本章教学提要

- **教学目标：**

✓理解网络层与网络互联的基本概念（功能、设计原则）

✓掌握**TCP/IP**网络层及其主要协议

✓掌握**IP**协议的基本内容

✓掌握**IP**地址的基本概念及相关技术（子网划分、**CIDR**和**NAT**）

✓掌握地址解析（**ARP**）的基本概念与实现方法

- ✓掌握IP分组的交付与路由选择的概念；理解路由实现的机理(主动路由与被动路由、路由表与路由选择、静态路由与动态路由、路由器)
- ✓掌握网络层中源到目标分组传输的实现机理
- ✓理解Internet控制报文协议ICMP
- ✓了解IPv6的产生背景及其主要特点
- **教学重/难点**：IP协议与IP地址规划，源到目标IP分组的传送机理。
- **教学时数**：理论8-10学时

本章教学特点与结构

- 特点：量大、内容复杂、要求高→高度重视
- 结构：

1. 关于网络层功能及其必要性的理解；
2. 围绕TCP/IP网络层的讨论包括：

- ✓ IP协议
- ✓ IP地址及其规划
- ✓ ARP协议
- ✓ 路由与路由协议
- ✓ ICMP协议

Section 1　网络层与网络互连的基本概念

本节关注/研讨问题之一

- 为什么在数据链路层之上需要网络层？
- 就源到目标的主机通信而言，网络层需要解决哪些问题？或提供哪些基本功能？

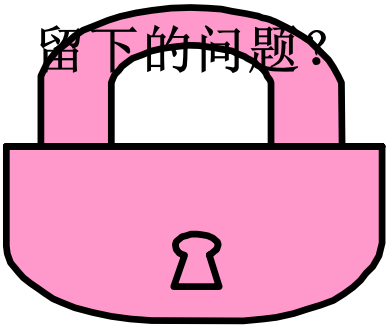
对物理层和数据链路层的回顾

物理层

- 原始比特流的传输
  - ✓ 传输介质、信号、噪音、复用；

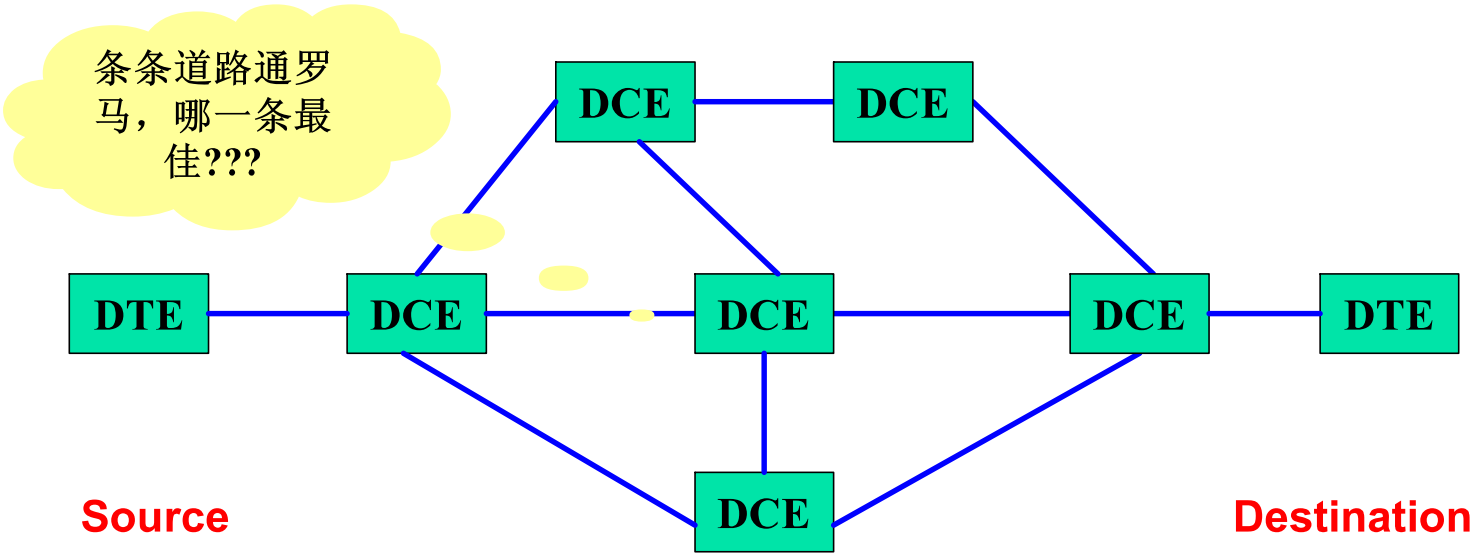
数据链路层

- 相邻节点间的(可靠)数据传输
  - ✓ 物理寻址
  - ✓ 差错控制、流量控制、介质访问控制



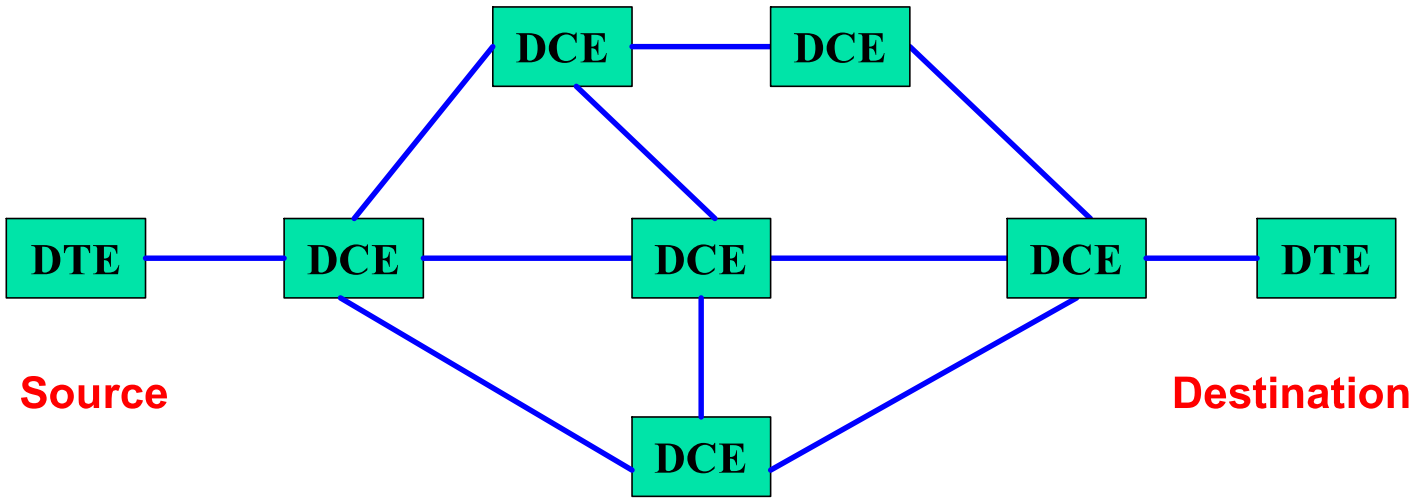
- 源到目标主机的数据传输
  - ✓ 网络互连(LAN-WAN-MAN)
    跨越网络的主机寻址
  - ✓ 最佳路径选择

第2层的问题之一



- 当网络互连规模增大时，从源端到目的端会存在许多的中间节点，这些中间节点构成了从源端到目的端的多条路径，数据包传输面临路径选择的问题。

第2层的问题之二



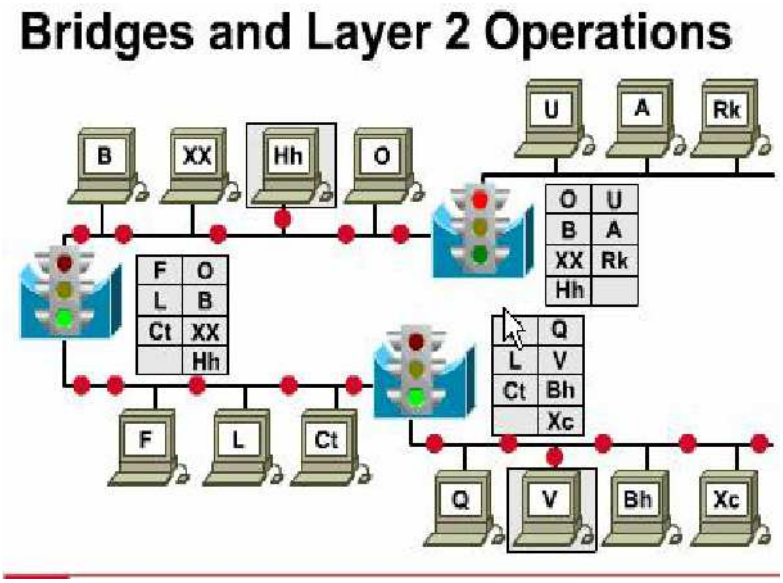
- 网络互连规模增大时，必然会涉及异构网络的互连。

✓类型的不同：**LAN**、**MAN**或**WAN**

✓实现技术的差异：同属**LAN**，**Ethernet**和**Token Ring**存在差异

✓通信协议、计算机系统体系结构、操作系统的不同

第2层的问题之三



- 第2层使用平面结构的寻址模式(如MAC地址)来唯一标识网络中的主机，地址编码中不含任何有关主机所在网络的结构信息。

第2层的问题之三

- 当网络规模变大时，利用数据链路层的物理寻址方法直接定位网络中的主机会因为网桥或交换机所产生的大量广播转发（洪泛，flooding）而可能导致网络瘫痪。

- **结论：**基于平面化物理地址的直接寻址方式只能适用于规模非常小的网络环境。当网络互连规模增大时，需要提供一种包含主机所在位置信息的结构化地址来实现跨越不同LAN、MAN和WAN的主机逻辑寻址。

网络层的功能

- 将源主机发出的分组经由各种网络路径送达目的主机。
- 具体地包含：

✓ 了解通信子网的拓扑结构，并通过一定的路由算法为分组实现进行最佳路径的选择→路由（**Routing)**；

✓ 在选择路径时注意既不要使某些路径或通信线路处于超负载状态，也不能让另一些路径或通信线路处于空闲状态→拥塞控制/负载平衡；

网络层的功能

- 当从源主机到目标主机所经历的网络不属于同一种类型时，协调好不同网络间的差异→异构网络互连。
- 网络层的功能及其实现机制由网络层协议来描述，并且集中体现在网络层协议数据单元－分组（packet）中。

  分组中包括实现网络层功能所必需的控制信息，如收发双方的网络地址等。

拥塞及其产生原因

- 当通信子网中的某一部分有太多的数据分组时，所导致的网络性能下降现象被称为网络拥塞。
- 拥塞会引起网络分组的丢失，在严重的情况下，会导致网络运行的瘫痪。
- 产生拥塞的原因是多样的：

