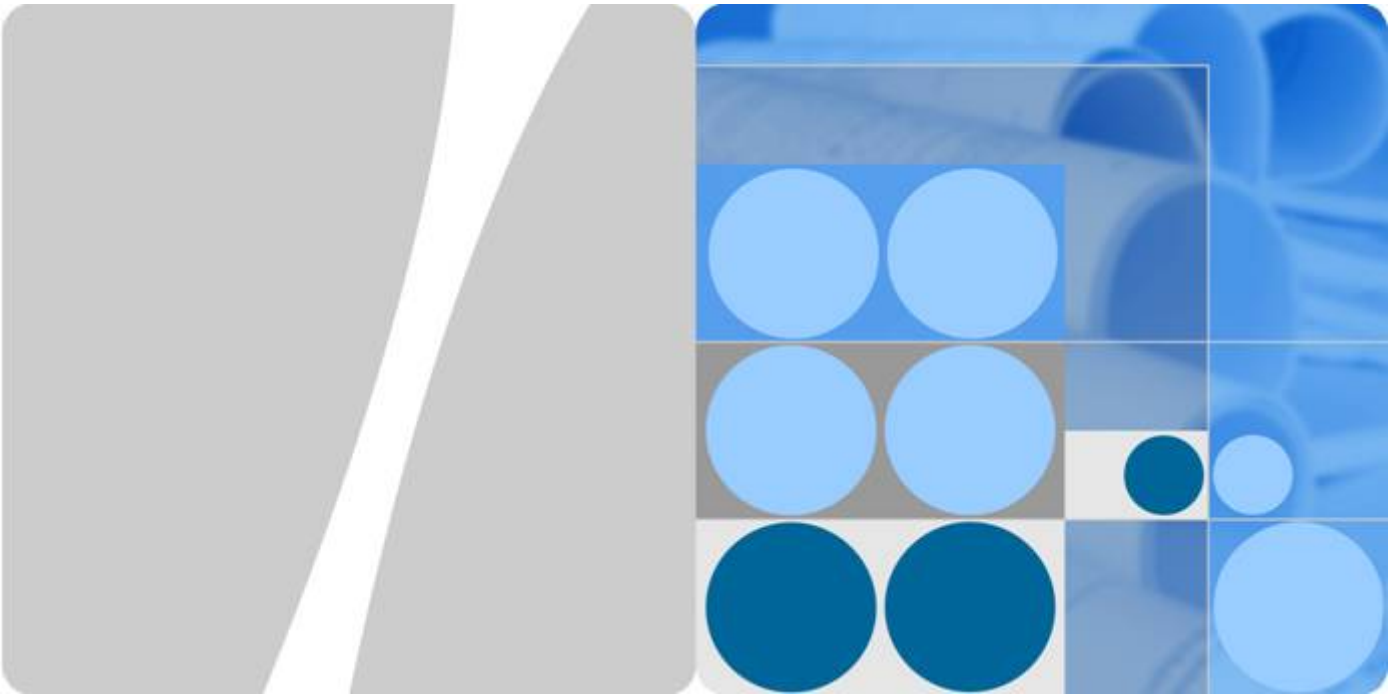


HUAWEI NetEngine20E-X6 高端业务路由器

V600R003C00

特性描述-IP 业务

文档版本　01

发布日期　2011-05-15

华为技术有限公司

商标声明


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

注意

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

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

华为技术有限公司

地址： 深圳市龙岗区坂田华为总部办公楼 邮编：518129

网址： http://www.huawei.com

客户服务邮箱： support@huawei.com

客户服务电话： 0755-28560000 4008302118

客户服务传真： 0755-28560111

前 言

概述

本文档针对 IP 业务特性，从简介、原理描述和应用三个方面介绍了 IP 业务特性。

本文档与其它类型手册相结合，便于读者深入掌握 IP 业务特性的实现原理。

读者对象

本文档主要适用于以下工程师：

- 网络规划工程师
- 调测工程师
- 数据配置工程师
- 系统维护工程师

符号约定

在本文中可能出现下列标志，它们所代表的含义如下。

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

修订记录

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

文档版本 01 (2011-05-15)

第一次正式归档。

目 录

插图目录

表格目录

1 IP 地址

关于本章

1.1 介绍

在 IP 网络上，需要为网络上的主机分配 IP 地址。如果用户要将一台计算机连接到 Internet 上，就需要向 ISP 申请一个 IP 地址。

IP 地址是在计算机网络中被用来唯一标识一台设备的一组数字，各个节点（设备）之间使用 IP 协议进行通信。IP 地址的层次是按网络结构进行划分，一个 IP 地址是由网络号和主机号两部分组成。

IP 地址由 32 位二进制数值组成，但为了便于用户识别和记忆，采用了“点分十进制表示法”。采用了这种表示法的 IP 地址由 4 个由点分隔的十进制整数来表示，每个十进制整数对应一个字节。例如，A 主机的 IP 地址使用二进制的表示形式为 00001010 00000001 00000001 00000010，采用点分十进制表示法表示为 10.1.1.2。

IP 地址由如下两部分组成：

- 网络号码字段（net-id）：用于区分不同网络。网络号码字段的前几位称为类别字段（又称为类别比特），用来区分 IP 地址的类型。
- 主机号码字段（host-id）：用于区分一个网络内的不同主机。

IP 地址的网络号码字段用来标识一个网络，主机号码字段用来标识网络中网络设备的一个连接。如果有多台网络设备，无论它们分别处于任何物理位置，只要它们具有相同的网络号，那他们就处在同一网络中。也就是说，在公共网络内的多台网络设备是否处于相同网络与它们所处的物理位置无关。

1.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC 1166	Internet Numbers	-
RFC 1918	Address Allocation for Private Internets	-

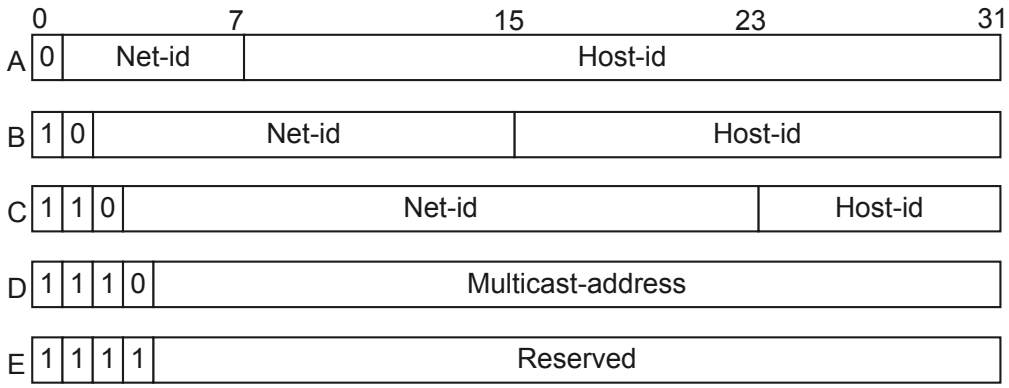
1.3 原理描述

1.3.1 IP 地址分类

为了方便 IP 地址的管理及组网，IP 地址分成五类，如**图 1-1** 所示。

通过网络号码字段的前几个比特就可以判断 IP 地址属于哪一类，这是区分各类地址最简单的方法。

图 1-1 五类 IP 地址



目前大量使用中的 IP 地址属于 A、B、C 三类 IP 地址中的一种。D 类地址是组播地址，E 类地址保留。在 IETF（Internet Engineering Task Force）发布的 RFC1166 Internet Numbers 中详细描述了各类 IP 地址。

在使用 IP 地址时要注意，一些 IP 地址是保留作为特殊用途的，一般的用户不能使用。表 1-1 列出各类 IP 地址的范围。

表 1-1 IP 地址分类及范围

网络类型	地址范围	用户可用的 IP 网络范围	说明
A	0.0.0.0 ～ 127.255.255.255	1.0.0.0 ～ 126.0.0.0	全 0 的主机号码表示该 IP 地址就是网络的地址，用于网络路由；全 1 的主机号码表示广播地址，即对该网络上所有的主机进行广播；IP 地址 0.0.0.0 仅在采用 DHCP 方式的系统启动时允许本主机利用它进行临时的通信，并且永远不是有效目的地址；网络号码为 0 的 IP 地址表示当前网络的主机，可以让机器引用自己的网络而不必知道其网络号；所有形如 127.X.Y.Z 的地址都保留作环回测试，发送到这个地址的分组不会输出到线路上，它们被内部处理并当作输入分组。
B	128.0.0.0 ～ 191.255.255.255	128.1.0.0 ～ 191.254.0.0	全 0 的主机号码表示该 IP 地址就是网络的地址，用于网络路由；全 1 的主机号码表示广播地址，即对该网络上所有的主机进行广播。
C	192.0.0.0 ～ 223.255.255.255	192.0.1.0 ～ 223.255.254.0	全 0 的主机号码表示该 IP 地址就是网络的地址，用于网络路由；全 1 的主机号码表示广播地址，即对该网络上所有的主机进行广播。

网络类型	地址范围	用户可用的IP 网络范围	说明
D	224.0.0.0 ～ 239.255.255.255	无	D 类地址是一种组播地址。
E	240.0.0.0 ～ 255.255.255.255	无	保留。255.255.255.255 用于局域网广播地址。

1.3.2 IP 地址的特点

IP 地址的主要特点有：

- IP 地址是一种非等级的地址结构，不同于电话号码的结构。也就是说，IP 地址不能反映任何有关主机位置的地理信息，只能通过网络号码字段判断出主机属于哪个网络。
- 当一个主机同时连接到两个网络上时（作路由设备用的主机即为这种情况），该主机就必须同时具有两个相应的 IP 地址，其网络号码 net-id 是不同的，这种主机称为多地址主机（Multihomed Host）。主机上的每个接口都对应着一个 IP 地址，因此多接口主机会有多个 IP 地址。
- 按照 Internet 的观点，用转发器或网桥连接起来的若干个局域网仍为一个网络，因此这些局域网都具有同样的网络号码 net-id。
- 在 IP 地址中，所有分配到网络号码 net-id 的网络（不管是小的局域网还是很大的广域网）都是平等的。

1.3.3 特殊 IP 地址

在实际使用过程中，有一些特殊的 IP 地址，其范围和描述如表 1-2 所示。

表 1-2 特殊情况的 IP 地址

IP 地址网络号	IP 地址子网号	IP 地址主机号	能否作为源端地址	能否作为目的端地址	描述
全 0	-	全 0	可以	不可以	用于网络上的主机
全 0	-	主机号	可以	不可以	用于网络上的特定主机
127	-	任何值	可以	可以	用于环回地址
全 1	-	全 1	不可以	可以	用于受限的广播（永远不被转发）
net-id	-	全 1	不可以	可以	用于向以 net-id 为目的的网络广播
net-id	subnet-id	全 1	不可以	可以	用于向以 net-id，subnet-id 为目的的子网广播
net-id	全 1	全 1	不可以	可以	用于向以 net-id 为目的的所有子网广播

说明

net-id，subnet-id 分别表示不全为 0 和不全为 1 的对应字段。

1.3.4 私有 IP 地址

为了解决 IP 地址短缺的问题，提出了私有地址的概念。私有地址是指内部网络或主机地址，这些地址只能用于某个内部网络，不能用于公共网络。RFC1918 描述了为私有网络预留的 3 个 IP 地址段。

IP 地址分配组织规定将下列的 IP 地址保留用作私有地址，如**表 1-3** 所示。

表 1-3 私有 IP 地址

网络类型	地址范围
A	10.0.0.0 ～ 10.255.255.255
B	172.16.0.0 ～ 172.31.255.255
C	192.168.0.0 ～ 192.168.255.255

1.4 应用

1.4.1 子网划分

IP 地址的网络部分称为网络地址，网络地址用于唯一的标识一个网段。通过将网络地址进一步划分为若干个子网，实现不同子网之间隔离广播报文。

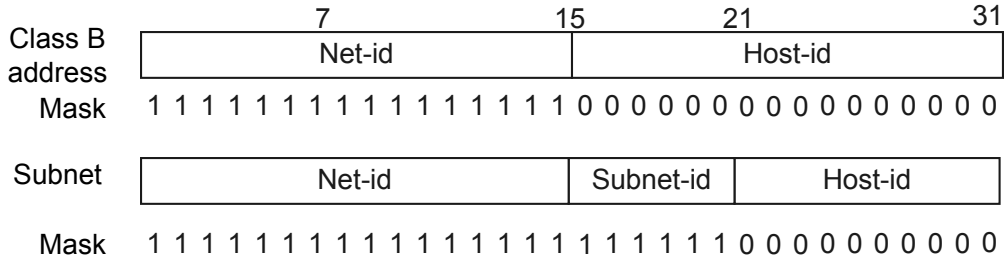
从地址分配的角度来看，子网是网段地址的扩充。为了使 IP 地址的使用更加灵活，只分配 IP 地址的网络号码 net-id，而后面的主机号码 host-id 则是受本单位控制。即某个单位申请到 IP 地址时，实际上只是拥有了一个网络号码 net-id，具体的各个主机号码 host-id 则由该单位自行分配，只要做到在该单位管辖的范围内无重复的主机号码即可。

当一个单位的主机很多而且分布在很大的地理范围时，为了便于管理，可将单位内部的主机号码再进一步划分为多个子网。通过子网划分，整个网络地址可以划分成更多的小网络。

子网的划分是网络内部的行为，从外部看，这个单位只有一个网络号码。只有当外部的报文进入到本单位范围后，本单位的路由设备才根据子网号码再进行选路，找到目的主机。

如图 1-2 所示，为一个 B 类 IP 地址子网划分情况，其中子网掩码由一串连续的“1”和一串连续的“0”组成。“1”对应于网络号码和子网号码字段，而“0”对应于主机号码字段。

图 1-2 IP 地址子网划分



将 32 位的 IP 地址和子网掩码的对应位作与运算可以确定 IP 地址的网络号。例如，IP 地址为 10.1.1.2，子网掩码为 255.255.0.0，那么将 IP 地址与其相应掩码位执行与运算的结果就是网络地址 10.1.0.0。

多划分出一个子网号码字段是要付出代价的。举例来说，本来一个 B 类 IP 地址可以容纳 65534 个主机号码。但划分出 6bits 长的子网字段后，最多可有 64 个子网，每个子网有 10bit 的主机号码，即每个子网最多可有 1022（2^{10}-2，去掉全 1 和全 0 的主机号码）个主机号码。因此主机号码的总数是（64 x 1022 = 65408）个，比不划分子网时要少 126 个。

若一个单位不进行子网的划分，则其子网掩码即为默认值，此时子网掩码中“1”的长度就是网络号码的长度。因此，对于 A、B 和 C 类的 IP 地址，其对应子网掩码的默认值分别为 255.0.0.0、255.255.0.0.和 255.255.255.0。

子网划分与 IP 地址规划时，通常需要综合考虑以下原则，实现合理高效的网络规划。

层次性

实现网络的层次性划分，需要综合考虑地域和业务因素，尽可能和网络层次相对应，采用自顶向下的方法划分。达到有效管理网络、简化路由表的目标。一般情况下：

- 对于大骨干网络和大城域网络相结合的网络，采用扁平化思路划分方式。
- 对于行政区类型的网络，采用多级网络分配方式。

连续性

连续地址在层次结构的网络中易于进行路由聚合，大大缩减路由表数量，提高路由查找的效率。

- 尽量为每个区域分配连续的 IP 地址空间。
- 尽量为具有相同业务和功能的设备分配连续的 IP 地址。
- 即使使用了支持地址重叠的 MPLS/VPN 技术，也尽量不要规划为相同的地址。

扩展性

分配地址时，在每一层次上都要留有余量。当网络规模扩展时能保证地址分配的连续性，实现网络的长远规划。

骨干网络应有足够的连续地址组成独立的自治域，并为今后的扩展留有余地。

高效性

划分子网时，要保证充分利用地址资源，使子网的划分满足主机个数的要求。

- 利用可变长子网掩码（VLSM）技术，分配 IP 地址，充分合理地利用地址资源。
- 与网络的路由机制设计相结合，合理使用已划分的地址空间，提高地址的利用率。

业务相关性

规划 IP 地址时，应该为具有类似功能的设备分配相同类型的 IP 地址。

- 对于高端路由器、IP 电话网关、IP 电话网守、各种 Internet 服务器、防火墙、边缘或接入路由器等设备，应分配公网 IP 地址。VPN 与采用 VPN 方式进行的服务可以在 VPN 内部分配私网地址。
- 对于作为设备管理地址的 Loopback 接口，应尽量为其分配单独的一段连续地址，掩码使用 32 位。
- 对于设备间的互联接口，应尽量为其分配单独的一段连续地址，掩码使用 30 位。

