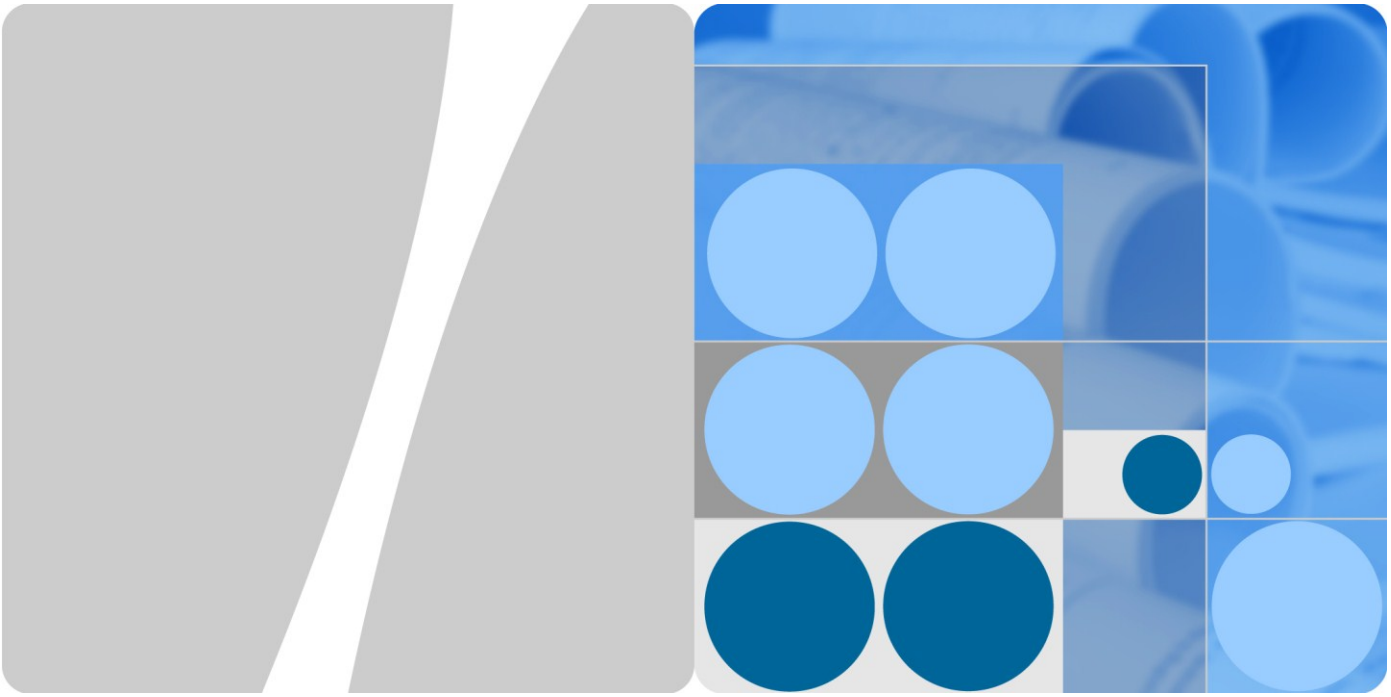


华为 Secoway USG9500 统一安全网关
特性描述

文档版本 01

发布日期 2012-10-22

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。

本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

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

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

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 4008302118

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
USG9500	V200R001C01

读者对象

本文档介绍了 USG9500 的功能特性。

本文档提供了 USG9500 的特性介绍、原理描述、应用场景和相关参考文档。

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。
 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。

符号	说明
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。

修订记录

文档版本 01 (2012-10-22)

第一次正式发布。

目 录

1 概述

关于本章

1.1 网络安全概述

1.1.1 威胁

随着互联网的迅速发展，越来越多的企业借助网络服务来加速自身的发展。如何在一个开放的网络环境中守卫自身的机密数据和资源已越来越为人们所关注。

目前，常见的网络安全威胁主要分为以下几类：

- 非法使用

  资源被未授权的用户（非法用户）或合法用户以未授权方式（非法权限）使用。例如，攻击者通过猜测帐号和密码，进入计算机系统，非法使用资源。
- 拒绝服务

  服务器拒绝合法用户正常访问信息或资源的请求。例如，攻击者短时间内使用大量数据包不断向服务器发起连接，致使服务器负荷过重而不能处理正常访问。
- 信息盗窃

  攻击者并不直接入侵目标系统，而是通过窃听网络来获取重要数据或信息。
- 数据篡改

  攻击者对系统数据或消息流进行有选择的修改、删除、延误、重排序及插入虚假消息等操作，破坏数据的一致性。

1.1.2 服务种类

针对各种安全威胁而采取的安全防护措施称为安全服务，它主要分为以下几类：

- 可用性服务

保证信息或数据在需要时能够被合法用户正常访问。

- 机密性服务
  保证敏感信息或数据不被泄漏给未授权的用户。
- 完整性服务
  保证信息或数据不被未经授权的用户改动或破坏。
- 鉴别服务
  保证某个通信实体身份的合法性。
- 授权服务
  对资源的使用实施控制，规定访问者的权限。

1.1.3 服务实现

加密

加密是将可读的明文文本转化为不可读的加密文本的过程。加密不仅为用户提供通信方面的安全保证，同时也是其他许多安全机制的基础。

加密的方式主要分为三种：

- 对称密码体制
  其特征是用于加密和解密的密钥是同一个，通信双方通过共享同一密钥来交换消息。密钥必须秘密保存。典型代表包括：DES（Data Encryption Standard）、3DES（Triple DES）等。
- 公钥密码体制
  不同于对称密码体制，公钥密码体制有两个不同密钥，可将加密功能和解密功能分开。一个密钥称为私钥，必须秘密保存；另一个称为公钥，可被公开分发。典型代表包括：DH（Diffie-Hellman）、RSA（Rivest，Shamir，Adleman）。
- 散列函数机制
  其特征是将一个变长的消息压缩到一个定长的编码字中，成为一个散列或消息摘要。典型代表包括：MD5（Message-Digest Algorithm 5）、SHA（Secure Hash Algorithm）。

加密技术能够应用在以下安全机制中：

- 认证口令设计
- 安全通信协议设计
- 数字签名设计

认证

认证通常在访问网络前或网络提供服务前进行，用于鉴别用户身份的合法性。

认证服务可以由网络上的设备在本地提供，也可以通过专用的认证服务器提供。相比较而言，后者具有更好的灵活性、可控性和可扩展性。目前，在异构网络环境中，认证服务主要使用 RADIUS（Remote Authentication Dial in User Service）这一开放的标准。

访问控制

访问控制是一种加强授权的方法，一般分为两种：

- 基于操作系统的访问控制

  对用户访问某计算机系统资源时的特定访问行为进行授权。可以基于身份、组、规则等属性配置访问控制策略。

- 基于网络的访问控制

  限制接入网络的权限。由于网络的复杂性，其机制远比基于操作系统的访问控制更为复杂。一般在访问请求者和访问目标之间的一些中间设备（例如 USG9500）上配置访问控制策略，从而实现基于网络的接入控制。

安全协议

安全协议是网络安全的重要内容。下面将从 TCP/IP（Transmission Control Protocol/Internet Protocol）的分层模型角度来介绍目前广泛使用的安全协议：

- 应用层安全

  提供从一台主机上的应用程序到另一台主机上的应用程序的端到端的安全保障。应用层安全机制必须根据具体的应用而定，因此不存在通用的应用层安全协议。

  例如，SSH（Secure Shell）协议可以建立安全的远程登录会话，为 Telnet、FTP 等提供安全的连接通道。

- 传输层安全

  提供同一台主机的进程之间，或不同主机的进程之间的安全保障。在传输层中提供安全服务的方法是强化通信实体双方的交互过程，具体包括通信实体的认证、数据加密、密钥的交换等。

  例如，SSL（Secure Socket Layer）可以在 TCP 的基础上提供安全保障。

- 网络层安全

  网络层安全是整个 TCP/IP 安全的基础，也是 Internet 安全的核心。即使上层协议没有实现安全性保障，通过对网络层报文进行保护，用户信息也能够从网络层得到安全保障。

  目前，网络层最重要的安全协议是 IPSec（IP Security Protocol）。IPSec 是一系列网络安全协议的总称，包括安全协议、加密协议等，能够为通讯双方提供访问控制、无连接的完整性、数据源认证、防重放、加密以及对数据流分类加密等服务。

- 数据链路层安全

  提供点到点的安全性，如在一个点到点链路或帧中继的永久虚链路上提供安全性。链路层安全的主要实现方法是在链路的每一端使用专用设备完成加密和解密。

1.2 防火墙概述

1.2.1 简介

在实际的网络环境中，单一的安全防护技术不足以确保网络的安全，多种安全防护技术的综合应用才能够将安全风险控制在尽量小的范围内。

一般而言，构建一个安全防范体系具体实施的第一项内容就是在内部网络和外部网络之间构筑一道防线，以抵御来自外部的绝大多数攻击。完成这项任务的网络产品的作用类似于建筑行业中用于防止火灾蔓延的隔断墙，因此我们称这种网络产品为防火墙。

防火墙是监控可信任网络（内部网络）和不可信任网络（外部网络）之间的访问通道。它一方面阻止来自外部网络的用户对内部网络的未授权访问，另一方面允许内部网络的用户对外部网络进行访问。防火墙也可以作为一个访问因特网的权限控制关口，如允许组织内的特定的主机可以访问因特网。现在的许多防火墙同时还具有一些其他特点，如进行身份鉴别，对信息进行安全处理（如加密）等等。

防火墙不单用于对因特网的连接，也可以用来保护内部网络的大型机和重要的资源（如数据）。对受保护数据的访问都必须经过防火墙的过滤，即使网络内部用户要访问受保护的数据，也要经过防火墙。

防火墙主要用于以下目的：

- 限制用户或信息由一个特定的被严格控制的站点进入。
- 阻止攻击者接近其他安全防御设施。
- 限制用户或信息由一个特定的被严格控制的站点离开。

1.2.2 发展历史

第一代防火墙－包过滤防火墙

包过滤机制是指设备在网络层对每一个数据包进行检查，根据配置的安全策略转发或丢弃数据包。

包过滤防火墙的基本原理是通过配置 ACL（Access Control List），根据源/目的 IP 地址、源/目的端口号、协议类型和报文传递的方向等信息制定规则，对匹配规则的报文采取相应（允许或拒绝）的操作。

包过滤防火墙设计简单，易于实现，而且价格便宜。但其缺点也不容忽视，主要表现在：

- 随着 ACL 数量的增加，防火墙过滤性能急剧下降。
- 静态配置的 ACL 灵活性差，难以适应动态的安全要求。
- 包过滤机制不检查会话状态也不分析数据内容，安全性低。

  例如，攻击者可以使用假冒地址进行欺骗，通过把自己主机 IP 地址设成一个合法主机 IP 地址，就能轻易地通过报文过滤器，达到攻击目的。

第二代防火墙－代理防火墙

代理服务作用于网络的应用层，其工作原理是接管内部网络和外部网络用户之间直接进行的业务。代理服务检查来自内部网络客户端的请求，认证通过后，将代表客户端与真正的外部网络服务器建立连接，转发客户端的请求，并将服务器返回的响应回送给客户端。

代理防火墙能够完全控制会话过程和信息交换，具有较高的安全性。但其缺点也同样突出，主要表现在：

- 代理服务由软件实现，限制了处理速度，容易遭受拒绝服务攻击。
- 需要针对每一种协议开发应用层代理，开发代价大并且升级困难。

第三代防火墙－状态检测防火墙

状态检测技术是包过滤技术的扩展（非正式的也可称为“动态包过滤”）。状态检测防火墙的基本原理如下：

- 通过各种状态表来追踪激活的 TCP（Transmission Control Protocol）会话和 UDP（User Datagram Protocol）伪会话，由 ACL 规则来决定哪些会话允许建立，只有与被允许建立的会话相关联的数据包才被转发。

说明

UDP 伪会话是指防火墙在处理基于 UDP 的数据包时为 UDP 建立虚拟连接，以便对 UDP 连接过程进行状态监控的会话过程。

- 状态检测防火墙在网络层截获数据包，然后从数据包中提取出安全策略所需要的状态信息，并保存到动态状态表中。通过分析这些状态表和与该数据包有关的后续连接请求来做出恰当决定。

状态检测防火墙具有以下优点：

- 速度快

  状态检测防火墙对数据包进行 ACL 检查的同时，可以将数据包连接状态记录下来，后续数据包则无需再进行 ACL 检查，只需根据状态表进行连接记录检查即可。检查通过后，该连接状态记录将被刷新，从而避免重复检查具有相同连接状态的数据包。连接状态表里的记录可以随意排列，防火墙可采用诸如二叉树或哈希（hash）等算法进行快速搜索，提高了处理效率。

- 安全性较高

  连接状态表是动态管理的，老化时间到期后，防火墙会删除连接状态表项，保障了内部网络的实时安全。同时，防火墙采用连接状态实时监控技术，通过在状态表中识别诸如应答响应等连接状态因素，增强了系统的安全性。

2 入门

关于本章

2.1 工作模式

2.1.1 分类

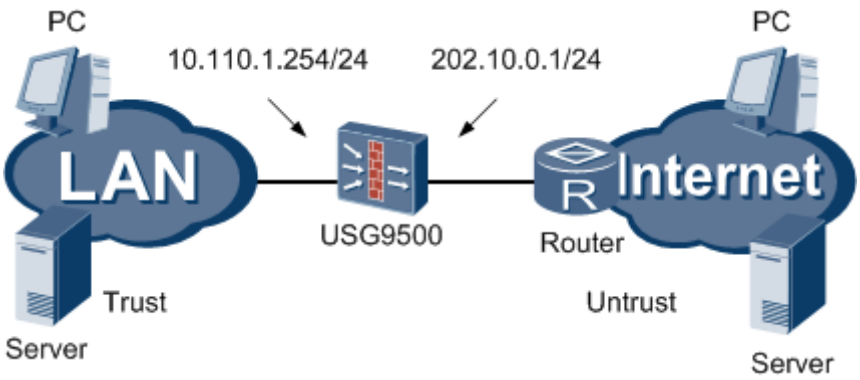
目前，USG9500 能够工作在三种模式下：路由模式、透明模式、混合模式。

- 路由模式

  路由模式下 USG9500 以第三层对外连接，所有接口都需要配置 IP 地址。此时需要将 USG9500 与内部网络、外部网络相连的接口分别配置成不同网段的 IP 地址，并重新规划原有的网络拓扑。此时的 USG9500 相当于一台路由器。

  如图 2-1 所示，USG9500 的 Trust 区域接口与公司内部网络相连，Untrust 区域接口与外部网络相连。值得注意的是，Trust 区域接口和 Untrust 区域接口分别处于两个不同的子网中。

图2-1 路由模式组网图



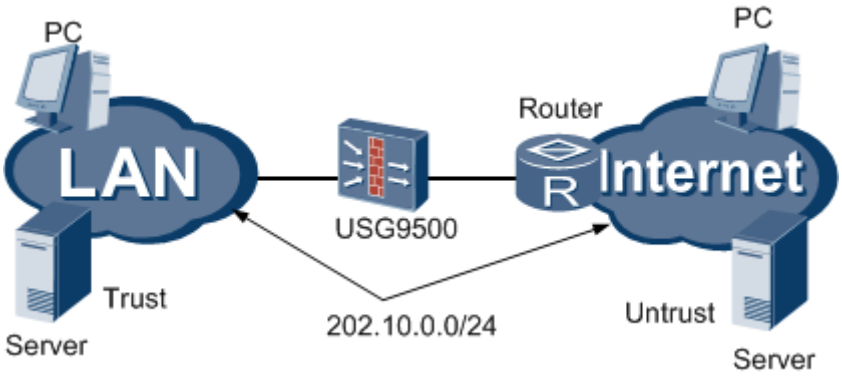
采用路由模式的优点是可以完成 ACL 包过滤的功能。缺点是需要对网络拓扑进行修改。例如，内部网络用户需要更改网关，路由器需要更改路由配置等。

- 透明模式

  透明模式下 USG9500 通过第二层与外界连接，所有接口都不能配置 IP 地址。此时 USG9500 对于子网用户和路由器来说是完全透明的，用户完全感觉不到 USG9500 的存在。

  如图 2-2 所示，USG9500 的 Trust 区域接口与公司内部网络相连，Untrust 区域接口与外部网络相连。需要注意的是内部网络和外部网络必须处于同一个子网。

图2-2 透明模式组网图



透明模式可以避免改变拓扑结构造成的麻烦。采用透明模式时，只需在网络中像放置网桥（bridge）一样插入 USG9500 即可，无需修改任何已有的配置。IP 报文同样会经过相关的过滤检查，内部网络用户依旧受到 USG9500 的保护。

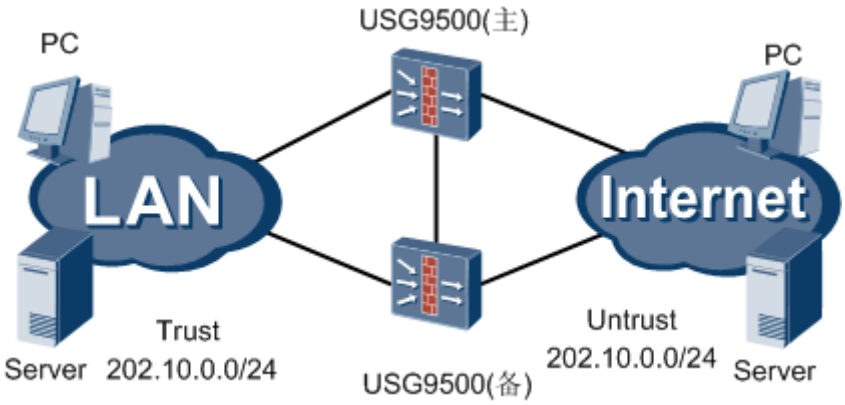
- 混合模式

  混合模式既存在工作在路由模式的接口（接口具有 IP 地址），又存在工作在透明模式的接口（接口无 IP 地址）。

  混合模式主要用于透明模式作双机热备份的情况，此时启动 VRRP（Virtual Router Redundancy Protocol）功能的接口需要配置 IP 地址，其他接口不配置 IP 地址。

  USG9500 混合模式的典型组网方式如图 2-3 所示。

图2-3 混合模式组网图



主/备 USG9500 的 Trust 区域接口与公司内部网络相连；Untrust 区域接口与外部网络相连；主/备 USG9500 互相连接，并运行 VRRP 进行备份。需要注意的是内部网络和外部网络必须处于同一个子网。

2.1.2 工作原理

三种工作模式的工作原理如下：

- 路由模式

  USG9500 工作在路由模式下时，所有接口都需要配置 IP 地址。不同的安全区域相关的接口连接的外部用户属于不同的子网。

  当报文在接口间进行转发时，根据报文的 IP 地址来查找路由表。此时 USG9500 表现为一个路由器。但是，USG9500 与路由器不同，USG9500 转发的 IP 报文还需要送到上层进行相关过滤等处理，通过检查会话表或 ACL 规则以确定是否允许该报文通过。除此之外，USG9500 还需要完成其他防攻击检查。

- 透明模式

  USG9500 工作在透明模式（也可以称为桥模式）下时，所有接口都不能配置 IP 地址。透明模式下所有相关接口连接的外部用户同属一个子网。

  当 USG9500 转发报文时，需要根据报文的 MAC（Media Access Control）地址寻找出接口。此时 USG9500 表现为一个透明网桥。但是，USG9500 与网桥不同，USG9500 转发的 IP 报文还需要送到上层进行相关过滤等处理，通过检查会话表或 ACL 规则以确定是否允许该报文通过。此外，USG9500 还需要完成其他防攻击检查。

  工作在透明模式下的 USG9500 在数据链路层连接局域网（LAN），网络终端用户无需为连接网络而对设备进行特别配置，就像 LAN Switch 进行网络连接。

- 混合模式

  USG9500 工作在混合模式下时，部分接口配置 IP 地址，部分接口不能配置 IP 地址。配置 IP 地址的接口，接口上启动 VRRP 功能，用于双机热备份；而未配置 IP 地址的接口，相关接口连接的外部用户同属一个子网。

  当报文在透明模式下的接口间进行转发时，转发过程与透明模式的工作过程完全相同。当 USG9500 进行双机热备份时，转发过程类似路由模式的工作过程。

2.2 安全区域

2.2.1 概述

区域（zone）是 USG9500 产品所引入的一个安全概念，是 USG9500 产品区别于路由器的主要特征。

对于路由器，各个接口所连接的网络在安全上可以视为是平等的，没有明显的内外之分，所以即使进行一定程度的安全检查，也是在接口上完成的。这样，一个数据流单方向通过路由器时有可能需要进行两次安全规则的检查：入接口的安全检查和出接口的安全检查，以便使其符合每个接口上独立的安全定义。

而这种思路对于 USG9500 不适合，因为 USG9500 放置于内部网络和外部网络之间，用于保护内部网络不受外部网络上恶意用户的侵害，有着明确的内外之分。当一个数据流通过 USG9500 的时候，根据其发起方向的不同，所引起的操作是截然不同的。由于这种安全级别上的差别，采用在接口上检查安全策略的方式已经不适用。因此，USG9500 提出了安全区域的概念。

一个安全区域是一个或多个接口的组合，具有一个安全级别。

安全区域有如下特点：

- 安全级别通过 1 ~ 100 的数字表示，数字越大表示安全级别越高。
- 不存在两个具有相同安全级别的安全区域。

2.2.2 划分

USG9500 缺省保留四个安全区域，划分如下：

- 非受信区域 Untrust
  低安全级别的安全区域，安全级别为 5。
- 非军事化区域 DMZ（Demilitarized Zone）
  中等安全级别的安全区域，安全级别为 50。
- 受信区域 Trust
  较高安全级别的安全区域，安全级别为 85。
- 本地区域 Local
  最高安全级别的安全区域，安全级别为 100。

这四个区域无需创建，也不能删除，同时其安全级别也不能重新设置。

根据实际组网需要，用户可以自行创建安全区域并定义其安全级别。

说明

- DMZ 这一术语起源于军方，指的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。
- USG9500 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区域。该区域可以放置需要对外提供网络服务的设备，如 WWW Server、FTP Server 等。上述服务器如果放置于外部网络，则 USG9500 无法保障它们的安全；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。DMZ 区域很好地解决了服务器的放置问题。

2.2.3 接口、网络与安全区域的关系

注意

- 系统不允许两个安全区域具有相同的安全级别。
- 系统不允许同一接口属于两个不同的安全区域。

接口、网络与安全区域的关系如下：

- 接口与安全区域的关系

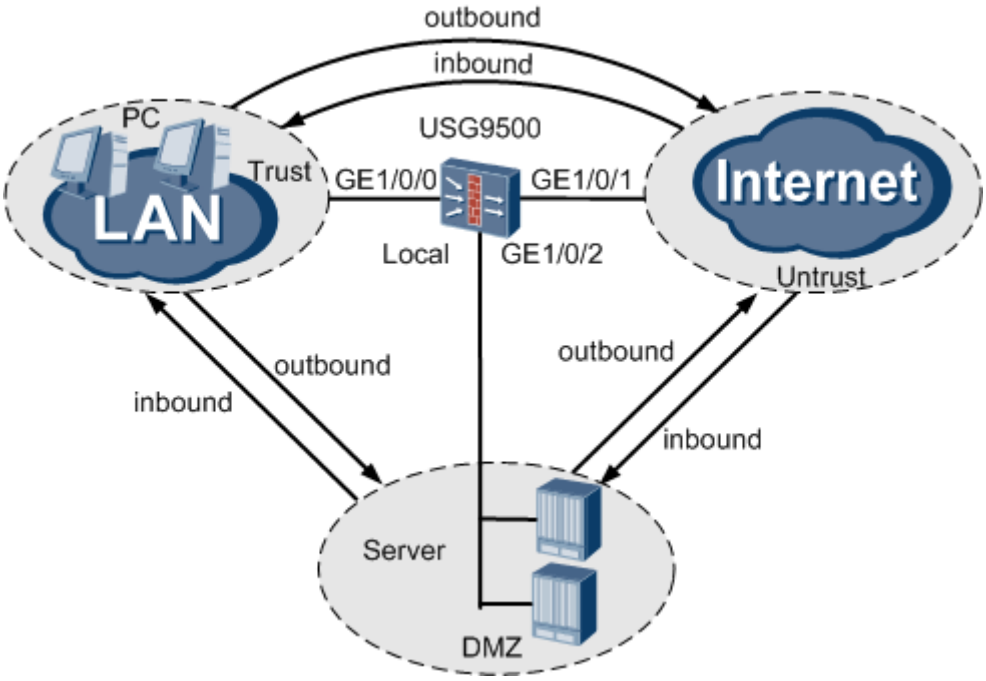
  一个安全区域可以包括一个或多个接口，具有一个安全级别。

  除 Local 区域外，使用其他安全区域前，都需要将安全区域分别与 USG9500 的特定接口关联，即将接口加入安全区域。
- 网络与安全区域的关系

  安全区域与各网络的关联遵循如下原则：
  - 需要保护的网络应安排在安全级别较高的区域，如 Trust 区域。
  - 外部网络应安排在安全级别较低的区域，如 Untrust 区域。
  - 对外提供有条件服务的网络应安排在中等安全级别的区域，如 DMZ 区域。
  - Local 区域不包含任何接口，USG9500 设备本身即可认为是 Local 区域。
- 接口、网络与安全区域之间的关系

  接口、网络与安全区域之间的关系如图 2-4 所示。

图2-4 接口、网络和安全区域的关系示意图



2.2.4 数据流方向

两个安全区域之间（简称安全域间）的数据流分两个方向：

- 入方向（inbound）

  数据由低安全级别的安全区域向高安全级别的安全区域传输的方向。
- 出方向（outbound）

  数据由高安全级别的安全区域向低安全级别的安全区域传输的方向。

不同安全级别的安全区域间的数据流动都将激发 USG9500 进行安全策略的检查。可以事先为同一安全域间的不同方向设置不同的安全策略，当有数据流在此安全域间的两个不同方向上流动时，将触发不同的安全策略检查。

在 USG9500 上，判断数据传输是出方向还是入方向，总是相对高安全级别的安全区域一侧而言。数据流的具体出、入方向如下：

- 从 Local 安全区域到 Trust 安全区域的数据流动方向为出方向，反之为入方向。
- 从 Local 安全区域到 DMZ 安全区域的数据流动方向为出方向，反之为入方向。
- 从 Local 安全区域到 Untrust 安全区域的数据流动方向为出方向，反之为入方向。
- 从 Trust 安全区域到 DMZ 安全区域的数据流动方向为出方向，反之为入方向。
- 从 Trust 安全区域到 Untrust 安全区域的数据流动方向为出方向，反之为入方向。
- 从 DMZ 安全区域到 Untrust 安全区域的数据流动方向为出方向，反之为入方向。

说明

- 在 USG9500 上，当报文从高安全级别的安全区域向低安全级别的安全区域发起连接时，如果允许高安全级别的安全区域用户自由访问外部网络，可以配置域间缺省过滤规则为允许报文通过。
- 路由器上数据流动方向的判定是以接口为主：由接口发送的数据方向称为出方向；由接口接收的数据方向称为入方向。这也是路由器有别于 USG9500 的重要特征。

3 IP 路由

关于本章

3.1 IP 路由概述

3.1.1 路由表和 FIB 表

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

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

路由表

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

- 协议路由表

  协议路由表中存放着该协议发现的路由信息。

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

- 本地核心路由表

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

说明

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

路由表中的内容

在 USG9500 中，通过执行命令 **display ip routing-table** 可以查看到路由器的路由表简表，如下：

```
<USG9500> display ip routing-table
Route Flags: R - relay, D - download to fib
------------------------------------------------------------------------------
Routing Tables: Public
        Destinations : 8        Routes : 8

Destination/Mask  Proto  Pre  Cost      Flags NextHop         Interface

        0.0.0.0/0  Static 60   0           D   1.1.4.2         Pos1/0/0
        1.1.1.0/24  Direct 0    0           D   1.1.1.1         GigabitEthernet2/0/0
        1.1.1.1/32  Direct 0    0           D   127.0.0.1       InLoopBack0
        1.1.4.0/30  OSPF   10   0           D   1.1.4.1         Pos1/0/0
        1.1.4.1/32  Direct 0    0           D   127.0.0.1       InLoopBack0
        1.1.4.2/32  OSPF   10   0           D   1.1.4.2         Pos1/0/0
      127.0.0.0/8   Direct 0    0           D   127.0.0.1       InLoopBack0
    127.0.0.1/32    Direct 0    0           D   127.0.0.1       InLoopBack0
```

路由表中包含了下列关键项：

- Destination：目的地址。用来标识 IP 包的目的地址或目的网络。
- Mask：网络掩码。与目的地址一起来标识目的主机或路由器所在的网段的地址。
  - 将目的地址和网络掩码“逻辑与”后可得到目的主机或路由器所在网段的地址。例如：目的地址为 1.1.1.1，掩码为 255.255.255.0 的主机或路由器所在网段的地址为 1.1.1.0。
  - 掩码由若干个连续“1”构成，既可以用点分十进制表示，也可以用掩码中连续“1”的个数来表示。例如掩码 255.255.255.0 长度为 24，即可以表示为 24。
- Proto：用来学习路由的协议。
- Pre：本条路由加入 IP 路由表的优先级。针对同一目的地，可能存在不同下一跳、出接口等的若干条路由，这些不同的路由可能是由不同的路由协议发现的，也可

以是手工配置的静态路由。优先级高（数值小）者将成为当前的最优路由。各协议路由优先级请参见表 3-1。

- Cost：路由开销。当到达同一目的地的多条路由具有相同的优先级时，路由开销最小的将成为当前的最优路由。

说明

Preference 用于不同路由协议间路由优先级的比较，Cost 用于同一种路由协议内部不同路由优先级的比较。

- NextHop：下一跳 IP 地址。说明 IP 包所经由的下一个路由器。
- Interface：输出接口。说明 IP 包将从该路由器哪个接口转发。

根据路由的目的地不同，可以划分为：

- 网段路由：目的地为网段
- 主机路由：目的地为主机

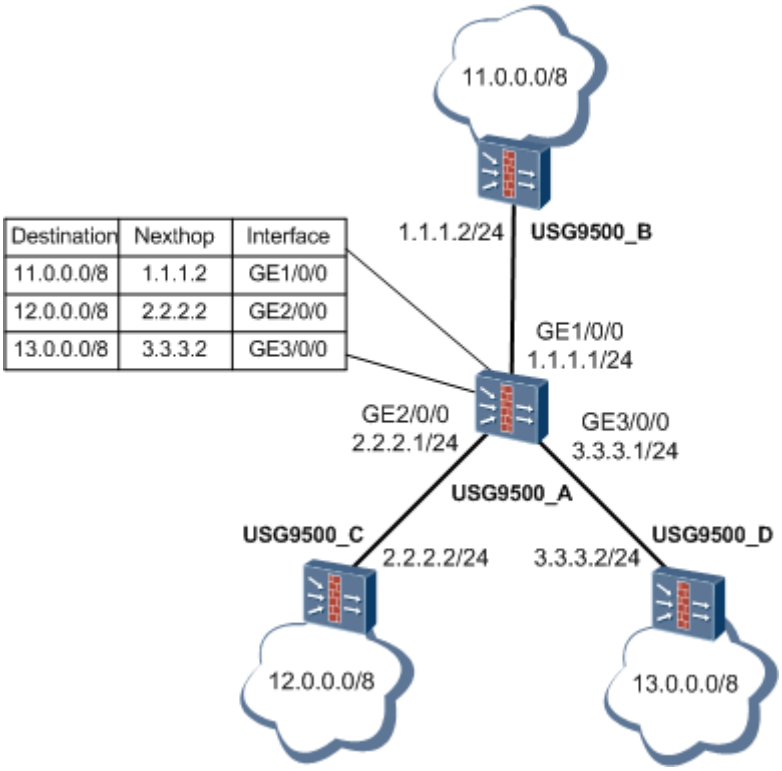
另外，根据目的地与该路由器是否直接相连，又可分为：

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

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

在图 3-1 所示的网络中，USG9500_A 与三个网络相连，因此有三个 IP 地址和三个出接口，其路由表如图所示。

图3-1 路由表示意图



FIB 表的匹配

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

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

例如，一台路由器上的路由表简表如下：

```
Routing Tables:
Destination/Mask    Proto  Pre  Cost      Flags NextHop         Interface
 0.0.0.0/0    Static   60   0           D   120.0.0.2       Pos1/0/0
 8.0.0.0/8    RIP      100  3           D   120.0.0.2       Pos1/0/0
 9.0.0.0/8    OSPF     10   50          D   20.0.0.2        Ethernet1/0/0
 9.1.0.0/16   RIP      100  4           D   120.0.0.2       Pos2/0/0
 20.0.0.0/8   Direct   0    0           D   20.0.0.1        Ethernet2/0/0
```

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]        Pos1/0/0       0x0
8.0.0.0/8          120.0.0.2      DU   t[37]        Pos1/0/0       0x0
9.0.0.0/8          20.0.0.2       DU   t[9992]      Ethernet1/0/0  0x0
9.1.0.0/16         120.0.0.2      DU   t[9992]      Pos2/0/0       0x0
20.0.0.0/8         20.0.0.1       U    t[9992]      Ethernet2/0/0  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，USG9500 会选择最长匹配 9.1.0.0/16 表项，从接口 Pos2/0/0 转发这条目的地址是 9.1.2.1 的报文。

3.1.2 静态路由与动态路由

USG9500 不仅支持静态路由，同时也支持 RIP（Routing Information Protocol）、OSPF、IS-IS 和 BGP（Border Gateway Protocol）等动态路由协议。

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

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

3.1.3 动态路由协议的分类

对动态路由协议的分类可以采用以下不同标准：

根据作用范围

根据作用的范围，路由协议可分为：

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

根据使用的算法

根据使用的算法，路由协议可分为：

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

以上两种算法的主要区别在于发现路由和计算路由的方法。

根据目的地址类型

根据目的地址的类型，路由协议可分成：

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

3.1.4 路由协议及路由优先级

路由优先级

对于相同的目的地，不同的路由协议（包括静态路由）可能会发现不同的路由，但这些路由并不都是最优的。事实上，在某一时刻，到某一目的地的当前路由仅能由唯一的路由协议来决定。为了判断最优路由，各路由协议（包括静态路由）都被赋予了一个优先级，当存在多个路由信息源时，具有较高优先级（取值较小）的路由协议发现的路由将成为最优路由。各种路由协议及其发现路由的缺省优先级如表 3-1 所示。

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

表3-1 路由协议及缺省时的路由优先级

路由协议或路由种类	相应路由的优先级
DIRECT	0
OSPF	10
IS-IS	15
STATIC	60
RIP	100
OSPF ASE	150
OSPF NSSA	150
IBGP	255
EBGP	255

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

USG9500 分别定义了外部优先级和内部优先级，外部优先级即前面提到的用户为各路由协议配置的优先级，缺省情况下如表 3-1 所示。

当不同的路由协议配置了相同的优先级后，系统会通过内部优先级决定哪个路由协议发现的路由将成为最优路由。路由协议的内部优先级如表 3-2 所示。

表3-2 路由协议内部优先级

路由协议或路由种类	相应路由的优先级
DIRECT	0
OSPF	10
IS-IS Level-1	15
IS-IS Level-2	18
STATIC	60
RIP	100
OSPF ASE	150
OSPF NSSA	150
IBGP	200
EBGP	20

例如，到达同一目的地 10.1.1.0/24 有两条路由可供选择，一条静态路由，另一条是 OSPF 路由，且这两条路由的协议优先级都被配置成 5。这时 USG9500 系统将根据表 3-2 所示的内部优先级进行判断。因为 OSPF 协议的内部优先级是 10，高于静态路由的内部优先级 60。所以系统选择 OSPF 协议发现的路由作为可用路由。

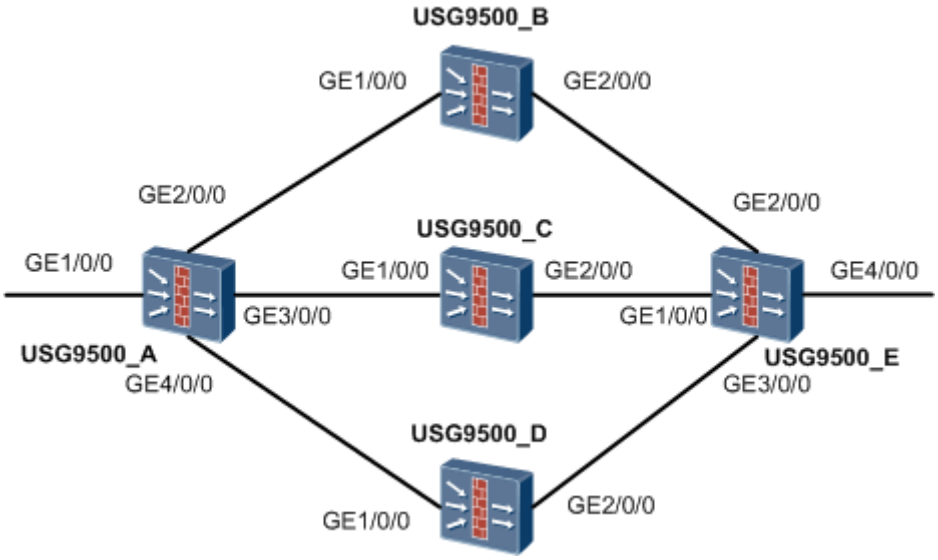
3.1.5 负载分担与路由备份

负载分担

USG9500 支持多路由模式，即允许配置多条目的地相同且优先级也相同的路由。当到同一目的地没有更高优先级路由时，这几条路由都被采纳，在转发去往该目的地报文时，由 IP 依次通过各条路径发送，从而实现网络的负载分担。

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

图3-2 路由负载分担示意图



如图 3-2 所示，负载分担的组网环境如下（以 OSPF 为例）：

- 在 USG9500_A、USG9500_B、USG9500_C、USG9500_D 和 USG9500_E 上配置 OSPF 路由协议，OSPF 会发现三条不同的路由。
- 目的从 GE1/0/0 接口进入 USG9500_A 去往 USG9500_E 的报文，会根据具体的负载分担方式，依次通过这三条路由发送，从而实现负载分担。

路由备份

设备支持路由备份功能，提高网络的可靠性。用户可根据实际情况，配置到同一目的地的多条路由，其中一条路由的优先级最高，做为主路由，其余的路由优先级较低，做为备份路由。

正常情况下，USG9500 采用主路由转发数据。当线路故障时，该路由变为非激活状态，USG9500 选择备份路由中优先级最高的转发数据。这样，也就实现了主路由到备份路由的切换。当主路由恢复正常时，USG9500 重新选择路由。由于该路由的优先级最高，USG9500 选择主路由来发送数据。

3.1.6 缺省路由

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

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

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

3.2 静态路由

3.2.1 介绍

定义

静态路由是一种需要管理员手工配置的特殊路由。

目的

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

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

USG9500 支持普通静态路由，也支持与 VPN 实例关联的静态路由，后者主要用于 VPN 路由的管理。

3.2.2 参考标准和协议

无

3.2.3 原理描述

3.2.3.1 静态路由的组成

在 USG9500 中，使用 **ip route-static** 命令配置静态路由，一条静态路由包含以下要素：

- 目的地址与掩码
- 出接口与下一跳地址

目的地址与掩码

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

出接口与下一跳地址

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

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

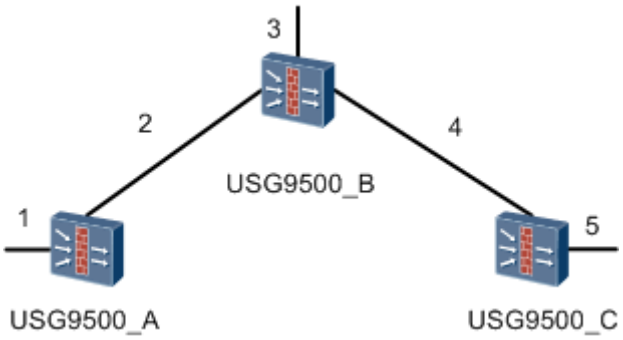
指定发送接口时需要注意：

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

3.2.3.2 静态路由的应用

如图 3-3 所示，该网络结构比较简单，可以使用静态路由实现网络互通。首先需要确定每个物理网络的地址，并为每个路由器标识出非直连的物理网络，最后为每个非直连的物理网络配置静态路由命令。

图3-3 静态路由组网图



此例中需要在 USG9500_A 上配置到网络 3、4、5 的静态路由，在 USG9500_B 上配置到网络 1、5 的静态路由，在 USG9500_C 上配置到网络 1、2、3 的静态路由。

缺省静态路由

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

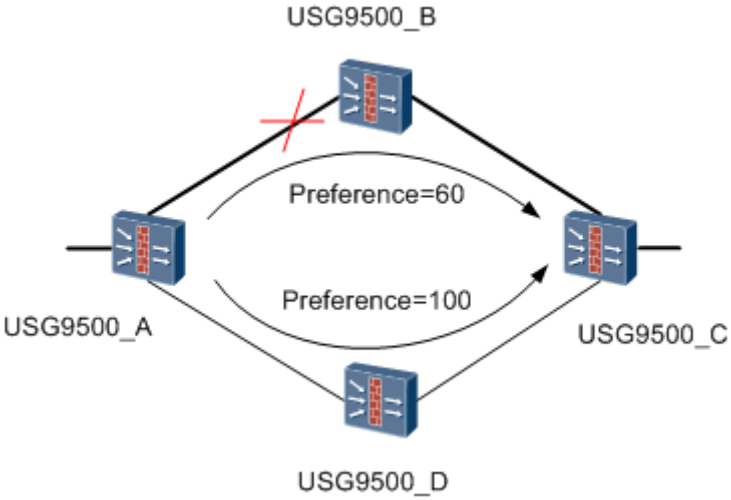
图 1 中，因为 USG9500_A 发往 3、4、5 网络的报文下一跳都是 USG9500_B，因此可在 USG9500_A 上配置一条缺省路由，代替上个例子中通往 3、4、5 网络的 3 条静态路由。同理，USG9500_C 也只需要配置一条到 USG9500_B 的缺省路由，代替上个例子中通往 1、2、3 网络的 3 条静态路由。

浮动静态路由

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

如图 3-4，从 USG9500_A 到 USG9500_C 有两条静态路由。在正常情况下，路由表上仅下一跳是 USG9500_B 的静态路由的状态为“Active”，因为这条路由具有更高的优先级。另一条下一跳是 USG9500_D 的静态路由则作为备份路由，只有在主链路上出现故障的时候，备份路由才会被激活，承担数据转发的业务。在主链路恢复正常后，路由表又恢复到原来的样子，即下一跳是 USG9500_B 的静态路由又成为活跃路由来承担数据转发，因此这条备份路由也叫做浮动静态路由。USG9500_B 和 USG9500_C 之间链路发生故障时，浮动静态路由就无能为力了。

图3-4 浮动静态路由

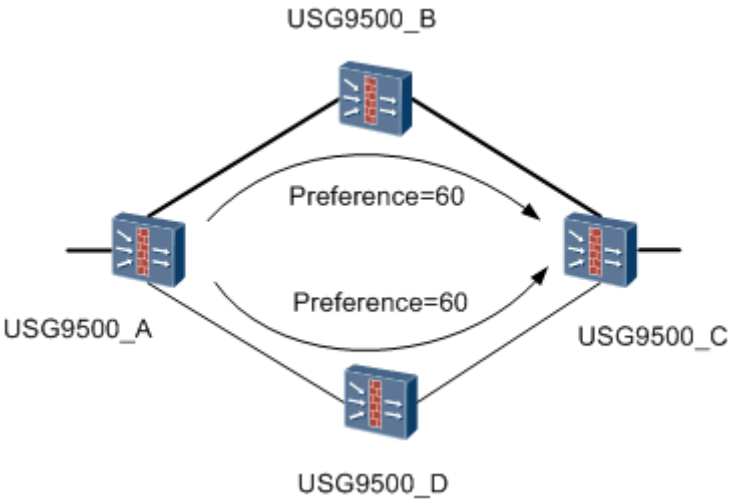


静态路由负载分担

配置到达相同目的地的多条路由，如果指定相同优先级，则可实现负载分担。

如图 3-5，从 USG9500_A 到 USG9500_C 有两条优先级相同的静态路由。两条路由都会出现在路由表上，同时进行数据的转发。

图3-5 静态路由负载分担



3.2.3.3 静态路由特性

IPv4 静态路由

USG9500 支持普通静态路由，也支持与 VPN 实例关联的静态路由，后者主要用于 VPN 路由的管理。

IPv6 静态路由属性及功能

IPv6 静态路由与 IPv4 静态路由类似，也需要管理员手工配置，适合于一些结构比较简单的 IPv6 网络。

它们之间的主要区别是目的地址和下一跳地址有所不同，IPv6 静态路由使用的是 IPv6 地址，而 IPv4 静态路由使用 IPv4 地址。

在配置 IPv6 静态路由时，如果指定的目的地址为::/0（掩码长度为 0），则表示配置了一条 IPv6 缺省路由。如果报文的目的地址无法匹配路由表中的任何一项，路由器将选择 IPv6 缺省路由来转发 IPv6 报文。

3.2.3.4 BFD for 静态路由

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

BFD for 静态路由可为每条静态路由绑定一个 BFD 会话。

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

BFD for 静态路由有单跳检侧和多跳检测两种方式。

- 单跳检测

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

- 多跳检测

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

3.3 OSPF

3.3.1 介绍

定义

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

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

目的

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

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

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

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

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

3.3.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC1587	This document describes a new optional type of OSPF area, somewhat humorously	

文档	描述	备注
	referred to as a "not-so-stubby" area (or NSSA). NSSAs are similar to the existing OSPF stub area configuration option but have the additional capability of importing AS external routes in a limited fashion.	
RFC1765	Proper operation of the OSPF protocol requires that all OSPF routers maintain an identical copy of the OSPF link-state database. However, when the size of the link-state database becomes very large, some routers may be unable to keep the entire database due to resource shortages; we term this "database overflow".	该 RFC 为 Experimental，非 Standard。
RFC2328	This memo documents version 2 of the OSPF protocol. OSPF is a link-state routing protocol.	
RFC2370	This memo defines enhancements to the OSPF protocol to support a new class of link-state advertisements (LSA) called Opaque LSAs. Opaque LSAs provide a generalized mechanism to allow for the future extensibility of OSPF.	
RFC3137	This memo describes a backward-compatible technique that may be used by OSPF (Open Shortest Path First) implementations to advertise unavailability to forward transit traffic or to lower the preference level for the paths through such a router.	该 RFC 为 Informational，非 Standard。
RFC3623	This memo documents an enhancement to the OSPF routing protocol, whereby an OSPF router can stay on the forwarding path even as	

文档	描述	备注
	its OSPF software is restarted.	
RFC3630	This document describes extensions to the OSPF protocol version 2 to support intra-area Traffic Engineering (TE), using Opaque Link State Advertisements.	
RFC3682	The use of a packet's Time to Live (TTL) (IPv4) or Hop Limit (IPv6) to protect a protocol stack from CPU-utilization based attacks has been proposed in many settings.	该 RFC 为 Experimental，非 Standard。
RFC3906	This document describes how conventional hop-by-hop link-state routing protocols interact with new Traffic Engineering capabilities to create Interior Gateway Protocol (IGP) shortcuts.	
RFC4576	This document specifies the necessary procedure, using one of the options bits in the LSA (Link State Advertisements) to indicate that an LSA has already been forwarded by a PE and should be ignored by any other PEs that see it.	
RFC4577	This document extends that specification by allowing the routing protocol on the PE/CE interface to be the Open Shortest Path First (OSPF) protocol.	
RFC4750	This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in TCP/IP-based internets. In particular, it defines objects for managing version 2 of the Open Shortest Path First Routing Protocol. Version 2	

文档	描述	备注
	of the OSPF protocol is specific to the IPv4 address family.	

3.3.3 原理描述

3.3.3.1 OSPF 基础

OSPF 协议具有以下特点：

- OSPF 把自治系统划分成逻辑意义上的一个或多个区域；
- OSPF 通过 LSA（Link State Advertisement)的形式发布路由；
- OSPF 依靠在 OSPF 区域内各路由设备间交互 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3-3 OSPF 报文类型

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

LSA 类型

表3-4 OSPF LSA 类型

LSA 类型	LSA 作用
Router-LSA（Type1）	每个路由设备都会产生，描述了路由设备的链路状态和开销，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，描述区域内某个网段的路由，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。
Opaque LSA（Type9/Type10/Type11）	Opaque LSA 提供用于 OSPF 的扩展的通用机制。其中： Type9 LSA 仅在接口所在网段范围内传播。用于支持 GR 的 Grace LSA 就是 Type9 LSA 的一种。 Type10 LSA 在区域内传播。用于支持 TE 的 LSA 就是 Type10 LSA 的一种。 Type11 LSA 在自治域内传播，目前还没有实际应用的例子。

路由器类型

OSPF 协议中常用到的路由设备类型如图 3-6 所示。

图3-6 路由器类型

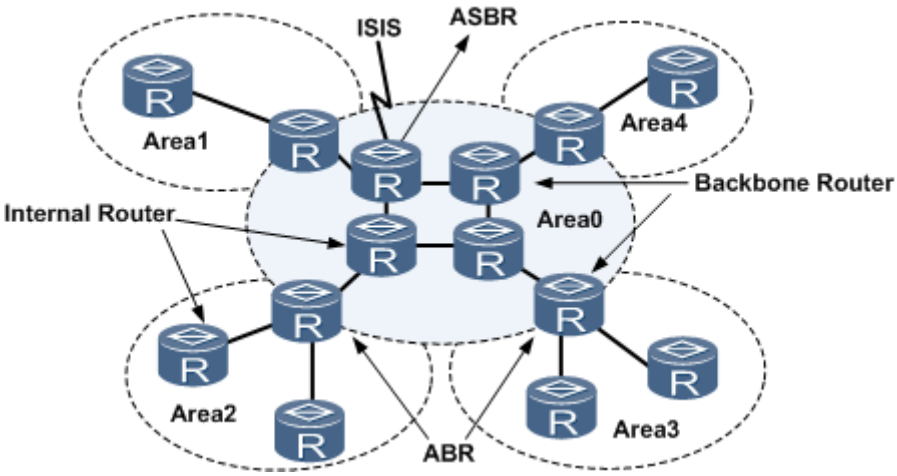


表3-5 OSPF 路由器类型

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

OSPF 路由类型

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

表 3-6 中按优先级从高到低顺序列出了路由类型。

表3-6 OSPF 路由类型

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

区域类型

表3-7 OSPF 区域类型

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

OSPF 支持的网络类型

OSPF 根据链路层协议类型，将网络分为如表 3-8 所列四种类型。

表3-8 OSPF 网络类型

网络类型	含义
广播类型（Broadcast）	当链路层协议是 Ethernet、FDDI 时，缺省情况下，OSPF 认为网络类型是 Broadcast。 在该类型的网络中：

网络类型	含义
	• 通常以组播形式发送 Hello 报文、LSU 报文和 LSAck 报文。其中，224.0.0.5 的组播地址为 OSPF 路由器的预留 IP 组播地址；224.0.0.6 的组播地址为 OSPF DR 的预留 IP 组播地址。 • 以单播形式发送 DD 报文和 LSR 报文。
NBMA 类型（Non-broadcast multiple access）	当链路层协议是帧中继、ATM 或 X.25 时，缺省情况下，OSPF 认为网络类型是 NBMA。 在该类型的网络中，以单播形式发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。
点到多点 P2M 类型（Point-to-Multipoint）	没有一种链路层协议会被缺省的认为是 Point-to-Multipoint 类型。点到多点必须是由其他的网络类型强制更改的。常用做法是将非全连通的 NBMA 改为点到多点的网络。 在该类型的网络中： • 以组播形式（224.0.0.5）发送 Hello 报文； • 以单播形式发送其他协议报文（DD 报文、LSR 报文、LSU 报文、LSAck 报文）。
点到点 P2P 类型（point-to-point）	当链路层协议是 PPP、HDLC 和 LAPB 时，缺省情况下，OSPF 认为网络类型是 P2P。 在该类型的网络中，以组播形式（224.0.0.5）发送协议报文（Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文）。

Stub 区域

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

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

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

配置 Stub 区域时需要注意下列几点：

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

OSPF 报文认证

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

USG9500 支持两种验证方式：

- 区域验证方式
- 接口验证方式

USG9500 支持的验证模式按加密算法不同分为 null、simple、MD5 以及 HMAC-MD5。

当两种验证方式都存在时，优先使用接口验证方式。

OSPF 路由聚合

路由聚合是指将具有相同前缀的路由信息聚合在一起，只发布一条路由到其它区域。

通过路由聚合，可以减少路由信息，从而减小路由表的规模，提高路由器的性能。

OSPF 有两种路由聚合方式：

- ABR 聚合

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

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

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

OSPF 缺省路由

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

OSPF 缺省路由通常应用于下面两种情况：

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

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

OSPF 缺省路由的发布原则如下：

- OSPF 路由器只有具有对外的出口时，才能够发布缺省路由 LSA。
- 如果 OSPF 路由器已经发布了缺省路由 LSA，那么不再学习其它路由器发布的相同类型缺省路由。即路由计算时不再计算其它路由器发布的相同类型的缺省路由 LSA，但数据库中存有对应 LSA。
- 外部缺省路由的发布如果要依赖于其它路由，那么被依赖的路由不能是本 OSPF 路由域内的路由，即不是本进程 OSPF 学习到的路由。因为外部缺省路由的作用是用于指导报文的域外转发，而本 OSPF 路由域的路由的下一跳都指向了域内，不能满足指导报文域外转发的要求。

不同区域缺省路由发布原则如表 3-9 所示。

表3-9 不同区域的缺省路由发布原则

区域类型	缺省路由发布原则
普通区域	缺省情况下，普通 OSPF 区域内的 OSPF 路由器是不会产生缺省路由的，即使它有缺省路由。 当网络中缺省路由通过其他路由进程产生时，路由器必须将缺省路由通告到整个 OSPF 自治域中。实现方法是在 ASBR 上手动通过命令进行配置，产生缺省路由。配置完成后，路由器会产生一个缺省 ASE LSA（Type5 LSA），并且通告到整个 OSPF 自治域中。 如果 ASBR 上没有缺省路由，则路由器不会通告缺省路由。
Stub Area	Stub 区域不允许自治系统外部的路由（Type5 LSA）在区域内传播。 区域内的路由器必须通过 ABR 学到自治系统外部的路由。实现方法是 ABR 会自动产生一条缺省的 Summary LSA（Type3 LSA）通告到整个 Stub 区域内。这样，到达自治系统的外部路由就可以通过 ABR 到达。
Totally Stub Area	Totally Stub 区域既不允许自治系统外部的路由（Type5 LSA）在区域内传播，也不允许区域间路由（Type3 LSA）在区域内传播。 区域内的路由器必须通过 ABR 学到自治系统外部和其他区域的路由。实现方法是配置 Totally Stub 区域后，ABR 会自动产生一条缺省的 Summary LSA（Type3 LSA）通告到整个 Stub 区域内。这样，到达自治系统外部的路由和其他区域间的路由都可以通过 ABR 到达。
NSSA Area	NSSA 区域允许引入少量通过本区域的 ASBR 到达的外部路由，但不允许其他区域的外部路由 ASE LSA（Type5 LSA）在区域内传播。即到达自治系统外部的路由只能通过本区域的 ASBR 到达。 只配置了 NSSA 区域是不会自动产生缺省路由的。 此时，有两种选择： • 如果希望到达自治系统外部的路由通过该区域的 ASBR 到达，而其它外部路由通过其它区域出去。则必须在 ABR 上

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

OSPF 路由过滤

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

OSPF 可以使用的路由策略包括 route-policy，访问控制列表（access-list），地址前缀列表（prefix-list）。

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3-10 区域间 LSA 学习与路由学习的差异

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

- 区域间 LSA 发布

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

OSPF 虚连接

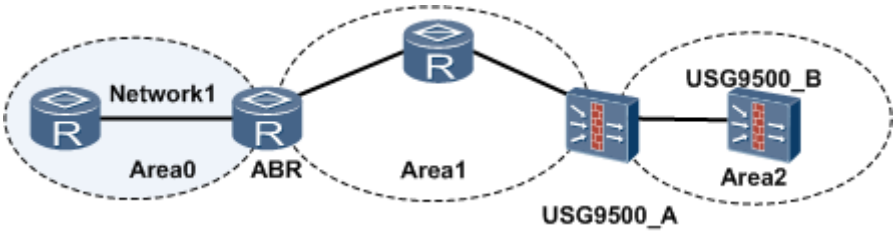
虚连接（Virtual link）是指在两台 ABR 之间通过一个非骨干区域建立的一条逻辑上的连接通道。

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

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

如图 3-7 中所示，Area2 没有连接到骨干区 Area0，USG9500_A 不是 ABR，因此不会向 Area2 生成 Area0 中 Network1 的路由信息，所以 USG9500_B 上没有到达 Network1 的路由。

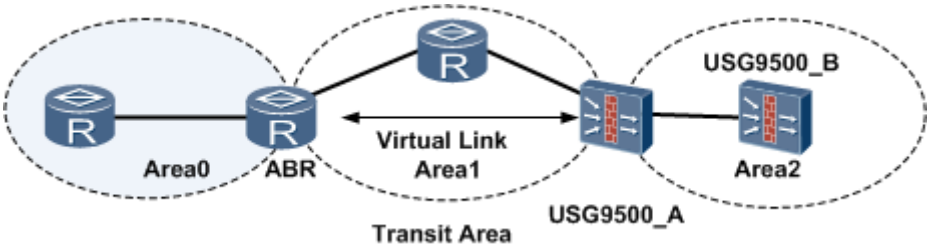
图3-7 OSPF 非骨干区没有连接骨干区



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

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

图3-8 OSPF 虚连接



如图 3-8 所示，通过虚连接，两台 ABR 之间直接传递 OSPF 报文信息，他们之间的 OSPF 路由设备只是起到一个转发报文的作用。由于 OSPF 协议报文的目的地址不是这些路由设备，所以这些报文对于他们而言是透明的，只是当作普通的 IP 报文来转发。

OSPF 多进程

OSPF 支持多进程，在同一台路由设备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。

路由设备的一个接口只能属于某一个 OSPF 进程。

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

3.3.3.2 OSPF GR

随着路由设备普遍采用了控制和转发分离的技术，在网络拓扑保持稳定的情况下，控制层面的重启并不会影响转发层面，转发层面仍然可以很好地完成数据转发任务，从而保证业务不受影响。 GR 技术保证了在重启过程中转发层面能够继续指导数据的转发，同时控制层面邻居关系的重建以及路由计算等动作不会影响转发层面的功能，从而避免了路由震荡引发的业务中断，提高了整网的可靠性。

基本概念

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

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

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

和 GR 相关的概念如下：

- Grace-LSA

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

- 路由器在 GR 中的角色

  Restarter：重启路由器。可以通过配置支持完全 GR 或者部分 GR。

  Helper：协助重启路由器。可以通过配置支持有计划 GR、无计划 GR 或者通过策略有选择支持 GR。

- GR 的原因

  Unknown：未知原因导致的 GR 操作。

  Software restart：通过命令行主动触发的 GR 操作。

  Software reload/upgrade：软件重启或升级导致的 GR 操作。

  Switch to redundant control processor：异常主备倒换导致的 GR 操作。

- GR 的持续时间

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

GR 的分类

完全 GR（Totally GR)：指当有一个邻居不支持 GR 功能时，整个路由器退出 GR 状态。

部分 GR（Partly GR)：指当有一个邻居不支持 GR 时，仅该邻居所关联的接口退出 GR，其它接口正常进行 GR 过程。

有计划 GR（Planned-GR)：指手动通过命令使路由器执行重启或主备倒换。在进行重启或主备倒换前 Restarter 会先发送 Grace-LSA。

非计划 GR（Unplanned-GR)：与 Planned-GR 的区别在于，路由器是由于故障等原因进行重启或主备倒换，并且在主备倒换前不会事先发送 Grace-LSA，而是直接开始主备倒换，在备板正常 Up 后才进入 GR 过程。以下的步骤同 Planned-GR。

GR 的过程

- GR 开始

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

  对于 Unplanned-GR，则不发送这个 Grace-LSA。

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

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

- GR 退出

表3-11 GR 退出原因

GR 执行情况	Restarter	Helper
GR 成功	Restarter 在 GR 超时前与主备倒换前的所有邻居都重新建立好邻居关系。	收到 Restarter 发送的 Age 为 3600 秒的 Grace-LSA 时与 Restarter 的邻居关系为 Full 状态。
GR 失败	• GR 超时并且邻居关系尚未完全恢复。 • Helper 发送的 Router-LSA 或 Network-LSA 导致 Restarter 端进行双向检查时失败。 • Restarter 接口状态变化。 • Restarter 收到 Helper 发送的 1-way Hello 报文。 • Restarter 收到同一网段上另一台路由器产生的 Grace-LSA。同一网段同一时间只能有一台路由器做 GR。 • Restarter 同一个网段的邻居之间存在 DR/BDR 不一致的情况（拓扑变化）。	• 在邻居关系超时前没有收到 Restarter 发送的 Grace-LSA。 • Helper 接口状态发生变化。 • 收到其它路由器发送的与 Helper 本地数据库不一致的 LSA。（可以通过配置不进行严格 LSA 检查排除这种情况。） • 同一网段上同一时间收到两台路由器发送的 Grace-LSA。 • 与其它路由器邻居关系变化。

有无 GR 技术的比较

表3-12 有无 GR 技术的比较

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

3.3.3.3 OSPF NSSA

定义

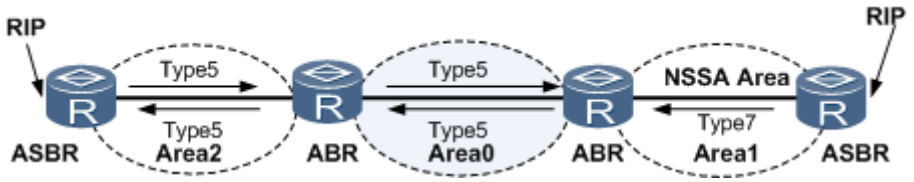
OSPF NSSA 区域（Not-So-Stubby Area）是 OSPF 新增的一类特殊的区域类型。

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

目的

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

图3-9 NSSA 区域



Type-7 LSA

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

Type-7 LSA 转化为 Type-5 LSA

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

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

缺省路由环路预防

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

3.3.3.4 BFD for OSPF

定义

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

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

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

目的

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

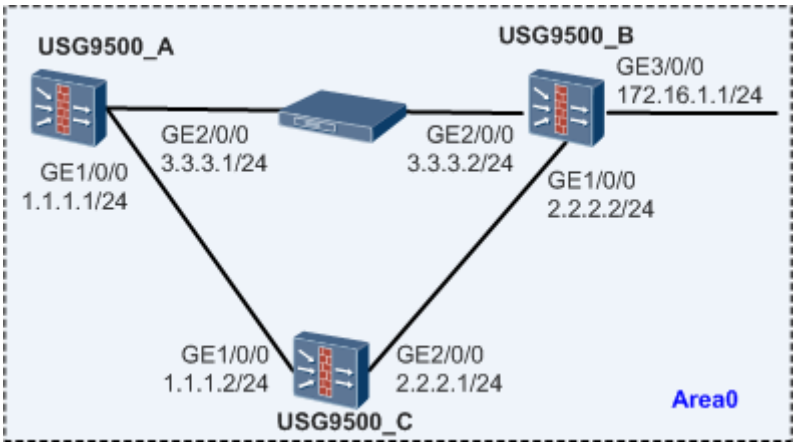
由于链路故障是无法完全避免的，因此，加快故障感知速度并将故障快速通告给路由协议是一种可行的方案。BFD 和路由协议相关联，一旦链路出现故障，BFD 的快速性能够加快路由协议的收敛速度。

表3-13 BFD for OSPF

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

原理

图3-10 BFD for OSPF



BFD for OSPF 的原理如图 3-10 所示：

1. 三台设备间建立 OSPF 邻居关系。
2. 邻居状态到达 Full 状态时通知 BFD 建立 BFD 会话。
3. USG9500_A 到 USG9500_B 的路由出接口为 GE2/0/0，当这两台设备间的链路出现故障后，BFD 首先感知到并通知 USG9500_A。
4. USG9500_A 处理邻居 Down 事件，重新进行路由计算，新的路由出接口为 GE1/0/0，经过 USG9500_C 到达 USG9500_B。

3.3.3.5 OSPF-BGP 联动

定义

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

通过使能 OSPF-BGP 联动特性可以解决这个问题。

目的

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

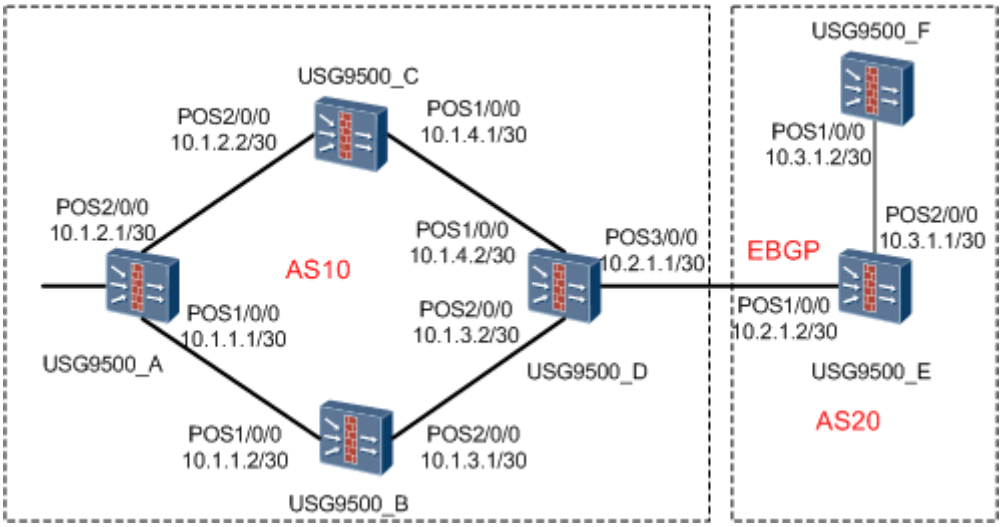
如图 3-11 所示，四台设备 USG9500_A、USG9500_B、USG9500_C、USG9500_D 之间运行 OSPF 协议，并建立 IBGP 连接。USG9500_C 为 USG9500_B 的备份设备。当网络环境稳定时，BGP 与 OSPF 在设备上是完全收敛的。

正常情况下，从 USG9500_A 到 10.3.1.0/30 的流量会途经 USG9500_B。当 USG9500_B 发生故障后，流量切换到 USG9500_C。USG9500_B 故障恢复以后，流量回切到 USG9500_B，此时会有流量丢失。

这是因为，在流量回切到 USG9500_B 的过程中，IGP 收敛速度比 BGP 快，因此 OSPF 先收敛，BGP 还没有完成收敛，导致 USG9500_B 不知如何到达 10.3.1.0/30。

这样，当从 USG9500_A 去往 10.3.1.0/30 的流量被发送给 USG9500_B 时，由于没有必要的路由选择信息，这些流量就会被丢弃。

图3-11 OSPF-BGP 联动



原理

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

图 3-11 中，在 USG9500_B 上使能 BGP 联动，这样，在 BGP 收敛完成前，USG9500_A 不把流量转发到 USG9500_B 上，而是继续使用备份链路 USG9500_C，直到 USG9500_B 上的 BGP 路由完成收敛

3.4 OSPFv3

3.4.1 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（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。

3.4.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC2740	This document describes the modifications to OSPF to support version 6 of the Internet Protocol (IPv6).	
RFC5187	This document describes the OSPFv3 graceful restart. The OSPFv3 graceful restart is identical to OSPFv2 except for the differences described in this document. These differences include the format of the grace Link State Advertisements (LSA) and other considerations.	
RFC5340	This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in IPv6-based internets. In particular, it defines objects for managing the Open Shortest Path First Routing Protocol for IPv6.	

