第5章 网 络 层

网络层是互联网体系结构中最重要的一层，其主要任务是向上层提供主机到主机的通信服务。网络层的关键设备是路由器，其主要功能包括分组转发和路由选择。

本章主要内容包括网络层的控制平面和数据平面功能概述，互联网协议(IP)以及IP数据报格式，IP分组转发的算法，互联网控制报文协议(ICMP)的报文格式、种类以及应用实例，路由选择协议的分类以及路由选择协议RIP、OSPF和BGP，网络地址转换(NAT)和虚拟专用网络(VPN)的基本概念，多协议标记交换(MPLS)的概念和典型应用。

5.1 网络层概述

在第3、4章中，已经介绍了应用层协议和传输层协议，它们都是在网络边缘部分的主机中实现的。本章主要介绍的网络层协议与之不同，不论在网络边缘部分的主机上，还是在网络核心部分的路由器上，都会实现网络层协议。

从第1章已经知道，互联网采用的交换方式是分组交换。分组交换是一种动态地按需分配通信资源的交换方式。实现分组交换的关键设备是网络核心部分的路由器，其任务是将收到的分组转发到下一个网络。路由器中的网络层是本章介绍的重点。

5.1.1 传统网络的控制平面和数据平面

路由器是一种具有多个接口的专用计算机，每个接口连接了不同的网络。每个网络可以采用不同的体系结构或不同的协议实现，也就是说路由器能够连接异构的网络。

路由器的主要功能包括分组转发和路由选择，其中分组转发功能属于数据平面，路由选择功能属于控制平面。每台路由器由实现路由选择功能的控制平面和实现分组转发功能的数据平面构成。一个传统的路由器结构如图5.1所示。

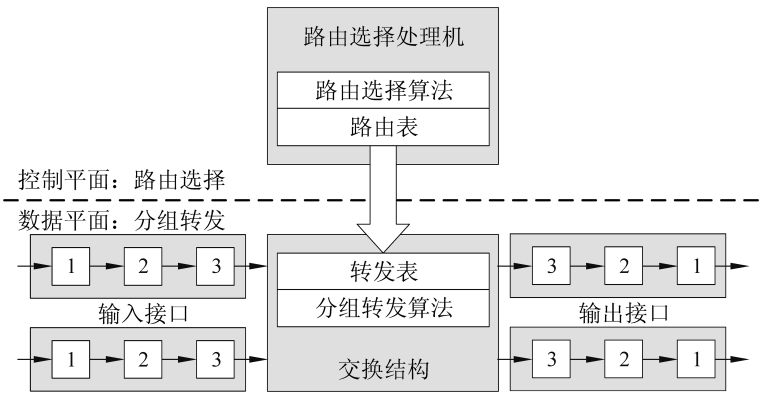
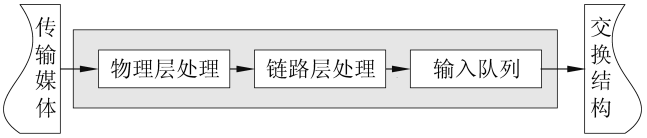


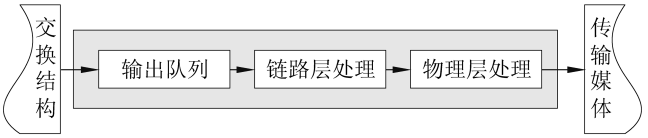
图5.1 路由器的结构

(1) 控制平面的核心构件是路由选择处理机。路由选择处理机的任务包括利用路由选择协议与其他的路由器通信,获得网络拓扑结构的相关信息;根据获取的信息,利用路由选择算法计算到目的网络的路由,构造和更新路由表。

(2) 数据平面由一组输入接口、一组输出接口和交换结构组成。交换结构是数据平面的核心构件,它的任务是通过分组转发算法查找转发表,将输入分组转发到适当的输出接口;输入接口在执行必要的物理层和数据链路层功能后,将收到的分组放入接口输入队列,如图 5.2(a)所示;输出接口从交换结构接收分组并将其放入接口输出队列,在执行必要的数据链路层和物理层功能后,将分组发送出去,如图 5.2(b)所示。



