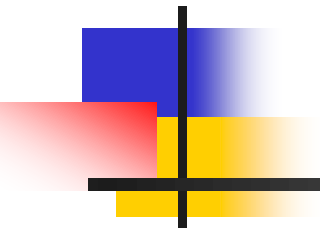


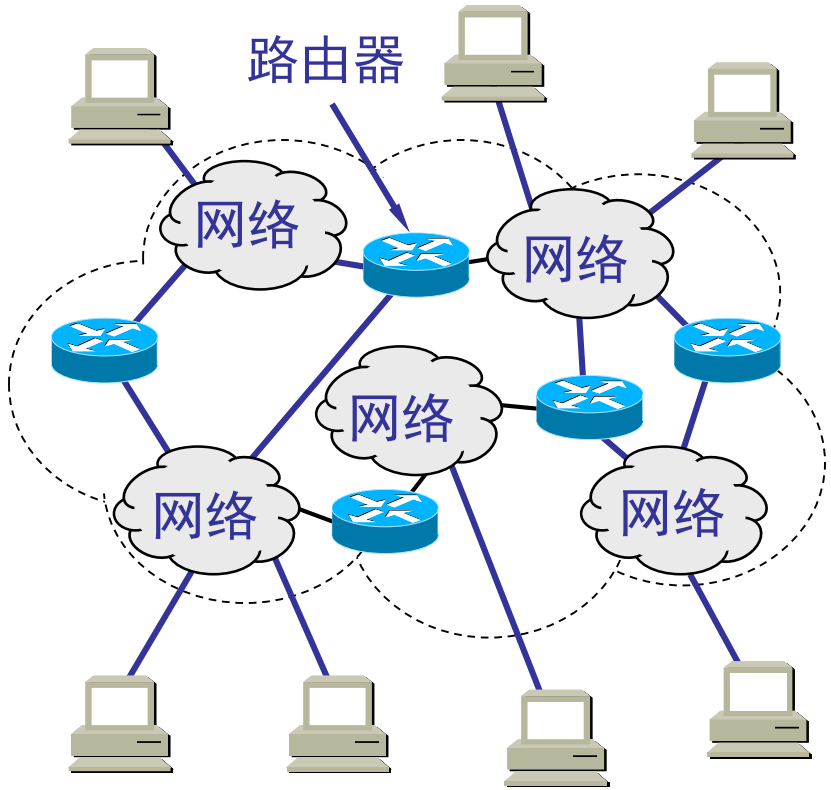
网络层：网络互联

刘志敏

liuzm@pku.edu.cn

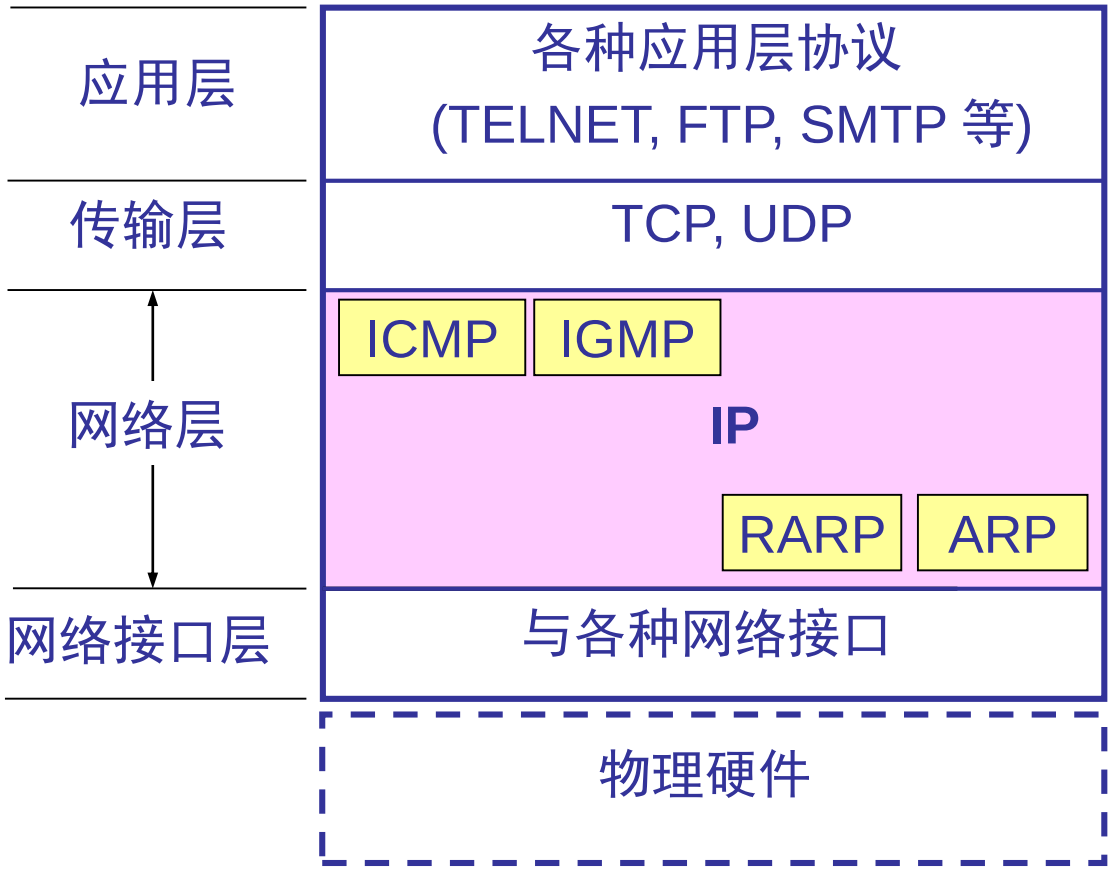
Internet是一种互联网

- 互联网：多个网络通过路由器互连
- 常见的互连设备
  - 物理层的转发器(repeater)；数据链路层的网桥或桥接器(bridge)；网络层的路由器(router)；网络层以上的网关(gateway)



网络层协议

- IP
- 地址解析协议 ARP (Address Resolution Protocol)
- 反向地址解析协议 RARP (Reverse Address Resolution Protocol)
- 互联网报文控制协议 ICMP (Internet Control Message Protocol)
- 互联网组管理协议 IGMP (Internet Group Management Protocol)



网络互联

- 需要解决的主要问题：
  - 地址分配
  - 分组传送
  - 路由与转发
  - 网络控制：超时控制、差错恢复、状态报告、拥塞检测与控制
- 互联网的核心协议是IP