3.4.3 原理描述

3.4.3.1 OSPFv3 基本原理

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

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

OSPFv3 报文类型

表3-14 OSPFv3 报文类型

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

LSA 类型

表3-15 LSA 类型

LSA 类型	LSA 作用
Router-LSA（Type1）	路由器会为每个运行 OSPFv3 接口所在的区域产生一个 LSA，描述了路由器的链路状态和开销，在所属的区域内传播。
Network-LSA（Type2）	由 DR 产生，描述本链路的链路状态，在所属的区域内传播。
Inter-Area-Prefix-LSA（Type3）	由 ABR 产生，描述区域内某个网段的路由，并通告给其他相关区域。
Inter-Area-Router-LSA（Type4）	由 ABR 产生，描述到 ASBR 的路由，通告给除 ASBR 所在区域的其他相关区域。

LSA 类型	LSA 作用
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）。
Link-LSA（Type8）	每个路由器都会为每个链路产生一个 Link-LSA，描述到此 Link 上的 link-local 地址、IPv6 前缀地址，并提供将会在 Network-LSA 中设置的链路选项，它仅在此链路内传播。
Intra-Area-Prefix-LSA（Type9）	每个路由器及 DR 都会产生一个或多个此类 LSA，在所属的区域内传播。 • 路由器产生的此类 LSA，描述与 Route-LSA 相关联的 IPv6 前缀地址。 • DR 产生的此类 LSA，描述与 Network-LSA 相关联的 IPv6 前缀地址。

路由器类型

图3-12 路由器类型

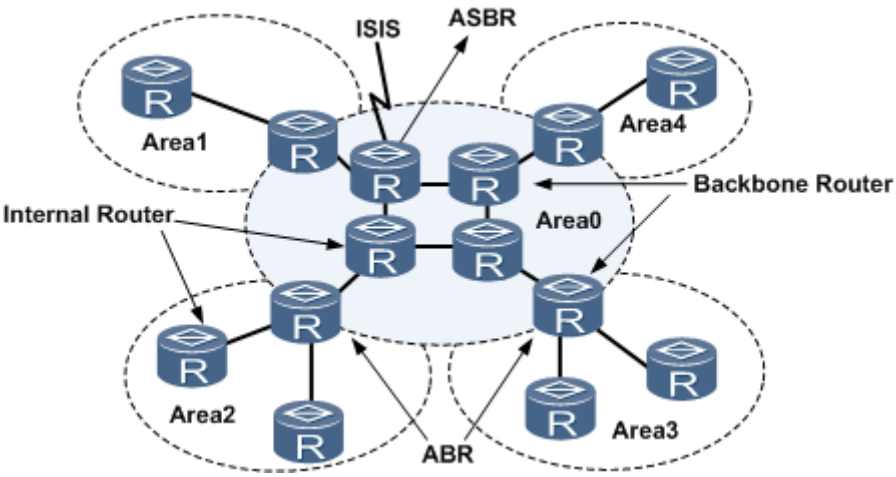


表3-16 路由器的类型及含义

路由器类型	含义
区域内路由器（Internal Router）	该类路由器的所有接口都属于同一个 OSPFv3 区域。
区域边界路由器 ABR（Area Border Router）	该类路由器可以同时属于两个以上的区域，但其中一个必须是骨干区域。 ABR 用来连接骨干区域和非骨干区域，它与骨干区域之间既可以是物理连接，也可以是逻辑上的连接。

路由器类型	含义
骨干路由器（Backbone Router）	该类路由器至少有一个接口属于骨干区域。 因此，所有的 ABR 和位于 Area0 的内部路由器都是骨干路由器。
自治系统边界路由器 ASBR（AS Boundary Router）	与其他 AS 交换路由信息的路由器称为 ASBR。 ASBR 并不一定位于 AS 的边界，它可能是区域内路由器，也可能是 ABR。只要一台 OSPFv3 路由器引入了外部路由的信息，它就成为 ASBR。

OSPFv3 路由类型

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

表 3-17 中按优先级从高到低顺序列出了路由类型。

表3-17 OSPFv3 路由类型

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

区域类型

表3-18 OSPFv3 区域类型

区域类型	作用
Totally Stub Area	允许 ABR 发布的 Type3 缺省路由，不允许自治系统外部路由和区域间的路由。
Stub Area	和 Totally Stub 区域的不同在于，该区域允许区域间路由。

OSPFv3 支持的网络类型

OSPFv3 根据链路层协议类型，将网络分为如表 3-19 所列四种类型。

表3-19 OSPFv3 网络类型

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

Stub 区域

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

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

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

配置 Stub 区域时需要注意下列几点：

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

OSPFv3 路由聚合

路由聚合是指将具有相同前缀的路由信息聚合在一起，只发布一条路由到其它区域。

通过路由聚合，可以减少路由信息，从而减小路由表的规模，提高路由器的性能。

OSPFv3 路由聚合过程如下：

- ABR 向其它区域发送路由信息时，以 IPv6 地址前缀为单位生成 Type3 LSA；
- 如果该区域中存在一些连续的 IPv6 地址前缀，则将这些连续的前缀聚合成一个前缀；
- ABR 只发送一条聚合后的 LSA，所有属于本命令指定的聚合前缀范围的 LSA 将不再会被单独发送出去。

OSPFv3 虚连接

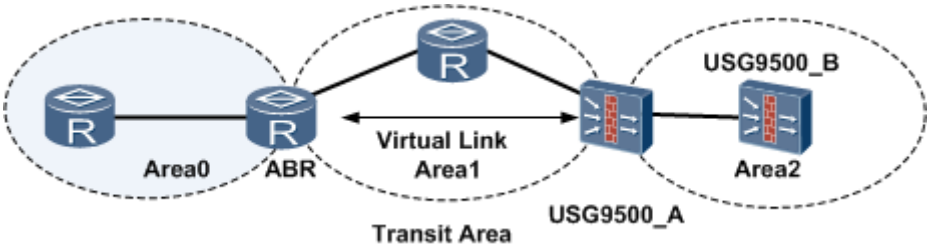
虚连接（Virtual link）是指在两台 ABR 之间通过一个非骨干区域建立的一条逻辑上的连接通道。

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

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

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

图3-13 OSPFv3 虚连接



如图 3-13 所示，通过虚连接，两台 ABR 之间直接传递 OSPFv3 报文信息，他们之间的 OSPFv3 设备只是起到一个转发报文的作用。由于 OSPFv3 协议报文的目的地址不是这些设备，所以这些报文对于他们而言是透明的，只是当作普通的 IP 报文来转发。

OSPFv3 多进程

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

路由器的一个接口只能属于某一个 OSPFv3 进程。

3.4.3.2 OSPFv3 GR

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

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

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

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

基本概念

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

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

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

GR 过程

图3-14 OSPFv3 Planned-GR 过程（reset ospfv3 graceful-restart）

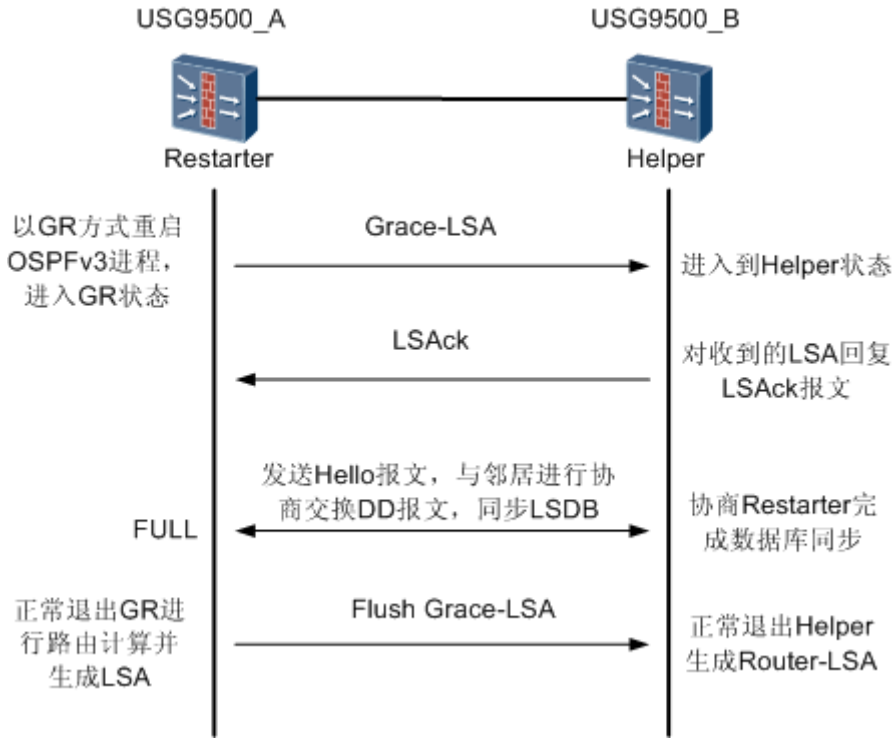
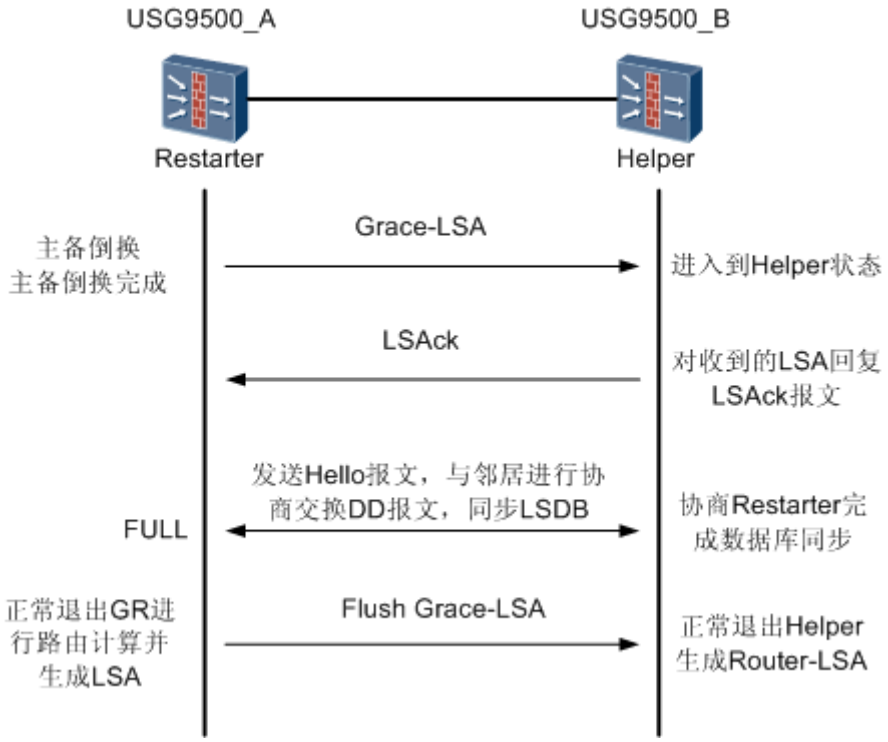


图3-15 OSPFv3 Unplanned-GR 过程（主备倒换）



- Restarter 端：

1. 对于 Planned-GR，Restarter 会首先向每个邻居发送一个 Grace-LSA 通知邻居 GR 的开始以及 GR 的周期、原因等。

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

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

- Helper 端：

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

5. 如果 GR 过程出错，则退出 Helper 状态，重新进行路由计算，删除失效的路由。

有无 GR 技术的比较

表3-20 有无 OSPFv3 GR 的比较

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

3.4.3.3 BFD for OSPFv3

定义

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

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

BFD for OSPFv3 就是将 BFD 和 OSPFv3 协议关联起来，将 BFD 对链路故障的快速感应通知 OSPFv3 协议，从而加快 OSPFv3 协议对于网络拓扑变化的响应。

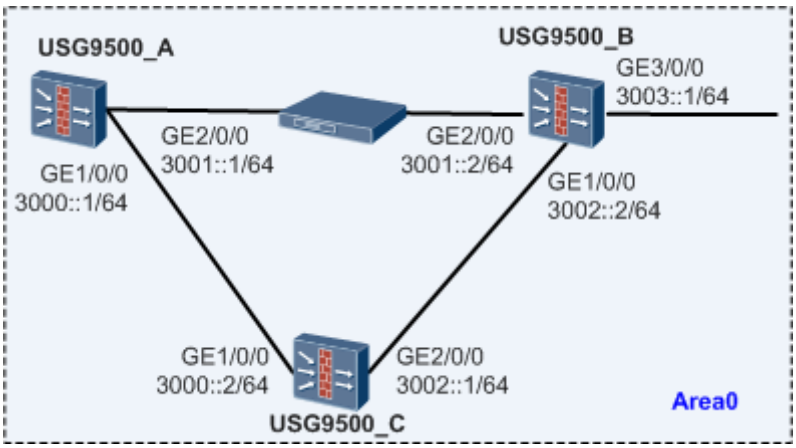
目的

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

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

原理

图3-16 BFD for OSPFv3



BFD for OSPFv3 的原理如图 3-16 所示：

1. 三台设备间建立 OSPFv3 邻居关系。
2. 邻居状态到达 Full 状态时通知 BFD 建立 BFD 会话。
3. USG9500_A 到 USG9500_B 的路由出接口为 GE2/0/0，当这两台设备间的链路出现故障后，BFD 首先感知到并通知 USG9500_A。
4. USG9500_A 处理邻居 Down 事件，重新进行路由计算，新的路由出接口为 GE1/0/0，经过 USG9500_C 到达 USG9500_B。

3.4.3.4 OSPFv3 和 OSPFv2 协议比较

相同点：

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

不同点：

- OSPFv3 基于链路，而不是网段

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

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

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

- OSPFv3 的报文及 LSA 格式发生改变
  - OSPFv3 报文不包含 IP 地址。
  - OSPFv3 的 Router LSA 和 Network LSA 里不包含 IP 地址。IP 地址部分由新增的两类 LSA（Link LSA 和 Intra Area Prefix LSA）宣告。
  - OSPFv3 的 Router ID、Area ID 和 LSA Link State ID 不再表示 IP 地址，但仍保留 IPv4 地址格式。
  - 广播、NBMA 及 P2MP 网络中，邻居不再由 IP 地址标识，只由 Router ID 标识。
- OSPFv3 的 LSA 报文里添加 LSA 的洪泛范围

  OSPFv3 在 LSA 报文头的 LSA Type 里，添加 LSA 的洪泛范围，这使得 OSPFv3 的路由器更加灵活，可以处理不能识别类型的 LSA：
  - OSPFv3 可存储或洪泛不识别报文，而 OSPF 只简单丢弃掉不识别报文。
  - OSPFv3 允许洪泛范围为区域或链路本地（Link-local），并且设置 U 位（报文可按洪泛范围为链路本地来处理）的不识别报文存储或通过 Stub 区域。

  例如，USG9500_A 和 USG9500_B 都可识别某类 LSA，它们之间通过 USG9500_C 连接，但 USG9500_C 不识别该类 LSA。这样，当 USG9500_A 洪泛此类 LSA 时，USG9500_C 虽然不识别，但还是可以洪泛给 USG9500_B，USG9500_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，USG9500_B 则不能收到此类 LSA。
- 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。

3.5 IS-IS

3.5.1 介绍

定义

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

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

本文所指的 IS-IS，如不加特殊说明，均指集成 IS-IS。

目的

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

3.5.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
ISO 10589	ISO IS-IS Routing Protocol	
ISO 8348/Ad2	Network Services Access Points	
RFC 1195	Use of OSI IS-IS for Routing in TCP/IP and Dual Environments	不支持配置多个认证密码
RFC 2763	Dynamic Hostname Exchange Mechanism for IS-IS	
RFC 2966	Domain-wide Prefix Distribution with Two-Level IS-IS	
RFC 2973	IS-IS Mesh Groups	

文档	描述	备注
RFC 3277	IS-IS Transient Blackhole Avoidance	
RFC 3373	Three-Way Handshake for IS-IS Point-to-Point Adjacencies	
RFC 3567	Intermediate System to Intermediate System (IS-IS) Cryptographic Authentication	
RFC 3719	Recommendations for Interoperable Networks using IS-IS	
RFC 3784	IS-IS extensions for Traffic Engineering	
RFC 3786	Extending the Number of IS-IS LSP Fragments Beyond the 256 Limit	
RFC 3787	Recommendations for Interoperable IP Networks using IS-IS	
RFC 3847	Restart signaling for IS-IS	
RFC 3906	Calculating Interior Gateway Protocol (IGP) Routes Over Traffic Engineering Tunnels	
RFC 4444	Management Information Base for IS-IS	
RFC 5308	Routing IPv6 with IS-IS	

3.5.3 原理描述

3.5.3.1 IS-IS 基本概念

IS-IS 的发展

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

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

OSI 协议的几个常用术语：

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

OSI 通过 CLNP 实现 CLNS，通过 CONP 实现 CMNS。

CLNS 由以下三个协议构成：

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

表3-21 OSI 与 IP 相对应的概念

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

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

IS-IS 的地址结构

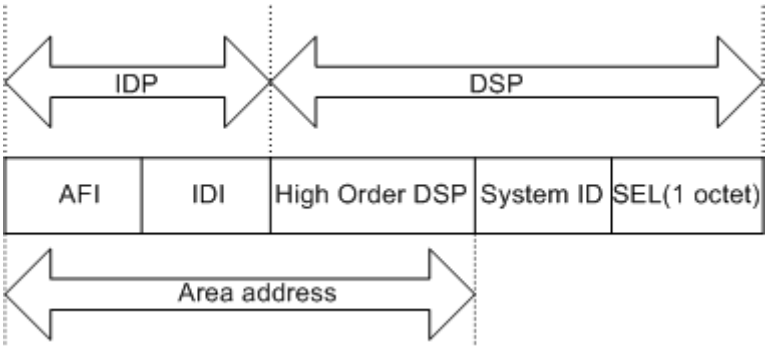
NSAP 是 OSI 协议中用于定位资源的地址。ISO 采用如图 3-17 所示的地址结构，即 NSAP，它由 IDP（Initial Domain Part）和 DSP（Domain Specific Part）组成。IDP 相当于 IP 地址中的主网络号，DSP 相当于 IP 地址中的子网号和主机地址。

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

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

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

图3-17 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。同一区域中路由设备的 Level-1 区域地址必须相同。

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

- System ID

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

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

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；
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；
3. 重新组合的 1680.1000.1001 就是 System ID。

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

- SEL

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

- NET

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

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

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

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

说明
位于同一区域内的路由器的区域地址必须相同。

IS-IS PDU 格式

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

表3-22 PDU 类型对应关系表

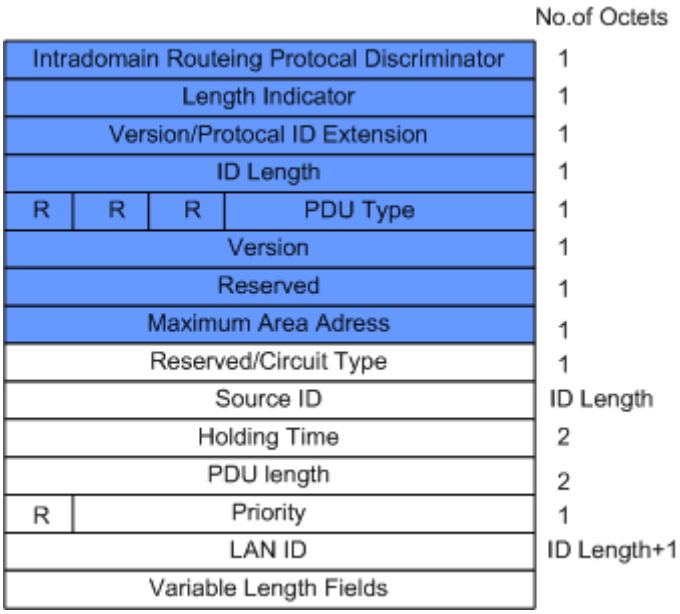
类型值	PDU 类型	简称
15	Level-1 LAN IS-IS Hello PDU	L1 LAN IIH
16	Level-2 LAN IS-IS Hello PDU	L2 LAN IIH
17	Point-to-Point IS-IS Hello PDU	P2P IIH
18	Level-1 Link State PDU	L1 LSP
20	Level-2 Link State PDU	L2 LSP
24	Level-1 Complete Sequence Numbers PDU	L1 CSNP
25	Level-2 Complete Sequence Numbers PDU	L2 CSNP
26	Level-1 Partial Sequence Numbers PDU	L1 PSNP
27	Level-2 Partial Sequence Numbers PDU	L2 PSNP

- Hello 报文格式

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

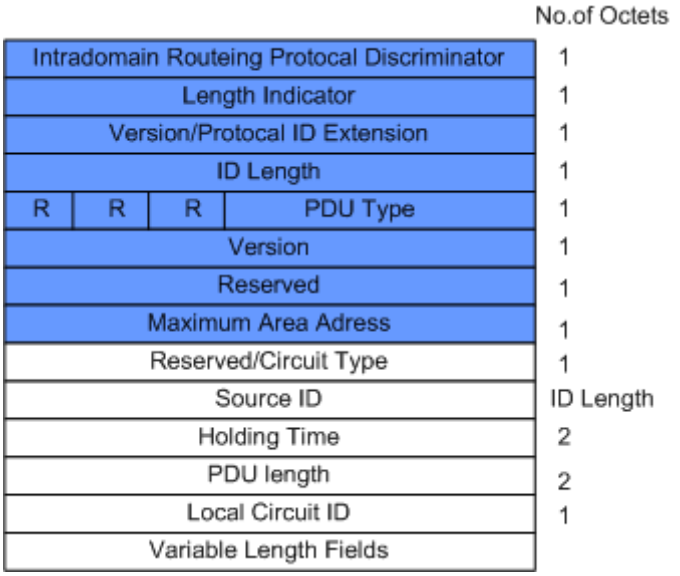
  广播网中的 Hello 报文格式如图 3-18 所示（蓝色部分是通用报文头）。

图3-18 Level-1/Level-2 LAN IIH 格式



P2P 网络中的 Hello 报文格式如图 3-19 所示。

图3-19 P2P IIH 格式



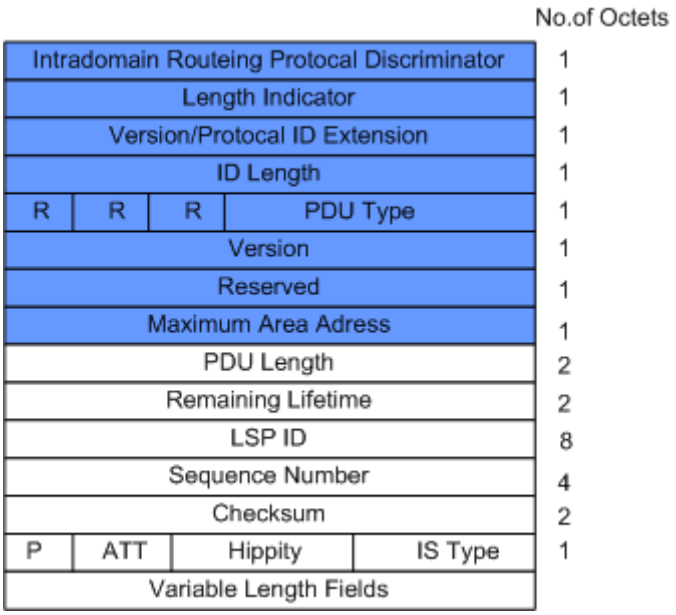
从图中可以看出，P2P IIH 中的多数字段与 LAN IIH 相同。不同的是没有 Priority 和 LAN ID 字段，而多了一个 Local Circuit ID 字段，表示本地链路 ID。

- LSP 报文格式

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

两类 LSP 有相同的报文格式，如图 3-20 所示。

图3-20 Level-1/Level-2 LSP 格式



SNP 格式

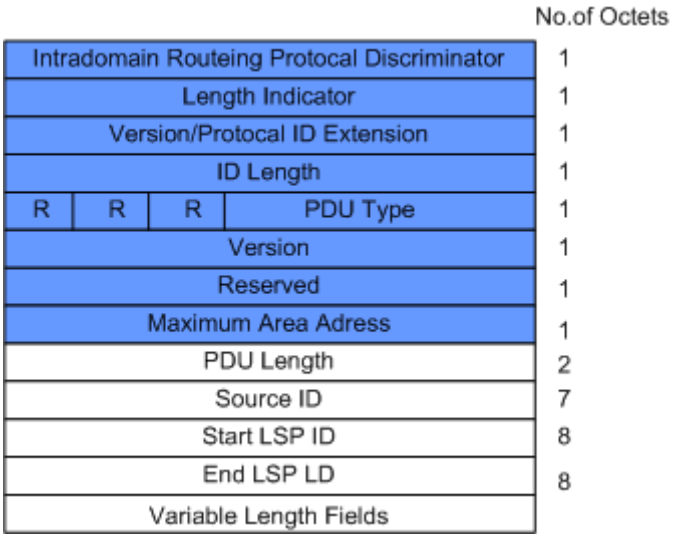
SNP（Sequence Number PDUs）通过描述全部或部分数据库中的 LSP 来同步各 LSDB（Link-State DataBase），从而维护 LSDB 的完整与同步。

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

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

CSNP 的报文格式如图 3-21 所示。

图3-21 Level-1/Level-2 CSNP 格式



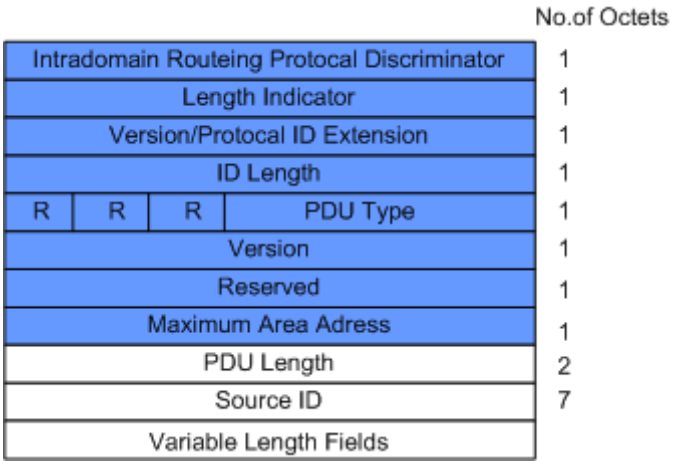
主要字段的解释如下：

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

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

PSNP 的报文格式如图 3-22 所示。

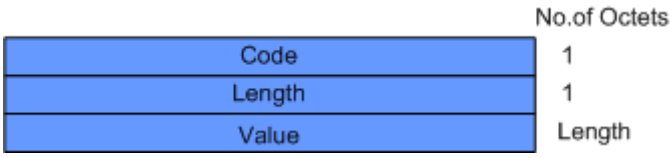
图3-22 Level-1/Level-2 PSNP 格式



CLV

PDU 中的变长字段部分是多个 CLV（Code-Length-Value）三元组。其格式如图 3-23 所示。CLV 也称为 TLV（Type-Length-Value）。

图3-23 CLV 格式



不同 PDU 类型所包含的 CLV 是不同的。如表 3-23 所示。

表3-23 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。

IS-IS 区域

- 两级结构

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

- Level-1 路由器

  Level-1 路由器负责区域内的路由，它只与属于同一区域的 Level-1 和 Level-1-2 路由器形成邻居关系，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 Level-2 或其它区域的 Level-1-2 路由器形成邻居关系，维护一个 Level-2 的 LSDB，该 LSDB 包含区域间的路由信息。

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

- Level-1-2 路由器

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

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

  说明

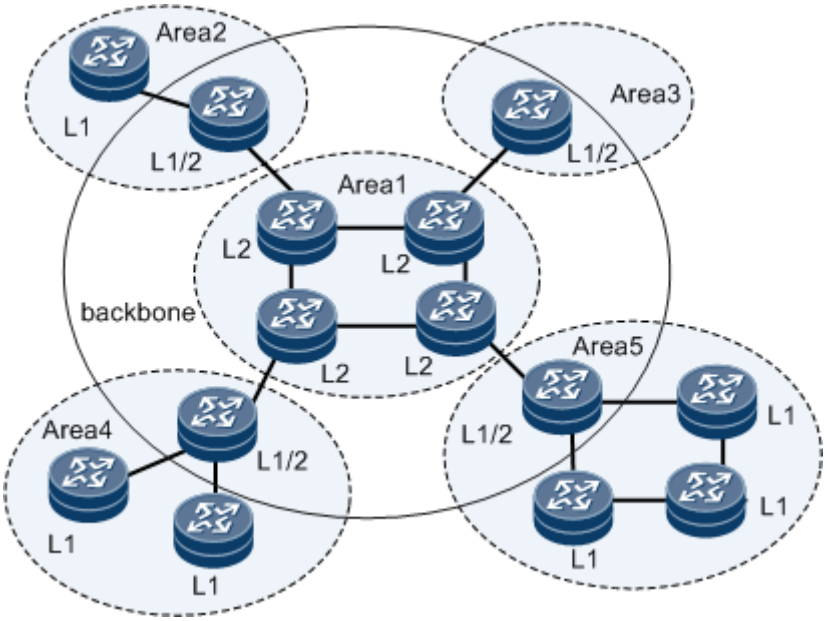
  属于不同区域的 Level-1 路由器不能形成邻居关系。Level-2 路由器之间可以直接形成邻居，与所在区域无关。

- 接口的级别

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

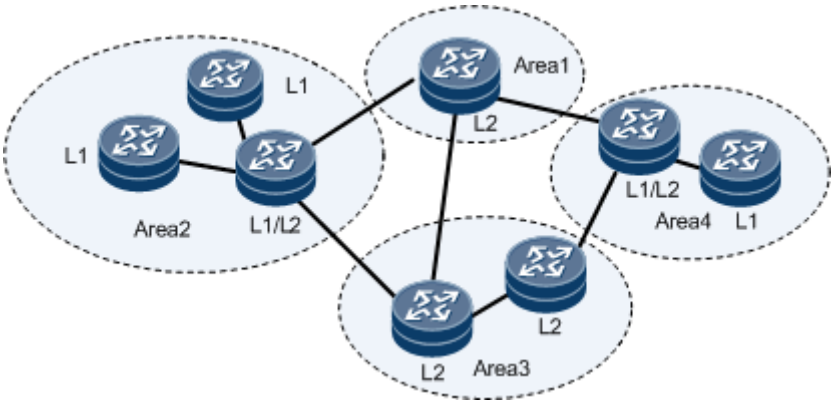
  如图 3-24 所示为一个运行 IS-IS 协议的网络，它与 OSPF 的多区域网络拓扑结构非常相似。整个 backbone 区域不仅包括 Area1 中的所有路由器，还包括其它区域的 Level-1-2 路由器。

图3-24 IS-IS 拓扑结构图一



如图 3-25 所示是 IS-IS 的另外一种拓扑结构图。所有连续的 Level-1-2 和 Level-2 路由器构成了 IS-IS 的骨干区域。在这个拓扑中，Level-2 / Level-1-2 级别的路由器分别属于不同的区域，并没有规定哪个区域是骨干区域。

图3-25 IS-IS 拓扑结构图二



说明

IS-IS 的骨干网（Backbone）指的不是一个特定的区域。

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

IS-IS 的网络类型

IS-IS 只支持两种类型的网络，根据物理链路不同可分为：

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

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

DIS 和伪节点

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

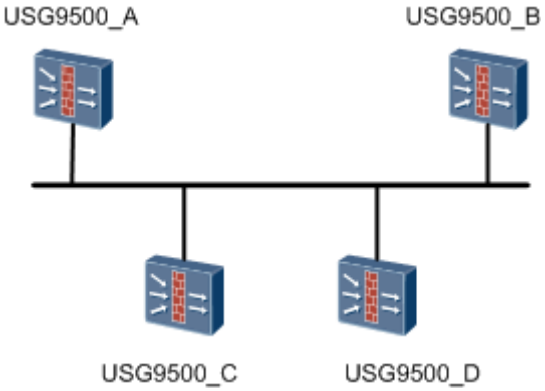
Level-1 和 Level-2 的 DIS 是分别选举的，用户可以为不同级别的 DIS 选举设置不同的优先级。DIS 优先级数值最大的被选为 DIS。如果优先级数值最大的路由器有多台，则其中 MAC 地址最大的路由器会被选中。不同级别的 DIS 可以是同一台路由器，也可以是不同的路由器。

与 OSPF 的不同点：

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

在 IS-IS 广播网中，同一网段上的同一级别的路由器之间都会形成邻接关系，包括所有的非 DIS 路由器之间也会形成邻接关系，这一点与 OSPF 是不同的。如图 3-26 所示。

图3-26 IS-IS 广播网的 DIS 和邻接关系



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

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

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

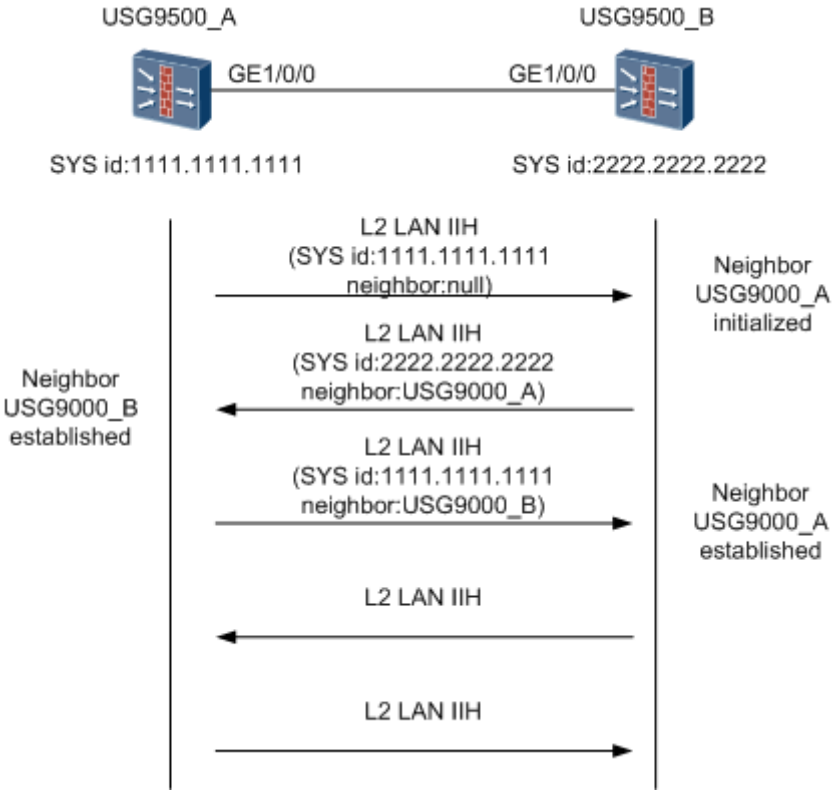
说明
IS-IS 广播网上所有的路由器之间都形成邻接关系，但 LSDB 的同步仍然依靠 DIS 来保证。

IS-IS 邻居关系的建立

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

- 广播链路邻居关系的建立

图3-27 广播链路组网图



如图 3-27 所示。USG9500_A 广播发送 Level-2 LAN IS-IS Hello PDU，USG9500_B 收到此报文后，将自己和 USG9500_A 的邻居状态标识为 Initial；然后，USG9500_B 再回复 L2 LAN IIH 报文，USG9500_A 收到这个带有 USG9500_A 为 USG9500_B 邻居信息的 IIH 报文后，USG9500_A 再将自己与 USG9500_B 的邻居状态标识为 Up。

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

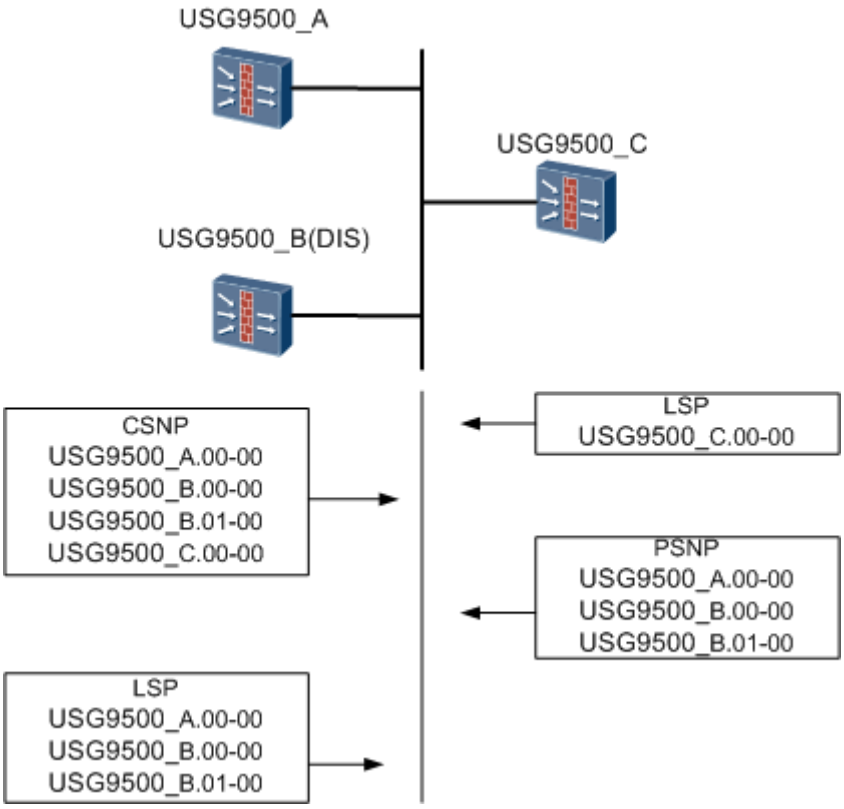
- P2P 链路邻居关系的建立

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

  - 2-way 方式

    即只要设备收到 IS-IS Hello 报文，就会单方向建立起邻居关系。如图 3-28 所示。

图3-28 P2P 链路邻居关系建立过程



- 3-way 方式

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

说明

对 IS-IS 三次握手机制特性有专门章节进行更详细的讨论。

IS-IS 按如下原则建立邻居关系：

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

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

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

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

- 对于 P2P 接口，可以配置接口忽略 IP 地址检查。

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

IS-IS 的 LSP 交互过程

- LSP 的“泛洪”（flooding）

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

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

- LSP 产生的原因

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

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

- 收到邻居新的 LSP 的处理过程
  - 将新的 LSP 安装到自己的 LSDB 数据库中标记为 flooding。
  - 发送新的 LSP 到除了收到该 LSP 的接口之外的接口。
  - 邻居再扩散到其他邻居。

- 新加入路由器与 DIS 同步 LSDB 数据库过程
  - 新加入的路由设备 USG9500_C 首先发送 Hello 报文，与该广播域中的路由设备建立邻居关系。（请参见“广播链路邻居关系的建立”）
  - 邻居关系建立起来后，USG9500_C 等待 LSP 定时器超时，然后将自己的 LSP 发送往组播地址：

    Level-1：01-80-C2-00-00-14

    Level-2：01-80-C2-00-00-15

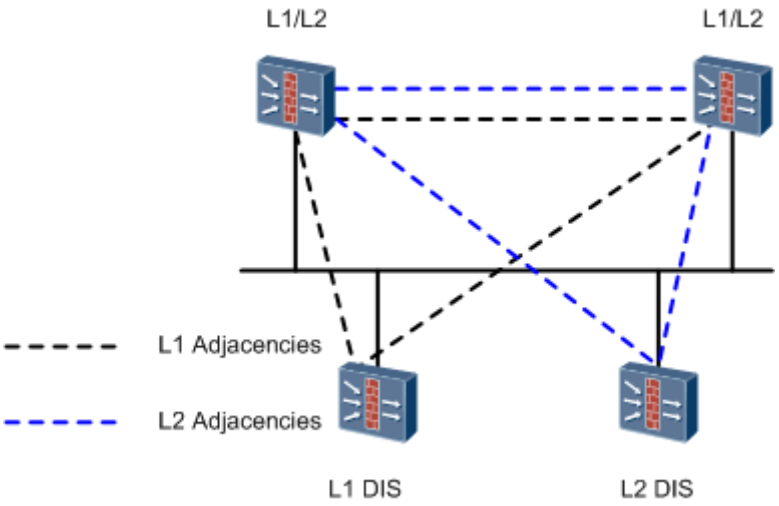
    网络上所有的邻居都将收到该 LSP。
  - 该网段中的 DIS 会把收到 USG9500_C 的 LSP 加入到 LSDB 中，并等待 CSNP 报文定时器超时并发送 CSNP 报文，进行该网络内的 LSDB 同步。CSNP 报文的发送间隔缺省值为 10 秒。
  - USG9500_C 收到 DIS 发来的 CSNP 报文，对比自己的 LSDB 数据库，发送 PSNP 报文请求自己没有的 LSP。
  - DIS 收到该 PSNP 报文请求后发送对应的 LSP 进行 LSDB 的同步。

- DIS 的 LSDB 更新过程
  - DIS 接收到 LSP，在数据库中搜索对应的记录。若没有该 LSP，则将其加入数据库，并广播新数据库内容。

- 若收到的 LSP 序列号大于本地 LSP 的序列号，就替换为新报文，并广播新数据库内容。
- 若收到的 LSP 序列号小本地 LSP 的序列号，就向入端接口发送本地 LSP 报文。
- 若两个序列号相等，则比较 Remaining Lifetime。若收到的 LSP 的 Remaining Lifetime 小于本地 LSP 的 Remaining Lifetime，就替换为新报文，并广播新数据库内容。
- 若两个序列号相等，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。

- P2P 链路上数据库的同步过程

图3-29 点到点链路数据库更新过程



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

说明
在 P2P 链路上 PSNP 有两种作用：

- 作为 Ack 应答以确认收到的 LSP。
- 用来请求所需的 LSP。

- 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。

3.5.3.2 IS-IS 多实例和多进程

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

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

  为了方便管理，提高控制效率，IS-IS 支持多进程和多实例特性。

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

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

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

配置 IS-IS 多实例和多进程时，有以下注意事项：

- 创建 IS-IS 多实例时，必须在创建 IS-IS 进程时绑定 VPN。如果在创建时没有进行绑定，后面无法通过配置将一个已存在的进程绑定到一个 VPN 上。
- 对一个已绑定了 VPN 的 IS-IS 进程，无法通过配置绑定到另一个 VPN 上。
- 多个 IS-IS 进程可以绑定到同一个 VPN 上，这就是 IS-IS 多进程。
- 需要使能 IS-IS 多实例的接口必须和 IS-IS 绑定相同的 VPN。
- 绑定了 VPN 的 IS-IS 进程从属于 VPN，所以当 VPN 删除时，IS-IS 进程也跟着删除。
- 不同 VPN 的路由不能相互引入。

3.5.3.3 IS-IS 路由渗透

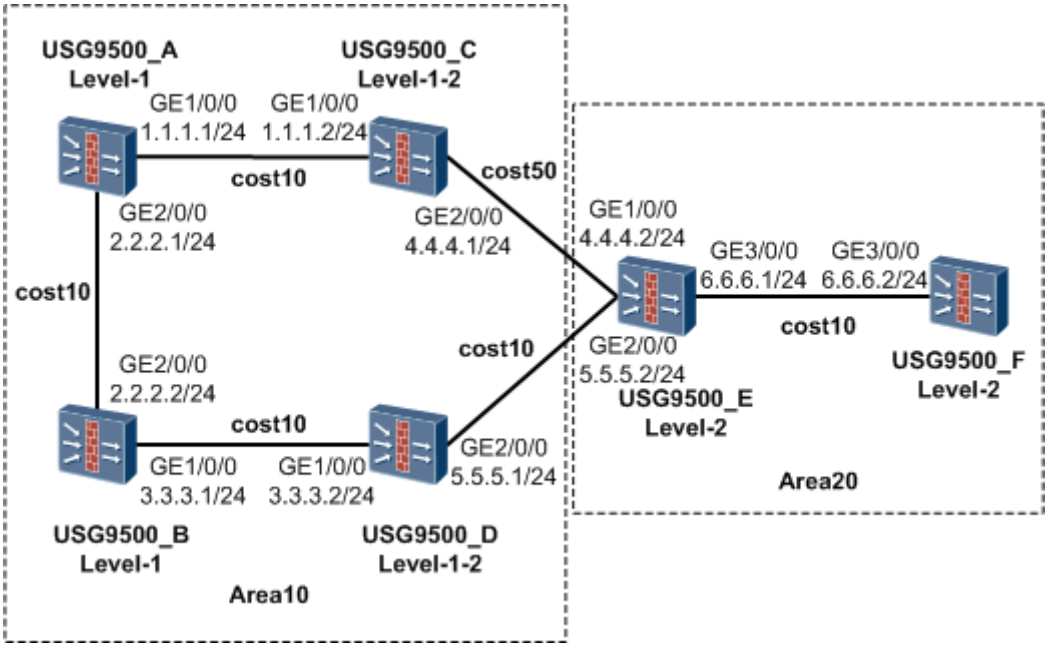
路由渗透特性是指 Level-1-2 IS-IS 将已知的其他 Level-1 区域以及 Level-2 区域的路由信息通报给指定的 Level-1 区域。

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

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

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

图3-30 路由渗透示例



- USG9500_A、USG9500_B、USG9500_C 和 USG9500_D 同属于 Area10 区域，USG9500_A 和 USG9500_B 为 Level-1 路由器，USG9500_C 和 USG9500_D 为 Level-1-2 路由器。
- USG9500_E、USG9500_F 同属于 Area20 区域，为 Level-2 路由器。

USG9500_A 发送报文给 USG9500_F，选择的最佳路径应该是 USG9500_A->USG9500_B->USG9500_D->USG9500_E->USG9500_F。因为这条链路上的 cost 值为 10+10+10+10=40，但在 USG9500_A 上查看发送到 USG9500_F 的报文选择的路径是：USG9500_A->USG9500_C->USG9500_E->USG9500_F，其 cost 值为 10+50+10=70，不是 USG9500_A 到 USG9500_F 的最优路由。

这是由于 USG9500_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 Level-1-2 路由器 USG9500_C 和 USG9500_D 上使能路由渗透。再查看报文选择的路径，发现路径是 USG9500_A->USG9500_B->USG9500_D->USG9500_E->USG9500_F，为 USG9500_A 到 USG9500_F 的最优路由。

3.5.3.4 IS-IS GR

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

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

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

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

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

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

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

IS-IS GR 基本概念

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

- GR-Restarter

  具备 GR 能力、并将要进行 GR 的路由器。
- GR-Helper

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

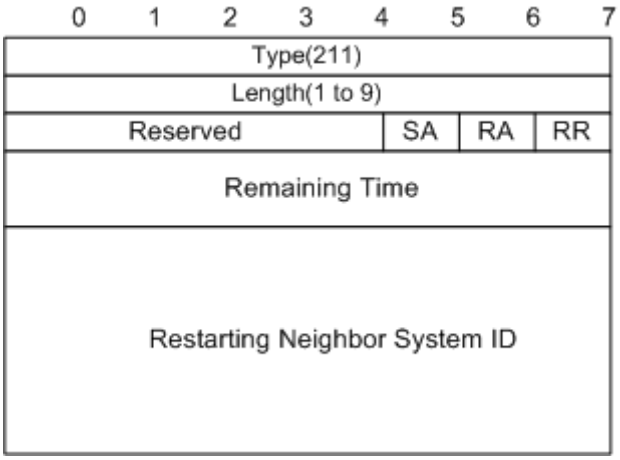
说明
设备默认支持 GR-Helper。

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

Restart TLV

Restart TLV 是包含在 IIH（IS-to-IS Hello PDUs）报文中的扩展部分。支持 IS-IS GR 能力的路由器的所有 IIH 报文都包含 Restart TLV。Restart TLV 中携带了协议重启的一些参数。其报文格式如图 3-31 所示。

图3-31 Restart TLV 格式



Restart TLV 各字段的含义如表 3-24 所示。

表3-24 Restart TLV 报文字段含义

字段名	长度	含义
Type	1 字节	TLV 的类型。值为 211 表示是 Restart TLV。
Length	1 字节	TLV 值的长度。
RR	1 比特	重启请求位（Restart Request）。路由器发送的 RR 置位的 Hello 报文用于通告邻居自己发生 Restarting/Starting，请求邻居保留当前的 IS-IS 邻接关系并返回 CSNP 报文。
RA	1 比特	重启应答位（Restart Acknowledgement）。路由器发送的 RA 置位的 Hello 报文用于通告邻居确认收到了 RR 置位的报文。
SA	1 比特	抑制发布邻接关系位（Suppress adjacency advertisement）。用于发生 Starting 的设备请求邻居抑制与自己相关的邻居关系的广播，以避免路由黑洞。
Remaining Time	2 字节	邻居保持邻接关系不重置的时间。长度是 2 字节，单位是秒。当 RA 置位时，这个值是必需的。
Restarting Neighbor System ID	6 字节	回应重启应答报文的邻居路由器的 System ID。

定时器

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

- T1

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

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

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

- T2

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

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

- T3

  整个系统维护一个 T3 定时器。

  T3 定时器可理解为成功完成 GR 所允许的最大时间。

  T3 定时器超时表示 GR 失败。

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

  T3 定时器只用于 Restarting 设备。

IS-IS GR 的会话机制

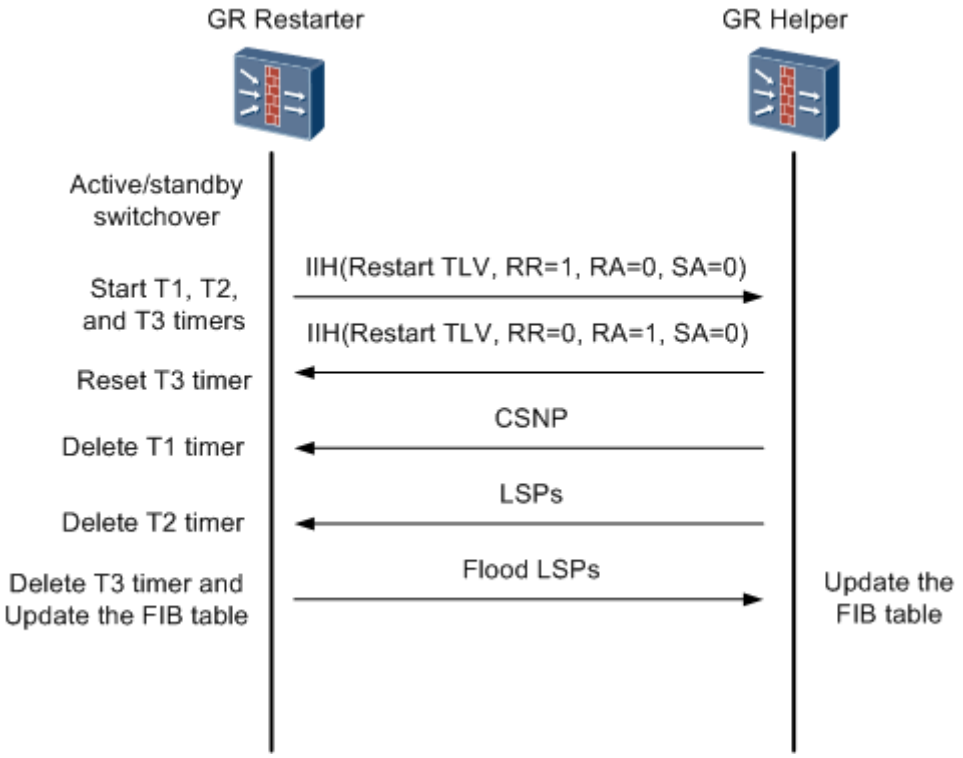
为了以示区别，主备倒换和重启 IS-IS 进程触发的 GR 过程称为 Restarting，FIB 表保持不变。路由器重启触发的 GR 过程称为 Starting，进行 FIB 表更新。

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

IS-IS Restarting

IS-IS Restarting 的过程如图 3-32 所示。

图3-32 IS-IS Restarting 过程



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

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

如果 GR Helper 不支持 GR，就忽略 Restart TLV，按正常的 IS-IS 过程处理，重置和 GR Restarter 的邻接关系。

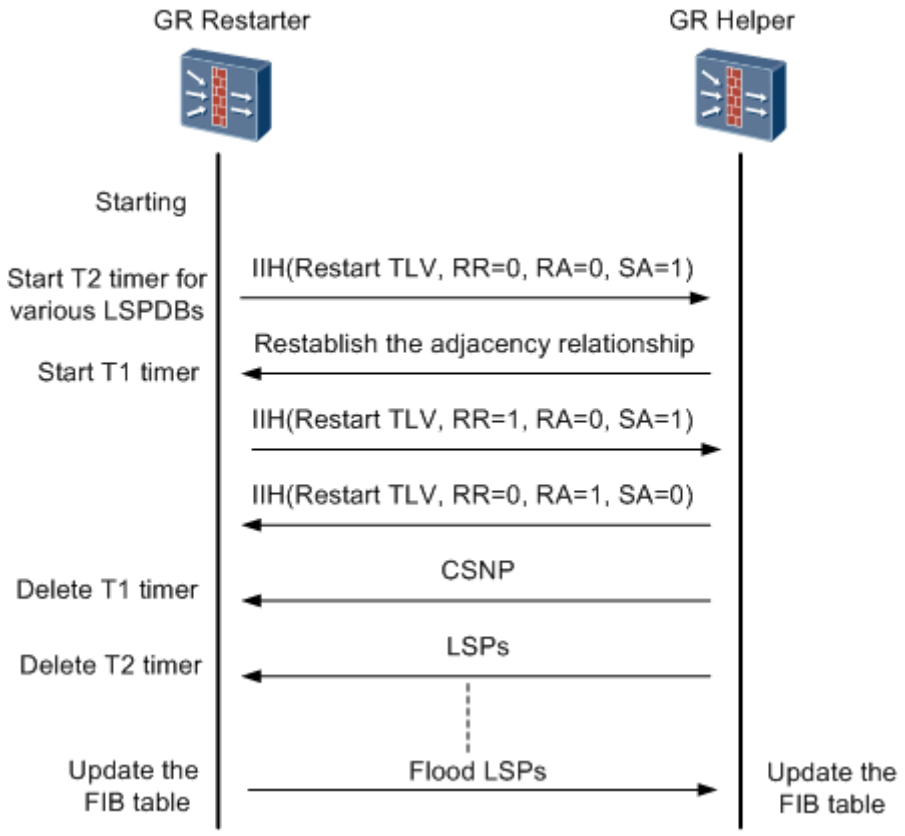
3. GR Restarter 接收到邻居的 IIH 回应报文（RR 清除、RA 置位），做如下处理：
   - 把 T3 的当前值和报文中 Remaining time 比较，取其中较小者作为 T3 的值。
   - 在接口收到确认报文和 CSNP 报文之后，取消该接口的 T1 定时器。
   - 如果该接口没有收到确认报文和 CSNP 报文，T1 会不停地重置，重发含 Restart TLV 的 IIH 报文。如果 T1 超时次数超过阈值，GR Restarter 强制取消 T1 定时器，启动正常的 IS-IS 处理流程。

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

IS-IS Starting

对于 Starting 设备，因为没有保留 FIB 表项，所以一方面希望在 Starting 之前和自己的邻接关系为“Up”的邻居重置和自己的邻接关系，同时希望邻居能在一段时间内抑制和自己的邻接关系的发布。其处理过程和 Restarting 不同，具体如图 3-33 所示。

图3-33 IS-IS Starting 过程



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

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

  RR 位清除表示是 Starting 完成。

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

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

     重新初始化邻接关系。

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

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

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

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

3.5.3.5 IS-IS for IPv6

IETF 的 RFC5308 中规定了 IS-IS 为支持 IPv6 所新增的内容。支持 IPv6 路由的处理和计算。主要是新添加的支持 IPv6 路由信息的两个 TLVs（Type-Length-Values）和一个新的 NLPID（Network Layer Protocol Identifier）。

新增的两个 TLV 分别是：

- IPv6 Reachability

  类型值为 236（0xEC），通过定义路由信息前缀、度量值等信息来说明网络的可达性。
- IPv6 Interface Address

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

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

3.5.3.6 BFD for IS-IS

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

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

BFD 包括静态 BFD 和动态 BFD。

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

- 静态 BFD

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

- 动态 BFD（包括 BFD for IPv4 和 BFD for IPv6）

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

  BFD for IPv4 的会话和 BFD for IPv6 的会话分开建立，互不影响。

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

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

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

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

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

静态 BFD

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

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

动态 BFD

动态建立 BFD 会话指的是由路由协议动态触发 BFD 会话建立。BFD for IPv4 会话是 IS-IS 在 IPv4 邻居建立的时候触发建立的；BFD for IPv6 会话是 IS-IS 在 IPv6 邻居建立的时候触发建立的。

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

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

BFD 会话的建立与删除

- 创建 BFD 会话的条件
  - 各路由器配置了 IS-IS 基本功能并且在接口下使能了 IS-IS。

  说明
  对于 IPv6 网络，还需要配置 IS-IS IPv6 基本特性。
  - 各路由器配置了全局 BFD 功能并且使能了接口或者进程的 BFD for IPv4 或者 BFD for IPv6 特性。
  - 使能了接口或者进程的 BFD for IPv4 或者 BFD for IPv6 特性，且相邻路由器的邻居状态为 Up（广播网中须等到 DIS 选举出来）。
  - 邻居的 IP 协议类型包含 IPv4 和 IPv6
- 创建 BFD 会话的过程
  - P2P 网络

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

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

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

  如果同一链路上的同一对路由器形成的是 Level-1-2 的类型的邻居，在广播网中 IS-IS 会针对这两个 Level 分别创建两个 BFD 会话，但在 P2P 网络中 IS-IS 只会创建一个 BFD 会话。
- 如果邻居的 IP 协议类型包含 IPv4 和 IPv6，那么 IS-IS 会分别创建两个会话，一个 BFD for IPv4 会话和一个 BFD for IPv6 会话。其中创建 BFD for IPv6 会话时，将采用对应接口的 IPv6 link-local 地址。
- 删除 BFD 会话的条件
  - P2P 网络

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

- 广播网络

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

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

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

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

- 响应 BFD 会话 Down 事件

  当 BFD 检测到链路发生故障并产生 Down 事件时，会通知 RM。RM 通知 IS-IS 删除此邻接。IS-IS 响应这个事件并重新进行路由计算，实现网络迅速收敛。BFD for IPv4 通知 IS-IS 链路故障后，IS-IS 只改变其 IPv4 路由；BFD for IPv6 通知 IS-IS 链路故障后，IS-IS 只改变其 IPv6 路由。

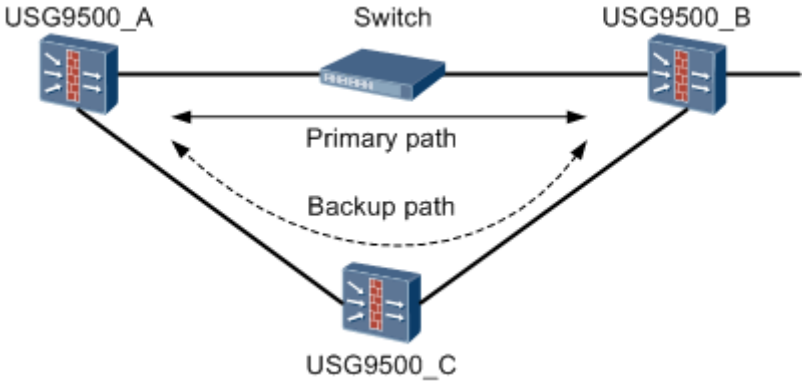
  当本地路由器与邻居路由器均为 Level-1-2 时，二者之间会针对不同的 Level 分别创建两个邻居，此时 IS-IS 也会创建两个不同 Level 的会话，在这种情况下，RM 会删除根据相应 Level 的邻接关系。

组网应用

说明
请根据网络环境配置 BFD，如果时间参数设置不当将会导致网络震荡。

BFD for IS-IS 可以快速感知链路变化实现路由收敛。

图3-34 IS-IS BFD 组网示意图



配置要求：

- 如所示在各路由器上使能 IS-IS 基本功能。

说明
对于 IS-IS BFD for IPv6，还需配置 IS-IS 的 IPv6 特性。

- 使能全局 BFD 特性。
- 在 USG9500_A 和 USG9500_B 上使能 IS-IS BFD 检测机制。

这样，当 USG9500_A 和 USG9500_B 之间的链路故障时，BFD 能够快速检测到故障并通告给 IS-IS 协议，IS-IS Down 掉此接口的邻居并删除邻接对应的 IP 协议类型，从而触发拓扑计算，同时更新 LSP 使得其他邻居（如 USG9500_B 的邻居 USG9500_C）及时收到 USG9500_B 的更新 LSP，实现了网络拓扑的快速收敛。

3.5.3.7 IS-IS 认证

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

根据报文的种类，认证可以分为以下三类：

- 区域认证

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

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

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

根据报文的认证方式，可以分为以下两类：

- 明文认证

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

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

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

- Type

  ISO 定义认证报文的类型值为 10，1 字节。
- Length

  认证 TLV 值的长度，1 字节。
- Value

  认证的具体内容，其中包括了认证的类型和认证的密码，1～254 字节。

  在 Value 中，认证的类型为 1 字节，具体定义如下：
  - 0：保留的类型
  - 1：明文认证
  - 54：MD5 认证
  - 255：路由域私有认证方式

认证密码的保存情况如下：

- 对于 IIH 报文，使用的认证密码保存在接口下，即前面提到的接口认证。

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

对于接口认证，有以下两种设置：

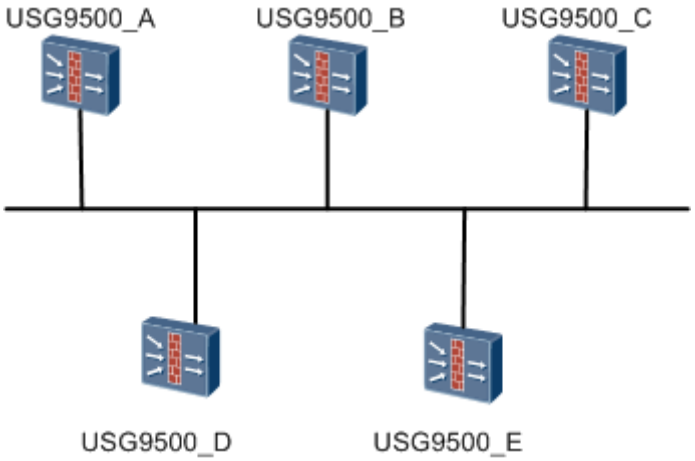
- 发送带认证 TLV 的认证报文，本地对收到的报文也进行认证检查。
- 发送带认证 TLV 的认证报文，但是本地对收到的报文不进行认证检查。

对于区域和路由域认证，可以设置为 SNP 和 LSP 分开认证。

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

组网应用

图3-35 广播网中的 IS-IS 认证



配置要求：

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

3.6 BGP

3.6.1 介绍

定义

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

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

当前使用的版本是 BGP-4（RFC4271）。

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

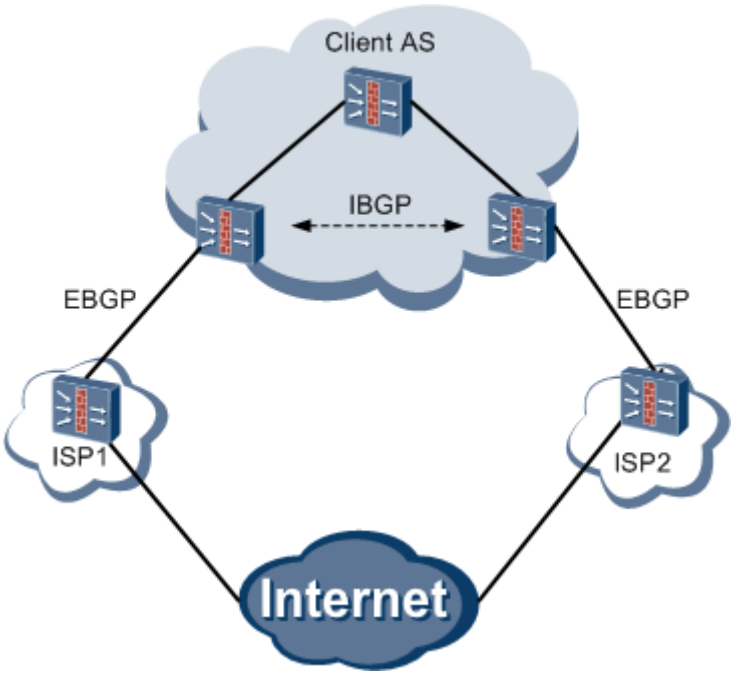
BGP 协议具有如下特点：

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

目的

BGP 用于在 AS 之间传递路由信息，并不是所有情况都需要运行 BGP。

图3-36 BGP 的应用场景



以下情况中需要使用 BGP 协议：

- 如图 3-36，用户需要同时与两个或者多个 ISP 相连，ISP 需要向用户提供部分或完全的 Internet 路由。这时可以通过 BGP 路由携带的 AS 信息来决定到达目的地，走哪一个 ISP 的 AS 更为经济。
- 不同组织下的用户之间需要传递 AS 路径信息。
- 用户需要通过三层 VPN 传播私网路由。

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。

3.6.2 参考标准和协议

本特性的参考资料清单如下所示：

表3-25 参考标准和协议

文档	描述	备注
RFC4271	A Border Gateway Protocol 4 (BGP-4)	
RFC4760	Multiprotocol Extensions for BGP-4	
RFC3392	Capabilities Advertisement	

文档	描述	备注
	with BGP-4	
RFC2918	Route Refresh Capability for BGP-4	
RFC2439	BGP Route Flap Damping	
RFC1997	BGP Communities Attribute	
RFC4456	BGP Route Reflection	
RFC3065	Autonomous System Confederations for BGP	
RFC3232	Assigned Numbers: RFC 1700 is Replaced by an On-line Database	
RFC827	Exterior Gateway Protocol (EGP)	
RFC3682	The Generalized TTL Security Mechanism (GTSM)	
RFC4724	Graceful Restart Mechanism for BGP	
RFC4486	Subcodes for BGP Cease Notification Message	

3.6.3 原理描述

3.6.3.1 BGP 基本原理

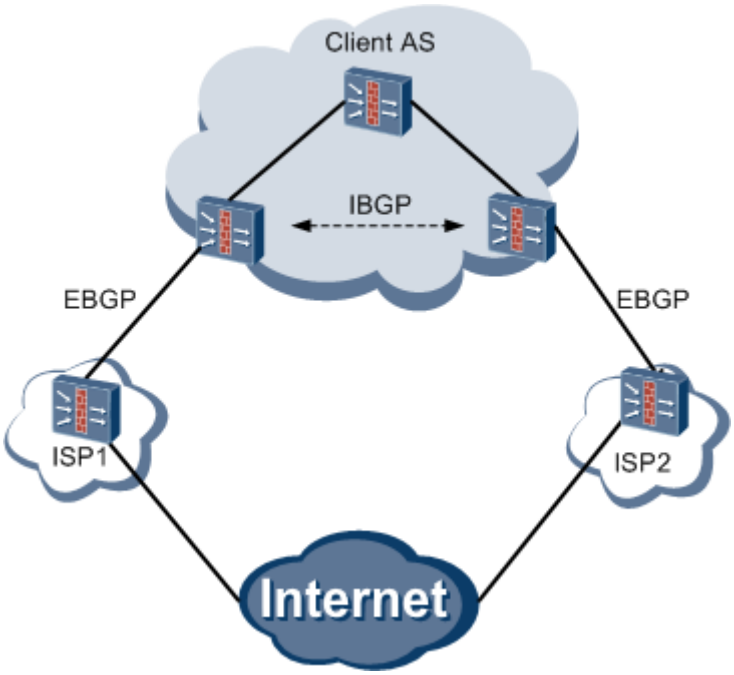
BGP 运行方式

BGP 在路由设备上以下列两种方式运行，如图 3-37 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3-37 BGP 的运行方式



BGP 消息中的角色

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

BGP 的消息

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

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

  - 一条 Update 消息可以撤销多条不可达路由。每一个路由通过目的地（用 IP 前缀表示），清楚的定义了 BGP Speaker 之间先前通告过的路由。
  - 一条 Update 消息可以只用于撤销路由，这样就不需要包括路径属性或者 NLRI。相反，也可以只用于通告可达路由，就不需要携带撤销路由信息了。
- Notification 消息：当 BGP 检测到错误状态时，就向对等体发出 Notification 消息，之后 BGP 连接会立即中断。
- Keepalive 消息：BGP 会周期性的向对等体发出 Keepalive 消息，用来保持连接的有效性。
- Route-Refresh 消息：Route-Refresh 消息用来通知对等体自己支持路由刷新能力（Route-Refresh capability）。

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

BGP 有限状态机

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

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

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

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

BGP 处理过程

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

BGP 属性

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

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

下面介绍几种常用的 BGP 路由属性：

- Origin 属性

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

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

- AS_Path 属性

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

  当 BGP Speaker 本地通告一条路由时：

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

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

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

- Next_Hop 属性

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

  - BGP Speaker 在向 EBGP 对等体发布某条路由时，会把该路由信息的下一跳属性设置为本地与对端建立 BGP 邻居关系的接口地址。

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

- MED

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

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

- Local_Pref 属性

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

  Local_Pref 属性用于判断流量离开 AS 时的最佳路由。当 BGP 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. 优选协议首选值（PrefVal）最高的路由；
2. 优选本地优先级（Local_Pref）最高的路由；
3. 优选聚合路由（聚合路由优先级高于非聚合路由）；
4. 本地手动聚合路由的优先级高于本地自动聚合的路由；
5. 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；
6. 优选 AS 路径（AS_Path）最短的路由；
7. 比较 Origin 属性，依次选择 Origin 类型为 IGP、EGP、Incomplete 的路由；
8. 优选 MED 值最低的路由；
9. 优选从 EBGP 学来的路由（EBGP 路由优先级高于 IBGP 路由）；
10. 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；
11. 优选 Cluster_List 最短的路由；
12. 优选 Originator_ID 最小的路由；
13. 优选 Router ID 最小的路由器发布的路由；
14. 比较对等体的 IP Address，优选从具有较小 IP Address 的对等体学来的路由。

