



“十二五”普通高等教育本科国家级规划教材

计 算 机 网 络

(第 7 版)

谢希仁 编著

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书自 1989 年首次出版以来, 曾于 1994 年、1999 年、2003 年、2008 年和 2013 年分别出了修订版。在 2006 年本书通过了教育部的评审, 被纳入普通高等教育“十一五”国家级规划教材; 2008 年出版的第 5 版获得了教育部 2009 年精品教材称号。2013 年出版的第 6 版是“十二五”普通高等教育本科国家级规划教材。现在的第 7 版又在第 6 版的基础上进行了一些修订。

全书分为 9 章, 比较全面系统地介绍了计算机网络的发展和原理体系结构、物理层、数据链路层(包括局域网)、网络层、运输层、应用层、网络安全、互联网上的音频/视频服务, 以及无线网络和移动网络等内容。各章均附有习题(附录 A 给出了部分习题的答案和提示)。全书课件(PowerPoint 文件)放在电子工业出版社悦学多媒体课程资源平台上(http://yx.51zhy.cn/mtcrsRes/phei_cnetwork.jsp), 供读者下载参考。

本书的特点是概念准确、论述严谨、内容新颖、图文并茂, 突出基本原理和基本概念的阐述, 同时力图反映计算机网络的一些最新发展。本书可供电气信息类和计算机类专业的大学本科生和研究生使用, 对从事计算机网络工作的工程技术人员也有参考价值。

未经许可, 不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有, 侵权必究。

图书在版编目(CIP)数据

计算机网络 / 谢希仁编著. —7 版. —北京: 电子工业出版社, 2017.1

“十二五”普通高等教育本科国家级规划教材

ISBN 978-7-121-30295-4

I. ①计… II. ①谢… III. ①计算机网络—高等学校—教材 IV. ①TP393

中国版本图书馆 CIP 数据核字(2016)第 269601 号

策划编辑: 郝志恒

责任编辑: 牛晓丽

印 刷: 三河市华成印务有限公司

装 订: 三河市华成印务有限公司

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱

邮编: 100036

开 本: 787×1092 1/16

印张: 29

字数: 742.4 千字

版 次: 1999 年 4 月第 2 版

2017 年 1 月第 7 版

印 次: 2017 年 1 月第 1 次印刷

定 价: 45.00 元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888, 88258888。

质量投诉请发邮件至 zltts@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式: QQ 9616328。

第4章 网络层

本章讨论网络互连问题。在介绍网络层提供的两种不同服务后，就进入本章的核心内容——网际协议 IP，这是本书的一个重点内容。只有深入地掌握了 IP 协议的主要内容，才能理解互联网是怎样工作的。本章还要讨论网际控制报文协议 ICMP，几种常用的路由选择协议，IPv6 的主要特点，IP 多播的概念。在讨论虚拟专用网 VPN 和网络地址转换 NAT 后，最后简单介绍多协议标记交换 MPLS。

本章最重要的内容是：

- (1) 虚拟互连网络的概念。
- (2) IP 地址与物理地址的关系。
- (3) 传统的分类的 IP 地址（包括子网掩码）和无分类域间路由选择 CIDR。
- (4) 路由选择协议的工作原理。

4.1 网络层提供的两种服务

在计算机网络领域，网络层应该向运输层提供怎样的服务（“面向连接”还是“无连接”）曾引起了长期的争论。争论焦点的实质就是：在计算机通信中，可靠交付应当由谁来负责？是网络还是端系统？

有些人认为应当借助于电信网的成功经验，让网络负责可靠交付。大家知道，传统电信网的主要业务是提供电话服务。电信网使用昂贵的程控交换机（其软件也非常复杂），用**面向连接**的通信方式，使电信网络能够向用户（实际上就是电话机）提供可靠传输的服务。因此他们认为，计算机网络也应模仿打电话所使用的面向连接的通信方式。当两台计算机进行通信时，也应当先建立连接（但在分组交换中是建立一条**虚电路 VC (Virtual Circuit)**^①），以预留双方通信所需的一切网络资源。然后双方就沿着已建立的虚电路发送分组。这样的分组的首部不需要填写完整的目的主机地址，而只需要填写这条虚电路的编号（一个不大的整数），因而减少了分组的开销。这种通信方式如果再使用可靠传输的网络协议，就可使所发送的分组无差错按序到达终点，当然也不丢失、不重复。在通信结束后要释放建立的虚电路。图 4-1(a)是网络提供虚电路服务的示意图。主机 H_1 和 H_2 之间交换的分组都必须在事先建立的虚电路上传送。

但互联网的先驱者却提出一种崭新的网络设计思路。他们认为，电信网提供的端到端可靠传输的服务对电话业务无疑是很合适的，因为电信网的终端（电话机）非常简单，没有智能，也没有差错处理能力。因此电信网必须负责把用户电话机产生的话音信号可靠地传送

① 注：虚电路表示这只是一条逻辑上的连接，分组都沿着这条逻辑连接按照存储转发方式传送，而并不是真正建立了一条物理连接。请注意，电路交换的电话通信是先建立了一条真正的连接。因此分组交换的虚连接和电路交换的连接只是类似，但并不完全一样。

到对方的电话机，使还原后的话音质量符合技术规范的要求。但计算机网络的端系统是有智能的计算机。计算机有很强的差错处理能力（这点和传统的电话机有本质上的差别）。因此，互联网在设计上就采用了和电信网完全不同的思路。

互联网采用的设计思路是这样的：**网络层向上只提供简单灵活的、无连接的、尽最大努力交付的数据报服务^①**。这里的“数据报”(datagram)是互联网的设计者最初使用的名词，其实数据报（或 IP 数据报）就是我们经常使用的“分组”。在本书中，数据报和分组是同义词，可以混用。

网络在发送分组时不需要先建立连接。每一个分组（也就是 IP 数据报）独立发送，与其前后的分组无关（不进行编号）。**网络层不提供服务质量的承诺**。也就是说，所传送的分组可能出错、丢失、重复和失序（即不按序到达终点），当然也不保证分组交付的时限。由于传输网络不提供端到端的可靠传输服务，这就使网络中的路由器比较简单，且价格低廉（与电信网的交换机相比较）。如果主机（即端系统）中的进程之间的通信需要是可靠的，那么就由网络的主机中的运输层负责（包括差错处理、流量控制等）。采用这种设计思路的好处是：网络造价大大降低，运行方式灵活，能够适应多种应用。互联网能够发展到今日的规模，充分证明了当初采用这种设计思路的正确性。

图 4-1(b)给出了网络提供数据报服务的示意图。主机 H_1 向 H_2 发送的分组各自独立地选择路由，并且在传送的过程中还可能丢失。

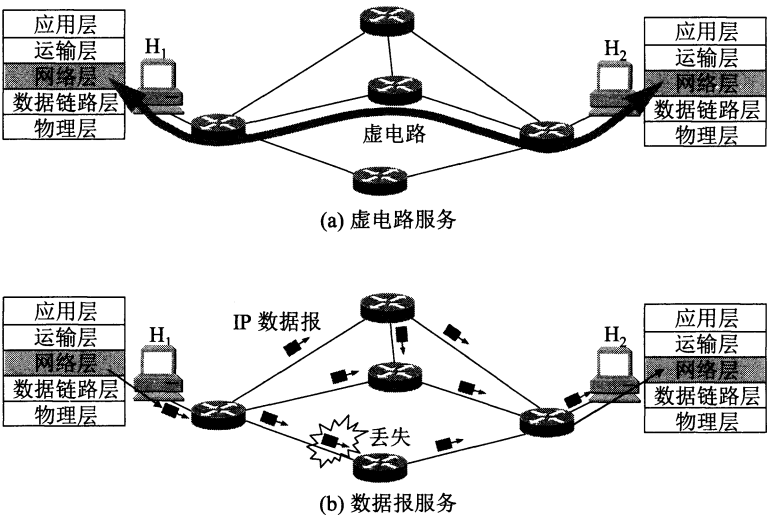


图 4-1 网络层提供的两种服务

OSI 体系的支持者曾极力主张在网络层使用可靠传输的虚电路服务，也曾推出过网络层虚电路服务的著名标准——ITU-T 的 X.25 建议书。但现在 X.25 早已成为历史了。

^① 注：尽最大努力交付(best effort delivery)虽然并不表示路由器可以任意丢弃分组，但在网络层上的这种交付实质上就是不可靠交付。顺便提一下，文献中也常使用“尽力而为”的译名。这个译名固然较为简洁，但似不够准确。

表 4-1 归纳了虚电路服务与数据报服务的主要区别。

表 4-1 虚电路服务与数据报服务的对比

对比的方面	虚电路服务	数据报服务
思路	可靠通信应当由网络来保证	可靠通信应当由用户主机来保证
连接的建立	必须有	不需要
终点地址	仅在连接建立阶段使用，每个分组使用短的虚电路号	每个分组都有终点的完整地址
分组的转发	属于同一条虚电路的分组均按照同一路由进行转发	每个分组独立选择路由进行转发
当结点出故障时	所有通过出故障的结点的虚电路均不能工作	出故障的结点可能会丢失分组，一些路由可能会发生变化
分组的顺序	总是按发送顺序到达终点	到达终点的时间不一定按发送顺序
端到端的差错处理和流量控制	可以由网络负责，也可以由用户主机负责	由用户主机负责

鉴于 TCP/IP 体系的网络层提供的是数据报服务，因此下面我们的讨论都是围绕网络层如何传送 IP 数据报这个主题。

4.2 网际协议 IP

网际协议 IP 是 TCP/IP 体系中两个最主要的协议之一[STEV94][COME06][FORO10]，也是最重要的互联网标准协议之一。网际协议 IP 又称为 Kahn-Cerf 协议，因为这个重要协议正是 Robert Kahn 和 Vint Cerf 二人共同研发的。这两位学者在 2005 年获得图灵奖（其地位相当于计算机科学领域的诺贝尔奖）。严格来说，这里所讲的 IP 其实是 IP 的第 4 个版本，应记为 IPv4。但在讲述 IP 协议的各种原理时，往往不在 IP 后面加上版本号。在后面的 4.6 节我们再介绍较新的版本 IPv6（版本 1~3 和版本 5 都未曾使用过）。

与 IP 协议配套使用的还有三个协议：

- 地址解析协议 ARP (Address Resolution Protocol)
- 网际控制报文协议 ICMP (Internet Control Message Protocol)
- 网际组管理协议 IGMP (Internet Group Management Protocol)