✓ 线路的带宽太小

✓ 网络上的流量不平衡

✓ 通信子网中的设备如路由器的**CPU**性能不够

拥塞控制与流量控制的区别

- 拥塞控制用于确保通信子网能运送所有等待传送的数据，是一个全局性(global)的问题。

  涉及所有主机、路由器，并与路由器的存储转发能力和其他影响通信子网负荷的因素有关。

- 流量控制只涉及发送者和接收者之间的点到点通信流量(local)。其任务是确保一个快速的发送者不要以高于接收者所能承受的速率发送数据。

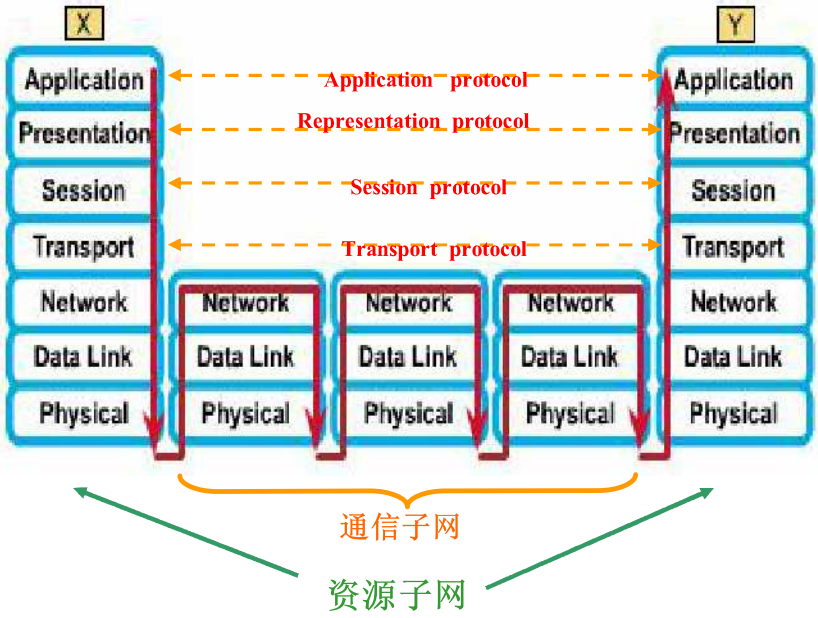
网络层的地位

- 利用数据链路层所提供的相邻节点之间的数据传输服务，向传输层提供源到目标的数据传输服务。
- 处理端到端（源→目标）数据传输的最低层

数据链路层仅涉及相邻节点间的数据传输。

- 通信子网的最高层

通信子网中的主机因为仅涉及通信问题而只拥有OSI的下三层。



- 通信子网与资源子网的接口

资源子网中的主机具备了OSI模型中所有七层的功能，但通信子网中的主机最多涉及OSI模型的下三层。

本节关注/研讨问题之二

- 网络层所提供的两类服务各有什么特点？分别采用了什么方式？
- TCP/IP的网际层采用了哪种方式？提供了什么服务？

网络层的基本服务类型

- **两种服务类型：**

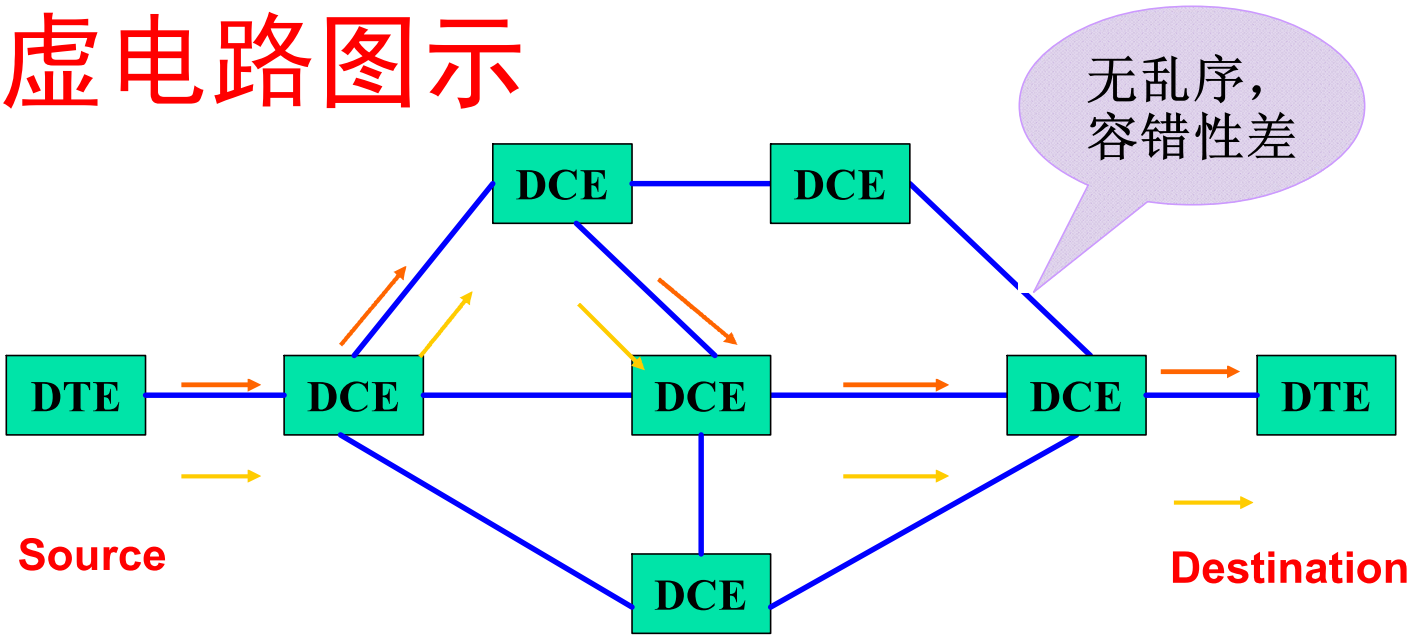
✓ 可靠的面向连接服务（虚电路）

✓ 不可靠的无连接服务（数据报）

面向连接服务与虚电路

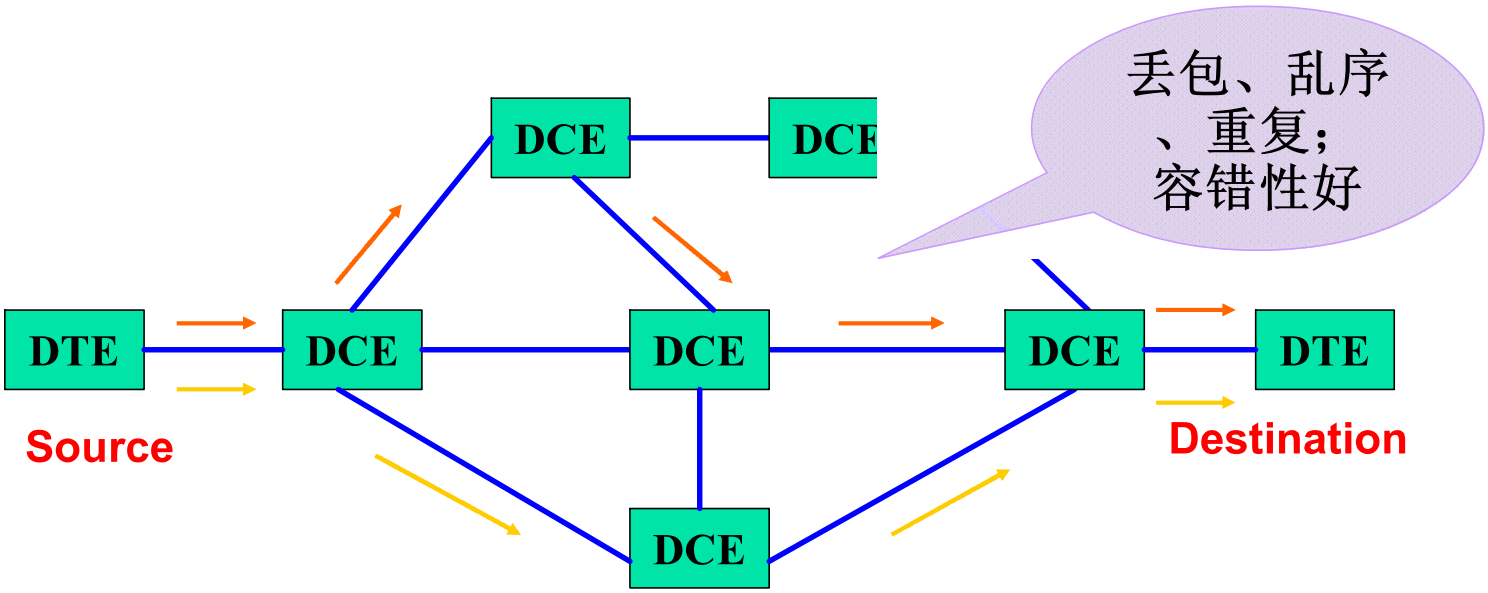
- 面向连接是指在数据传输之前双方需要为此建立一种连接，然后在该连接上实现有次序的分组传输，直到数据传送完毕连接才被释放。
- 通信子网以虚电路（Virtual Circuit）实现面向连接的服务。
- 涉及虚电路逻辑连接的三个阶段：虚电路建立、数据传输和虚电路拆除。

虚电路图示



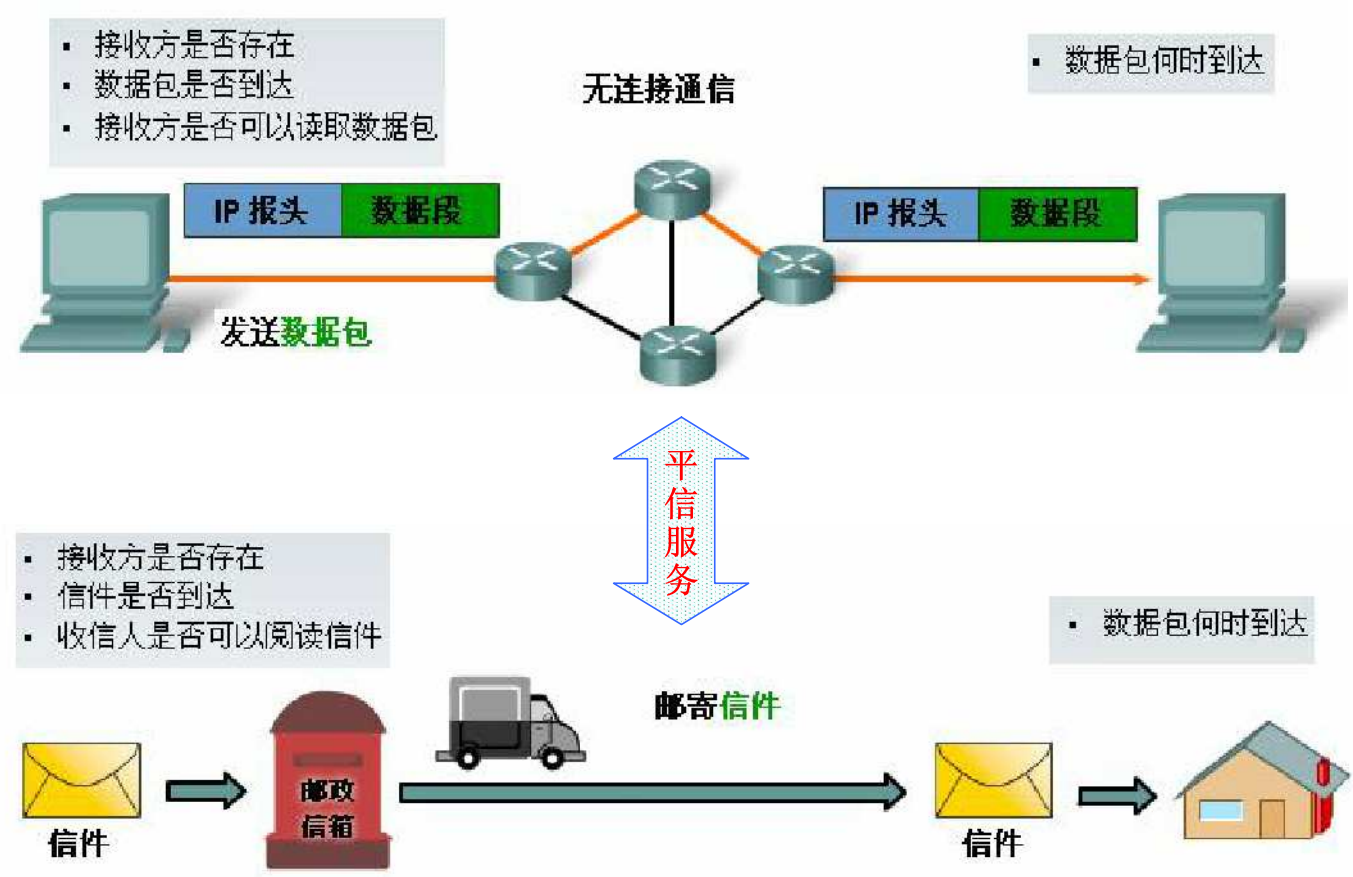
- 虚电路建立：将从源端机器到目标机器的路由作为连接建立的一部分加以保存。
- 数据传输：在虚电路上传送的分组不需要再携带目的地址，并取相同的路径（路由）通过通信子网。
- 传输完成后需要拆除连接。

无连接服务与数据报(Datagram)



- 不需要为数据传输事先建立连接，只提供简单的源和目标之间的数据发送与接收功能。
- 为每个分组选择独立的路由，不同的分组可以走不同的路由。

进一步理解无连接的数据报服务



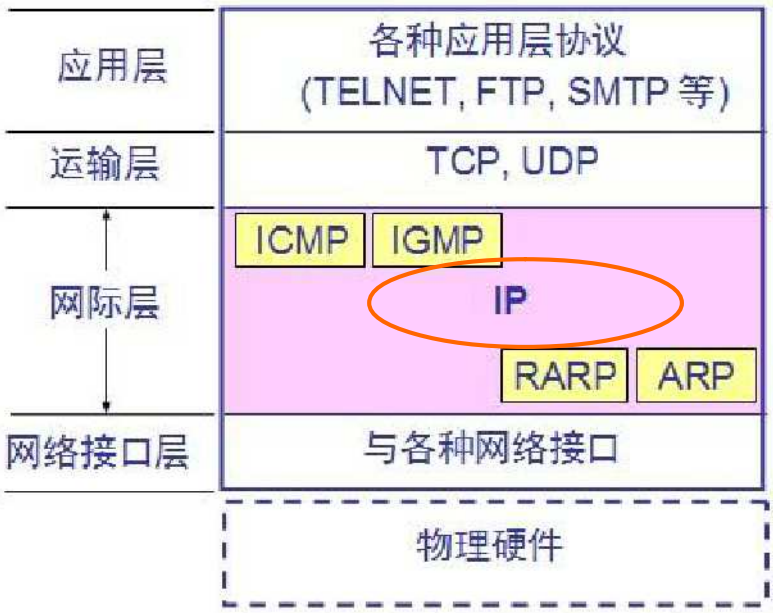
虚电路与数据报的比较

分组交换方式 / 比较项目	数据报	虚电路
连接设置	不需要	需要
地址	每个分组需要完整的源和目的地址	每个分组只需要包含一个虚电路号
状态信息	有路由表，无连接表	连接表
路由选择	每个包独立选择	虚电路建立后勿需路由
路由器失败的影响	丢失失败时的分组	所有经过失败路由器的 VC 失效
传输质量	同一报文的不同分组会出现乱序、重复或丢失	同一报文的不同分组不会出现乱序、重复或丢失
协议复杂度	相对低	相对高
通信效率	相对高	相对低