BGP 等价负载分担

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

形成 BGP 等价负载分担的条件是：“BGP 选择路由的策略”的 1 至 10 条规则中需要比较的属性完全相同。

BGP 发布路由的策略

BGP 发布路由时采用如下策略：

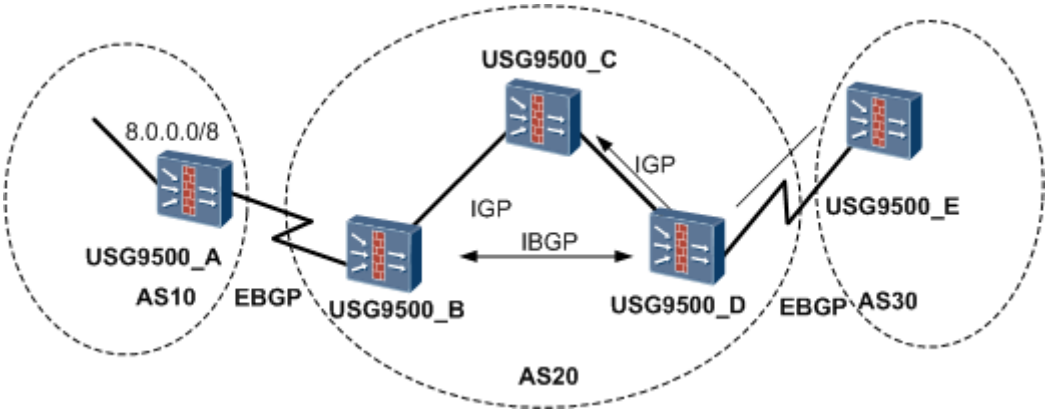
- 存在多条有效路由时，BGP Speaker 只将最优路由发布给对等体。
- BGP Speaker 从 EBGP 获得的路由会向它所有 BGP 对等体发布（包括 EBGP 对等体和 IBGP 对等体）。
- BGP Speaker 从 IBGP 获得的路由不向它的 IBGP 对等体发布。
- BGP Speaker 从 IBGP 获得的路由发布给它的 EBGP 对等体。
- 连接一旦建立，BGP Speaker 将把自己所有 BGP 路由发布给新对等体。

IBGP 和 IGP 同步

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

如果一个 AS 中有非 BGP 路由器提供转发服务，经该 AS 转发的 IP 报文将可能因为目的地址不可达而被丢弃。如图 3-38 所示，USG9500_E 通过 BGP 从 USG9500_D 可以学到 USG9500_A 的一条路由 8.0.0.0/8，于是将到这个目的地址的报文转发给 USG9500_D，USG9500_D 查询路由表，发现下一跳是 USG9500_B。由于 USG9500_D 从 IGP 学到了到 USG9500_B 的路由，所以通过路由迭代，USG9500_D 将报文转发给 USG9500_C。但 USG9500_C 并不知道去 8.0.0.0/8 的路由，于是将报文丢弃。

图3-38 IBGP 和 IGP 同步



如果设置了同步特性，在 IBGP 路由加入路由表并发布给 EBGP 对等体之前，会先检查 IGP 路由表。只有在 IGP 也知道这条 IBGP 路由时，它才会被加入到路由表，并发布给 EBGP 对等体。

在下面的情况中，可以安全地关闭同步特性。

- 本 AS 不是过渡 AS（图 3-38 中的 AS20 就属于一个过渡 AS）
- 本 AS 内所有路由器建立 IBGP 全连接

说明
缺省情况下，USG9500 的同步功能是关闭的。

3.6.3.2 路由引入

BGP 协议自身不能发现路由，所以需要引入其他协议的路由（如 IGP 或者静态路由等）注入到 BGP 路由表中，从而将这些路由在 AS 之内和 AS 之间传播。

BGP 引入路由时支持 Import 和 Network 两种方式：

- Import 方式是按协议类型，将 RIP 路由、OSPF 路由、ISIS 路由、静态路由和直连路由等某一协议的路由注入到 BGP 路由表中。
- Network 方式比 Import 方式更精确，将指定前缀和掩码的一条路由注入到 BGP 路由表中。

3.6.3.3 路由聚合

在大规模的网络中，BGP 路由表十分庞大，使用路由聚合（Routes Aggregation）可以大大减小路由表的规模。

路由聚合实际上是将多条路由合并的过程。这样 BGP 在向对等体通告路由时，可以只通告聚合后的路由，而不是通告所有的具体路由。

BGP 路由聚合支持两种方式：

- 自动聚合：对 BGP 引入的路由进行聚合。配置自动聚合后，对参加聚合的具体路由进行抑制，BGP 将按照自然网段聚合路由（如 10.1.1.1/24 和 10.2.1.1/24 将聚合为 A 类地址 10.0.0.0/8），并且 BGP 向对等体只发送聚合后的路由。
- 手动聚合：对 BGP 本地路由进行聚合。手动聚合可以控制聚合路由的属性，以及决定是否发布具体路由。

IPv4 支持自动聚合和手动聚合两种方式，而 IPv6 仅支持手动聚合。

3.6.3.4 路由衰减

路由不稳定的主要表现形式是路由振荡（Route Flapping），即路由表中的某条路由反复消失和重现。

说明
路由表中添加一条路由后，该路由又被撤销，这个过程称为一次路由震荡。

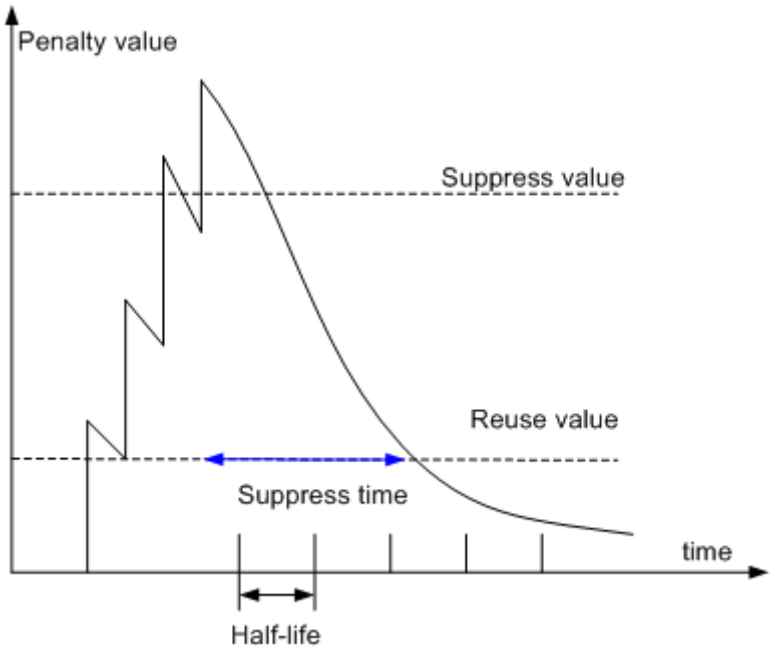
发生路由振荡时，设备就会向邻居发布路由更新，收到更新报文的设备需要重新计算路由并修改路由表。所以频繁的路由振荡会消耗大量的带宽资源和 CPU 资源，严重时会影响到网络的正常工作。

路由衰减（Route Dampening）用来解决路由不稳定的问题。多数情况下，BGP 协议都应用于复杂的网络环境中，路由变化十分频繁。为了防止持续的路由振荡带来的不利影响，BGP 使用路由衰减来抑制不稳定的路由。

BGP 衰减使用惩罚值（Penalty Value）来衡量一条路由的稳定性，惩罚值越高则说明路由越不稳定。路由每发生一次振荡（路由从激活状态变为未激活状态，称为一次路由振荡），BGP 便会给此路由增加一定的惩罚值（1000）。当惩罚值超过抑制阈值（Suppress Value）时，此路由被抑制，不加入到路由表中，也不再向其他 BGP 对等体发布更新报文。

被抑制的路由每经过一段时间，惩罚值便会减少一半，这个时间称为半衰期（Half-life）。当惩罚值降到再使用阈值（Reuse Value）时，此路由变为可用并被加入到路由表中，同时向其他 BGP 对等体发布更新报文。上文提到的惩罚值、抑制阈值和半衰期都可以手动配置。BGP 衰减的处理过程如图 3-39 所示。

图3-39 BGP 衰减示意图



路由衰减只适用于 EBGP 路由。对于从 IBGP 收来的路由不能进行衰减，因为 IBGP 路由经常含有本 AS 的路由，内部网络路由要求转发表尽可能一致，IGP 快速收敛就是为了达到信息同步，转发一致。如果衰减对 IBGP 路由起作用，不同设备的衰减参数不一致时，会导致转发表不一致。

3.6.3.5 团体属性

团体属性（Community）是一组有相同特征的目的地址的集合。团体属性用一组以 4 字节为单位的列表来表示，设备中团体属性的格式是 aa:nn 或团体号。

- aa:nn：aa 和 nn 的取值范围都是 0～65535，管理员可根据实际情况设置具体数值。通常 aa 表示自治系统 AS 编号，nn 是管理员定义的团体属性标识。例如，来自 AS100 的一条路由，管理员定义的团体属性标识是 1，则该路由的团体属性格式是 100:1。
- 团体号：团体号是 0～4294967295 的整数。RFC1997 中定义，0（0x00000000）～65535（0x0000FFFF）和 4294901760（0xFFFF0000）～4294967295（0xFFFFFFFF）是预留的。

团体属性用来简化路由策略的应用和降低维护管理的难度，利用团体可以使多个 AS 中的一组 BGP 设备共享相同的策略。团体是一个路由属性，在 BGP 对等体之间传播，且不受 AS 的限制。BGP 设备在将带有团体属性的路由发布给其它对等体之前，可以先改变此路由原有的团体属性。

对等体组可以使一组对等体共享相同的策略，而团体可以使一组路由共享相同的策略。

除了使用公认的团体属性外，用户还可以使用团体属性过滤器过滤自定义扩展团体属性，以便更为灵活的控制路由策略。

公认的团体属性

BGP 公认的团体属性见表 3-26。

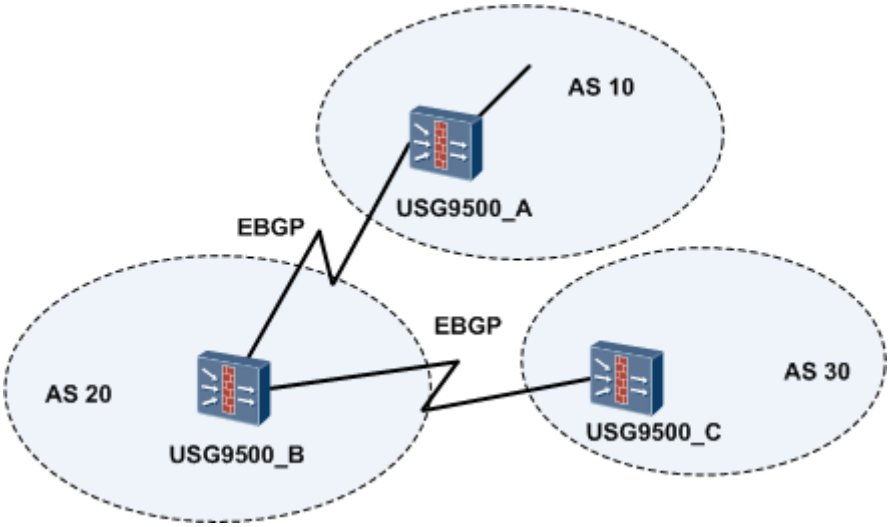
表3-26 BGP 公认团体属性

团体名称	团体标识	说明
Internet	0（0x00000000）	缺省情况下，所有的路由都属于 Internet 团体。具有此属性的路由可以被通告给所有的 BGP 对等体。
No_Export	4294967041（0xFFFFFF01）	具有此属性的路由在收到后，不能被发布到本地 AS 之外。如果使用了联盟，则不能被发布到联盟之外，但可以发布给联盟中的其他子 AS。
No_Advertise	4294967042（0xFFFFFF02）	具有此属性的路由在收到后，不能被通告给任何其他的 BGP 对等体。
No_Export_Subconfed	4294967043（0xFFFFFF03）	具有此属性的路由在收到后，不能被发布到本地 AS 之外，也不能发布到联盟中的其他子 AS。

组网应用

如图 3-40 所示，USG9500_B 分别与 USG9500_A、USG9500_C 之间建立 EBGP 连接。通过在 USG9500_A 上配置 No_Export 团体属性，使得 AS10 发布到 AS20 中的路由，不再被 AS20 向其他 AS 发布。

图3-40 配置 BGP 团体组网图



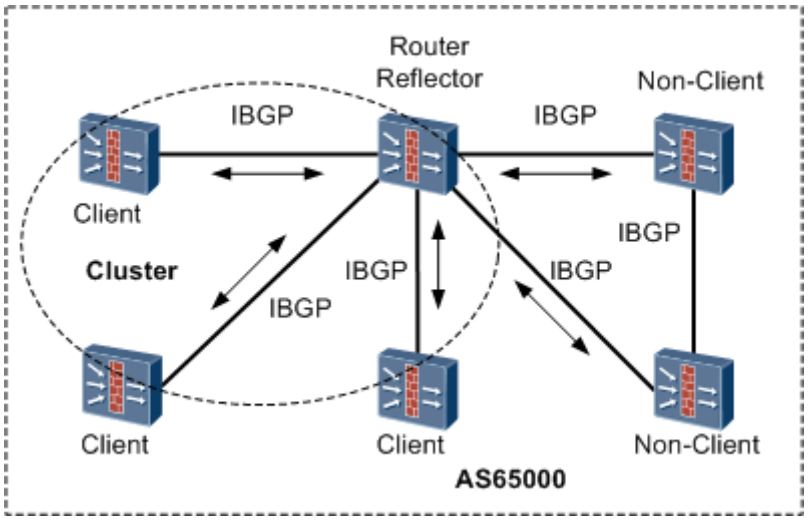
3.6.3.6 路由反射器

为保证 IBGP 对等体之间的连通性，需要在 IBGP 对等体之间建立全连接（Full-mesh）关系。假设在一个 AS 内部有 n 台路由器，那么应该建立的 IBGP 连接数就为 n(n-1)/2。当 IBGP 对等体数目很多时，对网络资源和 CPU 资源的消耗都很大。利用路由反射可以解决这一问题。

在一个 AS 内，其中一台路由器作为路由反射器 RR（Route Reflector），其它路由器作为客户机（Client）。客户机与路由反射器之间建立 IBGP 连接。路由反射器和它的客户机组成一个集群（Cluster）。路由反射器在客户机之间反射路由信息，客户机之间不需要建立 BGP 连接。

既不是反射器也不是客户机的 BGP 设备被称为非客户机（Non-Client）。非客户机与路由反射器之间，以及所有的非客户机之间仍然必须建立全连接关系。如图 3-41 所示。

图3-41 路由反射器示意图



应用

当 RR 收到对等体发来的路由，首先使用 BGP 选路策略来选择最佳路由。在向 IBGP 邻居发布学习到的路由信息时，RR 按照 RFC2796 中的规则发布路由。

- 从非客户机 IBGP 对等体学到的路由，发布给此 RR 的所有客户机。
- 从客户机学到的路由，发布给此 RR 的所有非客户机和客户机（发起此路由的客户机除外）。
- 从 EBGP 对等体学到的路由，发布给所有的非客户机和客户机。

RR 的配置方便，只需要对作为反射器的路由器进行配置，客户机并不需要知道自己是客户机。

在某些网络中，路由反射器的客户机之间已经建立了全连接，它们可以直接交换路由信息，此时客户机到客户机之间的路由反射是没有必要的，而且还占用带宽资源。USG9500 支持配置命令 **undo reflect between-clients** 来禁止客户机之间的路由反射，但客户机到非客户机之间的路由仍然可以被反射。缺省情况下，允许客户机之间的路由反射。

Originator_ID 属性

RFC2796 定义了 Originator_ID 属性和 Cluster_List 属性，用于检测和防止路由环路。

Originator_ID 属性长 4 字节，由路由反射器（RR）产生，携带了本地 AS 内部路由发起者的 Router ID。

- 当一条路由第一次被 RR 反射的时候，RR 将 Originator_ID 属性加入这条路由，标识这条路由的发起路由器。如果一条路由中已经存在了 Originator_ID 属性，则 RR 将不会创建新的 Originator_ID。
- 当其他 BGP Speaker 接收到这条路由的时候，将比较收到的 Originator_ID 和本地的 Router ID，如果两个 ID 相同，BGP Speaker 会忽略掉这条路由，不做处理。

Cluster_List 属性

对于 AS 之间，BGP 用于防止环路的主要措施是通过 AS_Path 属性记录途经的 AS 路径，带有本地 AS 号的路由将被路由器丢弃；对于 AS 之内，BGP 防止路由环路的方法是禁止 IBGP 对等体发布从 AS 内部学来的路由。

RR 的实现是基于放宽对“BGP 在 AS 内学到的路由不会在 AS 中转发”的要求，即允许 IBGP 对等体之间发布从 AS 内部学来的路由。在这种情况下，Cluster_List 属性被引入，用于防止 AS 内部的环路。

路由反射器和它的客户机组成一个集群（Cluster）。在一个 AS 内，每个路由反射器使用唯一的 CLUSTER_ID 作为标识。

为防止产生路由环路，路由反射器使用 CLUSTER_LIST，记录反射路由经过的所有 CLUSTER_ID。

Cluster_List 由一系列的 Cluster_ID 组成，描述了一条路由所经过的反射器路径，这和描述路由经过的 AS 路径的 AS_Path 属性有相似之处。Cluster_List 由路由反射器产生。

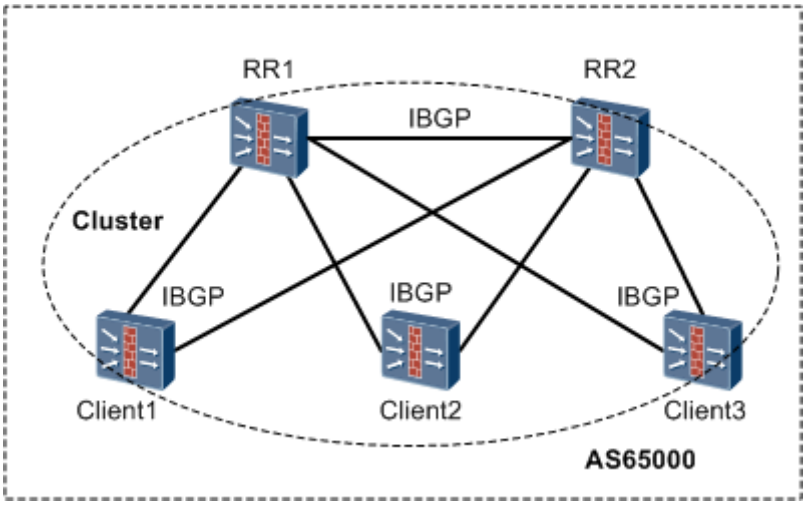
- 当 RR 在它的客户机之间或客户机与非客户机之间反射路由时，RR 会把本地 Cluster_ID 添加到 Cluster_List 的前面。如果 Cluster_List 为空，RR 就创建一个。
- 当 RR 接收到一条更新路由时，RR 会检查 Cluster_List。如果 Cluster_List 中已经有本地 Cluster_ID，丢弃该路由；如果没有本地 Cluster_ID，将其加入 Cluster_List，然后反射该更新路由。

备份 RR

为增加网络的可靠性，防止单点故障，有时需要在一个集群中配置一个以上的路由反射器。这时，相同集群中的路由反射器要共享相同的 Cluster_ID，以避免路由环路。USG9500 中需要使用命令 **reflector cluster-id** 给所有位于同一个集群内的路由反射器配置相同的 Cluster_ID。

在冗余的环境里，客户机会收到不同反射器发来的到达同一目的地的多条路由，这时客户机应用 BGP 选择路由的策略来选择最佳路由。

图3-42 备份路由反射器



如图 3-42，路由反射器 RR1 和 RR2 在同一个 Cluster 内。RR1 和 RR2 之间配置 IBGP 连接，即两个反射器互为非客户机。

- 当客户机 Client1 从外部对等体接收到一条更新路由后，它通过 IBGP 向 RR1 和 RR2 通告这条路由。
- RR1 接收到该更新路由后，它向其他的客户机（Client2、Client3）和非客户机（RR2）反射，同时将本地 Cluster_ID 添加到 Cluster_List 前面。
- RR2 接收到该反射路由后，检查 Cluster_List，发现自己的 Cluster_ID 已经包含在 Cluster_List 中。因此，它丢弃该更新路由，不再向自己的客户机反射。

说明
Cluster_List 的应用保证了同一 AS 内的不同 RR 之间不出现路由循环。

AS 内多个集群

一个 AS 中可能存在多个集群（Cluster）。各个 RR 之间是 IBGP 对等体的关系，一个 RR 可以把另一个 RR 配置成自己的客户机或非客户机。因此可以灵活的配置 AS 内部集群与集群之间的关系。

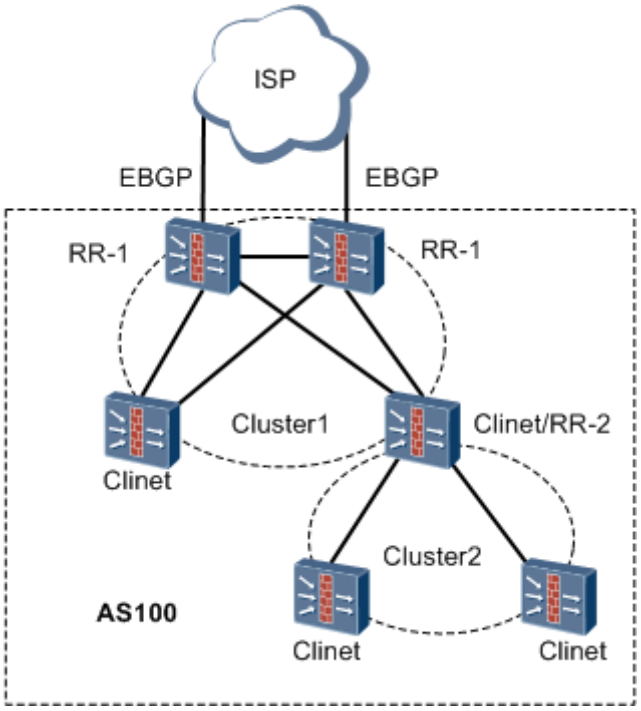
例如，一个骨干网被分成多个反射集群，每个 RR 将其它的 RR 配置成非客户机，各 RR 之间建立全连接。每个客户机只与所在集群的 RR 建立 IBGP 连接。这样该自治系统内的所有 BGP 路由器都会收到反射路由信息。

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 3-43，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。

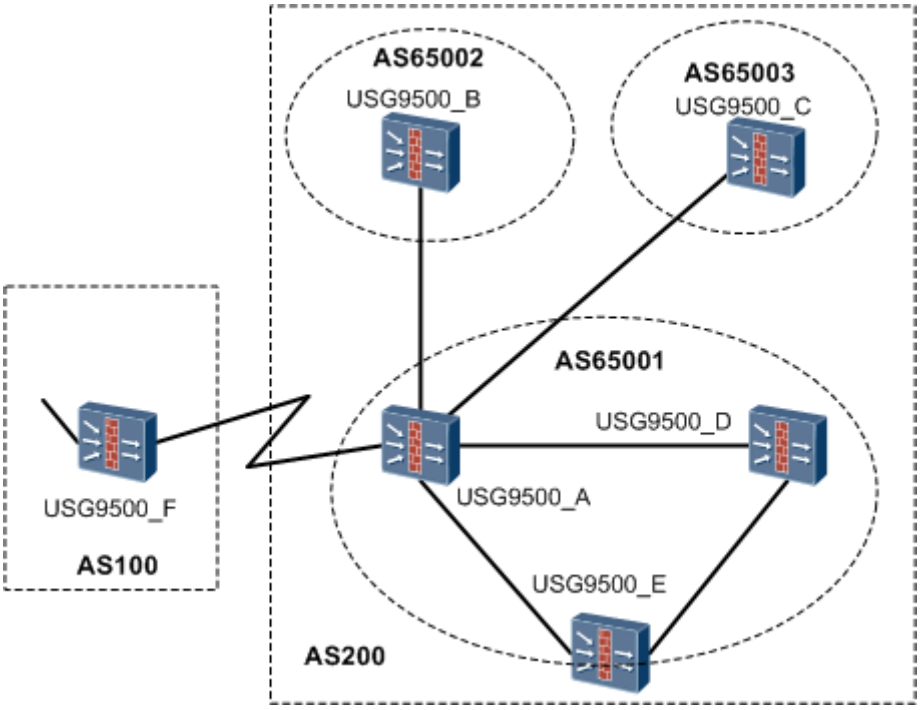
图3-43 分级反射器



3.6.3.7 BGP 联盟

联盟（Confederation）是处理 AS 内部的 IBGP 网络连接激增的另一种方法，它将一个 AS 划分为若干个子自治系统（Sub AS），每个子 AS 内部建立 IBGP 全连接关系，子 AS 之间建立 EBGP 连接关系。如图 3-44 所示。

图3-44 联盟示意图



在不属于联盟的 BGP Speaker（如 AS100 中的设备）看来，属于同一个联盟的多个子 AS（AS65001、AS65002、AS65003）是一个整体，外界不需要了解内部的子 AS 情况，联盟 ID 就是标识联盟这一整体的自治系统号，如图 3-44 中的 AS200 就是联盟 ID。

如图 3-44 所示，AS200 中有多台 BGP 设备，为了减少 IBGP 的连接数，现将他们划分为 3 个子自治系统：AS65001、AS65002 和 AS65003。其中 AS65001 内的三台设备建立 IBGP 全连接。

应用和限制

联盟需要在每个设备上进行配置，且要求加入联盟的设备具有联盟功能。

联盟的缺陷是：从非联盟向联盟方案转变时，要求设备重新进行配置，逻辑拓扑也要改变。

在大型 BGP 网络中，路由反射器和联盟可以被同时使用。

3.6.3.8 MP-BGP

传统的 BGP-4 只能管理 IPv4 单播路由信息，对于使用其它网络层协议（如 IPv6、组播等）的应用，在跨 AS 传播时就受到一定限制。

为了提供对多种网络层协议的支持，IETF 对 BGP-4 进行了扩展，形成 MP-BGP，目前的 MP-BGP 标准是 RFC4760（Multiprotocol Extensions for BGP-4，BGP-4 的多协议扩展）。MP-BGP 向前兼容，即支持 BGP 扩展的路由器与不支持 BGP 扩展的路由器可以互通。

MP-BGP 在现有 BGP-4 协议的基础上增强功能，使 BGP 能够为多种路由协议提供路由信息。

- MP-BGP 可以同时为单播和组播维护路由信息，将它们储存在不同的路由表中，保持单播和组播之间路由信息相互隔离。
- MP-BGP 可以同时支持单播和组播模式，为两种模式构建不同的网络拓扑结构。
- 原 BGP-4 支持的单播路由策略和配置方法大部分都可应用于组播模式，从而根据路由策略为单播和组播维护不同的路由。

扩展属性

BGP-4 使用的报文中，与 IPv4 相关的三处信息都由 Update 报文携带，这三处信息分别是：NLRI 字段、Next_Hop 属性、Aggregator 属性（该属性中包含形成聚合路由的 BGP Speaker 的 IP 地址）。

为实现对多种网络层协议的支持，BGP-4 需要将网络层协议的信息反映到 NLRI 及 Next_Hop。MP-BGP 中引入了两个新的路径属性：

- MP_REACH_NLRI：Multiprotocol Reachable NLRI，多协议可达 NLRI。用于发布可达路由及下一跳信息。
- MP_UNREACH_NLRI：Multiprotocol Unreachable NLRI，多协议不可达 NLRI。用于撤销不可达路由。

这两种属性都是可选非过渡（Optional non-transitive）的，因此，不提供多协议能力的 BGP Speaker 将忽略这两个属性的信息，不把它们传递给其它邻居。

3.6.3.9 BGP GR

当 BGP 协议重启时会导致对等体关系重新建立和转发中断，使能平滑重启 GR（Graceful Restart）功能后可以避免流量中断。

在系统进行 GR 时有两种角色：

- GR Restarter：以 GR 方式重启的设备，指由管理员触发或故障触发重启的设备，必须是有 GR 能力的设备，即路由协议使能并协商了 GR 能力。
- GR Helper：GR Restarter 的邻居，本身必须具备了 GR 能力，才能协助 GR Restarter 进行 GR。

系统进行 GR 时有如下会话和定时器的概念：

- GR Session：GR 会话，是 GR Restarter 和 GR Helper 之间的协议关系。通过控制协议的会话协商机制，GR Restarter 和 GR Helper 可以了解彼此的 GR 能力，建立有 GR 能力的会话。
- GR Time：是 GR Helper 发现 GR Restarter Down 后，保持转发信息不删除的时间。当 GR Helper 发现对端的 GR Restarter 处于 Down 状态时，在 GR Time 时间内仍保留从 GR Restarter 得到的拓扑信息或路由，不删除这些信息。

BGP GR 的基本原理是：

- 利用 BGP 的能力协商机制，在 Restarting 前建立进行 GR 能力协商，建立有 GR 能力的 BGP 会话。

- 当邻居检查到 GR Restarter 发生重启时，不删除和 GR Restarter 相关的路由和转发表项，而是等待重建 BGP 连接。
- GR Restarter 和邻居在新连接上完成 BGP 路由更新。

这样既可以保证转发不中断，也可以让 BGP 协议的震荡仅限于和 Restarting 设备相连邻居，不会扩散到整个路由域，这对 BGP 这种路由数据大的协议来说，尤其有意义。

3.6.3.10 BGP 安全性

BGP 验证

BGP 使用 TCP 作为传输层协议，为提高 BGP 的安全性，可以在建立 TCP 连接时进行 MD5 认证。但 BGP 的 MD5 认证并不能对 BGP 报文认证，它只是为 TCP 连接设置 MD5 认证密码，由 TCP 完成认证。如果认证失败，则不建立 TCP 连接。

BGP 的 GTSM

通用 TTL 安全保护机制 GTSM（Generalized TTL Security Mechanism）机制通过对 TTL（Time-to-Live：生存时间字段，设置数据报文可以经过的路由器的最多数量）的检测来达到防止攻击的目的，如果攻击者模拟真实的 BGP 协议报文，对一台路由器不断的发送报文，路由器收到这些报文后，发现是发送给本机的报文，则直接上送控制层面的 BGP 协议处理，而不加辨别其“合法性”，这样导致路由器控制层面因为处理这些“合法”报文，系统异常繁忙，CPU 占用率高。

GTSM 通过检测 IP 报文头中的 TTL 值是否在一个预先定义好的特定范围内，对 IP 层以上业务进行保护，增强系统的安全性。

使能 BGP 的 GTSM 策略后，接口板对所有 BGP 报文的 TTL 值进行检查。根据实际组网的需要，对于不符合 TTL 值范围的报文，GTSM 可以设置为通过或丢弃。配置 GTSM 缺省动作为丢弃时，可以根据网络拓扑选择合适的 TTL 有限值范围，不符合 TTL 值范围的报文会被接口板直接丢弃，这样就避免了网络攻击者模拟的“合法” BGP 报文占用 CPU。

对于丢弃的报文，可以通过 LOG 信息开关，控制是否对报文被丢弃的情况记录日志，以方便故障的定位。

3.6.3.11 BFD for BGP

BFD（Bidirectional Forwarding Detection）可为 BGP 协议提供更快速的链路故障检测。BGP 协议通过周期性的向对等体发送报文来实现邻居检测机制。但这种机制检测到故障所需时间比较长，超过 1 秒钟。当数据达到 Gbit/s 的速率等级时，这种机制的检测时间将导致大量数据丢失，无法满足电信级网络高可靠性的需求。

因此，BGP 协议通过引入 BFD for BGP 特性，利用 BFD 的快速检测机制（检测到故障的时间可以达到毫秒级）即迅速发现 BGP 对等体间链路的故障，并报告给 BGP 协议，从而实现 BGP 路由的快速收敛。

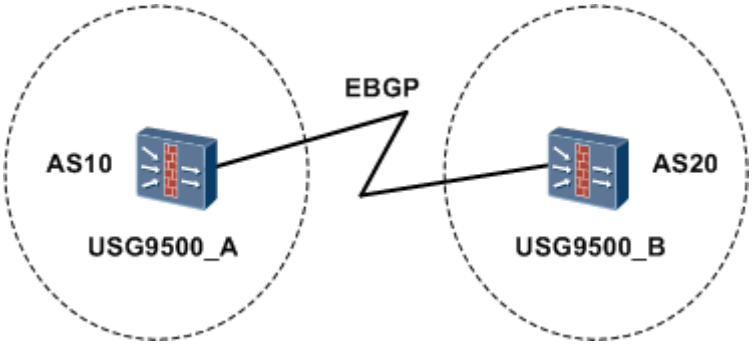
BFD for BGP 适用于 IPv4 和 IPv6 两种网络。

组网

如图 3-45 所示，USG9500_A 和 USG9500_B 分别属于 AS100 和 AS200，两台路由器直接相连并建立 EBGP 连接。

使用 BFD 检测 USG9500_A 和 USG9500_B 之间的 BGP 邻居关系，当 USG9500_A 和 USG9500_B 之间的链路发生故障时，BFD 能够快速检测到故障并通告给 BGP 协议。

图3-45 BFD for BGP 组网图



3.6.3.12 BGP4+

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，在跨自治系统传播路由信息时就受到一定限制。

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 RFC4760（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 协议的信息反映到 NLRI（Network Layer Reachable Information）属性及 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。

在 BGP4+中，BGP 协议原有的消息机制和路由机制并没有改变。

4 安全

关于本章

4.1 ACL

4.1.1 定义

USG9500 必须具备控制网络数据流的能力，用于满足安全性、QoS（Quality of Service）需求和各种策略制定等各个方面。实现数据流控制的手段之一是使用访问控制列表 ACL。

ACL 是由 permit 或 deny 语句组成的一系列有顺序的规则，这些规则主要通过数据包的源地址、目的地址、端口号、上层协议等信息来描述。

4.1.2 应用

作为数据流控制的一种常用的手段，ACL 被应用于很多方面。例如：

- 包过滤

  包过滤作为一种网络安全保护机制，用于在两个不同安全区域之间控制流入和流出的数据。为了实现包过滤，需要通过 ACL 定义一系列的过滤规则，然后将 ACL 规则应用于 USG9500 的不同安全区域之间。USG9500 转发数据包时，会将数据包的信息（例如源地址/目的地址、源端口/目的端口和上层协议等）与设定的 ACL 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。

- NAT

  NAT（Network Address Translation）是将数据报报头中的 IP 地址转换为另一个 IP 地址的过程，主要用于实现内部网络（私有 IP 地址）访问外部网络（公有 IP 地址）以及解决 IP 地址缺乏的问题。

在实际应用中，我们可能仅希望某些内部主机（具有私有 IP 地址）具有访问 Internet（外部网络）的权限，而其他内部主机则不允许。这种情况是通过将 ACL 和 NAT 地址池进行关联来实现的，即只有满足 ACL 条件的数据报文才可以进行地址转换，从而有效地控制地址转换的使用范围。

- IPSec

  IPSec 协议族是 IETF（Internet Engineering Task Force）制定的一系列协议，它通过 IP 层的加密与数据源验证机制，确保在 Internet 上参与通信的两个网络节点之间传输的数据包具有私有性、完整性和真实性。

  IPSec 能够对不同的数据流施加不同的安全保护，例如对不同的数据流使用不同的安全协议、算法和密钥。实际应用中，数据流首先通过 ACL 来定义，匹配同一个 ACL 的所有流量在逻辑上作为一个数据流。然后，通过在安全策略中引用该 ACL，从而确保指定的数据流受到保护。

- QoS

  QoS 用来评估服务方满足客户需求的能力。在 Internet 上保证 QoS 的有效办法是增加网络层在流量控制和资源分配上的功能，为有不同服务需求的业务提供有区别的服务。

  流分类是有区别地进行服务的前提和基础。实际应用中，首先制定流分类策略（规则），流分类规则既可以使用 IP 报文头的 ToS（Type of Service）字段内容来识别不同优先级特征的流量，也可以通过 ACL 定义流分类的策略，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。然后，在流量监管、流量整形、拥塞管理和拥塞避免等具体实施上引用流分类策略或 ACL。

- 路由策略

  路由策略是指在发送与接收路由信息时所实施的策略，它能够对路由信息进行过滤。

  路由策略有多种过滤方法。其中，ACL 作为它的一个重要过滤器被广泛使用，即用户使用 ACL 指定一个 IP 地址或子网的范围，作为匹配路由信息的目的网段地址、源网段地址或下一跳地址。

4.1.3 步长设定

配置 USG9500 的 ACL 时，可以为一个 ACL 规则组指定一个“步长”。步长的含义是：自动为 ACL 子规则分配编号的时候，每个 ACL 规则组的子规则编号之间的差值。如果步长设定为 5，子规则编号分配是按照 5、10、15……这样的规则分配的。默认情况下，ACL 规则组的步长为 5。

只有 ACL 规则组下没有子规则时，才能改变步长。ACL 规则组下有子规则时，必须删除已经存在的子规则，然后再使用 **step** 命令改变步长或者使用 **undo step** 命令将步长变为默认值。

使用步长设定的好处是，可以方便在子规则之间插入新的规则。例如配置好了 4 个规则，子规则编号为：5、10、15、20。此时希望能在第一条规则之后插入一条规则，则可以使用 **rule 6 xxxx** 命令在 5 和 10 之间插入一条编号为 6 的子规则。

4.1.4 USG9500 支持的 ACL

支持的 ACL 类型

USG9500 支持的 ACL 类型如表 4-1 所示。

表4-1 USG9500 支持的 ACL 类型

ACL 类型	数字范围	描述
IPv4、IPv6 基本 ACL	2000～2999	仅使用源地址信息定义数据流。
IPv4、IPv6 高级 ACL	3000～3999	可以根据报文的源 IP 地址、目的 IP 地址、源端口号、目的端口号、承载的协议信息（例如 ICMP/ICMPv6 协议的类型）等多种元素组合定义规则。

匹配顺序

一个 ACL 规则可以由多条 **permit** 或 **deny** 语句组成，每一条语句描述的规则是不相同的，这些规则可能存在重复或矛盾的地方。在将一个数据包和 ACL 规则进行匹配的时候，需要确定规则的匹配顺序。USG9500 按照如下原则进行匹配：

- 在同一 ACL 规则组中，rule-id 小的规则被优先匹配。
- 不同 ACL 规则组，按照用户配置 ACL 规则的先后顺序进行匹配。

数据流一旦与一条 rule 规则匹配成功，将不再继续向下一规则匹配。USG9500 将根据该 rule 规则的动作，对数据流进行后续操作。

源地址和通配符掩码

在使用 ACL 时，需要指定源地址。源地址可以指一台主机、一组主机、整个子网或整个网络。源地址的范围是由通配符掩码字段来确定的。

通配符掩码不同于子网掩码，它是以 0 表示必须匹配的位，以 1 表示无需匹配的位，将 source-wildcard 和 source-address 进行“与”运算，从而得出源地址范围。例如：

```
source-address = 192.168.15.16   11000000.10101000.00001111.00010000
source-wildcard = 0.0.0.255      00000000.00000000.00000000.11111111
源地址范围 = 192.168.15.0～255          11000000.10101000.00001111.00000000～11111111
```

any 的含义指来自任何地址的包都符合匹配条件，此时 *source-wildcard* 取 255.255.255.255，*source-address* 可以取任意地址。

基于时间段的 ACL 规则

在允许或拒绝用户对资源的访问方面，当今的网络安全策略需要有更大的控制灵活性。例如在某些情况下，系统管理员可能只想在某些特定时间段才允许某些数据流通过，或只允许用户在一天的某些时间段访问某些资源。此时可以使用基于时间段的 ACL 规则。

引用地址组和端口组的 ACL 规则

为简化 ACL 规则的配置和维护，USG9500 支持引用地址组和端口组的 ACL。

通过地址组和端口组描述的一条 rule 规则，在使用时体现为具有相同优先级的传统 rule 规则的集合。具体公式为：

新集合中相同优先级的 rule 规则元素个数=地址组 1 元素个数 × 地址组 2 元素个数 × 端口组 1 元素个数 × 端口组 2 元素个数

例如，配置两个地址组和一个端口组，分别包含两个元素，并在 ACL 3000 中应用。

```
<USG9500> system-view
[USG9500] ip address-set a1
[USG9500-address-set-a1] address 1 1.1.1.1 0
[USG9500-address-set-a1] address 2 2.2.2.1 0
[USG9500-address-set-a1] quit
[USG9500] ip address-set a2
[USG9500-address-set-a2] address 1 3.3.3.1 0
[USG9500-address-set-a2] address 2 4.4.4.1 0
[USG9500-address-set-a2] quit
[USG9500] ip port-set p1 protocol tcp
[USG9500-tcp-port-set-p1] port 1 eq 21
[USG9500-tcp-port-set-p1] port 2 eq 22
[USG9500-tcp-port-set-p1] quit
[USG9500] acl 3000
[USG9500-acl-adv-3000] rule permit tcp source address-set a1 destination address-set a2 destination-port port-set p1
```

上述命令的配置效果与如下几个 ACL 规则的配置效果相同：

```
<USG9500> system-view
[USG9500] acl 3000
[USG9500-acl-adv-3000] rule permit tcp source 1.1.1.1 0 destination 3.3.3.1 0 destination-port eq 21
[USG9500-acl-adv-3000] rule permit tcp source 1.1.1.1 0 destination 3.3.3.1 0 destination-port eq 22
[USG9500-acl-adv-3000] rule permit tcp source 1.1.1.1 0 destination 4.4.4.1 0 destination-port eq 21
[USG9500-acl-adv-3000] rule permit tcp source 1.1.1.1 0 destination 4.4.4.1 0 destination-port eq 22
[USG9500-acl-adv-3000] rule permit tcp source 2.2.2.1 0 destination 3.3.3.1 0 destination-port eq 21
[USG9500-acl-adv-3000] rule permit tcp source 2.2.2.1 0 destination 3.3.3.1 0 destination-port eq 22
[USG9500-acl-adv-3000] rule permit tcp source 2.2.2.1 0 destination 4.4.4.1 0 destination-port eq 21
```

```
[USG9500-acl-adv-3000] rule permit tcp source 2.2.2.1 0 destination 4.4.4.1 0
destination-port eq 22
```

4.2 安全策略

4.2.1 包过滤

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数据。USG9500 转发数据包时，先检查包头信息（例如包的源地址/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。

为了实现数据包过滤，需要配置一系列的过滤规则。采用 ACL 定义过滤规则，然后将 ACL 应用于 USG9500 不同区域之间，从而实现包过滤。

USG9500 还支持域内包过滤，通过 ACL 过滤规则，对安全区域内流动的数据进行控制。

4.2.2 会话表

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。

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，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。

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 USG9500 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，USG9500 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。

4.2.3 ASPF

ASPF 是针对应用层的包过滤，即基于状态的报文过滤。它和普通的静态防火墙协同工作，以便于实施内部网络的安全策略。ASPF 能够检测试图通过 USG9500 的应用层协议会话信息，阻止不符合规则的数据报文穿过。

为保护网络安全，基于 ACL 规则的包过滤可以在网络层和传输层检测数据包，防止非法入侵。

ASPF 能够检测应用层协议的信息，并对应用的流量进行监控。ASPF 通过维护会话的状态和检查会话报文的协议和端口号等信息，阻止恶意的入侵。

在 USG9500 中，ASPF 还提供以下功能：

- Java 阻断（Java Blocking），保护网络不受有害 Java Applets 的破坏。
- ActiveX 阻断（ActiveX Blocking），保护网络不受有害 ActiveX 的破坏。

域间 ASPF 支持的协议包括：DNS、FTP、H323、MGCP、MMS、MSN、PPTP、QQ、RTSP、SCCP、SIP、SQLNET、ILS、NETBIOS、RSH、user-defined 以及 IPv6 FTP、IPv6 RTSP。

域内 ASPF 支持的协议包括：FTP 和 RTSP。

QQ/MSN 聊天的检测

目前，为了节省有限的 IP 地址资源，绝大部分网络均部署了 NAT 设备以提供地址转换。对于纯文本聊天，由于在 QQ/MSN 服务器中保存了聊天用户的地址映射信息，信息交互可以顺利地通过 QQ/MSN 服务器中转。

聊天用户可能传送文件或进行音频/视频聊天，如果 QQ/MSN 服务器中转此类报文将消耗过多资源，无法保证对纯文本聊天报文的正常中转。QQ/MSN 服务器希望两个用户通过网络设备直接交互大流量的文件/音频/视频信息，但是由于一般 NAT 设备需要转换聊天用户的地址信息，无法实现该需求。

配置 USG9500 的 NAT 功能时，可以在相关安全域间启动 QQ/MSN 检测功能，USG9500 在 QQ/MSN 聊天启动时则会创建地址映射关系，从而使两个私网用户直接传送文件和进行音频/视频聊天。

三元组 ASPF

USG9500 相当于一个五元组的 NAT 设备，即 USG9500 上的每个会话的建立都需要五元组：源 IP 地址、源端口、目的 IP 地址、目的端口、协议号。只有这些元素都具备了，会话才能建立成功，报文才能通过。而一些象 QQ、MSN 等实时通讯工具，通过 NAT 设备，却需要按三元组处理：源 IP 地址，源端口、协议号。为了适配类似 QQ、MSN 等通讯机制，变五元组处理方式为三元组方式，让类似 QQ、MSN 等的通讯方式能够正常的穿越。

除 QQ、MSN 穿越 NAT 设备外，其他仅使用源 IP 地址、源端口、协议号的会话，如 TFTP（Trivial File Transfer Protocol），同样需要配置三元组 ASPF。

4.2.4 黑名单

黑名单是 USG9500 一个重要的安全特性，其特点为可以由 USG9500 动态地进行添加或删除。同基于 ACL 的包过滤功能相比，由于黑名单仅对 IP 地址进行匹配，可以以很高的速度实现黑名单表项匹配，从而快速有效地屏蔽特定 IP 地址的用户。

黑名单表项有如下两种创建方式：

- 通过命令行手工创建。
- 通过攻击防范模块动态创建。

当 USG9500 根据报文的行为特征察觉到特定 IP 地址的用户的攻击企图后，主动将其插入黑名单表项，过滤从该 IP 地址发送的报文，从而保障网络安全。

4.2.5 端口映射

应用层协议一般使用通用的端口号（知名端口号）进行通信。端口映射 PAM（Port to Application Mapping）允许用户针对不同的应用在系统定义的端口号之外定义一组新的端口号。端口映射提供了一些机制来维护和使用用户定义的端口配置信息。端口映射能够对不同的应用协议创建和维护一张系统定义（system-defined）和用户定义（user-defined）的端口映射表。

端口映射支持基于基本访问控制列表（ACL）的主机端口映射。

主机端口映射是对去往某些特定主机的报文建立自定义端口号和应用协议的映射，例如：将去往 10.110.0.0 网段的主机使用 8080 端口的 TCP 报文识别为 HTTP 报文。主机的范围可由基本 ACL 指定。

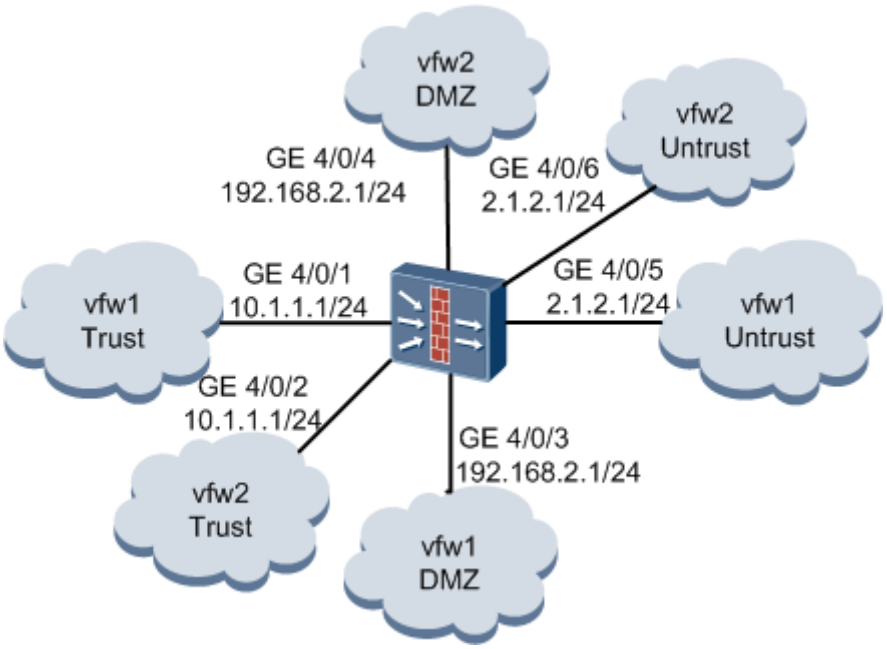
4.2.6 虚拟防火墙

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。

华为公司推出 USG9500 多实例解决方案。USG9500 多实例的配置案例组网图如图 4-1 所示。图 4-1 中将一台 USG9500 从逻辑上划分为多个 VPN 实例，向多个小型私有网络提供相对独立的安全保障。

图4-1 虚拟防火墙配置组网图



每台虚拟防火墙都是 VPN 实例（VPN-Instance）、安全实例和配置实例的综合体。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 路由将为各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 和 NAT 地址池，并能为虚拟防火墙用户提供地址绑定、黑名单、NAT、包过滤、统计、攻击防范、ASPF 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 路由和安全实例。

4.3 攻击防范

4.3.1 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。

在 USG9500 中，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。

4.3.2 网络攻击类型介绍

网络攻击可分为拒绝服务型攻击、扫描窥探攻击和畸形报文攻击三大类：

- 拒绝服务型攻击
  - DoS（Denial of Service）攻击是使用大量的数据包攻击系统，使系统无法接受正常用户的请求，或者主机挂起不能提供正常的工作。主要 DoS 攻击有 SYN Flood、UDP Flood、DNS Flood 等。拒绝服务攻击和其他类型的攻击不同之处在于：攻击者并不是去寻找进入内部网络的入口，而是阻止合法用户访问资源或设备。
  - DDoS（Distributed Denial of Service）攻击是一种 DoS 攻击。这种攻击是使用攻击者控制的几十台或几百台计算机攻击一台主机，使系统无法接受正常用户的请求，或者挂起不能正常的工作。
- 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 包时会出现崩溃，给目标系统带来损失。主要的畸形报文攻击有 Ping of Death、Teardrop 等。

4.3.3 典型网络攻击介绍

目前网络上的典型攻击有如下几种：

- 拒绝服务型攻击
  - SYN Flood 攻击

    由于资源的限制，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，它伪造一个 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报文，会在被攻击主机上出现大量的半连接，消耗尽其资源，使正常的用户无法访问。直到半连接超时。在一些创建连接不受限制的实现里，SYN Flood 具有类似的影响，它会消耗掉系统的内存等资源。
  - ICMP Flood 攻击

    攻击者通过向服务器发送大量的 ICMP 消息（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。
  - 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。
  - DNS-flood 攻击

    DNS-flood 攻击是一种 DDoS 攻击手段。攻击者在短时间内通过向 DNS（Domain Name System）服务器发送大量的查询报文，使得服务器不得不对所有的查询请求进行回应，进而，导致 DNS 服务器无法为合法用户提供服务。
- 扫描窥探攻击

  地址扫描与端口扫描攻击，即运用扫描工具探测目标地址和端口，对此作出响应的表示其存在，用来确定哪些目标系统确实存活着并且连接在目标网络上，这些主机使用哪些端口提供服务。
- 畸形报文攻击
  - IP 地址欺骗攻击

    为了获得访问权，入侵者生成一个带有伪造源地址的报文。对于使用基于 IP 地址验证的应用来说，此攻击方法可以导致未被授权的用户可以访问目的系统，甚至是以 root 权限来访问。即使响应报文不能达到攻击者，同样也会造成对被攻击对象的破坏。这就造成 IP 地址欺骗攻击。
  - Land 攻击

所谓 Land 攻击，就是把 TCP SYN 包的源地址和目标地址都配置成受害者的 IP 地址。这将导致受害者向它自己的地址发送 SYN-ACK 消息，结果这个地址又发回 ACK（ACKnowledgement）消息并创建一个空连接，每一个这样的连接都将保留直到超时掉。各种受害者对 Land 攻击反应不同，许多 UNIX 主机将崩溃，Windows NT 主机会变的极其缓慢。

- Smurf 攻击

  简单的 Smurf 攻击，用来攻击一个网络。方法是发 ICMP 应答请求，该请求包的目标地址配置为受害网络的广播地址，这样该网络的所有主机都对此 ICMP 应答请求作出答复，导致网络阻塞，这比 ping 大包的流量高出一或两个数量级。高级的 Smurf 攻击，主要用来攻击目标主机。方法是将上述 ICMP 应答请求包的源地址改为受害主机的地址，最终导致受害主机雪崩。攻击报文的发送需要一定的流量和持续时间，才能真正构成攻击。理论上讲，网络的主机越多，攻击的效果越明显。Smurf 攻击的另一个变体为 Fraggle 攻击。

- Fraggle 攻击

  UDP 端口 7（ECHO）和端口 19（Chargen）在收到 UDP 报文后，都会产生回应。在 UDP 的 7 号端口收到报文后，会像 ICMP Echo Reply 一样回应收到的内容；而 UDP 的 19 号端口在收到报文后，会产生一串字符流。就像 ICMP 一样，这两个 UDP 端口都会产生大量无用的应答报文，占满网络带宽。

  攻击者可以向攻击目标所在的网络发送源地址为被攻击主机、而目的地址为其所在子网的广播地址或子网网络地址的 UDP 报文，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受害主机发送回应报文，从而产生大量的流量，导致受害网络的阻塞或受害主机的崩溃。

  子网上没有启动这些功能的系统将产生一个 ICMP 不可达消息，因而仍然消耗带宽。也可将源端口改为 Chargen，目的端口为 ECHO，这样会自动不停地产生回应报文，其危害性更大。

- WinNuke 攻击

  WinNuke 攻击通常向装有 Windows 系统的特定目标的 NetBIOS 端口（139）发送 OOB（Out-Of-Band）数据包，引起一个 NetBIOS 片断重叠，致使目标主机崩溃。还有一种是 IGMP（Internet Group Management Protocol）分片报文，一般情况下，IGMP 报文是不会分片的，所以，不少系统对 IGMP 分片报文的处理有问题。如果收到 IGMP 分片报文，则可能是受到了 WinNuke 攻击。

- Large ICMP 攻击

  Large ICMP 攻击是指利用尺寸超大的 ICMP 报文对目标系统进行攻击。对于有些设备，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。

- Ping of Death 攻击

  IP 报文的长度字段为 16 位，这表明一个 IP 报文的最大长度为 65535 字节。对于 ICMP 回应请求报文，如果数据长度大于 65507，就会使 ICMP 数据+IP 头长度(20)+ICMP 头长度（8）>65535。对于有些设备，在接收到一个这样的报文后，由于处理不当，会造成系统崩溃、死机或重启。所谓 Ping of Death，就是利用一些尺寸超大的 ICMP 报文对系统进行的一种攻击。

4.3.4 攻击防范原理介绍

DNS-flood 攻击防范原理介绍

USG9500 通过在接口对 DNS 报文进行合法性检查及反向源认证，实现 DNS Flood 的攻击防范功能。

SYN Flood 攻击防范原理介绍

USG9500 通过限制 SYN 报文的速率来防范 SYN Flood 攻击。可以基于接口、IP 地址和安全区域来限制 SYN 报文的速率。

当报文的来回路径一致时，可以开启 TCP 代理（TCP Proxy）功能，对 SYN Flood 攻击进行防范。

当报文的来回路径不一致时，可以配置 TCP 反向源探测，通过对 TCP 协议的源 IP 进行反向探测技术，解决了虚假 IP 发起的 SYN-Flood 攻击防范。

TCP 反向源探测是对攻击者采用虚假 IP 进行攻击的一种有效防范。当启动了 TCP 反向源探测后，USG9500 对经过的 TCP SYN 报文进行源 IP 地址的反向探测，确定源 IP 地址为有效 IP 后方允许报文通过。

说明

TCP 反向源探测是基于实接口进行配置，所有虚接口（子接口、trunk、Vlanif 等）都不能配置该功能。

同时配置 TCP 代理功能和 TCP 反向源探测功能时，优先采用 TCP 反向源探测方式进行 SYN Flood 攻击防范。

UDP Flood 攻击防范原理介绍

USG9500 通过限制 UDP 报文的速率来防范 UDP Flood 攻击。可以基于接口、IP 地址和安全区域来限制 UDP 报文的速率。

ICMP Flood 攻击防范原理介绍

USG9500 通过限制 ICMP 报文的速率来防范 ICMP Flood 攻击。可以基于接口、IP 地址和安全区域来限制 ICMP 报文的速率。

IP 地址/端口扫描攻击防范原理介绍

USG9500 对某个 IP 地址或端口的连接速率进行检测，当速率超过阈值时，则认为发生了扫描攻击，将该 IP 地址或端口加入黑名单，禁止建立新的连接。当黑名单老化时间到期后，才允许此 IP 地址或端口建立新的连接。

其他协议报文攻击防范原理介绍

USG9500 处理非 TCP、UDP、ICMP 协议的报文时不创建会话表，每个报文都相当于首包报文。USG9500 通过限制此类报文的速率进行攻击防范，超过速率限制的报文将被丢弃。

基于会话的 TCP/UDP/ICMP 攻击防范原理介绍

当 USG9500 发现 TCP/UDP/ICMP 会话上的报文速度超过设定阈值时，则认为发生了攻击。此时 USG9500 锁定此会话，后续此会话上不再允许报文通过。当此会话连续 3 秒或者 3 秒以上没有流量时，解锁此会话，后续此会话上的报文可以继续通过。

4.4 认证与授权

4.4.1 概述

认证与授权一般采用客户端/服务器结构。客户端运行于被管理的资源这一侧，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。

认证功能

USG9500 支持如下认证方式：

- 本地认证

  当用户接入时，根据 USG9500 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

  当用户接入时，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，由 USG9500 作客户端，与 RADIUS 服务器通信或 HWTACACS 服务器通信。对于 RADIUS 协议可以采用标准 RADIUS 协议或华为公司的扩展 RADIUS 协议，与 iTELLIN/CAMS(Comprehensive Access Management Sever)等设备配合完成认证。

授权功能

USG9500 支持以下授权方式：

- 本地授权

  根据 USG9500 上为本地用户账号配置的相关属性进行授权。
- if-authenticated 授权

  如果用户通过了验证，则对用户授权通过。
- RADIUS 认证成功后授权

  由 RADIUS 服务器对用户认证通过后，即可授权。这是由于 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。
- HWTACACS 授权

  由 HWTACACS 服务器对用户进行授权。

4.4.2 RADIUS 协议简介

认证与授权可以用多种协议来实现，最常用的是 RADIUS 协议。RADIUS 协议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 NAS（Network Access Server）系统。

当用户想要通过某个网络（如电话网）与 NAS 建立连接，从而获得访问其他网络或取得使用某些网络资源的权利时，NAS 起到了验证用户或对应连接的作用。NAS 负责把用户的验证、授权信息传递给 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。

RADIUS 服务器负责接收用户的连接请求，完成验证，并把用户所需的授权信息返回给 NAS。NAS 和 RADIUS 之间的验证信息的传递通过密钥的参与来完成，避免了用户信息在不安全的网络上被窃取。

RADIUS 的消息流程

RADIUS 协议规定了客户/服务器间消息交互的消息流程和消息结构。采用 RADIUS 协议时，服务器就叫 RADIUS 服务器。RADIUS 协议规定的简单消息流程如图 4-2 所示。

图4-2 RADIUS 客户/服务器间消息流程

此时，USG9500 是作为接入服务器（Access Server）。当用户登录 USG9500 时，会遵循如下操作流程。

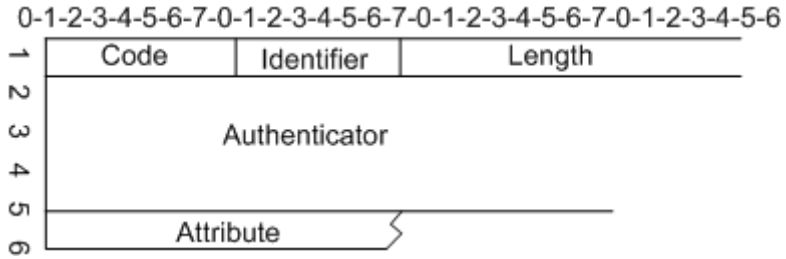
1. 当用户接入 USG9500 时，首先会将用户名和密码发送给 USG9500。
2. USG9500 作为 RADIUS 客户端接收用户名和密码，向 RADIUS 服务器发送认证请求。
3. RADIUS 服务器接收到合法的请求后，对用户名和密码进行认证。
4. 完成认证后，RADIUS 服务器把所需的用户授权信息返回给 USG9500。

登录的用户可以是使用网络资源的 PPP（Point-to-Point Protocol）用户，也可以是对网络设备进行配置、维护的管理用户。

RADIUS 的消息结构

RADIUS 的消息结构如图 4-3 所示。

图4-3 RADIUS 消息结构



具体介绍如下：

- Code

  用于表示 RADIUS 的消息类型，如接入请求、接入允许等。
- Identifier

  一般是顺序递增的数字，用于匹配请求包和回应包。
- 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，包括用户名、用户密码、NAS IP 地址等属性。

RADIUS 的特点

RADIUS 有如下特点：

- 使用 UDP 作为传输协议，具有良好的实时性。
- 支持重传机制和备用服务器机制，具有较好的可靠性。
- 实现比较简单，适用于大用户量时服务器端的多线程结构。

上述特点使得 RADIUS 协议得到了广泛的应用。

作为 RADIUS 协议客户端，NAS 能够实现以下功能：

- 标准 RADIUS 协议及扩充属性，包括 RFC2865、RFC2866。
- 华为扩展的 RADIUS + 1.1 协议。
- 对 RADIUS 服务器状态的主动探测功能。

  如果当前服务器的状态为 DOWN，USG9500 收到认证消息后，启动服务器探测处理，将消息转换为探测报文后向当前服务器发送。如果收到 RADIUS 服务器的回应，则认为该服务器重新可用。
- RADIUS 服务器的自动切换功能。

  当报文等待定时器超时的时候，如果当前发送的 Server 的状态为不可发送，或者发送次数超过当前 Server 的最大重传次数，则需要在配置的服务器组中选择另外的服务器发送报文。

4.4.3 HWTACACS 协议简介

HWTACACS 是在 TACACS（RFC 1492）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 Server/Client 模式实现多种用户的认证与授权功能，可用于 PPP 和 VPDN 接入用户及 login 用户的认证、授权和计费。

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 4-2 所示。

表4-2 HWTACACS 协议与 RADIUS 协议的比较

HWTACACS	RADIUS
使用 TCP 协议，网络传输更可靠	使用 UDP 协议
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
支持对配置命令进行授权使用	不支持

HWTACACS 消息流程

HWTACACS 协议的消息流程和 RADIUS 协议的消息流程类似。不同在于 HWTACACS 协议当用户认证通过之后，服务器返回的是认证回应，而不返回用户的权限，只有当授权流程完成后才会返回用户的权限。

HWTACACS 协议支持按命令行授权

用户通过 Console 口、Telnet 或者 SSH 登录到 USG9500 上后，如果该用户需要进行命令行授权，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，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，USG9500 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使用于因 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。下面的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 协议支持对用户级别提升进行认证

用户通过 Console 口、Telnet 或者 SSH 登录到 USG9500 后，可以通过在用户模式下使用 **super** 命令来提升或降低自己的级别。这时，USG9500 对用户的密码进行验证。

USG9500 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。用户级别更改的结果只影响本次登录。

如果在用户配置的超时时间内，USG9500 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
使用 USG9500 对用户级别提升进行验证的时候，各级别的密码可以不同；使用 HWTACACS 服务器对用户级别进行提升进行验证时，各级别密码必须相同。

4.4.4 域简介

USG9500 对用户的管理包括两个层次：

- 通过域进行管理
- 通过用户账号进行管理

所有用户都属于某个域。

域下可以进行缺省授权配置、RADIUS 模板配置、认证方案的配置等。

域下配置的授权信息较认证与授权服务器的授权信息优先级低。即，优先使用认证与授权服务器下发的授权属性，在服务器无该项授权或不支持该项授权时，域的授权属性生效。这样处理的优点是：可以凭借域管理灵活增加业务，而不必受限于服务器提供的属性。

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。

4.4.5 本地用户管理简介

在 USG9500 上，可以使用认证与授权建立本地用户数据库，维护用户信息，并对用户进行管理。除了可以建立本地用户账号外，还可以进行本地认证。

说明
用户信息在本地用户数据库上的用户称为本地用户。

在 USG9500 目前的实现中，可以单独配置本地用户。

5 NAT

关于本章

5.1 NAT 简介

NAT 是将 IP 数据报报头中的 IP 地址转换为另一个 IP 地址的过程，主要用于实现内部网络（私有 IP 地址）访问外部网络（公有 IP 地址）的功能。

在实际应用中，内部网络一般使用私有地址。RFC（Request For Comments）1918 为私有、内部的使用留出了三个 IP 地址块。具体如下：

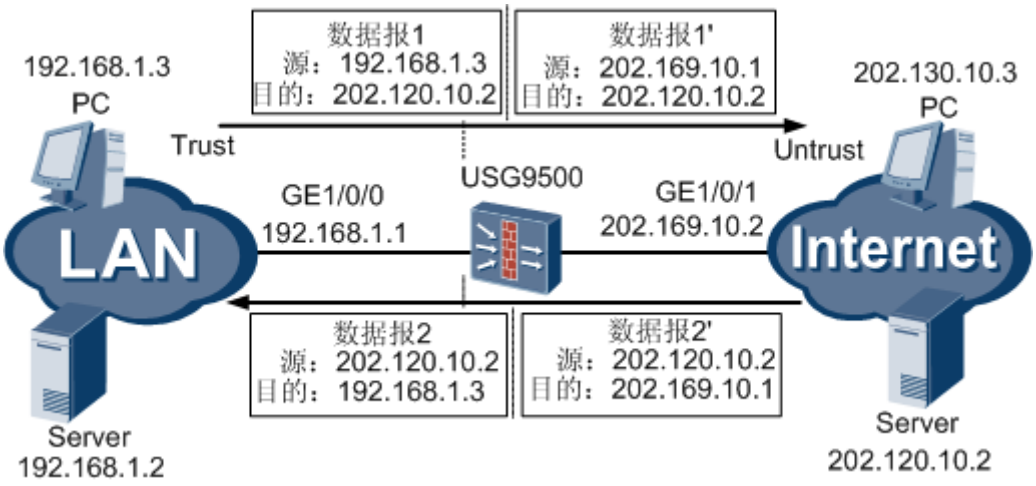
- A 类 10.0.0.0～10.255.255.255（10.0.0.0/8）
- B 类 172.16.0.0～172.31.255.255（172.16.0.0/12）
- C 类 192.168.0.0～192.168.255.255（192.168.0.0/16）

上述三个范围内的地址不会在 Internet 上被分配，因而可以不必向 ISP（Internet Service Provider）或注册中心申请而在公司或企业内部自由使用。

NAT 主要用于实现内部网络访问外部网络的功能。通过应用 NAT，能够使多数的私有 IP 地址转换为少数的公有 IP 地址，减缓可用 IP 地址空间枯竭的速度。

图 5-1 描述了一个基本的 NAT 应用。

图5-1 地址转换的基本过程



NAT 设备（如 USG9500）处于私有网络和公有网络的连接处。内部 PC 与外部服务器的交互报文全部通过该 NAT 设备。地址转换的过程如下。

1. 内部 PC（192.168.1.3）发往外部服务器（202.120.10.2）的数据报 1 到达 NAT 设备后，NAT 设备查看报头内容，发现该数据报头中的信息匹配上了某条 NAT 策略。
2. NAT 设备将数据报 1 的源地址字段的私有地址 192.168.1.3 换成一个可在 Internet 上选路的公有地址 202.169.10.1，发送到外部服务器，同时在网络地址转换表中记录这一地址转换映射。
3. 外部服务器收到数据报 1’后，向内部 PC 发送应答报文，即数据报 2’，初始目的地址为 202.169.10.1。
4. 数据报 2’到达 NAT 设备后，NAT 设备查看报头内容，查找当前网络地址转换表的记录，用私有地址 192.168.1.3 替换目的地址，发送给内部 PC。

