# 第1章 概述

## 1.1 计算机网络在信息时代的作用

**电信网络**：向用户提供电话、电报及传真等服务。

**有线电视网络**：向用户传送各种电视节目。

**计算机网络**：使用户能够在计算机之间传送数据文件。

互联网的两个基本特点：**连通性**和**共享性**。**连通性**就是互联网使上网用户之间，不管相距多远，都可以非常便捷、经济地交换各种信息，好像这些用户终端都彼此直接连通一样。**共享**可以是信息共享、软件共享和硬件共享。例如，互联网上有许多服务器存储了大量有价值的电子文档，可供上网的用户很方便地读取或下载。由于网络的存在，这些资源就在用户身边一样。

## 1.2 互联网概述

### 1.2.1 网络的网络

计算机网络由若干**结点**和连接这些结点的**链路**组成。网络中的结点可以是**计算机**、**集线器**、**交换机**或**路由器**等。网络之间还可以通过**路由器**互连起来，这就构成了一个覆盖范围更大的计算机网络。这样的网络称为**互连网**。因此互连网就是“**网络的网络**”。与网络相连的计算机常称为**主机**。

### 1.2.2 互联网基础结构发展的三个阶段

**第一个阶段**是**从单个网络ARPANET向互连网发展**的过程。1969年美国国防部创建的第一个分组交换网ARPANET最初只是一个单个的分组交换网（并不是一个互连的网络）。**所有要连接在ARPANET上的主机都直接与就近的结点交换机相连**。但到了20世纪70年代中期，人们已认识到不可能使用一个单独的网络来满足所有的通信问题。于是ARPA**开始研究多种网络互连的技术**，这就导致了互连网络的出现。这就成为现今互联网的雏形。**1983年TCP/IP协议成为ARPANET上的标准协议，使得所有使用TCP/IP协议的计算机都能利用互连网相互通信，因而人们就把1983年作为互联网的诞生时间**。1990年ARPANET正式宣布关闭，因为它的实验任务已经完成。

**internet（互连网）**：通用名词，泛指由多个计算机网络互连而成的计算机网络。在这些网络之间的通信协议可以任意选择，不一定非要使用TCP/IP协议。

**Internet（互联网）**：专用名词，指当前全球最大的、开放的、由众多网络相互连接而成的特定互连网，它采用TCP/IP协议族作为通信的规则，且其前身是美国的ARPANET。

**第二个阶段**的特点是**建成了三级结构的互联网**。从1985年起，美国国家科学基金会NSF就围绕六个大型计算机中心建设计算机网络，即国家科学基金网NSFNET。它是一个三级计算机网络，分为**主干网**、**地区网**和**校园网**（或**企业网**）。这种三级计算机网络覆盖了全美国主要的大学和研究所，并且成为互联网中的主要组成部分。1991年，NSF和美国的其他政府机构开始认识到，互联网必将扩大其使用范围，不应仅限于大学和研究机构。世界上的许多公司纷纷接入到互联网，网络上的通信量急剧增大，使互联网的容量已满足不了需要。于是美国政府决定将互联网的主干网转交给私人公司来经营，并开始对接入互联网的单位收费。

**第三个阶段**的特点是**逐渐形成了多层次ISP结构的互联网**。从1993年开始，由美国政府资助的NSFNET逐渐被若干个商用的互联网主干网替代，而政府机构不再负责互联网的运营。这样就出现了一个新的名词：互联网服务提供者ISP（Internet Service Provider）。例如，中国电信、中国联通和中国移动。

ISP可以从互联网管理机构申请到很多IP地址，同时拥有通信线路以及路由器等连网设备，因此任何结构和个人只要向某个ISP交纳规定的费用，就可从该ISP获取所需IP地址的使用权，并可通过该ISP接入到互联网。IP地址的管理机构不会把一个单个的IP地址分配给单个用户（**不“零售”IP地址**），而是把一批IP地址有偿租赁给经审查合格的ISP（**只“批发”IP地址**）。由此可见，**现在的互联网已不是某个单个组织所拥有而是全世界无数大大小小的ISP所共同拥有的**，这就是互联网也称为“网络的网络”的原因。



根据提供服务的覆盖面积大小以及所拥有的IP地址数目不同，ISP也分为不同层次的ISP：**主干ISP**、**地区ISP**和**本地ISP**。

主干ISP由几个专门的公司创建和维持，服务面积最大（国家范围），并且还拥有高速主干网。有一些地区ISP网络也可直接与主干ISP相连。

地区ISP是一些较小的ISP。这些地区ISP通过一个或多个主干ISP连接起来。它们位于等级中的第二层，数据率也低一些。

本地ISP给用户提供直接的服务。**本地ISP可以连接到地区ISP，也可直接连接到主干ISP**。绝大多数的用户都是连接到本地ISP的。本地ISP可以是一个仅仅提供互联网服务的公司，也可以是一个拥有网络并向自己的雇员提供服务的企业，或者是一个运行自己的网络的非营利机构。

从原理上讲，只要每一个本地ISP都安装了路由器连接到某个地区ISP，而每一个地区ISP也有路由器连接到主干ISP，那么在这些相互连接的ISP的共同合作下，就可以完成互联网中的所有的分组转发任务。但随着互联网上数据流量的急剧增长，人们开始研究如何更快地转发分组，以及如何更加有效地利用网络资源。于是，**互联网交换点IXP**（Internet eXchange Point）就应运而生了。

互联网交换点IXP的主要作用就是**允许两个网络直接相连并交换分组，而不需要再通过第三个网络来转发分组**。这样就使互联网上的数据流量分布更加合理，同时也减少了分组转发的迟延时间，降低了分组转发的费用。现在许多IXP在进行对等交换分组时，都互相不收费。但本地ISP或地区ISP通过IXP向高层的IXP转发分组时，则需要交纳一定的费用。IXP的结构非常复杂。典型的IXP由一个或多个网络交换机组成，许多ISP再连接到这些网络交换机的相关端口上。IXP常采用工作在数据链路层的网络交换机，这些网络交换机都用局域网互连起来。

### 1.2.3 互联网的标准化工作

所有的互联网标准都是以RFC（Request For Comments）的形式在互联网上发表的。但并非所有的RFC文档都是互联网标准。只有很少部分的RFC文档最后能变成互联网标准。RFC文档按发表时间的先后编上序号（即RFC xxxx，这里的xxxx是阿拉伯数字）。一个RFC文档更新后就使用一个新的编号，并在文档中指出原来老编号的RFC文档已成为陈旧的或被更新，但陈旧的RFC文档并不会被删除，而是永远保留着，供用户参考。

制定互联网的正式标准要经过以下三个阶段：

（1）**互联网草案**——互联网草案的有效期只有六个月，在这个阶段还不能算是RFC文档。

（2）**建议标准**——从这个阶段开始就成为RFC文档。

（3）**互联网标准**——达到正式标准后，每个标准就分配到一个编号STD xx。一个标准可以和多个RFC文档关联。

## 1.3 互联网的组成

**边缘部分**：由所有连接在互联网上的主机（端系统）组成。这部分是**用户直接使用**的，用来进行通信和资源共享。

**核心部分**：由大量网络和连接这些网络的路由器组成。这部分是**为边缘部分提供服务**的。

在互联网核心部分的路由器之间一般用高速链路相连接，而在网络边缘的主机接入到核心部分则以相对较低速率的链路相连接。

### 1.3.1 互联网的边缘部分

**计算机之间通信**：主机A的某个进程和主机B上的另一个进程进行通信。

在网络边缘的端系统之间的通信方式通常可划分为两大类：**客户-服务器**（C/S）方式和**对等**（P2P）方式。

#### 1. 客户-服务器方式

**客户程序**：

（1）被用户调用后运行，在通信时主动向远地服务器发起通信。客户程序必须知道服务器程序的地址。

（2）不需要特殊的硬件和很复杂的操作系统。

**服务器程序**：

（1）是一种专门用来提供某种服务的程序，可**同时处理**多个远地或本地客户的请求。

（2）系统启动后即自动调用并一直不断地运行着，被动地等待并接受来自各地的客户的通信请求。服务器程序不需要知道客户程序的地址。

（3）一般需要有强大的硬件和高级的操作系统支持。

#### 2. 对等连接方式

在通信时不区分哪一个是服务请求方哪一个是服务提供方。只要两台主机都运行了对等连接软件，它们就可以进行平等的、对等连接通信。实际上，对等连接方式从本质上看仍然是客户-服务器方式，只是对等连接中的每一台主机既是客户又同时是服务器。

### 1.3.2 互联网的核心部分

**路由器**是一种专用计算机（但不叫主机），是实现**分组交换**的关键构件，其任务是转发收到的分组，这是网络核心部分最重要的功能。路由器收到一个分组，先暂时存储一下（存储在**内存**中），检查其首部，查找转发表，按照首部中的目的地址，找到合适的接口转发出去，把分组交给下一个路由器。这样一步一步地以存储转发的方式，把分组交付最终的目的主机。**各路由器之间必须经常交换彼此掌握的路由信息，以便创建和动态维护路由器中的转发表**。

#### 1. 电路交换的主要特点

从通信资源的分配角度来看，交换就是按照某种方式动态地分配传输线路的资源。在使用电路交换通话之前，必须先拨号请求建立连接。当被叫用户听到交换机送来的振铃音并摘机后，从主叫端到被叫端就建立了一条连接，也就是专用的物理通路。这条连接保证了双方通话时所需的通信资源，而这些资源在双方通信时不会被其他用户占用。此后主叫和被叫双方就能互相通电话。通话完毕挂机后，交换机释放刚才使用的这条专用的物理通路。这种必须经过“**建立连接**（占用通信资源）→**通话**（一直占用通信资源）→**释放连接**（归还通信资源）”三个步骤的交换方式称为**电路交换**。电路交换的一个重要特点就是**在通话的全部时间内，通话的两个用户始终占用端到端的通信资源**。