(a) 输入接口



(b) 输出接口

图 5.2 输入接口和输出接口

交换结构可以采用多种方法实现。常见的方法包括内存交换、总线交换和互连网络交换。数据平面中的转发表是由控制平面中的路由表得到的。路由表一般由软件实现,其数据结构适用于网络拓扑变化后,对其进行高效地增、删和更新操作;而转发表一般由硬件实现,其数据结构适用于快速查找操作。虽然二者采用了不同的实现方式,但本书在介绍路由器原理和网络层原理时,对路由表和转发表不做区分。

传统网络的控制平面是分布式实现的。每台路由器中都包含控制平面,如图 5.3 所示。

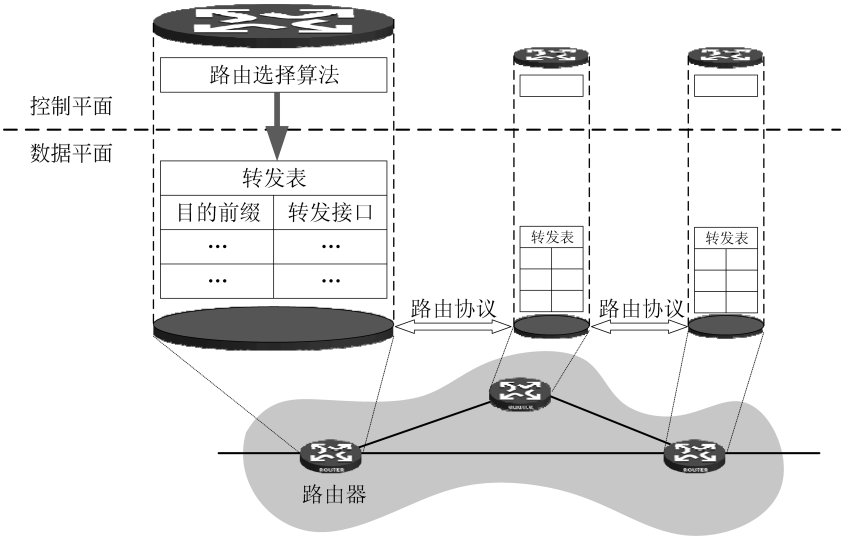


图 5.3 传统网络的控制平面与数据平面

每台路由器通过路由协议与其他路由器交换网络拓扑信息，独立维护路由表(转发表)。路由表包括目的前缀(CIDR 前缀)、掩码、转发接口或者下一跳 IP 地址等信息，每个路由表项指明到某一个 CIDR 地址块的路径信息。

传统网络的数据平面采用基于目的地址的转发策略。路由器根据收到分组的目的 IP 地址，查找转发表，转发分组。关于路由表和分组转发算法将在 5.3 节中介绍。如果要提供分组过滤、加密等服务，则需要在路由器上安装运行支撑软件，如虚拟专用网络(virtual private network，VPN)软件、网络地址转换(network address translation，NAT)软件、防火墙(firewall)软件等。

5.1.2 软件定义网络的控制平面和数据平面

软件定义网络(software defined network，SDN)是一种将控制平面和数据平面分离，构建可编程控制的网络体系结构。SDN 的网络交换设备仅需实现数据平面的功能，而控制平面的功能集中在远程控制器上实现。为区别于传统路由器，SDN 将受控网络交换设备称为 SDN 网元(network element，NE)或 SDN 交换机。SDN 的控制平面和数据平面如图 5.4 所示。