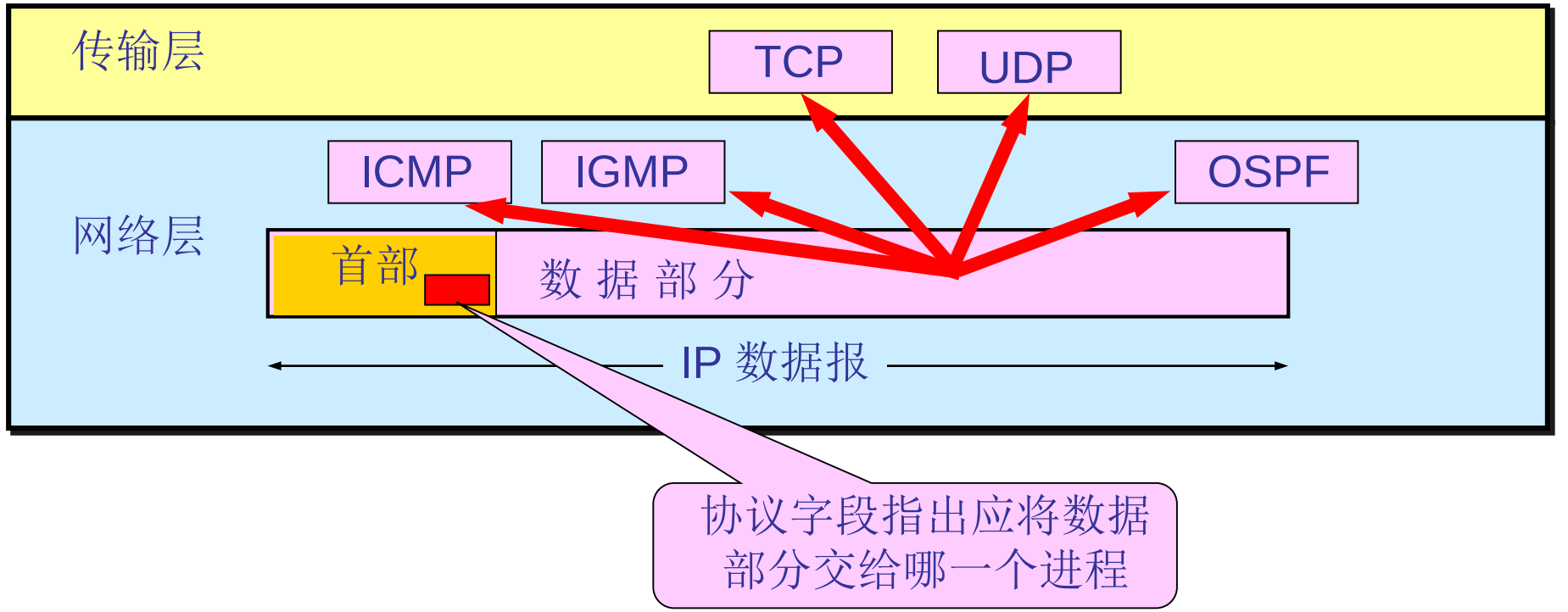
IP 分组格式

- 由首部和数据两部分组成，首部占 20 字节

IP 分组格式

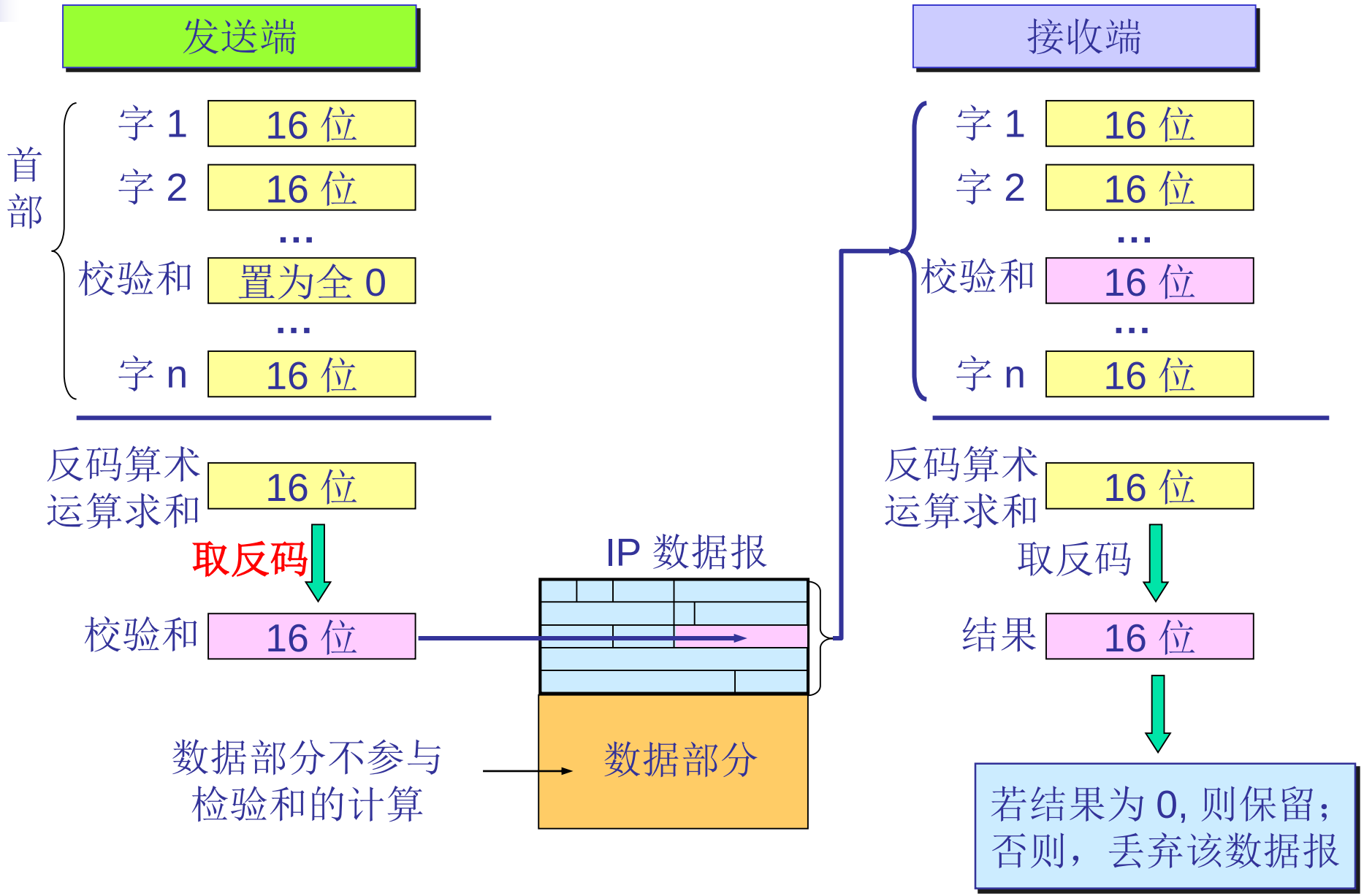
- 版本：占4位，IPV4为4
- 首部长度：占4位，**单位为4字节**
- **服务类型（TOS）：用于表示分组的类型，如实时性、可靠性、高吞吐量等**
- 总长度：占16位，指首部和数据的字节长度，总长度不超过MTU（最大传输单元，链路层上可传输的最大帧长度）
- 标识：占16 位，它是一个计数器，用来产生数据报的标识
- 标志：占3位，最低位 **MF** (More Fragment，1表示其后“有分片”，0表示最后分片）；中间位**DF** (Don‘t Fragment)，0表示允许分片
- 片偏移(13 位)：某片在原分组中的相对位置，单位为 8个字节
- 生存时间（TTL）：应大于0，经过路由器是TTL-1，为0时的分组被丢弃
- 协议：表示承载数据的协议类型
- 首部校验和：检验IP首部比特错，路由器校验并丢弃错误报文，重新计算
- 地址：注意网络序和主机序

IP首部中协议字段



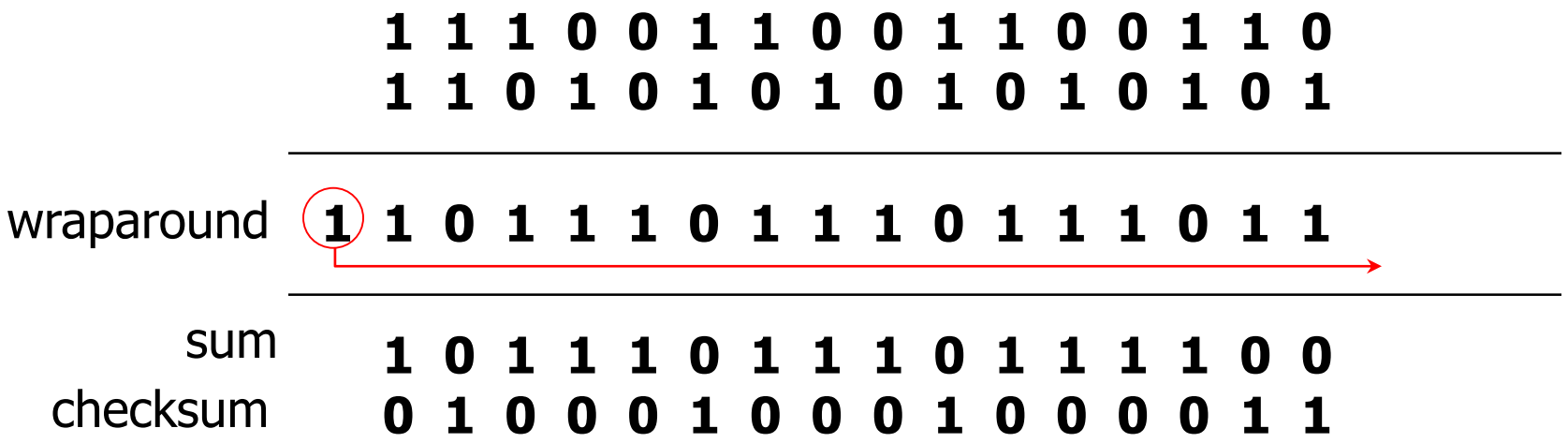
协议名	ICMP	IGMP	TCP	EGP	IGP	UDP	IPV6	OSPF
协议值	1	2	6	8	9	17	41	89

IP首部中校验和计算方法



校验和计算举例

- 反码算术运算求和：带进位的二进制加法运算，若最高位有进位，则结果+1，注意：最后一次运算若有溢出，就要回卷（在最低位+1）
- 例如：2个16位数的校验和



参考程序：计算校验和

```
unsigned short int checksum(unsigned short int *pBuffer, int length)
{
  //计算校验和
  unsigned int sum = 0;
  for (int i = 0; i < length; i++){
     sum += ntohs(pBuffer[i]); // calculate the sum
     sum = (sum >> 16) + (sum & 0xffff);
        // wrap around when overflow
   }
   return sum;
}
```

IP 地址

- 地址及标识：身份证、固定电话号码、学号等
  - 按一定的规则编码，编号唯一
- IP地址是连接在互联网上的主机（或路由器）的惟一标识；IPV4占32位，地址数为2^{32}，IPV6占128位，地址数为2^{128}
- IP地址的编址方法（以IPV4为例）
  - 分类IP地址：最基本编址方法
  - 子网划分：对最基本编址方法的改进
  - 构成超网：较新的无分类编址方法
- IP地址管理：由互联网域名和地址分配机构ICANN负责

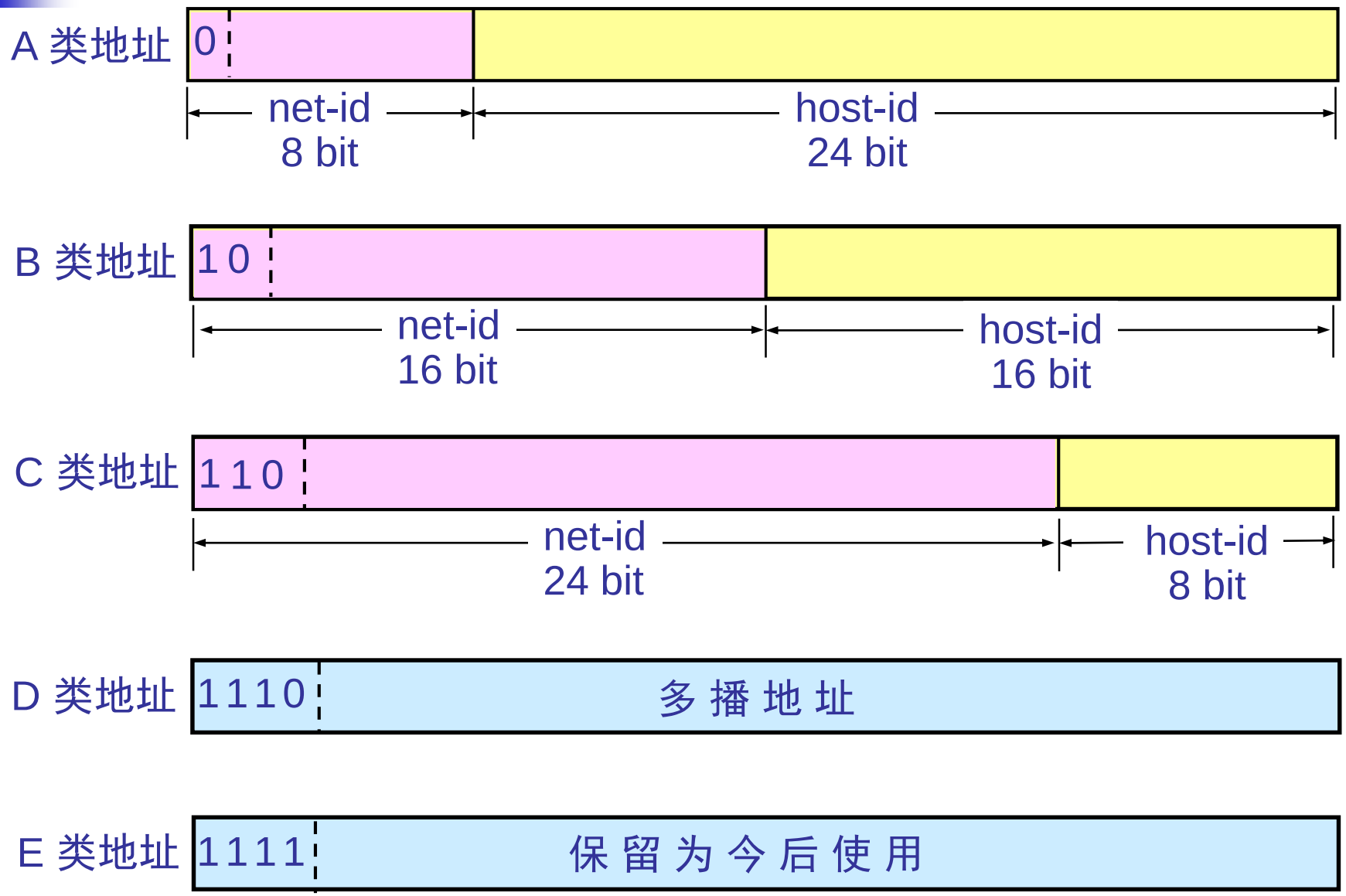
分类 IP 地址

- 将IP地址分为A、B、C、D、E类，每类地址都由两个固定长度的字段组成，网络号 net-id 标志主机所连接到的网络，主机号 host-id 标志该主机。
- 两级的 IP 地址记为：