1.4.2 IP 地址分配

用户访问 Internet 必须要有合法的 IP 地址，因此，用户地址的统一分配和管理是宽带接入服务器必须具备的基本功能。目前有以下几种主要的 IP 地址分配方式。

手工分配 IP 地址

可以直接在用户计算机上手工配置 IP 地址，这种方式一般用于固定用途的服务器或有特殊需要的用户，例如 Web 服务器、路由器等。为防止这类 IP 地址被盗用，可以在宽带接入服务器上配置 IP/VLAN、IP/PVC 绑定。

使用 DHCP 服务分配 IP 地址

DHCP 采用客户/服务器通信模式。网络管理员在 DHCP 服务器上设定一个 IP 地址范围，客户端向服务器提出配置申请（包括分配的 IP 地址、子网掩码、缺省网关等参数），服务器根据策略返回相应的配置信息。

1.4.3 IP 地址借用

一个接口如果没有 IP 地址就无法生成路由，也就无法转发报文。IP 地址借用（IP Address Unnumbered）就是在本接口没有 IP 地址的情况下，可以使用其它接口的 IP 地址。所谓“借用 IP 地址”，其实质就是：一个接口上没有配置 IP 地址，但是还想使用该接口。就向其它有 IP 地址的接口借一个 IP 地址过来，以使该接口能够正常使用。

IP 地址借用的主要目的是节省 IP 地址资源。有时某个接口只是偶尔使用，这种情况也可配置该接口借用其他接口的 IP 地址，而不必让其一直占用一个单独的 IP 地址。

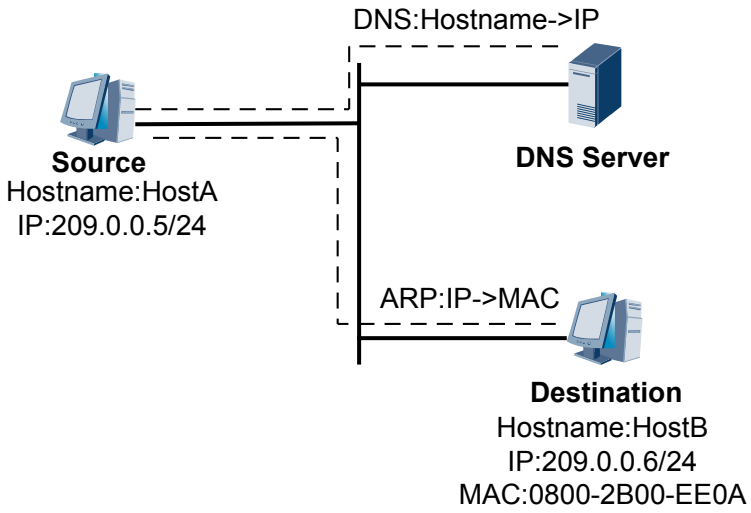
1.4.4 IP 地址解析

一台路由设备用来连接多个网络，具有多个网络的 IP 地址。上面讲的 IP 地址还不能直接用来进行通信。具体原因如下：

- IP 地址只是主机在网络层中的地址，若要将网络层中传送的数据报交给目的主机，必须知道该主机的物理地址。因此必须将 IP 地址解析为物理地址。
- 用户平时不愿意使用难于记忆的 IP 地址，而更愿意使用易于记忆的主机名，因此也需要将主机名解析为 IP 地址。

图 1-3 表示了主机名、IP 地址和物理地址之间的关系。在 Ethernet 上，主机的物理地址就是指 MAC 地址。将主机名解析为 IP 地址的操作是由 DNS 服务器来完成，而将 IP 地址解析为 MAC 地址的操作是由 ARP 来完成的。

图 1-3 主机名、IP 地址和物理地址之间的关系



1.4.5 广域网接口 IP 地址与链路层协议地址的映射

在路由设备中，除了维护以太网口 IP 地址到 MAC 地址的映射外，还需维护广域网口的 IP 地址与链路层协议地址的映射，这类映射有：

- 在封装帧中继接口上，IP 地址与 DLCI（Data Link Control Identifier）的映射。

1.4.6 VPN-Instance 中的 IP 地址空间重叠

VPN-Instance

VPN-Instance 的概念最初是在 BGP/MPLS VPN 中引入的，其主要作用是隔离 VPN 路由与公网路由，以及隔离不同 VPN 的路由。

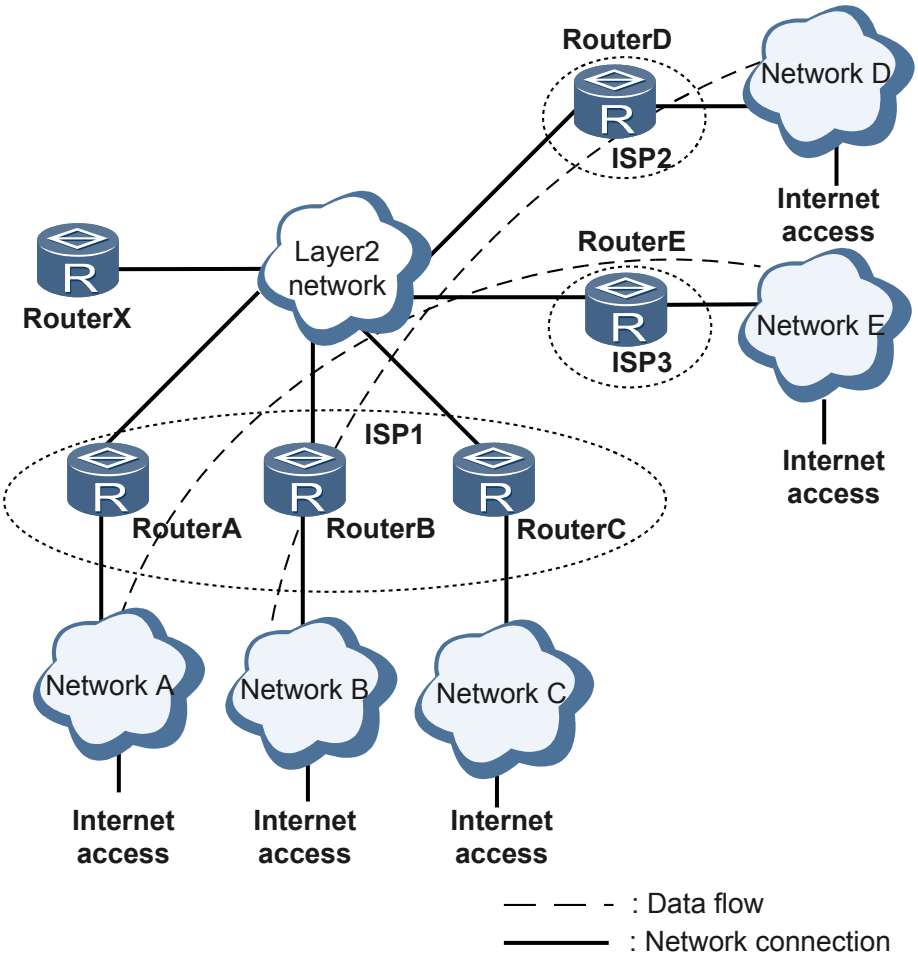
VPN-Instance 的这一功能使它在非 BGP/MPLS VPN 的环境中也可以有广泛的应用，使用 VPN-Instance，可以在一台路由设备上虚拟出多台相互独立的路由设备，在 IP 网络中实现路由隔离。

在 NE20E-X6 中，多种软件特性都实现对 VPN-Instance 的支持，通过绑定到不同的 VPN-Instance 提供“多实例”功能，例如：各种路由协议的多实例（RIP 多实例、OSPF 多实例、IS-IS 多实例、BGP 多实例）。

VPN-Instance 用于本地流量转发

图 1-4 是一种常见的本地流量转发组网方案：ISP1、ISP2、ISP3 接入到一个提供高速交换能力的二层网络；五个本地网络 Network A、B、C、D、E 分别连接到 ISP1、ISP2 和 ISP3；五个本地网络各自有接入 Internet 的专用链路，本地网络之间的流量互访通过二层网络进行。

图 1-4 本地流量转发组网方案



为了提高网络的安全性和降低对 ISP 路由设备的容量要求，二层网络连接到一台大容量路由设备 RouterX。

RouterX 与所有 ISP 路由设备分别建立 EBGP 邻居关系，不同 ISP 路由设备之间不建立任何 BGP 邻居关系。这样，RouterX 可以学到所有网络的路由，并将路由下发给各 ISP 路由设备，指导对本地流量的转发。

在图 1-4 中，ISP1 为三个本地网络提供到二层网络的连接。由于通常情况下，同一路由设备的不同接口不能配置相同网段的 IP 地址，而 ISP1 到二层网络可能只有一个网段的地址，这样，ISP1 必须配备三台路由设备分别接入，设备的投资比较高。

VPN-Instance 可以实现在一台路由设备的不同接口上配置同一网段的 IP 地址，从而节省 ISP1 的建设成本。

1.5 术语与缩略语

术语

术语	解释
点分十进制表示法	点分十进制表示法是一种书写格式。采用了点分十进制的 IP 地址，即 IP 地址被“.”分隔成四部分，每部分都由十进制数字来表示。
IP 地址借用	在本接口没有 IP 地址的情况下，使用其它接口的 IP 地址。
私有 IP 地址	指内部网络或主机地址，这些地址只能用于某个内部网络，不能用于公共网络。
子网掩码	子网掩码是 32 比特的二进制数字，使用子网掩码可以了解 IP 地址的网络号。

2 ARP

关于本章

2.1 介绍

定义

ARP（Address Resolution Protocol）是用来将 IP 地址解析为 MAC 地址的协议。ARP 表项可以分为动态和静态两种类型。另外 ARP 还有扩展应用功能，包括 Proxy ARP 功能、免费 ARP、ARP 安全、ARP 与接口状态联动以及 ARP-Ping。

目的

局域网中每台主机或路由器都有一个 32 位的 IP 地址，这个地址用于该主机的所有通信。IP 地址的分配是独立于机器的硬件地址的。而在以太网中，主机或路由器是根据 48 位的 MAC 地址来发送、接收以太网数据帧的，这个 MAC 地址又称为物理地址或硬件地址，是制造设备时分配到以太网接口中的。因而，在实际的网络互联中，需要一种地址解析的机制来为这两种不同的地址形式提供映射。

ARP 协议主要是解决以上问题，此外 ARP 特性中还包括如下的应用特性：

- 动态 ARP：利用 ARP 报文，ARP 动态执行并自动进行 IP 地址到以太网 MAC 地址的解析，无需网络管理员手工处理。
- 静态 ARP：建立 IP 地址和 MAC 地址之间固定的映射关系，在主机和路由器上不能动态调整此映射关系。需要网络管理员手工添加。
- Proxy ARP 功能：当主机上没有配置缺省网关地址（即不知道如何到达本网络的中介系统），它可以发送一个 ARP 请求（请求目的主机的 MAC 地址）。使能 Proxy ARP 功能的路由器收到这样的请求后，会使用自己的 MAC 地址作为该 ARP 请求的回应，使得处于不同物理网络但网络号相同的内部主机之间可以正常的相互通信。
- 免费 ARP：用于检查重复的 IP 地址和通告新的 MAC 地址。
- ARP 安全：过滤不信任的 ARP 报文以及对 ARP 报文进行时间戳抑制来保证网络设备的安全性和稳定性。
- ARP 与接口状态联动：ARP 与接口状态联动功能可以使路由器向对端发送 ARP 探测报文，然后根据是否收到回应来判断对端是否具有正常的报文转发功能。本端的协议状态会相应变为 Up 或 Down，进而可以触发路由的快速收敛。

2.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC826	Ethernet Address Resolution Protocol	
RFC903	Reverse Address Resolution Protocol	
RFC1027	Using ARP to Implement Transparent Subnet Gateways	
RFC1042	Standard for the Transmission of IP Datagrams over IEEE 802 Networks	

2.3 原理描述

ARP 是用来实现以太网中三层 IP 地址与二层 MAC 地址之间的映射，是以太网通信的基础。

2.3.1 ARP 原理

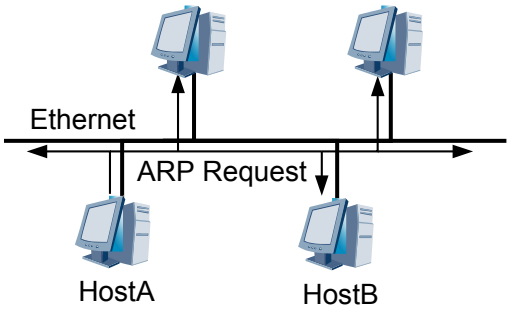
以太网的同一网段内以广播的方式查询某个 IP 地址对应的 MAC 地址，以实现三层 IP 地址与二层 MAC 地址之间的动态映射，这是任何以太网主机设备都支持的一个协议。我们有的时候称 ARP 为 2.5 层协议。

ARP 地址解析过程

TCP/IP 协议的设计人员根据以太网这种具有广播特性的网络开发出的 ARP 地址解析协议。主机在仅知道同一物理网络上的目的端的 IP 地址情况下，通过 ARP 解析到目的端的 MAC 地址。即使网络上的主机发生变化，比如主机的增加或减少、主机更换计算机的网卡等，仍可以完成从 IP 地址到 MAC 地址的转换，并且这个转换关系可以动态更新。

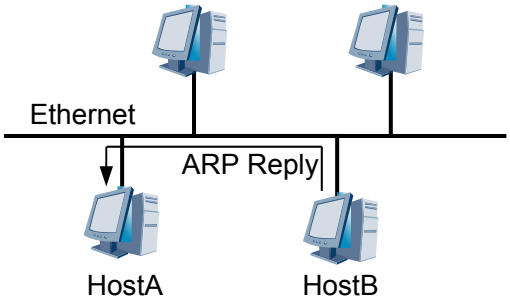
ARP 通过以下两个步骤完成地址解析过程。

图 2-1 ARP 请求过程



如图 2-1 所示，当主机 A 仅知道主机 B 的 IP 地址而不知道其 MAC 地址时，它广播一个 ARP 请求报文，请求得到主机 B 的 MAC 地址。

图 2-2 ARP 响应过程



因为主机 A 发送的是一个广播报文，所以包括主机 B 在内的所有主机都会接收到这个请求。由于 ARP 请求报文的目的 IP 地址是主机 B 的 IP 地址，所以只有主机 B 会响应这个 ARP 请求。如图 2-2 所示，主机 B 向主机 A 发出一个包含其 MAC 地址的 ARP 响应报文。

当主机 A 接收到主机 B 的响应报文后，就用这个 MAC 地址和主机 B 通信。

ARP 老化机制

- 高速缓存

  如果每次主机 A 向主机 B 发送一个分组前都要发送一个广播的 ARP 请求报文的话，会增加很多网络的通信量。而且网络上的所有机器都必须接受和处理这个广播的 ARP 请求报文，这也极大的影响了网络运行效率。

  为了解决以上问题，每台主机上都维护着一个高速缓存，这是 ARP 高效运行的一个关键。在这个高速缓存中，存放最近获得的 IP 地址到 MAC 地址的映射关系。

  发送方在每次发送分组时，都先在缓存中查找目标 IP 地址所对应的 MAC 地址。如果 ARP 缓存中有对应的 MAC 地址，主机就不会再发送 ARP 请求报文，而是直接将分组发至这个 MAC 地址。如果 ARP 缓存中没有对应的 MAC 地址时，主机才会发送广播的 ARP 请求报文。

- 动态 ARP 表项的老化超时时间

  如图 2-2 所示，当主机 B 回应了主机 A 的 ARP 请求后，在主机 A 的缓存中会形成主机 B 的 IP 地址和其 MAC 地址的映射关系。但是，如果主机 B 发生故障后或者更换了网卡时，主机 A 没有得到任何关于主机 B 的任何通告，于是主机 A 仍会继续将分组发送给主机 B。造成地址解析出现错误的原因就是主机 A 中的缓存表的信息没有得到及时的更新。

  为了减少地址解析过程中所出现的错误，ARP 高速缓存中的表项一般都会设定一个定时器。当达到定时器的动态 ARP 表项的老化超时时间后，删除掉这个表项。

  通过设置定时器，在地址解析过程中出现错误的现象得到了改善但并没有完全消除，其原因在于时延。如果定时器的动态 ARP 表项的老化超时时间是 N 秒，发送方只有等到 N 秒后才能检测到接收方出现了故障，在此期间发送方缓存表的信息还是没有得到及时的更新。