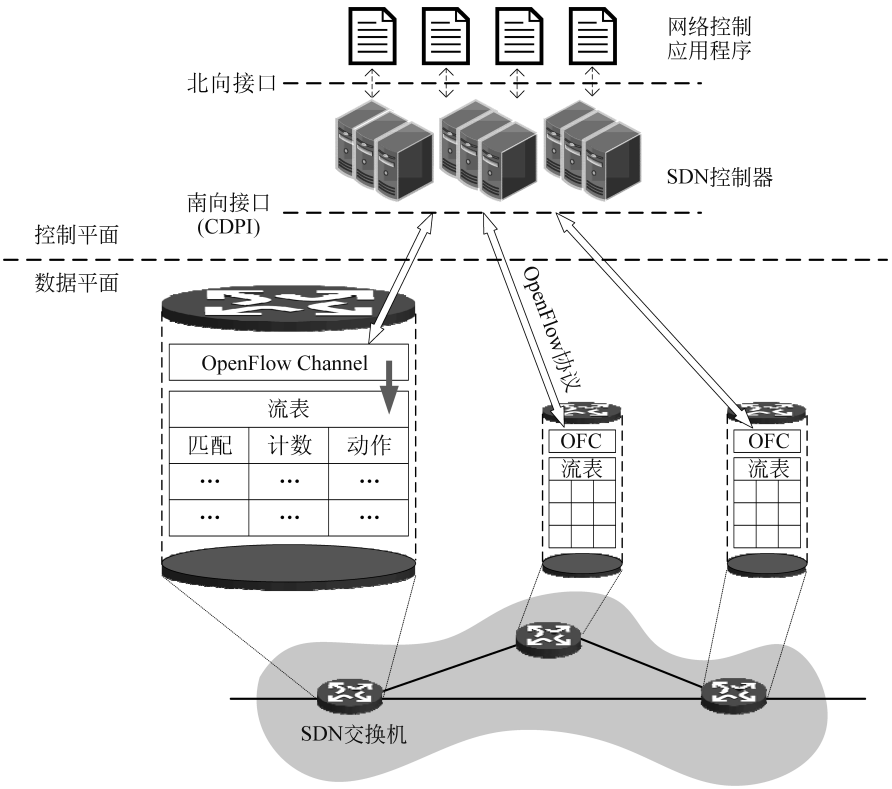


图 5.4 SDN 的控制平面与数据平面

SDN 控制平面的实现是集中式的。SDN 的控制逻辑全部在 SDN 控制器中实现，SDN 控制器通过控制数据平面接口(control to data-plane interface，CDPI)对 SDN 交换机进行控制和管理，控制数据平面接口也称为南向接口(southbound interface，SBI)。SDN 控制器

与 SDN 交换机之间的通信基于 OpenFlow 协议。SDN 交换机通过 OpenFlow 协议向 SDN 控制器传递本地观察到的事件;SDN 控制器利用收集到的这些事件,管理网络状态信息,维护流表(flow table),并通过 OpenFlow 协议将流表下发给 SDN 交换机。流表包括匹配域(match fields)、计数器集(counters)和动作(actions)等信息,每个流表项是一条规则,指明与流表项匹配的分组应该执行的动作。

流表的匹配域是首部字段的集合。OpenFlow 协议规范允许基于一组首部字段与收到的分组进行匹配操作,这些首部字段分别来自传输层协议首部、网络层协议首部和数据链路层协议首部。计数器集包括分组数、字节数、持续时间等已经与该表项匹配的分组的统计数据。动作是指当收到的分组与流表项匹配时应该采取的转发、丢弃、修改指定的首部字段等操作。

SDN 控制器通过北向接口(northbound interfaces,NBI)向网络控制应用程序开放编程能力。网络控制应用程序利用 SDN 控制器提供的 API 来定义和控制网络设备中的数据平面。例如,一个路由选择应用程序可以定义分组转发规则;一个防火墙应用程序可以定义分组通过或丢弃规则。如果需要在 SDN 中提供额外的服务,仅需要编写并部署新的网络控制应用程序,不需要在 SDN 控制器或 SDN 交换机中安装软件。

SDN 的数据平面采用通用转发策略,即基于流表的转发策略。SDN 交换机根据所收分组中的首部字段,匹配流表,执行匹配的动作。由于 SDN 的转发策略能够匹配协议栈中的多个首部字段,比传统路由器更“通用”,因此 SDN 的分组转发被称为通用转发。

SDN 控制平面与数据平面的分离会带来以下优点。

(1) 网络的全局优化。集中式的控制平面更易于进行网络的全局优化。

(2) 灵活性。通过开发新的网络控制应用程序可以部署新业务,更易于新业务的灵活和快速部署。

(3) 开放性。传统网络设备中的控制平面功能由网络设备厂商开发实现,与网络设备捆绑销售,SDN 的北向接口向软件企业开放了编程能力,可使更多的软件企业参与网络控制应用程序的开发。

SDN 的控制平面与数据平面分离也会带来以下问题。

(1) 服务能力问题。随着网络规模的扩大,集中式控制结点的服务能力有可能成为网络性能的瓶颈。

(2) 单点故障问题。SDN 控制器的故障会造成整个受控网络发生故障,因此 SDN 控制器通常以控制器集群的形式存在。

(3) 高可用性问题。传统网络设备的控制平面和数据平面集成在一起,数据平面与控制平面的通信延迟极小,数据平面具备高可用性,但 SDN 的控制平面与数据平面通过网络远程连接,网络的延迟可能会带来数据平面可用性问题。

