第5章 网络层与网络互连

前面讲述了两类底层物理网络技术，本章及后面两章将分别讲述因特网使用的 TCP/IP 体系结构中的网络层、传输层和应用层。在因特网中网络层的主要任务是将各种物理网络互连起来，使得不同物理网络的主机之间也可以相互通信。

本章主要讨论将各种物理网络通过路由器互连成为全球范围的互联网——因特网所需要面临的各种问题，以及在因特网中的解决方案。因特网中网络互连需要解决的问题主要包括网络层编址、数据报传送、差错处理、互联网路由维护和 IP 组播等，与此相关的协议或技术包括网际协议 IPv4、地址解析协议 ARP、网际控制报文协议 ICMP、无分类域间路由选择 CIDR、路由信息协议 RIP、开放最短路径优先路由选择协议 OSPF、边界网关协议 BGP、因特网组管理协议 IGMP 等。另外，还将简要讨论移动 IP、虚拟专用网 VPN、网络地址转换 NAT 等技术、下一代网际协议 IPv6 和网络层互连设备。

通过本章的学习，要掌握网络层编址、地址解析 ARP、数据报的交付与转发、IP 差错与控制机制、因特网路由维护机制等协议和方法，了解 IP 组播、移动 IP、虚拟专用网 VPN、网络地址转换 NAT 技术、IPv6 等主要特点和路由器的功能。

5.1 网络层概述

根据 OSI/RM，网络层为不同网络上的主机提供通信服务。数据链路层提供相邻节点间，以帧为单位的数据传输服务。网络层利用数据链路层提供的服务，向传输层提供主机间的分组传递服务。网络层主要需要解决网络层编址、路由选择和拥塞控制等问题。

在 TCP/IP 体系中，网络层也称 IP 层或网间互连层，为互联网主机提供无连接的通信服务。IP 层利用数据链路层提供的服务，向高层提供互联网主机之间的 IP 包传递服务。因特网（Internet）是一个庞大的计算机互联网，由不同的物理网络通过网络互连设备（路由器）相互连接而成。IP 层主要解决网络层编址和路由选择问题，为提高效率，将拥塞控制主要留给高层解决。

因特网为什么要考虑包容多种物理网络技术呢？原因是价格低廉的局域网只能提供短距离的高速通信，而能跨越长距离的广域网不能提供低费用的局部通信。没有哪种网络技术可以满足所有需求，因此需要考虑多种底层硬件技术。

为什么要实现网际互连呢？因为用户希望能够在任意两主机之间进行通信，各物理网络中的用户希望有一个不受任何物理网络边界限制的通信系统。网际互连的作用就是隐藏底层细节，使互联网可以看成是单一的**虚拟网络**，所有计算机都与它相连，而不管实际的物理连接如何。图 5-1 表示从用户的角度，互联网可看成是单个网络，虽然实际上它是多个物理网络通过路由器互连起来的集合。每个物理网络中的主机以及互连设备路由器必须运行 TCP/IP 软件，以允许应用程序可以把互联网当成一个单独的物理网络来使用。

在 TCP/IP 体系中，IP 层包含 IP 协议（网际协议）、ICMP 协议（网际控制报文协议）等若干协议，其中 IP 是该层中最重要的协议，也是 TCP/IP 体系中最重要的协议。IP 协议数据单元称为 IP 数据报，也可称为数据报、**IP 包或 IP 分组**。IP 层下面的网络接口层对应

各种物理网络协议栈，即各种局域网和广域网协议栈。要注意的是，即使广域网（例如 X.25 分组网）包含网络层协议（分组级），也是在 IP 层之下，即 IP 包将被封装在广域网的网络层协议数据单元中传送。

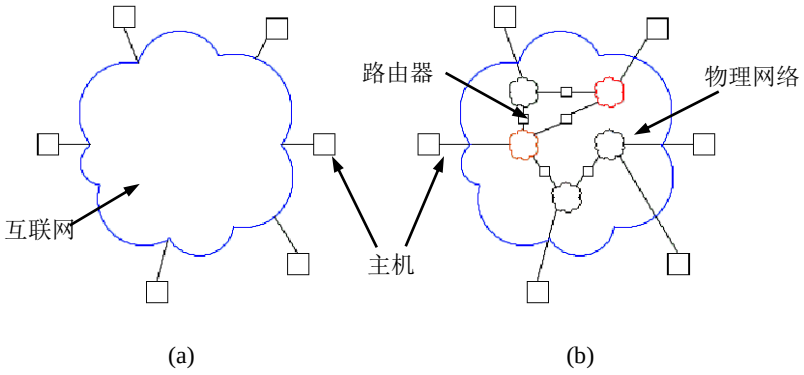


图 5-1 (a) 用户观点的互联网; (b) 互联网的实际连接示例

IP 层主要功能是负责为不同物理网络上的主机提供通信。为此，需要解决若干问题，主要包括：

（1）IP 编址和地址的分配。各物理网络有自己的编址方式，为方便任意主机之间的通信，互连各个物理网络的互联网需要统一标识所有主机。IPv4 要求给每个主机都分配一个 32 比特的整数地址，称为**网际协议地址**（Internet Protocol address，IP 地址）。互联网可以包含很多物理网络，给其中的每个主机都应分配 IP 地址，究竟该怎样分配和使用 IP 地址，才能够方便传送数据报过程中的路由查找，提高互联网的运行效率呢？这是 IP 层要解决的主要问题之一。相关内容包括分类 IP 地址、子网划分、构造超网、无分类编址和 CIDR 等。

（2）IP 包的转发。互联网中通信双方可能位于不同的物理网络中，怎样才能使 IP 包从源主机抵达目的主机呢？这需要依靠工作在网络层互连物理网络的设备路由器，路由器中保存着到各个物理网络的路由信息。路由器通过查找路由表为经过的每个 IP 包选择一条正确的路由，再进行逐跳转发，使 IP 包不断接近目的主机。路由查找算法与 IP 编址方案相关，路由查找结果是下一跳（next hop）的 IP 地址。IP 包必须封装在各物理网络协议包（以下统一称为帧）中发送或转发，帧的目的地址可利用地址解析协议 ARP 来获悉。

（3）路由表的产生和动态刷新。主机发送 IP 包以及路由器转发 IP 包都需要查找路由表确定路由，因此维持路由表的正确性很关键。因特网是个大型网络，为有效地维护路由表，目前已提出了若干路由选择算法和协议，以及自治系统概念。

（4）差错处理。IP 包转发过程中会发生差错，IP 层需要对此进行一些差错处理。这由 IP 协议和 ICMP 协议共同完成。

（5）IP 组播 有许多应用需要一对多的通信，例如网络电视。IP 协议、网际组管理协议（IGMP）和组播路由选择协议共同实现 IP 组播。

IP 层利用物理网络所提供的服务，加上本层的协议功能，向高层提供无连接的 IP 包交付服务。IP 层通过 IP 数据报和 IP 地址实现对物理网络的抽象，隐藏底层网络体系结构和技术细节，向高层提供统一的 IP 数据报，使得各种物理帧的差异性对上层不复存在。IPv4

因特网是一个很大的互联网，它由大量的通过路由器互连起来的物理网络构成。IP 编址和 IP 数据报是支持 TCP/IP 软件隐藏物理网络细节，使构成的互联网看起来是一个统一实体的基础。IP 提供无连接的数据报交付服务。本节介绍：（1）IP 编址方案，包括因特网先后采用的最初的分类编址、子网编址和无分类编址方案；（2）一直在使用的 IPv4 数据报的格式；（3）无连接 IP 数据报的传送，包括直接交付与间接交付概念、各种编址情况下转发 IP 数据报的算法、动态完成 IP 地址到物理地址映射的 ARP 协议、报告传送过程中发生的异常情况的网际控制报文协议 ICMP。

5.2 IPv4

5.2.1 分类的 IP 地址

互联网是一个由设计者抽象出来的虚拟结构，IP 层及以上的协议功能完全由软件实现。设计人员可以自由地选择 IP 编址方案、IP 数据报格式以及交付技术等，不受底层网络硬件的支配。IP 地址实现对各种物理地址的统一，即 IP 层以上各层均使用 IP 地址。TCP/IP 的设计者选择了一种类似于物理网络的编址方案，给因特网上每个主机分配一个 32 比特的唯一值作为单播地址，该地址用在与该主机的所有通信中。

理想的地址应该比较短，因为作为分组控制信息的一部分，地址越长分组的开销会越大。但理想的地址也要足够大，以便能够标识更多的主机，对于网际协议地址，最好能够标识全世界所有主机。此外，IP 地址应支持高效率的路由选择，比如根据目的主机的 IP 地址，就可分辨能否与之直接通信，或者该选择哪个路由器作为下一跳以使 IP 包逐跳向目的主机转发。

最初的 IP 编址方案将 IP 地址分为两部分：前缀和后缀。前缀标识主机所属的物理网络，称为网络号（network ID）。后缀用于区分物理网络内的主机，即标识主机，后缀也称为主机号（host ID）。那么前缀与后缀各应包含多少比特呢？前缀越长，支持的物理网络数越多，但网络内的主机数就越少。反之，长后缀和短前缀，则意味着支持规模较大的物理网络（主机数多），但仅能支持较少的这样的网络。最初的 IP 编址方案兼顾了这两种情形，没有采用单一界限划分前缀和后缀，而是采用三种界限划分，因此称为分类编址方案（classful addressing）。

最初的分类编址方案中包含 5 种形式的 IP 地址，见图 5-2。其中的 A、B、C 类是三种主要类别，用于标识主机和路由器。D 类地址为组播地址。E 类地址为保留地址，留作以后使用。自 1993 年起为了充分利用 IP 地址空间，因特网采用无类 IP 编址方案分配尚未分配的分类 IP 地址，将在 5.2.6 节介绍。虽然分类 IP 地址已不再广泛使用，但这是 IP 编址技术的基础，也是后续发展的根源。

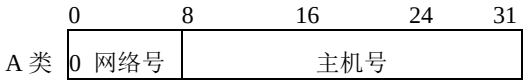


图 5-2 最初的分类编址方案中 IP 地址的五种形式

分类 IP 地址是自标识的（self-identifying），仅从地址本身就能够确定前缀和后缀之间的边界，不用参考其他信息。从地址的最高 2 比特可以区分三种主要类别，从地址的最高 3 比特可以区分 A、B、C、D 四类。路由器在决定一个分组发往何处时要使用地址的网络号部分进行路由选择，地址的自标识特性使得网络号的抽取非常方便，有助于提高路由器的效率。

A 类地址包含 8 比特的网络号部分和 24 比特的主机号部分，B 类地址包含 16 比特的网络号和 16 比特的主机号，C 类地址包含 24 比特的网络号和 8 比特的主机号。

在应用程序或技术文档中，为方便，一般采用**点分十进制记法**书写 IP 地址。将 IP 地址写成小数点分隔的 4 个十进制整数，每个整数给出 IP 地址的一个八比特组的值。例如某主机 32 比特的 IP 地址：

10000001 00000001 01000110 00001111

可写成 129.1.70.15。由于该地址最高两位是 10，根据分类规定可知该地址是一个 B 类地址，并且网络号为 129.1，主机号为 70.15。该主机所在物理网络的 IP 网络地址也可写成 32 位 IP 地址形式：129.1.0.0，注意网络地址的主机号部分用全 0 表示。

5.2.2 IP 地址的分配与使用

在最初的 IP 编址方案中，因特网中的每个物理网络都必须被分配一个唯一的网络号（IP 地址前缀），而该物理网络上的每个主机都使用该网络号作为主机 IP 地址的前缀，网络上各主机的主机号互不相同。

为确保地址的网络部分在因特网上是唯一的，所有因特网地址都有一个中央管理机构进行分配。从因特网出现到 1998 年秋天，一直由 IANA（因特网赋号管理局）控制着 IP 地址的分配，并制定政策。注意**全球统一分配的是 IP 地址的网络号部分**，而主机号由用户组织自行分配，必须保证同一物理网络中各主机的主机号互不相同。位于不同物理网络中的主机，其 IP 地址的主机号可以一样。1998 年底，组建了 ICANN（因特网名字与号码指派协会），它负责指定政策，分配地址，并为协议中使用的名字和其他常量分配值。

ICANN 是顶级的地址管理机构，它授权了一些地址注册商 ARIN、RIPE、APNIC、LATNIC、AFRINIC 管理地址。一般单位可以从它的 ISP（因特网服务提供商）申请 IP 地址，本地 ISP 将单位联入因特网，并为用户网络提供有效的地址前缀。本地 ISP 还很可能是更大型 ISP 的用户，本地 ISP 向它的 ISP 申请地址前缀。因此，一般只有最大型的 ISP 需要和地址注册商联系。

申请和分配分类 IP 地址时，应充分考虑物理网络的大小，根据网络中已经或将要包含的主机数申请合适类别的 IP 地址。表 5-1 总结了每个 IP 地址类的点分十进制值的范围。

表 5-1 每个 IP 地址类的点分十进制值的范围

类别	最低地址	最高地址	备注
A	1.0.0.0	127.0.0.0	网络号 127 用于回送地址
B	128.0.0.0	191.255.0.0	128.0.0.0 不会被分配
C	192.0.0.0	223.255.255.0	192.0.0.0 不会被分配
D	224.0.0.0	239.255.255.255	组播地址

每个类中的地址值并不是全都可供分配。例如 A 类的网络号 127 保留用于**回送地址**，用于测试 TCP/IP 协议软件以及本机进程间的通信。发送到网络号 127 的分组永远不会出现在任何网络上。B 类地址中最先被分配的网络号是 128.1，C 类地址中最先被分配的网络号是 192.0.1。

由于 IP 地址是一个网络标识符和该网络上一个主机标识符的编码，因此 IP 地址不仅指明单个计算机，还指明了计算机到一个网络的连接。例如连接 2 个物理网络的路由器，有 2 个 IP 地址，其中的网络号互不相同，分别标识网络连接所属的物理网络。因此，准确地说，IP 地址标识的是网络连接。

【例 5-1】 设某单位有 3 个物理网络，分别分配了 128.9.0.0、128.10.0.0、128.11.0.0 三个 B 类 IP 地址，连接情况如图 5-3 所示，请给图中的主机和路由器分配 IP 地址。

解：主机和路由器的 IP 地址分配示例见图 5-3。例如，路由器 R2 互连了 128.9 和 128.10 两个网络，其两个接口的 IP 地址分别设置为 128.9.0.21 和 128.10.0.20。路由器 R3 互连了 128.9 和 128.11 两个网络，其两个接口的 IP 地址分别设置为 128.9.0.22 和 128.11.0.20。连接到 128.11 网络的主机 H3 分配了 IP 地址 128.11.0.3。

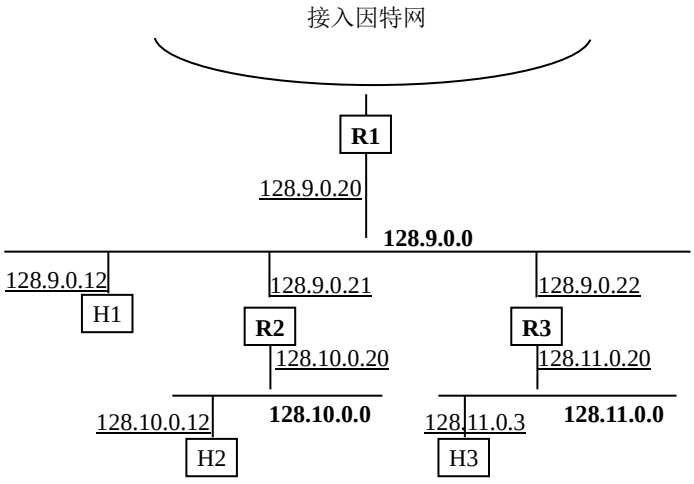


图 5-3 IP 地址分配示例

另外，有些特殊形式的 IP 地址只能在特定情况下使用，见表 5-2。

表 5-2 特殊形式的 IP 地址

net-id	host-id	用作源地址	用作目的地址	说明
0	0	可以	不可以	启动时源站地址
全 1	全 1	不可以	可以	本地网受限广播
net-id	全 1	不可以	可以	定向广播
127	任意(常为 1)	可以	可以	回送地址

5.2.3 IP 数据报

TCP/IP 技术是为包容物理网络技术的多样性而设计的，而这种包容性主要体现在 IP 层中。TCP/IP 的重要思想之一就是通过 IP 数据报和 IP 地址将物理网络统一起来，达到隐藏底层物理网络细节、提供一致性的目的。IP 数据报（简称数据报）是因特网的基本传送单元，它提供对物理网络帧的统一。

与典型的物理网络帧类似，数据报划分为首部和数据区，而且也包含源地址和目的地址，当然数据报首部中包含的是 IP 地址，而物理帧首部中包含的是物理地址。数据报要封装在物理帧中作为帧的数据传送，对于以太网，帧类型字段值为 0x0800 表示帧数据区存放的是 IP 数据报。IPv4 数据报的格式如图 5-4 所示。

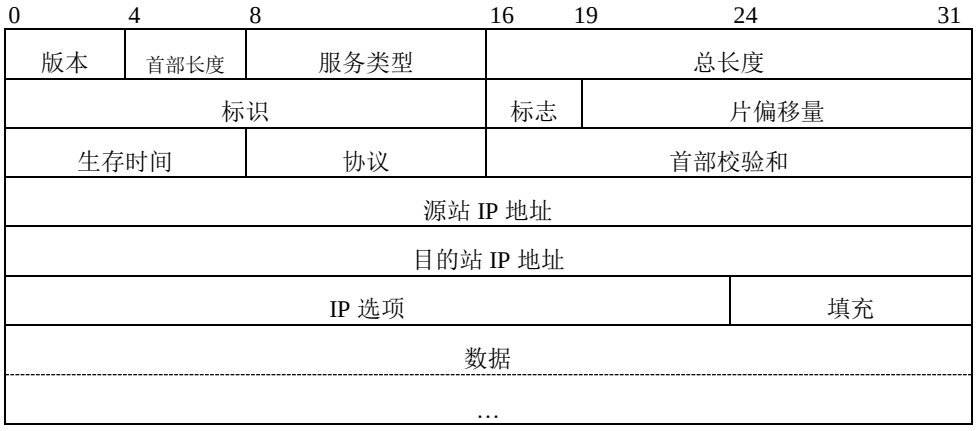


图 5-4 IPv4 数据报格式

数据报首部包含 20 字节的固定部分和可选的 IP 选项部分。下面介绍首部各字段的含义。

（1）版本　占 4 比特，包含了创建数据报所用的 IP 协议的版本信息。目前广泛使用的版本号是 4，IPv4 即表示版本 4 的 IP 协议。IPv6 网络目前也在发展，IPv6 有相同的版本字段，其余字段有所不同。本章中的 IP 除非特别说明，都是指 IPv4。

（2）首部长度　占 4 比特，以 4 字节为单位的 IP 首部长度。IP 首部的最大长度为 15*4 个字节。数据报首部长度必须是 4 字节的整数倍，有 IP 选项时可能需要在填充字段中填 0 来保证。由于选项字段很少使用，所以最常见的首部长度是 20 字节，字段值为 5。

