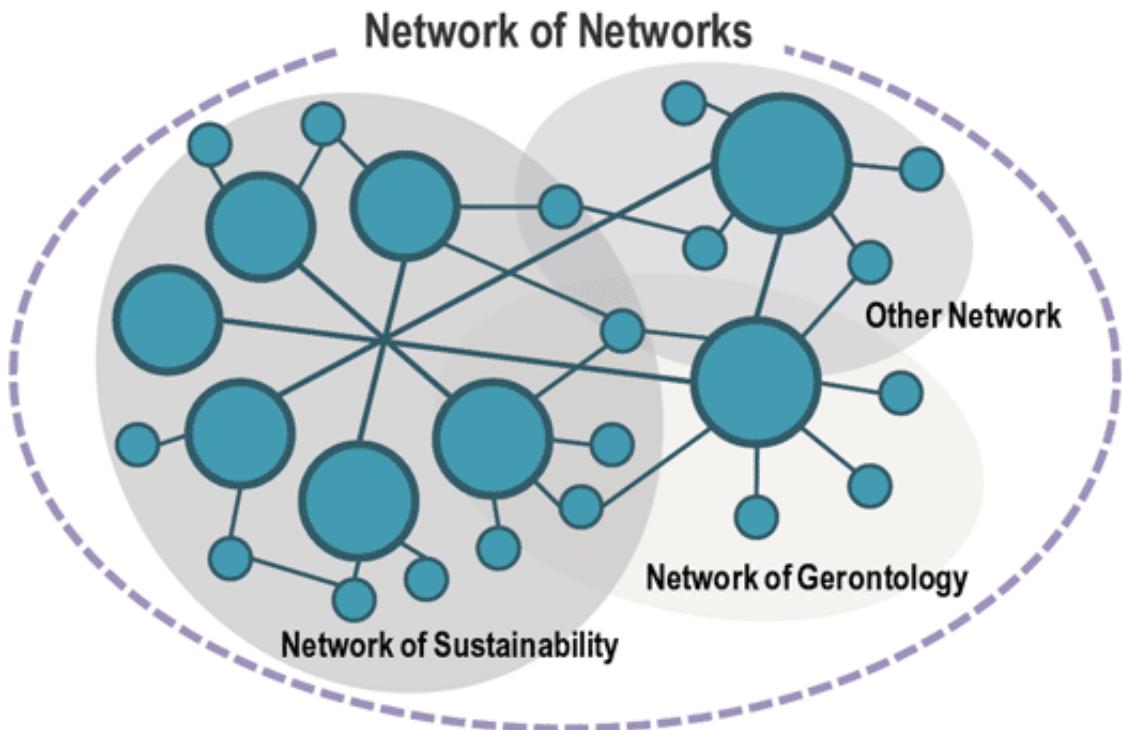


一、概述

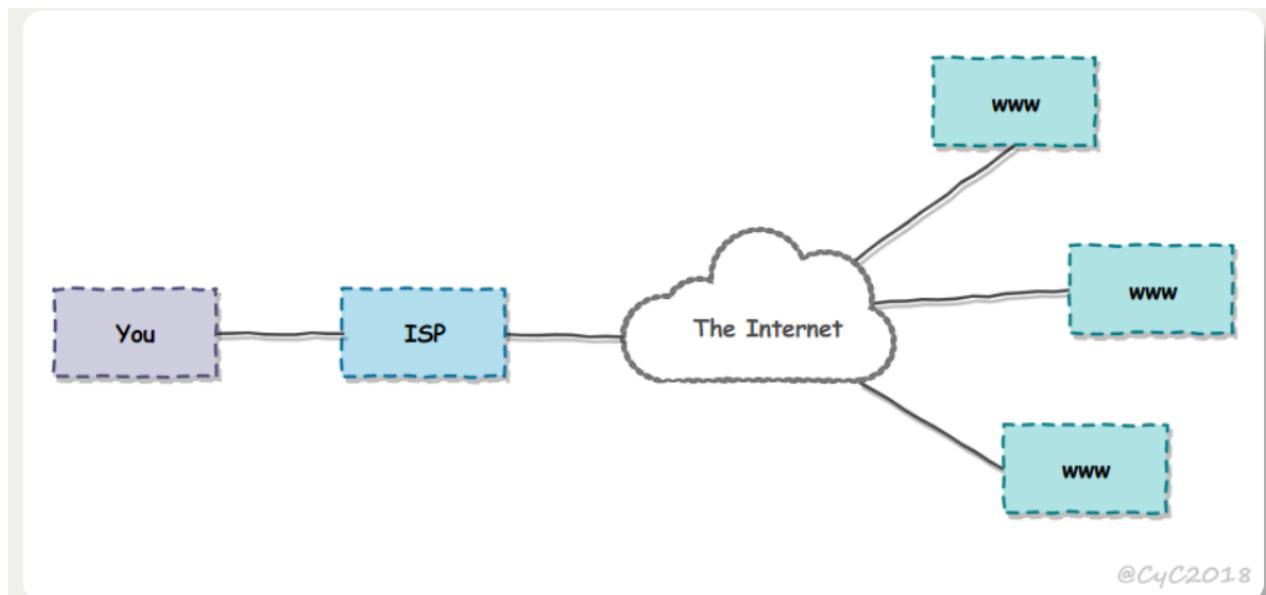
网络的网络

网络把主机连接起来，而互联网是把多种不同的网络连接起来，因此互联网是网络的网络。

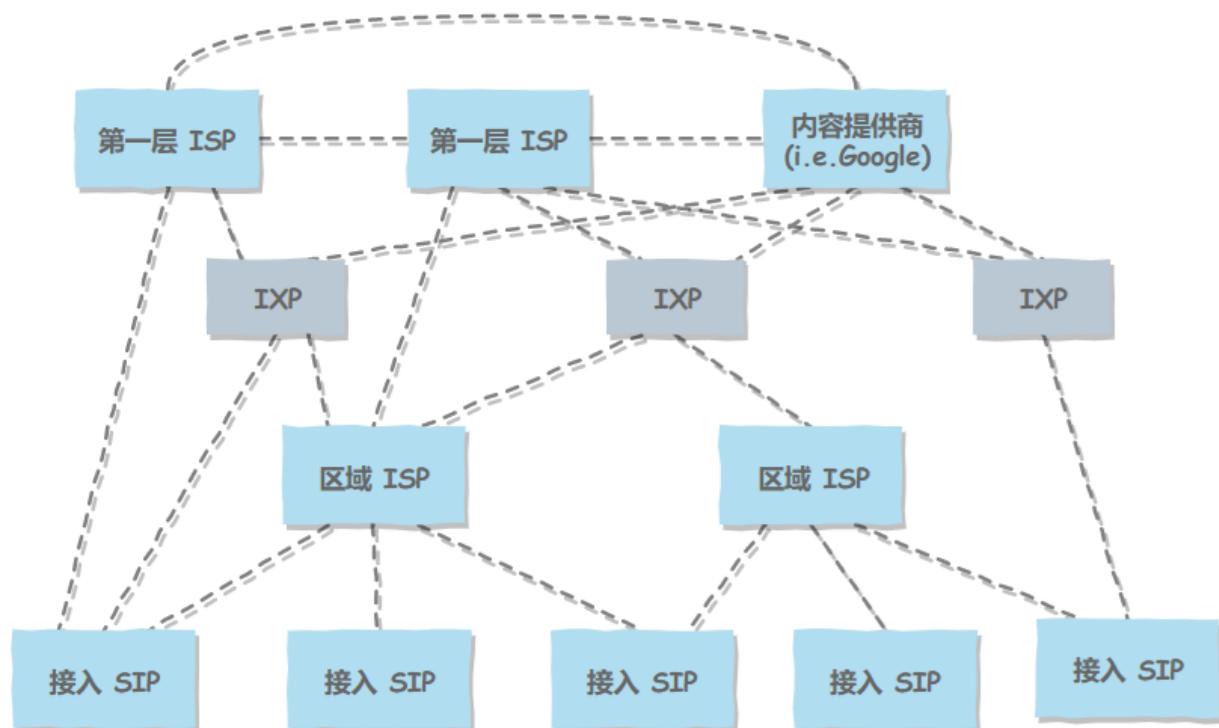


ISP (Internet Service Provider)

互联网服务提供商 ISP 可以从互联网管理机构获得许多 IP 地址，同时拥有通信线路以及路由器等联网设备，个人或机构向 ISP 缴纳一定的费用就可以接入互联网。



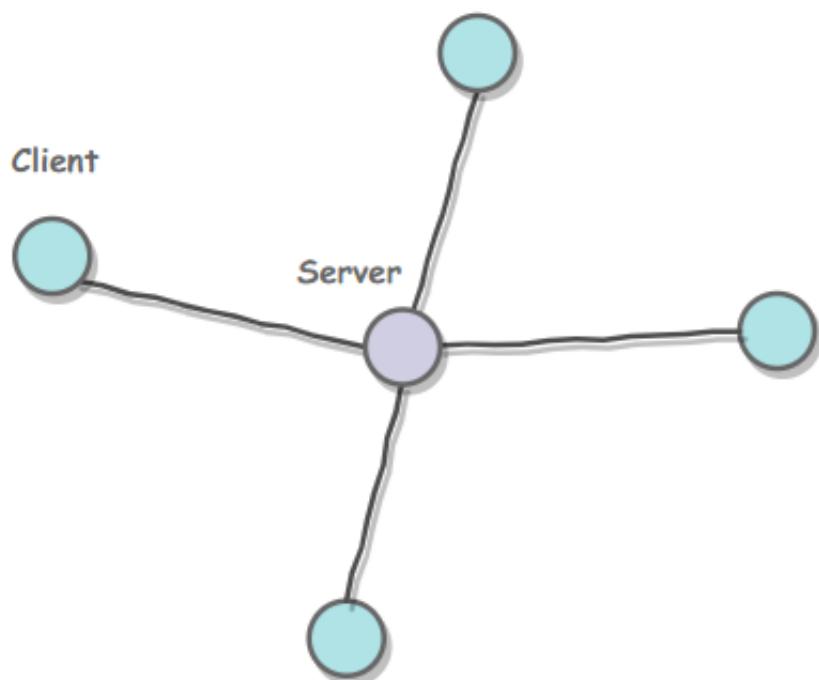
目前的互联网是一种多层次 ISP 结构，ISP 根据覆盖面积的大小分为第一层 ISP、区域 ISP 和接入 ISP。互联网交换点 IXP 允许两个 ISP 直接相连而不用经过第三个 ISP。



@Cyc2018

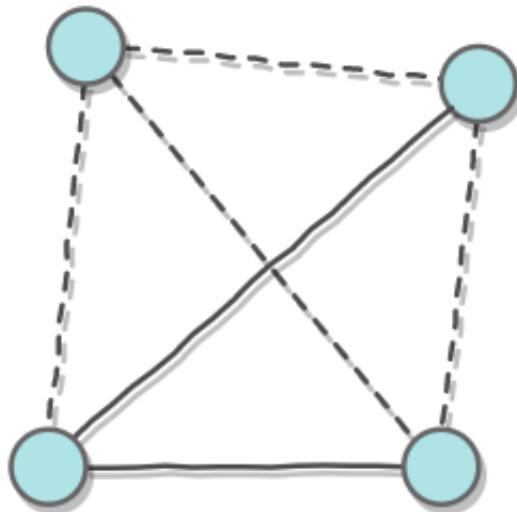
主机之间通信方式

- 客户-服务器 (C/S)：客户是服务的请求方，服务器是服务的提供方。



@Cyc2018

- 对等 (P2P)：不区分客户和服务器。



@Cyc2018

电路交换与分组交换

1. 电路交换

电路交换用于电话通信系统，两个用户要通信之前需要建立一条**专用的物理链路**，并且在整个通信过程中始终占用该链路。由于通信的过程中不可能一直在使用传输线路，因此电路交换对线路的利用率很低，往往不到 10%。

2. 分组交换

每个分组都有首部和尾部，包含了源地址和目的地址等控制信息，在同一个传输线路上同时传输多个分组互相不会影响，因此在同一条传输线路上允许同时传输多个分组，也就是说分组交换不需要占用传输线路。

在一个邮局通信系统中，邮局收到一份邮件之后，先存储下来，然后把相同目的地的邮件一起转发到下一个目的地，这个过程就是**存储转发过程**，分组交换也使用了存储转发过程。

时延

$$\text{总时延} = \text{排队时延} + \text{处理时延} + \text{传输时延} + \text{传播时延}$$



@Cyc2018

1. 排队时延

分组在路由器的输入队列和输出队列中排队等待的时间，取决于网络当前的通信量。

2. 处理时延

主机或路由器收到分组时进行处理所需要的时间，例如分析首部、从分组中提取数据、进行差错检验或查找适当的路由等。

3. 传输时延

主机或路由器传输数据帧所需要的时间。

$$delay = \frac{l(bit)}{v(bit/s)}$$

其中 l 表示数据帧的长度， v 表示传输速率。

4. 传播时延

电磁波在信道中传播所需要花费的时间，电磁波传播的速度接近光速。

$$delay = \frac{l(m)}{v(m/s)}$$

其中 l 表示信道长度， v 表示电磁波在信道上的传播速度。

计算机网络体系结构



@Cyc2018

1. 五层协议

- **应用层**：为特定应用程序提供数据传输服务，例如 HTTP、DNS 等协议。（数据单位）报文。
- **传输层**：为进程提供通用数据传输服务。由于应用层协议很多，定义通用的传输层协议就可以支持不断增多的应用层协议。运输层包括两种协议：**传输控制协议 TCP**，提供面向连接、可靠的数据传输服务，数据单位为报文段；**用户数据报协议 UDP**，提供无连接、尽最大努力的数据传输服务，数据单位为用户数据报。TCP 主要提供完整性服务，UDP 主要提供及时性服务。（数据单位）报文段或者用户数据报
- **网络层**：为主机提供数据传输服务。而传输层协议是为主机中的进程提供数据传输服务。网络层把传输层传递下来的报文段或者用户数据报封装成组（数据单位）。
- **数据链路层**：网络层针对的还是主机之间的数据传输服务，而主机之间可以有很多链路，链路层协议就是为同一链路的主机提供数据传输服务。数据链路层把网络层传下来的分组封装成帧（数据单位）。
- **物理层**：考虑的是怎样在传输媒体上上传输数据比特流（数据单位），而不是指具体的传输媒体。物理层的作用是尽可能屏蔽传输媒体和通信手段的差异，使数据链路层感觉不到这些差异。

2. OSI (开放式系统互连通信参考模型，Open System Interconnection Reference Model)

其中表示层和会话层用途如下：

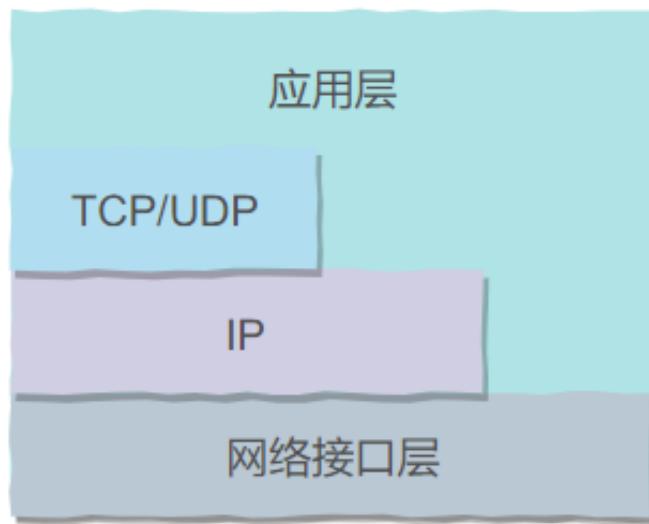
- **表示层**：数据压缩、加密以及数据描述，这使得应用程序不必关心在各台主机中数据内部格式不同的问题。
- **会话层**：建立及管理会话。