本来还有一个协议叫做逆地址解析协议 RARP (Reverse Address Resolution Protocol)，是和 ARP 协议配合使用的。但现在已被淘汰不使用了。

图 4-2 画出了这三个协议和网际协议 IP 的关系。在这一层中，ARP 画在最下面，因为 IP 经常要使用这个协议。ICMP 和 IGMP 画在这一层的上部，因为它们要使用 IP 协议。这三个协议将在后面陆续介绍。由于网际协议 IP 是用来使互连起来的许多计算机网络能够进行通信的，因此 TCP/IP 体系中的网络层常常被称为网际层(internet layer)，或 IP 层。使用“网际层”这个名词的好处是强调

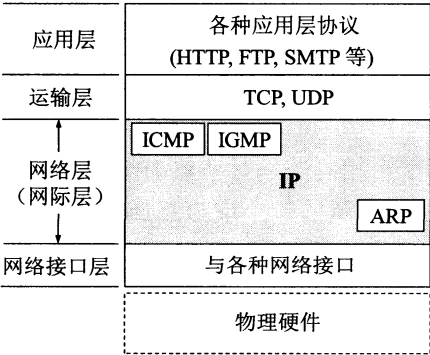


图 4-2 网际协议 IP 及其配套协议

这是由很多网络构成的互连网络。

在讨论网际协议 IP 之前，必须了解什么是虚拟互连网络。

4.2.1 虚拟互连网络

我们知道，如果要在全世界范围内把数以百万计的网络都互连起来，并且能够互相通信，那么这样的任务一定非常复杂。其中会遇到许多需要解决的问题，如：

- 不同的寻址方案；
- 不同的最大分组长度；
- 不同的网络接入机制；
- 不同的超时控制；
- 不同的差错恢复方法；
- 不同的状态报告方法；
- 不同的路由选择技术；
- 不同的用户接入控制；
- 不同的服务（面向连接服务和无连接服务）；
- 不同的管理与控制方式；等等。

能不能让大家都使用相同的网络，这样可使网络互连变得比较简单。答案是不行的。因为用户的需求是多种多样的，没有一种单一的网络能够适应所有用户的需求。另外，网络技术是不断发展的，网络的制造厂家也要经常推出新的网络，在竞争中求生存。因此在市场上总是有很多种不同性能、不同网络协议的网络，供不同的用户选用。

从一般的概念来讲，将网络互相连接起来要使用一些中间设备。根据中间设备所在的层次，可以有以下四种不同的中间设备：

- (1) 物理层使用的中间设备叫做转发器(repeater)。
- (2) 数据链路层使用的中间设备叫做网桥或桥接器(bridge)。
- (3) 网络层使用的中间设备叫做路由器(router)^①。
- (4) 在网络层以上使用的中间设备叫做网关(gateway)。用网关连接两个不兼容的系统需要在高层进行协议的转换。

当中间设备是转发器或网桥时，这仅仅是把一个网络扩大了，而从网络层的角度看，这仍然是一个网络，一般并不称之为网络互连。网关由于比较复杂，目前使用得较少。因此现在我们讨论网络互连时，都是指用路由器进行网络互连和路由选择。路由器其实就是一台专用计算机，用来在互联网中进行路由选择。由于历史的原因，许多有关 TCP/IP 的文献曾经把网络层使用的路由器称为网关（本书有时也这样用），对此请读者加以注意。

图 4-3(a)表示有许多计算机网络通过一些路由器进行互连。由于参加互连的计算机网络都使用相同的网际协议 IP (Internet Protocol)，因此可以把互连以后的计算机网络看成如图 4-3(b)所示的一个虚拟互连网络(internet)。所谓虚拟互连网络也就是逻辑互连网络，它的意思就是互连起来的各种物理网络的异构性本来是客观存在的，但是我们利用 IP 协议就可以

^① 注：还有一种网桥和路由器的混合物桥路由器(brouter)，它是兼有网桥和路由器的功能的产品。实际上，严格的网桥或严格的路由器产品是较少见的。不过桥路由器名词用得普遍。

使这些性能各异的网络在网络层上看起来好像是一个统一的网络。这种使用 IP 协议的虚拟互连网络可简称为 **IP 网**（IP 网是虚拟的，但平常不必每次都强调“虚拟”二字）。使用 IP 网的好处是：当 IP 网上的主机进行通信时，就好像在一个单个网络上通信一样，它们看不见互连的各网络的具体异构细节（如具体的编址方案、路由选择协议，等等）。如果在这种覆盖全球的 IP 网的上层使用 TCP 协议，那么就是现在的互联网(Internet)。

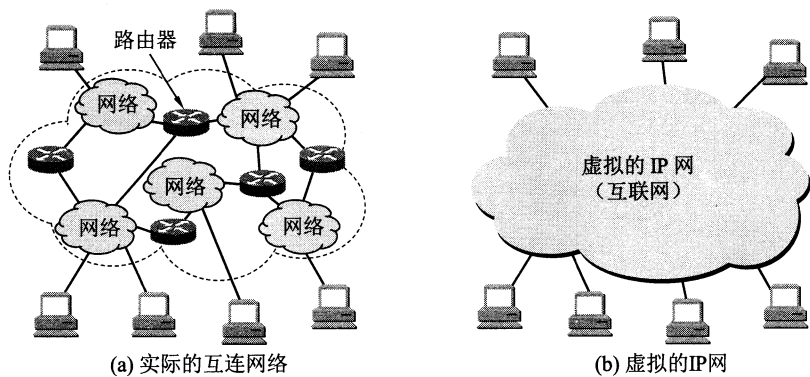


图 4-3 IP 网的概念

当很多异构网络通过路由器互连起来时，如果所有的网络都使用相同的 IP 协议，那么在网络层讨论问题就显得很方便。现在用一个例子来说明。

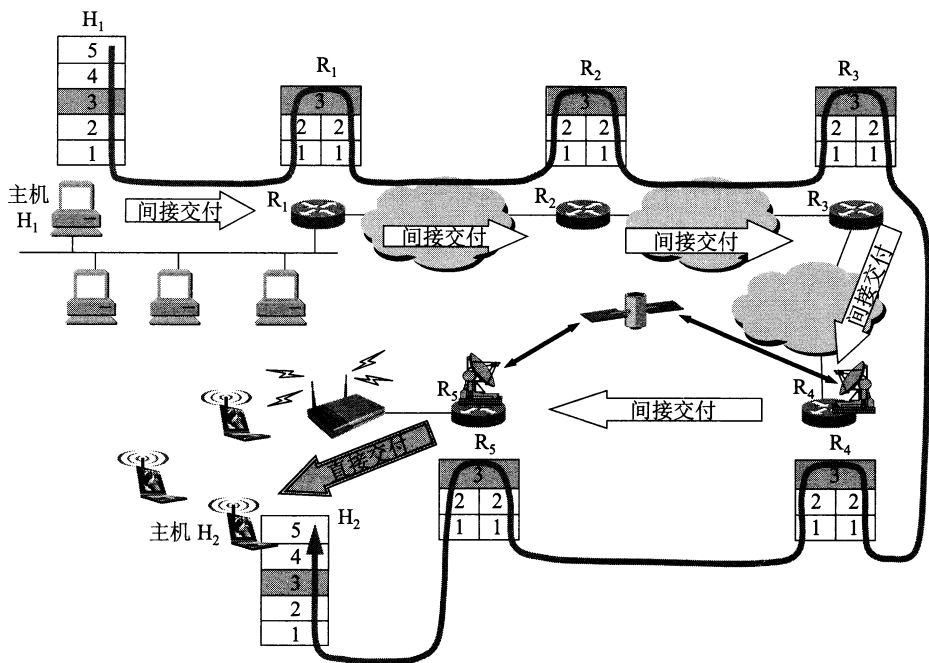
在图 4-4 所示的互联网中的源主机 H_1 要把一个 IP 数据报发送给目的主机 H_2 。根据第 1 章中讲过的分组交换的存储转发概念，主机 H_1 先要查找自己的路由表，看目的主机是否就在本网络上。如是，则不需要经过任何路由器而是**直接交付**，任务就完成了。如不是，则必须把 IP 数据报发送给某个路由器（图中的 R_1 ）。 R_1 在查找了自己的路由表^①后，知道应当把数据报转发给 R_2 进行**间接交付**。这样一直转发下去，最后由路由器 R_5 知道自己是和 H_2 连接在同一个网络上，不需要再使用别的路由器转发了，于是就把数据报**直接交付**目的主机 H_2 。图中画出了源主机、目的主机以及各路由器的协议栈。我们注意到，主机的协议栈共有五层，但路由器的协议栈只有下三层。图中还画出了数据在各协议栈中流动的方向（用黑色粗线表示）。我们还可注意到，在 R_4 和 R_5 之间使用了卫星链路，而 R_5 所连接的是个无线局域网。在 R_1 到 R_4 之间的三个网络则可以是任意类型的网络。总之，这里强调的是：**互联网可以由多种异构网络互连组成**。

如果我们只从网络层考虑问题，那么 IP 数据报就可以想象是在网络层中传送，其传送路径是：

$$H_1 \rightarrow R_1 \rightarrow R_2 \rightarrow R_3 \rightarrow R_4 \rightarrow R_5 \rightarrow H_2$$

这样就不必画出许多完整的协议栈，使问题的描述更加简单。
有了虚拟互连网络的概念后，我们再讨论在这样的虚拟网络上如何寻址。

^① 注：更准确些说应是转发表。路由表和转发表的区别见后面 4.5.5 节的讨论。



图中的协议栈中的数字 1~5 分别表示物理层、数据链路层、网络层、运输层和应用层

图 4-4 分组在互联网中的传送

4.2.2 分类的 IP 地址

在 TCP/IP 体系中，IP 地址是一个最基本的概念，一定要把它弄清楚。有关 IP 最重要的文档就是互联网的正式标准 RFC 791。

1. IP 地址及其表示方法

整个的互联网就是一个单一的、抽象的网络。IP 地址就是给互联网上的每一台主机（或路由器）的每一个接口分配一个在全世界范围内是唯一的 32 位的标识符。IP 地址的结构使我们可以互联网上很方便地进行寻址。IP 地址现在由互联网名字和数字分配机构 ICANN (Internet Corporation for Assigned Names and Numbers) 进行分配^①。