（3）服务类型(TOS)　占 8 比特，指明应当如何处理数据报。这个字段最初用来指定数据报的优先级和期望的路径特征（低时延、高吞吐量或高可靠性）。在上世纪 90 年代 IETF 重新定义了该字段的含义，用于提供对分组的区分服务（Differentiated Service，简称 DiffServ）。新定义将 TOS 前 6 比特作为码点（codepoint，也称 DSCP），后 2 位保留未用。一个码点值被映射到一个底层服务定义。无论使用最初的 TOS 解释还是修改后的区分服务解释，在数据报中指明某种服务级别，仅仅是提供给转发算法作参考，转发软件必须在当前可用的底层物理网络技术中进行选择，并且必须符合本地策略，并不能保证沿途路由器都接受并响应这种服务级别的请求。

（4）总长度 占 16 比特，以字节来单位的整个数据报的长度。

IP 数据报总长度理论上可以达到 65535 字节。但数据报从一台机器传送到另一台，总是要通过底层的物理网络进行传输。而每种分组交换技术都规定了一个物理帧所能传送的最大数据量，称为最大传送单元（MTU）。例如以太帧的 MTU 为 1500 字节，即一个以太帧至多传送 1500 字节的数据；有些硬件技术的传送限制是 128 字节。为了使互联网传输更高效，一般尽量使每个数据报尽可能长并且能封装在一个独立的物理帧中发送。一个数据报在从源站到目的站的过程中，可能会穿过 MTU 不尽相同的多个物理网络。如果把数据报的大小限制成互联网中最小可能的 MTU，会令所有能够运载更大长度帧的网络不能充分发挥作用。

为隐藏底层网络技术并方便用户通信，TCP/IP 软件并没有设计受物理网络限制的数据报。而是根据源站所在网络的 MTU，以及高层协议数据的大小，选择一个合适的初始数据报大小，所谓合适指在源站所在物理网络上能进行最大限度的封装。此外，提供一种机制，在数据报需要经过 MTU 小于数据报长度的网络时，把数据报分解成若干较小的片（fragment），数据报分解的过程称为分片（fragmentation）。每个数据报片都封装在单个物理帧中发送，并且作为独立的数据报进行传输。而且在数据报片到达目的站之前，如果需要还可被再次（多次）分片，但在沿途路由器上不进行重装（reassemble，也称重组）。TCP/IP 规定所有的片重装在目的站进行。

（5）标识　16 比特整数，源主机赋予数据报的惟一标识符。实现方法比如：在源主机的内存中保持一个全局计数器，每产生一个新数据报，计数器加 1，值达到 65536 时置为 0，将计数器的值分配给新数据报。总之要保证（在较长一段时间内）同一主机发出的各数据报的标识是惟一的。一个数据报分片，其实是**分割数据报的数据部分**，数据报片的首部主要从初始数据报首部中复制，仅做少量修改，标识字段必须不加修改地复制到各个分片中，以方便重装时识别属于同一初始数据报的所有分片。

（6）标志　占 3 比特，只有低两位有效。中间一位为“不分片”（Don't Fragment flag，简称 DF）比特，置 1 时表示数据报不能被分片，为 0 时表示数据报允许被分片。当路由器必须对数据报分片才能转发，而该数据报的 DF 又被置位(为 1)时，路由器将抛弃该数据报，并向其源主机发送一个 ICMP 差错报告。3 个比特中的最低位是“更多分片”（More

Fragments flag，简称 MF），置位时说明该数据报不是最初始数据报的最后一个分片，该位复位（为 0）时表示是最后一个分片。

（7）片偏移量　占 13 比特，指出本数据报片中数据相对于最初始数据报中数据的偏移量，以 8 个字节为单位计算偏移量。还没被分片的数据报或者第 1 个数据报片的偏移量为 0。 由于各片按独立数据报的方式传输，无法保证按序到达目的主机，而目的主机能够根据分片中的源站 IP 地址、标识、偏移量以及 MF 字段重装出最初始数据报的完整副本，除非没能收齐所有分片。

注意，因为片偏移量以 8 字节为单位，所以除最后一个分片外，其余分片的数据部分的大小应尽量接近但不超过网络 MTU 并且是 8 字节的整数倍，最后一个分片可以较其他片小。图 5-5 给出一种可能发生分片的互联网，其中 A 和 B 两主机分别直连到 MTU 为 1500 字节的以太网上，A 和 B 之间通信需要穿越 MTU 为 660 字节的网络。如果 A 向 B 发送一个长度超过 660 字节的数据报，则路由器 R1 需要把数据报分片，反之类似。

图 5-5 可能发生分片的情形示例

【例 5-2】假定上图中 A 向 B 发送了一个首部 20 字节，数据区 1400 字节长，DF 为 0 的数据报，R1 向 R2 转发时要先把数据报分片，再分别封装在物理帧中发送，请写出片结果。

解：分片结果如图 5-6 所示。

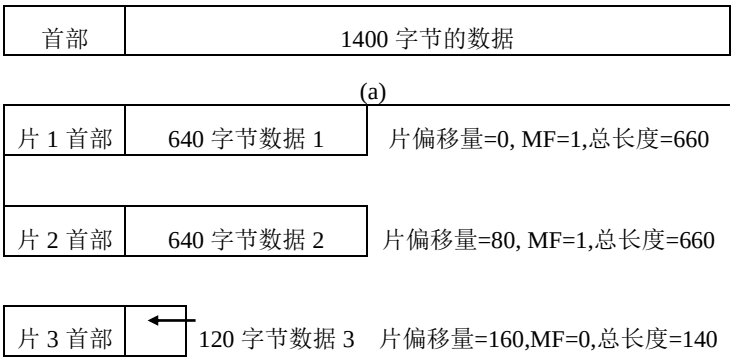


图 5-6 分片示例 (a)初始数据报; (b)在 MTU=660 字节的网络上的 3 个分片

（8）生存时间（Time To Live，简称 TTL）　占 1 个字节，设计初衷是用来指明数据报在互联网系统中允许保留的时间（以秒为单位）。由于路由器刚出现时速度慢，所以过去标准规定，如果路由器让一个数据报滞留了 K 秒，则应把 TTL 字段的值减去 K。但现在的路由器和网络完成一个数据报的转发一般仅需要几毫秒。因此，现在 TTL 实际起着“跳数限制”的作用，而不是延迟时间的估计。数据报每经过一个路由器，路由器就将其 TTL 值就递减 1，并且一旦 TTL 减为 0，路由器就不再转发该数据报，而是予以丢弃，并向数

据报的源站发送一个 ICMP 差错报告。

（9）协议　1 个字节的整数，指明数据报数据区的格式，即数据报封装了哪个协议的协议数据单元，以便目的站的 IP 软件知道应将数据交由哪个（高层）协议软件处理。协议和协议字段值的映射由中央管理机构（NIC）统一管理，确保在整个因特网内保持一致。表 5-3 列出了一些协议与规定的协议编号。

表 5-3 指定的网际协议编号

协议字段值	1	2	3	4	6	8	17	88	89
协议名	ICMP	IGMP	GGP	IP	TCP	EGP	UDP	IGRP	OSPF

（10）首部校验和　占 16 比特，用于首部的校验。校验和算法：设校验和字段初值为 0，再把首部看成一个 16 位整数序列，对所有整数进行反码求和（其规则是从低位到高位逐位进行计算。0＋0＝0；0＋1＝1；1＋1＝0，但要产生一个进位。如果最高位产生进位，则结果要加 1），得到的和的二进制反码就是校验和的值。数据报从源站发出后，沿途路由器及目的站都要检验首部校验和，如果检验失败，数据报将被立即丢弃。检验方法同校验和的计算，运算结果为 0 表示首部没有变化，否则表示有错。校验和要随首部任何字段的变更而重新计算，例如分片后要为各数据报片算校验和，再如所有数据报的 TTL 字段在转发节点处都要被减 1，因此路由器对每个被转发的数据报都要重算校验和。

网际协议不提供可靠通信功能，端到端或点到点之间没有确认，也没有对数据的差错控制，只检验首部，并且没有重传，没有流控。只有首部校验和的优点是大大节约了路由器处理每个数据报的时间，符合 IP“尽力传递”的思想。缺点是给高层软件留下了数据不可靠的问题，增加了高层协议的负担。不过 IP 数据报首部和数据区的分开校验允许高层协议选择自己的校验方法。

（11）源站 IP 地址和目的站 IP 地址　也称为源 IP 地址和目的 IP 地址，各占 4 字节，分别指明本数据报最初发送者和最终接收者的 IP 地址。数据报经路由器转发时，这两个字段的值始终保持不变，即使被分片转发。路由器总是提取目的站 IP 地址与路由表中的表项进行匹配，以决定把数据报发往何处。

（12）IP 选项　长度可变，主要用于控制和测试两大目的。要求主机和路由器的 IP 模块均支持 IP 选项功能。每个数据报中选项字段是可选的。为保证数据报首部长度是 32 位的整数倍，可能需要填充字段包含一些 0 比特。IP 选项不常用，因此 IPv4 数据报首部长度一般都为 20 字节。

一个数据报中可以包含多个选项。有实际意义的选项含有 1 字节的选项类型（option-type）、1 字节的选项长度（option-length）和若干字节的选项数据（option-data）。有 2 个仅含有选项类型的单字节选项。一个用于放在选项表的末尾使 IP 数据报首部长度是 32 比特的整数倍，可放多个。还有一个单字节选项，用于放在选项之间，对齐选项使其长度都是 32 比特的倍数。

选项类型八位组分为拷贝标志(copied flag)、选项类(option class)、选项号(option number) 3 个字段：

拷贝标志	1 bit	1: 表示分片时本选项拷贝到所有分片中； 0: 表示分片时本选项仅复制到第 1 个片中
选项类	2 bits	0: 控制；2: 诊断和测量；1 或 3: 保留暂未使用
选项号	5 bits	指明选项类中某个具体选项

可用的选项类与选项号列表可参见 RFC791。这里简单介绍 2 个较受关注的选项功能：

（1）记录路由选项　选项类型为 135，该选项允许源主机在选项部分创建一个 IP 地址列表空间，沿途路由器处理该数据报时将其 IP 地址填入选项的地址列表中，以此跟踪数据报所走的路线。

（2）严格源路由选项（Strict Source Routing）选项类型为 137，该选项允许源站指明本数据报的确切路径。选项数据是一条路线的数据，即各跳的 IP 地址列表。

例 5-3 在某主机（IP 地址为 10.10.1.95）上用网络监听工具监测网络流量，获取的一个 IP 包的前 28 字节用十六进制表示如下：

45 00 00 47 E6 EE 00 00 67 11

19 2A 75 4E D2 D6 0A 0A 01 5F

A4 CA 0D 4B 00 33 6B 26

请解析 IP 包各字段。

解：根据 IP 数据报的格式分析，以上各字节与各字段的对应关系如图 5-7 所示。第 1 个字节为版本号和单位为 4 字节的 IP 首部长度，因此算得版本号为 4，IP 首部长度为 5×4=20 字节；总长度为$(00\ 47)_{16}$=71 字节；生存时间为$(67)_{16}$=103；协议为$(11)_{16}$=17，表示 IP 包数据部分是 UDP 报文；源 IP 地址：75 4E D2 D6，也即 117.78.210.214；目的 IP 地址为 0A 0A 01 5F，也即 10.10.1.95，可见这是主机收到的 IP 包。参考下一章的 UDP 报文格式可以进一步分析下面的 8 个字节。

图 5-7 IP 包解析结果

5.2.4 因特网地址到物理地址的映射

互联网使用 TCP/IP 软件实现物理网络的互连。TCP/IP 软件都使用 IP 地址标识通信主机。IP 地址将不同的物理地址“统一”起来，不过，地址统一的代价是需要建立 IP 地址和物理地址之间的映射。因为 IP 层以上各层均使用 IP 地址，但在物理网络内仍使用各自

的物理地址，互联网并不做任何改动。

考虑连接到同一物理网络的主机 A 和 B，设 A 和 B 分配得到的 IP 地址分别为 I_A 和 I_B，物理地址分别为 P_A 和 P_B。TCP/IP 的设计目标是隐藏物理网络细节，高层的程序仅利用 IP 地址进行通信，因此 A 上应用程序要向 B 的应用程序发送 IP 分组，只需知道 B 的 IP 地址。不过，IP 分组由 A 传到 B 必须依靠物理网络来实现，而物理网络中两台机器之间的通信必须使用硬件地址（物理地址）。由此产生了问题，即 A 如何将 B 的 IP 地址 I_B 映射为 B 的物理地址 P_B 呢？

如果通信双方 A 和 B 不在同一个物理网络，则 IP 分组从 A 发到 B 需要依赖沿途的路由器进行转发。每个主机和路由器都有路由表，指明到目的网络的路由。例如，通过查询本机的路由表，A 知道应该将分组发给本地路由器 R1，其 IP 地址为 I_{R1}，由 R1 再进行转发。由上一段的讨论可以知道 A 向 R1 发送 IP 分组需要获悉 R1 的物理地址。同样 R1 通过查路由表可以知道下一个路由器 R2 的 IP 地址，然后也需要进行地址映射，获悉 R2 的物理地址。同理，A 至 B 路径上的最后一个路由器需要由 B 的 IP 地址获悉 B 的物理地址。总之，协议软件需要一种机制将一个 IP 地址映射为相对应的硬件地址，这种把高层地址映射为物理地址的问题称为**地址解析**问题。

1. 直接映射方法

TCP/IP 采用 2 种地址解析技术：**直接映射法**和**动态绑定法**。通过直接映射进行解析适用于物理地址是易配置的短地址的情形。而对于固定长度的长物理地址，例如以太网地址，则通过动态绑定进行解析。TCP/IP 采用**地址解析协议**（Address Resolution Protocol，简称 ARP）完成动态地址解析。

如果网络硬件的硬件地址是可配置的，而且可以使用小整数，那么可以给网络内计算机顺序分配地址，例如给网络上的第一台计算机分配地址 1，给第二台分配地址 2，依此类推。我们已经知道，给一个网络内的计算机分配 IP 地址的要点是，使用相同的网络号，主机号部分任意分配，互不重复即可。那么假定一个网络的网络号是 202.119.211.0，则可以给其中的硬件地址为 1 的计算机分配 IP 地址 202.119.211.1，给硬件地址为 2 的计算机分配 IP 地址 202.119.211.2。也就是，将计算机的硬件地址编码到 IP 地址的低 8 位中。由于 IP 地址包含了硬件地址编码，因此地址解析极其简单，只需通过提取 IP 地址的低 8 位就可获得相应的物理地址，这样完成的地址解析称为**直接映射**。还可以采用不同于上述的方法将硬件地址融入 IP 地址，只要能够从 IP 地址计算出物理地址即可，不过，IP 地址和硬件地址之间的关系越简单，直接映射的效率越高。

2. 动态绑定方法——ARP

虽然直接映射是高效的，但将 48 位的以太网地址编入 32 位的 IP 地址实在不可行。因此对有广播能力的以太网，使用 ARP 通过动态绑定进行地址解析。基本思路很简单：当主机 A 需要解析本网络内主机 B 的 IP 地址 I_B 时，A 先广播一个特殊的分组，请求 IP 地址为 I_B 的主机 B 将其物理地址告诉 A。网内所有主机都接收到这个请求，但只有主机 B 发现是在问自己（分组中指明了 I_B），所以向 A 单播发出一个含有自己物理地址的分组作为响应。

并非 A 每次向 B 发送分组前，都要先广播一个 ARP 请求以获悉 B 的物理地址，再利用物理网络发送 IP 分组。实际上，为降低通信费用，使用 ARP 的计算机各维护着一张 ARP

表，ARP 表在高速缓存中，存放着最近获得的 IP 地址与物理地址的绑定。为防止绑定陈旧，每个绑定都设有超时计时器，典型的超时时间是 20 分钟，过时的表项将被删除。当 A 要向其他主机发送 IP 分组时，总是先在 ARP 高速缓存中寻找所需的绑定，如果找不到，才向网络广播 ARP 请求，响应到达时再发送所有等待该解析结果的 IP 分组。

ARP 报文格式相当通用，能够适用于任何物理地址和协议地址。图 5-8 给出了 ARP 报文格式，其中的 4 个地址字段占用字节数不固定，取决于硬件类型和协议类型，并由地址长度字段明确指出。ARP 报文中的硬件类型字段指明物理网络类型，值为 1 表示是以太网。协议字段指明高层协议地址类型，值为 0x0800 表示是 IP 地址。操作类型字段指明本 ARP 分组是 ARP 请求（值为 1）、ARP 响应（值为 2）、RARP 请求（值为 3），还是 RARP 响应（值为 4）。硬件地址长度和协议地址长度字段分别指出了硬件地址和高层协议地址的长度，这使得 ARP 能够在任意网络中使用。以太网硬件地址为 6 个八位组（字节）长，IP 地址为 4 个八位组长。

设 A 为了向 B 发送 IP 分组而要解析 B 的 IP 地址，则 A（发送方）应在 ARP 请求报文的目标协议地址中填上 B 的 IP 地址 I_B，为使 B 能够向 A 单播 ARP 响应，A 还应在 ARP 请求报文的发送方硬件地址和协议地址中分别填上 P_A 和 I_A。

2 字节的硬件类型	2 字节的协议类型	1 字节硬件地址长度	1 字节协议地址长度	2 字节的操作类型	发送方硬件地址	发送方协议地址	目标硬件地址	目标协议地址
1: 以太网 6: IEEE 802 网络 15:帧中继	0x0800: IP	6: 以太网硬件地址长度	4: IP 地址长度	1: ARP 请求 2: ARP 响应 3: RARP 请求 4: RARP 响应				

图 5-8 ARP 报文格式

ARP 请求报文封装在物理网络帧中（作为帧的数据）广播出去以后，网络上的任何计算机都能收到。对于以太网，帧的类型为 0x0806（表示封装了 ARP 报文），目的地址是广播地址，源地址是 A 的物理地址 P_A。接收方的 ARP 软件将首先提取发送方的硬件地址和 IP 协议地址，并检查本地高速缓存，查看 ARP 表中是否已存在该发送方的地址绑定，如果有，则用 ARP 请求中的发送方硬件地址覆盖该表项中的物理地址，并复位该表项的计时器。

接着，接收方检查 ARP 请求中的目标协议地址是不是与本机 IP 地址匹配，如果不是，则可停止处理该 ARP 请求。如果匹配，则将本机的物理地址 P_B 填入报文中所缺的目标硬件地址字段，并交换发送方和目标地址对，然后把操作类型字段值改成 2（响应），再将该 ARP 响应报文封装在物理网络帧中单播发给 A。对于以太网，帧的类型为 0x0806，目的地址为 P_A，源地址为 P_B。

ARP 响应报文仅有一个接收方，接收方 A 将 ARP 响应中发送方的 IP 地址和物理地址绑定写入 ARP 高速缓存中。然后 A 就可以用该绑定中的物理地址作为帧的目的地址封装