- 动态 ARP 表项的老化探测次数

除了设置定时器中动态 ARP 表项的老化超时时间，还可以通过设置动态的探测次数来减少地址的解析错误。在一个动态 ARP 表项老化之前，系统先进行探测，如果超过设置的探测次数后仍没有应答，则此 ARP 表项将被删除。

- 动态 ARP 表项的老化探测模式

  ARP 表项老化之前，接口会发送 ARP 老化探测报文。老化探测报文可以是单播报文，也可以是广播报文。缺省情况下，接口以广播模式发送 ARP 老化探测报文。

  当对端设备的 IP 地址不变化而 MAC 地址频繁更新时，建议使用广播模式发送 ARP 老化探测报文。

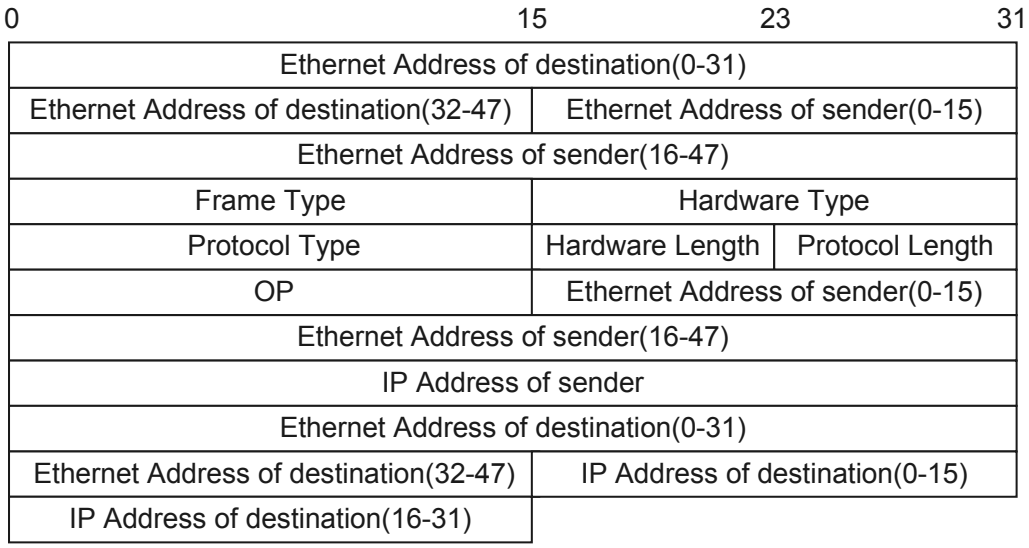
  当对端设备 MAC 地址不变，当前网络带宽资源特别紧缺，且 ARP 表项的老化时间设置的比较小时，建议使用单播模式发送 ARP 老化探测报文。

  当其他厂商设备与华为设备互联时，其他厂商设备接收到目的 MAC 地址为广播地址的 ARP 老化探测报文后，若 ARP 表项中已存在华为设备的 IP 地址与 MAC 地址映射，则不响应该广播 ARP 老化探测报文。这种特殊情况下华为设备需要配置成以单播方式发送 ARP 老化探测报文，即单播的 ARP 老化探测模式。

2.3.2 ARP 报文格式

ARP 请求和应答的报文格式如图 2-3 所示。

图 2-3 ARP 请求和应答报文格式



ARP 报文的长度是 42 字节。最前面 14 字节的字段表示以太网首部，后面的 28 字节内容是 ARP 请求或响应的报文格式。

主要字段的解释如表 2-1 所示：

表 2-1 ARP 报文各字段的含义

字段	长度	含义
Ethernet Address of destination	48 比特	目的以太网地址。发送 ARP 请求时，为广播的 MAC 地址，0xFF.FF.FF.FF.FF.FF。

字段	长度	含义
Ethernet Address of sender	48 比特	以太网源地址。
Frame Type	16 比特	表示后面数据的类型。对于 ARP 请求或应答来说，该字段的值为 0x0806。
Hardware Type	16 比特	表示硬件地址的类型。对于以太网，该类型的值为“1”。
Protocol Type	16 比特	表示发送方要映射的协议地址类型。对于 IP 地址，该值为 0x0800。
Hardware Length	8 比特	表示硬件地址的长度，单位是字节。对于 ARP 请求或应答来说，该值为 6。
Protocol Length	8 比特	表示协议地址的长度，单位是字节。对于 ARP 请求或应答来说，该值为 4。
OP	16 比特	操作类型。OP 的值与操作类型的关系如**表 2-2** 所示。
Ethernet Address of sender	48 比特	发送方以太网地址。这个字段和 ARP 报文首部的源以太网地址字段是重复信息。
IP Address of sender	32 比特	发送方的 IP 地址。
Ethernet Address of destination	48 比特	接收方的以太网地址。发送 ARP 请求时，该处填充值为 0x00.00.00.00.00.00。
IP Address of destination	32 比特	接收方的 IP 地址。

表 2-2 OP 值与操作类型的对应关系

操作	操作类型
1	ARP 请求
2	ARP 应答
3	RARP 请求
4	RARP 应答

说明

对于一个 ARP 请求来说，除目的端硬件地址外的所有其他的字段都有填充值。

2.3.3 动态 ARP

ARP 表项的创建与更新

依据 ARP 协议描述，几乎所有的以太网通信都以 ARP 开始，所以任何以太网主机设备都支持这个协议，而且 IP 地址到以太网 MAC 地址的解析主要也是动态生成，无须网络管理员手工处理。

一般实现中，如果收到的 ARP 报文满足以下条件中的任何一条，系统将创建或更新 ARP 表项：

- ARP 报文的源 IP 地址与入接口 IP 地址在同一网段，且不是广播地址，目的 IP 地址是本接口 IP 地址。
- ARP 报文的源 IP 地址与入接口 IP 地址在同一网段，且不是广播地址，目的 IP 地址是本接口的 VRRP（Virtual Router Redundancy Protocol）虚拟 IP 地址。

如果收到的 ARP 报文的源 IP 地址在入接口的 ARP 表中已经存在对应表项，也将对 ARP 表项进行更新。

ARP 抑制功能

在特殊组网或者遭受到 ARP 攻击时，系统在同一时间内会接收到多个源 IP 地址相同的 ARP 报文，这就需要系统对 ARP 表项进行重复更新。为了维护系统性能，系统可以启动 ARP 抑制功能，即在 1s 内收到多次源 IP 地址相同的 ARP 报文，系统将只通知发送 ARP 报文的设备已收到 ARP 报文，而不更新设备的 ARP 表。

如果对所有接口都做 ARP 抑制会造成某些接口的 ARP 表项暂时无法正常更新。ARP 抑制只针对 Eth-Trunk 接口，

说明

Eth-Trunk 在跨接口板绑定成员接口后，在该逻辑接口上生成 ARP 表后会同步到其成员接口的接口板上。如果 ARP 报文数量过大，会使 CPU 处理繁忙。故对跨接口板的逻辑接口进行了 ARP 相同报文的抑制功能，而普通的物理接口不存在该问题，所以没有此抑制功能。

2.3.4 静态 ARP

静态 ARP 是指 IP 地址和 MAC 地址之间有固定的映射关系，在主机和设备上不能动态调整此映射关系。

静态 ARP 主要用来解决如下问题：

- 为了将目的 IP 地址不在本网段的报文，穿过本网段的某个网关，使得到该 IP 地址的报文能通过该网关进行转发。
- 当用户需要过滤掉一些非法的报文时，可以将这些非法报文的目的 IP 地址绑定到某个不存在的 MAC 地址。

静态 ARP 由网络管理员手动配置生成。

2.3.5 Proxy ARP

Proxy ARP 主要是通过代理的方式来解决网络互通问题的 ARP 实现功能。

Proxy ARP 有以下特点：

- 所有处理在 ARP 子网网关（ARP Subnet Gateways）进行，所连网络中的主机不必做任何改动；
- 在主机端看不到子网，只是一个标准 IP 网络；
- Proxy ARP 只影响主机的 ARP 高速缓存，对网关的 ARP 高速缓存和路由表没有影响；
- 使用 Proxy ARP 后，主机应该减小 ARP 老化时间，以尽快使无效 ARP 项失效，减少发给路由器而路由器却不能转发的报文。

下表为三种 Proxy ARP：

Proxy ARP 方式	解决的问题
路由式 Proxy ARP	解决同一网段不同物理网络上计算机的互通问题。
VLAN 内 Proxy ARP	解决相同 VLAN 内，且 VLAN 配置用户隔离后的网络上计算机互通问题。
VLAN 间 Proxy ARP	解决不同 VLAN 之间对应计算机的二层互通问题。

路由式 Proxy ARP

路由式 Proxy ARP 就是使那些在同一网段却不在同一物理网络上的计算机或路由器能够相互通信的一种功能。

在实际应用中，如果连接路由器的当前主机上没有配置缺省网关地址（即不知道如何到达本网络的中介系统），此时将无法进行数据转发。

路由式 Proxy ARP 可以解决这个问题，主机发送一个 ARP 请求（请求目的主机的 MAC 地址），使能 Proxy ARP 功能的路由器收到这样的请求后，会使用自己的 MAC 地址作为该 ARP 请求的回应，以此进行数据转发。

使能 Proxy ARP 功能的路由器还可隐藏物理网络的细节，使得处于不同物理网络但网络号相同的两个 Ethernet 的内部主机之间可以正常的相互通信。

VLAN 内 Proxy ARP

如果两个用户属于相同的 VLAN，但 VLAN 内配置了用户隔离。此时用户间要互通，需要在关联了 VLAN 的接口上启动 VLAN 内 Proxy ARP 功能。

若路由器的接口使能了 VLAN 内 Proxy ARP 功能，接口在接收到目的地址不是自己的 ARP 请求报文后，路由器并不立即丢弃该报文，而是查找该接口的 ARP 表项。如果满足代理条件，则将路由器的 MAC 地址发送给 ARP 请求方。

VLAN 内 Proxy ARP 主要用于配置了用户隔离的 VLAN 内的用户间互通。

VLAN 间 Proxy ARP

如果两个用户属于不同的 VLAN，用户间要进行二层互通，需要在关联了 VLAN 的接口上启动 VLAN 间 Proxy ARP 功能。

若路由器的接口使能了 VLAN 间 Proxy ARP 功能，接口在接收到目的地址不是自己的 ARP 请求报文后，路由器并不立即丢弃该报文，而是查找该接口的 ARP 表项。如果满足代理条件，则将路由器的 MAC 地址发送给 ARP 请求方。

VLAN 间 Proxy ARP 主要用于：

- 处于不同 VLAN 的用户进行二层通信。

2.3.6 免费 ARP

主机主动使用自己的 IP 地址作为目标地址发送 ARP 请求，此种方式称免费 ARP。免费 ARP 有三方面的作用：

- 用于检查重复的 IP 地址：正常情况下不会收到 ARP 回应，如果收到，则表明本网络中存在与自身 IP 地址重复的地址。
- 用于通告一个新的 MAC 地址：发送方更换了网卡，MAC 地址变了，为了能够在 ARP 表项老化前通告所有主机，发送方可以发送一个免费 ARP。
- 在 VRRP 备份组中用来通告主备发生变换。

2.3.7 ARP 安全

ARP 安全配置是一种基于 ARP 的安全特性，通过过滤不信任的 ARP 报文以及对某些 ARP 报文进行时间戳抑制来保证网络设备的安全性和健壮性。

ARP 安全配置不仅能够防范针对 ARP 协议的攻击方式，还能够防范其它基于 ARP 协议的攻击方式，ARP 安全主要应用如下表所示：

攻击手段	防攻击功能
通过发送大量伪造的 ARP 请求报文，路由器不断学习，最终造成路由器的 ARP 缓存溢出，从而无法缓存正常的 ARP 表项，进而阻碍正常转发。	严格学习 ARP 表项。
利用路由器 ARP 缓存的有限性，通过发送大量伪造的 ARP 请求、应答报文，路由器不断学习，最终造成路由器的 ARP 缓存溢出，从而无法缓存正常的 ARP 表项，进而阻碍正常转发。	基于接口的 ARP 表项限制。
利用路由器计算能力的有限性，通过发送大量伪造的 ARP 请求、应答报文或其他能够触发路由器 ARP 处理的报文，造成路由器的计算资源长期忙于 ARP 处理，影响其他业务的处理，进而阻碍正常转发。	对 ARP 报文进行时间戳抑制。

严格学习 ARP 表项

严格学习 ARP 表项是指路由器仅学习自己发送的 ARP 请求报文的应答报文，并不学习其它设备向路由器发送的 ARP 请求报文。通过严格学习 ARP 表项可以拒绝掉大部分的 ARP 请求和应答报文攻击。

基于接口的 ARP 表项限制

限制每个接口 ARP 表项学习总数目，可以有效的防止 ARP 缓存溢出，保证 ARP 表项的安全性。

对 ARP 报文进行时间戳抑制

时间戳抑制就是指路由器会对 ARP 报文进行数量统计，如果在一定时间内，ARP 报文数量超出了配置的阈值，超出部分的 ARP 报文将被忽略，路由器不作任何处理，目前只支持基于目的地址的 ARP报文时间戳抑制。

2.3.8 ARP 与接口状态联动

ARP 与接口状态联动是指通过接口发送 ARP 探测报文和应答流程来决定接口的协议状态的功能。

此功能主要是解决设备与设备间互通时，如果一方设备不支持 BFD，无法进行 BFD 的链路检测，从而无法快速指导路由的切换的问题。

ARP 与接口状态联动功能可以使路由器向对端发送 ARP 探测报文，然后根据是否收到回应来判断对端是否具有正常的报文转发功能，从而决定本端的协议状态为 Up 或 Down，进而可以触发路由的快速收敛。

只能对 GE 及其子接口配置 ARP 与接口状态联动功能，主要应用于不支持 BFD 与接口联动功能的设备。

2.3.9 ARP-Ping

ARP-Ping：包括 ARP-Ping IP 和 ARP-Ping MAC，用于部署二层特性时方便维护。

ARP-Ping IP

ARP-Ping IP 的标准是 ARP 协议。通过配置管理平面获得用户输入的 IP 地址和出接口（出接口是可选项），然后构造 ARP Request 报文，在出接口广播该报文。在指定超时时间内，若收不到回复报文，则向用户显示该 IP 地址无人使用；若收到回复报文，则将回复报文中的对端 MAC 地址取出，显示给用户。

ARP-Ping IP 是利用 ARP 报文在局域网内探测 IP 地址是否被其它的设备使用的一种方法。

用户对设备配置 IP 地址前，需要确认该 IP 地址有没有被网络上的其他设备使用，可以通过发送 ARP 报文（二层），确认该 IP 的使用情况，以便做出相应调整。

通过 **ping** 命令也可以探测该 IP 地址是否被网络上的其他设备使用。但是如果带有防火墙功能的目的主机和路由设备设置了对 Ping 报文不进行回复的功能时，就不会响应 Ping 报文，造成该 IP 没有被使用的假象。由于 ARP 报文是二层协议，大多数情况下可以透过设置了对 **ping** 报文不进行回复的防火墙，从而避免了此类情况的发生。

ARP-Ping IP 原理

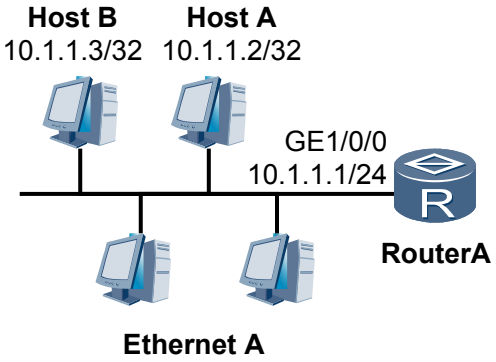
ARP-Ping IP 发送的是 ARP 请求报文。以下是 ARP-Ping IP 的具体实现过程。

1. 用户通过命令行设置指定的 IP 地址后，发送 ARP 请求报文并且启动 ARP Reply 报文的超时定时器。
2. 局域网内路由设备或主机收到 ARP 请求报文后，回复 ARP Reply 报文。
3. 源路由设备收到 ARP Reply 报文后将 Reply 报文中的源 IP 地址和命令行中输入的 IP 地址进行比较。若匹配，则向用户显示与所输入 IP 地址相对应的 MAC 地址并且关闭 Reply 报文的超时定时器，本次操作结束。

