 地址池结构

DHCP 服务器中的地址池采用树状结构：树根是自然网段地址，分支是该网段的子网地址，叶节点是手工绑定的客户端地址。

图 5-24 地址池树状结构示意图



如图 5-24 所示：

- A 为 B、C 的父节点。反之，B 和 C 为 A 的子节点。B 和 C 互为兄弟节点。
- B 为 D、E 的父节点。反之，D 和 E 为 B 的子节点。D 和 E 互为兄弟节点。其中 E 为叶节点。
- C 为 F 的父节点。反之，F 为 C 的子节点。

这种树状结构实现了配置的继承性，即子网（子节点）配置继承自然网段（父节点）的配置，客户端（孙子节点）的配置继承子网（子节点）的配置。这样，对于一些通用参数（如 DNS 服务器地址），只需要在自然网段或者子网上配置即可。具体的继承情况如下：

1. 在父子关系建立时，子地址池将会继承父地址池的已有配置。
2. 在父子关系建立后，对父地址池进行的配置，子地址池是否会继承，则有以下两种情况：
  - 如果子地址池没有该项配置，则继承父地址池的配置。
  - 如果子地址池已有该项配置，则不会继承父地址池的配置。

## 地址池选取原则

DHCP 服务器从地址池中为客户端选择并分配 IP 地址及其相关参数。当客户端向服务器发出 DHCP 请求报文后，DHCP 服务器选择地址池的原则如下：

1. 如果存在将客户端 MAC 地址与 IP 地址静态绑定的地址池，则选择该地址池，并将静态绑定的 IP 地址分配给客户端。
2. 如果不存在静态绑定的地址池，则选择包含 DHCP 请求报文接收接口的 IP 地址（客户端与服务器在同一网段时）或 DHCP 请求报文中 Relay Agent IP Address 字段指定的 IP 地址（客户端与服务器不在同一网段，客户端通过 DHCP 中继获取 IP 地址时）、地址范围最小的地址池。如果该地址池中没有可供分配的 IP 地址，则服务器无法为客户端分配地址，服务器不会将父地址池中的 IP 地址分配给客户端。

例如，DHCP 服务器上配置了两个地址池，动态分配的 IP 地址范围分别是 1.1.1.0/24 和 1.1.1.0/25，如果接收到的 DHCP 请求报文的接口 IP 地址为 1.1.1.1/25，服务器将从 1.1.1.0/25 地址池中选择 IP 地址分配给客户端，1.1.1.0/25 地址池中如果没有可供分配的 IP 地址，则服务器无法为客户端分配地址；如果接收 DHCP 请求报文的接口 IP 地址为 1.1.1.130/25，服务器将从 1.1.1.0/24 地址池中选择 IP 地址分配给客户端。

## DHCP 服务器分配 IP 地址的优先次序

DHCP 服务器为客户端分配 IP 地址的优先次序如下：

1. 在 DHCP 地址池中，与客户端 MAC 地址静态绑定的 IP 地址。
2. DHCP 服务器记录的曾经分配给客户端的 IP 地址。
3. 客户端发送的 DHCPDISCOVER 报文中请求 Option 50（IP 地址选项）字段指定的 IP 地址。
4. 在 DHCP 地址池中，顺序查找可供分配的 IP 地址，最先找到的 IP 地址。
5. 如果在 DHCP 地址池中未找到可供分配的 IP 地址，则依次查询超过租期、发生冲突的 IP 地址，如果找到可用的 IP 地址，则进行分配，否则报告错误。

### 5.4.5.5 租期与定时器

租期指 DHCP 客户端从服务器端获得 IP 地址后可以使用该 IP 地址的最长时期，是 DHCP 一个重要概念。租期的更新、到期是由 DHCP 客户端上的租期更新、重绑定、到达租期三个定时器进行控制的。

DHCP 服务器可以为不同的地址池指定不同的地址租期，但同一 DHCP 地址池中的地址都具有相同的租期。

定时器的取值如表 5-14 所示。

表 5-14 定时器的取值

| 定时器  | 取值                        |
|------|---------------------------|
| 租期更新 | 总租期的 50%                  |
| 重绑定  | 总租期的 87.5%                |
| 到达租期 | 总租期。在 DHCP 服务器上配置，缺省为 1 天 |

当 DHCP 客户端获得 IP 地址时，会进入到绑定状态，此后三个定时器的作用如下：

- 当“租期更新定时器”到期时，DHCP 客户端必须进行 IP 地址的更新。DHCP 客户端会自动地向曾经为自己分配过的 IP 地址的 DHCP 服务器发送 DHCPREQUEST 报文，此时客户端进入到更新状态。如果之前分配的 IP 地址有效，则 DHCP 服务器回应 DHCPACK 报文，通知 DHCP 客户端已经获得新 IP 租约，客户端会重新进入到绑定状态。如果客户端收到 DHCP 服务器返回的 DHCPNAK 报文，则进入到初始化状态。
- 当客户端发送延长租期的 DHCPREQUEST 报文后，保持在更新状态等待响应。若直到“重绑定定时器”到期，客户端还没有收到服务器的响应，客户端会假定原来的 DHCP 服务器不可用，并开始发送广播的 DHCPREQUEST 报文。

网络上的任何 DHCP 服务器均可以响应此客户端的请求，并向此客户端发送 DHCPACK 报文或者 DHCPNAK 报文。

如果客户端收到一个 DHCPACK 报文，那么就返回到绑定状态，且重新设置“租期更新”和“重绑定”定时器。

如果客户端收到的都是 DHCPNAK 报文，那么就返回到初始化状态。此时客户端必须立即停止使用此 IP 地址，并且返回到初始化状态，重新申请新的 IP 地址。

- 如果客户端在“到达租期定时器”到期前都没有收到响应，客户端必须立即停止使用此 IP 地址，并且返回到初始化状态，重新申请新的 IP 地址。

#### 5.4.5.6 地址冲突探测

为防止 IP 地址重复分配导致地址冲突，DHCP 服务器为客户端分配地址前，需要先对该地址进行探测。

地址冲突探测是通过 ping 命令实现的，检测是否能在指定时间内得到 Ping 应答来判断是否存在地址冲突。DHCP 服务器发送目的地址为待分配 IP 地址的 Ping 报文，如果在等待 Ping 响应的最长时间没有得到应答，则继续发送 Ping 报文，直到发送 Ping 包数量达到最大值，如果仍然没有收到应答，则将该 IP 地址分配给客户端，从而确保分配给客户端的 IP 地址是唯一的。

## 5.5 DHCP Snooping

介绍了 DHCP Snooping 安全的基本原理、基本概念和应用。

### 5.5.1 介绍

DHCP Snooping 是一种 DHCP 安全特性，可以过滤不信任的 DHCP 消息并建立和维护一个 DHCP Snooping 绑定表。该绑定表包括 MAC 地址、IP 地址、租约时间、绑定类型、VLAN ID、接口信息。DHCP Snooping 的作用就如同在 DHCP Client 和 DHCP Server 之间建立一道防火墙。

DHCP Snooping 主要是解决网络中应用 DHCP 时设备遇到 DHCP DoS 攻击、DHCP Server 仿冒者攻击、ARP 中间人攻击及 IP/MAC Spoofing 攻击的问题。

为此，设备已经实现了 MAC 地址限制、DHCP Snooping 安全绑定、IP + MAC 绑定、OPTION82 特性等安全性方面的功能，为设备在网络中应用 DHCP 功能提供更高的安全性。

DHCP Snooping 既可以应用在设备的二层接口上，又可应用在设备三层接口上，分别如图 5-25 和图 5-26 所示。

图 5-25 在设备的二层接口下应用 DHCP Snooping 的典型组网

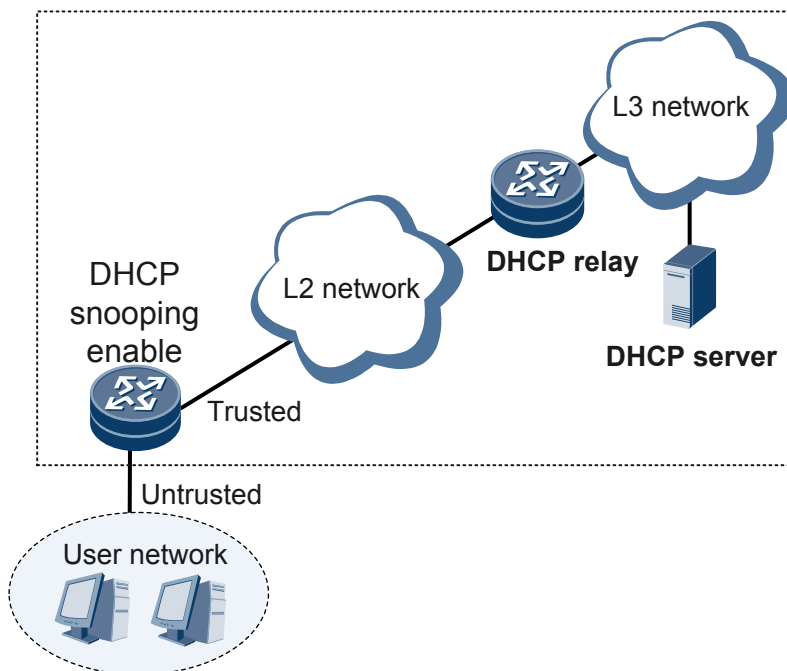
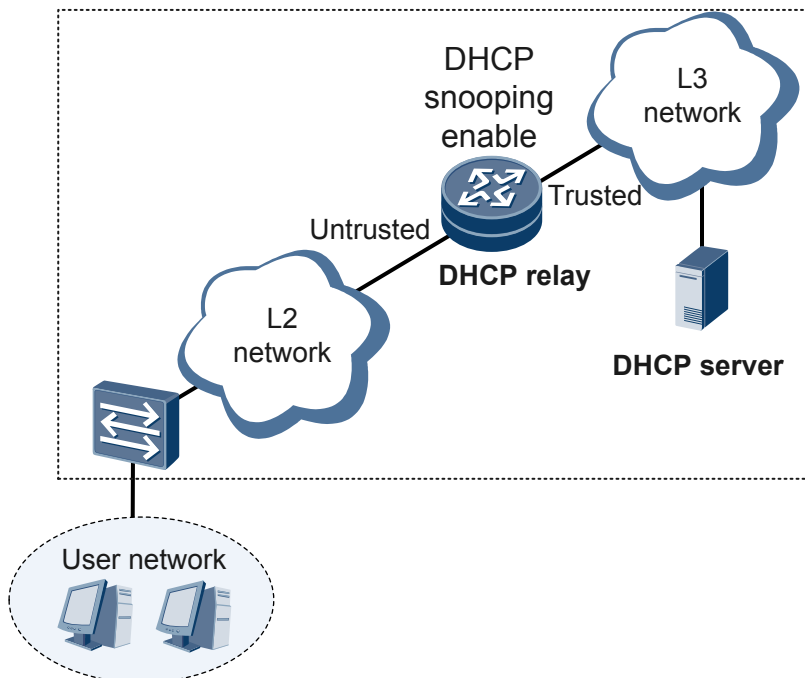


图 5-26 在设备的三层接口下应用 DHCP Snooping 的典型组网



DHCP Snooping 功能用于防止:

- DHCP Server 仿冒者攻击



- 中间人攻击与 IP/MAC Spoofing 攻击
- 改变 CHADDR 值的 DoS 攻击

根据不同的攻击类型，DHCP Snooping 提供不同的工作模式，见表 5-15。

表 5-15 攻击类型与 DHCP Snooping 工作模式对应表

| 攻击类型                     | DHCP Snooping 工作模式         |
|--------------------------|----------------------------|
| DHCP Server 仿冒者攻击        | 信任（Trusted）/不信任（Untrusted） |
| 中间人攻击/IP/MAC Spoofing 攻击 | DHCP Snooping 绑定表          |
| 改变 CHADDR 值的 DoS 攻击      | 检查 DHCP 报文的 CHADDR 字段      |

### 5.5.2 参考标准和协议

如果您想了解更多关于 DHCP Snooping 的信息，请参考以下文档。

- RFC 3046: DHCP Relay Agent Information Option
- RFC 2132: DHCP Options and BOOTP Vendor Extensions

### 5.5.3 可获得性

#### License 支持

不需要 License 支持。

#### 版本支持

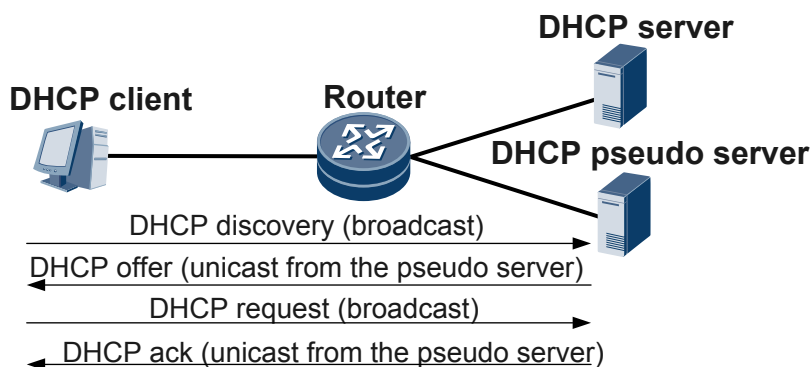
| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 5.5.4 原理描述

#### 5.5.4.1 DHCP Server 仿冒者攻击

由于 DHCP 请求报文以广播形式发送，所以 DHCP Server 仿冒者可以侦听到此报文。DHCP Server 仿冒者回应给 DHCP Client 仿冒信息，如错误的网关地址、错误的 DNS 服务器、错误的 IP 等，达到 DoS（Deny of Service）的目的。如图 5-27 所示。

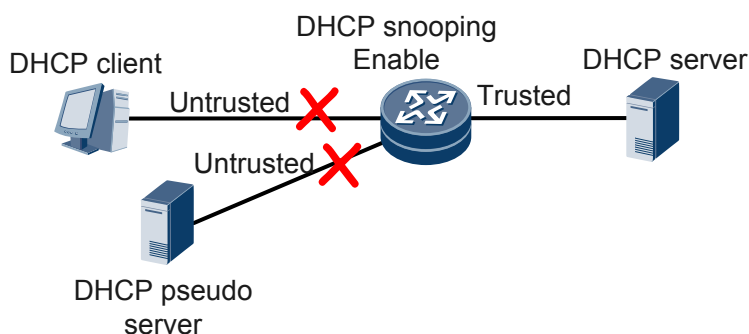
图 5-27 DHCP Server 仿冒者攻击示意图



为防止 DHCP Server 仿冒者攻击，可使用 DHCP Snooping 的“信任（Trusted）/不信任（Untrusted）”工作模式。

把某个物理接口或者 VLAN 的接口设置为“信任（Trusted）”或者“不信任（Untrusted）”。凡是从“不信任（Untrusted）”接口上收到的 DHCP Reply（Offer、ACK、NAK）报文直接丢弃，这样可以隔离 DHCP Server 仿冒者攻击。如图 5-28 所示。

图 5-28 Trusted/Untrusted 工作模式示意图



## 5.5.4.2 中间人攻击与 IP/MAC Spoofing 攻击

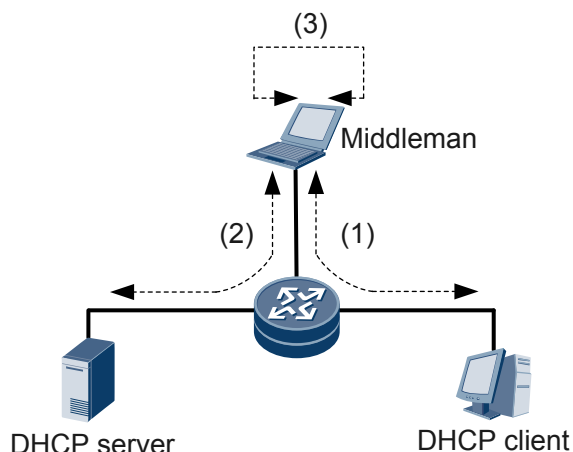
### 中间人攻击

首先，中间人向客户端发送带有自己的 MAC 地址和服务器 IP 地址的报文，让客户端学到中间人的 IP 和 MAC，达到仿冒 DHCP Server 的目的。达到目的后，客户端发往 DHCP 服务器的报文都会经过中间人；然后，中间人向服务器发送带有自己 MAC 和客户端 IP 的报文，让服务器学到中间人的 IP 和 MAC，达到仿冒客户端的目的，如图 5-29 所示。

中间人完成服务器和客户端的数据交换。在服务器看来，所有的报文都是来自或者发往客户端；在客户端看来，所有的报文也都是来自或者发往服务器端。但实际上这些报文都是经过了中间人的“二手”信息。



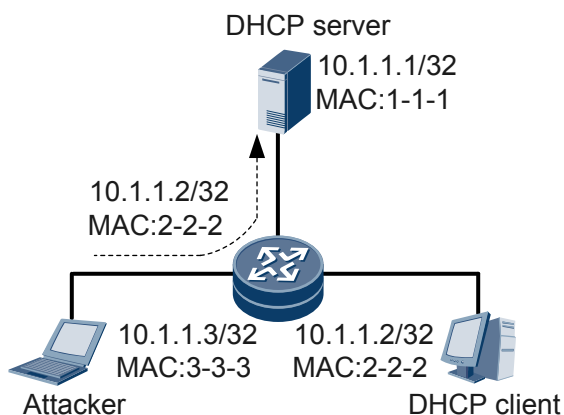
图 5-29 中间人攻击示意图



## IP/MAC Spoofing 攻击

攻击者向服务器发送带有合法用户 IP 和 MAC 的报文，令服务器误以为已经学到这个合法用户的 IP 和 MAC，但真正的合法用户不能从服务器获得服务。如图 5-30 所示。

图 5-30 IP/MAC Spoofing 攻击示意图



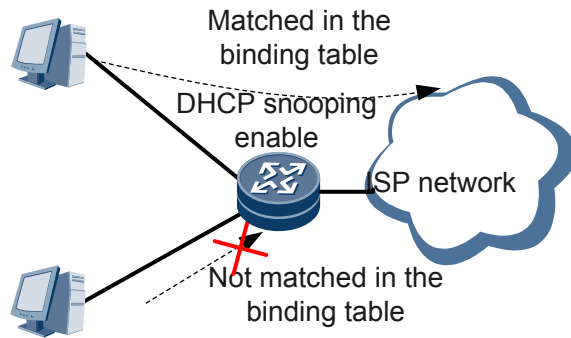
为隔离中间人攻击与 IP/MAC Spoofing 攻击，可使用 DHCP Snooping 绑定表工作模式。

当接口接收到 ARP 或者 IP 报文，使用 ARP 或者 IP 报文中的“源 IP+源 MAC”匹配 DHCP Snooping 绑定表。如果匹配就进行转发，如果不匹配就丢弃。如图 5-31 所示。

对于配置静态 IP 的用户，由于没有通过 DHCP 请求而获得 IP，所以，没有对应的 DHCP Snooping 绑定表项，该用户发出的 ARP、IP 报文会被丢弃，从而防止该用户非法使用网络。只能通过配置静态 DHCP Snooping 绑定表来允许静态 IP 用户访问网络。

对于盗用其他合法用户 IP 地址的用户，由于同样不是自己通过 DHCP 请求而获得 IP 的，IP 对应的 DHCP Snooping 绑定表项中的 MAC 以及接口与盗用者的不一致，盗用者发出的 ARP、IP 报文会被丢弃，从而防止该盗用者非法使用网络。

图 5-31 应用 IP 与 MAC 绑定表示意图



DHCP Snooping 绑定表中的表项分为两种：

- 通过命令行配置的静态表项，只能通过命令行删除。
- 通过 DHCP Snooping 功能自动学习到的动态表项，并根据 DHCP 租期进行老化。

DHCP Snooping 绑定表中的动态表项是根据 DHCP server 回的 ACK 报文自动生成的。根据网络设备层次不同，动态表项的建立过程也不同：

- 在二层设备上
  - 如果使能了 Option82 选项，二层设备通过监听 DHCP 报文，在 DHCP 请求报文中插入 Option82，DHCP Server 会在 Reply 报文中携带 Option82 选项。二层设备分析 Option82 选项即可确定 DHCP Reply 报文回应给哪个接口，从而建立对应的 DHCP Snooping 绑定表项。
  - 如果未使能 Option82 选项，二层设备根据 MAC 地址表来识别不同的接口信息。
- 在三层设备上
  - 对于配置为“不信任”的接口，设备通过监听 DHCP Reply 消息，获取 DHCP Server 分配给用户的 IP 地址、用户 MAC 地址、报文通过的接口等信息，建立不信任侧的 IP 与 MAC 绑定表项。该表项与分配给用户的 IP 地址具有同样的租期，租约到期或者用户释放该 IP 地址则自动删除该表项。

### 5.5.4.3 改变 CHADDR 值的 DoS 攻击

如果攻击者改变的不是数据帧头部的源 MAC，而是改变 DHCP 报文中的 CHADDR（Client Hardware Address）值来不断申请 IP 地址，如图 5-32 所示，而设备仅根据数据帧头部的源 MAC 来判断该报文是否合法，那么“MAC 地址限制”方案不能起作用。

图 5-32 改变 CHADDR 的 DOS 攻击

|                                              |   |               |       |                 |      |
|----------------------------------------------|---|---------------|-------|-----------------|------|
| 0                                            | 7 | 15            | 23    | 31              |      |
| OP Code                                      |   | Hardware Type |       | Hardware Length | HOPS |
| Transaction ID (XID)                         |   |               |       |                 |      |
| Seconds                                      |   |               | Flags |                 |      |
| Client IP Address (CIADDR)                   |   |               |       |                 |      |
| Your IP Address (YIADDR)                     |   |               |       |                 |      |
| Server IP Address (SIADDR)                   |   |               |       |                 |      |
| Gateway IP Address (GIADDR)                  |   |               |       |                 |      |
| Client Hardware IP Address (CHADDR)-16 bytes |   |               |       |                 |      |
| Server Name (SNAME)-64 bytes                 |   |               |       |                 |      |
| Filename-128 bytes                           |   |               |       |                 |      |
| DHCP Options                                 |   |               |       |                 |      |

这时，可以使用 DHCP Snooping 检查 DHCP 请求报文中 CHADDR 字段的功能。如果该字段跟数据帧头部的源 MAC 相匹配，便转发报文；否则，丢弃报文。

5.5.4.4 Option82

Option82 报文格式

Option82 是 DHCP 报文的一个特殊字段，供 DHCP Relay 设置，用来标识 DHCP Request 报文的发送路径。

客户端发送的 DHCP 请求报文，经过 DHCP Relay 时，DHCP Relay 为 DHCP Request 报文填充 Option82 字段，DHCP Server 收到含有 Option82 字段的 DHCP 请求报文时，在 DHCP Relay 报文中原样添加该字段返回给 DHCP Relay，DHCP Relay 根据 relay 报文中携带的 Option82 字段确定发送给那个接口。

如图 5-33 所示，Option82 的 Code 字段为 82，Length 字段标识代理信息字段的字节数，iN 为代理信息字段子选项。子选项由“子选项编号/长度值/子选项值”的元组序列组成。如图 5-34 所示，SubOpt 为子选项的编号，Length 字段标识代理信息子选项内容的长度。在 Option82 字段中，必须要定义一个子选项，子选项可以为空，因此 Option82 字段最短长度为 2。

图 5-33 Option82 报文格式

|             |   |                         |    |    |    |    |     |    |
|-------------|---|-------------------------|----|----|----|----|-----|----|
| Code Length |   | Agent Information Field |    |    |    |    |     |    |
| 82          | N | i1                      | i2 | i3 | i4 | i5 | ... | iN |

图 5-34 Option82 报文格式

| SubOpt Length |   | Sub-Option Value |    |    |    |    |     |    |
|---------------|---|------------------|----|----|----|----|-----|----|
| 1             | N | a1               | a2 | a3 | a4 | a5 | ... | aN |
| 2             | N | b1               | b2 | b3 | b4 | b5 | ... | bN |
| 9             | N | c1               | c2 | c3 | c4 | c5 | ... | cN |

最初分配的 DHCP 中继代理子选项为代理路径 ID 子选项和代理远程 ID 子选项。DHCP 服务器可以根据路径 ID 作为 IP 地址分配和其他参数指定的策略。

设备除了支持子选项 1 之外，还支持子选项 9。用来标识本设备添加的路径 ID 信息。

子选项 9 的作用：

- 对于接口转发的 DHCP 应答报文，如果 Option82 中含有子选项 9，并且子选项 9 中有厂商信息，设备就可以解析该字段，解析成功后，取出接口信息，把子选项 9 中的厂商信息字段剥掉，进行报文转发。
- 对于建立 DHCP Snooping 绑定表，首先解析该 DHCP 报文是不是 DHCP 应答报文，如果是 DHCP 应答报文，根据子选项 9 中取到的接口信息建立 DHCP Snooping 绑定表；否则，不能建立 DHCP Snooping 绑定表。

Option82 既可应用在二层设备上，又可应用在三层设备上。三层设备可以通过 DHCP 服务器使用 Option82 选项执行地址分配策略或其他策略；二层设备可以通过分析 Option82 选项确定 DHCP Reply 报文回应给哪个接口，从而建立对应的 IP 与 MAC 绑定表项。下面分别描述这两种情况。

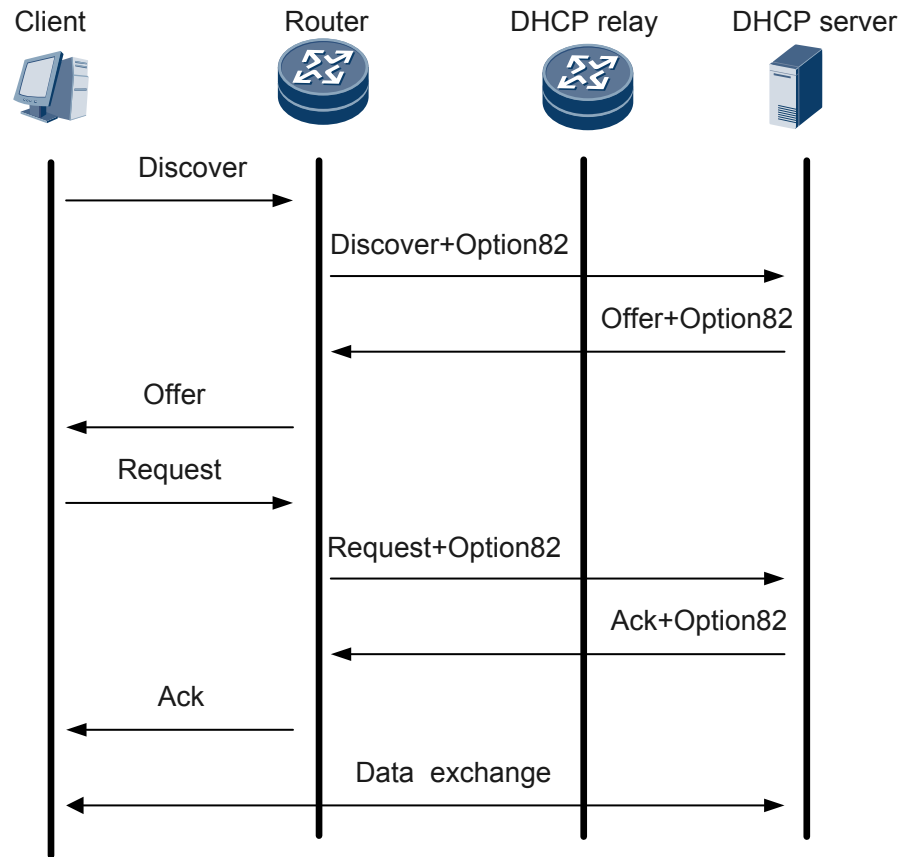
在二层设备上插入 Option82

如图 5-35 所示，用户设备首先接入二层设备，然后通过一个二层网络连接到 DHCP Relay 或者 DHCP Server。

如果在这个二层设备上启用了 DHCP Snooping，DHCP Reply 报文可能是广播报文。二层设备可以根据报文的 MAC 地址查找 VLAN，确定 DHCP 回应报文应该回给哪个接口，建立对应的 IP 与 MAC 地址绑定表项。

如果需要在二层设备上监听 DHCP 报文，在 DHCP Discover 报文中插入 Option82，DHCP Server 会在回应报文中携带 Option82 选项。二层设备可以通过分析 Option82 选项确定 DHCP Offer 报文回应给哪个接口，从而建立对应的 IP 与 MAC 绑定表项。并且将 DHCP Reply 报文中的 Option82 剥离后发给用户。

图 5-35 在二层设备上插入 Option82



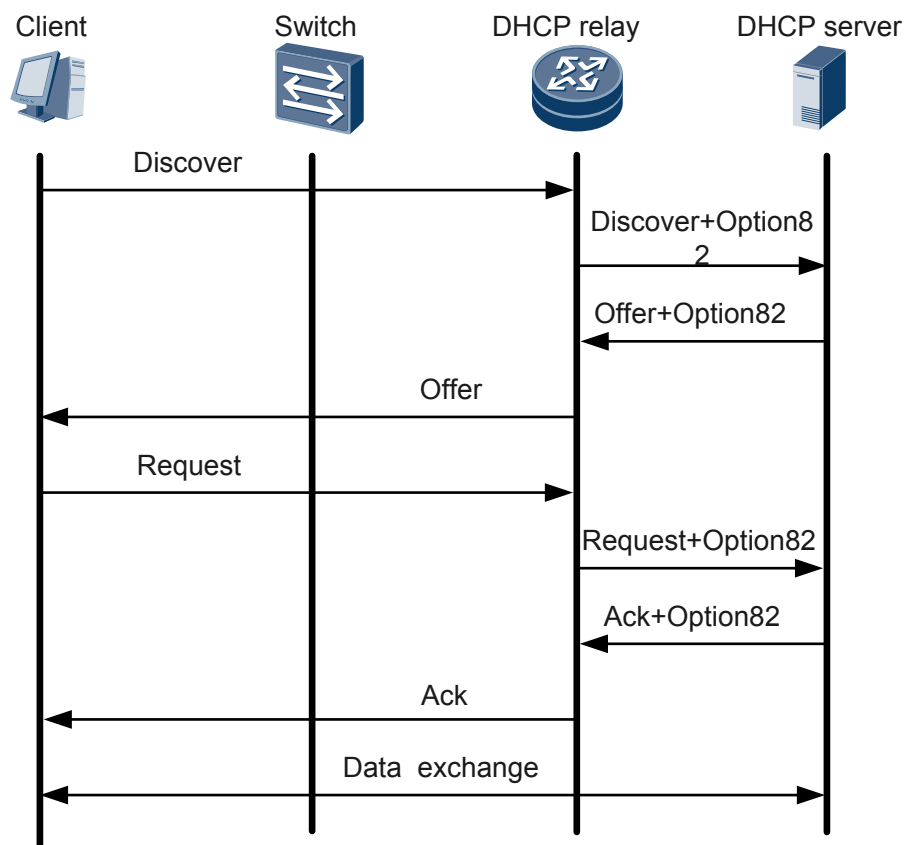
### 在三层设备上插入 Option82

在三层设备上插入 Option82 时，三层设备通常是指 DHCP Relay Agent。

如图 5-36 所示，DHCP Relay Agent 使能 Option82 功能后将 Option82 插入到用户发送的 Discover 和 Request 报文中，DHCP 服务器通过识别 Option82 来执行 IP 地址分配策略或其它策略。

DHCP 服务器的响应报文也带有 Option82，DHCP Relay Agent 收到带有 Option82 的响应报文后，将 Option82 剥离，然后发给用户。

图 5-36 在三层设备上插入 Option82



## Option82 的实现

客户端发送的 DHCP 请求报文，DHCP Realy 检查其是否携带 Option82 域。

- 如果有 Option82 域。  
检查其插入配置，插入 Option82 的配置分为 Insert 和 Rebuild 两种：
  - 如果当前接口配置是 Option82 Rebuild，说明该接口对已经携带的 Option82 不信任，修改 Option82，修改后的 Option82 子选项为 1。
  - 如果当前接口配置是 Option82 Insert，说明该接口对已经携带的 Option82 信任，不能对于该 Option82 的子选项 1 修改。需要检查 Option82 中是否有子选项 9，如果没有子选项 9，构造子选项 9。如果已有子选项 9，检查是否有厂商信息字段，如果没有厂商信息字段，构造厂商信息字段，添加到其它厂商信息字段的后面。
- 如果没有 Option82 域。  
不区分接口针对 Option82 的插入配置是 Insert 还是 Rebuild，都插入子选项 1 的 Option82。

DHCP 应答报文转发过程中，如果有子选项 1 或 9，并且子选项 1 或 9 中有厂商信息字段，转发设备成功解析该字段后，把子选项 1 或 9 中的厂商信息字段剥掉，然后转发。

## 5.6 IP 性能

在一些特定的网络环境里，需要调整 IP 参数，以便网络性能达到最佳。

### 5.6.1 介绍

#### 定义

在特定的网络环境中，可以配置接口最大传输单元（MTU）、配置 ICMP 属性、配置 TCP 属性等来调整设备的 IP 性能。

#### 目的

在不同的网络环境下，对特定的影响 IP 性能的指标进行调整，使网络业务可以更稳定的运行。

### 5.6.2 参考标准和协议

与 IP 性能相关的参考标准与协议如下：

- RFC 791: Internet Protocol
- RFC 792: Internet Control Message Protocol
- RFC 950: Internet Standard Subnetting Procedure
- RFC 1122: Requirements for Internet Hosts - Communication Layers
- RFC 1256: ICMP Router Discovery Messages

### 5.6.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 5.6.4 原理描述

### 5.6.4.1 IP 报文

#### IP 协议简介

IP 协议（Internet Protocol）是 TCP/IP 协议族中最核心的协议，提供不可靠的、无连接的、尽力而为的数据报文传输服务。

- 不可靠的意思是 IP 并不能保证数据报文能成功到达目的地，可靠性由上层来提供。
- 尽力而为表明互联网并不轻易的丢弃分组。
- 无连接表示每个分组都是被独立对待的，一系列的分组序列从一台主机发送到另外一台主机的过程中，可能经过的是不同的传输路径，甚至在这过程中会出现有的分组丢失的现象。

#### IP 报文格式

IP 报文是互联网上数据传输的基本单元，分为首部和数据区，IP 报文的格式如[图 5-37](#)所示。

图 5-37 IP 报文格式

|                     |  |          |  |                 |  |                 |         |    |  |
|---------------------|--|----------|--|-----------------|--|-----------------|---------|----|--|
| 0                   |  | 7        |  | 15              |  | 23              |         | 31 |  |
| Version             |  | HLEN     |  | Service Type    |  | Total Length    |         |    |  |
| Identification      |  |          |  | Flags           |  | Fragment Offset |         |    |  |
| TTL                 |  | Protocol |  | Header Checksum |  |                 |         |    |  |
| Source Address      |  |          |  |                 |  |                 |         |    |  |
| Destination Address |  |          |  |                 |  |                 |         |    |  |
| Options (Variable)  |  |          |  |                 |  |                 | Padding |    |  |
| Data<br>(Variable)  |  |          |  |                 |  |                 |         |    |  |

报文中各个字段的解释如[表 5-16](#)所示：

表 5-16 IP 报文各个字段的含义

| 字段      | 长度   | 含义                                                                                  |
|---------|------|-------------------------------------------------------------------------------------|
| Version | 4 比特 | 版本，该字段表示该数据报文的 IP 协议的版本信息： <ul style="list-style-type: none"><li>● 4：IPv4</li></ul> |



| 字段                  | 长度       | 含义                                                                                                                                                                                                                                                         |
|---------------------|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| HLEN                | 4 比特     | 首部长度字段，表示以 32 比特为单位的首部长度。如图 5-37 所示，除去 IP 选项（IP Options）和填充（Padding）字段外，IP 报文首部的长度是固定的。一般地 IP 报文首部不含 IP 选项和填充字段，因此首部长度为 20 字节，那么首部长度字段的值为 0101（即二进制表示的 5）。IP 报文首部的最大长度是 60 字节，此时首部长度字段的值为 1111（即二进制表示的 15）。                                                |
| Service Type        | 8 比特     | 服务类型字段，表示数据报文的处理方式。                                                                                                                                                                                                                                        |
| Total Length        | 16 比特    | 总长度字段，表示以字节为单位的 IP 报文总长度。由于总长度字段的长度为 16 比特，因而 IP 报文的最大长度为 65535 字节。                                                                                                                                                                                        |
| Identification      | 16 比特    | 标识字段，唯一的标识主机发送的每一份数据报文。主机每发送一份数据报文该字段的值就加 1。                                                                                                                                                                                                               |
| Flags               | 3 比特     | 标志字段，该字段第 2 位和第 3 位控制分片。该字段的详细格式如“图 5-38”所示。其中最高位保留未用，置 0。 <ul style="list-style-type: none"> <li>● DF（Don't Fragment）：该字段表示是否允许数据报文被分片。该位置 0 时表示该数据报文可以被分片，否则不允许分片。</li> <li>● MF（More Fragments）：该字段表示本片的数据是否为原始数据报文中的最后一块。该位置 0 时表示这是最后一个分片。</li> </ul> |
| Fragment Offset     | 13 比特    | 片偏移量字段，该字段表示此分片中的数据在原始数据报文中的偏移量。为偏移量可达到数据包最大值 65535 字节，使用 8 字节作为本字段单位。                                                                                                                                                                                     |
| TTL                 | 8 比特     | 生存时间字段，单位为秒。该字段表示该数据报文在互联网中允许存在的时间。在当前的实现中，TTL 表示跳数限制，每经过一个 USG，其值会减 1。当 USG 接收到一个 TTL 字段为 1 且目的地址不是自己的数据报文时，会丢弃该数据报文，并向源端发出一个差错报告。正常情况下，设备不会接收到 TTL 字段为 0 的数据报文。                                                                                          |
| Protocol            | 8 比特     | 协议字段，表示用来创建该数据报文的数据区中报文的高层协议。                                                                                                                                                                                                                              |
| Header Checksum     | 16 比特    | 首部校验和字段，用于保证 IP 报文头部信息的完整。                                                                                                                                                                                                                                 |
| Source Address      | 32 比特    | 源 IP 地址字段，表示数据报文的发送方的 IP 地址。数据报文在互联网传递过程中，此字段的值都始终不变。                                                                                                                                                                                                      |
| Destination Address | 32 比特    | 目的 IP 地址字段，表示数据报文的接收方的 IP 地址。数据报文在互联网传递过程中，此字段的值都始终不变。                                                                                                                                                                                                     |
| Options             | 长度取决于所选项 | 选项字段。                                                                                                                                                                                                                                                      |

| 字段      | 长度       | 含义                                       |
|---------|----------|------------------------------------------|
| Padding | 长度取决于所选项 | 填充字段，若数据报文首部不是 32 比特的整数倍，则会通过该字段填充“0”补齐。 |

图 5-38 Flags 字段的格式

|        |    |    |   |
|--------|----|----|---|
| 0      | 1  | 2  | 3 |
| Unused | DF | MF |   |

IP 报文的 DF 位

IP 报文头中包含一个用于指示是否允许对报文进行分片的标志位 DF（Don't Fragment）。通常情况下，如果 DF 被置为 1，则 USG 不能再对该报文分片。当 IP 报文长度大于接口的 MTU（Maximum Transfer Unit）时，该报文将被丢弃并发送 ICMP 差错报文，这样处理的结果是容易引起数据的丢失。

当 DF 置位的报文长度已经超过出接口的 MTU 时，可通过一条命令开关，控制在指定输出接口下对这种报文的处理。通过这种命令决定是丢弃报文，还是清除不分片标志 DF 并分片发送报文。

IP 源地址校验

IP 源地址校验功能用来对 IP 报文的源地址进行合法性校验，丢弃源地址不合法的报文。如下几种 IP 地址均为非法源地址。

- 全 0（且目的地址不为全 1）
- 全 1
- 组播地址（D 类地址）
- E 类地址
- 非本机产生的环回地址（形式为 127.x.x.x）
- A、B、C 类广播地址
- 与入接口地址在同一网段的子网广播地址

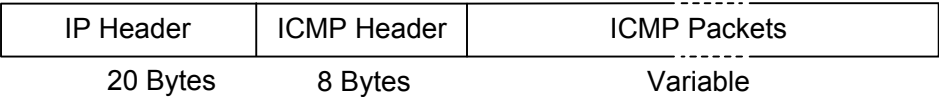
当该功能的命令开关在某接口打开时，会对该接口收到的所有 IP 报文按以上原则进行源地址合法性校验，如果发现是非法源地址，则丢弃该报文，并统计到计数器中。

5.6.4.2 ICMP 报文

ICMP 简介

ICMP（Internet Control Message Protocol）是一个差错报告机制，通常被 IP 层或更高层协议（TCP 或 UDP）使用。如图 5-39 所示，ICMP 报文被封装在 IP 数据报内部，作为 IP 数据报的数据部分通过互联网传递。

图 5-39 ICMP 报文



当数据报产生差错时，ICMP 只向数据报的源端报告这个差错，既不会去纠正这个差错也不会通知中间的网络设备。网络中多数差错产生于最初的源端，而当差错产生于中间的设备时，即使源端收到了差错报告，它也无法判断出这个差错是由中间的哪个网络设备引起的。

IP 数据报中的字段仅包含源端和最终的目的站，并没记录报文在网络传递中的全部路径（除非 IP 数据报中设置了路由记录选项）。因此当设备检测到差错时，它无法通知中间的网络设备，只能向源端发送差错报告。

ICMP 报文格式

每个 ICMP 报文都有自己的格式，但是所有的 ICMP 报文的前 3 个字段的格式都是一样的，如图 5-40 所示。

图 5-40 ICMP 报文格式

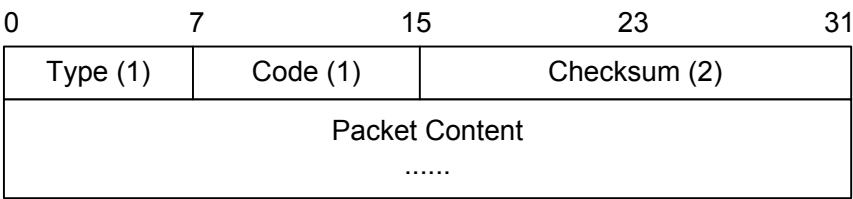


图 5-40 中括号里的数字表示字段的长度，单位是字节。

报文中各个字段的解释如表 5-17 所示。

表 5-17 ICMP 报文中各个字段的含义

| 字段             | 长度   | 含义                                               |
|----------------|------|--------------------------------------------------|
| Type           | 1 字节 | 报文类型，用来标识报文，Type 字段的格式和含义如表 5-18 所示。             |
| Code           | 1 字节 | 代码，提供报文类型的进一步信息，Code 字段的格式和含义如表 5-18 所示。         |
| Checksum       | 2 字节 | 校验和，使用 and IP 相同的加法校验和算法，但是 ICMP 校验和仅覆盖 ICMP 报文。 |
| Packet Content | 可变   | ICMP 的报文内容，这部分内容因不同的 ICMP 类型和代码而不同。              |

## ICMP 报文类型

各种类型的 ICMP 报文如表 5-18 所示，报文中的类型字段和代码字段共同决定 ICMP 报文的类型。表格中的最后两列表示此 ICMP 报文为查询报文或者差错报文。规定在对 ICMP 差错报文进行响应时，不会生成另一份 ICMP 差错报文。否则就会造成一个差错报文产生另一个差错的情况，要是这个差错报文再产生差错的话，就会出现无休止循环下去的状况。

在下列情况下，不会再产生 ICMP 差错报文：

- ICMP 差错报文
- 目的地址为广播地址或者多播地址的 IP 数据报
- 链路层广播的数据报
- 不是 IP 分片的第一片
- 源地址为零或者源地址为环回地址、广播地址、多播地址的 IP 数据报

表 5-18 ICMP 报文的类型

| Type | Code | 描述               | 查询报文 | 差错报文 |
|------|------|------------------|------|------|
| 0    | 0    | 回送应答             | 是    | -    |
| 3    |      | 目的站不可达           | -    | -    |
|      | 0    | 网络不可达            | -    | 是    |
|      | 1    | 主机不可达            | -    | 是    |
|      | 2    | 协议不可达            | -    | 是    |
|      | 3    | 端口不可达            | -    | 是    |
|      | 4    | 需要进行分片但设置了不分片比特  | -    | 是    |
|      | 5    | 源站选路失败           | -    | 是    |
|      | 6    | 目的网络不认识          | -    | 是    |
|      | 7    | 目的主机不认识          | -    | 是    |
|      | 8    | 源主机被隔离           | -    | 是    |
|      | 9    | 目的网络被强制禁止        | -    | 是    |
|      | 10   | 目的主机被强制禁止        | -    | 是    |
|      | 11   | 由于服务类型 TOS，网络不可达 | -    | 是    |
|      | 12   | 由于服务类型 TOS，主机不可达 | -    | 是    |
|      | 13   | 由于过滤，通信被强制禁止     | -    | 是    |
|      | 14   | 主机越权             | -    | 是    |
|      | 15   | 优先权中止生效          | -    | 是    |
| 4    | 0    | 源站抑制             | -    | 是    |

| Type | Code | 描述              | 查询报文 | 差错报文 |
|------|------|-----------------|------|------|
| 5    |      | 重定向             | -    | 是    |
|      | 0    | 对网络重定向          | -    | 是    |
|      | 1    | 对主机重定向          | -    | 是    |
|      | 2    | 对服务类型和网络重定向     | -    | 是    |
|      | 3    | 对服务类型和主机重定向     | -    | 是    |
| 8    | 0    | 回送请求            | 是    | -    |
| 9    | 0    | 路由器通告           | 是    | -    |
| 10   | 0    | 路由器请求           | 是    | -    |
| 11   |      | 数据报超时           | -    | -    |
|      | 0    | 传输期间生存时间为 0     | -    | 是    |
|      | 1    | 在数据报组装期间生存时间为 0 | -    | 是    |
| 12   |      | 数据报参数错          | -    | -    |
|      | 0    | 坏的 IP 首部        | -    | 是    |
|      | 1    | 缺少必需的选项         | -    | 是    |
| 13   | 0    | 时间戳请求           | 是    | -    |
| 14   | 0    | 时间戳应答           | 是    | -    |
| 15   | 0    | 信息请求（已作废不用）     | 是    | -    |
| 16   | 0    | 信息应答（已作废不用）     | 是    | -    |
| 17   | 0    | 地址掩码请求          | 是    | -    |
| 18   | 0    | 地址掩码应答          | 是    | -    |

回送请求和回送应答报文

图 5-41 回送请求和回送应答报文格式

|                        |          |                 |    |    |
|------------------------|----------|-----------------|----|----|
| 0                      | 7        | 15              | 23 | 31 |
| Type (0 or 8)          | Code (0) | Checksum        |    |    |
| Identifier             |          | Sequence Number |    |    |
| Optional Data<br>..... |          |                 |    |    |

图 5-41 中括号里的数字表示字段的取值。

报文中各个字段的解释如表 5-19 所示。

表 5-19 回送请求和回送应答报文中字段的解释

| 字段              | 长度       | 含义                                                                                  |
|-----------------|----------|-------------------------------------------------------------------------------------|
| Type            | 1 字节     | <ul style="list-style-type: none"><li>● 0：表示回送应答报文。</li><li>● 8：表示回送请求报文。</li></ul> |
| Code            | 1 字节     | 取值为 0。                                                                              |
| Checksum        | 2 字节     | 校验和。                                                                                |
| Identifier      | 2 字节     | 标识符。                                                                                |
| Sequence Number | 2 字节     | 序列号。                                                                                |
| Optional Data   | 长度取决于所选项 | 可选数据，包含返回给发送者的数据。回送应答返回的数据通常与其收到的请求完全相同。                                            |

5.6.4.3 差错和重定向 ICMP 报文处理

目的站不可达报文

IP 是一个尽力而为的交付机制，不会轻易丢弃报文。当 USG 无法转发或者交付 IP 报文时，会向源站发送一个目的站不可达的报文，然后丢弃该报文。其报文格式如图 5-42 所示。

图 5-42 目的站不可达报文的格式

|                                                              |             |          |    |    |
|--------------------------------------------------------------|-------------|----------|----|----|
| 0                                                            | 7           | 15       | 23 | 31 |
| Type (3)                                                     | Code (0-15) | Checksum |    |    |
| Unused                                                       |             |          |    |    |
| Internet Header + 64 bits of Original Data Datagram<br>..... |             |          |    |    |

图 5-42 中括号里的数字表示字段的取值。

关于 Type 和 Code 字段的取值请参见 ICMP 报文类型。报文中 4 字节的字段保留未用，全部置 0。

网络不可达意味着选路失败，主机不可达意味着交付失败。通过报文中的 IP 首部和原始报文的前 64 比特（Internet Header + 64 bits of Original Data Datagram）字段，源站可以得知是哪个地址不可达。

如果 USG 在转发报文时发现报文同时满足以下条件：

- 报文的目地址没有对应的路由
- 该报文也不是到本机的报文

那么 USG 将丢弃该报文，同时向源地址返回一个 ICMP 主机不可达报文，以通知源主机停止发送到该目的地的报文。

重定向报文

重定向报文的格式如图 5-43 所示。

图 5-43 重定向报文的格式

|                                                     |            |          |    |    |
|-----------------------------------------------------|------------|----------|----|----|
| 0                                                   | 7          | 15       | 23 | 31 |
| Type (5)                                            | Code (0-3) | Checksum |    |    |
| Gateway Internet Address                            |            |          |    |    |
| Internet Header + 64 bits of Original Data Datagram |            |          |    |    |
| .....                                               |            |          |    |    |

图 5-43 中括号里的数字表示字段的取值。

若 USG 在转发报文时同时满足以下条件，这说明主机错误地选择了报文发送的第一跳 USG，则需要向源主机发送 ICMP 重定向报文，通知主机重新选择正确的 USG 进行报文下一跳的发送。

- 接收和转发报文的接口是同一接口
- 被选择的路由本身没有被 ICMP 重定向报文创建或修改过
- 被选择的路由也不是到默认目的地（0.0.0.0）的路由
- 报文源地址的子网掩码位和输出接口的子网掩码位相同

ICMP 报文发送开关



如果关闭系统 ICMP 重定向报文发送功能，则 USG 在任何情况下都不会再发出 ICMP 重定向报文。

如果关闭 ICMP 主机不可达报文发送功能，则 USG 在任何情况下都不会再发出 ICMP 主机不可达报文。

在正常情况下，可以正确发送 ICMP 主机不可达报文和 ICMP 重定向报文。但是，当网络流量较大时，如果以上的现象出现较多，则 USG 会因此发送大量的 ICMP 报文，增大网络的流量负担。尤其是当遇到恶意攻击时，容易产生恶性循环，从而加剧网络的拥塞。

USG 提供在 ICMP 报文的出接口增加两个控制开关，分别用来打开或关闭 ICMP 主机不可达报文和 ICMP 重定向报文的发送开关。如果关闭这两个开关，则 USG 不会发送这两种报文，从而起到减小网络流量负担、防止遭到恶意攻击的作用。

#### 5.6.4.4 Ping 和 Tracert

##### Ping

**Ping** 命令通过发送 ICMP 请求报文和 ICMP 应答报文来验证与另一台 TCP/IP 设备的 IP 级连接。因为请求和应答都是用 IP 封装和发送的，所以应答的成功接收就证实互联网传输系统的主要部分是正常的。**Ping** 是用于检测网络连接性、可到达性和名称解析等疑难问题的主要 TCP/IP 命令，可以帮助网络工程师或用户定位网络问题。

使用 **Ping** 命令的主机或 USG 向目的站发送 ICMP 请求报文后，等待接收目的主机发回的 ICMP 响应报文。发送方可以设置选项来指定发送到目的地址的回应请求报文数以及报文大小。超时选项可用于指定程序等待响应的时间。如果指定了目的主机名称，先进行 DNS 查找获取主机的 IP 地址，再向目的主机发送回应请求。

##### Tracert

**Traceroute** 命令用来跟踪从本地 USG 到目的地址的路由。跟踪路由使用扩展的跳数限制搜索来确定路由。从本地到达目的地址的每一跳都会显示到控制台上。

**Traceroute** 命令使用 ICMP 报文和 IP 首部中的 TTL 字段，TTL 字段的值由发送方指定。在通信过程中，每个处理数据报的 USG 都需要把 TTL 的值减去 1。当 USG 接收到的 TTL 字段为 0 或 1 时，丢弃该数据报，并向发送方发送一个 ICMP “超时” 信息。通过这个 ICMP 信息，发送方可以知道 USG 的 IP 地址。

**Traceroute** 命令使用的是 UDP 报文，并且选择一个不可达的端口号（大于 30000）。当数据报到达主机时，目的主机会发送一个“端口不可达”的 ICMP 报文给发送方。所以，发送方根据接收到的 ICMP 报文是超时还是端口不可达来结束 **Traceroute** 命令程序。

#### 5.6.4.5 报文保序

##### 乱序报文

如图 5-44 所示，接收端接收到的报文顺序和发送端发出的报文顺序不一致。此时我们称接收端收到的报文为乱序报文。

乱序报文可能会因网络状况不稳定、Trunk 技术、启用逐包负载分担等原因导致。

图 5-44 乱序报文示意图



##### 报文保序

对报文顺序要求严格的业务系统在处理乱序报文时，可能会引起解析失败，从而影响业务正常运行。此时，需开启设备的报文保序功能。

设备通过建立保序队列，记录接口接收报文的先后顺序，从而保证报文按照接收的顺序进行转发。



启用报文保序功能后，三层接口上采用队列调度技术的相关功能（例如 LR、PQ、CQ、WFQ、RTPQ、CBWFQ、GTS）将失效。

## 5.7 IPv6 基础

IPv6 是新版本的网络层协议。

### 5.7.1 介绍

#### 定义

IPv6（Internet Protocol Version 6）是新版本的网络层协议，它是 Internet 工程任务组（IETF）设计的一套规范。IPv6 和 IPv4 之间最显著的区别就是 IP 地址的长度从 32 位升为 128 位。

IPv6 可以较为彻底的解决 IP 地址缺乏问题。同时，使用 IPv6 后，网络中路由设备的路由表项将会变少，可以提高路由设备转发报文的速率。

#### 目的

以 IPv4 为核心技术的 Internet 获得巨大成功，促使 IP 技术得到广泛应用。然而，随着因特网的迅猛发展，IPv4 设计的不足也日益明显，主要有以下几点：

- IPv4 地址空间不足

IPv4 地址采用 32 比特标识，理论上能够提供的地址数量是 43 亿。但由于地址分配的原因，实际可使用的数量不到 43 亿。另外，IPv4 地址的分配也很不均衡：美国占全球地址空间的一半左右，而欧洲则相对匮乏；亚太地区则更加匮乏。与此同时，移动 IP 和宽带技术的发展需要更多的 IP 地址。IPv4 地址资源紧张直接限制了 IP 技术应用的进一步发展。

针对 IPv4 的地址短缺问题，也曾先后出现过几种解决方案。比较有代表性的是 CIDR(Classless Inter-Domain Routing)和 NAT(IP Network Address Translator)。但是 CIDR 和 NAT 都有各自的弊端和不能解决的问题，由此推动了 IPv6 的发展。

- 骨干路由器维护的路由表表项数量过大

由于 IPv4 发展初期的分配规划问题，造成许多 IPv4 地址分配不连续，不能有效聚合路由。日益庞大的路由表耗用较多内存，对设备成本和转发效率产生影响，这一问题促使设备制造商不断升级其路由器产品，以提高路由寻址和转发性能。

- 不易进行自动配置和重新编址

由于 IPv4 地址只有 32 比特，并且地址分配不均衡，导致在网络扩容或重新部署时，经常需要重新分配 IP 地址。因此需要能够进行自动配置和重新编址以减少维护工作量。

- 不能解决日益突出的安全问题

随着因特网的发展，安全问题越来越突出。IPv4 协议制定时并没有仔细针对安全性进行设计，因此固有的框架结构并不能支持端到端的安全。IPv6 将 IPSec 作为它的标准扩展头实现，可以提供端到端的安全特性。

IPv6 技术从根本上解决了 IP 地址短缺的问题；且易于部署，能够兼容当前的各种应用，方便用户的平滑过渡；同时可实现与 IPv4 网络的共存和互通。由于 IPv4 存在以上种种弊端和不足，IPv6 技术的优越性显而易见，因此 IPv6 技术得以迅猛发展。

## 5.7.2 参考标准和协议

与 IPv6 特性相关的参考标准与协议如下：

- RFC 1981: Path MTU Discovery for IP version 6
- RFC 2292: Advanced Sockets API for IPv6
- RFC 2460: Internet Protocol, Version 6 (IPv6) Specification
- RFC 2462: Transmission of IPv6 Packets over Ethernet Networks
- RFC 2463: Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification
- RFC 2464: Transmission of IPv6 Packets over Ethernet Networks
- RFC 2465: Management Information Base for IP Version 6: Textual Conventions and General Group
- RFC 2472: IP Version 6 over PPP
- RFC 2641: Cabletron's VlanHello Protocol Specification Version 4
- RFC 2784: Generic Routing Encapsulation (GRE)
- RFC 2893: Transition Mechanisms for IPv6 Hosts and Routers
- RFC 3056: Connection of IPv6 Domains via IPv4 Clouds
- RFC 3513: Internet Protocol Version 6 (IPv6) Addressing Architecture
- RFC 3542: Advanced Sockets Application Program Interface (API) for IPv6
- RFC 3493: Basic Socket Interface Extensions for IPv6

## 5.7.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 5.7.4 原理描述

### 5.7.4.1 IPv6 的特点

IPv6 作为新一代的网络协议，具有以下特点：

- 充足的地址空间

IPv6 的源地址与目的地址长度都是 128 位（16 字节）。它可以提供超过  $3.4 \times 10^{38}$  种可能的地址空间，可以满足多层次的地址划分需要，以及公有地址和机构内部私有网络地址的分配。在 IPv6 网络中可以不再使用一些地址节约技术，如 NAT 等。

- 层次化的地址结构

IPv6 的地址空间采用了层次化的地址结构，利于路由快速查找，同时借助路由聚合，可减少 IPv6 路由表的大小，提高路由设备的转发效率。

- 地址自动配置

为了简化主机配置，IPv6 支持有状态地址配置（Stateful Address Autoconfiguration）和无状态地址配置（Stateless Address Autoconfiguration）。

- 对于有状态地址配置，主机通过服务器获取地址信息和配置信息。
- 对于无状态地址配置，主机自动配置地址信息，地址中带有本地路由设备通告的前缀和主机的接口标识。如果链路上没有路由设备，主机只能自动配置链路本地地址，实现与本地节点的互通。

- 支持 QoS

IPv6 报头的新字段定义了流量应该被如何标识和处理。通过报文头里的流标签（Flow Label）字段完成流量标识，允许路由设备对某一流中的报文进行识别并提供特殊处理。

由于 IPv6 报头可对流量进行识别，即使是带有 IPSec 加密的报文载荷也可对其 QoS 进行保证。



流是从源地址到目的地址之间符合某种分类的一组报文。

- 内置安全性

IPv6 将 IPSec 作为它的扩展报头实现，提供端到端的安全特性。这一特性为解决网络安全问题提供了标准，并提高了不同 IPv6 实现的互操作性。

- 增强的邻居发现机制

IPv6 的邻居发现（Neighbor Discovery）协议通过一组 ICMPv6（Internet Control Message Protocol for IPv6）消息，管理邻居节点（即同一链路上的节点）的交互。

IPv6 邻居发现协议用高效的组播和单播消息代替了 ARP（Address Resolution Protocol）、ICMPv4 路由设备发现（Router Discovery）和 ICMPv4 重定向（Redirect）消息，并提供了一系列其他功能。

## 5.7.4.2 IPv6 地址

### IPv6 地址的书写格式

IPv6 的 128 位 IP 地址有以下两种表示形式：

- X:X:X:X:X:X:X:X

在这种形式中，128 位的 IP 地址被分为 8 组，每组的 16 位用 4 个十六进制字符（0～9，A～F）来表示，组和组之间用冒号（:）隔开。其中每个“X”代表一组十六进制数值。比如下面这个 IPv6 地址：

2031:0000:130F:0000:0000:09C0:876A:130B

为了书写方便，每组中的前导“0”都可以省略，所以上述地址可写为：

2031:0:130F:0:0:9C0:876A:130B

另外，地址中包含的连续两个或多个均为 0 的组，可以用双冒号 “::” 来代替，这样可以压缩 IPv6 地址书写时的长度，所以上述地址又可以进一步简写为：

2031:0:130F::9C0:876A:130B

#### 说明

在一个 IPv6 地址中只能使用一次双冒号 “::”，否则当计算机将压缩后的地址恢复成 128 位时，无法确定每段中 0 的个数。

#### ● X:X:X:X:X:X:d.d.d.d

分为如下两种类型：

- IPv4 兼容 IPv6 地址。地址格式为：0:0:0:0:0:IPv4-address，其高阶 96bits 均为 0，其低阶 32bits 是一个 IPv4 地址。该 IPv4 地址必须是 IPv4 网络中可达的 IPv4 地址，且不能是组播地址、广播地址、环回地址或未指定的地址（0.0.0.0）。
- IPv4 映射 IPv6 地址。地址格式为：0:0:0:0:FFFF:IPv4-address。该地址用来将 IPv4 节点的地址表示为 IPv6 地址。

其中 IPv4 兼容 IPv6 地址用于配置 IPv6 over IPv4 隧道。

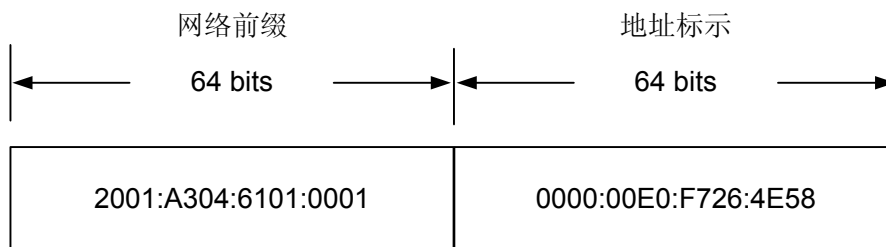
其中 “X” 代表高阶的六组数字，用十六进制数来表示每组的 16 比特。“d” 代表低阶的四组数字，用十进制数表示每组的 8 比特。后边的部分（d.d.d.d）其实就是一个标准的 IPv4 地址。

一个 IPv6 地址可以分为如下两部分：

- 网络前缀：n 比特，相当于 IPv4 地址中的网络 ID。
- 接口标识：128-n 比特，相当于 IPv4 地址中的主机 ID。

地址 2001:A304:6101:1::E0:F726:4E58 /64 的构成如图 5-45 所示。

图 5-45 地址 2001:A304:6101:1::E0:F726:4E58 /64 的构成示意图



## IPv6 的地址分类

IPv6 主要有三种地址：单播（Unicast）地址、任播（Anycast）地址和组播（Multicast）地址。

#### 说明

目前设备不支持任播地址。

- 单播地址：唯一标识一个接口，类似于 IPv4 的单播地址。发送到单播地址的数据包将被传输到此地址所标识的唯一接口。

单播地址还可以分为五种，如下：

- 链路本地单播地址

- 站点本地单播地址
- 环回地址
- 未指定地址
- 全球单播地址

上述五种地址如表 5-20 所示。

- 任播地址：用来标识一组接口（通常这组接口属于不同的节点）。发送到任播地址的数据包被传输给此地址所标识的一组接口中距离源节点最近的一个接口（最“近”的一个，是指根据路由协议的距离度量）。

应用场合：当移动主机需要与它的“home”子网上的移动代理之一通信时，它将用该子网路由设备的任播地址。

具体地址规定：任播地址没有独立的地址空间，它们可使用任何单播地址的格式。因此，需要一种语法来区别任播地址和单播地址。

- 组播地址：用来标识属于不同节点的一组接口，类似 IPv4 的组播地址。发送到组播地址的数据包被传输给此地址所标识的所有接口。

IPv6 不包括广播地址，广播地址的功能均由组播地址来提供。

## 单播地址的类型

单播只能进行一对一的传输，它只能识别一个接口（类似 IPv4 单播地址）。IPv6 的单播地址类型如表 5-20 所示。

表 5-20 IPv6 单播地址类型

| 地址类型     | 二进制前缀             | IPv6 前缀标识 |
|----------|-------------------|-----------|
| 链路本地单播地址 | 1111111010        | FE80::/10 |
| 站点本地单播地址 | 1111111011        | FEC0::/10 |
| 环回地址     | 00...1 (128 bits) | ::1/128   |
| 未指定地址    | 00...0 (128 bits) | ::/128    |
| 全球单播地址   | 其他                | -         |

表 5-20 中各类地址的意义如下：

- 链路本地单播地址用于邻居发现协议和无状态自动配置进程中链路本地节点之间的通信。使用链路本地地址作为源或目的地址的数据包不会被转发到其他链路上。使用链路本地前缀 FE80::/10(1111 1110 10)和 IEEE EUI-64 格式的接口标识符（EUI-64 可来源于 EUI-48）可在任意接口对其进行自动配置。
- 站点本地单播地址：根据 RFC 4291，站点本地单播地址被完全当作全球单播地址使用。
- 环回地址 0:0:0:0:0:0:0:1 或::1，不会被分配给任何接口。它的作用与在 IPv4 中的 127.0.0.1 相同，即节点将 IPv6 报文发送给自己。
- 未指定地址 (::)，不能被分配给任何节点，也不能作为目的地址。在主机初始化且没有取得自己的地址时，未指定地址可以用在 IPv6 报文的源地址字段，例如重复地址探测时，NS 报文的源地址就是未指定地址。

- 全球单播地址等同于 IPv4 公网地址。用于可以聚合的链路，最后提供给网络服务提供商。这种地址类型的结构允许路由前缀的聚合，从而满足全球路由表项的数量限制。地址包括运营商管理的 48 位路由前缀和本地站点管理的 16 位子网 ID，以及 64 位接口 ID。如无特殊说明，全球单播地址包括站点本地单播地址。

## IEEE EUI-64 格式的接口标识符

IPv6 地址中的 64 位接口标识符（Interface ID）用来标识链路上的唯一接口。这个地址是从接口的链路层地址（如 MAC 地址）变化而来的。IPv6 地址中的接口标识符是 64 位，而 MAC 地址是 48 位，因此需要在 MAC 地址的中间位置插入十六进制数 FFFE（1111 1111 1111 1110）。然后将 U/L 位（从高位开始的第 7 位）设置为“1”，这样就得到了 EUI-64 格式的接口 ID。如图 5-46 所示。

图 5-46 MAC 地址到 EUI-64 格式的转换过程

```
MAC:                0012:3400:ABCD

Binary:
00000000 00010010 00110100 00000000 10101011 11001101

Insert FFFE:
00000000 00010010 00110100 11111111 11111110 00000000
10101011 11001101

Set U/L bit:
00000010 00010010 00110100 11111111 11111110 00000000
10101011 11001101

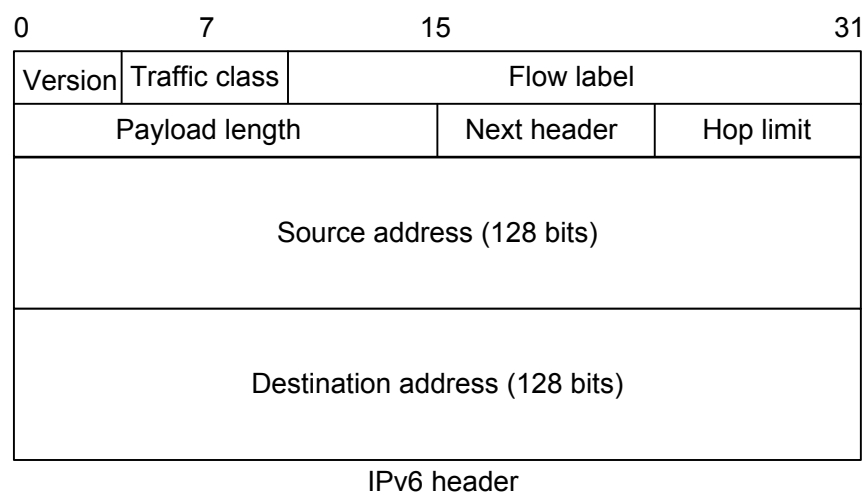
EUI-64:             0012:34FF:FE00:ABCD
```

### 5.7.4.3 IPv6 报文

#### IPv6 的报文头格式

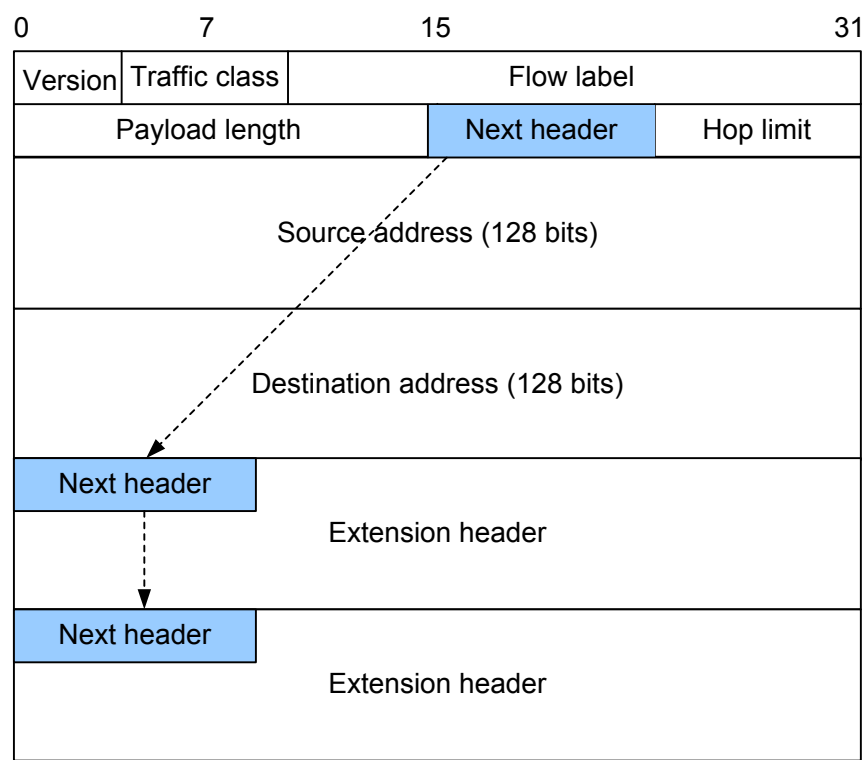
图 5-47 所示为 IPv6 报文头结构。

图 5-47 IPv6 报文头格式



- Version（版本）：该字段表示 IP 版本，值为 6。
- Traffic class（流量类别）：该字段及其功能类似于 IPv4 的业务类型字段。该字段以区分业务编码点（DSCP）标记一个 IPv6 数据包，以此指明数据包应当如何处理。
- Flow label（流标签）：该字段用来标记 IP 数据包的一个流，当前的标准中没有定义如何管理和处理流标签的细节。
- Payload length（有效载荷长度）：该字段表示有效载荷的长度，有效载荷是指紧跟 IPv6 基本报头的数据包，包含 IPv6 扩展报头。
- Next header（下一报头）：该字段指明了跟随在 IPv6 基本报头后的扩展报头的信息类型。如[图 5-48](#)所示。

图 5-48 Next header 在 IPv6 报文头中的作用



- **Hop limit（跳数限制）：**该字段定义了 IPv6 数据包所能经过的最大跳数，这个字段和 IPv4 中的 TTL 字段非常相似。
- **Source address（报文源地址）：**该字段表示该报文的源地址。
- **Destination address（报文目的地址）：**该字段表示该报文的目的地地址。

和 IPv4 报文头的比较

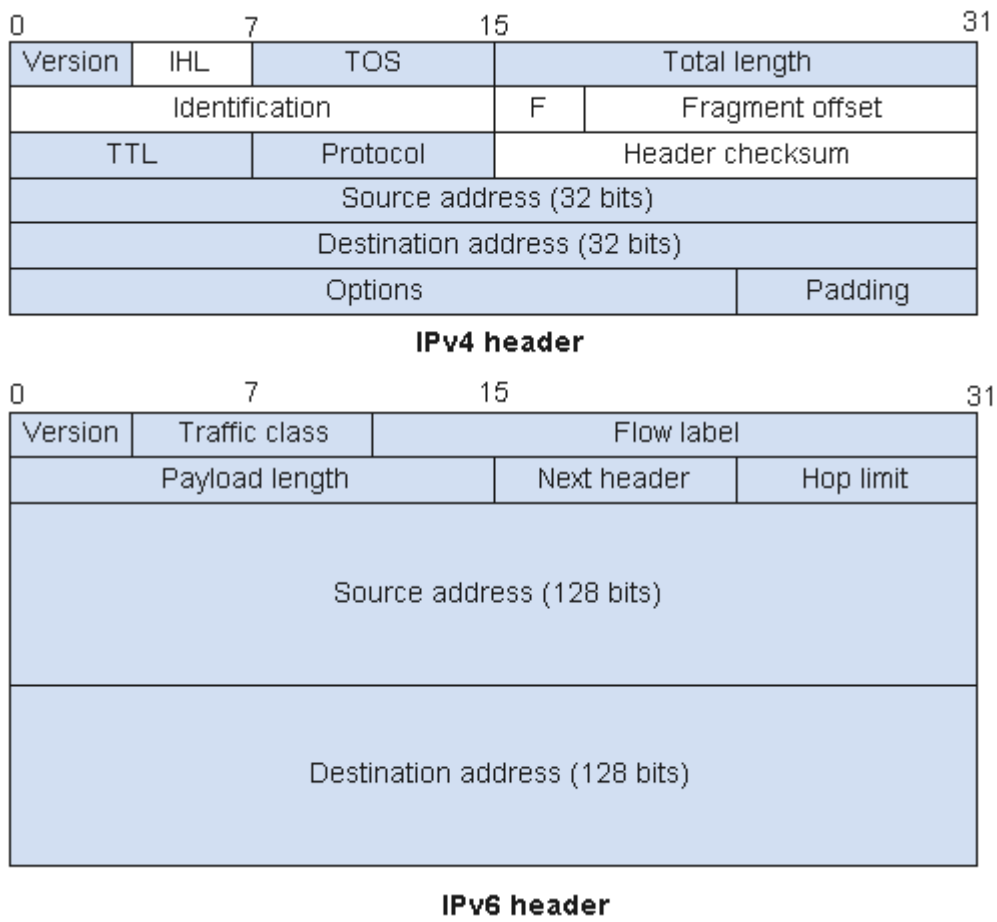
IPv6 通过将不重要的字段和选项字段移入扩展报头来减小报头的负载，使中间路由设备对报文的处理更有效。

尽管 IPv6 地址长度是 IPv4 地址长度的四倍，但 IPv6 基本报头的长度只有 IPv4 报头的两倍。

IPv4 和 IPv6 报头不具有互操作性，而且 IPv6 协议不能向后兼容 IPv4 协议。为了识别和处理两种报头格式，主机或路由设备必须同时运行 IPv4 和 IPv6 两种协议。如图 5-49 所示。



图 5-49 IPv4 和 IPv6 报文头格式比较

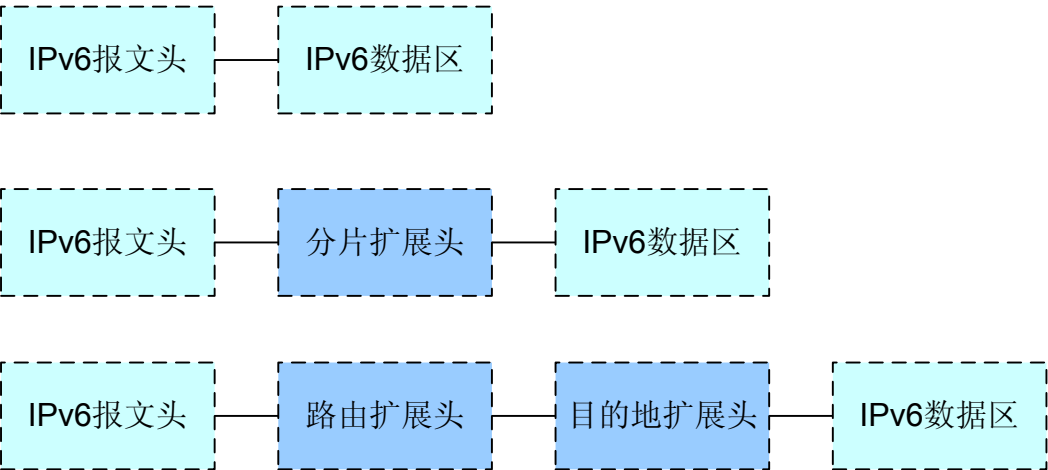


IPv6 的扩展报文头

IPv4 报头只能支持 40 字节的选项，而 IPv6 扩展报头的大小只受到 IPv6 报文大小的限制。

IPv6 取消了 IPv4 报头中的选项字段，并引入了多种扩展报文头，在提高处理效率的同时还增强了 IPv6 的灵活性，为 IP 协议提供了良好的扩展能力。如图 5-50 所示。

图 5-50 IPv6 扩展报文头



当超过一种扩展报头被用在同一个分组里时，报头必须按照下列顺序出现：

- IPv6 基本报头
- 逐跳选项扩展报头
- 目的选项扩展报头
- 路由扩展报头
- 分片扩展报头
- 授权扩展报头
- 封装安全有效载荷扩展报头
- 目的选项扩展报头（指那些将被分组报文的最终目的地处理的选项。）
- 上层扩展报头

不是所有的扩展报头都需要被转发路由设备查看和处理的。路由设备转发时根据基本报头中 Next Header 值来决定是否要处理扩展头。

除了目的选项扩展报头出现两次（一次在路由扩展报头之前，另一次在上层扩展报头之前），其余扩展报头只出现一次。

5.7.4.4 ICMPv6 报文

ICMPv6 报文格式

ICMPv6（Internet Control Message Protocol for the Internet Protocol Version 6）是 IPv6 的基础协议之一，具有差错报文和信息报文两种，用于 IPv6 节点报告报文处理过程中的错误和信息。ICMPv6 报文的报文格式如图 5-51 所示。

图 5-51 ICMPv6 报文格式

|                         |          |              |    |    |
|-------------------------|----------|--------------|----|----|
| 0                       | 7        | 15           | 23 | 31 |
| Type (1)                | Code (1) | Checksum (2) |    |    |
| Packet Content<br>..... |          |              |    |    |

图 5-51 中括号里的数字表示字段的长度，单位是字节。

报文中各个字段的解释如下：

- Type 字段表明消息的类型，数值为 0 至 127 表示差错报文类型，数值为 128 至 255 表示消息报文类型。
- Code 字段表示此消息类型细分的类型。
- Checksum 表示 ICMPv6 报文的校验和。

## ICMPv6 错误报文的分类

ICMPv6 错误报文分为以下几种：

- 目的不可达错误报文  
在 IPv6 节点转发 IPv6 报文过程中，发现目的地址不可达时，就会向发送报文的源节点发送 ICMPv6 目的不可达错误报文。同时报文中会携带引起该错误报文的具体原因。目的不可达错误报文又细分为以下几种：
  - 没有到目的地的路由
  - 地址不可达
  - 端口不可达
- 数据包过大错误报文  
在 IPv6 节点转发 IPv6 报文过程中，发现报文超过出接口的链路 MTU 时，则向发送报文的源节点发送 ICMPv6 数据包过大错误报文，其中携带出接口的链路 MTU 值。数据包过大错误报文是 Path MTU 发现机制的基础。
- 时间超时错误报文  
在 IPv6 报文收发过程中，当路由器收到 Hop Limit 值等于 0 的数据包，或者当路由器将 Hop Limit 值减为 0 时，会向报文的源节点发送 ICMPv6 超时错误报文。对于分段重组报文的操作，如果超过定时时间，也会产生一个 ICMPv6 超时报文。
- 参数错误报文  
当目的节点收到一个 IPv6 报文时，会对报文进行有效性检查，如果发现以下问题会向报文的源节点回应一个 ICMPv6 参数错误报文。
  - IPv6 基本头或扩展头的某个域有错误
  - IPv6 基本头或扩展头的 Next Header 值不可识别
  - 扩展头中出现未知的选项

## ICMPv6 信息报文的分类

ICMPv6 信息报文分为请求信息（Echo Request）和应答信息（Echo Reply）。路由器可以利用 ICMPv6 报文实现网络故障诊断、PMTU 发现和邻居发现等功能。在两节点的互通性检测中，收到 Echo Request 报文的节点向源节点回应 Echo Reply 报文，实现两节点间报文的收发。

## 5.8 IPv6 邻居发现

邻居发现 ND 是确定邻居节点之间关系的一组消息和进程，安全邻居发现 SEND 协议是 ND 的增强。

## 5.8.1 介绍

### 定义

邻居发现 ND (Neighbor Discovery) 是确定邻居节点之间关系的一组消息和进程。邻居发现协议替代了 IPv4 的 ARP (Address Resolution Protocol)、ICMP 路由器发现 (Router Discovery) 和 ICMP 重定向 (Redirect) 消息，并提供了其他功能。

安全邻居发现 SEND 协议是 ND 的增强，通过引入新的消息类型和选项字段，在地址所有权证明、消息保护和路由器授权方面提高了 ND 的安全性。

### 目的

邻居发现协议 ND 解决了 IPv6 网络中同一链路上不同节点之间的信息交互和地址自动配置问题。

由于 ND 建立在完全可信网络的基础上，其前提是假定所有节点都严格按照协议标准发送正常的 ND 消息，因此存在一定的安全威胁，常见的安全威胁有以下几种：

- **NS/NA 欺骗**  
攻击者发送包含虚假链路层地址的 NS/NA 消息更新被攻击者的邻居缓存，导致被攻击者将报文转发到虚假的链路层地址，类似于 IPv4 中的 ARP 欺骗。
- **DAD 攻击**  
攻击者通过响应所有的 DAD 消息，声称已使用被攻击者请求的地址，被攻击者将无法获取 IPv6 地址，无法与其他节点正常通信。
- **重定向攻击**  
攻击者使用被攻击者的缺省网关（第一跳路由器）的链路层地址作为源地址，发送重定向报文给被攻击者，将其下一跳路由定向到一个不存在的地址，从而造成被攻击者的通信中断。
- **参数欺骗**  
攻击者冒充本地路由器发送伪造的 RA 消息，消息中包含一虚假网段的地址前缀，并设置 Autonomous 标志。被攻击者将使用此地址前缀进行无状态地址自动配置，生成一个虚假网段的 IPv6 地址。当被攻击者以此地址为源地址与外界通信时，外界回应的报文将被本地路由器丢弃，导致被攻击者无法正常通信。
- **重放攻击**  
攻击者截获节点发送的消息，并在一段时间后重新发送该消息，使被攻击者接收过期的消息。

通过应用 SEND 特性，可以有效地防范上述的安全威胁，提高 ND 的安全性。

## 5.8.2 参考标准和协议

与 SEND 特性相关的参考标准与协议如下：

- RFC2461  
Neighbor Discovery for IP Version 6 (IPv6)
- RFC2462  
IPv6 Stateless Address Autoconfiguration

- RFC3756  
IPv6 Neighbor Discovery (ND) Trust Models and Threats
- RFC3971  
SEcure Neighbor Discovery (SEND)
- RFC3972  
Cryptographically Generated Addresses (CGA)

### 5.8.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

#### 特性依赖

SEND 建立在 ND 的基础上，依赖于 ND 的正常运行。

### 5.8.4 原理描述

#### 5.8.4.1 IPv6 邻居发现

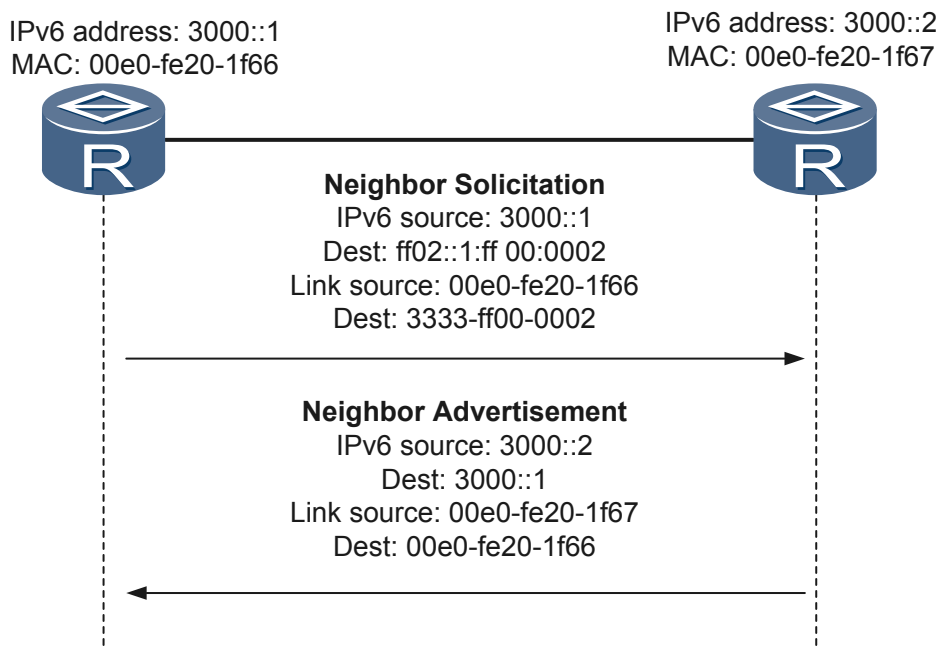
对于一个节点而言，当其配置一个 IPv6 地址之后，首先会确定此地址是否可用、不冲突。当节点是主机时，路由器需要通知主机向特定目的地址转发报文的更好的下一跳地址；当节点是另一台路由器时，路由器需要发布自己的地址、地址前缀和其他配置参数以指导主机进行参数配置。在 IPv6 报文转发过程中，节点需要确定邻居节点的链路层地址和其可达性。IPv6 邻居发现机制提供了 5 种不同类型的 ICMPv6 报文：

- 路由器请求报文 RS（Router Solicitation）：主机启动后，通过 RS 报文向路由设备发出请求，路由设备则会以 RA 报文响应。
- 路由器通告报文 RA（Router Advertisement）：路由设备周期性的发布 RA 报文，其中包括前缀和一些标志位的信息。
- 邻居请求报文 NS（Neighbor Solicitation）：IPv6 节点通过 NS 报文可以得到邻居的链路层地址，检查邻居是否可达，也可以进行重复地址检测。
- 邻居通告报文 NA（Neighbor Advertisement）：NA 报文是 IPv6 节点对 NS 报文的响应，同时 IPv6 节点在链路层变化时也可以主动发送 NA 报文。

- 重定向报文（Redirect）：路由设备发现报文的入接口和出接口相同时，可以通过重定向报文通知主机选择另外一个更好的下一跳地址。

图 5-52 所示为 IPv6 邻居发现过程。

图 5-52 IPv6 邻居发现过程示意图



IPv6 邻居发现协议主要包括以下功能：

## 地址冲突检测功能

地址冲突检测 DAD（Duplicate address detect）是确定 IPv6 地址是否可用的一种探测机制。具体执行过程如下：

1. 当一个节点配置了 IPv6 地址，为了查看该地址是否被其他邻居节点所使用，会即时发送邻居请求报文来确定其可用性。
2. 当其他邻居节点收到该报文后会查找本地的 IPv6 地址中是否存在相同的 IPv6 地址，若存在会回应一个邻居通告报文给源节点，并携带此 IPv6 地址信息。
3. 源节点收到邻居的回应报文则认为该 IPv6 地址已被邻居使用。反之，如果源节点发出的邻居请求报文没有收到相应的回应报文，则表示配置的 IPv6 地址是可用的。

## 邻居发现功能

邻居发现功能和 IPv4 中的 ARP 功能类似，主要实现对邻居地址的解析和邻居可达性的探测，依赖于邻居请求和邻居通告报文完成。

当一个节点需要得到同一本地链路上另外一个节点的链路层地址时，就会发送 ICMPv6 类型为 135 的邻居请求报文。此报文类似于 IPv4 中的 ARP 请求报文，不过使用组播地址而不使用广播地址，只有被请求节点的最后 24 比特和此组播地址相同的节点才会收到此报文，减少了广播风暴的可能。目的节点在响应报文中填充其链路层地址。

邻居请求报文也用来在邻居的链路层地址已知时，验证邻居的可达性。IPv6 邻居通告报文是对 IPv6 邻居请求报文的响应。收到邻居请求报文后，目的节点通过在本地链路上发送 ICMPv6 类型为 136 的邻居通告报文进行响应。收到邻居通告后，源节点和目的节点可以进行通信。当一个节点的本地链路上的链路层地址改变时也会主动发送邻居通告报文。

## 路由器发现功能

路由器发现功能用来定位邻居路由设备，同时学习和地址自动配置有关的前缀和配置参数。IPv6 路由发现由下面两种机制实现：

- 路由器请求

当主机没有配置单播地址时（例如系统刚启动），就会发送路由器请求报文 RS。路由器请求报文有助于主机迅速进行自动配置而不必等待 IPv6 路由设备的周期性路由器通告报文。IPv6 路由器请求也是 ICMPv6 报文，类型为 133。

- 路由器通告

每个 IPv6 路由设备的接口在配置了 IPv6 RA 去抑制的前提下会周期发送路由器通告报文。在本地链路上收到 IPv6 节点的路由器请求报文后，路由设备也会回应路由器通告报文。IPv6 路由器通告报文发送到所有节点多播地址（FF02::1）或发送路由器请求报文节点的 IPv6 单播地址。路由器通告为 ICMPv6 报文，类型为 134，包含以下内容：

- 是否使用地址自动配置
- 标记支持的自动配置类型（无状态或有状态自动配置）
- 一个或多个本地链路前缀（本地链路上的节点可以使用这些前缀完成地址自动配置）
- 通告的本地链路前缀的生存期
- 发送路由器通告的路由设备是否可作为缺省路由设备，如果可以，还包括此路由设备可作为缺省路由设备的时间（用秒表示）
- 和主机相关的其他信息，如跳数限制、主机发起的报文可以使用的最大 MTU

本地链路上的 IPv6 节点接收路由器通告报文，并用其中的信息得到更新的缺省路由设备、前缀列表以及其他配置。

## 地址自动配置功能

通过使用路由器通告报文和针对每一前缀的标记，路由设备可以通知主机如何进行地址自动配置。例如，路由设备可以指定主机是使用有状态（DHCPv6）地址配置还是无状态地址自动配置进行地址配置。

对于无状态地址自动配置而言，当主机收到路由器通告报文后，使用其中的前缀信息和本地接口 ID 自动形成 IPv6 地址，同时还可以根据其中的默认路由设备信息设置默认路由设备。

## 重定向功能

重定向报文用来通知主机去往目的地的更好的下一跳 IPv6 地址。和 IPv4 类似，IPv6 路由设备发送重定向报文的目的仅限于把报文重新路由到更合适的路由设备。收到重定向报文的节点随后会把后续报文发送到更合适的路由设备。路由设备只针对单播流发送重定向报文，重定向报文只发送给引起重定向的报文的节点（主机），并被处理。

### 5.8.4.2 IPv6 安全邻居发现

SEND 在 ND 的基础上引入了 CGA（Cryptographically Generated Addresses）、RSA（Rivest Shamir and Adleman）、Timestamp 和 Nonce 选项字段和 CPS（Certification Path Solicitation）、CPA（Certification Path Advertisement）两种消息类型。通过这些新的选项字段和消息类型，SEND 可以提供如下的安全增强功能：

- 地址所有权证明  
CGA 实现了 IPv6 地址和报文的绑定，避免 IPv6 地址被恶意盗用。通信双方通过生成和验证 CGA，可以防止地址欺骗，有效地抵御了 NS/NA 欺骗和 DAD 攻击。
- 消息保护  
通过 RSA 签名和验证，实现了消息完整性保护。同时，通信双方通过检查 Timestamp 和 Nonce 选项，增强了消息的时效性，有效地抵御了重放攻击。
- 路由器授权  
通过证书验证机制，实现了路由器的身份验证，防止攻击者冒充路由器发送恶意报文，有效地抵御了重定向攻击和参数欺骗。

## CGA

CGA 是通过公钥结合 HASH 算法生成的一个 IPv6 地址，节点通过验证 CGA 地址，丢弃与 CGA 不符的报文，防范欺骗性攻击。结合 RSA 签名机制，还可以实现报文的完整性保护。

CGA 和 RSA 签名生成过程如下：

1. 获取 RSA 密钥对。
2. 生成 CGA 参数表，包含公钥等信息。
3. 对 CGA 参数表进行 HASH 运算，输出散列值，取后 64 位作为网络 ID。
4. 将前缀信息和网络 ID 组合，生成 CGA 地址。
5. 构造报文，源地址使用 CGA 地址，同时将 CGA 参数表填充到 CGA 选项中。然后使用私钥对报文进行签名，将签名填充到 RSA 选项中。

当节点收到带有 CGA 选项和 RSA 选项的报文后，验证过程如下：

1. 从报文的 CGA 选项中获取 CGA 参数表。
2. 对 CGA 参数表进行 HASH 运算，输出散列值，取后 64 位作为网络 ID。
3. 检查生成的网络 ID 是否与报文源地址的网络 ID 匹配。
4. 从 CGA 参数表中获取公钥，验证 RSA 签名。

生成 CGA 地址后，接口发送 ND 报文时将会遵循如下规则：

- 接口发送的 NS（不包含 DAD 消息）、NA、RA 和 Redirect 消息的源地址都将使用 CGA 地址。
- 接口发送的 NS、NA、RA 和 Redirect 消息中将会带有 CGA 选项，内容为 CGA 参数表。
- 接口发送的 NS、NA、RA 和 Redirect 消息中将会带有 RSA 选项，内容为签名。
- 接口发送的 NS、NA、RA 和 Redirect 消息中会带有 Timestamp 选项，内容为设备当前的时间。



- 接口发送的 NS 消息中会带有 Nonce 选项，内容为随机数。接口回复的 NA 消息中也会带有 Nonce 选项，内容为接口收到的 NS 消息中的 Nonce 值。

## Timestamp 和 Nonce

Timestamp 指的是 ND 报文中的时间标签，主要用于非 NS/NA 消息交互过程中对重放攻击进行防范。开启 SEND 功能后，节点维护了 Delta 和 Fuzz 参数。当节点收到 ND 报文时，根据 RFC3971 中定义的公式，对报文的时效性进行检测，丢弃不附合要求的报文。

Nonce 是一个随机数，可以看作是当前会话的标签，主要用于 NS/NA 消息交互过程中对重放攻击进行防范。节点发送 NS 消息请求其他节点的链路层地址时，生成一个随机数置于 NS 消息中。接收节点在响应该请求时，回复的 NA 消息中必须带有该随机数，以表明回复的 NA 消息是针对当前的 NS 消息。

## 路由器授权

为了防止路由器被攻击者冒充，SEND 引入了两个新的消息类型 CPS 和 CPA，用于路由器的身份认证。

路由器必须先向 CA（Certificate Authority）服务器申请证书，证书中包含路由器的身份信息、公钥信息以及 CA 的数字签名信息。在无状态自动配置环境中，当主机收到 RA 消息后，发送 CPS 消息请求路由器的证书。路由器通过 CPA 消息发送自己的证书，响应主机的请求。主机收到 CPA 消息后，对消息中的证书进行验证，只有通过验证的路由器才会被主机作为默认路由器。

## 5.9 IPv6 性能

在一些特定的网络环境里，需要调整 IPv6 参数，以便网络性能达到最佳。

### 5.9.1 介绍

#### 定义

在特定的网络环境中，可以配置 TCP6、UDP6 和 PMTU 等属性等来调整设备的 IPv6 性能。

#### 目的

在不同的网络环境下，对特定的影响 IPv6 性能的指标进行调整，使网络业务可以更稳定的运行。

### 5.9.2 参考标准和协议

与 IPv6 性能相关的参考标准与协议如下：

- RFC 4022: Management Information Base for the Transmission Control Protocol (TCP)
- RFC 768: User Datagram Protocol
- RFC 4727: Experimental Values in IPv4, IPv6, ICMPv4, ICMPv6, UDP, and TCP Headers

## 5.9.3 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 5.9.4 原理描述

### TCP6

TCP6 提供了在两个端点的进程间建立虚电路的机制，一个 TCP6 虚连接如同在系统间承载数据的全双工电路。由于 TCP6 中提供了进程间数据的可靠传输，因此被称为可靠协议，它还提供了根据当前网络状态来优化传输性能的机制。在所有数据均可收到和确认的情况下，传输速率可以逐渐增加。延时将导致发送主机在收到进一步的确认前降低发送速率。

TCP6 通常用于交互式应用，如 Web 之类，某些数据接收差错将影响正常的工作能力。TCP6 使用了“三次握手”机制来建立虚电路，所有的虚电路都需使用“四次握手”拆除。这种连接方式可以提供多种校验和及其他可靠性功能，但是增加了使用 TCP6 的开销并导致其效率低于 UDP6。

### UDP6

UDP6 是用来在互连网络环境中提供包交换的计算机通信协议。有如下特点：

- 只使用源和目的信息，主要用于简单的请求/响应式结构。
- 不可靠，即没有任何控制能确定 UDP6 数据报是否已被接收。
- 无连接，即在主机间传输数据时，不需要任何类型的虚电路。

UDP6 的无连接特性使得 UDP6 可以向广播地址发送数据；而 TCP6 则不同，它要求特定的源地址和目的地址。

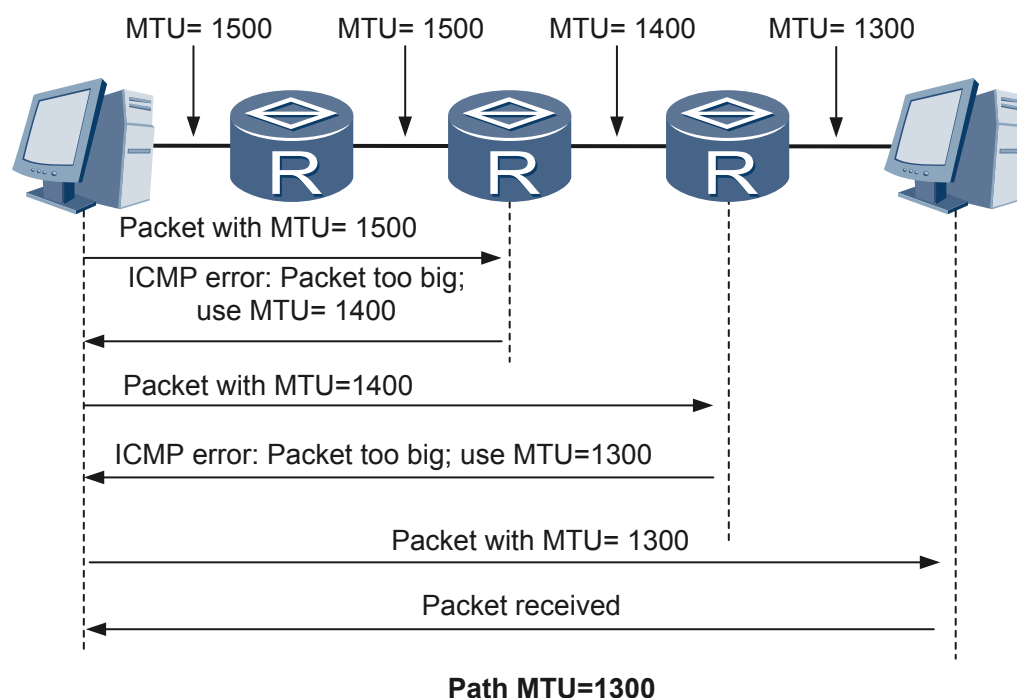
### PMTU

由于 IPv6 报文在传输过程中不允许在中间节点分片转发，所以在转发过程中经常会出现报文长度大于路径 IPv6 MTU 的情形，这就需要源节点不断的进行重传，降低了传输的效率，如果在源节点使用最小链接 IPv6 MTU（1280）作为分片的最大长度，在大多数情况下，路径的 IPv6 MTU 是大于最小链接的 IPv6 MTU 的，一个节点发出的分片远小于路径 IPv6 MTU，这是对网络资源的一种浪费，为了解决这个问题，提出了路径 MTU 发现协议。

Path MTU（以下简称 PMTU）发现，是确定从源端到目的端路径上合适的 IPv6 MTU 值的一种机制。PMTU 发现协议描述了一种动态发现任意路径的 PMTU 的方法。当一个 IPv6 节点发送大量数据到另一节点时，数据通过一系列 IPv6 分片传送。当这些分片具有从源节点到宿节点能够成功传送所允许的最大长度时，我们认为它达到理想状态，这个分片长度被称为路径 MTU。

如图 5-53 所示，PMTU 发现的工作过程是：源端主机先使用自己的 MTU 值向目的主机发送报文，如果中间路由设备给源端返回一个错误消息，其中包括该网络的 MTU 值，源端主机使用该 MTU 值来重新发送这个报文，如此反复，直到目的端主机收到这个报文，从而确定网络中两台主机之间能够处理的最大报文的大小。

图 5-53 PMTU 发现的工作过程



## 5.10 DHCPv6

在 IPv6 网络中，客户端可以通过 DHCPv6 技术动态获取 IPv6 地址等参数。

### 5.10.1 介绍

#### 定义

DHCPv6（Dynamic Host Configuration Protocol for IPv6）是针对 IPv6 网络中的动态编址方案设计的，可以为客户端分配 IPv6 地址/前缀和其他网络配置信息。

## 目的

DHCPv6 提供了一种机制来简化 IPv6 地址/前缀的配置，并实现 IPv6 地址/前缀的集中式管理。通过 DHCPv6 技术，把手工配置 IPv6 地址所导致的错误减少到最低程度，并使网络管理员可以集中管理全局的配置信息。

## 5.10.2 参考标准和协议

与 DHCPv6 特性相关的参考标准与协议如下：

- RFC 3315: Dynamic Host Configuration Protocol for IPv6 (DHCPv6)
- RFC 2461: Neighbor Discovery for IP Version 6 (IPv6)
- RFC 3736: Stateless Dynamic Host Configuration Protocol (DHCP) Service for IPv6
- RFC 2462: IPv6 Stateless Address Autoconfiguration
- RFC 3633: IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6

## 5.10.3 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本   |
|-------------------------------------------------------------------------|----------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | V300R001 |

## 5.10.4 原理描述

### 5.10.4.1 DHCPv6 概述

DHCPv6 是针对 IPv6 网络中的动态编址方案设计的，可以为客户端分配 IPv6 地址/前缀和其他配置信息等。

### 客户端获得 IPv6 地址/前缀和配置信息的方式

客户端可以通过以下几种方式获得 IPv6 地址/前缀和配置信息：

- 手工配置：由用户静态配置 IPv6 地址等信息。
- 无状态地址自动配置：客户端通过邻居发现协议从本链路上的路由器设备获取 IPv6 前缀和相关配置信息。
- DHCPv6 - PD：由服务器和客户端分别充当 Delegating Router 和 Requesting Router。Delegating Router 为 Requesting Router 分配一段前缀以及配置信息，

Requesting Router 收到前缀以及配置信息后将这些信息当作本地拥有的资源再分配给其他客户端使用。

- 有状态 DHCPv6: 客户端通过与 DHCPv6 服务器的交互获取 IPv6 地址和配置信息。
- 无状态 DHCPv6: 客户端通过无状态地址自动分配方式获取 IPv6 前缀和跳数限制等信息, 再通过 DHCPv6 方式获取其他配置信息, 如 DNS 服务器地址等。

其中, 后面三种分配方式是 DHCPv6 在不同应用场景下的具体表现形式。

与手工配置方式、无状态地址自动配置方式相比, DHCPv6 方式具有以下特点:

1. 可以更好的管理网络资源
  - a. DHCPv6 服务器可以集中维护 IPv6 地址/前缀等信息, 并对这类信息进行租约更新等服务。
  - b. DHCPv6 服务器可以为客户端分配配置信息, 如 DNS 服务器地址, SIP 服务器地址等。
2. 可以更好的控制地址分配
  - a. 在 DHCPv6 - PD 场景中, DHCPv6 服务器可以为客户端分配一段前缀, 使客户端可以将这段前缀划分为多个子网并继续向下分配, 便于网络的自动配置和管理。
  - b. 可以将特定前缀与指定客户端绑定, 为客户端提供定制化前缀分配服务。

## DHCPv6 的通信模式

DHCPv6 采用客户端/服务器通信模式, 由客户端向服务器提出配置申请, 服务器返回为客户端分配的 IPv6 地址等配置信息, 以实现 IPv6 地址、前缀等信息的动态配置。

DHCPv6 信息交互时使用 UDP 端口 546 和 547。其中, 客户端通过 UDP 端口 546 监听 DHCPv6 消息, 而服务器和中继使用 UDP 端口 547 监听 DHCPv6 信息。

## DHCPv6 唯一标志符 DUID (DHCP Unique Identifier)

在 DHCPv6 的信息交互过程中, 客户端和服务端都需要拥有一个唯一的标志符 DUID, 用于互相标志对方。服务器可以用 Client DUID 标识客户端, 并可以将其用于本地的地址分配策略; 客户端可以用 Server DUID 来标识服务器。

DUID 包括四种生成方式:

- 手工配置: 由用户手工配置 DUID 值。
- DUID-LLT (DUID Based on Link-layer Address Plus Time): 基于时间和链路层地址生成 DUID 值。
- DUID-EN (DUID Assigned by Vendor Based on Enterprise Number): 基于 IANA 注册的企业号码生成 DUID 值。
- DUID-LL (DUID Based on Link-layer Address): 基于链路层地址生成 DUID 值。

## DHCPv6 协议组播地址

与 DHCPv4 类似, DHCPv6 客户端不用配置 DHCPv6 服务器的 IPv6 地址, 而是通过发送目的地址为组播的 Solicit 报文来定位 DHCPv6 服务器。在有多台服务器的情况下, 客户端可以通过一定的策略选择一个服务器 (如根据 Preference 选项)。

协议中定义了两个组播地址:

- All\_DHCP\_Relay\_Agents\_and\_Servers(FF02::1:2): 用于链路范围内所有的服务器和中继。当一个客户端发起 DHCPv6 交互时, 使用该地址与所有的服务器和中继进行交互。
- All\_DHCP\_Servers(FF05::1:3): 用于站点范围内所有的服务器。当中继需要转发报文到服务器, 但不知道服务器的单播地址时, 可以使用该组播地址转发报文。

5.10.4.2 DHCPv6 报文

服务器/客户端之间报文格式

服务器/客户端之间报文格式如图 5-54 所示。

图 5-54 服务器/客户端之间报文格式

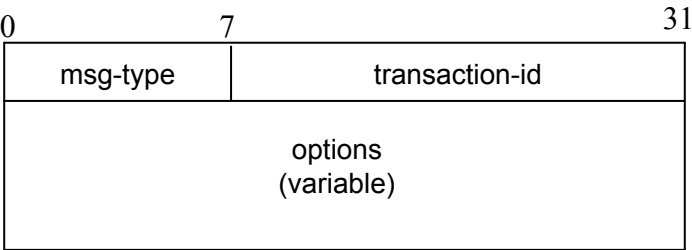


表 5-21 服务器/客户端之间报文中各个字段的含义

| 字段             | 长度   | 含义                                                                                                                                                                                                                                                                                                                                                                        |
|----------------|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| msg-type       | 1 字节 | 表示报文的类型： <ul style="list-style-type: none"><li>● Solicit (1)</li><li>● Advertise (2)</li><li>● Request (3)</li><li>● Confirm (4)</li><li>● Renew (5)</li><li>● Rebind (6)</li><li>● Reply (7)</li><li>● Release (8)</li><li>● Decline (9)</li><li>● Reconfigure (10)</li><li>● Information-Request (11)</li><li>● Relay-Forward (12)</li><li>● Relay-Reply (13)</li></ul> |
| transaction-id | 3 字节 | 用户选取的随机数, 使服务器的回复与客户端的报文相关联。                                                                                                                                                                                                                                                                                                                                              |
| Options        | 可变   | 用于服务器和客户端之间交互参数信息。                                                                                                                                                                                                                                                                                                                                                        |

## 服务器/客户端之间的报文类型

根据报文的发送方和接收方的不同，可以将这些报文进行如下分类：

- 客户端->服务器
  - Solicit (1): 客户端发送 Solicit 消息来定位服务器。
  - Request (3): 客户端发送 Request 消息向指定服务器请求配置信息，如 IPv6 地址。
  - Confirm (4): 客户端发送 Confirm 消息来确认自己当前的 IPv6 地址是否仍然有效。
  - Renew (5): 客户端向最初为自己提供地址和配置信息的服务器发送 Renew 消息来延长已获得地址的使用周期和更新其他配置信息。
  - Rebind (6): 如果客户端发出的 Renew 消息没有收到反馈，则客户端会发送 Rebind 消息到任意服务器来申请延长已获得地址的使用周期和更新其他配置信息；Requesting Router 发送 Rebind 消息来确认自己和 Delegating Router 的绑定关系是否依然有效。
  - Release (8): 客户端向为自己分配地址的服务器发送 Release 消息，声明不再继续使用该地址。
  - Decline (9): 客户端发送 Decline 消息，声明检测到服务器分配给自己的地址在链路上有冲突。
  - Information-Request (11): 客户端发送 Information-request 消息来请求除地址以外的配置信息。
- 服务器->客户端
  - Advertise (2): 服务器发送 Advertise 消息来回应客户端发出的 Solicit 消息，声明自己可以提供 DHCPv6 服务。
  - Reply (7): 服务器发送带有地址和配置信息的 Reply 消息来回应从客户端收到的 Solicit、Request、Renew 和 Rebind 消息；发送带有配置信息的 Reply 消息来回应从客户端收到的 Information-request 消息；发送 Reply 消息来回应 Confirm 消息，确认该地址在此链路上是否仍然有效；发送 Reply 消息回应 Release 或 Decline 消息。
  - Reconfigure (10): 当服务器上有新的配置信息或者配置信息有变更时，服务器向客户端发送 Reconfigure 消息来通知客户端进行更新。
- 中继->中继/服务器
  - Relay-Forward (12): 中继将从客户端或中继接收到的消息以 Relay-forward 消息方式发送给中继或服务器。
- 中继/服务器->中继
  - Relay-Reply (13): 服务器或中继将发送给客户端的信息以 Relay-reply 消息方式发送给中继。

## 服务器/客户端之间报文中的 Options 格式

Options 报文格式如图 5-55 所示。

图 5-55 Options 报文格式

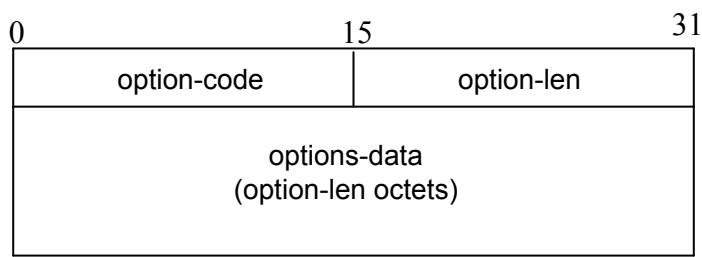


表 5-22 Options 报文中各个字段的含义

| 字段           | 长度   | 含义                       |
|--------------|------|--------------------------|
| option-code  | 2 字节 | Option 编码。               |
| option-len   | 2 字节 | 表示 options-data 区域的信息长度。 |
| options-data | 可变   | 此 Option 的具体数据。          |

常用 Options 举例如表 5-23 所示

表 5-23 Options 举例

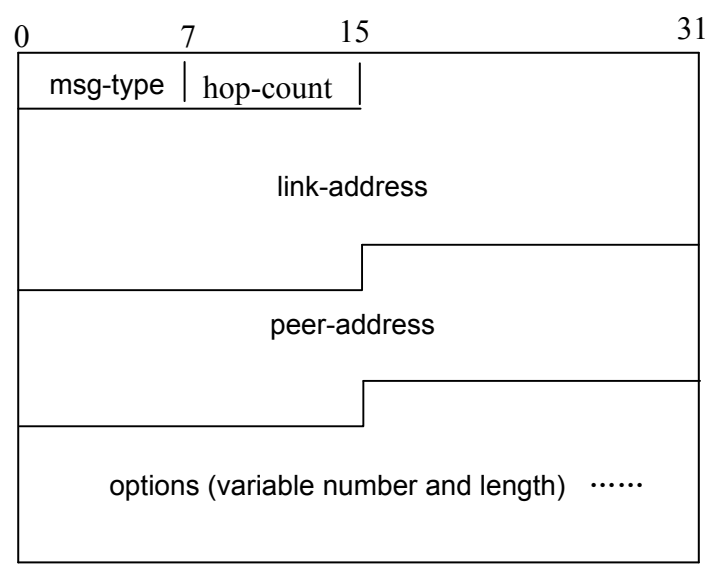
| option-code            | option-len          | options-data |
|------------------------|---------------------|--------------|
| OPTION_CLIENTID<br>(1) | 客户端 DUID 的长度，单位是字节。 | 客户端 DUID 的值。 |
| OPTION_SERVERID<br>(2) | 服务器 DUID 的长度，单位是字节。 | 服务器 DUID 的值。 |

服务器/中继之间报文格式

服务器/中继之间报文格式如图 5-56 所示。



图 5-56 服务器/中继之间报文格式



根据 msg-type 取值的不同，报文中各字段的取值也有所不同。具体可以分为表 5-24 和表 5-25 两种情况：

表 5-24 服务器/中继之间报文中各个字段的含义（msg-type 取值为 RELAY-FORWARD）

| 字段           | 长度    | 含义                               |
|--------------|-------|----------------------------------|
| msg-type     | 1 字节  | 表示报文的类型，此处取值为 RELAY-FORWARD。     |
| hop-type     | 3 字节  | 中继过此消息的中继个数。                     |
| link-address | 16 字节 | 一个全球地址或本地站点地址，服务器用该地址来确定客户端所在链路。 |
| peer-address | 16 字节 | 该消息的上一跳的地址。                      |
| options      | 可变    | 用于服务器和中继之间交互参数信息。                |

表 5-25 服务器/中继之间报文中各个字段的含义（msg-type 取值为 RELAY-REPLY）

| 字段           | 长度    | 含义                        |
|--------------|-------|---------------------------|
| msg-type     | 1 字节  | 表示报文的类型，此处取值为 RELAY-REPL。 |
| hop-type     | 3 字节  | 从 Relay-froward 消息中拷贝。    |
| link-address | 16 字节 | 从 Relay-froward 消息中拷贝。    |

| 字段           | 长度    | 含义                               |
|--------------|-------|----------------------------------|
| peer-address | 16 字节 | 从 Relay-froward 的 options 消息中拷贝。 |
| options      | 可变    | 用于服务器和中继之间交互参数信息。                |

服务器/中继之间报文中的 Options 格式

Options 报文格式如图 5-57 所示。

图 5-57 Options 报文格式

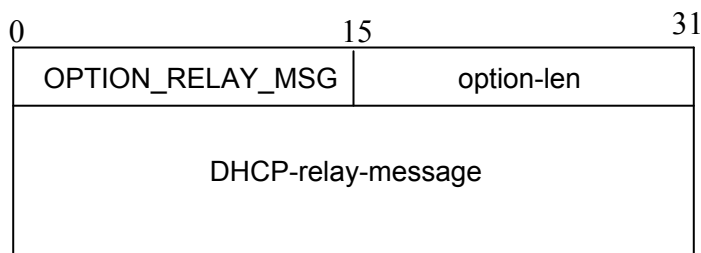


表 5-26 Options 报文中各个字段的含义

| 字段                 | 长度   | 含义                             |
|--------------------|------|--------------------------------|
| OPTION_RELAY_MSG   | 2 字节 | Option 编码。                     |
| option-len         | 2 字节 | 表示 DHCP-relay-message 区域的信息长度。 |
| DHCP-relay-message | 可变   | 此 Option 的具体数据。                |

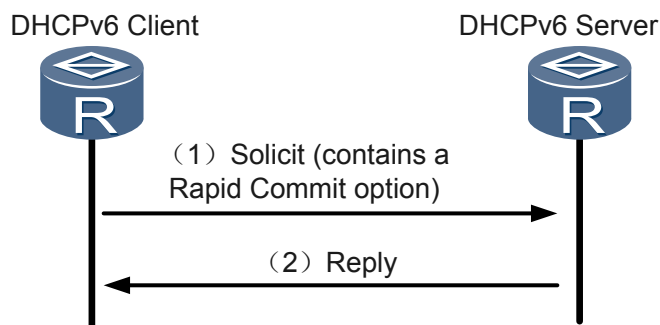
5.10.4.3 DHCPv6 工作原理

DHCPv6 客户端与服务器之间的交互模式

DHCPv6 采用客户/服务器通信模式，客户端为了获取合法的动态 IPv6 地址、前缀和配置信息，会向服务器发送配置申请报文，服务器根据策略返回携带相应配置信息的报文。在不同应用场景中，客户端与服务器之间采用不同的交互模式，交互不同的信息。主要包括以下几种交互模式。

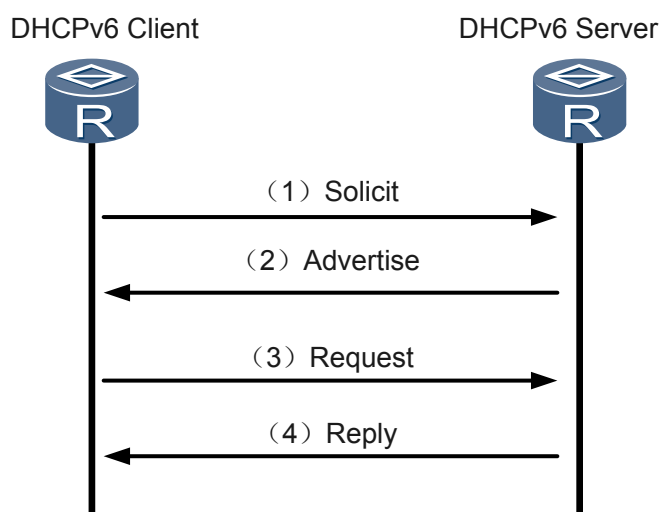
- 两步交互模式

图 5-58 两步交互模式流程图



1. DHCPv6 客户端首次登录网络时，以组播方式发送 DHCPv6 Solicit 报文，报文中携带 Rapid commit 选项。
  2. DHCPv6 服务器接收到客户端的 DHCPv6 Solicit 报文后，若 DHCPv6 服务器支持两步交互，则从 IPv6 地址/前缀池中挑选一个尚未分配的 IPv6 地址/前缀后，向该客户端以单播方式发送包含 IPv6 地址/前缀和其它配置信息的 DHCPv6 Reply 报文；否则以单播方式发送 DHCPv6 Advertise 报文，并进入四步交互模式。
- 四步交互模式

图 5-59 四步交互模式流程图



1. 发现阶段：客户端以组播方式发送 DHCPv6 Solicit 报文寻找 DHCPv6 服务器。
2. 提供阶段：DHCPv6 服务器接收到客户端的 DHCPv6 Solicit 报文后，从 IPv6 地址/前缀池中挑选一个尚未分配的 IPv6 地址/前缀分配给客户端，以单播方式向该客户端发送包含出租 IPv6 地址/前缀和其它配置的 DHCPv6 Advertise 报文。
3. 选择阶段：当网络中存在多台 DHCPv6 服务器向该客户端发来 DHCPv6 Advertise 报文时，客户端会根据配置策略选择其中的一个服务器，并以单播方式向该服务器发送 DHCPv6 Request 报文，申请希望获得的 IPv6 地址/前缀等内容。

4. 确认阶段：当 DHCPv6 服务器收到客户端反馈的 DHCPv6 Request 报文后，会以单播方式向客户端发送 DHCPv6 Reply 报文，报文中包含它所提供的 IPv6 地址/前缀和其它配置信息。

### 客户端延长 IPv6 地址/前缀的租约

DHCPv6 服务器分配给客户端的 IPv6 地址/前缀通常有一定的租借期限，期满后服务器会收回该地址/前缀。如果客户端希望继续使用地址/前缀，需要更新租约。

DHCPv6 服务器给客户端分配 IPv6 地址/前缀时，会携带首选生命周期值、有效生命周期值、续租时间和重绑定时间，这些周期和时间用来确定客户端获得的 IPv6 地址/前缀所处的状态和客户端需要进行的动作。其中续租时间 < 重绑定时间 < 首选生命周期 < 有效生命周期。

当客户端获得 IPv6 地址/前缀时，会进入到绑定状态，客户端会设置 3 个定时器，分别用来控制租期更新、重绑定和判断是否已经到达租期。定时器的缺省值如表 5-27 所示。

表 5-27 定时器的缺省值

| 定时器  | 默认值          |
|------|--------------|
| 租期更新 | 首选生命周期的 50%。 |
| 重绑定  | 首选生命周期的 80%。 |
| 到达租期 | 有效生命周期。      |

DHCPv6 服务器为客户端分配 IPv6 地址/前缀时，可以为定时器配置特定的值。若服务器没有设置定时器的值，客户端就使用缺省值。

- 当续租更新时间到达时，客户端会自动向服务器发送 DHCPv6 Renew 报文，以完成 IPv6 地址/前缀租约的更新。  
DHCPv6 服务器收到 DHCPv6 Renew 报文后，若 IPv6 地址/前缀有效且允许续租，则在回应给客户端的 DHCPv6 Reply 报文中携带新的 IPv6 地址/前缀租期。客户端收到回应的 DHCPv6 Reply 报文后，更新 IPv6 地址/前缀租期。
- 当重绑定时间到达时，若客户端未完成 IPv6 地址/前缀的续租，则以组播报文形式发送 DHCPv6 Rebind 报文到所有可达的 DHCPv6 服务器。  
DHCPv6 服务器收到 DHCPv6 Rebind 报文后，若 IPv6 地址/前缀有效且允许续租，则在回应给客户端的 DHCPv6 Reply 报文中携带新的 IPv6 地址/前缀租期。客户端收到回应的 DHCPv6 Reply 报文后，更新 IPv6 地址/前缀租期。
- 当 IPv6 地址/前缀到达租期时，DHCPv6 服务器会收回 IPv6 地址/前缀。如果客户端希望继续使用地址/前缀，需要在有效生命周期到达前，完成租期的更新。

### 客户端所处链路发生变化时，确认 IPv6 地址/前缀是否仍可用

客户端所处链路发生变化时，例如发生网线插拔等情况时，客户端需要向 DHCPv6 服务器发送报文确认之前获取的 IPv6 地址/前缀是否仍然可用。

根据确认的内容不同，可以分为以下两种情况：

- 若需要确认的是 IPv6 地址，则 DHCPv6 客户端以组播形式发送 DHCPv6 Confirm 报文，报文中携带需要确认的 IPv6 地址。

DHCPv6 服务器收到 DHCPv6 Confirm 报文后，若 IPv6 地址仍然可以使用，则在回应给客户端的 Reply 报文中设置该地址的状态为成功。客户端收到 Reply 报文后，可以继续使用该地址。

- 若需要确认的是 IPv6 前缀，则客户端以组播形式发送 DHCPv6 Rebind 报文，报文中携带需要确认的 IPv6 前缀。DHCPv6 服务器收到并处理 Rebind 报文，回应给客户端 Reply 报文。客户端收到 Reply 报文后，若报文中包含的前缀生命周期不为 0，则继续使用该前缀，并更新租期。

## 客户端发现 IPv6 地址冲突

若客户端检测到从 DHCPv6 服务器获取的 IPv6 地址存在冲突，则向 DHCPv6 服务器发送 DHCPv6 Decline 报文，通知 IPv6 地址冲突。

客户端向 DHCPv6 服务器发送携带包含冲突的 IPv6 地址的 DHCPv6 Decline 报文，DHCPv6 服务器收到 DHCPv6 Decline 报文后，将该地址标志为冲突状态，则此地址将不再继续被分配给任何客户端。

## 客户端释放 IPv6 地址/前缀

若客户端不再需要继续使用从 DHCPv6 服务器获取的 IPv6 地址/前缀时，可以向 DHCPv6 服务器发送 DHCPv6 Release 报文来释放该 IPv6 地址/前缀。

客户端向 DHCPv6 服务器发送携带包含需要释放的 IPv6 地址/前缀的 DHCPv6 Release 报文，DHCPv6 服务器收到 DHCPv6 Release 报文后，则释放该 IPv6 地址/前缀，并回应 Reply 报文。

# 5.11 IPv6 over IPv4 隧道

IPv6 over IPv4 隧道是为了在 IPv4 网络中传递 IPv6 节点间信息而出现的隧道技术。

## 5.11.1 介绍

### 定义

在从 IPv4 向 IPv6 的过渡中，IPv6 over IPv4 隧道作为一种隧道技术，用于在 IPv4 的网络中实现 IPv6 网络之间的互联。

### 目的

在从 IPv4 向 IPv6 发展的初期，IPv4 网络占据主导地位，IPv6 网络是一些孤岛，这些孤岛可能处于网络拓扑的边缘，也可能处于 IPv4 网络的包围中。IPv6 网络可能是一个由若干 IPv6 路由器和主机组成的网络，也可能只有一台 IPv6 主机。在这个阶段，IPv6 与 IPv4 共存技术和互通技术分别解决了两类过渡问题：多个 IPv6 孤岛互连，IPv6 和 IPv4 互通。

## 5.11.2 参考标准和协议

与 IPv6 over IPv4 隧道特性相关的参考标准与协议如下：

- RFC 2784: Generic Routing Encapsulation (GRE)

- RFC 2893: Transition Mechanisms for IPv6 Hosts and Routers
- RFC 3056: Connection of IPv6 Domains via IPv4 Clouds

5.11.3 可获得性

License 支持

本特性无需 License 支持。

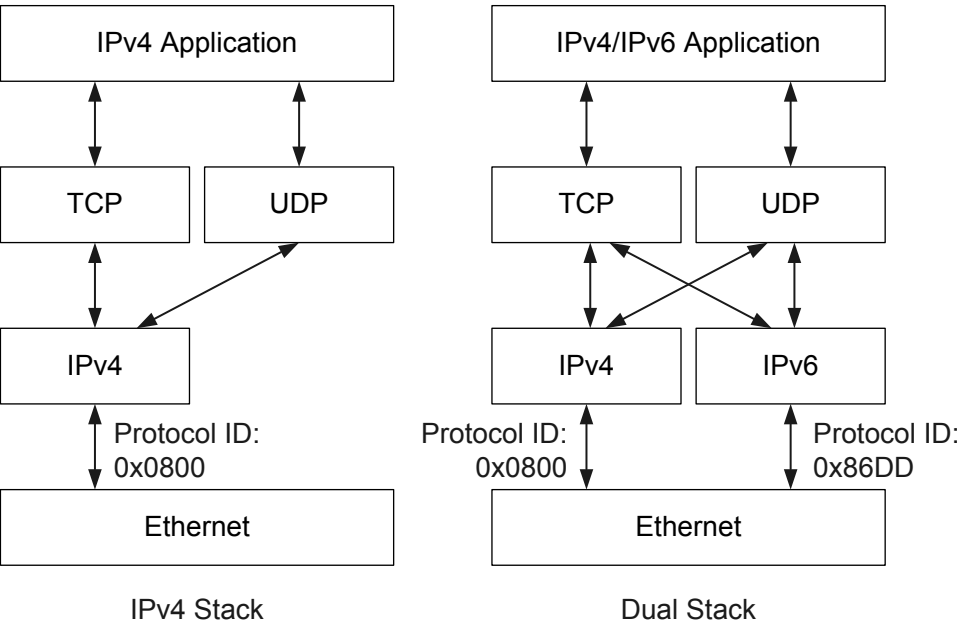
版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

5.11.4 原理描述

对于 IPv6 节点来说，兼容 IPv4 的最直接有效的办法就是保留一个完整的 IPv4 协议栈，这样的节点即为双协议栈节点。单协议栈和双协议栈结构示例如图 5-60 所示。

图 5-60 单协议栈与双协议栈结构（以太网）



双协议栈具有以下特点：

- 多种链路协议支持双协议栈  
多种链路协议（如以太网）支持双协议栈。图中的链路层是以太网，在以太网帧上，如果协议 ID 字段的值为 0x0800，表示网络层收到的是 IPv4 报文，如果为 0x86DD，表示网络层是 IPv6 报文。
- 多种应用支持双协议栈  
多种应用（如 DNS/FTP/Telnet 等）支持双协议栈。上层应用（如 DNS）可以选用 TCP 或 UDP 作为传输层的协议，但优先选择 IPv6 协议栈，而不是 IPv4 协议栈作为网络层协议。

## 5.12 IPv4 over IPv6 隧道

IPv4 over IPv6 隧道是在 IPv4 向 IPv6 过渡的后期用于联系 IPv4 孤岛的一种隧道技术。

### 5.12.1 介绍

#### 定义

在 IPv6 网络上用于连接 IPv4 孤岛的隧道技术，称为 IPv4 over IPv6 隧道技术。

#### 目的

在 IPv4 Internet 向 IPv6 Internet 过渡的后期，IPv6 网络已被大量部署，此时可能出现 IPv4 孤岛。利用隧道技术可在 IPv6 网络上创建隧道，从而实现 IPv4 孤岛的互连。

### 5.12.2 参考标准和协议

与 IPv4 over IPv6 隧道特性相关的参考标准与协议如下：

- RFC 2473: Generic Packet Tunneling in IPv6 Specification

### 5.12.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

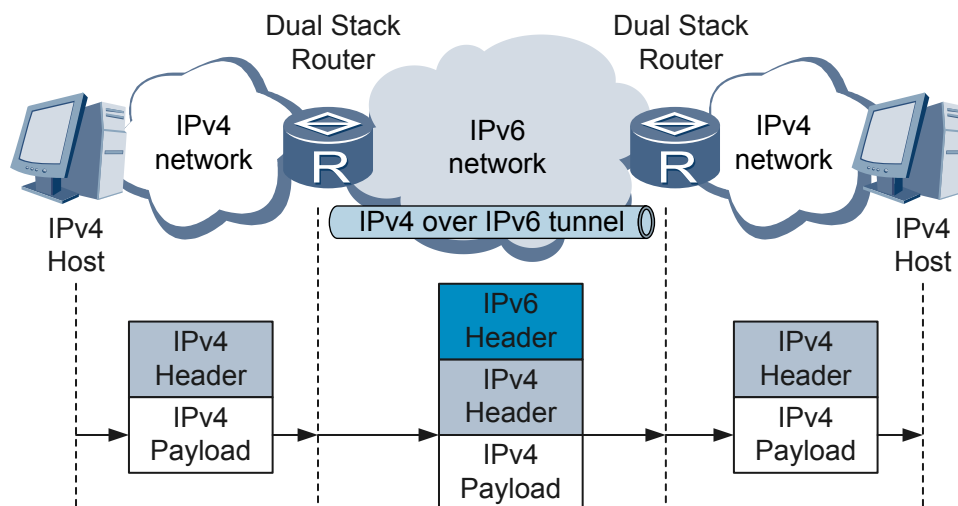
| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 5.12.4 原理描述

在 IPv4 Internet 向 IPv6 Internet 过渡后期，IPv6 网络被大量部署后，而 IPv4 网络只是散布在世界各地的一些孤岛。利用隧道技术可以在 IPv6 网络上创建隧道，从而实现 IPv4 孤岛的互联，IPv4 孤岛能通过 IPv6 公网访问其他 IPv4 网络。

IPv4 over IPv6 隧道技术的原理如图 5-61 所示，

图 5-61 IPv4 over IPv6 隧道原理图



1. 边界路由设备启动 IPv4/IPv6 双协议栈，并配置 IPv4 over IPv6 隧道。
2. 边界路由设备在收到从 IPv4 网络侧发来的报文后，如果报文的目的地址不是自身，就要把收到的 IPv4 报文作为负载，加上 IPv6 报文头，封装到 IPv6 报文里。
3. 在 IPv6 网络中，封装后的报文被传递到对端的边界路由设备。
4. 对端边界路由设备对报文解封装，去掉 IPv6 报文头，然后将解封装后的 IPv4 报文发送到对端的 IPv4 网络。



# 6 IP 路由

## 关于本章

### 6.1 静态路由

静态路由多使用在结构较简单的 IPv4 和 IPv6 网络中。仔细设计静态路由可以改进网络性能，并为重要的应用保证带宽。

### 6.2 RIP

RIP 协议主要用于规模较小的网络中，比如校园网以及结构较简单的地区性网络。对于更为复杂的环境和大型网络，一般不使用 RIP。RIP 的实现较为简单，配置和维护也比 OSPF 和 IS-IS 容易，在实际组网中有广泛的应用。

### 6.3 RIPng

RIPng 多用于规模较小和结构较简单的 IPv6 网络环境中，是基于距离矢量的路由协议，使用跳数（Hop Count）来衡量到达目的地址的距离。

### 6.4 OSPF

OSPF 多用于支持大型 IPv4 网络，是基于链路状态的路由协议，与基于距离矢量的路由协议相比具有更快的收敛速度。

### 6.5 OSPFv3

OSPFv3 多用于支持大型 IPv6 网络，是基于链路状态的路由协议，与基于距离矢量的路由协议相比具有更快的收敛速度。

### 6.6 IS-IS

IS-IS 协议在运营商中得到广泛的应用，是直接运行在链路层的内部网关路由协议，具有收敛速度快，分层结构清晰的特点，非常适合大规格网络。

### 6.7 BGP

BGP 是用于自治系统之间的外部网关协议，广泛用于 ISP 之间。

### 6.8 IP 单播策略路由

策略路由是一种依据用户制定的策略进行路由选择的机制，可应用于安全、负载分担等目的。

### 6.9 IPv6 单播策略路由

策略路由是一种依据用户制定的策略进行路由选择的机制，可应用于安全、负载分担等目的。

#### 6.10 路由策略

通过配置路由策略来影响网络流量经过的途径。目前实施路由策略的手段主要是对路由信息进行过滤。

## 6.1 静态路由

静态路由多使用在结构较简单的 IPv4 和 IPv6 网络中。仔细设计静态路由可以改进网络性能，并为重要的应用保证带宽。

### 6.1.1 介绍

#### 定义

静态路由是一种特殊的路由，它由网络操作人员手工输入到设备的配置信息中。

#### 目的

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

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

### 6.1.2 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 6.1.3 原理描述

静态路由的原理比较简单，管理者通过配置途径（如 CLI、SNMP）将路由添加到路由表中即可。转发模块采用最长匹配（Longest Match Algorithm）的原则找到相应的路由表项，并将报文转发到下一跳地址。

#### 缺省路由

如果到达某个指定网络的数据报文在路由器的路由表里找不到对应的表项，那么该报文将被路由器丢弃。

通过给当前路由器配置一条缺省路由，那些在路由表里找不到匹配路由表项的数据报文将会转发给另外一台路由器，由另外一台路由器进行报文的转发。

缺省路由是在路由器没有找到匹配的路由表入口项时才使用的路由：

- 如果报文的地址不能与路由表的任何项相匹配，那么该报文将选取缺省路由。
- 如果没有缺省路由且报文的地址不在路由表中，那么该报文将被丢弃，路由器将向源端返回一个 ICMP 报文报告该目的地址或网络不可达。

缺省路由有两种生成方式：

- 第一种是通过网络管理员在路由器上配置目的地址为 0.0.0.0（掩码也为 0.0.0.0）的 IPv4 静态路由，或者目的地址为 0:0::0:0（掩码长度为 0）的 IPv6 静态路由，对于一个到来的数据报文，如果路由器在当前路由表里找不到匹配的路由表项，将会把报文发给在该条静态路由里指定的下一跳路由器。
- 第二种是通过动态路由协议生成（如 OSPF、IS-IS 和 RIP 等），由路由能力比较强的路由器将缺省路由发布给其它路由器，其它路由器在自己的路由表里生成指向那台路由器的缺省路由。

## 静态路由应用

### 1. 目的地址与掩码

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

在 **ipv6 route-static** 命令中，IPv6 地址为 X:X::X:X 形式，掩码采用掩码长度表示。

### 2. 出接口和下一跳地址

在配置静态路由时，可指定出接口。对于 NULL0 和 Loopback 接口，配置了出接口即可。

对于广播链路和 NBMA 链路，所有的路由项都必须明确下一跳地址。在发送报文时，首先根据报文的地址寻找路由表中与之匹配的路由。只有指定了下一跳地址，链路层才能找到对应的链路层地址，并转发报文。下一跳地址不能为本地接口 IP 地址，否则路由不会生效。

对于点对点链路，只需指定出接口即可。

### 3. 其它属性

对于不同的静态路由，可以为它们配置不同的优先级，从而更灵活地应用路由管理策略。例如：配置到达相同目的地的多条路由，如果指定相同优先级，则可实现负载分担，如果指定不同优先级，则可实现路由备份。

## 6.2 RIP

RIP 协议主要用于规模较小的网络中，比如校园网以及结构较简单的地区性网络。对于更为复杂的环境和大型网络，一般不使用 RIP。RIP 的实现较为简单，配置和维护也比 OSPF 和 IS-IS 容易，在实际组网中有广泛的应用。

### 6.2.1 介绍

#### 定义

RIP（Routing Information Protocol）是路由信息协议的简称。它是一种较为简单的内部网关协议 IGP（Interior Gateway Protocol），主要应用于规模较小的网络中，例如校园

网以及结构较简单的地区性网络。对于更为复杂的环境和大型网络，一般不使用 RIP 协议。

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

为防止产生路由循环：

- RIP 使用跳数（Hop Count）来衡量到达目的地址的距离，称为度量值。在 RIP 中，缺省情况下，路由器到与它直接相连网络的跳数为 0，通过一个路由器可达的网络的跳数为 1，其余依此类推。也就是说，度量值等于从本网络到达目的网络经过的路由器数量。为限制收敛时间，RIP 规定度量值取 0～15 之间的整数，大于或等于 16 的跳数被定义为无穷大，即目的网络或主机不可达。由于这个限制，使得 RIP 不可能在大型网络中得到应用。
- 支持水平分割（Split Horizon）和毒性反转（Poison Reverse）功能。

目的

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

6.2.2 参考标准和协议

与 RIP 相关的参考标准与协议如下：

- RFC 1058: Routing Information Protocol
- RFC 1724: RIP Version 2 MIB Extension
- RFC 2083: PNG (Portable Network Graphics) Specification Version 1.0
- RFC 1723: RIP Version 2 - Carrying Additional Information
- RFC 2453: RIP Version 2
- RFC 2082: RIP-2 MD5 Authentication

6.2.3 可获得性

License 支持

本特性无需 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

6.2.4 原理描述

6.2.4.1 RIP-1 报文格式

RIP 报文由头部（Header）和多个路由表项（Route Entries）部分组成。RIP 是一个基于 UDP 协议的，并且 RIP-1 的数据包不能超过 512 字节。

RIP-1 的报文格式如图 6-1 所示。

图 6-1 RIP-1 的报文格式

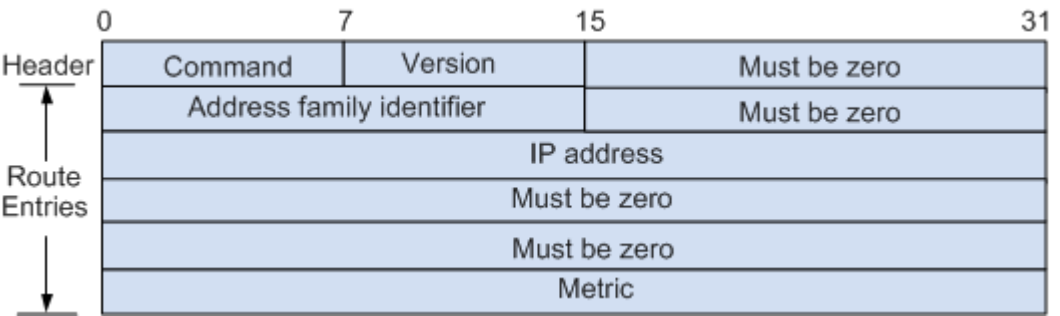


表 6-1 RIP-1 的报文格式字段解释

| 字段名                            | 长度       | 含义                                                                                                                                                                 |
|--------------------------------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Command                        | 8 比特     | 标识报文的类型： <ul style="list-style-type: none"><li>● 1: Request 报文，发送广播报文向邻居请求全部或部分路由信息。</li><li>● 2: Reponse 报文，发送自己全部或部分路由信息，一个 Response 报文中最多包含 25 个路由表项。</li></ul> |
| Version                        | 8 比特     | RIP 的版本号： <ul style="list-style-type: none"><li>● 1: RIP-1。</li><li>● 2: RIP-2。</li></ul>                                                                          |
| Must be zero                   | 16/32 比特 | 必须为零字段。                                                                                                                                                            |
| AFI（Address family identifier） | 16 比特    | 地址族标识，其值为 2 时表示 IP 协议。对于 Request 报文，此字段值为 0。                                                                                                                       |
| IP Address                     | 32 比特    | 该路由的目的 IP 地址，可以是自然网段的地址，也可以是子网地址或主机地址。                                                                                                                             |
| Metric                         | 32 比特    | 路由的开销值。对于 Request 报文，此字段值为 16。                                                                                                                                     |

6.2.4.2 RIP-2 报文格式

RIP-2 的报文格式与 RIP-1 类似，如图 6-2 所示。

图 6-2 RIP-2 的报文格式

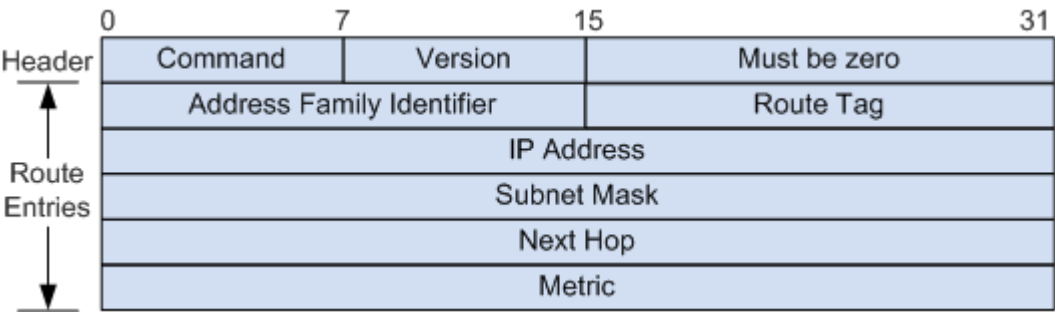


表 6-2 RIP-2 的报文格式字段解释

| 字段名                            | 长度    | 含义                                                                                                                                                                                                           |
|--------------------------------|-------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Command                        | 8 比特  | 标识报文的类型： <ul style="list-style-type: none"><li>● 1: Request 报文，发送组播报文（224.0.0.9）向邻居请求全部或部分路由信息。</li><li>● 2: Reponse 报文，发送自己全部或部分路由信息，一个 Response 报文中最多包含 25 个路由表项。如果 AFI 指明为认证信息，则最多只能有 24 个路由表项。</li></ul> |
| Version                        | 8 比特  | RIP 的版本号。                                                                                                                                                                                                    |
| Must be zero                   | 16 比特 | 必须为零字段。                                                                                                                                                                                                      |
| AFI（Address Family Identifier） | 16 比特 | 地址族标识，其值为 2 时表示 IP 协议。对于 Request 报文，此字段值为 0。如果 AFI 为 0xFFFF，后面紧跟着就是认证信息。                                                                                                                                     |
| Route Tag                      | 16 比特 | 外部路由标记。                                                                                                                                                                                                      |
| IP Address                     | 32 比特 | 该路由的目的 IP 地址，可以是自然网段的地址，也可以是子网地址或主机地址。                                                                                                                                                                       |
| Subnet Mask                    | 32 比特 | 目的地址的掩码。                                                                                                                                                                                                     |
| Next Hop                       | 32 比特 | 提供一个更好的下一跳地址。如果为 0.0.0.0，则表示发布此路由的路由器地址就是最优下一跳地址。                                                                                                                                                            |
| Metric                         | 32 比特 | 路由的开销值。对于 Request 报文，此字段为 16。                                                                                                                                                                                |

6.2.4.3 RIP-2 的验证报文格式

RIP-2 为了支持报文验证，使用第一个路由表项（Route Entry）作为验证项，并将 AFI 字段的值设为 0xFFFF 作为标识，如图 6-3 所示。

图 6-3 RIP-2 的验证报文格式

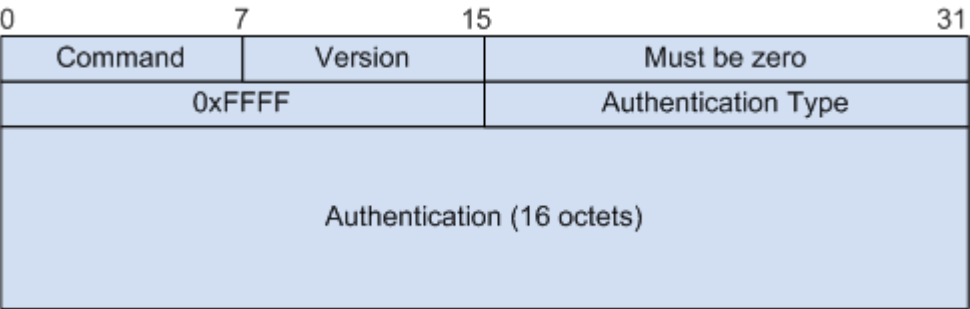


表 6-3 RIP-2 的验证报文格式字段解释

| 字段名                 | 长度    | 含义                                                                                                                                                                                          |
|---------------------|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Command             | 8 比特  | 标识报文的类型： <ul style="list-style-type: none"><li>● 1: Request 报文，向邻居请求全部或部分路由信息。</li><li>● 2: Reponse 报文，发送自己全部或部分路由信息，一个 Response 报文中最多包含 25 个路由表项。如果 AFI 指明为认证信息，则最多只能有 24 个路由表项。</li></ul> |
| Version             | 8 比特  | RIP 的版本号。                                                                                                                                                                                   |
| Must be zero        | 16 比特 | 必须为零字段。                                                                                                                                                                                     |
| 0xFFFF              | 16 比特 | 验证项标识，表示整个路由报文需要验证。                                                                                                                                                                         |
| Authentication Type | 16 比特 | 验证类型： <ul style="list-style-type: none"><li>● 2: 明文验证。</li><li>● 3: MD5 验证。</li></ul>                                                                                                       |
| Authentication      | 16 字节 | 验证口令，当使用明文验证时该字段才会包含密码信息。                                                                                                                                                                   |

6.2.4.4 RIP 的路由数据库

每个运行 RIP 的路由器管理一个路由数据库，该路由数据库包含了到网络所有可达信宿的路由项，这些路由项包含下列信息：

- 目的地址：主机或网络的地址。



- 下一跳地址：为到达目的地，RIP 报文需要经过的相邻路由器的接口 IP 地址。
- 接口：转发报文的接口。
- Metric 值：本路由器到达目的地的开销，是一个 0 ~ 15 之间的整数。
- 路由时间：从路由项最后一次被修改到现在所经过的时间，路由项每次被修改时，路由时间重置为 0。
- 路由标记：区分内部路由协议路由和外部路由协议路由的标记。

#### 6.2.4.5 RIP 定时器

RIP 受三个定时器的控制，分别是 Update、Age 和 Garbage-Collect。

- Update 为更新定时器，它定时触发更新报文的发送。
- Age 为超时定时器，RIP 路由器如果在该段时间内没有收到邻居发来的路由更新报文，则认为该路由不可达。
- Garbage-Collect 为垃圾收集时间，在这段时间内不可达路由没有收到来自同一邻居的更新，则该路由被从路由表中彻底删除。

三个定时器之间的关系：

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

表 6-4 定时器之间的关系

| 定时器名称           | 定时期时间                                                                                                                      |
|-----------------|----------------------------------------------------------------------------------------------------------------------------|
| Update          | 默认 30 秒钟发送更新报文。                                                                                                            |
| Age             | 在 Age 定时器时间内即使没有收到路由更新也按照 Update 发送更新报文。<br>在 Age 定时器超时且 Garbage-Collect 定时器没有超时，如果还没有收到更新报文，则按照 Update 定时发送度量值为 16 的更新报文。 |
| Garbage-Collect | 当 Garbage-Collect 定时器超时后，从路由表中删除该路由表项。                                                                                     |

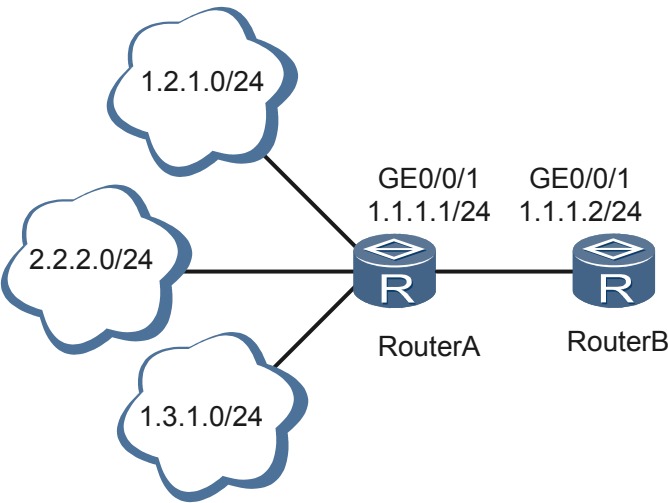
#### 6.2.4.6 RIP 运行和路由计算的过程

##### RIP 运行过程

1. 路由器启动 RIP 后，便会向相邻的路由器发送请求报文（Request message）。
  - 如果为 RIP-2，则请求报文发往组播地址 224.0.0.9。
  - 如果为 RIP-1，则请求报文发往广播地址。RIP 路由需要与接口上配置的掩码进行比较后才能发布，具体的 RIP 路由出策略是：
    - 当要发布的路由地址和接口的 IP 地址在同一自然网段内，并且掩码位相同，则按照此掩码位发送；若掩码位不同，则按照自然掩码发送。

- 如果该路由与发送接口的 IP 地址不在同一自然网段，则按照自然掩码聚合后发布。
2. 相邻的 RIP 路由器收到请求报文后，响应该请求，回送包含本地路由表信息的响应报文（Response message），并进行路由计算。
- 如果配置为 RIP-1，需要按照接口的入策略对路由进行过滤：
- 接收的路由同样需要与接口上配置的掩码进行比较。如果该路由与接收接口的 IP 地址属于同一自然网段，则按接口的掩码位接收该路由；如果该路由与接收接口的 IP 地址不属于同一自然网段，则按照自然网段接收。
  - 如果本机的路由表中存在相同网段的直连子网路由，则接收该路由；如果本机的路由表中不存在相同网段的直连子网路由，则拒绝该路由。
3. 路由器收到邻居路由器响应报文后，修改本地路由表。
- 如果配置为 RIP-1，在以下情况可以看到路由的掩码。如图 6-4 所示。

图 6-4 RIP-1 组网图



RouterA 和 RouterB 之间运行 RIP-1 协议。因为 RIP-1 报文中不携带掩码信息，所以只能根据出接口的地址来获得要发送的路由的掩码，但在 RouterB 上面查看路由表：

```
<RouterB> display rip 1 route
Route Flags: R - RIP, T - TRIP
              P - Permanent, A - Aging, S - Suppressed, G - Garbage-collect
-----
Peer 1.1.1.1 on GigabitEthernet0/0/1
  Destination/Mask    Nexthop    Cost    Tag    Flags    Sec
    2.0.0.0/8         1.1.1.1    1        0      RA       12
    1.2.1.0/24         1.1.1.1    1        0      RA       12
    1.3.1.0/24         1.1.1.1    1        0      RA       12
```

可以看到，在 RouterB 上的路由表中，1.2.1.0/24 网段和 1.3.1.0/24 网段不是按照 1.0.0.0/8 自然网段发布的，而会显示子网掩码。

这是因为在 RouterA 上 GigabitEthernet 0/0/1 接口的掩码为 24，要发送网段的掩码也为 24 且在同一个自然网段（1.0.0.0），所以会发布该路由的同时附带掩码。

如果 RouterA 和 RouterB 均配置为 RIP-2，再查看 RouterB 的路由表：

```
<RouterB> display rip 1 route
Route Flags: R - RIP, T - TRIP
              P - Permanent, A - Aging, S - Suppressed, G - Garbage-collect
```

```
Peer 1.1.1.1 on GigabitEthernet0/0/1
  Destination/Mask    Nexthop    Cost    Tag    Flags    Sec
  1.2.1.0/24          1.1.1.1    1        0      RA       26
  1.3.1.0/24          1.1.1.1    1        0      RA       26
  2.2.2.0/24          1.1.1.1    1        0      RA       26
```

路由表项均带有掩码，这是因为 RIP-2 报文中带有掩码信息，可以更精确的表示出路由。

## RIP 路由计算过程

路由器收到响应报文后，修改本地路由表，同时向相邻路由器发送触发更新的报文，广播路由修改信息。相邻路由器收到触发更新报文后，又向其各自的相邻路由器发送触发更新报文。在一连串的触发修改广播后，各路由器都能得到并保持最新的路由信息。

RIP 采用老化机制对超时的路由进行老化处理，以保证路由的有效性。因此，RIP 每隔一定时间周期性的向邻居路由器发布本地的路由表，相邻路由器在收到报文后，对其本地路由进行更新。所有 RIP 路由器都会重复此过程。

### 6.2.4.7 RIP 发送请求报文和响应报文

配置不同的版本，RIP 发送的 Request 报文和响应的 Response 报文的版本不同：

- 如果 RIP 的版本为 RIP-1，则只广播发送 RIP-1 报文，接收广播的 RIP-1 报文。
- 如果 RIP 的版本为 RIP-2，则默认使用组播发送和接收组播 RIP-2 报文。
- 如果 RIP 的版本为 RIP-2（广播模式），广播发送 RIP-2 报文，接收 RIP-1 和 RIP-2 的报文。

### 6.2.4.8 RIP 报文的认证

RIP 报文的认证过程：

- 如果路由器没有配置 RIP-2 报文认证：
  - RIP-1 报文和没有配置认证的 RIP-2 报文可以被接受。
  - 配置认证的 RIP-2 报文将被丢弃。
- 如果路由器配置了 RIP-2 报文认证：
  - 通过认证的 RIP-2 报文可以被接受。
  - 没有配置认证和认证错误的 RIP-2 报文将被丢弃。

 说明

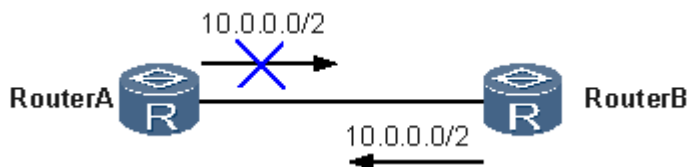
RFC2453 中只定义了明文验证方式，关于 MD5 验证的详细信息，请参考 RFC2082 “RIP-2 MD5 Authentication”。

### 6.2.4.9 水平分割

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

如图 6-5 所示，RouterB 目的地址是 10.0.0.0 的路由信息通告给 RouterA 后，RouterA 不会再把到网络 10.0.0.0 的路由发回给 RouterB。

图 6-5 水平分割原理图



#### 6.2.4.10 毒性逆转

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

RIP 毒性逆转可以防止产生路由环路。

图 6-6 毒性逆转原理图



如图 6-6 所示，在不配置水平分割的情况下，RouterB 会向 RouterA 发送从 RouterA 学到的路由，并且 RouterA 到网络 10.0.0.0 的路由开销值为 1。如果 RouterA 到网络 10.0.0.0 的路由变成不可达，同时 RouterB 没有收到 RouterA 的更新报文，而继续向 RouterA 发送到达网络 10.0.0.0 路由信息，则会导致路由环路。

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

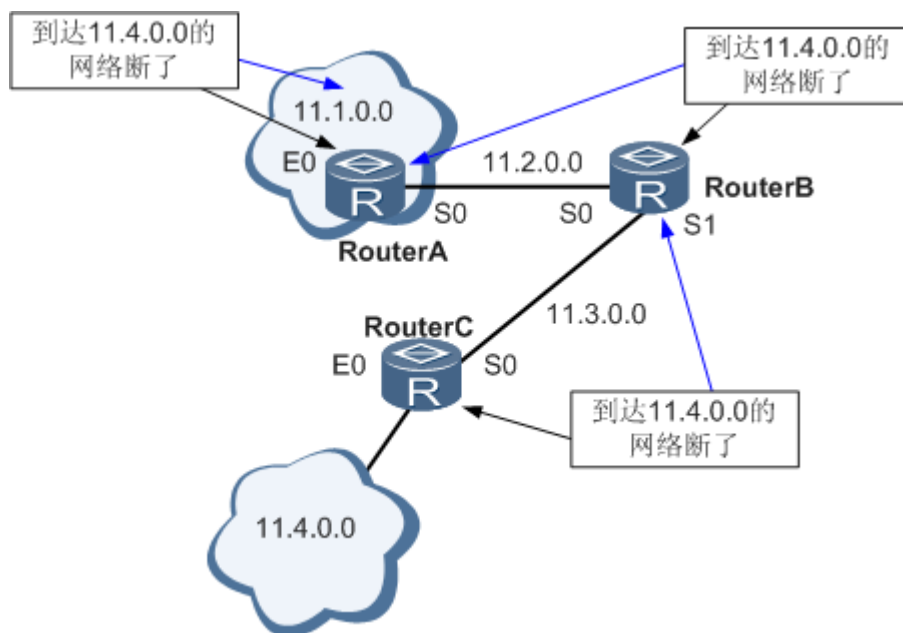
如果毒性逆转和水平分割都配置了，简单的水平分割行为（从某接口学到的路由再从从这个接口发布时将被抑制）会被毒性逆转行为代替。

#### 6.2.4.11 触发更新

触发更新（Triggered Updates）的原理是，路由信息发生变化时，立即向邻居路由器发送触发更新报文，通知变化的路由信息。

触发更新可以缩短网络收敛时间，在路由表项变化时立即向其他路由器发送该信息，而不必等待定时更新。

图 6-7 触发更新原理图



如图 6-7 所示，网络 11.4.0.0 不可达时，RouterC 最先得知这一信息。通常，更新路由信息会定时发送给相邻路由器（RIP 协议每隔 30 秒发送一次）。但如果在 RouterC 等待更新周期到来的时候，RouterB 的更新报文传到了 RouterC，RouterC 就会学到 RouterB 的去往网络 11.4.0.0 的错误路由。这样 RouterB 和 RouterC 上去往网络 11.4.0.0 的路由都指向对方从而形成路由环路。如果 RouterC 发现网络故障之后，不再等待更新周期到来，就立即发送路由更新信息给路由器 B，使路由器 B 的路由表及时更新，则可以避免产生上述问题。

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

#### 6.2.4.12 路由聚合

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

在 RIP-2 中进行路由聚合可提高大型网络的可扩展性和效率，缩减路由表。

路由聚合有两种方式：

- 基于 RIP 进程的有类聚合：

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

比如，对于 10.1.1.0/24 (metric=2) 和 10.2.2.0/24 (metric=3) 这两条路由，会聚合成自然网段路由 10.0.0.0/8 (metric=2)。RIP-2 聚合是按类聚和的，聚合得到最优的 metric 值。

- 基于接口的聚合：

用户可以指定聚合地址。

比如，对于 10.1.1.0/24 (metric=2) 和 10.2.2.0/24 (metric=3) 这两条路由，可以在此接口上配置聚合路由 10.1.0.0/16 (metric=2)。

### 6.2.4.13 多进程和多实例

为了方便管理，提高控制效率，RIP 支持多进程和多实例特性。

多进程允许为一个指定的 RIP 进程关联一组接口，从而保证该进程进行的所有协议操作都仅限于这一组接口。这样，就可以实现一台路由器有多个 RIP 协议进程，每个进程负责唯一的一组接口。而且每个 RIP 进程的路由数据也是相互独立的，但进程之间可以相互引入路由。

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

## 6.3 RIPng

RIPng 多用于规模较小和结构较简单的 IPv6 网络环境中，是基于距离矢量的路由协议，使用跳数 (Hop Count) 来衡量到达目的地址的距离。

### 6.3.1 介绍

#### 定义

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

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

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

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

#### 目的

RIPng 是为了支持 IPv6 而对 RIP 协议进行的扩展。

### 6.3.2 参考标准和协议

与 RIPng 相关的参考标准与协议如下：

- RFC 2080: RIPng for IPv6
- RFC 1058: Routing Information Protocol
- RFC 2453: RIP Version 2

### 6.3.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 6.3.4 原理描述

#### 6.3.4.1 RIPng 报文格式

##### 基本格式

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

RIPng 报文基本格式如[图 6-8](#)所示。

图 6-8 RIPng 报文格式

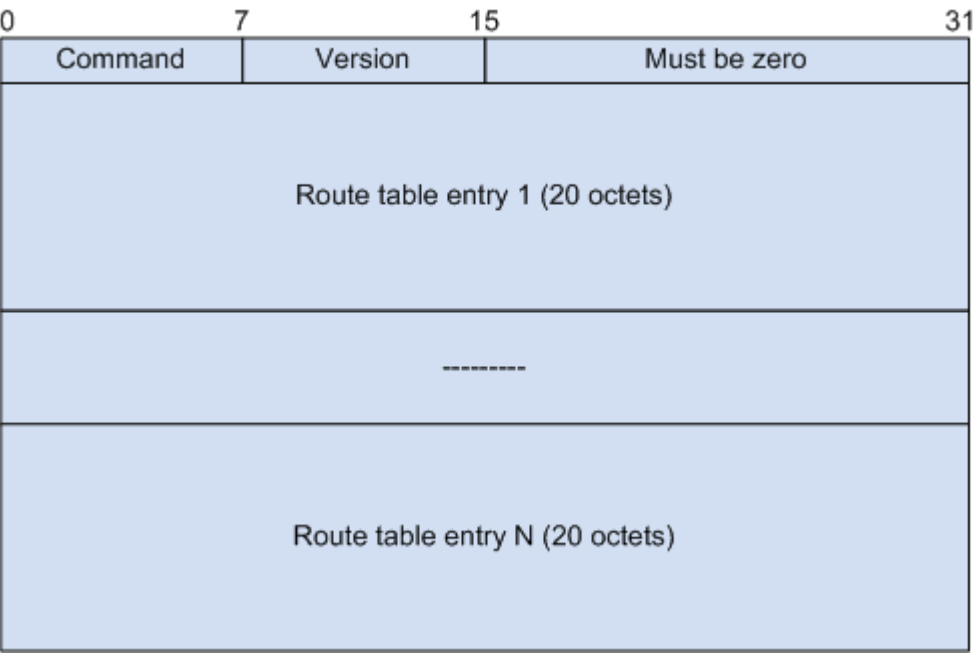


表 6-5 RIPng 报文格式字段解释

| 字段名                    | 长度    | 含义                                                                                                                                             |
|------------------------|-------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Command                | 8 比特  | 标识报文的类型： <ul style="list-style-type: none"><li>● 1: Request 报文，向邻居请求全部或部分路由信息。</li><li>● 2: Reponse 报文，发送自己全部或部分路由信息。路由条目与链路 MTU 有关。</li></ul> |
| Version                | 8 比特  | RIPng 的版本号：其值为 1。                                                                                                                              |
| Must be zero           | 16 比特 | 必须为零字段。                                                                                                                                        |
| RTE（Route table entry） | 20 字节 | 路由表项。                                                                                                                                          |

RTE 的格式

在 RIPng 里有两类 RTE，分别是：

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



下一跳 RTE 的格式如图 6-9 所示。

图 6-9 下一跳 RTE 格式

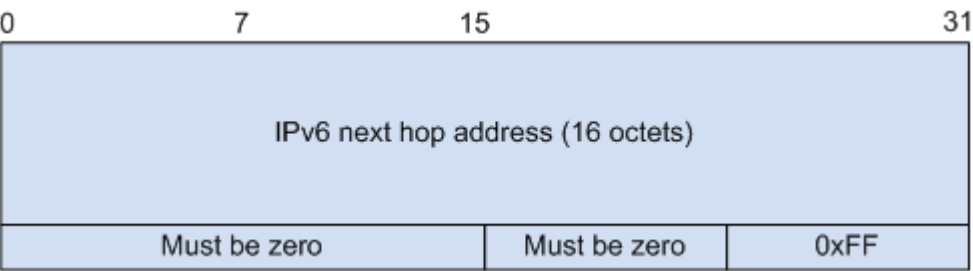


表 6-6 下一跳 RTE 格式字段解释

| 字段名                   | 长度      | 含义            |
|-----------------------|---------|---------------|
| IPv6 next hop address | 16 字节   | 下一跳的 IPv6 地址。 |
| Must be zero          | 16/8 比特 | 必须为零字段。       |

IPv6 前缀 RTE 的格式如图 6-10 所示。

图 6-10 IPv6 前缀 RTE 格式

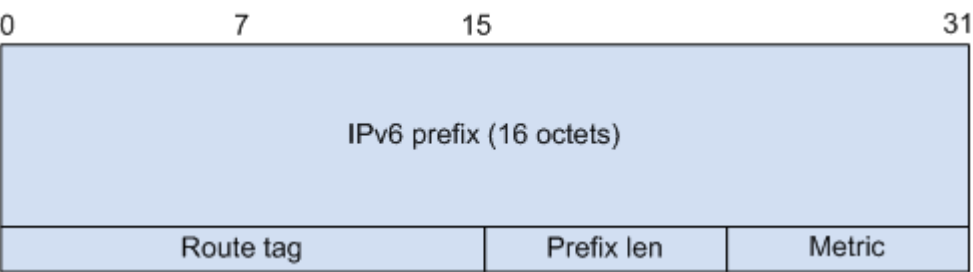


表 6-7 下一跳 RTE 格式字段解释

| 字段名         | 长度    | 含义             |
|-------------|-------|----------------|
| IPv6 prefix | 16 字节 | 目的 IPv6 地址的前缀。 |
| Route tag   | 16 比特 | 路由标记，用来区分外部路由。 |
| Prefix len  | 8 比特  | IPv6 地址的前缀长度。  |
| Metric      | 8 比特  | 路由的度量值（开销）。    |

### 6.3.4.2 定时器

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

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

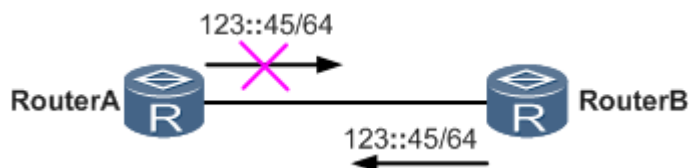
三个定时器之间的关系：

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

### 6.3.4.3 水平分割

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

图 6-11 水平分割原理图



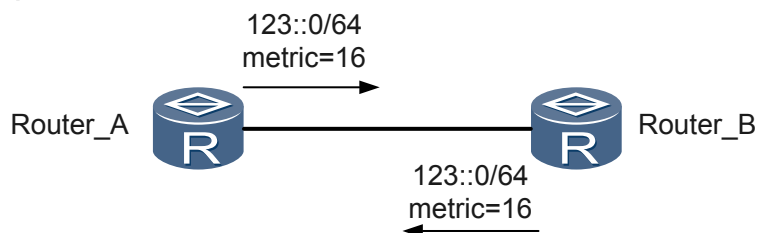
如图 6-11 所示路由器 B 发过来的到网络 123::45 的路由到路由器 A 上后，路由器 A 不会再把到网络 123::45 的路由发回给路由器 B。

### 6.3.4.4 毒性逆转

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

RIPng 毒性逆转也是为了防止产生路由环路。

图 6-12 毒性逆转原理图



如图 6-12 所示，在不配置毒性逆转的情况下，Router\_B 会向 Router\_A 发送从 Router\_A 学到的路由。Router\_A 到网络 123::0/64 的路由开销为 1，如果 Router\_A 到网络 123::0/64 的路由变成不可达，同时 Router\_B 没有收到 Router\_A 的更新报文，而继续向 Router\_A 发送到网络 123::0/64 的路由信息，则会导致路由环路。

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

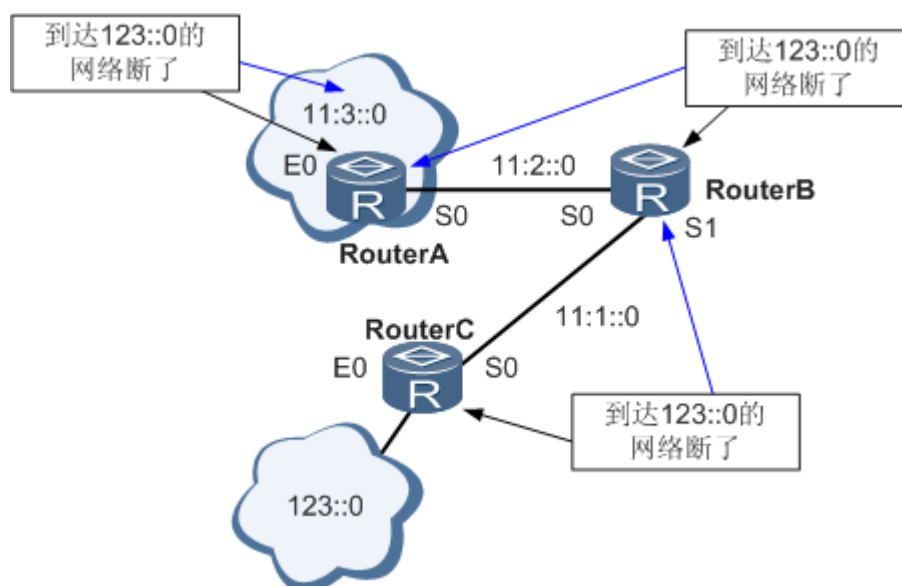
如果同时配置了毒性逆转和水平分割，简单的水平分割行为（从某接口学到的路由再从该接口发布时将被抑制）会被毒性逆转行为代替。

### 6.3.4.5 触发更新

触发更新（Triggered Updates）是指路由信息发生变化时，立即向邻居路由器发送触发更新报文，通知变化的路由信息。

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

图 6-13 触发更新原理图



如图 6-13 所示，网络 123::0 不可达时，路由器 C 最先得得到这一信息。通常，更新路由信息会定时发送给相邻路由器。例如，RIPng 协议每隔 30 秒发送一次。但如果在路由器 C 等待更新周期到来的时候，路由器 B 的更新报文传到了路由器 C，路由器 C 就会学到路由器 B 的去往网络 123::0 的错误路由。这样路由器 B 和 C 上去往网络 123::0 的路由都指向对方从而形成路由环路。如果路由器 C 发现网络故障之后，不再等待更新周期到来，就立即发送路由更新信息给路由器 B，使路由器 B 的路由表及时更新，则可以避免产生上述问题。

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

### 6.3.4.6 路由聚合

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

RIPng 路由聚合可提高大型网络的可扩展性和效率，缩减路由表。

路由聚合的实现原理：

例如，RIPng 可以从这个接口发布出去的路由有两条。

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

## 6.4 OSPF

OSPF 多用于支持大型 IPv4 网络，是基于链路状态的路由协议，与基于距离矢量的路由协议相比具有更快的收敛速度。

### 6.4.1 介绍

#### 定义

OSPF（Open Shortest Path First）是 IETF 组织开发的一个基于链路状态的内部网关协议。目前针对 IPv4 协议使用的是 OSPF Version 2（RFC2328）。OSPF 是不属于任何一个厂商或组织私有的路由协议，使用 Dijkstra 的最短路径（SPF）算法计算路由。在 OSPF 中引入了区域的概念，有效的减少了路由选择协议对路由器 CPU 和内存的占用，同时也降低了路由选择协议的通信量，使得构建大型层次化网络成为可能。

#### 目的

随着网络中路由器数目的增加，Rip 路由协议由于自身的一些问题，如路由自环、最大跳数限制等，难以为大型网络提供良好的支持。IETF 组织（Internet Engineering Task Force）开发了 OSPF 路由协议，对大型网络的发展起到了重要的作用。

OSPF 路由协议具有如下特性：

- 适应范围广——支持各种规模的网络，最多可支持几百台路由器。
- 快速收敛——在网络的拓扑结构发生变化后立即发送更新报文，使这一变化在自治系统中同步。
- 无自环——由于 OSPF 根据收集到的链路状态用最短路径树算法计算路由，从算法本身保证了不会生成自环路由。
- 区域划分——允许自治系统的网络被划分成区域来管理，路由器的链路状态数据库仅需和所在区域的其他路由器保持一致。链路状态数据库的减小大大降低了对路由器内存的占用和 CPU 的消耗。同时，需要在区域间传送的路由信息的减小，降低了网络带宽的占用。
- 路由分级——使用 4 类不同的路由，按优先顺序来说分别是：区域内路由、区域间路由、第一类外部路由、第二类外部路由。
- 等价路由——支持到同一目的地址的多条等价路由。
- 支持验证——支持基于区域和接口的报文验证，以保证报文交互的安全性。

- 组播发送——在某些类型的链路上以组播地址发送协议报文，减少对其他设备的干扰。

### 6.4.2 参考标准和协议

与 OSPF 相关的参考标准与协议如下：

- RFC 2328: OSPF Version 2
- RFC 1850: OSPF Version 2 Management Information Base
- RFC 1587: The OSPF NSSA Option
- RFC 2370: The OSPF Opaque LSA Option
- RFC 3137: OSPF Stub Router Advertisement
- RFC 1765: OSPF Database Overflow

### 6.4.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 6.4.4 原理描述

#### 6.4.4.1 OSPF 基础概念

OSPF 协议具有以下特点：

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

## OSPF 报文类型

表 6-8 OSPF 报文类型

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

## LSA 类型

表 6-9 OSPF LSA 类型

| LSA 类型                      | LSA 作用                                                |
|-----------------------------|-------------------------------------------------------|
| Router-LSA (Type1)          | 每个路由器都会产生，描述了路由器的链路状态和开销，在所属的区域内传播。                   |
| Network-LSA (Type2)         | 由 DR 产生，描述本网段的链路状态，在所属的区域内传播。                         |
| Network-summary-LSA (Type3) | 由 ABR 产生，描述区域内所有的路由，并通告给其他相关区域。                       |
| ASBR-summary-LSA (Type4)    | 由 ABR 产生，描述到 ASBR 的路由，通告给除 ASBR 所在区域的其他相关区域。          |
| AS-external-LSA (Type5)     | 由 ASBR 产生，描述到 AS 外部的路由，通告到所有的区域（除了 Stub 区域和 NSSA 区域）。 |
| NSSA LSA (Type7)            | 由 ASBR 产生，描述到 AS 外部的路由，仅在 NSSA 区域内传播。                 |

| LSA 类型                          | LSA 作用                                                                                                                                                                                                          |
|---------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Opaque LSA（Type9/Type10/Type11） | <p>Opaque LSA 提供用于 OSPF 的扩展的通用机制。其中：</p> <p>Type9 LSA 仅在接口所在网段范围内传播。用于支持 GR 的 Grace LSA 就是 Type9 LSA 的一种。</p> <p>Type10 LSA 在区域内传播。用于支持 TE 的 LSA 就是 Type10 LSA 的一种。</p> <p>Type11 LSA 在自治域内传播，目前还没有实际应用的例子。</p> |

路由器类型

OSPF 协议中常用到的路由器类型如图 6-14 所示。

图 6-14 路由器类型

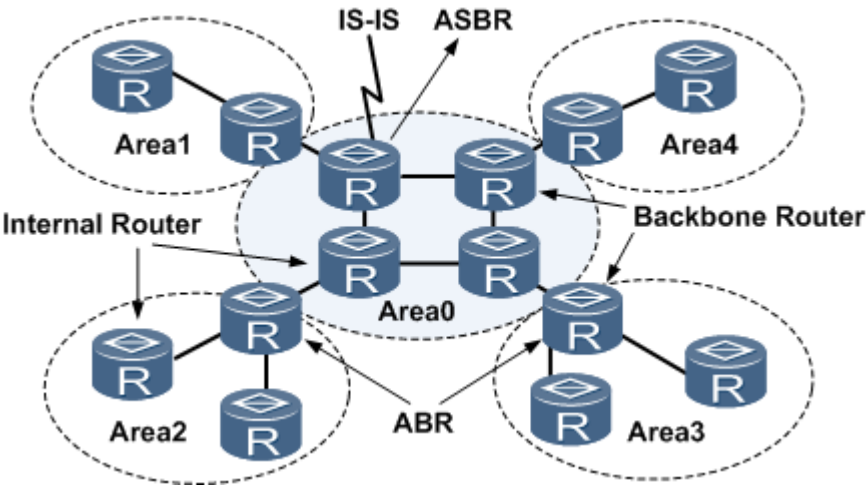


表 6-10 OSPF 路由器类型

| 路由器类型                           | 含义                                                                                             |
|---------------------------------|------------------------------------------------------------------------------------------------|
| 区域内路由器（Internal Router）         | 该类路由器的所有接口都属于同一个 OSPF 区域。                                                                      |
| 区域边界路由器 ABR（Area Border Router） | <p>该类路由器可以同时属于两个以上的区域，但其中一个必须是骨干区域。</p> <p>ABR 用来连接骨干区域和非骨干区域，它与骨干区域之间既可以是物理连接，也可以是逻辑上的连接。</p> |
| 骨干路由器（Backbone Router）          | <p>该类路由器至少有一个接口属于骨干区域。</p> <p>所有的 ABR 和位于 Area0 的内部路由器都是骨干路由器。</p>                             |

| 路由器类型                                  | 含义                                                                                                     |
|----------------------------------------|--------------------------------------------------------------------------------------------------------|
| 自治系统边界路由器 ASBR<br>(AS Boundary Router) | 与其他 AS 交换路由信息的路由器称为 ASBR。<br>ASBR 并不一定位于 AS 的边界，它可能是区域内路由器，也可能是 ABR。只要一台 OSPF 路由器引入了外部路由的信息，它就成为 ASBR。 |

OSPF 路由类型

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

表 6-11 中按优先级从高到低顺序列出了路由类型。

表 6-11 OSPF 路由类型

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

区域类型

表 6-12 OSPF 区域类型

| 区域类型              | 作用                                        |
|-------------------|-------------------------------------------|
| Totally Stub Area | 允许 ABR 发布的 Type3 缺省路由，不允许自治系统外部路由和区域间的路由。 |
| Stub Area         | 和 Totally Stub 区域的不同在于该区域允许区域间路由。         |



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

OSPF 支持的网络类型

OSPF 根据链路层协议类型，将网络分为如表 6-13 所列四种类型。

表 6-13 OSPF 网络类型

| 网络类型                                   | 含义                                                                                                                                                                                                                                                                                 |
|----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 广播类型（Broadcast）                        | <p>当链路层协议是 Ethernet、FDDI 时，缺省情况下，OSPF 认为网络类型是 Broadcast。</p> <p>在该类型的网络中：</p> <ul style="list-style-type: none"><li>● 通常以组播形式发送 Hello 报文、LSU 报文和 LSAck 报文。其中，224.0.0.5 的组播地址为 OSPF 路由器的预留 IP 组播地址；224.0.0.6 的组播地址为 OSPF DR 的预留 IP 组播地址。</li><li>● 以单播形式发送 DD 报文和 LSR 报文。</li></ul> |
| NBMA 类型（Non-broadcast multiple access） | <p>当链路层协议是帧中继、ATM 或 X.25 时，缺省情况下，OSPF 认为网络类型是 NBMA。</p> <p>在该类型的网络中，以单播形式发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。</p>                                                                                                                                                        |
| 点到多点 P2M 类型（Point-to-Multipoint）       | <p>没有一种链路层协议会被缺省的认为是 Point-to-Multipoint 类型。点到多点必须是由其他的网络类型强制更改的。常用做法是将非全连通的 NBMA 改为点到多点的网络。</p> <p>在该类型的网络中：</p> <ul style="list-style-type: none"><li>● 以组播形式（224.0.0.5）发送 Hello 报文。</li><li>● 以单播形式发送其他协议报文（DD 报文、LSR 报文、LSU 报文、LSAck 报文）。</li></ul>                            |
| 点到点 P2P 类型（point-to-point）             | <p>当链路层协议是 PPP、HDLC 和 LAPB 时，缺省情况下，OSPF 认为网络类型是 P2P。</p> <p>在该类型的网络中，以组播形式（224.0.0.5）发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。</p>                                                                                                                                            |

OSPF 路由过滤

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

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

OSPF 路由过滤可以应用于以下几个方面：

- 路由引入  
OSPF 可以引入其它路由协议学习到的路由。在引入时可以通过配置路由策略来过滤路由，只引入满足条件的路由。
- 引入路由发布  
OSPF 引入了路由后会向其它邻居发布引入的路由信息。  
可以通过配置过滤规则来过滤向邻居发布的路由信息。该过滤规则只在 ASBR 上配置才有效（只有 ASBR 才能引入路由）。
- 路由学习  
通过配置过滤规则，可以设置 OSPF 对接收到的区域内、区域间和自制系统外部的路由进行过滤。  
该过滤只作用于路由表项的添加与否，即只有通过过滤的路由才被添加到本地路由表中，但所有的路由仍可以在 OSPF 路由表中被发布出去。
- 区域间 LSA 学习  
通过命令可以在 ABR 上配置对进入本区域的 Summary LSA 进行过滤。该配置只在 ABR 上有效（只有 ABR 才能发布 Summary LSA）。

表 6-14 区域间 LSA 学习与路由学习的差异

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

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

OSPF 路由引入限制

路由引入限制，是指 OSPF 在引入其它路由协议路由时，对引入路由的数量进行限制。  
在实际应用中，由于操作不当等原因，可能将大量外部路由引入到 OSPF 中，导致网络拥塞。而且，网络中的低性能节点在引入大量路由的情况下，存在无法进行有效处理的风险。通过限制引入路由数量，可以消除潜在的风险，提高网络的稳定性。

由于外部路由被 OSPF 引入后会以 5 类或者 7 类 LSA 的形式发布，所以路由引入限制是通过限制 LSDB 中 5 类和 7 类 LSA 的数量实现的。其实现方式如下：

- 只对路由器本身产生的 5 类和 7 类 LSA 的总数量进行限制。
- 不限制缺省路由产生的 LSA。这是因为，缺省路由可以匹配所有的目的地，这样做可以保留尽可能多的外部路由。
- 如果配置了外部路由聚合命令，则按聚合后生成的 LSA 进行限制。例如，10 条外部路由由聚合生成一条 LSA 时，数量按一条计算。

缺省情况下，不对路由引入进行限制。

随着引入路由和配置路由引入限制的具体操作不同，OSPF 的处理流程也不同，如表 6-15 所示。

表 6-15 OSPF 路由引入限制

| 场景                                   | OSPF 处理流程                                                                                                                                                                          |
|--------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 先配置路由引入限制，再引入外部路由                    | <ul style="list-style-type: none"><li>● 需要引入的路由数量小于配置的路由引入限制数量，则按正常引入流程处理。</li><li>● 需要引入的路由数量大于配置的路由引入限制数量，则只生成最大限制数量路由的 LSA。此时，LSDB 里的 5 类和 7 类 LSA 的总数量等于配置的路由引入限制数量。</li></ul> |
| 已经引入了外部路由，再配置路由引入限制                  | <ul style="list-style-type: none"><li>● 引入的路由数量小于配置的路由引入限制数量，则不做任何处理。</li><li>● 引入的路由数量大于配置的路由引入限制数量，则先将已经引入的 5 类或 7 类 LSA 清除，再重新生成最大限制数量路由的 LSA。</li></ul>                        |
| 当需要再引入的路由数量已经大于配置的限制数量时              | 如果再引入新的路由，则不做引入处理。<br>如果需要删除之前的路由，则有以下两种情况： <ul style="list-style-type: none"><li>● 之前引入过这些待删除的路由，则重新引入。</li><li>● 之前并没有引入过这些待删除的路由，则忽略此路由。</li></ul>                              |
| 需要引入的路由数量大于已经配置的限制数量，引入路由后，再重新配置引入限制 | 先将已经引入的 Type5 或 Type7 LSA 清除，再重新引入，然后根据引入的路由数量和配置的路由引入限制数量的大小关系进一步处理。                                                                                                              |