上述的 NAT 过程对 PC 和外部服务器来说是透明的。内部 PC 认为与外部服务器的交互报文没有经过 NAT 设备的干涉；外部服务器认为内部 PC 的 IP 地址就是 202.169.10.1，并不知道存在 192.168.1.3 这个地址。

USG9500 支持的 NAT 功能包括对源 IP 地址进行转换，和对目的 IP 地址进行转换两种方式。

其中，基于源 IP 地址的转换可以从以下两个方面进行划分：

- 转换的方向。
- 端口是否转换。

基于源 IP 地址的转换还包括三元组 NAT（Full-cone NAT），具体介绍请参见 5.5 三元组 NAT。

按照转换的方向可以将源 IP 地址转换划分为以下两类：

- Inbound 方向

  数据包由低安全级别的安全区域向高安全级别的安全区域方向传输时，基于源 IP 地址进行的转换。

- Outbound 方向

  数据包由高安全级别的安全区域向低安全级别的安全区域方向传输时，基于源 IP 地址进行的转换。

按照端口是否转换可以将源 IP 地址转换划分为以下两类：

- No-PAT (Port Address Translation)方式的 NAT

  主要用于一对一的 IP 地址的转换，端口不进行转换。

- NAPT(Network Address Port Translation)方式的 NAT

  主要用于多对一或多对多的地址转换，转换时地址和端口号同时进行转换。

按照功能不同，可以将基于目的 IP 地址的转换分为以下两类：

- NAT Server

  主要应用于实现私网服务器以公网 IP 地址对外提供服务的功能。

- 目的 NAT

  主要应用于实现手机用户上网时，需要修改目的网关地址的功能。

5.2 NAT 地址池/NAT 地址池组及转换控制

NAT 地址池是一些连续的 IP 地址集合，当来自私网的报文通过地址转换到公网 IP 时，将会选择地址池中的某个地址作为转换后的地址。

NAT 地址池中的地址可以是一个公网 IP 地址，也可以是多个公网 IP 地址。USG9500 的一个地址池中最多可以包含 4096 个地址。

USG9500 还支持将多个地址池合并到一个地址池组中，可以灵活的规划公网 IP 地址的数据。在配置域间 NAT 或域内 NAT 时使用地址池组，可以简化配置过程，减少配置的工作量。

在配置域间 NAT 或域内 NAT 时，需要首先配置 NAT 地址池/NAT 地址池组，然后将 NAT 地址池/NAT 地址池组与 ACL 绑定，通过选择不同的参数，实现不同功能的 NAT。

在实际应用中，用户可能希望其内部网络中某些主机具有访问 Internet 的权利，而某些主机没有。即当 NAT 进程查看数据报报头内容时，如果发现源 IP 地址是为那些不允许访问外部网络的内部主机所拥有的，将不进行 NAT 转换。这就是一个对地址转换进行控制的问题。

将一个地址池/NAT 地址池组和一个 ACL 规则关联起来，即指定了“具有某些特征的 IP 报文”才可以使用“这个地址池/地址池组中的地址”。当报文到达 USG9500 时，USG9500 首先根据访问列表判定是否是允许的数据包，然后根据转换关联找到与之对应的地址池/地址池组，这样就把一个地址转换成这个地址池/地址池组中的另一个地址，完成地址转换过程。

5.3 NAT No-PAT

NAT No-PAT 也可以称为“一对一地址转换”，在地址转换过程中，数据包的源 IP 地址由私网地址转换为公网地址，但端口号不做转换。

例如，地址池中的公网 IP 地址只有两个，当所有私网的主机访问公网时，只能拥有两个公网 IP 地址，因此，这种情况只允许最多有两台私网主机同时访问公网，其他的私网主机要等到公网 IP 地址被释放后，才可以再做地址转换访问公网。

当配置了 No-PAT 方式的 NAT 后，USG9500 会为有流量的私网 IP 地址分配一个公网地址，同时建立 Server-map 表。Server-map 表用于存放私网 IP 地址与公网 IP 地址的映射关系。后续从该私网 IP 发出的所有报文，都将命中 Server-map 表转换成该公网地址，这种地址转换关系是一一对应的。

5.4 NAPT

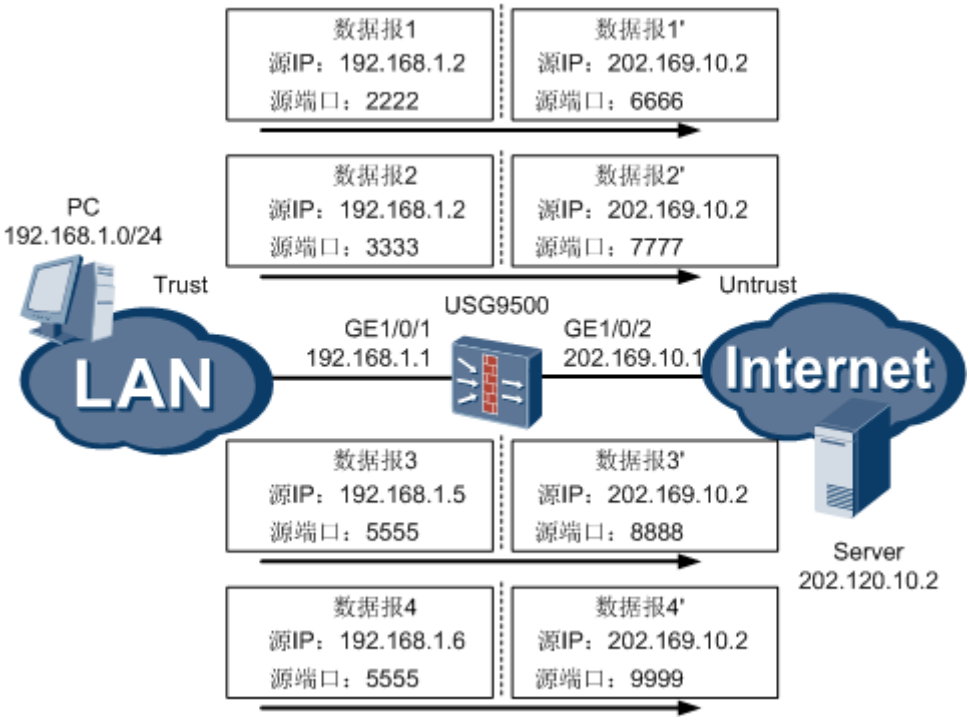
NAT No-PAT 功能可以实现 USG9500 一对一的地址转换，但是在实际应用中，在公网 IP 地址非常有限的情况下，如果大量的私网用户需要同时访问 Internet，则 NAT No-PAT 功能无法满足用户需求。

USG9500 的 NAPT 功能可以解决多个私网 IP 地址同时映射为少个或一个公网 IP 地址的问题。

NAPT 也可以称之为“地址复用”。通过配置 NAPT 功能，USG9500 同时对端口号和 IP 地址进行映射，允许多个私网 IP 地址同时映射到同一个公网 IP 地址，相同的公网 IP 地址通过不同的端口号区分映射不同的私网 IP 地址，从而实现多对一或多对多的地址转换。

下面对 NAPT 转换原理进行说明。

图5-2 NAPT 转换原理示意图



如图 5-2 所示，四个带有私网 IP 地址的数据报到达 USG9500，其中数据报 1 和 2 来自同一个私网 IP 地址但有不同的源端口号，数据报 3 和 4 来自不同的私网 IP 地址但具有相同的源端口号。通过 NAT 转换，四个数据报都被转换到同一个公网 IP 地址，但每个数据报都赋予了不同的源端口号，因而仍保留了报文之间的区别。当回应报文到达 USG9500 时，NAT 进程仍能够根据回应报文的目的地址和端口号来区别该报文应转发到的相应的内部主机。

5.5 三元组 NAT

NAT 可以解决 IP 地址短缺的问题，也可以隐藏内部主机的 IP 地址，使内部主机的安全性得到提高。但是 NAT 设备建立起来的非对称体系与目前广泛应用于文件共享、语音通信、视频传输等方面的 P2P 技术不能很好的共存，位于内部网络的主机没有永久可用的公网 IP 地址和端口供 Internet 上的其他用户访问。

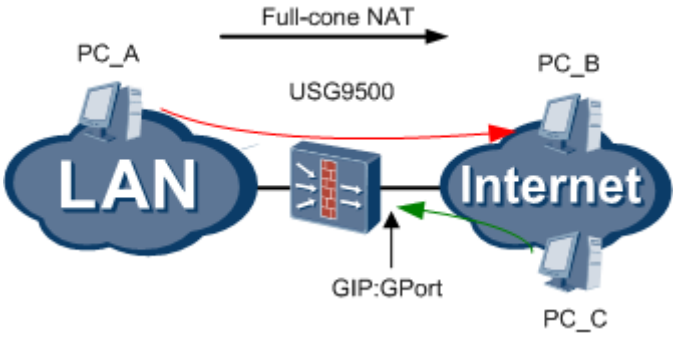
为了解决上述问题，USG9500 支持 Full-cone NAT，也称之为三元组 NAT。USG9500 会为匹配 Full-cone NAT 的报文生成源 Server-map 表和目的 Server-map 表：

- 源 Server-map 表

  表项中的元素包括源地址、源端口和协议，目的是保证内部网络中相同的源地址、源端口和协议发送至 Internet 的报文能够转换成相同的地址和端口。
- 目的 Server-map 表

  表项中的元素包括地址转换后的源地址、源端口和协议，目的是保证 Internet 中的任何主机都能通过此地址、端口和协议访问内部网络中对应的主机。

图5-3 三元组 NAT 示意图



如图 5-3 所示，内部网络中的主机 PC_A 访问 Internet 上的主机 PC_B 时，在 USG9500 上会生成源 Server-map 表和目的 Server-map 表。当 Internet 上的主机 PC_C 访问 PC_A 的公网地址（GIP:GPort）时，将会匹配目的 Server-map 表。USG9500 根据目的 Server-map 表中的转换关系，将 PC_A 的公网地址转换为 PC_A 的私网地址，使 PC_C 主动访问 PC_A 的报文可以穿越 NAT 设备。

此外，NAT64、DS-Lite 功能也支持三元组 NAT，关于 NAT64 和 DS-Lite 功能的具体介绍请参见《*配置指南 IPv6 配置*》中的过渡技术部分。

5.6 NAT Server

NAT 隐藏了内部网络的结构，具有“屏蔽”内部主机的作用。但是在实际应用中，可能需要提供给外部一个访问内部主机的机会，如提供给外部一台 Web 服务器，或是一台 FTP 服务器。使用 NAT 可以灵活地添加 NAT Server。USG9500 提供两种方式为 NAT Server 指定外部地址：

- 可以使用 202.169.10.10 作为 Web 服务器的外部地址。
- 可以使用 202.110.10.12:8080 作为 Web 服务器的外部地址。

USG9500 的 NAT 能够为外部网络用户提供访问的 NAT Server。外部用户访问 NAT Server 时，有如下两部分操作：

- USG9500 将外部用户的请求报文的目的地址转换成 NAT Server 的私有地址。
- USG9500 将 NAT Server 的回应报文的源地址（私网地址）转换成公网地址。

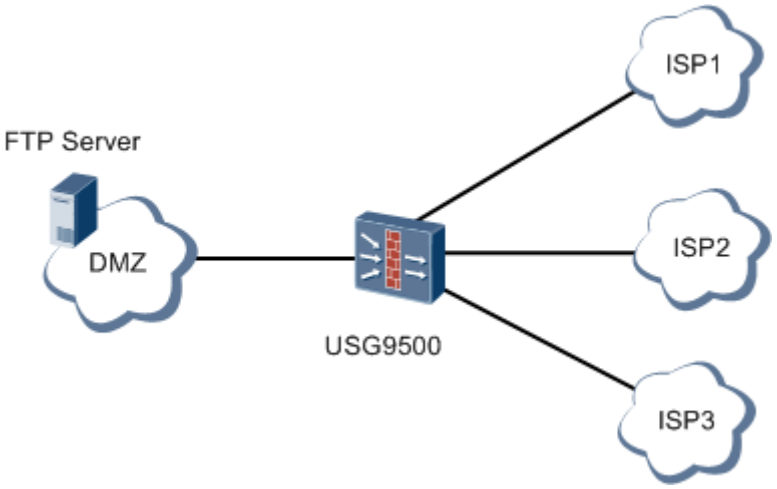
USG9500 支持为外部用户提供多台同样的服务器，例如，提供多台 Web 服务器。

USG9500 还支持为 NAT Server 配置描述信息，建议用户使用有意义的名字标识 NAT Server，便于记忆和管理。

说明
允许外部用户访问的 NAT Server 通常置于 USG9500 的 DMZ 区。正常情况下不允许这个区域中的设备主动向外发起连接。

在实际应用中，经常会出现同一个内部服务器提供给多个外部网络访问的情况，即 NAT Server 多出口场景。如图 5-4 所示，USG9500 连接多个外部网络，每一个外部网络都使用各自的 Global 地址访问内部服务器。

图5-4 NAT Server 多出口示意图



USG9500 支持将外部网络划分到不同的安全区域，为每一个外部网络配置一条 NAT Server，同时在配置时指定 **zone** 参数来实现 NAT Server 多出口。

需要注意的是，上述方式可能会产生报文的来回路径不一致的问题。以图 5-4 为例，如果 Internet 中的用户通过 USG9500 发布给 ISP1 的 Global 地址访问内部服务器 FTP Server，FTP Server 响应的报文到达 USG9500 后，根据目的地址查找路由表，响应报文可能会由 ISP2 发送出去，导致业务中断。

为了避免上述问题，可以在报文进入的接口上配置 **redirect-reverse** 命令，当 USG9500 转发响应报文时，直接使用入接口作为响应报文的出接口，而不是通过查找路由表来确定出接口，从而实现报文从同一接口进入和发出。

5.7 目的 NAT

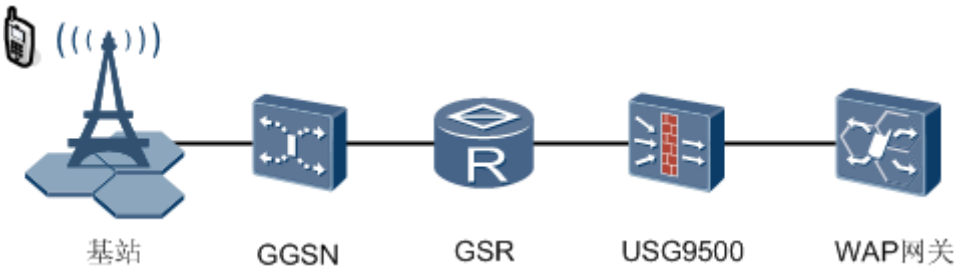
手机用户需要通过登录 WAP（Wireless Application Protocol）网关来实现上网的功能。目前，大量用户直接从国外购买手机使用，这些手机出厂时，缺省设置的 WAP 网关地址与本国 WAP 网关地址不符，且无法自行修改，从而导致用户不能移动上网。

为解决这一问题，无线网络中，在 WAP 网关与用户之间部署 USG9500。通过在 USG9500 上配置目的 NAT 功能，使这部分手机用户能够正常获取网络资源。

如图 5-5 所示，当手机用户上网时，目的 NAT 处理过程如下：

1. 当手机用户上网时，请求报文经过基站及其他中间设备到达 USG9500。
2. 到达 USG9500 的报文如果匹配 USG9500 上所配置的目的 NAT 策略，则将此数据报文的目的 IP 地址转换为已配置好的 WAP 网关的 IP 地址，并送往 WAP 网关。
3. WAP 网关对手机客户端提供相应的业务服务（如视频服务、网页服务等），并将回应报文发往 USG9500。
4. 回应报文在 USG9500 上命中会话，USG9500 转换该报文的源 IP 地址，并将该报文发往手机用户，完成一次通信。

图5-5 手机用户上网目的 NAT 组网图



5.8 域内 NAT

典型应用一

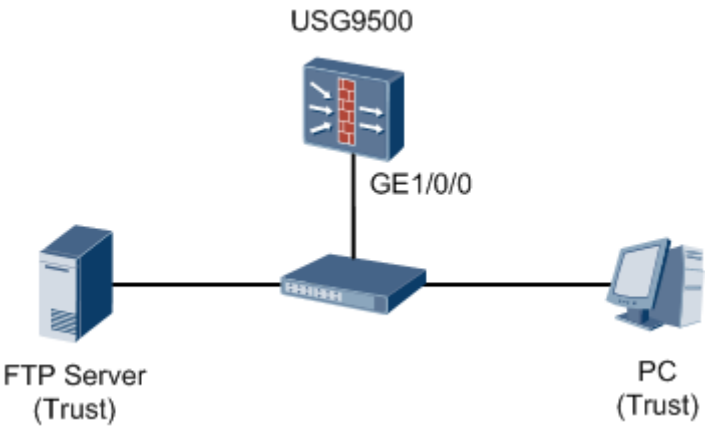
在配置 NAT Server 过程中，可能会遇到这样的情况，当用户和内部服务器处于同一个安全区域，并且 IP 地址在同一网段时，当用户访问服务器时，请求报文会不经过 USG9500 而直接到达内部服务器。

如图 5-6 所示，FTP 服务器和 PC 均在防火墙的 Trust 安全区域，二者通过交换机与防火墙同一个接口相连。FTP 服务器通过公网地址对用户提供 FTP 服务，如果 PC 和 FTP 服务器都具有内网地址，且位于同一个网段，那么当 PC 访问 FTP 服务器时，请求报文会直接到达 FTP 服务器，而不经过 USG9500。

为避免这一情况，保证统一安全区域的用户访问 FTP 服务器的报文也要经过 USG9500，那么，就需要在 USG9500 上配置域内 NAT 功能。

当 PC 访问 FTP 服务器的公网地址时，PC 的地址也进行地址转换，由私网地址转换为公网地址，以保证 PC 和 FTP 服务器交互的所有报文都经过 USG9500，同时保证 USG9500 对报文正确处理。

图5-6 内部服务器供域内用户访问典型组网图

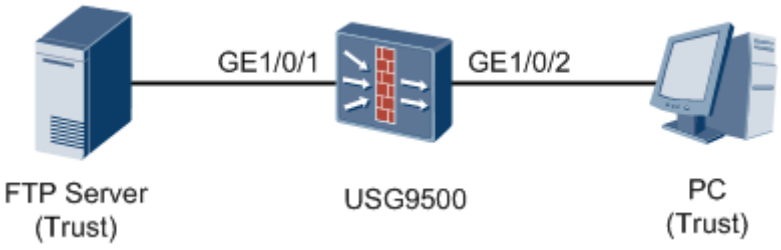


典型应用二

如图 5-7 所示，FTP 服务器和 PC 均在防火墙的 Trust 安全区域，二者分别与防火墙的不同接口相连。PC 访问 FTP 服务器时，私网地址需要转换为公网地址。

通过配置域内 NAT 功能，可以实现由 PC 发出的报文源 IP 地址私网地址转换为公网地址，从而实现同一安全区域内的地址转换。

图5-7 内部服务器供域内用户访问典型组网图二



5.9 双向 NAT

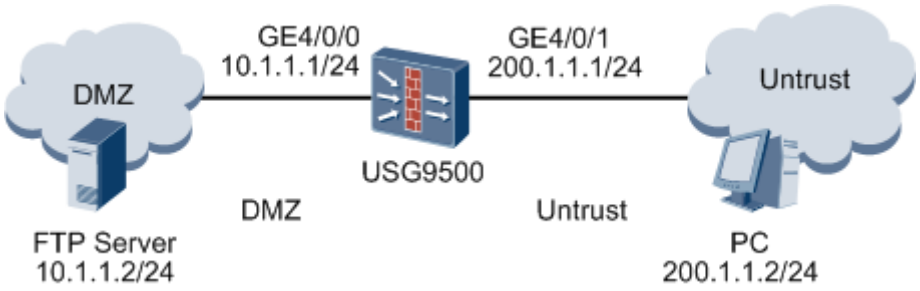
当既要对报文的源 IP 地址做 NAT，又要对此报文的目的 IP 地址做 NAT 时，则需要配置双向 NAT。

双向 NAT 应用环境如下：

- 低优先级区域的用户访问公网地址时将报文的目的地址转换为 NAT Server 服务器的私网地址，但 NAT Server 服务器需要配置到该公网地址的路由。也可以配置从低优先级区域到高优先级区域方向的 NAT，即 inbound 方向的 NAT，实现低优先级区域的用户访问公网地址。此方法可以简化配置，避免配置到公网地址的路由。
- 同一个安全区域内的访问需要作 NAT，则需要配置域内 NAT 和 NAT Server 功能。

如图 5-8 所示，在 USG9500 上配置从低优先级区域到高优先级区域方向的 NAT。以配置从 Untrust 区域到 DMZ 区域方向的 NAT 为例，进行说明。

图5-8 从低优先级区域到高优先级区域的 NAT 组网图



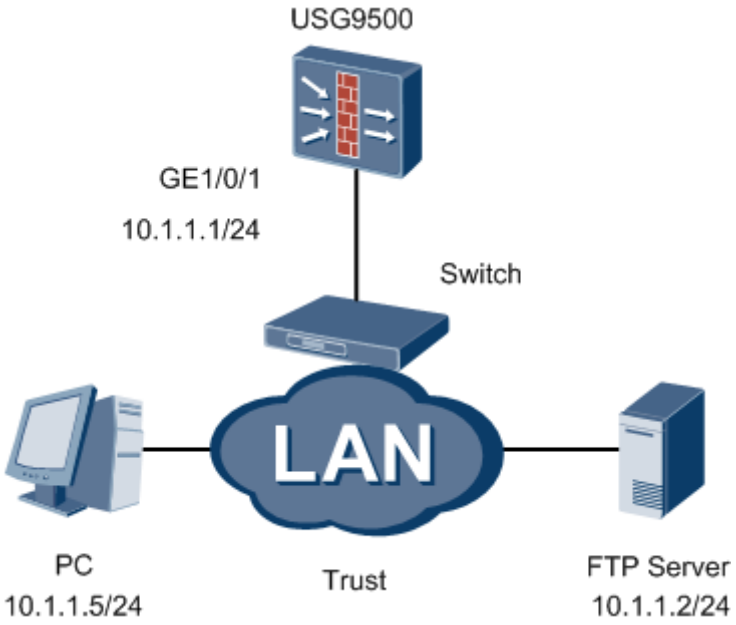
当 Untrust 区域的用户访问 DMZ 的服务器时，有如下两部分操作：

- USG9500 将外部用户的请求报文的目的地址转换成内部服务器的私有地址，源地址转换成地址池中的地址（私网地址）。
- USG9500 将内部服务器的回应报文的源地址（私网地址）转换成公网地址，目的地址（私网地址）转换为公网地址。

说明
允许外部用户访问的内部服务器通常置于 USG9500 的 DMZ 区域。正常情况下不允许这个区域中的设备主动向外发起连接。

如图 5-9 所示，在 USG9500 上配置同一个区域内的 NAT。以配置 Trust 域内的 NAT 为例，进行说明。

图5-9 域内 NAT 组网图



当 Trust 域的用户访问 Trust 域的服务器时，有如下两部分操作：

- USG9500 将内部用户的请求报文的目的地址转换成内部服务器的私有地址，源地址转换成地址池中的地址（公网地址）。
- USG9500 将内部服务器的回应报文的源地址（私网地址）转换成公网地址，目的地址（公网地址）转换为私网地址。

5.10 NAT ALG

注意
仅当配置域内 NAT 时，NAT ALG 功能需要在安全区域视图下配置，即域内视图下配置。

通常 NAT 只能对 IP 报文的头部地址和 TCP/UDP 头部的端口信息进行转换，但是对于一些特殊的协议，比如 FTP 等协议，则需要由数据连接和控制连接共同完成，而且数据连接的建立要由控制连接载荷字段中的报文信息动态的决定，这就需要能够根据控制连接的载荷字段中的报文解析出数据连接要使用的地址和端口号，也就是说 USG9500 必须能够辨识 FTP 应用载荷字段中包含的端口号和地址信息，才能进行有效的 NAT 处理，否则可能导致 NAT 功能失败。

为解决这一问题，当 NAT 功能与 FTP、MSN、PPTP、QQ、RTSP、TFTP 等协议共同使用时，需要在 USG9500 上配置 NAT ALG（Application Level Gateway）功能，通过配置 NAT ALG 功能，USG9500 对数据包进行深度解析，并改变封装在 IP 报文数据部分中的 IP 地址和端口号信息，从而实现 NAT 功能。

域间 NAT ALG 支持的协议包括：DNS、FTP、H323、MGCP、MMS、MSN、PPTP、QQ、RTSP、SCCP、SIP、SQLNET、ILS、NETBIOS、RSH、user-defined 以及 IPv6 FTP、IPv6 RTSP。

域内 NAT ALG 支持的协议包括：FTP 和 RTSP。

6 VPN

关于本章

6.1 概述

6.1.1 VPN 简介

VPN（Virtual Private Network）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，用于实现在公用网络上构建私人专用网络。“虚拟”主要指这种网络是一种逻辑上的网络。

伴随企业和公司的不断扩张，员工出差日趋频繁，驻外机构及客户群分布日益分散，合作伙伴日益增多，越来越多的现代企业迫切需要利用公共 Internet 资源来进行促销、销售、售后服务、培训、合作及其它咨询活动，这为 VPN 的应用奠定了广阔市场。

VPN 的特点

VPN 的特点如下：

- VPN 有别于传统网络，它并不实际存在，而是利用现有公共网络，通过资源配置而成的虚拟网络，是一种逻辑上的网络。
- VPN 只为特定的企业或用户群体所专用。

  从 VPN 用户角度看来，使用 VPN 与传统专网没有区别：
  - VPN 作为私有专网，与底层承载网络之间保持资源独立性，即在一般情况下，VPN 资源不会被承载网络中的其它 VPN 或非该 VPN 用户的网络成员所使用。
  - VPN 提供足够安全性，确保 VPN 内部信息不受外部的侵扰。
- VPN 不是一种简单的高层业务。

该业务建立专网用户之间的网络互联，包括建立 VPN 内部的网络拓扑、路由计算、成员的加入与退出等，因此 VPN 技术比各种普通的点对点的应用机制要复杂得多。

VPN 的优势

VPN 的优势如下：

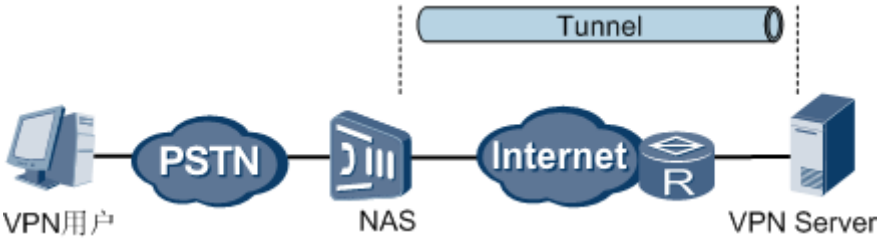
- 在远端用户、驻外机构、合作伙伴、供应商与公司总部之间建立可靠的安全连接，保证数据传输的安全性。

  这一优势对于实现电子商务或金融网络与通讯网络的融合将有特别重要的意义。
- 利用公共网络进行信息通讯，一方面使企业以明显更低的成本连接远地办事机构、出差人员和业务伙伴，另一方面极大的提高了网络的资源利用率，有助于增加 ISP 的收益。
- 只需要通过软件配置就可以增加、删除 VPN 用户，无需改动硬件设施。这使得 VPN 的应用具有很大灵活性。
- 支持驻外 VPN 用户在任何时间、任何地点的移动接入，这将满足不断增长的移动业务需求。
- 构建具有服务质量保证的 VPN，可为 VPN 用户提供不同等级的服务质量保证，通过收取不同的业务使用费用可获得更多的利润。

6.1.2 VPN 原理和实现

VPN 的原理如图 6-1 所示。

图6-1 VPN 接入示意图



VPN 用户通过 PSTN/ISDN 或局域网拨入 ISP 的 NAS，NAS 通过用户名或接入号码识别出该用户为 VPN 用户后，就和用户的目的 VPN 服务器建立一条连接，称为隧道（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 服务器收到数据包并拆封后就可以读到真正有意义的报文了。反向的处理也一样。对用户来说，隧道是其 PSTN/ISDN 链路的逻辑延伸，操作起来和实际物理链路相同。

隧道可以通过隧道协议来实现。根据是在 OSI（Open Systems Interconnection）模型的第二层还是第三层实现隧道，隧道协议分为第二层隧道协议和第三层隧道协议，其说明如下：

- 第二层隧道协议

  第二层隧道协议是将整个 PPP 帧封装在内部隧道中。现有的第二层隧道协议有三种。

- PPTP（Point-to-Point Tunneling Protocol）

  PPTP 在 Windows NT 4.0 以上版本中支持。该协议支持 PPP 在 IP 网络上的隧道封装，PPTP 作为一个呼叫控制和管理协议，使用一种增强的 GRE（Generic Routing Encapsulation）技术为传输的 PPP 报文提供流控和拥塞控制的封装服务。

- L2F（Layer Two Forwarding）协议

  L2F 协议支持对更高级协议链路层的隧道封装，实现了拨号服务器和拨号协议连接在物理位置上的分离。

- L2TP（Layer 2 Tunneling Protocol）

  L2TP 结合了上述两个协议的优点，为众多公司所接受。并且已经成为标准 RFC。L2TP 既可用于实现拨号 VPN 业务（VPDN 接入），也可用于实现 VPN 业务。

- 第三层隧道协议

  第三层隧道协议的起点与终点均在 ISP 内，PPP 会话终止在 NAS 处，隧道内只携带第三层报文。现有的第三层隧道协议主要有两种。

  - GRE 协议

    GRE 用于实现任意一种网络层协议在另一种网络层协议上的封装。

  - IPSec 协议

    IPSec 协议不是一个单独的协议，它给出了 IP 网络上数据安全的一整套体系结构，包括 AH（Authentication Header）、ESP（Encapsulating Security Payload）、IKE（Internet Key Exchange）等协议。

  GRE 和 IPSec 主要用于实现 VPN 业务。

- 第二、三层隧道协议之间的异同

  第三层隧道与第二层隧道相比，优势在于它的安全性、可扩展性与可靠性。

  - 从安全性的角度看，由于第二层隧道一般终止在用户侧设备上，对用户网的安全及防火墙技术提出十分严峻的挑战；而第三层隧道一般终止在 ISP 网关上，因此一般情况下不会对用户网的安全技术提出较高要求。
  - 从扩展性的角度看，第二层隧道内封装了整个 PPP 帧，这可能产生传输效率问题。其次，PPP 会话贯穿整个隧道并终止在用户侧设备上，导致用户侧网关必须要保存大量 PPP 会话状态与信息，这将对系统负荷产生较大的影响，也会影响到系统的扩展性。此外，由于 PPP 的 LCP（Link Control Protocol）及 NCP（Network Control Protocol）协商都对时间非常敏感，这样隧道的效率降低会造成 PPP 对话超时等一系列问题。相反，第三层隧道终止在 ISP 的网关内，PPP 会话终止在 NAS 处，用户侧网关无需管理和维护每个 PPP 对话的状态，从而减轻了系统负荷。
  - 一般地，第二层隧道协议和第三层隧道协议都是独立使用的，如果合理地将这两层协议结合起来使用，将可能为用户提供更好的安全性（如将 L2TP 和 IPSec 协议配合使用）和更佳的性能。

6.2 L2TP

6.2.1 介绍

L2TP 是目前使用最广泛的 VPDN（Virtual Private Dial Network）隧道协议。L2TP 功能可以简单描述为在非点对点的网络上建立点对点的 PPP 会话连接，主要用于实现企业驻外机构或出差人员经由公共网络，通过虚拟加密隧道实现和企业总部之间的网络连接。

定义

为了更好的说明 L2TP，需要首先了解 VPDN 和 PPP 的概念。

VPDN 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 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。

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（Network Access Server）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 功能可以简单描述为在非点对点的网络上建立点对点的 PPP 会话连接。L2TP 协议由 IETF（Internet Engineering Task Force）起草，微软等公司参与，结合了 PPTP 和 L2F 两个协议的优点，成为 IETF 有关二层隧道协议的工业标准。

目的

L2TP 协议具有以下优势：

- 灵活的身份验证机制以及高度的安全性
  - L2TP 本身并不保证连接的安全性，但它可利用 PPP 提供的认证机制（如 CHAP、PAP），因此具有 PPP 的所有安全特性。
  - 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安全性。例如，L2TP 通常与 IPSec 结合起来实现数据安全，这使得通过 L2TP 所传输的数据更难被攻击。
- 支持 RADIUS 服务器的验证

LAC（L2TP Access Concentrator，L2TP 访问集中器）端支持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（L2TP Network Server，L2TP 网络服务器）可以对远端用户的地址进行动态的分配和管理，可支持私有地址应用（RFC1918，Address Allocation for Private Internets）。为远端用户所分配的地址不是 Internet 地址而是企业内部的私有地址，这样方便了地址的管理并可以增加安全性。

6.2.2 参考标准和协议

通过学习 L2TP 相关的参考标准和协议可以加深对 L2TP 特性的了解。

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

6.2.3 可获得性

介绍 L2TP 特性与 License 的关系以及哪些版本支持 L2TP 特性。L2TP 特性不需要 License 支持。

License 支持

本特性无须 License 支持。

版本支持

产品	支持版本
HUAWEI Secoway USG9500	V100R001、V100R003、V200R001

6.2.4 特性增强

介绍 L2TP 特性在不同版本的变化，包括增加、修改和删减等情况。

版本	特性增强
V200R001	增加“L2TP 多实例”功能，即增加虚拟防火墙下的 L2TP 功能。 修改“缺省 L2TP 组”。V100R003 中缺省 L2TP 组为“l2tp-group 1”，修改后缺省 L2TP 组为“default-lns”。

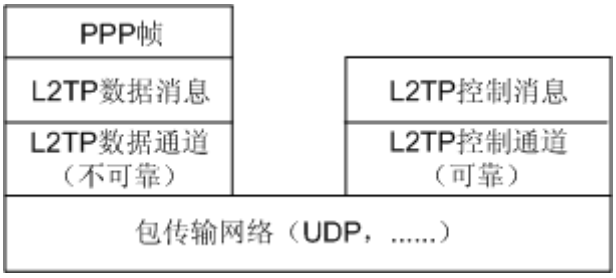
6.2.5 原理描述

主要介绍 L2TP 协议的结构，L2TP 隧道的两种发起模式，L2TP 隧道及会话的建立过程。

6.2.5.1 L2TP 协议结构

通过了解 L2TP 协议结构可以更好的了解 L2TP 的数据和控制消息如何在 L2TP 隧道和会话中传输。

图6-2 L2TP 协议结构



如图 6-2 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 隧道发起方任选一个空闲端口（未必是 1701）向接收方的 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

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，该会话将被清除。

控制消息和数据消息

L2TP 中存在控制消息和数据消息，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个隧道上，报文头中的隧道标识符与会话标识符由对端分配。

6.2.5.2 L2TP 隧道发起模式

介绍使用 L2TP 协议构建 VPDN 的两种典型的隧道模式，并对 LAC、LNS 和用户的概念进行说明。

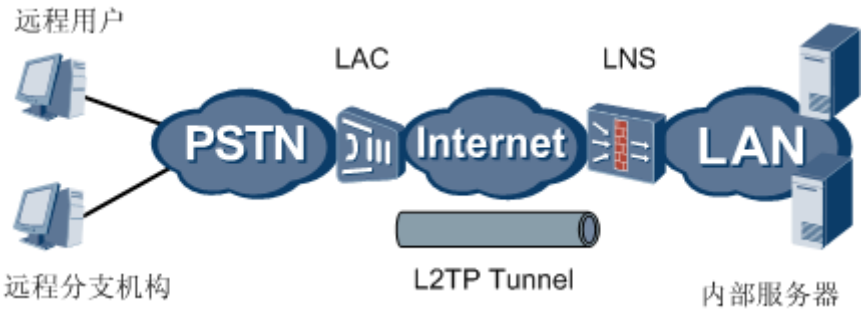
使用 L2TP 协议构建 VPDN 有两种典型的隧道发起模式：

- NAS-Initialized

  远端系统通过 PPPoE 拨入 LAC，由 LAC（NAS）通过 Internet 向 LNS 发起建立隧道连接请求。由 LNS 为拨号用户分配私有 IP 地址，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 完成。

  典型组网如图 6-3 所示。

图6-3 配置 NAS-Initialized VPN 组网图

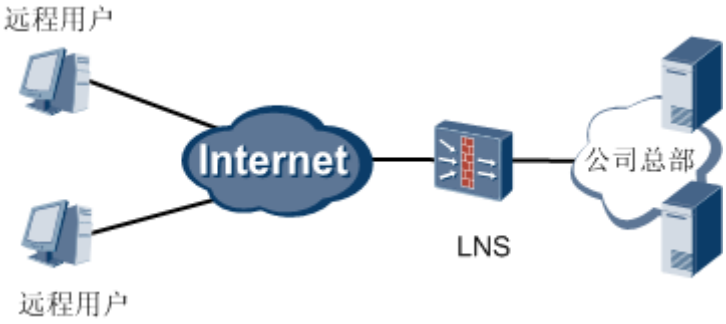


- Client-Initialized

  LAC 客户可直接向 LNS 发起隧道连接请求，无需再经过一个单独的 LAC 设备。在 LNS 设备收到了 LAC 客户的请求之后，根据用户名、密码进行验证，并且在验证通过后为 LAC 客户分配私有 IP 地址。

  典型组网如图 6-4 所示。

图6-4 配置 Client-Initialized VPN 组网图



基本概念

对 LAC、LNS 和用户的说明如下：

- LAC 位于 LNS 和远端系统（远程用户和远程分支机构）之间。LAC 是交换网络上具有 PPP 端系统和 L2TP 处理能力的设备，一般是本地 ISP 的接入设备，如网络接入服务器 NAS，通过 PSTN/ISDN 网络为用户提供接入服务。

  LAC 在 LNS 和远端系统之间传递数据：把从远端系统收到的数据进行 L2TP 封装并送往 LNS；将从 LNS 收到的数据进行解封装并送往远端系统。

  LAC 与远端系统间可采用本地连接或 PPP 链路，VPDN 应用中通常使用 PPP 链路。LAC 是直接接受用户呼叫的一端，也是 PPP 二层链路一端。NAS 可以和用户合并为一个 LAC 端点，也可以单独作为 LAC 端点。
- LNS 是接受 PPP 会话的一端，通过 LNS 验证，用户就可以登录到私网上，访问私网资源。同时，LNS 作为 L2TP 隧道的另一侧端点，是 LAC 的对端设备，是通过 LAC 进行隧道传输的 PPP 会话的逻辑终止端点。

  LNS 位于私网与公网边界，通常是企业网关设备，实施网络接入功能及 LNS 功能。必要时，LNS 还兼有网络地址转换（NAT）功能，对企业总部网络内的专用 IP 地址与 IP 网公用 IP 地址进行转换。LNS 可以放在企业总部网络内，也可以是 IP 公共网络的 PE（Provider Edge）。
- 用户

  L2TP 组网模型中，用户是需要登录私网的设备（如 PC）。VPDN 用户的特征是接入的方式和地点不固定。用户可以通过 PSTN 或 ISDN 网络与 LAC 连接，或者接入 Internet 直接与总部服务器建立连接。

  用户是发起 PPP 协商的终端设备，既是 PPP 二层链路一端又是 PPP 会话的一端。

  从 LNS 和 LAC 角度看，有三类用户：
  - 使用 VPN 服务号码的用户

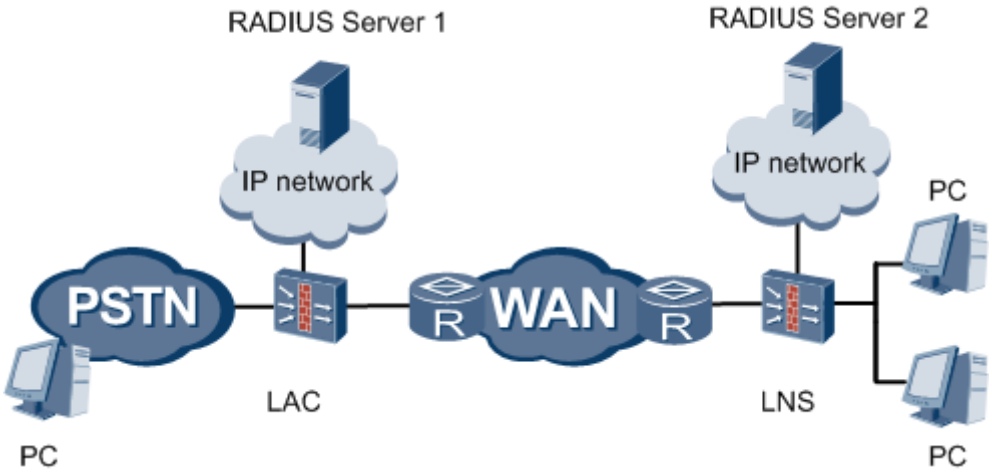
    每个申请 VPN 服务的公司都分别配置一个 VPN 号码，缺点是浪费号码资源。
  - 域名接入用户
  - 全名接入用户

6.2.5.3 L2TP 隧道及会话的建立过程

以 NAS-Initialized 隧道发起模式为例讲解 L2TP 隧道会话的建立过程。

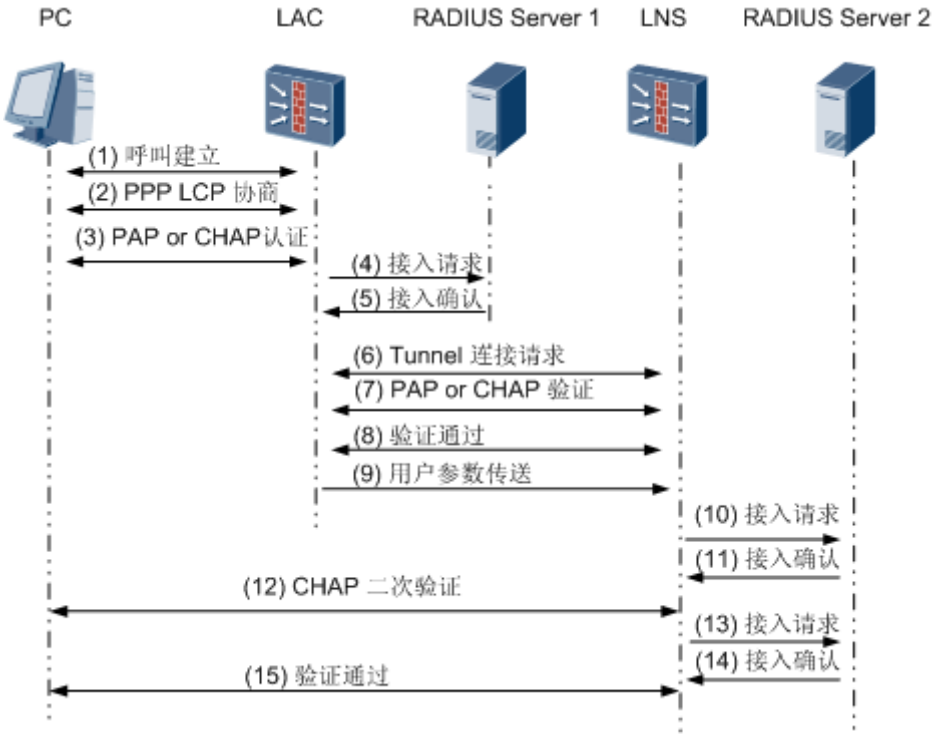
NAS-Initialized VPN 典型组网如图 6-5 所示。下面以 NAS-Initialized 隧道发起模式为例讲解 L2TP 隧道会话的建立过程。对于 Client-Initialized 隧道发起模式中 PC 作为 LAC 客户端直接向 LNS 发起隧道连接。

图6-5 NAS-Initialized VPN 的典型组网示意图



L2TP 隧道的呼叫建立流程如图 6-6 所示。

图6-6 L2TP 隧道的呼叫建立流程



L2TP 隧道的呼叫建立流程过程如下。

1. 用户端 PC 发起呼叫连接请求。
2. PC 和 LAC 端进行 PPP LCP 协商。
3. LAC 对 PC 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（RADIUS Sever 1）进行认证。
5. RADIUS 服务器（RADIUS Sever 1）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. LNS 对 LAC 进行 PAP 或 CHAP 验证。

   USG9500 支持 PPP 的 PAP 和 CHAP 两种验证方式。

   - PAP 验证为两次握手验证，口令为明文。

     LAC 端向 LNS 端发送用户名和口令，LNS 根据本端用户表查看用户名和口令是否正确，然后返回不同的响应（Acknowledge or Not Acknowledge）。
   - CHAP 验证为三次握手验证，口令为密文（密钥）。

     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（RADIUS Sever 2）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. 若用户在 LNS 侧配置强制本端 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。

说明
USG9500 不支持 LAC（L2TP Access Concentrator）功能。可以选用其他支持 LAC 的设备。

6.2.6 应用场景

介绍 USG9500 作为 LNS 设备在防火墙和虚拟防火墙下的典型应用。

6.2.6.1 整机作为 LNS 设备

介绍 USG9500 在防火墙下作为 LNS 的应用。

说明
USG9500 不支持 LAC 功能。只支持 LNS 功能。

USG9500 作为 LNS 设备：

- 支持移动设备通过拨号软件远程接入 VPN，拨号软件既可以采用 Windows 自带的拨号软件也可以采用 VPN Client 软件。

- 支持分支机构通过 LAC 接入 VPN。

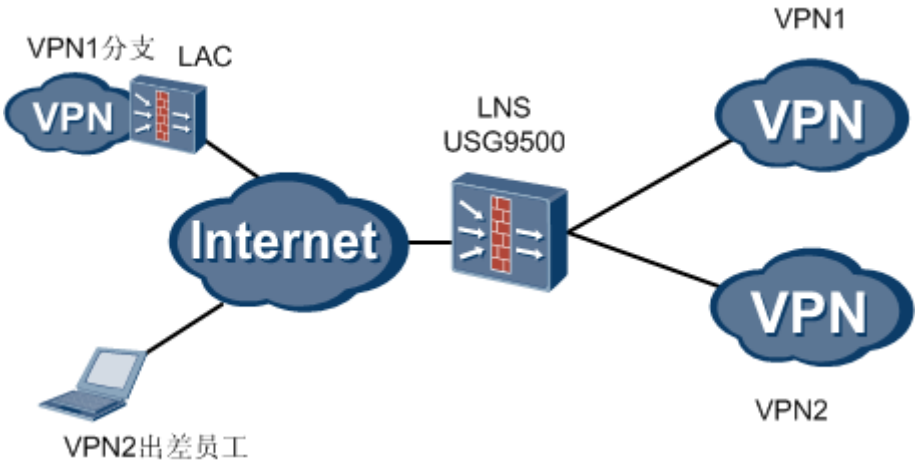
典型组网请参见 6.2.5.2 L2TP 隧道发起模式。

6.2.6.2 L2TP 多实例

介绍 USG9500 在虚拟防火墙下作为 LNS 的应用。

L2TP 多实例允许不同企业的用户通过 L2TP 接入不同企业网 VPN 网络。L2TP 多实例的典型组网如图 6-7 所示。

图6-7 L2TP 多实例典型组网



USG9500 的 GE 接口划分多个子接口，每一个子接口对应一个企业 VPN。每一个 GE 子接口对应 1 个 IP 地址池（相同的 VRF)。远程用户设置 LNS 的 IP 地址为防火墙与 internet 相连接口的 IP 地址，远程用户作为 LAC 发起呼叫，从而建立从 LAC 到 USG9500 的隧道，解封装后，由于呼叫用户与企业网络处于相同的 VPN，呼叫用户可以访问企业网内部资源。

6.3 GRE

6.3.1 介绍

GRE 协议用来对某些网络层协议的数据报文进行封装，使这些被封装的报文能够在另一网络层协议中传输。

定义

GRE（General Routing Encapsulation，通用路由封装）是对某些网络层协议的数据报文进行封装，使这些被封装的报文能够在另一网络层协议中传输。

GRE 可以作为 VPN 的第三层隧道协议，在协议层之间采用了一种被称之为隧道（Tunnel）的技术。Tunnel 是一个虚拟的点对点的连接，在实际中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文能够在这个通路上传输，并且在一个 Tunnel 的两端分别对数据报文进行封装及解封装。

目的

GRE 作为 VPN 的第三层隧道协议，为 VPN 数据提供透明传输通道。

GRE 主要有以下特点：

- 机制简单，对隧道两端设备的 CPU 负担小。
- 本身不提供数据的加密，可以与 IPSec 结合使用。
- 不提供流量控制和 QoS。

6.3.2 参考标准和协议

通过学习 GRE 相关的参考标准和协议加深对 GRE 特性的了解。

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)

6.3.3 可获得性

介绍 GRE 特性与 License 的关系以及哪些版本支持 GRE 特性。GRE 特性不需要 License 支持。

License 支持

本特性无须 License 支持。

版本支持

产品	支持版本
HUAWEI Secoway USG9500	V100R001、V100R003、V200R001

6.3.4 原理描述

主要介绍 GRE 的报文传输过程、报文头和安全机制。

6.3.4.1 报文传输过程

介绍 GRE 报文加封装与解封装的过程。

一个数据报文要在 Tunnel 中传输，必须经过封装与解封装两个过程，下面以图 6-8 的网络为例说明这两个过程：

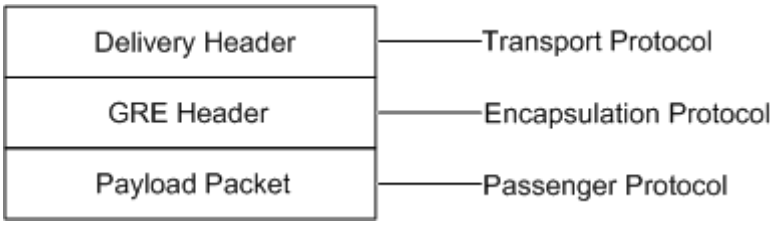
图6-8 私有 IP 网络通过 GRE 隧道互连

- 加封装过程

  USG9500_A 连接 IP group1 的接口收到 IP 数据报后首先交由 IP 协议处理，IP 协议检查 IP 报头中的目的地址域来确定如何转发此包。若报文的目的地址要经过 Tunnel 接口的 IP 地址，则将此报文发给 Tunnel 接口。Tunnel 口收到此包后进行 GRE 封装，封装完成后交给 IP 模块处理，在封装 IP 报文头后，根据此包的目的地址及路由表交由相应的网络接口处理。

  封装后的 GRE 报文格式如图 6-9 所示。

图6-9 封装好的 GRE 报文格式



其中：

- 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）

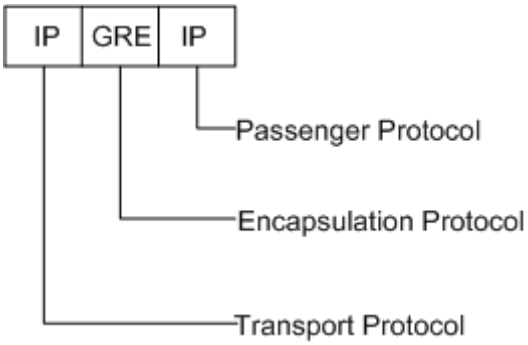
  上述的 GRE 协议称为封装协议，也称为运载协议（Carrier Protocol）。

- 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。

举例来说，一个封装在 IP Tunnel 中的 IP 传输报文的格式如图 6-10 所示。

图6-10 封装在 IP Tunnel 中的 IP 传输报文格式



- 解封装的过程

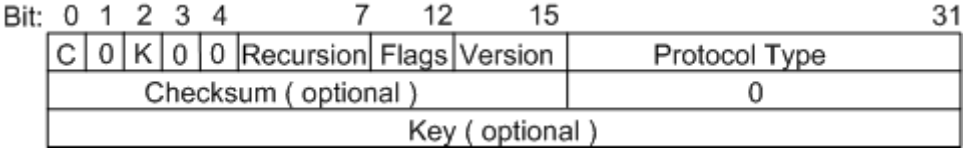
  解封装过程和加封装过程相反。USG9500_B 从 Tunnel 接口收到 IP 报文后检查目的地址，若发现目的地为本设备，则 USG9500_B 去掉此报文的 IP 头，然后交给 GRE 协议模块处理（进行检验识别关键字、检查校验等）；GRE 协议模块完成相应的处理后，去掉 GRE 报头，再交由 IP 协议模块处理，IP 协议模块象对待一般数据报一样对此数据报进行处理。

6.3.4.2 GRE 报文头

介绍 GRE 的报文头格式。

GRE 实现遵循 RFC 标准。GRE 头格式如图 6-11 所示。

图6-11 GRE 头格式



各字段解释如下：

- C

  校验和验证位。如果该位置 1，表示 GRE 头插入了校验和（Checksum）字段；该位为 0 表示 GRE 头不包含校验和字段。

- K

  关键字位。如果该位置 1，表示 GRE 头插入了关键字（Key）字段；该位为 0 表示 GRE 头不包含关键字字段。

- Recursion

  用来表示 GRE 报文被封装的层数。完成一次 GRE 封装后将该字段加 1。该字段的作用是防止报文被无限次的封装。

  USG9500 只支持一次 GRE 封装。

- Flags

  预留字段。当前必须设为 0。

- Version
  版本字段，必须置为 0。Version 为 1 是使用在 RFC2637 的 PPTP 中。
- Protocol Type
  乘客协议的协议类型。
- Checksum
  GRE 头及其负载的校验和字段。
- Key
  关键字字段，隧道接收端用于对收到的报文进行验证。

说明
GRE 头不包含源路由字段，Bit 1、Bit3 和 Bit 4 都置为 0。

6.3.4.3 安全机制

介绍 GRE 保证报文安全的两种机制。

GRE 本身提供两种安全机制：

- 校验和验证
- 识别关键字验证

校验和验证

校验和验证是指对封装的报文进行端到端校验。

RFC1701 中规定：如果 GRE 报文头中的 C 位置位，则校验和有效。校验和是 GRE 头中的可选字段。如果 C 位置 1，则发送方将根据 GRE 头及 payload 信息计算校验和，在报文头的 Checksum 字段的位置插入校验和，将包含校验和的报文发送给对端。接收方对接收到的报文计算校验和，并与报文中的校验和进行比较。如果计算出来的校验和与报文中的校验和一致，则对报文进一步处理，否则丢弃报文。

隧道两端可以根据实际需要选择是否配置校验和，从而决定是否触发校验功能。如果本端配置了校验和而对端没有配置，则本端将不会对接收到的报文进行校验和检查；相反本端没有配置校验和而对端已配置，本端将对从对端发来的报文进行校验和检查。

因校验和配置不同，对收发报文的处理方式也不同，请参见表 6-1。

表6-1 校验和与报文处理

本端	对端	本端对接收报文的处理	本端对发送报文的处理
配置校验和	没有配置校验和	不检查校验和	计算校验和
没有配置校验和	配置校验和	检查校验和	不计算校验和

识别关键字验证

识别关键字（key）是指对 Tunnel 接口进行校验。通过这种安全机制，可以防止错误识别、接收其它隧道来的报文。

RFC1701 中规定：若 GRE 报文头中的 K 位置 1，则在 GRE 头中插入关键字字段，收发双方将进行遂道识别关键字的验证。

关键字字段是一个四字节长的数，在报文封装时被插入 GRE 头。关键字的作用是认证隧道。属于同一流量的报文使用相同的关键字。在报文解封装时，隧道端将基于关键字来识别属于相同流量的数据报。

只有 Tunnel 两端设置的识别关键字完全一致时才能通过验证，否则将报文丢弃。这里的“完全一致”是指两端都不设置识别关键字；或者两端都设置关键字，且关键字的值相等。

6.3.5 应用场景

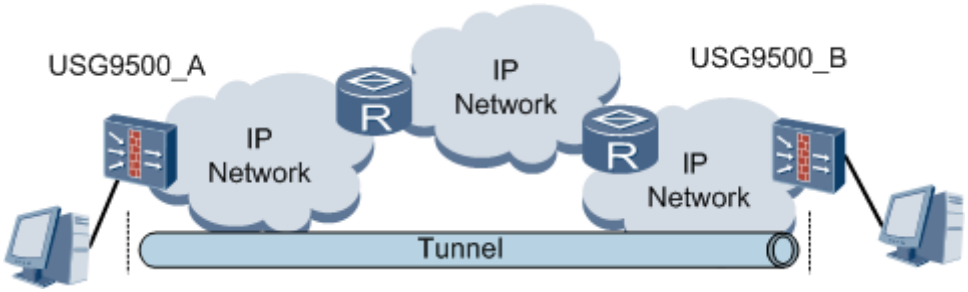
介绍 GRE 的典型应用。

6.3.5.1 扩大跳数受限的网络工作范围

在网络中通过使用 GRE 隧道可以扩大网络的工作范围。

如图 6-12 所示，网络运行 IP 协议，假设 IP 协议限制的跳数为 255。如果两台 PC 之间的跳数超过 255，它们将无法通信。在网络中使用隧道（Tunnel）可以隐藏一部分跳数，从而扩大网络的工作范围。

图6-12 扩大网络工作范围

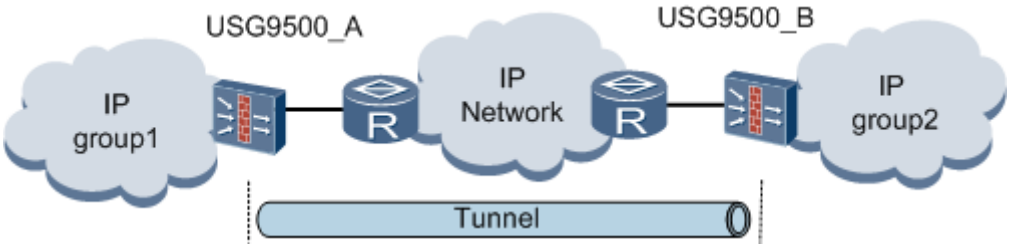


6.3.5.2 将不连续的子网连接起来组建 VPN

使用 GRE 隧道可以将不连续的子网连接起来，实现跨越广域网的 VPN。

例如，两个 VPN 子网 group1 和 group2 位于不同的城市，通过在网络边界设备之间建立 GRE 隧道，可以把这两个子网连接成一个连续的 VPN 网络。

图6-13 Tunnel 连接不连续子网



如图 6-13 所示，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。

6.4 IPSec

6.4.1 介绍

IPSec 是一系列为 IP 网络提供完整安全性的协议和服务的集合。IPSec 工作在 IP 层，为上层协议和应用提供透明的安全服务。

定义

IPSec（Internet Protocol Security）是 IETF 制定的一系列为了保护 IP 层通信安全的协议的组合。IPSec 是在 IPv6 构思框架下设计的。因为 IPv6 的开发与推广需要时间，而网络安全性的需求却迫在眉睫，于是被设计为同时支持 IPv4 和 IPv6 的 IPSec 应运而生。

IPSec 主要通过 AH（Authentication Header）协议和 ESP（Encapsulating Security Payload）协议提供安全服务。AH 和 ESP 是 IPSec 协议族的两个主要协议。AH 和 ESP 协议主要依赖相应的验证算法和加密算法提供验证和加密的安全服务。AH 协议和 ESP 协议都支持两种报文封装模式，即传输模式和隧道模式。这两种报文封装模式就是将 ESP 和 AH 应用到 IP 数据包中的两种不同的方法。

IPSec 提供安全服务需要用到共享密钥。共享密钥既可以通过手工方式增加，也可以密钥管理交换协议 IKE（Internet Key Exchange）来动态生成。

目的

由于 IP 包本身并不集成任何安全特性，很容易便可伪造出 IP 包的地址、修改其内容、重播以前的包以及在传输途中拦截并查看包的内容。针对这些问题，IPSec 可有效地保护 IP 数据报文的安全。

IPSec 为 IP 数据报文提供了高质量的、可互操作的、基于密码学的安全性。特定的通信方之间在 IP 层通过加密与数据源验证等方式，来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。具体介绍如下：

- 私有性（Confidentiality）

  指对用户数据进行加密保护，用密文的形式传送。
- 完整性（Data integrity）

指对接收的数据进行验证，以判定报文是否被篡改。

- 真实性（Data Authenticity）
  指验证数据源，以保证数据来自真实的发送者。
- 防重放（Anti-replay）
  指防止恶意用户通过重复发送捕获到的数据包所进行的攻击，即接收方会拒绝重复的数据包。

6.4.2 规格

介绍 IPSec 的性能。

IPSec 性能

USG9500 的 IPSec 性能说明如表 6-2 所示。

表6-2 IPSec 性能

参数	性能
IKE Peer 数目	USG9500 上最多可配置 2000 个
IPSec 安全策略数目	USG9500 上最多可配置 2000 个
安全联盟（SA）数目	单个 CPU 上最多支持 4 万个
IKE 提议数目	USG9500 上最多可配置 100 个
IPSec 提议数目	USG9500 上最多可配置 50 个

6.4.3 参考标准和协议

通过学习 IPSec 相关的参考标准和协议可以深入了解 IPSec 特性。

与 IPSec 特性相关的参考标准与协议如下：

- RFC 2403：The Use of HMAC-MD5-96 within ESP and AH
- RFC 2409：The Internet Key Exchange (IKE)
- RFC 2857：The Use of HMAC-RIPEMD-160-96 within ESP and AH
- RFC 3625：More Modular Exponential (MODP) Diffie-Hellman groups for Internet Key Exchange (IKE)
- RFC 3664：The AES-XCBC-PRF-128 Algorithm for the Internet Key Exchange Protocol (IKE)
- RFC 3706：A Traffic-Based Method of Detecting Dead Internet Key Exchange (IKE) Peers
- RFC 3748：Extensible Authentication Protocol (EAP)
- RFC 3947：Negotiation of NAT-Traversal in the IKE
- RFC 4109：Algorithms for Internet Key Exchange version 1 (IKEv1)

- RFC 3948：UDP Encapsulation of IPsec ESP Packets
- RFC 4305：Cryptographic Algorithm Implementation Requirements for Encapsulating Security Payload (ESP) and Authentication Header (AH)
- RFC 4306：Internet Key Exchange (IKEv2) Protocol
- RFC 4307：Cryptographic Algorithms for Use in the Internet Key Exchange Version 2 (IKEv2)
- RFC 4322：Opportunistic Encryption using the Internet Key Exchange (IKE)
- RFC 4359：The Use of RSA/SHA-1 Signatures within Encapsulating Security Payload (ESP) and Authentication Header (AH)
- RFC 4434：The AES-XCBC-PRF-128 Algorithm for the Internet Key Exchange Protocol (IKE)
- RFC 4478：Repeated Authentication in Internet Key Exchange (IKEv2)

6.4.4 可获得性

介绍 IPSec 特性与 License 的关系以及哪些版本支持 IPSec 特性，以及处理 IPSec 协议的硬件。IPSec 性能受 License 文件控制。

License 支持

缺省情况下，USG9500 支持 100 个 IPSec 隧道。用户可以购买并激活 License 来增加 IPSec 隧道数。

只有当设备的 ESN（Equipment Serial Number）编号与 License 文件完全匹配的情况下，License 才能成功被激活。用户可以使用 **display license** 命令查看设备的 ESN 编号。 License 文件被激活后请保存配置，之后 License 将长期生效。 当新激活的 License 包含的资源比旧的 License 包含的资源少时，设备重新启动前，原有资源仍能继续使用。

版本支持

产品	支持版本
HUAWEI Secoway USG9500	V100R001、V100R003、V200R001

硬件要求

实际应用中，使用 IPSec 软件进行加密/解密运算会占用大量的 CPU 资源，影响整机性能。为解决这一问题，USG9500 本身集成了 SAE(Security Accelerate Engine)加密引擎，以硬件方式完成数据的加/解密运算，消除了软件处理 IPSec 协议对性能的影响，提高了 USG9500 工作效率。

加密引擎与 IPSec 模块对数据的处理机制完全相同，区别在于加密引擎是通过硬件实现加密/解密处理，而 IPSec 模块是通过软件实现加/解密处理。

6.4.5 特性增强

介绍 IPSec 特性在不同版本的变化，包括增加、修改和删减等情况。

版本	特性增强
V100R003	增加“IPSec 隧道化”、“DHCP over IPSec”、“IPSec 双机热备”功能。 增加“RSA 签名方式”。 修改“IPSec NAT 穿越功能缺省状态”。修改后 IPSec NAT 穿越功能缺省打开。
V200R001	增加“IPv6 IPSec”、“IPSec 多实例”、“L2TP over IPSec 多实例”功能。 修改缺省情况下支持的 IPSec 隧道数。缺省情况下，USG9500 支持 100 个 IPSec 隧道。

6.4.6 IPSec 原理描述

介绍 IPSec 的安全协议、封装模式、密钥管理和 IPSec 安全联盟。

6.4.6.1 安全协议

介绍 IPSec 的两种主要安全协议：AH 协议和 ESP 协议。以及为 AH 协议和 ESP 协议所依赖的验证算法和加密算法。

IPSec 包括两种主要的安全协议：AH 协议和 ESP 协议。IPSec 通过这两个安全协议对数据进行加密和验证：

- AH 协议

  AH 是报文头验证协议，主要提供的功能有数据源验证、数据完整性校验和防报文重放功能。然而，AH 并不加密所保护的数据报文。

- ESP 协议

  ESP 是封装安全载荷协议。它除提供 AH 协议的所有功能外（但其数据完整性校验不包括 IP 头），还可提供对 IP 报文的加密功能。

AH 协议和 ESP 协议既可以单独使用，也可以同时使用。

AH 和 ESP 协议所提供的加密或验证保护完全依赖于它们采用的加密算法或验证算法。

验证算法

AH 和 ESP 都能够对 IP 报文的完整性进行验证，以判别报文在传输过程中是否被篡改。验证算法主要是通过杂凑函数来完成验证。杂凑函数是一种能够接受任意长的消息输入，并产生固定长度输出的算法。杂凑函数的输出被称为消息摘要。在发送报文和接收报文的两端分别进行消息摘要的计算，如果两个消息摘要是相同的，则表示报文是完整未经篡改的。

一般来说 IPSec 使用两种验证算法：

- MD5（Message Digest 5）
  输入任意长度的消息，产生 128bit 的消息摘要。
- SHA-1（Secure Hash Algorithm）
  输入长度小于 2^{64}bit 的消息，产生 160bit 的消息摘要。

SHA-1 比 MD5 具有更高的安全性。

加密算法

ESP 能够对 IP 报文内容进行加密保护，防止报文内容在传输过程中被窥探。加密算法实现主要通过对称密钥系统，它使用相同的密钥对数据进行加密和解密。

一般来说 IPSec 使用加密算法有以下几种：

- DES（Data Encryption Standard）
  使用 56bit 的密钥对一个 64bit 的明文块进行加密。
- 3DES（Triple Data Encryption Standard）
  使用三个 56bit 的 DES 密钥（共 168bit 密钥）对明文进行加密。
- AES（Advanced Encryption Standard）
  使用 AES 密钥对明文进行加密。密钥的长度分为 128bit、192bit、256bit。

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。

AES 算法是一种更新的算法，AES 的安全性比 DES 和 3DES 算法都高，而且 AES 的计算复杂度低。一般情况下 128bit 就可以充分满足安全需求。