网络层的设计目标

- 所提供的服务与通信子网所采用的技术无关；
- 通信子网的数量、类型和拓朴结构对于传输层是透明的，即不影响网络层向传输层提供的服务。
- 网络层所定义的地址应采用统一的方式，与底层的网络无关，从而能跨越不同的LAN和WAN实现逻辑寻址。

TCP/IP的网络层



- 位于TCP/IP模型的第二层；
- 提供无连接的数据报服务
- 包括了IP、ARP、RARP协议、ICMP和若干路由协议。

Protocols for the Internet Layer

- **IP协议（因特网协议）：**
  Internet Protocol
- **ICMP协议 （因特网消息控制协议）**
  Internet Control Message Protocol
- **ARP协议 （地址解析协议）**
  Address Resolution Protocol
- **RARP协议（反向地址解析协议）**
  Reverse Address Resolution Protocol
- **Routing protocols （路由协议）**

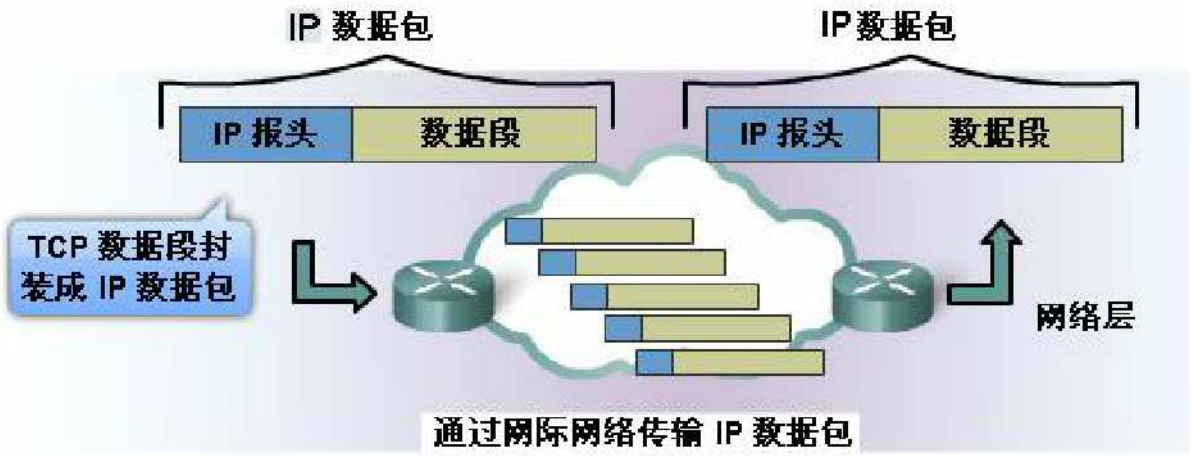
Section 2 IP协议

本节关注/研讨问题之一

- 为什么IP协议被认为是**TCP/IP**网络层的核心协议？
- **IP**协议的基本功能与特点是什么？如何在**IP**分组中体现？

IP协议的地位与作用

- 地位：TCP/IP网络层的核心协议，也是整个TCP/IP模型中的核心协议，所有其他协议如TCP、UDP、ICMP等都以它为基础。
- 协议数据单元被称为IP分组。
- 作用：通过网际(互连)网络传输IP分组

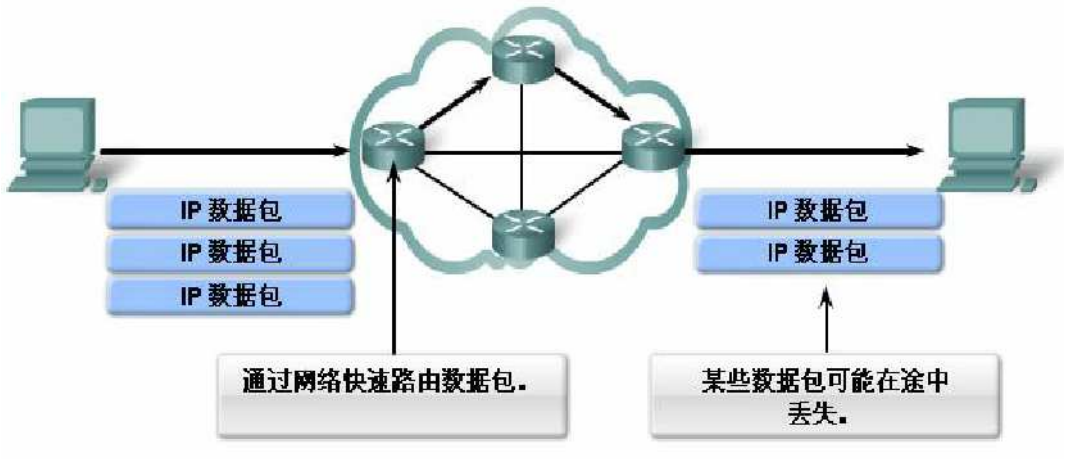


IP协议的特点

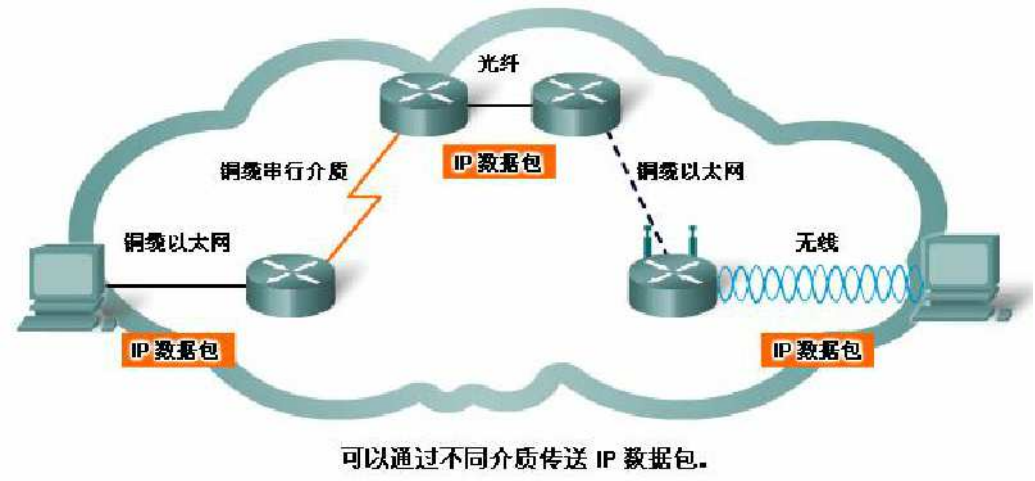
- **提供无连接的、不可靠的数据报传输服务**

✓ “无连接”：传输分组前不需要建立连接，也不维护IP分组发送后的任何状态信息，每个分组的处理相对独立，不同分组可以走不同的路径。

✓ “不可靠”：不提供差错控制和确认机制，不保证每个IP分组能被送达目的节点或被正确接收，也不保证IP分组传输顺序的正确性。→“尽力而为(best-effort)”

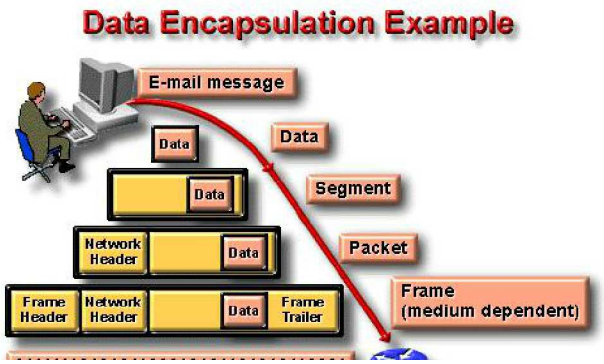


支持异构网络的互连—介质、通信网络的无关性



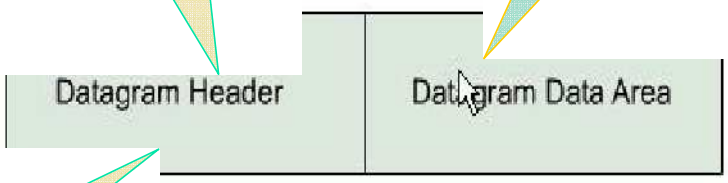
- ✓ 以统一的IP分组传输提供对异构网络互连的支持；
- ✓ 向传输层屏蔽了底层通信子网中的不同网络技术在物理层和数据链路层的差异；
- ✓ IP编址模式实现了跨越不同LAN、MAN和WAN的主机寻址。