待发 IP 分组，并将其发送了。

3. 逆地址解析协议——RARP

逆地址解析（Reverse Address Resolution Protocol，简称 RARP）用于将物理地址映射为 IP 地址。RARP 目前在因特网中基本不再使用，但它过去曾是无硬盘工作站自引导系统所使用的重要协议。RARP 允许系统在启动时获得一个 IP 地址，过程如下：在系统启动时，广播发送一个 RARP 请求，请求中包含本机的硬件地址，然后等待 RARP 服务器的响应，响应报文中给出请求方的 IP 地址。RARP 报文的格式与 ARP 报文的格式一样，仅仅是操作类型字段值不同。RARP 请求报文与响应报文各字段值的设置与 ARP 的类似。另外，封装了 RARP 报文的以太帧类型字段值应为 0x8035。

5.2.5 差错与控制报文（ICMP）

本小节介绍 IP 数据报传送过程中的差错监测机制。采用分组交换技术的网络不可能在任何时候都能运转正常，错误总是难免的。对于互联网，除了存在通信线路和处理器故障外，主机或路由器临时或永久的网络连接断开、路由器拥塞得无法接收或处理数据报、路由表有误导致出现了路由环路（routing cycle）都可能导致数据报交付的失败。因此互联网需要差错检查与纠正机制。

为提供高效率的尽力而为服务，IP 协议仅通过 IP 首部校验和提供一种传输差错检测手段，并没有提供差错纠正机制，而是让高层协议（如 TCP）解决各种差错。虽然不直接纠错，但网际互连层有个 IP 的补充协议 ICMP，它提供一种差错报告机制，用于路由器或目的主机把发生的交付问题或路由问题通告（发送 ICMP 报文）给源站。源站必须将差错告诉给某单独的应用程序，或者采取其他措施来纠错。此外，ICMP 还包括提供信息功能。在每个 IP 实现中都必须包含 ICMP。

为什么 ICMP 报文仅发给引起问题的数据报的源站呢？原因是数据报只含有源、目的站的 IP 地址，并不包含所走路径的完整记录（除非数据报使用了记录路由选项），而且实在无法确定究竟路径上的哪个节点该对问题负责。

ICMP 报文的传递需要 IP 的支持，即每个 ICMP 报文要封装在 IP 数据报中，源 IP 地址为发送报告的机器的 IP 地址，目的 IP 地址为出现差错的数据报的源站地址。因为一个 ICMP 报告可能要经过多个物理网络才能到达目的地，所以必须封装在 IP 数据报中，进而封装在帧中发送出去。ICMP 的两级封装如图 5-9 所示，其中帧的类型字段值为 0x0800，IP 数据报的协议字段值为 1，表示数据是 ICMP 报文。ICMP 是 IP 的必要组成部分，因此不把它当成高层协议。

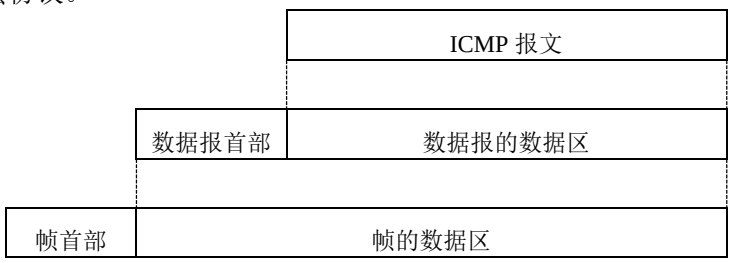


图 5-9 两级 ICMP 封装

ICMP 报文一般报告在数据报处理中遇到的差错。但为避免对报告再产生报告，携带

了 ICMP 报文的数据报发生差错时不再发送 ICMP 报文。另外，仅在对片偏移量为 0 的分片处理过程中才可能发送 ICMP 报文，对其余分片处理不发差错报告。

ICMP 报文分为两大类：ICMP **差错报告报文**和**提供信息的报文**。每个 ICMP 报文有自己的格式，前 3 个字段格式统一：1 字节的类型、1 字节的代码和 2 字节的校验和。类型用于标识报文类型，代码表示有关本类型的更多信息。校验和算法与 IP 首部校验和相同，不过是计算整个 ICMP 报文的校验和。此外，报告差错的 ICMP 报文总是复制了产生问题的数据报的首部和前 64 比特数据（包含重要信息），以便让接收方能够更准确地判断应由哪个协议及应用程序对已发生的差错负责。下面介绍部分差错报告报文和提供信息的报文，更多内容请参见 RFC792。

（1）目的不可达报文

目的不可达 ICMP 报文的格式如图 5-10 所示，类型为 3。代码进一步描述问题：

0：网络不可达　　1：主机不可达

2：协议不可达　　3：端口不可达

4：需要分片但 DF 被置位　　5：源路由失败

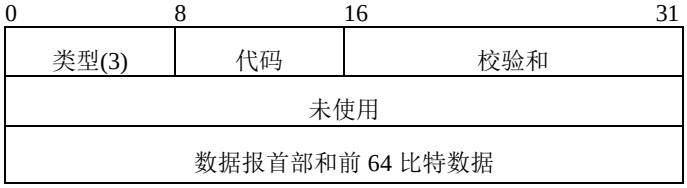


图 5-10 ICMP 目的不可达报文格式

如果根据路由器的路由表，一个数据报的目的站 IP 地址所指定的网络是不可达的，比如到那个网络的距离是无穷的，则路由器可向该数据报的源主机发送代码为 0 的目的不可达报文。

在目的主机中，如果数据报指定的协议模块不在活动，IP 模块将无法交付数据报中的数据，则目的主机会向源主机发送代码为 2 的目的不可达报文。

当路由器必须对一个数据报进行分片才能将其转发，而数据报的 DF（不分片）标志为 1 时，路由器将丢弃数据报，并向源主机返回一个代码为 4 的目的不可达差错报告。

代码为 0, 1, 4 和 5 的目的不可达报文一般由路由器发出，而代码为 2 和 3 的一般由主机发出。

（2）超时报文

ICMP 超时报文的格式与目的不可达报文的相同，只是类型值为 11。代码说明超时的性质：0 表示转发中 TTL（生存时间）超时，1 表示分片重装超时。

路由选择协议用于维护更新路由表，路由表难免有时会有差错，差错可能导致数据报“兜圈子”，例如数据报被路由器 R1 转发给 R2，经过数跳又转发给了 R1。为了避免数据报在因特网中无休止地兜圈子而到不了目的站，IP 规定：路由器在转发数据报前要先将其 TTL 字段减 1，一旦 TTL 为 0，则丢弃之，并借助超时报文（代码为 0）通知数据报的源主机。

目的主机负责分片重装，即收集一个数据报的所有分片并组装成完整的数据报。主机在收到某数据报的第 1 个分片后就启动一个重装计时器。如果在计时器超时前没有收齐所有的分片，则主机将丢弃已收到的分片。发生超时时如果已收到片偏移量为 0 的分片则向源主机发送超时报文（代码为 1），否则不发。

还有一些差错报告报文：参数问题报文（类型 12）、源站抑制报文（类型 4）、重定向报文（类型 5）。注意对 ICMP 差错报告报文是不需要进行反馈的，仅仅起着报告的作用。

（3）回应请求与应答报文

回应请求与应答是格式相同的一对报文，见图 5-11。它们仅类型值不同，类型字段值为 8 表示报文是回应请求，为 0 表示是回应应答。数据字段长度可变，可以是任何数据，回应应答返回的数据总是与收到的回应请求中的数据完全相同。标识符（Identifier）字段和序号（Sequence Number）字段被发送方用来匹配应答与请求。回应应答报文可以来自路由器或主机。

图 5-11 ICMP 回应请求与应答报文的格式

回应请求与应答主要用于测试目的站的可达性，也可以通过计算发出请求和收到应答之间的时间差来估计源和目的主机之间的往返时延。另外，通过适当设置封装回应请求报文的数据报的 TTL 值，还可以实现路径跟踪功能。

提供信息的 ICMP 报文对还有：时间戳请求与应答、地址掩码请求与应答等。

例 5-4 分析 Windows 操作系统上提供的路径跟踪工具的实现原理。

解：在 Win2K 上运行路径跟踪工具 tracert 跟踪从本机到 Web 服务器 www.edu.cn 的路径，同时运行监听工具，如 Sniffer，监测 tracert 产生的流量。注意观察 IP 包首部中的协议和 TTL，以及 ICMP 报文中的类型和代码取值。实现原理请自行推理，并请完成有关习题。

5.2.6 子网编址

最初的 IPv4 编址方案把 IP 地址分成 2 部分，前缀作为网络部分，后缀作为主机部分，并规定每个物理网络都要被分配一个惟一的网络地址。一个物理网络上，每个主机的 IP 地址都有共同的前缀。因特网设计之初个人计算机还不曾出现，因此设计人员没有预见到因特网的发展速度：每隔 9~15 个月，其物理网络数（已分配的分类 IP 网络地址数）就翻一番。

到 20 世纪 80 年代中，就发现了分类 IP 网络地址将不够用。此外，已分配的地址并没有得到充分利用。例如一个 B 类 IP 网络地址可以给 6 万多个主机编址，但实际上为了网络性能较好，避免网络拥塞，一个 LAN 并不能连接如此多的主机，因而给一个 LAN 使用一个 B 类 IP 网络地址会导致地址空间的利用率极低。在不摒弃分类编址的情况下，如何

适应网络增长的需要呢？设计人员主要提出了 3 种技术：子网编址、代理 ARP、无编号的点对点链路，它们的动机都是减少使用网络前缀数量。

20 世纪 90 年代，又创造了无分类编址方案进一步提高 32 比特地址空间的利用率，以及提高路由查找效率。

下面阐述子网编址技术和支持子网编址的 IP 转发算法。

1. 子网划分

对于一个中等大小的组织，比如有若干大楼的大学或公司，鉴于 LAN 技术的限制，一般需要构建若干 LAN 来覆盖本地区域。对于这种情况，TCP/IP 设计人员想到可以给这样的**网点**分配一个 IP 网络地址，再从主机号部分借用几比特来标识各个子网（各个 LAN）。这种允许一个分类网络地址供多个物理网络使用的技术称为子网编址（subnet addressing）或划分子网（subnetting），相应更新的 IP 转发技术称为子网转发（subnet forwarding）。最初的 IP 编址方案中没有子网的概念，现在划分子网的思想已经融入到当前的无分类编址方案中了。

划分子网技术使得多个物理网络可以共用一个网络前缀。将 IP 地址的后缀分成 2 个字段，分别用于标识物理网络和网络上的主机。具体方法如图 5-12 所示。IP 地址原有的前缀解释为因特网部分，用于标识网点，该网点可能包含多个物理网络。而原有的后缀解释为本地部分，因特网中的路由器在作转发决策时照例只用看网络前缀。本地部分的具体分配与后缀一样留给本地网点，网点上的所有主机和路由器知道本网点的子网划分方案，而这些对于网点之外的路由器可以是透明的，即它们可以认为这个网点仅有一个物理网络。本地部分的一部分用于标识网点上的物理网络，剩余部分用于标识给定物理网络上的主机。

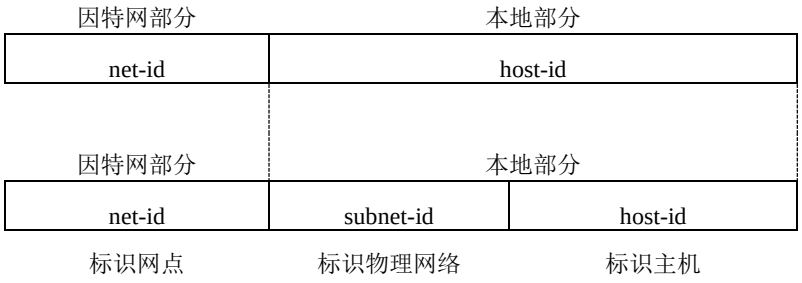


图 5-12 划分子网时 IP 地址结构

TCP/IP 的子网编址的标准允许各网点根据具体情况灵活选择子网划分方案。应当根据网点的拓扑及每个网络的主机数决定如何分割后缀部分。

【例 5-5】一个包含 5 个物理网络的单位拥有一个 B 类网络地址 130.27.0.0，每个网络中主机不超过 1000 台，该如何划分 B 类 IP 地址的主机号部分呢？

解：在分类编址方案中，默认情况是不划分子网的，即一个分类网络地址仅用于 1 个子网，设某分类 IP 地址的后缀有 y 位（B 类地址的 y=16），则惟一的子网中的主机数最高可达 2^y-2。若从主机号字段划分出 3 比特作为子网号，则一个 B 类网络地址可用于 2^3-2 个子网，子网中的主机数最高可达 2^{16-3}-2。同理，假定各子网的子网号长度一样，设子网号

占 x（x≥2 且 x≤y-2）比特，则最多允许有 2^x-2 个子网，每个子网最多有 2^{y-x}-2 台主机。注意一般要求避免使用全 0 和全 1 的子网号和主机号。所以子网号位数至少为 2，以免没有可分配的子网号；子网号位数必须小于等于 y-2，也即主机号位数大于等于 2，否则没有可分配的主机号。

一个 B 类地址的所有定长子网划分方法如表 5-4 所示。对于本例，查表 5-4 可知满足条件（能包含 5 个子网），且每个子网的主机可达 1000 台的共有 4 种选择，见表中带阴影的 4 行，即子网号字段占 3~6 位都可以满足条件。若选择子网号长度为 3，则子网号可以为 001、010、011、100、101、110 中的任意 5 个。

表 5-4 一个 B 类地址的所有定长子网划分方案

子网号长度	子网数	每个子网的主机数
0	1	65 534
2	2	16 382
3	6	8190
4	14	4094
5	30	2046
6	62	1022
7	126	510
8	254	254
9	510	126
10	1022	62
11	2046	30
12	4094	14
13	8190	6
14	16 382	2

大多数划分子网的网点都采用了定长的分配方案。具体确定子网号占几比特由各网点自己确定，各子网号所占位数一致，各子网所能容纳的主机数一致。有时候，一个网点内的物理网络大小也很不均衡，有的主机多，有的包含很少的主机，采用固定长度的子网划分就显得地址空间利用得不够合理。TCP/IP 子网标准允许使用变长划分子网（Variable-Length Subnet Masks, VLSM）技术，允许为一个网点的各个物理网络挑选长度不一的子网号。采用 VLSM 分配地址，比较困难，容易出现地址二义性；优点是灵活，支持网点内大小网络的混合，并能够更充分利用地址空间。

标准要求用 32 比特的**子网掩码**来表示划分方案。无论使用定长或变长的配置方案，使用子网编址的网点必须为每个网络设置一个子网掩码。子网掩码中的 1 表示主机 IP 地址的对应比特是网络号和子网号部分。标准没有规定必须从主机号的高位起选择连续相邻的若干比特作为子网标识标志物理网络，但实践中还是推荐如此，并且建议在所有共享同一 IP 网络地址的各物理网络中使用相同的掩码。

【例 5-6】一个包含 5 个物理网络的单位拥有一个 B 类网络地址 130.27.0.0，每个网络中主机不超过 1000 台，请划分子网，并写出每个子网的子网地址、子网掩码、子网中的最小/最大主机地址及子网广播地址。

解：不妨选用子网号占用 3 比特的方案，并将原主机号部分的前 3 比特作为子网号部分，则各子网的子网掩码为 255.255.224.0。给每个物理网络指定一个子网号，子网地址和每个子网可用的主机地址见表 5-5，注意子网号也可选用“110”。

表 5-5 子网编址示例：一个 B 类地址用于包含 5 个物理网络的网点

子网号	子网地址	子网掩码	子网中最小主机地址	子网中最大主机地址	子网广播地址
001	130.27.32.0	255.255.224.0	130.27.32.1	130.27.63.254	130.27.63.255
010	130.27.64.0	255.255.224.0	130.27.64.1	130.27.95.254	130.27.95.255
011	130.27.96.0	255.255.224.0	130.27.96.1	130.27.127.254	130.27.127.255
100	130.27.128.0	255.255.224.0	130.27.128.1	130.27.159.254	130.27.159.255
101	130.27.160.0	255.255.224.0	130.27.160.1	130.27.191.254	130.27.191.255

5.2.7 无分类编址与 CIDR

虽然子网编址和无编号网络能够节省 IP 网络地址，但到 1993 年，因特网的增长速度还是让人们感觉这些技术无法阻止地址空间的耗尽。此外，因特网还即将面临 B 类网络地址空间的耗尽和路由信息过量等问题。因缺乏适于中等大小组织所需要的网络类而导致 B 类地址消耗得快，毕竟一个 C 类地址仅有 254 个主机地址，所以一般单位更愿意申请 B 类地址。然而很少单位有 6 万多主机，因此导致即使划分子网 B 类地址也未能得以充分利用。另外，随着大量网络前缀的分配，路由器的路由表大小和增长速率也即将使当时的软件无法有效管理。

于是人们开始定义含有更多地址的新版 IP 协议 IPv6，并发明了一种称为**无分类域间路由选择**（Classless Inter-Domain Routing，简称 CIDR，读作“sider”）的新技术作为在新版 IP 被正式采纳前的过渡方案。1993 年发布的有关 CIDR 的 RFC 文档为 RFC1517~RFC1520。使用 CIDR 可以更加有效地分配 IPv4 的地址空间，另外可以减缓路由表的增长速度和降低对新 IP 网络地址的需求的增长速度，使得因特网在一定时期内仍能持续增长并高效地运转。

CIDR 最大的特点是采用无分类编址机制，与分类编址相同的是将地址分成前缀和后缀两部分，不同的是前后缀之间的边界不是仅主要有 3 种（1 字节、2 字节、3 字节长的前缀），而是任意的，前缀长度不一，可以是 1 到 32 之间的任意值。与子网编址类似，CIDR 使用 32 比特的地址掩码来指明前缀与后缀之间的边界。掩码中连续相邻的 1 比特对应于前缀，掩码中的 0 比特与后缀相对应。

1. CIDR 地址块

对尚未分配的分类 IP 地址，CIDR 将其看作是一些地址块，每个块内的地址连续。例如 A 类地址 58.0.0.0 和 59.0.0.0 可以看成是一个大小为 2^{25} 的 CIDR 地址块（也称为“25

位的块”），掩码为 254.0.0.0，也可使用 CIDR 记法表示为 58.0.0.0/7，见图 5-13。这个地址块被 IANA 分配给了地址注册商 APNIC(亚太地区网络信息中心)，由它把这些地址划分为若干地址块分配给一些大型 ISP。这些 ISP 会将申请到的地址块根据用户的要求划分成更小的地址块，分给单位或小型 ISP。一个单位将拥有的地址块再根据需要分成若干块（可以大小不等），分配给物理网络。使用 CIDR 后，为了方便路由聚类，减少路由表的项数，应尽量按照网络拓扑和网络所在地理位置来划分地址块。

	点分十进制记法	32 比特的二进制地址
最低地址	58.0.0.0	00111010 00000000 00000000 00000000
最高地址	59.255.255.255	00111011 11111111 11111111 11111111

