

# 计算机网络（第 5 版）

## 第 4 章 网络层



# 第 4 章 网络层

---

4.1 网络层提供的两种服务

4.2 网际协议 IP

4.2.1 虚拟互连网络

4.2.2 分类的 IP 地址

4.2.3 IP 地址与硬件地址

4.2.4 地址解析协议 ARP 与逆地址解析协议  
RARP

4.2.5 IP 数据报的格式

4.2.6 IP 层转发分组的流程



## 第 4 章 网络层（续）

---

### 4.3 划分子网和构造超网

#### 4.3.1 划分子网

#### 4.3.2 使用子网时分组转发

#### 4.3.3 无分类编址 CIDR（构造超网）

### 4.4 网际控制报文协议 ICMP

#### 4.4.1 ICMP 报文的种类

#### 4.4.2 ICMP 的应用举例

### 4.5 因特网的路由选择协议

#### 4.5.1 有关路由选择协议的几个基本概念

#### 4.5.2 内部网关协议 RIP



# 第 4 章 网络层（续）

---

4.5.3 内部网关协议 OSPF

4.5.4 外部网关协议 BGP

4.5.5 路由器的构成

4.6 IP 多播

4.6.1 IP 多播的基本概念

4.6.2 在局域网上进行硬件多播

4.6.2 因特网组管理协议 IGMP 和多播路由选择协议

4.7 虚拟专用网 VPN 和网络地址转换 NAT

4.7.1 虚拟专用网 VPN

4.7.2 网络地址转换 NAT



# 本章重点

---

- (1) 虚拟互连网络的概念
- (2) IP 协议、IP 地址及与物理地址的关系
- (3) 传统的分类的 IP 地址（包括子网掩码）和无分类域间路由选择 CIDR
- (4) 路由选择协议的工作原理



# 复习 ----- 交换技术

---

- 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。
- 交换技术分类：
  - 电路交换 (Circuit switching)
  - 报文交换 (Message switching)
  - 分组交换 (Packet switching)
- 分组交换又分为两种：数据报服务和虚电路服务



## 4.1 网络层提供的两种服务

- 网络层所提供的服务有**两大类**：
- **面向连接的网络服务**
  - 先建立连接，再传输数据，最后拆除连接。
  - 可以协商控制参数、可选服务类型、服务质量等。
  - 特点：传输层简单。
- **无连接的网络服务**
  - 不需建立连接，直接传输数据。
  - 前提：通信子网是不可靠的，流量、差错等的控制由主机自己负责进行。
  - 特点：传输层设计复杂。