IP 地址 ::= { <网络号>, <主机号>}

::= 代表“定义为”

IP 地址中的网络字段和主机字段

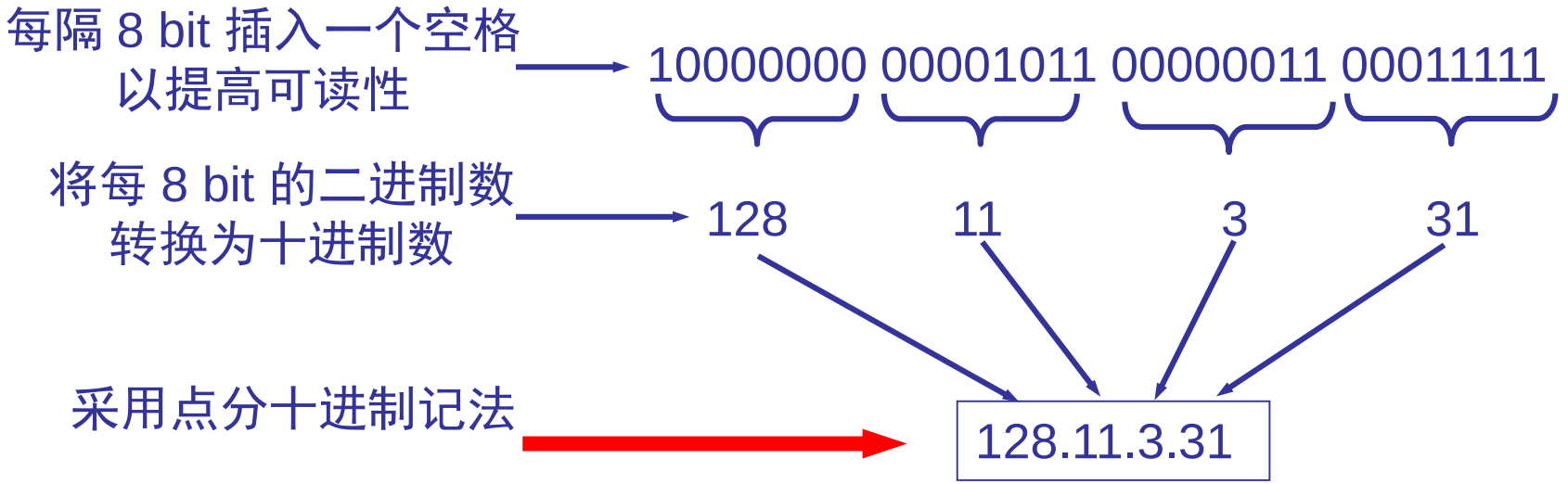


IP 地址的特点

- IP地址是一种分级结构，只分配网络号，主机号由网络所属单位分配
- IP地址标志主机与链路的接口，路由器至少连接两个网络，有两个以上的IP地址
- 不使用的特殊IP

网络号	主机号	源地址	目的地址	含义
0	0	可以	不可以	本网络的本主机，用于DHCP
0	Host-id	可以	不可以	本网络的主机Host-d
全1	全1	不可以	可以	本网络上广播
Net-id	全1	不可以	可以	对Net-id的所有主机广播
127	非全0或非全1	可以	可以	用作本地软件环回测试

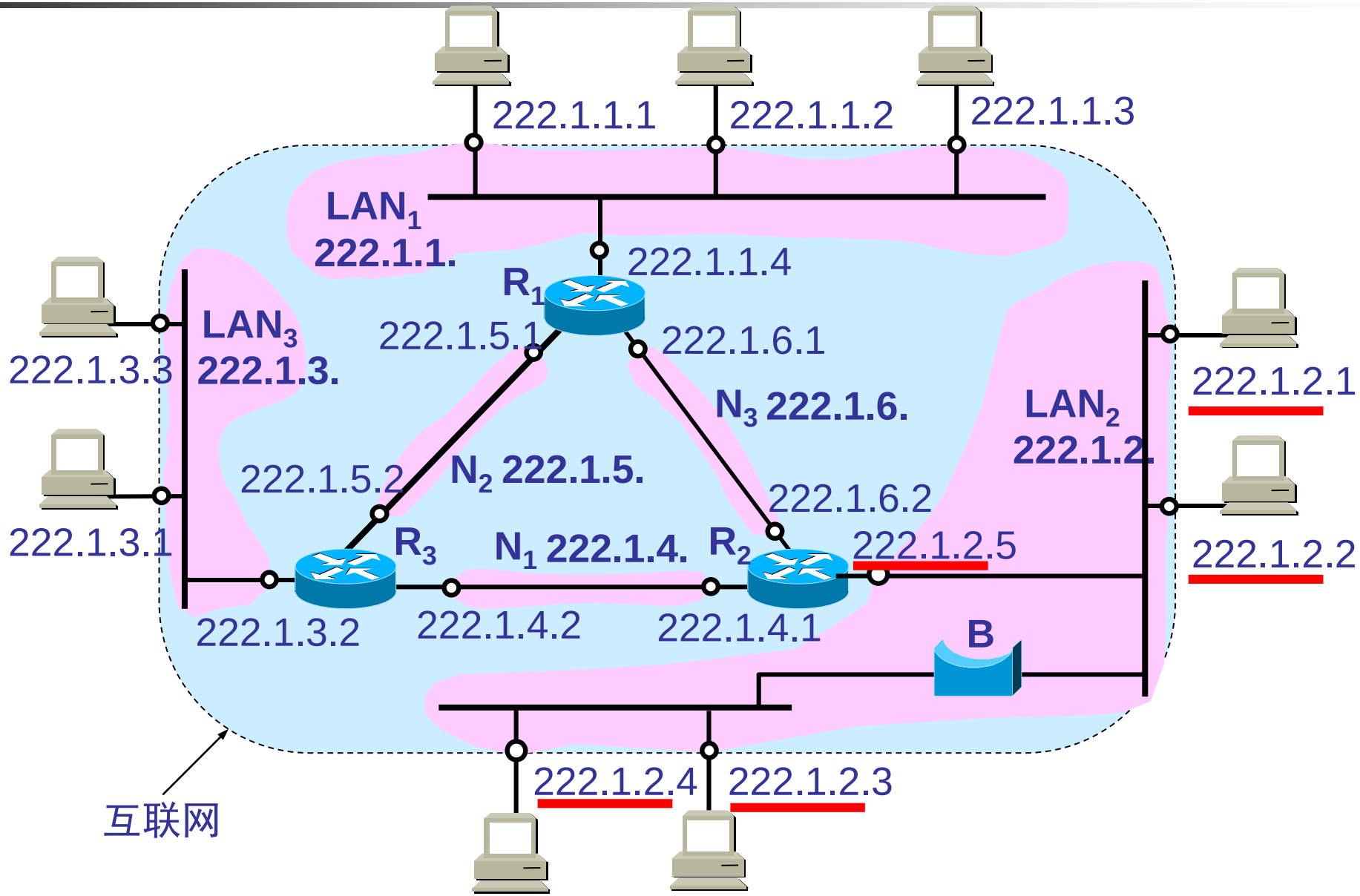
常用的三类 IP 地址



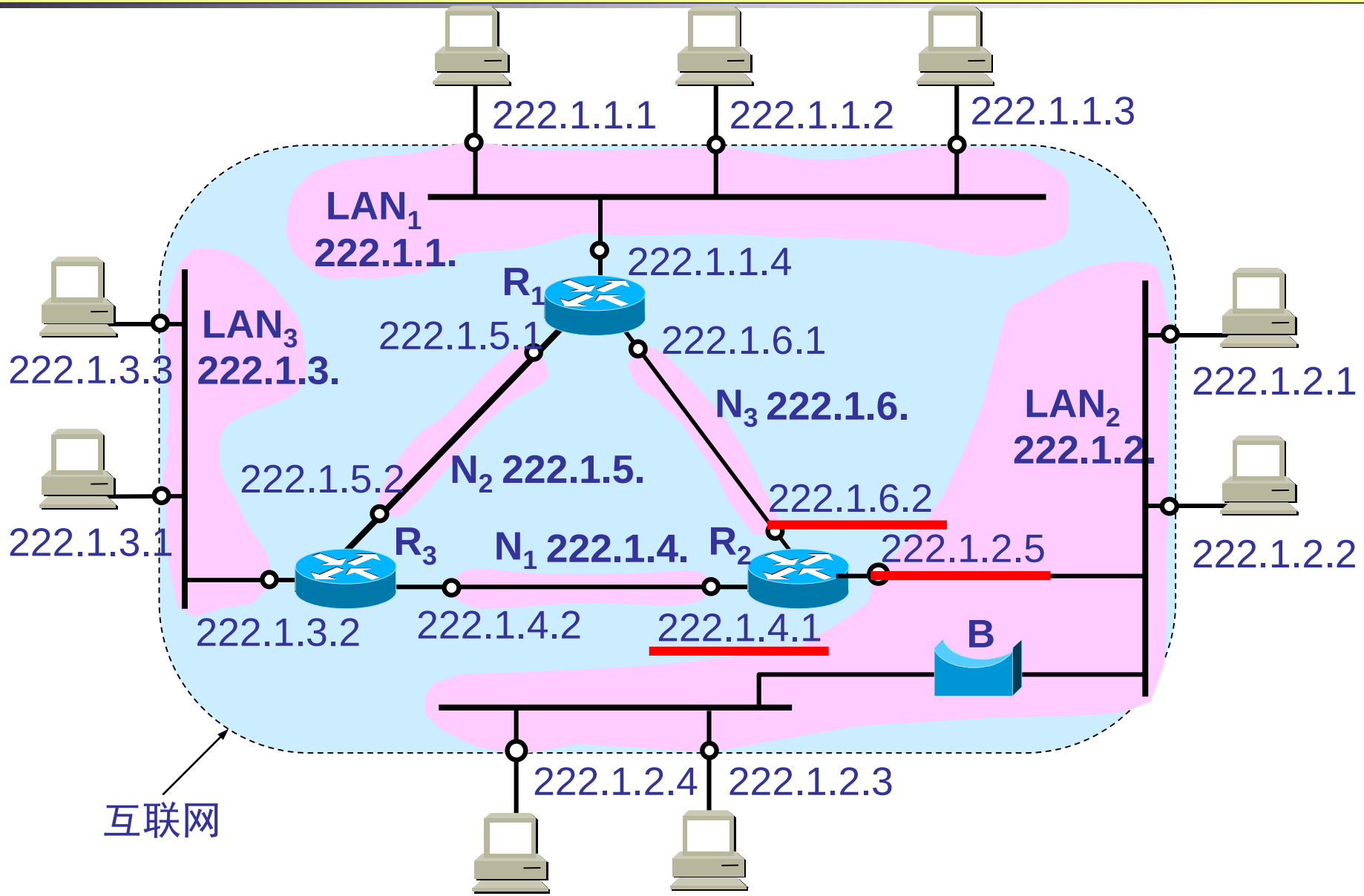
网络类别	最大网络数	第1个网络号	最后1个网络号	每个网络最多主机数
A	126 (=2^7 – 2)	1	126	16,777,214 (=2^{24} – 2)
B	16,384(=2^{14}-1)	128.1	191.255	65,534 (=2^{16} – 2)
C	2,097,152(=2^{21}-1)	192.0.1	223.255.255	254 (=2^8 – 2)