图 5-13 CIDR 地址块 58.0.0.0/7 示例

地址块的 CIDR 记法也称斜线记法。斜线“/”后的数值 N 表示网络前缀的长度，确切地说有两种含义。对于一个主机的 IP 地址，N 表示地址的前 N 比特是一个具体的网络前缀，惟一标识了主机所在的物理网络；如果作为一个地址块，表示地址块拥有者可以自由分配 32-N 比特的后缀，前 N 比特标识地址块。如果一个 ISP 拥有 N 比特长前缀的 CIDR 块，它可以选择给用户分配前缀长大于 N 比特的任意地址块。这是无分类编址的一个主要优点：能够灵活分配各种大小的块。

【例 5-7】某个 ISP 拥有地址块 202.118.0.0/15。先后有 5 个单位申请地址块，单位 A 需要 1800 个地址，单位 B 需要 900 个，单位 C 需要 900 个地址，单位 D 需要 400 个地址，单位 E 需要 3500 个地址，该怎样分配地址块呢？

解：首先分析各单位的需求，如果不使用 CIDR，则应给每个单位都分配一个 B 类网络地址（这将浪费很多地址）或若干 C 类地址。而使用 CIDR，对于单位 A，1800 个地址需要 11 比特标识主机，因此这个单位的 IP 地址前缀长度应是 32-11=21。同理，单位 B、C、D、E 的网络前缀长度应分别为 22、22、23、20。另外应保证各个单位的前缀是可区分的，不会引起二义性。一种可能的分配方案如图 5-14 所示。

此外，各个单位内部可以根据需要再进行划分，直到给每个物理网络分配一个具体的网络前缀。CIDR 地址块划分机制可以大大缩减路由表的大小。例如若采用分类编址，可以给 A 单位分配 8 个 C 类网络地址，在 ISP 内路由器的路由表中，则需要包含 8 个表项表示到单位 A 的路由。而采用无分类编址，则在 ISP 内路由器的路由表中，仅需要使用一个“超网”路由 202.118.0.0/21。

ISP/单位	地址块	前缀的二进制表示	地址数
ISP	202.118.0.0/15	11001010 0111011*	2^{17}=131072
单位 A	202.118.0.0/21	11001010 01110110 00000*	2^{11}=2048
单位 B	202.118.8.0/22	11001010 01110110 000010*	2^{10}=1024
单位 C	202.118.12.0/22	11001010 01110110 000011*	2^{10}=1024

单位 D	202.118.16.0/23	11001010 01110110 0001000*	2^9=512
单位 E	202.118.32.0/20	11001010 01110110 0010*	2^{12}=4096

图 5-14 CIDR 地址块划分示例

5.2.8 无连接的数据报传送

IP 的目的是提供包含多个物理网络的一个虚拟网络，并提供无连接的数据报交付服务。本节关注 IP 数据报的传送。下面先对互连物理网络完成数据报转发任务的设备——路由器作一简单介绍。

1. 网络层互连设备——路由器

每个路由器与两个以上的物理网络有直接的连接。路由器的每个网络接口（network interface）提供双向通信，包含输入和输出端口。整个路由器结构可分为两大部分：路由选择部分和分组转发部分。路由选择部分简单地说就是按照选定的路由选择协议构造并维护路由表，将在后面介绍。分组转发部分由三个部分组成：交换结构、一组输入端口和一组输出端口。

路由器在输入端口接收 IP 分组，首先按照物理层协议进行比特流的接收，再按照数据链路层协议接收传送 IP 分组的帧，再将帧中的数据报交由网络层模块处理，若网络层模块在忙（查路由表），则数据报被暂存在输入队列中等待处理，排队结束后，网络层模块根据数据报首部中的目的站 IP 地址查找路由表（实质上是匹配目的网络地址），根据查找结果（包括下一跳 IP 地址和输出端口），经过交换结构到达合适的输出端口。

输出端口也设有队列，当交换结构传送过来的分组的到达速率超过输出链路的发送速率时，来不及发送的数据报就暂存在队列中。排队结束后，输出端口中的数据链路层处理模块给 IP 分组加上帧头和帧尾，交给物理层实体后发送到线路上。

路由器中输入或输出队列的溢出是造成分组丢失的重要原因。

2. 直接交付与间接交付

互联网中，每个路由器至少互连 2 个物理网络，即至少与 2 个物理网络有直接的连接。主机通常直接与一个物理网络连接，或者说属于一个物理网络。但也有直接与多个物理网络相连的多穴主机（multi-homed host）。

主机和路由器都要参与到 IP 数据报传送过程。当一个主机上的应用程序试图进行通信时，TCP/IP 协议将产生若干数据报。无论是只有一个网络连接的主机还是多穴主机，都要做出最初的转发决策，即决定把数据报发往何处。

源主机首先根据目的主机的 IP 地址判断目的主机与本机是否在同一个物理网络上。对于最初的 IP 编址方案，可以根据分类编址规则，很容易地从目的 IP 地址中抽取出网络前缀，再与本机 IP 地址的网络前缀作比较。

如果匹配，则意味着数据报可以**直接交付**。可通过地址解析获取目的主机的物理地址，再将 IP 数据报封装在物理帧中直接发给目的主机。

如果不匹配，则应将数据报交给本地路由器的本地网络连接（在源主机路由表中指定）。这时要先通过地址解析获取路由器该网络连接的物理地址，再将数据报封装在帧中发给路

由器。这种交付称为**间接交付**。每个路由器将数据报间接交付给下一个路由器，直到数据报到达路径上最接近目的主机的路由器，由该路由器将数据报直接交付给目的主机。

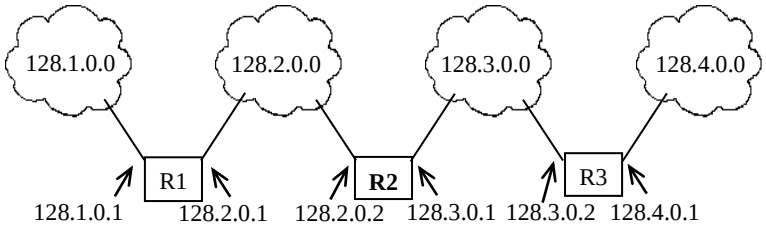
上述表明，TCP/IP 互联网中的路由器形成了一个相互协作的互连结构。对于源宿主机不在一个物理网络上的数据报，先被源主机传递到本地路由器，再经过若干次间接交付后抵达可进行直接交付的路由器，最后被直接交付。直接交付是任何数据报传输的最后一步。

3. 采用分类编址方案时的 IP 数据报转发算法

IP 转发是基于路由表驱动的。路由表存储有关怎样到达目的网络的信息。主机和路由器都有路由表。当主机或路由器中的 IP 转发软件需要传输数据报时，它就查询路由表来决定把数据报发往何处（下一路由器或目的主机）。

路由表一般存储目的网络地址以及如何到达该网络的信息，并不保存主机地址信息。这有助于大大缩减路由表大小，因为网络数量远小于主机数量。此外，还有利于提高路由表查询效率以及降低路由表维护开销。

一个互联网及路由表的示例见图 5-15。示例的互联网有 4 个 B 类网络，用 3 个路由器将它们互连起来。图 5-15(b)是路由器 R2 的路由表。路由表一般包含多对（N，R），N 是目的网络 IP 地址，R 是通往网络 N 的路径上的下一跳（next hop）的 IP 地址，此外，实际的路由表中还会指明数据报被发送到下一个路由器时所用的网络接口。当 R2 接收到一个目的网络地址为 128.2.0.0 的数据报，根据路由表，R2 将直接交付该数据报。当 R2 接收到一个目的网络地址为 128.4.0.0 的数据报，逐条查询路由表项，路由选择结果是下一跳地址为 128.3.0.2。注意路由表中的下一跳总是与本路由器的某网络连接属于同一个物理网络。



(a)

目的网络	下一跳
128.2.0.0	直连
128.3.0.0	直连
128.1.0.0	128.2.0.1
128.4.0.0	128.3.0.2

(b)

图 5-15 (a) 一个互联网示例；(b)路由器 R2 的路由表

图 5-15 例示的是一个小型互联网，如果互联网包含的物理网络很多，让路由表包含所有网络将使路由表表项数很多，不利于查找。有一种非常常用的用来隐藏信息和保持路由表容量较小的技术是把多个表项合并成一个表项，即**默认路由**。例如对于只有一个网络连接的主机，除了和直连（同一物理网络内）的主机通信，其余情况都应通过惟一的路由

器通向互联网的其余部分，因此主机路由表中一般只需 2 个表项即可。对于一个网点（例如一个包含多个物理网络的单位互联网）内的路由器，路由表中可以包含网点内的各网络的网络 IP 地址，最后加一个到所有其他目的网络的默认路由。

尽管 IP 转发是基于网络而不是基于个别主机的，但是多数 IP 转发软件允许为某个特定的目的主机特别指定路由。这主要用于测试，还可以出于安全的考虑。在调试网络连接或路由表时，尤其可能需要为单个主机指定一条特殊路由（**特定主机路由**）。

考虑上述所有情况，采用分类编址方案时 IP 数据报转发算法如图 5-16 所示。

```
采用最初的分类编址方案时的 IP 数据报转发（数据报 DG, 路由表 T）
从数据报 DG 中取出目的站 IP 地址 ID;
if 表 T 中含有 ID 的一个特定路由, 则
    把 DG 发送到该表项指明的下一跳
    (包括完成下一跳 IP 地址到物理地址的映射, 将 DG 封装入帧并发送);
    return.
根据分类地址规则, 从 ID 中提取出网络前缀, 得到网络地址 N;
if N 与任何一个直接相连的网络地址匹配, 则
    通过该网络把 DG 直接交付给目的站
    (包括解析 ID 得到对应的物理地址, 将 DG 封装入帧并发送);
else if 表 T 中包含一个到网络 N 的路由, 则
    把 DG 发送到该表项指明的下一跳
    (包括完成下一跳 IP 地址到物理地址的映射, 将 DG 封装入帧并发送);
else if 表 T 中包含一个默认路由, 则
    把 DG 发送到该表项指明的下一跳（默认路由器）;
else
    向 DG 的源站发送一个目的不可达差错报告;
```

图 5-16 采用分类编址方案时的 IP 数据报转发算法

4. 对传入数据报的处理

前面讨论了 IP 数据报的传送过程，并详细介绍了如何基于路由表进行 IP 数据报的转发。下面讨论 IP 软件对传入数据报的处理。分为两种情况，一种是主机收到数据报，另一种是路由器收到。

当一个数据报到达主机时，网络接口软件就把它交给 IP 模块进行处理：

```
从数据报 DG 中取出目的站 IP 地址 ID;
if ID 与主机的 IP 地址(单播或广播地址)匹配, 则
    接受 DG, 根据 DG 中的协议指示将 DG 的数据交给高层协议软件进一步处理;
else
    丢弃 DG;
```

注意，不作为路由器使用的主机应避免完成路由器的功能，所以当收到不是发给自己

的数据报时，选择丢弃而不是转发。

当一个数据报到达路由器某网络连接上的输入端口时，网络接口软件把它交给 IP 模块进行处理：

```
从数据报 DG 中取出目的站 IP 地址 ID;
if (ID 与路由器的任一个物理网络连接的 IP 地址匹配) ||
   (ID 是受限 IP 广播地址, 或目标是路由器的某直连网络的定向 IP 广播地址), 则
      接受 DG, 根据 DG 中的协议指示将 DG 的数据交给相应协议软件进一步处理;
      对于定向广播, 在指定的网络上广播该数据报;
else
      把 DG 首部中的生存时间 TTL 减 1;
      if TTL 为 0, 则
            丢弃 DG, 向 DG 的源站发送一个超时差错报告;
      else
            重新计算校验和, 并转发该数据报
```

在网际协议的控制下，通过主机和路由器的 IP 实体间以及相邻路由器的 IP 实体间的通信，网际互连层能够向上一层提供无连接的数据报传送服务。

5. 支持子网编址的 IP 转发算法

在一个使用子网编址的网络上，必须适当修改主机和路由器上使用的标准 IP 转发算法。

在标准 IP 转发算法中，特定主机路由和默认路由属于特例，必须专门检查，对其他路由则按常规方式进行表查询，路由表中普通路由的表项形式如下：

（目的网络地址，下一跳地址）

其中，下一跳地址字段指明了一个路由器的地址。

不划分子网时，根据分类地址规定可以很容易地从待转发数据报的目的 IP 地址中提取出网络地址。使用子网编址时，仅从目的 IP 地址无法判断出其中哪些比特对应网络部分（含子网部分），哪些比特对应主机部分。因此子网转发算法要求在路由表的每个表项中增加一个字段，指明该表项中的网络（子网）所使用的子网掩码：

（子网掩码，目的网络地址，下一跳地址）

在查找路由时，修改过的算法使用表项中地址掩码与目的 IP 地址按比特位进行布尔与运算，再把结果与表项中的目的网络地址相比较，若相等，表明匹配，应把数据报转发到该表项的下一跳地址。

通过巧妙设置“子网掩码”，路由查找时不必区分特定主机路由、普通网络路由和默认路由。例如为 202.119.220.10 指定一条特定路由，在路由表中可以表达成：（255.255.255.255，202.119.220.10，下一跳地址）；在路由表中可以这样表达默认路由：（0.0.0.0，0.0.0.0，下一跳地址），与其他所有路由都不匹配时再选择默认路由。对与路由器直接相连的网络可以分别添加一个表项，不过下一跳地址字段不应是具体的地址，而应标明按直接交付方式转发。对于到达没有划分子网的分类网络的路由，可以使用默认掩码，例如对于到达 C 类网

络 202.119.230.0 的路由，在路由表中可表示为（255.255.255.0，202.119.230.0，下一跳地址）。如果给路由表排序，应将最长掩码（掩码中的 1 比特最多）的表项排在最前面。

支持子网编址的统一的 IP 转发算法如图 5-17 所示。

```
采用子网编址方案时的 IP 数据报转发（数据报 DG，路由表 T）
从数据报 DG 中取出目的 IP 地址 ID;
for 表 T 中的每一表项 do
    将 ID 与表项中的子网掩码按位相“与”，结果为 N;
    if N 等于该表项中的目的网络地址，则
        if 下一跳指明应直接交付，则
            把 DG 直接交付给目的站
            (包括解析 ID 得到对应的物理地址，将 DG 封装入帧并发送);
        else
            把 DG 发往本表项指明的下一跳地址
            (包括完成下一跳地址到物理地址的映射，将 DG 封装入帧并发送);
        return.
for_end
因没有找到匹配的表项，向 DG 的源站发送一个 ICMP 目的不可达差错报告.
```

图 5-17 支持子网编址的统一的 IP 转发算法

6. 使用 CIDR 时的路由查找算法

使用 CIDR，路由表的每个表项应由“网络前缀/掩码”和“下一跳”组成。观察 ISP 给用户分配地址块，会发现用户地址块的网络前缀长度总是比 ISP 的长，网络前缀越长，其地址块就越小。路由表中可能会混合含有到 ISP 的路由、到 ISP 的某用户单位网络的路由，及到单位内某物理网络甚至到某主机的路由。显然，查找路由时，目的地越具体的路由越值得采纳。因此路由表查找的目标是**最长前缀匹配**（longest prefix match）。也就是，查找路由表时，即使找到了匹配表项，查找还不能结束，必须查找完所有的表项，在所有的匹配表项中再选择具有最长前缀的路由。

为了提高查找下一跳的速度，在分类编址情况下，IP 查找使用散列方法，路由表项的存放地址取决于以网络前缀作为关键字的散列函数值。分类地址是自标识的，容易提取出网络前缀。采用无分类编址时，散列就不能很好发挥作用了。

为了避免低效率的搜索，无分类查找使用分层的数据结构。使用最广泛的是一种二叉线索（binary trie）的变形。方法是将路由表中的各个路由信息存放在一棵二叉线索树中。具体地说，就是将各表项中的网络前缀写成比特串（取前缀长度个比特），表项中网络前缀的比特串决定从根结点逐层向下的路径，可以令 0 比特对应左分支，1 比特对应右分支，在每个地址路径的终止节点中应包含相应表项信息（网络前缀/掩码以及下一跳地址）。如果包含特定主机路由，理论上二叉线索树应为 33 层（含根层）。

下面通过例子说明使用二叉线索存储结构实现无分类路由查找的基本原理。例如路由表中有图 5-18 所示的一组路由，构建的二叉线索树如图 5-19 所示。由于各路由开头有共同的“128.10”，因此可对线索树做适当优化，使根节点之下的连续 16 层的单分支合并为一个分支。同样对特定主机地址的第 4 个字节所对应的线索也进行了压缩。图 5-19 中加粗的节点表示路由表中某个网络前缀路径的终止。

给定一个目的 IP 地址 128.10.4.3 （I_D），从线索树的根节点开始，首先将 I_D 的前 16 比特与分支上的“128.10”比较，相等则转到下一节点，该节点中存放路由信息，表示找到一个匹配项。但仍需继续往下查找，经过“00000”、“1”、“0”、“0”等分支，到达一个包含路由信息的节点，表示又找到一个匹配路由，覆盖较早发现的匹配，因为较晚的匹配对应一个更长的前缀。继续与“00000011”比较，相等，转到下一节点，该节点中包含路由，这表示又匹配了，再覆盖先前发现的匹配，另外由于该节点是叶子节点，所以查找结束。最长前缀的匹配所对应的下一跳地址是最终路由查找结果。如果 I_D 是 128.10.4.5，也将查找到叶子节点，但最长前缀的匹配项存储在叶子节点的上一个节点中。

网络前缀/前缀长度	下一跳
128.10.0.0/16	10.0.0.2
128.10.2.0/24	10.0.0.4
128.10.3.0/24	10.1.0.5
128.10.4.0/24	10.0.0.6
128.10.4.3/32	10.0.0.3
128.10.5.0/24	10.0.0.6
128.10.5.1/32	10.0.0.3

图 5-18 含有同一网络的一般路由和特殊路由的路由表示例

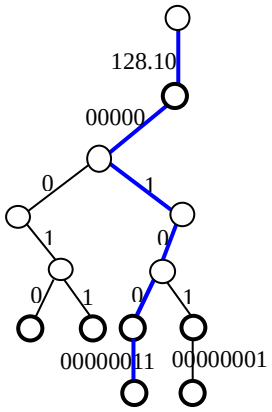


图 5-19 图 5-18 中路由表所构成的二叉线索树

7. 专用 IP 地址

IP 地址资源是有限的，为了节约地址的使用，IANA 保留了三个只能用于专用互联网（private internet）内部通信的 IP 地址块[RFC1918]，见表 5-6。任何机构可以使用 TCP/IP 技术并且使用保留的专用地址构建专用互联网。

表 5-6 保留用于专用互联网的 CIDR 地址块

前缀	最低地址	最高地址
10/8	10.0.0.0	10.255.255.255
172.16/12	172.16.0.0	172.31.255.255
192.168/16	192.168.0.0	192.168.255.255