当使用电路交换来传送计算机数据时，其线路的传输效率往往很低。这是因为计算机数据时突发式地出现在传输线路上的，因此线路上真正用来传送数据的时间往往不到10%。已被用户占用的通信线路资源在绝大部分时间里都是空闲的。

#### 2. 分组交换的主要特点



分组交换采用**存储转发**技术。通常把要发送的整块数据称为一个**报文**。在发送报文之前，先把较长的报文划分成为一个个更小的等长数据段。在每一个数据前面，加上一些由必要的控制信息组成的**首部**后，就构成了一个**分组**。分组又称为“**包**”，而分组的首部也可称为“**包头**”。分组是在互联网中传送的数据单元。**分组的首部包含了诸如目的地址和源地址等重要控制信息**。

分组交换在传送数据之前不必先占用一条端到端的链路的通信资源。分组在哪段链路上传送才占用这段链路的通信资源。分组到达一个路由器后，先暂时存储下来，查找转发表，然后从一条合适的链路转发出去。分组在传输时就这样一段一段地断续占用通信资源，而且还省去了建立连接和释放连接的开销，因而数据的传输效率更高。

分组交换的缺点：

（1）分组在各**路由器存储转发时需要排队**，会造成一定的时延。

（2）不像电路交换那样通过建立连接来保证通信时所需的各种资源，**无法保证通信时端到端的带宽**。

（3）**各分组必须携带的控制信息**造成一定的开销。

（4）**整个分组交换网需要专门的管理和控制机制**。

* 能否说“电路交换就是面向连接，分组交换就是面向无连接”？

不能。

电路交换是在A和B通信开始前，必须先建立一条从A到B的连接。当A到B的连接建立后，通信就沿着这条路径进行。A和B在通信期间始终占用这条信道。通信完毕时就释放所占用的信道，即断开连接。因此电路交换是面向连接的。

分组交换也可以是面向连接的。例如X.25网络、帧中继网络或ATM网络都属于分组交换网。然而这种面向连接的分组交换网在传送用户数据之前必须先建立连接。数据传送完毕后必须释放连接。

因此**使用面向连接的服务可以是电路交换，也可以是分组交换**。电路交换肯定是面向连接的，但面向连接的也可以是分组交换。

现在的互联网所使用的分组交换采用IP协议，IP协议使用无连接的IP数据报来传送数据，即不需要先建立连接就可以立即发送数据。当数据发送完毕后也不存在释放连接的问题。

面向连接和无连接往往可以在不同的层次上来讨论。例如，在数据链路层，**HDLC和PPP协议是面向连接的**，而**以太网使用的CSMA/CD是无连接的**。在网络层，**X.25协议是面向连接的**，而**IP协议是无连接的**。在运输层，**TCP是面向连接的**，而**UDP是无连接的**。

* 试在下列条件下比较电路交换和分组交换。要传送的报文共x（bit）。从源点到终点共经过k段链路，每段链路的传播时延为d（s），数据率为b（bit/s）。在电路交换时电路的建立时间为s（s）。在分组交换时分组长度为p（bit），且各结点的排队等待时间可忽略不计。分别计算电路交换时延和分组交换时延。

电路交换必须先建立连接，需要的时间是s秒。

发送x比特的报文所需的时间是报文长度除以数据率b，因此发送时延是x/b。

总的传播时延是链路数乘以每段链路的传播时延，即kd。

因此，电路交换的时延由以下三项组成：s + x/b + kd。



分组交换不需要先建立连接（这里假定题目中的分组交换使用数据报传送。如果使用虚电路传送，则需要先建立连接）。

先计算分组交换的传播时延，这和电路交换是一样的，也是kd。

再计算n个分组所需的发送时延，n为，故该时延为。

在一段链路上发送一个分组的发送时延是p/b，k-1段链路的发送时延是。

因此，总时延为。

## 1.4 计算机网络在我国的发展

## 1.5 计算机网络的类别

### 1.5.1 计算机网络的定义

#### 1. 按照网络的作用范围进行分类

**广域网WAN**（Wide Area Network）、**城域网MAN**（Metropolitan Area Network）、**局域网LAN**（Local Area Network）、**个人区域网PAN**（Personal Area Network）

#### 2. 按照网络的使用者进行分类

**公用网**、**专用网**

#### 3. 用来把用户接入到互联网的网络

**接入网AN**（Access Network），又称为**本地接入网**或**居民接入网**。接入网本身既不属于互联网的核心部分，也不属于互联网的边缘部分。接入网是**从某个用户端系统到互联网中的第一个路由器之间的一种网络**。在互联网发展初期，用户多用电话线拨号接入到互联网，速率很低，因此那时并没有使用接入网这个名词。直到最近，由于出现了多种宽带接入技术，宽带接入网才成为互联网领域中的一个热门课题。

## 1.6 计算机网络的性能

### 1.6.1 计算机网络的性能指标

#### 1. 速率

一个**比特**就是二进制数字中的一个1或0。网络技术中的**速率**指的是数据的传送速率，它也称为**数据率**或**比特率**，单位是**比特每秒**。

#### 2. 带宽

在计算机网络中，带宽用来表示网络中某通道传送数据的能力，因此网络带宽表示**在单位时间内网络中的某信道所能通过的“最高数据率”**，单位是比特每秒。

#### 3. 吞吐量

吞吐量表示在单位时间内通过某个网络的**实际**的数据量，单位是比特每秒。

#### 4. 时延

**时延**是指数据从网络的一端传送到另一端所需的时间。

（1）**发送时延**

发送时延是主机或路由器发送数据帧所需要的时间，也就是从发送数据帧的第一个比特算起，到该帧的最后一个比特发送完毕所需的时间。

（2）**传播时延**

传播时延是电磁波在信道中传播一定的距离需要的时间。

**发送时延发生在机器内部的发送器（网络适配器）中，与传输信道的长度没有任何关系。传播时延发生在机器外部的传输信道媒体上，而与信号的发送速率无关。信号传送的距离越远，传播时延就越大。**

（3）**处理时延**

主机或路由器在收到分组时要花费一定的时间进行处理。

（4）**排队时延**

**分组在进入路由器后要先在输入队列中排队等待处理**。**在路由器确定了转发接口后，还要在输出队列中排队等待转发**。排队时延的长短往往取决于网络当时的通信量。当网络的通信量很大时会发生队列溢出，使分组丢失，这相当于排队时延为无穷大。

“**光纤信道的传输速率高**”是指可以用很高的速率向光纤信道发送数据，而光纤信道的传播速率实际上还要比铜线的传播速率略低一点。

#### 5. 时延带宽积（以比特为单位的链路长度）

#### 6. 往返时间RTT

#### 7. 利用率

**信道利用率**指出某信道有百分之几的时间是被利用的（有数据通过）。

**网络利用率**是全网络的信道利用率的加权平均值。

**信道或网络的利用率过高会产生非常大的时延**。

### 1.6.2 计算机网络的非性能特征

#### 1. 费用

#### 2. 质量

#### 3. 标准化

#### 4. 可靠性

#### 5. 可扩展性和可升级性

#### 6. 易于管理和维护

## 1.7 计算机网络体系结构

### 1.7.1 计算机网络体系结构的形成

**开放系统互连基本参考模型OSI/RM**（Open System Interconnection Reference Model）是法律上的国际标准，TCP/IP是事实上的国际标准。

### 1.7.2 协议与划分层次

网络协议三要素：（1）**语法**，即**数据与控制信息的结构或格式**；（2）**语义**，即**需要发出何种控制信息，完成何种动作以及做出何种响应**；（3）**同步**，即**事件实现顺序的详细说明**。

### 1.7.3 具有五层协议的体系结构



#### （1）应用层

应用层的任务是**通过应用进程间的交互来完成特定网络应用**。应用层协议定义的是**应用进程间通信和交互的规则**。不同的网络应用需要有不同的应用层协议，如域名系统DNS，支持万维网应用的HTTP协议，支持电子邮件的SMTP协议等。应用层交互的数据单元称为**报文**。

#### （2）运输层

运输层的任务就是**负责向两台主机中进程之间的通信提供通用的数据传输服务**。应用进程利用该服务传送应用层报文。所谓“通用的”，是指并不针对某个特定网络应用，而是多种应用可以使用同一个运输层服务。由于一台主机可同时运行多个进程，因此运输层有复用和分用的功能。**复用**就是**多个应用层进程可同时使用下面运输层的服务**，**分用**是**运输层把收到的信息分别交付上面应用层中的相应进程**。

运输层主要使用以下两种协议：

* **传输控制协议TCP**（Transmission Control Protocol）——提供**面向连接的**、**可靠的**数据传输服务，其数据传输的单位是**报文段**。
* **用户数据报协议UDP**（User Datagram Protocol）——提供**无连接的**、**尽最大努力的**数据传输服务（不保证数据传输的可靠性），其数据传输的单位是**用户数据报**。
* 在运输层应根据什么原则来确定使用TCP还是UDP？

根据上层应用程序的性质来确定使用哪种连接服务。

例如，在传送**文件**时要使用文件传送协议FTP，而文件的传送必须是可靠的，因此在运输层就必须使用面向连接的TCP协议。但若应用程序要传送**分组话音**或**视频点播信息**，那么为了保证信息传输的实时性，在运输层就必须使用无连接的UDP协议。