IP 地址的编址方法共经过了三个历史阶段。

- (1) 分类的 IP 地址。这是最基本的编址方法，在 1981 年就通过了相应的标准协议。
- (2) 子网的划分。这是对最基本的编址方法的改进，其标准 RFC 950 在 1985 年通过。
- (3) 构成超网。这是比较新的无分类编址方法。1993 年提出后很快就得到推广应用。

本节只讨论最基本的分类的 IP 地址。后两种方法将在 4.3 节中讨论。

所谓“分类的 IP 地址”就是将 IP 地址划分为若干个固定类，每一类地址都由两个固定长度的字段组成，其中第一个字段是网络号(net-id)，它标志主机（或路由器）所连接到的网络。一个网络号在整个互联网范围内必须是唯一的。第二个字段是主机号(host-id)，它标志

^①注：我国用户可向亚太网络信息中心 APNIC (Asia Pacific Network Information Center) 申请 IP 地址（要缴费）。

该主机（或路由器）。一台主机号在它前面的网络号所指明的网络范围内必须是唯一的。由此可见，一个 IP 地址在整个互联网范围内是唯一的。

这种两级的 IP 地址可以记为：

IP 地址 ::= { <网络号>, <主机号> }

(4-1)

式(4-1)中的符号“::=”表示“定义为”。图 4-5 给出了各种 IP 地址的网络号字段和主机号字段，这里 A 类、B 类和 C 类地址都是单播地址（一对一通信），是最常用的。

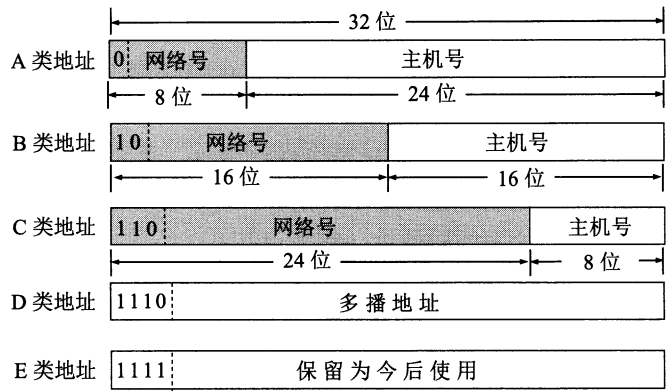


图 4-5 IP 地址中的网络号字段和主机号字段

从图 4-5 可以看出：

- A 类、B 类和 C 类地址的网络号字段（在图中这个字段是灰色的）分别为 1 个、2 个和 3 个字节长，而在网络号字段的最前面有 1~3 位的类别位，其数值分别规定为 0，10 和 110。
- A 类、B 类和 C 类地址的主机号字段分别为 3 个、2 个和 1 个字节长。
- D 类地址（前 4 位是 1110）用于多播（一对多通信）。我们将在 4.6 节讨论 IP 多播。
- E 类地址（前 4 位是 1111）保留为以后用。

这里要指出，由于近年来已经广泛使用无分类 IP 地址进行路由选择，A 类、B 类和 C 类地址的区分已成为历史[RFC 1812]，但由于很多文献和资料都还使用传统的分类的 IP 地址，而且从概念的演进上更清晰，因此我们在这里还要从分类的 IP 地址讲起。

从 IP 地址的结构来看，IP 地址并不仅仅指明一台主机，而是还指明了主机所连接到的网络。

把 IP 地址划分为 A 类、B 类、C 类三个类别，当初是这样考虑的。各种网络的差异很大，有的网络拥有很多主机，而有的网络上的主机则很少。把 IP 地址划分为 A 类、B 类和 C 类是为了更好地满足不同用户的要求。当某个单位申请到一个 IP 地址时，实际上是获得了具有同样网络号的一块地址。其中具体的各台主机号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即可。

对主机或路由器来说，IP 地址都是 32 位的二进制代码。为了提高可读性，我们常常把 32 位的 IP 地址中的每 8 位插入一个空格（但在机器中并没有这样的空格）。为了便于书写，可用其等效的十进制数字表示，并且在这些数字之间加上一个点。这就叫做点分十进制记法(dotted decimal notation)。图 4-6 是一个 B 类 IP 地址的表示方法。显然，128.11.3.31 比 10000000 00001011 00000011 00011111 书写起来要方便得多。

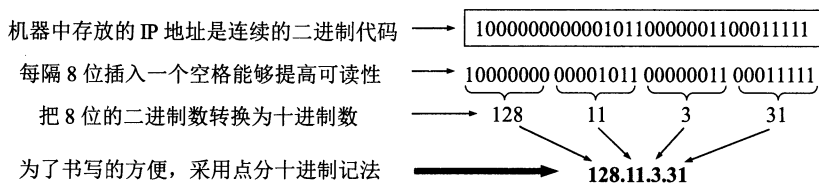


图 4-6 采用点分十进制记法能够提高可读性

2. 常用的三类别的 IP 地址

A 类地址的网络号字段占 1 个字节, 只有 7 位可供使用 (该字段的第一位已固定为 0), 但可指派的网络号是 126 个 (即 $2^7 - 2$)。减 2 的原因是: 第一, IP 地址中的全 0 表示“这个(this)”。网络号字段为全 0 的 IP 地址是个保留地址, 意思是“本网络”; 第二, 网络号为 127 (即 01111111) 保留作为本地软件环回测试(loopback test)本主机的进程之间的通信之用。若主机发送一个目的地址为环回地址 (例如 127.0.0.1) 的 IP 数据报, 则本主机中的协议软件就处理数据报中的数据, 而不会把数据报发送到任何网络。目的地址为环回地址的 IP 数据报永远不会出现在任何网络上, 因为网络号为 127 的地址根本不是一个网络地址。

A 类地址的主机号占 3 个字节, 因此每一个 A 类网络中的最大主机数是 $2^{24} - 2$, 即 16777214。这里减 2 的原因是: 全 0 的主机号字段表示该 IP 地址是“本主机”所连接到的单个网络地址 (例如, 一主机的 IP 地址为 5.6.7.8, 则该主机所在的网络地址就是 5.0.0.0), 而全 1 表示“所有的(all)”, 因此全 1 的主机号字段表示该网络上的所有主机^①。

IP 地址空间共有 2^{32} (即 4294967296) 个地址。整个 A 类地址空间共有 2^{31} 个地址, 占整个 IP 地址空间的 50%。

B 类地址的网络号字段有 2 个字节, 但前面两位 (1 0) 已经固定了, 只剩下 14 位可以进行分配。因为网络号字段后面的 14 位无论怎样取值也不可能出现使整个 2 字节的网络号字段成为全 0 或全 1, 因此这里不存在网络总数减 2 的问题。但实际上 B 类网络地址 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 [COME06]。因此 B 类地址可指派的网络数为 $2^{14} - 1$, 即 16383。B 类地址的每一个网络上的最大主机数是 $2^{16} - 2$, 即 65534。这里需要减 2 是因为要扣除全 0 和全 1 的主机号。整个 B 类地址空间共约有 2^{30} 个地址, 占整个 IP 地址空间的 25%。

C 类地址有 3 个字节的网络号字段, 最前面的 3 位是 (1 1 0), 还有 21 位可以进行分配。C 类网络地址 192.0.0.0 也是不指派的, 可以指派的 C 类最小网络地址是 192.0.1.0 [COME06], 因此 C 类地址可指派的网络总数是 $2^{21} - 1$, 即 2097151。每一个 C 类地址的最大主机数是 $2^8 - 2$, 即 254。整个 C 类地址空间共约有 2^{29} 个地址, 占整个 IP 地址的 12.5%。

这样, 我们就可得出表 4-2 所示的 IP 地址的指派范围。

^① 注: 关于全 1 和全 0 还可以再举两个例子。例如, B 类地址 128.7.255.255 表示“在网络 128.7.0.0 上的所有主机”。而 A 类地址 0.0.0.35 则表示“在这个网络上主机号为 35 的主机”。

表 4-2 IP 地址的指派范围

网络类别	最大可指派的网络数	第一个可指派的网络号	最后一个可指派的网络号	每个网络中的最大主机数
A	$126 (2^7 - 2)$	1	126	16777214
B	$16383 (2^{14} - 1)$	128.1	191.255	65534
C	$2097151 (2^{21} - 1)$	192.0.1	223.255.255	254

表 4-3 给出了一般不使用的特殊 IP 地址，这些地址只能在特定的情况下使用。

表 4-3 一般不使用的特殊 IP 地址

网络号	主机号	源地址使用	目的地址使用	代表的意义
0	0	可以	不可	在本网络上的本主机（见 6.6 节 DHCP 协议）
0	host-id	可以	不可	在本网络上的某台主机 host-id
全 1	全 1	不可	可以	只在本网络上进行广播（各路由器均不转发）
net-id	全 1	不可	可以	对 net-id 上的所有主机进行广播
127	非全 0 或全 1 的任何数	可以	可以	用于本地软件环回测试

IP 地址具有以下一些重要特点。

(1) 每一个 IP 地址都由网络号和主机号两部分组成。从这个意义上说，IP 地址是一种分等级的地址结构。分两个等级的好处是：第一，IP 地址管理机构在分配 IP 地址时只分配网络号（第一级），而剩下的主机号（第二级）则由得到该网络号的单位自行分配。这样就方便了 IP 地址的管理；第二，路由器仅根据目的主机所连接的网络号来转发分组（而不考虑目的主机号），这样就可以使路由表中的项目数大幅度减少，从而减小了路由表所占的存储空间以及查找路由表的时间。

(2) 实际上 IP 地址是标志一台主机（或路由器）和一条链路的接口。当一台主机同时连接到两个网络上时，该主机就必须同时具有两个相应的 IP 地址，其网络号必须是不同的。这种主机称为多归属主机(multihomed host)。由于一个路由器至少应当连接到两个网络，因此一个路由器至少应当有两个不同的 IP 地址。这好比一个建筑正好处在北京路和上海路的交叉口上，那么这个建筑就可以拥有两个门牌号码。例如，北京路 4 号和上海路 37 号。

(3) 按照互联网的观点，一个网络是指具有相同网络号 net-id 的主机的集合，因此，用转发器或网桥连接起来的若干个局域网仍为一个网络，因为这些局域网都具有同样的网络号。具有不同网络号的局域网必须使用路由器进行互连。