完全隔离的专用互联网通常也不是人们所希望的，可以使用 NAT 技术（后面介绍）将专用互联网联入因特网。

5.3 因特网的路由选择协议

本节讨论因特网中路由器的路由表是怎样初始化和动态更新的，将讨论几种常用的路由选择协议。

5.3.1 自治系统与路由选择协议分类

路由选择协议的核心是路由选择算法，也即路由计算与更新算法。一个理想的路由选择算法应具有如下一些特点：

（1）算法必须是正确的。所谓正确是指：沿着各路由表所指引的路由，分组一定能够最终到达目的网络和目的主机。

（2）算法在计算上应简单。更新路由表的计算要占用路由器的处理器资源，为计算路由，需要路由器之间交换信息，这还将占用网络带宽。因此，路由选择算法应简单，以免对路由器的关键任务——数据报的转发产生大的影响。

（3）算法应能适应通信量和网络拓扑的变化。这就是说，要有自适应性。当网络中的通信量发生变化时，算法应能自适应地改变路由以均衡各链路的负载。当某些路由器、链路发生故障不能工作时，或者设备或链路修复再投入运行时，算法应能及时地改变路由。

（4）算法应具有稳定性。当网络通信量和网络拓扑相对稳定时，路由选择算法计算得出的路由应比较稳定，不应不停地变化。

（5）算法应是公平的。算法应平等对待具有相同优先级的用户。

从能否随网络通信量或拓扑的变化进行自适应地调整来看，路由选择算法可以划分为两大类，即静态路由选择策略与动态路由选择策略。静态路由选择也叫做非自适应路由选择，其特点是简单和开销较小，但不能及时适应网络状态的变化。动态路由选择也叫做自适应路由选择，其特点是能较好地适应网络状态的变化，但实现起来较为复杂，开销也较大。

一个实际的路由选择算法应尽可能地接近于理想的算法。在不同的应用条件下，可以对以上几个方面有不同的侧重。实际上，因特网路由选择是个非常复杂的问题，因为它需要因特网中路由器共同协调工作。其次，路由选择的环境往往是不断变化的，而且这种变化有时是无法事先知道的。

因特网采用的路由选择协议主要是自适应的、分布式的路由选择协议。因特网采用分层次的路由选择协议，原因有：

（1）因特网是全球范围的互联网，规模很大，已有几百万个路由器将很多物理网络互连在一起。如果让所有路由器知道到所有物理网络应怎样到达，则路由表将非常大，查询和更新起来都很费时，而且所有路由器之间交换路由信息的通信量就会使因特网的通信链路饱和。

（2）许多单位不愿意外界了解自己单位互联网的拓扑细节，以及本单位采用的路由选

择协议，但同时还希望连到因特网上。

整个因特网被划分为许多自治系统（autonomous system，简称 AS）。传统定义的 AS 是在单一技术管理下的一组路由器，使用一个内部网关协议（Interior Gateway Protocol，IGP）和共同的测度确定如何在 AS 内路由分组，并使用一个 AS 间路由选择协议决定如何将分组发送到其他自治系统。不过，AS 的定义有了发展，现在单个 AS 可以使用多个内部网关协议，有时还使用几组测度。现在使用自治系统术语强调的是即使使用了多个内部网关协议和几组测度，一个 AS 在其他自治系统看来应具有单个一致的内部路由选择策略，对可通过该 AS 到达的目的网络具有一致的描绘。

一个 AS 是一组连通的有单一、明确定义的路由策略的 IP 网络，包含由一个或多个网络提供商管理的一个或多个 IP 前缀（CIDR 地址块）。AS 是一个互联网，或更确切地说，是连接网络的路由器的集合，这些路由器共享相同的路由策略。自治系统的路由策略表述的是网络前缀如何在自治系统之间进行交换。

因特网可看作是随意连接的 AS 的集合，一个 AS 与一个或多个其他的 AS 连接。每个 AS 都有一个自治系统号(AS 号)，一个与自治系统相关联的 16 比特整数，作为与其他自治系统交换动态路由信息的标识符。每个与外界连接的 AS 必须指定本 AS 内的一台或几台路由器，使用某外部网关协议（Exterior Gateway Protocol，EGP）向其他 AS 通告网络可达性。当前因特网中使用的外部网关协议是版本号为 4 的边界网关协议（Border Gateway Protocol version 4, BGP-4）。AS 之间使用 BGP-4 交换路由信息。

AS 号空间与 IP 地址空间一样是有限的，因此现在并不建议把 AS 号作为管理的一种形式，而是把 AS 号作为路由策略的表示，仅当存在不同于边界路由器对等端所用的路由策略时才需要。由 IANA 下属的 APNIC 等地址注册商负责管理 AS 号的统一分配，这有助于限制全球路由表的扩展。因为一个自治系统将汇集本 AS 内相邻的 IP 地址前缀，并与其他 AS 交换信息，AS 划分过细不利于将公布于全球互联网的路由数量减至最低。

对于一个单接入网点（Single-homed site），一般都不需要作为单独的 AS，因为网点的 1 个或多个前缀（1 个前缀表示一个 CIDR 地址块）通常都是由网点的 ISP 分配的，并且网点的前缀通常与网点服务提供者的其他客户有相同的路由策略。

一个网点如果满足下列条件，可以分配 AS 号：（1）是多宿主的（Multi-homed site，也称为多接入网点）；（2）有单一的、明确定义的路由策略，并且不同于提供商的路由策略。

一个 AS 有权自主地决定在本系统中采用何种内部路由更新机制。一个 AS 内的路由器可以使用一个或多个内部网关协议与本 AS 内其他路由器交换路由信息。在互联网中常用的 IGP 有：RIP、OSPF 和 IGRP。IGRP（Interior Gateway Routing Protocol）是 Cisco 公司 20 世纪 80 年代开发的，是一种动态的最大可支持 255 跳的路由选择协议，使用一组测度确定到达一个网络的最佳路由，测度包括网络延迟、带宽、可靠性和负载等，Cisco IOS(Internetwork Operating System)允许路由器管理员为 IGRP 的每种测度设置权重。IGRP 是一种距离向量型的内部网关协议，协议要求每个路由器以规则的时间间隔向其相邻的路由器发送其路由表的全部或部分。随着路由信息在网络上扩散，路由器就可以计算到所有目的网络或目的站的距离。

总之，因特网路由选择协议可划分为两大类：

（1）内部网关协议（IGP） 把一个自治系统内部路由器交换路由信息所用的任何协议统称为内部网关协议。每个自治系统可自主选择具体的 IGP 协议。目前因特网中常用的 IGP 有 RIP、OSPF 和 IGRP。

（2）外部网关协议（EGP） 两个自治系统之间传递网络可达性信息所用的协议称为外部网关协议。每个自治系统内都指定一个或多个路由器除了运行本系统的 IGP 外，还运行 EGP 与其他的自治系统交换信息。目前因特网中惟一在用的 EGP 协议是 BGP-4。BGP 称运行 BGP 的路由器为边界网关（border gateway）或边界路由器（border router）。在图 5-20 中，路由器 R1 收集自治系统 AS1 中的网络有关信息，并使用 EGP 把信息报告给 AS2 中的路由器 R2。同样，R2 把 AS2 的网络可达性信息报告给 R1。

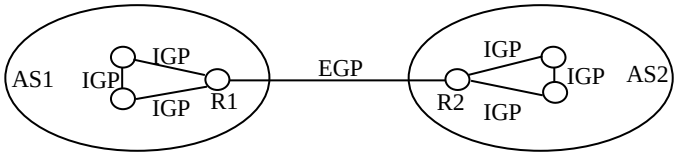


图 5-20 自治系统、内部网关协议和外部网关协议

下面介绍因特网中最常用的两种 IGP 协议 OSPF 和 RIP，以及惟一在用的 EGP 协议 BGP-4。

5.3.2 内部网关协议 RIP

路由信息协议（Routing Information Protocol，RIP）是内部网关协议中最先得到广泛使用的协议，RIP 使用一种距离向量算法更新路由表，常用于小型的自治系统。

距离向量算法要求每个路由器，在路由表中列出到所有已知目的网络的最佳路由，并且定期把自己的路由表副本发送给与其直接相连的其他路由器。为了确定最佳路由，使用测度度量路由优劣。可以使用表示数据报到目的网络必须经过的路由器的个数，即跳数作为测度，也可以使用表示数据报经历的时延、发送数据报的开销等作为测度。RIP 使用跳数作为测度，这样，所谓最佳路由就是能够以最少跳数到达某目的网络的路由。

路由器启动时对路由表进行初始化，为与自己直接相连的每个网络生成一个表项。表项包括一个目的网络，到该网络的最短距离（最少跳数），及路由（下一跳）。每个路由器根据相邻路由器定期发来的路由信息更新自己的路由表，获悉更多目的网络以及到各网络的最佳路由。下面通过例子说明。设路由器 K 与两个网络相连，K 初始的距离向量路由表如图 5-21 所示。

目的网络	距离	路由
网络 1	0	直接
网络 2	0	直接

图 5-21 一个初始的距离向量路由表示例

每个路由器只和相邻路由器（数量非常有限）交换路由信息并更新路由表，一个自治系统中的所有路由器经过若干次路由通告与更新后，最终都会知道到达本 AS 中任何一个

网络的最短距离和路径中下一跳路由器的地址。

【例 5-8】假如经过数次路由更新后路由器 K 的路由表如图 5-22(a)所示。当相邻路由器 J 的路由信息报文（如图 5-22(b)所示）到达路由器 K 后，请问 K 的路由表将如何更新。

解： K 检查报文中的（目的网络，到该网络的距离）列表（J 的路由表副本）。如果 J 知道去某目的网络更短的路由，或者 J 列出了 K 中不曾有的目的网络，或者 K 目前到某目的网络的路由经过 J，而 J 到达该网络的距离有所改变，则 K 就会替换自己的路由表中的相应表项。更新后的 K 的路由表如图 5-22(c)所示。

目的网络	距离	路由
网络 1	0	直接
网络 2	0	直接
网络 4	4	路由器 L
网络 17	7	路由器 M
网络 24	6	路由器 J
网络 30	2	路由器 Q
网络 42	2	路由器 J

(a)路由器 K 的路由表

目的网络	距离
网络 1	2
网络 4	3
网络 17	**5**
网络 21	**6**
网络 24	**4**
网络 30	9
网络 42	**3**

(b)来自路由器 J 的路由信息

目的网络	距离	路由
网络 1	0	直接
网络 2	0	直接
网络 4	4	路由器 L
网络 17	**6**	**路由器 J**
网络 21	**7**	**路由器 J**
网络 24	**5**	**路由器 J**
网络 30	2	路由器 Q
网络 42	**4**	**路由器 J**

(c)更新后的 K 的路由表

图 5-22 基于距离向量算法的路由更新示例

图 5-22(b)中加粗的表项将引起 K 的路由表的更新。原本从 K 经路由器 M 至目的网络 17 的距离为 7，但邻居 J 声称它到网络 17 的距离为 5，这个路由更短，因此 K 路由表中至网络 17 的距离更新为 5+1（从 J 到目的网络的距离加上 K 到 J 的距离），路径上的下一跳指定为 J。J 声称从它能够到达网络 21，K 路由表中无此目的网络，因此新增一个到网络 21 的表项。从 K 到网络 24 的路由原本经过 J，距离为 6，但 J 声称从它到网络 24 的距离（由 5 变）为 4 了，因此更新距离为 4+1。同理，K 的路由表中至网络 42 的路由也要做类似更新。

注意，如果 J 报告到某目的网络的距离是 N，并且 K 根据该信息需要添加或更新自己路由表中的某个表项时，则该表项的距离为 N+1，下一跳指定为路由器 J。

虽然距离向量算法易于实现，但它们也有缺点。当路由迅速发生变化（例如链路出现故障）时，相应的信息缓慢地从一个路由器传到另一个路由器，算法可能无法稳定下来，出现路由表的不一致问题和慢收敛问题。

RIP 和下一小节要介绍的 OSPF 都是分布式路由选择协议。它们共同的特点是每一个路由器都要不断地和其他一些路由器交换路由信息。RIP 路由信息交换与更新有以下 3 个特点：

（1）RIP 路由器仅和本自治系统内与自己相邻的路由器交换信息。RIP 规定，信息仅在相邻的路由器之间交换，所谓相邻指在一个网络上。此外主机可以参与接收 RIP 广播并更新自己的路由表，但主机不发送路由更新报文。

（2）RIP 支持 2 种信息交换方式。一种是**定期的路由更新**，即路由器按固定的时间间隔，例如每 30 秒向所有邻居发送一个更新报文，其中包含路由器当前所知道的全部路由信息，即自己的路由表。另一种是**触发的路由更新**，无论何时只要路由表中有路由发生改变，路由器就可立即向与其直连的主机和路由器发送触发更新报文。

（3）路由表更新的原则是按照距离向量算法，确定并记录到各目的网络的最短距离（以跳数计）和路径上的下一跳。

RIP 规定距离 16 表示无路由或不可达，还规定路由超时时间为 180 秒。例如假设某路由器 X 到网络 n 的当前路由以路由器 G 为下一跳，如果 X 有 180 秒都没有收到来自 G 的路由更新信息，则可以认为 G 崩溃了或 X 连到 G 的网络不可用了，此时 X 可以标记至网络 n 的距离为 16。

RIP 存在 2 个版本，版本 1（RIP-1）出现于上世纪 80 年代[RFC 1058]，较新的版本 2（RIP-2）发布于 90 年代[RFC 1388，RFC 2453]。RIP-1 中交换的路由信息仅包含一组（网络地址，到网络的距离），而 RIP-2 的更新报文中还增加了下一跳信息，这有助于解决慢收敛问题和防止出现路由环路。RIP-2 的更新报文中还增加了子网掩码信息，以支持变长子网地址或无分类地址。总之，RIP-2 更新报文包含 4 元组（网络地址，网络掩码，到网络的下一跳，到网络的距离）列表，格式见图 5-23。

<table>
<tr><td>命令</td><td>版本</td><td>为 0</td></tr>
<tr><td colspan="2">网络 1 的协议族</td><td>网络 1 的路由标记</td></tr>
<tr><td colspan="3">网络 1 的 IP 地址</td></tr>
<tr><td colspan="3">网络 1 的子网掩码</td></tr>
<tr><td colspan="3">到网络 1 的下一跳</td></tr>
<tr><td colspan="3">到网络 1 的距离</td></tr>
<tr><td colspan="2">网络 2 的协议族</td><td>网络 2 的路由标记</td></tr>
<tr><td colspan="3">网络 2 的 IP 地址</td></tr>
<tr><td colspan="3">网络 2 的子网掩码</td></tr>
<tr><td colspan="3">到网络 2 的下一跳</td></tr>
<tr><td colspan="3">到网络 2 的距离</td></tr>
<tr><td colspan="3">……</td></tr>
</table>

图 5-23 RIP-2 报文格式

命令字段指明一种操作，例如为 1 表示请求，请求响应系统发送路由表所有或部分信息，为 2 表示响应，一个响应报文包含发送者路由表的全部或部分信息，该报文可以是为响应一个请求而发送，也可能是由发送者产生的一个更新报文。

路由标记（Route Tag）字段用于支持 EGP，用于传播路由来源之类的额外信息。例如，如果 RIP-2 路由器从另一个自治系统得知一个路由，可以使用路由标记字段携带那个自治系统的编号。

此外，为了防止不必要地增加不监听 RIP-2 分组的主机的负担，RIP-2 的周期广播使

用一个固定的组播地址 224.0.0.9。使用固定的组播地址意味着不需要依赖 IGMP（因特网组管理协议）。RIP-2 比 RIP-1 还增加了认证机制。

RIP 基于 UDP，使用 UDP 端口 520（有关端口的意义请参阅下一章）。虽然可以在其他 UDP 端口发起 RIP 请求，但请求报文的 UDP 目的端口总是 520，并且 RIP 广播报文的源端口也是 520。

RIP 作为内部网关协议，存在一些限制。第一，用一个小的跳数值表示无穷大，限制了使用 RIP 的互联网规模。使用 RIP 的互联网中，任意 2 台主机之间最多有 15 跳。第二，路由器周期地向邻居广播完整的路由表，随着网络规模的增大，开销会增大，路由更新的收敛时间也会延长。第三，RIP 只使用跳数测度，不支持负载均衡，路由选择相对固定不变。

5.3.3 内部网关协议 0SPF

1. 协议概述

OSPF 是 IETF 的一个工作组设计的一个内部网关协议，它使用链路状态（Link State）算法，或称最短路径优先（Shortest Path First, SPF）算法。OSPF 也即开放的 SPF 协议，所谓开放是指协议规范可在公开发表的文献中找到。

在链路状态路由选择协议中，每个路由器维护一个描述自治系统拓扑的数据库。该数据库称为链路状态数据库。每个参与的路由器有相同的数据库。数据库的每一项是单个路由器的本地状态（例如，路由器接口所连网络、与接口输出端关联的代价（cost，也称开销）、不可用的接口及可达的邻居等）。路由器利用洪泛法（flooding）向整个自治系统发布自己的本地状态。

所有路由器并行地运行着相同的算法。每个路由器根据链路状态数据库，使用 Dijkstra 最短路径算法，构建一个以自己为根的最短路径树。最短路径树给出了到自治系统中每个网络的路由。从自治系统外部得到的路由信息在树中作为叶子出现。

如果到一个目的站存在若干条代价相同的路由，则把流量均匀地分配给这些路由。路由的代价用单个无量纲测度描述。因此 OSPF 能提供负载均衡（load balancing）功能。而 RIP 对每个目的站只计算一条路由。

OSPF 允许将 AS 中的网络分成若干组，每个组称为一个**区域**（area）。一个区域的拓扑相对于 AS 的其他部分来说是隐藏的。信息隐藏能够使路由信息流量显著减少。此外，在区域内的路由选择仅取决于区域自己的拓扑，从而保护区域不受外界坏路由数据的影响。区域是子网化 IP 网络的推广。

OSPF 允许灵活配置 IP 子网，支持特定主机的路由、特定子网的路由、无分类路由和特定分类网络的路由。OSPF 分发的每个路由都含有目的地和掩码。相同 IP 网络号的两个不同子网可能具有不同的掩码，即变长子网划分。主机路由被当作是掩码为全 1 的子网。IP 分组被转发到最佳匹配所指定的下一跳。

所有 OSPF 协议交换都要被鉴别。保证只有可信的路由器可以参与自治系统的路由选择。OSPF 支持各种鉴别机制，而且允许每个区域配置不同的鉴别机制。

从外部得到的路由选择数据（例如从一个外部网关协议如 BGP 获得的路由）要在整个

自治系统中通告。这些数据将与 OSPF 协议的链路状态数据分开存放。

2. OSPF 区域和路由器类别

为使每个区域能够和同一 AS 内的其他区域进行通信，每个区域都设有**边界路由器**，所有区域边界路由器都属于特别的区域 0，也称为 OSPF 骨干（backbone）。骨干负责在非骨干区域之间分发路由信息。骨干必须是连续的，但物理上不必是连续的，骨干连通性可以通过配置虚拟链路（virtual link）建立与维持。在任意两个有接口连接到普通（非骨干）区域的骨干路由器之间，可以配置虚拟链路，虚拟链路属于骨干。协议把由虚拟链路连接的两个路由器，当作就像是由一个无编号点到点骨干网络连接的一样处理。