虽然 SDN 已经提出并发展多年,目前也已经有很多支持 SDN 的网络设备面世并应用,部分基于 SDN 的网络已经商用,但是在互联网领域,由于以下几点原因,SDN 仍不可能完全取代传统网络。

(1) SDN 仍然没有统一的国际标准。

(2) 互联网上已经部署了大量的传统网络设备。

(3) 互联网中自治系统之间所用的路由协议为边界网关协议(border gateway

protocol,BGP),它的功能和作用仍不可替代。

在目前的互联网中,传统网络仍然占据较大市场。本书的介绍依然以传统网络为主,关注 SDN 的读者可以参考相关专业书籍进行学习。

5.1.3 本章的主要协议

本章主要介绍以下协议。

(1) 互联网协议(Internet protocol,IP)。网络层核心协议以及传输层的 TCP、UDP 等协议的数据都通过 IP 数据报传输。

(2) 互联网控制报文协议(Internet control message protocol,ICMP)。它提供与网络配置信息和 IP 数据报处置相关的诊断和控制信息。

ICMP 报文直接封装在 IP 数据报内传输,从封装层次看与 TCP、UDP 等传输层协议一致,但是 ICMP 是 IP 的辅助协议,必须与 IP 一起实现。通常 ICMP 被认为是网络层协议,也有人将其归为 3.5 层协议,即网络层和传输层之间的协议。

(3) 路由协议(routing protocol)。它是路由器之间用来交换路由信息、链路状态信息或网络拓扑信息的协议,主要包括路由信息协议(routing information protocol,RIP)、开放最短通路优先(open shortest path first,OSPF)协议和边界网关协议(border gateway protocol,BGP)。

RIP 报文封装在 UDP 数据报中传输;OSPF 协议报文直接封装在 IP 数据报中传输;BGP 报文封装在 TCP 报文段中传输。从封装层次上看,RIP 和 BGP 与应用层协议一致,而 OSPF 与传输层协议一致。如果将路由选择看作一种特殊的应用,将路由器看作特殊的主机,路由协议可以归为应用层协议。

从功能上看,各路由协议都是为路由器控制平面中的路由选择算法提供数据支持的。由于路由选择功能属于网络层,本书将路由协议放在本章后面介绍。

(4) 多协议标记交换(multi-protocol label switching,MPLS)。它为 IP 等网络层协议提供面向连接的服务质量,支持流量工程(traffic engineering,TE)和负载均衡(load balance,LB),支持 MPLS VPN,在运营商和 ISP 中得到广泛应用。

MPLS 首部位于数据链路层首部和网络层首部之间,因此也可以将其归为 2.5 层协议。

与网络层 IP 相关的协议还有地址解析协议(address resolution protocol,ARP)。从封装层次上看,ARP 与 IP 一致,封装在数据链路层的数据帧内传输,因此 ARP 通常被归为网络层协议。但由于 ARP 的功能是将 IP 地址解析为数据链路层地址,与数据链路层关系紧密,故有人将其归为 2.5 层协议。学习 ARP 需要用到数据链路层相关的知识,本书将在第 6 章中介绍。

5.2 互联网协议

5.2.1 互联网协议概述

互联网协议(internet protocol,IP)是 TCP/IP 协议族中两个最重要的协议之一,是互联网的正式标准。IP 的协议数据单元通常称为 IP 分组或 IP 数据报。目前有两个版本的

IP正在使用，分别是IPv4和IPv6。IPv4由RFC791规定，在RFC2474、RFC3168和RFC6864等文档中做了更新。IPv6由RFC8200规定。本章介绍IPv4，IPv6的相关知识将在第7章中介绍。本书中，如未特别说明，“IP”均代表IPv4。

IP是为了实现网络互连才设计的协议。利用IP，可以实现在不同网络之间转发分组。当多个异构网络通过路由器互相连接起来后，IP屏蔽了底层网络的实现细节（如编址方案、协议格式等），向上层协议实体提供了统一的接口和服务。

如图5.5所示，路由器的每个接口连接一个网络，这些网络的实现各不相同，可以是无线局域网（wireless local area network，WLAN）、以太网（Ethernet）、点对点网络或移动网络等。采用IP后，路由器转发的都是统一的IP数据报，网络层地址都采用统一的IP地址，网络层之上的协议实体都无须再考虑具体网络的实现细节，每一个具体网络的细节由数据链路层协议实现。统一采用了IP的网络，也称为IP网络或简称IP网。