注释：
主机地址：全0表示本主机，全1表示所有主机，不能作为主机地址
网络地址：全0表示本地网络，不能作为网络地址；127用于本地软件环路测试

在同一网络上的主机或路由器，其IP 地址的网络号必须相同

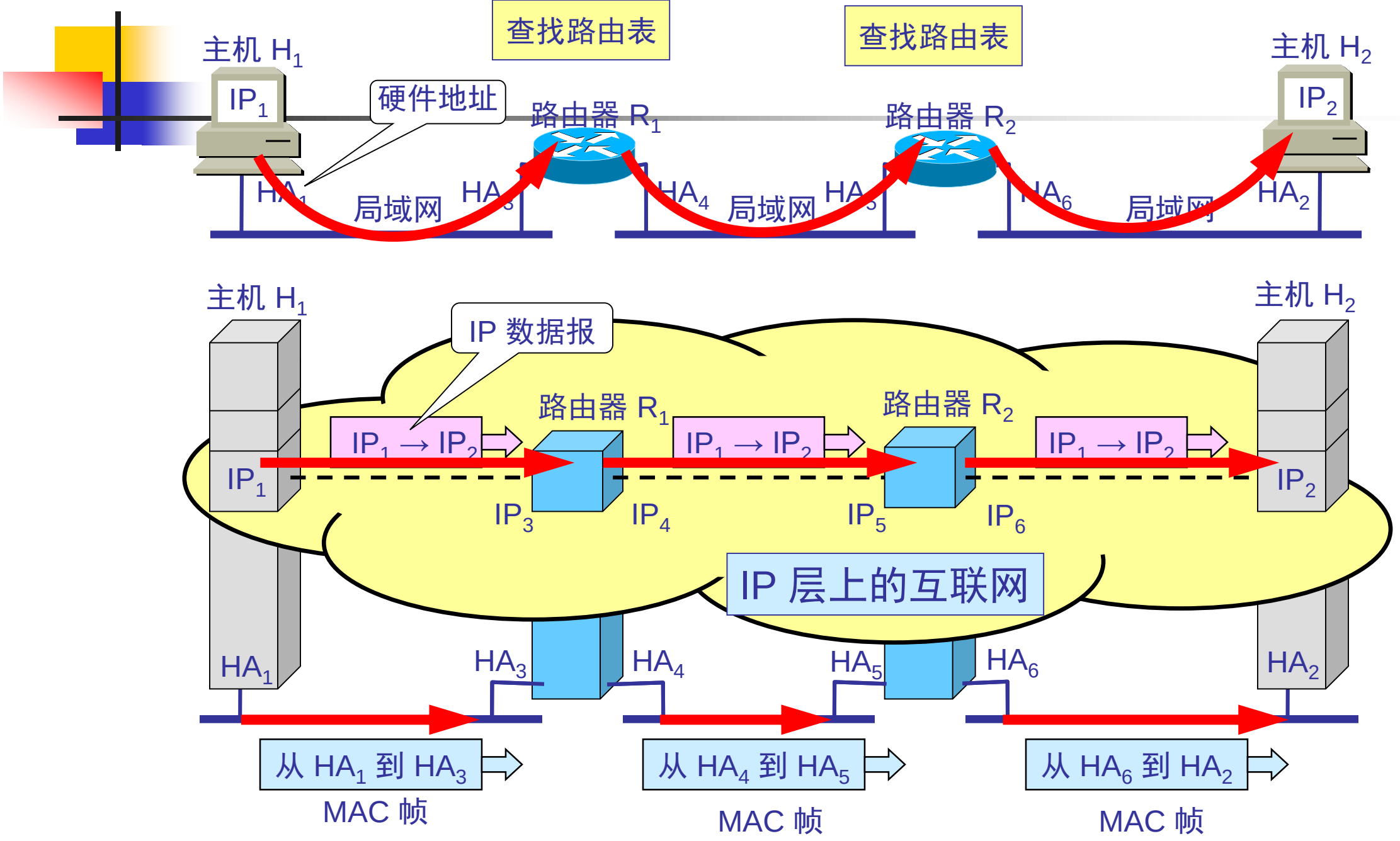


路由器有两个以上的接口，每个接口的IP地址的网络号不同



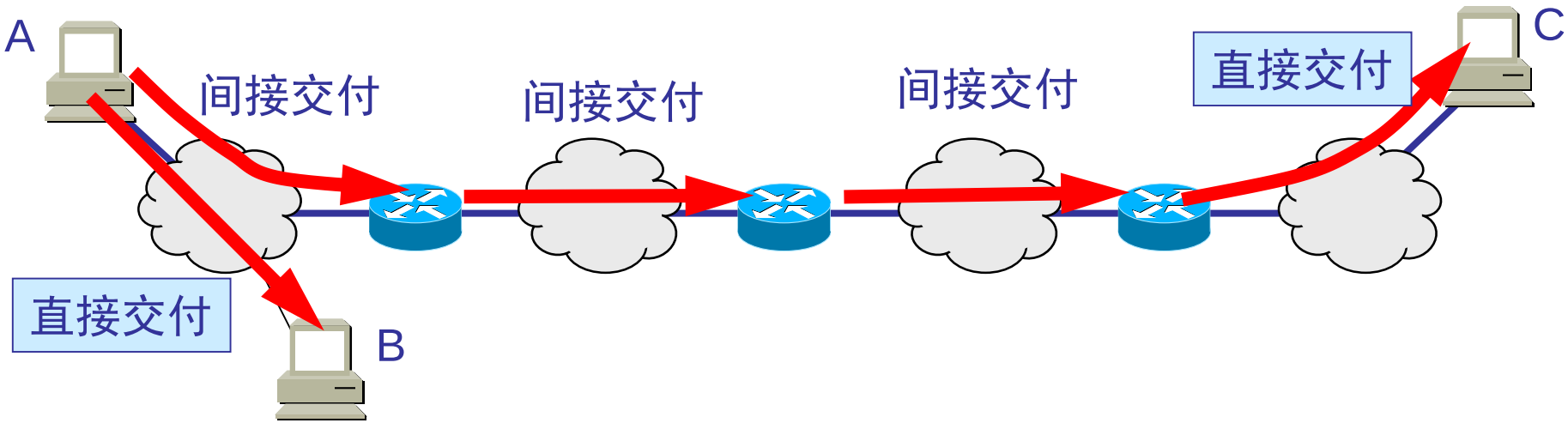
网络互联

- 需要解决的主要问题：
  - 地址分配
  - 分组传送
  - 路由与转发
  - 网络控制：超时控制、差错恢复、状态报告、拥塞检测与控制
- 互联网的核心协议是IP

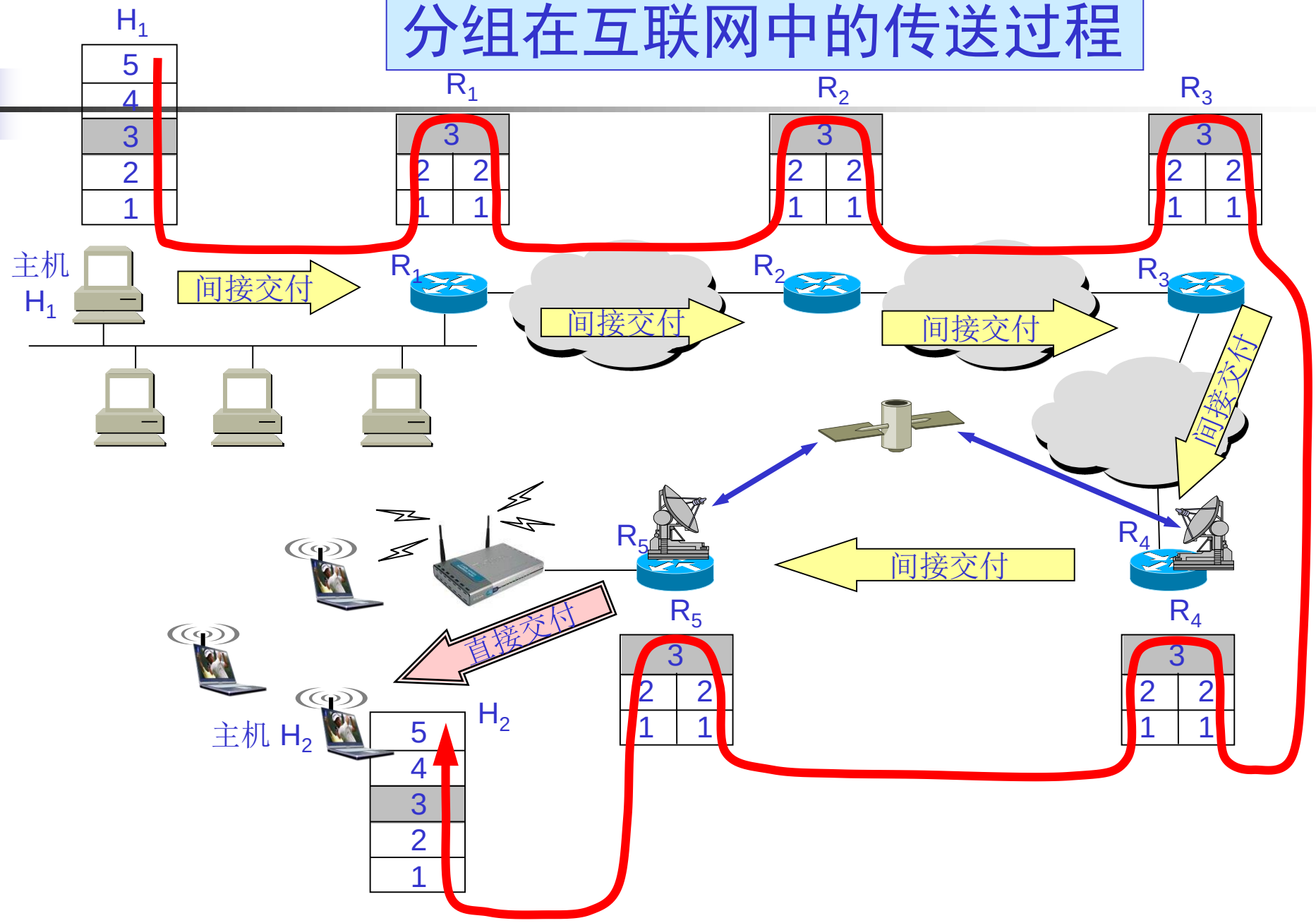

查找路由表
查找路由表
主机 H_1
主机 H_2
IP_1
IP_2
硬件地址
路由器 R_1
路由器 R_2
HA_1
局域网
HA_3
HA_4
局域网
HA_5
HA_6
局域网
HA_2
主机 H_1
主机 H_2
IP 数据报
路由器 R_1
路由器 R_2
$IP_1 \rightarrow IP_2$
$IP_1 \rightarrow IP_2$
$IP_1 \rightarrow IP_2$
IP_1
IP_2
IP_3
IP_4
IP_5
IP_6
IP 层上的互联网
HA_1
HA_3
HA_4
HA_5
HA_6
HA_2
从 HA_1 到 HA_3
MAC 帧
从 HA_4 到 HA_5
MAC 帧
从 HA_6 到 HA_2
MAC 帧