6.4.6.2 封装模式

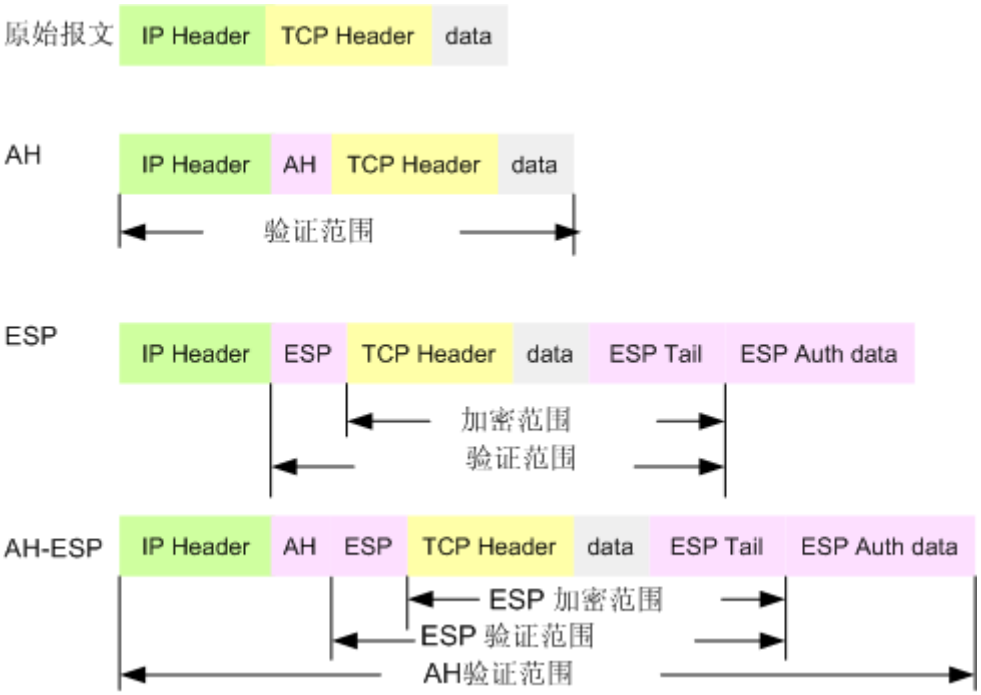
介绍 IPSec 支持的两种封装模式，以及 AH 协议和 ESP 协议在两种封装模式下提供的验证和加密范围。

AH 协议和 ESP 协议都支持两种封装模式：传输模式和隧道模式。

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。以 TCP 为例，原始报文经过 AH 协议或 ESP 协议的传输模式封装后，报文格式如图 6-14 所示。

图6-14 传输模式下报文封装



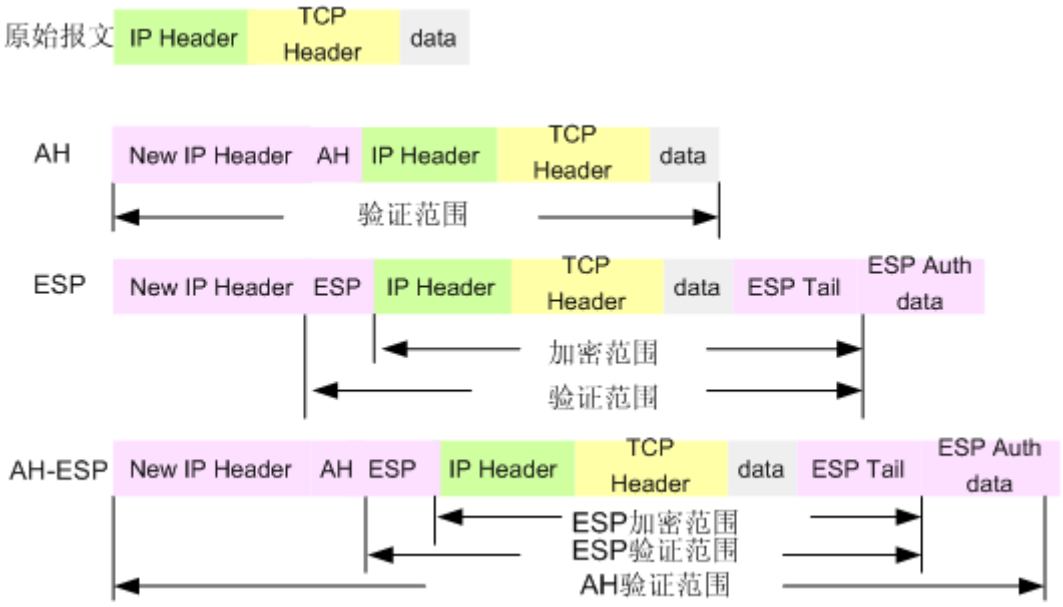
传输模式下，AH 协议的完整性验证范围为整个 IP 数据包。ESP 协议验证报文的完整性检查部分包括 ESP 包头、传输层协议头、数据和 ESP 报尾，但不包括 IP 头，因此 ESP 协议无法保证 IP 头的安全。ESP 的加密部分包括传输层协议头、数据和 ESP 报尾。

说明
AH-ESP 表示同时采用 AH 协议和 ESP 协议。

隧道模式

在隧道模式下，AH 或 ESP 插在原始 IP 头之前，另外生成一个新的报文头放到 AH 或 ESP 之前。以 TCP 为例，如图 6-15 所示。

图6-15 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 数据包。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。

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。
- 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。

6.4.6.3 密钥管理

IPSec 提供安全服务需要用到共享密钥。密钥既可以通过手工方式创建，也可以密钥管理交换协议 IKE（Internet Key Exchange）来动态生成。

手工创建密钥

当采用手工方式创建 IPSec 时，所有的安全参数都必须手工创建。手工方式只适用于小型、静态的网络环境。

IKE 自动协商密钥

IKE 能够为 IPSec 提供自动协商交换密钥、建立安全联盟的服务，以简化 IPSec 的使用和管理。只需要配置好 IKE 协商安全策略的信息，就可以由 IKE 自动协商来创建和维护安全联盟。此种方式适合于复杂的大中型或动态网络。推荐采用 IKE 自动协商密钥。

采用 IKE 自动协商时，可以采用如下密钥认证方式：

- 预共享密钥

  需要为每个对端配置预共享密钥认证字。预共享密钥认证字是验证双方身份的关键。 配置 IPSec 的两个对端的预共享密钥认证字必须一致。

- RSA 签名方式

  RSA 签名方式适用于大型网络和网络扩展的情况。此种方式不需要对每个对等体进行配置，但是用户必须从 CA 获取证书，然后通过 PKI 域指定 IPSec 隧道协商时使用的证书。此种方式通过证书进行验证，相对于预共享方式来说，更加安全。

6.4.6.4 IPSec 安全联盟

IPSec 安全联盟是 IPSec 对等体间对某些要素的约定。本节介绍 IPSec 安全联盟的概念、特点、创建方式和生存周期。

概念

IPSec 安全联盟即 IPSec SA（Security Association）与对等体和隧道存在密切的关系。IPSec 安全联盟、对等体和隧道的定义如下：

- 对等体

  IPSec 在两个端点之间提供安全通信，IPSec 一侧的端点称为另一侧端点的对等体（peer）。

- IPSec 安全联盟

  IPSec 安全联盟是 IPSec 对等体间对某些要素的约定，例如，使用哪种协议（AH、ESP 或 AH-ESP）、报文的封装模式（传输模式或隧道模式）、加密算法（DES、3DES 或 AES）、特定数据流中保护数据的共享密钥以及密钥的生存周期等。

- 隧道

  两个 IPSec 对等体之间的数据传输通道被称为 IPSec 隧道。通过隧道传输的数据只能从一个端点传输到其对等体。

  在 IPSec 隧道创建过程中，必须首先在对等体之间创建安全联盟。安全联盟建立后，隧道也就建立了。

特点

IPSec 安全联盟由一个三元组来唯一标识，这个三元组包括：

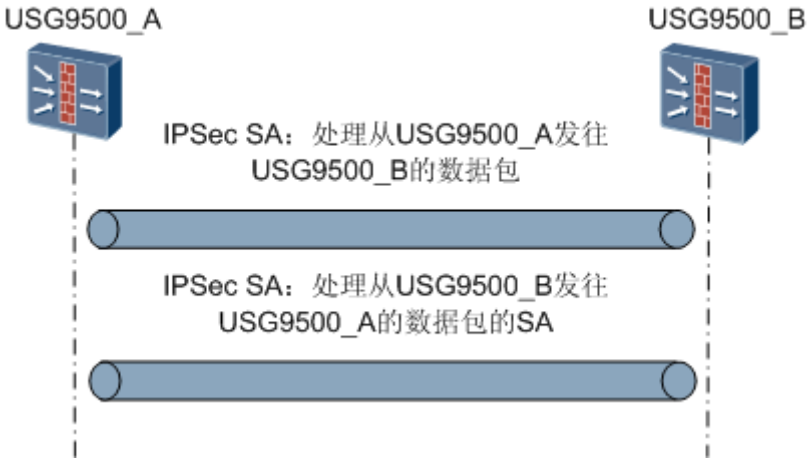
- SPI（Security Parameter Index）

  SPI 是为唯一标识 SA 而生成的一个 32bit 的数值，它在 AH 和 ESP 头中传输。

- 目的 IP 地址
- 安全协议号（AH 或 ESP）

IPSec 安全联盟是单向的逻辑连接。这就要求 IPSec 一端的设备上至少需要一对 SA 来分别传输设备发送的数据和设备接收的数据。传输发送数据的 SA 被称为出方向 SA，传输接收到的数据的 SA 被称为入方向 SA。如图 6-16 所示，USG9500_A 和 USG9500_B 存在一对 SA，USG9500_A 的入方向 SA 和出方向 SA 分别对应 USG9500_B 出方向的 SA 和入方向 SA。

图6-16 IPSec 安全联盟示意图



对等体之间建立的 IPSec 安全联盟的个数还与协议相关。当 USG9500_A 和 USG9500_B 使用 AH 或 ESP 协议进行安全通信，只需要一对 IPSec SA。如果 USG9500_A 和 USG9500_B 同时使用 AH 和 ESP 进行安全通信，则需要两对 IPSec SA，AH 协议需要一对 SA（出方向和入方向各一个）和 ESP 协议需要一对 SA（出方向和入方向各一个）。

创建方式

IPSec 安全联盟支持两种创建方式：

- 手工方式（**manual**）

  手工方式配置比较复杂，创建安全联盟所需的全部信息都必须手工配置，并且不支持 IPSec 的一些高级特性（例如定时更新密钥）。

  通过手工方式创建安全联盟，不需要进行安全联盟协商，只要安全策略在接口上应用后就直接创建。手工方式安全联盟一旦建立就永远不失效，无法设置生存周期。

- IKE 自动协商方式（**isakmp**）

  IKE 自动协商方式相对比较简单，只需要配置好 IKE 协商安全策略的信息，由 IKE 自动协商和维护安全联盟。

  通过 IKE 自动协商方式建立安全联盟，安全联盟由 IKE 进行隧道协商，IKE 协商安全联盟可以设置生存周期，从而使安全联盟更加的安全。

当与 USG9500 进行通信的对等体设备数量较少时，或是在小型网络环境中，手工配置安全联盟是可行的。对于中、大型的网络环境中，推荐使用 IKE 协商建立安全联盟。

生存周期

IKE 自动协商生成的 IPSec 安全联盟可以设置生存周期。生存周期的定义包括两种方式：

- 以时间为基准定义。每隔指定的时间就更新安全联盟。
- 以流量为基准定义。每传输指定的数据量（字节）就更新安全联盟。

6.4.7 IKE 原理描述

介绍 IKE 的原理。

6.4.7.1 介绍

介绍 IKE 的定义、IKE 与 IPSec 的关系、使用 IKE 可以达到的目的。

定义

IKE 协议是 Oakley 协议和 SKEME 协议的一种混合，并建立在由 Internet 安全联盟和密钥管理协议 ISAKMP（Internet Security Association and Key Management Protocol）定义的框架上。它能够为 IPSec 提供自动协商交换密钥、建立安全联盟的服务，以简化 IPSec 的使用和管理。

IKE 与 IPSec 的关系

IKE 并非 IPSec 专用，其他的协议也可以通过 IKE 来协商安全服务。IKE 采用的规范是在“解释域 DOI（Domain of Interpretation）”中制定的。针对 IPSec 存在一个名为 RFC2407 的 DOI，它定义了 IKE 具体如何与 IPSec 安全联盟进行协商。如果其他协议要用到 IKE，每种协议都要定义各自的 DOI。

作为 IPSec VPN 实现中的首选密钥交换协议，IKE 保证了安全关联 SA 建立过程的安全性和动态性。IKE 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。新版的 IKEv2 协议保留了传统 IKE 的基本功能，并针对 IKE 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 IKE 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。

目的

IKE 具有一套自保护机制，可以在不安全的网络上安全地分发密钥、验证身份、建立 IPSec 安全联盟。

6.4.7.2 安全联盟协商过程

介绍 IKE 为 IPSec 协商安全联盟的过程，以及采用 IKEv2 进行安全联盟协商的优势。

采用 IKE（即 IKEv1）协商安全联盟的过程

IKE 使用了两个阶段为 IPSec 进行密钥和安全联盟协商：

- 第一阶段，建立 IKE 安全联盟（IKE SA）。

  IKE 安全联盟是一个保密和验证无误的通信信道。IKE 安全联盟利用验证过的密钥为通信双方的 IKE 通信提供机密性、消息完整性以及消息源验证服务。

  在 RFC 2409 中规定，IKE 第一阶段的协商可以采用两种交换模式：

  - 主模式（Main Mode）

    主模式分为三次交换，总共用到六条消息，最终建立 IKE SA。这三次交换是：

1. 策略协商
2. DH 和 nonce 交换
3. 对对方身份的验证

   主模式被设计成将密钥交换信息与身份、认证信息相分离。这种分离保护了身份信息；交换的身份信息受已生成的 Diffie-Hellman 共享密钥的保护，但这增加了开销。

   - 野蛮模式（Aggressive Mode）

     野蛮模式则允许同时传送与 SA、密钥交换和认证相关的载荷。将这些载荷组合到一条消息中减少了消息的往返次数，但是无法提供身份保护。

     虽然野蛮模式存在一些功能限制，但可以满足某些特定的网络环境需求。例如：远程访问时，如果响应者（服务器端）无法预先知道发起者（终端用户）的地址、或者发起者的地址总在变化，而双方希望采用预共享密钥验证方法来创建 IKE SA。

- 第二阶段，建立 IPSec 安全联盟（IPSec SA）。

  第二阶段，利用在第一阶段建立的安全联盟（IKE SA）为 IPSec 协商安全服务，即为 IPSec 协商具体的安全联盟，建立 IPSec SA。IPSec SA 用于最终的 IP 数据安全传送。

  第二阶段只存在一种交换模式，即快速模式。这种交换在特定的 IKE SA 保护下进行。

具体安全联盟的建立过程如图 6-17 所示。

图6-17 安全联盟建立过程

安全联盟建立的过程解析如下：

1. 当一个报文从某接口外出时，如果此接口应用了 IPSec 安全策略，会进行安全策略的匹配。
2. 如果找到匹配的安全策略，会查找相应的 IPSec SA。如果 IPSec SA 还没有建立，则触发 IKE 进行协商。IKE 首先建立第一阶段的安全联盟，即 IKE SA。
3. 在 IKE SA 的保护下协商 IPSec SA。

4. 使用 IPSec SA 保护通讯数据。

采用 IKEv2 协商安全联盟的优势

IKE 协商中要建立一对 IPSec SA，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”。前者需要交换至少 9 条消息，后者也至少需要 6 条消息。

而 IKEv2 建立一对 IPSec SA，正常情况使用两次交换也就是 4 条消息就可以完成一个 IKE SA 和一对 IPSec SA 的协商建立，如果要求建立的 IPSec SA 大于一对时，每一对 SA 只需额外增加一次交换，也就是两条消息就可以完成。这比 IKEv1 要简化很多。

6.4.7.3 IKE 安全联盟

IKE 安全联盟是在安全联盟协商的第一阶段生成的。IPSec 安全联盟必须在 IKE 安全联盟的保护下生成。

同 IPSec 一样，IKE 也存在“安全联盟”的概念。但是与 IPSec 安全联盟不同，IKE 安全联盟定义了双方的通信形式。包括，用哪种算法来加密 IKE 通信；怎样对远程通信方的身份进行验证等。

IKE 安全联盟是采用 IKE 协商安全联盟的第一阶段。IPSec 安全联盟是采用 IKE 协商安全联盟的第二阶段。IPSec SA 必须在 IKE SA 的保护下创建。IKE SA 是双向的逻辑连接，不用来传输数据。

IKE SA 可以在通信双方之间提供任意数量的 IPSec SA。通过一次 IKE 密钥交换，可创建多对 IPSec SA。而且单独一个 IKE SA 可进行任意数量的这种交换。

IKE SA 也存在“生存周期”。IKE SA 的生存周期也存在基于时间和基于流量两种方式进行定义。对于 IKE SA 来说，除非生存周期耗尽或由于某种外部原因而被删除，否则它会一直保持活动状态。

6.4.7.4 IKE 的安全机制

介绍 IKE 的安全机制。

IKE 可以提供如下安全机制：

- DH（Diffie-Hellman）交换及密钥分发

  Diffie-Hellman 算法是一种公共密钥算法。通信双方在不传送密钥的情况下通过交换一些数据，计算出共享的密钥。加密的前提是交换加密数据的双方必须要有共享的密钥。IKE 的精髓在于它永远不在不安全的网络上直接传送密钥，而是通过一系列数据的交换，最终计算出双方的密钥。即使第三者（如黑客）截获了双方用于计算密钥的所有交换数据，也不足以计算出真正的密钥。

- 完善的前向安全性

  PFS（Perfect Forward Secrecy）是一种安全特性，指一个密钥被破解，并不影响其他密钥的安全性，因为这些密钥间没有派生关系。PFS 是由 DH 算法保障的。此特性是通过在 IKE 阶段 2 的协商中增加密钥交换来实现的。

- 身份验证

  身份验证确认通信双方的身份。支持预共享密钥认证和证书认证-RSA 签名方式。

- 身份保护

  身份数据在密钥产生之后加密传送，实现了对身份数据的保护。

6.4.7.5 IKEv2 的安全性分析

本节主要介绍 IKEv2 的安全性问题。

IKEv2 对传统 IKE 存在的安全漏洞进行了修订，提高了密钥协商的安全性，并明确规定了所有的消息必须以请求/响应对的形式存在，有效的解决了使用 UDP 作为传输层协议的不可靠性问题。

以下从三方面来讨论 IKEv2 的安全性问题。

抵御中间人攻击

中间人攻击（Man-in-the-middle Attack）是一种主动攻击，指攻击者对通信双方进行窃听，截获通信双方的消息并进行任意插入、删除或篡改消息，之后返回消息给发送者，或者重放旧消息以及重定向消息，是最危险的攻击。IKEv2 中抵御中间人攻击的机制和方法：

- 密钥材料生成方式

  与传统 IKE 相比，IKEv2 的密钥材料发生了变化，双方用于后继交互使用的加密密钥与认证密钥都是不同的。这些密钥是从 prf+输出流中依次提取，从而增加了攻击者猜测密钥的难度，减少了密钥泄漏的可能性，增强了传输的安全性，一定程度上可以抵御中间人攻击。

- 身份认证

  IKEv2 使用预共享密钥和数字签名方式进行身份认证。身份认证方式具有交互性，参与协商的实体彼此都对对方的身份进行认证；具有对称性，参与协商的双方都使用相同的机制或方法对对方的身份进行认证。双向的身份认证可以有效地抵御中间人攻击。同时 IKEv2 定义了扩展认证交互，即使用扩展认证协议（EAP）描述的方法对 IKEv2 协商进行身份认证，支持非对称双向认证，进一步加强了认证的灵活性和协商的可扩展性。

- 消息交换

  IKEv2 将传统 IKE 主模式交换的六条消息修订为四条消息，将 SA 载荷和 KE 载荷、nonce 载荷一同发送，这样，消息中包含随机的 nonce 值，如果攻击者伪装成响应方进行应答，将收到的发起方的消息基本上不做改变，再发回给发起方，发起方可以根据消息内容判断消息的真假，在一定程度上可以抵御重放攻击。每个 IKEv2 消息的头都包含了一个消息 ID，用于匹配对应的请求和响应消息以及识别消息重传。当发送和接收到请求时，必须对消息 ID 值顺序增加，且除了 IKE_SA_INIT 交互外其值受加密和完整性保护，使得它能够防重放攻击。同时 IKEv2 加入了滑动窗口机制，使交互能够更加有效地抵御重放攻击的威胁。

抵御 DDoS 攻击

IKEv2 中抵御 DDoS 攻击的机制和方法：

- SPI 值

  IKEv2 消息头部有发起方 SPIi 和响应方 SPIr，它们是内核产生的 8 字节的随机数，用来标识 SA，同时也可以标识进行消息交换的一对节点。具有相同 SPI 值的请求处理一次（重传消息除外），而把其他请求作为重复数据报丢弃，可以在一定程度上防止 DDoS 攻击。

- 带 Cookie 交互

IKEv2 中使用 N 载荷携带 Cookie 的辅助交换来抵御拒绝服务攻击。在通信过程中，响应方认为自己正受到 DDoS 攻击时，可以向发起方请求回复一个无状态 cookie。

响应方收到对方发来的第一条消息后并不急于进行 IKE_SA_INIT 交互，而是再产生一个新的 cookie，封装在通知载荷中发送给对方。如果发起方不是攻击者，就可以收到这条消息，然后重新开始协商，并将响应方的 cookie 封装在该消息中，其它载荷内容保持不变。

- 重传约定

  IKEv2 中所有消息都是成对出现，在每对消息中，发起方负责重传事件，响应方不必对其响应消息进行重传，除非收到对方的一个重传请求。避免了双方同时发起重传，造成资源的浪费，同时也可以防止攻击者截获消息后，伪装成协商者不断地发送重传消息，耗费协商双方的资源。

- 丢弃半开（half-open）连接

  IKEv2 只能通过两种情况判断对方是否失效：一种是重复尝试联系对方，直到应答时间过期；另一种是收到对方的不同 IKE SA 加密保护下的 INITIAL CONTACT 通知消息。IKEv2 发起方允许多个响应方响应第一条消息，并把所有的响应方视为合法并作回应。发起方发送一些消息后，一旦收到一个有效的加密的响应消息，将其他的响应消息忽略，并将其他所有的无效的半连接丢弃。这样在协商开始时就可以避免受到 DDoS 攻击。

完善的前向安全性 PFS

完善的前向安全性，即限制单密钥只能解密受该单密钥保护的数据。即使攻击者攻克了一个密钥，也只能破解这个密钥保护的数据，而不能破解受其它密钥保护的数据。对于 IPSec VPN 来说，是指在 IKE 协商阶段所用的加密密钥同 IPSec 使用的加密密钥，源于不同密钥衍生材料，即使攻击者攻克 IKE 协商阶段密钥，也并不能破解 IPSec 加密消息。

IKEv2 初始交互的密钥衍生材料不被用于衍生供 IPSec SA 使用的相关密钥，而是通过在 CREATE_IPSec_SA 交互中引入可选的 KE 载荷重新生成密钥材料，以此有效完成 PFS 服务要求。

6.4.7.6 EAP 认证

IKEv2 中支持采用 EAP 对协商的发起方(Initiator)进行第三方认证。

EAP（Extensible Authentication Protocol）是一种支持多种认证方法的认证协议。可扩展性是 EAP 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。采用 EAP 方式认证，可以方便的继承系统原有的认证机制。

IKEv2 中支持采用 EAP 对协商的发起方(Initiator)进行第三方认证。响应方根据发起方消息中有无 AUTH 载荷来判断是否需要 EAP 认证。

如果没有 AUTH (Authentication) 载荷则表示发起方请求 EAP 认证，在响应方发回的 Response 消息指定自己允许的 EAP 认证方法。发起方的下一个 Request 消息携带了对应于该 EAP 方法的认证信息，收到该消息后响应方向第三方的 EAP 认证服务器按照 RFC 3748（Extensible Authentication Protocol）的规范进行认证。然后在 Response 消息中发回认证成功或失败的信息。

在实现中响应方完全不用知道具体的认证方法和过程，而仅充当发起方和 EAP 认证服务器的中转(pass through 模式)，由发起方和 EAP 认证服务器来完成认证的全过程而响应方只需要得到认证结果。这样可以支持很多的认证方式，包括很多高强度的认证算法而不用增加响应方的软件复杂度。

6.4.8 应用场景

介绍 IPSec 的典型应用场景。

6.4.8.1 网关到网关场景

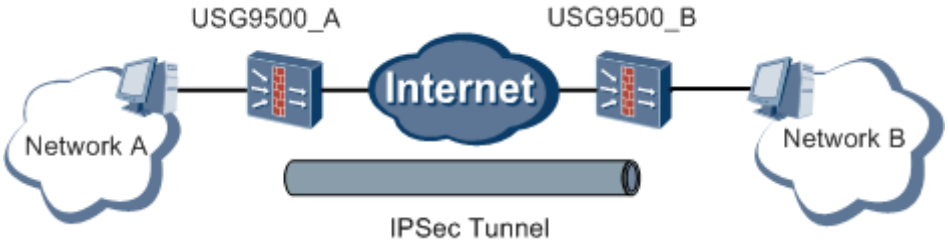
两个网关之间的点到点通信模式下，两个网关之间的连接是加密的。但从客户到客户的网关之间的连接，服务器和服务器的网关之间的连接是未加密的。两侧的网关都必须支持 IPSec。

网关对网关模式的典型组网如图 6-18 所示。要在两端网关之间建立 IPSec 隧道，每端的网关上都必须配置 IP 地址。通过两个网关建立 IPSec 隧道，网络 A 和网络 B 可以在 IPSec 保护下互访。

USG9500 不支持动态获取 IP 地址。当两端网关的 IP 地址都为固定 IP 地址时，两端网关都可以采用 USG9500。

当一端 IP 地址不固定时，可以采用其他支持动态获取 IP 地址的 IPSec 网关设备。

图6-18 两个网关之间的点到点通信组网

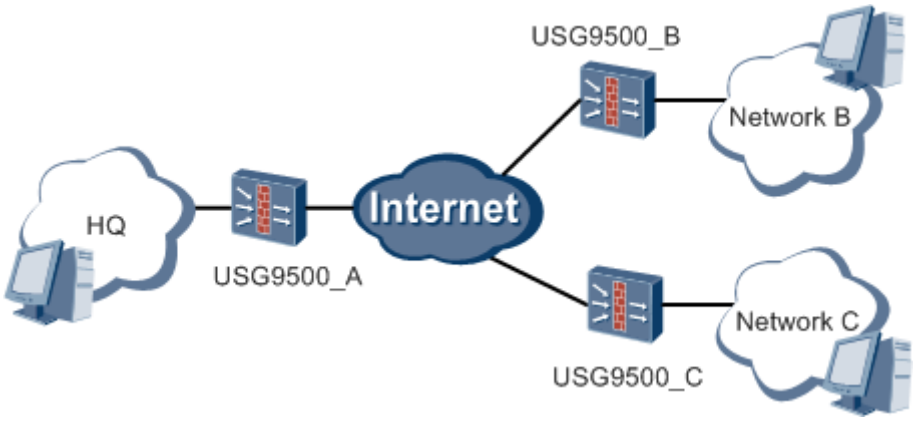


6.4.8.2 Hub to Spoke 场景

介绍一个网关和多个对端网关之间的通信场景，也是最常见的 IPSec 组网之一。

Hub to Spoke 组网模式适用于公司总部与分公司之间进行通信。其典型组网如图 6-19 所示。要求总部网关 IP 地址为固定的公网 IP。分支机构网关 IP 既可以是固定公网 IP 也可以是动态获取的 IP 地址。IP 地址可以通过 3G、ADSL、PPPoE 拨号或 DHCP 方式获取。

图6-19 Hub to Spoke 典型组网



说明

USG9500 不支持动态获取 IP 地址。需要动态接入的分公司网关可以采用其他支持动态获取 IP 地址且支持 IPSec 功能的网关设备。

6.4.8.3 网关之间 L2TP over IPSec 场景

介绍两个网关之间 L2TP 与 IPSec 结合使用的场景，可以将远程登录与加密和认证的功能结合起来。

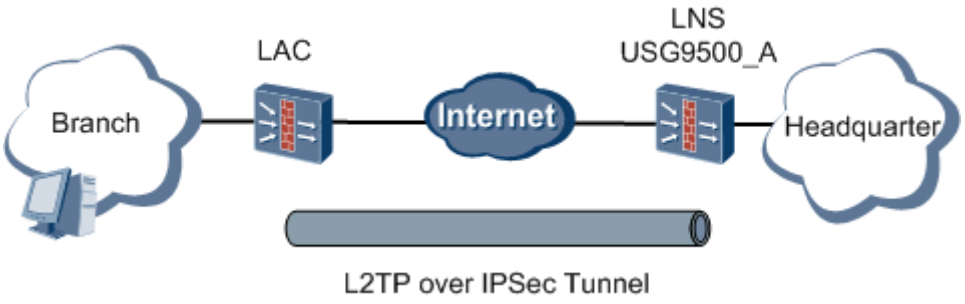
两个网关之间 L2TP over IPSec 的典型场景如图 6-20 所示。适用于分支网络通过 LAC 拨号接入 VPN 并进行安全访问的场景。

分支网络的用户通过 L2TP 拨号获取可以访问总部网络的内网 IP 地址。L2TP 拨号用户可以采用本地认证也可以采用 AAA 服务器进行认证。

在 L2TP over IPSec 中先进行 L2TP 封装，再进行 IPSec 封装，可以实现报文的安全传输。

USG9500 不支持 LAC，只支持 LNS。

图6-20 两个网关之间的 L2TP over IPSec 组网



6.4.8.4 移动设备通过 L2TP over IPSec 方式远程接入 VPN

介绍移动设备通过 L2TP over IPSec 方式远程接入 VPN 的场景。

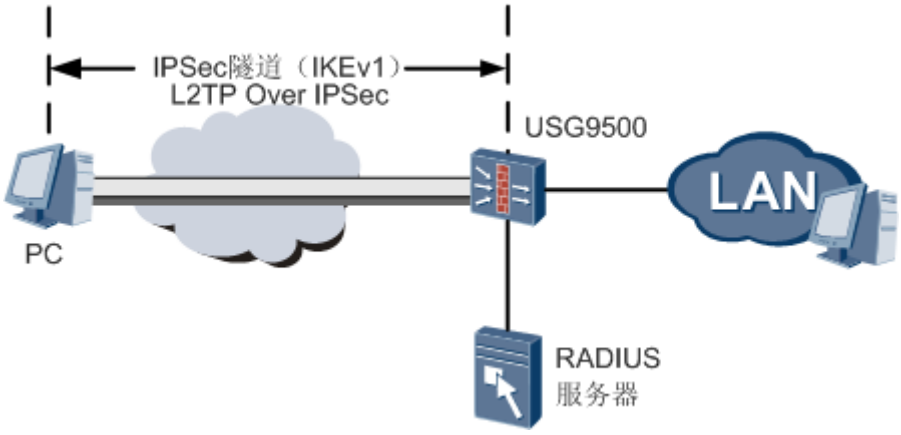
移动设备远程接入场景适用于出差员工或其他移动设备访问公司网络的情况。

移动设备可以通过 L2TP over IPSec 方式接入公司网络。此时采用 L2TP 进行拨号，采用 IKEv1 进行 IPSec 协商。用户认证方式可以采用本地认证或 AAA 服务器认证（用户较少时，可采用本地认证），认证通过的用户可以通过隧道安全访问总部服务器。

在移动设备可以使用 Windows 自带拨号软件或 VPN Client 进行拨号。VPN Client 软件为本公司提供的拨号软件，推荐安装 VPN Client 并使用 VPN Client 进行拨号。

移动设备 L2TP over IPSec 方式远程接入 VPN 的典型组网如图 6-21 所示。

图6-21 采用 L2TP Over IPSec 方式接入 IPSec 网关



6.4.8.5 移动设备通过 EAP 方式远程接入 VPN

介绍移动设备通过 EAP 方式远程接入 VPN 的场景。

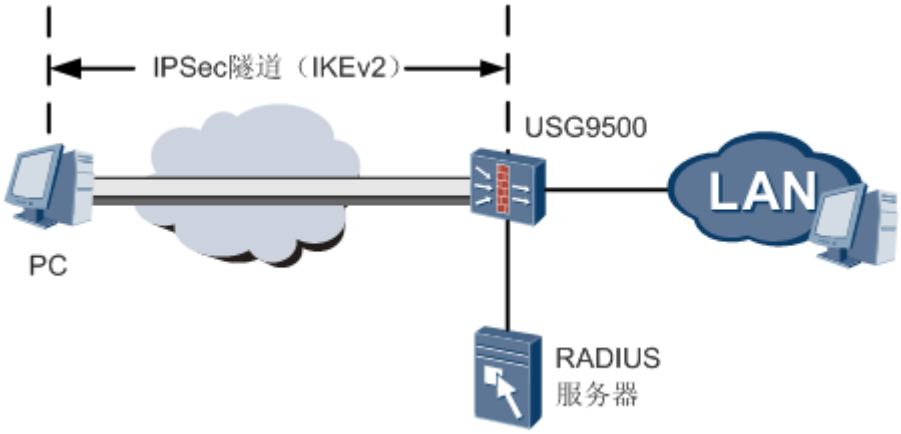
移动设备远程接入场景适用于出差员工或其他移动设备访问公司网络的情况。

移动设备或 AP 可以通过 IKEv2 进行 IPSec 协商，此时可以通过 EAP 进行用户认证，认证通过后给用户分配可以访问 VPN 网络的内网 IP 地址。通过 EAP 认证的移动设备或 AP 设备可以通过隧道安全访问总部服务器。

此种场景下，网关和 AP 设备都需要支持 IKEv2 方式的 IPSec，以及 EAP 认证功能。

典型组网如图 6-22 所示。

图6-22 采用 IKEv2 接入 IPSec 网关



6.4.8.6 IPSec NAT 穿越

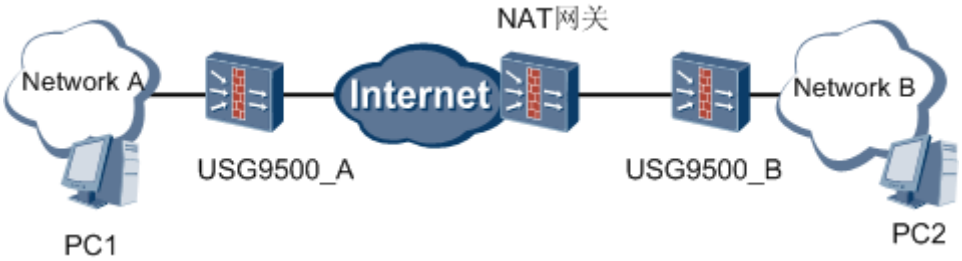
介绍 IPSec 隧道支持 NAT 穿越的场景，此场景和其他场景结合使用的情况较多。当建立 IPSec 隧道的两端之间存在 NAT 设备时，需要在两端 IPSec 网关上同时开启 NAT 穿越功能。

NAT 网关可能出现在如下几种位置：

- 分支网关侧，分支网络采用内网地址，经过 NAT 访问外网。
- 总部网关侧，总部网络为保护重要的服务器，采用 NAT Server 隐藏内部服务器的 IP 地址。
- 分支网关侧和总部网关侧都存在 NAT 网关。

NAT 网关在分支网络一侧的典型场景如图 6-23 所示。

图6-23 IPSec NAT 穿越



隧道中存在 NAT 网关时，若要建立 IPSec 隧道，就必须在 IKE 协商过程中发现两个端点之间的 NAT 网关，并使 ESP 报文可以正常穿越 NAT 网关。

- 若要发现两个端点间的 NAT 网关，必须进行 NAT 穿越能力协商。NAT 穿越能力协商在 IKE 协商的前两个消息中进行，通过 Vendor ID 载荷指明的一组数据来标示。
- IKE 通过 NAT-D(NAT Discovery)载荷来发现 NAT 网关，NAT-D 载荷用于两个目的，即发现 NAT 的存在，并确定 NAT 设备在 IKE peer 的哪一侧。

NAT 侧的 Peer 作为发起者，需要定期发送 NAT-Keepalive 报文，以使 NAT 网关确保安全隧道处于激活状态。

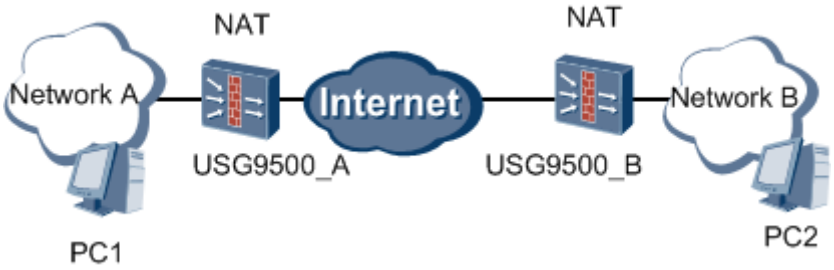
IPSec NAT 穿越的处理如下：要使 ESP 报文正常穿越 NAT 网关，需要在原报文的 IP 头和 ESP 头间增加一个标准的 UDP 报头。当 ESP 报文穿越 NAT 网关时，NAT 对该报文的外层 IP 头和增加的 UDP 报头进行地址和端口号转换；转换后的报文到达 IPSec 隧道对端时，与普通 IPSec 处理方式相同，但在发送响应报文时也要在 IP 头和 ESP 头之间增加一个 UDP 报头。

6.4.8.7 IPSec 网关同时作为 NAT 设备

介绍建立 IPSec 隧道的一端或两端网关同时作为 NAT 设备的应用场景。对访问外网和访问对端的数据流可以分别进行配置，互不影响。

IPSec 网关同时作为 NAT 设备时的典型组网如图 6-24 所示。IPSec 网关与 NAT 网关同为一个设备，此时不需要开启 NAT 穿越功能。

图6-24 IPSec 与 NAT 特殊组网



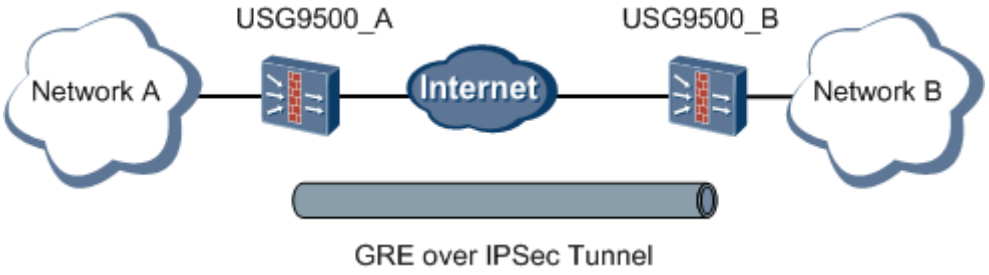
网络 A 和网络 B 之间通过 IPSec 隧道互相访问。同时，网络 A 和网络 B 还可以经过 NAT 转换后正常访问 Internet。在配置 NAT 时需要拒绝 IPSec 数据流，从而使此部分数据流不经过 NAT 而是通过 IPSec 隧道传输。

6.4.8.8 GRE over IPSec

介绍 IPSec 与 GRE 结合的场景，一般用来传输组播数据。

GRE over IPSec 解决了 IPSec 不支持组播、广播和非 IP 报文的缺点。对于这些报文数据，首先采用 GRE 进行封装，然后再进行 IPSec 封装。在 IPSec 隧道中就可以把 GRE 封装的这些报文当作普通的单播报文进行处理。

图6-25 GRE over IPSec 组网图



6.4.8.9 DHCP over IPSec

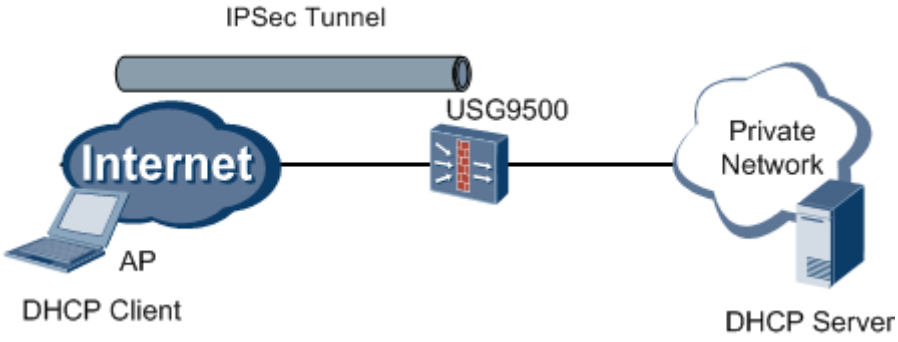
DHCP over IPSec 适用于远程移动设备访问公司内部网络的场景。

DHCP over IPSec 应用场景

如图 6-26 所示，AP 已经存在一个公网的 IP 地址，要访问公司内部网络可以通过向 DHCP 服务器发送请求获取相应的私有 IP 地址来实现。

此场景下网关和 AP 设备（DHCP Client）必须同时支持 DHCP over IPSec。

图6-26 DHCP over IPSec 典型应用场景

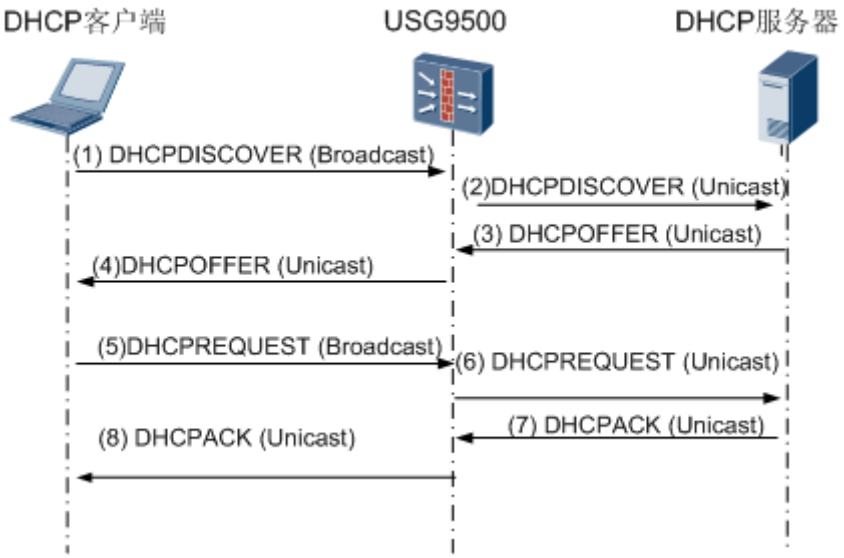


下面对 DHCP over IPSec 的报文交互过程和隧道建立过程进行说明。

报文交互过程

USG9500 不支持 DHCP 功能，但是支持 DHCP over IPSec，此时 USG9500 的作用类似于 DHCP 中继的功能。对此种情况下的 DHCP 交互过程说明如下。

图6-27 交互过程



1. DHCP 客户端发送 DISCOVER 广播报文给 USG9500。
2. USG9500 将此报文发送给 DHCP 服务器。
3. DHCP 服务器收到 DHCPDISCOVER 报文后，发送 DHCPOFFER 报文给 USG9500。
4. USG9500 将此报文发送给 DHCP 客户端。
5. DHCP 客户端发送 DHCPREQUEST 给 USG9500。
6. 由 USG9500 为 DHCP 客户端向 DHCP 服务器请求配置参数。
7. DHCP 服务器确认请求合法时，发送 DHCPACK 报文给 USG9500。
8. 通过 USG9500 将报文传给 DHCP 客户端。DHCP 客户端从 DHCPACK 报文中获取相应的配置参数，包括 IP 地址等。

DHCP over IPSec 的隧道建立过程

在 DHCP 客户端与 USG9500 之间会建立两种 IPSec SA。首先建立的 IPSec SA 此处称为 DHCP SA。说明如下：

- 建立 DHCP SA

1. 由 DHCP Client 发起协商，和 USG9500 建立 IKE SA。
2. Client 和 USG9500 建立 DHCP SA (IPSec SA)，保护 Client 和 DHCP 服务器信息的传输。
3. DHCP 四条消息加密交互，交互完毕后 DHCP Client 获得 DHCP Server 分配的地址以及其他信息。

- 建立 IPSec SA

1. DHCP Client 在获得 DHCP 信息之后，删除已建立的 DHCP SA，在以后需要更新 DHCP 消息时再建立新的 DHCP SA。
2. DHCP Client 根据分配的 IP 与 USG9500 协商一条新的 IPSec SA，用于传输数据。后续客户端发送的所有 DHCP 消息也都会通过新建立的 IPSec SA 进行传输。

USG9500 的作用

DHCP over IPSec 中，USG9500 的作用如下：

- USG9500 用 DHCP SA 解密客户端发送过来的 DHCP 报文，把广播报文转换成单播报文，修改 DHCP 报文内容后发送给 DHCP 服务器。
- 服务器对报文进行响应后，USG9500 把单播报文转换为广播报文或单播报文，修改 DHCP 报文内容后加密转发。
- 由于 DHCP relay 是一种无状态的转发，USG9500 必须加入适当的信息到 DHCP 报文里，才能对返回报文使用正确的 DHCP SA 加密，发送给相应的客户端。

6.4.8.10 IPSec 隧道化

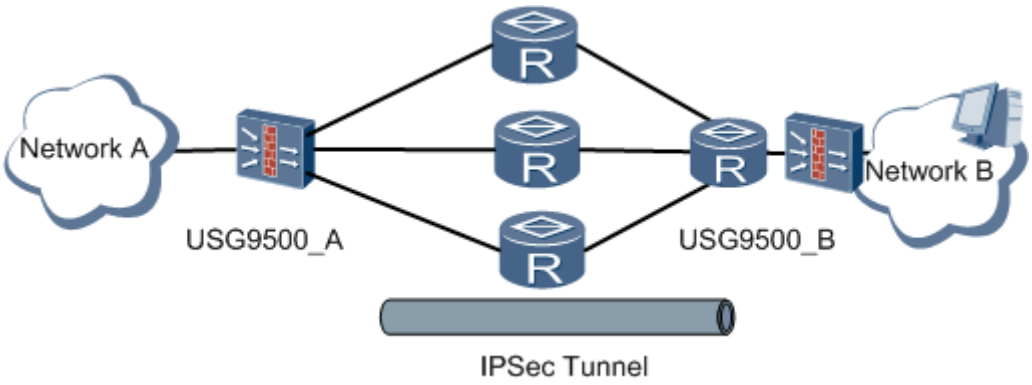
IPSec 隧道化就是将 IPSec 策略应用到虚拟隧道接口上。IPSec 隧道化可以实现出接口链接备份。

USG9500 除了支持直接在物理接口应用 IPSec 安全策略还支持在虚拟隧道接口上应用 IPSec 安全策略（也就是 IPSec 隧道化）。

IPSec 隧道化提供了一种简单的 IPSec 隧道端点的实现方式，IPSec 策略跟具体的物理接口没有任何绑定关系，将 IPSec 策略应用到 Tunnel 逻辑接口上，并且通过静态路由的方式将需要 IPSec 保护的报文引导到 Tunnel 接口进行 IPSec 处理。任何普通接口收到的报文都可以被引导到 Tunnel 接口，来实现在该 IPSec 隧道端点进行 IPSec 处理。经 IPSec 处理后的 IPSec 报文，通过路由来选择出接口，可以实现出接口的链路备份。

IPSec 隧道化的典型组网如图 6-28。

图6-28 IPSec 隧道化典型组网



配置 IPsec 隧道化具有以下优点：

- 简化配置

  通过配置路由可以将任何物理接口的报文引导到 Tunnel 接口进行 IPSec 处理。简化了 IPSec 策略的配置，并且使 IPSec 配置不会受到网络规划的影响，增强了网络规划的可扩展性。
- 业务应用更灵活

  IPSec 隧道化在实施过程中明确地区分出“加密前”和“加密后”两个阶段，用户可以根据不同的组网需求灵活选择其它业务（例如 NAT、QoS）实施的阶段。
- 增强了链路可靠性

  IPSec 隧道化可以实现出接口链路的备份，增强了链路的可靠性。

6.4.8.11 IPSec 双机热备

介绍 IPSec 特性在双机热备环境中的使用场景。

IPSec 双机热备适用于对链路可靠性要求高的场合。

如图 6-29 和图 6-30 所示，网络中使用两台 USG9500 进行双机热备配置，可以将 IPSec 的配置信息、隧道建立信息等从主用设备备份到备用设备上，保证当主用设备断开后隧道也不会拆除，提高了网络的可靠性。

图 6-29 为网关到网关模式下建立 IPSec 隧道，可靠性要求非常高的一端采用双机热备。

图6-29 IPSec 双机热备典型组网（与对端网关建立隧道）

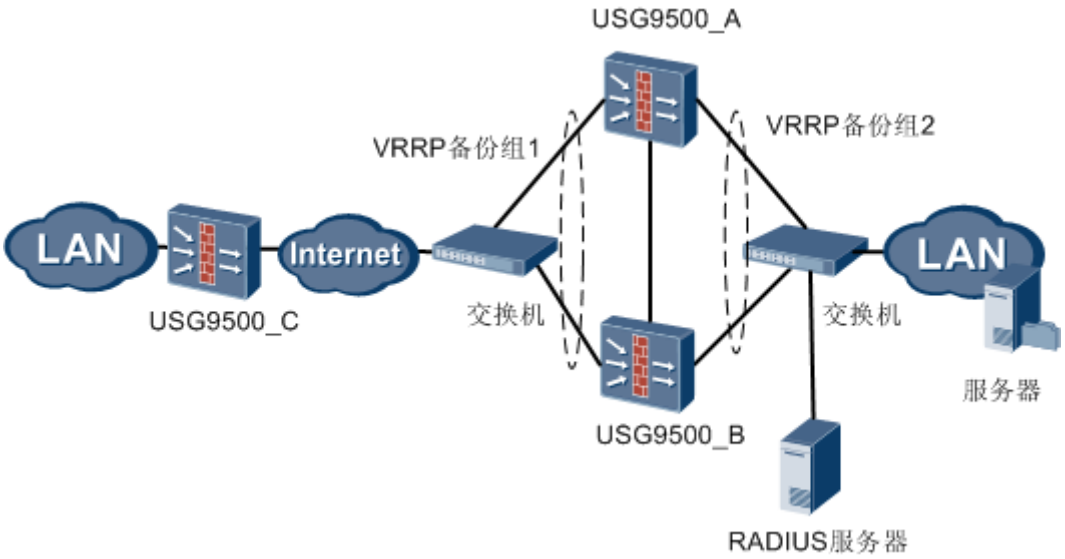
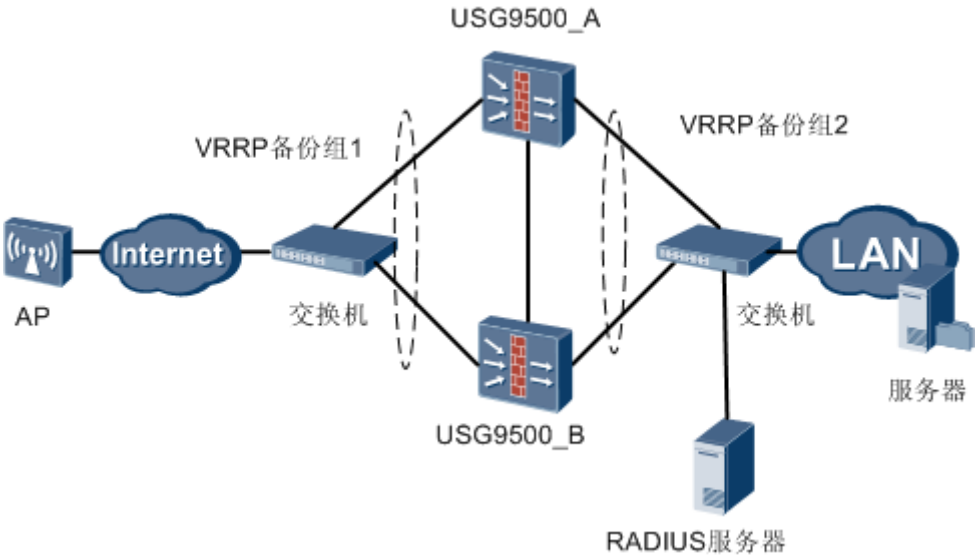


图 6-30 为 AP 设备接入 VPN 网络，VPN 的 IPSec 网关采用双机热备，提高可靠性。

图6-30 IPSec 双机热备典型组网（AP 设备远程接入）



此时 USG9500 A、USG9500 B 以及 AP 设备上都要使用 IKEv2 建立 IPSec 隧道，且都支持 EAP 认证功能。

AP 设备在通过 EAP 认证后，与参与双机热备的两台设备之间建立 IPSec 隧道，从而实现隧道两端内网设备通过隧道互相访问。在 USG9500 A、USG9500 B 其中之一断开后，不会影响隧道通信。

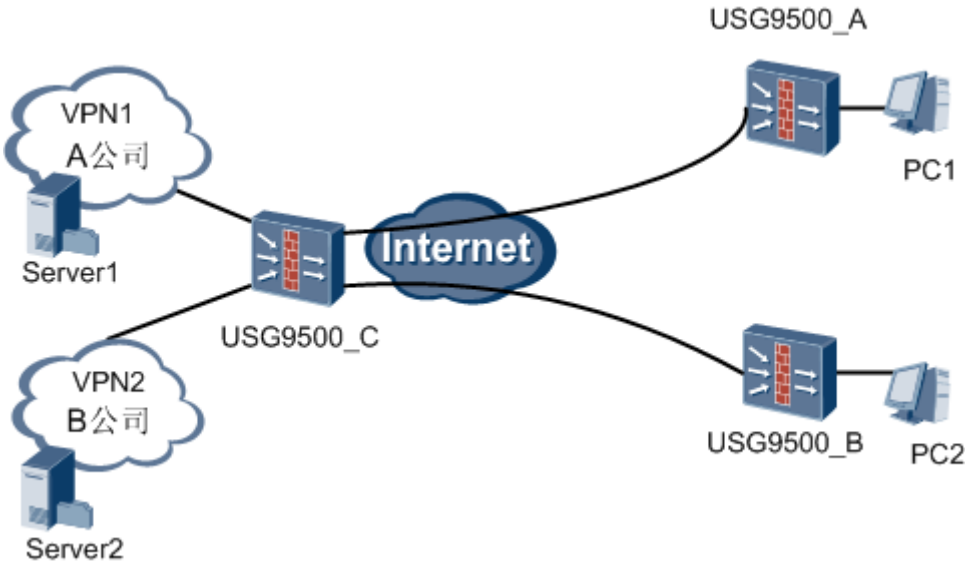
6.4.8.12 IPSec 多实例

介绍虚拟防火墙模式下 IPSec 的典型应用。

IPSec 多实例的典型应用场景如图 6-31 所示。其中 USG9500 C 划分为虚拟防火墙租用给不同的公司，每个公司的 VPN 网络保持独立。外地员工通过 IPSec 隧道接入到各自公司，IPSEC 隧道属于不同的 VPN 实例。

图 6-31 中 PC1 为 A 公司分部的员工，PC2 为 B 公司分部的员工。

图6-31 IPSec 多实例典型应用场景



虚拟防火墙作为独立的租用设备，各自规划的子网是独立的，有可能出现地址重叠的现象，即 VPN1 的网络和 VPN2 的网络中 IP 规划可能出现一样的情况，而这些相同的 IP 地址都需要 IPSEC 隧道进行保护。

6.4.8.13 IPSec 在 IPv6 中的应用

介绍 IPSec 在 IPv6 网络中的应用和限制。

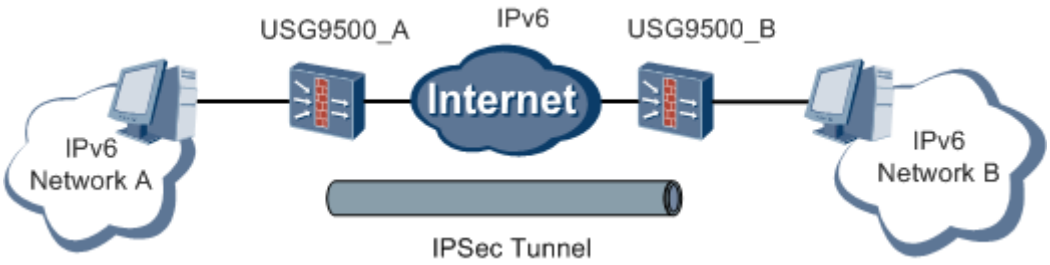
IPSec6 可以应用在纯 IPv6 网络中，即 IPSec 网关和被保护的网络都采用 IPv6 地址。

IPSec6 的功能限制请参见 6.4.2 规格。

IPSec6 的原理与 IPv4 IPSec 类似，不再进行说明。IPSec6 的封装模式与 IPv4 IPSec 类似，只不过其 IP 头为 IPv6 头。

IPv6 网络中的典型组网如图 6-32 所示。

图6-32 纯 IPv6 环境下网关到网关典型组网



7 证书

关于本章

7.1 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

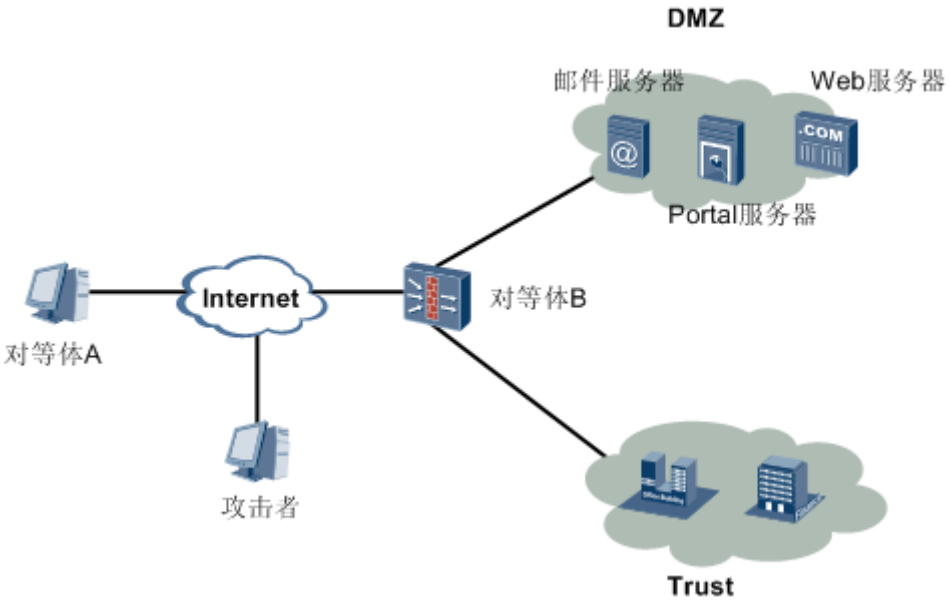
随着电子商务、网上银行及网上证券交易的飞速发展，Internet 的安全性越来越重要。部分不道德的个人会截取应用的明文数据，进行中间人攻击。

以图 7-1 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7-1 中间人攻击例子



受益

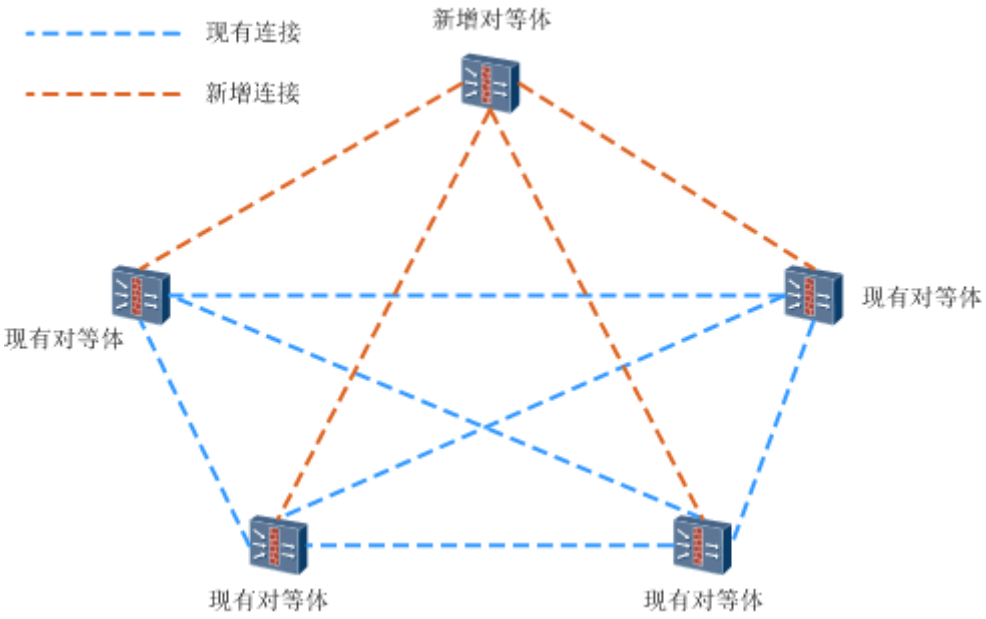
在 IPSec VPN 中，预共享密钥和证书是验证对等体身份的两种常用方法。

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 7-2 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7-2 使用预共享密钥验证方式示意图



7.2 规格

证书特性的相关规格如下：

- USG9500 支持创建的 PKI 实体的最大个数为 10 个。
- USG9500 支持创建的 PKI 域的最大个数为 64 个。
- USG9500 支持的 CA 证书的最大个数为 64 个。
- USG9500 支持的 CRL 的最大个数为 64 个。
- CA 证书文件最大不能超过 1M。
- 本地证书文件最大不能超过 1M。
- CRL 文件最大不能超过 1M。

7.3 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

7.4 可获得性

版本支持

产品	支持版本
HUAWEI Secoway USG9500	V200R001CO1

特性依赖

在 USG9500 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。

7.5 原理描述

7.5.1 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。

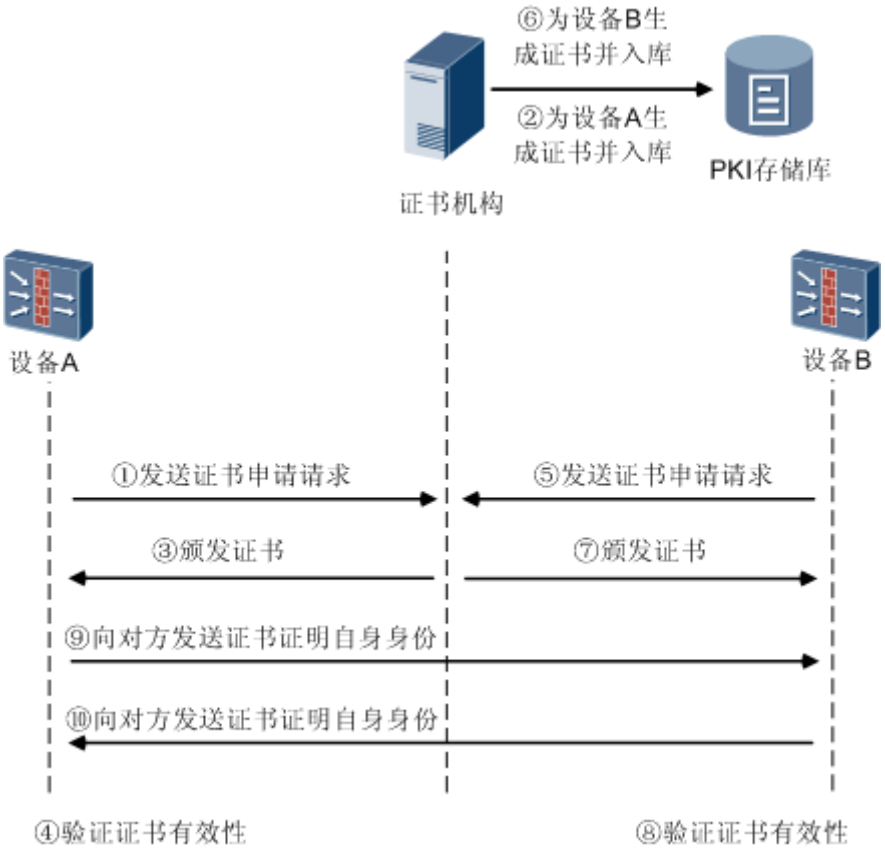
一个 PKI 体系由证书机构、终端实体、PKI 存储库等部分组成，如表 7-1 所示。

表7-1 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 USG9500。
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 7-3 所示。

图7-3 PKI 体系示意图



1. 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 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 B 建立实体证书。
6. CA 对收到的设备信息进行验证，为设备 B 生成证书并保存到 PKI 存储库里。
7. CA 将根证书和设备证书同时颁发给设备 B。
8. 设备 B 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。
9. 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。

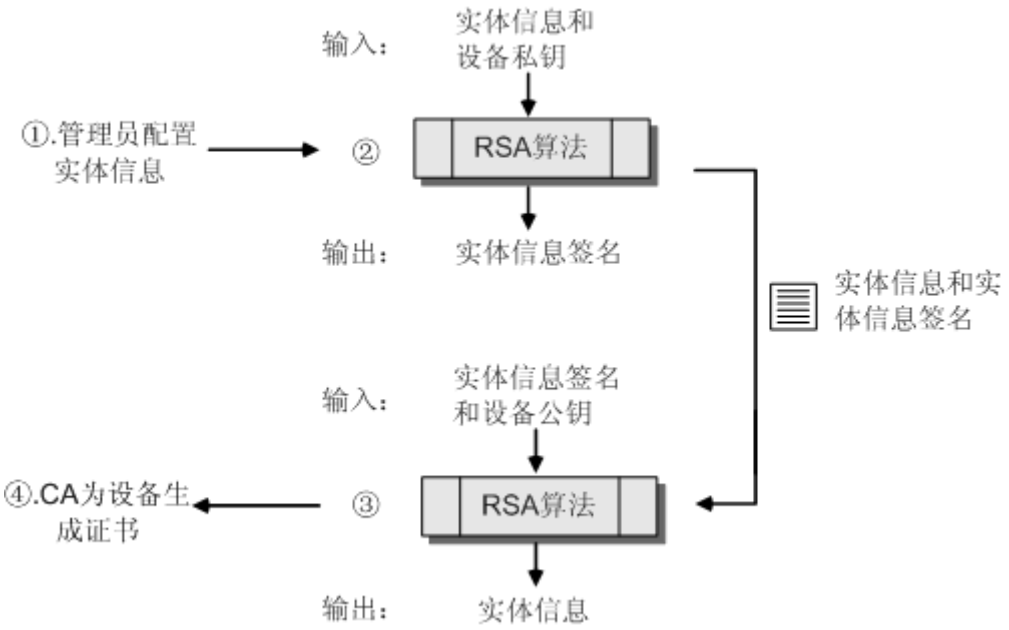
7.5.2 证书申请

CA 通过对表示设备特征的实体信息进行签名来为设备产生实体证书的，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。

证书申请过程如图 7-4 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7-4 证书申请过程示意



目前，USG9500 支持在线和离线两种方式申请证书：

- 在线方式（带内方式）

  USG9500 使用 SCEP（Simple Certification Enrollment Protocol）或 CMP（Certificate Management 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）

  USG9500 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。

7.5.3 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。

USG9500 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，USG9500 使用 SCEP 或 CMP 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，USG9500 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 USG9500 的 CF 卡中。

  然后还需要将证书导入到设备中才能生效。

7.5.4 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。

7.5.5 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。

7.6 证书应用

7.6.1 证书在 IPSec VPN 中的应用

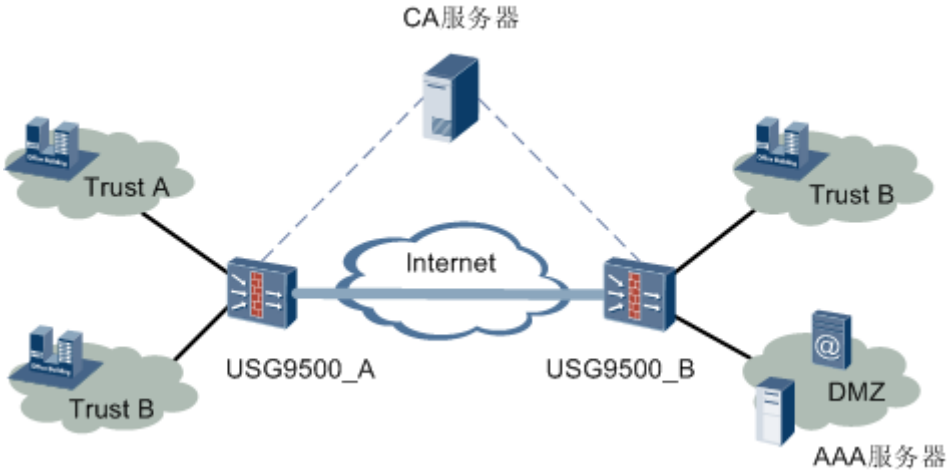
在站点到站点和远程访问的 VPN 应用中，当两台设备彼此通信时，用数字证书来检验身份。

以图 7-5 为例，USG9500_A 和 USG9500_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 USG9500_A 和 USG9500_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，它们将共享各自的实体证书。
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，USG9500_A 和 USG9500_B 就可以建立 IPSec VPN。