五层协议没有表示层和会话层，而是将这些功能留给应用程序开发者处理。

3. TCP/IP

它只有四层，相当于五层协议中数据链路层和物理层合并为网络接口层。

TCP/IP 体系结构不严格遵循 OSI 分层概念，应用层可能会直接使用 IP 层或者网络接口层。



@Cyc2018

4. 数据在各层之间的传递过程

在向下的过程中，需要添加下层协议所需要的首部或者尾部，而在向上的过程中不断拆开首部和尾部。

路由器只有下面三层协议，因为路由器位于网络核心中，不需要为进程或者应用程序提供服务，因此也就不需要传输层和应用层。

二、物理层

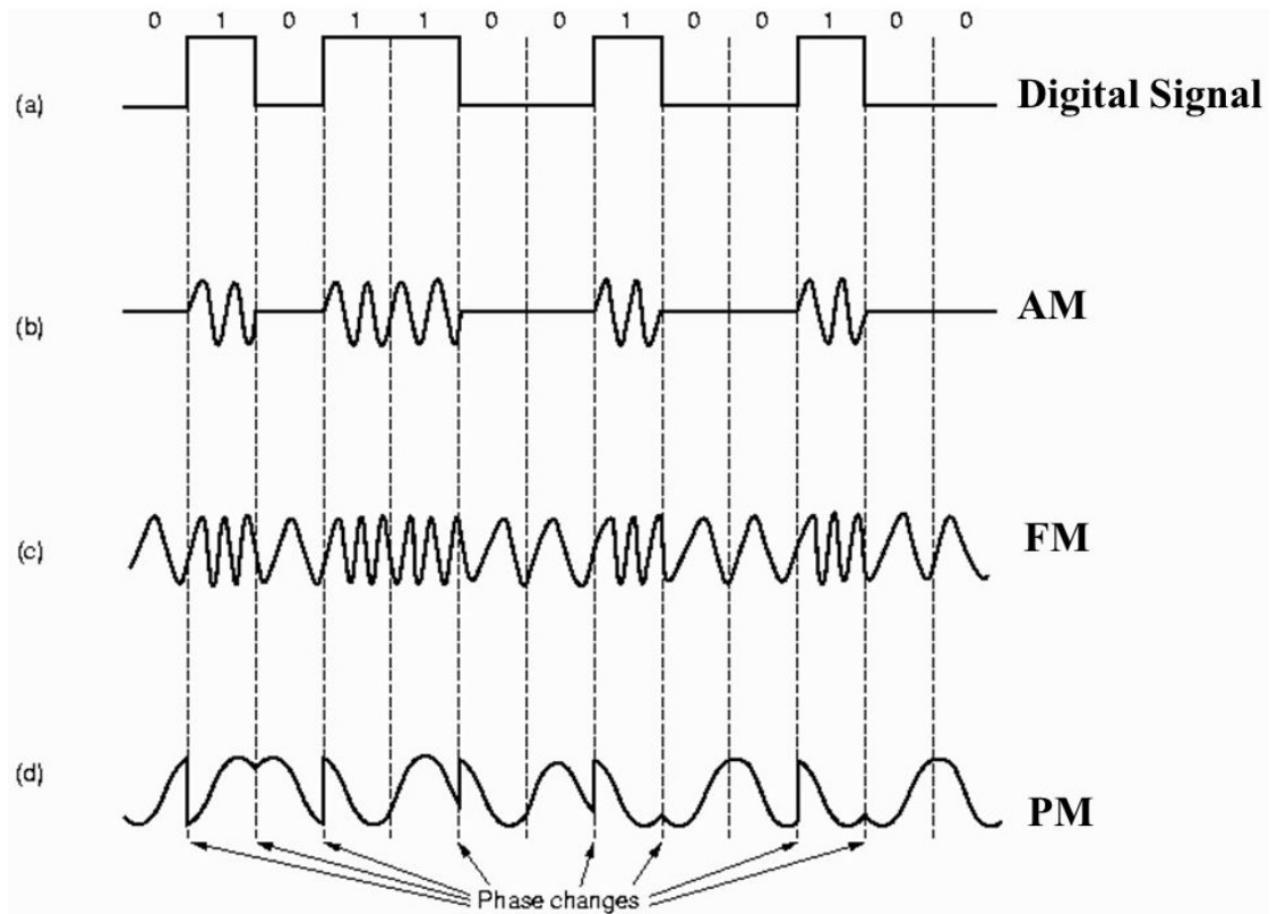
通信方式

根据信息在传输线上的传送方向，分为以下三种通信方式：

- 单工通信：单向传输
- 半双工通信：双向交替传输
- 全双工通信：双向同时传输

带通调制

模拟信号是连续的信号，数字信号是离散的信号。带通调制把数字信号转换为模拟信号。



三、链路层

基本问题

1. 封装成帧

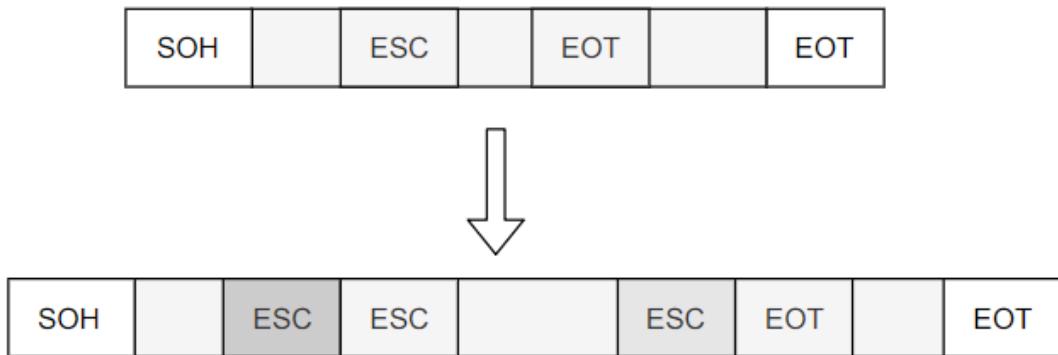
将网络层传下来的分组添加首部和尾部，用于标记帧的开始和结束。



2. 透明传输

透明表示一个实际存在的事物看起来好像不存在。

帧使用首部和尾部进行定界，如果帧的数据部分含有和首部尾部相同的内容，那么帧的开始和结束位置就会被错误的判定。需要在数据部分出现首部尾部相同的内容前面插入转义字符。如果数据部分出现转义字符，那么就在转义字符前面再加个转义字符。在接收端进行处理之后可以还原出原始数据。这个过程透明传输的内容是转义字符，用户察觉不到转义字符的存在。



3. 差错检测

目前数据链路层广泛使用了循环冗余检验（**CRC**）来检查比特差错。

信道分类

1. 广播信道

一对多通信，一个节点发送的数据能够被广播信道上所有的节点接收到。

所有的节点都在同一个广播信道上发送数据，因此需要有专门的控制方法进行协调，避免发生冲突（冲突也叫碰撞）。

主要有两种控制方法进行协调，一个是使用信道复用技术，一是使用 **CSMA/CD** 协议。

2. 点对点信道

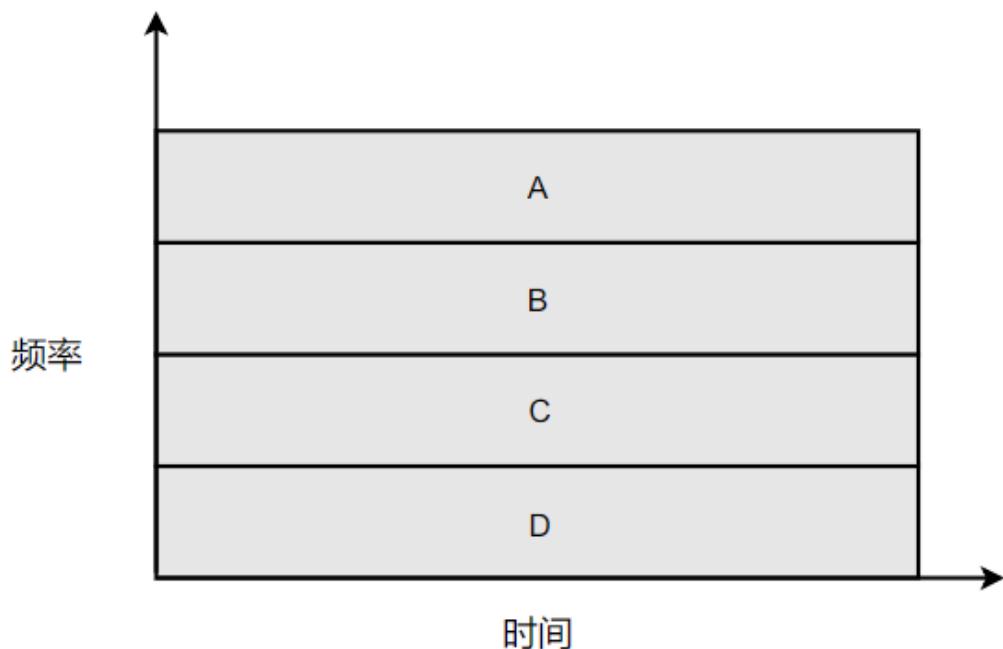
一对一通信。

因为不会发生碰撞，因此也比较简单，使用 **PPP** 协议进行控制。

信道复用技术

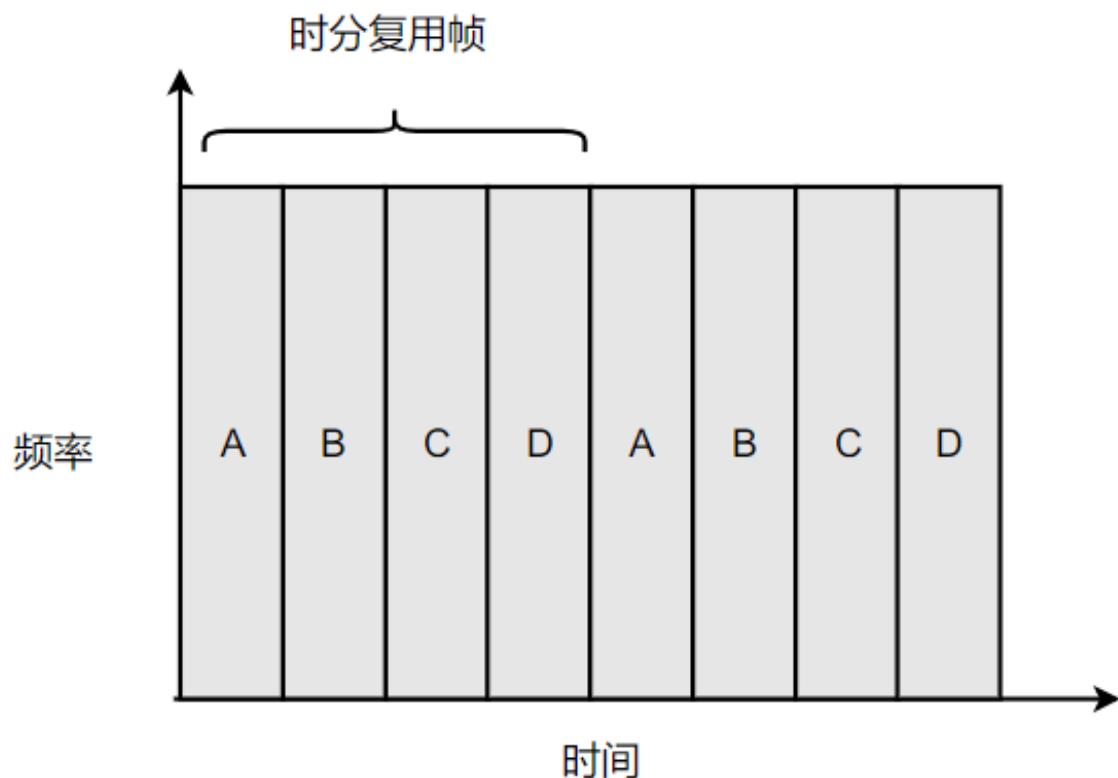
1. 频分复用

频分复用的所有主机在相同的时间占用不同的频率带宽资源。



2. 时分复用

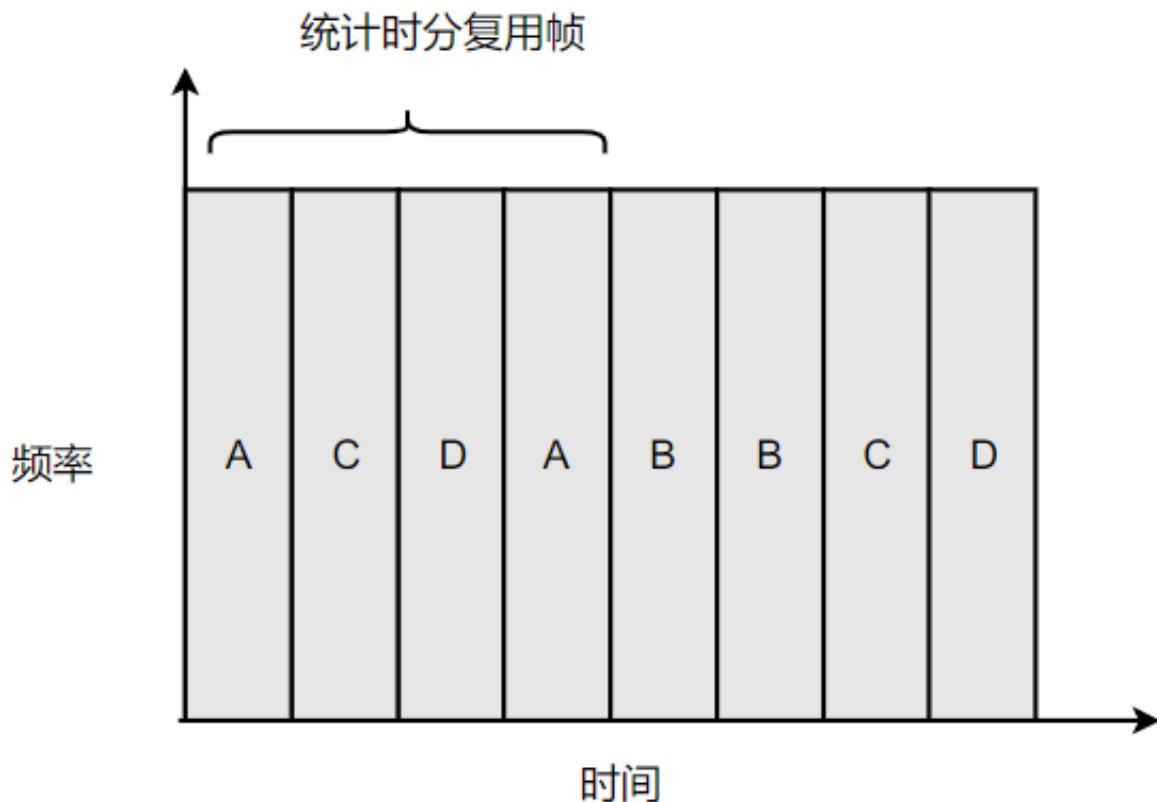
时分复用的所有主机在不同的时间占用相同的频率带宽资源。



使用频分复用和时分复用进行通信，在通信的过程中主机会一直占用一部分信道资源。但是由于计算机数据的突发性质，通信过程没必要一直占用信道资源而不出给其它用户使用，因此这两种方式对信道的利用率都不高。

3. 统计时分复用

是对时分复用的一种改进，不固定每个用户在时分复用帧中的位置，只要有数据就集中起来组成统计时分复用帧然后发送。



4. 波分复用

光的频分复用。由于光的频率很高，因此习惯上用波长而不是频率来表示所使用的光载波

5. 码分复用 (TBD)