图 5-24 给出了一个 OSPF 区域划分示例，该图实质上是个有向图。其中，每个路由器接口的输出端都有一个代价，如果没标定则表示代价为 0，注意从网络到路由器的代价总为 0。代价可由系统管理员配置。代价越小，接口越有可能用于转发流量。从外部获得的路由数据（例如 BGP-获得的路由）也有代价与之关联。

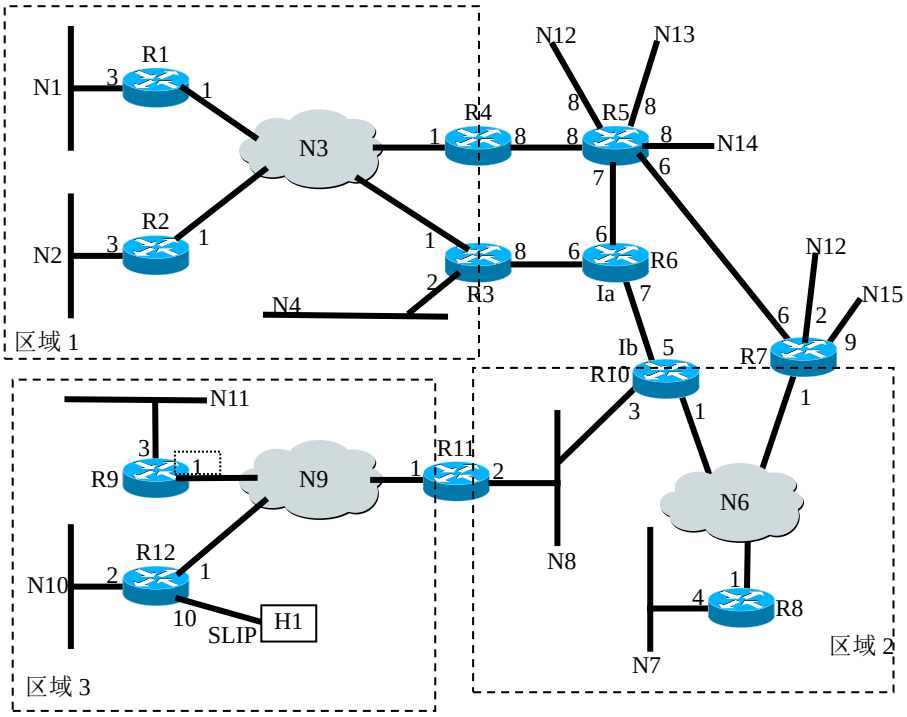


图 5-24 一个 AS 的 OSPF 区域配置示例

如果不引入区域，惟一具有专门功能的是那些通告外部路由信息的 OSPF 路由器。如果划分区域，AS 中的路由器按照功能可进一步分为 4 个有重叠的类别：

（1）内部路由器（Internal router） 一个所有直连网络属于同一区域的路由器。这些路由器运行单个一份基本路由选择算法。

（2）区域边界路由器（Area border router） 连接到多个区域的路由器。这些路由器运行多份基本路由选择算法，每份用于一个连接的区域。区域边界路由器浓缩它们所附属的区域的拓扑信息并散布到骨干，骨干反过来再将信息分发到其他区域。

（3）骨干路由器（Backbone router） 有接口到骨干区域的路由器。包括所有连接到

不止 1 个区域的路由器。不过，骨干路由器不一定是区域边界路由器，可以是内部路由器，其所有接口都连接到骨干区域。

（4）AS 边界路由器（AS boundary router） 与属于其他 AS 的路由器交换路由信息的路由器。这样的路由器将 AS 外部路由信息传遍本自治系统。AS 中的每个路由器都知道到每一 AS 边界路由器的路径。AS 边界路由器可能是内部或区域边界路由器，并且也许加入也许没加入骨干。

图 5-24 中，路由器 R1, R2, R5, R6, R8, R9 和 R12 是内部路由器；R3, R4, R7, R10 和 R11 是区域边界路由器；R5 和 R7 是 AS 边界路由器。

自治系统中网络类型可分为 3 种：（1）点到点网络（Point-to-point networks），指连接一对路由器的网络；（2）广播网络（Broadcast networks），指支持连接多个（超过 2 个）路由器，并且具有广播能力的网络。可使用 OSPF 的 Hello 协议动态发现网络上的相邻路由器，广播网上的每一对路由器之间都假定能直接通信，例如以太网；（3）非广播网络（Non-broadcast networks），指支持连接多个路由器，但没有广播能力的网络，例如 X.25 公用数据网。对这种网络可能必需作适当配置以帮助发现邻居，使用 Hello 协议维持邻居关系。每个通常被组播的 OSPF 协议分组需要被依次发送到每个邻近路由器(neighboring router)。非广播网络在 OSPF 中分为 2 种模式：NBMA（non-broadcast multi-access），模拟广播网上的 OSPF 操作，另一种是点到多点（Point-to-MultiPoint）网络，把网络视为点到点链路的集合。

图 5-24 中，惟一的点到点网络连接 R6 和 R10，并已被指定了接口地址 Ia 和 Ib。点到点网络的接口可以不指定 IP 地址。当指定了接口地址时，接口就作为末梢链路，每个路由器向另一个路由器的接口地址通告一个末梢连接。网络 N6 是一个连接了 3 个路由器的广播网络。

3. OSPF 基本路由选择算法

在每个区域运行单独的一份 OSPF 基本路由选择算法。有若干接口连到多个区域的路由器运行多份算法。算法简单概括如下：

- 当路由器启动时，首先初始化路由选择协议数据结构。然后等待低级协议指示其接口已经起作用了。
- 然后路由器使用 OSPF 的 Hello 协议获得邻居。路由器发送 Hello 分组给它的邻居，接着收到邻居返回的 Hello 分组。在广播和点到点网络上，路由器通过发送 Hello 分组到组播地址 AllSPFRouters（224.0.0.5）动态地探测它的邻居。在非广播网络上，为发现邻居可能必需一些配置信息。在广播和 NBMA 网络上，Hello 协议还为网络选举一个指定路由器（Designated Router）。例如图 5-24 中网络 N6 就会选举出一个指定路由器，由它产生网络 N6 的 LSA（Link State Advertisement，链路状态通告）。
- 路由器将尝试与其新获得的邻居中的一些形成**邻接**关系（Adjacency）。一对邻接的路由器之间的链路状态数据库是同步的。在广播和 NBMA 网络上，指定路由器决定哪些路由器应成为邻接的。邻接关系控制路由信息的分发，仅在邻接路由器上收发路由更新。

- 路由器周期地通告它的状态，也称为链路状态。路由器也在其状态改变时通告链路状态。路由器的邻接关系反映在它的 LSA 的内容中。邻接和链路状态之间的关系使得协议能够及时地察觉不工作的路由器。
- LSA 在整个**区域**内洪泛发送。OSPF 使用可靠的洪泛算法，确保一个区域中的所有路由器有完全相同的链路状态数据库。该数据库由属于该区域的各个路由器发起的 LSA 的集合组成。从这个数据库，每个路由器计算一个以自己为根的最短路径树（shortest-path tree）。再由最短路径树产生一个 OSPF 路由表。

上述描写的是单个区域内协议的运转。对于**区域内部路由选择**（intra-area routing），不需要其他路由信息。

对于**区域间路由选择**（inter-area routing），则需要另外的路由信息。为了能够路由到区域外的目的地，区域边界路由器要给区域注入附加的路由信息。附加的路由信息是对除本区域外的自治系统拓扑其余部分的提炼，提炼方法是：每个区域边界路由器（根据定义是连至骨干的）汇总其所连接的非骨干区域的拓扑，在骨干上传输到达所有其他区域边界路由器。结果，一个区域边界路由器就有了关于骨干及来自各个其他区域边界路由器的区域摘要的完整拓扑信息。根据这个信息，该路由器计算到所有跨区域目的地（inter-area destinations）的路径。这使该区域的内部路由器在转发到跨区域目的地的流量时，能够选择最佳的出口路由器。

关于 AS 外部路由（AS external routes），有关于其他 AS 的信息的路由器可以将信息洪泛至整个 AS。外部路由选择信息（external routing information）一般被分发到每个参与的路由器，有个例外是：不洪泛到末梢区域（"stub" areas）。当一个区域仅有一个出口点，或者当出口点不基于每个外部目的地来选择时，区域可以配置为末梢区域。为利用外部路由选择信息，到所有通告外部信息的路由器的路径必须传遍 AS（末梢区域除外）。因此非末梢区域的边界路由器要汇总 AS 边界路由器的位置。

4. 最短路径树和路由表生成示例

假如图 5-24 中的自治系统没有划分区域，即 AS 只有 1 个区域，其中 R5 和 R7 是 AS 边界路由器，则路由器 R6 构建的最短路径树如图 5-25 所示。

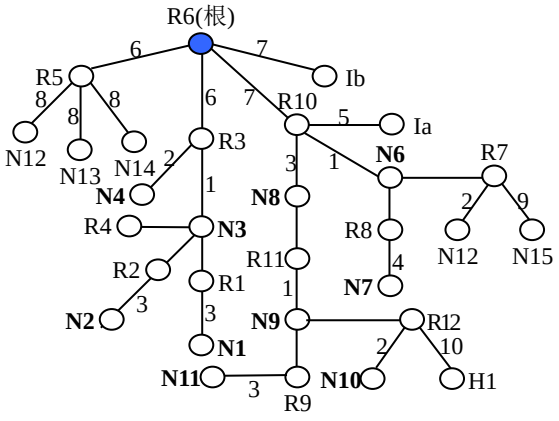


图 5-25 图 5-24 中的自治系统不分区域时路由器 R6 的 SPF 树

依据该树可以得到 R6 的路由表，见表 5-7。其中计算了至 R5 和 R7 的区域内部路由（intra-area routes），并进一步计算了到 R5 和 R7 所通告的目的网络 N12~N15 的外部路由（external routes）。

表 5-7 图 5-24 中的自治系统不分区域时 R6 的路由表

目的类型	目的	区域	路径类型	代价	下一跳	通告路由器
Network	N1	0	intra-area	10	R3	*
Network	N2	0	intra-area	10	R3	*
Network	N3	0	intra-area	7	R3	*
Network	N4	0	intra-area	8	R3	*
Network	Ib	0	intra-area	7	*	*
Network	Ia	0	intra-area	12	R10	*
Network	N6	0	intra-area	8	R10	*
Network	N7	0	intra-area	12	R10	*
Network	N8	0	intra-area	10	R10	*
Network	N9	0	intra-area	11	R10	*
Network	N10	0	intra-area	13	R10	*
Network	N11	0	intra-area	14	R10	*
Network	H1	0	intra-area	21	R10	*
Router	R5	0	intra-area	6	R5	*
Router	R7	0	intra-area	8	R10	*
Network	N12	*	external	10	R10	R7
Network	N13	*	external	14	R5	R5
Network	N14	*	external	14	R5	R5
Network	N15	*	external	17	R10	R7

下面考虑划分了区域的图 5-24 中的自治系统。区域 1 由网络 N1-N4 以及路由器 R1-R4 组成。区域 2 由网络 N6-N8 以及路由器 R7,R8,R10 和 R11 组成。区域 3 由网络 N9-N11 和主机 H1 以及路由器 R9, R11 和 R12 组成。区域 3 被配置成在向区域外部通告路由信息时，能够将网络 N9-N11 和主机 H1 组合成一条路由。R5 和 R7 是 AS 边界路由器，R3, R4, R7, R10 和 R11 是区域边界路由器。

路由器 R4 连至区域 1 和骨干（区域 0），R4 的路由表如表 5-8 所示。注意区域边界路由器 R3 有两个路由表项，因为它有和 R4 一样的两个区域。表中计算了到所有区域边界路由器的骨干路径，因为在决定区域间路由（inter-area routes）时要用到。表中所有的区域间路由都与骨干关联，计算路由的路由器本身是区域边界路由器时总是这样。路由选择信息在区域边界被压缩。例如，路由器 R11 向骨干通告区域 3 路由时，区域 3 中到网络 N9-N11 的路由和到 H1 的主机路由全被压缩成单条路由，并且这条路由的代价是到各个组成部分的代价集合中的最大值。

本例中有两条到网络 N12 的等代价路径，不过它们都使用相同的下一跳（路由器 R5）。另外，路由器 R10 和 R11 之间配置了虚拟链路，没有该虚拟链路，R11 就不能向骨干通告用于网络 N9-N11 和主机 H1 的路由了。如果在 R4 和 R3 之间也配置一条虚拟链路，R4 的路由表中的某些路径将变得短一些，这里从略。

表 5-8 划分区域的图 5-24 的 AS 中路由器 R4 的路由表

目的类型	目的	区域	路径类型	代价	下一跳	通告路由器
N	N1	1	intra-area	4	R1	*
N	N2	1	intra-area	4	R2	*
N	N3	1	intra-area	1	*	*
N	N4	1	intra-area	3	R3	*
R	R3	1	intra-area	1	*	*
N	Ib	0	intra-area	22	R5	*
N	Ia	0	intra-area	27	R5	*
R	R3	0	intra-area	21	R5	*
R	R5	0	intra-area	8	*	*
R	R7	0	intra-area	14	R5	*
R	R10	0	intra-area	22	R5	*
R	R11	0	intra-area	25	R5	*
N	N6	0	inter-area	15	R5	R7
N	N7	0	inter-area	19	R5	R7
N	N8	0	inter-area	18	R5	R7
N	N9-N11,H1	0	inter-area	36	R5	R11
N	N12	*	external	16	R5	R5,R7
N	N13	*	external	16	R5	R5
N	N14	*	external	16	R5	R5
N	N15	*	external	23	R5	R7

5. OSPF 分组和链路状态通告

为减少未参与系统的负载，OSPF 通过组播发送报文。为了消除对 IGMP 的依赖，协议预设了两个 IP 组播地址：224.0.0.5 用于所有路由器（AllSPFRouters），224.0.0.6 用于所有**指定路由器**（AllDRouters）。为避免将 OSPF 报文送出区域，要对路由器进行配置，防止它将发送给上述两地址的报文转发出去。OSPF 分组直接封装在 IP 数据报中发送，协议号为 89。

OSPF 共有 5 种分组类型：

（1） Hello 分组 在每个运行的路由器接口上发送。用于发现和维持路由器的邻居关系。在广播和 NBMA 网络上，Hello 分组还要用于选举**指定路由器**和候补指定路由器。

（2） 数据库描述（Database description）分组，汇总数据库内容。数据库描述和下面的链路状态请求分组用于形成**邻接关系**。数据库描述分组的发送取决于邻居的状态。

（3） 链路状态请求（Link State Request），用于下载数据库。

（4） 链路状态更新（Link State Update），用于数据库更新。每个链路状态更新分组携带一组新的链路状态通告（LSAs）。单个链路状态更新分组可能包含不同路由器的 LSAs。每个 LSA 用发起路由器的 ID 和链路状态内容的校验和标记。每个 LSA 还有类型字段标识 LSA 的类型，LSA 分为如表 5-9 所示的 5 种类型。

（5） 链路状态确认（Link State Ack），用于对洪泛的确认。OSPF 的可靠更新机制通过链路状态更新和链路状态确认分组实现。

除了 Hello 分组，OSPF 路由选择分组都仅在邻接路由器上发送，分组的 IP 源地址是邻接的一端，目的地址是邻接的另一端或者是 IP 组播地址。

每个 LSA 描绘 OSPF 路由域的一部分。每个路由器发起一个 router-LSA。无论何时一个路由器被选为指定路由器，它就发起一个 network-LSA。区域边界路由器为每个已知的区域间目的地(inter-area destination)发起单个 summary-LSA。AS 边界路由器为每个已知的 AS 外目的地发起单个 AS-external-LSA。

例如考虑图 5-24 中的路由器 R4，它是一个区域边界路由器，连到区域 1 和骨干。R4 向骨干区域发起 5 个不同的 LSAs，1 个 router-LSA 和 4 个 summary-LSAs（为到网络 N1-N4 各发起 1 个）。R4 还要向区域 1 发起 8 个不同的 LSAs，1 个 router-LSA 和 7 个 summary-LSAs（其中为到网络 N6-N8 的路由各发起 1 个 summary-LSA，为到 AS 边界路由器 R5 和 R7 的路由各发起 1 个，为到主机 Ia 和 Ib 的路由合并发起 1 个，另发起 1 个通告到网络 N9-N11 和主机 H1 的路由）。如果 R4 被选为网络 N3 的指定路由器，它还将向区域 1 为 N3 发起一个 network-LSA。

再如图 5-24 中的 AS 边界路由器 R5。R5 将发起 3 个不同的 AS-external-LSAs（网络 N12-N14 各 1 个）。这些 LSAs 将被洪泛遍及整个 AS，假如没有区域被配置为末梢区域。不过，假如区域 3 被配置为末梢区域，网络 N12-N14 的 AS-external-LSAs 就不会洪泛到该区域中。而路由器 R11 将发起一个默认 summary-LSA，该 LSA 将被洪泛传遍区域 3，指示所有区域 3 的内部路由器把到 AS 外的流量发送给 R11，由它再转发。

表 5-9 OSPFv2 链路状态通告(LSAs)

LS 类型	LSA 名字	LSA 描述
1	Router-LSAs	区域中的每个路由器发起一个 router-LSA，描述路由器到本区域的接口的状态，仅洪泛遍及单个区域。
2	Network-LSAs	区域中的每个广播和 NBMA 网络由其指定路由器发起一个 network-LSA，该 LSA 包含连到该网络上的路由器列表，仅洪泛遍及单个区域。
3, 4	Summary-LSAs	由区域边界路由器发起，洪泛遍及与本 LSA 相关的区域。每个 Summary-LSA 描述一条到区域外且还在 AS 内的一个目的地的路由（即一个 inter-area route）。类型 3

		描述到网络的路由。类型 4 描述到 AS 边界路由器的路由。
5	AS-external-LSAs	由 AS 边界路由器发起，洪泛传遍 AS。该 LSA 描述至另一 AS 中的一个目的地的路由。AS 的默认路由也可以由 AS-external-LSAs 描述。

在洪泛过程中，许多 LSAs 可以包含在单个链路状态更新分组中运送。然后所有 LSAs 被洪泛传遍 OSPF 路由域。洪泛算法是可靠的，保证所有路由器拥有相同的 LSAs 集合，即链路状态数据库。

注意，唯有 AS-external-LSAs 要被洪泛传遍整个自治系统；所有其他类型的 LSAs 仅在单个区域内洪泛。不过，AS-external-LSAs 不被洪泛到末梢区域，这样可以减少末梢区域内路由器的链路状态数据库的大小。