(4) 在 IP 地址中，所有分配到网络号的网络（不管是范围很小的局域网，还是可能覆盖很大地理范围的广域网）都是平等的。所谓平等，是指互联网同等对待每一个 IP 地址。

图 4-7 画出了三个局域网（LAN₁, LAN₂ 和 LAN₃）通过三个路由器（R₁, R₂ 和 R₃）互连起来所构成的一个互联网（此互联网用虚线圆角方框表示）。其中局域网 LAN₂ 是由两个网段通过网桥 B 互连的。图中的小圆圈表示需要有一个 IP 地址。

我们应当注意到：

- 在同一个局域网上的主机或路由器的 IP 地址中的网络号必须是一样的。图中所示的网络号就是 IP 地址中的网络号字段的值，这也是文献中常见的一种表示方法。另一种表示方法是用主机号为全 0 的网络 IP 地址。

- 用网桥（它只在链路层工作）互连的网段仍然是一个局域网，只能有一个网络号。
- 路由器总是具有两个或两个以上的 IP 地址。即路由器的每一个接口都有一个不同网络号的 IP 地址。
- 当两个路由器直接相连时（例如通过一条租用线路），在连线两端的接口处，可以分配也可以不分配 IP 地址。如分配了 IP 地址，则这一段连线就构成了一种只包含一段线路的特殊“网络”（如图中的 N₁、N₂ 和 N₃）。之所以叫做“网络”是因为它有 IP 地址。但为了节省 IP 地址资源，对于这种仅由一段连线构成的特殊“网络”，现在也常常不分配 IP 地址。通常把这样的特殊网络叫做无编号网络(unnumbered network)或无名网络(anonymous network)[COME06]。

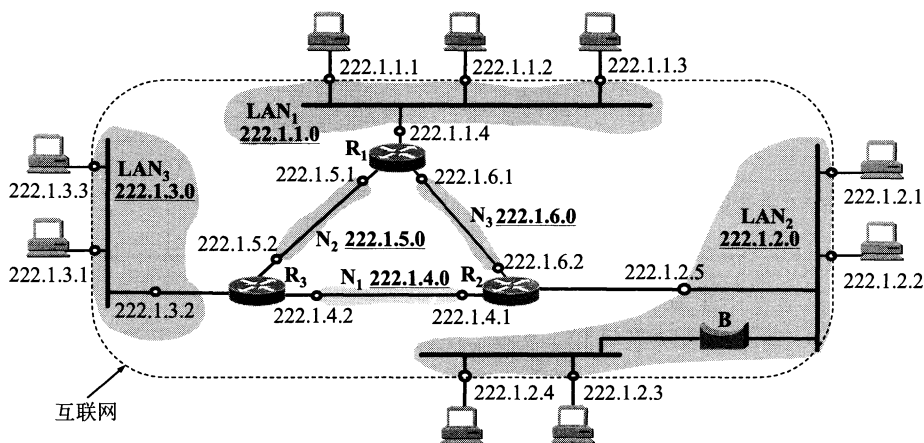


图 4-7 互联网中的 IP 地址

4.2.3 IP 地址与硬件地址

在学习 IP 地址时，很重要的一点就是要弄清主机的 IP 地址与硬件地址^①的区别。

图 4-8 说明了这两种地址的区别。从层次的角度看，物理地址是数据链路层和物理层使用的地址，而 IP 地址是网络层和以上各层使用的地址，是一种逻辑地址（称 IP 地址为逻辑地址是因为 IP 地址是用软件实现的）。

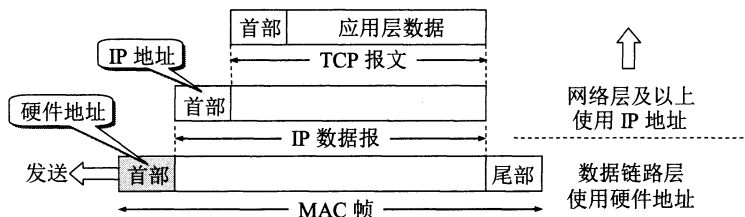


图 4-8 IP 地址与硬件地址的区别

在发送数据时，数据从高层下到低层，然后才到通信链路上传输。使用 IP 地址的 IP 数

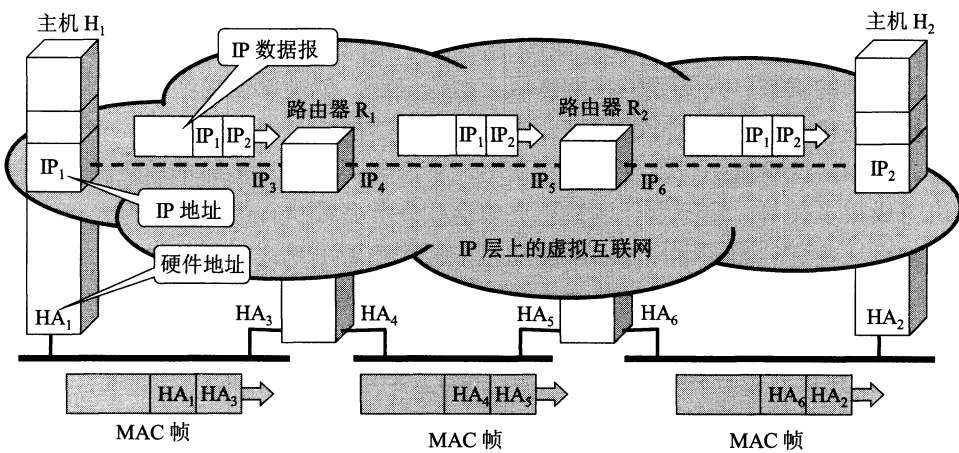
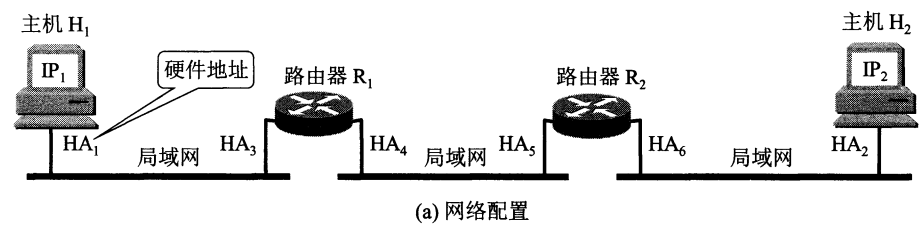
① 注：在局域网中，由于硬件地址已固化在网卡上的 ROM 中，因此常常将硬件地址称为物理地址。因为在局域网的 MAC 帧中的源地址和目的地址都是硬件地址，因此硬件地址又称为 MAC 地址。在本书中，物理地址、硬件地址和 MAC 地址常常作为同义词出现。

据报一旦交给了数据链路层，就被封装成 MAC 帧了。MAC 帧在传送时使用的源地址和目的地址都是硬件地址，这两个硬件地址都写在 MAC 帧的首部中。

连接在通信链路上的设备（主机或路由器）在收到 MAC 帧时，根据 MAC 帧首部中的硬件地址决定收下或丢弃。只有在剥去 MAC 帧的首部和尾部后把 MAC 层的数据上交给网络层后，网络层才能在 IP 数据报的首部中找到源 IP 地址和目的 IP 地址。

总之，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。在网络层和网络层以上使用的是 IP 地址，而数据链路层及以下使用的是硬件地址。在图 4-8 中，当 IP 数据报放入数据链路层的 MAC 帧中以后，整个的 IP 数据报就成为 MAC 帧的数据，因而在数据链路层看不见数据报的 IP 地址。

图 4-9(a)画的是三个局域网用两个路由器 R_1 和 R_2 互连起来。现在主机 H_1 要和主机 H_2 通信。这两台主机的 IP 地址分别是 IP_1 和 IP_2 ，而它们的硬件地址分别为 HA_1 和 HA_2 （HA 表示 Hardware Address）。通信的路径是： $H_1 \rightarrow$ 经过 R_1 转发 \rightarrow 再经过 R_2 转发 $\rightarrow H_2$ 。路由器 R_1 因同时连接到两个局域网，因此它有两个硬件地址，即 HA_3 和 HA_4 。同理，路由器 R_2 也有两个硬件地址 HA_5 和 HA_6 。



(b) 不同层次、不同区间的源地址和目的地址

图 4-9 从不同层次上看 IP 地址和硬件地址

图 4-9(b)特别强调了 IP 地址与硬件地址的区别。表 4-4 归纳了这种区别。

表 4-4 图 4-9(b)中不同层次、不同区间的源地址和目的地址

	在网络层		在数据链路层	
	写入 IP 数据报首部的地址		写入 MAC 帧首部的地址	
	源地址	目的地址	源地址	目的地址
从 H ₁ 到 R ₁	IP ₁	IP ₂	HA ₁	HA ₃
从 R ₁ 到 R ₂	IP ₁	IP ₂	HA ₄	HA ₅
从 R ₂ 到 H ₂	IP ₁	IP ₂	HA ₆	HA ₂

这里要强调指出以下几点：

(1) 在 IP 层抽象的互联网上只能看到 IP 数据报。虽然 IP 数据报要经过路由器 R₁ 和 R₂ 的两次转发，但在它的首部中的源地址和目的地址始终分别是 IP₁ 和 IP₂。图中的数据报上写的“从 IP₁ 到 IP₂”就表示前者是源地址而后者是目的地址。数据报中间经过的两个路由器的 IP 地址并不出现在 IP 数据报的首部中。

(2) 虽然在 IP 数据报首部有源站 IP 地址，但路由器只根据目的站的 IP 地址的网络号进行路由选择。

(3) 在局域网的链路层，只能看见 MAC 帧。IP 数据报被封装在 MAC 帧中。MAC 帧在不同网络上传送时，其 MAC 帧首部中的源地址和目的地址要发生变化，见图 4-9(b)。开始在 H₁ 到 R₁ 间传送时，MAC 帧首部中写的是从硬件地址 HA₁ 发送到硬件地址 HA₃，路由器 R₁ 收到此 MAC 帧后，在数据链路层，要丢弃原来的 MAC 帧的首部和尾部。在转发时，在数据链路层，要重新添加上 MAC 帧的首部和尾部。这时首部中的源地址和目的地址分别便成为 HA₄ 和 HA₅。路由器 R₂ 收到此帧后，再次更换 MAC 帧的首部和尾部，首部中的源地址和目的地址分别变成为 HA₆ 和 HA₂。MAC 帧的首部的这种变化，在上面的 IP 层上是看不见的。