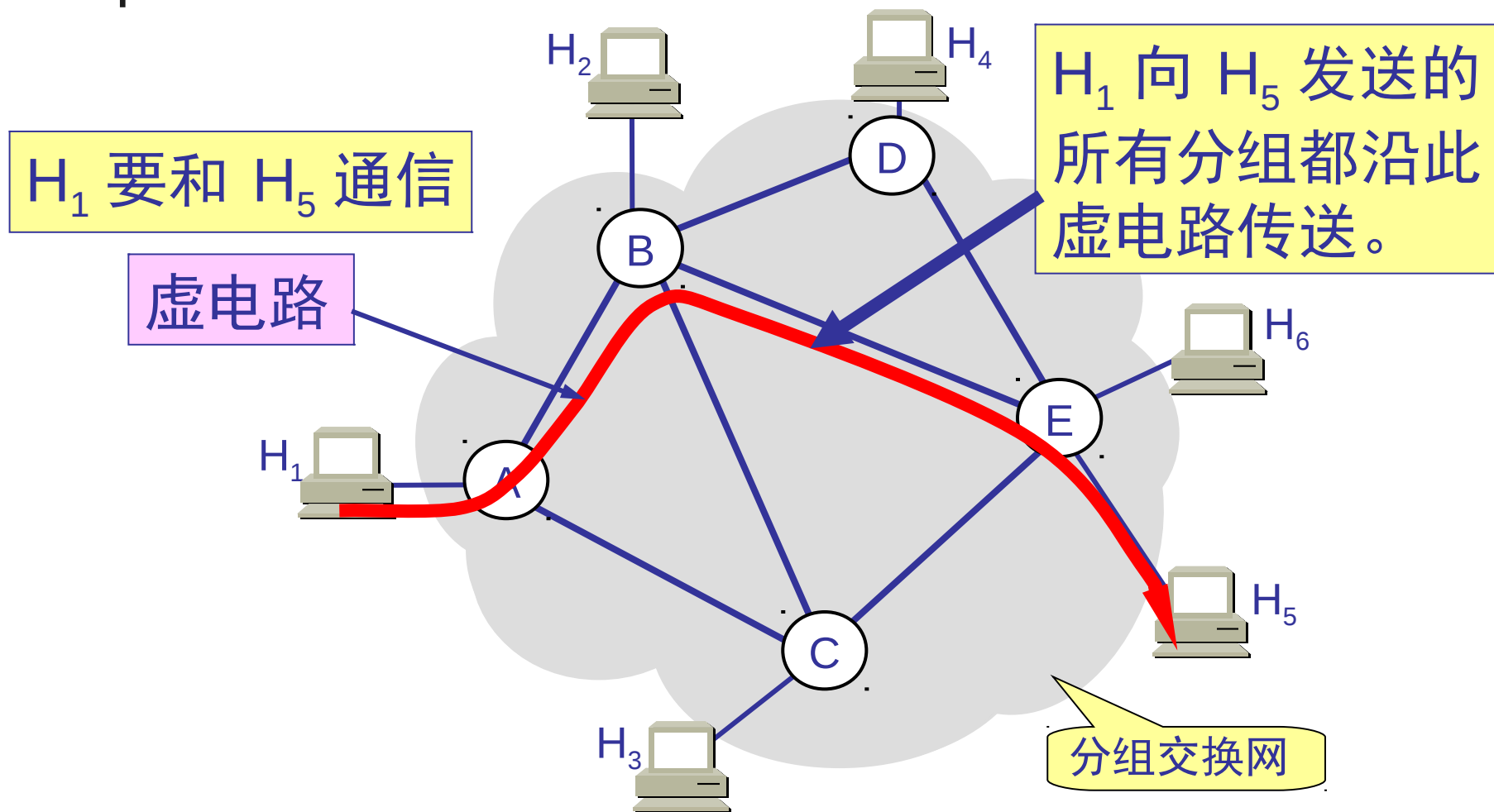


# 虚电路分组交换的原理

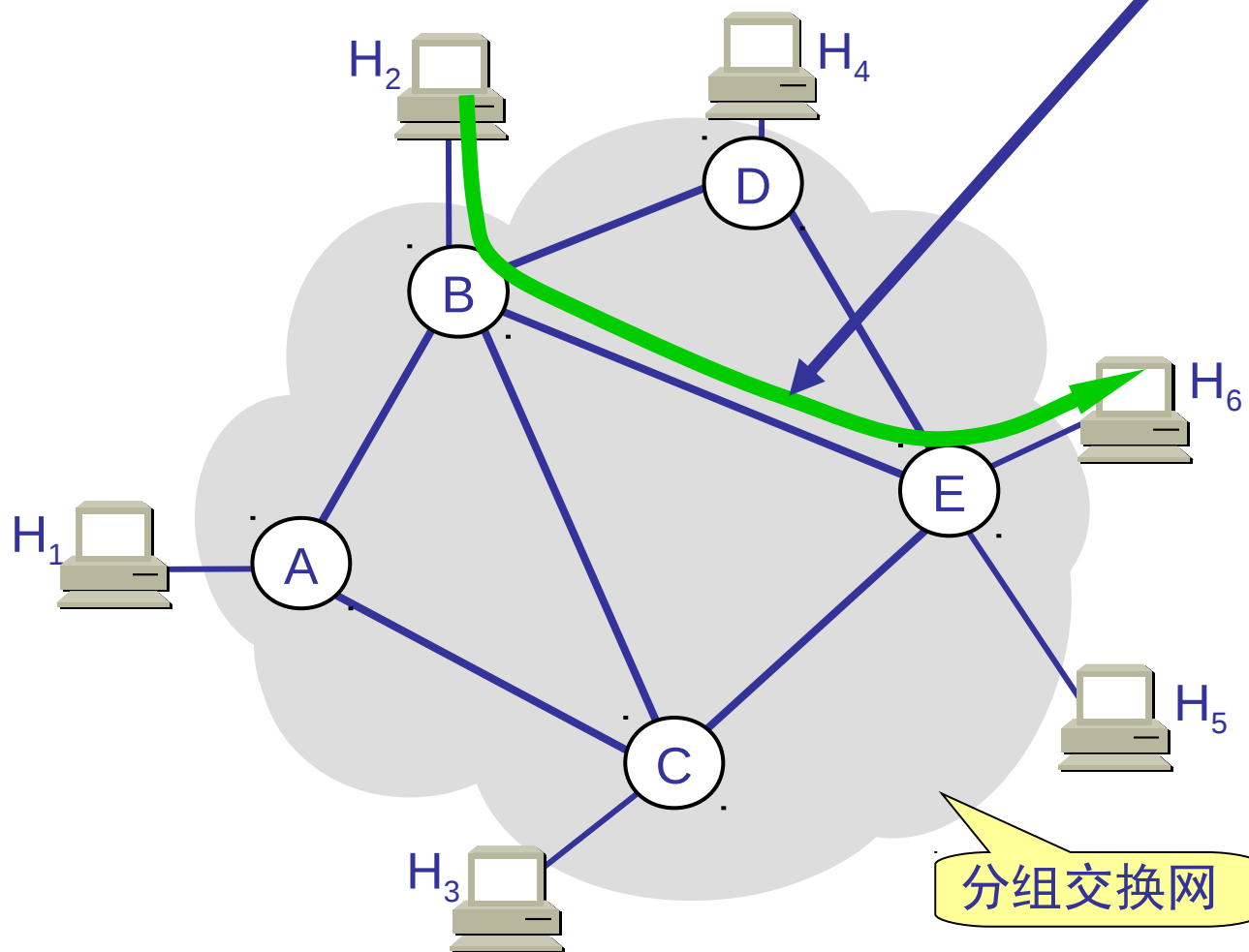
- 虚电路服务：面向连接的网络服务
  - 在传送数据之前，首先通过**虚呼叫**建立一条虚电路（源端到目的端的各分段的逻辑信道组成，虚电路只是一条**逻辑上通路**）。
  - 每个分组除了包含数据之外需包含**虚电路标识符**。所有分组沿同一条路径按顺序传送到达目的主机。
  - 通信结束后，由某一个站用**清除请求分组**来结束这次连接
  - 虚电路服务对通信的服务质量 QoS (Quality of Service) 有较好的保证。



