# 第5章 网络互联技术

网络互联技术是计算机网络技术中的核心技术，其中涉及硬件、软件及算法等多方面的知识。本章将从介绍网络互联的基本概念入手，逐步深入地探讨IP协议族、路由协议、多播技术、专用网络互联技术等。随着互联网的不断发展，IPv6技术已经成为一种必然的趋势，移动IP技术也受到了越来越广泛的重视，本章也将对这两种技术进行探究。

## 5.1 网络互联的基本概念

不同的数据通信网络在很多方面都有所不同，存在很大的差异性。这种差异性称为异构性(Heterogeneity)，主要表现在：①不同的网络类型(如广域网、城域网、局域网)；②不同的数据链路层协议(Ethernet、Token Bus、ATM、WLAN)；③不同的计算机系统及操作系统平台。

为了隐藏所有底层网络细节的不同，实现异构网络中任意两台计算机之间的通信，可以利用网关或路由器将两个或两个以上不同的网络相互连接起来构成一个更大的“网络”，这种互联方式称为网络互联(Internetwork)，互联后的网络称为互联网(internet)，如图5-1所示。当使用TCP/IP协议进行互联之后，将在网络层上形成一个单一的虚拟网络。网络互联的目标是建立一个支持通用通信服务的统一、协调的互联网络，处于该网络中的主机之间通信就像是在单个网络中进行的一样。

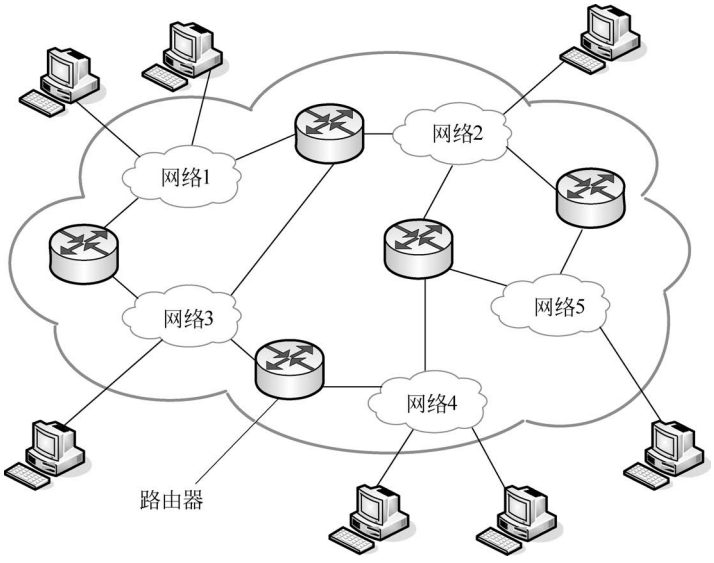



图5-1 网络互联示意图

需要注意，人们常说的因特网（Internet）和互联网（internet）是有区别的。因特网是指使用 TCP/IP 协议相互连接起来的全世界范围的一个互联网，即全球互联网，也可以说因特网是世界上最大的互联网（Internet is the largest internet）。

为了实现多个网络之间的互联，首先应该确保两个网络在物理上的互通，这是网络互联的前提，如果两个网络在物理上没有相互连接在一起，是没有办法实现网络间互联的。但是，仅有物理连接还远远不够，还需要使用一些特殊的计算机，这些计算机分别与两个网络连接在一起，它们可以将分组从一个网络传递到另一个网络。将这样起着特殊功能作用的计算机称为互联网网关（Internet Gateway）或者互联网路由器（Internet Router）。

应该指出，网关和路由器本来是指不同的网络互联设备。网关是指在网络层以上（主要是应用层）使用的互联设备，也称为高层协议转发器，它能在不同的高层协议间提供协议转换和数据重分组功能，但由于比较复杂，目前使用得较少，本书不进行详细讨论。路由器是指在网络层使用的互联设备，其实就是一台特殊的计算机，用来在互联网中进行路由选择。由于历史的原因，有很多文献都曾经将网络层使用的路由器称为网关（因为真正的网关很少用到），这样网关就成了路由器的代名词，二者可以互换使用。但是，从当前的使用情况来看，人们还是比较认可将网关和路由器区别对待，所以本书将严格区分使用这两个名词。

在物理层使用的连接设备中继器（Repeater）和在数据链路层使用的连接设备网桥并不属于网络互联设备。当使用中继器时，只是在物理层将信号进行了复制、调整和放大，以此来延长网络的长度，并没有实现异种网络间的连接；当使用网桥时，只是将连接部分的局域网的范围扩大了，从网络层的角度来看，依然属于同一种网络，因此也不属于网络互联。由此可见，网络互联已经是网络层及其以上各层的范畴了。能够实现网络互联的设备只有路由器及高层使用的网关，这一点需要大家特别注意。有一些资料将中继器、集线器、交换机（包括二层和三层）、网桥、路由器及网关统称为网络互联设备，本书并不赞成这样做。

## 5.2 网际协议（IPv4）

最基本的互联网服务被定义为不可靠的（Unreliable）、尽最大努力交付的（Best-effort delivery）、无连接的分组交付系统。这种服务不能保证交付（但这并不意味着可以任意丢弃分组），分组可能会丢失、重复或乱序，但服务检测不到这些情况，也不会通知发送方或接收方。实现这种交付功能的协议称为网际协议（Internet Protocol，IP）。当前使用较多的是该协议的第 4 个版本，通常用 IPv4 代表。IPv4 协议的研究与发展的过程如图 5-2 所示。

IP 协议定义了在整个 TCP/IP 互联网中使用的数据传送基本单元（IP 数据报）、IP 软件完成的转发（Forwarding）功能，同时还规定了主机和路由器应当如何处理分组、何时及如何产生差错报文，以及在什么情况下可以丢弃分组等。图 5-3 说明了 TCP/IP 协议的层次关系（包括网络层、传输层和应用层），从图 5-3 中可以看出，IP 协议是互联网中最基本的组成部分。

本节将讨论与 IP 协议相关的内容，包括 IP 地址、地址解析协议 ARP、IPv4 数据报、ICMPv4 协议及整个的 IP 数据报的转发过程。IP 协议不是独立工作的，它常常会与 ARP

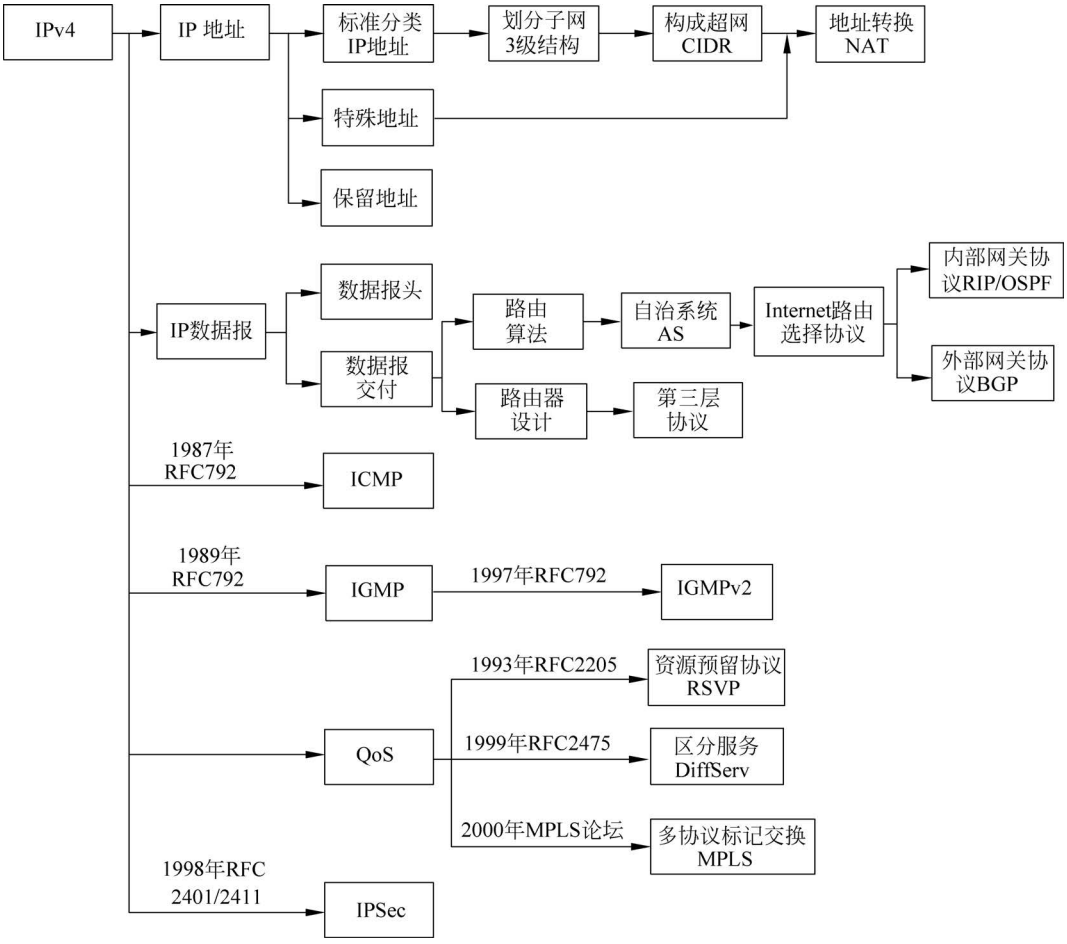



图 5-2　IPv4 协议的研究与发展的过程

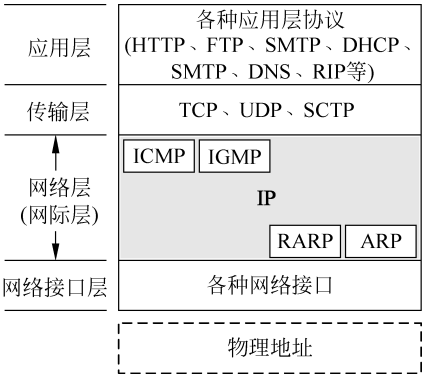



图 5-3　TCP/IP 协议的层次关系

和 ICMP 协议配套使用,相互之间构成了一个完整的系统。另外,在介绍 IP 协议相关的内容时,将不加区分地使用 IP 数据报、IP 包、IP 分组 3 个概念。

## 5.2.1 IP 地址


在互联网当中有数以亿计的主机和路由器，为了能够实现彼此间的无障碍通信，需要有一种方法来标识这些设备。在 TCP/IP 协议栈中，规定为每台设备分配一个统一格式、全球唯一的地址，这个地址称为 IP 地址（有时也可以称为互联网协议地址、互联网地址、逻辑地址、网络层地址）。IP 地址是 IP 协议使用的标识，工作于网络层。一台设备至少拥有一个 IP 地址（多则不限），任何两台设备的 IP 地址不允许相同。IP 地址由 ICANN 负责分配，实际上 ICANN 只负责将地址分配给国家或者地区，具体的分配工作由相应的国家或者地区的网络管理机构负责完成。

为了确保地址的网络部分在 Internet 上的唯一性，所有的地址都由一个中央管理机构进行分配。最初由因特网赋号管理局（Internet Assigned Number Authority，IANA）负责分配并制定相应的政策。从因特网诞生到 1998 年秋天，一直由 Jon Postel 一个人负责 IANA 的运转及地址分配。1998 年年底，在 Jon Postel 去世以后，组建了一个新的组织来承接 IANA 的工作，这个组织就是因特网名称与号码指派协会（Internet Corporation for Assigned Names and Numbers，ICANN）。ICANN 负责制定政策、分配地址等工作。

IP 地址由 32 位二进制数构成，通常有 3 种常用的表示方法：二进制记法、点分十进制记法和十六进制记法。

（1）二进制记法。IP 地址表示为 32 位，为了有更好的可读性，可以在每字节（8 位）之间加上一个（或多个）空格，将 32 位的 IP 地址分隔成 4 个 8 位组，如 01110001 10100101 00001110 10111100。

（2）点分十进制记法。是一种更为常用、更加简洁的 IP 地址表示方法。这种方法将 32 位的二进制 IP 地址用小数点分隔开，分隔的时候以 8 位（字节）为单位进行。由于每字节仅有 8 位，因此在点分十进制记法中的每个数目一定为 0～255，如 192.168.0.1。将二进制记法表示的 IP 地址转换为点分十进制记法表示的 IP 地址的方法是：把二进制记法中的每一组 8 位转换成等效的十进制数，并增加分隔用的小数点，如将 11100111 11011011 10001011 01101111 转换成点分十进制记法表示的 IP 地址为 231.219.139.111。

（3）十六进制记法。每一个十六进制数字等效于 4 位二进制数字，这样一来，一个 32 位的 IP 地址可以表示为 8 个十六进制数字，这种记法常用于网络编程中。需要说明的是，十六进制记法中通常不需要加入空格或小数点进行分隔，但在开始处可以加入 0X（或 0x），或者在最后加入下标 16 以表示这个数是十六进制的，如 0X910C1B26 或者 $910C1B26_{16}$。

Internet 早期使用的是分类编址的方式；在 20 世纪 90 年代中期，出现了一种称为无分类编址的方式，这已经是当前 Internet 中主流的编址机制。所以，本节将结合当前的应用情况，分别展开讨论。

**1. 分类编址**

IP 地址由网络号和主机号两部分组成，这样的 IP 地址是两级地址结构，如图 5-4 所示。

IP 地址的层次结构便于在 Internet 上实现寻址。可以先按 IP 地址的网络号（Net-ID）找到相应的网络，进而按主机号（Host-ID）找到对应的主机。所以，IP 地址这样的层次结构有利于快速准确地定位主机。反之，像 MAC 地址这样的平面地址结构是不利于寻址的。

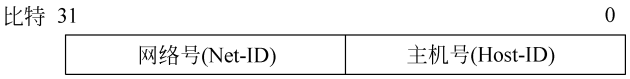



图 5-4　IP 地址结构

在分类编址中,IP 地址按最高 1～5 位的值分成 5 类：A、B、C、D 和 E 类,如图 5-5 所示。事实上,大量使用的 IP 地址是 A、B、C 三类。其中,网络号部分是由相关机构分配的。当一个单位申请到一个 IP 地址时,实际上只是获得了一个网络号(Net-ID),具体的主机号(Host-ID)由本单位内部自行进行分配。


图 5-5　分类编址的 IP 地址

(1) A 类地址：8 位网络号,网络号的第一位为 0,其余 7 位可以分配,但可以指派的网络号是 $2^7-2=126$ 个,之所以减去了两个网络号,是由于 IP 地址中的全 0 表示“此网络(this)”或者“本网络”,是一个保留地址,另外,网络号为 127(01111111)的地址保留作为本地软件回环测试地址。

A 类网络具有 24 位主机号,可以容纳的最大主机数是 $2^{24}-2=16\ 777\ 214$ 个,这里减去两个主机号的原因是主机位全 0 的地址是本主机连接到的网络地址；而主机位全 1 的地址则表示该网络上的所有主机(即本网络的广播地址)。A 类网络可以容纳的主机数较多,一般用于大型网络。

(2) B 类地址：16 位网络号,网络号的前两位为 10,其余 14 位可以分配。由于 B 类地址中前两位已经是固定的了(10),因此网络位后面的 14 位无论怎样变化都不可能使整个 2B 的网络号字段成为全 0 或全 1,因此不存在网络总数减 2 的问题。但实际上 B 类网络地址中 128.0.0.0 是不指派的,而可以指派的 B 类最小网络地址是 128.1.0.0,因此 B 类地址可指派的网络数为 $2^{14}-1=16\ 383$ 个。