OSPF RFC1583 兼容

RFC1583 是 OSPFv2 协议比较早的版本。

OSPF 在计算外部路由时，由于 RFC2328 和 RFC1583 的路由计算规则不一致，可能会导致路由环路。为了避免路由环路的发生，RFC2328 中提出了 RFC1583 兼容特性。

- 启用 RFC1583 兼容后，OSPF 采用 RFC1583 的路由计算规则。
- 不启用 RFC1583 兼容时，OSPF 采用 RFC2328 的路由计算规则。

OSPF 是根据 5 类 LSA 来计算外部路由的。RFC1583 兼容特性主要用于路由器收到 5 类 LSA 后：

- 选择到达产生该 LSA 的 ASBR 或该 LSA 所描述的转发地址（Forwarding Address）的路径。
- 选择到达相同目的地的外部路径。

缺省情况下，OSPF 兼容 RFC1583。

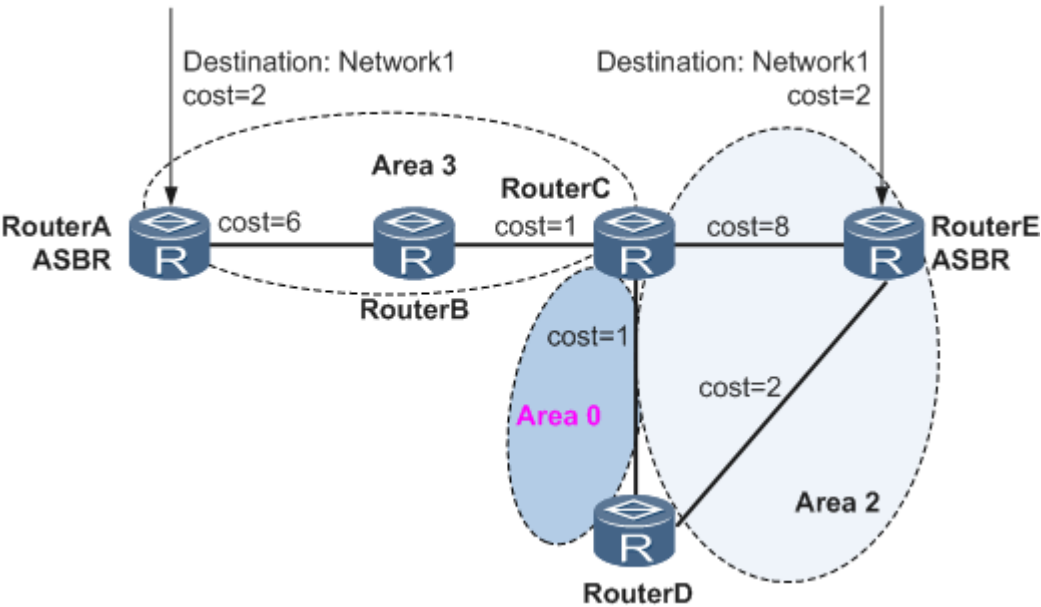
表 6-16 中分别描述了启用和不启用 RFC1583 兼容特性时，外部路由计算的情况。

表 6-16 启用和不启用 RFC1583 兼容特性的区别

| 启用 RFC1583 兼容的计算规则                                                                                       | 不启用 RFC1583 兼容的计算规则                                                                                                                                                     |
|----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>● 优选 cost 值最小的路径。</li><li>● 对于 cost 值相同的路径，选择关联区域号较大的路径。</li></ul> | <ul style="list-style-type: none"><li>● 优选非骨干区域的区域内路径。</li><li>● 如果没有非骨干区域的区域内路径，则在骨干区域中的区域内路径和所有区域间路径中，优选 cost 值最小的路径。</li><li>● 对于 cost 值相同的路径，选择关联区域号大的路径。</li></ul> |

如图 6-15 所示，RouterA 和 RouterE 为 ASBR，同时引入到达外部网络 Network1 的路由，并发布 Network1 的 5 类 LSA，cost 值为 2。RouterB 使用 RFC1583 中规定的选路规则，RouterC 使用 RFC2328 中规定的选路规则。

图 6-15 OSPF RFC1583 兼容特性



根据表 6-16 中所描述的路由计算规则，在计算到达外部网络 Network1 的路由时，RB 会选择 cost 值最小的路径 RouterB->RouterC->RouterD->RouterE->Network1，而 RouterC 则优先选择非骨干区域 cost 最小的区域内路径 RouterC->RouterB->RouterA->Network1。这样，RouterB 上去往 Network1 的下一跳为 RouterC，而同样 RouterC 上去往 Network1 的下一跳为 RouterB，导致路由环路。

在 RouterC 上启用 RFC1583 兼容后，RouterC 选择到达 Network1 的路由时，会选择 cost 值最小的路径 RouterC->RouterD->RouterE->Network1，这样，RouterC 上去往 Network1 的下一跳不再是 RouterB，从而避免了环路的产生。

## OSPF 洪泛策略控制

OSPF 洪泛策略控制指的是，通过对特定的接口或邻居设置过滤规则，来过滤特定的 LSA。

缺省情况下，OSPF 会在所有符合条件的接口下进行 LSA 洪泛。

OSPF 洪泛策略控制具有以下优点：

- 当网络振荡时，可以减少 LSA 的洪泛，减少网络负担。
- 通过不向邻居发送无用的 LSA，可以减少邻居 LSDB 的大小，提高网络收敛速度。
- 当两台路由器之间存在多条链路时，可以在某些链路上过滤 LSA 的传送，减少不必要的重传，节省带宽资源。
- 可以用较少的资源实现路由器的特殊路由需求。

OSPF 洪泛策略控制的主要应用有如下两种：

- 在广播、NBMA 以及 P2P 网络中，需要对特定的接口配置洪泛策略来过滤 LSA。
- 在 P2MP 网络中，需要对特定的邻居配置洪泛策略来过滤 LSA。

## OSPF 路由计算过程

OSPF 协议的路由计算过程可简单描述如下：

- 每台支持 OSPF 协议的路由器都维护着一份描述整个自治系统拓扑结构的链路状态数据库 LSDB（Link State Database）。每台路由器根据自己周围的网络拓扑结构生成链路状态广播 LSA（Link State Advertisement），通过相互之间发送协议报文将 LSA 发送给网络中其它路由器。这样每台路由器都收到了其它路由器的 LSA，所有的 LSA 一起组成链路状态数据库。
- 由于 LSA 是对路由器周围网络拓扑结构的描述，那么 LSDB 则是对整个网络的拓扑结构的描述。路由器很容易将 LSDB 转换成一张带权的有向图，这张图便是对整个网络拓扑结构的真实反映。显然，各台路由器得到的是一张完全相同的图。
- 每台路由器都使用 SPF 算法计算出一棵以自己为根的最短路径树，这棵树给出了到自治系统中各节点的路由，外部路由信息为叶子节点，外部路由可由广播它的路由器进行标记以记录关于自治系统的额外信息。显然，各台路由器各自得到的路由表是不同的。

### 6.4.4.2 OSPF 报文格式

#### 报文头格式

OSPF 的五种报文具有相同的报文头格式，长度为 24 字节。如图 6-16 所示。

图 6-16 OSPF 报文头格式

|                |      |               |    |
|----------------|------|---------------|----|
| 0              | 7    | 15            | 31 |
| Version        | Type | Packet length |    |
| Router ID      |      |               |    |
| Area ID        |      |               |    |
| Checksum       |      | AuType        |    |
| Authentication |      |               |    |

表 6-17 OSPF 报文头格式字段解释

| 字段名            | 长度    | 含义                                                                                                                                                              |
|----------------|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Version        | 8 比特  | OSPF 的版本号。对于 OSPFv2 来说，其值为 2。                                                                                                                                   |
| Type           | 8 比特  | OSPF 报文的类型。 <ul style="list-style-type: none"><li>● 1: Hello 报文。</li><li>● 2: DD 报文。</li><li>● 3: LSR 报文。</li><li>● 4: LSU 报文。</li><li>● 5: LSAck 报文。</li></ul> |
| Packet length  | 16 比特 | OSPF 报文的总长度，包括报文头在内，单位为字节。                                                                                                                                      |
| Router ID      | 32 比特 | 发送该报文的路由器标识。                                                                                                                                                    |
| Area ID        | 32 比特 | 发送该报文的路由器所属区域。                                                                                                                                                  |
| Checksum       | 16 比特 | 包含除了认证字段的整个报文的校验和。                                                                                                                                              |
| AuType         | 16 比特 | 验证类型： <ul style="list-style-type: none"><li>● 0: 不验证。</li><li>● 1: 简单认证。</li><li>● 2: MD5 认证。</li></ul>                                                         |
| Authentication | 64 比特 | 其数值根据验证类型而定。 <ul style="list-style-type: none"><li>● 0: 不对此字段作定义。</li><li>● 1: 此字段为密码信息。</li><li>● 2: 此字段包括 Key ID、MD5 验证数据长度和序列号的信息。</li></ul>                 |

 说明

MD5 验证数据添加在 OSPF 报文后面，不包含在 Authenticaiton 字段中。

Hello 报文

最常用的一种报文，其作用为建立和维护邻接关系，周期性的在启用了 OSPF 的接口上发送。报文内容包括一些定时器的数值、DR、BDR 以及自己已知的邻居。Hello 报文格式如图 6-17 所示。

图 6-17 Hello 报文格式

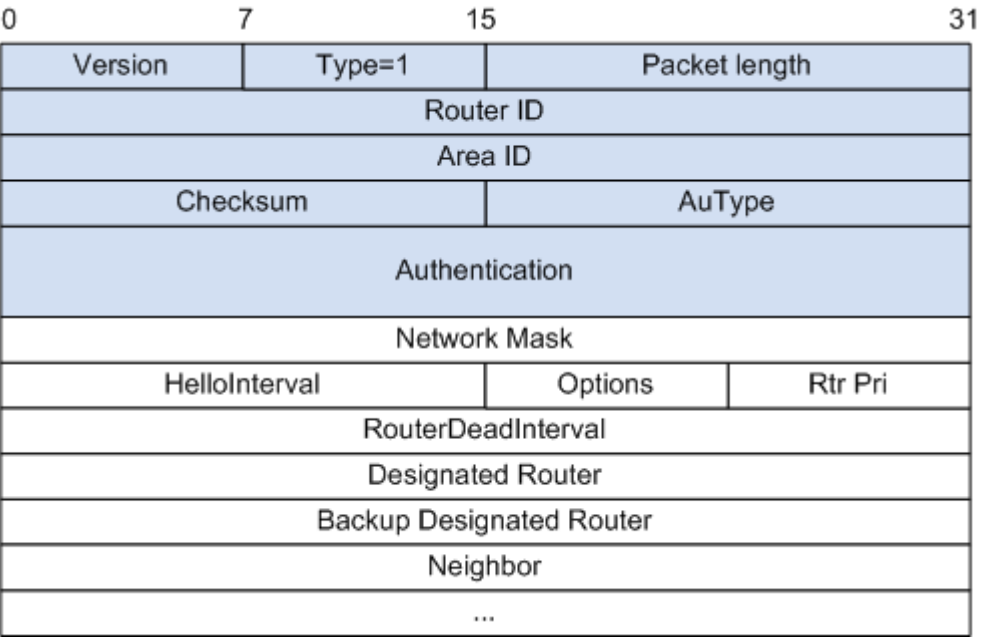


表 6-18 Hello 报文格式字段解释

| 字段名                      | 长度    | 含义                                                                                                                                                          |
|--------------------------|-------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Network Mask             | 32 比特 | 发送 Hello 报文的接口所在网络的掩码。                                                                                                                                      |
| HelloInterval            | 16 比特 | 发送 Hello 报文的时间间隔。                                                                                                                                           |
| Options                  | 8 比特  | 可选项： <ul style="list-style-type: none"><li>● E：允许 Flood AS-External-LSAs</li><li>● MC：转发 IP 组播报文</li><li>● N/P：处理 Type-7 LSAs</li><li>● DC：处理按需链路</li></ul> |
| Rtr Pri                  | 8 比特  | DR 优先级。默认为 1。<br><b>说明</b><br>如果设置为 0，则路由器不能参与 DR 或 BDR 的选举。                                                                                                |
| RouterDeadInterval       | 32 比特 | 失效时间。如果在此时间内未收到邻居发来的 Hello 报文，则认为邻居失效。                                                                                                                      |
| Designated Router        | 32 比特 | DR 的接口地址。                                                                                                                                                   |
| Backup Designated Router | 32 比特 | BDR 的接口地址。                                                                                                                                                  |
| Neighbor                 | 32 比特 | 邻居，以 Router ID 标识。                                                                                                                                          |



在同一网段上的路由器，其 HelloInterval 和 RouterDeadInterval 必须分别一致，否则不能形成邻居关系。

Hello 报文在不同链路上的发送：

- 在广播链路上，Hello 报文以 HelloInterval 周期发送往组播地址。
- 在虚连接链路上，Hello 报文以单播形式每 HelloInterval 发送到对端。
- 在 P2P 链路上，Hello 报文每 HelloInterval 发送往组播地址。
- 在 P2MP 链路上，Hello 报文以单播形式每 HelloInterval 发送到每个单独的邻居路由器。
- 在 NBMA 链路上，Hello 报文以单播发送。其他具体规则如下：
  - DR、BDR、有能力成为 DR 的路由器会每隔 HelloInterval 发送 Hello 报文给每个邻居。
  - Hello 报文发送的间隔取决于邻居的状态：只有当邻居的状态为 Down 时，发送间隔为 Pollinterval。其它情况为 HelloInterval。

说明

缺省情况下，P2P、广播链路类型接口发送 Hello 报文的时间间隔的值为 10 秒；P2MP、NBMA 类型接口发送 Hello 报文的时间间隔的值为 30 秒。NBMA 网络上发送 Pollinterval 缺省时间间隔为 120 秒。

DD 报文

两台路由器在邻接关系初始化时，用 DD 报文（Database Description Packet）来描述自己的 LSDB，进行数据库的同步。报文内容包括 LSDB 中每一条 LSA 的 Header（LSA 的 Header 可以唯一标识一条 LSA）。LSA Header 只占一条 LSA 的整个数据量的一小部分，这样可以减少路由器之间的协议报文流量，对端路由器根据 LSA Header 就可以判断出是否已有这条 LSA。在两台路由器交换 DD 报文的过程中，一台为 Master，另一台为 Slave。由 Master 规定起始序列号，每发送一个 DD 报文序列号加 1，Slave 方使用 Master 的序列号作为确认。

DD 报文格式如图 6-18 所示。

图 6-18 DD 报文格式

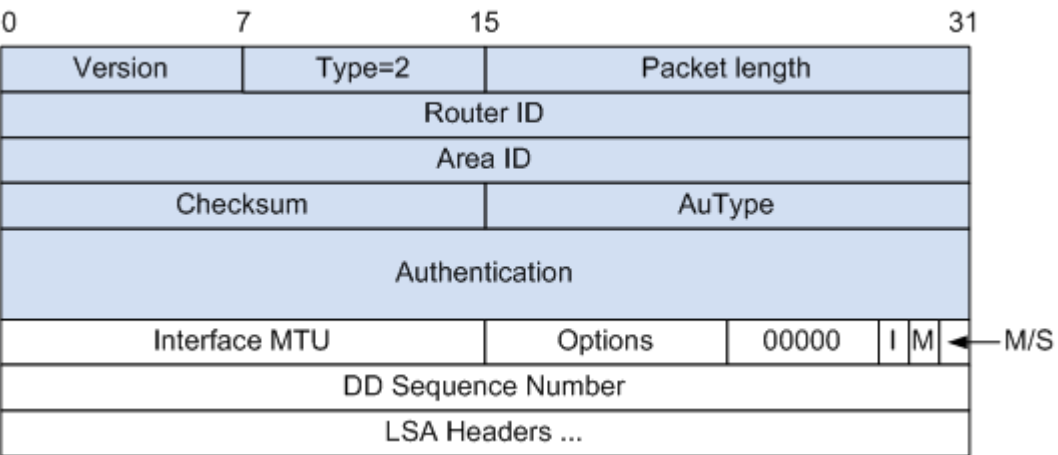




表 6-19 DD 报文格式字段解释

| 字段名                | 长度    | 含义                                                                                                                                                                   |
|--------------------|-------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Interface MTU      | 16 比特 | 在不分片的情况下，此接口最大可发出的 IP 报文长度。                                                                                                                                          |
| Options            | 8 比特  | 可选项： <ul style="list-style-type: none"> <li>● E：允许 Flood AS-External-LSAs。</li> <li>● MC：转发 IP 组播报文。</li> <li>● N/P：处理 Type-7 LSAs。</li> <li>● DC：处理按需链路。</li> </ul> |
| I                  | 1 比特  | 当发送连续多个 DD 报文时，如果这是第一个 DD 报文，则置为 1，否则置为 0。                                                                                                                           |
| M（More）            | 1 比特  | 当发送连续多个 DD 报文时，如果这是最后一个 DD 报文，则置为 0。否则置为 1，表示后面还有其他的 DD 报文。                                                                                                          |
| M/S（Master/Slave）  | 1 比特  | 当两台 OSPF 路由器交换 DD 报文时，首先需要确定双方的主从关系，Router ID 大的一方会成为 Master。当值为 1 时表示发送方为 Master。                                                                                   |
| DD sequence number | 32 比特 | DD 报文序列号。主从双方利用序列号来保证 DD 报文传输的可靠性和完整性。                                                                                                                               |
| LSA Headers        | -     | 该 DD 报文中所包含的 LSA 的头部信息。                                                                                                                                              |

## LSR 报文

两台路由器互相交换过 DD 报文之后，知道对端的路由器有哪些 LSA 是本地的 LSDB 所缺少的和哪些 LSA 是已经失效的，这时需要发送 LSR 报文（Link State Request Packet）向对方请求所需的 LSA。内容包括所需要的 LSA 的摘要。LSR 报文格式如图 6-19 所示。LS type、Link State ID 和 Advertising Router 可以唯一标识出一个 LSA，当两个 LSA 一样时，需要根据 LSA 中的 LS sequence number、LS checksum 和 LS age 来判断出所需要 LSA 的新旧。

图 6-19 LSR 报文格式

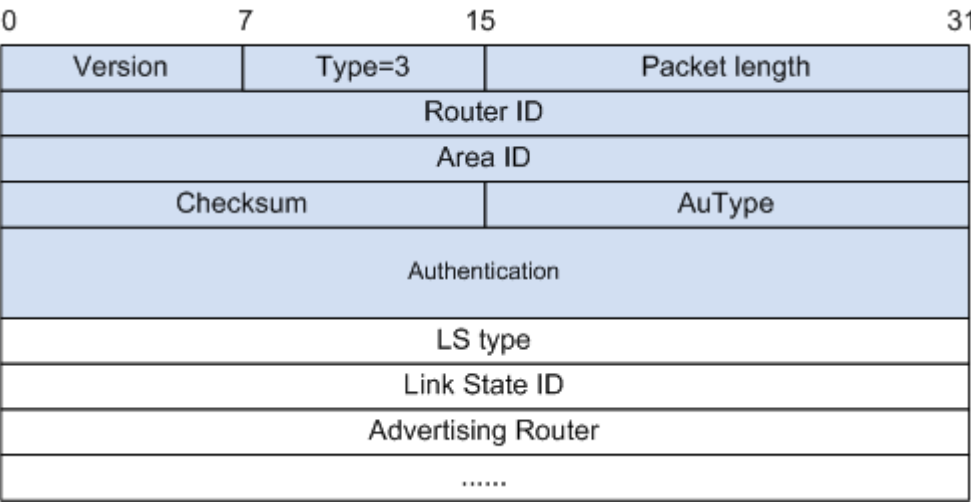


表 6-20 LSR 报文格式字段解释

| 字段名                | 长度    | 含义                                                 |
|--------------------|-------|----------------------------------------------------|
| LS type            | 32 比特 | LSA 的类型号。                                          |
| Link State ID      | 32 比特 | 根据 LSA 中的 LS Type 和 LSA description 在路由域中描述一个 LSA。 |
| Advertising Router | 32 比特 | 产生此 LSA 的路由器的 Router ID。                           |

LSU 报文

用来向对端路由器发送其所需要的 LSA 或者泛洪自己更新的 LSA，内容是多条 LSA（全部内容）的集合。LSU 报文（Link State Update Packet）在支持组播和广播的链路上是以组播形式将 LSA 泛洪出去。为了实现 Flooding 的可靠性传输，需要 LSAck 报文对其进行确认。对没有收到确认报文的 LSA 进行重传，重传的 LSA 是直接发送到邻居的。LSU 格式如图 6-20 所示。

图 6-20 LSU 报文格式

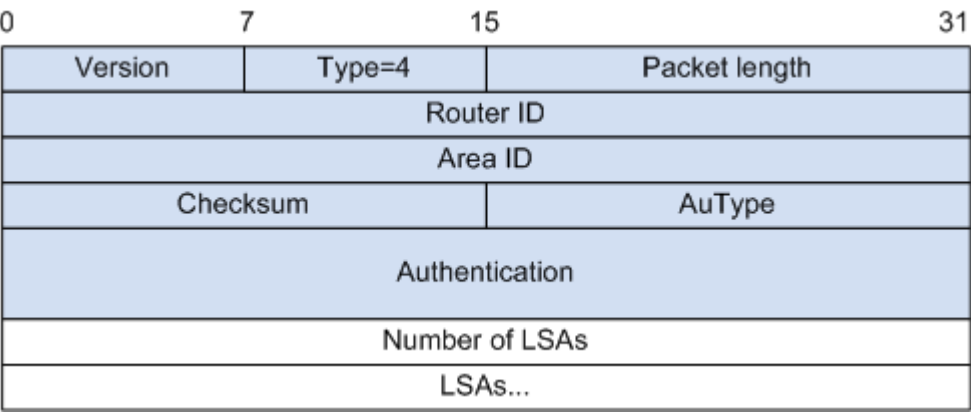


表 6-21 LSU 报文格式字段解释

| 字段名            | 长度    | 含义       |
|----------------|-------|----------|
| Number of LSAs | 32 比特 | LSA 的数量。 |

LSAck 报文

用来对接收到的 LSU 报文进行确认。内容是需要确认的 LSA 的 Header（一个 LSAck 报文可对多个 LSA 进行确认）。LSAck（Link State Acknowledgment Packet）报文根据不同的链路以单播或组播的形式发送，格式如图 6-21 所示。

图 6-21 LSAck 报文格式

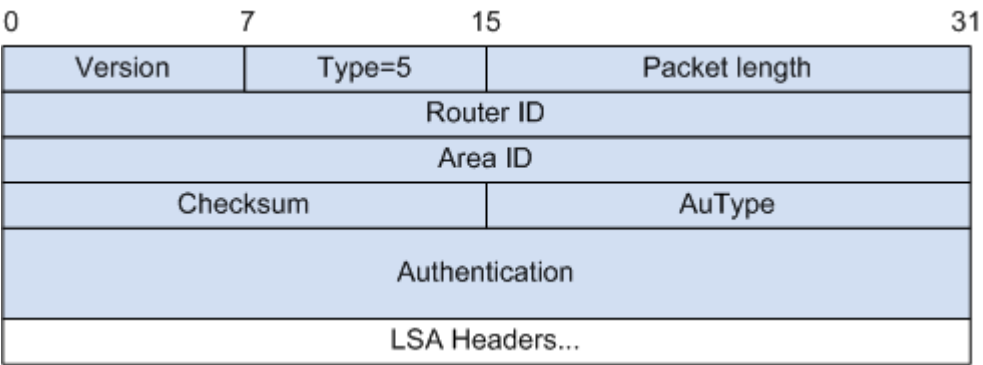


表 6-22 LSAck 报文格式字段解释

| 字段名          | 长度 | 含义                     |
|--------------|----|------------------------|
| LSAs Headers | -  | 通过 LSA 的头部信息确认收到该 LSA。 |

6.4.4.3 OSPF 的 LSA 格式

LSA 头部信息

所有的 LSA 都有相同的报文头，其格式如图 6-22 所示。

图 6-22 LSA 的头格式

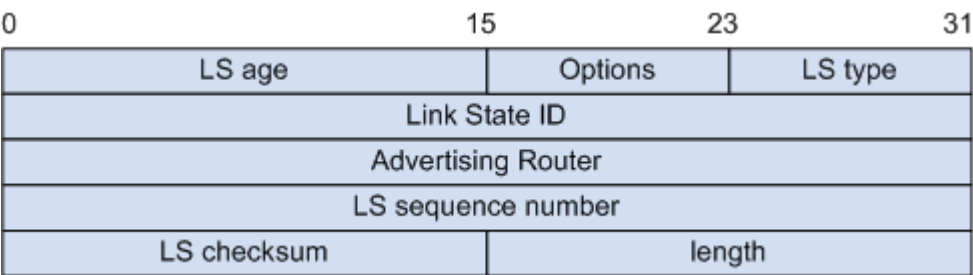


表 6-23 LSA 的头格式字段解释

| 字段名                | 长度    | 含义                                                                                                                                                                                                                                                    |
|--------------------|-------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| LS age             | 16 比特 | LSA 产生后所经过的时间，以秒为单位。无论 LSA 是在链路上传送，还是保存在 LSDB 中，其值都会在不停的增长。                                                                                                                                                                                           |
| Options            | 8 比特  | 可选项：<br><ul style="list-style-type: none"> <li>● E：允许泛洪 AS-External-LSA。</li> <li>● MC：转发 IP 组播报文。</li> <li>● N/P：处理 Type-7 LSA。</li> <li>● DC：处理按需链路。</li> </ul>                                                                                     |
| LS type            | 8 比特  | LSA 的类型：<br><ul style="list-style-type: none"> <li>● Type1：Router-LSA。</li> <li>● Type2：Network-LSA。</li> <li>● Type3：Network-summary-LSA。</li> <li>● Type4：ASBR-summary-LSA。</li> <li>● Type5：AS-External-LSA。</li> <li>● Type7：NSSA-LSA。</li> </ul> |
| Link State ID      | 32 比特 | 与 LSA 中的 LS Type 和 LSA description 一起在路由域中描述一个 LSA。                                                                                                                                                                                                   |
| Advertising Router | 32 比特 | 产生此 LSA 的路由器的 Router ID。                                                                                                                                                                                                                              |
| LS sequence number | 32 比特 | LSA 的序列号。其他路由器根据这个值可以判断哪个 LSA 是最新的。                                                                                                                                                                                                                   |
| LS checksum        | 16 比特 | 除了 LS age 外其它各域的校验和。                                                                                                                                                                                                                                  |
| length             | 16 比特 | LSA 的总长度，包括 LSA Header，以字节为单位。                                                                                                                                                                                                                        |

## Router-LSA

Router-LSA（Type1）：每个路由器都会产生，描述了路由器的链路状态和花费，在所属的区域内传播。Router-LSA 格式如图 6-23 所示。

图 6-23 Router-LSA 格式

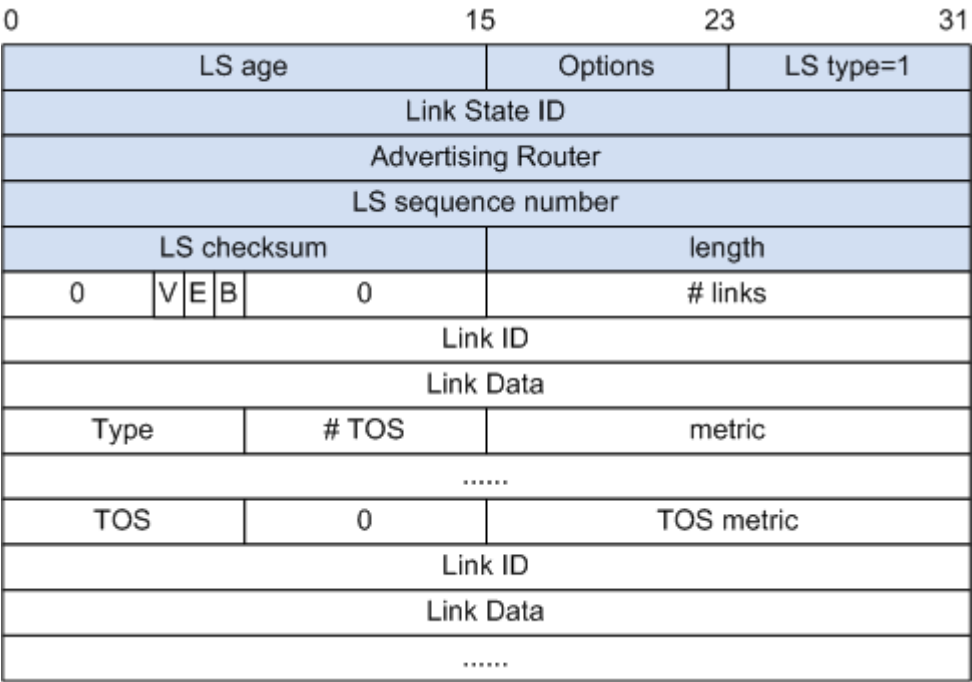


表 6-24 Router-LSA 格式字段解释

| 字段名             | 长度    | 含义                                                                                                                                                                  |
|-----------------|-------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Link State ID   | 32 比特 | 生成 LSA 的 Router ID。                                                                                                                                                 |
| V（Virtual Link） | 1 比特  | 如果产生此 LSA 的路由器是虚连接的端点，则置为 1。                                                                                                                                        |
| E（External）     | 1 比特  | 如果产生此 LSA 的路由器是 ASBR，则置为 1。                                                                                                                                         |
| B（Border）       | 1 比特  | 如果产生此 LSA 的路由器是 ABR，则置为 1。                                                                                                                                          |
| # links         | 16 比特 | LSA 中所描述的链路信息的数量，包括路由器上处于某区域中的所有链路和接口。                                                                                                                              |
| Link ID         | 32 比特 | 路由器所接入的目标，其值取决于连接的类型： <ul style="list-style-type: none"><li>● 1：Router ID。</li><li>● 2：DR 的接口 IP 地址。</li><li>● 3：网段 / 子网号。</li><li>● 4：虚连接中对端的 Router ID。</li></ul> |
| Link Data       | 32 比特 | 连接数据，其值取决于连接的类型： <ul style="list-style-type: none"><li>● unnumbered P2P：接口的索引值。</li><li>● stub 网络：子网掩码。</li><li>● 其它连接：路由器接口的 IP 地址。</li></ul>                      |

| 字段名        | 长度    | 含义                                                                                                                                            |
|------------|-------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| Type       | 8 比特  | 路由器连接的基本描述： <ul style="list-style-type: none"><li>● 1：点到点连接到另一台路由器。</li><li>● 2：连接到传输网络。</li><li>● 3：连接到 stub 网络。</li><li>● 4：虚拟链路。</li></ul> |
| # TOS      | 8 比特  | 连接不同的 TOS 数量。                                                                                                                                 |
| metric     | 16 比特 | 链路的开销值。                                                                                                                                       |
| TOS        | 8 比特  | 服务类型。                                                                                                                                         |
| TOS metric | 16 比特 | 和指定 TOS 值相关联的度量。                                                                                                                              |

Network-LSA

Network-LSA（Type2）：由广播网或 NBMA 网络中的 DR 产生，Network-LSA 中记录了这一网络上所有路由器的 Router ID，描述本网段的链路状态，在所属的区域内传播。Network-LSA 格式如图 6-24 所示。

图 6-24 Network-LSA 格式

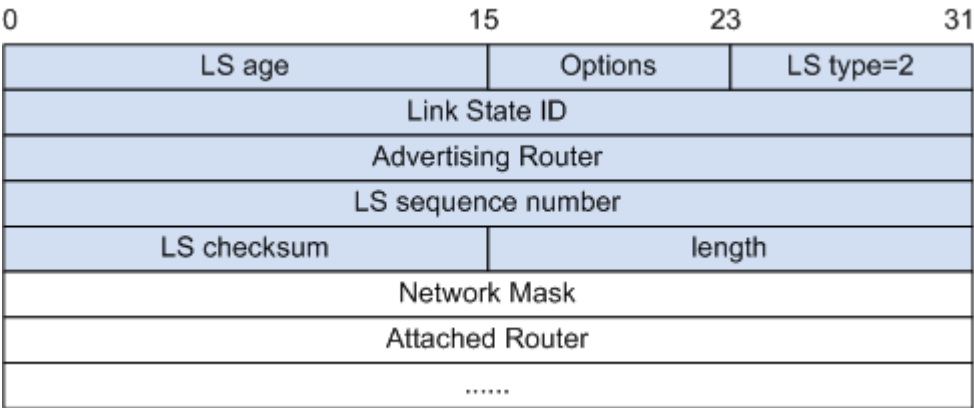


表 6-25 Network-LSA 格式字段解释

| 字段名             | 长度    | 含义                                             |
|-----------------|-------|------------------------------------------------|
| Link State ID   | 32 比特 | DR 的接口 IP 地址。                                  |
| Network Mask    | 32 比特 | 该广播网或 NBMA 网络地址的掩码。                            |
| Attached Router | 32 比特 | 连接在同一个网络上的所有路由器的 Router ID，也包括 DR 的 Router ID。 |

Summary-LSA

Network-summary-LSA（Type3）：描述区域内所有网段的路由，并通告给其他相关区域。

ASBR-summary-LSA（Type4）：描述到 ASBR 的路由，通告给除 ASBR 所在区域的其他相关区域。

Type3 和 Type4 的 LSA 有相同的格式，它们都是由 ABR 产生。Summary-LSA 格式如图 6-25 所示。

图 6-25 Summary-LSA 格式

|                    |            |         |                |
|--------------------|------------|---------|----------------|
| 0                  | 15         | 23      | 31             |
| LS age             |            | Options | LS type=3 or 4 |
| Link State ID      |            |         |                |
| Advertising Router |            |         |                |
| LS sequence number |            |         |                |
| LS checksum        |            | length  |                |
| Network Mask       |            |         |                |
| 0                  | metric     |         |                |
| TOS                | TOS metric |         |                |
| .....              |            |         |                |

表 6-26 Network-summary-LSA 格式字段解释

| 字段名           | 长度    | 含义                  |
|---------------|-------|---------------------|
| Link State ID | 32 比特 | 通告的网络地址。            |
| Network Mask  | 32 比特 | 该广播网或 NBMA 网络地址的掩码。 |
| metric        | 24 比特 | 到目的地址的路由开销。         |
| TOS           | 8 比特  | 服务类型。               |
| TOS metric    | 24 比特 | 和指定 TOS 值相关联的度量。    |

 说明

通告缺省路由时，Link State ID 和 Network Mask 都设置为 0.0.0.0。

表 6-27 ASBR Summary LSA 格式字段解释

| 字段名           | 长度    | 含义                |
|---------------|-------|-------------------|
| Link State ID | 32 比特 | ASBR 的 Router ID。 |

| 字段名          | 长度    | 含义               |
|--------------|-------|------------------|
| Network Mask | 32 比特 | 无意义，设置为 0.0.0.0。 |
| metric       | 24 比特 | 到目的地址的路由开销。      |
| TOS          | 8 比特  | 服务类型。            |
| TOS metric   | 24 比特 | 和指定 TOS 值相关联的度量。 |

AS-External-LSA

AS-External-LSA（Type5）：由 ASBR 产生，描述到 AS 外部的路由，这是五种 LSA 中，唯一一种通告到所有区域（除了 Stub 区域和 NSSA 区域）的 LSA。AS-External-LSA 格式如图 6-26 所示。

图 6-26 AS-External-LSA 格式

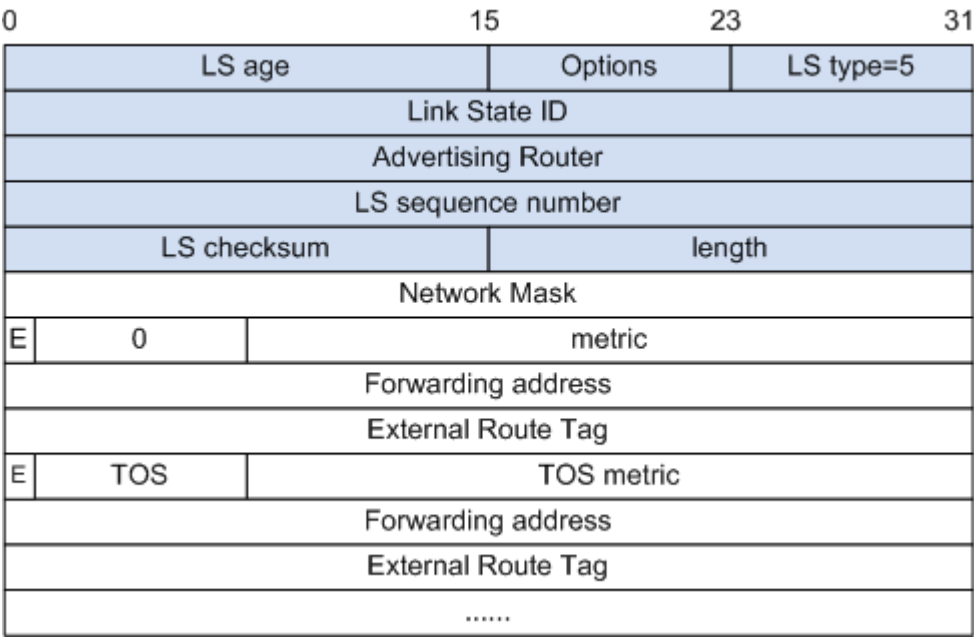


表 6-28 AS-External-LSA 格式字段解释

| 字段名           | 长度    | 含义                                                                                         |
|---------------|-------|--------------------------------------------------------------------------------------------|
| Link State ID | 32 比特 | 通告的网络地址。                                                                                   |
| Network Mask  | 32 比特 | 通告的目的地址的掩码。                                                                                |
| E             | 1 比特  | 外部度量值类型： <ul style="list-style-type: none"><li>● 0：第一类外部路由。</li><li>● 1：第二类外部路由。</li></ul> |



| 字段名                | 长度    | 含义                                          |
|--------------------|-------|---------------------------------------------|
| metric             | 24 比特 | 到目的地址的路由开销。                                 |
| Forwarding Address | 32 比特 | 到所通告的目的地址的报文将被转发到这个地址。                      |
| External Route Tag | 32 比特 | 添加到外部路由上的标记。OSPF 本身并不使用这个字段，它可以用来对外部路由进行管理。 |
| TOS                | 8 比特  | 服务类型。                                       |
| TOS metric         | 24 比特 | TOS 附加距离信息。                                 |

说明

Type5 的 LSA 可以用来通告缺省路由，此时 Link State ID 和 Network Mask 都设置为 0.0.0.0。

6.4.4.4 OSPF 区域特性

区域划分前

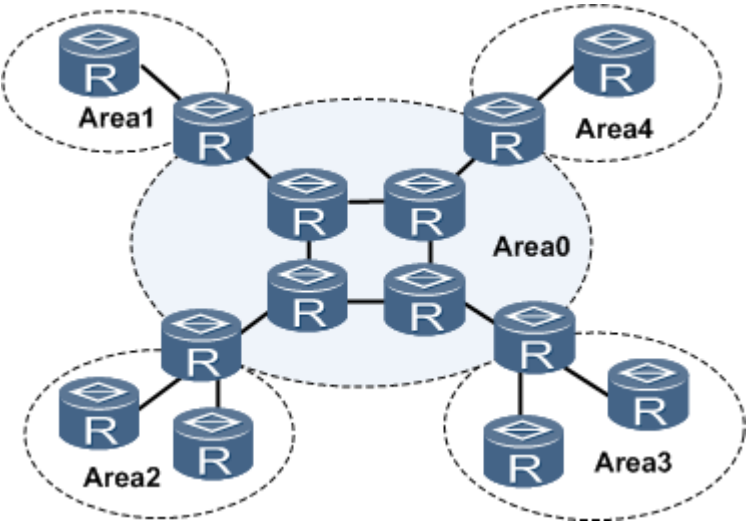
随着网络规模日益扩大，当一个大型网络中的路由器都运行 OSPF 路由协议时，路由器数量的增多会导致 LSDB 非常庞大，占用大量的存储空间，并使得运行 SPF 算法的复杂度增加，导致路由器负担很重。

在网络规模增大之后，拓扑结构发生变化的概率也增大，网络会经常处于“动荡”之中，造成网络中会有大量的 OSPF 协议报文在传递，降低了网络的带宽利用率。更为严重的是，每一次变化都会导致网络中所有的路由器重新进行路由计算。

区域划分后

OSPF 协议通过将自治系统划分成不同的区域（Area）来解决上述问题。区域是从逻辑上将路由器划分为不同的组，每个组用区域号（Area ID）来标识。区域的边界是路由器，而不是链路。一个网段（链路）只能属于一个区域，或者说每个运行 OSPF 的接口必须指明属于哪一个区域。如图 6-27 所示。

图 6-27 OSPF 区域划分



## 骨干区域

OSPF 划分区域之后，并非所有的区域都是平等的关系。其中有一个区域其区域号（Area ID）是 0，通常被称为骨干区域。骨干区域负责区域之间的路由，非骨干区域之间的路由信息必须通过骨干区域来转发。

对此，OSPF 有两个规定：

- 所有非骨干区域必须与骨干区域保持连通。
- 骨干区域自身也必须保持连通。

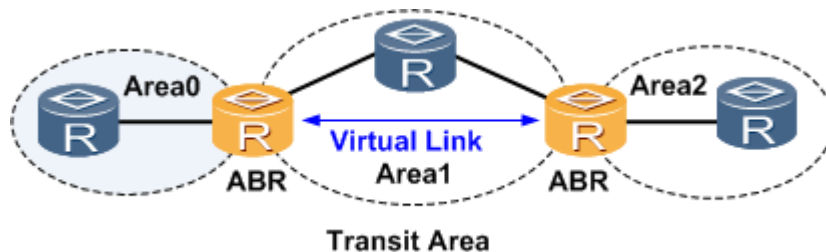
## 虚连接

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

虚连接是指在两台 ABR 之间通过一个非骨干区域而建立的一条逻辑上的连接通道。虚连接必须在两端同时配置方可生效。为虚连接两端提供一条非骨干区域内部路由的区域称为传输区域（Transit Area）。

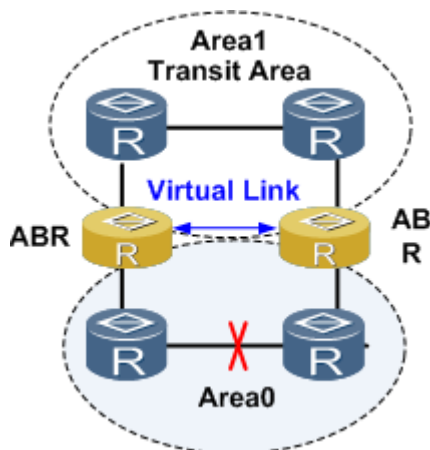
在图 6-28 中，Area2 与骨干区域之间没有直接相连的物理链路，但可以在 ABR 上配置虚连接，使 Area2 通过一条逻辑链路与骨干区域保持连通。

图 6-28 虚连接示意图之一



虚连接的另外一个应用是提供冗余的备份链路，当骨干区域因链路故障将被分割时，通过虚连接仍然可以保证骨干区域在逻辑上的连通性。如图 6-29 所示。

图 6-29 虚连接示意图之二



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

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

## Stub 区域

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

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

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

配置 Stub 区域时需要注意下列几点：

- 骨干区域不能配置成 Stub 区域。
- Stub 区域内不能存在 ASBR，即自治系统外部的路由不能在本区域内传播。
- 虚连接不能穿过 Stub 区域。

## NSSA 区域

在 RFC1587 中增加一类新的区域：NSSA（Not-So-Stubby Area），同时增加一类新的 LSA：NSSA-LSA（或称为 Type7 LSA）。

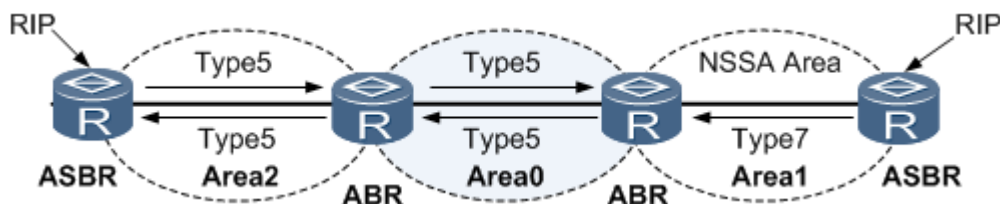
NSSA 其实是 Stub 区域的一个变形，它和 Stub 区域有许多相似的地方。NSSA 区域不允许存在 5 类 LSA。Type7 LSA 由 NSSA 的 ASBR 产生，仅在本 NSSA 内传播。当 Type7 LSA 到达 NSSA 的 ABR 时，由 ABR 将 Type7 LSA 转换成 Type5 LSA，传播到其他区域，负责转换 LSA 的 ABR 也称为转换路由器。

例如在图 6-30 所示的组网中，运行 OSPF 协议的自治系统包括 3 个区域：区域 0、区域 1 和区域 2，另外两个自治系统运行 RIP 协议。区域 1 被定义为 NSSA，区域 1 接收的 RIP 路由传播到 NSSA 的 ASBR 后，由 NSSA ASBR 产生 Type7 LSA 在区域 1 内传播，当 Type7 LSA 到达 NSSA 的 ABR 后，转换成 Type5 LSA 传播到区域 0 和区域 2。

另一方面，区域 2 中运行 RIP 的自治系统的 RIP 路由通过区域 2 的 ASBR 产生 Type-5 LSA 在 OSPF 自治系统中传播。但由于区域 1 是 NSSA，所以 Type-5 LSA 不会到达区域 1。

与 Stub 区域一样，NSSA 也不能配置虚连接。

图 6-30 NSSA

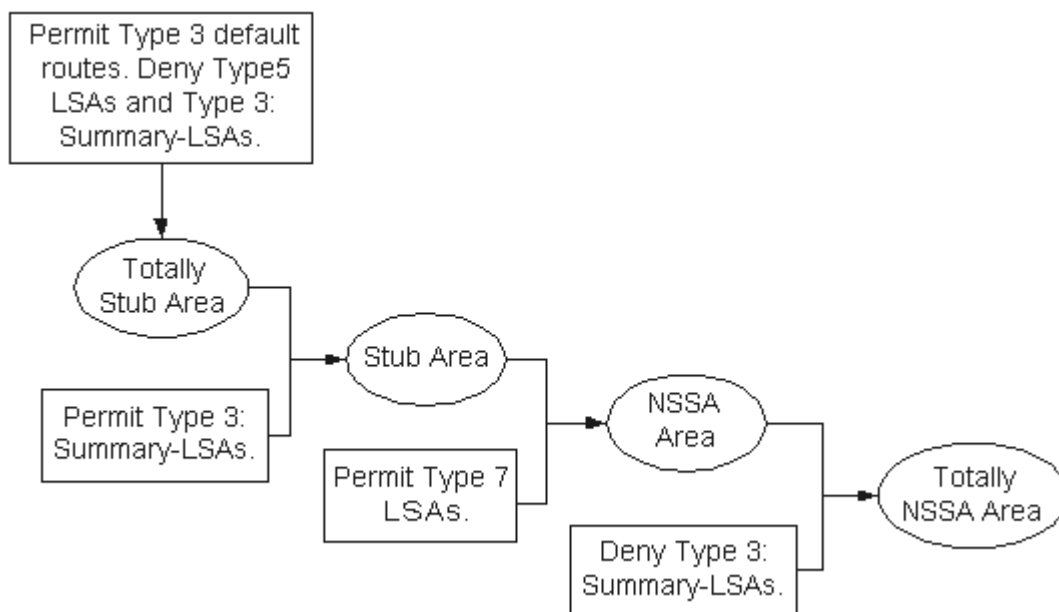


## 各区域特性

如图 6-31 所示。

下图表示了各区域（Totally Stub area、Stub area、NSSA area 和 Totally NSSA area）之间的区别。

图 6-31 区域特性示意图



各区域特性：

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

### 6.4.4.5 邻居和邻接

#### 邻居和邻接的区别

在 OSPF 中，邻居（Neighbors）和邻接（Adjacencies）是两个不同的概念。

OSPF 路由器启动后，便会通过 OSPF 接口向外发送 Hello 报文。收到 Hello 报文的 OSPF 路由器会检查报文中所定义的一些参数，如果双方一致就会形成邻居关系。

形成邻居关系的双方不一定都能形成邻接关系，这要根据网络类型而定。只有当双方成功交换 DD 报文，并交换 LSA 之后，才形成真正意义上的邻接关系。

## 邻居状态机

OSPF 共有 8 种邻居状态机，分别是：Down、Attempt、Init、2-way、Exstart、Exchange、Loading、Full。

- Down: 邻居会话的初始阶段，表明没有在邻居失效时间间隔内收到来自邻居路由器的 Hello 数据包。
- Attempt: 该状态仅发生在 NBMA 网络中，表明对端在 DeadInterval 超时后仍然没有回复 Hello 报文。在 Attempt 状态下，路由器每隔 Hello-Interval 定义的时间间隔向对端发送 Hello 报文。
- Init: 收到 Hello 报文后状态为 Init。
- 2-way: 收到的 Hello 报文中包含有自己的 Router ID，则状态为 2-way；如果不形成邻接关系则邻居状态机就停留在此状态，否则进入 Exstart 状态。
- Exstart: 如果形成邻接关系，则从 Init 状态转到 Exstart 状态，开始协商主从关系，并确定 DD 的序列号。
- Exchange: 使用第一个 DD 报文协商主从关系后继续交换 DD 报文，此时状态为 Exchange。
- Loading: DD 报文交换完成即 Exchange done，状态为 Loading。（如果路由器不需要邻接路由器的 LSA 则直接将状态转换为 Full）
- Full: LSR 重传列表为空，此时状态为 Full。

## 接口状态机

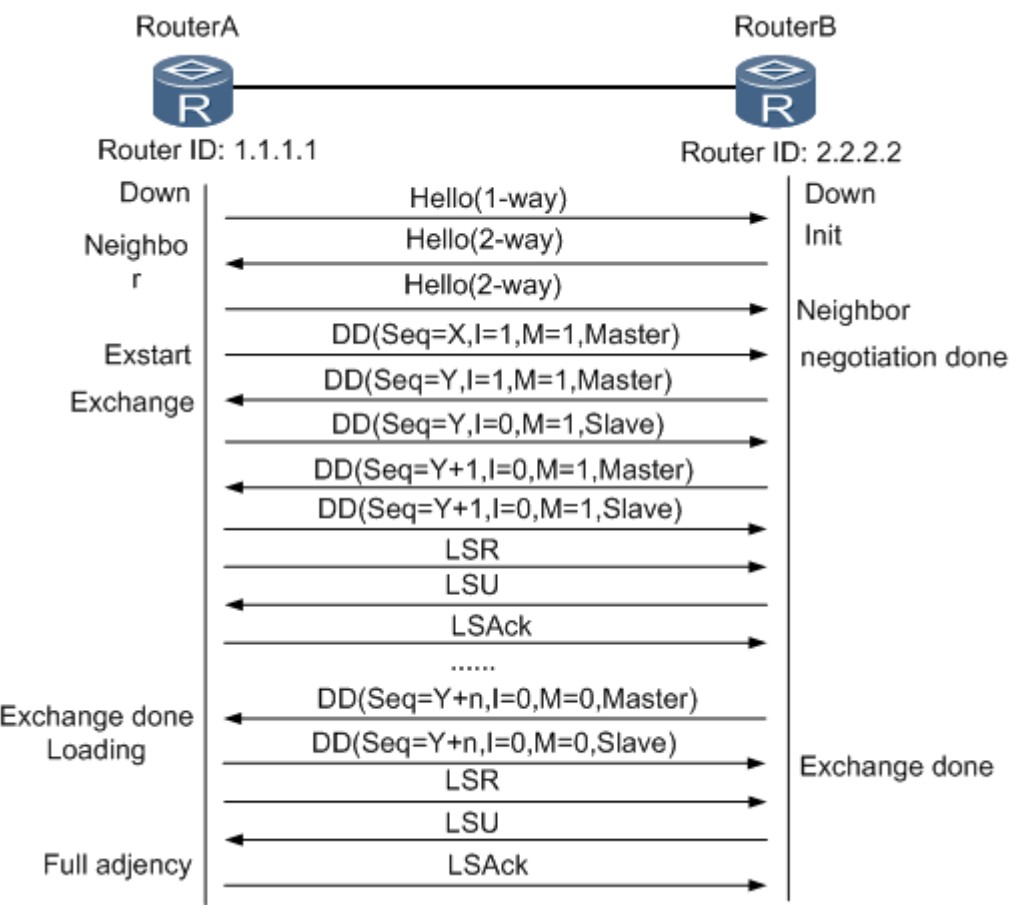
OSPF 共有 7 种接口状态机，分别是：Down、Point-to-Point、Waiting、DR、Backup、DRother、Loopback。

- Down: 接口的初始阶段，表明此接口没有数据进行发送和接收。
- Point-to-Point: 此接口状态仅适用于点对点、点对多点和虚电路等网络类型，表明接口尝试和另一端路由器建立邻接关系。
- Waiting: 此接口状态仅适用于广播型和 NBMA 等网络类型，表明接口开始发送和接收 Hello 报文，尝试识别网络上是否存在 DR 和 BDR。
- DR: 表明该路由器是所连接网段的指定路由器（DR），和该网段其他路由器建立邻接关系。
- Backup: 表明该路由器是所连网段的备份指定路由器（BDR），和该网段其他路由器建立邻接关系。
- DRother: 表明该路由器不是该网段的 DR 或 BDR，仅和该网段的 DR 和 BDR 建立邻接关系。
- Loopback: 表明该接口通过软件或硬件方式设置为环回接口，多为测试网络连通性使用。

## 广播链路

广播链路邻接关系建立过程如图 6-32 所示。

图 6-32 广播链路邻接关系建立过程



在广播网络中，DR、BDR 和网段内的每一台路由器都形成邻接关系，但 DR other 之间只形成邻居关系。

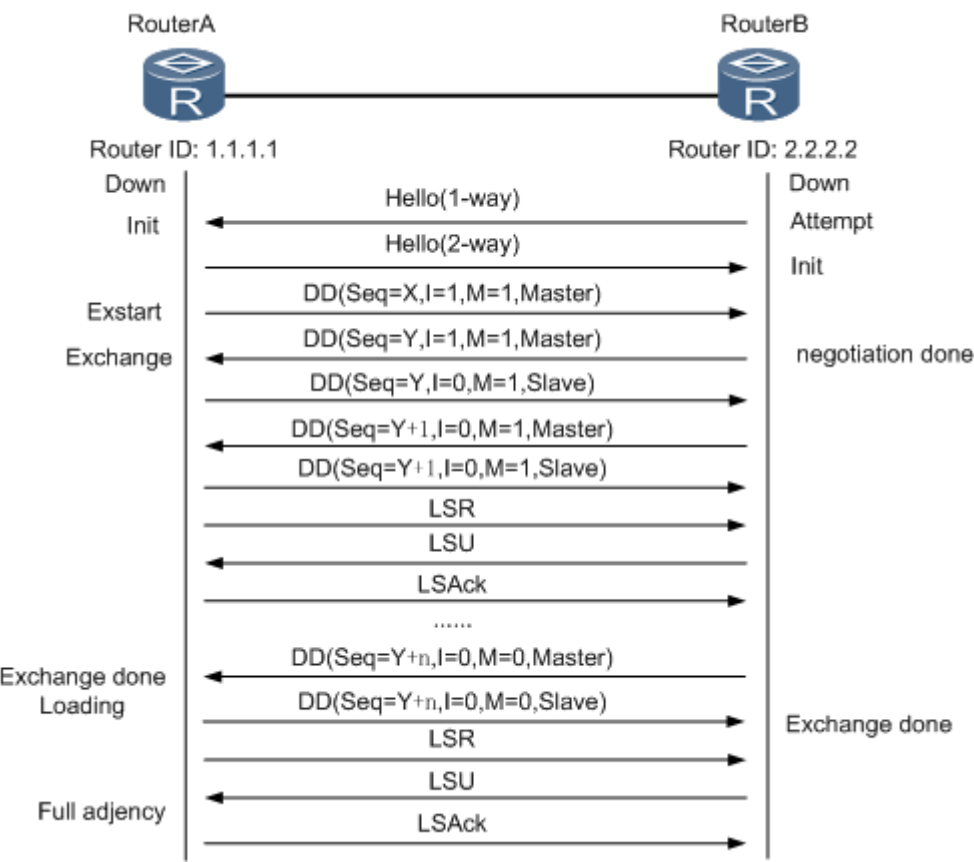
1. 当一台路由器刚加入该网段时，发送 Hello 报文到组播地址（224.0.0.5），与每台路由器建立邻居关系。  
1-way 与 2-way 的区别：路由器收到 Hello 报文中包含有自己的 Router ID 就是 2-way 状态，如果没有包含就是 1-way。
2. 路由器之间形成邻居后达到 Exstart 状态，路由器开始和邻居协商 Master/Slave 关系以确定 DD 报文的序列号。DD 报文的序号由 Master 决定，并每交换完成一个 DD 报文后序号加一，主从关系由路由器的 Router ID 大小决定。当没有报文继续交换时，将报文字段中的 M 位置零，表示这是最后一个报文。DD 报文是单播发送的。
3. 在交换 DD 报文的过程中，路由器知道需要从和自己邻接的路由器获取哪些 LSA，也知道有哪些 LSA 需要更新，这时发送 LSR 报文，请求相关的 LSAs。当 DD 报文交换完成后，即 Exchange done，此时状态为 Loading。
4. 当接收到这些 LSA，由接收路由器回复 LSAck 报文确认收到 LSA。
5. 当路由器交换完 DD 报文并更新所有的 LSA 后，此时邻接关系建立完成。

NBMA 链路

NBMA 链路上邻接关系建立过程如图 6-33 所示。



图 6-33 NBMA 链路上邻接关系建立过程



在 NBMA 链路上，每台 DR 优先级非零的路由器之间都会定时发送 Hello 报文，这是为了确保网段的 DR、BRD 的正确选举。所有路由器定期向 DR 和 BDR 发送 Hello 报文，并在收到 DR 优先级非零的路由器发来的 Hello 报文后进行回复。所以，在一个 NBMA 网络上，需要控制有能力成为 DR 的路由器数量。所有路由器只与 DR 和 BDR 之间形成邻接关系，并不是所有的路由器都形成邻居关系。

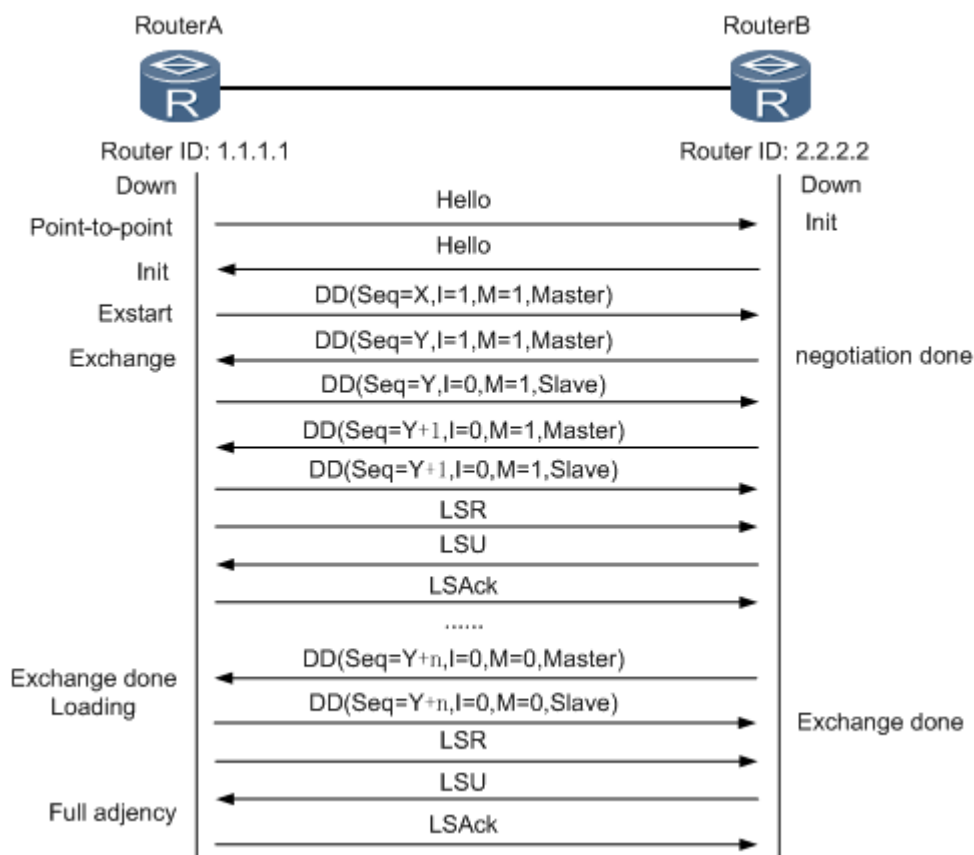
NBMA 链路和广播链路的邻接关系建立过程只在交换 DD 报文前不太一样。

1. 如果接口状态为 Down，网段上的 DR 和 BDR 依然会向该接口发送 Hello 报文，但时间间隔为 PollInterval。
2. 向一个 Down 接口发送 Hello 报文后邻居的状态变为 Attempt，当路由器收到 Hello 报文后会状态变为 Init，然后再回复一个 Hello 报文。
3. 以下过程同广播链路邻接建立。

点到点/点到多点链路

点到点、点到多点链路上邻接关系建立过程如图 6-32 所示。

图 6-34 点到点/点到多点链路上邻接关系建立过程



在点到点、点到多点和虚连接的链路上，邻接关系的建立过程和广播链路一样。唯一不同的是不需要选举 DR 和 BDR，DD 报文是组播发送的。

1. 当接口 UP 后，路由器试图与邻居路由器形成邻接关系。发送 Hello 报文的时间间隔为 HelloInterval。
2. 当路由器收到 Hello 报文后状态转为 Init，这点与广播链路和 NBMA 链路不同。（广播链路和 NBMA 链路中路由器收到 Hello 报文将状态转为 Attempt）
3. 以下过程同广播链路邻接建立。

## 邻接关系建立的条件

在上述邻居状态机的变化中，有两处决定是否建立邻接关系：

- 当与邻居的双向通讯初次建立时。
- 当网段中的 DR 和 BDR 发生变化时。

如果不形成邻接，邻居状态就保持 2-way。

当至少满足以下一种情况时，就会形成邻接关系：

- 链路类型为点到点
- 链路类型为点到多点



- 链路类型为虚连接
- 路由器自身为 DR
- 路由器自身为 BDR
- 邻居路由器为 DR
- 邻居路由器为 BDR

#### 6.4.4.6 网络类型和 DR 选举

### OSPF 支持的网络类型

OSPF 根据链路层协议类型将网络分为下列四种类型：

- 广播（Broadcast）类型：当链路层协议是 Ethernet、FDDI 时，OSPF 缺省认为网络类型是 Broadcast。在该类型的网络中，通常以组播形式（224.0.0.5：含义是 OSPF 路由器的预留 IP 组播地址；224.0.0.6：含义是 OSPF DR 和 BDR 的预留 IP 组播地址）发送 Hello 报文、LSU 报文和 LSAck 报文；以单播形式发送 DD 报文和 LSR 报文。
- NBMA（Non-Broadcast Multi-Access）类型：当链路层协议是帧中继、ATM 或 X.25 时，OSPF 缺省认为网络类型是 NBMA。在该类型的网络中，以单播形式发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。
- 点到多点 P2MP（point-to-multipoint）类型：没有一种链路层协议会被缺省的认为是 Point-to-Multipoint 类型。点到多点必须是由其他的网络类型强制更改的。常用做法是将非全连通的 NBMA 改为点到多点的网络。在该类型的网络中，以组播形式（224.0.0.5）发送 Hello 报文，以单播形式发送其他协议报文（DD 报文、LSR 报文、LSU 报文、LSAck 报文）。
- 点到点 P2P（point-to-point）类型：当链路层协议是 PPP、HDLC 和 LAPB 时，OSPF 缺省认为网络类型是 P2P。在该类型的网络中，以组播形式（224.0.0.5）发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。

### NBMA 网络的配置原则

NBMA 网络是指非广播、多点可达的网络，比较典型的有 ATM 和帧中继网络。

对于接口类型为 NBMA 的网络需要进行一些特殊的配置。由于无法通过广播 Hello 报文的形式发现相邻路由器，必须手工为该接口指定相邻路由器的 IP 地址，以及该相邻路由器是否有选举权等。NBMA 网络中，即使接口的状态为 Down，有能力成为 DR 的路由器依然会向该接口发送 Hello 报文，但时间间隔为 PollInterval。

NBMA 网络必须是全连通的，即网络中任意两台路由器之间都必须直接可达。如果部分路由器之间没有直接可达的链路时，应将接口配置成 P2MP 方式。如果路由器在 NBMA 网络中只有一个对端，也可将接口类型改为 P2P 方式。

NBMA 与 P2MP 网络之间的区别：

- NBMA 是指那些全连通的、非广播、多点可达网络。而点到多点的网络，则并不需要一定是全连通的。
- 在 NBMA 上需要选举 DR 与 BDR，而在点到多点网络中没有 DR 与 BDR。
- NBMA 是一种缺省的网络类型，点到多点必须是由其它的网络强制更改的。最常见的做法是将非全连通的 NBMA 改为点到多点的网络。

- NBMA 用单播发送报文，需要手工配置邻居。点到多点网络支持组播方式发送报文。

## DR 和 BDR 的意义

在广播网和 NBMA 网络中，任意两台路由器之间都要传递路由信息。如果网络中有  $n$  台路由器，则需要建立  $n \times (n-1)/2$  个邻接关系。这使得任何一台路由器的路由变化都会导致多次传递，浪费了带宽资源。为解决这一问题，OSPF 协议定义了指定路由器 DR (Designated Router)，所有路由器都只将信息发送给 DR，由 DR 将网络链路状态广播出去。

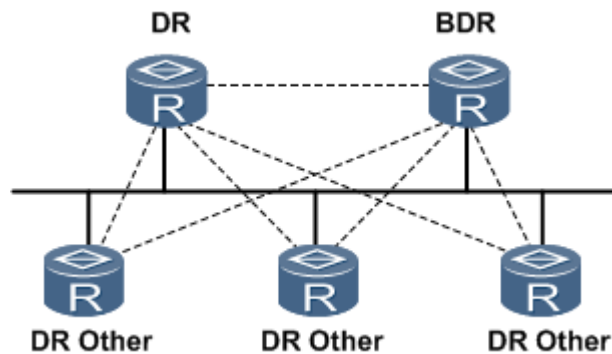
如果 DR 由于某种故障而失效，则网络中的路由器必须重新选举 DR，并与新的 DR 同步。这需要较长的时间，在这段时间内，路由的计算有可能是不正确的。为了能够缩短这个过程，OSPF 提出了 BDR (Backup Designated Router) 的概念。

BDR 实际上是对 DR 的一个备份，在选举 DR 的同时也选举出 BDR，BDR 也和本网段内的所有路由器建立邻接关系并交换路由信息。当 DR 失效后，BDR 会立即成为 DR。由于不需要重新选举，并且邻接关系事先已建立，所以这个过程是非常短暂的。当然这时还需要再重新选举出一个新的 BDR，虽然一样需要较长的时间，但并不会影响路由的计算。

除 DR 和 BDR 之外的路由器（称为 DR Other）之间将不再建立邻接关系，也不再交换任何路由信息。这样就减少了广播网和 NBMA 网络上各路由器之间邻接关系的数量。

在图 6-35 中，用实线代表以太网物理连接，虚线代表建立的邻接关系。可以看到，采用 DR/BDR 机制后，5 台路由器之间只需要建立 7 个邻接关系就可以了。

图 6-35 DR 和 BDR 示意图



## DR 和 BDR 的选举过程

DR 和 BDR 是由本网段中所有的路由器共同选举出来的。路由器接口的 DR 优先级决定了该接口在选举 DR、BDR 时所具有资格。本网段内 DR 优先级大于 0 的路由器都可作为“候选人”。

选举中使用的“选票”就是 Hello 报文。每台路由器将自己选出的 DR 写入 Hello 报文中，发给网段上的其他路由器。当处于同一网段的两台路由器同时宣布自己是 DR 时，DR 优先级高者胜出。如果优先级相等，则 Router ID 大者胜出。如果一台路由器的优先级为 0，则它不会被选举为 DR 或 BDR。

需要注意的是：

- 只有在广播或 NBMA 类型接口时才会选举 DR，在点到点或点到多点类型的接口上不需要选举 DR。
- DR 是指某个网段的概念，是针对路由器的接口而言的。某台路由器在一个接口上的角色可能是 DR，但在另一个接口上有可能是 BDR，或者是 DR Other。
- 若 DR、BDR 已经选择完毕，当一台新路由器加入后，即使它的 DR 优先级值最大，也不会立即成为该网段中的 DR。
- DR 并不一定就是 DR 优先级最大的路由器；同理，BDR 也并不一定就是 DR 优先级第二大的路由器。

### 6.4.4.7 OSPF 缺省路由和路由聚合

#### 缺省路由

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

OSPF 缺省路由通常应用于下面两种情况：

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

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

不同区域缺省路由发布原则如表 6-29 所示。

表 6-29 OSPF 缺省路由发布原则

| 区域类型              | 作用                                                                                                                                                                         |
|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 普通区域              | 缺省情况下，普通 OSPF 区域内的 OSPF 路由器是不会产生缺省路由的，即使它有缺省路由。必须手动通过命令进行配置才能产生缺省路由。配置完成后，路由器会产生一个缺省 LSA（Type3 LSA），并且通告到整个 OSPF 自治域中。<br>当网络中缺省路由通过其他路由进程产生时，路由器必须能够将缺省路由通告到整个 OSPF 自治域中。 |
| Stub Area         | Stub 区域不允许自治系统外部的路由（Type5 LSA）在区域内传播，区域内的路由器通过 ABR 学到自治系统外部的路由。<br>ABR 会自动产生一条缺省 LSA（Type3 LSA）通告到整个 Stub 区域内。                                                             |
| Totally Stub Area | Totally Stub 区域与 Stub 区域的区别仅在于前者不允许区域间路由（Type3 LSA）。区域内的路由器通过 ABR 学到其他区域和自治系统外部的路由。<br>ABR 会自动产生一条缺省 LSA（Type3 LSA）通告到整个 Stub 区域内。                                         |

| 区域类型              | 作用                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| NSSA Area         | <p>NSSA 区域与 Stub 区域的区别在于前者可以引入自治系统外部的路由，但一样不允许 ASE LSA（Type5 LSA）在区域内传播。</p> <ul style="list-style-type: none"><li>● 如果希望到达自治系统外部的路由通过该区域的 ASBR 到达，而其它外部路由通过其它区域出去的话，ABR 会产生一条缺省的 NSSA LSA（Type7 LSA），通告到整个 NSSA 区域内。</li><li>● 如果希望所有的外部路由只通过本区域 NSSA 的 ASBR 到达，要在 ASBR 上手动通过命令进行配置，ASBR 会产生一条缺省的 NSSA LSA（Type7 LSA），通告到整个 NSSA 区域内。</li></ul> <p>在 ABR 上无论路由表中是否存在路由 0.0.0.0，都会产生 Type7 LSA 缺省路由，在 ASBR 上当路由表中存在路由 0.0.0.0，才会产生 Type7 LSA 缺省路由。</p> <p>因为缺省路由只是在本 NSSA 区域内泛洪，并没有泛洪到整个 OSPF 域中，所以本 NSSA 区域内的路由器在找不到路由之后可以从该 NSSA 的 ASBR 出去，但不能实现其他 OSPF 域的路由从这个出口出去。</p> |
| Totally NSSA Area | <p>Totally NSSA 与 NSSA 区域的区别仅在于前者不允许区域间路由（Type3 LSA）注入。区域内的路由器通过 ABR 知道其他区域的路由。</p> <p>ABR 会自动产生一条缺省的 Type3 LSA 通告到整个 NSSA 区域内。</p>                                                                                                                                                                                                                                                                                                                                                                                                                                        |

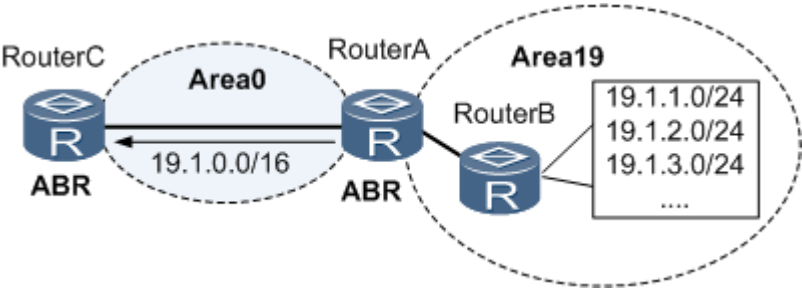
路由聚合

路由聚合是指：具有相同前缀的路由信息，ABR 可以将它们聚合在一起，只发布一条路由到其它区域。

AS 被划分成不同的区域后，区域间可以通过路由聚合来减少路由信息，从而减小路由表的规模，提高路由器的性能。

例如，图 6-36 中，Area19 内有三条区域内路由 19.1.1.0/24，19.1.2.0/24，19.1.3.0/24，如果此时在路由器 A 上配置了路由聚合，将三条路由聚合成一条 19.1.0.0/16，则路由器 A 就只生成一条聚合后的 LSA，并发布给 Area0 中的其他路由器。

图 6-36 路由聚合示意图



OSPF 有两类聚合：

- ABR 聚合

ABR 向其它区域发送路由信息时，以网段为单位生成 Type3 LSA。如果该区域中存在一些连续的网段，则可以使用 **abr-summary** 命令将这些连续的网段聚合成一个网段。这样 ABR 只发送一条聚合后的 LSA，所有属于本命令指定的聚合网段范围的 LSA 将不再会被单独发送出去，这样可减少其它区域中 LSDB 的规模。

一旦将某一网络的聚合网段加入到区域中，该区域中所有属于这一聚合网段的 IP 地址的内部路由都不再被独立地广播到别的区域，而只广播整个聚合网段的路由信息。

- ASBR 聚合

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

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

#### 6.4.4.8 OSPF 报文验证

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

支持两种验证方式：

- 区域验证方式
- 接口验证方式

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

当两种验证方式都存在时，优先使用接口验证方式。

#### 6.4.4.9 多进程和多实例

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

路由器的一个接口只能属于某一个 OSPF 进程。

OSPF 支持多实例，可以运行在 VPN 网络中的 PE-CE 之间。

在 BGP MPLS VPN 网络中，属于同一个 VPN 的多个 Site 可以使用 OSPF 作为内部路由协议。然而，它们会被看作属于不同的自治系统来处理。这样，在一个节点学到的 OSPF 路由，将被作为外部路由传送给另一节点。这种处理方式导致了比较高的 OSPF 路由协议流量，并带来了一些原本可以避免的网络管理问题。

在目前的实现中，可以在 PE 上通过配置域 ID 来区分 Site 所在的 VPN。属于同一个 VPN 的不同 Site 之间彼此看做是直接相连的。这样，PE 路由器之间交换 OSPF 路由信息时就好像是通过一条专线相连，改善了网络管理并使 OSPF 的应用更为有效。

 说明

OSPF VPN 多实例的详细介绍请参见《特性描述 MPLS 与 VPN》。

## 6.5 OSPFv3

OSPFv3 多用于支持大型 IPv6 网络，是基于链路状态的路由协议，与基于距离矢量的路由协议相比具有更快的收敛速度。

### 6.5.1 介绍

#### 定义

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

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

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

#### 目的

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

OSPFv3 与 OSPFv2 的不同在于：

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

### 6.5.2 参考标准和协议

与 OSPFv3 相关的参考标准与协议如下：

- RFC 4552: Authentication/Confidentiality for OSPFv3
- RFC 4970: Extensions to OSPF for Advertising Optional Router Capabilities
- RFC 2740: OSPF for IPv6

### 6.5.3 可获得性

#### License 支持

本特性无需 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                      |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br><br>USG5500 V200R001C00 |

6.5.4 原理描述

6.5.4.1 OSPFv3 报文格式

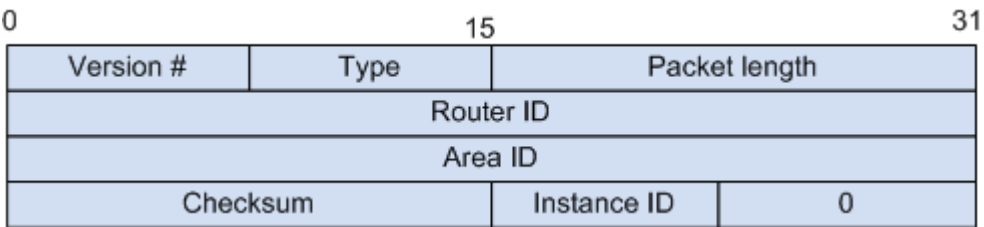
和 OSPF 一样，OSPFv3 也有五种报文类型，分别是 Hello 报文、DD 报文、LSR 报文、LSU 报文和 LSAck 报文。

OSPF 报文格式请参见 6.4.4.2 OSPF 报文格式。

这五种报文有相同的报文头，但是它和 OSPF 的报文头有一些区别，没有认证字段，这是因为由 IPv6 的扩展报文头 AH 和 ESP 来保证安全性。另外，还多了一个 Instance ID 字段，用来支持在同一条链路上运行多个实例。

OSPFv3 的报文头如图 6-37 所示。

图 6-37 OSPFv3 报文头示意图



6.5.4.2 OSPFv3 的 LSA 类型

LSA 是 OSPFv3 协议计算和维护路由信息的主要来源。在 RFC2740 中定义了七类 LSA，描述如下：

- **Router-LSA：**由每个路由器生成，描述本路由器的链路状态和花费，只在路由器所处区域内传播。
- **Network-LSA：**由广播网络和 NBMA 网络的 DR 生成，描述本网段的链路状态，只在 DR 所处区域内传播。
- **Inter-Area-Prefix-LSA：**和 OSPFv2 中的 Type-3 LSA 类似，该 LSA 由区域边界路由器 ABR 生成，在与该 LSA 相关的区域内传播。每一条 Inter-Area-Prefix-LSA 描述

了一条到达本自治系统内其他 OSPFv3 区域的 IPv6 地址前缀（IPv6 Address Prefix）路由。

- **Inter-Area-Router-LSA:** 和 OSPFv2 中的 Type-4 LSA 类似，该 LSA 由区域边界路由器 ABR 生成，在与该 LSA 相关的区域内传播。每一条 Inter-Area-Router-LSA 描述了一条去往本自治系统内其他区域的 ASBR 的路由。
- **AS-External-LSA:** 由自治系统边界路由器 ASBR 生成，描述到达其它 AS 的路由，传播到整个 AS（Stub 区域除外）。AS 的缺省路由也可以用 AS-External-LSA 来描述。
- **Link-LSA:** 路由器为每一条链路生成一个 Link-LSA，在本地链路范围内传播。每一个 Link-LSA 描述了该链路上所联系的 IPv6 地址前缀，包含 Link-local 地址。
- **Intra-Area-Prefix-LSA:** 每个 Intra-Area-Prefix-LSA 包含路由器上的 IPv6 前缀信息，Stub 区域信息或传输区域（Transit Area）的网段信息，该 LSA 在区域内传播。由于 Router-LSA 和 Network-LSA 不再包含地址信息，导致了 Intra-Area-Prefix-LSA 的引入。

### 6.5.4.3 OSPFv3 的网络类型和区域类型

#### OSPFv3 支持的网络类型

| 网络类型                                   | 含义                                                                                                                                                                                                                                                                                    |
|----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 广播类型（Broadcast）                        | <p>当链路层协议是 Ethernet、FDDI 时，缺省情况下，OSPFv3 认为网络类型是 Broadcast。</p> <p>在该类型的网络中：</p> <ul style="list-style-type: none"> <li>● 通常以组播形式发送 Hello 报文、LSU 报文和 LSAck 报文。其中，FF02::5 为 OSPFv3 路由器的预留 IPv6 组播地址；FF02::6 为 OSPFv3 DR/BDR 的预留 IPv6 组播地址。</li> <li>● 以单播形式发送 DD 报文和 LSR 报文。</li> </ul> |
| NBMA 类型（Non-broadcast multiple access） | <p>当链路层协议是帧中继、ATM 或 X.25 时，缺省情况下，OSPFv3 认为网络类型是 NBMA。</p> <p>在该类型的网络中，以单播形式发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。</p>                                                                                                                                                         |
| 点到多点 P2M 类型（Point-to-Multipoint）       | <p>没有一种链路层协议会被缺省的认为是 Point-to-Multipoint 类型。点到多点必须是由其他的网络类型强制更改的。常用做法是将非全连通的 NBMA 改为点到多点的网络。</p> <p>在该类型的网络中：</p> <ul style="list-style-type: none"> <li>● 以组播形式（FF02::5）发送 Hello 报文。</li> <li>● 以单播形式发送其他协议报文（DD 报文、LSR 报文、LSU 报文、LSAck 报文）。</li> </ul>                              |
| 点到点 P2P 类型（point-to-point）             | <p>当链路层协议是 PPP、HDLC 和 LAPB 时，缺省情况下，OSPFv3 认为网络类型是 P2P。</p> <p>在该类型的网络中，以组播形式（FF02::5）发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。</p>                                                                                                                                               |



## OSPFv3 区域类型

| 区域类型                          | 作用                                        |
|-------------------------------|-------------------------------------------|
| 完全 Stub 区域（Totally Stub Area） | 允许 ABR 发布的 Type3 缺省路由，不允许自治系统外部路由和区域间的路由。 |
| Stub 区域（Stub Area）            | 和 Totally Stub 区域的不同在于，该区域允许区域间路由。        |

## OSPFv3 路由类型

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

表 6-30 中按优先级从高到低顺序列出了路由类型。

表 6-30 路由类型

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

### 6.5.4.4 OSPFv3 和 OSPF 协议比较

相同点：

- 网络类型和接口类型
- 接口状态机和邻居状态机
- 链路状态数据库（LSDB）
- 洪泛机制（Flooding mechanism）

- 相同类型的报文：Hello 报文、DD 报文、LSR 报文、LSU 报文和 LSAck 报文
- 路由计算基本相同

不同点：

- OSPFv3 基于链路，而不是网段  
OSPFv3 运行在 IPv6 协议上，IPv6 是基于链路而不是网段的。  
这样，在配置 OSPFv3 时，不需要考虑是否配置在同一网段，只要在同一链路，就可以不配置 IPv6 全局地址而直接建立联系。
- OSPFv3 上移除了 IP 地址的意义  
这样做的目的是为了使“拓扑与地址分离”。OSPFv3 可以不依赖 IPv6 全局地址的配置来计算出 OSPFv3 的拓扑结构。IPv6 全局地址仅用于 Vlink 接口及报文的转发。
- OSPFv3 的报文及 LSA 格式发生改变
  - OSPFv3 报文不包含 IP 地址。
  - OSPFv3 的 Router LSA 和 Network LSA 里不包含 IP 地址。IP 地址部分由新增的两类 LSA（Link LSA 和 Intra Area Prefix LSA）宣告。
  - OSPFv3 的 Router ID、Area ID 和 LSA Link State ID 不再表示 IP 地址，但仍保留 IPv4 地址格式。
  - 广播、NBMA 及 P2MP 网络中，邻居不再由 IP 地址标识，只由 Router ID 标识。
- OSPFv3 的 LSA 报文里添加 LSA 的洪泛范围  
OSPFv3 在 LSA 报文头的 LSA Type 里，添加 LSA 的洪泛范围，这使得 OSPFv3 的路由器更加灵活，可以处理不能识别类型的 LSA：
  - OSPFv3 可存储或洪泛不识别报文，而 OSPF 只简单丢弃掉不识别报文。
  - OSPFv3 允许洪泛范围为区域或链路本地（Link-local），并且设置 U 位（报文可按洪泛范围为链路本地来处理）的不识别报文存储或通过 Stub 区域。例如，路由器 A 和 B 都可识别某类 LSA，它们之间通过路由器 C 连接，但路由器 C 不识别该类 LSA。这样，当路由器 A 洪泛此类 LSA 时，路由器 C 虽然不识别，但还是可以洪泛给路由器 B，B 收到后继续处理。  
如果运行的是 OSPF 协议，只会丢弃不能识别的报文，B 则不能收到此类 LSA。
- OSPFv3 支持一个链路上多个进程  
OSPF 的一个物理接口，只能配置一个进程。  
OSPFv3 中，同一物理接口，可以配置多个进程，这些进程用不同的 Instance ID 区分。也就是说，在一条物理链路上可运行多个 OSPFv3 实例，分别与链路对端路由器建立邻居及发送报文，且互不干扰。  
这样可以充分共享同一链路资源，模拟多台 OSPFv3 路由器，更高效的利用有限的路由器资源。
- OSPFv3 利用 IPv6 链路本地地址  
IPv6 使用链路本地（Link-local）地址在同一链路上发现邻居及自动配置等。运行 IPv6 的路由器不转发目的地址为链路本地地址的 IPv6 报文，此类报文只在同一链路有效。链路本地单播地址从 FE80/10 开始。  
OSPFv3 是运行在 IPv6 上的路由协议，同样使用链路本地地址来维持邻居，同步 LSA 数据库。除 Vlink 外的所有 OSPFv3 接口都使用链路本地地址作为源地址及下一跳来发送 OSPFv3 报文。  
这样的好处是：

- 不需要配置 IPv6 全局地址，就可以得到 OSPFv3 拓扑，实现拓扑与地址分离。
- 通过在链路上泛洪的报文不会传到其他链路上，来减少报文不必要的泛洪来节省带宽。
- OSPFv3 移除所有认证字段  
OSPFv3 的认证直接使用 IPv6 的认证及安全处理，不再需要其自身来完成认证，使用协议时只需关注协议本身即可。
- 新增两种 LSA  
Link LSA：用于路由器宣告各个链路上对应的链路本地地址及其所配置的 IPv6 全局地址，仅在链路内洪泛。  
Intra Area Prefix LSA：用于向其他路由器宣告本路由器或本网络（广播网及 NBMA）的 IPv6 全局地址信息，在区域内洪泛。
- OSPFv3 只通过 Router ID 来标识邻居  
OSPF 在广播网，NBMA 及 P2MP 网络中是通过 IPv4 接口地址来标识的。  
OSPFv3 只通过 Router ID 来标识邻居，这样即使没有配置 IPv6 全局地址，或是 IPv6 全局地址配置都不在同一网段，OSPFv3 的邻居还是可以建立并维护的，以达到“拓扑与地址分离”的目的。

## 6.6 IS-IS

IS-IS 协议在运营商中得到广泛的应用，是直接运行在链路层的内部网关路由协议，具有收敛速度快，分层结构清晰的特点，非常适合大规格网络。

### 6.6.1 介绍

#### 定义

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

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

本文所指的 IS-IS，如不加特殊说明，均指集成 IS-IS。

IS-IS 属于内部网关协议 IGP（Interior Gateway Protocol），用于自治系统内部。IS-IS 是一种链路状态协议，使用最短路径优先 SPF（Shortest Path First）算法进行路由计算，与 OSPF 协议有很多相似之处，但是 IS-IS 是直接运行于链路层之上的路由协议。

#### 目的

IS-IS 支持 IP, OSI 两种路由，可以在 IP 和 OSI 双重环境中工作，支持灵活的 TLV 编址方式，具有良好的协议扩展性。

IS-IS 路由收敛速度快，结构清晰，适合于大规模网络。近年来在运行商网络中得到越来越广泛的应用。

## 6.6.2 参考标准和协议

与 IS-IS 相关的参考标准与协议如下：

- RFC 1195: Use of OSI IS-IS for routing in TCP/IP and dual environments
- RFC 3567: Intermediate System to Intermediate System (IS-IS) Cryptographic Authentication
- RFC 2973: IS-IS Mesh Groups
- RFC 2763: Dynamic Hostname Exchange Mechanism for IS-IS

## 6.6.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 6.6.4 原理描述

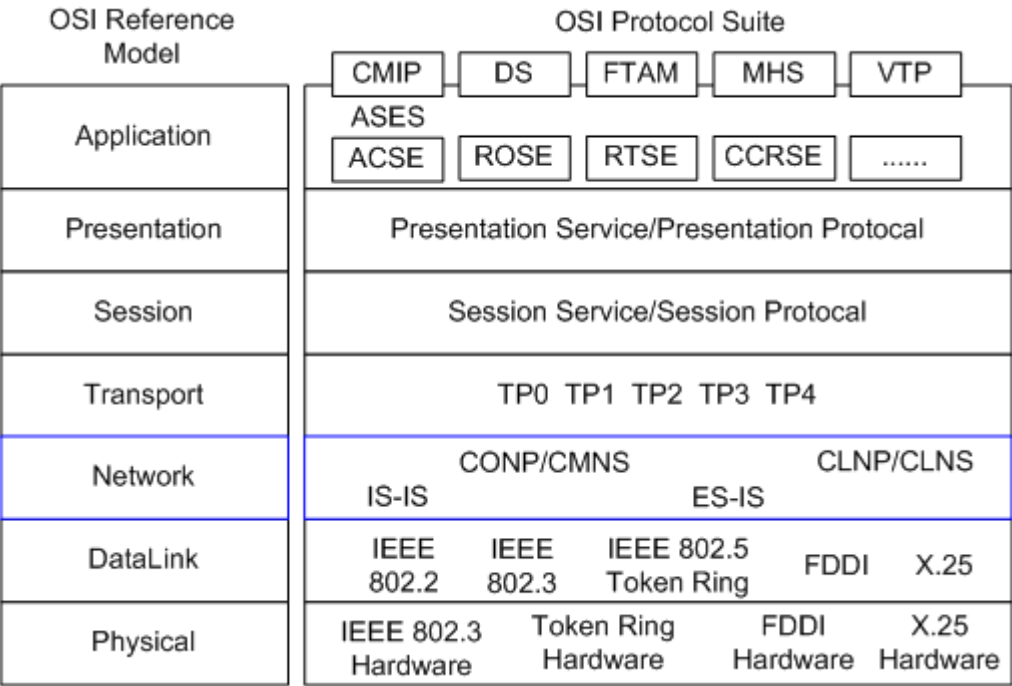
### 6.6.4.1 IS-IS 的发展

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

随着 TCP/IP 协议的流行，IS-IS 在 RFC1195 中也加入了对 IP 协议的支持，实现了 IP 路由能力，因此 IS-IS 也被称为集成化 IS-IS（Integrated IS-IS）或者称为 Dual IS-IS。

IS-IS 是 ISO 定义的 OSI 协议栈中无连接网络服务 CLNS（Connectionless Network Service）的一部分。

图 6-38 OSI 结构模型



OSI 协议采用体系化（或层次化 Hierarchical）编址，我们能通过 NSAP（Network Service Access Point）来寻址 OSI 网络中处于传输层的各种服务，它类似于 TCP/IP 协议中的 IP 地址+端口号。

OSI 协议的几个常用术语：

CLNS（Connectionless Network Service）：无连接网络服务。

CLNP（Connectionless Network Protocol）：无连接网络协议。

CMNS（Connection-Mode Network Service）：连接模式网络服务。

CONP（Connection-Oriented Network Protocol）：面向连接网络协议。

OSI 通过 CLNP 实现 CLNS，通过 CONP 实现 CMNS。

CLNS 由以下三个协议构成：

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

OSI 协议和 IP 协议中的一些术语所指代的作用是基本一致的，对应关系如表 6-31 所示。

表 6-31 OSI 与 IP 相对应的概念

| 缩略语 | OSI 中的概念                 | IP 中对应的概念  |
|-----|--------------------------|------------|
| IS  | Intermediate System 中间系统 | Router 路由器 |
| ES  | End System 端系统           | Host 主机    |

| 缩略语   | OSI 中的概念                                    | IP 中对应的概念                         |
|-------|---------------------------------------------|-----------------------------------|
| DIS   | Designated Intermediate System 选举中间系统       | Designated Router(DR)OSPF 中的选举路由器 |
| SysID | System ID 系统 ID                             | OSPF 中的 Router ID                 |
| PDU   | Protocol Data Unit 报文数据单元                   | IP 报文                             |
| LSP   | Link State Protocol Data Unit 链路状态协议数据单元    | OSPF 中的 LSA 用来描述链路状态              |
| NSAP  | Network Service Access Point 网络服务访问点(网络层地址) | IP 地址                             |

6.6.4.2 IS-IS 协议地址结构

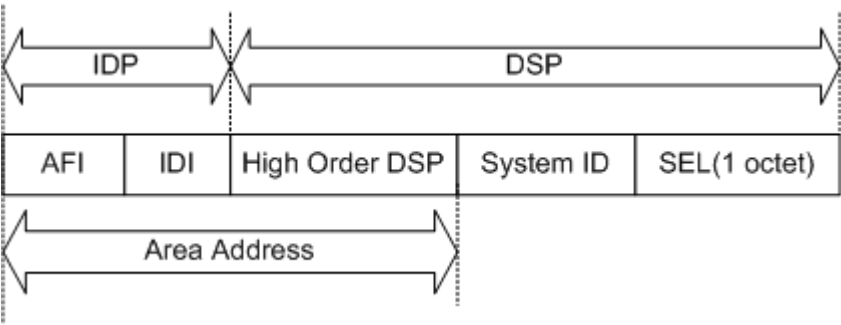
NSAP 是 OSI 协议中用于定位资源的地址。ISO 采用如图 6-39 所示的地址结构，即 NSAP，它由 IDP（Initial Domain Part）和 DSP（Domain Specific Part）组成。IDP 相当于 IP 地址中的主网络号，DSP 相当于 IP 地址中的子网号和主机地址。

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

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

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

图 6-39 IS-IS 协议的地址结构示意图



- 区域地址
  - IDP 和 DSP 中的 HODSP（High Order DSP）一起，既能够标识路由域，也能够标识路由域中的区域，因此，它们一起被称为区域地址（Area Address），相当于 OSPF 中的区域编号。同一个路由域中不允许有相同的区域地址。同一区域中路由器的 Level-1 区域地址必须相同。
  - 一般情况下，一台路由器只需要配置一个区域地址，且同一区域中所有节点的区域地址都要相同。为了支持区域的平滑合并、分割及转换，在 USG 的实现中，一台路由器最多可配置 3 个区域地址。

- System ID

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

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

1. 将 IP 地址 168.10.1.1 的每个十进制数都扩展为 3 位，不足 3 位的在前面补 0，得到扩展后地址：168.010.001.001。
2. 将扩展后的地址 168.010.001.001 分为 3 部分，每部分由 4 位数字组成，重新组合为 System ID：1680.1000.1001。

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

- SEL

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

- 路由方式

由于这种地址结构明确地定义了区域，Level-1 路由器很容易识别出发往它所在的区域之外的报文，这些报文是需要转交给最近的 Level-1-2 路由器的。

- Level-1 路由器利用 System ID 进行区域内的路由，如果发现报文的目的地地址不属于自己所在的区域，就将报文转发给最近的 Level-1-2 路由器。
- Level-2 路由器根据区域地址进行区域间的路由。

## NET

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

通常情况下，一台路由器配置一个 NET 即可，当区域需要重新划分时，例如将多个区域合并，或者将一个区域划分为多个区域，这种情况下配置多个 NET 可以在重新配置时仍然能够保证路由的正确性。由于一台路由器在一个 IS-IS 进程中区域地址最多可配置 3 个，所以 NET 最多也只能配 3 个。在配置多个 NET 时，必须保证它们的 System ID 都相同。

例如有 NET 为：ab.cdef.1234.5678.9abc.00，则其中区域地址为 ab.cdef，System ID 为 1234.5678.9abc，SEL 为 00。



说明

位于同一区域内的路由器区域地址必须相同。

### 6.6.4.3 IS-IS 的报文结构

IS-IS 的 PDU 有以下类型：Hello 报文，LSP，CSNP，PSNP。

## PDU 报文格式

IS-IS 报文是直接封装在数据链路层的帧结构中的。PDU 可以分为两个部分，报文头和变长字段部分。其中头部又可分为通用报文头和专用报文头。对于所有 PDU 来说，通用报文头都是相同的，但专用报文头根据 PDU 类型不同而有所差别。如图 6-40 所示。

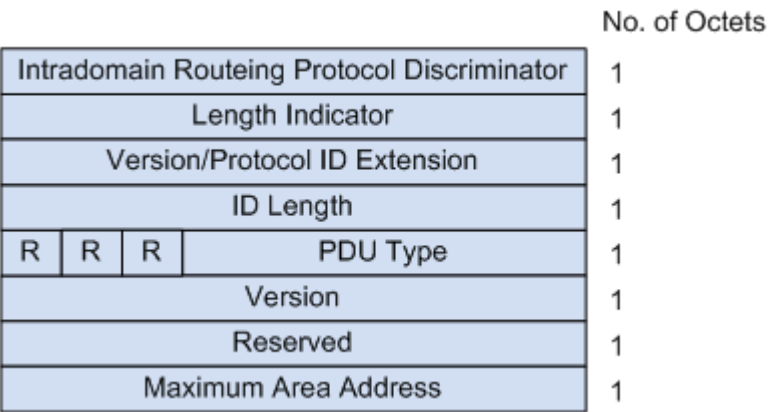
图 6-40 PDU 报文格式



通用报文格式

所有的 PDU 都有相同的通用报文头格式，如图 6-41 所示。

图 6-41 PDU 通用报文头格式



主要字段的解释如下：

- Intradomain Routing Protocol Discriminator：域内路由选择协议鉴别符，设置为 0x83。
- Length Indicator：PDU 报文头的长度（包括通用报文头和专用报文头），以字节为单位。
- Version/Protocol ID Extension：版本/协议标识扩展，设置为 1（0x01）。
- ID Length：NSAP 地址或 NET 中 System ID 区域的长度。
  - 值为 0 时，表示 System ID 区域的长度为 6 字节。
  - 值为 255 时，表示 System ID 区域为空（即长度为 0）。
- R（Reserved）：保留，设置为 0。
- PDU Type：PDU 的类型。IS-IS PDU 共有 9 种类型，详细信息请参考表 6-32。
- Version：设置为 1（0x01）。
- Maximum Area Address：支持的最大区域个数。设置为 1～254 的整数，表示该 IS-IS 进程实际所允许的最大区域地址数；设置为 0，表示该 IS-IS 进程最大只支持 3 个区域地址数。

表 6-32 PDU 类型对应关系表

| 类型值 | PDU 类型                      | 简称         |
|-----|-----------------------------|------------|
| 15  | Level-1 LAN IS-IS Hello PDU | L1 LAN IIH |



| 类型值 | PDU 类型                                | 简称         |
|-----|---------------------------------------|------------|
| 16  | Level-2 LAN IS-IS Hello PDU           | L2 LAN IIH |
| 17  | Point-to-Point IS-IS Hello PDU        | P2P IIH    |
| 18  | Level-1 Link State PDU                | L1 LSP     |
| 20  | Level-2 Link State PDU                | L2 LSP     |
| 24  | Level-1 Complete Sequence Numbers PDU | L1 CSNP    |
| 25  | Level-2 Complete Sequence Numbers PDU | L2 CSNP    |
| 26  | Level-1 Partial Sequence Numbers PDU  | L1 PSNP    |
| 27  | Level-2 Partial Sequence Numbers PDU  | L2 PSNP    |

Hello 报文格式

Hello 报文用于建立和维持邻居关系，也称为 IIH（IS-to-IS Hello PDUs）。其中，广播网中的 Level-1 路由器使用 Level-1 LAN IIH；广播网中的 Level-2 路由器使用 Level-2 LAN IIH；非广播网络中则使用 P2P IIH。它们的报文格式有所不同。  
广播网中的 Hello 报文格式如图 6-42 所示（蓝色部分是通用报文头）。

图 6-42 L1/L2 LAN IIH 格式

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

主要字段的解释如下：

- **Reserved/Circuit Type:** 高位的 6 比特保留，值为 0。低位的 2 比特表示路由器的类型（01 表示 L1，10 表示 L2，11 表示 L1/L2）。
- **Source ID:** 发出 Hello 报文的路由器的 System ID。
- **Holding Time:** 保持时间。在此时间内如果没有收到邻居发来的 Hello 报文，则中止已建立的邻居关系。
- **PDU Length:** PDU 的总长度，单位是字节。
- **Priority:** 选举 DIS 的优先级，取值范围为 0 ~ 127。数值越大，优先级越高。
- **LAN ID:** 包括 DIS 的 System ID 和一字节的伪节点 ID。

点到点网络中的 Hello 报文格式如图 6-43 所示。

图 6-43 P2P IIH 格式

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

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

LSP 报文格式

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

两类 LSP 有相同的报文格式，如图 6-44 所示。

图 6-44 L1/L2 LSP 格式

|                                             |     |   |          |    | No. of Octets |   |
|---------------------------------------------|-----|---|----------|----|---------------|---|
| Intradomain Routeing Protocol Discriminator |     |   |          |    | 1             |   |
| Length Indicator                            |     |   |          |    | 1             |   |
| Version/Protocol ID Extension               |     |   |          |    | 1             |   |
| ID Length                                   |     |   |          |    | 1             |   |
| R                                           | R   | R | PDU Type |    | 1             |   |
| Version                                     |     |   |          |    | 1             |   |
| Reserved                                    |     |   |          |    | 1             |   |
| Maximum Area Address                        |     |   |          |    | 1             |   |
| PDU Length                                  |     |   |          |    | 2             |   |
| Remaining Lifetime                          |     |   |          |    | 2             |   |
| LSP ID                                      |     |   |          |    | ID Length+2   |   |
| Sequency Number                             |     |   |          |    | 4             |   |
| Checksum                                    |     |   |          |    | 2             |   |
| P                                           | ATT |   |          | OL | IS Type       | 1 |
| Variable Length Fields                      |     |   |          |    |               |   |

主要字段的解释如下：

- PDU Length: PDU 的总长度，以字节为单位。
- Remaining Lifetime: LSP 的生存时间，以秒为单位。
- LSP ID: 由三部分组成，System ID、伪节点 ID（一字节）和 LSP 分片后的编号（一字节）。
- Sequency Number: LSP 的序列号。
- Checksum: LSP 的校验和。
- P (Partition Repair): 仅与 L2 LSP 有关，表示路由器是否支持自动修复区域分割。
- ATT (Attachment): 由 Level-1-2 路由器产生，用来指明始发路由器是否与其它区域相连。虽然此标志位也存在于 Level-1 和 Level-2 的 LSP 中，但实际上此字段只和 Level-1-2 路由器始发的 L1 LSP 有关。此字段有 4bit，用来表明相连的区域所使用的度量方式。

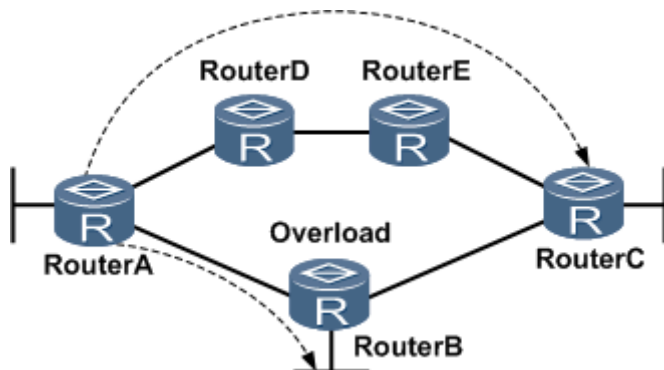
从右至左这 4 位依次表示如下所示：

- 第 4 位：缺省度量。
- 第 5 位：时延度量。
- 第 6 位：代价度量。
- 第 7 位：差错度量。
- OL (LSDB Overload): 过载标志位。设置了过载标志位的 LSP 虽然还会在网络中扩散，但是在计算通过超载路由器的路由时不会被采用。即，对路由器设置过载位后，其它路由器在进行 SPF 计算时不会考虑这台路由器。当路由器内存不足时，系统自动在发送的 LSP 报文中设置过载标志位。

如图 6-45 所示，假设正常情况下 RouterA 到 RouterC 的报文都是经过 RouterB 转发，但如果 RouterB 的 OL 位置 1，则 RouterA 会认为 RouterB 的路由不完整，从

而将报文通过 RouterD、RouterE 转发给 RouterC，但到 RouterB 直连地址的报文不受影响。

图 6-45 LSDB Overload 示意图



- IS Type: 生成 LSP 的路由器的类型。用来指明是 Level-1 还是 Level-2 路由器（01 表示 Level-1，11 表示 Level-2）。

## SNP 格式

时序报文 SNP（Sequence Number PDU）通过描述全部或部分数据库中的 LSP 来同步各 LSDB（Link-State DataBase），从而维护 LSDB。

SNP 包括 CSNP（Complete SNP，全时序报文）和 PSNP（Partial SNP，部分时序报文），进一步又可分为 L1 CSNP、L2 CSNP、L1 PSNP 和 L2 PSNP。

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

CSNP 的报文格式如图 6-46 所示。

图 6-46 L1/L2 CSNP 格式

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

主要字段的解释如下：

- Source ID：发出 SNP 报文的路由器的 System ID。
- Start LSP ID：CSNP 报文中第一个 LSP 的 ID 值。
- End LSP ID：CSNP 报文中最后一个 LSP 的 ID 值。

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

PSNP 的报文格式如图 6-47 所示。

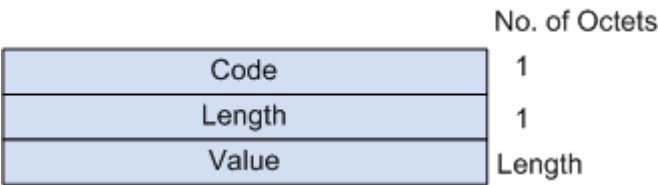
图 6-47 L1/L2 PSNP 格式

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

CLV 格式

PDU 中的变长字段部分是多个 CLV（Code-Length-Value）三元组。其格式如图 6-48 所示。CLV 也称为 TLV（Type-Length-Value）。

图 6-48 CLV 格式



不同 PDU 类型所包含的 CLV 是不同的。如表 6-33 所示。

表 6-33 PDU 类型和包含的 CLV 名称

| CLV Code | 名称                                        | 所应用的 PDU 类型 |
|----------|-------------------------------------------|-------------|
| 1        | Area Addresses                            | IIH、LSP     |
| 2        | IS Neighbors（LSP）                         | LSP         |
| 4        | Partition Designated Level2 IS            | L2 LSP      |
| 6        | IS Neighbors（MAC Address）                 | LAN IIH     |
| 7        | IS Neighbors（SNPA Address）                | LAN IIH     |
| 8        | Padding                                   | IIH         |
| 9        | LSP Entries                               | SNP         |
| 10       | Authentication Information                | IIH、LSP、SNP |
| 128      | IP Internal Reachability Information      | LSP         |
| 129      | Protocols Supported                       | IIH、LSP     |
| 130      | IP External Reachability Information      | L2 LSP      |
| 131      | Inter-Domain Routing Protocol Information | L2 LSP      |
| 132      | IP Interface Address                      | IIH、LSP     |

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

6.6.4.4 IS-IS 区域

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

## Level-1 与 Level-2

- Level-1 路由器

Level-1 路由器负责区域内的路由，它只与属于同一区域的 Level-1 和 Level-1-2 路由器形成邻居关系，维护一个 Level-1 的 LSDB，该 LSDB 包含本区域的路由信息，到区域外的报文转发给最近的 Level-1-2 路由器。

- Level-2 路由器

Level-2 路由器负责区域间的路由，可以与本区域和其它区域的 Level-2 和 Level-1-2 路由器形成邻居关系，维护一个 Level-2 的 LSDB，该 LSDB 包含区域间的路由信息。所有 Level-2 级别（即形成 Level-2 邻居关系）的路由器组成路由域的骨干网，负责在不同区域间通信，路由域中 Level-2 级别的路由器必须是连续的，以保证骨干网的连续性。只有 Level-2 级别的路由器才能直接与区域外的路由器交换数据报文或路由信息。

- Level-1-2 路由器

同时属于 Level-1 和 Level-2 的路由器称为 Level-1-2 路由器，可以与同一区域的 Level-1 和 Level-1-2 路由器形成 Level-1 邻居关系，也可以与其他区域的 Level-2 和 Level-1-2 路由器形成 Level-2 的邻居关系。Level-1 路由器必须通过 Level-1-2 路由器才能连接至其他区域。Level-1-2 路由器维护两个 LSDB，Level-1 的 LSDB 用于区域内路由，Level-2 的 LSDB 用于区域间路由。

 说明

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

图 6-49 所示为一个运行 IS-IS 协议的网络，它与 OSPF 的多区域网络拓扑结构非常相似。整个 backbone 区域不仅包括 Area1 中的所有路由器，还包括其它区域的 Level-1-2 路由器。

图 6-49 IS-IS 拓扑结构图之一

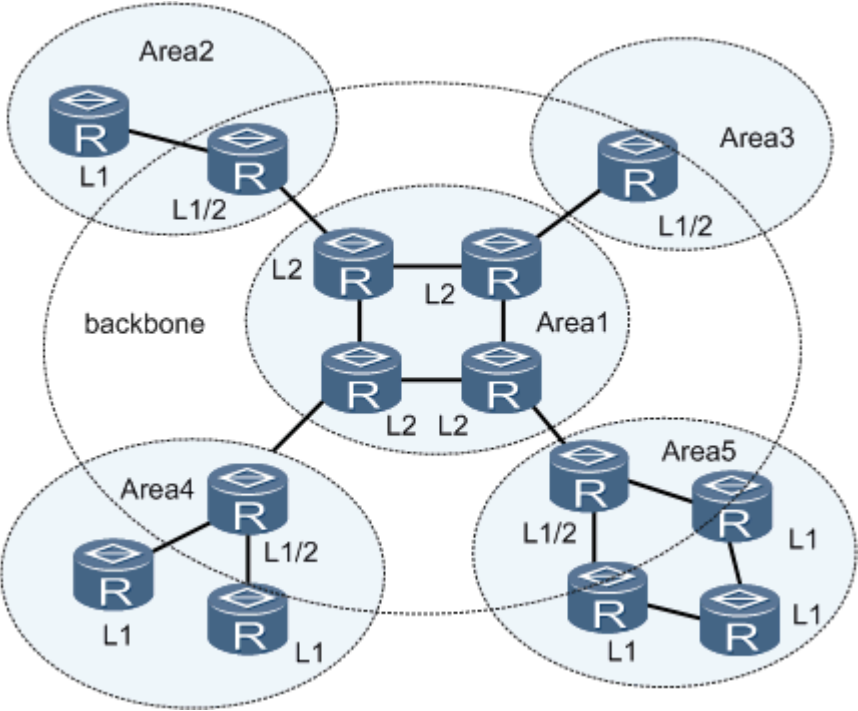
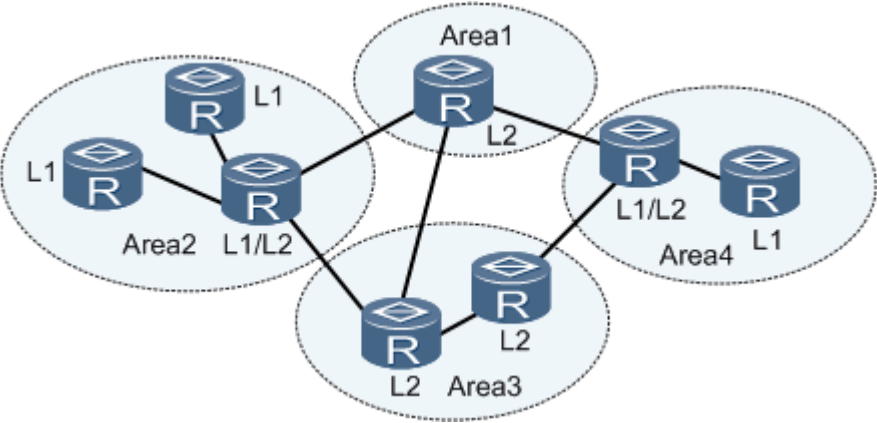


图 6-50 是 IS-IS 的另外一种拓扑结构图。所有连续的 Level-1-2 和 Level-2 路由器构成了 IS-IS 的骨干区域。在这个拓扑中，Level-2 级别的路由器分别属于不同的区域，并没有规定哪个区域是骨干区域。

图 6-50 IS-IS 拓扑结构图之二



说明

IS-IS 的骨干网（Backbone）指的不是一个特定的区域。

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



## 接口的 Level 级别

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

## 路由渗透

通常情况下，区域内的路由通过 Level-1 的路由器进行管理。所有的 Level-2 和 Level-1-2 路由器构成一个连续的骨干区域。

Level-1 区域必须且只能与 Level-2 区域相连，不同的 Level-1 区域之间并不相连。

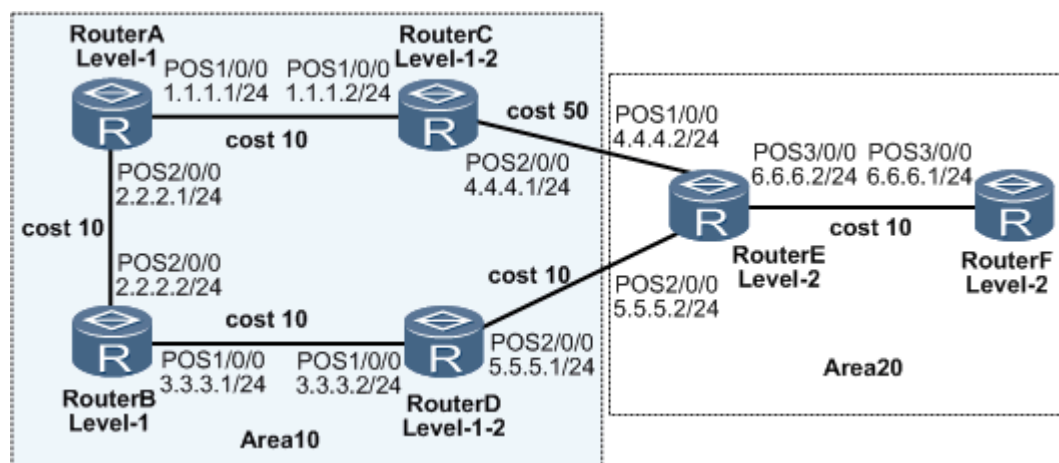
Level-1 区域内的路由信息通过 Level-1-2 路由器通报给 Level-2 区域，因此，Level-1-2 和 Level-2 路由器知道整个 IS-IS 路由域的路由信息。但是，在缺省情况下，Level-2 路由器并不将自己知道的其他 Level-1 区域以及骨干区域的路由信息通报给 Level-1 区域。这样，Level-1 路由器将不了解本区域以外的路由信息，可能导致对本区域之外的目的地无法选择最佳的路由。

为解决上述问题，IS-IS 提供了路由渗透功能，使 Level-1-2 路由器可以将已知的其他 Level-1 区域以及 Level-2 区域的路由信息通报给指定的 Level-1 区域。

如图 6-51 所示：

- 路由器 A、路由器 B、路由器 C 和路由器 D 同属于 Area10 区域，路由器 A 和路由器 B 为 Level-1 路由器，路由器 C 和路由器 D 为 Level-1-2 路由器。
- 路由器 E、路由器 F 同属于 Area20 区域，为 Level-2 路由器。

图 6-51 路由渗透示例



路由器 A 发送报文给路由器 F，选择的最佳路径应该是路由器 A—路由器 B—路由器 D—路由器 E—路由器 F。因为这条链路上的 cost 值为 10+10+10+10=40，但在路由器 A 上执行 **tracert** 命令，查看发送到路由器 F 的报文选择的路径是：路由器 A—路由器 C—路由器 E—路由器 F，其 cost 值为 10+50+10=70，不是路由器 A 到路由器 F 的最优路由。

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

此时分别在路由器 C 和路由器 D（Level-1-2）上执行 **import-route isis level-2 into level-1** 命令，启用路由渗透。

再执行 **tracert** 命令，发现路径是路由器 A—路由器 B—路由器 D—路由器 E—路由器 F，为路由器 A 到路由器 F 的最优路由。

#### 6.6.4.5 IS-IS 网络类型

IS-IS 只支持两种类型的网络，根据物理链路不同可分为：

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

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

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

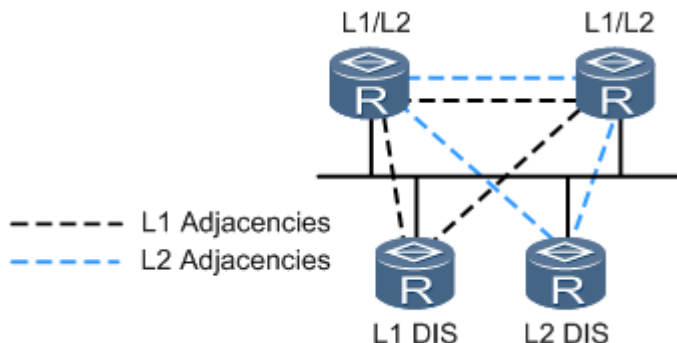
Level-1 和 Level-2 的 DIS 是分别选举的，用户可以为不同级别的 DIS 选举设置不同的优先级。DIS 优先级数值越高，被选中的可能性就越大。如果优先级最高的路由器有多台，则其中 MAC 地址最大的路由器会被选中。不同级别的 DIS 可以是同一台路由器，也可以是不同的路由器。

与 OSPF 的不同点：

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

在 IS-IS 广播网中，同一网段上的同一级别的路由器之间都会形成邻接关系，包括所有的非 DIS 路由器之间也会形成邻接关系，这一点与 OSPF 是不同的。如图 6-52 所示。

图 6-52 IS-IS 广播网的 DIS 和邻接关系



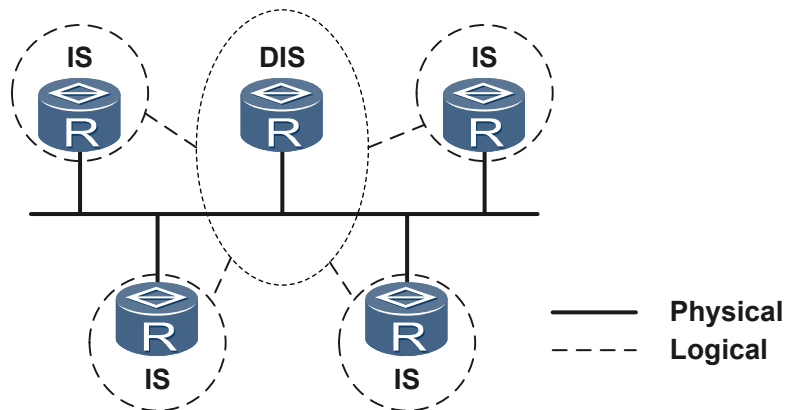
DIS 用来创建和更新伪节点（Pseudonodes），并负责生成伪节点的 LSP，用来描述这个网络上有哪些路由器。

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

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

伪节点示意图如图 6-53 所示。

图 6-53 伪节点示意图



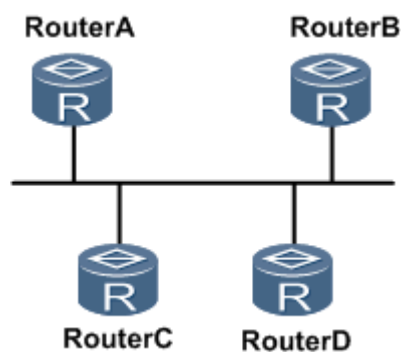
说明

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

#### 6.6.4.6 IS-IS 邻居建立过程

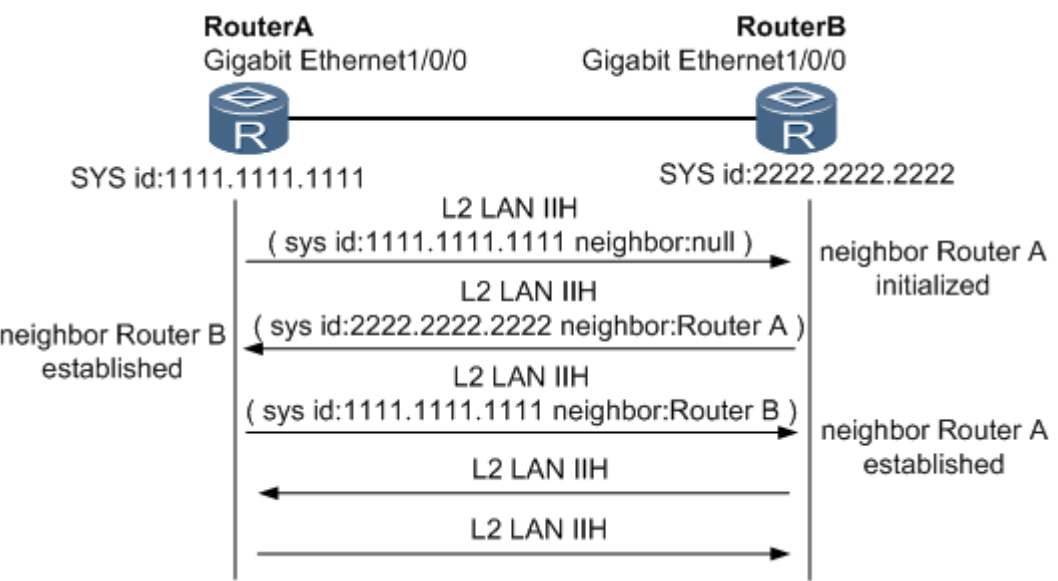
##### 广播链路的邻居建立

图 6-54 广播链路组网图



路由器 A、路由器 B、路由器 C 和路由器 D 都是 Level-2 路由器。路由器 A 新加入到此广播网络中。以下只列出路由器 A 和路由器 B 建立邻居的过程，与路由器 C 和路由器 D 建立邻居的过程相同。

图 6-55 广播链路邻居关系建立过程



路由器 A 广播发送 Level-2 LAN IS-IS Hello PDU，路由器 B 收到此报文后，将自己和路由器 A 的邻居状态标识为 Initial；然后，路由器 B 发送 L2 LAN IIH 报文给路由器 A，路由器 A 收到这个带有路由器 A 为路由器 B 邻居信息的 IIH 报文后，路由器 A 再将自己与路由器 B 的邻居状态标识为 Up。然后，路由器 A 发送 L2 LAN IIH 报文给路由器 B，路由器 B 收到这个带有路由器 B 为路由器 A 邻居信息的 IIH 报文后，将自己与路由器 A 的邻居状态标识为 Up。

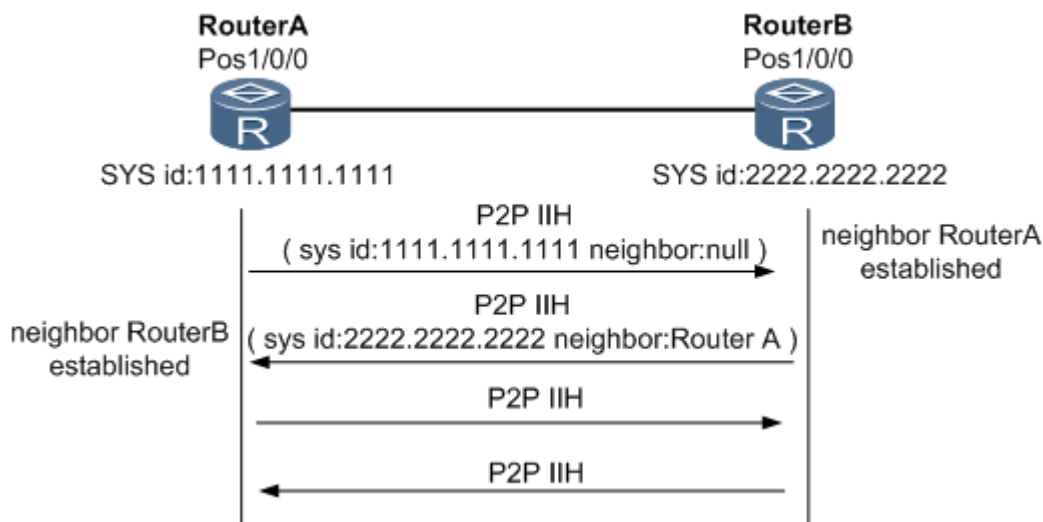
因为是广播网络，需要选举 DIS，所以在邻居关系建立后路由器会等待两个 Hello 报文间隔，以进行 DIS 的选举。IIH 报文中包含 Priority 字段，Priority 值最大的将被选举为该广播网的 DIS。若优先级相同，接口 MAC 地址较大的被选举为 DIS。

点到点链路邻居关系的建立

在点到点链路上，邻居关系的建立不同于广播链路。

在 USG 的实现中，使用 2-way 方式，即只要路由器收到 IS-IS Hello 报文，就会单方面建立起邻居关系。如图 6-56 所示。

图 6-56 点到点链路邻居关系建立过程



## IS-IS 邻居关系建立的基本原则

IS-IS 需满足以下条件才能建立邻居关系：

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

由于 IS-IS 是直接运行在数据链路层上的协议，并且最早设计是给 CLNP 使用的，IS-IS 邻居关系的形成与 IP 地址无关。但在 USG 的实现中，由于只在 IP 上运行 IS-IS，在没有配置 IP unnumber 时要检查对端 IP 地址。如果接口配置了从 IP，那么只要双方有某个 IP（主 IP 或者从 IP）在同一网段，就能建立邻居。在没有配置 IP 地址借用的情况下，如果对方的 IP 地址不和自己收到报文的接口 IP 地址在同一网段上，将不形成邻居关系，这样可以避免 IP 的不可达性。

如果接口配置了对接收的 Hello 报文不作 IP 地址检查，就可以建立邻居关系。对于 P2P 接口，可以通过配置 IP unnumber 启用接口忽略 IP 地址检查，接口将不检查 IP 地址是否在同一网段。但对于以太网接口，需要将以太网接口模拟成 P2P 接口，然后才可以配置接口忽略 IP 地址检查。

### 6.6.4.7 IS-IS 的 LSP 交互过程

#### LSP 的“泛洪”（flooding）

LSP 报文的“泛洪”指当一个路由器向相邻路由器报告自己的 LSP 后，相邻路由器再将同样的 LSP 报文传送到除发送该 LSP 的路由器外的其它邻居，并这样逐级将 LSP 传送到整个层次内的一种方式。通过这种“泛洪”，整个层次内的每一个路由器就都可以拥有相同的 LSP 信息，并保持 LSDB 的同步。每一个 LSP 都拥有其自己的一个 4 字节的序列号。在路由器启动时所发送的第一个 LSP 报文中的序列号为 1，以后当需要生成新的 LSP 时，新 LSP 的序列号在前一个 LSP 序列号的基础上加 1。更高的序列号意味着更新的 LSP。

每个 LSP 在 LSDB 中都有一个最大生存时间 (MaxAge)，当这个时间减为 0 之前如果没有接收到带有相同 LSP ID 和更大序列号的 LSP 来更新本地的 LSP，则这个 LSP 会从 LSDB 中清除。在旧的 LSP 被从 LSDB 中清除后，它还会再保留一段时间 (ZeroAgeLifetime，这期间只保留这个 LSP 的报文头，如果有认证字段的话也保留)，当这个时间也达到时它将会被真正删除。

一般缺省的 MaxAge 是 1200 秒，ZeroAgeLifetime 是 60 秒。当一个 IS 所发出的 LSP 的序列号达到 0xFFFFFFFF 时，这个路由器会将 IS-IS 进程暂停 MaxAge + ZeroAgeLifetime 秒，以便在整个路由域内将与这个路由器对应的 LSP 都被删除掉。随后 IS-IS 进程再重新启动，并从序列号 1 开始发送 LSP。这种情况其实不会出现，因为即使每秒钟产生一个 LSP，也要到 136 年后序列号才能从 1 达到 0xFFFFFFFF。

## LSP 产生或更新的原因

IS-IS 路由域内的所有路由器都会产生 LSP，以下事件会触发 LSP 的产生或更新：

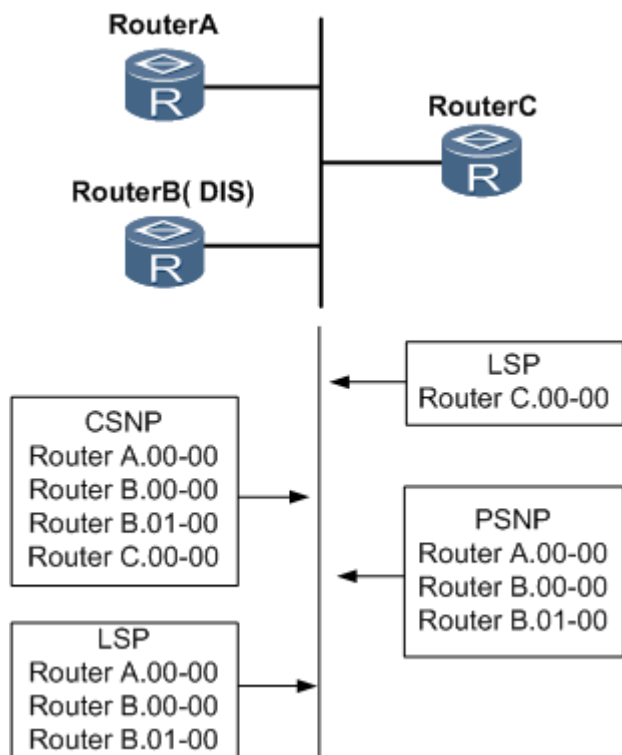
- 邻居 Up 或 Down。
- IS-IS 相关接口 Up 或 Down。
- 引入的 IP 路由发生变化。
- 区域间的 IP 路由发生变化。
- 接口被赋了新的 metric 值。
- 周期性更新。

## 收到邻居新的 LSP 的处理过程

- 将新的 LSP 安装到自己的 LSDB 数据库中标记为 flooding。
- 发送新的 LSP 到除了发送该 LSP 的邻居之外的邻居。
- 邻居再扩散到其它邻居。

## 新加入路由器与 DIS 同步 LSDB 数据库过程

图 6-57 广播链路数据库更新过程



1. 新加入的路由器 RouterC 首先发送 Hello 报文，与该广播域中的路由器建立邻居关系。
2. 邻居关系建立起来后，RouterC 等待 LSP 定时器超时然后将自己的 LSP 发送往组播地址：
  - Level-1: 01-80-C2-00-00-14
  - Level-2: 01-80-C2-00-00-15即网络上所有的邻居都将收到该 LSP。
3. 该网段中的 DIS 会把收到 RouterC 的 LSP 加入到 LSDB 中，并等待 CSNP 报文定时器超时并发送 CSNP 报文，进行该网络内的 LSDB 同步。CSNP 报文的发送间隔缺省值为 10 秒。
4. RouterC 收到 DIS 发来的 CSNP 报文，对比自己的 LSDB 数据库，发送 PSNP 报文请求自己没有的 LSP。
5. DIS 收到该 PSNP 报文请求后发送对应的 LSP 进行 LSDB 的同步。

## DIS 的 LSDB 更新过程

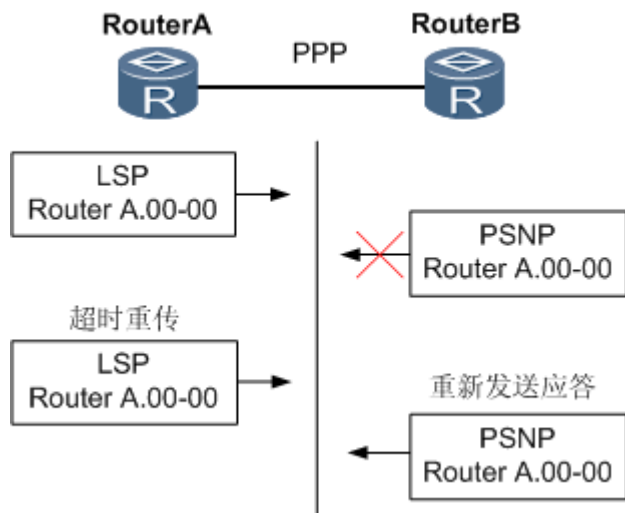
- DIS 接收到 LSP，在数据库中搜索对应的记录。若没有该 LSP，则将其加入数据库，并广播新数据库内容。
- 若收到的 LSP 序列号大于本地 LSP 的序列号，就替换为新报文，并广播新数据库内容。
- 若收到的 LSP 序列号小本地 LSP 的序列号，就向入端接口发送本地 LSP 报文。



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

## 点到点链路上数据库的同步过程

图 6-58 点到点链路数据库更新过程



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

 说明

在点到点 (P2P) 链路上 PSNP 有两种作用, 一种是作为 Ack 应答以确认收到的 LSP, 另一种是用来请求所需的 LSP。

## LSDB 更新的过程

- 若收到的 LSP 比本地的序列号更大，则将这个新的 LSP 存入自己的 LSDB，再通过一个 PSNP 报文来确认收到此 LSP，最后再将这个新 LSP 发送给除了发送该 LSP 的邻居以外的邻居。

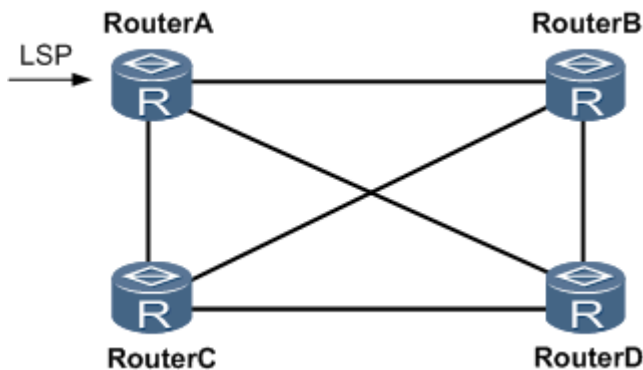


- 若收到的 LSP 比本地的序列号更小，则直接给对方发送本地的 LSP，然后等待对方给自己一个 PSNP 报文作为确认。
- 若收到的 LSP 序列号和本地相同，则比较 Remaining Lifetime，若收到 LSP 的 Remaining Lifetime 小于本地 LSP 的 Remaining Lifetime，则将收到的 LSP 存入 LSDB 中并发送 PSNP 报文来确认收到此 LSP，然后将该 LSP 发送给除了发送该 LSP 的邻居以外的邻居。
- 若收到的 LSP 序列号和本地相同，则比较 Remaining Lifetime，若收到 LSP 的 Remaining Lifetime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

## Mesh Group

如图 6-59 所示，在 NBMA 网络中运行 IS-IS 协议，需要将子接口配置为点到点类型。即图中 RouterA、RouterB、RouterC 和 RouterD 之间相连接的链路都配置为 P2P 类型。

图 6-59 Mesh Group



当 RouterA 收到一个新的 LSP 时，按照点到点链路的数据库更新过程，RouterA 会将该 LSP 分别发往 RouterB、RouterC 和 RouterD。同样，对于 RouterB、RouterC 和 RouterD，收到该 LSP 时也会将其再发送出去，这种处理方式会造成 LSP 重复扩散，导致带宽的浪费。

为避免这种情况的发生，可以将一些接口组成 Mesh Group，一个组中的接口不把从本组接口接收的 LSP 扩散到同组中的其它接口，而只扩散到其它组的接口以及没有配置 Mesh Group 的接口。所有加入到 Mesh Group 中的接口，通过 CSNP 和 PSNP 机制来保证整个网段的 LSDB 的同步。

### 6.6.4.8 IS-IS 对 IPv6 的支持

IETF 的 draft-ietf-isis-ipv6-05.txt 中规定了 IS-IS 为支持 IPv6 所新增的内容。主要是新添加的支持 IPv6 路由信息的两个 TLVs（Type-Length-Values）和一个新的 NLPID（Network Layer Protocol Identifier）。IS-IS 与 RIP 和 OSPF 不同，这两个协议有单独的版本 RIPng 和 OSPFv3 支持 IPv6。

TLV 是在 LSP（Link State PDUs）中的一个可变长字段值。新增的两个 TLV 分别是：

- IPv6 Reachability:

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

- IPv6 Interface Address:

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

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

### 6.6.4.9 多进程和多实例

为了方便管理，提高控制效率，IS-IS 支持多进程和多实例特性。多进程允许为一个指定的 IS-IS 进程关联一组接口，从而保证该进程进行的所有协议操作都仅限于这一组接口。这样，就可以实现一台路由器有多个 IS-IS 协议进程，每个进程负责唯一的一组接口。

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

## 6.7 BGP

BGP 是用于自治系统之间的外部网关协议，广泛用于 ISP 之间。

### 6.7.1 介绍

#### 定义

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

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

当前使用的版本是 BGP-4（RFC4271）。

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

BGP 协议具有如下特点：

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

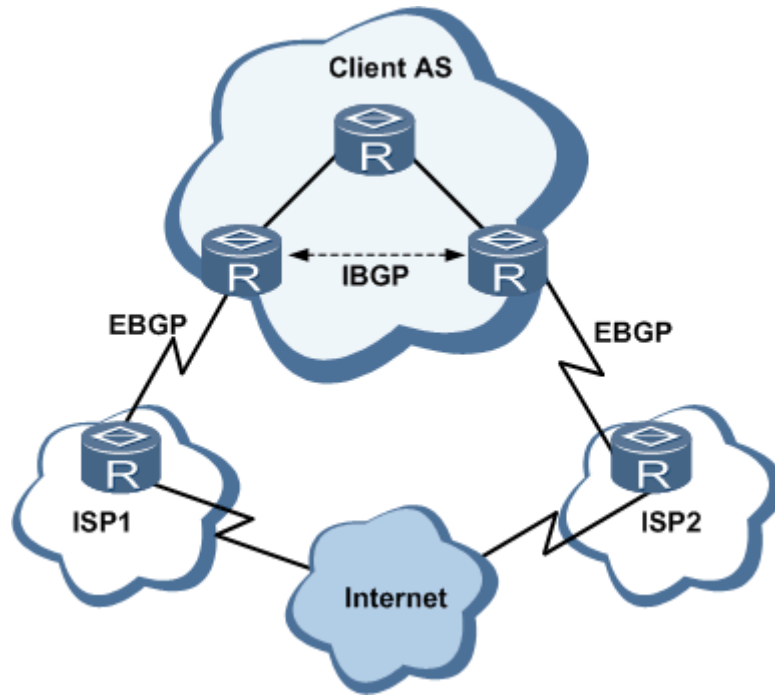
## 目的

BGP 用于在 AS 之间传递路由信息，并不是所有情况都需要运行 BGP。

以下情况中需要使用 BGP 协议：

- 如 [图 6-60](#)，用户需要同时与两个或者多个 ISP 相连，ISP 需要向用户提供部分或完全的 Internet 路由。这时可以通过 BGP 路由携带的 AS 信息来决定到达目的地，走哪一个 ISP 的 AS 更为经济。

图 6-60 BGP 的应用场景



- 不同 AS 间需要传递 AS 路径信息。
- 用户需要通过三层 VPN 传播私网路由，请参见《特性描述-MPLS&VPN 部分》。
- 用户需要在二层应用中（如 komPELLa 方式的 VPLS）以 BGP 为信令传播二层信息，请参见《特性描述-MPLS 与 VPN》。
- 用户需要传播组播路由构造组播拓扑，请参见《特性描述-IP 组播》。

以下情况不需要使用 BGP 协议：

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

## 6.7.2 参考标准和协议

与 BGP 相关的参考标准与协议如下：

- RFC 1771: A Border Gateway Protocol 4 (BGP-4)
- RFC 1997: BGP Communities Attribute
- RFC 1998: An Application of the BGP Community Attribute in Multi-home Routing
- RFC 2796: BGP Route Reflection - An Alternative to Full Mesh IBGP
- RFC 1965: Autonomous System Confederations for BGP
- RFC 2439: BGP Route Flap Damping
- RFC 2918: Route Refresh Capability for BGP-4

### 6.7.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 6.7.4 原理描述

#### 6.7.4.1 BGP 协议基本原理

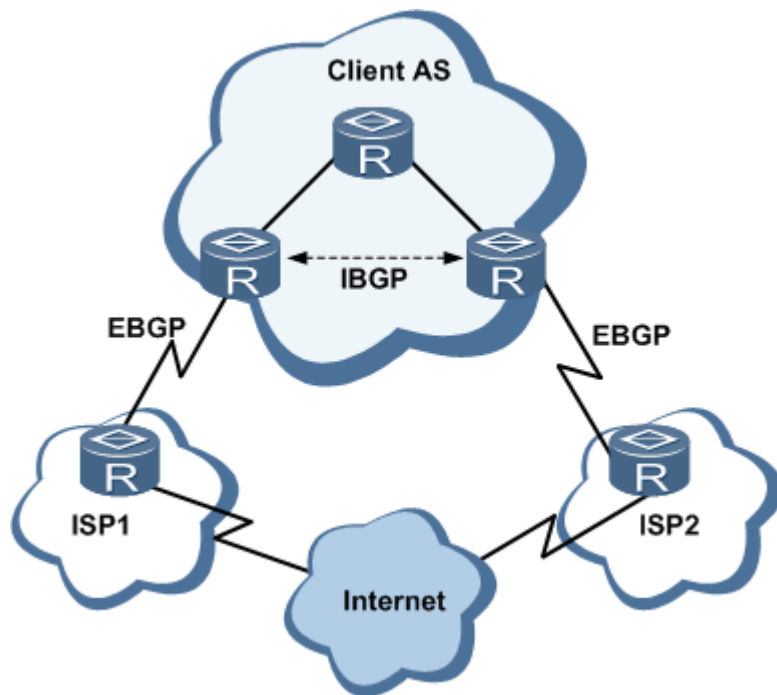
#### BGP 运行方式

BGP 在路由器上以下列两种方式运行，如[图 6-61](#) 所示：

- IBGP（Internal BGP）
- EBGp（External BGP）

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

图 6-61 BGP 的运行方式



## BGP 报文中的角色

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

## BGP 的报文

BGP 的运行是通过报文驱动的，共有 Open、Update、Notification、Keepalive 和 Route-refresh 等 5 种报文类型。

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

- 一条 Update 报文可以撤销多条不可达路由。每一个路由通过目的地（用 IP 前缀表示），清楚的定义了 BGP Speaker 之间先前通告过的路由。
- 一条 Update 报文可以只用于撤销路由，这样就不需要包括路径属性或者 NLRI。相反，也可以只用于通告可达路由，就不需要携带撤销路由信息了。
- Notification 报文：当 BGP 检测到错误状态时，就向对等体发出 Notification 报文，之后 BGP 连接会立即中断。
- Keepalive 报文：BGP 会周期性的向对等体发出 Keepalive 报文，用来保持连接的有效性。
- Route-refresh 报文：Route-refresh 报文用来通知对等体自己支持路由刷新能力（Route-refresh capability）。

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

## BGP 有限状态机

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

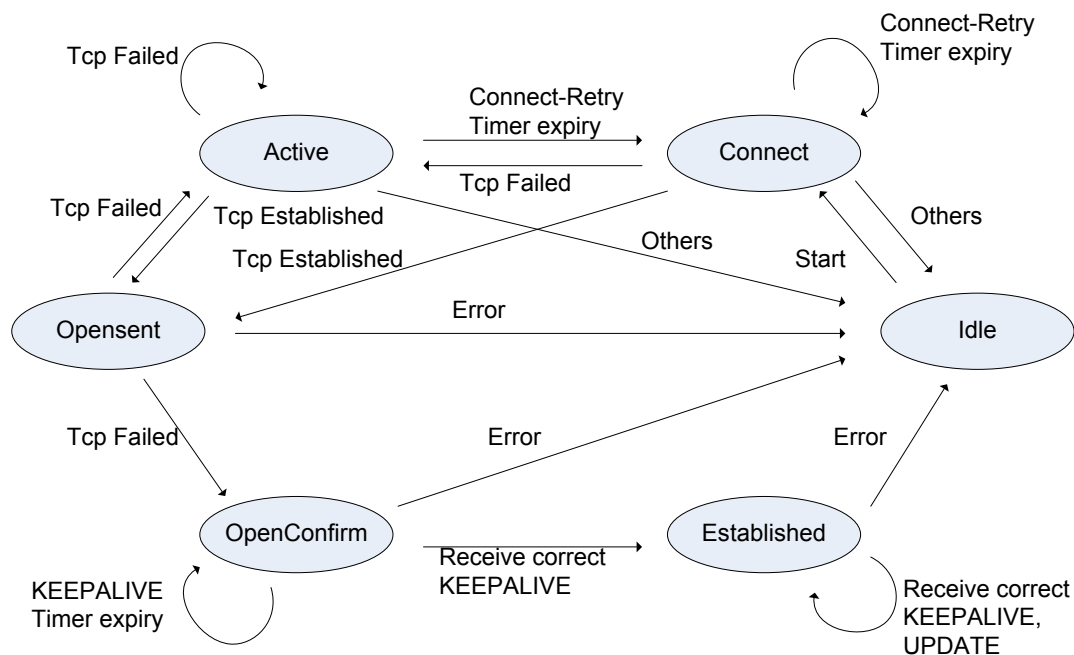
在 BGP 对等体建立的过程中，通常可见的三个状态是：Idle、Active、Established。

- Idle 状态下，BGP 拒绝任何进入的连接请求，是 BGP 初始状态。
- Active 状态下，BGP 将尝试进行 TCP 连接的建立，是 BGP 的中间状态。
- Established 状态下，BGP 对等体间可以交换 Update 报文、Route-refresh 报文、Keepalive 报文和 Notification 报文。

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

BGP 有限状态机详细转换过程如图 6-62 所示。

图 6-62 BGP 有限状态机



## BGP 处理过程

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

## BGP 属性

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

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

下面介绍几种常用的 BGP 路由属性：

- Origin 属性

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

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

- AS\_Path 属性

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

当 BGP Speaker 本地通告一条路由时：

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

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

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



- 当 BGP Speaker 将这条路由通告到本地 AS 时，不会改变这条路由相关的 AS\_Path 属性。
- Next\_Hop 属性

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

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

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

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

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

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

## BGP 选择路由的策略

当到达同一目的地存在多条路由时，BGP 采取如下策略进行路由选择：

1. BGP 优选 Preference 值较低的本地始发路由。这个优先级是 IP 路由表（管理路由表）中各个协议路由（包括直连路由和静态路由）的 Preference 值，Preference 值越小，优先级越高。
2. 如果不同协议路由的 Preference 值相同，系统会按照如下的顺序优选一种协议路由：OSPF、IS-IS Level-1、IS-IS Level-2、EBGP（包含 BGP Aggregate）、Static、RIP、OSPF\_ASE、IBGP。
3. 首先丢弃下一跳 (Next\_Hop) 不可达的路由。
4. 无条件优选带 Label 的 IPv4 路由。
5. 优选协议首选值 (PrefVal) 最高的路由。
6. 优选本地优先级 (Local\_Pref) 最高的路由。
7. 优选聚合路由（聚合路由优先级高于非聚合路由）。
8. 本地手动聚合路由的优先级高于本地自动聚合的路由。
9. 本地通过 **network** 命令引入的路由的优先级高于本地通过 **import-route** 命令引入的路由。
10. 优选 AS 路径 (AS\_Path) 最短的路由。
11. 比较 Origin 属性，依次选择 Origin 类型为 IGP、EGP、Incomplete 的路由。

12. 优选 MED 值最低的路由。
13. 优选从 EBGp 学来的路由（EBGP 路由优先级高于 IBGP 路由）。
14. 优选 AS 内部 IGP 的 Metric 最低的路由。如果配置了负载分担，并且有多条 As\_Path 完全相同的外部路由，则根据配置的路由条数选择多条路由进行负载分担。



说明

只有以上条件都没有比较出更优的路由，才会负载均衡。

15. 优选 Cluster\_List 最短的路由。
16. 优选 Originator\_ID 最小的路由。
17. 优选 Router ID 最小的路由器发布的路由。
18. 比较对等体的 IP Address，优选从具有较小 IP Address 的对等体学来的路由。

## 应用 BGP 负载分担时的选路策略

在 BGP 中，由于协议本身的特殊性，它产生的路由的下一跳地址可能不是当前路由器直接相连的邻居。常见的一个原因是，IBGP 之间发布路由信息时不改变下一跳。这种情况下，为了能够将报文正确转发出去，路由器必须先找到一个直接可达的地址，通过这个地址到达路由表中指示的下一跳。在上述过程中，去往直接可达地址的路由被称为依赖路由，BGP 路由依赖于这些路由指导报文转发。根据下一跳地址找到依赖路由的过程就是路由迭代。

USG 支持基于迭代的 BGP 负载分担，即，如果依赖路由本身是负载分担的（假设有三个下一跳地址），则 BGP 也会生成相同数量的下一跳地址来指导报文转发。需要说明的是，基于迭代的 BGP 负载分担并不需要命令配置，这一特性在 USG 上始终启用。



说明

- BGP 只对 AS\_Path 属性完全相同的路由进行负载分担。
- BGP 负载分担特性同样适用于联盟内部的自治系统之间。

BGP 的负载分担与 IGP 的负载分担在实现方法上有所不同：

- 在 IGP 中，对到达同一目的地址的不同路由，IGP 根据本身的路由算法计算路由的度量值（metric），在度量值相等的路由间进行负载分担。
- BGP 由于本身并没有路由算法，不能根据一个明确的度量值决定是否对路由进行负载分担。但 BGP 有很多路由属性，这些属性在 BGP 选路策略中的优先顺序是不同的。对 BGP 负载分担的处理规则是加入到这些选路策略中的，即在所有高优先级路由属性相同的情况下，BGP 根据所配置的最大负载分担的路由条数进行负载分担。

## BGP 发布路由的策略

BGP 发布路由时采用如下策略：

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

6.7.4.2 BGP 报文格式

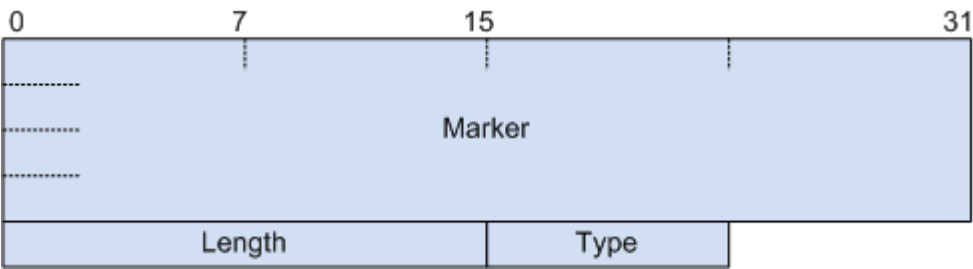
报文中的角色

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

报文头格式

BGP 的运行是通过报文驱动的，共有 5 种报文类型。这些报文通过 TCP 协议进行传播 (端口号是 179)。报文最长为 4096 字节，最短为 19 字节 (只包含报文头)。这些报文有相同的报文头，其格式如图 6-63 所示。

图 6-63 BGP 报文的报文头格式



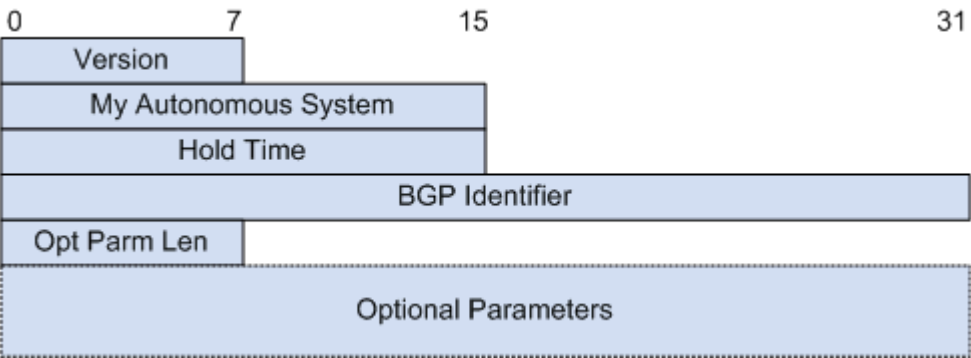
主要字段的解释如下：

- **Marker:** 用于检查 BGP 对等体的同步信息是否完整，以及用于 BGP 验证的计算，共 16 字节。不使用验证时所有比特均为 1。
- **Length:** BGP 报文总长度 (包括报文头在内)，以字节为单位。长度范围是 19 ~ 4096。
- **Type:** BGP 报文的类型。其取值从 1 到 5，分别表示 Open、Update、Notification、Keepalive 和 Route-refresh 报文。其中，前四种报文是在 RFC4271 中定义的，而 Type5 的报文则是在 RFC2918 中定义的。

Open

Open 报文是 TCP 连接建立后发送的第一个报文，用于建立 BGP 对等体之间的连接关系。其报文格式如图 6-64 所示。

图 6-64 BGP Open 报文格式



主要字段的解释如下：

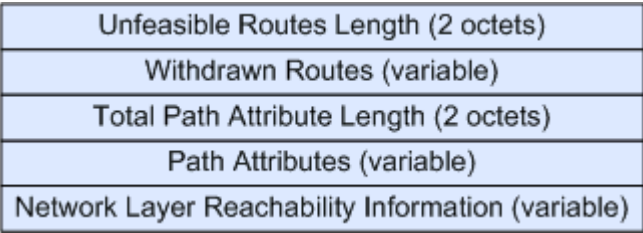
- **Version:** BGP 的版本号。对于 BGP-4 来说，其值为 4。
- **My Autonomous System:** 本地 AS 编号。通过比较两端的 AS 编号可以确定是 EBGp 连接还是 IBGP 连接。
- **Hold Time:** 在建立对等体关系时两端要协商 Hold time，并保持一致。如果两端所配置的 Hold time 时间不同，则 BGP 会选择较小的值作为协商的结果。如果在这个时间内未收到对端发来的 Keepalive 报文或 Update 报文，则认为 BGP 连接中断。
- **BGP Identifier:** BGP 路由器的 Router ID，以 IP 地址的形式表示，用来识别 BGP 路由器。在 USG 系统中，如果没有通过命令 **router id** 进行配置，则按照如下规则进行选择：优选 Loopback 接口地址中最大的地址作为 Router ID，如果没有配置了 IP 地址的 Loopback 接口，则从其它接口的 IP 地址中选择一个最大的作为 Router ID。
- **Opt Parm Len (Optional Parameters Length):** 可选参数的长度。如果为 0 则没有可选参数。
- **Optional Parameters:** 用于 BGP 验证或多协议扩展 (Multiprotocol Extensions) 等功能的可选参数。每一个参数为一个 (Parameter Type-Parameter Length-Parameter Value) 三元组。

对等体在接收到 Open 报文后，将发送 Keepalive 报文确认并保持连接的有效性。确认后，对等体间可以进行 Update、Notification、Keepalive 和 Route-refresh 报文的交换。

Update

Update 报文用于在对等体之间交换路由信息。Update 报文可以发布多条属性相同的可达路由信息，也可以撤销多条不可达路由信息。同样，Update 报文中包含固定的 BGP 报文头，如图 6-63 所示。报文中其他可选字段格式如图 6-65 所示。

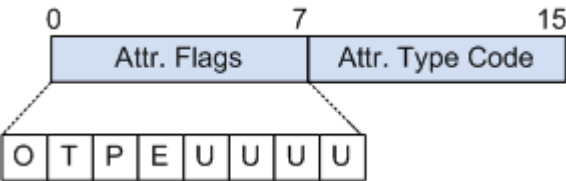
图 6-65 BGP Update 报文格式



主要字段的解释如下：

- **Unfeasible Routes Length:** 不可达路由字段的长度，单位为字节。如果为 0 则说明没有 Withdrawn Routes 字段。
- **Withdrawn Routes:** 不可达路由的列表。每一个列表项包含一个 IP 地址前缀（Length-Prefix）二元组。
  - **Length** 表示 IP 地址前缀的长度，单位为比特。如果为 0，则说明该前缀匹配所有的 IP 地址。
  - **Prefix** 包含了被撤销的 IP 地址的前缀，后面则是填充比特，用于保证字段结尾符合字节边界，填充比特的值无意义。
- **Total Path Attribute Length:** 路径属性字段的长度，单位为字节。如果为 0 则说明没有网络可达信息，即 Path Attributes 和 NLRI（Network Layer Reachability Information）字段。
- **Path Attributes:** 与 NLRI 相关的所有路径属性列表，每个路径属性由一个 TLV（Type, Length, Value）三元组构成，该三元组的各字段解释如下：
  - **Type:** 属性类型（2 字节）包括了属性标志字节（Attribute Flags）和属性类型码字节（Attribute Type Code）。

图 6-66 TLV 结构—Type



属性标志字段由高位比特到低位比特的意义如表 6-34。

表 6-34 属性类型各比特意义

| 符号 | 名称             | 描述                                                                                                         |
|----|----------------|------------------------------------------------------------------------------------------------------------|
| O  | Optional bit   | 可选比特，定义属性是否为可选： <ul style="list-style-type: none"><li>● 0：公认</li><li>● 1：可选</li></ul>                      |
| T  | Transitive bit | 转发比特，定义属性是否是转发的： <ul style="list-style-type: none"><li>● 0：不转发</li><li>● 1：转发</li></ul> 对于公认属性，T 比特必须设为 1。 |

| 符号 | 名称                  | 描述                                                                                                                                                           |
|----|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|
| P  | Partial bit         | 不完整比特，定义包含在可选过渡属性内的信息是否完整：<br><ul style="list-style-type: none"> <li>● 0：完整</li> <li>● 1：不完整</li> </ul> 对于公认属性和可选非过渡属性，P 比特必须设为 0。                           |
| E  | Extended Length bit | 扩展长度比特，定义是否扩展属性的长度：<br><ul style="list-style-type: none"> <li>● 0：表示不扩展属性长度，属性长度是 1 字节。</li> <li>● 1：表示扩展属性长度，属性长度是 2 字节。</li> </ul> 当属性值超过 255 字节时需要扩展属性长度。 |
| U  | Unused bit          | 该比特没有被使用，发送时填充为 0（接收时不处理）。                                                                                                                                   |

- Length: 路径属性长度。  
Length 在 Path Attributes 中所占的字节数取决于 Type 的扩展长度（E）比特。  
如果 E 比特值为 0，则 Path Attributes 的第 3 字节是 Length。  
如果 E 比特值为 1，则 Path Attributes 的第 3 和第 4 字节是 Length。
- Value: 路径属性值。  
USG 所常用属性类型码和他们的属性值如表 6-35。

表 6-35 路径属性类型码和属性值

| 属性类型                | 属性值                                                                                                                                      |
|---------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| 1: Origin           | <ul style="list-style-type: none"> <li>● IGP</li> <li>● EGP</li> <li>● Incomplete</li> </ul>                                             |
| 2: As_Path          | <ul style="list-style-type: none"> <li>● AS_SET</li> <li>● AS_SEQUENCE</li> <li>● AS_CONFED_SET</li> <li>● AS_CONFED_SEQUENCE</li> </ul> |
| 3: Next_Hop         | 下一跳的 IP 地址                                                                                                                               |
| 4: Multi_Exit_Disc  | MED 用于判断流量进入 AS 时的最佳路由                                                                                                                   |
| 5: Local_Pref       | Local_Pref 用于判断流量离开 AS 时的最佳路由                                                                                                            |
| 6: Atomic_Aggregate | BGP Speaker 选择聚合后的路由，而非具体的路由                                                                                                             |
| 7: Aggregator       | 发起聚合的路由器 ID 和 AS 号                                                                                                                       |
| 8: Community        | 团体属性                                                                                                                                     |
| 9: Originator_ID    | 反射路由发起者的 Router ID                                                                                                                       |
| 10: Cluster_List    | 反射路由经过的反射器列表                                                                                                                             |

| 属性类型                     | 属性值         |
|--------------------------|-------------|
| 14: MP_REACH_NLRI        | 多协议可达 NLRI  |
| 15: MP_UNREACH_NLRI      | 多协议不可达 NLRI |
| 16: Extended Communities | 扩展团体属性      |

- NLRI（Network Layer Reachability Information）： 由一个或多个二元组<前缀长度，可达路由的前缀>组成。

一条 Update 报文可以发布多条具有相同路由属性的可达路由，这些路由可共享一组路由属性。所有包含在一个给定的 Update 报文里的路由属性适用于该 Update 报文中的 NLRI 字段里的所有目的地（用 IP 前缀表示）。

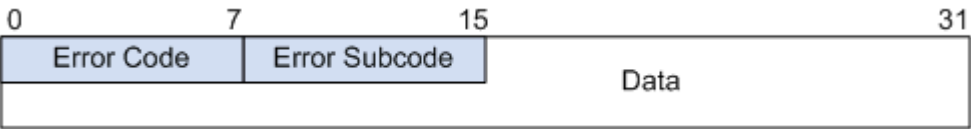
一条 Update 报文可以撤销多条不可达路由。每一个路由通过目的地（用 IP 前缀表示），清楚的定义了 BGP Speaker 之间先前通告过的路由。

一条 Update 报文可以只用于撤销路由，这样就不需要包括路径属性或者网络可达信息。相反，也可以只用于通告可达路由，就不需要携带 Withdrawn Routes 了。

Notification

当 BGP 检测到错误状态时，就向对等体发出 Notification 报文，之后 BGP 连接会立即中断。其报文格式如 图 6-67 所示。

图 6-67 BGP Notification 报文格式



主要字段的解释如下：

- Error Code： 错误码，指定错误类型。
- Error Subcode： 错误子码，描述错误的详细信息。
- Data： 用来诊断错误原因，长度不固定。

表 6-36 Notification 报文错误码和错误子码

| 错误码      | 错误子码                                                                                                 |
|----------|------------------------------------------------------------------------------------------------------|
| 1: 报文头错误 | <ul style="list-style-type: none"><li>● 1: 连接未同步</li><li>● 2: 错误的报文长度</li><li>● 3: 错误的报文类型</li></ul> |

| 错误码              | 错误子码                                                                                                                                                                                                                                                                                        |
|------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 2: Open 报文错误     | <ul style="list-style-type: none"><li>● 1: 不支持的版本号</li><li>● 2: 错误的对等 AS</li><li>● 3: 错误的 BGP 标识符</li><li>● 4: 不支持的可选参数</li><li>● 5: 认证失败</li><li>● 6: 不可接受的保持时间</li><li>● 7: 不支持的能力</li></ul>                                                                                              |
| 3: Update 报文错误   | <ul style="list-style-type: none"><li>● 1: 畸形属性列表</li><li>● 2: 不可识别的公认属性</li><li>● 3: 缺少公认属性</li><li>● 4: 属性标志错误</li><li>● 5: 属性长度错误</li><li>● 6: 无效 Origin 属性</li><li>● 7: AS 路由环路</li><li>● 8: 无效 Next_Hop 属性</li><li>● 9: 可选属性错误</li><li>● 10: 无效网络字段</li><li>● 11: 畸形 AS_Path</li></ul> |
| 4: Hold Timer 溢出 | 0: 没有特别的错误子码定义。                                                                                                                                                                                                                                                                             |
| 5: 有限状态机错误       | 0: 没有特别的错误子码定义。                                                                                                                                                                                                                                                                             |
| 6: 终止            | <ul style="list-style-type: none"><li>● 1: 前缀超过最大值。</li><li>● 2: 管理关闭</li><li>● 3: 删除邻居</li><li>● 4: 管理重置</li><li>● 5: 连接失败</li><li>● 6: 其他配置改变</li><li>● 7: 连接冲突</li><li>● 8: 资源短缺</li></ul>                                                                                               |

Keepalive

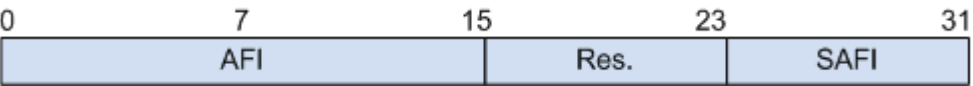
BGP 会周期性的向对等体发出 Keepalive 报文，用来保持连接的有效性。其报文格式中只包含报文头，没有附加其他任何字段。

Route-refresh

Route-refresh 报文用来通知对等体自己支持路由刷新能力。其报文格式如图 6-68 所示。



图 6-68 BGP Route-refresh 报文格式



主要字段的解释如下：

- AFI（Address Family Identifier）：地址族标识符（16bit）。
- Res.（Reserved field）：保留区域（8bit），发送方应将其设置为 0，接收方应当忽略该区域的信息。
- SAFI（Subsequent Address Family Identifier）：子地址族标识符（8bit）。

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

6.7.4.3 BGP 属性分类

BGP 属性分类

BGP 路由属性是一套参数，它对特定的路由进一步的描述，使得 BGP 能够对路由进行过滤和选择。

事实上，所有的 BGP 路由属性都可以分为以下 4 类：

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

BGP 路由属性和对应的类别如表 6-37 所示。

表 6-37 路由属性和类别

| 属性名称             | 类别                       |
|------------------|--------------------------|
| Origin           | Well-known mandatory     |
| AS_Path          | Well-known mandatory     |
| Next_Hop         | Well-known mandatory     |
| Local_Pref       | Well-known discretionary |
| Atomic_Aggregate | Well-known discretionary |

| 属性名称                 | 类别                      |
|----------------------|-------------------------|
| Aggregator           | Optional transitive     |
| Community            | Optional transitive     |
| Multi_Exit_Disc(MED) | Optional non-transitive |
| Originator_ID        | Optional non-transitive |
| Cluster_List         | Optional non-transitive |

## Origin 属性

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

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

## AS\_Path 属性含义

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

- 当 BGP Speaker 本地通告一条路由时：
  - 当 BGP Speaker 将这条路由通告到其他 AS 时，便会将本地 AS 编号添加在 AS\_Path 列表中，并通过 Update 消息通告给邻居路由器。
  - 当 BGP Speaker 将这条路由通告到本地 AS 时，便会在 Update 消息中创建一个空的 AS\_Path 列表。
- 当 BGP Speaker 传播从其他 BGP Speaker 的 Update 消息中学习到的路由时：
  - 当 BGP Speaker 将这条路由通告到其他 AS 时，便会把本地 AS 编号添加在 AS\_Path 列表的最前面（最左面）。收到此路由的 BGP 路由器根据 AS\_Path 属性就可以知道去目的地址所要经过的 AS。离本地 AS 最近的相邻 AS 编号排在前面，其他 AS 编号按顺序依次排列，如 [图 6-69](#) 所示。
  - 当 BGP Speaker 将这条路由通告到本地 AS 时，不会改变这条路由相关的 AS\_Path 属性。

综上所述，IBGP 在向对等体通告路由时，并不改变 AS\_Path 属性。只有在 EBGP 对等体间通告路由时，AS 编号才会被添加到 AS\_Path 列表中。

图 6-69 AS\_Path 属性

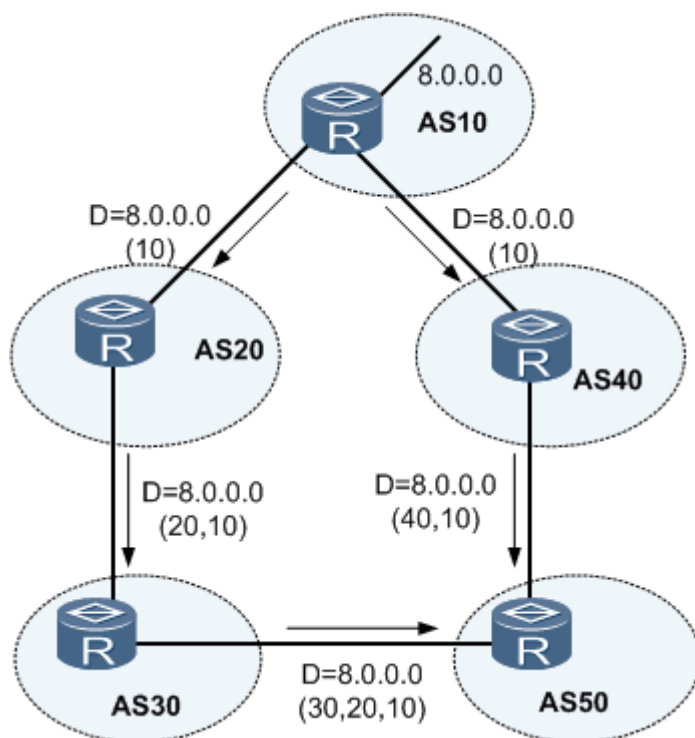


图 6-69 中，BGP 要将自治系统 AS10 的路由 D=8.0.0.0 通告到其他 AS 时，首先把 AS10 的 AS 编号 10 放到 AS\_Path 列表中（10）；在经过 AS20 后，AS20 又将自己的 AS 编号 20 放到 AS\_Path 列表的最左边（20，10），以此类推。当 AS50 的路由器收到这条路由后，就可以根据 AS\_Path 列表知道要去 AS10 可以有两条路径供选择，它们分别是经过 AS30，AS20 的路径和经过 AS40 的路径。

## AS\_Path 属性的作用

- 防止路由环路

AS\_Path 属性可以防止路由环路。通常情况下，BGP 不会接收 AS\_Path 中已包含本地 AS 编号的路由，从而避免了形成路由环路的可能。

在 Hub&Spoke 组网方案中，需要允许 AS 编号重复。详细内容请参见《特性描述-VPN》。

- 控制路由选择

AS\_Path 属性可用于控制路由的选择。在其他因素相同的情况下，BGP 会优先选择 AS 路径较短的路由。比如在图 6-69 中，AS50 中的 BGP 路由器会选择从 AS40 收到的路由作为到目的地 8.0.0.0 去的最优路由。

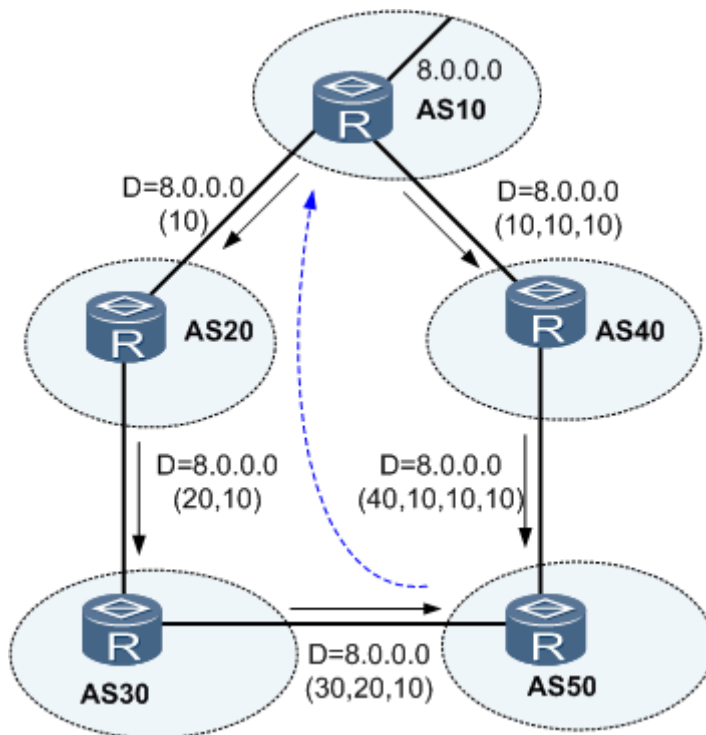
在某些应用中，可以使用路由策略人为的增加 AS 路径的长度，以便更为灵活的控制 BGP 路径的选择。比如在图 6-69，如果到目的地 8.0.0.0，经过 AS30 的路径上带宽较大，而经过 AS40 的路径上带宽较小；或者经过 AS30 的路径是主链路，经过 AS40 的路径是备份链路。这些情况下，AS10 可以通过改变其通告路由的 AS\_Path 属性，来控制 BGP 选择实际中最优的路径。

如图 6-70，AS10 在其发送给 AS40 的 AS\_Path 列表中附加多个本地 AS 编号，使 AS50 认为到达 8.0.0.0 最短的路径是（30，20，10）。

🔑 窍门

在 USG 的实现中，通过配置 `apply as-path` 命令可以附加 AS 编号。

图 6-70 通过增加 AS 路径长度控制路由选择



- 路由过滤

通过配置 AS 路径过滤器，可以针对 AS\_Path 属性中所包含的 AS 编号对路由进行过滤。如图 6-69，到目的地 8.0.0.0，通过配置 AS\_Path 过滤器（例如，只接收 AS\_Path 列表中包含“30”的路由），可以使 AS50 选择经过 AS30 的较长路径，而不选择经过 AS40 的较短路径。

## AS\_Path 属性分类

前面提到 AS\_Path 属性按一定次序记录了某条路由从本地到目的地址所要经过的所有 AS 编号。实际上 AS\_Path 属性包含四种类型：

- As\_Sequence: AS\_Path 有序集。
- As\_Set: AS\_Path 无序集，应用在路由聚合中。
- As\_Confed\_Sequence
- As\_Confed\_Set

📖 说明

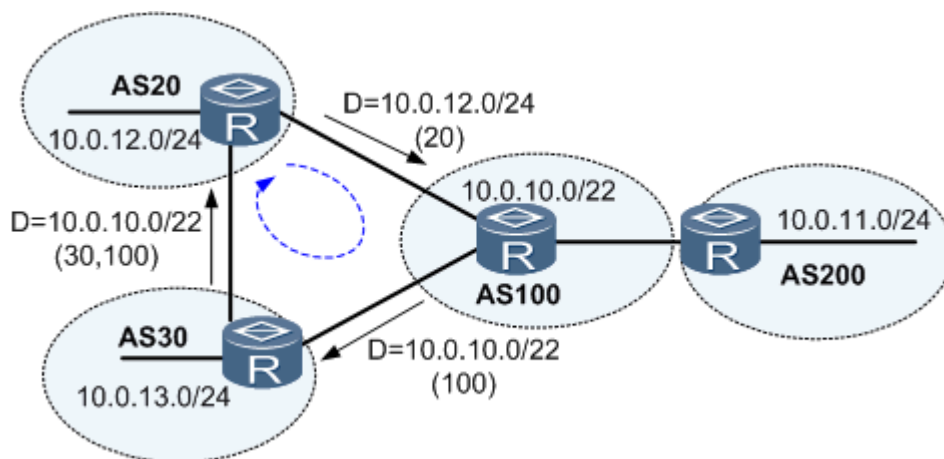
As\_Confed\_Sequence 和 As\_Confed\_Set 在联盟中使用，详细内容请参见 6.7.4.8 BGP 联盟。

AS\_Path 属性的一个主要作用是防止路由环路产生，如果 BGP 路由器从来自对等体的路由中发现了自己的 AS 号，就会放弃这条路由以避免环路的产生。但是仅应用有序列表在路由聚合的情况下会存在环路的可能。

如图 6-71，AS20 发布路由 10.0.12.0/24 给 AS100，该路由在 AS100 处聚合后发布给 AS30，最后又被 AS30 发回给 AS20。而此时 AS20 收到了聚合后的路由，并不包含 AS20 的路径信息，因为详细的路径信息在 AS100 处聚合时丢失了。

这时如果 AS20 内部有一个网段（例如 10.0.12.224/27）发生了故障，AS20 就会与来自 AS30 的聚合后的路由 10.0.10.0/22 匹配，发生路由环路。

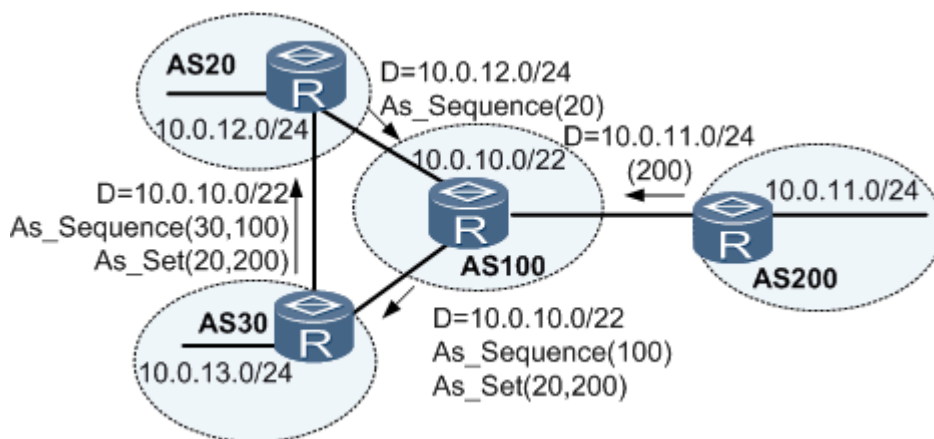
图 6-71 聚合导致的环路



解决以上问题的方法就是在聚合路由中加入无序集 AS\_Set。

如图 6-72，在 AS100，聚合路由开始了一个新的 AS\_Sequence 序列（100），同时在聚合路由中加入 AS\_Set 序列（20，200）记录丢失的路径信息。这里 AS\_Set 是一个无序集，是因为被聚合路由 10.0.12.0/24 和 10.0.11.0/24 之间是没有顺序关系的。

图 6-72 As\_Set 和 As\_Sequence



#### 说明

Atomic\_Aggregate 和 AS\_Set 都用来表示路由发生了聚合。当一个 AS\_Path 中包含了 AS\_Set 时，在聚合路由中就不再需要加入 Atomic\_Aggregate 属性。因为 AS\_Set 不仅表示该路由发生了聚合，同时包含了聚合前的详细路径信息。

## Next\_Hop 属性

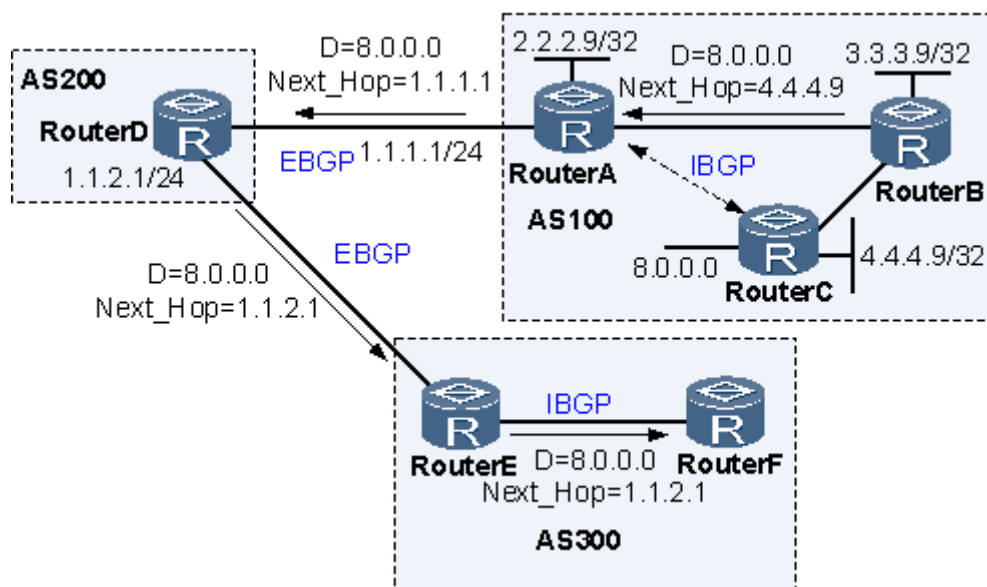
BGP 的下一跳属性和 IGP 的有所不同，不一定就是邻居路由器的 IP 地址。

通常情况下，Next\_Hop 属性遵循下面的规则：

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

图 6-73 中通过一条路由的发布，描述了 BGP 下一跳属性的三个规则。

图 6-73 下一跳属性



- AS100 内部 IBGP 对等体间传递 IBGP 内部路由。

在 AS100 中，三台路由器上运行 IGP 协议，RouterA 与 RouterC 之间通过 Loopback 接口建立 IBGP 邻居，则 RouterA 接收到 Update 报文中目的地址是 8.0.0.0 的路由的 Next\_Hop 属性是 4.4.4.9，也就是到目的地的始发路由器 RouterC 的接口地址。

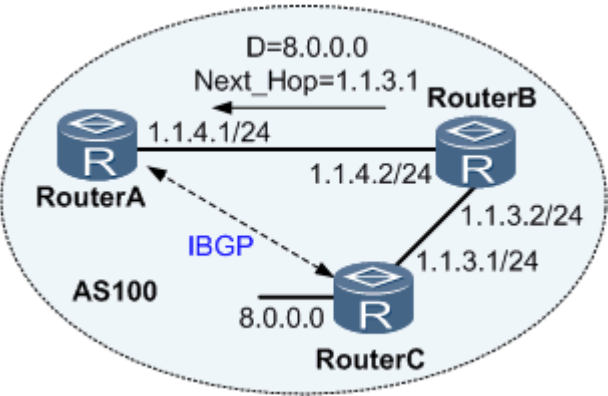
但是目的地 8.0.0.0 不属于与 RouterA 相连的子网地址中的任何一个，如图 6-74，RouterA 是通过路由迭代（iterative）找到了去 8.0.0.0 的路径。查看 RouterA 的 BGP 路由表，可以看到，RouterA 通过 IGP 路由查找到了实际的下一跳是 1.1.4.2。

说明

路由迭代的概念，请参见[应用 BGP 负载分担时的选路策略](#)。



图 6-74 下一跳属性—IBGP 对等体传递 AS 内路由



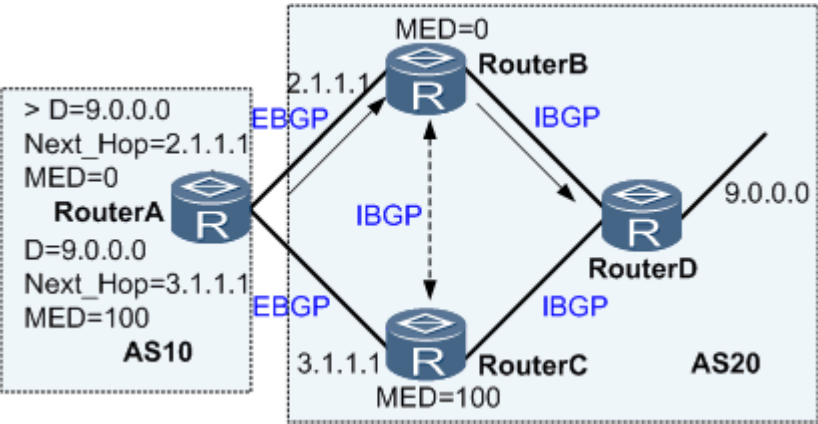
- EBGP 对等体间发布路由。  
如 图 6-73，AS200 的路由器接收到目的地址是 8.0.0.0 的路由，下一跳属性是 EBGP 对等体 RouterA 的接口地址 1.1.1.1。AS300 的 RouterE 接收到目的地址是 8.0.0.0 的路由，下一跳属性是 EBGP 对等体接口地址 1.1.2.1。
- IBGP 对等体间传递 EBGP 路由。  
RouterF 接收到目的地是 8.0.0.0 的路由，下一跳属性是 EBGP 对等体的接口地址 1.1.2.1，可见 RouterE 在传递此路由时，没有改变下一跳地址为自身的接口地址。  
但是在 RouterF 并不知道 AS 外部网络 1.1.2.0 的存在，这个外部网络并没有在 AS300 内部发布。因此 RouterF 学到的 D=8.0.0.0 的路由是不可达的，RouterF 发往 8.0.0.0 的数据包将被丢弃。在这种情况下，为保证 IBGP 邻居能够找到正确的下一跳，可以配置在向 IBGP 对等体发布路由时，改变下一跳地址为自身地址。

MED 属性

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

MED 属性相当于 IGP 使用的度量值（Metrics），它用于判断流量进入 AS 时的最佳路由。当一个运行 BGP 的路由器通过不同的 EBGP 对等体得到目的地址相同但下一跳不同的多条路由时，在其它条件相同的情况下，将优先选择 MED 值较小者作为最佳路由。如 图 6-75 所示，从 AS10 到 AS20 的流量将选择 RouterB 作为入口。