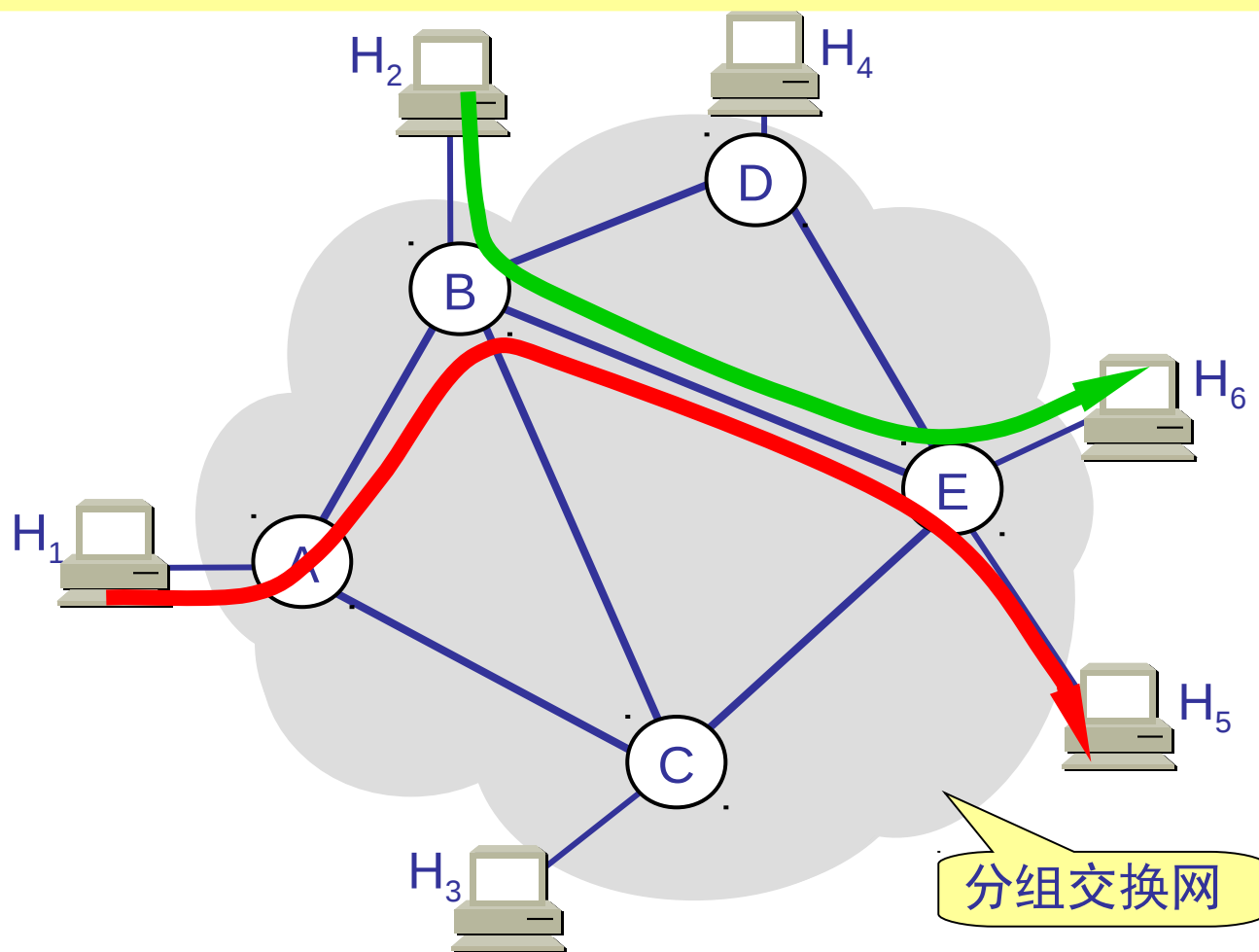
主机  $H_1$  先向主机  $H_5$  发出一个特定格式的控制信息分组，要求进行通信，同时寻找一条合适路由。若主机  $H_5$  同意通信就发回响应，然后双方就建立了虚电路。