分组传送：直接交付或间接交付

- 当主机A要向主机B发送分组时，先检查主机B是否与其在同一网络上。如果是，就直接交付；否则，将间接交付，即将分组发送给本网络上某个路由器，由路由器负责转发。
- 如何判决主机A与B是否在同一网络上？
  - 检查主机A与主机B的网络地址是否相同
- 谁负责判决？发送分组的主机

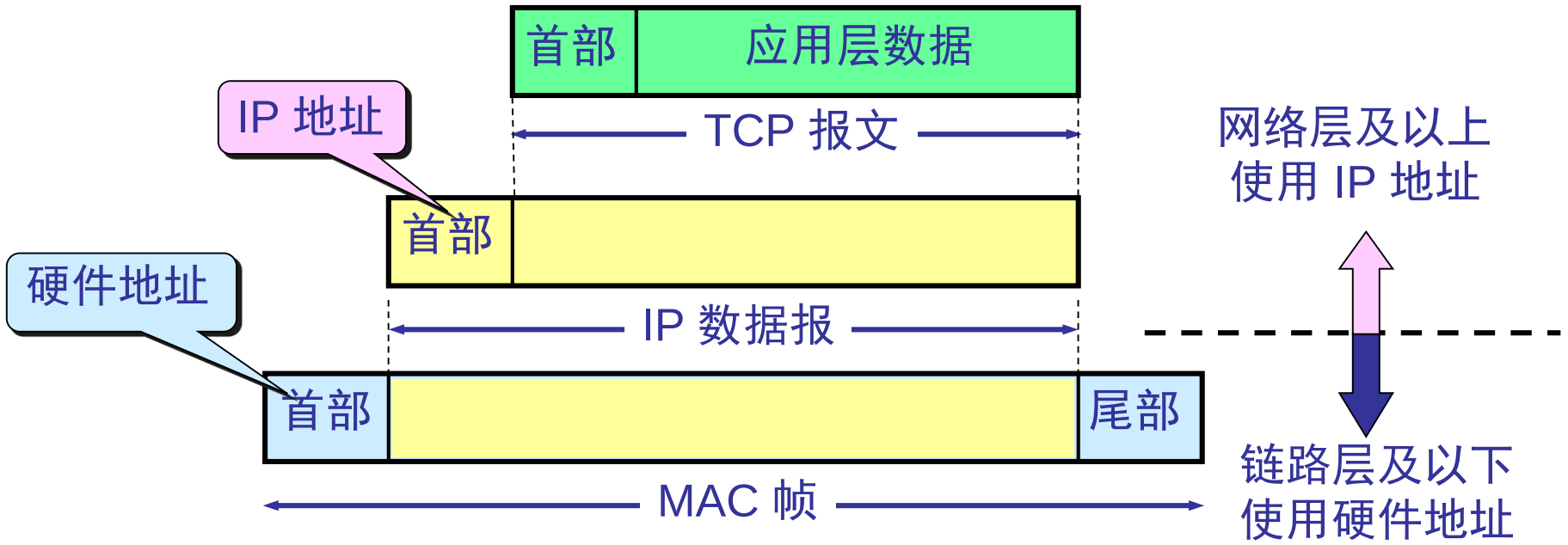


分组在互联网中的传送过程



IP 地址与硬件地址

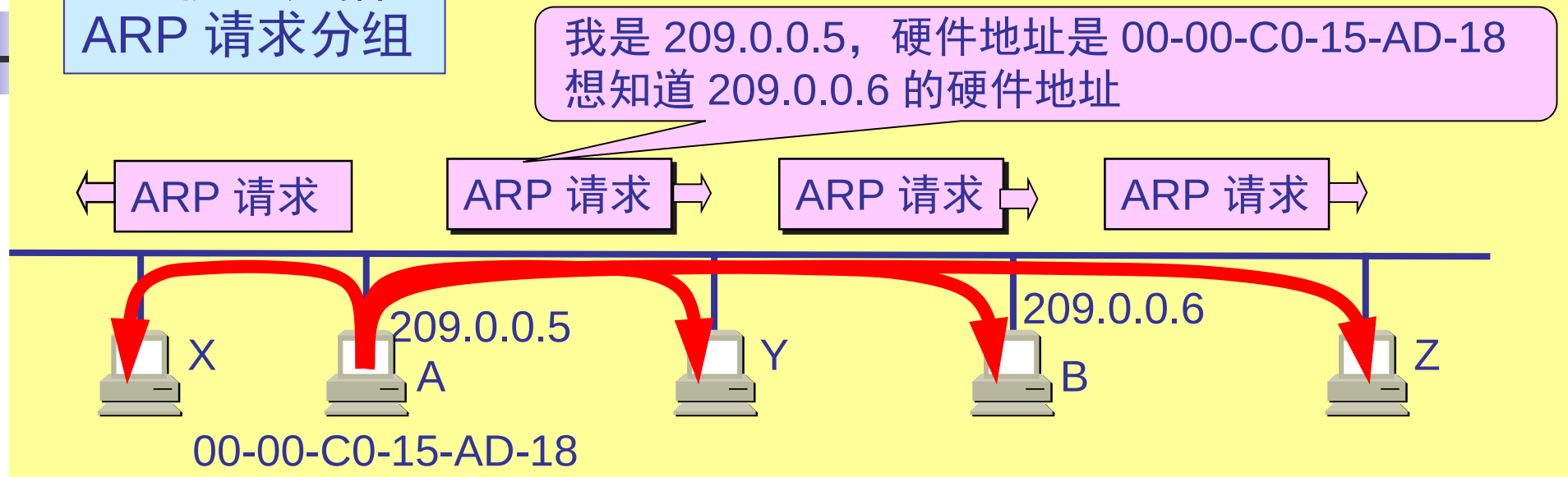
- 分组传送：或直接交付或间接交付
- 如何将IP分组交付给接收设备？将分组封装到帧中
- 已知接收设备的IP地址，需要知道该IP设备对应的MAC地址，采用ARP



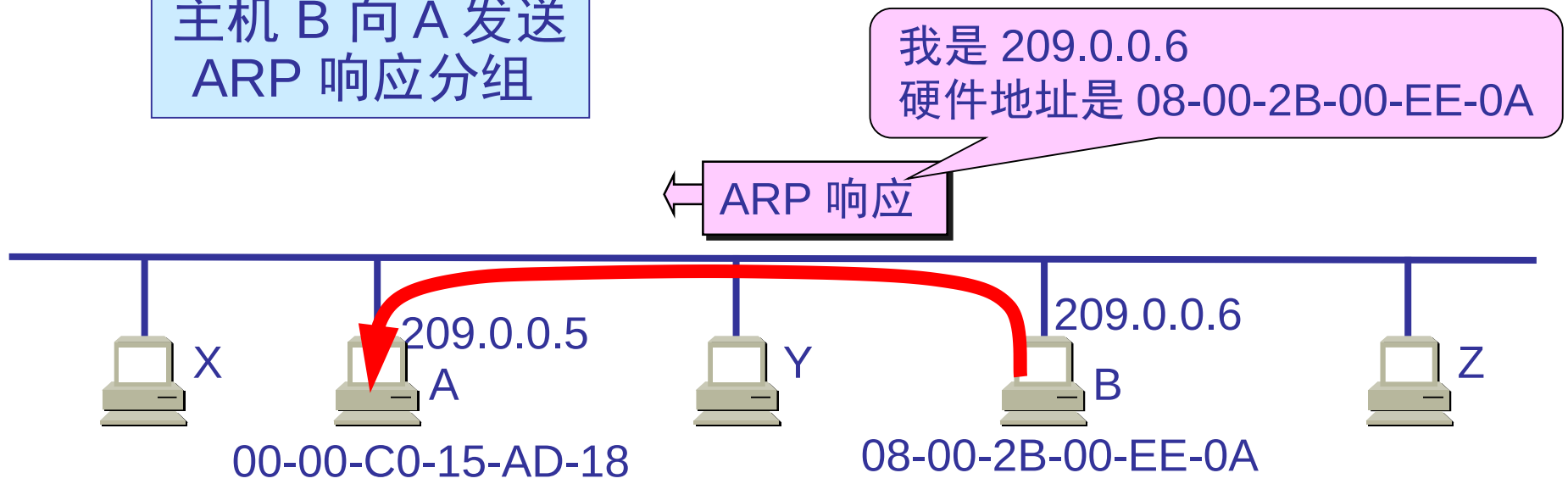
ARP

- 在网络层上传输IP分组，用IP地址；而在链路上传送数据帧，必须使用硬件地址。
- 需要建立网络地址与硬件地址的映射关系
- 地址解析协议 ARP：解决在同一局域网（子网）上主机 IP 地址与网卡硬件地址，即MAC地址之间的映射。根据IP地址找其对应的MAC。
- 反向地址解析协议 RARP：已知主机的硬件地址，而要找到其 IP 地址。这种主机往往是无盘工作站。