另外，选择TCP或UDP时还需考虑对连接资源的控制。若应用程序不希望在服务器端同时建立太多的TCP连接，可考虑使用UDP。

#### （3）网络层

网络层负责**为分组交换网上的不同主机提供通信服务**。在发送数据时，网络层把运输层产生的报文段或用户数据报封装成**分组**或**包**进行传送。在TCP/IP体系中，由于网络层使用IP协议，因此分组也叫做**IP数据报**。

网络层的另一个任务就是要**选择合适的路由**，使源主机运输层所传下来的分组，能通过网络中的路由器找到目的主机。

互联网是由大量的**异构网络**通过**路由器**连接起来的。互联网使用的网络层协议是无连接的网际协议IP和许多路由选择协议，因此互联网的网络层也叫做**网际层**或**IP层**。

#### （4）数据链路层

两台主机之间的数据传输是在一段一段的链路上传送的，这就需要使用专门的链路层的协议。在两个相邻结点之间传送数据时，数据链路层将网络层交下来的IP数据报组装成**帧**，在两个相邻结点间的链路上传送帧。每一帧包括数据和必要的控制信息（同步信息、地址信息、差错控制等）。

在接收数据时，控制信息使**接收端能够知道一个帧从哪个比特开始和到哪个比特结束**。这样，数据链路层在收到一个帧后，就可从中提取出数据部分，上交给网络层。

控制信息还使**接收端能够检测到所收到的帧中有无差错**。如发现有差错，数据链路就简单地**丢弃**这个出现了差错的帧，以免继续在网络中传送下去白白浪费网络资源。如果需要改正数据在数据链路层传输时出现的差错，那么就要采用可靠传输协议来纠正出现的差错，这会使数据链路层的协议更复杂。

#### （5）物理层

在物理层上所传数据的单位是比特。物理层要考虑用多大的电压代表“1”和“0”，以及接收方如何识别出发送方所发送的比特。物理层还要确定连接电缆的插头应当有多少根引脚以及各引脚应如何连接。解释比特的意思不是物理层的任务。传递信息所利用的一些物理媒体，如双绞线、光缆、无线信道等不在物理层协议之内而是在物理层协议的下面。

**协议数据单元PDU**（Protocol Data Unit）指对等层次之间传送的数据单位。

**服务数据单元SDU**（Service Data Unit）指层与层之间交换的数据单位。

### 1.7.4 实体、协议、服务和服务访问点

在协议的控制下，两个对等实体间的通信使得本层能够向上一层提供服务。要实现本层协议，还需要使用下面一层所提供的服务。上层使用下层所提供的服务必须通过与下层交换一些命令，这些命令在OSI中称为**服务原语**。在同一系统中相邻两层的实体进行交互的地方称为**服务访问点**。

### 1.7.5 TCP/IP体系结构



路由器在转发分组时**最高只用到网络层**而没有使用运输层和应用层。



实际上现在某些应用程序可以直接使用IP层，或甚至直接使用最下面的网络接口层。



TCP/IP协议可以为各式各样的应用提供服务（everything over IP），同时TCP/IP协议允许IP协议在各式各样的网络构成的互联网上运行（IP over everything）。

# 第2章 物理层

## 2.1 物理层的基本概念

物理层考虑的是**怎样才能在连接各种计算机的传输媒体上传输数据比特流**，而不是指具体的传输媒体。在传输媒体中传输的是信号，但传输媒体并不知道所传输的信号代表什么意思。也就是说，传输媒体不知道所传输的信号什么时候是1什么时候是0。现有的计算机网络中的硬件设备和传输媒体的种类非常多，物理层的作用是要尽可能地屏蔽掉这些传输媒体和通信手段的差异，使物理层上面的数据链路层感觉不到这些差异，这样就可使数据链路层只需要考虑如何完成本层的协议和服务，而不必考虑网络具体的传输媒体和通信手段是什么。

可以将物理层的主要任务描述为确定与传输媒体的接口有关的一些特性，即：

（1）**机械特性**——指明**接口所用接线器的形状和尺寸、引脚数目和排列、固定和锁定装置**等。

（2）**电气特性**——指明**在接口电缆的各条线上出现的电压的范围**。

（3）**功能特性**——指明**某条线上出现的某一电平的电压的意义**。

（4）**过程特性**——指明**对于不同功能的各种可能事件的出现顺序**。

数据在计算机内部所采用并行传输方式，但数据在通信线路（传输媒体）上的传输方式一般都是串行传输，即逐个比特按照时间顺序传输，因此物理层还要完成传输方式的转换。

## 2.2 数据通信的基础知识

### 2.2.1 数据通信系统的模型

一个数据通信系统可划分为**源系统**、**传输系统**和**目的系统**三大部分。

源系统一般包括以下两个部分：

（1）**源点**——产生要传输的数据，例如，从计算机的键盘输入汉字，计算机产生输出的数字比特流。

（2）**发送器**——源点生成的数字比特流要通过发送器编码后才能够在传输系统中进行传输。典型的发送器是**调制器**。现在很多计算机使用内置的**调制解调器**（包含调制器和解调器），用户在计算机外面看不见调制解调器。

目的系统一般包括以下两个部分：

（1）**接收器**——接收传输系统传送过来的信号，并把它转换为能够被目的设备处理的信息。典型的接收器就是**解调器**，它把来自传输线路上的模拟信号进行解调，提取出在发送端置入的消息，还原出发送端产生的数字比特流。

（2）**终点**——终点设备从接收器获取传送来的数字比特流，然后把信息输出，例如，把汉字在计算机屏幕上显示出来。

### 2.2.2 有关信道的几个基本概念

**信道**表示向某一个方向传送信息的媒体。**一条通信电路往往包含一条发送信道和一条接收信道**。

从通信的双方信息交互的方式来看，可以有以下三种基本方式：

（1）**单向通信**（单工通信）——只能有一个方向的通信而没有反方向的交互，如无线电广播、有线电广播和电视广播。

（2）**双向交替通信**（半双工通信）——通信的双方都可以发送信息，但不能双方同时发送。

（3）**双向同时通信**（全双工通信）——通信的双方可以同时发送和接收信息。

**单向通信只需要一条信道，而双向交替通信或双向同时通信都需要两条信道**。

来自信源的信号常称为**基带信号**，如计算机输出的代表各种文字或图像文件的数据信号。**基带信号往往包含有较多的低频成分，甚至有直流成分，而许多信道并不能传输这种低频分量或直流分量**。为了解决这一问题，必须对基带信号进行**调制**。

调制可分为两大类。一类是仅仅**对基带信号的波形进行变换**，使它能够与信道特性相适应。变换后的信号仍然是基带信号。这类调制称为**基带调制**。由于这种基带调制是把数字信号转换为另一种形式的数字信号，因此这种过程又称为**编码**。另一类调制则需要使用**载波**进行调制，把基带信号的频率范围搬移到较高的频段，并转换为模拟信号，这样就能够更好地在模拟信道中传输。经过载波调制后的信号称为**带通信号**（仅在一段频率范围内能够通过信道），而使用载波的调制称为**带通调制**。

### 2.2.3 信道的极限容量

## 2.3 物理层下面的传输媒体

在**导引型传输媒体**中，电磁波被导引沿着固体媒体（铜线或光纤）传播，而**非导引型传输媒体**就是指自由空间，在非导引型传输媒体中电磁波的传输常称为无线传输。

### 2.3.1 导引型传输媒体

#### 1. 双绞线

#### 2. 同轴电缆

#### 3. 光缆

### 2.3.2 非导引型传输媒体

## 2.4 信道复用技术

### 2.4.1 频分复用、时分复用和统计时分复用

**频分复用**的所有用户**在同样的时间占用不同的带宽资源**。

**时分复用**（Time Division Multiplexing）的所有用户**在不同的时间占用同样的频带宽度**。

当使用时分复用系统传送计算机数据时，由于计算机数据的突发性质，一个用户对已经分配到的子信道的利用率一般是不高的。当用户在某一段时间暂时无数据传输时，那就只能让已经分配到手的子信道空闲着，而其他用户也无法使用这个暂时空闲的线路资源。

**统计时分复用**（Statistic TDM）是一种改进的时分复用，它能明显地提高信道的利用率。集中器常使用这种统计时分复用。

统计时分复用使用STDM帧来传送复用的数据。但**每一个STDM帧中的时隙数小于连接在集中器上的用户数**。各用户有了数据就随时发往集中器的输入缓存，然后集中器按顺序依次扫描输入缓存，把缓存中的输入数据放入STDM帧中。对没有数据的缓存就跳过去。**当一个帧的数据放满了，就发送出去**。因此，STDM帧不是固定分配时隙，而是按需动态地分配时隙。因此统计时分复用可以提高线路的利用率。在输出线路上，某一个用户所占用的时隙不是周期性地出现，因此统计时分复用又称为**异步时分复用**，而普通的时分复用称为**同步时分复用**。虽然统计时分复用的输出线路上的数据率小于各输入线路数据率的总和，但从平均的角度来看，这二者是平衡的。假定所有的用户都不间断地向集中器发送数据，那么集中器肯定无法应付，它内部设置的缓存都将溢出。所以集中器能够正常工作的前提是假定各用户都是间歇地工作。



由于STDM帧中的时隙并不是固定地分配给某个用户，因此**在每个时隙中还必须有用户的地址信息**，这是统计时分复用必须要有的和不可避免的一些开销。上图中输出线路上每个时隙之前的短时隙（白色）就是放入这样的地址信息。使用统计时分复用的集中器也叫做**智能复用器**，它能**提供对整个报文的存储转发能力**（但大多数复用器一次只能存储一个字符或一个比特），通过排队方式使各用户更合理地共享信道。此外，许多集中器还可能具有路由选择、数据压缩、前向纠错等功能。