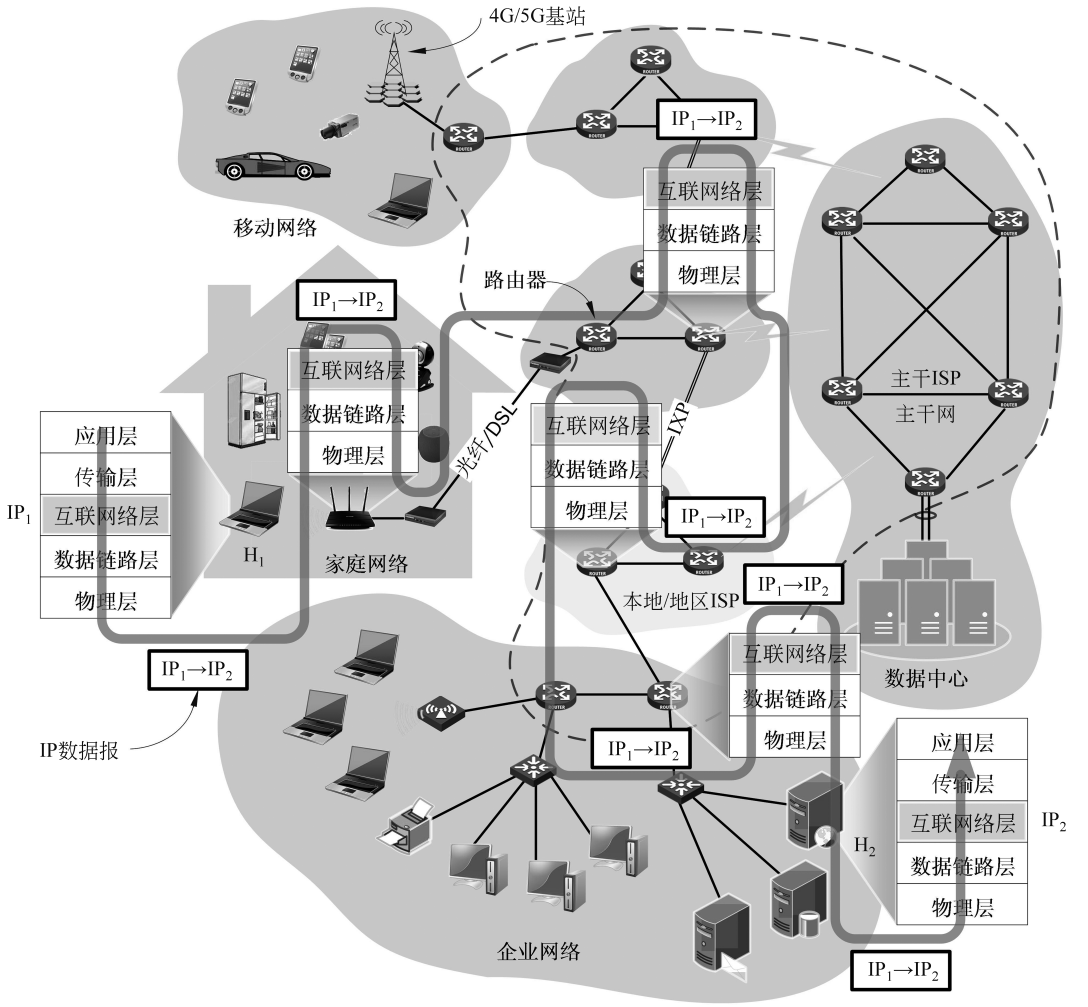


图5.5　IP数据报在互联网中的传送

在IP网络中，路由器的每个接口具有一个IP地址，这些IP地址属于不同的网络。在第2章已经介绍了IP地址的相关知识，已经知道IP地址目前是按照CIDR的方案进行编

址和管理的。在 CIDR 编址方案中，相同网络的含义为，在掩码作用下，具有相同网络前缀的 IP 地址属于相同网络。与之类似，在掩码作用下，具有不同网络前缀的 IP 地址属于不同网络。在本书后续的介绍中，如无特殊说明，“相同网络”和“不同网络”两个名词均指上述含义。

在图 5.5 中，源主机 H_1 产生的 IP 数据报，经过多个路由器的转发，最终到达目的主机 H_2。IP 数据报的源 IP 地址 IP_1 和目的 IP 地址 IP_2 在传送过程中均不发生变化（暂不考虑 NAT）。IP 地址唯一地标识了互联网上的一台主机，更确切地说，唯一地标识了该主机的一个网络接口。因此，IP 的作用范围是源主机的网络接口到目的主机的网络接口，如图 5.6 所示。