IP分组的结构与IP数据报封装



由 IP协议定义

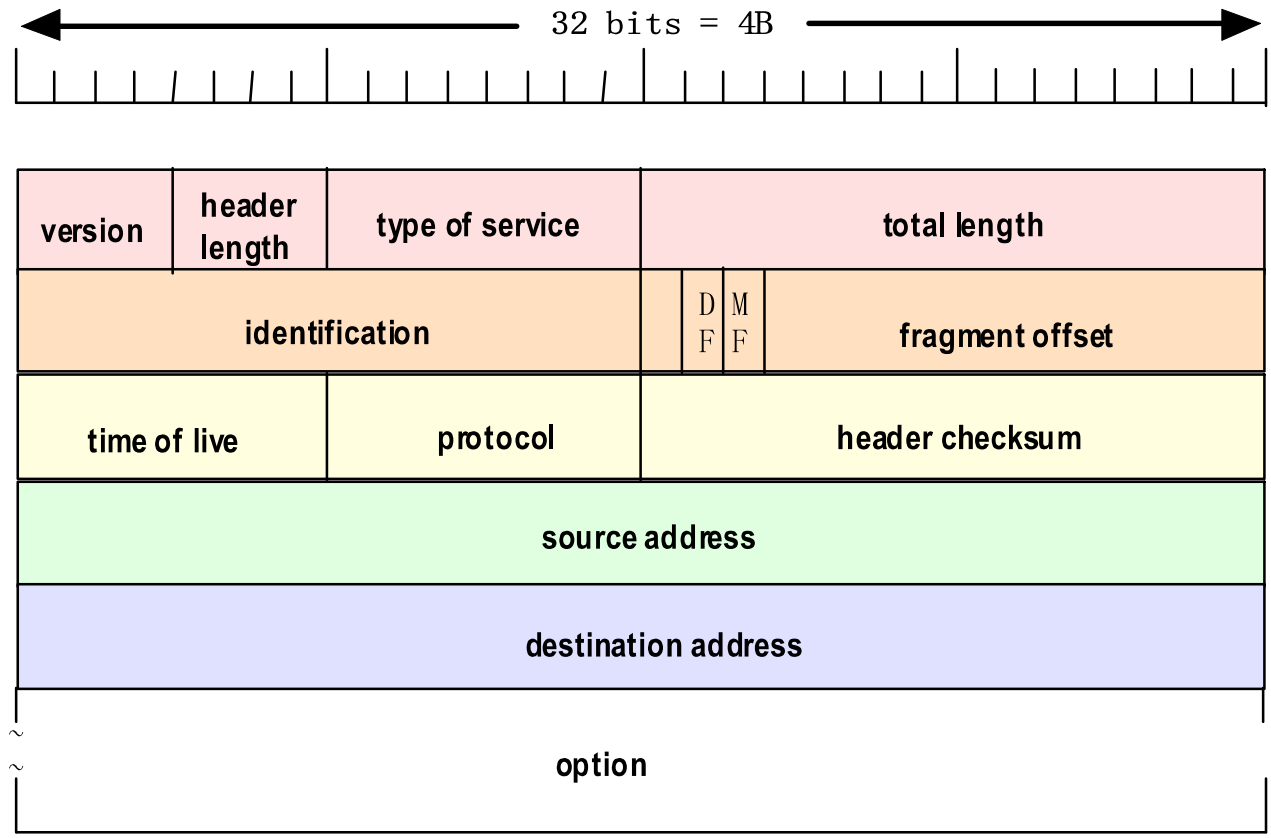
来自传输层的分段



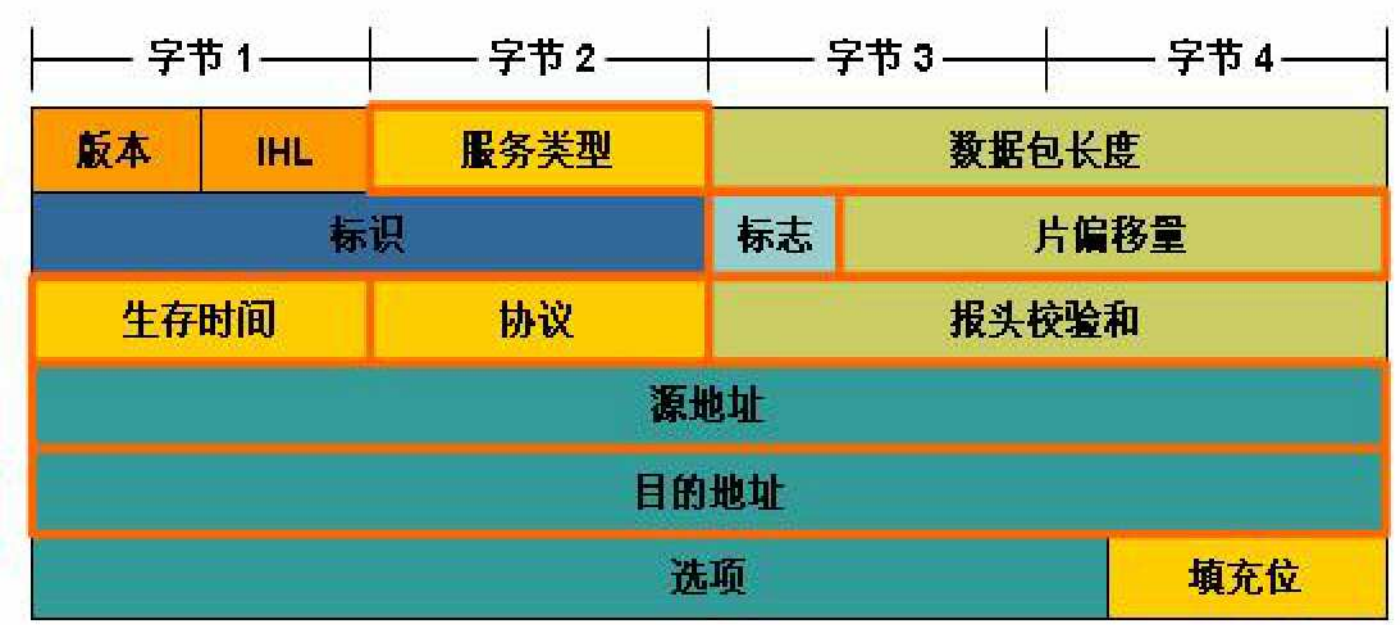
IP数据报

无连接的IP数据报传输服务及其机制就体现在IP报头字段的定义中。

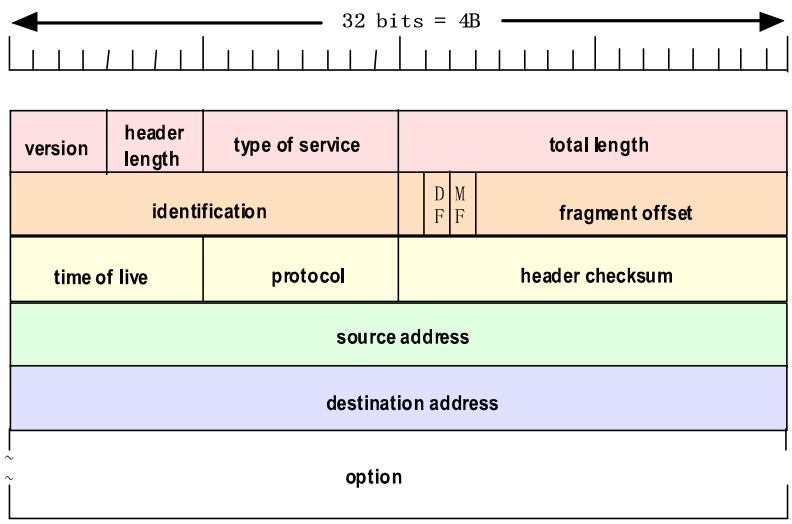
IP数据报的报头格式(英文)



IP数据报的报头格式(中文)

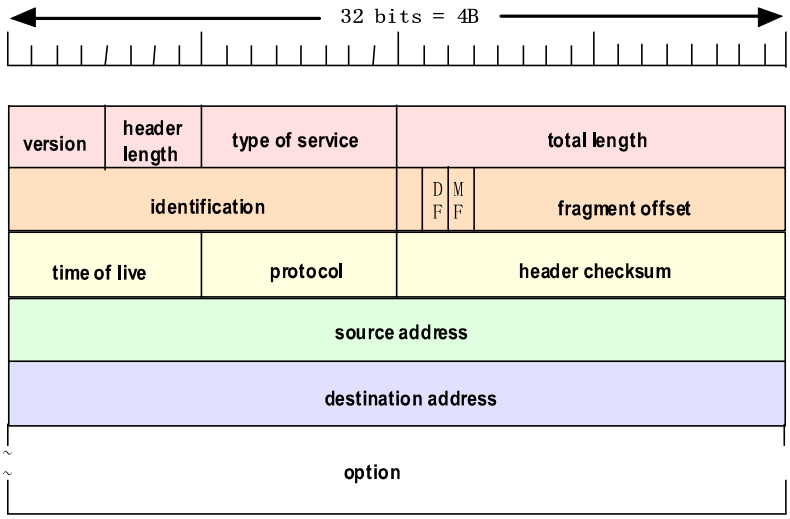


IP数据报中的域（fields）



- 版本(version)：长度为4位，表示数据报协议的版本，如4.0和6.0；
- 报头长度(header length)：长度为4位，表示数据报报头的长度，以32位(相当于4byte)长度为单位。当报头中无可选项时，其基本长度为5(相当于20byte)；报头长度的最大值为15 (相当于60byte).

- 服务类型(type of service)：长度为8位，主机要求通信子网提供的服务类型。包括：
  - ✓ 3位长度的优先级，共分为8级，数值越大优先级越高；
  - ✓ 4位长度的服务类型，标志分别为:D-延迟/delay、T-吞吐量/ throughput、R-可靠性/ reliability、C-开销/ cost）和1个保留位。



- 总长(total length)：长度为16位，表示数据报的总长度，包括头部和数据，以字节为单位。

  数据报的最大长度为2^{16}-1字节，即65535字节。

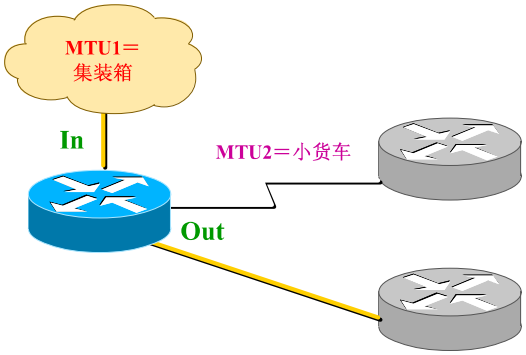
- 标识(identification):用以标识属于同一数据报的分片。

  当数据报长度超出网络最大传输单元（Maximum transfer unit, MTU）时，必须进行数据报的分片，并且要为各分片(fragment)提供标识。

  所有属于同一数据报的分片被赋予相同的标识。
- 标志(flags) :长度为3位，低2 位表示该数据报是否可分片。DF值为“0”时表示可以分片，为“1”时表示不可分片；MF值为“0”表示所接收的是最后一个分片，为“1”表示还有进一步的分片；
- 片偏移：若有分片时，用以指出该分片在数据报中的位置。片偏移值以8B为单位来计数，因此各分片的长度必须为8B的整数倍。13位的偏移长度意味着一个长数据报至多可被分为2^{13}个分片。

MTU与IP数据报的分片

- 每一种物理网络都规定了各自帧的数据域最大字节长度，即最大传输单元（MTU），如：
  - ✓ Ethernet:1500B（RFC894）
  - ✓ Token Ring:17914B（RFC1042）
  - ✓ FDDI:4352B（RFC1188)
  - ✓ PPP:296B（RFC1144)；



- IP数据报作为网络层数据在底层网络要以帧的形式来传输→每一个路由节点都要将接收到的帧进行拆包和处理，然后封装成转发端口所对应的帧
- 一个数据报可能要通过多个不同的(异构的)物理网络;
- 一个来自具有较大MTU局域网的数据报，在通过另一个具有较小MTU的局域网或广域网时，必须对该IP数据报进行分片。

IP 数据报分片的举例

P256: 一个数据报的数据部分为3800字节长（使用固定首部），需要分片为长度不超过1420字节的数据报片。

因固定首部长度为20字节，因此每个数据报片的数据部分长度不能超过1400字节，3800÷1400≈2.7。

于是分为3个数据报片，其数据部分的长度分别为1400，1400和1000字节。原始数据报首部被复制为各数据报片的首部，但必须修改有关字段的值。图8-16表示分片的结果。表8-7是各数据报的首部中与分片有关的字段中的数值，其中标识字段的值是任意给定的。具有相同标识的数据报片在目的站就可无误地重装成原来的数据报。

IP 数据报分片的举例

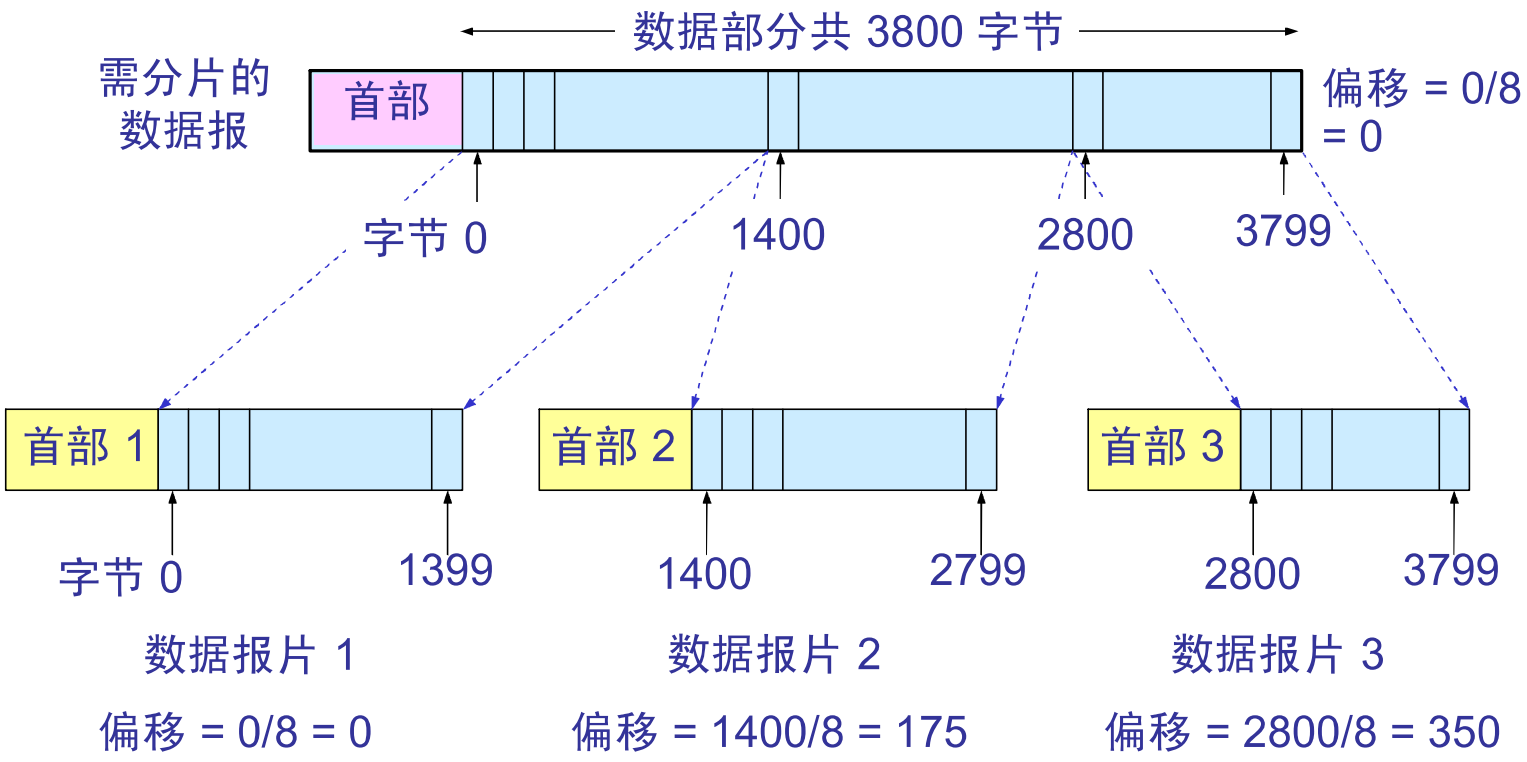

数据部分共 3800 字节
需分片的
数据报
首部
偏移 = 0/8
= 0
字节 0
1400
2800
3799
首部 1
首部 2
首部 3
字节 0
1399
1400
2799
2800
3799
数据报片 1
偏移 = 0/8 = 0
数据报片 2
偏移 = 1400/8 = 175
数据报片 3
偏移 = 2800/8 = 350