STDM“帧”和数据链路层的“帧”是完全不同的概念。

### 2.4.2 波分复用

波分复用就是光的频分复用。

### 2.4.3 码分复用

**码分复用CDM**（Code Division Multiplexing）又叫**码分多址CDMA**（Code Division Multiple Access），每一个用户可以在同样的时间使用同样的频带进行通信。**由于各用户使用经过特殊挑选的不同码型，因此各用户之间不会造成干扰**。

## 2.5 数字传输系统

## 2.6 宽带接入技术

### 2.6.1 ADSL技术

**非对称数字用户线**ADSL（Asymmetric Digital Subscriber Line）技术是**用数字技术对现有的模拟电话用户线进行改造**，使它能够承载宽带数字业务。虽然标准模拟电话信号的频带被限制在300~3400Hz的范围内，但用户线本身实际可通过的信号频率却超过1MHz。ADSL技术把0~4kHz低端频谱留给传统电话使用，而把原来没有被利用的高端频谱留给用户上网使用。由于用户在上网时主要是从互联网下载各种文档，而向互联网发送的信息量一般都不太大，因此**ADSL的下行（从ISP到用户）带宽都远远大于上行（从用户到ISP）带宽**。“非对称”这个名词就是这样得出的。

ADSL在用户线（铜线）的两端各安装一个**ADSL调制解调器**。这种调制解调器的实现方案有许多种。我国目前采用的方案是**离散多音调**DMT（Discrete Multi-Tone）调制技术。这里的“多音调”就是“多载波”或“多子信道”的意思。DMT调制技术采用频分复用的方法，把40kHz以上一直到1.1MHz的高端频谱划分为许多子信道，其中**25个子信道用于上行信道，而249个子信道用于下行信道**，并使用不同的载波（即不同的音调）进行数字调制。这种做法相当于在一对用户线上使用许多小的调制解调器并行地传送数据。

### 2.6.2 光纤同轴混合（HFC，Hybrid Fiber Coex）网

**光纤同轴混合网**是在目前覆盖面很广的**有线电视网**的基础上开发的一种居民宽带接入网，除可传送电视节目外，还能提供电话、数据和其他宽带交互型业务。最早的有线电视网是树形拓扑结构的同轴电缆网络，它采用模拟技术的频分复用对电视节目进行单向广播传输。但以后有线电视网进行了改造，变成了现在的光纤同轴混合网。

### 2.6.3 FTTx技术

**光纤到户**FTTH（Fiber To The Home）就是把光纤一直铺设到用户家庭，只有在光纤进入用户的家门后，才把光信号转换为电信号，这样做就可以使用户获得最高的上网速率。

但光纤到户FTTH有两个问题：首先是目前的价格还不够便宜；其次是一般的家庭用户也并没有这样高的数据率的需求。在这种情况下，就出现了多种宽带光纤接入方式，称为FTTx，表示Fiber To The…。这里字母x可代表不同的光纤接入地点。实际上，FFTx就是把光电转换的地方，从用户家中（这时x就是H）向外延伸到离用户家门口有一定距离的地方。

# 第3章 数据链路层

数据链路层使用的信道主要有以下两种类型：

（1）**点对点信道**。这种信道使用一对一的点对点通信方式。

（2）**广播信道**。这种信道使用一对多的广播通信方式。广播信道上连接的主机很多，因此必须使用专用的共享信道协议来协调这些主机的数据发送。

数据链路层的三个基本问题：**封装成帧**、**透明传输**和**差错检测**。

## 3.1 使用点对点信道的数据链路层

### 3.1.1 数据链路和帧

**链路**是从一个结点到相邻结点的一段物理链路（有线或无线），而中间没有任何其他的交换结点。

当需要在一条线路上发送数据时，除了必须有一条物理线路外，还必须有一些必要的通信协议来控制这些数据的传输。若把实现这些协议的硬件和软件加到链路上，就构成了**数据链路**。现在最常用的方法是使用**网络适配器**（既有硬件，也包括软件）来实现这些协议。**一般的适配器都包括了数据链路层和物理层这两层的功能**。

数据链路层把网络层交下来的数据构成**帧**发送到链路上，以及把接收到的帧中的数据取出并上交给网络层。

点对点信道的数据链路层在进行通信时的主要步骤如下：

（1）结点A的数据链路层把网络层交下来的IP数据报添加**首部**和**尾部**封装成帧。

（2）结点A把封装好的帧发送给结点B的数据链路层。

（3）若结点B的数据链路层收到的帧无差错，则从收到的帧中提取出IP数据报交给上面的网络层；否则丢弃这个帧。

### 3.1.2 三个基本问题

#### 1. 封装成帧



所有在互联网上传送的数据都以分组（IP数据报）为传送单位。网络层的IP数据报传送到数据链路层就成为帧的数据部分。在帧的数据部分的前面和后面分别添加上首部和尾部，构成了一个完整的帧。这样的帧就是数据链路层的数据传送单元。**一个帧的帧长等于帧的数据部分长度加上帧首部和帧尾部的长度**。首部和尾部的一个重要作用就是进行**帧定界**。此外，首部和尾部还包括许多必要的控制信息。**在发送时，是从帧首部开始发送的**。各种数据链路层协议都对帧首部和帧尾部的格式有明确的规定。**为了提高帧的传输效率，应当使帧的数据部分长度尽可能地大于首部和尾部的长度，但是，每一种链路层协议都规定了所能传送的帧的数据部分长度上限**——**最大传送单元MTU**。

* 如果在数据链路层不进行封装成帧，会发生什么问题？

如果在数据链路层不进行封装成帧，那么数据链路层在收到一些数据时，就无法知道对方传送的数据中哪些是数据，哪些是控制信息，甚至数据中有没有差错也不清楚（因为无法进行差错检测）。数据链路层也无法知道数据传送结束了没有，因此不知道应当在什么时候把收到的数据交给上一层。

#### 2. 透明传输

由于帧的开始和结束的标记使用专门指明的控制字符，因此，所传输的数据中的任何8比特的组合一定不允许和用作帧定界的控制字符的比特编码一样，否则就会出现帧定界的错误。

“在数据链路层透明传送数据”表示**无论什么样的比特组合的数据，都能够按照原样没有差错地通过这个数据链路层**。

#### 3. 差错检测

**比特差错**：比特在传输过程中1变成0或0变成1。传输错误的比特占所传输比特总数的比率称为**误码率**BER（Bit Error Rate）。

**循环冗余检验**CRC（Cyclic Redundancy Check）的原理：

在发送端，先把数据划分为组，假定每组k个比特。现假定待传送的数据M=101001（k=6）。CRC运算就是在数据M的后面添加供差错检测用的n位冗余码，然后构成一个帧发送出去，一共发送（k+n）位。

这n位冗余码可用以下方法得出。用二进制的**模2运算**进行2n乘M的运算，这相当于在M后面添加n个0得到的（k+n）位的数除以收发双方事先商定的长度为（n+1）位的除数P，得出商是Q而余数是R（n位，比P少一位）。这个余数R就作为冗余码拼接在数据M的后面发送出去。这种为了进行检错而添加的冗余码常称为**帧检验序列**FCS（Frame Check Sequence）。

在接收端把接收到的数据以帧为单位进行CRC检验：把收到的每一个帧都除以同样的除数P（**模2运算**），然后检查得到的余数R。如果在传输过程中无差错，那么经过CRC检验后得出的余数R肯定是0。

* **模2运算进行加法时不进位，例如，1111+1010=0101。减法和加法一样，按加法规则计算**。

循环冗余检验CRC（Cyclic Redundancy Check）可以做到对帧的无差错接受，即“**无比特差错**”，但这还不是可靠传输，因此无法做到“**无传输差错**”，即无法避免**帧丢失**、**帧重复**和**帧失序**。

对于通信质量良好的有线传输链路，数据链路层协议不使用**确认和重传机制**，即不要求数据链路层向上提供可靠传输的服务。如果在数据链路层传输数据时出现了差错并且需要进行改正，那么改正差错的任务就由上层协议（如TCP）来完成。

对于通信质量较差的无线传输链路，数据链路层协议使用确认和重传机制，数据链路层向上提供可靠传输的服务。

* 当数据链路层使用PPP协议或CSMA/CD协议时，既然不保证可靠传输，那么为什么对所传输的帧进行差错检验呢？

如果在接收端不进行差错检测，那么接收端上交给主机的帧就可能包括在传输中出了差错的帧，这样的帧对接收端主机是没有用处的。

## 3.2 点对点协议PPP

### 3.2.1 PPP协议的特点

PPP协议是用户计算机和ISP进行通信时所使用的数据链路层协议。

#### 1. PPP协议应满足的需求

（1）简单。**接收方每接收一个帧，就进行CRC检验。如CRC检验正确，就收下这个帧；反之，就丢弃这个帧，其他什么也不做**。

（2）封装成帧。PPP协议必须规定特殊的字符作为帧定界符。

（3）透明性。如果数据中碰巧出现了和帧定界符一样的比特组合，就要采取有效的措施来解决这个问题。

（4）多种网络层协议。**PPP协议必须能够在同一条物理链路上同时支持多种网络层协议的运行**。

（5）多种类型链路。**PPP协议必须能够在多种类型的链路上运行**。

（6）差错检测。

（7）检测连接状态。**PPP协议必须具有一种机制能够及时自动检测出链路是否处于正常工作状态**。

（8）最大传送单元。如果高层协议发送的分组过长并超过MTU的数值，PPP就要丢弃这样的帧，并返回差错。

（9）网络层地址协商。**PPP协议必须提供一种机制使通信的两个网络层的实体能够通过协商知道或能够配置彼此的网络层地址**。这对拨号连接的链路特别重要，因为如果仅仅在链路层建立了连接而不知道对方网络层地址，则还不能够保证网络层可以传送分组。