(4) 尽管互连在一起的网络的硬件地址体系各不相同，但 IP 层抽象的互联网却屏蔽了下层这些很复杂的细节。只要我们在网络层上讨论问题，就能够使用统一的、抽象的 IP 地址研究主机和主机或路由器之间的通信。上述的这种“屏蔽”概念是一个很有用、很普遍的基本概念。例如，计算机中广泛使用的图形用户界面使得用户只需简单地点击几下鼠标就能让计算机完成很多任务。实际上计算机要完成这些任务必须执行很多条指令。但这些复杂的过程全都被设计良好的图形用户界面屏蔽掉了，使用户看不见这些复杂过程。

以上这些概念是计算机网络的精髓所在，对这些重要概念务必仔细思考和掌握。

细心的读者会发现，还有两个重要问题没有解决：

- (1) 主机或路由器怎样知道应当在 MAC 帧的首部填入什么样的硬件地址？
- (2) 路由器中的路由表是怎样得出的？

第一个问题就是下一节所要讲的内容，而第二个问题将在后面的 4.5 节详细讨论。

4.2.4 地址解析协议 ARP

在实际应用中，我们经常会遇到这样的问题：已经知道了一个机器（主机或路由器）的 IP 地址，需要找出其相应的硬件地址。地址解析协议 ARP 就是用来解决这样的问题的。图 4-10 说明了 ARP 协议的作用。

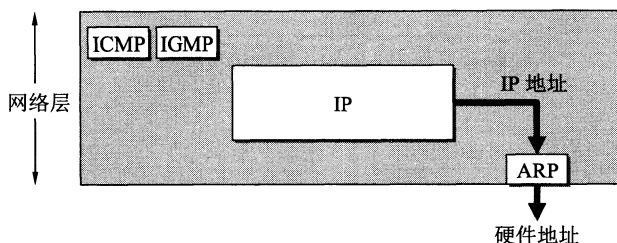


图 4-10 ARP 协议的作用

由于是 IP 协议使用了 ARP 协议，因此通常就把 ARP 协议划归网络层。但 ARP 协议的用途是为了从网络层使用的 IP 地址，解析出在数据链路层使用的硬件地址。因此，有的教科书就按照协议的所用，把 ARP 协议划归在数据链路层。这样做当然也是可以的。

还有一个旧的协议叫做逆地址解析协议 RARP，它的作用是使只知道自己硬件地址的主机能够通过 RARP 协议找出其 IP 地址。现在的 DHCP 协议（见第 6 章的 6.6 节）已经包含了 RARP 协议的功能。因此本书不再介绍 RARP 协议。

下面就介绍 ARP 协议的要点。

我们知道，网络层使用的是 IP 地址，但在实际网络的链路上传送数据帧时，最终还是必须使用该网络的硬件地址。但 IP 地址和下面的网络的硬件地址之间由于格式不同而不存在简单的映射关系（例如，IP 地址有 32 位，而局域网的硬件地址是 48 位）。此外，在一个网络上可能经常会有新的主机加入进来，或撤走一些主机。更换网络适配器也会使主机的硬件地址改变。地址解析协议 ARP 解决问题的方法是在主机 ARP 高速缓存中存放一个从 IP 地址到硬件地址的映射表，并且这个映射表还经常动态更新（新增或超时删除）。

每一台主机都设有一个 ARP 高速缓存(ARP cache)，里面有本局域网上的各主机和路由器的 IP 地址到硬件地址的映射表，这些都是该主机目前知道的一些地址。那么主机怎样知道这些地址呢？我们可以通过下面的例子来说明。

当主机 A 要向本局域网上的某台主机 B 发送 IP 数据报时，就先在其 ARP 高速缓存中查看有无主机 B 的 IP 地址。如有，就在 ARP 高速缓存中查出其对应的硬件地址，再把这个硬件地址写入 MAC 帧，然后通过局域网把该 MAC 帧发往此硬件地址。

也有可能查不到主机 B 的 IP 地址的项目。这可能是主机 B 才入网，也可能是主机 A 刚刚加电，其高速缓存还是空的。在这种情况下，主机 A 就自动运行 ARP，然后按以下步骤找出主机 B 的硬件地址。

(1) ARP 进程在本局域网上广播发送一个 ARP 请求分组（具体格式可参阅[COME06]的第 23 章）。图 4-11(a)是主机 A 广播发送 ARP 请求分组的示意图。ARP 请求分组的主要内容是：“我的 IP 地址是 209.0.0.5，硬件地址是 00-00-C0-15-AD-18。我想知道 IP 地址为 209.0.0.6 的主机的硬件地址。”

(2) 在本局域网上的所有主机上运行的 ARP 进程都收到此 ARP 请求分组。

(3) 主机 B 的 IP 地址与 ARP 请求分组中要查询的 IP 地址一致，就收下这个 ARP 请求分组，并向主机 A 发送 ARP 响应分组（其格式见[COME06]），同时在这个 ARP 响应分组中写入自己的硬件地址。由于其余的所有主机的 IP 地址都与 ARP 请求分组中要查询的 IP 地址不一致，因此都不理睬这个 ARP 请求分组，见图 4-11(b)。ARP 响应分组的主要内容是：“我的 IP 地址是 209.0.0.6，我的硬件地址是 08-00-2B-00-EE-0A。”请注意：虽然 ARP 请求分组是广播发送的，但 ARP 响应分组是普通的单播，即从一个源地址发送到一个目的地址。

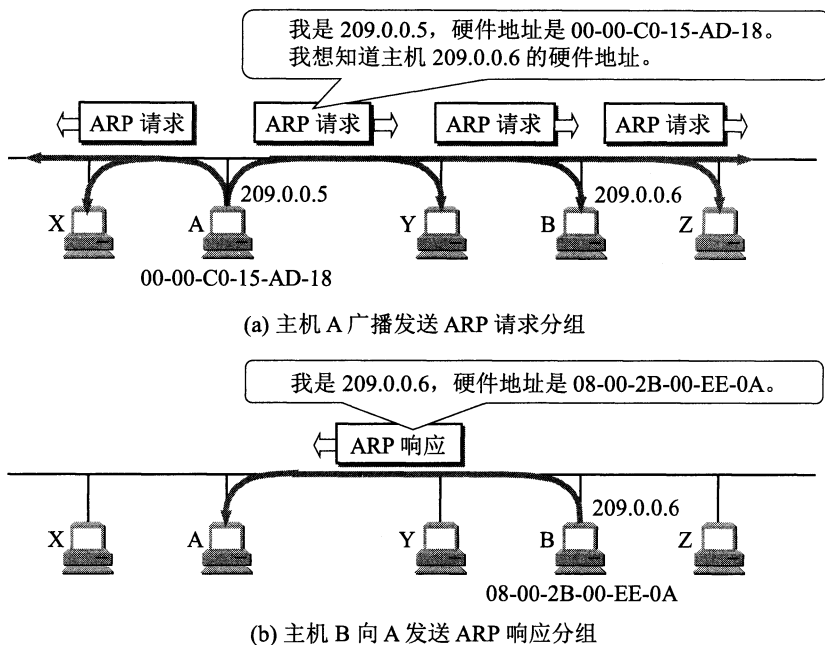


图 4-11 地址解析协议 ARP 的工作原理

(4) 主机 A 收到主机 B 的 ARP 响应分组后，就在其 ARP 高速缓存中写入主机 B 的 IP 地址到硬件地址的映射。

当主机 A 向 B 发送数据报时，很可能以后不久主机 B 还要向 A 发送数据报，因而主机 B 也可能要向 A 发送 ARP 请求分组。为了减少网络上的通信量，主机 A 在发送其 ARP 请求分组时，就把自己的 IP 地址到硬件地址的映射写入 ARP 请求分组。当主机 B 收到 A 的 ARP 请求分组时，就把主机 A 的这一地址映射写入主机 B 自己的 ARP 高速缓存中。以后主机 B 向 A 发送数据报时就很方便了。

可见 ARP 高速缓存非常有用。如果不使用 ARP 高速缓存，那么任何一台主机只要进行一次通信，就必须在网络上用广播方式发送 ARP 请求分组，这就使网络上的通信量大大增加。ARP 把已经得到的地址映射保存在高速缓存中，这样就使得该主机下次再和具有同样目的地址的主机通信时，可以直接从高速缓存中找到所需的硬件地址而不必再用广播方式发送 ARP 请求分组。

ARP 对保存在高速缓存中的每一个映射地址项目都设置生存时间（例如，10 ~ 20 分钟）。凡超过生存时间的项目就从高速缓存中删除掉。设置这种地址映射项目的生存时间是很重要的。设想有一种情况。主机 A 和 B 通信。A 的 ARP 高速缓存里保存有 B 的硬件地址。但 B 的网络适配器突然坏了，B 立即更换了一块，因此 B 的硬件地址就改变了。假定 A 还要和 B 继续通信。A 在其 ARP 高速缓存中查找到 B 原先的硬件地址，并使用该硬件地址向 B 发送数据帧。但 B 原先的硬件地址已经失效了，因此 A 无法找到主机 B。但是过了一段不长的生存时间，A 的 ARP 高速缓存中已经删除了 B 原先的硬件地址，于是 A 重新广播发送 ARP 请求分组，又找到了 B。

请注意，ARP 是解决同一个局域网上的主机或路由器的 IP 地址和硬件地址的映射问题。如果所要找的主机和源主机不在同一个局域网，例如，在前面的图 4-9 中，主机 H₁

就无法解析出另一个局域网上主机 H_2 的硬件地址（实际上主机 H_1 也不需要知道远程主机 H_2 的硬件地址）。主机 H_1 发送给 H_2 的 IP 数据报首先需要通过与本站 H_1 连接在同一个局域网上的路由器 R_1 来转发。因此主机 H_1 这时需要把路由器 R_1 的 IP 地址 IP_3 解析为硬件地址 HA_3 ，以便能够把 IP 数据报传送到路由器 R_1 。以后， R_1 从转发表找出了下一跳路由器 R_2 ，同时使用 ARP 解析出 R_2 的硬件地址 HA_5 。于是 IP 数据报按照硬件地址 HA_5 转发到路由器 R_2 。路由器 R_2 在转发这个 IP 数据报时用类似方法解析出目的主机 H_2 的硬件地址 HA_2 ，使 IP 数据报最终交付主机 H_2 。