图 6-75 MED 属性



通常情况下，BGP 只比较来自同一个 AS 的路由的 MED 属性值。如果 AS10 的路由器同时从 AS10 和另外一个 AS 收到来自 AS20 的路由，不会比较 MED 值。

#### 🔗 窍门

在 USG 的实现中，通过配置 **compare-different-as-med** 命令，可以强制 BGP 比较来自不同 AS 的路由的 MED 属性值。

#### 📖 说明

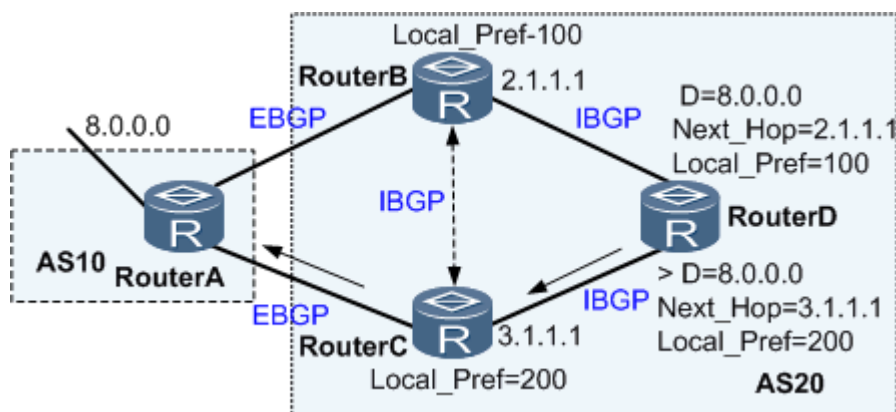
- MED 属性仅在相邻的 AS 间传递，如 [图 6-75](#)，AS10 的 1 路由器收到了 AS20 路由的 MED 属性，仅在 AS10 内部传播。如果目的地是 9.0.0.0 的路由被 AS10 发送到另外一个 AS，AS20 的 MED 属性不会携带在此路由中。
- **AS\_Path 属性的作用**可在多个 AS 间传递。

## Local\_Pref 属性

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

Local\_Pref 属性用于判断流量离开 AS 时的最佳路由。当 BGP 的路由器通过不同的 IBGP 对等体得到目的地址相同但下一跳不同的多条路由时，将优先选择 Local\_Pref 属性值较高的路由。如 [图 6-76](#) 所示，从 AS20 到 AS10 的流量将选择 RouterC 作为出口。

图 6-76 Local\_Pref 属性



#### 📖 说明

MED 属性和 Local\_Pref 属性的区别在于：MED 用于判断流量进入 AS 时的最佳路由，Local\_Pref 属性用于判断流量离开 AS 时的最佳路由。

## Community 属性

团体属性用来简化路由策略的应用和降低维护管理的难度。它是一组有相同特征的目的地址的集合，没有物理上的边界，与其所在的 AS 无关。这些目的地址共享一个或多个共同的属性。

团体属性用一组以 4 字节为单位的列表来表示，路由器中团体属性的格式是 aa:nn 或团体号。

- aa:nn: aa 和 nn 的取值范围都是 0 ~ 65535，管理员可根据实际情况设置具体数值。通常 aa 表示自治系统 AS 编号，nn 是管理员定义的团体属性标识。例如，来



自 AS100 的一条路由，管理员定义的团体属性标识是 1，则该路由的团体属性格式是 100:1。

- 团体号：团体号是 0 ~ 4294967295 的整数。RFC1997 中定义，0（0x00000000）~ 65535（0x0000FFFF）和 4294901760（0xFFFF0000）~ 4294967295（0xFFFFFFFF）是预留的。

路由器中，BGP 公认的团体属性见表 6-38。

表 6-38 BGP 公认团体属性

| 团体名称                | 团体标识                  | 说明                                                                |
|---------------------|-----------------------|-------------------------------------------------------------------|
| Internet            | 0（0x00000000）         | 缺省情况下，所有的路由都属于 Internet 团体。具有此属性的路由可以被通告给所有的 BGP 对等体。             |
| No_Export           | 4294967041（0xFFFFF01） | 具有此属性的路由在收到后，不能被发布到本地 AS 之外。如果使用了联盟，则不能被发布到联盟之外，但可以发布给联盟中的其他子 AS。 |
| No_Advertise        | 4294967042（0xFFFFF02） | 具有此属性的路由在收到后，不能被通告给任何其他 BGP 对等体。                                  |
| No_Export_Subconfed | 4294967043（0xFFFFF03） | 具有此属性的路由在收到后，不能被发布到本地 AS 之外，也不能发布到联盟中的其他子 AS。                     |

Atomic\_Aggregate

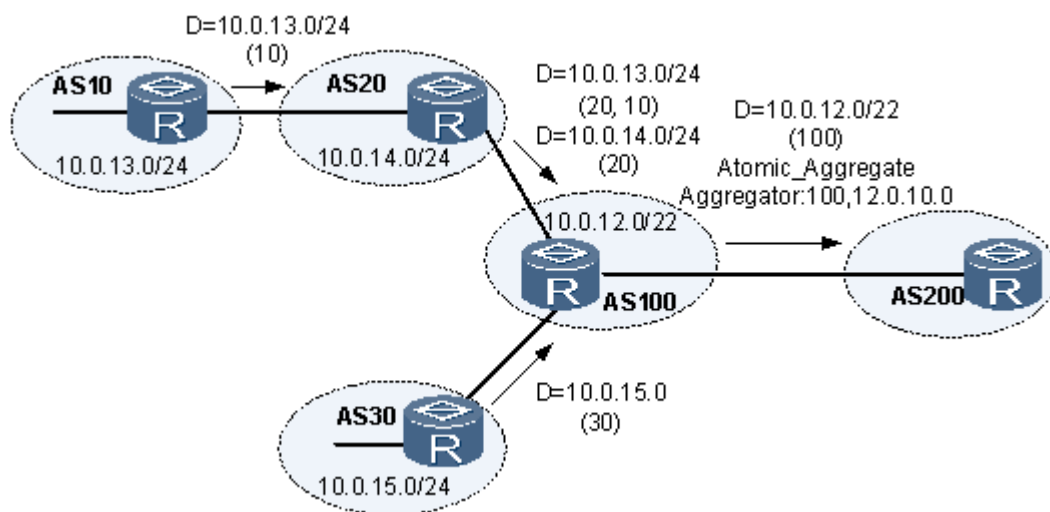
当遇到重叠路由（overlapping）时，BGP Speaker 会选择掩码较短的路由，而非更具体的路由进行发布。这时聚合路由会标记 Atomic\_Aggregate 属性告知下游路由器出现了 AS\_Path 属性信息丢失。

- 当 BGP Speaker 接收到带有 Atomic\_Aggregate 属性的路由时，无法得知也不允许把 NLRI 修改的更加详细。
- 当 BGP Speaker 要发送这条路由到其他 BGP Speaker 时，Atomic\_Aggregate 属性必须继续存在于该路由中。

如图 6-77，三条路由 10.0.13.0/24、10.0.14.0/24、10.0.15.0/24 与路由 10.0.12.0/22 是重叠路由，尽管这四条路由是不同的路由，但是 10.0.12.0/22 包含了前三条路由。

在三条路由到达 AS100 时，AS100 上对路由进行了聚合，只发布 10.0.12.0/22 给 AS200。此时的 AS\_Path 信息里只包含 AS100。之前的详细路径信息都丢失了。所以在其发出的路由中加入 Atomic\_Aggregate 属性标识已经发生了路径信息的丢失。

图 6-77 Atomic\_Aggregate



## Aggregator

可以选择加入 Aggregator 属性标识路径信息丢失的位置。Aggregator 属性包含发起聚合的路由器的 ID 和 AS 号。如图 6-77，Aggregator:100,10.0.12.0 表示发起此聚合的路由器在 AS100 内，该路由器的 IP 地址是 10.0.12.0。

### 说明

Atomic\_Aggregate: 标识发生了路径信息的丢失。

Aggregator: 标识路径信息丢失发生的位置。

## Originator\_ID

在路由反射器上使用，当从源路由器收到路由向目的路由器反射时，使用 Originator\_ID 标识源路由器 Router ID。

详细内容请参见 6.7.4.7 BGP 路由反射器。

## Cluster\_List

在路由反射器上使用，用于检测和预防 AS 内部产生环路。

详细内容请参见 6.7.4.7 BGP 路由反射器。

## 6.7.4.4 BGP 路由聚合

在大规模的网络中，BGP 路由表十分庞大，使用路由聚合（Routes Aggregation）可以大大减小路由表的规模。

路由聚合实际上是将多条路由合并的过程。这样 BGP 在向对等体通告路由时，可以只通告聚合后的路由，而不是将所有具体的路由都通告出去。

USG 支持自动聚合和手动聚合方式：

- 自动聚合：对 BGP 引入的路由进行聚合。配置自动聚合后，对参加聚合的具体路由进行抑制。
- 手动聚合：对 BGP 本地路由进行聚合。通常情况下，手动聚合的优先级高于自动聚合的优先级，可以控制聚合路由的属性，以及决定是否发布具体路由。

 说明

本地聚合后的路由不再参与本地的路由聚合。

### 6.7.4.5 BGP 对等体组

对等体组（Peer Group）是一些具有某些相同策略的对等体的集合。当一个对等体加入对等体组中时，此对等体将获得与所在对等体组相同的配置。当对等体组的配置改变时，组内成员的配置也相应改变。

在大型 BGP 网络中，对等体的数量会很多，其中，很多对等体需要配置相同的策略，通过配置对等体组并将对等体加入到对等体组，可以使对等体获得与所在对等体组相同的配置，而且当对等体组的配置改变时，组内成员的配置也相应改变，从而简化配置。

根据对等体所在的 AS，对等体组可分为：

- IBGP 对等体组：对等体组中的对等体与当前路由器位于同一 AS。
- EBGP 对等体组：对等体组中的对等体与当前路由器位于不同 AS。

对于 EBGP 对等体组，根据所包括的对等体是否属于同一个外部 AS，又可分为纯 EBGP 对等体组和混合 EBGP 对等体组。

### 6.7.4.6 BGP 团体

对等体组可以使一组对等体共享相同的策略，而团体可以使一组路由共享相同的策略。团体是一个路由属性（Community 属性），在 BGP 对等体之间传播，它并不受到 AS 范围的限制。

BGP 路由器在将带有团体属性的路由发布给其它对等体之前，可以改变此路由原有的团体属性。

除了使用公认的团体属性外，用户还可以使用团体属性过滤器过滤自定义扩展团体属性，以便更为灵活的控制路由策略。

### 6.7.4.7 BGP 路由反射器

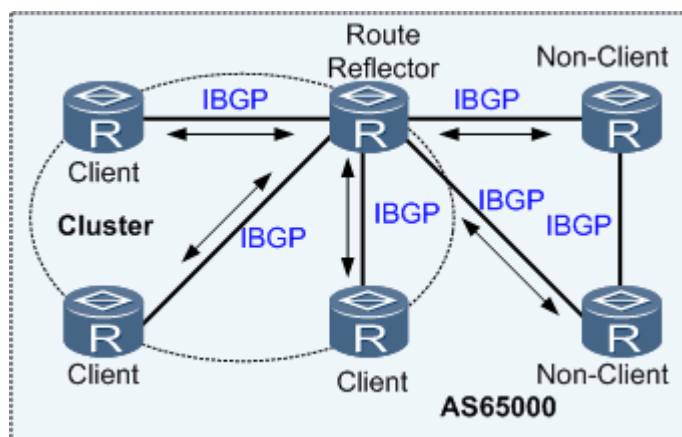
为保证 IBGP 对等体之间的连通性，需要在 IBGP 对等体之间建立全连接关系。假设在一个 AS 内部有  $n$  台路由器，那么应该建立的 IBGP 连接数就为  $n(n-1)/2$ 。当 IBGP 对等体数目很多时，对网络资源和 CPU 资源的消耗都很大。

### RR 的定义

利用路由反射可以解决这一问题。在一个 AS 内，其中一台路由器作为路由反射器 RR（Route Reflector），其它路由器做为客户机（Client）。客户机与路由反射器之间建立 IBGP 连接。路由反射器和它的客户机组成一个集群（Cluster）。路由反射器在客户机之间传递（反射）路由信息，而客户机之间不需要建立 BGP 连接。

既不是反射器也不是客户机的 BGP 路由器被称为非客户机（Non-Client）。非客户机与路由反射器之间，以及所有的非客户机之间仍然必须建立全连接关系。如图 6-78 所示。

图 6-78 路由反射器示意图



## RR 的操作

当 RR 收到 IBGP 对等体发来的路由，首先使用 **BGP 选择路由的策略** 选择最佳路由。在公布学习到的路由信息时，RR 按照 RFC2796 中的规则发布路由：

- 从非客户机 IBGP 对等体学到的路由，发布给此 RR 的所有客户机。
- 从客户机学到的路由，发布给此 RR 的所有非客户机和客户机（发起此路由的客户机除外）。
- 从 EBGP 对等体学到的路由，发布给所有的非客户机和客户机。

RR 的一个好处就是配置方便，因为只需要在反射器上配置，客户机不需要知道自己是客户机。

在某些网络中，路由反射器的客户机之间已经建立了全连接，它们可以直接交换路由信息，此时客户机到客户机之间的路由反射是没有必要的，而且还占用带宽资源。USG 支持配置命令 **reflect between-clients** 来控制在客户机之间反射路由。缺省情况下，允许客户机之间的路由反射，客户机到非客户机之间的路由仍然可以被反射。

## Originator\_ID

RFC2796 定义了 Originator\_ID 属性和 Cluster\_List 属性，用于检测和防止路由环路。

Originator\_ID 属性长 4 字节，是由路由反射器（RR）产生的，携带了本地 AS 内部路由发起者的 Router ID。

- 当一条路由第一次被 RR 反射的时候，RR 将 Originator\_ID 属性加入这条路由，标识这条路由的发起路由器。如果一条路由中已经存在了 Originator\_ID 属性，则 RR 将不会创建新的 Originator\_ID。
- 当其它 BGP Speaker 接收到这条路由的时候，将比较收到的 Originator\_ID 和本地的 Router ID，如果两个 ID 相同，BGP Speaker 会忽略掉这条路由，不做处理。

📖 说明

Originator\_ID 属性保证路由在反射器和客户机/非客户机之间的不出现循环。

## Cluster\_List

对于 AS 之间，BGP 用于防止环路的主要措施是通过 AS\_Path 属性记录途经的 AS 路径，带有本地 AS 号的路由将被路由器丢弃；对于 AS 之内，BGP 防止路由环路的方法是禁止 IBGP 对等体发布从 AS 内部学来的路由。

路由反射器的实现是基于放宽对“BGP 在 AS 内学到的路由不会在 AS 中转发”的要求，即允许 IBGP 对等体之间发布从 AS 内部学来的路由。在这种情况下，Cluster\_List 属性被引入，用于防止 AS 内部的环路。

Cluster\_List 由一系列的 Cluster\_ID 组成，描述了一条路由所经过的反射器路径，这和描述路由经过的 AS 路径的 AS\_Path 属性有相似之处。Cluster\_List 由路由反射器产生。

- 当 RR 在它的客户机之间或客户机与非客户机之间反射路由时，RR 会把本地 Cluster\_ID 添加到 Cluster\_List 的前面。如果 Cluster\_List 为空，RR 就创建一个。
- 当 RR 接收到一条更新路由时，RR 会检查 Cluster\_List。如果 Cluster\_List 中已经有本地 Cluster\_ID，丢弃该路由；如果没有本地 Cluster\_ID，将其加入 Cluster\_List，然后反射该更新路由。

### 说明

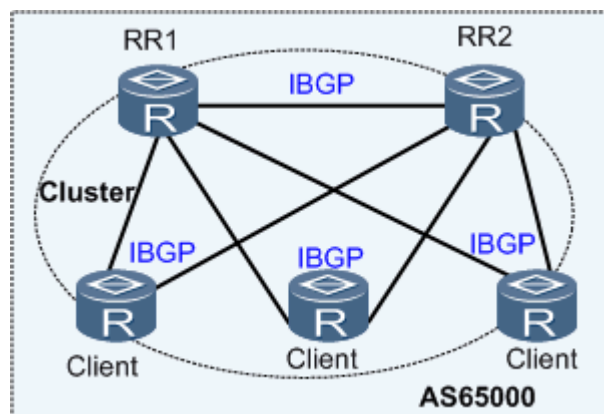
- 缺省情况下，每个反射器是使用自己的路由器 ID 作为 Cluster\_ID。
- 为了保证客户机可以学习到反射器发来的路由，Cluster\_ID 不能和客户机的 Router ID 相同。如果相同，客户机将会将收到的路由丢弃。
- Cluster\_List 只在 AS 内部传播，从 EBGP 对等体收到的含有 Cluster\_List 的路由将被丢弃。

## 备份 RR

为增加网络的可靠性，防止单点故障，有时需要在一个集群中配置一个以上的路由反射器。这时，相同集群中的路由反射器要共享相同的 Cluster\_ID，以避免路由环路。路由器中需要使用命令 **reflector cluster-id** 给所有位于同一个集群内的路由反射器配置相同的 Cluster\_ID。

在冗余的环境里，客户机收到不同反射器发来的到达同一目的地的多条路由，这时客户机应用 **BGP 选择路由的策略** 来选择最佳路由。

图 6-79 备份路由反射器



如图 6-79，路由反射器 RR1 和 RR2 在同一个 Cluster 内。RR1 和 RR2 之间配置 IBGP 连接，即两个反射器互为非客户机。

- 当客户机 Client1 从外部对等体接收到一条更新路由后，它通过 IBGP 向 RR1 和 RR2 通告这条路由。
- RR1 接收到该更新路由后，它向其他的客户机（Client2、Client3）和非客户机（RR2）反射，同时将本地 Cluster\_ID 添加到 Cluster\_List 前面。
- RR2 接收到该反射路由后，检查 Cluster\_List，发现自己的 Cluster\_ID 已经包含在 Cluster\_List 中。因此，它丢弃该更新路由，不再向自己的客户机反射。

如果 RR1 和 RR2 配置不同的 Cluster\_ID，RR2 接收到 RR1 反射的更新路由后，检查 Cluster\_List，发现没有本地 Cluster\_ID，它会将自己的 Cluster\_ID 添加到 Cluster\_List 前面，然后将该更新路由反射给自己的客户机。这样，Client1 会接收到自己转发出去的路由并更新该路由，造成路由环路。

#### 说明

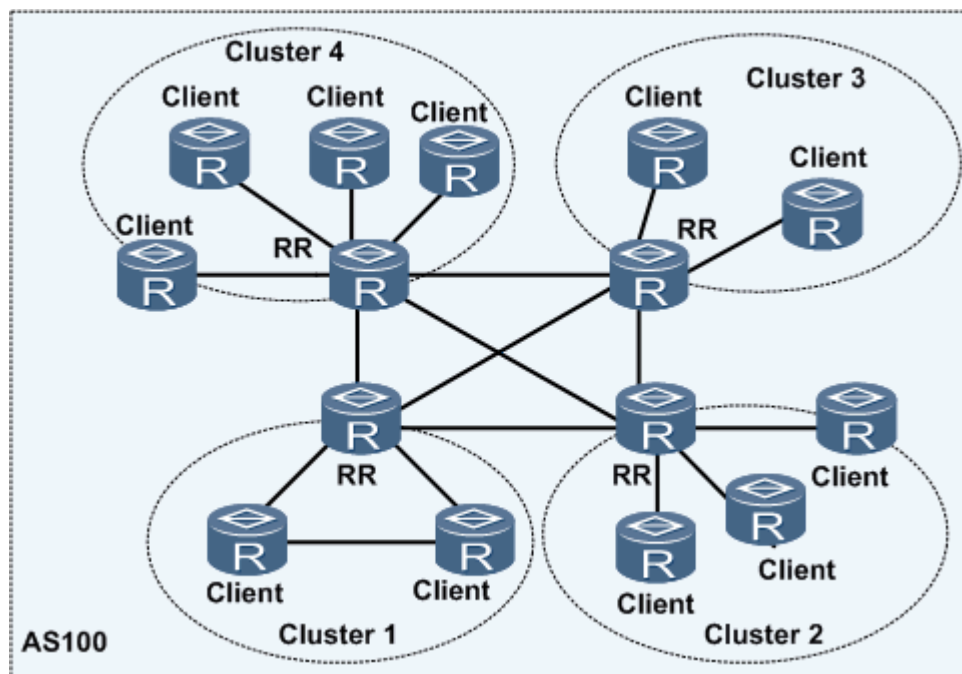
Cluster\_List 的应用保证了同一反射集群内的 RR 之间不出现路由循环。

## AS 内多个集群

一个 AS 中可能存在多个集群（Cluster）。各个 RR 之间是 IBGP 对等体的关系，一个 RR 可以把另一个 RR 配置成自己的客户机或非客户机。因此可以灵活的配置 AS 内部集群与集群之间的关系。

例如，一个骨干网被分成多个反射集群，每个 RR 将其它的 RR 配置成非客户机，各 RR 之间建立全连接。每个客户机只与所在集群的 RR 建立 IBGP 连接。这样该自治系统内的所有 BGP 路由器都会收到反射路由信息。如图 6-80 所示。

图 6-80 AS 内多个集群



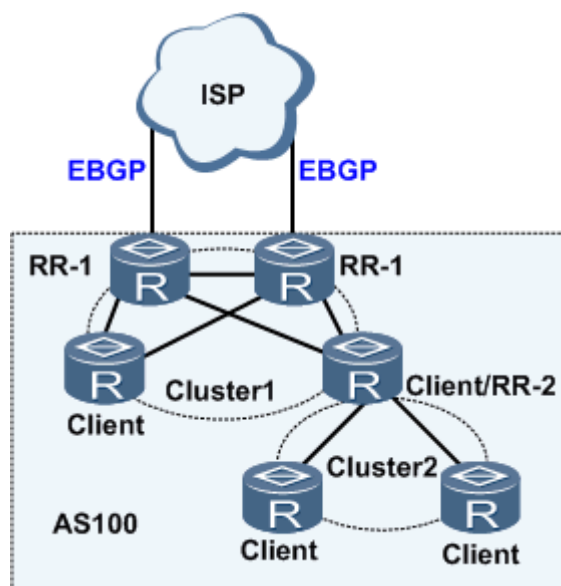


## 分级反射器

在实际的反射器部署中，常用的是分级反射器的场景。如图 6-81，ISP 为 AS100 提供 Internet 路由，ISP 与 AS100 内建立双出口 EBGP 连接。AS100 内部分为两个集群。Cluster1 内的四台路由器是核心路由器。

- Cluster1 中部署了两个一级 RR（RR-1），这种冗余结构保证了 AS100 内部网络核心层的可靠性。核心层其余两台路由器作为 RR-1 的客户机。
- Cluster2 中部署了一个二级 RR（RR-2），这个 RR-2 同时也是 RR-1 的客户机。

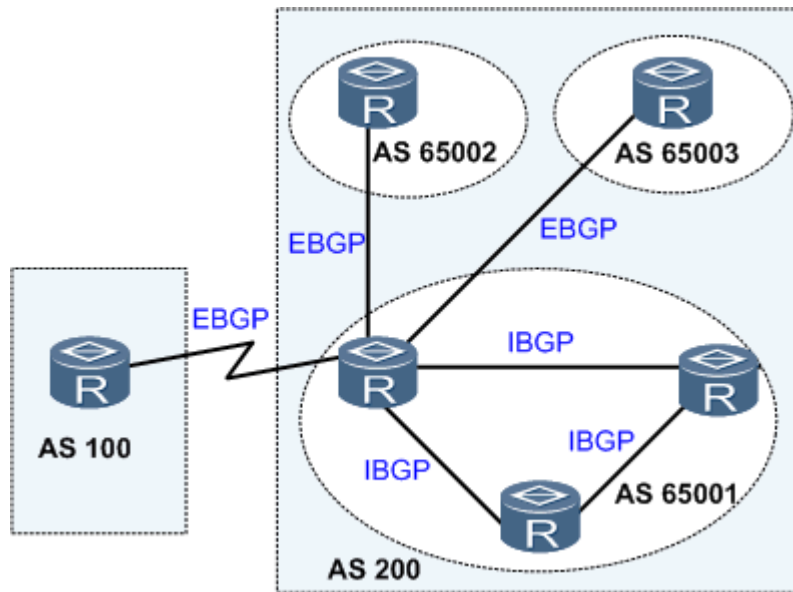
图 6-81 分级反射器



### 6.7.4.8 BGP 联盟

联盟（Confederation）是处理 AS 内部的 IBGP 网络连接激增的另一种方法，它将一个自治系统划分为若干个子自治系统，每个子自治系统内部建立 IBGP 全连接关系，子自治系统之间建立 EBGP 连接关系。如图 6-82 所示。

图 6-82 联盟示意图



在不属于联盟的 BGP Speaker 看来，属于同一个联盟的多个子自治系统是一个整体，外界不需要了解内部的子自治系统情况，联盟 ID 就是标识联盟这一整体的自治系统号，如上图中的 AS200 就是联盟 ID。

## 联盟的 AS\_Path 属性

RFC3065 扩展了 AS\_Path 属性，应用于联盟：

- As\_Confed\_Sequence：联盟内部 AS 号的有序集。用法同 As\_Sequence。
- As\_Confed\_Set：联盟内部 AS 号的无序集。用法同 As\_Set。

当 BGP Speaker 向对等体发布一条学习到的路由时，AS\_Path 属性是否需要修改，取决于对等体的位置。

- 对等体与 BGP Speaker 在联盟内的同一 AS 内部，不修改路由的 AS\_Path 属性。
- 对等体与 BGP Speaker 不在联盟内的同一 AS 内部，但对等体所在的 AS 属于本地联盟，将按照以下规则更新 AS\_Path 属性：
  - 如果 AS\_Path 属性中存在 As\_Confed\_Sequence，BGP 路由器将自己的 AS 号添加到序列的最左端。
  - 如果 AS\_Path 属性中不存在 As\_Confed\_Sequence，创建一个包含本地 AS 号的新的 As\_Confed\_Sequence。
- 对等体与 BGP Speaker 不在同一 AS 内，且对等体所在 AS 不属本地联盟。

如果 AS\_Path 属性中包含 As\_Confed\_Sequence，或同时包含 As\_Confed\_Sequence 和 As\_Confed\_Set，在 AS\_Path 属性中删除这两个类型的序列，并按照下面的步骤执行：

- 如果在删除后，AS\_Path 属性中包含 As\_Sequence，BGP 路由器将自己所在联盟的 ID 添加到序列的最左端。
- 如果在删除后，AS\_Path 属性中不包含 As\_Sequence，创建一个新的序列（包含所在联盟的 ID 号）。



当 BGP Speaker 产生一条新的路由时，

- 发往 AS 内部对等体的路由中，AS\_Path 属性为空。
- 发往联盟内不同 AS 的路由中，AS\_Path 属性的 As\_Confed\_Sequence 中要包含本地的 AS 号。
- 发往联盟外 AS 的路由中，AS\_Path 属性的 As\_Sequence 中要包含本地的联盟 ID。

## 联盟的限制和应用

反射器的配置比较简单，只在反射器上配置，仅需充当反射器的路由器具有反射功能即可。而联盟则需要每个路由器上进行配置，且要求加入联盟的路由器具有联盟功能。

联盟的缺陷是：从非联盟向联盟方案转变时，要求路由器重新进行配置，逻辑拓扑也要改变。

在大型 BGP 网络中，路由反射器和联盟可以被同时使用。

### 6.7.4.9 BGP 衰减

路由衰减（Route Dampening）用来解决路由不稳定的问题。路由不稳定的主要表现形式是路由振荡（Route flapping），即路由表中的某条路由反复消失和重现。

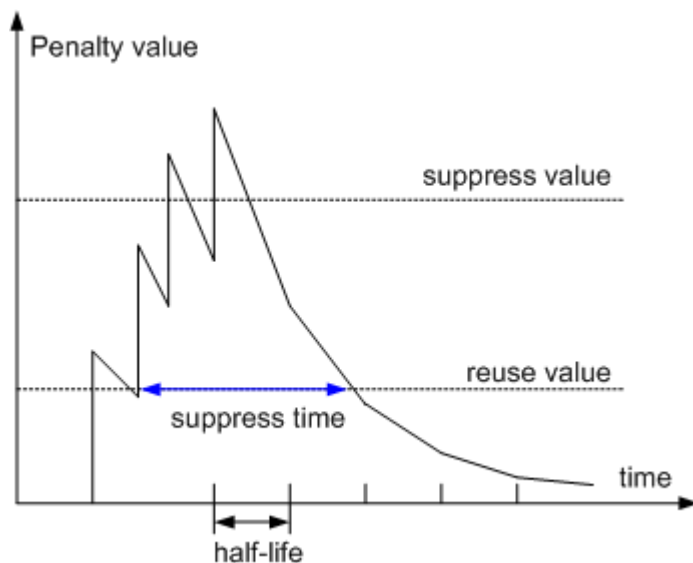
发生路由振荡时，路由协议就会向邻居发布路由更新，收到更新报文的路由器需要重新计算路由并修改路由表。所以频繁的路由振荡会消耗大量的带宽资源和 CPU 资源，严重时会影响网络的正常工作。

在多数情况下，BGP 协议都应用于复杂的网络环境中，路由变化十分频繁。为了防止持续的路由振荡带来的不利影响，BGP 使用衰减来抑制不稳定的路由。

BGP 衰减使用惩罚值来衡量一条路由的稳定性，惩罚值越高则说明路由越不稳定。路由每发生一次振荡（路由从激活状态变为未激活状态，称为一次路由振荡），BGP 便会给此路由增加一定的惩罚值（1000）。当惩罚值超过抑制阈值时，此路由被抑制，不加入到路由表中，也不再向其他 BGP 对等体发布更新报文。

被抑制的路由每经过一段时间，惩罚值便会减少一半，这个时间称为半衰期（Half-life）。当惩罚值降到再使用阈值时，此路由变为可用并被加入到路由表中，同时向其他 BGP 对等体发布更新报文。BGP 衰减的处理过程如图 6-83 所示。

图 6-83 BGP 衰减示意图





说明

只对 EBGp 路由进行衰减。

## 等级 Dampening

在规划国际关口局或大型网络中的核心路由器的 BGP 协议时，需要应用 Dampening 特性来防止由于链路不稳定造成不必要的路由震荡。假设路由震荡达到 3 次，需要对该路由实行 10 分钟的抑制，这段时间内该路由不会发布给任何邻居。由于该路由是其他自治域内一台普通路由器或大型核心路由器所产生的影响是不一样的，因此需要对不同前缀的路由采用不同的 Dampening 策略。

等级 Dampening（Specified-requirement Dampening）实现了 BGP 采用不同的 Dampening 抑制参数针对不同的 EBGp 邻居路由有选择的进行抑制。通过等级 Dampening，可按策略区分路由，BGP 对不同的路由采用不同的 Dampening 参数进行抑制处理。同一路由策略的不同节点下，可以根据需要配置不同的 Dampening 抑制参数。路由振荡时，BGP 会根据路由通过策略的结果，使用不同的 Dampening 参数进行抑制处理。例如，实际网络中，对掩码较长的路由设置较长的抑制时间，而对掩码较短的（例如 8 位掩码长度）路由，则采用相对较短的抑制时间。

### 6.7.4.10 BGP 安全性

#### BGP 验证

BGP 使用 TCP 作为传输层协议，为提高 BGP 的安全性，可以在建立 TCP 连接时进行 MD5 认证。但 BGP 的 MD5 认证并不能对 BGP 报文认证，它只是为 TCP 连接设置 MD5 认证密码，由 TCP 完成认证。如果认证失败，则不建立 TCP 连接。

### 6.7.4.11 BGP4+特性

BGP4+是一种用于自治系统 AS（Autonomous System）之间的动态路由协议，它是对 BGP 的扩展。

传统的 BGP4 只能管理 IPv4 的路由信息，对于使用其它网络层协议（如 IPv6 等）的应用，在跨自治系统传播时就受到一定限制。

为了提供对多种网络层协议的支持，IETF 对 BGP4 进行了扩展，形成 BGP4+，目前的 BGP4+标准是 RFC2858（Multiprotocol Extensions for BGP-4，BGP-4 多协议扩展）。

为了实现对 IPv6 协议的支持，BGP4 需要将 IPv6 网络层协议的信息反映到 NLRI（Network Layer Reachable Information）及 Next\_Hop 属性中。

BGP4+中引入的两个 NLRI 属性分别是：

- MP\_REACH\_NLRI: Multiprotocol Reachable NLRI，多协议可达 NLRI。用于发布可达路由及下一跳信息。
- MP\_UNREACH\_NLRI: Multiprotocol Unreachable NLRI，多协议不可达 NLRI。用于撤销不可达路由。

BGP4+中的 Next\_Hop 属性用 IPv6 地址来表示，可以是 IPv6 全球单播地址或者下一跳的链路本地地址。

BGP IPv6 是利用 BGP 的多协议扩展属性，来达到在 IPv6 网络中应用的目的，BGP 协议原有的消息机制和路由机制并没有改变。

## BGP4+参考协议

BGP4+参考的协议和标准如下：

- RFC 1745: BGP4/IDRP for IP---OSPF Interaction
- RFC 2270: Using a Dedicated AS for Sites Homed to a Single Provider

## BGP4+与 BGP 的规格差异

- 支持引入直连路由、静态路由、RIPng 路由、OSPFv3 路由和 IS-IS 路由。
- BGP4+使用 IPv6 的报文验证，不支持 MD5 验证。
- 不支持路由聚合。

## 6.8 IP 单播策略路由

策略路由是一种依据用户制定的策略进行路由选择的机制，可应用于安全、负载分担等目的。

### 6.8.1 介绍

#### 定义

策略路由（Policy-based routing，简称 PBR）是一种依据用户制定的策略进行路由选择的机制。

IP 单播策略路由用来对 IP 单播报文进行策略路由。

#### 目的

普通的报文转发是依据报文的目地址查询 FIB 表来实现的。但很多大型复杂的网络，如高校校园网，对外往往使用多出口，由不同的 ISP 提供网络服务。基于访问速度、费用、资源利用等考虑，如何对不同报文进行路由控制，是这些大型网络期望解决的问题。

策略路由就是为解决这个问题而提出的，通过策略路由，可以根据以下信息，将报文分类，然后控制将这些报文按照不同的路由进行转发，从而达到提高网络安全性能和负载分担的目的。

- 基于报文的源 IP、目的 IP、协议、端口号、优先级、ToS、时间段等各种丰富的信息
- 基于发起网络访问的用户/用户组
- 基于应用协议

### 6.8.2 参考标准和协议

与策略路由相关的参考标准与协议如下：

- RFC 791: Internet Protocol

### 6.8.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 6.8.4 特性增强

| 版本       | 特性增强                                              |
|----------|---------------------------------------------------|
| V300R001 | 策略路由新增用户/用户组/应用协议作为匹配条件，将报文分类，灵活地控制报文按照不同的路由转发出去。 |

### 6.8.5 应用场景

#### 应用环境

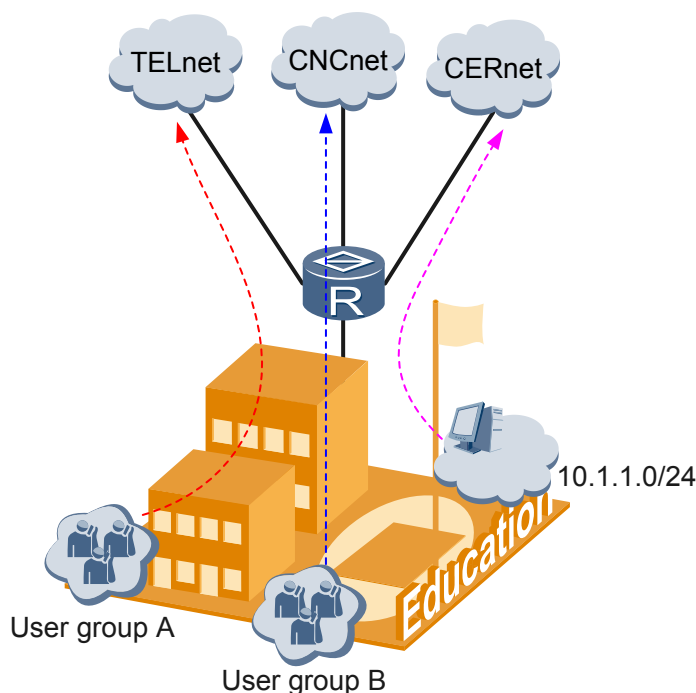
内部网络通过设备与外部网络连接，设备上存在多个到外部网的出口。通过策略路由，可以控制某些报文通过指定的出口转发出去。

#### 典型应用

如图 6-84 所示，在某高校校园网中，设备作为出口网关，存在多个出口，将内部网络连接到外部网络。外部网络提供的服务及其特点如下所示：

- 电信网（TELnet）：较高付费的网络资源，是内网用户访问互联网的最主要通道。
- 联通网（CNCnet）：较高付费的网络资源，是内网用户访问互联网的最主要通道。
- 教育网（CERnet）：主要是方便内网用户访问教育网资源和接受来自教育网用户访问，带宽利用率远低于电信或联通网络。带宽采用包月收费，即无论使用流量多大，每月费用固定不变。

图 6-84 IP 单播策略路由典型组网应用



基于访问速度、费用、资源利用等考虑，可以通过策略路由，控制报文的转发：

- 基于用户的策略路由可以控制指定用户/用户组只能通过指定的链路访问互联网，例如用户组 A 只能通过电信网链路访问互联网，用户组 B 只能通过联通网链路访问互联网。
- 基于 ACL 的策略路由可以根据报文的源地址、目的地址等信息，控制报文按照指定链路转发，例如源 IP 地址为 10.1.1.0/24 的用户只能通过教育网链路访问互联网。
- 基于应用的策略路由就可以使教育网链路主要分担内网 P2P 的需求，这样既能满足内网用户 P2P 应用的需要，可以不用过于的打压，又可以释放宝贵的电信、联通外网带宽资源。

## 6.8.6 原理描述

### 6.8.6.1 策略路由

#### 策略路由的分类

根据作用对象的不同，策略路由可分为本地策略路由和接口策略路由：

- 本地策略路由：对设备自身所发的报文进行策略路由，它只对设备自身产生的报文起作用，对转发的报文不起作用。
- 接口策略路由：对到达该接口的报文进行策略路由，它只对转发的报文起作用，对设备自身产生的报文不起作用。

对于一般转发和安全等方面的使用需求，大多数情况下使用的是接口策略路由。

策略路由的组成

策略路由功能基于用户制定的策略进行路由选择，一个策略可以包含一个或多个节点。

节点

每个节点由唯一的顺序号 *node-id* 来标识。*node-id* 的值越小优先级越高，优先级高的节点优先被执行。

每个节点的具体内容由 **if-match** 和 **apply** 子句来指定：

- **if-match** 子句定义该节点的匹配规则。系统根据这些匹配规则将报文进行分类。
- **apply** 子句定义该节点的动作。系统根据这些动作决定如何处理满足匹配规则的报文、从哪个接口转发出去。

每个节点对报文的处理方式由匹配模式 **permit** 和 **deny** 决定。

一个节点的匹配模式与这个节点的 **if-match** 子句、**apply** 子句的关系如表 6-39 所示。

表 6-39 节点的匹配模式、if-match 子句、apply 子句三者之间的关系

| 报文是否满足 if-match 子句 | permit                           | deny                                                            |
|--------------------|----------------------------------|-----------------------------------------------------------------|
| 是（所有匹配项都满足）        | 执行此节点 <b>apply</b> 子句，不再匹配下一节点。  | 不执行此节点 <b>apply</b> 子句，不再匹配下一节点，报文按普通路由（根据报文中的目的地址查 FIB 表）进行转发。 |
| 否（任意一个匹配项不满足）      | 不执行此节点 <b>apply</b> 子句，继续匹配下一节点。 | 不执行此节点 <b>apply</b> 子句，继续匹配下一节点。                                |

说明

如果策略路由的某节点没有配置 **if-match** 和 **apply** 子句，设备对报文的处理方式如下：

- 如果某一节点不配置 **if-match** 子句，则所有报文都不会通过该节点的过滤，继续匹配下一节点。
- 如果某一 **permit** 模式的节点不配置 **apply** 子句，当报文满足此节点的 **if-match** 子句时，将不会执行任何动作，且不再继续匹配下一节点，报文按普通路由进行转发。

if-match 子句

IP 单播策略路由提供了 3 类 **if-match** 子句：

- **if-match acl**：基于 ACL 规则匹配报文，即可以根据报文的源 IP 地址、目的 IP 地址、协议、端口号等信息将报文分类。
- **if-match { category | app-set }**：基于应用协议匹配报文。当指定的应用协议是应用大类或应用协议集时，则报文只要满足其中的 1 个应用协议，就算匹配通过。该 **if-match** 子句只在应用控制（DPI）功能识别出对应的应用协议才有效。
- **if-match { user-group | user } \***：基于用户/用户组匹配报文。当用户和用户组同时配置时，两者之间是“或”的关系，即报文只要满足 1 个匹配条件，就算该 **if-match** 子句匹配通过。该 **if-match** 子句只在对应的用户通过设备的身份认证后才有效。

在一个节点中，允许 3 类 **if-match** 子句共存，但每类只能存在 1 条。各 **if-match** 子句之间是“与”的关系，报文必须同时满足各个 **if-match** 子句，才算匹配通过。

### apply 子句

IP 单播策略路由提供了六种 apply 子句。

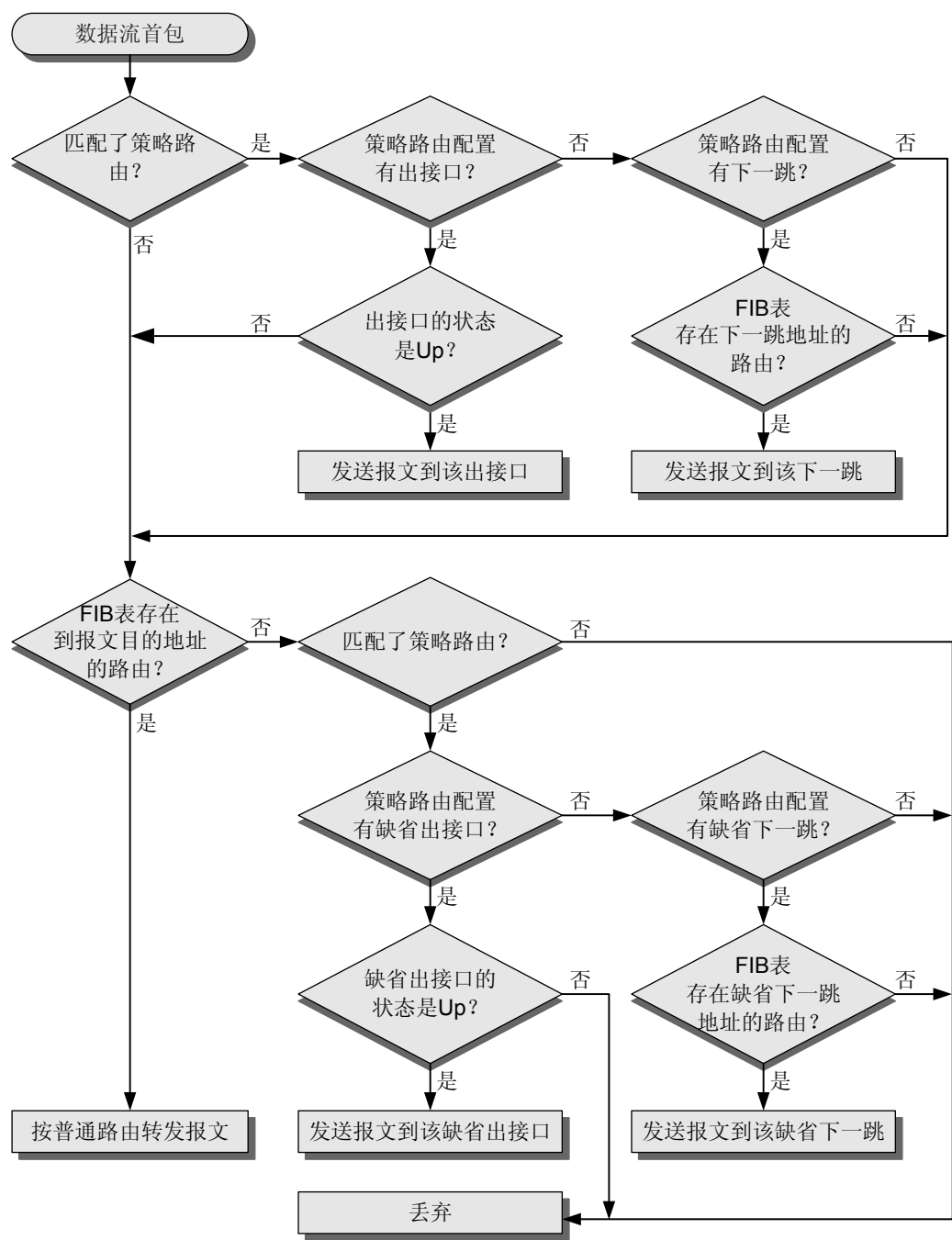
同一个节点中的各 apply 子句的执行优先级情况如下：

- 配置转发报文的 VPN 实例：**apply access-vpn vpn-instance**，只要配置了该子句，就不会执行其他 **apply** 子句（除 **apply ip-precedence** 外）。报文如果匹配了其中一个 VPN 实例下的转发表，报文将在该 VPN 实例中进行转发，如果对于所有配置的 VPN 实例，报文都未能匹配其下的转发表，该报文匹配策略路由当前节点不成功，继续匹配下一节点。
- 配置报文的优先级：**apply ip-precedence**，只要配置了该子句，该子句就一定会执行。
- 配置策略路由出接口和下一跳：**apply output-interface** 和 **apply ip-address next-hop**，其中 **apply output-interface** 的优先级高于 **apply ip-address next-hop**。当两条 **apply** 子句都有效时，系统只会执行 **apply output-interface**。
- 配置策略路由缺省出接口和下一跳：**apply default output-interface** 和 **apply ip-address default next-hop**，同样 **apply default output-interface** 的优先级高于 **apply ip-address default next-hop**。当两条 **apply** 子句都有效时，系统只会执行 **apply default output-interface**。

当策略路由中指定的出接口（下一跳）时，其优先级要高于普通路由；而对于配置了缺省出接口（下一跳）的情况，则是先按照普通路由进行转发，如果无法匹配，再进行策略路由转发。具体转发流程如 [图 6-85](#) 所示。有关 FIB 的详细介绍请参见 [6.8.6.2 FIB](#)。

[图 6-85](#) 所示为首包转发流程。后续报文命中该数据流后，将直接以缓存出接口转发不再走策略路由转发流程。

图 6-85 报文转发流程



6.8.6.2 FIB

FIB 简介

路由器转发报文的关键是 FIB 表，在系统中报文转发时查找的是 FIB 表而非路由表。这是因为路由表表示所有的有效路由所形成的表项，并不指导转发。FIB 表是网络层用来控制数据报发送的。



FIB 表中每条转发项都指明报文到某网段或某主机应通过路由器的哪个物理接口发送。发送后，报文可到达该路径的下一个路由器，或者不再经过别的路由器而传送到直接相连的网络中的目的主机。

FIB 中包含了路由器在转发报文时所必需的一组最小信息。一个 FIB 条目中一般包括目的地址、前缀长度、传输端口、下一跳地址、标明路由特征的标志以及时间戳。路由器使用 FIB 的各项来转发报文。

## FIB 工作原理

首先，为了连接不同的网络拓扑需要运行不同的路由协议，这样就产生了 RIB（Routing Information Base）。RIB 是创建 FIB 的基础，路由器会根据路由管理策略，从 RIB 中提取出最小转发信息并放入 FIB。用户还可通过路由管理向 FIB 中增加静态路由。

FIB 表包含以下元素：

- **Destination address:** 报文发送的目的网络地址或主机地址。
- **Prefix length:** 目的地址前缀长度，可确定目的地址是否对应网络或主机。
- **Nexthop:** 为了将报文发送到目的地址所要经过的紧邻的下一跳地址。
- **Flag(s):** 标明路由特征。有以下一些标志：
  - B: 黑洞路由，即下一跳是空接口（Null）。
  - D: 动态路由。
  - G: 网关。
  - H: 下一跳为主机。
  - S: 静态路由。
  - U: 表明该路由状态为 Up。
- **Interface:** 输出接口。
- **Timestamp:** 时间戳，表示 FIB 生成时间与系统启动时间的相对值。它不用于转发，但在分布式系统中可作为故障诊断与排除时的参考信息，可以验证 FIB 项是否从主控板到 I/O 板定时刷新，同时还可以用于内部老化功能。

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

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

例如，一台路由器上的路由表如下：

Route Flags: R - relied, D - download to fib

| Routing Tables: Public |        |     |            |       |           |                      |
|------------------------|--------|-----|------------|-------|-----------|----------------------|
| Destinations : 5       |        |     | Routes : 5 |       |           |                      |
| Destination/Mask       | Proto  | Pre | Cost       | Flags | NextHop   | Interface            |
| 0.0.0.0/0              | Static | 60  | 0          | RD    | 120.0.0.2 | GigabitEthernet0/0/5 |
| 8.0.0.0/8              | RIP    | 100 | 3          | D     | 120.0.0.2 | GigabitEthernet0/0/5 |
| 9.0.0.0/8              | OSPF   | 10  | 50         | D     | 20.0.0.2  | Ethernet4/0/0        |
| 9.1.0.0/16             | RIP    | 100 | 4          | D     | 120.0.0.2 | GigabitEthernet5/0/1 |
| 20.0.0.0/8             | Direct | 0   | 0          | D     | 20.0.0.1  | Ethernet1/0/0        |

 说明

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

一个目的地址是 9.1.2.1 的报文进入路由器，查找对应的 FIB 表。

FIB Table:

Total number of Routes : 5

| Destination/Mask | Nexthop   | Flag | TimeStamp | Interface            | TunnelID |
|------------------|-----------|------|-----------|----------------------|----------|
| 0.0.0.0/0        | 120.0.0.2 | SU   | t[37]     | GigabitEthernet0/0/5 | 0x0      |
| 8.0.0.0/8        | 120.0.0.2 | DU   | t[37]     | GigabitEthernet0/0/5 | 0x0      |
| 9.0.0.0/8        | 20.0.0.2  | DU   | t[9992]   | Ethernet4/0/0        | 0x0      |
| 9.1.0.0/16       | 120.0.0.2 | DU   | t[9992]   | GigabitEthernet5/0/1 | 0x0      |
| 20.0.0.0/8       | 20.0.0.1  | U    | t[9992]   | Ethernet1/0/0        | 0x0      |

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

最终，路由器会选择最长匹配 9.1.0.0/16 表项，从接口 GigabitEthernet 5/0/1 转发这条目的地址是 9.1.2.1 的报文。

## FIB 的机制

FIB 的操作包括两个单独的部分，控制平面（FibAgent）和转发平面（FibContainer）。

控制平面负责在创建 FIB 表前与路由管理的接口工作，将 FIB 下载到转发引擎，对于分布式系统，还需要将 FIB 下载到 I/O 板。转发平面是利用 FIB 模块提供的接口进行 FIB 表项查找，然后根据查找的结果指导转发。每个 FIB 表项中都指明报文到某网段或者某主机的出接口，及下一跳地址或者目的主机地址，这些都是通过转发平面的指导转发作用实现的。

## 6.9 IPv6 单播策略路由

策略路由是一种依据用户制定的策略进行路由选择的机制，可应用于安全、负载分担等目的。

### 6.9.1 介绍

#### 定义

策略路由（Policy-based routing，简称 PBR）与单纯依照 IPv6 报文的目的地址查找 IPv6 FIB 表进行转发不同，是一种依据用户制定的策略进行路由选择的机制，可应用于安全、负载分担等目的。策略路由支持基于到达报文的源地址、报文长度等信息进行路由选择，能够灵活地指定路由。

报文到达后，系统首先根据策略路由转发，若没有配置策略路由或者配置了策略路由但找不到匹配的表项，再根据 IPv6 FIB（Forwarding Information Base）表来转发报文，详细介绍请参见 [6.9.4.1 IPv6 FIB](#)。

#### 目的

策略路由支持基于 IPv6 ACL 进行报文过滤，灵活地指定路由，可以根据报文的源 IPv6 地址、目的 IPv6 地址、协议、端口号、优先级、时间段等各种丰富的信息将报文分类，然后控制将这些报文按照不同的路由进行转发。

## 6.9.2 参考标准和协议

与策略路由相关的参考标准与协议如下：

- RFC 2460: Internet Protocol, Version 6 (IPv6) Specification

## 6.9.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 6.9.4 原理描述

### 6.9.4.1 IPv6 FIB

#### IPv6 FIB 简介

设备转发 IPv6 报文的关键是 IPv6 FIB 表，在系统中报文转发时查找的是 FIB 表而非路由表。这是因为路由表表示所有的有效路由所形成的表项，并不指导转发。IPv6 FIB 表是网络层用来控制数据报发送的。

IPv6 FIB 表中每条转发项都指明 IPv6 报文到某网段或某主机应通过设备的哪个物理接口发送。发送后，报文可到达该路径的下一个设备，或者不再经过别的设备而传送到直接相连的网络中的目的主机。

IPv6 FIB 中包含了设备在转发报文时所必需的一组最小信息。一个 FIB 条目中一般包括目的地址、前缀长度、下一跳地址、标明路由特征的标志以及时间戳。

#### IPv6 FIB 工作原理

IPv6 FIB 表主要包含以下元素：

- Destination address: IPv6 报文发送的目的网络地址或主机地址。
- Prefix length: 目的地址前缀长度，可确定目的地址是否对应网络或主机。
- Nexthop: 为了将 IPv6 报文发送到目的地址所要经过的紧邻的下一跳地址。
- Flag(s): 标明路由特征。有以下一些标志：

- B: 黑洞路由, 即下一跳是空接口 (Null)。
- D: 动态路由。
- G: 网关。
- H: 下一跳为主机。
- S: 静态路由。
- U: 表明该路由状态为 Up。
- Interface: 出接口。

在 IPv6 路由表选择出路由后, 路由表会将激活路由下发到 IPv6 FIB 表中。当 IPv6 报文到达设备时, 会通过查找 IPv6 FIB 表进行转发。

### 6.9.4.2 策略路由工作原理

#### 策略路由概述

普通的 IPv6 报文转发是依据报文的目的地址查询 IPv6 FIB 表来实现的。但是在某些网络中, 有如下需求:

- 根据源 IPv6 地址来控制报文转发。即控制来自不同主机的报文通过不同的出接口进行转发。
- 根据 IPv6 报文的其他属性, 如协议、端口号、目的地址等来控制报文转发。

此时即可应用策略路由来满足以上需求。

策略路由基于转发平面, 为转发策略服务, 可以逐跳进行配置, 进而达到对报文转发的控制。策略路由与单纯依照 IPv6 报文的目的地址查找 IPv6 FIB 表进行转发不同, 是一种依据用户制定的策略进行路由选择的机制, 可应用于安全、负载分担等目的。

策略路由支持基于到达报文的源 IPv6 地址、目的 IPv6 地址等信息进行路由选择, 能够灵活地指定路由。

IPv6 报文到达后, 系统首先根据策略路由转发, 若没有配置策略路由或者配置了策略路由但找不到匹配的表项, 再根据 IPv6 FIB 表来转发报文。

#### 策略路由分类

策略路由可分为接口策略路由和本地策略路由。

- 接口策略路由在接口视图下配置, 作用于到达该接口的 IPv6 报文。对于一般的转发和安全等方面的需求, 使用接口策略路由。
- 本地策略路由在系统视图下配置, 对本设备产生的 IPv6 报文进行策略路由。

### 6.9.4.3 策略路由的应用

策略路由在网络中有很广泛的应用, 可以通过 **policy-based-route ipv6** 命令来实现策略路由的配置。配置策略路由分为两个步骤: 首先定义需要使用策略路由的 IPv6 报文; 然后为这些 IPv6 报文指定下一跳地址。

1. 使用 **if-match acl6** 子句定义需要使用策略路由的 IPv6 报文, 设置匹配条件。可以使用基本 IPv6 ACL 或者高级 IPv6 ACL。
2. 对于匹配 **if-match acl6** 子句条件的报文, 执行 **apply** 子句完成 IPv6 报文的转发。

可执行的动作如下：

- 设置 IPv6 报文转发的缺省下一跳。
- 设置 IPv6 报文转发的下一跳。

 说明

下一跳地址不能是设备本地 IPv6 地址。

## 6.10 路由策略

通过配置路由策略来影响网络流量经过的途径。目前实施路由策略的手段主要是对路由信息进行过滤。

### 6.10.1 介绍

#### 定义

路由策略（Routing Policy）主要实现了对路由的过滤、设置等控制功能，通过改变路由属性（包括可达性）改变网络流量所经过的途径。

#### 目的

USG 在发布、接收和引入路由信息时，根据实际组网需要实施一些策略，以便对路由信息进行过滤和改变路由信息的属性，如：

- 控制路由的发布  
只发布满足条件的路由信息。
- 控制路由的接收  
只接收必要、合法的路由信息，以控制路由表的容量，提高网络的安全性。
- 过滤和控制引入的路由  
一种路由协议在引入其它路由协议发现的路由信息丰富自己的路由知识时，只引入一部分满足条件的路由信息，并对所引入的路由信息的某些属性进行设置，以使其满足本协议的要求。
- 设置特定路由的属性  
为通过路由策略过滤的路由设置相应的属性。

### 6.10.2 可获得性

#### License 支持

本特性无需 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                      |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br><br>USG5500 V200R001C00 |

## 6.10.3 原理描述

### 6.10.3.1 路由策略的基本原理

路由策略的实现分为两个步骤：

1. 定义规则  
首先要定义将要实施路由策略的路由信息的特征，即定义一组匹配规则，可以以路由信息中的不同属性作为匹配依据进行设置，如目的地址、AS（Autonomous System）号等。
2. 应用规则  
根据设置的匹配规则，再将它们应用于路由的发布、接收和引入等过程的路由策略中。

目前提供了如下几种过滤器供路由协议引用：

- 访问控制列表  
访问控制列表包括针对 IPv4 报文的 ACL（Access Control List）和针对 IPv6 报文的 ACL6。用户在定义 ACL 时可以指定 IP 地址和子网范围用于匹配路由信息的目的网段地址或下一跳地址。
- 地址前缀列表  
地址前缀列表包括 IPv4 和 IPv6 地址前缀列表。  
一个地址前缀列表由前缀列表名标识。每个前缀列表可以包含多个表项，每个表项可以独立指定一个网络前缀形式的匹配范围，并用一个索引号（index-number）来标识，索引号指明了进行匹配检查的顺序。  
在匹配的过程中，路由器按升序依次检查由索引号标识的各个表项。只要有某一表项满足条件，就意味着本次匹配过程结束，而不再进行下一个表项的匹配。
- AS 路径过滤器  
BGP 的路由信息中，包含一个自治系统路径域。As-path-filter 就是针对自治系统路径域指定匹配条件。AS 路径过滤器仅应用于 BGP 协议。  
AS 路径属性参考文献：RFC1965。
- 团体属性过滤器  
团体属性过滤器 Community-filter 仅用于 BGP。BGP 的路由信息中，包含一个 community 属性域，用来标识一个团体。Community-filter 就是针对团体属性域指定匹配条件。

- 团体属性的参考文献：RFC1997。
- 扩展团体属性过滤器  
扩展团体属性过滤器 Extcommunity-filter 仅用于 BGP。目前只支持对 VPN 的 RT (Route-Target) 扩展团体属性的过滤。
  - 路由标识属性过滤器  
路由标识属性列表 Rd-filter 仅用于 BGP。路由标识属性过滤器就是针对 VPN 的 RD (Route Distinguisher) 属性指定匹配条件。

## 路由策略

路由策略最核心的内容就是它的匹配规则。

Route-policy 可以选择使用上述 6 种过滤器定义自己的匹配规则。一个 Route-policy 可以由多个节点 (node) 构成，不同节点之间是“或”的关系。系统按节点序号依次检查各个节点，如果通过了其中一节点，就意味着通过该策略，不再对其它节点进行匹配。

每个节点可以由一组 If-match 和 Apply 子句组成。If-match 子句定义匹配规则，匹配对象是路由信息的一些属性。同一节点中的不同 If-match 子句是“与”的关系，只有满足节点内所有 If-match 子句指定的匹配条件，才能通过该节点的匹配。Apply 子句指定动作，也就是在通过节点的匹配后，对路由信息的一些属性进行设置。

节点的匹配模式有两种：

- 允许模式 (permit)：当路由项满足该节点的所有 If-match 子句时被允许通过该节点的过滤，并执行该节点的 Apply 子句，如路由项不满足该节点的 If-match 子句，将继续匹配下一个节点。
- 拒绝模式 (deny)：当路由项满足该节点的所有 If-match 子句时，被拒绝通过并且不再匹配其他节点。

### 6.10.3.2 路由策略与策略路由的比较

路由策略 (Routing Policy) 是为了改变网络流量所经过的途径而对路由信息采用的方法。主要通过应用路由属性 (包括可达性) 来实现。路由器在发布和接收路由时应用的策略，目的是通过应用路由属性而过滤路由。目前实施路由策略的手段主要是对路由信息进行过滤。

策略路由 PBR (Policy Based Routing) 与单纯依照 IP 报文的目的地址查找 FIB 表进行转发不同，是一种依据用户制定的策略而进行路由选择的机制。PBR 支持基于到达报文的源地址等信息，依据用户制定的策略进行路由选择，可应用于安全、负载分担等目的。

路由策略与策略路由是两种不同的机制，主要区别如表 6-40。

表 6-40 路由策略与策略路由的比较

| 特性   | 特点                                                                                                                                                |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| 路由策略 | <ul style="list-style-type: none"><li>● 基于目的地址按路由表转发。</li><li>● 基于控制平面，为路由协议和路由表服务。</li><li>● 与路由协议结合完成策略。</li><li>● 应用命令 route-policy。</li></ul> |

| 特性   | 特点                                                                                                                                                                        |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 策略路由 | <ul style="list-style-type: none"><li>● 基于策略的转发，失败后再查找路由表转发。</li><li>● 基于转发平面，为转发策略服务。</li><li>● 需要手工逐跳配置，以保证报文按策略转发。</li><li>● 应用命令 <b>policy-based-route</b>。</li></ul> |

6.10.3.3 过滤器的比较

综合上面的简介，表 6-41 给出各种过滤器的应用范围和匹配条件的对比。

表 6-41 过滤器的比较

| 过滤器       | 应用范围    | 匹配条件                                |
|-----------|---------|-------------------------------------|
| 访问控制列表    | 各动态路由协议 | 源地址、目的地址、下一跳                        |
| 地址前缀列表    | 各动态路由协议 | 源地址、目的地址、下一跳                        |
| 路由策略      | 各动态路由协议 | 目的地址、下一跳、度量值、接口信息、路由类型、ACL、地址前缀列表等。 |
| AS 路径过滤器  | BGP 协议  | AS 路径属性                             |
| 团体属性过滤器   | BGP 协议  | 团体属性                                |
| 扩展团体属性过滤器 | VPN     | 扩展团体属性 RT                           |
| RD 属性过滤器  | VPN     | RD 属性                               |

6.10.3.4 地址前缀列表的组成和匹配规则

一个地址前缀列表由前缀列表名标识。每个前缀列表可以包含多个表项，每个表项可以单独指定一个网络前缀形式的匹配范围，并用一个索引号来标识，索引号指明了进行匹配检查的顺序。

在 USG 中，用来配置地址前缀列表的命令如下：

```
ip ip-prefix ip-prefix-name [ index index-number ] { permit | deny } ip-address mask-length  
[ greater-equal greater-equal-value ] [ less-equal less-equal-value ]
```

前缀列表名和列表项

列表名 *ip-prefix-name* 用来标识一个地址前缀列表。一个地址前缀列表里可以包含多个列表项。例如，名称为 *abcd* 的地址前缀列表包含三个列表项。其中索引号是 10 的表项允许地址前缀是 1.0.0.0，掩码长度是 8 的路由通过匹配；索引号是 20 的表项允许前缀是 2.0.0.0，掩码长度是 24 的路由通过匹配；索引号是 30 的表项拒绝前缀是 3.0.0.0，掩码长度是 8 的路由通过匹配。



```
ip ip-prefix abcd index 10 permit 1.0.0.0 8
ip ip-prefix abcd index 20 permit 2.0.0.0 24
ip ip-prefix abcd index 30 deny 3.0.0.0 8
```

## 索引号

每一个列表项都用索引号 `index` 进行了标识。索引号指明了列表匹配的顺序。

如上面的例子中，表项进行匹配检查的顺序是：按索引号的升序依次执行。

索引号可以手工指定，也可以由系统自动生成。系统自动生成的第一个索引号是 10，该序号值按照配置的顺序依次递增，步长是 10。例如，某地址前缀列表当前的索引号是 10、20、30、40。

- 新创建的表项如果手工指定索引号是 37，则系统会生成索引号是 37 的表项，当前索引号的顺序变为：10、20、30、37、40。
- 新创建的表项如果没有指定索引号，则系统会自动生成索引号是 50 的表项，当前索引号的顺序变为：10、20、30、40、50。

## 地址前缀

地址前缀列表进行匹配的依据有两个：掩码长度和掩码长度范围。

- 掩码长度

地址前缀列表匹配的对象是 IP 地址前缀，前缀由 IP 地址和掩码长度共同定义，即由 `ip-address mask-length` 确定。

例如 10.1.1.1 16，掩码长度是 16，这个地址的有效前缀为 16 位，即 10.1.0.0。

- 掩码长度范围

对于前缀相同，掩码不同的路由，可以通过设置掩码长度范围进行匹配。即根据 `[ greater-equal greater-equal-value ] [ less-equal less-equal-value ]` 所定义的掩码长度范围进行匹配。**greater-equal** 的意思是“大于等于”；**less-equal** 的意思是“小于等于”。需要按照以下关系指定掩码长度范围： $greater-equal-value \leq mask-length \leq less-equal-value \leq 32$ 。

- 如果只配置了 **greater-equal**，则掩码长度范围在 `greater-equal-value` 和 32 之间。
- 如果只配置了 **less-equal**，则掩码长度范围在 `mask-length` 和 `less-equal-value` 之间。
- 如果同时配置了 **greater-equal** 和 **less-equal**，则掩码长度范围在 `greater-equal-value` 和 `less-equal-value` 之间。

例如：

- 指定 `ip-address mask-length` 为 **0.0.0.0 0**，只匹配缺省路由。
- 指定 `ip-address mask-length` 为 **10.1.1.1 16**，匹配网段 10.1.0.0。
- 指定的地址前缀范围是 **0.0.0.0 0 less-equal 32**，匹配所有路由。
- 指定的地址前缀范围是 **10.1.1.0 16 great-equal 24**，则匹配 10.1.0.0 网段内，掩码长度在 24 到 32 之间的路由。
- 指定的地址前缀范围是 **10.1.1.0 16 less-equal 24**，则匹配 10.1.0.0 网段内，掩码长度在 16 到 24 之间的路由。
- 指定的地址前缀范围是 **10.1.1.0 16 great-equal 18 less-equal 30**，则匹配 10.1.0.0 网段内，掩码长度在 18 到 30 之间的路由。

## 地址前缀列表的匹配规则

- 顺序匹配  
在匹配过程中，系统按索引号升序依次检查各个表项。因此在指定表项索引号的时候，要注意符合期望匹配的顺序。
- 唯一匹配  
只要有一个表项满足条件，就认为通过该过滤列表，不再去匹配其他表项。
- 默认拒绝
  - 在配置了 **permit** 模式的表项后，默认其它所有不匹配的路由都为 **deny** 模式。则只有通过匹配的路由可以通过该过滤列表。
  - 在配置了 **deny** 模式的表项后，默认其它所有路由也都是 **deny** 模式，则即使有路由通过匹配，但都不能通过该过滤列表。因此，需要在多条 **deny** 模式的表项后定义一条 **permit 0.0.0.0 0 less-equal 32** 表项，允许其它所有路由信息通过。



说明

如果定义了一个以上的表项，则至少应该有一个表项的匹配模式为 **permit** 模式。

### 6.10.3.5 地址前缀列表的应用

在定义了地址前缀列表后：

- Route-Policy 过滤器可引用地址前缀列表作为匹配条件。
- 各路由协议可应用 **filter-policy** 命令引用地址前缀列表（也可以引用 ACL 列表），对接收和发布的路由进行过滤，仅接收或发布满足条件的部分路由。
  - 过滤接收的路由：**filter-policy import**
  - 过滤发布的路由：**filter-policy export**

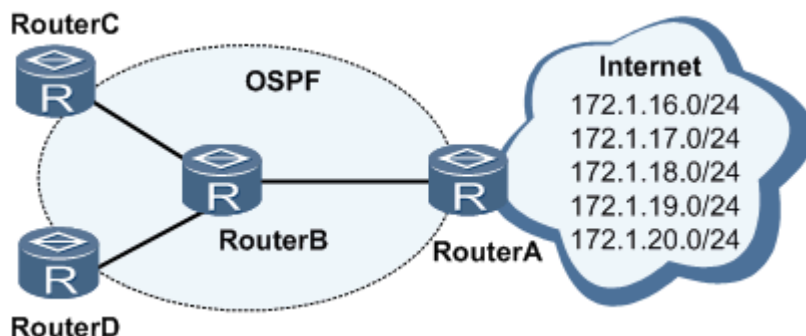
对于距离矢量协议和链路状态协议，**filter-policy** 命令的操作过程是不同的：

- 距离矢量协议  
距离矢量协议是基于路由表生成路由的。因此过滤器会影响：
  - 从邻居接收的路由
  - 向邻居发布的路由
- 链路状态协议  
链路状态协议是基于链路状态数据库来生成路由的，**filter-policy** 不影响链路状态通告或链路状态数据库的完整性，因此命令 **filter-policy import** 和 **filter-policy export** 的影响是不同的。
  - 对于 **filter-policy import**，该命令只能决定哪些路由从协议路由表加入到本地 IP 路由表，即只影响本地 IP 路由表，不影响协议路由表。
  - 对于 **filter-policy export**，该命令用来控制是否发布协议引入的其他路由协议的路由（如引入的 RIP 路由），但不影响向其他路由器发布的链路状态通告。

如图 6-86，运行 OSPF 协议的网络中，RouterA 从 Internet 网络接收路由，并为 RouterB 提供了部分 Internet 路由。

- RouterA 仅提供 172.1.17.0/24、172.1.18.0/24 和 172.1.19.0/24 给 RouterB
- RouterC 仅接收路由 172.1.18.0/24
- RouterD 接收 RouterB 提供的全部路由

图 6-86 对接收和发布的路由过滤组网图



在 RouterA 上配置地址前缀列表 a2b，该地址前缀列表包含三个表项，允许 172.1.17.0/24、172.1.18.0/24、172.1.19.0/24 通过匹配。

```
ip ip-prefix a2b index 10 permit 172.1.17.0 24
ip ip-prefix a2b index 20 permit 172.1.18.0 24
ip ip-prefix a2b index 30 permit 172.1.19.0 24
```

在 RouterA 上，在 OSPF 协议中使用 **filter-policy** 命令配置路由发布策略，应用地址前缀列表 a2b：

```
filter-policy ip-prefix a2b export
```

经过以上配置，RouterA 在发布路由的时候，就会对路由分别进行过滤，只有通过匹配的三条路由可以发布给 RouterB。

在 RouterC 上配置地址前缀列表 in，允许 172.1.18.0/24 通过匹配：

```
ip ip-prefix in index 10 permit 172.1.18.0 24
```

在 RouterC 上配置路由接收策略，应用地址前缀列表 in。

```
filter-policy ip-prefix in import
```

经过上面的配置，RouterC 在接收从 RouterB 发来的路由时，会应用策略进行过滤，只有符合条件的路由 172.1.18.0/24 才会被接收。

RouterD 上没有配置路由策略，可以接收 RouterB 发来的全部路由。

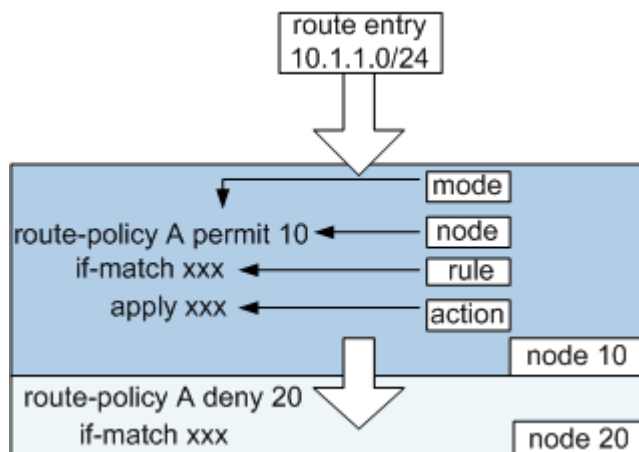
### 6.10.3.6 Route-Policy 的组成和匹配规则

Route-Policy 用来匹配给定的路由信息或者路由信息的某些属性（如下一跳、出接口等），并在条件满足时改变这些路由信息的属性。

一个 Route-Policy 的匹配条件可以使用其它过滤器（ACL、地址前缀列表、AS 路径过滤器、团体属性过滤器、扩展团体属性过滤器）。

如图 6-87 所示，一个 Route-Policy 可以由多个节点（node）构成，不同节点之间是“或”的关系（节点号的指定方法与地址前缀列表的索引号相同）。当一条路由由表项进入 Route-Policy 过滤器，系统按节点序号依次检查各个节点，如果通过了其中某一节点，就意味着通过该过滤器，不再对其他节点进行匹配。

图 6-87 Route-Policy 的组成



每个节点可以由一组 **if-match** 子句和 **apply** 子句组成。

- **if-match** 子句

定义匹配条件，匹配对象是路由信息的一些属性。

- 同一节点中的不同 **if-match** 子句是“与”的关系，只有满足节点内所有 **if-match** 子句指定的匹配条件，才能通过该节点的匹配。
- 匹配多个相同的匹配条件时，这多个相同的条件是“或”的关系。如：执行 **if-match interface** 的时候，如果匹配多个 interface，那么这些 interface 之间是“或”的关系，即只要匹配其中的一个 interface，就认为匹配上了。

- **apply** 子句

指定动作，满足 **if-match** 子句指定的过滤条件且匹配模式为 **permit** 的时候，对路由信息的一些属性进行设置。

**if-match** 子句和 **apply** 子句可根据应用进行设置，是可选的。

- 如果只需要过滤路由，不需要设置路由的属性，则不使用 **apply** 子句。
- 通常在多个 **deny** 节点后设置一个不含 **if-match** 子句和 **apply** 子句的 Route-Policy，用于允许其它的路由通过，请参见[默认拒绝](#)。
- 如果某个 **permit** 节点没有配置任何 **if-match** 子句，则该节点匹配所有的路由。

## 顺序匹配

如[图 6-87](#)所示，在匹配过程中，系统按节点号升序依次检查各个表项，因此在指定节点号时，要注意符合期望的匹配顺序。

## 唯一匹配

Route-Policy 各节点号之间是“或”的关系，只要通过一个节点的匹配，就认为通过该过滤器，不再进行其它节点的匹配。

## 匹配模式和匹配条件

节点的匹配模式有两种：**permit** 和 **deny**。

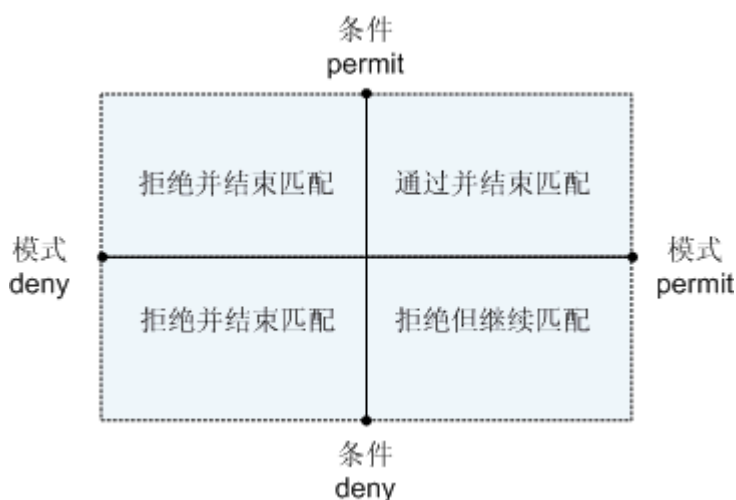
- **permit** 指定节点的匹配模式为允许。当路由项通过该节点的过滤后，将执行该节点的 **apply** 子句，不进入下一个节点；如果路由项没有通过该节点过滤，将进入下一个节点继续匹配。
- **deny** 指定节点的匹配模式为拒绝，这时 **apply** 子句不会被执行。当路由项满足该节点的所有 **if-match** 子句时，将被拒绝通过该节点，不进入下一个节点；如果路由项不满足该节点的 **if-match** 子句，将进入下一个节点继续匹配。

匹配条件通常是 **if-match** 子句，最终的匹配结果要综合以下两点：

- 节点的匹配模式（**permit** 或 **deny**）
- **if-match** 子句（如引用地址前缀列表）中包含的匹配条件（**permit** 或 **deny**）。

匹配的结果如图 6-88 所示：

图 6-88 Route-Policy 匹配结果



## 默认拒绝

USG 默认所有未匹配的路由将被拒绝通过 Route-Policy。

如果 Route-Policy 中定义了一个以上的节点，应保证各节点中至少有一个节点的匹配模式是 **permit**。因为 Route-Policy 用于路由信息过滤时：

- 如果某路由信息没有通过任一节点，则认为该路由信息没有通过该 Route-Policy。
- 如果 Route-Policy 的所有节点都是 **deny** 模式，则没有路由信息能通过该 Route-Policy。

## 匹配过程举例

Route-Policy 的名称是 **abc**，该过滤器有三个节点 10、20、30。

```
route-policy abc permit node 10
if-match ip-prefix prefix-a
if-match interface Ethernet6/0/0
apply cost 100
route-policy abc permit node 20
if-match cost 100
apply cost 200
```

```
route-policy abc deny node 30
  if-match as-path-filter 20
```

一条路由进入该过滤器，首先对节点 10 的条件进行匹配。

- 如果同时满足两个 **if-match** 子句设置的地址前缀列表 **prefix-a** 和接口 Ethernet6/0/0，那么 Route-Policy 会执行 **apply** 子句将这条路由的 **cost** 值修改为 100。并且路由不会再执行节点 20 和 30 的匹配。也就是说，这条路由已经通过过滤器，不会匹配节点 20 的条件，**cost** 值不会被修改为 200。
- 如果不满足节点 10 的匹配条件，路由才会执行节点 20 的匹配，如果满足条件，路由的 **cost** 值会修改为 200，通过策略过滤。
- 如果路由没有通过前两个节点的匹配，将执行节点 30 的匹配，如果匹配 AS 路径过滤器，这条路由将被拒绝，无法通过策略 **abc** 的过滤。
- 如果以上三个节点都没有通过匹配，那么这条路由被拒绝掉，无法通过这个过滤器。

### 6.10.3.7 Route-Policy 应用简介

路由协议在引入其它路由协议发现的路由时，通过应用 Route-Policy 过滤器只引入满足条件的路由信息，并为这些路由信息设置特定的属性。

USG 中提供了丰富的应用于路由协议的 **if-match** 条件和 **apply** 动作，见表 6-42 和表 6-43。灵活应用条件和动作的组合，可以实现不同的过滤效果。

表 6-42 If-match 条件

| 语句                                  | 规则              | 应用范围                            |
|-------------------------------------|-----------------|---------------------------------|
| <b>if-match acl</b>                 | ACL             | 全部路由协议                          |
| <b>if-match as-path-filter</b>      | AS 路径过滤器        | BGP                             |
| <b>if-match community-filter</b>    | 团体属性过滤器         | BGP                             |
| <b>if-match cost</b>                | 路由开销            | 全部路由协议                          |
| <b>if-match extcommunity-filter</b> | 扩展团体属性过滤器       | VPN                             |
| <b>if-match rd-filter</b>           | RD 属性过滤器        | VPN                             |
| <b>if-match interface</b>           | 出接口             | 全部路由协议                          |
| <b>if-match ip next-hop</b>         | IPv4 路由信息的下一跳地址 | 全部路由协议                          |
| <b>if-match ip route-source</b>     | IPv4 路由信息的源地址   | 动态路由协议                          |
| <b>if-match ip-prefix</b>           | 地址前缀列表          | BGP、IS-IS、OSPF、RIP 等            |
| <b>if-match ipv6 address</b>        | IPv6 路由信息的目的地址  | BGP4+、IS-IS (v6)、OSPFv3、RIPng 等 |
| <b>if-match ipv6 next-hop</b>       | IPv6 路由信息的下一跳地址 | BGP4+、IS-IS (v6)、OSPFv3、RIPng 等 |

| 语句                                | 规则            | 应用范围                            |
|-----------------------------------|---------------|---------------------------------|
| <b>if-match ipv6 route-source</b> | IPv6 路由信息的源地址 | BGP4+、IS-IS (v6)、OSPFv3、RIPng 等 |
| <b>if-match mpls-label</b>        | MPLS 标签       | BGP/MPLS IP VPN                 |
| <b>if-match route-type</b>        | 路由类型          | OSPF、IS-IS                      |
| <b>if-match tag</b>               | 路由标记 Tag 字段   | 全部路由协议                          |

表 6-43 Apply 动作

| 语句                               | 动作                             | 应用范围            |
|----------------------------------|--------------------------------|-----------------|
| <b>apply as-path</b>             | 替换或增加 AS_Path 属性               | BGP             |
| <b>apply backup-interface</b>    | 配置备份出接口                        | IP FRR          |
| <b>apply backup-nexthop</b>      | 配置备份下一跳                        | IP FRR、VPN FRR  |
| <b>apply behavior</b>            | 设置 QoS 流行为                     | VPN QoS         |
| <b>apply comm-filter delete</b>  | 删除团体属性过滤器中指定的 BGP 团体           | BGP             |
| <b>apply community</b>           | 设置团体属性                         | BGP             |
| <b>apply cost</b>                | 设置路由开销                         | 全部路由协议          |
| <b>apply cost-type</b>           | 设置路由开销类型                       | OSPF、IS-IS、BGP  |
| <b>apply extcommunity</b>        | 设置扩展团体属性                       | VPN             |
| <b>apply ip-address next-hop</b> | 设置 IPv4 路由的下一跳地址               | 全部路由协议          |
| <b>apply ip-precedence</b>       | 设置 QoS 参数 <b>ip-precedence</b> | VPN QoS         |
| <b>apply ipv6 next-hop</b>       | 设置 IPv6 路由的下一跳地址               | 全部路由协议          |
| <b>apply isis</b>                | 设置路由引入到 IS-IS 协议的 Level 级别     | IS-IS           |
| <b>apply local-preference</b>    | 设置 BGP 路由的本地优先级                | BGP             |
| <b>apply mpls-label</b>          | 分配 MPLS 标签                     | BGP/MPLS IP VPN |
| <b>apply origin</b>              | 设置 BGP 路由源                     | BGP             |
| <b>apply ospf</b>                | 设置路由引入到 OSPF 的特定区域             | OSPF            |
| <b>apply preference</b>          | 设置路由优先级                        | 全部路由协议          |
| <b>apply preferred-value</b>     | 指定 BGP 路由首选值                   | BGP             |
| <b>apply qos-local-id</b>        | 设置 QoS ID 值                    | VPN QoS         |

| 语句                         | 动作            | 应用范围   |
|----------------------------|---------------|--------|
| <b>apply tag</b>           | 设置路由标记        | 全部路由协议 |
| <b>apply traffic-index</b> | 设置 BGP 流量统计索引 | BGP    |



# 7 防火墙

## 关于本章

### 7.1 安全区域

引入“安全区域”的概念是为了对流量来源进行安全等级的划分，以判断何时对流量进行检测。本节将介绍安全区域的基本概念、规格及原理。

### 7.2 ACL

ACL 是设备实现流量控制的重要手段，它可以应用到包过滤、IPSec、QoS（Quality of Service）、路由策略等功能中。

### 7.3 ACL6

ACL6 包括基本 ACL6 和高级 ACL6，支持按照配置的顺序进行规则匹配和按照深度优先的顺序进行规则匹配。

### 7.4 安全策略

安全策略用于对流经设备或访问设备的流量进行安全检查、控制哪些流量可以通过设备或访问设备。本节将介绍安全策略的基本概念、应用场景及原理。

### 7.5 IPv6 包过滤

IPv6 包过滤主要是为了对网络中各种不同的 IPv6 流量是否转发做一个最基本的控制。本节将介绍 IPv6 包过滤的基本概念、规格及原理。

### 7.6 限流策略

限流策略通过为不同的 IP 分配不同的流量，达到优化网络的目的。本节介绍限流策略的基本概念、规格及原理。

### 7.7 NAT

NAT 能够实现 IP 报文中源地址或目的地址的转换，缓解 IPv4 地址资源枯竭的问题。本节介绍 NAT 的基本概念、规格、参考文档及原理。

### 7.8 NAT64

NAT64 技术可以实现 IPv6 与 IPv4 协议和地址的转换。

### 7.9 负载均衡

负载均衡可以分别利用各个服务器的处理能力，达到流量分担的目的，而且保障了服务器的可用性，得到最佳的网络扩展性。通常配置在内部网络的出口处。

### 7.10 攻击防范

攻击防范功能能够检测出多种类型的网络攻击，并能采取相应的措施保护内部网络免受恶意攻击，保证内部网络及系统的正常运行。本节介绍攻击防范的基本概念、规格及原理。

#### 7.11 IP 与 MAC 地址绑定

将 IP 地址与 MAC 地址进行绑定是防范 IP 地址冒用以及 ARP 欺骗等攻击的重要手段。此类攻击的特点就是攻击者的 IP 地址与 MAC 地址的对应关系不符合实际。本节将介绍 IP 与 MAC 地址绑定的基本概念、规格及原理。

#### 7.12 黑名单

黑名单是对 IP 地址进行屏蔽的最简单也是最高效的手段。本节将介绍黑名单的基本概念、规格及原理。

#### 7.13 会话表与长连接

会话表是一个记录 TCP、UDP、ICMP 等协议连接状态，实现报文正常转发的重要表项。长连接是保持部分特殊会话超长存活的手段。本节将介绍会话表和长连接的基本概念、规格及原理。

#### 7.14 Server-map 表与 ASPF

由于会话表对哪些报文属于同一条流量的标准过于严格，会导致一些特殊协议不能正确匹配会话表。Server-map 表可以解决这一问题。而 ASPF 功能通过对应用层协议的检测，最终也是生成 Server-map 表来实现特殊协议的转发。本节将介绍 Server-map 表项和 ASPF 的基本概念、规格及原理。

#### 7.15 端口映射

端口映射又称为“端口识别”，是为了系统能够将一些发给不知名端口的报文识别为知名协议，可结合 ASPF 使用。本节将介绍端口映射的基本概念、规格及原理。

#### 7.16 分片报文处理

网络设备在转发报文时，如果报文长度超过了链路 MTU，通常会对报文进行分片。本节介绍设备对分片报文的处理机制。

#### 7.17 GTP

GTP 是用于 GPRS 网络中 GSN 节点之间的接口协议，通过 GTP 协议实现 GPRS 网络中分组数据的传输。

#### 7.18 IDS 联动

IDS 联动指 IDS 设备能自动侦听整个网络中是否存在恶意攻击、入侵或其他安全隐患行为，通过下发指令的方式（如动态黑名单）通知 USG,由 USG 对攻击报文进行丢弃或其他处理。

#### 7.19 URPF

介绍了 URPF 的基本原理、特点和应用。

#### 7.20 终端安全管理（TSM 联动）

设备与 TSM 系统联动，主要是为了解决内网终端用户带来的安全隐患。本节介绍 TSM 联动的基本概念、规格、可获得性及原理。

## 7.1 安全区域

引入“安全区域”的概念是为了对流量来源进行安全等级的划分，以判断何时对流量进行检测。本节将介绍安全区域的基本概念、规格及原理。

### 7.1.1 介绍

#### 定义

一个安全区域是若干接口所连网络的集合，这些网络中的用户具有相同的安全属性。

#### 目的

在网络安全的应用中，如果网络安全设备对所有报文都进行逐包检测，会导致设备资源的大量消耗和性能的急剧下降。而这种对所有报文都进行检查的机制也是没有必要的。所以在网络安全领域出现了基于安全区域的报文检测机制。

引入安全区域的概念之后，网络管理员可以将具有相同安全级别的网络设备划入同一个安全区域。由于同一安全区域内的网络设备是“同样安全”的，USG 认为在同一安全区域内部发生的数据流动是不存在安全风险的，不需要实施任何安全策略。只有当不同安全区域之间发生数据流动时，才会触发设备的安全检查，并实施相应的安全策略。

所以 USG 在支持报文直接转发的基础上，还支持了安全区域的创建，并且允许网络管理员在安全区域的基础上实施各种特殊的报文检测与安全功能。

### 7.1.2 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

#### 特性依赖

如果需要创建除公网 VPN 实例外的 VPN 实例的安全区域，必须首先创建该 VPN 实例。

### 7.1.3 原理描述

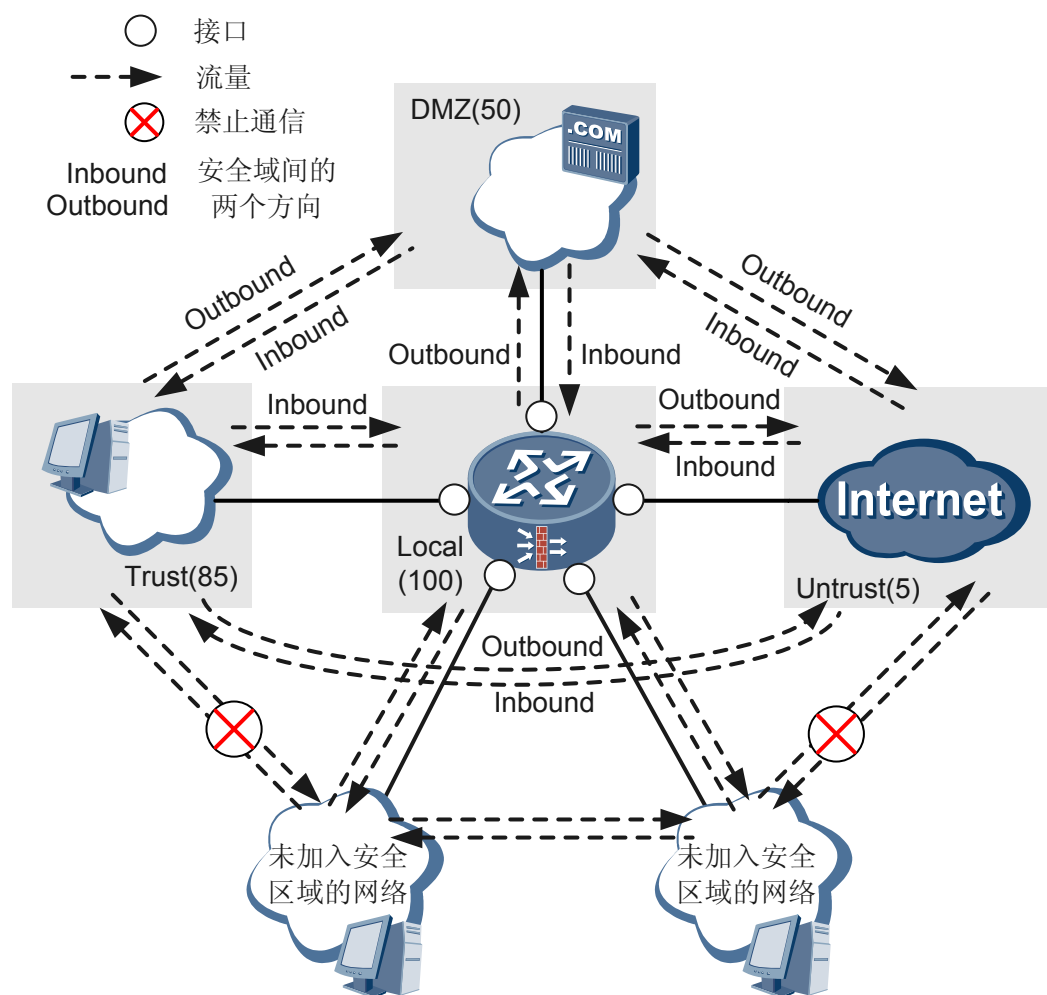
## 安全区域与安全级别

将不同的网络划入不同的安全区域，其实是从安全性的角度上对网络中的用户进行的一个分类。在 USG 中，必须以接口为单位来进行分类，即同一个接口所连网络的所有网络设备一定位于同一安全区域中，而一个安全区域可以包含多个接口所连的网络。

这里的接口既可以是物理接口，也可以是逻辑接口。所以可以通过子接口或者 VLAN IF 等逻辑接口实现将同一物理接口所连的不同网段的用户划入不同安全区域的功能。

需要注意的是，将接口加入安全区域这个操作，实际上意味着将该接口所连网络加入到安全区域中，而该接口本身仍然属于系统预留用来代表设备本身的 Local 安全区域，如图 7-1 所示。

图 7-1 安全区域及安全域间示意图



正因为安全区域的划分是基于对网络用户的安全性的判断，所以每个安全区域都有一个唯一的安全级别来对该区域的安全性做标识。安全级别通过 0 ~ 100 的数字表示，数字越大，则代表该区域内的网络设备越可信。

## 系统缺省预留的安全区域

如图 7-1 所示，系统缺省已经创建了四个安全区域。这些安全区域不可被删除，也不能修改其安全级别，其中 Local 域代表的是设备本身。如表 7-1 所示。

表 7-1 设备缺省划分的安全区域

| 区域名称            | 安全级别                  | 说明                                                                                                                                                                                                                                                                                                                                                                             |
|-----------------|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 非受信区域 (Untrust) | 低安全级别的安全区域，安全级别为 5。   | 通常用于定义 Internet 等不安全的网络。                                                                                                                                                                                                                                                                                                                                                       |
| 非军事化区域 (DMZ)    | 中等安全级别的安全区域，安全级别为 50。 | <p>通常用于定义内网服务器所在区域。因为这种设备虽然部署在内网，但是经常需要被外网访问，存在较大安全隐患，同时一般又不允许其主动访问外网，所以将其部署一个安全级别比 Trust 低，但是比 Untrust 高的安全区域中。</p> <p><b>说明</b></p> <p>DMZ 起源于军方，是介于严格的军事管制区和松散的公共区域之间的一种有着部分管制的区域。USG 设备引用了这一术语，指代一个逻辑上和物理上都与内部网络和外部网络分离的安全区域。</p> <p>DMZ 安全区域很好地解决了服务器的放置问题。该安全区域可以放置需要对外提供网络服务的设备，如 WWW 服务器、FTP 服务器等。上述服务器如果放置于内部网络，外部恶意用户则有可能利用某些服务的安全漏洞攻击内部网络；如果放置于外部网络，则无法保障它们的安全。</p> |
| 受信区域 (Trust)    | 较高安全级别的安全区域，安全级别为 85。 | 通常用于定义内网终端用户所在区域。                                                                                                                                                                                                                                                                                                                                                              |

| 区域名称            | 安全级别                               | 说明                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-----------------|------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 本地区域<br>(Local) | 最高安全<br>级别的安全区域，<br>安全级别<br>为 100。 | <p>Local 区域定义的是设备本身，包括设备的各接口本身。凡是由设备构造并主动发出的报文均可认为是从 Local 区域中发出，凡是需要设备响应并处理（而不仅是检测或直接转发）的报文均可认为是由 Local 区域接受。</p> <p>用户不能改变 Local 区域本身的任何配置，包括向其中添加接口。</p> <p><b>说明</b></p> <p>由于 Local 区域的特殊性，在很多需要设备本身进行报文收发中的应用中，需要开放对端所在安全区域与 Local 区域之间的安全策略。包括：</p> <ul style="list-style-type: none"><li>● 需要对设备本身进行管理的情况下。例如 Telnet 登录、Web 登录、接入 SNMP 网管等。</li><li>● 设备本身作为某种服务的客户端或服务器，需要主动向对端发起请求或处理对端发起的请求的情况下。例如 FTP，PPPoE 拨号，NTP 以及 IPSec VPN 等。</li></ul> <p>向安全区域中添加接口，只是认为该接口所连的网络属于该安全区域，而接口本身还是属于 Local 区域。</p> <p>例如，一台设备通过接口 1 与 Trust 区域相连。如果一个属于 Trust 区域的用户希望在 ping 接口 1 时得到设备处理和回应，以验证网络连通性时，就需要开放 Trust 区域与 Local 区域之间的包过滤策略，否则这些 ping 报文将被设备丢弃。</p> |

## 安全域间及域间方向

如图 7-1 所示，每两个安全区域又可以构成一个域间（Interzone）关系。一个安全区域创建之后，这个安全区域与其他安全区域之间的域间关系也就自动创建好了。

安全域间这个概念用来描述流量的传输通道。它是两个“区域”之间的唯一“道路”，如果希望对经过这条通道的流量进行控制，就必须在通道上设立“关卡”，也就是安全策略。但是对于不需要经过这条通道的流量，例如在同一个安全域间内转发的流量，在安全域间下发的安全策略是不起作用的。

任意两个安全区域都构成一个安全域间（Interzone），并具有单独的安全域间视图，大部分的安全策略都需要在安全域间视图下配置。

安全域间的数据流动具有方向性，包括入方向（Inbound）和出方向（Outbound）。

- 入方向：数据由低优先级的安全区域向高优先级的安全区域传输。
- 出方向：数据由高优先级的安全区域向低优先级的安全区域传输。

### 说明

在判断公网 VPN 实例与其他 VPN 实例的域间方向时，其他 VPN 实例的优先级总比公网 VPN 实例高。即入方向是指流量由公网 VPN 实例的区域流向其他 VPN 实例的区域，出方向是指流量由其他 VPN 实例的区域流向公网 VPN 实例的区域。

通常情况下，通信双方一定会交互报文，即安全域间的两个方向上都有报文的传输。而判断一条流量的方向应以发起该条流量的第一个报文为准。

## 安全区域对通信的影响

如图 7-1 所示，一个网络是否加入安全区域对这个网络和其他网络之间的通信有着非常重要的影响。首先介绍一下 Local 区域和一个接口所连网络的关系。由于 Local 区域代表的是设备本身，所以不能将一个设备所连的网络划入到 Local 区域中。

- 一个加入到安全区域的网络和本地区域：两者之间缺省不可通信，必须开放相应的域间包过滤规则才能实现流量在两者之间的流动。在开放规则之后，两者之间的流量也受到域间应用的安全策略的影响。关于包过滤的详细信息，请参见 7.4 安全策略。

除 Local 区域之外，其他任意两个网络之间的关系可以分为以下几种情况：

- 两个加入到同一安全区域的网络：两者可以直接通信，不受安全策略的影响。
- 两个加入到不同安全区域的网络：两者之间缺省不可通信，必须开放相应的域间包过滤规则才能实现流量在两者之间的流动。在开放规则之后，两者之间的流量也受到域间应用的安全策略的影响。关于包过滤的详细信息，请参见 7.4 安全策略。
- 一个加入到安全区域的网络和一个未加入安全区域的网络：两者之间不可通信。

### 说明

所有设备所连网络在缺省情况下未加入任何安全区域。必须通过将该网络与设备相连的接口加入到安全区域中，才能实现将这个网络划入到某个安全区域里。

## VPN 实例中的安全区域

VPN 实例是对设备的一个虚拟化，所以在创建了一个 VPN 实例后，相应的这个 VPN 实例的几个缺省安全区域也就创建好了。在使用安全区域时，如果不附加 VPN 实例名这个参数，都是指公网 VPN 实例的安全区域。

由于 USG 只支持公网 VPN 实例与其他 VPN 实例之间的报文转发，并不支持其他 VPN 实例之间的报文转发，所以在进入安全域间视图的时候，也就有以下原则：

- 可以进入公网 VPN 实例的任意两个安全区域的域间视图
- 可以进入公网 VPN 实例的一个安全区域与其他 VPN 实例的一个安全区域的域间视图
- 可以进入其他 VPN 实例的任意两个安全区域的域间视图
- 不可以进入分别属于除公网 VPN 实例外的两个 VPN 实例的两个安全区域的域间视图

## 7.2 ACL

ACL 是设备实现流量控制的重要手段，它可以应用到包过滤、IPSec、QoS（Quality of Service）、路由策略等功能中。

### 7.2.1 介绍

#### 定义

访问控制列表是一系列有顺序的规则组的集合，这些规则根据数据包的源地址、目的地址、端口号等来描述。ACL（Access Control List）通过规则对数据包进行分类，这些规则应用到 USG 上，USG 根据这些规则判断哪些数据包可以接收，哪些数据包需要拒绝。



例如可以用访问列表描述：拒绝任何用户终端使用 Telnet 登录本机。允许每个用户终端经由 SMTP 向本机发送电子邮件。

每个 ACL 中可以定义多个规则，ACL 规则是一个匹配选项的集合，由用户根据不同业务进行选择配置。

 说明

针对不同的业务，ACL 的匹配选项支持情况不相同。

分类

为了满足不同类型的安全功能的使用，设备支持多种不同类型的 ACL，如表 7-2 所示。

表 7-2 ACL 的分类

| 分类             | 编号范围        | 支持的过滤选项                                                                                                                                                                   |
|----------------|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 基本 ACL         | 2000 ~ 2999 | 匹配条件较少，只能通过源 IP 地址和时间段来进行流量匹配，在一些只需要进行简单匹配的功能可以使用。                                                                                                                        |
| 高级 ACL         | 3000 ~ 3999 | 匹配条件较为全面，通过源 IP 地址、目的 IP 地址、ToS、时间段、协议类型、优先级、ICMP 报文类型和 ICMP 报文码等多个维度来对进行流量匹配，在大部分功能中都可使用高级 ACL 来进行精确流量匹配。                                                                |
| 基于 MAC 地址的 ACL | 4000 ~ 4999 | 由于数据链路层使用 MAC 地址来进行寻址，所以在控制数据链路层帧时需要通过 MAC 地址来对流量进行分类。基于 MAC 地址的 ACL 就可以通过源 MAC 地址、目的 MAC 地址、CoS、协议码等维度来进行流量匹配。                                                           |
| 硬件包过滤 ACL      | 9000 ~ 9499 | 硬件过滤 ACL 是一种特殊的 ACL，主要用于硬件包过滤功能。将硬件包过滤 ACL 下发到接口卡上后，接口卡通过硬件实现包过滤功能，比普通的软件包过滤速度更快，消耗系统资源更少。硬件包过滤 ACL 的匹配条件比较全面，可以通过源 IP 地址、目的 IP 地址、源 MAC 地址、目的 MAC 地址、CoS、协议类型等维度来进行流量匹配。 |

目的

各种业务通过引用 ACL，对路由或报文进行规则匹配后处理。

7.2.2 可获得性

License 支持

不需要 License 支持。



## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 7.2.3 原理描述

### 步长

步长的含义是：自动为 ACL 规则分配编号时，规则编号之间的差值。例如，如果步长设定为 5，规则编号分配是按照 5、10、15 这样的规则分配的。使用步长设定可以方便在规则之间插入新的规则。例如配置了 4 条规则，规则编号为：5、10、15、20。此时希望在第一条规则（即编号为 5 的规则）之前插入一条规则，则可以使用 **rule 1** 命令在规则 5 之前插入一条编号为 1 的规则。

缺省情况下，ACL 规则的步长为 5。当步长改变的时候，ACL 规则组下的规则编号会自动重新排列。例如，本来规则编号为：5、10、15、20。如果把步长设定改为 2，则规则编号变成：2、4、6、8。

### ACL 规则的匹配过程

1. 首先查找用户是否配置了该 ACL（因为有些业务可能允许引用不存在的 ACL）。
2. 根据 ACL 的配置情况：
  - 如果存在 ACL 且需要根据规则来检查报文，则查找该 ACL 中所有规则，只要有一条规则和报文匹配，就直接通知给业务规则匹配的动作，不再继续查找后续的规则。
  - 如果存在 ACL 且业务只匹配源地址信息、目的地址信息、IP 承载的协议类型、TCP 的源端口、目的端口、ICMP 协议的类型中的某些选项，则根据业务要求查找对应存在的 ACL，进行规则匹配。匹配的第一条就通知给业务，不再继续查找后续的规则。

### ACL 规则的匹配顺序

规则匹配方式有两种：配置顺序和自动顺序。

- 配置顺序  
配置顺序根据 ACL 规则的 ID 进行排序，ID 小的规则排在前面，优先进行匹配。当找到第一条匹配条件的规则时，查找结束。系统按照该规则对应的动作处理。
- 自动顺序  
自动顺序也叫深度优先匹配。此时 ACL 规则的 ID 由系统自动分配，规则中指定数据包范围小的排在前面，优先进行匹配。当找到第一条匹配条件的规则时，查找结束。系统按照该规则对应的动作处理。
  - 对于基本访问控制规则的语句，直接比较源地址通配符，通配符相同的则按配置顺序。

- 对于高级访问控制规则，首先比较协议范围，再比较源地址通配符，相同时再比较目的地址通配符，仍相同时则比较端口号的范围，范围小的排在前面，如果端口号范围也相同则按配置顺序。



说明

如果所有规则都没有匹配，系统不进行任何处理（等同于没有配置 ACL）。

## 7.3 ACL6

ACL6 包括基本 ACL6 和高级 ACL6，支持按照配置的顺序进行规则匹配和按照深度优先的顺序进行规则匹配。

### 7.3.1 介绍

#### 定义

IPv6 ACL 指用于过滤 IPv6 报文的访问控制列表。

每个 IPv6 ACL 中可以定义多个规则，根据规则的功能分为基本 ACL 规则和高级 ACL 规则。IPv6 ACL 规则是一个匹配选项的集合，由用户根据不同业务进行选择配置。

IPv6 ACL（Access Control List IPv6）的类型划分方式有两种，如表 7-3。

表 7-3 IPv6 ACL 的分类

| ACL 类型划分依据   | ACL 类型                                                                                                                                                                                                                                               |
|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 按照 ACL 规则的功能 | <ul style="list-style-type: none"><li>● 基本 IPv6 ACL：使用源地址信息做为定义 IPv6 ACL 规则的元素,编号的取值范围是 2000 ~ 2999。</li><li>● 高级 IPv6 ACL：使用数据包的源地址信息、目的地址信息、IPv6 承载的协议类型、针对协议的特性，例如 TCP 的源端口、目的端口，ICMPv6 协议的类型、ICMPv6 Code 等内容定义规则，编号的取值范围是 3000 ~ 3999。</li></ul> |

#### 目的

各种业务通过引用 IPv6 ACL，对路由或报文进行规则匹配后再处理。

### 7.3.2 可获得性

#### License 支持

不需要 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 7.3.3 IPv6 ACL

### IPv6 ACL 的规则匹配过程

1. 首先查找用户是否配置了 IPv6 ACL。
2. 根据 IPv6 ACL 的配置情况：
  - 如果存在 IPv6 ACL 且需要根据规则来检查报文，则查找该 IPv6 ACL 中规则，只要有一条规则和报文匹配，就直接通知给业务规则匹配的动作，不再继续查找后续的规则。
  - 如果存在 IPv6 ACL 且业务只匹配源地址信息、目的地址信息、IPv6 承载的协议类型、源端口、目的端口、ICMP6 协议的类型中的某些选项，则根据业务要求查找所有 IPv6 ACL，进行规则匹配。匹配的第一条就通知给业务，不再继续查找后续的规则。

### IPv6 ACL 的规则匹配顺序

一个 IPv6 ACL 中可以包含多个规则，而每个规则都指定不同的匹配选项。在将一个报文和 IPv6 ACL 的规则进行匹配的时候，首先需要确定规则的匹配顺序。

IPv6 ACL 支持两种匹配顺序：

- 配置顺序：按照用户配置规则的先后顺序进行匹配。
- 自动排序：按照“深度优先”的顺序进行规则匹配。

在报文匹配规则时，会按照匹配顺序去匹配定义的规则，一旦有一条规则被匹配，报文就不再继续匹配其它规则了，设备将对该报文执行第一次匹配的规则指定的动作。

### IPv6 ACL 与 ACL 的区别

IPv6 ACL 和 ACL 的基本原理是一样的，差别很小，如表 7-4

表 7-4 IPv6 ACL 与 IPv4 ACL 的区别

| IPv4 ACL | IPv6 ACL |
|----------|----------|
| 支持 MIB。  | 无 MIB。   |

## 7.4 安全策略

安全策略用于对流经设备或访问设备的流量进行安全检查、控制哪些流量可以通过设备或访问设备。本节将介绍安全策略的基本概念、应用场景及原理。

### 7.4.1 介绍

#### 定义

安全策略是按一定规则检查数据流是否可以通过防火墙的基本安全控制机制。

#### 目的

防火墙的基本作用是保护特定网络免受“不信任”的网络的攻击，但是同时还必须允许两个网络之间可以进行合法的通信。安全策略的作用就是对通过防火墙的数据流进行检验，符合安全策略的合法数据流才能通过防火墙，主要应用如下：

- 对跨防火墙的网络互访进行控制  
例如控制内网访问外网的权限、控制内网不同安全级别的子网间的访问权限等。
- 对设备本身的访问进行控制  
例如限制哪些 IP 地址可以通过 Telnet 和 Web 等方式登录设备，控制网管服务器、NTP 服务器等与设备的互访等。

安全策略有多种类型，适用场合不同：

- 域间或域内安全策略：用于控制域间或域内的流量，此时的安全策略既有传统包过滤功能，也有对流量进行 IPS、AV、应用控制等进一步的应用层检测的作用。域间或域内安全策略是包过滤、UTM 应用层检测等多种安全检查同时实施的一体化策略。
- 应用在接口上的包过滤规则：用于控制接口的流量，就是传统的包过滤功能，基于 IP、MAC 地址等二、三层报文属性直接允许或拒绝报文通过。包括基于 MAC 地址的包过滤和硬件包过滤。

### 7.4.2 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持



说明

不同产品支持的安全策略不同，请参见《特性支持》中的防火墙特性。

| 产品                                                                          | 功能                                                  | 最低支持版本                                                                                                           |
|-----------------------------------------------------------------------------|-----------------------------------------------------|------------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/<br>HSR | 域间安全策略<br>域内安全策略<br>接口包过滤<br>基于 MAC 地址的包过滤<br>硬件包过滤 | USG2110-X/<br>2100/2200/5100<br>V100R005C00、<br>USG2100/2200/5100 BSR/<br>HSR V100R005C00<br>USG5500 V200R001C00 |

## 硬件要求

硬件包过滤只有 5FSW，8FE+2GE，18FE+2SFP，16GE+4SFP 二层接口卡上的接口支持。

## 7.4.3 特性增强

| 版本       | 特性增强                                                                                                                                               |
|----------|----------------------------------------------------------------------------------------------------------------------------------------------------|
| V300R001 | 域间或域内策略中新增用户/用户组作为匹配条件，匹配上安全策略的用户将获取策略中指定的权限，例如 Web 过滤中允许用户访问哪些网站、邮件过滤中允许用户接收哪些邮件等。<br><br>域间或域内策略中新增对应用控制策略（DPI 策略）的引用，也就是应用控制功能也通过在域间或域内策略上应用生效。 |

## 7.4.4 应用场景：通过域间安全策略进行安全检查

### 7.4.4.1 控制流量转发

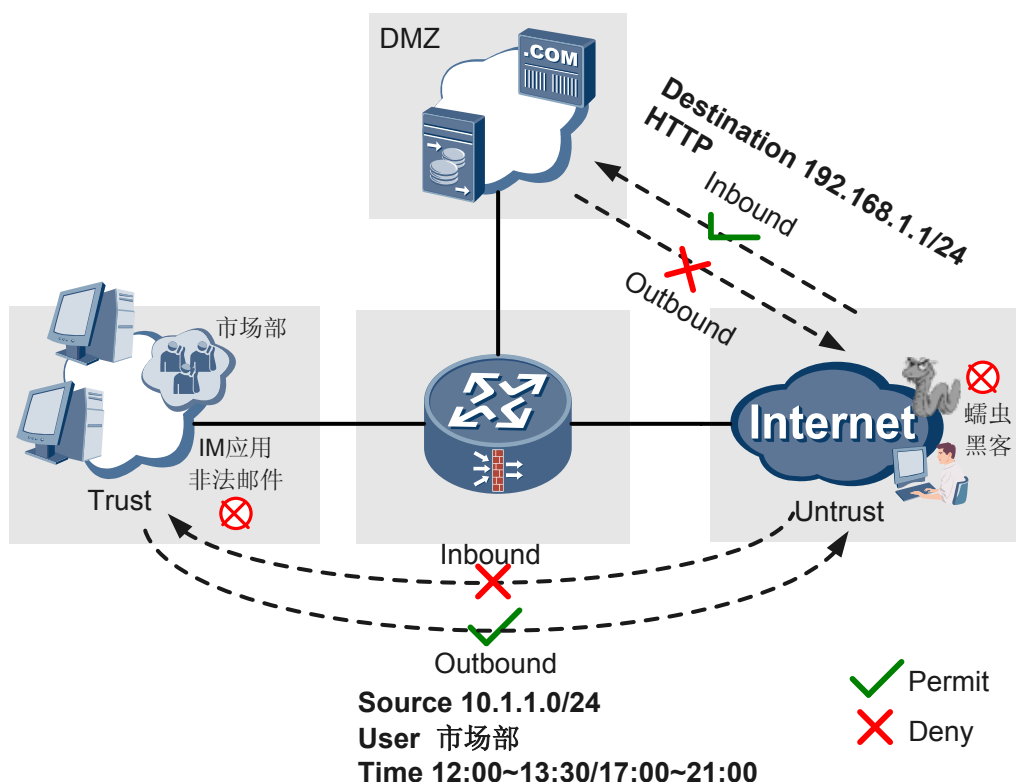
#### 应用环境

域间安全策略用于控制域间流量的转发（此时称为转发策略），适用于接口加入不同安全区域的场景。域间安全策略可以按 IP 地址、时间段和服务（端口或协议类型）、用户等多种方式匹配流量，并对符合条件的流量进行包过滤控制（permit/deny）或高级的 UTM 应用层检测。

#### 典型应用

图 7-2 是对企业上网、服务访问流量进行安全检查，防御威胁的典型应用。域间安全策略是包过滤、UTM 应用层检测等多种安全检查同时实施的一体化策略。

图 7-2 域间转发策略典型应用



转发策略为企业 provide 全方位的安全防护：

- 企业需要限制只有市场部员工可以上网，而且只能在休息时间上网，同时外网不可以主动访问内网。  
可以在 Trust 和 Untrust 域间的 Outbound 方向上应用基于源 IP、用户帐号和时间段的转发策略，同时在 Inbound 方向上应用禁止所有流量通过的转发策略。
- 企业需要限制外网只可以访问企业 DMZ 区域的对外提供 Web 服务的服务器，同时 Web 服务器不能主动访问外网。  
可以在 DMZ 和 Untrust 域间的 Inbound 方向上应用基于目的 IP 和服务协议（HTTP）的转发策略，同时在 Outbound 方向上应用禁止所有流量通过的转发策略。
- 在限制基本访问的同时，还要进行阻断 IM 应用、拦截非法邮件、禁止访问非法网站、防御蠕虫黑客攻击等高级的应用层控制，保护企业上网环境。  
可以在转发策略中引用应用控制策略、邮件过滤策略、URL 过滤策略和 IPS 策略等 UTM 策略进行应用层检测。

## 7.4.4.2 控制设备访问

### 应用环境

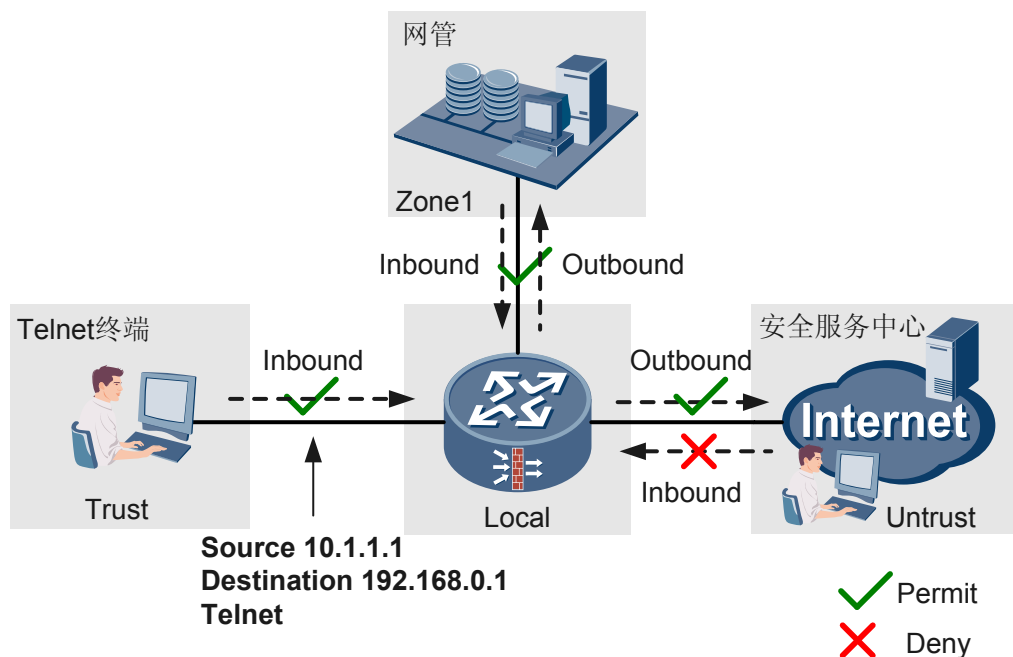
域间安全策略用于控制外界与设备本身的互访（此时称为本地策略），适用于接口已经加入安全区域的场景。域间安全策略可以按 IP 地址、时间段和服务（端口或协议类型）等多种方式匹配流量，并对符合条件的流量进行包过滤控制（permit/deny），允许或拒绝与设备本身的互访。

本地策略就是与 Local 域相关的域间包过滤功能。

## 典型应用

控制对设备本身的安全访问，如图 7-3 所示。

图 7-3 域间本地策略典型应用



与设备本身进行交互的场景都需要配置 Local 域与对端所在安全区域的本地策略：

- 控制登录设备的权限：图 7-3 中控制只有 10.1.1.1 这个 IP 地址可以使用 Telnet 方式登录设备，需要在 Trust 到 Local 的 Inbound 方向应用限制指定 IP 地址和 Telnet 协议的本地策略。同时还要控制不允许外网登录设备，在 Untrust 到 Local 的 Inbound 方向应用禁止任何流量访问设备的本地策略。
- 控制与设备的交互访问：升级特征库需要在 Local 到 Untrust 的 Outbound 方向应用允许设备访问安全服务中心的本地策略；使用网管系统管理设备时需要在 Local 到网管服务器所在安全区域的 Inbound 和 Outbound 方向都应用本地策略，允许网管和设备互访。

### 说明

很多需要与设备本身进行交互的应用，都需要配置 Local 域与对端所在安全区域的域间策略。除图 7-3 中提到的情况外，还包括设备本身作为某种服务的客户端或服务，需要主动向对端发起请求或处理对端发起的请求的情况，例如 FTP、NTP 以及 IPSec VPN 等。另外还有一些特性涉及到将逻辑接口加入安全区域并配置本地策略的步骤，具体参见对应特性的配置过程，不再一一列举。

## 7.4.5 应用场景：通过域内安全策略进行安全检查

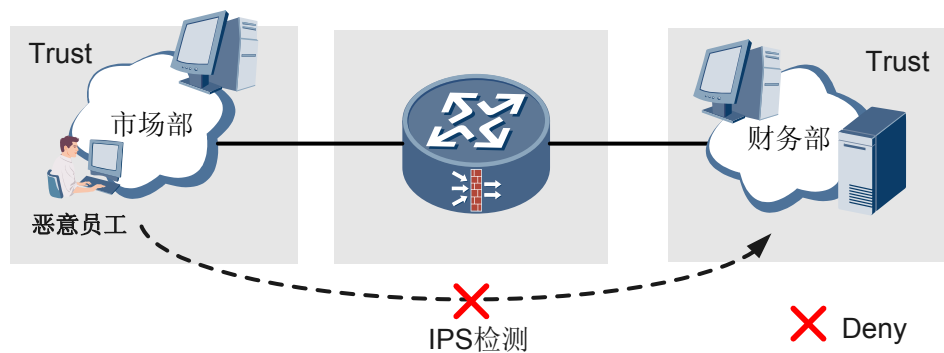
### 应用环境

缺省情况下域内数据流动不受限制，如果需要进行安全检查可以应用域内安全策略。与域间安全策略一样可以按 IP 地址、时间段和服务（端口或协议类型）、用户等多种方式匹配流量，然后对流量进行安全检查。

### 典型应用

如图 7-4 所示，市场部和财务部都属于内网所在的安全区域 Trust，可以正常互访。但是财务部是企业重要数据所在的部门，需要防止内部员工对服务器、PC 等的恶意攻击。所以在域内应用安全策略进行 IPS 检测，阻断恶意员工的非法访问。

图 7-4 域内安全策略典型应用



## 7.4.6 应用场景：通过接口上的包过滤规则控制报文的收发

### 7.4.6.1 控制接口收发 IP 报文

### 应用环境

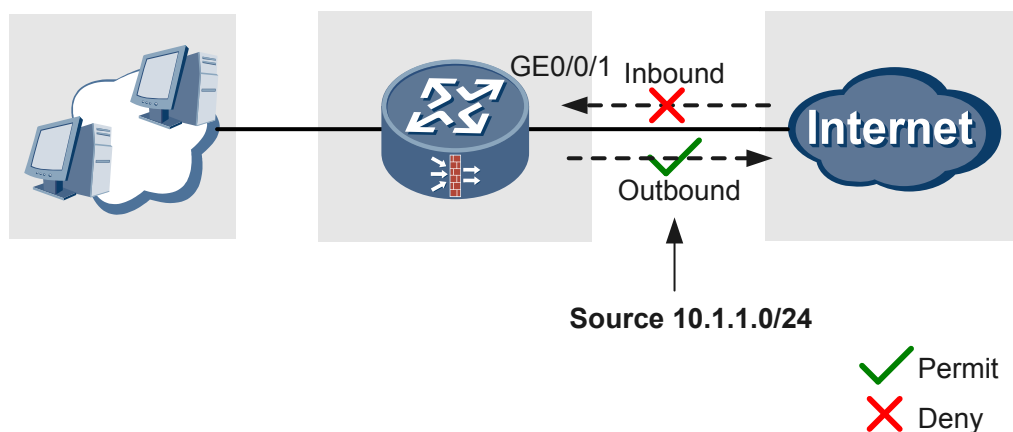
当接口未加入安全区域的情况下，可以通过接口包过滤控制接口接收和发送的 IP 报文，可以按 IP 地址、时间段和服务（端口或协议类型）等多种方式匹配流量并执行相应动作（permit/deny）。

### 典型应用

接口包过滤的典型应用如图 7-5 所示。



图 7-5 接口包过滤典型应用



在 GigabitEthernet 0/0/1 的 Inbound 方向上，也就是接口接收报文的方向上应用接口包过滤，拒绝接收外网对内网的访问流量。

在 GigabitEthernet 0/0/1 的 Outbound 方向上，也就是接口发送报文的方向上应用接口包过滤，只允许发送 10.1.1.0/24 这个网段访问外网的流量。

## 7.4.6.2 控制接口接收的以太网帧

### 应用环境

基于 MAC 地址的包过滤用来控制接口可以接收哪些以太网帧，可以按 MAC 地址、帧的协议类型和帧的优先级匹配流量并执行相应动作（permit/deny）。

如果在已经应用了域间包过滤的情况下，还需要拒绝某些 MAC 地址的报文，此时可以在接口上应用基于 MAC 地址的包过滤。如果报文被基于 MAC 地址的包过滤拒绝就直接丢弃了，不再进行域间包过滤处理。

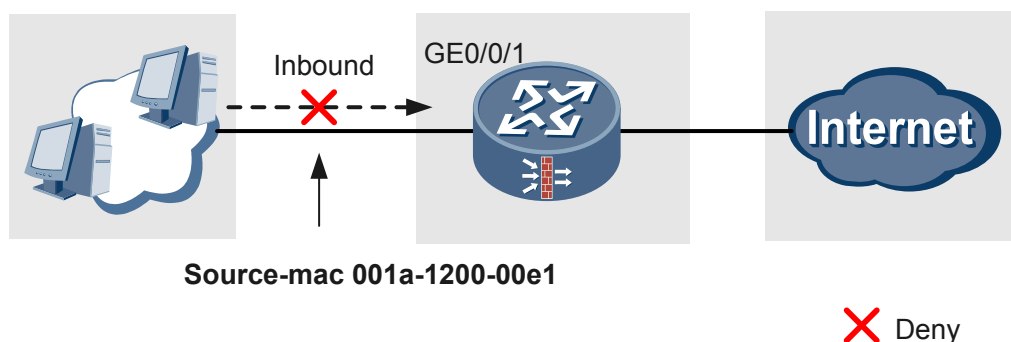
### 典型应用

基于 MAC 地址的包过滤的主要应用如下：

- 通过 MAC 地址控制访问权限，比如限制某些 MAC 地址的 PC 才能上网。
- 避免攻击，比如某个主机经常发 ARP 攻击，可以通过 MAC 过滤将指定 MAC 地址的主机屏蔽掉。
- 根据以太网帧中的协议类型（TPID 字段）过滤某些报文，比如不允许 VLAN、MPLS 报文通过等。

通过 MAC 地址控制上网权限的典型应用如图 7-6 所示。

图 7-6 通过基于 MAC 地址的包过滤控制上网权限的典型应用



在 GigabitEthernet 0/0/1 的 Inbound 方向上应用基于 MAC 地址的包过滤，禁止指定 MAC 地址的内网用户访问外网。

基于 MAC 地址的包过滤只支持应用在接口的 Inbound 方向上，也就是只支持对接口接收的报文进行过滤。

### 7.4.6.3 控制二层接口卡接收报文

#### 应用环境

硬件包过滤是在特定的二层硬件接口卡上实现的，用来控制接口卡上的接口可以接收哪些流量。硬件包过滤直接通过硬件实现，所以过滤速度更快。支持硬件包过滤的接口卡请参见 [7.4.2 可获得性](#)。

硬件包过滤既可以过滤二层的以太网帧，也可以过滤三层的 IP 报文。

硬件包过滤一般用于二层接口卡内的访问控制。

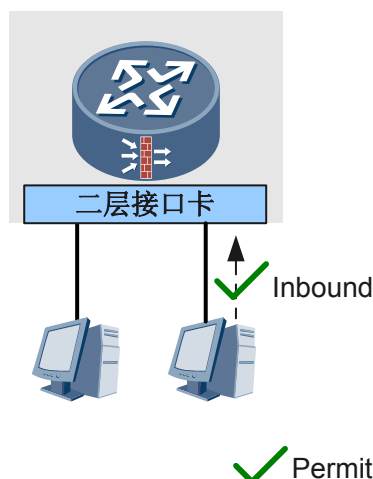
硬件包过滤和基于 MAC 地址的包过滤都可以过滤以太网帧，主要区别在于：

- 硬件包过滤在二层接口卡的物理口上应用，主要用于控制二层接口卡接收的以太网帧。
- 基于 MAC 地址的包过滤可应用在三层口，或由三层口通过 **portswitch** 命令切换的二层口上，用于控制三层接口卡上的以太网帧接收。

#### 典型应用

硬件包过滤只能应用在接口的 Inbound 方向上，也就是只支持对接口接收的报文进行过滤。典型应用如 [图 7-7](#) 所示。

图 7-7 通过硬件包过滤控制局域网访问控制典型应用



二层接口卡下 PC 属于同一网段，可以通过硬件包过滤过滤以太网帧或 IP 报文。

## 7.4.7 原理描述

### 7.4.7.1 安全策略的实现

安全策略中最关键的两个问题是“如何匹配需要控制的报文”以及“对报文进行哪些控制或检查”。也就是先根据报文中的字段、时间、发起访问的用户帐号等条件匹配报文，然后对匹配的报文执行相应的控制动作或安全检查。

当策略中定义了多种条件时，报文必须同时满足所有条件才会被认为匹配成功。如果策略中没有定义任何条件，只定义了动作，则该策略会与任意报文匹配。

安全策略分为两大类：

- 域间或域内安全策略：用于控制域间或域内的流量，此时的安全策略既有传统包过滤功能，也有对流量进行 IPS、AV、Web 过滤、应用控制等进一步的应用层检测的作用。域间或域内安全策略是包过滤、UTM 应用层检测等多种安全检查同时实施的一体化策略。
- 应用在接口上的包过滤规则：用于控制接口的流量，就是传统的包过滤功能，基于 IP、MAC 地址等二、三层报文属性直接允许或拒绝报文通过。

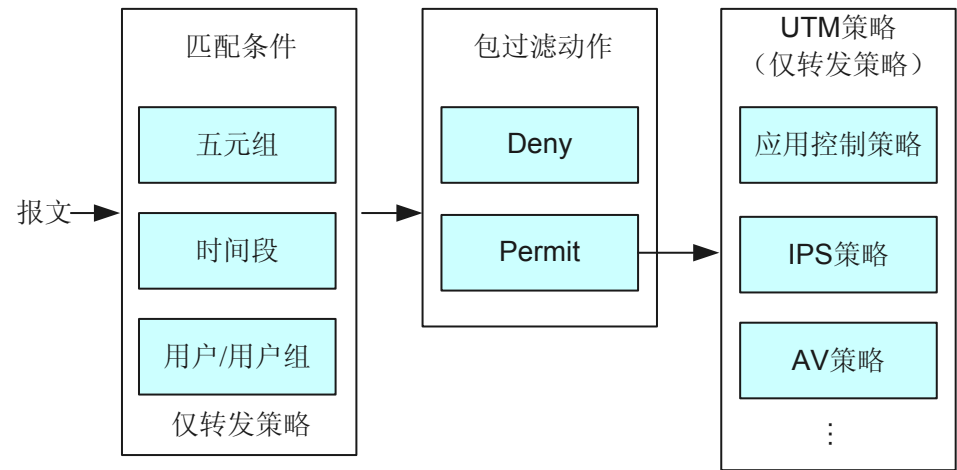
## 域间安全策略

域间安全策略控制域间数据流动，根据适用场景分为两类：

- 转发策略：控制设备转发的流量，包括传统的包过滤和应用层的 UTM 检测。
- 本地策略：控制外界与设备的互访，只根据五元组进行控制。

域间安全策略的基本处理过程如图 7-8 所示。

图 7-8 域间安全策略示意图



域间还提供缺省包过滤，指定报文没有匹配到任何安全策略时的控制动作。设备将首先查找域间的 Policy，如果没有找到匹配项将匹配缺省包过滤进行处理。对报文的处理流程如表 7-5 所示。

表 7-5 域间安全策略匹配过程

| 条件              |                      | 操作    |
|-----------------|----------------------|-------|
| 报文匹配到域间 Policy  | Policy 允许报文通过        | 转发该报文 |
|                 | Policy 禁止报文通过        | 丢弃该报文 |
| 报文未匹配到域间 Policy | 该域间的缺省包过滤规则为默认允许报文通过 | 转发该报文 |
|                 | 该域间的缺省包过滤规则为默认禁止报文通过 | 丢弃该报文 |

域内安全策略

缺省情况下域内允许任何流量通过，如果需要对域内流量进行控制可以通过域内安全策略实现。

域内安全策略与域间安全策略类似，区别就是域内安全策略没有 Inbound 和 Outbound 方向的区分。

域内安全策略不支持对 Local 域内流量的控制。

接口包过滤

USG 的接口包过滤用于对没有加入安全区域的接口收发的 IP 报文进行控制。

在接口包过滤中，主要通过基本 ACL 或高级 ACL 来对流量进行选择，然后在接口上应用 ACL。之后该接口接收或发送的报文中，如果在 ACL 中对应动作为 **permit**，将允许被转发；如果对应动作为 **deny**，将被丢弃。

## 基于 MAC 地址的包过滤

基于 MAC 地址的包过滤主要用于对接口收到的以太网帧进行控制。

在基于 MAC 地址的包过滤中，主要通过基于 MAC 地址的 ACL 来对流量进行选择，然后在接口上应用 ACL。之后该接口接收的报文中，如果在 ACL 中对应动作为 **permit**，将允许被转发；如果对应动作为 **deny**，将被丢弃。

## 硬件包过滤

硬件包过滤用于对二层接口接收的 IP 报文或以太网帧进行控制，只有特定的二层接口卡支持。硬件包过滤可以对特定接口卡接收的报文，直接进行由硬件芯片实现的包过滤。相比于软件包过滤，其匹配和过滤的速度更快，效率更高，消耗系统资源更少。

硬件包过滤由二层接口卡的芯片直接处理，不会经过设备的 CPU 处理。

在硬件包过滤中，主要通过硬件包过滤 ACL 来对流量进行定义，然后在接口上应用 ACL。之后该接口接收的报文中，如果在 ACL 中对应动作为 **permit**，将允许被转发；如果对应动作为 **deny**，将被丢弃。

## 安全策略的匹配

一个安全域间或接口上可以对不同的流量进行不同的控制，不同类型安全策略的实现方法不同：

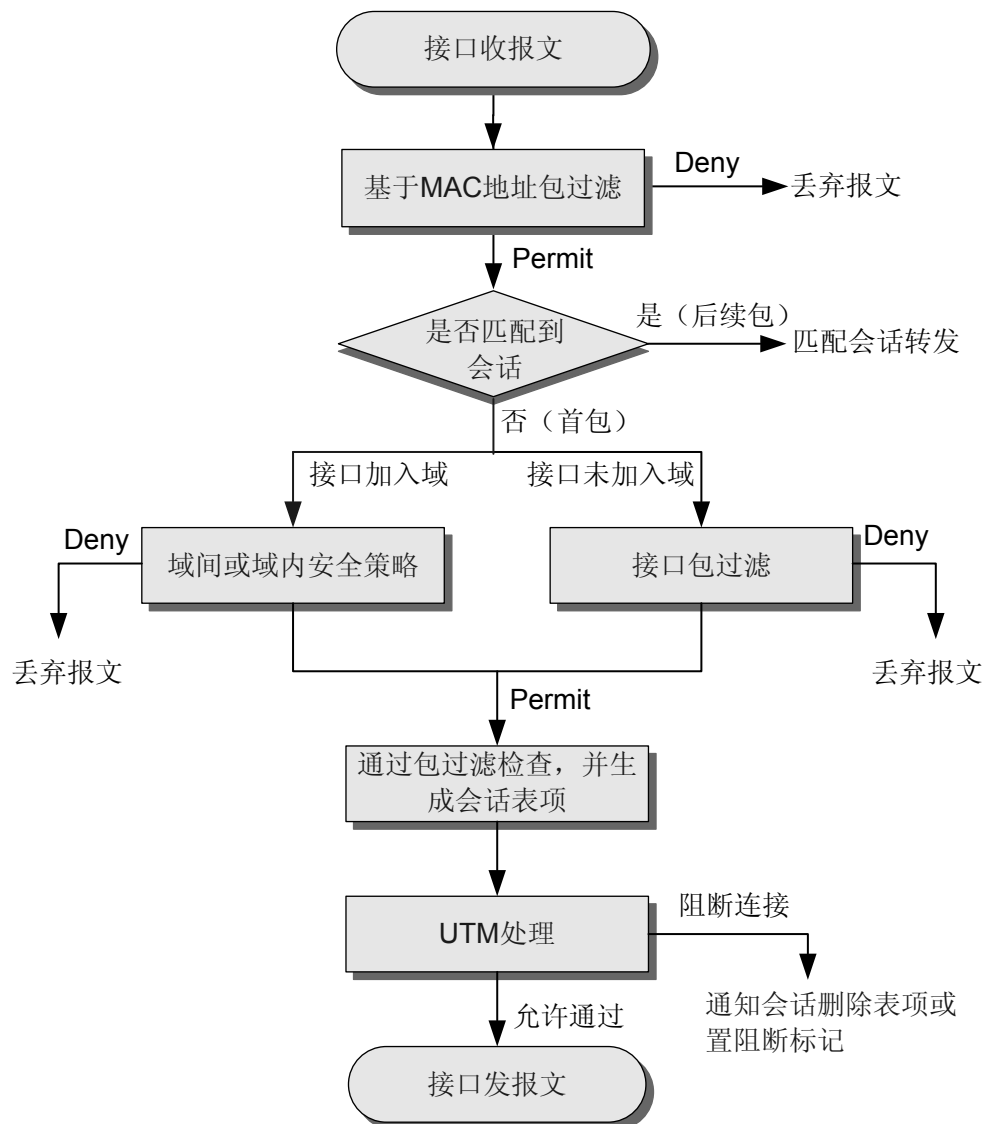
- 域间或域内安全策略通过在域间应用多条策略实现，每条策略匹配不同的数据流并指定控制动作。如果安全策略不是以自动排序方式配置的，策略的优先级按照配置顺序进行排列，越先配置的策略，优先级越高，越先匹配报文。但是也可以手工调整策略之间的优先级。如果安全策略是以自动排序方式配置的，策略的优先级按照策略 ID 的大小进行排列，策略 ID 越小，优先级越高，越先匹配报文。此时，策略之间的优先级关系不可调整。  
另外域间还提供缺省包过滤，指定报文没有匹配到任何安全策略时的控制动作。设备将首先查找域间的 Policy，如果没有找到匹配项将匹配缺省包过滤进行处理。
- 应用在接口上的包过滤，每个接口上只能应用一条包过滤策略。但是 ACL 中可以配置多条 rule，每条 rule 匹配不同的数据流并指定控制动作。缺省情况下，越先创建的 rule 优先级越高，越先用来匹配。

所有类型的包过滤都是一旦匹配到一条策略或 rule，就直接按照该策略或 rule 的定义处理报文，不再继续往下匹配。

## 安全策略处理流程

安全策略处理流程如图 7-9 所示。

图 7-9 安全策略处理流程



1. 基于 MAC 地址的包过滤为逐包过滤，对每个收到的报文的源 MAC 地址、目的 MAC 地址等进行匹配，丢弃不允许通过的以太网帧。基于 MAC 地址的包过滤为二层过滤，以太网帧被丢弃将不再继续下一步三层包过滤处理。
2. 如果没有基于 MAC 地址的包过滤，将首先查找会话，如果是首包需要进行域间安全策略处理；如果是后续包直接匹配会话转发，关于安全策略与会话的关系请参见 [7.4.7.2 安全策略与状态检测机制的关系](#)。
3. 首包如果匹配到安全策略的 permit 动作，将生成会话表项；如果匹配到 deny 动作报文被丢弃。  
同一条数据流的后续报文（包括反向应答报文）不再进行安全策略处理直接匹配会话转发。
4. 如果安全策略中应用了 UTM 策略，安全策略 permit 的流量将继续进行 UTM 处理，根据 UTM 处理结果转发。  
如果 UTM 阻断了连接将通知会话表删除会话表项或在会话表项上置“阻断”标记。

 说明

还有一类比较特殊的硬件包过滤，只在二层卡的物理接口上支持，控制二层接口卡接收的流量。硬件包过滤一般用于控制二层卡内互访的流量。

当需要在二层接口卡的接口上进行三层转发时（跨网段转发），要将二层物理口加入 VLAN，然后将 Vlanif 接口和三层口加入安全区域并配置域间安全策略。例如二层接口卡用作局域网接入，三层口接入互联网，控制内网的上网权限的场景就是这种情况。

硬件包过滤也是逐包过滤。

## 7.4.7.2 安全策略与状态检测机制的关系

不同于传统防火墙的逐包检测机制，USG 采用基于流的状态检测。设备根据五元组（源 IP 地址、目的 IP 地址、源端口号、目的端口号和协议类型）来定义流，只要这五个元素相同的报文即可认为属于同一条流。USG 通过会话表来记录各条流的五元组信息，每条流的信息是会话表中的一个会话表项。设备只对首包按照安全策略进行检查，并将允许通过的流的“状态”记录下来，建立会话表项，后续报文直接根据会话表进行转发。这样的处理大大提高了检测速率和转发效率。

 说明

基于 MAC 地址的包过滤和硬件包过滤是逐包检测。

建立会话表项后如果安全策略配置发生变化，后续报文到来时将根据新的安全策略配置刷新会话表项。

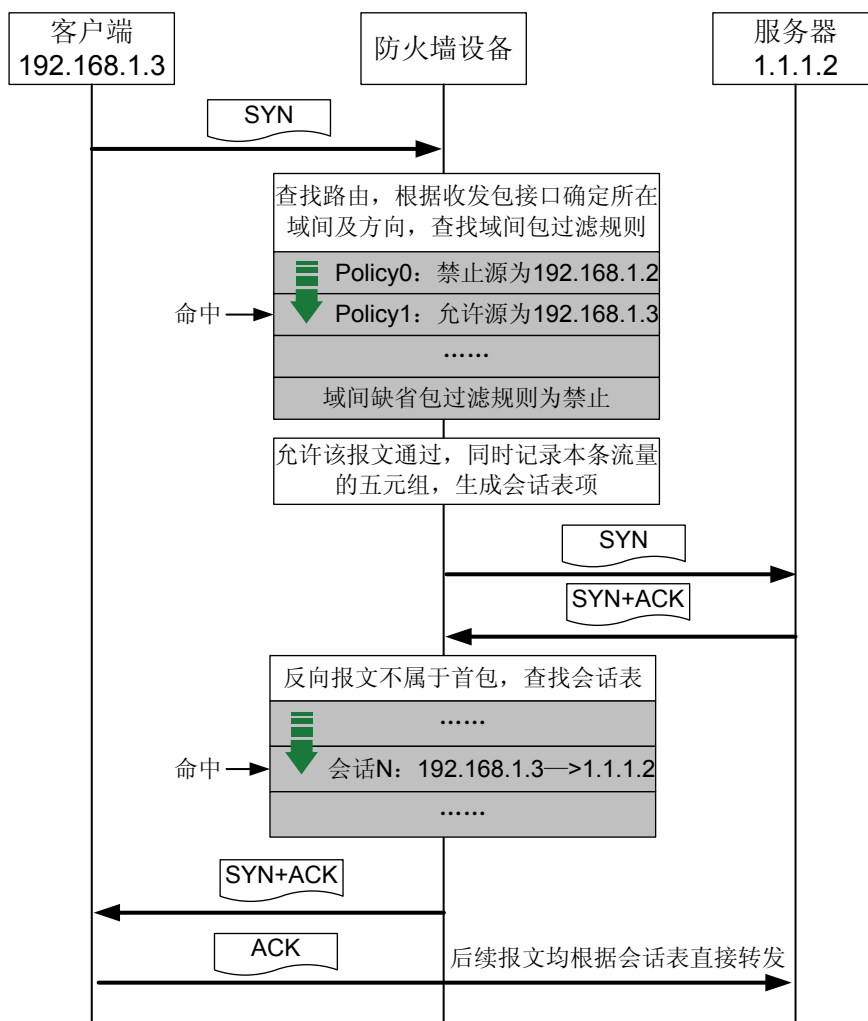
请根据需要选择开启或关闭状态检测机制：

- 状态检测机制开启状态下，只有首包通过设备才能建立会话表项，后续包直接匹配会话表项进行转发。
- 状态检测机制关闭状态下，即使首包没有经过设备，后续包只要通过设备就可以生成会话表项。在来回路径不一致的组网中，就必须关闭状态检测功能，否则可能由于首包不经过防火墙导致会话表项不能正常建立。但是关闭状态检测机制会导致 TCP 代理方式的 SYN Flood 攻击防范、TCP 全连接攻击防范、UTM、ASPF 等功能不可用。

下面以建立 TCP 连接过程中的域间安全策略检查为例，来对这种机制予以说明。

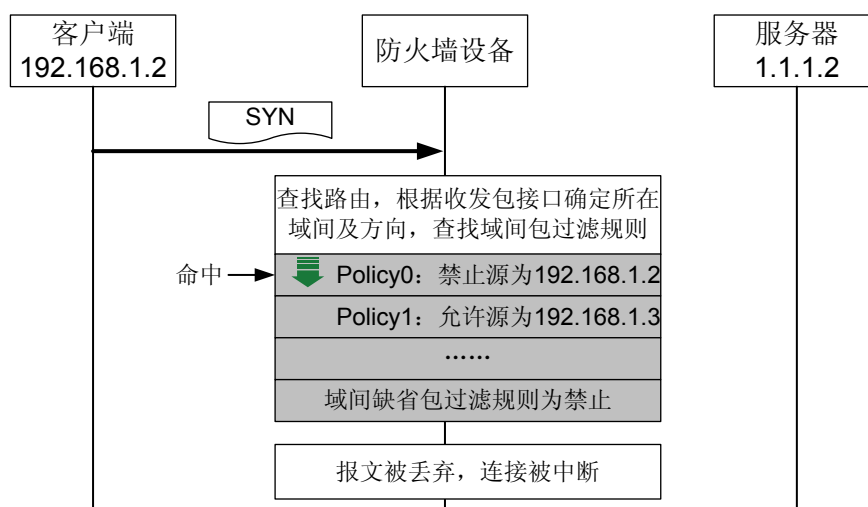
一个被正常转发的 TCP 流量的检查过程如[图 7-10](#)所示。

图 7-10 安全策略允许流量通过示例



一个由于被域间安全策略禁止的 TCP 流量的检查过程如图 7-11 所示。

图 7-11 安全策略禁止流量通过示例





## 7.5 IPv6 包过滤

IPv6 包过滤主要是为了对网络中各种不同的 IPv6 流量是否转发做一个最基本的控制。本节将介绍 IPv6 包过滤的基本概念、规格及原理。

### 7.5.1 介绍

#### 定义

IPv6 包过滤是一个对网络中的数据报文进行有选择地控制的过程。

#### 目的

IPv6 包过滤控制了不同用户在安全域间的 IPv6 网络访问权限，是最基础的安全策略。设备在实施其他安全策略之前，都会先进行 IPv6 包过滤规则的检查。

### 7.5.2 原理描述

IPv6 包过滤作为一种网络安全保护机制，用于在两个不同安全级别的网络之间控制流入和流出网络的 IPv6 数据。在转发数据包时，先检查包头信息（例如 IPv6 包的源地址/目的地址、源端口/目的端口和上层协议等），然后与设定的规则进行比较，根据比较的结果决定对该数据包进行转发还是丢弃处理。

为了实现 IPv6 数据包过滤，需要配置一系列的过滤规则。采用 ACL6 定义过滤规则，然后将 ACL6 应用于不同安全区域之间，从而实现 IPv6 包过滤。

## 7.6 限流策略

限流策略通过为不同的 IP 分配不同的流量，达到优化网络的目的。本节介绍限流策略的基本概念、规格及原理。

### 7.6.1 介绍

#### 定义

流量限制是指通过对指定 IP 或者网段的带宽或者连接数进行限制，来达到合理分配流量，避免网络拥堵的目的。

流量限制的方式包括：

- 连接数限制：对指定的 IP 地址或者网络发起的连接数量或接收的连接数量进行限制。
- 带宽限制：对指定的 IP 地址或者网络的带宽进行限制，可同时设置带宽的上下限。

流量限制的场景可分为如下三种：

- 每 IP 限流是指对命中策略的每个 IP 进行带宽或连接数的限制。

- 整体限流是指对命中策略的所有 IP 进行整体的带宽或连接数的限制。
- 保证带宽是指对命中策略的 IP 固定至少分配一定带宽的流量，以保证其业务不受影响。

## 目的

流量限制为了帮助管理员为不同用户分配带宽、优化网络，通过限制非关键网络业务的带宽，来保证重点网络业务的带宽。

## 7.6.2 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                       |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100 V300R001、<br>USG2100/2200/5100 BSR/HSR V300R001<br>USG5500 V300R001 |

## 7.6.3 应用场景

### 7.6.3.1 每 IP 限流

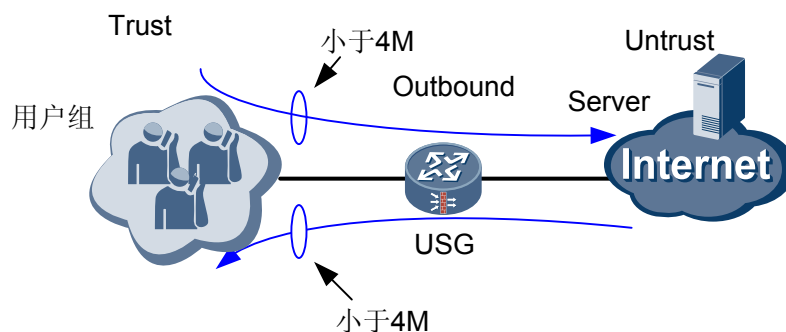
#### 应用环境

当需要同时进行针某个网络中每个 IP 进行流量控制时，可采用每 IP 限流策略，实现对每个 IP 地址进行带宽和连接数的限制，也可以具体针对应用协议、服务（基于端口）等进行限流控制。

#### 典型应用

如图 7-12 所示，当需要对某个网络用户组内的每个用户带宽限制在 4M 以内，可采用每 IP 限流，实现对每个用户进行带宽和连接数的限制。同时，也可以具体针对用户访 Server 的应用协议进行限流控制，如 P2P 下载、IM 即时通信等协议。

图 7-12 每 IP 限流



在配置每 IP 限流时，需要结合流量方向来对流量进行限制和管理，配置哪个方向的限流策略，取决于流量的方向。例如：由于 Trust 安全域比 Untrust 安全域级别高，所以 Trust 区域的用户访问 Untrust 区域的服务器、用户上传文件至服务器，都是属于域间 Outbound 方向；用户从服务器下载文件至本地 PC，是属于 Inbound 方向。

在同一策略中，不能实现同时对用户的收发流量进行限制，当实现同时对用户的收发流量同时进行限制时，需要配置 Outbound 和 Inbound 方向上的两条策略。对于每 IP 限流的同一个方向，可以选择针对源或者目的 IP 来进行限流，例如：对于 Outbound 方向，有如下两种限流形式：

- 基于源 IP 的限流，就是限制 PC 发出的流量带宽和连接数。
- 基于目的 IP 的限流，就是限制 Server 收到的流量带宽和连接数。

### 7.6.3.2 整体限流

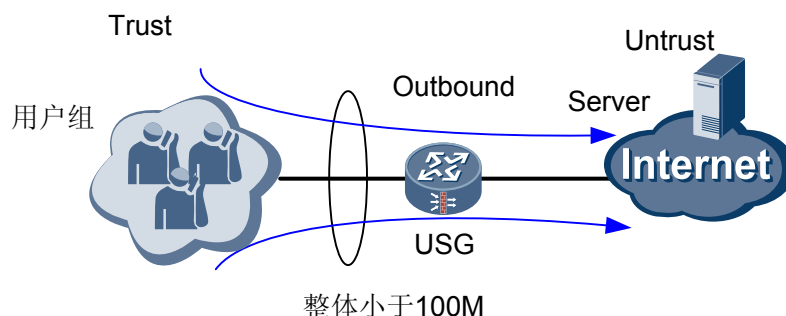
#### 应用环境

当需要对某一网段的用户进行整体限流控制，保证整个网络的带宽和连接数不超过限制，也可以具体针对应用协议、服务（基于端口）等进行限流控制。

#### 典型应用

如图 7-13 所示，当需要对整个用户组进行 100M 的带宽限流，对网络内用户连接数不超过 10 个时，可通过整体限流来实现对整个网络内带宽和连接数的控制。同时，也可以具体针对用户访 Server 的应用协议进行限流控制，如 P2P 下载、IM 即时通信等协议。

图 7-13 整体限流



同每 IP 限流一样，在配置整体限流时，需要结合流量方向来对流量进行限制和管理，配置哪个方向的限流策略，取决于流量的方向。例如：由于 Trust 安全域比 Untrust 安全域级别高，所以 Trust 区域的用户访问 Untrust 区域的服务器、用户上传文件至服务器，都是属于域间 Outbound 方向；用户从服务器下载文件至本地 PC，是属于 Inbound 方向。

在同一策略中，不能实现同时对 PC 的收发流量进行限制，当需要实现同时对 PC 的收发流量进行限制时，需要配置 Outbound 和 Inbound 方向上的两条策略。

### 7.6.3.3 保证带宽

#### 应用环境

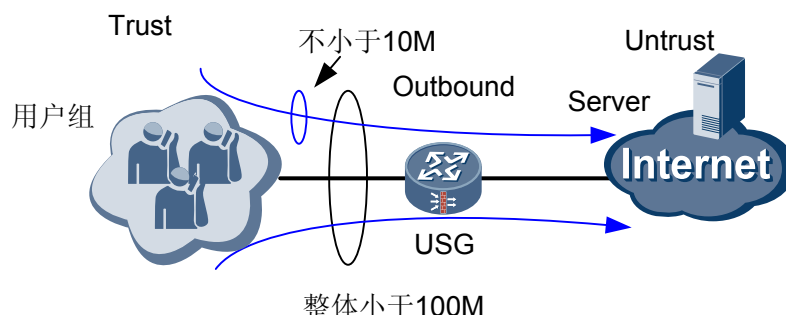
当在某个网络中有些重要的业务，需要有一定的流量带宽才能保证其正常运行时，即便是在网络繁忙时，其带宽流量也不会受影响；同时，在网络空闲时，也可以分配到多余的带宽。这种情况下可以通过配置保证带宽功能来解决该问题。

#### 典型场景

如图 7-14 所示，当需要对用户组进行整体带宽限流 100M，对组内每个用户限流最大为 20M，同时，由于用户组正在运行某项重要业务，需要一定的带宽流量保证每个用户的业务不受影响，这时可以通过保证带宽来满足需求。

例如：给整个网络中的每个用户配置保证带宽为 10M，当用户进行通信时，首先会固定先获得 10M 流量，保证其业务流量。然后整体网络中的所有用户再通过抢占的方式，分配整体剩下的带宽，但每个用户最带宽最大不超过 20M。

图 7-14 保证带宽



## 7.6.4 原理描述

流量限制的方式包括：

- 连接数限制：对指定 IP 地址或者网络发起的连接数量或接收的连接数量进行限制。  
此处的连接是指使用 IP 协议建立的连接，既可以是 TCP 连接，又可以是 UDP 连接和 ICMP 连接。一个连接通常由源 IP、源端口、目的 IP、目的端口和协议五个元素（五元组）唯一确定。所以一个 IP 地址可以和其他 IP 地址建立多条连接，只要这五个元素不同即可认为是不同的连接。由于设备在实现上是以会话表个数为依据统计连接的个数，所以对于多通道协议，两个网络设备之间可能会建立多个连接。  
在实际应用中，有部分协议会在一次通信过程中建立多个连接，大量占用设备的会话资源，其中最典型的是 P2P 协议。所以可以通过连接数限制功能控制此类协议的连接数，以保证其他正常业务的转发。
- 带宽限制：对指定 IP 地址或者网络的带宽进行限制。  
带宽限制能起到优化网络流量、保证用户正常访问速率、辅助防范网络攻击的作用。在实际应用中，可以使用带宽限制功能进行网络带宽的分配，对不同用户进行差异化的管理。

### 每 IP 限流策略（一级 CAR）

每 IP 限流策略，是针对每个 IP 进行 CAR(Committed Access Rate)的，其中对于策略约束条件包括五元组，用户身份以及应用协议。具体内容如下：

流量管控方式：通过针对源或目的 IP 进行流量控制或管理，即针对对匹配策略流量的源地址或者目的地址来进行限流。

- 带宽限制：
  - 对每个 IP 地址进行最大带宽限制。
  - 对每个 IP 地址进行保证带宽限制。
- 连接数限制：对每个 IP 地址进行最大连接数限制。

## 整体限流策略（二级 CAR）

整体限流是针对一个域间或域内的所有数据流进行整体管控，其中约束条件包括五元组，用户身份，应用协议。具体内容如下：

流量管控的内容：

- 带宽限制：对整体网段进行最大带宽管控。
- 连接数限制：对整体网段的连接数进行最大连接数管控。

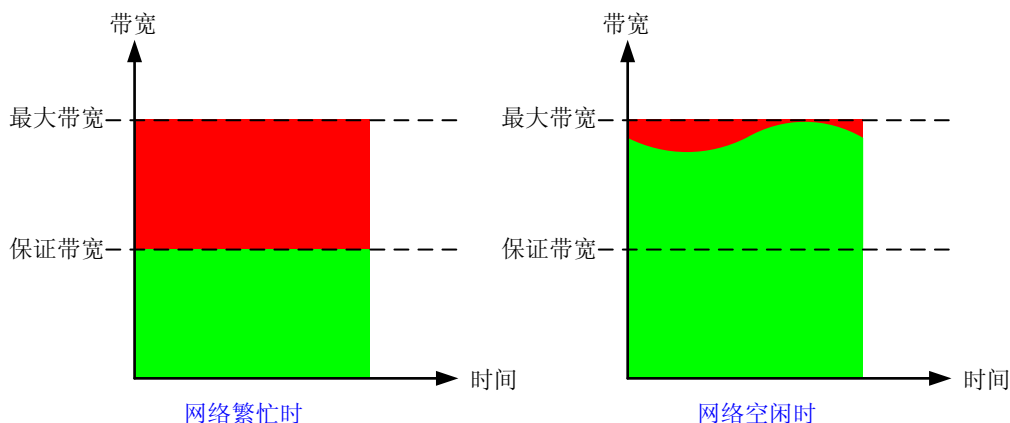
## 保证带宽

保证带宽是指如果同时配置了每 IP 限流策略和整体限流策略，如果每 IP 策略放行流量后，在整体限流时，如果多个 IP 的流量总和超出了整体限流值，这时每个 IP 获得的流量减少，但仍然首先让保证带宽内的流量通过（保证带宽必须不大于每 IP 获得的最大带宽，所有 IP 保证带宽的总和不超过整体带宽）。而且当整体带宽有空闲时，还可以使用多余的带宽。如果还配置了每 IP 的最大带宽，那多余带宽上限不超过：最大带宽减去保证带宽。

如图 7-15 对报文处理实现如下：

- 网络繁忙时：保证带宽对报文进行了分类，并不丢弃报文，在保证带宽内的绿色报文优先允许通过，红色报文暂不通过。
- 网络空闲时：对于绿色报文无论如何都放过，也就是保证带宽始终是最优先的。然后对于红色报文，只有当二级 CAR 还有多余带宽的时候，红色报文通过抢占剩余带宽的方式通过。

图 7-15 保证带宽



## 7.7 NAT

NAT 能够实现 IP 报文中源地址或目的地址的转换，缓解 IPv4 地址资源枯竭的问题。本节介绍 NAT 的基本概念、规格、参考文档及原理。

### 7.7.1 介绍

## 定义

NAT (Network Address Translation) 一种 IP 地址转换技术，通过 NAT 可以将 IP 报头中的 IP 地址转换为另一个 IP 地址。

## 目的

引入 NAT 机制的主要原因是随着计算机网络的发展，IPv4 地址资源被迅速消耗。私网地址的引入有效缓解了这个问题，但是私网用户并不能直接使用私网地址访问 Internet。所以必须通过 NAT 技术，对 IP 地址进行转换，实现大量私网用户接入互联网的需求。

## 7.7.2 参考标准和协议

本特性的参考资料清单如下：

- RFC3022: Traditional IP Network Address Translator (Traditional NAT)
- RFC3519: Mobile IP Traversal of Network Address Translation (NAT) Devices
- RFC5508: NAT Behavioral Requirements for ICMP
- RFC5382: NAT Behavioral Requirements for TCP
- RFC4787: Network Address Translation (NAT) Behavioral Requirements for Unicast UDP

## 7.7.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

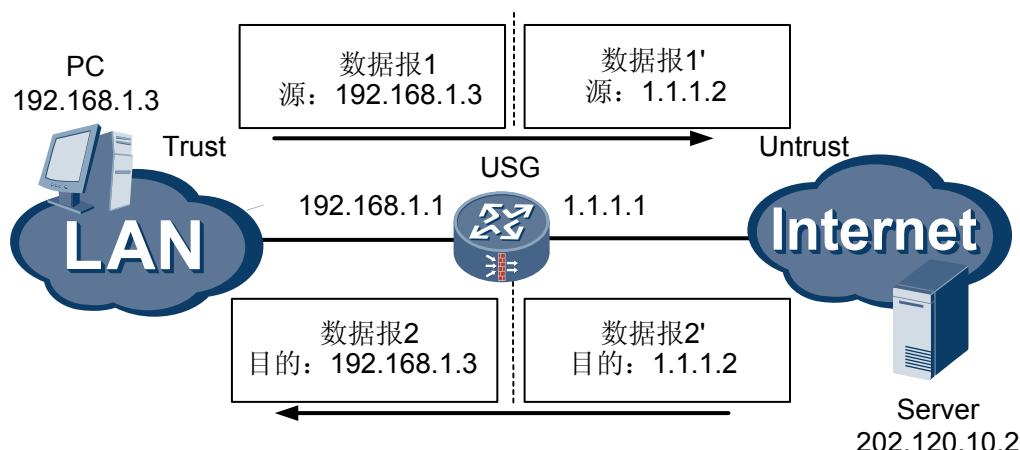
## 7.7.4 原理描述

### 7.7.4.1 NAT 的实现

#### NAT No-pat 转换机制

图 7-16 描述了一个基本的基于源 IP 地址的 NAT 应用。当 USG 收到报文时，首先根据域间策略判定是否允许该报文通过，然后再根据 NAT 相关配置判定该报文是否需要做 NAT，如果需要做 NAT，则根据 NAT 配置对地址进行地址转换。

图 7-16 NAT No-pat 应用示意图



USG 作为 NAT 服务器处于私有网络和公共网络的连接处，私网 PC 与公网服务器的交互报文全部通过该 NAT 服务器，地址转换的过程如下：

1. 私网 PC（192.168.1.3）发往公网服务器（202.120.10.2）的数据报 1 到达 NAT 服务器后，NAT 服务器查看报头内容，发现该数据报为发往公网网络的报文。
2. NAT 服务器从 NAT 地址池中选取一个公网地址 1.1.1.2，然后将数据报 1 的源地址字段的私有地址 192.168.1.3 换成 1.1.1.2，作为数据报 1' 发送到公网服务器，同时在会话表中记录这一地址转换映射。
3. 公网服务器收到数据报 1' 后，向私网 PC 发送应答报文，即数据报 2'，目的地址为 1.1.1.2。
4. 数据报 2' 到达 NAT 服务器后，NAT 服务器查看报头内容，查找当前会话表的记录，用私有地址 192.168.1.3 替换目的地址，作为数据报 2 发送给私网 PC。

上述的 NAT 过程对私网 PC 和公网服务器来说是透明的。私网 PC 认为与公网服务器的交互报文没有经过 NAT 服务器的干涉，公网服务器认为私网 PC 的 IP 地址就是 1.1.1.2，并不知道存在 192.168.1.3 这个地址。

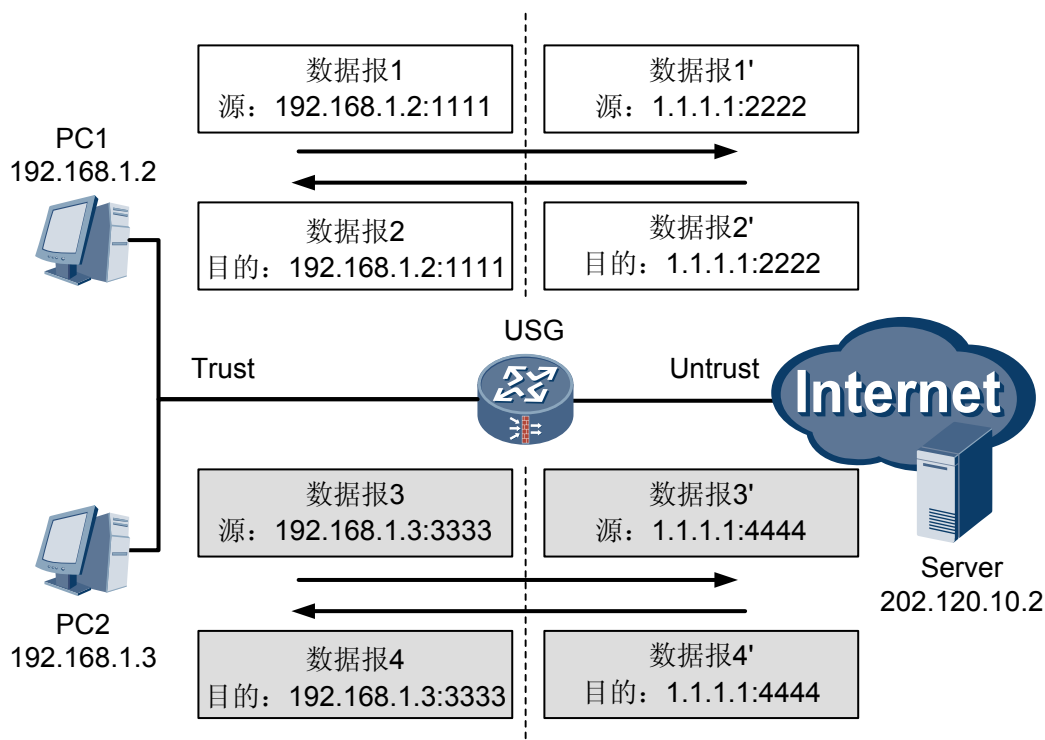
## NAPT 转换机制以及 Easy-IP 机制

NAT No-pat 是最基础的 NAT 转换方式。但是在公网地址匮乏的情况下，这种“一对一”的转换往往不能满足需求。更多的私网需要实现多个私网用户共享一个公网 IP 访问 Internet，即“多对一”的需求。NAPT 转换机制通过进行地址转换的同时，对端口也进行转换解决了这一问题。

图 7-17 描述了一个基于源 IP 地址的 NAPT 应用。当 USG 收到报文时，首先根据域间策略判定是否允许该报文通过，然后再根据 NAT 相关配置判定该报文是否需要做 NAT，如果需要做 NAT，则根据 NAT 配置对地址进行地址转换。



图 7-17 NAPT 应用示意图



- 对于 PC1
    1. 私网 PC1 (192.168.1.2:1111) 发往公网服务器 (202.120.10.2) 的数据报 1 到达 NAT 服务器后，NAT 服务器查看报头内容，发现该数据报为发往公网网络的报文。
    2. NAT 服务器从 NAT 地址池中选取一个可以在 Internet 上路由的公网地址 1.1.1.1，然后将数据报 1 的源地址字段的私有地址 192.168.1.2 换成 1.1.1.1，将源端口号字段 1111 替换为随机抽取的端口号 2222，作为数据报 1' 发送到公网服务器，同时在会话表中记录这一地址转换映射。
    3. 公网服务器收到数据报 1' 后，向私网 PC 发送应答报文，即数据报 2'，目的地址为 1.1.1.1。
    4. 数据报 2' 到达 NAT 服务器后，NAT 服务器查看报头内容，查找当前会话表的记录，用私有地址 192.168.1.2 替换目的地址，使用端口号 1111 替换目的端口，作为数据报 2 发送给私网 PC。
  - 对于 PC2 的转换过程基本和 PC1 一致。在替换发往公网报文的源地址时，采用的是和 PC1 一样的 1.1.1.1，但是替换源端口时，采用的是 PC1 不一样的 4444。通过端口号将两台 PC 的两个转换关系区分开来。
- 当 NAT 服务器收到 Server 回应给 PC2 的报文时，根据会话表信息，相应地将报文的目的地址转换为 192.168.1.3，目的端口转换为 3333。

另外，设备在进行地址转换时还可以直接使用设备与 Internet 相连接口的 IP 地址作为转换后的地址。这种应用称为“Easy-IP”。“Easy-IP”特别适用于需要私网内部主机较少，出接口的 IP 地址是通过拨号或 DHCP 方式获取的临时 IP 地址，经常会发生变化的情况。

需要注意的是，当采用 Easy-IP 时，由于 NAT 转换会占用设备接口的部分端口，所以当这个端口被用做 NAT 转换后，就不能再通过这个端口访问设备本身了。

例如：设备与 Internet 相连接口的 IP 地址是 1.1.1.1。当内网用户 A 访问外网时，设备选用了 1.1.1.1 的 8000 端口供其使用。如果此时，管理员又配置设备使用 8000 端口对外提供基于 HTTPS 协议的 Web 管理服务，那么实际上 Internet 中的远程管理员是不能通过访问 https://1.1.1.1:8000 来登录设备的 Web 界面的。因为当设备收到发往 1.1.1.1 的 8000 端口的报文就会根据会话表将其转发给内网用户 A。

NAT Server 转换机制

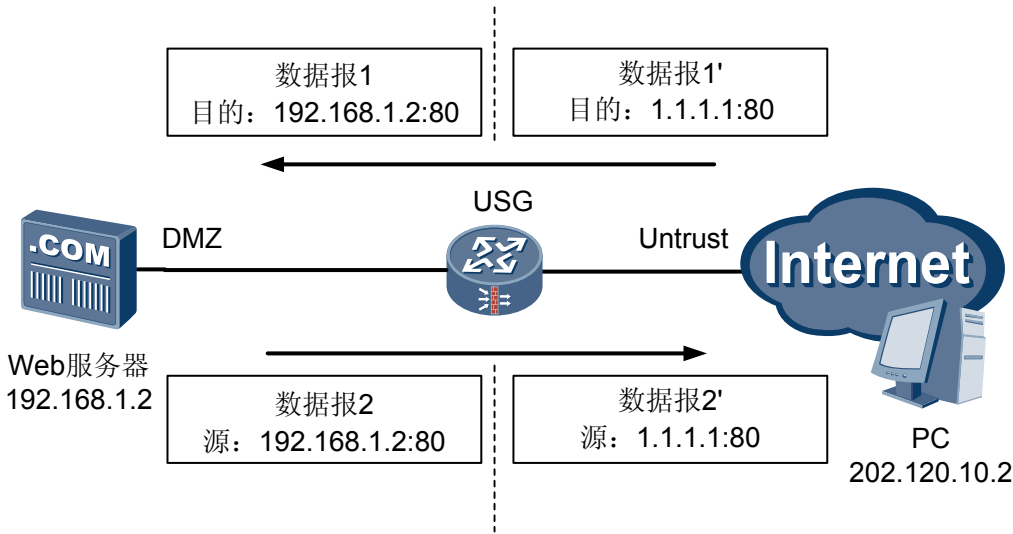
在 NAT 应用中，通常还有这样一个需求：在私网中有一台服务器，其真实 IP 地址是私网地址。但是这台服务器同时还向 Internet 用户提供服务，需要被公网用户主动访问。这种应用称为“NAT Server”或“内部服务器”。

在这种应用中主要需要解决两个问题：

- 外网用户必须通过公网地址来访问服务器，所以必须将发往服务器的报文中的目的地址由公网地址转换为服务器真实的私网地址。
- 在严格包过滤情况下，为了保证安全性，通常不允许外网用户主动向内网主机发起连接。而在 NAT Server 应用中，这种情况是必须允许的。

解决第一个问题的方法是配置 NAT Server 功能，在设备上建立一个服务器的私网地址与公网地址的映射关系（保存在 Server-map 表项中）。当设备收到一个发往服务器公网地址的报文时，将报文的目的 IP 地址转换为私网地址。如图 7-18 所示。

图 7-18 NAT Server 应用示意图



1. 公网 PC（202.120.10.2）发往 Web 服务器（1.1.1.1）的数据报 1'到达 NAT 服务器后，NAT 服务器查看报头内容，并比对 Server-map 表中的映射关系，发现该数据报为发往 Web 服务器的报文。
2. NAT 服务器将数据报 1'的目的地址字段由公网地址 1.1.1.1 换成 192.168.1.2，作为数据报 1 发送到 Web 服务器，同时在会话表中记录这一访问链接。
3. Web 服务器收到数据报 1 后，向公网 PC 发送应答报文，即数据报 2，源地址为 192.168.1.2。
4. 数据报 2 到达 NAT 服务器后，NAT 服务器查看报头内容，查找当前会话表的记录，用公网地址 1.1.1.1 替换源地址，作为数据报 2'发送给公网 PC。

5. 后续 PC 继续发送给 Web 服务器的报文，设备都会直接根据会话表项的记录对其进行转换，而不会再去查找 Server-map 表项。

解决第二个问题的方法是配置合适的包过滤策略。应当允许 Untrust 的用户访问 Web 服务器。需要注意的是，由于设备处理流程中，是先进行 NAT Server 地址转换，再进行包过滤的检查。所以在进行流量匹配时，需要匹配的流量的目的地址应该是服务器的私网地址，即 NAT Server 转换后的地址。例如在图 7-18 中，在 Untrust 和 DMZ 的 Inbound 方向上，应当配置允许目的地址为 192.168.1.2 的流量通过的包过滤规则。

在 NAT Server 应用中，如果私网拥有的公网地址资源较少，可以直接使用设备连接 Internet 的接口的 IP 地址作为服务器的公网 IP 地址。此时存在以下两种情况：

- 如果配置 NAT Server 时不指定公网和私网的端口映射关系，则设备会将发往公网地址所有端口的报文都转发给私网地址对应端口。这是因为配置 NAT Server 后创建的 Server-map 表项仅记录了转换前后的 IP 地址，没有记录端口信息。这样设备就将对所有端口进行转换。  
这样造成的影响是，设备管理不能在 Internet 上通过这个公网 IP 地址访问设备本身，例如通过 Telnet 或 SSH 登录设备的命令行配置界面，或者通过浏览器登录设备的 Web 配置界面。
- 如果配置 NAT Server 时指定公网和私网的端口映射关系，则只有指定的端口才会被占用。所以此时管理员如果仍然希望可以从 Internet 上访问设备本身，应当避免将内部服务器的端口映射为设备本身提供服务的端口，例如 FTP 协议的 21 端口、Telnet 协议的 23 端口、HTTP 协议的 80 端口等。

## 通过 NAT Inbound 简化 NAT Server 应用中的路由配置

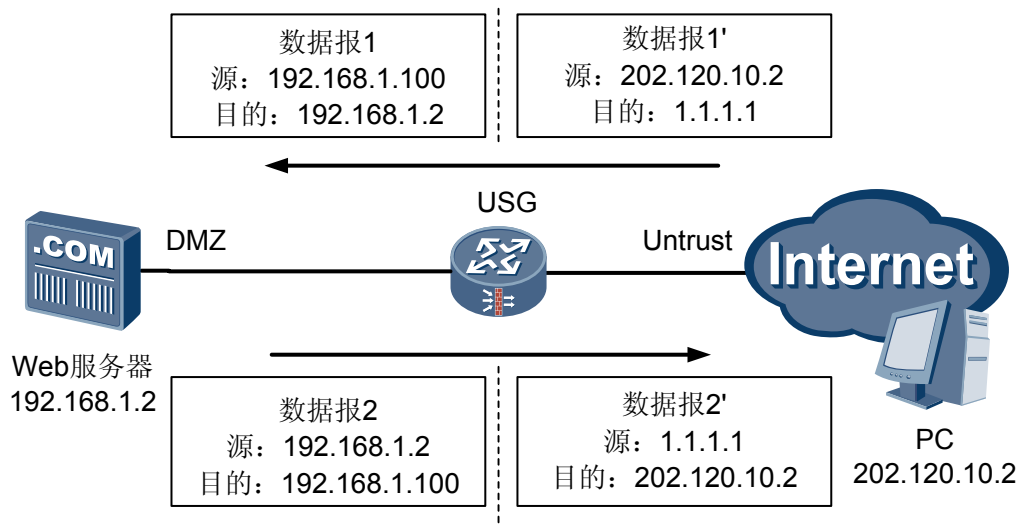
在 NAT Server 应用中，为了保证内部服务器将回应给公网用户的报文发给设备进行处理，必须在内部服务器上配置相应的路由。但是由于内部服务器部署在内网，通常不能通过动态路由协议自动学习到达公网某个 IP 地址的路由信息，而逐条手工添加静态路由又非常繁琐。有时也不能直接将内部服务器的缺省网关配置为设备内网接口的 IP 地址，所以内部服务器也不能将找不到路由的报文都缺省发给设备。

这种情况下，通过配置一个 Inbound 方向上的 NAT，可以解决这一问题。

NAT Inbound 实际上是指对公网用户访问内部服务器的报文的源 IP 地址也进行转换的一种方法。由于公网用户所在安全区域一般比内网要低，所以这个 NAT 转换关系通常应用在域间的 Inbound 方向上，因此得名。

NAT Inbound 解决上述的问题的主要原理是：将公网用户访问内部服务器的报文的源 IP 地址转换为与内部服务器私网 IP 在同一网段内的 IP 地址。由于 NAT 服务器的内网接口的 IP 地址同时也是这个网段的缺省网关，所以内部服务器一定会将回应报文发送给 NAT 服务器。从而使内部服务器即使不配复杂的路由信息也可以正常回应报文。如图 7-19 所示。

图 7-19 NAT Inbound 应用示意图

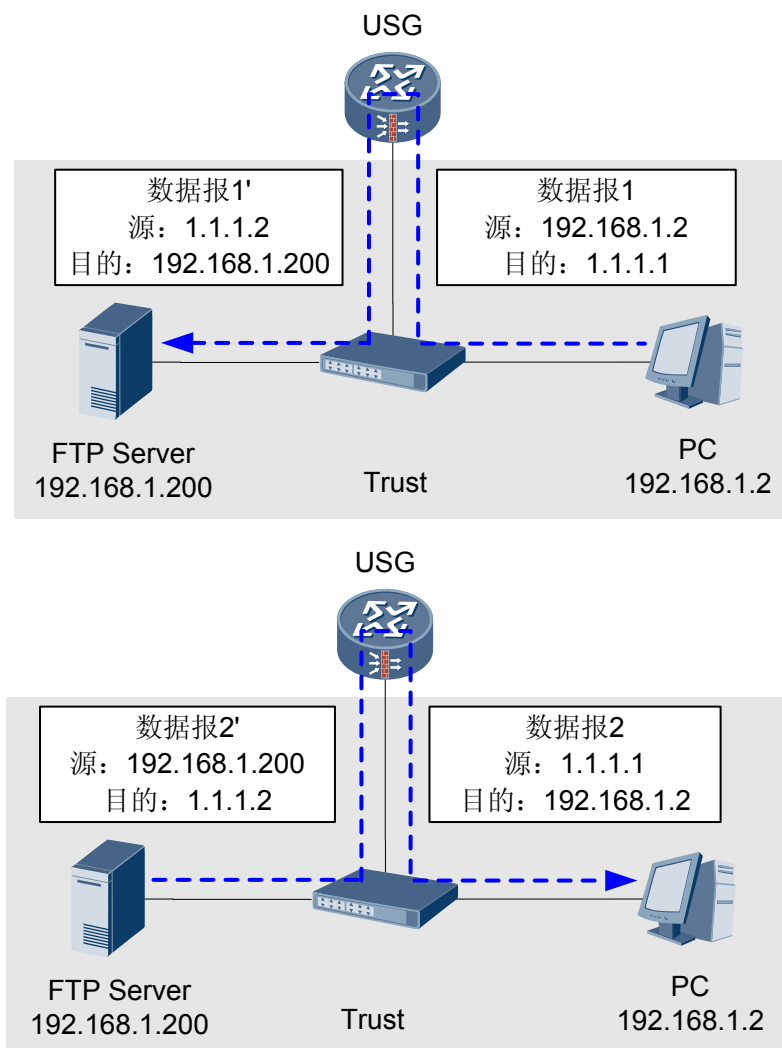


1. 公网 PC (202.120.10.2) 发往 Web 服务器 (1.1.1.1) 的数据报 1' 到达 NAT 服务器后, NAT 服务器查看报头内容, 并比对 Server-map 表中的映射关系, 发现该数据报为发往 Web 服务器的报文。
2. NAT 服务器将数据报 1' 的目的地址字段由公网地址 1.1.1.1 换成 192.168.1.2, 将源地址字段由公网地址 202.120.10.2 转换为 192.168.1.100, 作为数据报 1 发送到 Web 服务器, 同时在会话表中记录这一访问链接。
3. Web 服务器收到数据报 1' 后, 向私网 PC 发送应答报文, 即数据报 2', 源地址为 192.168.1.2, 目的地址为 192.168.1.100。由于目的地址和自己位于同一网段, 所以它将该报文发给了该网段的网关, 也就是 NAT 服务器。
4. 数据报 2 到达 NAT 服务器后, NAT 服务器查看报头内容, 查找当前会话表的记录, 用公网地址 1.1.1.1 替换源地址, 用 202.120.10.2 替换目的地址, 作为数据报 2' 发送给公网 PC。
5. 后续 PC 继续发送给 Web 服务器的报文, 设备都会直接根据会话表项的记录对其进行转换, 而不会再去查找 Server-map 表项。

## NAT Server 的一种特殊应用: 域内 NAT

在 NAT Server 的应用中, 对于一些小型局域网还存在一个特殊应用, 即内网用户和内部服务器部署在同一个安全区域中, 但是为了安全性, 希望内网用户也通过公网地址来访问该服务器。此时, 可通过配置域内 NAT 来实现。如图 7-20 所示。

图 7-20 域内 NAT 示意图



1. 私网 PC (192.168.1.2) 发往内部服务器 (1.1.1.1) 的数据报 1 到达 NAT 服务器后, NAT 服务器查看报头内容, 并比对 Server-map 表中的映射关系, 发现该数据报为发往 FTP 服务器的报文。
2. NAT 服务器从 NAT 地址池中选取一个可以在 Internet 上路由的公网地址 1.1.1.2, 然后将数据报 1 的源地址字段的私有地址 192.168.1.2 换成 1.1.1.2, 将目的地址字段由公网地址 1.1.1.1 换成 192.168.1.200。作为数据报 1 发送到 FTP 服务器, 同时在会话表中记录这一地址转换映射。
3. FTP 服务器收到数据报 1' 后, 向 PC 发送应答报文, 即数据报 2', 目的地址为 1.1.1.2。
4. 数据报 2' 到达 NAT 服务器后, NAT 服务器查看报头内容, 查找当前会话表的记录, 用私有地址 192.168.1.2 替换目的地址, 用公网地址 1.1.1.1 替换源地址, 作为数据报 2 发送给私网 PC。
5. 双方不断交互报文, 当时这个地址转换过程对两者都是透明的。PC 认为自己在与一台公网服务器通信, 而 FTP 服务器则认为自己在与一个公网用户通信。

在这个应用中, 可以看到在进行地址转换时, 同一个报文的源 IP 地址和目的 IP 地址都被替换了。所以域内 NAT 属于“双向 NAT”的一种典型应用。

在域内 NAT 的应用中，需要解决两个路由上的问题：

- 必须保证内网用户访问内网服务器的报文经过设备，而不能被交换机直接转发。这样才能进行报文的地址转换。
- 必须保证内网服务器回应给内网用户的报文经过设备，而不能被交换机直接转发。这样才能将回应报文的目的 IP 转换回用户真实的私网 IP 地址，否则用户将接收不到回应报文。

解决第一个问题的方法是，不向内网用户提供内部服务器的真实私网 IP 地址，只向他们提供内部服务器的公网 IP 地址。同时，保证内网用户将访问内部服务器公网 IP 地址的报文发给设备。这可以通过将内网 PC 的缺省网关设置为设备内网接口的 IP 地址来实现。

解决第二个问题的方法同样是保证内部服务器将访问 NAT 地址池的 IP 地址的报文发给设备。这可以通过在内部服务器上配置静态路由来实现，也可以直接将内部服务器的缺省网关配置为设备内网接口的 IP 地址。