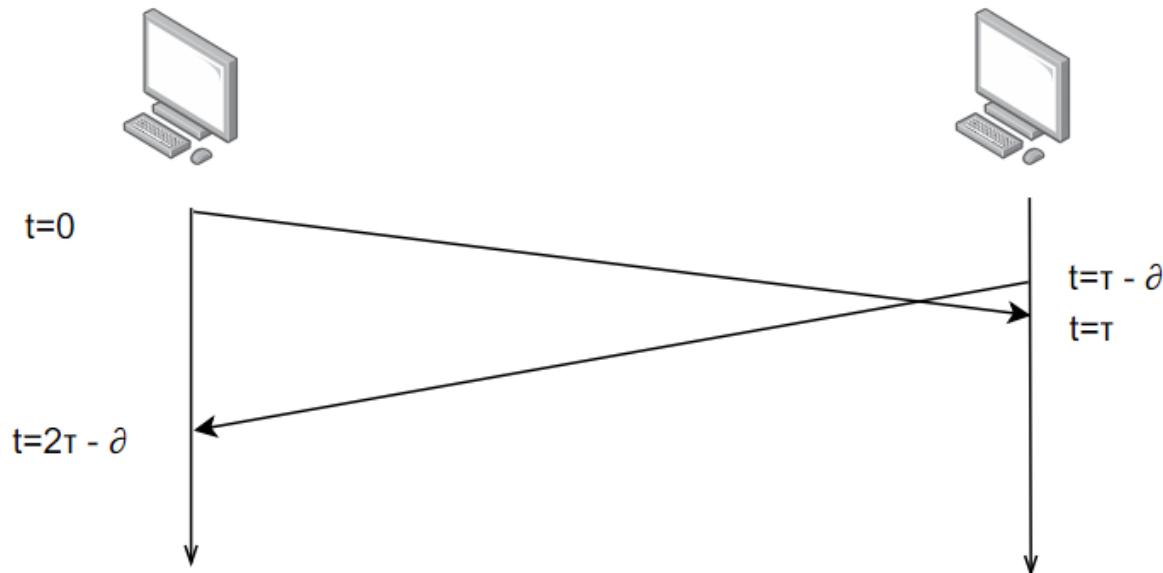
CSMA/CD协议（解决广播信道碰撞）

CSMA/CD 表示载波监听多点接入 / 碰撞检测。

- 多点接入：说明这是总线型网络，许多主机以多点的方式连接到总线上。
- 载波监听：每个主机都必须不停地监听信道。在发送前，如果监听到信道正在使用，就必须等待。
- 碰撞检测：在发送中，如果监听到信道已有其它主机正在发送数据，就表示发生了碰撞。虽然每个主机在发送数据之前都已经监听到信道为空闲，但是由于电磁波的传播时延的存在，还是有可能会发生碰撞。

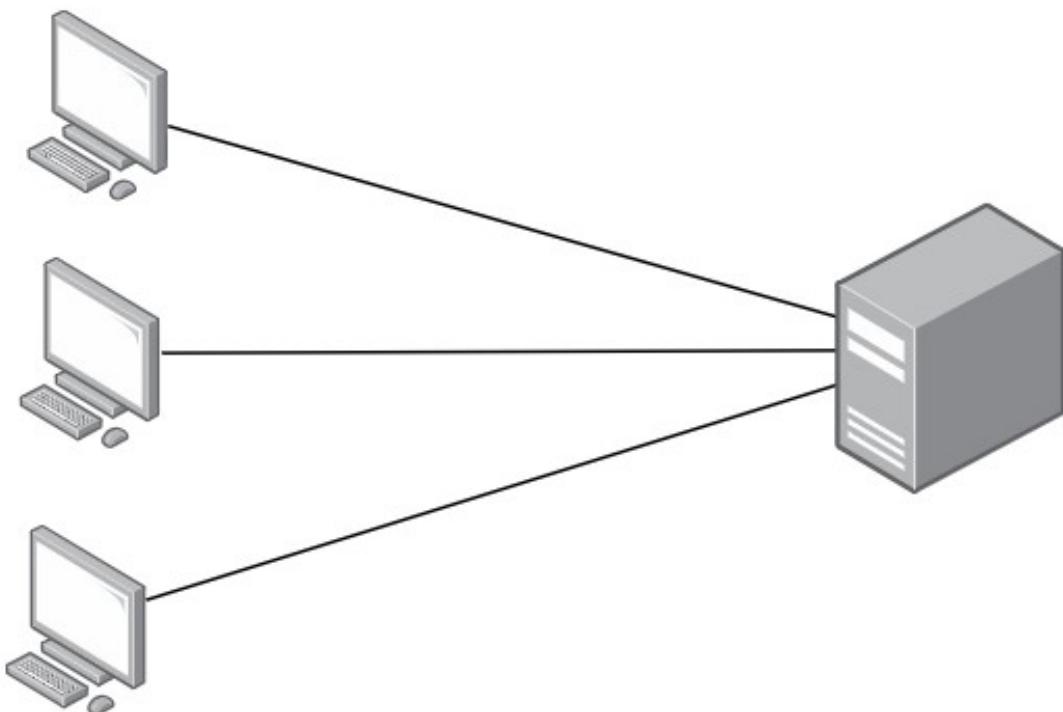
记端到端的传播时延为 τ ，最先发送的站点最多经过 2τ 就可以知道是否发生了碰撞，称 2τ 为 争用期。只有经过争用期之后还没有检测到碰撞，才能肯定这次发送不会发生碰撞。

当发生碰撞时，站点要停止发送，等待一段时间再发送。这个时间采用 截断二进制指数退避算法 来确定。从离散的整数集合 $\{0, 1, \dots, (2k-1)\}$ 中随机取出一个数，记作 r ，然后取 r 倍的争用期作为重传等待时间。



PPP协议

互联网用户通常需要连接到某个 ISP 之后才能接入到互联网，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。



PPP 的帧格式：

- F 字段为帧的定界符
- A 和 C 字段暂时没有意义
- FCS 字段是使用 **CRC** 的检验序列
- 信息部分的长度不超过 1500

F	A	C	协议	IP 数据报	FCS	F
---	---	---	----	--------	-----	---

MAC地址

MAC 地址是链路层地址，长度为 **6 字节（48 位）**，用于唯一标识网络适配器（网卡）。

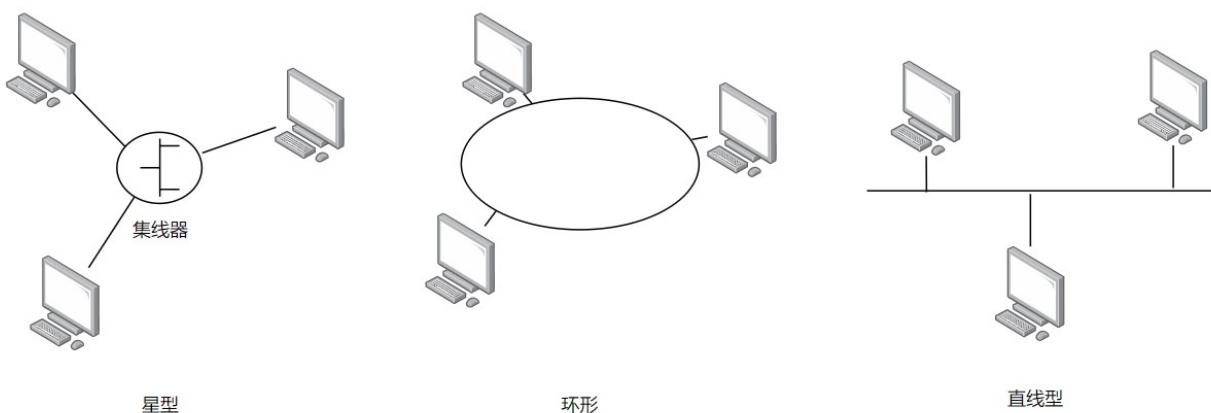
一台主机拥有多少个网络适配器就有多少个 MAC 地址。例如笔记本电脑普遍存在无线网络适配器和有线网络适配器，因此就有两个 MAC 地址。

局域网

局域网是一种典型的广播信道，主要特点是网络为一个单位所拥有，且地理范围和站点数目均有限。

主要有以太网、令牌环网、FDDI 和 ATM 等局域网技术，目前以太网占领着有线局域网市场。

可以按照网络拓扑结构对局域网进行分类：



以太网（主流局域网技术）

以太网是一种星型拓扑结构局域网。

早期使用集线器进行连接，集线器是一种「物理层设备」，作用于比特而不是帧，当一个比特到达接口时，集线器重新生成这个比特，并将其能量强度放大，从而扩大网络的传输距离，之后再将这个比特发送到其它所有接口。如果集线器同时收到两个不同接口的帧，那么就发生了碰撞。

目前以太网使用交换机替代了集线器，交换机是一种「链路层设备」，它不会发生碰撞，能根据 MAC 地址进行「存储转发」。

以太网帧格式：

- **类型**：标记上层使用的协议；
- **数据**：长度在 46-1500 之间，如果太小则需要填充；
- **FCS**：帧检验序列，使用的是 CRC 检验方法；

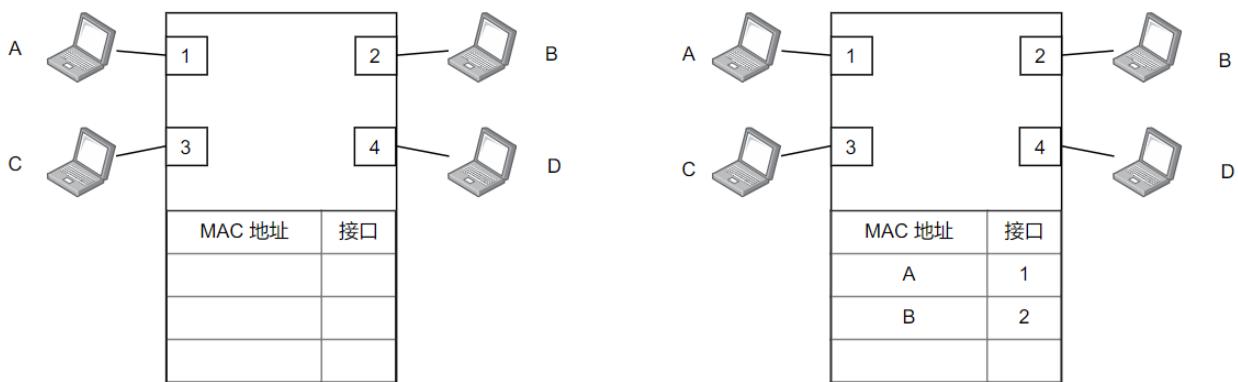
目的地址	源地址	类型	IP 数据报	FCS
------	-----	----	--------	-----

交换机

交换机具有自学习能力，学习的是交换表的内容，交换表中存储着 MAC 地址到接口的映射。

正是由于这种自学习能力，因此交换机是一种即插即用设备，不需要网络管理员手动配置交换表内容。

下图中，交换机有 4 个接口，主机 A 向主机 B 发送数据帧时，交换机把主机 A 到接口 1 的映射写入交换表中。为了发送数据帧到 B，先查交换表，此时没有主机 B 的表项，那么主机 A 就发送广播帧，主机 C 和主机 D 会丢弃该帧，主机 B 回应该帧向主机 A 发送数据包时，交换机查找交换表得到主机 A 映射的接口为 1，就发送数据帧到接口 1，同时交换机添加主机 B 到接口 2 的映射。

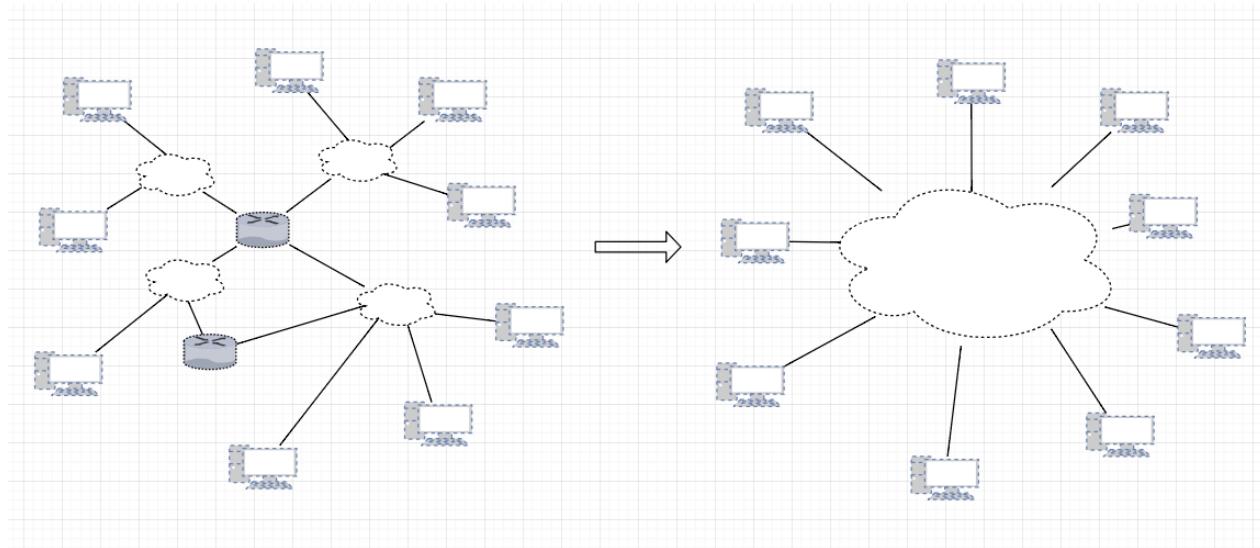


四、网络层

概述

因为网络层是整个互联网的核心，因此应当让网络层尽可能简单。网络层向上只提供简单灵活的、无连接的、尽最大努力交互的数据报服务。

使用 IP 协议，可以把异构的物理网络连接起来，使得在网络层看起来好像是一个统一的网络。



与 IP 协议配套使用的还有三个协议：

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

IP数据报格式

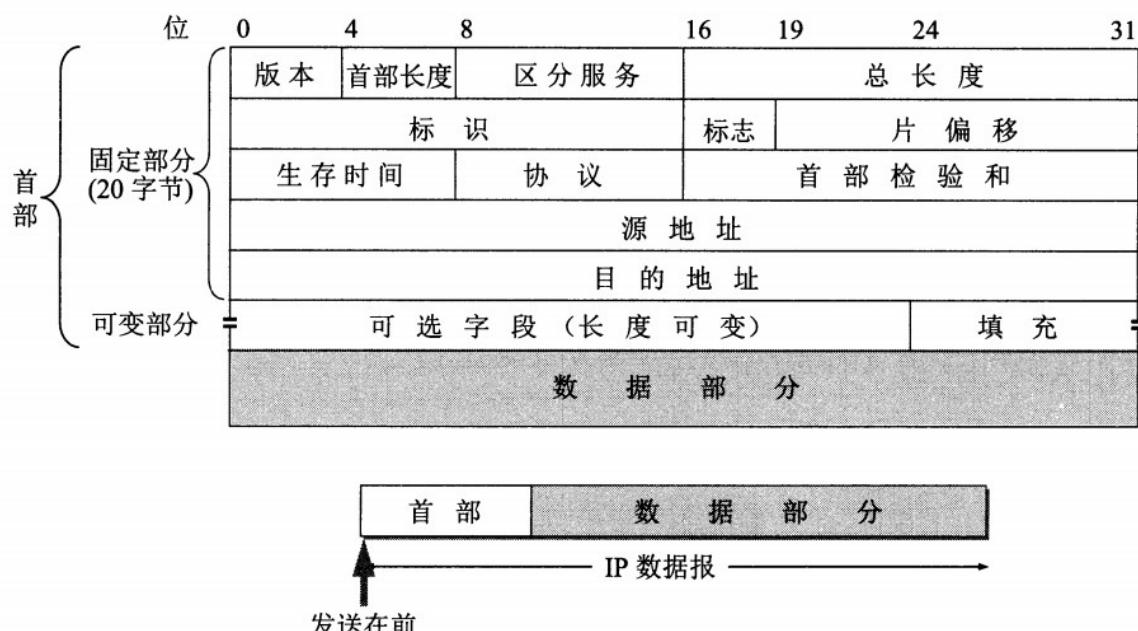


图 4-13 IP 数据报的格式

- 版本：有 4 (IPv4) 和 6 (IPv6) 两个值；
- 首部长度：占 4 位，因此最大值为 15。值为 1 表示的是 1 个 32 位字的长度，也就是 4 字节。因为首部固定长度为 20 字节，因此该值最小为 5。如果可选字段的长度不是 4 字节的整数倍，就用尾部的填充部分来填充。
- 区分服务：用来获得更好的服务，一般情况下不使用。
- 总长度：包括首部长度和数据部分长度。
- 生存时间：TTL (Time-To-Live)，它的存在是为了防止无法交付的数据报在互联网中不断兜圈子。以路由器跳数为单位，当 TTL 为 0 时就丢弃数据报。
- 协议：指出携带的数据应该上交给哪个协议进行处理，例如 ICMP、TCP、UDP 等。