   若 ARP Reply 报文的超时定时器超时，输出该 IP 地址无设备使用的显示信息。

如图 2-4 所示，RouterA 可通过 ARP-Ping IP 来获知 10.1.1.2 这个 IP 地址是否被使用。RouterA 收到网络内 IP 地址为 10.1.1.2 的主机 A 的 ARP Reply 报文后，将这个主机的 MAC 地址显示出来。通过显示信息可得知这个 IP 地址被网络内的主机使用。

图 2-4 ARP-Ping IP 的实现过程



ARP-Ping MAC

ARP-Ping MAC 和普通 Ping 处理一样，但 ARP-Ping MAC 只应用在直连以太局域网。发送 ICMP 回显请求报文，接收 ICMP Reply 报文，解析报文，把保存在报文数据区的源 MAC 地址和本机保存的 MAC 地址相比较。如果相同则显示该报文的 IP 地址，并提示该 MAC 地址已被使用，关闭超时定时器，本次操作结束。当 ICMP Request 报文响应时间超时，输出该 MAC 地址无设备使用的信息。

ARP-Ping MAC 的基本原理

ARP-Ping MAC 发送的是广播的 ICMP 请求（ECHO Request）报文。以下是 ARP-Ping MAC 的具体实现过程：

1. 用户从命令行通过设置指定的 MAC 地址后，发送广播的 ICMP 报文并且启动超时定时器。
2. 局域网内各路由设备或主机收到 ICMP 请求报文后，回复 ICMP 应答（ECHO Reply）报文。
3. 源路由设备收到 ICMP 应答报文后，将 Reply 报文中的源 MAC 地址和命令行中输入的 MAC 地址相比较。若匹配，显示出该报文的 IP 地址，并提示该 MAC 地址已被使用并且关闭超时定时器，本次操作结束。

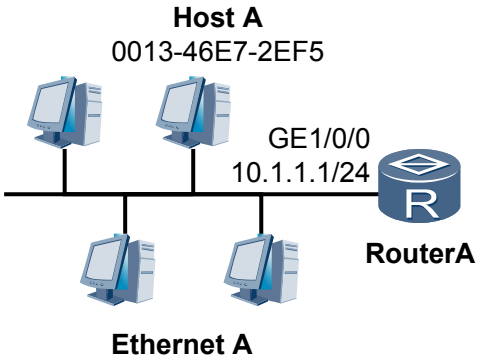
   若 ICMP 应答报文的响应时间定时器超时，输出该 MAC 地址无设备使用的信息。

说明

如果系统关闭了回复网段地址的报文请求，发送方是收不到 ICMP 响应报文的。

如图 2-5 所示，RouterA 可通过 ARP-Ping MAC 来获知 0013-46E7-2EF5 这个 MAC 地址是否被使用。RouterA 收到网络内所有主机回复的 ICMP 的响应报文后，将 MAC 地址为 0013-46E7-2EF5 的主机的 IP 地址显示出来。通过显示信息可得知这个 MAC 地址所对应的 IP 地址。

图 2-5 ARP-Ping MAC 的实现过程

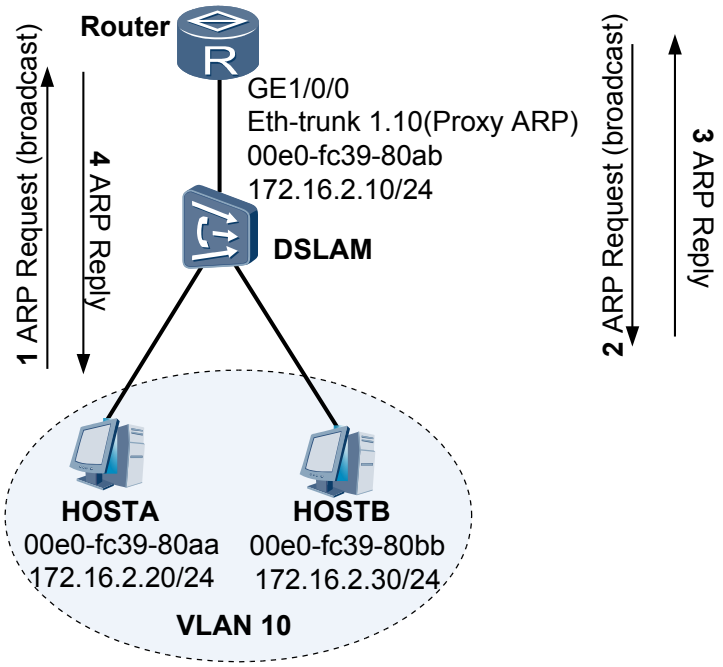


2.4 应用

VLAN 内 Proxy ARP

如图 2-6 所示，HOST A 和 HOST B 是 DSLAM 设备下的两个用户。连接 HOST A 和 HOST B 的两个接口在 DSLAM 上属于同一个 VLAN10。由于在 DSLAM 上配置了 VLAN 内不同接口彼此隔离，因此 HOST A 和 HOST B 不能直接在二层互通。

图 2-6 VLAN 内 Proxy ARP 典型组网图

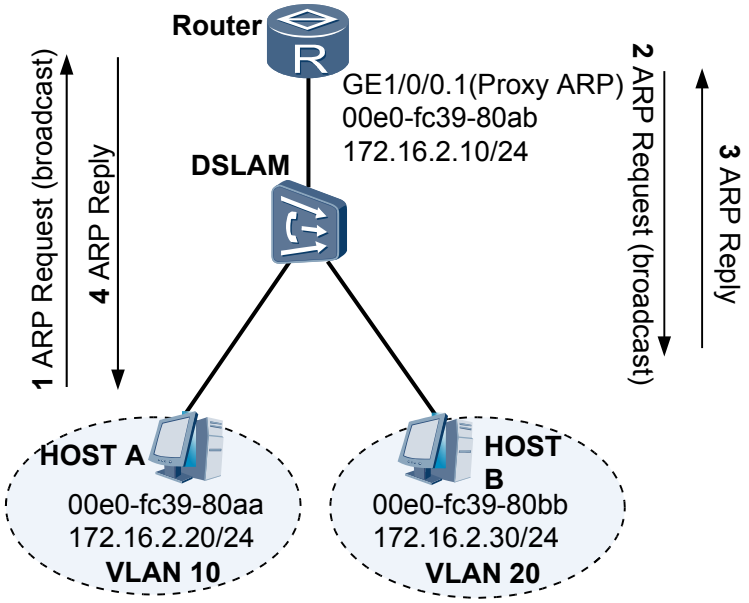


如果在 Router 上创建子接口 Eth-trunk1.10，使子接口关联 VLAN10。在 Router 的子接口 Eth-trunk1.10 上使能 VLAN 内 Proxy ARP，HOST A 和 HOST B 就可以在二层互通了。子接口 Eth-trunk1.10 的 IP 地址与 VLAN10 中的主机 IP 地址必须在同一个网段。

VLAN 间 Proxy ARP

如图 2-7 所示，HOST A 和 HOST B 是 DSLAM 设备下的两个用户。由于连接 HOST A 和 HOST B 的两个接口在 DSLAM 上属于不同的 VLAN，因此 HOST A 和 HOST B 不能直接实现二层互通。

图 2-7 VLAN 间 Proxy ARP 典型组网图



如果在 Router 上创建子接口 GE1/0/0.1，在子接口上关联 VLAN10 和 VLAN20，并且在 GE1/0/0.1 上使能 VLAN 间 Proxy ARP，HOST A 和 HOST B 就可以实现二层互通了。子接口 GE1/0/0.1 的 IP 地址与 VLAN10 和 VLAN20 中的主机 IP 地址在同一个网段。

ARP 安全的应用

ARP 安全配置主要应用在网络的接入层和汇聚层。如图 2-8 和图 2-9 所示。ARP 安全可以有效防止 ARP 表项攻击和网段扫描攻击。

图 2-8 ARP 安全配置在接入层的典型组网图

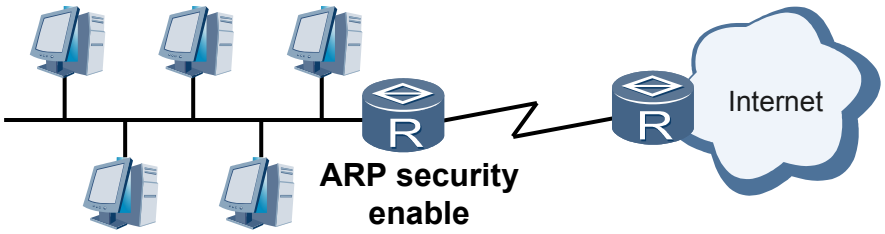
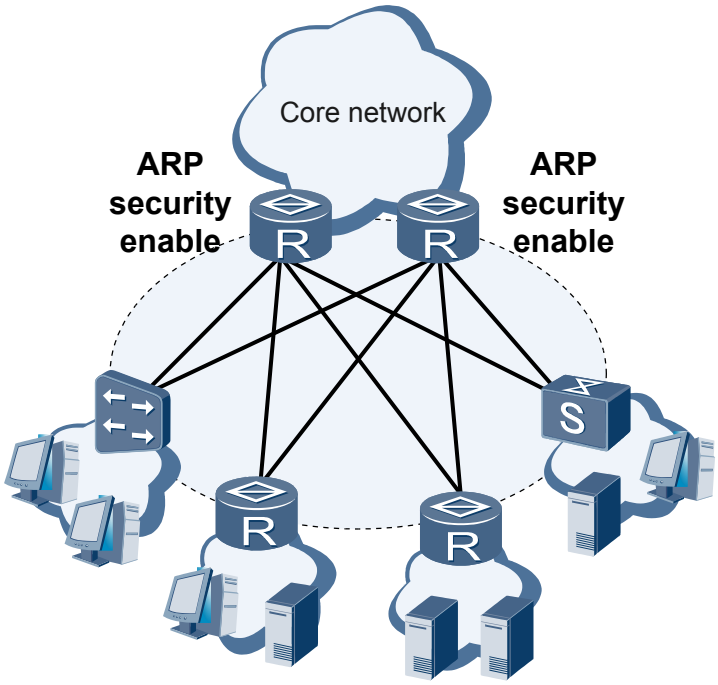


图 2-9 ARP 安全配置在汇聚层的典型组网图



在汇聚边缘启用 ARP 安全功能，可以过滤大量不信任的 ARP 报文以及对某些 ARP 报文进行时间戳抑制，以此来保证核心网络设备的安全性和稳定性。

2.5 术语与缩略语

术语/缩略语

缩略语	英文全称	中文全称
ARP	Address Resolution Protocol	地址解析协议
VRRP	Virtual Router Redundancy Protocol	虚拟路由冗余协议
VLAN	Virtual Local Area Netw	虚拟局域网
IPoEoA	IP over Ethernet over AAL5	AAL5 上承载以太网帧

3 DNS

关于本章

3.1 介绍

定义

TCP/IP 提供了通过 IP 地址来确定设备的功能，但对用户来讲，记住某台设备的 IP 地址是相当困难的，因此专门设计了一种字符串形式的主机命名机制，这些主机名与 IP 地址一一对应。在 IP 地址与主机名之间需要有一种转换和查询机制，提供这种机制的系统就是域名系统 DNS（Domain Name System）。

目的

域名系统 DNS 使用一种有层次的命名方式，为网上的设备指定一个有意义的名字，并且在网络上设置域名解析服务器，建立域名与 IP 地址的对应关系。这样用户就可以使用便于记忆的、有意义的域名，而不必去记忆复杂的 IP 地址。

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
RFC1034	DOMAIN NAMES - CONCEPTS AND FACILITIES	
RFC1035	DOMAIN NAMES - IMPLEMENTATION AND SPECIFICATION	不支持 DNS 服务器功能，只支持 A 类查询（即请求获得域名对应的 IP 地址） 注：后续版本支持 PTR 查询（即请求获得一个 IP 地址对应的域名）

3.3 原理描述

域名解析分为动态解析和静态解析，二者可以相辅相成。在解析域名时，先采用静态解析的方法，如果静态解析不成功，再采用动态解析的方法。将一些常用的域名放入静态域名解析表中，可以提高域名解析效率。

3.3.1 静态 DNS

3.3.2 动态 DNS

3.3.1 静态 DNS

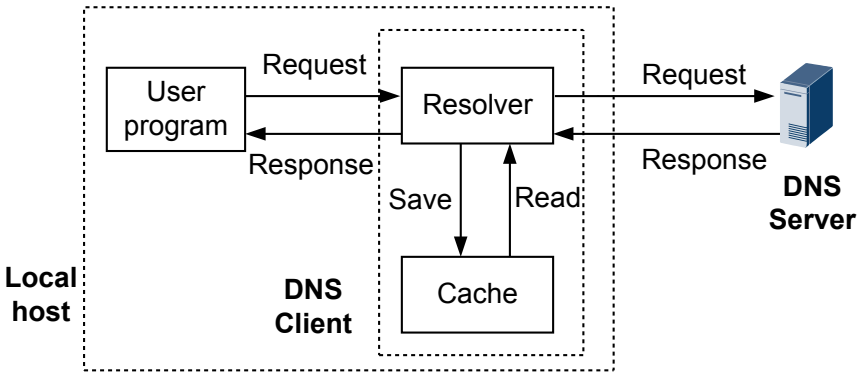
静态域名解析通过静态域名解析表进行，即手动建立域名和 IP 地址之间的对应关系表，将一些常用的域名放入表中。当客户机需要域名所对应的 IP 地址时，首先到静态域名解析表中查找指定的域名，从而获得所对应的 IP 地址，提高域名解析的效率。

3.3.2 动态 DNS

用户程序（例如 Ping、Tracert）对域名服务器（DNS Server）的访问是通过 DNS 客户端（DNS Client）的一个地址解析器（Resolver）完成的。

用户程序（例如 Ping、Tracert）、解析器和域名服务器以及解析器上的缓存区关系如图 3-1 所示。

图 3-1 动态 DNS



其中解析器和缓存区集成在一起构成 DNS Client，它的作用是接受用户程序的 DNS 咨询，并对其做出反应。一般来说，用户程序（例如 Ping、Tracert）、缓存区和解析器是在同一台主机上，域名服务器和它们在不同的主机上。

动态 DNS 的工作过程

1. 用户程序（例如 Ping、Tracert）首先向 DNS Client 发出请求。
2. DNS Client 收到请求后，首先查询本机数据库/缓存，如果没有发现所要查找的映射项，就向域名服务器发送查询报文。
3. 域名服务器收到查询报文后，首先判断请求的域名是否处于自己被授权管理的子域里，再根据不同的判断结果，向 DNS Client 发送相应的响应报文。
4. DNS Client 收到响应后，解析域名服务器发回来的响应报文，并根据响应报文的内容决定下一步的操作。

域名后缀列表功能

动态域名解析支持域名后缀列表功能，用户可以预先设置一些域名后缀，在域名解析的时候，用户只需要输入域名的部分字段，系统会自动将输入的域名加上不同的后缀进行解析。

域名解析方式

动态域名解析需要专用的域名解析服务器，该服务器运行域名解析服务器程序，提供从域名到 IP 地址的映射关系，负责处理客户提出的域名解析请求。

域名解析服务器接收到客户端提出的域名解析请求后，首先判断请求的域名是否处于自己被授权管理的子域里。如果是，就查询数据库，把域名转换为 IP 地址，并将转换结果发送给客户端。如果域名解析服务器不能解析出域名，它就根据客户在查询报文中所指明的解析方式（递归解析或者迭代解析）来进行下一步操作。

有以下两种域名解析方式：

- 递归解析

  域名解析服务器和其他能解析该域名的服务器联系，并将查询结果即域名所对应的 IP 地址返回给客户端。

- 迭代解析

  若该域名解析服务器不能提供解析结果，会在给客户端的响应报文中指明客户端应联系的下一个域名解析服务器。客户端会向指明的下一个域名解析服务器再次发出查询请求。

查询类型

目前，设备支持 DNS 客户端功能，支持的查询类型是 A 类查询。

A 类查询是最常用的查询类型，用于请求获得域名对应的 IP 地址。例如在 ping 和 tracert 的时候，可以 ping 或 tracert 一个域名，此时 ping 或 tracert 作为用户程序会向系统中 DNS 客户端查询该域名对应的 IP 地址。如果系统中没有该域名对应的 IP 地址信息，DNS 客户端就会向 DNS 服务器发起 A 类查询，获取该域名对应的 IP 地址，完成 ping 和 tracert 的功能。