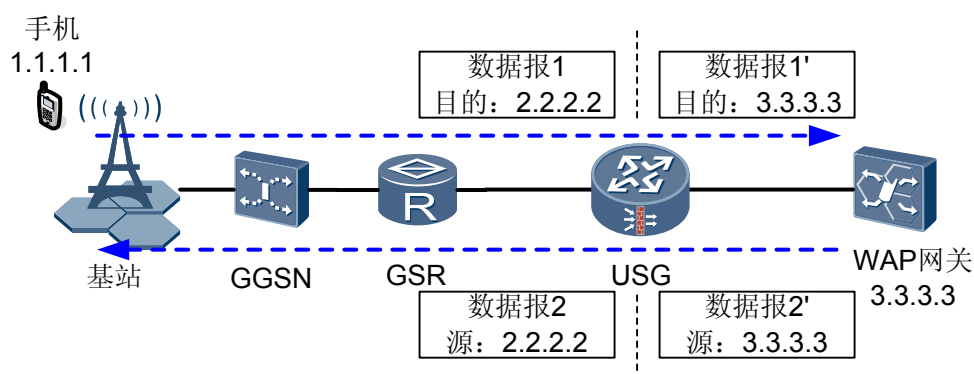
## NAT 在无线网络中的一种应用：目的 NAT

现在越来越多的手机用户直接从其他国家购买手机。而手机的上的缺省 WAP 网关地址一般均为发售地区运营商的网关地址。所以这类用户所拿到的手机的缺省 WAP 网关地址和自己所在地的运营商的 WAP 网关地址并不一致。如果手机用户不知道这一原理，也不会修改手机的缺省 WAP 网关，那么他将无法使用 WAP 上网服务。

为了解决这一问题，提出了对报文的目的地址进行转换的 NAT 转换机制。这种应用又称为“目的 NAT”。

目的 NAT 的主要原理和 NAT No-PAT 这种对报文源 IP 地址进行转换的 NAT 转换机制基本一致，区别就是此时 NAT 服务器通常部署在手机终端与运营商的 WAP 网关之间，同时目的 NAT 仅对报文的目的地址进行转换。如图 7-21 所示。

图 7-21 目的 NAT 应用示意图



在使用目的 NAT 时需要注意一点的就是允许转换的数据流一定要定义精确，以免不需要被转换的数据流也被转换。

### 7.7.4.2 NAT ALG 的实现



## 应用协议在 NAT 转换过程中的问题

通常情况下，NAT 只能对 IP 报文头中的 IP 地址及端口进行转换。但是很多协议会通过 IP 报文的数据载荷进行新端口甚至新 IP 地址的协商。协商完成之后，通信双方会根据协商结果建立新的连接进行后续报文的传输。而这些协商出来的端口和 IP 地址往往是随机的，管理员并不能为其提前配置好相应的 NAT 规则，这些协议在 NAT 转换过程中就会出现这个问题。

提出 NAT ALG 正是为了解决这一问题。NAT ALG 的主要原理就是设备通过对应用层协议的协商报文的分析，在双方的新通道建立之前，就提前获取到相关的信息，并为其自动建立相应的 NAT 转换规则。等到该通道上有数据报文传输时为其建立会话表项，使这个通道能够正常使用。

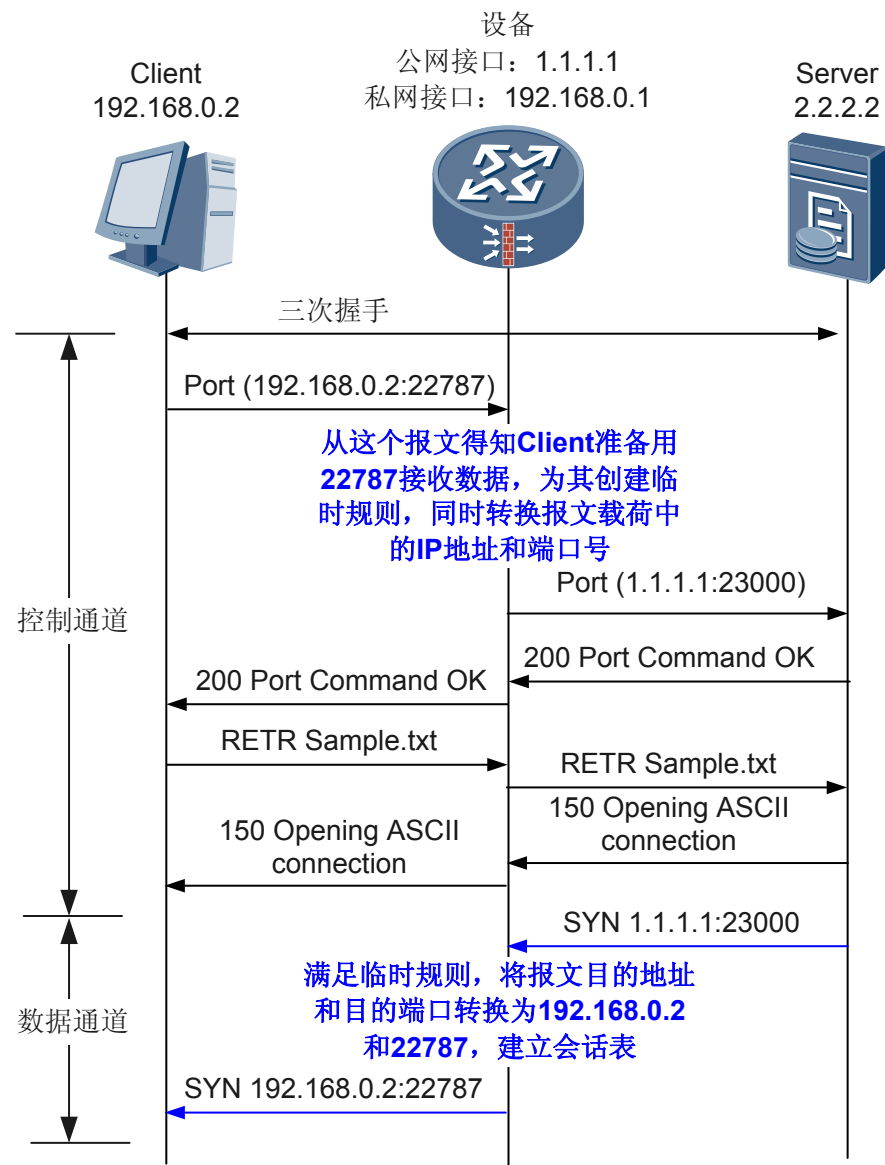
在 [7.14 Server-map 表与 ASPF](#) 中曾经提到为了实现应用层协议的包过滤而提出的 ASPF 功能。ASPF 功能的主要目的是通过对应用层协议的报文分析，为其开放相应的包过滤规则，而 NAT ALG 的主要目的，是为其开放相应的 NAT 规则。由于两者通常都是结合使用的，所以使用同一条命令就可以将两者同时开启。

## FTP 应用中的 NAT ALG 的实现

下面以典型的多通道协议 FTP 的应用中，NAT ALG 的实现为例说明 NAT ALG 的工作原理。

采用主动模式（又称为 Port 模式）的 FTP 协议的情况，如 [图 7-22](#) 所示。

图 7-22 FTP 主动模式下的 NAT ALG 应用示意图

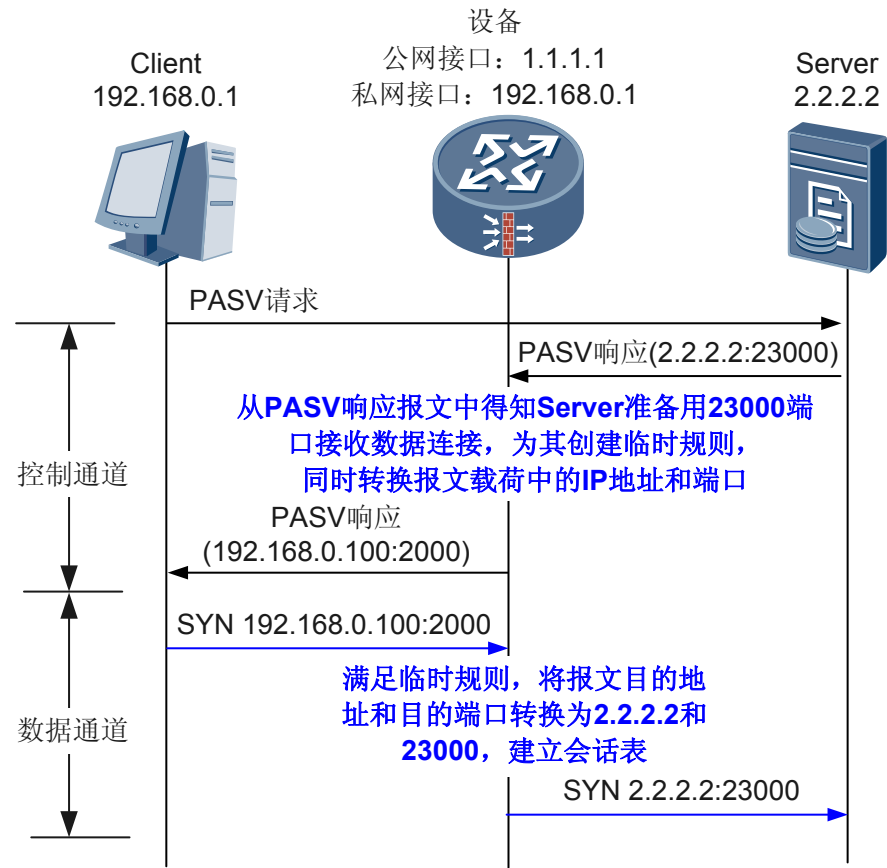


在 NAT 的情况下，Client 所选用的端口号可能已被其他用户所占用，所以设备不仅对其端口号进行了识别，还对载荷中的端口号进行了替换。以保证所有业务的正常运行。

采用被动模式（又称为 Passive 模式）的 FTP 协议的情况，如图 7-23 所示。



图 7-23 FTP 被动模式下的 NAT ALG 应用示意图



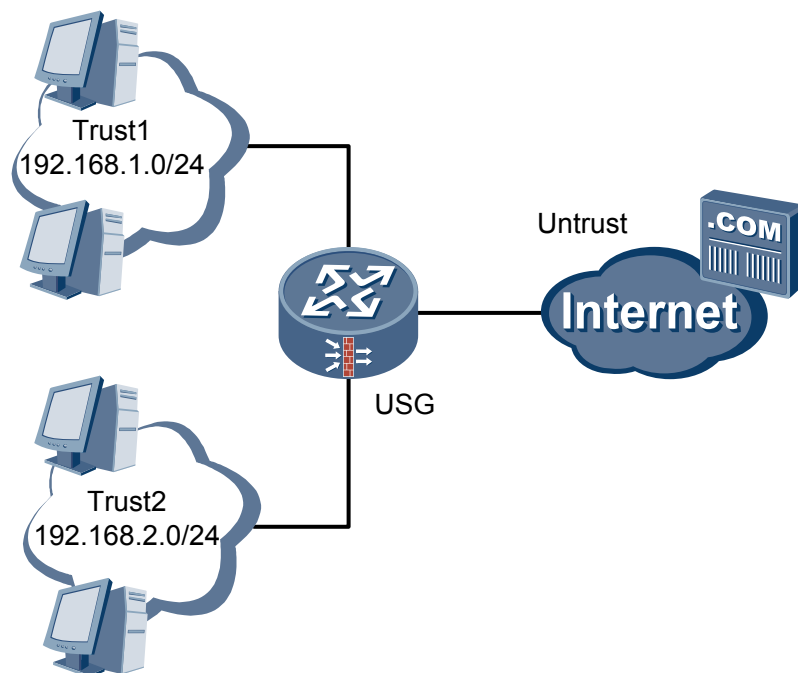
本例中配置了 NAT Inbound 以对外网报文的源 IP 地址进行转换。所以设备对 PASV 响应报文载荷中的 IP 地址和端口号都进行了转换。

7.7.4.3 NAT 组网应用

内网用户访问公网服务器

在小区、学校、公司中经常会有多个用户共享少量公网 IP 访问 Internet 中的网站、服务器的需求。这种组网通常都可以用 NAPT 来实现。将内网用户和公网服务器划入不同的安全区域，在域间应用 NAPT。如果希望对不同的用户分配不同的公网地址，可以将不同的用户划分到不同的安全区域。如图 7-24 所示。

图 7-24 内网用户访问公网服务器组网示意图

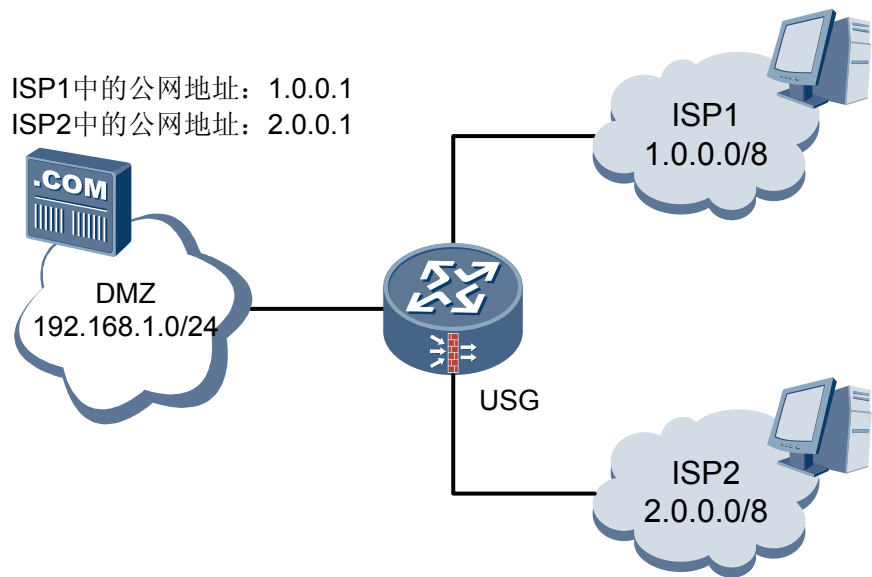


在应用该组网时，需要注意如果希望访问公网的一些特殊协议服务，需要在域间开启相应的 NAT ALG 功能。

## 公网用户访问内部服务器

部分学校或者公司都会对外提供 Web 服务、FTP 服务等业务。当时为了保证服务器本身的安全，必须将其部署在内网中。通过配置 NAT Server 可以使公网用户仍然可以访问到这些位于内网的服务器。如果这个服务器还同时对不同的运营商网络提供服务，并且在每个运营商网络中都拥有一个公网地址。可以将不同的运营商网络划分到不同的安全区域，然后对不同的安全区域发布不同的服务器地址。如图 7-25 所示。

图 7-25 公网用户访问内部服务器组网示意图

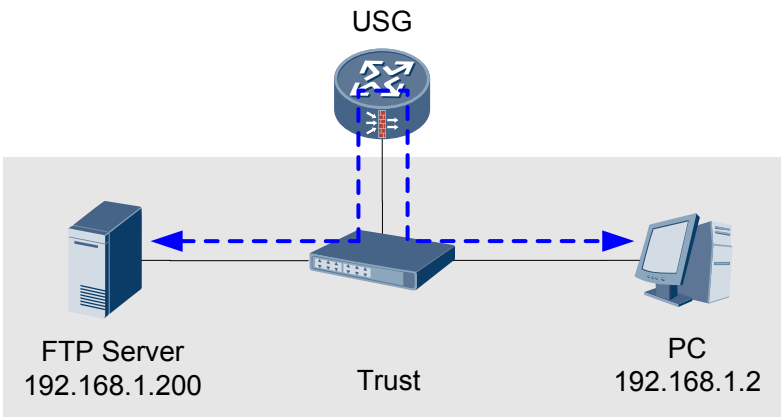


在应用该组网时，需要注意设备上的路由配置。通常跨运营商之间的访问速率比较慢，所以在配置路由时，要使服务器回应给 ISP1 的报文一定是从设备连接 ISP1 的接口发出，回应给 ISP2 的报文一定是从设备连接 ISP2 的接口发出。同时，还要根据服务器所提供的服务开启相应的 NAT ALG 功能。

### 内网用户通过公网 IP 访问位于同一安全区域的内部服务器

如果内网用户和内部服务器都是通过交换机汇聚后接入到设备上，那么就不能将内网用户和内部服务器划分到不同的安全区域中。所以此时只能通过配置域内 NAT 来实现内网用户通过公网 IP 访问内部服务器的需求。如图 7-26 所示。

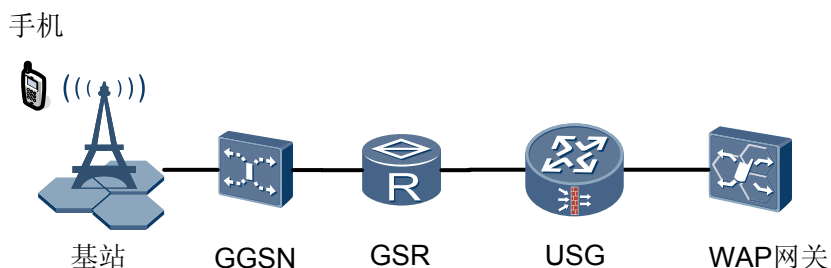
图 7-26 内网用户通过公网 IP 访问位于同一安全区域的内部服务器组网示意图



## 自动转换无线用户的 WAP 网关地址

当无线用户所使用的无线终端的缺省 WAP 网关地址和所在地运营商的 WAP 网关地址不一致时，可能导致无线用户无法访问 WAP 网络。此时可以通过配置目的 NAT 来将无线用户的报文的地址转换为正确的 WAP 网关地址。如图 7-27 所示。

图 7-27 自动转换无线用户的 WAP 网关地址组网示意图



在该组网中，一定要注意确定需转换流量时一定要精确匹配，以免不需要转换的流量被误转发给 WAP 网关。

## 7.8 NAT64

NAT64 技术可以实现 IPv6 与 IPv4 协议和地址的转换。

### 7.8.1 介绍

#### 定义

NAT64 (Stateful NAT64: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers) 是 IPv4 与 IPv6 共存时期的一种过渡技术，是一种有状态的地址与协议转换机制。

#### 目的

在 IPv4 网络向 IPv6 网络迁移的过程中，NAT64 作为一种过渡技术，主要实现了 IPv6 网络主动访问 IPv4 网络。

### 7.8.2 参考标准和协议

与 NAT64 特性相关的参考标准与协议如下：

- RFC 6052: IPv6 Addressing of IPv4/IPv6 Translators
- RFC 4443: Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification
- RFC 6144: Framework for IPv4/IPv6 Translation
- RFC 6145: IP/ICMP Translation Algorithm
- RFC 6146: Network Address and Protocol Translation from IPv6 Clients to IPv4 Servers

- RFC 6147: DNS Extensions for Network Address Translation from IPv6 Clients to IPv4 Servers

## 7.8.3 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                      |
|-------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2200/5100 V300R001、<br>USG2200/5100 BSR/HSR V300R001<br>USG5500 V300R001 |

## 7.8.4 原理描述

### 7.8.4.1 NAT64 概述

IPv4 网络在向 IPv6 网络迁移的过程中，NAT64 作为一种过渡技术，能够实现 IPv6 网络与 IPv4 网络的共存以及数据交互。

 说明

NAT64 中描述的 IPv4 网络为纯 IPv4 网络，IPv6 网络为纯 IPv6 网络。

### NAT64 转换机制

NAT64 转换机制分为以下两种方式：

- 地址转换机制  
地址转换是 IPv6 传输地址（IPv6 地址与端口，简称 IPv6 传输地址）与 IPv4 传输地址（IPv4 地址与端口，简称 IPv4 传输地址）的转换。
- 协议转换机制  
包括对 IPv6 报文头与 IPv4 报文头的转换，也包括对 TCP、UDP、ICMP 的协议转换。

NAT64 设备（支持 NAT64 功能的设备，简称为 NAT64 设备）至少要存在两个接口，分别连接 IPv6 网络和 IPv4 网络，接口分别称为 IPv6 接口和 IPv4 接口。

### NAT64 地址池

为了创建 IPv6 与 IPv4 的地址映射关系，NAT64 设备需要存在 IPv6 地址池与 IPv4 地址池。

- IPv6 地址池

IPv6 地址池是 NAT64 设备上配置的一个或多个 IPv6 前缀。缺省已经存在知名前缀 64:ff9b::/96。

- IPv4 地址池  
IPv4 地址池是手工配置的一组 IPv4 地址。由于 IPv4 资源稀少，因此，通常多个 IPv6 地址映射到一个 IPv4 地址上，可以有效提高 IPv4 地址的利用率。

NAT64 映射方式

NAT64 设备能够实现静态映射和动态映射两种映射方式。

- NAT64 静态映射方式  
NAT64 的静态映射是针对地址的转换机制。NAT64 静态映射方式属于“一对一的地址转换方式”。
- NAT64 动态映射方式  
NAT64 动态映射基于 NAT64 的域间策略实现。IPv6 传输地址与 IPv4 传输地址的映射关系动态创建。IPv6 转换后的 IPv4 地址既可以是 IPv4 地址池中的 IPv4 地址，也可以是接口的 IPv4 地址（Easy IP 功能）。NAT64 动态映射方式属于“多对一的地址转换方式”。

7.8.4.2 NAT64 地址转换原理

NAT64 地址转换时应用的地址格式为 IPv4 嵌入的 IPv6 地址格式。

通常情况下，IPv6 客户端借助 DNS64 的域名解析功能，获得 IPv6 目的地址，其地址格式为 IPv4 嵌入的 IPv6 地址格式。

IPv4 嵌入的 IPv6 地址

IPv4 嵌入的 IPv6 地址是由前缀，嵌入的 IPv4 地址和后缀组成。格式如图 7-28 所示。

图 7-28 IPv4 嵌入的 IPv6 地址

| PL | 0      | 32 | 40 | 48      | 56     | 64  | 72 | 80      | 88     | 96      | 104 |
|----|--------|----|----|---------|--------|-----|----|---------|--------|---------|-----|
| 32 | prefix |    |    | V4 (32) |        |     | U  | suffix  |        |         |     |
| 40 | prefix |    |    | V4 (24) |        |     | U  | (8)     | suffix |         |     |
| 48 | prefix |    |    |         | V4(16) |     | U  | (16)    | suffix |         |     |
| 56 | prefix |    |    |         |        | (8) | U  | V4 (24) |        | suffix  |     |
| 64 | prefix |    |    |         |        | U   |    | V4 (32) |        | suffix  |     |
| 96 | prefix |    |    |         |        |     |    |         |        | V4 (32) |     |

- prefix 可以是知名前缀或自定义前缀。
- PL（Prefix Length）表示前缀长度。前缀长度的可选值为：32，40，48，56，64 或 96，其中知名前缀的前缀长度只能选用 96。
- U 为保留位，取值始终为零。当前缀长度取值为 96 时，需要确保 64 位～71 位设置为 0。
- v4 表示嵌入的 IPv4 地址。根据前缀长度的不同，32 位的 IPv4 地址可能会被保留位 U 分割。

- suffix 为保留位，取值始终为零。当前缀长度为 96 时，因为地址中没有剩余位，因此该地址没有后缀。而在前缀长度为其他可选情况时，剩余位将由后缀构成。

IPv4 地址嵌入 IPv6 地址时，根据前缀长度不同，嵌入的位置存在差异。具体差异如图 7-28 所示：

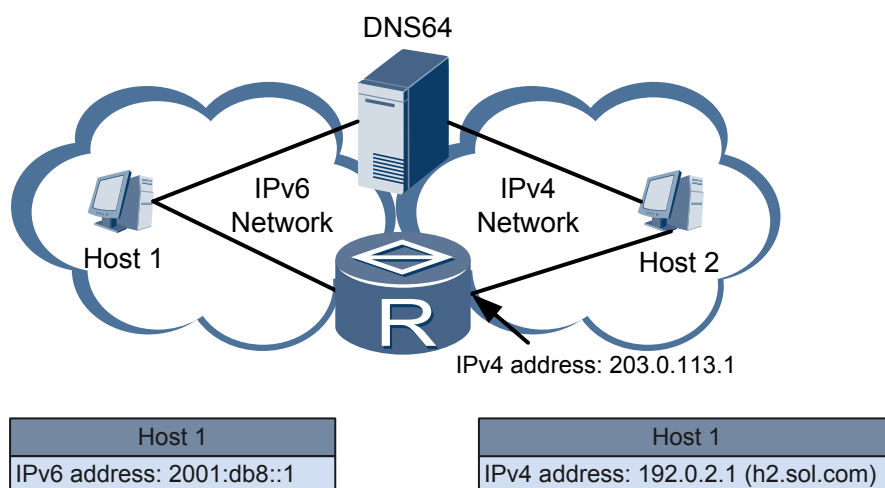
- 当前缀长度为 32 位时，IPv4 地址嵌入 IPv6 地址的位置为 32 位～63 位。
- 当前缀长度为 40 位时，24 位的 IPv4 地址被嵌入到 IPv6 地址的 40 位～63 位，剩余 8 位的 IPv4 地址被嵌入到 IPv6 地址的 72 位～79 位。
- 当前缀长度为 48 位时，16 位的 IPv4 地址被嵌入到 IPv6 地址的 48 位～63 位，剩余 16 位的 IPv4 地址被嵌入到 IPv6 地址的 72 位～87 位。
- 当前缀长度为 56 位时，8 位的 IPv4 地址被嵌入到 IPv6 地址的 56 位～63 位，剩余 24 位的 IPv4 地址被嵌入到 IPv6 地址的 72 位～95 位。
- 当前缀长度为 64 位时，IPv4 地址被嵌入到 IPv6 地址的 72 位～103 位。
- 当前缀长度为 96 位时，IPv4 地址被嵌入到 IPv6 地址的 96 位～127 位。

### 7.8.4.3 NAT64 工作原理

#### NAT64 转换

以 IPv6 网络中的 IPv6 主机主动建立到 IPv4 网络的连接为例。如图 7-29 所示，Host 1 主动建立到 Host 2 的连接。设备连接 IPv6 网络和 IPv4 网络。NAT64 的前缀为知名前缀 64:ff9b::/96。接口 IPv4 地址为 203.0.113.1，采用 Easy IP 方式的 NAT64 转换。IPv6 地址将转换为接口的 IPv4 地址。接口 IPv4 地址与 Host 2 存在可达的路由。DNS64 上存在同样的知名前缀 64:ff9b::/96。

图 7-29 NAT64 场景图



在 NAT64 转换过程中，Host 1 与 Host 2 通信工作步骤为：

1. Host 1 发送访问 h2.sol.com 的 DNS 请求后，经 DNS64 处理后，收到由知名前缀和 IPv4 地址组成的 IPv6 地址，为 64:ff9b::192.0.2.1。

2. Host 1 发送报文到 Host 2。报文从源传输地址[2001:db8::1,1500]发送到目的传输地址[64:ff9b::192.0.2.1,80]。
3. 报文被路由转发至 NAT64 设备的 IPv6 接口。
4. 设备收到报文后的处理流程:
  - a. 报文匹配 NAT64 知名前缀 64:ff9b::/96, 根据地址转换算法提取出 IPv4 目的地址。
  - b. 由于采用 Easy IP 方式, 设备会选择接口 IPv4 地址 203.0.113.1 作为 IPv4 源地址。同时会选择一个未使用的端口(例如为 2000)作为 IPv4 源端口。
  - c. NAT64 将 IPv6 报文转换为 IPv4 报文。
  - d. NAT64 转换后, (203.0.113.1,2000)作为源传输地址, (192.0.2.1,80)为目的传输地址。
5. NAT64 设备从 IPv4 接口发送转换后的报文到 Host 2。
6. Host 2 收到报文后发送响应报文, 此时源传输地址为(192.0.2.1,80), 目的传输地址为(203.0.113.1,2000)。
7. 203.0.113.1 作为设备的 IPv4 接口, 因此报文能够被路由到设备。由于设备上已经存在映射关系(2001:db8::1,1500)<--> (203.0.113.1,2000), 因此能够进行如下操作:
  - a. NAT64 将 IPv4 报文转换为 IPv6 报文。
  - b. NAT64 转换后, (2001:db8::1,1500)作为目的传输地址, (64:ff9b::192.0.2.1,80)作为源传输地址。
8. 转换后的报文从设备的 IPv6 接口发送给 Host 1。

## 7.9 负载均衡

负载均衡可以分别利用各个服务器的处理能力, 达到流量分担的目的, 而且保障了服务器的可用性, 得到最佳的网络扩展性。通常配置在内部网络的出口处。

### 7.9.1 介绍

#### 定义

负载均衡就是 USG 按照事先配置的负载均衡算法, 将访问同一个 IP 地址的用户流量分配到不同的服务器上。

负载均衡的典型的应用是将 USG 放在私网出口上。

负载均衡采用以下方法, 将用户流量分配到多台服务器:

- 虚服务技术
- 服务器健康性检测
- 基于流的转发

#### 目的

负载均衡有如下作用:

配置负载均衡后, 在访问用户看来, 他们访问的是同一个服务器, 而实际上 USG 将他们的请求分送给了不同的服务器进行处理。这样不但可以分别利用各个服务器的处理能力, 达到流量分担的目的, 而且保障了服务器的可用性, 得到最佳的网络扩展性。



## 7.9.2 可获得性

### License 支持

负载均衡功能不需要 License 支持。

### 版本支持

负载均衡的版本支持如表 7-6 所示。

表 7-6 产品及最低版本支持

| 产品                                                                      | 最低支持版本                                                                                                   |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100 BSR/<br>HSR V100R005C00<br>USG5500 V200R001C00 |

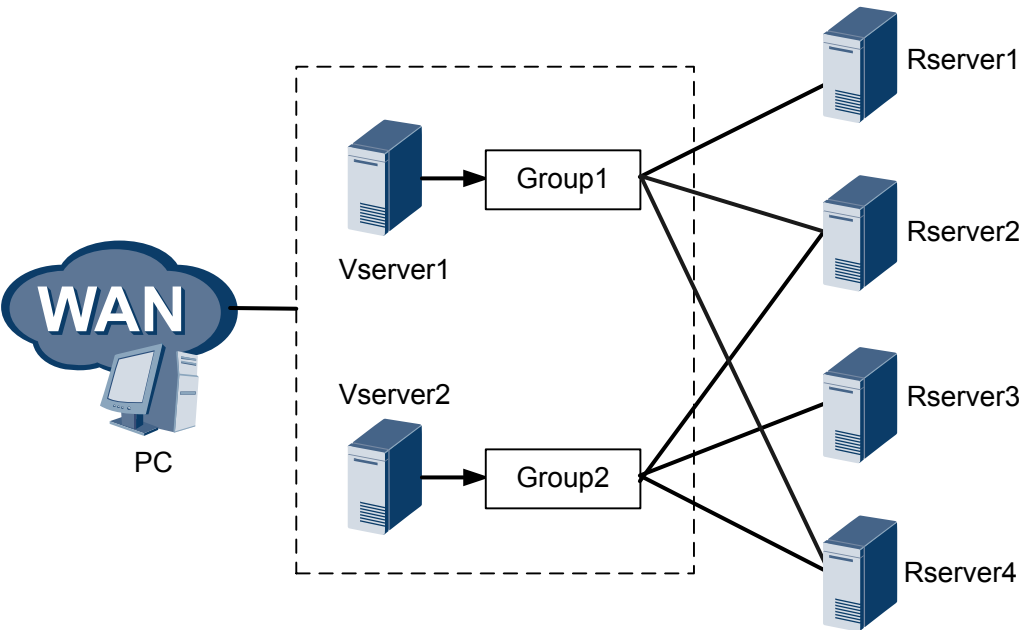
## 7.9.3 原理描述

### 7.9.3.1 虚服务技术

实际的服务器被称作真实服务器，每一个真实服务器有不同的私网 IP 地址（即实 IP 地址），但是所有真实服务器对外表现为一个公网 IP 地址。这个公网 IP 地址对应一个虚服务器，USG 将访问虚服务器的流量按照预先配置的负载均衡算法分配到每一个真实服务器。

为方便管理，在虚服务器（Vserver）和真实服务器（Rserver）之间通过服务器组 Group 进行衔接。Group 是一个逻辑概念，USG 通过 Group 对真实服务器进行管理，提供网络服务。VServer、Group 和 Rserver 之间对应关系如图 7-30 所示。

图 7-30 虚服务示意图



采用虚服务技术有如下优点：

- 节约公网 IP 地址
- 提高系统的安全性
- 提高系统的可扩展性

### 7.9.3.2 服务器健康性检测

USG 周期性探测真实服务器，向真实服务器发送健康性检测报文。如果能收到真实服务器回应的报文，说明真实服务器可用；如果多次收不到服务器的回应报文，USG 将禁止使用该真实服务器，将流量按配置好的策略分配到其他真实服务器上。

### 7.9.3.3 基于流的转发

USG 可以通过指定的算法，将数据流分发到各个真实服务器进行处理。USG 支持源地址哈希算法、简单轮询算法和加权轮询算法三种负载均衡算法。

## 7.10 攻击防范

攻击防范功能能够检测出多种类型的网络攻击，并能采取相应的措施保护内部网络免受恶意攻击，保证内部网络及系统的正常运行。本节介绍攻击防范的基本概念、规格及原理。

### 7.10.1 介绍

#### 定义

攻击防范是一种安全机制。它通过对报文内容或者行为的分析,判断报文是否具有攻击特征，并根据配置对攻击报文执行防范措施。

#### 目的

TCP/IP 在设计之初主要是为了实现网络的互通，在安全性方面考虑较少。所以随着计算机网络的发展，很多黑客利用 TCP/IP 协议族和网络设备固有的缺陷，通过各种手段对网络设备进行攻击，以达成个人的目的。

所以攻击防范就成为网关设备的一个基本功能，同时也是网关设备提升网络安全性的一个重要组成部分。

### 7.10.2 可获得性

#### License 支持

本特性无需 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 特性依赖

TCP 全连接攻击防范、IP 地址扫描攻击防范、端口地址扫描攻击防范必须在黑名单功能开启情况下才能防范成功。

## 7.10.3 原理描述

### 攻击防范概述

通常的网络攻击，一般是侵入或破坏网上的服务器（主机），盗取服务器的敏感数据或占用网络带宽，干扰破坏服务器对外提供的服务。也有直接破坏网络设备的网络攻击，这种破坏影响较大，会导致网络服务异常，甚至中断。

网络攻击主要分为流量型攻击，扫描窥探攻击，畸形报文攻击和特殊报文攻击四大类：

- 流量型攻击  
流量型攻击是指攻击者通过大量的无用数据占用过多的资源以达到服务器拒绝服务的目的。这类攻击典型特征是通过发出海量数据包，造成设备负载过高，最终导致网络带宽或是设备资源耗尽。通常，被攻击的路由器、服务器和防火墙的处理资源都是有限的，攻击负载之下它们就无法处理正常的合法访问，导致正常服务被拒绝。流量型攻击最通常的形式是 Flood 方式，这种攻击把大量看似合法的 TCP、UDP、ICMP 包发送至目标主机，甚至，有些攻击者还利用伪造源地址来绕过检测系统的监控，以达到攻击的目的。
- 扫描窥探攻击  
扫描窥探攻击主要包括 IP 地址扫描和端口扫描，IP 地址扫描是指攻击者发送目的地址不断变化的 IP 报文（TCP/UDP/ICMP）来发现网络上存在的主机和网络，从而准确的发现潜在的攻击目标。端口扫描是指通过扫描 TCP 和 UDP 的端口，检测被攻击者的操作系统和潜在服务。攻击者通过扫描窥探就能大致了解目标系统提供的服务种类和潜在的安全漏洞，为进一步侵入系统做好准备。
- 畸形报文攻击  
畸形报文攻击是指通过向目标系统发送有缺陷的 IP 报文，使得目标系统在处理这样的 IP 报文时发生错误，或者造成系统崩溃，影响目标系统的正常运行。主要的畸形报文攻击有 Ping of Death、Teardrop 等。
- 特殊报文攻击  
特殊报文攻击是指攻击者利用一些合法的报文对网络进行侦察或者数据检测，这些报文都是合法的应用类型，只是正常网络很少用到。

USG 通过多层过滤技术可以有效防御各种攻击，精准区别各种攻击流量和正常流量，并能采取相应的措施保护内部网络免受恶意攻击，保证内部网络及系统的正常运行。

## 各类攻击攻防原理简介

下面简单介绍一下各种攻击的简介原理，如表 7-7 所示。

表 7-7 攻防原理简介

| 分类    | 攻击类型         | 攻击原理                                                                                                                                                                                                                   | 防范原理                                                                                                                                                                                                                                                                                                                                                                   |
|-------|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 流量型攻击 | SYN Flood 攻击 | 由于网络设备资源的限制，TCP/IP 栈的实现只能允许有限个 TCP 连接。而 SYN Flood 攻击正是利用这一点，伪造一个 SYN 报文（其源地址是伪造的或者是一个不存在的地址）向服务器发起连接，服务器在收到报文后用 SYN-ACK 应答，而此应答发出去后，不会收到 ACK 报文，造成一个半连接。如果攻击者发送大量这样的 SYN 报文，会在被攻击主机上出现大量的半连接，耗尽其资源，使正常的用户无法访问，直到半连接超时。 | <ul style="list-style-type: none"> <li>● 超过阈值丢弃：为 SYN 报文的接收速率设定上限，如果超过上限则进行丢弃。</li> <li>● TCP 反向源探测：原理是攻击者使用伪造的 IP 地址发送报文，所以当设备接收到 SYN 报文时，会对源 IP 是否存在进行探测，如果源 IP 不存在，则说明可能是攻击报文，将其予以丢弃。如果源 IP 有效，则将正确的源 IP 地址加入白名单，后续一段时间里，此源 IP 地址的 TCP 报文直接转发。</li> <li>● TCP 代理：设备充当两端之间的 TCP 连接的代理，当一端发起连接后，必须先与设备完成 TCP 的三次握手，设备才允许其与另一端建立 TCP 连接。这样可以防止在被保护对象上形成半连接。</li> </ul> |
|       | TCP 全连接攻击    | 与 SYN Flood 不同，TCP 全连接攻击是指攻击者与被攻击对象正常建立了 TCP 全连接，但是却没有后续的报文，占用被攻击者的资源的攻击类型。                                                                                                                                            | <ol style="list-style-type: none"> <li>1. 异常识别：设定报文速率下限，将 TCP 连接建立后报文传输速率低于下限的连接识别为异常连接。</li> <li>2. 将攻击者加入黑名单：设定异常连接上限，将建立多个异常连接的 IP 加入黑名单，直接丢弃其报文。</li> </ol>                                                                                                                                                                                                        |

| 分类 | 攻击类型          | 攻击原理                                                                           | 防范原理                                                                                                                                                                                                                                                                                         |
|----|---------------|--------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | HTTP Flood 攻击 | HTTP Flood 攻击是指攻击者直接或者间接向目标服务器发起大量的 HTTP 报文，导致服务器消耗过重，无法响应正常的请求，严重时会导致主干链路拥塞。  | <p>对接收的 HTTP 报文速率进行限制，当接收的 TCP 报文速率超过设定的阈值时，则视为攻击，并给予丢弃。</p> <p>针对 HTTP Flood 的防范可以设置重定向检测功能。将 HTTP 报文速率超过告警阈值的访问者重定向至一个虚拟页面，如果访问者合法，它将做出合理回应，如果访问者非法则不会做回应。以此来判断 HTTP 报文的来源是否真实存在。</p> <p>对合法 IP 会再次重定向回正常页面，整个过程由于时间很短，对于用户来说是不可见的，后续该 IP 的 HTTP 报文也能得到正常转发；而对于虚假 IP 发送的 HTTP 报文则予以丢弃。</p> |
|    | UDP Flood 攻击  | 攻击者通过向服务器发送大量的 UDP 报文，占用服务器的链路带宽，导致服务器负担过重而不能正常向外提供服务。                         | 设定 UDP 报文速率上限，对超过部分进行丢弃。                                                                                                                                                                                                                                                                     |
|    | ICMP Flood 攻击 | ICMP Flood 攻击是指攻击者通过向服务器发送大量的 ICMP 报文（如 Ping），占用服务器的链路带宽，导致服务器负担过重而不能正常向外提供服务。 | 设定 ICMP 报文速率上限，对超过部分进行丢弃。                                                                                                                                                                                                                                                                    |
|    | ARP Flood 攻击  | 由于 ARP 协议中存在多种广播类型的报文，所以 ARP 报文经常被用来作流量攻击。                                     | 设定 ARP 报文速率上限，对超过部分进行丢弃。                                                                                                                                                                                                                                                                     |

| 分类     | 攻击类型         | 攻击原理                                                                                                  | 防范原理                                                                                                                                                                                                                                                                                                                                          |
|--------|--------------|-------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | SIP Flood 攻击 | 攻击者在短时间内向服务器发送大量的 SIP 报文，从而达到耗尽服务器 CPU 或带宽的目的。                                                        | <ul style="list-style-type: none"> <li>● 超过阈值丢弃：为 SIP 报文的接收速率设定上限，如果超过上限则进行丢弃。</li> <li>● SIP 反向源探测：原理是攻击者使用伪造的 IP 地址发送报文，所以当设备接收到 SIP 报文时，会对源 IP 是否存在进行探测，如果源 IP 不存在，则说明可能是攻击报文，将其予以丢弃。如果源 IP 有效，则将正确的源 IP 地址加入白名单，后续一段时间里，此源 IP 地址的 SIP 报文直接转发。</li> </ul>                                                                                  |
| 扫描窥探攻击 | IP 地址扫描攻击    | 攻击者运用 ICMP 报文探测目标地址，或者使用 TCP/UDP 报文对一定地址发起连接，通过判断是否有应答报文，以确定哪些目标系统确实存活并且连接在目标网络上。IP 地址扫描通常是为其其他攻击做准备。 | <p>配置 IP 地址扫描攻击防范参数后，设备对进入的 TCP，UDP，ICMP 报文进行检测，并以某个源 IP 地址为索引，判断该源 IP 地址发送报文的源 IP 与前一报文的源 IP 地址是否不同，如果是则异常次数加 1。当异常频率超过预定义的阈值时，则认为该源 IP 地址的报文为 IP 地址扫描攻击，并将该源 IP 地址加入黑名单，然后做如下处理：</p> <ul style="list-style-type: none"> <li>● 若开启黑名单功能，则从该源 IP 地址发出的报文命中黑名单被全部丢弃。</li> <li>● 若没有开启黑名单功能，则从该源发出的报文，如果扫描速率超过了设定的阈值，系统会产生告警，但不丢弃报文。</li> </ul> |

| 分类     | 攻击类型     | 攻击原理                                                                                                                               | 防范原理                                                                                                                                                                                                                                                                                                                                   |
|--------|----------|------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | 端口扫描攻击   | 攻击者通常使用 Port Scan 攻击软件，向大范围的主机的一系列 TCP/UDP 端口发起连接，根据应答报文判断主机是否使用这些端口提供服务。端口扫描通常是为其他攻击做准备。                                          | 配置端口扫描攻击防范参数后，设备对进入的 TCP，UDP，ICMP 报文进行检测，并以某个源 IP 地址为索引，判断该源 IP 地址发送报文的目的 MAC 与前一报文的目的 MAC 地址是否不同，如果是则异常次数加 1。当异常频率超过预定义的阈值时，则认为该源 IP 地址的报文为 IP 地址扫描攻击，并将该源 IP 地址加入黑名单，然后做如下处理： <ul style="list-style-type: none"> <li>● 若开启黑名单功能，则从该源 IP 地址发出的报文命中黑名单被全部丢弃。</li> <li>● 若没有开启黑名单功能，则从该源发出的报文，如果扫描速率超过了设定的阈值，系统会产生告警，但不丢弃报文。</li> </ul> |
| 畸形报文攻击 | Smurf 攻击 | Smurf 攻击的原理是攻击者以被攻击者的 IP 地址向网络中的广播地址发送 ICMP 请求报文，使得网络中的所有主机向被攻击者回应 ICMP 应答报文，造成受害者系统繁忙，链路拥塞。                                       | 检查 ICMP 请求报文的目的地址是否为网络的 A、B、C 类广播地址或子网广播地址，如果是则丢弃。                                                                                                                                                                                                                                                                                     |
|        | Land 攻击  | 攻击者发送 SYN 报文的源地址和目标地址同为被攻击者的 IP 地址，导致被攻击者向其自己的地址发送 SYN-ACK 消息，因此存在大量的空连接。不同被攻击者对 Land 攻击反应不同，UNIX 主机将崩溃，Windows NT 主机的运行状态会变的极其缓慢。 | 检查 TCP 报文的源地址和目的地址是否相同，或者 TCP 报文的源地址是否为环回地址，如果是则丢弃。                                                                                                                                                                                                                                                                                    |

| 分类 | 攻击类型       | 攻击原理                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 防范原理                                                                                                                                                                                                 |
|----|------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | Fraggle 攻击 | <p>主机在收到目的端口为 7（ECHO）和 19（Chargen）的 UDP 报文后，都会产生回应。主机在收到目的端口为 7 的 UDP 报文后，会像 ICMP Echo Reply 一样回应收到的内容；而当主机收到目的端口为 19 的 UDP 报文后，会产生一串字符流。就像 ICMP 一样，这两个 UDP 端口都会产生大量的应答报文，占据网络带宽。</p> <p>攻击者可以向攻击目标所在的网络发送 UDP 报文，报文的源地址为被攻击主机的地址，目的地址为被攻击主机所在子网的广播地址或子网网络地址，目的端口号为 7 或 19。子网中启用了此功能的每个系统都会向被攻击主机发送回应报文，从而产生大量的流量，占满带宽，导致受害网络的阻塞或受害主机的崩溃。</p> <p>即使子网上没有启动这些功能的系统也将产生一个 ICMP 不可达消息，因而仍然消耗带宽。若攻击者将 UDP 报文的源端口改为 19，目的端口为 7，这样会不停地产生大量回应报文，其危害性更大。</p> | <p>进入的 UDP 报文进行检测，若目的端口号为 7 或 19，则直接拒绝，并将攻击记录到日志。</p>                                                                                                                                                |
|    | IP 分片报文攻击  | <p>IP 报文头中的 DF 和 MF 标志位用于分片控制，攻击者通过发送分片控制非法的报文，从而导致主机接收报文时产生故障，报文处理异常，甚至导致主机崩溃。</p>                                                                                                                                                                                                                                                                                                                                                                  | <p>将检测到报文控制位是下列情况之一时，直接丢弃并记录攻击日志：</p> <ul style="list-style-type: none"> <li>● DF 位为 1，而 MF 位也为 1。</li> <li>● DF 位为 1，而 Offset &gt; 0。</li> <li>● DF 位为 0，而分片 Offset + Length &gt; 65535。</li> </ul> |



| 分类 | 攻击类型             | 攻击原理                                                                                                                                                                                                                                                               | 防范原理                                                              |
|----|------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|
|    | IP 欺骗攻击          | <p>IP 欺骗攻击是一种常用的攻击方法，同时也是其他攻击方法的基础。这是由 IP 协议自身的特点决定的，IP 协议依据 IP 头中的目的地址来发送 IP 报文，如果 IP 报文是本网络内的地址，则被直接发送到目的地址；如果该 IP 地址不是本网络地址，则被发送到网关，而不对 IP 包中提供的源地址做任何检查，默认为 IP 包中的源地址就是发送 IP 包主机的地址。</p> <p>攻击者通过向目标主机发送源 IP 地址伪造的报文，欺骗目标主机，从而获取更高的访问和控制权限。该攻击导致危害资源，信息泄漏。</p> | 对报文的源 IP 地址进行 FIB 表反查，如果反查该 IP 地址的出接口与报文的入接口不相同，则视为 IP 欺骗攻击，给予处理。 |
|    | ARP 欺骗攻击         | 攻击者为了改变网关设备上的 ARP 表项，会向网关设备发送虚假 ARP 请求报文。如果设备将该报文中的 IP 地址与 MAC 地址对应关系学习至 ARP 表项，则会误将其他主机的报文转发给攻击源。                                                                                                                                                                 | 只学习正确的 ARP 报文，丢弃虚假的 ARP 报文。                                       |
|    | Ping of Death 攻击 | 网络设备对数据包的大小是有限制的，IP 报文的长度字段为 16 位，即 IP 报文的最大长度为 65535 字节。如果遇到大小超过 65535 字节的报文，会出现内存分配错误，从而使接收方死机。攻击者只需不断的通过 Ping 命令向攻击目标发送超过 65535 字节的报文，就可以使目标对象的 TCP/IP 堆栈崩溃，致使接收方死机。                                                                                            | 检测数据包的大小是否大于 65535 字节，对大于 65535 字节的数据包，则丢弃。                       |

| 分类 | 攻击类型        | 攻击原理                                                                                                                                                                                                                                             | 防范原理                                                                                                                                                                                                                       |
|----|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | TCP 报文标志位攻击 | TCP 报文标志位包括 URG、ACK、PSH、RST、SYN、FIN。攻击者通过发送非法 TCP flag 组合的报文，对主机造成危害。                                                                                                                                                                            | 检查 TCP 报文的各个标志位，若出现以下情况，则视为攻击，予以丢弃并记录攻击日志： <ul style="list-style-type: none"> <li>● 6 个标志位全为 1。</li> <li>● 6 个标志位全为 0。</li> <li>● SYN 和 FIN 位同时为 1。</li> <li>● SYN 和 RST 位同时为 1。</li> <li>● FIN 位为 1，而 ACK 位为 0</li> </ul> |
|    | UDP 短头报文攻击  | 正常情况下，UDP 报文头的最短长度为 8 个字节。报文头长度小于 8 字节的称为 UDP 短头报文，这些报文通常是攻击报文。                                                                                                                                                                                  | UDP 短头报文攻击防范的原理就是检测进入的 UDP 报文长度。如果报文长度不符合要求，则认为该 UDP 报文是攻击报文。                                                                                                                                                              |
|    | Teardrop 攻击 | 对于一些大的 IP 数据包，为了满足链路层的 MTU (Maximum Transmission Unit) 的要求，需要传送过程中对其进行分片，分成几个 IP 包。在每个 IP 报头中有一个偏移字段和一个拆分标志 (MF)，其中偏移字段指出了这个片段在整个 IP 包中的位置。如果攻击者截取 IP 数据包后，把偏移字段设置成不正确的值，接收端在收到这些分拆的数据包后，就不能按数据包中的偏移字段值正确组合出被拆分的数据包，这样，接收端会不停的尝试，以至操作系统因资源耗尽而崩溃。 | 对接收到的分片数据包进行分析，计算数据包的偏移量是否有误，如果有误则丢弃，并记录攻击日志。                                                                                                                                                                              |
|    | WinNuke 攻击  | WinNuke 攻击又称“带外传输攻击”，它的特征是攻击目标端口，被攻击的目标端口通常是 139，而且 URG 位设为 1，即紧急模式。WinNuke 攻击是利用 Windows 操作系统的漏洞，向端口发送一些携带 TCP 带外 (OOB) 数据报文，但这些攻击报文与正常携带 OOB 数据报文不同，其指针字段与数据的实际位置不符，即存在重合，这样 Windows 操作系统在处理这些数据时，就会崩溃。                                        | 首先判断数据包目标端口是否为 139，并判断 URG 位是否为 1，以及 URG 指针不为空，同时满足以上三个条件的报文将被丢弃，并记录攻击日志。                                                                                                                                                  |

| 分类     | 攻击类型            | 攻击原理                                                                                                                                   | 防范原理                                                                       |
|--------|-----------------|----------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| 特殊报文攻击 | 超大 ICMP 报文攻击    | 超大 ICMP 报文攻击是指利用长度超大的 ICMP 报文对目标系统进行攻击。对于有些系统，在接收到超大 ICMP 报文后，由于处理不当，会造成系统崩溃、死机或重启。                                                    | 根据实际网络需要配置允许通过的 ICMP 报文的长度。当实际 ICMP 报文的长度超过该值时，设备认为发生了超大 ICMP 报文攻击，将丢弃该报文。 |
|        | ICMP 重定向报文攻击    | 网络设备通常通过向同一个子网的主机发送 ICMP 重定向报文来请求主机改变路由。一般情况下，设备仅向主机而不向其他设备发送 ICMP 重定向报文。但一些恶意的攻击可能跨越网段向另外一个网络的主机发送虚假的重定向报文，以改变主机的路由表，干扰主机正常的 IP 报文发送。 | 对 ICMP 重定向报文进行丢弃，并记录攻击日志。                                                  |
|        | ICMP 不可达报文攻击    | 有的系统在收到网络或主机不可达的 ICMP 报文后，对于后续发往此目的地的报文直接认为不可达，从而切断了目的地与主机的连接。攻击者利用这一点，伪造不可达 ICMP 报文，切断受害者与目的地的连接，造成攻击。                                | 对 ICMP 不可达报文进行丢弃并记录攻击日志。                                                   |
|        | 带路由记录项的 IP 报文攻击 | 在 IP 路由技术中，提供了路由记录选项，用来记录 IP 报文从源地址到目的地址过程中所经过的路径，也就是一个处理此报文的路由器的列表。IP 路由记录选项通常用于网络路径的故障诊断，但也会被恶意攻击者利用，窥探网络结构。                         | 检测进入路由器的报文是否设置 IP 路由记录选项，如果是则丢弃报文。                                         |

| 分类 | 攻击类型            | 攻击原理                                                                                                                                                                                                                        | 防范原理                                                    |
|----|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------|
|    | 带源路由选项的 IP 报文攻击 | 在 IP 路由技术中，一个 IP 报文的传递路径是由网络中的路由器根据报文的目地地址来决定的，但也提供了一种由报文的发送方决定报文传递路径的方法，这就是源站选路选项。源站选路选项允许源站明确指定一条到目的地的路由，覆盖掉中间路由器的路由选项。源站选路选项通常用于网络路径的故障诊断和某种特殊业务的临时传送。由于 IP 源站选路选项忽略了报文传输路径中的各个设备的中间转发过程，而不管转发接口的工作状态，可能被恶意攻击者利用，窥探网络结构。 | 检测进入的报文是否设置 IP 源站选路选项，如果是则丢弃该报文。                        |
|    | Tracert 报文攻击    | Tracert 报文攻击是攻击者利用 TTL 为 0 时返回的 ICMP 超时报文，和达到目的地址时返回的 ICMP 端口不可达报文来发现报文到达目的地所经过的路径，它可以窥探网络的结构。                                                                                                                              | 对于检测到的超时的 ICMP 报文或 UDP 报文，或者目的端口不可达报文，给予丢弃。             |
|    | 带时间戳选项的 IP 报文攻击 | 在 IP 路由技术中，提供了时间戳选项，记录 IP 报文从源到目的过程中所经过的路径和时间，也就是一个处理过此报文的路由器的列表。IP 时间戳选项通常用于网络路径的故障诊断，但也会被恶意攻击者利用，窥探网络结构。                                                                                                                  | 将检测进入的报文是否设置 IP 时间戳记录选项，如果设置了，则丢弃报文，并记录日志，如果没有设置，就正常转发。 |

## 7.11 IP 与 MAC 地址绑定

将 IP 地址与 MAC 地址进行绑定是防范 IP 地址冒用以及 ARP 欺骗等攻击的重要手段。此类攻击的特点就是攻击者的 IP 地址与 MAC 地址的对应关系不符合实际。本节将介绍 IP 与 MAC 地址绑定的基本概念、规格及原理。

### 7.11.1 介绍

#### 定义

将 IP 与 MAC 地址绑定是指在设备上建立 IP 地址与 MAC 地址的对应关系。设备在收到报文后，根据报文中的源 IP 地址查找关系表中相应的 MAC 地址，并且与帧头部包含

的实际源 MAC 地址进行比较。如果两者不一致，则认为该报文是非法报文，并将其丢弃。

目的

将 IP 与 MAC 地址绑定主要是为了防止 IP 地址冒用或者 ARP 欺骗等攻击行为的发生。

- IP 地址冒用是指攻击者使用自己的 MAC 地址，但是冒用他人的 IP 地址进行通信，以获取被攻击者的权限或者本应发送给被攻击者的报文。
- ARP 欺骗是指攻击者冒用他人的 MAC 地址发送 ARP 报文，同样可以获取到本应发送给被攻击者的报文；或者冒用网关的 MAC 地址，使网络内所有主机都将报文发给自己。

7.11.2 可获得性

License 支持

本特性无需 License 支持。

版本支持

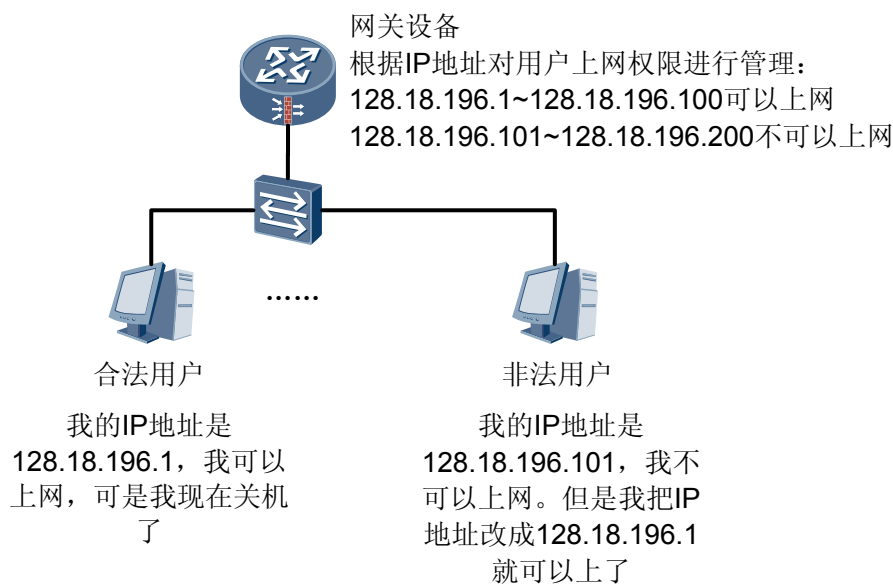
| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

7.11.3 原理描述

IP 地址冒用攻击原理

IP 地址冒用是指攻击者冒用网络中其他主机的 IP 地址，来绕过网关设备上一些基于 IP 地址实现的安全策略，例如访问控制，计费等。一个典型的攻击示例如图 7-31 所示。

图 7-31 IP 地址冒用攻击示例



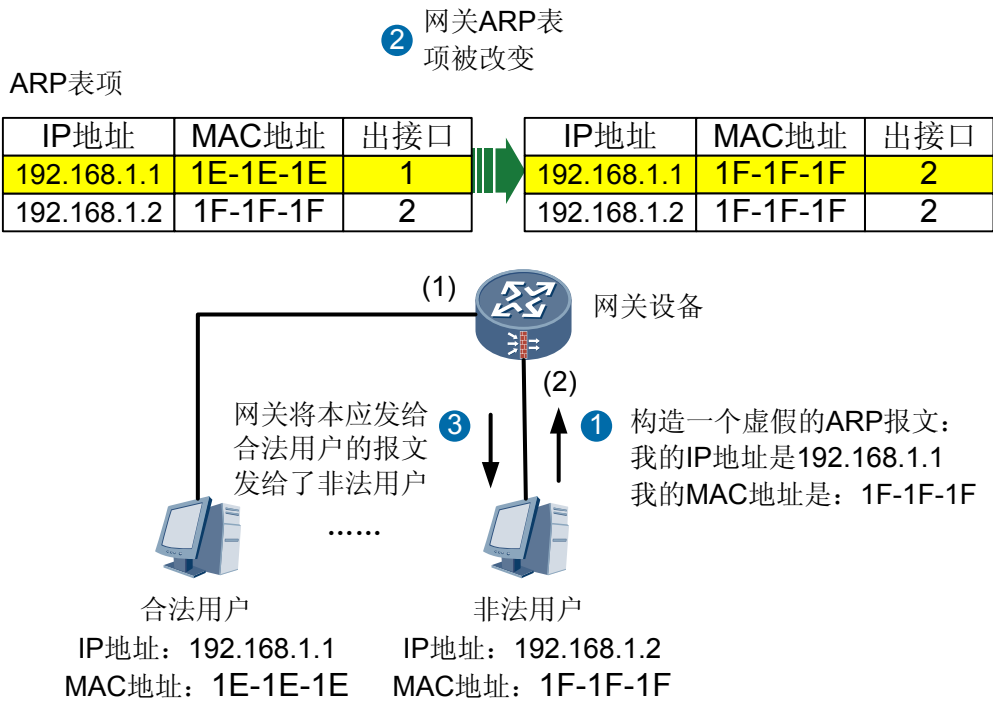
## ARP 欺骗攻击原理

ARP 欺骗是指攻击者通过构造虚假的 ARP 报文来改变网关设备或者是网络中其他主机的 ARP 表项，以获取到原本应该发给网关设备或者其他主机的报文。

由于在数据链路层转发中，交换机是根据 MAC 地址来判断出接口的，所以这种攻击的主要目的是改变交换设备上的 ARP 表项，以使交换设备在根据 IP 地址解析 MAC 地址的过程中，将被攻击者的 IP 解析为攻击者的 MAC 地址，从而将本应发给被攻击者的报文发送给攻击者。这样攻击者就可以从报文中获取到被攻击者的隐私信息。

一个典型的攻击示例如图 7-32 所示。

图 7-32 ARP 欺骗攻击示例



IP 与 MAC 地址绑定的实现原理与防范攻击的原理

实现 IP 与 MAC 地址绑定的前提是网络管理员应知道网络中正确的 IP 地址与 MAC 地址的对应关系，从而可以在设备上建立 IP 地址与 MAC 地址的对应关系。设备支持通过 ARP 探测自动收集 IP 地址与 MAC 地址的对应关系并进行绑定。

IP 和 MAC 地址绑定的实现原理是，设备上会建立一个地址绑定表记录 IP 与 MAC 地址的对应关系，管理员可以通过配置向其中添加表项。当设备收到一个 IP 报文时，会提取报文头中的源 IP 地址查找关系表中相应的 MAC 地址，并且与帧头部包含的实际源 MAC 地址进行比较。如果两者不一致，则认为该报文是非法报文，并将其丢弃。

通过 IP 与 MAC 地址绑定防范以上两种攻击的原理如图 7-33 和图 7-34 所示。

图 7-33 通过地址绑定防范 IP 地址冒用原理

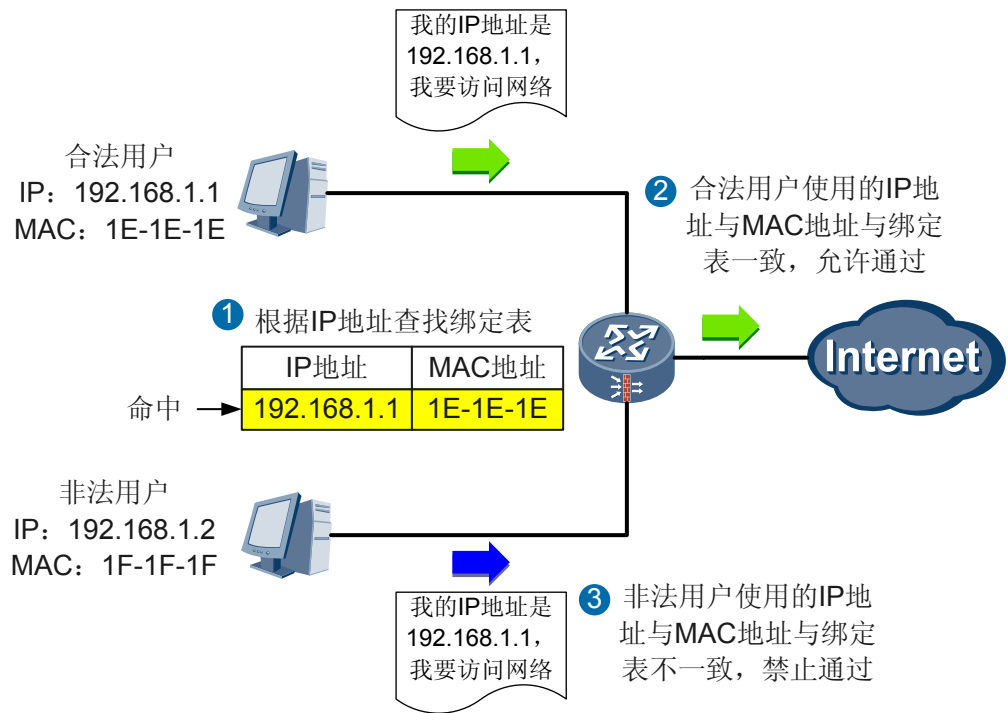
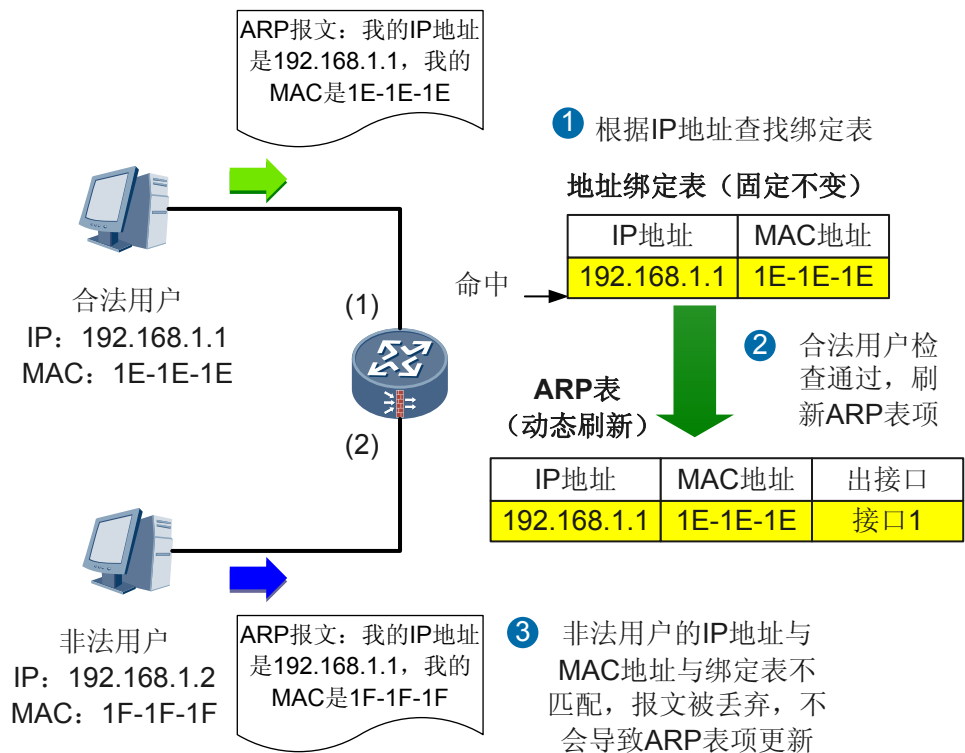


图 7-34 通过地址绑定防范 ARP 欺骗原理



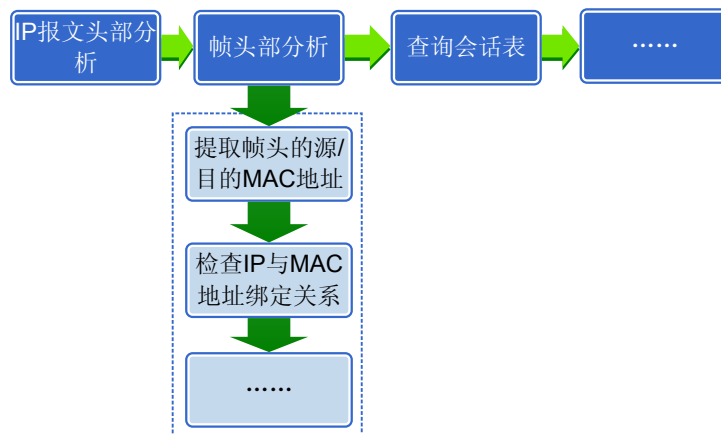
通过地址绑定表实现对这两种攻击的防范的主要缺点在于：



- 绑定关系需要管理员手工逐条添加，当网络中主机较多时比较费时费力。
- 攻击者有可能同时伪造 IP 地址和 MAC 地址。

IP 与 MAC 绑定关系的检查在转发流程中的位置如图 7-35 所示，在链路层帧的分析过程中就会进行 IP 与 MAC 绑定关系的检查，所以该项检查能够很早发现非法报文。

图 7-35 IP 与 MAC 绑定关系检查在转发流程中的位置



## 地址绑定表与 ARP 表、MAC 转发表的关系

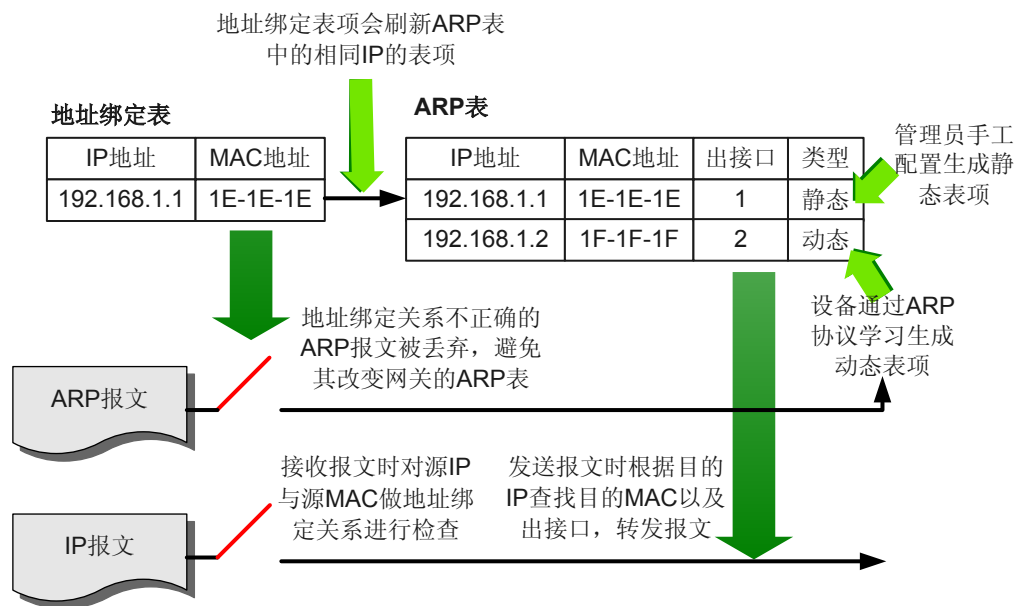
在设备上存在三种与 IP 地址、MAC 地址相关的表项：

- 地址绑定表：记录由管理员配置的 IP 地址与 MAC 地址的对应的关系，只能通过配置更改。  
设备在收到报文时根据该表核对报文的源 IP 地址与源 MAC 是否正常对应。
- ARP 表：记录通过 ARP 协议自动学习到的 IP 地址与 MAC 地址的对应关系。管理员也可以通过配置手工向表中添加静态表项。  
设备在发送报文时，要查找该表，以根据报文的目的 IP 地址确定目的 MAC 地址。
- MAC 转发表：记录网络中其他设备的 MAC 地址与出接口的对应关系。  
设备通过交换口直接转发链路层帧时，要查找该表，以根据报文的目的 MAC 地址确定应该将这个报文从设备的哪个接口发送出去。

从以上分析可以看出，地址绑定表主要用于接收报文时对报文的合法性进行核对，而 ARP 表和 MAC 转发表都是用于发送报文时，查找目的 MAC 地址和出接口。

地址绑定表与 ARP 表的关系如图 7-36 所示。

图 7-36 地址绑定表与 ARP 表的关系



## 7.12 黑名单

黑名单是对 IP 地址进行屏蔽的最简单也是最高效的手段。本节将介绍黑名单的基本概念、规格及原理。

### 7.12.1 介绍

### 定义

黑名单是指一系列 IP 地址的集合，设备丢弃来自黑名单中的 IP 地址的报文。

## 目的

黑名单是一种安全防护措施，最大优点就是可以以很高的匹配和处理效率，实现基于 IP 地址的最简单的报文控制。

### 7.12.2 可获得性

## License 支持

本特性无需 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 7.12.3 原理描述

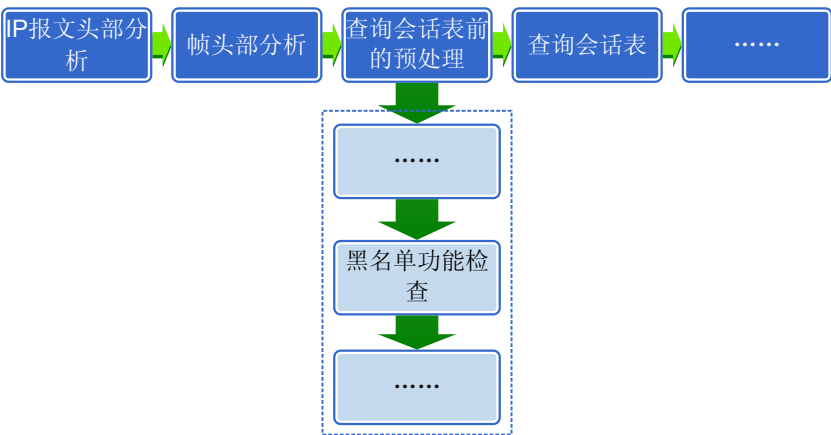
### 黑名单的实现原理

黑名单的实现原理就是：设备上建立一个黑名单表。对于接收到的报文，设备提取其 IP 报文头部中的源 IP 地址。如果该 IP 地址存在于黑名单中，且没有相应的放行规则，就将该报文予以丢弃。

由于黑名单的检测元素和处理机制都非常简单，所以黑名单的匹配和处理效率非常高。

黑名单检查在转发流程中的位置如图 7-37 所示。在查询会话表之前的预处理过程中就会进行黑名单的检查。

图 7-37 黑名单在转发流程中的位置



### 将 IP 地址加入黑名单的方法

对于如何将一个 IP 地址加入到黑名单中，以实现对该 IP 的屏蔽，设备提供了两种方法：

- 静态黑名单  
管理员可以通过命令行或 Web 方式手工逐个将 IP 地址添加到黑名单中。
- 动态黑名单  
设备上的部分安全机制可以自动识别危险对象，并将其加入到黑名单中。主要可以分为：

- 设备与 IDS 联动：IDS 可以识别危险对象，并将其加入到设备的黑名单中。
- 攻击防范功能：在针对 TCP 全连接攻击、IP 地址扫描攻击和端口扫描攻击这三种攻击的检测和防范中，设备会将检测出来的危险对象的 IP 地址加入到黑名单中。
- 网络用户在试图登录设备的命令行或 Web 管理界面时，连续多次输入错误用户名和密码，会被设备加入到黑名单，禁止其反复尝试登录。

在 IDS 联动的情况下，哪些 IP 地址会被加入到黑名单是由 IDS 设备的检测机制来决定的。在攻击防范和用户登录失败这两种情况下，将 IP 地址加入黑名单的原则是：

- 攻击防范

- TCP 全连接：管理员设定一个 TCP 连接建立后报文传输速率的最低值。低于该值的连接被认为是异常连接。当一个 IP 地址在一定时间内发起的异常连接数超过一定上限时，设备就将该 IP 地址加入到黑名单中。判断是否加入黑名单的时间以及异常连接数上限都可以由管理员进行配置。
- IP 地址扫描和端口扫描：管理员设定一个源 IP 连续对不同 IP 地址或端口发起连接的上限。超过该值的源 IP 被认为是攻击者，设备将该 IP 地址加入到黑名单中。

- 登录失败

当网络用户通过 Telnet、SSH、Web 方式登录设备的管理界面时，缺省情况下，连续三次用户名和密码不匹配，该网络用户所使用的 IP 地址就会被加入到黑名单中。10 分钟后，该 IP 地址才会从黑名单中删除。该用户才可以继续尝试登录设备。连续输错密码的次数可以由管理员配置。

当添加静态黑名单或动态黑名单时，相同的 IP 地址已经在黑名单中存在的情况下，设备将保留最新插入的那个黑名单表项。

## 将 IP 地址从黑名单中删除的方法

根据黑名单生成的机制不同，将 IP 地址从黑名单中删除的方法也不同：

- 静态黑名单

在管理员手工添加静态黑名单时，可以为每个黑名单表项配置老化时间。该表项在老化时间后会被自动删除。没有配置老化时间的表项不会被自动删除，只能由管理员使用命令行或 Web 界面手工将其删除。

- 动态黑名单

- 设备与 IDS 联动：由 IDS 设备决定何时将其从黑名单中删除。
- 攻击防范功能：在配置各种攻击防范功能时，可以配置黑名单表项的老化时间。一个 IP 在被自动加入到黑名单后，经过老化时间之后，会被自动从黑名单中删除。
- 登录失败：老化时间固定为 10 分钟。即某个 IP 在登录失败被加入到黑名单 10 分钟后，会被自动从黑名单中删除。

## 对黑名单过滤进行进一步控制

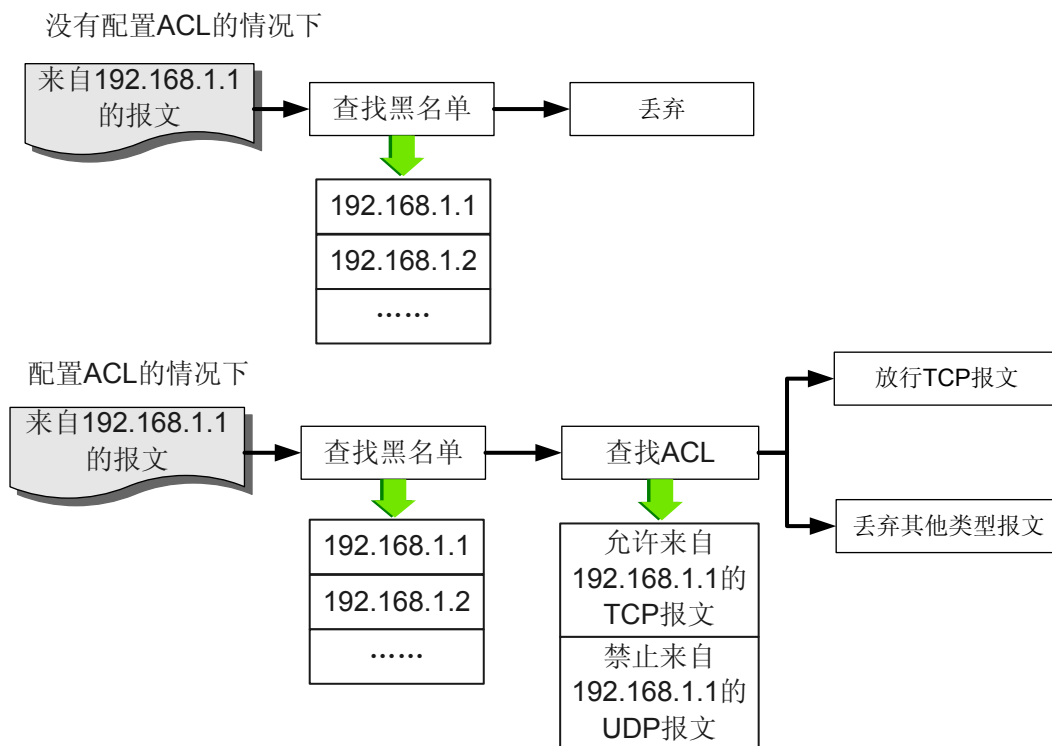
黑名单的基于 IP 地址的处理机制虽然很高效，但是也存在粒度过粗的问题。所以 USG 提供一个利用 ACL 进一步控制来自黑名单中 IP 地址的报文的机制。

实现的原理是在配置黑名单功能时可以引用一条 ACL。对于来自黑名单中 IP 地址的报文，

- 如果 ACL 中定义的动作为 permit，则不会丢弃该报文，而是继续转发
- 如果 ACL 中定义的动作为 deny，才会丢弃该报文
- 如果 ACL 中没有定义对该报文的动作，则丢弃该报文

如图 7-38 所示。

图 7-38 通过 ACL 对黑名单过滤做进一步控制



## 7.13 会话表与长连接

会话表是一个记录 TCP、UDP、ICMP 等协议连接状态，实现报文正常转发的重要表项。长连接是保持部分特殊会话超长存活的手段。本节将介绍会话表和长连接的基本概念、规格及原理。

### 7.13.1 介绍

#### 定义

会话表是一个记录系统中存在的 TCP、UDP、ICMP 等协议连接状态的表项。

长连接是指为部分特殊会话配置超长老化时间，使其长时间不从会话表中删除。

#### 目的

为了改进上一代包过滤防火墙的检测和转发效率低下的问题，目前主流防火墙都采用了基于“状态”的报文控制机制：只对首包或者少量报文进行检测就确定一条连接的状态。

态，大量报文直接根据所属连接的状态进行控制。这种状态检测机制迅速提高了防火墙产品的检测和转发效率。关于包过滤的详细原理请参见 [7.4 安全策略](#)。

而会话表正是为了记录连接的状态而存在的。设备在转发 TCP、UDP 和 ICMP 报文时都需要查询会话表，来判断该报文所属的连接以及相应的处理措施。

由于会话表对设备转发的关键影响，所以会话表项的管理也非常重要。由于大量会话的存在会对设备资源造成很大的消耗，当超过规格限制时还会导致新会话无法建立，业务不能正常运行。所以必须及时对无用的连接进行清理。

当一条会话在长时间没有被任何报文匹配，则说明该条会话所对应的连接可能已经关闭了，这条会话也就没有存在的必要了。所以 USG 为各种协议设定了会话老化机制。当一条会话在老化时间内没有被任何报文匹配，则会被从会话表中删除。

但是有一些特殊业务即使长时间没有报文传输也属于正常现象，例如数据库服务。这个时候如果其会话表项被删除，则该业务会中断。所以为了解决这些特殊业务的特殊问题，USG 设定了长连接机制。通过长连接机制可以给部分连接设定超长的老化时间。

7.13.2 可获得性

License 支持

本特性无需 License 支持。

版本支持

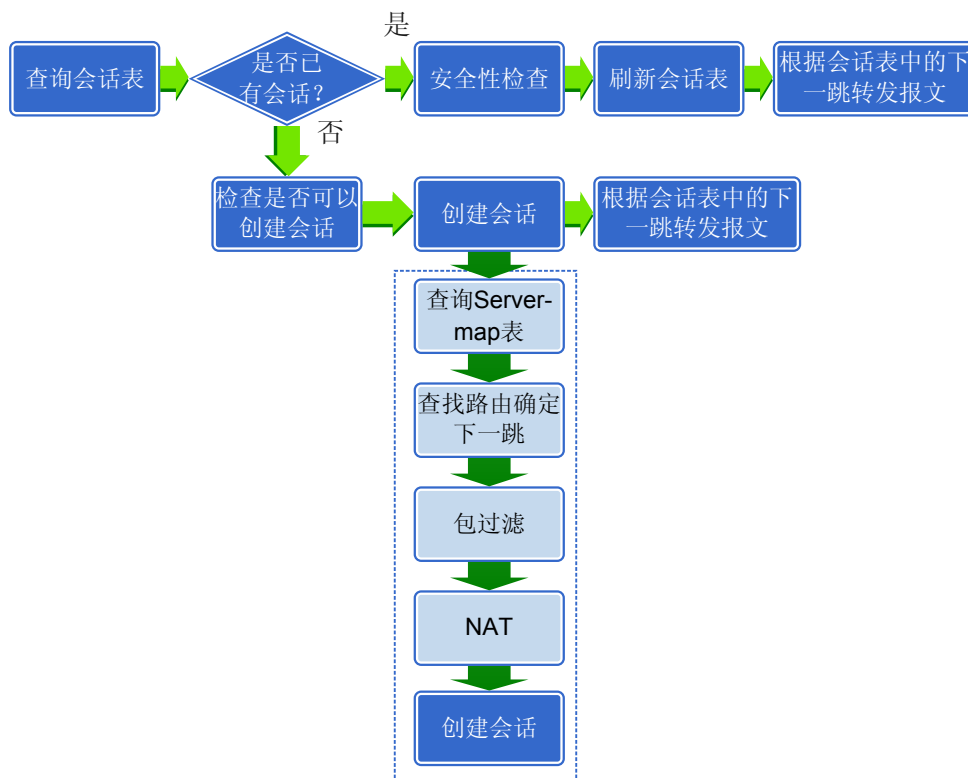
| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

7.13.3 原理描述

会话表在转发流程中的位置

会话表是控制报文转发的一个关键表项。除了同一 VLAN 内的链路层帧直接根据 MAC 转发表转发外，绝大部分 TCP、UDP、ICMP 报文都需要根据会话表来转发。在经过对报文的 IP 首部和帧首部的分析之后，设备就进行会话表的查询。查询和建立会话表的流程如[图 7-39](#)所示。

图 7-39 查询和建立会话表的流程



从图 7-39 中可以看出，对于已经存在会话表的报文的检测过程比没有会话表的报文要短很多。而通常情况下，通过对一条连接的首包进行检测并建立会话后，该条连接的绝大部分报文都不再需要重新检测。这就是状态检测防火墙的“状态检测机制”相对于包过滤防火墙的“逐包检测机制”的改进之处。这种改进使状态检测防火墙在检测和转发效率上有迅速提升。

在检查是否可以创建会话表的过程中，状态检测机制起到了关键的作用。会话表项的生成与协议首包、状态检测机制之间的关系可以总结为：

- 对于 TCP 报文  
开启状态检测机制时，首包（SYN 报文）建立会话表项。对除 SYN 报文外的其他报文，如果没有对应会话表项（设备没有收到 SYN 报文或者会话表项已老化），则予以丢弃，也不会建立会话表项。  
关闭状态检测机制时，任何格式的报文在没有对应会话表项的情况下，只要通过各项安全机制的检查，都可以为其建立会话表项。
- 对于 UDP 报文  
UDP 是基于无连接的通信，任何 UDP 格式的报文在没有对应会话表项的情况下，只要通过各项安全机制的检查，都可以为其建立会话表项。
- 对于 ICMP 报文  
开启状态检测机制时，没有对应会话的 ICMP 应答报文将被丢弃。  
关闭状态检测机制时，没有对应会话的应答报文以首包形式处理。

## 会话表匹配原则

会话表中记录的信息很多，但是一般是根据五个关键元素（源 IP 地址、源端口、目的 IP 地址、目的端口和协议号）来判断一个报文是否和一条会话表项匹配。这五个关键元素又被称为“五元组”。

### 说明

在使用 VPN 实例的情况下，会结合五元组和 VPN 实例号一共六个元素来判断报文是否和会话表项匹配。

由于一条连接往往是存在来回两个方向的报文，所以在匹配时源 IP 地址和目的 IP 地址、源端口和目的端口相互对换后的报文也可以认为与同一个会话表项匹配。

例如，主机 A 的 IP 地址是 192.168.1.1，使用 20000 端口向服务器 B 发起 TCP 连接。服务器 B 的 IP 地址是 1.1.1.1，提供服务的端口号是 30000。那么由该条连接首包建立的会话表表项将记录如下数据：

| 源 IP 地址     | 源端口   | 目的 IP 地址 | 目的端口  | 协议  |
|-------------|-------|----------|-------|-----|
| 192.168.1.1 | 20000 | 1.1.1.1  | 30000 | TCP |

服务器 B 在收到 A 的连接申请后，会向 A 发送一个应答报文，此时该报文的五元组应该是：

| 源 IP 地址 | 源端口   | 目的 IP 地址    | 目的端口  | 协议  |
|---------|-------|-------------|-------|-----|
| 1.1.1.1 | 30000 | 192.168.1.1 | 20000 | TCP |

这个报文的源 IP 地址和目的 IP 地址、源端口和目的端口和前一个发起连接的报文以及会话表项中的记录是相反的，但是实际上这个报文应该认为和前一个报文是属于同一条连接的，不需要重新检测。

所以在设备实现上，这样的报文也会被认为与会话表项匹配。

## 来回路径不一致对会话表的影响

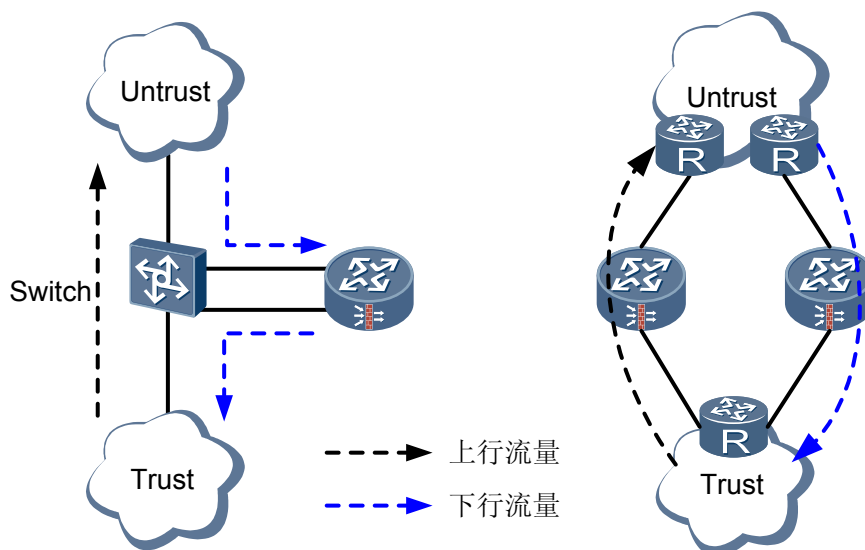
为了保证业务的正常运行，有必要解决合法的 TCP 连接由于“设备没有收到首包”而无法转发的问题。关闭状态检测机制可以解决这个问题。

在两个网络设备在通信时一定在进行报文的交互，即两者之间存在着来回两个方向的数据链路。报文来回路径一致，实际上是指一次通信过程中来回两个方向的数据都能经过同一个 USG 的处理。在这种情况下，USG 可以得到一次通信过程中所有的报文，才可以实现很多安全功能的检测。但是由于目前的 TCP/IP 网络主要进行的是逐包转发，事先并不协商实际的物理链路，所以有可能存在一次通信过程中，部分报文由一个设备转发，另外一部分报文由另一设备转发的情况。

当 USG 作为内网唯一出口的情况下，所有报文都必须流经设备才能被转发到其他网络中，设备没有收到首包这种情况很少出现。所以“设备没有收到首包”通常是在网络中存在多个通信通道的情况下出现，通常把这种网络环境称为“来回路径不一致的组网”，如图 7-40 所示。



图 7-40 两种来回路径不一致的组网



在这两种情况，都有可能出现设备没有收到首包这个问题。而这个问题完全是由网络规划导致的，并非流量本身存在不安全的因素。所以在这种情况下，为了使业务能够正常通过，唯一的解决方法是关闭设备的状态检测机制。

当前组网是否存在来回路径不一致的情况，主要与 USG 的部署位置有关，如果 USG 部署在内外网唯一的通路上，则报文来回路径应该是一致的，如果内外网在 USG 之外，还存在其它的通信链路，则很有可能出现报文来回路径不一致的情况。

## 老化时间对会话表的影响

为了保证业务的正常运行，有必要解决合法的 TCP 连接由于“会话表项已经老化”而无法转发的问题。长连接可以解决这个问题。

### 说明

状态检测机制关闭时，非首包也可以建立会话表，所以此时不需使用长连接功能也可保持业务的正常运行。

通常情况下，设备上缺省对各种协议设定的老化时间已经可以满足各种协议的转发需求了。在不同的网络环境下管理员也可以通过调整各种协议的老化时间来保障业务的正常运行。但是对于某些特殊业务中，一条会话的两个连续报文可能间隔时间很长。例如：

- 用户通过 FTP 下载大文件，需要间隔很长时间才会在控制通道继续发送控制报文。
- 用户需要查询数据库服务器上的数据，这些查询操作的时间间隔远大于 TCP 的会话老化时间。

如果只靠延长这些业务所属协议的老化时间来解决这个问题，会导致一些同样属于这个协议，但是其实并不需要这么长的老化时间的会话长时间不能得到老化。这会导致系统资源被大量占用，性能下降，甚至无法再为其他业务建立会话。

所以必须缩小延长老化时间的流量范围。长连接功能可以解决这一问题。在长连接功能中，可以通过 ACL 来指定流量，为这些特殊流量设定超长的老化时间。目前仅支持对 TCP 协议配置长连接。

## 7.14 Server-map 表与 ASPF

由于会话表对哪些报文属于同一条流量的标准过于严格，会导致一些特殊协议不能正确匹配会话表。Server-map 表可以解决这一问题。而 ASPF 功能通过对应用层协议的检测，最终也是生成 Server-map 表来实现特殊协议的转发。本节将介绍 Server-map 表项和 ASPF 的基本概念、规格及原理。

### 7.14.1 介绍

#### 定义

Server-map 表是一个通过少量关键元素来记录部分特殊服务连接状态的特殊表项。

ASPF（Application Specific Packet Filter）是指系统为了转发一些多通道协议报文，通过解析报文数据载荷，识别多通道协议自动协商出来的端口号，并自动生成相应的 Server-map 表项的功能。

#### 目的

在严格包过滤的情况下，设备通常只允许内网用户单方向主动访问外网。但是在使用 NAT 或 ASPF 功能的情况下，可能在外网用户通过随机端口主动访问内网服务器的情况。由于会话表的五元组限制过于严格，会导致这些特殊服务不能正常运行。引入 Server-map 表是为了解决这一问题。

ASPF 是为了解决多通道协议这种特殊服务的转发而引入的。这些协议会在通信过程中自动协商一些随机端口，在严格包过滤的情况下，这些随机端口发出的报文同样不能得到正常转发。通过 ASPF 功能可以对这些协议的应用层数据进行解析，识别这些协议协商出来的端口号，从而自动为其开放相应的访问规则，解决这些协议不能正常转发的问題。

### 7.14.2 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 7.14.3 原理描述

## Server-map 表的引入

通常情况下，如果在设备上配置严格包过滤，那么设备将只允许内网用户单方向主动访问外网。但在实际应用中，例如使用 FTP 协议的 port 方式传输文件时，既需要客户端主动向服务器端发起控制连接，又需要服务器端主动向客户端发起服务器数据连接，如果设备上配置的包过滤为允许单方向上报文主动通过，则 FTP 文件传输不能成功。

为了解决这一类问题，USG 设备引入了 Server-map 表，Server-map 用于存放一种映射关系，这种映射关系可以是控制数据协商出来的数据连接关系，也可以是配置 NAT 中的地址映射关系，使得外部网络能透过设备主动访问内部网络。

生成 Server-map 表之后，如果一个数据连接匹配了 Server-map 表项，那么就能够被设备正常转发，而不需要去查会话表，这样就保证了某些特殊应用的正常转发。

USG 设备上生成 Server-map 表项目目前总共有四种情况：

- 配置 ASPF 后，转发 FTP、RTSP 等多通道协议时生成的 Server-map 表项。
- 配置 ASPF 后，转发 QQ/MSN、TFTP 等 STUN 类型协议时生成的三元组 Server-map 表项。
- 配置 NAT Server 或 SLB 时生成的静态 Server-map。
- 配置 NAT No-PAT 时生成的动态 Server-map。

下面分别对这四种 Server-map 表进行介绍。

## 转发多通道协议时生成的 Server-map 表项

转发 FTP、RTSP 等多通道协议的数据会生成 Server-map 表项。

多通道协议会由客户端和服务端之间的控制通道动态协商出数据通道，即通信双方的端口号是不固定的。而在配置 ASPF 功能后，设备检测到控制通道的协商，根据关键报文载荷中的地址信息动态创建 server-map 表项，用于数据通道发起连接时进行查找。这个 server-map 表项包含了多通道协议报文中协商的数据通道的信息。

例如在 FTP 的 port 模式中，在 FTP 客户端随机选择一个数据通道端口号（本例中的 2165），并通过控制通道将该端口号发送给 FTP 服务器后，FTP 服务器会直接向该端口发起连接。如果此时配置了 ASPF 功能，会生成如下 Server-map 表项：

```
ASPF: 40.0.0.5 -> 40.0.0.10:2165 , Zone: ---  
Protocol: tcp(Appro: ftp-data), Left-Time: 00:00:05, Addr-Pool: ---  
Vpn: public -> public
```

其中，tcp(Appro: ftp-data)表示该条 Server-map 表项用于 FTP 协议的数据通道的转发。40.0.0.5 为源 IP 地址，即发起连接的 FTP 服务器的 IP 地址，40.0.0.10 为目的 IP 地址，即 FTP 客户端的 IP 地址。

在采用严格包过滤情况下，由于事先没有配置对 2165 端口允许的包过滤策略，所以如果没有 Server-map 表项，会导致该连接被阻断。ASPF 自动生成 Server-map 表项后，由于只需要匹配四元组信息，所以就可以实现正常的数据传输了。

## 转发 STUN 类型协议时生成的三元组 Server-map 表项

转发 QQ/MSN 等 STUN（Simple Traversal of UDP over NATs，NAT 的 UDP 简单穿越）类型会生成的三元组 Server-map 表项。

QQ/MSN 等协议中，当用户登录之后，用户的 IP 地址和端口就固定下来了，可是会向该用户发起对话的另一方的 IP 地址和端口号是不固定的。通过配置 STUN 类型的 ASPF，当 QQ 或者 MSN 等用户连接服务器时，设备会记录下用户的 IP 地址和端口信

息，并动态生成 STUN 类型的 Server-map。这个 server-map 表项中仅包含三元组信息，即通信一方的 IP 地址，端口号和协议号。这样其他用户可以直接通过该 IP 和端口与该用户进行通信。只要有相关的流量存在，STUN 动态的 server-map 就将一直存在。在所有相关流量结束后，server-map 表开始老化。

例如：

```
STUN : ANY -> 40.0.0.10:4967, Zone: ---  
Protocol: udp(Appro: qq-derived), Left-Time: 00:00:05, Addr-Pool: ---  
Vpn: public -> public
```

在本例中，由于源地址为 ANY，即不限制，这样任意用户都可以主动向 40.0.0.10 的 4967 端口发起连接，以实现未知用户向内网的 QQ/MSN 用户发起对话的情况。

## 配置 NAT Server 或 SLB 时生成的静态 Server-map

在使用 NAT Server 功能时，外网的用户向内部服务器主动发起访问请求，该用户的 IP 地址和端口号都是不确定的，唯一可以确定的是内部服务器的 IP 地址和所提供服务的端口号。

所以在配置 NAT Server 成功后，设备会自动生成 Server-map 表项，用于存放 Globble 地址与 Inside 地址的映射关系。设备根据这种映射关系对报文的地址进行转换，并转发。

每个生效的 NAT Server 都会生成正反方向两个静态的 Server-map。用户删除 NAT Server 时，Server-map 也同步被删除。

NAT Server 生成的 Server-map 表项实例如下所示。

```
Nat Server, ANY -> 202.169.10.2:21[10.1.1.2:21], Zone:---  
Protocol: tcp(Appro: ftp), Left-Time:---, AddrPool: ---  
Vpn: public -> public  
  
Nat Server Reverse, 10.1.1.2[202.169.10.2] -> ANY, Zone:---  
Protocol: ANY(Appro: ---), Left-Time:---, AddrPool: ---  
Vpn: public -> public
```

- **Nat Server** 表示正向（即外网客户端主动访问内网服务器方向）的 Server-map 表项。
- **Nat Server Reverse** 表示反向（即内网服务器主动访问外网客户端方向）Server-map 表项。
- **Protocol** 表示配置 Nat Server 时指定的协议类型。
- **202.169.10.2** 表示 Nat Server 的 Globble 地址，即对外公布的公网 IP 地址。
- **10.1.1.2** 表示 Nat Server 的 Inside 地址，即转换前的私网 IP 地址。

在 SLB 功能中，由于需要将内网的多个服务器以同一个 IP 地址对外发布，所以也会建立与 NAT Server 类似的 Server-map 表项，只不过根据内网服务器的个数 N，需要建立 1 个正向表项和 N 个反向表项。

## 配置 NAT No-PAT 时生成的动态 Server-map

在使用 NAT 功能时，如果配置了 No-PAT 参数，那么设备会对内网 IP 和公网 IP 进行一对一的映射，而不进行端口转换。此时，内网 IP 的所有端口号都可以被映射为公网地址的对应端口，外网用户也就可以向内网用户的任意端口主动发起连接。

所以配置 NAT No-PAT 后，设备会为有实际流量的数据流建立 Server-map 表，用于存放私网 IP 地址与公网 IP 地址的映射关系。设备根据这种映射关系对报文的地址进行转换，然后进行转发。

NAT No-PAT 生成的 Server-map 表项实例如下所示。

No-Pat: 10.1.1.2[202.169.10.2] -> ANY, Zone: untrust  
Protocol: ANY (Appro: ---), Left-Time: 00:00:03, Addr-Pool: 61  
Vpn: public -> public

No-Pat Reverse, ANY -> 202.169.10.2[10.1.1.2], Zone: untrust  
Protocol: ANY (Appro: ---), Left-Time: 00:00:03, Addr-Pool: ---  
Vpn: public -> public

- **No-Pat** 表示正向 Server-map，即指从私网 IP 发起的连接。
- **No-Pat Reverse** 表示反向 Server-map，即指外网设备通过分配的公网 IP 来访问私网 IP 对应设备所发起的连接。
- **10.1.1.2** 表示映射前的私网 IP 地址。
- **202.169.10.2** 表示映射后的公网 IP 地址。

Server-map 表在转发流程中的位置

Server-map 表在转发流程中的位置请参考图 7-39。

当一个没有会话表项的报文在通过系统的安全性检查之后，系统准备为其查找路由之前，系统会查询 Server-map 表。此时查找 Server-map 表主要是为了两个目的：

- 对该报文所属连接的状态再次检查和确认  
由于该报文没有会话表项，所以它有可能是属于需要建立 Server-map 的特殊业务的报文。通过查询 Server-map 可以确认这个报文是否属于一个已经存在的连接。如果这个报文既没有会话表项，又不属于一个已经存在的特殊业务，那么在严格包过滤情况下，这个报文就会被丢弃。
- 根据 Server-map 表对报文中一些元素进行转换  
Server-map 表项除了会记录报文的关键元素，还会记录系统对该报文的一些处理措施。对于一个和 Serve-map 表项匹配的报文，系统在查询 Server-map 表后，会根据表项中的记录对报文进行转换。例如在 NAT Server 应用中，系统会将一个访问内部服务器的报文的地址从公网地址转换为私网地址。

由于 Server-map 中并不会记录报文的下一跳等信息，所以 Server-map 通常只是用检查第一个报文。在这个报文通过检查之后，系统会生成一条会话表用于后续报文的转发。即 Server-map 只是用于新通道的建立，通道建立后的报文还是根据会话表来转发。

以 FTP 的 Port 模式为例。主机 A 的 IP 地址是 192.168.1.1，使用 20000 端口向服务器 B 发起 FTP 连接。服务器 B 的 IP 地址是 1.1.1.1，提供服务的端口号是 21。系统已经通过配置包过滤允许了 192.168.1.1 向 1.1.1.1 的 21 号端口发起连接。

1. 在发起连接的报文通过后，系统中创建了一条会话表项，通信双方的控制通道也就建立起来了。

| 类型  | 源 IP 地址     | 源端口   | 目的 IP 地址 | 目的端口 | 协议  |
|-----|-------------|-------|----------|------|-----|
| 会话表 | 192.168.1.1 | 20000 | 1.1.1.1  | 21   | FTP |

2. 经过控制通道的协商，A 随机选择一个数据通道端口号 2165，并通过控制通道将该端口号发送给 FTP 服务器。系统在获取到这个报文后为其建立 Server-map 表项。

| 类型           | 源 IP 地址 | 目的 IP 地址    | 目的端口 | 协议       |
|--------------|---------|-------------|------|----------|
| Server-map 表 | 1.1.1.1 | 192.168.1.1 | 2165 | FTP-Data |

3. B 在收到 A 的协商报文之后，使用 20 端口主动向 192.168.1.1 的 2165 端口发起连接。这个连接在会话表中没有对应的表项，如果域间缺省包过滤策略为禁止，则可能会被丢弃。所以此时，系统就会查询 Server-map，以确定该报文是否属于特殊业务。

由于这个报文的源 IP 地址、目的 IP 地址、目的端口和协议都与 Server-map 表项符合，系统允许这个报文通过。而且此时这个新的数据通道的五元组都已经确定，所以系统同时创建一条会话表，通信双方的数据通道也就建立起来了。

| 类型  | 源 IP 地址 | 源端口 | 目的 IP 地址    | 目的端口 | 协议       |
|-----|---------|-----|-------------|------|----------|
| 会话表 | 1.1.1.1 | 20  | 192.168.1.1 | 2165 | FTP-Data |

4. 后续数据通道上传输的数据报文都可以匹配这条会话表项并得到转发，不再需要重新匹配 Server-map 表项。Server-map 表项由于一直没有报文匹配，经过老化时间后就会被删除。这种机制保证了 Server-map 表项这种较为宽松的检测条件能够及时被删除，保证了网络的安全性。

## ASPF 的实现以及 ASPF 与 Server-map 表的关系

在上述举例中，A 随机选择一个数据通道端口号 2165，并通过控制通道将该端口号发送给 FTP 服务器。“2165”这个端口信息是包含在协商报文的应用层载荷中的。系统在收到这个报文之后，能够将其解析出来，正是因为开启了 ASPF 功能。

ASPF 的主要作用就是对报文的应用层数据进行解析，识别出包含在 IP 报文数据载荷中的各种协商信息，以供设备自动为其创建相应的访问规则。

由于各种协议的报文结构和协商方式各不相同，所以 USG 目前只能支持特定种类的协议的识别。

ASPF 识别完成之后，系统通常会生成一条 Server-map 表项来记录识别结果。之所以不直接生成会话表项，是因为识别完成时，通常通信双方都只是在协商阶段，新通道上还没有实际的数据传输，或者新通道的五元组信息还没有完全确定，所以还不能建立会话表。

通过 ASPF 识别，建立 Server-map 表项后，当系统收到新通道上的第一个报文，就会生成相应的会话表项。后续该通道上的报文就都可以根据会话表进行转发了。

## 7.15 端口映射

端口映射又称为“端口识别”，是为了系统能够将一些发给不知名端口的报文识别为知名协议，可结合 ASPF 使用。本节将介绍端口映射的基本概念、规格及原理。

### 7.15.1 介绍

#### 定义

端口映射是指将发往非知名端口的报文识别为知名协议的功能。

## 目的

对于一些知名协议，由于其有特殊的报文结构和协商机制，所以系统会对其进行特殊处理。但是系统通常是通过端口号来对协议进行识别的。所以如果一种知名协议采用了非知名端口来提供服务，系统就不能对其进行相应的处理，这些业务也就不能正常运行。

端口映射功能可以解决这一问题。通过管理员配置，将系统中建立一张端口号与协议类型的映射表，使得系统将发往这些端口的报文也可以当做相应的协议报文来进行处理。

通过端口映射，使非知名端口可以提供知名服务，这样可以避免一些专门针对知名端口发起的扫描、攻击行为，起到安全防范的作用。

## 7.15.2 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 7.15.3 原理描述

应用层协议一般使用知名端口号进行通信，例如 FTP 使用 20 和 21 端口，Web 服务通常使用 80 端口。但是在某些情况下，内网服务器需要通过其他自定义端口对外提供知名服务时，可以通过端口映射功能来完成相应的转换。

端口映射的原理是，创建一个协议与映射端口的关联表，当收到发往映射端口的报文时，会将其识别为对应的知名服务的报文，并进行相应的应用层识别等功能。

在创建协议与端口的映射关系时，支持以下两种多重映射关系：

- 一个协议可以映射为多个端口。  
例如，可以通过配置，将发往 20001、20002、20003 的报文都识别为 FTP 协议报文。
- 一个端口可以映射为多个协议  
例如，可以通过配置，将主机 A 发往 20001 的报文识别为 FTP 协议报文，将主机 B 发往 20001 的报文识别为 HTTP 协议报文。  
需要注意的是，在进行这种配置时必须通过 ACL 将流量区分开来。即通过 ACL 定义哪些流量识别为 FTP 协议，哪些流量识别为 HTTP 协议。

用户自定义的映射关系并不会替代系统预定义的映射关系，而是两者同时有效。区别是系统预定义映射对所有流量有效，用户自定义映射只对 ACL 所匹配的流量有效。系统预定义的映射包括：

- FTP: 21
- SMTP: 25
- HTTP: 80
- RTSP: 554
- H.323: 1719
- SQL.NET: 1521

## 7.16 分片报文处理

网络设备在转发报文时，如果报文长度超过了链路 MTU，通常会对报文进行分片。本节介绍设备对分片报文的处理机制。

### 7.16.1 介绍

#### 定义

一条链路上所能传输的最大帧长度被称为 MTU（Maximum Transmission Unit，最大传输单元）。报文分片是指网络设备为了满足通信过程中报文所经过的所有网络的 MTU，而将一个较长报文分割为多个较短报文的处理机制。被分割之后的报文被称为分片报文。

#### 目的

各种网络设备对分片报文的处理机制都有不同。而分片报文由于其特殊性，也经常被用来做攻击报文。所以在 USG 中引入三种处理机制是为了满足不同用户对网络连通性和安全性的不同要求。

### 7.16.2 参考标准和协议

本特性的参考资料清单如下：

- RFC0791: Internet Protocol
- RFC1191: Path MTU discovery

### 7.16.3 可获得性

#### License 支持

本特性无需 License 支持。



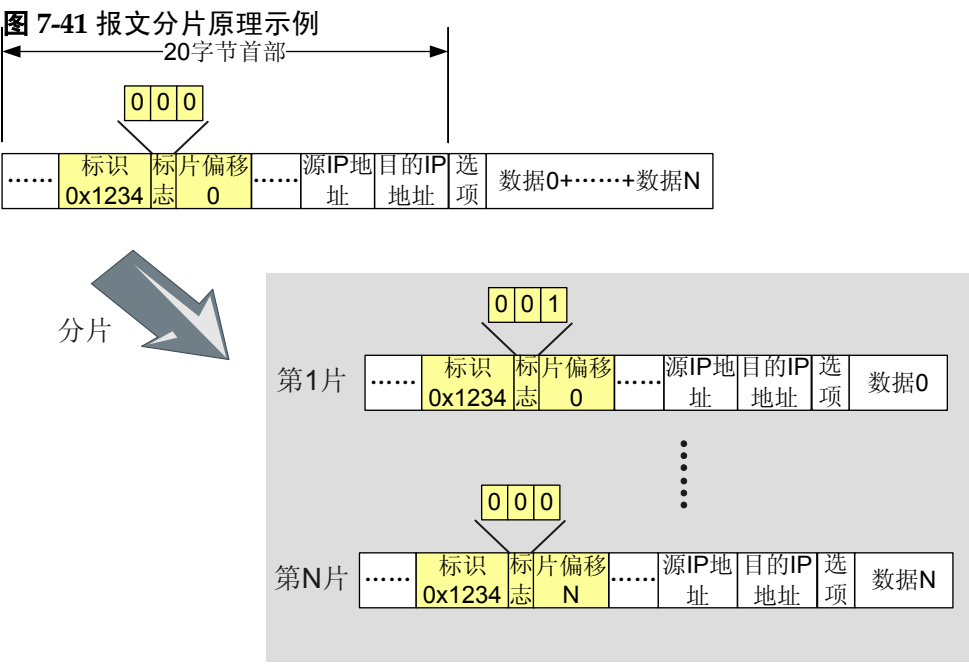
版本支持

| 产品                                                                      | 最低支持版本                                                                                                      |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br><br>USG5500 V200R001C00 |

7.16.4 原理描述

报文的分片与重组

一个普通的 IP 报文由于长度超过 MTU 值而被分片的例子如图 7-41 所示。



报文分片后，通过 IP 报文首部中的标识字段识别哪些分片是属于同一个报文的，通过片偏移字段记录该分片在整个报文中的位置，标志字段用最后一个比特来表示“更多分片”。除了最后一个分片外，其他每个组成数据报的分片都要把该比特置 1。

一个网络设备在收到所有分片报文之后，就可以根据标识字段和片偏移字段来将所有分片重组为所有原来的完整报文。

由于分片报文的特殊性，分片报文经常会被利用来做网络攻击。同时分片报文的传输也经常因为某个分片的丢失而导致整个报文重传。所以在实际应用中，应尽量避免分片报文的出现。RFC1191 描述了一种“路径 MTU 发现”机制，该机制用于确定两台主机间的路径 MTU，这样就可以避免 IP 分片。

## USG 对分片报文的处理

在实际传输过程中，由于网络环境的复杂，可能存在以下情况：

- 后续分片报文先于首片报文到达某个网络设备
- 后续报文需要在某个中间设备上重组后才能继续传输，例如中间设备需要解析报文载荷后才能判断如何转发。

在 USG 上，这两种情况都可能存在。由于 USG 提供的安全机制对报文检测非常严格，为了满足攻击防范和 NAT 转换等特殊功能的需要，设备上对分片报文的处理分为三种机制：

- 分片缓存  
设备缺省机制。设备会将非首片的分片报文缓存至分片散列表，等待首片到来建立会话后，将所有分片报文进行转发。在分片散列表已经放满，但是首片还未到来的情况下，后续分片报文会被丢弃，直至收到首片报文。如果分片散列表的老化时间已到，还未收到首片报文，所有缓存的分片报文将被全部丢弃。通过 **firewall session aging-time fragment** 命令可以修改分片散列表的老化时间。
- 分片丢弃  
不信任所有分片报文，收到分片报文即丢弃。可以提供最大的安全性，但是可能会影响正常业务的转发。
- 分片直接转发  
信任所有分片报文，收到分片立即直接转发，不会缓存到分片散列表等待首包的到来。在同一报文的分片可能通过不同设备的情况下，该机制将保证所有分片都能正常转发，以被最终接收者接收。但是存在一定的安全风险。

## USG 的分片缓存机制

在 USG 的分片缓存机制的缓存机制中，涉及到几个数值：

- 分片散列表最大表项数
- 每个报文最多可缓存分片数
- 最多可缓存分片报文数

下面介绍一下这几个数值之间的关系。

- 分片缓存机制中，设备每收到一个非首片的分片报文，就会创建一条散列表项。以后凡是属于同一原始报文的分片报文，都会按照该表项缓存。所以分片散列表最大表项数表示最多可缓存的原始报文数。
- 为了保证缓存资源的有效利用，设备对一个原始报文只能缓存有限数量的分片。当缓存分片报文数量超过这个上限，后续收到属于同一个原始报文的非首片分片报文，都会将其丢弃。直至收到首片，将所有报文进行转发，或者表项被老化删除，将所有报文丢弃。
- 可以看出，“分片散列表最大表项数×每个报文最多可缓存分片数”大于最多可缓存分片报文数。这是因为通常缓存分片数还未达到“每个报文最多可缓存分片数”，设备就已收到首片，所有报文被正常转发。另外缓存分片对设备资源和性能有较大影响，所以设备不能保证每个报文的分片都可以存至最大值。当设备缓存的分片数超过最大值，所有后续收到的非首片分片报文都将被丢弃。

## 对于需要设备重组的分片报文的处理

对于发送给设备本身的分片报文，不受上述规格的影响。设备会缓存所有分片以进行处理。

对于不是发给设备本身，但是需要设备重组成原始报文之后才能转发或处理的分片报文，必须将设备配置到分片缓存状态才能保证业务的正常运行。

例如在 VPN 应用中(主要指 IPSEC 和 GRE)，由于需要设备对分片报文进行重组后解密或者解封装，设备才能进行后续处理，所以必须将设备配置成分片缓存状态，完成原始报文重组之后，才可以进行相应的加密解密处理。在 NAT 应用中，需要设备对分片报文进行重组后才能正常解析和转换报文中的 IP 地址，所以也必须将设备配置成分片缓存状态，才可以正常进行 NAT。

这些需要设备重组的分片报文的缓存个数也是按照上述规格来进行限制。所以如果设备收到分片过多，会影响这些业务的运行。

## 7.17 GTP

GTP 是用于 GPRS 网络中 GSN 节点之间的接口协议，通过 GTP 协议实现 GPRS 网络中分组数据的传输。

### 7.17.1 介绍

#### 定义

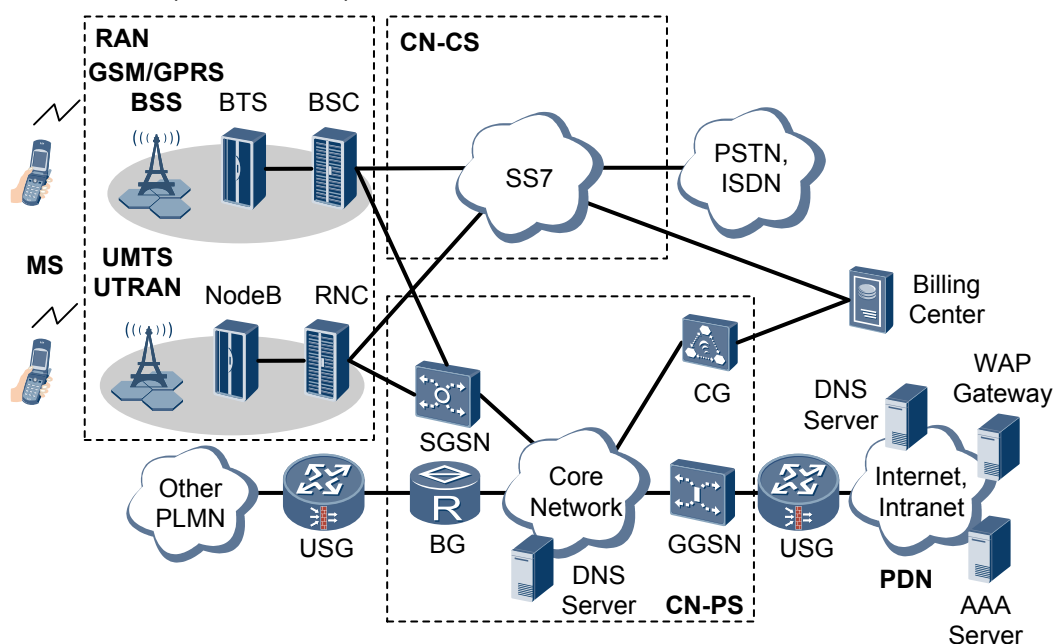
随着 3GPP（the 3rd Generation Partnership Project）协议的演进，移动通信组网及其提供的业务也在不断发展变化：

- 移动通信网络  
从 2G GSM（Global System for Mobile Communications）经过 2.5G GPRS（General Packet Radio Service）演进到 3G UMTS（Universal Mobile Telecommunications System），移动通信网络实现了广域覆盖、高速无线数据传输和与因特网的融合。
- 提供的业务  
提供语音、数据、视频等丰富多彩的业务，极大满足了用户随时随地多种方式相互通信的需求。

其中，GPRS 是为了支持分组业务而在 GSM 网络的基础上发展而来的 2.5G 网络；TD-SCDMA（Time Division-Synchronous Code Division Multiple Access）/WCDMA（Wideband Code Division Multiple Access）是 3G 网络，支持 CS（Circuit Switching）和 PS（Packet Switching）业务。TD-SCDMA/WCDMA 分组域与 GPRS 网络的结构相同。

GPRS/TD-SCDMA/WCDMA 网络结构如图 7-42 所示，其中核心网 CS 域中有一系列的设备，在图 7-42 中没有体现。

图 7-42 GPRS/TD-SCDMA/WCDMA 网络结构



MS: Mobile Station, 移动台

CN-CS: Core Network-Circuit Switching, 核心网电路交换域

BSS: Base Station System, 基站子系统

BTS: Base Transceiver Station, 基站收发信台

NodeB: UMTS 基站

SS7: CCITT Signaling System No.7, CCITT 7 号信令系统

SGSN: Serving GPRS Support Node, 服务 GPRS 支持节点

BG: Border Gateway, 边缘网关

DNS Server: 域名服务器

WAP Gateway: Wireless Access Protocol Gateway, 无线接入协议网关

PDN: Public Data Network, 公众数据网

RAN: Radio Access Network, 无线接入网络

CN-PS: Core Network-Packet Switching, 核心网分组交换域

UTRAN: Universal Terrestrial Radio Access Network, UMTS 陆地无线接入网

BSC: Base Station Controller, 基站控制器

RNC: Radio Network Controller, 无线网络控制器

CG: Charging Gateway, 计费网关

GGSN: Gateway GPRS Support Node, 网关 GPRS 支持节点

Billing Center: 计费中心

AAA Server: 认证、授权和计费服务器

FW: Firewall, 防火墙

PLMN: Public Land Mobile Network, 公用陆地移动通信网

在 GPRS 网络中, 数据从手机到达 Internet, 需要经过四个设备: MS、BSS、SGSN 和 GGSN。这四个设备的功能分别描述如下:

- MS

MS 是移动用户设备, 可以通过空中接口发起、接收呼叫。当进行数据业务时, MS 和核心网分组域建立逻辑链路。

- BSS

包括一系列设备, 负责分配空中的信道资源, 并在 MS 和 SGSN 之间转发信息。

- SGSN

SGSN 是为提供分组数据业务功能而引入的一个新的网元设备, 主要的作用是为本 SGSN 服务区域的 MS 转发输入/输出的 IP 分组。

SGSN 主要提供用户分组数据包的路由与转发、加密与鉴权、会话管理、移动性管理、逻辑链路管理和计费信息产生和输出等功能。

- GGSN

GGSN 也是为提供分组数据业务功能而引入的一个新的网元设备，提供数据包在 GPRS/TD-SCDMA/WCDMA 网和外部数据网之间的路由和封装。

GGSN 是 MS 接入外部分组网络的网关，从外部网络来看，GGSN 类似是可寻址 GPRS/TD-SCDMA/WCDMA 网络中所有用户 IP 地址的路由器。GGSN 接收 MS 发送的数据，选路到相应的外部网络；或者接收外部网络的数据，根据其目的地址选择 GPRS/TD-SCDMA/WCDMA 网中的传输通道，发送给相应的 SGSN。

GPRS 网络是在现有的 GSM 网络中，通过增加一些网元（如 SGSN、GGSN），改造而成的承载网。这些网元统称为 GSN（GPRS Supporting Node）。

为了保证新增的网元之间以及这些网元与其它网络之间的通信，增加了统称为 G 接口的新接口，这些接口分别是 Gb、Gn、Gp、Gi 等数据接口，Gr、Gs、Gd、Gf 等信令接口。在 USG 的应用场景中，需要重点关注 Gn、Gp、Gi 三个接口。

- Gn

同一个 PLMN 中的不同 GSN 之间的接口。Gn 接口运行 GTP（GPRS Tunneling Protocol）协议，确保在同一个 PLMN 中的 SGSN 和 GGSN 之间的互通。

- Gp

不同 PLMN 之间的 GSN 之间的接口，用来实现不同 PLMN 之间的数据漫游业务。Gp 接口运行 GTP 协议，确保在不同的 PLMN 之间的 SGSN 和 GGSN 之间的互通。

- Gi

GGSN 和 PDN 之间的接口。该接口实现 GPRS 网络和外部数据网的互连。Gi 接口运行 IP 协议，保证 GGSN 与外部网络的数据传输。

GTP 是为 Gn 和 Gp 接口定义的一个隧道协议，用来支持 GSN 之间的数据通信。GTP 基于 UDP 协议，包括 GTP 控制平面（GTP-C）和 GTP 用户平面（GTP-U）。

- GTP 控制平面（GTP-C）

在控制平面，使用信令机制创建、修改和删除隧道。

- GTP 用户平面（GTP-U）

在用户平面，使用隧道机制传送用户数据包。

GTP 包括 Version 0 和 Version 1 两个版本。前者属于 3GPP Release 98 协议，用于 GPRS 网络；后者属于 3GPP Release 99 协议，用于 3G 网络。GTP Version 1 可兼容 Version 0，两者可以通过 GTP 报文头的版本字段取值来区分。

除此之外，GTP 还包括用于计费的 GTP' 协议。

## 目的

GTP 控制平面和用户平面使用了多种消息进行报文交互，例如：

- 路径管理消息
- 隧道管理消息
- 位置管理消息
- 移动性管理消息
- 信息元素 IE（Information Element）等

由于 GPRS 网络的数据业务都是承载在 GTP 隧道中，因此 USG 的主要应用是对 GTP 协议进行解析，根据预先设置的规则，对 GTP 报文进行过滤，确保 GPRS 网络的安全性。

## 收益

USG 在核心网 PS 域中主要有以下几种的应用：

- 工作在 Gn 接口，过滤掉恶意报文，保证同一个 PLMN 内网元的安全。
- 工作在 Gp 接口，当一个 PLMN 网络与其他 PLMN 网络相连时，过滤掉来自其他 PLMN 网络的恶意报文，保证 PLMN 网络内网元的安全。
- 工作在 Gi 接口，当 PLMN 网络与外部 IP 网络相连时，过滤掉来自外部 IP 网络的恶意报文，保证 PLMN 网络内网元的安全。
- 对 GTP 计费溢出攻击进行防范。

## 7.17.2 参考标准和协议

与 GTP 特性相关的参考标准与协议如下：

- 3GPP TS 23.060: General Packet Radio Service (GPRS) Service description Stage 2
- 3GPP TS 29.060: GPRS Tunnelling Protocol (GTP) across the Gn and Gp interface
- 3GPP TS 29.061: Interworking between the Public Land Mobile Network (PLMN) supporting packet based services and Packet Data Networks (PDN)

## 7.17.3 可获得性

### License 支持

USG5500 的 GTP 功能需要 License 支持。USG2110-X/2100/2200/5100、USG2100/2200/5100 BSR/HSR 的 GTP 功能不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 7.17.4 原理描述

### 7.17.4.1 GTP 策略

#### 内容过滤

GTP 控制层面的消息可以包含多个信息元素（Information Element），例如：

- APN (Access Point Name)  
接入点名称，在 GGSN 中用于标识一个指定的外部网络和一种服务的 ISP，在 SGSN 中可根据 APN 通过 DNS 域名解析得到与此 APN 对应的 GGSN 地址。
- IMSI (International Mobile Subscriber Identity)  
国际移动用户识别码，区别移动用户的标志。IMSI 由三部分元素构成：MCC (Mobile Country Code)、MNC (Mobile Network code) 和 MSIN (Mobile Subscriber Identification Number)。MCC 和 MNC 购成了 IMSI 的前缀，标识移动用户的本地网或 PLMN。
- MSISDN (Mobile Station International ISDN Number)  
移动台国际 ISDN 号码，用于标识公用交换电信网 (PSTN) 或综合业务数字网 (ISDN) 拨向 GSM 系统的号码。
- RAI (Routing Area Identity)  
路由区标识，用来标识 SGSN 的属性，也包括 MCC 和 MNC。

USG 支持 APN、IMSI 前缀、MSISDN、RAI 前缀以及以上四种信息元素组合的方式对 GTP 报文进行过滤。

## 类型过滤

USG 支持通过判断消息类型对 GTP 报文进行过滤，有以下几种方式：

- 过滤指定类型的消息  
对所有的 GTP 消息、GTP'消息以及未知的 GTP 消息进行过滤。另外，GTP 的 Version 0 和 Version 1 版本中包含多个消息类型，还可以配置对指定的消息类型进行过滤。
- 过滤非法的 RELOCATION 消息  
由于 RELOCATION 消息只会在相邻的 SGSN 之间产生，因此通过严格限制 SGSN 的源 IP 地址，可以对不合法的 RELOCATION 消息进行过滤，防止攻击。
- 过滤非法的位置管理消息  
当 SGSN 为 GGSN 提供位置查询功能时，一般只会为本地的 GGSN 服务，因此通过严格限制 GGSN 的源 IP 地址，可以对不合法的位置管理消息进行过滤，防止攻击。

## 长度过滤

GTP 消息长度是以字节为单位的净荷长度，即分组中除了 GTP 头的必选部分外剩余部分的长度（即除去前面的 8 个字节），不包括 GTP 报文头长度、UDP 头长度和 IP 头长度。

USG 通过解析 GTP 消息头中的长度字段，得到 GTP 报文的消息长度，如果该长度处于配置的最小长度和最大长度范围之内，则允许 GTP 报文通过，否则 GTP 报文将被丢弃。

## IE 过滤

GTP 消息中可以包含很多个不同的 IE，这些 IE 有三种状态：必选、可选、有条件选择，USG 支持对必选 IE、可选 IE 和重复 IE 进行检测，避免恶意报文攻击：

- 必选 IE 检测

当 USG 收到 GTP 消息后，会自动对必选 IE 进行检测，如果 GTP 消息中没有完全包含协议规范要求的必选 IE，则丢弃此 GTP 消息。此功能无需使用命令开启。

- 可选 IE 检测

用户可以通过命令配置 USG 对某个可选 IE 进行检测，当 USG 收到的 GTP 消息中没有包含该可选 IE，则丢弃此 GTP 消息。

- 重复 IE 检测

GTP 协议规范要求如果网元收到非协议规定的重复 IE 时，只处理第一个 IE。USG 支持对重复的必选 IE 进行检测，当 GTP 消息中必选 IE 重复出现时，丢弃此 GTP 消息。

## 扩展头过滤

在 GTP 协议 v1 版本中，GTP 消息头中可以添加扩展头，容易引起攻击，因此需要针对扩展头进行检测。

USG 支持以下两种方式对扩展头进行检测：

- 对扩展头的类型进行过滤，只有带有指定类型扩展头的 GTP 消息可以通过。
- 过滤带有重复扩展头的 GTP 消息。

## 序列号检测

序列号是 GTP 报文头中的一个字段，序列号的值逐个报文递增 1。GTP 协议通过序列号来维护报文传送时的可靠性。

GTP 隧道建立后，USG 记录序列号的值。后续的 GTP 报文到来时，将报文的序列号与记录的值进行对比，如果相差为 1，则允许报文通过，并将记录值增加 1，继续检测下一个报文；如果相差不为 1，则丢弃报文。



### 注意

配置序列号检测功能后，如果 USG 检测出一个 GTP 报文的序列号错误而将其丢弃，后续所有的 GTP 报文都有可能被丢弃，将会对业务造成很大的影响，因此一般情况下不建议配置该功能。

---

## 日志

为了监控 GTP 的工作状态，USG 提供以下日志功能：

- 根据对 GTP 报文不同的处理结果，记录日志。  
一个 GTP 协议报文经过 USG 处理后其结果状态可能为：
  - **forward**  
报文匹配 GTP 通过规则，USG 转发该报文。
  - **prohibit**  
报文匹配 GTP 拒绝规则，USG 丢弃该报文。
  - **rate-limit**  
报文超过 GTP 隧道最大速率限制，USG 丢弃该报文。



- **state-invalid**  
报文的 GTP 状态检测失败，USG 丢弃该报文。
- **tunnel-limit**  
GTP 隧道已经超过最大限制数量，USG 丢弃该报文。
- 根据指定的条件，记录日志。  
通过指定 APN、IMSI、MSISDN 的值或者消息的类型，USG 对匹配的 GTP 报文记录日志。

日志采用 SYSLOG 方式由 GTP 模块发送给信息中心，有两种信息格式：

- **basic**  
基本的 log 信息，包括：时间戳、源 IP 地址、目的 IP 地址、消息类型、报文状态 (forwarded, prohibited, rate-limited, state-invalid, tunnel-limited, matchmsgtype, matchimsi, matchmsisdn, matchapn)、接口等信息。
- **extended**  
扩展的 log 信息，除了包含 basic 的信息外，还包括：IMSI、MSISDN、APN、选择模式、信令 SGSN 地址、数据 SGSN 地址、信令 GGSN 地址、数据 GGSN 地址等信息。

#### 7.17.4.2 GTP 计费溢出攻击防范

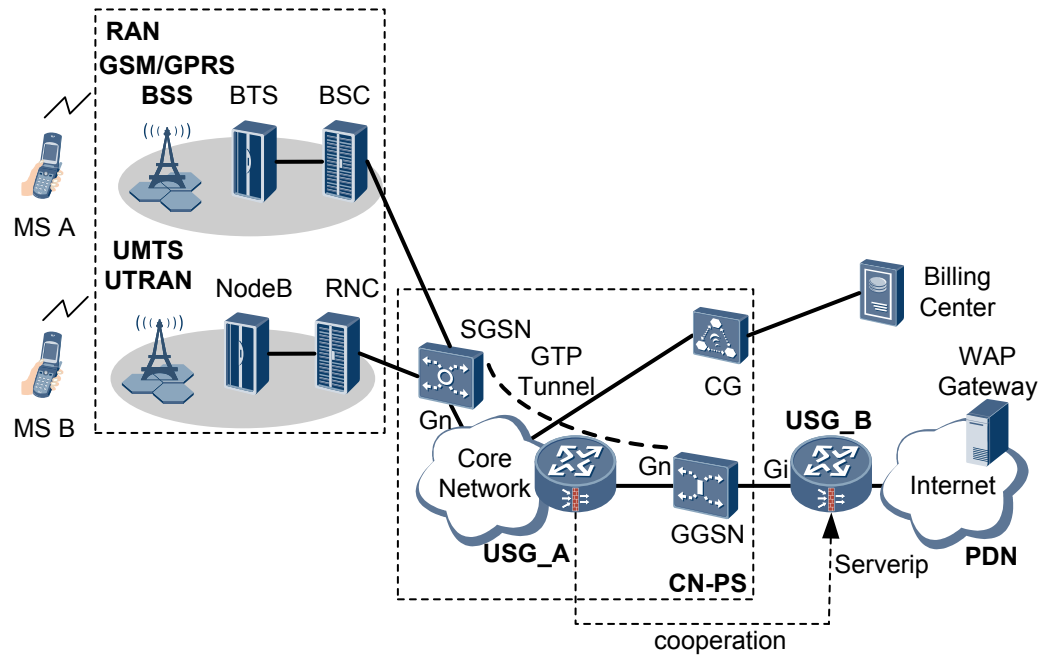
在移动分组业务中，类似于传统 IP 网络通信，MS 和 PDN 中的主机服务器都有一个 IP。PDN 中主机服务器的 IP 通常是固定并且对 GPRS 可见，可以是 Web 服务器、流媒体服务器等，为 MS 提供各种分组数据业务。MS 的 IP 地址通常都是动态获得的。

GTP 计费溢出攻击的产生过程如下：

1. MS A 激活 GTP 业务后，从 GGSN 获得 IP 地址 10.10.10.10，并向 Internet 上的主机请求大量分组数据。
2. MS A 请求大量数据后，不等待数据传输完毕就立即退出。此时 MS A 把 IP 地址 10.10.10.10 归还给 GGSN，GGSN 因查询不到 GTP 隧道而把来自 Internet 的分组数据丢弃。
3. MS B 激活 GTP 业务后，获得 IP 地址 10.10.10.10，此时 GGSN 可以重新查询到 GTP 隧道，因此把来自 Internet 的分组数据发给 MS B，并对 MS B 计费。虽然这些分组数据并不是 MS B 请求的，但 MS B 将为所有的分组数据付费，此时开始发生计费溢出。

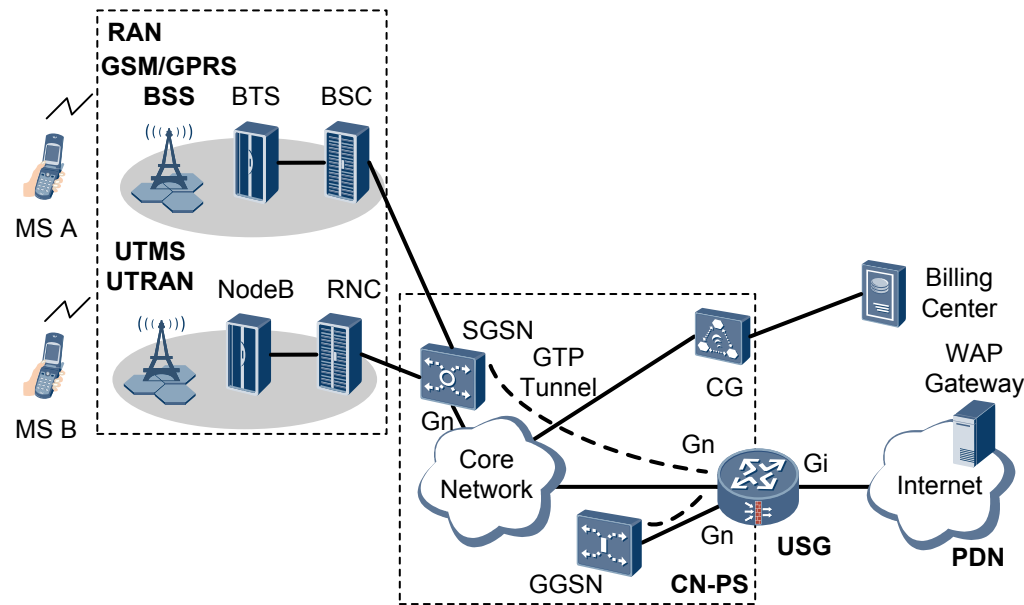
防范计费溢出攻击的关键是阻止攻击流量（来自 Internet 的分组数据）到达 GGSN，这就需要在 GGSN 接收到分组数据之前识别并丢弃流量，即在 Gi 接口进行防范。由于 GTP 协议运行于 Gn 和 Gp 接口，Gi 接口并不运行 GTP 协议，用户的会话激活和去激活信息在 Gi 接口无法识别。因此需要在 USG\_A 上对用户会话去激活信息进行识别，然后通知位于 Gi 接口的 USG\_B 对此用户请求的数据进行过滤，如图 7-43 所示。

图 7-43 联动方式的 GTP 计费溢出攻击防范组网图



此外，USG 还支持在一台设备上对计费溢出攻击进行防范，不需要两台设备进行联动。如图 7-44 所示，USG 上既支持 Gn 接口又支持 Gi 接口，Gi 接口和 Gn 接口处于不同的虚拟防火墙中，USG 在逻辑上还是相当于两台设备。当 Gn 接口收到 MS 下线信息后，将 MS 的 IP 地址记录到下线 IP 表中。然后 Gi 接口对收到的非 GTP 报文进行检测，如果报文的 IP 地址匹配了下线 IP 表中的 IP 地址，则将报文丢弃。

图 7-44 单机方式的 GTP 计费溢出攻击防范组网图



### 7.17.4.3 GTP 全局参数

#### 老化时间

USG 支持对状态表定时检测，老化超时的状态信息。老化时间缺省为 3600 秒，用户可以根据实际需要，设置不同的老化时间。



#### 注意

在 GTP 正在运行的情况下更改老化时间后，对于新建立的 GTP 隧道，将会按照新的老化时间进行老化，已经建立的 GTP 隧道不会受到影响。

---

#### GTP-in-GTP 过滤

USG 支持对 GTP-U 报文的载荷进行检测，如果 GTP-U 报文中又包含 GTP 报文，则将改报文丢弃。

#### 流量和隧道数限制

USG 支持以下两种方式对 GTP 进行限制：

- 限制 GTP 流量  
限制 USG 上的 GTP 控制平面和 GTP 用户平面的流量，保护 GPRS 网络内的 GSN 节点不受到过量 GTP 数据包的攻击。此外，USG 还支持通过限制 Supported Extension Headers Notification 消息的通过速度来防止路径拥塞攻击。
- 限制 GTP 隧道数  
检测 USG 上的 GTP 隧道数是否超过设置的阈值，丢弃超过阈值的 GTP 报文。

#### MNC 所占位数

移动国家码 MCC（Mobile Country Code）由 3 位数组成，唯一地识别移动用户所属的国家，例如，中国的 MCC 规定为 460。移动网络码 MNC（Mobile Network code）由 2 或 3 位数组成，识别移动用户所属的移动通信网。

在信息元素 IMSI 和 RAI 中都存在 MCC 和 MNC 字段，所以 USG 在对信息元素进行解码时会根据 MCC 确定 MNC 的位数。如果没有指定 MNC 的位数，缺省认为 MNC 为两位。

#### 状态检测

在 GTP 隧道建立初始，USG 为此隧道的上下文建立状态表。当后续的消息（响应类和相关的消息）到来时，针对 GTP 报文的交互过程，USG 检索状态表，对 GTP 报文进行状态的合法性检测。如果状态合法，则允许该报文通过；反之，则丢弃该报文。

同时，USG 定时检测状态表，老化超时的状态信息。



### 注意

配置 GTP 状态检测功能后，如果 USG 上的 GTP 状态和 SGSN、GGSN 上的 GTP 状态表项没有同步，USG 将会根据协议规范丢弃掉 create-pdp-context、GTP-U 等报文。而 create-pdp-context、GTP-U 等报文对 SGSN 和 GGSN 来说都属于关键报文，这些报文被丢弃后将会对业务造成很大的影响，因此一般情况下不建议配置该功能。

## 统计

USG 支持对收到的 GTP 总的报文数、允许通过的 GTP 报文数、被丢弃的 GTP 报文数、GTP 隧道数、GTP 隧道持续时间等信息进行统计，更好的监控 GTP 的运行状态。

## 隧道日志

当 GTP 隧道被清除时，USG 支持对 GTP 隧道的状态记录日志，便于用户查询。

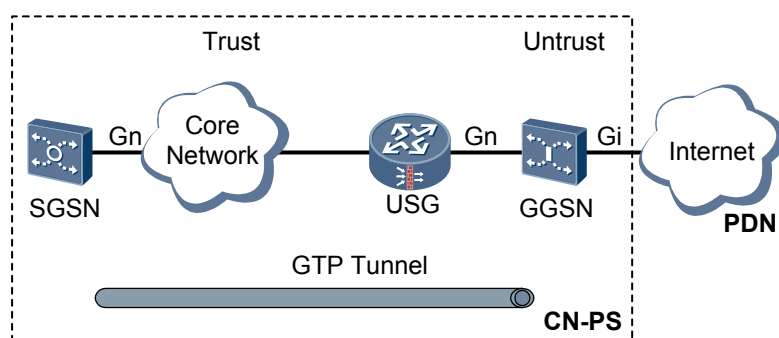
### 7.17.4.4 GTP 策略在核心网中的应用

USG 部署在核心网中，根据实际需求，应用相应的 GTP 策略，保护 SGSN 和 GGSN 设备之间数据传输的安全性。

以图 7-45 为例，USG 工作在 Gn 接口，通过配置并应用 GTP 策略，实现如下需求：

- 过滤 APN 为 test 的 GTP 报文。
- 过滤 GTP'报文。
- 限制 GTP 消息的长度为 0 ~ 1400 字节，过滤不符合长度要求的 GTP 报文。
- 对重复 IE 进行检测，过滤非协议规定的重复 IE。
- 对命中允许通过规则和拒绝通过规则的 GTP 报文记录日志，日志信息格式为 extended。

图 7-45 GTP 策略在核心网中的应用



## 7.18 IDS 联动

IDS 联动指 IDS 设备能自动侦听整个网络中是否存在恶意攻击、入侵或其他安全隐患行为，通过下发指令的方式（如动态黑名单）通知 USG,由 USG 对攻击报文进行丢弃或其他处理。

## 7.18.1 介绍

### 定义

IDS 联动功能是指 USG 与放置于内部网络的 IDS 设备联合工作，检查和防范内部网络用户的非法操作和已侵入内部网络的攻击行为。

### 目的

IDS 联动有如下作用：

- 限制用户或信息进入特定的被严格控制的站点，监控可信任网络和不可信任网络之间的访问通道，以防止外部网络的危险蔓延到内部网络上。
- 限制用户或信息从某个特定的被严格控制的站点离开，通过有效控制外部用户对内部资源的访问，确保信息安全。

## 7.18.2 可获得性

### License 支持

IDS 联动功能不需要 License 支持。

### 版本支持

IDS 联动的版本支持如表 7-8 所示。

表 7-8 产品及最低版本支持

| 产品                                                                      | 最低支持版本                                                                                                   |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100 BSR/<br>HSR V100R005C00<br>USG5500 V200R001C00 |

## 7.18.3 原理描述

### 7.18.3.1 IDS 联动的特点

USG 与 IDS 设备联动，可以充分利用专用 IDS 软件的功能，对流经网络的报文进行详细的分析与检查，探测各种可能的异常情况和攻击行为，并通过 USG 进行实时的响应。

IDS 联动有如下特点：

- IDS 设备可以对所在网络的传输进行监视。
- IDS 设备能够通过不断的软件升级，支持对网络最新的攻击行为的探测。
- IDS 设备检查通过的每个报文，从而尽早处理可疑的网络传输。

- 采用 IDS 联动方式进行攻击防范，入侵检测和攻击处理两个过程有效分离，充分发挥了各设备的优势，改善了系统性能。

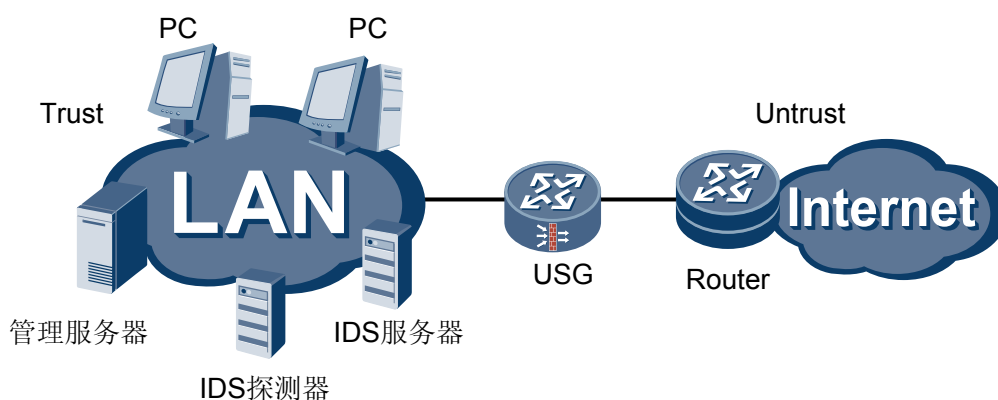
### 7.18.3.2 IDS 联动的典型应用

由于 USG 产品自身的特点，其应用存在一个局限性：USG 检查的颗粒度较粗，难以对众多协议的细节进行深入的分析 and 检查。

为解决上述问题，USG 开放了相关接口，通过与其他安全软件进行联动，从而构建统一的安全网络。

USG 与 IDS（Intrusion Detect System）设备联动组网。USG 位于内部 LAN 网络和外部网络之间，IDS 服务器和管理服务器位于内部网络。如图 7-46 所示。

图 7-46 IDS 联动典型组网图



网络中的 IDS 系统就像在网络上装备了网络分析器，对网络传输进行监视。该系统熟悉最新的攻击手段，检查通过的每个报文，从而尽早处理可疑的网络传输。具体采取的检查策略由用户使用的特定 IDS 系统和配置情况决定。

USG 与 IDS 联动，可以充分利用专用 IDS 软件的功能，对流经网络的报文进行详细的分析与检查，探测各种可能的异常情况和攻击行为，并通过 USG 进行实时的响应。如果 IDS 系统探测到异常情况或攻击行为，可以通过下发指令的方式（如动态黑名单）通知 USG，由 USG 对攻击报文进行丢弃或其他处理。

## 7.19 URPF

介绍了 URPF 的基本原理、特点和应用。

### 7.19.1 介绍

介绍 URPF 特性的定义和目的。

#### 定义

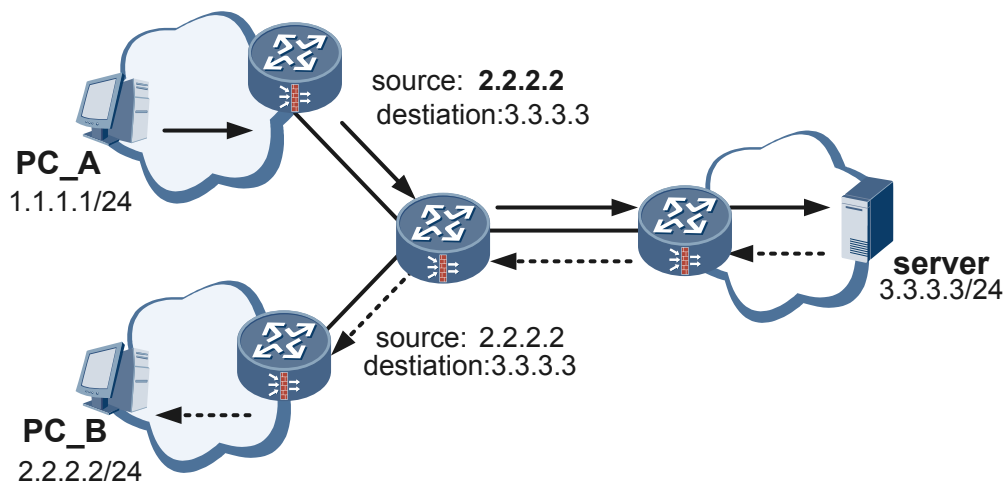
URPF（Unicast Reverse Path Forwarding）是单播逆向路径转发的简称，其主要功能是防止基于源地址欺骗的网络攻击行为。

## 目的

基于源地址欺骗发起的网络攻击，已经成为 Internet 上一种非常普遍的攻击形式。使用单播逆向路径转发（URPF）技术可以解决源地址欺骗造成的网络安全问题。

## 源地址欺骗造成的网络安全问题

图 7-47 源地址欺骗攻击示意图

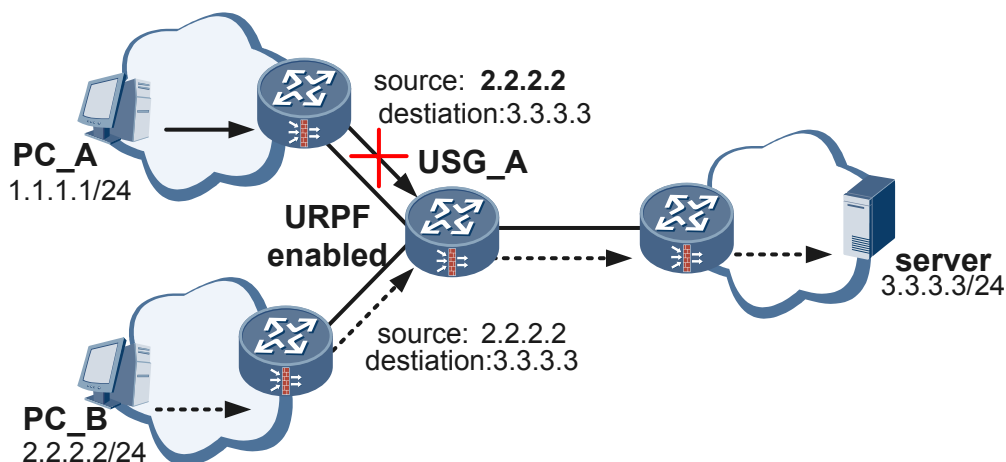


如图 7-47 所示，PC\_A 伪造了源地址为 2.2.2.2 的报文，向 server 发起请求，server 在收到请求报文后就向 PC\_B (2.2.2.2) 发送回应报文。PC\_A 发起的这种伪造报文对 Server 以及 PC\_B 都造成了攻击。

大量的这种伪造报文会造成 DoS (Denial of Service) 攻击，对网络的影响极大。

## 在网络入口处解决源地址欺骗

图 7-48 URPF 防止基于源地址欺骗示意图



如图 7-48 所示，在 USG\_A 上使能 URPF 功能后，PC\_A 发送的伪造报文就会在 USG\_A 处被直接丢掉，而 PC\_B 的报文的被正常的转发。

7.19.2 可获得性

介绍 URPF 特性的 License 支持信息。

License 支持

本特性无需 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

7.19.3 原理描述

介绍了 URPF 在网络中的应用场景及处理流程。

USG 实现了 URPF 的两种模式:

- 严格模式
- 松散模式

严格模式

建议在路由对称的环境下使用 URPF 严格模式，即：不仅要求在转发表中存在相应表项，还要求接口一定匹配才能通过 URPF 检查。

如果两个网络边界路由器（此处为 USG）之间只有一条路径的话，这时，路由能够保证是对称的，使用严格模式能够最大限度的保证网络的安全性。

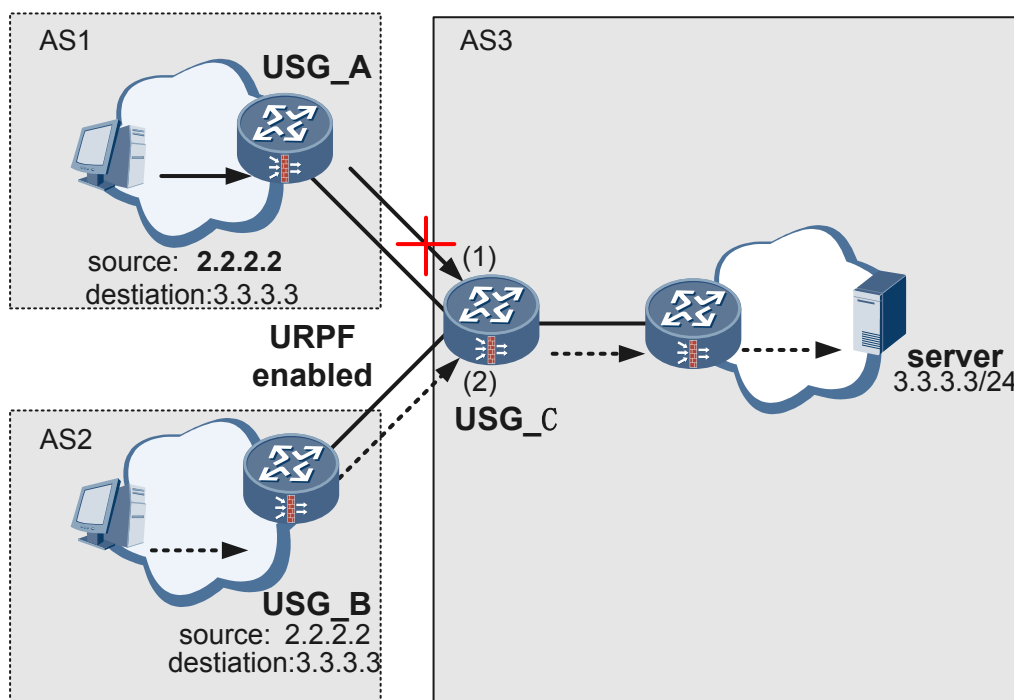
如图 7-49 所示，AS1 和 AS2 与 AS3 之间为单连接。在 USG\_C 的(1)接口和(2)接口上启动 URPF，可以保护 AS3 免受来自 AS1 和 AS2 的源地址欺骗攻击。

如果，AS1 中的一台 PC 伪造了一个源地址为 2.2.2.2 的报文，向 AS3 中的 Server 发送请求。USG\_C 在接受到这个报文后，对其进行入接口检查，发现源地址为 2.2.2.2 的报文应该从(2)接口进入，而不应该从(1)接口进入。USG\_C 认为该报文源地址是伪造的，直接丢弃该伪造报文。

从 AS2 发向 Server 的正常报文，检查通过后，被正常的转发。



图 7-49 URPF 单宿主客户应用环境示意图



## 松散模式

在不能保证路由对称的环境下使用 URPF 的松散模式，即：不检查接口是否匹配，只要存在针对源地址的路由，报文就可以通过。

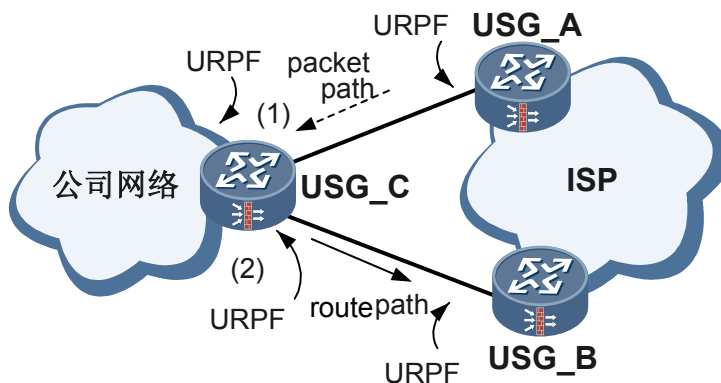
不能保证路由对称的情况有两种，即：多宿主单 ISP 客户和多宿主多 ISP 客户，如下。

- 多宿主单 ISP 客户

如图 7-50 所示，一个公司网络和某个 ISP 之间，为了保证可靠性，有多条连接。这时，就不能够保证公司网络和 ISP 之间路由的对称性，必须使用松散模式。

URPF 的松散模式可以保证较强的安全性。如果在 USG C 的接口 (1) 上启动松散模式的 URPF，则当有报文通过 USG A 到达 USG C 时，USG C 发现以此报文的源地址为目的地址的路由在 USG C 上存在，但出接口不是接口 (1)，而是接口 (2)，此报文就不会被丢弃；USGC 发现以此报文的源地址为目的地址的路由在 USG C 上不存在，则丢弃该报文。

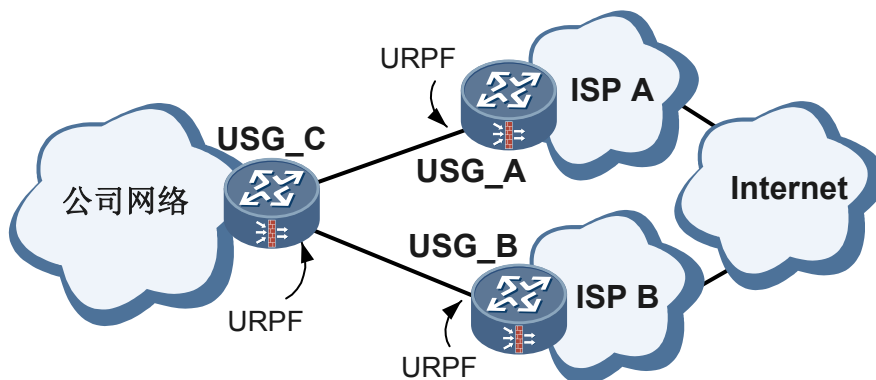
图 7-50 URPF 多宿主单 ISP 客户应用环境示意图



- 多宿主多 ISP 客户

在图 7-51 环境中，客户与多个 ISP 连接，很难保证公司网络和两个 ISP 之间路由的对称性，必须使用 URPF 松散模式。

图 7-51 URPF 多宿主多 ISP 客户应用环境示意图



另外，URPF 应用具有以下特点。

- 如果用户希望某些特殊报文任何情况都可以通过 URPF 的检查，即不受 URPF 丢弃的控制，可以在 ACL 中指定这些特殊的源地址。
- 许多用户的 USG 可能只有一条缺省路由指向 ISP，此时，需要配置允许匹配缺省路由选项。

## URPF 的处理流程

URPF 的处理流程如下：

1. 如果报文的源地址在 USG 的 FIB 表中存在。
  - 对于 strict 型检查，反向查找报文出接口，若只有一个出接口和报文的入接口一一匹配，则报文通过检查；否则报文将被拒绝。当有多个出接口和报文的入接口相匹配时，必须使用 loose 型检查。（反向查找是指查找以该报文源 IP 地址为目的 IP 地址的报文的出接口）。
  - 对于 loose 型检查，当报文的源地址在 USG 的 FIB 表中存在（不管反向查找的出接口和报文的入接口是否一致），报文就通过检查；否则报文将被拒绝。
2. 如果报文的源地址在 USG 的 FIB 表中不存在，则检查缺省路由及 URPF 的 allow-default-route 参数。
  - 对于配置了缺省路由，但没有配置参数 allow-default-route 的情况。  
不管是 strict 型检查还是 loose 型检查，只要报文的源地址在 USG 的 FIB 表中不存在，该报文都将被拒绝。
  - 对于配置了缺省路由，同时又配置了参数 allow-default-route 的情况。
    - 如果是 strict 检查，只要缺省路由的出接口与报文的入接口一致，则报文将通过 URPF 的检查，进行正常的转发。如果缺省路由的出接口和报文的入接口不一致，则报文将拒绝。
    - 如果是 loose 型检查，报文都将通过 URPF 的检查，进行正常的转发。

3. 当且仅当报文被拒绝后，才去匹配 ACL 列表。如果被 ACL 允许通过，则报文继续进行正常的转发；如果被 ACL 拒绝，则报文被丢弃。

## 7.20 终端安全管理（TSM 联动）

设备与 TSM 系统联动，主要是为了解决内网终端用户带来的安全隐患。本节介绍 TSM 联动的基本概念、规格、可获得性及原理。

### 7.20.1 介绍

#### 定义

TSM 联动是指设备通过在 TSM 系统中承担 SACG 的角色，协助 TSM 系统实现内网终端用户的安全控制。

#### 目的

随着网络的发展与防火墙设备的普及，企业的网络安全威胁的主要来源正逐渐从外网转移到内网。TSM 方案的提出主要就是为了解决内网终端用户带来的安全隐患。USG 可以在 TSM 方案中承担 SACG 的角色，主要起到不同区域之间的隔离，以及对终端用户访问权限进行实际控制的作用。

### 7.20.2 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 7.20.3 原理描述

#### 7.20.3.1 终端安全理念与 TSM 简介

随着网络的发展与防火墙设备的普及，企业的网络安全威胁的主要来源正逐渐从外网转移到内网，包括：

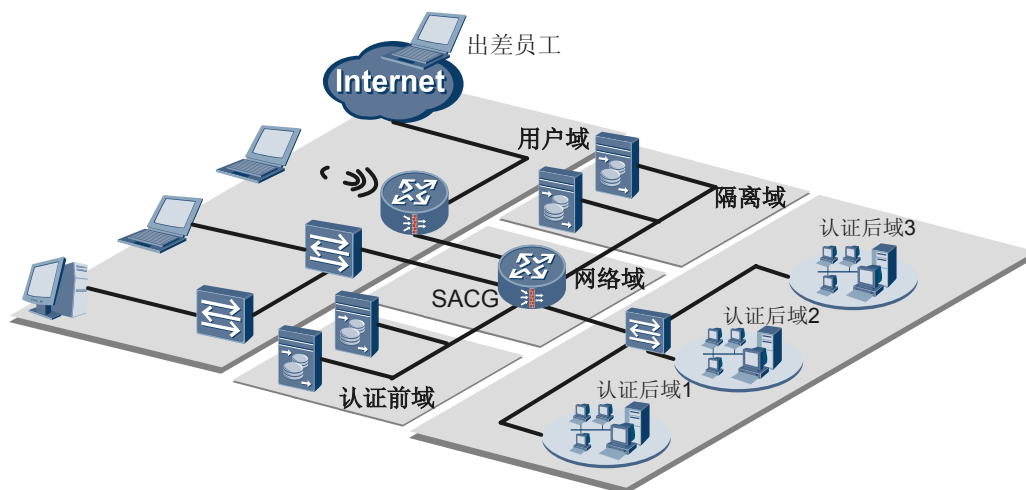
- 非法终端和非授权终端对业务系统的访问，主要包括以下几种情况：
  - 非自有的终端设备接入企业内网
  - 不合法人员利用企业设备接入企业内网
  - 合法人员越权访问网络资源
- 终端不安全导致病毒的扩散
- 终端数量大、系统复杂、员工的违规行为得不到监控

TSM(Terminal Security Management)方案正是为了解决这一问题而提出来的。它将网络划分为以下几类区域：

- 用户域  
所有接入企业内网的终端设备，例如台式机、便携机以及通过 Internet 接入的出差员工、驻外机构、合作伙伴等。
- 网络域  
进行流量转发的网络设备所组成的域，承载业务流量，实现各网络的互联。SACG 就部署在这个域中。
- 认证前域  
认证前域是终端设备在完成认证之前可以访问的区域。该区域主要用于对终端设备和用户进行认证、授权、策略管理、补丁下发等。TSM 方案的中大部分组件都部署在该区域中。
- 受控域  
受控域是终端在完成认证之后才可以访问的区域。包括两种：
  - 隔离域  
隔离域是指在终端用户通过了身份认证但未通过授权时可以访问的区域。通常将能够帮助终端用户消除安全隐患的相关资源（如补丁服务器、防病毒服务器等）部署在本区域。
  - 认证后域  
认证后域是企业真正的核心资源所在的区域，终端设备和用户都必须进行认证和授权之后才能访问自己的权限相对应的安全区域。

TSM 方案的典型组网图如图 7-52 所示。

图 7-52 TSM 方案典型组网图



TSM 的主要工作流程是：

1. 终端设备接入网络，发起认证请求。
2. SACG 将接收到的认证请求转发给服务器中的 TSM 控制器，同时阻止未通过认证和授权的终端设备访问认证后域。
3. TSM 系统进行权限以及终端安全性核查。如果终端设备存在安全漏洞，TSM 系统会首先开放该设备访问隔离域的权限，使其可以下载补丁或病毒库。修复安全漏洞成功后，TSM 系统确定是否允许其访问认证后域，并且确定允许其访问哪个认证后域。
4. TSM 系统将针对该终端生成的用户访问控制策略下发到 SACG 上。
5. SACG 根据收到的策略生成相应的转发策略，判断是否允许相关终端设备通过 SACG 访问认证后域。

### 7.20.3.2 SACG 工作原理简介

USG 在 TSM 方案中承担 SACG 这个角色，主要作用是进行网络访问权限控制。在 TSM 方案中，对各类用户进行访问控制的策略不需要手工创建，而是由 TSM 系统根据相关的策略配置，自动下发到 USG 上。

终端通过 SACG 进行认证请求的方式主要有以下两种：

- TSM 代理认证

TSM 代理是 TSM 的一个组件，作为 TSM 的客户端安装在终端设备上。主要承担用户身份认证、终端设备安全性检查以及与 TSM 后台服务器的自动交互等工作。安装 TSM 代理后，与 TSM 控制器联动的所有过程都是由 TSM 代理自动完成，实现了对用户的透明。

- 无代理 Web 认证

如果终端设备上不能安装 TSM 代理软件，例如终端设备的操作系统为 UNIX，那么在用户没有经过认证就访问 Web 的情况下，会被强制跳转到一个用于认证的 Web 界面。用户可以通过该界面进行相应的认证请求。无代理 Web 认证不能实现 TSM 方案的所有功能，所以不推荐使用。

TSM 服务器将访问控制策略下发给 SACG 的原理是：

- 通过在 SACG 上配置 TSM 联动功能，将 SACG 上的 ACL3099 ~ 3999 作为接纳 TSM 下发控制策略的容器，这 901 条 ACL 分别对应 ID 号为 0 ~ 900 的 901 种角色，每种角色对应 TSM 系统中的一个受控域。其中 ACL3099 对应角色 0，是所有接入用户的缺省角色，可以用来允许未经认证的用户访问认证前域。剩余的 ACL3100 ~ 3999 对应角色 1 ~ 900 号角色，这些角色是由 TSM 系统下发的普通角色。

执行 **default acl 3099** 命令指定缺省 ACL 后，ACL3099 中的原有规则会被自动清除，并且不能被人工配置。

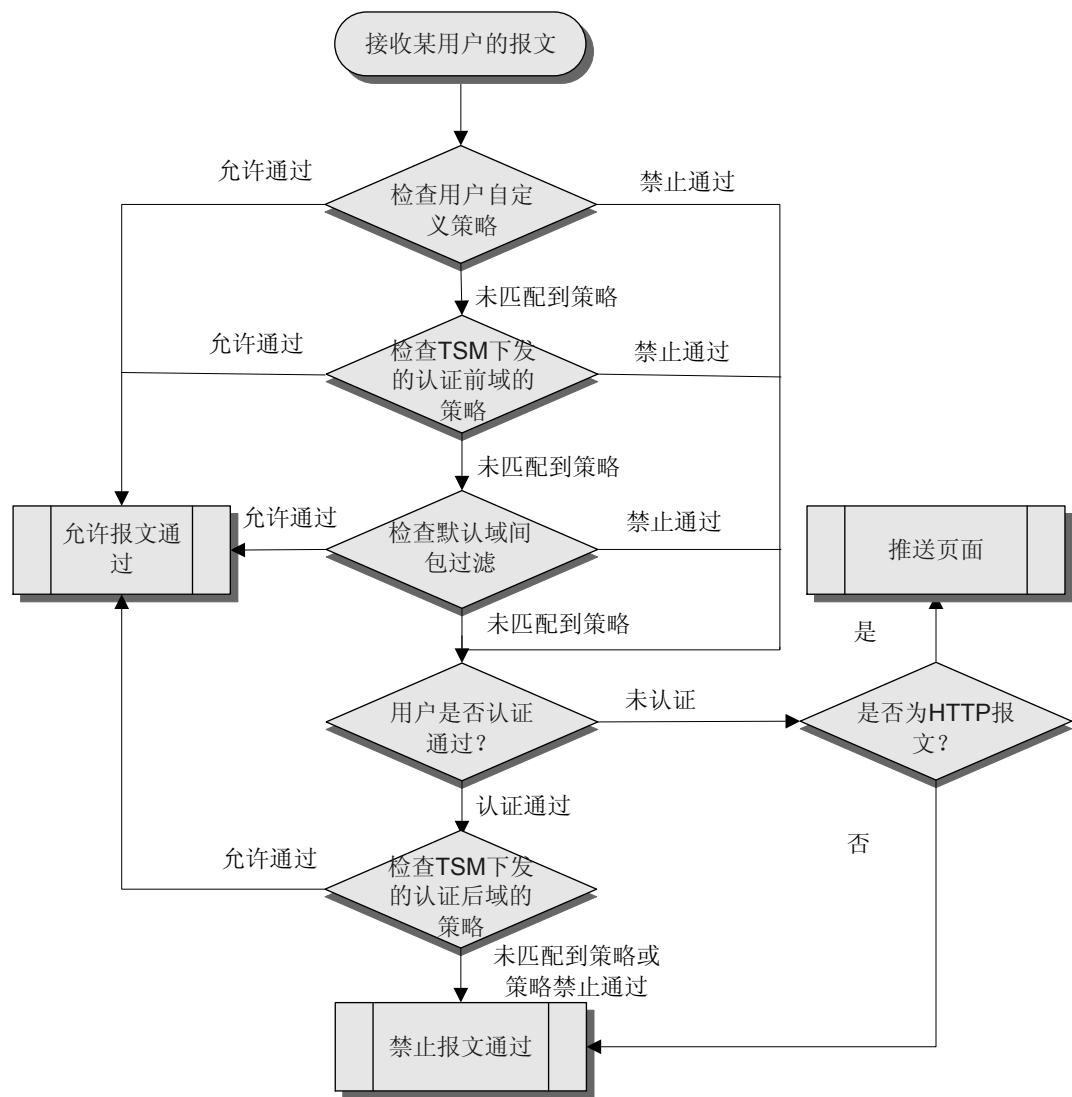
执行 **right-manager server-group enable** 开启与 TSM 服务器的连接后，ACL3100 ~ ACL3999 将不能被人工配置。如果在开启之前，这些 ACL 已经包含规则或者正在被其他功能使用，必须先手工清除规则，并且取消在其他功能中的使用，才能开启与 TSM 服务器的连接。

- 一个终端设备接入后，通过 SACG 向 TSM 服务器发起认证请求，认证通过后，TSM 服务器将该终端设备的 IP 地址以及对应的角色信息发给 SACG。

- SACG 根据 TSM 服务器下发的各条 ACL 中的信息，建立一个源 IP 监控表，记录 IP 地址、角色、角色的权限以及可访问的资源等信息的对应关系。

SACG 收到一个来自终端设备的报文后，在进行域间包过滤时，按如图 7-53 所示流程进行处理：

图 7-53 TSM 联动下的策略查找流程



**说明**

当安全域间同时启用 TSM 联动与 UTM 检测功能时，TSM 联动相关策略和默认域间包过滤用来做报文转发控制；严格的域间包过滤策略不再做报文转发控制，只控制报文是否需要做 UTM 检测。图 7-53 流程处理完之后，报文的具体处理如下。

- 报文允许通过，且命中严格的域间包过滤策略的 permit 规则，报文做 UTM 检测，UTM 检测结果为允许通过则放行报文，UTM 检测结果为阻断报文，则丢弃报文；
- 报文允许通过，但没有命中严格的域间包过滤策略的 permit 规则，报文放行，不做 UTM 检测。

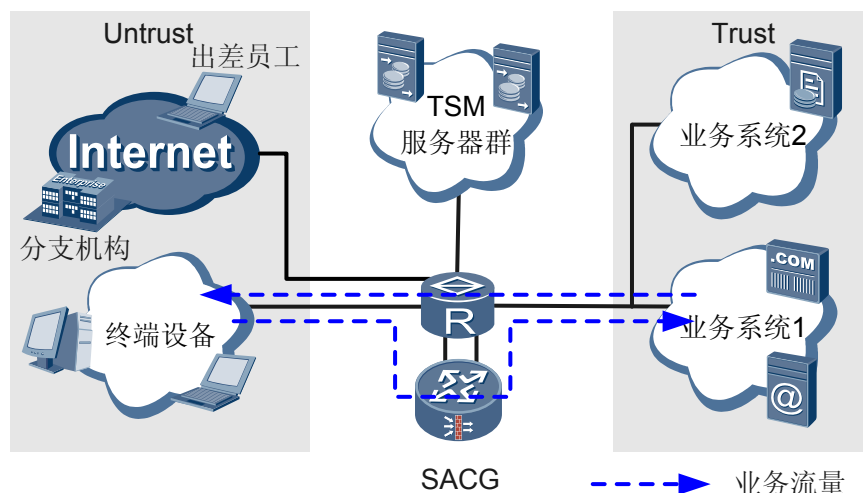
7.20.3.3 SACG 的部署方式

## 旁挂模式

通常情况下，使用 TSM 功能时内网环境已经存在，终端用户和业务系统已经通过一个核心转发设备（三层交换机或路由器）连接在了一起。如果在部署 TSM 系统时，使用 SACG 替换掉这个转发设备，就会涉及到重新构建内网环境和路由的重新学习等问题。为了避免这些问题，可以使用旁挂模式来部署 SACG。同时通过旁挂模式也可以避免由于 SACG 故障导致网络不通的问题。

旁挂模式的典型的组网图如图 7-54 所示。

图 7-54 旁挂模式的 TSM 联动组网图

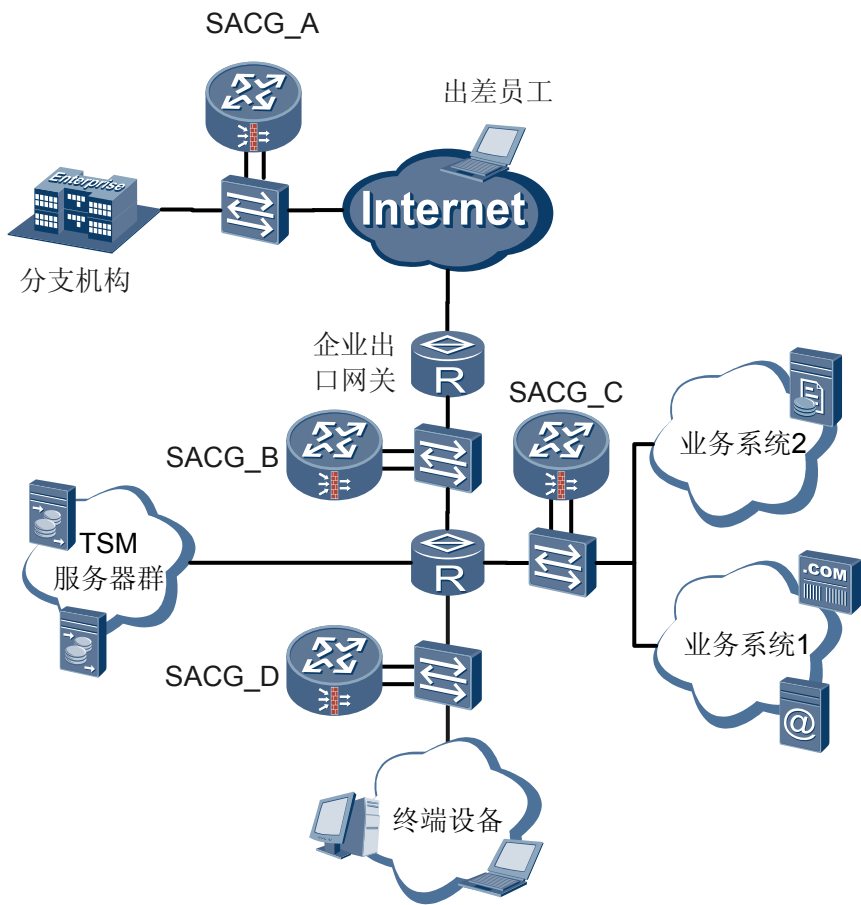


在该组网中，连接 Trust 和 Untrust 区域的路由器也可以使用交换机来代替。通过在路由器或交换机上配置重定向或者策略路由，将 Untrust 区域的终端设备访问 Trust 区域的业务系统的流量从路由器或交换机转发到 SACG 上，由 SACG 对其进行处理。期间 SACG 会和 TSM 服务器进行一系列交互，以确定该流量所属的用户的权限，并进行相应的处理。当这些流量被允许通过 SACG 后，由业务系统回应的流量会直接通过路由器或交换机转发给 Untrust 区域。

这种组网是在用户原有组网基础上直接添加 SACG 和 TSM 服务器群，不需要破坏原有网络结构。而且由于只有一个方向的流量会经过 SACG，所以对 SACG 的负载要求较小。推荐使用这种模式部署 TSM 系统。

在图 7-54 中，从 Internet 接入的出差员工和分支机构也都作为不信任的终端，划入到 SACG 的 Untrust 区域中，但是在某些场景下，Internet 本身就是一种需要认证后才可访问的资源，应当作为认证后域加以保护。所以在这种情况下，可以采用 TSM 系统的多 SACG 的组网，如图 7-55 所示。

图 7-55 多 SACG 的 TSM 联动旁挂模式组网



在该组网中，各个 SACG 的主要作用如下：

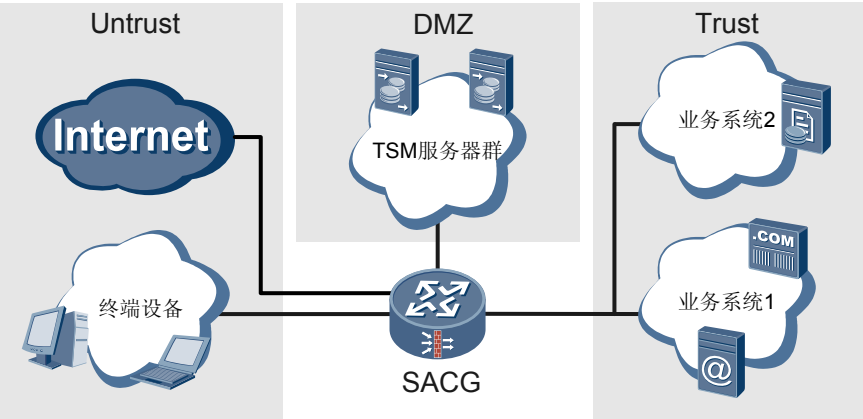
- SACG\_A  
分支机构的 SACG，部署在分支机构出口，用于控制分支机构访问 Internet 以及总部内网的行为。如果分支机构不需要单独控制，可以不部署此 SACG。
- SACG\_B  
企业总部的 SACG，部署在企业出口，用于控制内网用户访问 Internet 的行为。如果不需要控制此类行为，即不将 Internet 看做认证后域，可以不部署此 SACG。
- SACG\_C  
企业总部业务系统的 SACG，部署在进入业务系统的入口，用于控制所有终端设备访问业务系统的行为。
- SACG\_D  
内网用户的 SACG，部署在内网用户汇聚的交换机旁，用于控制内网用户之间互访的行为。如果不部署此 SACG，内网用户就可以通过交换机直接相互访问，不受控制。

直挂模式

直挂模式是指将设备直接串接在用户原有组网中。典型的组网图如图 7-56 所示。



图 7-56 直挂模式的 TSM 联动组网图



在该组网模式下，有两种使用方式：

- 使用二层接口连接 Untrust、DMZ、Trust 三个安全区域  
这种方式适合于使用 TSM 前，SACG 所在网络位置是一个交换机的情况。相当于通过 USG 的二层口替代了之前的交换机。  
此处的二层接口为三层接口工作在二层模式下的接口。
- 使用三层接口连接 Untrust、DMZ、Trust 三个安全区域  
这种方式适合于使用 TSM 前，SACG 所在网络位置是一个路由器的情况。相当于通过 USG 的三层接口替代了之前的路由器。这种方式中，SACG 上需要正确配置内网的路由。

两种部署模式在配置上的区别

表 7-9 TSM 的两种部署模式在配置上的区别

| 项目 | 旁挂模式                                                              | 直挂模式                                                                                                                                                                                                                                           |
|----|-------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 接口 | SACG 通过两个三层接口与交换机相连。<br>交换机的两个接口需要划入不同的 VLAN，使用 VLANIF 与 SACG 通信。 | <ul style="list-style-type: none"><li>● 可以使用三个二层接口与各个区域相连。<br/>这种情况下相当于在网络中串接一个交换机。此时各个二层接口应该属于同一 VLAN，所以不能位于同一接口卡，否则不能触发 TSM 功能检查。</li><li>● 可以使用三个三层接口与各个区域相连。<br/>这种情况下相当于在网络中串接一个路由器。三个接口必须属于不同网段。所以这些接口既可以属于同一个接口卡，也可以属于不同的接口卡。</li></ul> |

| 项目        | 旁挂模式                                                                                                                                  | 直挂模式                                                                                                                  |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| 安全区域      | SACG 通过交换机与终端设备相连的接口划入 Untrust 区域，与业务系统相连的接口划入 Trust 区域。                                                                              | SACG 与终端设备相连的接口划入 Untrust 区域，与业务系统相连的接口划入 Trust 区域，与 TSM 服务器相连的接口划入 DMZ 区域。                                           |
| 重定向与路由配置  | <p>需要在与 SACG 相连的交换机或路由器上配置端口重定向，将从终端设备收到的流量全部转发到 SACG 上。</p> <p>需要在 SACG 上配置一条指向交换机或路由器的缺省路由，以便将检测完成的流量回注到交换机或路由器上，使其可以被转发到业务系统中。</p> | <p>需要在 SACG 上配置路由协议，保证内网网络的连通。</p> <p>如果由于 SACG 的部署，导致原有网络的 IP 地址分配和路由发生变化，需要在与 SACG 相连的路由器上修改相应的路由表项，以保证内网网络的连通。</p> |
| 其他安全功能的使用 | 由于旁挂模式中，只有一个方向的流量会经过 SACG，所以需要关闭 SACG 的状态检测功能。USG 上基于状态检测实现的功能在旁挂模式中均不可用。                                                             | 直挂模式中，USG 的大部分安全功能均可正常使用，但是由于直挂模式下两个方向的流量都需要经过设备的检测，流量负载较大，所以开启其他安全功能时要考虑设备的性能承受情况。                                   |

# 8 UTM (USG2200/5100/5500、USG2200/5100 BSR/HSR)

---

## 关于本章

### 8.1 反病毒 (AV)

AV (Anti-Virus) 特性可以检测包括病毒、间谍软件、蠕虫、木马病毒、后门和其它恶意代码的各种威胁，并根据策略进行响应，有效保护网络。

### 8.2 入侵防御 (IPS)

入侵防御 (Intrusion Prevention System) 能够有效防御来自应用层的攻击，例如缓冲区溢出攻击、木马、后门攻击、蠕虫等。

### 8.3 邮件过滤

邮件过滤分为 RBL 过滤和邮件内容过滤。

### 8.4 Web 过滤

通过配置 Web 过滤，您可以防止内网用户进行以下操作：访问不合适的网站、查看非法网页内容或在上传/下载不符合要求的文件、通过搜索引擎搜索敏感信息。

### 8.5 FTP 过滤

通过 FTP 过滤，可以对 FTP 操作（上传、下载、删除）、上传/下载的文件名、文件类型、文件大小进行控制。

### 8.6 应用控制 (DPI)

DPI (Deep Packet Inspection) 提供对 TCP/UDP 传输层以上具体应用协议的识别功能，基于该功能，支持对网络流量按协议进行细分统计，对某些具体协议流量进行流量阻断或放行。