由链路状态数据库，每个路由器构建以自己为根的最短路径树。根据树可以构建路由表，算法略，可以参见 RFC2328。

5.3.4 外部网关协议 BGP

1. BGP 概述

BGP（Border Gateway Protocol）是设计用于 TCP/IP 互联网自治系统之间的路由选择协议。它的创建是基于 EGP 及其使用经验，这里 EGP 表示一个具体的协议，定义于 RFC904 中。BGP 最初版本 BGP-1 于 1989 年在 RFC1105 中发布。后来又分别在 RFC1163、RFC1267、RFC1771 中发布了 BGP-2、BGP-3、BGP-4，最新 BGP-4 发布在 RFC4271 中。BGP-4（以后简称 BGP）增加了对 CIDR 的支持。而早期版本缺乏对 CIDR 的支持，所以都过时了，不能用于当今的因特网。

每个自治系统中需要配置一个或多个路由器运行 BGP，这些路由器称为 BGP 发言人（BGP speaker）。一对通信的 BGP 发言人也可互称为 BGP 对端（BGP peer）。BGP 发言系统的主要功能是与其他的 BGP 系统交换网络可达性信息。网络可达性信息包括可到达的网络信息以及到达网络所经过的一系列自治系统的信息。这些信息足够构造一个 AS 连通图，由图可以删除路由回路（routing loop），并可以在 AS 级别上实施一些策略决策（policy decisions）。

2. BGP 特点

BGP 特点包括：

（1）BGP 是一个自治系统之间的路由选择协议。

（2）BGP 发言系统的主要功能是与其他 BGP 系统交换网络可达性信息。BGP 通告下一跳和路径信息。

与距离向量路由选择协议类似，BGP 通告可到达的目的地和到达这些目的地各自的下一跳信息。BGP 发言人一般仅向其对端通告它自己使用的路由（指最首选的 BGP 路由，并且在转发中使用）。此外，BGP 还通告到达目的地的路径信息，允许接收方了解到达目的地的路径上的一系列自治系统，以避免路由环路以及执行路由策略。

（3）经由 BGP 交换的路由选择信息仅支持基于目的的转发模式（destination-based forwarding paradigm）。

BGP 假设路由器转发分组仅基于分组中的目的 IP 地址。这反过来决定了能够使用 BGP 实施的策略决策集。有些策略，基于目的的转发模式不支持，因而需要使用比如源路由技术来实施。这样的策略不能使用 BGP 实施。BGP 能够支持任何符合基于目的转发模式的策略。

（4）BGP 提供一组机制支持无类域间路由 CIDR。

这些机制包括支持将一组目的地作为一个 IP 前缀通告，并在 BGP 内部消除网络“类”的概念。BGP 还引入机制允许路由聚合，包括 AS 路径的聚合。

（5）BGP 假定一个 AS 内部的路由选择由 IGP 完成，BGP 对各个自治系统使用什么 IGP 没有特别的要求，对自治系统之间的互连拓扑不作限制。BGP 强调即使使用了多个 IGPs 和测度，一个 AS 的管理从其他自治系统看来应具有单个一致的内部路由选择规划并呈现一致的对通过它可达的目的地的描述。

（6）BGP 使用 TCP 作为传输协议，在 TCP 端口 179 上监听。TCP 提供可靠传输服务，因此 BGP 不需要执行显式的 BGP 报文分段、重传、确认和排序。

（7）BGP 采用增量更新以节约网络带宽。

在两个 BGP 系统之间建立一个 TCP 连接，连接上最初的数据流是输出策略所允许的 BGP 路由表（routing table）的一部分。以后当路由表有改变时，再发送增量（变化的部分）更新。BGP 不要求周期地刷新路由表。

（8）BGP 支持策略，不是简单地通告本地路由表中的路由，而是能够执行本地管理员选择的策略。例如，BGP 路由器经过配置，能够把自治系统内可达的目的地和通告给其他自治系统的目的地区分开来。

（9）BGP 需要周期地发送保活报文确保连接是活跃的；当连接发生错误时，发送通知报文并关闭 TCP 连接。

（10）BGP 提供鉴别机制，允许接收方对报文进行鉴别，即确认发送方的身份。

3. BGP 报文

BGP 定义了五种基本报文类型：OPEN（打开）、UPDATE（更新）、NOTIFICATION（通知）、KEEPALIVE（保活）和 ROUTE-REFRESH（路由刷新）。每个 BGP 报文都有固定大小的首部：

16 字节的标记	2 字节的长度	1 字节类型

在初始的报文中，标记值为全 1，如果 BGP 双方同意使用鉴别机制，标记就可以包含鉴别信息。由于 BGP 基于将高层协议数据看成是流式数据的 TCP，TCP 不提供相邻 BGP 报文之间的边界，因此需要标记字段识别报文的起始。在任何情况下，BGP 双方必须就标记值达成一致，这样才能使双方保持同步。首部中的长度字段指明以字节为计量单位的报文总长度，最小为 19 字节（不含数据部分），允许的最大报文长度是 4096 字节。类型字段用于标识报文的类型，为 1 表示是 OPEN 报文， 为 5 表示是 ROUTE-REFRESH 报文。

（1）OPEN 报文

两个 BGP 对等端一旦建立了 TCP 连接，就分别发送一个 OPEN 报文。OPEN 报告中声明发送者自己的 AS 号，并设置其他操作参数。如果 OPEN 报文被接受，对等端就会确认 OPEN 而发回 KEEPALIVE 报文。除了固定长度的 BGP 首部外，OPEN 报文还包含如图

5-26 所示的一些字段。

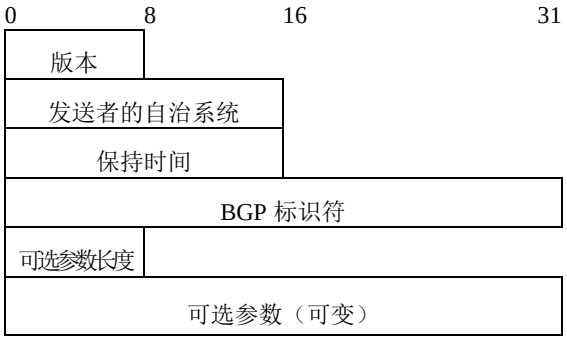


图 5-26 除 BGP 首部之外的 BGP OPEN 报文格式

其中版本字段指示报文的协议版本号，当前 BGP 版本号为 4。保持时间（Hold Time）字段用来设定保持计时器，该计时器定义了 BGP 收到来自对等端的连续的 KEEPALIVE 和/或 UPDATE 报文可以经过的最大秒数。如果保持计时器超时，则推断对等端不再可用。保持时间必须为 0 或者至少 3 秒。若为 0 表示不使用 KEEPALIVE 报文。如果保持计时器的值大于 0，标准建议把 KEEPALIVE 间隔时间设置为保持时间的 1/3，且不能小于 1 秒。4 字节的 BGP 标识符惟一标识发送方，协议规定使用 BGP 发言人的某个 IP 地址作为 BGP 标识符的值，在启动时确定 BGP 标识符的值，并且用在每个本地接口及其与 BGP 对等端的通信中。

可选参数长度指示可选参数字段以字节为单位的总长度。OPEN 报文的可选参数字段是可选的，可以没有（只需将可选参数长度字段设为 0）。如果有，可选参数包含一个参数列表。现有的参数用于有关鉴别、能力、允许 32 位 AS 号等的协商。

（2）UPDATE 报文

两个 BGP 对等端发送 OPEN 报文并得到确认后，使用 UPDATE 报文相互传送路由选择信息。UPDATE 报文中的信息可用于构建一幅描述各个自治系统关系的图。通过应用规则，可以发现路由环路和一些其他异常，以便从 AS 间路由（inter-AS routing）中删除。UPDATE 报文用于向对等端通告可行的且具有相同路径属性的路线，或者当目的地变得不可用时用来撤销曾通告过但现在不可行的路线。除固定长度的 BGP 首部外，UPDATE 报文还可能包括如图 5-27 所示的字段，其中长度可变的字段并不出现在每个更新报文中。

图 5-27 除 BGP 首部之外的 BGP UPDATE 报文格式

每个 UPDATE 报文分成两个部分，前一部分列出正准备撤销的目的地。一个 UPDATE

报文可以列出以前被通告过但现在要被撤销的多个路由。如果没有要撤销的，则撤销路由长度字段为 0，并省略撤销路由字段。撤销路由长度指出后面撤销路由字段以字节为单位的长度。后一部分给出要通告的新目的地的相关内容。如果没有新的目的地要通告，则 TPAL 字段为 0，此时报文不包含路径属性和 NLRI 字段。

撤销路由和 NLRI 都是变长字段，都包含 IP 地址前缀的列表。为了适用于无分类编址，每个 IP 地址前缀编成一个 2 元组的形式<长度，前缀>。长度字段占 1 个字节，指明 IP 地址前缀的二进制位的长度。长度为 0 表示匹配所有 IP 地址的前缀，此时没有前缀字段。前缀字段包含一个 IP 地址前缀，后面接若干个使本字段位数为 8 的倍数的填充比特，这些后缀比特的值任意。

总的路径属性长度字段指出要通告的路由相关的路径属性的长度，以字节为单位。由此可以确定 NLRI 字段的长度=BGP 报文长度-19-4-撤销路由长度-TPAL。

路径属性字段包含路径属性列表，每个路径属性是个变长的三元组<属性类型，属性长度，属性值>。属性类型占 2 个字节，由属性标志八位组和属性类型码八位组组成。属性标志主要用来标识属性是熟知的（well-known）还是可选的（optional），是可传递的还是不可传递的。属性类型码表示属性的类型，RFC4271 中定义了 7 种属性类型，见表 5-10。目前已定义了 22 种路径属性，可参见 IANA 的在线发布。

表 5-10 BGP 路径属性

属性名称	类型码	含义
ORIGIN	1	熟知的强制的属性，定义路径信息的来历，可以来源于 IGP、EGP 或其他
AS_PATH	2	熟知的强制的属性，由一系列 AS 路径段组成，每个 AS 路径段包含一个或多个 AS 号
NEXT_HOP	3	熟知的强制的属性，定义应该用作到 NLRI 字段中列出的目的地的下一跳的路由器的（单播）IP 地址
MULTI_EXIT_DISC	4	可选的非传递的属性，一个 BGP 发言人的决策处理可能用该属性的值来区别到一个相邻自治系统的多个进入点
LOCAL_PREF	5	熟知的属性，一个 BGP 发言人用它来通知其他内部对等端它对被通告路由的优先等级
ATOMIC_AGGREGATE	6	熟知的任意的属性，表示被聚合的路由中不含路由环路
AGGREGATOR	7	可选可传递的属性，6 字节长，包含形成聚合路由的最后一个 AS 号，以及形成聚合路由的 BGP 发言人的 IP 地址

一个 UPDATE 报文至多通告一组路径属性，但可以通告多个共享这些属性的目的地。给定 UPDATE 报文中包含的所有路径属性适用于该报文 NLRI 字段中所携带的所有目的地。

（3）KEEPALIVE 报文

BGP 并不基于 TCP 的保活机制来确定对等端是否可达，而是在对等端之间通过足够

多次地交换 KEEPALIVE 报文避免保持计时器超时。KEEPALIVE 报文仅包含首部，19 字节长。

（4）NOTIFICATION 报文

当检测到错误状况时，BGP 发送通知（NOTIFICATION）报文，然后立即关闭 BGP 连接。除固定长度的 BGP 首部外，通知报文还包含如图 5-28 所示的内容。差错码指明通知的类型，目前定义了 6 个差错码：报文首部差错、OPEN 报文差错、UPDATE 报文差错、保持计时器超时、有限状态机差错和停止（Cease），差错码值分别为 1 到 6。差错子码提供更具体的有关错误的信息，每个差错码可能有若干差错子码，如果没定义子码，则子码字段置为 0。数据字段是变长的，用于诊断通知的原因，数据字段的内容取决于具体的错误。

差错码(1 字节)	差错子码(1 字节)	数据（可变）

图 5-28 除 BGP 首部之外的 BGP 通知报文格式

（5）ROUTE-REFRESH 报文

BGP 发言人之间可以动态地交换路由刷新请求，然后再重新通告各自的 Adj-RIB-Out（允许通告的路由信息）。BGP 不要求周期刷新路由表，为允许本地策略改变时不用复位 BGP 连接，BGP 发言人应该保留其对等端向它通告的当前版本的路由信息，或者利用路由刷新功能。利用路由刷新功能，可以避免维护开销。一个 BGP 发言人若愿意接收来自对等方的路由刷新报文，在 BGP 会话建立时可使用能力通告（OPEN 报文的能力可选参数）向对等方通告路由刷新能力。路由刷新报文的格式略，可参见 RFC2918。专用网络与互连（VPN 和 NAT）

5.4 专用网络互连（VPN 和 NAT）

5.4.1 虚拟专用网 VPN

前面说过，因特网可以看成是单一的虚拟网络，所有的计算机都与它相连，这是一种单层抽象结构。因特网也可以看成一种双层结构。在这种结构中，每个机构有一个专用互联网，另外有一个中央互联网连接各个专用互联网。

专用互联网内主机之间的通信相对于外界应该是不可见的，即私密的。如果一个机构仅由一个网点组成，容易保证私密性。如果一个机构由分散的多个网点构成，为了保证私密性，最直接的方法是租用数字线路或帧中继永久虚电路来连接各个网点，不过成本较高。虚拟专用网（Virtual Private Network，VPN）技术提供了一种低成本的替代方法，允许机构使用因特网互连多个网点，并用加密来保证网点之间的通信量的私密性。

实现 VPN 有两种基本技术：隧道传输技术和加密技术。VPN 定义的是一条从某网点的一个路由器到另一个网点的一个路由器之间的通过因特网的隧道，使用 IP-in-IP 封装要经过隧道转发的数据报。为了防止经过因特网时被窥视，在将外发数据报封装到另一个数

据报之前，先要将整个数据报进行加密。VPN 使用的 IP-in-IP 封装如图 5-29 所示。

图 5-29 VPN 使用的 IP-in-IP 封装

当发自一个网点的数据报通过隧道到达接收路由器时，路由器先将数据区解密，还原出内层数据报，再将其转发给另一网点内的某台主机。

下面简单了解一下 VPN 的路由选择技术。图 5-30 所示为一个 VPN 以及处理隧道的一个路由器的路由表。考虑从 128.9.2.0/24 网络上某主机向 128.9.4.0/24 网络上某主机发送数据报。发送主机首先将数据报转发给 R2，R2 再把数据报转发给 R1。根据路由表，R1 应将数据报通过隧道转发给 R3。因此，R1 先对数据报做加密处理，再把它封装在外层数据报中（源宿分别为 R1 和 R3）。然后，R1 通过本地 ISP 转发外层数据报，经过因特网的外层数据报到达 R3 并被识别后，R3 先将其数据区进行解密，还原出原始数据报，再取出目的地址在本地路由表中查找，然后将原始数据报转发至 R4，由它进行最后的交付。

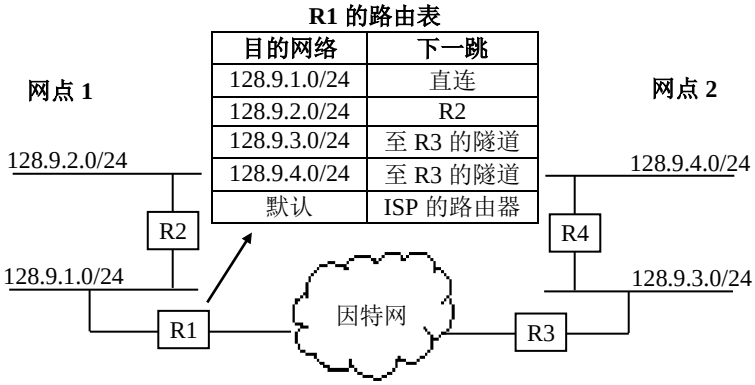


图 5-30 包含 2 个网点的 VPN 示例

VPN 能够为机构的各网点之间提供成本不高且能保密的通信服务，网点中可以只有一台主机。例如，现在不少公司向其员工提供 VPN 软件，以便员工能够利用因特网相对安全（防窃听）地访问本公司的网络。

学习 CIDR 时，我们提到过为节约 IPv4 地址，各专用互联网可以使用专用地址。当采用专用地址时，每个网点只需要一个全球有效的 IP 地址，用于 VPN 需要的隧道传输。

专用互联网内主机一般不仅需要和本互联网内主机通信，还需要访问因特网的其他部分。如果专用互联网没有为网点中每个主机都分配全球有效 IP 地址，那么可以使用应用网关（application gateway）或 NAT（Network Address Translation）技术实现网点内使用专用地址的主机和因特网上本专用互联网外的主机通信。下面简单介绍 NAT 技术。

5.4.2 网络地址转换 NAT

NAT（网络地址转换）提供一种机制将使用专用 IP 地址的域和使用全球唯一注册 IP

地址的外部域连通。NAT 要求网点具有一条到因特网的连接，至少有一个全球唯一 IP 地址 G。可在互连网点和因特网的路由器上运行 NAT 软件，将 G 分配给该路由器。运行 NAT 软件的计算机称为 NAT 盒（NAT box）。有两种传统 NAT 方法：基本网络地址转换（基本 NAT）和网络地址和端口转换（NAPT）。

基本 NAT 对传入数据报和外发数据报中的地址进行转换。用 G 替代每个外发数据报中的源地址，同时 NAT 在 NAT 转换表中记录外发数据报的源和目的地址。这样，从外部主机的角度看，所有数据报都来自 NAT 盒。传入数据报从因特网到达 NAT 时，NAT 在转换表中查找传入数据报的源地址，提取相应的内部主机地址，用主机的地址替换数据报中的目的地址，再通过网点内互联网把数据报转发给内部主机。整个过程对于通信双方是透明的。缺点是如果 NAT 盒仅有一个全球唯一地址，则不允许网点内同时有多台主机并发访问给定的某个外部地址。对此，多地址 NAT（持有多个全球唯一地址）可提供多个内部主机并发访问给定的某外部主机。

NAPT 通过转换 TCP 或 UDP 协议端口号以及地址允许并发访问。NAPT 对 NAT 转换表要做扩展，除了一对源地址以外，还要包含一对源和目的协议端口号（端口号将在下一章讨论），以及转化后的本地端口号（NAPT 使用的协议端口号）。

例 5-9 有 5 个内网主机在与外部主机通信，已知 NAPT 的全球有效地址为 G，NAPT 使用的网络地址与端口转换表如表 5-11 所示，请写出 NAPT 转换前后各 TCP 连接的五元组标识。

表 5-11 NAPT 的转换表举例

内部地址 (专用地址)	内部端口	NAPT 端口	外部地址	外部端口	所用协议
10.10.8.27	21043	14007	211.23.33.12	80	tcp
10.10.9.23	43572	14012	211.23.33.12	80	tcp
10.10.9.12	21043	14013	211.23.33.12	80	tcp
10.10.12.124	9542	14015	130.126.13.45	21	tcp
10.10.1.10	5112	14018	202.115.232.57	6919	udp

解：TCP 连接可以用五元组（本地端点的 IP 地址，端口号；对端主机的 IP 地址，端口号；tcp）来标识（参见下一章）。表中包括 4 个 TCP 连接，其中前 3 个表示主机正在访问同一外部主机的同一 TCP 端口。这 4 个连接在网点内部和网点外部（经过 NAPT 转换后）的五元组标识见表 5-12。