B 类网络具有 16 位主机号,因此每个网络上的最大主机数是 $2^{16}-2=65\ 534$ 个,减去的两个地址为主机位全 0 的地址(网络地址)和主机位全 1 的地址(本网络的广播地址)。B 类地址一般用于大、中型网络。

(3) C 类地址：24 位网络号,网络号的前三位为 110,其余 21 位可以分配,8 位主机号。C 类网络地址中的 192.0.0.0 也是不能指派的,可以指派的 C 类最小网络地址是 192.0.1.0,因此 C 类网络可以指派的网络地址总数是 $2^{21}-1=2\ 097\ 151$ 个。

每一个 C 类网络中可以容纳的主机数是 $2^8-2=254$ 个,同样减去的两个地址为主机位

全 0 的地址(网络地址)和主机位全 1 的地址(本网络的广播地址)。C 类网络可以容纳的主机数较少,一般用于小型网络。

(4) D 类地址:前四位为 1110,用于多播。

(5) E 类地址:前五位为 11110,保留未使用。

IP 地址中,全 0、全 1 的地址一般不能当作普通地址使用。各类地址的可指派范围如表 5-1 所示。

**表 5-1 各类 IP 地址的可指派范围**

| 类别 | 范围 | 可用首网络号 | 可用末网络号 | 最大可指派网络数 | 网络主机数 |
|---|---|---|---|---|---|
| A | 0.0.0.0～127.255.255.255 | 1 | 126 | 126($2^7-2$)① | 16 777 214 |
| B | 128.0.0.0～191.255.255.255 | 128.1 | 191.255 | 16 383($2^{14}-1$)② | 65 534 |
| C | 192.0.0.0～223.255.255.255 | 192.0.1 | 223.255.255 | 2 097 151($2^{21}-1$)③ | 254 |
| D | 224.0.0.0～239.255.255.255 | | | | |
| E | 240.0.0.0～255.255.255.254 | | | | |

**注**:① 如果除去 A 类地址中的 1 个私有网络地址 10.0.0.0(该地址本来是分配给 ARPANET 的,由于 ARPANET 已经关闭停止运行了,因此这个地址就用作专用地址了),实际可指派的地址变为 125 个。

② 如果除去 B 类地址中的 16 个私有网络地址,实际可指派的地址变为 16 367 个。

③ 如果除去 C 类地址中的 256 个私有网络地址,实际可指派的地址变为 2 096 895 个。

另外,IP 定义了一套特殊地址(有时也称为保留地址)。这些特殊地址有着特殊的用途,如表 5-2 所示。

**表 5-2 特殊 IP 地址**

| Net-ID | Host-ID | 源地址 | 目的地址 | 含义 |
|---|---|---|---|---|
| 0 | 0 | 可以 | 不可以 | 本网段上的本主机 |
| 0 | ×× | 不可以 | 可以 | 本网段上的某主机 |
| 全 1 | 全 1 | 不可以 | 可以 | 本网内广播 |
| ×× | 全 1 | 不可以 | 可以 | 对目的的网络广播 |
| 127 | ×× | 可以 | 可以 | Loopback 测试 |
| 169.254 | ××.×× | | | DHCP 因故障分配的地址 |
| 10 | ××.××.×× | | | 私有地址,用于内部网络 |
| 172.16～172.31 | ××.×× | | | |
| 192.168.×× | ×× | | | |

分类的 IP 地址有一些缺点,如 IP 地址的空间利用率不高,数以百万计的 A 类地址、大量的 B 类地址都被浪费了;与此同时,C 类地址空间对于大多数机构而言是不够用的。另外,给每一个物理网络都分配一个网络号会导致路由表变得太大而无法正常工作。基于这样的原因,希望能够有一种更为灵活地使用 IP 地址的方式。

为此,可以考虑将一个规模较大的网络划分成为相对独立的子网(Subnet),子网间的通信类似于不同网络之间的通信。划分子网后的网络,对外仍然表现为一个网络。本网络外的网络并不知道这个网络是由多少子网构成的。也就是说,划分子网完全是网络内部的事情,进行子网划分的优势是明显的。


视频讲解

(1) 减轻网络的拥塞状况：通过路由器的分隔，可以将原本属于同一个网络的主机分散到多个子网中，从而大大减少了每个子网内的通信量，同时分隔了广播域。

(2) 缩减了路由表的内容：划分子网以后，路由表只需要记录没有划分子网前的网络情况，而不需要将每一个子网的信息都记录在路由表中，这样就可以大大减少路由表中的记录数；如果不进行子网划分，路由表需要记录所有的网络情况，记录数巨大。

(3) 便于网络管理：网络范围减小后，排错更为容易，安全性相对更高，管理更加灵活。

划分子网的方法是：将主机号部分进一步分成子网号和新的主机号两部分。这样，IP地址就由三部分组成：网络号(Net-ID)、子网号(SubNet-ID)、主机号(Host-ID)，如图5-6所示。

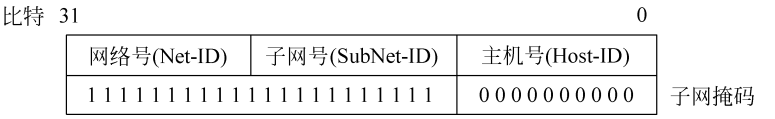



图5-6 三级IP地址结构及子网掩码

在子网划分后的IP地址中，子网号有几位？怎样判定哪些位是网络号，哪些位是子网号？这就需要引入子网掩码(Subnet Mask)。子网掩码可以区分网络号和子网号，有时也可以称为屏蔽码。子网掩码的格式和IP地址相同，对应网络号(Net-ID)和子网号(SubNet-ID)部分全部为1，对应主机号部分全部为0。子网掩码的表示形式同IP地址，如255.255.255.0。

划分子网后，同一个子网内的所有主机的网络地址、子网地址、子网掩码是相同的。在使用过程中，将网络地址和子网地址合并在一起，称为广义网络地址。广义网络地址的计算方式是：广义网络地址=(IP地址) AND (子网掩码)。

当前使用的Internet中规定，所有的网络都必须使用子网掩码，在路由器的路由表中也必须有子网掩码这一项内容。子网掩码已经成为一个网络不可或缺的属性。对于一些没有进行子网划分的分类IP地址来说，它们的子网掩码可以使用默认子网掩码，即与IP地址的网络号对应的内容置为1(没有子网号部分)，与IP地址的主机号对应的内容置为0。显然，A类IP地址的默认子网掩码是255.0.0.0，B类IP地址的默认子网掩码是255.255.0.0，C类IP地址的默认子网掩码是255.255.255.0。

**【例题】** 设计一个网络时，分配给其中一台主机的IP地址为192.55.12.120，子网掩码为255.255.255.240。

(1) 确定该主机的网络号、子网号和主机号。

(2) 确定该主机所在网络在该子网掩码下子网号的范围。

**【解答】**

(1) 主机的网络号和子网号可以通过主机的IP地址与子网掩码进行逻辑“与”运算得到，而IP地址剩余的部分即为主机号。

建议大家在进行逻辑运算时，首先将点分十进制记法表示的IP地址转换为二进制记法表示的IP地址。当熟练掌握以后，可以省略这一步骤。

192.55.12.120转换为二进制形式后为11000000 00110111 00001100 01111000；

255.255.255.240转换为二进制形式后为11111111 11111111 11111111 11110000；

广义网络地址=(192.55.12.120) AND (255.255.255.240)=192.55.12.112。

主机分配到的是一个 C 类 IP 地址，C 类地址的前 24 位均为网络号；由子网掩码可以知道，从原主机号中拿出了 4 位(25～28 位)充当子网号；剩余的位为主机号。

所以，网络号为 192.55.12；子网号为 112；主机号为 8。

(2) 由(1)的分析可知，IP 地址中的第 25 位到第 28 位为子网号。因此，子网号范围为 192.55.12.0，192.55.12.16，192.55.12.32，192.55.12.48，…，192.55.12.240。

在子网掩码的使用过程中，虽然没有明确规定子网掩码中的 1 要连续，但建议大家选用连续的 1，以避免不必要的麻烦。

在进行子网规划的过程中，会涉及选用几位主机号来充当子网号。选取的原则就是够用即可。例如，选取两位主机号充当子网号后，原则上就可以表示 4($2^2$)个不同的子网了。问题在于，子网号部分全 0 和全 1 的 IP 地址能否使用。不同的资料对此问题的回答不尽相同。准确的答案是，在分类 IP 地址中，子网号不能为全 0 或全 1；但随着无分类地址的普及应用，现在全 0 和全 1 的子网号也可以使用了。使用全 0 或全 1 子网号的 IP 地址时要谨慎，要清楚当前网络中的路由器和主机是否支持这一应用。传统的教材认为使用全 0 或全 1 子网号的做法是错误的，这里并不这样认为。当然，主机号部分全 0 或全 1 的 IP 地址是不能分配给主机使用的。

**2. 无分类编址**

分类编址方式有一些明显的缺陷。

(1) 已有的 IP 地址已经快分配完毕。A 类地址早已分配完，B 类地址也将近分配完毕。

(2) 因特网主干网上的路由表中的项目数急剧增长。

(3) 整个 IPv4 的地址空间将被全部耗尽。

一些不同的策略已经开始实施，以应对地址空间不足的问题。1987 年，RFC1009 指明在一个划分子网的网络中可以同时使用几个不同的子网掩码。使用变长子网掩码(Variable Length Subnet Mask，VLSM)可以进一步提高 IP 地址资源的利用率。

在 VLSM 的基础上又进一步研究出无分类编址方法，其正式名称为无分类域间路由选择(Classless Inter-Domain Routing，CIDR)。目前 CIDR 已经成为 Internet 的建议标准，得到了广泛的应用。

CIDR 消除了传统的 A、B、C 类地址及划分子网的概念，可以更加有效地使用 IPv4 的地址空间。CIDR 把 32 位的 IP 地址划分成为两部分，即网络前缀(Network-Prefix)和主机号(Host-ID)。网络前缀用来标明网络，主机号用来指明主机。可以看出，CIDR 使 IP 地址从三级编址(使用子网掩码)又回到了两级编址，但这是无分类的两级。

CIDR 使用“斜线记法”表示，也称为 CIDR 记法，即在 IP 地址后面加上斜线“/”，然后写上网络前缀所占的位数，如 201.18.5.0/19。

CIDR 把网络前缀相同的连续的 IP 地址组成的地址区间称为 CIDR 地址块。根据 CIDR 地址的表示方法，可以比较容易地知道一个 CIDR 地址块中的最小地址和最大地址，以及地址块当中的地址数。其方法是，首先计算出主机地址的位数 $n$(32－网络前缀位数)，再将 IP 地址中后 $n$ 位分别置为全 1 和全 0，就得到了该地址块的最大地址和最小地址。

例如，128.14.35.7/20 这个 CIDR 形式表示的 IP 地址，主机地址为 12 位，即第三个字节中的前 4 位为网络前缀，后 4 位为主机地址，依据上述方法计算可知，该 CIDR 地址块的

最小地址为128.14.32.0,最大地址为128.14.47.255。

由此可以看出,CIDR方法分配IP地址比较灵活,可以依据实际需要进行分配。不会出现分类地址中那样的浪费情况。CIDR可以分配多个传统网络规模的地址块,也可以分配$1/n$个传统网络规模的地址块,这样的方式称为"超网"或"地址聚合"。通过地址聚合,可以有效地减少路由表中的记录数。

需要说明的是,"CIDR不使用子网"并不意味着单位内部不能划分子网。CIDR不使用子网是指在32位IP地址中没有指明若干位用于子网号标识。但分配到一个CIDR地址块的单位,仍然可以在本单位内部依据实际需要划分子网,当然,相应的网络前缀部分的长度也会发生变化。

## 5.2.2 地址解析协议

### 1. ARP的提出

在互联网中通信时,网络层与数据链路层使用的是不同的地址。数据链路层使用的是物理地址(48位的MAC地址),而在网络层使用的是逻辑地址(32位的IP地址)。如图5-7所示,当应用层的数据在传输层经过处理变成TCP报文后,向下交给网络层。网络层再对报文加上含有IP地址的首部后就交给了数据链路层。在数据链路层,IP数据报会被封装为MAC帧,而MAC帧是通过首部含有的硬件地址实现寻址的。在数据链路的另一端的设备是根据数据帧首部的硬件地址来接收MAC帧的,只有在数据链路层向上交付给网络层后,IP地址才在IP数据报中被找出来进行应用。从图5-7中可以看出,IP地址放在IP数据报的首部,而硬件地址则放在MAC帧的首部。

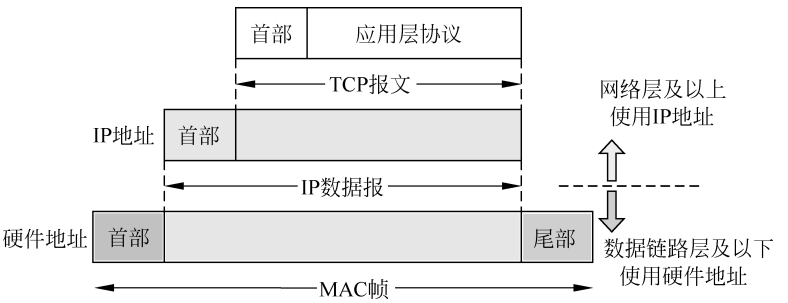



图5-7 IP地址和MAC地址的封装关系

当然有人会问,为什么不在网络层和数据链路层使用同一个地址,而在不同的层要使用不同的地址呢?因为IP地址在互联网内是唯一的主机标识,而硬件地址在全世界也是独一无二的。实际上,最终也是依靠硬件地址找到目标主机的。那么,为什么不直接使用硬件地址进行寻址,而是要使用抽象的IP地址并调用ARP来寻找出相应的硬件地址呢?似乎看起来只使用硬件地址是可以的,其实不然。

由于IP地址是分层次的结构,而MAC地址是一种平面的结构。就如主机A要和主机B进行通信时,核心网络上的路由设备必须清楚地知道全世界所有主机及网络设备是连接在主干网的哪一个端口上。由于MAC地址的这种平面结构不能提供任何信息给路由设备,这些路由设备就必须存储世界上所有主机的MAC地址。如果这样做,就会使得路由器承担极其繁重的工作任务,而且网络性能也是相当糟糕。就好比你要寄一封信给一个人,不

写具体的通信地址，只写了收信人的姓名便交给了邮局。可以想象，邮局必须得知道全世界所有人的个人信息，并且还得假设这个世界上所有的人都有不同的名字，才能保证将信送到正确的收信人手中，显然这是不现实的。而 IP 地址的分层结构可以很好地解决这样的问题，就像平时正常寄信时的情况一样，会以层次性的结构方式写上收信人在世界的具体位置，以方便邮递员投递。

另一个问题是网络异构问题。互联网是由各种网络设备（路由器、网关等）将很多异构型网络相互连接而组成的，这些各式各样的网络可能出自不同的组织，运行不同的协议，使用不同的物理地址（这些网络可能是 Ethernet、令牌环网、令牌总线网、ATM 或其他类型网络）。当两个主机之间进行通信时，它们的分组从源主机到目的主机将有可能经过各种各样的异构网络。要使这些异构网络能够互相通信就必须进行非常复杂的硬件地址转换工作，这些几乎是不可能做到的事情。因此，需要对这种差异性进行屏蔽，使之具有透明的特性；如果在上层统一使用 IP 地址就可以解决这个问题。连接到 Internet 的主机都拥有统一的网络层地址，方便了相互之间的通信。