## 8.1 反病毒（AV）

AV（Anti-Virus）特性可以检测包括病毒、间谍软件、蠕虫、木马病毒、后门和其它恶意代码的各种威胁，并根据策略进行响应，有效保护网络。

### 8.1.1 介绍

#### 定义

AV 是一种安全机制，它通过比对扫描的文件内容与病毒特征来检测病毒，并通过一定的响应方式对感染病毒的文件进行处理。

#### 目的

保护系统和网络免受多种动态威胁的侵扰。AV 可以防范的威胁包括：病毒、间谍软件、蠕虫、木马、后门及恶意代码等。

### 8.1.2 可获得性

#### License 支持

只有获取了 License，才能使用 AV 升级功能。当 License 过期后，AV 功能仍可以使用，但不能升级病毒库和引擎。

#### 版本支持

| 产品                                             | 最低支持版本                                                              |
|------------------------------------------------|---------------------------------------------------------------------|
| Secoway USG2200/5100/5500 和<br>USG2200/5100HSR | USG2200/5100、USG2200/5100 HSR<br>V100R005C00<br>USG5500 V200R001C00 |

### 8.1.3 原理描述

#### 原理

病毒是一种可执行代码，可感染或附在其他可执行代码上。有些病毒可直接导致删除文件或锁住系统；有些病毒感染其他文件后，使这些文件在传播时产生大量的伪造数据来耗尽受感染主机或网络的资源。

USG 将接收到需要进行病毒扫描的文件的内容与病毒特征库进行比较，判断扫描内容中是否含匹配病毒特征的信息。如果检测到病毒，USG 根据 AV 策略进行响应，对含病毒的内容进行删除或告警。如果没有检测到病毒，则 USG 放行该文件。

#### 支持处理的文件

USG 能够对使用以下协议进行传输的文件进行病毒扫描和相应处理：

- HTTP (Hypertext Transfer Protocol)，超文本传输协议
- FTP (File Transfer Protocol)，文件传输协议
- SMTP (Simple Mail Transfer Protocol)，简单邮件传输协议
- POP3 (Post Office Protocol 3)，邮局协议版本 3

## 扫描方式

USG 对文件进行扫描的方式分为两种：

- 智能扫描  
根据文件的真实类型进行扫描。  
此方式下 USG 扫描所有文件。
- 指定扩展名扫描  
根据文件扩展名表示的文件类型进行扫描。  
此方式下 USG 只扫描带系统预定义的扩展名和用户自定义的扩展名的文件，不扫描其他文件。

## 8.2 入侵防御 (IPS)

入侵防御 (Intrusion Prevention System) 能够有效防御来自应用层的攻击，例如缓冲区溢出攻击、木马、后门攻击、蠕虫等。

### 8.2.1 介绍

#### 定义

IPS 是一种安全机制，它通过监控或者分析系统事件，检测入侵，并通过一定的响应方式，实时地中止入侵行为。

#### 目的

保护系统或系统资源不受未经授权的访问。IPS 可以防范的威胁包括：蠕虫、病毒、广告软件、来自 Internet 的无授权访问、内部用户的安全违规等。

### 8.2.2 可获得性

#### License 支持

只有获取了 License，才能使用 IPS 升级功能。当 License 过期后，IPS 功能仍可以使用，但不能升级签名库和引擎。

#### 版本支持

| 产品                                             | 最低支持版本                                                              |
|------------------------------------------------|---------------------------------------------------------------------|
| Secoway USG2200/5100/5500 和<br>USG2200/5100HSR | USG2200/5100、USG2200/5100 HSR<br>V100R005C00<br>USG5500 V200R001C00 |

## 8.2.3 原理描述

### 产生背景

随着攻击手段和工具的不断涌现、攻击技术的日趋成熟，传统防火墙的网络层攻击检测已无法满足安全需要。在此背景下，出现了 IPS 技术。

IPS 设备与传统防火墙以及 IDS (Intrusion Detection System) 设备相比主要有以下不同：

- 传统防火墙很难对基于应用层的攻击进行预防和阻止。IPS 设备能够有效防御应用层攻击，如：缓冲区溢出攻击、木马、后门攻击、蠕虫等。
- IDS 设备只检测和告警。IPS 设备不仅能够检测入侵的发生，而且能通过一定的响应方式，实时地中止入侵行为的发生和发展，实时地保护信息系统。

#### 说明

USG 有两种运行模式：防火墙模式和 UTM 模式。当 USG 运行模式为防火墙时，只能检测网络层的攻击行为。当 USG 运行模式为 UTM 时，可以配置 IPS 功能检测应用层的攻击行为，此时 USG 能够同时检测网络层和应用层的攻击行为。

### IPS 处理流程

IPS 的基本处理流程如下：

1. 重组应用数据  
USG 支持对 IP 分片报文重组以及 TCP 流重组，确保了应用数据的连续性，防止躲避 IPS 检测的攻击行为。
2. 协议识别  
与传统的根据端口识别协议不同，USG 能根据报文内容识别多种常见应用层协议，并对这些数据进行检测，提高了攻击行为的检测率。
3. 匹配签名  
USG 采用基于状态的匹配引擎。USG 将报文内容与签名进行比较，只有匹配签名的报文才会进行下一步 IPS 处理。
4. 完成检测后，USG 根据用户配置的策略和响应方式对匹配到 IPS 签名的报文进行处理。

#### 说明

关于 IPS 签名、策略和响应方式的概念请参见下文中的 IPS 签名、策略定制和响应方式。

### IPS 签名

IPS 签名用来描述网络中存在的攻击行为的特征，USG 通过将报文内容和 IPS 签名进行比较来检测和防范攻击。

USG 的 IPS 签名分为两类：

- 预定义签名  
USG 中预先定义的签名。用户购买了带 IPS 升级功能的 License 后即可获得包含预定义签名的签名库，并且能够不断地从安全服务中心获取新的 IPS 版本来更新签名库。

- 自定义签名  
用户根据网络流量特点对特定的入侵行为自行定义的签名，自定义签名的攻击特征使用正则表达式定义。

策略定制

传统的 IPS 策略基本上不区分受保护对象，对所有流量都进行检测。因此会出现将一些不需要检测的流量也进行了检测，导致检测性能低下。USG 支持用户根据实际网络情况定制细致的 IPS 策略，并应用在特定的流量上，提高了检测性能。

设备可以通过以下方式定制 IPS 策略：

- 签名集  
用户根据需要配置各种过滤条件来过滤签名集中包含的签名，并配置签名集的启用状态和响应方式。
- 覆盖签名  
用户可以根据需要在 IPS 策略里修改某个签名的响应方式，称为覆盖签名。配置了覆盖签名之后，这条签名在该策略中的响应方式临时被修改了，当用户 undo 覆盖签名之后，这条签名在此策略中的响应方式恢复成签名的原响应方式。  
 **说明**  
在同一条 IPS 策略中，覆盖签名优先级高于签名集的配置，如果配置的覆盖签名和签名集中的配置有冲突，覆盖签名的配置生效。
- 同步（复制）策略/策略模板  
设备提供了策略模板，模板内容为针对指定场景预先定义的签名集及其启用状态和响应方式。同步策略/策略模板后，新建的策略将具有与原策略/策略模板相同的配置。通过同步已经创建的策略或策略模板，将已有策略或策略模板中的签名引入到新建策略中，然后可以在此基础上进一步修改，方便配置。

攻击响应方式

一个签名包含一种攻击特征，当报文命中签名时，USG 将该报文识别为攻击报文，然后按照签名的攻击响应方式处理该报文。

IPS 攻击响应方式如表 8-1 所示。

表 8-1 攻击响应方式

| 工作模式 | 处理策略 | 实际动作               |
|------|------|--------------------|
| 告警模式 | 告警   | USG 不对报文进行处理，记录日志。 |
|      | 阻断   |                    |
| 防护模式 | 阻断   | USG 阻断报文，记录日志。     |
|      | 告警   | USG 不对报文进行处理，记录日志。 |



说明

当报文命中多个签名，对该报文的响应方式如下：

- 如果这些签名的响应方式都为 **Alert** 时，响应方式为告警。
- 如果这些签名中至少有一个签名的响应方式为 **Block** 时，响应方式为阻断。

## 8.3 邮件过滤

邮件过滤分为 RBL 过滤和邮件内容过滤。

### 8.3.1 介绍

#### 定义

邮件过滤包括 RBL 过滤和邮件内容过滤。

- RBL (Real-time Blackhole List) 过滤，即反垃圾邮件

RBL 通过定义一个黑名单列表，在该列表中的 IP 地址对外发布的邮件即为垃圾邮件。同时，提供 RBL 服务的机构会实时更新黑名单列表来保证可以过滤最新的垃圾邮件。

设备检测到 SMTP 连接请求后，提取 SMTP 连接的源 IP 地址，与第三方组织提供的动态更新的远程实时黑名单以及本地黑白名单进行比对，对列入黑名单的邮件进行阻断。

- 邮件内容过滤

当内网用户通过 Webmail 或 SMTP/POP3 客户端收发电子邮件时，邮件内容过滤对邮件地址、主题（标题）、正文、附件大小和数量、附件名或附件扩展名等进行监控，防止数据泄漏或敏感信息传输。



说明

邮件内容过滤与 RBL 过滤的区别在于：

- 邮件内容过滤支持对 Webmail 或通过 SMTP/POP3 传输的邮件进行过滤，而 RBL 过滤只过滤 SMTP 传输的邮件。
- 邮件内容过滤对邮件内容进行过滤，而 RBL 过滤根据 SMTP 连接的源 IP 地址进行过滤。

#### 目的

- 保护用户或邮件服务器免受垃圾邮件的侵扰。
- 防止企业内部的机密信息通过邮件泄漏到外部。
- 防止内网用户发送和接收包含敏感信息的邮件，避免触犯法律法规和不良内容的侵扰。

### 8.3.2 参考标准和协议

与邮件过滤特性相关的参考标准与协议如下：

- RFC 1939: Post Office Protocol - Version 3
- RFC 1957: Some Observations on Implementations of the Post Office Protocol (POP3)
- RFC 2449: POP3 Extension Mechanism



- RFC 2821: Simple Mail Transfer Protocol
- RFC 1945: Hypertext Transfer Protocol -- HTTP/1.0
- RFC 2616: Hypertext Transfer Protocol -- HTTP/1.1

8.3.3 可获得性

License 支持

RBL 过滤功能受 License 控制。当 License 过期后，RBL 功能不可用。

RBL 远程查询服务受激活码控制。当激活码过期后，用户只能使用本地黑白名单，如果要使用远程实时黑名单，必须输入或导入有效的激活码。

邮件过滤不受 License 控制。

版本支持

| 产品                                             | 最低支持版本                                                                  |
|------------------------------------------------|-------------------------------------------------------------------------|
| Secoway USG2200/5100/5500 和<br>USG2200/5100HSR | USG2200/5100、USG2200/5100 HSR<br>V100R005C00<br><br>USG5500 V200R001C00 |

8.3.4 应用场景

8.3.4.1 RBL 过滤

垃圾邮件不仅会浪费网络资源，还会对企业的邮件服务器和个人电脑的系统安全构成威胁。RBL 通过定义一个黑名单列表，在该列表中的 IP 地址对外发布的邮件即为垃圾邮件。同时，提供 RBL 服务的机构会实时更新黑名单列表来保证可以过滤最新的垃圾邮件。

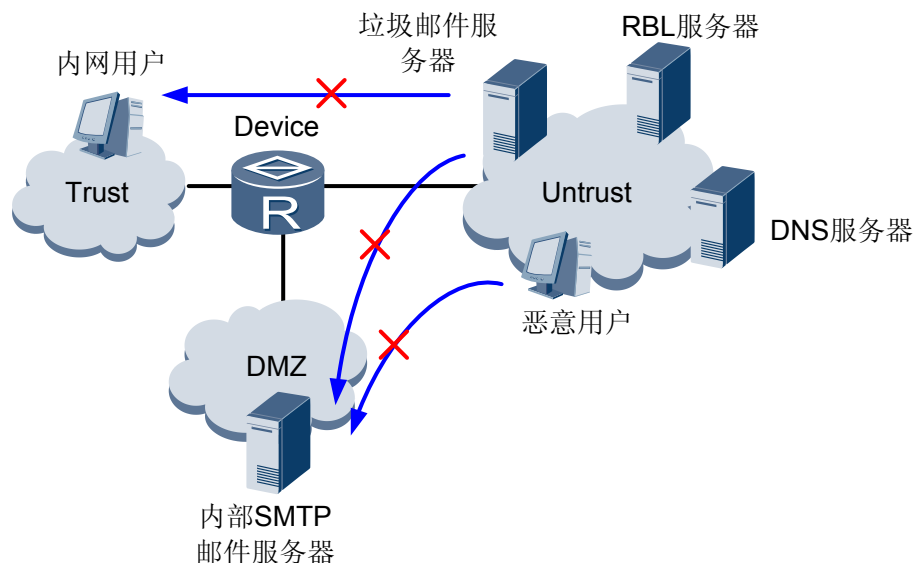
RBL 过滤只支持对 SMTP 协议传输的邮件进行过滤。设备检测到 SMTP 连接请求后，提取 SMTP 连接的源 IP 地址，与第三方组织提供的动态更新的远程实时黑名单以及本地黑白名单进行比对，对匹配黑名单的邮件进行阻断。

 说明

RBL 过滤依赖于远程黑名单的完整性，当远程黑名单不包含 SMTP 请求的源 IP 地址时，无法过滤该邮件。此时，可以通过补充本地 RBL 黑名单来完成过滤。

RBL 的典型应用场景如图 8-1 所示。

图 8-1 RBL 过滤应用场景



当垃圾邮件服务器发送邮件给内部邮件服务器或内网用户时，如果该垃圾邮件的服务器 IP 地址命中了本地黑名单或 RBL 服务器上的远程黑名单，USG 阻断该邮件。

当恶意用户向内部邮件服务器发送邮件时，如果该恶意用户的 IP 地址命中了本地黑名单或 RBL 服务器上的远程黑名单，USG 阻断该邮件。

#### 说明

RBL 过滤只支持对 SMTP 协议传输的邮件进行过滤。

如果垃圾邮件服务器和恶意用户的 IP 地址在传输过程中发生改变，且发送到 USG 进行 RBL 过滤的 SMTP 报文的源 IP 地址不在黑名单内时，邮件还是可以照常转发。

### 8.3.4.2 Webmail 邮件过滤

#### Webmail 介绍

Webmail 是指利用网络浏览器来收发电子邮件的服务或技术。Webmail 不需要借助邮件客户端来收发邮件，只要能上网、有 Webmail 的账号就能通过使用网络浏览器从邮件服务器（电子邮箱）上收发邮件，极大地方便了用户。

Webmail 邮件内容过滤是指当内网用户通过电子邮箱收发邮件时，对邮件地址、主题、正文、附件大小和数量、附件名或附件扩展名进行监控。



#### 注意

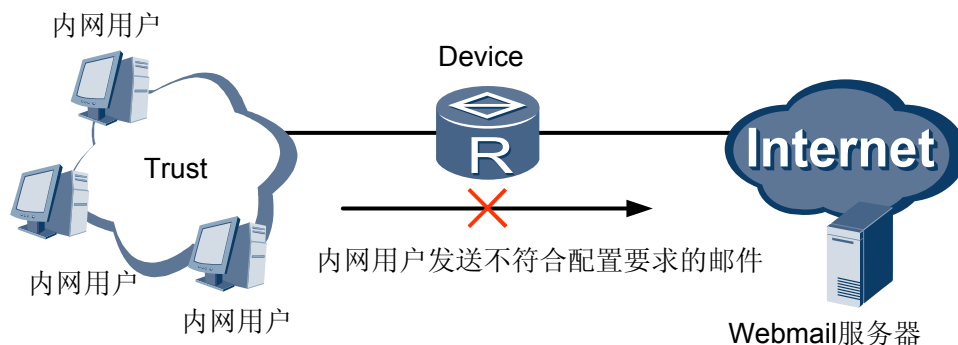
Webmail 过滤不支持内网用户收邮件的过滤。

目前 Webmail 过滤只支持网易邮箱（@126.com 和@163.com）、雅虎中文邮箱（@yahoo.com.cn）。

## 应用场景

Webmail 邮件内容过滤应用场景如图 8-2 所示，邮件的发送需要经过 Device。当内网用户登录到 Web 邮箱中，向外网用户发送邮件时，Device 监控发送的邮件，当发送的邮件不符合配置要求时，阻断邮件。

图 8-2 Webmail 邮件内容过滤应用场景



### 8.3.4.3 SMTP/POP3 邮件过滤

#### SMTP/POP3 邮件介绍

SMTP (Simple Mail Transfer Protocol) 主要负责在 Internet 上对电子邮件进行传输。POP3 (Post Office Protocol) 允许客户端从邮件服务器上取得邮件。

通过 SMTP/POP3 传输的邮件需要在用户 PC 上安装一个客户端，常见的有 Outlook 和 Foxmail。

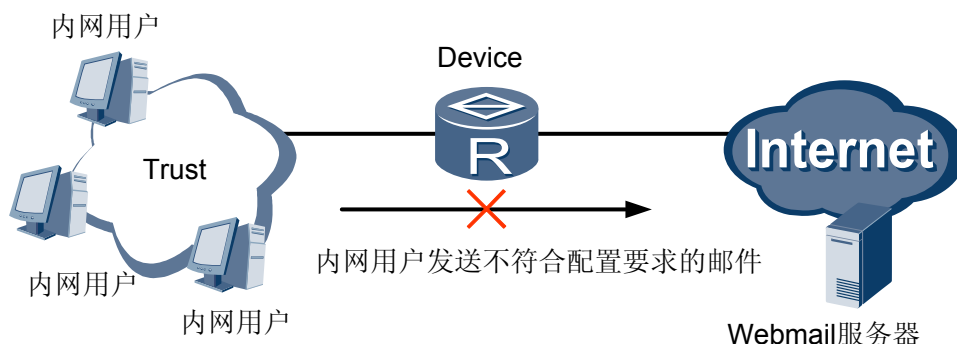
SMTP/POP3 邮件内容过滤是指当内网用户通过邮件客户端收发邮件时，对邮件地址、标题、正文、附件大小和数量、附件名或附件扩展名进行监控。

## 应用场景

SMTP/POP3 邮件内容过滤应用场景如图 8-3 所示，邮件收发需要经过 Device。

- 当用户登录到邮件客户端，从邮件服务器上发送邮件时，Device 监控发送的邮件，当发送的邮件不符合配置要求时，阻断邮件。
- 当用户登录到邮件客户端，从邮件服务器上接收邮件时，Device 监控接收的邮件，当接收的邮件不符合配置要求时，阻断邮件。

图 8-3 SMTP/POP3 邮件内容典型应用



## 8.3.5 原理描述

### 8.3.5.1 电子邮件介绍

#### 电子邮件概念

电子邮件是一种通过网络提供信息交换的通信方式。通过电子邮件，用户可以迅速地和网络上的用户发送或接受信息，这些信息可以是文字、文件、图片等。电子邮件的特点是：使用方便、传输迅速、费用低廉、易于保存、一对多发送，使得电子邮件被广泛地应用。

完整的电子邮件一般包括邮件地址、主题、正文、附件。

邮件地址包括收件人地址（主送人、抄送人和密送人）和发件人地址，格式一般为 host@domainname.type，例如：abc@yahoo.com。

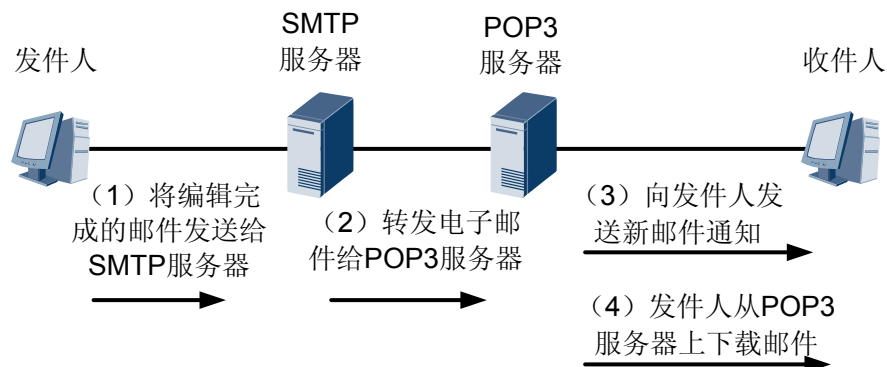
- domainname 为域名的标识符，也就是邮件必须要发送到的邮件目的地的域名。
- host 为在该域名上的邮箱地址。
- type 一般则代表了该域名的性质或地区的代码。例如：com、edu.cn、gov、org。

#### 电子邮件的工作过程

电子邮件的工作过程采用客户端/服务器模式。

图 8-4 为 SMTP/POP3 电子邮件的工作过程。

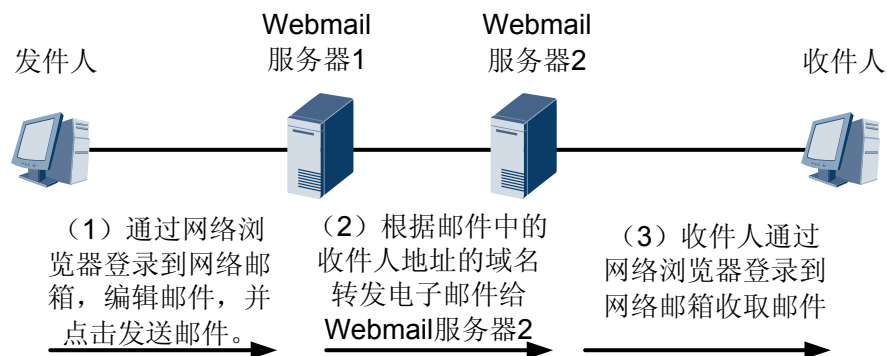
图 8-4 SMTP/POP3 邮件的工作过程



1. 发件人通过邮件客户端程序（如 Outlook，foxmail 等），将编辑好的电子邮件发送给 SMTP 服务器。
2. SMTP 服务器识别收件人的邮件地址，并向管理该邮件地址的 POP3 服务器转发该电子邮件。
3. POP3 服务器将电子邮件存放在接收者的电子信箱内。
4. 收件人通过邮件客户端程序连接到 POP3 服务器后，就会看到 POP3 服务器的通知，从 POP3 服务器上下载邮件。

图 8-5 为 Webmail 电子邮件的工作过程。

图 8-5 Webmail 邮件的工作过程



说明

Webmail 服务器 1 和 Webmail 服务器 2 表示网络中收发邮件的邮箱，Webmail 服务器 1 和 Webmail 服务器 2 可以是不同品牌的 Webmail 服务器，也可以是同一个品牌的 Webmail 服务器。例如，A 通过 163 邮箱，向使用雅虎邮箱的 B 发送邮件时，那么 Webmail 服务器 1 是 163 邮箱，而 Webmail 服务器 2 是雅虎邮箱；当 A 通过 163 邮箱，向使用 163 邮箱的 B 发送邮件时，那么 Webmail 服务器 1 和 Webmail 服务器 2 均为 163 邮箱。

1. 发件人通过网络浏览器登录到网络邮箱（即 Webmail 服务器 1），编辑邮件后，点击发送邮件按钮。
2. Webmail 服务器 1 识别收件人的邮件地址，并向该邮件地址所在区域的 Webmail 服务器 2 转发该电子邮件。
3. 收件人通过网络浏览器登录到网络邮箱（即 Webmail 服务器 2），从网络邮箱上下载邮件。

说明

USG 部署在发件人和 Webmail 服务器 1 之间，过滤发送的邮件。

### 8.3.5.2 RBL 过滤

RBL 过滤的匹配优先级顺序为：本地白名单、本地黑名单、远程实时黑名单。

- 本地白名单

当某个邮件的源地址命中了本地白名单，那么本地黑名单和远程黑名单不再对此源 IP 地址进行匹配，该邮件被直接转发。白名单的设置有利于提高匹配效率。

- 本地黑名单

本地黑名单的优先级比远程黑名单高，当某个邮件的源地址命中了本地黑名单，那么远程黑名单不再对此源 IP 地址进行匹配，该邮件直接阻断。

RBL 过滤依赖于远程黑名单的完整性，当远程黑名单不包含 SMTP 请求的源 IP 地址时，无法过滤该邮件。此时，可以通过补充本地 RBL 黑名单来完成过滤。

#### ● 远程实时黑名单

远程实时黑名单是由第三方组织提供的动态更新的黑名单。第三方组织实时更新和维护 RBL 并通过 RBL 服务器提供查询服务。

#### 📖 说明

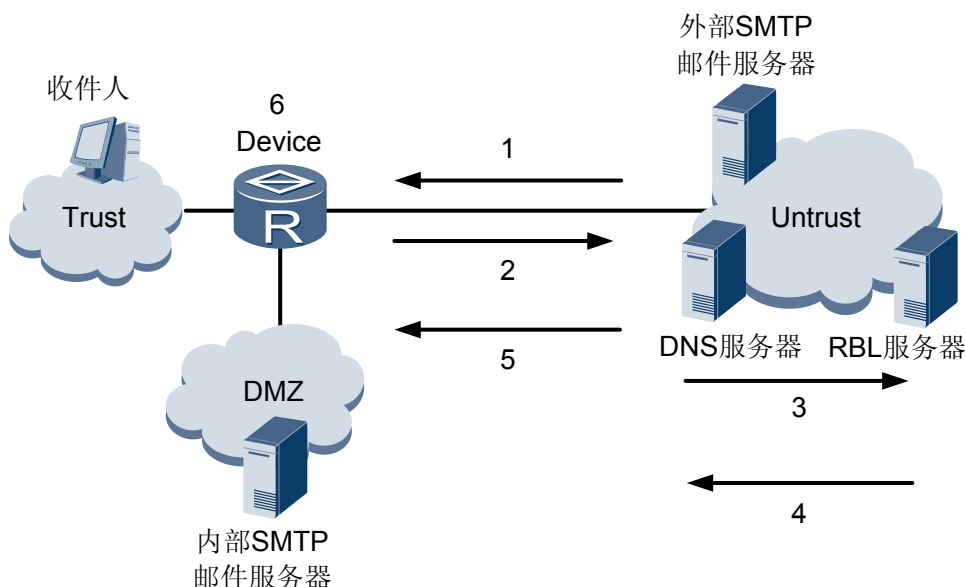
激活码控制 RBL 远程查询服务的有效期限。当激活码过期后，用户只能使用本地黑白名单，如果要使用远程实时黑名单，必须输入或导入有效的激活码。输入或导入激活码的步骤请参见 2。

设备允许用户选择以下两种类型的 RBL 服务器：

- 预定义 RBL 服务器：使用系统预置的 RBL 服务器。
- 自定义 RBL 服务器：用户自己选定需要使用的第三方 RBL 服务器。

远程实时黑名单的处理流程如图 8-6 所示。

图 8-6 远程实时黑名单的处理流程



1. 外部邮件服务器向内网发送 SMTP 请求。
2. Device 收到 SMTP 请求后，提取 SMTP 连接的源 IP 地址并将查询集合和提取的源 IP 地址组合后发送给 DNS 服务器。



#### 注意

使用自定义 RBL 服务器的情况下，查询集合需要在 Device 上进行配置，查询集合由 RBL 服务提供商提供。使用预定义 RBL 服务器不需要配置查询集合。

请使用没有被 DNS 劫持的 DNS 服务器，否则会导致 RBL 失效。DNS 劫持是指服务提供商为引导用户访问其增值站点或合作站点，对 DNS 服务器做修改，在接收到一个不存在结果的 DNS 查询时，总是返回一些特定的 IP 地址，使用户访问到这些增值站点。

3. DNS 服务器根据查询集合向 RBL 服务器转发 RBL 查询请求。
4. RBL 服务器以应答码（例如：127.0.0.1）的形式将查询结果返回给 DNS 服务器。

说明

使用自定义 RBL 服务器的情况下，需要在 Device 上配置应答码。应答码通常是 IP 地址的形式，由 RBL 服务机构提供，例如：中国反垃圾邮件联盟和 Spamhaus。
5. DNS 服务器向 Device 转发应答码。
6. Device 根据应答码判断该 SMTP 连接的邮件是否为垃圾邮件，并进行相应处理。

● 如果 RBL 服务器回复的应答码和设备侧配置的应答码一致，该邮件将被视为垃圾邮件。

● 如果 RBL 服务器不回复应答码或回复的应答码与设备侧配置的应答码不一致，放行邮件。

响应方式

如果 RBL 服务器回复的应答码和 Device 上配置的应答码一致，那么该 SMTP 邮件为垃圾邮件。此时，可以通过设置响应方式来对垃圾邮件进行处理。响应方式如表 8-2 所示。

表 8-2 响应方式

| 响应方式      | 处理方式              |
|-----------|-------------------|
| 告警（Alert） | 正常转发邮件，并发出日志告警信息。 |
| 阻断（Block） | 阻断邮件，并发出日志告警信息。   |

8.3.5.3 邮件内容过滤

当内网用户通过 Webmail 或 SMTP/POP3 客户端收发电子邮件时，邮件内容过滤对邮件地址、主题（标题）、正文、附件大小和数量、附件名或附件扩展名等进行监控，防止数据泄漏或敏感信息传输。

说明

SMTP/POP3 的邮件与附件作为一个整体进行发送和接收，如果附件不符合设备的策略要求，整封邮件会被阻断；与 SMTP/POP3 不同，Webmail 的附件上传与邮件发送是两个相对独立的操作，对于 Webmail 附件上传行为，如果附件不符合设备的策略要求，会对当前附件上传行为进行阻断，但不影响符合策略要求的邮件正文发送。

邮件内容过滤支持的所有过滤配置项以及匹配方式如表 8-3 所示。

表 8-3 邮件内容过滤方式明细表

| 过滤配置项        | 匹配方式        | 引用的公共模式组类型 |
|--------------|-------------|------------|
| 收件人或发件人的邮件地址 | 前缀/后缀/任意/精确 | 邮件地址       |

| 过滤配置项               | 匹配方式          | 引用的公共模式组类型 |
|---------------------|---------------|------------|
| 邮件主题<br>邮件正文<br>附件名 | 任意            | 关键字        |
| 附件扩展名（文件扩展名）        | 精确            | 文件扩展名      |
| 附件大小                | 直接配置，不引用公共模式组 | -          |
| 附件数量                | 直接配置，不引用公共模式组 | -          |

## 8.4 Web 过滤

通过配置 Web 过滤，您可以防止内网用户进行以下操作：访问不合适的网站、查看非法网页内容或在上传/下载不符合要求的文件、通过搜索引擎搜索敏感信息。

### 8.4.1 介绍

#### 定义

Web 过滤包括 URL 过滤、搜索关键字过滤、Web 内容过滤。

- URL 过滤功能对用户的 HTTP 请求进行访问控制，允许或禁止用户访问某些网络资源，可以达到规范上网行为的目的。
- 搜索关键字过滤是指对指定搜索引擎中的关键字进行过滤，控制内网用户的搜索内容，防止用户获得敏感信息。
- Web 内容过滤对用户访问 Web 页面的内容进行控制。

#### 目的

精确管理用户的上网行为。

### 8.4.2 参考标准和协议

与 Web 过滤特性相关的参考标准与协议如下：

- RFC 1945: Hypertext Transfer Protocol -- HTTP/1.0
- RFC 2616: Hypertext Transfer Protocol -- HTTP/1.1

### 8.4.3 可获得性

#### License 支持

URL 过滤预定义分类查询服务功能受 License 控制。当 License 过期后，URL 过滤预定义分类查询服务功能不可用。



其他功能不受 License 控制。

版本支持

| 产品                                             | 最低支持版本   |
|------------------------------------------------|----------|
| Secoway USG2200/5100/5500 和<br>USG2200/5100HSR | V300R001 |

8.4.4 应用场景

应用环境

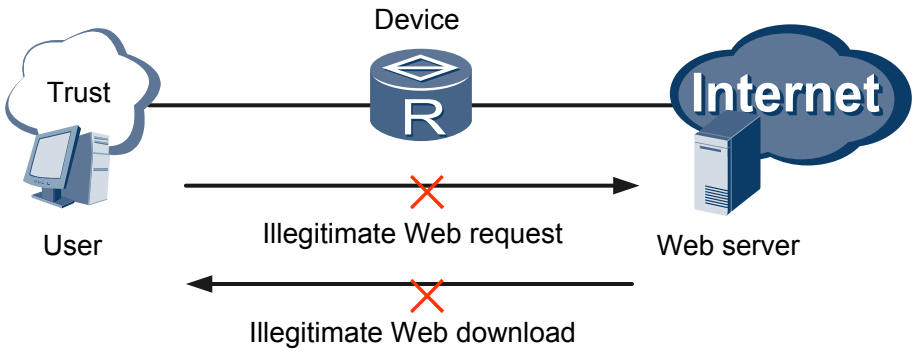
随着互联网应用的迅速发展，计算机网络在经济和生活的各个领域迅速普及，使得信息的获取、共享和传播更加方便，但同时也给企业带来了前所未有的威胁：员工随意不受控地访问非法网站，不仅严重影响工作效率而且威胁企业网络安全。

URL 过滤、搜索关键字过滤、Web 内容过滤分别从 URL、搜索关键字、Web 页面等维度控制内网用户的上网行为。

典型应用

Web 过滤应用场景如图 8-7 所示，当用户向 Web 服务器发起 URL 请求、Web 内容访问或使用搜索引擎时，设备会根据 Web 过滤策略过滤不符合要求的 Web 页面或文件。

图 8-7 Web 过滤典型应用



8.4.5 原理描述

8.4.5.1 URL 过滤

URL 过滤功能对用户的 HTTP 请求进行访问控制，允许或禁止用户访问某些网络资源，可以达到规范上网行为的目的。

URL 简介

Internet 上的每一个网页都具有一个唯一的标识，称为 URL（Uniform Resource Locator）地址。

URL 的一般格式为：“protocol://hostname[:port]/path[?query]”。各参数含义如表 8-4 所示。

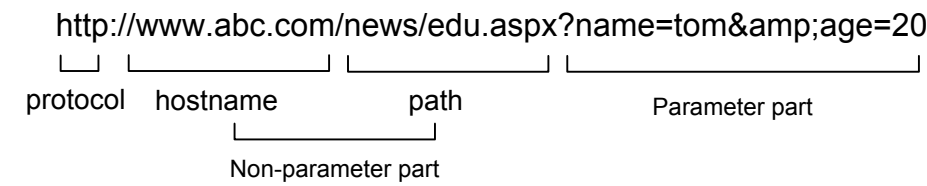
表 8-4 URL 参数解释

| 字段       | 含义                                                                                               |
|----------|--------------------------------------------------------------------------------------------------|
| protocol | 使用的应用协议，最常用的是 HTTP 协议。对于 HTTP 协议，一般可不输入“http://”。<br><b>说明</b><br>URL 过滤功能只支持过滤 HTTP 协议的 URL 请求。 |
| hostname | Web 服务器的域名或者 IP 地址。                                                                              |
| port     | 可选，通信端口。各种应用协议都有默认的端口号，如 HTTP 协议的默认端口为 80。当 Web 服务器采用非标准端口时，URL 中不能省略端口号。                        |
| path     | 由零或多个“/”符号隔开的字符串，一般用来表示主机上的一个目录或文件地址。                                                            |
| ?query   | 可选，用于给动态网页传递参数。                                                                                  |

例如，如图 8-8 所示，对于“http://www.abcd.com/news/education.aspx?name=tom&age=20”：

- www.abcd.com 为 hostname。
- news/education.aspx 为 path。
- ?name=tom&age=20 为参数部分。

图 8-8 URL 格式介绍



URL 过滤匹配顺序

URL 过滤只对 HTTP 协议的 URL 请求进行过滤。

设备通过分析用户的上网请求中的 URL，将 URL 和黑白名单、URL 分类进行比较，根据策略对 HTTP 请求进行放行或阻断。URL 过滤的匹配优先级顺序为：白名单、黑名单、自定义分类、预定义分类。当 URL 与某项匹配后，不会再进行后续的匹配。

- 黑白名单

设备将 HTTP 上网请求的 URL 与黑白名单进行匹配，如果匹配白名单则允许该 HTTP 请求，如果匹配黑名单则阻止该 HTTP 请求。

当上网请求的 URL 与白名单匹配时，不会再进行后续的处理（包括 AV、IPS 等）。设置白名单有利于提高匹配效率。

- URL 自定义分类

URL 自定义分类由用户自行配置和维护。URL 自定义分类将具有相同特征的 URL 进行分类，用户可以根据业务配置策略，允许或拒绝各个分类 URL 的访问。相对于预定义 URL 分类，用户可以使用自定义分类对 URL 进行更为精细化的控制。

- URL 预定义分类，即远程 URL 过滤

URL 预定义分类由安全服务中心提供并维护。URL 预定义分类中定义了各种 URL 的分类，例如教育类 URL、新闻类 URL，用户可以根据业务使用策略，允许或拒绝各个分类网站的访问。使用 URL 预定义分类的情况下，需要建立到安全服务中心的连接。相对于需要用户自行配置的自定义分类，预定义分类已经预先对大量常见的 URL 进行了分类，用户可以根据这些分类轻松地控制内网用户禁止访问哪些类别的 URL、允许访问哪些类别的 URL。

防火墙策略支持对属于特定地址集的用户在特定时间段内进行访问控制。USG 可以利用防火墙策略实现时间和地址的 URL 分类过滤：

- 时间对象

USG 支持在特定时间段内对 HTTP 请求进行 URL 过滤。

- 地址对象

USG 支持对特定地址或者地址组发起的 HTTP 请求进行 URL 过滤。

## URL 匹配方式

URL 过滤的黑白名单、自定义分类通过引用公共模式组来实现 URL 的匹配。例如，自定义分类 uc:game 引用公共模式组 game，game 里面有模式 aaa.com、bbb.net，方式为任意匹配，那么 URL 中包括 aaa.com、bbb.net 的所有 URL 将被列入自定义分类 uc:game。

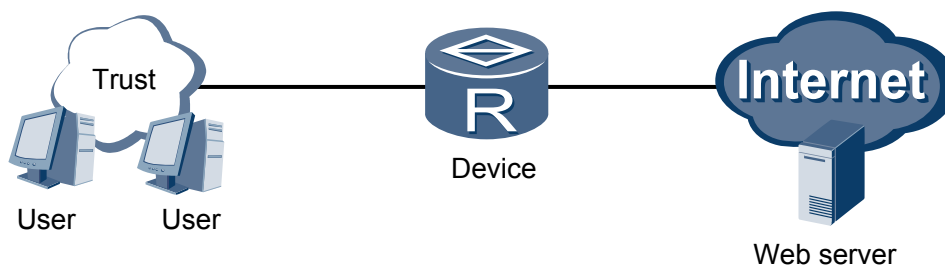
URL 过滤只支持过滤 URL 的非参数部分，绑定的公共模式组类型为 URL，匹配方式为前缀/后缀/任意/精确。

### 8.4.5.2 搜索关键字过滤

搜索关键字过滤是指对指定搜索关键字进行过滤，控制内网用户的搜索内容，防止用户获得敏感信息。

目前支持进行关键字过滤的搜索引擎有 Google、Yahoo、Bing、百度。搜索关键字过滤的典型组网如图 8-9 所示。

图 8-9 搜索关键字过滤组网图



Device 上开启搜索关键字过滤功能后，当用户在搜索引擎上搜索内容时，Device 会对该 HTTP 报文进行解析：

- 如果匹配搜索关键字策略，执行过滤策略（阻断或告警）。
- 如果没有匹配搜索关键字策略，则允许该 HTTP 报文通过。用户可以正常进行搜索。

### 8.4.5.3 Web 内容过滤

Web 内容过滤可以对 HTTP 协议传输的 Web 页面内容进行控制，过滤选项有：

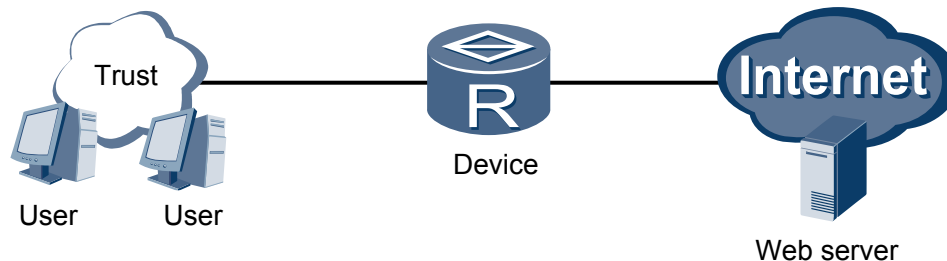
- 网页浏览关键字：过滤网页上的内容。
- HTTP POST 过滤：过滤 HTTP POST 操作。
- HTTP POST 文件过滤：过滤 HTTP POST 文件操作。
- 上传/下载文件名关键字：根据上传或下载的文件名进行过滤。
- 上传/下载文件类型关键字：根据上传或下载的文件类型进行过滤。
- 文件大小过滤：根据上传或下载的文件大小进行过滤。

 说明

HTTP POST 一般用于向网页、BBS、邮件群组 and 数据库发送信息，例如发帖、上传内容等。

Web 内容过滤的典型组网如图 8-10 所示。

图 8-10 Web 内容过滤组网图



下面 Device 上开启上传文件名关键字过滤为例，介绍 Web 内容过滤的处理过程：

1. 用户发起上传附件的操作，触发发送 HTTP 报文。
2. HTTP 报文经过 Device 时，Device 会对其进行过滤：
  - 如果匹配上传文件名关键字策略，执行过滤策略（阻断或告警）。
  - 如果没有匹配上传文件名关键字策略，则允许该 HTTP 报文通过。用户可以正常上传该文件。

## 8.5 FTP 过滤

通过 FTP 过滤，可以对 FTP 操作（上传、下载、删除）、上传/下载的文件名、文件类型、文件大小进行控制。

## 8.5.1 介绍

### 定义

通过 FTP 过滤，可以对 FTP 操作（上传、下载、删除）、上传/下载的文件名、文件类型、文件大小进行控制。

### 目的

对 FTP 上传和下载的文件、FTP 操作进行控制。

## 8.5.2 参考标准和协议

与 FTP 过滤特性相关的参考标准与协议如下：

- RFC 0959: File Transfer Protocol
- RFC 2228: FTP Security Extensions
- RFC 2640: Internationalization of the File Transfer Protocol
- RFC 3659: Extensions to FTP

## 8.5.3 可获得性

### License 支持

本功能不受 License 控制。

### 版本支持

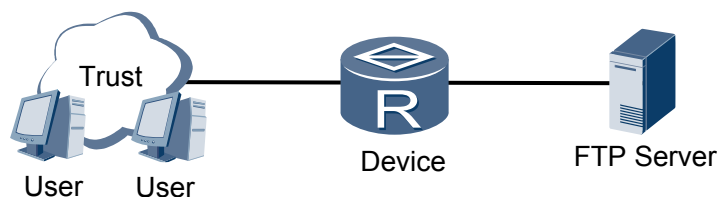
| 产品                                             | 最低支持版本   |
|------------------------------------------------|----------|
| Secoway USG2200/5100/5500 和<br>USG2200/5100HSR | V300R001 |

## 8.5.4 应用场景

FTP 主要功能是向用户提供本地和远程主机之间的文件传输。在进行版本升级、日志下载、文件传输和配置保存等业务操作时，FTP 功能被广泛地使用。如果 FTP 操作不受控，会给网络带来较多不可控的威胁。通过 FTP 过滤，可以对 FTP 操作（上传、下载、删除）、上传/下载的文件名、文件类型、文件大小进行控制。

FTP 过滤的典型应用如图 8-11 所示。

图 8-11 FTP 过滤应用场景



Device 上开启 FTP 过滤功能后，当用户进行 FTP 的 upload、download 或 delete 操作时，Device 会对 FTP 操作进行解析：

- 如果用户的 FTP 操作匹配 FTP 过滤策略，执行过滤策略（阻断或告警）。
- 如果用户的 FTP 操作没有匹配 FTP 过滤策略，则允许该 FTP 操作。

## 8.6 应用控制（DPI）

DPI（Deep Packet Inspection）提供对 TCP/UDP 传输层以上具体应用协议的识别功能，基于该功能，支持对网络流量按协议进行细分统计，对某些具体协议流量进行流量阻断或放行。

### 8.6.1 介绍

#### 定义

应用控制是一种通过 DPI 技术对网络中的报文进行识别控制的应用功能。DPI 即深度报文检测，通过对报文的应用层数据进行内容检测，分析报文或数据流在 IP 和 TCP/UDP 层以上的应用业务类型，从而进行流量识别和流量管理以及其他增值服务。

#### 目的

DPI 通过对数据流进行深度检测，可以识别应用层业务，并对指定类型的数据流量进行控制。通过分析 USG 收到的数据包并和 DPI 特征库进行比对，对 P2P、IM、VoIP 等类型的网络数据流量进行分类，并对不同类型的协议进行相应的控制。

### 8.6.2 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                 | 最低支持版本                                                                       |
|----------------------------------------------------|------------------------------------------------------------------------------|
| Secoway USG2200/5100/5500、<br>USG2200/5100 BSR/HSR | USG2200/5100、USG2200/5100 BSR/<br>HSR V100R005C00<br><br>USG5500 V200R001C00 |

8.6.3 特性增强

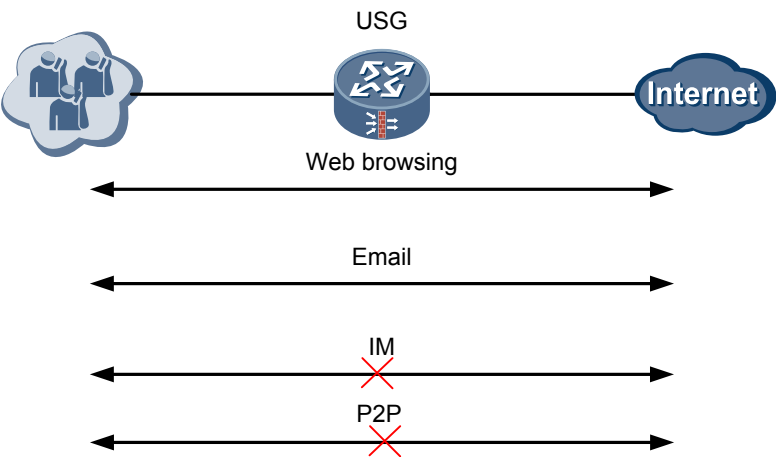
| 版本       | 特性增强                                                      |
|----------|-----------------------------------------------------------|
| V300R001 | 应用控制（DPI）功能改进配置方法，能将应用控制策略（DPI 策略）应用在域间或域内，被域间或者域内策略引用生效。 |

8.6.4 应用场景

典型应用

在实际应用中，可以对网络中检测到的各种类型的应用协议的流量进行阻断和放行，确保信息安全，保证正常业务的运行。如图 8-12 所示，对于一般的网络浏览和邮件流量予以放行，对某些不允许使用的 IM 类型应用和影响带宽的 P2P 下载应用可以进行阻断。

图 8-12 DPI 应用场景

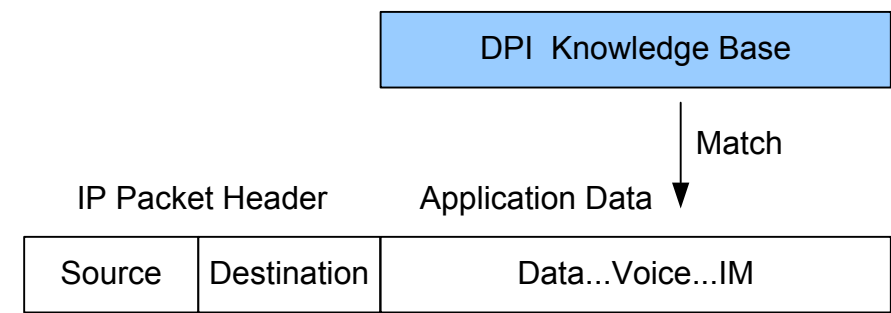


8.6.5 原理描述

8.6.5.1 DPI 工作原理

原理概述

图 8-13 原理概述示意图



如图 8-13 所示，设备根据 DPI 知识库或者用户自定义的应用协议，对报文进行匹配，并将结果用于对报文的控制。

与传统的只对报文中的 IP 报头进行检查的简单机制不同，DPI 设备将网络上的数据报文根据五元组（源地址、源端口号、目的地址、目的端口号、传输层协议类型）分为各个应用数据流，并通过识别技术对应用数据流中的特定的数据报文进行探测，从而确定应用数据流对应的应用。

特征识别

特征识别是 DPI 技术的核心。对于不同类型的应用，所依赖的应用协议各不相同，而不同的应用协议具有各自的特征，DPI 知识库即各种不同协议特征的集合。

对于不同的协议，特征的表现形式也不同，可能是特定的端口、特定的字符串和特定的比特序列。

协议的特征不仅在单个报文中体现，某些协议报文的特征是分布在多个报文中的，需要对多个报文进行采集分析，对报文在内容或报文长度的变化规律或趋势上进行分析，才能够判断协议类型。

8.6.5.2 关联协议识别

原理概述

某些数据流，控制通道和数据通道是分开的（如 FTP、SIP、H.323 等），一般的处理方式是对这些会话分别进行识别，这可能导致分析速度较慢或影响对数据通道的识别。

关联协议识别，是将来自同一源地址的控制通道和后续的数据通道进行关联的技术，从而达到快速识别的目的。

关联过程

关联协议识别可以将同一应用协议的控制通道流和数据通道流关联起来。通过对控制通道流的分析，可以分析出通讯双方将要在哪个通道上建立何种类型的数据流，并在协议识别时将控制通道流和该控制通道流协商出来的数据通道流关联起来。



DPI 在对控制通道流进行深度解析时，提取出其中协商的数据通道流的源三元组和目的三元组信息，并加入关联表。在后续识别过程中，可以通过该关联表项对数据通道流快速识别。

### 8.6.5.3 全包检测

一般情况下，DPI 对报文的特定字节进行协议特征检测，对于多数已知协议可以做到成功检测。

某些协议的报文是承载在其他协议上进行传输的，如 MSN 信息可能通过 HTTP 协议承载、VoIP 信息可能通过 IM 类型协议承载。全包检测即对这种类型的报文的所有字节进行检测，以提高对协议的识别准确率。

USG 支持对指定的协议启用全包检测功能。

# 9 UTM (USG2110-X/2100、USG2100BSR/HSR)

## 关于本章

### 9.1 反病毒 (AV)

AV (Anti-Virus) 特性可以检测包括病毒、间谍软件、蠕虫、木马病毒、后门和其它恶意代码的各种威胁，并根据策略进行响应，有效保护网络。

### 9.2 协议识别

随着新型网络协议以及软件的出现，传统的基于标准端口来识别报文协议类型的方法的局限性越来越大。USG 不仅支持基于标准端口的协议识别，也支持基于非标准端口的协议识别功能。

### 9.3 入侵防御 (IPS)

入侵防御 (Intrusion Prevention System) 能够有效防御来自应用层的攻击，例如缓冲区溢出攻击、木马、后门攻击、蠕虫等。

### 9.4 协议检测

通过对应用层协议的检测，可以发现基于协议异常的溢出攻击等攻击，同时，通过对这些协议的详细解析可以提高对病毒、蠕虫、木马等有害数据的检测速度及准确度。

### 9.5 反垃圾邮件 (RBL 过滤)

RBL (Real-time Blackhole List) 过滤可以通过本地黑白名单和远程黑名单，有效过滤垃圾邮件，保护邮件服务器。

### 9.6 URL 过滤

URL 过滤对用户的 URL 请求进行访问控制，允许或禁止用户访问某些网络资源，可以达到规范上网行为的目的。

### 9.7 应用控制 (DPI)

DPI (Deep Packet Inspection) 提供对 TCP/UDP 传输层以上具体应用协议的识别功能，基于该功能，支持对网络流量按协议进行细分统计，对某些具体协议流量进行流量阻断或放行。

## 9.1 反病毒 (AV)

AV (Anti-Virus) 特性可以检测包括病毒、间谍软件、蠕虫、木马病毒、后门和其它恶意代码的各种威胁，并根据策略进行响应，有效保护网络。

### 9.1.1 介绍

#### 定义

AV 是一种安全机制，它通过比对扫描的文件内容与病毒特征来检测病毒，并通过一定的响应方式对感染病毒的文件进行处理。

#### 目的

保护系统和网络免受多种动态威胁的侵扰。AV 可以防范的威胁包括：病毒、间谍软件、蠕虫、木马、后门及恶意代码等。

### 9.1.2 可获得性

#### License 支持

只有获取了支持 AV 功能的 License，才能使用 AV 功能。

#### 版本支持

| 产品                                     | 最低支持版本      |
|----------------------------------------|-------------|
| Secoway USG2110-X/2100 和<br>USG2100HSR | V100R005C00 |

### 9.1.3 原理描述



#### 注意

使用 AV 功能前，请确认已购买支持 AV 升级服务的 License，并在 USG 上激活。

#### 原理

病毒是一种可执行代码，可感染或附在其他可执行代码上。有些病毒可直接导致删除文件或锁住系统；有些病毒感染其他文件后，使这些文件在传播时产生大量的伪造数据来耗尽受感染主机或网络的资源。

USG 将接收到需要进行病毒扫描的文件的内容与病毒特征库进行比较，判断扫描内容中是否含匹配病毒特征的信息。如果检测到病毒，USG 根据 AV 策略进行响应，对含病毒的内容进行删除或告警。如果没有检测到病毒，则 USG 放行该文件。

## 支持处理的文件

USG 能够对使用以下协议进行传输的文件进行病毒扫描和相应处理：

- HTTP（Hypertext Transfer Protocol），超文本传输协议
- SMTP（Simple Mail Transfer Protocol），简单邮件传输协议
- POP3（Post Office Protocol 3），邮局协议版本 3

## 特殊文件的处理策略

对于超大文件、密码保护文件、多层压缩文件，HTTP 传输文件和 SMTP/POP3 传输文件的处理策略有差异，具体处理策略如表 9-1 所示。

表 9-1 特殊文件处理策略

| 协议类型      | 处理策略                                                                                                                                                    |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| HTTP      | <ul style="list-style-type: none"><li>● 允许（permit）<br/>USG 不对文件进行处理，记录日志。</li><li>● 拒绝（deny）<br/>USG 断开与 HTTP 服务器的连接并阻断文件，向客户端推送 Web 页面并记录日志。</li></ul> |
| SMTP/POP3 | <ul style="list-style-type: none"><li>● 允许（permit）<br/>USG 在邮件正文中添加宣告并记录日志。</li><li>● 拒绝（deny）<br/>USG 删除邮件中的附件，在邮件正文中添加宣告并记录日志。</li></ul>              |

## 病毒文件处理策略

当文件检测出病毒时，HTTP 传输文件和 SMTP/POP3 传输文件的病毒文件处理策略有差异，具体差异请参见表 9-2。

表 9-2 病毒文件处理策略

| 协议类型 | 处理策略                                                                                                                                                          |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| HTTP | <ul style="list-style-type: none"><li>● 告警（alert）<br/>USG 只产生日志，不对文件进行处理就发送出去。</li><li>● 阻断（block）<br/>USG 断开与 HTTP 服务器的连接并阻断文件，向客户端推送 Web 页面并产生日志。</li></ul> |

| 协议类型      | 处理策略                                                                                                                                                                                                             |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SMTP/POP3 | <ul style="list-style-type: none"><li>● 告警（alert）<br/>USG 只产生日志，不对邮件进行处理就发送出去。</li><li>● 添加宣告（declare）<br/>USG 在邮件正文中添加宣告并产生日志。</li><li>● 删除附件（attachment-detach）<br/>USG 删除邮件中附件里的内容，在邮件正文中添加宣告并产生日志。</li></ul> |

## 9.2 协议识别

随着新型网络协议以及软件的出现，传统的基于标准端口来识别报文协议类型的方法的局限性越来越大。USG 不仅支持基于标准端口的协议识别，也支持基于非标准端口的协议识别功能。

### 9.2.1 介绍

#### 定义

协议识别支持识别使用非标准端口的 FTP、HTTP、IMAP、POP3、SMTP 协议。

#### 目的

协议识别功能可防止使用非标准端口的应用服务报文的漏报和误报。

例如：一般情况下，HTTP 服务使用 80 端口，当 HTTP 服务使用非标准端口如 8000 端口时，协议识别可以对 HTTP 协议进行识别，减少应用层攻击的漏报和误报。

### 9.2.2 可获得性

#### License 支持

只有获取了支持 IPS 功能的 License，才能使用协议识别功能。

#### 版本支持

| 产品                                  | 最低支持版本      |
|-------------------------------------|-------------|
| Secoway USG2110-X/2100 和 USG2100HSR | V100R005C00 |

## 9.2.3 原理描述

### 背景信息

随着使用动态协商端口的网络协议的出现，传统的基于默认标准端口来识别报文协议类型的方法的局限性越来越大。USG 的协议识别功能解决了协议识别过程中对使用非标准端口的应用服务报文的漏报和误报。

USG 的协议识别功能支持基于非标准端口的协议识别功能。

基于非标准端口的协议识别可以解决以下常见网络问题：

- 仅基于标准端口的协议识别无法识别在协议运行过程中动态协商端口的协议。
- 仅基于标准端口的协议识别对运行在非标准端口上的服务进行攻击检测时，由于无法判断报文的协议类型而使攻击行为躲避检测。

### 工作原理

USG 采用如下步骤进行协议识别：

1. 会话建立时，USG 截取报文。
2. USG 根据截取的报文的特征自动识别报文的协议类型，保证协议识别的实时性。
3. 根据 USG 的协议验证规则验证协议识别结果正确性，保证协议识别的正确性。
4. 将协议识别结果保存到固化表，为后续的协议识别积累报文特征。

 说明

固化表中保存的识别结果包括以下内容：

- 服务器 IP 地址，如 Web 服务器地址等。
- 服务器开启服务的非标准端口号。
- 使用非标准端口进行会话的协议类型，包括 HTTP、FTP、IMAP、SMTP 和 POP3。

## 9.3 入侵防御（IPS）

入侵防御（Intrusion Prevention System）能够有效防御来自应用层的攻击，例如缓冲区溢出攻击、木马、后门攻击、蠕虫等。

### 9.3.1 介绍

#### 定义

IPS 是一种安全机制，它通过监控或者分析系统事件，检测入侵，并通过一定的响应方式，实时地中止入侵行为。

#### 目的

保护系统或系统资源不受未经授权的访问。IPS 可以防范的威胁包括：蠕虫、病毒、广告软件、来自 Internet 的无授权访问、内部用户的安全违规等。

## 9.3.2 可获得性

### License 支持

只有获取了支持 IPS 功能的 License，才能使用 IPS 功能。

### 版本支持

| 产品                                  | 最低支持版本      |
|-------------------------------------|-------------|
| Secoway USG2110-X/2100 和 USG2100HSR | V100R005C00 |

## 9.3.3 原理描述

### 产生背景

随着攻击手段和工具的不断涌现、攻击技术的日趋成熟，传统防火墙的网络层攻击检测已无法满足安全需要。在此背景下，出现了 IPS 技术。

IPS 设备与传统防火墙以及 IDS（Intrusion Detection System）设备相比主要有以下不同：

- 传统防火墙很难对基于应用层的攻击进行预防和阻止。IPS 设备能够有效防御应用层攻击，如：缓冲区溢出攻击、木马、后门攻击、蠕虫等。
- IDS 设备只检测和告警。IPS 设备不仅能够检测入侵的发生，而且能通过一定的响应方式，实时地中止入侵行为的发生和发展，实时地保护信息系统。

#### 说明

USG 有两种运行模式：防火墙模式和 UTM 模式。当 USG 运行模式为防火墙时，只能检测网络层的攻击行为。当 USG 运行模式为 UTM 时，可以配置 IPS 功能检测应用层的攻击行为，此时 USG 能够同时检测网络层和应用层的攻击行为。

### IPS 处理流程

IPS 的基本处理流程如下：

1. 重组应用数据  
USG 支持对 IP 分片报文重组以及 TCP 流重组，确保了应用数据的连续性，防止躲避 IPS 检测的攻击行为。
2. 协议识别  
与传统的根据端口识别协议不同，USG 能根据报文内容识别多种常见应用层协议，并对这些数据进行检测，提高了攻击行为的检测率。
3. 匹配签名  
USG 将报文内容与签名进行比较，只有匹配签名的报文才会进行下一步 IPS 处理。
4. 完成检测后，USG 根据用户配置的策略和响应方式对匹配到 IPS 签名的报文进行处理。

#### 说明

关于 IPS 签名、策略和响应方式的概念请参见下文中的 IPS 签名、策略定制和响应方式。

## IPS 签名

IPS 签名用来描述网络中存在的攻击行为的特征，USG 通过将报文内容和 IPS 签名进行比较来检测和防范攻击。

USG 的 IPS 签名分为两类：

- 预定义签名  
USG 中预先定义的签名。用户购买了带 IPS 升级功能的 License 后即可获得包含预定义签名的签名库，并且能够不断地从安全服务中心获取新的 IPS 版本来更新签名库。
- 自定义签名  
用户根据网络流量特点对特定的入侵行为自行定义的签名。

## 策略定制

用户可通过引用签名和配置覆盖签名来定制 IPS 策略，满足不同的安全应用场景。

完成 IPS 策略配置后，需要将 IPS 策略应用在指定的域间才能使 IPS 策略生效。

IPS 策略经过编译后才能生效，USG 支持手动提交配置来编译 IPS 策略。另外，升级 IPS 版本也会触发 IPS 策略的编译，而且与手动提交配置的编译效果相同。

## 攻击响应方式

一个签名包含一种攻击特征，当报文命中签名时，USG 将该报文识别为攻击报文，然后按照签名的攻击响应方式处理该报文。

IPS 攻击响应方式如表 9-3 所示。

表 9-3 攻击响应方式

| 工作模式 | 处理策略 | 实际动作               |
|------|------|--------------------|
| 告警模式 | 告警   | USG 不对报文进行处理，记录日志。 |
|      | 阻断   |                    |
| 防护模式 | 阻断   | USG 阻断报文，记录日志。     |
|      | 告警   | USG 不对报文进行处理，记录日志。 |

### 说明

当报文命中多个签名，对该报文的响应方式如下：

- 如果这些签名的响应方式都为 **Alert** 时，响应方式为告警。
- 如果这些签名中至少有一个签名的响应方式为 **Block** 时，响应方式为阻断。

## 9.4 协议检测

通过对应用层协议的检测，可以发现基于协议异常的溢出攻击等攻击，同时，通过对这些协议的详细解析可以提高对病毒、蠕虫、木马等有害数据的检测速度及准确度。



## 9.4.1 介绍

### 定义

协议检测是对报文应用层载荷数据进行分析检测。

### 目的

协议检测功能可防止隐藏在数据载荷部分的数据攻击。

## 9.4.2 参考协议和标准

如果要更详细了解 HTTP 协议的原理，请参考以下文档：

- RFC1945: Hypertext Transfer Protocol HTTP/1.0
- RFC2616: Hypertext Transfer Protocol HTTP/1.1
- RFC3986: Uniform Resource Identifier (URI) Generic Syntax

如果要更详细了解 FTP 协议的原理，请参考 RFC0959: File Transfer Protocol

如果要更详细了解 SMTP 协议的原理，请参考以下文档：

- RFC2821: Simple Mail Transfer Protocol
- RFC2822: Internet Message Format
- RFC2045: Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies
- RFC2046: Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types
- RFC2047: MIME (Multipurpose Internet Mail Extensions) Part Three: Message Header Extensions for Non-ASCII Text

如果要更详细了解 IMAP 协议的原理，请参考 RFC2060: INTERNET MESSAGE ACCESS PROTOCOL - VERSION 4rev1。

如果要更详细了解 POP3 协议的原理，请参考 RFC1939: Post Office Protocol - Version 3。

如果要更详细了解 DNS 协议的原理，请参考以下文档：

- RFC1034: DOMAIN NAMES - CONCEPTS AND FACILITIES
- RFC1035: DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION
- RFC1101: DNS Encoding of Network Names and Other Types

## 9.4.3 可获得性

### License 支持

只有获取了支持 IPS 功能的 License，才能使用协议检测功能。

版本支持

| 产品                                  | 最低支持版本      |
|-------------------------------------|-------------|
| Secoway USG2110-X/2100 和 USG2100HSR | V100R005C00 |

9.4.4 原理描述

9.4.4.1 检测类型和响应策略

介绍协议检测特性的检测类型和响应策略。

检测类型

USG 支持对 HTTP、FTP、POP3、IMAP、SMTP 和 DNS 会话进行协议检测。协议检测过程中，USG 主要支持以下类型的攻击检测：

- 溢出攻击检测  
应用服务器一般通过解析请求报文信息头中的不同字段来作出响应的。当请求报文超大时，请求报文可能会对应用服务器造成溢出攻击。USG 通过溢出攻击检测来检测该类型的攻击，并通过限制指定字段的长度阈值来防止此类攻击。
- 危险行为检测及控制  
主要用来检测对应用服务器进行的危险操作，如删除服务器上的重要文件等行为。USG 的行为检测通常包括以下类型：
  - 限制危险命令，如禁止 Delete 命令、Put 命令。
  - 限制一个会话中的用户登录次数。
  - 限制请求报文的协议版本号。
  - 限制 HTTP Chunk 传输块长度。
  - 限制异常回复报文。

响应策略

USG 进行协议检测后，对检测到的异常报文按照如表 9-4 所示的响应策略进行处理。

表 9-4 响应策略

| 检测结果   | 响应策略  | 报文处理方式              |
|--------|-------|---------------------|
| 检测结果异常 | alert | 报文照常转发，发送告警信息到信息中心  |
|        | block | 断开当前会话，并发送告警信息到信息中心 |

9.4.4.2 HTTP 协议检测

介绍 HTTP 协议检测的具体内容、工作原理、及其参考信息。通过 HTTP 协议检测，USG 可以实现对 WWW 服务器和客户端的保护。

## HTTP 的溢出检测

USG 进行 HTTP 溢出检测时，对以下内容进行检测，从而避免利用 HTTP 报文，对 WWW 服务器的溢出攻击：

- Authorization 字段的长度
- Content-type 字段的长度
- Cookie 字段的长度
- Header-line 字段的长度
- Host 字段的长度
- Referer 字段的长度
- 请求行的长度
- User-agent 字段的长度

## HTTP 的危险行为检测和控制

- 限制 HTTP 报文使用的命令，可以选择是否禁止使用 GET、POST、HEAD、DELETE、OPTIONS、PUT、TRACE、CONNECT 及 RFC 未定义的命令。
- 限制 HTTP 报文使用的版本，可以选择是否禁止使用 0.9、1.0、1.1 或其他未知版本的 HTTP 报文。
- 当 HTTP 报文采用 Chunk 方式传输时，限制 Chunk 块的大小。某些攻击报文会包装成大 Chunk 块对 WWW 服务器发起攻击，使 Web 服务器的缓冲区溢出。
- 限制接收异常的 HTTP 回复报文。某些攻击报文会伪装成 HTTP 回复报文对 HTTP 客户端发起攻击。

## CGI 攻击检测

USG 除可以进行溢出检测和异常检测外，还可以进行 CGI（Common Gateway Interface）攻击检测。

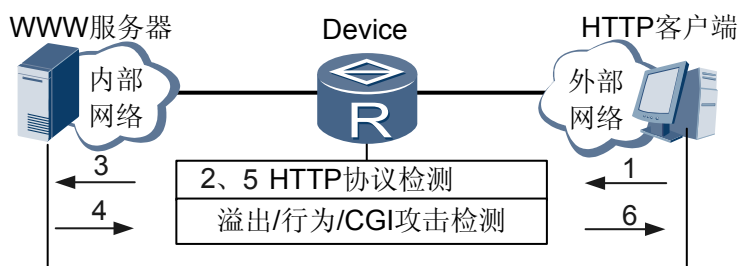
由于不同的 Web 服务器会对 URI（Uniform Resource Identifier）按照不同方式进行解码，非法用户会利用该特点逃避检测。USG 使用 URI 标准化来避免对上述非法用户的漏检。USG 的 CGI 攻击检测按照下面步骤来实现：

1. USG 根据用户设置的 Web 服务器的编码类型（包括 Apache 和 IIS）对 URI 进行不同解码。
2. USG 对解码后的数据进行特征检测，从而防止 CGI 攻击。

## 工作原理

USG 的 HTTP 协议检测工作过程如图 9-1 所示。

图 9-1 HTTP 协议检测工作原理



下面以 HTTP 客户端访问 WWW 服务器为例说明：

1. HTTP 客户端向 WWW 服务器发送 HTTP 连接请求。
2. USG 截取该请求，并进行 HTTP 协议检测。
3. USG 根据检测结果，处理该请求。
  - 如果未检测到异常报文，则将请求报文转发到 WWW 服务器。
  - 如果检测到异常报文，则根据用户配置的响应策略对该请求报文进行处理。
4. HTTP 服务器收到 HTTP 请求后发出响应报文。
5. USG 截取该响应报文，并进行 HTTP 协议检测。
6. USG 根据检测结果，处理该响应。
  - 如果未检测到异常报文，则将响应报文转发到 HTTP 客户端。
  - 如果检测到异常报文，则根据用户配置的响应策略对该响应报文进行处理。

### 9.4.4.3 FTP 协议检测

通过 FTP 协议检测，USG 可以实现对 FTP 服务器的保护。

#### FTP 的溢出检测

USG 进行 FTP 溢出检测时，对以下内容进行检测，从而防止利用 FTP 报文对 FTP 服务器缓冲区的溢出攻击：

- FTP 命令行的长度
- FTP 用户的用户名长度
- FTP 用户的密码长度
- SITE 字段的长度

#### FTP 的危险行为检测和控制

- 限制 FTP 报文使用的命令，可以选择是否禁止使用 appe、petr、cdup、rmd、dele、rnfr、stor、rnto、help、site、mkd、stou 命令。
- 限制一个 FTP 会话中允许 FTP 用户连续登录失败的次数。
- 通过启用 FTP 跳板检测，检查并控制通过一个 FTP 服务器对另一个 FTP 服务器进行的攻击。

#### 隐藏 FTP 服务器信息

USG 支持在发送 FTP 报文时隐藏 FTP 服务器信息。

USG 可以通过在以下消息中隐藏 FTP 服务器信息，来实现对 FTP 服务器的保护：

- 隐藏欢迎消息  
当建立 FTP 会话连接后，FTP 服务器会发送欢迎消息，欢迎消息中包含了 FTP 服务器的名称和版本信息等。
- 隐藏 SYST 命令回复信息  
当建立 FTP 会话连接后，FTP 服务器会发送 SYST 命令回复消息。SYST 命令回复消息中包含了 FTP 服务器的名称和版本信息等。

## 工作原理

USG 的 FTP 协议检测工作过程如图 9-2 所示。

图 9-2 FTP 协议检测工作原理

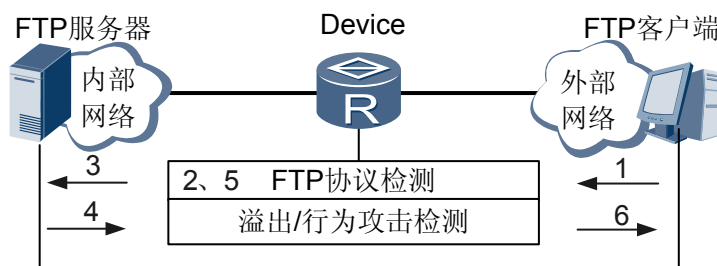


图 9-2 以 FTP 客户端访问 FTP 服务器为例说明了 USG 进行 FTP 协议检测的工作过程：

1. FTP 客户端向 FTP 服务器发送文件传输请求。
2. USG 截取该请求。
3. USG 根据检测结果，处理该请求。
  - 如果未检测到异常报文，则开启 FTP 客户端和服务器协商的端口。
  - 如果检测到异常报文，则根据用户配置的响应策略对该请求报文进行处理。
4. FTP 服务器收到请求后，向 FTP 客户端发送请求的数据。
5. USG 截取该数据。
6. USG 根据检测结果，处理该数据。
  - 如果未检测到异常报文，则将数据转发到 FTP 客户端。
  - 如果检测到异常报文，则根据用户配置的响应策略对该数据进行处理。

### 9.4.4.4 SMTP 协议检测

介绍 SMTP 协议检测的内容及其工作原理。通过 SMTP 协议检测，USG 可以实现对邮件服务器或客户端的保护。

#### SMTP 的溢出检测

USG 进行 SMTP 溢出检测时，对以下内容进行检测：

- 命令行的长度
- 应答行的长度
- 发件人邮箱地址的长度
- 收件人邮箱地址的长度

#### SMTP 的危险行为检测和控制

- 限制 SMTP 报文使用的命令，可以选择是否禁止使用 vrfy 命令、expn 命令、turn 命令。
- 限制一个 SMTP 会话中包含的错误请求次数。

- 限制一个 SMTP 会话进行重新建立的次数。

## 工作原理

USG 的 SMTP 协议检测工作过程如图 9-3 所示。

图 9-3 SMTP 协议检测工作原理

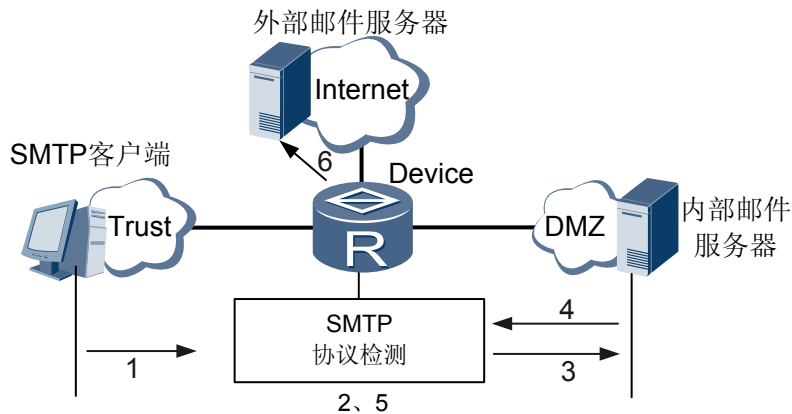


图 9-3 说明了 SMTP 客户端向外发送邮件时，USG 进行 SMTP 协议检测的过程：

1. SMTP 客户端向内部的 SMTP 服务器发送邮件。
2. USG 截取该邮件，并进行协议检测。
3. USG 根据检测结果，处理该邮件。
  - 如果未检测到异常报文，则将该邮件发送到内部邮件服务器。
  - 如果检测到异常报文，则用户配置的响应策略对该请求报文进行处理。
4. 内部邮件服务器将收到的邮件发送到外部邮件服务器。
5. USG 截取该邮件，并进行协议检测。
6. USG 根据检测结果，处理该邮件。
  - 如果未检测到异常报文，则将该邮件发送到外部邮件服务器。
  - 如果检测到异常报文，则根据用户配置的响应策略对该请求报文进行处理。

### 9.4.4.5 IMAP/POP3 协议检测

当 IMAP 或 POP3 客户端从内部邮件服务器上下载邮件时，USG 对 IMAP 或 POP3 报文进行协议检测。

## IMAP/POP3 的溢出检测

USG 进行 IMAP/POP3 溢出检测时，对以下内容进行检测，从而防止利用 IMAP/POP3 报文对 IMAP/POP3 客户端的溢出攻击：

- IMAP/POP3 命令行的长度
- IMAP/POP3 用户的用户名长度
- IMAP/POP3 用户的密码长度

- IMAP 报文中 TAG 字段的长度
- IMAP 报文中邮箱名的长度
- POP3 中的 APOP 命令字段的长度

POP3 中的 APOP 命令用来加密 POP3 用户的用户名和密码。通过 APOP 命令可以保护用户信息，但是可能会造成该命令行过长从而使邮件服务器缓冲区溢出。

## IMAP/POP3 的危险行为检测和控制

IMAP/POP3 的危险行为检测和控制，通过限制一个 IMAP/POP3 会话中 IMAP/POP3 用户连续登录失败的次数来实现。

### 工作原理

USG 的 IMAP/POP3 协议检测工作过程如图 9-4 所示。

图 9-4 IMAP/POP3 协议检测工作原理

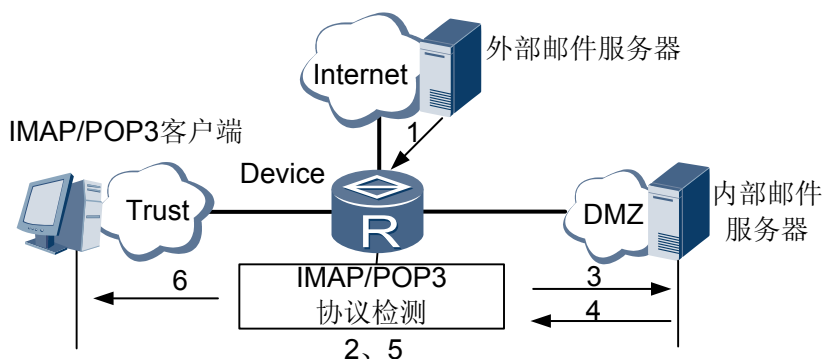


图 9-4 以内部网络中的 IMAP/POP3 客户端从外部邮件服务器下载邮件为例说明 IMAP/POP3 协议检测的过程：

1. 外部邮件服务器向内部邮件服务器发送邮件。
2. USG 截取该邮件，并进行协议检测。
3. USG 根据检测结果，处理该邮件。
  - 如果未检测到异常报文，则将该邮件发送到内部邮件服务器。
  - 如果检测到异常报文，则根据用户配置的响应策略对该请求报文进行处理。
4. 内部邮件服务器向 IMAP/POP3 客户端发送邮件。
5. USG 截取该邮件，并进行协议检测。
6. USG 根据检测结果，处理该邮件。
  - 如果未检测到异常报文，则将该邮件发送到 IMAP/POP3 客户端。
  - 如果检测到异常报文，则根据用户配置的响应策略对该请求报文进行处理。

### 9.4.4.6 DNS 协议检测

通过 DNS 协议检测，USG 可以防止来自应用层，通过 DNS 报文进行的网络攻击。

USG 的 DNS 协议检测通过限制基于 UDP 的 DNS 报文的最大长度来实现。

## 9.5 反垃圾邮件（RBL 过滤）

RBL（Real-time Blackhole List）过滤可以通过本地黑白名单和远程黑名单，有效过滤垃圾邮件，保护邮件服务器。

### 9.5.1 介绍

#### 定义

RBL 过滤一种安全技术，它通过本地黑白名单或第三方组织提供的动态更新的黑名单库，对垃圾邮件进行过滤。

#### 目的

过滤垃圾邮件，保护用户或邮件服务器免受垃圾邮件的侵扰。

### 9.5.2 可获得性

#### License 支持

RBL 过滤功能受 License 控制。当 License 过期后，RBL 功能不可用。

RBL 远程查询服务受激活码控制。当激活码过期后，用户只能使用本地黑白名单，如果要使用远程实时黑名单，必须输入或导入有效的激活码。

#### 版本支持

| 产品                                  | 最低支持版本      |
|-------------------------------------|-------------|
| Secoway USG2110-X/2100 和 USG2100HSR | V100R005C00 |

### 9.5.3 原理描述

RBL 过滤的匹配优先级顺序为：本地白名单、本地黑名单、远程实时黑名单。

- 本地白名单  
当某个邮件的源地址命中了本地白名单，那么本地黑名单和远程黑名单不再对此源 IP 地址进行匹配，该邮件被直接转发。白名单的设置有利于提高匹配效率。
- 本地黑名单  
本地黑名单的优先级比远程黑名单高，当某个邮件的源地址命中了本地黑名单，那么远程黑名单不再对此源 IP 地址进行匹配，该邮件直接阻断。  
RBL 过滤依赖于远程黑名单的完整性，当远程黑名单不包含 SMTP 请求的源 IP 地址时，无法过滤该邮件。此时，可以通过补充本地 RBL 黑名单来完成过滤。
- 远程实时黑名单



远程实时黑名单是由第三方组织提供的动态更新的黑名单。第三方组织实时更新和维护 RBL 并通过 RBL 服务器提供查询服务。

 说明

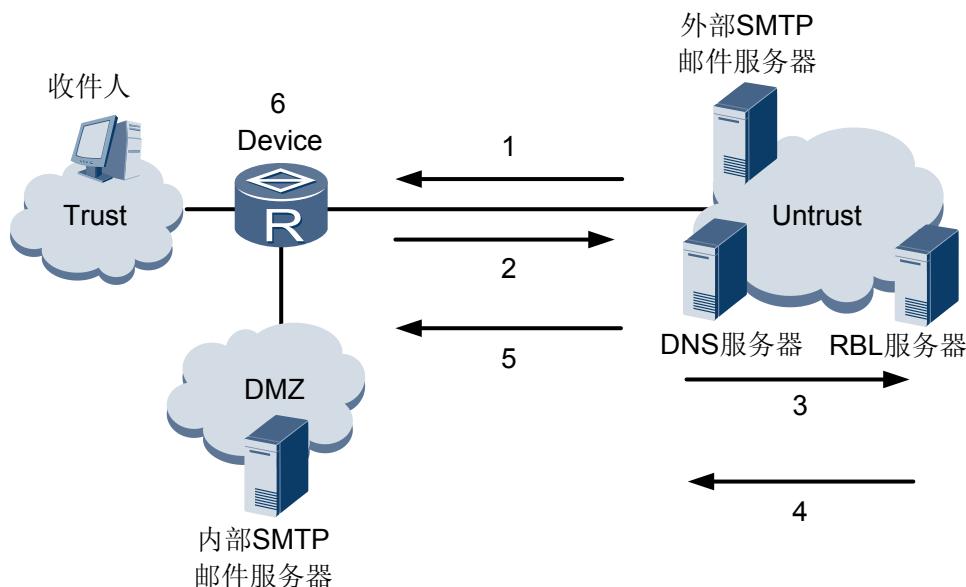
激活码控制 RBL 远程查询服务的有效期限。当激活码过期后，用户只能使用本地黑白名单，如果要使用远程实时黑名单，必须输入或导入有效的激活码。输入或导入激活码的步骤请参见 2。

设备允许用户选择以下两种类型的 RBL 服务器：

- 预定义 RBL 服务器：使用系统预置的 RBL 服务器。
- 自定义 RBL 服务器：用户自己选定需要使用的第三方 RBL 服务器。

远程实时黑名单的处理流程如图 9-5 所示。

图 9-5 远程实时黑名单的处理流程



1. 外部邮件服务器向内网发送 SMTP 请求。
2. Device 收到 SMTP 请求后，提取 SMTP 连接的源 IP 地址并将查询集合和提取的源 IP 地址组合后发送给 DNS 服务器。



注意

使用自定义 RBL 服务器的情况下，查询集合需要在 Device 上进行配置，查询集合由 RBL 服务提供商提供。使用预定义 RBL 服务器不需要配置查询集合。

请使用没有被 DNS 劫持的 DNS 服务器，否则会导致 RBL 失效。DNS 劫持是指服务提供商为用户引导访问其增值站点或合作站点，对 DNS 服务器做修改，在接收到一个不存在结果的 DNS 查询时，总是返回一些特定的 IP 地址，使用户访问到这些增值站点。

3. DNS 服务器根据查询集合向 RBL 服务器转发 RBL 查询请求。
4. RBL 服务器以应答码（例如：127.0.0.1）的形式将查询结果返回给 DNS 服务器。



使用自定义 RBL 服务器的情况下，需要在 Device 上配置应答码。应答码通常是 IP 地址的形式，由 RBL 服务机构提供，例如：中国反垃圾邮件联盟和 Spamhaus。

5. DNS 服务器向 Device 转发应答码。
6. Device 根据应答码判断该 SMTP 连接的邮件是否为垃圾邮件，并进行相应处理。
  - 如果 RBL 服务器回复的应答码和设备侧配置的应答码一致，该邮件将被视为垃圾邮件。
  - 如果 RBL 服务器不回复应答码或回复的应答码与设备侧配置的应答码不一致，放行邮件。

响应方式

如果 RBL 服务器回复的应答码和 Device 上配置的应答码一致，那么该 SMTP 邮件为垃圾邮件。此时，可以通过设置响应方式来对垃圾邮件进行处理。响应方式如表 9-5 所示。

表 9-5 响应方式

| 响应方式      | 处理方式              |
|-----------|-------------------|
| 告警（Alert） | 正常转发邮件，并发出日志告警信息。 |
| 阻断（Block） | 阻断邮件，并发出日志告警信息。   |

9.6 URL 过滤

URL 过滤对用户的 URL 请求进行访问控制，允许或禁止用户访问某些网络资源，可以达到规范上网行为的目的。

9.6.1 介绍

定义

URL 过滤是一种安全技术，它根据不同的时间对象和地址对象，对用户进行 URL 访问控制。

目的

精确管理用户上网行为。

9.6.2 可获得性

License 支持

只有获取了支持 URL 过滤的 License，才能使用 URL 过滤功能，实现上网行为管理。超出期限后，无法进行预定义分类查询，但 URL 的豁免 IP、黑白名单和自定义分类过滤功能仍可以使用。

版本支持

| 产品                                  | 最低支持版本      |
|-------------------------------------|-------------|
| Secoway USG2110-X/2100 和 USG2100HSR | V100R005C00 |

9.6.3 原理描述

9.6.3.1 URL 过滤概述

介绍 URL 过滤的产生背景和作用。

随着企业 Web 访问控制问题不断涌现，传统的防火墙设备已不能很好地管理用户的上网行为，在此背景下，出现了 URL 过滤技术。

随着互联网应用的迅速发展，计算机网络在经济和生活的各个领域迅速普及，使得信息的获取、共享和传播更加方便，但同时也给企业带来了前所未有的威胁：员工随意不受控地访问非法网站，不仅严重影响工作效率而且威胁企业网络安全。

URL 过滤技术可以根据不同的时间对象和地址对象，对用户进行 URL 访问控制，达到精确管理用户上网行为的目的。

USG 采用多种方式实现 URL 过滤，并支持 HTTP 访问日志记录功能。管理员通过监控和记录 HTTP 访问日志，可以分析出内网用户访问 Web 的习惯，从而更好地进行上网行为管理。

 说明

URL 过滤功能仅支持 HTTP 协议。

9.6.3.2 URL 格式

Internet 上的每一个网页都具有一个唯一的标识，称为 URL 地址。URL 地址用来完整地描述 Internet 上的网页或者其他资源的地址。

URL 的一般格式为：“protocol://hostname[:port]/path[?query]”。各参数含义如表 9-6 所示。

表 9-6 URL 格式介绍

| 字段       | 含义                                                                        |
|----------|---------------------------------------------------------------------------|
| protocol | 使用的应用协议，最常用的是 HTTP 协议。对于 HTTP 协议，一般可不输入“http://”。                         |
| hostname | Web 服务器的域名或者 IP 地址。                                                       |
| port     | 可选，通信端口。各种应用协议都有默认的端口号，如 HTTP 协议的默认端口为 80。当 Web 服务器采用非标准端口时，URL 中不能省略端口号。 |

| 字段     | 含义                                    |
|--------|---------------------------------------|
| path   | 由零或多个“/”符号隔开的字符串，一般用来表示主机上的一个目录或文件地址。 |
| ?query | 可选，用于给动态网页传递参数。                       |

例如，对于“http://www.abcd.com/news/education.aspx?name=tom&age=20”：

- www.abcd.com 为 hostname。
- news/education.aspx 为 path。
- ?name=tom&age=20 为参数部分。

9.6.3.3 URL 匹配方式

URL 匹配方式包括前缀匹配、后缀匹配、关键字匹配、精确匹配、参数匹配。

其中，前缀匹配、后缀匹配、关键字匹配和精确匹配只针对 URL 的域名（或者 IP 地址）和路径部分，参数匹配只针对 URL 的 query 部分。[表 9-7](#) 对几种匹配方式做了简单比较。

表 9-7 URL 匹配方式

| 匹配方式  | 条目                    | 匹配结果                                                                                                                                                                          |
|-------|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 前缀匹配  | www.test.com          | 匹配所有以 www.test.com 开头的 URL，如：<br>www.test.com、www.test.com/solutions.do。                                                                                                      |
| 后缀匹配  | aspx                  | 匹配所有以 aspx 结尾的 URL，如：www.test.com/<br>news/solutions.aspx、www.test.com/it/price.aspx。                                                                                         |
| 关键字匹配 | test.com              | 匹配所有包含 test.com 的 URL，如：<br>www.test.com/news/solutions.aspx、www.test.com/<br>it/。                                                                                            |
| 参数匹配  | a-param               | 匹配所有参数中包含 a-param 的 URL，如：<br>www.test.com/education.aspx?#\$\$%a-param^&、<br>www.test.com/news/en/political?a-parameter，但是<br>不会匹配 www.a-param.com、www.test.com/a-<br>param。 |
| 精确匹配  | www.test.com/<br>news | 只匹配 www.test.com/news。www.test.com/news/<br>solutions.aspx、www.test.com/news/en/等不会匹配<br>该条目。                                                                                 |

 说明

- 对于所有匹配方式，USG 会去除输入字符串中包含的前缀"http://", 如果是精确匹配，对于 Host 字段后没有"/"的字符串会在末尾增加"/"。
- URL 条目不区分大小写。

匹配方式主要应用于黑白名单和 URL 自定义分类的配置。用户配置的 URL 条目与匹配方式一起，决定了黑白名单或自定义分类的范围。表 9-8 列出了黑白名单和自定义分类对匹配方式的支持情况。

表 9-8 匹配方式的支持情况

| 匹配方式  | 黑名单 | 白名单 | 自定义分类 |
|-------|-----|-----|-------|
| 前缀匹配  | 支持  | 支持  | 不支持   |
| 后缀匹配  | 支持  | 支持  | 不支持   |
| 关键字匹配 | 支持  | 支持  | 支持    |
| 参数匹配  | 支持  | 不支持 | 不支持   |
| 精确匹配  | 支持  | 支持  | 支持    |

### 9.6.3.4 URL 过滤方式

USG 提供基于豁免 IP、黑白名单和 URL 分类的过滤方式。通过配置时间对象和地址对象实现在特定时间内对特定用户进行 URL 过滤，达到精确管理用户上网行为的目的。

### URL 过滤策略

URL 过滤支持豁免 IP、黑白名单和 URL 分类的过滤策略。

#### 说明

URL 过滤的五种匹配方式中优先级顺序为：豁免 IP、白名单、黑名单、自定义分类、预定义分类。

- 豁免 IP

豁免 IP 发起的请求不经过 URL 过滤，也不会被记录 HTTP 访问日志，因此可以通过配置豁免 IP 为特殊用户配置所有的上网权限。

- 黑白名单

USG 将 HTTP 上网请求的 URL 与黑白名单进行匹配，如果匹配白名单则允许该 HTTP 请求，如果匹配黑名单则阻止该 HTTP 请求，同时显示 Web 推送页面。

- URL 分类

用户可以对 URL 进行分类，并且指定每个 URL 分类的控制动作，从而达到 URL 过滤的目的。

URL 分类包括两种，分别是：

- 预定义分类，由安全服务中心提供并维护。
- 自定义分类，由用户自行配置和维护。

防火墙策略支持对属于特定地址集的用户在特定时间段内进行访问控制。USG 可以利用防火墙策略实现时间和地址的 URL 分类过滤：

- 时间对象

USG 支持在特定时间段内对 HTTP 请求进行 URL 过滤。

- 地址对象

USG 支持对特定地址或者地址组发起的 HTTP 请求进行 URL 过滤。

## 控制动作

URL 过滤的控制动作包括以下两个：

- 允许  
指允许用户访问请求的 URL。
- 阻止  
指阻止用户访问请求的 URL。

如果启用了预定义分类过滤，但是 USG 连接安全服务中心失败或者查询超时，则按照查询预定义分类失败处理方式处理。

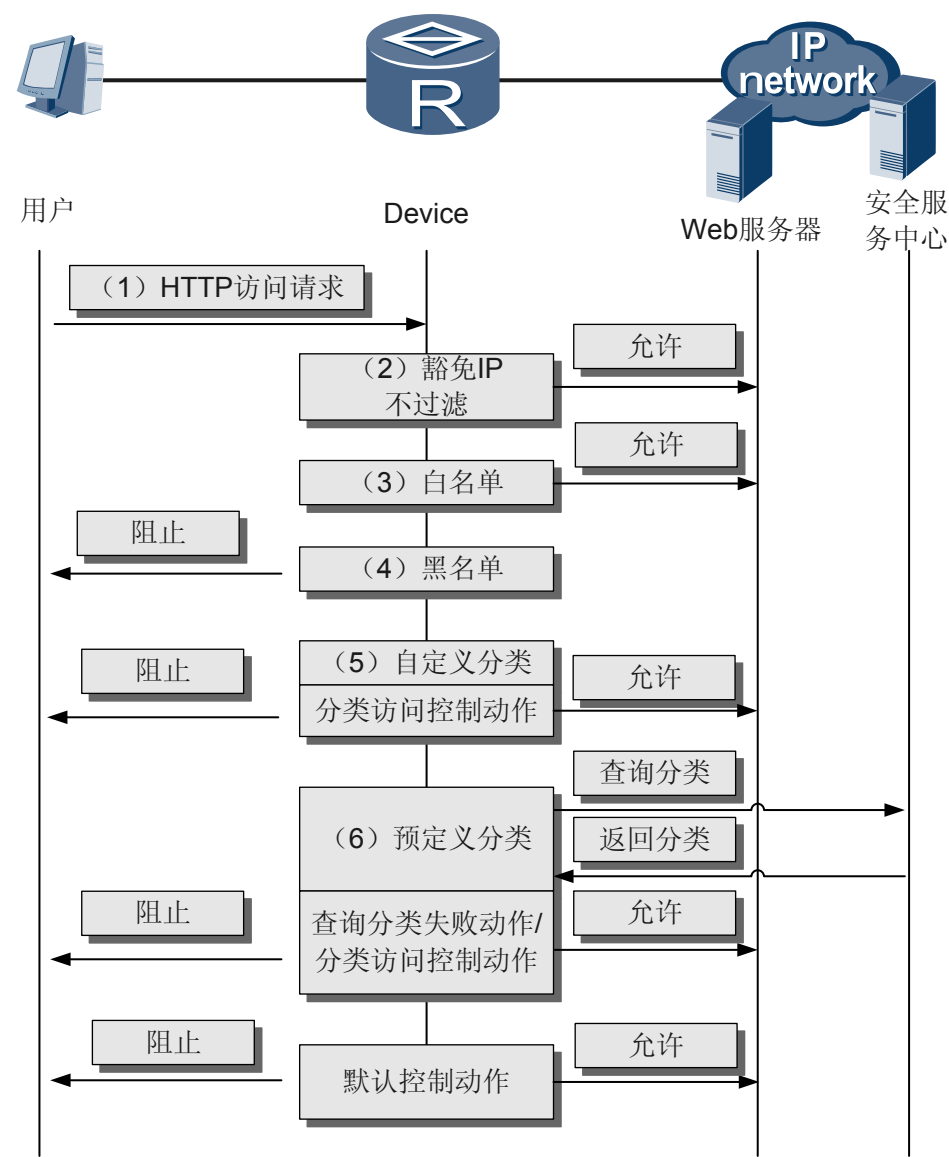
如果内部用户的 URL 请求没有匹配黑名单和白名单，也不属于任何一个分类，或者黑白名单过滤和分类过滤功能均未启用，则 USG 按照 URL 过滤默认访问控制动作处理此 URL 请求。

### 9.6.3.5 URL 过滤工作原理

介绍 USG 进行 URL 过滤的原理。

如[图 9-6](#)所示，以 Trust-Untrust 域间的 URL 过滤为例，说明 USG 进行 URL 过滤的原理。

图 9-6 URL 过滤原理图



在 USG 启用 URL 过滤功能的情况下，当用户通过 USG 使用 HTTP 访问某个网络资源时，USG 将对 URL 进行过滤。工作流如下：

- 说明**
- 以下流程是在启用了豁免 IP、黑白名单和自定义分类、预定义分类过滤功能的前提下。
1. 用户发起 HTTP 请求。
  2. USG 将发起 HTTP 请求的用户 IP 地址与豁免 IP 进行匹配。
    - 如果匹配豁免 IP，则不需过滤，直接允许。
    - 如果未匹配豁免 IP，则进行下一步检测。
  3. USG 将 HTTP 请求的 URL 与白名单进行匹配。
    - 如果匹配白名单，则允许该 HTTP 访问请求。
    - 如果未匹配白名单，则进行下一步检测。

4. USG 将 HTTP 请求的 URL 与黑名单进行匹配。
  - 如果匹配黑名单，则阻止该 HTTP 访问请求。
  - 如果未匹配黑名单，则进行下一步检测。
5. USG 将 HTTP 请求的 URL 与自定义分类进行匹配。
  - 如果匹配自定义分类，则根据自定义分类的访问控制动作允许或者阻止该 HTTP 访问请求。
  - 如果未匹配自定义分类，则进行下一步检测。
6. USG 将 HTTP 请求的 URL 向安全服务中心发起分类查询，安全服务中心返回查询结果。
  - 如果结果是该 URL 属于安全服务中心上的某个分类，则根据分类的控制动作允许或者阻止该 HTTP 访问请求。
  - 如果查询失败，则根据查询预定义分类失败的处理方式允许或者阻止该 HTTP 访问请求。



说明

如果查询的 URL 在安全服务中心上未具体分类，则返回 Other-Sites 分类，用户可以配置 Other-Sites 分类的访问控制动作。

7. 对于未匹配黑白名单、自定义分类和预定义分类的 URL，根据 URL 的默认访问控制动作进行处理。

如果启用了 HTTP 访问日志记录功能，USG 将对访问符合配置扩展名资源的 HTTP 请求进行记录。豁免 IP 的发起的 HTTP 请求不会被记录访问日志。

## 9.7 应用控制（DPI）

DPI（Deep Packet Inspection）提供对 TCP/UDP 传输层以上具体应用协议的识别功能，基于该功能，支持对网络流量按协议进行细分统计，对某些具体协议流量进行流量阻断或放行。

### 9.7.1 介绍

#### 定义

应用控制是一种通过 DPI 技术对网络中的报文进行识别控制的应用功能。DPI 即深度报文检测，通过对报文的应用层数据进行内容检测，分析报文或数据流在 IP 和 TCP/UDP 层以上的应用业务类型，从而进行流量识别和流量管理以及其他增值服务。

#### 目的

DPI 通过对数据流进行深度检测，可以识别应用层业务，并对指定类型的数据流量进行控制。通过分析 USG 收到的数据包并和 DPI 特征库进行比对，对 P2P、IM、VoIP 等类型的网络数据流量进行分类，并对不同类型的协议进行相应的控制。

### 9.7.2 可获得性

#### License 支持

本特性无需 License 支持。



版本支持

| 产品                                     | 最低支持版本      |
|----------------------------------------|-------------|
| Secoway USG2110-X/2100、USG2100 BSR/HSR | V100R005C00 |

9.7.3 特性增强

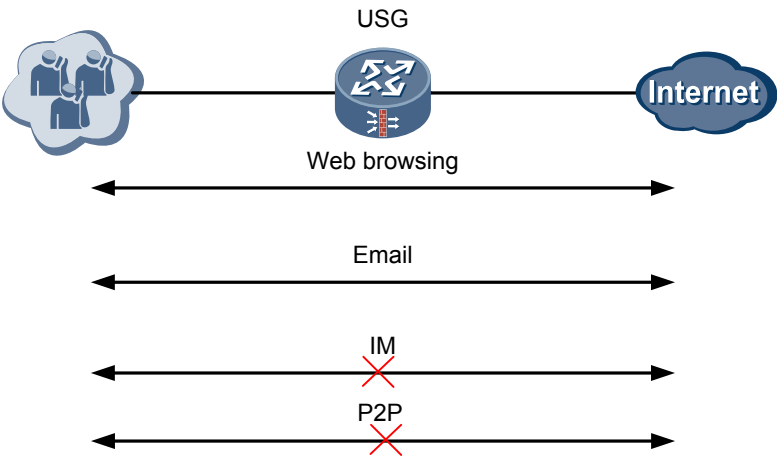
| 版本       | 特性增强                                                      |
|----------|-----------------------------------------------------------|
| V300R001 | 应用控制（DPI）功能改进配置方法，能将应用控制策略（DPI 策略）应用在域间或域内，被域间或者域内策略引用生效。 |

9.7.4 应用场景

典型应用

在实际应用中，可以对网络中检测到的各种类型的应用协议的流量进行阻断和放行，确保信息安全，保证正常业务的运行。如图 9-7 所示，对于一般的网络浏览和邮件流量予以放行，对某些不允许使用的 IM 类型应用和影响带宽的 P2P 下载应用可以进行阻断。

图 9-7 DPI 应用场景

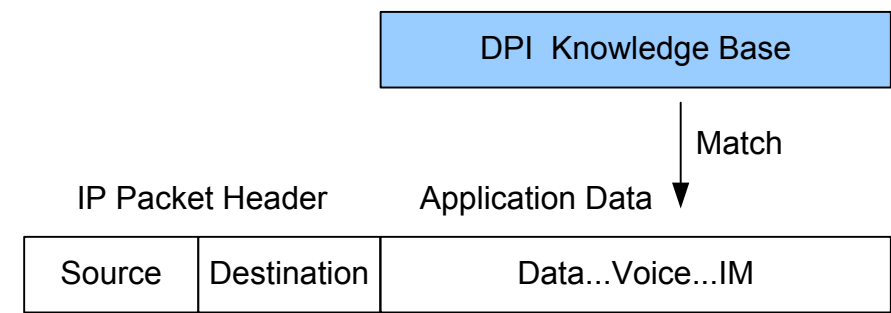


9.7.5 原理描述

9.7.5.1 DPI 工作原理

原理概述

图 9-8 原理概述示意图



如图 9-8 所示，设备根据 DPI 知识库或者用户自定义的应用协议，对报文进行匹配，并将结果用于对报文的控制。

与传统的只对报文中的 IP 报头进行检查的简单机制不同，DPI 设备将网络上的数据报文根据五元组（源地址、源端口号、目的地址、目的端口号、传输层协议类型）分为各个应用数据流，并通过识别技术对应用数据流中的特定的数据报文进行探测，从而确定应用数据流对应的应用。

特征识别

特征识别是 DPI 技术的核心。对于不同类型的应用，所依赖的应用协议各不相同，而不同的应用协议具有各自的特征，DPI 知识库即各种不同协议特征的集合。

对于不同的协议，特征的表现形式也不同，可能是特定的端口、特定的字符串和特定的比特序列。

协议的特征不仅在单个报文中体现，某些协议报文的特征是分布在多个报文中的，需要对多个报文进行采集分析，对报文在内容或报文长度的变化规律或趋势上进行分析，才能够判断协议类型。

9.7.5.2 关联协议识别

原理概述

某些数据流，控制通道和数据通道是分开的（如 FTP、SIP、H.323 等），一般的处理方式是对这些会话分别进行识别，这可能导致分析速度较慢或影响对数据通道的识别。

关联协议识别，是将来自同一源地址的控制通道和后续的数据通道进行关联的技术，从而达到快速识别的目的。

关联过程

关联协议识别可以将同一应用协议的控制通道流和数据通道流关联起来。通过对控制通道流的分析，可以分析出通讯双方将要在哪个通道上建立何种类型的数据流，并在协议识别时将控制通道流和该控制通道流协商出来的数据通道流关联起来。

DPI 在对控制通道流进行深度解析时，提取出其中协商的数据通道流的源三元组和目的三元组信息，并加入关联表。在后续识别过程中，可以通过该关联表项对数据通道流快速识别。

### 9.7.5.3 全包检测

一般情况下，DPI 对报文的特定字节进行协议特征检测，对于多数已知协议可以做到成功检测。

某些协议的报文是承载在其他协议上进行传输的，如 MSN 信息可能通过 HTTP 协议承载、VoIP 信息可能通过 IM 类型协议承载。全包检测即对这种类型的报文的所有字节进行检测，以提高对协议的识别准确率。

USG 支持对指定的协议启用全包检测功能。

# 10 MPLS 与 VPN

## 关于本章

### 10.1 VPN

VPN（Virtual Private Network）用于在公用网络上构建私人专用虚拟网络。可以把现有的 IP 网络分解成逻辑上隔离的网络。在保证网络的安全性、可靠性、可管理性的同时提供更强的扩展性和灵活性。

### 10.2 IPSec

通信双方在 IP 层通过隧道加密方式进行通信，其优点是能够保证数据报在网络上传输时的私有性、完整性、真实性和防重放。

### 10.3 L2TP

L2TP 是二层 VPN 技术，提供了对 PPP 链路层数据包的通道（Tunnel）传输支持，允许二层链路端点和 PPP 会话点驻留在不同设备上，并且采用包交换网络技术进行信息交互，扩展了 PPP 模型。

### 10.4 GRE

GRE 是三层隧道的封装协议，对某些网络层协议的数据报文进行封装并在另一网络层协议中传输。GRE 提供了将一种协议的报文封装在另一种协议报文中的机制，使报文能够在隧道中传输。

### 10.5 MPLS

MPLS（Multiprotocol Label Switching）支持多层标签和面向连接特性。MPLS LDP 协议提供了环路检测机制，可以预防 LSP 环路的出现。如果管理员不需要严格控制 LSP 建立的过程，且不需要在 MPLS 网络中部署流量工程，则推荐采用 LDP 协议来创建 LSP。

### 10.6 BGP MPLS IP VPN

BGP/MPLS IP VPN 使用 BGP 在服务提供商骨干网上发布 VPN 路由，使用 MPLS 在服务提供商骨干网上转发 VPN 报文。BGP/MPLS IP VPN 组网方式灵活、可扩展性好，并能够方便地支持 MPLS QoS。

### 10.7 SSL VPN

SSL 建立在传输层之上，可实现加密算法、通信密钥的协商以及服务器的认证功能。主要提供业务有 Web 代理、网络扩展、文件共享和端口转发。

## 10.1 VPN

VPN（Virtual Private Network）用于在公用网络上构建私人专用虚拟网络。可以把现有的 IP 网络分解成逻辑上隔离的网络。在保证网络的安全性、可靠性、可管理性的同时提供更强的扩展性和灵活性。

### 10.1.1 介绍

#### 定义

VPN 是近年来随着 Internet 的广泛应用而迅速发展起来的一种新技术，用于在公用网络上构建私人专用虚拟网络。“虚拟”主要指这种网络是一种逻辑上的网络。

#### 目的

VPN 用于用户和企业网络之间的安全接入或企业分部之间的安全连接，保证经济、安全、有效地进行网络互联。

### 10.1.2 原理描述

#### 10.1.2.1 VPN 的原理

VPN 的基本原理是利用隧道技术，把 VPN 报文封装在隧道中，利用 VPN 骨干网建立专用数据传输通道，实现报文的透明传输。

隧道技术使用一种协议封装另外一种协议报文，而封装协议本身也可以被其他封装协议所封装或承载。对用户来说，隧道是其 PSTN/ISDN 链路的逻辑延伸，在使用上与实际物理链路相同。

VPN 隧道需要完成的功能包括：

- 封装用户数据
- 实现隧道两端的连通性
- 定时检测 VPN 隧道的连通性
- VPN 隧道的安全性
- VPN 隧道的 QoS 特性

图 10-1 VPN 接入示意图

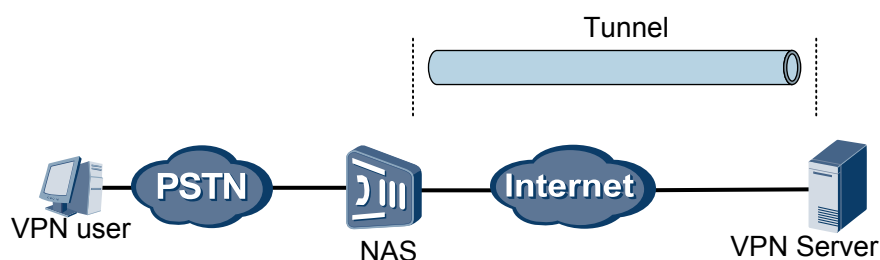


图 10-1 中，VPN 用户通过 PSTN/ISDN 网拨入 ISP 的 NAS（Network Access Server），NAS 通过用户名或接入号码识别出该用户为 VPN 用户后，发起与目的 VPN 服务器的连接，建立 NAS 与 VPN 服务器之间的隧道（Tunnel）。然后将用户数据包封装成 IP 报文后通过该隧道传送给 VPN 服务器，VPN 服务器收到数据包后进行解封装，获得原始报文。反向的处理也一样。隧道两侧可以对报文进行加密处理，使 Internet 上的其他用户无法读取，因而通信是安全可靠的。

### 10.1.2.2 VPN 的特征

VPN 具有以下两个基本特征：

- 专用（Private）：对于 VPN 用户，使用 VPN 与使用传统专网没有区别。VPN 与底层承载网络之间保持资源独立，即 VPN 资源不被网络中非该 VPN 的用户所使用；VPN 能够提供足够的安全保证，确保 VPN 内部信息不受外部侵扰。
- 虚拟（Virtual）：VPN 用户内部的通信是通过公共网络进行的，而这个公共网络同时也可以被其他非 VPN 用户使用，VPN 用户获得的只是一个逻辑意义上的专网。这个公共网络称为 VPN 骨干网（VPN Backbone）。

利用 VPN 的专用和虚拟的特征，可以把现有的 IP 网络分解成逻辑上隔离的网络。这种逻辑隔离的网络应用丰富：可以用在解决企业内部的互连、相同或不同办事部门的互连；也可以用来提供新的业务，如为 IP 电话业务专门开辟一个 VPN，以此解决 IP 网络地址不足、QoS 保证以及开展新的增值服务等问题。

### 10.1.2.3 VPN 的优势

从客户角度看，VPN 和传统的数据专网相比具有如下优势：

- 安全：在远端用户、驻外机构、合作伙伴、供应商与公司总部之间建立可靠的连接，保证数据传输的安全性。这对于实现电子商务或金融网络与通讯网络的融合特别重要。
- 廉价：利用公共网络进行信息通讯，企业可以用更低的成本连接远程办事机构、出差人员和业务伙伴。
- 支持移动业务：支持驻外 VPN 用户在任何时间、任何地点的移动接入，能够满足不断增长的移动业务需求。
- 服务质量保证：构建具有服务质量保证的 VPN（如 MPLS VPN），可为 VPN 用户提供不同等级的服务质量保证。

从运营商角度看，VPN 具有如下优势：

- 可运营：提高网络资源利用率，有助于增加 ISP 的收益。
- 灵活：通过软件配置就可以增加、删除 VPN 用户，无需改动硬件设施。在应用上具有很大灵活性。
- 多业务：SP 在提供 VPN 互连的基础上，可以承揽网络外包、业务外包、客户化专业服务的多业务经营。

VPN 以其独具特色的优势赢得了越来越多的企业的青睐，使企业可以较少地关注网络的运行与维护，从而更多地致力于企业的商业目标的实现。另外，运营商可以只管理、运行一个网络，并在一个网络上同时提供多种服务（如 Best-effort IP 服务、VPN、流量工程、差分服务(Diffserv)），从而减少建设、维护和运行的费用。

VPN 在保证网络的安全性、可靠性、可管理性的同时提供更强的扩展性和灵活性。在全球任何一个角落，只要能够接入到 Internet，即可使用 VPN。

### 10.1.2.4 VPN 的分类

USG 支持以下几类 VPN:

- IPsec VPN

IPsec (IP Security) 是 IETF (Internet Engineering Task Force) 制定的一系列协议, 提供了一种建立和管理安全隧道的方式, 通过对要传输的数据报文提供认证和加密服务来防止数据在网络内或通过公网传输时被非法查看或篡改, 相当于为位于不同地域的用户创建了一条安全的通信隧道。

IPsec VPN 适用于 Site to Site 的安全访问。

IPsec VPN 的具体介绍请参见 [10.2 IPsec](#)。

- L2TP VPN

PPP 协议定义了一种封装技术, 可以在二层点到点链路上传输多种协议数据包, 这时, 用户与 NAS 之间运行 PPP, 二层链路端点与 PPP 会话点在相同硬件设备上。

L2TP 协议提供了对 PPP 链路层数据帧的隧道 (Tunnel) 传输支持, 允许二层链路端点和 PPP 会话点驻留在不同设备上, 并采用包交换技术进行信息交互, 从而扩展了 PPP 模型。

L2TP 协议结合了 L2F 协议和 PPTP 协议的优点, 成为 IETF 有关二层隧道协议的工业标准。

L2TP VPN 的具体介绍请参见 [10.3 L2TP](#)。

- GRE VPN

通用路由封装 GRE (Generic Routing Encapsulation) 用来对某些网络层协议如 IPX (Internet Packet Exchange) 的报文进行封装, 使这些被封装的报文能够在另一网络层协议 (如 IP) 中传输。GRE 可以作为 VPN 的第三层隧道协议, 为 VPN 数据提供透明传输通道。

为了防止攻击, 可以在 GRE 封装的基础上, 再使用 IPsec 协议加密。即使用 IP 的安全机制来保护 GRE 头及其负载, 例如 ESP (Encapsulating Security Payload) 或 AH (Authentication Header)。ESP 保护整个 GRE 报文。AH 认证除了可变域整个 IP 报文。

GRE VPN 的具体介绍请参见 [10.4 GRE](#)。

- BGP/MPLS IP VPN

BGP/MPLS IP VPN 是一种 L3VPN (Layer 3 Virtual Private Network)。它使用 BGP (Border Gateway Protocol) 在服务提供商骨干网上发布 VPN 路由, 使用 MPLS (Multiprotocol Label Switch) 在服务提供商骨干网上转发 VPN 报文。这里的 IP 是指 VPN 承载的是 IP 报文。

MPLS 无缝地集成了 IP 路由技术的灵活性和 ATM 标签交换技术的简捷性。MPLS 在无连接的 IP 网络中增加了面向连接的控制平面, 为 IP 网络增添了管理和运营的手段。

BGP/MPLS IP VPN 的具体介绍请参见 [10.5 MPLS](#) 和 [10.6 BGP MPLS IP VPN](#)。

- SSL VPN

SSL (Secure Sockets Layer) VPN 无需安装客户端, 用户只要使用支持 HTTPS (以 SSL 为基础的 HTTP) 的 Web 浏览器, 可以方便地建立标准安全通道访问远程应用, 因此大幅度减轻 VPN 系统管理员负担。SSL VPN 适用于下述情况:

- 企业需要通过 Web 远程接入互联网。
- 客户端与目标服务器之间有防火墙, 允许 HTTPS 数据包通过, 但不允许 IKE (Internet Key Exchange) 或 IPsec 数据包通过。



- 需要精细访问控制能力的场合。

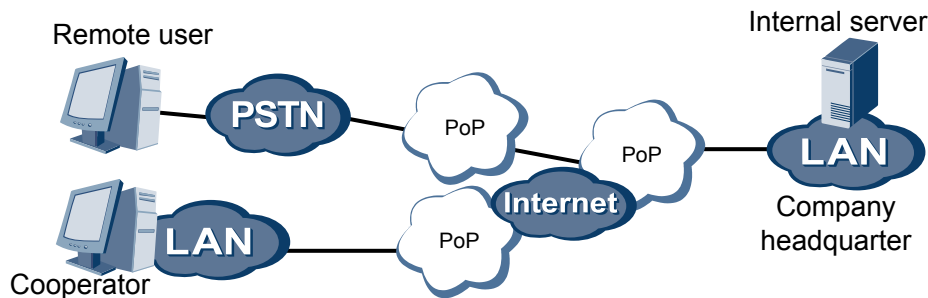
SSL VPN 适用于 Point to Site 的安全访问。

SSL VPN 的具体介绍及其相关特性请参见 [10.7 SSL VPN](#)。

### 10.1.2.5 VPN 基本组网应用

以某企业为例，通过 VPN 建立的企业内部网如图 10-2 所示。

图 10-2 VPN 组网示意图



从图 10-2 可以看出，企业内部资源享用者通过 PSTN（Public Switched Telephone Network）/ISDN（Integrated Services Digital Network）网或局域网就可以连入本地 ISP 的 POP（Point Of Presence）服务器，从而访问公司内部资源。而利用传统的 WAN 组建技术，相互之间要有专线相连才可以达到同样的目的。虚拟网组成后，远端用户和外地客户甚至不必拥有本地 ISP 的上网权限就可以访问企业内部资源，这对于流动性很大的出差员工和分布广泛的客户来说是很有帮助的。

企业应用 VPN 业务所需的设备很少，只需在资源共享处放置一台支持 VPN 的服务器（如一台 Windows NT 服务器或支持 VPN 的防火墙）就可以了。资源享用者通过 PSTN/ISDN 网或局域网连入本地 POP 服务器后，直接呼叫企业的远程服务器（VPN 服务器），呼叫接续过程由 ISP 的接入服务器（Access Server）与 VPN 服务器共同完成。

## 10.2 IPSec

通信双方在 IP 层通过隧道加密方式进行通信，其优点是能够保证数据报在网络上传输时的私有性、完整性、真实性和防重放。

### 10.2.1 介绍

主要介绍 IPSec 的由来、定义和作用。

由于 IP 报文本身并不集成任何安全特性，恶意用户很容易便可伪造 IP 报文的地址、修改报文内容、重播以前的 IP 数据包以及在传输途中拦截并查看数据包的内容。因此，我们不能担保收到的 IP 数据包的安全。在应用层保证网络安全的方法只对特定的应用有效，不够通用。人们迫切需要能够在 IP 层提供安全服务的协议，这样可以使 TCP/IP 高层的所有协议受益。IPSec（Internet Protocol Security）正是用来解决 IP 层安全性问题的技术。IPSec 被设计为同时支持 IPv4 和 IPv6 网络。

IPSec 是 Internet 工程任务组(IETF)制定的一个开放的网络层安全框架协议。它并不是一个单独的协议，而是一系列为 IP 网络提供安全性的协议和服务的集合。IPSec 主要包括



安全协议 AH（Authentication Header）和 ESP（Encapsulating Security Payload），密钥管理交换协议 IKE（Internet Key Exchange）以及用于网络认证及加密的一些算法等。

IPSec 主要通过加密与验证等方式，为 IP 数据包提供安全服务。IPSec 可以提供的安全服务包括：

- 用户数据加密  
通过用户数据加密提供数据私密性。
- 数据完整性验证  
通过数据完整性验证确保数据在传输路径上未经过篡改。
- 数据源验证  
通过对发送数据的源进行身份验证，保证数据来自真实的发送者。
- 防止数据重放  
通过在接收方拒绝重复的数据包防止恶意用户通过重复发送捕获到的数据包进行攻击。

## 10.2.2 参考标准和协议

介绍 IPSec 特性相关的参考标准和协议。

与 IPSec 特性相关的参考标准与协议如下：

- RFC 4301: Security Architecture for the Internet Protocol  
定义了 IPSec 的架构，IPSec 能够提供的安全服务，以及各个组件之间如何配置工作的。
- RFC 4302: IP Authentication Header  
定义了数据连接中，使用验证头（AH）协议验证和核实数据包的完整性。
- RFC 4303: IP Encapsulating Security Payload (ESP)  
定义了数据连接中，使用封装安全载荷（ESP）协议来提供保密性、数据包的验证，和数据包的完整性。
- RFC 1829: The ESP DES-CBC Transform  
定义了数据连接中 DES 作为加密算法的使用。
- RFC 1851: The ESP Triple DES Transform  
定义了数据连接中 3DES 作为加密算法的使用。
- RFC 2403: The Use of HMAC-MD5-96 within ESP and AH  
定义了 IPSec 数据连接中作为 HMAC 功能的 MD5 的使用。
- RFC 2404: The Use of HMAC-SHA-1-96 within ESP and AH  
定义了 IPSec 数据连接中作为 HMAC 功能的 SHA-1 的使用。
- RFC 2407: The Internet IP Security Domain of Interpretation for ISAKMP  
定义了 IPSec DOI，它解释了如何使用 ISAKMP 协商安全联盟。
- RFC 2408: Internet Security Association and Key Management Protocol (ISAKMP)  
定义了 ISAKMP 是如何来构造安全连接的。
- RFC 2409: The Internet Key Exchange (IKE)  
定义了 Internet 密钥交换（IKE）协议，它是用于协商和验证密钥信息来保护连接的。

- RFC 5996: Internet Key Exchange Protocol Version 2 (IKEv2)  
定义了 IKEv2。
- RFC 2410: The NULL Encryption Algorithm and Its Use With IPsec  
这个 RFC 允许在 ESP 中使用一个“空”的加密算法，允许使用 ESP 而无需加密。
- RFC 3706: A Traffic-Based Method of Detecting Dead Internet Key Exchange (IKE) Peers  
定义了使用 DPD 方法检测失效对等体。
- RFC 3947: Negotiation of NAT-Traversal in the IKE  
定义了如何检测到 IPsec 设备中间的 NAT 设备以及如何使 UDP 封装的 IPsec 报文穿越 NAT。
- RFC 3948: UDP Encapsulation of IPsec ESP Packets  
定义了如何 NAT 环境下使用 UDP 封装 ESP 报文的方法
- RFC 4434: The AES-XCBC-PRF-128 Algorithm for the Internet Key Exchange Protocol (IKE)  
定义了数据连接中 AES（128 位）作为加密算法的使用。

### 10.2.3 可获得性

IPsec 特性的 License 支持情况，以及功能支持情况。

#### License 支持

USG5500 需要 License 支持，用户能够创建的隧道数由所购买的 License 决定。无 License 情况下 IPsec 功能不可用，且 IPsec 相关的命令行和 Web 界面均不可见。USG2110-X/2100/2200/5100 和 USG2100/2200/5100 BSR/HSR 无需 License 支持。

#### 功能支持

IPsec 协议扩展（L2TP over IPsec/GRE over IPsec）、IPsec 安全协议（AH/ESP/AH-ESP）、IKE 协商模式（主模式/野蛮模式）、NAT 穿越、IKE 对等体检测等功能是 IPsec 的基本功能，各系列都支持。下面仅对不同系列有差别的功能项进行对比。

表 10-1 USG 系列功能支持

| 功能分类   | 功能项       | V300R001                                                                                         |
|--------|-----------|--------------------------------------------------------------------------------------------------|
| 隧道数    | 缺省支持      | USG5500: 0。USG2110-X/2100/2200 和 USG2100/2200 BSR/HSR: 2000。USG5100 和 USG5100 BSR/HSR: 4000。     |
|        | 最多支持      | USG5500: 15000。USG2110-X/2100/2200 和 USG2100/2200 BSR/HSR: 2000。USG5100 和 USG5100 BSR/HSR: 4000。 |
| 多实例    | IPsec 多实例 | Y                                                                                                |
| IKE 版本 | IKEv1     | Y                                                                                                |
|        | IKEv2     | Y                                                                                                |

| 功能分类         | 功能项            | V300R001 |
|--------------|----------------|----------|
| 身份认证         | 预共享密钥          | Y        |
|              | RSA 签名         | Y        |
|              | RSA 验证字        | Y        |
|              | 数字信封认证（HSS-DE） | 依赖加密卡支持  |
|              | EAP 认证         | Y        |
| 本端 ID 类型     | IP             | Y        |
|              | DN             | Y        |
|              | FQDN           | Y        |
|              | USER-FQDN      | Y        |
| 指定对端方式       | IP 地址          | Y        |
|              | 域名（Domain）     | Y        |
|              | DN             | Y        |
|              | FQDN           | Y        |
|              | USER-FQDN      | Y        |
| 封装模式         | 传输             | Y        |
|              | 隧道             | Y        |
| IPSec 隧道自动协商 |                | Y        |
| IPSec 隧道化    |                | Y        |
| IPSec 双机热备   |                | Y        |

10.2.4 应用场景

分别从接入方式、冗余设计和虚拟防火墙下 VPN 应用这几个维度介绍 IPSec 的应用场景。

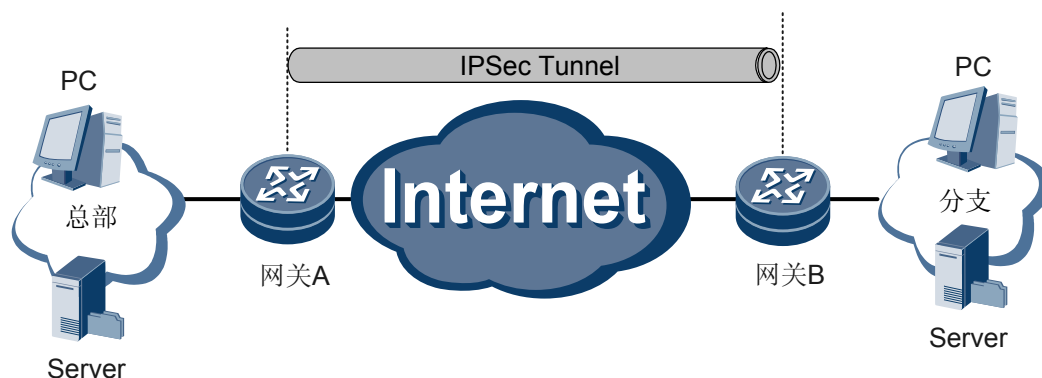
10.2.4.1 局域网通过 VPN 互通

用于指导公司之间通过 IPSec VPN 网络互连。

点到点 VPN（Site-to-Site VPN）

点到点 VPN 也称局域网到局域网的 VPN（LAN to LAN VPN），网关到网关 VPN（Gateway to Gateway VPN），主要用于公司总部与分支机构之间建立 IPSec 隧道，从而实现局域网之间互通。典型组网如图 10-3 所示。

图 10-3 点到点方式组网



网关的 IP 地址可以是静态公网 IP 地址，也可以通过 3G、ADSL、PPPoE 拨号或 DHCP 方式动态获取。USG 支持两端 IP 地址都不固定的场景，此时至少一端需要有固定的域名，就能够通过 DDNS 解析到固定的 IP 地址。

说明：

- USG 支持 IPSec 网关和 NAT 网关合一。
- 当两个 IPSec 网关之间存在 NAT 设备时，USG 支持 IPSec NAT 穿越。

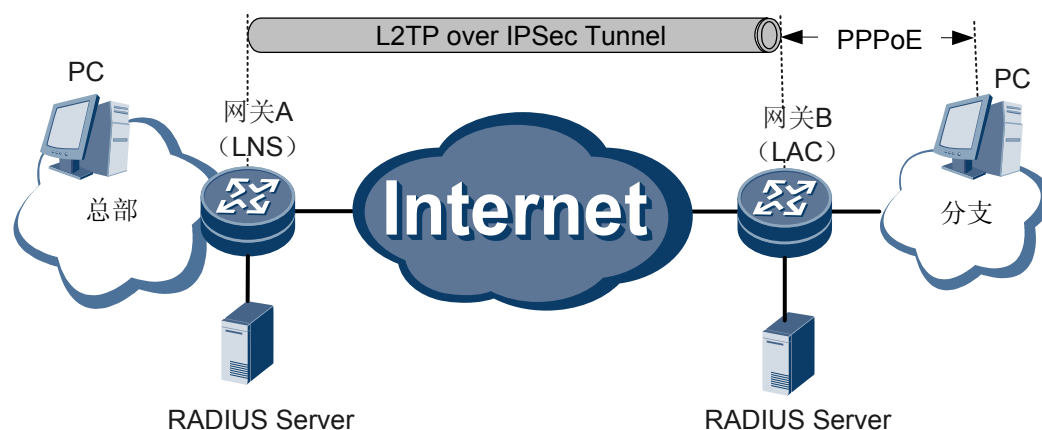
## 点到点 VPN 扩展（L2TP over IPSec）

L2TP over IPSec，即先用 L2TP 封装报文后再用 IPSec 封装，这样可以综合两种 VPN 的优势，通过 L2TP 实现用户验证和地址分配，并利用 IPSec 保障安全性。

如图 10-4 所示，LNS 和 LAC 的出接口 IP 地址固定，分支用户通过 PPPoE 拨号到 LAC。由 LAC 通过 Internet 向 LNS 发起建立隧道连接请求。在 LNS 和 LAC 之间建立 L2TP over IPSec 隧道。LAC 侧对用户的身份进行验证，LNS 侧还可以对用户的身份进行二次验证。验证通过后由 LNS 为接入用户分配总部内网的 IP 地址。

L2TP 能够提供数据传输的出入报文数、字节数以及连接的起始、结束时间等计费数据，可根据这些数据方便地进行网络计费。可以在 LAC 和 LNS 两处同时对用户进行上网计费，LAC 处用于产生账单，LNS 处用于付费及审计。

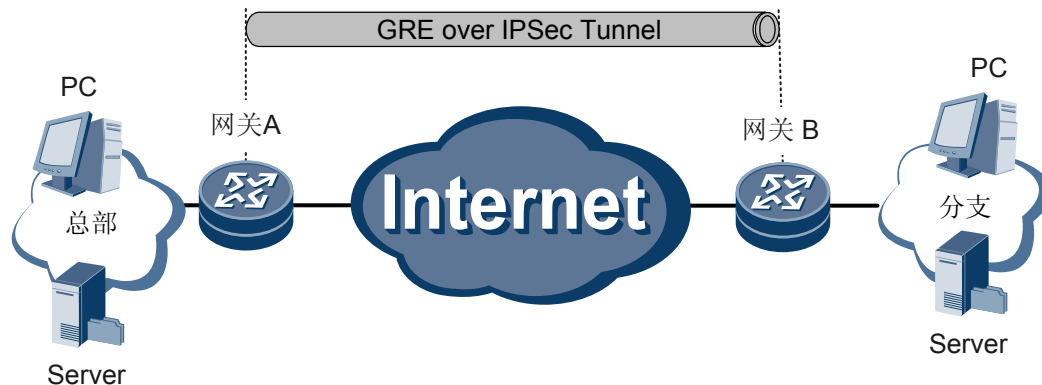
图 10-4 L2TP over IPSec 组网



## 点到点 VPN 扩展（GRE over IPSec）

IPSec 本身不支持封装组播、广播和非 IP 报文。当需要在 IPSec VPN 上传输这些报文时，可以使用 GRE 先将其封装为 IP 报文，然后再进行 IPSec 封装。如图 10-5 所示。

图 10-5 GRE over IPSec 组网图



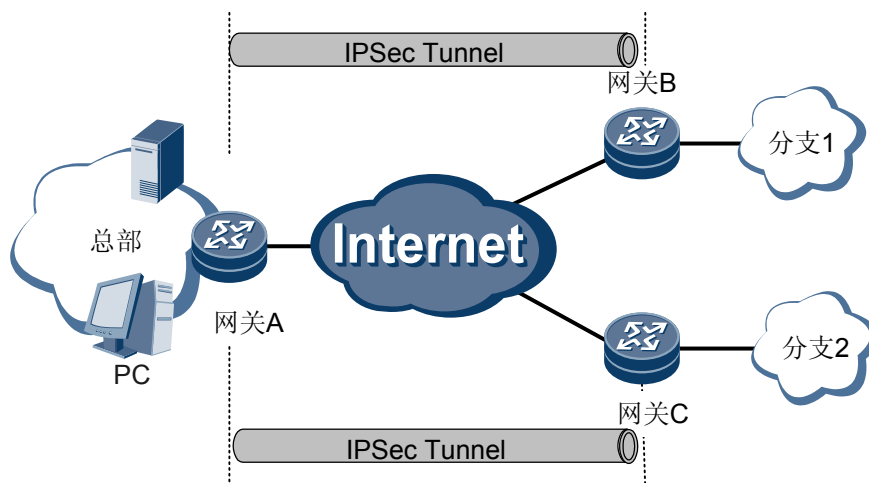
GRE over IPSec VPN 可以提供在总部和分支之间传送广播、组播的业务，例如视频会议或动态路由协议消息等等。

## 点到多点 VPN（Hub-Spoke VPN）

实际组网中最常见的是公司总部与多个分支机构通过点到多点 IPSec VPN 互通，典型组网如图 10-6 所示。

在这种组网中各个分支的 IP 地址可以配置为固定公网地址，也可以通过 3G、ADSL、PPPoE 拨号或 DHCP 方式动态获取。请根据实际需求配置 IPSec、L2TP over IPSec 或 GRE over IPSec。如图 10-6 所示。

图 10-6 总部与分支网关之间的点到多点通信组网

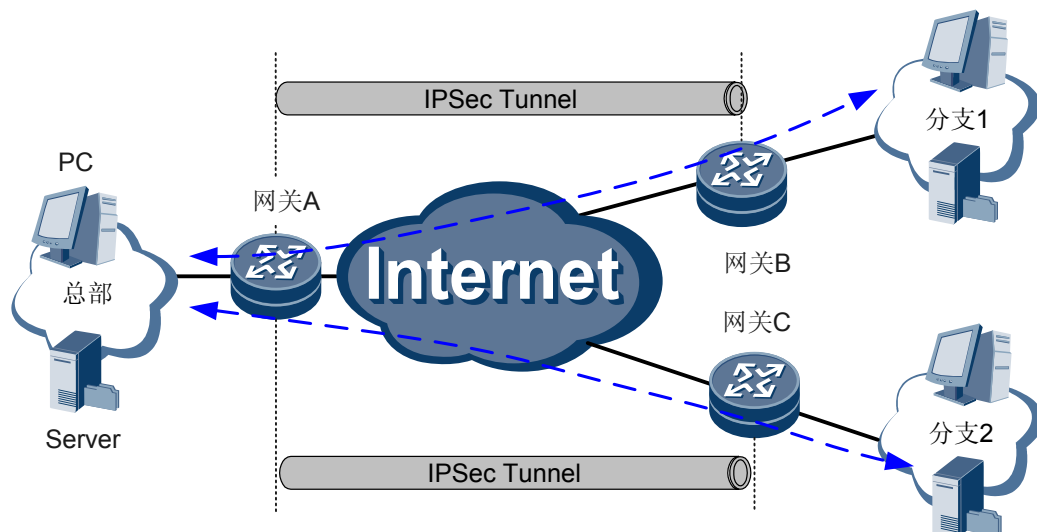


此时网络内数据流量可能存在如下几种情况：

- 各分支机构之间不需要通信

只有总部和分支之间部署 IPsec VPN，也只有总部和分支之间存在业务流量。如图 10-7 所示。

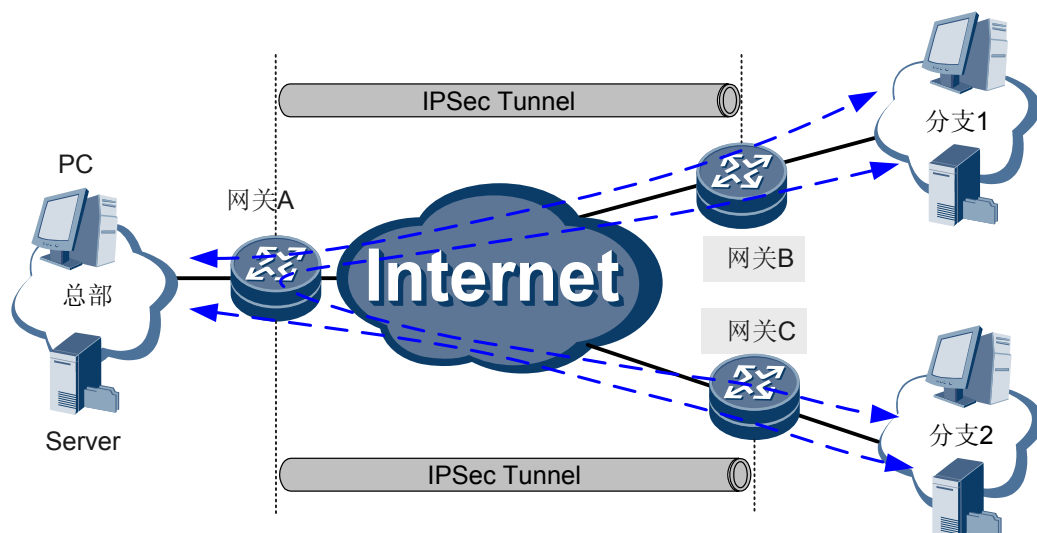
图 10-7 Hub-Spoke VPN（1）



- 各分支机构之间需要通信

分支机构之间通过总部进行通信。根据分支机构之间是否需要进行流量过滤，将分支结构配置在同一安全区域或不同的安全区域下。如图 10-8 所示。

图 10-8 Hub-Spoke VPN（2）



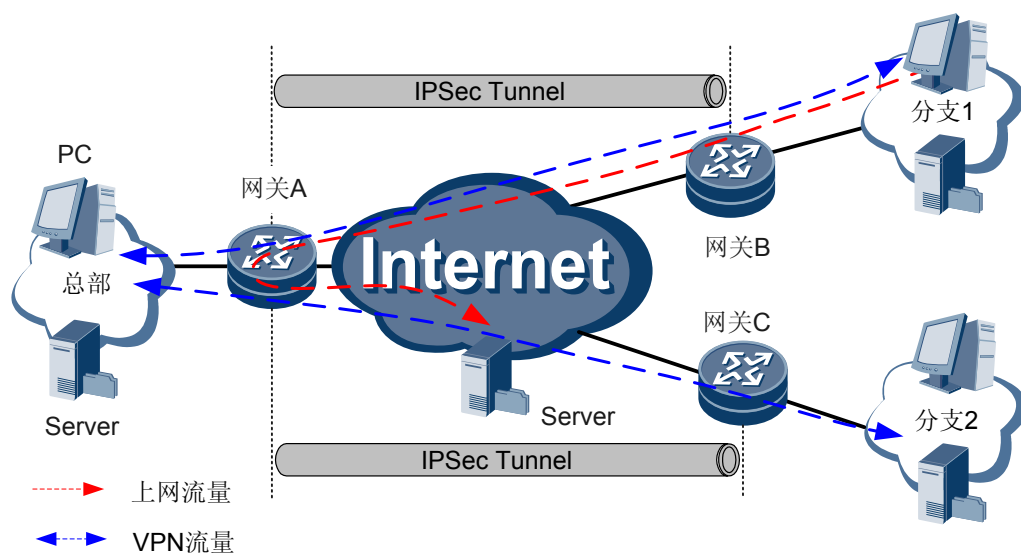
## 分支机构用户上网控制

在点到点或点到多点 VPN 中，分支机构用户上网可以通过两种方式来实现。

- 分支机构用户通过总部接入 Internet

为进行统一管理和监控，分支机构用户不允许通过自身的网关接入 Internet，只能通过 VPN 接入到总部，然后通过总部访问 Internet。分支访问 Internet 的流量一般需要在总部做 NAT 才能到公网。如图 10-9 所示。

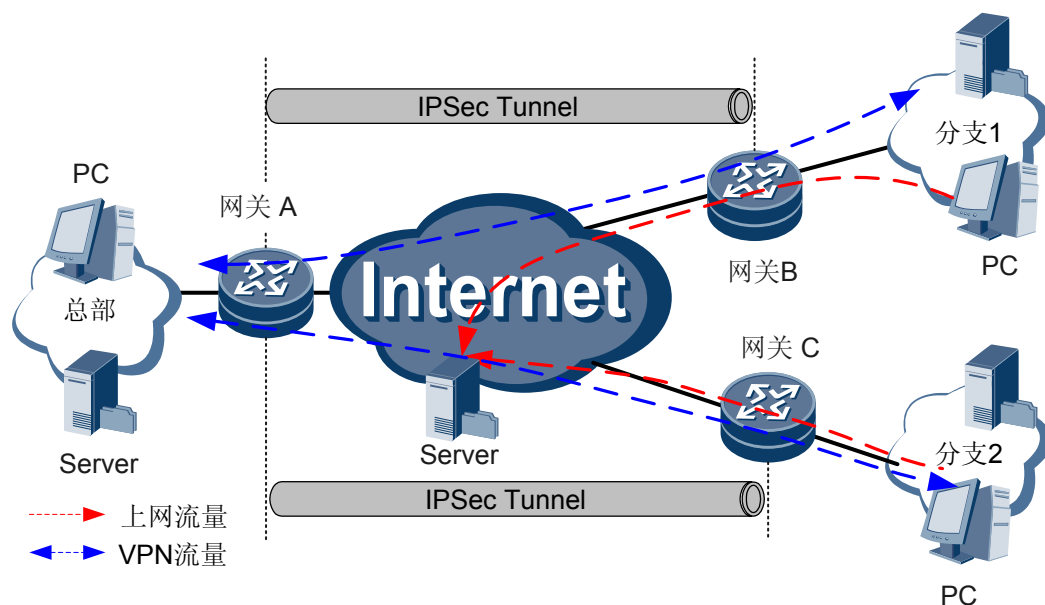
图 10-9 分支机构用户通过总部接入 Internet



- 分支机构用户自行接入 Internet

对分支机构的上网流量不做控制，如图 10-10 所示。

图 10-10 分支机构用户自行接入 Internet



#### 10.2.4.2 移动用户远程接入 VPN

移动设备的 IP 地址不固定，当其远程访问总部网络时，为避免不安全网络设备的攻击，要在拨入用户和总部网关之间建立一条 IPsec 安全隧道，且总部网关需要对拨入用户进行认证，认证通过的用户方能访问总部网络。

移动用户远程接入总部 VPN 有两种实现方式。

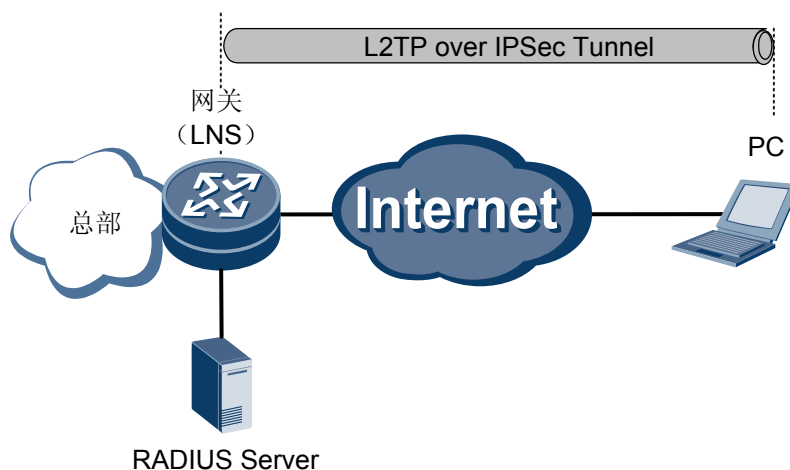
#### L2TP over IPsec

移动设备通过 Windows 自带拨号软件或其他拨号软件进行拨号。本公司提供 VPN Client 软件安装简单，使用方便。

如图 10-11 所示，在 USG 和 PC 之间建立 L2TP over IPsec 隧道。在总部部署 AAA 服务器（以 RADIUS 服务器为例），对接入的用户进行身份认证，认证通过后由网关为 PC 分配总部内网的 IP 地址。



图 10-11 移动用户远程接入总部（L2TP over IPSec）

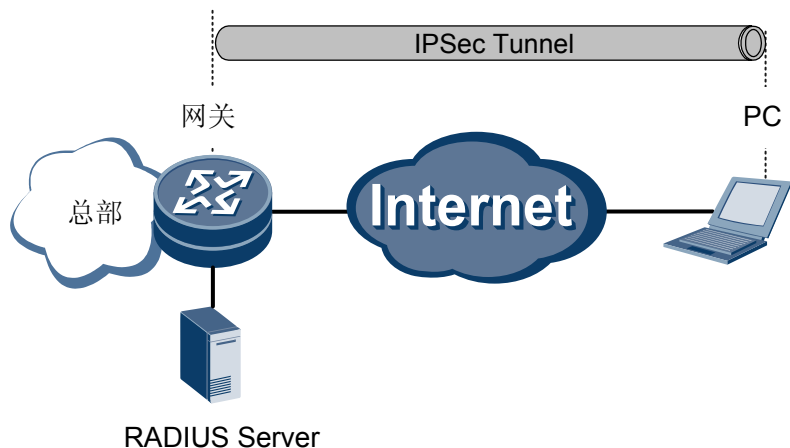


## IPSec(采用 IKEv2)

移动设备通过支持 IKEv2 的拨号软件进行拨号。IKEv2 支持 EAP 扩展认证，用户通过 EAP 方式拨号。

如图 10-12 所示，在 USG 和 PC 之间建立 IPSec 隧道。在总部部署支持 EAP 认证的 AAA 服务器（以 RADIUS 服务器为例），对 PC 进行身份认证，认证通过后由 IPSec 网关为 PC 分配总部内网的 IP 地址。

图 10-12 移动用户远程接入总部（IPSec+IKEv2）



### 10.2.4.3 IPSec 冗余设计

介绍提高 IPSec 可靠性常用的几种措施。

为了提高 VPN 网络的可靠性可以采用以下几种技术：

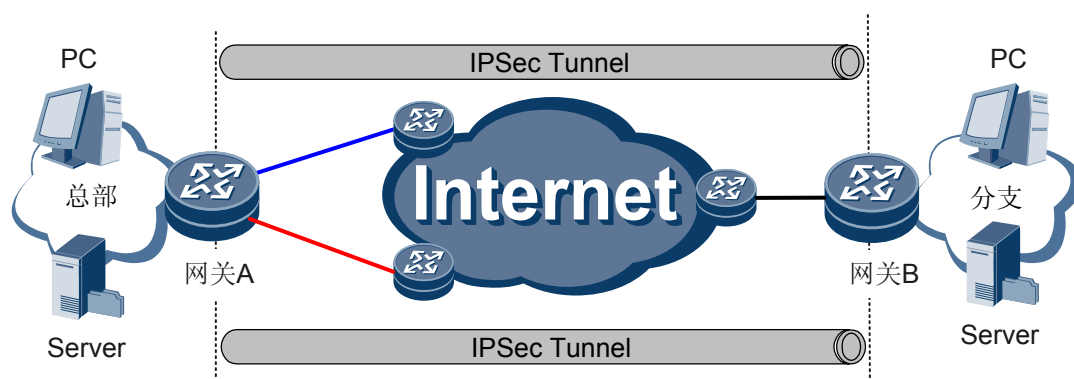
- 链路冗余
  - 主备链路备份：一般情况下，主用 IPSec 隧道采用以太网链路固定 IP 接入，备用 IPSec 隧道采用以太网链路或拨号链路（PPPoE/ADSL/3G）接入均可。主用链路故障时业务倒换到备用链路。

- 隧道化链路备份：将 IPSec 策略应用到 Tunnel 接口上，IPSec 策略跟具体的物理接口没有绑定关系，从而实现出接口链路的备份。当一条链路出现问题时，可直接路由到其他链路传输。
  - 全网状网络：每个设备都与其它设备之间建立 IPSec 隧道，这是最为安全的网络结构，但需要较多的资金投入。
- 设备冗余  
IPSec VPN 网关采用主备备份机制，当一台设备出现故障时，业务可以平滑的切换到备用设备上。

## 主备链路备份

如图 10-13 所示，网关 A 通过主备两条链路连接网关 B。在网关 A 的两个物理接口分别应用 IPSec 策略，可以创建主备两条 IPSec 隧道。当主链路故障时，将在备份链路上建立新的 IPSec 隧道，并拆除旧的 IPSec 隧道，完成流量和隧道的切换。

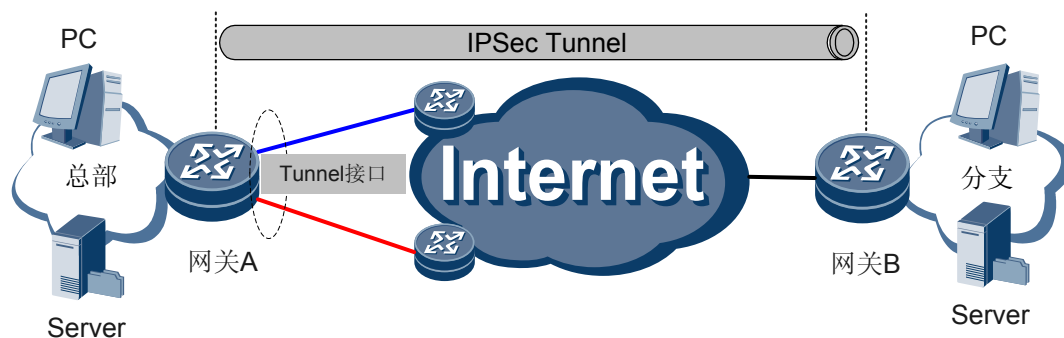
图 10-13 主备链路备份



## 隧道化链路备份

如图 10-14 所示，网关 A 通过主备两条链路连接网关 B。在网关 A 的 Tunnel 接口和网关 B 的物理接口之间创建 IPSec 隧道，VPN 流量通过 Tunnel 接口进行 IPSec 处理，然后通过路由表选择物理接口发送。当主物理链路失效时，其路由变为不可达，流量自然切换到备用链路。这种情况下，IPSec 隧道不需要进行重协商，故可快速完成流量切换。

图 10-14 IPSec VPN 隧道化链路备份

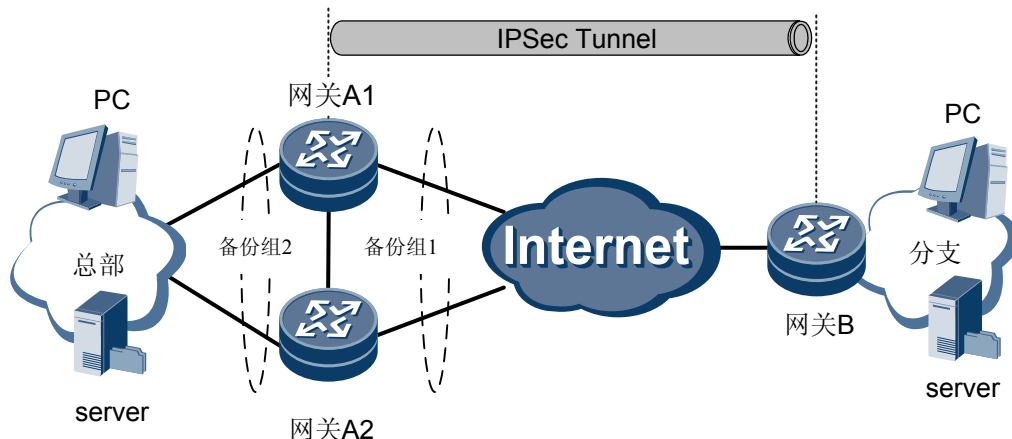


隧道化链路备份可以实现多条链路的备份，而且与主备链路备份相比，配置更简单，流量切换速度更快。

## IPSec 网关主备备份

如图 10-15 所示，网关 A1 和 A2 对外配置虚拟接口 1，并在虚拟接口 1 和分支网关 B 的物理接口之间建立 IPSec 隧道。当主用网关 A1 物理接口、链路或主机故障时，流量被引导到备用网关 A2 进行 IPSec 和转发处理。这种情况下，原有的 IPSec 隧道并不会被拆除，切换速度更具优势。

图 10-15 IPSec 网关主备备份

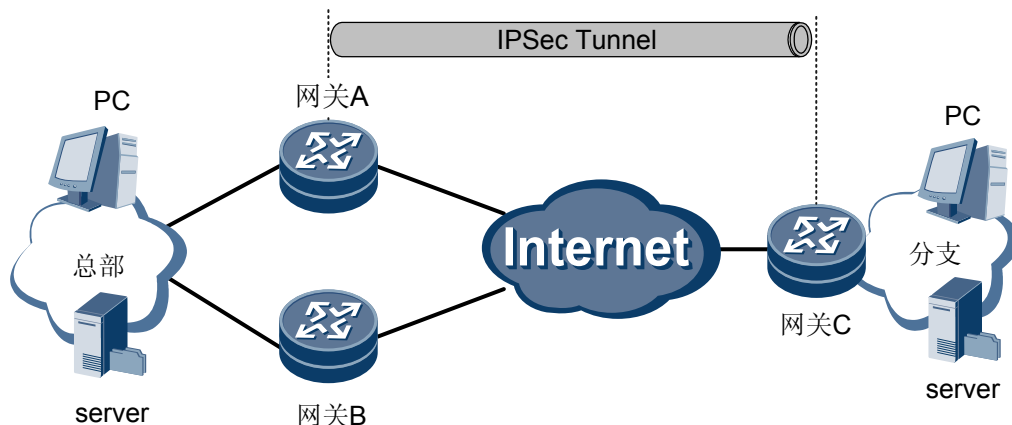


此种场景下，要求两台网关的软硬件配置完全一致，包括硬件插板的数量、位置，硬件版本和驱动软件版本、主机软件版本等。

## IKE Peer 备份

如图 10-16 所示，公司异地部署两台接入设备进行容灾备份，两台设备连接的内网网段相同。当其中一台设备连接不通后（如机房断电等），另一台设备可作为备份供分支接入。分支设备的安全策略中配置两个 IKE Peer，其 IP 地址分别对应总部的两台设备。当分支发现总部主接入设备不通（隧道建立失败）后，切换 IPSec 策略的 PEER 为公司的备份接入设备，以建立新的隧道。

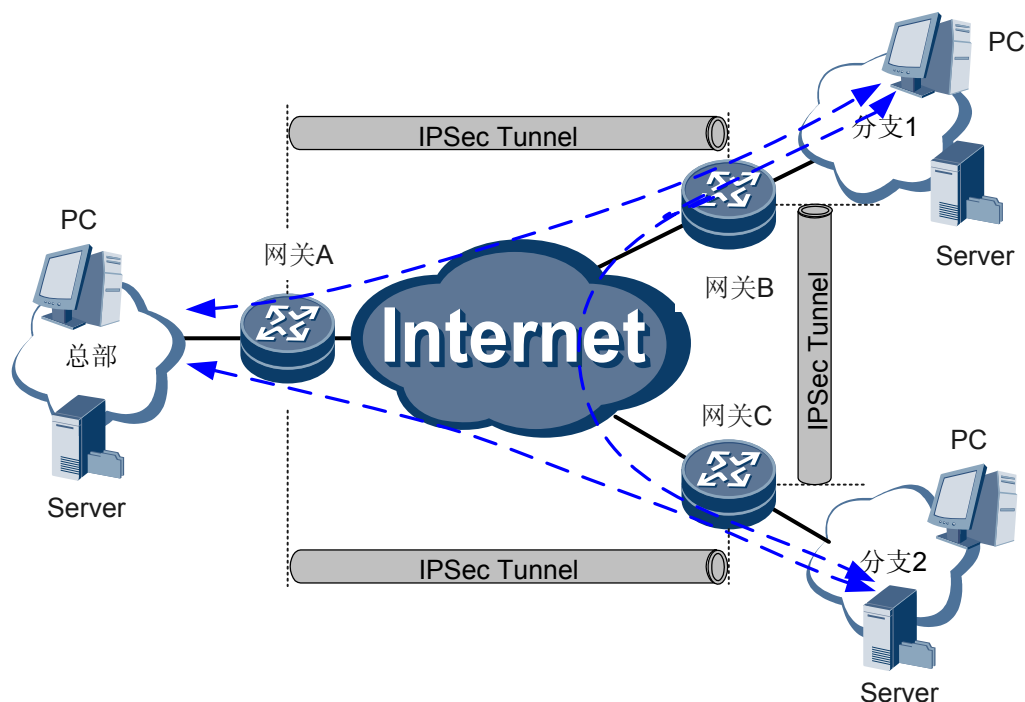
图 10-16 IKE Peer 备份



## 全网状 VPN (Full-Mesh)

如图 10-17 所示，当网络中的所有节点之间都需要通信或者通信流量较大时，可采用网状网结构建立 VPN。所有节点都可以自行上网，任何节点 Down 掉都不影响其它节点。这种网络结构适用于多个重要且地位均等的节点之间的互联。

图 10-17 Full-Mesh VPN

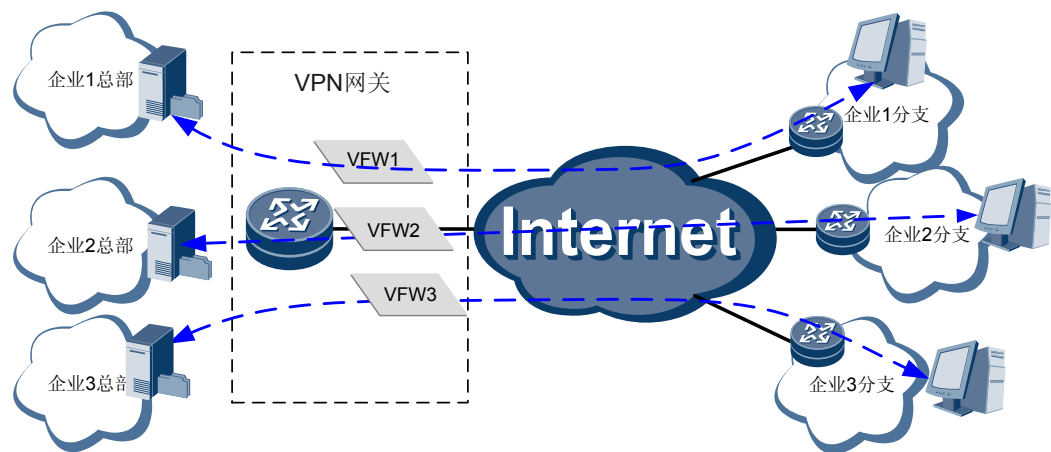


### 10.2.4.4 IPSec 多实例应用

IPSec 多实例主要用于向小企业提供防火墙租赁业务和实现企业内部网络隔离。

如图 10-18 所示，三个企业分支共用一台 VPN 网关。三个企业之间需要保持网络独立，互相之间不会产生影响。三个企业分支内部网络 IP 地址是各自规划的，可能出现 IP 地址空间重叠（即不同私网出现使用相同 IP 地址的情况），此时需要在网关上应用 IPSec 多实例功能，将三个企业的 IPSec 隧道与三个 IPSec 多实例绑定，保证相同 IP 地址的报文能够正确转发。

图 10-18 IPSec 多实例应用



10.2.5 IPSec 协议框架

IPSec 是 Internet 工程任务组(IETF)制定的一个开放的网络层安全框架协议。它并不是一个单独的协议，而是一系列为 IP 网络提供安全性的协议和服务的集合。

10.2.5.1 协议框架简介

IPSec 是在 IETF RFC 中定义的一个框架标准，是一系列为 IP 网络提供安全性的协议的集合。

IPSec 协议框架如图 10-19 所示。

图 10-19 IPSec 协议框架

|      |                  |       |     |     |       |
|------|------------------|-------|-----|-----|-------|
| 安全协议 | ESP              |       |     | AH  |       |
| 加密   | DES              | 3DES  | AES |     |       |
| 验证   | MD5              | SHA-1 |     | MD5 | SHA-1 |
| 密钥交换 | IKE (ISAKMP, DH) |       |     |     |       |

IPSec 通过 AH（Authentication Header，验证头）和 ESP（Encapsulating Security Payload，封装安全载荷）两个安全协议实现 IP 报文的封装/解封装。

- AH 是报文头验证协议，主要提供数据源验证、数据完整性验证和防报文重放功能，不提供加密功能。
- ESP 是封装安全载荷协议，主要提供加密、数据源验证、数据完整性验证和防报文重放功能。

AH 和 ESP 协议提供的安全功能依赖于协议采用的加密、验证算法。

- ESP 能够对原 IP 报文内容进行加密，使用的加密算法为对称加密算法，包括 DES（Data Encryption Standard）、3DES（Triple Data Encryption Standard）、AES（Advanced Encryption Standard）。三种算法的安全性依次递增。
- AH 和 ESP 都能够提供数据源验证和数据完整性验证，使用的验证算法为 MD5（Message Digest 5）、SHA-1（Secure Hash Algorithm）。

IPSec 网关的加密和验证算法所使用的密钥可以手工配置，也可以动态协商。为了提升密钥的安全性，降低 IPSec VPN 管理复杂度，IPSec 协议框架中引入 IKE 协议实现安全联盟动态协商和密钥管理功能。

IKE 协议建立在 Internet 安全联盟和密钥管理协议 ISAKMP（Internet Security Association and Key Management Protocol）框架之上，采用 DH（Diffie-Hellman）算法在不安全的网络上安全地分发密钥、验证身份，以保证数据传输的安全性。

10.2.5.2 安全协议

IPSec 使用安全协议来传输和封装用户数据，提供安全服务。

虽然 AH 协议和 ESP 协议都可以提供数据源验证和数据完整性校验服务，但是 ESP 协议并无法取代 AH 协议。两者之间的差别在于验证报文的范围不同，验证范围请参见 [10.2.5.3 封装模式](#)。

AH 和 ESP 协议的简单比较如[表 10-2](#) 所示。

表 10-2 AH 与 ESP 比较

| 安全特性                | AH             | ESP          |
|---------------------|----------------|--------------|
| IP 协议号              | 51             | 50           |
| 数据完整性校验             | 支持（验证整个 IP 报文） | 支持（不验证 IP 头） |
| 数据源验证               | 支持             | 支持           |
| 数据加密                | 不支持            | 支持           |
| 防报文重放攻击             | 支持             | 支持           |
| IPSec NAT-T（NAT 穿越） | 不支持            | 支持           |

从上表中可以看出两个协议各有优缺点，AH 协议不提供数据加密功能，ESP 的验证范围不包括 IP 头，故在安全性要求较高的场景中可以考虑联合使用 AH 协议和 ESP 协议。

AH 协议

AH 是一种 IP 协议，IP 协议号为 51。AH 头部结构如[图 10-20](#) 所示。

图 10-20 AH 头部结构



下面对 AH 头部结构进行说明：

- 下一头部：8 比特，标识 AH 头后面的负载类型。传输模式下，是被保护的上层协议（TCP 或 UDP）或 ESP 的编号；隧道模式下，是 IP 协议或 ESP 协议的编号。

 说明

当 AH 与 ESP 协议同时使用时，AH 头的下一头部为 ESP 头。

- 负载长度：8 比特，表示以 32 比特为单位的 AH 头部长度减 2，缺省为 4。
- 保留字段：16 比特，保留将来使用，缺省为 0。
- SPI：32 比特，用于标识有相同 IP 地址和相同安全协议的不同 IPsec 安全联盟（IPsec SA）。

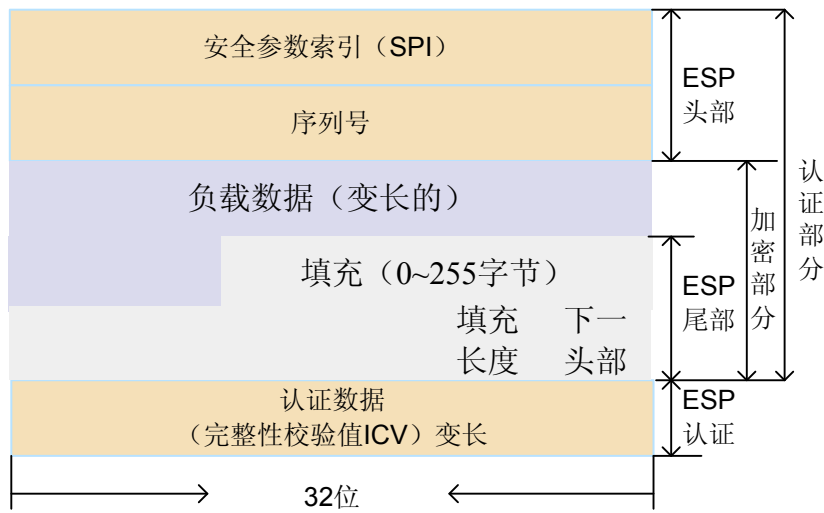
IPsec 安全联盟是用来传输数据的安全连接。详细说明请参见 [10.2.6 IPsec 工作原理](#)。

- 序列号：32 比特，是一个从 1 开始的单项递增的计数器，唯一地标识每一个数据包，用于防止重放攻击。
- 认证数据：一个变长字段，长度为 32 比特的整数倍，通常为 96 比特。该字段包含数据完整性校验值 ICV（Integrity Check Value），用于接收方进行完整性校验，校验过程请参见 [10.2.5.5 验证](#)。

## ESP 协议

ESP 是一种 IP 协议，IP 协议号为 50。ESP 头部结构如 [图 10-21](#) 所示。

图 10-21 ESP 头部结构



下面对 ESP 头部结构进行说明：

- **SPI:** 32 比特，用于标识有相同 IP 地址和相同安全协议的不同 IPSec SA。
- **序列号:** 32 比特，是一个从 1 开始的单项递增的计数器，唯一地标识每一个数据包，用于防止重放攻击。
- **负载数据:** 包含由下一头部字段给出的变长数据。
- **填充字段:** 用于增加 ESP 包头的位数。填充字段的长度与负载数据的长度和算法有关。大多数加密算法要求输入数据包含整数个分组，因此需要填充。
- **填充长度:** 8 比特，给出前面填充字段的长度，置 0 时表示没有填充。
- **下一头部:** 8 比特，标识 ESP 头后面的下一个负载类型。传输模式下，是被保护的上层协议（TCP 或 UDP）的编号；隧道模式下，是 IP 协议的编号。
- **认证数据:** 该字段包含数据完整性校验值 ICV，用于接收方进行完整性校验，校验过程请参见 [10.2.5.5 验证](#)。

ESP 的验证功能是可选的，如果启动了数据包验证，会在加密数据的尾部添加一个 ICV 数值。

10.2.5.3 封装模式

针对不同应用场景，AH 和 ESP 协议可以采用两种封装模式封装数据，故在深入了解 AH 和 ESP 协议前先介绍一下安全协议的封装模式。

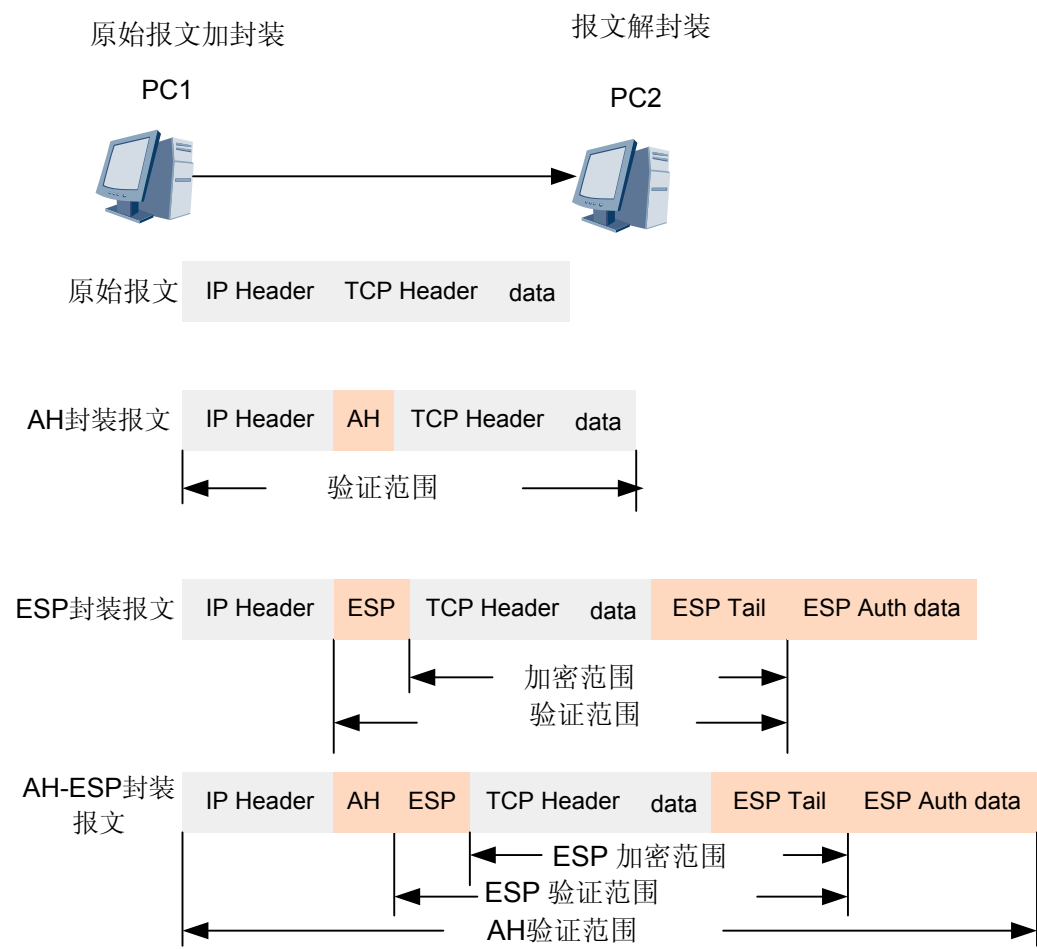
传输模式

在传输模式中，AH 头或 ESP 头被插入到 IP 头与传输层协议头之间，保护 TCP/UDP/ICMP 负载。传输模式不改变报文头，故隧道的源和目的地址必须为 IP 报文头中的源和目的地址，所以只适合两台主机之间通信。

以 TCP 报文为例，原始报文经过传输模式封装后，报文格式如 [图 10-22](#) 所示。



图 10-22 传输模式下报文封装



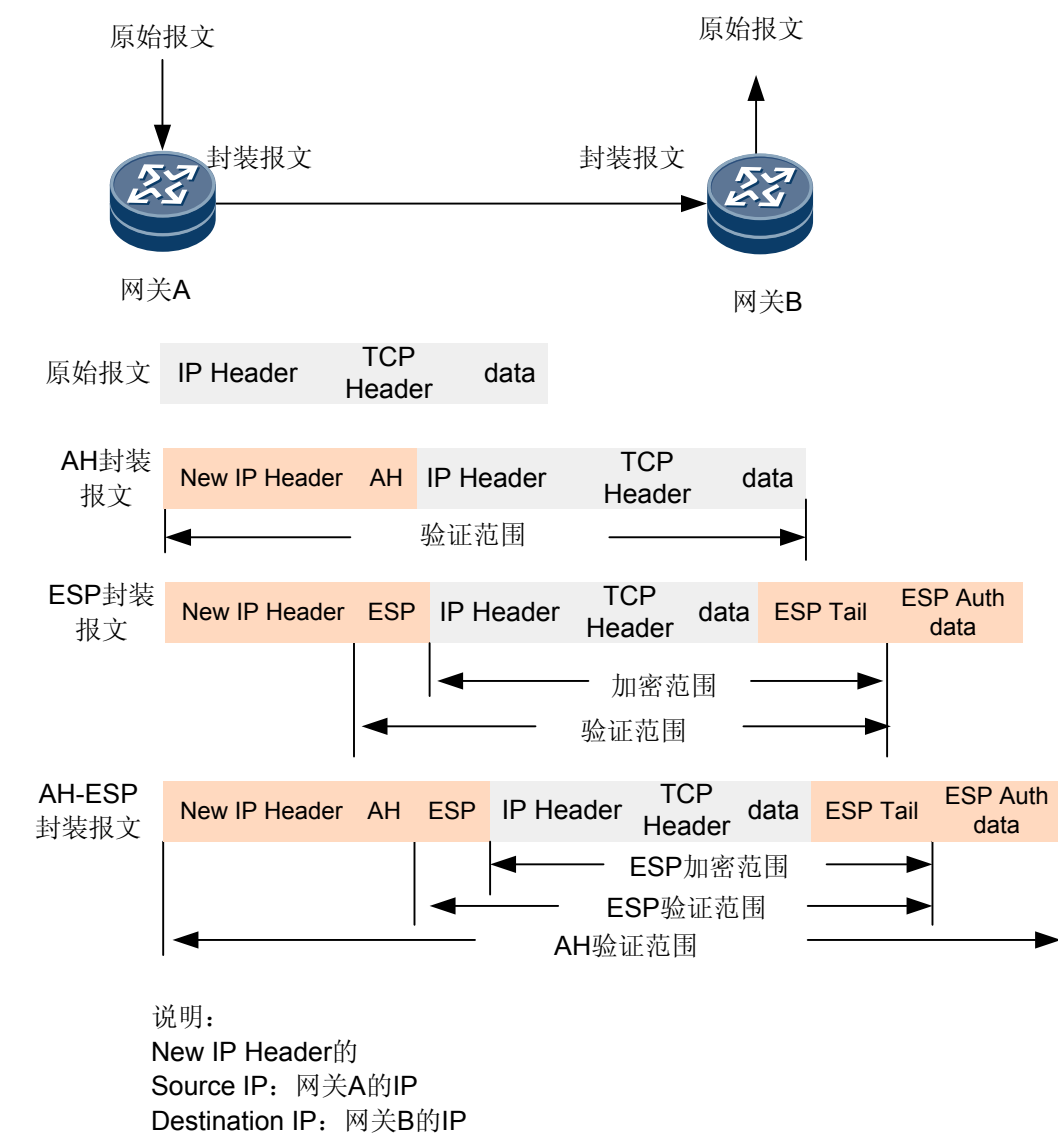
传输模式下，AH 协议的完整性验证范围为整个 IP 报文。ESP 协议验证报文的完整性检查部分包括 ESP 头、传输层协议头、数据和 ESP 报尾，但不包括 IP 头，因此 ESP 协议无法保证 IP 头的安全。ESP 的加密部分包括传输层协议头、数据和 ESP 报尾。

隧道模式

在隧道模式下，AH 头或 ESP 头被插到原始 IP 头之前，另外生成一个新的报文头放到 AH 头或 ESP 头之前，保护 IP 头和负载。隧道模式主要应用于两台 VPN 网关之间或一台主机与一台 VPN 网关之间的通信。

以 TCP 报文为例，原始报文经隧道模式封装后的报文结构如图 10-23 所示。

图 10-23 隧道模式下报文封装



隧道模式下，AH 协议的完整性验证范围为包括新增 IP 头在内的整个 IP 报文。ESP 协议验证报文的完整性检查部分包括 ESP 包头、原 IP 头、传输层协议头、数据和 ESP 报尾，但不包括新 IP 头，因此 ESP 协议无法保证新 IP 头的安全。ESP 的加密部分包括传输层协议头、数据和 ESP 报尾。

当安全协议同时采用 AH 和 ESP 时，AH 和 ESP 协议必须采用相同的封装模式。

传输模式和隧道模式比较

传输模式和隧道模式比较：

- 从安全性来讲，隧道模式优于传输模式。它可以完全地对原始 IP 数据报进行验证和加密。隧道模式下可以隐藏内部 IP 地址，协议类型和端口。

- 从性能来讲，隧道模式因为有一个额外的 IP 头，所以它将比传输模式占用更多带宽。

## 10.2.5.4 加密

加密是一种将数据从明文转换成无法读懂的密文的过程，接收方只有在拥有正确的密钥的情况下才能对密文进行解密，从而保证了数据的私密性。

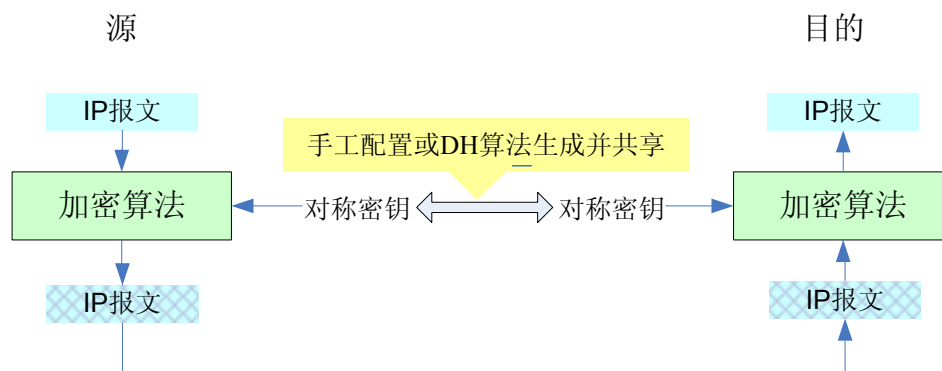
IPSec VPN 工作过程中涉及数据加密（IP 报文加密）和协议消息加密（ISAKMP 消息加密）两种情况。

## 数据加密

IPSec 采用对称加密算法对数据进行加密和解密。对称加密算法是指数据发送方和接收方使用相同的密钥进行加密、解密。

采用对称加密算法进行数据加密和解密的过程如图 10-24 所示。

图 10-24 加密和解密的过程



用于加密的对称密钥可以手工配置，也可以通过 DH 算法生成并在两端设备共享。有关 DH 算法具体能够生成哪些密钥及密钥的作用请参见 10.2.7.1 IKEv1 协商阶段 1。

常用的对称加密算法包括：

- 数据加密标准 DES（Data Encryption Standard）  
DES 是由美国国家标准与技术研究院（NIST）开发的。它使用 56 位的密钥对一个 64 位的明文块进行加密。
- 3DES（Triple Data Encryption Standard）  
3DES 是一种增强型的 DES 标准，它在需要保护的数据上使用 3 次 DES，即使用三个不同的 56 位的 DES 密钥（共 168bit 密钥）对明文进行加密。  
3DES 与 DES 相比，3DES 具有更高的安全性，但其加密数据的速度要比 DES 慢得多。
- 先进加密标准 AES（Advanced Encryption Standard）  
AES 被设计用来替代 3DES，提供更快和更安全的加密功能。AES 可以采用三种密钥，其密钥的长度分为 128bit、192bit、256bit。  
一般情况下 128bit 就可以充分满足安全需求。

## 协议消息加密

协议消息加密用于 IKE 协商阶段。协议消息加密所用的算法也是 DES、3DES 和 AES。用于加密的对称密钥通过 DH 算法生成。有关 DH 算法具体能够生成哪些密钥及密钥的作用请参见 [10.2.7.1 IKEv1 协商阶段 1](#)。

### 10.2.5.5 验证

IPSec 数据验证包括完整性验证和真实性验证两部分。

IPSec 采用 HMAC（Keyed-Hash Message Authentication Code）功能进行验证。HMAC 利用 Hash 函数，以一个对称密钥和一个数据包作为输入，生成一个固定长度的输出，这个输出被称为完整性校验值 ICV（Integrity Check Value）。接收方通过比较自身生成的 ICV 和对端发送的 ICV 来判断数据的完整性和真实性。

数据源验证和数据完整性验证统一被称为验证。因为这两种安全服务总是绑定在一起提供的。数据完整性验证是基于每个 IP 数据包进行计算；将完整性验证密钥同 IPSec 对端身份绑定的结果间接提供了数据源验证。

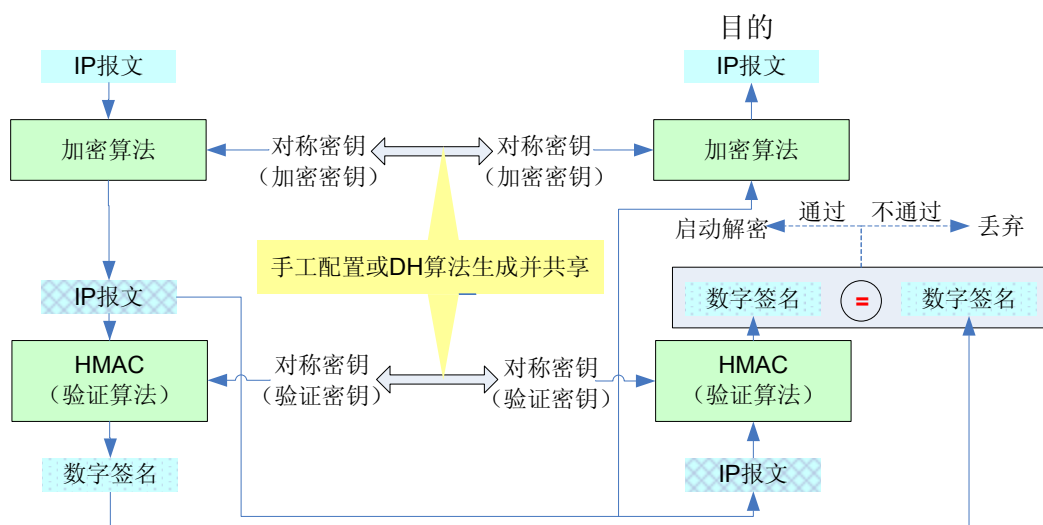
虽然加密后的数据只能通过原始的加密密钥进行解密，但是我们无法验证解密后的信息是否是原始发送的信息。另外加密和解密的过程非常的消耗 CPU，恶意用户可能会通过发送欺骗数据包，占用 CPU 资源。HMAC 功能通过比较数字签名进行数据包完整性和真实性验证，这个过程消耗的 CPU 资源非常少，效率非常高。

故在 IPSec VPN 发送设备中，加密和验证通常配合使用。加密后的报文经 HMAC 生成数字签名，IP 报文和数字签名同时发给对端（数字签名填写在 AH 和 ESP 报文头的完整性校验值 ICV 字段，请参见 [10.2.5.2 安全协议](#)）；在接收设备中，通过比较数字签名进行数据完整性和真实性验证，验证不通过的报文直接丢弃，验证通过的报文再进行解密。

用于验证的对称密钥可以手工配置，也可以通过 DH 算法生成并在两端设备共享。有关 DH 算法具体能够生成哪些密钥及密钥的作用请参见 [10.2.7.1 IKEv1 协商阶段 1](#)。

加密和 HMAC 验证配合使用的过程如图 10-25 所示。

图 10-25 HMAC 验证过程



常用的 hash 函数包括：

- MD5（Message Digest 5）  
消息摘要 5（MD5）在 RFC 1321 中有明文规定。输入任意长度的消息，MD5 产生 128 位的签名。  
MD5 比 SHA 更快，但是安全性稍差。
- SHA-1（Secure Hash Algorithm）  
安全的散列算法（SHA）是由 NIST 开发的。在 1994 年对原始的 HMAC 功能进行了修订，被称为 SHA-1。SHA-1 在 RFC 2404 中描述。输入长度小于  $2^{64}$ bit 的消息，SHA-1 产生 160 位的消息摘要。  
SHA-1 比 MD5 要慢，但是更安全。因为它的签名比较长，可以更有效的防止强大的攻破功能并发现共享的密钥。

### 10.2.5.6 密钥交换

IKEv1 和 IKEv2 中，隧道两端需要进行密钥材料的交换，以便使用相同密钥正确的加密和解密。

### 密钥交换方法

使用对称密钥进行加密、验证时，如何安全地共享密钥是一个很重要的问题。有两种方法解决这个问题：

- 带外共享密钥  
在发送、接收设备上手工配置静态的加密、验证密钥。双方通过带外共享的方式（例如通过电话或邮件方式）保证密钥一致性。这种方式的缺点是可扩展性差，在点到多点组网中配置密钥的工作量成倍增加。另外，为提升网络安全需要周期性修改密钥，这种方式下也很难实施。
- 使用一个安全的连接分发密钥  
IPSec 使用 IKE 协议在发送、接收设备之间安全地协商密钥。IKE 采用 DH 算法在不安全的网络上安全地交换密钥信息，并生成加密、验证密钥。这种方式配置简单，可扩展性好，特别是在大型动态的网络环境下此优点更加突出。

### IKEv1 协议

IKE 提供密钥交换，自动协商建立安全联盟等服务。采用 IKE 协议可以使 IPSec 配置和管理更简单、更灵活。

- Internet 安全联盟和密钥管理协议 ISAKMP（Internet Security Association and Key Management Protocol）是 IKE 的基础，IKE 使用 ISAKMP 协议定义密钥交换的过程。ISAKMP 提供了对安全服务进行协商的方法，密钥交换时交换信息的方法，以及对对等体身份进行验证的方法。
- IKE 的精髓在于它永远不在不安全的网络上传送密钥，而是通过一些数据的交换，通信双方最终计算出共享的密钥，并且即使第三方截获了双方用于计算密钥的所有交换数据，也无法计算出真正的密钥。其中的核心技术就是 DH（Diffie Hellman）交换技术。

## DH

DH 被称为公共密钥交换方法，它用于产生密钥材料，并通过 ISAKMP 消息在发送和接收设备之间进行密钥材料交换。然后，两端设备各自计算出完全相同的对称密钥。该对称密钥用于加密和验证密钥的计算。在任何时候，双方都不交换真正的密钥。

MD5、SHA-1、DES、3DES、AES 等算法都采用 DH 方法来共享对称密钥。

DH 使用密钥组来定义自己产生的密钥长度。密钥长度越长，产生的密钥就越强壮。

表 10-3 DH 密钥组

| 密钥组 | 长度     |
|-----|--------|
| 1   | 768 位  |
| 2   | 1024 位 |
| 5   | 1536 位 |