3.4 术语与缩略语

术语

术语	解释
DNS Server	DNS 服务器。能在网络上给客户端提供域名解析服务的设备。
DNS Client	向 DNS 服务器发出请求并等待响应的设备。

缩略语

缩略语	英文全称	中文全称
DNS	Domain Name System	域名系统

4 ACL

关于本章

4.1 介绍

定义

访问控制列表是一系列有顺序的规则组的集合，这些规则根据数据包的源地址、目的地址、端口号等来描述。ACL 通过规则对数据包进行分类，这些规则应用到路由设备上，路由设备根据这些规则判断哪些数据包可以接收，哪些数据包需要拒绝。

例如可以用访问列表描述：拒绝任何用户终端使用 Telnet 登录本机。允许每个用户终端经由 SMTP 向本机发送电子邮件。

每个 ACL 中可以定义多个规则，根据规则的功能分为接口 ACL 规则、基本 ACL 规则和高级 ACL 规则。ACL 规则是一个匹配选项的集合，由用户根据不同业务进行选择配置。

说明

针对不同的业务，ACL 的匹配选项支持情况不相同。

ACL（Access Control List）的类型划分方式有四种，如表 4-1。

表 4-1 ACL 的分类

ACL 类型划分依据	ACL 类型
按照对 IPv4 和 IPv6 的支持情况	● ACL4 ● ACL6
按照 ACL 规则的功能	● 接口 ACL：限制数据包“允许”或“拒绝”通过接口。数字范围是 1000 ～ 1999，即支持 1000 个接口 ACL。 ● 基本 ACL：限制数据包的源地址。数字范围是 2000 ～ 2999，即支持 1000 个基本 ACL。 ● 高级 ACL：限制数据包的源地址、目的地址、协议号（TCP、UDP）、源端口和目的端口号的五元组。包括数字型高级 ACL 和命名型 ACL： – 数字型高级 ACL 的编号范围是 3000 ～ 3999，即支持 1000 个数字型高级 ACL。 – 命名型 ACL 的编号范围是 42768 ～ 75535，即支持 32768 个命名型 ACLL。 ● MPLS ACL:限制 MPLS 报文的 Exp 值、Lable 值、TTL 值。编号范围是 10000 ～ 10999，即支持 999 个 MPLS ACL。

根据 ACL 功能划分的四种类型 ACL，分别支持的过滤选项如表 4-2。

表 4-2 不同类型 ACL 所支持的过滤选项

ACL 规则类型	支持的过滤选项
接口 ACL	接口名：指定数据包是从该接口进入的。或者用“any”代表所有的接口。 生效时间段：指定规则生效的时间范围。如果不配置，表示规则配置后马上生效。
基本 ACL	源 IP 地址：指定 ACL 规则的源地址信息。如果不配置，表示任何源地址的报文都匹配。 生效时间段：指定规则生效的时间范围。如果不配置，表示规则配置后马上生效。
高级 ACL	协议类型：用名字或数字表示的协议类型。如果用整数形式，取值范围是 1 ～ 255；如果用字符串形式，可以选取：gre、icmp、igmp、ip、ipinip、ospf、tcp、udp。对不同的协议类型，有不同的参数组合，TCP 和 UDP 有源端口和目的端口可选项，其它协议类型没有。 源 IP 地址：指定 ACL 规则的源地址信息。如果不配置，表示报文的任何源地址都匹配。 目的 IP 地址：指定 ACL 规则的目的地址信息。如果不配置，表示报文的任何目的地址都匹配。 源端口和目的端口：指定 UDP 或者 TCP 报文的目的端口信息，仅仅在规则指定的协议号是 TCP 或者 UDP 时有效。如果不指定，表示 TCP/UDP 报文的任何目的端口信息都匹配。 dscp：指定区分服务代码点（Differentiated Services Code Point，IP 头 TOS 字段的高 7 位）的取值，取值范围是 0 ～ 63。 分片报文类型：指定该规则是否仅对非首片分片报文有效。当包含此参数时表示该规则仅对非首片分片报文有效。 优先级：数据包可以依据优先级字段（IP 包 TOS 字段的高 3 位）进行过滤。用关键字或数字表示，数字的取值范围是 0 ～ 7 的整数。 TCP flag：指定 TCP-FLAG 的值，取值范围是 0 ～ 63。 TOS：数据包可以依据服务类型字段进行过滤。 ICMP：ICMP 包可以依据 ICMP 的消息名称、消息类型或消息码进行过滤，仅仅在报文协议是 ICMP 的情况下有效。如果不配置，表示任何 ICMP 类型的报文都匹配。 生效时间段：指定规则生效的时间范围。如果不配置，表示规则配置后马上生效。
MPLS ACL	MPLS 报文的 Exp 值。如果不配置，表示 Exp 为任何值的 MPLS 报文都匹配。 MPLS 报文的 Lable 值。如果不配置，表示 Lable 为任何值的 MPLS 报文都匹配。 MPLS 报文的 TTL 值。如果不配置，表示 TTL 为任何值的 MPLS 报文都匹配。

目的

各种业务通过引用 ACL，对路由或报文进行规则匹配后处理。

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
RFC4314	Defines several new access control rights and clarifies which rights are required for different IMAP commands.	

4.3 原理描述

ACL 负责管理用户配置的所有规则，并提供规则匹配算法。业务根据匹配的规则动作（“允许”或“拒绝”）进行操作。

ACL 的管理规则

每个 ACL 作为一个规则组，可以保存多个规则。当添加超规格的 ACL 组和规则时，提示用户配置不成功。

ACL 的规则的匹配

规则匹配的情况：指存在 ACL 且 ACL 中有符合条件的规则，不论匹配的动作是“允许”或“拒绝”，都是匹配的。

规则不匹配的情况：指不存在 ACL 或 ACL 中无规则或查找了 ACL 下所有规则都不符合匹配条件。

ACL 的规则匹配过程

1. 首先查找用户是否配置了该 ACL（因为有些业务可能允许引用不存在的 ACL，例如 QoS 和 OSPF）。
2. 根据 ACL 的配置情况：
   - 如果存在 ACL 且需要根据规则来检查报文，则查找该 ACL 中所有规则，只要有一条规则和报文匹配，就直接将规则匹配的动作通知给业务，不再继续查找后续的规则。
   - 如果存在 ACL 且业务只匹配源地址信息、目的地址信息、IP 承载的协议类型、TCP 的源端口、目的端口、ICMP 协议的类型中的某些选项，则根据业务要求查找所有 ACL，进行规则匹配。匹配的第一条就通知给业务，不再继续查找后续的规则。

ACL 的规则匹配顺序

规则显示顺序决定匹配顺序。规则匹配时，从 ACL 中显示的第一条规则开始查找，当找到一条符合匹配条件的规则时，结束查找。即规则越靠前越容易被匹配。

决定规则显示顺序的因素有两个：规则 ID 和规则匹配方式。

规则匹配方式有两种：配置顺序和自动顺序。

- 如果是配置顺序，按照用户配置 ACL 规则的先后进行匹配。可以由用户配置规则 ID，也可以由系统根据步长自动生成（步长可以方便用户进行规则维护，方便插入新规则。例如：ACL4 默认步长为 5，当用户不输入规则 ID 时，系统自动生成的第一条规则的 ID 就是 5，当用户想在规则 5 前面插入新规则时，只需要输入比 5 小的规则 ID 即可，排序后新规则就成了第一条规则）。
- 如果是自动顺序，由系统自动分配规则 ID，按照“深度优先”规则把精度最高的规则排在最前面。这一点可以通过比较地址的通配符来实现，通配符越小，则指定的主机的范围就越小。

  比如 129.102.1.1 0.0.0.0 指定了一台主机：129.102.1.1，而 129.102.1.1 0.0.0.255 则指定了一个网段：129.102.1.1 ～ 129.102.1.255，显然前者指定的主机范围小，在访问控制规则中排在前面。具体标准如下。
  - 对于基本访问控制规则的语句，直接比较源地址通配符，通配符相同的则按配置顺序；
  - 对于基于接口的访问控制规则，配置了“**any**”的规则排在后面，其它按配置顺序；
  - 对于高级访问控制规则，首先比较协议范围，再比较源地址通配符，相同时再比较目的地址通配符，仍相同时则比较端口号的范围，范围小的排在前面，如果端口号范围也相同则按配置顺序。

规则通过规则 ID 来标识，规则 ID 可以由用户进行配置，也可以由系统自动根据步长生成。一个 ACL 中所有规则均按照规则 ID 从小到大排序。

规则 ID 之间会留下一定的空间，具体空间大小由“ACL 的步长”来设定。例如步长设定为 5，ACL 规则 ID 分配是按照 5、10、15……这样来分配的。如果步长值是 2，自动生成的规则 ID 从 2 开始。这样做是为了便于用户在第一条规则前面插入新规则。

- 在“配置顺序”的情况下
  - 如果配置规则的时候没有指定“规则 ID”，则系统会根据“ACL 步长”照用户配置规则的先后顺序，自动为规则分配规则编号。例如：用户配置了 3 条没有指定“规则 ID”的规则，如果 ACL 步长为 5，则系统按照这 3 条规则的配置顺序为它们分别分配规则编号：5，10，15。
  - 如果配置规则的时候指定了“规则 ID”，则会按照“规则 ID”的位置决定该规则的插入位置。例如系统现在的规则编号是：5、10、15。如果指定“规则 ID”为 3，创建一条 ACL 规则，则规则的规则顺序就为：3、5、10、15，相当于在规则 5 之前插入了一条规则。

  因此，在“配置顺序”的情况下，系统会按照用户配置规则的先后顺序进行匹配。但本质上，系统是按照规则编号的顺序，由小到大进行匹配，后插入的规则有可能先执行。
- 在“自动排序”的情况下

  在“自动排序”的情况下，无法为规则指定“规则 ID”。系统会按照“深度优先”原则自动为规则分配规则编号，用户无法通过指定规则 ID 插入规则。指定数据包范围较小的规则将获得较小的规则编号。系统将按照规则编号的顺序，由小到大进行匹配。

  说明

  对已经存在的 ACL 规则中不相关字段进行编辑，相当于追加新的规则，不会影响已经存在的规则。

4.3.1 ACL4 和 ACL6 的区别

4.3.1 ACL4 和 ACL6 的区别

ACL4 和 ACL6 的基本原理是一样的，只有很小差别，如表 4-3。

表 4-3 ACL4 和 ACL6 的区别

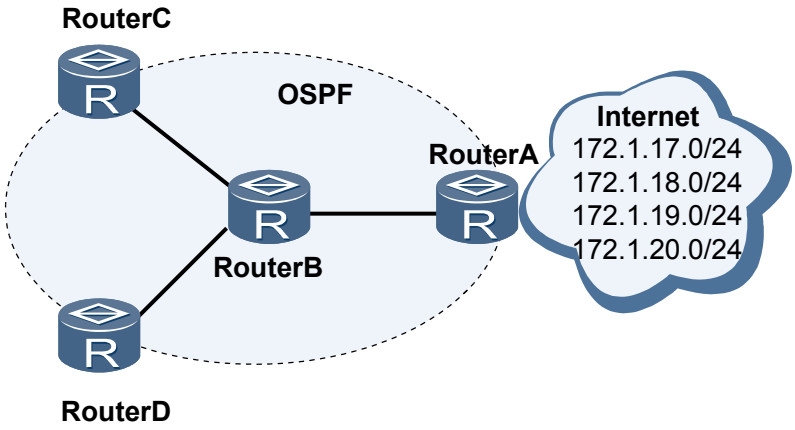
ACL4	ACL6
能配置步长，能对规则进行排序。	不能配置步长，不能对规则进行排序。
支持 MIB。	无 MIB。

4.4 应用

在路由过滤中使用 ACL

ACL 可以应用在各种动态路由协议中，对路由协议发布和接收的路由信息进行过滤。

图 4-1 在路由过滤中使用 ACL



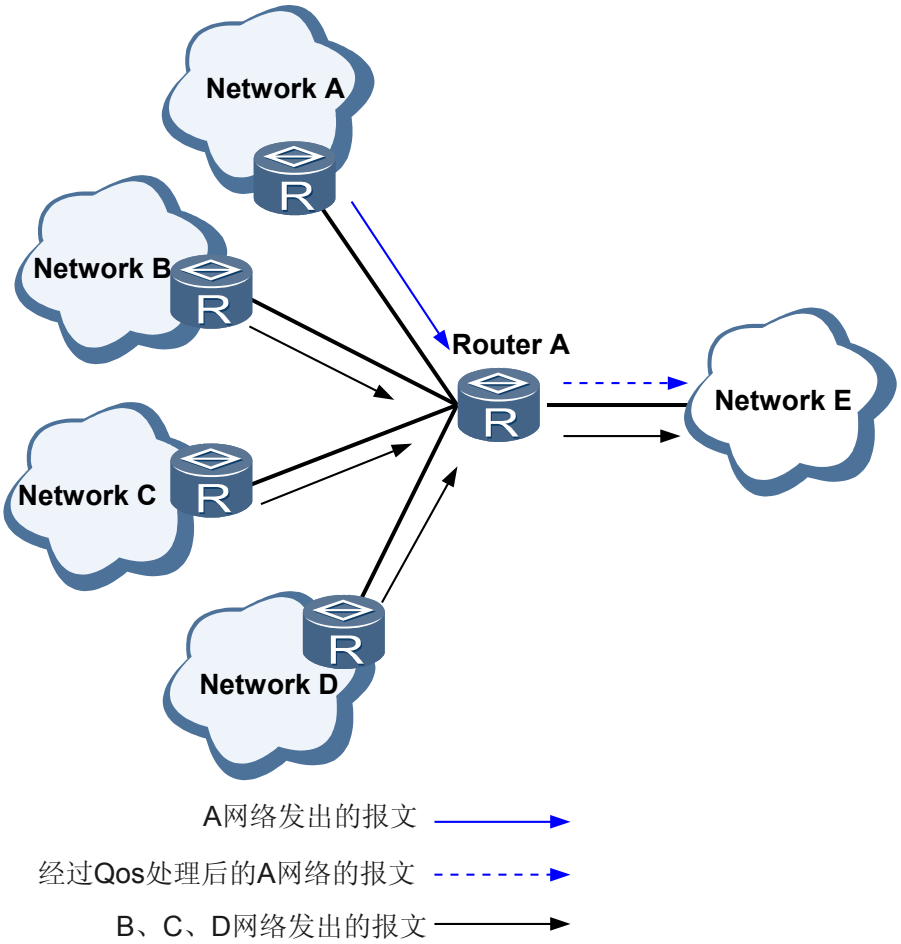
如图 4-1 所示，在运行 OSPF 协议的网络中，RouterA 从 Internet 网络接收路由，并为 RouterB 提供了部分 Internet 路由。在 RouterA 上定义 ACL 列表，并在 OSPF 协议中应用 ACL 过滤，可控制路由的发布和接收，如：

- RouterA 仅提供 172.1.17.0/24、172.1.18.0/24 和 172.1.19.0/24 给 RouterB；
- RouterC 仅接收路由 172.1.18.0/24；
- RouterD 接收 RouterB 提供的全部路由。

在 QOS 中使用 ACL

利用 ACL 对具有某种属性的报文进行 QOS 处理。

图 4-2 在 QOS 中使用 ACL



如图 4-2 所示，在 RouterA 上使用 ACL，标识来自 A 网络的所有报文，然后在 QoS 策略中引用这个 ACL，这样，所有来自 A 网络的报文都会被 RouterA 进行 QoS 处理后转发，而来自其它网络的所有报文，因为没有匹配 ACL 而正常的转发。

4.5 术语与缩略语

术语

术语	解释
基于接口的 ACL（Interface-based ACL）	基于接口的访问控制列表可以根据接收报文的接口指定规则。
基本 ACL（Basic ACL）	基本访问控制列表只能使用源地址信息作为定义访问控制列表规则的元素。
高级 ACL（Advanced ACL）	高级访问控制列表可以使用数据包的源地址信息、目的地址信息、协议类型、TCP 的源端口、目的端口、ICMP 协议的类型、ICMP 报文的消息码等元素定义规则。

术语	解释
基于 MPLS 的 ACL	基于 MPLS 的 ACL 根据 MPLS 报文的 Exp 值、Label 值、TTL 值过滤报文。

缩略语

缩略语	英文全称	中文全称
ACL	Access Control List	访问控制列表

5 IPv4

关于本章

5.1 介绍

定义

IPv4（Internet Protocol Version 4）是 TCP/IP 协议族中最为核心的协议。它工作在 TCP/IP 栈的互联网络层。该层与 OSI 参考模型的网络层相对应。IP 层提供了无连接数据传输服务，即将信息分割成数据单元，以数据报的形式从网络的一个地方传送到另一个地方。