表 5-12 经过 NAPT 转换前后的 TCP 连接的五元组标识

在网点内部	经过 NAPT 转换后
（10.10.8.27, 21043, 211.23.33.12, 80, tcp）	（G, 14007, 211.23.33.12, 80, tcp）
（10.10.9.23, 43572, 211.23.33.12, 80, tcp）	（G, 14012, 211.23.33.12, 80, tcp）
（10.10.9.12, 21043, 211.23.33.12, 80, tcp）	（G, 14013, 211.23.33.12, 80, tcp）
（10.10.12.124, 9542, 130.126.13.45, 21, tcp）	（G, 14015, 130.126.13.45, 21, tcp）

从示例可以发现，NAPT 的优点是能够仅用一个全球有效地址获得通用性、透明性和并发性。主要缺点是通信仅限于 TCP 和 UDP。对于 ICMP，NAT 需要另做处理以维持透明性。此外，如果需要在应用协议数据中传递地址或端口信息，也不能使用 NAT，除非使 NAT 能够识别应用，并对协议数据做必要的修改。绝大多数 NAT 实现只能识别很少几个应用。

5.5 下一代网际协议 IPv6

上世纪 90 年代初，研究人员认为 32 位 IPv4 地址空间很快就会耗尽。然而，现在的事实是根据最新预测，IPv4 地址还够我们使用十几甚至二十年。这主要归功于无分类编址 CIDR 技术使 IP 地址的分配更加合理，提高了地址的利用率；并且网络地址转换 NAT 技术极大地缓解了 IP 地址空间的消耗。

IPv4 出现于上世纪 70 年代末，这是第一个被实际应用的版本。下一个可能替代的新版是 IPv6。更新 IP 的主要动机是提供一个比 IPv4 大得多的全局地址空间。

5.5.1 IPv6 的主要特点

IPv6 协议保留了 IPv4 赖以成功的许多特点，包括无连接交付、允许发送方选择数据报的大小、要求发方指明数据报在到达目的地前允许经过的最大跳数，以及允许分片和支持源路由功能。

尽管许多概念类似，IPv6 还是改变了许多协议细节，IPv6 完全修订了数据报的格式。IPv6 与 IPv4 相比主要的变化有：

（1）更大的地址空间。IPv6 的地址字段为 128 位，IPv6 地址空间增大为 IPv4 地址空间的 2^{128-32} 倍。

（2）扩展的地址层次。IPv6 地址空间可以划分为更多的层次。

（3）灵活的首部格式。IPv6 定义了一组可选的扩展首部（extension header），用于取代 IPv4 中可变长度的选项字段。每个 IPv6 数据报除包含固定长度（40 字节）的基本首部外，还可以包含若干扩展首部。路由器一般仅处理基本首部，对扩展首部不处理（逐跳扩展首部除外），这有利于提高路由器处理数据报的速度。

（4）增强的选项。IPv6 提供了一些 IPv4 所不具备的新选项功能。

（5）对协议扩展的保障。IPv6 没有指明所有的细节，允许协议适应底层网络或新应用的变化。

（6）支持自动配置和重新编号。IPv6 允许孤立网络上的主机自动分配本地地址，还允许管理员动态地给网络重新编号。

（7）支持资源预分配。IPv6 支持流（flow）的抽象，通过将流与资源分配相关联来支持区分服务（DiffServ）功能。

IPv6 数据报格式如图 5-31 所示，包含一个 40 字节的基本首部和 0 个或多个扩展首部，以及数据部分。注意扩展首部和数据合起来称为 IPv6 数据报的有效载荷。

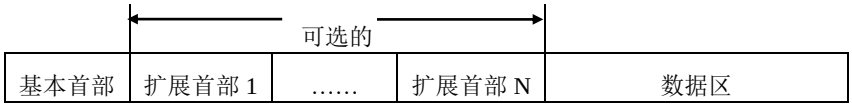


图 5-31 IPv6 数据报的一般形式

5.5.2 IPv6 基本首部格式

如图 5-32 所示，IPv6 基本首部包含了更长的地址字段，但所包含的字段数比 IPv4 的还少一些。首部包含信息的变化反映了协议的变化。

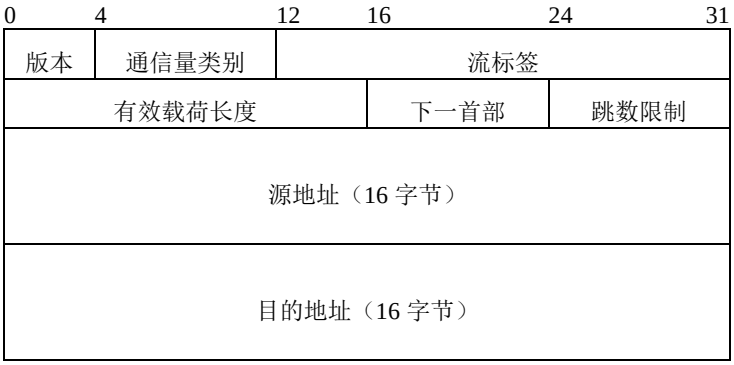


图 5-32 40 字节 IPv6 基本首部的格式

与 IPv4 首部的固定部分相比，IPv6 基本首部主要有下列变化：

（1）由于基本首部长度固定，取消了 v4 中的首部长度字段，v4 中的数据报总长度字段被有效载荷长度字段所取代。

（2）源、目的地址由 4 字节增大到 16 字节（128 位）。

（3）分片有关字段被转移到了“分片扩展首部”中。

（4）生存时间字段改名为跳数限制（hop limit）字段。

（5）服务类型字段改名为通信量类别（traffic class）字段，并增加了流标签（flow label）字段，一并用于支持资源的预分配。

（6）协议字段由指明后续内容格式的下一首部字段替代，注意下一首部可能是 IPv6 数据报的扩展首部，也有可能是 ICMP、TCP、UDP、IGMP、OSPF 等首部。

IPv6 定义了下列扩展首部：逐跳选项（Hop-by-Hop Options）、源路由（Routing）、分片（Fragment）、目的地选项（Destination Options）、认证（Authentication）和封装安全净荷（Encapsulating Security Payload）扩展首部。这里从略，下面简单介绍 IPv6 的编址。

5.5.3 IPv6 编址

1. 编址模型

IPv6 地址有三种类型：

（1）单播（Unicast） 单个接口的标识符。发向一个单播地址的分组被交付给由该地址标识的接口。

（2）任播（Anycast） 一组接口（一般属于不同的节点）的标识符。发向一个任播地址的分组被交付给该地址标识的其中一个接口（最近的那个接口，根据路由选择协议的距

离度量）。

（3）组播（Multicast）一组接口（一般属于不同的节点）的标识符。发向一个组播地址的分组被交付给由该地址标识的所有接口。IPv6 中没有广播地址，广播被看作是组播的一个特例。

与 IPv4 地址一样，所有类型的 IPv6 地址被分配给接口，而不是节点。一个 IPv6 单播地址涉及单个接口。每个接口属于单个节点，节点的任何一个接口的单播地址可以用作节点的标识符。

所有接口要求拥有至少一个本地链路单播地址。单个接口可以有多个任意类型或范围的 IPv6 地址。另外，和 IPv4 模型一样，子网前缀与一个链路相关联。允许给同一链路分配多个子网前缀。此外，IPv6 编址模型中有一个例外：一个单播地址或一组单播地址可以被分配给多个物理接口，如果实现把多个物理接口作为一个接口来处理。这对于基于多个物理接口的负载分配是有用的。

2. 地址的表示

IPv6 地址较长，为表示简洁些，使用冒号分十六进制记法。另外由于 IPv6 地址中常包含长的 0 比特串，因此允许**零压缩**，即使用″::″表示一个或多个连续的 16 比特 0。为避免搞不清″::″表示几个 16 比特 0，规定在任何一个地址中只能使用一次零压缩。

【例 5-10】将下列地址记法进行零压缩：

单播地址　　2001:DB8:0:0:8:800:200C:417A

组播地址　　FF01:0:0:0:0:0:0:101

回送地址　　0:0:0:0:0:0:0:1

未指定地址　0:0:0:0:0:0:0:0

解：上述地址记法零压缩后可以写成：

2001:DB8::8:800:200C:417A

FF01::101

::1

::

IPv6 还支持冒号分隔和点分隔混合记法：x:x:x:x:x:x:d.d.d.d，其中高位的 6 个 x 表示 6 个十六进制数，低位的 4 个 d 表示 4 个十进制数。这种记法适用于表示 IPv4 兼容或映射的 IPv6 地址，例如：0:0:0:0:0:0:13.1.68.3、0:0:0:0:0:FFFF:129.144.52.38，相应的压缩表示为：::13.1.68.3、::FFFF:129.144.52.38。

此外，CIDR 的斜线表示法仍然可用。IPv6 地址前缀表示为：ipv6 地址/前缀长度。

3. 地址空间的分配

IPv6 地址的高位标识 IPv6 地址的类型，见表 5-15。

表 5-13 IPv6 地址类型

地址类型	二进制前缀	IPv6 记法	解释
非特指	00...0 (128 bits)	::/128	不可分配给任何节点,仅用作源地址,且路由

(Unspecified)			器不转发源地址为非特指地址的 IPv6 分组
回送(Loopback)	00...1 (128 bits)	::1/128	回送地址,不可分配给任何物理接口
组播	11111111	FF00::/8	
本地链路单播	1111111010	FE80::/10	用于单个链路,仅在本地范围有意义
全球单播	其他		

以上除组播地址外，都是单播地址。取决于担当的角色，IPv6 节点可以非常了解也可以不了解 IPv6 地址的内部结构。一般的单播地址可分为 2 部分：子网前缀和接口 ID。接口 ID 用于标识链路上的接口。在某些情况下，接口标识符可直接由接口的链路层地址派生出来。同一接口标识符可以用于一个节点的多个接口上，只要它们连接到不同的子网。所有单播地址，除了以 000 比特开头的，接口 ID 的长度要求是 64 比特。

全球单播地址的一般格式如图 5-33 所示。其中全球选路前缀是分配给一个网点（一群子网/链路）的值，子网 ID 是网点内链路的标识符。接口 ID 见前面说明。非 000 比特开头的全球单播地址的接口 ID 长 64 比特，以 000 比特开头的则不受此限。

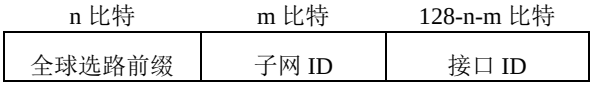


图 5-33 IPv6 全球单播地址的一般格式

以 000 比特开头的全球单播地址的例子是在低 32 比特嵌入 IPv4 地址的 IPv6 地址。有 2 种： IPv4 兼容（IPv4-Compatible）IPv6 地址和 IPv4 映射（IPv4- mapped）IPv6 地址，具体格式可参见 RFC4291。

下面简单了解一下任播地址。任播地址从单播地址空间分配，使用任何已定义的单播地址格式，因此从地址本身无法区分二者。当单播地址分配给不止一个接口时，就转成了任播地址，被赋予该地址的节点必须被显式地配置了解这是一个任播地址。

任播地址可以用于标识连接到一个特别子网上的路由器集合，标识提供到一个特别路由域入口的路由器集合等。目前预定义了子网-路由器任播地址（Subnet-Router anycast address），其格式如图 5-34 所示。其中"子网前缀"标识一个特定的链路，发向子网-路由器任播地址的 IPv6 分组将被交付给该子网上一个路由器。所有路由器要求支持其直连子网的子网-路由器任播地址。该地址拟用于节点需要和一组路由器中任何一个通信的应用。

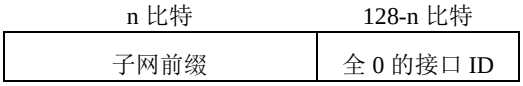


图 5-34 IPv6 子网-路由器任播地址格式

本章小结

TCP/IP 体系的核心层是网络层和传输层，相应的核心协议是 IP 和 TCP 两大协议。本章主要讨论 TCP/IP 体系的网络层，也称 IP 层。IP 层负责为不同物理网络上的主机提供通信功能，另一任务是路由选择，主要内容包括：

（1）网络层编址。IP 编址屏蔽了物理网络编址细节。介绍了最初的分类编址方案，能够给多个子网分配相同分类 IP 网络地址的子网划分方案，以及现在正在使用的能够进一步提高地址使用率的无分类编址方案。

（2）IP 地址到物理地址的解析 ARP 协议。ARP 报文被直接封装在物理帧中发送，查询主机通过发送 ARP 请求和接收 ARP 响应解析本物理网络上另一主机的物理地址。

（3）IP 协议的三大功能：无连接的数据报交付、数据报的转发、IP 差错与控制（ICMP）。数据报屏蔽了物理网络帧的细节。IP 软件负责数据报的转发，根据每个数据报中的目的地址查找源主机或路由器上的路由表决定把数据报发往何处。ICMP 是 IP 的一个组成部分，当一个数据报产生差错时，使用 ICMP 向其源站报告差错情况，ICMP 也用于提供信息或网络测试。

（4）因特网路径建立与刷新机制。介绍了自治系统（AS）的概念，用于 AS 内部的路由选择协议 RIP 和 OSPF，以及用于 AS 之间的路由选择协议 BGP-4。RIP 使用距离向量算法。而 OSPF 使用链路状态算法，并支持将自治系统划分区域，因此适用于较大的 AS。一个 AS 中的 BGP 发言人使用 BGP 与位于另一 AS 中的对等端通信，双方相互通告自己的网络可达性信息，从而允许不同自治系统中的主机能够相互通信。

此外，本章还简单讨论了虚拟专用网 VPN、网络地址转换 NAT、下一代网际协议 IPv6 和网络互连设备：

（1）虚拟专用网 VPN 技术提供了一种低成本的方法，允许一个拥有分散的多网点的机构，使用因特网互连多个网点，并保证网点之间通信的私密性。VPN 主要利用加密和隧道技术实现。

（2）网络地址转换 NAT 提供一种机制将使用专用 IP 地址的域和因特网连通。主要方法是在网点边界设置一个 NAT 盒，由它将外出数据报的源 IP 地址和源端口转换为一个全球唯一地址和一个 NAT 本地唯一的端口号，并在转换中登记，以便对进入数据报作对应逆操作。

（3）IPv6 数据报由固定长度（40 字节）的基本首部、若干个扩展首部和一个数据区组成。IPv6 拥有 128 比特长的地址，分为单播、任播和多播三种。

（4）网络互连设备按照其工作的层次，可以分为物理层、数据链路层、网络层、高层互连设备。路由器基本功能包括路由查找与数据报转发和路由维护。

练习题

5.1 网络互连有何实际意义？进行网络互连时，有哪些共同的问题需要解决？

5.2 转发器、网桥和路由器都有何区别？

5.3 试简单说明IP、ARP、RARP和ICMP协议的作用。

5.4 分类IP地址共分几类？各如何表示？单播分类IP地址如何使用？

5.5 试说明IP地址与硬件地址的区别，为什么要使用这两种不同的地址？

5.6 简述以太网上主机如何通过ARP查询其默认路由器的物理地址。

5.7 试辨认以下IP地址的网络类别：

（1）138.56.23.13 （2）67.112.45.29 （3）198.191.88.12 （4）191.62.77.32

5.8 IP数据报中的首部检验和并不检验数据报中的数据，这样做的最大好处是什么？坏处是什么？

5.9 当某个路由器发现一数据报的检验和有差错时。为什么采取丢弃的办法而不是要求源站重传此数据报？计算首部检验和为什么不采用CRC检验码？

5.10 在因特网中分片传送的IP数据报在哪儿进行组装，这样做的优点是什么？

5.11 假设互联网由两个局域网通过路由器连接起来。第一个局域网上某主机有一个400字节长的TCP报文传到IP层，加上20字节的首部后成为IP数据报，要发向第二个局域网。但第二个局域网所能传送的最长数据帧中的数据部分只有150字节。因此数据报在路由器处必须进行分片。试问第二个局域网向其上层要传送多少字节的数据？

5.12 一个数据报长度为4000字节（包含固定长度的首部）。现在经过一个网络传送，但此网络能够传送的最大数据长度为1500字节。试问应当划分为几个短些的数据报片？各数据报片的数据字段长度、片偏移字段和MF标志应为何数值？

5.13 如何利用ICMP报文实现路径跟踪？

5.14 划分子网有何意义？子网掩码为255.255.255.0代表什么意思？某网络的现在掩码为255.255.255.248，问该网络能够连接多少台主机？某一A类网络和一B类网络的子网号分别占16比特和8比特，问这两个网络的子网掩码有何不同？

5.15 设某路由器建立了如下表所示的路由表:

目的网络	子网掩码	下一跳
128.96.39.0	255.255.255.128	接口 0
128.96.39.128	255.255.255.128	接口 1
128.96.40.0	255.255.255.128	R2
192.4.153.0	255.255.255.192	R3
*（默认）	—	R4

此路由器可以直接从接口0和接口1转发分组，也可通过相邻的路由器R2、R3和R4进行转发。现共收到5个分组，其目的站IP地址分别为：

（1）128.96.39.10 （2）128.96.40.12 （3）128.96.40.151 （4）192.4.153.17 （5）192.4.153.90

试分别计算其下一站。

5.16 某单位分配到一个B类IP地址，其网络号为129.250.0.0。该单位有4000台机器，平均分布在16个不同的地点。如选用子网掩码为255.255.255.0，试给每一个地点分配一个子网号码，并算出每个地点主机号码的最小值和最大值。

5.17 设某ISP（因特网服务提供者）拥有CIDR地址块202.192.0.0/16。先后有四所大学（A、B、C、D）向该ISP分别申请大小为4000、2000、4000、8000个IP地址的地址块，试为ISP

给这四所大学分配地址块。

5.18 简述采用无分类编址时的IP数据报转发算法。

5.19 试简述RIP、OSPF和BGP路由选择协议的主要特点。

5.20 有个IP数据报从首部开始的部分内容如右所示(16进制表示)，请标出IP首部和传输层首部，并回答：

45 00 02 79 1C A4 40 00
80 06 00 00 0A 0A 01 5F
DA 1E 73 7B 07 38 00 50
19 71 85 77 7F 25 2B AA
50 18 FF FF 5B 6E 00 00
47 45 54 20 2F 73 2F 62
6C 6F 67 5F 34 62 63 66
64 64 63 64

（1）数据报首部长度和总长度各为多少字节？

（2）数据报的协议字段是多少，表示什么意思？

（3）源站IP地址和目的站IP地址分别是什么？(用点分十进制表示)

（4）TTL、校验和字段是多少？

（5）源端口和宿端口是什么？并请推测所用的应用层协议是什么？

5.21 以下地址前缀中的哪一个与2.52.90.140匹配？

（1）0/4　　（2）32/4　　（3）4/6　（4）80/4

5.22 分析划分子网、无分类编址以及NAT是如何推迟IPv4地址空间的耗尽的？

5.23 简述NAPT的优缺点。

5.24 简述VPN主要作用及其技术要点。

5.25 IPv6没有首部检验和。这样做的优缺点是什么？

5.26 IPv6地址有几种基本类型？