从 IP 地址到硬件地址的解析是自动进行的，主机的用户对这种地址解析过程是不知道的。只要主机或路由器要和本网络上的另一个已知 IP 地址的主机或路由器进行通信，ARP 协议就会自动地把这个 IP 地址解析为链路层所需要的硬件地址。

下面我们归纳出使用 ARP 的四种典型情况（图 4-12）。

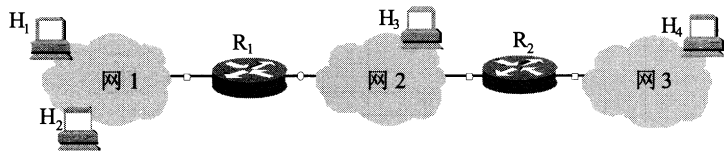


图 4-12 使用 ARP 的四种典型情况

(1) 发送方是主机（如 H_1 ），要把 IP 数据报发送到同一个网络上的另一台主机（如 H_2 ）。这时 H_1 发送 ARP 请求分组（在网 1 上广播），找到目的主机 H_2 的硬件地址。

(2) 发送方是主机（如 H_1 ），要把 IP 数据报发送到另一个网络上的一台主机（如 H_3 或 H_4 ）。这时 H_1 发送 ARP 请求分组（在网 1 上广播），找到网 1 上的一个路由器 R_1 的硬件地址。剩下的工作由路由器 R_1 来完成。 R_1 要做的事情是下面的(3)或(4)。

(3) 发送方是路由器（如 R_1 ），要把 IP 数据报转发到与 R_1 连接在同一个网络（网 2）上的主机（如 H_3 ）。这时 R_1 发送 ARP 请求分组（在网 2 上广播），找到目的主机 H_3 的硬件地址。

(4) 发送方是路由器（如 R_1 ），要把 IP 数据报转发到网 3 上的一台主机（如 H_4 ）。 H_4 与 R_1 不是连接在同一个网络上。这时 R_1 发送 ARP 请求分组（在网 2 上广播），找到连接在网 2 上的一个路由器 R_2 的硬件地址。剩下的工作由这个路由器 R_2 来完成。

在许多情况下需要多次使用 ARP。但这只是以上几种情况的反复使用而已。

有的读者可能会产生这样的问题：既然在网络链路上传送的帧最终是按照硬件地址找到目的主机的，那么为什么我们还要使用抽象的 IP 地址，而不直接使用硬件地址进行通信？这样似乎可以免除使用 ARP。

这个问题必须弄清楚。

由于全世界存在着各式各样的网络，它们使用不同的硬件地址。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，因此由用户或用户主机来完成这项工作几乎是不可能的事。但 IP 编址把这个复杂问题解决了。连接到互联网的主机只需各自拥有一个唯一的 IP 地址，它们之间的通信就像连接在同一个网络上那样简单方便，因为上述的调用 ARP 的复杂过程都是由计算机软件自动进行的，对用户来说是看不见这种调用过程的。

因此，在虚拟的 IP 网络上用 IP 地址进行通信给广大的计算机用户带来很大的方便。

4.2.5 IP 数据报的格式

IP 数据报的格式能够说明 IP 协议都具有什么功能。在 TCP/IP 的标准中，各种数据格式常常以 32 位（即 4 字节）为单位来描述。图 4-13 是 IP 数据报的完整格式。

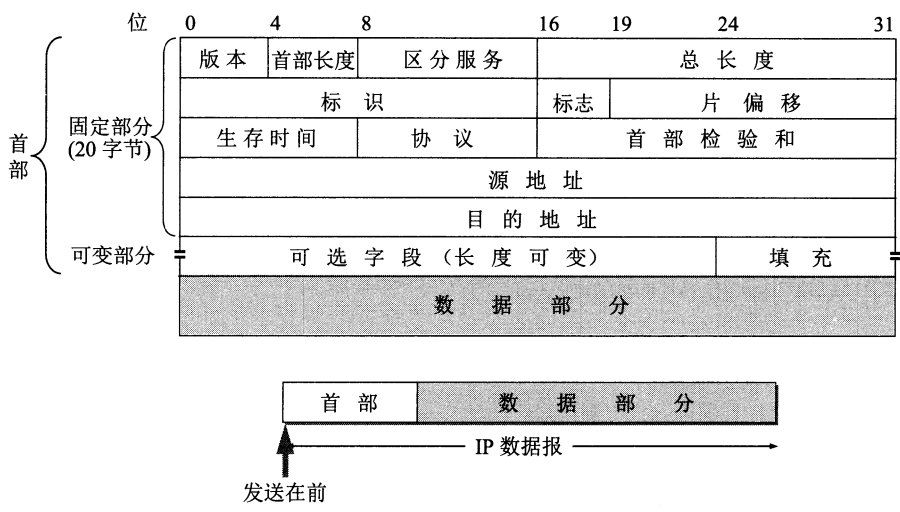


图 4-13 IP 数据报的格式

从图 4-13 可看出，一个 IP 数据报由首部和数据两部分组成。首部的前一部分是**固定长度**，共 20 字节，是所有 IP 数据报必须具有的。在首部的固定部分的后面是一些**可选字段**，其长度是可变的。下面介绍首部各字段的意义。

1. IP 数据报首部的固定部分中的各字段

(1) **版本** 占 4 位，指 IP 协议的版本。通信双方使用的 IP 协议的版本必须一致。目前广泛使用的 IP 协议版本号为 4（即 IPv4）。关于以后要使用的 IPv6（即版本 6 的 IP 协议），我们将在后面的 4.6 节讨论。

(2) **首部长度** 占 4 位，可表示的最大十进制数值是 15。请注意，首部长度字段所表示数的单位是 32 位字（1 个 32 位字长是 4 字节）。因为 IP 首部的固定长度是 20 字节，因此首部长度字段的最小值是 5（即二进制表示的首部长度是 0101）。而当首部长度为最大值 1111 时（即十进制数的 15），就表明首部长度达到最大值 15 个 32 位字长，即 60 字节。当 IP 分组的首部长度不是 4 字节的整数倍时，必须利用最后的填充字段加以填充。因此 IP 数据报的数据部分永远在 4 字节的整数倍时开始，这样在实现 IP 协议时较为方便。首部长度限制为 60 字节的缺点是有可能不够用。但这样做是希望用户尽量减少开销。最常用的首部长度是 20 字节（即首部长度为 0101），这时不使用任何选项。

(3) **区分服务** 占 8 位，用来获得更好的服务。这个字段在旧标准中叫做**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 DS (Differentiated Services)**。只有在使用区分服务时，这个字段才起作用（见 8.4.4 节）。在一般的情况下都不使用这个字段[RFC 2474, 3168, 3260]。

(4) **总长度** 总长度指首部和数据之和的长度，单位为字节。总长度字段为 16 位，因此数据报的最大长度为 $2^{16} - 1 = 65535$ 字节。然而实际上传送这样长的数据报在现实中是极少遇到的。

我们知道，在 IP 层下面的每一种数据链路层协议都规定了一个数据帧中的**数据字段的最大长度**，这称为**最大传送单元 MTU (Maximum Transfer Unit)**。当一个 IP 数据报封装成链路层的帧时，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层所规定的 MTU 值。例如，最常用的以太网就规定其 MTU 值是 1500 字节。若所传送的数据报长度超过数据链路层的 MTU 值，就必须把过长的数据报进行分片处理。

虽然使用尽可能长的 IP 数据报会使传输效率得到提高（因为每一个 IP 数据报中首部长度占数据报总长度的比例就会小些），但数据报短些也有好处。每一个 IP 数据报越短，路由器转发的速度就越快。为此，IP 协议规定，在互联网中所有的主机和路由器，必须能够接受长度不超过 576 字节的数据报。这是假定上层交下来的数据长度有 512 字节（合理的长度），加上最长的 IP 首部 60 字节，再加上 4 字节的富余量，就得到 576 字节。当主机需要发送长度超过 576 字节的数据报时，应当先了解一下，目的主机能否接受所要发送的数据报长度。否则，就要进行分片。

在进行分片时（见后面的“片偏移”字段），数据报首部中的“总长度”字段是指分片后的每一个分片的首部长度与该分片的数据长度的总和。

(5) **标识(identification)** 占 16 位。IP 软件在存储器中维持一个计数器，每产生一个数据报，计数器就加 1，并将此值赋给标识字段。但这个“标识”并不是序号，因为 IP 是无连接服务，数据报不存在按序接收的问题。当数据报由于长度超过网络的 MTU 而必须分片时，这个标识字段的值就被复制到所有的数据报片的标识字段中。相同的标识字段的值使分片后的各数据报片最后能正确地重装成为原来的数据报。

(6) **标志(flag)** 占 3 位，但目前只有两位有意义。

- 标志字段中的最低位记为 **MF (More Fragment)**。MF = 1 即表示后面“还有分片”的数据报。MF = 0 表示这已是若干数据报片中的最后一个。
- 标志字段中间的一位记为 **DF (Don't Fragment)**，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。片偏移指出：较长的分组在分片后，某片在原分组中的相对位置。也就是说，相对于用户数据字段的起点，该片从何处开始。片偏移以 8 个字节为偏移单位。这就是说，每个分片的长度一定是 8 字节（64 位）的整数倍。

下面举一个例子。

【例 4-1】 一数据报的总长度为 3820 字节，其数据部分为 3800 字节长（使用固定首部），需要分片为长度不超过 1420 字节的数据报片。因固定首部长度为 20 字节，因此每个数据报片的数据部分长度不能超过 1400 字节。于是分为 3 个数据报片，其数据部分的长度分别为 1400, 1400 和 1000 字节。原始数据报首部被复制为各数据报片的首部，但必须修改有关字段的值。图 4-14 给出分片后得出的结果（请注意片偏移的数值）。