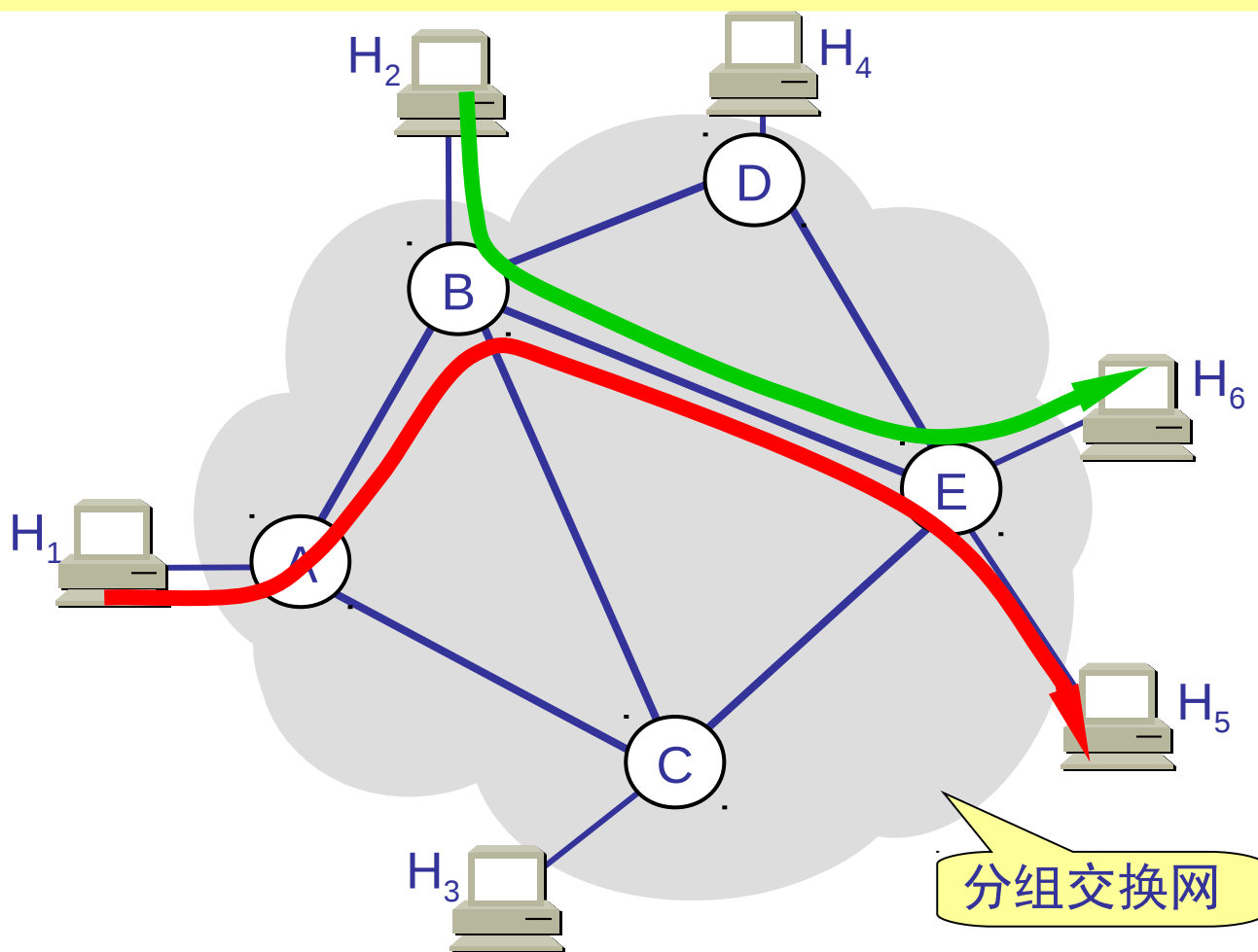
同理，主机  $H_2$  和主机  $H_6$  通信之前，也要建立虚电路。



在虚电路建立后，网络向用户提供的服务就好像在两个主机之间建立了一对穿过网络的**数字管道**。所有发送的分组都按顺序进入管道，然后按照先进先出的原则沿着此管道传送到目的站主机。



到达目的站的分组顺序就与发送时的顺序一致，  
因此网络提供虚电路服务对通信的  
**服务质量** QoS (Quality of Service) 有较好的保证。