主机 A 广播
ARP 请求分组



主机 B 向 A 发送
ARP 响应分组



ARP

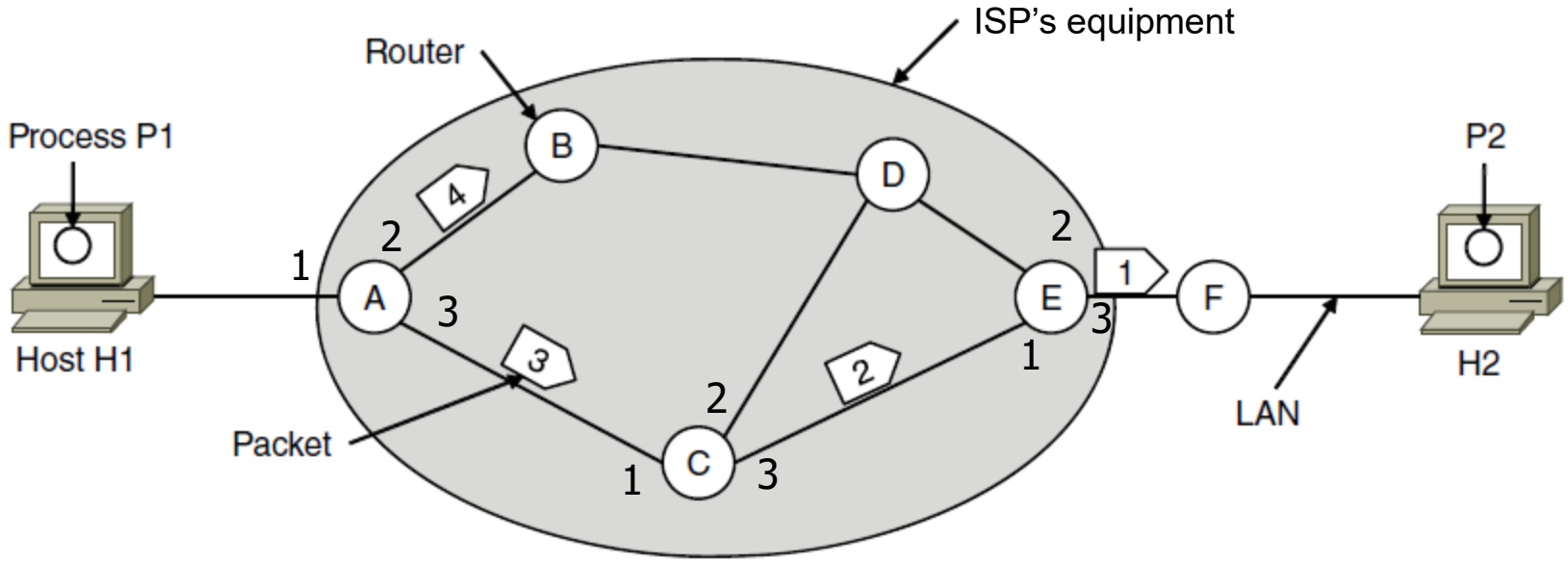
- ARP解决同一个网络上主机或路由器的IP地址和硬件地址的映射问题。若目的主机与源主机位于不同的网络，则源主机发送分组给其路由器，由路由器转发。
- ARP高速缓存：主机存储IP地址到硬件地址的映射表，减少发送ARP请求的机会。
- IP地址到硬件地址的解析是自动进行的
- 互联网为何不直接用硬件地址通信？
  - 网络种类多，硬件种类也多，转换很复杂
    - 交换机基于硬件地址进行数据交换，自学习过程需要泛洪flooding
  - 用IP地址，用户可设置，使用方便

网络互联

- 需要解决的主要问题：
  - 地址分配
  - 分组传送
  - 路由与转发
  - 网络控制：超时控制、差错恢复、状态报告、拥塞检测与控制
- 互联网的核心协议是IP

网络层提供无连接服务

- 只要更新转发表，分组的路由就改变了



A's table (initially)

目的	接口
A	--
B	2
C	3
D	2
E	3
F	3

A's table (later)

目的	接口
A	--
B	2
C	3
D	2
E	2
F	3

C's Table

目的	接口
A	1
B	1
C	-
D	3
E	3
F	3

E's Table

目的	接口
A	1
B	2
C	1
D	2
E	--
F	3

路由表：目的地址 接口

Destination Address Range	Link Interface
11001000 00010111 00010000 00000000 through 11001000 00010111 00010111 11111111	0
11001000 00010111 00011000 00000000 through 11001000 00010111 00011000 11111111	1
11001000 00010111 00011001 00000000 through 11001000 00010111 00011111 11111111	2
otherwise	3

地址数量多，记录数多，占用更多的内存，查表时间长

更小的路由表：匹配前缀 接口

Prefix Match	Link Interface
11001000 00010111 00010	0
11001000 00010111 00011000	1
11001000 00010111 00011	2
otherwise	3

路由器基于目的地址的匹配前缀（而非目的地址）查找转发端口；大大减少路由表记录数，加快查表速度

Examples

DA: 11001000 00010111 00010110 10100001 Which interface?

DA: 11001000 00010111 00011000 10101010 Which interface?

最长匹配原则：选择更长匹配前缀项

路由与转发

- 转发表如何产生？
  - 路由器执行路由算法及路由协议
- 距离矢量路由算法，RIP支持
  - 每个路由器维护一张距离矢量路由表
  - 在邻居路由器之间交换表，路由表得到更新
  - 按照距离矢量路由算法，计算最短路径路由
- 链路状态路由算法，OSPF支持
  - 使用扩散法向所有路由器发送信息
  - 每个路由器获得完整的拓扑结构
  - 按照最短路径算法计算最短路径

RIP (Routing Information Protocol)

- RIP 是一种基于距离矢量的分布式路由协议
- 每个路由器维护一张路由表，其结构为“目的网络　跳数　下一跳”，记录该路由器到每个目的网络的距离（即跳数）以及路径（即下一跳）
  - 跳数：到直连网络的为1，到非直连网络的为所经过的路由器数加1
- 路由器只选择具有最少跳数的路由
- 一条路径最多包含15个路由器，“跳数”为16时表示不可达
- 仅与相邻路由器交换路由表信息
- 按固定时间间隔（例如30秒）交换路由信息

距离矢量算法

收到相邻路由器地址为X的RIP报文：

(1) 修改RIP报文中的所有项：将“下一跳”改为X，将“距离”值加1

(2) 对RIP报文中的每一项，重复以下步骤：

若“目的网络”不在路由表中，则将新表项加入路由表；

否则，

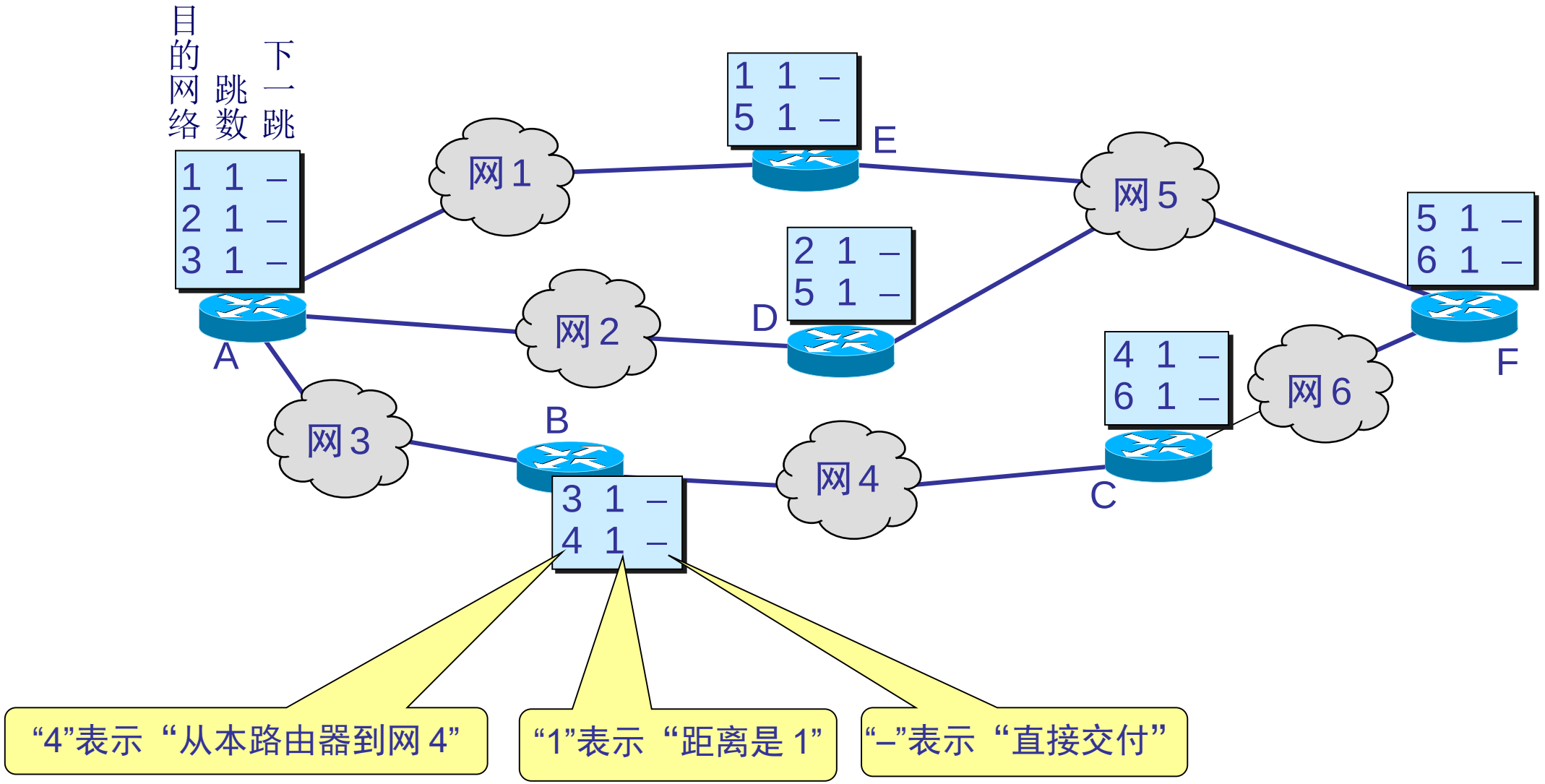
　　若“下一跳”地址相同，则用新表项替换原表项；

　　若新表项中的跳数更小，则更新；否则，忽略

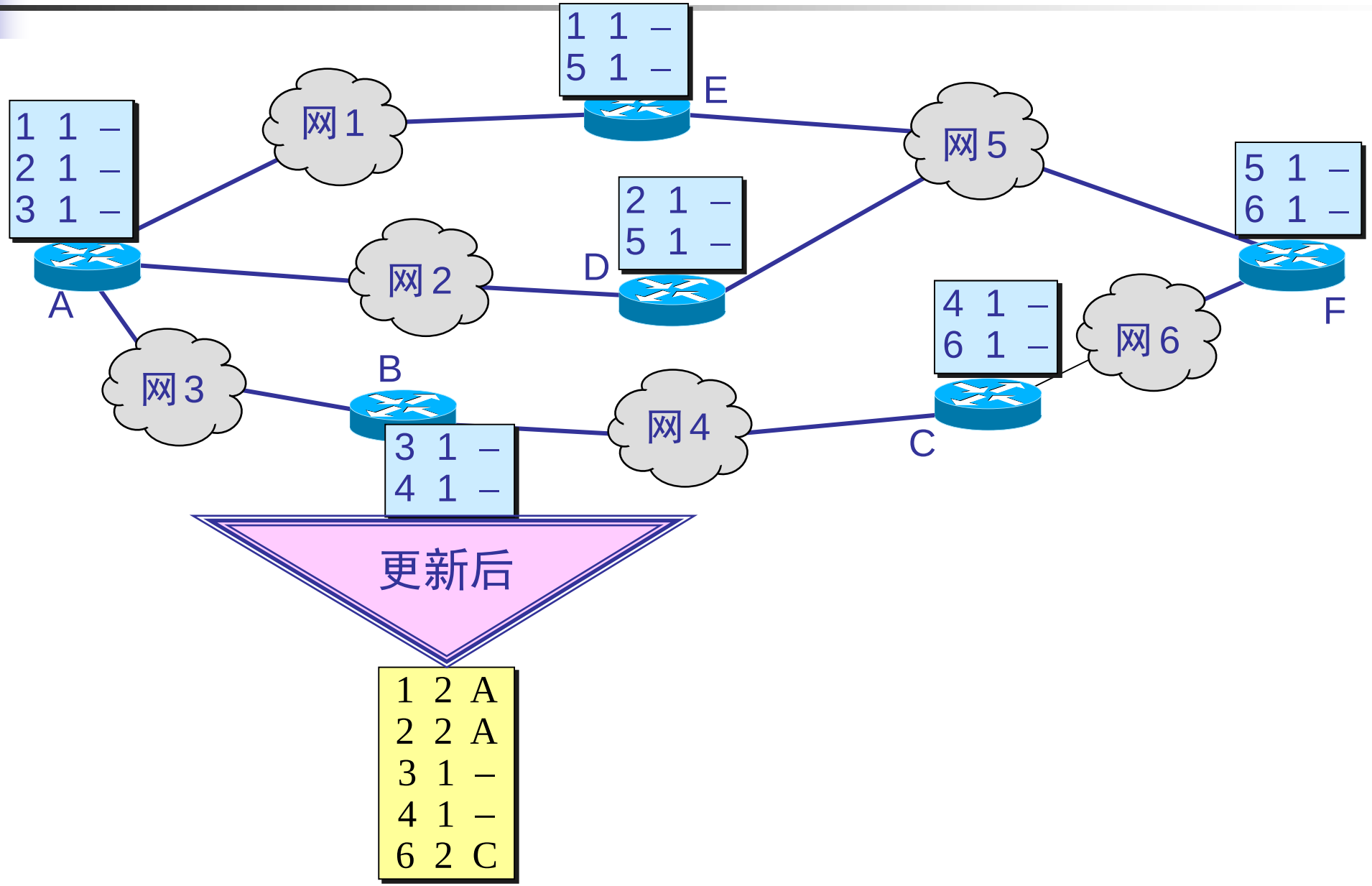
(3) 若3分钟未收到相邻路由器的RIP报文，则将此路由器记为不可达，即将距离置为16；