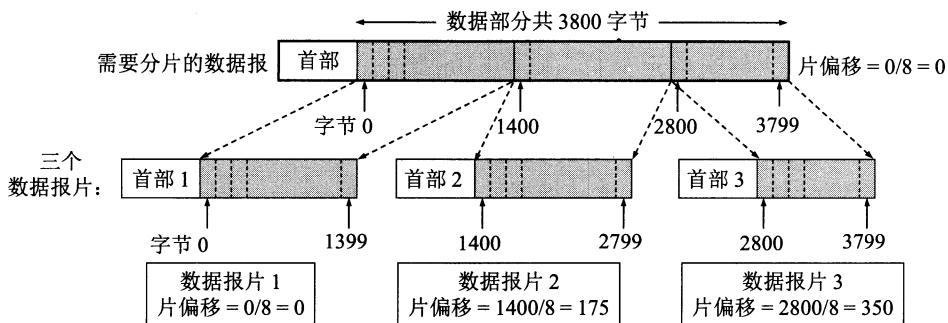


图 4-14 数据报的分片举例

表 4-5 是本例中数据报首部与分片有关的字段中的数值，其中标识字段的值是任意给定的（12345）。具有相同标识的数据报片在目的站就可无误地重装成原来的数据报。

表 4-5 IP 数据报首部中与分片有关的字段中的数值

	总长度	标识	MF	DF	片偏移
原始数据报	3820	12345	0	0	0
数据报片 1	1420	12345	1	0	0
数据报片 2	1420	12345	1	0	175
数据报片 3	1020	12345	0	0	350

现在假定数据报片 2 经过某个网络时还需要再进行分片，即划分为数据报片 2-1（携带数据 800 字节）和数据报片 2-2（携带数据 600 字节）。那么这两个数据报片的总长度、标识、MF、DF 和片偏移分别为：820, 12345, 1, 0, 175；620, 12345, 1, 0, 275。

(8) 生存时间 占 8 位，生存时间字段常用的英文缩写是 TTL (Time To Live)，表明这是数据报在网络中的寿命。由发出数据报的源点设置这个字段。其目的是防止无法交付的数据报无限制地在互联网中兜圈子（例如从路由器 R_1 转发到 R_2 ，再转发到 R_3 ，然后又转发到 R_1 ），因而白白消耗网络资源。最初的设计是以秒作为 TTL 值的单位。每经过一个路由器时，就把 TTL 减去数据报在路由器所消耗掉的一段时间。若数据报在路由器消耗的时间小于 1 秒，就把 TTL 值减 1。当 TTL 值减为零时，就丢弃这个数据报。

然而随着技术的进步，路由器处理数据报所需的时间不断在缩短，一般都远远小于 1 秒，后来就把 TTL 字段的功能改为“跳数限制”（但名称不变）。路由器在每次转发数据报之前就把 TTL 值减 1。若 TTL 值减小到零，就丢弃这个数据报，不再转发。因此，现在 TTL 的单位不再是秒，而是跳数。TTL 的意义是指明数据报在互联网中至多可经过多少个路由器。显然，数据报能在互联网中经过的路由器的最大数值是 255。若把 TTL 的初始值设置为 1，就表示这个数据报只能在本局域网中传送。因为这个数据报一传送到局域网上的某个路由器，在被转发之前 TTL 值就减小到零，因而就会被这个路由器丢弃。

(9) 协议 占 8 位，协议字段指出此数据报携带的数据是使用何种协议，以便使目的主机的 IP 层知道应将数据部分上交给哪个协议进行处理。

常用的一些协议和相应的协议字段值如下^①：

协议名	ICMP	IGMP	IP ^②	TCP	EGP	IGP	UDP	IPv6	ESP	OSPF
协议字段值	1	2	4	6	8	9	17	41	50	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）。不检验数据部分可减少计算的工作量。为了进一步减小计算检验和的工作量，IP 首部的检验和不采用复杂的 CRC 检验码而采用下面的简单计算方法：在发送方，先把 IP 数据报首部划分为许多 16 位字的序列，并把检验和字段置零。用反码算术运算^③把所有 16 位字相加后，将得到的和的反码写入检验和字段。接收方收到数据报后，将首部的所有 16 位字再使用反码算术运算相加一次。将得到的和取反码，即得出接收方检验和的计算结果。若首部未发生任何变化，则此结果必为 0，于是就保留这个数据报。否则即认为出差错，并将此数据报丢弃。图 4-15 说明了 IP 数据报首部检验和的计算过程。

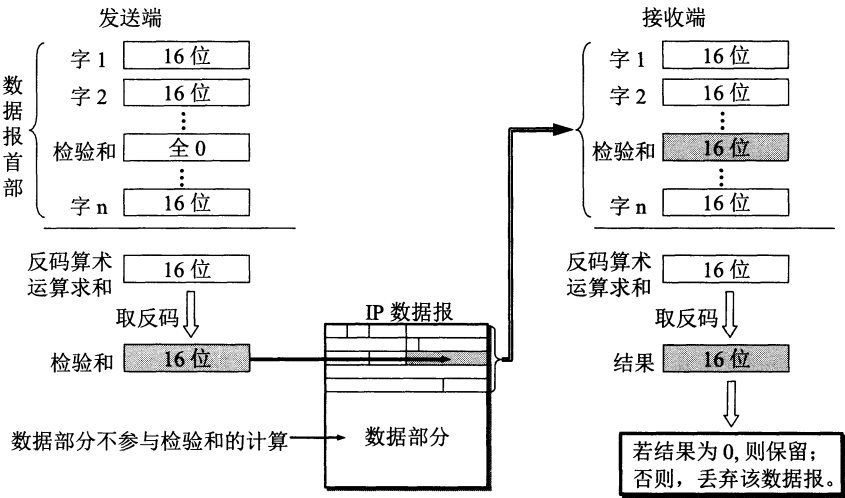


图 4-15 IP 数据报首部检验和的计算过程

(11) 源地址 占 32 位。

(12) 目的地址 占 32 位。

① 注：原来如协议字段值这样的数值都是由互联网赋号管理局 IANA 负责制定，并公布在有关的 RFC 文档中。其实 IANA 并不是一个庞大的机构，而仅仅由 Jon Postel 一个人来负责管理。由于 Jon Postel 于 1998 年去世，同时也由于互联网的商业化和国际化，美国决定用一个新的、私营的、非营利的国际公司——互联网名称与数字地址分配机构 ICANN [W-ICANN]取代 IANA。但后来 ICANN 并没有取消 IANA，而是保留了 IANA，并且和 IANA 进行了分工。因此现在就出现了 IANA/ICANN 或 ICANN/IANA 这样的写法。这两个机构都负责 IP 地址和一些重要参数的管理。现在有关互联网上的重要的参数已经不在 RFC 文档公布[RFC 3232]，而改为在网址 www.iana.org 上查询一个联机数据库。

② 注：这里的 IP 表示特殊的 IP 数据报——IP 数据报再封装到 IP 数据报中。

③ 注：两个数进行二进制反码求和的运算很简单。它的规则是从低位到高位逐列进行计算。0 和 0 相加是 0，0 和 1 相加是 1，1 和 1 相加是 0 但要产生一个进位 1，加到下一列。若最高位相加后产生进位，则最后得到的结果要加 1。请注意，反码 (one's complement) 和补码 (two's complement) 是不一样的。

2. IP 数据报首部的可变部分

IP 数据报首部的可变部分就是一个选项字段。选项字段用来支持排错、测量以及安全措施，内容很丰富。此字段的长度可变，从 1 个字节到 40 个字节不等，取决于所选择的项目。某些选项项目只需要 1 个字节，它只包括 1 个字节的选项代码。而有些选项需要多个字节，这些选项一个个拼接起来，中间不需要有分隔符，最后用全 0 的填充字段补齐成为 4 字节的整数倍。

增加首部的可变部分是为了增加 IP 数据报的功能，但这同时也使得 IP 数据报的首部长度成为可变的。这就增加了每一个路由器处理数据报的开销。实际上这些选项很少被使用。很多路由器都不考虑 IP 首部的选项字段，因此新的 IP 版本 IPv6 就把 IP 数据报的首部长度做成固定的。这里就不讨论这些选项的细节了。有兴趣的读者可参阅 RFC 791。

4.2.6 IP 层转发分组的流程

下面我们先用一个简单例子来说明路由器是怎样转发分组的。图 4-16(a)是一个路由表的简单例子。有四个 A 类网络通过三个路由器连接在一起。每一个网络上都可能有成千上万台主机（图中没有画出这些主机）。可以想象，若路由表指出到每一台主机应怎样转发，则所得出的路由表就会过于庞大（如果每一个网络有 1 万台主机，四个网络就有 4 万台主机，因而每一个路由表就有 4 万个项目，即 4 万行。每一行对应于一台主机）。但若路由表指出到某个网络应如何转发，则每个路由器中的路由表就只包含 4 个项目（即只有 4 行，每一行对应于一个网络）。以路由器 R_2 的路由表为例。由于 R_2 同时连接在网络 2 和网络 3 上，因此只要目的主机在网络 2 或网络 3 上，都可通过接口 0 或 1 由路由器 R_2 直接交付（当然还要利用地址解析协议 ARP 才能找到这些主机相应的硬件地址）。若目的主机在网络 1 中，则下一跳路由器应为 R_1 ，其 IP 地址为 20.0.0.7。路由器 R_2 和 R_1 由于同时连接在网络 2 上，因此从路由器 R_2 把分组转发到路由器 R_1 是很容易的。同理，若目的主机在网络 4 中，则路由器 R_2 应把分组转发给 IP 地址为 30.0.0.1 的路由器 R_3 。我们应当注意到，图中的每一个路由器都有两个不同的 IP 地址。

可以把整个的网络拓扑简化为图 4-16(b)所示的那样。在简化图中，网络变成了一条链路，但每一个路由器旁边都注明其 IP 地址。使用这样的简化图，可以使我们不必关心某个网络内部的具体拓扑以及连接在该网络上有多少台主机，因为这些对于研究分组转发问题并没有什么关系。这样的简化图强调了在互联网上转发分组时，是从一个路由器转发到下一个路由器。

总之，在路由表中，对每一条路由最主要的是以下两个信息^①：

（目的网络地址，下一跳地址）

^① 注：一个实际的路由表还会有其他的一些信息。例如，标志、参考计数、使用情况以及接口等。“标志”可以设置多个字符以说明不同的意思。如 U 表示该路由是可用的，G 表示下一跳地址是一个路由器，因而是间接交付（如不设置 G，则表示直接交付），H 表示该路由是到一台主机（如不设置 H，则表示该路由是到一个网络）。“参考计数”是给出正在使用该路由的 TCP 连接数。“使用情况”显示出通过该路由的分组数。“接口”是本地接口的名字，指出分组应当从哪一个接口转发。