## IKEv2 协议

IKEv1 协议是一个混合型协议，其自身的复杂性不可避免地带来一些安全及性能上的缺陷，已经成为目前实现的 IPSec 系统的瓶颈。IKEv2 协议保留了 IKEv1 的基本功能，并针对 IKEv1 研究过程中发现的问题进行修订，同时兼顾简洁性、高效性、安全性和健壮性的需要，整合了 IKEv1 的相关文档，由 RFC4306 单个文档替代。通过核心功能和默认密码算法的最小化规定，新协议极大地提高了不同 IPSec VPN 系统的互操作性。

IKEv2 与 IKEv1 相比有以下优点：

- 简化了安全联盟的协商过程，提高了协商效率。
- 修复了多处公认的密码学方面的安全漏洞，提高了安全性能。
- 加入对 EAP（Extensible Authentication Protocol）身份认证方式的支持，提高了认证方式的灵活性和可扩展性。

EAP 是一种支持多种认证方法的认证协议，可扩展性是其最大的优点，即若想加入新的认证方式，可以像组件一样加入，而不用变动原来的认证体系。当前 EAP 认证已经广泛应用于拨号接入网络中。

## 10.2.6 IPSec 工作原理

本节介绍 IPSec 的基本概念和工作原理。

在介绍 IPSec 的工作原理之前先介绍一下 IPSec 的重要概念：

- IPSec 安全联盟（IPSec SA）

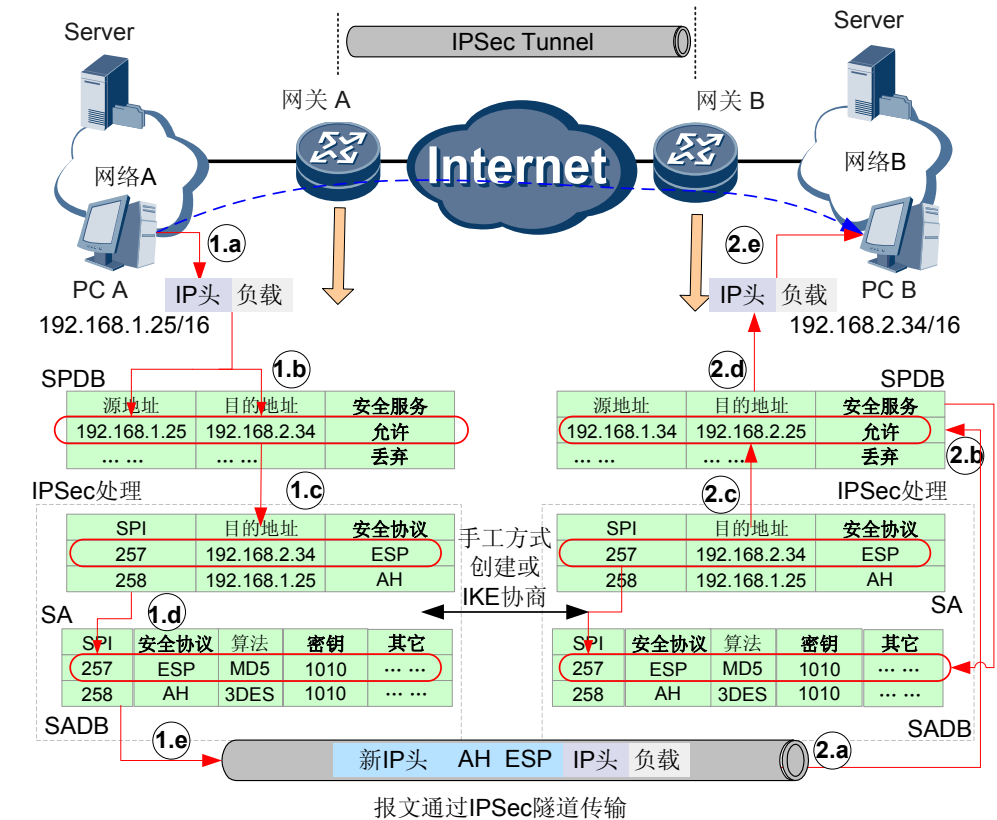
通过 IPSec VPN 传输数据的前提是在 IPSec 对等体（IPSec 协议的两个端点）之间成功建立 IPSec SA。

IPSec SA 是为安全目的创建的单向逻辑连接，这就要求在每一个方向上都需要拥有一个 IPSec SA。另外 SA 的个数还与安全协议相关。如果您只使用 AH 或 ESP 来保护两个对等体之间的流量，则对等体之间就有两个 SA，每个方向上一个。如果对等体同时使用了 AH 和 ESP，那么对等体之间就需要四个 SA，每个方向上两个，分别对应 AH 和 ESP。因此不能将 IPSec SA 等同于一个“连接”。

- IPsec SA 分类  
IPsec SA 有两种类型：手工方式创建的 SA 和 IKE 自动协商方式（isakmp）创建的 SA。二者的主要区别为：
  - 密钥生成方式不同  
手工方式下，创建 IPsec SA 所需的全部参数，包括加密、验证密钥，都需要用户手工配置，也只能手工刷新。  
IKE 方式下，创建 IPsec SA 需要的加密、验证密钥是通过 DH 算法生成的，可以动态刷新。密钥管理成本低，安全性较高。
  - IPsec SA 生存周期不同  
手工方式创建的 IPsec SA 一经创建永久存在。  
IKE 方式下，由数据流触发建立 IPsec SA，SA 的生存周期由双方配置的生存周期参数控制。
- SPI  
SPI 是为唯一标识 SA 而配置或生成的一个 32bit 的数值，它在 AH 和 ESP 头中传输。SPI 和目的 IP 地址、安全协议号一起，唯一标识一个 IPsec SA。
- SPDB 和 SADB  
安全策略数据库 SPDB（Security Policy Database）是 IPsec SA 创建的前提，它定义哪些数据流受安全协议保护。安全关联数据库 SADB（Security Associations Database）用于存放 IPsec SA 的所有属性参数。

下面以点到点 VPN 单向发送数据（采用隧道模式）为例，对 IPsec VPN 的工作原理进行详细介绍。

图 10-26 IPsec VPN 工作原理



1. 网关 A:
  - a. PC A 向网关 A 发送 IP 报文（目的地址为 PC B）。
  - b. 网关 A 查找 SPDB，根据报文的源目的地址确定是否对数据流进行 IPSec 处理。
    - 动作为“允许”数据流需要进行 IPSec 处理。
    - 动作为“丢弃”的数据流不需要进行 IPSec 处理。
  - c. 根据数据流的目的地址查找 SA。
    - 若有相应的表项，则转到 1.4（手工方式下，配置完成后 SA 和 SADB 即已经生成）。
    - 若没有相应表项，则发起 IKE 协商。协商成功后，双方设备各自生成 SA 和 SADB。

使用 IKEv1 和 v2 版本的进行协商的过程不同，详细内容请参见 [10.2.7 IKEv1 协商安全联盟的过程](#)和 [10.2.8 IKEv2 协商安全联盟的过程](#)。
  - d. 根据 SPI 在 SADB 中找到相应 IPSec SA，根据 SA 的各项参数对数据流进行加密、Hash 及封装。
  - e. 物理接口发送 IPSec 报文。
2. 网关 B:
  - a. 物理接口接收 IPSec 报文。
  - b. （可选）对 IP 报文（未加密的明文）进行 IPSec 前反查。IPSec 反查功能缺省情况下开启，特殊情况下可以关闭。

若需要加密的报文没有加密则丢弃报文。
  - c. 根据 SPI 和数据流的目的 IP 地址、安全协议号查找 SADB 和 SA，根据 SA 的各项参数对数据流进行解封装、验证、解密。
  - d. （可选）对解密后的 IP 报文进行 IPSec 后反查。

若不需要加密的报文进行了加密则丢弃报文。
  - e. 将通过 IPSec 处理后的 IP 报文投递给 PC B。

## 10.2.7 IKEv1 协商安全联盟的过程

采用 IKEv1 协商安全联盟主要分为两个阶段。

### 10.2.7.1 IKEv1 协商阶段 1

IKEv1 阶段 1 的目的是建立 IKE SA。IKE SA 建立后对等体间的所有 ISAKMP 消息都将通过加密和验证，这条安全通道可以保证 IKEv1 阶段 2 的协商能够安全进行。

两个对等体间仅有一个 IKE SA，它是一个双向逻辑连接。

IKEv1 阶段 1 支持两种协商模式：

- 主模式（Main Mode）
- 野蛮模式（Aggressive Mode）

IKEv1 阶段 1 主要协商下面三个任务：

1. 协商建立 IKE SA 所使用的参数。



加密算法、完整性验证算法、身份认证方法和认证字、DH 组、IKE SA 生存周期等等。这些参数在 IKE 安全提议中定义。

2. 使用 DH 算法交换与密钥相关的信息（生成各种密钥的材料）。  
对等体双方设备能够使用这些密钥信息各自生成用于 ISAKMP 消息加密、验证的对称密钥。
3. 对等体之间验证彼此身份。  
使用预共享密钥或数字证书来验证设备身份。

这三个任务都协商成功后，IKE SA 就建立成功了。

主模式

主模式采用三个步骤来完成上述三个任务。每个步骤需要 2 个 ISAKMP 报文，共 6 个 ISAKMP 报文。交换密钥相关信息的操作在身份认证之前完成，故设备的身份信息受到已生成的共享密钥的保护。

采用主模式时 IKEv1 阶段 1 的协商过程如图 10-27 所示。

图 10-27 主模式下 IKEv1 阶段 1 的协商过程



下面对这三个步骤进行详细说明：

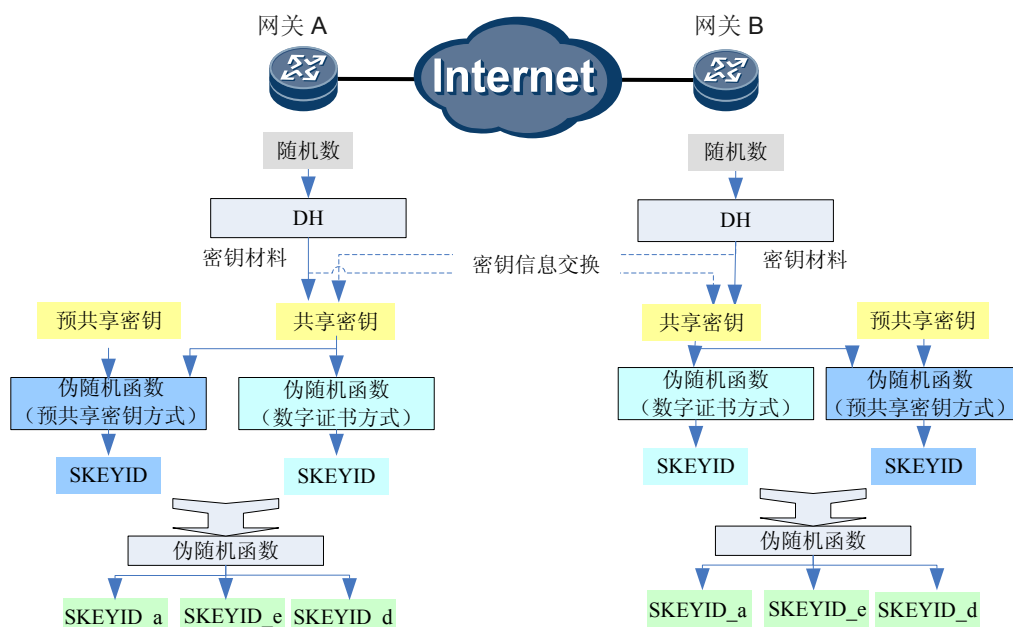
1. 协商对等体之间使用的 IKE 安全提议。
  - a. 网关 A 发送 ISAKMP 消息，携带建立 IKE SA 所使用的参数（由 IKE 安全提议定义）。
  - b. 网关 B 对网关 A 的 IKE 安全提议进行协商。

在协商时将从优先级最高的提议开始匹配，协商双方必须至少有一条匹配的 IKE 安全提议才能协商成功。匹配的原则为协商双方具有相同的加密算法、认证算法、认证方法和 Diffie-Hellman 组标识。

- c. 网关 B 响应 ISAKMP 消息，携带经协商匹配的安全提议及参数。如果没有匹配的安全提议，网关 B 将拒绝发起方的安全提议。
2. 使用 DH 算法交换与密钥相关的信息，并生成密钥。
    - a. 两个对等体通过两条 ISAKMP 消息（3、4）交换与密钥相关的信息。
    - b. 由获得的密钥信息推导出 4 个密钥。其中 SKEYID 为基础密钥，通过它可以推导出 SKEYID\_a，为 ISAKMP 消息完整性验证密钥；可以推导出 SKEYID\_e，为 ISAKMP 消息加密密钥；可以推导出 SKEYID\_d，用于衍生出 IPSec 报文加密、验证密钥。

预共享密钥方式和数字证书方式下 SKEYID 的计算公式不同。

图 10-28 DH 交换及密钥生成



3. 对等体之间验证彼此身份。
  - a. 两个对等体通过两条 ISAKMP 消息（5、6）交换身份信息（预共享密钥方式下为 IP 地址或名称，数字证书方式下还需要传输证书的内容），身份信息通过 SKEYID\_e 加密，故可以保证身份信息的安全性。
  - b. 两个对等体使用 IKE 安全提议中定义的加密算法、验证算法、身份验证方法和 SKEYID\_a、SKEYID\_e 对 IKE 消息进行加解密和验证。

IKEv1 支持如下几种身份验证方法：

- 预共享密钥

这种方法要求对等体双方必须要有相同的预共享密钥（该密钥直接参与 SKEYID 的生成计算）。对于设备数量较少的 VPN 网络来说易于配置，在大型 VPN 网络中，不建议采用预共享密钥来做身份验证。

- RSA 签名（通常称为数字证书）

数字证书需要由 CA 服务器来颁发。这种方法适用于大型动态的 VPN 网络。证书验证和预共享密钥验证的主要区别在于 SKEYID 的计算和交换的身份信息，其他的交换和计算过程和预共享密钥验证方式相同。

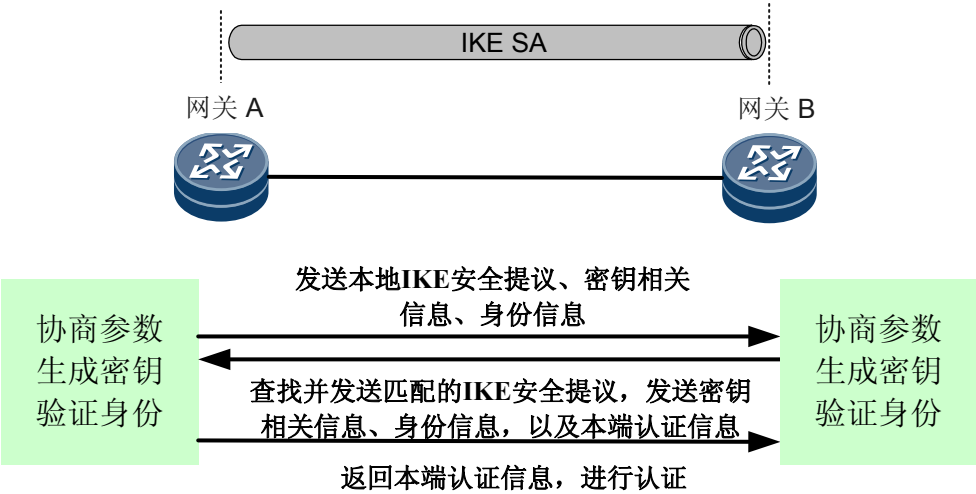
- RSA 验证字
- 当 IKEv1 协商过程中需要使用 RSA 加密 Nonce 认证时，可采用此方法。

野蛮模式

野蛮模式仅交换 3 个消息就可以完成 IKE SA 的建立。

采用野蛮模式时 IKEv1 阶段 1 的协商过程如图 10-29 所示。

图 10-29 野蛮模式下 IKEv1 阶段 1 的协商过程



野蛮模式时 IKEv1 阶段 1 的协商过程：

1. 网关 A 发送 ISAKMP 消息，携带建立 IKE SA 所使用的参数、与密钥生成相关的信息和身份验证信息。
2. 网关 B 对收到的第一个数据包进行确认，查找并返回匹配的参数、密钥生成信息和身份验证信息。
3. 网关 A 回应验证结果，并建立 IKE SA。

与主模式相比，野蛮模式的优点是建立 IKE SA 的速度较快。但是由于密钥交换与身份认证一起进行，野蛮模式无法提供身份保护。

10.2.7.2 IKEv1 协商阶段 2

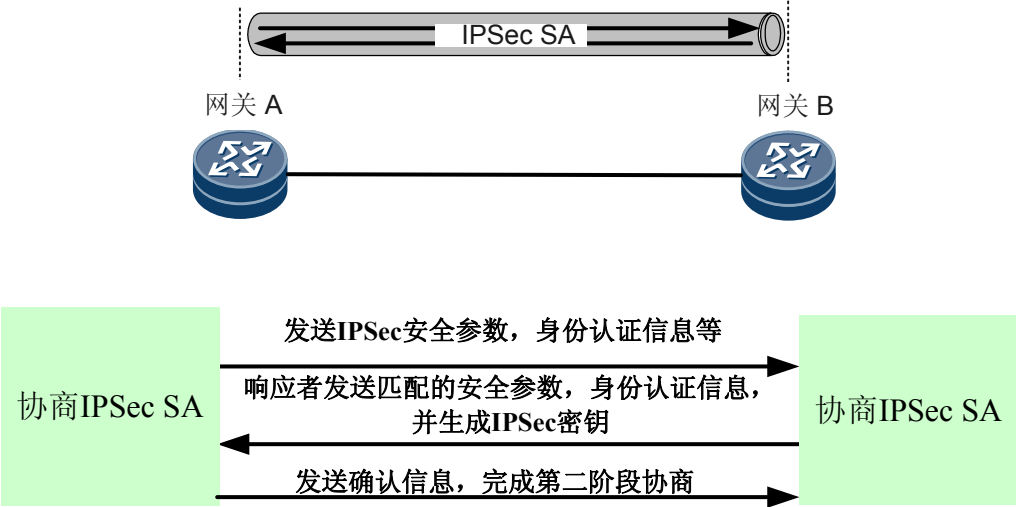
IKEv1 阶段 2 的目的就是建立用来传输数据的 IPSec SA 。

IKEv1 阶段 2 通过快速交换模式完成。由于快速交换模式使用 IKEv1 阶段 1 中生成的密钥 SKEYID\_a 对 ISAKMP 消息的完整性和身份进行验证，使用密钥 SKEYID\_e 对 ISAKMP 消息进行加密，故保证了交换的安全性。

在快速交换模式中，对等体两端协商 IPSec SA 的各项参数，并为数据传输衍生出密钥。

快速模式的协商过程如图 10-30 所示。

图 10-30 IKEv1 阶段 2 协商过程



快速模式共有 3 条消息完成双方 IPSec SA 的建立。

1. 消息 1 发送本端的安全参数和身份认证信息。  
安全参数包括被保护的数据流和 IPSec 安全提议的参数等需要协商的参数。身份认证信息包括第一阶段计算出的密钥和第二阶段产生的密钥材料等，可以再次认证对等体。
2. 消息 2 响应消息 1，发送响应方的安全参数和身份认证信息并生成新的密钥。  
对等体双方通过交换密钥材料生成新的共享密钥，并最终衍生出 IPSec 的加密密钥。此时响应者和发送者各有两个 SA。  
IPSec SA 数据传输需要的加密、验证密钥由 SKEYID\_d、SPI、协议等参数衍生得出，以保证每个 IPSec SA 都有自己独一无二的密钥。  
当启用 PFS 时，要再次应用 DH 算法计算出一个共享密钥，然后参与上述计算，因此在参数协商时要为 PFS 协商 DH 密钥组。
3. 消息 3 响应消息 2，确认与响应方可以通信，协商结束。

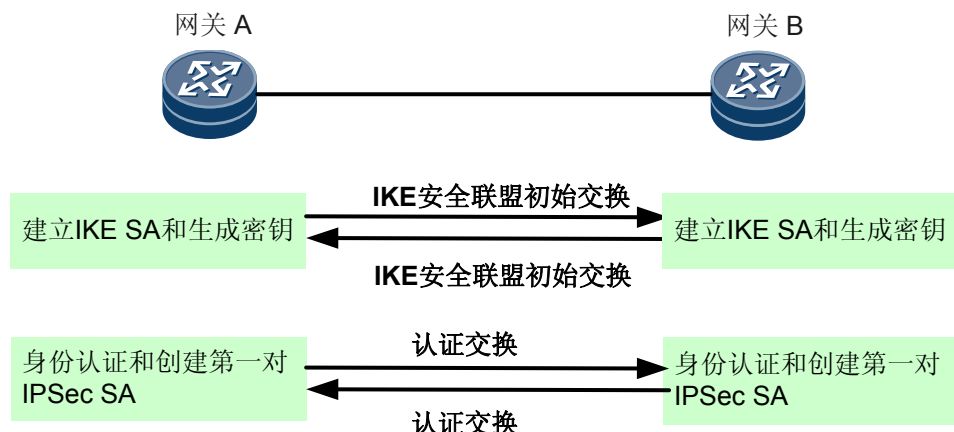
10.2.8 IKEv2 协商安全联盟的过程

要建立一对 IPSec SA，IKEv1 需要经历两个阶段：“主模式+快速模式”或者“野蛮模式+快速模式”。前者至少需要交换 9 条消息，后者也至少需要 6 条消息。而 IKEv2，正常情况使用 2 次交换共 4 条消息就可以完成一个 IKE SA 和一对 IPSec SA，如果要求建立的 IPSec SA 大于一对时，每一对 SA 只需额外增加 1 次交换，也就是 2 条消息就可以完成。这比 IKEv1 要简化很多。

IKEv2 定义了三种交换：初始交换、创建子 SA 交换以及通知交换。

1. 初始交换  
IKE 通信总是从 IKE 安全联盟初始交换（IKE\_SA\_INIT 交换）和 IKE 认证交换（IKE\_AUTH 交换）开始。这 2 个交换通常由 4 条消息组成，在某些场景下消息数目可能会增加。所有使用 IKE 的通信都有请求/响应对组成。IKE 安全联盟初始交换和 IKE 认证交换完成后可以建立一个 IKE 安全联盟和第一对 CHILD\_SA（即 IPSec SA）。

图 10-31 IKEv2 初始阶段的协商过程



详细说明如下：

a. 第一个消息对（IKE\_SA\_INIT）

这个消息对负责进行 IKE 安全联盟参数的协商，包括协商加密和验证算法，交换临时随机数（nonces）和 DH 交换。

IKE\_SA\_INIT 交换后可以生成一个共享密钥材料。通过这个共享密钥材料可以生成其他相关密钥。

b. 第二个消息对（IKE\_AUTH）

从 IKE\_AUTH 交换开始，所有报文都必须加密再交换。IKE\_AUTH 交换至少需要两个消息。在这两个报文的交互中完成了身份认证以及一个 CHILD\_SA 的创建过程。

RSA 签名认证和预共享密钥认证方式下，认证载荷（AUTH 载荷）的计算方式不同。在 IKE\_SA\_INIT 交换阶段，生成 IPsec SA 的密钥材料，并通过这个密钥材料衍生出 IPsec SA 的所有密钥。

IKEv2 除支持 RSA 签名和预共享密钥认证外，还支持在 RFC3748 中定义的扩展认证方法 EAP（Extensible Authentication Protocol）。EAP 认证是作为附加的 IKE\_AUTH 交换在 IKE 中实现的。发起者通过在消息 3 中省去 AUTH 载荷来表明需要使用 EAP 认证。

2. 创建子 SA 交换

当一个 IKE SA 需要创建多个 IPsec SA 时，需要使用创建子 SA 交换来协商多于一个的 SA。另外创建子 SA 交换还可以用于进行 IKE SA 的重协商。

创建子 SA 交换包含一个交换两个消息。在 IKEv1 中这个交换称为阶段 2 交换（快速模式交换）。这个交换必须在 IKE 初始交换完成之后才能进行，交换的发起者可以是 IKE 初始交换的发起者，也可以是 IKE 初始交换的响应者。在交换中的两个消息需要由 IKE 初始交换协商的密钥进行保护。

类似与 IKEv1 的 PFS，创建子 SA 交换阶段可以重新进行一次 DH 交换，生成新的密钥材料。生成密钥材料后，子 SA 的所有密钥都从这个密钥材料衍生出来。

3. 通知交换

运行 IKE 协商的两端有时会传递一些控制信息，例如错误信息或者通告信息，这些信息在 IKEv2 中是通过通知交换完成的。

通知交换必须在 IKE SA 保护下进行，也就是说通知交换只能发生在初始交换之后。

## 10.2.9 EAP 认证

IKEv2 中支持采用 EAP 方式对协商的发起方进行第三方（比如 AAA 服务器）认证。

EAP（Extensible Authentication Protocol）运行于数据链路层，它本身不是一个认证机制，而是一个用来传输实际认证协议的通用架构。目前有超过 20 种 EAP 协议，不同的 EAP 协议的差异在于认证机制与密钥管理的不同。USG 支持的 EAP 协议包括：最基本的 EAP-MD5；基于 SIM 卡的 EAP-AKA 与 EAP-SIM。

EAP 支持单向认证，这种方法可以用于服务器认证用户的身份。在 IKEv2 中这种单向认证方法适用于 IPSec 网关认证远程接入用户。此时 EAP 认证服务器一般采用 AAA 服务器。

EAP 在基于用户的接入认证方面提供了较好的安全机制。IKEv2 中的 EAP 认证基于 RFC3748，通过在 IKE 报文中传递 EAP 载荷实现 EAP 认证。

- EAP-MD5

当远程接入终端是普通 PC 时，EAP 认证一般采用 EAP-MD5 协议。

- EAP-SIM

当远程接入终端是 2G 手机时，EAP 认证采用 EAP-SIM 协议。

EAP-SIM 协议在 WLAN 上提供了基于 GSM-SIM（即 2G 手机 SIM 卡）的认证和密钥分发机制。

- EAP-AKA（Authenticantion and Key Agreement）

当远程接入终端是 3G 手机或安装了 UICC（USIM Integrated Circuit Card，USIM 集成电路卡）的 PC 时 EAP 认证采用 EAP-AKA 协议。

EAP-AKA 协议在 WLAN 上提供了基于 USIM（Universal Subscriber Identity Module，通用用户表示模块，即 3G 手机 SIM 卡）的认证和密钥分发机制。

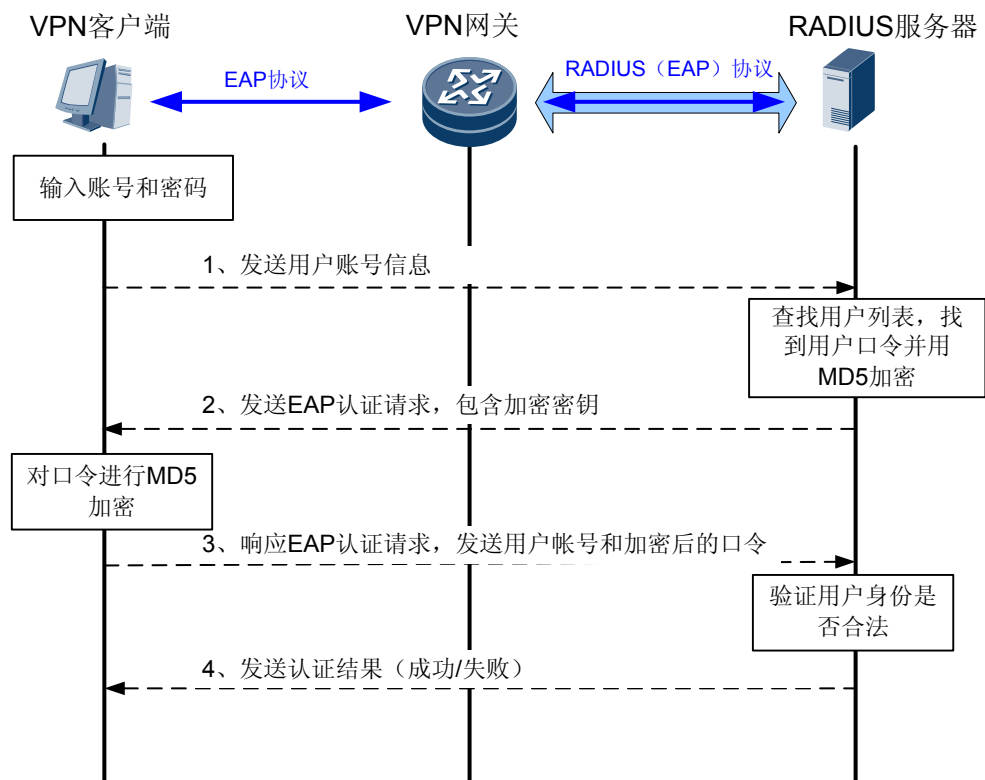
EAP-SIM 和 EAP-AKA 同样采用了 EAP 协议框架，认证流程与 EAP-MD5 基本相同，但使用的用户身份信息和加密算法与 EAP-MD5 不同，需要在终端上安装 EAP-SIM 或 EAP-AKA 客户端软件；认证服务器也不是企业网内部的服务器，而是移动运营商核心网的认证服务器。

## EAP 认证过程

下面以 EAP-MD5 为例说明 EAP 认证过程。如图 10-32 所示，VPN 客户端与 RADIUS 之间交互 EAP 认证消息，实现身份认证。VPN 网关在中间不需要对消息进行解析，仅需要将 EAP 认证消息封装为 RADIUS 消息发往 RADIUS 服务器，起中转作用即可；反之亦然。



图 10-32 EAP-MD5 认证流程



EAP 认证是作为附加的 IKE\_AUTH 交换在 IKE 中实现的。

1. 用户接入 VPN 网关时，通过在 IKE\_AUTH 交换阶段消息 3 中省去 AUTH 载荷来表明需要使用 EAP 认证。此时发起者发送了没有经过验证的用户帐号信息。
2. RADIUS 服务器收到 VPN 网关转发的用户帐号信息后，查找此用户名对应的口令信息，查找到用户名口令后进行加密处理，并发送 EAP 认证请求和加密字，要求客户端响应。
3. 客户端用该加密字对口令进行加密后，将加密字、加密后的密码和用户名发送给 VPN 网关，再由 VPN 网关发送给 RADIUS 服务器进行认证。
4. RADIUS 服务器将 VPN 网关发来的加密后的口令信息和服务器上经过加密运算后的口令信息进行对比，判断用户是否合法。若用户合法，则发送认证成功报文给 VPN 客户端。

## 10.3 L2TP

L2TP 是二层 VPN 技术，提供了对 PPP 链路层数据包的通道（Tunnel）传输支持，允许二层链路端点和 PPP 会话点驻留在不同设备上，并且采用包交换网络技术进行信息交互，扩展了 PPP 模型。

### 10.3.1 介绍

## 定义

PPP 协议定义了一种封装技术，可以在二层点到点链路上传输多种协议数据包，这时，用户与 NAS 之间运行 PPP，二层链路端点与 PPP 会话点在相同硬件设备上。

L2TP 协议提供了对 PPP 链路层数据帧的隧道（Tunnel）传输支持，允许二层链路端点和 PPP 会话点驻留在不同设备上，并采用包交换技术进行信息交互，从而扩展了 PPP 模型。

L2TP 协议结合了 L2F 协议和 PPTP 协议的优点，成为 IETF 有关二层隧道协议的工业标准。

## 目的

L2TP 可以在非点对点的网络上建立点对点的 PPP 会话连接，用于在用户和企业的服务器之间透明传输 PPP 报文。

### 10.3.2 参考标准和协议

与 L2TP 特性相关的参考标准与协议如下：

- RFC 1661: The Point-to-Point Protocol (PPP)
- RFC 1918: Address Allocation for Private Internets
- RFC 2661: Layer Two Tunneling Protocol "L2TP"
- RFC 2809: Implementation of L2TP Compulsory Tunneling via RADIUS
- RFC 2888: Secure Remote Access with L2TP

### 10.3.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 10.3.4 原理描述

#### 10.3.4.1 VPDN 简介



L2TP (Layer 2 Tunneling Protocol) 属于 VPDN 隧道协议的一种。为了更好的理解 L2TP, 下面先简单介绍 VPDN。

VPDN (Virtual Private Dial Network) 采用专用的网络加密通信协议, 在公共网络上为企业建立安全的虚拟专网。企业驻外机构和出差人员可以远程经由公共网络, 通过虚拟加密隧道实现和企业总部之间的网络连接, 而公共网络上其他用户则无法穿过虚拟隧道访问企业网内部的资源。

VPDN 有以下两种实现方式:

- 接入服务器发起 VPDN 连接

NAS (Network Access Server, 网络接入服务器) 通过使用 VPDN 隧道协议, 将客户的 PPP 连接直接连到企业的 VPDN 网关上, 从而与 VPDN 网关建立隧道。

其优势在于: 对用户透明, 用户只需要登录一次就可以接入企业网络, 由企业网进行用户认证和地址分配, 不占用公共地址, 用户可使用各种平台上网。

这种方式需要 NAS 支持 VPDN 协议, 需要认证系统支持 VPDN 属性, 网关一般使用 USG 或 VPN 专用服务器。

- 用户发起 VPDN 连接

客户端与 VPDN 网关建立隧道。这种方式由客户端先建立与 Internet 的连接, 再通过专用的客户软件 (如 Windows 2000 支持的 L2TP 客户端) 与 VPDN 网关建立隧道连接。优缺点如下:

- 优点: 用户上网的方式和地点没有限制, 不需 ISP 介入。
- 缺点: 用户需要安装专用的软件 (一般都是 Win2000 平台的 L2TP 客户端), 限制了用户使用的平台。

VPDN 隧道协议可分为:

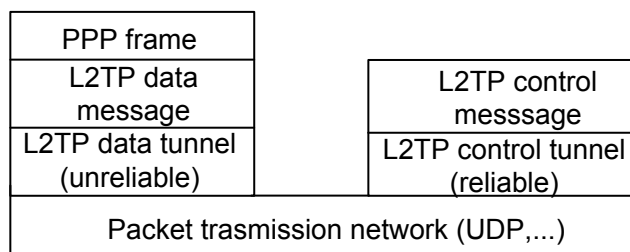
- PPTP (Point-to-Point Tunneling Protocol, 点到点隧道协议)
- L2F (Layer 2 Forwarding, 二层转发)
- L2TP (Layer 2 Tunneling Protocol, 二层隧道协议)

目前使用最广泛的是 L2TP。

## 10.3.4.2 L2TP 协议简介

### L2TP 协议结构

图 10-33 L2TP 协议结构



如图 10-33 所示, L2TP 协议结构描述了 PPP 帧、控制消息和控制通道以及数据消息、数据通道之间的关系: PPP 帧在不可靠的 L2TP 数据通道上进行传输, 控制消息在可靠的 L2TP 控制通道内传输。

通常 L2TP 数据以 UDP 报文形式发送。L2TP 注册了 UDP 端口 1701，但是这个端口仅用于初始的隧道建立过程。L2TP 隧道发起方（LAC）任选一个空闲端口（未必是 1701）向接收方（LNS）的 1701 端口发送报文；LNS 收到报文后，也任选一个空闲端口（未必是 1701），给 LAC 的指定端口回送报文。至此，双方的端口选定，并在隧道保持连通的时间段内不再改变。

## 隧道和会话的概念

在 LNS（L2TP Network Server，L2TP 网络服务器）和 LAC（L2TP Access Concentrator，L2TP 访问集中器）对之间存在着两种类型的连接：

- 隧道（Tunnel）连接：它定义了互相通信的两个实体 LNS 和 LAC。
- 会话（Session）连接：它复用在隧道连接之上，用于表示承载在隧道连接中的每个 PPP 会话过程。

在同一对 LAC 和 LNS 之间可以建立多个 L2TP 隧道，隧道由一个控制连接和至少一个会话（Session）组成。会话连接必须在隧道建立（包括身份保护、L2TP 版本、帧类型、硬件传输类型等信息的交换）成功之后进行，每个会话连接对应于 LAC 和 LNS 之间的一个 PPP 数据流。控制消息和 PPP 数据报文都在隧道上传输。

L2TP 使用 Hello 报文来检测隧道的连通性。LAC 和 LNS 定时向对端发送 Hello 报文，如果在一段时间内未收到 Hello 报文的应答，隧道将被清除。

## 控制消息和数据消息的概念

L2TP 中存在控制消息和数据消息两种消息，其中：

- 控制消息用于隧道和会话连接的建立、维护以及传输控制。  
控制消息的传输是可靠传输，并且支持对控制消息的流量控制和拥塞控制。
- 数据消息则用于封装 PPP 帧并在隧道上传输。  
数据消息的传输是不可靠传输，如果数据报文丢失，不予重传，不支持对数据消息的流量控制和拥塞控制。

控制消息和数据消息共享相同的报文头。L2TP 报文头中包含隧道标识符（Tunnel ID）和会话标识符（Session ID）信息，用来标识不同的隧道和会话。隧道标识相同、会话标识不同的报文将被复用在一条隧道上，报文头中的隧道标识符与会话标识符由对端分配。

## 三种 L2TP 隧道模式

USG 有三种方式可以建立连接：

- NAS-Initialized  
由远程拨号用户发起。远程系统通过 PSTN/ISDN 拨入 LAC，由 LAC 通过 Internet 向 LNS 发起建立隧道连接请求。拨号用户地址由 LNS 分配；对远程拨号用户的验证与计费既可由 LAC 侧的代理完成，也可在 LNS 侧完成。
- Client-Initialized  
直接由 LAC 客户端（指可在本地支持 L2TP 协议的 PC）发起。此时 LAC 客户端可直接向 LNS 发起隧道连接请求，无需再经过一个单独的 LAC 设备。LAC 客户端地址的分配由 LNS 来完成。
- LAC 自主拨号

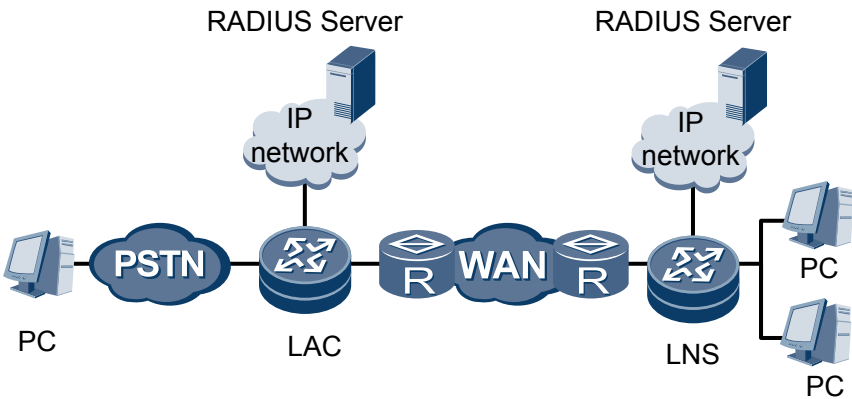
用户通过配置命令触发建立 LAC 与 LNS 之间的永久性 L2TP 会话。LAC 使用存储在本地的用户名通过虚拟模板接口和 LNS 建立一个永久存在的 L2TP 隧道，此时的 L2TP 隧道就想当于一个物理连接，出接口是虚拟模板接口。

用户与 LAC 之间的连接就不受限于 PPP 连接，而只要是一个 IP 连接即可，这样 LAC 能够将用户的 IP 报文转发到 LNS。

L2TP 隧道会话的建立过程

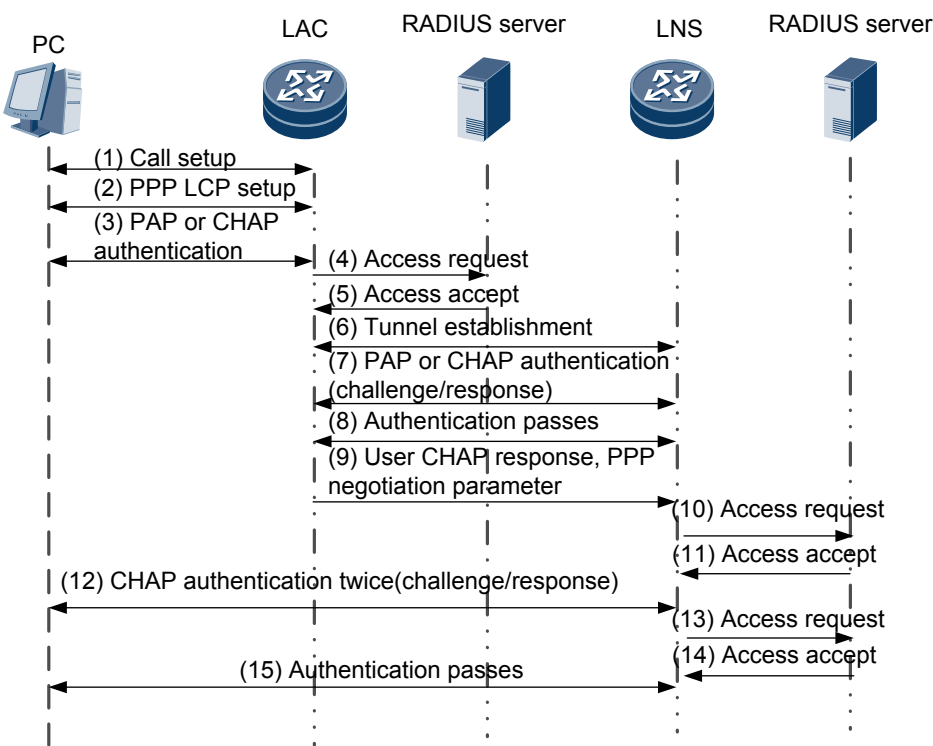
L2TP 隧道的典型应用组网如图 10-34 所示。

图 10-34 L2TP 隧道的应用组网图



L2TP 隧道的呼叫建立流程可如图 10-35 所示。

图 10-35 L2TP 隧道的呼叫建立流程



L2TP 隧道的呼叫建立流程过程为：

1. 用户端 PC 发起呼叫连接请求。
2. PC 和 LAC 端（USG A）进行 PPP LCP 协商。
3. LAC 对 PC 机提供的用户信息进行 PAP（Password Authentication Protocol）或 CHAP（Challenge Handshake Authentication Protocol）认证。
4. LAC 将认证信息（用户名、密码）发送给 RADIUS 服务器进行认证。
5. RADIUS 服务器认证该用户，如果认证通过则 LAC 准备发起 Tunnel 连接请求。
6. LAC 端向指定 LNS 发起 Tunnel 连接请求。
7. LAC 端向指定 LNS 发送 CHAP challenge 信息，LNS 回送该 challenge 响应消息 CHAP response，并发送 LNS 侧的 CHAP challenge，LAC 返回该 challenge 的响应消息 CHAP response。
8. 隧道验证通过。
9. LAC 端将用户 CHAP response、response identifier 和 PPP 协商参数传送给 LNS。
10. LNS 将接入请求信息发送给 RADIUS 服务器进行认证。
11. RADIUS 服务器认证该请求信息，如果认证通过则返回响应信息。
12. 如果用户在 LNS 侧配置强制本端进行 CHAP 认证，则 LNS 对用户进行认证，发送 CHAP challenge，用户侧回应 CHAP response。
13. LNS 再次将接入请求信息发送给 RADIUS 服务器进行认证。
14. RADIUS 服务器认证该请求信息，如果认证通过则返回响应信息。
15. 验证通过，用户访问企业内部资源。

## L2TP 协议的特点

L2TP 协议有如下特点：

- 灵活的身份验证机制以及高度的安全性
  - L2TP 协议本身并不提供连接的安全性，但它可依赖于 PPP 提供的认证（比如 CHAP、PAP 等），因此具有 PPP 所具有的所有安全特性。
  - L2TP 可以与 IPSec 结合，使通过 L2TP 所传输的数据更难被攻击。
  - 可根据特定的网络安全要求，在 L2TP 之上采用通道加密技术、端对端数据加密或应用层数据加密等方案来提高数据的安全性。
- 多协议传输

L2TP 传输 PPP 数据包，PPP 本身可以传输多协议，而不仅仅是 IP。可以在 PPP 数据包内封装多种协议。
- 支持 RADIUS 服务器的验证

LAC 端将用户名和密码发往 RADIUS 服务器进行验证申请，由 RADIUS 服务器负责接收用户的验证请求，完成验证。
- 支持内部地址分配

LNS 可放置于企业网的 USG 之后，对远端用户地址进行动态分配和管理，可支持私有地址应用（RFC 1918）。
- 网络计费的灵活性

可在 LAC 和 LNS 两处同时计费，即 ISP 处（用于产生帐单）及企业网关（用于付费及审计）。L2TP 能够提供数据传输的出入包数、字节数以及连接的起始、结束时间等计费数据，可根据这些数据方便地进行网络计费。

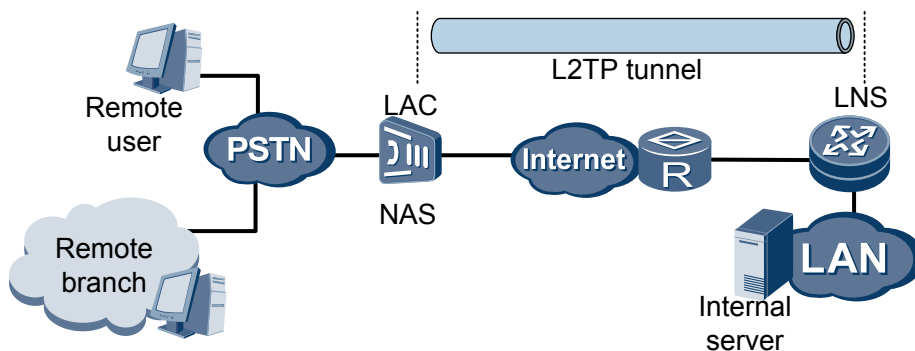
- 可靠性

L2TP 协议支持备份 LNS，当一个主 LNS 不可达之后，LAC 可以重新与备份 LNS 建立连接，增加了 VPN 服务的可靠性和容错性。

### 10.3.4.3 L2TP 典型应用

使用 L2TP 协议构建的 VPDN 应用的典型组网如图 10-36 所示。

图 10-36 应用 L2TP 构建的 VPDN 服务



在 L2TP 构建的 VPDN 中，网络组件包括以下三个部分：

- 远端系统

远端系统是要接入 VPDN 网络的远地用户和远地分支机构，通常是一个拨号用户的主机或私有网络的一台路由设备。

- LAC

LAC 是附属在交换网络上的具有 PPP 端系统和 L2TP 协议处理能力的设备，通常是一个当地 ISP 的 NAS，主要用于为 PPP 类型的用户提供接入服务。

LAC 位于 LNS 和远端系统之间，用于在 LNS 和远端系统之间传递信息包。它把从远端系统收到的信息包按照 L2TP 协议进行封装并送往 LNS，同时也将从 LNS 收到的信息包进行解封并送往远端系统。

LAC 与远端系统之间采用本地连接或 PPP 链路，VPDN 应用中通常为 PPP 链路。

- LNS

LNS 既是 PPP 端系统，又是 L2TP 协议的服务器端，通常作为一个企业内部网的边缘设备。

LNS 作为 L2TP 隧道的另一侧端点，是 LAC 的对端设备，是 LAC 进行隧道传输的 PPP 会话的逻辑终止端点。通过在公网中建立 L2TP 隧道，将远端系统的 PPP 连接的另一端由原来的 LAC 在逻辑上延伸到了企业网内部的 LNS。

## 10.4 GRE

GRE 是三层隧道的封装协议，对某些网络层协议的数据报文进行封装并在另一网络层协议中传输。GRE 提供了将一种协议的报文封装在另一种协议报文中的机制，使报文能够在隧道中传输。

### 10.4.1 介绍

#### 定义

GRE（General Routing Encapsulation，通用路由封装）是对某些网络层协议（如 IPv4、IPv6）的报文进行封装，使这些被封装的报文能够在另一网络层协议（如 IPv4）中传输。

GRE 提供了将一种协议的报文封装在另一种协议报文中的机制，使报文能够在异种网路中传输，封装后报文的传输通道称为 tunnel。

Tunnel 是一个虚拟的点对点的连接，可以看成仅支持点对点连接的虚拟接口，这个接口提供了一条通路，使封装的数据报能够在这个通路上传输，并在一个 Tunnel 的两端分别对数据报进行封装及解封装。

#### 目的

为了使某些网路层协议（如 ATM、IPV6、AppleTalk 等）的报文能够在 IPv4 网络中传输，可以将某些网络层协议的报文进行封装，以此解决了异种网络的传输问题。

GRE 也可以作为 VPN 的第三层隧道协议，为 VPN 数据提供透明传输通道。

### 10.4.2 参考标准和协议

与 GRE 特性相关的参考标准与协议如下：

- RFC1701: Generic Routing Encapsulation (GRE)
- RFC1702: Routing Encapsulation over IPv4 networks
- RFC2784: Generic Routing Encapsulation (GRE)
- draft-ietf-l3vpn-greip-2547-02: Use of PE-PE GRE or IP in BGP/MPLS IP VPNs
- draft-ietf-mpls-in-ipor-gre-08: Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)

### 10.4.3 可获得性

#### License 支持

本特性无需 License 支持。

版本支持

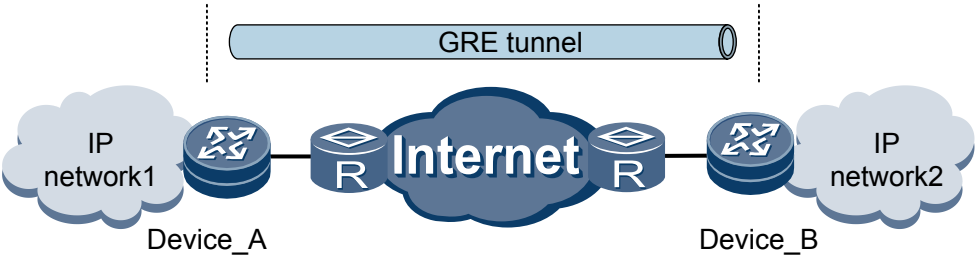
| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

10.4.4 原理描述

10.4.4.1 GRE 的实现

下面以图 10-37 的网络为例说明封装和解封装过程。

图 10-37 IP 网络通过 GRE 隧道互连



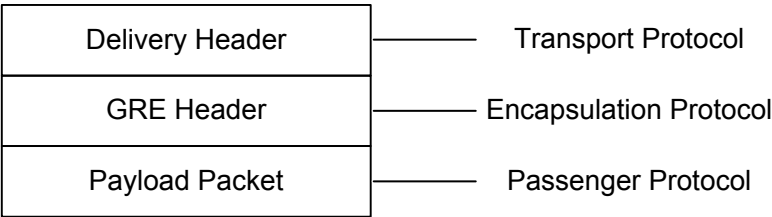
加封装过程

Device\_A 连接 IP network1 的接口收到 IP 数据包后首先交由 IP 协议模块处理，IP 协议模块检查 IP 报头中的目的地址来确定如何转发此包。若报文的目的地址为经过 Tunnel 接口的虚拟网络地址，则将此报文发给 Tunnel 接口。Tunnel 接口收到此报文后进行 GRE 封装，封装完成后交给 IP 协议模块处理。IP 协议模块封装新的 IP 报文头后，根据目的地址查找路由表交由相应的接口处理。

系统收到一个需要封装和路由的数据包，称之为净荷（Payload），这个净荷首先被加上 GRE 封装，成为 GRE 报文；再被封装在 IP 报文中，这样就可完全由 IP 层负责此报文的向前传输（Forwarded）。人们常把这个负责向前传输 IP 报文的协议称为传输协议（Delivery Protocol 或者 Transport Protocol）。

封装后的报文的形式如图 10-38 所示。

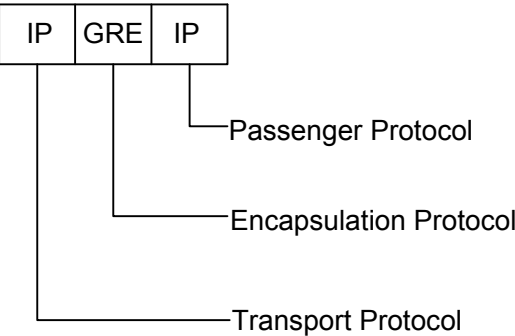
图 10-38 封装后的 GRE 报文格式



- 净荷（Payload）：系统收到的需要封装和路由的数据报称为净荷。
- 乘客协议（Passenger Protocol）：封装前的报文协议称为乘客协议。
- 封装协议（Encapsulation Protocol）：上述的 GRE 协议称为封装协议，也称为运载协议（Carrier Protocol）。
- 传输协议（Transport Protocol 或者 Delivery Protocol）：负责对封装后的报文进行转发的协议称为传输协议。

例如，一个封装在 IP Tunnel 中的传输报文的格式可以表示为如图 10-39 所示。

图 10-39 Tunnel 中传输报文的格式



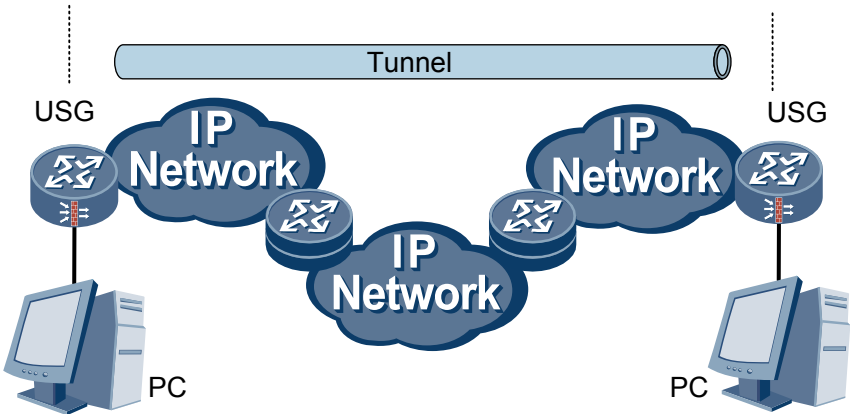
## 解封装过程

解封装过程和加封装过程相反。Device\_B 从连接 Internet 的接口收到 IP 报文后检查目的地址，若发现目的地址为 Device\_B，且协议字段值为 47，表示协议为 GRE。此时交给 GRE 协议模块处理（去掉报文的 IP 报头、检查识别关键字、校验和等）。GRE 协议模块完成相应的处理后，去掉 GRE 报头，再交由 IP 协议模块处理，IP 协议模块将此数据包转发到 IP network2。

### 10.4.4.2 GRE 典型应用

#### 扩大跳数受限的网络工作范围

图 10-40 扩大网络工作范围

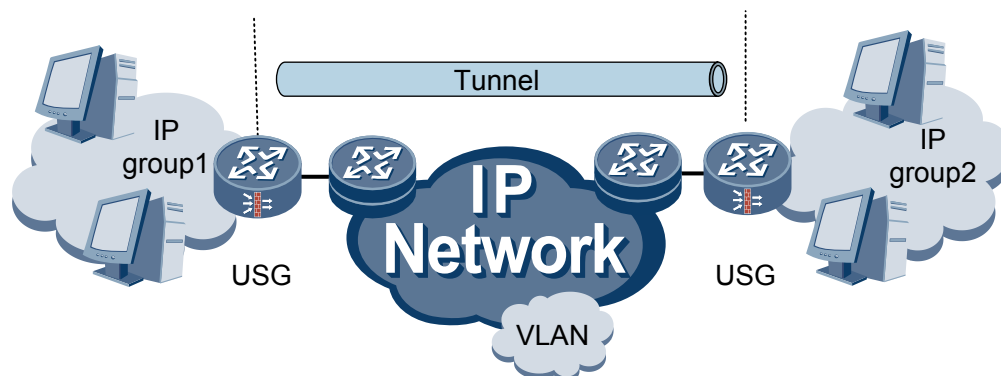




在图 10-40 中，网络运行 IP 协议，假设 IP 协议限制的跳数为 15。如果两台 PC 之间的跳数超过 15，他们将无法通信。在网络中使用隧道（Tunnel）可以隐藏一部分步跳，从而扩大网络的工作范围。

## 将不连续的子网连接起来，用于组建 VPN

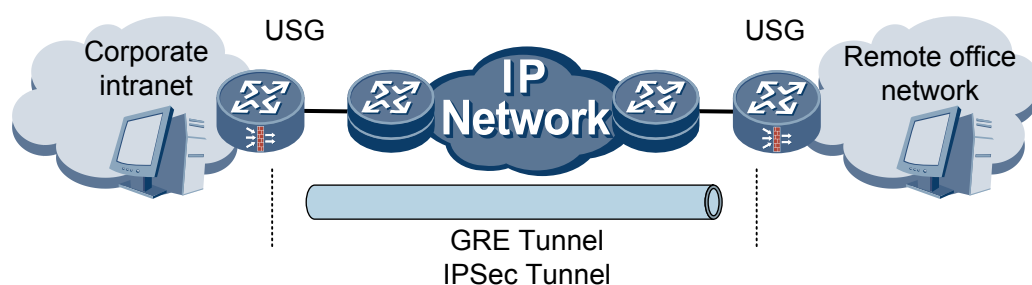
图 10-41 Tunnel 连接不连续子网



在图 10-41 中，运行 IP 协议的两个子网 group1 和 group2 分别在不同的城市，通过使用隧道可以实现跨越广域网的 VPN。

## 与 IPSec 结合，保护组播数据

图 10-42 GRE-IPSec 隧道应用



在图 10-42 中，GRE 可以封装组播数据并在 GRE 隧道中传输。而 IPSec 目前只能对单播数据进行加密保护，因此对于诸如路由协议、语音、视频等组播数据需要在 IPSec 隧道中传输的情况，可以通过建立 GRE 隧道，并对组播数据进行 GRE 封装，然后再对封装后的报文进行 IPSec 的加密处理，就实现了组播数据在 IPSec 隧道中的加密传输。

另外，GRE 还支持由用户选择记录 Tunnel 接口的识别关键字和对封装的报文进行端到端校验。

由于 GRE 收发双方加封装、解封装处理以及由于封装造成的数据量增加等因素的影响，这就导致使用 GRE 会造成 USG 数据转发效率有一定程度的下降。

## 10.5 MPLS

MPLS（Multiprotocol Label Switching）支持多层标签和面向连接特性。MPLS LDP 协议提供了环路检测机制，可以预防 LSP 环路的出现。如果管理员不需要严格控制 LSP 建立的过程，且不需要在 MPLS 网络中部署流量工程，则推荐采用 LDP 协议来创建 LSP。

### 10.5.1 介绍

#### 定义

多协议标签交换 MPLS 起源于 IPv4（Internet Protocol version 4，因特网协议版本 4），最初是为了提高转发速度而提出的，其核心技术可扩展到多种网络协议，包括 IPv6（Internet Protocol version 6，因特网协议版本 6）和 CLNP（Connectionless Network Protocol，无连接网络协议）等。MPLS 中的“M”指的就是支持多种网络协议。

MPLS 能工作在任何链路协议(Ethernet、ATM、FR、PPP 等)之上，能够提供优质 QoS 保证，支撑更大规模的网络应用，并可以替代 IP 寻址转发的协议载体。MPLS 是一种比 ATM 更简单，灵活，更易于扩展的标签交换技术，通常工作在链路层协议之上，网络层 IP 协议之下。

LDP（Label Distribution Protocol，标签分发协议）是 MPLS 的一种控制协议，相当于传统网络中的信令协议，负责 FEC（Forwarding Equivalence Class，转发等价类）的分类、标签的分配以及 LSP（Label Switched Path，标签交换路径）的建立和维护等操作。LDP 规定了标签分发过程中的各种消息以及相关处理过程。

#### 目的

MPLS 技术集二层的快速交换和三层的路由转发于一体，可以满足各种新应用对网络的要求。

MPLS 为 VPN 的实现提供了一种简单、灵活、高效的隧道机制，在 VPN 的不同站点(site)之间建立 LSP，用来传递 VPN 报文。每个站点将到达自己的路由发送给骨干网，这些路由在骨干网中加上标签以及 VPN 的成员信息进行分布。

MPLS 支持多层标签，并且转发平面面向连接，故具有良好的扩展性，使在统一的 MPLS/IP 基础网络架构上为客户提供各类服务成为可能。通过 LDP 协议，LSR（Label Switched Router，标签交换路由器）可以把网络层的路由信息直接映射到数据链路层的交换路径上，建立起网络层的 LSP。目前，LDP 广泛地应用在 VPN 服务的提供上，具有组网、配置简单、支持路由拓扑驱动建立 LSP、支持大容量 LSP 等优点。

### 10.5.2 参考标准和协议

与 MPLS 特性相关的参考标准与协议如下：

- RFC 3031: Multiprotocol Label Switching Architecture
- RFC 3036: LDP Specification
- RFC 3032: MPLS Label Stack Encoding
- RFC 3443: Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks

- RFC 3034: Use of Label Switching on Frame Relay Networks Specification
- RFC 2547: BGP/MPLS VPNs
- draft-bonica-icmpmpls-02: ICMP Extensions for MultiProtocol Label Switching
- draft-ietf-mpls-lspping-07: Detecting MPLS Data Plane Failures
- RFC 3031: Multiprotocol Label Switching Architecture
- RFC 3032: MPLS Label Stack Encoding
- RFC 3033: The Assignment of the Information Field and Protocol Identifier in the Q.2941 Generic Identifier and Q.2957 User-to-user Signaling for the Internet Protocol
- RFC 3034: Use of Label Switching on Frame Relay Networks Specification
- RFC 5036: LDP Specification
- RFC3037: LDP Applicability
- RFC3270: Multi-Protocol Label Switching (MPLS) Support of Differentiated Services
- draft-jork-ldp-igpsync-02: LDP IGP Synchronization

### 10.5.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

### 10.5.4 原理描述

#### 10.5.4.1 MPLS 产生背景

90 年代初，随着 Internet 的快速普及，由于当时硬件技术的限制，采用最长匹配查找路由的 IP 技术已经成为限制网络转发性能的一大瓶颈。而 ATM 技术采用定长标签，并且只需要维护比路由表规模小得多的标签表，能够提供比 IP 路由方式高得多的转发性能。

传统的 IP 技术简单，但性能受到限制。ATM 技术性能高，但其复杂的控制信令和高昂的部署成本使得它很难普及。MPLS 技术就是在这种背景下产生的，它把 IP 和 ATM 的优势很好地结合在一起。

随着 ASIC 技术的发展，路由查找速度已经不是阻碍网络发展的瓶颈。这使得 MPLS 在提高转发速度方面不再具备明显的优势。但是 MPLS 支持多层标签和面向连接的特性，使其在 VPN、流量工程、QoS 等方面得到广泛应用。

MPLS 技术集二层的快速交换和三层的路由转发于一体，可以满足各种新应用对网络的要求。

### 10.5.4.2 MPLS 基本概念

#### 转发等价类（FEC）

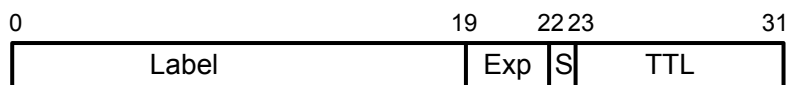
MPLS 作为一种分类转发技术，将具有相同转发处理方式的分组归为一类，称为 FEC。相同 FEC 的分组在 MPLS 网络中将获得完全相同的处理。FEC 的划分方式非常灵活，可以是以源地址、目的地址、源端口、目的端口、协议类型或 VPN 等为划分依据的任意组合。例如，在传统的采用最长匹配算法的 IP 转发中，到同一个目的地址的所有报文就是一个 FEC。

#### 标签（Label）

标签是一个长度固定、只具有本地意义的短标识符，用于唯一标识一个分组所属的 FEC。在某些情况下，例如要进行负载分担，对应一个 FEC 可能会有多个标签，但是一个标签只能代表一个 FEC。

标签由报文的头部所携带，不包含拓扑信息，只具有局部意义。标签的长度为 4 个字节（32bits），封装结构如图 10-43 所示。

图 10-43 标签的封装结构



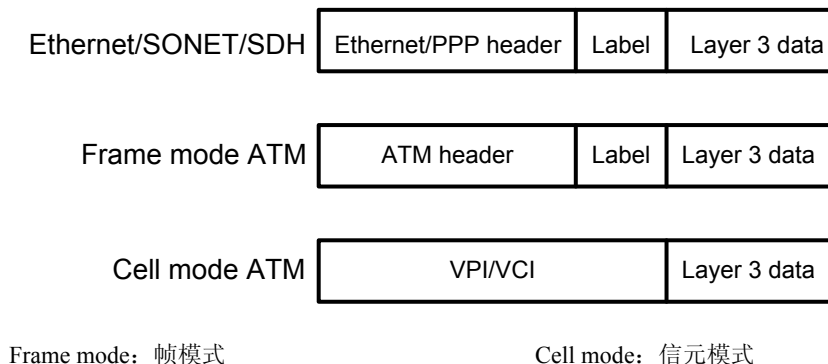
标签共有 4 个域：

- Label: 标签值字段，长度为 20bits，用于转发的指针；
- Exp: 3bits，用于 QoS；
- S: 1bit，用于标识该标签是否是栈底标签，值为 1 时表明为最底层标签。主要应用于 MPLS 标签的多重嵌套；
- TTL: 8bits，和 IP 分组中的 TTL（Time To Live，生存时间）意义相同。

标签与 ATM 的 VPI/VCI 以及 Frame Relay 的 DLCI 类似，是一种连接标识符。如果链路层协议具有标签域，如 ATM 的 VPI/VCI 或 Frame Relay 的 DLCI，则标签封装在这些域中；如果不支持，则标签封装在链路层和 IP 层之间的一个垫层中。这样，标签能够被任意的链路层所支持。

标签在各种介质中的封装位置如图 10-44 所示：

**图 10-44 标签在各种介质中的封装位置**

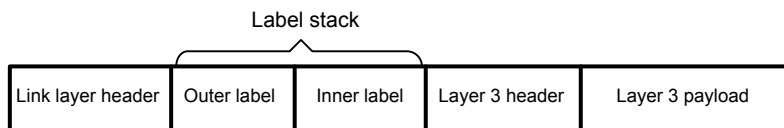


对于以太网、PPP 的分组，标签堆栈象“垫层”一样，位于二层报头与数据之间，对于 ATM 信元模式的分组，借用 VPI/VCI 来作为标签使用。

## 标签栈 (Label stack)

标签栈是指标签的排序集合。MPLS 报文支持同时携带多个标签，靠近二层首部的标签称为栈顶标签或外层标签；靠近 IP 首部的标签称为栈底标签，或内层标签。理论上，MPLS 标签可以无限嵌套。如图 10-45 所示。

图 10-45 标签栈



标签栈按后进先出 (Last In First Out) 方式组织标签, 从栈顶开始处理标签。

## 标签操作类型

标签的操作类型包括标签压入（Push）、标签交换（Swap）和标签弹出（Pop），它们是标签转发的基本动作，是标签转发信息表的组成部分。

- **Push:** 指当 IP 报文进入 MPLS 域时, MPLS 边界设备在报文二层首部 and IP 首部之间插入一个新标签; 或者 MPLS 中间设备根据需要, 在标签栈顶增加一个新的标签 (即标签嵌套封装)。
- **Swap:** 当报文在 MPLS 域内转发时, 根据标签转发表, 将 MPLS 报文的栈顶标签删除, 然后添加一层下一跳分配的标签。
- **Pop:** 当报文离开 MPLS 域时, 将 MPLS 报文的标签去掉; 或者 MPLS 倒数第二跳节点处去掉栈顶标签, 减少标签栈中的标签数目。

## 倒数第二跳弹出

在最后一跳节点，标签已经没有使用价值。这种情况下，可以利用倒数第二跳弹出特性 PHP（Penultimate Hop Popping），在倒数第二跳节点处将标签弹出，减少最后一跳的负担。最后一跳节点直接进行 IP 转发或者下一层标签转发。

PHP 在 Egress 节点上配置。支持 PHP 的 Egress 节点分配给倒数第二跳节点的标签有以下两种：

- 标签值 0：表示 IPv4 显式空标签（explicit-null）。被分配了显式空标签的报文在 Egress 节点将被直接进行 Pop 操作，处理 EXP 和 TTL 后进行 IP 转发或下一层标签转发，不再查询标签转发表。
- 标签值 3：表示隐式空标签（implicit-null），这个值不会出现在标签栈中。当一个 LSR 发现自己被分配了隐式空标签时，它并不用这个值替代栈顶原来的标签，而是直接执行 Pop 操作。Egress 节点直接进行 IP 转发或下一层标签转发。

## 标签交换路由器（LSR）

标签交换路由器 LSR 是指可以进行 MPLS 标签交换和报文转发的网络设备，也称为 MPLS 节点。

LSR 是 MPLS 网络中的基本元素，所有 LSR 都支持 MPLS 技术。

## LER

位于 MPLS 域边缘的 LSR 称为 LER（Label Edge Router）。如果一个 LSR 有一个不运行 MPLS 的相邻节点，那么该 LSR 就是 LER。

LER 负责对进入 MPLS 域的报文划分 FEC，并为这些 FEC 压入标签，进行 MPLS 转发。当报文离开 MPLS 域时弹出标签，恢复成原来的报文，再进行相应的转发。

## 上游和下游

根据数据传送的方向，LSR 可以分为上游和下游。

- 上游：以指定的 LSR 为视角，根据数据传送的方向，所有往本 LSR 发送 MPLS 报文的 LSR 都可以称为上游 LSR。
- 下游：以指定的 LSR 为视角，根据数据传送的方向，本 LSR 将 MPLS 报文发送到的所有下一跳 LSR 都可以称为下游 LSR。

如图 10-46 所示，对于发往 192.168.1.0/24 的数据流来说，LSR-A 是 LSR-B 的上游节点，LSR-B 是 LSR-A 的下游节点。同理，LSR-B 是 LSR-C 上游节点。LSR-C 是 LSR-B 的下游节点。

图 10-46 上游和下游概念



## 标签交换路径(LSP)

一个转发等价类在 MPLS 网络中经过的路径称为 LSP（Label Switched Path）。

LSP 在功能上与 ATM 和帧中继（Frame Relay）的虚电路相同，是从 MPLS 网络的入口到出口的一个单向路径。LSP 中的每个节点由 LSR 组成。

## 入节点、中间节点和出节点

标签交换路径 LSP 是一个单向路径，LSP 中的 LSR 可以分为：

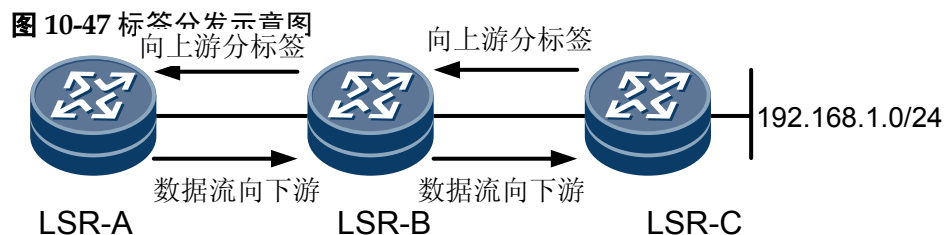
- 入节点（Ingress）：LSP 的起始节点，一条 LSP 只能有一个 Ingress。  
Ingress 的主要功能是给报文压入一个新的标签，封装成 MPLS 报文进行转发。
- 中间节点（Transit）：LSP 的中间节点，一条 LSP 可能有多个 Transit。  
Transit 的主要功能是查找标签转发信息表，通过标签交换完成 MPLS 报文的转发。
- 出节点（Egress）：LSP 的末节点，一条 LSP 只能有一个 Egress。  
Egress 的主要功能是弹出标签，恢复成原来的报文进行相应的转发。

其中 Ingress 和 Egress 既是 LSR，又是 LER；Transit 是 LSR。

## 标签分发

将目的地址相同的分组报文划分为一个 FEC，然后从 MPLS 标签资源池中取出一个标签，分配给这个 FEC。LSR 记录该标签和 FEC 的对应关系，并将该对应关系封装成消息报文，通告给上游的 LSR，这个过程称为标签分发。

如图 10-47，LSR-B 和 LSR-C 对去往 192.168.1.0/24 的报文划分为一个 FEC，然后为该 FEC 分配标签，并向上游通告。因此标签是由下游分配的。



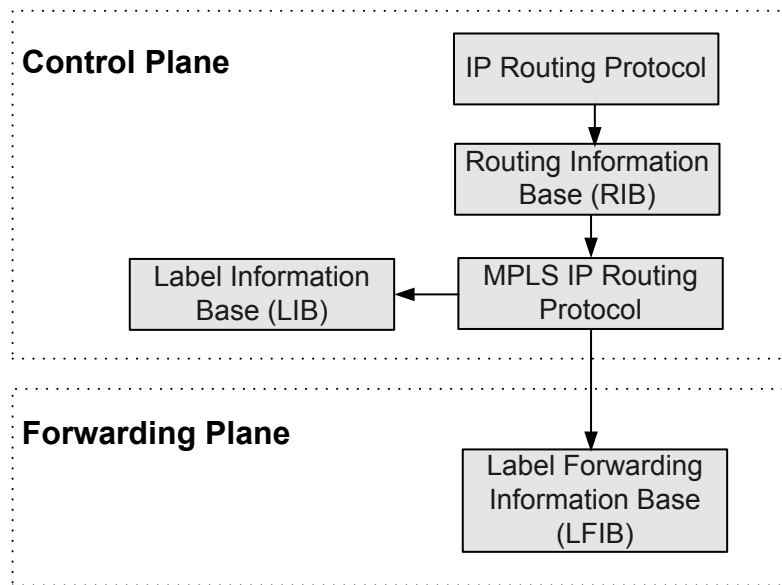
## 标签分发协议（LDP）

LDP 是 MPLS 中的信令控制协议，负责 FEC 的分类、标签的分配以及 LSP 的建立和维护等一系列操作。

### 10.5.4.3 MPLS 体系结构

MPLS 的体系结构由控制平面（Control Plane）和转发平面（Forwarding Plane）组成。MPLS 体系结构如图 10-48。

图 10-48 MPLS 体系结构示意图



- 控制平面是无连接的，主要功能是负责标签的分配、标签转发表的建立、标签交换路径的建立、拆除等工作。
- 转发平面也称为数据平面（Data Plane），是面向连接的，可以使用 ATM、帧中继、Ethernet 等二层网络。转发平面的主要功能是对 IP 包进行标签的添加和删除，同时依据标签转发表对收到的分组进行转发。

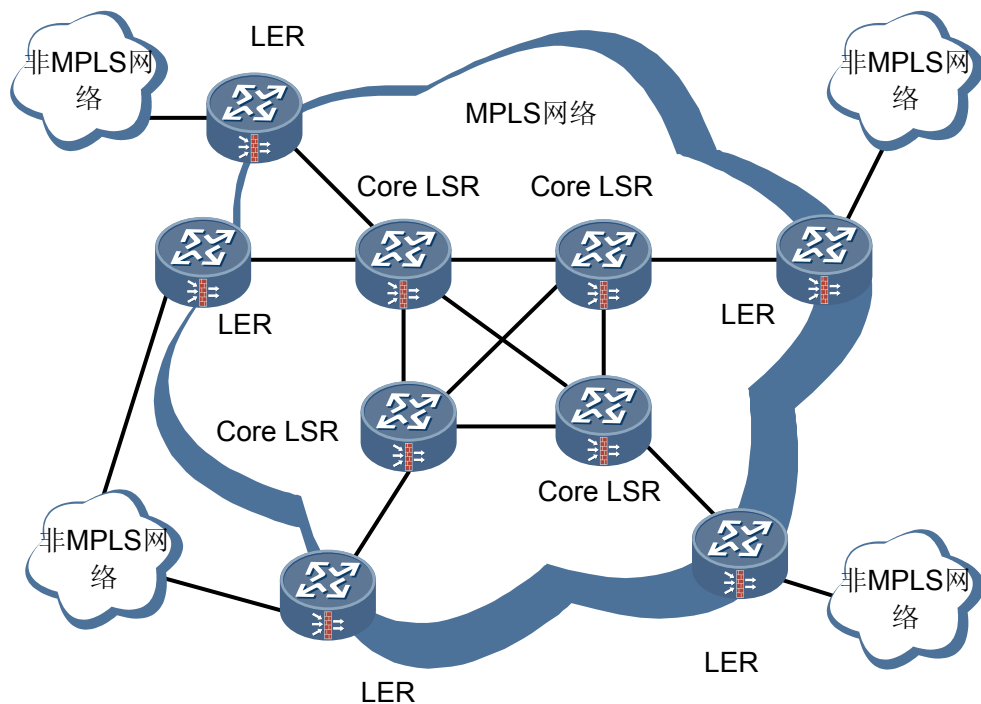
#### 10.5.4.4 MPLS 网络结构

MPLS 网络的典型结构如图 10-49，MPLS 网络的基本组成单元是 LSR，由 LSR 构成的网络区域称为 MPLS 域（MPLS Domain）。位于 MPLS 域边缘、连接其它网络的 LSR 称为 LER，区域内部的 LSR 称为核心 LSR（Core LSR）。如果一个 LSR 的相邻节点都运行 MPLS，则该 LSR 就是核心 LSR。

MPLS 基于标签进行转发。IP 包进入 MPLS 网络时，MPLS 入口的 LER 分析 IP 包的内容并且为这些 IP 包添加合适的标签，所有 MPLS 网络中的节点都是依据标签来转发数据的。当该 IP 包离开 MPLS 网络时，标签由出口边缘路由器删除。



图 10-49 MPLS 网络结构



结合上图简要介绍 MPLS 的基本工作过程：

1. LDP 和传统路由协议（如 OSPF、ISIS 等）一起，在各个 LSR 中为有业务需求的 FEC 建立路由表和 LIB（Label Information Base，标签信息表）；
2. 入口 LER 接收分组，完成第三层功能，判定分组所属的 FEC，并给分组加上标签，形成 MPLS 标签分组；
3. 在 LSR 构成的网络中，LSR 根据分组上的标签以及 LFIB（Label Forwarding Information Base，标签转发表）进行转发，不对标签分组进行任何第三层处理；
4. 在 MPLS 出口 LER 去掉分组中的标签，继续进行后面的 IP 转发。

由此可以看出，MPLS 并不是一种业务或者应用，它实际上是一种隧道技术，也是一种将标签交换转发和网络层路由技术集于一身的路由与交换技术平台。这个平台不仅支持多种高层协议与业务，而且，在一定程度上可以保证信息传输的安全性。

#### 10.5.4.5 MPLS Ping/Traceroute

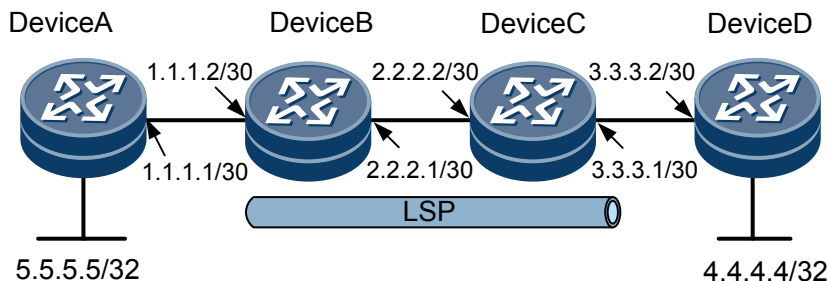
在 MPLS 中，如果 LSP 转发数据失败，负责建立 LSP 的 MPLS 控制平面将无法检测到这种错误，这会给网络维护带来困难。MPLS Ping/Traceroute 为用户提供了发现 LSP 错误、并及时定位失效节点的机制。

类似于普通 IP 的 Ping/Traceroute，MPLS Ping/Traceroute 使用 MPLS Echo Request 和 MPLS Echo Reply 检测 LSP 的可用性。这两种消息以 UDP 报文格式发送，端口号为 3503。接收端通过 UDP 端口号识别出 MPLS Echo Request 和 MPLS Echo Reply 报文。

MPLS Echo Request 中携带需要检测的 FEC（Forwarding Equivalence Class）信息，和其他属于此 FEC 的报文一样沿 LSP 发送，从而实现对 LSP 的检测。MPLS Echo Request 通过 MPLS 转发给目的端，而 MPLS Echo Reply 则通过 IP 转发给源端。为了防止消息

到达 Egress 节点后又被转发给其他节点，Echo Request 消息的 IP 头中目的地址设置为 127.0.0.1/8（本机环回地址）。

图 10-50 MPLS 网络



- MPLS Ping

如图 10-50，DeviceA 上建立了一条目的地为 DeviceD 的 LSP。从 DeviceA 对该 LSP 进行 MPLS Ping 时的处理如下：

1. DeviceA 查找该 LSP 是否存在。如果不存在，返回错误信息，停止 Ping。如果存在，则继续进行以下操作。
2. DeviceA 构造 MPLS Echo Request 报文，IP 首部目的地址为 127.0.0.1/8，IP TTL = 1。查找相应的 LSP，压入 LSP 的标签（标签的 TTL = 255），将报文发送给 DeviceB。
3. 中间节点 DeviceB 和 DeviceC 对 MPLS Echo Request 报文进行普通 MPLS 转发。
4. 如果中间节点 MPLS 转发失败，则中间节点返回带有错误码的应答消息。
5. 当 MPLS 转发路径无故障，则 MPLS Echo Request 报文到达 LSP 的出节点 DeviceD。DeviceD 处理后返回 MPLS Echo Reply 报文。

- MPLS Traceroute

如图 10-50，从 DeviceA 对 4.4.4.4/32 进行 MPLS traceroute 时的处理如下：

1. DeviceA 检查 LSP 是否存在，如果不存在，返回错误消息，停止 Traceroute，否则继续进行如下处理。
2. DeviceA 构造 MPLS Echo Request 报文，IP 首部目的地址为 127.0.0.1/8，IP TTL = 1。查找相应的 LSP，压入 LSP 的标签（标签的 TTL = 1），将报文发送给 DeviceB。DeviceB 收到此报文，标签的 TTL 超时，因此返回 MPLS Echo Reply 消息。MPLS Echo Reply 消息的目的 UDP 端口和目的 IP 地址就是 MPLS Echo Request 报文的源 UDP 端口和源 IP 地址，IP TTL = 255。
3. DeviceA 收到 MPLS Echo Reply 消息后发送 MPLS Echo Request 报文，其中标签的 TTL = 2。DeviceB 对该报文进行普通 MPLS 转发。DeviceC 收到此报文，标签的 TTL 超时，返回 MPLS Echo Reply 消息。
4. DeviceA 收到 MPLS Echo Reply 消息后发送 MPLS Echo Request 报文，其中标签的 TTL = 3。DeviceB 和 DeviceC 对该报文进行普通 MPLS 转发。DeviceD 收到此报文，发现报文目的地址为本机环回地址，返回 MPLS Echo Reply 消息。

#### 10.5.4.6 MPLS LDP 的基本概念

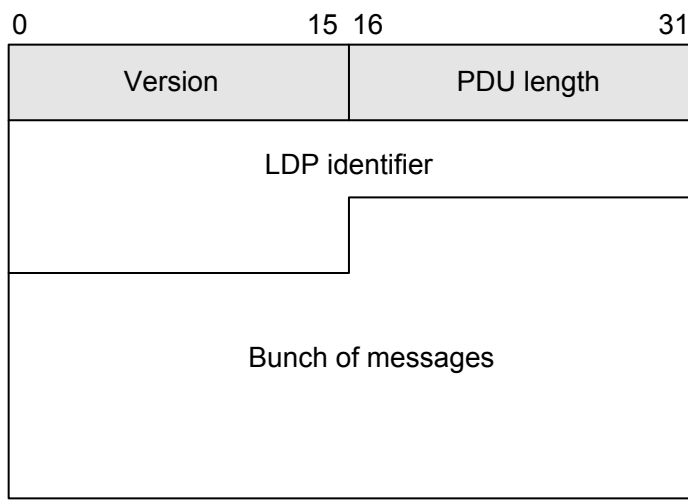
MPLS 体系有多种标签发布协议，LDP（Label Distribution Protocol）是其中使用较广的一种标签发布协议。

LDP 规定了标签分发过程中的各种消息以及相关的处理进程，主要用于 LSR 之间协商标签的分配，进而建立起标签交换路径（LSP）。LSR 之间将依据本地转发表中对应于一个特定 FEC 的入标签、下一跳节点、出标签等信息连接在一起，从而形成跨越整个 MPLS 域的标签交换路径。

LDP 报文

LDP 报文由 LDP PDU（Protocol Data Unit）头部和一组 LDP 消息组成，LDP 报文的格式如图 10-51 所示。

图 10-51 LDP 报文格式



- 其中，LDP PDU 头部包括 3 个域：
- **Version:** 16 比特，表示版本号。目前 LDP 的版本号始终为 1。
  - **PDU length:** 16 比特，表示整组 LDP 消息的总长度，不包括 LDP PDU 头部的长度。  
例如某个 LDP 报文中包含 3 个 Hello 消息，则该报文的 PDU length = 3 \* Message length。
  - **LDP identifier:** 48 比特，表示 LDP 标识符。前 32 比特为 LSR ID，后 16 比特为标签空间。  
例如 LDP 标识符为 192.168.1.1: 0，表示 LSR ID 为 192.168.1.1，标签空间为全局标签空间。
- Bunch of messages 是一组 LDP 消息的集合，可以是一个或者多个 LDP 消息。
- 当 LDP 报文以 UDP 方式传输时，“Bunch of messages”只能是 Hello 消息的集合。
  - 当 LDP 报文以 TCP 方式传输时，“Bunch of messages”可以是除 Hello 消息外任意类型的 LDP 消息的集合。

标签空间（Label Space）

标签空间用于指定标签分配和分发的方式，可以分为：

- 全局标签空间（Per-Platform Label Space）：整个 LSR 只能为特定 FEC 生成唯一的标签。目前主要使用全局标签空间。
- 接口标签空间（Per-Interface Label Space）：LSR 的每个接口都可以为特定 FEC 生成一个标签。

### LDP 标识符（LDP identifier）

LDP 标识符用于标识特定 LSR 的标签空间范围。LDP 标识符的格式为<LSR ID>:<Label space ID>，长度为六字节，其中：

- LSR ID：表示 LSR 编号，占四字节。
- Label space ID：表示标签空间编号，占两字节。全局标签空间为“0”，接口标签空间为“1”。

例如 LDP 标识符为 192.168.1.1:0，表示 LSR ID 为 192.168.1.1，标签空间为全局标签空间。

### LDP 消息类型

LDP 协议主要使用以下四类消息，如表 10-4 所示。

表 10-4 LDP 消息类型

| 消息类型                | 功能                     | 消息名称                                                                                                                                         | 传输类型 |
|---------------------|------------------------|----------------------------------------------------------------------------------------------------------------------------------------------|------|
| 发现（Discovery）消息     | 通告和维护网络中 LSR 的存在。      | Hello 消息                                                                                                                                     | UDP  |
| 会话（Session）消息       | 建立、维护和终止 LDP 对等体之间的会话。 | Initialization 消息<br>Keepalive 消息                                                                                                            | TCP  |
| 通告（Advertisement）消息 | 创建、改变和删除 FEC 的标签映射。    | Address 消息<br>Address withdraw 消息<br>Label request 消息<br>Label mapping 消息<br>Label withdraw 消息<br>Label release 消息<br>Label abort request 消息 | TCP  |
| 通知（Notification）消息  | 提供建议性的消息和差错通知。         | Notification 消息                                                                                                                              | TCP  |

### 10.5.4.7 LDP 会话基本概念

#### LDP 对等体

LDP 对等体是指相互之间存在 LDP 会话、使用 LDP 来交换标签或 FEC 映射关系的两个 LSR。

LDP 对等体通过它们之间的 LDP 会话获得对方的标签映射消息。

## LDP 会话

LDP (Label Distribution Protocol) 会话用于在 LSR (Label Switching Router) 之间进行标签交换。LDP 会话分为以下两种：

- 本地 LDP 会话 (Local LDP Session)：本地 LDP 会话只能在相邻的 LSR 之间建立。
- 远端 LDP 会话 (Remote LDP Session)：远端 LDP 会话主要建立在不相邻的 LSR 之间，也可以建立在相邻的 LSR 之间。

## LDP 使用的定时器

LDP 会话主要使用以下几个定时器：

- Hello 发送定时器：LSR 使用 Hello 定时器周期性地发送 Hello 消息，向邻居 LSR 通告它在网络中的存在，并建立 Hello 邻接关系。Hello 定时器的取值为 Hello 保持定时器的三分之一。
- Hello 保持定时器：建立了 Hello 邻接关系的 LDP 对等体之间，通过周期性发送 Hello 报文表明自己希望继续维持这种邻接关系。如果 Hello 保持定时器超时，没有收到新的 Hello 报文，则拆除 Hello 邻接关系。
- 会话保持定时器：LDP 对等体之间通过 LDP 会话连接上传送的 LDP 协议报文 (PDU) 维持 LDP 会话，如果会话保持定时器超时时没有收到任何 LDP PDU，则关闭连接，结束 LDP 会话。

### 10.5.4.8 LDP 的发现机制

希望建立会话的 LSR 向相邻 LSR 周期性地发送 Hello 消息，通知相邻节点的本地对等关系的机制称为 LDP 发现机制。

LDP 有两种发现机制：

- 基本发现机制

基本发现机制用于发现本地的 LDP 对等体，即通过链路层直接相连的 LSR，建立本地 LDP 会话。

这种方式下，LSR 周期性以 UDP 报文形式从接口发送 LDP 链路 Hello 消息 (LDP Link Hello)，发往标识“子网内所有路由器”的组播地址。

LDP 链路 Hello 消息带有接口的 LDP 标识符及其他相关信息，如果 LSR 在某个接口收到了 LDP 链路 Hello 消息，则表明在该接口 (链路层) 可能存在 LDP 对等体。

- 扩展发现机制

扩展发现机制用于发现远端的 LDP 对等体，即非直连的 LSR，建立远端 LDP 会话。

这种方式下，LSR 周期性以 UDP 报文形式向指定的 IP 地址发送 LDP 目标 Hello 消息 (LDP Targeted Hello)。

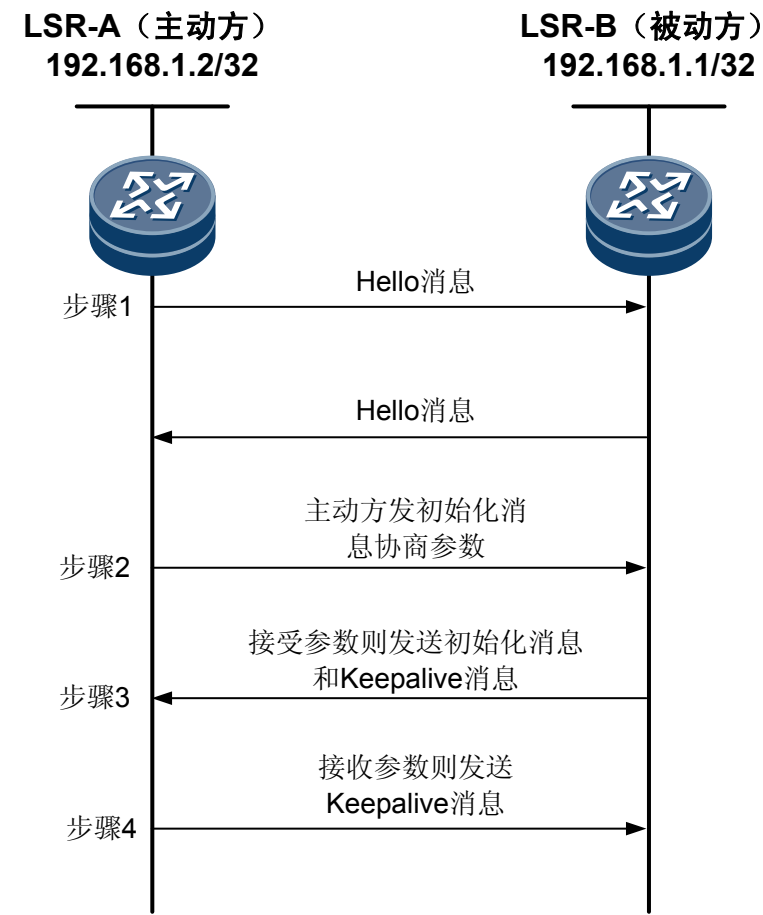
LDP 目标 Hello 消息带有 LSR 的 LDP 标识符及其他相关信息，如果 LSR 收到 LDP 目标 Hello 消息，则表明在网络中可能存在 LDP 对等体。

LSR 之间在 LDP 发现阶段建立 Hello 邻接关系，之后通过进一步的消息交换建立 LDP 会话。

10.5.4.9 LDP 会话的建立

LDP 会话的建立过程如图 10-52 所示。

图 10-52 LDP 会话的建立



下文对 LDP 会话的建立过程进行了详细描述：

1. 两个 LSR 之间互相发送 Hello 消息。Hello 消息中携带传输地址，双方使用传输地址建立 LDP 会话。首先选择传输地址较大的一方作为主动方，发起建立 TCP 连接。如图 10-52 所示，LSR-A 作为主动方发起建立 TCP 连接，LSR-B 作为被动方等待对方发起连接。
2. TCP 连接建立成功后，由主动方 LSR-A 发送 Initialization 消息，协商建立 LDP 会话的相关参数，包括 LDP 协议版本、标签分发方式、定时器值和标签空间等。
3. 被动方 LSR-B 收到 Initialization 消息后，如果不能接受相关参数，则发送 Notification 消息终止 LDP 会话的建立；如果被动方 LSR-B 能够接受相关参数，则发送 Initialization 消息，同时发送 Keepalive 消息。
4. 主动方 LSR-A 收到 Initialization 消息后，如果不能接受相关参数，则发送 Notification 消息终止 LDP 会话的建立；如果能够接受相关参数，则发送 Keepalive 消息。

当双方都收到对端的 Keepalive 消息后，LDP 会话建立成功。

#### 10.5.4.10 LDP 会话的维护

##### 维护 Hello 邻接关系

两个相邻的 LSR 之间可以建立多个 Hello 邻接关系，但是只能建立一个 LDP 会话。LSR 通过周期性地发送 Hello 消息，向邻居 LSR 通告它在网络中的存在，并建立 Hello 邻接关系。然后 LSR 为每个邻居建立一个 Hello 保持定时器，用于维护 Hello 邻接关系，每收到一个 Hello 消息时刷新 Hello 保持定时器。

如果在收到新的 Hello 消息之前 Hello 保持定时器超时，则 LSR 认为 Hello 邻接关系中断。由于一个 LDP 会话可能包含多个 Hello 邻接关系，因此只有所有的 Hello 邻接关系都中断时，LSR 会关闭相应的传输层连接，终止会话进程。

##### 维护 LDP 会话

LDP 通过检测会话连接上传输的 LDP PDU 来判断会话的完整性。

LSR 为每个会话建立一个“生存状态”定时器，每收到一个 LDP PDU 时刷新该定时器。如果在收到新的 LDP PDU 之前定时器超时，LSR 认为会话中断，对等关系失效。LSR 将关闭相应的传输层连接，终止会话进程。

### 10.6 BGP MPLS IP VPN

BGP/MPLS IP VPN 使用 BGP 在服务提供商骨干网上发布 VPN 路由，使用 MPLS 在服务提供商骨干网上转发 VPN 报文。BGP/MPLS IP VPN 组网方式灵活、可扩展性好，并能够方便地支持 MPLS QoS。

#### 10.6.1 介绍

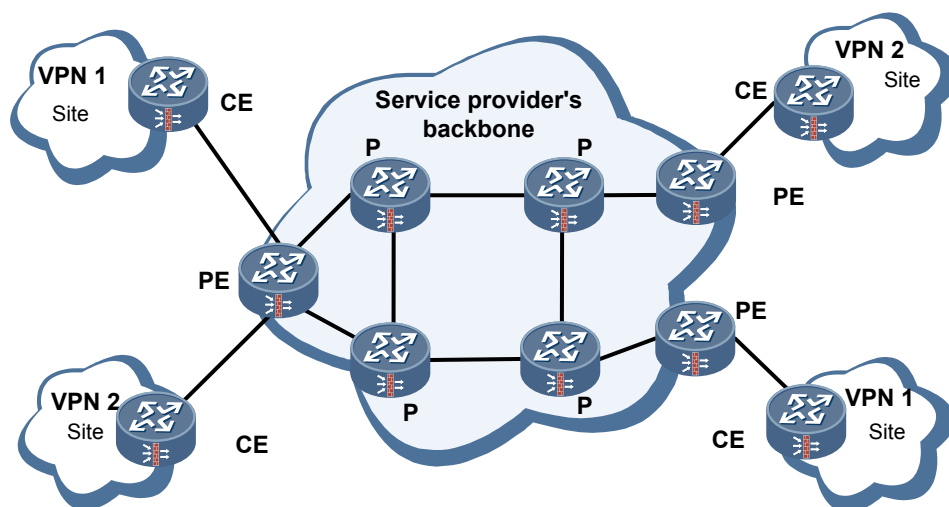
##### 定义

BGP/MPLS IP VPN 是一种 L3VPN（Layer 3 Virtual Private Network）。它使用 BGP（Border Gateway Protocol）在服务提供商骨干网上发布 VPN 路由，使用 MPLS（Multiprotocol Label Switch）在服务提供商骨干网上转发 VPN 报文。这里的 IP 是指 VPN 承载的是 IP 报文。

BGP/MPLS IP VPN 的基本模型如图 10-53 所示。



图 10-53 BGP/MPLS IP VPN 模型



BGP/MPLS IP VPN 的基本模型由三部分组成：CE、PE 和 P。

- CE (Customer Edge)：用户网络边缘设备，有接口直接与服务提供商 SP (Service Provider) 网络相连。CE 可以是路由器或交换机，也可以是一台主机。通常情况下，CE “感知” 不到 VPN 的存在，也不需要支持 MPLS。
- PE (Provider Edge)：是服务提供商网络的边缘设备，与 CE 直接相连。在 MPLS 网络中，对 VPN 的所有处理都发生在 PE 上，对 PE 性能要求较高。
- P (Provider)：服务提供商网络中的骨干设备，不与 CE 直接相连。P 设备只需要具备基本 MPLS 转发能力，不维护 VPN 信息。

PE 和 P 设备仅由 SP 管理；CE 设备仅由用户管理，除非用户把管理权委托给 SP。

一台 PE 设备可以接入多台 CE 设备。一台 CE 设备也可以连接属于相同或不同服务提供商的多台 PE 设备。

## 目的

MPLS 无缝地集成了 IP 路由技术的灵活性和 ATM 标签交换技术的简捷性。MPLS 在无连接的 IP 网络中增加了面向连接的控制平面，为 IP 网络增添了管理和运营的手段。

因此，使用基于 MPLS 的 IP 网络作为骨干网的 VPN (MPLS VPN) 成为在 IP 网络运营商提供增值业务的重要手段，越来越被运营商看好。

BGP 与 IGP 不同，其着眼点不在于发现和计算路由，而在于控制路由的传播和选择最佳路由。VPN 本身就是利用公共网络传递 VPN 数据，而公共网络通常已经应用 IGP 发现和计算自身的路由。构建 VPN 的关键在于控制 VPN 路由的传播，及如何在两个 PE 之间选择最佳的路由。

BGP 使用 TCP 作为其传输层协议 (端口号 179)，提高了协议的可靠性。可以利用这一点来进行跨 USG 的两个 PE 设备之间交换 VPN 路由。

BGP 可以承载附加在路由后的任何信息，作为可选的 BGP 属性，任何不了解这些属性的 USG 都将透明的转发它们。这为在 PE 间传播 VPN 路由提供了便利。

路由更新时，BGP 只发送更新的路由，减少了传播路由所占用的带宽，提供了在公共网络上传播大量的 VPN 路由的可能。



BGP 是一种外部网关协议（EGP），因此实现跨运营商的 VPN 更加容易。

## 10.6.2 参考标准和协议

与 BGP/MPLS IP VPN 特性相关的参考标准与协议如下：

- RFC2858: Multiprotocol Extensions for BGP-4
- RFC2547: BGP/MPLS VPNs
- RFC 4364: BGP/MPLS IP Virtual Private Networks (VPNs)
- RFC2764: A Framework for IP Based Virtual Private Networks
- RFC3392: Capabilities Advertisement with BGP-4
- RFC2917: A Core MPLS IP VPN Architecture
- RFC3107: Carrying Label Information in BGP-4
- RFC4026: Provider Provisioned Virtual Private Network (VPN) Terminology
- RFC4577: OSPF as the Provider/Customer Edge Protocol for BGP/MPLS IP Virtual Private Networks (VPNs)

## 10.6.3 可获得性

### License 支持

USG 系列支持的 VPN 实例个数由购买的虚拟防火墙 License 决定；USG BSR/HSR 系列不受 License 控制，USG2100 BSR/HSR 最多支持 10 个 VPN 实例，USG2200/5100 BSR/HSR 最多支持 100 个 VPN 实例。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 10.6.4 原理描述

### 10.6.4.1 BGP/MPLS IP VPN 基本概念

#### 站点（Site）

在介绍 VPN 时经常会提到“Site”，Site（站点）的含义可以从下述几个方面理解：

- Site 是指相互之间具备 IP 连通性的一组 IP 系统，并且，这组 IP 系统的 IP 连通性不需通过服务提供商网络实现。

- Site 的划分是根据设备的拓扑关系，而不是地理位置，尽管在大多数情况下一个 Site 中的设备地理位置相邻。地理位置隔离的两组 IP 系统，如果它们使用专线互联，不需要通过服务提供商网络就可以互通，那么这两组 IP 系统也组成一个 Site。
- 一个 Site 中的设备可以属于多个 VPN，换言之，一个 Site 可以属于多个 VPN。
- Site 通过 CE 连接到服务提供商网络，一个 Site 可以包含多个 CE，但一个 CE 只属于一个 Site。

根据 Site 的情况，建议 CE 设备选择方案如下：

- 如果 Site 只是一台主机，则这台主机就作为 CE 设备；
- 如果 Site 是个子网，则使用交换机作为 CE 设备；
- 如果 Site 是多个子网，则使用路由器作为 CE 设备。

对于多个连接到同一服务提供商网络的 Sites，通过制定策略，可以将它们划分为不同的集合（set），只有属于相同集合的 Sites 之间才能通过服务提供商网络互访，这种集合就是 VPN。

## VPN 实例

VPN 实例（VPN-instance）也称为 VPN 路由转发表 VRF（VPN Routing and Forwarding table）。PE 上存在多个路由转发表，包括一个公网路由转发表，以及一个或多个 VPN 路由转发表。也就是说，PE 上存在多个实例，包括一个公网实例和一个或多个 VPN 实例。

公网路由转发表与 VPN 路由转发表的不同在于：

- 公网路由表包括所有 PE 和 P 路由器的 IPv4 路由，由骨干网的路由协议或静态路由产生。
- VPN 路由表包括属于该 VPN 实例的所有 site 的路由，通过 CE 与 PE 之间或者两个 PE 之间的 VPN 路由信息交互获得。
- 公网转发表是根据路由管理策略从公网路由表提取出来的最小转发信息；而 VPN 转发表是根据路由管理策略从对应的 VPN 路由表提取出来的最小转发信息。

可以看出，PE 上的各 VPN 实例之间相互独立，并与公网路由转发表相互独立。

可以将每个 VPN 实例看作一台虚拟的路由器：维护独立的地址空间、有连接到该路由器的接口。

在 RFC2547（BGP/MPLS VPNs）中，VPN 实例被称为 per-site forwarding table，顾名思义，VPN 实例与 site 对应。更准确的描述是：每条 CE 与 PE 的连接对应一个 VPN 实例（但不是一一对应关系），实现这种对应关系的方法是将 VPN 实例和 PE 上与 CE 直接相连的接口关联（或称为绑定），这需要手工设置。

VPN 实例通过路由标识符 RD（Route Distinguisher）实现地址空间独立，通过 VPN Target 属性实现直连 site 的 VPN 成员关系和路由规则控制。

## VPN、Site 和 VPN 实例的关系

VPN 实例、Site、VPN 之间的关系如下：

- VPN 是多个 Site 的组合。一个 Site 可以属于多个 VPN。
- 每一个 Site 在 PE 上都关联一个 VPN 实例。VPN 实例综合了它所关联的 Site 的 VPN 成员关系和路由规则。多个 Site 根据 VPN 实例的规则组合成一个 VPN。
- VPN 实例与 VPN 不是一一对应的关系。

## 地址空间重叠

PE 从 CE 接收到私网路由后，需要将这些路由发布给其他 PE。

VPN 是一种私有网络，不同的 VPN 独立管理自己的地址范围，也称为地址空间（address space）。不同 VPN 的地址空间可能会在一定范围内重合，例如，VPN1 和 VPN2 都使用 10.110.10.0/24 网段地址，这就发生了地址空间的重叠（address spaces overlapping）。

以下两种情况允许 VPN 使用重叠的地址空间：

- 两个 VPN 没有共同的 site；
- 两个 VPN 有共同的 site，但此 site 中的设备不与两个 VPN 中使用重叠地址空间的设备互访。

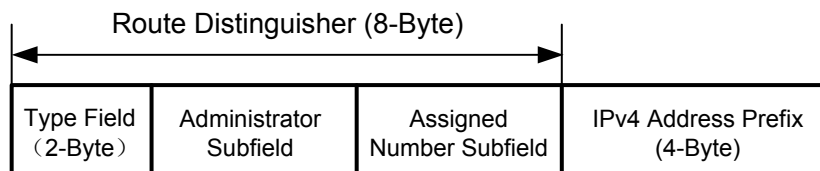
## VPN-IPv4 地址

传统 BGP 无法正确处理地址空间重叠的 VPN 的路由。假设 VPN1 和 VPN2 都使用了 10.110.10.0/24 网段的地址，并各自发布了一条去往此网段的路由。不同 VPN 的路由之间不进行负载分担，因此 BGP 根据自己的选路规则只选择其中一条路由，从而导致去往另一个 VPN 的路由丢失。

产生上述问题的原因是 BGP 无法区分不同 VPN 中相同的 IP 地址前缀，为解决这一问题，BGP/MPLS IP VPN 使用了 VPN-IPv4 地址族。

VPN-IPv4 地址共有 12 个字节，包括 8 字节的路由标识符 RD（Route Distinguisher）和 4 字节的 IPv4 地址前缀，如图 10-54 所示。

图 10-54 VPN-IPv4 地址结构



增加了 RD 的 IPv4 地址称为 VPN-IPv4 地址。PE 从 CE 接收到 IPv4 路由后，转换为全局唯一的 VPN-IPv4 路由，并在公网上发布。RD 为零的 VPN-IPv4 地址相当于普通 IPv4 地址。

## RD

如图 10-54 所示，RD 有两种格式：

- Type 为 0 时，Administrator 子字段占 2 字节，Assigned Number 子字段占 4 字节。其中，Administrator 子字段是 16bits 自治系统号；Assigned Number 子字段是 32bits 用户自定义数字。
- Type 为 1 时，Administrator 子字段占 4 字节，Assigned Number 子字段占 2 字节。其中，Administrator 子字段是 32bits IPv4 地址；Assigned Number 子字段是 16bits 用户自定义数字。



说明

配置 RD 时，只需要指定 RD 的 Administrator 子字段和 Assigned Number 子字段。RD 的配置格式有两种：

- 16bits 自治系统号:32bits 用户自定义数字（例如：100:1）。
- 32bits IPv4 地址:16bits 用户自定义数字（例如：172.1.1.1:1）。

RD 用于区分使用相同地址空间的 IPv4 前缀。RD 的结构使得每个服务供应商可以独立地分配 RD，但为了在 CE 双归属的情况下保证路由正常，必须保证 RD 全局唯一。

如果 RD 非全局唯一，在 CE 双归属的情况下将不能正确路由。

## VPN Target

BGP/MPLS IP VPN 使用 32 位的 BGP 扩展团体属性—VPN Target（也称为 Route Target）来控制 VPN 路由信息的发布。

每个 VPN 实例关联一个或多个 VPN Target 属性。有两类 VPN Target 属性：

- **Export Target:** 本地 PE 从直接相连 site 学到 IPv4 路由后，转换为 VPN IPv4 路由，并为这些路由设置 Export Target 属性。Export Target 属性作为 BGP 的扩展团体属性随路由发布。
- **Import Target:** PE 收到其它 PE 发布的 VPN-IPv4 路由时，检查其 Export Target 属性。当此属性与 PE 上某个 VPN 实例的 Import Target 匹配时，PE 就把路由加入到该 VPN 实例的路由表。

也就是说，VPN Target 属性定义了一条 VPN 路由可以为哪些 Site 所接收，以及 PE 可以接收哪些 Site 发送来的路由。

当收到直连 CE 传过来的路由时，PE 将该路由与一个或多个 Export Target 属性关联。Export Target 属性将和 VPN-IPv4 路由一起由 BGP 发布给其他相关的 PE。当这些相关的 PE 收到该 VPN-IPv4 路由时，将其 Export Target 属性与本设备所有的 VPN 实例的 Import Target 属性值比较。如果相等，就将该路由注入到该 VPN 路由表。

### 10.6.4.2 Hub and Spoke

如果希望在 VPN 中设置中心访问控制设备，其它用户的互访都通过中心访问控制设备进行，可以使用 Hub & Spoke 组网方案。其中，中心访问控制设备所在站点称为 Hub 站点，其他用户站点称为 Spoke 站点。Hub 站点侧接入 VPN 骨干网的设备叫 Hub-CE；Spoke 站点侧接入 VPN 骨干网的设备叫 Spoke-CE。VPN 骨干网侧接入 Hub 站点的设备叫 Hub-PE，接入 Spoke 站点的设备叫 Spoke-PE。

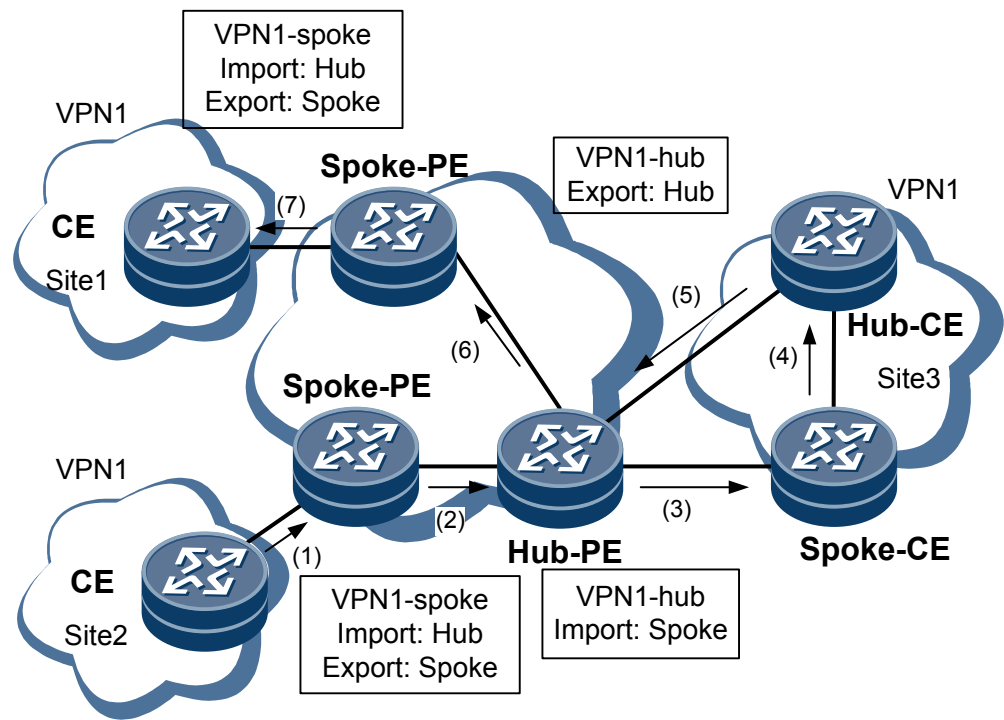
Spoke 站点需要把路由发布给 Hub 站点，再通过 Hub 站点发布给其他 Spoke 站点。Spoke 站点之间不直接通过路由。Hub 站点对 Spoke 站点之间的通讯进行集中控制。

对于这种组网情况，需要设置两个 VPN Target，一个表示“Hub”，另一个表示“Spoke”。

各 Site 在 PE 上的 VPN 实例的 VPN Target 设置规则为：

- 连接 Spoke 站点的 PE（Spoke-PE）：Export Target 为“Spoke”，Import Target 为“Hub”；
- 连接 Hub 站点的 PE（Hub-PE）：Hub-PE 上需要使用两个接口或子接口，一个用于接收 Spoke-PE 发来的路由，其 VPN 实例的 Import Target 为“Spoke”；另一个用于向 Spoke-PE 发布路由，其 VPN 实例的 Export Target 为“Hub”。

图 10-55 Hub & Spoke 组网 Site2 到 Site1 的路由发布途径

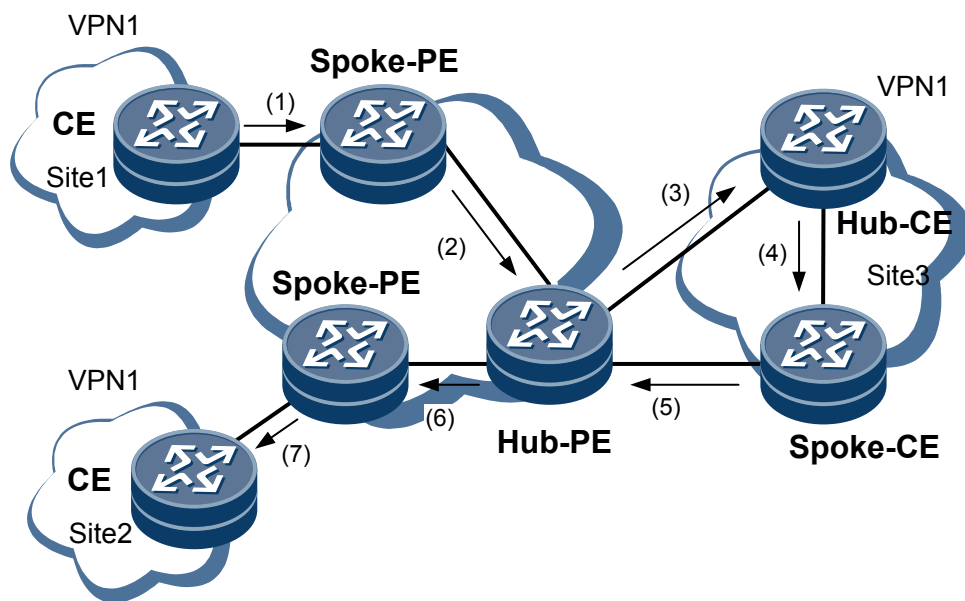


在图 10-55 中，Spoke 站点之间的通信通过 Hub 站点进行（图中箭头所示为 Site2 的路由向 Site1 的发布过程）：

- Hub-PE 能够接收所有 Spoke-PE 发布的 VPN-IPv4 路由；
- Hub-PE 发布的 VPN-IPv4 路由能够为所有 Spoke-PE 接收；
- Hub-PE 将从 Spoke-PE 学到的路由发布给 Spoke-CE，并将从 Hub-CE 学到的路由发布给所有 Spoke-PE。因此，Spoke 站点之间可以通过 Hub 站点互访。
- 任意 Spoke-PE 的 Import Target 属性不与其它 Spoke-PE 的 Export Target 属性相同。因此，任意两个 Spoke-PE 之间不直接发布 VPN-IPv4 路由，Spoke 站点之间不能直接互访。

图 10-55 中的 Site1 和 Site2 之间通讯数据的传输路径请参见图 10-56（图中箭头所示为数据传输方向）。

图 10-56 Site1 到 Site2 的数据传输途径



### 10.6.4.3 跨域 VPN

随着 MPLS VPN 解决方案的广泛应用，国内运营商的不同城域网之间，或相互协作的不同运营商的骨干网之间都存在着跨越不同自治域的情况。

一般的 MPLS VPN 体系结构都是在一个自治系统内运行，任何 VPN 的路由信息都是可以在一个自治系统内按需扩散，没有提供自治系统内的 VPN 信息向其他自治系统扩散的功能。因此，为了支持运营商之间的 VPN 路由选择信息交换，就需要扩展现有的协议和修改 MPLS VPN 体系框架，提供一个不同于基本的 MPLS VPN 体系结构所提供的互连模型——跨域（Inter-AS）的 MPLS VPN，以便可以穿过运营商间的链路来发布路由前缀和标签信息。

RFC2547bis 中提出了三种跨域 VPN 解决方案，分别是：

- 跨域 VPN-OptionA（Inter-Provider Backbones Option A）方式：需要跨域的 VPN 在 ASBR 间通过专用的接口管理自己的 VPN 路由，也称为 VRF-to-VRF；
- 跨域 VPN-OptionB（Inter-Provider Backbones Option B）方式：ASBR 间通过 MP-EBGP 发布标签 VPN-IPv4 路由，也称为 EBGP redistribution of labeled VPN-IPv4 routes；
- 跨域 VPN-OptionC（Inter-Provider Backbones Option C）方式：PE 间通过 Multi-hop MP-EBGP 发布标签 VPN-IPv4 路由，也称为 Multihop EBGP redistribution of labeled VPN-IPv4 routes。

### 跨域 VPN-OptionA 方式

跨域 VPN-OptionA 实际是基本 BGP/MPLS IP VPN 在跨域环境下的应用，不需要专门配置。这种方式下，两个 AS 的边界路由器 ASBR 直接相连，ASBR 同时也是各自所在自治系统的 PE（这里被称为 ASBR PE）。两个 ASBR PE 都把对端 ASBR PE 看作自己的 CE 设备，使用 EBGP 方式向对端发布 IPv4 路由。

VRF-to-VRF 方式实现跨域 VPN 的优点是简单：ASBR PE 之间不需要运行 MPLS，也不需要为跨域进行特殊配置。

缺点是可扩展性差：ASBR PE 需要管理所有 VPN 路由，为每个 VPN 创建 VPN 实例。这将导致 PE 上的 VPN-IPv4 路由数量过大。并且，由于 ASBR PE 间是普通的 IP 转发，要求为每个跨域的 VPN 使用不同的接口（可以是子接口、物理接口、捆绑的逻辑接口），从而提高了对 PE 设备的要求。如果跨越多个自治域，中间域必须支持 VPN 业务，不仅配置量大，而且对中间域影响大。在需要跨域的 VPN 数量比较少的情况，可以考虑使用。

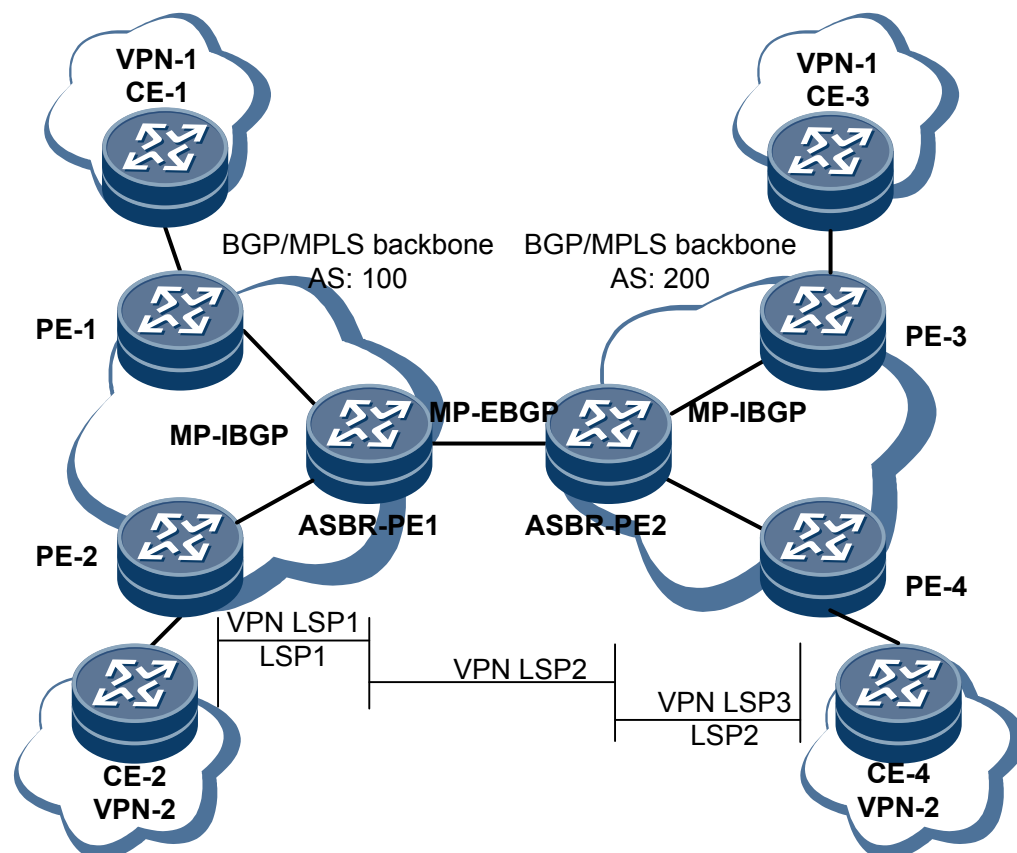
## 跨域 VPN-OptionB 方式

跨域 VPN-OptionB 中，两个 ASBR 通过 MP-EBGP 交换它们从各自 AS 的 PE 路由器接收的标签 VPN-IPv4 路由。

路由发布过程如下：

1. AS100 内的 PE 先通过 MP-IBGP 方式把标签 VPN-IPv4 路由发布给 AS100 的 ASBR PE，或发布给为 PE 反射路由的路由反射器 RR（Route Reflector）。
2. ASBR-PE1 通过 MP-EBGP 方式把标签 VPN-IPv4 路由发布给 ASBR-PE2。
3. ASBR-PE2 再通过 MP-IBGP 方式把标签 VPN-IPv4 路由发布给 AS200 内的 PE，或发布给为 PE 反射路由的路由反射器。

图 10-57 ASBR 间通过跨域 VPN-OptionB 方式发布标签 VPN-IPv4 路由组网图



跨域 VPN-OptionB 方案中, ASBR 接收本域内和域外传过来的所有跨域 VPN 的路由, 再把 VPN 发布出去。但 MPLS VPN 的基本实现中, PE 上只保存与本地 VPN 实例的 VPN-Target 相匹配的 VPN 路由。因此, 如果 ASBR 上没有配置对应的 VPN 实例, 可采取以下三种方法:

- ASBR 对标签 VPN-IPv4 路由进行特殊处理, 让 ASBR 把收到的 VPN 路由全部的保存下来, 而不管本地是否有和它匹配的 VPN 实例。

采用该方案时, 需要注意:

- ASBR 之间不对接收的 VPN-IPv4 路由进行 VPN Target 过滤, 因此, 交换 VPN-IPv4 路由的各 AS 服务提供商之间需要就这种路由交换达成信任协议;
- VPN-IPv4 路由交换仅发生在私网对等点之间, 不能与公网交换 VPN-IPv4 路由, 也不能与没有达成信任协议的 MP-EBGP 对等体交换 VPN-IPv4 路由。

这种方案的优点是所有的流量都经过 ASBR 转发, 流量的可控性较好, 但 ASBR 的负担重。

- 使用 BGP 路由策略 (如对 RT 的过滤) 控制 VPN-IPv4 路由信息的收发。
- 在 ASBR 上配置需要通过该 ASBR 传递路由的 VPN 实例, 但不绑定任何接口。

与 Option A 相比, 不受 ASBR 之间互连链路数目的限制。

此方案也有局限性: VPN 的路由信息是通过 AS 之间的 ASBR 路由器来保存和扩散的, 当 VPN 路由较多时, ASBR 负担重, 容易成为故障点。因此在 MP-EBGP 方案中, 需要维护 VPN 路由信息的 ASBR 一般不再负责公网 IP 转发。

## 跨域 VPN-OptionC 方式

前面介绍的两种方式都能够满足跨域 VPN 的组网需求, 但这两种方式也都需要 ASBR 参与 VPN-IPv4 路由的维护和发布。当每个 AS 都有大量的 VPN 路由需要交换时, ASBR 就很可能成为阻碍网络进一步扩展的瓶颈。

解决上述问题的方案是: ASBR 不维护或发布 VPN-IPv4 路由, PE 之间直接交换 VPN-IPv4 路由。

- ASBR 通过 MP-IBGP 向各自 AS 内的 PE 路由器发布标签 IPv4 路由, 并将到达本 AS 内 PE 的标签 IPv4 路由通告给它在对端 AS 的 ASBR 对等体, 过渡自治系统中的 ASBR 也通告带标签的 IPv4 路由。这样, 在入口 PE 和出口 PE 之间建立一条 LDP LSP;
- 不同 AS 的 PE 之间建立 Multihop 方式的 EBGP 连接, 交换 VPN-IPv4 路由;
- ASBR 上不保存 VPN-IPv4 路由, 相互之间也不通告 VPN-IPv4 路由。

跨域 VPN-OptionC 组网方案有如下优势:

- VPN 路由在入口 PE 和出口 PE 之间直接交互, 不需要中间设备的保存和转发。
- VPN 的路由信息只出现在 PE 设备上, 而 P 和 ASBR 路由器只负责报文的转发, 使得中间域的设备可以不支持 MPLS VPN 业务, 只需支持 IP 转发, ASBR 设备不再成为性能瓶颈。因此跨域 VPN-OptionC 更适合在跨越多个 AS 时使用。
- 更适合支持 MPLS VPN 的负载分担。

### 10.6.4.4 HoVPN

## HoVPN 的产生

在 BGP/MPLS IP VPN 中, PE 设备最为关键, 它完成两方面的功能:



- 为用户提供接入功能，这需要 PE 具有大量接口；
- 管理和发布 VPN 路由，处理用户报文，这需要 PE 设备具有大容量内存和高转发能力。

目前的网络设计大多采用经典的分层结构，例如，城域网的典型结构是三层模型：核心层、汇聚层、接入层。从核心层到接入层，对设备的性能要求依次下降，网络规模则依次扩大。

而 BGP/MPLS IP VPN 是一种平面模型，对网络中所有 PE 设备的性能要求相同，当网络中某些 PE 在性能和可扩展性方面存在问题时，整个网络的性能和可扩展性将受到影响。

由于 BGP/MPLS IP VPN 的平面模型与典型的分层网络模型不相符，在每一个层次上部署 PE 都会遇到扩展性问题，不利于大规模部署 VPN。

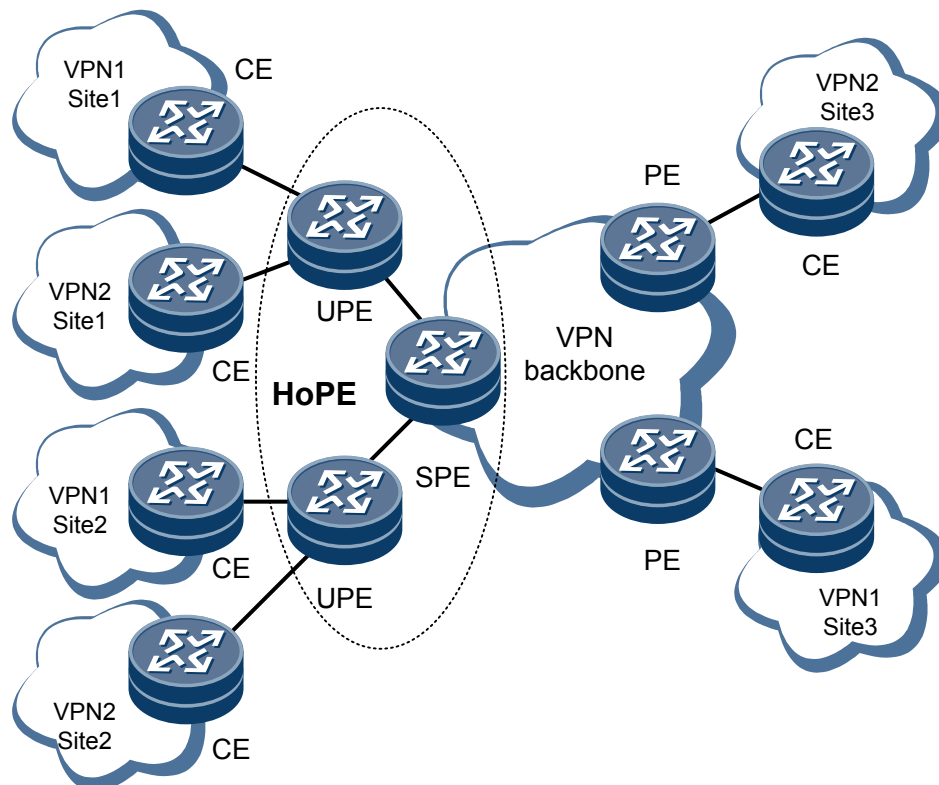
为解决可扩展性问题，BGP/MPLS IP VPN 必然要从平面模型转变为分层模型。

分层 VPN（Hierarchy of VPN，简称 HoVPN）解决方案将 PE 的功能分布到多个 PE 设备上，多个 PE 承担不同的角色，并形成层次结构，共同完成一个 PE 的功能。因此，这种解决方案有时也被称为分层 PE（Hierarchy of PE，HoPE）。

HoVPN 对处于较高层次的设备的路由能力和转发性能要求较高，而对处于较低层次的设备的相应要求也较低，符合典型的分层网络模型。

## HoVPN 的基本结构

图 10-58 HoVPN 的基本结构



在图 10-58 中，直接连接用户的设备称为下层 PE（Underlayer PE）或用户侧 PE（Userend PE），简称 UPE；连结 UPE 并位于网络内部的设备称为上层 PE（Superstratum PE）或运营商侧 PE（Service Provider-end PE），简称 SPE。

SPE 与 UPE 的关系是：

- UPE 主要完成用户接入功能。UPE 维护其直接相连的 VPN site 的路由，但不维护 VPN 中其它远端 site 的路由或仅维护它们的聚合路由；UPE 为其直接相连的 site 的路由分配内层标签，并通过 MP-BGP 随 VPN 路由发布此标签给 SPE；
- SPE 主要完成 VPN 路由的管理和发布。SPE 维护其通过 UPE 连接的 VPN 所有路由，包括本地和远端 site 的路由，但 SPE 不发布远端 site 的路由给 UPE，只发布 VPN 实例的缺省路由或聚合路由，并携带标签；
- UPE 和 SPE 之间采用标签转发，只需要一个接口连接，SPE 不需要使用大量接口来接入用户。UPE 和 SPE 之间的接口可以是物理接口、子接口（如 VLAN，PVC）或隧道接口（如 GRE、LSP）。采用隧道接口时，SPE 和 UPE 之间可以相隔一个 IP 网络或 MPLS 网络，UPE 或 SPE 发出的标签报文经过隧道传递。如果是 GRE 隧道，要求 GRE 支持对 MPLS 报文的封装。

由于分工的不同，对 SPE 和 UPE 的要求也不同：SPE 的路由表容量大，转发性能强，但接口资源较少；UPE 的路由容量和转发性能较低，但接入能力强。

需要说明的是，SPE 和 UPE 是相对的。在多个层次的 PE 结构中，上层 PE 相对于下层就是 SPE，下层 PE 相对于上层就是 UPE。

分层式 PE 可以和普通 PE 共存于一个 MPLS 网络。

## SPE-UPE

SPE 和 UPE 之间运行 MP-BGP，根据 UPE 和 SPE 是否属于同一个 AS，可以是 MP-IBGP，也可以是 MP-EBGP。

采用 MP-IBGP 时，为了在 IBGP 对等体之间通告路由，SPE 可以作为多个 UPE 的路由反射器。SPE 作为 UPE 的路由反射器时，为了减少 UPE 上的路由条数，建议 SPE 不再作为其它 PE 的路由反射器。

## HoVPN 的优势

HoVPN 组网方案具有以下优势：

- BGP/MPLS VPN 可以逐层部署。当 UPE 的性能不够的时候，可以添加一个 SPE，将 UPE 的位置下移。当 SPE 的接入能力不足的时候，可以为其添加 UPE。
- UPE 和 SPE 之间采用标签转发，因而只需要一个(子)接口相互连接，节约有限的接口资源。
- 若 UPE 和 SPE 之间相隔一个 IP/MPLS 网络，采用 GRE 或 LSP 等隧道连接。在分层部署 MPLS VPN 时，有良好的可扩展性。
- UPE 上只需维护本地接入的 VPN 路由，所有远端路由都用一条缺省或聚合路由替代，大大减轻了 UPE 的负担。
- SPE 和 UPE 通过动态路由协议 MP-BGP 交换路由、发布标签。每一个 UPE 只要运行一个 MP-BGP 对等体，协议开销小，配置工作量小。

### 10.6.4.5 OSPF Sham link

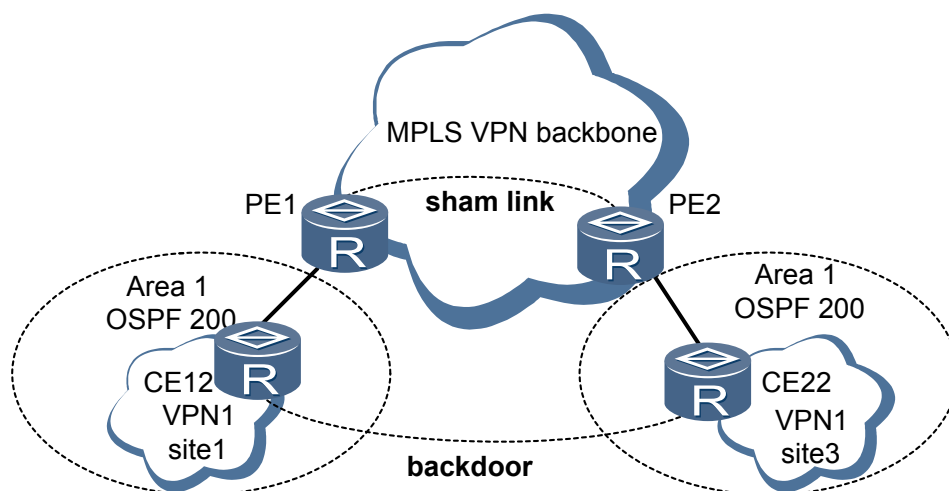
## Sham link 的产生

通常情况下，BGP 对等体之间通过 BGP 扩展团体属性在 MPLS VPN 骨干网上承载路由信息。另一端 PE 上运行的 OSPF 可利用这些信息来生成 PE 到 CE 的 Type-3 LSA，这些路由是区域间路由。

如图 10-59 所示：在 OSPF 的 PE-CE 连接中，同一个 OSPF 区域中有两个 site 连接到不同的 PE，两个 site 属于同一个 VPN，它们之间存在一条区域内 OSPF 链路（backdoor link）。可以将这两个 site 看成是一个 site（多归属）。

这种情况下，通过 PE 连接两个 site 的路由将作为区域间路由（Inter-Area Route），其优先级低于经过 backdoor 链路的区域内路由（Intra-Area Route），不被 OSPF 优选。

图 10-59 Sham link 应用示意图



上述情况导致 VPN 流量总是通过后门路由转发，而不走骨干网。为了避免这一问题，可以在 PE 设备之间建立 OSPF 伪连接（Sham link），使经过 MPLS VPN 骨干网的路由也成为 OSPF 区域内路由。

## Sham link 的实现

Sham link 作为区域内的一条点到点链路，包含在 Type1 LSA 中发布。用户可以通过调整度量值在 Sham link 和 backdoor 之间进行选路。

Sham link 被看成是两个 VPN 实例之间的链路，每个 VPN 实例中必须有一个 Sham link 的端点地址，它是 PE 设备上 VPN 地址空间中的一个有 32 位掩码的 Loopback 接口地址。同一个 OSPF 进程的多条 Sham link 可以共用端点地址，但不同 OSPF 进程不能拥有两条端点地址完全相同的 Sham link。

Sham link 的端点地址被 BGP 作为 VPN-IPv4 地址发布。如果路由经过了 Sham link，它就不能再以 VPN-IPv4 路由的形式被引入到 BGP。

## Sham link 的路由发布过程

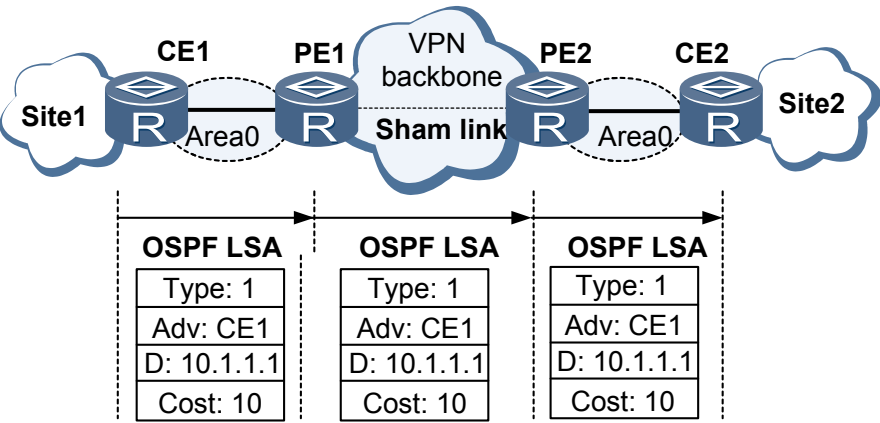
Sham link 在两个 PE 之间建立起来后，这两个 PE 将成为 Sham link 邻居，交互 LSDB。此时骨干网将被看成是点到点的区域内部链路，PE 通过其发布 Type1 LSA。

此时，Sham link 中的路由发布过程如图 10-60。其中，Adv 表示路由发布者（Advertiser）；D 表示报文的目的地址，Cost 表示 Sham link 的路由开销。

**说明**

如果把 Sham link 看成一条链路，则 Sham link 的路由开销相当于该链路接口的 OSPF 路由开销。为了使 VPN 流量通过 MPLS 骨干网转发，Sham link 的路由开销应小于 backdoor 链路的 OSPF 路由开销。

图 10-60 Sham link 路由发布



从图 10-60 可以看出，Sham link 路由发布整个过程，LSA 的类型、发布者及 Cost 值始终不变，相当于 LSA 从 PE1 透明传送到 PE2。

10.6.4.6 Multi-VPN-Instance CE

Multi-VPN-Instance CE 的产生

单纯使用传统路由器很难实现局域网中不同业务的完全隔离。目前，局域网不同业务的隔离一般是通过交换机的 VLAN 功能实现的。但交换机的路由功能相对较弱。因此，为了保证局域网的安全隔离并提高局域网的路由能力，需要同时布置路由器和交换机，这对小网络而言是不经济的。

在一台路由器上运行多个 OSPF 进程，将不同的进程绑定到不同的 VPN 实例。就像在 PE 上使用 OSPF 多实例一样，在 CE 通过为不同的业务建立各自的 OSPF 实例，相当于不同的业务使用不同的虚拟 CE 设备。这样，可以用较低成本实现不同业务的隔离，保证各自的安全性。

OSPF 多实例通常运行在 PE 设备上，在用户局域网内部运行 OSPF 多实例的设备称为 Multi-VPN-Instance CE，即多实例 CE。与 PE 上的 OSPF 多实例相比，Multi-VPNInstance CE 不需要支持 BGP/OSPF 互操作功能。

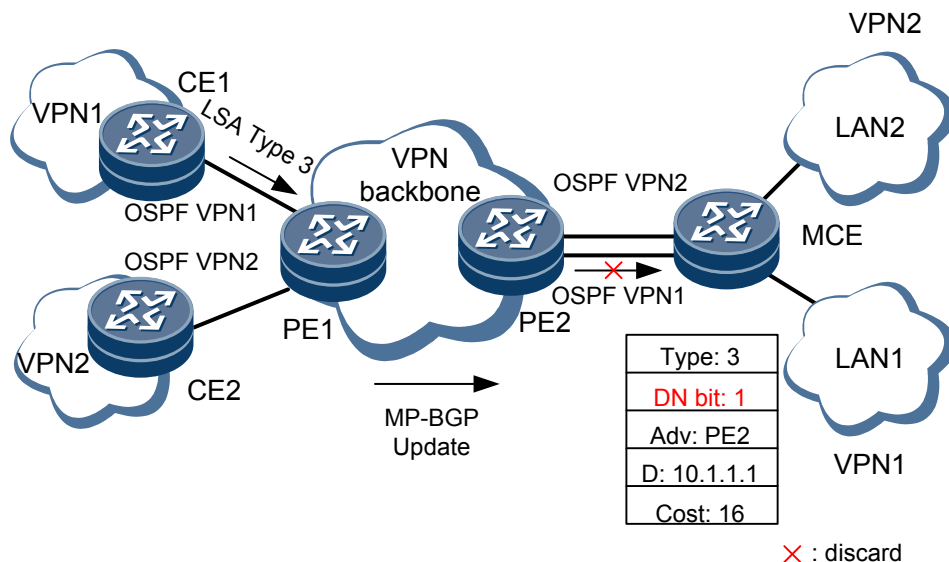
Multi-VPN-Instance CE 主要用于以较低的成本解决局域网的安全问题。

实现原理

Multi-VPN-Instance CE 的实现原理较简单，就是在同一台 CE 和 PE 设备上实现不同的 OSPF 多实例。关键的实现在于禁止路由环的检查，直接进行路由计算。

通常情况下，如果 PE-CE 之间运行 OSPF，则 PE 设备对接收的 Summary LSA 进行 DN 位检查，以防止 PE 又从 CE 学到发出的 LSA 而引起环路。但在运行 OSPF VPN 多实例的非 PE 设备上，对路由环的检测反而不能正确进行路由选路。

图 10-61 MCE 进行环路检测的情况



如图 10-61 所示，PE2 向 MCE 发布的 OSPF Type3 LSA 时，由于使用 OSPF 多实例发布，该 3 类 LSA 中的 DN 被置位。当 MCE 收到该 LSA 时，发现其 DN 位置位，就会丢弃该 LSA。这样，MCE 上就没有到 CE1 站点的路由。

如果在 MCE 上禁止路由环的检查，对 PE2 来的 3 类 LSA，不管其 DN 位是否置位，都使用它来进行路由计算。这样，MCE 就可以正确接收来自 CE1 站点的路由。

#### 10.6.4.7 VPN 与 Internet 互联

一般 VPN 内的用户只能相互通信，不能与 Internet 用户通信，也不能接入 Internet。但 VPN 的各个 Site 可能有访问 Internet 的需要。为了实现 VPN 与 Internet 互联，需要满足以下条件：

- 要访问 Internet 的用户设备必须有到达 Internet 目的地址的路由；
- 有从 Internet 返回的路由；
- 如果需要访问公网的 VPN 用户的地址是私网地址，则需要网络地址转换（NAT），把私网地址转换成公网地址。
- 像非 VPN 用户与 Internet 互联方式一样，必须采用一定的安全机制（如使用防火墙）。

有三种实现方法：

- 在骨干网边缘路由器 PE 侧实现，该 PE 负责区别两种不同的数据流，并分别转发至 VPN 及 Internet。同时，在 VPN 与 Internet 两个域之间提供防火墙功能。

在 PE 侧实现，与 VPN 接入使用同一个接口，节约接口资源，并且不同的 VPN 可以共享一个公有 IP 地址；缺点是在 PE 上实现复杂，且存在安全隐患：

- 如果 CE 使用逻辑链路接入 PE 访问 Internet，来自 Internet 的恶意的流量攻击会使得 PE-CE 链路饱和，从而使得正常的 VPN 数据包无法传输。
- 不论 CE 使用逻辑链路还是物理链路接入 PE 访问 Internet，该 PE 设备都有可能受到 Internet 的 DoS（Denial of Service）攻击。
- 在用户侧实现。此时，由私网边缘路由器 CE 区分两种不同的数据流，并分别引导两个不同的域：一个通过 PE 边缘路由器接入 VPN，一个通过不包含在 VPN 内的 ISP 路由器接入 Internet。同时，CE 设备提供防火墙功能。

在用户侧实现，其实现方法简单，公网和私网路由隔离，安全可靠；但缺点是需使用单独的接口，占用接口资源，并且每个 VPN 都需要单独使用一个公网地址。
- 在 Internet 网关侧实现。这里的 Internet 网关是指接入 Internet 的运营商设备，必须具备 VPN 路由管理功能，可以是不接入任何 VPN 用户的 PE 设备。

在 Internet 网关处实现，比在 PE 侧实现安全性高，但 Internet 网关要创建多个 VPN 实例，负担重。且 Internet 网关要使用多个接口接入 Internet，占用多个公有 IP 地址，每个 VPN 使用一个接口和一个公有 IP 地址。

三种 VPN 与 Internet 互联的实现方法比较如表 10-5 所示。

表 10-5 三种 VPN 与 Internet 互联的实现方法比较

| 实现方法             | 安全性  | 使用接口                                | 使用公有 IP 地址              | NAT 转换             | 实现难易程度 |
|------------------|------|-------------------------------------|-------------------------|--------------------|--------|
| 在用户侧实现           | 相对较高 | 每个 VPN 单独使用一个接口，占用用户接口资源            | 每个 VPN 单独使用一个公有 IP 地址   | 在 CE 侧进行           | 实现简单   |
| 在 PE 侧实现         | 相对较低 | Internet 接入与 VPN 接入使用同一个接口，节约接口资源   | PE 上多个 VPN 共用一个公有 IP 地址 | 一般在 PE 侧进行         | 实现复杂   |
| 在 Internet 网关侧实现 | 相对较高 | 每个 VPN 单独使用一个接口，占用 Internet 网关的接口资源 | 每个 VPN 单独使用一个公有 IP 地址   | 一般在 Internet 网关侧进行 | 实现复杂   |

## 10.7 SSL VPN

SSL 建立在传输层之上，可实现加密算法、通信密钥的协商以及服务器的认证功能。主要提供业务有 Web 代理、网络扩展、文件共享和端口转发。

### 10.7.1 介绍



## 定义

SSL (Secure Sockets Layer) 是一个安全协议, 为基于 TCP 的应用层协议提供安全连接, 如 SSL 可以为 HTTP 协议提供安全连接。SSL 协议广泛应用于电子商务、网上银行等领域, 为网络上数据的传输提供安全性保证。

SSL VPN 是以 SSL/TLS 协议为基础, 利用标准浏览器都内置支持 SSL/TLS 的现实优势, 对其应用功能进行扩展而产生的新型 VPN。除了 TCP 应用, SSL VPN 还能够对 IP 通信进行保护。SSL VPN 通信基于标准 TCP, 不受 NAT 限制, 使用户在任何地方都能够通过 SSL VPN 虚拟网关访问内网资源, 使远程安全接入更加灵活简单, 大大降低了企业部署维护 VPN 的费用。

## 目的

为连接和数据文件传输提供安全性保障。

SSL 提供的安全连接可以实现:

- 连接的私密性  
在 SSL 握手阶段生成密钥后, 用对称加密算法对传输数据进行加密。
- 身份认证  
对服务器和客户端进行基于证书的身份认证, 其中客户端认证是可选的。
- 连接的可靠性  
消息传输过程中使用基于密钥的 MAC (Message Authentication Code) 来检验消息的完整性。

## 10.7.2 参考标准和协议

与 SSL VPN 特性相关的参考标准与协议如下:

RFC5246: The Transport Layer Security (TLS) Protocol Version 1.2

## 10.7.3 可获得性

### License 支持

在没有 License 的情况下, 系统默认的最大并发用户数是 3 个。用户可以根据自己的业务需求购买 License 增加并发 SSL VPN 用户的数量。

License 所允许的 SSL VPN 的最大并发用户数是在所有虚拟网关中共享的。用户需要根据自己的情况来配置单个虚拟网关所允许的最大并发数。所有虚拟网关的最大并发数的总数量不能超过 License 所允许的最大并发用户数。

当 License 过期或者被注销时, 所有虚拟网关的最大并发用户数将被重置为 3, 用户需要重新配置每个虚拟网关的最大并发用户数。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100、<br>USG2100/2200/5100 HSR V100R005C00<br>USG5500 V200R001C00 |

## 10.7.4 原理描述

### 10.7.4.1 SSL 协议简介

SSL 是一个安全协议，为基于 TCP 的应用层协议提供安全连接，如 SSL 可以为 HTTP 协议提供安全连接。SSL 协议广泛应用于电子商务、网上银行等领域，为网络上数据的传输提供安全性保证。

SSL 与 IPSec 安全协议一样，提供加密和身份验证。但是，SSL 协议只对通信双方传输的应用数据进行加密，而不是对从一个主机到另一主机的所有数据进行加密。

要使用 SSL 协议进行 VPN 通信，通信双方必须支持 SSL。目前常见的应用一般都支持 SSL，如 IE、Netscape 浏览器、Outlook、Eudora 邮件应用等。

### SSL 协议分层

SSL 协议可以分为两层：

- 底层为 SSL 记录协议（SSL record protocol）  
该协议负责对上层的数据进行分块、压缩、计算并添加 MAC、加密，最后把记录块传输给对方。
- 上层为 SSL 握手协议（SSL handshake protocol）、SSL 密码变化协议（SSL change cipher spec protocol）和 SSL 警告协议（SSL alert protocol）
  - SSL 握手协议：客户端和服务端通过握手协议建立一个会话。会话包含一组参数，主要有会话 ID、对方的证书、加密算法列表（包括密钥交换算法、数据加密算法和 MAC 算法）、压缩算法以及主密钥。SSL 会话可以被多个连接共享，以减少会话协商开销。
  - SSL 密码变化协议：客户端和服务端通过密码变化协议通知接收方，随后的报文都将使用新协商的加密算法列表和密钥进行保护和传输。
  - SSL 警告协议：用来允许一方对另一方报告告警信息。消息中包含告警的严重级别和描述。

### SSL 协议通信的握手步骤

SSL 协议通信的握手步骤如下：

1. SSL 客户端向 SSL 服务器发起连接，并要求服务器验证自身的身份。
2. 服务器通过发送自身的数字证书证明身份。
3. 服务器发出一个请求，对客户端的证书进行验证。



4. 验证通过后，协商用于加密的消息加密算法和用于完整性检查的哈希函数。通常由客户端提供它支持的所有算法列表，然后由服务器选择最强大的加密算法。
5. 客户端和服务器通过以下步骤生成会话密钥：
  - a. 客户端生成一个随机数，并使用服务器的公钥（从服务器证书中获取）对它加密，以送到服务器上。
  - b. 服务器用随机数据（客户端的密钥可用时则使用客户端密钥，否则以明文方式发送数据）响应。
  - c. 使用哈希函数从随机数据中生成密钥。

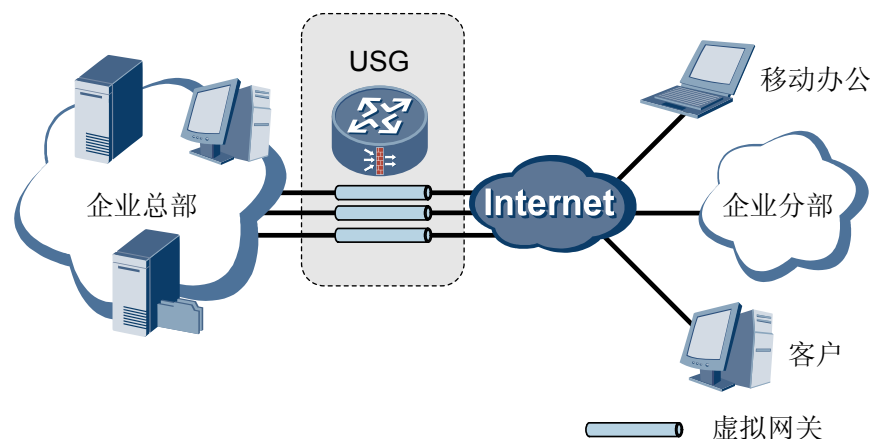
### 10.7.4.2 虚拟网关

在 USG 中，SSL VPN 建立的通道称为虚拟网关。

#### 工作原理

USG 可支持多个虚拟网关，各虚拟网关的配置和提供的服务互相独立。虚拟网关工作原理如图 10-62 所示。远程用户如移动办公用户、企业分部用户、客户等都可以通过与 USG 建立虚拟网关来访问指定的内网资源。

图 10-62 虚拟网关工作原理示意图



虚拟网关按照 IP 地址和域名的分配情况分为两种：

- 独占型  
独占型虚拟网关独占 IP 地址和域名。客户可以通过域名或者 IP 地址访问独占型虚拟网关。
- 共享型  
多个共享型虚拟网关共享同一个 IP 地址，具有相同的父域名，通过子域名来区分各虚拟网关。客户只能通过域名访问共享型虚拟网关。

#### 运营模式

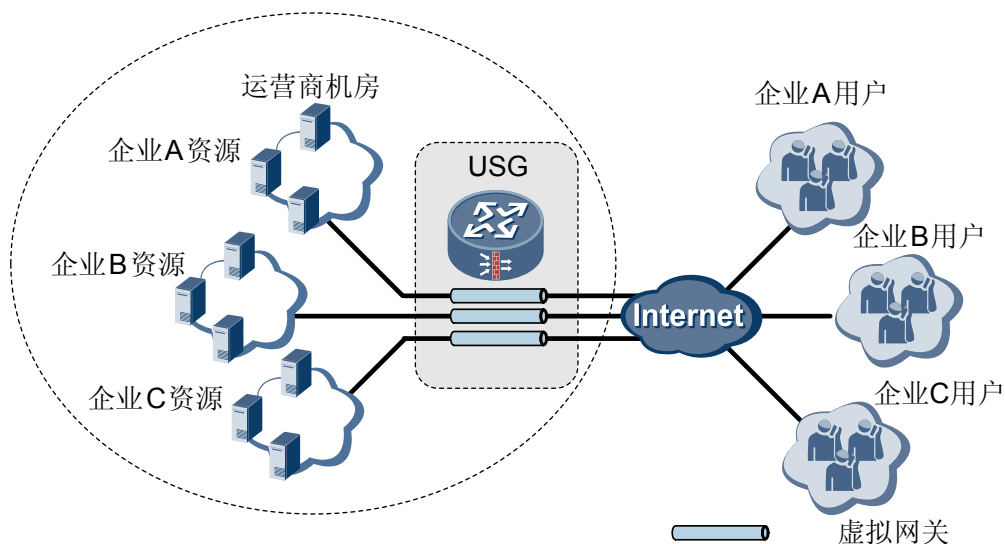
USG 可以部署在企业内部，由企业自行管理，如图 10-62 所示。

USG 也可以由运营商租用给企业，企业租用 USG 服务之后，外部用户如出差员工、企业分部可以通过专用的门户、专线，去访问企业总部的资源。USG 租用服务在实施的时候有两种情况：

1. 数据后台在运营商机房。

这种情况下，运营商直接提供 VPN 的服务。组网应用请参见图 10-63。

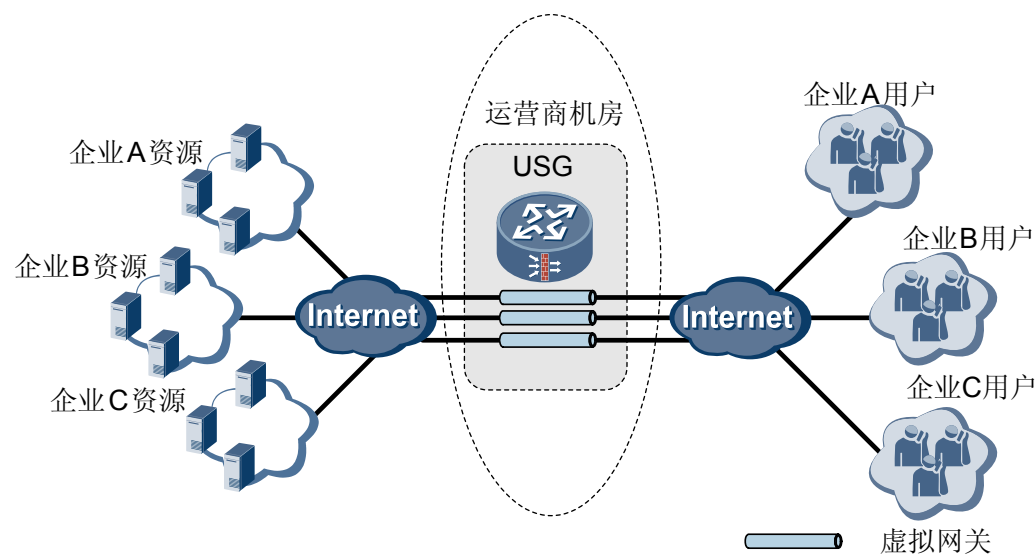
图 10-63 运营模式一组网图



2. 整个的数据在各自企业自己的机房。

这种情况下，运营商向客户公司的机房拉一条专线，或者建立 SSL VPN、IPSec VPN 线路，具体情况要看客户现有网络连接的情况。组网应用请参见图 10-64。

图 10-64 运营模式二组网图



### 10.7.4.3 SSL VPN 业务

SSL VPN 通过 SSL 协议和代理功能，实现远程用户安全地访问内网资源。SSL VPN 业务主要有以下四种业务：

- Web 代理
- 网络扩展
- 端口转发
- 文件共享

#### Web 代理

Web 代理支持用户通过 USG 访问内网的 Web 服务器资源，为用户提供基于 HTTP 的 Web 应用服务。当 USG 接收到用户的 HTTP 请求时，USG 作为 Web 代理从内网 Web 服务器上获取正确的 Web 资源返回给用户。

启用 Web 代理功能后，可以启用 Web 链接功能。启用 Web 链接功能后，用户只要在 USG 的虚拟网关客户端 Web 页面上安装控件后即可访问 Web 资源。



#### 注意

某些流行软件会与客户端软件发生冲突，导致 Web 代理功能不能正常运行。这些软件包括 QQ 旋风下载插件和微软 ISA（作为代理服务器），应避免同时使用。

---

#### 网络扩展

网络扩展是指用户在本地 PC 上安装 USG 的网络扩展客户端后，生成一个虚拟网卡，用户通过该虚拟网卡与企业内网进行 SSL 数据通信。

网络扩展客户端自动从虚拟网关上获得虚拟 IP 地址（IP 由企业内网管理员指定分配）并激活虚拟网卡进行 SSL 通信，使用户的 PC 如同工作在企业内网一样，快速、安全地访问企业内网的所有资源（在没有进行策略访问控制限制的情况下），并保证客户端和虚拟网关之间的通信采用 SSL 安全协议。



#### 说明

网络扩展通过用户目的 IP 型策略进行访问控制。

网络扩展功能将终端用户系统连接到企业网上（包含两层含义：1、对企业内网而言，就像终端用户在本地局域网真实存在一样；2、对终端用户而言，就像是接入了企业内网的局域网一样），并根据网络层信息进行接入控制。启用网络扩展功能一定程度上降低了安全级别，但在复杂拓扑结构下实现方便地管理网络。



#### 注意

某些流行软件会与客户端软件发生冲突，导致网络扩展功能不能正常运行。这些软件包括网络快车（VnetClient）、QQ 旋风下载插件、Sygate 和 D-link 网卡（开启 Adaptive Interrupt 功能），应避免同时使用。

---

用户使用网络扩展功能前，首先需要在虚拟网关客户端 Web 页面上下载并安装网络扩展客户端，网络扩展客户端安装成功后，机器上的网络扩展服务就会自动向 USG 申请一个虚拟 IP 地址，用来激活虚拟网卡并进行通信工作。

USG 采用 IP 地址池方式分配客户端虚拟 IP 地址。在 USG 上配置自定义的 IP 地址池，当遇到申请 IP 地址的请求时，USG 从自定义的 IP 地址池上申请 IP 地址，然后分配给客户端。

路由模式决定了客户端发送报文的方式。网络扩展功能支持三种路由模式：全路由模式（Full Tunnel）、分离模式（Split Tunnel）、手动模式（Manual Tunnel）。详细介绍请参见表 10-6。

表 10-6 网络扩展功能的路由模式

| 路由模式  | 说明                                                              |
|-------|-----------------------------------------------------------------|
| 全路由模式 | 用户只允许访问远端企业内网，不能访问 Internet 和本地局域网。                             |
| 分离模式  | 用户可以访问远端企业内网和本地局域网，不能访问 Internet。                               |
| 手动模式  | 用户可以访问远端企业内网特定网段的资源，对其它 Internet 和本地局域网的访问不受影响。网段冲突时优先访问远端企业内网。 |

## 端口转发

端口转发业务是提供基于 TCP 的应用程序的安全接入，是一种非 Web 的应用方式。端口转发在应用级对用户访问进行控制，控制是否提供各种应用的服务，如：Telnet、远程桌面、FTP、Email 等服务。

端口转发需要在客户端上运行一个 ActiveX 控件作为端口转发器，监听某个端口上的连接。当数据包进入这个端口时，它们通过虚拟网关被传送到 USG，USG 解开封装的数据包，将它们转发给目的应用服务器。



### 注意

某些流行软件会与客户端软件发生冲突，导致端口转发功能不能正常运行。这些软件包括 QQ 旋风下载插件和微软防火墙，应避免同时使用。

USG 的端口转发按照服务和端口对应关系分为四种情况：

- 单端口单服务  
一个服务对应一个端口。例如：Windows 远程桌面、Telnet、SSH、VNC、ERP、SQL Server、iNotes、OWA、BOSS。
- 多端口单服务  
一个服务对应多个端口。例如：POP3 Email（SMTP：25、POP3：110 等）
- 动态端口服务  
一个服务对应多个动态变化的端口。例如：FTP 被动模式、Oracle Manager。
- 单端口多服务

多个服务对应一个端口。例如：Notes（多个数据库服务器对应一个端口）

在 TCP/IP 协议中，端口是传输层的内容，端口与应用服务一一对应。协议里面端口号小于 1024 的端口都有确切的定义和规定，称为知名端口。端口转发配置中常用的端口如表 10-7 所示。

表 10-7 常用端口列表

| 应用服务         | 端口号  |
|--------------|------|
| FTP          | 21   |
| SSH          | 22   |
| Telnet       | 23   |
| SMTP         | 25   |
| HTTP         | 80   |
| POP3         | 110  |
| IMAP         | 143  |
| SQL Server   | 1433 |
| Windows 远程桌面 | 3389 |
| VNC          | 5900 |

## 文件共享

文件共享的主要功能是将不同系统的服务器(如支持 SMB(System management board)协议的 Windows 系统，支持 NFS(Network File System)协议的 Linux 系统)的共享资源以网页的形式提供给用户访问。

文件共享模块主要是将文件共享协议（SMB，NFS）转换成基于 SSL 的 HTTP（Hypertext Transfer Protocol）。

### 10.7.4.4 认证与授权

认证（Authentication）与授权（Authorization）提供对用户的认证和授权两种安全功能。

认证是指认证哪些用户可以访问网络。授权是指授权用户可以使用哪些服务。

认证与授权一般采用客户/服务器结构。客户端运行于被管理的资源侧，服务器上则集中存放用户信息。这种结构既具有良好的可扩展性，又便于用户信息的集中管理。

## 认证方式

USG 支持以下认证方式：

- VPND 本地认证

VPNDB 用于本地 VPN 数据库认证，管理员通过管理用户和组来维护 VPNDB。当用户接入时，根据本地配置的用户信息（包括用户名、密码及其他属性）进行认证。

本地认证的优点是速度快，可以降低运营成本；缺点是存储信息量受设备硬件条件限制。

- 远端认证
  - 支持通过 RADIUS（Remote Authentication Dial In User Service）协议进行远端认证，由网络接入服务器 NAS（Network Access Server）作 Client 端，与 RADIUS 服务器通信。
  - 支持通过 LDAP（Lightweight Directory Access Protocol）协议进行远端认证。

 说明

USG 作为 NAS 进行远端认证。

- 证书认证

支持对证书进行有效性认证。证书认证包括证书匿名认证和证书挑战认证。

 说明

USG 支持英文证书和 UTF8 格式编码的中文证书的认证。

证书认证通过 CA 证书来对虚拟网关用户身份进行验证，包括证书匿名和证书挑战：

- 证书匿名通过证书提取字段来验证虚拟网关用户身份。
- 证书挑战除了使用证书提取字段来验证虚拟网关用户身份外，还结合 VPNDB、RADIUS 或 LDAP 服务器的认证来辅助验证用户身份。

## 授权方式

USG 支持以下授权方式：

- VPNDB 本地授权

VPNDB 用于本地 VPN 数据库授权，管理员通过管理用户和组来维护 VPNDB。当用户接入时，根据本地配置的用户信息（包括用户名、密码及其他属性）进行授权。

分组易于对用户进行管理，可以按组对用户集体授权。
- 远端授权

支持通过 RADIUS 和 LDAP 进行远端授权。

### 10.7.4.5 策略

策略主要用来过滤用户可以访问的资源 and 进行流控制。管理员可以配置一系列的匹配规则，以识别需要过滤或放行的资源和报文。

策略是由 permit 或 deny 语句组成的一系列有顺序的规则，这些规则针对资源或报文的源地址、目的地址、端口号或其他信息来描述。

## 策略分类

虚拟网关的策略包括如下四种类型：

- 虚拟网关源 IP 型

通过限制客户端源 IP 地址定义对当前虚拟网关的访问控制。

- 用户源 IP 型

通过限制源 IP 地址定义用户对虚拟网关的访问控制。

- 用户目的 IP 型

通过限制目的 IP 和端口定义用户对虚拟网关内网资源的访问控制。

- 用户 URL 型

通过限制目的 URL 定义用户对虚拟网关内网资源的访问控制。

USG 还提供了针对组的策略，其本质是对组内用户进行集体策略控制。

默认策略行为是指默认情况下某一类型策略统一的行为。

## 策略的匹配规则

根据策略类型进行匹配，匹配顺序为：

1. 虚拟网关源 IP 型。
2. 用户源 IP 型。
3. 用户目的 IP 型/用户 URL 型。

针对每种类型的策略，匹配顺序为：

1. 按照“深度优先”规则，即掩码长的策略优先。如果掩码长度相同，固定端口的策略比任意端口的策略优先。
2. 与默认策略行为相反的策略。
3. 默认策略。

## 策略应用

虚拟网关源 IP 型策略用于限制指定用户 IP 网段的访问。

用户源 IP 型策略用于限制指定用户的访问。

用户目的 IP 型/用户 URL 型策略的具体应用如下：

- 目的 IP 型策略用于限制用户可以访问的端口转发或网络扩展资源。
- URL 型策略用于限制用户可以访问的文件共享或 Web 代理资源。

### 10.7.4.6 DNS

DNS（Domain name system）既是一个分布式数据库，是在由名称服务器主机构成的层次结构中实现的；同时，又是一个应用层协议，允许客户主机和名称服务器主机通信使用域名转换服务，提供域名与 IP 地址之间的转换。通过 DNS，用户可以直接使用便于记忆的、有意义的域名，而由网络中的 DNS 将域名解析为正确的 IP 地址。

# 11 用户与认证

---

## 关于本章

### 11.1 上网用户管理

上网用户是指经设备身份认证通过后访问网络资源的用户。通过上网用户管理，可以实现基于用户/用户组的网络行为控制。

### 11.2 接入用户管理

AAA 是 Authentication（认证）、Authorization（授权）、Accounting（计费）的简称，是一种安全管理框架，提供了授权部分实体访问特定资源并记录特定操作的一种安全机制。设备使用 AAA 对接入用户进行管理。

### 11.3 证书

证书建立了用户身份信息与用户公钥的关联，用于通信双方的身份认证。

### 11.4 802.1X

802.1X 提供一种更为安全的接入方式，通过对用户接入端口的认证控制，对接入用户进行管理。



## 11.1 上网用户管理

上网用户是指经设备身份认证通过后访问网络资源的用户。通过上网用户管理，可以实现基于用户/用户组的网络行为控制。

### 11.1.1 介绍

#### 定义

上网用户就是通过设备访问网络资源的用户，包括内网访问外网资源的用户（公司内部员工）和外网访问内网资源的用户（出差员工）。

上网用户管理是指管理员根据企业组织结构，为不同的上网用户定制差异化的认证方案，在他们访问网络资源前进行身份认证，为实现基于用户的网络行为控制和网络权限分配奠定了基础。

#### 说明

下文中提到的用户，如无特别指明，均指上网用户。

#### 目的

当前网络环境中，网络安全的威胁更多的来源于应用层，这也使得企业对于网络访问控制提出更高的要求。如何精确的识别出用户，保证用户的合法应用正常进行，阻断用户有安全隐患的应用等问题，已成为现阶段企业对网络安全关注的焦点。但 IP 不等于用户、端口不等于应用，传统防火墙基于 IP/端口的五元组访问控制策略已不能有效的应对现阶段网络环境的巨大变化。

上网用户管理通过对不同用户定制差异化的认证方案，在他们访问网络资源前进行身份认证，为网络行为控制和网络权限分配提供了基于用户/用户组的管理维度，实现了比状态防火墙更加精细的管理：

- 根据用户/用户组进行策略的可视化制定（对哪些用户的哪些应用行为进行策略控制），提高策略制定的易用性。
- 报表中包含用户信息，可以根据用户进行应用类型、威胁、数据传输行为的统计分析，解决以前只能根据 IP 地址进行的统计分析，以达到对用户行为的追踪审计。
- 可以解决同一用户 IP 地址变化带来的策略控制问题，即以用户的不变应对 IP 的万变。

### 11.1.2 参考标准和协议

与上网用户管理相关的参考标准与协议如下：

- RFC 1823: The LDAP Application Program Interface
- RFC 2865: Remote Authentication Dial In User Service (RADIUS)
- RFC 2866: RADIUS Accounting
- RFC 4510: Lightweight Directory Access Protocol (LDAP): Technical Specification Road Map
- RFC 4511: Lightweight Directory Access Protocol (LDAP): The Protocol

- RFC 4512: Lightweight Directory Access Protocol (LDAP): Directory Information Models
- RFC 4513: Lightweight Directory Access Protocol (LDAP): Authentication Methods and Security Mechanisms

### 11.1.3 可获得性

#### License 支持

本特性无需 License 支持。

#### 版本支持

| 产品                                                                      | 最低支持版本   |
|-------------------------------------------------------------------------|----------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | V300R001 |

#### 特性依赖

为提高策略的可视化，安全策略、限流策略、策略路由等功能均支持以用户/用户组作为匹配条件进行相关的控制。当策略以用户组作为匹配条件时，则该用户组下的所有用户将继承用户组的策略，拥有策略授予用户组的网络权限。

例如，在安全策略中将用户/用户组作为一个匹配条件，流量在匹配五元组的基础上，还要匹配用户身份。匹配上安全策略的用户将获取策略中指定的权限，例如 Web 过滤中允许用户访问哪些网站、邮件过滤中允许用户接收哪些邮件等。

要实现基于用户/用户组的其他功能（指基于用户/用户组的安全策略等），必须同时满足以下条件：

- 数据流源 IP 地址必须属于某个用户认证策略的 IP 地址范围内。
- 用户必须认证通过。

 说明

当报文匹配不到用户相关的策略时，设备不会将其丢弃，而是采用基于五元组的相关策略（例如安全策略、限流策略等），对报文进行控制。

### 11.1.4 应用场景

#### 11.1.4.1 免认证

#### 应用环境

针对以下类型用户，可通过免认证方式进行身份认证：

- 公司来访人员

这类用户，不是该公司员工，只是临时到访该公司，需要借用该公司网络上网，他们没有属于自己的账号，无法进行认证，但设备要对他们的网络权限进行控制。

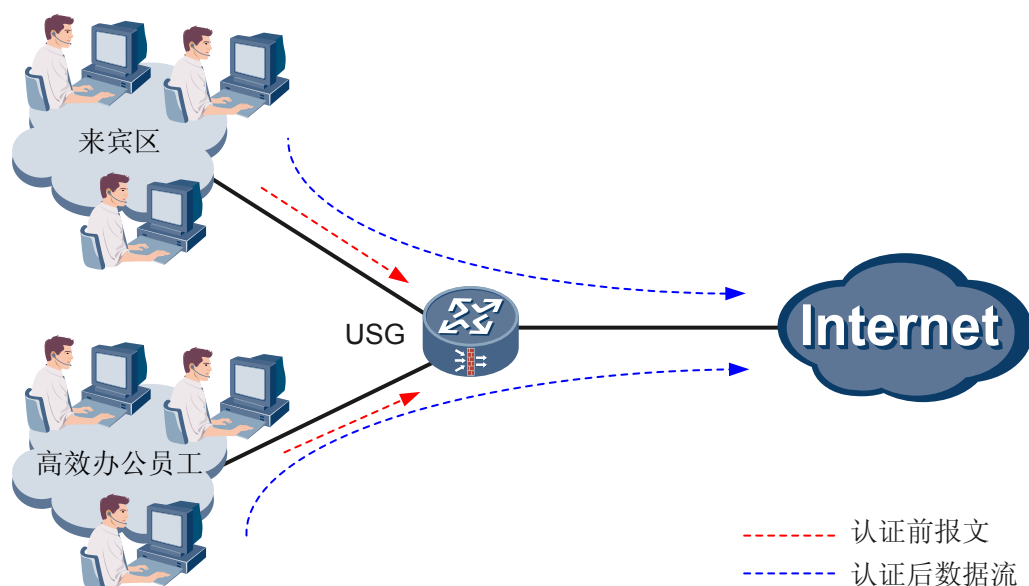
- 高效及高安全性要求用户

比如公司高层管理者，一方面他们需要高效率办公，希望省略认证过程，另一方面，他们可以访问公司机密数据，认证过程要求又更加严格。为此，管理员可将这类用户与指定的 IP/MAC 地址绑定，允许这类用户免认证，但只能通过指定的 IP 或者 MAC 地址进行登录。

## 典型应用

如图 11-1 所示，设备作为某公司的出口网关，部署在公司网络出口处。内网用户的来宾区和高效办公员工均通过免认证方式向设备进行身份认证。

图 11-1 内网用户进行免认证的典型组网应用



### 11.1.4.2 密码认证

#### 应用环境

密码认证包括本地密码认证和服务器密码认证：

- 在没有配置第三方认证服务器时，在设备上创建用户名和密码等信息，可对用户进行本地密码认证。
- 对于大型企业来说，用户规模较大，往往部署了专门的第三方认证服务器。设备可结合第三方认证服务器，对用户进行服务器密码认证，避免在设备上再管理用户的账号信息，实现用户的统一管理，减少配置及维护的工作量。

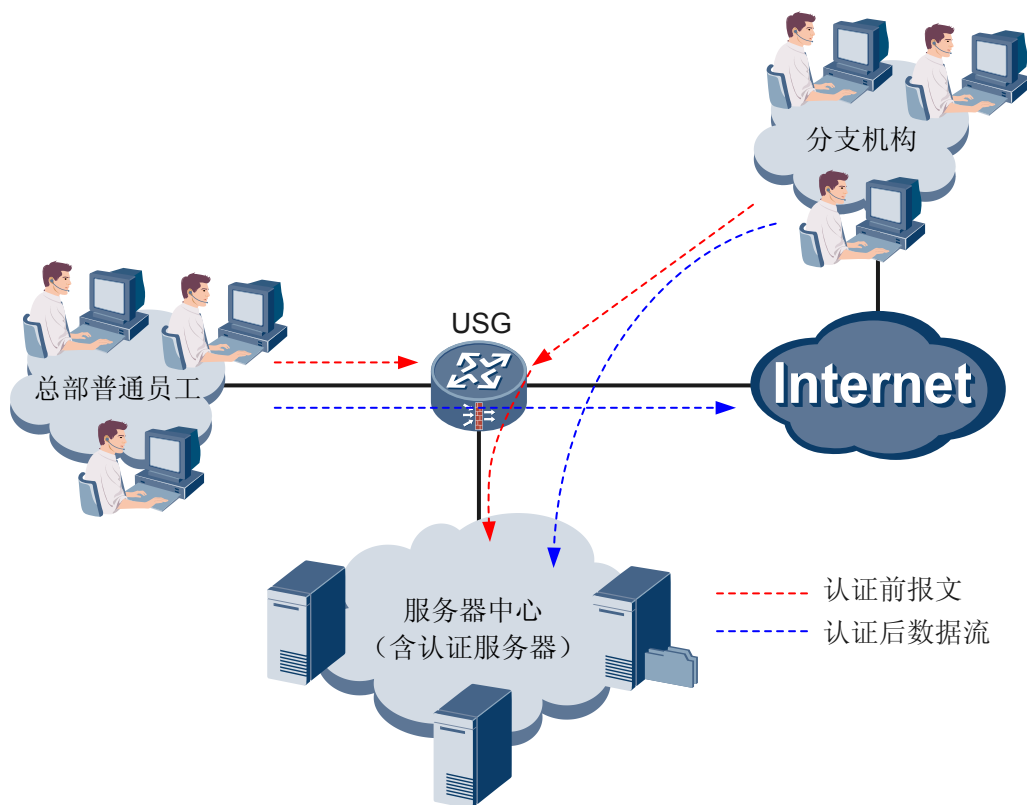
#### 典型应用

如图 11-2 所示，设备作为某公司的出口网关，部署在公司网络出口处。公司总部普通员工信息在设备上已有记录，分支机构信息都保存在认证服务器中。

- 当公司总部普通员工发起业务访问时，设备对这类员工进行本地密码认证。
- 当分支机构访问公司总部的服务器中心资源时，设备先在本地查找是否存在该用户账号及其密码信息。当用户信息或其密码信息在本地不存在时，设备将该用户的认证申请发送给第三方认证服务器，由服务器进行密码认证。

用户认证通过后，才允许按照相关策略（如安全策略等）指定的网络权限访问相关资源。

图 11-2 内网用户和分支结构员工进行密码认证的典型组网应用



### 11.1.4.3 单点登录

#### 应用环境

如果在当前网络中已经部署了其他的身份认证系统，则设备通过单点登录功能，就可以识别出经过这些身份认证系统认证通过的用户，避免用户上网时再次要求输入用户名/密码，降低对上网用户的影响。

目前，设备仅支持 AD 域和 TSM 单点登录。

#### 典型应用

如图 11-3 和图 11-4 所示，设备均作为某两公司的出口网关，部署在公司网络出口处。这两公司均在内网部署了 AD 服务器，对用户进行域账号认证，但差异点就在于两公司的 AD 服务器的部署位置有所不同。为避免用户反复输入用户名/密码进行认证，设备支持这两种场景下的单点登录功能：

图 11-3 内网用户进行单点登录的典型应用（用户和认证服务器在设备同一侧）

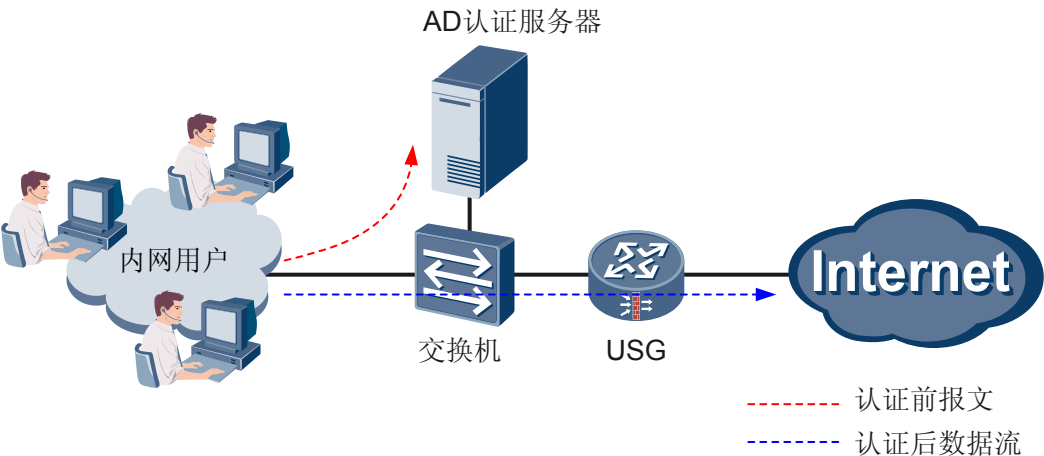
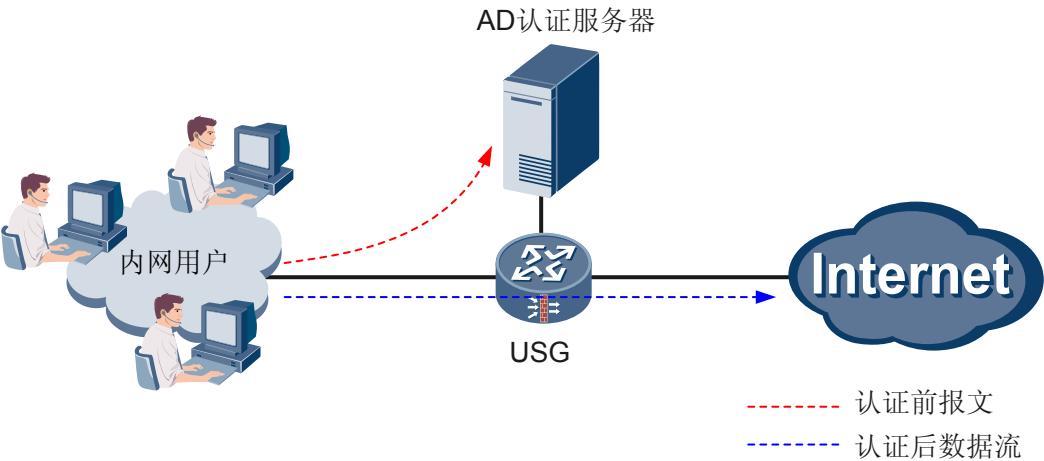


图 11-4 内网用户进行单点登录的典型应用（用户和认证服务器在设备两侧）



- 图 11-3 中，用户的认证报文不经过设备，AD 认证服务器将用户的认证结果通知设备。
- 图 11-4 中，用户的认证报文经过设备。为实现单点登录功能，要求 AD 认证服务器的 IP 地址必须在用户认证豁免 IP 地址范围内，并将用户的认证结果通知设备。

## 11.1.5 原理描述

### 11.1.5.1 组/用户

#### 树形组织结构

设备的用户（组）采用类似于企业的行政架构，以树型结构进行组织，如图 11-5 所示。

图 11-5 树形组织结构

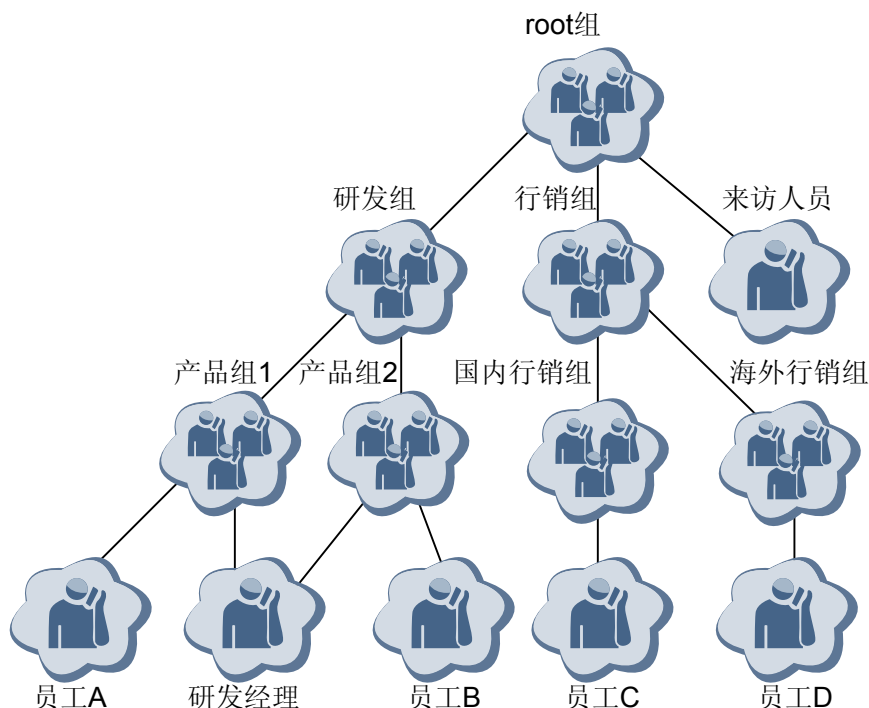


图 11-5 中：

- root 组是一级组，研发组和行销组是二级组，以此类推。
- 研发组和行销组是 root 组的子组。反之，root 组是研发组和行销组的父组。
- 产品组 1 和产品组 2 是研发组的子组，国内行销组和海外行销组是行销组的子组，产品组 1、产品组 2、国内行销组和海外行销组都是 root 组子组的子组。
- 员工 A ~ D、来访人员分别是产品组 1、产品组 2、国内行销组、海外行销组、root 组下的用户。反之，产品组 1、产品组 2、国内行销组、海外行销组、root 组分别是员工 A ~ D 和来访人员的父组（所属组）。
- 研发经理同时管理产品组 1 和产品组 2，其父组是产品组 1 和产品组 2。