由此可以看出，逻辑地址和物理地址分别有各自的用途。逻辑地址由网络层使用，而物理地址则在数据链路层使用。这就要求在通信过程中，有一种方法可以实现相互之间的映射，使这两种地址可以对应起来。地址解析协议（Address Resolution Protocol，ARP）正是在这样的需求下产生的。

**2. ARP 的报文结构**

ARP 报文的格式如图 5-8 所示。


图 5-8　ARP 报文的格式

以太网帧首部的前两个字段是目的硬件地址和源硬件地址，当要在网络中进行数据帧的广播时，可将目的地址全部置为 1；帧类型字段在分组中占 2B，用来说明后面数据的类型，ARP 请求/应答分组所对应的该字段的值为 0x0806；接下来是 ARP 报文部分，其中硬件类型字段指明了发送方想知道的硬件接口类型，以太网的值为 1，在分组中占 2B；协议类型字段则指明了发送方提供的高层协议类型，IP 协议为 0x0800，占 2B，它的值与包含 IP 数据报的以太网数据帧中的类型字段的值是相同的；硬件地址长度和协议地址长度是用来指明硬件地址和高层协议地址的长度，它们分别各占 1B，这样做的目的是让 ARP 报文可以在任意硬件和任意协议的网络中使用，对于 Ethernet 上的 IP 地址的 ARP 请求/应答分组来说，它们的值分别是 6 和 4；操作字段是指操作类型字段，用来表示这个报文的类型，因为 ARP 请求和 ARP 应答报文的帧类型字段值是一样的，所以必须靠这个字段来进行区别，一

般应用的4种操作类型有ARP请求、ARP响应、RARP请求、RARP响应,对应的字段值分别为1、2、3、4;最后的4个字段是发送方的硬件地址、发送方的协议地址、目的方的硬件地址、目的方的协议地址。

**3. ARP的工作过程**

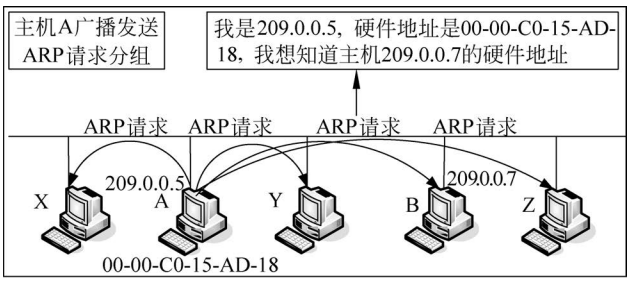
ARP是一种动态地址解析协议,因为动态的方法有着很好的扩展性与灵活性。假设在一个只有A、B两台主机组成的小型局域网中,A与B之间进行通信时,可以考虑使用静态映射的方法,使A与B的IP地址与MAC地址相对应,形成IP地址与MAC地址对应表。毫无疑问,这样做可以保证A与B之间的正常通信。但当一台新的主机C加入这个网络中时,这张对应表中则没有主机C的信息,C并不能与A、B之间实现通信。另外,若A或B其中的一台主机因为某些原因更换了网卡,则它们的MAC地址会发生相应变化,A与B之间此时也不能实现正常通信,因为IP地址与MAC地址对应表中记录的还是没有更换网卡前的MAC地址。同样,当主机A或主机B从当前网络移动到了一个新的网络,在MAC地址没有发生变化的前提下,IP地址变了,造成IP地址与MAC地址对应表出现错误。由此可见,单纯的静态方法是行不通的,ARP需要采用动态方法实现通信。

为了提高工作效率,ARP一般会采用静态映射与动态映射相结合的方法,这样可以实现一次请求、多次使用的良好效果。实现"动静结合"的关键在于每台主机都建立了一个ARP高速缓存表(ARP Cache),里面存储了本地局域网上一些主机和路由器的IP地址与MAC地址的关系映射,且这些关系映射是随时间动态更新的。当主机A要向本局域网上的主机B发送IP数据报时,首先查找自己的ARP高速缓存表,如果找到对应的记录,则取出MAC地址写入相应的MAC帧;如果找不到,便启用ARP服务进行地址解析。

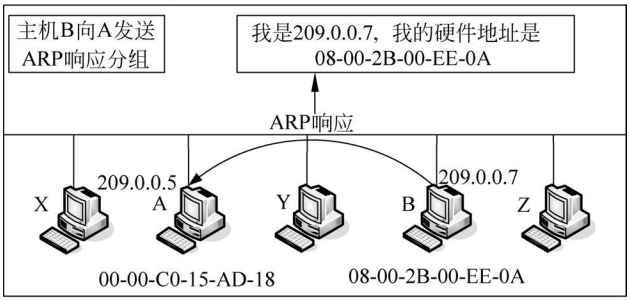
如图5-9所示,实现地址解析的第一步是产生ARP请求帧,在ARP请求帧的相应字段写入本地物理地址、IP地址、待侦测的目的IP地址,在目的物理地址字段填入0,并在操作类型字段填入1,用来表示本数据帧是一个ARP请求帧。该ARP请求帧以本地物理地址作为源地址,以物理广播地址(FF-FF-FF-FF-FF-FF)作为目的地址,在本局域网内进行广播。

本局域网当中的所有主机都会接收到该ARP请求帧,除目的主机外,所有接收到该ARP请求帧的主机和设备都会丢弃该ARP请求帧,因为目的主机能够识别ARP消息中的IP地址与自己的IP地址是不是相同。目的主机需要构造ARP应答帧以回应ARP请求。在ARP应答帧中,以请求分组中源物理地址、源IP地址作为其目的物理地址、目的IP地址,并将自身的物理地址、IP地址填入应答帧的源物理地址、源IP地址字段,并在操作字段中写入2,表示本ARP数据帧是一个应答数据帧。源主机接收到ARP应答帧后,获得目的主机的物理地址,并将它作为一条新记录加入ARP高速缓存表。由此,ARP高速缓存表中的内容便可以不断地添加及更新,当以后要有信息发送到同一主机时便会在ARP高速缓存表中找到相应的记录,不再需要进行ARP请求。这样可以减少网络流量,提高处理效率。如果不使用ARP高速缓存,那么任何一个主机只要进行一次通信,就必须在网络上用广播的方式发送ARP请求分组,这会使网络上的通信量大大增加。

此外,也需要考虑ARP高速缓存表中内容的正确性,ARP高速缓存表应当进行实时更新。所以,为每一条记录都设置了一个计时器,每一条记录在高速缓存中的生存时间一般为10～20min,起始时间从被创建时算起,超过生存时间的记录会从高速缓存中删除。




(a) 主机广播发送ARP请求分组




(b) 主机B向A发送ARP响应分组

图 5-9　ARP 地址解析的过程

由于 ARP 协议工作机制的缺陷，如果源主机没有发送 ARP 请求而收到其他主机的 ARP 响应数据帧，源主机也会在本地 ARP 高速缓存表中存储该主机物理地址和 IP 地址的对应关系。这样设计的初衷是减少网络通信量，提高传输效率。但是，这却成为 ARP 攻击的主要原因。由于不加验证就存储相应的记录，使得黑客可以很轻易地构造一个假的物理地址和 IP 地址的对应关系，造成用户没有和真正想通信的主机进行通信，而是在和一台傀儡主机进行通信，从而导致机密信息的泄露。

## 5.2.3　IPv4 数据报


### 1. IPv4 数据报格式

IPv4 数据报的格式如图 5-10 所示，一个 IPv4 数据报由首部和数据两部分组成。首部的前一部分长度固定(20B)，是所有 IP 数据报必须具有的。在首部的固定部分后面是一些可选字段，其长度是可变的。下面讨论首部各字段的意义。

(1) 版本：占 4 位，指 IP 协议的版本，IPv4 协议版本号为 4。

(2) 首部长度：占 4 位，可表示的最大数值是 15 个单位(一个单位为 4B)，因此，IP 的首部长度的最大值是 60B。当 IP 分组的首部长度不是 4B 的整数倍时，必须利用最后的填充字段加以填充。

(3) 区分服务：占 8 位，用来表示不同的服务质量。这个字段在旧标准中称为服务类型，但实际上一直没有被使用过。1998 年 IETF 把这个字段改名为区分服务。

(4) 总长度：占 16 位，指首部和数据之和的长度，单位为字节。因此，数据报的最大长度为 65 535B(64KB)。在网络层下面的每一种数据链路层都有其自己的帧格式，其中包括帧格式中的数据字段的最大长度，这称为最大传送单元 MTU。当一个 IP 数据报封装成数据链路层的帧时，此数据报的总长度(即首部加上数据部分)一定不能超过下面的数据链路

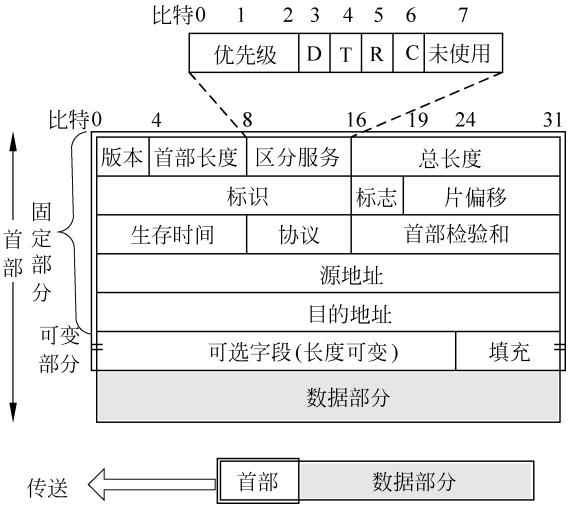



图 5-10 IPv4 数据报的格式

层的 MTU 值。

(5) 标识：占 16 位，可以看作 IP 数据报的编号。仅仅只是一个编号，而不是序号，因为 IP 数据报是无序的。

(6) 标志：占 3 位，目前只有两位有定义。标志字段中的最低位记为 MF，MF＝1 表示后面“还有分片”的数据报；MF＝0 表示这已是若干数据报分片中的最后一个。标志字段中间的一位记为 DF＝1，意思是“不能分片”，只有当 DF＝0 时才允许分片。

(7) 片偏移：占 13 位，表示本分片在原分组中的相对位置。片偏移以 8B 为偏移单位。

(8) 生存时间：占 8 位，记为 TTL，即数据报在网络中的寿命。初始值设置为允许经过的跳数，一般根据发送端的操作系统类型不同而具有不同的值，常见的值为 32、64、128 和 255，意味着一个数据报在网络中最多允许经过多少段链路。每经过一个路由器，将该值减 1，当该值减到 0 时就将该数据报丢弃，这样可以防止数据报在因特网中“兜圈子”，消耗大量网络资源。

(9) 协议：占 8 位，协议字段指出此数据报携带的数据使用哪种协议，以便目的主机的网络层知道应将数据部分上交给哪个处理进程。具体的对应关系有 ICMP 1、IGMP 2、TCP 6、EGP 8、IGP 9、UDP 17、IPv6 41、OSPF 89。

(10) 首部检验和：占 16 位，这个字段只检验数据报的首部，但不包括数据部分。这是因为数据报每经过一个路由器，路由器都要重新计算一下首部检验和(一些字段，如生存时间、标志、片偏移等都可能发生变化)，如果把数据部分一起检验，计算的工作量就太大了。其生成方法是 16 位为一个字，按字进行补码加法，再将和取反。

(11) 源地址：占 4B，源主机的 IP 地址。

(12) 目的地址：占 4B，目的主机的 IP 地址。

(13) 可选字段：用来支持排错、测量及安全等措施，此字段长度可变，从 1B 到 40B 不等，取决于所选择的项目。实际上这些选项很少被使用。

(14) 填充：任意数据，使头部的总长度为 32 位的整数倍。

(15) 数据部分：具体需要传输的数据。

**2. 数据报的封装与分片**

IP 数据报需要进行分片有以下两个原因。

(1) 不同的网络 MTU 值不同,如以太网的 MTU 为 1500B,PPP 的 MTU 为 296B,FDDI 的 MTU 为 4352B,令牌环的 MTU 为 4464B。

(2) 太长的 IP 数据报不能够封装到较短的数据链路层帧中进行传送。

所以,需要对 IP 数据报进行分片。如果 IP 数据报进行了分片,IP 首部中的"总长度"字段是指分片后的每片的首部长度与数据长度的总和,而不是未分片前的数据报长度。图 5-11 给出了一个数据部分长度为 3800 字节的 IP 数据报的分片与封装的情况。

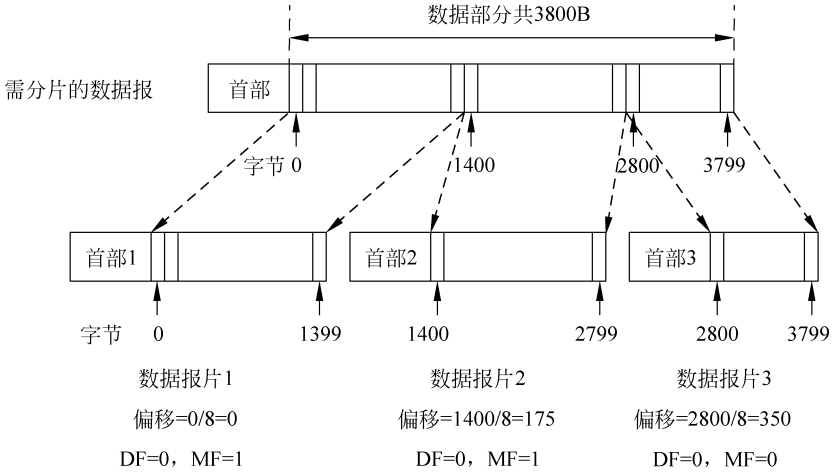



图 5-11 IP 数据报的分片与封装

## 5.2.4 ICMP 协议

在当今如此复杂的网络环境中,可能会有各种各样影响数据报传输的问题出现。例如,通信线路可能会断、处理器可能出现故障、路由器可能负载太高、目的主机可能出现临时或永久的断链、计时器出现超时等情况,这些状况的发生都无法确保 IP 能够进行正常通信。由于 IP 协议提供的是一种无连接的、不可靠、尽力而为的服务,它并不会关心网络服务是否处于正常状态,因此,需要设计某种机制来侦测或通知各种各样可能发生的状况,包括路由、拥塞和服务质量等问题。利用这种机制来帮助人们对网络的状态有一些了解。Internet 控制报文协议(Internet Control Message Protocol,ICMP)由此提出。

ICMP 协议是 TCP/IP 协议族中的一个子协议,与 IP 协议同属于网络层,但是它不能够独立于 IP 协议而存在。ICMP 主要是通过差错报告与查询、控制机制来保证 IP 协议的可靠运行。ICMP 不仅是一个管理性协议,并且也是一个 IP 信息服务的提供者,它的报文是被封装在 IP 数据报中进行传送的,因而也不保证可靠地提交。ICMP 报文的格式如图 5-12 所示。