(4) 返回

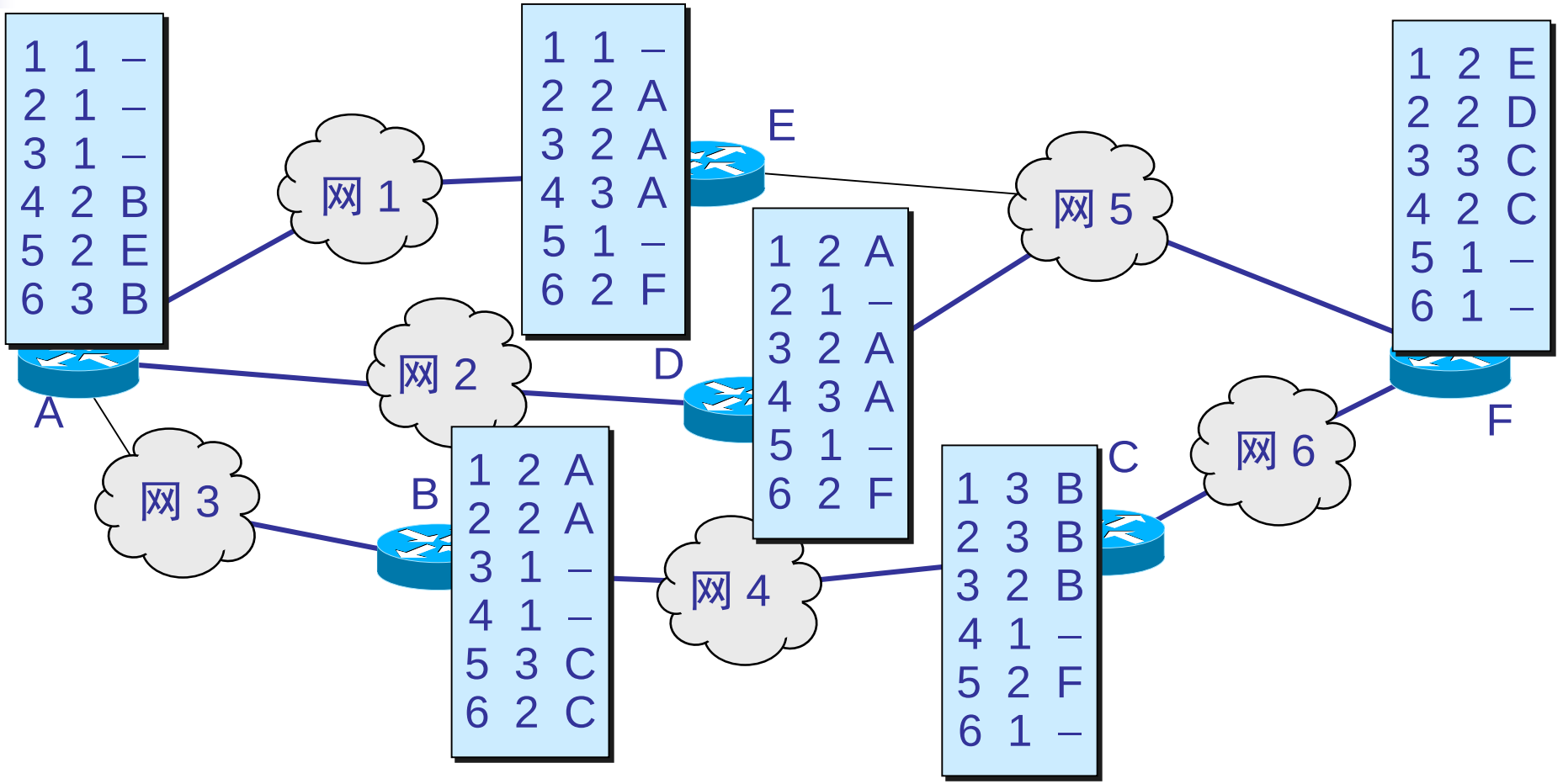
开始，各路由表只有到相邻路由器的信息



路由器 B 收到相邻路由器 A 和 C 的路由表



最终，所有路由器上路由表都更新了



路由器之间如何交换路由信息？

- 在路由器上运行的RIP进程，进程间通信使用UDP520端口
- RIP采用广播或组播方式交换路由消息
  - RIPv1使用广播，RIPv2使用组播224.0.0.9
- 主机只接收路由RIP报文但不发送
- RIP支持默认路由
- RIP的网络不超过15跳，适合于中小型网络
- RIPv1是有类（IP地址分类）路由协议，RIPv2是无类路由协议（，RIPv2报文中含有掩码信息）

路由器的转发算法

(1) 提取分组的目的IP地址 *D*, 得到目的网络地址 *N*

(2) 若 *N* 与路由器直连，则直接交付；否则，间接交付，执行(3)

(3) 若在转发表中有到目的地址 *D* 的特定路由，则将分组发送到对应的下一跳；否则，执行(4)

(4) 若在转发表中有到达 *N* 的路由，则将分组发送到对应的下一跳；否则，执行(5)

(5) 若在转发表中有一个默认路由，则将分组发送到默认路由器；否则，执行(6)

(6) 用ICMP报告转发分组出错

再议IP地址

- IP地址是一种分级结构，只分配网络号，主机号则由网络所属单位分配
- 路由器仅根据目的主机的网络号（而非目的地址）转发分组，使路由表项数大大减少
- 降低路由表项数，提升了路由器的查表速度
- IP地址的编址方法
  - 分类IP地址 是最基本编址方法
  - 子网划分 是对最基本编址方法的改进
  - 构成超网 是较新的无分类编址方法，得到推广应用。

划分子网

- A类B类网络，主机数过多；直接交付需要ARP，存在ARP广播风暴
- 在IP地址中增加“子网号”，称为划分子网
- 将主机地址的若干比特作为子网号
- 划分子网是将IP地址的主机地址再划分

  IP地址 ::= {<网络号>, <子网号>, <主机号>}
- 子网掩码：为1的部分表示子网地址，为0的部分表示主机地址
  - 例如：162.105.75.1 255.255.255.0

使用子网掩码的转发算法

(1) 提取接收分组的首部目的 IP 地址 *D*

(2) 用各网络的子网掩码与 *D* 相“与”，看是否与相应的网络地址匹配；若匹配，则直接交付；否则，间接交付，执行(3)

(3) 若转发表中有目的地址为 *D* 的特定主机路由，则将分组传送给指明的下一跳；否则，执行(4)

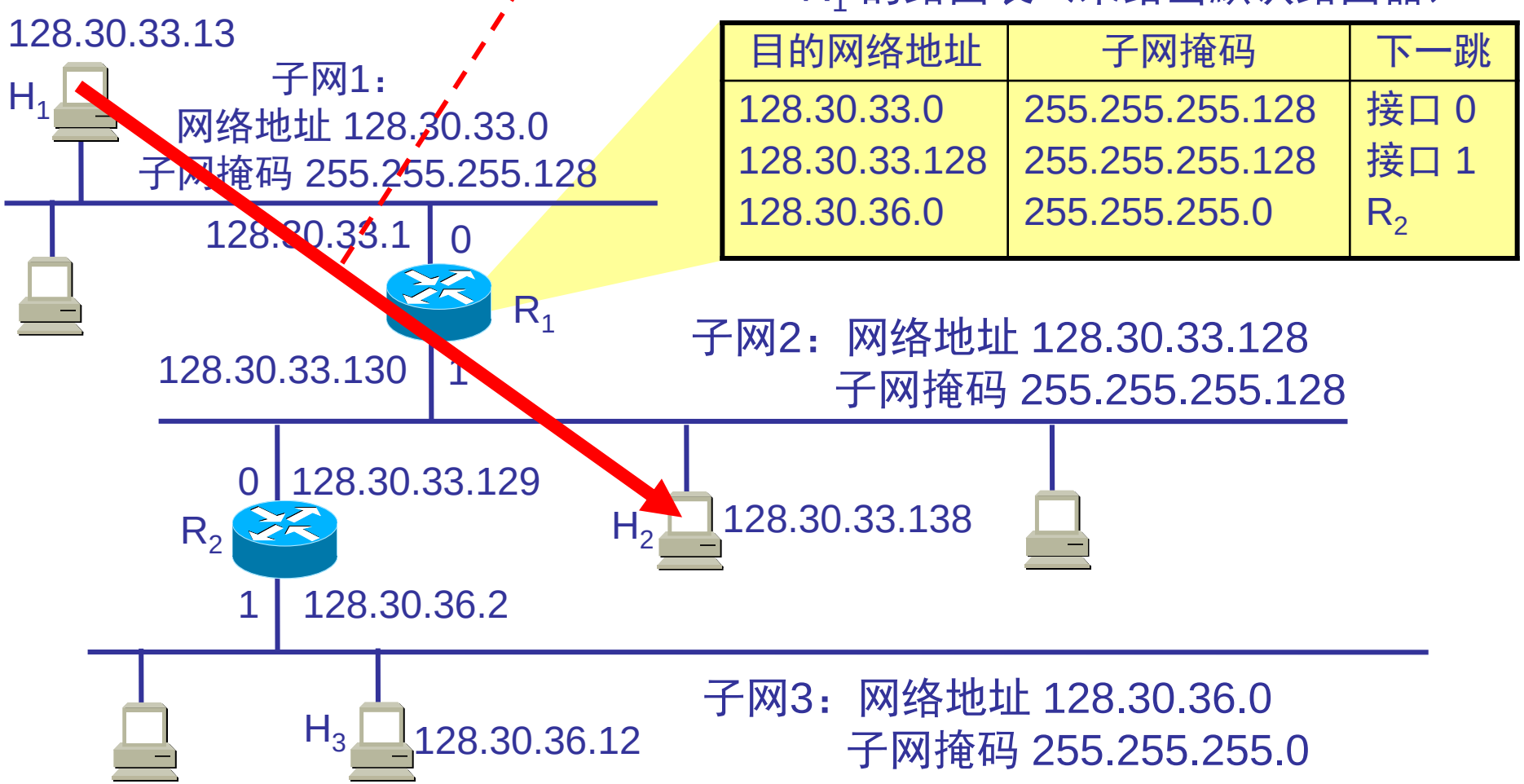
(4) 对转发表中的每一行的子网掩码和*D*逐位相“与”，若结果与该行的目的网络地址匹配，则将分组传送给下一跳；否则，执行(5)