- **首部检验和**：因为数据报每经过一个路由器，都要重新计算检验和，因此检验和不包含数据部分可以减少计算的工作量。
- **标识**：在数据报长度过长从而发生分片的情况下，相同数据报的不同分片具有相同的标识符。
- **片偏移**：和标识符一起，用于发生分片的情况。片偏移的单位为 8 字节。

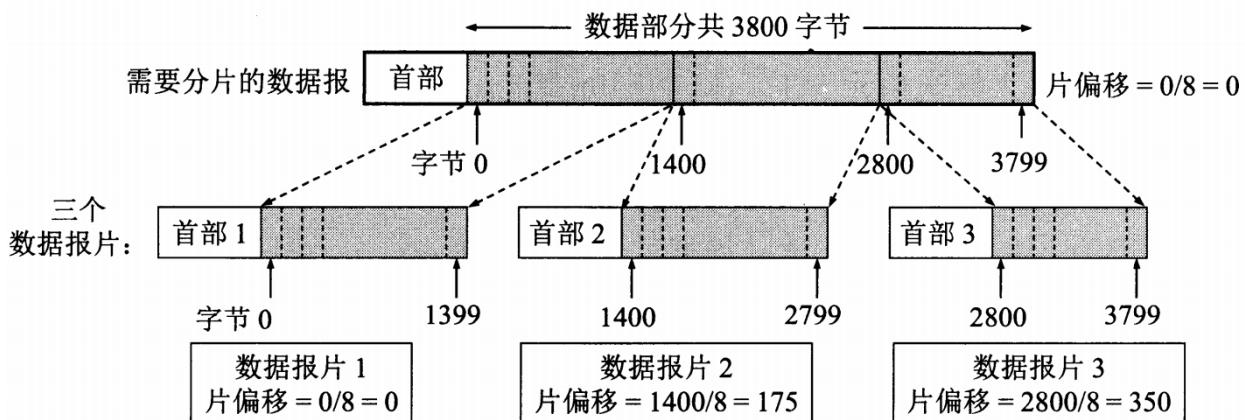


图 4-14 数据报的分片举例

IP地址编址方式

IP 地址的编址方式经历了三个历史阶段：

- 分类
- 子网划分
- 无分类

1. 分类

由两部分组成，网络号和主机号，其中不同分类具有不同的网络号长度，并且是固定的。

IP 地址 ::= {< 网络号 >, < 主机号 >}

- **IP 地址管理机构在分配 IP地址 时，分配网络号，而剩下的主机号由得到该网络号的单位自行分配。从而方便管理。**
- 路由器仅根据目的主机的网络号来转发分组，从而减小路由表所占用的存储空间以及查找路由表的时间。
- **主机号为全 0 的地址为子网的网络号，主机号为全 1 的地址为子网的广播地址，都不能被指派。**

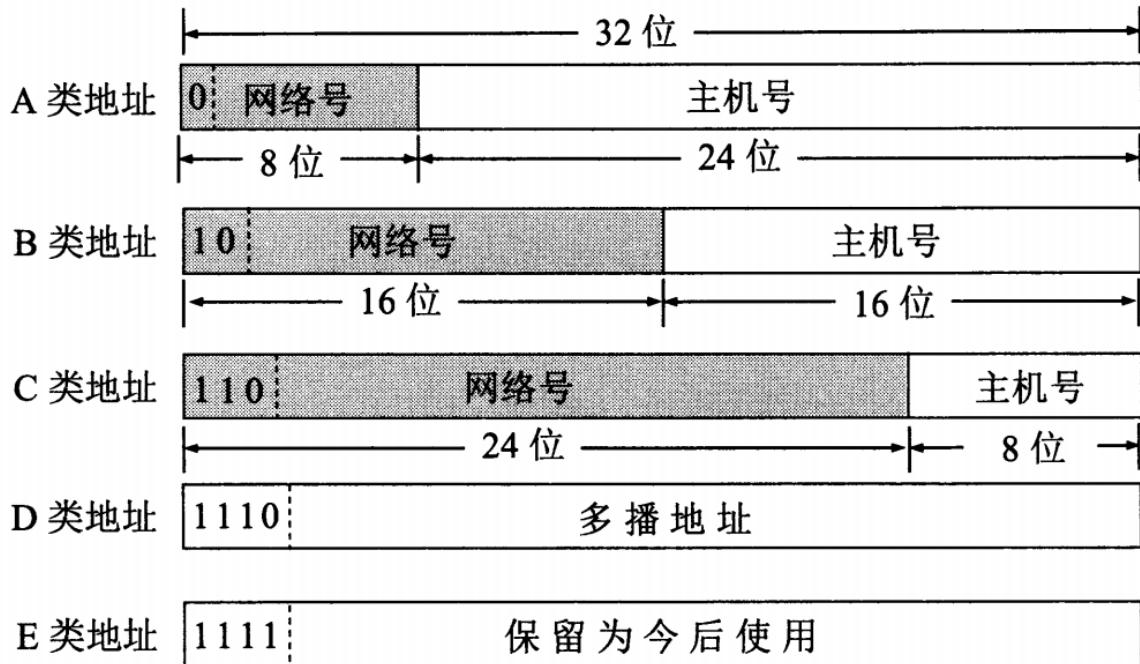


图 4-5 IP 地址中的网络号字段和主机号字段

可以看出，两级 IP 地址不够灵活，对 IP 地址空间的利用率比较低。如，C 类地址的局域网最多分配 254（0~255，去掉255广播和0这个网络号）个主机号，B 类地址的局域网最多分配 65534 个主机号。如果有单位有 255 台主机，则只能为其分配一个 B 类地址的 网络号。这样就会浪费很多 IP 地址。

https://blog.csdn.net/qq_36523667/article/details/79029794

2. 子网划分

通过在主机号字段中拿一部分作为子网号，把两级 IP 地址划分为三级 IP 地址。

IP 地址 ::= {< 网络号 >, < 子网号 >, < 主机号 >}

要使用子网，必须配置子网掩码。一个 B 类地址的默认子网掩码为 255.255.0.0，如果 B 类地址的子网占两个比特，那么子网掩码为 11111111 11111111 11000000 00000000，也就是 255.255.192.0。

注意，外部网络看不到子网的存在。

C类网络主机数

子网位数 子网掩码 主机数 可用主机数

1	255.255.255.128	128	126
2	255.255.255.192	64	62
3	255.255.255.224	32	30
4	255.255.255.240	16	14
5	255.255.255.248	8	6
6	255.255.255.252	4	2

3. 无分类

无分类编址 CIDR（无类别域间路由，Classless Inter-Domain Routing）消除了传统 A 类、B 类和 C 类地址以及划分子网的概念，使用网络前缀和主机号来对 IP 地址进行编码，网络前缀的长度可以根据需要变化。

IP 地址 ::= {< 网络前缀号 >, < 主机号 >}

CIDR 的记法上采用在 IP 地址后面加上网络前缀长度的方法，例如 128.14.35.7/20 表示前 20 位为网络前缀。

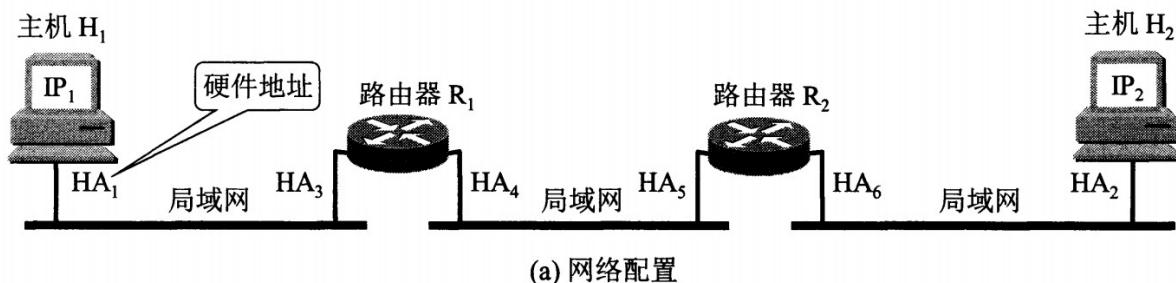
CIDR 的地址掩码可以继续称为子网掩码，子网掩码首 1 长度为网络前缀的长度。

一个 CIDR 地址块中有很多地址，一个 CIDR 表示的网络就可以表示原来的很多个网络，并且在路由表中只需要一个路由就可以代替原来的多个路由，减少了路由表项的数量。把这种通过使用网络前缀来减少路由表项的方式称为路由聚合，也称为 构成超网。

在路由表中的项目由“网络前缀”和“下一跳地址”组成，在查找时可能会得到不止一个匹配结果，应当采用最长前缀匹配来确定应该匹配哪一个。

地址解析协议ARP (Address Resolution Protocol, 因为MAC地址随链路改变而改变, IP->MAC)

网络层实现主机之间的通信，而链路层实现具体每段链路之间的通信。因此在通信过程中，IP 数据报的源地址和目的地址始终不变，而 MAC 地址随着链路的改变而改变。



ARP 实现由 IP 地址得到 MAC 地址。

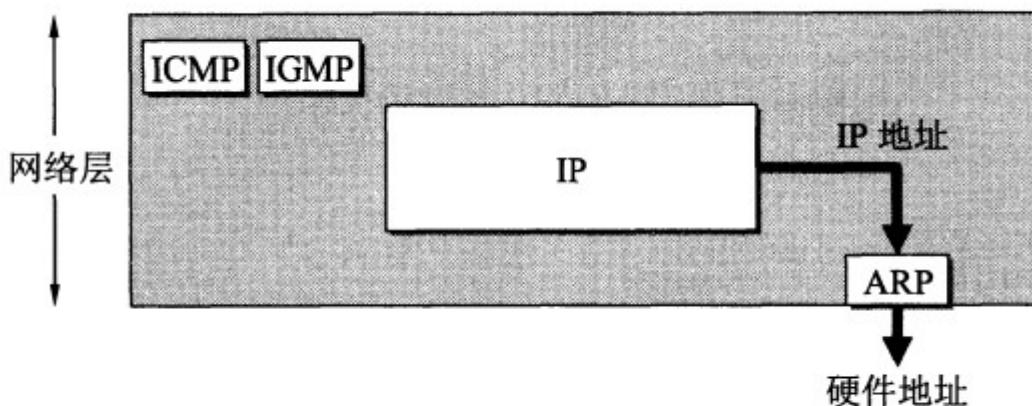


图 4-10 ARP 协议的作用

每个主机都有一个 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 MAC 地址的映射表。

如果主机 A 知道主机 B 的 IP 地址，但是 ARP 高速缓存中没有该 IP 地址到 MAC 地址的映射，此时主机 A 通过广播的方式发送 ARP 请求分组，主机 B 收到该请求后会发送 ARP 响应分组给主机 A 告知其 MAC 地址，随后主机 A 向其高速缓存中写入主机 B 的 IP 地址到 MAC 地址的映射。

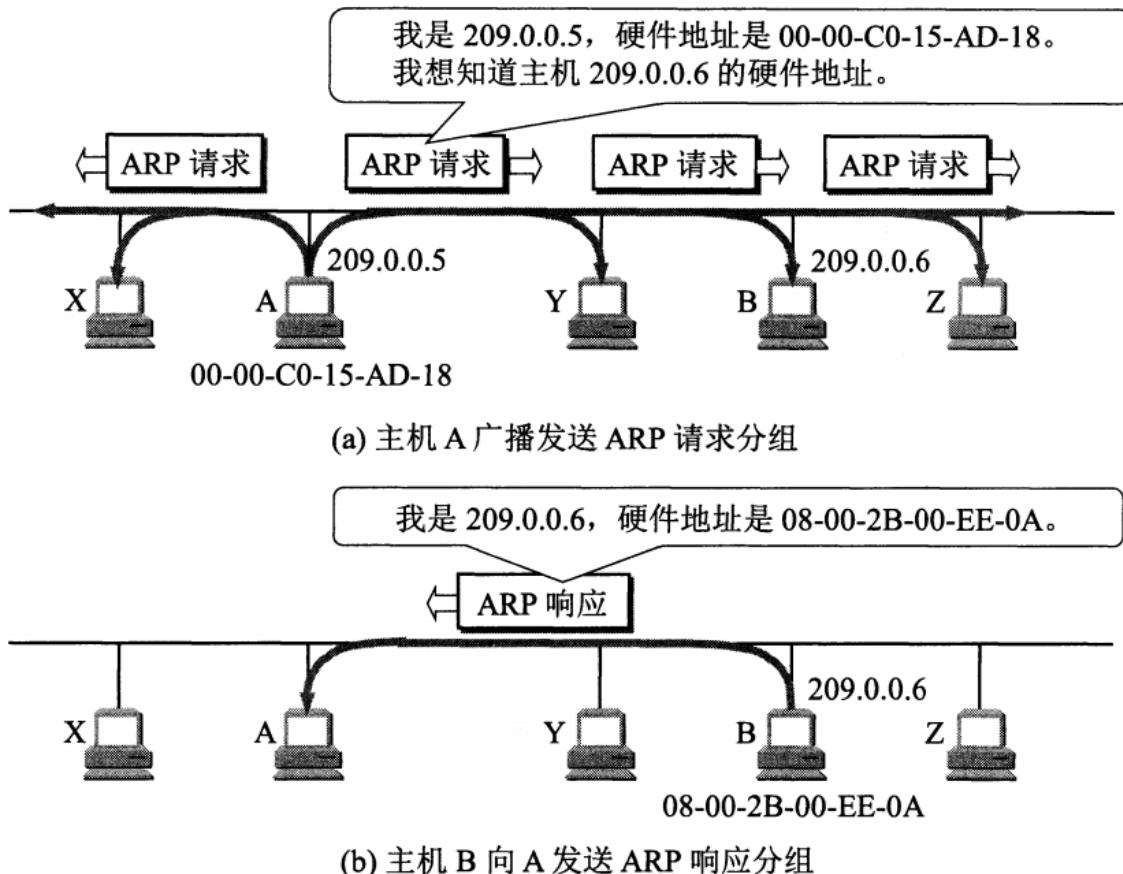


图 4-11 地址解析协议 ARP 的工作原理

网际控制报文协议ICMP (Internet Control Messages Protocol)