ICMP 报文各个字段的具体含义如下。

(1) 类型字段。类型字段表示 ICMP 报文的类型。

(2) 代码字段。代码字段表示报文的少量参数。

(3) 校验和字段。校验和字段用于进行 ICMP 报文的校验,覆盖整个 ICMP 报文。使

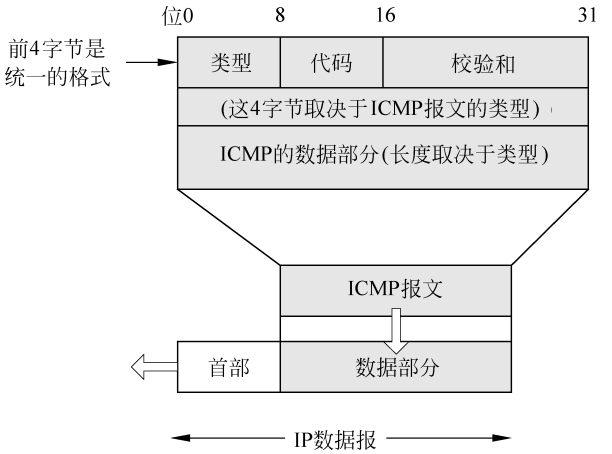



图 5-12　ICMP 报文的格式

用的算法和 IP 首部校验和算法相同。ICMP 的校验和字段是必须的。

ICMP 报文的前 4B 是统一的格式,共有 3 个字段,即类型、代码和校验和。接着的 4B 的内容与 ICMP 的类型有关。最后面是数据字段,其长度取决于 ICMP 的类型。类型字段可以有 15 个不同的值,以便描述特定类型的 ICMP 报文。某些 ICMP 报文还使用代码字段的值来进一步描述不同的条件。ICMP 类型字段的具体含义如表 5-3 所示。

**表 5-3　ICMP 类型字段的具体含义**

| 类 型 字 段 | ICMP 报文类型 |
|---|---|
| 0 | 回送应答 |
| 3 | 目的地不可达 |
| 4 | 源站抑制(Source Quench) |
| 5 | 重定向(改变路由) |
| 8 | 回送请求 |
| 9 | 路由器通告(Advertisement) |
| 10 | 路由器请求(Solicitation) |
| 11 | 超时 |
| 12 | 数据报参数错 |
| 13 | 时间戳请求 |
| 14 | 时间戳应答 |
| 15 | 信息请求(已过时) |
| 16 | 信息应答(已过时) |
| 17 | 地址掩码(Address Mask)请求 |
| 18 | 地址掩码(Address Mask)应答 |

ICMP 报文种类很多,可大致分为两类,即 ICMP 差错报告报文和 ICMP 查询报文。ICMP 差错报告报文分为 5 类:目标不可达、源站抑制、超时、参数问题、路由重定向。ICMP 查询报文分为 4 类:回送请求与应答、时间戳请求与应答、地址掩码请求与应答、路由询问和报告。其对应关系如图 5-13 所示。

ICMP 各类报文的含义如下。

(1) 目标不可达。一般出现这种报文的情况大致可分为两类:①路由器寻址失败。如果路由器发现找不到送达 IP 数据报到达目标主机的路径,就丢弃该 IP 数据报,然后这个路

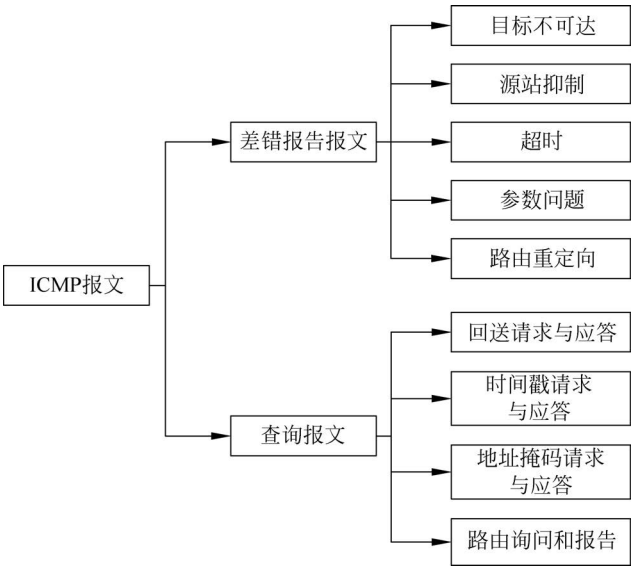



图 5-13　ICMP 报文的种类

由器就向源主机返回 ICMP 差错报文。②路由器寻址成功，但是目标主机找不到有关的用户协议或上层服务访问点。出现这种情况的原因可能是 IP 头中的字段不正确；也可能是路由器必须把数据报分段，但 IP 头中的 D 标志已置位。

（2）源站抑制。由于 IP 协议中没有流量控制机制，当路由器的处理速度太慢或者路由器传入数据速率大于传出数据速率时就有可能造成拥塞。为了控制拥塞，IP 软件采用了“源站抑制”技术，利用 ICMP 源抑制报文抑制源主机发送 IP 数据报的速率。路由器对每个接口进行密切监视，一旦发现拥塞，立即向相应源主机发送 ICMP 源抑制报文，请求源主机降低发送 IP 数据报的速率。

（3）超时。在 IP 网络中经常会出现由于路由表的错误而导致网络的转发也出现错误，由此可能发生某些数据报在网络内的路由器间出现“兜圈子”的情况。为了避免 IP 数据报在网络内无休止地循环传输，从而占用网络流量，在 IP 协议中设置了每个数据报的生存周期（TTL）；路由器如果发现 IP 数据报的生存时间已超时，或者目标主机在一定时间内无法完成重装配，就发回一个 ICMP 超时差错报告，通知源主机该数据报已被抛弃。

（4）参数问题。当数据报在传输时，路由器或主机会自动判断数据报头或报头选项是否出现错误；如果报头缺少某个域、IP 头中的字段或语义出现错误等，路由器便向源主机发送 ICMP 参数出错报文，报告错误的 IP 数据报报头和错误的 IP 数据报选项参数等情况。

（5）路由重定向。在互联网中，主机可以在数据传输过程中不断地从相邻的路由器获得新的路由信息。通常，主机在启动时都具有一定的路由信息，这些信息可以保证主机将 IP 数据报发送出去，但经过的路径不一定是最优的。路由器一旦检测到某 IP 数据报经非优路径传输，它一方面继续将该数据报转发出去，另一方面将向主机发送一个路由重定向 ICMP 报文，通知主机去往相应目标主机的最优路径。这样主机经过不断积累便能掌握越来越多的路由信息。ICMP 重定向机制的优点是，保证主机拥有一个动态的、既小且优的路由表。

（6）回送请求与应答。回送请求主要是测试两个网络结点之间的线路是否畅通。它是

由主机或路由器向一个特定的目的主机发出询问。收到此报文的主机必须给源主机发送ICMP回送应答报文。有些资料中也形象地将此过程称为回声。通常为了测试两个主机之间的连通性,会使用一种称为PING的命令,PING的过程实际就使用了ICMP回送请求与回送应答报文。不过值得注意的是,PING是应用层直接使用网络层ICMP的一个例子。它没有通过传输层的TCP或UDP。图5-14所示的是一个用PING进行测试的具体实例。

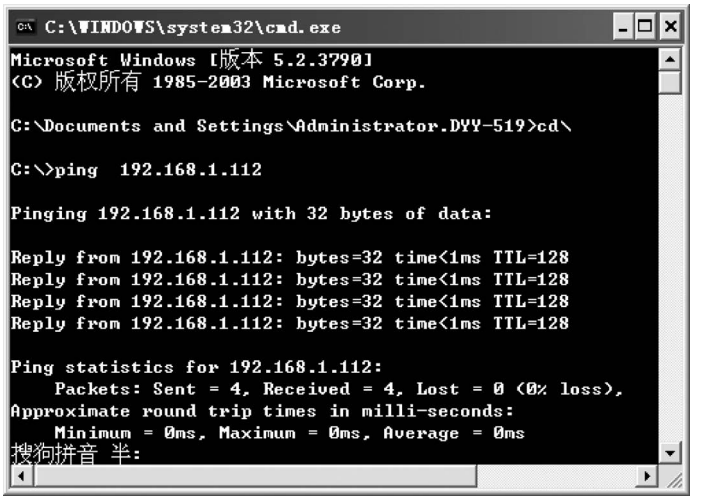



图5-14 PING测试实例

(7) 时间戳请求与应答。时间戳请求与应答可用来进行时钟同步和测量时间。请求方发出本地的发送时间,应答方返回自己的接收时间和发送时间。这种应答过程如果结合强制路由的数据报实现,则可以测量出指定线路上的延迟。

(8) 地址掩码请求与应答。当主机不知道自己所在的局域网中的子网掩码时则可以使用地址掩码请求报文,最常见的是无盘系统在系统引导时获取自己的子网掩码,具体过程非常类似于无盘系统使用RARP获取自己的IP地址。首先主机向同一局域网内的路由器发送地址掩码请求报文,路由器在获得请求后以地址掩码响应报文回答,告诉对方所需要的子网掩码。知道子网掩码主要是为了判断出数据包的目标结点与源结点是否在同一个局域网中。

(9) 路由询问和报告。主要是为了掌握在局域网中的路由器的工作信息,简单地说就是为了测试路由器是否工作正常;主机将路由器询问报文进行广播(或多播)。收到询问报文的一个或几个路由器就使用路由器通告报文广播其路由选择信息;另外,当主机没有进行询问时,路由器也会自动地、周期性地发送路由通告报文。路由器在通告报文中不仅会报告自己的存在,同时也会通告它所知道的在这个局域网内的所有路由器。

## 5.3 路由选择算法与路由协议

Internet是由许多分布于世界各地的互联网(internet)相互连接构成的,在Internet上负责实现连接并进行数据包转发的设备就是路由器。路由器以网络(间接地是指负责网络连接的路由器)而不是以主机作为路由选择的单位。路由器根据自己选用的路由协议生成

路由表，而路由协议是依据特定的路由选择算法实现的。工作过程中，路由器通过查找路由表中的记录来决定 IP 数据报的转发路径。那么，当数据包要到达目的地时，很可能要经过多个网络。问题是应该如何选择数据包的路由路径呢？选择的依据又是什么？这就是路由选择算法需要解决的问题。在 Internet 中，因为网络规模及要求的不同，路由选择算法及协议也有所差别，本节主要介绍几个典型的路由选择协议（RIP、OSPF 和 BGP）。

首先，学习一下路由选择算法（经常简称为路由算法）。路由选择算法是路由器产生和不断更新路由表的依据，它是路由选择协议的核心内容。一个理想的路由选择算法应该具有如下 5 个特点。

（1）正确性。正确是指沿着路由表所给出的路径，分组一定能够到达目的地。

（2）简单性。路由算法的计算必然会消耗一定的软、硬件资源，从而增加了分组转发的延时，所以算法要尽量简单，才可能更有实用价值。

（3）健壮性。路由算法应该能适应网络拓扑结构及流量的变化，在外部条件发生变化时仍能正确地实现预定的数据转发功能，而不发生剧烈的振荡变化，实现各链路之间的负载均衡。

（4）公平性。路由算法应对所有用户（一些优先级较高的用户除外）都是平等的，不能因为一些原因忽视了一些源结点的转发需求。

（5）最佳性。所谓“最佳”，是指分组转发过程中的最低开销。开销中涉及的考虑因素有很多，如数据传输速率、传播时延、占用带宽、通信费用、安全可靠等。

在实际的路由过程中，由于网络的复杂多变性和用户需求的多样性，想寻找一条能够满足各个方面需求的路径往往是比较困难甚至是不可能的。例如，找到了一条传输时间最短的路径，但这条路径不一定是费用最省、安全性最好的路径；同样的道理，费用最省的路径也不一定是传输时间最短的路径。所以在实际应用中，只能依据用户的需求，找出一条相对较为合理的路径，但不一定是各方面都最优的路径。可以说，“最佳”路径只是相对于某一种特定要求下得出的较为合理的选择。

从路由选择算法对网络通信量和网络拓扑变化的自适应能力的角度划分，可以将路由选择算法分为静态路由选择算法和动态路由选择算法两大类。静态路由选择算法也称为非自适应路由选择算法，其特点是简单、易于实现，开销较小，但是不能随着网络状况的变化进行动态调整；动态路由选择算法也称为自适应路由选择算法，其特点是能够较好地适应网络状况的变化，但实现过程相对复杂，开销较大。在使用过程中，可以依据实际需要选择决定路由算法的种类。一般地，一些小型的、变化不是很频繁的网络宜选用静态路由选择算法；而一些大型的、变化较为频繁的网络则应该选用动态路由选择算法。

在路由协议中，自治系统是一个常用的概念。Internet 规模的不断扩大给路由技术提出了巨大的挑战。试想，让每一台路由器都保存一张具有全网信息的路由表是一件多么困难的事情，即便是勉强做到了，这张路由表中的记录数也是巨大的，给后期的查找与更新带来许多不便。由此可见，让每一台路由器都保存一张具有全部网络信息的路由表是不现实的，也是低效的。同时，许多连接到 Internet 上的部门及网络在享用各种网络服务的同时，并不愿意让外界了解本单位的网络布局细节及采用何种路由选择协议等信息。因此，将 Internet 划分成许多较小的自治系统（Autonomous System，AS）。

自治系统（AS）是在单一技术管理下的一组路由器，这些路由器使用同一种 AS 内部的

路由选择协议和共同的度量以确定分组在AS内的路径,同时还使用一种AS之间的路由选择协议用以确定分组在AS之间的路由。这样一来,每个AS可以使用与其他AS不同的路由协议,具有较高的自主性。一般地,将自治系统内部使用的路由协议称为内部网关协议(如RIP、OSPF),把自治系统内部的路由选择称为域内路由选择(Intradomain Routing);将自治系统之间使用的路由协议称为外部网关协议(如BGP),把自治系统之间的路由选择称为域间路由选择(Interdomain Routing)。自治系统的结构如图5-15所示。

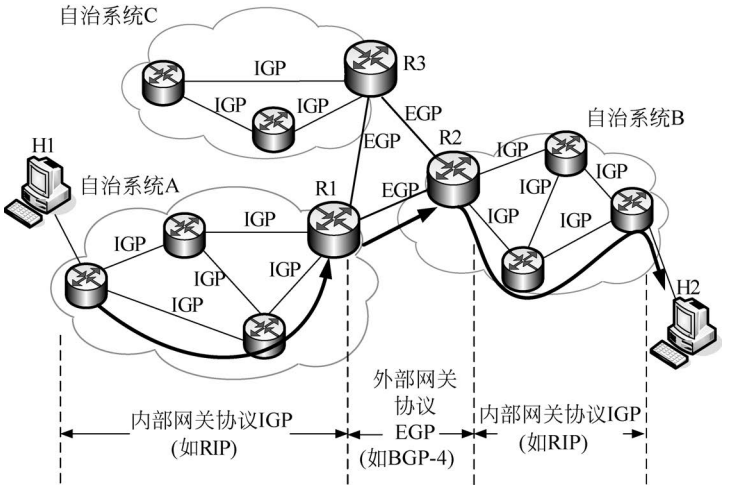



图5-15 自治系统的结构

每一个自治系统都有一个唯一的标识,称为AS号。它是由IANA(Internet Assigned Numbers Authority)来授权分配的。这是一个16位的二进制数,数值范围为1～65 535,其中65 412～65 535为AS专用组(RFC2270)。自治系统概念的提出,实际上是将Internet分成了两层。第一层是自治系统内部的网络,另一层是它外部的主干网络。自治系统内部的路由器完成自治系统中主机之间的分组交换;而整个自治系统又通过一个主干路由器连接到外部的主干网络。