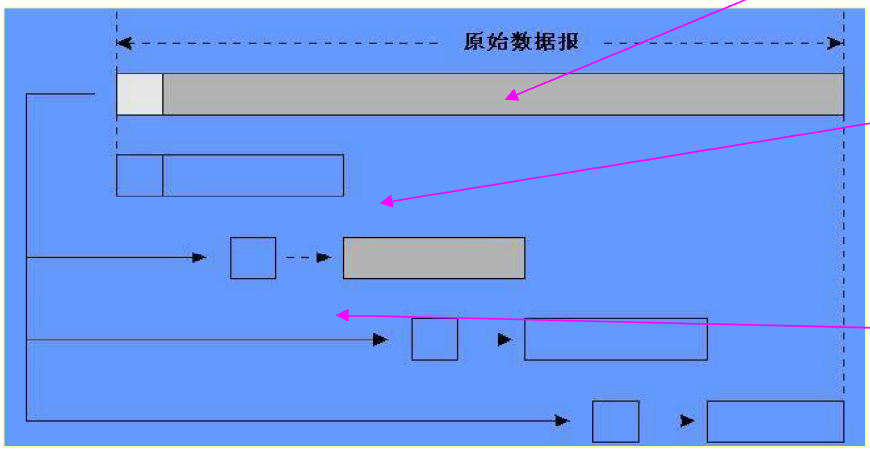
表8-7　IP数据报首部中与分片有关的字段中的数值。

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

IP数据报分片的示例

- 问题：某路由器的接口1，收到一个总长度为2220字节的IP数据报，查询路由表后，决定将该IP数据报由接口2送出，但接口2所连的网段的MTU数据字段只有820字节，请问该路由器将如何对该数据报进行分片？

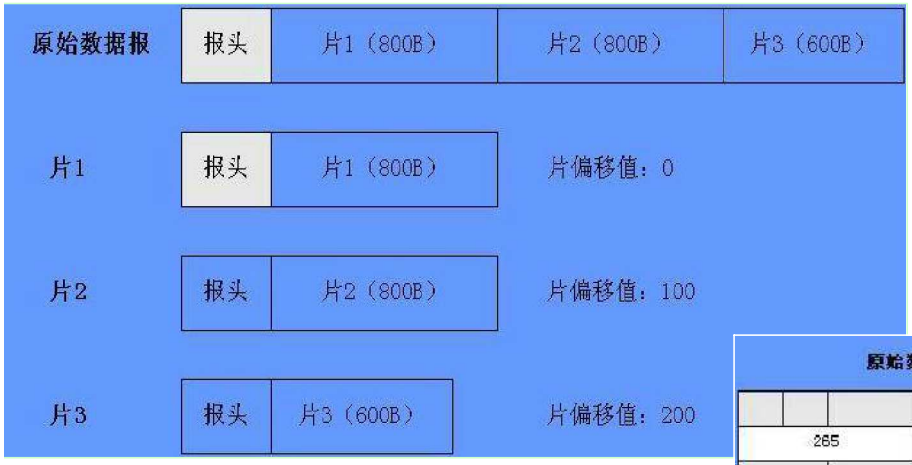
分析：

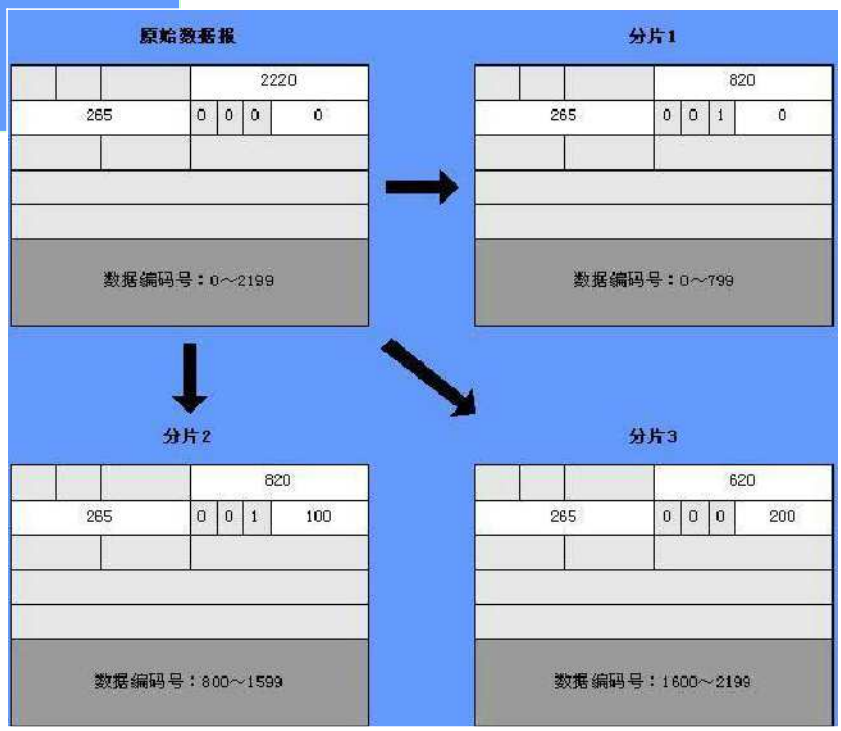


总长度为2220B的IP数据报＝20BIP头＋2200B数据

总长度为2220B的IP数据报超出了网段2的MTU（820B）的承载能力

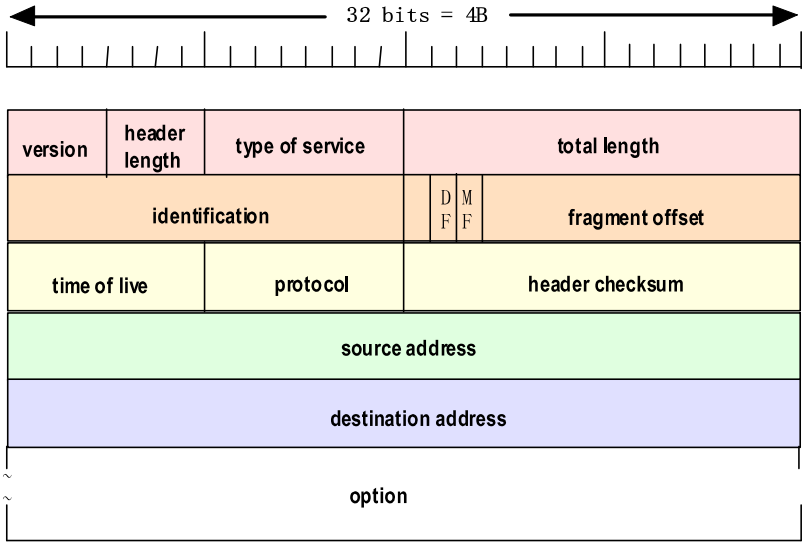
分片时必须考虑为每个分片重新封装IP头所付出的长度820＝20B的IP头＋分片数据


原始数据报
报头
片1（800B）
片2（800B）
片3（600B）
片1
报头
片1（800B）
片偏移值：0
片2
报头
片2（800B）
片偏移值：100
片3
报头
片3（600B）
片偏移值：200

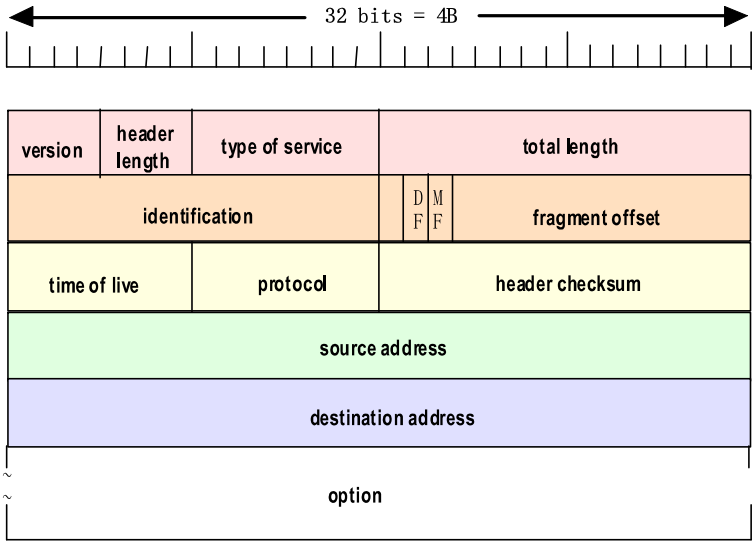

原始数据报
2220
265
0 0 0 0
数据编码号：0～2199
分片1
820
265
0 0 1 0
数据编码号：0～799
分片2
820
265
0 0 1 100
数据编码号：800～1599
分片3
620
265
0 0 0 200
数据编码号：1600～2199

- 生存时间(TTL)：长度为8位，限定数据报生存期的计时器

  推荐以秒来计数，最长为$2^8-1=255s$，实际中常以经过的路由器数目来计，生存时间每经过一个路由节点都要递减，当生存时间减到零时，分组就要被丢弃。目的是防止数据报在网络中无限制地漫游。



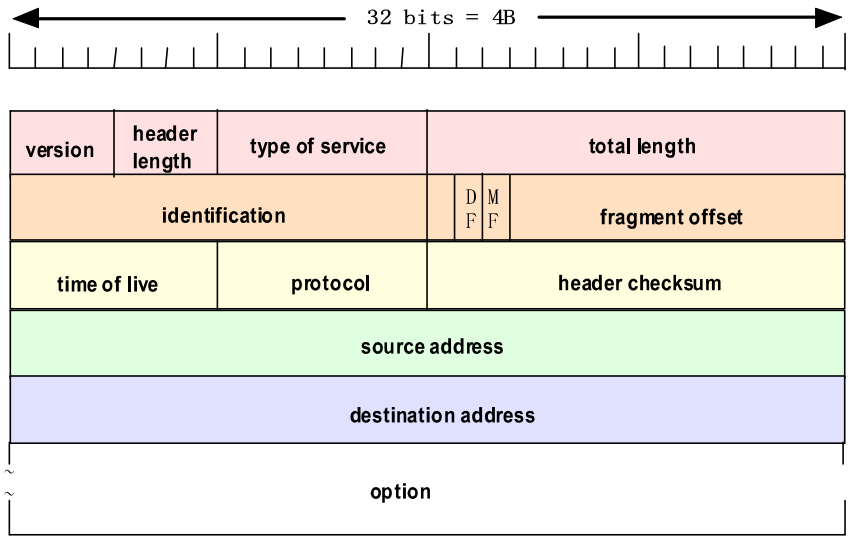
- 协议（**protocol**）：长度为8位，指示上层所采用的协议，如**TCP**、**UDP**或**ICMP**等。