图7-5 证书在 IPSec VPN 中的应用



7.6.2 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 7-6 为例：

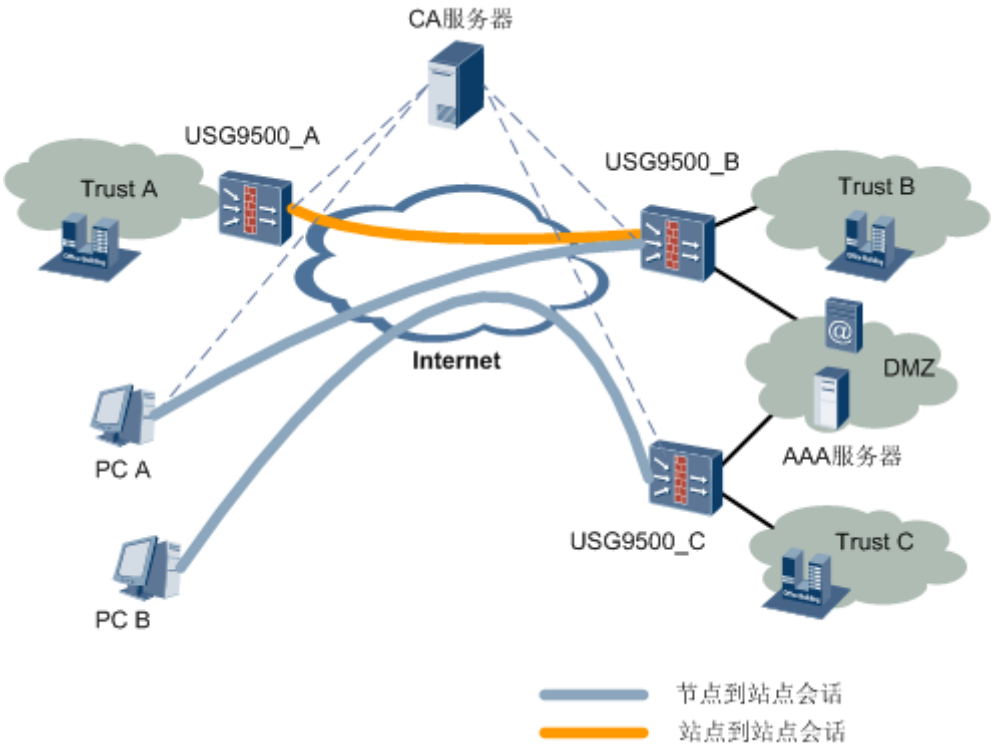
- USG9500_A 通过与 USG9500_B 建立站点到站点的 VPN 会话来实现区域 Trust A 与区域 Trust B 的互访，USG9500_A 和 USG9500_B 通过证书方式进行设备验证。
- PC A 通过与 USG9500_B 建立节点到站点的 VPN 会话来实现对区域 Trust B 的远程访问，其中，PC A 和 USG9500_B 通过证书方式来进行设备验证，通过 AAA 实现用户验证。
- PC B 通过与 USG9500_C 建立节点到站点的 VPN 会话来实现对区域 Trust C 的访问，PC B 和 USG9500_C 通过证书方式来进行设备验证，通过 AAA 实现用户验证。
- USG9500_B 和 USG9500_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 USG9500_B 和 USG9500_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 USG9500_B 和 USG9500_C 使用同一台 AAA 服务器，并且 USG9500_B 和 USG9500_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 USG9500_B 的 IPSec 远程访问会话。

为了在 USG9500_B 上限制 PC B 对 USG9500_B 发起的远程访问会话，可以在 USG9500_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。

USG9500 支持根据证书主题名、证书颁发者名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7-6 基于证书属性的 VPN 访问控制



8 IPS

关于本章

8.1 介绍

定义

IPS 是通过监控或者分析系统事件，检测入侵的发生，并通过一定的响应方式，实时地中止入侵行为的一种安全机制。

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 规格

IPS 特性的规格如下：

- 使用 Symantec 入侵检测和防御技术
- 支持分片重组
- 支持流重组
- 支持基于特征的检测和基于协议异常的检测
- 支持协议识别
- 支持自定义签名和预定义签名

- 支持细粒度的 IPS 策略设置
- 支持特权策略
- 支持直路和旁路的工作方式
- 支持签名按类别显示
- 支持签名的查询
- IPS 签名库的升级，包括自动升级、手动升级、本地升级、版本回退

8.3 可获得性

只有获取了支持 IPS 升级服务的 License，才能使用 IPS 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。

8.4 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，传统的防火墙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不受实质性的攻击。

IPS 处理流程

IPS 的基本处理流程如下：

1. 重组应用数据

   USG9500 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，USG9500 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

   USG9500 采用基于状态的匹配引擎，通过将报文内容与签名进行比较，能够准确识别出真实的应用层协议类型以及攻击行为。

4. 完成检测后，USG9500 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，USG9500 将报文内容和 IPS 签名进行比较，来检测和防范攻击。

USG9500 的签名分为两类：

- 预定义签名

  USG9500 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。USG9500 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。

USG9500 可以通过以下方式定制 IPS 策略：

- 引用模板

  USG9500 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。如果模板内容可满足安全需求，用户可以直接在 IPS 策略中引用模板，从而减少配置工作。
- 配置签名集

  签名集是满足指定过滤条件的签名的集合。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。
- 配置覆盖签名

  当需要对某个签名配置指定的启用状态和响应方式时，可以采用在 IPS 策略中配置覆盖签名的方式实现。自定义签名必须通过覆盖签名方式应用到 IPS 策略中。

说明
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时以覆盖签名为准。

完成 IPS 策略配置后，需要将 IPS 策略应用在指定的域内或域间才能使 IPS 功能生效。

USG9500 还支持从已有的 IPS 策略中选取一个作为特权策略。特权策略替换所有已经在域内或域间应用的 IPS 策略，没有应用 IPS 策略的域内或域间不添加特权策略。

IPS 策略经过编译后才能生效，USG9500 支持手动提交配置来编译 IPS 策略。另外，以下三种情况也会触发 IPS 策略的编译，而且与手动提交配置的编译效果相同。

- 域内或域间应用 IPS 策略（当修改后未提交配置进行编译时）

- 配置特权策略或取消已配置的特权策略（当修改后未提交配置进行编译时）
- 升级 IPS 版本（一定会触发 IPS 策略的编译）

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，USG9500 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。

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。

IPS 攻击响应方式如表 8-1 所示。

表8-1 攻击响应方式

处理策略	工作模式	实际动作
告警	防护模式	USG9500 不对文件进行处理，记录日志。
	告警模式	
阻断	防护模式	USG9500 阻断文件，记录日志。
	告警模式	USG9500 不对文件进行处理，记录日志。

9 DPI

关于本章

9.1 介绍

定义

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，可以识别几乎所有的应用层业务，并对指定类型的数据流量进行控制。通过分析 USG9500 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。

9.2 可获得性

License 支持

本特性无须 License 支持。

版本支持

产品	支持版本
HUAWEI Secoway USG9500	V200R001CO1

9.3 原理描述

9.3.1 DPI 起源

随着互联网的发展，服务提供商提供的服务越来越多样化，随之也给企业带来了新的网络问题。如：

- 过多的 P2P 流量将占用大量带宽，影响公司业务的开展。
- 工作时间段内的 IM 的过度使用，降低了工作效率，并给企业带来了一定的安全隐患。

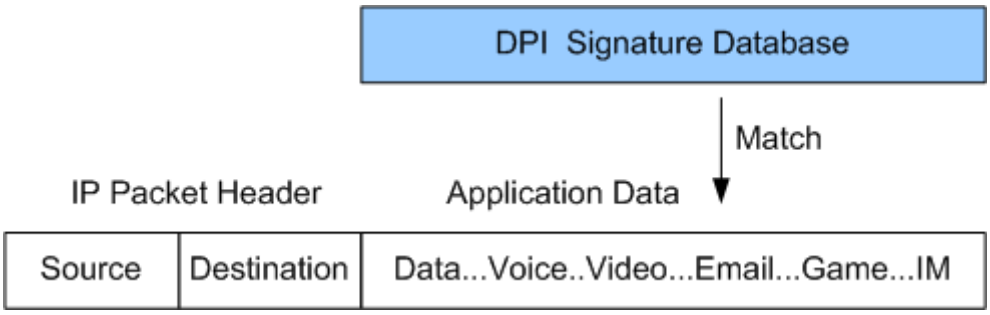
P2P 下载、流媒体等新的应用随着互联网新趋势而兴起，传统的基于数据流的分析技术，如 NetFlow 技术已无法准确识别出网络流量中的应用协议的类型。

由于这种需求的驱使，DPI 技术应运而生。

9.3.2 DPI 工作原理

原理概述

图9-1 原理示意图



如图 9-1 所示，设备根据 DPI 特征库，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，从而确定应用流对应的应用或者用户的动作。

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 特征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。

自定义规则

除了使用 DPI 特征库之外，还可以使用自定义规则来识别网络中的流量。

在实际网络环境中，可能会存在某些流量没有包含于 DPI 特征库之中，使用 DPI 特征库无法准确识别的情况。此时，在了解流量特征的前提下，用户可以创建自定义规则匹配这些流量，更加灵活地实施 DPI 控制策略。

9.3.3 关联协议识别

原理概述

对于某些数据流，控制通道和数据通道是分开的（如 FTP、SIP、H.323 等），将会在网络中建立两个会话连接，一般的协议识别方式由于对两个会话分开进行识别，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。

9.3.4 全包检测

一般情况下，DPI 对报文的特定部位进行协议特征检测，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。

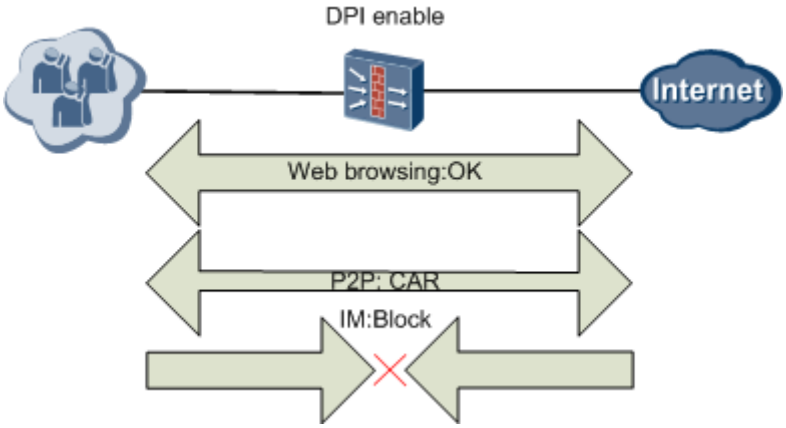
USG9500 支持对指定的协议启用全包检测功能。详细了解检测哪些协议需要启用此项功能，请查看《*配置指南 DPI*》中的应用协议速查表。

9.3.5 DPI 的应用

通过使用 DPI，可以对网络中检测到的各种类型的应用协议加以控制，确保信息安全，保证正常业务的运行。

在实际应用中，可以有多种控制效果。如图 9-2 所示，对于一般的网络浏览予以放行，对于 P2P 类型等影响带宽的应用程序做速率限制。在某些不允许使用 IM 类型应用程序的情况下可以进行阻断。

图9-2 DPI 应用示意图



10 QoS

关于本章

10.1 介绍

定义

服务质量 QoS（Quality of Service）概念普遍存在于各种拥有服务供需关系的场合中，它评估服务方满足客户服务需求的能力。

在 Internet 中，QoS 所评估的是网络投递分组的服务能力。由于网络提供的服务是多样的，因此对 QoS 的评估可以基于不同方面。通常所说的 QoS，是对分组投递过程中为延迟、延迟抖动、丢包率等核心需求提供支持的服务能力的评估。

目的

传统的 IP 网络无区别地对待所有的报文，设备采用先入先出 FIFO（First In First Out）的策略处理报文，依照报文到达时间的先后顺序分配转发所需要的资源。这种服务策

略称作 Best-Effort（尽力而为），它尽最大的努力将报文送到目的地，但对分组投递的延迟、延迟抖动、丢包率和可靠性等需求不提供任何承诺和保证。

新业务的不断涌现对 IP 网络的服务能力提出了更高的要求，用户已不再满足于能够简单地将报文送达目的地，而是还希望在投递过程中得到更好的服务。诸如为用户提供专用带宽、减少报文的丢失率、管理和避免网络拥塞、调控网络的流量、设置报文的优先级等服务。

QoS 就是针对各种不同的需求，提供不同的服务质量。

实现

QoS 根据网络质量和用户需求，通过不同的服务模型为用户提供服务。通常 QoS 提供以下三种服务模型：

- Best-Effort Service 模型

  尽力而为服务模型。适用于对时延、可靠性等性能要求不高的业务质量保证，它通过先入先出（FIFO）队列来实现。Best-Effort Service 模型是现在 Internet 的缺省服务模型，它适用于绝大多数网络应用，如 FTP、E-Mail 等。

- Integrated Service 模型

  综合服务模型。应用程序在发送报文前，向网络申请特定的服务，确认网络已经为这个应用程序的报文预留了资源后开始发送报文。

- Differentiated Service 模型

  区分服务模型。应用程序在发送报文前不必预先向网络提出资源申请，通过设置 IP 报文头部的 QoS 参数信息，来告知网络节点它的 QoS 需求。

实现区分服务的主要技术包括：

- 流分类

  依据一定的匹配规则识别出对象。流分类是有区别地实施服务的前提。

- 流量监管

  对进入网络的特定流量的速率进行监管，当流量超出规格时，采取限制或惩罚措施。只有流量的速率在合理的范围内才可以进入网络，以保护用户的网络资源不受损害。

- 流量整形

  一种主动调整流量输出速率的流控措施，通常是为了使流量适配下游设备可供给的网络资源，避免不必要的报文丢弃和拥塞。

- 拥塞管理

  网络拥塞时必须采取的解决资源竞争的措施。通常是将报文放入队列中缓存，并采取某种调度算法安排报文的转发次序。

- 拥塞避免

  拥塞避免监控网络资源的使用情况，当发现拥塞有加剧的趋势时采取主动丢弃报文的策略，通过调整流量来解除网络的过载。

10.2 规格

QoS 特性的相关规格如下：

- USG9500 支持简单流分类和复杂流分类。
- USG9500 支持流量监管、流量整形、策略路由、重标记 IPv4 报文的 Precedence 值、DSCP 值和 VLAN 报文的 802.1p 值、将 IPv4 网络中报文的 DSCP 值映射为 VLAN 网络中报文的 802.1p 值。
- USG9500 支持创建的流分类的最大个数为 4096 个。
- USG9500 支持创建的流行为的最大个数为 4096 个。
- USG9500 支持创建的流量策略的最大个数为 1024 个。
- 每个流量策略中最多支持配置 255 条流分类和流行为的关联。
- USG9500 支持 HQoS 上下行各五级调度。
- USG9500 支持创建的流队列 WRED 对象的最大个数为 511 个。
- USG9500 支持创建的流队列模板的最大个数为 2048 个。
- USG9500 支持创建的流队列到类队列映射模板的最大个数为 15 个。
- USG9500 支持创建的用户组队列的最大个数为 8191 个。
- USG9500 支持创建的类队列 WRED 对象的最大个数为 7 个。

10.3 参考标准和协议

与 QoS 特性相关的参考标准与协议如下：

- RFC 791
  Internet Protocol
- RFC 1122
  Requirements for Internet Hosts - Communication Layers
- RFC 1349
  Type of Service in the Internet Protocol Suite
- RFC 2474
  Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers
- RFC 2597
  Assured Forwarding PHB
- RFC 2598
  Expedited Forwarding PHB
- RFC 2697
  A Single Rate Three Color Marker
- RFC 2698
  A Two Rate Three Color Marker
- TR-059
  HQoS Model of the DSL Forum

- IEEE 802.1Q

  IEEE Standards for Local and Metropolitan Area Networks: Virtual Bridged Local Area Networks

10.4 可获得性

License 支持

本特性无须 License 支持。

版本支持

产品	支持版本
HUAWEI Secoway USG9500	V200R001CO1

10.5 流分类

在采用 Diff-Serv 模型实施 QoS 时，需要设备识别各种流，因此需要对报文进行流分类。根据分类规则的不同，流分类分为简单流分类 BA（Behavior Aggregation）和复杂流分类 MF（Multiple Field）。

- 简单流分类

  简单流分类是指采用简单的规则，如只根据 IPv4 报文的 DSCP（Differentiated Services CodePoint）值、VLAN 报文的 802.1p 值对报文进行粗略的分类，以识别出具有不同优先级或服务等级特征的流量，属于同一分类的报文集合称为 BA。通常在 Diff-Serv 模型中的核心设备上对流量进行简单流分类。

- 复杂流分类

  复杂流分类是指采用复杂的规则，例如 IPv4/IPv6 报文的源地址、目的地址、协议类型或应用程序的 TCP/UDP 端口号等，对报文进行精细的分类。通常在 Diff-Serv 模型中的边界设备上对流量进行复杂流分类。

10.6 流量监管和整形

在 Diff-Serv 模型中，流量监管与整形功能由流量控制器（Traffic Conditioner）来完成。流量控制器由四个部件构成：Meter、Marker、Shaper 和 Dropper，如图 10-1 所示。

- Meter

  测量流量，判断信息流是否遵循流量规格的定义。设备根据流量测量的结果，通过 Marker、Shaper 和 Dropper 实施动作。

- Marker

重新标记（Re-marking）报文的优先级字段，并将重新标记过的报文放入特定的BA 中。

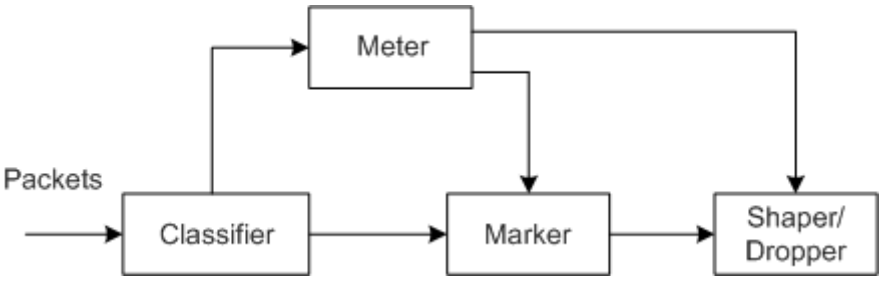
- Shaper

  即流量整形器，它具有一定的缓冲区，控制发出报文的速率不超出承诺的规格。

- Dropper

  流量监管中的动作，通过丢弃一些报文来控制流量使其符合流量规格。

图10-1 流量控制器结构图



10.7 拥塞管理和避免

传统网络所面临的服务质量问题，主要是由网络拥塞引起的。所谓拥塞，是指由于供给资源的相对不足而造成服务速率下降（引入了额外的延迟）的一种现象。

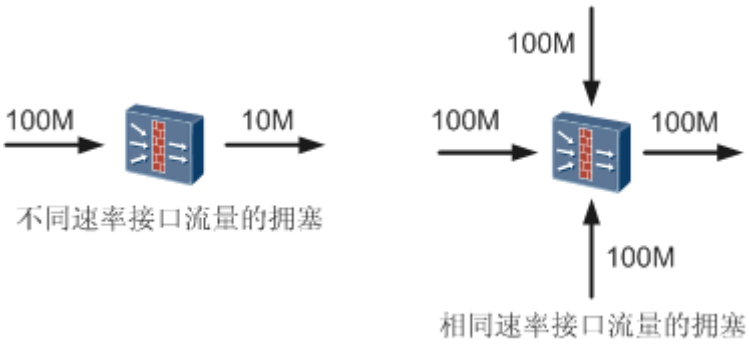
拥塞的产生

在 Internet 分组交换的复杂环境下，拥塞极为常见。以图 10-2 两种情况为例：

- 分组流从高速链路进入设备，由低速链路转发出去。
- 分组流从相同速率的多个接口同时进入设备，由一个相同速率的接口转发出去。

如果流量以线速到达，那么就会遭遇资源的瓶颈而导致拥塞。

图10-2 流量拥塞示意图



不仅仅是链路带宽的瓶颈会导致拥塞，任何用以正常转发处理的资源的不足，如可分配的处理器时间、缓冲区、内存资源的不足，都会造成拥塞。此外，在某个时间内对所到达的流量控制不力，使之超出了可分配的网络资源，也是引发网络拥塞的一个因素。

拥塞的影响

拥塞有可能会引发一系列的负面影响：

- 拥塞增加了报文传输的延迟和延迟抖动。
- 过高的延迟会引起报文重传。
- 拥塞使网络的有效吞吐率降低，造成网络资源的损害。
- 拥塞加剧会耗费大量的网络资源（特别是存储资源），不合理的资源分配甚至可能导致系统陷入资源死锁而崩溃。

可见，拥塞使流量不能及时获得资源，是造成服务性能下降的源头。然而在分组交换以及多用户业务并存的复杂环境下，拥塞又是常见的。因此采取有效的避免拥塞以及防止拥塞加剧的方法是必需的。

拥塞管理和对策

拥塞管理的处理包括队列的创建和报文的分类，将报文送入不同队列，队列调度等。通过采用排队技术，使得报文在设备中按一定的策略暂时排队，然后再按一定的调度策略把报文从队列中取出，在接口上发送出去。

拥塞避免一般通过丢弃报文来实现。在拥塞发生和拥塞加剧时，通过采用特定的分组丢弃策略，为属于不同转发业务类别的流量权衡资源的分配。常用的分组丢弃策略包括：

- 尾丢弃（Tail Drop）

  尾丢弃是在队列满之后，最后到达的报文将被丢弃。
- RED（Random Early Detection）

  RED 算法在队列到达一定长度后开始随机丢弃，它可以避免由于 TCP 的慢启动机制导致的全局同步现象。
- WRED（Weighted Random Early Detection）

  与 RED 相比，WRED 在丢弃报文时，考虑了队列长度和报文的优先级，低优先级较早开始丢弃且丢弃概率较大。

10.8 优先级重标记

优先级重标记实现了 IPv4 报文中 Precedence 值和 DSCP 值、VLAN 报文中 802.1p 值的重新标记。设备根据源地址、目的地址、协议类型或应用程序的 TCP/UDP 端口号等信息对 IPv4 报文或 VLAN 报文进行复杂流分类，为不同分类的报文设置新的 Precedence 值、DSCP 值或 802.1p 值。

通常 Diff-Serv 模型中的边界设备需要对进入网络的报文进行优先级重标记，核心设备只需按照边界设备所标记的优先级提供相应等级的服务即可，或者按自己的标准重新进行标记。

10.9 优先级映射

优先级映射实现外部优先级和内部优先级之间的映射。设备根据 IPv4 报文的 DSCP 值、VLAN 报文的 802.1p 值对报文进行简单流分类，建立不同网络之间报文优先级的映射关系。

优先级映射分为上行和下行两种情况，如图 10-3 所示。

- 上行映射

  根据 DSCP 值或 802.1p 值将 IPv4 报文或 VLAN 报文分为八种业务等级 CoS（Class of Service），包括 CS7、CS6、EF、AF4～AF1、BE）；三种颜色，包括 green、yellow、red。其中，当报文的业务等级为 EF、BE、CS6 或者 CS7 时，报文只能标记为绿色。

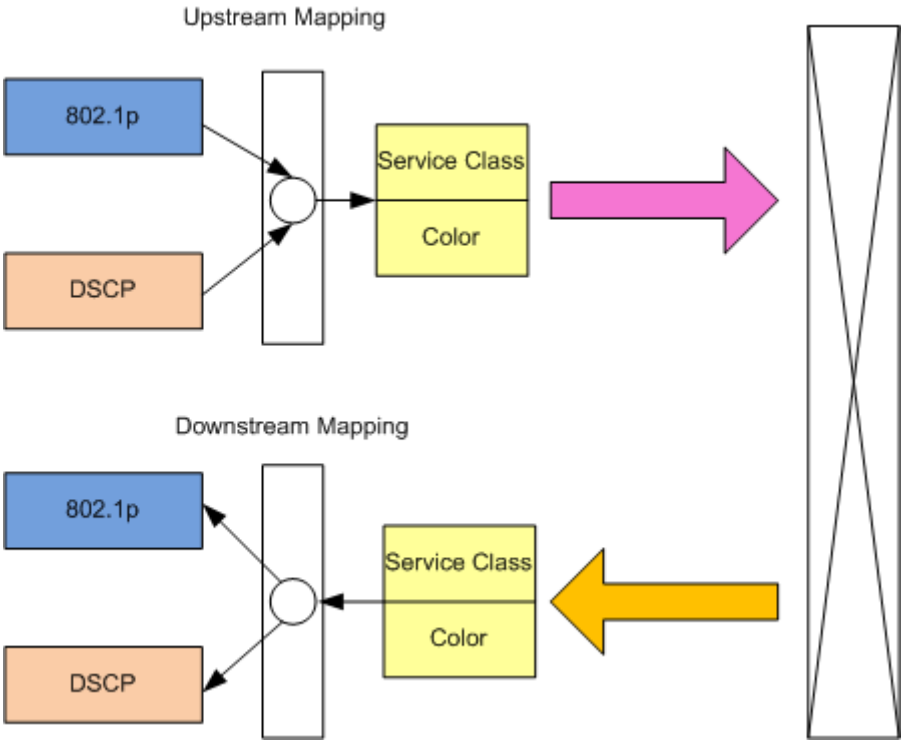
  通过上行映射区分不同的业务（如语音、视频、数据等）。拥塞管理、队列调度时，不同业务进入不同的队列，从而得到差异化的调度。例如语音可以进入高优先级的队列，保证低延时。

- 下行映射

  根据内部业务等级（CS7、CS6、EF、AF4～AF1、BE）、三种颜色（green、yellow、red），重新设置 IPv4 报文的 DSCP 值或 VLAN 报文的 802.1p 值。

  通过下行映射实现了重标记的功能，重新标记 IPv4 报文的 DSCP 值或 VLAN 报文的 802.1p 值。

图10-3 上下行映射关系图



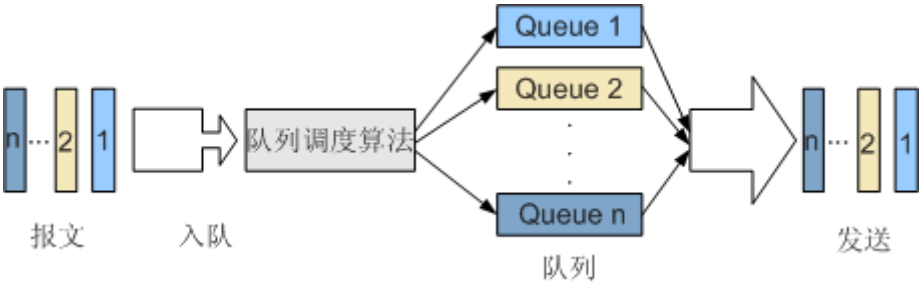
10.10 HQoS

HQoS（Hierarchical Quality of Service）即层次化 QoS，是一种通过队列调度机制，解决 Diff-Serv 模型下多用户多业务带宽保证的技术。

概述

传统的 QoS 基于端口进行流量调度。单个端口只能区分业务优先级，无法区分用户和不同用户的业务。只要属于同一优先级的流量，使用同一个端口队列，彼此之间竞争同一个队列资源，无法对端口上单个用户的单个流量进行区分服务。如图 10-4 所示。

图10-4 传统 QoS 的队列调度原理图



例如：有两个用户同时发送 AF4 的流量，用户 1 发送 10M，用户 2 发送 1G。但 AF4 的流量限速为 10M。传统的 QoS 不区分用户，由于用户 2 发送的 AF4 流量大，用户 2 的报文有很大几率进入队列，而用户 1 的报文则被丢弃的概率非常大。因此用户 1 的流量就受到了其它用户的影响。

随着网络用户数量的持续增长和网络业务的不断丰富，用户都希望能够提供区分用户和用户业务的服务，以获得更好的服务质量和更多的利润。HQoS 既能为高级用户提供精细化的服务质量保证，又能够从整体上节约网络运行维护成本，具有很高的市场需求。

基本概念

HQoS 中的基本概念包括：流队列、用户队列、用户组队列、类队列。

- 流队列 FQ（Flow Queue）

  HQoS 可以针对每个用户的业务流进行队列调度，每个用户的业务都可以细分为 8 个流队列。每个流队列可以配置 PQ（Priority Queue）、WFQ（Weighted Fair Queue）或 LPQ（Low Priority Queue）队列调度方式，还可以配置 WRED 丢弃机制以及流量整形的速率。

  调度的时候，优先调度 PQ 队列的报文，其次是 WFQ，最后是 LPQ。

- 用户队列 SQ（Subscriber Queue）

  SQ 为虚拟队列，队列中不存在实际的缓存单元，数据进入和离开队列没有延迟。

  每个 SQ 对应 8 种 FQ 业务等级，可配置 1～8 个 FQ。实际应用中，一个 SQ 对应一个用户，一个用户可使用 1～8 个 FQ。每个 SQ 可定义其 CIR（Committed Information Rate）和 PIR（Peak Information Rate）。

- 用户组队列 GQ（Group Queue）

  HQoS 可以将多个 SQ 绑定到一个 GQ 实现第三级队列调度。GQ 也是虚拟队列，每个 SQ 最多只能绑定到一个 GQ 内，也可以不绑定 GQ，跨过第三级队列调度。

  GQ 用来对多个用户的流量进行整体限速，其 PIR 值建议不要小于 GQ 中所有 SQ 的 CIR 之和，否则单个用户（SQ）的流量就无法得到保证。

- 类队列 CQ（Class Queue）

  HQoS 调度时，流队列报文经过用户队列调度后，要同普通报文同时进入接口中的 CQ。流队列报文入 CQ 时，可以有两种优先级映射方式：

  - Uniform 模型

    SQ 中 8 个等级的 FQ 与同接口的 8 个 CQ 有系统预定义的映射关系。

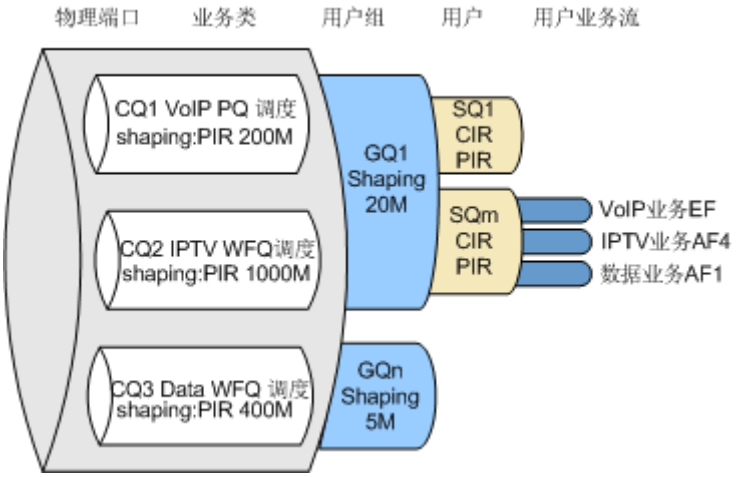
  - Pipe 模型

    SQ 中 8 个等级的 FQ 与同接口的 8 个 CQ 的映射关系可以由用户自行配置，但 Pipe 模型不会改变报文中自身携带的优先级。

实现

HQoS 通过上下行各五级调度的方式，来实现更加精细化的调度，为用户 QoS 业务层面提供丰富的业务支撑。如图 10-5 所示。

图10-5 HQoS 的调度原理图



HQoS 的部署需求可能在用户侧，也可能在网络侧，因此设备在上行转发和下行转发均实现了 HQoS 功能。针对不同的用户需求，可在不同的位置上进行配置，选择单独实施 HQoS 或联合实施 HQoS。

说明

配置 HQoS 时，建议同时配置基于简单流分类的优先级映射功能，使报文进入不同的优先级队列进行调度。

- 上行 HQoS

  上行 HQoS 分为五级调度。
  - 第一级调度：流队列 FQ
  - 第二级调度：用户队列 SQ
  - 第三级调度：用户组队列 GQ
  - 第四级调度：目的板 TB（Target Blade）
  - 第五级调度：类队列 CQ
- 下行 HQoS

  下行 HQoS 分为五级调度。
  - 第一级调度：流队列 FQ
  - 第二级调度：用户队列 SQ
  - 第三级调度：用户组队列 GQ
  - 第四级调度：类队列 CQ
  - 第五级调度：目的端口 TP（Target Port）

11 IPv6

关于本章

11.1 介绍

定义

IPv6（Internet Protocol Version 6）是网络层协议的第二代标准协议，也被称为 IPng（IP Next Generation）。它是 IETF（Internet 工程任务组）设计的一套规范，是 IPv4（Internet Protocol Version 4）的升级版本。IPv6 和 IPv4 之间最显著的区别就是 IP 地址长度从原来的 32 位升级为 128 位。

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。

11.2 规格

IPv6 特性的相关规格如下：

- 支持 IPv6 重定向报文。

- 支持 IPv6 基本报文头和扩展报文头。
- 支持发送差错和信息 ICMPv6 报文。
- 支持 ICMPv6 校验和。
- 支持自动/手动配置 Link-Local 地址。
- 支持动态 PMTU 表项老化。
- 支持基本和高级 ACL6。
- 支持基于 IPv6 地址的 Ping、Tracert 和 Telnet。
- 支持 IPv6 over IPv4 隧道。
- 支持 IPv4 over IPv6 隧道。
- 支持 6RD 隧道。
- 支持 DS-Lite 特性。
- 支持 NAT64 特性。

11.3 参考协议和标准

本特性的参考资料清单如下：

- RFC793：Transmission Control Protocol
- RFC768：User Datagram Protocol
- RFC1981：Path MTU Discovery for IP version 6
- RFC2461：Neighbor Discovery for IP Version 6 (IPv6)
- RFC2463：Internet Control Message Protocol for the Internet Protocol Version 6 Specification
- RFC2465：Management Information Base for IP Version 6:Textual Conventions and General Group
- RFC2466：Management Information Base for IP Version 6:ICMPv6 Group
- RFC2893：Transition Mechanisms for IPv6 Hosts and Routers
- RFC3056：Connection of IPv6 Domains via IPv4 Clouds
- RFC4214：Intra-Site Automatic Tunnel Addressing Protocol(ISATAP)

11.4 可获得性

License 支持

DS-Lite、NAT64 和 6RD 等功能受 License 控制，未加载 License 情况下，USG9500 不支持 DS-Lite、NAT64 和 6RD 等功能，并且相关命令行不可见。

版本支持

产品	支持版本
HUAWEI Secoway USG9500	V200R001CO1

11.5 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 IPv6 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：0:0:0:0:0:0: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：0:0:0:0:0: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。

IPv6 地址的结构

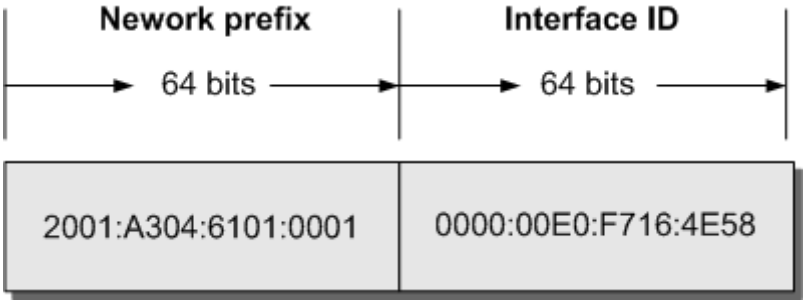
一个 IPv6 地址可以分为如下两部分：

- 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 11-1 所示。

图11-1 地址 2001:A304:6101:1::E0:F726:4E58 /64 的构成示意图



IPv6 的地址分类

IPv6 主要有三种类型的地址：单播地址、组播地址和任播地址。

- 单播地址：用来唯一标识一个接口，类似于 IPv4 的单播地址。发送到单播地址的数据报文将被传送给此地址所标识的接口。
- 组播地址：用来标识一组接口（通常这组接口属于不同的节点），类似于 IPv4 的组播地址。发送到组播地址的数据报文被传送给此地址所标识的所有接口。
- 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报文被传送给此地址所标识的一组接口中距离源节点最近（根据使用的路由协议进行度量）的一个接口。

说明
IPv6 中没有广播地址，广播地址的功能通过组播地址来实现。

IPv6 地址类型是由地址前面几位（称为格式前缀）来指定的，主要地址类型与格式前缀的对应关系如表 1 所示。

表11-1 IPv6 单播地址类型

地址类型		二进制前缀	IPv6 前缀标识
单播地址	链路本地单播地址	1111111010	FE80::/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	-
组播地址		11111111	FF00::/8
任播地址		从单播地址空间中进行分配，使用单播地址的格式	

表中各类地址的意义如下：

IPv6 单播地址的类型可有多种，包括全球单播地址、链路本地地址和站点本地地址等。

- 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上节点之间的通信。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。

- 环回地址：

  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

  地址“::”称为未指定地址，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

  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 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。

- 组播地址

  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。

  表 11-2 所示的组播地址，是预留的特殊用途的组播地址。

表11-2 预留的 IPv6 组播地址列表

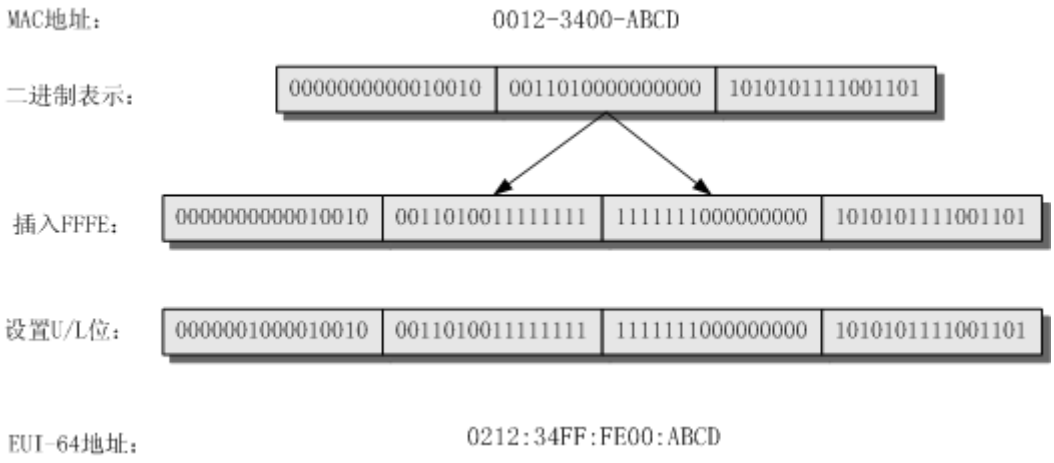
地址	应用
FF01::1	节点本地范围所有节点组播地址
FF02::1	链路本地范围所有节点组播地址
FF01::2	节点本地范围所有路由器组播地址
FF02::2	链路本地范围所有路由器组播地址
FF05::2	站点本地范围所有路由器组播地址

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。具体转换过程如图 11-2。

图11-2 MAC 地址到 EUI-64 格式接口标识符的转换过程



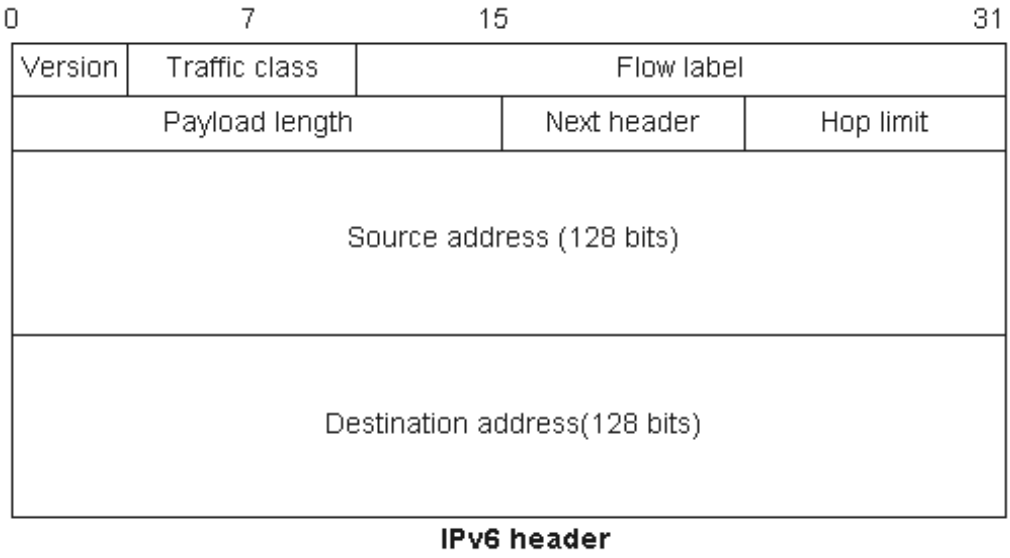
11.6 IPv6 报文格式

IPv6 的报文头格式

IPv6 报文的头部信息和一般的 IP 报文（即 IPv4 报文）有一定差异。

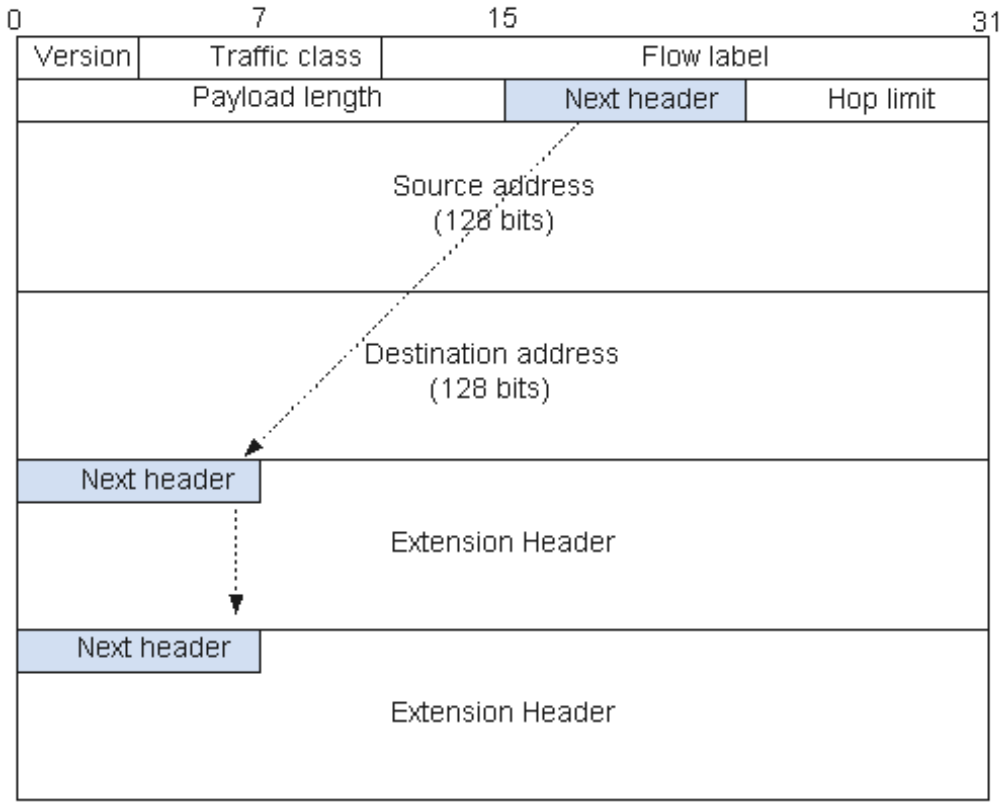
图 11-3 所示为 IPv6 报文头结构。

图11-3 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 11-4 所示。

图11-4 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。

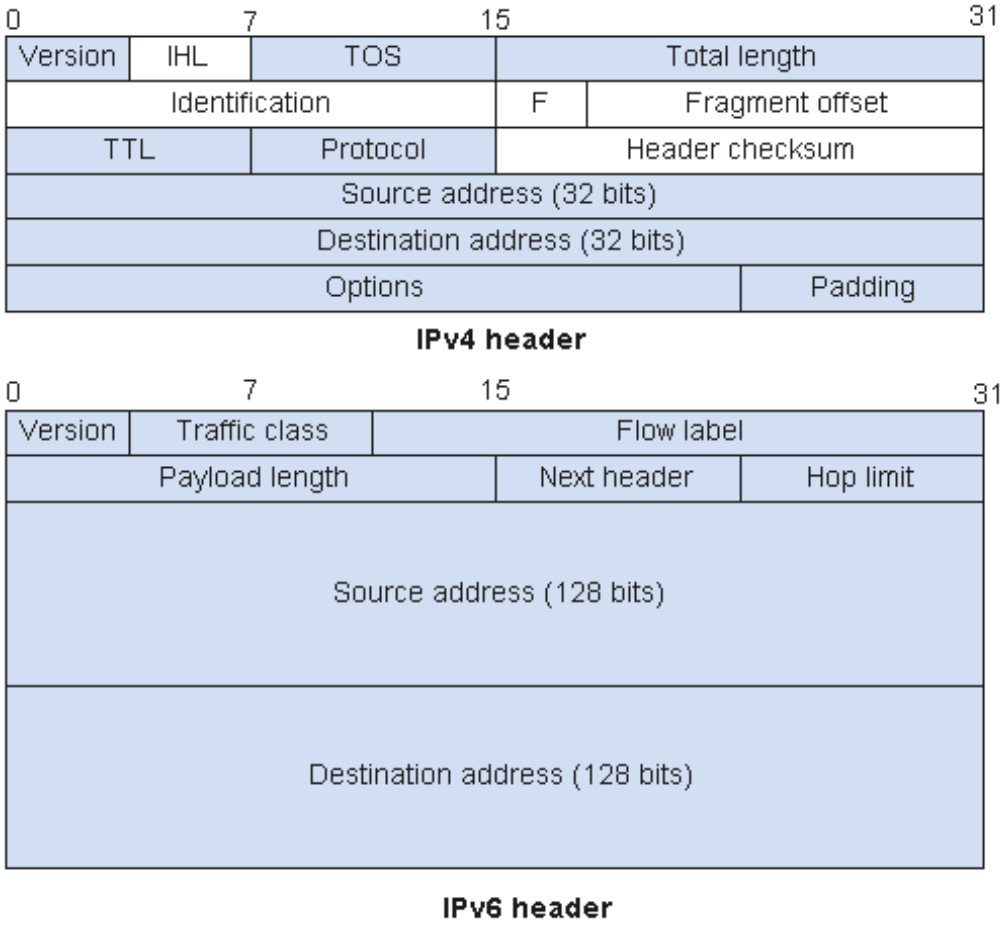
和 IPv4 报文头的比较

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。

图11-5 IPv4 和 IPv6 报文头格式比较



11.7 IPv6 的特点

简化的报文头格式

通过将 IPv4 报文头中的某些字段裁减或移入到扩展报文头，减小了 IPv6 基本报文头的长度。IPv6 使用固定长度的基本报文头，从而简化了转发设备对 IPv6 报文的处理，提高了转发效率。尽管 IPv6 地址长度是 IPv4 地址长度的四倍，但 IPv6 基本报文头的长度只有 40 字节，为 IPv4 报文头长度（不包括选项字段）的两倍。

充足的地址空间

IPv6 的源地址与目的地址长度都是 128 比特（16 字节）。它可以提供超过 3.4×10^{38} 种可能的地址空间，完全可以满足多层次的地址划分需要，以及公有网络和机构内部私有网络的地址分配。

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

内置安全性

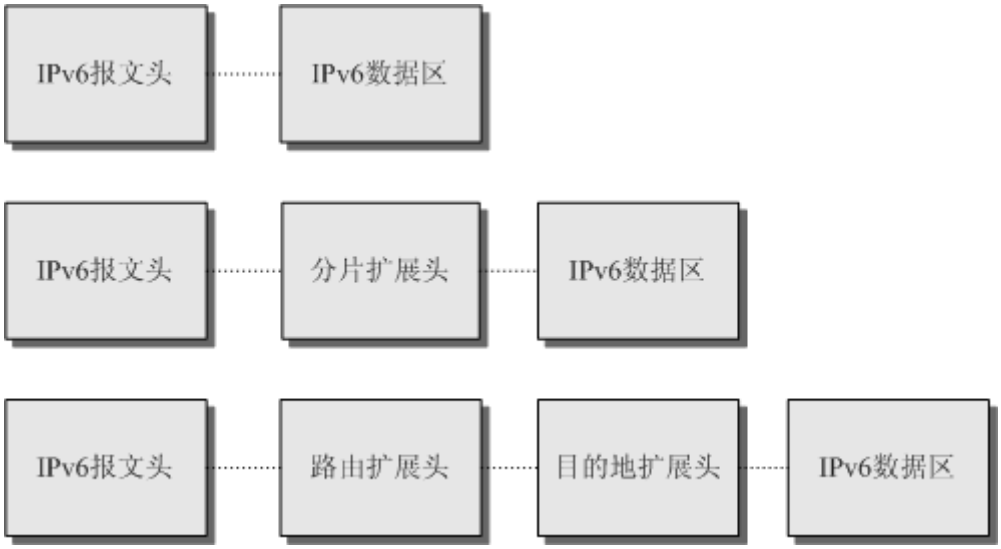
IPv6 将 IPSec 作为它的扩展报头实现，提供端到端的安全特性。这一特性为解决网络安全问题提供了标准，并提高了不同 IPv6 实现的互操作性。

灵活的扩展报文头

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 11-6 所示。

图11-6 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。

增强的邻居发现机制

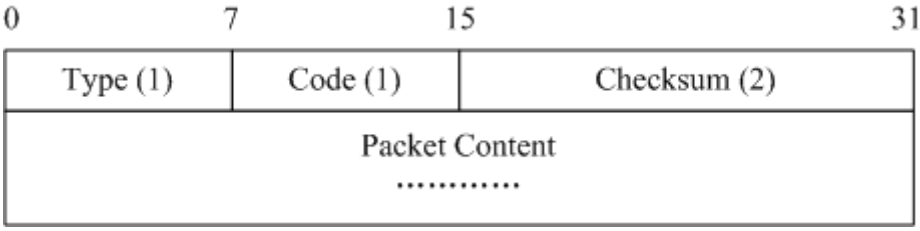
IPv6 的邻居发现协议是通过一组 ICMPv6（Internet Control Message Protocol for IPv6，IPv6 的因特网控制报文协议）消息实现的，管理着邻居节点间（即同一链路上的节点）信息的交互。它代替了 ARP（Address Resolution Protocol，地址解析协议）、ICMPv4 路由器发现和 ICMPv4 重定向消息，并提供了一系列其他功能。

11.8 ICMPv6

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 11-7 所示

图11-7 ICMPv6 报文格式



报文中各个字段的解释如下：

- Type 字段表明消息的类型，0 至 127 位表示差错报文类型，128 至 255 位表示消息报文类型。
- 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

- 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 HopLimit 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 参数错误差错报文。

  - IPv6 基本头或扩展头的某个域有错误

- IPv6 基本头或扩展头的 NextHeader 值不可识别
- 扩展头中出现未知的选项

ICMPv6 信息报文的分类

请求信息（Echo Request）和应答信息（Echo Reply）。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。

11.9 ACL6

ACL6 的分类

根据应用目的，可将 ACL6 分为两种：

- 基本 ACL6，基本 ACL6 只能使用源地址信息做为定义 ACL6 规则的元素。
- 高级 ACL6，高级 ACL6 可以使用数据包的源地址信息、目的地址信息、IP 承载的协议类型、针对协议的特性定义规则，例如 TCP 的源端口、目的端口，ICMPv6 协议的类型、ICMPv6 Code 等。可以利用高级 ACL6 定义比基本 ACL6 更准确、更丰富、更灵活的规则。

ACL6 的匹配顺序

一个 ACL 中可以包含多个规则，而每个规则都指定不同的报文匹配选项，这些规则可能存在重复或矛盾的地方，在将一个报文和 ACL 的规则进行匹配的时候，到底采用哪些规则呢？就需要确定规则的匹配顺序。

ACL6 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规则匹配。
- 自动排序：按照“深度优先”的顺序进行规则匹配。
  - 基本 ACL6 的“深度优先”顺序判断原则如下

1. 先比较源 IPv6 地址范围，源 IPv6 地址范围小（前缀长）的规则优先。
2. 如果源 IPv6 地址范围相同，则先配置的规则优先。

  - 高级 ACL6 的“深度优先”顺序判断原则如下

1. 先比较协议范围，指定了 IPv6 协议承载的协议类型的规则优先。
2. 如果协议范围相同，则比较源 IPv6 地址范围，源 IPv6 地址范围小（前缀长）的规则优先。
3. 如果协议范围、源 IPv6 地址范围相同，则比较目的 IPv6 地址范围，目的 IPv6 地址范围小（前缀长）的规则优先。
4. 如果协议范围、源 IPv6 地址范围、目的 IPv6 地址范围相同，则比较四层端口号（TCP/UDP 端口号）范围，四层端口号范围小的规则优先。
5. 如果上述范围都相同，则先配置的规则优先。

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。

ACL6 步长

配置 USG9500 的 ACL6 时，可以为一个 ACL6 规则组指定一个“步长”。步长的含义是：自动为 ACL6 子规则分配编号的时候，每个 ACL6 规则组的子规则编号之间的差值。ACL6 规则组的步长固定为 1，且不能通过 **step** 命令改变步长。

11.10 邻居发现

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。

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。当一个节点是主机时，路由器需要通知主机向特定目的地址转发报文的理想下一跳地址；当一个节点是路由器时，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。

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

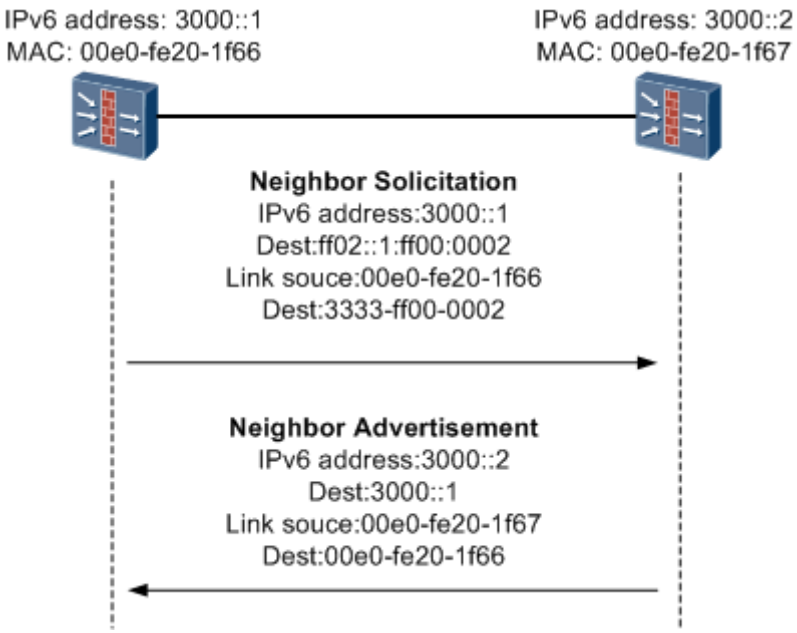
邻居发现功能和 IPv4 中的 ARP 功能类似，主要实现对邻居地址的解析和邻居可达性的探测，依赖于邻居请求和邻居通告报文完成。

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。

图 11-8 所示为 IPv6 邻居发现过程。

图11-8 IPv6 邻居发现过程



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）
- 和主机相关的其它信息，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、前缀列表以及其它配置。

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理想下一跳 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。

11.11 SEND

SEND（SEcure Neighbor Discovery）协议在 ND 的基础上进行了扩展，新增了如下信息：

- 选项字段

  CGA（Cryptographically Generated Addresses）、RSA（Rivest Shamir and Adleman）、Timestamp 和 Nonce。
- 消息类型

  CPS（Certification Path Solicitation）和 CPA（Certification Path Advertisement）。

通过这些新的消息类型和选项字段，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。

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。

关于证书的详细介绍请参见证书部分。

11.12 Path MTU

网络上的 MTU 问题

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 的工作原理

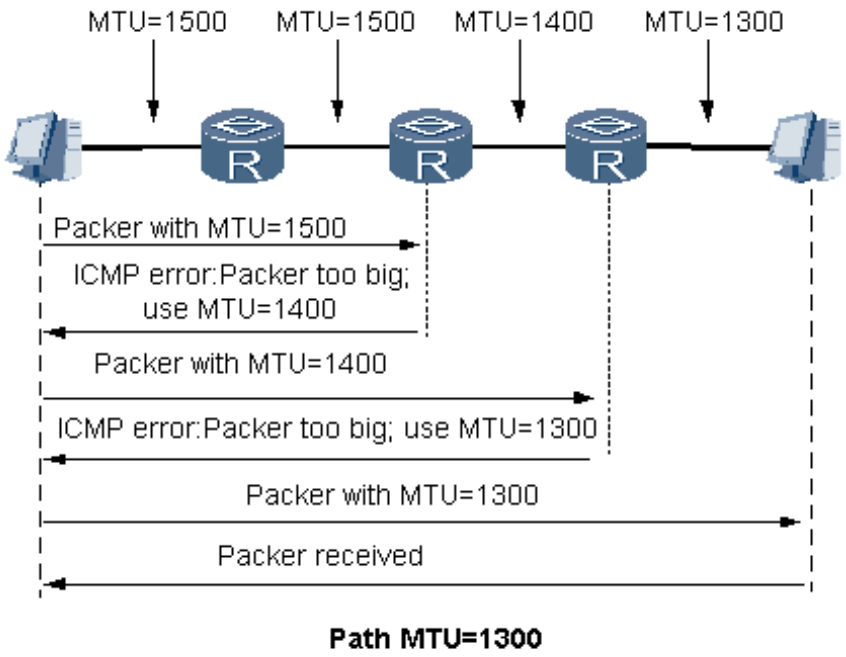
Path MTU（以下简称 PMTU），是确定从源端到目的端路径上合适的 IPv6 MTU 值的一种机制。PMTU 发现协议描述了一种动态发现任意路径的 PMTU 的方法。当一个 IPv6 节点发送大量数据到另一节点时，数据通过一系列 IPv6 分片传送。当这些分片具有从源节点到信宿节点能够成功传送所允许的最大长度时，我们认为它达到理想状态，这个分片长度被称为路径 MTU。

一个源节点开始会假设一个路径的 PMTU 是路径中第一跳的已知的 IPv6 MTU，如果从那个路径发出的报文太大以至于不能沿着路径转发，中间节点将丢弃此报文并返回一个 ICMPv6 数据过大差错报文给源节点，根据数据过大消息中的 IPv6 MTU 值来设置此路径的 PMTU 值。

当节点学习到的 PMTU 值小于或者等于实际的 PMTU 时，PMTU 的发现过程结束。注意在 PMTU 发现过程结束之前，可能会出现反复发送报文和收到报文太大消息，这是因为可能会不断发现更远的路径链路有更小的 IPv6 MTU。

如图 11-9 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。

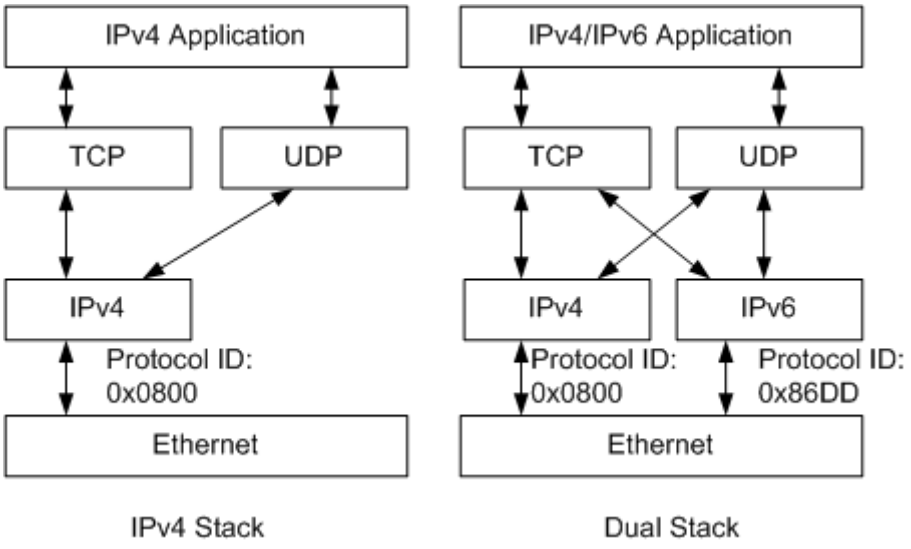
图11-9 PMTU 发现的工作过程



11.13 双协议栈

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 11-10 所示。

图11-10 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，如果协议类型字段的值为 0x0800，表示网络层是 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。

11.14 IPv6 over IPv4 隧道

在 IPv4 网络向 IPv6 网络过渡的初期，IPv4 网络已被大量部署，而 IPv6 网络只是散步在世界各地的一些孤岛。利用隧道技术可以在 IPv4 网络上创建隧道，从而实现 IPv6 孤岛之间的互连。在 IPv4 网络上用于连接 IPv6 孤岛的隧道称为 IPv6 over IPv4 隧道。为了实现 IPv6 over IPv4 隧道，需要在 IPv4 网络与 IPv6 网络交界的边界路由设备上启动 IPv4/IPv6 双协议栈。

IPv6 over IPv4 隧道技术的原理如图 11-11 所示。

1. 启动 IPv4/IPv6 双协议栈

   边界路由设备启动 IPv4/IPv6 双协议栈。

2. 封装 IPv6 报文

   边界路由设备在收到从 IPv6 网络侧来的报文后，根据路由表判定该报文要通过隧道进行转发，就把收到的 IPv6 报文作为负载，加上 IPv4 报文头，封装到 IPv4 报文里。

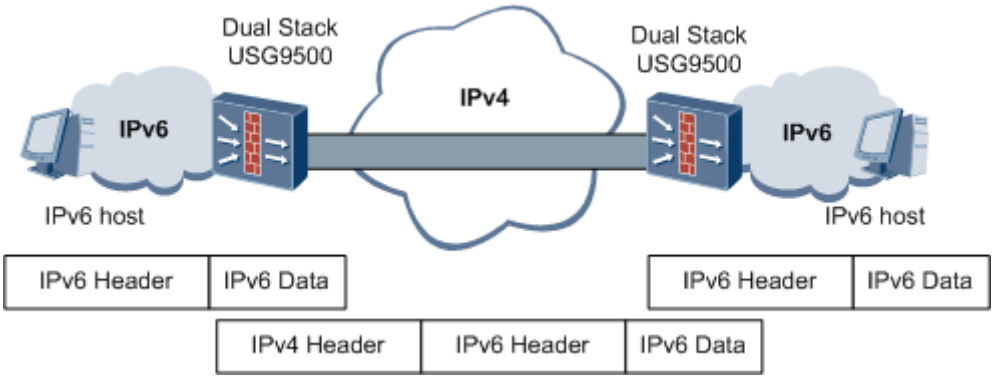
3. 传递封装后的报文

   在 IPv4 网络中，封装后的报文被传递到对端的边界路由设备上。

4. 对报文解封装

   对端边界路由设备对报文解封装，去掉 IPv4 报文头，然后将解封装后的 IPv6 报文转发到对端的 IPv6 网络中。

图11-11 IPv6 over IPv4 隧道原理图



在两个边界路由设备之间用来传递 IPv6 报文的虚拟的通道就是 IPv6 over IPv4 隧道。根据创建隧道的方式，可以对隧道进行分类。目前，常用的 IPv6 over IPv4 隧道模式有以下几种。

- IPv6 over IPv4 手动隧道
- IPv6 over IPv4 GRE 隧道（简称 GRE 隧道）
- IPv6 over IPv4 自动隧道（简称自动隧道）
- 6to4 隧道
- ISATAP 隧道

IPv6 over IPv4 手动隧道

IPv6 over IPv4 手动隧道是在隧道两端的边界路由设备上通过人工配置而创建的。它需要静态指定隧道的源 IPv4 地址和目的 IPv4 地址。

手动隧道相当于通过 IPv4 骨干网连接的两个 IPv6 域的永久链路，是边界路由设备之间进行定期安全通信的固定通道。

手动隧道可用于 IPv6 孤岛之间的通信，也可在边界路由设备与主机之间配置。隧道两端的主机和路由设备均需支持 IPv4 和 IPv6 协议栈。

IPv6 over IPv4 GRE 隧道

使用 IPv4 的 GRE（Generic Routing Encapsulation）隧道也可以承载 IPv6 报文，此时的 GRE 隧道称为 IPv6 over IPv4 GRE 隧道。与 IPv6 over IPv4 手动隧道相同，GRE 隧道也是两点之间的链路，每条链路都是一条单独的隧道。GRE 隧道不与特定的乘客或传输协议绑定，只把 IPv6 作为乘客协议，把 GRE 作为承载协议。

GRE 隧道也是在隧道两端的边界路由设备上通过人工配置而创建的，也需要静态指定隧道的源 IPv4 地址和目的 IPv4 地址。与手动隧道不同的是，GRE 隧道为了增强隧道的安全性，可以设置对 GRE 报文头进行校验以及对隧道的关键字进行验证。

GRE 隧道可用于边界路由设备之间，或者用于边界路由设备与主机系统之间。隧道两端的主机和路由设备均需支持 IPv4 和 IPv6 协议栈。

有关 GRE 配置的详细介绍请参见《配置指南 VPN 分册》。

IPv6 over IPv4 自动隧道

要创建 IPv6 over IPv4 自动隧道，需要使用一类特殊的 IPv6 地址，即兼容 IPv4 的 IPv6 地址。兼容 IPv4 的 IPv6 地址格式为：

0:0:0:0:0:0: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）。

在配置自动隧道时，只需要在边界路由设备或主机上指定隧道源地址，不需要指定隧道的目的地址。隧道目的地址是从原始的 IPv6 报文的目的地址中获取的。

IPv6 over IPv4 自动隧道通常用于孤立的 IPv4/IPv6 双协议栈主机需要穿过 IPv4 网络访问远端 IPv6 网络的情况。在孤立的 IPv4/IPv6 主机和 IPv4/IPv6 路由设备之间需要配置自动隧道。

在建立自动隧道时，需要隧道两端都要配置兼容 IPv4 的 IPv6 地址，兼容 IPv4 的 IPv6 地址又依赖于隧道的物理接口的 IPv4 地址，受到 IPv4 地址短缺的限制，因而有一定的局限性。

6to4 隧道

6to4 隧道也是一种将多个 IPv6 孤岛通过 IPv4 网络互连的机制。6to4 隧道可在孤立的 IPv6 网络和 IPv4 网络之间的边界路由设备上配置。6to4 隧道两端的边界路由设备必须同时支持 IPv4 和 IPv6 双协议栈。

6to4 隧道与手动配置隧道的主要区别在于：6to4 隧道可以是点到多点的连接，而手动隧道仅是点到点的连接。所以 6to4 隧道的路由设备并不是成对配置的。

6to4 隧道与自动隧道类似，它可自动查找隧道的另一端点，但它不需要指定兼容 IPv4 的 IPv6 地址。

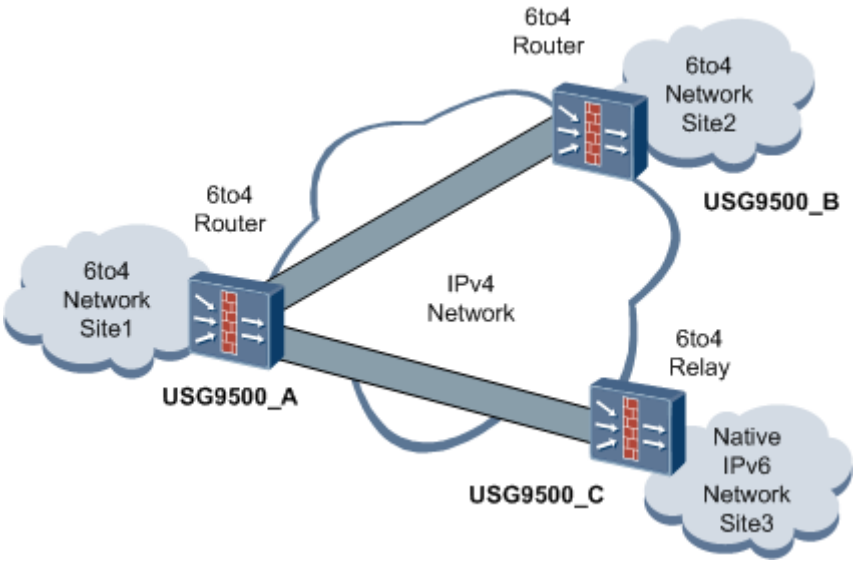
6to4 隧道使用了一种特殊的 IPv6 地址，即 6to4 地址，其格式为：

2002:IPv4 地址:子网 ID:接口 ID

6to4 地址的前缀是 **2002:IPv4 地址**，前缀长度为 48bits。其中 IPv4 地址是为 IPv6 孤岛申请的一个全球唯一的 IPv4 地址。在 IPv6/IPv4 边界路由设备与 IPv4 网络链接的物理接口上必须配置该 IPv4 地址。子网 ID 的长度为 16bits，接口 ID 的长度为 64bits，均由用户在 IPv6 孤岛内分配。