目的

屏蔽各链路层差异，为上层提供统一的网络层传输标准。

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
RFC793	Transmission Control Protocol	
RFC768	User Datagram Protocol	

5.3 原理描述

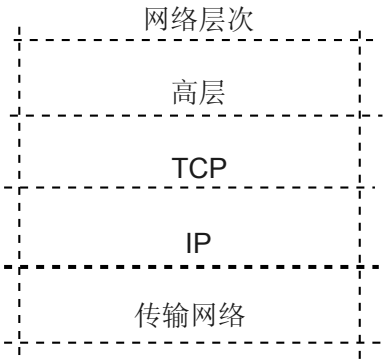
5.3.1 TCP 原理描述

传输控制协议（TCP）由 RFC793 定义，用于在主机间实现面向连接的可靠性服务。TCP 协议为用户进程定义了一个可靠的、面向连接的、全双工的服务。

TCP 是面向连接的端到端的可靠协议。它支持多种网络应用程序。TCP 假定下层只能提供不可靠的数据报服务，它可以在多种硬件构成的网络上运行。

图 5-1 表示了 TCP 在层次式结构中的位置，它的下层是 IP 协议，TCP 可以根据 IP 协议提供的服务传送大小不定的数据，IP 协议负责对数据进行分段、重组，在多种网络中传送。

图 5-1 层次式结构



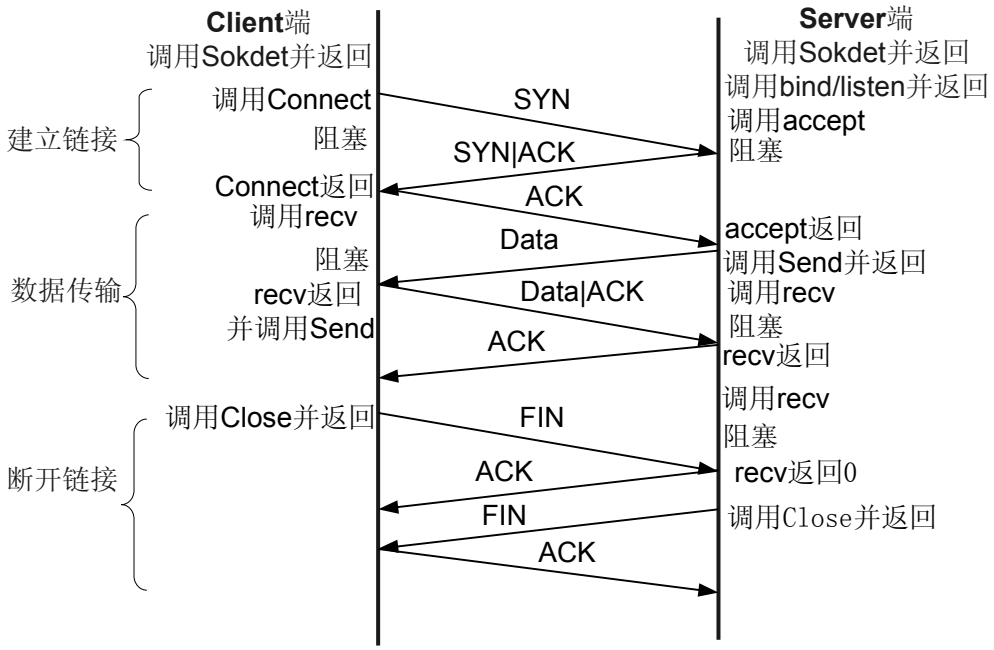
在 ISO 层次结构中，TCP 的上层是应用程序，下层是 IP 协议。

对于上层应用程序，TCP 应该能够异步传送数据。下层接口假定为 IP 协议接口。为了在并不可靠的网络上实现面向连接的可靠的传送数据，TCP 必须：

- 解决可靠性、流量控制的问题
- 为上层应用程序提供多个接口
- 为多个应用程序提供数据
- 解决连接问题
- 解决通信安全性的问题

图 5-2 表示了 TCP 连接建立和拆除过程。

图 5-2 TCP 连接建立和拆除过程



5.3.2 UDP 原理描述

UDP（用户数据报协议）是用来在互连网络环境中提供包交换的计算机通信协议。此协议默认为网路协议（IP）是其下层协议，提供了向另一用户程序发送信息的最简便的协议机制。UDP 是面向操作的，未提供数据提交和复制保护。如果应用程序要求可靠的数据传送应该使用传输控制协议（TCP）。数据报格式如图 5-3 所示。

图 5-3 UDP 协议报文格式

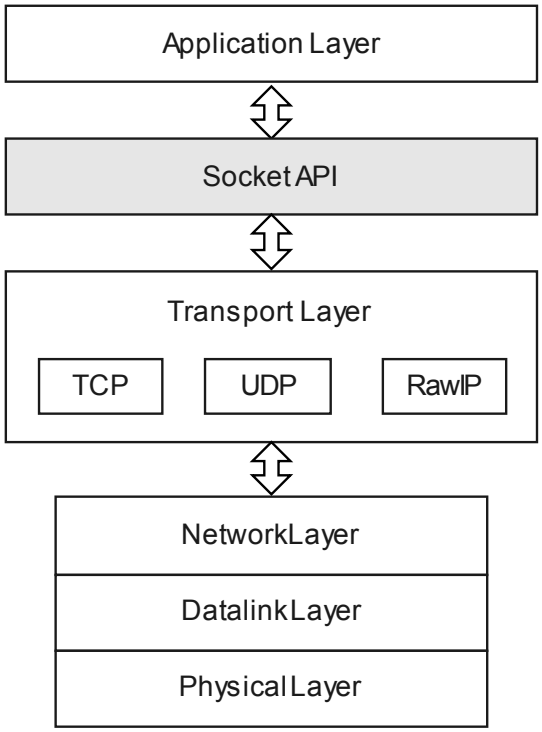
5.3.3 RawIP 原理描述

RawIP 只填充 IP 首部的有限几个字段，但它允许应用进程提供自己的 IP 首部。它与 UDP 类似，是不可靠的，即没有任何控制能确定 RawIP 数据报是否已被接收。它是无连接的，即在主机间传输数据时，不需要任何类型的电路。RawIP 相比 UDP 的区别在于：RawIP 允许应用程序直接通过 Socket 接口操作 IP 层。对于许多需要跟下层直接交互的应用，非常方便。

5.3.4 Socket 原理描述

Socket 是一组编程接口（API），介于传输层与应用层之间，屏蔽传输层差异，向应用层提供统一的编程接口。应用层可以不必了解 TCP/IP 协议的细节，直接通过对 Socket 接口函数的调用完成数据在 IP 网络中的传输。图 5-4 表示了 Socket 在 TCP/IP 协议栈中的位置。

图 5-4 Socket 分层模型



基于传输层差异，目前支持四种类型的 Socket：

- 基于 TCP 的 Socket，向应用层提供的一种可靠的流式数据通讯服务。
- 基于 UDP 的 Socket，向应用层提供给一种无连接的，不可靠的数据传输，但是这种基于数据报的传输可以提供报文边界。
- 基于 RawIP 的 Socket，也叫 Raw Socket。与基于 UDP 的 Socket 类似，也是无连接的，不可靠的数据传输，同样可以提供报文边界。但是它的特点是能够使应用程序直接访问网络层。
- 基于链路层的 Socket，这是为 IS-IS 路由协议提供的 Socket 接口，使 IS-IS 路由协议可以通过该 Socket 接口直接访问链路层。

5.4 应用

ICMP 报文发送开关

在正常情况下，设备可以正确发送 ICMP 主机不可达报文和 ICMP 重定向报文。但是，当网络流量较大时，则设备会发送大量的 ICMP 报文，增大网络的流量负担。同时，ICMP 差错报文经常被利用发起网络攻击，容易产生恶性循环，从而加剧网络的拥塞。如重定向报文可以被用来使路由频繁变更。

NE20E-X6 提供在 ICMP 报文的出接口增加两个控制开关，分别用来打开或关闭 ICMP 主机不可达报文和 ICMP 重定向报文的发送开关。如果关闭这两个开关，则路由设备不会发送这两种报文，从而起到减小网络流量、降低设备负担、防止遭到恶意攻击的作用。

5.5 术语与缩略语

术语

无

缩略语

缩略语	全称
TCP	Transmission Control Protocol
UDP	User Datagram Protocol

6 IPv6

关于本章

6.1 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 Engineering Task Force，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。IPv6 以其简化的报文头格式、充足的地址空间、层次化的地址结构、灵活的扩展头、增强的邻居发现机制将在未来的市场竞争中充满活力。

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。

- 骨干设备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。

6.2 参考标准和协议

本特性的参考资料清单如下：

文档	描述	备注
RFC793	Transmission Control Protocol	
RFC768	User Datagram Protocol	
RFC1981	Path MTU Discovery for IP version 6	
RFC2460	Version 6 of the Internet Protocol (IPv6), also sometimes referred to as IP Next Generation or IPng.	
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	
RFC2473	Generic Packet Tunneling in IPv6 Specification	
RFC2711	IPv6 Router Alert Option	
RFC2893	Transition Mechanisms for IPv6 Hosts and Routers	
RFC3056	Connection of IPv6 Domains via IPv4 Clouds	
RFC3068	An Anycast Prefix for 6to4 Relay Routers	
RFC3484	Default Address Selection for Internet Protocol Version 6 (IPv6) Section 2.1	
RFC3971	SEcure Neighbor Discovery (SEND)	
RFC3972	Cryptographically Generated Addresses (CGA)	
RFC4191	Default Router Preferences and More-Specific Routes	

文档	描述	备注
RFC4214	Intra-Site Automatic Tunnel Addressing Protocol(ISATAP)	
RFC4291	Internet Protocol Version 6 (IPv6) Addressing Architecture	
RFC4443	Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification)	
RFC4861	Neighbor Discovery for IP version 6 (IPv6)	

6.3 原理描述

IPv6 基本功能主要包括 IPv6 邻居发现、IPv6 路径 MTU 发现。邻居发现和 Path MTU 发现机制均是基于 ICMPv6 协议报文实现的。

6.3.1 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。

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。

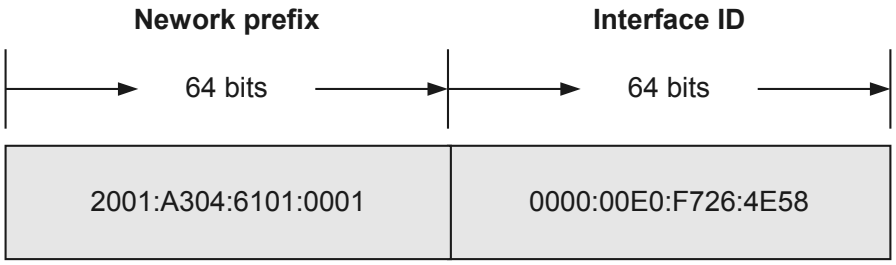
IPv6 地址的结构

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 6-1 所示。

图 6-1 地址 2001:A304:6101:1::E0:F726:4E58 /64 的构成示意图



IPv6 的地址分类

IPv6 主要有三种地址：

- 单播地址（Unicast）：唯一标识一个接口，类似于 IPv4 的单播地址。发送到单播地址的数据包将被传输到此地址所标识的唯一接口。

  单播地址还可以分为四种，如表 6-1 所示。

表 6-1 IPv6 单播地址类型

地址类型	二进制前缀	IPv6 前缀标识
链路本地单播地址	1111111010	FE80::/10

地址类型	二进制前缀	IPv6 前缀标识
环回地址	00...1 (128 bits)	::1/128
未指定地址	00...0 (128 bits)	::/128
全球单播地址	其他	-

表中各类地址的意义如下：

- 链路本地单播地址：用于邻居发现协议和无状态自动配置进程中链路本地上节点之间的通信。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 0:0:0:0:0:0:0:1 或::1，不会被分配给任何接口。它的作用与在 IPv4 中的 127.0.0.1 相同，即节点将 IPv6 报文发送给自己。
- 未指定地址（::），不能被分配给任何节点，也不能作为目的地址。在主机初始化且没有取得自己的地址时，未指定地址可以用在 IPv6 报文的源地址字段，例如重复地址探测时，NS 报文的源地址就是未指定地址。
- 全球单播地址等同于 IPv4 公网地址。用于可以聚合的链路，最后提供给网络服务提供商。这种地址类型的结构允许路由前缀的聚合，从而满足全球路由表项的数量限制。地址包括运营商管理的 48 位路由前缀和本地站点管理的 16 位子网 ID，以及 64 位接口 ID。如无特殊说明，全球单播地址包括站点本地单播地址。

- 任播地址（Anycast）：用来标识一组接口（通常这组接口属于不同的节点）。发送到任播地址的数据包被传输给此地址所标识的一组接口中距离源节点最近的一个接口（最“近”的一个，是指根据路由协议的距离度量）。

  应用场合：当移动主机需要与它的“home”子网上的移动代理之一通信时，它将用该子网路由设备的任播地址。

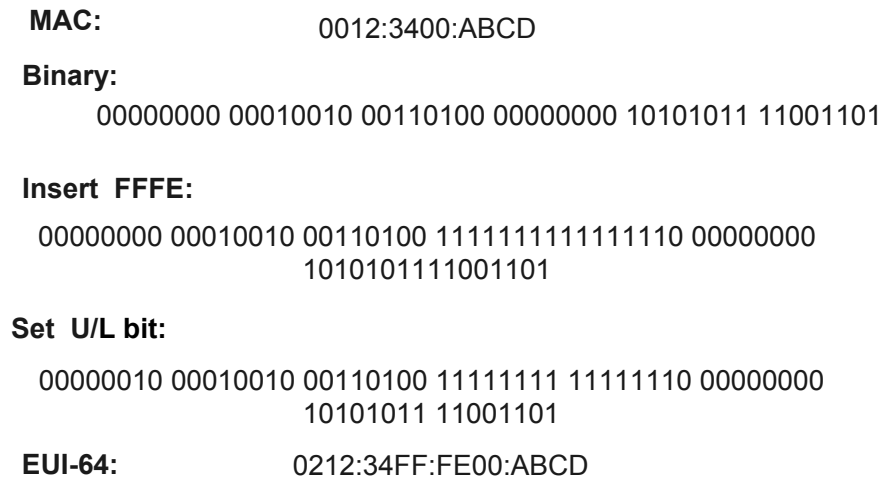
  具体地址规定：任播地址没有独立的地址空间，它们可使用任何单播地址的格式。因此，需要一种语法来区别任播地址和单播地址。
- 组播地址（Multicast）：用来标识属于不同节点的一组接口，类似 IPv4 的组播地址。发送到组播地址的数据包被传输给此地址所标识的所有接口。

  IPv6 不包括广播地址，广播地址的功能均由组播地址来提供。

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 6-2。

图 6-2 MAC 地址到 EUI-64 格式的转换过程



6.3.2 IPv6 的特点

- 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。

- 源/目的地址选择

  当网络管理者需要指定和预知系统发送报文的源/目的地址时，可以定义一组地址选择规则，这些规则构成地址选择策略表。该表类似于路由表，使用最长匹配原则查找规则。地址选择的结果是由源地址和目的地址共同决定的。

  依次根据以下规则进行源地址选择，规则的编号越小，优先级越高。

  1. 源地址和目的地址相同
  2. 合适的生效范围
  3. 避免使用已经废弃的地址
  4. 家乡地址（home address）
  5. 出接口地址
  6. 源地址的 *label* 值和目的地址的 *label* 值相同
  7. 最长匹配原则

  说明

  备选的源地址可以限定为配置在出接口上的单播地址，如果在出接口上没有找到和目的地址具有相同 *label* 值和范围的源地址，则可以选择其他接口上具有相同 *label* 值和范围的地址作为源地址。

依次根据以下规则进行目的地址选择，规则的编号越小，优先级越高。

1. 避免使用不可用的目的地址
2. 合适的生效范围
3. 避免使用已经废弃的地址
4. 家乡地址（home address）
5. 目的地址的 *label* 值和源地址的 *label* 值相同
6. 较高的 *precedence* 值
7. 在本地转发报文，不需要使用 6over4 或 6to4 隧道
8. 更小的生效范围
9. 最长匹配原则
10. 遵循原来的顺序