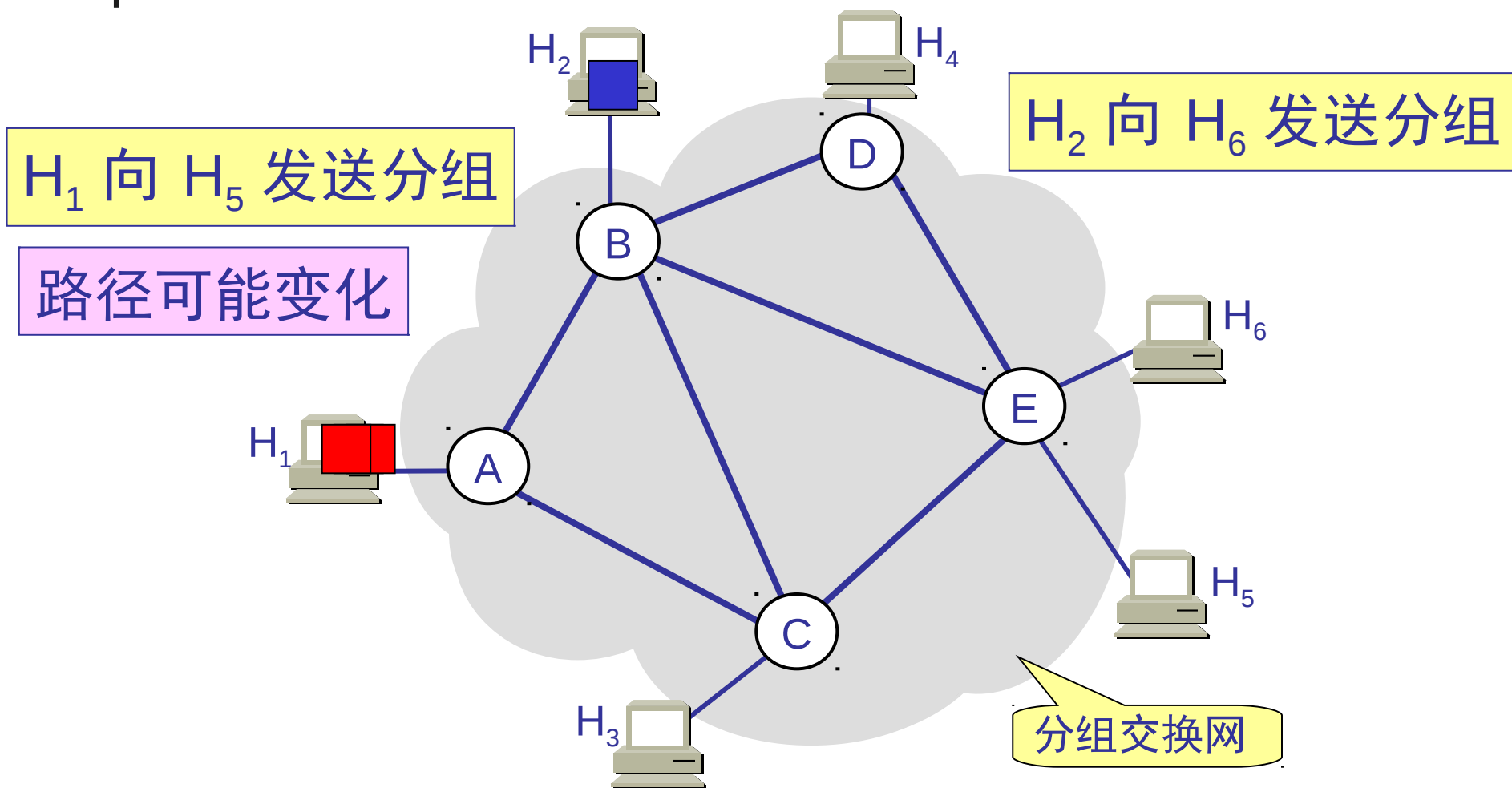




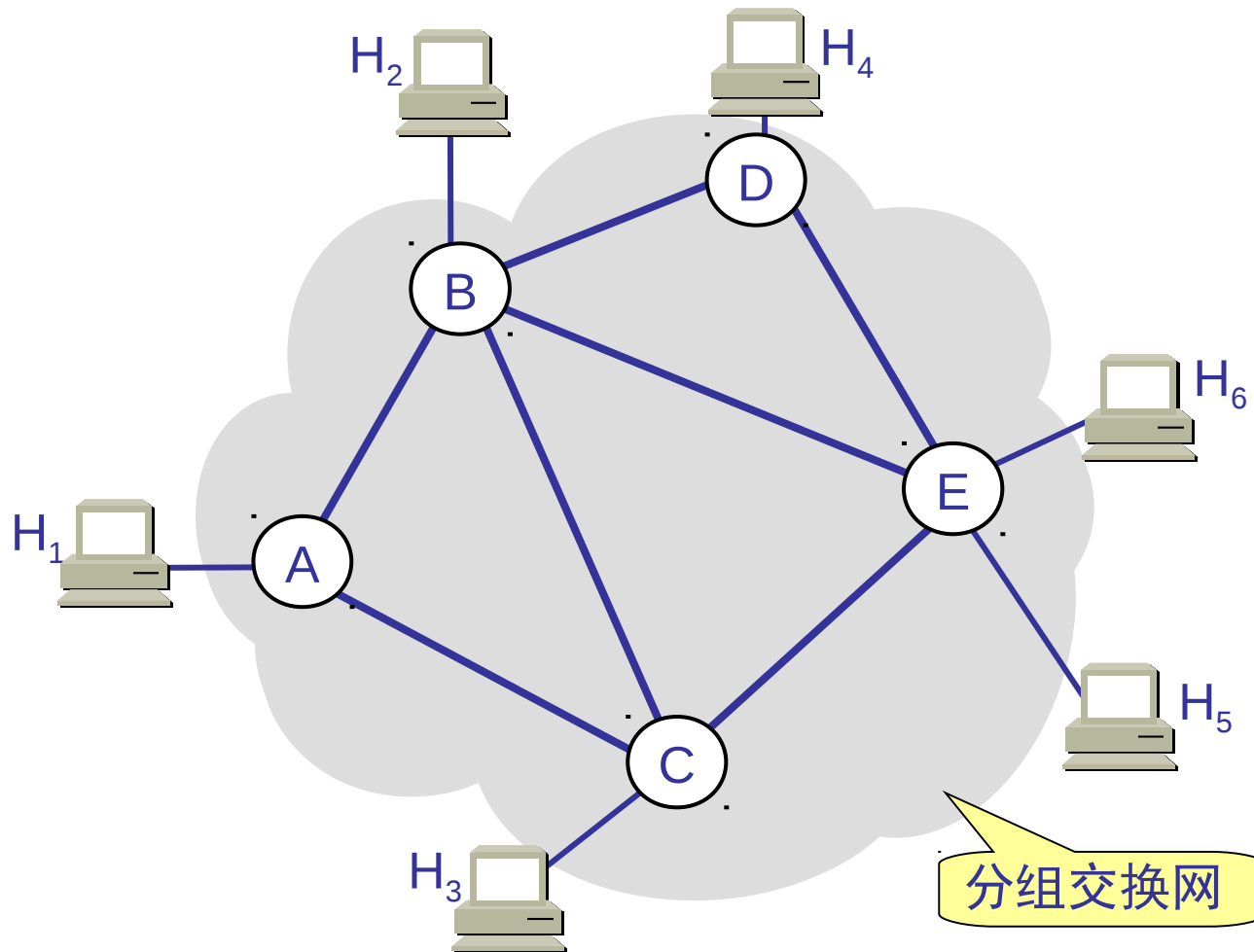
# 数据报分组交换的原理

- 数据报服务：无连接的网络服务
  - 每个分组**单独传送**。
  - 分组需自身携带**完整的地址信息**。
  - 网络为每个分组单独选路，路径可能不同。
  - 不能保证各个分组**按顺序**到达目的地，有的分组甚至会中途丢失。
  - 网络只是**尽最大努力**地将分组交付给目的主机，它不能保证服务质量。