如图 11-12 所示，Site1 和 Site2 均为 6to4 网络，6to4 网络内的主机和路由设备被分配了 6to4 地址。Site1 内的主机和路由设备的 6to4 地址内嵌入的 IPv4 地址就是 USG9500_A 到 IPv4 网络接口的 IPv4 地址。Site2 内的主机和路由设备的 6to4 地址内嵌入的 IPv4 地址就是 USG9500_B 到 IPv4 网络接口的 IPv4 地址。USG9500_A 和 USG9500_B 均为 6to4 路由设备。

图11-12 6to4 隧道和 6to4 中继



Site1 内的主机要访问 Site2 内的主机时，其工作原理如下：

1. IPv6 报文被传送到 USG9500_A；
2. USG9500_A 检查 IPv6 报文的目的地址，发现是 6to4 地址，从该 6to4 地址中获得 6to4 隧道对端的 IPv4 地址；
3. USG9500_A 将该 IPv6 报文封装到 IPv4 报文中，IPv4 报文头的目的地址就是隧道对端的 IPv4 地址，源地址就是隧道本端的 IPv4 地址；
4. USG9500_A 将该 IPv4 报文通过 IPv4 网络转发到 USG9500_B；
5. USG9500_B 进行解封装操作，获得原来的 IPv6 报文，然后将该 IPv6 报文在 Site2 内被送到目的主机。

上面介绍的过程可以实现 6to4 网络之间的通信。为了实现 6to4 网络与本真（Native）IPv6 网络之间的通信，就需要 6to4 中继路由设备了。所谓本真 IPv6 网络，就是它内部的主机或路由设备均不配置 6to4 地址。

6to4 中继路由设备是 6to4 网络与本真 IPv6 网络之间的网关。6to4 中继路由设备的一侧连接本真 IPv6 网络，其另一侧连接 IPv4 网络，并与 6to4 路由设备建立 6to4 隧道。如图 11-12 所示，6to4 网络内的主机要访问 IPv6 Internet 时，其工作过程如下。

1. IPv6 报文被传送到 USG9500_A；
2. USG9500_A 检查 IPv6 报文的目的地址，发现是本真（Native）IPv6 网络地址。通过查找路由表，下一跳是指向 6to4 隧道，从 6to4 隧道地址中获取要封装的 IPv4 目的地址；
3. USG9500_A 将该 IPv6 报文封装到 IPv4 报文中，IPv4 报文头的目的地址就是隧道对端的 IPv4 地址，源地址就是隧道本端的 IPv4 地址；
4. USG9500_A 将该 IPv4 报文通过 IPv4 网络转发到 USG9500_C；
5. USG9500_C 进行解封装操作，获得原来的 IPv6 报文，然后将该 IPv6 报文送到 Site3 内的目的主机。

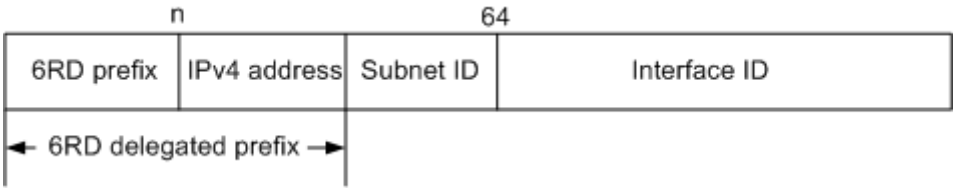
6RD 隧道

6RD（IPv6 Rapid Deployment）隧道技术是一种在已有的 IPv4 网络基础上，为用户提供 IPv6 接入服务的快速部署的技术。6RD 隧道技术是对原有 6to4 方案的改进，主要区别在于 6to4 定义的地址格式中使用知名的 2002::/16 前缀，而 6RD 的地址前缀则是可以由企业从自己的 IPv6 地址空间中划分得到的。

6RD 隧道技术主要是满足 IPv6 用户发送报文穿越 IPv4 网络，访问 IPv6 服务和资源的问题，其核心思想是通过在 CE（Customer Edge）与 CE 或者 CE 与网关之间自动建立、拆解隧道，实现 IPv6 报文跨越 IPv4 的网络。而自动隧道的建立是依靠预先定义的 6RD 前缀完成的。

6RD 地址格式包括：6RD 前缀、IPv4 地址、子网 ID 和接口标识符。如图 11-13 所示

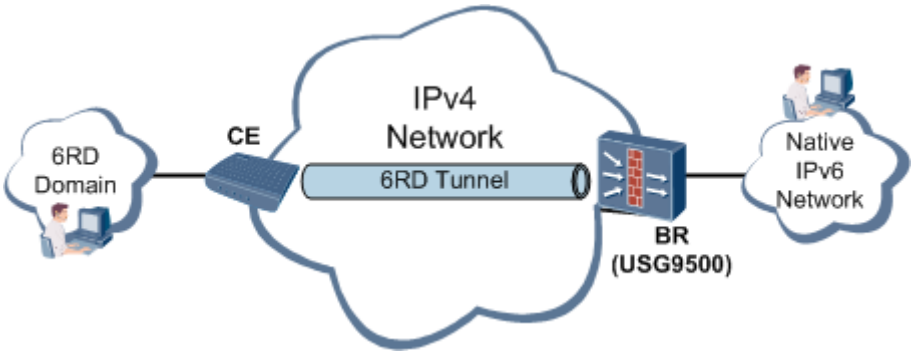
图11-13 6RD 地址格式



6RD 委托前缀（6RD delegated prefix）包含 6RD 前缀和部分或全部 IPv4 地址，由两部分内容计算得出 6RD 委托前缀值。其中 6RD 委托前缀中 IPv4 地址的长度取决于 6RD 隧道中配置的 IPv4 前缀长度。

如图 11-14 所示，6RD 隧道具体业务处理流程如下：

图11-14 6RD 隧道



1. 当 CE 接收到来自 IPv6 终端发送的报文时，将 IPv6 报文封装到 IPv4 隧道中，发送至 6RD Border Relay，隧道外层源地址为 CE 的 IPv4 地址，目的地址为 6RD Border Relay 的 IPv4 地址；
2. 6RD Borer Relay 将收到的 IPv4 隧道报文解封装，然后将 IPv6 报文进行转发。

ISATAP 隧道

ISATAP（Intra-site Automatic Tunnel Addressing Protocol）隧道用于 IPv4 网络中的 IPv4/IPv6 主机访问 IPv6 网络的情况，可以在 ISATAP 主机与 ISATAP 路由设备之间建立 ISATAP 隧道。

建立 ISATAP 隧道时，需要使用 ISATAP 格式地址，其结构如下：

Prefix(64bit)::5EFE:IPv4-Address

在创建 ISATAP 隧道时，由于 IPv4/IPv6 主机和 ISATAP 路由设备在同一个 IPv4 网络里，ISATAP 地址中嵌入的 IPv4 地址可以是公网地址，也可以是私网地址。

如图 11-15 所示，IPv4/IPv6 主机获得 IPv6 地址的过程如下：

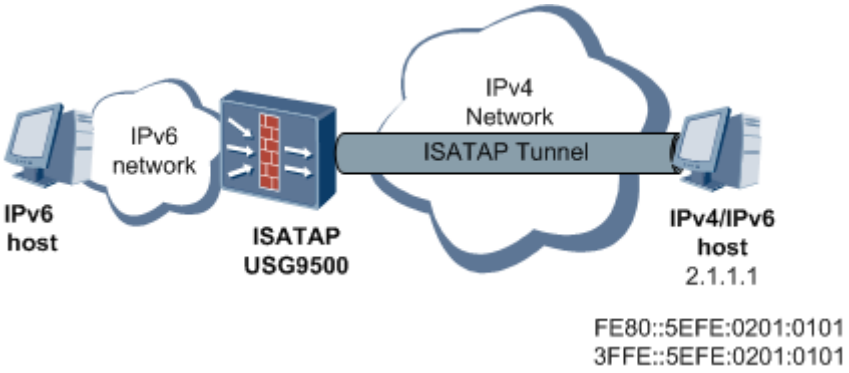
1. IPv4/IPv6 主机发送路由设备请求消息

   IPv4/IPv6 主机使用 ISATAP 格式的链路本地地址向 ISATAP 路由设备发送路由设备请求消息，该路由设备请求消息被封装在 IPv4 报文中。
2. ISATAP 路由设备响应请求

   ISATAP 路由设备使用路由设备通告消息响应主机的路由设备请求。路由设备通告消息中包含 ISATAP 前缀（ISATAP 前缀在路由设备上通过人工配置）。
3. IPv4/IPv6 主机得到自己的 IPv6 地址

   IPv4/IPv6 主机将 ISATAP 前缀与 **5EFE:IPv4-Address** 组合得到自己的 IPv6 地址，并用此地址访问 IPv6 主机。

图11-15 ISATAP 隧道



IPv4/IPv6 主机要访问 IPv6 Internet 时，其工作原理如下。

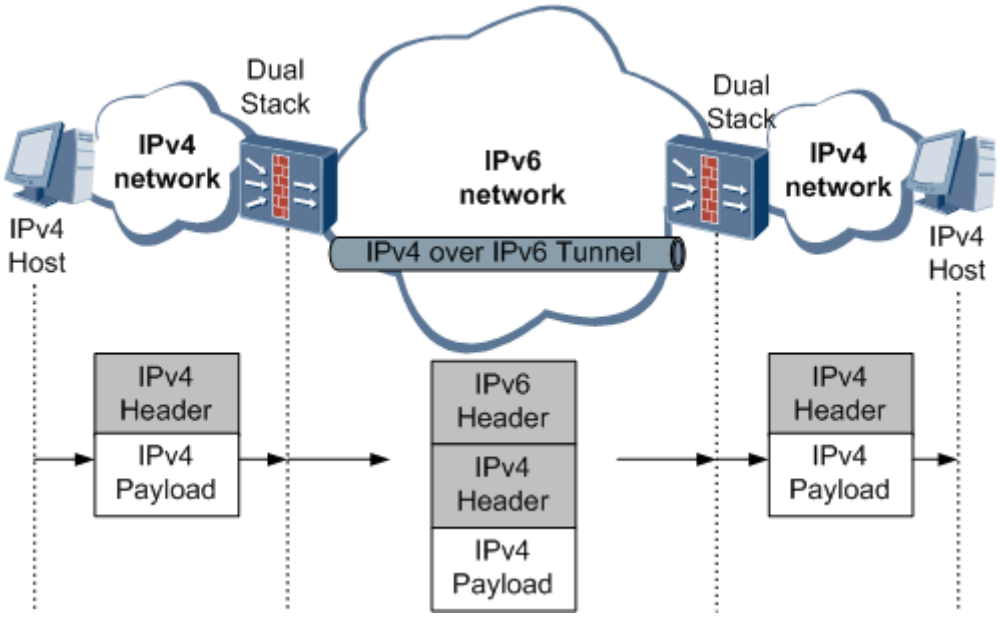
1. IPv4 网络中的 IPv4/IPv6 主机按照上面的过程获得自己的 IPv6 地址。
2. IPv4/IPv6 主机发送访问 IPv6 网络的 IPv6 主机的报文，该报文封装在 IPv4 中。
3. ISATAP 路由设备接收该 IPv4 报文后执行解封装操作，将其中的 IPv6 报文发送到 IPv6 网络中的 IPv6 主机。

11.15 IPv4 over IPv6 隧道

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。这类似于在 IP 网络上利用隧道技术部署 VPN。在 IPv6 网络上用于连接 IPv4 孤岛的隧道，称为 IPv4 over IPv6 隧道。

IPv4 over IPv6 技术原理

图11-16 IPv4 over IPv6 隧道组网图



IPv4 over IPv6 隧道技术的原理如图 11-16 所示。

1. 启动 IPv4/IPv6 双协议栈

   边界路由设备启动 IPv4/IPv6 双协议栈。

2. 封装 IPv6 报文

   边界路由设备在收到从 IPv4 网络侧来的报文后，如果报文的目的地不是自身，就把收到的 IPv4 报文作为净荷，加上 IPv6 报文首部，封装到 IPv6 报文里。

3. 传递封装后的报文

   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. 对报文解封装

   对端边界路由设备对报文解封装，去掉 IPv6 报文首部，然后将解封装后的 IPv4 报文转发到 IPv4 网络中。

11.16 NAT64

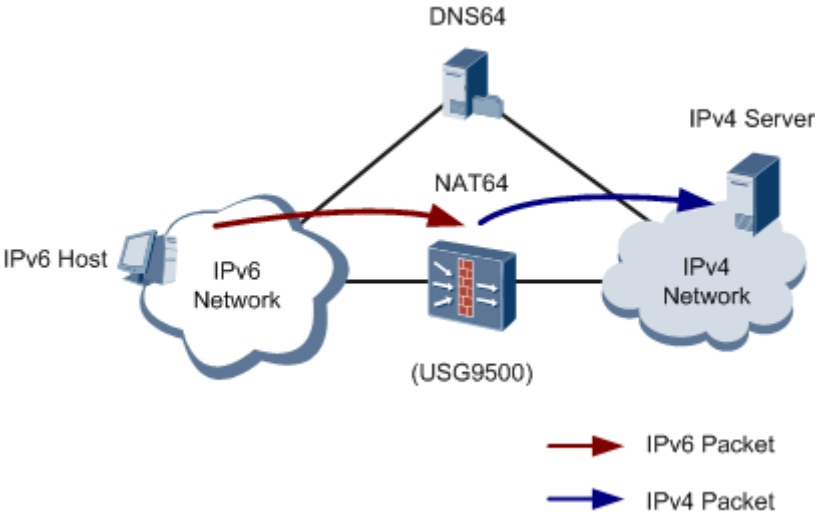
说明
NAT64 中描述的 IPv4 网络为单协议栈 IPv4 网络，IPv6 网络为单协议栈 IPv6 网络。

概述

NAT64 主要应用于 IPv6 网络中单协议栈主机访问 IPv4 网络资源的场景。如图 11-17 所示，USG9500 部署在 IPv6 网络和 IPv4 网络之间，网络中必须存在支持 IPv4 和 IPv6 域名解析功能的 DNS64 设备。

说明
DNS64 设备提供域名与 IPv6 地址的对应关系，它使用 USG9500 上配置的 NAT64 前缀与 IPv4 网络中 Server 的 IPv4 地址组合生成 IPv6 地址，并生成相应的 AAAA 记录。

图11-17 NAT64 示意图



以 NAT64 动态映射为例，当 IPv6 网络中的 Host 使用域名方式访问 IPv4 网络中的 Server 时，具体流程如下：

1. 借助 DNS64 的域名解析功能，Host 获得了 Server 的域名对应的 IPv6 地址（该地址由特定的前缀和 Server 的 IPv4 地址组成），Host 使用此 IPv6 地址为目的地址发起请求。
2. USG9500 收到 Host 发出的 IPv6 报文后，发现 IPv6 报文的目的地址中包含特定的 NAT64 前缀（该前缀在 USG9500 上事先设置），则进行 NAT64 处理。
3. USG9500 使用地址转换算法提取出 IPv6 报文中的 IPv4 地址，以此 IPv4 地址作为 IPv4 报文的目的地址。然后根据安全域间配置的 NAT64 动态映射，以 NAT 地址池/地址池组中的地址为 IPv4 报文的源地址，将 IPv6 报文转换为 IPv4 报文，同时生成会话表。
4. USG9500 将转换后的 IPv4 报文发送至 Server。
5. USG9500 收到 IPv4 网络中 Server 的响应报文后，根据会话表将 IPv4 报文转换为 IPv6 报文，然后发送至 Host。

NAT64 前缀

USG9500 通过判断 IPv6 报文的目的地址中是否包含 NAT64 前缀来决定是否对该 IPv6 报文进行 NA64 处理，NAT64 前缀分为以下两种形式：

- 知名前缀

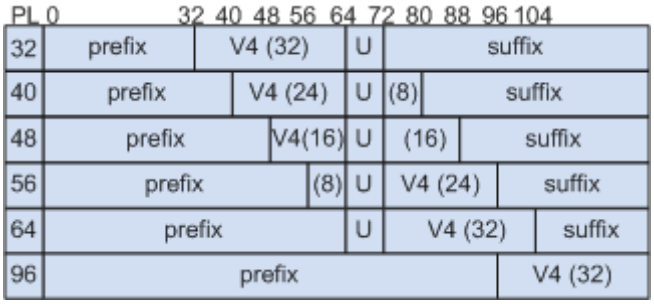
  即 64:FF9B::/96，缺省情况下已存在，无需配置。

- 自定义前缀

  前缀长度为 32、40、48、56、64 或 96。

  根据前缀长度不同，IPv4 地址嵌入 IPv6 地址时，嵌入的位置存在差异，具体差异如图 11-18 所示，其中 PL（Prefix Length）表示前缀长度；suffix 表示后缀，可以任意取值，USG9500 不处理该字段；U 为保留位，取值必须为 0。

图11-18 IPv4 地址嵌入 IPv6 地址



  - 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。

配置 DNS64 设备时，需要确保 NAT64 前缀和前缀长度与 USG9500 上的配置相同。

IPv4 NAT 地址池/地址池组

USG9500 将 IPv6 报文转换为 IPv4 报文时，使用 NAT 地址池/地址池组中的地址作为 IPv4 报文的源地址，与 IPv4 网络中的设备通信。需要确保 NAT 地址池/地址池组中的地址在 IPv4 网络中路由可达。

关于 NAT 地址池/地址池组的详细介绍请参见配置 NAT 地址池/NAT 地址池组。

NAT64 静态映射

除了在安全域间配置 NAT64 动态映射外，USG9500 还支持配置 NAT64 静态映射，将 IPv6 地址转换为特定的 IPv4 地址，不需要使用 NAT 地址池/地址池组。NAT64 静态映射方式属于“一对一的转换方式”。

配置 NAT64 静态映射后，USG9500 会生成 IPv6 和 IPv4 报文的 Server-Map 表，该 Server-Map 表既可以用于 IPv6 网络中的主机主动访问 IPv4 网络的场景，也可以用于 IPv4 网络中的主机主动访问 IPv6 网络的场景：

- 当 IPv6 网络中的主机主动访问 IPv4 网络时，USG9500 检查收到的 IPv6 报文，匹配 Server-Map 表的报文会进行源地址转换，由 IPv6 地址转换为 IPv4 地址。同时 USG9500 提取出目的 IPv6 地址中内嵌的 IPv4 地址，以此 IPv4 地址作为 IPv4 报文的目的地址。
- 当 IPv4 网络中的主机主动访问 IPv6 网络时，USG9500 检查收到的 IPv4 报文，匹配 Server-Map 表的报文会进行目的地址转换，由 IPv4 地址转换为 IPv6 地址。同时 USG9500 会在前缀列表中随机选择一个前缀，与报文原有的 IPv4 地址组合生成新的 IPv6 地址，以此 IPv6 地址作为 IPv6 报文的源地址。如果没有配置前缀，则会使用知名前缀 64:FF9B::/96。

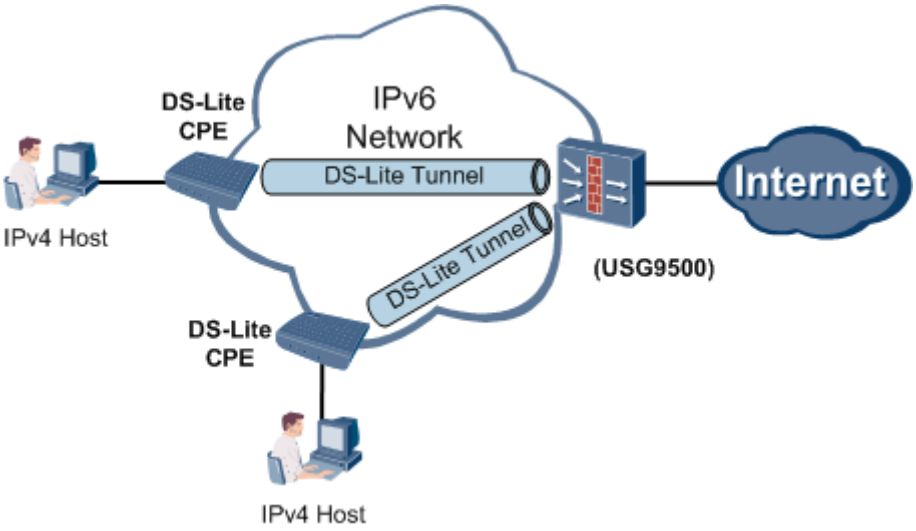
11.17 DS-Lite

在 IPv4 Internet 向 IPv6 Internet 过渡的中后期，IPv6 网络已被大量部署，甚至已经建成了 IPv6 城域网，但是仍有 IPv4 用户需要发展，典型组网如图 11-19 所示。为了实现 CPE（Customer Premise Equipment）下挂私网 IPv4 用户与 IPv4 Internet 互访，需要解决如下问题:

1. IPv4 报文在 IPv6 城域网传送的问题。
2. 因为 CPE 设备下用户是私网 IPv4 地址，所以在 USG9500 上需要对报文进行地址转换，同时要考虑到多个 CPE 设备下私网地址重叠的问题。

DS-Lite（Dual-stack Lite）技术可以解决这些问题。在 USG9500 上 DS-Lite 技术实现方式概括起来就是 IPv4 over IPv6 隧道＋NAT。IPv4 over IPv6 隧道解决 IPv4 报文在 IPv6 网络传送的问题；NAT 功能对报文进行地址转换，同时在 Server-map 表和会话表中增加 CPE IP 和 Tunnel ID 等字段，其中 CPE IP 字段用于解决私网地址重叠问题；Tunnel ID 字段用于反向报文找到对应的隧道接口。

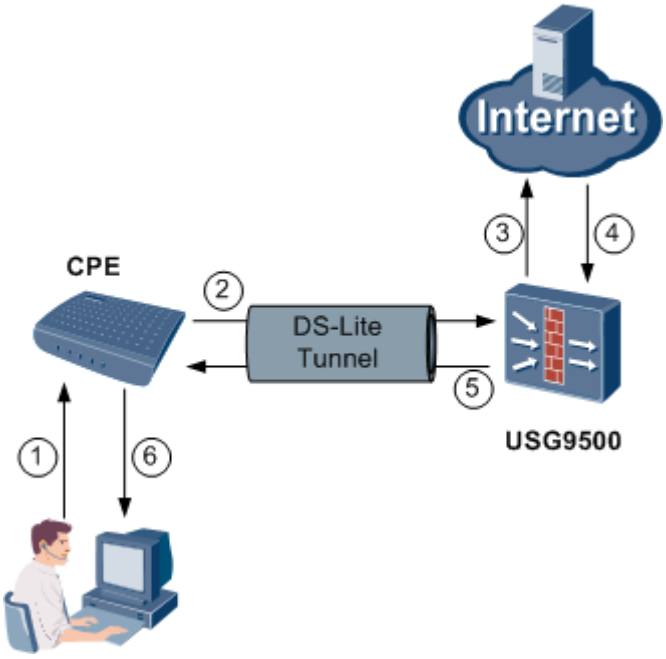
图11-19 DS-lite 组网图



DS-Lite 技术详细的处理流程与场景有关，不同的场景处理流程略有不同。

- CPE 接入的私网 IPv4 用户访问 IPv4 Internet，处理流程如图 11-20 所示。

图11-20 CPE 接入私网 IPv4 用户访问 IPv4 Internet 处理流程



1. 私网 IPv4 用户要访问 IPv4 Internet，首先发送 IPv4 报文到网关设备 CPE。
2. CPE 设备接收到私网 IPv4 报文后，封装成 4over6 报文，通过 IPv6 网络传送到 USG9500。
3. USG9500 对报文进行解封装，查询域间或域内 NAT 策略访问 IPv4 Internet，如果是三元组 NAT，要创建源和目的 Server-map，并把报文中携带的 CPE IP 和 Tunnel

ID 记录到 Server-map 里。同时创建会话表，记录出接口、CPE IP 和 Tunnel ID 等信息，用于给 IPv4 Internet 回应报文加封装。

说明

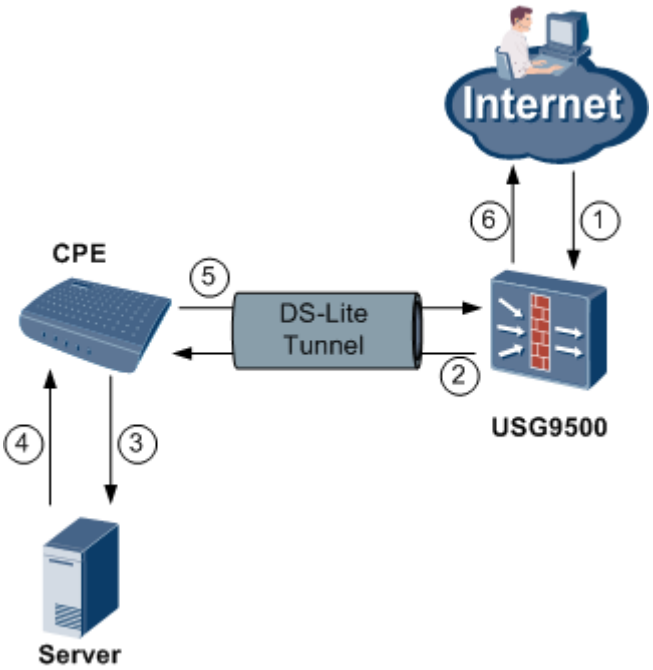
CPE IP 可以解决私网地址重叠问题。当不同 CPE 下私网地址相同时，USG9500 根据不同 CPE IP（此 IP 为 CPE 连接 IPv6 网络的 IPv6 地址，即 CPE 建立隧道的源地址）区分重叠的私网地址。

4. 外网来的如果是后续报文则根据 USG9500 上会话表记录的 CPE IP 查找路由找到去往 CPE 的出接口进行 4over6 加封装；如果是首报文且命中目的三元组 Server-map 表，则根据 Server-map 表记录的 CPE IP 查找路由找到去往 CPE 的出接口进行 4over6 加封装。
5. 进行 4over6 加封装时，用会话中记录的 CPE IP 做为隧道目的 IP 和 Tunnel 接口的源 IP 进行封装，并用 CPE IP 查找 IPv6 路由，找到实际的出接口发送封装的报文给 CPE 设备。
6. CPE 设备对接收到 4over6 报文解封装，发送给私网 IPv4 用户。

- 外网用户访问 CPE 下 IPv4 私网服务器，处理流程如图 11-21 所示。

  CPE 内部的私网需要对外部提供访问服务时，需要在 USG9500 上配置 NAT 的内部服务器功能，即 NAT Server。DS-Lite 场景下私网是可以重叠的，外网访问内部服务器的报文在封装隧道时要知道隧道源和目的 IP，因此在配置 NAT Server 时需要增加 CPE IP，并绑定 Tunnel 接口。CPE IP 用于私网隔离，也是外网访问内部服务器报文封装隧道的目的 IP；Tunnel ID 用于外网访问内部服务器报文加封装的隧道源 IP。

图11-21 外网用户访问 CPE 下 IPv4 私网服务器处理流程



1. 当外网用户访问 CPE 下 IPv4 私网服务器的报文进入 USG9500 后根据 Server-map 表记录的公网地址和私网地址对应关系、CPE IP 和 Tunnel ID 找到目的出接口，

并对报文进行 4over6 封装，同时创建会话表记录出接口、CPE IP 和 Tunnel ID 等信息。

2. USG9500 把封装的 4over6 报文，通过 IPv6 网络传送到 CPE 设备端。
3. CPE 设备对报文进行解封装后传给 IPv4 私网服务器。
4. IPv4 私网服务器回应报文给 CPE 设备。
5. CPE 设备接收到私网 IPv4 报文后，封装成 4over6 报文，通过 IPv6 网络传送到 USG9500。
6. USG9500 根据会话表记录信息找到外网出口。

12 可靠性

关于本章

12.1 双机热备份

12.1.1 介绍

定义

双机热备份是指双机状态信息和配置命令备份。当两台设备，在确定主用（Master）设备和备用（Backup）设备后，由主用设备进行业务的转发，而备用设备处于监控状态，同时主用设备定时向备用设备发送状态信息和需要备份的信息，当主用设备出现故障后，备用设备及时接替主用设备的业务运行。

双机热备份包含以下三种协议：

- VRRP（Virtual Router Redundancy Protocol）

  由 RFC2338 定义的一种容错协议，通过对物理设备和逻辑设备的分离，实现在多个出口网关之间进行选路。

- VGMP（VRRP Group Management Protocol）

  为防止 VRRP 状态不一致现象的发生，USG9500 在 VRRP 的基础上增加了 VGMP 扩展协议的功能。该协议负责统一管理加入其中的各 VRRP 备份组的状态。

- HRP（Huawei Redundancy Protocol）

  实现对 USG9500 的动态状态数据和关键配置命令进行实时备份的协议。

目的

USG9500 作为内外网的一个接入点，高可靠性至关重要。为了防止因为一台设备出现故障而导致网络业务中断的现象，USG9500 采用双机热备份技术，从而大大提高整个系统的稳定性和可靠性。

12.1.2 规格

双机热备份特性的相关规格如下：

- 支持来回路径不一致组网
- 支持实时备份
- 支持自动备份
- 支持手工备份连接状态
- 支持会话快速备份

12.1.3 可获得性

License 支持

本特性无须 License 支持。

版本支持

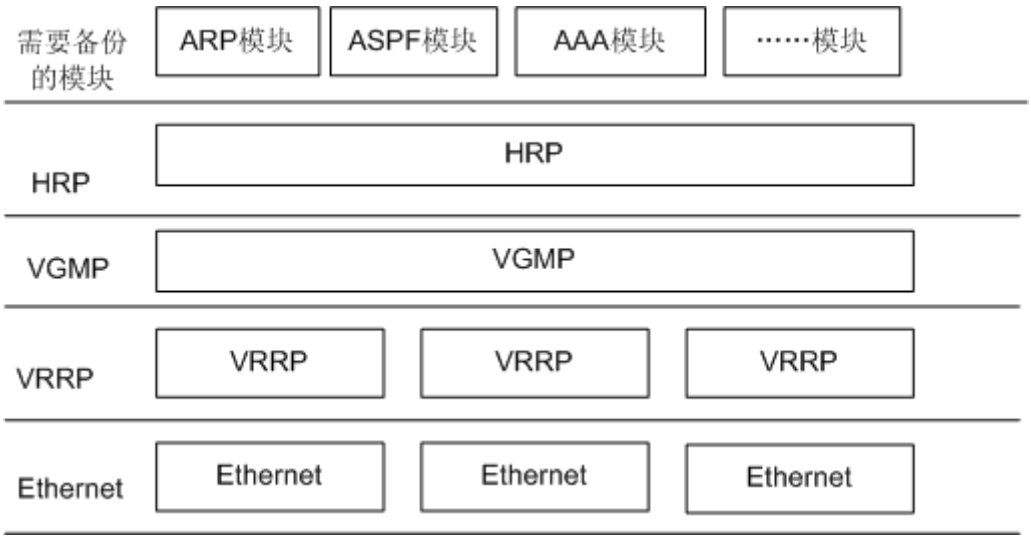
产品	最低支持版本
HUAWEI Secoway USG9500	V200R001

12.1.4 原理描述

12.1.4.1 双机热备份的协议体系结构

双机热备份包含的三个协议 VRRP、VGMP、HRP 协议的体系结构如图 12-1 所示。

图12-1 双机热备协议体系结构



VRRP 是标准的协议，负责监控单个链路的状态。

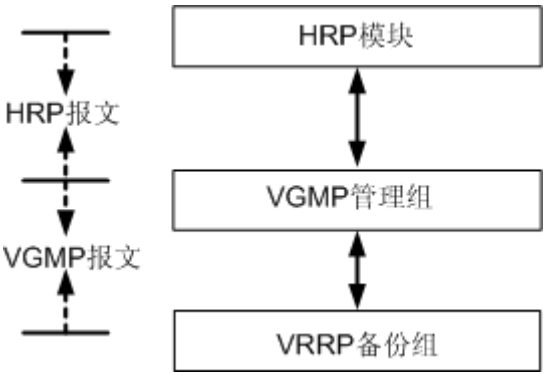
VGMP 是对设备上的所有 VRRP 备份组进行管理，负责监控整个设备的状态，并统一控制设备上所有 VRRP 备份组的状态。

HRP 负责在主备设备之间进行关键配置命令和连接状态信息的备份，保证主用设备出现故障后备用设备能够平滑接替工作。

12.1.4.2 双机热备份的协议层次关系

VRRP 备份组、VGMP 管理组和 HRP 之间的协议层次关系如图 12-2 所示。

图12-2 VRRP 备份组、VGMP 管理组和 HRP 之间的协议层次关系



当 VRRP 备份组状态变化时，由 VGMP 管理组来决定其本身的状态是否需要变化。

当 VGMP 管理组状态变化时，系统将通知 HRP 状态和配置主/从设备的状态发生相应的变化，从而确保两台设备之间配置命令和会话状态信息得到及时备份。

VGMP 管理组状态也要受 HRP 状态影响，即 VGMP 会根据 HRP 状态切换的结果来调整自身优先级，并进行 VRRP 状态切换。

12.2 VRRP

12.2.1 介绍

定义

VRRP（Virtual Router Redundancy Protocol）虚拟路由冗余协议，是一种容错协议。该协议通过把几台路由设备联合组成一台虚拟的路由设备，使用一定的机制保证当主机的下一跳 USG9500 出现故障时，及时将业务切换到其他 USG9500，从而保持通讯的连续性和可靠性。

以下是与 VRRP 协议相关的基本概念：

- 虚拟路由器（Virtual Router）：由 VRRP 管理的抽象设备，又称为 VRRP 备份组，被当作一个共享局域网内主机的缺省网关。它包括了一个虚拟路由器标识符和一组虚拟 IP 地址。
- 虚拟 IP 地址(Virtual IP Address)：虚拟路由器的 IP 地址，一个虚拟路由器可以有一个或多个 IP 地址，由用户配置。
- 虚拟 MAC 地址：是虚拟路由器根据虚拟路由器 ID 生成的 MAC 地址。一个虚拟路由器拥有一个虚拟 MAC 地址，格式为：00-00-5E-00-01-{VRID}。当虚拟路由器回应 ARP 请求时，使用虚拟 MAC 地址，而不是接口的真实 MAC 地址。
- 主 IP 地址（Primary IP Address）：从接口的真实 IP 地址中选出来的一个主用 IP 地址，通常选择配置的第一个 IP 地址。VRRP 广播报文使用主 IP 地址作为 IP 报文的源地址。
- Master USG9500（Virtual Router Master）：是承担转发报文或者应答 ARP 请求的 VRRP USG9500，转发报文都是发送到虚拟 IP 地址的。
- Backup USG9500（Virtual Router Backup）：一组没有承担转发任务的 VRRP USG9500，当 Master 设备出现故障时，它们将通过竞选成为新的 Master。

目的

随着 Internet 的发展，人们对网络的可靠性的要求越来越高。对于局域网用户来说，能够时刻与外部网络保持联系非常重要。

通常情况下，内部网络中的所有主机都设置一条相同的缺省路由，指向出口网关，实现主机与外部网络的通信。当出口网关发生故障时，主机与外部网络的通信就会中断。

配置多个出口网关是提高系统可靠性的常见方法，但局域网内的主机设备通常不支持动态路由协议，如何在多个出口网关之间进行选路是一个需要解决的问题。

VRRP 协议由 IETF（Internet Engineering Task Force，因特网工程任务组）推出，旨在解决局域网主机访问外部网络的可靠性问题，包括如下应用特性：

- 主备备份：这是 VRRP 提供 IP 地址备份功能的基本方式。主备备份方式需要建立一个虚拟路由器，该虚拟路由器包括一个 Master 设备和若干 Backup 设备，这些设备构成一个备份组。正常情况下，业务全部由 Master 承担。Master 出现故障时，Backup 接替工作。

- VRRP 负载分担：负载分担方式是指多台 USG9500 同时承担业务，单个 VRRP 备份组是不具备负载分担功能的，只有在多台设备上建立两个或更多的备份组，所有备份组均匀分担 Master 状态，此时就每台设备只承担了部分的业务，从而达到负载分担的作用。
- 虚拟 IP 地址 Ping 开关：提供了控制 Ping 通虚拟 IP 地址的开关命令。
- VRRP 的安全功能：对于安全程度不同的网络环境，可以在报头上设定不同的认证方式和认证字。
- VRRP 平滑倒换：CE 设备作为业务系统的网关，需要启用 VRRP 冗余备份功能，此时当设备处于进行主备倒换的过程中，本端和对端设备都不会出现 VRRP 状态切换，从而防止业务丢包。

12.2.2 规格

VRRP 特性的相关规格如下：

- 支持地址备份：当主用 USG9500 发生故障时，备用 USG9500 自动成为网关路由。
- 支持接口使用虚拟 MAC 地址发送报文。
- 支持主机 Ping 虚拟 IP 地址。

12.2.3 参考标准和协议

与 VRRP 特性相关的参考标准与协议如下：

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

12.2.4 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
HUAWEI Secoway USG9500	V200R001

12.2.5 原理描述

12.2.5.1 主备备份

这是 VRRP 提供 IP 地址备份功能的基本方式。主备备份方式需要建立一个虚拟备份组，功能上相当于一台虚拟路由器。该虚拟备份组包括一个 Master 和一个 Backup 设备。

- 正常情况下，业务全部由 Master 承担。
- Master 出现故障时，Backup 设备接替工作。

12.2.5.2 VRRP 负载分担

在虚拟备份组中，允许一台 Router 为多个 VRRP 备份组作备份。通过多虚拟 Router 设置可以实现负载分担。负载分担方式是指多台 Router 同时承担业务，因此需要建立两个或更多的备份组。

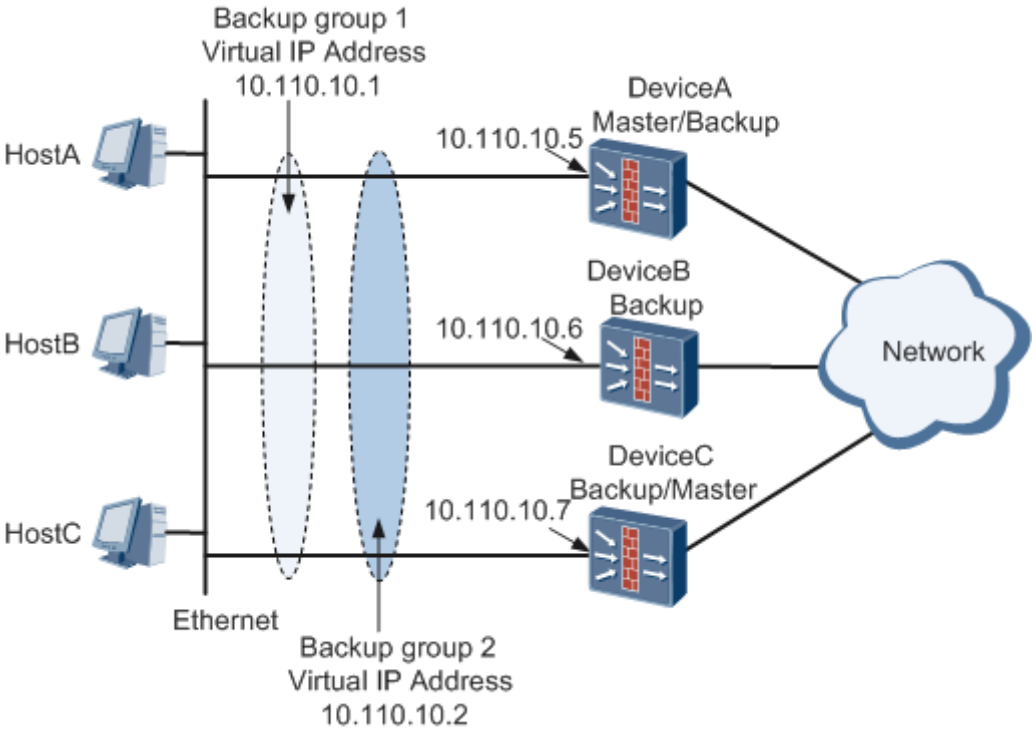
负载分担方式具有以下特点：

- 每个备份组都包括一个 Master 设备和若干 Backup 设备。
- 各备份组的 Master 设备可以不同。
- 同一台 Router 上的不同接口可以加入多个备份组，在不同备份组中有不同的优先级。

注意

USG9500 只支持两台设备进行 VRRP 负载分担。

图12-3 VRRP 负载分担示意图



如图 12-3 所示，配置两个备份组：备份组 1 和备份组 2。

- DeviceA 在备份组 1 中作为 Master，在备份组 2 中作为 Backup。

- DeviceB 在备份组 1 和 2 中都作为 Backup。
- DeviceC 在备份组 2 中作为 Master，在备份组 1 中作为 Backup。
- 一部分主机使用备份组 1 作网关，另一部分主机使用备份组 2 作为网关。

这样，可以达到分担数据流而又相互备份的目的。

12.2.5.3 虚拟 IP 地址 Ping 开关

VRRP 备份组使用虚拟 IP 地址，在 USG9500 上能够 Ping 通虚拟 IP 地址，由此可以比较方便的监控虚拟路由器的工作情况，但是带来可能遭到 ICMP 攻击的隐患。在 USG9500 中，提供了控制 Ping 通虚拟 IP 地址的开关命令，用户可以选择是否打开。

12.2.5.4 VRRP 安全

对于安全程度不同的网络环境，可以在报文头上设定不同的认证方式和认证字。

在一个安全的网络中，可以采用缺省设置：USG9500 对要发送的 VRRP 报文不进行任何认证处理，收到 VRRP 报文的 USG9500 也不进行任何认证，认为收到的都是真实的、合法的 VRRP 报文。这种情况下，不需要设置认证字。

12.3 VGMP

12.3.1 介绍

定义

VRRP 组管理协议 VGMP（VRRP Group Management Protocol）在 VRRP 基础上进行了扩展，弥补了 VRRP 在使用时存在的局限。VGMP 提出 VGMP 管理组的概念，将同一台 USG9500 上的多个 VRRP 备份组都加入到一个 VGMP 管理组。由管理组统一管理所有 VRRP 备份组。

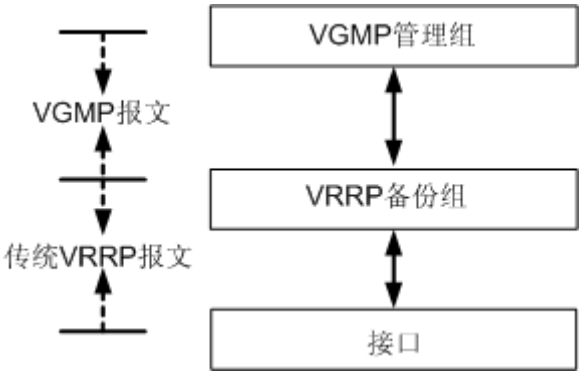
加入 VGMP 管理组中的 VRRP 分为 Master 和 Slave 状态：

- Master 状态的 VRRP：USG9500 上被管理的 VRRP 备份组状态是 Master（因接口 Down 而变成 Initialize 的除外），承担流量传输的任务，并定时发送 Hello 报文。
- Slave 状态的 VRRP：USG9500 上被管理的 VRRP 备份组状态是 Slave（因接口 Down 而变成 Initialize 的除外），不传输流量，处于监听状态，一旦 Master USG9500 出现故障，Slave USG9500 将竞选成为 Master USG9500。

VGMP 管理组相当于在 VRRP 备份组的基础上叠加了一层逻辑层。VGMP 管理组之间通过 VGMP 报文进行信息交互，VRRP 备份组和接口之间通过传统 VRRP 报文进行交互。

VGMP 管理组、VRRP 备份组和接口的协议层次关系如图 12-4 所示。

图12-4 VGMP 管理组、VRRP 备份组和接口的协议层次关系



VRRP 备份组向 VGMP 管理组汇报自己的状态，并接受 VGMP 管理组的管理。例如某备份组中某一接口或相关链路出现故障，导致备份组状态发生改变，此时备份组状态可能会影响到 VGMP 管理组状态。

VGMP 管理组提供以下功能：

- 状态一致性管理
  - 当两台 USG9500 优先级相同时，VGMP 管理组根据设备的配置决定 USG9500 的主备状态。
  - 当两台 USG9500 优先级不同时，优先级高的 USG9500 成为主用状态，优先级低的 USG9500 成为备用状态。
- 抢占管理

  对于加入 VGMP 管理组的 VRRP 备份组来说，无论各备份组内 USG9500 是否启动了抢占功能，抢占行为发生与否必须由 VGMP 管理组统一决定。
- 通道管理

  配置专门的数据通道传输 VGMP 报文，提高 VGMP 报文传输的可靠性。

目的

在配置大量 VRRP 备份组时，存在以下问题：

- 过多 VRRP 协议报文占用较大的链路带宽。
- 大量 VRRP 报文的处理对系统造成一定的负担。
- 每个 VRRP 备份组都要维护协议定时器，对系统来说也是个很大的开销。

按照传统的 VRRP 机制，VRRP 均为相对独立，且单独工作。由此，无法保证同一 USG9500 上各接口的 VRRP 状态都为主用或都为备用，即传统 VRRP 方式将无法实现 USG9500 的 VRRP 状态的一致性。引入 VGMP 后，优先级相同时，根据配置决定 USG9500 的主备状态，优先级不同时，根据优先级决定 USG9500 的主备状态，实现了状态的一致性。

12.3.2 规格

VGMP 特性的相关规格如下：

- 支持对 VGMP 管理组成员状态的统一管理和同时切换。
- 支持强制抢占、强制抢占时延和不抢占。

12.3.3 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
HUAWEI Secoway USG9500	V200R001

12.3.4 原理描述

12.3.4.1 VGMP 管理组之间的通讯

主备设备上的 VGMP 管理组依靠 VGMP 报文进行通讯，交换各自的运行状态信息，以维持主备状态的稳定，并在必要时协调主备状态的切换。

VGMP 报文是 VRRP 报文的扩展。VGMP 报文主要包括 HELLO 报文、状态切换请求和应答报文：

- HELLO 报文

  主用和备用设备定期互相向对端发送 HELLO 报文，通知对端设备本身的运行状态（包括优先级、VRRP 成员状态等）。

  VGMP HELLO 报文发送周期缺省为 1 秒。当备用设备在五个 HELLO 报文周期没有收到对端发送的 HELLO 报文时，会认为对端出现故障，从而将自己切换到主用状态。

- 状态切换请求和应答报文

  当主用设备上一个备份组成员出现故障时，VGMP 能立即感知到这个故障。此时 VGMP 会调整自己的优先级，并立即发送一个状态切换请求报文到对端。对端收到该报文后，会比较报文中的优先级和本身的优先级。如果本身优先级比报文中携带的优先级高，立即将自己切换到主用状态；发生故障的设备会立即将自己切换到备用状态。在管理组状态切换的同时，也会强制将管理组中的所有 VRRP 备份组成员的状态一起切换。

  如果对端的优先级比报文中的优先级低，则会回应一个拒绝状态切换的应答报文（NACK）。这样两端都不会进行状态切换。

  当 VGMP 感知到端口故障后，能立即主动发送状态切换请求报文，不再依赖于五次 HELLO 报文超时，因此大大提高了 USG9500 的故障响应速度。

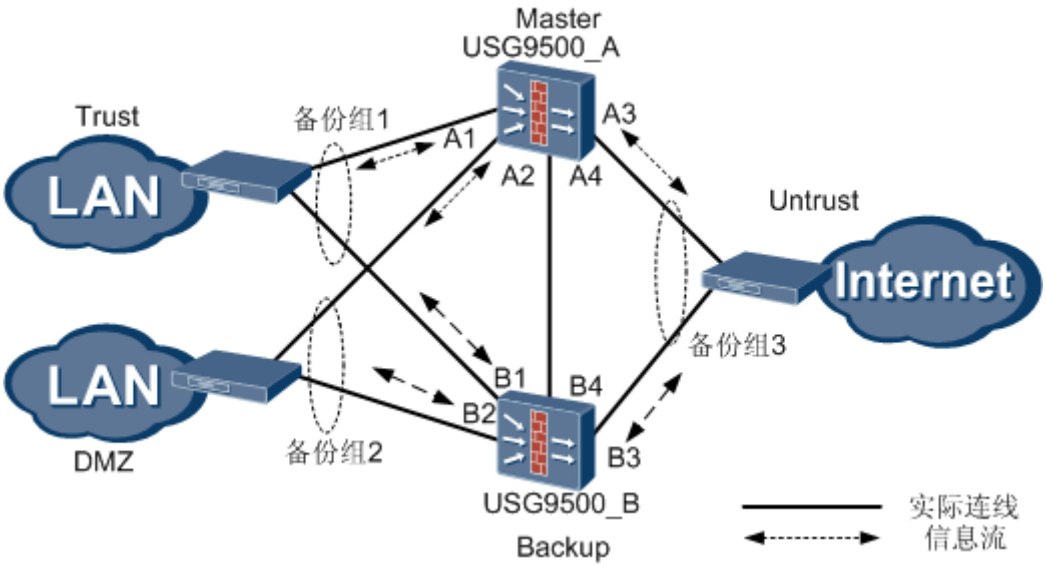
说明

如果是 USG9500 整机故障（如死机、掉电、重启等），还是只能依赖于五次 HELLO 报文超时的机制来发现故障。

12.3.4.2 VGMP 管理组、备份组、接口之间的关系

VGMP 管理组负责统一管理加入其中的所有 VRRP 备份组。每个 USG9500 上的管理组和备份组之间的关系如图 12-5 所示。

图12-5 VGMP 管理组和备份组之间的关系



USG9500_A 的接口	A1、A2、A3、A4
USG9500_B 的接口	B1、B2、B3、B4

- 两个 USG9500 上各接口之间的关系

  两个 USG9500 上的接口和安全区域的连接必须严格对应，包括接口插槽、类型、编号、相关配置等（IP 地址除外）。这就是说 USG9500_A 上的 A1 接口必须和 USG9500_B 上的 B1 接口完全一样，例如都为以太网接口、编号都为 1/0/0、都和备份组 1 关联，以此类推。
- 两个 USG9500 上 VRRP 备份组之间的关系

  两个 USG9500 上的备份组编号、构成必须完全一样。即 USG9500_A 上的 A1 接口关联备份组 1，A2 接口关联备份组 2，A3 接口关联备份组 3；则 USG9500_B 上的 B1、B2、B3 接口也必须分别关联备份组 1、2、3。
- 两个 USG9500 上各 VGMP 管理组之间的关系

  每个 USG9500 上可以配置一个 Master 管理组和一个 Slave 管理组。双机热备的两个 USG9500，一个 USG9500 上配置的 Master 管理组可以和另一个 USG9500 上配置的 Slave 管理组进行通信，这两个管理组的构成必须完全一样。如图 12-5 所示，在 USG9500_A 上配置了一个 Master 管理组，包括备份组 1、备份组 2、备份组 3。在 USG9500_B 上配置了一个 Slave 管理组，包括备份组 1、备份组 2、备份组 3。
- 同一 USG9500 上接口、备份组、管理组之间的关系

同一 USG9500（例如 USG9500_A）上，一个物理接口可以关联多个 VRRP 备份组。一个备份组能关联多个物理接口，对应多个虚拟 IP 地址。同一 VGMP 管理组可以包含多个备份组，但是相同备份组不能同时隶属于 Master 管理组和 Slave 管理组。

12.3.4.3 备份方式分类

通过接口、备份组、管理组之间的不同组合，实现了两台设备主备备份、负载分担两种备份方式，保证了业务不中断。

USG9500 双机热备的备份方式分为主备备份和负载分担两种方式：

设备内置 1 个 VGMP 管理组：

- 当两台 USG9500 都正常工作时，初始优先级相同，此时根据设备的配置确定备份方式是主备备份或负载分担。
- 当其中一台 USG9500 发生故障，优先级降低时，此时根据优先级高低确定两台设备的主备状态。

优先级计算

USG9500 内置一个 VGMP 管理组，初始优先级为 45000。每块单板的优先级计算如下：

- 每块 LPUG 接口板的优先级为 1000，若有 n 块 LPUG 接口板在位，则设备的优先级为 45000 + 1000 * n。
- 每块 LPUF-21、LPUF-40-A 接口板的优先级计算为：1000 * （接口板上插卡个数）。若有 n 块接口板在位，则设备的优先级为 45000 + n 块接口板的优先级之和。
- 每块 SPUA 板的优先级计算为：2 * （业务板上 CPU 个数），若有 n 块 SPUA 板在位，则设备的优先级为 45000 + n 块 SPUA 板的优先级之和。
- 每块 ESPU 板的优先级计算为：2 * 2 = 4。若有 n 块 ESPU 板在位，则设备的优先级为 45000 + 4 * n。
- SFU 板默认不参加优先级计算，但是为了提高系统的可靠性，当 SFU 板在位个数小于指定的在位个数时，优先级会降低，具体如下：
  - USG9310 和 USG9320 SFU 是 n+1 备份(n 默认值为 3)。
  - USG9520 无 SFU，不参与 SFU 优先级计算。
  - USG9560 SFU 是 n+1 备份(n 默认值为 2)。
  - USG9580 SFU 是 n+1 备份(n 默认值为 3)。

  当 SFU 板在位个数 m 小于指定在位个数 n 时，设备优先级会降低，降低的优先级计算公式为 45000 -2 * （n-m）。

每个链路或接口故障时，监控该链路或接口的 VGMP 管理组的优先级将会降低 2。

单板故障时，监控该单板的 VGMP 管理组的优先级降低的数值计算方法与上面优先级计算方法一致。

整机故障时，即 USG9500 在 5 个 HELLO 报文周期内没有收到对端发送的 HELLO 报文时，会认为对端出现故障，从而将自己切换到主用状态。

以 USG9580 为例：

1. 执行命令 **display device pic-status**，查看业务子卡的信息，然后计算 LPU 的优先级之和。

```
[USG9500] display device pic-status
Pic-status information in Chassis 1:
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
SLOT PIC Status    Type               Port_count Init_result   Logic down
1    0   Registered ETH_10G_LAN         1        SUCCESS      SUCCESS
2    0   Registered LAN_WAN_10G_CARD     1        SUCCESS      SUCCESS
2    1   Registered ETH_12XGE_CARD       12       SUCCESS      SUCCESS
16   1   Registered ETH_20XGF_NB_CARD    20       SUCCESS      SUCCESS
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

   1 号,16 号槽位的 LPU 各有一个插卡，2 号槽位的 LPU 有两个插卡，因此 LPU 的优先级之和为 1000+1000+1000*2=4000。

2. 执行命令 **display device**，查看 SPU 的信息，然后计算 SPU 的优先级。

```
[USG9500] display device

Slot #    Type      Online    Register     Status     Primary
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
1         LPU       Present   Registered   Normal     NA
2         LPU       Present   Registered   Normal     NA
7         SPU       Present   Registered   Normal     NA
15        SPU       Present   Registered   Normal     NA
16        LPU       Present   Registered   Normal     NA
17        MPU       Present   NA           Normal     Master
18        MPU       Present   Registered   Normal     Slave
19        SFU       Present   Registered   Normal     NA
20        SFU       Present   Registered   Normal     NA
21        SFU       Present   Registered   Normal     NA
22        SFU       Present   Registered   Normal     NA
23        CLK       Present   Registered   Normal     Master
24        CLK       Present   Registered   Normal     Slave
27        FAN       Present   Registered   Normal     NA
28        FAN       Present   Registered   Normal     NA
29        FAN       Present   Registered   Normal     NA
30        FAN       Present   Registered   Normal     NA
31        CMU       Present   Registered   Normal     NA
```

   执行命令 **display version slot 7/0** 以及 **display version slot 15/0**，查看两块 SPU 上 CPU 的个数。

   7 号槽位 SPU 有 4 个 CPU，优先级为 8（4*2），15 号槽位 SPU 有 2 个 CPU，优先级为 4（2*2），因此 SPU 优先级之和为 12。

3. USG9580 默认需要 3 块 SFU 在位，如上显示信息所示，目前有 4 块 SFU 在位，此时 SFU 不参与优先级计算。

经过上述三步计算，设备的优先级为 45000+4000+12=49012。

执行命令 **display hrp state**，查看设备的优先级为 49012，与计算结果一致。

```
[USG9500] display hrp state
 Role: active, peer: standby
 Running priority: 49012, peer: 49012
 Core state: load-balance, peer: load-balance
```