（10）数据压缩协商。**PPP协议必须提供一种方法来协商使用数据压缩算法**。

在TCP/IP协议族中，可靠传输由运输层的TCP协议负责，因此数据链路层的**PPP协议不需要进行纠错，不需要设置序号，也不需要进行流量控制**。**PPP协议不支持多点线路**（即一个主站轮流和链路上的多个从站进行通信），而只支持点对点的链路通信。此外，**PPP协议只支持全双工链路**。

#### 2. PPP协议的组成

（1）一个将IP数据报封装到串行链路的方法。

（2）一个用来建立、配置和测试数据链路连接的**链路控制协议**LCP（Link Control Protocol）。

（3）一套**网络控制协议**NCP（Network Control Protocol），其中的每一个协议支持不同的网络层协议。

### 3.2.2 PPP协议的帧格式

#### 1. 各字段的意义

#### 2. 字节填充

#### 3. 零比特填充

在发送端扫描整个信息字段，只要发现有5个连续1，则立即填入一个0。（因为帧开始和结束的标志字段都是01111110，有6个连续的1。）

### 3.2.3 PPP协议的工作状态



PPP链路的起始和终止状态都是“**链路静止**”。当用户个人电脑通过**调制解调器**呼叫路由器时，路由器就能够检测到调制解调器发出的载波信号。在双方建立了物理层连接后，PPP就进入“**链路建立**”状态，其目的是建立链路层的LCP连接。这时LCP开始协商一些配置选项，即发送**LCP的配置请求帧**。LCP配置选项包括链路上的最大帧长、所使用的鉴别协议的规约等。

协商结束后双方就建立了LCP链路，接着就进入“**鉴别**”状态。在这一状态，只允许传送**LCP协议的分组**、**鉴别协议的分组**以及**监测链路质量的分组**。

在“**网络层协议**”状态，**PPP链路的两端的网络控制协议NCP根据网络层的不同协议互相交换网络层特定的网络控制分组**。

当网络层配置完毕后，链路就进入可进行数据通信的“链路打开”状态。链路的两个PPP端点可以彼此向对方发送分组。两个PPP端点还可发送**回送请求LCP分组**和**回送回答LCP分组**，以检查链路状态。

数据传输结束后，可以由链路的一端发出**终止请求LCP分组**请求终止链路连接，在收到对方发来的**终止确认LCP分组**后，转到“链路终止”状态。

**PPP协议已不是纯粹的数据链路层的协议，它还包含了物理层和网络层的内容**。

## 3.3 使用广播信道的数据链路层

### 3.3.1 局域网的数据链路层

局域网工作的层次跨越了**数据链路层**和**物理层**。

共享信道要着重考虑的一个问题就是如何使众多用户能够合理而方便地共享通信媒体资源。这在技术上有两种方法：

（1）**静态划分信道**。如频分复用、波分复用和码分复用。

（2）**动态媒体接入控制**（多点接入）。

* **随机接入**——所有用户可随机地发送信息。但如果恰巧有两个或更多的用户在同一时刻发送信息，那么在共享媒体上就要产生**碰撞**，使得这些用户的发送都失败。因此，必须有解决碰撞的网络协议。（以太网）
* **受控接入**——用户不能随机地发送信息而必须服从一定的控制。典型代表有**分散**控制的**令牌环局域网**和**集中**控制的**多点线路探询**（轮询）。

#### 1. 以太网的两个标准

#### 2. 适配器的作用



计算机与外界局域网的连接是通过通信**适配器**进行的。适配器和局域网之间的通信是通过电缆或双绞线以串行传输方式进行的，而适配器和计算机之间的通信则是通过计算机主板上的I/O总线以并行传输方式进行的。因此，适配器的一个重要功能就是要进行**数据串行传输和并行传输的转换**。由于网络上的数据率和计算机总线上的数据率并不相同，因此在适配器中必须装有**对数据进行缓存的存储芯片**。在主板上插入适配器时，还必须把**管理该适配器的设备驱动程序**安装在计算机的操作系统中。这个驱动程序以后会告诉适配器，应当**从存储器的什么位置上把多长的数据块发送到局域网**，或者应当**在存储器的什么位置上把局域网传送过来的数据块存储下来**。适配器还要能够实现以太网协议。

适配器实现的功能包含了**数据链路层**和**物理层**这两个层次的功能。**适配器在接收和发送各种帧时，不使用计算机的CPU**。这时计算机的CPU可以处理其他任务。当适配器收到有差错的帧时，就把这个帧直接丢弃而不必通知计算机。当适配器收到正确的帧时，它就使用**中断**来通知计算机，并交付协议帧中的网络层。当计算机要发送IP数据报时，就由协议栈把IP数据报向下交给适配器，组装成帧后发送到局域网。

**计算机的硬件地址就在适配器的ROM中，而计算机的软件地址——IP地址在计算机的存储器中**。

### 3.3.2 CSMA/CD协议

以太网是**无连接**的，提供**尽最大努力交付**的服务，即**不可靠交付**。

**载波监听多点接入/碰撞检测**（Carrier Sense Multiple Access with Collision Detection）协议的要点：

（1）“**多点接入**”——总线型网络，许多计算机以多点接入的方式连接在一根总线上。

（2）“**载波监听**”——用电子技术检测总线上有没有其他计算机也在发送。

（3）“**碰撞检测**”——适配器边发送数据边检测信道上的信号电压的变化情况，以便判断自己再发送数据时其他站是否也在发送数据。任何一个正在发送数据的站，一旦发现总线上出现了碰撞，其适配器就要立即停止发送，免得继续进行无效的发送，白白浪费网络资源，然后等待一段随机事件后再次发送。

既然每一个站在发送数据之前已经监听到信道为“空闲”，为什么还会出现数据在总线上的碰撞呢？因为**电磁波在总线上总是以有限的速率传播的**。

在使用CSMA/CD协议时，一个站不可能同时进行发送和接收。因此**使用CSMA/CD协议的以太网不可能进行全双工通信而只能进行半双工通信**。

以太网的**发送不确定性**：每一个站在自己发送数据之后的一小段时间内，存在着遭遇碰撞的可能性。这一小段时间是不确定的，它取决于另一个发送数据的站到本站的距离。因此，以太网不能保证某一时间之内一定能够把自己的数据帧成功地发送出去。

经过**争用期**（碰撞窗口）这段时间还没有检测到碰撞，才能肯定这次发送不会发生碰撞。

以太网使用**截断二进制指数退避**算法让发生碰撞的站在停止发送数据后，不是等待信道变为空闲后就立即再发送数据，而是推迟一个随机的时间：

（1）基本退避时间为争用期512比特时间（发送1比特所需的时间）。

（2）从离散的整数集合中随机取出一个数，记为r。重传应推后的时间就是r倍的争用期。

（3）当重传达16次仍不能成功时，则丢弃该帧，并向高层报告。

若连续多次发生冲突，就表明可能有较多的站参与争用信道。但使用上述退避算法可使重传需要推迟的平均时间随重传次数而增大（动态退避），因而减小发生碰撞的概率，有利于整个系统的稳定。

适配器每发送一个新的帧，就要执行一次CSMA/CD算法。适配器对过去发生过的碰撞并无记忆功能。当好几个适配器正在执行指数退避算法时，很可能有某一个适配器发送的新帧能够碰巧立即地插入到信道中，得到了发送权，而已经推迟好几次发送的站，有可能很不巧，还要继续执行退避算法，继续等待。

某个站发送了一个很短的帧，在发送完毕之前没有检测出碰撞。假定这个帧在继续向前传播到达目的站之前和别的站发送的帧发生了碰撞，因而目的站将收到有差错的帧（丢弃）。可使发送站却不知道这个帧发生了碰撞，因而不会重传这个帧。为了避免发生这种情况，以太网规定了一个**最短帧长64字节**。

以太网在发送数据时，**如果在争用期（共发送了64字节）没有发生碰撞，那么后续发送的数据就一定不会发生冲突**。如果发生碰撞，就一定是在发送的前64字节之内。由于一检测到冲突就立即中止发送，这时已经发送出去的数据一定小于64字节，因此**凡长度小于64字节的帧都是由于冲突而异常中止的无效帧**。

**强化碰撞**：当发送数据的站一旦发现发生了碰撞时，除了立即停止发送数据外，还要**再发送32比特或48比特的人为干扰信号，以便让所有用户都知道现在已经发生了碰撞**。

以太网还规定了**帧间最小间隔**为9.6μs，相当于96比特时间。这是为了使刚刚收到数据帧的站的接收缓存来得及清理，做好接收下一帧的准备。

以太网每发送完一帧，一定要把已发送的帧暂时保留一下。如果在争用期内检测出发生了碰撞，那么还要在推迟一段时间后再把这个暂时保留的帧重传一次。

* 是什么原因使以太网有一个最小帧长和最大帧长？

在以太网上发送一个帧时，并非全网都能瞬时地检测出这个帧的发送。这是因为电磁波在以太网上的传播速度是有限的。当某个站刚刚发送数据后不久就检测到发生冲突时，就必须丢弃这个受到损伤的帧。这种帧就是因发生碰撞而异常中止的短帧。如果一个站只有非常少的数据要发送，帧就会很短。这种帧就是因发生碰撞而异常中止的短帧。如果一个站只有非常少的数据要发送，帧就会很短。如果让这种短帧在以太网上传送，那么就无法和因发生碰撞而异常中止的短帧相区分。前者是有用的帧，而后者是无用的帧。

设置最小帧长就是为了区分开这两种情况。对于有用的非常短的帧，以太网标准规定要在数据部分进行填充，使其长度达到以太网标准规定的最小帧长的长度。**因发生碰撞而异常中止的短帧的长度肯定都是小于最小帧长的长度**。