- 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

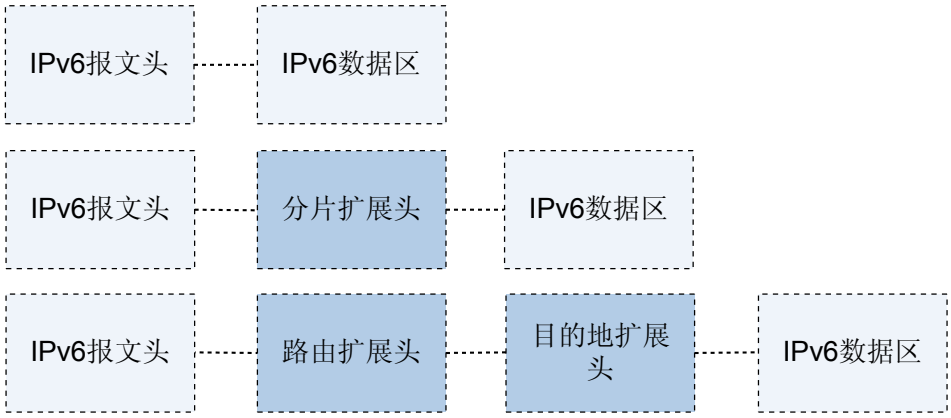
  IPv6 将 IPSec 作为它的扩展报头实现，提供端到端的安全特性。这一特性为解决网络安全问题提供了标准，并提高了不同 IPv6 实现的互操作性。

- 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 6-3 所示。

图 6-3 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

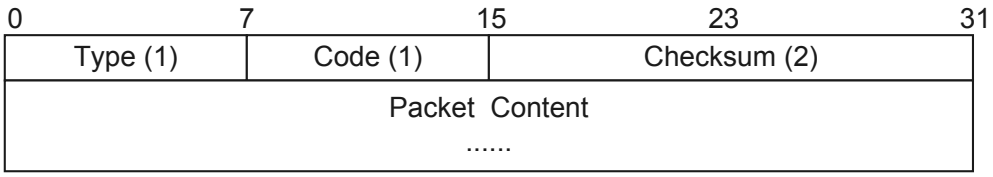
不是所有的扩展报头都需要被转发路由设备查看和处理的。路由设备转发时根据基本报头中 Next Header 值来决定是否要处理扩展头。

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。

6.3.3 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 6-4 所示。

图 6-4 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 表示差错报文类型，128 至 255 为消息报文类型。
- 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，当设备收到 Hop Limit 值等于 0 的数据包，或者当设备将 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。

6.3.4 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。

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。

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。

IPv6 安全邻居发现功能

IPv6 邻居发现协议（NDP，Neighbor Discovery Protocol）用来保证本链路内邻居的可达性，因此非常有必要保护 NDP 的安全。IPSec 方法可以在一定程度上保护 NDP，但是这种方法需要大量且复杂的手工配置。此时可以通过简单配置 IPv6 安全邻居发现（SEND，Security Neighbor Discovery）特性，实现对 IPv6 邻居发现协议的保护。

SEND 特性用来解决在 NDP 中涉及的安全问题，如：

- 重定向攻击：NS/NA 欺骗、恶意的最后一跳路由器、虚假的重定向报文、重放攻击。

  攻击节点可以使用携带不同源/目的链路层地址选项的 NS 报文，通过 NS/NA 欺骗，使合法节点的报文发往其他的链路层地址达到攻击的目的。
- 拒绝服务攻击（DoS，Denial of Service）：NUD 失败、DAD 攻击、虚假的地址配置前缀、参数欺骗。

  攻击者持续发送虚假的 NA 响应 NUD 的 NS，主机经过几次重试失败后就会删除被攻击者的邻居表项记录，造成被攻击者无法通信。攻击者还可以通过响应所有的 DAD 过程，通告已使用了被攻击者请求的地址，造成被攻击者因获取不到 IP 地址而无法正常运行。

为了解决上面的安全问题，SEND 中引入两种新的选项：CGA 选项（Cryptographically Generated Addresses）和 RSA 选项（Rivest Shamir Adleman）。

- CGA 是一种新的地址自动生成机制，可以用来验证 ND 报文的发送者对报文源地址拥有权的合法性（此地址指 ND 报文的源地址）。
- RSA 是对 ND 报文的数字签名，用来验证报文的完整性和发送者的真实性。

SEND 特性同时还定义了 ND 报文中的两种选项，用以解决 NDP 的安全问题：

- 随机数（Nonce）选项：用来在请求和回应交互中防止重放攻击，比如在 NS 和 NA 报文的交互中，NS 报文中携带 Nonce 选项，回应的 NA 报文中也携带此选项，发送者根据收到的选项判断是否为合法的回应报文。
- 时间戳（Timestamp）选项：用来保护非请求的通告报文和重定向报文。接收者应确保每个收到的报文其时间戳都比上一个收到的报文要新。

当接口需要拒绝接收非安全的 ND 报文时，可以配置 IPv6 安全邻居发现功能。如果满足以下条件中的任何一条，即为非安全的 ND 报文：

- 接收到的 ND 报文没有携带 CGA 和 RSA 选项，即发送该报文的接口没有配置 CGA 地址。
- 接收到的 ND 报文的密钥长度超出本接口可以接受的长度范围。
- 接收 ND 报文的速率超出系统接受的速率范围。
- 接收到 ND 报文的时间与发送 ND 报文的时间差超出本接口可以接受的时间差范围。

默认路由器优先级和路由信息

在邻居发现协议的 RA 报文中，定义了默认路由器优先级和路由信息两个字段，帮助主机在发送报文时选择合适的转发路由器。

当主机所在的链路中存在多个路由器时，主机需要根据报文的目的地址选择转发路由器。在这种情况下，路由器通过发布默认路由器优先级和特定路由信息给主机，提高主机根据不同的目的地选择合适的转发路由器的能力。

主机收到包含路由信息的 RA 报文后，会更新自己的路由表。当主机向其他设备发送报文时，通过查询该列表的路由信息，选择合适的路由发送报文。

主机收到包含默认路由器优先级信息的 RA 报文后，会更新自己的默认路由器列表。当主机向其他设备发送报文时，如果没有路由可选，则首先查询该列表，然后选择本链路内优先级最高的路由器发送报文；如果该路由器故障，主机根据优先级从高到低的顺序，依次选择其他路由器。

6.3.5 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。

6.3.6 TCP6

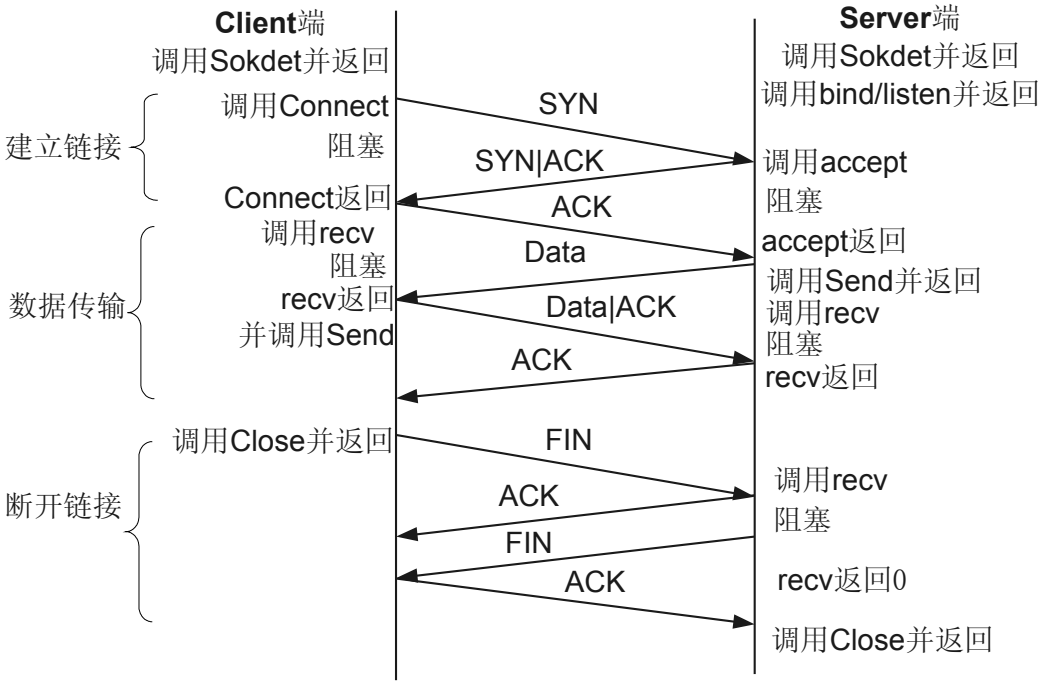
TCP6 提供了在两个端点的进程间建立虚电路的机制，一个 TCP6 虚连接如同在系统间承载数据的全双工电路。由于 TCP6 中提供了进程间数据的可靠传输，因此被称为可靠协议，它还提供了根据当前网络状态来优化传输性能的机制。在所有数据均可收到和确认的情况下，传输速率可以逐渐增加。延时将导致发送主机在收到进一步的确认前降低发送速率。

TCP6 通常用于交互式应用，如 WEB 之类，某些数据接收差错将影响正常的工作能力。TCP6 使用了“三次握手”机制来建立虚电路，所有的虚电路都需使用“四次握

手”拆除。这种连接方式可以提供多种校验和及其他可靠性功能，但是增加了使用 TCP6 的开销并导致其效率低于 UDP6。

如图 6-5 表示了 TCP6 连接建立和拆除的过程。

图 6-5 TCP6 连接建立和拆除过程示意图



6.3.7 UDP6

UDP6 是用来在互连网络环境中提供包交换的计算机通信协议。有如下特点：

- 只使用源和目的信息，主要用于简单的请求/响应式结构。
- 不可靠，即没有任何控制能确定 UDP6 数据报是否已被接收。
- 无连接，即在主机间传输数据时，不需要任何类型的虚电路。

UDP6 的无连接特性使得 UDP6 可以向广播地址发送数据；而 TCP6 则不同，它要求特定的源地址和目的地址。

6.3.8 RawIP6

RawIP6 较为简单，只填充 IPv6 首部的有限几个字段，允许应用进程提供自己的 IPv6 首部。

RawIP6 类似于 UDP6：

- 不可靠，即没有任何控制能确定 RawIP6 数据报是否已被接收。
- 无连接，即在主机间传输数据时，不需要任何类型的虚电路。

RawIP6 相比 UDP6 的区别在于，RawIP6 允许应用程序直接通过 Socket 接口操作 IP 层。对于许多需要跟下层直接交互的应用来说，非常方便。

6.4 术语与缩略语

术语

术语	解释
IPv6	Internet Protocol Version 6，下一代网际协议
ND	Neighbor Discovery，邻居发现，在 IPv6 报文转发过程中，用于地址冲突检测、邻居地址解析、确定邻居可达性，以及进行主机地址配置的一组协议和进程。由不同的 ICMPv6 报文实现路由器发现和邻居发现功能。
ICMPv6	Internet Control Management Protocol Version 6，Internet 互联网控制报文协议第 6 版，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 结点报告报文处理过程中的错误和信息。
PMTU	Path MTU，路径 MTU，利用 ICMPv6 数据包过大差错报文确定路径支持的最大传输单元的方法。

缩略语

缩略语	英文全称	中文全称
IPv6	Internet Protocol Version 6	网际协议第 6 版
ICMPv6	Internet Control Management Protocol Version 6	Internet 互联网控制报文协议第 6 版
ND	Neighbor Discovery	邻居发现
RS	Router Solicitation	路由器请求
RA	Router Advertisement	路由器通告
NS	Neighbor Solicitation	邻居请求
NA	Neighbor Advertisement	邻居通告
ARP	Address Resolution Protocol	地址解析协议
PMTU	Path MTU	路径 MTU
IPng	IP Next Generation	网络层协议的第二代标准协议
TCP6	Transmission Control Protocol 6	传输控制协议 6
UDP6	User Datagram Protocol 6	用户数据报协议 6
RawIP6	Raw IP6	原始 IP6

7 负载分担

关于本章

7.1 介绍

定义

负载分担是指在去往同一个目的地址有多条路由链路的情况下，将流量分担到多条路由。

表 7-1 负载分担类型

基于协议的负载分担	Trunk 负载分担	二级 Hash	二级负载分担
等值负载分担	三层单播报文情况	支持	支持
非等值负载分担	二层单播报文情况	-	-
TCP/UDP 情况	TCP/UDP 情况	-	-
MPLS 情况	组播转发情况	-	-
VLL 情况	MPLS 情况	-	-

目的

部署负载分担有以下优点：

- 将流量分担到不同的链路上，可以充分利用网络资源。
- 将流量分担到多条链路上，可以增加链路总带宽。
- 将流量分担到多条链路上，当负载分担链路中的一条链路出现故障后，流量可以从其它链路继续转发，从而提高链路的可靠性。

受益

运营商受益

- 负载分担可以使数据流量分担到多条链路上，可以增加总带宽。可以使运营商在不更换设备的情况下，得到更大的带宽。
- 可以提高链路的可靠性，当一条链路出现故障后，如果存在负载分担链路，流量可以切到负载分担链路上，保证业务不中断。

用户受益

- 使用负载分担功能，可以向用户提供更大的接入带宽。
- 用户可以得到更可靠的服务，当一条链路出现故障后，用户流量可以切换到负载分担链路上，用户可以继续使用业务。

7.2 参考标准和协议

本特性的暂无相关标准和协议。

7.3 原理描述

7.3.1 负载分担的基本原理

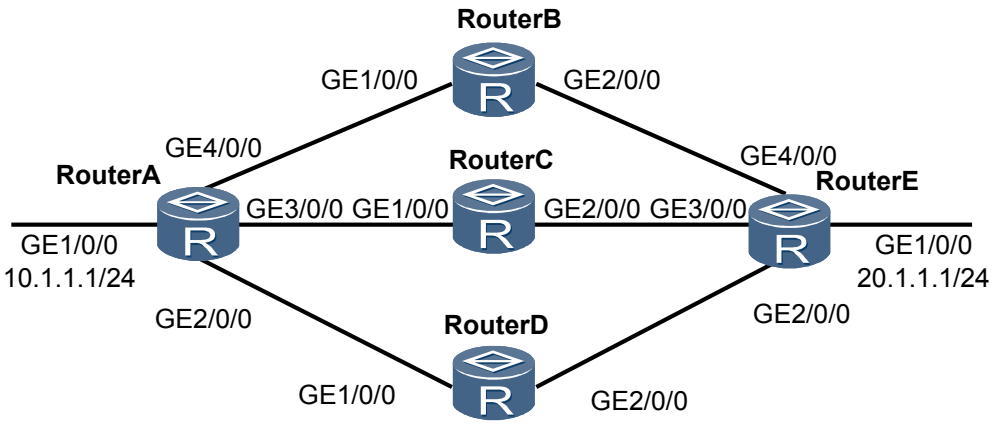
基于协议的负载分担

- 等值负载分担

  NE20E-X6 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。当没有到同一目的地的更高优先级路由时，这几条路由都被采纳，在转发去往该目的地报文时，依次通过各条路径发送，从而实现网络的负载分担。

  路由协议对负载分担的支持：对于同一目的地，特定的路由协议也可能会发现几条不同的路由，如果该路由协议在所有活跃的路由协议中优先级最高，那么这几条不同的路由都被看作当前有效的路由。这样，在路由协议层面上，保证了 IP 流量的负载分担。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 7-1 基于协议的负载分担示意图



  如图 7-1 所示，负载分担的组网环境如下（以 OSPF 为例）：

  - 在 RouterA、RouterB、RouterC、RouterD 和 RouterE 上配置 OSPF 路由协议，OSPF 会发现三条不同的路由。
  - 从 GE1/0/0 接口进入 RouterA 去往 RouterE 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。

- 非等值负载分担

  对于等值负载分担，流量是在这些负载分担路径上进行平分的，不会考虑链路带宽的差异问题。这种方式可能造成一些低带宽链路拥塞，而另一些高带宽链路空闲。如果能够按照出接口的带宽进行流量的负载分担，则能解决这个问题，这就是非等值负载分担。

  图 7-1 所示，如果在 RouterA 上使能了非等值负载分担，流量就会按照 RouterA 上的三个出接口的带宽比例进行负载分担。例如：如果三个出接口的带宽分别为 0.5G、1G 和 2.5G，那么去往三个出接口的流量比例就为 1：2：5。