视频讲解

## 5.3.1 路由信息协议

**1. 基本工作原理**

路由信息协议(Routing Information Protocol,RIP)是内部网关协议IGP中最先得到广泛应用的一个协议。现在较新的RIP版本是1998年11月公布的RIP2版本,较之早先的版本,RIP2本身并没有多大变化,但性能上出现了一些改进。RIP2支持变长子网掩码(VLSM)和CIDR,而且还提供简单的鉴别过程支持多播。

RIP是一个基于距离向量选择的路由协议,使用Bellman Ford算法。Bellman Ford算法也称为距离矢量算法,简称V-D算法。其工作原理是:路由器周期性地向外广播最新的路径信息,主要包括由(V,D)序偶表组成的路由更新报文,其中V代表可到达的信宿,D代表该路由器到达信宿所经过的距离。距离D按照经过的路由器的个数计算,其他路由器收到此更新报文后,按照最短路径原则对各自的路由表进行更新。

RIP中的距离是指“跳数”,也可以简单地理解为链路数,每经过一个路由器,跳数就增加1。与该路由器直接相连的网络,“距离”定义为1。RIP仅以跳数作为度量标准,它允许

的最大跳数为15(16表示不可达),任何超过15个站点的目的地的"距离"均被标记为16,即不可达,这也决定了RIP只适用于规模较小的网络。当系统变大后受到无穷计算问题的困扰,且往往收敛很慢,现已被OSPF所取代。现在一些新的路由器,所允许的最大距离为31。

RIP协议会在每个路由器上保存一张从本地路由器到其他每个网络的地址表,其表结构如表5-4所示。RIP规定,在路由表中只能为每一个目的网络保存一条路由记录,即使存在有多条路径。

**表5-4 简化的RIP路由表**

| 目的网络 | 下一跳路由器 | 距离 |
|---|---|---|
| 219.230.80.0 | 219.230.81.2 | 3 |
| … | … | … |

RIP报文格式如图5-16所示,从图5-16中可以看到,RIP报文借助UDP协议进行封装,使用UDP协议的520端口。事实上,RIP协议是一个应用层协议。

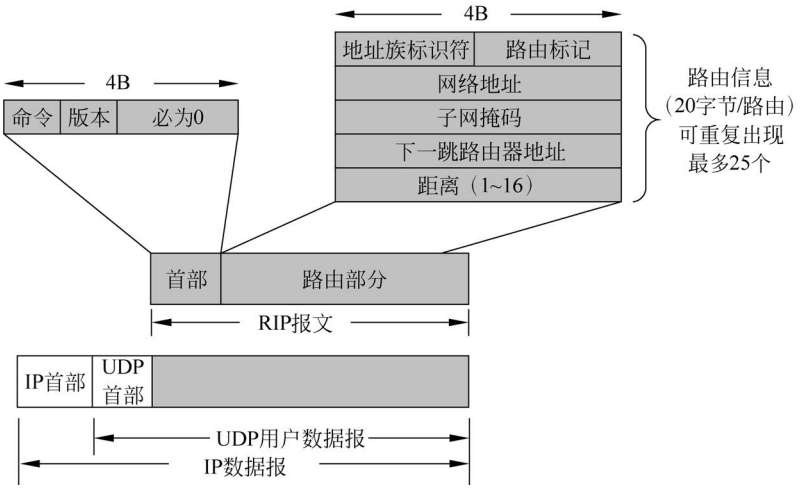



图5-16 RIP2报文格式

RIP报文由首部和路由部分组成。首部中的命令字段指出RIP报文的类型。如果是1,表示请求路由信息;如果是2,表示对请求的响应。路由部分中的地址族标识符字段用来标识所使用的地址协议,如使用IP地址时,该字段值为2;路由标记字段填入的是AS号,指明了始发报文来自哪个自治系统;网络地址字段标识要到达的目的网络的地址;子网掩码字段用来标识目的网络的子网掩码;下一跳路由器地址指明数据包应该被送往的下一跳的路由器的地址;距离字段标识到达目的网络的跳数值。一个RIP报文最多可包含25条路由信息,每条路由信息占20B。因此,一个RIP报文最大长度应该为504B(包含4B的首部及最多可重复出现25个的路由信息,每个路由信息占20B)。

使用RIP协议的网络中,路由器在交换信息时,遵循以下3点规则。

(1) 每一个路由器都与自己的邻居路由器共享自己的路由表信息,不相邻的路由器不准交换路由信息。

(2) 按固定的时间间隔交换路由信息。每一个路由器在经过一定的时间间隔后将其信

息发送给邻居路由器,一般的时间间隔为30s。但当网络拓扑发生变化时,路由器要及时地向邻居路由器发送变化后的路由更新信息。如果一个路由器在180s内未收到邻居路由器的状态信息,就可以将该邻居路由器标记为不可达。

(3) 路由器交换的信息是自己所知道的所有信息,也就是自身全部的路由表。

**2. 路由表的生成与更新**

每个路由器在开始工作时,都需要首先将自己的路由表进行初始化。初始化过程中规定,每个与当前路由器直接相连的网络的距离值为1。如前所述,每个路由器周期性地向与之直接相连的邻居路由器广播自己的路由表,告知邻居路由器自己到达各个网络所需要经过的距离(链路数)。当一个路由器接收到来自其他路由器的路由表时,需要对自己的路由表进行必要的更新,更新算法如下。

(1) 收到邻居路由器A的路由表A_table后,将其中的"下一跳路由器地址"字段都更改为A,并将所有的"距离"都加1。这样做的原因是对于当前接收路由器而言,下一跳转发路由器就是A,相对于路由器A来说,当前路由器到达目的网络的距离是路由器A到达目的网络的距离加上本路由器到达路由器A的距离1。

(2) 依据更改后的路由表A_table,依次检查其中的每一行记录。若当前行中的目的网络不在本地路由表中,则将该行添加到本地路由表中;否则,若下一跳中记录的内容与本地路由表当中的内容相同,则替换本地路由表中对应的记录;若该行的"距离"小于本地路由表中相应行的"距离",则用该行更新本地路由表当中对应的记录;若该行的"距离"大于本地路由表中相应行的"距离",且与"下一跳路由器地址"中的内容不同时,不做任何处理返回。

(3) 若超过180s后仍未收到邻居路由器A的路由表,则将到达邻居路由器A的距离设置为16(不可达)。

图5-17说明了RIP协议的路由表更新过程。在开始时,所有路由器中的路由表只记录与路由器直接相连的网络的情况,包括目的网络的地址和相应的距离;图5-17中"下一站路由器"项目中有符号"-",表示直接交付,因为路由器和同一网络上的主机可直接通信而不需要再经过其他路由器进行转发。接着,各路由器都向其相邻路由器广播自己路由表中的信息,接收到来自邻居路由器的更新信息后,依据上述算法更新自己的路由表。R2收到了路由器R1和R3的路由表更新信息,随后更新自己的路由表;更新后的路由表再发送给路由器R1和R3,路由器R1和R3依据相同的更新算法对自己的路由表进行更新。

RIP协议最大的优点是实现简单,开销较小;但它也有一些固定的缺点,主要是好消息传播得快,而坏消息传播得慢,以及存在无穷计算的问题。好消息传播得快,坏消息传播得慢主要是指当网络出现故障时,要经过比较长的时间才能将此故障消息传送至所有的路由器,产生这一现象的根本原因在于RIP的工作机制;相反,如果一个路由器发现了一条更短的路由,那么这种更新信息可以迅速得到传播。

针对这些问题,RIP有一些解决方案,主要包括设置最大距离值(一般设置为15)、触发更新、水平分割、毒性逆转等。遗憾的是,这些办法在解决一些问题的同时,又带来了一些新的问题,如由触发更新技术引发的广播雪崩等。所以,当网络规模较大时,RIP协议就已经不是最好的选择,OSPF协议可以较好地运行在较大规模的网络中。

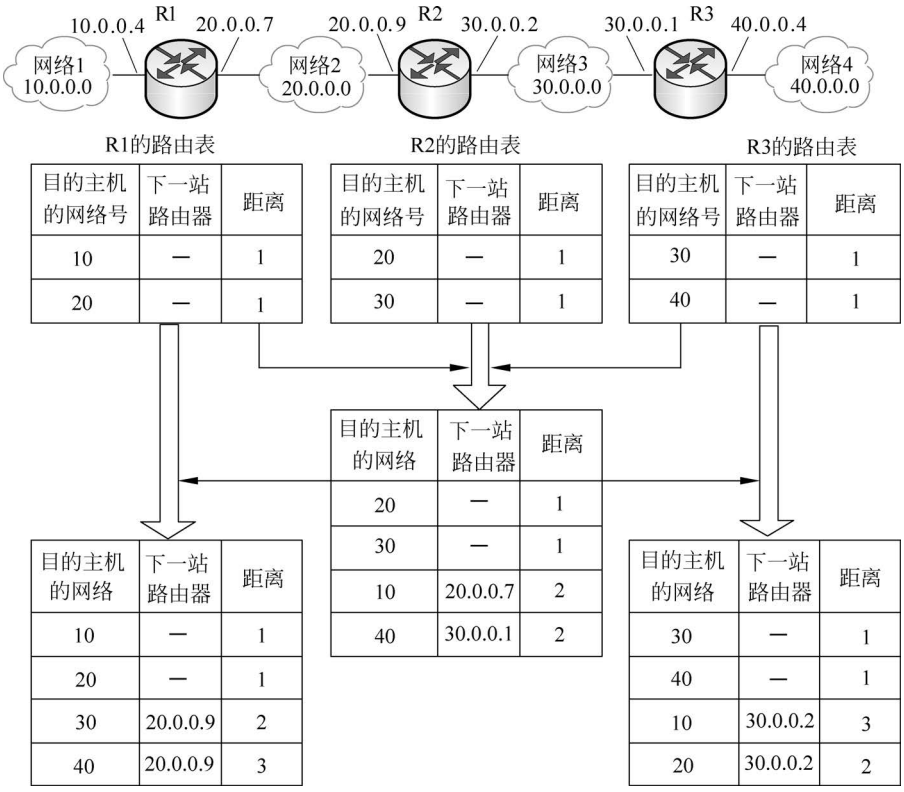



图 5-17　RIP 协议的路由表更新过程

## 5.3.2　开放最短路径优先协议

开放最短路径优先协议(Open Shortest Path First,OSPF)是一个基于链路状态算法的路由协议。OSPF 是为克服 RIP 的缺点于 1989 年开发出来的。OSPF 的原理很简单,但实现起来却较复杂。OSPF 协议是通过使用 Dijkstra 提出的最短路径算法 SPF 来工作的。首先,构建一个以本路由器为根的最短路径树,然后根据最短路径树来组建路由表。OSPF 的第二个版本 OSPF2 已成为因特网标准协议。事实上,Internet 上的路由协议基本上都是基于最短路径算法的,只是 OSPF 使用了所谓的"最短路径优先"的名称而已。OSPF 规定每个路由器保存一个链路状态数据库,实质上就是一张链路状态表,其中链路状态值(Cost)一般设置为链路通断,用 1 表示链路是连通的,用∞表示链路不存在或者不通。

**1. 基本工作原理**

OSPF 遵循如下的规定。

(1) 每个路由器向本自治系统(AS)内的所有路由器广播路由信息,而不仅仅是只向邻居路由器广播路由信息。

(2) 发送给其他路由器的是与本路由器相邻的所有路由器的链路状态信息。

(3) 只有当链路状态发生变化时,路由器才用洪泛法向所有路由器发送信息。

(4) 不同的链路可以使用不同的成本度量值,一般都选用链路进行度量。

以上规定(3)中提到的洪泛法是指路由器通过所有的输出端口向它所有的邻站发送信

息,而所有站又将其信息发送给自己的所有相邻路由器(但不再发送给刚刚发信息的那个路由器)。以上4点规定的目的是要让每个路由器都有整个网络或AS每一时刻的准确拓扑图,从而计算出到达目的网络的最短路径。经过路由器之间频繁地交换链路状态信息,网络中的所有路由器最终都能建立一个链路状态数据库(Link-state Database),这个链路状态数据库反映的正是全网的拓扑结构图。我们期望的是,链路状态数据库中的信息在全网所有路由器中是一致的,因为每一时刻的网络拓扑图只有一个。所以,OSPF协议在工作过程中需要经常交换各自的链路状态信息以期尽快实现全网同步。根据链路状态数据库中的数据可以构造出自己的路由表,进而计算出到达目的网络的最短路径。

随着网络规模的不断增大,全网链路状态信息也会急剧增加,频繁的链路状态信息交换势必会给网络带来沉重甚至是不可接受的负担,同时也增大了计算工作量。为了能够让OSPF协议工作于大规模网络,OSPF将一个自治系统AS再划分为一些更小的范围,称为区域(Area)。每一个区域都有一个32位的区域标识符,一般采用点分十进制记法表示。上层区域称为主干区域(Backbone Area),标识符规定为0.0.0.0,用来连通其他下层的区域。区域不能太大,通常在一个区域中路由器的数目不超过200个。

区域划分的示例图如图5-18所示。划分区域后,洪泛法交换链路状态信息的范围将局限于每一个区域而不是整个的自治系统,这就减少了整个网络上的通信量。相应地,每个路由器也只知道本区域的完整网络拓扑图,而不知道其他区域的网络拓扑情况。如果需要与其他区域中的路由器进行通信,必须借助主干区域中的路由器。在一个区域的边界把有关本区域的信息汇总起来发送到其他区域的路由器称为区域边界路由器,区域边界路由器至少要有一个接口在主干区域中;在主干区域中的路由器称为主干路由器,主干路由器也可以同时是区域边界路由器。在图中的R3、R4和R7都是区域边界路由器,R3、R4、R5、R6和R7都是主干路由器。在主干区域内,还应专门有一个路由器负责与其他自治系统的信息交换,这个路由器称为AS边界路由器,如R6就是一个AS边界路由器。

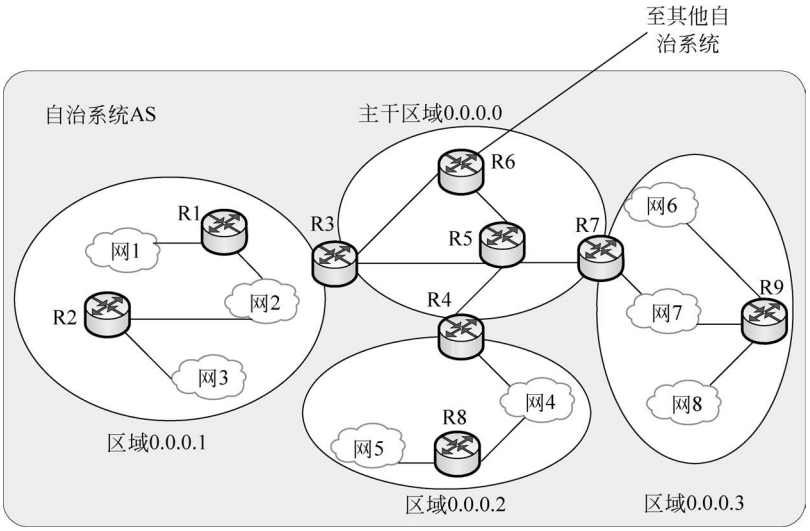



图5-18 OSPF区域划分示例

OSPF协议数据直接使用IP数据报进行封装,使用IP中的协议字段值是89,称为

OSPF 分组。这一点不同于 RIP 路由协议(RIP 使用 UDP 协议进行数据封装),从这一角度分析,OSPF 是网络层的协议。绝大多数资料都将路由协议放在网络层这一部分进行讲解,但读者需要明白,这并不意味着所有的路由协议都位于网络层中。事实上,RIP 协议和后面要讲到的 BGP 协议,都是应用层的协议。一些新型的路由协议,也都工作于应用层中而非网络层。熟悉网络分层结构的读者一定会有疑问,路由协议是工作在路由器中的,而路由器明确是一种网络层设备,网络层的设备怎么会运行应用层的协议呢?以至于大多数的读者都认定路由协议是网络层的协议。实际上,路由器中运行了相关的应用层进程,由这些进程负责路由协议数据的封装及处理。换言之,路由器中加载了需要运行路由协议的应用层进程。