**设置最大帧长是为了保证各个站都能公平竞争接入到以太网**。因为如果某个站发送特长的数据帧，则其他的站就必须等待很长的时间才能发送数据。

### 3.3.3 使用集线器的星形拓扑

使用**集线器**的以太网**在逻辑上仍是一个总线网**，各站共享逻辑上的总线，使用的还是CSMA/CD协议（各站中的适配器执行CSMA/CD协议）。网络中的各站必须竞争对传输媒体的控制，并且在同一时刻至多只允许一个站发送数据。

集线器工作在**物理层**，它的每个接口仅仅简单地转发比特——收到1就转发1，收到0就转发0，不进行碰撞检测。若两个接口同时有信号输入（碰撞），那么所有的接口都收不到正确的帧。

### 3.3.4 以太网的信道利用率

### 3.3.5 以太网的MAC层

#### 1. MAC层的硬件地址

MAC地址有48位，固化在适配器的ROM中。

当路由器通过适配器连接到局域网时，适配器上的硬件地址就用来标志路由器的某个接口。路由器如果同时连接到两个网络上，那么它就需要两个适配器和两个硬件地址。

#### 2. MAC帧的格式



当数据字段的长度小于46字节时，MAC子层就会在数据字段的后面加入一个整数字节的填充字段，以保证以太网的MAC帧长不小于64字节。MAC帧的首部并没有指出数据字段的长度是多少。在有填充字段的情况下，接收端的MAC子层在剥去首部和尾部后就把数据字段和填充字段一起交给上层协议。**上层协议如何知道填充字段的长度呢**？（IP层应该丢弃没有用处的填充字段。）可见，上层协议必须具有识别有效的数据字段长度的功能。当上层使用IP协议时，其首部就有一个“总长度”字段。因此，“总长度”加上填充字段的长度，应当当于MAC帧数据字段的长度。例如，**当IP数据报的总长度为42字节时，填充字段共有4字节。当MAC帧把46字节的数据上交给IP层后，IP层就把其中最后4字节的填充字段丢弃**。

在以太网上传送数据时是以帧为单位传送的。**以太网在传送帧时，各帧之间还必须有一定的间隙**。接收端只要找到帧开始定界符，其后面的连续到达的比特流就都属于同一个MAC帧。可见**以太网不需要使用帧结束定界符，也不需要使用字节插入来保证透明传输**。

### 3.4 扩展的以太网

### 3.4.1 在物理层扩展以太网

集线器

### 3.4.2 在数据链路层扩展以太网

#### 1. 以太网交换机的特点

**以太网交换机**实质上就是一个**多接口网桥**，通常都有十几个或更多的接口，和工作在物理层的转发器、集线器有很大差别。**以太网交换机的每个接口都直接与一个单台主机或另一个以太网交换机相连，并且一般都工作在全双工方式**。以太网交换机还具有并行性，即能同时连通多对接口，使多对主机能同时通信（而网桥只能一次分析和转发一个帧）。**相互通信的主机都是独占传输媒体，无碰撞地传输数据**。

以太网交换机的接口还有**存储器**，能**在输出端口繁忙时把到来的帧进行缓存**。如果连接在以太网交换机上的两台主机，同时向另一台主机发送帧，那么当这台主机的接口繁忙时，发送帧的这两台主机的接口会把收到的帧暂存一下，以后再发送出去。

以太网交换机是一种即插即用的设备，其内部的**帧交换表**（地址表）是通过**自学习算法**自动地逐渐建立起来的。以太网交换机使用了专用的交换结构芯片，用硬件转发，其转发速率要比使用软件转发的网桥快很多。

**对于传统的10Mbit/s的共享式以太网，若共有10个用户，则每个用户占有的平均带宽只有1Mbit/s，但由于一个用户在通信时是独占而不是和其他网络用户共享传输媒体的带宽，因此对于拥有10个接口的交换机的总容量则为100Mbit/s**。这正是交换机的最大优点。

从共享总线以太网转到交换式以太网，所有接入设备的软件和硬件、适配器等都不需要作任何改动。

虽然许多以太网交换机对收到的帧采用存储转发方式进行转发，但也有一些交换机采用**直通**的交换方式。直通交换不必把整个数据帧先缓存后再进行处理，而是在接收数据帧的同时就立即按数据帧的目的MAC地址决定该帧的转发接口，因而提高了帧的转发速度，缺点是不检查差错就直接转发。

* 假定一个以太网E1上的通信量中的80%是在本局域网上进行的，而其余的20%的通信量是在本局域网和互联网之间进行的。另一个以太网E2的情况则反过来。这两个以太网一个适用以太网集线器，而另一个使用以太网交换机。你认为以太网交换机应当用在哪一个网络上？

使用集线器的以太网最主要的特点是：这种局域网在逻辑上仍是一个总线网，局域网上的各主机共享逻辑上的总线，使用的还是CSMA/CD协议（各主机中的适配器执行CSMA/CD协议）。网络中的各主机必须竞争对传输媒体的控制，同一时刻至多只允许一个主机发送数据。

使用交换机的以太网最主要的特点是：交换机的每个接口都直接与一个单个主机或另一个集线器相连，并且一般都工作在全双工方式。当主机需要通信时，交换机能同时连通许多对接口，使每一对相互通信的主机都能像独占通信媒体那样，无碰撞地传输数据。

如果以太网的管理者认为，确保局域网上的主机相互通信是最为重要的，那么就应当把交换机安装在以太网E1上。这样可以保证能够有多对局域网上的主机同时进行通信。反之，如果把集线器安装在以太网E1上，那么在同一时间，在局域网上就只能有一对主机进行通信。

如果以太网的管理者认为，确保局域网上的主机能够上互联网是最为重要的，那么就应当把交换机安装在以太网E2上。局域网上的主机要上互联网，首先必须经过本局域网上的路由器。如果以太网E2上只有一个路由器，那么在同一时间还是只有一个主机能够通过这个路由器上网。当某个主机通过这个路由器连接到互联网时，这个局域网上的其他主机就无法再和这个路由器相连。除非这个局域网上有好几个路由器，而通过这些路由器都能够连接到互联网。那么在这种情况下，在这个局域网中使用交换机就可以使多个主机同时接入到互联网。在使用集线器的以太网中，不管这个局域网上有多少个路由器，在同一时间，只能有一个主机与某一个路由器相连。

* 网桥的工作原理和特点是什么？网桥与转发器以及以太网交换机有何异同？

网桥工作在数据链路层，它根据MAC帧的目的地址对收到的帧进行转发和过滤。当网桥收到一个帧时，并不是向所有的接口转发此帧，而是先检查此帧的目的MAC地址，然后再确定将该帧转发到哪一个接口，或者是把它丢弃，网桥依靠转发表来转发帧。

网桥的优点：

（1）过滤通信量，增大吞吐量。

（2）扩大了物理范围，增加了整个以太网上工作站的最大数目。

（3）提高了可靠性。当网络出现故障时，一般只影响个别网段。

（4）可互连不同物理层、不同MAC子层和不同速率的以太网。

网桥的缺点：

（1）由于网桥对接收的帧要先存储和查找转发表，然后才转发，而转发之前，还必须执行CSMA/CD算法，这就增加了时延。

（2）在MAC子层并没有流量控制功能。当网络上的负荷很重时，网桥中的缓存空间可能不够而发生溢出，以致帧丢失。

（3）网桥只适合于用户数不太多（几百个）和通信量不太大的局域网，否则有时还会因传播过多的广播信息而产生网络拥塞（广播风暴）。

网桥与转发器最大的区别就是工作的层次不同。**网桥工作在数据链路层**，它根据MAC帧的目的地址对收到的帧进行转发和过滤。而**转发器工作在物理层**，用来连接以太网的不同网段，以便扩展以太网的覆盖范围。转发器的特点是收到一个比特就转发一个比特，而不进行碰撞检测，也不管整个的帧是有效帧还是无效帧。

以太网交换机实质上就是一个多接口网桥，它的每个接口都直接与一个单个主机或另一个集线器相连（普通网桥的接口往往是连接到以太网的一个网段），并且一般都工作在全双工方式。

#### 2. 以太网交换机的自学习功能

网桥中的转发表是用自学习算法建立的。如果有的站点总是不发送数据而仅仅接收数据，那么在转发表中就没有与这样的站点相对应的项目。如果要向这个站点发送数据帧，那么网桥还是能够把数据帧正确转发到目的地址。这是因为局域网有**广播**的功能。

#### 3. 从总线以太网到星形以太网

总线以太网使用CSMA/CD协议，以半双工方式工作。但以太网交换机不使用共享总线，没有碰撞问题，因此不使用CSMA/CD协议，而是以**全双工**方式工作。但由于其仍采用**以太网的帧结构**，故仍叫作以太网。

### 3.4.3 虚拟局域网

**虚拟局域网**VLAN是由一些局域网网段构成的与物理位置无关的逻辑组，而这些网段具有某些共同的需求。每一个VLAN的帧都有一个明确的标识符，指明发送这个帧的计算机属于哪一个VLAN。

## 3.5 高速以太网

### 3.5.1 100BASE-T以太网

### 3.5.2 吉比特以太网

### 3.5.3 10吉比特以太网（10GE）和更快的以太网

### 3.5.4 使用以太网进行宽带接入

# 第4章 网络层

## 4.1 网络层提供的两种服务

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。

## 4.2 网际协议IP



**地址解析协议**ARP（Address Resolution Protocol）画在IP层最下面，因为IP经常要使用这个协议。**网际控制报文协议**ICMP（Internet Control Message Protocol）和**网际组管理协议**IGMP（Internet Group Management Protocol）画在IP层上部，因为它们要使用IP协议。