设备对用户与用户组的对应关系规定如下：

- root 组是设备默认自带的根用户组，不能删除，也无需创建，且组名不能修改。
- 其余所有用户组都是 root 组的子组，或者子组的子组。
- 设备最多支持 15 层用户结构，包括 root 组 and 用户，即 root 组和用户之间最多允许存在 13 层用户组。
- 每个用户组可以包括多个用户和用户组，但每个用户组只能属于一个父组。一个用户最多属于三个父组。
- 用户组名允许重名，但所在组织结构路径必须确保唯一性。
- 用户不允许重名。
- 策略（如安全策略、限流策略等）可以引用用户（组），如果是引用用户组，则组下的用户继承其父组 and 所有祖先节点的策略。
- 设备支持 VPN 实例，每个 VPN 实例相当于上述的一个“系统”，拥有上述的“组织结构”。

## 用户类型

用户类型可根据用户来源和认证方式进行划分。

根据用户来源划分，可分为以下几类：

- 管理员手动创建  
由管理员在设备本地手工创建用户组和用户，并配置用户属性。例如，管理员可根据企业的组织架构创建用户组，然后再在各用户组下创建用户信息。
- 本地导入  
本地导入支持从 CSV 格式文件和数据库 dbm 文件批量导入用户。
  - 从 CSV 格式文件中批量导入用户  
CSV 格式文件导入是指：
    - 将用户信息按照指定格式的 CSV 表格文件预先编辑完成，再将 CSV 格式文件中的用户信息导入到设备内存中。
    - 将之前从设备上导出的 CSV 格式文件中的用户信息导入到设备内存中。
  - 从数据库 dbm 文件中批量导入用户  
数据库 dbm 文件导入是指将管理员 PC 上保存的用户和用户组的数据库文件导入到设备，当不保存配置的情况下重启设备后，系统按照新导入的数据库进行用户（组）组织结构恢复。  
但是，导入的数据库文件必须是之前从设备的内存数据库导出的数据库文件，否则设备无法识别文件格式。且导入的数据库文件中的内容为不可见字符，无法对其进行修改。
- 服务器导入  
网络中，使用第三方认证服务器的情况非常多，很多公司的网络都存在认证服务器，认证服务器上存放着所有用户和用户组信息。从认证服务器上批量导入用户是指通过执行服务器导入策略，将认证服务器上用户（组）信息导入到设备上。  
设备只支持从 AD、LDAP 和 TSM 服务器批量导入用户，导入类型支持如下几种：
  - 只导入用户  
当不使用认证服务器上的组织结构，在本地重新组建组织结构时，可采用这种方式。
  - 只导入用户组  
当认证服务器上用户量太大，而又只有部分用户具有上网权限时，如果导入所有用户到本地，会导致同步了很多无效用户的问题。此时，建议只导入用户组。
  - 用户及用户组一起导入  
当认证服务器用户量不大，或者用户大多为上网用户，且某些组下的用户比较特殊需要单独配置策略来管理时，可采用这种方式。
- 设备自动发现并创建  
设备通过配置认证策略的新用户选项，可以将新用户（通过认证但本地目前不存在的账号）自动添加到指定的本地用户组中。  
例如，图 11-5 中的“来访人员”，他们没有属于自己的账号，又要通过公司网络访问内网和外网资源。对于这类用户，公司可以划定某块单独的区域供其上网。设备收到这类用户的业务报文后，判断该报文的“源”是否来自“指定的单独区域”，若是，设备自动创建该类用户，并指定其 IP 地址为该类用户的用户名。  
有关认证策略的介绍，请参见 11.1.5.2 认证原理。

根据用户上网时的认证方式，可分为以下几类：

- 免认证
- 本地密码认证
- 服务器密码认证
- 单点登录

有关认证方式的详细介绍，请参见 [11.1.5.2 认证原理](#)。

## 用户属性

用户属性及其说明如表 11-1 所示。

表 11-1 用户属性

| 用户属性          | 说明                                                                                                                                       | 备注        |
|---------------|------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 登录名           | 用户的登录名，作为用户认证的标识，即用户使用登录名发起认证请求。<br>用户登录名不允许重名。如果是密码认证，用户需要在认证页面中输入该名称。<br>登录名不区分大小写，不管输入大写还是小写字母，设备均识别为小写字母。                            | 必选        |
| 显示名           | 用户的显示名，方便识别不同的用户。<br>为了方便记忆和管理，显示名作为区分不同用户的一个标识。但显示名不能作为用户认证的标识，即用户不能使用显示名发起认证请求。例如，某企业通过域帐号（用户登录名）对员工进行用户认证，为了方便识别不同用户，可将用户显示名设置为用户的姓名。 | 可选        |
| 描述            | 用户的描述信息，方便管理员对该用户进行说明。                                                                                                                   | 可选        |
| 所属组           | 用户所在的父组。<br>一个用户最多属于三个父组。                                                                                                                | 必选        |
| 本地密码          | 用户采用本地密码认证时的密码。                                                                                                                          | 本地密码认证时必选 |
| 账号过期时间        | 用户的账号过期时间。<br>当用户账号过期时间到期后，用户不能被认证通过。但是，对于已经在线用户，不强制其下线。                                                                                 | 可选        |
| 允许多人同时使用该账号登录 | 是否允许多人同时使用该用户的登录名登录，即允许该登录名同时在多台计算机上登录。<br>如果不允许，则同一时刻仅允许该登录名在一台计算机上登录。                                                                  | 可选        |



| 用户属性        | 说明                                                                                                                                                                                                                                                                                                                                                      | 备注 |
|-------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| IP/MAC 绑定方式 | <p>该功能用来限制该用户账号只能在特定的 IP/MAC 地址上登录。并且根据绑定方式的设定，绑定的效果也不同。</p> <ul style="list-style-type: none"><li>● 不绑定：用户与 IP/MAC 地址不绑定，只要符合认证策略的 IP 地址范围的计算机上都能使用该用户账号登录。</li><li>● 单向绑定：用户只能使用指定的 IP/MAC 地址进行认证，但同时允许其他用户也使用该 IP/MAC 地址进行认证。</li><li>● 双向绑定：用户只能使用指定的 IP/MAC 地址进行认证，并且指定的 IP/MAC 地址仅供该用户使用。当一个 IP/MAC 地址被双向绑定后，其他单向绑定此 IP/MAC 地址的用户将无法登录。</li></ul> | 可选 |

### 11.1.5.2 认证原理

管理员通过配置认证策略，指定某一个或几个 IP 地址范围内的用户发起的 HTTP 报文经过设备时，采用哪种认证方式进行身份认证。

根据认证的时间顺序，用户认证可分为：

- 事前认证
- 会话认证

根据认证方式，用户认证可分为：

- 免认证
- 本地密码认证
- 服务器密码认证
- 单点登录

有关认证相关概念及其原理，介绍如下。

### 认证策略

当某个用户上网前，设备根据报文的源 IP 地址查找适用于此用户的认证策略。根据认证策略中的认证方式和新用户选项的不同，后续处理流程也有所不同。

设备支持的认证方式有以下几种：

- 免认证：表示用户不需要进行基于 Web 的用户名密码认证。如果设备同时启用了单点登录，则优先通过单点登录功能对用户进行认证。在没有单点登录或者单点登录不成功的情况下，设备根据数据包的源 IP 地址、源 MAC 地址来识别用户。因此，免认证也可表述为“免认证/单点登录”。
- 密码认证：表示用户需要进行基于 Web 的用户名密码认证，包括本地密码认证和服务器认证。
  - 如果设备同时启用了单点登录，则优先通过单点登录功能对用户进行认证。
  - 在没有单点登录或者单点登录不成功的情况下，设备在本地查找用户：
    - 如果用户在本地存在且已配置本地密码，则进行本地密码认证。

- 如果用户在本地不存在，或者在本地存在但是没有配置本地密码，则向认证服务器认证。

因此，密码认证也可表述为“本地密码认证/服务器认证/单点登录”。

- 只允许单点登录：表示只允许用户通过单点登录方式进行认证。

新用户是指设备中不存在的用户。目前，新用户有以下来源：

- 如果认证方式为免认证，则新用户指通过免认证，但在本地中不存在的用户。这类新用户将以 IP 地址作为用户名。
- 如果认证方式为密码认证，则新用户指通过了服务器认证，但在本地中不存在的用户。
- 如果认证方式为单点登录，则新用户指通过单点登录认证，但在本地中不存在的用户。

对于新用户，设备根据数据包中的源 IP 地址匹配到对应的用户认证策略，再根据新用户认证选项决定处理方式。设备对新用户的认证选项如下：

- 不允许新用户登录  
不管第三方服务器是否认证通过，设备都不允许新用户登录。
- 添加到指定的用户组中  
新用户认证通过后允许自动添加到指定的本地用户组中。新用户添加到指定组后，就可以拥有该组的上网权限。  
但对于通过 LDAP、AD 服务器认证的新用户，设备可以根据配置选择是否自动将该用户在 LDAP 或 AD 认证服务器上的组织结构按照对应的服务器导入策略导入到本地。如果选择自动导入，此时指定的本地用户组不起作用。
- 仅作为临时用户，不添加到本地用户列表中  
新用户认证通过后只能作为临时用户，不添加到本地用户列表，但可以享有指定本地用户组的上网权限。  
免认证新用户只能作为临时用户。

设备对新用户的三种认证选项是互斥的，只能选其一。

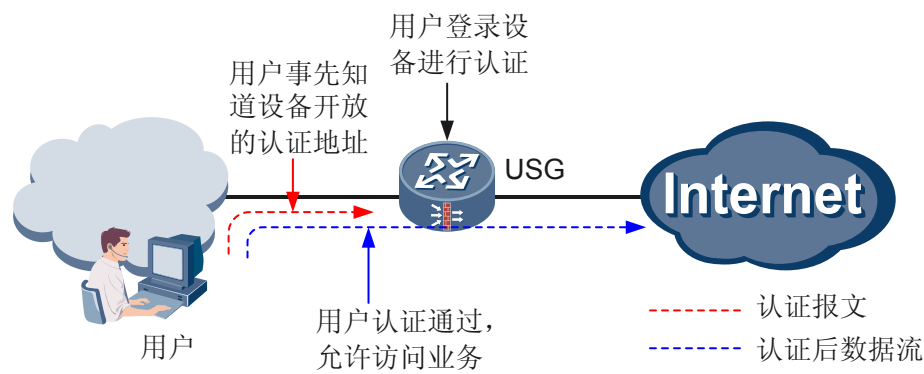
## 事前认证

事前认证是指用户在进行业务访问之前，主动先进行身份认证，认证成功后，再访问业务。

如图 11-6 所示，用户事先知道设备开放的认证地址，主动向设备发起认证请求。设备收到认证请求后，对用户进行身份认证。用户认证通过后，就可以访问业务。

目前，只要用户知道设备的接口 IP 地址，并与设备路由可达，向该接口 IP 地址发起的 HTTP 报文后，设备会弹出内置的 Web 认证页面，对用户进行认证。

图 11-6 事前认证流程



会话认证

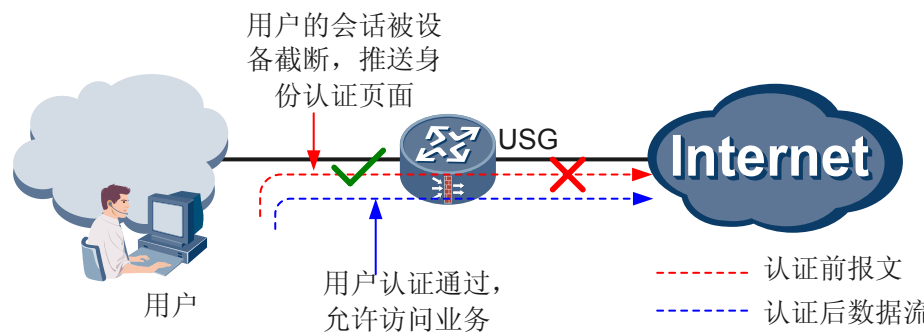
会话认证是指用户不主动进行身份认证，先进行业务访问，设备收到业务请求后，重定向到内置的认证页面，让用户进行认证。目前，设备只支持目的端口为 80 的 HTTP 方式业务访问推送认证页面，其余拒绝其访问。

**说明**

对于 DHCP 报文，设备做放行处理；对于 DNS 报文，在用户身份认证之前，做放行处理。

如图 11-7 所示，用户在进行业务访问之前，没有进行事前认证，当设备收到用户的第一条 HTTP 访问数据流时，设备自动将页面重定向到设备认证页面，触发用户进行身份认证。用户认证通过后，就可以访问业务。

图 11-7 会话认证流程

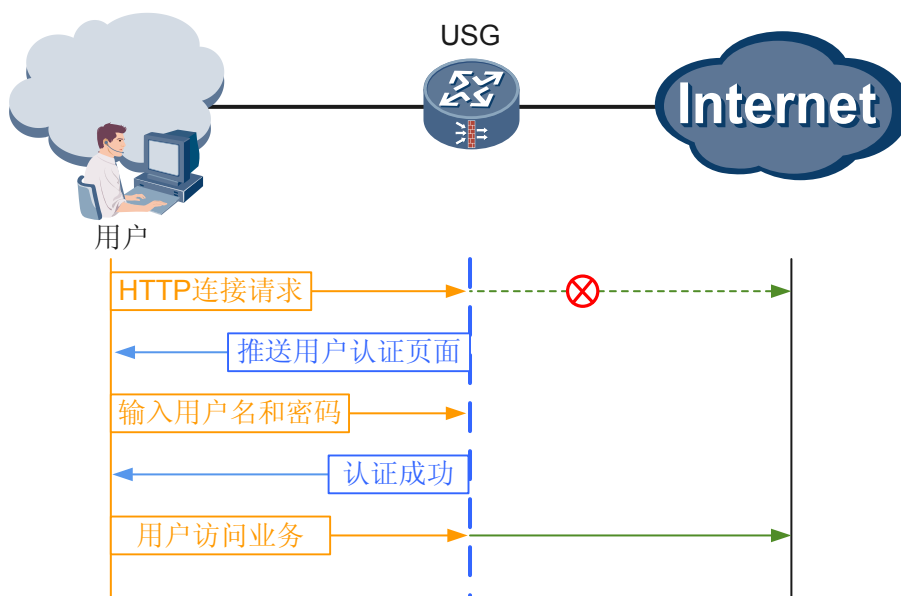


本地密码认证

本地密码认证是指用户在进行业务访问前，首先需要在设备上进行本地密码认证。本地密码认证要求用户信息在设备上存在且已配置本地密码。

如图 11-8 所示，以内网用户进行本地密码认证访问外网资源为例，介绍本地密码认证的流程。

图 11-8 本地密码认证流程



如图 11-8 所示：

1. 管理员在设备上配置对应的用户账号和密码等相关信息。
2. 用户发起的 HTTP 业务访问经过设备。
3. 设备收到 HTTP 连接请求后，向用户推送内置的用户认证页面，要求用户进行认证。
4. 用户获取到认证页面后，输入用户名和密码，向设备发起认证申请。
5. 设备收到用户认证请求后，校验用户名和密码。用户名和密码合法则根据配置跳转到用户最近使用的 Web 页面或跳转到自定义的 URL 页面。否则，则发送“认证失败”。
6. 用户认证成功后，能访问业务资源。同时设备记录用户 IP 地址与用户账号之间的关联关系，当该用户再次向外网发起业务访问时，设备根据业务数据流的源 IP 地址查询对应的用户账号。如果未查询到，则认为该用户还未认证通过，不能访问；如果查询到，则认为该用户已认证成功，根据该用户对应的策略（如安全策略等）对数据流进行处理。

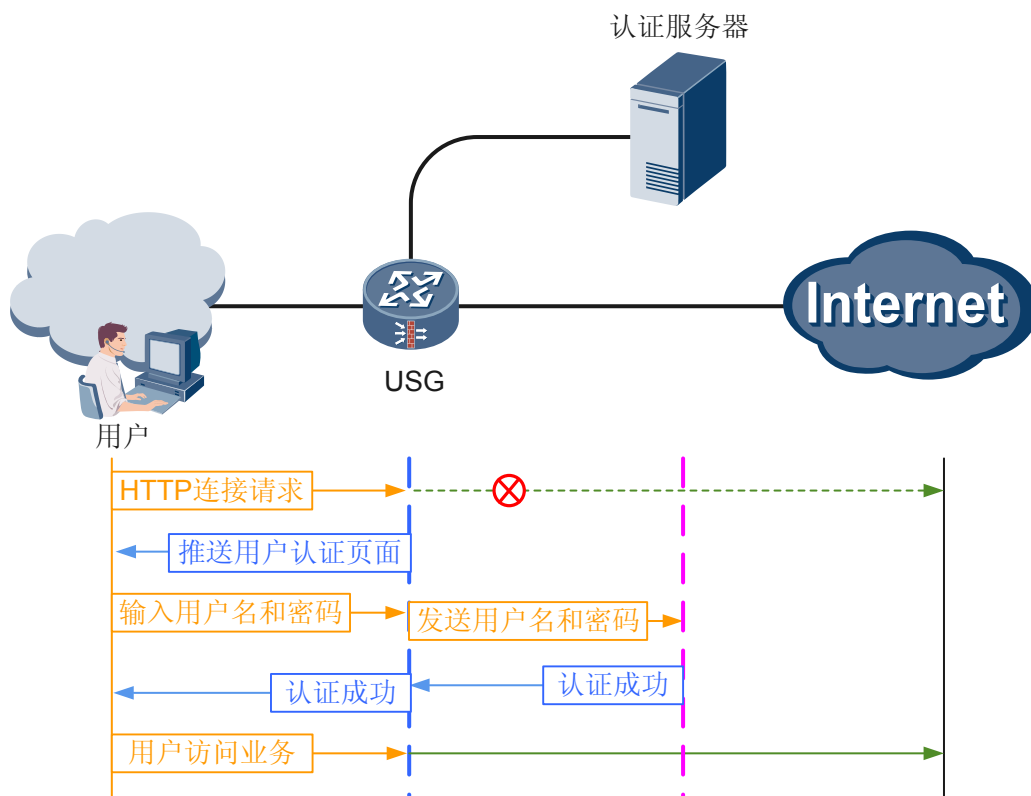
## 服务器密码认证

服务器密码认证是指设备把用户输入的用户名、密码发送到第三方认证服务器上以检查用户名、密码是否正确。如果组织内已经部署有认证服务器，则可以利用此功能，避免在设备上再管理用户的账号信息，实现用户的统一管理，减少配置及维护的工作量。

认证服务器支持 RADIUS、LDAP、AD 认证服务器。

如图 11-9 所示，以内网用户访问外网资源之前，通过设备代理，向第三方认证服务器进行身份认证为例，介绍服务器密码认证的流程。

图 11-9 服务器密码认证流程



如图 11-9 所示：

1. 管理员将用户所使用的 PC、设备、认证服务器连通，并在认证服务器上配置对应的用户账号和密码等相关信息。
2. 用户发起的 HTTP 业务访问经过设备。
3. 设备收到 HTTP 连接请求后，向用户推送内置的用户认证页面，要求用户进行认证。
4. 用户获取到认证页面后，输入用户名和密码，向设备发起认证申请。
5. 设备收到用户认证请求后，先在设备本地查找用户：
  - 如果用户在本地存在且已配置本地密码，则进行本地密码认证。
  - 如果用户在本地不存在，或者在本地存在但是没有配置本地密码，则向认证服务器认证。

本例中，设备是作为认证服务器的代理客户端，则将用户名和密码发送给认证服务器进行认证。认证服务器对用户名和密码进行校验，并将校验结果发送给设备。

6. 设备收到认证结果后，判断是否认证成功，如果是，则根据配置跳转到用户最近使用的 Web 页面或跳转到自定义的 URL 页面。否则，则发送“认证失败”。
7. 用户认证成功后，能访问业务资源。同时设备根据认证策略中的新用户选项中的配置，决定是否将该用户 IP 地址与用户账号之间的关联关系导入到本地指定组中或者将该用户及其在服务器上的组织结构自动导入到本地。当该用户再次向外网发起业务访问时，设备根据业务数据流的源 IP 地址查询对应的用户账号。如果未查询到，则认为该用户还未认证通过，不能访问；如果查询到，则认为该用户已认证成功，根据该用户对应的策略（如安全策略等）对数据流进行处理。

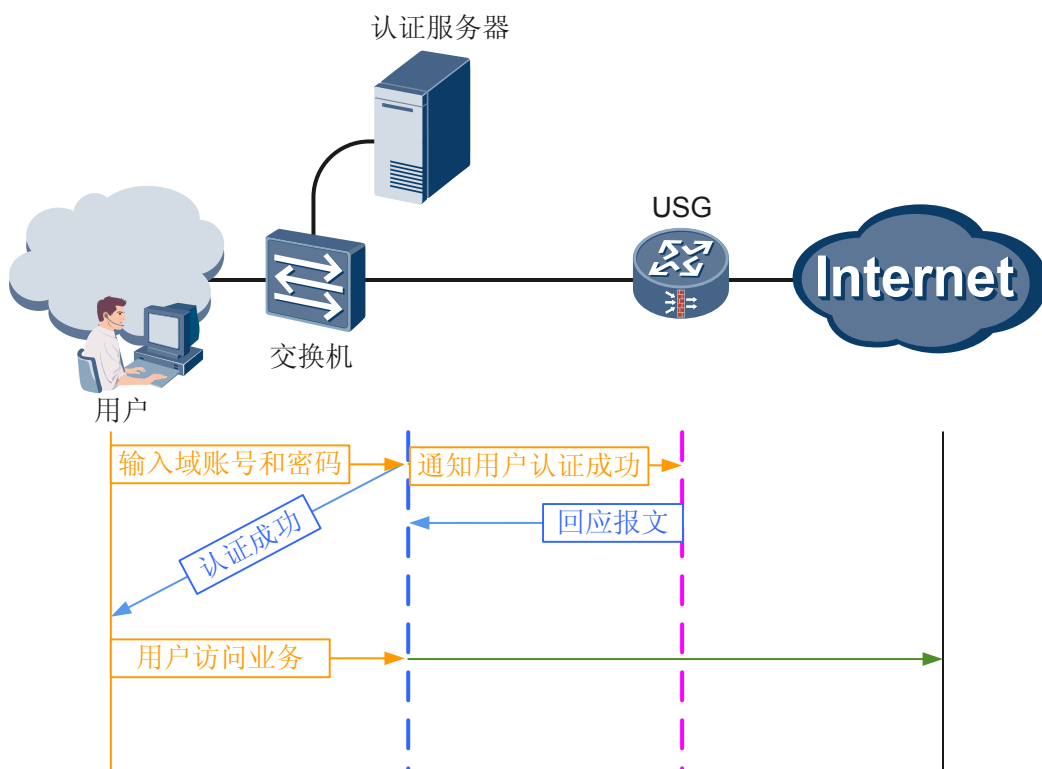
## 单点登录

本地密码和服务器密码认证均要求用户在上网前必须输入用户名和密码，对用户的身份进行认证后才能访问业务。如果在当前网络中已经部署了其他的身份认证系统，则设备通过单点登录功能，就可以识别出经过这些身份认证系统认证通过的用户，避免用户上网时再次要求输入用户名/密码，降低对上网用户的影响。

目前，设备仅支持 AD 域和 TSM 单点登录。

如图 11-10 所示，以内网用户通过 AD 域单点登录访问外网资源为例，介绍单点登录认证报文不经过设备时的处理流程。

图 11-10 单点登录流程（用户和 AD 服务器在设备同一侧）



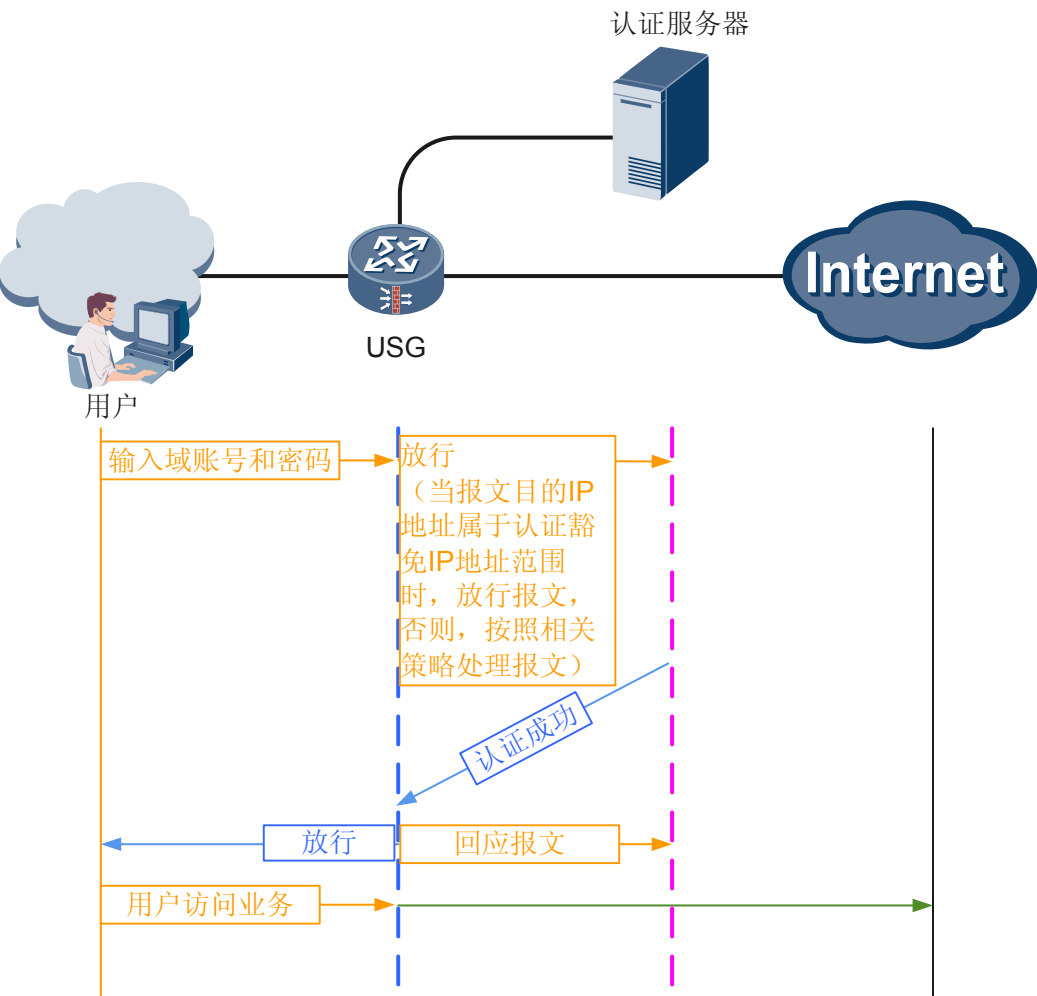
如图 11-10 所示：

1. 用户向认证服务器发起认证请求。  
对于 AD 单点登录来说，就是用户登录 AD 域；对于 TSM 单点登录来说，就是用户通过各种认证客户端向 TSM 控制器发起认证请求。
2. 认证服务器收到认证请求后，对域账号和密码进行校验，并将校验结果同时发送给设备和用户。
3. 设备收到认证服务器发送过来的用户认证成功通知后，向认证服务器发送回应报文，并记录用户已上线的信息，并根据认证策略中的新用户选项中的配置，决定是否将该用户 IP 地址与用户账号之间的关联关系导入到本地指定组中或者将该用户及其在服务器上的组织结构自动导入到本地。
4. 用户感知不到设备的存在，收到认证服务器发过来的认证成功通知后，开始访问业务。

5. 设备根据业务数据流的源 IP 地址查询对应的用户账号。如果未查询到，则认为该用户还未认证通过，不能访问；如果查询到，则认为该用户已认证成功，根据该用户对应的策略（如安全策略等）对数据流进行处理。

如图 11-11 所示，以内网用户通过 AD 域单点登录访问外网资源为例，介绍单点登录认证报文经过设备时的处理流程。

图 11-11 单点登录流程（用户和 AD 认证服务器在设备两侧）



如图 11-11 所示：

1. 用户输入域账号和密码，向认证服务器发起认证请求。
2. 设备收到认证报文后，判断报文的目的 IP 地址是否属于认证豁免 IP 地址范围，如果是，则放行报文，不对其处理用户管理流程。否则，该报文因匹配不了用户认证策略而被丢弃。

说明

当设备对目的 IP 地址属于认证豁免 IP 地址范围的报文不处理用户管理流程后，如果有其他策略（如域间安全策略）将该报文丢弃，则该用户通过单点登录失败。

3. 认证服务器收到设备发过来的用户认证请求后，对域账号和密码进行校验，并将校验结果同时发送给设备。



4. 设备收到认证服务器发送过来的用户认证成功通知后，向认证服务器发送回应报文，并放行“认证成功”报文。同时记录用户已上线的信息，并根据认证策略中的新用户选项中的配置，决定是否将该用户 IP 地址与用户账号之间的关联关系导入到本地指定组中或者将该用户及其在服务器上的组织结构自动导入到本地。
5. 用户感知不到设备的存在，收到认证服务器发过来的认证成功通知后，开始访问业务。
6. 设备根据业务数据流的源 IP 地址查询对应的用户账号。如果未查询到，则认为该用户还未认证通过，不能访问；如果查询到，则认为该用户已认证成功，根据该用户对应的策略（如安全策略等）对数据流进行处理。

## 免认证

免认证是指用户在访问业务前，不需要像事前认证那样主动先进行身份认证。设备在收到用户的业务报文时，根据报文源 IP 地址，查找对应的用户认证策略，当认证策略是免认证时，直接认为用户认证通过，让该用户上线。

免认证有两种方式：

- 双向绑定免认证：对应报文的源 IP 地址或 MAC 或 IP+MAC 对，与某个本地用户进行了双向绑定，则该用户可免认证上线，对应使用的用户身份为双向绑定的本地用户身份。
- 自动创建免认证：对应报文的源 IP 地址或 MAC 或 IP+MAC 对，均没有与任何用户进行双向绑定，则设备自动取其 IP 地址作为用户名让其上线，且该类用户只能作为临时用户，临时享有指定组的上网权限。

对于匹配到免认证认证策略的用户来说，设备先检测其是否满足双向绑定免认证条件，如果满足，则使用该用户使用对应的本地用户身份上线；如果不满足，则使用该用户使用自动创建免认证方式上线。

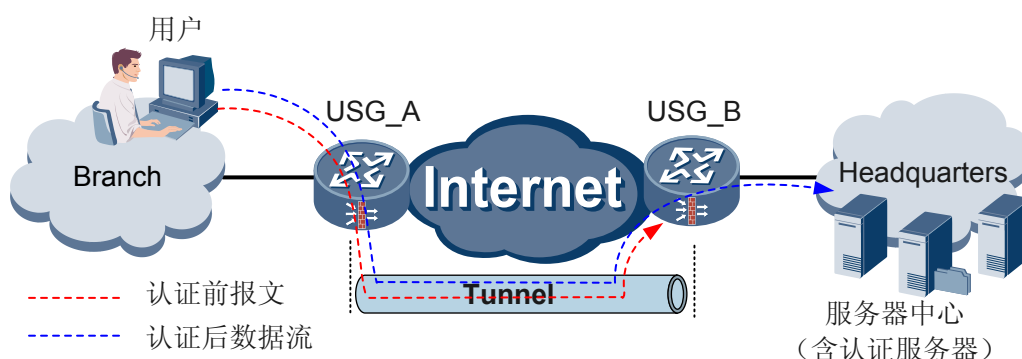
## VPN 用户认证

为了提高安全性，如企业出差员工、分支机构员工等用户一般通过 VPN 隧道访问公司资源。设备支持对这类 VPN 用户进行身份认证。用户身份认证与隧道建立过程中的认证没有关联，是两个相互独立的过程。设备只对成功接入隧道的用户才能进行身份认证。

设备对 VPN 用户的认证方式支持免认证、本地密码认证、服务器密码认证以及单点登录。不同认证方式下的 VPN 用户的身份认证流程同内网用户的认证流程类似，请参见本文之前的描述，下文仅介绍差异点。

如图 11-12 所示，用户通过 VPN 隧道访问内网资源。

图 11-12 VPN 用户认证





以会话认证方式为例，VPN 用户的认证流程如下：

1. 用户通过 VPN 隧道的接入认证，VPN 隧道建立完成。
2. 用户经 VPN 隧道向内网发起的 HTTP 业务访问经过 USG\_B。
3. USG\_B 收到 HTTP 连接请求后，根据配置的认证策略，对用户进行认证。

认证过程与非 VPN 用户的认证过程类似，需要注意的是，USG\_B 记录用户 IP 地址和用户账号之间的关联关系时，记录的是内层报文，即 VPN 用户的私网 IP 地址，也就是加封装前的报文或者解封装后的报文的源 IP 地址。

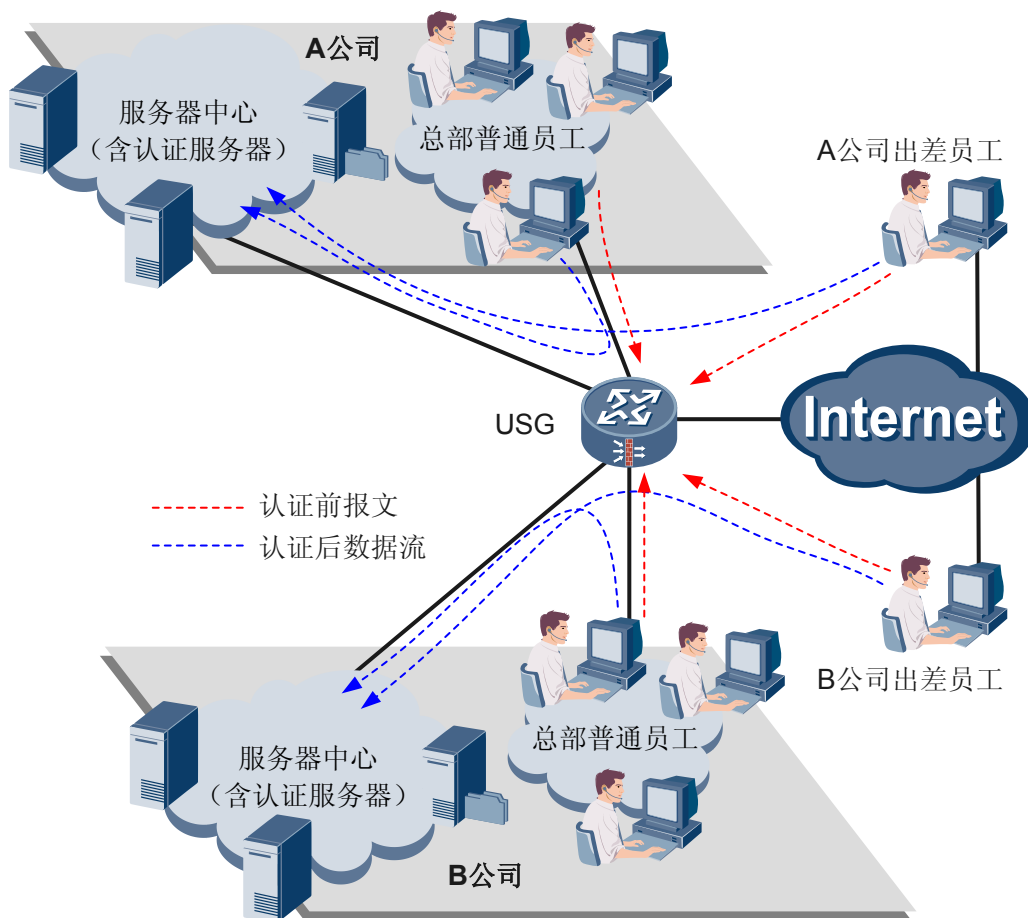
4. 用户认证成功后，能访问业务资源。

## VPN 实例下的用户认证

设备支持对不同 VPN 实例下的用户进行身份认证。不同 VPN 实例下的身份认证过程相互独立。

如图 11-13 所示，公司 A 购买了设备，将设备虚拟成连个多个 VPN 实例，自己使用其中一个 VPN 实例的同时，将另一个 VPN 实例出租给了公司 B。公司 A 和公司 B 有各自的用户群，两个用户群中的用户可能重名。这种场景下，设备需要支持分析各个公司的用户群，即使两个公司存在重名用户，也可以正确区分用户的身份，并对用户发出的业务访问进行合法的安全，带宽和 UTM 相关特性的管理。

图 11-13 不同 VPN 实例下的用户认证



## 11.2 接入用户管理

AAA 是 Authentication（认证）、Authorization（授权）、Accounting（计费）的简称，是一种安全管理框架，提供了授权部分实体访问特定资源并记录特定操作的一种安全机制。设备使用 AAA 对接入用户进行管理。

### 11.2.1 介绍

#### 定义

AAA（Authentication Authorization Accounting）是一种提供认证、授权和计费的技术。

- 认证（Authentication）：验证用户是否可以获得访问权，确定哪些用户可以访问网络。
- 授权（Authorization）：授权用户可以使用哪些服务。
- 计费（Accounting）：记录用户使用网络资源的情况。

#### 目的

提供对用户进行认证、授权和计费三种安全功能。

### 11.2.2 参考标准和协议

AAA 特性的参考资料清单如下：

- RFC 2809: Implementation of L2TP Compulsory Tunneling via RADIUS
- RFC 2865: Remote Authentication Dial In User Service (RADIUS)
- RFC 2866: RADIUS Accounting
- RFC 2867: RADIUS Accounting Modifications for Tunnel Protocol Support
- RFC 2868: RADIUS Attributes for Tunnel Protocol Support
- RFC 2869: RADIUS Extensions
- RFC 0927: TACACS user identification Telnet option
- RFC 1492: An Access Control Protocol, Sometimes Called TACACS
- RFC 5080: Common Remote Authentication Dial In User Service (RADIUS) Implementation Issues and Suggested Fixes
- RFC 5997: Use of Status-Server Packets in the Remote Authentication Dial In User Service (RADIUS) Protocol

### 11.2.3 可获得性

#### License 支持

AAA 功能不需要 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                       |
|-------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100 BSR/<br>HSR V100R005C00<br><br>USG5500 V200R001C00 |

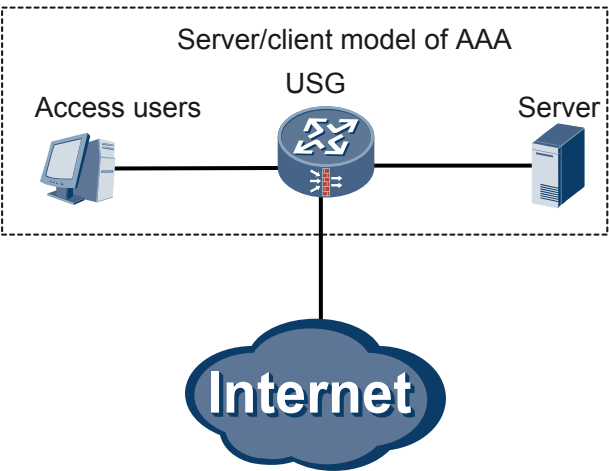
11.2.4 原理描述

11.2.4.1 AAA 协议

AAA 的基本构架

AAA 通常采用“客户端—服务器”结构。这种结构既具有良好的可扩展性，又便于集中管理用户信息。如图 11-14 所示。

图 11-14 AAA 的基本构架示意图



认证

AAA 支持以下认证方式：

- 不认证：对用户非常信任，不对其进行合法检查，一般情况下不采用这种方式。
- 本地认证：将用户信息（包括本地用户的用户名、密码和各种属性）配置在网络接入服务器上。本地认证的优点是速度快，可以为运营降低成本；缺点是存储信息量受设备硬件条件限制。
- 远端认证：将用户信息（包括本地用户的用户名、密码和各种属性）配置在认证服务器上。AAA 支持通过 RADIUS（Remote Authentication Dial In User Service）协议或 HWTACACS（HuaWei Terminal Access Controller Access Control System）协议

进行远端认证。网络接入服务器 NAS（Network Access Server）作为客户端，与 RADIUS 服务器或 HWTACACS 服务器通信。

## 授权

AAA 支持以下授权方式：

- 不授权：不对用户进行授权处理。
- 本地授权：根据网络接入服务器为本地用户账号配置的相关属性进行授权。
- HWTACACS 授权：由 HWTACACS 服务器对用户进行授权。
- if-authenticated 授权：如果用户通过了认证，而且使用的认证模式不是不认证，则用户授权通过。
- RADIUS 认证成功后授权：RADIUS 协议的认证和授权是绑定在一起的，不能单独使用 RADIUS 进行授权。

## 计费

AAA 支持以下计费方式：

- 不计费：不对用户计费。
- 远端计费：支持通过 RADIUS 服务器或 HWTACACS 服务器进行远端计费。

### 11.2.4.2 RADIUS 协议

AAA 可以用多种协议来实现，最常用的是 RADIUS 协议。RADIUS 最初用来管理使用串口和调制解调器的大量分散用户，后来广泛应用于网络接入服务器 NAS（Network Access Server）系统。

当用户想要通过某个网络（如电话网络）与 NAS 建立连接从而取得访问其他网络的权利或取得使用某些网络资源权利时，NAS 起到了认证用户或对应连接的作用。

NAS 负责把用户的认证和计费信息传递给 RADIUS 服务器。RADIUS 协议规定了 NAS 与 RADIUS 服务器之间如何传递用户信息和计费信息以及认证和计费结果，RADIUS 服务器负责接收用户的连接请求，完成认证，并把认证结果和用户所需的配置信息返回给 NAS。

NAS 和 RADIUS 之间的验证信息的传递是通过密钥的参与来完成的，以避免用户的密码在不安全的网络上传输时被窃取。

## 协议实现

RADIUS 使用 UDP（User Datagram Protocol）作为传输协议，具有良好的实时性；同时也支持重传机制和备用服务器机制，从而具有较好的可靠性。

RADIUS 的实现比较简单，适用于大用户量时服务器端的多线程结构。

网络接入服务器作为 RADIUS 协议的客户端，实现以下功能：

- 标准 RADIUS 协议及扩充属性，包括 RFC2865、RFC2866。
- 华为扩展的私有属性。
- 计费结束报文的本地缓存重传功能：计费结束报文在重传发送失败次数超过配置次数后，将计费结束报文保存到计费结束报文缓存队列；系统定时器周期扫描该队列，如果存在计费结束缓存报文，则取出报文内容，向指定的服务器发送并启动定

时器等待，如果发送失败或在超时时间内没有收到服务器回应，则重新入缓存队列。

认证和计费

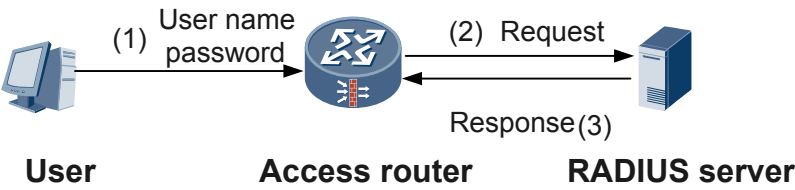
RADIUS 服务器通过建立一个唯一的用户数据库，存储用户名、密码来对用户进行验证。

RADIUS 客户端与服务器间的消息流程如图 11-15 所示。

- 用户登录 USG 或接入服务器等网络设备时，会将用户名和密码发送给该网络接入服务器；
- 该网络设备中的 RADIUS 客户端（网络接入服务器）接收用户名和密码，并向 RADIUS 服务器发送认证请求；
- RADIUS 服务器接收到合法的请求后，完成认证，并把所需的用户授权信息返回给客户端；对于非法的请求，RADIUS 服务器返回认证失败的信息给客户端。

客户端和 RADIUS 服务器之间发送的用户密码信息经过加密以后才在网络上传递，以避免用户密码在不安全的网络上被窃取。

图 11-15 RADIUS 客户端与服务器间的消息流程

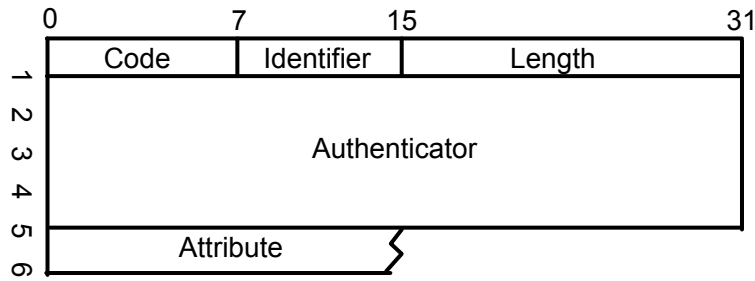


计费的消息流程和认证/授权的消息流程类似。

RADIUS 消息结构

RADIUS 的消息结构如图 11-16 所示。

图 11-16 RADIUS 消息结构图



各字段的含义简单说明如下：

- Code  
消息类型，如接入请求、接入允许等。

- Identifier  
一般是顺序递增的数字，请求报文和响应报文中该字段必须匹配。
- Length  
所有域的总长度。
- Authenticator  
验证字，用于验证 RADIUS 的合法性。
- Attribute  
消息的内容主体，主要是用户相关的各种属性。

### 11.2.4.3 HWTACACS 协议

HWTACACS 是在 TACACS (RFC1492) 基础上进行了功能增强的一种安全协议。该协议与 RADIUS 协议类似，主要是通过“客户端—服务器”模式与 HWTACACS 服务器通信来实现多种用户的 AAA 功能，可用于 PPP、VPDN (Virtual Private Dial Network) 接入用户和 Login 用户的认证、授权和计费。

### HWTACACS 协议原理

- 按命令行授权
- 用户通过 Telnet 或者 SSH 登录到 USG 上后，如果需对该用户输入的命令行进行认证，可以将该级别用户的命令行授权方法设置为 HWTACACS，该用户输入的每一条命令都要通过 HWTACACS 服务器授权。如果授权通过，命令就可以被执行。否则，HWTACACS 服务器输出信息，通知用户该命令的授权失败，命令不能执行。
- 命令行授权可以使用本地授权的方法作为备选方法，这样，如果因为服务器的问题（服务器 Down、不可达或回应超时）导致命令行授权失败时，可以将命令行授权转入本地授权处理。
- 如果在用户配置的超时时间内，USG 没有接收到 HWTACACS 服务器的授权结果，则授权超时，该命令不能被执行。
- 用户还可以配置服务器无响应或本地未配置用户时命令授权失败的策略，可以选择让用户继续在线，也可以选择授权失败次数超过阈值后下线。

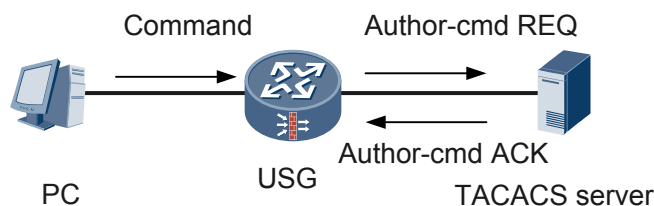
#### 说明

按命令授权失败的策略仅仅适用于因 HWTACACS 服务器不可用或本地未配置用户而导致的按命令行授权失败情况。以下两种情况不能触发命令行授权失败时的策略：

- 服务器正常时，所执行的命令行未能通过在 HWTACACS 服务器端的授权。
- 服务器不可用后，按命令行授权转入本地授权后，因执行的命令级别高于本地配置的级别而授权失败。

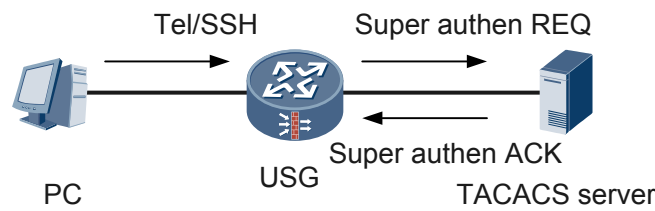
HWTACACS 按命令行授权的执行流程如图 11-17 所示。

图 11-17 HWTACACS 按命令行授权的执行流程



- 对用户级别提升进行认证
- 用户通过 Telnet 或者 SSH 登录到 USG 后，可以通过在用户模式下使用 **super** 命令来提升自己的级别。这时，USG 对用户的密码进行验证。
- HWTACACS 可以对用户级别的提升进行认证，其执行流程如图 11-18 所示。USG 将用户的密码发送到 HWTACACS 服务器上进行认证，如果认证通过，用户的权限就可以得到提升，否则，用户的权限不能提升。特权等级更改的结果只影响本次登录。

图 11-18 HWTACACS 的用户级别提升的执行流程



如果在用户配置的超时时间内，USG 没有接收到 HWTACACS 服务器用户级别提升的认证结果，则认证超时，用户不能提升权限。

**说明**

当 USG 对用户级别提升进行验证的时候，各级别的密码可以不同；在用 HWTACACS 对用户级别提升进行验证时，各级别密码是相同的。

HWTACACS 协议和 RADIUS 协议的比较

与 RADIUS 相比，HWTACACS 具有更加可靠的传输和加密特性，更加适合于安全控制。HWTACACS 协议与 RADIUS 协议的主要区别如表 11-2 所示。

表 11-2 HWTACACS 协议与 RADIUS 协议的比较

| HWTACACS                       | RADIUS            |
|--------------------------------|-------------------|
| 使用 TCP，网络传输更可靠                 | 使用 UDP            |
| 除了标准的 HWTACACS 报文头，对报文主体全部进行加密 | 只是对认证报文中的密码字段进行加密 |
| 认证与授权分离                        | 认证与授权一起处理         |
| 适于进行安全控制                       | 适于进行计费            |
| 支持对 USG 上的配置命令进行授权使用           | 不支持               |

11.2.4.4 基于域的用户管理

在目前 AAA 的实现中，所有用户都属于某个域。用户属于哪个域是由用户名中带的“@”后的字符串来决定的，比如“user@hua”，就属于“hua”域；如果用户名中没有带“@”，就属于系统缺省的 default 域。

## USG 对接入用户的管理

USG 通过域来进行用户管理，域下可以进行缺省授权配置、RADIUS/HWTACACS 模板配置、认证和计费方案的配置等。

所有对于接入用户的认证、授权、计费都是在域视图下应用认证方案、授权方案、计费方案来实现的，为此必须先在 AAA 视图下分别配置相应的认证方案、授权方案、计费方案。

AAA 有缺省的认证方案、授权方案、计费方案，分别采用本地认证、本地授权、不计费。如果新创建一个域，没有在域下应用认证方案、授权方案、计费方案，那么 AAA 对该域将采用缺省的认证方案、授权方案和计费方案。此外，如果要对用户采用 RADIUS/HWTACACS 方案，必须预先在系统视图下配置 RADIUS/HWTACACS 服务器模板，然后在用户所属域的视图下应用该服务器模板。

当域和域下的用户同时配置了某一属性时，基于用户的配置优先级高于域的配置优先级。

USG 最多可配置 1024 个本地用户。

## AAA 对 PPP 用户的地址分配原则

采用 PPP 方式接入的用户，可以利用 PPP 地址协商功能从网络接入服务器获得本端接口的 IP 地址。在网络接入服务器上，AAA 对 PPP 用户的地址分配原则如下：

- 对于不认证的用户：
  - 如果接口下配置为对端分配 IP 地址，直接把该地址分配给对端。
  - 如果接口下配置从地址池中为对端分配地址，则从指定的全局地址池中为对端分配地址。
- 对于认证通过的 default 域用户（包括两种：不加@的，如“aaa”；用户名后加@default，如“aaa@default”）：
  - 如果服务器下发了 IP 地址，则直接把该地址分配给对端。
  - 如果服务器下发了地址池号，则用该地址池号通过域地址池或者全局地址池给对端分配地址。
  - 在服务器既没有下发 IP 地址，也没有下发地址池号的情况下，如果接口下配置了给对端的 IP 地址，直接把该地址分配给对端。如果接口下配置了从地址池中为对端分配地址，则从指定的全局地址池中为对端分配地址。
- 对于认证通过的普通域用户：
  - 如果服务器下发了 IP 地址，直接把该地址分配给对端。
  - 如果服务器下发了地址池号，则用该地址池号通过域地址池给对端分配地址。
  - 在服务器既没有下发 IP 地址，也没有下发地址池号的情况下，从域下的第一个地址池开始遍历分配。
- 以上从地址池中分配地址的情况，全局地址池和域地址池都只遍历一遍。当指定的全局地址池或域地址池分配完后，直接返回无效地址 0，不去全局地址池遍历查找可用地址。
- 在配置地址池的时候，A 类地址中的 XXX.255.255.255 和 XXX.0.0.0，B 类地址中的 XXX.XXX.255.255 和 XXX.XXX.0.0，C 类地址中的 XXX.XXX.XXX.255 和 XXX.XXX.XXX.0 都不能够作为合法的地址池起始、结束地址。如果地址池里面含有这些地址，这些地址也不会被分配。



说明

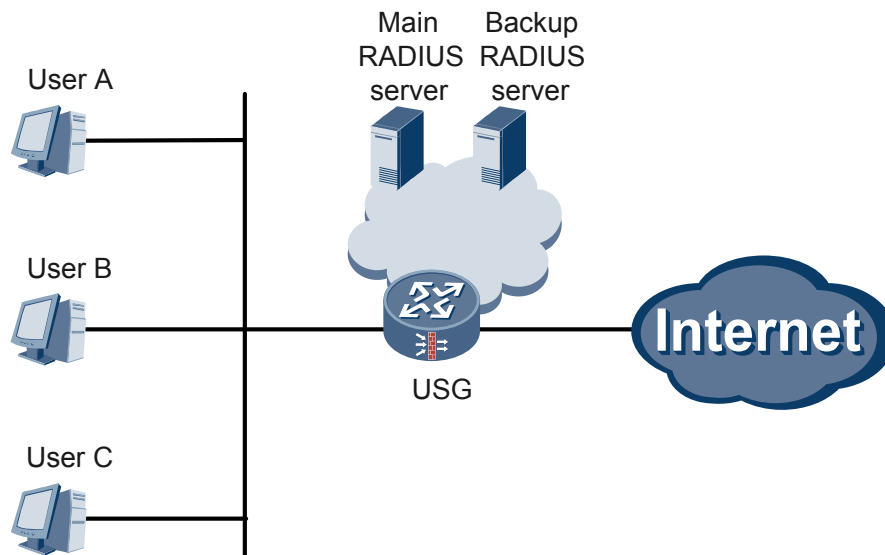
IP 地址协商的配置需要在客户端与服务器端分别进行。



## 11.2.5 应用

### 使用 RADIUS 对接入用户进行管理

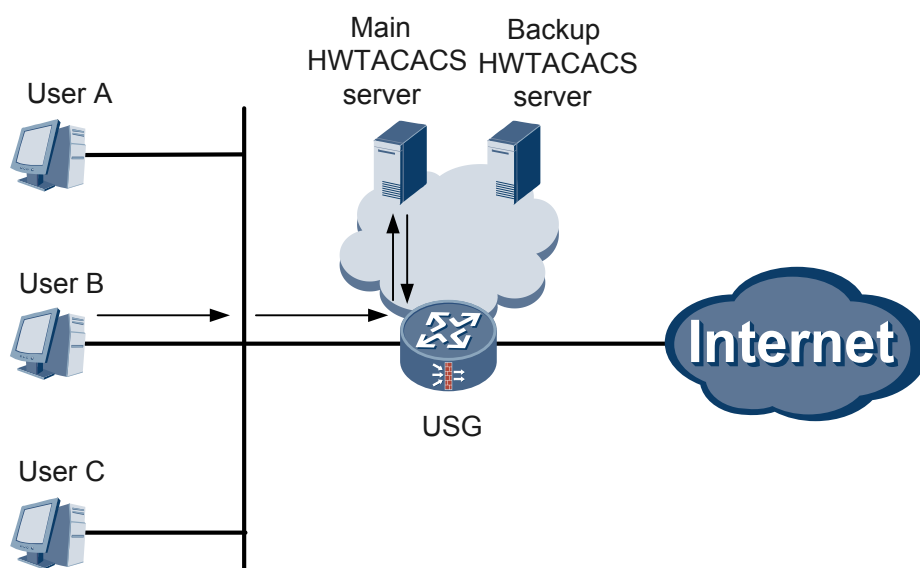
图 11-19 使用 RADIUS 对接入用户进行管理



使用 RADIUS 对接入用户进行管理，如图 11-19 所示。用户 A、B、C 都属于某个域下的用户，通过 USG 接入 Internet。USG 根据用户所属的域，选择具体的 RADIUS 服务器对其进行认证，如果认证通过，用户就可以访问 Internet。

### 使用 HWTACACS 对接入用户进行管理

图 11-20 使用 HWTACACS 对接入用户进行管理



使用 HWTACACS 对接入用户进行管理，如图 11-20 所示。用户 A、C 属于某个域下的用户，通过 USG 接入 Internet。USG 根据用户所属的域，选择具体的 TACACS 服务器对其进行认证，如果认证通过，用户 A、C 就可以访问 Internet。

用户 B 是管理型用户，需要登录 USG 对其进行配置。使用 HWTACACS 对其进行管理，以获得较高的安全性。可以使用 HWTACACS 对用户 B 在 USG 上配置的每一条命令进行授权，同时，如果用户 B 需要提升自己的用户权限，HWTACACS 也可以对其进行认证。

## 11.3 证书

证书建立了用户身份信息与用户公钥的关联，用于通信双方的身份认证。

### 11.3.1 介绍

#### 定义

数字证书简称证书，用来证明一台设备的身份，解决了通信双方之间的信任问题，同时可以保证信息在传输过程中的安全性、完整性和不可否认性。

证书文件通常由第三方通常是数字证书认证中心颁发，包含设备信息、公开密钥以及颁发者签名。

#### 目的

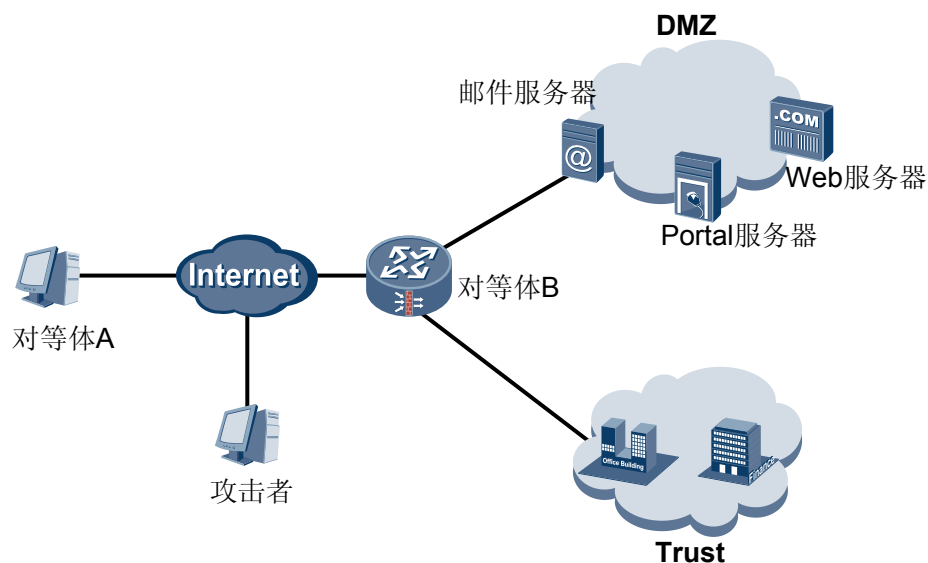
随着电子商务、网上银行及网上证券交易的飞速发展，Internet 的安全性越来越重要。部分不道德的个人，会截取应用的明文数据，进行中间人攻击。

以图 11-21 为例，对等体 A 要与对等体 B 创建 IPSec VPN，但是攻击者中途拦截了对等体 A 发送给对等体 B 的信息，冒充对等体 B 与之建立连接。考虑到这种安全问题，对等体 A 和对等体 B 必须在彼此建立通信之前检验双方身份。

数字证书为 VPN 实施提供了设备验证方式。

对等体事先向证书颁发机构申请证书，在建立 VPN 连接之前，会共享它们的证书，并验证对方证书是否合法，只有通过了合法性认证，才会与对方建立连接，从而有效防止中间人攻击。

图 11-21 中间人攻击例子



## 受益

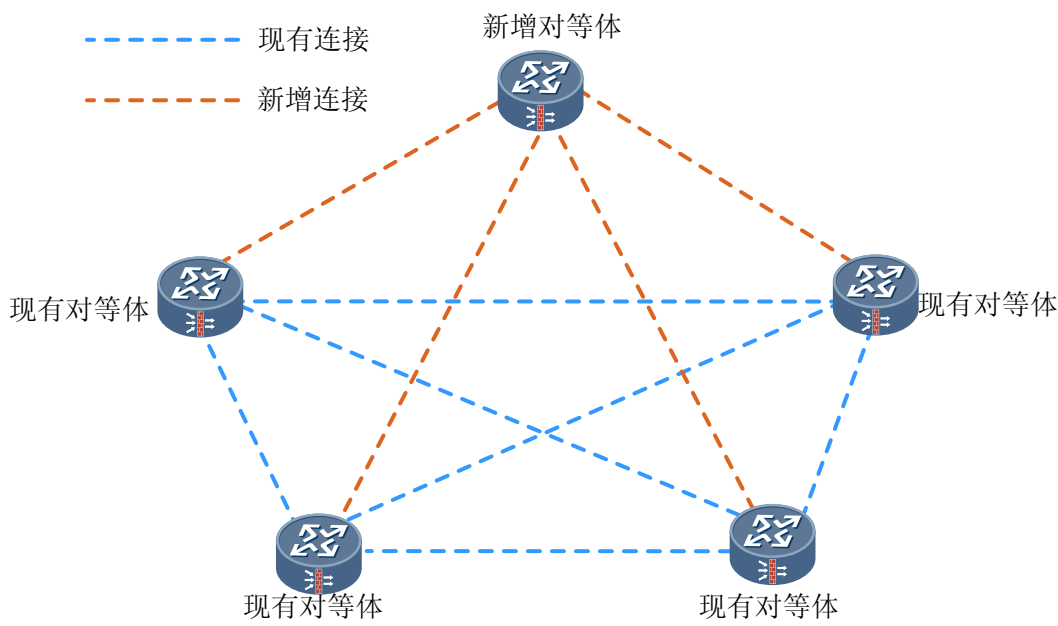
在 IPsec VPN 中，预共享密钥和证书是验证对等体身份的两种常用方法。

预共享密钥是指两台设备事先配置相同密钥，通过检查对方密钥与自己密钥是否一致来完成验证，易于配置。

但当网络中设备数量较多时，每新增一台设备，都需要在所有设备上重新配置预共享密钥，工作量会以指数级速度增长，以图 11-22 为例，有 4 个对等体配置了预共享密钥，如果新增了第 5 台设备，需要从那 4 个站点建立到这台新设备的 VPN 连接。为了确保安全，不得不对每一个对等体建立一个不同密钥。在这台新设备上，需要对其他 4 台设备建立 4 个密钥，在其它 4 台设备上，也需要配置到这台新设备的密钥。

为了帮助用户将设备验证扩展到很大范围内的设备，并且减少中间人攻击的风险，在大型 VPN 中，可以使用证书来进行设备验证。

图 11-22 使用预共享密钥验证方式示意图



11.3.2 参考标准和协议

与证书特性相关的参考标准与协议如下：

- RFC2585  
Internet X.509 Public Key Infrastructure Operational Protocols: FTP and HTTP
- RFC3280  
Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile
- RFC2560  
X.509 Internet Public Key Infrastructure Online Certificate Status Protocol - OCSP
- RFC5019  
The Lightweight Online Certificate Status Protocol (OCSP) Profile for High-Volume Environments

11.3.3 可获得性

版本支持

| 产品                                                                      | 支持版本                                                                                                    |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

特性依赖

在 USG 上，证书特性可以与 IPSec 特性配合使用，为建立 IPSec VPN 会话提供证书的设备认证方式。

11.3.4 原理描述

11.3.4.1 PKI 体系

PKI（Public Key Infrastructure，公钥基础设施）是一个利用公共密钥理论和技术来实现信息安全服务并具有通用安全性的安全基础设施。

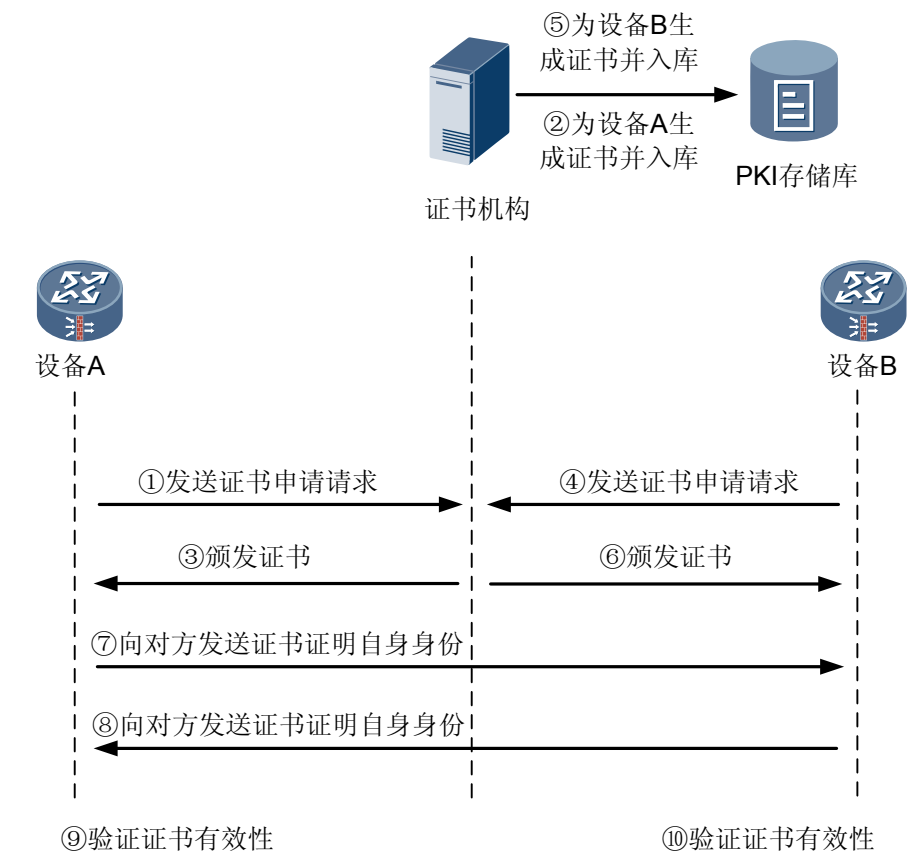
一个 PKI 体系由证书机构、终端实体、PKI 存储库等部分组成，如表 11-3 所示。

表 11-3 PKI 体系组件

| 组件名称                           | 组件作用                                                                                    |
|--------------------------------|-----------------------------------------------------------------------------------------|
| 终端实体                           | 证书申请者和使用者，也就是 USG。                                                                      |
| 证书机构（CA，Certificate Authority） | 用于签发并管理数字证书的第三方机构，其作用包括：发放证书、规定证书的有效期和通过发布 CRL（Certificate Revocation List）确保必要时可以废除证书。 |
| PKI 存储库                        | 各个终端实体证书以及 CRL 列表等信息的集中存放地，提供公众查询，可以是 LDAP 服务器或普通数据库。                                   |

各个组件的交互过程如图 11-23 所示。

图 11-23 PKI 体系示意图



1. 以设备 A 为例，设备 A 向 CA 发送证书建立请求，在请求中包括描述终端特征的实体信息，CA 将使用这些信息来为设备 A 建立实体证书。
2. CA 对收到的设备信息进行验证，为设备 A 生成证书并保存到 PKI 存储库里。

**说明**  
有两种类型的证书：根证书（CA 证书）和实体证书。根证书代表 CA，用来证明 CA 身份，实体证书代表在 CA 域内的设备，用来证明设备身份。在一个 CA 域内，每一个证书都将有一个唯一的序列号，来核实这个证书是否有效或者被吊销。

3. CA 将根证书和设备证书同时颁发给设备 A。
4. 设备 A 使用根证书来验证设备证书的有效性，即验证设备证书是由合法 CA（而不是冒充 CA 的攻击者）颁发的。
5. 设备 B 向 CA 申请证书过程与设备 A 相同。
6. 当设备想使用证书来证明自己身份时，例如通过 IKE 方式建立 IPSec VPN 时，设备可以将自己的证书发送给对等体来进行身份证明。

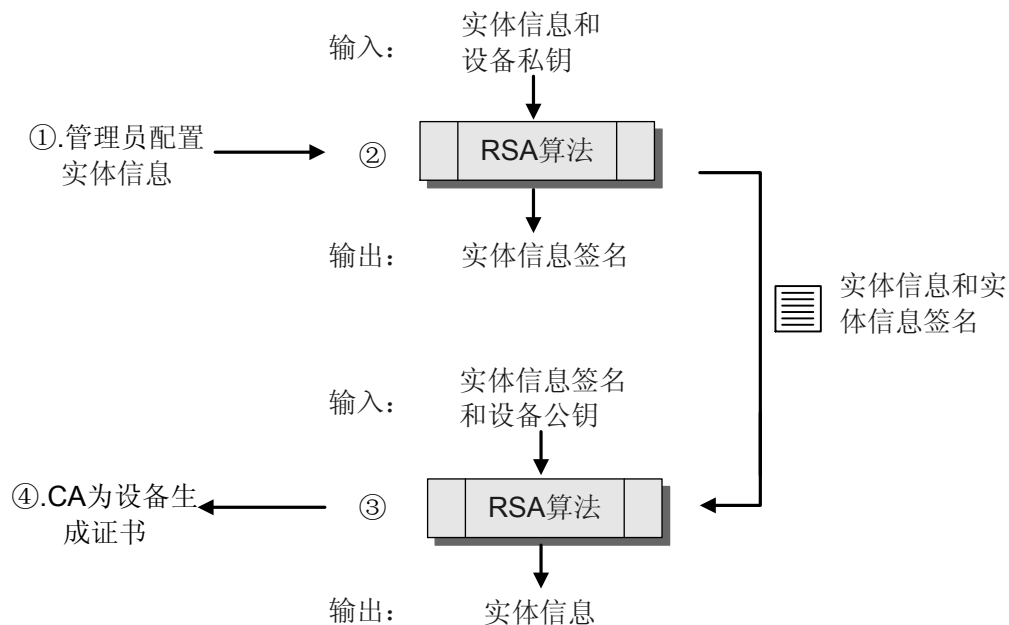
11.3.4.2 证书申请

CA 通过对表示设备特征的实体信息进行签名来为产生实体证书的，因此设备在向 CA 请求建立个人证书时，必须向 CA 提供实体信息。

证书申请过如图 11-24 所示，为了保证实体信息在传输过程中没有被篡改，设备会先使用自己的私钥对包括设备公钥在内的实体信息进行签名，然后，将实体信息和产生的签名一起用于产生一个证书请求发送给 CA。

CA 在收到设备的证书申请请求后，使用包含在实体信息里的公钥来验证实体信息签名，只有签名通过验证，CA 才会为设备建立证书。

图 11-24 证书申请过程示意



目前，USG 支持在线和离线两种方式申请证书：

- 在线方式（带内方式）  
USG 使用 SCEP（Simple Certification Enrollment Protocol）与 CA 服务器通信，在线申请证书，然后将获取到的证书保存在 CF 卡中。
- 离线方式（带外方式）  
USG 生成证书请求文件，管理员通过磁盘、电子邮件等方式将证书申请文件发送给 CA，向 CA 申请证书。

### 11.3.4.3 证书获取

CA 服务器生成证书后，设备需要从 CA 服务器获取根证书和实体证书。

USG 支持使用如下方式获取已经生成的根证书和实体证书：

- 申请证书的方式为在线方式  
在该方式下，USG 使用 SCEP 与 CA 服务器通信，将证书自动保存在 CF 卡中。然后还需要将证书导入到设备中才能生效。
- 申请证书的方式为离线方式  
在该方式下，USG 可以使用 HTTP 或 LDAP 协议与存放证书的服务器通信，将证书下载到 CF 卡中。  
此外，管理员也可以通过磁盘、电子邮件等方式获得证书后，上传到 USG 的 CF 卡中。

然后还需要将证书导入到设备中才能生效。

#### 11.3.4.4 证书吊销列表

证书吊销是指当证书过期或作废时，CA 会吊销证书的使用，有许多原因导致实体证书被吊销，例如：

- 用户的信息变更。
- 用户的私钥泄漏。
- CA 的私钥泄漏
- 证书已经过期，设备需要一个新的实体证书。
- 安全策略改变，对签名功能需要更长（或者更短）的密钥，因此需要一个新的公钥/私钥对、一个新的签名和一个新的实体证书。

当收到来自对方的证书后，设备需要验证这个证书是否被 CA 吊销，证书吊销列表 CRL 就是一种检验证书有效性的方式。

1. CA 把吊销证书的序列号保存到证书吊销列表中。
2. 设备将对方证书缓存到本地，并周期性地从 CA 下载证书吊销列表。
3. 当设备需要验证对方时，由于在缓存中存有对方证书，只需要将证书序列号和 CRL 中的序列号进行对比：如果找到了匹配，则对方证书已经被吊销了，应当向对方和 CA 重新请求证书；如果没有找到匹配，则证明证书是有效的，可以使用存储在本地缓存中的对方证书。

#### 11.3.4.5 OCSP

OCSP 可以实时获取证书的吊销状态，与 CRL 相比，具有更高的时效性。

OCSP（Online Certificate Status Protocol）表示在线证书状态协议。设备之间使用证书机制验证双方身份时，可以通过 OCSP 协议与 OCSP 服务器通信，以在线的方式获取某个证书的吊销状态，以此来检查对方的证书是否被吊销。

与使用证书吊销列表（CRL）检查证书是否被吊销的方式相比，OCSP 方式实时查询证书的吊销状态，因此具有更高的时效性。OCSP 方式可以作为 CRL 方式的补充，应用在金融、股票等涉及大量资金交易的场合中，使交易双方能够及时获得对方的证书状态，确保交易的安全性。

### 11.3.5 证书应用

#### 11.3.5.1 证书在 IPsec VPN 中的应用

在站点到站点和远程访问的 VPN 应用中，当两台设备彼此通信时，用数字证书来检验身份。

以图 11-25 为例，USG\_A 和 USG\_B 都向同一个 CA 申请设备证书，并将根证书和实体证书都下载到了本地。当 USG\_A 和 USG\_B 有数据流量要建立 IPsec VPN 时，它们执行如下的验证过程：

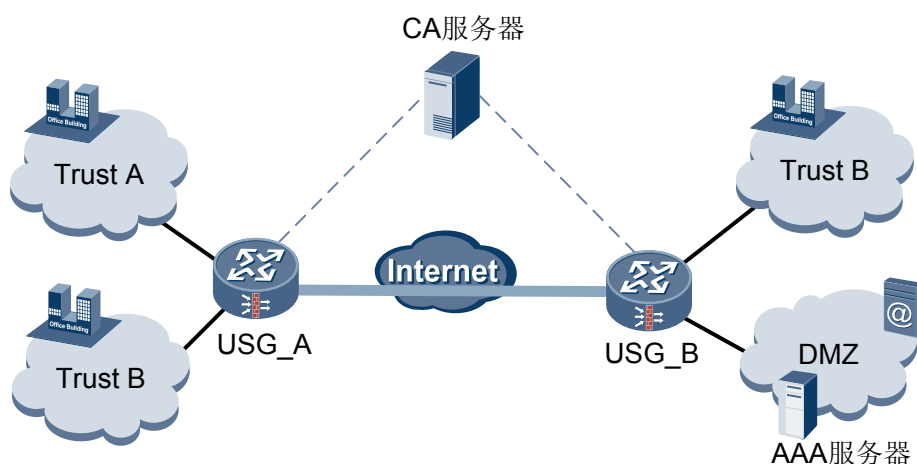
1. 双方设备初始联系之后，它们将共享他们的实体证书。



2. 使用存储在本地的根证书的公钥来核实对方的实体证书签名，CA 在给设备颁发实体证书时，会在实体证书后附加一个签名，我们可以通过根证书中的 CA 公钥来核实对方实体证书的签名。
3. 如果通过了证书签名真实性验证，设备将当前时间和证书上的开始和终止时间做对比。如果设备时间在这两个值的范围内，有效期验证通过，否则，验证失败。
4. 如果开启了 CRL 验证（依赖于设备的配置），设备将会在 CRL 中查找对方证书中的序列号，如果找到了序列号，则认为证书是无效的，验证失败；如果没有找到序列号，则证书验证通过。

一旦证书验证通过，USG\_A 和 USG\_B 就可以建立 IPSec VPN。

图 11-25 证书在 IPSec VPN 中的应用



### 11.3.5.2 基于证书属性的 VPN 访问控制

基于属性的证书访问控制，允许在验证证书有效性之前执行额外的步骤。只有符合特定条件的证书才能通过验证，进而对用户的访问权限进行精细化控制。

设备可以为证书的特定字段制定匹配条件，当设备收到一个证书时，会先查看证书上的特定字段。如果符合匹配条件，设备将接受这个证书，并检查 CA 的签名来核实它的真实性、检查有效日期和吊销状态（最后一项可选，取决于是否配置 CRL 检查），否则将直接拒绝这个证书。

以图 11-26 为例：

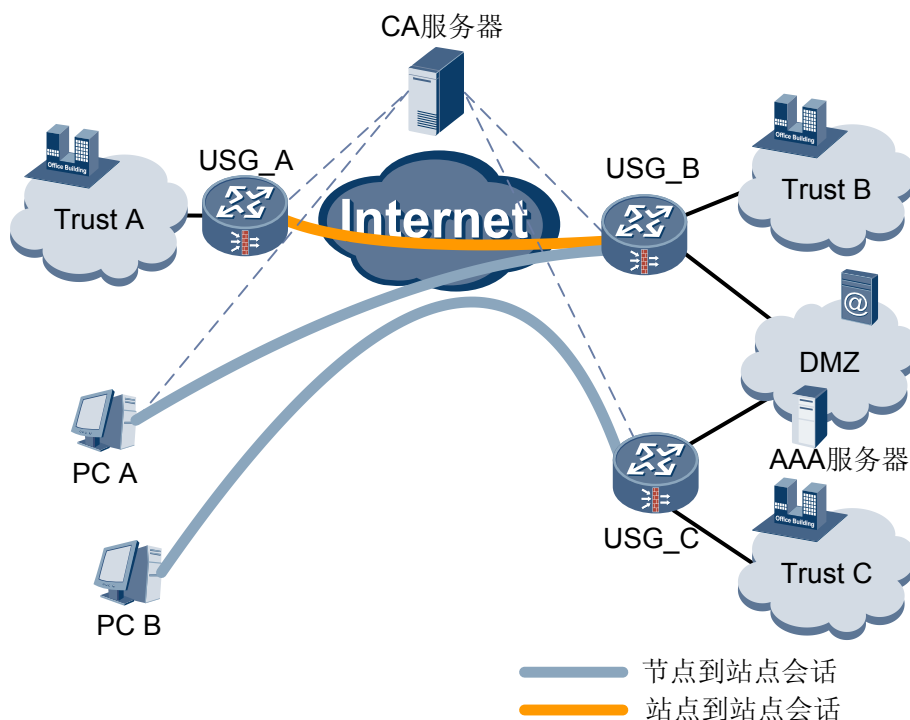
- USG\_A 通过与 USG\_B 建立站点到站点的 VPN 会话来实现区域 Trust A 与区域 Trust B 的互访，USG\_A 和 USG\_B 通过证书方式进行设备验证。
- PC A 通过与 USG\_B 建立节点到站点的 VPN 会话来实现对区域 Trust B 的远程访问，其中，PC A 和 USG\_B 通过证书方式来进行设备验证，通过 AAA 实现用户验证。
- PC B 通过与 USG\_C 建立节点到站点的 VPN 会话来实现对区域 Trust C 的访问，PC B 和 USG\_C 通过证书方式来进行设备验证，通过 AAA 实现用户验证。
- USG\_B 和 USG\_C 使用 DMZ 区域中的同一个 AAA 服务器来对远端用户进行验证，并且 USG\_B 和 USG\_C 的证书是由同一个 CA 服务器颁发的。

这样会导致一个问题，由于 USG\_B 和 USG\_C 使用同一台 AAA 服务器，并且 USG\_B 和 USG\_C 的实体证书是由同一个 CA 服务器颁发的，因此 PC B 也可以建立到 USG\_B 的 IPSec 远程访问会话。

为了在 USG\_B 上限制 PC B 对 USG\_B 发起的远程访问会话，可以在 USG\_B 上定义一个基于证书属性的访问控制策略，只有当证书的主题名 DN 字段等于“pca”时，才会启动证书的有效性检查，否则直接认为证书无效，设备验证失败。

USG 支持根据证书主题名、证书颁发者和证书备用主题名为内容，以包含（ctn）、相等（equ）、不包含（nctn）、不等（nequ）为判断条件，对证书进行过滤。

图 11-26 基于证书属性的 VPN 访问控制



## 11.4 802.1X

802.1X 提供一种更为安全的接入方式，通过对用户接入端口的认证控制，对接入用户进行管理。

### 11.4.1 介绍

#### 定义

802.1X 是一种基于接口的网络接入控制协议，用户使用 EAP（Extensible Authentication Protocol 可扩展的认证协议）报文对局域网接入设备进行认证和控制。

客户端如果能通过认证，就可以访问局域网中的资源；如果不能通过认证，则无法访问局域网中的资源。

## 目的

802.1X 提供一种更为安全的接入方式，通过对用户接入端口的认证控制，对接入用户进行管理。

## 11.4.2 参考标准和协议

- RFC 2284: PPP Extensible Authentication Protocol(EAP)
- RFC 3748: Extensible Authentication Protocol (EAP)
- IEEE Std 802.1X-2001: IEEE Standard for Local and metropolitan area networks--Port-Based Network Access Control

## 11.4.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                           |
|-------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00 |

## 11.4.4 原理描述

### 11.4.4.1 802.1X 的基本原理

#### 802.1X 标准概述

IEEE 802.1X 标准（以下简称 802.1X）是一种基于接口的网络接入控制（Port Based Network Access Control）协议。802.1X 的提出起源于 IEEE 802.11 标准—无线局域网用户接入协议标准，其最初目的主要是解决无线局域网用户的接入认证问题。后来，802.1X 协议作为局域网接口的一个普通接入控制机制应用于以太网中，主要用于解决以太网内认证和安全方面的问题。

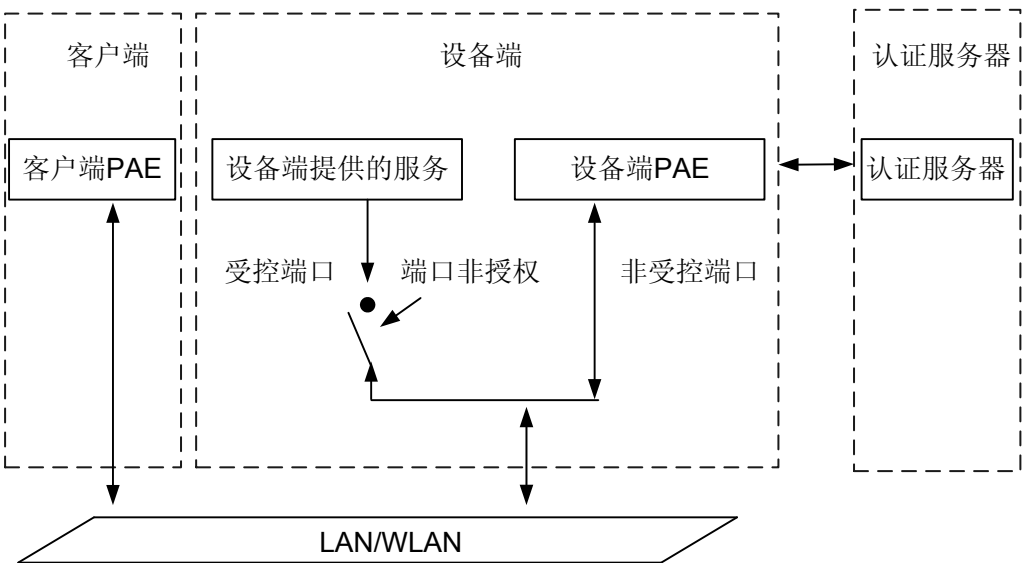
“基于接口的网络接入控制”是指在局域网接入控制设备的接口对接入的设备进行认证和控制。用户设备如果能通过认证，就可以访问局域网中的资源；如果不能通过认证，则无法访问局域网中的资源。

#### 802.1X 认证系统体系结构

如图 11-27 所示。802.1X 的系统为典型的 Client/Server 体系结构，包括三个实体：

- **Supplicant（客户端）**  
客户端一般为用户终端设备，由设备端对其进行认证。客户端需要安装 802.1X 的客户端软件，如 Windows 自带的 802.1X 客户端。客户端必须支持局域网上的可扩展认证协议 EAPoL（Extensible Authentication Protocol over LAN）。
- **Authenticator（设备端）**  
设备端通常为支持 802.1X 协议的网络设备，它为客户端提供接入局域网的接口。
- **Authentication Server（认证服务器）**  
认证服务器是为设备端提供认证服务的实体。认证服务器用于实现对用户进行认证、授权和计费，通常为 RADIUS（Remote Authentication Dial-In User Service，远程认证拨号用户服务）服务器。

图 11-27 802.1X 认证系统体系结构



- **PAE**  
PAE（Port Access Entity，接口访问实体）是 802.1x 认证机制中负责执行算法和协议操作的实体。
- **受控端口**  
受控端口在授权状态下处于双向连通状态，用于传递业务报文；在非授权状态下禁止设备从客户端接收任何报文。
- **非受控端口**  
非受控端口始终处于双向连通状态，主要用来传递 EAPoL 协议帧，保证客户端始终能够发出或接收认证报文。

## 认证实体间的信息交换关系

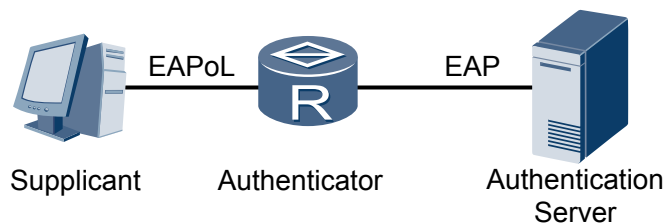
三个认证实体的信息交换关系：

- **Authenticator 与 Authentication Server 之间**通过 EAP（Extensible Authentication Protocol，可扩展认证协议）帧交换信息。
- **Supplicant 与 Authenticator 间**通过 IEEE 802.1X 所定义的 EAPoL（EAP over LANs）帧交换信息。

EAP 帧中封装了认证数据，该认证数据将被封装在其它 AAA 上层协议（如 RADIUS）的报文中以穿越复杂的网络到达认证服务器。

如图 11-28 所示。

图 11-28 认证实体的信息交换关系



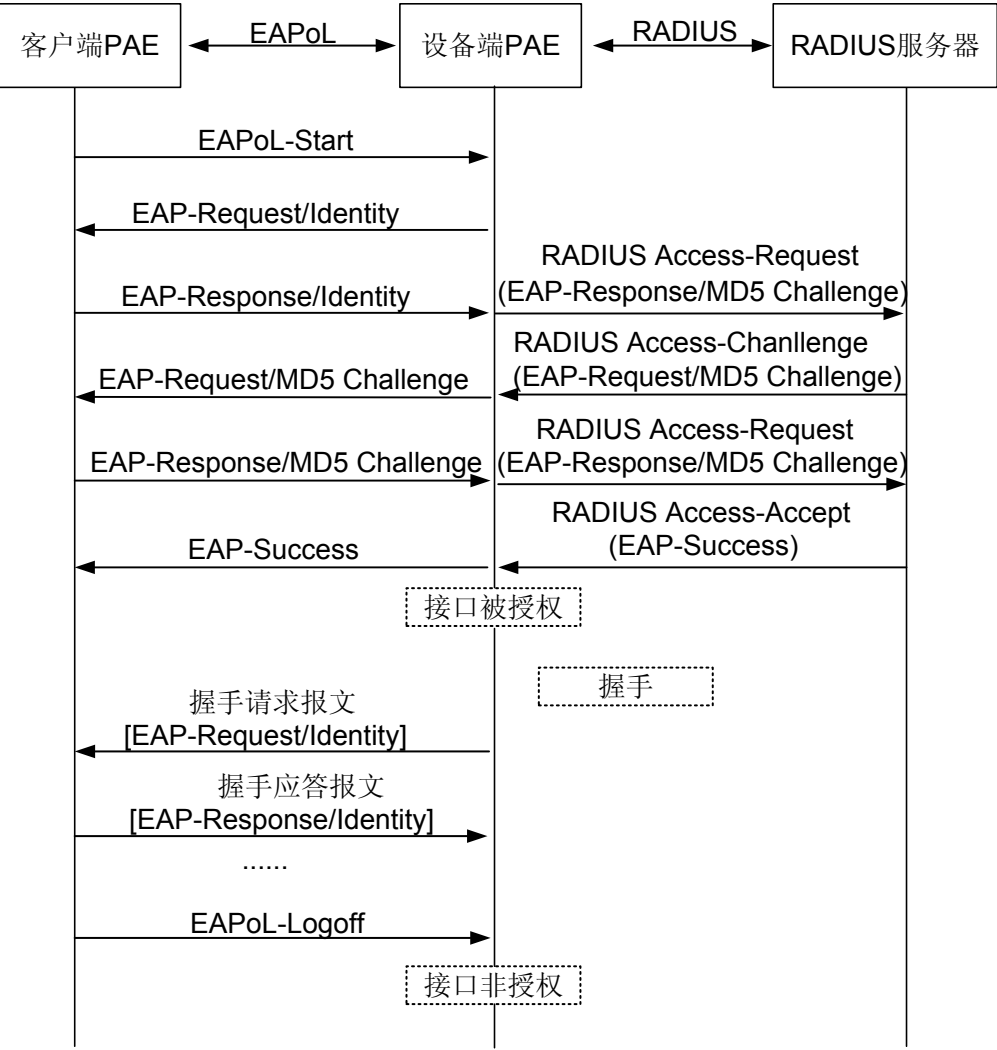
#### 11.4.4.2 802.1X 的认证过程

##### 802.1X 认证过程

EAP 中继方式是 IEEE 802.1X 标准规定的。EAP（扩展认证协议）承载于其它高层协议中，如 EAP over RADIUS，从而使扩展认证协议报文穿越复杂的网络到达认证服务器。EAP 中继方式需要 RADIUS 服务器支持 EAP 属性：EAP-Message 和 Message-Authenticator。

以设备端 PAE 对 EAP 报文进行中继转发为例，IEEE 802.1X 认证系统的基本业务流程如图 11-29 所示。

图 11-29 802.1X 认证系统认证过程



- 认证过程如下：
1. 用户打开 802.1X 客户端程序，输入已经申请、登记过的用户名和密码，发起认证请求（EAPOL-Start 报文）。
  2. 设备端收到请求认证的数据帧后，将发出一个请求帧（EAP-Request/Identity 报文），要求用户的客户端程序发送输入的用户名。
  3. 客户端程序响应设备端发出的请求，将用户名信息通过数据帧（EAP-Response/Identity 报文）发送给设备端。
  4. 设备端将客户端发送的数据帧经过封包处理后（RADIUS Access-Request 报文）送给认证服务器进行处理。
  5. 认证服务器收到设备端转发的用户名信息后，在数据库中找到该用户名对应的密码信息，用随机生成的一个加密字对密码进行加密处理，同时也将此加密字通过 RADIUS Access-Challenge 报文发送给设备端。
  6. 客户端程序收到由设备端传来的加密字（EAP-Request/MD5 Challenge 报文）后，用该加密字对密码进行加密处理，生成 EAP-Response/MD5 Challenge 报文，并通过设备端传给认证服务器。

7. 认证服务器将收到的加密后的密码信息（RADIUS Access-Request 报文）和本地加密运算后的密码信息进行对比，如果相同，则发送认证通过的消息（RADIUS Access-Accept 报文和 EAP-Success 报文）。
8. 设备收到认证通过消息后，接口成为授权状态，允许用户通过接口访问网络。

## 用户下线

当满足下列条件之一时，用户下线：

- 客户端认证失败
- 管理员将受控端口强制置为非授权状态
- 与端口关联的 MAC 不可操作（硬件故障或者被管理员禁用）
- 客户端与设备端之间的物理连接出现故障，导致受控端口的授权状态老化
- 重新认证过程失败
- 客户端 PAE 无法响应设备端 PAE 发出的认证请求
- 客户端 PAE 发送 EAPoL-Logoff 下线请求

客户端 PAE 可以在任何时刻、任何情况下发起 EAPoL-Logoff 下线请求。客户端退出或者客户端需要重新启动时，建议客户端 PAE 主动发送一个 EAPoL-Logoff 下线请求以确保接入安全。

# 12 IP 组播

---

## 关于本章

### [12.1 组播概述](#)

### [12.2 IGMP](#)

IGMP 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。

### [12.3 IGMP Snooping](#)

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。

### [12.4 PIM](#)

PIM 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。

### [12.5 MSDP](#)

MSDP 是为了解决多个 PIM-SM 域之间的互连而开发的一种域间组播解决方案，用来发现其它 PIM-SM 域内的组播源信息。

### [12.6 组播路由与转发](#)

介绍组播路由和转发的基本原理、“控制平面”与“转发平面”之间的信息交互过程，以及用户控制组播转发功能的方式。



## 12.1 组播概述

### 12.1.1 介绍

介绍了 IP 组播的基础知识、协议功能、实现原理和典型应用。

#### 定义

组播（Multicast）技术能够有效地解决单点发送、多点接收的问题，从而实现了网络中点到多点的高效数据传送，能够节约大量网络带宽、降低网络负载。

#### 目的

IP 组播技术可以提供对带宽和数据交互的实时性要求较高的信息服务，包括在线直播、网络电视、远程教育、远程医疗、网络电台、实时视/音频会议等。

### 12.1.2 原理描述

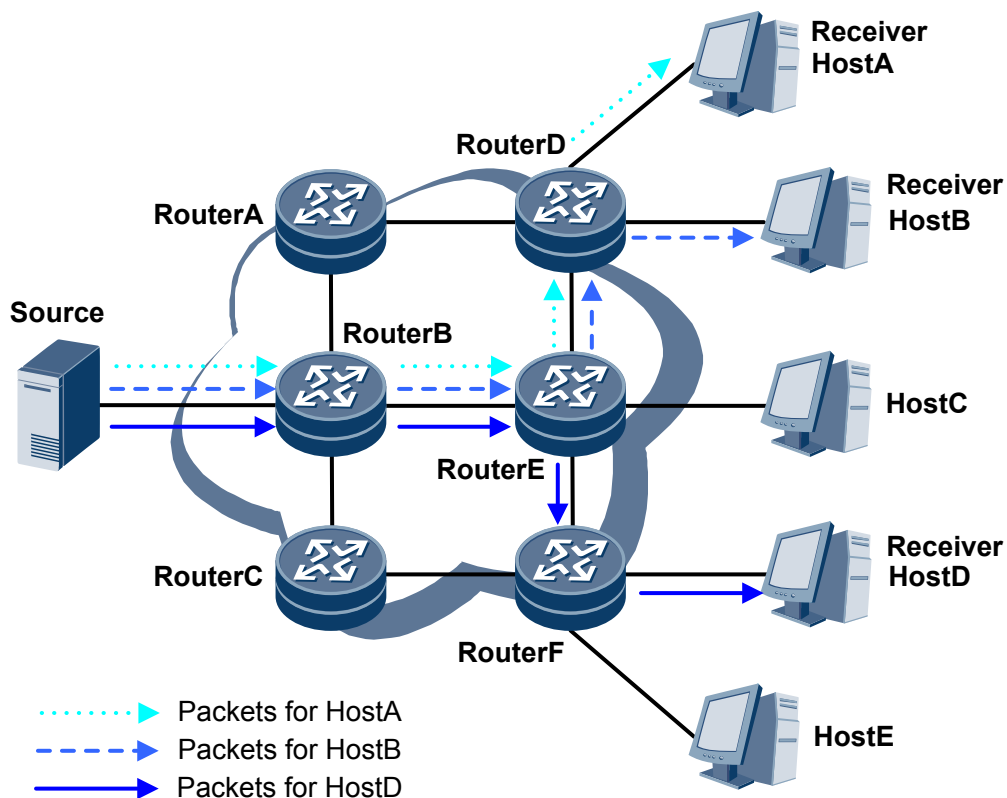
#### 12.1.2.1 组播特点和应用

作为一种与单播（Unicast）和广播（Broadcast）并列的通信方式，组播（Multicast）技术能够有效地解决单点发送、多点接收的问题，从而实现了网络中点到多点的高效数据传送，能够节约大量网络带宽、降低网络负载。

#### 单播方式

如图 12-1 所示，网络中采用单播方式传输信息，信息源（即 Source）要为每个需要信息的主机（即 Receiver）都发送一份独立的拷贝信息。

图 12-1 单播方式传输信息



假设 HostA、HostB 和 HostD 需要信息，则 Source 要与 HostA、HostB 和 HostD 分别建立一条独立的信息传输通道。

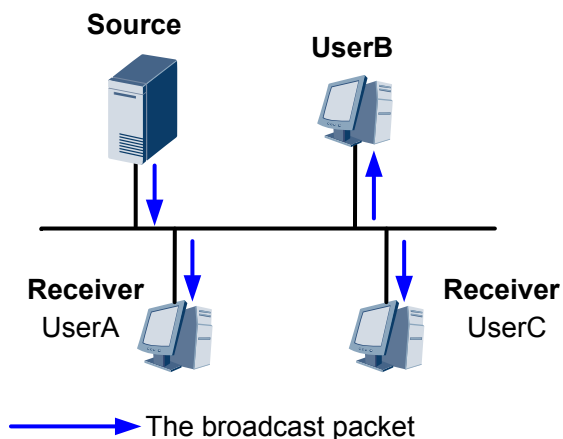
单播方式下，网络中传输的信息量和需求该信息的用户量成正比，当需求该信息的用户量较大时，信息源需要将多份内容相同的信息发送给不同的用户，这对信息源以及网络带宽都将造成巨大的压力。

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

如图 12-2 所示，在一个网段中若采用广播的方式，信息源（即 Source）将把信息传送给该网段中的所有主机，而不管其是否需要该信息。

图 12-2 广播方式传输信息



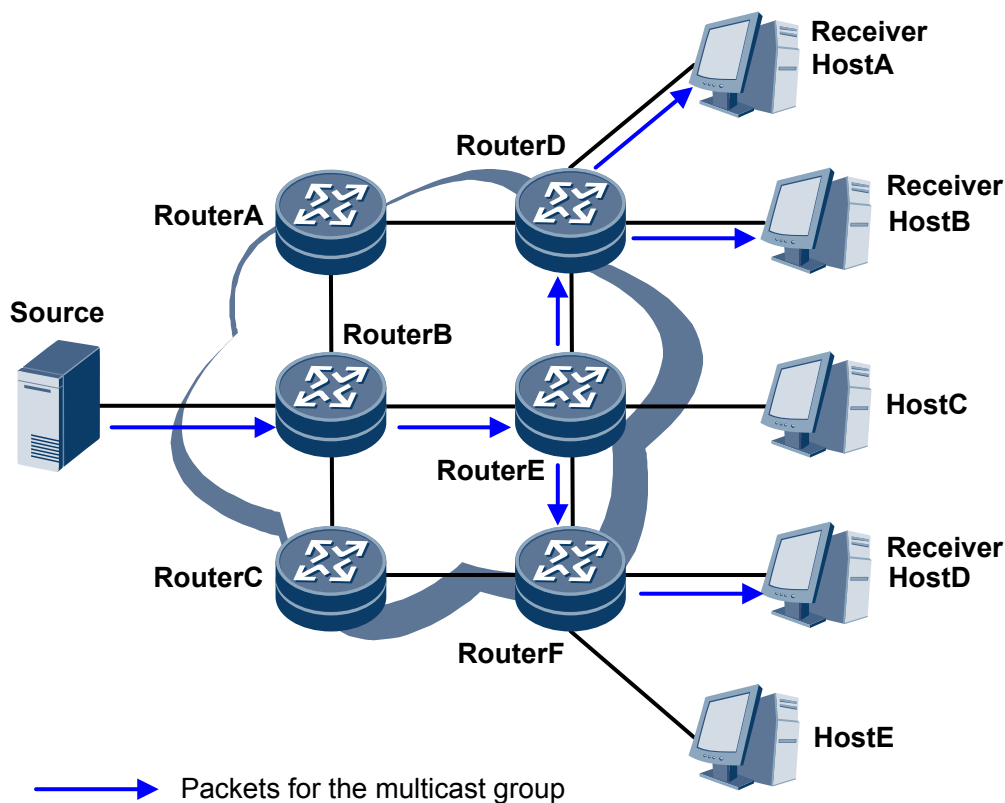
假设只有 UserA 和 UserC 需要信息，若将该信息在网段中进行广播，则原本不需要信息的 UserB 也将收到该信息。

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，可能会造成同一网段中信息的泛滥。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 12-3 所示，网络中采用组播的方式传输信息。当网络中的某些主机（即 Receiver）需要信息时，组播源（即 Source）仅需向一个组播组发送一份信息，组播路由协议为此信息建立组播分发树。这样，被传递的信息在距离组播源尽可能远的网络节点（如 RouterF）才开始复制和分发。

图 12-3 组播方式传输信息



假设只有 HostA、HostB 和 HostD 需要信息，可以让这些主机加入同一个组播组，组播源只需向该组播组发送一份信息，并由网络中各路由器根据组播协议对该信息进行复制和转发，最后该信息会准确地发送给 HostA、HostB 和 HostD。

## 组播的优势

组播的优势归纳如下：

- 相比单播来说，组播的优势在于：由于被传递的信息在距信息源尽可能远的网络节点才开始被复制和分发，所以用户的增加不会导致信息源负载的加重以及网络资源消耗的显著增加。
- 相比广播来说，组播的优势在于：由于被传递的信息只会发送给需要该信息的接收者，所以不会造成网络资源的浪费，并能提高信息传输的安全性。另外，广播只能在同一网段中进行，而组播可以实现跨网段的传输。

### 12.1.2.2 组播基本概念

#### 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，就成为了该组播组的成员，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，发送组播信息的信源称为组播源。

- 一个组播源可以同时向多个组播组发送报文。
- 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。

组播源作为组播数据的发送者，通常不会同时是组播数据的接收者，因此不属于组播组成员。

为了更好地理解，可以将组播方式的传输信息过程类比于电视节目的传送过程，如表 12-1 所示。

表 12-1 组播信息传输与电视节目传输的类比

| 步骤 | 电视节目的传送过程                        | 组播方式的传输信息过程                     |
|----|----------------------------------|---------------------------------|
| 1  | 电视台 S 通过频道 G 传送电视节目              | 组播源 S 向组播组 G 发送组播数据             |
| 2  | 用户 U 将电视机的频道调至频道 G               | 接收者 U 加入组播组 G                   |
| 3  | 用户 U 能够收看到由电视台 S 通过频道 G 传送的电视节目了 | 接收者 U 能够收到由组播源 S 发往组播组 G 的组播数据了 |
| 4  | 用户 U 关闭电视机或切换到其它频道               | 接收者 U 离开组播组 G                   |

## 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”。本手册中的组播路由器指的是运行了组播功能的 USG。

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，组播路由协议为组播数据在网络中转发建立树型路由。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。

## 组播中常用表示方法

在组播中，经常出现以下两种表示方式：

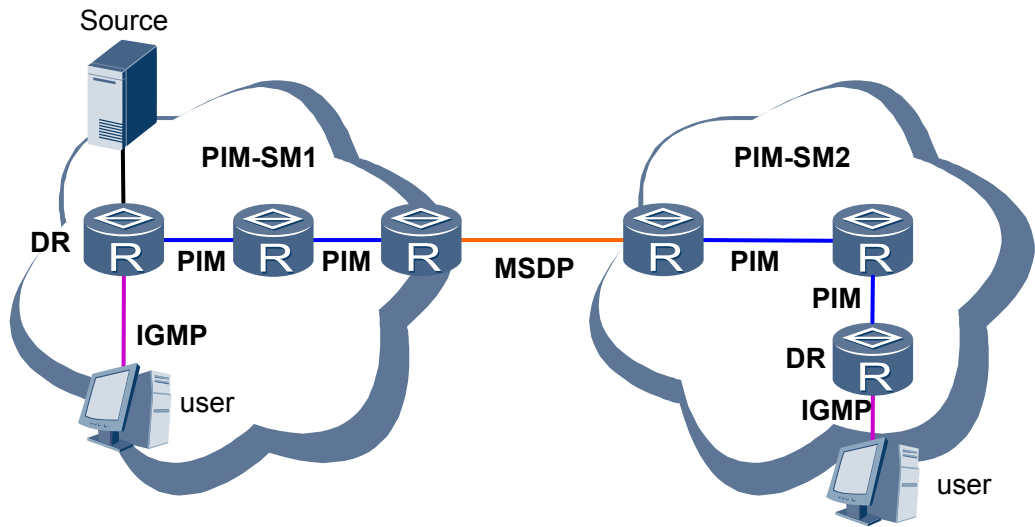
- (\*, G)：通常用来表示共享树（RPT），或者由任意组播源发往组播组 G 的组播报文。其中的“\*”代表任意组播源，“G”代表特定组播组 G。

- (S, G)：称为“组播源组”，通常用来表示最短路径树（SPT），或者由组播源 S 发往组播组 G 的组播报文。其中的“S”代表特定组播源 S，“G”代表特定组播组 G。

12.1.2.3 组播协议

实现一套完整的组播服务，需要部署在网络各个位置的多种组播协议相互配合，共同运作，如图 12-4 所示。

图 12-4 组播协议的应用位置



针对不同的应用位置与目的，USG 支持功能各异的多种组播路由协议，如表 12-2 所示：

表 12-2 组播协议

| 组播协议                                                      | 应用位置         | 目的                                                                                                                                  |
|-----------------------------------------------------------|--------------|-------------------------------------------------------------------------------------------------------------------------------------|
| IGMP（Internet Group Management Protocol）组播组管理协议           | 用户主机与组播路由器之间 | 将用户主机接入组播网络： <ul style="list-style-type: none"><li>● 在主机侧实现组播组成员动态加入与离开。</li><li>● 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。</li></ul> |
| PIM（Protocol Independent Multicast）协议无关组播，包括 DM 和 SM 两种模式 | 域内组播路由器之间    | 组播路由与转发： <ul style="list-style-type: none"><li>● 按需创建组播路由。</li><li>● 动态响应网络拓扑变化，维护组播路由表。</li><li>● 按照路由表项执行转发。</li></ul>            |

| 组播协议                                             | 应用位置      | 目的                                                                                                                  |
|--------------------------------------------------|-----------|---------------------------------------------------------------------------------------------------------------------|
| MSDP（Multicast Source Discovery Protocol）组播源发现协议 | 域间组播路由器之间 | 域间组播源信息共享： <ul style="list-style-type: none"><li>● 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。</li><li>● 在不同域的路由器之间传递源信息。</li></ul> |

 说明

- PIM-DM 属于密集模式的组播路由协议，适用于组成员分布相对密集的小型网络。
- PIM-SM 属于稀疏模式的组播路由协议，适用于组成员分布相对分散、范围较广、大规模的网络。

12.1.2.4 组播地址

IP 组播地址

IPv4 组播地址范围是：224.0.0.0 ~ 239.255.255.255，范围及含义见表 12-3。

组播数据业务可用的组地址范围是：224.0.1.0 ~ 239.255.255.255。任何用户主机（或其他接收设备）加入此范围内的某组播组，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。网络中的组成员可能广泛分布在网络中的任何地方，用户主机可以在任何时刻加入和离开组播组。

表 12-3 组播地址的范围及含义

| 地址范围                        | 含义                                                                                                                                                                                                                                                                            |
|-----------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 224.0.0.0 ~ 224.0.0.255     | 永久组地址。除 224.0.0.0 保留不做分配外，其它地址供路由协议、拓扑查找和协议维护等使用，不用于组播转发。                                                                                                                                                                                                                     |
| 224.0.1.0 ~ 238.255.255.255 | 用户组地址，全网范围内有效。其中： <ul style="list-style-type: none"><li>● 224.0.1.0 ~ 224.0.1.255 保留地址。</li><li>● 232.0.0.0 ~ 232.255.255.255 为组播源（SSM）地址。</li><li>● 233.0.0.0 ~ 238.255.255.255 为 GLOP 组地址。GLOP 是一种 AS 之间的组播地址分配机制，将 AS 号填入该范围内组播地址的中间两个字节中，每个 AS 都可以得到 255 个组播地址。</li></ul> |
| 239.0.0.0 ~ 239.255.255.255 | 本地管理组播地址。缺省的 BSR 管理域组地址范围，仅在 BSR 管理域内有效，属于私有地址。在不同的 BSR 管理域内使用相同的地址不会冲突。                                                                                                                                                                                                      |

- 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 12-4。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 12-4 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              |
| 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                  |

## 以太网组播 MAC 地址

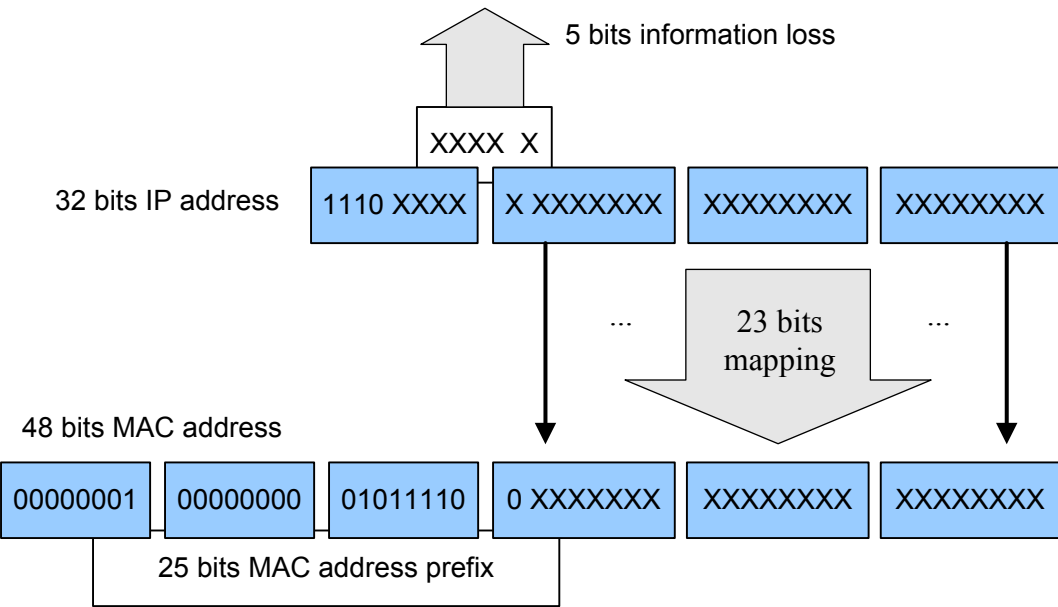
组播 MAC 地址用于在链路层上标识属于同一组播组的接收者。



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 12-5 所示。

图 12-5 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

本手册着重介绍 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。

12.1.2.5 组播模型

根据接收者对组播源处理方式的不同，组播模型分为以下三类：

ASM 模型

ASM（Any-Source Multicast，任意信源组播）模型就是任意源组播模型。

在 ASM 模型中，任意一个发送者都可以作为组播源向某组播组地址发送信息。众多接收者通过加入由该组播组地址标识的组播组以获得发往该组播组的组播信息。

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

ASM 模型是目前最常用的组播模型。PIM-DM 和 PIM-SM 协议都可以构建 ASM 模型。

SFM 模型

SFM（Source-Filtered Multicast，信源过滤组播）模型继承了 ASM 模型，从发送者角度来看，两者的组播组成员关系完全相同。

SFM 模型在功能上对 ASM 模型进行了扩展。在 SFM 模型中，上层软件对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。因此，接收者只能收到来自部分组播源的组播数据。从接收者的角度来看，只有部分组播源是有效的，组播源经过了筛选。

## SSM 模型

在现实生活中，用户可能只对某些组播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（Source-Specific Multicast，指定信源组播）模型为用户提供了一种能够在客户端指定组播源的传输服务。

SSM 模型与 ASM 模型的根本区别在于：SSM 模型中的接收者已经通过其它手段预先知道了组播源的具体位置。SSM 模型使用与 ASM/SFM 模型不同的组播地址范围，直接在接收者与其指定的组播源之间建立专用的组播转发路径。

SSM 模型应用于接收者预先知道组播源具体位置的情况。PIM-SM 协议可以构建 SSM 模型。SSM 模型需要 IGMPv3 的支持。

### 12.1.2.6 组播报文转发

在组播模型中，IP 报文的目地址字段为组播组地址，组播源向此目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，需要将从一个入接口接收到的报文转发到多个出接口。与单播模型相比，组播模型的复杂性就在于此。

组播报文转发基本原理如下：

- 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。

 说明

关于组播路由与转发的更多内容请参见 [12.6 组播路由与转发](#)。

## 12.2 IGMP

IGMP 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。

### 12.2.1 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。

## 12.2.2 参考标准和协议

与 IGMP 相关的参考标准和协议如下:

- RFC3228: IANA Considerations for IPv4 Internet Group Management Protocol (IGMP)
- RFC2236: Internet Group Management Protocol, Version 2
- RFC3376: Internet Group Management Protocol, Version 3

## 12.2.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 12.2.4 原理描述

### 12.2.4.1 IGMPv1

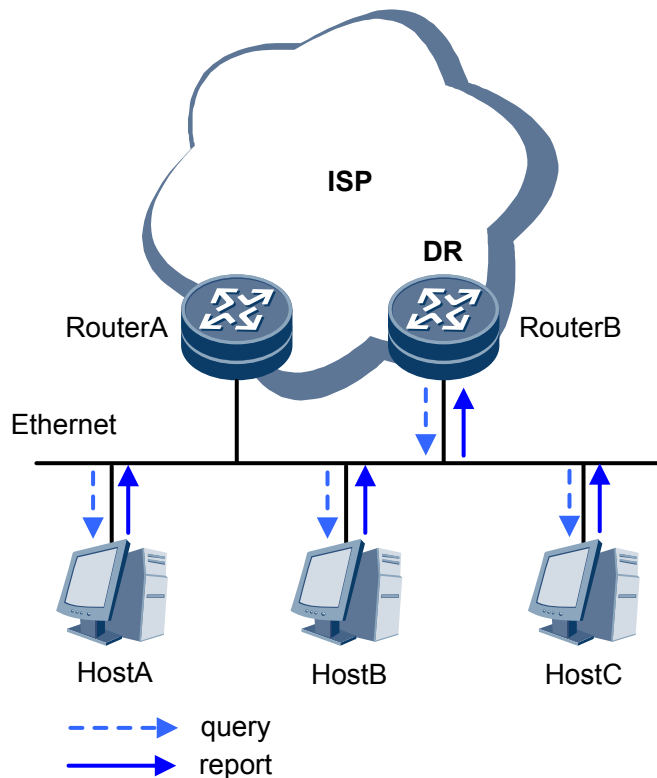
IGMPv1 协议主要基于查询和响应机制完成组播组管理。在连接用户主机的多路由器共享网段上, 由组播路由协议选举出唯一的组播信息转发者 (DR), 并作为 IGMPv1 的查询器, 负责该网段的组成员关系查询。

以图 12-6 为例说明 IGMPv1 的基本功能。

 说明

有关 DR 的介绍请参见 12.4.4.3 PIM-SM。

图 12-6 IGMP 典型组网图



## 组成员关系维护

1. IGMP 查询器（RouterB）周期性地发送 General-Query 消息。发送周期可以使用命令配置，缺省值是 60 秒。
2. 网段内所有主机都接收到该 General-Query 消息，组成员在本地启动针对该组的定时器，超时时间为 0 ~ 10 秒之间的随机值。HostA 不是任何组播组的成员，不做任何响应；HostB 是组播组 G1 成员，则在本地启动定时器 Timer-G1。HostC 同时是组播组 G1 和 G2 的成员，则在本地启动定时器 Timer-G1 和 Timer-G2。
3. 第一个定时器超时的主机发送针对该组的 Report 消息。HostB 上的 Timer-G1 首先超时，向该网段发送 Report 消息，目的地址为 G1。
4. 网段内的组 G1 的其它成员 HostC 侦听到 Report 消息，则停止定时器 Timer-G1，不再发送针对 G1 的 Report 消息。即 Report 消息被抑制。
5. HostC 上的 Timer-G2 超时，向该网段发送 Report 消息，目的地址为 G2。
6. 网段中的路由器接收到 Report 消息后，刷新对应的 Group Timer，了解到本网段内存在组播组 G1 和 G2 的成员，则由组播路由协议生成 (\*, G1) 和 (\*, G2) 组播转发表项，并添加对应的下游接口。网络中一旦有组播组 G1 和 G2 的数据到达路由器，将向该网段转发。

## 新组成员加入

1. HostA 加入组播组 G3，立即发送针对 G3 的 Report 消息。
2. 网段中的路由器接收到 Report 消息后，了解到本网段内出现了组播组 G3 的成员，则由组播路由协议生成 (\*, G3) 组播转发表项，并添加对应的下游接口。网络中一旦有组播组 G3 的数据到达路由器，将向该网段转发。

## 组成员离开

1. 退出组播组的主机将不再反馈 Report 消息。HostA 退出组播组 G3。当 General-Query 消息的到来时，HostA 不启动 Timer-G3，不再发送针对 G3 的 Report 消息。
2. 由于该网段不存在其它的组 G3 成员，则路由器无法收到 G3 的 Report 消息，定时器 Timer-Membership 超时后，删除 (\*, G3) 表项。如果这时网络中组播组 G3 的数据到达路由器，则不会向该网段转发。

### 12.2.4.2 IGMPv2

与 IGMPv1 相比，IGMPv2 能够不依赖上层组播路由协议选举查询器和及时获知组播组成员的离开。

## 查询器选举机制

在 IGMPv1 中，当某共享网段上存在多个组播路由器时，由组播路由协议（如 PIM）选举的指定组播路由器（DR）充当查询器。

在 IGMPv2 中，增加了独立的查询器选举机制，可以不依赖于上层组播路由协议选举查询器。其选举过程如下：

1. 所有 IGMPv2 路由器在初始时都认为是查询器，并向本地网段内的所有主机和路由器（目的地址为 224.0.0.1）发送 IGMP 普遍组查询（General-Query）报文。
2. 本地网段中的其它 IGMPv2 路由器在收到该报文后，将报文的源 IP 地址与自己的接口地址作比较。通过比较，IP 地址最小的路由器将成为查询器，其它路由器成为非查询器（Non-Querier）。
3. 所有非查询器上都会启动一个定时器（即其它查询器存在时间定时器，Other Querier Present Timer）。在该定时器超时前，如果收到了来自查询器的 IGMP 查询报文，则重置该定时器；否则，就认为原查询器失效，并发起新的查询器选举过程。

## 离开组机制

在 IGMPv1 中，主机离开组播组时不会向组播路由器发出任何通知，导致组播路由器只能依靠组播组成员查询的响应超时来获知组播组成员的离开。

而在 IGMPv2 中，当一个主机离开某组播组时：

1. 该主机向本地网段内的所有组播路由器（目的地址为 224.0.0.2）发送离开组（Leave Group）报文。
2. 当查询器收到该报文后，向该主机所声明要离开的那个组播组发送特定组查询（Group-Specific Query）报文（目的地址字段和组地址字段均填充为所要查询的组播组地址）。
3. 如果该网段内还有该组播组的其它成员，则这些成员在收到特定组查询报文后，会在该报文中所设定的最大响应时间（Max Response Time）内发送成员关系报告报文。
4. 如果在最大响应时间内收到了该组播组其它成员发送的成员关系报告报文，查询器就会继续维护该组播组的成员关系；否则，查询器将认为该网段内已无该组播组的成员，于是不再维护这个组播组的成员关系。

### 12.2.4.3 IGMPv3

## 主机指定源/组加入

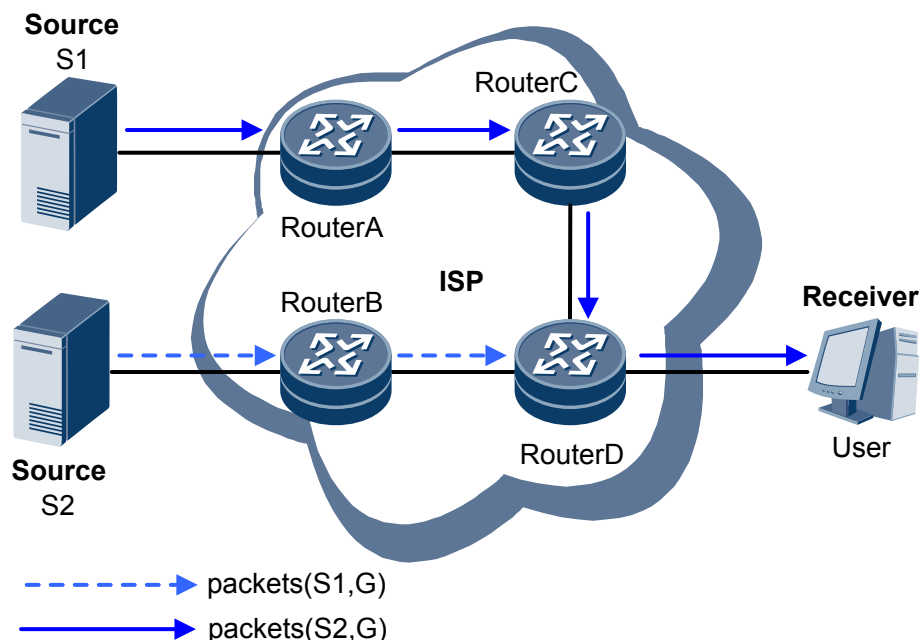
IGMPv3 增加了针对组播源的过滤模式 (INCLUDE/EXCLUDE)，使主机在加入某组播组 G 的同时，能够明确要求接收或拒绝来自某特定组播源 S 的组播信息。

当主机加入组播组时：

- 若要求只接收来自指定组播源如 S1、S2、……的组播信息，则其 Report 消息中可以标记为 (G, INCLUDE, (S1、S2...))。
- 若拒绝接收来自指定组播源如 S1、S2、……的组播信息，则其 Report 消息中可以标记为 (G, EXCLUDE, (S1、S2...))。

如图 12-7 所示，网络中存在 S1 和 S2 两个组播源，均向组播组 G 发送组播报文。User 仅希望接收从组播源 S1 发往组播组 G 的信息，而对来自 S2 的信息没有兴趣。

图 12-7 指定源/组的组播流路经



如果主机和路由器之间应用 IGMPv1 或 IGMPv2，User 加入到某组播组 G 后，将同时接收到来自组播源 S1 和 S2 的组播信息。如果采用 IGMPv3，主机 User 可以选择仅接收 S1 组播信息。

- 方法一：User 发送 IGMPv3 报告 (G, IS\_IN, (S1))，仅接收源 S1 向组播组 G 发送的数据。
- 方法二：User 发送 IGMPv3 报告 (G, IS\_EX, (S2))，退出指定源 S2 对应的组播组 G，从而仅有来自 S1 的组播信息才能传递到 User。

## 携带源地址的查询消息

IGMPv3 不仅支持 IGMPv1 的普遍组查询和 IGMPv2 的特定组查询，而且还增加了对特定源组查询的支持：

- 普遍组查询消息中，既不携带组地址，也不携带源地址。
- 特定组查询消息中，携带组地址，但不携带源地址。

- 特定源组查询消息中，既携带组地址，还携带一个或多个源地址。

## 包含多个组记录的报告（Report）消息

IGMPv3 的 Report 消息的目的地址为 224.0.0.22，可以携带一个或多个组记录。在每个组记录中，包含有组播组地址和组播源地址列表，可以过滤此组播组地址的组播源。

组记录可以分为多种类型，如下：

- IS\_IN：表示组播组与组播源列表之间的过滤模式为 INCLUDE，即只接收从指定组播源列表发往该组播组的组播数据。
- IS\_EX：表示组播组与组播源列表之间的过滤模式为 EXCLUDE，即只接收从指定组播源列表之外的组播源发往该组播组的组播数据。
- TO\_IN：表示组播组与组播源列表之间的过滤模式由 EXCLUDE 转变为 INCLUDE。
- TO\_EX：表示组播组与组播源列表之间的过滤模式由 INCLUDE 转变为 EXCLUDE。
- ALLOW：表示在现有状态的基础上，还希望从某些组播源接收组播数据。如果当前的对应关系为 INCLUDE，则向现有组播源列表中添加这些组播源；如果当前的对应关系为 EXCLUDE，则从现有组播源列表中删除这些组播源。
- BLOCK：表示在现有状态的基础上，不再希望从某些组播源接收组播数据。如果当前的对应关系为 INCLUDE，则从现有组播源列表中删除这些组播源；如果当前的对应关系为 EXCLUDE，则向现有组播源列表中添加这些组播源。

### 说明

对于 USG，SSM 范围符合标准协议（RFC4607）；但 ASM 范围仅支持处理 IS\_EX、TO\_EX，而且忽略报文中的源列表。

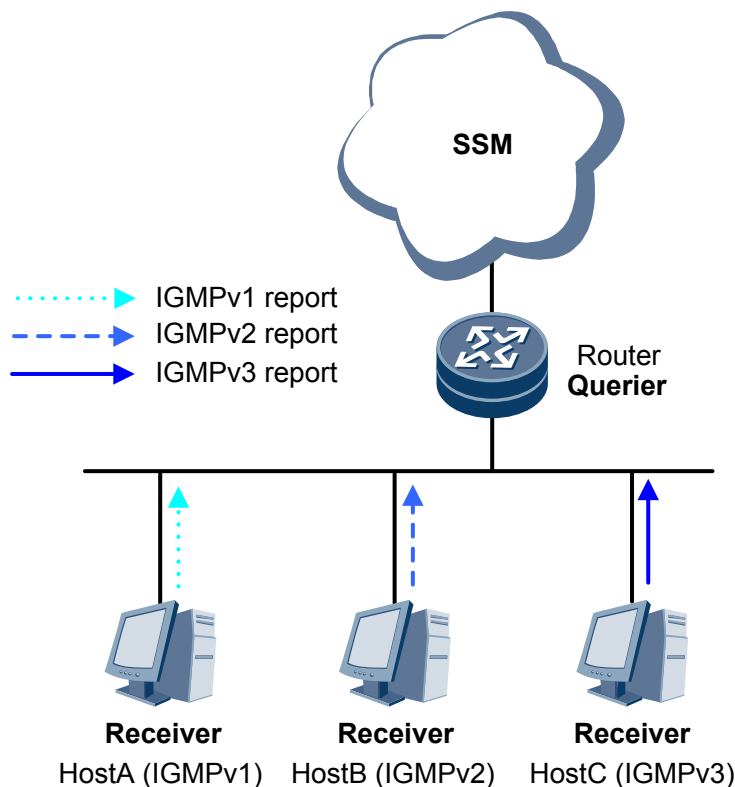
## 12.2.4.4 SSM Mapping

IGMP SSM Mapping 通过在路由器上配置 SSM 静态映射规则，从而为运行 IGMPv1 或 IGMPv2 的接收者主机提供对 SSM 模型的支持。

SSM 模型要求在接收者主机所在的网段，路由器能够了解主机加入组播组时所指定的组播源。

- 如果接收者主机上运行的是 IGMPv3，则可以在 IGMPv3 的报告（Report）消息中直接指定组播源的地址。
- 如果某些接收者主机只能运行 IGMPv1 或 IGMPv2，则在 IGMPv1 或 IGMPv2 的 Report 消息中无法指定组播源的地址。这种情况下需要通过在与主机相连的路由器上配置 IGMP SSM Mapping 功能，将 IGMPv1 或 IGMPv2 的 Report 消息中所包含的（\*，G）信息映射为（G，INCLUDE，（S1，S2...））信息。

图 12-8 SSM Mapping 组网图



如图 12-8 所示，SSM 网络的用户网段中 HostA 运行 IGMPv1、HostB 运行 IGMPv2、HostC 运行 IGMPv3。

在不允许将 HostA 和 HostB 升级为 IGMP 版本为 v3 的情况下，为该网段中的所有主机提供 SSM 组播服务，需要在 Router 上进行如下配置：

- 在连接主机网段的接口上启用 IGMPv3。
- 启用 SSM Mapping。
- 配置将组地址与源地址建立静态映射关系的 SSM Mapping 规则。

配置完成后，当 Router 接收到来自主机的 Report 消息时，首先检查 Report 消息中携带的组地址 G，然后根据检查结果进行如下处理：

- 如果 G 不在 SSM 组地址范围内，则无法为组 G 提供 SSM 模式的组播服务。如果网络允许运行 ASM，则组 G 将采用 ASM 模式。
- 如果 G 在 SSM 组地址范围内，但是 Report 消息中没有指定组播源，Router 上也没有组 G 对应的 SSM Mapping 规则，则无法为组 G 提供组播服务。此情况针对于 IGMPv1 和 IGMPv2。
- 如果 G 在 SSM 组地址范围内，且 Report 消息指定了组播源，则可以为组 G 提供 SSM 组播服务。此情况针对于 IGMPv3。
- 如果 G 在 SSM 组地址范围内，且 Router 上存在组 G 对应的 SSM Mapping 规则，则依照映射策略将 Report 消息中包含的 (\*, G) 信息，转换成 (G, INCLUDE, (S1, S2...)) 信息，则可以为组 G 提供 SSM 组播服务。此情况针对于 IGMPv1 和 IGMPv2。

#### 12.2.4.5 IGMP 消息



以下分类介绍了 IGMPv1、IGMPv2 和 IGMPv3 的各种消息。

## IGMPv1 消息

表 12-5 IGMPv1 消息

| 名称                     | 定义                                                 | 说明                                                                                                                              |
|------------------------|----------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| 普遍组查询消息（General-Query） | 普遍组查询消息是查询器定期向共享网段内所有主机以组播方式发送的查询消息，用于查询哪些组播组存在成员。 | 封装该消息的 IP 报文头的目的地址字段为 224.0.0.1，本地网段上的所有路由器和主机都能识别并接收。<br>消息中的组地址字段为全 0，表示不指定组播组。                                               |
| 成员报告消息（Report）         | 成员报告消息是主机向组播路由器发送的报告消息，用于申请加入某个组播组或者应答查询消息。        | 封装该消息的 IP 报文头的目的地址字段为主机要加入的组播组地址，本地网段上的所有路由器及属于该组的所有主机都能够识别并接收。同一组的其它成员侦听到该报文后，不再发送针对相同组播组的报告，即报告被抑制。<br>消息中的组地址字段为主机要加入的组播组地址。 |

## IGMPv2 消息

表 12-6 IGMPv2 消息

| 名称                            | 定义                                            | 说明                                                                                   |
|-------------------------------|-----------------------------------------------|--------------------------------------------------------------------------------------|
| 普遍组查询消息（General-Query）        | 同 IGMPv1                                      | 同 IGMPv1                                                                             |
| 特定组查询消息（Group-Specific Query） | 特定组查询消息是查询器向共享网段内特定组播组成员发送的消息，用于查询该组播组是否存在成员。 | 封装该消息的 IP 报文头的目的地址字段为被查询的组播组的 IP 地址，网络中属于该组播组的成员才能识别并响应。<br>消息中的组地址字段为被查询的组播组 IP 地址。 |
| 成员报告消息（Report）                | 同 IGMPv1                                      | 同 IGMPv1                                                                             |
| 离开消息（Leave）                   | 离开消息是主机主动离开组播组时向组播路由器发送的消息，用于宣告自己离开了某个组播组。    | 封装该消息的 IP 报文头的目的地址字段为 224.0.0.2，本地网段上的所有路由器都能识别并接收。<br>消息中的组地址字段为主机要离开的组播组地址。        |

IGMPv3 消息

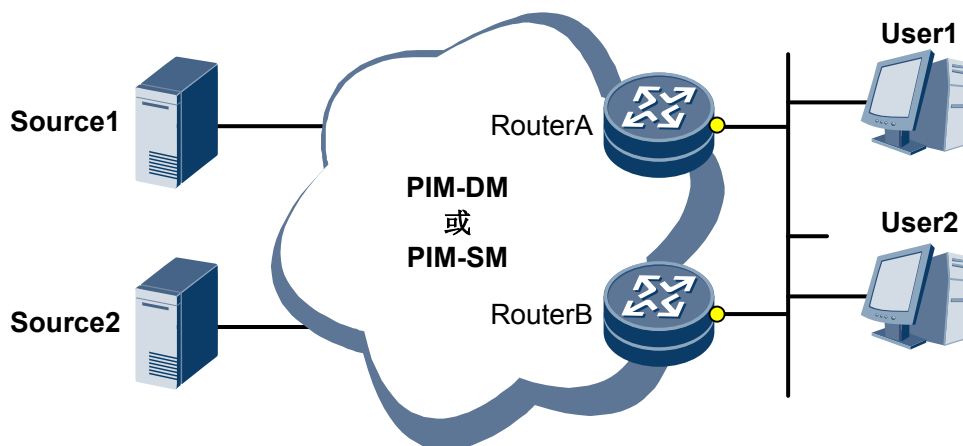
表 12-7 IGMPv3 消息

| 名称                                         | 定义                                                    | 说明                                                                                                                                          |
|--------------------------------------------|-------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| 普遍组查询消息（General-Query）                     | 同 IGMPv1、IGMPv2                                       | 同 IGMPv1、IGMPv2                                                                                                                             |
| 特定组查询消息（Group-Specific Query）              | 同 IGMPv2                                              | 同 IGMPv2                                                                                                                                    |
| 特定组/源查询消息（Group-and-Source-Specific Query） | 特定组查询消息是查询器向共享网段内特定组播组成员发送的消息，用于查询该组成员是否愿意接收特定源发送的数据。 | 封装该消息的 IP 报文头的目的地址字段为被查询的组播组的 IP 地址，本地网段上的所有路由器及属于该组的所有主机都能够识别并接收。<br><br>特定组/源查询消息既携带组地址，还携带一个或多个源地址。                                      |
| IGMPv3 成员报告消息（Report）                      | 成员报告消息是主机向组播路由器发送的报告消息，用报告加入某组播组并只接收由指定组播源发往该组的数据。    | 封装该消息的 IP 报文头的目的地址字段为 224.0.0.22，本地网段上的所有 IGMPv3 路由器都能识别并接收。<br><br>IGMPv3 的 Report 消息可以携带一个或多个组记录。在每个组记录中，包含有组播组地址和组播源地址列表，可以过滤此组播组地址的组播源。 |

12.2.4.6 IGMP 典型应用

## ASM 主机动态接入

图 12-9 ASM 主机动态接入



●与用户主机相连的接口，启用IGMPv1或IGMPv2

在 ASM 组播模型中，接收者无法预先知道组播源的位置，只能接收发往组播组的所有数据。

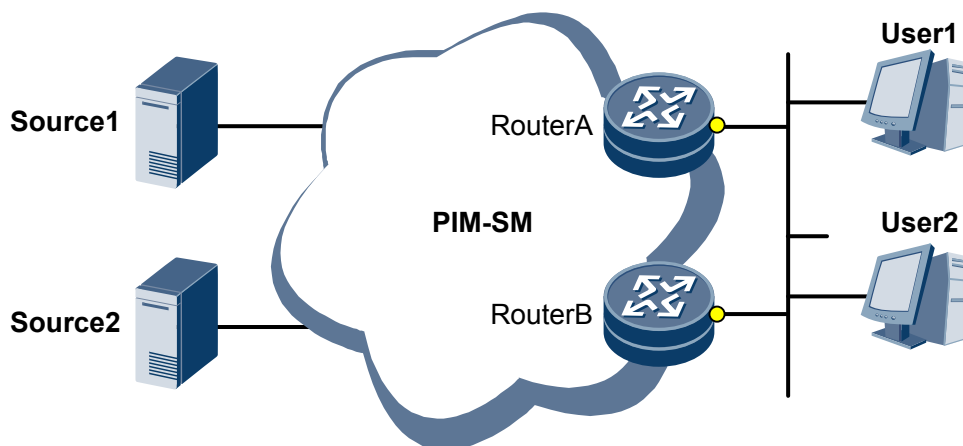
适用 IGMPv1 或 IGMPv2。

实现方案如下：

- 在接收者主机上启用 IGMPv1 或 IGMPv2 协议。
- 在路由器上与接收者主机相连的接口上启用 IGMPv1 或 IGMPv2 协议。
- 网络中路由器之间启用 PIM-DM 或 PIM-SM 协议。

## SSM 主机动态接入

图 12-10 SSM 主机动态接入



●与用户主机相连的接口，启用IGMPv3

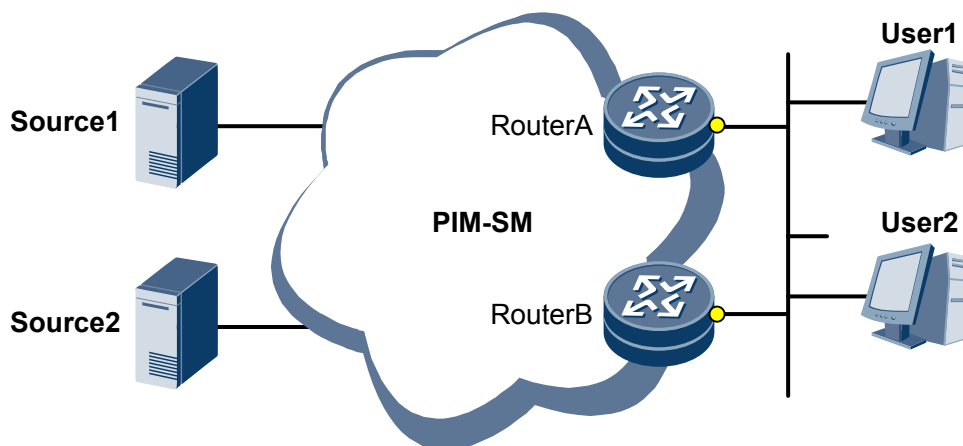
在 SSM 组播模型中，接收者预先知道组播源的位置，可以指定接收具体的源/组数据。  
适用 IGMPv3。

实现方案如下：

- 在接收者主机上启用 IGMPv3 协议。
- 在与接收者主机处于同一网段的路由器接口上启用 IGMPv3 协议。
- 网络中路由器之间启用 PIM-SM 协议。

## SSM Mapping 主机动态接入

图 12-11 SSM Mapping 主机动态接入



●与用户主机相连的接口，启用IGMPv3、SSM Mapping

在网络中运行 SSM，但是用户主机无法升级到 IGMPv3 的情况下，适用 SSM Mapping 功能。在路由器端定制用户需要的源/组对应关系，向用户转发符合要求的组播数据。

主机适用 IGMPv1 或 IGMPv2。路由器接口适用 IGMPv3。

实现方案如下：

- 在接收者主机上启用 IGMPv1 或 IGMPv2 协议。
- 在与接收者主机处于同一网段的路由器接口上启用 IGMPv3 协议，启用 SSM Mapping 功能，配置源/组映射关系。

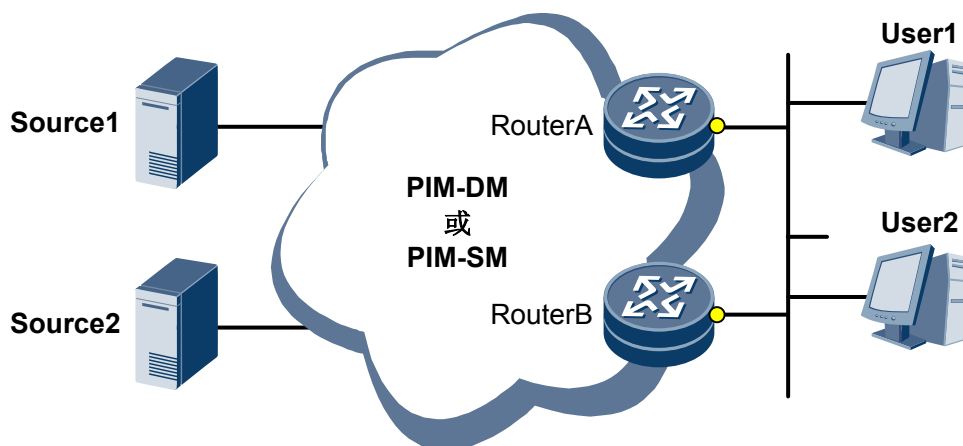
### 说明

只有当同一网络上同时存在 IGMPv3 和 IGMPv1/IGMPv2 主机，才需要配置 IGMPv3，否则使能 SSM Mapping 功能不要求接口必须配置 IGMPv3。

- 网络中路由器之间启用 PIM-SM 协议。

## 静态加入组播组

图 12-12 路由器接口静态加入



- 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

当网段中的组播组成员关系比较稳定时，适用静态加入组播组。静态加入某组播组后，无论网段上是否存在该组成员，路由器将持续转发该组数据。

静态加入组播组与接口配置的 IGMP 版本没有关系。

实现方案如下：

- 在接收者主机上启用 IGMP 协议。
- 在与接收者主机处于同一网段的路由器接口上启用 IGMP 协议，配置静态加入组播组。当使用 IGMPv3 时，可以同时指定组播源。
- 网络中路由器之间启用 PIM 协议。

## 12.3 IGMP Snooping

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。

### 12.3.1 介绍

#### 定义

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。

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。

目的

当二层设备没有运行 IGMP Snooping 时，组播数据在二层被广播。这样不但浪费了网络带宽，而且影响了组网安全。当二层设备运行了 IGMP Snooping 后，已知组播组的组播数据不会在二层被广播，而在二层被组播给指定的接收者。

12.3.2 参考标准和协议

- 与 IGMP Snooping 相关的参考标准和协议如下：
- RFC4541: Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches

12.3.3 可获得性

License 支持

本特性无需 License 支持。

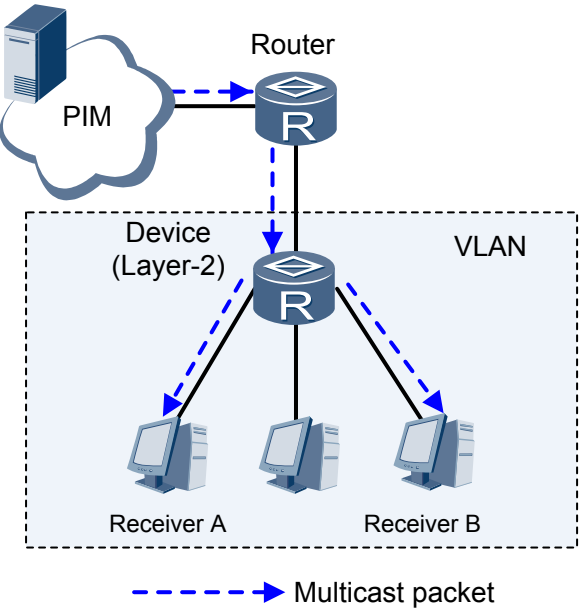
版本支持

| 产品                                                                      | 最低支持版本                                                                            |
|-------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2200/5100 V100R005C00、<br>USG2200/5100 BSR/HSR V100R005C00<br>USG5500 V300R001 |

12.3.4 应用场景

如图 12-13 所示，网络边缘的组播路由器 Router 启用了 IGMP 功能，负责维护组播成员关系。Device 位于组播路由器和用户主机之间，提供 IGMP Snooping 功能，负责在二层网络将组播数据只转发给有需要的接收者（组播成员）。

图 12-13 IGMP Snooping 应用场景



应用 IGMP Snooping 的优点主要有：

- 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。

## 12.3.5 原理描述

### 12.3.5.1 IGMP Snooping 基本原理和概念

#### 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。

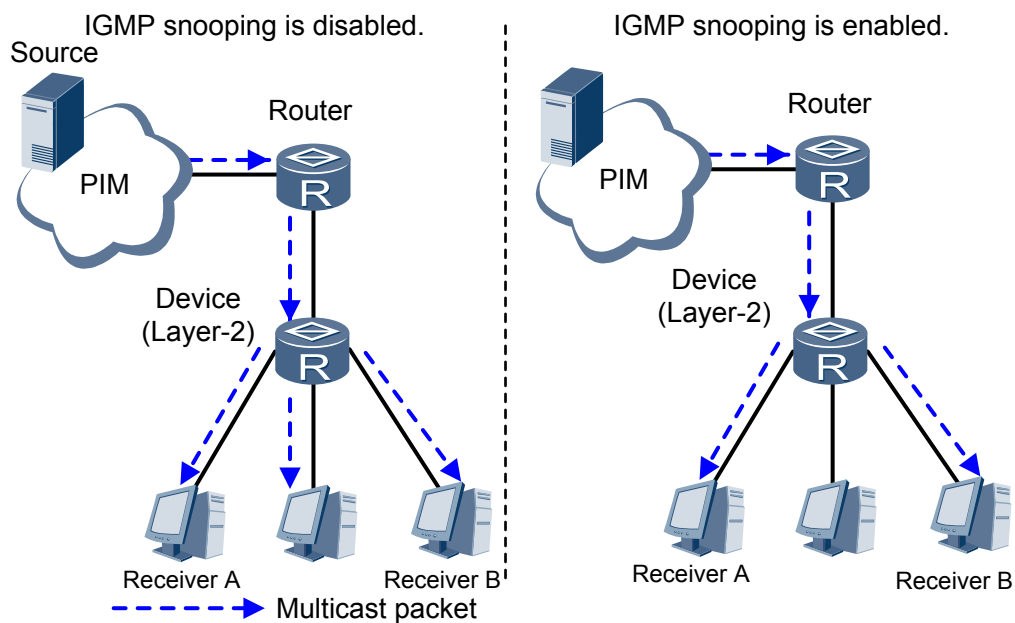
根据 IGMP 协议，主机加入组播组时需要向上游设备发送 IGMP 报文，向其报告加入组播组消息，这样上游设备才可以将组播报文发送给主机。由于 IGMP 报文封装在 IP 报文内，属于三层协议报文，而二层设备不处理报文的三层信息，所以这个过程二层设备并不知道，而且通过对二层数据帧的源 MAC 地址的学习也学不到组播 MAC 地址（数据帧的源 MAC 地址不是组播 MAC 地址）。这样当二层设备在接收到一个目的 MAC 地址为组播 MAC 地址的数据帧时，在以前学习的 MAC 地址表中就不会找到对应的表项。那么它就会采用广播方式发送接收到的组播报文，这不但会造成带宽的极大浪费，而且影响网络安全。

IGMP Snooping 是实现数据链路层组播的一种解决方案。当运行 IGMP Snooping 的设备收到主机和上游设备之间传递的 IGMP 报文时，设备的 IGMP Snooping 模块分析报文携带的信息，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，MAC 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。

如图 12-14 所示，当二层设备没有运行 IGMP Snooping 时，组播数据在二层被广播；当二层设备运行 IGMP Snooping 后，已知组播组的组播数据不会在二层被广播，而在二层被组播给指定的接收者。



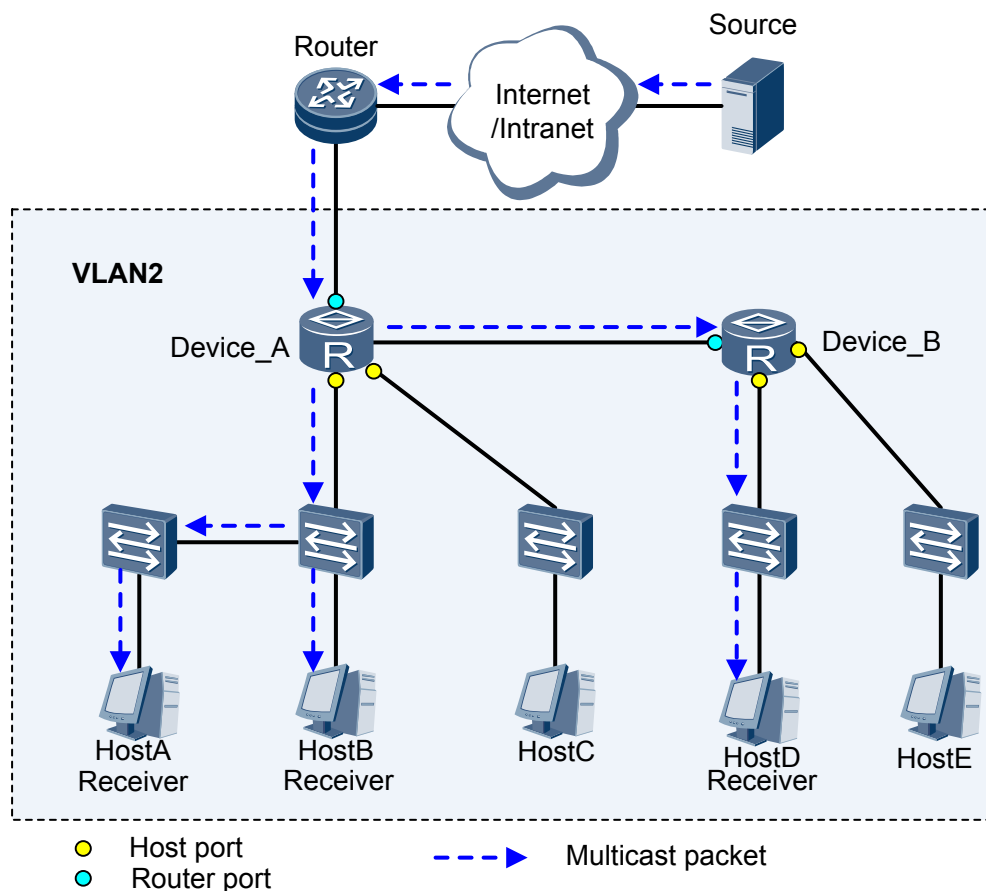
图 12-14 二层设备运行 IGMP Snooping 前后的对比



## IGMP Snooping 相关端口

如图 12-15 所示，Router 连接组播源，在 Device\_A 和 Device\_B 上分别运行 IGMP Snooping，HostA、HostB 和 HostD 为接收者主机（即组播组成员）。

图 12-15 IGMP Snooping 相关端口



结合图 12-15，介绍一下 IGMP Snooping 相关的端口概念：

- 路由器端口（Router Port）：二层设备上朝向三层组播设备（DR 或 IGMP 查询器）一侧的端口，如图 12-15 中所示接口。

路由器端口分为如下两类：

- 动态路由器端口：能够接收到源地址不为 0.0.0.0 的 IGMP 普遍组查询（IGMP Query）报文或 PIM Hello 报文的端口。设备为其每个动态路由器端口都启动一个路由器端口老化定时器，其超时时间就是动态路由器端口老化时间。在定时器超时前如果没有收到源地址不为 0.0.0.0 的 IGMP 普遍组查询（IGMP Query）报文或 PIM Hello 报文，则该路由器端口失效；如果收到则重置定时器。
- 静态路由器端口：用户使用配置命令指定的路由器端口，不会老化。

- 成员端口（Member Port）：二层设备上朝向组播组成员一侧的端口，如图 12-15 中所示接口。

成员端口分为如下两类：

- 动态成员端口：能够接收到 IGMP 成员关系报告（IGMP Report）报文的端口。当一个端口动态加入某组播组时，设备为该端口启动一个成员端口老化定时器，其超时时间就是动态成员端口老化时间。在定时器超时前如果没有收到 IGMP 成员关系报告（IGMP Report）报文，则该成员端口失效；如果收到则重置定时器。
- 静态成员端口：用户使用配置命令指定的成员端口，不会老化。

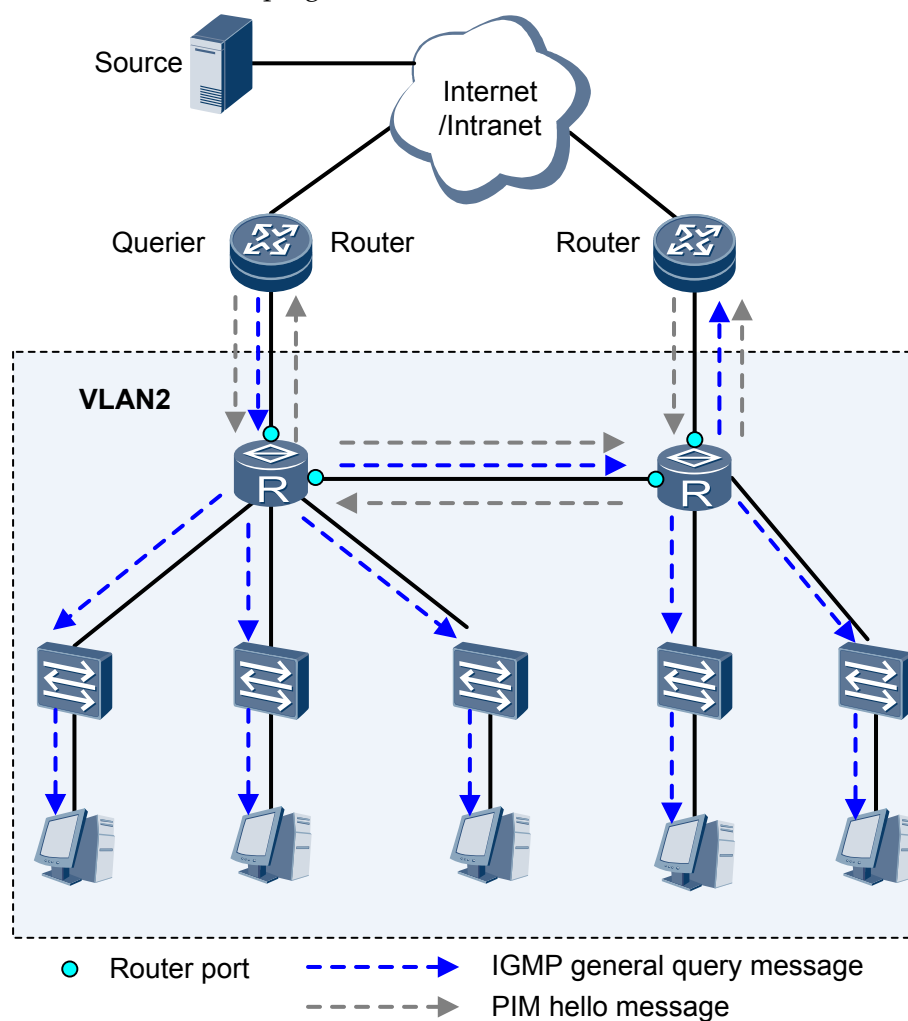
### 12.3.5.2 IGMP Snooping 实现机制

运行了 IGMP Snooping 的设备对不同 IGMP 动作的具体处理方式如下：

#### 普遍组查询

如图 12-16 所示，IGMP 查询器（Querier）定期向本地网段内的所有主机与路由器（224.0.0.1）发送 IGMP 普遍组查询（IGMP Query）报文，以查询该网段有哪些组播组的成员。

图 12-16 IGMP Snooping 处理 IGMP 普遍组查询报文和 PIM Hello 报文



在收到 IGMP 普遍组查询报文时，设备将其通过 VLAN 内除接收端口以外的其它所有端口转发出去，并对该报文的接收端口做如下处理：

- 如果该端口是组播组对应的路由器端口，则刷新该路由器端口的老化时间。
- 如果该端口不是组播组对应的路由器端口，则该端口作为路由器端口生效，同时启动该路由器端口的老化定时器。

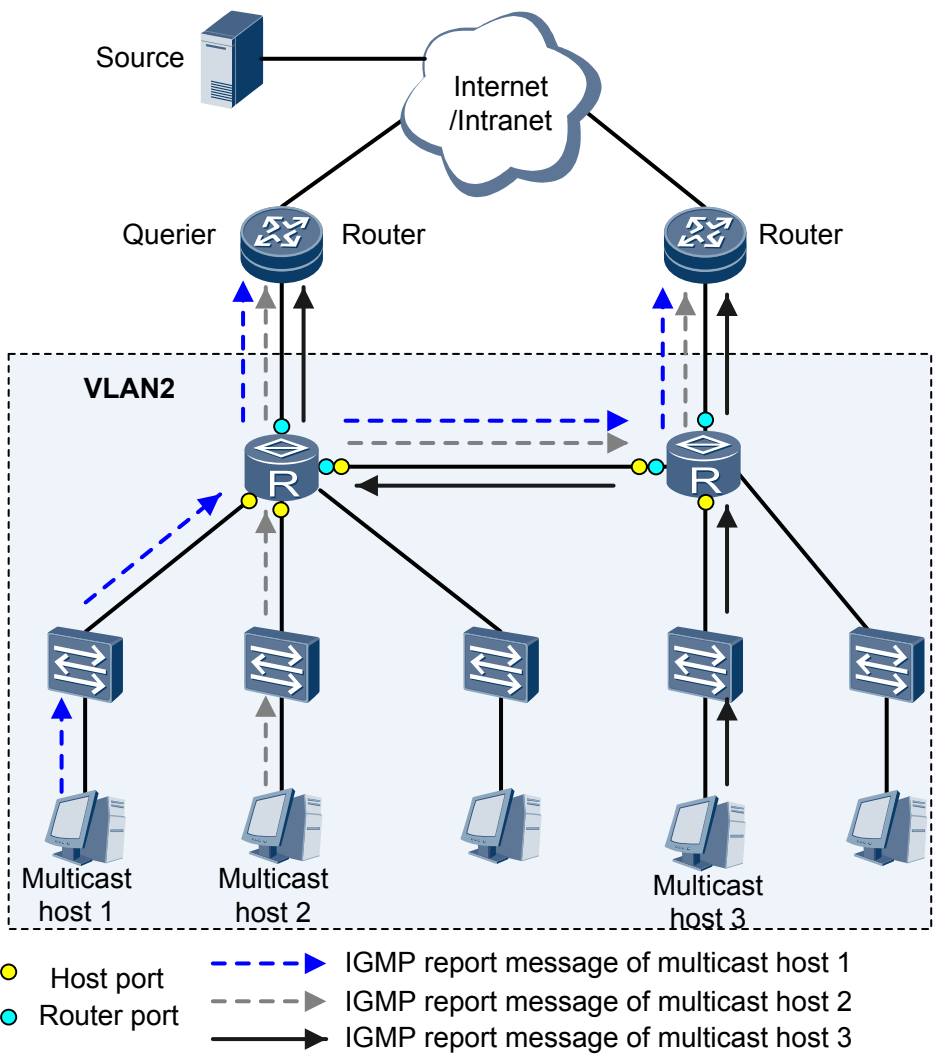
如图 12-16 所示，与用户网段相连的路由器（Router）会定期发送 PIM Hello 报文，以建立 PIM 邻居，竞选 DR。

当设备接收到 PIM Hello 报文时，向 VLAN 内的其它路由器端口转发。设备对接收到 PIM Hello 报文的端口的处理方式与接收到 IGMP 普遍组查询报文时相同。

报告成员关系

如图 12-17 所示，当用户主机需要加入某个组播组时，或组播组的成员主机接收到 IGMP 普遍组查询报文时，会向 IGMP 查询器（Querier）发送 IGMP 成员关系报告（IGMP Report）报文。

图 12-17 IGMP Snooping 处理 IGMP 成员关系报告报文



在收到 IGMP 成员关系报告报文时，设备将其通过 VLAN 内的所有路由器端口转发出去，从该报文中解析出主机要加入的组播组地址，并对该报文的接收端口做如下处理：

- 如果接收端口是该组播组对应的成员端口，则刷新该成员端口的老化时间。
- 如果接收端口不是该组播组对应的成员端口，则该端口作为成员端口生效，同时启动该成员端口的老化定时器。

## 离开组播组

运行 IGMPv1 的主机离开组播组时不会发送 IGMP 离开组报文，因此设备无法立即获知主机离开的信息。但是，由于主机离开组播组后不会再发送 IGMP 成员关系报告报文，因此当其对应的动态成员端口的老化定时器超时后，设备就会将该端口对应的转发表项从转发表中删除。

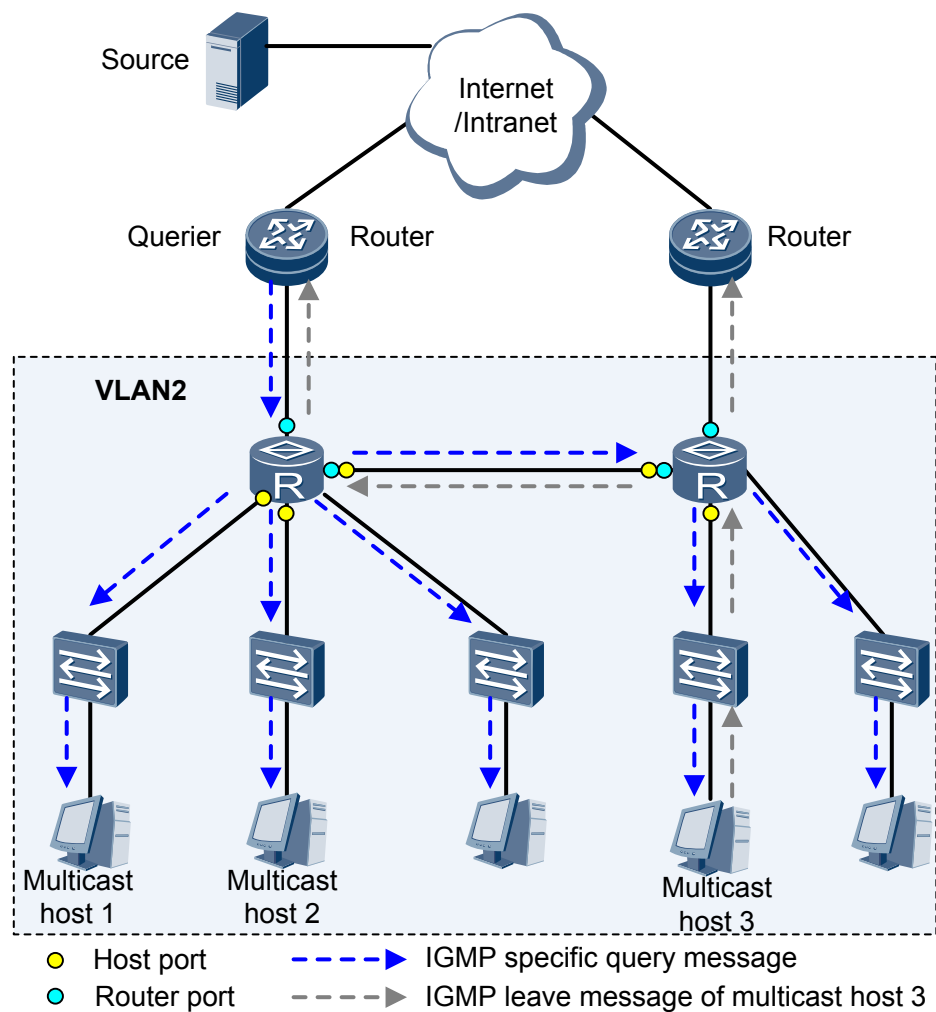
如图 12-18 所示，运行 IGMPv2 的主机离开组播组时，会通过发送 IGMP 离开组报文，以通知组播路由器（Router）自己离开了某个组播组。当设备从某动态成员端口上收到 IGMP 离开组报文时，首先判断要离开的组播组所对应的转发表项是否存在：

- 如果不存在该组播组对应的转发表项，设备不会向任何端口转发该报文，而将其直接丢弃。
- 如果存在该组播组对应的转发表项，设备会将该报文通过 VLAN 内的所有路由器端口转发出去。同时，由于并不知道该接收端口下是否还有该组播组的其它成员，所以设备不会立刻把该端口从该组播组所对应转发表项中删除，而是重置该成员端口的老化定时器。

当 IGMP 查询器收到 IGMP 离开组报文后，从中解析出主机要离开的组播组的地址，并通过接收端口向该组播组发送 IGMP 特定组查询报文。设备在收到 IGMP 特定组查询报文后，将其通过 VLAN 内的所有路由器端口和该组播组的所有成员端口转发出去。对于 IGMP 离开组报文的接收端口（假定为动态成员端口），设备在该端口的老化时间内：

- 如果从该端口收到了主机响应该特定组查询的 IGMP 成员关系报告报文，则表示该端口下还有该组播组的成员，于是重置该端口的老化定时器。
- 如果没有从该端口收到主机响应特定组查询的 IGMP 成员关系报告报文，则表示该端口下已没有该组播组的成员，则在其老化时间超时后则该成员端口自动失效。

图 12-18 IGMP Snooping 处理离开组播组报文和 IGMP 特定组查询报文



快速离开

端口快速离开是指当交换机从某端口收到主机发送的离开某组播组的 IGMP 离开组报文时，直接使该成员端口失效。此后，当设备收到对该组播组的 IGMP 特定组查询报文时，交换机将不再向该端口转发。

只有当 VLAN 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 内配置允许成员端口快速离开。否则，当端口下有多台接收者主机时，一个主机离开，可能会造成同一端口下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，在这种场景中成员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。

端口快速离开有以下优点：

- 减小响应延迟
- 节省因各种消息而占用的网络带宽

12.3.5.3 IGMP Proxy

## IGMP Proxy 的作用

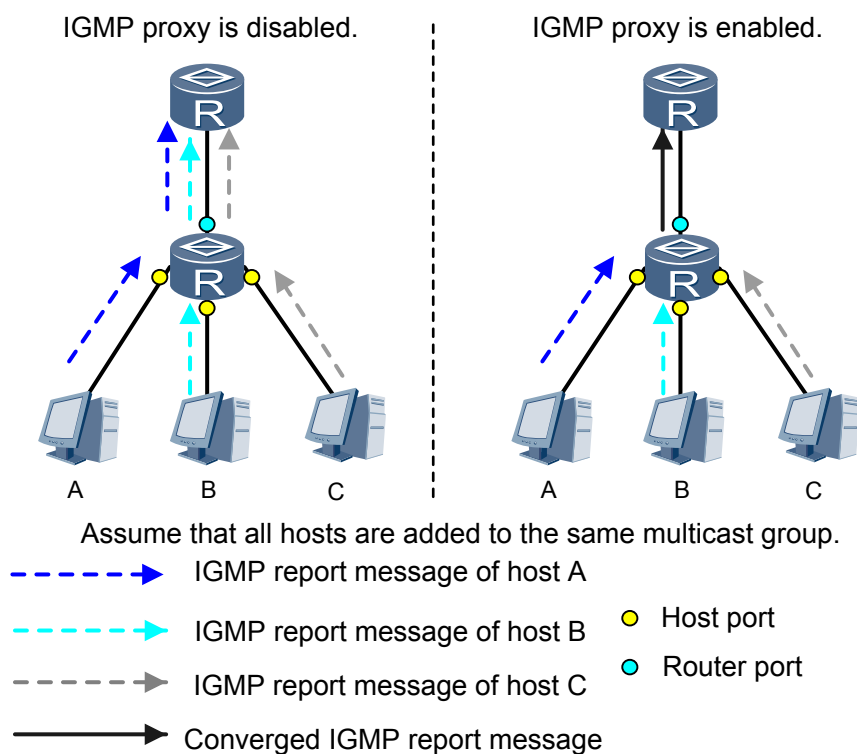
IGMP Proxy 运行在二层交换设备上，基于 IGMP snooping 实现，依赖 IGMP snooping 实现组播数据转发，拦截并汇聚上下游设备发送的组播协议报文，以降低上游三层组播网络的负担。

### 拦截并汇聚 IGMP 成员关系报告报文

如图 12-19 所示，二层设备配置了 IGMP Snooping 功能之后，将侦听下游主机发送的 IGMP Report 报文，建立并维护组成员关系。若设备同时启用了 IGMP Proxy，则设备将拦截并分析 IGMP Report 报文，将分析后的组播组信息汇聚并缓存，然后丢弃该报文。设备会根据缓存的组播组信息是否变化，决定是否向路由端口发送 IGMP Report 报文。

- 若该 IGMP Report 报文为 VLAN 内首个加入此组播组的请求，则需要向路由端口发送此组播组的 IGMP Report 报文。
- 若 VLAN 内已经存在加入该组播组的成员，则只记录信息到缓存中，不发送 IGMP Report 报文。
- 该 VLAN 内后续加入同一组播组的 IGMP Report 报文都会被抑制，不会向路由端口发送 IGMP Report 报文。

图 12-19 拦截并汇聚 IGMP 成员关系报告报文



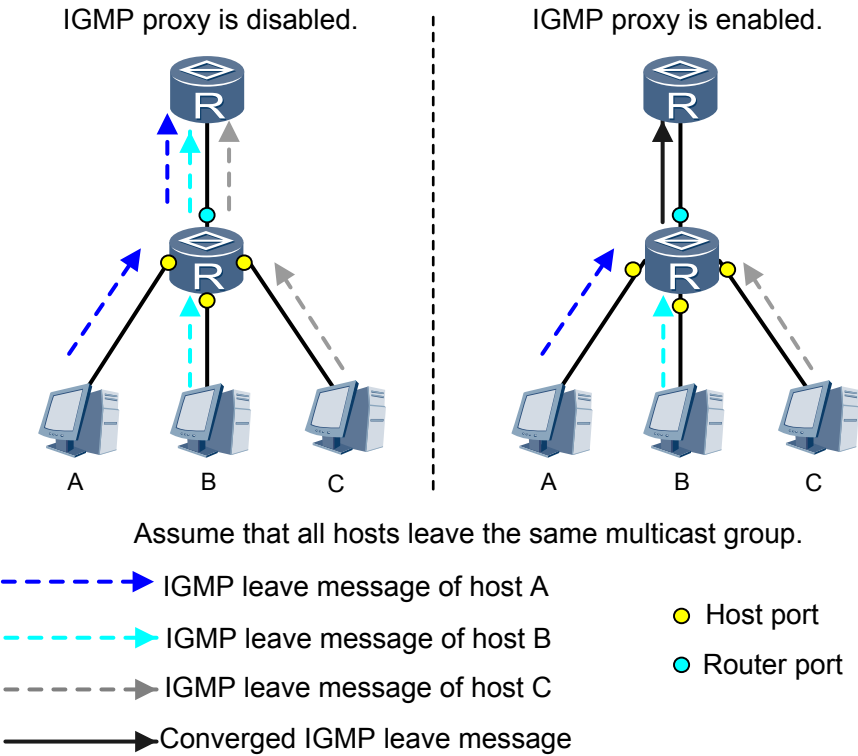
### 拦截并汇聚 IGMP 离开组报文

如图 12-20 所示，二层设备配置了 IGMP Snooping 功能之后，将侦听下游主机发送的 IGMP Leave 报文，建立并维护组成员关系。若设备同时启用了 IGMP Proxy，则设备将拦截并分析 IGMP Leave 报文，将分析后的组播组信息汇聚并缓存，然后丢弃该报文。

设备会根据缓存的组播组信息判定 VLAN 内是否还存在该组播组的成员，以决定是否向路由端口发送 IGMP Leave 报文。

- 若 VLAN 内存在该组播组成员，则设备不会向路由器端口发送该组播组的 IGMP Leave 报文。
- 若 VLAN 内不存在该组播组成员，则设备向路由器端口发送该组播组的 IGMP Leave 报文。

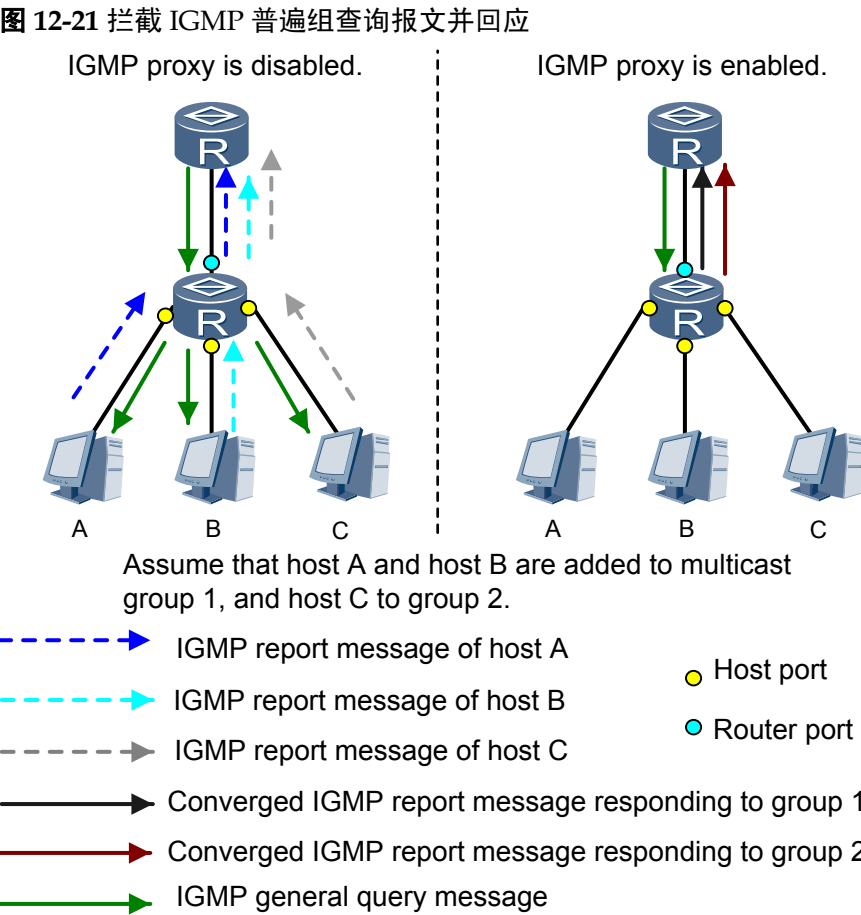
图 12-20 拦截并汇聚 IGMP 离开组报文



### 拦截 IGMP 普遍组查询报文并回应

如图 12-21 所示，二层设备配置了 IGMP Snooping 功能之后，将侦听上游设备发送的 IGMP General Query 报文，建立并维护路由端口。若设备同时启用了 IGMP Proxy，则设备会拦截并丢弃 IGMP General Query 报文。同时设备会查找该 VLAN 内缓存的所有组播组信息，针对每个组播组，都构造 IGMP Report 报文回应 IGMP General Query 报文。

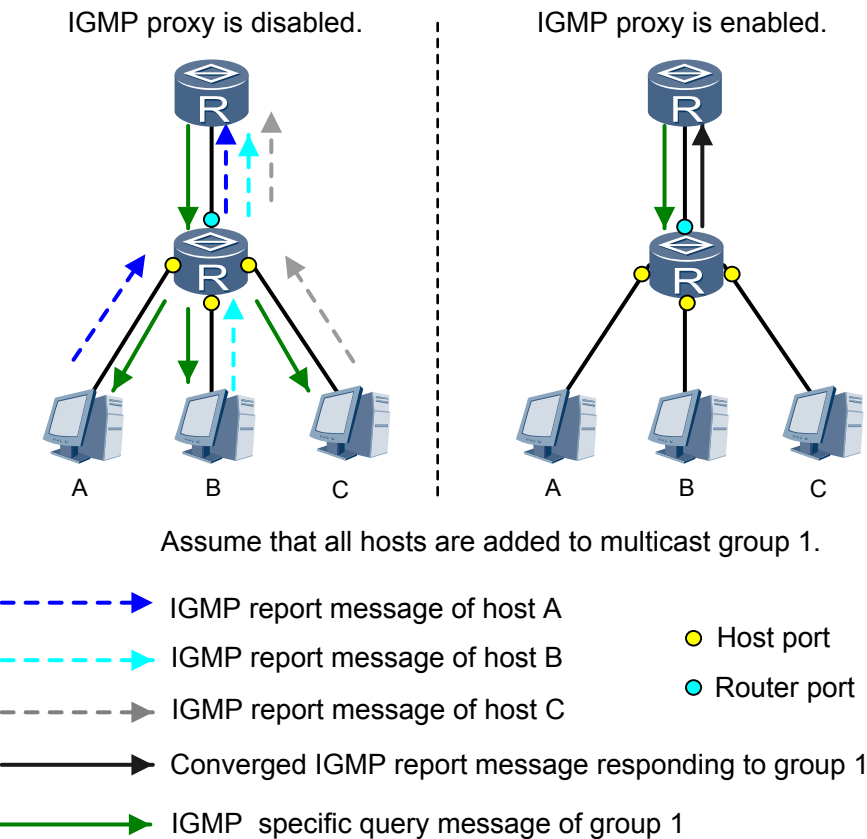




拦截 IGMP 特定组查询报文并回应

如图 12-22 所示，二层设备配置了 IGMP Snooping 功能之后，将侦听上游设备发送的 IGMP Specific Query 报文，建立并维护路由端口。若设备同时启用了 IGMP Proxy，则设备会拦截并丢弃 IGMP Specific Query 报文。同时设备会在缓存的组播组信息中查找，若存在此特定组，则构造特定组的 IGMP Report 报文回应。

图 12-22 拦截 IGMP 特定组查询报文并回应



12.4 PIM

PIM 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。

12.4.1 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。

PIM 不维护专门的单播路由，也不依赖某具体的单播路由协议，而是直接利用单播路由的结果。为 PIM 提供单播路由信息的可以是静态路由、RIP、OSPF、IS-IS、BGP 等任何一种单播路由协议。

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：

- 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 协议利用现有的单播路由结果实现组播报文的复制和转发。

## 12.4.2 参考标准和协议

与 PIM 相关的参考标准和协议如下：

- RFC4601: Protocol Independent Multicast - Sparse Mode (PIM-SM)
- draft-ietf-pim-sm-bsr-05: Bootstrap Router (BSR) Mechanism for PIM
- RFC3973: Protocol Independent Multicast - Dense Mode
- draft-ietf-ssm-arch-01: Source-Specific Multicast for IP
- RFC3569: An Overview of Source-Specific Multicast (SSM)

## 12.4.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 12.4.4 原理描述

### 12.4.4.1 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 路由器组成的网络称为 PIM 网络。

通过在路由器接口上设置“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。

## 第一跳路由器

与组播源直接相连且负责转发该组播源发出的组播数据的 PIM 路由器。

## 最后一跳路由器

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。

## DR

Designated Router，称为指定路由器。只在 PIM-SM 中使用。在连接组播源 S 的共享网段，由 DR 负责向 RP 发送注册（Register）消息；在连接组成员的共享网段，由 DR 负责向 RP 发送加入（Join）消息。

## RP

Rendezvous Point，称为汇聚点。只在 PIM-SM 中使用。RP 是 PIM-SM 网络的转发核心。组成员向 RP 发起加入，构建以 RP 为根的 RPT；组播源向 RP 注册，在 RP 上创建（S，G）表项，通过 RP 向组成员传输组播报文。

## 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 路由器。路由器根据 RP-Set 计算出特定组播组对应的 RP。

## RPF

Reverse Path Forwarding, 称为反向路径转发, 是组播路由的基础。路由器对接收到的组播报文执行 RPF 检查, 从而创建并维护组播路由表项, 确保组播数据沿正确的路径转发。路由器接收到组播报文后, 依据“报文源”, 查找单播路由表、MBGP 路由表和组播静态路由表, 选定 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, 转发该报文; 否则 RPF 检查失败, 丢弃该报文。

### 12.4.4.2 PIM-DM

PIM-DM (Protocol Independent Multicast Dense Mode) 协议无关组播一密集模式, 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。

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)。

PIM-DM 的工作过程可以概括为: 邻居发现、扩散、剪枝、嫁接、断言。下面将分别具体介绍。

#### 邻居发现

在 PIM-DM 网络中, 组播路由器需要使用 Hello 消息来发现邻居, 并维护邻居关系。通过比较 Hello 消息上携带的优先级和 IP 地址, 各路由器为多路由器网段选举指定路由器 DR, 由 DR 来充当 IGMPv1 的查询器。通过各路由器之间周期性地使用 Hello 消息保持联系, 从而构建和维护了 SPT 树。

#### 扩散 (Flooding)

PIM-DM 假设网络上的所有主机都准备接收组播数据。

当路由器接收到组播报文后, 首先根据单播路由表进行 RPF 检查, 如果检查通过则创建一个 (S, G) 表项, 然后将数据向网络上所有下游 PIM-DM 节点转发, 这个过程称为扩散 (Flooding)。经过这个过程, PIM-DM 组播域内每个路由器上都会创建 (S, G) 表项。

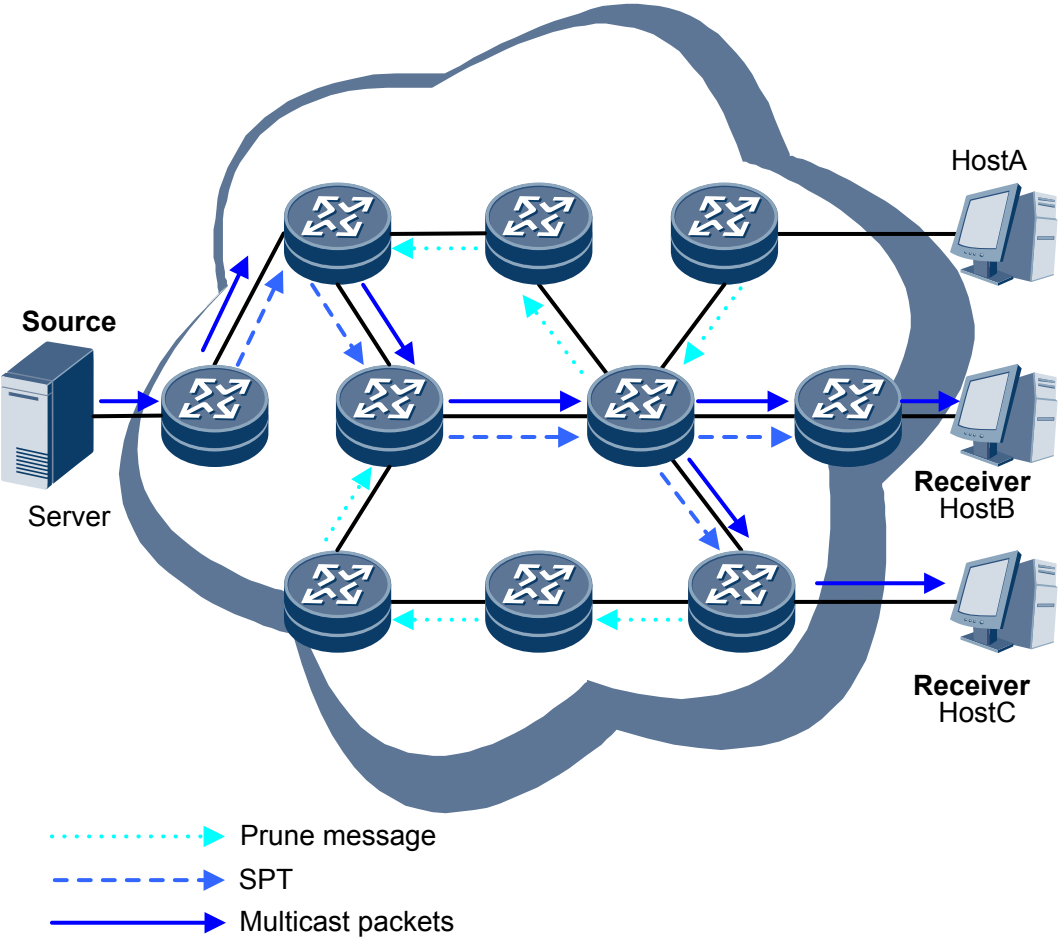
如果没有通过 RPF 检查, 则将报文丢弃。

#### 剪枝 (Prune)

如果下游节点没有组播组成员, 则向上游节点发 Prune 剪枝消息, 通知上游节点不用再转发数据到该分支。上游节点收到 Prune 剪枝消息后, 就将相应的接口从其组播转发表项 (S, G) 对应的出接口列表中删除。剪枝过程继续直到 PIM-DM 中仅剩下了必要的分支, 这就建立了一个以组播源 S 为根的 SPT (最短路径树)。