OSPF 分组的首部固定为 24B,分组很短,便于在网络中快速传播,其格式如图 5-19 所示。OSPF 分组中各部分的含义如下。

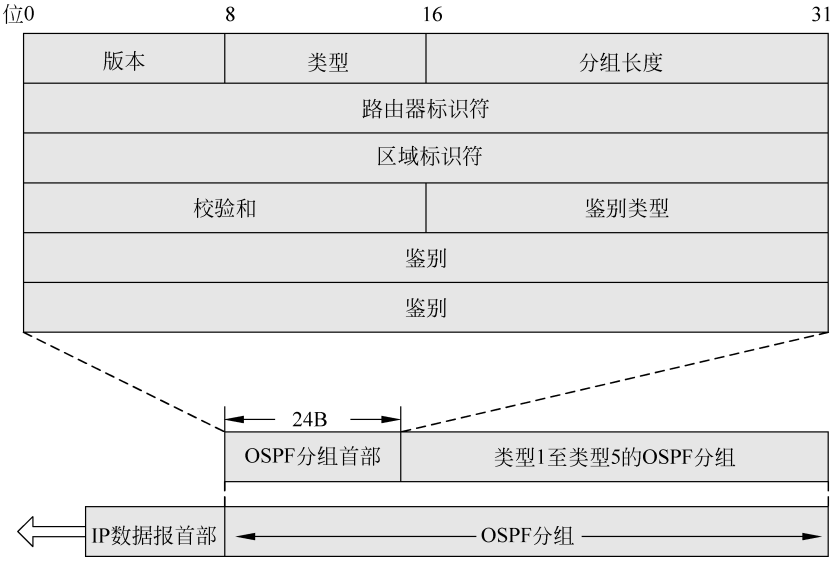



图 5-19　OSPF 分组格式

(1) 版本。定义所使用的 OSPF 协议的版本,对于 OSPFv2 来说其值为 2。

(2) 类型。说明 OSPF 数据包的类型,其数值为 1～5。OSPF 数据包共有 5 种分组类型,依次为 Hello 报文、DD 报文、LSR 报文、LSU 报文、LSAck 报文。

① 问候(Hello)分组。用来发现和维持邻站的可达性,以建立和维护相邻的两个 OSPF 路由器的关系。

② 数据库描述(Database Description,DD)分组。向邻站给出自己的链路状态数据库中的所有状态项目的摘要信息。

③ 链路状态请求(Link State Request,LSR)分组。向对方请求发送某些链路状态项目的详细信息。

④ 链路状态更新(Link State Update,LSU)分组。用洪泛法对全网更新链路状态,这种分组是最重要、也是最复杂的一种分组。

⑤ 链路状态确认(Link State Acknowledgment,LSAck)分组。对链路更新分组的确认。

(3) 分组长度。说明包括首部在内的整个分组的长度,单位是字节。

(4) 路由器标识符。描述数据包的源地址,以IP地址的形式表示,是始发该数据包的路由器的ID。

(5) 区域标识符。用于区分OSPF数据包所属的区域号,它是始发该LSA的路由器所在的区域号。

(6) 校验和。用补码加法生成的校验和,用来检测分组中的差错。

(7) 鉴别类型。目前只有两种,0表示不用,1表示口令。

(8) 鉴别。包含OSPF鉴别类型,其值按鉴别类型所定,共有8B。当鉴别类型为0时不作定义,类型为1时此字段为密码信息,类型为2时此字段包括Key ID、MD5验证数据长度和序列号的信息。MD5鉴别数据添加在OSPF报文后面,不包含在鉴别字段中。

OSPF使用洪泛法发送链路状态信息,收到信息的路由器需要回送确认信息,因此这种洪泛法是可靠可信的。

**2. 链路状态表的生成与更新**

路由器初始化时,每个路由器的链路状态表中只有与其直接相连的路由器的链路状态信息,该状态信息是对路由器进行初始配置时进行设置的,并不是当前网络的最新状态信息。路由器运行过程中,会不断地更新链路状态表中的内容,以期与当前网络拓扑结构一致。OSPF协议规定,每两个相邻路由器每隔10s要交换一次问候(Hello)分组,这样就能知道哪些邻站是可达的。在正常情况下,网络中传送的绝大多数OSPF分组都是问候分组。如果40s都还没有收到邻居路由器的Hello分组信息,则认为该邻居路由器是不可达的,随即修改链路状态数据库中所对应的记录,用洪泛法传播相应的链路状态信息,并重新计算路由表。

由于一个路由器发送出的更新信息不可能同时到达所有的路由器,这样就有可能造成各路由器中保存的链路状态数据库不一致。在这种情况下,就需要进行同步。所谓同步,就是指不同路由器的链路状态数据库的内容是一样的。在OSPF协议中,使用数据库描述分组(Database Description)交换状态信息,实现各方的同步。即使没有链路状态的变化,每隔30min也要进行一次全网同步,这样可以保证各方路由器的一致。

### 5.3.3 边界网关协议

边界网关协议(Border Gateway Protocal,BGP)是在1989年问世的,现今最新的版本是BGP-4,它已成为因特网草案标准协议。BGP完成了AS之间的路由选择,可以说它是现今整个因特网的支架。

这里首先需要说明一个问题,就是AS之间的路由能否使用内部网关协议(如RIP、OSPF),答案是否定的。RIP和OSPF主要用于AS内部,在AS内部选择最佳路由。它们没有什么特别的限制性条件,也没有人为因素,只是依据相关路由信息和事先制定好的路由策略寻找最佳路由。然而,Internet是一个规模巨大、环境复杂的互联网,在AS之间进行路由选择是一件非常困难的事情。更为重要的是,自治系统间的路由选择要受到很多的约束限制,包括人为因素,如存在政治原因或者经济原因,可能会规定从A到B的信息必须经由哪些路由器而不能经由哪些路由器等,这一系列的原因都说明RIP和OSPF协议不适合工作于自治系统之间。通常,在自治系统之间一般只需要找到可行的路由而不一定是最佳的路由,这也与RIP和OSPF的工作机制不同。这就要求有一种能够较好地工作于自治系统

之间的外部网关协议。

**1. BGP-4 的工作原理**

每一个自治系统需要选择至少一个路由器作为该自治系统的“BGP 发言人”，BGP 发言人一般是 AS 的边界路由器。两个 BGP 发言人通过一个共享网络连接在一起，如图 5-20 所示。在 5-20 图中画出了 3 个 AS 中的 5 个 BGP 发言人，每一个 BGP 发言人除了必须运行 BGP 协议外，也需要运行 AS 内部的路由协议，如 RIP、OSPF 等。

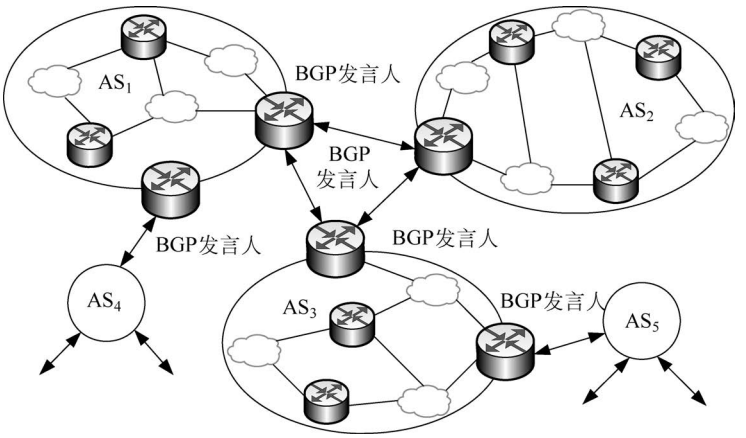



图 5-20　BGP 发言人

一个 BGP 发言人与其他自治系统中的 BGP 发言人要交换路由信息，就要先建立 TCP 连接(端口号为 179)，然后在此连接上交换 BGP 报文以建立 BGP 会话(Session)，利用 BGP 会话交换路由信息。使用 TCP 连接能够提供可靠的服务，简化了路由选择协议，BGP 协议中不再使用差错控制和重传机制。使用 TCP 连接交换路由信息的两个 BGP 发言人，彼此成为对方的邻站或对等站(Peer)。

BGP 路由表类似于 RIP 的路由表，但 BGP 计算出的路由与 RIP 不同，RIP 只是指出下一跳地址，而 BGP 指明的是一条完整的路径，因此也将 BGP 这样的协议称为路径向量协议。BGP 所交换的网络可达性信息就是要到达某个网络(用网络前缀表示)所要经过的一系列的自治系统。BGP 发言人互相交换从本 AS 到邻居 AS 的可达信息，随着该信息的传播，从一个 AS 到其他 AS 的可达信息被记录下来，进而得到了不同 AS 之间的一条可达路径信息。由此可以看到，AS 之间的路由包含了一系列的 AS 地址，表示从源 AS 到目的 AS 之间经过的 AS 列表。如图 5-21 所示，表示了图 5-20 中的 $AS_1$ 上的 BGP 发言人构造出的 AS 连通图，据此可以实现 $AS_1$ 至其他 AS 间的数据传播。注意，这个连通图是树状结构，不存在回路。

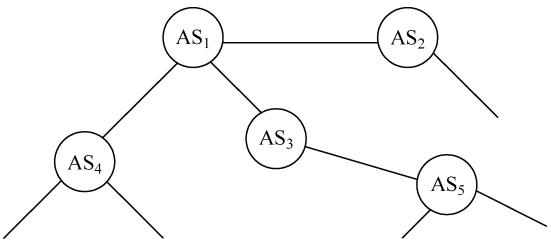



图 5-21　AS 的连通图

BGP具备负载均衡能力,可以将负载合理地分配到多条路径上进行传输,从而更好地利用网络带宽,实现QoS路由。

BGP-4协议的报文格式如图5-22所示,各字段的含义如下。

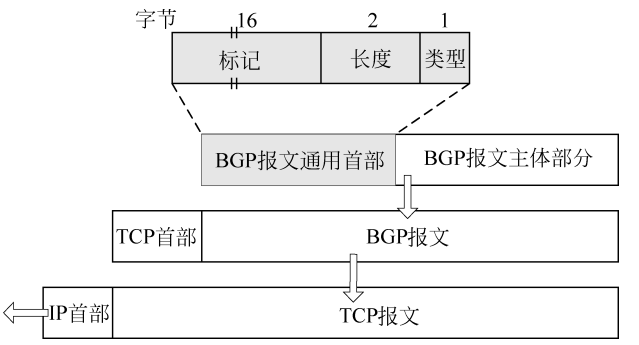



图5-22 BGP-4协议报文格式

(1) 标记。16B,用于鉴别收到的BGP报文,当这个字段不使用时,全部置1。

(2) 长度。2B,定义包括首部在内的报文的长度,最小值为19,最大值为4096。

(3) 类型。1B,定义分组的类型,其值为1～4。

BGP-4使用4种不同类型的报文:打开(Open)报文、更新(Update)报文、保活(Keepalive)报文和通知(Notification)报文。打开报文用来与相邻的另一个BGP发言人建立联系,初始化通信;更新报文用来发送某一路由信息,以及列出要撤销的多条路由;保活报文用来确认打开报文和周期性地证实邻站关系;通知报文用来发送检测到的差错。

**2. BGP-4的工作过程**

在BGP刚刚运行时,BGP的邻站要交换整个的BGP路由表,但以后只需要在发生变化时更新有变化的部分,这样做对节省网络带宽和减少路由器的处理开销方面都有好处。具体的工作过程如下所述。

(1) 当一个BGP发言人需要与另一个自治系统中的BGP发言人进行通信时,首先向其发送打开报文,对方进行确认后成为对等站(邻站),之后便可以相互发送路由信息。

(2) 对等站(邻站)关系建立后,对方需要确认对方是存活的。为此,两个BGP发言人需要周期性地交换保活报文(一般间隔为30s)。保活报文只有19字节长(仅含有BGP报文的通用首部),不会给网络造成太大的开销。

(3) 一个BGP发言人可以向对等站(邻站)发送更新报文,告知对方路由信息,同时也可以撤销先前的路由。撤销路由可以一次撤销多条,但增加新路由时,每个更新报文只能增加一条。

(4) 收到更新报文的BGP发言人更新路由信息,并将这些信息发送给自己的其他对等站(邻站)。

(5) 每个BGP发言人根据自己保存的路由信息,为两个不同的AS之间确定一条可行的路由。

## 5.4 路由器与第三层交换技术

路由器是一种有多个输入和输出端口的专用计算机,其主要任务是实现路由协议,完成

数据转发。数据分组在网络中经由路由器的连续逐跳转发最终交付给目的网络。路由器工作于网络层中，依靠网络层的逻辑地址实现寻址。路由器的常见连接形式有如下 3 种。

（1）连接交换机。计算机首先组成小范围的局域网（一般通过交换机实现），局域网通过路由器连接成为广域网，如图 5-23 所示。

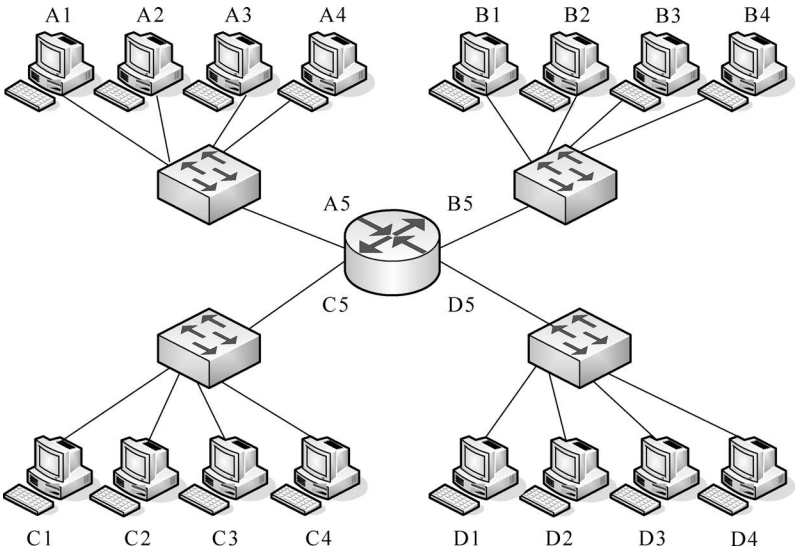



图 5-23　路由器连接交换机

（2）连接远程计算机。远程计算机（如家庭中的计算机、某单位的服务器等）通过调制解调器、远程线路连接到接入路由器（前端有 Modem 池）上实现联网，如图 5-24 所示。