IP 分组中的协议域值与数据报所采用的高层协议类型间的对应关系举例

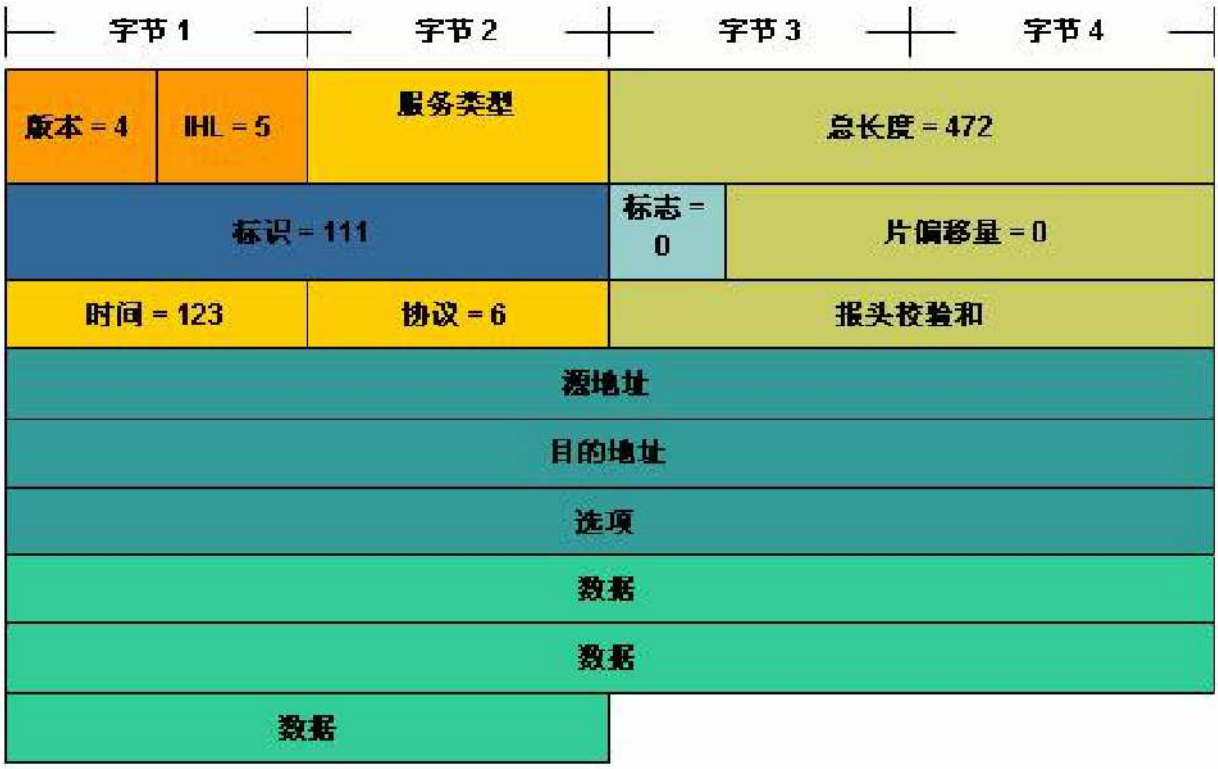
高层协议类型	协议域值	高层协议类型	协议域值
TCP	6	UDP	17
ICMP	1	BGP	8
OSPF	89	RIP	9

- 头校验和（**header checksum**）：长度为16位，用于校验头标。采用累加求补再取其结果补码的校验方法。若正确到达时，校验和应为零。
- 任选字段（**options**）：长度范围为0-40Bytes(可变长),支持各种选项，提供扩展余地。

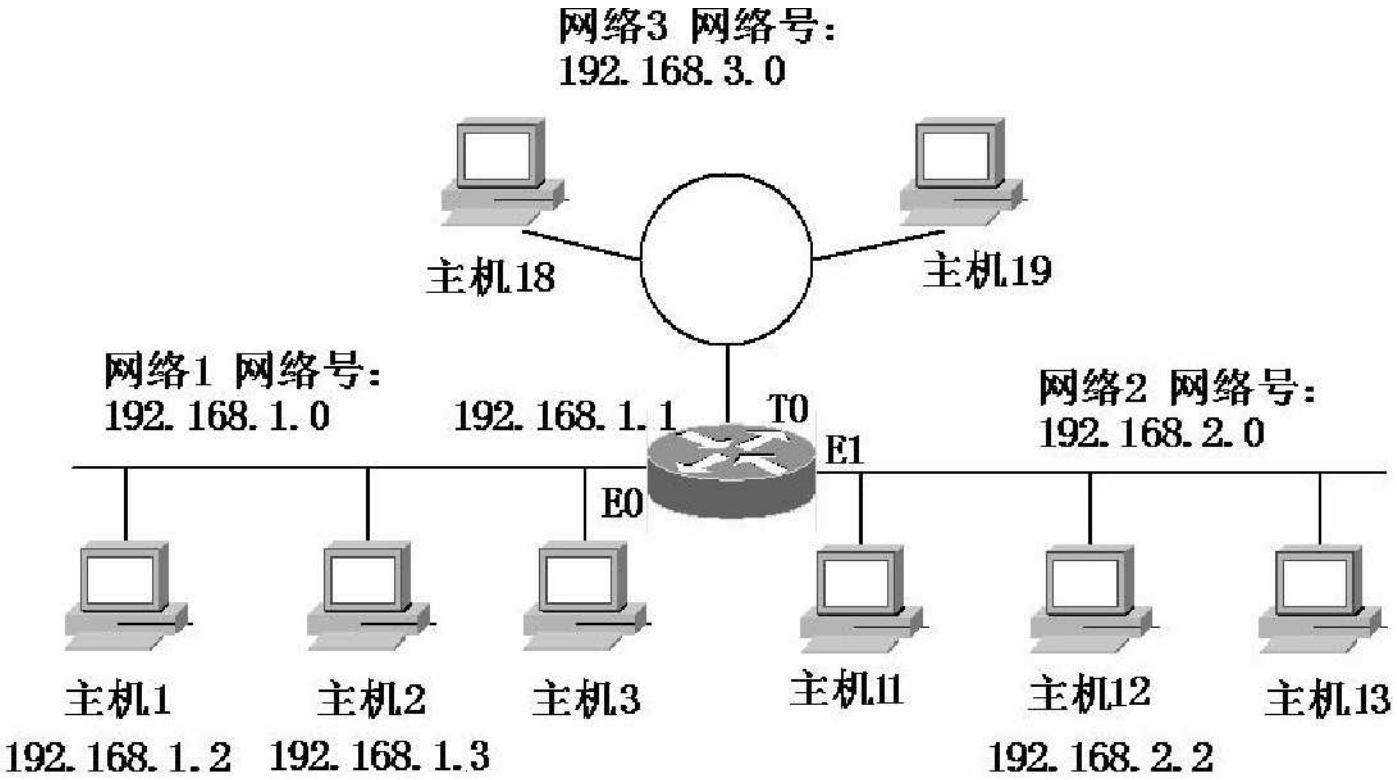


- 源地址/目标地址（**source /destination address**）：长度各为32位，分别表示源主机和目标主机的网络地址。

IP数据报示例


字节 1
字节 2
字节 3
字节 4
版本 = 4
IHL = 5
服务类型
总长度 = 472
标识 = 111
标志 = 0
片偏移量 = 0
时间 = 123
协议 = 6
报头校验和
源地址
目的地址
选项
数据
数据
数据

问题的引入



问题：若主机1 要将数据送到主机2 ，两台主机都有各自的物理地址和IP地址，发送数据时是如何实现寻址的？

ARP的产生背景

- **IP地址只是一种在网际范围内标识主机的逻辑地址，不能直接利用它们在物理上发送分组→**

  逻辑地址不可被直接用于主机寻址，无法在物理上实现IP分组的传输。

- **为了在物理上实现IP分组的传输，所有网络层的分组在数据链路层必须都以帧的方式传送，以借助数据链路层的物理寻址功能**

  例如，以太网中的主机通过网卡连接到以太网中，网络层的分组要封装到以太网帧的数据字段中，以帧的形式进行发送。

- 数据链路层硬件只能对帧进行处理→

  不能识别分组中的三层逻辑地址，它们只能识别帧中的二层物理地址。

  例如，网卡只能识别48位的MAC地址，不能识别IP地址。

- 需要在网络互连层提供从IP地址到物理地址或MAC地址的映射功能→

  地址解释协议(address resolution protocol，简称ARP)提供了该功能，该协议在RFC865中定义。

ARP相关的术语

- **ARP 表：**

✓ 主机上用于存储 IP 地址及其经过解析的物理地址的数据表；

✓ 存储在 RAM中，掉电后会丢失；

✓ 自动维护(两种方式：流量监控、ARP广播请求)。

- **ARP request** (ARP请求）

  用于在网络中请求ARP解析(IP地址→MAC地址)的广播包。

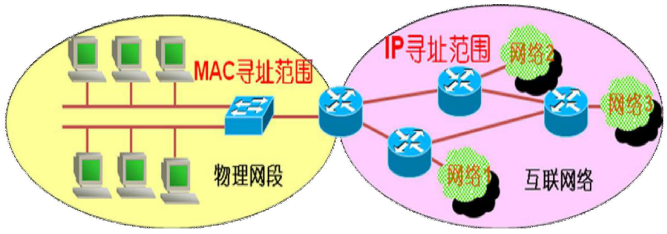
- **ARP reply** (ARP回应）

  用于在网络中回应关于ARP请求的包。

- **ARP update** (ARP更新）

  当主机收到一个当前ARP 缓存中不存在的ARP数据项或回应时，要在ARP中增加一个新的表项；若当前ARP 缓存已经存在相应的ARP表项时，则更新时间标记。

- ARP必须在某一局域网中进行，一台主机在向同局域网上的另一台计算机发送数据时，应先做地址解析，然后按MAC地址直接发送数据帧。



为了减少网络内通信流量，每台主机在内存中都维护着一个ARP表，其初始值为空，这个ARP表就是ARP高速缓存 (ARP cache)，里面有所在的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。

ARP 高速缓存

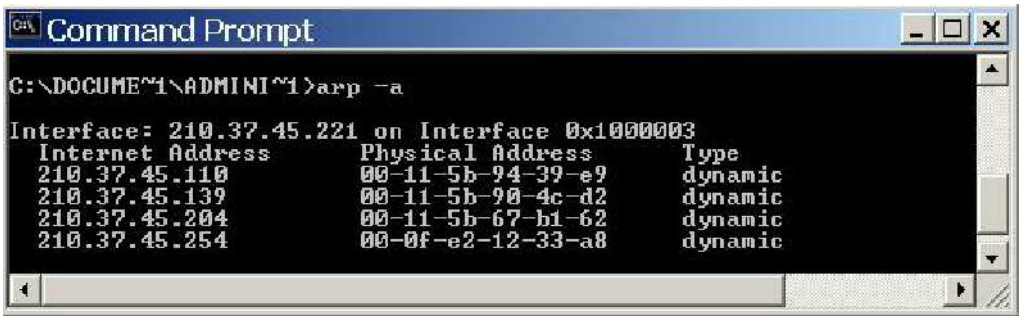
- 每台主机都有一个 ARP 高速缓存，存放着最近使用过的IP地址到硬件地址的映射记录，这些记录的生存时间一般为10~20分钟。

这个时间设置得太大或太小会出现什么问题？

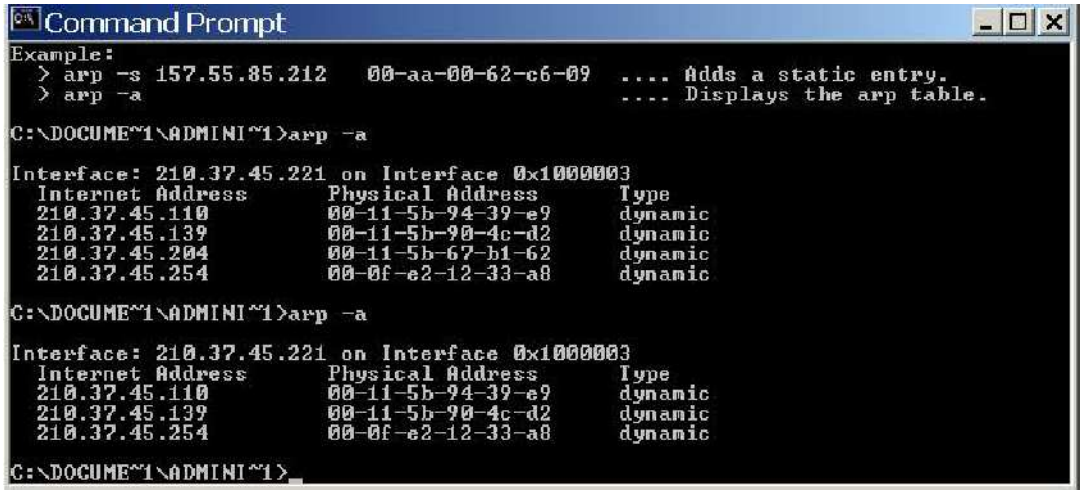
当网络中的某个**IP**地址和硬件地址的映射发生变化时，**ARP**缓存中的相应项目就要改变。例如，更换以太网网卡就会发生这样的事情。**10~20**分钟更换一块网卡是合理的。生存时间太短会使**ARP**请求和响应分组的通信量太频繁，而生存时间太长会使更换网卡后的主机迟迟无法和网络上的其他主机通信。