ICMP 是为了更有效地转发 IP 数据报和提高交付成功的机会。它「封装在 IP 数据报中」，但是「不属于高层协议」。

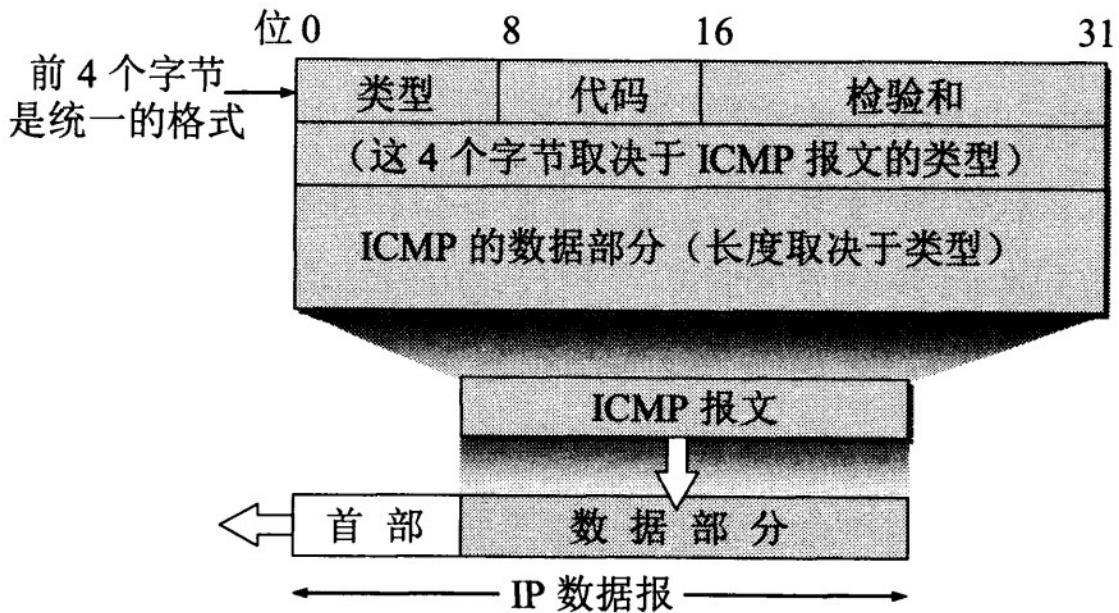


图 4-27 ICMP 报文的格式

ICMP 报文分为差错报告报文和询问报文。

表 4-8 几种常用的 ICMP 报文类型

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

1. Ping

Ping 是 ICMP 的一个重要应用，主要用来测试两台主机之间的连通性。

Ping 的原理是通过向目的主机发送 ICMP Echo 请求报文，目的主机收到之后会发送 Echo 回答报文。Ping 会根据时间和成功响应的次数估算出数据包往返时间以及丢包率。

2. Traceroute (路由跟踪，通过不停修改TTL发送无法交付UDP来跟踪路由路径、往返时间)

Traceroute 是 ICMP 的另一个应用，用来跟踪一个分组从源点到终点的路径。

Traceroute 发送的 IP 数据报封装的是无法交付的 UDP 用户数据报，并由目的主机发送终点不可达差错报告报文。

- 源主机向目的主机发送一连串的 IP 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，R1 收下它并把 TTL 减 1，此时 TTL 等于 0，R1 就把 P1 丢弃，

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 R1，R1 收下后把 TTL 减 1 再转发给 R2，R2 收下后也把 TTL 减 1，由于此时 TTL 等于 0，R2 就丢弃 P2，并向源主机发送一个 ICMP 时间超过差错报文。
- 不断执行这样的步骤，直到最后一个数据报刚刚到达目的主机，主机不转发数据报，也不把 TTL 值减 1。但是因为数据报封装的是无法交付的 UDP，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。
- 之后源主机知道了到达目的主机所经过的路由器 IP 地址以及到达每个路由器的往返时间。

虚拟专用网VPN（Virtual Private Network）

由于 IP 地址的紧缺，一个机构能申请到的 IP 地址数往往远小于本机构所拥有的主机数。并且一个机构并不需要把所有的主机接入到外部的互联网中，机构内的计算机可以使用仅在本机构有效的 IP 地址（专用地址）。

有三个专用地址块：

- 10.0.0.0 ~ 10.255.255.255
- 172.16.0.0 ~ 172.31.255.255
- 192.168.0.0 ~ 192.168.255.255

VPN 使用「公用的互联网」作为本机构各专用网之间的通信载体。专用指机构内的主机只与本机构内的其它主机通信；虚拟指好像是专用网络，而实际上并不是，它有经过公用的互联网。

下图中，场所 A 和 B 的通信经过互联网，如果场所 A 的主机 X 要和另一个场所 B 的主机 Y 通信，IP 数据报的源地址是 10.1.0.1，目的地址是 10.2.0.3。数据报先发送到与互联网相连的路由器 R1，R1 对内部数据进行加密，然后重新加上数据报的首部，源地址是路由器 R1 的全球地址 125.1.2.3，目的地址是路由器 R2 的全球地址 194.4.5.6。路由器 R2 收到数据报后将数据部分进行解密，恢复原来的数据报，此时目的地址为 10.2.0.3，就交付给 Y。

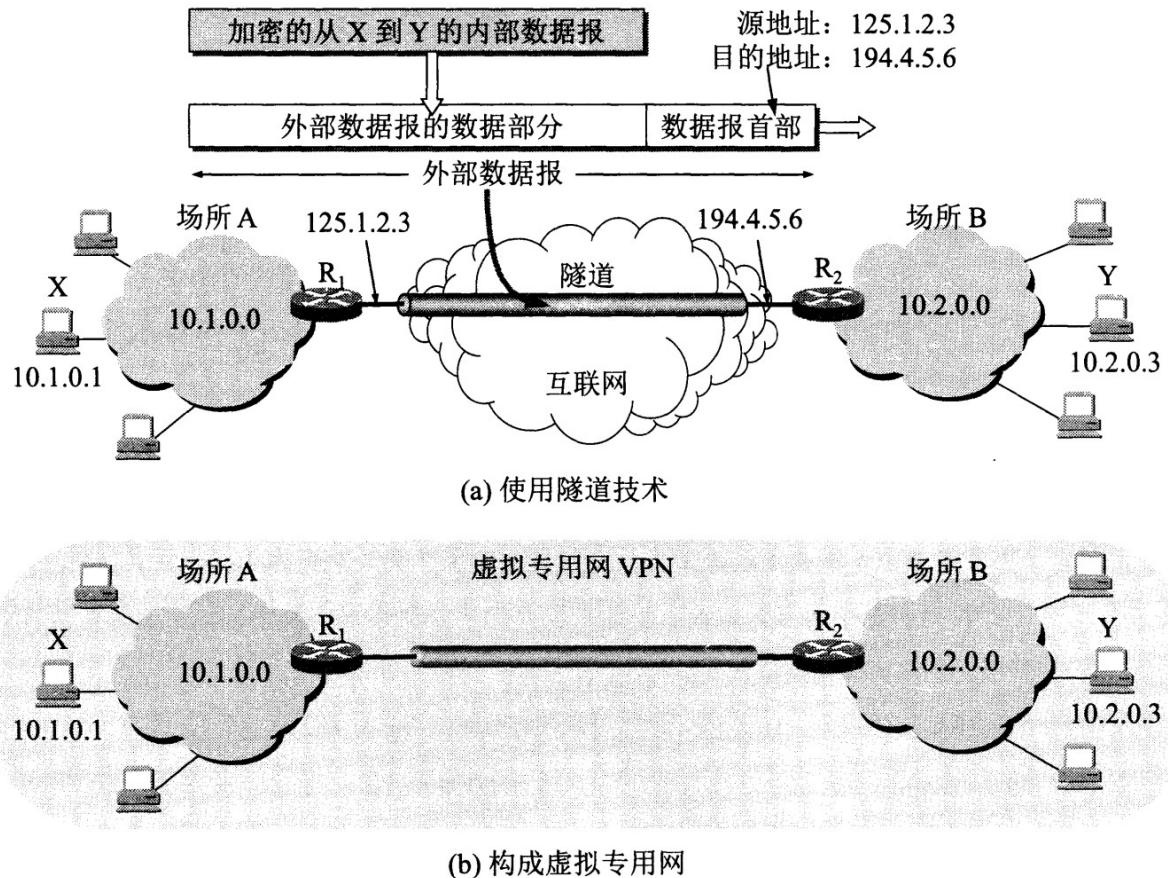


图 4-59 用隧道技术实现虚拟专用网

网络地址转换 (Network Address Translation, Local IP -> Universal IP)

专用网内部的主机使用本地 IP 地址又想和互联网上的主机通信时，可以使用 NAT 来将本地 IP 转换为全球 IP。

在以前，NAT 将本地 IP 和全球 IP 一一对应，这种方式下拥有 n 个全球 IP 地址的专用网内最多只可以同时有 n 台主机接入互联网。为了更有效地利用全球 IP 地址，现在常用的 NAT 转换表把传输层的端口号也用上了，使得多个专用网内部的主机共用一个全球 IP 地址。使用端口号的 NAT 也叫做网络地址与端口转换 NAPT。

路由器的结构

路由器从功能上可以划分为：路由选择和分组转发。

分组转发结构由三个部分组成：交换结构、一组输入端口和一组输出端口。

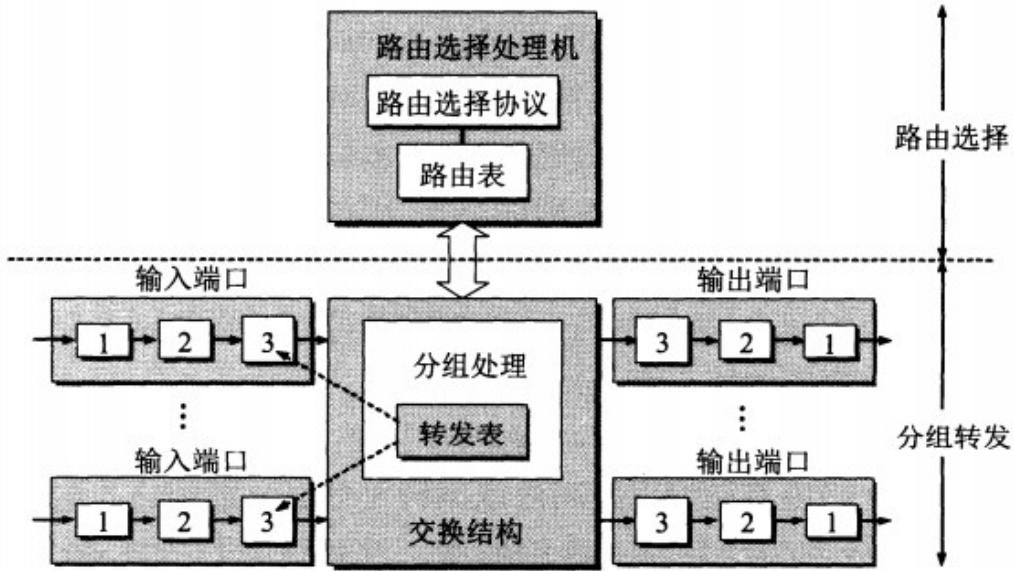


图 4-42 典型的路由器的结构（图中的数字 1~3 表示相应层次的构件）

路由表的结构（注意：路由表主机和路由器都有！！！）

可以使用`netstat -rn`命令查看本机路由表

Routing Table: IPv4						
Destination	Gateway	Flags	Ref	Use	Interface	
default	172.20.1.10	UG	1	532	ce0	
224.0.0.0	10.0.5.100	U	1	0	bge0	
10.0.0.0	10.0.5.100	U	1	0	bge0	
127.0.0.1	127.0.0.1	UH	1	57	lo0	

Destination: 主机地址或网络地址

Gateway: 网关地址

Flags:

G和H十分重要，G区分间接路由和直接路由。H区分目的地址是网络地址（主机号可以为0）还是主机地址

U 该路由可以使用

G 该路由是到一个网关（路由器）。如果没有设置该标志，说明目的地是直接相连的

H 该路由是到一个主机，也就是说，目的地址是一个完整的主机地址。如果没有设置该标志，说明该路由是到一个网络，而目的地址是一个网络地址：一个网络号，或者网络号与子网号的组合（CIDR）。

D 该路由是由重定向报文创建的

M 该路由已被重定向报文修改

Refcnt (Reference count) : 列给出的是正在使用路由的活动进程个数。面向连接的协议如TCP在建立连接时要固定路由。如果在主机sr4和slip之间建立Telnet连接，可以看到参考记数值变为1。建立另一个Telnet连接时，它的值将增加为2，依此类推。

Use: 显示的是通过该路由发送的分组数。如果我们是这个路由的唯一用户，那么运行ping程序发送5个分组后，它的值将变为5。

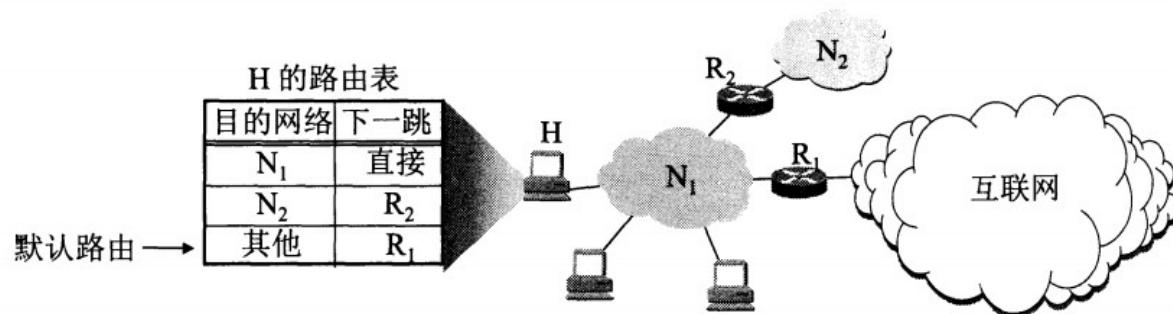
Interface: 本地接口的名字。环回接口的名字始终为lo0

静态路由的系统不依赖于路由协议来获取路由信息及更新路由表

动态路由的系统依赖路由协议（如用于 IPv4 网络的 RIP 和用于 IPv6 网络的 RIPng）来维护其路由表

路由器分组转发

- 从数据报的首部提取目的主机的 IP 地址 D，得到目的网络地址 N。
- 若 N 就是与此路由器直接相连的某个网络地址，则进行直接交付；
- 若路由表中有目的地址为 D 的特定主机路由，则把数据报传送给表中所指明的下一跳路由器；
- 若路由表中有到达网络 N 的路由，则把数据报传送给路由表中所指明的下一跳路由器；
- 若路由表中有一个默认路由，则把数据报传送给路由表中所指明的默认路由器；
- 报告转发分组出错。



路由选择协议

路由选择协议都是自适应的，能随着网络通信量和拓扑结构的变化而自适应地进行调整。

互联网可以划分为许多较小的自治系统 AS，一个 AS 可以使用一种和别的 AS 不同的路由选择协议。

可以把路由选择协议划分为两大类：

- 自治系统内部的路由选择：RIP 和 OSPF
- 自治系统间的路由选择：BGP

1. 路由信息协议，内部网关协议 (RIP, Routing Information Protocol)

RIP 是一种基于距离向量的路由选择协议。距离是指跳数，直接相连的路由器跳数为 1。跳数最多为 15，超过 15 表示不可达。

RIP 按固定的时间间隔仅和相邻路由器交换自己的路由表，经过若干次交换之后，所有路由器最终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地址。

距离向量算法：

- 对地址为 X 的相邻路由器发来的 RIP 报文，先修改报文中的所有项目，把下一跳字段中的地址改为 X，并把所有的距离字段加 1；
- 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：若下一跳路由器地址是 X，则把收到的项目替换原来路由表中的项目；否则：若收到的项目的距离 d 小于路由表中的距离，则进行更新（例如原始路由表项为 Net2, 5, P，新表项为 Net2, 4, X，则更新）；否则什么也不做。
- 若 3 分钟还没有收到相邻路由器的更新路由表，则把该相邻路由器标为不可达，即把距离置为 16。

RIP 协议实现简单，开销小。但是 RIP 能使用的最大距离为 15，限制了网络的规模。并且当网络出现故障时，要经过比较长的时间才能将此消息传送到所有路由器。

2. 开放最短路径优先 (OSPF, Open Shortest Path First)