(5) 若转发表中有一个默认路由，则将分组传送给默认路由器；否则，执行(6)

(6) 用ICMP报告分组转发出错

划分子网场景下的分组转发过程

发送分组的目的 IP 地址：128.30.33.138



R_1 的路由表（未给出默认路由器）

目的网络地址	子网掩码	下一跳
128.30.33.0	255.255.255.128	接口 0
128.30.33.128	255.255.255.128	接口 1
128.30.36.0	255.255.255.0	R_2

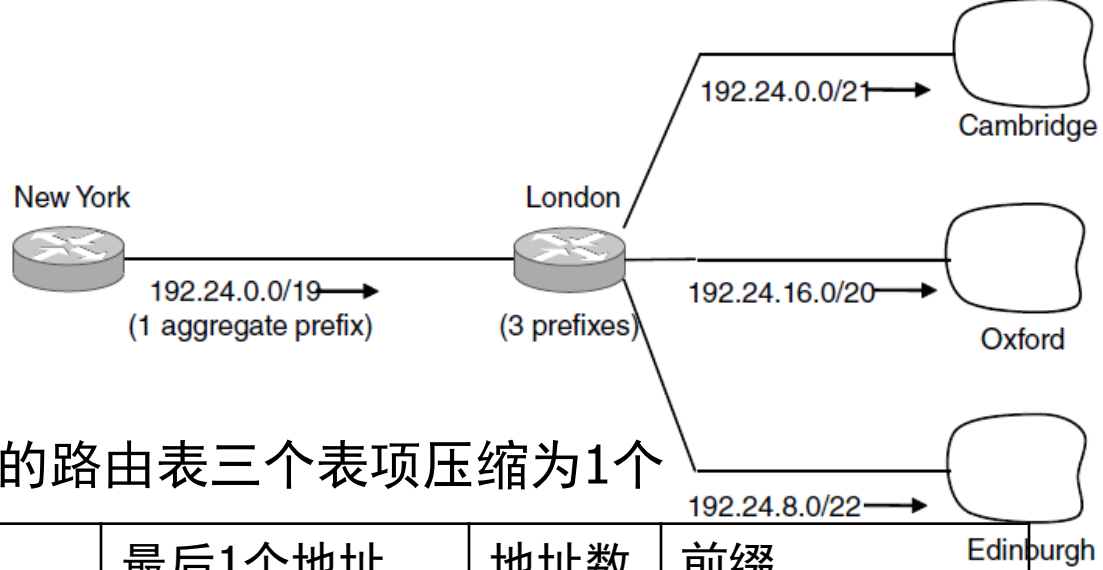
无分类编址CIDR

- CIDR(Classless Inter-Domain Routing)：不采用分类地址及划分子网的概念
- CIDR用网络前缀代替地址中的网络号和子网号

IP地址 ::= {<网络前缀>, <主机号>}

- CIDR 使用“斜线记法”，又称为CIDR记法
  - 例如162.105.75.1/16， 162.105.75.1/24
- CIDR 将网络前缀相同的连续IP地址组成“CIDR 地址块”，这样路由表中一项可表示多个分类地址
- 称这种地址聚合为路由聚合，也称为构成超网

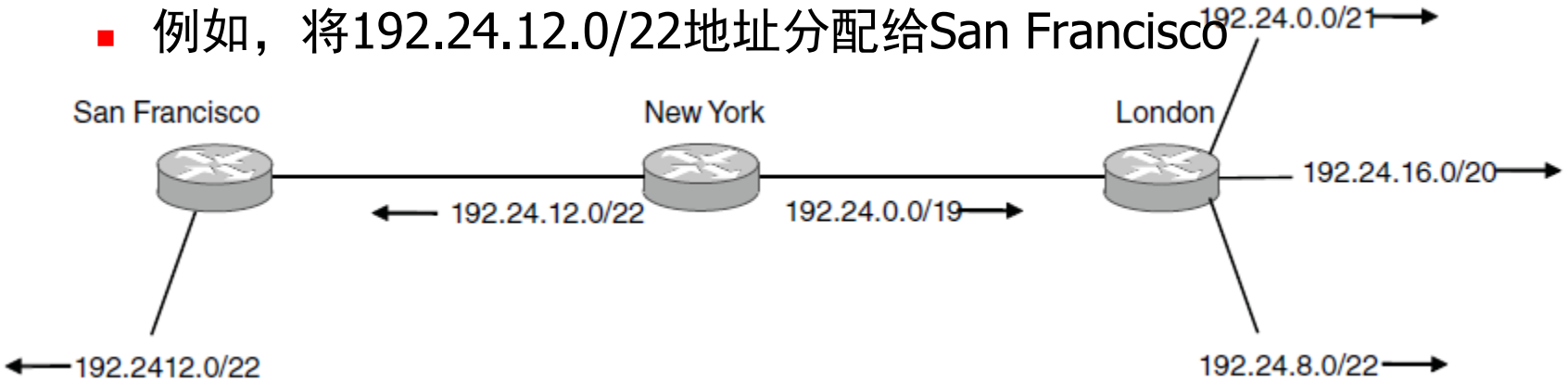
CIDR举例



- 路由聚合，减少了表项数
  - 例如，将NewYork的路由表三个表项压缩为1个

大学	第1个地址	最后1个地址	地址数	前缀
Cambridge	192.24.0.0	192.24.7.255	2048	192.24.0.0/21
Edinburgh	192.24.8.0	192.24.11.255	1024	192.24.8.0/22
保留	192.24.12.0	192.24.15.255	1024	192.24.12.0/22
OXford	192.24.16.0	192.24.31.255	4096	192.24.16.0/20

- 最长匹配前缀，利于灵活调整地址
  - 例如，将192.24.12.0/22地址分配给San Francisco



小结

- 地址分配：
  - IP地址，三种编址方式；
  - 如何分配IP地址？
  - IP地址数量不够如何解决？
- 分组传送
  - ARP：IP 地址到MAC的映射
  - 各段链路的帧长度不同，如何确定IP分组长度？
- 路由与转发：
  - RIP及距离矢量路由算法
  - 其他的路由算法及路由协议
- 网络控制：超时控制、差错恢复、状态报告、拥塞检测与控制？

练习题

一个路由器的转发表如下：

地址/掩码	下一跳
135.46.56.0/22	接口0
135.46.60.0/22	接口1
192.53.40.0/23	路由器1
default	路由器2

若到达的分组，其目的地址有下述IP地址，问路由器如何处理

（1）135.46.63.10（2）135.46.57.14

（3）135.46.52.2（4）192.53.40.7（5）192.53.56.7

- 解：56=0x38,60=0x3C, 63=0x3F 57=0x39 52=0x34

  （1）接口1　　（2）接口0　（3） 路由器2

  （4）路由器1　　（5） 路由器2

思考题：

- 路由器与交换机有何不同？

	路由器	交换机
基于目的地址	网络地址	MAC地址
转发表的记录数	为端口数的数量级	为主机数的数量级
广播频次	与路由器数量有关	与主机数有关
路由更新	定时或拓扑变化	与交换表的定时器有关

- 主机与路由器在处理IP数据报时，其行为有何不同？

作业2：IP收发实验（主机行为）

- IP接收
  - 检查IPV4头部字段：版本号（4），头部长度(>4)，生存时间（>0）及头部校验和，丢弃错误的分组并说明错误类型
  - 检查目的地址，为本机或广播地址，则接收，否则丢弃
    - 地址在使用时需要进行大小端（网络序—主机序）转换
  - 提取协议类型
- IP发送
  - 提取数据长度，分配存储空间
  - 产生头部信息，转换为网络字节序
  - 封装分组并发送

要求：4月26日前提交报告及源程序

提示：网络字节序

- **小端模式**：数据的低字节保存在内存的低地址中
- **大端模式**：数据的高字节保存在内存的低地址中
- **主机的字节序与CPU有关，可以是小端模式或大端模式**
- 网络字节序是大端模式
- 因此需要进行大端到小端(函数ntohl、ntohs)或小端到大端(函数htonl、htons)；
  - 收到IP分组转换为主机序后再处理，如IP地址、长度
  - 产生IP分组，填写IP地址、长度、校验和等，转换为网络序后发送
  - 转换方式分为2字节和4字节

作业3：IP转发实验（路由器行为）

- 路由器的主要任务是分组转发， 接收的多数分组需要转发，而不像主机协议栈中IPv4模块只接收发送给本机的分组； 另外，也要接收发给本机的一些分组，如路由协议分组、ICMP分组等
- 路由信息包括地址段、距离、下一跳地址、操作类型等。 在接收IPv4分组后，通过其目的地址匹配地址段来判断是否为本机地址，如果是则接收；如果不是，则通过其目的地址段查找路由表信息，得到进一步的操作类型，转发情况下要获得下一跳IPv4地址。发送IPv4分组时，要用目的地址来查找路由表，得到下一跳IPv4地址，然后调用发送接口函数做进一步处理。在转发路径中， 本路由器可能是路径上的最后一跳， 可直接转发给目的主机； 而非最后一跳情况下，下一跳地址是从对应的路由信息中获取。因此，在路由表中转发类型要区分最后一跳和非最后一跳的情况。

作业3：IP转发实验（路由器行为）

- 1) 向上层协议上交目的地址为本机地址的分组；
- 2) 根据路由查找结果， 丢弃查不到路由的分组；
- 3) 根据路由查找结果，向相应接口转发不是本机接收的分组
- 实验内容：
  - 设计路由表数据结构
  - IPv4分组的转发：
    - 如果目的地址为本机地址， 本地接收
    - 如果TTL为0， 则丢弃该分组
    - 查表，实施最长匹配
    - 如果路由表中存在该目的地址， 转发（TTL减一，重新计算校验和）

要求：5月17日前提交报告及源程序