高速缓存中的ARP表

使用 arp –a 命令查看ARP表



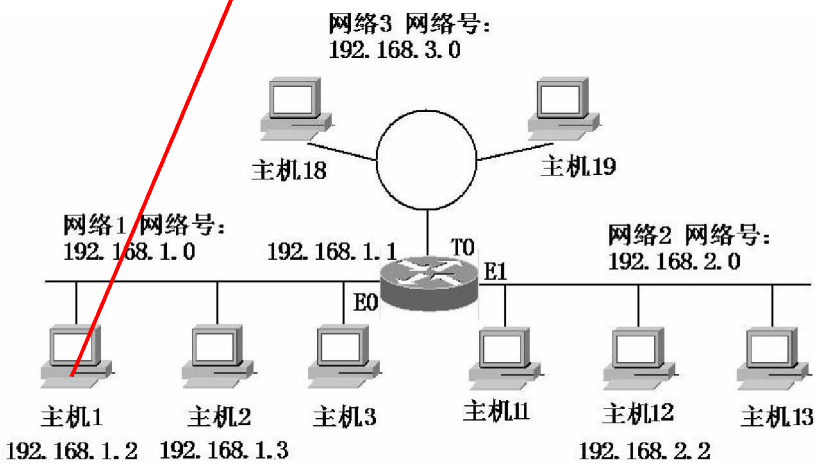
几分钟后再执行 arp –a 命令



本地ARP工作原理示例

- 主机1以主机2的IP地址为目标IP地址，以自己的IP地址为源IP地址封装了一个IP数据包；
- 数据包发送以前，主机1通过将子网掩码和源IP地址及目标IP地址进行求“与”操作，判断出源和目标在同一网络中；
- 主机1转向查找本地的ARP缓存，以确定在缓存中是否有主机2的IP地址与MAC地址的映射信息；
- 若在缓存中存在主机2的MAC地址信息，则主机1的网卡立即以主机2的MAC地址为目标MAC地址、以其自己的MAC地址为源MAC地址进行帧的封装并启动帧的发送；

物理地址	IP地址
02-60-8c-01-d1-10	192.168.1.2
????????	192.168.1.3

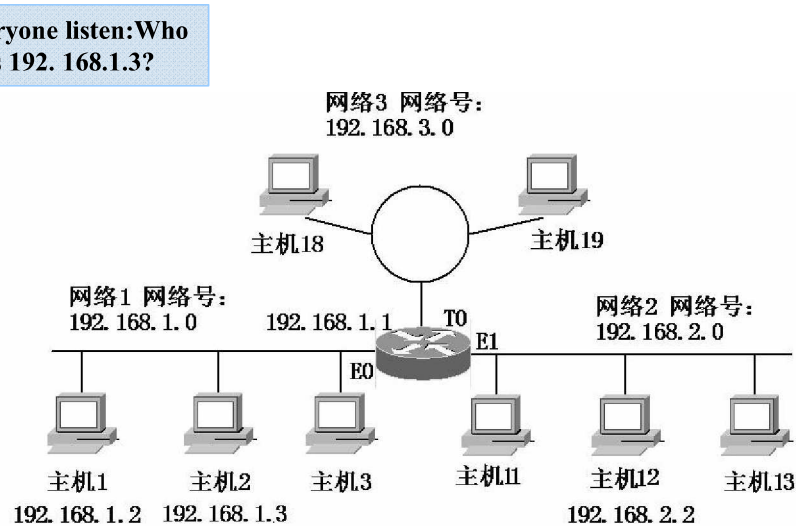


Look up ARP table:Do you know the MAC of 192.168.1.3?

- 若在缓存中不存在主机2的MAC地址映射信息，则主机1以广播帧形式发送一个ARP请求(ARP request)，位于同一网络中的所有节点都能够接收：

- 该广播帧中48位的目标MAC地址以全“1”即“ffffffffffff”表示，并在数据部分发出关于“谁的IP地址是192.168.1.3”的询问。

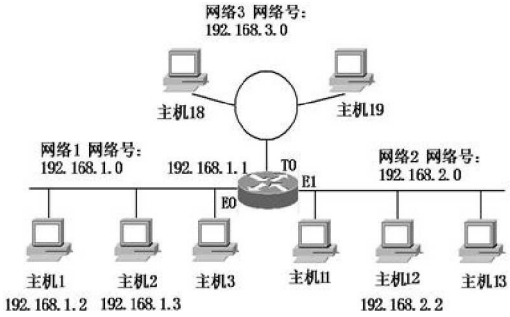
- 网络1中的所有主机都会收到该广播帧，并利用其中的源IP与MAC地址信息来更新自己的ARP表（ARP学习），同时检查自己的IP地址以判断自己是否为目标主机。



MAC头	IP头	数据信息
目标：ff-ff-ff-ff-ff-ff 源：02-60-8c-01-02-03	目标：192.168.1.3 源： 192.168.1.2	你的MAC地址是多少?

- 只有作为目标主机的主机2以自己的MAC地址信息为内容发给主机1一个ARP回应（单播方式）。
- 网络1中的所有主机都会收到该应答帧，并以其中的源IP与MAC地址信息来更新自己的ARP表(关于主机2的表项)，同时检查自己的IP地址以判断自己是否是目标主机。
- 主机1收到该回应后，首先将该其中的MAC地址信息加入到本地ARP缓存中，即ARP更新。

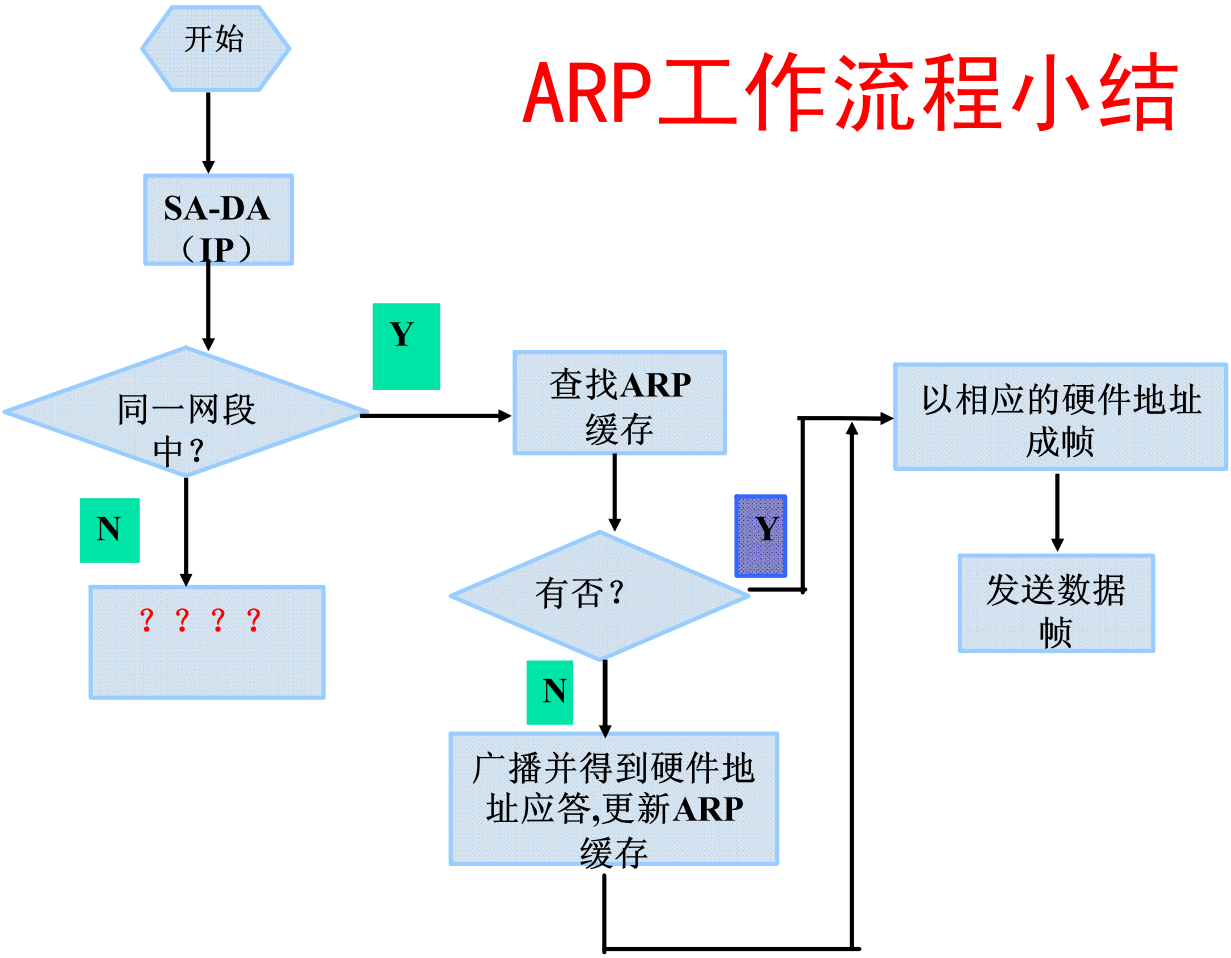
物理地址	IP地址
02-60-8c-01-d1-10	192.168.1.2
02-60-8c-01-a1-08	192.168.1.3



Hi! I'm 02-60-8c-01-a1-08

- 然后，主机1以主机2的MAC地址为目标MAC，以自己的MAC地址为源MAC，将要发送给主机2的IP数据包封装成帧，并启动发送。

ARP工作流程小结


开始
SA-DA
（IP）
同一网段
中？
Y
N
？ ？ ？ ？
查找ARP
缓存
有否？
Y
N
广播并得到硬件地
址应答,更新ARP
缓存
以相应的硬件地址
成帧
发送数据
帧

问题举例

- 若主机1要给位于另一个网段中的主机5发送数据，是否也是采用本地ARP的方式？

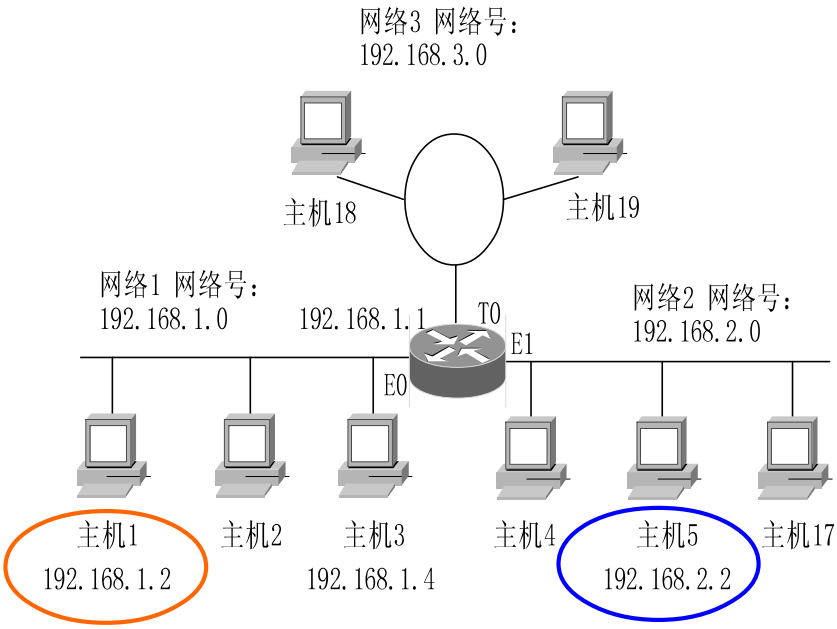
- 回答是否定的→

  第二层广播(如以太网帧的广播)是不可能被第三层设备路由器转发的。所有的第二层广播都会被路由器丢弃→路由器不处理和转发第二层的广播帧。

- 解决方案：

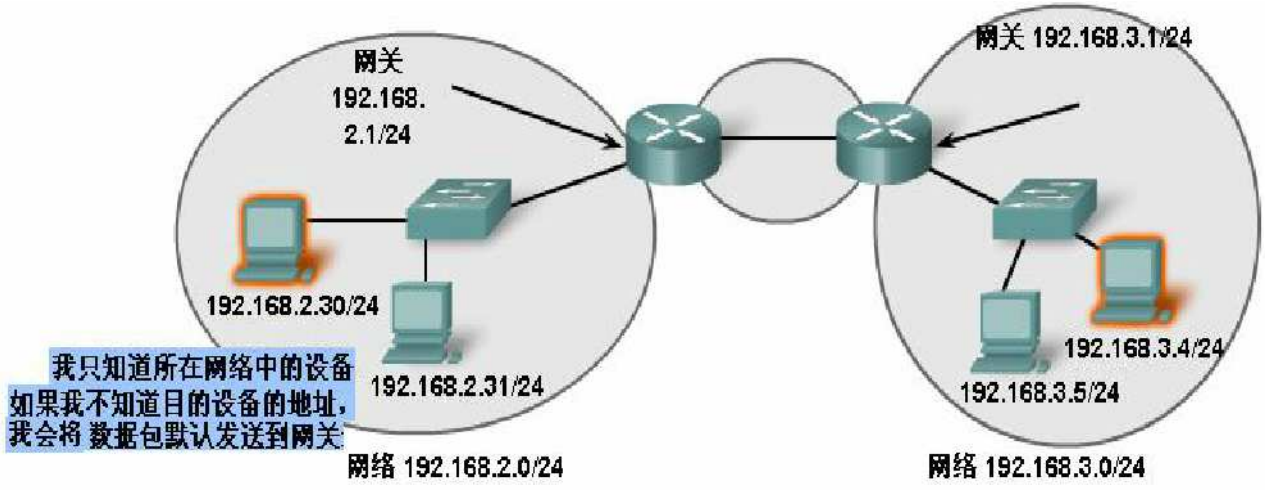
  Default gateway

  （默认网关）



缺省网关

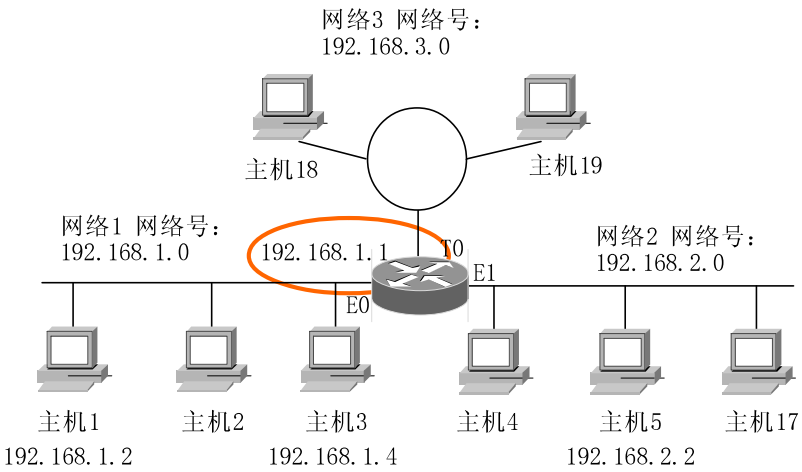
- default gateway：与主机位于同一网段中的某个路由器接口(或交换机虚接口)的IP地址 。