开放最短路径优先 OSPF，是为了克服 RIP 的缺点而开发出来的。

开放表示 OSPF 不受某一家厂商控制，而是公开发表的；最短路径优先表示使用了 Dijkstra 提出的最短路径算法 SPF。

OSPF 具有以下特点：

- 向本自治系统中的所有路由器发送信息，这种方法是洪泛法。
- 发送的信息就是与相邻路由器的链路状态，链路状态包括与哪些路由器相连以及链路的度量，度量用费用、距离、时延、带宽等来表示。
- 只有当链路状态发生变化时，路由器才会发送信息。

所有路由器都具有全网的拓扑结构图，并且是一致的。相比于 RIP，OSPF 的更新过程收敛的很快。

3. 外部网关协议(BGP, Border Gateway Protocol)

BGP，边界网关协议)

AS 之间的路由选择很困难，主要是由于：

- 互联网规模很大；
- 各个 AS 内部使用不同的路由选择协议，无法准确定义路径的度量；
- AS 之间的路由选择必须考虑有关的策略，比如有些 AS 不愿意让其它 AS 经过。

BGP 只能寻找一条比较好的路由，而不是最佳路由。

每个 AS 都必须配置 BGP 发言人，通过在两个相邻 BGP 发言人之间建立 TCP 连接来交换路由信息。

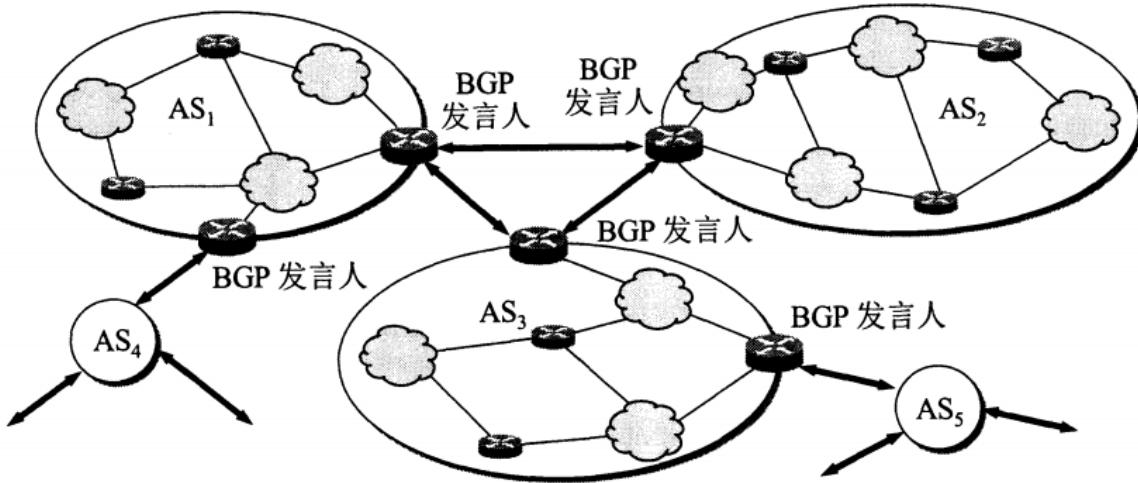


图 4-38 BGP 发言人和自治系统 AS 的关系

五、传输层

网络层只把分组发送到目的主机，但是真正通信的并不是主机而是主机中的进程。传输层提供了进程间的逻辑通信，传输层向高层用户屏蔽了下面网络层的核心细节，使应用程序看起来像是在两个传输层实体之间有一条端到端的逻辑通信信道。

UDP和TCP特点

- 用户数据报协议 UDP (User Datagram Protocol) 是无连接的，尽最大可能交付，没有拥塞控制，面向报文（对于应用程序传下来的报文不合并也不拆分，只是添加 UDP 首部），支持一对一、一对多、多对一和多对多的交互通信。最大只支持 512 字节
- 传输控制协议 TCP (Transmission Control Protocol) 是面向连接的，提供可靠交付，有流量控制，拥塞控制，提供全双工通信，面向字节流（把应用层传下来的报文看成字节流，把字节流组织成大小不等的数据块），每一条 TCP 连接只能是点对点的（一对一）。

UDP首部格式

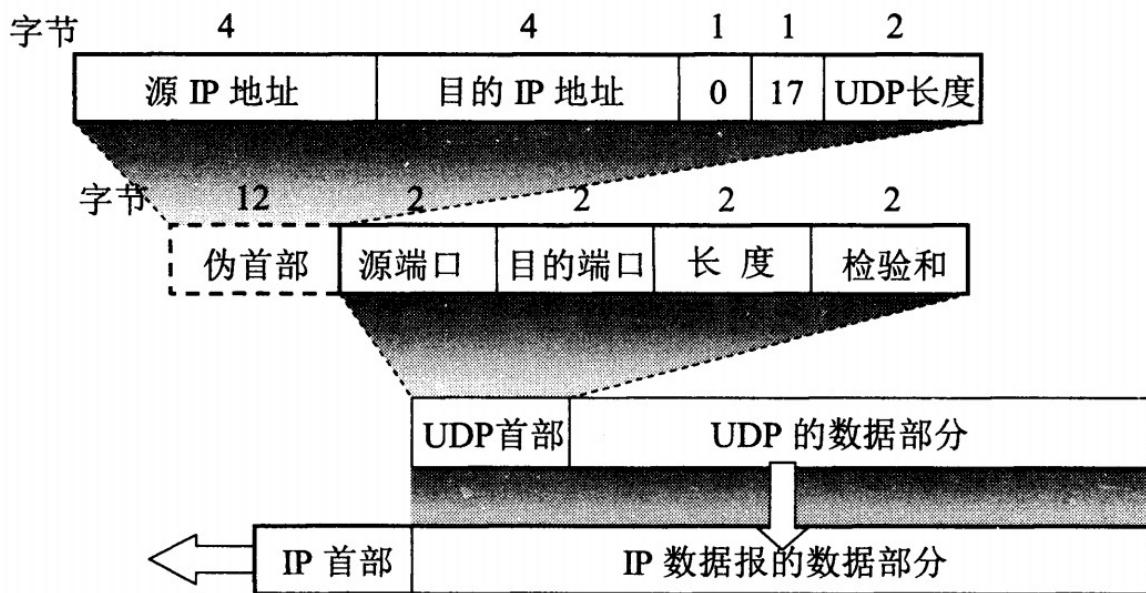


图 5-5 UDP 用户数据报的首部和伪首部

首部字段只有 8 个字节，包括源端口、目的端口、长度、检验和。12 字节的伪首部是为了计算检验和临时添加的。

TCP首部格式

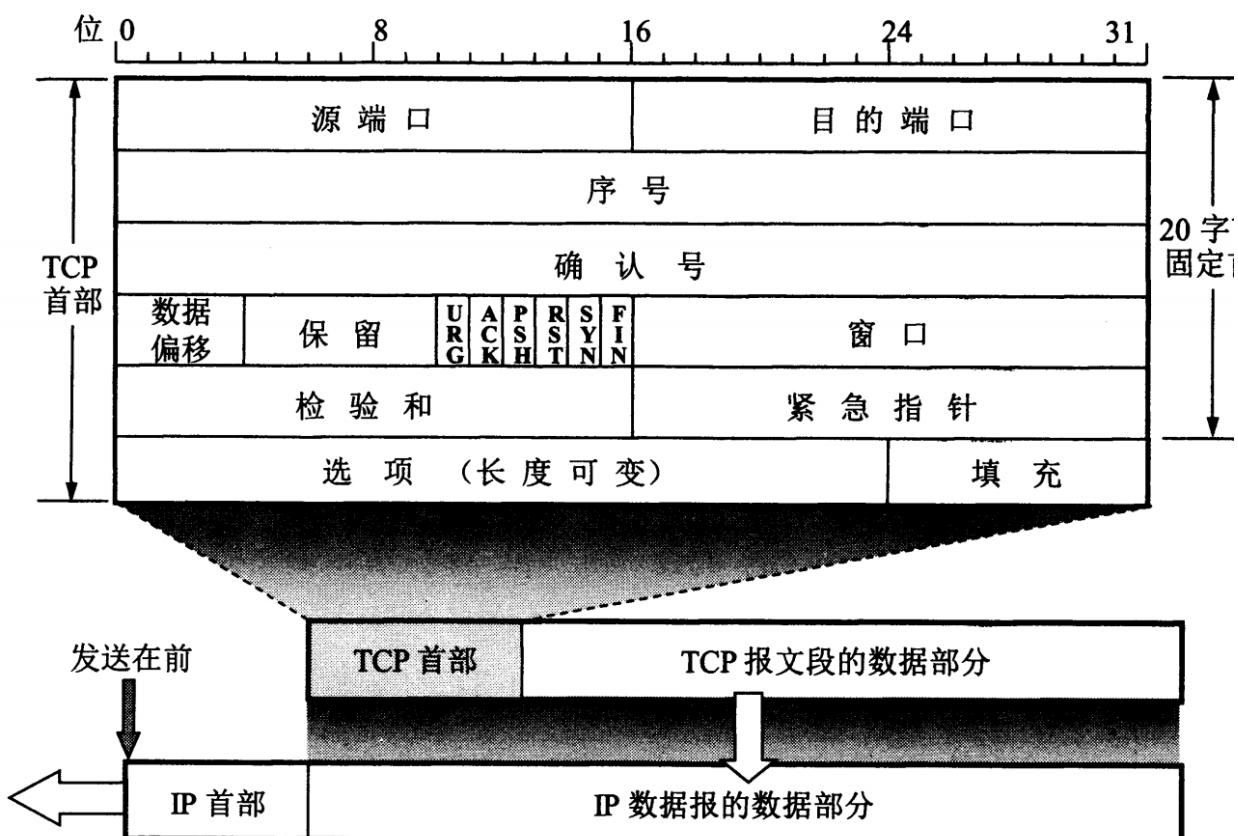


图 5-14 TCP 报文段的首部格式

- 序号：用于对字节流进行编号，例如序号为 301，表示第一个字节的编号为 301，如果携带的数据长度为 100 字节，那么下一个报文段的序号应为 401。
- 确认号：期望收到的下一个报文段的序号。例如 B 正确收到 A 发送来的一个报文段，序号为 501，携带的数据长度为 200 字节，因此 B 期望下一个报文段的序号为 701，B 发送给 A 的确认报文段中确认号就为 701。
- 数据偏移：指的是数据部分距离报文段起始处的偏移量，实际上指的是首部的长度。
- 确认 ACK：当 ACK=1 时确认号字段有效，否则无效。TCP 规定，在连接建立后所有传送的报文段都必须把 ACK 置 1。
- 同步 SYN：在连接建立时用来同步序号。当 SYN=1，ACK=0 时表示这是一个连接请求报文段。若对方同意建立连接，则响应报文中 SYN=1，ACK=1。
- 终止 FIN：用来释放一个连接，当 FIN=1 时，表示此报文段的发送方的数据已发送完毕，并要求释放连接。
- 窗口：窗口值作为接收方让发送方设置其发送窗口的依据。之所以要有这个限制，是因为接收方的数据缓存空间是有限的。

TCP三次握手

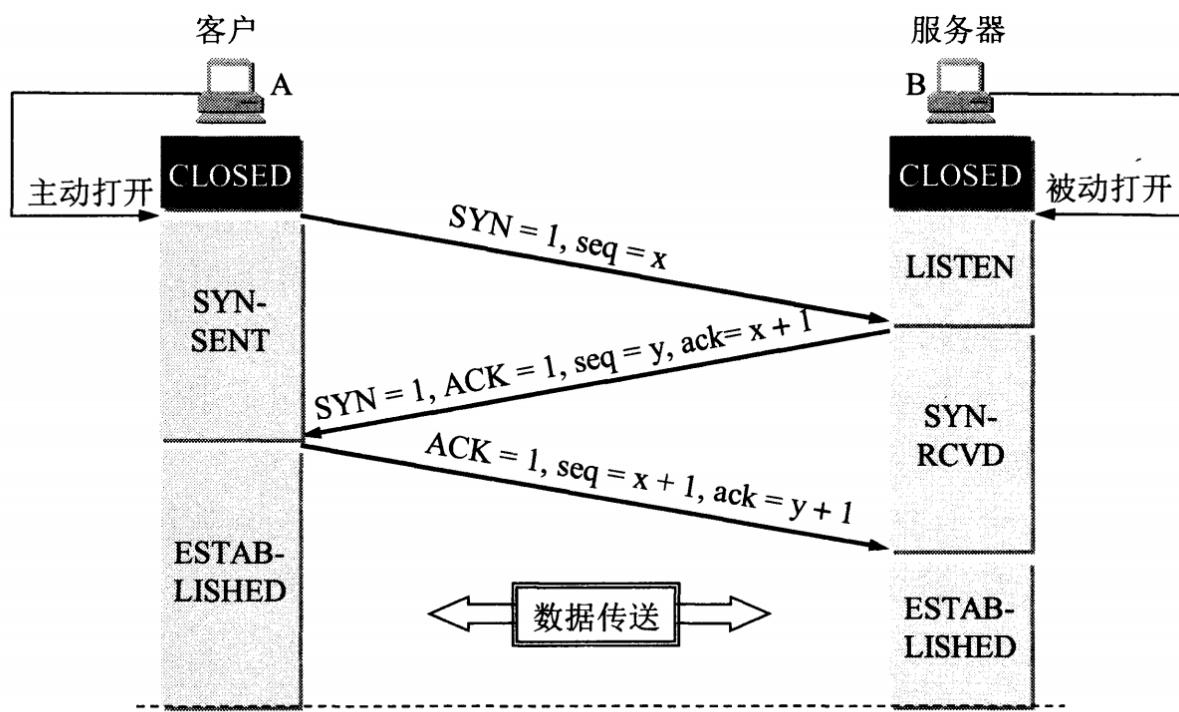


图 5-28 用三报文握手建立 TCP 连接

假设 A 为客户端，B 为服务器端。

- 首先 B 处于 LISTEN（监听）状态，等待客户的连接请求。
- A 向 B 发送连接请求报文，SYN=1，ACK=0，选择一个初始的序号 x 。
- B 收到连接请求报文，如果同意建立连接，则向 A 发送连接确认报文，SYN=1，ACK=1，确认号为 $x+1$ ，同时也选择一个初始的序号 y 。
- A 收到 B 的连接确认报文后，还要向 B 发出确认，确认号为 $y+1$ ，序号为 $x+1$ 。
- B 收到 A 的确认后，连接建立。

三次握手的原因

第三次握手是为了防止失效的连接请求到达服务器，让服务器错误打开连接。

客户端发送的连接请求如果在网络中滞留，那么就会隔很长一段时间才能收到服务器端发回的连接确认。客户端等待一个超时重传时间之后，就会重新请求连接。但是这个滞留的连接请求最后还是会到达服务器，如果不进行三次握手，那么服务器就会打开两个连接。如果有第三次握手，客户端会忽略服务器之后发送的对滞留连接请求的连接确认，不进行第三次握手，因此就不会再次打开连接。