```
Backup channel usage: 0%
Stable time: 0 days, 0 hours, 53 minutes
```

主备备份

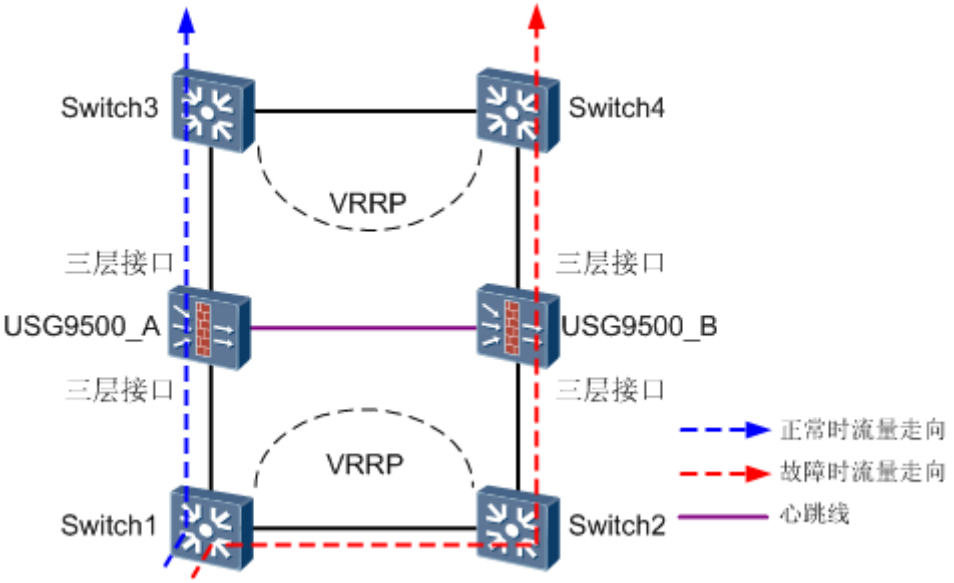
在网络入口处同时部署两台 USG9500 设备，USG9500_A 为主用设备，处理业务流量；USG9500_B 为备用设备，不处理业务流量，只通过心跳线同步 USG9500_A 的配置命令和状态信息。当 USG9500_A 的接口、链路或整机发生故障时，USG9500_B 立即切换为主用设备接替 USG9500_A 处理业务流量，从而保证业务的不中断。

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几种组网。

- **业务接口工作在三层，上下行连接交换机**

  如图 12-6 所示，USG9500 的上、下行业务接口工作在三层，与二层交换机直连。

图12-6 业务接口工作在三层，上下行连接交换机的组网



此组网是 USG9500 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。

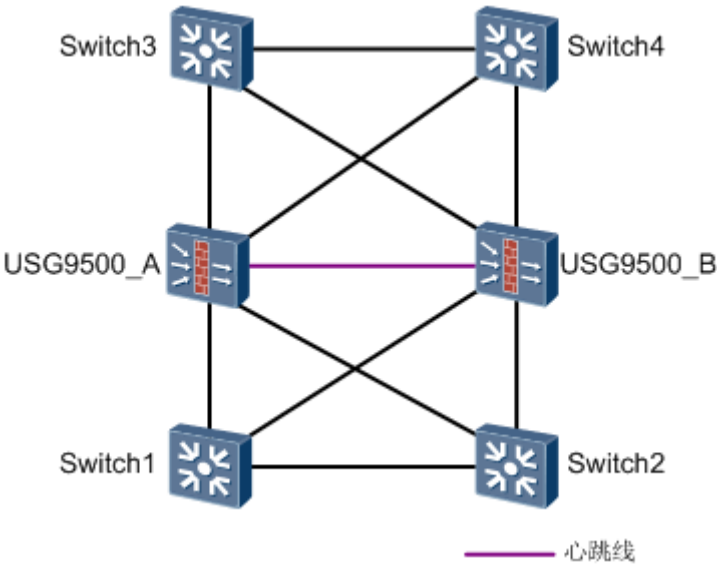
在此组网中，两台 USG9500 的上、下行业务接口各配置一组 VRRP 备份组，状态分别为 Master 和 Slave，由 VGMP 管理组监控 VRRP 备份组状态，以监控上下行链路状态。上、下行网络中的主机设备通过静态路由，将下一跳分别指向 VRRP 备份组的虚拟 IP 地址。

交换机之间会转发数据报文，所以交换机之间需要保证链路的畅通和足够的带宽。

在图 12-6 的基础上，将 USG9500_A 的上行接口与 Switch4 相连，下行接口与 Switch2 相连；将 USG9500_B 的上行接口与 Swith3 相连，下行接口与 Swith1 相连。

这样就组成了双机热备的全冗余组网，如图 12-7 所示。

图12-7 双机热备全冗余组网

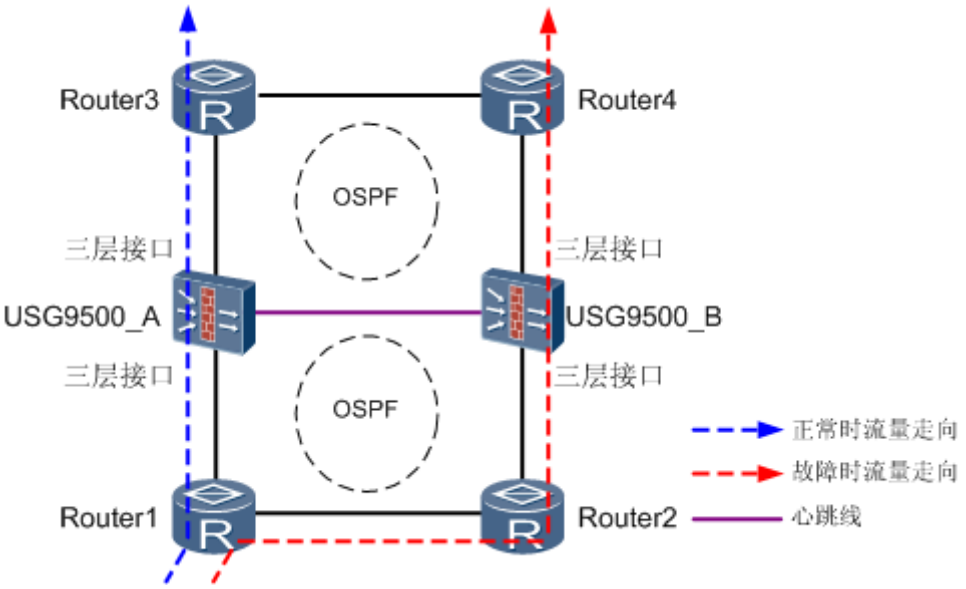


双机热备的全冗余组网能够进一步提升网络可靠性，避免多条链路故障时业务中断。

- **业务接口工作在三层，上下行连接路由器**

  如图 12-8 所示，USG9500 的上、下行业务接口工作在三层，与路由器直连。USG9500 与上、下行路由器之间运行 OSPF 协议。

图12-8 业务接口工作在三层，上下行连接路由器的组网



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 USG9500 作为防火墙可以采用这种组网方案。

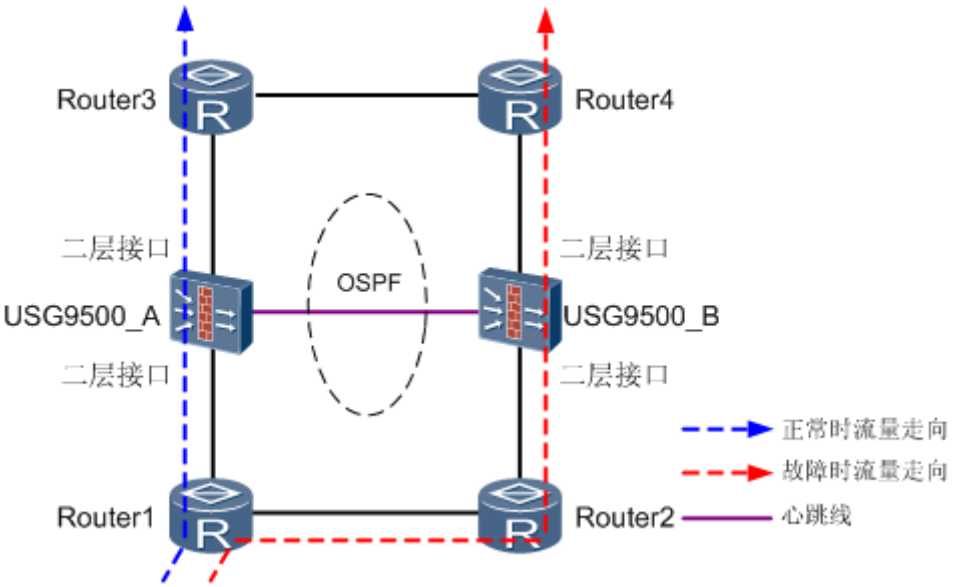
在此组网中，由于 USG9500 上下行直连设备为三层路由器，而路由器不能透传 VRRP 报文，所以接口上不能配置 VRRP 备份组。在这种情况下，由 USG9500 的 VGMP 管理组直接监控接口状态，USG9500 与上下行设备之间运行 OSPF 协议，同时在备用设备上配置 **hrp standby-device** 命令，备用设备上 OSPF COST 值会调整到 65500，流量由主用设备承担。

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。

- **业务接口工作在二层，上下行连接路由器**

  如图 12-9 所示，USG9500 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 USG9500 的上下行业务接口加入到同一个 VLAN。

图12-9 业务接口工作在二层，上下行连接路由器的组网



USG9500 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 USG9500，可以采用此种组网。

在此组网中，USG9500 透明接入到网络中，USG9500 自身不运行 OSPF 路由协议。所以如果需要实现主备备份，则需要在 USG9500 的上下行路由器上配置不同 COST 值的 OSPF 路由；如果需要实现负载分担，则需要在 USG9500 的上下行路由器上配置相同 COST 值的 OSPF 路由。

USG9500 业务接口工作在二层，在备用设备上配置 **hrp standby-device** 命令，则备用设备上的 VLAN 会被禁用，不能转发流量，流量油主用设备上的 VLAN 转发。

负载分担

在网络入口处同时部署两台 USG9500 设备。正常情况下，业务流量分别送到两台 USG9500 上进行处理。每台 USG9500 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 USG9500 的配置及状态信息。

如果 USG9500_A 的接口、链路或整机发生故障，USG9500_B 会立即承担全部业务流量的转发。

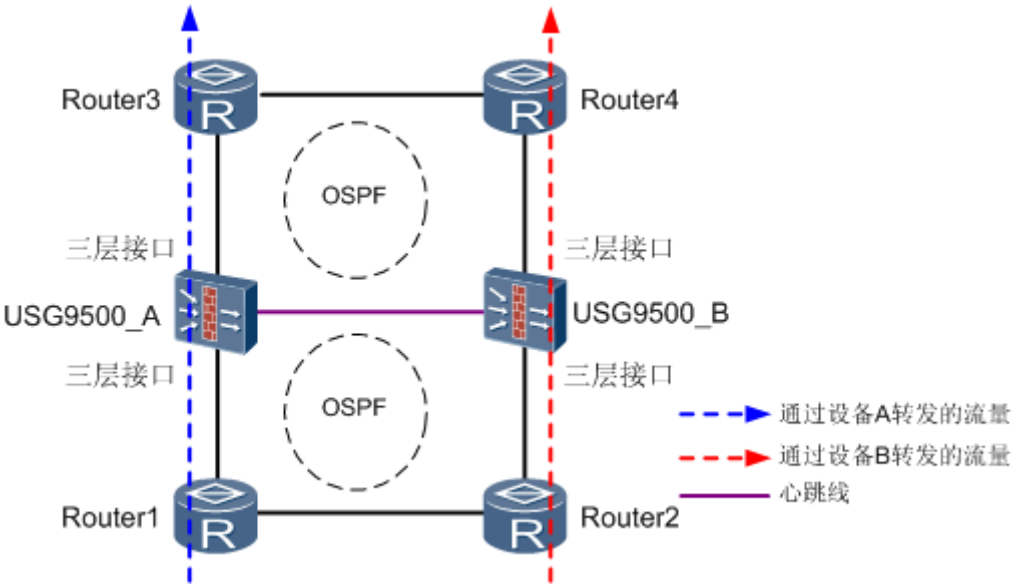
根据主备设备的上下行业务接口类型，以及上下行设备的类型，负载分担的双机热备场景可以分成以下两种组网。

说明
在负载分担场景中，我们推荐您的 USG9500 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 12-10 所示，USG9500 的上、下行业务接口工作在三层，与路由器直连。USG9500 与上、下行路由器之间运行 OSPF 协议，使业务流量能通过 USG9500_A 和 USG9500_B 共同转发。

图12-10 业务接口工作在三层，上下行连接路由器的组网

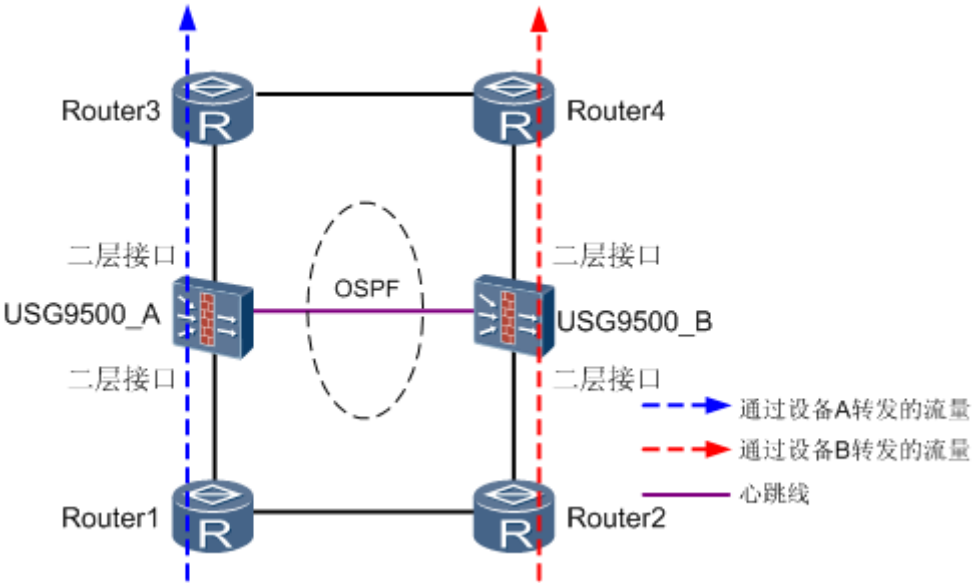


此组网网络拓扑简单，是比较常用的一种组网方式。对已经存在的都是路由器的网络，如果需要增加 USG9500 作为防火墙且希望两台设备共同处理流量时，可以采用这种组网方案。

- **业务接口工作在二层，上下行连接路由器**

  如图 12-11 所示，USG9500 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 USG9500_A 和 USG9500_B 共同转发。每台 USG9500 的上下行业务接口加入到同一个 VLAN。

图12-11 业务接口工作在二层，上下行连接路由器的组网



在此组网中，USG9500 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 USG9500，且希望两台 USG9500 共同处理业务时，可以采用此种组网。

12.4 HRP

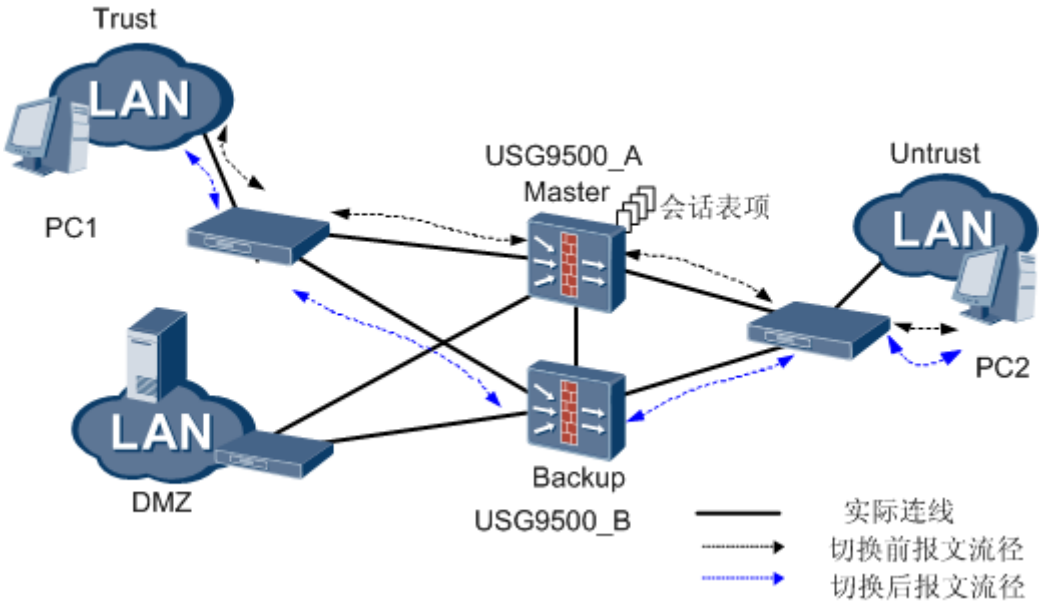
12.4.1 介绍

定义

USG9500 是状态防火墙，对于每一个动态生成的会话连接，其上都有一个会话表表项与之对应。如图 12-12 所示。

说明
USG9500 只支持两台设备之间运行 VRRP 和 HRP 协议。

图12-12 USG9500 主备备份的典型数据路径组网图



假设采用主备备份方式，USG9500_A 作为主用设备并承担所有数据传输任务，其上创建了很多动态会话表项；而 USG9500_B 由于处于备份状态，没有任何流量经过。

如果 USG9500_A 出现故障或相关链路出现问题，USG9500_B 将会切换状态而变成新的主用，并开始承担传输任务。

如图 12-12 所示，如果状态切换前，会话表项和配置命令没有备份到 USG9500_B，则先前经过 USG9500_A 的所有会话都会因为无法匹配 USG9500_B 的会话表而中断，从而影响业务正常进行。

为此，华为公司推出了 HRP（Huawei Redundancy Protocol）。HRP 是承载在 VGMP 报文上进行传输的，用于在主用设备和备用设备之间备份关键配置命令和会话表状态信息。

目的

为了实现主用设备出现故障时能由备用设备平滑地接替工作，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。启动 HRP 双机热备份功能后，关键配置命令和会话表状态信息会实时同步备份到备用 USG9500。如果主用 USG9500 发生故障，备用 USG9500 能够平滑地接替工作。

12.4.2 规格

HRP 特性的相关规格如下：

- 支持实时备份
- 支持自动备份
- 支持手工批量备份连接状态信息
- 支持会话快速备份

- 支持来回路径不一致组网
- 支持根据 HRP 状态调整 OSPF、OSPFv3 以及 BGP 的 Cost 值功能

12.4.3 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
HUAWEI Secoway USG9500	V200R001

12.4.4 原理描述

12.4.4.1 配置设备的主从划分

当采用负载分担方式时，网络中存在两台主用设备，用户可能在两台主用设备上输入了很多命令。当其中一台主用设备出现故障时，如何在这两台 USG9500 之间备份信息、需要备份哪些命令以及备份方向都是需要考虑的问题。

为了避免备份时混乱，USG9500 中引入了配置主设备、配置从设备概念。发送配置备份内容的 USG9500 称为配置主设备，接收配置备份内容的 USG9500 称为配置从设备。一台 USG9500 要想成为配置主设备，必须具备如下条件：

- 只有 VGMP 管理组中状态为主用的 USG9500 才有机会成为配置主设备。
- 在负载分担方式下，参与双机热备份的两台 USG9500 都是主用，此时先配置启动 HRP 功能的 USG9500 会成为配置主设备。

除非配置主设备出现故障或者退出 VRRP 备份组，否则配置主设备与配置从设备之间不进行转换，从而保证了配置主设备的稳定性。

12.4.4.2 配置命令和状态信息的备份

目前，USG9500 的双机热备份功能支持配置命令和连接状态信息的备份，可以通过自动备份、手工批量备份（只能备份连接状态信息）和快速备份三种方式实现。

备份内容

USG9500 备份的内容包括：

- 配置命令

  USG9500 备份的配置命令包括以下方面。
  - ACL（Access Control List）包过滤命令
  - 攻击防范命令
  - 黑名单命令

包括黑名单的启动命令、手工添加黑名单表项命令。

- NAT 命令

  包括 NAT 地址池、NAT Server 以及 NAT 在域间应用的命令。

- 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
- IPsec 命令
- AAA 命令
- 清除会话表项的命令
- IPS 命令

  其中 IPS 的升级服务只支持在线自动升级命令的备份，不支持本地升级、出厂默认升级包安装及版本回退等命令的备份。

- GTP 命令
- 部分 IPv6 配置命令

说明

- SSH 的相关命令不备份，需要在主备 USG9500 上分别配置。
- 接口绑定虚拟防火墙的命令不备份，需要在主备 USG9500 上分别配置。
- L2TP、GRE 的相关命令不备份，需要在主备 USG9500 上分别配置。
- 二进制日志的相关命令不备份，需要在主备 USG9500 上分别配置。
- 配置双机热备份前，需要在备 USG9500 上删除以下特性相关命令的配置：
- AAA（认证方案、计费方案、授权方案、域）。
- radius 服务器模板、hwtacacs 服务器模板。

- 状态信息

  USG9500 备份的状态信息包括以下方面。

  - USG9500 生成的会话表表项
  - 源 IP 监控表
  - Servermap 表项，包括使用 QQ、MSN、STUN（Simple Traversal of UDP Through Network Address Translators）协议、NAT Server、NO-PAT 地址分配表项、快速备份时的 ASPF 生成的 Servermap 表项

备份方向

连接状态会互相备份，由系统决定需要备份的连接状态的内容。

配置命令则只能进行单向备份，即备份方向只能从配置主用 USG9500 到配置备用 USG9500，不能反向备份。

自动备份

自动备份方式为自动实时备份。下面针对配置命令的备份和连接状态信息的备份两方面分别进行说明：

- 配置命令的备份

在主用 USG9500、备用 USG9500 都正常工作的情况下，如果启动自动实时备份功能，则在主用 USG9500 上每输入一条双机热备份需要的命令时，此配置命令将被传送到备用 USG9500 并执行；如果在主用 USG9500 上输入双机热备份不需要的命令，则该命令仅在主用 USG9500 上执行，不会被传送到备用 USG9500。对于在备用 USG9500 上执行的命令，不会被传送到主用 USG9500。当备用 USG9500 未工作或工作异常时，自动备份无法进行。

- 连接状态信息的备份

  在 USG9500 运行过程中，当一台设备上产生了需要备份的连接状态信息时，USG9500 会自动将连接状态信息备份到另外一台设备进行状态处理和更新。

手工批量备份

手工批量备份连接状态信息如下：

在一台设备上执行手工批量备份的命令后，USG9500 会向对端设备进行连接状态备份，对端进行连接状态处理和更新。

快速备份

当 USG9500 工作于双机热备份组网环境下，如果报文的来回路径不一致，即业务的来回两个方向的报文分别从不同的 USG9500 经过，主用 USG9500 的信息没有备份到备用 USG9500，备用 USG9500 会将到达的报文丢弃。

为防止上述现象发生，可以通过快速备份会话，将主用 USG9500 的相应的会话表表项快速备份到备用 USG9500，使返回报文在备用 USG9500 上能够查找到会话表表项，从而能够通过备用 USG9500，保证内外部用户的会话不中断。

12.5 IP-link

12.5.1 介绍

定义

IP-Link，即链路可达性检查，通过 USG9500 定时地向指定的目的 IP 进行 ICMP 回显请求，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 USG9500 直接相连或不直接相连的链路状态，保证业务持续通畅。

12.5.2 规格

IP-Link 特性的相关规格如下：

- 支持通过 ICMP 方式检测链路可达性。
- 支持双机热备份环境下自动侦测链路，触发 USG9500 进行 HRP 主备切换。
- 支持根据检测结果进行静态路由优先级调整。
- 支持根据检测结果触发策略路由切换。

12.5.3 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
HUAWEI Secoway USG9500	V200R001

特性依赖

IP-Link 应用于双机热备份环境自动侦测链路，需要启动 HRP 功能，且配置 HRP 绑定 IP-Link 时，USG9500 会对 HRP 的优先级进行相关调整，触发主备 USG9500 切换，从而保证业务能够持续通畅。

12.5.4 原理描述

实现原理

通过 USG9500 定时地向指定的目的 IP 进行 ICMP 回显请求，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

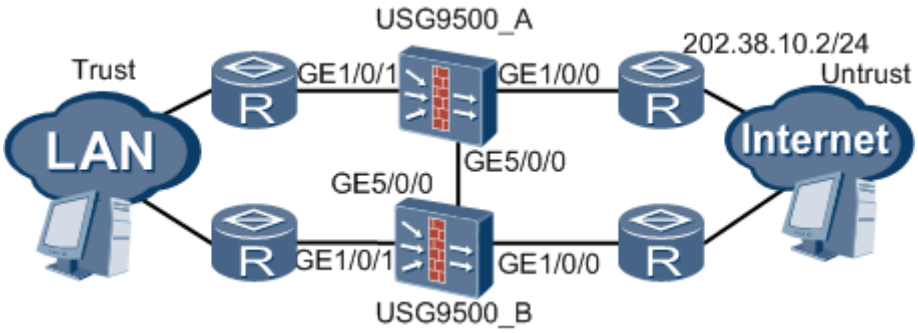
IP-Link 主要应用在以下三种环境：

- 双机热备份环境

  当 USG9500 工作于双机热备份环境时，IP-Link 自动检查后发现链路故障影响主备业务，通过配置 HRP 绑定 IP-Link，USG9500 会对 HRP 的优先级进行相关调整，触发主备 USG9500 切换，从而保证业务能够持续流通。

  配置 HRP 绑定 IP-Link 后，可以检测到与 USG9500 直接相连或不直接相连的接口或链路状态。如图 12-13 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 HRP 主备切换，保证业务正常进行。

图12-13 双机热备环境下的 IP-Link 链路可达性检查

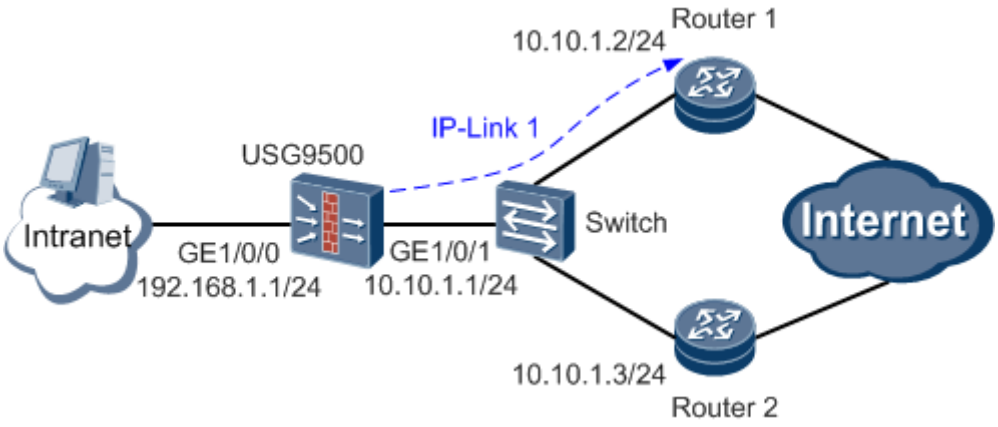


- 静态路由环境

  当 IP-Link 自动检查发现链路故障时，USG9500 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 12-14 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。

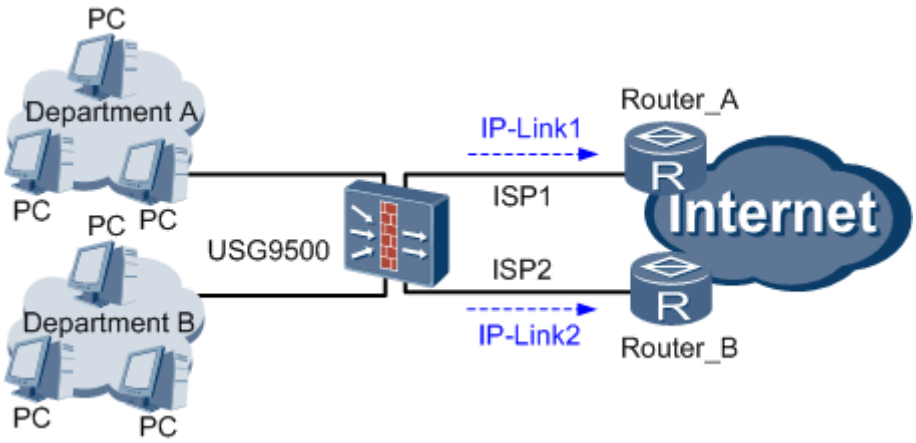
图12-14 静态路由环境下的 IP-Link 链路可达性检查



- 策略路由环境

  如图 12-15 所示，当 IP-Link 自动检查发现链路故障时，系统可以触发链路绑定的策略路由失效，这样 USG9500 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12-15 策略路由环境下的 IP-Link 链路可达性检查



12.6 Link-group

12.6.1 介绍

定义

Link-group 是指将多个物理接口的状态相互绑定，组成一个逻辑组。如果组内任意接口因故障而状态变为 fault，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。

通过配置 Link-group 与防火墙业务板、IPS 业务板或 Anti-DDoS 业务板绑定，当 USG9500 上的防火墙业务板、IPS 业务板或 Anti-DDoS 业务板被拔出，或者三种板卡发生故障，状态变为 Abnormal 时，USG9500 将 Link-group 组中有效接口的状态变为 DOWN。与 USG9500 相连的路由设备即可感知到接口状态的变化，随之刷新路由表，重新进行路由选路，避免业务流量继续送往业务板故障的 USG9500。

可以加入到 Link-group 的物理接口类型有：GigabitEthernet 和 POS。允许将不同类型的接口加入到同一个 Link-group 中。

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。

12.6.2 规格

Link-group 特性的相关规格如下：

- 支持将不同类型的物理接口加入同一个 Link-group 管理组
- USG9500 最多支持 64 个 Link-group 管理组

12.6.3 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
HUAWEI Secoway USG9500	V200R001

12.6.4 原理描述

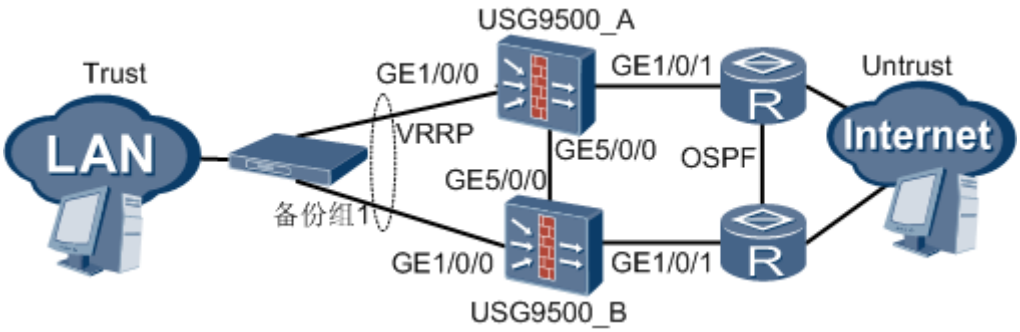
组网应用

如图 12-16 所示，在 OSPF（Open Shortest Path First）和双机热备份混合组网中，USG9500_A 和 USG9500_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，USG9500 使用 HRP 协议监视与路由器相连的接口 GE1/0/1。

当 USG9500_A 与交换机相连的接口 GE1/0/0 出现故障时，触发主备倒换，主用设备由 USG9500_A 转换成 USG9500_B。然而，由 USG9500 和路由器组成的 OSPF 区域无法感知到接口 GE1/0/0 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 USG9500_A，造成业务中断。

将接口 GE1/0/0 与接口 GE1/0/1 加入同一 Link-group 管理组后。如果接口 GE1/0/0 因故障而状态变为 fault，将会触发接口 GE1/0/1 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12-16 OSPF 和双机热备份混合组网图



12.7 BFD

12.7.1 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应用。

BFD 用于检测转发引擎之间的通信故障。具体来说，BFD 对系统间的、同一路径上的一种数据协议的连通性进行检测，这条路径可以是物理链路或逻辑链路，包括隧道。

可以把 BFD 看作是系统提供的一种服务：

- 上层应用向 BFD 提供检测地址、检测时间等参数。
- BFD 根据这些信息创建、删除或修改 BFD 会话，并把会话状态通告给上层应用。

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。

现有的故障检测机制主要包括：

- 硬件检测：例如通过 SDH（Synchronous Digital Hierarchy）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，其目标如下：

- 为相邻转发引擎之间的通道提供轻负荷的、快速的故障检测。这些故障包括接口、数据链路、甚至有可能是转发引擎本身的故障。
- 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。

12.7.2 规格

BFD 特性的相关规格如下：

- USG9500 单板支持创建的 BFD 会话最大个数为 512 个。
- USG9500 整机支持创建的 BFD 会话最大个数为 8192 个。

12.7.3 参考标准和协议

与 BFD 特性相关的参考标准与协议如下：

- draft-ietf-bfd-base-08
  Bidirectional Forwarding Detection
- draft-ietf-bfd-generic-04.txt
  Generic Application of BFD
- draft-ietf-bfd-multihop-06.txt
  BFD for Multihop Paths
- draft-ietf-bfd-v4v6-1hop-08.txt
  BFD for IPv4 and IPv6 (Single Hop)

12.7.4 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
HUAWEI Secoway USG9500	V200R001

12.7.5 原理描述

12.7.5.1 BFD 机制

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，如果一方在规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为事件间隔在彼此之间的路径上定时发送 BFD 控制报文。

说明
为满足快速检测的需求，BFD 草案规定发送间隔和接收间隔的时间单位是微秒。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。USG9500 支持的最小检测时间为 30 毫秒。

BFD 提供异步检测模式：

异步模式是 BFD 的主要操作模式。在这种模式下，BFD 会话建立起来后，两个系统之间相互周期性地发送 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报文，就认为此 BFD 会话的状态是 Down。

BFD 会话状态

BFD 会话有四种状态：Down、Init、Up 和 AdminDown。

- Down：会话处于 Down 状态或刚刚创建。
- Init：已经能够与对端系统通信，本端希望使会话进入 Up 状态。
- Up：会话已经建立成功。
- AdminDown：会话处于管理性 Down 状态。

会话状态通过 BFD 控制报文的 State 字段传递，系统根据自己本地的会话状态和接收到的对端会话状态驱动状态改变。

BFD 会话的建立方式

BFD 会话的建立有两种方式，即静态配置 BFD 会话和动态建立 BFD 会话。

BFD 通过控制报文中的 My Discriminator 和 Your Discriminator 区分不同的会话。静态和动态创建 BFD 会话的主要区别在于 My Discriminator 和 Your Discriminator 的配置方式不同。

- 静态配置 BFD 会话

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

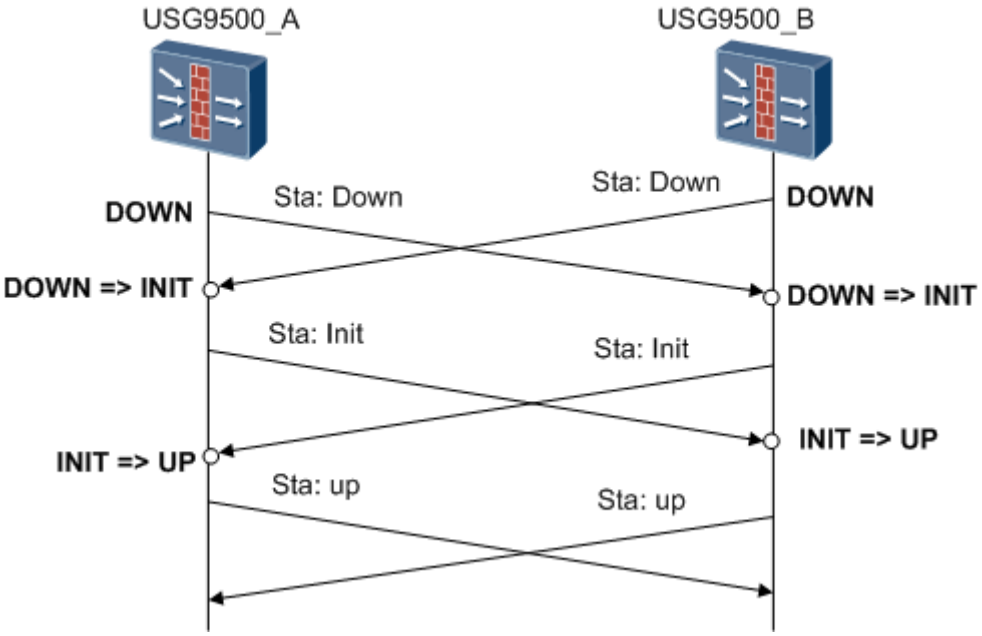
    当应用程序触发动态创建 BFD 会话时，系统分配属于动态会话标识符区域的值作为 BFD 会话的本地标识符。然后向对端发送 Your Discriminator 的值为 0 的 BFD 控制报文，进行会话协商。
  - 自学习远端标识符

    当 BFD 会话的一端收到 Your Discriminator 的值为 0 的 BFD 控制报文时，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 报文中 My Discriminator 的值，获取远端标识符。

BFD 会话的建立过程

BFD 状态机的建立和拆除都采用三次握手机制，以确保两端系统都能知道状态的变化。图 12-17 以 BFD 会话建立为例，简单介绍状态机的迁移过程。

图12-17 BFD 会话连接建立



1. USG9500_A 和 USG9500_B 各自启动 BFD 状态机，初始状态为 Down，发送状态为 Down 的 BFD 报文。对于静态配置 BFD 会话，报文中的 Your Discriminator 的值是用户指定的；对于动态创建 BFD 会话，Your Discriminator 的值是 0。
2. USG9500_B 收到状态为 Down 的 BFD 报文后，状态切换至 Init，并发送状态为 Init 的 BFD 报文。
3. USG9500_B 本地 BFD 状态为 Init 后，不再处理接收到的状态为 Down 的报文。
4. USG9500_A 的 BFD 状态变化同 USG9500_B。
5. USG9500_B 收到状态为 Init 的 BFD 报文后，本地状态切换至 Up。
6. USG9500_A 的 BFD 状态变化同 USG9500_B。

USG9500_A 和 USG9500_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 报文，则本地状态自动切换回 Down。

12.7.5.2 BFD for IP

在 IP 链路上建立 BFD 会话，利用 BFD 检测机制快速检测故障。

BFD for IP 支持单跳检测和多跳检测：

- BFD 单跳检测是指对两个直连系统进行 IP 连通性检测，这里所说的“单跳”是 IP 的一跳。在进行 BFD 单跳检测的两个系统中，对于一种给定的数据协议，在指定接口上只存在一个 BFD 会话。
- BFD 多跳检测是指 BFD 可以检测两个系统间的任意路径，这些路径可能跨越很多跳，也可能在某些部分发生重叠。

组网应用

典型应用一：

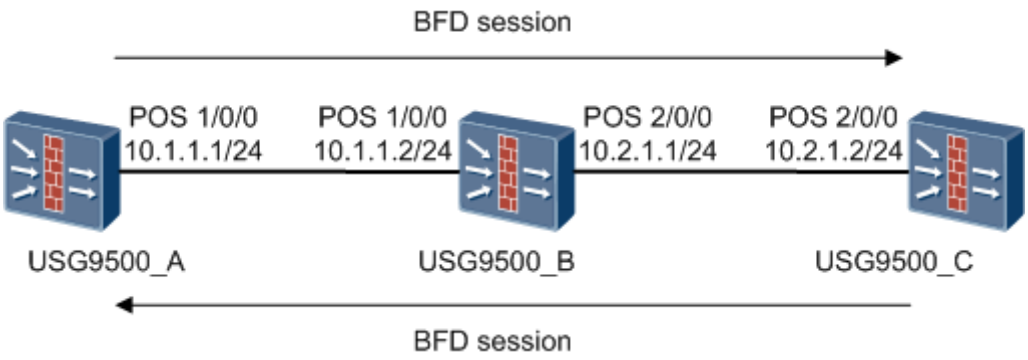
如图 12-18 所示，BFD 检测两台设备之间的单跳路径，BFD 会话绑定出接口。

图12-18 单跳 BFD for IP

典型应用二：

如图 12-19 所示，BFD 检测 USG9500_A 和 USG9500_C 之间的多跳路径，BFD 会话绑定对端 IP 但不绑定出接口。

图12-19 多跳 BFD for IP



12.7.5.3 组播 BFD

组播 BFD 用于检测无 IP 地址等三层属性的接口之间的链路连通性，达到链路故障快速检测。

通过将检测报文通过 IP 层发送组播检测报文，在所需检测链路之间的 USG9500 配上配置组播检测。本端发送组播报文，如果链路连通，则对端接口也可以收到这个组播报文，上送对端 BFD 应用，感知链路正常。对于二层 Trunk 链路，由于发送的是组播报文，IP 层转发不需要三层属性，直接下发链路层发送，快速检测链路的连通性。这里的 IP 是 BFD 模块配置的公认的组播地址 Default-IP，任何收到此 IP 的接口都将此报文上送 BFD 应用，完成 IP 转发。

组网应用

图12-20 组播 BFD 组网示意图

如图 12-20 所示，组播 BFD 可以快速检测接口之间的链路连通性。在 USG9500_A、USG9500_B 上配置 BFD 会话，使用缺省组播地址对绑定 GE1/0/0 接口的单跳链路进行检测，这样就能快速检测接口之间的链路连通性。

12.7.6 应用

12.7.6.1 BFD for HRP

在双机热备份组网环境下，当 USG9500 的上下行链路发生故障时，需要进行 HRP 主备状态的切换，以确保业务正常进行。

通过配置 BFD，可以快速检测到 USG9500 上下行链路的故障，也可以快速检测与 USG9500 不直接相连的链路的故障。

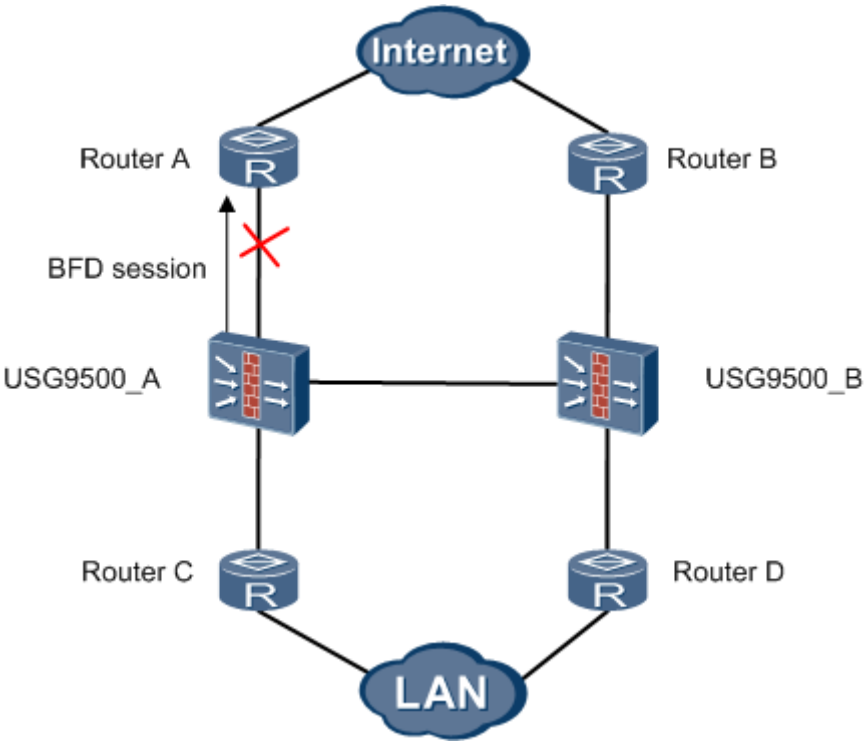
BFD for HRP 就是通过配置 HRP 绑定 BFD，在 BFD 会话快速检测到链路 DOWN 时，立即降低 USG9500 上 VGMP 管理组对应的优先级，从而触发 HRP 主备状态的快速切换。链路状态恢复正常时，被绑定的 BFD 能够检测到链路状态的变化，恢复 USG9500 上 VGMP 管理组的优先级。

说明
当 VGMP 管理组绑定的 BFD 会话 DOWN 时，对应的 VGMP 管理组的优先级减 2。HRP 主备状态是否切换是根据 VGMP 管理组的优先级计算规则来统一管理的。

组网应用

如图 12-21 所示，USG9500_A 和 USG9500_B 工作在双机热备环境下，正常情况下，USG9500_A 的 HRP 状态为主，USG9500_B 的 HRP 状态为备，业务流量通过 USG9500_A 进行转发。在 USG9500_A 上配置 BFD 会话来检测 USG9500_A 到 Router A 的链路，USG9500_A 上的 VGMP 管理组绑定 BFD 会话。当 USG9500_A 到 Router A 的链路发生故障时，BFD 会话检测到链路 DOWN，会立即降低 USG9500_A 上 VGMP 管理组的优先级，从而触发 HRP 主备状态的快速切换。切换后 USG9500_A 的 HRP 状态为备，USG9500_B 的 HRP 状态为主，业务流量通过 USG9500_B 进行转发。

图12-21 BFD for HRP 组网图



12.7.6.2 BFD for USR

BFD for USR（Unicast Static Route）用于支持 IPv4 单播静态路由，支持 IPv4 单播静态路由绑定后快速感知链路状态。

与动态路由协议不同，单播静态路由自身没有检测机制，当网络发生故障的时候，需要管理员介入。BFD for USR 特性可为公网 IPv4 单播静态路由绑定 BFD 会话，利用 BFD 会话来检测单播静态路由所在链路的状态。

BFD for USR 可为每条 IPv4 单播静态路由绑定一个 BFD 会话，当这条 USR 上绑定的 BFD 会话检测到链路故障（由 Up 转为 Down）后，BFD 会将故障上报路由管理模块，由路由管理模块将这条路由设置为“非激活”状态（此条路由不可用，从 IP 路由表中删除）。

当这条 USR 上绑定的 BFD 会话成功建立或者从故障状态恢复后（由 Down 转为 Up），BFD 会上报路由管理模块，由路由管理模块将这条路由设置为“激活”状态（此路由可用，加入 IP 路由表）。

12.7.6.3 BFD for OSPF

网络上的链路故障或拓扑变化都会导致 USG9500 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 for OSPF 就是将 BFD 和 OSPF 协议关联起来，通过 BFD 对链路故障的快速感应进而通知 OSPF 协议，从而加快 OSPF 协议对于网络拓扑变化的响应。

表 12-1 显示了 OSPF 协议在有、无 BFD 协议下收敛速度的数据。

表12-1 OSPF 协议收敛速度的数据

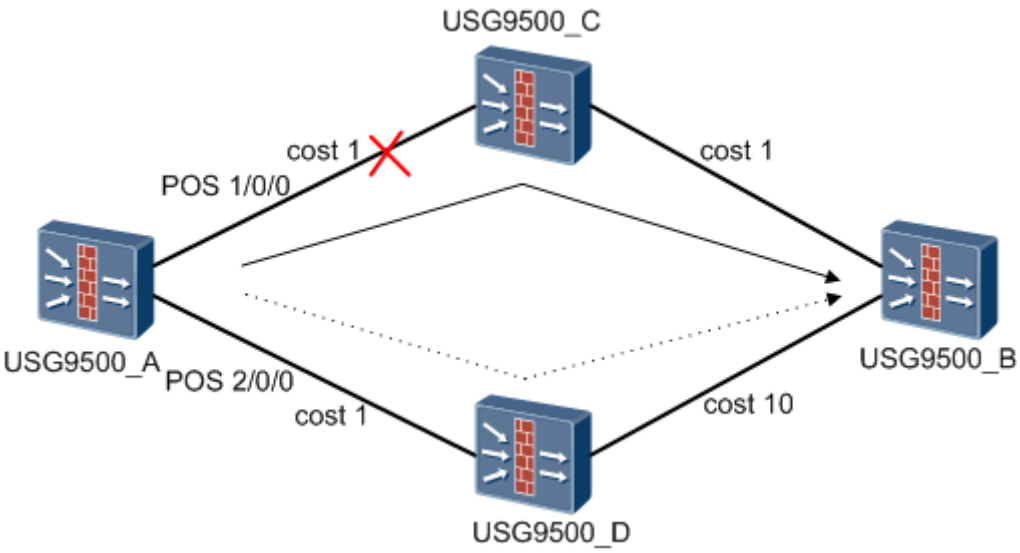
有无 BFD	链路故障检测机制	收敛速度
无 BFD	OSPF Hello Keepalive 定时器超时	秒级
有 BFD	BFD 会话 Down	毫秒级

如图 12-22 所示，USG9500_A 分别与 USG9500_C 和 USG9500_D 建立 OSPF 邻居关系，根据 OSPF 路由的选路规则，USG9500_A 到 USG9500_B 的路由出接口为 POS1/0/0，经过 USG9500_C 到达 USG9500_B。邻居状态到达 FULL 状态时通知 BFD 建立 BFD 会话。

1.当 USG9500_A 和 USG9500_C 之间链路出现故障，BFD 首先感知到并通知 USG9500_A。

2.USG9500_A 处理邻居 Down 事件，重新进行路由计算，新的路由出接口为 POS2/0/0，经过 USG9500_D 到达 USG9500_B。

图12-22 BFD for OSPF 组网图



12.7.6.4 BFD for BGP

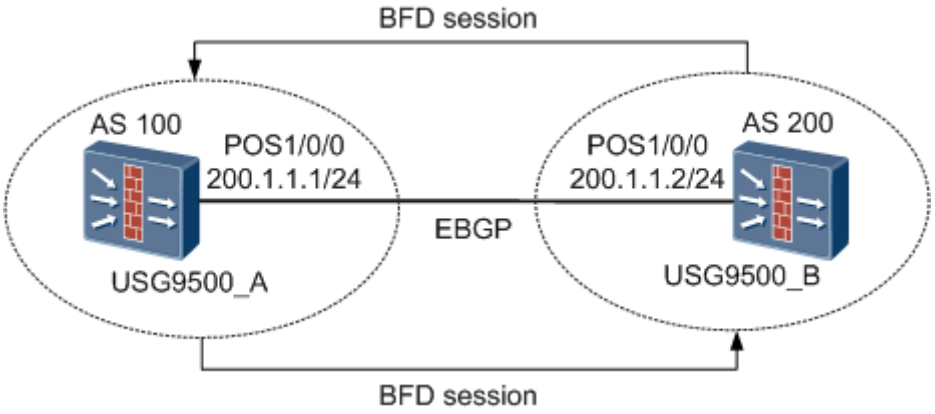
BGP 协议通过周期性的向对等体发送 Keepalive 报文来实现邻居检测机制。但这种机制检测到故障所需时间比较长，超过 1 秒钟，当数据达到吉比特速率级别时，将会导致大量的数据丢失，从而无法满足电信级网络高可靠性的需求。

因此，BGP 协议通过引入 BFD for BGP 特性，利用 BFD 的快速检测机制，迅速发现 BGP 对等体间链路的故障，并报告给 BGP 协议，从而实现 BGP 路由的快速收敛。

组网应用

如图 12-23 所示，USG9500_A 和 USG9500_B 分别属于 AS100 和 AS200，两台设备直接相连并建立 EBGP 连接。使用 BFD 检测 USG9500_A 和 USG9500_B 之间的 BGP 邻居关系，当 USG9500_A 和 USG9500_B 之间的链路发生故障时，BFD 能够快速检测到故障并通告给 BGP 协议。

图12-23 BFD for BGP 组网图



12.7.6.5 BFD for ISIS

通常情况下，ISIS 设定发送 Hello 报文的时间间隔为 10 秒钟，宣告邻居 Down 的时间即相邻设备失效的时间一般配置为 Hello 报文间隔的 3 倍。若在相邻设备失效时间内没有收到邻居发来的 Hello 报文，将会删除邻居。设备能感知到邻居故障的时间最小也是秒级。在高速的网络环境中，这将导致报文大量丢失。

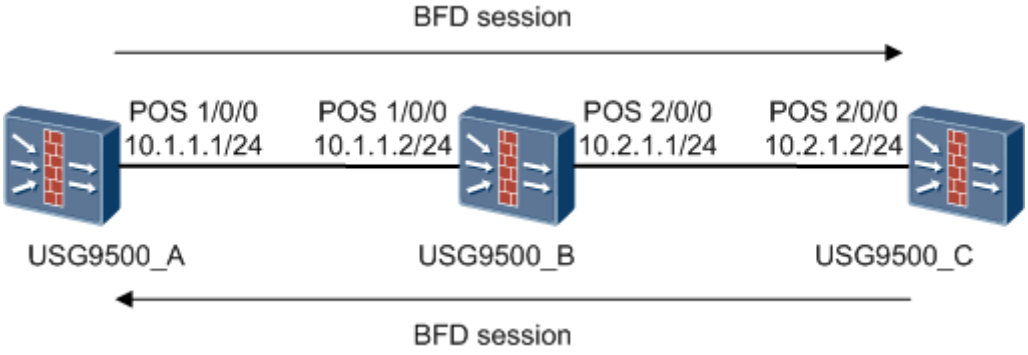
BFD for ISIS 是指 BFD 会话由 ISIS 协议动态创建，不再依靠手工配置，当 BFD 检测到故障时，通过路由管理通知 ISIS 协议，由协议进行相应邻居 Down 处理，快速更新 LSP 信息和进行增量路由计算，从而实现 ISIS 路由的快速收敛。

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

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

组网应用

图12-24 BFD for ISIS 组网图



在 USG9500_A、USG9500_B 和 USG9500_C 上使能 BFD 后，当 USG9500_A 和 USG9500_B 之间的链路故障时，BFD 快速检测到故障并通过 RM 模块通告给 ISIS 协议，ISIS 将此接口关联的邻居的状态置为 Down，从而触发 ISIS 拓扑计算，同时更新 LSP 使得其他邻居（如 USG9500_B 的邻居 USG9500_C）及时收到 USG9500_B 的更新 LSP，最终实现了网络拓扑的快速收敛。

13 系统管理

关于本章

13.1 信息中心

13.1.1 介绍

定义

信息中心是路由设备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路由设备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）。

   说明
   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13-1 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 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 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。

13.1.2 参考标准和协议

本特性的参考资料清单如下：

- RFC 3164: The BSD syslog

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
HUAWEI Secoway USG9500	V200R001SPC00

13.1.4 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路由设备内部运行的轨迹。

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 16 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 512 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。

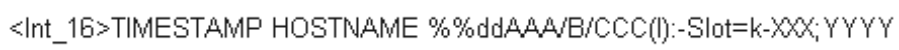
  缺省情况下，诊断日志向诊断日志文件方向输出。

- 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 13-1 所示：

图13-1 日志输出格式



各字段的详细说明见表 13-2。

表13-2 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	时间戳有 5 种格式可供选择。 • boot 型：相对时间类型。调试信息缺省采用 boot 型时间戳。 • 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。 • short-date 型：系统时间类型。不含有年份信息。 • format-date 型：另一种系统时间形式。 • none 型：信息中不包含时间戳。 时间戳与主机名之间用一个空格隔开。
HOSTNAME	本地的系统名	缺省是“USG9500”。 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	• l：日志信息 • T：告警信息 • d：debugging 信息 • D：诊断日志信息
-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 信息是路由设备发送到网管设备的信息。在路由设备上使能了 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”），路由设备会生成 trap 信息并发往指定的目的地址。如果路由设备与网管主机之间路由可达，网管软件就能够接收到由路由设备发出的 trap 信息。

另外在路由设备中有一个告警缓冲区，用于储存路由设备产生的 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路由设备产生的 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13-2 告警输出格式

TimeStamp HostName ModuleName/Severity/Brief:Description

各字段的详细说明见表 13-3。

表13-3 告警记录格式说明

字段	字段含义	说明
TimStamp	时间戳，信息输出的时间	时间戳有 5 种格式可供选择。 • boot 型：相对时间类型。调试信息缺省采用 boot 型时间戳。 • 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。 • short-date 型：与 date 型的唯一区别是，short-date 型时间戳不含年份。 • format-date 型：另一种系统时间形式。 • none 型：信息中不包含时间戳。 时间戳与主机名之间用一个空格隔开。
HostName	本地的系统名	缺省是“USG9500”。 主机名与模块名之间用一个空格隔开。
ModuleName	模块名	用来表示产生告警的模块名。

字段	字段含义	说明
Severity	严重级别	告警的严重级别。 • Critical（紧急） • Major（重要） • Minor（次要） • Warning（提示） • indeterminate（不确定）
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

调试信息是系统对路由设备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，路由设备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 方式登录到路由设备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路由设备进行调试时，可以控制调试信息输出的内容。

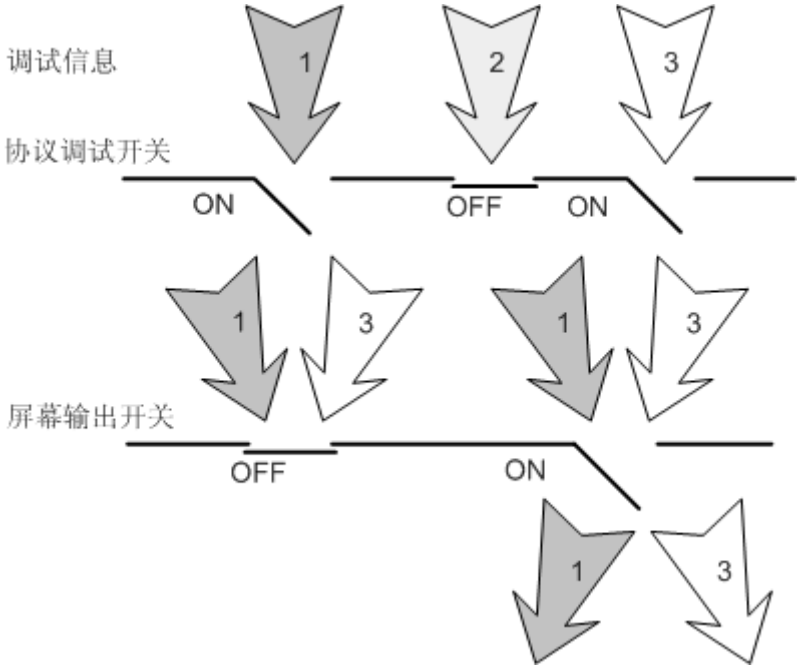
系统的命令行接口提供了种类丰富的调试功能，对于路由设备所支持的各种协议和功能，基本上都提供了相应的调试功能，可以帮助用户对错误进行诊断和定位。

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 13-3 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13-3 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 13-4 所示。

表13-4 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。
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，BFD 探测出设备不可达，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，BFD 探测的一次报文丢失，检测出错误协议报文等。

显示值	严重等级	描述
5	Notice	用于设备正常运转的关键操作信息。如用户对接口的 **shutdown** 命令，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 **display** 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。

13.2 SNMP

13.2.1 介绍

定义

SNMP（Simple Network Management Protocol，简单网络管理协议）的简称，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（如路由设备、交换机等）的方法。

目前，设备支持 SNMP v3 版本，并兼容 SNMP v1 版本和 SNMP v2c 版本。

- SNMP v1 采用团体名（Community Name）认证。团体名定义了 NM Station（Network Management Station）和 Agent 的关系，用来限制 NM Station 对 Agent 的访问。
- SNMP v2c 也采用团体名认证。它在兼容 SNMP v1 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。
- SNMP v3 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 NM Station 和 Agent 之间的传输报文进行加密，以免被窃听。

目的

SNMP 的主要目的是网络管理。

网络管理分为两类：

- 第一类是对网络应用程序、用户帐号（例如文件的使用）和存取权限（许可）的管理。它们都是与软件有关的网络管理问题。这里不作深入解释。

- 第二类是对构成网络的硬件即网元的管理，包括工作站、服务器、网卡、路由设备、网桥和集线器等等。通常情况下，这些设备与网络管理员所在的中心机房所在地距离很远。当这些设备有问题发生时，如果网络管理员可以自动地被通知，那么无疑是最佳的。但是路由设备不会像用户那样，当有一个应用程序发生问题时会打电话通知。

为了解决这个问题，设备制造商已经在一些设备中提供了网络管理的功能，这样网络管理工作站就可以远程询问设备的状态，同样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警告。

网络管理通常被分为四个部分：

- 被管理节点：即被监视的设备，简称网元。
- 代理：用来跟踪被管理设备状态的特殊软件或固件(firmware)。
- 网络管理工作站：与不同的被管理节点中的代理通信，并且显示这些代理状态的中心设备。
- 网络管理协议：被网络管理工作站和代理用来交换信息的协议。

13.2.2 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 rameworks

- RFC 3412: Message Processing and Dispatching for the Simple NetworkManagement 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).

13.2.3 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
HUAWEI Secoway USG9500	V200R001SPC00

13.2.4 原理描述

SNMP 的工作机制

SNMP 是建立在 TCP/IP 基础上的应用层协议。SNMP 为 NM Station 提供了一套简单的命令集，依据 BER 规则形成报文，使用 UDP 实现网管系统和被管设备间的通信。

SNMP 网络元素分为 NM Station 和 Agent 两种。

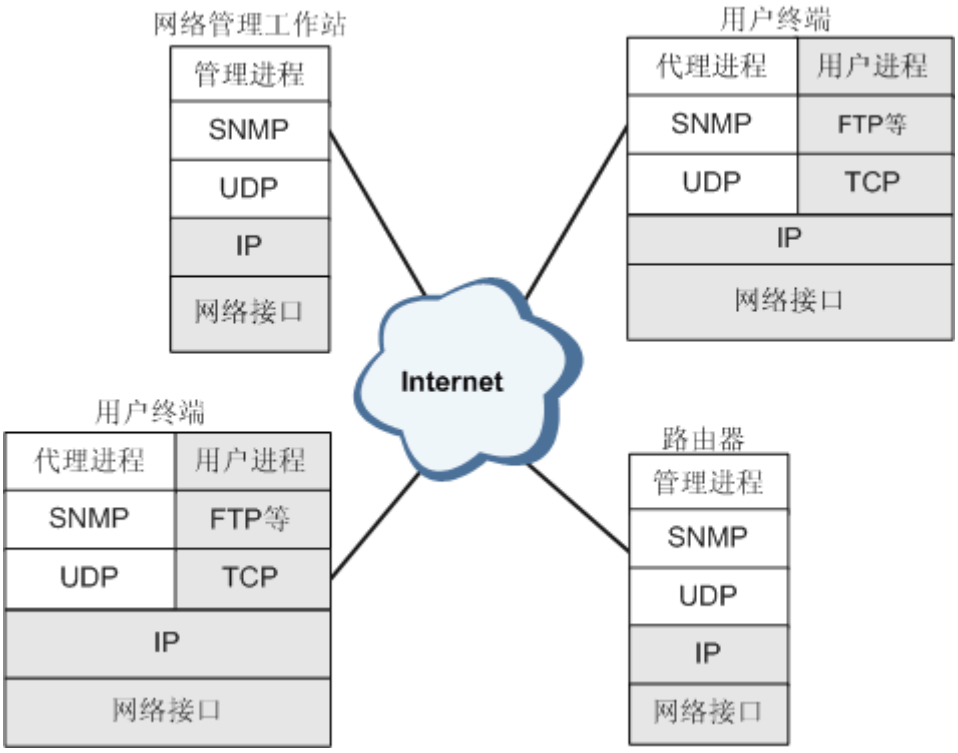
- NM Station 是运行 SNMP 客户端程序的工作站，能够提供非常友好的人机交互界面，方便网络管理员完成绝大多数的网络管理工作。
- Agent 是驻留在设备上的一个进程，负责接收、处理来自 NM Station 的请求报文。在一些紧急情况下，如接口状态发生改变等，Agent 也会主动通知 NM Station。

NM Station 是 SNMP 网络的管理者，Agent 是 SNMP 网络的被管理者。NM Station 和 Agent 之间通过 SNMP 协议来交互管理信息。

如图 13-4 所示是使用 SNMP 的典型配置。整套系统必须有一个网络管理工作站，它是整个网络的网管中心，在它之上运行管理进程。

每个被管对象中一定要有代理进程。管理进程和代理进程利用 SNMP 报文进行通信，SNMP 报文使用 UDP 作为运载协议。

图13-4 SNMP 典型配置



SNMP 有以下几种操作方式：

- 管理工作站通过 Get、Get-Next、Get-Bulk 操作来获取网络资源信息。
- 管理工作站通过 Set 操作来设置网络资源。
- 管理代理主动上报 Trap 报文给网络管理工作站，可使管理工作站及时获取网管代理的工作状态，从而使网络管理员能够及时采取响应措施。

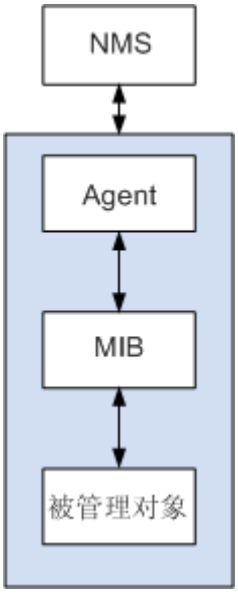
管理模型

SNMP 的管理体系，在 NM Station 和 Agent 两侧进行信令交互。

- 网管端工作站上的 NM Station 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 NM Station 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 NM Station 发送 Trap 消息，告知设备侧的出现的情况，这样便于网络管理人员及时的对网络中出现的情况进行处理。

网络管理模型如图 13-5 所示。

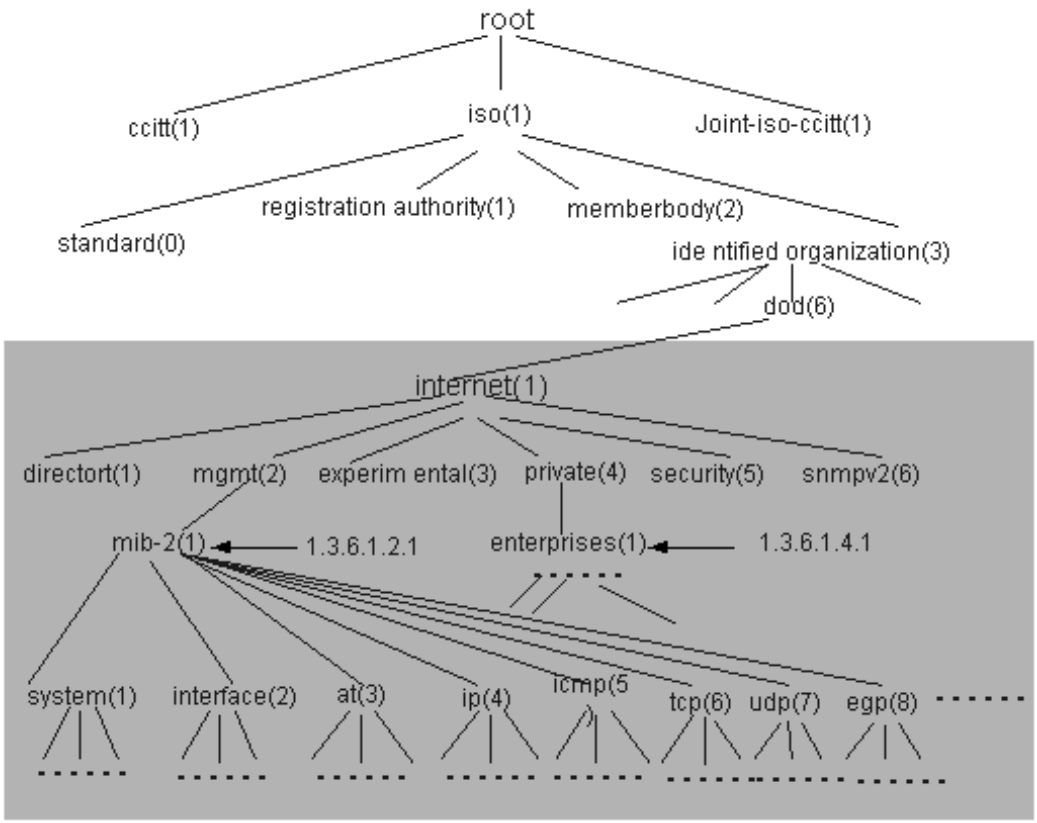
图13-5 SNMP 管理模型



MIB

MIB（Management Information Base，管理信息库）指明了网络元素所维护的变量（即能够被管理进程查询和设置的信息）。MIB 给出了一个数据结构，包含了网络中所有可能的被管理对象的集合。SNMP 的管理信息库采用和域名系统 DNS 相似的树型结构，它的根在最上面，根没有名字。如图 13-6 所示的是管理信息库的一部分，它又称为对象命名（object naming）树。

图13-6 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。在其下面有一个美国国防部 DoD（Department of Defense）的子树（标号是 6），再下面就是 Internet（标号是 1）。在只讨论 Internet 中的对象时，可只画出 Internet 以下的子树（图中带阴影的虚线方框），并在 Internet 结点旁边标注上{1.3.6.1}即可。

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。

最初的结点 MIB 将其所管理的信息分为 8 个类别，见表 13-5。现在的 MIB-2 所包含的信息类别已超过 40 个。

表13-5 MIB 节点的管理信息类别

类别	标号	所包含的信息
system	1	主机或路由设备的操作系统
interfaces	2	各种网络接口及它们的测定通信量
address translation	3	地址转换（例如 ARP 映射）
ip	4	Internet 软件（IP 分组统计）

类别	标号	所包含的信息
icmp	5	ICMP 软件（已收到 ICMP 消息的统计）
Tcp	6	TCP 软件（算法、参数和统计）
udp	7	UDP 软件（UDP 通信量统计）
egp	8	EGP 软件（外部网关协议通信量统计）

MIB 的定义与具体的网络管理协议无关。设备制造商可以在产品（如路由设备）中包含 SNMP 代理软件，并保证在定义新的 MIB 项目后该软件仍遵守标准。用户可以使用同一网络管理客户软件来管理具有不同版本的 MIB 的多个路由设备。当然，若一台路由设备上不支持此 MIB，那么就无法提供相应的功能。

SMI

SMI（Structure of Management Information，管理信息结构）为命名和定义管理对象指定了一套规则，可定义被管对象的对象标识、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
- OCTER STRING
- DisplayString
- OBJECT IDENTIFIER
- NULL
- IpAddress
- PhysAddress
- Counter
- Gauge
- TimeTicks
- SEQUENCE
- SEQUENDEOF

SNMP v1 工作原理

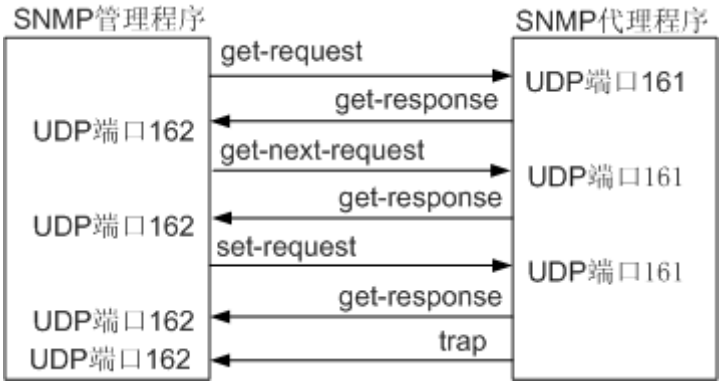
SNMP 规定了 5 种协议数据单元 PDU（也就是 SNMP 报文），用来在管理进程和代理之间的交换。

- get-request 操作：从代理进程中提取一个或多个参数值。
- get-next-request 操作：从代理进程中按照字典序提取下一个参数值。
- set-request 操作：设置代理进程的一个或多个参数值。
- get-response 操作：返回的一个或多个参数值。这个操作是由代理进程发出的，它是前面三种操作的响应操作。
- trap 操作：代理进程主动发出的报文，通知管理进程有某些事情发生。

前面的 3 种操作是由管理进程向代理进程发出的，后面的 2 个操作是代理进程发给管理进程的，为了简化起见，前面 3 个操作今后叫做 get、get-next 和 set 操作。如图 13-7 所示，描述了 SNMP 的这 5 种报文操作。

说明
在代理进程端是用熟知端口 161 来接收 get 或 set 报文，而在管理进程端是用熟知端口 162 来接收 trap 报文。

图13-7 SNMP 的报文操作



SNMPv2c 工作原理

SNMPv2 已被作为 Internet 的推荐标准予以公布。

简单性是 SNMP 标准取得成功的主要原因。因为在大型的、多厂商产品构成的复杂网络中，管理协议的明晰是至关重要的，但同时这又是 SNMP 的缺陷所在——为了使协议简单易行，SNMP 对部分的功能进行了裁减，如：

1. 没有提供成批存取机制，对大块数据进行存取效率很低；
2. 只在 TCP/IP 协议上运行，不支持别的网络协议；
3. 没有提供 manager 与 manager 之间通信的机制，只适合集中式管理，而不利于进行分布式管理；
4. 只适于监测网络设备，不适于监测网络本身。

针对这些问题，对 SNMP 改进工作一直在进行。如 1991 年 11 月，推出了 RMON（Remote Netw ork Monitoring）MIB，加强 SNMP 对网络本身的管理能力。它使得 SNMP 不仅可管理网络设备，还能收集局域网和互联网上的数据流量等信息。1992 年 7 月，针对 SNMP 缺乏安全性的弱点，又公布了 SNMPv2。1996 年，IETF 的工作组发布了一系列 RFC 文档。在这系列新文档中，舍弃了原有 SNMPv2 的安全方面特性，最终形成 SNMPv2c。

SNMPv2c 对 SNMPv1 的主要增强可以分为以下几类：

- SMI（管理信息结构）
- 管理站与管理站之间通信
- 协议控制

SNMPv3 工作原理

RFC 2271 定义的 SNMPv3 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、本地处理模块、用户安全模块以及基于视图的访问控制模块。

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 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 trap（v3）的目的。

SNMP 错误码协议栈对错误码的支持

SNMP 的错误码标识了网元和网管进行通信时，网元对于网管请求的处理情况（例如报文过长、索引不存在等）。由 SNMP 协议规定的错误码称为标准错误码。

SNMP 协议栈提供了 21 种标准错误码：

- SNMPv1 协议专用的 5 种错误码。
- SNMPv2 与 SNMPv3 共有的 16 种错误码。

由于目前系统中支持的特性和场景日益增加，SNMP 提供的标准错误码已经难以满足系统中日益多元化的场景需求。在这种情况下，网管无法正确了解网元在处理上发生的错误场景。因此引入了扩展错误码来解决这一问题。

网元在报文处理出错时返回一个当时场景对应的错误码，若出错场景不在以上标准错误码范围内，可以返回通用场景错误（GEN ERROR）或者一个用户自己定义的错误码，即扩展错误码。扩展错误码可以定义更多的场景。网管（只限于华为网管）能够根据和网元的约定，来正确解析当前网元的错误场景。由于网管和路由器采用了相同的机制来定义错误码；扩展错误码可以通过命令行使能，也可以通过网管执行操作来使能。当扩展错误码被使能后，SNMP 会将特性返回的内部错误码根据一定的规则转换为不同的扩展错误码后，再发送到网管。如果特性发送的是标准错误码，则 SNMP 向网管发送标准错误码；如果扩展错误码功能没有使能，则标准错误码以及模块内部定义的内部错误码都会直接向网管发送。

系统对扩展错误码依据模块号以及模块下注册的扩展错误码进行统一生成和管理。网管侧按照相同的规则解析后，可以把详尽的错误呈现给用户。

SNMP 协议在安全性方面的比较

表13-6 SNMP 协议安全性比较

协议版本	用户校验	加密	鉴权
v1	NO，采用团体字	NO	NO
v2c	NO，采用团体字	NO	NO
v3	YES，基于用户名的加解密	YES	YES

13.3 NTP

13.3.1 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。

13.3.2 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

13.3.3 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
HUAWEI Secoway USG9500	V200R001SPC00

13.3.4 原理描述

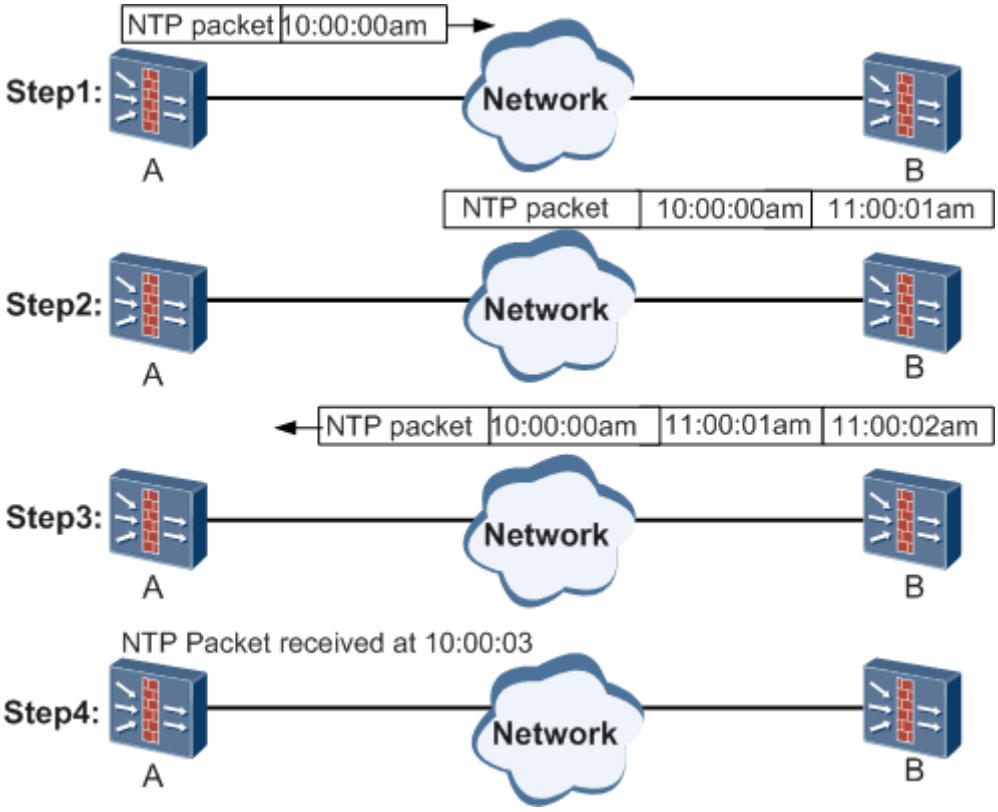
NTP 主要用于在分布式时间服务器和客户端之间进行时间同步，把主机的时间同步到某些时间标准。服务器和客户端的概念是相对而言的，提供时间标准的设备称为时间服务器，接收时间服务的设备称为客户端。对于运行 NTP 的本地系统，既可以接受来自其他时钟源的同步，也可以作为时钟源去同步别的时钟，并且可以通过交换 NTP 报文互相同步。NTP 基于 UDP 传输，使用端口号 123。

实现过程

NTP 的实现过程如图 13-8 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13-8 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：Delay=（T4-T1）-（T3-T2）。
- 设备 A 相对设备 B 的时间差：Offset=（（T2-T1）+（T3-T4））/2。

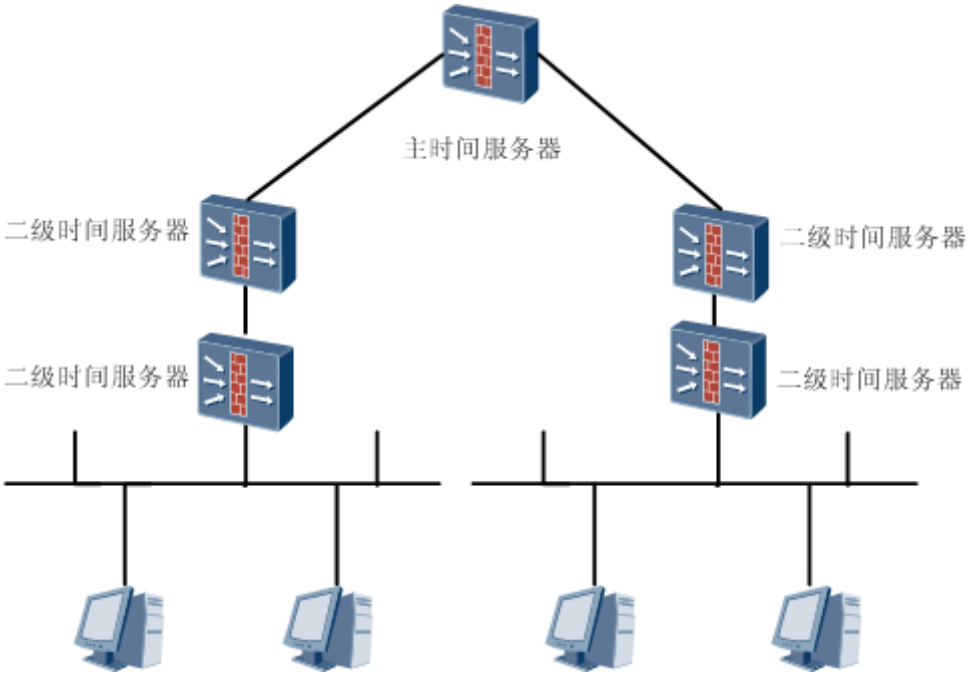
设备 A 根据这些信息来设定自己的时钟，实现与设备 B 的时钟同步。

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 13-9 所示。

图13-9 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，USG9500 提供的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：

- 客户端和服务器端均需要完整配置，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。