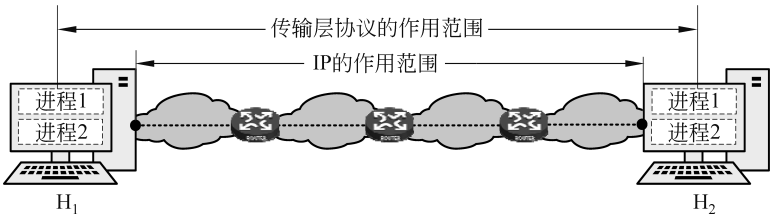


图 5.6　IP 的作用范围

IP 向上层仅提供简单灵活的、无连接的、尽力而为服务的数据报服务。在发送 IP 数据报时不需要先建立连接。每一个 IP 数据报独立发送，与其前后的 IP 数据报无关。IP 不提供服务质量的承诺。也就是说，所传送的 IP 数据报可能出错、丢失、重复和失序，当然也不保证 IP 数据报交付的时限。

5.2.2　IP 数据报格式

为了观察 IP 数据报，在 Linux 虚拟网络环境中构建如图 5.7 所示网络拓扑[①]，然后利用 Linux 的 nc 命令从主机 ns56A 向主机 ns57C 发起 UDP 通信[②]，观察 IP 数据报的首部格式。

执行如下 Linux 命令，利用 UDP，从主机 ns56A 向主机 ns57C 发送长度 3500B 的文件。

```
#ip netns exec ns57C nc -lvu 4499>3500.1
#ip netns exec ns56A nc -u 192.168.57.254 4499<3500.0
```

在主机 ns56A 上，启动 Wireshark 软件截获 IP 数据报如图 5.8 所示。3500B 数据经过 UDP 和 IP 的处理后，被分为 3 个 IP 数据报发送，Wireshark 为其编号 1～3。图 5.8 中显示的是 2 号 IP 数据报。

IP 数据报封装在数据链路层帧内部，由 IP 首部和数据部分组成。IP 首部长度为 20～60B，其中前 20B 是固定首部，其余是不超过 40B 的选项部分，如图 5.9 所示。

---

① 本实验网络拓扑的配置脚本可以参考本书配套电子资源。

② 执行命令前，需要关闭网卡 GSO 功能。网卡 GSO 功能允许将 IP 分片操作交给网卡执行，如果不关闭 GSO，用 Wireshark 截获数据时，在发送方不能观察到 IP 分片。