非等值负载分担的工作机制如下：

- 非等值负载分担和等值负载分担工作机制基本相同，但是非等值负载分担会把带宽信息带到 FIB，并根据带宽比例生成 NHP 表，从而实现按照带宽比例的流量负载分担。

- TCP/UDP 情况

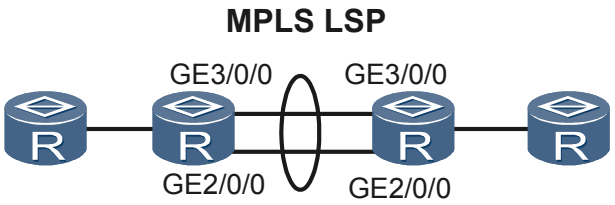
  对于 TCP/UDP 报文，支持根据 TCP/UDP 端口号进行 hash。

  TCP/UDP 情况负载分担的工作机制如下：

  - 对与 TCP/UDP 报文，支持将 TCP/UDP 的源、目的端口号也作为 hash 的 key 值进行负载分担。

- MPLS 负载分担

图 7-2 MPLS 负载分担示意图



如图 7-2：两台 NE20E-X6 之间有两条等价的 LSP，报文在这两个 LSP 上进行负载分担。

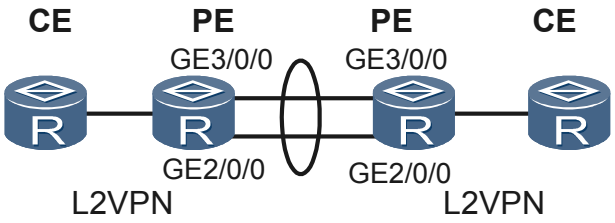
MPLS LSP 负载分担的工作机制如下：

在转发引擎查询负载分担表，然后分别把报文 Hash 到不同的负载分担项。

- VLL 情况

  当公网存在多条等价链路时，VLL 可以进行等值负载分担，充分利用网络资源。

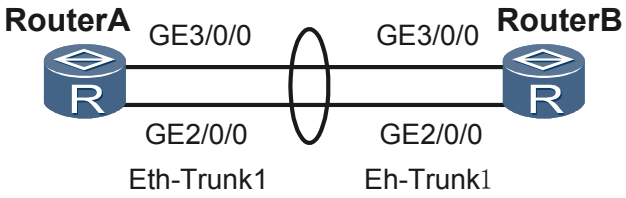
图 7-3 VLL 负载分担示意图



Trunk 负载分担

TRUNK 是将几个相同类型的物理端口捆绑在一起，作为一个逻辑接口使用。TRUNK 的好处是可以动态的增加带宽，提高连接的冗余度，并可将流量分担到各个链路上去。

图 7-4 TRUNK 负载分担示意图：



- 三层单播报文情况

  默认为逐流转发，可以配置为逐包，不支持逐 MAC。

  - 转发流程

    单播 IP 转发支持按照权重进行负载分担，当用户配置了成员接口的备份接口的时候，支持组外快速备份，收敛时间小于 50ms，当没有配置成员端口的备份口时，支持组内快速备份，同样要求收敛时间小于 50ms。

    IP 报文的上行流程如图 7-4 所示，如果是报文从捆绑端口出，查 FIB 后，其 TB 为某特殊值（这个特殊值用宏定义，目前定为 253），TP 对应相应的 Trunk ID，然后查 Trunk 转发表，使用 Hash 算法选出一个分担项，即得到真正的 TB 和 TP，然后进行转发。

- 二层单播报文情况

  默认为逐 MAC 转发，可以配置为逐 IP 和逐包转发。

  - 转发流程

    二层转发上行查找 MAC 表，而不是 FIB 表，其他流程同 IP 报文转发。

  - 负载分担方式

    - 逐 MAC 分担：SMAC（48bit）和 DMAC（48bit）进行 XOR 算法 Hash
    - 逐 IP 分担：同三层报文 IP 负载分担
    - 逐包分担：同三层报文逐包负载分担

- 组播转发

  - 转发流程

    先在 NP 中查 MFIB 表，以报文的 SIP、组播组地址和 VPN 实例作为 KEY 值，得到（S，G）表项，然后得到 MID。当出接口为 TRUNK 接口时，在 587 查 TRUNK 转发表，使用在 NP 中查到的 MID 的后四位在 TRUNK 的成员口上进行 Hash，确定出接口。

  - 负载分担方式

    组播对于 trunk 成员口而言，只支持逐流负载分担。以 SIP、组播组地址和 VPN 实例三者作为 key 值进行 Hash。

    对于一个相同的流，最终 Hash 到 trunk 的哪个成员口依赖于此流（即（S, G））申请 MID 的后四位。

- ?TCP/UDP 情况

  对于 TCP/UDP 报文，支持根据 TCP/UDP 端口号进行 hash。

  TCP/UDP 情况负载分担的工作机制如下：

  - 对与 TCP/UDP 报文，支持将 TCP/UDP 的源、目的端口号也作为 hash 的 key 值进行负载分担。

- MPLS 情况

- 转发流程

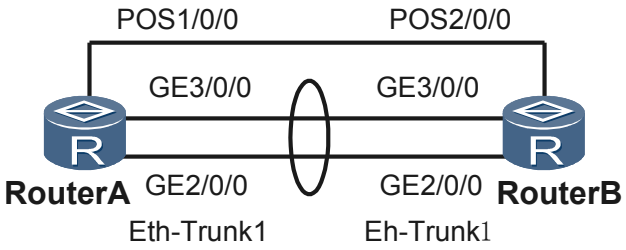
  在 NP 查 insegment 表后，得到出接口，如果该出接口为 TRUNK，然后查 Trunk 转发表，在 TRUNK 转发表上进行 Hash。

- 负载分担方式

二级 Hash

当多个下一跳中存在 TRUNK 链路时，流量在进行一次基于协议的负载分担 Hash 以后，还需要在 TRUNK 转发表上进行第二次 Hash，就存在两级 Hash，这个时候也能做正常的负载分担。

图 7-5 二级 Hash 场景示意图：



1. POS 链路和 TRUNK 链路先进行 Hash。
2. TRUNK 链路上的流量会在两个成员口上进行 Hash。

二级 Hash 的工作机制如下：

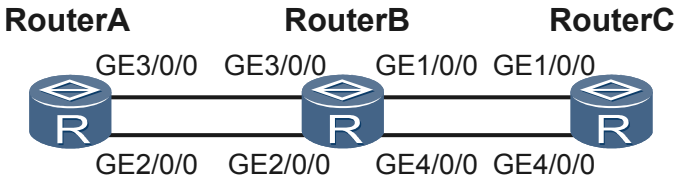
TRUNK 作为基于协议的负载分担的一条链路和其它链路进行第一级负载分担，第一级负载分担和基于协议的负载分担机制相同。TRUNK 上的流量在查完 NHP 表后，再查 TRUNK 转发表进行 Hash，完成第二级负载分担。

负载分担方式:

第一级 Hash 同基于协议的负载分担，第二级 Hash 同 Trunk 负载分担。

二级负载分担

图 7-6 二级负载分担场景示意图：



如图 7-6 所示，流量在 RouterA 到 RouterB 之间会进行一次负载分担，在 RouterB 到 RouterC 之间也会做一次负载分担。如果两次负载分担的算法相同，两次负载分担 Hash 结果就是相同的，同一条流总是 Hash 到同一条链路上，流量均衡效果不好。

二级负载分担的工作机制如下：

引入随机数作为 Hash 因子，这样在不同的设备上由于随机数不同，Hash 产生的结果也不同，使得流量可以在二级负载分担设备上流量均衡。

7.4 术语与缩略语

缩略语

缩略语	英文全称	中文全称
ECMP	Equal-cost multipath	等值负载分担
UCMP	Unequal-cost multipath	非等值负载分担
NHP	Nexthop	下一跳
RE	Prefix	前缀表
PST	Port state table	端口状态表
FIB	Forwarding Information Base	转发信息库

8 UCMP

关于本章

8.1 介绍

定义

非等值负载分担 UCMP（Unequal Cost Multiple Path），是指如果到达目的地有多条带宽不同但优先级相同的链路，则流量会根据带宽按比例分担到每条链路上。这样所有链路可根据带宽不同而分担不同比例的流量，使流量转发更合理。

目的

当等价路由有多个出接口，这些出接口中同时存在高速链路和低速链路，由于目前负载分担模式是等价负载分担 ECMP（Equal Cost Multiple Path），也就是说到目的地有多条负载分担路径可达时，流量不考虑链路带宽的差异，会在这些负载分担路径上进行平分。当不同路径带宽差异比较大时，可能出现低速路径流量阻塞的问题，不能有效的利用高速链路的带宽。为了使各链路带宽得到充分合理的利用，需要所有链路能够根据带宽不同分担不同比例的流量，实现根据带宽的比例进行流量的分配。

8.2 参考标准和协议

无

8.3 原理描述

8.3.1 UCMP 的基本原理

如果到达目的地有多个出接口，形成多条等价路由，底层硬件会根据各个接口的带宽比例申请资源，使经由这多个出接口输出的流量比例接近或等于各个接口的带宽比例。当某一接口的带宽变化时，流量能够自动按照变化后的带宽比例进行负载分担。

8.3.2 基于接口的 UCMP

支持基于接口的 UCMP 的接口类型

- Ethernet
- Serial
- GigabitEthernet

说明

说明：逻辑口不支持接口 UCMP。

使能基于接口的 UCMP

需要存在多个支持接口 UCMP 的物理出接口，并且要在每个接口下使能 UCMP。如果其中有一个接口不支持 UCMP，便忽略其他接口的带宽，流量在这些负载分担路径上平均分配。

为了使系统接口板能记录各个负载分担接口的带宽信息，要求用户 **shutdown** 和 **undo shutdown** 接口，从而触发路由管理重新下发路由给接口板，主控板 FIB 模块在下发路由时，检查路由的出接口，判读此接口是否已经使能 UCMP，将使能 UCMP 的接口的带宽记录到下发消息包中，接口板硬件根据负载分担的各个接口的带宽比，计算流量分配比例。

接口带宽变化的处理

带宽取自接口物理链路信息，不可更改。

注意事项

- 使能接口 UCMP 后，则不能再使能全局 UCMP；同理，全局 UCMP 使能后则无法再使能接口 UCMP。
- 给接口板下发的带宽精度为 Mbit/s，主要是支持高速链路。
- 由于接口使能 UCMP 之后要 **shutdown** 和 **undo shutdown** 接口，会使流量中断，促使了全局 UCMP 的产生，全局 UCMP 的功能比它更加完善。

8.3.3 全局 UCMP

支持全局 UCMP 的接口

- Ethernet
- Serial
- GigabitEthernet
- MP-Group
- ETH-TRUNK

使能全局 UCMP

需要存在多个支持全局 UCMP 的出接口。如果其中有一个出接口不支持 UCMP，全局使能 UCMP 后，也忽略其他接口带宽，流量依然会在这些负载分担路径上平均分配。

全局下使能 UCMP 后，使能所有支持 UCMP 的接口，触发整机路由下发，携带各个出接口的负载分担带宽给接口板，带宽携带方式同基于接口的 UCMP 方式，接口板根据负载分担的各个接口的带宽计算流量分配比例。

大容量路由情况下，为了避免频繁整机路由下发给系统带来压力，设置全局命令，命令使能/去使能抑制时间间隔为 5 分钟。

接口带宽变化的处理

- 当接口带宽变化，如逻辑接口成员加入退出或成员 UP/DOWN，逻辑接口相关模块如 TRUNK、MP 等需要通知接口管理模块新的接口带宽。接口管理模块在接口带宽变化后需要以事件方式通知 FIB 模块。FIB 模块在经过判断后将接口带宽变化的

消息通知到所有接口板，接口板获取消息后，重新计算流量比例。在接口带宽变化时，如果引起了路由计算，则会重新下发 FIB。

- 由于接口带宽变化的处理需要花费一定时间，为了防止接口带宽频繁变化导致 CPU 繁忙，对接口带宽变化通知的下发进行时间限制，设置下发生效时间默认为 5 秒。接口带宽变化，5 秒后把带宽变化通知下发接口板。在这 5 秒内如果接口带宽继续发生变化，最终会把最新的带宽变化通知下发接口板。此下发生效时间可以进行设置。

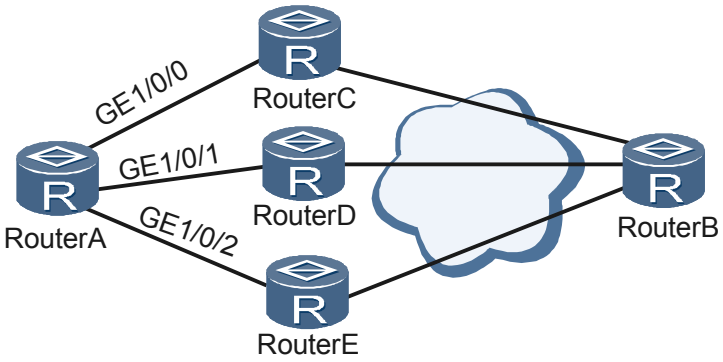
注意事项

- 使能全局 UCMP 后，则不能再使能接口非等值负载分担，同理，接口 UCMP 使能后则无法再使能全局 UCMP。
- 由于涉及低速接口，使用 Mbit/s 为单位不能满足所有情况，所以采用以 Kbit/s 为单位的带宽。

8.4 应用

8.4.1 基于接口的 UCMP 场景描述

图 8-1 基于接口的 UCMP 组网图



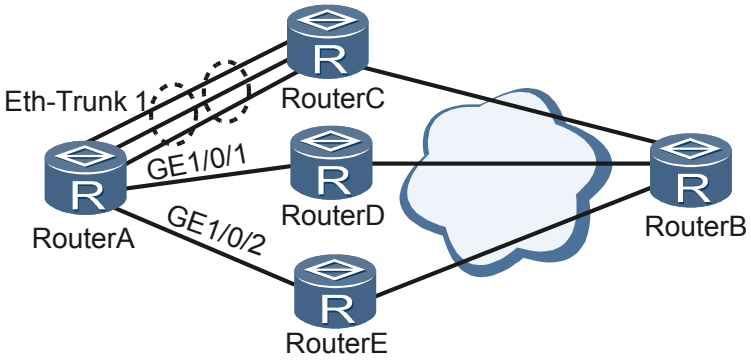
RouterA 上有 3 个物理出接口：GE1/0/0、GE1/0/1 和 GE1/0/2。GE1/0/0 带宽为 10G，GE1/0/1 为 1G，GE1/0/2 为 1G。RouterA 到 RouterB 形成三条 IPv4 等价路由。

在这个 3 个接口没有全部使能 UCMP 的情况下，流量是按 1:1:1 的比例进行分担的。

在每个接口下使能 UCMP 后，由 RouterA 到 RouterB 的流量在三个出接口上实现非等值逐流负载分担，经由 GE1/0/0、GE1/0/1、GE1/0/2 输出的流量比例接近或等于三个接口的带宽比例，即 10:1:1。

8.4.2 全局 UCMP 场景描述

图 8-2 基于全局 UCMP 组网图



RouterA 上有 3 个出接口：Eth-Trunk1、GE1/0/1 和 GE1/0/2，其中 Eth-Trunk1 为逻辑接口，包含 3 个 GE 口。Eth-Trunk1 带宽为 3G，GE1/0/1 为 1G，GE1/0/2 为 1G。RouterA 到 RouterB 形成三条 IPv4 等价路由。

在没有使能全局 UCMP 的情况下，流量是按 1:1:1 的比例进行分担的，不区分这 3 个接口之间的带宽比。

Eth-Trunk 1、GE1/0/1 和 GE1/0/2 都支持 UCMP，在使能全局 UCMP 之后，由 RouterA 到 RouterB 的流量在三个出接口上实现 UCMP 逐流负载分担，经由 Eth-Trunk1、GE1/0/1、GE1/0/2 输出的流量比例接近或等于三个接口的带宽比例，即 3:1:1。

当逻辑口 Eth-Trunk1 的一个成员口 shutdown 后，Eth-Trunk1 的带宽变为 2G，流量能够自动按照变化后的带宽比 2:1:1，进行负载分担。

支持 UCMP 的情况下，查看 FIB 表信息时，显示各个等价路由的相应带宽，通过计算接口带宽比，进而确认和实际流量负载分担比是否接近，从而判断功能是否正常。

8.5 术语与缩略语

缩略语

缩略语	英文全称	中文全称
ECMP	Equal Cost Multiple Path	等值负载分担
UCMP	Unequal Cost Multiple Path	非等值负载分担