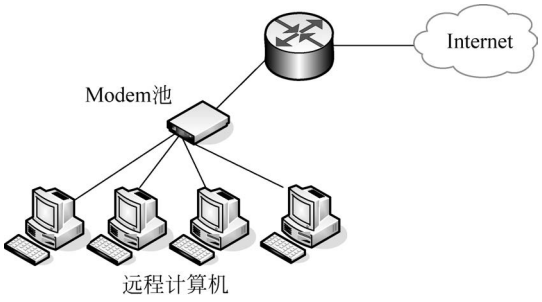



图 5-24　路由器连接远程计算机

（3）混合式连接。这种连接方式可以组合使用前两种连接方式，也是实际中使用较多的一种连接方式。

一般说来，路由器可以实现以下 3 种主要功能。

（1）网络互联。实现局域网-局域网、局域网-广域网、广域网-广域网 3 种类型的网络互联。

（2）分组转发。对接收到的数据分组，路由器检查分组中的源地址与目的地址，然后根据路由表中的相关信息，决定该数据分组的输出路径。

（3）数据记录。路由器会记录用户访问网络的相关数据，如每次访问的时间、发送/接收的字节数、访问的源地址/目的地址等信息。

### 5.4.1 路由器的构成

整个路由器可以划分为两大部分：路由选择部分和分组转发部分，如图5-25所示。

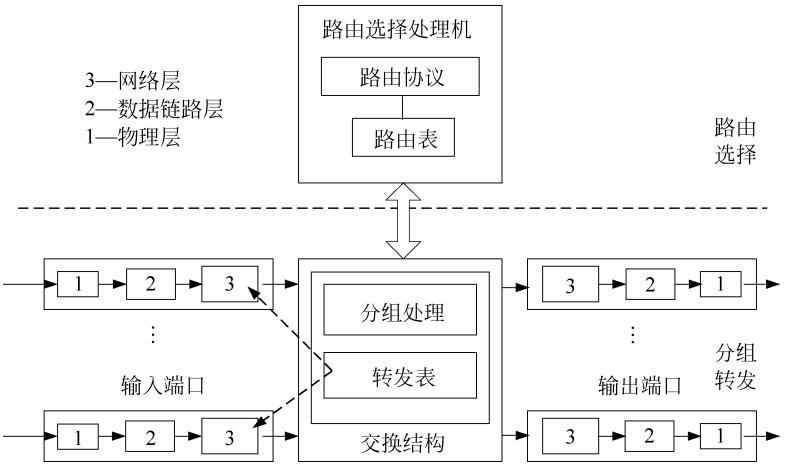



图5-25 路由器的构成

路由选择部分也称控制部分，核心部件是路由选择处理机。路由选择处理机依据一定的路由协议生成并维护路由表。分组转发部分由3部分组成，即交换结构、一组输入端口和一组输出端口。交换结构根据转发表对分组进行处理，将某个输入端口进入的分组从一个合适的输出端口转发出去。交换结构本身就是一种网络，但这种网络完全包含在路由器之中。

从网络术语的角度来讲，“转发”和“路由选择”是有区别的。在互联网中，“转发”就是路由器根据转发表把收到的IP分组从本路由器合适的端口转发出去，仅仅只涉及当前路由器；而“路由选择”则涉及多个路由器，路由表也是由许多路由器进行协同工作后的结果，是一个全局性的工作，考虑到了整个网络当前的状况，路由选择的目标是依据路由协议选出一条较为合理的传输路径。转发表是从路由表得到的，转发表包含了完成本次转发所必需的信息(如MAC地址等)。虽然在讨论路由算法工作原理时，可以不区分两者，但是作为学习，还是有必要将两者进行区别的。转发表和路由表使用不同的数据结构，转发表的结构应当使查找过程最优化，但路由表则需要对网络拓扑变化的计算进行最优化。路由表是用软件实现的，转发表可以用特殊的硬件来实现。

路由器接收到数据分组后，根据网络物理接口的类型，调用相应的数据链路层功能模块对数据分组进行处理，完成CRC校验、帧长度检查等工作；通过数据链路层的完整性验证后，路由器开始处理数据分组的网络层协议头部分，根据分组头中的目的IP地址，路由器在路由表中查找下一跳的IP地址；路由器将数据分组头部中的TTL字段值减1，并重新计算校验和；根据已确定的下一跳的IP地址，将IP数据分组送往相应的输出端口，进而封装上相应的数据链路层协议头部，最后由网络物理接口发送出去。

### 5.4.2 路由器的分类

路由器产品有多种分类方法，常见的分类方式如下。

**1. 按性能档次划分**

按性能档次可将路由器划分为高、中、低档，主要的依据是背板交换速率和包转发速率（吞吐量）。通常高档路由器其背板交换能力能够达到上百 Tb/s，包转发能力能够达到上万 Mp/s；中档路由器背板交换能力能够达到几百 Gb/s，包转发能力可以达到几十 Mp/s；性能更低的路由器则可以看作低档路由器。当然这只是一种宏观上的划分标准，各厂家划分并不完全一致，这种标准也会随着技术的进步而不断变化。

事实上，仅以背板带宽作为划分依据也是不全面的，常常需要考虑多种因素，如端口数量、支持的协议种类多少、支持的传输介质类型、安全性、网络管理功能等。以华为公司的路由器产品为例，NE9000 系列路由器为高端路由器，AR3200、AR2200 和 AR1200 系列路由器为中低端路由器。

按转发包的速率，路由器可以划分为线速路由器和非线速路由器。线速（Line Speed）是指完全可以按传输介质带宽进行包的转发，没有间断及时延。一般地，线速路由器是高端路由器，具有较为理想的传输效率，能以介质速率进行分组转发；非线速路由器是中低端路由器。

**2. 按结构划分**

从结构上分为模块化路由器和非模块化路由器。模块化结构可以灵活地配置路由器，以适应企业不断增加的业务需求，非模块化的路由器就只能提供固定的端口。通常中高端路由器为模块化结构，低端路由器为非模块化结构。

**3. 按功能划分**

从功能上划分，可将路由器分为核心层（骨干级）路由器、分发层（企业级）路由器和访问层（接入级）路由器。

核心层路由器是实现企业级网络互联的关键设备，数据吞吐量大。对核心层路由器的基本性能要求是高速率和高可靠性。为了获得高可靠性，网络系统普遍采用诸如热备份、双电源、双数据通路等传统冗余技术，从而使得核心层路由器的可靠性一般不成问题。分发层路由器连接许多中小型网络，连接对象较多，但系统相对简单，且数据流量较小，对这类路由器的要求是以适中的价格实现尽可能多的网络互联，同时还要求能够支持不同的服务质量。访问层路由器主要应用于连接家庭或 ISP 内的小型企业客户群体。

**4. 按所处网络位置划分**

按路由器所处的网络位置，通常把路由器划分为边界路由器和中间结点路由器。边界路由器处于网络边缘，用于不同网络路由器的连接；中间结点路由器处于网络的中间，用于连接不同网络，起到一个数据转发的桥梁作用。由于各自所处的网络位置有所不同，其主要性能也就有相应的侧重，如中间结点路由器因为要面对各种各样的网络，需要识别这些网络当中的结点，选择中间结点路由器时就需要在 MAC 地址记忆功能上更加注重，选择缓存更大的路由器。边界路由器由于它可能要同时接收来自许多不同网络路由器发来的数据，要求边界路由器的背板带宽要足够大。

## 5.4.3 第三层交换

**1. 第三层交换的基本概念**

20 世纪 90 年代中期，网络设备制造商提出了“第三层交换”的概念。简单地说，第三层交换技术是二层交换技术与三层路由技术的结合。目标是实现快速转发分组，保证 QoS 服

务质量,提高交换结点的稳定性及可靠性等。实现了第三层交换技术的交换机称为第三层交换机。

在第三层交换技术的发展过程中,不少公司都提出了自己的交换技术,主要有以下几种:Ipsilon公司首倡的Ipsilon IP交换技术,这也是最早的第三层交换技术,该技术通过识别数据包流,使得数据包尽量在第二层进行交换,以绕过路由器,改善网络性能;Cisco公司研发的Cisco标签交换技术,通过给数据包贴上标签实现,此标签在交换结点读出,判断包传送路径,该技术适用于大型网络和Internet;3Com公司提出的Fast IP技术,该技术侧重数据策略管理、优先原则和服务质量,保证实时音频或视频数据流能得到所需的带宽;IBM公司提出的ARIS技术(Aggregate Route-based IP Switching),该技术与Cisco的标签交换技术相似,包上附上标记,借以穿越交换网,一般用于ATM网络;Toshiba公司的信元交换技术,Cascade公司的IP导航器等也都是第三层交换技术的代表。当时,Cisco、3Com、北电网络、朗讯、Cabletron、Foundry和Extreme等公司都有比较成熟的第三层交换产品和模块推出。这些技术各有所长,都在一定程度上提高了IP分组的转发速率,改善了网络性能。

第三层交换技术是在市场需求和设备制造商的共同推动下产生发展起来的。那么,为什么有了集线器、网桥、二层交换机、路由器等设备后,还要研发第三层交换机呢?早期的局域网大都使用集线器将计算机连接在一起,所有的计算机共享同一个“冲突域”,因此在介质争用的过程中浪费了网络的共享带宽。

为解决冲突域问题,可以考虑使用网桥来分隔网段中的流量,网桥根据硬件地址过滤和转发数据帧,网桥建立了分离的冲突域。但是网桥存在“广播风暴”的问题,同时网桥让网络上的所有计算机共享同一个“广播域”。

为解决广播域的问题,引入了路由器,为互联网之间的数据转发提供了路由,路由器可以建立分离的广播域,根据分组报头中的IP地址决定分组的转发路径,但是路由器处理数据分组的速率较低,因为路由器需要使用软件来实现处理功能,不同的数据分组经由不同的路由器所需要的处理时间也大不相同。

在网桥的基础上,结合硬件交换技术,人们制造出了第二层交换机(简称为交换机)。第二层交换机实现了网桥的基本功能,同时提高了传输性能。同样地,人们考虑能否将硬件交换技术与路由器相结合,研发第三层交换机,使之具有路由器和交换机共同的优点。第三层交换机本质上是一种用硬件实现了的高速路由器,工作于网络层,依据IP地址实现数据分组的转发。较之一般的路由器,第三层交换机转发数据分组的速率较高,但实现的额外功能要比路由器少,工作时不如路由器灵活、容易控制。

第三层交换机通常可以提供如下功能。

(1) 分组转发。分组转发是第三层交换机的主要功能,第三层交换机依据目的IP地址决定转发路径。

(2) 路由处理。第三层交换机通过内部路由选择协议(如RIP、OSPF)创建并维护路由表。

(3) 安全服务。第三层交换机也会提供一些简单的安全服务,如防火墙、分组过滤等服务功能。

除此之外,第三层交换机一般还可以实现流量工程、拥塞控制等额外功能。

**2. 第三层交换机的工作原理**

一个具有三层交换功能的设备，是一个带有第三层路由功能的第二层交换机，但它是二者的有机结合，并不是简单地把路由器设备的硬件及软件叠加在局域网交换机上。与二层交换机类似，三层交换机也需要建立 MAC 地址转发表，记录 MAC 地址与端口的映射关系。

假设发送站 A 要向目的站 B 发送信息，三层交换机的工作原理如下。

(1) 发送站 A 在开始发送时，把自己的 IP 地址与目的站 B 的 IP 地址进行比较，判断目的站 B 是否与自己在同一子网内。

(2) 若目的站 B 与发送站 A 在同一子网内，则进行二层转发。

(3) 若两个站不在同一子网内，发送站 A 要向“默认网关”发出 ARP 请求包，而“默认网关”的 IP 地址其实是三层交换机的三层交换模块；如果三层交换模块在以前的通信过程中已经知道目的站 B 的 MAC 地址，则向发送站 A 回复目的站 B 的 MAC 地址，否则三层交换模块根据路由信息向目的站 B 广播一个 ARP 请求，目的站 B 得到此 ARP 请求后向三层交换模块回复其 MAC 地址，三层交换模块保存此 MAC 地址并回复给发送站 A，同时将目的站 B 的 MAC 地址发送到二层交换引擎的 MAC 地址表中。

(4) 自此之后，发送站 A 向目的站 B 发送的数据包便全部交给二层交换处理，信息得以高速传输。

由此可以看出，第三层交换机在工作时是“一次路由，处处交换”。由于仅仅在路由过程中才需要三层处理，绝大部分数据都通过二层交换转发，因此三层交换机的速度很快，接近二层交换机的速度。

特别指出，第三层交换技术在不同时期、不同公司有着不同的实现，如 Cisco 公司的 MPLS 技术与 IBM 的 ARIS 技术在实现细节上就有所不同；即便同是 Cisco 公司的三层交换技术(有早期的标签交换 Tag Switching 和后期的多协议标签交换 MPLS)，在实现细节上也有所区别。所以，这里只是介绍第三层交换的一般性原理，具体细节就不进行详细讨论了。

**3. 第三层交换机的应用**

在一些需要高速转发而对网络管理和安全又有很高要求的场合，使用第三层交换机是最好的选择，如一些重要部门的内部主干网络。然而，当需要接入 Internet 并对网络进行必要的控制管理时，路由器仍是最好的选择。

图 5-26 给出了一个以路由器为核心的网络结构示意图。从图 5-26 中可以看到，路由器在这个网络中的地位非常重要，不但担负着接入 Internet 的任务，同时也担负着内网通信的任务。路由器不仅仅是连接的中心点，也是全网的瓶颈。如果路由器出现故障或者路由器的性能不良，不但会影响到全网的 Internet 接入，还会影响到内网之间的通信。

为了解决这样的问题，可以考虑在主干结点部分增加一个第三层交换机，如图 5-27 所示。采用这种连接方式可以有效提高全网的性能。因为第三层交换机承担了主要的内网通信任务，只有当需要接入 Internet 时，才使用到路由器。也就是说，第三层交换机主要负责了内网的通信工作，而路由器主要负责与外网的通信。通过这样合理地分工，提高了整个系统的运作效率。

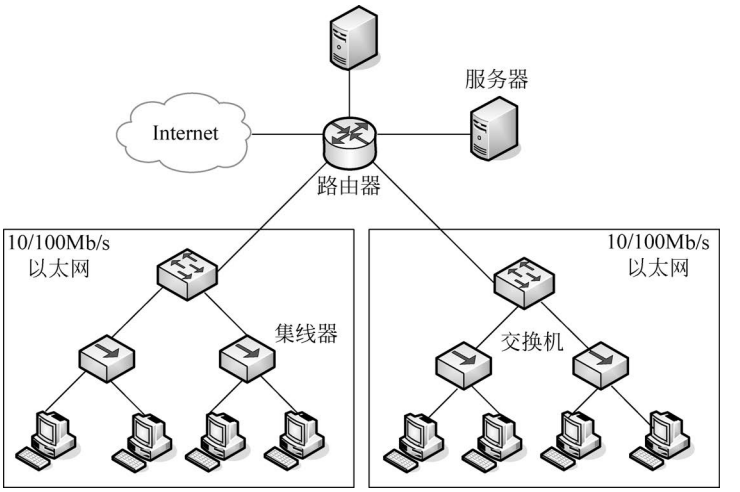



图 5-26　以路由器为核心的网络结构示意图

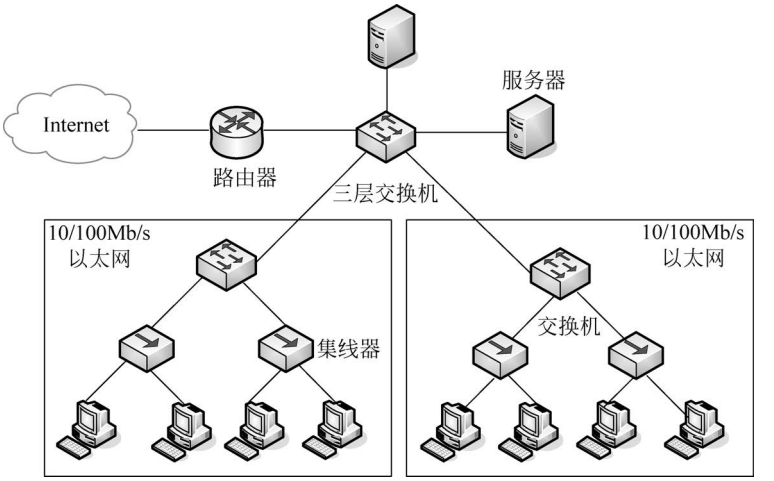



图 5-27　第三层交换机加入主干结点的网络结构

## 5.5　IP 多播与 IGMP 协议

IP 多播(Multicast,曾译为组播)已成为当今网络应用领域中的一个热点话题。IP 多播的概念首次于 1988 年提出,1992 年 3 月 IETF 在 Internet 范围内尝试了会议声音的多播,当时一共有 20 个结点可以同时接收到会议的声音。IP 多播在网络电话、视频会议、股市行情、网络教学等领域中都有应用。