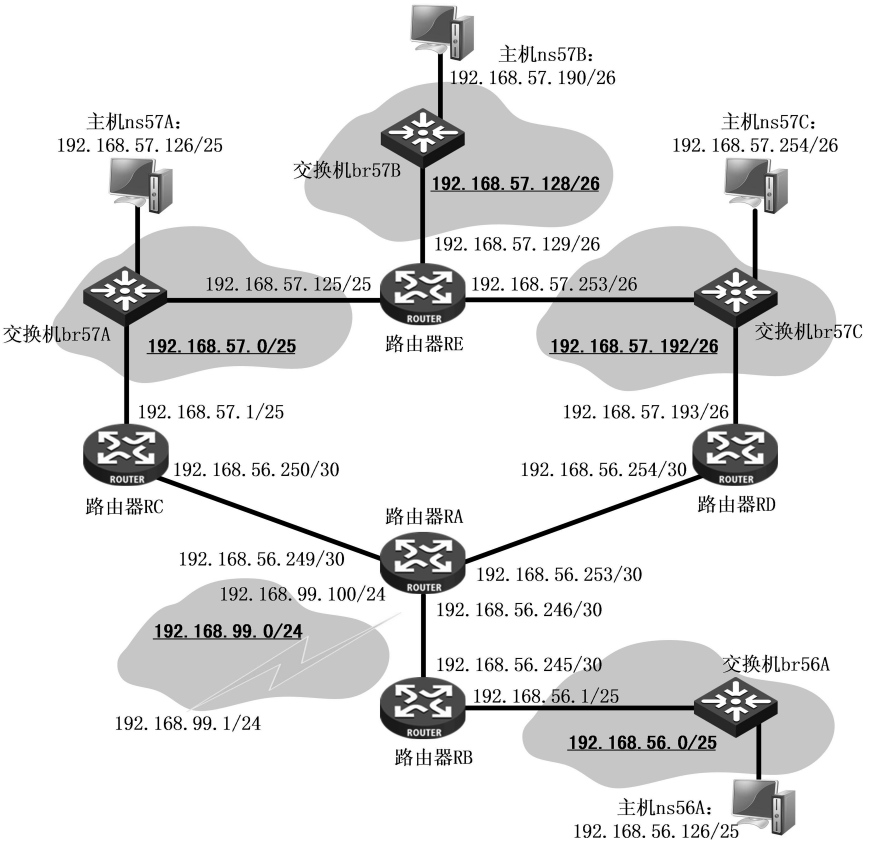


图 5.7　网络层通信实例拓扑图

No.	源IP	目的IP	总长度	标识	禁止分片	更多分片	片偏移	TTL
1	192.168.56.126	192.168.57.254	1500	0x8076 (32886)	Not set	Set	0	64
2	192.168.56.126	192.168.57.254	1500	0x8076 (32886)	Not set	Set	1480	64
3	192.168.56.126	192.168.57.254	568	0x8076 (32886)	Not set	Not set	2960	64

```
> Frame 2: 1514 bytes on wire (12112 bits), 1514 bytes captured (12112 bits) on interface tap56A
> Ethernet II, Src: da:8b:42:1c:a4:3d, Dst: 5e:52:40:50:47:88
v Internet Protocol Version 4
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
  v Differentiated Services Field: 0x00 (DSCP: CS0, ECN: Not-ECT)
      0000 00.. = Differentiated Services Codepoint: Default (0)
      .... ..00 = Explicit Congestion Notification: Not ECN-Capable Transport (0)
    Total Length: 1500
    Identification: 0x8076 (32886)
  v Flags: 0x20b9, More fragments
      0... .... .... .... = Reserved bit: Not set
      .0.. .... .... .... = Don't fragment: Not set
      ..1. .... .... .... = More fragments: Set
    Fragment offset: 1480
    Time to live: 64
    Protocol: UDP (17)
    Header checksum: 0xe014 [correct]
    [Header checksum status: Good]
    [Calculated Checksum: 0xe014]
    Source: 192.168.56.126
    Destination: 192.168.57.254
> Data (1480 bytes)
```

主机 ns56A：192.168.56.126
主机 ns57C：192.168.57.254

```
0000  5e 52 40 50 47 88 da 8b  42 1c a4 3d 08 00 45 00   ^R@PG··· B··=··E·
0010  05 dc 80 76 20 b9 40 11  e0 14 c0 a8 38 7e c0 a8   ···v ·@· ····8~··
0020  39 fe 20 20 20 20 20 20  20 20 20 20 20 20 20 20   9·
```

图 5.8　IP 数据报实例

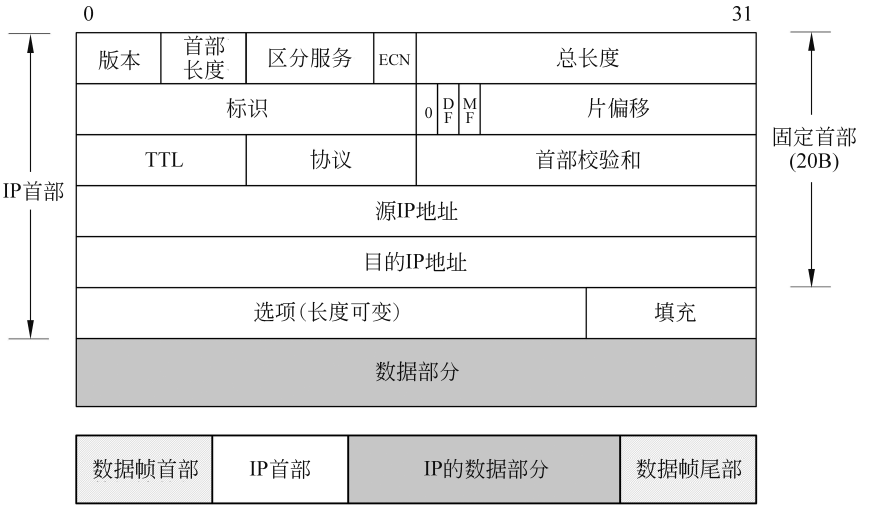


图 5.9　IP 数据报格式

IP 数据报的固定首部各字段含义如下。

1. 版本

版本字段指 IP 的版本号，占 4 位。

图 5.8 所示 IP 数据报实例的 2 号数据报中，版本字段值为 4，代表 IPv4。

2. 首部长度

首部长度字段指 IP 首部长度，占 4 位。首部长度以 4B(32 位)为单位，因此 IP 首部长度必须是 4B 的整数倍。首部长度最大取值为 15，对应 IP 首部最大长度为 60B，因此 IP 首部的选项部分长度不超过 40B。