TCP的四次挥手

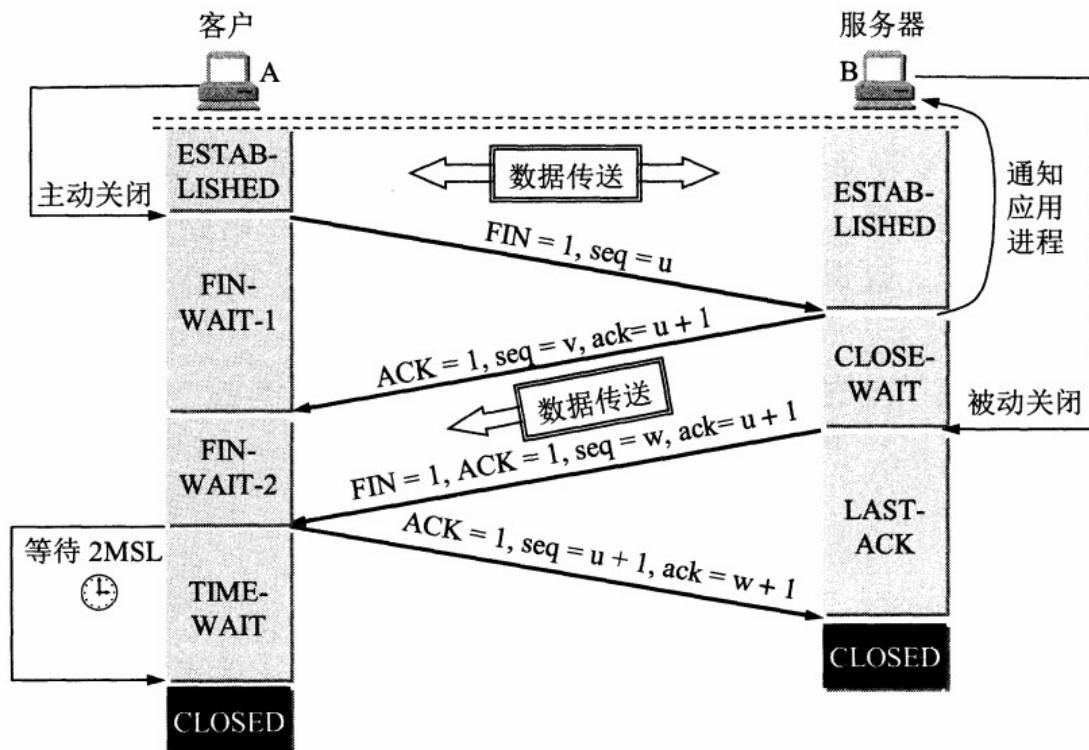


图 5-29 TCP 连接释放的过程

以下描述不讨论序号和确认号，因为序号和确认号的规则比较简单。并且不讨论 ACK，因为 ACK 在连接建立之后都为 1。

- A 发送连接释放报文，FIN=1。
- B 收到之后发出确认，此时 TCP 属于半关闭状态，B 能向 A 发送数据但是 A 不能向 B 发送数据。
- 当 B 不再需要连接时，发送连接释放报文，FIN=1。
- A 收到后发出确认，进入 TIME-WAIT 状态，等待 2 MSL (Maximum Segment Lifetime, 最大报文存活时间，一个报文的发送和收到回复的最大时间) 后释放连接。
- B 收到 A 的确认后释放连接。

四次挥手的原因

客户端发送了 FIN 连接释放报文之后，服务器收到了这个报文，就进入了 CLOSE-WAIT 状态。这个状态是为了让服务器端发送还未传送完毕的数据，传送完毕之后，服务器会发送 FIN 连接释放报文。

TIME_WAIT

客户端接收到服务器端的 FIN 报文后进入此状态，此时并不是直接进入 CLOSED 状态，还需要等待一个时间计时器设置的时间 2MSL。这么做有两个理由：

- 确保最后一个确认报文能够到达。如果 B 没收到 A 发送来的确认报文，那么就会重新发送连接释

放请求报文，如果A重新发来FIN报文，说明最后一个确认报文没到达，则B再次发确认报文，并再等待2 MSL，**A**等待一段时间就是为了处理这种情况的发生。

- 等待一段时间是为了让本连接持续时间内所产生的所有报文都从网络中消失，使得下一个新的连接不会出现旧的连接请求报文。

TCP可靠传输

TCP 使用超时重传来实现可靠传输：如果一个已经发送的报文段在超时时间内没有收到确认，那么就重传这个报文段。

一个报文段从发送再到接收到确认所经过的时间称为往返时间 RTT

TCP滑动窗口

窗口是缓存的一部分，用来暂时存放字节流。发送方和接收方各有一个窗口，接收方通过 TCP 报文段中的窗口字段告诉发送方自己的窗口大小，发送方根据这个值和其它信息设置自己的窗口大小。

发送窗口内的字节都允许被发送，接收窗口内的字节都允许被接收。如果发送窗口左部的字节已经发送并且收到了确认，那么就将发送窗口向右滑动一定距离，直到左部第一个字节不是已发送并且已确认的状态；接收窗口的滑动类似，接收窗口左部字节已经发送确认并交付主机，就向右滑动接收窗口。

按序接受：接收窗口只会对窗口内最后一个按序到达的字节进行确认，例如接收窗口已经收到的字节为 {31, 34, 35}，其中 {31} 按序到达，而 {34, 35} 就不是，因此只对字节 31 进行确认。发送方得到一个字节的确认之后，就知道这个字节之前的所有字节都已经被接收。

TCP拥塞控制

如果网络出现拥塞，分组将会丢失，此时发送方会继续重传，从而导致网络拥塞程度更高。因此当出现拥塞时，应当控制发送方的速率。这一点和流量控制很像，但是出发点不同。流量控制是为了让接收方能来得及接收，而拥塞控制是为了降低整个网络的拥塞程度。

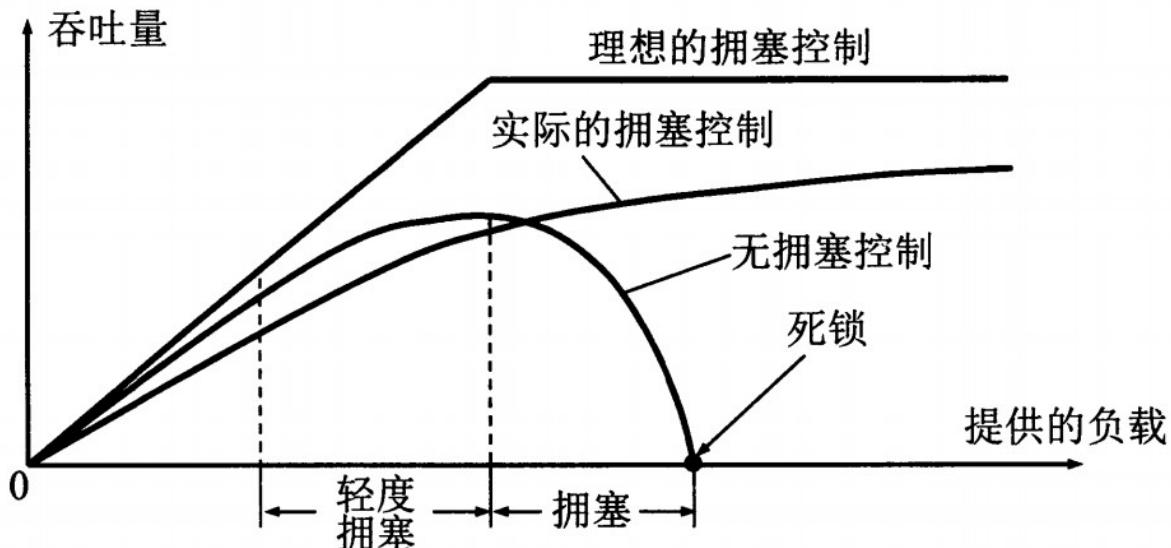


图 5-23 拥塞控制所起的作用

TCP 主要通过四个算法来进行拥塞控制：慢开始、拥塞避免、快重传、快恢复。

发送方需要维护一个叫做拥塞窗口（**cwnd**, **congestion window**）的状态变量，注意拥塞窗口与发送方窗口的区别：拥塞窗口只是一个状态变量，实际决定发送方能发送多少数据的是发送方窗口。

为了便于讨论，做如下假设：

- 接收方有足够的接收缓存，因此不会发生流量控制；
- 虽然 TCP 的窗口基于字节，但是这里设窗口的大小单位为报文段。

1. 慢开始与拥塞避免

发送的最初执行慢开始，令 $cwnd = 1$ ，发送方只能发送 1 个报文段；当收到确认后，将 $cwnd$ 加倍，因此之后发送方能够发送的报文段数量为：2、4、8 ...

注意到慢开始每个轮次都将 $cwnd$ 加倍，这样会让 $cwnd$ 增长速度非常快，从而使得发送方发送的速度增长速度过快，网络拥塞的可能性也就更高。设置一个慢开始门限 **ssthresh** (**Slow-Start Threshold**)，当 $cwnd \geq ssthresh$ 时，进入拥塞避免，每个轮次只将 $cwnd$ 加 1。

如果出现了超时，则令 $ssthresh = cwnd / 2$ ，然后重新执行慢开始。

2. 快重传与快恢复

在接收方，要求每次接收到报文段都应该对最后一个已收到的有序报文段进行确认。例如已经接收到 M_1 和 M_2 ，此时收到 M_4 ，应当发送对 M_2 的确认。

在发送方，如果收到三个重复确认，那么可以知道下一个报文段丢失，此时执行快重传，立即重传下一个报文段。例如收到三个 M_2 ，则 M_3 丢失，立即重传 M_3 。

在这种情况下，只是丢失个别报文段，而不是网络拥塞。因此执行快恢复，令 $ssthresh = cwnd / 2$ ， $cwnd = ssthresh$ ，注意到此时直接进入拥塞避免。

慢开始和快恢复的快慢指的是 $cwnd$ 的设定值，而不是 $cwnd$ 的增长速率。慢开始 $cwnd$ 设定为 1，而快恢复 $cwnd$ 设定为 $ssthresh$ 。

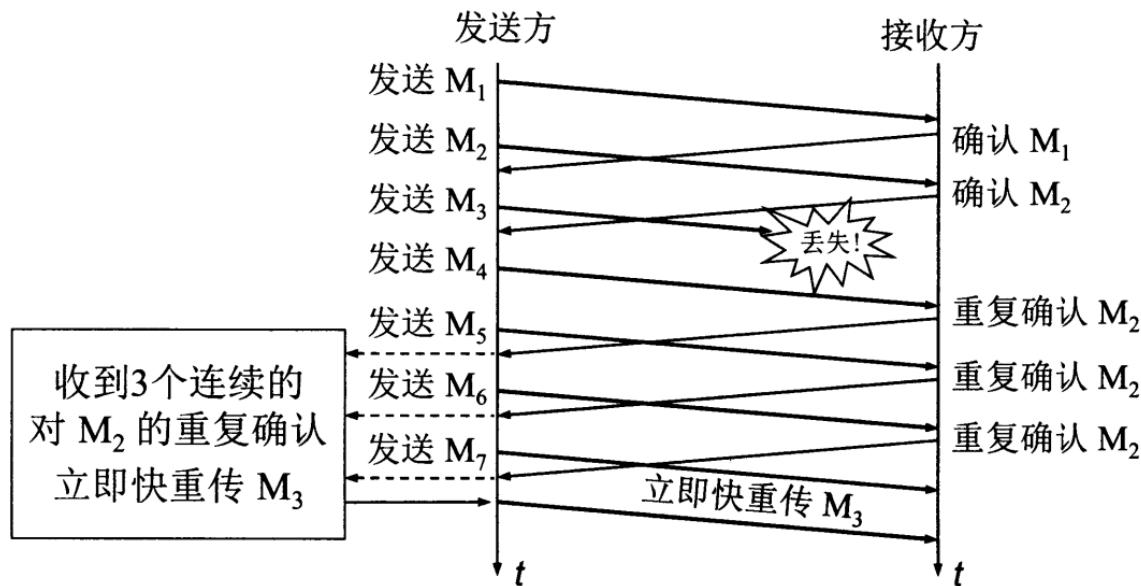


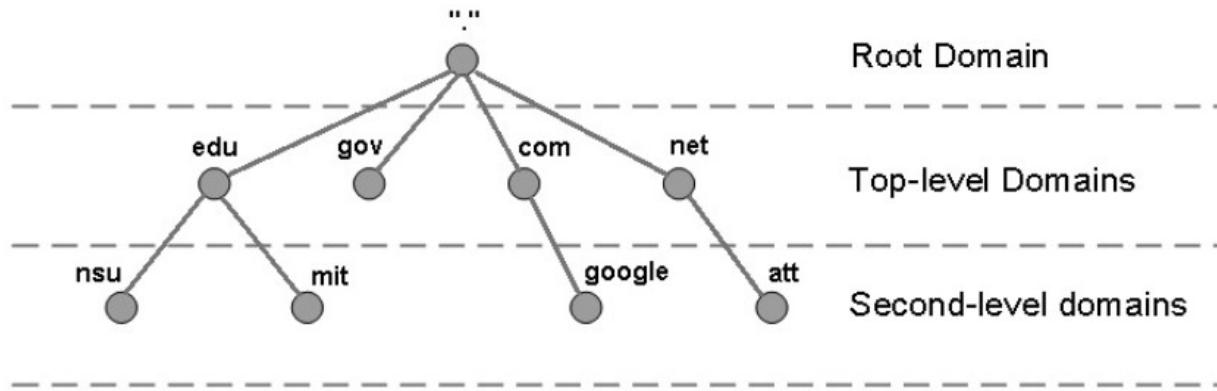
图 5-26 快重传的示意图

六、应用层

域名系统DNS (Domain Name System, 主机名->IP)

DNS 是一个分布式数据库，提供了主机名和 IP 地址之间相互转换的服务。这里的分布式数据库是指，每个站点只保留它自己的那部分数据。

域名具有层次结构，从上到下依次为：根域名、顶级域名、二级域名。



DNS 可以使用 UDP 或者 TCP 进行传输，使用的端口号都为 53。大多数情况下 DNS 使用 UDP 进行传输，这就要求域名解析器和域名服务器都必须自己处理超时和重传来保证可靠性。在两种情况下会使用 TCP 进行传输：

- 如果返回的响应超过的 512 字节（UDP 最大只支持 512 字节的数据）。
- 区域传送（区域传送是主域名服务器向辅助域名服务器传送变化的那部分数据）。

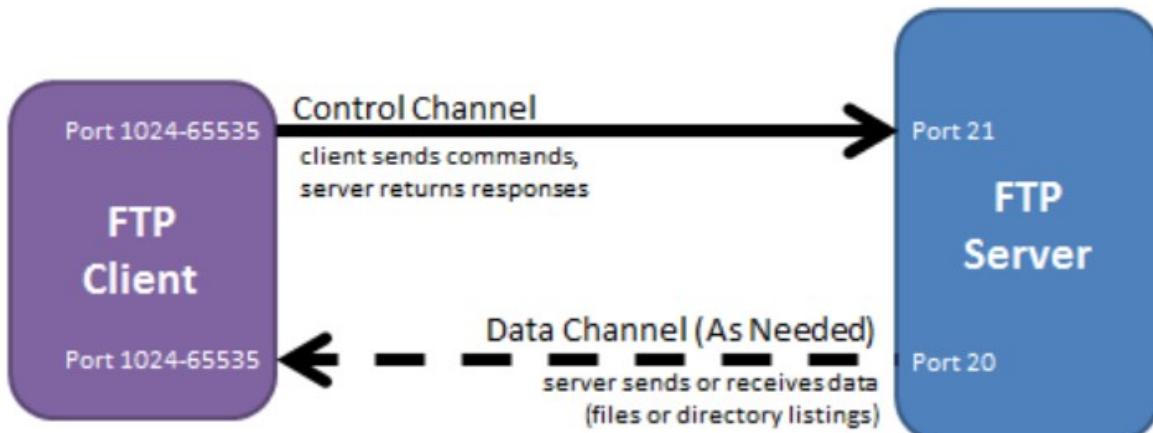
文件传送协议 (FTP)