### 4.2.1 虚拟互连网络

物理层使用的中间设备叫做**转发器**。

数据链路层使用的中间设备叫做**网桥**或**桥接器**。

网络层使用的中间设备叫做**路由器**。

在网络层以上使用的中间设备叫做**网关**。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。

### 4.2.2 分类的IP地址

#### 1. IP地址及其表示方法

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**。（2）**子网的划分**。（3）**构成超网**。

分类的IP地址：

IP地址={<网络号>，<主机号>}



#### 2. 常用的三种类别的IP地址

A类网络可指派的网络号是126个（27-2）。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**本主机的进程之间的通信之用。

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224-2。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”，因此**全1的主机号字段表示该网络上的所有主机**。

IP地址空间共有232个地址，整个A类地址空间共有231个地址，占整个IP地址空间的50%。

B类地址可指派的网络数为214-1，减1的原因是：B类网络地址128.0.0.0是不指派的，可以指派的B类最小网络地址是128.1.0.0。B类地址的每一个网络上的最大主机数是216-2，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230个地址，占整个IP地址空间的25%。

C类地址可指派的网络数位221-1，减1的原因是：C类网络地址192.0.0.0是不指派的，可指派的C类网络最小地址是192.0.1.0。每一个C类地址的最大主机数是28-2。整个C类地址空间共约有229个地址，占整个IP地址的12.5%。

IP地址的特点：

（1）IP地址是一种分等级的地址结构。IP地址管理机构在分配IP地址时只分配网络号，剩下的主机号由得到该网络号的单位自行分配。**路由器仅根据目的主机所连接的网络号来转发分组**，减小了路由表所占的存储空间以及查找路由表的时间。

（2）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**。

（3）**一个网络是指具有相同网络号的主机的集合**，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**。

（4）在IP地址中，所有分配到网络号的网络都是平等的。

### 4.2.3 IP地址与硬件地址



硬件地址（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。

### 4.2.4 地址解析协议ARP



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**。

每一台主机都设有一个**ARP高速缓存**，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。

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。如果有，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一个**ARP请求分组**（**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，把主机A的IP地址到硬件地址的映射写入ARP高速缓存中，并向主机A发送**ARP响应分组**（**我的IP地址是209.0.0.6，我的硬件地址是08-00-2B-00-EE-0A**）。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。**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。

（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。

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（10~20分钟），凡超过生存时间的项目就从高速缓存中删除掉。

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。



（1）发送方是主机（H1），要把IP数据报发送到同一个网络上上另一台主机（H2）。这时H1发送ARP请求分组（在网1上广播），找到目的主机H2的硬件地址。

（2）发送方是主机（H1），要把IP数据报发送到另一个网络上的一台主机（H3/H4）。这时H1发送ARP请求分组（在网1上广播），找到网1上的一个路由器R1的硬件地址。剩下的工作由路由器R1来完成。R1要做的事情是下面的（3）或（4）。

（3）发送方是路由器（R1），要把IP数据报转发到与R1连接在同一个网络（网2）上的主机（H3）。这时R1发送ARP请求分组（在网2上广播），找到目的主机H3的硬件地址。

（4）发送方是路由器（R1），要把IP数据报转发到网3上的一台主机（H4）。H4与R1不是连接在同一个网络上。这时R1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R2的硬件地址。剩下的工作由这个路由器R2来完成。

* 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，为什么要用抽象的IP地址，而不直接使用硬件地址进行通信？

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**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。

### 4.2.5 IP数据报的格式



#### IP数据报首部的固定部分中的各字段

（1）**版本**。占4位。目前广泛使用的版本是IPv4。

（2）**首部长度**。占4位。首部长度字段所表示数的单位时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。**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**。

（3）**区分服务**。占8位。一般不用。

（4）**总长度**。占16位。首部和数据之和的长度，单位为字节。

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（最大传送单元MTU）。当一个IP数据报封装成链路层的帧时，此数据报的总长度（首部+数据部分）一定不能超过下面的数据链路层所规定的MTU值。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**。

使用尽可能长的IP数据报会使传输效率得到提高（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）。每一个数据报越短，路由转发的速度就越快。

在进行分片时（“片偏移”），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**。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”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**。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“不能切片”。只有当DF=0时才允许切片。

（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字节的整数倍**。

（8）**生存时间**TTL（Time To Live）。占8位，表明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，从而白白消耗网络资源。TTL中记录的是“**跳数限制**”，指明数据报在互联网中至多可经过多少个路由器。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。**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抛弃。

（9）**协议**。占8位。指出此数据报携带的数据是使用何种协议（TCP, UDP, ...），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。

（10）**首部检验和**。占16位。**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（生存时间、标志、片偏移等字段可能发生变化）。不检验数据部分可减少计算的工作量。

（11）**源地址**。占32位。

（12）**目的地址**。占32位。

#### IP数据报首部的可变部分

### 4.2.6 IP层转发分组的流程

* 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（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。

## 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

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。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**。

划分子网的方法是从网络的主机号借用若干位作为子网号，当然主机号也就相应减少了同样的位数。

IP地址={<网络号>，<子网号>，<主机号>}

凡是从其他网络发送给本单位某台主机的IP数据报，仍然是根据IP数据报的目的网络号找到连接在本单位网络上的路由器。但此路由器在收到IP数据报后，再按目的网络号和子网号找到目的子网，把IP数据报交付目的主机。

#### 2. 子网掩码

从IP数据报的首部无法看出源主机或目的主机所连接的网络是否进行了子网的划分。这是因为32位的IP地址本身以及数据报的首部都没有包含任何有关子网划分的信息。因此必须另外想办法，这就是使用**子网掩码**。

子网掩码是一个网络或一个子网的重要属性。路由器在和相邻路由器交换路由信息时，必须把自己所在网络（或子网）的子网掩码告诉相邻路由器。在路由器的路由表中的每一个项目，除了要给出目的网络地址外，还必须同时给出该网络的子网掩码。若一个路由器连接在两个子网上就拥有两个网络地址和两个子网掩码。

**同样的IP地址和不同的子网掩码可以得出相同的网络地址**。

### 4.3.2 使用子网时分组的转发

使用子网划分后，路由表必须包含以下三项内容：**目的网络地址**、**子网掩码**和**下一跳地址**。

### 4.3.3 无分类编址CIDR（构造超网）

#### 1. 网络前缀

**无分类域间路由选择**CIDR（Classless Inter-Domain Routing）的特点：

（1）CIDR消除了传统的A类、B类和C类地址以及划分子网的概念。

IP地址={<网络前缀>，<主机号>}

（2）CIDR把网络前缀都相同的连续的IP地址组成一个**“CIDR”地址块**。

#### 最长前缀匹配

在使用CIDR时，由于采用了网络前缀这种记法，IP地址由网络前缀和主机号这两个部分组成，因此在路由表中的项目也要有相应的改变。这时，每个项目由“**网络前缀**”和“**下一跳地址**”组成。但是在查找路由表时可能会得到不止一个匹配结果。这样就带来一个问题：我们应当从这些匹配结果中选择哪一条路由呢？

正确的答案是：**应当从匹配结果中选择具有最长网络前缀的路由**。这叫做**最长前缀匹配**，这是因为网络前缀越长，其地址块就越小，因而路由就越具体。

#### 使用二叉线索查找路由表

## 4.4 网际控制报文协议ICMP



为了更有效地转发IP数据报和提高交付成功的机会，在网际层使用了**网际控制报文协议**ICMP。ICMP**允许主机或路由器报告差错情况和提供有关异常情况的报告**。ICMP是互联网的标准协议。但ICMP不是高层协议（看起来好像是高层协议，因为ICMP报文是装在IP数据报中，作为其中的数据部分），而是IP层的协议。ICMP报文作为IP层数据报的数据，加上数据报的首部，组成IP数据报发送出去。

### 4.4.1 ICMP报文的种类

|  |  |  |
| --- | --- | --- |
| ICMP报文种类 | 类型的值 | ICMP报文的类型 |
| 差错报告报文 | 3 | 终点不可达 |
| 11 | 时间超过 |
| 12 | 参数问题 |
| 5 | 改变路由（Redirect） |
| 询问报文 | 8或0 | 回送（Echo）请求或回答 |
| 13或14 | 时间戳（Timestamp）请求或回答 |

ICMP报文的代码字段是为了进一步区分某种类型中的几种不同情况。检验和字段用来检验整个ICMP报文。**IP数据报首部的检验和并不检验IP数据报的内容**，因此不能保证经过传输的ICMP报文不产生差错。

ICMP差错报告报文共有四种：

（1）**终点不可达**。当路由器或主机不能交付数据报时就向源点发送终点不可达报文。

（2）**时间超过**。当路由器收到生存时间为零的数据报时，除丢弃该数据报外，还要向源点发送时间超过报文。当**终点在预先规定的时间内不能收到一个数据报的全部数据报片时，就把已收到的数据报片都丢弃，并向源点发送时间超过报文**。

（3）**参数问题**。当路由器或目的主机收到的数据报的首部中有的字段的值不正确时，就丢弃该数据报，并向源点发送参数问题报文。

（4）**改变路由（重定向）**。路由器把改变路由报文发送给主机，让主机知道下次应将数据报发送给另外的路由器（可通过更好的路由。

在互联网的主机有一个路由表。当主机要发送数据报时，首先是查找主机自己的路由表，看应当从哪一个接口把数据报发送出去。在互联网中主机的数量远大于路由器的数量，出于效率的考虑，这些主机不和连接在网络上的路由器定期交换路由信息。在主机刚开始工作时，一般都在路由表中设置一个**默认路由器**的IP地址。不管数据报要发送到哪个目的地址，都一律先把数据报传送给这个默认路由器，而这个默认路由器知道到每一个目的网络的最佳路由。如果默认路由器发现主机发往某个目的地址的数据报的最佳路由应当经过网络上的另一个路由器R时，就用改变路由报文把这情况高速主机。于是，该主机就在其路由表中增加一个项目：到某某目的地址应经过路由器R（而不是默认路由器）。