剪枝过程最先由叶子路由器发起, 如图 12-23 所示。

图 12-23 PIM-DM 中 SPT 建立示意图



各个被剪枝的节点提供超时机制，当剪枝超时重新开始扩散-剪枝过程。PIM-DM 的扩散-剪枝机制周期性进行。

说明

对于图 12-23 中的中间组播设备，当 RPF 检查失败后会立即向上游设备发起剪枝。

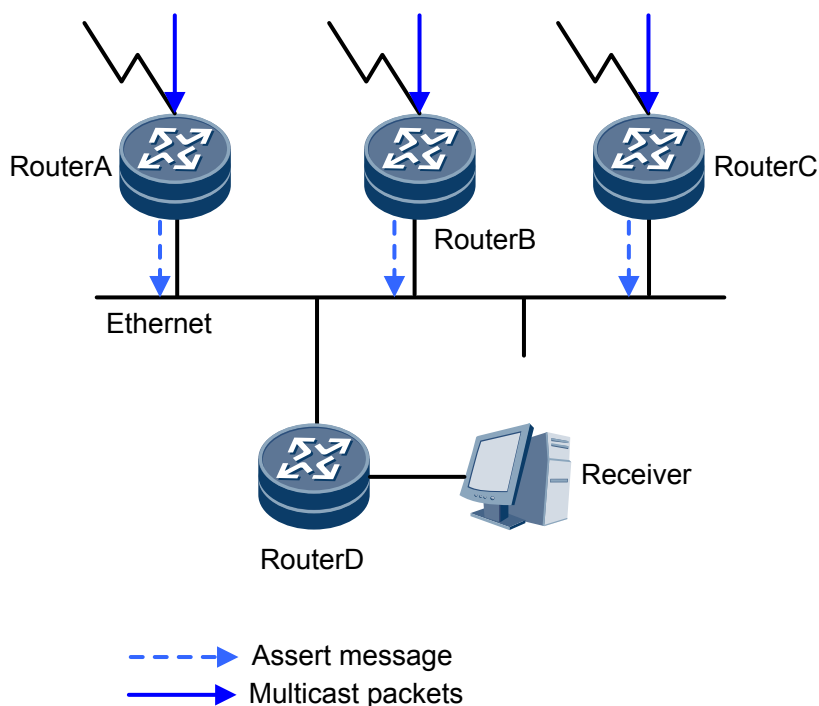
嫁接（Graft）

当被剪枝的下游节点需要恢复到转发状态时，该节点使用 Graft 嫁接消息通知上游节点。例如图 12-23 中的 HostA 恢复接收组播数据，则 Graft 嫁接消息逐跳向组播源 S 传递，中间节点接收到 Graft 嫁接消息后回应确认，从而先前被剪掉的分支恢复信息传输。

断言（Assert）

在共享网络（如 Ethernet）中会出现相同报文的重复发送。例如 LAN 网段上存在多台组播路由器 A、B、C，A、B 和 C 各自都能收到组播源 S 发送的报文。如图 12-24 所示。

图 12-24 Assert 机制示意图



当路由器 A、B 和 C 都从上游接收到组播源 S 发出的组播数据报文后，都会向 Ethernet 网络上转发该组播报文，这样下游节点组播路由器 D 就会收到三份完全相同的组播报文。为了避免这种情况，就需要通过 Assert 机制来选定一个唯一的转发者。

如图 12-24 所示，路由器 A、B、C 会分别从下游接口向 Ethernet 发送 Assert 消息。Assert 消息中携带：该 PIM 路由器到组播源 S 的开销、所采用的单播路由协议的优先级、组播组地址 G。各路由器（A、B、C）在收到 Assert 消息后，会将自身条件与对方 Assert 消息中携带的信息进行比较，这个过程称为 Assert 竞选。

其具体规则如下：

- 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下游接口，只传输一份组播报文。所有 Assert Loser 可以周期性地恢复组播数据转发，从而引发周期性的 Assert 竞选。

#### 12.4.4.3 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。

PIM-SM 转发组播数据的关键是建立 RPT 和 SPT。

PIM-SM 的工作过程主要有：邻居发现、DR 选举、RP 发现、BSR 选举、构建 RPT、组播源注册、SPT 切换、断言。下面将分别具体介绍。

## 邻居发现

在 PIM-SM 网络中，组播路由器需要使用 Hello 消息来发现邻居，并维护邻居关系，协商协议参数。

通过比较 Hello 消息上携带的优先级和 IP 地址，各路由器为多路由器网段选举指定路由器 DR。

说明

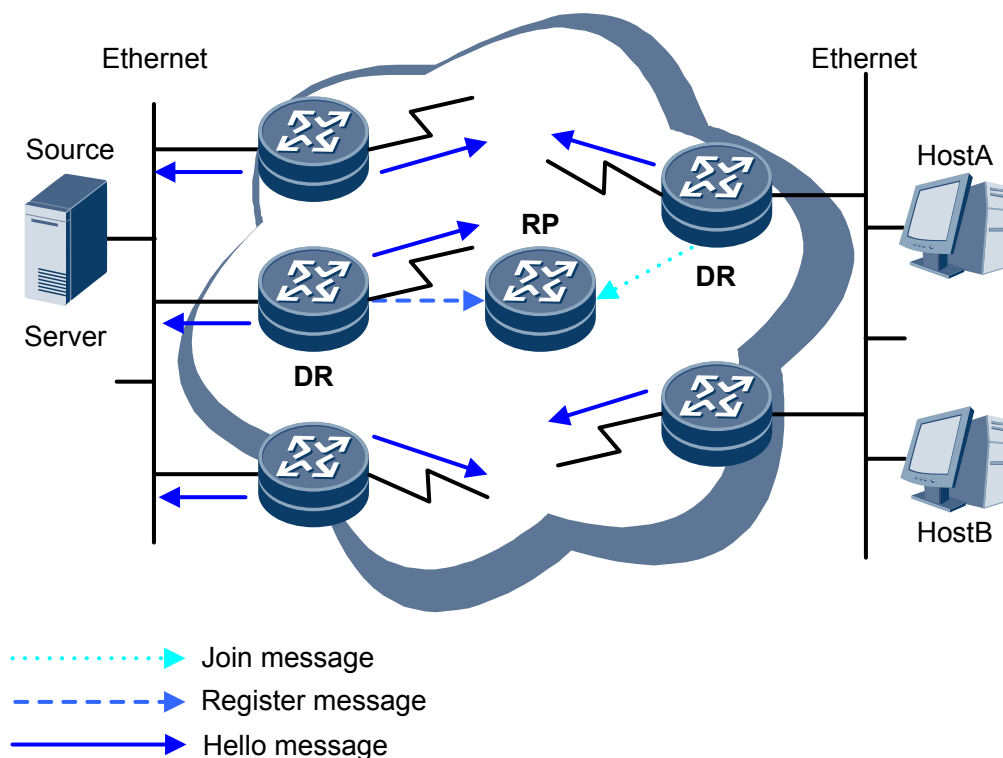
如果用户主机网段运行 IGMPv1，则由 DR 充当查询器。

## DR 选举

借助 Hello 消息可以为共享网络（如 Ethernet）选举 DR，DR 将作为本网段中组播信息的唯一转发者。

无论是和组播源 S 连接的网络，还是和接收者连接的网络，只要网络为共享网络则需要选举 DR，接收者侧 DR 向 RP 发送 Join 加入消息；组播源侧 DR 向 RP 发送 Register 注册消息。

图 12-25 DR 选举示意图





如图 12-25 所示，DR 选举过程如下：

1. 共享媒介网络上的各路由器相互之间发送 Hello 消息（携带 DR 优先级选项），拥有最高 DR 优先级路由器将被选举为本网络中的 DR。
2. 如果优先级相同或网络中至少有一台路由器不支持在 Hello 报文中携带优先级，则拥有最大 IP 地址的路由器被选举为 DR。

当 DR 出现故障时，接收 Hello 消息将会超时，邻居路由器之间会触发新的 DR 选举过程。

#### 说明

选举出的 DR 对于 PIM-SM 有实际的意义；而对于 PIM-DM 来说，其本身其实并不需要 DR，但如果 PIM-DM 域中的共享网络上运行了 IGMPv1，则需要选举出 DR 来充当共享网络上的 IGMPv1 查询器。

## RP 发现

PIM-SM 实现组播转发的核心任务就是构造并维护 RPT，RPT 选择 PIM 域中某台路由器作为公用的根节点 RP，组播数据通过 RP 沿着 RPT 转发给接收者。RP 是 PIM-SM 域中的核心路由器。在小型并且简单的网络中，组播信息量少，全网络仅依靠一个 RP 进行信息转发即可，此时可以在 SM 域中各路由器上静态指定 RP 位置。

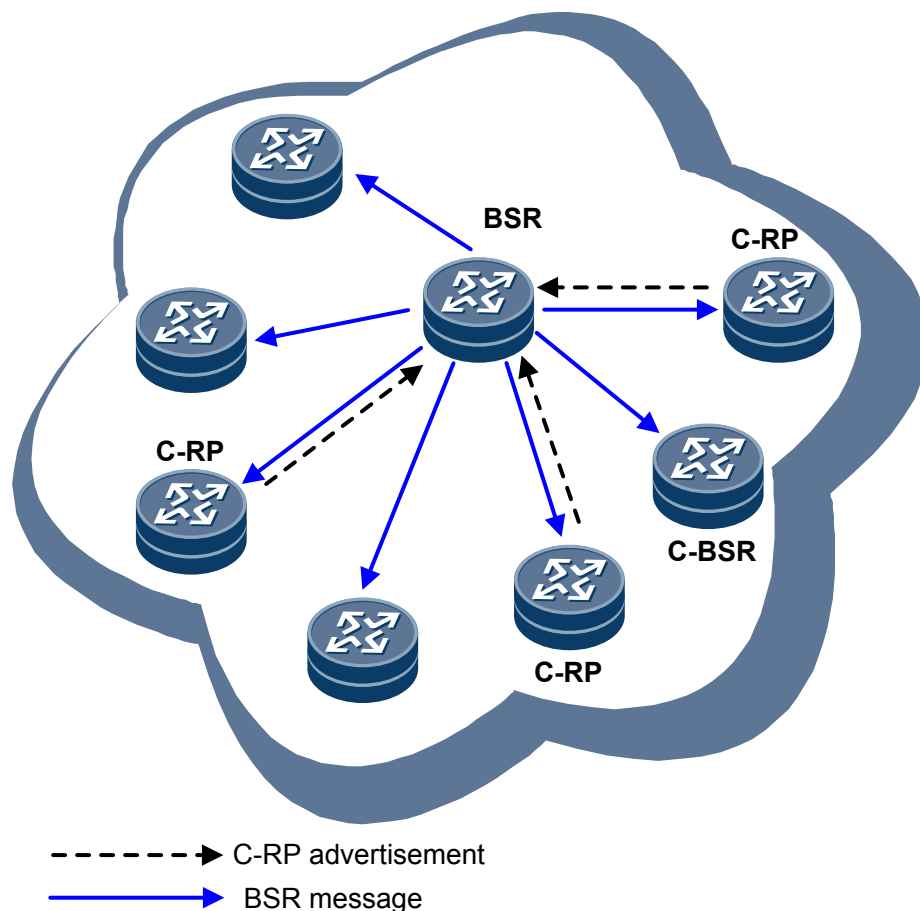
但是在更多的情况下，PIM-SM 网络规模都很大，通过 RP 转发的组播信息量巨大。为了缓解 RP 的负担同时优化共享树的拓扑结构，可以配置多个 C-RP（Candidate-RP），使不同组播组应该对应不同的 RP。此时需要在网络中配置自举路由器 BSR（Bootstrap Router），从而应用自举机制从 C-RP 中动态选举 RP。

BSR 是 PIM-SM 网络里的管理核心，负责收集网络中 C-RP 发来的 Advertisement 信息，消息中携带 C-RP 地址、服务的组范围和 C-RP 优先级。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，从而网络内的所有路由器（包括 DR）都会知道所有 C-RP 的相关信息，并计算出特定组播组范围所对应的 RP。

一个网络（或某管理域）内部只能选举出一个 BSR，但可以配置多个 C-BSR（Candidate-BSR）。这样，一旦 BSR 发生故障后，其余 C-BSR 能够通过自动选举产生新的 BSR，从而确保业务免受中断。

RP 和 BSR 在网络中的位置如图 12-26 所示。

图 12-26 RP 与 BSR 信息交互示意图



网络中的各路由器将依据 RP-Set 提供的信息，从对应某个组播组的多个 C-RP 中选出该组的 RP，具体规则如下：

1. 首先比较 C-RP 的优先级，优先级较高者获胜。
2. 若优先级相同，则使用哈希（Hash）函数计算哈希值，该值较大者获胜。
3. 若优先级和哈希值都相同，则 C-RP 的 IP 地址较大者获胜。

## BSR 选举

BSR 是从众多 C-BSR 中竞选产生的，负责收集并发布网络中的 C-RP 信息。在 PIM-SM 路由器上配置 C-BSR 的同时，指定 C-BSR 所在接口（该接口地址作为 C-BSR 地址），Hash 掩码长度、C-BSR 的优先级。

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了所有 C-BSR 发出的 Bootstrap 消息。通过比较这些 C-BSR 信息，从中竞选产生 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。

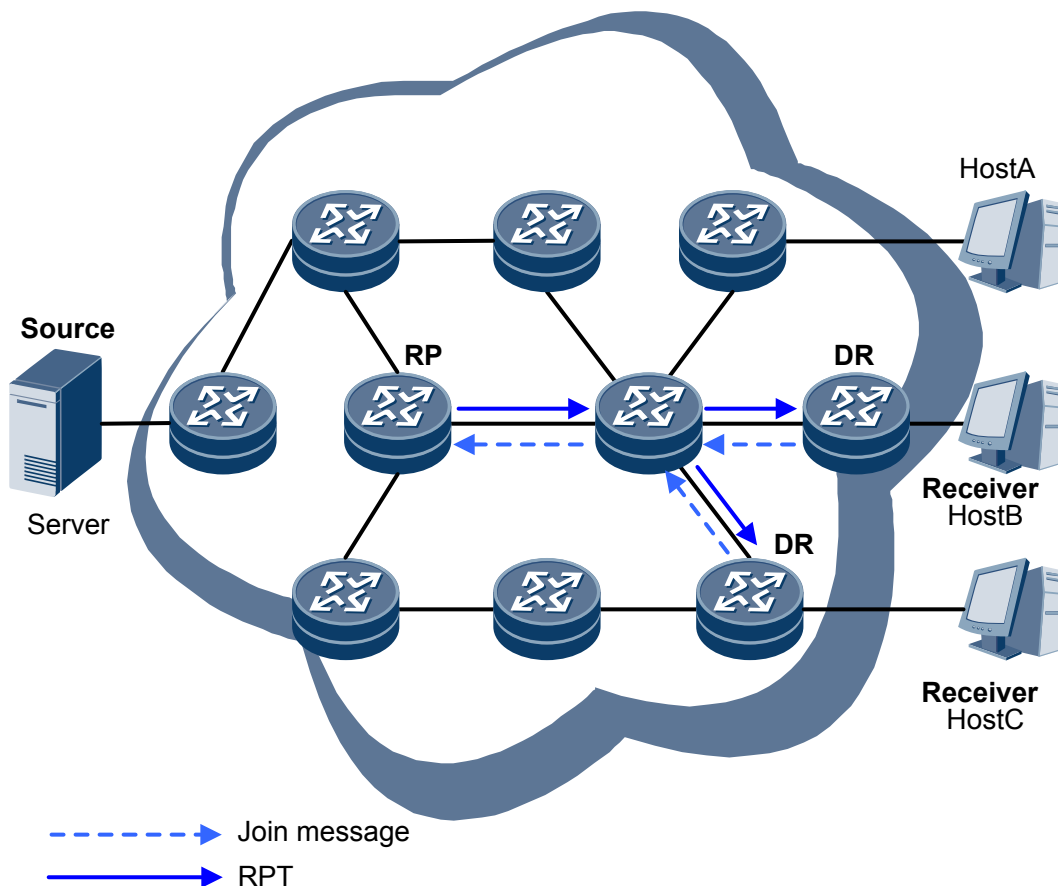
由于所有路由器使用相同的竞选规则，所以得到的当选 BSR 也相同。

根据 BSR 竞选结果，C-BSR 将执行不同的操作：

- 如果获胜，则周期性地向网络发送 Bootstrap 消息，消息中携带自己的 IP 地址和 RP-Set 信息。
- 如果落败，则抑制 Bootstrap 消息的发送。
- 一旦当选 BSR 发生故障，落败 C-BSR 自发执行竞选产生新的 BSR，从而确保业务免受中断。

## 构建 RPT

图 12-27 PIM-SM 中构建 RPT 示意图



如图 12-27 所示，RPT 的构建过程如下：

1. 当接收者加入一个组播组 G 时，先通过 IGMP 报文通知与其直连的 DR。
2. DR 掌握了组播组 G 的接收者的信息后，向该组所对应的 RP 方向逐跳发送加入消息。
3. 从 DR 到 RP 所经过的路由器就形成了 RPT 的分支，这些路由器都在其转发表中生成了 (\*, G) 表项，这里的 “\*” 表示来自任意组播源。RPT 以 RP 为根，以 DR 为叶子。

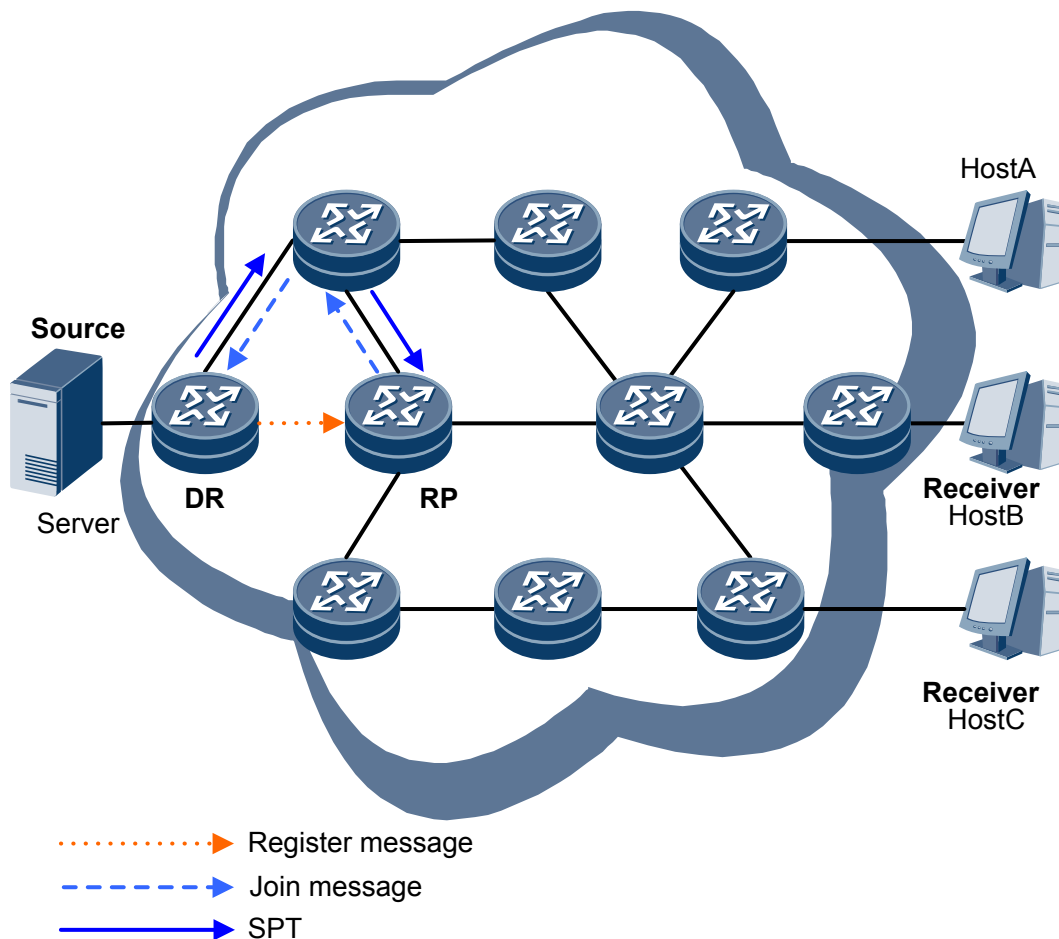
当发往组播组 G 的组播数据流经 RP 时，数据就会沿着已建立好的 RPT 到达 DR，进而到达接收者。

当某接收者对组播组 G 的信息不再感兴趣时，与其直连的 DR 会逆着 RPT 向该组的 RP 方向逐跳发送剪枝报文；上游节点收到该报文后在其出接口列表中删除与下游节点相连的接口，并检查自己是否拥有该组播组的接收者，如果没有则继续向其上游转发该剪枝报文。

## 组播源注册

组播源注册的目的是向 RP 通知组播源的存在。

图 12-28 PIM-SM 中组播源注册示意图



如图 12-28 所示，组播源向 RP 注册的过程如下：

1. 当组播源 S 向组播组 G 发送了一个组播报文时，与组播源直连的 DR 在收到该报文后，就将其封装成注册消息，并通过单播方式发送给相应的 RP。
2. 当 RP 收到该消息后，一方面解封装注册消息并将封装在其中的组播报文沿着 RPT 转发给接收者，另一方面向组播源逐跳发送 (S, G) 加入消息。  
这样，从 RP 到组播源所经过的路由器就形成了 SPT 的分支，这些路由器都在其转发表中生成了 (S, G) 表项。SPT 以组播源为根，以 RP 为叶子。
3. 组播源发出的组播数据沿着已建立好的 SPT 到达 RP，然后由 RP 把组播数据沿着 RPT 向接收者进行转发。

当 RP 收到沿着 SPT 转发来的组播数据后，通过单播方式向与组播源直连的 DR 发送注册停止消息 (Register-Stop Message)，组播源注册过程结束。

## SPT 切换

当接收者侧的 DR 发现从 RP 发往组播组 G 的组播数据速率超过了一定的阈值时，将由其发起从 RPT 向 SPT 的切换，过程如下：

1. 接收者侧 DR 向组播源 S 逐跳发送 (S, G) 加入消息，并最终送达组播源侧 DR，沿途经过的所有路由器在其转发表中都生成了 (S, G) 表项，从而建立了 SPT 分支。
2. 接收者侧 DR 向 RP 逐跳发送包含 RP 位的剪枝消息，RP 收到该报文后会向组播源方向继续发送剪枝消息（假设此时只有这一个接收者），从而最终实现从 RPT 向 SPT 的切换。

从 RPT 切换到 SPT 后，组播数据将直接从组播源发送到接收者。通过由 RPT 向 SPT 的切换，PIM-SM 能够以比 PIM-DM 更经济的方式建立 SPT。

断言（Assert）

PIM-SM 使用与 PIM-DM 类似的断言机制，具体请参见 12.4.4.2 PIM-DM 中的“断言（Assert）”。

12.4.4.4 PIM-SM 的 BSR 管理域

在传统模式下，一个 PIM-SM 网络内只有一个 BSR（BootStrap Router），由该 BSR 负责在整个网络内收集并宣告 RP-Set 集信息，所有组播组的信息都在此 BSR 管理的网络范围内进行转发。

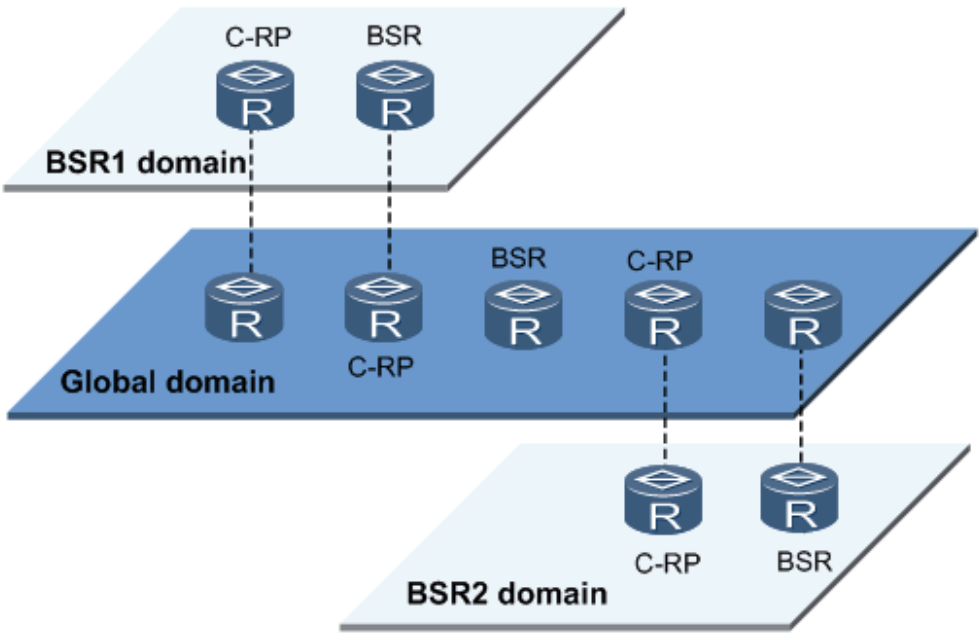
出于网络管理精细化的目的，可以将一个统一的 PIM-SM 网络划分为多个 BSR 管理域和一个 Global 域。这样一方面可以有效的分担单一 BSR 的管理压力，另一方面可以使用私有组地址为特定区域的用户提供专门服务。

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。

以下将从地域空间、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 12-29 地域空间上 BSR 管理域和 Global 域的关系

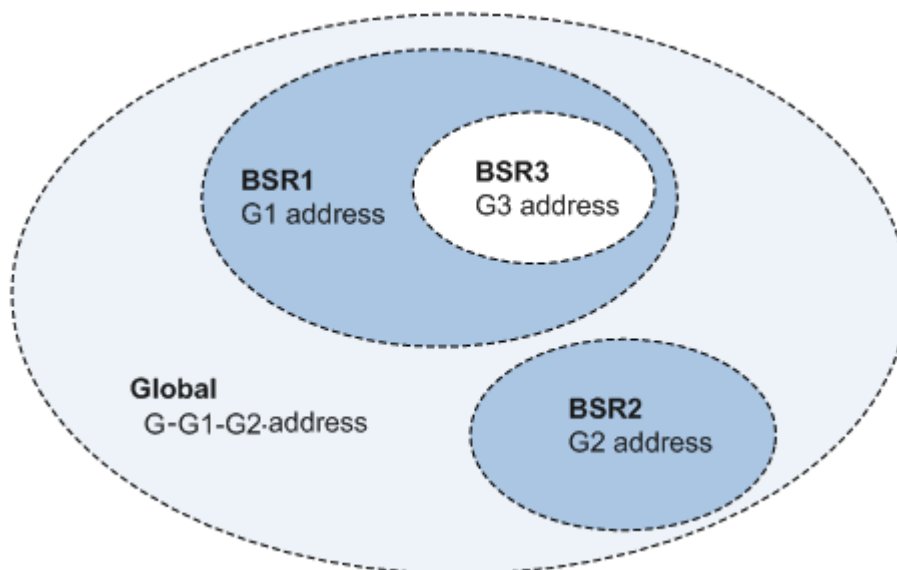


如图 12-29 所示，对于有相同组地址的不同管理域，各 BSR 管理域所包含的路由器互不相同，同一路由器不能从属于多个 BSR 管理域。各 BSR 管理域在地域上相互独立，且相互隔离。BSR 管理域是针对特定地址范围的组播组的管理区域，属于此范围的组播报文只能在本管理域内传播，无法通过 BSR 管理域边界。

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 12-30 组地址范围上 BSR 管理域和 Global 域的关系



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 12-30 所示，BSR1 域和 BSR3 域对应的组地址范围出现重叠（BSR3 域的组地址是 BSR1 域组地址的子集）。

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（“-”为减号）。也就是说，Global 域和所有 BSR 管理域之间就组地址范围来说是互补关系。

## 组播功能

如图 12-29，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，可以穿越任一 BSR 管理域。

### 12.4.4.5 PIM-SSM

SSM 模型为指定源组播提供了解决方案，通过 IGMPv3 来维护主机与路由器之间的关系。在实际应用中，通常采用 PIM-SM 模式的一部分技术来实现 SSM 模型。



由于接收者已经通过其它渠道（如广告咨询等）知道了组播源的具体位置，因此在 SSM 模型中无需 RP，无需构建 RPT，无需组播源注册过程，也无需通过 MSDP（Multicast Source Discovery Protocol，组播源发现协议）来发现其它 PIM 域内的组播源。

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

与 ASM 模型相比，SSM 模型仅需要 IGMPv3 和 PIM-SM 部分子集的支持，SSM 模型在 PIM 域内的工作机制可以简单概括如下：

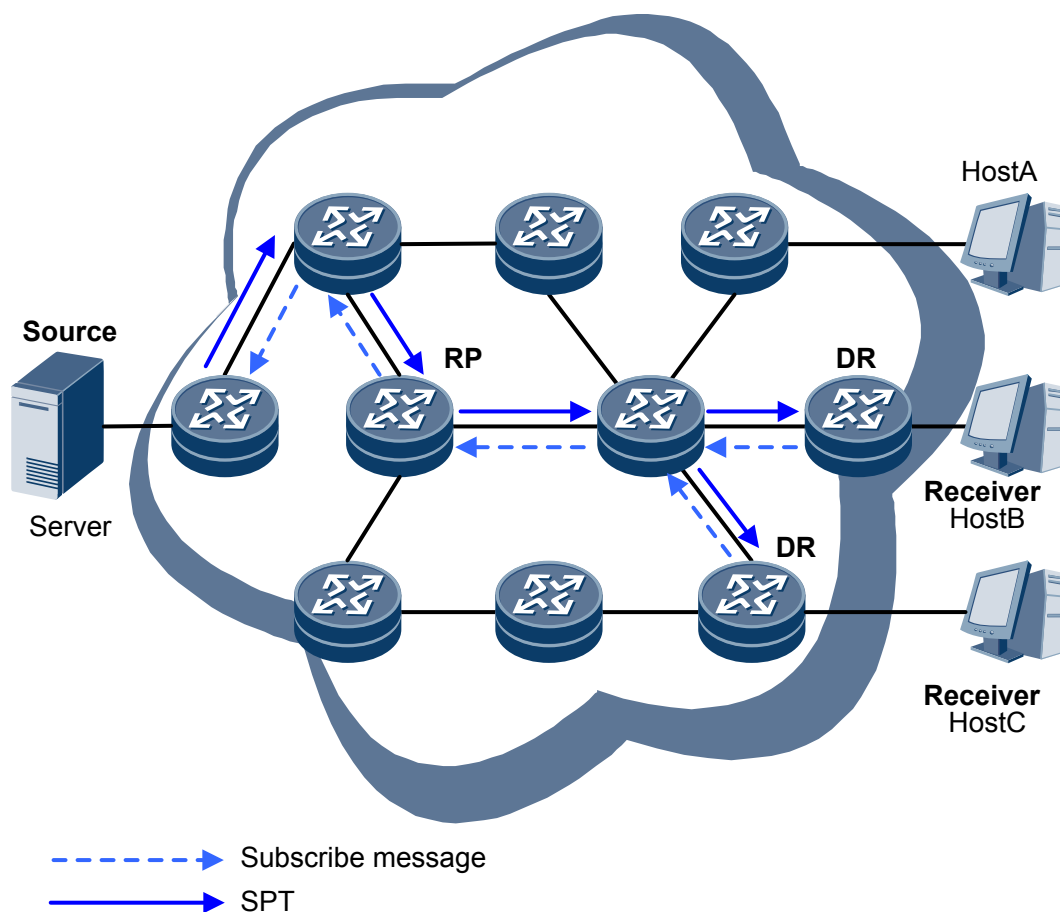
- 邻居发现
- DR 选举
- 构建 SPT

邻居发现机制、DR 选举过程都与 PIM-SM 中的描述相同，也是通过各路由器之间彼此发送 Hello 消息实现的，具体请参见 [12.4.4.3 PIM-SM](#)。

## 构建 SPT 树

无论是构建为 ASM 模型服务的 RPT 树，还是构建为 SSM 模型服务的 SPT 树，关键在于接收者准备加入的组播组是否属于 SSM 组地址范围（IANA 保留的 SSM 组地址范围为 232.0.0.0/8）。

图 12-31 PIM-SSM 中构建 SPT 示意图



如图 12-31 所示，如果 HostB 和 HostC 需要接收组播源 S 发送的信息，则借助 IGMPv3 的 Report 消息向最近的 DR 报告自己要加入指定源 S 的组播组 G，标记为（G，

INCLUDE, (S))；如果需要接收除指定源 S 外的其他源的信息，则其发送的 Report 消息可标记为 (G, EXCLUDE, (S))。

接收到该消息的 DR 判断该消息中的组地址 G 是否在 SSM 组地址范围内：

- 如果在 SSM 组地址范围内，则构建 PIM-SSM，并向指定组播源 S 逐跳发送通道订阅的 Join 消息。沿途所有路由器上都创建 (S, G) 表项，从而在网络内构建了一棵以指定组播源 S 为根，以接收者为叶子节点的 SPT，该 SPT 就是 SSM 模型中的传输通道。
- 如果 DR 发现组地址 G 不在 SSM 组地址范围内，则应用 ASM 模型，此时 DR 需要向 RP 发送 (\*, G) 加入消息，同时需要组播源注册过程。

PIM-SSM 应用

一个 PIM-SM 网络可以同时采用 ASM 模式和 SSM 模式为用户主机提供服务。

- 当用户加入的组播组属于 SSM 组地址范围内，则适用 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。

使用 SSM 实现组播业务，需要满足如下条件：

- 用户加入的组播组在 SSM 组地址范围内。
- 用户加入组播组的同时明确指定组播源的位置。
- 路由器之间运行 PIM-SM 协议。

12.4.4.6 PIM 协议比较

表 12-8 协议比较

| 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。 |
| 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。         |



### 12.4.4.7 PIM 消息

#### PIM 消息类型

PIM 消息类型和适用的协议如表 12-9 所示。

表 12-9 PIM 消息类型

| 消息类型                       | 适用协议                |
|----------------------------|---------------------|
| Hello                      | PIM-DM 与 PIM-SM 都适用 |
| Register                   | 只适用于 PIM-SM         |
| Register-Stop              | 只适用于 PIM-SM         |
| Join/Prune                 | PIM-DM 与 PIM-SM 都适用 |
| Bootstrap                  | 只适用于 PIM-SM         |
| Assert                     | PIM-DM 与 PIM-SM 都适用 |
| Graft                      | 只适用于 PIM-DM         |
| Graft-Ack                  | 只适用于 PIM-DM         |
| Candidate-RP-Advertisement | 只适用于 PIM-SM         |
| State Refresh              | 只适用于 PIM-DM         |

#### Hello 消息

PIM 路由器从所有 PIM 接口周期性的向外发送 Hello 消息。PIM 路由器之间通过交互 Hello 消息，发现 PIM 邻居并维护邻居关系。

封装 Hello 消息的 IP 报文源地址为本地接口地址，目的地址为 224.0.0.13，TTL 值为 1。使用组播方式发送。

#### Register 消息

当 PIM-SM 网络中出现活跃组播源时，源端 DR 向 RP 发送 Register 消息，进行源注册。

封装 Register 消息的 IP 报文源地址为源端 DR，目的地址为 RP。使用单播方式发送。

#### Register-Stop 消息

在 PIM-SM 网络中，在以下三种情况下，RP 将会向组播源端 DR 发送 Register-Stop 消息。

- 接收者不再通过 RP 接收发往某组播组的数据
- RP 不再为某组播组服务
- 组播数据转发路径已经由 RPT 切换到 SPT

组播源端 DR 收到 Register-Stop 消息后，停止使用 Register 注册消息封装组播数据报文，并进入注册抑制状态。

封装 Register-Stop 消息的 IP 报文源地址为 RP，目的地址为源端 DR。使用单播方式发送。

## Join/Prune 消息

一条 Join/Prune 消息中可以同时包含 Join 信息和 Prune 信息。只包含 Join 信息的 Join/Prune 消息称为 Join 消息。只包含 Prune 信息的 Join/Prune 消息称为 Prune 消息。

- 当 PIM 路由器下游没有组播需求时，从上游接口发出 Prune 消息，通知上游停止向该网段转发组播报文。
- 当 PIM-SM 网络中出现组成员时，组成员端 DR 从朝向 RP 的 RPF 接口，发出 Join 消息，通知上游邻居向该网段转发组播报文。Join 消息逐跳向上，构建 RPT。
- 当 RP 触发 SPT 切换时，RP 从朝向组播源的 RPF 接口，发送 Join 消息，通知上游邻居向该网段转发组播报文。Join 消息逐跳向上，构建 RP-源树。
- 当组成员端 DR 触发 SPT 切换时，组成员端 DR 从朝向组播源的 RPF 接口，发送 Join 消息，通知上游邻居向该网段转发组播报文。Join 消息逐跳向上，构建 SPT。

封装 Join/Prune 消息的 IP 报文源地址为本地接口地址，目的地址为 224.0.0.13，TTL 值为 1。使用组播方式发送。

## Bootstrap 消息

当 PIM-SM 网络中使用动态 RP 时，配置了 C-BSR 的路由器从所有 PIM 接口周期性的发送 Bootstrap 消息，参与 BSR 竞选。竞选获胜者，继续发送 Bootstrap 消息，向域内所有 PIM 路由器发布 RP-Set 信息。

封装 Bootstrap 消息的 IP 报文源地址为 C-BSR 地址，目的地址为 224.0.0.13，TTL 为 1，使用组播方式发送。

可以在 PIM 接口上配置 BSR 边界。众多 BSR 边界接口将网络划分成不同 PIM-SM 域，Bootstrap 消息无法通过 BSR 边界。

## Assert 消息

在共享网段上，如果 PIM 路由器从 (S, G) 或 (\*, G) 表项的下游接口收到 (S, G) 报文，则表示该网段存在其他的转发者。路由器从该下游接口发出 Assert 消息，参与竞选。竞选落败者停止下游接口的转发。

封装 Assert 消息的 IP 报文源地址为本地接口地址，目的地址为 224.0.0.13，TTL 值为 1。使用组播方式发送。

## Graft 消息

在 PIM-DM 网络中，路由器上出现组成员时，如果本身不在 SPT 上，则从对应的 (S, G) 表项的上游接口发送 Graft 消息。上游邻居立即恢复下游接口的转发。如果上游邻居不在 SPT 上，则继续向上游发送 Graft 消息。

封装 Graft 消息的 IP 报文源地址为上游接口地址，目的地址为 RPF 邻居。使用单播方式发送。

## Graft-Ack 消息

在 PIM-DM 网络中，路由器收到下游发来的 Graft 消息后，恢复相应下游的转发。同时从该下游接口发出 Graft-Ack 消息，表示已经接受嫁接请求。

如果发出 Graft 消息的路由器在一个设定时间内没有收到 Graft-Ack 消息，则认为上游未收到 Graft 消息，重发 Graft 消息。

封装 Graft-Ack 消息的 IP 报文源地址为下游接口地址，目的地址为 Graft 消息的发出者。使用单播方式发送。

## C-RP Advertisement 消息

当 PIM-SM 网络中使用动态 RP 时，配置了 C-RP 的路由器周期性的向 BSR 发送 Advertisement 消息，通告希望服务的组范围。

封装 Advertisement 消息的 IP 报文源地址为源端 C-RP，目的地址为 BSR。使用单播方式发送。

## State-Refresh 消息

在 PIM-DM 网络中，为了避免被剪枝的接口因为“剪枝定时器”超时而恢复转发，离组播源最近的第一跳路由器会周期性地触发 State-Refresh 消息。State-Refresh 消息在全网扩散，刷新所有路由器上的剪枝定时器状态。

封装 State-Refresh 消息的 IP 报文源地址为下游接口地址，目的地址为 224.0.0.13，TTL 值为 1。使用组播方式发送。

## 12.4.4.8 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。

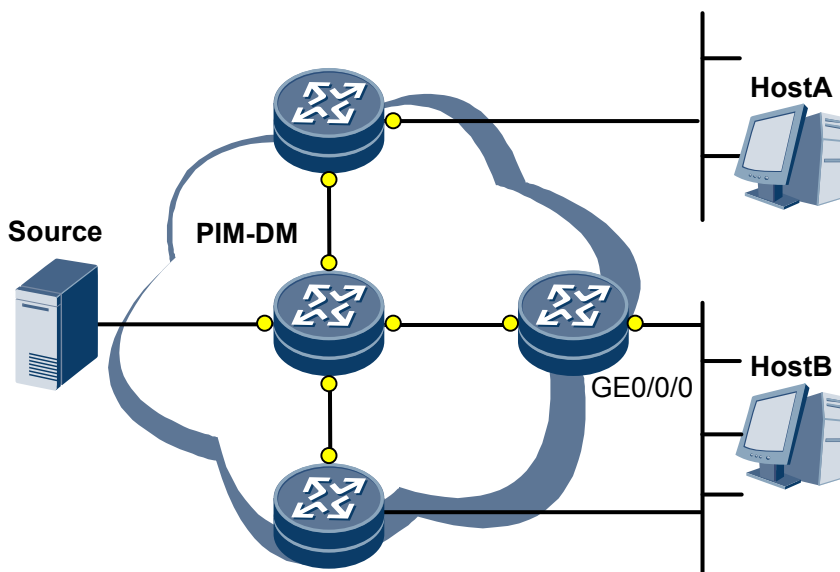
默认情况下，接收或发送 Join/Prune 消息和 Assert 消息时，不检查该消息是否来自 PIM 邻居或发送给 PIM 邻居。

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。

### 12.4.4.9 PIM 典型应用

## PIM-DM 网络

图 12-32 PIM-DM 典型组网图



● 需要启用PIM-DM的接口

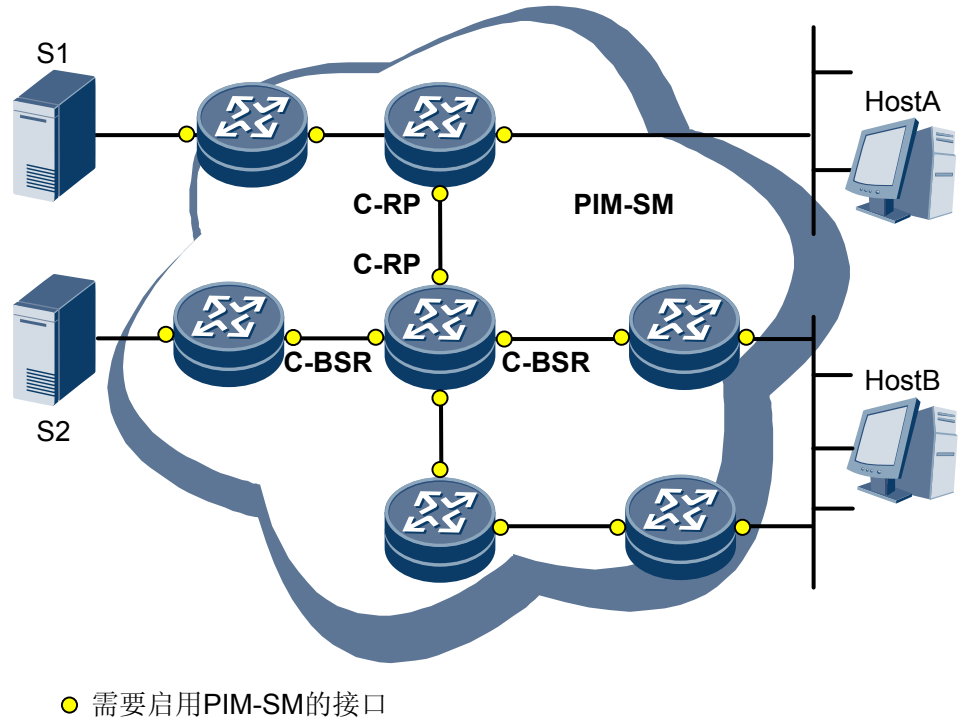
PIM-DM 称为协议独立组播-密集模式。适合规模较小、组播组成员相对比较密集的组播网络。

实现方案如下：

- 配置单播路由协议，确保任意网段路由可达。
- 在接收者主机上启用 IGMP 协议。
- 在与接收者主机处于同一网段的路由器接口上启用 IGMP 协议。
- 在所有路由器接口上启用 PIM-DM 协议。

PIM-SM 网络

图 12-33 PIM-SM 典型组网图



PIM-SM 适合网络中的组成员相对比较稀疏，分布广泛的大型网络。

实现方案如下：

- 配置单播路由协议，确保任意网段路由可达。
- 在接收者主机上启用 IGMP 协议。
- 在与接收者主机处于同一网段的路由器接口上启用 IGMP 协议。
- 在所有路由器接口上启用 PIM-SM 协议。
- 在某些 PIM-SM 接口上配置 C-BSR 和 C-RP。

PIM-DM 和 PIM-SM 的选择

密集（DM）和稀疏（SM）在定义上并没有严格的界限。下面的实验结果可以作为 PIM 协议两种模式网络应用环境的一个初步参考。

在 47 个组播节点，5 个组播源的网络环境下，分别应用 PIM-DM 和 PIM-SM，测量了数据报文和控制报文占用的网络带宽、路由器的处理开销。实验的结果如表 12-10 所示。

表 12-10 组播协议实验结果

| 观察对象    | 组播成员所占百分比 | 试验结果                     |
|---------|-----------|--------------------------|
| 占用的网络带宽 | 大于 42%    | PIM-SM 发出的数据包比 PIM-DM 多。 |

| 观察对象        | 组播成员所占百分比 | 试验结果                     |
|-------------|-----------|--------------------------|
|             | 小于 42%    | PIM-SM 发出的数据包比 PIM-DM 少。 |
| 路由器上的路由表项个数 | 小于 32%    | PIM-SM 协议维护的路由表项个数少。     |
|             | 32% ~ 58% | 两种路由协议情况下，路由表项个数相当。      |
|             | 大于 58%    | PIM-DM 协议维护的路由表项个数少。     |

## PIM-SSM 网络

PIM-SSM 网络适用于网络用户能够预先知道组播源的具体位置的情况。

PIM-SSM 网络无需维护 RP、无需构建 RPT、无需注册组播源，可以直接从源直连路由器到接收端 DR 建立 SPT。

实现方案如下：

- 配置单播路由协议，确保任意网段路由可达。
- 在接收者主机上启用 IGMP 协议。
- 在与接收者主机处于同一网段的路由器接口上启用 IGMPv3。SSM 模型需要 IGMPv3 版本的支持。如果用户主机必须运行 IGMPv1 或 IGMPv2，则需要在路由器接口上配置 SSM Mapping。
- 在所有路由器接口上启用 PIM-SM 协议。
- 配置将路由器加入 SSM 组播组地址范围（232.0.0.0/8）。

## 12.5 MSDP

MSDP 是为了解决多个 PIM-SM 域之间的互连而开发的一种域间组播解决方案，用来发现其它 PIM-SM 域内的组播源信息。

### 12.5.1 介绍

#### 定义

MSDP 是 Multicast Source Discovery Protocol（组播源发现协议）的简称，是为了解决多个 PIM-SM（Protocol Independent Multicast Sparse Mode，协议无关组播—稀疏模式）域之间的互连而开发的一种域间组播解决方案，用来发现其它 PIM-SM 域内的组播源信息。

## 目的

在基本的 PIM-SM 模式下，组播源只向本 PIM-SM 域内的 RP（Rendezvous Point，汇聚点）注册，且各域的组播源信息是相互隔离的，因此 RP 仅知道本域内的组播源信息，只能在本域内建立组播分发树，将本域内组播源发出的组播数据分发给本地用户。

如果能够有一种机制，将其它域内的组播源信息传递给本域内的 RP，则本域内的 RP 就可以向其它域内的组播源发起加入过程并建立组播分发树，实现组播数据跨域传输，从而使本域内的组成员主机接收到其他域的组播源发出的数据。

基于这一设想，MSDP 通过在网络中选取适当的路由器建立 MSDP 对等体关系，以连通各 PIM-SM 域的 RP。通过在各 MSDP 对等体之间交互 SA（Source Active，信源有效）消息来共享组播源信息。

## 12.5.2 参考标准和协议

与 MSDP 相关的参考标准和协议如下：

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

## 12.5.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

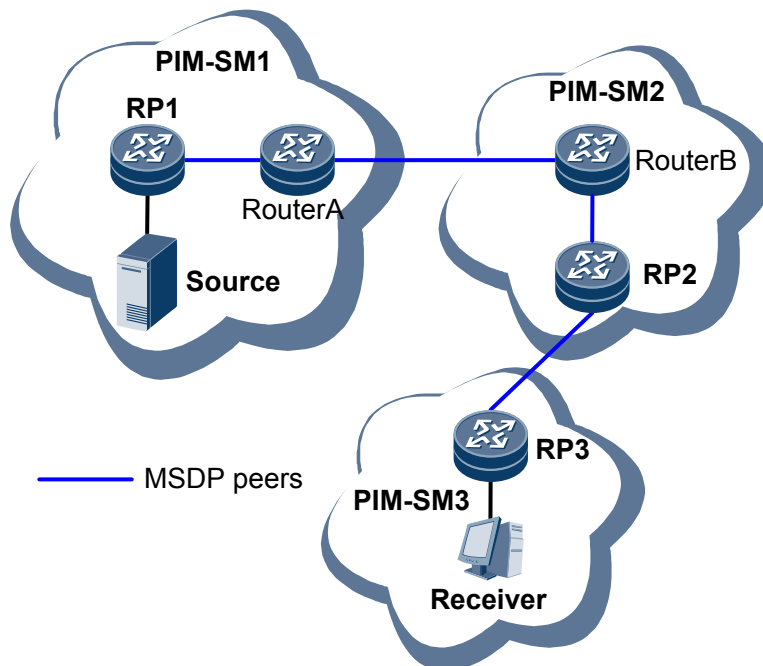
## 12.5.4 原理描述

### 12.5.4.1 MSDP 对等体

通过在网络中配置一对或多对 MSDP 对等体，形成彼此相连的一张“MSDP 连通图”，以连通各个 PIM-SM 域的 RP。通过这些 MSDP 对等体之间的接力，可以把某 RP 发出的 SA 消息传递给其它所有的 RP。



图 12-34 MSDP 对等体的位置



如图 12-34 所示，MSDP 对等体可以创建在任意的 PIM-SM 路由器上，在不同角色的 PIM-SM 路由器上所创建的 MSDP 对等体的功能有所不同：

- 在 RP 上创建的 MSDP 对等体
  - 源端 MSDP 对等体：即离组播源（Source）最近的 MSDP 对等体（通常也就是源端 RP，如 RP1）。源端 RP 创建 SA 消息并发送给远端 MSDP 对等体，通告在本 RP 上注册的组播源信息。源端 MSDP 对等体必须配置在 RP 上，否则将无法向外发布组播源信息。
  - 接收者端 MSDP 对等体：即离接收者（Receiver）最近的 MSDP 对等体（如 RP3）。接收者端 MSDP 对等体在收到 SA 消息后，根据该消息中所包含的组播源信息，跨域加入以该组播源为根的 SPT；当来自该组播源的组播数据到达后，再沿 RPT 向本地接收者转发。
  - 中间 MSDP 对等体：即拥有多个远端 MSDP 对等体的 MSDP 对等体（如 RP2）。中间 MSDP 对等体把从一个远端 MSDP 对等体收到的 SA 消息转发给其它远端 MSDP 对等体，其作用相当于传输组播源信息的中转站。
- 在普通的 PIM-SM 路由器（非 RP）上创建的 MSDP 对等体  
如 RouterA 和 RouterB，其作用仅限于将收到的 SA 消息转发出去。

## 推荐配置方案

MSDP 对等体之间使用 TCP 连接。因此，不直接相连的路由器之间也可以建立 MSDP 对等体连接。配置 MSDP 对等体的目的是为了连通网络中的 RP，因此最直接有效的配置方法就是在网络中所有 RP（包括静态 RP 和 C-RP）上配置 MSDP 对等体。

- 如果 PIM-SM 域内同时存在多个 RP，为不同的组播组服务，则推荐在所有的 RP 之间建立 MSDP 对等体关系。
- 如果 PIM-SM 域内从多个 C-RP 中动态选举产生一个 RP，为了保证选举获胜的 C-RP 始终位于“MSDP 连通图”上，建议在所有的 C-RP 之间建立 MSDP 对等体关系。选举落败的 C-RP 在“MSDP 连通图”上担当的角色相当于非 RP 路由器。

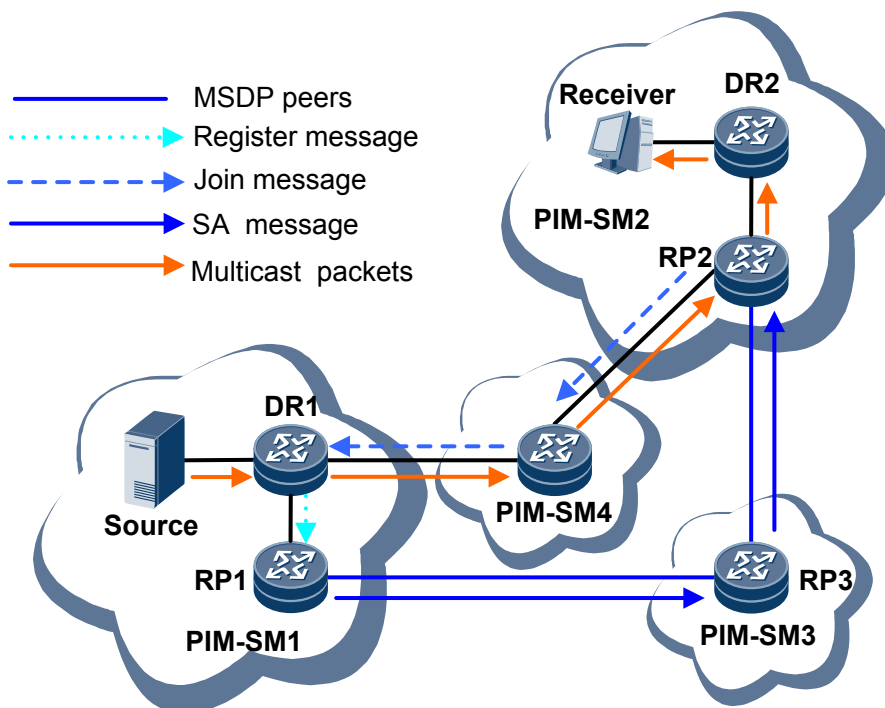
- 在 PIM-SM 域间建立 MSDP 对等体关系，把各个 PIM-SM 域内的“MSDP 连通图”连接起来。

#### 12.5.4.2 MSDP 实现 PIM-SM 域间组播

在组播网络被划分成多个 PIM-SM 域时，使用 MSDP 沟通各个域内的 RP，共享组播源信息。最终使一个域内的用户接收到其它域内组播源发送的组播数据。

如图 12-35 所示，PIM-SM1 域内存在激活的组播源（Source），RP1 通过组播源注册过程了解到了该组播源的存在。如果 PIM-SM2 和 PIM-SM3 域也希望知道该组播源的具体位置，进而能够从该组播源获取组播数据，则需要在 RP1 与 RP3、RP2 与 RP3 之间分别建立 MSDP 对等体关系。

图 12-35 MSDP 对等体示意图



借助 MSDP 对等体实现 PIM-SM 域间组播的工作过程如下：

1. 当 PIM-SM1 域内的组播源向组播组 G 发送第一个组播数据包时，DR1 将该组播数据封装在注册消息（Register Message）中，并发给 RP1。RP1 因此获知了该组播源的相关信息。
2. RP1 作为源端 RP，创建 SA 消息，并周期性地向其它 MSDP 对等体发送。SA 消息中包含组播源的地址 S、组播组的地址 G 以及创建该 SA 消息的源端 RP（即 RP1）的地址。
3. MSDP 对等体对收到的 SA 消息进行 RPF（Reverse Path Forwarding，逆向路径转发）检查，以及各种转发策略的过滤，从而只接受和转发来自正确路径并通过过滤的 SA 消息，以避免 SA 消息传递环路。

另外，可以在 MSDP 对等体之间配置 MSDP 全连接组（12.5.4.4 Mesh Group），以避免 SA 消息在 MSDP 对等体之间的泛滥。

- 4. SA 消息在 MSDP 对等体之间转发，最终该组播源的相关信息将传遍所有建立了 MSDP 对等体关系的 PIM-SM 域（即 PIM-SM2 和 PIM-SM3）。
- 5. PIM-SM2 中的 RP2 在收到该 SA 消息后，检查本域内是否有组播组 G 的接收者（Receiver）存在：
  - 如果有接收者，RP2 与接收者之间维护组播组 G 的 RPT。RP2 创建（S，G）表项，向源端的 DR1 逐跳发送（S，G）加入消息（Join Message），从而跨越各 PIM-SM 域直接加入以该组播源为根的 SPT。组播数据沿 SPT 到达 RP2 后，再沿 RPT 向接收者转发。当接收者端的 DR2 收到来自组播源的组播数据后，可以自行决定是否发起从 RPT 向 SPT 的切换。
  - 如果没有接收者，RP2 不会创建（S，G）表项，也不加入以该组播源为根的 SPT。

12.5.4.3 MSDP 实现 Anycast RP

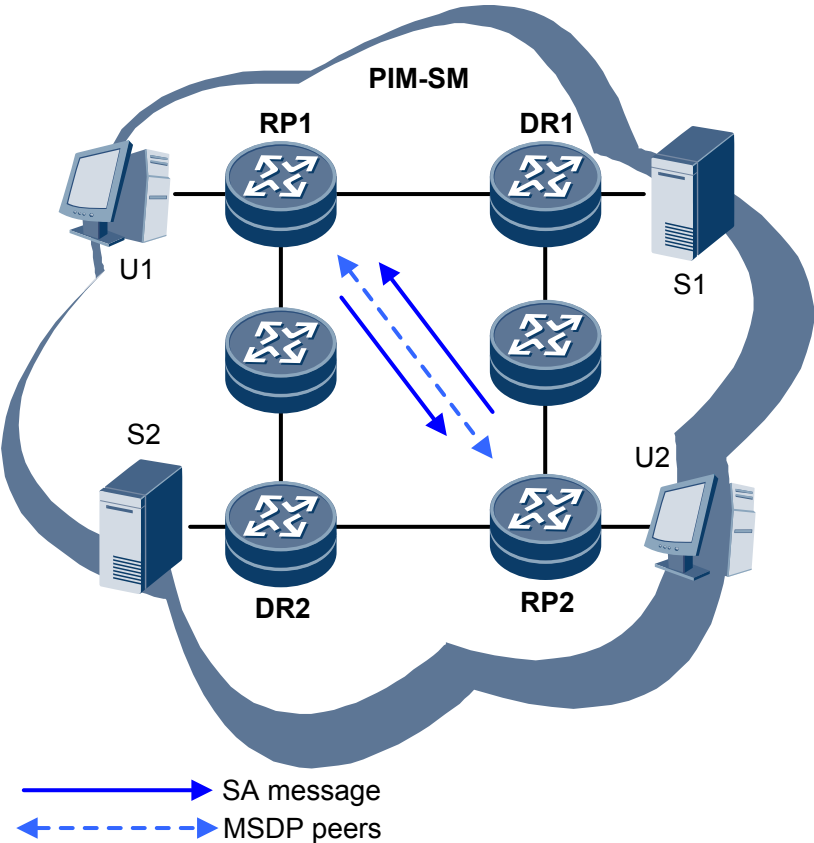
Anycast RP 应用环境

在传统的 PIM-SM 域中，每个组播组都只能映射到一个 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。

Anycast RP 是指在同一个 PIM-SM 域内设置多个具有相同地址的 RP，并且在这些 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。

如图 12-36 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 12-36 Anycast RP 典型组网图



在 PIM-SM 域内应用 Anycast RP，配置方法如下：

- 在 RP1 和 RP2 上配置相同的 Loopback 接口地址作为 RP1 和 RP2 的 RP 地址。
- 在 RP1 和 RP2 上配置不同的逻辑 RP 地址，不能与实际 RP 接口相同。建议配置为 MSDP 对等体接口地址。
- 在 RP1 和 RP2 两个路由器之间建立 MSDP 对等体关系。

## Anycast RP 实现过程

以图 12-36 为例，借助 MSDP 对等体实现 PIM-SM 域内 Anycast RP 的工作过程如下：

1. 接收者选择距离最近的 RP 发送加入消息以构建 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 之间交互 SA 消息，共享组播源信息。
  - RP1 创建一个 SA 消息，发送到 RP2。SA 消息中携带 (S1, G)。
  - RP2 创建一个 SA 消息，发送到 RP1。SA 消息中携带 (S2, G)。
4. RP 加入以源端 DR 为根的 SPT，接收者沿 RPT 接收到组播数据。
  - 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. 接收者接收到组播数据后，自行决定是否发起 SPT 切换。

## Anycast RP 的功能

使用 Anycast RP 的组播网络与传统的 PIM-SM 网络相比，具有如下功能：

- RP 路径最优
  - 接收者向距离最近的 RP 发起加入，建立路径最优的 RPT。
  - 组播源向距离最近的 RP 发起注册，建立路径最优的 SPT。
- RP 负载分担

每个 RP 上只维护 PIM-SM 域内的部分源/组信息，转发部分的组播数据。
- RP 冗余备份

一个 RP 失效后，原先注册的组播源和加入的接收者会自动选择另一个就近 RP 进行注册和加入操作。

### 12.5.4.4 Mesh Group

Mesh Group 内部成员不对来自内部的 SA 消息进行 RPF 检查。因此将多个 MSDP 对等体加入同一个 Mesh Group，可以减轻系统负担。

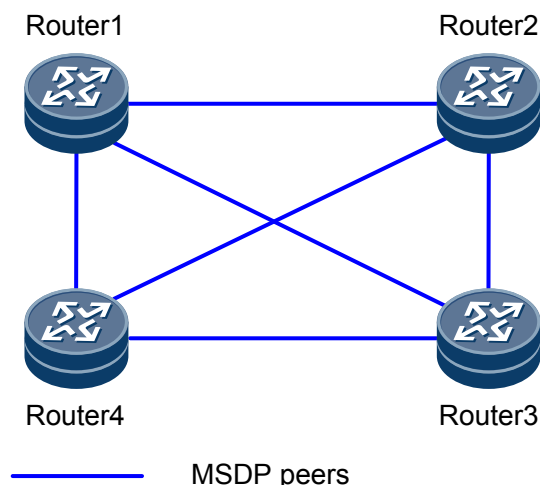
网络中存在多个 MSDP 对等体时，很容易导致 SA 消息在对等体之间泛滥。尤其在同一个 PIM-SM 域中配置了多个 MSDP 对等体的时候，RPF 规则不能过滤很多无用的 SA 消息。MSDP peer 对每一个到来的 SA 报文进行 RPF 检查，给系统造成很大的负担。

将多个 MSDP Peer 加入同一个 Mesh Group，就可以大幅度减少在这些 MSDP Peer 之间传递的 SA 消息。

Mesh Group 成员可以都属于同一个 PIM-SM 域，也可以分布在多个 PIM-SM 域中；可以都位于同一个 AS，也可以位于多个 AS 中。

属于同一个 Mesh Group 的所有成员之间必须两两建立 MSDP 对等体连接，并承认对方为该 Mesh Group 的成员。如图 12-37 中的 Router1、Router2、Router3 和 Router4，加入同一个 Mesh Group，则必须在每台路由器上配置与其他三台路由器建立 MSDP 对等体关系。

图 12-37 Mesh Group 内部成员之间的 MSDP 对等体连接



当 Mesh Group 内部成员接收到 SA 消息后，首先检查该 SA 消息的来源：

- 如果该 SA 消息来自 Mesh Group 外部的某个 MSDP Peer，则对该 SA 消息进行 RPF 检查。如果检查通过，向 Mesh Group 内其他所有成员转发。
- 如果该 SA 消息来自 Mesh Group 内部成员，则不进行 RPF 检查，直接接收。同时也不再向 Mesh Group 内其他成员转发。

#### 12.5.4.5 SA 消息的 RPF 检查

##### SA 消息的 RPF 规则

SA 消息的 RPF 规则主要有以下 6 条：

- 规则 1：发出 SA 消息的 MSDP 对等体就是源 RP（即创建该 SA 消息的 RP）时，则他的 MSDP 对等体接收该 SA 消息并向其他对等体转发。
- 规则 2：接收来自静态 RPF 对等体的 SA 消息。一台路由器可以同时与多个路由器建立 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：如果一台路由器只拥有一个远端 MSDP 对等体，则该远端对等体自动成为 RPF 对等体，路由器接收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 消息的对等体与本地路由器属于同一 Mesh Group 时，本地路由器接收该 SA 消息并不再向属于该 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 消息的对等体是本地路由器到源 RP 的“路由”下一跳，则本地路由器接收该 SA 消息并向其他对等体转发。“路由”包括：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（Autonomous System，自治系统）时，接收从下一跳 AS（以 AS 为单位）中的对等体发出的 SA 消息，如果该 AS 中存在多个远端 MSDP 对等体，则接收从 IP 地址最高的对等体发来的 SA 消息。

## 推荐配置方案

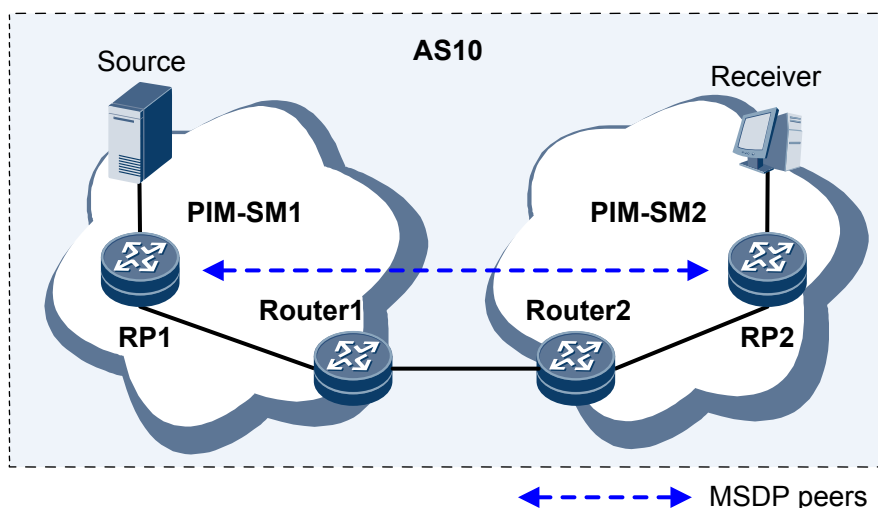
为了保证 MSDP 对等体之间传递的 SA 消息不被 RPF 规则阻断，且尽量减少冗余流量，推荐的配置方案如下：

- 将属于同一个 AS 的所有 MSDP 对等体加入同一个 Mesh Group。
- 在跨 AS 的 MSDP 对等体之间相互指定对方为静态 RPF 对等体。

### 12.5.4.6 MSDP 典型应用

#### AS 内 PIM-SM 域间组播

图 12-38 AS 内 PIM-SM 域间组播



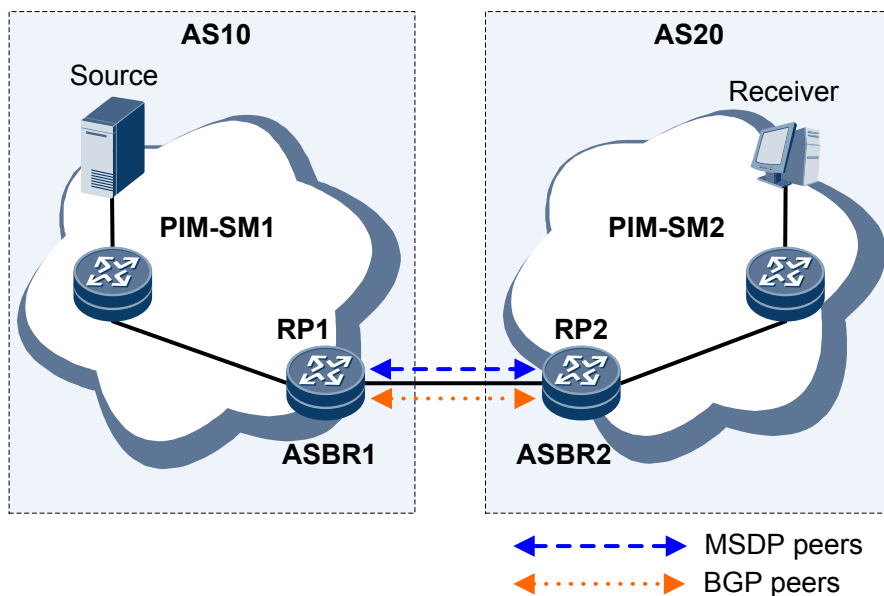
应用场景：当一个大型的组播网络划分成多个 PIM-SM 域时，需要使用 MSDP 连通各个 PIM-SM 域内的 RP，共享组播源信息。最终使一个 PIM-SM 域内的用户能够接收到其它 PIM-SM 域内组播源发送的组播数据。

解决方案：在两个 PIM-SM 域的 RP 之间配置 MSDP 对等体，并加入同一个 Mesh Group。



## 借助 BGP 对等体实现 AS 间组播

图 12-39 借助 BGP 对等体实现 AS 间组播

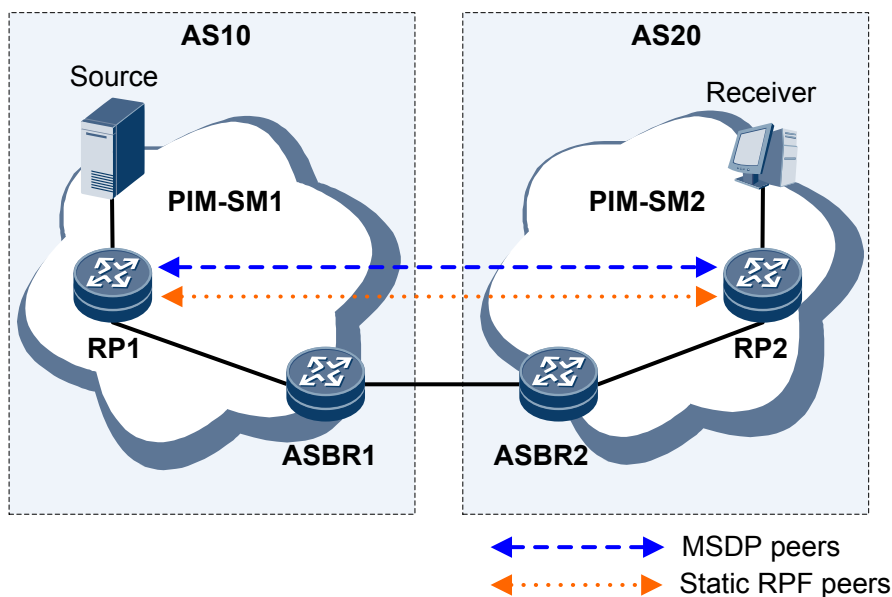


应用场景：连通处于不同 AS 的 PIM-SM 域时，如果 RP 与单播 ASBR（自治系统边界路由器）在同一对路由器上，可以借助 BGP 对等体实现 AS 间组播。

解决方案：在不同 AS 的两个作为 ASBR 的 RP 上配置 BGP 对等体和 MSDP 对等体。建立 MSDP 对等体的接口地址与建立 EBGP 对等体的接口地址相同。

## 借助静态 RPF 对等体实现 AS 间组播

图 12-40 借助静态 RPF 对等体实现 AS 间组播

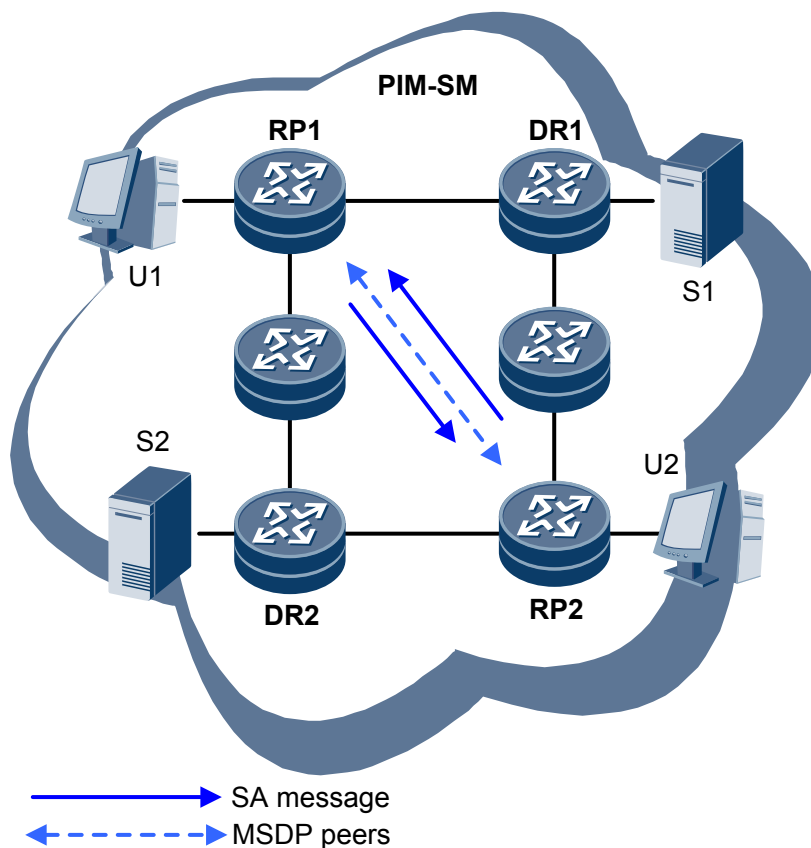


应用场景：当跨 AS 的 PIM-SM 域之间没有运行 BGP 路由或 RP 与单播 ASBR 不在同一对路由器上时，可以借助静态 RPF 对等体实现 AS 间组播。

解决方案：在不同 AS 的两个 RP 之间配置 MSDP 对等体并相互指定对方为自己的静态 RPF 对等体。

## Anycast RP

图 12-41 Anycast RP 典型组网图



应用场景：在 PIM-SM 域内应用 Anycast RP，可以实现组播源就近注册和接收者就近加入。既缓解了单个 RP 的负担，也实现了 RP 备份、优化了转发路径。

在 PIM-SM 域内应用 Anycast RP，配置方法如下：

- 在 RP1 和 RP2 上配置相同的 Loopback 接口地址作为 RP1 和 RP2 的 RP 地址。
- 在 RP1 和 RP2 上配置不同的逻辑 RP 地址，不能与实际 RP 接口相同。建议配置为 MSDP 对等体接口地址。
- 在 RP1 和 RP2 两个路由器之间建立 MSDP 对等体关系。

## 12.6 组播路由与转发

介绍组播路由和转发的基本原理、“控制平面”与“转发平面”之间的信息交互过程，以及用户控制组播转发功能的方式。

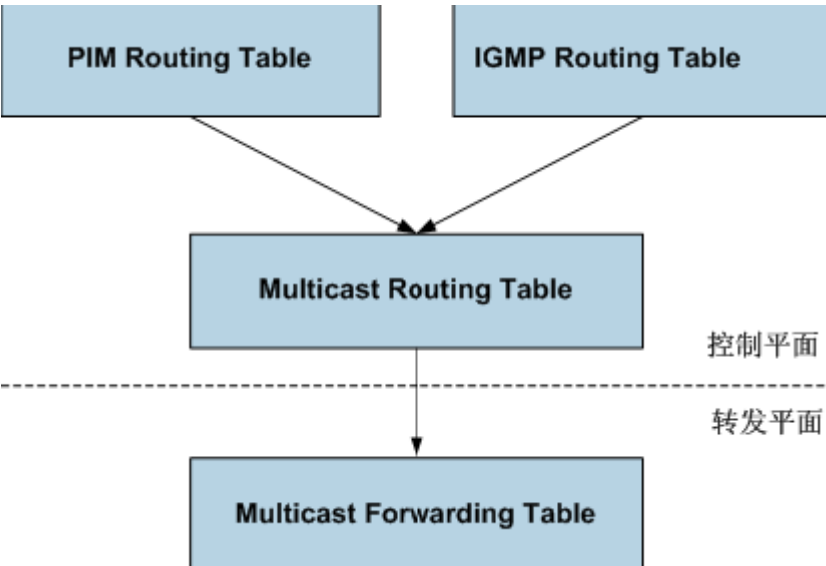
### 12.6.1 介绍



## 定义

组播路由器可以同时维护多种组播路由协议，通过“控制平面”与“转发平面”之间的信息交互，实现组播路由和转发。如图 12-42 所示。

图 12-42 组播路由表和转发表



- 控制平面负责收集路由信息，创建并维护路由表项。
- 转发平面继承控制平面的路由表项，指导报文转发。

组播路由协议必须在 RPF（Reverse Path Forwarding）检查通过后，才可以创建组播路由表项、构建组播分发树，从而转发组播报文。

## 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。
- 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。

## 12.6.2 参考标准和协议

与组播路由与转发相关的参考标准和协议如下：

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

## 12.6.3 可获得性

### License 支持

本特性无需 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 12.6.4 原理描述

### 12.6.4.1 控制平面

#### 主要功能

控制平面的核心任务是：生成并维护组播路由表。

- 每个组播路由协议维护一个协议自身的路由表，如 PIM Routing-Table 和 IGMP Routing-Table。
- 各个组播路由协议的组播路由信息经过综合形成一个总的组播路由表，即 Multicast Routing-Table。
- 当控制平面收到转发平面上送的 NOCACHE 事件时，执行 RPF 检查。根据协议情况，创建路由表项。
- 当控制平面收到转发平面上送的 WRONGIF 事件时，执行 RPF 检查。根据协议情况，更新路由表项。

#### 组播路由表

Multicast Routing-Table 是组播路由管理模块中的路由表，由一组（S，G）表项组成。（S，G）表示由源 S 向组播组 G 发送的组播数据的路由信息。

如果组播路由管理支持多种组播路由协议，则路由表中将汇集由多种协议生成的组播路由表项。路由表信息如下所示：

```
<USG> display multicast routing-table
Multicast routing table of VPN-Instance: public net
Total 1 entry
00001. (172.168.0.2, 227.0.0.1)
  Uptime: 00:00:28
  Upstream Interface: GigabitEthernet0/0/0
  List of 2 downstream interfaces
    1: GigabitEthernet0/0/1
    2: GigabitEthernet5/0/0
```

display multicast routing-table 命令输出信息描述如表 12-11 所示。

表 12-11 display multicast routing-table 命令输出信息描述

| 项目                                                                                     | 描述                |
|----------------------------------------------------------------------------------------|-------------------|
| 00001                                                                                  | (S,G)项的序号         |
| (172.168.0.2, 227.0.0.1)                                                               | 组播路由表的(S,G)项      |
| Uptime                                                                                 | (S,G)项已经存在的时间     |
| Upstream interface                                                                     | (S,G)项的上游接口       |
| List of 2 downstream interfaces:<br>1: GigabitEthernet0/0/1<br>2: GigabitEthernet5/0/0 | 下游接口列表，包含接口序号和接口名 |

12.6.4.2 转发平面

主要功能

转发平面的核心任务是：依照转发表项执行组播报文转发。

- 保持转发表项信息与路由表项信息一致。系统周期性地将路由表项下刷至转发表中。
- 设备接收到组播报文后，会查看转发表，决定是否转发该报文：
  - 如果转发表中存在与报文源组匹配的表项，且报文实际到达接口与转发表项信息一致，则执行报文转发。
  - 如果转发表中存在与报文匹配的表项，且报文实际到达接口与转发表项信息不一致，则丢弃报文，同时向控制平面上送 WRONGIF 事件。
  - 如果转发表中不存在与报文匹配的转发表项，缓存组播报文，同时向控制平面上送 NOCACHE 事件。

组播转发表

Multicast Forwarding-Table 指导组播报文转发。转发表项直接继承路由表项信息。转发表信息如下所示：

```
<USG> display multicast forwarding-table
Multicast Forwarding table of VPN-Instance: public net
Total 1 entry, 1 matched
00001. (172.168.0.2, 227.0.0.1), MID: 0, Flags: 0x0:0
  Uptime: 00:08:32, Timeout in: 00:03:26
  Incoming interface: GigabitEthernet0/0/0
  List of 1 outgoing interfaces:
    1: GigabitEthernet0/0/1
  Matched 18696 packets(523488 bytes), Wrong If 0 packets
  Forwarded 18696 packets(523488 bytes)
```

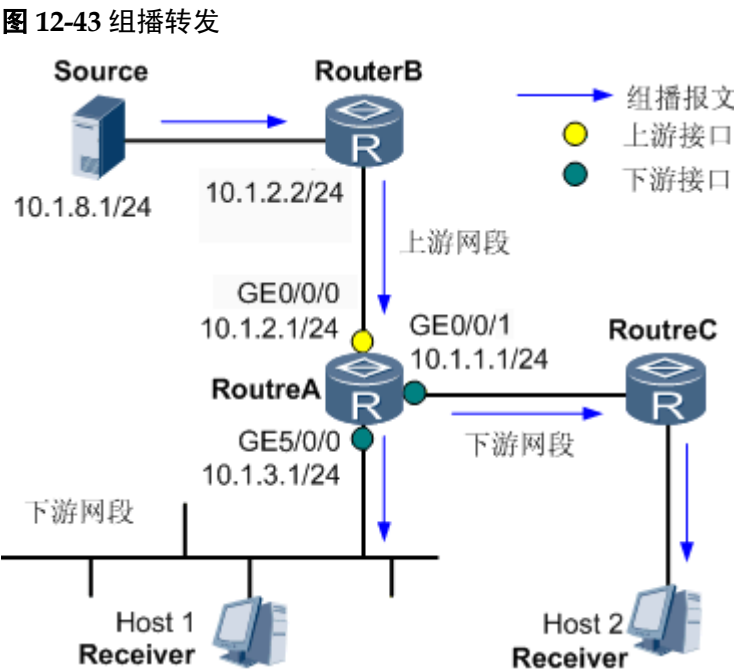
display multicast forwarding-table 命令输出信息描述如表 12-12 所示。

表 12-12 display multicast forwarding-table 命令输出信息描述

| 项目                                                                                                                                                                  | 描述                                                                                 |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| 00001                                                                                                                                                               | (S,G)项的序号                                                                          |
| (172.168.0.2, 227.0.0.1)                                                                                                                                            | 组播路由表的(S,G)项                                                                       |
| Uptime                                                                                                                                                              | (S,G)项已经存在的时间                                                                      |
| Timeout in                                                                                                                                                          | (S,G)项还有多少时间超时                                                                     |
| Incoming interface                                                                                                                                                  | (S,G)项的上游接口                                                                        |
| List of 2 downstream interfaces:<br>1: GigabitEthernet0/0/1<br>Matched 18696 packets(523488 bytes),<br>Wrong If 0 packets<br>Forwarded 18696 packets (523488 bytes) | 下游接口列表，包含接口序号、接口名、<br>(S,G)项匹配的报文个数(字节数)、发生<br>wrong if 的报文个数、已转发的组播报文<br>个数(字节数)。 |

12.6.4.3 组播转发流程

从设备的角度出发，组播转发流程可以简化为：组播报文从唯一的上游接口到达，从一个或多个下游接口发送出去。图 12-43 显示了以 RouterA 为行为主体的组播转发（组地址为 225.1.1.1）。



上游

- 上游：组播数据到达的方向。通常，将组播数据转发过来的路由器或组播源称为本路由器的上游。以图 12-43 为例，RouterA 的上游为 RouterB。

- 上游接口：本地路由器上接收到组播数据的接口。
- 上游网段：上游接口所在网段。以图 12-43 为例，RouterA 的上游网段为 10.1.2.0/24。

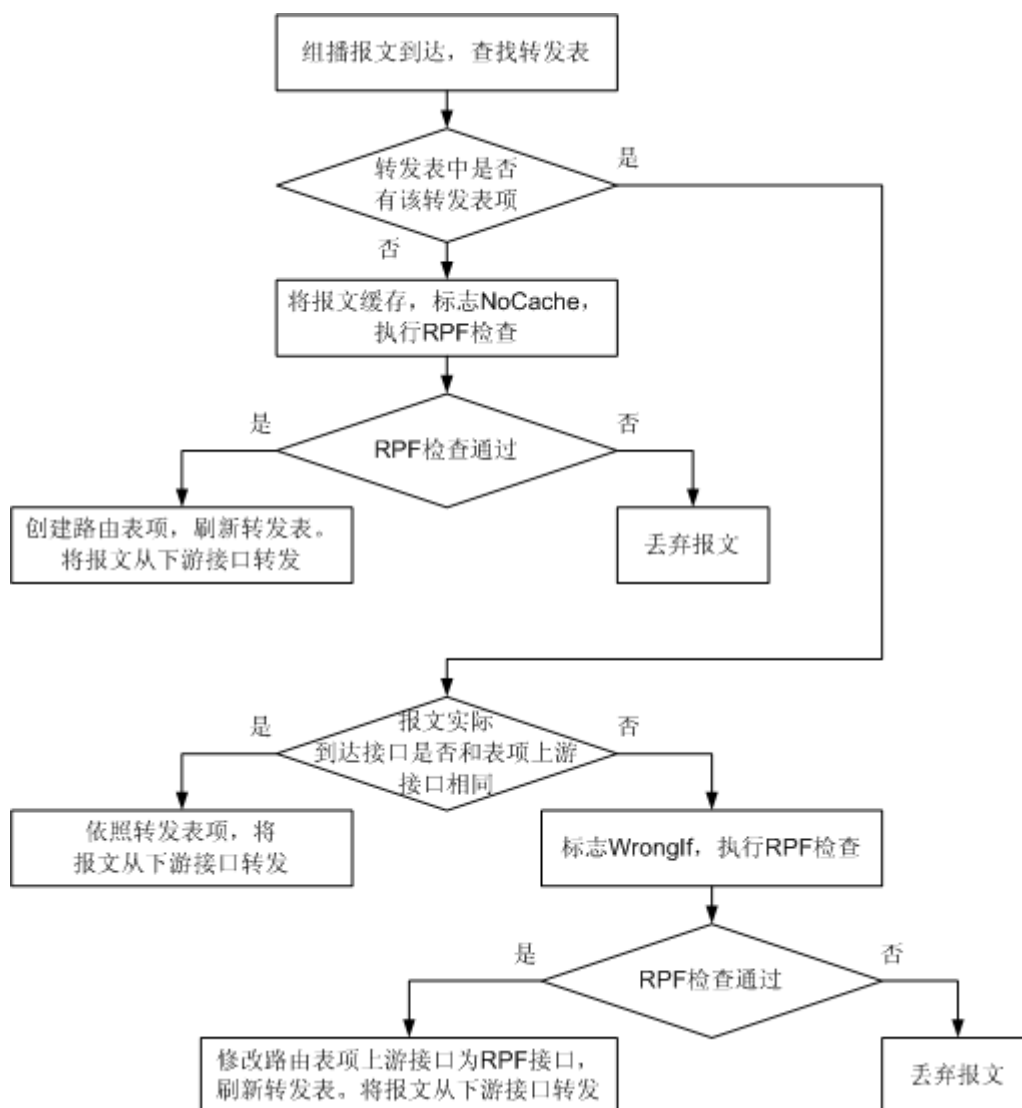
## 下游

- 下游：将组播数据发送出去的方向。通常，发送出去的组播数据到达的路由器或接收者主机称为本路由器的下游。以图 12-43 为例，RouterA 的下游为 RouterC 和 Host1。
- 下游接口：将组播数据转发出去的接口。
- 下游网段：下游接口所在网段。以图 12-43 为例，RouterA 的下游网段为 10.1.1.0/24 和 10.1.3.0/24。

## 组播转发流程

当组播路由器接收到由源 S 向组播组 G 发送的组播报文后，按照图 12-44 所示的流程，执行转发。

图 12-44 组播转发流程



在建立组播路由和转发表时，会把组播数据报文（S，G）的 RPF 接口记录为（S，G）表项的上游接口。

RPF 接口是由 RPF 路由（单播路由或组播静态路由）确定的。选取最优路由和 RPF 路由的具体过程请参见 [12.6.4.4 RPF 检查](#)。

当路由器收到组播数据报文（S，G）后，查找组播转发表：

1. 如果组播转发表中不存在（S，G）表项，则对该报文执行 RPF 检查，将其 RPF 接口作为上游接口，结合相关路由信息创建相应的表项，并下发到组播转发表中：
  - 若该报文实际到达的接口正是其 RPF 接口，则 RPF 检查通过，向所有的下游接口转发该报文。
  - 若该报文实际到达的接口不是其 RPF 接口，则 RPF 检查失败，丢弃该报文。
2. 如果组播转发表中已存在（S，G）表项，且该报文实际到达的接口与上游接口相匹配，则向所有的下游接口转发该报文。
3. 如果组播转发表中已存在（S，G）表项，但该报文实际到达的接口与上游接口不匹配，则对此报文执行 RPF 检查：
  - 若其 RPF 接口与上游接口一致，则说明（S，G）表项正确，丢弃这个来自错误路径的报文。
  - 若其 RPF 接口与上游接口不符，则说明（S，G）表项已过时，于是把上游接口更新为 RPF 接口。
  - 如果该报文实际到达的接口正是其 RPF 接口，则向所有的下游接口转发该报文，否则将其丢弃。

#### 12.6.4.4 RPF 检查

组播路由协议在创建组播路由表项时，运用了 RPF（Reverse Path Forwarding，逆向路径转发）检查机制，以确保组播数据能够沿正确的路径传输，同时还能避免由于各种原因而造成的环路。

执行 RPF 检查的过程如下：

- 依据“报文源”，查找单播路由表和组播静态路由表。从这些路由表中各选出一条最优路由。
- 从两条最优路由中选取一条作为 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。

#### 选取最优路由

本设备执行 RPF 检查的依据是单播路由或组播静态路由。单播路由表中汇集了到达各个目的网段的最短路径；组播静态路由表中列出了用户通过手工静态配置指定的 RPF 路由信息。

组播路由协议并不独立维护某种单播路由，而是依赖于网络中现有的单播路由信息或组播静态路由来创建组播路由表项。可以说，单播路由可达是组播路由可达的基础。

在执行 RPF 检查时，路由器同时查找单播路由表和组播静态路由表，分别从中选出一条最优路由，具体过程如下：

- 以“报文源”的 IP 地址为目的地址查找单播路由表，自动选取一条最优单播路由。对应表项中的下游接口为 RPF 接口，下一跳为 RPF 邻居。路由器认为来自 RPF 邻居且由该 RPF 接口收到的组播报文所经历的路径是从源 S 到本地的最短路径。

- 以“报文源”的 IP 地址为指定源地址查找组播静态路由表，自动选取一条最优组播静态路由。对应表项明确指定了 RPF 接口和 RPF 邻居。

## 选取 RPF 路由

如上所述，路由器已经从组播静态路由表和单播路由表中各选出一条最优路由。现在路由器需要从这两条路由中选取一条作为 RPF 路由，选取规则如下：

- 如果配置了按照最长匹配选择路由，则从这两条路由中选出掩码最长匹配的那条路由；如果这两条路由的掩码一样，则选择其中优先级较高的那条路由；如果它们的优先级也相同，则按照组播静态路由、单播路由的顺序进行选择。
- 如果没有配置按照最长匹配选择路由，则从这两条路由中选出优先级较高的那条路由；如果它们的优先级相同，则按照组播静态路由、单播路由的顺序进行选择。

## RPF 检查在组播转发中的应用

对每一个收到的组播数据报文都进行 RPF 检查会给路由器带来较大负担，而利用组播转发表可以解决这个问题。

在建立组播路由和转发表时，会把组播数据报文（S，G）的 RPF 接口记录为（S，G）表项的上游接口。

当路由器收到组播数据报文（S，G）后，查找组播转发表：

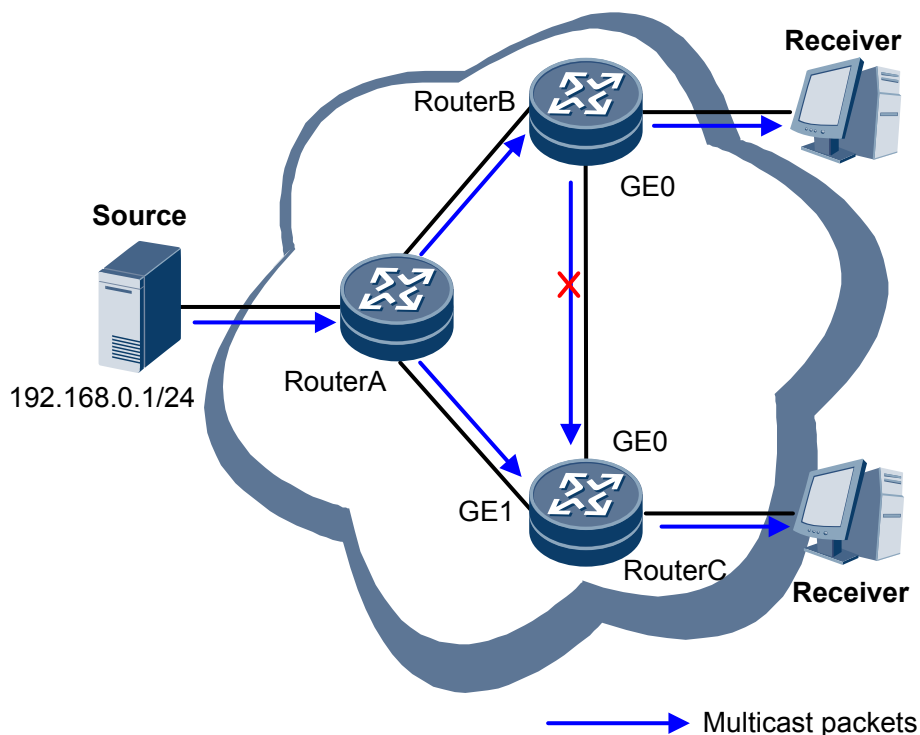
1. 如果组播转发表中不存在（S，G）表项，则对该报文执行 RPF 检查，将其 RPF 接口作为上游接口，结合相关路由信息创建相应的表项，并下发到组播转发表中：
  - 若该报文实际到达的接口正是其 RPF 接口，则 RPF 检查通过，向所有的下游接口转发该报文。
  - 若该报文实际到达的接口不是其 RPF 接口，则 RPF 检查失败，丢弃该报文。
2. 如果组播转发表中已存在（S，G）表项，且该报文实际到达的接口与上游接口相匹配，则向所有的下游接口转发该报文。
3. 如果组播转发表中已存在（S，G）表项，但该报文实际到达的接口与上游接口不匹配，则对此报文执行 RPF 检查：
  - 若其 RPF 接口与上游接口一致，则说明（S，G）表项正确，丢弃这个来自错误路径的报文。
  - 若其 RPF 接口与上游接口不符，则说明（S，G）表项已过时，于是把上游接口更新为 RPF 接口。
  - 如果该报文实际到达的接口正是其 RPF 接口，则向所有的下游接口转发该报文，否则将其丢弃。

## 典型示例

如图 12-45 所示，网络中单播路由可达，RouterC 上未配置组播静态路由。组播报文（S，G）沿从组播源（Source）到接收者（Receiver）的 SPT 进行传输。假定 RouterC 上的组播转发表中已存在（S，G）表项，其记录的上游接口为 GE1。



图 12-45 RPF 检查过程



RouterC 上的 IP 路由表:

- Destination/Mask: 192.168.0.0/24
- Interface: GE1
- 如果该组播报文从接口 GE1 到达 RouterC, 与 (S, G) 表项的上游接口相匹配, 则向所有的下游接口转发该报文。
- 如果该组播报文从接口 GE0 到达 RouterC, 与 (S, G) 表项的上游接口不匹配, 则对其执行 RPF 检查: 通过查找单播路由表发现到达 Source 的下游接口 (即 RPF 接口) 是 GE1, 与 (S, G) 表项的上游接口一致。这说明 (S, G) 表项是正确的, 该报文来自错误的路径, RPF 检查失败, 于是丢弃该报文。

### 12.6.4.5 组播静态路由

组播静态路由是 RPF 检查的重要依据。通过配置组播静态路由, 用户可以在当前路由器上为特定“报文源”指定 RPF 接口和 RPF 邻居。

组播静态路由的作用只在于影响 RPF 检查, 而不能用于指导组播数据转发, 故又称为 RPF 静态路由。

组播静态路由仅在所配置的组播路由器上生效, 不会以任何方式被广播或者引入给其它路由器。

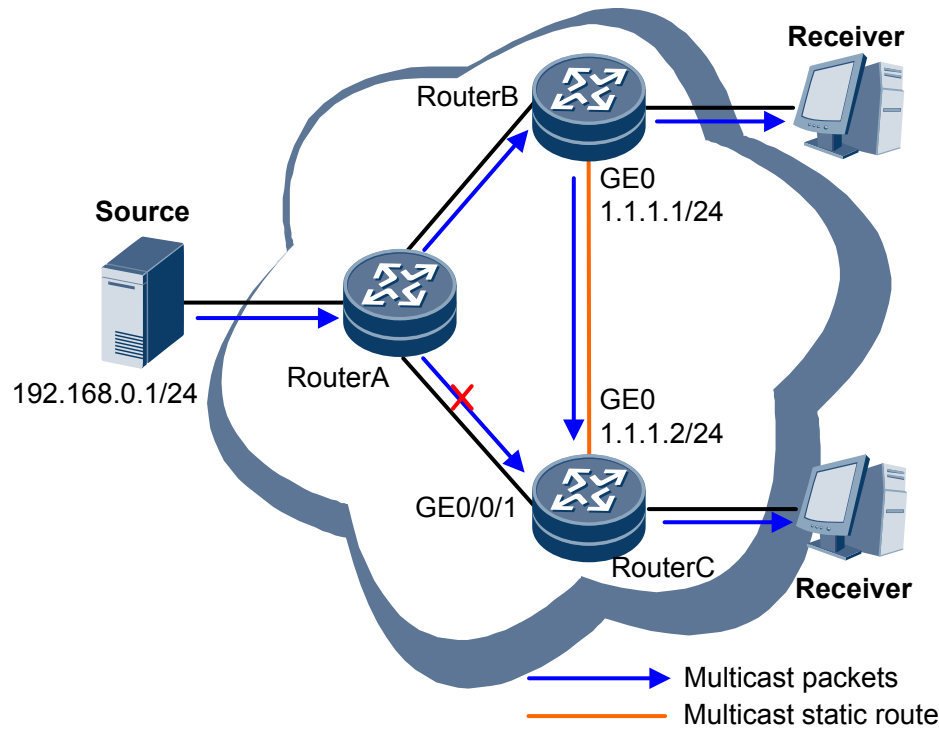
### 改变 RPF 路由

通常, 组播的网络拓扑结构与单播相同, 组播数据的传输路径也与单播相同。



可以通过配置组播静态路由以改变 RPF 路由，从而为组播数据创建一条与单播不同的传输路径。

图 12-46 改变 RPF 路由示意图



| RPF neighbor/Mask                         | Interface | RPF neighbor/Mask |
|-------------------------------------------|-----------|-------------------|
| Multicast Routing Table Static on RouterC |           |                   |
| 192.168.0.0/24                            | GE0       | 1.1.1.1/24        |

如图 12-46 所示，当网络中没有配置组播静态路由时，RouterC 到组播源（Source）的 RPF 邻居为 RouterA，从 Source 发出的组播信息沿 RouterA—RouterC 的路径传输，与单播路径一致。

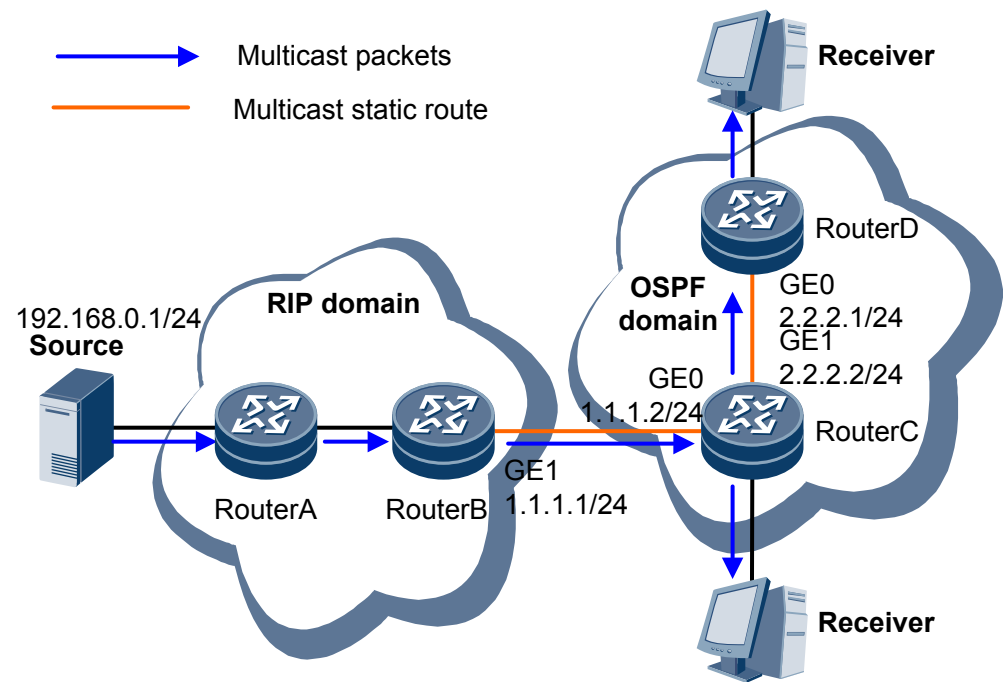
当在 RouterC 上配置了组播静态路由，指定从 RouterC 到 Source 的 RPF 邻居为 RouterB 之后，从 Source 发出的组播信息将改变传输路径，沿 RouterA—RouterB—RouterC 的新路径传输。

衔接 RPF 路由

当网络中的单播路由被阻断时，由于没有 RPF 路由而无法进行包括组播数据在内的数据转发。

可以通过配置组播静态路由以生成 RPF 路由，从而创建组播路由表项以指导组播数据的转发。

图 12-47 衔接 RPF 路由示意图



| RPF neighbor/Mask                         | Interface | RPF neighbor/Mask |
|-------------------------------------------|-----------|-------------------|
| Multicast Routing Table Static on RouterC |           |                   |
| 192.168.0.0/24                            | GE0       | 1.1.1.1/24        |
| Multicast Routing Table Static on RouterD |           |                   |
| 192.168.0.0/24                            | GE0       | 2.2.2.2/24        |

如图 12-47 所示，RIP 域与 OSPF 域之间实行单播路由隔离。当网络中没有配置组播静态路由时，OSPF 域内的接收者（Receiver）不能收到 RIP 域内的组播源（Source）所发出的组播信息。

当在 RouterC 和 RouterD 上均配置了组播静态路由，分别指定从 RouterC 到 Source 的 RPF 邻居为 RouterB、从 RouterD 到 Source 的 RPF 邻居为 RouterC 之后，Receiver 便能收到 Source 发出的组播信息了。

12.6.4.6 组播路由匹配策略

当组播路由器选择上游接口时，若存在多个开销相同的单播路由，根据具体的网络需要，用户可以从以下三种方法中任选其一，设置组播路由器。

- 缺省情况下，选取下一跳地址最大的路由。
- 配置按照最长匹配原则，选取目的地址与“报文源”地址匹配最精确的路由。

- 配置各路由之间的负载分担。根据组播源、组播组或组播源/组来进行组播流量的负载分担，可以优化存在多个组播数据流时的网络流量。

#### 12.6.4.7 控制组播转发

##### 定义组播转发范围

组播信息在网络中的转发并不是漫无边际的，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。目前 USG 有三种方式定义组播转发范围：

- 在接口上配置组播转发边界，以形成一个封闭的组播转发区域。当组播路由器接口配置了针对某组播组的转发边界之后，将不能再发出或接收该组播组的报文。
- 在接口上配置组播转发 TTL 阈值，以限定组播报文被转发的距离。接口只转发 TTL 不小于阈值的报文（包括本机创建的），若报文 TTL 值小于阈值，则丢弃该报文。
- 禁止转发平面接收组播数据报文，则组播路由器不能在指定接口上接收组播数据报文，从而限制了组播转发范围。

##### 限制组播转发表中的表项数量

组播路由器为每一个接收到的组播报文都维护有对应的转发表项。但是，过量的组播转发表项可能会耗尽路由器内存。

USG 允许用户自己定义设备组播转发表项的最大数量。根据实际组网情况和业务性能对表项数量进行适当的限制，可以避免由此引发的设备故障。

##### 限制单个转发表项的下游接口数量

组播路由器为每一个下游接口复制一份组播报文，并发送出去。每一个下游接口形成组播分发树的一路分支。下游接口数决定了组播分发树可能达到的最大规模，和组播服务范围。

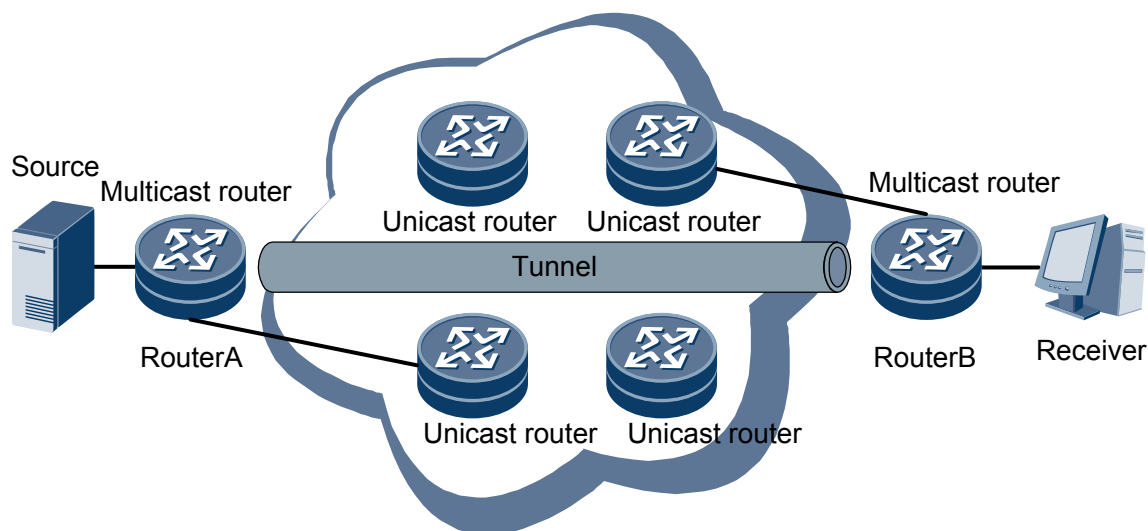
USG 允许用户自己定义单个转发表项的下游接口数量。根据实际组网情况和业务性能对下游接口数量进行适当的限制，可以缓解设备的处理压力，并控制组播服务范围。

#### 12.6.4.8 GRE 隧道在组播转发中的应用

网络中可能存在不支持组播协议的路由器。从组播源（Source）发出的组播数据沿组播路由器逐跳转发，当下一跳路由器不支持组播协议时，组播转发路径将被阻断。

此时，通过在处于单播网段两端的组播路由器之间建立 GRE（Generic Routing Encapsulation，通用路由封装）隧道，可以实现跨越单播网段的组播数据交换。有关 GRE 隧道的详细介绍，请参见《配置指南 MPLS 与 VPN》中的“配置 GRE”。

图 12-48 使用隧道方式传送组播数据



如图 12-48 所示，在 RouterA 和 RouterB 之间建立起 GRE 隧道。RouterA 将组播数据包封装在单播 IP 报文中，经由单播路由器转发，传送到隧道另一端的 RouterB。然后，RouterB 将单播 IP 报文头解封装，继续进行组播传输。

如果在隧道的两端配置了单播静态路由，则任意单播数据包都可以通过该隧道传输。为了将该隧道专用于组播数据包的传输，可以在隧道两端只配置组播静态路由，从而使单播数据包的传输不能再使用此隧道。

# 13 QoS

---

## 关于本章

### 13.1 流量监管、流量整形和接口限速

介绍了流量监管、流量整形、接口限速的基本概念原理。

### 13.2 拥塞管理

拥塞管理是一种当拥塞发生时通过制定资源的调度策略，从而决定报文转发时的处理次序的机制。

### 13.3 拥塞避免

拥塞避免通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞有加剧的趋势时，主动丢弃报文，通过调整网络的流量来解除网络过载。

### 13.4 基于类的 QoS

介绍基于类的 QoS 的基本概念，以及如何实现基于类的流量监管、流量整形、拥塞管理、重标记等功能。

### 13.5 链路效率机制

链路效率机制通过 LFI 和 IPHC 两种链路效率机制改善链路的性能，间接提高网络的 QoS。

### 13.6 VPN QoS

VPN QoS 应用在 Tunnel 接口上，用于在报文未被 GRE 等隧道封装之前，对报文进行预分类，并对分类后的报文进行 QoS 处理，保证隧道中报文的服务质量。

### 13.7 HQoS

HQoS 是一种既能控制用户的流量，又能同时根据用户的业务进行调度的 QoS 技术，多级的 QoS 调度提供了精细化的服务质量保证。

### 13.8 二层 QoS

介绍 QoS 应用于二层接口卡上的接口限速、优先级映射和拥塞管理技术。

## 13.1 流量监管、流量整形和接口限速

介绍了流量监管、流量整形、接口限速的基本概念原理。

### 13.1.1 介绍

#### 定义

流量监管 TP (Traffic Policing) 监督流入或流出网络的某一流量的规格, 把它限制在一个合理的范围之内, 或者对超出的部分流量进行“惩罚”, 以保护网络资源和运营商的利益。

流量整形 TS (Traffic Shaping) 限制流出某一网络的某一连接的流量, 使这一流量的报文以比较均匀的速度向外发送, 是一种主动调整流量输出速率的措施。

接口限速 LR (Limit Rate) 可以限制接口上发送报文的总速率。可分为二层 LR 和三层 LR:

- 二层 LR: 在链路层的限速, 对二层接口入方向和出方向流量进行限速。
- 三层 LR: 用于对三层接口出方向流量的限速, 和拥塞管理结合使用, 配置拥塞管理时, 需在接口上同步配置 LR。

#### 目的

流量监管: 通过设置特定数据流的规格, 对超规格数据流进行限制和惩罚, 具体包括将流量的规格限制在一个合理的范围内或根据设置的动作进行转发、丢弃等处理, 防止由超规格突发数据流引发的网络拥塞问题。

流量整形: 限制流出某一网络的某一连接的流量, 将超出规格的报文缓存在队列中, 使用令牌桶进行发送控制, 使这一流量的报文以比较均匀的速度向外发送。

接口限速: 对接口上通过的所有报文的总速率进行限制, 如果超过限制, 报文将进入 QoS 队列进行拥塞管理。

### 13.1.2 参考标准和协议

本特性的参考资料清单如下:

- RFC 2697: A Single Rate Three Color Marker

### 13.1.3 可获得性

#### License 支持

不需要 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                      |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br><br>USG5500 V200R001C00 |

13.1.4 原理描述

13.1.4.1 令牌桶

令牌桶可以看作是一个存放一定数量令牌的容器，通过衡量桶中的令牌数量是否足够满足报文的转发来评估流量规格。

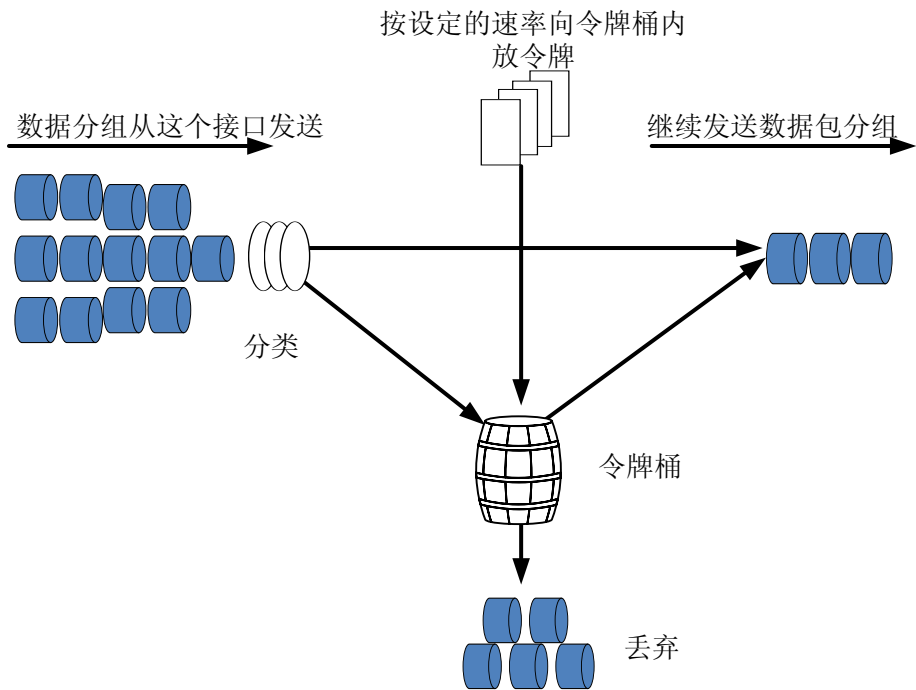
令牌用于控制数据的传输，拥有令牌就意味着设备可以传输信息。令牌桶功能就是通过按一定速率向桶中放置令牌，实现对数据传输的控制。

流量监管、流量整形和接口限速的实现过程中，都使用到了令牌桶。

令牌桶工作原理

图 13-1 为利用令牌桶评估报文流量的过程。

图 13-1 利用令牌桶评估报文流量



系统按设定的速率向令牌桶中放置令牌，如果桶中存在的令牌足够用来发送报文，则报文允许通过，令牌桶中的令牌量随报文的发送而相应减少。如果桶中的令牌少到不足以发送报文时，报文将被丢弃。直到桶中新放置了令牌，报文才可以被发送出去。

当桶中令牌满时，多出的令牌将溢出，桶中令牌不再增加。

通过令牌桶技术，不仅可以对流量进行限制，限制报文的流量不能大于令牌生成的速度，同时起到了允许突发流量通过的目的。

## 令牌桶常用参数

令牌桶中包括以下几种参数用于评估流量：

- CIR（Committed Information Rate，承诺信息速率）：系统向桶中放置令牌的速率，即允许的流的平均速度。
- CBS（Committed Burst Size，承诺突发尺寸）：每次突发所允许的最多的流量尺寸。

### 13.1.4.2 流量监管

流量监管 TP（Traffic Policing）通过设置特定数据流的规格，对超规格数据流进行限制和惩罚，防止由超规格突发数据流引发的网络拥塞问题。

通常的做法是使用承诺访问速率 CAR（Committed Access Rate）来限制某类报文的流量，首先，根据预先设置的匹配规则来对报文进行分类，如果是符合流量规定的报文，就直接继续发送，如果是超出流量规定的报文，则会被丢弃或重新标记。

## 流量监管处理过程

1. 使用承诺访问速率 CAR 限制某类报文的流量。
2. 设置动作，分别关联匹配规则的报文和不匹配规则的报文。具体动作包括转发、丢弃、改变优先级并转发等。

### 13.1.4.3 流量整形

流量整形 TS（Traffic Shaping）限制流出某一网络的某一连接的流量，使这一流量的报文以比较均匀的速度向外发送。

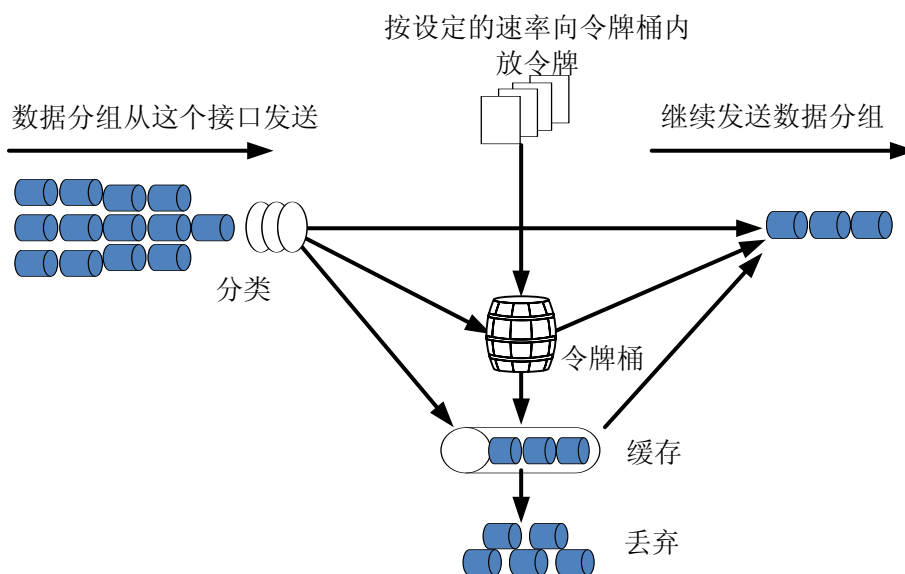
当网络中上下游的接口带宽不匹配，尤其是上游接口的带宽资源大于下游的带宽时，容易在下游网络形成拥塞。

如果不希望下游网络因为上游发送数据流量过大造成拥塞，可在上游的出接口配置流量整形，使上游发送的流量与下游接收的能力匹配。配置流量整形功能后，报文可以均匀向外发送，而且超出规格的部分不会直接被丢弃，而是被缓存起来，到可以发送的时再被发送出去。

流量整形是一种主动调整流量输出速率的措施，通常使用缓冲区和令牌桶来完成，如果对需要丢弃的报文进行缓冲，将可以减少报文的丢弃，同时满足报文的流量特性。这就是通用流量整形（General Traffic Shaping，通用流量整形，以后简称 GTS）。当报文的发送速度过快时，首先在缓冲区进行缓存，在令牌桶的控制下，再均匀的发送这些被缓存的数据。其处理过程如图 13-2 所示。



图 13-2 利用令牌桶实现流量整形



## 流量整形处理过程

图 13-2 为流量整形的基本处理过程，其所处理的流量为三层流量。

当数据到来的时候，首先对数据进行分类，按照以下几种情况进行处理。

- 不参与流量整形处理的分组，直接发送。
- 对于参与流量整形的分组，当不存在 GTS 队列时，将分组规格与令牌桶内的令牌数目进行比较。
  - 如果令牌桶中有足够的令牌可以用来发送报文，则报文直接被继续发送下去。
  - 如果令牌桶中没有足够的令牌，则产生 GTS 队列，分组缓存在队列中。令牌桶按用户设定的速度向桶中放置令牌，在分组被发送的同时，令牌做相应的减少。
- 对于参与流量整形的分组，当有 GTS 队列存在的时候，直接进入队列，等待 GTS 队列按照一定周期对队列中的分组进行调度转发。
- 当分组到来而发现 GTS 队列已满时，分组被丢弃。

## 流量整形与流量监管的区别

流量整形与流量监管的主要区别在于：

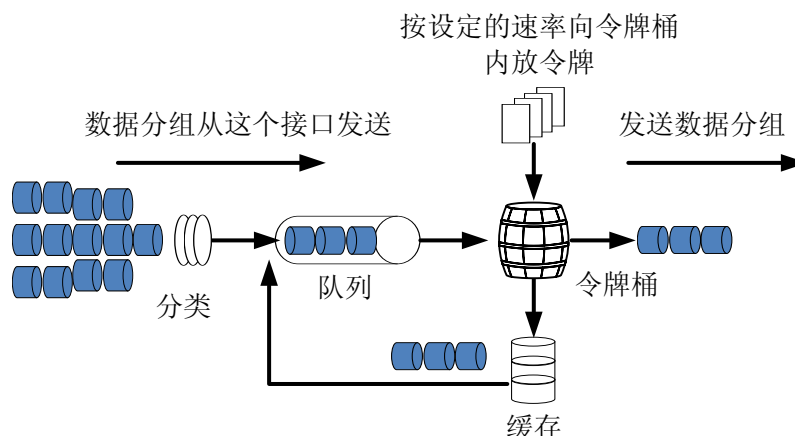
- 流量整形对流量监管中超出流量规格的报文进行缓存。当令牌桶有足够的令牌时，再均匀的向外发送这些被缓存的报文。
- 流量整形可能会增加延迟，而流量监管几乎不引入额外的延迟。

### 13.1.4.4 接口限速

接口限速 LR (Limited Rate) 是对接口上通过的所有报文的总速率进行限制的一种机制。

接口限速采用令牌桶进行流量控制，其处理过程如图 13-3 所示。

图 13-3 利用令牌桶实现接口限速



### 三层接口限速处理过程

在设备的某个接口上配置了 LR，则所有需要通过此接口进行发送的报文都要经过 LR 的令牌桶进行处理。

- 当令牌桶中的令牌足够时，报文可以顺利发送。
- 当令牌桶中没有足够的令牌时，报文进入 QoS 队列进行拥塞管理。

### 接口限速与流量监管、流量整形的对比

- 与流量监管相比，接口限速能够限制在物理接口上通过的所有报文。而流量监管是在 IP 层实现的，所以对于不经过 IP 层处理的报文不起作用。
- 与流量整形相比，接口限速不但能够对超过流量限制的报文进行缓存，而且报文进入了 QoS 队列机制进行处理，队列调度机制更灵活。而流量整形是在 IP 层实现的，所以对于不经过 IP 层处理的报文不起作用。
- 在用户只要求对所有报文进行限速时，可以使用配置操作比较简单的接口限速功能。对于网络建设投资者，使用接口限速可以对客户隐藏实际带宽，客户只能严格按照所购买的带宽来使用。

## 13.2 拥塞管理

拥塞管理是一种当拥塞发生时通过制定资源的调度策略，从而决定报文转发时的处理次序的机制。

### 13.2.1 介绍

#### 定义

拥塞管理是一种当拥塞发生时通过制定资源的调度策略，从而决定报文转发时的处理次序的机制。主要调度策略包括 FIFO、CQ、PQ、WFQ、RTPQ、CBWFQ 等。

#### 目的

对于网络单元，当分组到达的速度大于出接口发送分组的速度时，在出接口处就会产生拥塞。如果没有足够的存储空间来保存这些分组，它们其中的一部分就会被丢弃。分组

的丢弃又可能会导致发送该分组的主机或设备因超时而重传此分组，这样将会形成恶性循环。

拥塞管理的中心内容就是当拥塞发生时如何制定一个资源的调度策略，以决定报文转发时的处理次序。拥塞管理的处理包括队列的创建、报文的分类、将报文送入不同的队列、队列调度等。

13.2.2 可获得性

License 支持

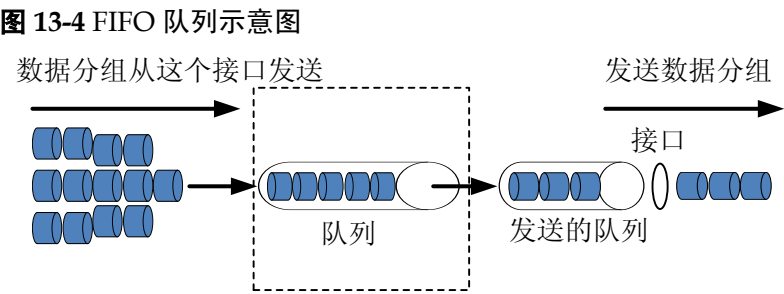
不需要 License 支持。

版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

13.2.3 原理描述

13.2.3.1 FIFO 队列



当网络中传输的业务相对简单，对实时性、可靠性要求不高时，可以采用先进先出队列 FIFO（First In First Out）作为队列调度机制。

FIFO 按照分组到达的先后决定报文的转发次序，先进的先出，后进的后出，不进行流分类和队列调度。

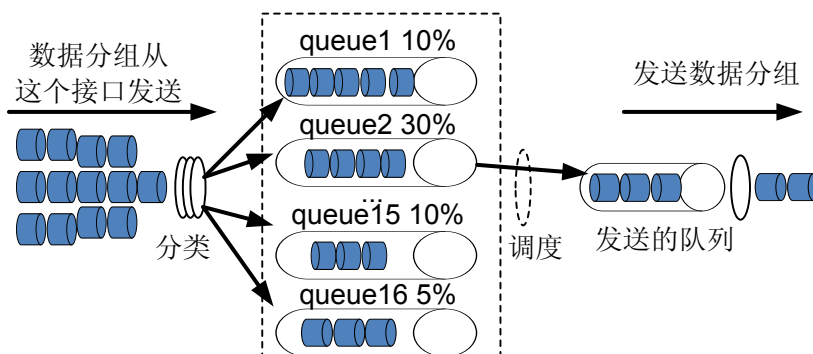
FIFO 关心的只是队列的长度以及队列的长度对延迟和丢包率的影响。用户的业务在某个设备上能够获得的资源取决于报文到达的时机和当时的负载情况。

传统网络服务采用的尽力而为（Best-Effort）策略的报文传送方式采用的就是 FIFO。

FIFO 是三层接口缺省使用的队列调度机制。

### 13.2.3.2 CQ 队列

图 13-5 CQ 队列示意图



当网络发生拥塞时，如果希望关键业务获得较多的带宽，保证服务质量，同时保证定制小带宽用户的非关键业务也可以获得一定的带宽，可以使用定制队列 CQ（Custom Queue）作为队列调度机制。

CQ 按照一定的规则将分组分成 16 类（对应于 16 个队列），分组根据自己的类别按照先进先出的策略进入相应的 CQ 队列。

CQ 的 16 个队列，如图 13-5 所示。用户可以配置流分类的规则，指定每个队列调度时的字节数，根据配置的字节数可以计算出每个队列占用带宽的比例。

在队列调度时，采用轮询的方式按照预先配置的字节数依次从 1 到 16 号队列中取出一定数量的分组发送出去。对每一个队列每次都要取出足够的字节数甚至是超出配置的字节数，队列内的字节数小于队列调度配置的字节数时，将队列取空为止。这样，就可以使不同业务的分组获得不同的带宽，既可以保证关键业务能获得较多的带宽，又不至于使非关键业务得不到带宽。缺省情况下，数据流进入 1 号队列。

#### 说明

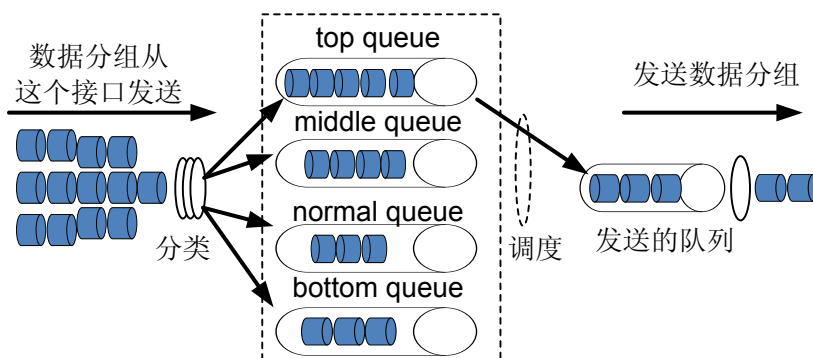
CQ 不能对报文进行分片，CQ 在轮询调度时，每次要取够配置的字节数，当前报文字节数不够时，要从下一个报文来取，因为不能分片，所以要将下一个报文整体取出，这样有可能导致每次取得字节数大于原来设置的字节数。

定制队列的另一个优点是：可根据业务的繁忙程度分配带宽，适用于对带宽有特殊需求的应用。虽然 16 个用户队列的调度是轮询进行的，但对每个队列不是固定地分配服务时间片——如果某个队列为空，那么马上换到下一个队列调度，如果某个队列一段时间为空，这个空的队列所占用的带宽资源被剩余队列重新按比例划分。因此，当没有某些类别的报文时，CQ 调度机制能自动增加现存类别的报文可占的带宽。举个例子来说：当前配置了三个队列 1，2，3，配置三个队列调度的字节数为 200，300，500，三个队列发送的字节数占用的带宽比例为：20%、30%、50%，假如某一段时间队列 1 为空，那么队列 2 和 3 占用的实际带宽就为  $1000 \times 30\% / (30\% + 50\%) = 375$  和  $1000 \times 50\% / (30\% + 50\%) = 625$ ，当没有关键业务时，非关键业务可以占用整个带宽。

因为 CQ 是采用 16 个队列轮询调度，不像 PQ（Priority Queue）那样将绝对优先权赋予优先级高的报文，所以在传送语音等实时性较高的业务流时，并不能保证低延迟，不宜采用。对于带宽的保证也没有 CBWFQ 那样灵活、准确。

### 13.2.3.3 PQ 队列

图 13-6 PQ 队列示意图



PQ (Priority Queue) 队列是针对关键业务应用而设计的。关键业务有一个重要的特点，即在拥塞发生时要求优先获得服务以减小响应的延迟。PQ 可以根据网络协议（比如 IP，MPLS）、接受报文的接口、报文长短、源地址/目的地址等灵活地指定优先次序。

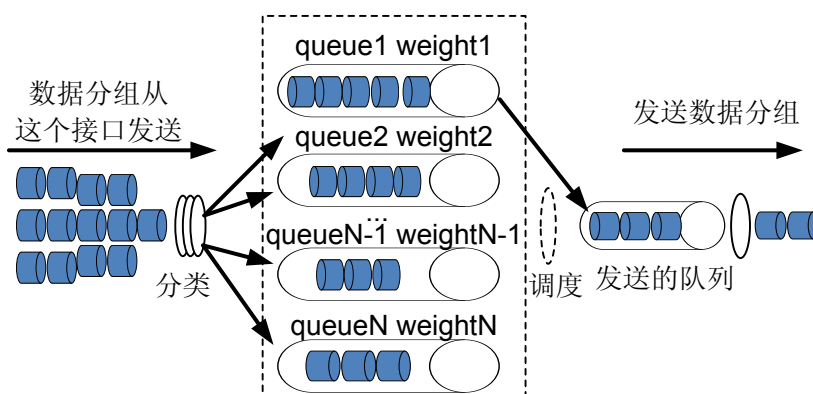
如图 13-6，优先队列将报文分入 4 类队列，分别为高优先队列（top queue）、中优先队列（middle queue）、正常优先队列（normal queue）和低优先队列（bottom queue），它们的优先级依次降低。缺省情况下，数据流进入 normal 队列。每个队列内部又遵循 FIFO 原则。

在队列调度时，PQ 严格按照优先级从高到低的次序，优先发送较高优先级队列中的分组，当较高优先级队列为空时，再发送较低优先级队列中的分组。这样，将关键业务的分组放入较高优先级的队列，将非关键业务的分组放入较低优先级的队列，可以保证关键业务的分组被优先传送，使关键业务获得较大带宽、较短延迟、较小抖动，非关键业务的分组在处理关键业务数据的空闲间隙被传送。这样既保证了关键业务的优先，又充分利用了网络资源。

PQ 的缺点是如果较高优先级队列中长时间有分组存在，那么低优先级队列中的分组将可能一直得不到服务。

### 13.2.3.4 WFQ 队列

图 13-7 WFQ 队列示意图



在介绍加权公平队列（Weighted Fair Queue）前，先要理解公平队列 FQ（Fair Queue）。FQ 是为了公平地分享网络资源，尽可能使所有流的延迟和延迟抖动达到最优而设计的。它照顾了各方面的利益，主要表现在：

- 不同的队列获得公平的调度机会，从总体上均衡各个流的延迟。
- 短报文和长报文获得公平的调度。如果不同队列间同时存在多个长报文和短报文等待发送，应当顾及短报文的利益，让短报文优先获得调度，从而在总体上减少各个流的报文间的延迟抖动。

与 FQ 相比，WFQ 在计算报文调度次序时增加了优先级方面的考虑。从统计上，WFQ 使高优先权的报文获得优先调度的机会多于低优先权的报文。

WFQ 是基于流的“会话”信息自动进行流分类，WFQ 不允许用户根据 ACL 对报文进行分类，只能根据系统定义类的流匹配规则来进行流分类，对于 IP 网络，具有相同源 IP 地址、目的 IP 地址、源端口号、目的端口号和 IP 优先级的报文属于同一个流，而对于 MPLS 网络，具有相同的标签和 EXP 域值的报文属于同一个流。

每一个流被分配到一个队列，该过程称为散列，采用 HASH 算法来自动完成，尽量将不同的流分入不同的队列，从而在总体上均衡各个流的延迟。在出队的时候，WFQ 按流的优先级（precedence）来分配每个流应占有出口的带宽。优先级的数值越小，所得的带宽越少。

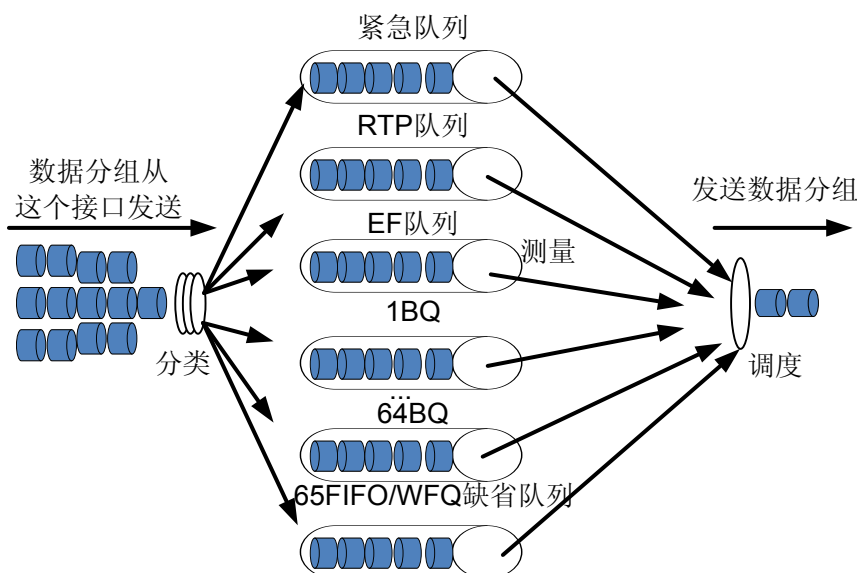
例如：接口中当前共有 5 个流，它们的优先级分别为 0、1、2、3、4，则带宽总配额为：所有（流的优先级+1）的和，即：1 + 2 + 3 + 4 + 5 = 15。

每个流所占带宽比例为：（自己的优先级数 + 1） / （所有(流的优先级 + 1)的和）。即每个流可得的带宽分别为：1/15, 2/15, 3/15, 4/15, 5/15。

由于 WFQ 在拥塞发生时能均衡各个流的延迟和延迟抖动，所以 WFQ 在一些特殊场合得到了有效的应用。但是在需要严格保证语音等实时业务服务质量的情况下，不建议采用 WFQ。

### 13.2.3.5 CBWFQ 队列

图 13-8 CBWFQ 队列示意图



CBWFQ（Class Based WFQ）可以根据复杂的分类规则将数据流分为最多 64 个类。当网络发生拥塞时，不同类别的数据报文缓存进入相应的队列。对于不匹配 AF(Assured Forwarding)、EF(Expedited Forwarding)队列类的报文，将被送入系统定义的缺省队列。



CBWFQ 中包含如下队列：

- 一个低时延队列 EF 队列，用来支撑 EF 类业务，被绝对优先发送，保证时延。
- 最多 64 个 BQ（Bandwidth Queue）队列，用来支撑 AF 类业务，可以保证每一个队列的带宽及可控的时延。
- 一个 WFQ，作为缺省队列用来支撑 BE 业务，使用接口剩余带宽进行发送。

图 13-8 所示的 EF 队列是一个具有较高优先级的队列，它的优先级仅次于紧急队列与 RTP 优先队列。一个或多个配置了 EF 类的数据流可以被设定进入 EF 队列。不同类别的报文可设定占用不同的带宽。在调度出队的时候，若 EF 队列中有报文，则总是优先发送 EF 队列中的报文，直到 EF 队列中没有报文时，或者超过为 EF 队列配置的最大带宽时才调度发送其他队列中的报文。

在接口没有发生拥塞的时候，所有对应 EF 队列类的报文都可以被发送。在接口发生拥塞的时候（此时队列中有报文），进入 EF 队列的报文被限速，超出规定流量的报文将被丢弃。这样，在接口不发生拥塞的情况下，可以使属于 EF 队列的报文能够获得空闲的带宽；在接口拥塞的情况下，又可以保证属于 EF 队列的数据流量不会占用超出规定的带宽，保护了其他数据流的应得带宽。

由于只要 EF 队列中有报文，系统就会发送 EF 队列中的报文，所以 EF 队列中的报文被发送的延迟最多是接口发送一个最大长度报文的时间。无论是延迟还是延迟抖动，EF 队列都可以将之降低为最低限度。这为对延迟敏感的应用如 VoIP 业务提供了良好的服务质量保证。

图 13-8 中 1 到 64 的队列是为各种 AF 类的报文配置的队列，每个 AF 类报文占一个队列，我们称它们为 BQ。在系统调度报文出队的时候，按用户为各类报文设定的带宽进行调度。这种队列技术应用了先进的队列调度算法，可以实现各个类的队列的公平调度。属于 1 到 N 号 BQ 队列的报文可以被确保得到用户设定的带宽。当接口中某些类别的报文没有时，BQ 队列的报文还可以公平地得到空闲的带宽，和时分复用系统相比，大大提高了线路的利用率。同时，在接口拥塞的时候，仍然能保证各类数据流得到用户设定的最小带宽。

当数据流不匹配用户设定的所有类别时，报文则被送入系统定义的缺省类。虽然允许为缺省类配置带宽，使其作为 AF 类进行基于类的队列调度，但是更多的情况是为缺省类配置 WFQ，使所有进入缺省类的报文进行基于流的队列调度。

CBWFQ 最多允许配置 64 个 BQ 类，缺省类的 WFQ 的队列个数可以由用户设定。

对于缺省类的 WFQ 和 BQ，当队列的长度达到队列的最大长度时，缺省采用尾丢弃的策略。但用户还可以选择用加权随机早期检测 WRED（Weighted Random Early Detection）的丢弃策略。

对于 EF 队列，由于在接口拥塞的时候流量限制开始起作用，所以用户不必设置队列的长度。另外，由于优先队列中的报文一般是 VoIP 语音报文，采用的是 UDP 报文，所以 WRED 的丢弃策略也不需要。

CBWFQ 可根据报文的输入接口、满足 ACL 情况、IP Precedence、DSCP、EXP、Label 等规则对报文进行复杂的分类、进入相应队列，分类规则通过构件化 QoS 命令行（CQC）进行手工配置。

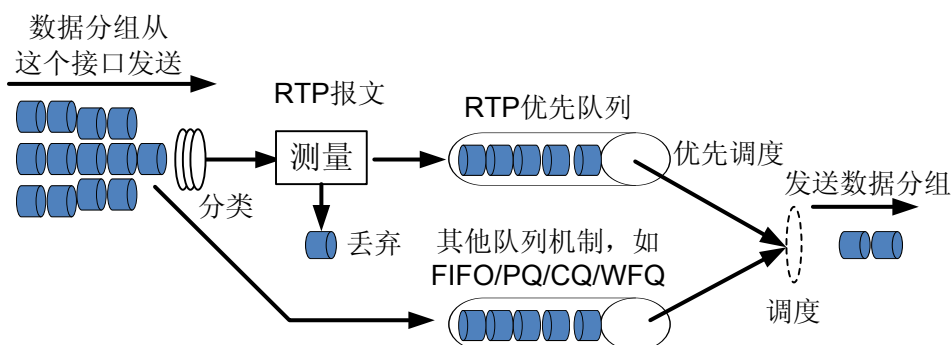
#### 说明

考虑到链路层控制报文的发送、链路层封装开销及物理层开销，建议 RTP 队列和 CBWFQ 占用接口的总带宽不要超过接口带宽的 80%。

CBWFQ 可为不同的业务提供不同的服务质量保证，为带宽、时延等指标提供完美的保证。但是由于涉及到复杂的流分类，启用 CBWFQ 特性，系统资源存在一定的开销。

### 13.2.3.6 RTP 队列

图 13-9 RTP 队列示意图



RTP 队列（Real-time Transport Protocol）是一种解决实时业务（包括语音与视频业务）服务质量的简单的队列技术。其原理就是将承载语音或视频的 RTP 报文送入高优先级队列，使其得到优先发送，保证延迟和抖动降低为最低限度，从而保证了语音或视频这种对时延敏感业务的服务质量。

如图 13-9 所示，RTP 队列将 RTP 报文送入一个具有较高优先级的队列，RTP 报文是端口号在一定范围内为偶数的 UDP 报文，端口号的范围可以配置，一般为 16384 ~ 32767。RTP 队列可以和前面所述的任何一种队列（包括 FIFO、PQ、CQ、WFQ 与 CBWFQ）结合使用，它的优先级是最高的。

由于对进入 RTP 优先队列的报文进行了限速，超出规定流量的报文将被丢弃，这样在接口拥塞的情况下，可以保证属于 RTP 优先队列的报文不会占用超出规定的带宽，保护了其他报文的应得带宽，解决了当较高优先级队列中长时间有分组存在，低优先级队列中的分组将可能一直得不到服务的问题。

## 13.3 拥塞避免

拥塞避免通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞有加剧的趋势时，主动丢弃报文，通过调整网络的流量来解除网络过载。

### 13.3.1 介绍

#### 定义

拥塞避免监视网络资源的使用情况，在拥塞趋势加剧时通过主动丢弃报文来避免网络过载情况的发生。

#### 目的

过度的拥塞会对网络资源造成极大危害，必须采取某种措施加以解除。拥塞避免，是指通过监视网络资源（如队列或内存缓冲区）的使用情况，在拥塞有加剧的趋势时，主动丢弃报文，通过调整网络的流量来解除网络过载的一种流量控制机制。

与源端的流量控制相比，这里的流量控制有更广泛的意义，它影响到设备中更多的业务流的负载。设备在丢弃报文时，需要与源端的流量控制动作配合（比如 TCP 流量控制），共同调整网络的流量到一个合理的负载状态。



## 13.3.2 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 13.3.3 原理描述

### 传统的丢包策略

传统的丢包策略采用尾部丢弃（Tail-Drop）的方法，当队列的长度达到最大值后，所有新到达的报文都将被丢弃。这种丢弃策略会引发 TCP 全局同步现象。

#### 说明

TCP 全局同步现象：当队列同时丢弃多个 TCP 连接的报文时，将造成多个 TCP 连接同时进入拥塞避免和慢启动状态以降低并调整流量，而后又会在某个时间同时出现流量高峰。如此反复，使网络流量忽大忽小。

### RED 与 WRED

使用 RED（Random Early Detection，随机早期检测）或 WRED（Weighted Random Early Detection，加权随机早期检测）可以避免 TCP 全局同步现象的发生。

USG 只支持 WRED。

在 RED 算法中，为每个队列都设定一对低限值和高限值，并规定：

- 当报文占用队列的长度小于低限值时，不丢弃报文。
- 当报文占用队列的长度超过高限值时，丢弃所有到来的报文。
- 当报文占用队列的长度在低限值和高限值之间时，开始随机丢弃到来的报文。

丢弃方法：用当前队列的丢弃概率与理论的丢弃概率（系统根据用户配置的参数计算出的理论的丢弃概率参考值）相比较，如果大于理论值则被丢弃。

队列越长，被丢弃的概率越高，但有一个最大的被丢弃概率。

与 RED 不同，WRED 丢弃概率的计算是基于优先权的。它引入 IP 优先级区别丢弃策略，考虑了高优先级报文的利益并使其被丢弃的概率相对较小。

RED 和 WRED 通过随机丢弃报文避免了 TCP 的全局同步现象。当某个 TCP 连接的报文被丢弃，开始减速发送的时候，其他的 TCP 连接仍然有较高的发送速度。这样，无论什么时候，总有 TCP 连接在进行较快的发送，提高了线路带宽的利用率。

## 队列平均长度

直接采用队列的长度（这是设置队列门限的绝对长度）和低限值、高限值比较并进行丢弃，会对突发性的数据流造成不公正的待遇，不利于数据流的传输。WRED 使用平均长度与低限值、高限值比较并进行丢弃。

队列的平均长度是队列长度被低通滤波后的结果。它既反映了队列的变化趋势，又对队列长度的突发变化不敏感，避免了对突发性数据流的不公正待遇。

## WRED 和 WFQ 队列配合使用

当 WRED 和 WFQ 队列配合使用时，可以实现基于流的 WRED。这是因为，在进行分类的时候，不同的流有自己的队列。对于流量小的流，由于其队列长度总是比较小，所以丢弃的概率将比较小。而流量大的流将会有较大的队列长度，从而丢弃较多的报文，保护了流量较小的流的利益。

## 13.4 基于类的 QoS

介绍基于类的 QoS 的基本概念，以及如何实现基于类的流量监管、流量整形、拥塞管理、重标记等功能。

### 13.4.1 介绍

#### 定义

基于类的 QoS 是指通过对流量按照某种规则进行分类，并将匹配到的流量关联某种动作，形成流量策略，将该策略应用后实现基于类的流量监管、流量整形、拥塞管理、重标记等功能。

配置基于类的 QoS 需要分为下面几个步骤：

1. 配置流分类：根据一定的规则将符合某类特征的流量分为一类。
2. 配置流行为：对某类流量指定的具体动作。  
USG 支持的流行为包括：重标记、流量监管、流量整形、基于类的 WRR、CBWFQ 和拥塞避免。
3. 配置流量策略：将相应的流和相应的流行为进行关联。
4. 应用流量策略：将流量策略应用到接口上。

#### 目的

配置基于类的 QoS，可以对网络中的流量进行分类管理，如对语音、视频、数据等分别对待，分配不同的带宽、保证不同的时延等。

### 13.4.2 可获得性

#### License 支持

不需要 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 13.4.3 原理描述

### 13.4.3.1 流分类

流分类是有区别地进行服务的前提和基础。

流分类规则可以使用 IP 报文头的 ToS（Type of Service，服务类型）字段的优先级位或差分服务编码点 DSCP（Differentiated Services Codepoint）识别出有不同优先级特征的流量，也可以由网络管理员设置流分类的策略，例如综合源地址、目的地址、MAC 地址、IP 协议或应用程序的端口号等信息对流进行分类。

分类的结果是没有范围限制的，它可以是一个由五元组（源地址、源端口号、目的地址、目的端口号、协议号码）确定的狭小范围，也可以是到某网段的所有报文。

其中，系统预定义了缺省类(default-class),匹配的是缺省数据流，用户定义策略时可直接使用此类，用户定义类名不允许为系统预定义类。

### 13.4.3.2 标记

配置优先级标记，可以对 IP 报文、MPLS 报文或 VLAN 报文中的优先级进行设置。

几种常见报文的优先级

- IP 优先级和 DSCP 优先级

- IP 优先级：使用 ToS 字段的前 3 个 bit，取值范围是 0 ~ 7。
- DSCP 优先级：在 RFC 2474 中，重新定义了 IP 报文头部的 ToS 域，称之为 DS（Differentiated Services，差分服务）域，其中 DSCP 优先级使用该域的前 6 个 bit（0 ~ 5）表示，取值范围为 0 ~ 63，后 2 位（6、7 位）是保留位。

Diff-Serv（Differentiated Services，差分服务）网络定义了三类流量，设备根据报文中的 DSCP 优先级对报文执行相应的动作：

- EF 类（Expedited Forwarding，加速转发）：这种方式不用考虑其他流量是否分享其链路，适用于低时延、低丢包率、低抖动、高带宽的优先业务。这类分组具有最高的优先级，可以得到最优先转发处理。
- AF 类（Assured Forwarding，确保转发）：又分为四个小类（AF1、AF2、AF3、AF4），每个 AF 小类又分为三个丢弃优先级，可以细分 AF 业务的等级。这类分组具有较高优先级，可以得到保证转发处理。AF 类的 QoS 等级低于 EF 类。

- BE 类（Best Effort，尽力转发）：没有任何保证，现有 IP 网络流量也都默认为此类。这类分组优先级最低，只作尽力而为的转发。
- **802.1p 优先级**  
802.1p 优先级位于二层报文头部，适用于不需要分析三层报头，而需要在二层环境下保证 QoS 的场合。  
802.1Q 标签头包含了 2 个字节的 TCI（Tag Control Information，标签控制信息），其中 TCI 中的前 3 个 bit 即为 802.1p 优先级，取值范围 0 ~ 7，共 8 个优先级。
- **EXP 字段**  
在 MPLS 标签中，3bit 的 EXP 字段用于优先级标记，取值范围是 0 ~ 7。
- **IPv6 TC 字段**  
IPv6 头部 8bit 的 TC（Traffic Class）字段，用于对 IPv6 报文进行匹配的场景。

针对不同报文的优先级标记动作

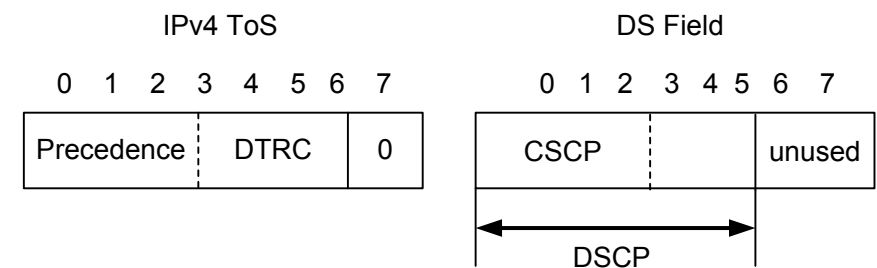
- 对于 IP 报文，优先级标记是对 IP 报文的 IP 优先级或者 DSCP 优先级进行设置。
- 对于 VLAN 报文，优先级标记是对 VLAN 报文中的 8021P 域进行设置。
- 对于 MPLS 报文，优先级标记是对 MPLS 报文中的 EXP 域进行设置。
- 对于 IPv6 报文，优先级标记是对 IPv6 报文的 TC 字段进行设置。

13.4.3.3 DSCP

在 RFC791、RFC134 和 RFC1349 中定义了 IPv4 报文头的 ToS（Type of Service）字段，如图 13-10 左所示，ToS 字段包含 3bits 的优先级（Precedence）、D bit、T bit、R bit 和 C bit，ToS 字段的最高位 bit 必须为 0。D bit 代表延迟（Delay），T bit 代表吞吐量（Throughput），R bit 代表可靠性（Reliability），C bit 代表花费（Cost）。在实施 QoS 时，设备会检查报文的优先级。其余的比特位未使用。

在 RFC2474 中对 IPv4 报文头的 ToS 字段进行了重新定义，称为 DS（Differentiated Services）字段。如图 13-10 右所示，DS 字段的低 6 位（0 ~ 5 位）用作区分服务代码点 DSCP（DS CodePoint），高 2 位（6、7 位）是保留位。DS 字段的低 3 位（0 ~ 2 位）是类选择代码点 CSCP（Class Selector CodePoint），相同的 CSCP 值代表一类 DSCP。DS 节点设备根据 DSCP 的值选择相应的 PHB（Per-Hop Behavior）。

图 13-10 ToS 字段和 DS 字段



13.4.3.4 标准的 PHB

PHB（Per Hop Behavior）用来描述拥有相同 DSCP 值的报文的下一步转发动作。一般 PHB 包括延时、丢包率等流量特性。

目前，IETF 定义了三种标准的 PHB：加速转发 EF（Expedited Forwarding）、确保转发 AF（Assured Forwarding）和尽力而为 BE（Best-Effort），BE 是缺省的 PHB。

EF 被定义为这样的一种转发处理：从任何 DS 节点发出的信息流速率在任何情况下必须获得等于或大于设定的速率。被标记为 EF PHB 的流量在同一个 DS 域内的节点上不能重新标记 DSCP 值。仅允许在 DS 域的边界节点上重新标记报文的 DSCP 值，并且要求新的 DSCP 满足 EF PHB 的特性。

AF 的推出是为了满足这样的需求：用户在与 ISP 订购带宽服务时，ISP 允许业务量超出所订购的规格。对不超出所订购规格的流量确保转发的质量，对超出规格的流量将降低服务待遇继续转发，而不只是简单地被丢弃。

BE 即传统的 IP 分组投递服务，只关注可达性，其他方面不做任何要求。任何路由器必须支持 BE PHB。

### 13.4.3.5 基于类的 CBWFQ

CBWFQ 为用户提供了定义类的支持。在网络拥塞时，CBWFQ 对输出报文根据用户定义类规则进行匹配，并使其进入相应的队列，在入队列之前进行拥塞避免机制（尾部丢弃或 WRED）和带宽限制的检查。在报文出队列时，调度每个类对应的队列中的报文。

CBWFQ 提供三种类型的队列：EF、AF 和 BE 队列，分别满足低时延业务，需要带宽保证的关键数据业务，以及不需要严格 QoS 保证的尽力发送业务。

### 13.4.3.6 策略的实现

基于类的 QoS 的实现，体现在策略的应用。一个 QoS 策略中关联了流分类和流行为。将策略应用在接口上，就可以提供相应的服务质量保证。

实现策略的具体步骤如下：

1. 定义流分类
2. 定义流行为
3. 定义策略，关联流分类和流行为
4. 在接口上应用策略

## 流分类

流分类规则可以使用 IP 优先级或 DSCP 值等，识别出有不同优先级特征的流量；也可以由网络管理者设置流分类的规则，例如综合源 IP 地址、目的 IP 地址、MAC 地址、IP 协议或应用程序的端口号等信息对流进行分类。要对网络中的流量进行基于类的 QoS 应用，就需要先定义流分类。

具体实现如下：

1. 定义流分类的名称
2. 设定流分类的匹配规则

其中，系统预定义了缺省类(default-class),匹配的是缺省数据流，用户定义策略时可直接使用此类，用户定义类名不允许为系统预定义类。

定义流分类匹配规则不能递归使用。例如，traffic classifier A 定义了匹配 traffic classifier B 的规则，traffic classifier B 不能再直接或者间接的引用 traffic classifier A。

## 流行为

基于流分类的动作，可以是重新标记报文的优先级，也可以是流量监管、流量整形、队列调度、拥塞避免。

具体实现如下：

1. 定义行为名
2. 定义具体流行为

## 定义策略

定义策略就是要将定义的流分类和定义的动作关联起来。

具体实现如下：

1. 定义策略名
2. 关联流分类和流行为

## 应用策略

应用策略，是要将定义好的策略应用到具体的接口上，分为出接口和入接口，可以根据具体的行为决定是将策略应用在接口的出方向还是入方向，比如流量监管，既可以是出方向也可以是入方向，流量整形只能应用在接口的出方向上。

# 13.5 链路效率机制

链路效率机制通过 LFI 和 IPHC 两种链路效率机制改善链路的性能，间接提高网络的 QoS。

## 13.5.1 介绍

### 定义

链路效率机制通过 LFI（Link Fragmentation and Interleaving）和 IPHC（IP Header Compression）两种链路效率机制改善链路的性能，间接提高网络的 QoS。

#### 链路分片与交叉（LFI）

在低速串行链路上进行实时交互式通信，如 Telnet 和 VoIP，往往会由于发送大型分组而导致阻塞延迟。

为了在相对低速的链路上缩短实时报文的延迟时间，需要将大型报文进行分片，将小报文和大报文的分片一起加入到队列。

#### IP 报文头压缩（IPHC）

IPHC 支持对 IP/TCP 头和 IP/UDP/RTP 头的压缩。

在 NGN（Next Generation Network）承载网络中，很多运营商都存在传输资源缺乏的情况。在 IP NGN 的业务中，RTP（Real Time Transport Protocol）/UDP/IP 报文头有接近 40 字节，但实际每个报文中的语音数据（采用好的语音压缩算法）一般都小于 30 字节。这样，报文开销字节太大，传输效率较低。

使用 IP 报文头压缩协议对 RTP 头进行压缩，可以将 IP/UDP/RTP 头从 40 字节压缩到 2 ~ 4 字节，压缩比为  $(40+40)/(40+4)$ ，约为 1.82。

## 目的

应用 LFI 技术可以将大型数据报文分割成小型报文，与其他小片的报文一起发送，从而减少在速度较慢的链路上的延迟和抖动。

将 IPHC 用在 IP 网络上承载语音、视频等实时多媒体业务，可以提高链路的利用效率。

## 13.5.2 参考标准和协议

本特性的参考资料清单如下：

- RFC1661: The Point-to-Point Protocol (PPP)
- RFC1055: A NONSTANDARD FOR TRANSMISSION OF IP DATAGRAMS OVER SERIAL LINES: SLIP
- RFC1144: Compressing TCP/IP headers for low-speed serial links
- RFC1717: The PPP Multilink Protocol (MP)
- RFC1332: The PPP Internet Protocol Control Protocol (IPCP)
- RFC1962: The PPP Compression Control Protocol (CCP)

## 13.5.3 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本                                                        |
|-------------------------------------------------------------------------|---------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2200/5100 V100R005C00、<br>USG2200/5100 BSR/HSR V100R005C00 |

## 13.5.4 原理描述

### 链路分片与交叉(LFI)的实现原理

链路分片与交叉可以将大型数据帧分割成小型数据帧后与其他小片的报文一起发送，从而减少在速度较慢的链路上的延迟和抖动。

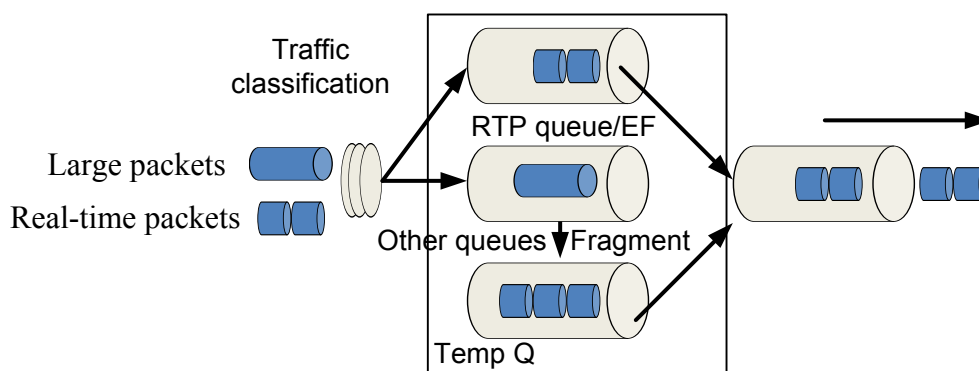
每一个分片重新构成一个报文，并加上序号，以保证设备在接收到这些分片后能够按顺序重新组装原报文。

分片长度不能配置的过小，否则每个报文分片后产生的分片报文数就会越多。每个分片报文都要加上 6 个字节长度的包头，这样就会降低报文收发的效率。另外，在组装报文

时，只要属于某个报文的一个分片报文没有收到，组装就不会成功，如果分片数过多，报文被丢弃的可能性就会增大。

图 13-11 描述了链路分片与交叉的处理过程。大报文和实时报文一起到达接口时，除了 RTP 实时队列和 EF 队列中的报文外，其他队列中的大型报文将被分成若干小包放入分片队列进行发送。如果 RTP 实时队列和 EF 中有缓存的报文，优先调度 RTP 实时队列和 EF 队列，否则调度分片队列，这样就避免了在低速链路上传送大包对实时报文造成的时延与抖动。

图 13-11 链路分片与交叉处理过程

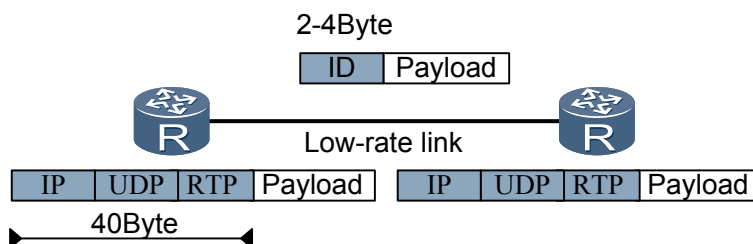


## IP 报文头压缩（IPHC）的实现原理

IPHC 是通过 IPCP 来协商压缩算法和压缩的参数的，要通过 PPP 链路实现 IP 数据包的压缩，必须在链路的两端设置压缩的参数。

IP 头压缩的处理过程如图 13-12 所示。

图 13-12 IP 头压缩处理过程



## 13.6 VPN QoS

VPN QoS 应用在 Tunnel 接口上，用于在报文未被 GRE 等隧道封装之前，对报文进行预分类，并对分类后的报文进行 QoS 处理，保证隧道中报文的服务质量。

### 13.6.1 介绍

#### 定义

VPN QoS 应用在 Tunnel 接口上，用于在报文未被 GRE 等隧道封装之前，对报文进行预分类，并对分类后的报文进行 QoS 处理，保证隧道中报文的服务质量。



## 目的

GRE 等 VPN 技术是通过在 Tunnel 接口上建立 VPN 隧道来实现的，为了保证隧道内流量的服务质量，需要在 Tunnel 接口上应用 VPN QoS。

由于经过 Tunnel 接口的报文是经过 GRE 等隧道封装的，传统 QoS 无法对隧道内的报文进行识别，从而不能对报文进行正确的流分类。而 VPN QoS 则可以在报文被隧道封装前，对报文进行分类，保证隧道内的报文也可以接受 QoS 提供的服务质量保证。

## 13.6.2 参考标准和协议

与 VPN QoS 特性相关的参考标准与协议如下：

- RFC1701: Generic Routing Encapsulation (GRE)
- RFC2784: Generic Routing Encapsulation (GRE)

## 13.6.3 可获得性

### License 支持

不需要 License 支持。

### 版本支持

| 产品                                                                      | 最低支持版本   |
|-------------------------------------------------------------------------|----------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | V300R001 |

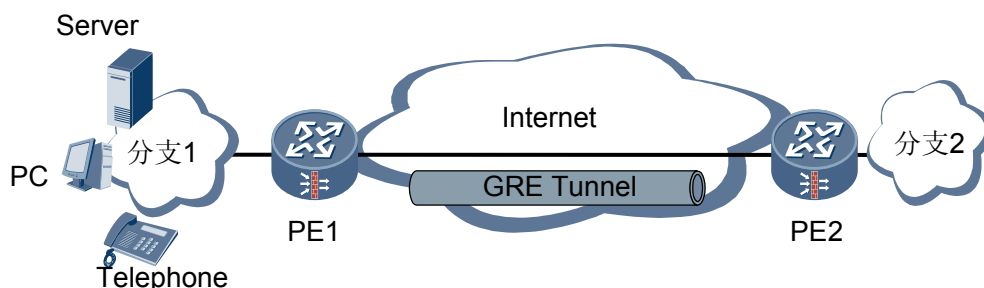
## 13.6.4 应用场景

VPN QoS 的典型应用场景有以下三种：

- 单通道多业务共享带宽

如图 13-13 所示，某公司的两个分支机构（分支 1 和分支 2）分别部署在两个不同的城市，他们之间通过 PE1 和 PE2 跨越 Internet 相连，PE1 和 PE2 之间使用 GRE 隧道实现互联。分支 1 中有视频、语音和数据等多种业务。由于 Tunnel 隧道有带宽限制，可能使报文在整个链路传输过程中出现拥塞，例如 PE1 或者 PE2 的物理接口带宽过小，或者中间网络出现拥塞。因此通过 VPN QoS 来实现对多业务的拥塞管理，保证两个部门之间的视频、语音和数据等多种业务的传输质量。

图 13-13 单通道多业务共享带宽

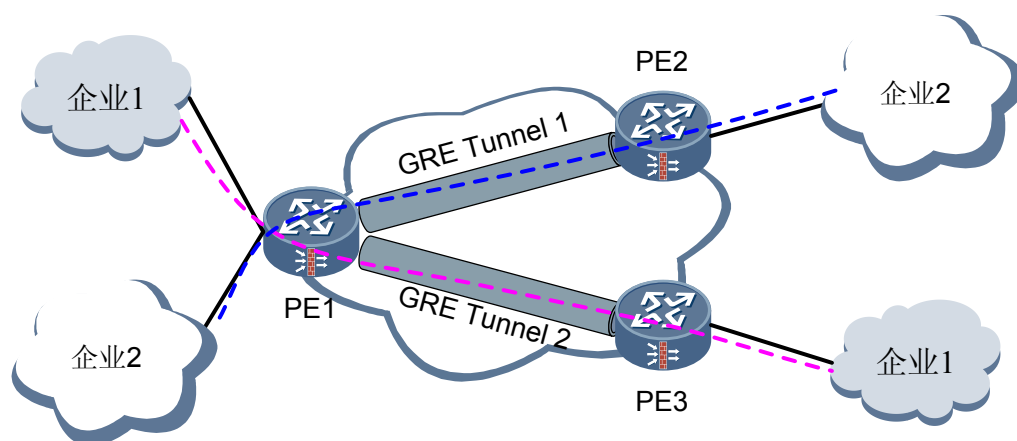


- 多个隧道共享带宽

如图 13-14 所示，PE1 分别与 PE2 和 PE3 之间通过 GRE 建立了两条隧道：GRE Tunnel1 和 GRE Tunnel2。Tunnel1 和 Tunnel2 共用 PE1 的出接口，企业 1 和企业 2 分别通过这两条隧道建立通信。

由于企业 1 和企业 2 共用 PE1 的出接口，如果其中一个企业的流量过大，将会挤压另一个企业的带宽。同时，企业内部的视频、语音和数据等多种业务也需要保证服务质量。因此，需要在 PE1 的出接口处应用 VPN QoS 来保证服务质量。

图 13-14 多个隧道共享带宽示意图

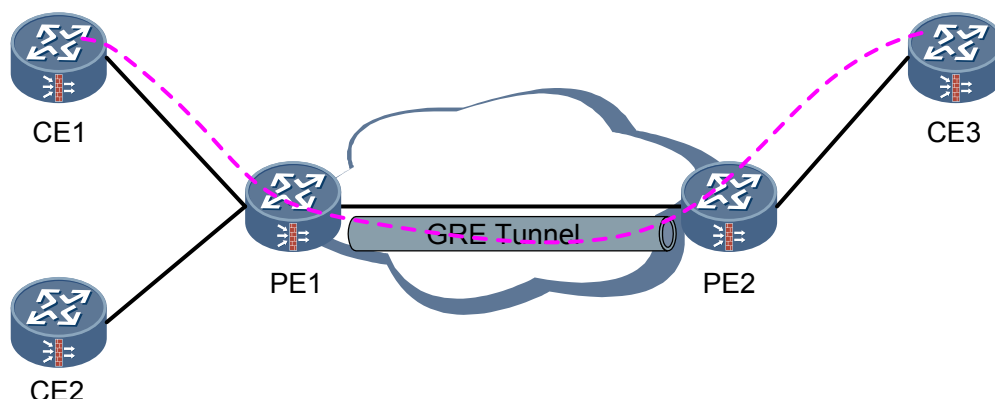


- 隧道流量与非隧道流量共享带宽

如图 13-15 所示，PE1 和 PE2 之间通过 GRE 建立了隧道。CE1 为企业总部，与分支机构 CE3 之间通过隧道通信。CE2 为普通上网数据流量，不走隧道，但与隧道共用 PE1 的出接口。

流量过大时，隧道流量和上网流量会相互抢占，导致隧道丢包严重或者上网时断时续。此时，需要对上网流量进行限流，保证隧道流量带宽。同时对隧道内部各业务进行拥塞管理保证关键业务的服务质量。

图 13-15 隧道流量与非隧道流量共享带宽示意图

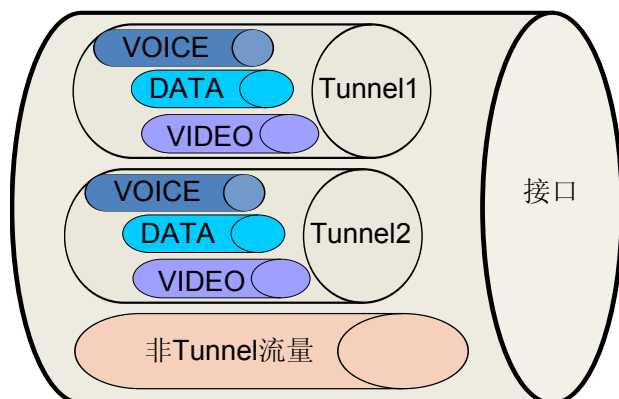


## 13.6.5 原理描述

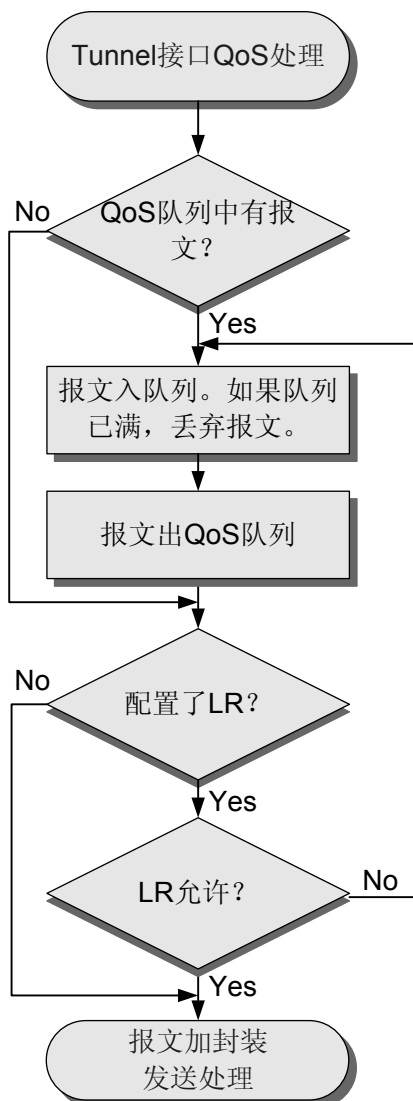
### VPN QoS 的需求模型

VPN QoS 的需求模型如图 13-16 所示，在同一个物理口上配置多个 Tunnel 接口建立隧道，或者 Tunnel 流量与非隧道流量共享带宽时，通过启用 Tunnel 口上的 QoS，有限保证隧道流量，并对隧道中不同业务报文进行拥塞管理。

图 13-16 VPN QoS 的需求模型示意图



## VPN QoS 拥塞管理处理流程



## 13.7 HQoS

HQoS 是一种既能控制用户的流量，又能同时根据用户的业务进行调度的 QoS 技术，多级的 QoS 调度提供了精细化的服务质量保证。

### 13.7.1 介绍

#### 定义

HQoS 是一种既能控制用户的流量，又能同时根据用户业务优先级进行调度的 QoS 技术。多级的 QoS 调度提供了精细化的服务质量保证。

## 目的

HQoS 弥补了传统 QoS 无法对接口上多个用户的多种业务流量进行多级区分服务的不足，在区分了用户的同时，还对用户的流量进行了业务区分。既提供了精细化的服务质量保证，又可以从整体上节约网络运维成本。

## 13.7.2 可获得性

### License 支持

不需要 License 支持。

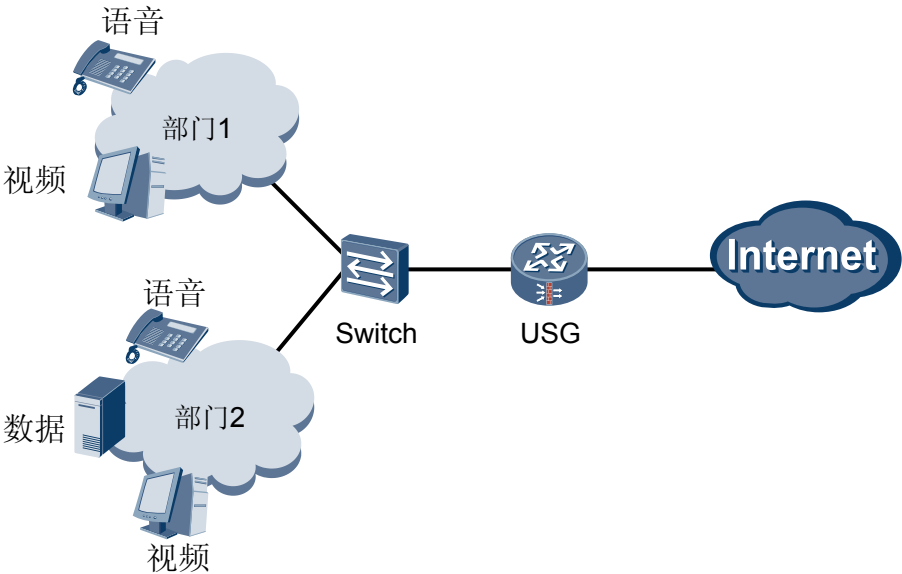
### 版本支持

| 产品                                                                      | 最低支持版本   |
|-------------------------------------------------------------------------|----------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | V300R001 |

## 13.7.3 应用场景

如图 13-17 所示，HQoS 主要应用在网络中存在多个独立的部门，而各个部门中的不同业务也有关键业务和非关键业务之分的情况。为了保证重要部门得到更多的带宽，而重要部门中的关键业务可以得到优先的服务，就需要使用 HQoS 技术进行层次化的流量调度。

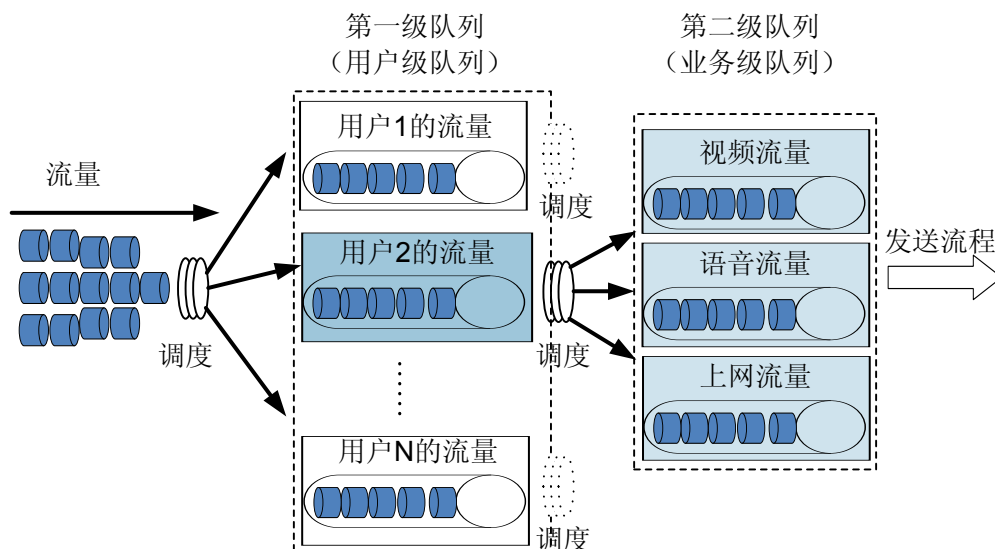
图 13-17 HQoS 常见应用场景图



## 13.7.4 原理描述

如图 13-18 所示,HQoS 可以进行两级的队列调度。第一级用来区分用户, 第二级用来区分业务, 每一级都对应某一特定的队列调度方式。

图 13-18 HQoS 调度方式示意图



HQoS 目前支持的调度队列包括 CQ 和 PQ 两种。其中, 用户级队列只支持使用 CQ 队列, 业务级队列既可以单独使用 CQ 队列或 PQ 队列, 也可以同时使用两种队列。

使用 CQ 队列可以在保证关键业务获得较多带宽的同时, 使得非关键业务也可以获得一定的带宽, 避免由于关键业务长期占用带宽而造成的非关键业务饿死的现象。

使用 PQ 队列可以在网络发生拥塞时, 保证关键业务的报文得到优先处理, 非关键业务的报文只有在处理完关键业务的报文后才可以进行处理。

## HQoS 的具体实现方式

HQoS 使用两级队列绑定的方式实现对流量的层次化调度。其中, 第一级队列指用户级队列, 目前支持的队列为 CQ 队列; 第二级队列指业务级队列, 目前支持的队列包括 CQ 队列和 PQ 队列。具体实现流程如下所示。

1. 配置第一级队列 (用户级队列): 设置按照不同用户对流量进行区分的匹配规则和匹配后的入队列动作。
2. 配置第二级队列 (业务级队列): 设置将用户流量按照业务进行区分的匹配规则和匹配后的入队列动作。
3. 配置两级队列的绑定策略: 设置用户级队列和业务级队列的绑定关系。
4. 在接口上应用这些策略: 在接口上通过应用用户级队列来实现 HQoS 策略。

实际上, 以上四步中只有步骤 3 是 HQoS 特有的配置步骤, 其他步骤的配置方法和传统 QoS 中的配置方法完全相同。

## 13.8 二层 QoS

介绍 QoS 应用于二层接口卡上的接口限速、优先级映射和拥塞管理技术。

## 13.8.1 简介

介绍二层接口上的 LR、优先级映射和拥塞管理的基本概念。

支持二层 QoS 的接口卡有：18FE+2SFP，16GE+4SFP，5FSW，8FE+2GE

### 二层接口限速

**二层 LR：**在链路层的限速，对二层接口入方向和出方向流量进行限速。

### 优先级映射

在网络的入口需要为网络的流量打上一定的区分标记，这种标记用以标识流量的调度权重或者转发处理优先级别的高低。在网络中间节点处理报文时，就可以根据报文的优先级来进行相应的调度。报文在进入设备以后，设备会根据自身支持的情况和相应的规则给报文分配包括 802.1p 优先级、本地优先级、端口优先级等在内的一系列参数。

设备提供 dot1p-lp 缺省映射关系表，802.1p 优先级到本地优先级的缺省映射关系。

**802.1p 优先级：**位于二层报文头部，适用于不需要分析三层报头，而需要在二层环境下保证 QoS 的情况。

**本地优先级：**USG 为报文分配的具有本地意义的优先级，对应于出接口队列序号。

**端口优先级：**预先在端口上配置的 QoS 优先级。

### 拥塞管理

与三层接口的拥塞管理实现不同的是，二层接口的拥塞管理运用的是 SP 队列、WRR 队列和 SP+WRR 队列。

**SP 队列：**SP（Strict-Priority）严格优先级队列，在队列调度时，严格按照优先级从高到低的顺序发送分组。

**WRR 队列：**WRR（Weighted Round Robin）加权轮询队列，使用 WRR 来对各个队列进行轮询调度。这里的轮询调度并不是为每个队列分配固定的服务时间，而是为每个队列配置了一个权重值，并用这个权重值来表示该队列可以获取到的资源的比重。

**SP+WRR 队列：**二层接口上的队列部分使用 SP 队列，部分使用 WRR 队列，实现 SP+WRR 的调度功能。SP 调度组内的各个队列执行严格调度方式，WRR 调度组中的队列执行加权轮循调度方式。

## 13.8.2 可获得性

### License 支持

不需要 License 支持。

## 版本支持

| 产品                                                                      | 最低支持版本                                                                                                  |
|-------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|
| Secoway USG2110-X/<br>2100/2200/5100/5500、<br>USG2100/2200/5100 BSR/HSR | USG2110-X/2100/2200/5100<br>V100R005C00、USG2100/2200/5100<br>BSR/HSR V100R005C00<br>USG5500 V200R001C00 |

## 13.8.3 原理描述

### 13.8.3.1 优先级映射

#### USG 提供的优先级信任模式

USG 对报文的优先级信任模式：默认为信任端口优先级模式，可配置为信任报文的 802.1p 优先级。

当信任模式为信任报文的 802.1p 优先级时，USG 根据报文中是否携带 802.1Q 标签头，采取两种方式处理优先级映射关系：

- 报文不带 802.1Q 标签头：  
USG 使用端口优先级作为该端口接收到的报文的 802.1p 优先级。根据此优先级查找 802.1p 优先级到本地优先级映射表，为报文标记本地优先级。
- 报文带有 802.1Q 标签头：  
根据报文的 802.1p 优先级，查找 802.1p 优先级（dot1p）到本地优先级（lp）映射表，为报文标记本地优先级。

#### 优先级映射表

设备中存在 802.1p 优先级到本地优先级的缺省映射关系。

802.1p 优先级到本地优先级缺省映射关系如表 13-1 所示。

表 13-1 dot1p-lp 缺省映射关系

| 802.1p 优先级（dot1p） | 本地优先级（lp） |
|-------------------|-----------|
| 0                 | 2         |
| 1                 | 0         |
| 2                 | 1         |
| 3                 | 3         |
| 4                 | 4         |
| 5                 | 5         |



| 802.1p 优先级 (dot1p) | 本地优先级 (lp) |
|--------------------|------------|
| 6                  | 6          |
| 7                  | 7          |

用户可以根据需要，修改设备中的优先级映射表。

### 13.8.3.2 SP 队列

为保证关键业务在拥塞发生时可以先获得服务以减小相应的延迟，可以使用严格优先级队列 SP（Strict-Priority）作为队列调度机制。

在队列调度时，SP 严格按照优先级从高到低的顺序发送分组。

USG 的 18FE+2SFP 和 16GE+4SFP 接口卡的接口支持 8 个队列，5FSW 和 8FE+2GE 接口卡的接口支持 4 个队列。以 5FSW 接口卡为例，配置 SP 队列，将端口的输出队列分为高优先级队列、中优先级队列、正常优先级队列和低优先级队列，这些队列的优先级依次降低。发送分组时，只有当高优先级队列为空时，才会发送较低优先级队列中的分组。这样，通过将关键业务的分组放入较高优先级的队列，将非关键性业务的分组放入较低优先级的队列，就可以保证关键业务的分组被优先传送。

SP 队列依据进入设备的报文中的优先级对数据进行区分，然后根据优先级到本地优先级（LP）的映射关系使数据进入相应队列，最后依据 SP 队列的规则对数据进行发送。

### 13.8.3.3 WRR 队列

USG 的 18FE+2SFP 和 16GE+4SFP 接口卡的接口支持 8 个队列，5FSW 和 8FE+2GE 接口卡的接口支持 4 个队列。以 18FE+2SFP 接口卡为例，队列 ID 分别为 0～7。为了保证每个队列都可以得到一定程度的服务，可以使用 WRR 来对各个队列进行轮询调度。这里的轮询调度并不是为每个队列分配固定的服务时间，而是为每个队列配置了一个权重值，并用这个权重值来表示该队列可以获取到的资源的比重。

例如，在某个带宽为 100M 的接口上应用 WRR 队列，并分别为 0 号到 7 号队列配置它们的权重值为 10、10、10、10、10、20、10、20。通过计算可以发现，最终即使是最低优先级的队列也可以获得 10Mbit/s 的带宽（计算公式： $100 \times 10 / (10 + 10 + 10 + 10 + 10 + 20 + 10 + 20) = 10$ ）。通过 WRR 调度机制，保证了带宽资源可以得到充分的利用。

### 13.8.3.4 SP+WRR 队列

根据用户的需要，可以配置端口上的队列部分使用 SP 队列，部分使用 WRR 队列，实现 SP+WRR 的调度功能。SP 调度组内的各个队列执行严格调度方式，WRR 调度组中的队列执行加权轮询调度方式。

在实际队列调度时，优先保证 SP 调度组内的队列调度，只有当 SP 调度组内的列表中没有报文需要发送时，才会执行 WRR 调度组中的队列调度。