FTP 使用 TCP 进行连接，它需要两个连接来传送一个文件：

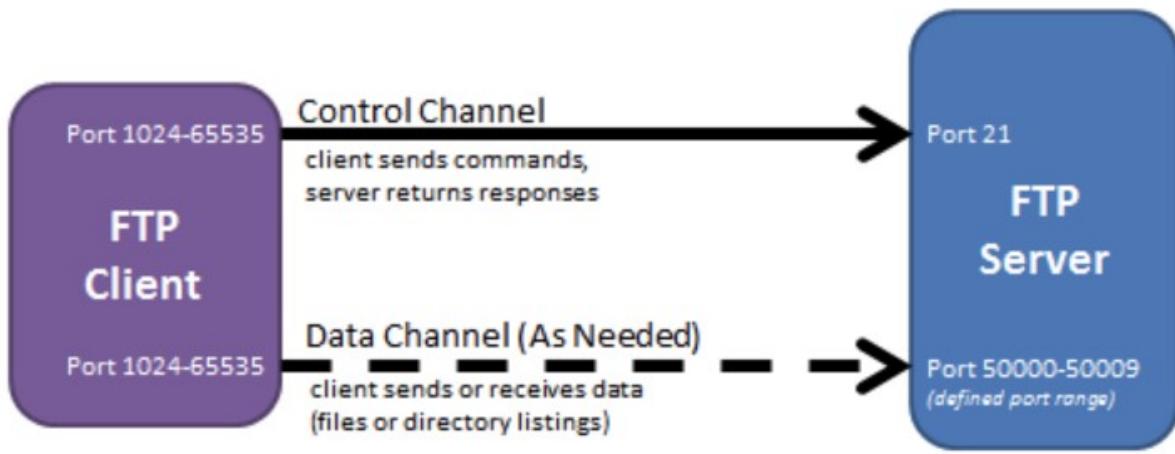
- 控制连接：服务器打开端口号 21 等待客户端的连接，客户端主动建立连接后，使用这个连接将客户端的命令传送给服务器，并传回服务器的应答。
- 数据连接：用来传送一个文件数据。

根据数据连接是否是服务器端主动建立，FTP 有主动和被动两种模式：

- 主动模式：服务器端主动建立数据连接，其中服务器端的端口号为 20，客户端的端口号随机，但是必须大于 1024，因为 0~1023 是熟知端口号。



- 被动模式：客户端主动建立数据连接，其中客户端的端口号由客户端自己指定，服务器端的端口号随机。



主动模式要求客户端开放端口号给服务器端，需要去配置客户端的防火墙。被动模式只需要服务器端开放端口号即可，无需客户端配置防火墙。但是被动模式会导致服务器端的安全性减弱，因为开放了过多的端口号。

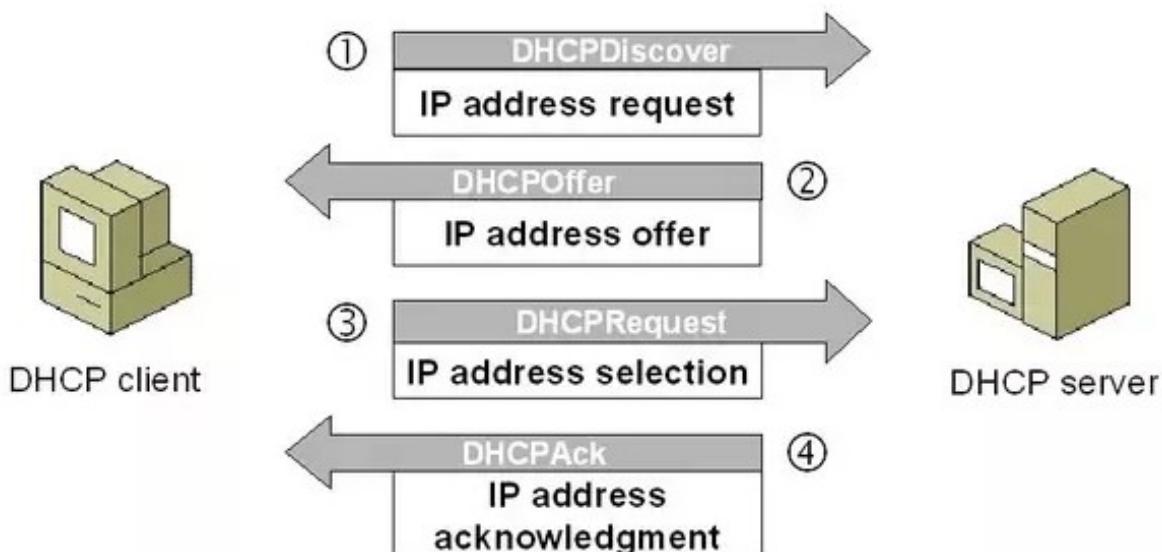
动态主机配置协议 (DHCP, Dynamic Host Configuration Protocol)

提供了即插即用的连网方式，用户不再需要去手动配置 IP 地址等信息。

DHCP 配置的内容不仅是 IP 地址，还包括子网掩码、网关 IP 地址。

DHCP 工作过程如下：

1. 客户端发送 **Discover** 报文，该报文的目的地址为 **255.255.255.255:67**，源地址为 **0.0.0.0:68**，被放入 **UDP** 中，该报文被广播到同一个子网的所有主机上。如果客户端和 DHCP 服务器不在同一个子网，就需要使用中继代理。
2. DHCP 服务器收到 Discover 报文之后，发送 **Offer** 报文给客户端，该报文包含了客户端所需要的信息。因为客户端可能收到多个 DHCP 服务器提供的信息，因此客户端需要进行选择。
3. 如果客户端选择了某个 DHCP 服务器提供的信息，那么就发送 **Request** 报文给该 DHCP 服务器。
4. DHCP 服务器发送 **Ack** 报文，表示客户端此时可以使用提供给它的信息。



远程登录协议TELNET

TELNET 用于登录到远程主机上，并且远程主机上的输出也会返回。

TELNET 可以适应许多计算机和操作系统的差异，例如不同操作系统系统的换行符定义。

电子邮件协议

一个电子邮件系统由三部分组成：用户代理、邮件服务器以及邮件协议。

邮件协议包含发送协议和读取协议，发送协议常用 SMTP，读取协议常用 POP3 和 IMAP。

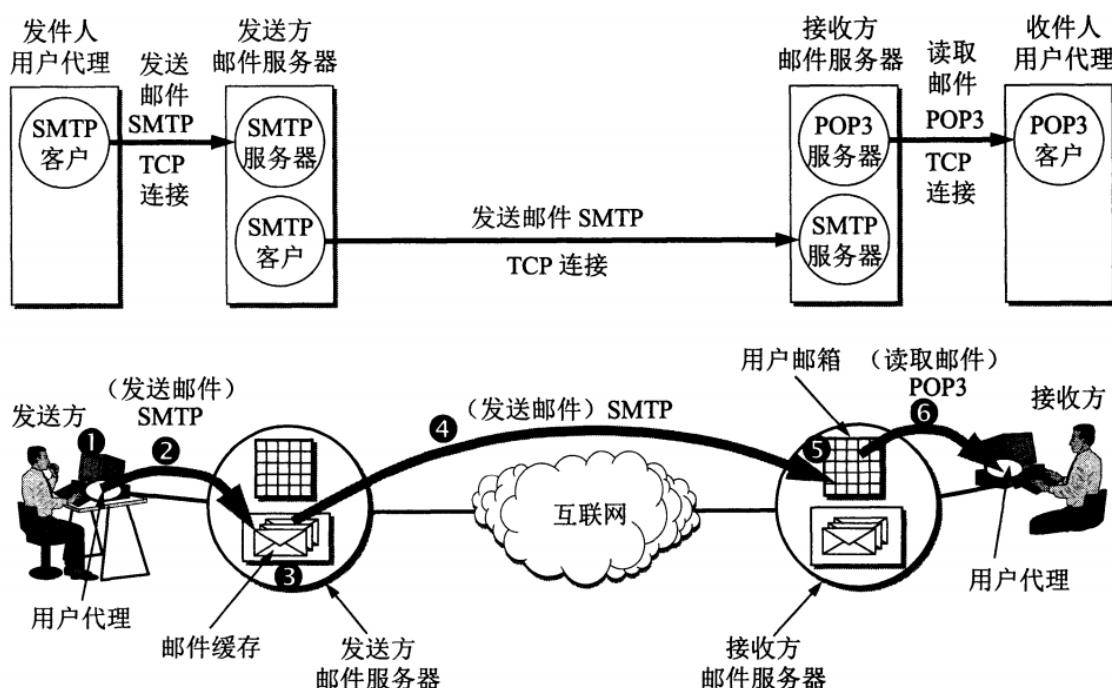


图 6-17 电子邮件的最主要的组成构件

1. SMTP (*Simple Mail Transfer Protocol*, 简单邮件传输协议)

SMTP 只能发送 ASCII 码，而互联网邮件扩充 MIME 可以发送二进制文件。MIME 并没有改动或者取代 SMTP，而是增加邮件主体的结构，定义了非 ASCII 码的编码规则。

2. POP3 (邮局协议Post Office Protocol, POP)

POP3 的特点是只要用户从服务器上读取了邮件，就把该邮件删除。

3. IMAP (因特网信息访问协议, *Internet Message Access Protocol*)

IMAP 协议中客户端和服务器上的邮件保持同步，如果不手动删除邮件，那么服务器上的邮件也不会被删除。IMAP 这种做法可以让用户随时随地去访问服务器上的邮件。

常用端口

应用	应用层协议	端口号	传输层协议	备注
域名解析	DNS	53	UDP/TCP	长度超过 512 字节时使用 TCP
动态主机配置协议	DHCP	67/68	UDP	
简单网络管理协议	SNMP	161/162	UDP	
文件传送协议	FTP	20/21	TCP	控制连接 21, 数据连接 20
远程终端协议	TELNET	23	TCP	
超文本传送协议	HTTP	80	TCP	
简单邮件传送协议	SMTP	25	TCP	
邮件读取协议	POP3	110	TCP	
网际报文存取协议	IMAP	143	TCP	

Web 页面请求过程

1. DHCP 配置主机信息

- 假设主机最开始没有 IP 地址以及其它信息，那么就需要先使用 DHCP 来获取。
- 主机生成一个 DHCP 请求报文，并将这个报文放入具有目的端口 67 和源端口 68 的 UDP 报文段中。
- 该报文段则被放入在一个具有广播 IP 目的地址(255.255.255.255) 和源 IP 地址 (0.0.0.0) 的 IP 数据报中。
- 该数据报则被放置在 MAC 帧中，该帧具有目的地址 FF:FF:FF:FF，将广播到与交换机连接的所有设备。
- 连接在交换机的 DHCP 服务器收到广播帧之后，不断地向上分解得到 IP 数据报、UDP 报文段、DHCP 请求报文，之后生成 DHCP ACK 报文，该报文包含以下信息：IP 地址、DNS 服务器的 IP 地址、默认网关路由器的 IP 地址和子网掩码。该报文被放入 UDP 报文段中，UDP 报文段有被放入 IP 数据报中，最后放入 MAC 帧中。
- 该帧的目的地址是请求主机的 MAC 地址，因为交换机具有自学习能力，之前主机发送了广播帧之后就记录了 MAC 地址到其转发接口的交换表项，因此现在交换机就可以直接知道应该向哪个接口发送该帧。
- 主机收到该帧后，不断分解得到 DHCP 报文。之后就配置它的 IP 地址、子网掩码和 DNS 服务器的 IP 地址，并在其 IP 转发表中安装默认网关。

2. ARP 解析 MAC 地址

- 主机通过浏览器生成一个 **TCP 套接字**, **套接字向 HTTP 服务器发送 HTTP 请求**。为了生成该套接字, 主机需要知道网站的域名对应的 IP 地址。
- 主机生成一个 **DNS 查询报文**, 该报文具有 53 号端口, 因为 DNS 服务器的端口号是 53。
- 该 DNS 查询报文被放入目的地址为 DNS 服务器 IP 地址的 IP 数据报中。
- 该 IP 数据报被放入一个以太网帧中, 该帧将发送到网关路由器。
- DHCP 过程只知道网关路由器的 IP 地址, 为了获取网关路由器的 MAC 地址, **需要使用 ARP 协议**。
- 主机生成一个包含目的地址为网关路由器 IP 地址的 ARP 查询报文, 将该 ARP 查询报文放入一个具有广播目的地址 (FF:FF:FF:FF) 的以太网帧中, 并向交换机发送该以太网帧, 交换机将该帧转发给所有的连接设备, 包括网关路由器。
- 网关路由器接收到该帧后, 不断向上分解得到 ARP 报文, **发现其中的 IP 地址与其接口的 IP 地址匹配**, 因此就发送一个 ARP 回答报文, 包含了它的 MAC 地址, 发回给主机。

3. DNS解析域名

- 知道了网关路由器的 MAC 地址之后, 就可以继续 DNS 的解析过程了。
- 网关路由器接收到包含 DNS 查询报文的以太网帧后, 抽取出 IP 数据报, 并根据转发表决定该 IP 数据报应该转发的路由器。
- 因为路由器具有内部网关协议 (RIP、OSPF) 和外部网关协议 (BGP) 这两种路由选择协议, 因此路由表中已经配置了网关路由器到达 DNS 服务器的路由表项。
- 到达 DNS 服务器之后, DNS 服务器抽取出 DNS 查询报文, 并在 DNS 数据库中查找待解析的域名。
- 找到 DNS 记录之后, 发送 DNS 回答报文, 将该回答报文放入 **UDP 报文段中**, 然后放入 IP 数据报中, 通过路由器反向转发回网关路由器, 并经过以太网交换机到达主机。

4. HTTP请求页面

- 有了 HTTP 服务器的 IP 地址之后, 主机就能够生成 **TCP 套接字**, 该套接字将用于向 Web 服务器发送 HTTP GET 报文。
- 在生成 TCP 套接字之前, 必须先与 HTTP 服务器进行三次握手来建立连接。生成一个具有目的端口 80 的 TCP SYN 报文段, 并向 HTTP 服务器发送该报文段。
- HTTP 服务器收到该报文段之后, 生成 TCP SYN ACK 报文段, 发回给主机。
- 连接建立之后, **浏览器生成 HTTP GET 报文, 并交付给 HTTP 服务器**。
- HTTP 服务器从 TCP 套接字读取 HTTP GET 报文, 生成一个 HTTP 响应报文, 将 Web 页面内容放入报文主体中, 发回给主机。
- 浏览器收到 HTTP 响应报文后, 抽取出 Web 页面内容, 之后进行渲染, 显示 Web 页面。

Web页面请求过程——另一种说法

1、查询DNS，获取域名对应的IP。

- (1) 检查本地hosts文件是否有这个网址的映射, 如果有, 就调用这个IP地址映射, 解析完成。
- (2) 如果没有, 则查找本地DNS解析器缓存是否有这个网址的映射, 如果有, 返回映射, 解析完成。
- (3) 如果没有, 则查找填写或分配的首选DNS服务器, 称为本地DNS服务器。服务器接收到查询时:
 - 如果要查询的域名包含在本地配置区域资源中, 返回解析结果, 查询结束, 此解析具有权威性。
 - 如果要查询的域名不由本地DNS服务器区域解析, 但服务器缓存了此网址的映射关系, 返回解析结果, 查询结束, 此解析不具有权威性。

(4) 如果本地DNS服务器也失效：

- 如果未采用转发模式（迭代），本地DNS就把请求发至13台根DNS，根DNS服务器收到请求后，会判断这个域名（如.com）是谁来授权管理，并返回一个负责该顶级域名服务器的IP，本地DNS服务器收到顶级域名服务器IP信息后，继续向该顶级域名服务器IP发送请求，该服务器如果无法解析，则会找到负责这个域名的下一级DNS服务器（如<http://baidu.com>）的IP给本地DNS服务器，循环往复直至查询到映射，将解析结果返回本地DNS服务器，再由本地DNS服务器返回解析结果，查询完成。
- 如果采用转发模式（递归），则此DNS服务器就会把请求转发至上一级DNS服务器，如果上一级DNS服务器不能解析，则继续向上请求。最终将解析结果依次返回本地DNS服务器，本地DNS服务器再返回给客户机，查询完成。

2、客户机发送HTTP请求报文：

- (1) 应用层：客户端发送HTTP请求报文
- (2) 传输层：切分长数据，并确保可靠性。
- (3) 网络层：进行路由
- (4) 数据链路层：传输数据
- (5) 物理层：物理传输bit

3、服务器端经过物理层→数据链路层→网络层→传输层→应用层，解析请求报文，发送HTTP响应报文。

4、客户端解析HTTP响应报文

5、浏览器开始显示HTML

6、浏览器重新发送请求获取图片、CSS、JS的数据。

7、如果有AJAX，浏览器发送AJAX请求，及时更新页面。