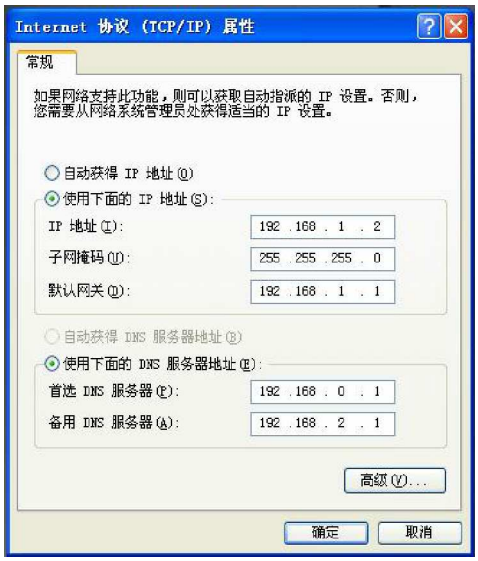


- 主机的IP配置选项，一旦被配置，则参数设置被作为主机配置的一部分保存起来。

- **主机1、2和3的缺省网关该如何设置？→**

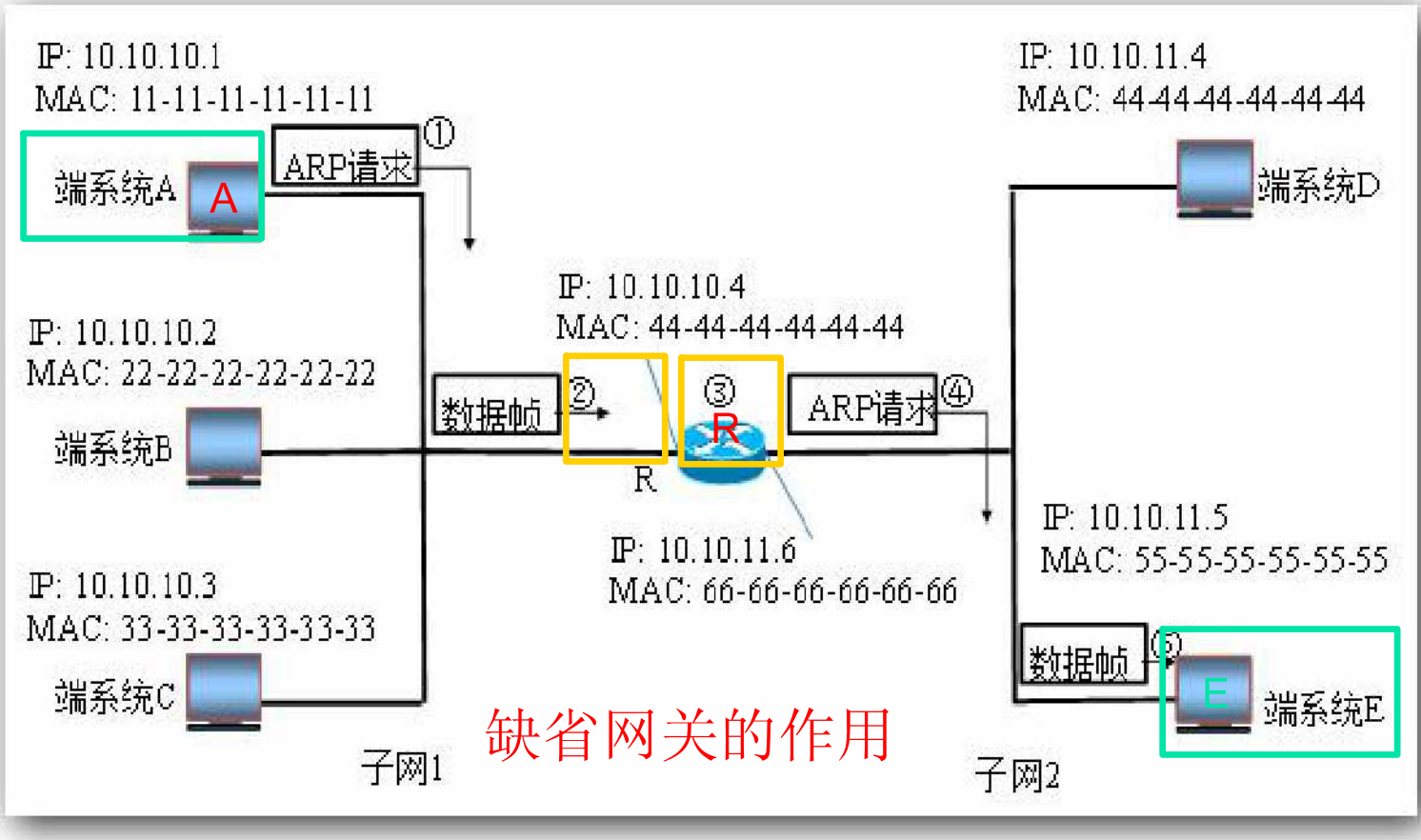
  路由器的以太网接口**E0**的**IP**地址，即**192.168.1.1**。

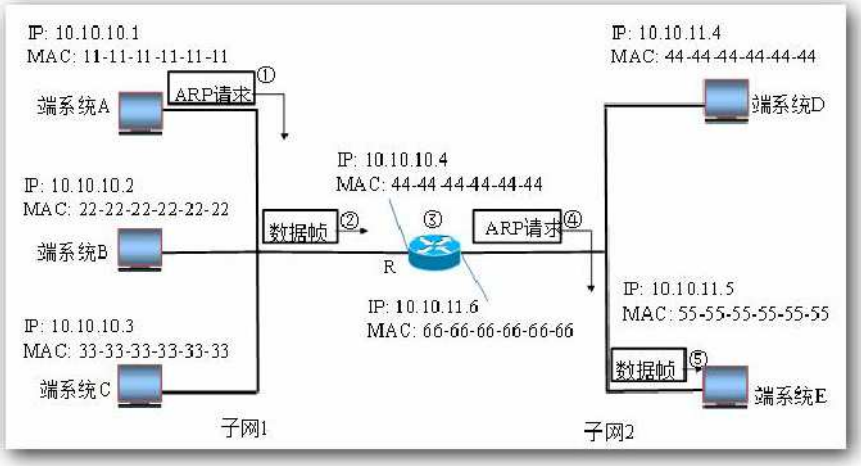




- 若主机上未进行网关设置则任何数据包不可能被送到缺省网关。

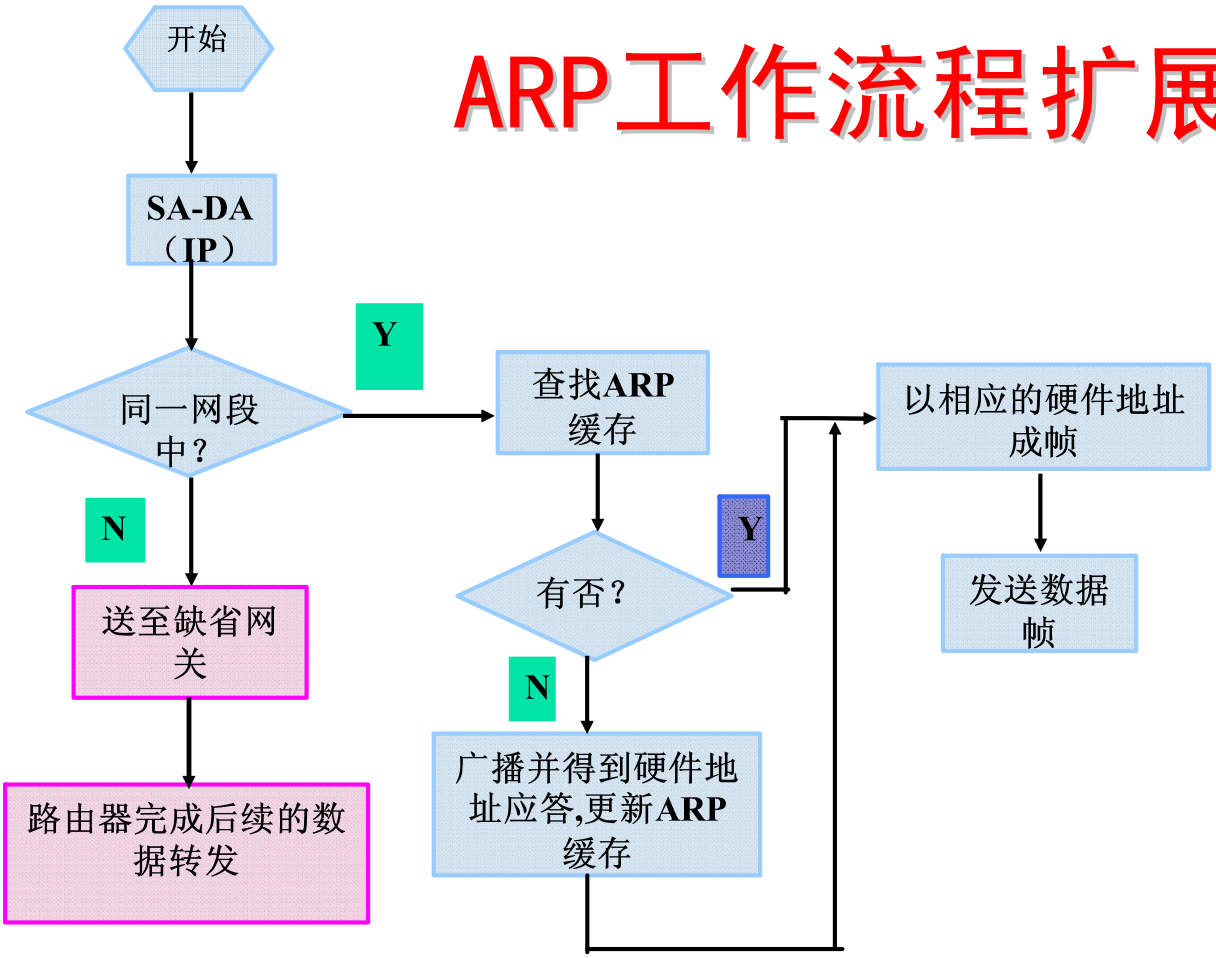
ARP：结点在不同LAN





1. A比较E的网络地址，发现不在相同网络，送路由器R(间接)
2. A使用ARP从10.10.10.4得到R的MAC地址
3. A生成以R的MAC地址作为目的地的链路层帧,帧包含A到E IP 数据报
4. A的适配器发送帧，R的适配器接收帧
5. R从帧中看到它目的地是E，使用选路协议确定转发端口
6. R出端口发现E在右侧网络，用ARP得到E的MAC(直接)
7. R生成包含A到E IP数据报的帧向E发送
8. E收到来自A的IP分组

ARP工作流程扩展


开始
SA-DA（IP）
同一网段中？
Y
N
查找ARP缓存
有否？
Y
N
以相应的硬件地址成帧
发送数据帧
送至缺省网关
路由器完成后续的数据转发
广播并得到硬件地址应答,更新ARP缓存