所有的ICMP差错报告报文中的数据字段都具有同样的格式。把收到的需要进行差错报告的**IP数据报的首部**和**数据字段的前8个字节**提取出来，作为ICMP报文的数据字段。再加上相应的ICMP差错报告报文的前8个字节，就构成了ICMP差错报告报文。提取收到的数据报的数据字段前8个字节是为了得到运输层的端口号（对于TCP和UDP）以及运输层报文的发送序号（对于TCP）。这些信息对源点通知高层协议是有用的。**整个ICMP报文作为IP数据报的数据字段发送给源点**。

对ICMP差错报告报文，不再发送ICMP差错报告报文。

对第一个分片的数据报片的所有后续数据报片，都不发送ICMP差错报告报文。

对具有多播地址的数据报，都不发送ICMP差错报告报文。

对具有特殊地址（127.0.0.0或0.0.0.0）的数据报，不发送ICMP差错报告报文。

ICMP询问报文共有两种：

（1）**回送请求和回答**。ICMP回送请求报文是由主机或路由器向一个特定的目的主机发出的询问。收到此报文的主机必须给源主机或路由器发送ICMP回送回答报文。这种询问报文用来**测试目的站是否可达以及了解其有关的状态**。

（2）**时间戳请求和回答**。ICMP时间戳请求报文是请某台主机或路由器回答当前的日期和时间。在ICMP时间戳回答报文中有一个32位的字段，其中写入的整数代表从1900年1月1日起到当前时刻一共有多少秒。时间戳请求和回答可用于**时钟同步和时间测量**。

### 4.4.2 ICMP的应用举例

**分组网间探测**PING（Packet InterNet Groper）用来测试两台主机之间的连通性。PING使用了**ICMP回送请求与回送回答报文**。PING是**应用层直接使用网络层ICMP**的一个例子，它没有通过TCP或UDP。

**Traceroute**用来跟踪一个分组从源点到终点的路径。Traceroute从源主机向目的主机发送一连串的IP数据报，数据报中封装的是无法交付的UDP数据报。第一个数据报P1的生存时间TTL设置为1。当P1到达路径上的第一个路由器R1时，路由器R1先收下它，接着把TTL的值减1。由于TTL等于零了，R1就把P1丢弃了，并向源主机发送一个**ICMP时间超过差错报告报文**。

源主机接着发送第二个数据报P2，并把TTL设置为2。P2先到达路由器R1，R1收下后把TTL减1再转发给路由器R2。R2收到P2时TTL为1，但减1后TTL变为零了。R2就丢弃P2，并向源主机发送一个**ICMP时间超过差错报告报文**。这样一直持续下去。当最后一个数据报到达目的主机时，数据报的TTL是1。主机不转发数据报，也不把TTL值减1。但因IP数据报中封装的是无法交付的运输层的UDP用户数据报，因此目的主机要向源主机发送**ICMP终点不可达差错报告报文**。

这样，源主机达到了自己的目的，因为这些路由器和最后目的主机发来的ICMP报文正好给出了源主机想知道的路由信息——到达目的主机所经过的路由器的IP地址，以及到达其中的每一个路由器的往返时间。

## 4.5 互联网的路由选择协议

### 4.5.1 有关路由选择协议的几个基本概念

#### 1. 理想的路由算法

**静态路由选择**（非自适应路由选择）：简单、开销小，不能及时适应网络状态的变化，适用于简单的小网络。

**动态路由选择**（自适应路由选择）：能较好地适应网络状态的变化，实现复杂、开销大。

#### 2. 分层次的路由选择协议

自治系统AS（Autonomous System）是在单一技术管理下的一组路由器，而这些路由器使用一种自治系统内部的路由选择协议和共同度量。一个大的ISP就是一个自治系统。

**内部网关协议**IGP（Interior Gateway Protocol）：在一个自治系统内部使用的路由选择协议，而这与在互联网中的其他自治系统选用什么路由选择协议无关，如RIP和OSPF。

**外部网关协议**EGP（External Gateway Protocol）：若源主机和目的主机处在不同的自治系统中（这两个自治系统可能使用不同的内部网关协议），当数据报传到一个自治系统的边界时，就需要使用一种协议将路由选择信息传递到另一个自治系统中，这样的协议就是外部网关协议IGP，如BGP。

**域间路由选择**：自治系统之间的路由选择。

**域内路由选择**：自治系统内部的路由选择。

### 4.5.2 内部网关协议RIP

#### 1. 工作原理

**路由信息协议**RIP（Routing Information Protocol）是一种**分布式**的基于**距离向量**的路由选择协议。

**距离**（跳数）：从一路由器到直接连接的网络的距离定义为1。从一路由器到非直接连接的网络的距离定义为所经过的路由器数加1。“加1”是因为到达网络后就进行直接交付，而到直接连接的网络的距离已经定义为1。

RIP认为好的路由就是它通过的路由器的数目少。RIP允许一条路径最多只能包含15个路由器，**“距离”等于16时相当于不可达**，故RIP只适用于小型互联网。

RIP不能在两个网络之间同时使用多条路由。RIP选择一条具有最少路由器的路由（**最短路由**），哪怕还存在另一条高速（低时延）但路由器较多的路由。

* 和哪些路由器交换信息？

**仅和相邻路由器交换信息**。

* 交换什么信息？

当前本路由器所知道的全部信息，即**自己现在的路由表**。

* 在什么时候交换信息？

**按固定的时间间隔交换路由信息**。

**路由器在刚刚开始工作时，路由表是空的**。然后路由器就得出到直接相连的几个网络的距离。接着，每一个路由器也只和数目非常有限的相邻路由器交换并更新路由信息。但经过若干次的更新后，所有的路由器最终都会知道到达本自治系统中任何一个网络的最短距离和下一跳路由器的地址。

#### 2. 距离向量算法

对每一个相邻路由器发送过来的RIP报文，进行以下步骤：

（1）对地址为X的相邻路由器发来的RIP报文，先修改此报文中的所有项目（每一个项目都有三个关键数据，即：到目的网络N，距离是d，下一跳路由器是X）：把“下一跳”字段中的地址都改为X，并把所有的“距离”字段的值加1。

（2）对修改后的RIP报文中的每一个项目，进行以下步骤：

若原来的路由表中没有目的网络N，则把该项目添加到路由表中。

否则（即在路由表中有目的网络N，这时就再查看下一跳路由器地址）

若下一跳路由器地址是X，则把收到的项目替换原路由表中的项目（**因为这是最新的消息，要以最新的消息为准**）。

否则（即这个项目时：到目的网络N，但下一跳路由器不是X）

若收到的项目中的距离d小于路由表中的距离，则进行更新，

否则说明也不做。

（3）若3分钟还没有收到相邻路由器的更新路由表，则把此相邻路由器记为不可达的路由器，即把距离置为16。

（4）返回。

**Bellman-Ford算法**：设X是结点A到B的最短路径上的一个结点。若把路径A→B拆成两段路径A→X和X→B，则每一段路径A→X和X→B也都分别是结点A到X和结点X到B的最短路径。

#### 3. RIP协议的报文格式



RIP协议使用运输层的**用户数据报UDP**进行传送（使用UDP的端口520）。

RIP的首部占4字节，其中的命令字段指出报文的意义，1表示请求路由信息，2表示对请求路由信息的响应或未被请求而发出的路由更新报文。首部后面的“必为0”是为了4字节的对齐。

**地址族标识符**（地址类别）字段用来标志所使用的地址协议，如采用IP地址就令这个字段的值为2（原来考虑RIP也可用于其他非TCP/IP协议的情况）。**路由标记**填入自治系统号ASN（Autonomous System Number），这是考虑使RIP有可能收到本自治系统以外的路由选择信息。

RIP存在的一个问题是**当网络出现故障时，要经过比较长的时间才能将此信息传送到所有的路由器**。

### 4.5.3 内部网关协议OSPF

#### 1. OSPF协议的基本特点

**开放最短路径优先**OSPF（Open Shortest Path First）使用了Dijkstra最短路径算法，使用**分布式**的**链路状态**协议。

* 和哪些路由器交换信息？

**向本自治系统中的所有路由器发送信息**。**洪泛法**：路由器通过所有输出端口向所有相邻的路由器发送信息。而每一个相邻路由器又再将此信息发往其所有的相邻路由器（但不再发送给刚刚发来信息的那个路由器）。这样，最终整个区域中所有的路由器都得到了这个信息的一个副本。

* 交换什么信息？

发送的信息是**与本路由器相邻的所有路由器的链路状态**（本路由器都和哪些路由器相邻，以及该链路的“**度量**”）。

* 在什么时候交换信息？

只有**当链路状态发生变化时**，路由器才向所有路由器使用洪泛法发送此信息。

由于各路由器之间频繁地交换链路状态信息，因此所有的路由器最终都能建立一个**链路状态数据库**，这个数据库实际上就是**全网的拓扑结构图**。这个拓扑结构图在全网范围内是一致的（**链路状态数据库的同步**）。每一个路由器都知道全网共有多少个路由器，以及哪些路由器是相连的，其代价是多少。每一个路由器使用链路状态数据库中的数据，使用**Dijkstra最短路径算法**构造出自己的路由表。

为了使OSPF能够用于规模很大的网络，将一个自治系统再划分为若干个更小的范围，叫做**区域**。一个预取内的路由器最少不超过200个。



#### 2. OSPF的五种分组类型

### 4.5.4 外部网关协议BGP

### 4.5.5 路由器的构成