如图 5.8 所示，在 IP 数据报实例的 2 号数据报中，首部长度字段值为 5，代表 IP 首部长度为 20B。这说明该 IP 数据报仅包含 20B 固定首部，未包含选项。

3. 区分服务

区分服务(differentiated services，DS)字段和紧随其后的 ECN 字段共占 8 位。这 8 位最初被 RFC791 定义为服务类型(type of service，ToS)字段，可以为 IP 数据报定义优先级，用来为 IP 提供差异化服务。IETF 在 RFC2474 中重新定义了服务类型字段，但仅使用了前 6 位，命名为区分服务(DS)字段。重新定义的 DS 字段与早期的服务类型字段能够部分兼容。

支持区分服务(DS)功能的结点称为 DS 结点。DS 结点根据 DS 字段的值来处理 IP 数据报的转发。因此，利用 DS 字段的不同值就可提供不同等级的服务质量。互联网建议标准 RFC2474 中将 DS 字段的取值称为区分服务码点(differentiated services codepoint，DSCP)，6 位 DS 字段可以定义 64 个 DSCP。

RFC2474 和 RFC8436 按照用途将 64 个 DSCP 分为 3 个池(pool)，如表 5.1 所示，表中 x 可取值 0 或 1。

DS 结点依据 DSCP 值对 IP 数据报采取的转发处理行为称为每跳行为(per-hop behavior，PHB)。RFC2474 要求对每跳行为的描述应该足够清晰。因具有公共约束条件，而能够同时实现的一组 PHB 被称为 PHB 组(PHB group)。

表 5.1　DSCP 用途

池	DSCP 空间	用　　途
1	xxxxx0	标准用途
2	xxxx11	实验或本地用途
3	xxxx01	标准用途①

① RFC2474 中规定将池 3 初始定义为实验或本地用途，当池 1 的空间耗尽时，转为标准用途。2018 年 RFC8436 发布时，虽然池 1 空间尚未耗尽，但是已经定义了 22 个标准 DSCP，因此将池 3 的用途转为标准用途。2019 年，RFC8622 定义了第一个属于池 3 的 DSCP。

1）默认 PHB

DSCP 的默认值为全“0”，该 DSCP 值对应的 PHB 称为默认 PHB（default PHB，DF PHB）。默认 PHB 采用常规的尽力而为服务（best effort）的 IP 数据报转发策略。

2）类别选择 PHB 组

RFC2474 规定，按照从高位到低位的顺序，DS 字段的第 0～2 位与早期服务类型字段中的优先级定义保持兼容，DS 字段中的第 3～5 位均为“0”的 DSCP 值称为类别选择码点（class selector codepoint），其对应的 PHB 称为类别选择 PHB（class selector PHB，CS PHB）组。CS PHB 的定义如表 5.2 所示。

表 5.2　DSCP 定义、服务类型和典型应用

服 务 类 型	DSCP 名称	DSCP 值	PHB 定义	典 型 应 用	是否应用 AQM
保留（未定义）	CS7	111000	RFC2474		
网络控制	CS6	110000	RFC2474	BGP 路由协议	是
电话语音（容量许可）	VOICE-ADMIT	101100	RFC5865	IP 电话语音	否
电话语音	EF	101110	RFC3246	IP 电话语音	否
电话信令	CS5	101000	RFC2474	IP 电话信令	否
多媒体会议	AF41 AF42 AF43	100010 100100 100110	RFC2597	视频会议	是 （每 PHB）
实时交互	CS4	100000	RFC2474	交互式游戏	否
多媒体流	AF31 AF32 AF33	011010 011100 011110	RFC2597	音频或视频点播	是 （每 PHB）
广播视频	CS3	011000	RFC2474	IPTV 广播	否
低延迟数据	AF21 AF22 AF23	010010 010100 010110	RFC2597	基于 Web 的客户-服务器应用	是 （每 PHB）
操作和维护	CS2	010000	RFC2474	网络维护	是