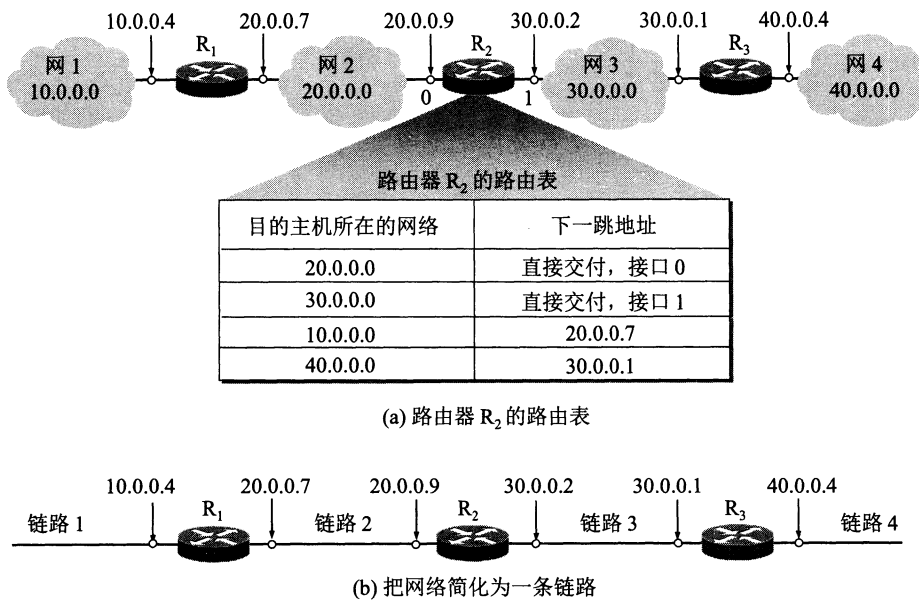


图 4-16 路由表举例

于是，我们就根据目的网络地址来确定下一跳路由器，这样做可得出以下的结果。

(1) IP 数据报最终一定可以找到目的主机所在目的网络上的路由器（可能要通过多次的间接交付）。

(2) 只有到达最后一个路由器时，才试图向目的主机进行直接交付。

虽然互联网所有的分组转发都是基于目的主机所在的网络，但在大多数情况下都允许有这样的特例，即对特定的目的主机指明一个路由。这种路由叫做**特定主机路由**。采用特定主机路由可使网络管理人员更方便地控制网络和测试网络，同时也可在需要考虑某种安全问题时采用这种特定主机路由。在对网络的连接或路由表进行排错时，指明到某一台主机的特殊路由就十分有用。

路由器还可采用**默认路由**(default route)以减小路由表所占用的空间和搜索路由表所用的时间。这种转发方式在一个网络只有很少的对外连接时是很有用的。实际上，默认路由在主机发送 IP 数据报时往往更能显示出它的好处。我们在前面的 4.2.1 节已经讲过，主机在发送每一个 IP 数据报时都要查找自己的路由表。如果一台主机连接在一个小网络上，而这个网络只用一个路由器和互联网连接，那么在这种情况下使用默认路由是非常合适的。例如，在图 4-17 的互联网中，连接在网络 N₁ 上的任何一台主机中的路由表只需要三个项目即可。第一个项目就是到本网络主机的路由，其目的网络就是本网络 N₁，因而不需要路由器转发，而是直接交付。第二个项目是到网络 N₂ 的路由，对应的下一跳路由器是 R₂。第三个项目就是**默认路由**。只要目的网络是其他网络（不是 N₁ 或 N₂），就一律选择默认路由，把数据报先间接交付路由器 R₁，让 R₁ 再转发给互联网中的下一个路由器，一直转发到目的网络上的路由器，最后进行直接交付。在实际上的路由器中，像图 4-17 路由表中所示的“直接”和“其他”的几个字符并没有出现在路由表中，而是被记为 0.0.0.0。

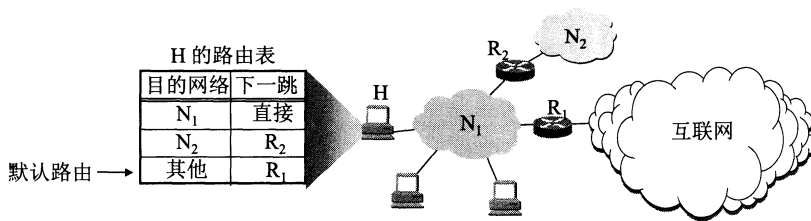


图 4-17 路由器 R₁ 充当网络 N₁ 的默认路由器

这里我们应当强调指出，在 IP 数据报的首部中没有地方可以用来指明“下一跳路由器的 IP 地址”。在 IP 数据报的首部写上的 IP 地址是源 IP 地址和目的 IP 地址，而没有中间经过的路由器的 IP 地址。既然 IP 数据报中没有下一跳路由器的 IP 地址，那么待转发的数据报又怎样能够找到下一跳路由器呢？

当路由器收到一个待转发的数据报，在从路由表得出下一跳路由器的 IP 地址后，不是把这个地址填入 IP 数据报，而是送交数据链路层的网络接口软件。网络接口软件负责把下一跳路由器的 IP 地址转换成硬件地址（必须使用 ARP），并将此硬件地址放在链路层的 MAC 帧的首部，然后根据这个硬件地址找到下一跳路由器。由此可见，当发送一连串的数据报时，上述的这种查找路由表、用 ARP 得到硬件地址、把硬件地址写入 MAC 帧的首部等过程，将不断地重复进行，造成了一定的开销。

那么，能不能在路由表中不使用 IP 地址而直接使用硬件地址呢？不行。我们一定要弄清楚，使用抽象的 IP 地址，本来就是为了隐蔽各种底层网络的复杂性而便于分析和研究问题，这样就不可避免地要付出些代价，例如在选择路由时多了一些开销。但反过来，如果在路由表中直接使用硬件地址，那就会带来更多的麻烦。

根据以上所述，可归纳出**分组转发算法**如下：

- (1) 从数据报的首部提取目的主机的 IP 地址 D ，得出目的网络地址为 N 。
- (2) 若 N 就是与此路由器直接相连的某个网络地址，则进行**直接交付**，不需要再经过其他的路由器，直接把数据报交付目的主机（这里包括把目的主机地址 D 转换为具体的硬件地址，把数据报封装为 MAC 帧，再发送此帧）；否则就是间接交付，执行(3)。
- (3) 若路由表中有目的地址为 D 的特定主机路由，则把数据报传送给路由表中所指明的下一跳路由器；否则，执行(4)。
- (4) 若路由表中有到达网络 N 的路由，则把数据报传送给路由表中所指明的下一跳路由器；否则，执行(5)。
- (5) 若路由表中有一个默认路由，则把数据报传送给路由表中所指明的默认路由器；否则，执行(6)。
- (6) 报告转发分组出错。

这里我们要再强调一下，路由表并没有给分组指明到某个网络的完整路径（即先经过哪一个路由器，然后再经过哪一个路由器，等等）。路由表指出，到某个网络应当先到某个路由器（即下一跳路由器），在到达下一跳路由器后，再继续查找其路由表，知道再下一步应当到哪一个路由器。这样一步一步地查找下去，直到最后到达目的网络。

可以用一个简单的比喻来说明查找路由表的作用。例如，从家门口开车到机场，但没有地图，不知道应当走哪条路线。好在每一个道路岔口都有一个警察可以询问。因此，每到一岔口（相当于到了一个路由器），就问：“到机场应当走哪个方向？”（相当于查找路由

表)。该警察既不指明到下一个岔口以后再应当如何走，也不指明还要经过几个岔口才到达机场。他仅仅指出下一个岔口的方向。其回答可能是：“向左转。”到了下一个岔口，再讯问到机场该走哪个方向？回答可能是：“直行。”这样，每到一个岔口，就询问下一步该如何走。这样，即使我们没有地图，但最终一定可以到达目的地——机场。

上面所讨论的是 IP 层怎样根据路由表的内容进行分组转发，而没有涉及到路由表一开始是如何建立的以及路由表中的内容应如何进行更新。但是在进一步讨论路由选择之前，我们还要先介绍划分子网和构造超网这两个非常重要的概念。

4.3 划分子网和构造超网

4.3.1 划分子网

1. 从两级 IP 地址到三级 IP 地址

在今天看来，在 ARPANET 的早期，IP 地址的设计确实不够合理。

第一，IP 地址空间的利用率有时很低。

每一个 A 类地址网络可连接的主机数超过 1000 万，而每一个 B 类地址网络可连接的主机数也超过 6 万。有的单位申请到了一个 B 类地址网络，但所连接的主机数并不多，可是又不愿意申请一个足够使用的 C 类地址，理由是考虑到今后可能的发展。IP 地址的浪费，还会使 IP 地址空间的资源过早地被用完。

第二，给每一个物理网络分配一个网络号会使路由表变得太大从而使网络性能变坏。

每一个路由器都应当能够从路由表查出应怎样到达其他网络的下一跳路由器。因此，互联网中的网络数越多，路由器的路由表的项目数也就越多。这样，即使我们拥有足够多的 IP 地址资源可以给每一个物理网络分配一个网络号，也会导致路由器的路由表中的项目数过多。这不仅增加了路由器的成本（需要更多的存储空间），而且使查找路由时耗费更多的时间，同时也使路由器之间定期交换的路由信息急剧增加，因而使路由器和整个互联网的性能都下降了。

第三，两级 IP 地址不够灵活。

有时情况紧急，一个单位需要新的地点马上开通一个新的网络。但是在申请到一个新的 IP 地址之前，新增加的网络是不可能连接到互联网上工作的。我们希望有一种方法，使一个单位能随时灵活地增加本单位的网络，而不必事先到互联网管理机构去申请新的网络号。原来的两级 IP 地址无法做到这一点。

为解决上述问题，从 1985 年起在 IP 地址中又增加了一个“子网号字段”，使两级 IP 地址变成为三级 IP 地址，它能够较好地解决上述问题，并且使用起来也很灵活。这种做法叫做划分子网(subnetting) [RFC 950]，或子网寻址或子网路由选择。划分子网已成为互联网的正式标准协议。

划分子网的基本思路如下：

(1) 一个拥有许多物理网络的单位，可将所属的物理网络划分为若干个子网(subnet)。划分子网纯属一个单位内部的事情。本单位以外的网络看不见这个网络是由多少个子网组成，因为这个单位对外仍然表现为一个网络。