### 5.5.1　IP 多播的基本概念

IP 多播是指由一个源点发送数据,多个终点同时接收数据的通信方式,即一对多的通信。在 Internet 中,实现一对多通信可以有两种方式:一种是由源结点采用点对点的方式,依次向目标结点发送同一分组;另一种方法就是采用多播。采用多播方式可以节省网络资

源，简化传输过程。能够运行多播协议的路由器称为多播路由器(Multicast Router)。多播路由器可以是单独的路由器，也可以是运行多播软件的普通路由器。所有的多播路由器都同时可以转发普通的单播 IP 数据报。

IP 多播实现了一种高效的一点对多点传输的工作模式，主要包括以下几方面的内容。

(1) 定义了一个组地址，每个组代表了一个或多个发送者与一个或多个接收者的一个会话(Session)。

(2) 接收者可以自主地选择自己所希望加入或者退出的多播组，可以用多播地址通知相应的多播路由器来实现。

(3) 发送者使用多播地址作为目的地址以发送分组，发送者不需要了解接收者的位置与状态等信息。

(4) 多播路由器建立一棵从发送者分支出去的多播路由传递树，这棵树延伸至所有的 IP 多播成员涉及的网络中。利用这棵树，多播路由器将多播数据分组转发至所有的相关网络中。

**1. IP 多播地址**

Internet 中的主机都有一个全球唯一的 IP 地址。那么，当主机需要加入多播组工作时，应该使用什么样的地址呢?

D 类 IP 地址是为 IP 多播专门定义的。每个多播地址都会在 224.0.0.0 到 239.255.255.255 之间。每一个 D 类 IP 地址代表一个多播组，如此一来，D 类地址一共可以表示 $2^{28}$ 个多播组。不过，其中也有一些地址被保留用于一些特殊的用途，如 224.0.0.1 用来表示在本子网上所有加入多播的主机，224.0.0.2 用来表示在本子网上所有加入多播的路由器，224.0.0.3 没有进行指派，224.0.0.11 用来当作移动代理的地址，224.0.1.1～224.0.1.18 的地址预留给电视会议等多播应用。完整的多播地址表可以从 IANA 授权的网站(http://www.iana.org/numbers)上获取。

显然，多播地址只能用作目的地址，而不能用作源地址。在多播数据分组的目的地址字段写入多播地址(而不是每一单个主机的 IP 地址)，然后设法让加入到这个多播组的主机 IP 地址与多播地址相关联。

**2. IP 多播的工作过程**

多播数据报也是采用“尽最大努力交付”的原则，在传输过程中，并不保证能够将多播数据报交付给多播组内的所有成员。在多播数据报的传输过程中，不产生 ICMP 差错报文。因此，若在 PING 命令后面输入多播地址，将永远不会收到响应。

图 5-28 给出了 IP 多播的工作过程。主机 A 需要发送多播数据报时，只需要发送一个多播数据报，多播路由器 1 进行复制后，发送至多播路由器 2 和多播路由器 3；多播路由器 2 和多播路由器 3 接收后再分别进行复制，将多播数据报发送至多播路由器 4、多播路由器 5 和多播路由器 6；进而将多播数据报交付给所有的多播组目标主机。

IP 多播在工作时需要运行两种不同的协议，即网际组管理协议(IGMP)和多播路由选择协议。图 5-29 所示为 IGMP 协议的工作过程。图 5-29 中，标有 IP 地址的 4 台主机都参加了多播组 230.0.0.1，多播数据报应传送到路由器 R1、R3 和 R4，而不应传送到路由器 R2，因为与 R2 连接的局域网上目前没有该多播组的成员。而 IGMP 协议的作用就是要让这些路由器知道它们所连接的网络上有没有多播组的成员。

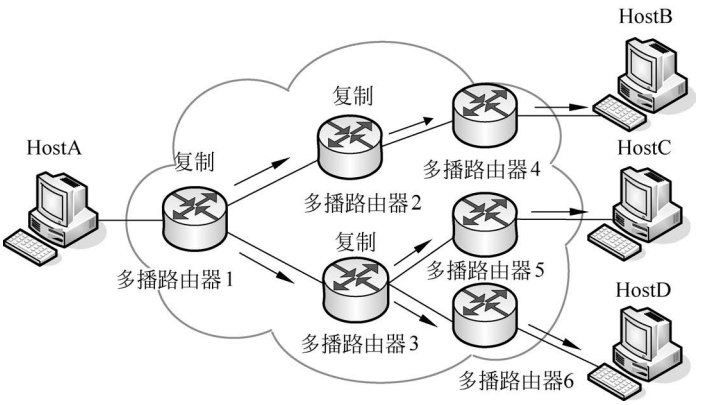



图 5-28　IP 多播的工作过程

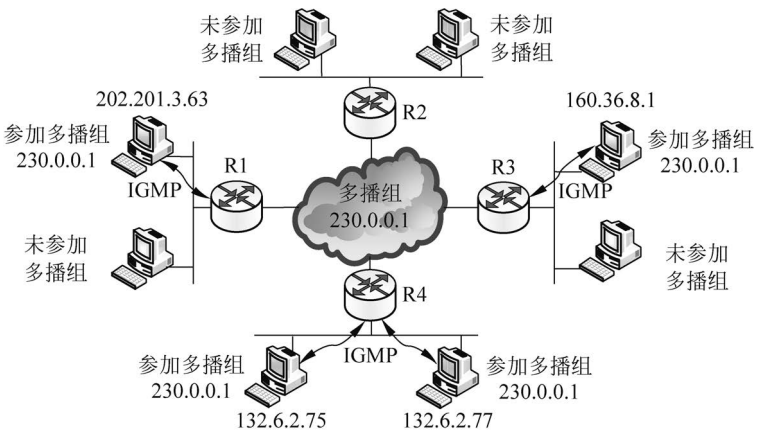



图 5-29　IGMP 协议的工作过程

由此可以看出,加入多播组 230.0.0.1 的成员有{R1[202.201.3.63],R3[160.36.8.1],R4[132.6.2.75,132.6.2.77]}。此处需要注意的是,IGMP 并不知道 IP 多播组所包含的成员的具体数量,也不知道这些成员分布在哪些网络上,IGMP 协议只是让连接在本地局域网上的多播路由器知道本网络上是否有加入多播组的主机(更具体地说,是主机上的某个进程)。当某台主机加入了新的多播组时,该主机会向多播组的多播地址发送一个 IGMP 报文,声明自己要成为该多播组的成员。本地的多播路由器收到 IGMP 报文后,还要利用多播路由选择协议(如距离矢量多播路由选择协议 DVMRP),把该组的成员关系转发给互联网上的其他多播路由器。

多播数据报的发送者和接收者永远都不知道(也无法找出)一个多播组的成员有多少个,以及这些成员是哪些主机。互联网中的路由器和主机都不知道哪个应用程序进程将要向哪个多播组发送多播数据报,因为任何一个应用程序进程都可以在任何时刻向任何一个多播组发送多播数据报,而这个应用程序进程并不需要加入这个多播组。

仅有 IGMP 协议仍然无法完成多播任务,连接在局域网上的多播路由器还必须和互联网上的其他多播路由器协同工作,以便把多播数据报用最小的代价传送给所有的组成员,这就需要使用多播路由选择协议[常见的有距离矢量多播路由协议(Distance Vector Multicast Routing Protocol,DVMRP)、多播开放最短路径优先协议(Multicast Open Shortest Path

First，MOSPF）和密集模式独立多播协议（Protocol-Independent Multicast-Dense Mode，PIM-DM）等]。多播路由选择协议要比单播路由选择协议复杂得多，这是因为多播的转发必须动态地适应多播组成员的动态变化（而此时网络拓扑并未发生改变）。与多播路由选择协议不同，单播路由选择协议通常在网络拓扑发生变化时才会更新路由。目前，还没有在整个互联网范围内使用的多播路由选择协议，多播路由选择协议也尚未标准化。

为了适应交互式音频和视频信息的多播，从 1992 年起，在 Internet 上开始试验虚拟的多播主干网（Multicast Backbone On the Internet，Mbone）。Mbone 可以将分组发送到不在一个网络，但属于同一个多播组的多个主机。现在多播主干网的规模已经很大，拥有几千台多播路由器。

## 5.5.2 以太网物理多播

在网络层中，IP 多播数据报可以借助 D 类 IP 地址实现。但是，由于多播数据报使用的是多播 IP 地址，ARP 协议将无法找出相对应的物理地址，而在数据链路层转发数据报时需要用到物理地址。现在大部分主机都是通过局域网接入到 Internet 中的，在 Internet 进行多播的最后阶段，需要使用硬件多播将多播数据报交付给多播组的所有成员。

以太网支持物理多播编址。以太网的物理地址（MAC）长度为 6B（48 位），若 MAC 地址中的前 25 位是 00000001000000000010111100，则这个地址定义 TCP/IP 协议中的多播地址，剩下的 23 位用来定义一个多播组。要把 IP 多播地址转换为以太网地址，多播路由器需要提取 D 类 IP 地址中的最低的 23 位，将它们放到多播以太网物理地址中，如图 5-30 所示。由此可以看到，以太网多播物理地址块的范围为 01-00-5E-00-00-00～01-00-5E-7F-FF-FF。

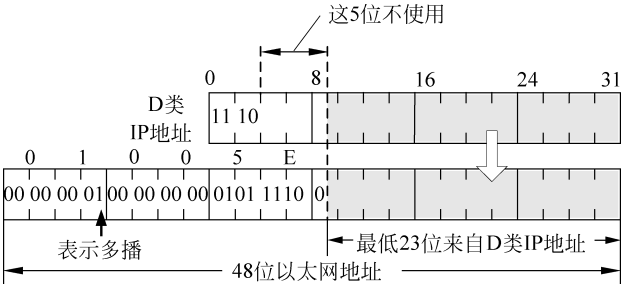



图 5-30 D 类 IP 地址映射为以太网物理地址

在这个映射过程中，有一点问题需要指出：D 类 IP 地址的可用长度为 28 位（前 4 位为固定的 1110），而每一个物理多播地址中只有 23 位可用作多播，这意味着有 5 位没有使用。也就是说，将会有 32（$2^5$）个 IP 多播地址映射为单个的物理多播地址。整个的映射过程是多对一的，而并非一对一。例如，IP 多播地址 224.128.60.5（E0-80-3C-05）和 224.0.60.5（E0-0-3C-05）转换为物理多播地址后都是 01-00-5E-00-3C-05。由于映射关系的不唯一性，因此收到多播数据报的主机还需要在网络层利用软件进行过滤，把不是本主机要接收的数据报丢弃。当加入多播的位于以太网中的主机接收到基于转换后的 MAC 地址为目的地址的帧后，由于在加入多播组时链路层已经被通知接收该目的 MAC 地址的帧，所以该帧就会被主机的数据链路接收并处理。

### 5.5.3 Internet 组管理协议

Internet 组管理协议(Internet Group Management Protocol,IGMP)是在多播环境下使用的协议,正在发挥着越来越重要的作用。IGMP 用来帮助多播路由器识别加入一个多播组的成员主机。IGMP 协议共有 3 个版本:1989 年公布的 RFC1112 文档是 IGMP 的第 1 个版本;1997 年公布的 RFC2236 文档是 IGMP 协议的第 2 个版本,已成为 Internet 的建议标准协议;2002 年 10 月公布的 RFC3376 文档是 IGMP 协议的第 3 个版本,宣布 RFC2236(IGMPv2)是陈旧的版本。

类似于 ICMP 协议,IGMP 也被当作 IP 协议的一部分,IGMP 使用 IP 数据报传递其报文,IGMP 报文通过 IP 首部中协议字段值为 2 来指明。IGMP 有固定的报文长度(8B),没有可选数据项。图 5-31 表示了 IGMPv1 的报文格式。

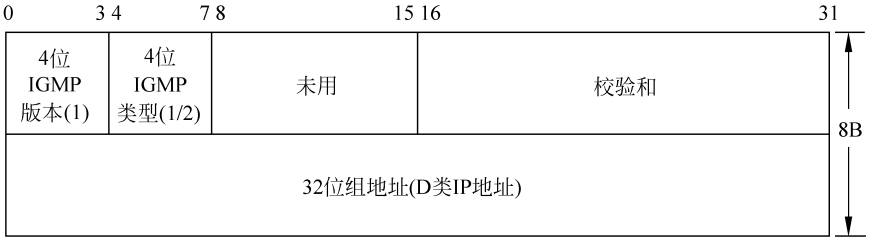



图 5-31 IGMPv1 的报文格式

IGMP 协议执行过程可以分为以下两个阶段。

(1) 当有主机加入新的多播组时,该主机按照多播组的多播地址发送一个 IGMP 报文,声明自己想加入该多播组。本地多播路由器接收到 IGMP 报文后,将组成员关系转发给 Internet 上的其他多播路由器。

(2) 由于组成员关系是动态变化的,本地多播路由器需要周期性地探询本地网络上的主机,以确认这些主机是否还继续是组成员。只要有一个主机对某个组响应,多播路由器就认为这个组是活跃的。但一个组在经过几次探询后仍然没有一个主机响应,多播路由器就认为本网络上的主机都已经离开了这个组,因此也就不再将该组的成员关系转发给其他的多播路由器。

多播路由器在探询组成员关系时,只需要对所有的组发送一个请求信息的询问报文,而不需要对每一个组发送一个询问报文。默认的询问频率是每 125s 发送一次。当同一个网络上连接有多个多播路由器时,它们能够迅速且有效地选择其中的一个来探询主机间的成员关系。因此,网络上的多个多播路由器或者多个多播组不会带来过大的通信量。

在 IGMP 的询问报文中有一个数值 $N$,它指明一个最长的响应时间(默认值为 10s)。当收到询问时,主机在 0～$N$ 之间随机选择一个数值作为发送响应报文的时延。因此,若一个主机同时加入了几个多播组,则主机对每一个多播组选择不同的随机数,对应于最小时延的响应报文最先被发送。

同一个组内的每一个主机都要监听响应,只要有本组的其他主机先发送了响应,自己就不必再发送响应了。这样可以有效地控制通信量。

多播数据报的发送者和接收者都不知道一个多播组的成员数有多少,以及这些成员是