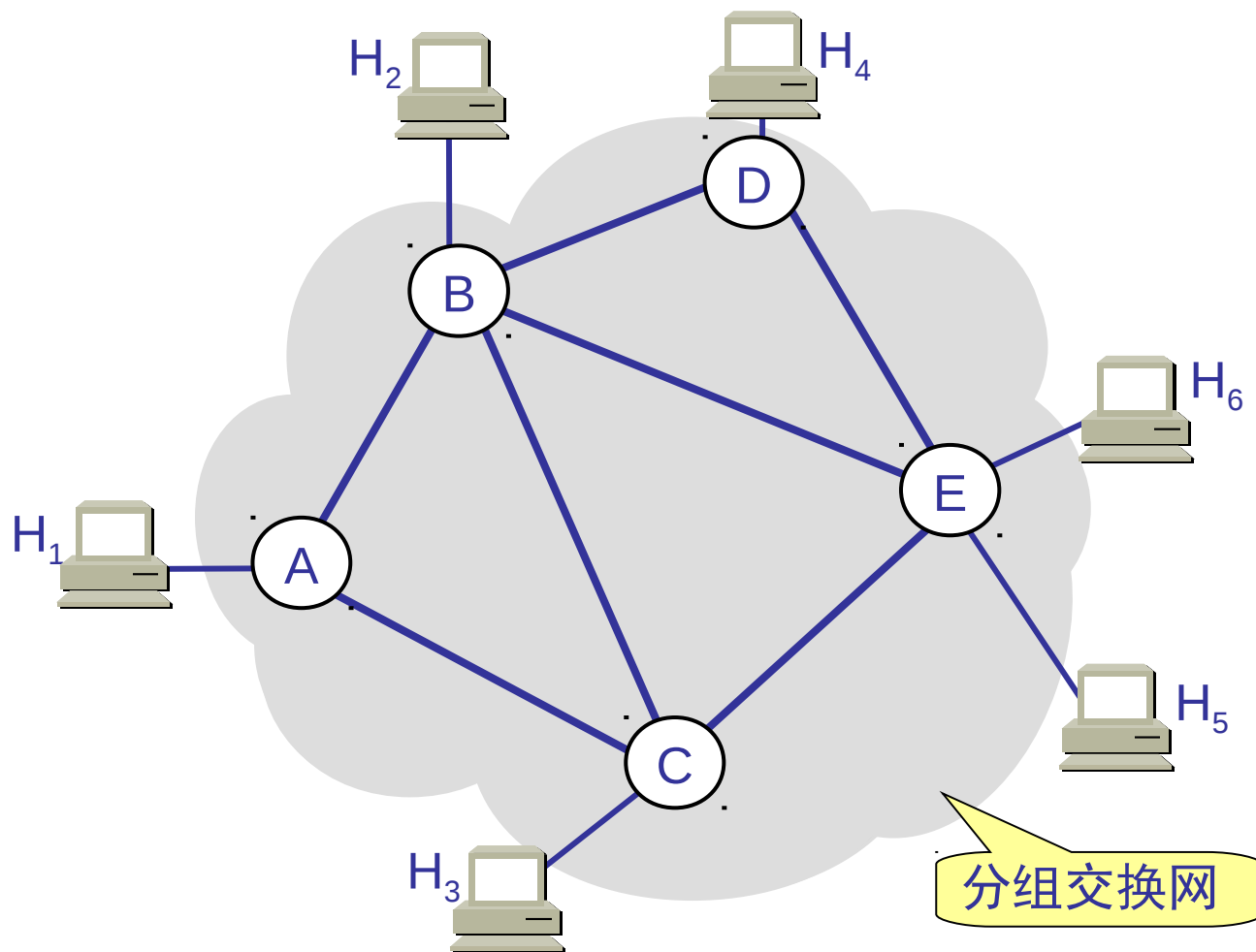
网络随时接受主机发送的分组（即数据报）  
网络为每个分组独立地选择路由。



网络尽最大努力地将分组交付给目的主机，  
但网络对源主机没有任何承诺。

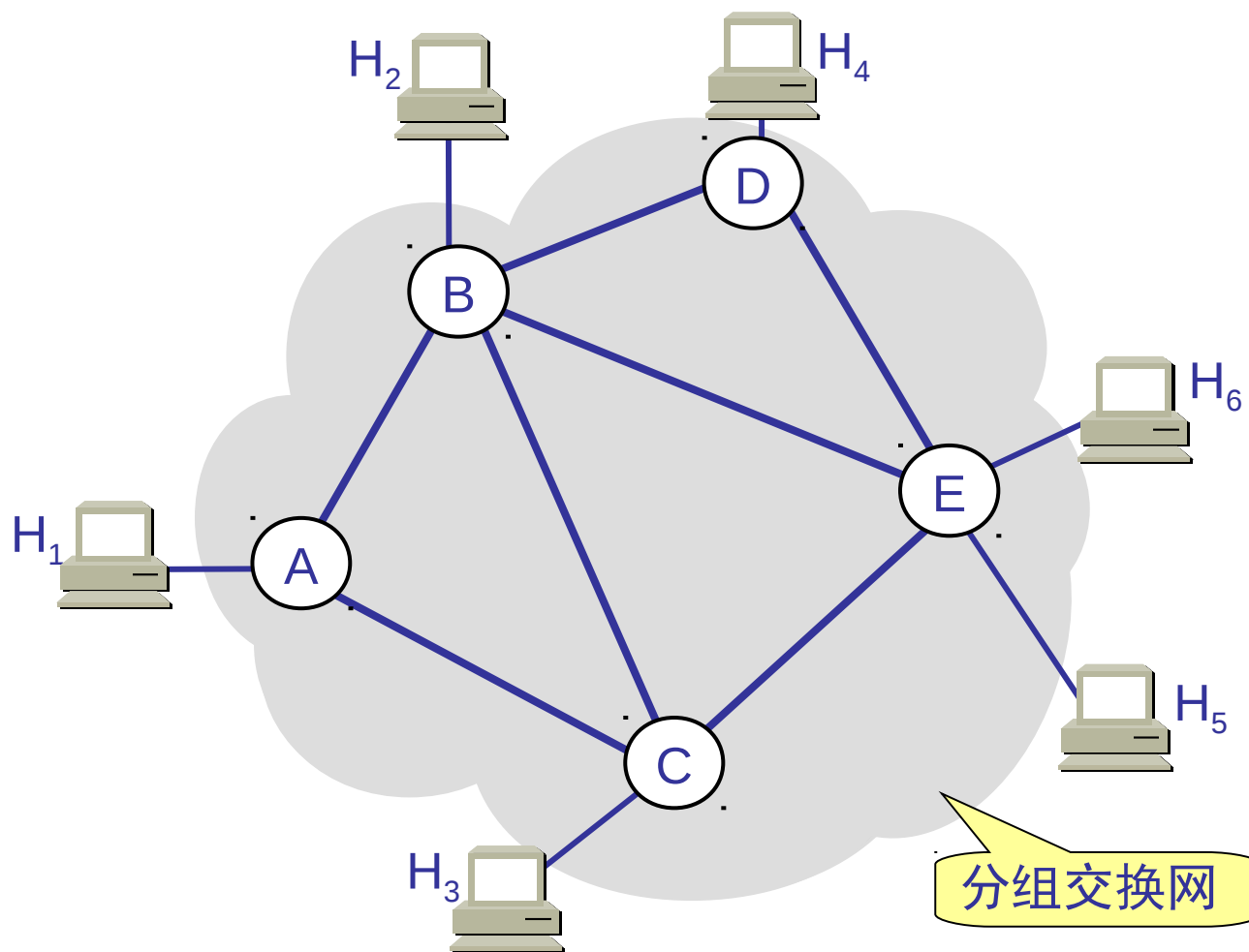


网络不保证所传送的分组不丢失  
也不保证按源主机发送分组的先后顺序  
以及在时限内必须将分组交付给目的主机

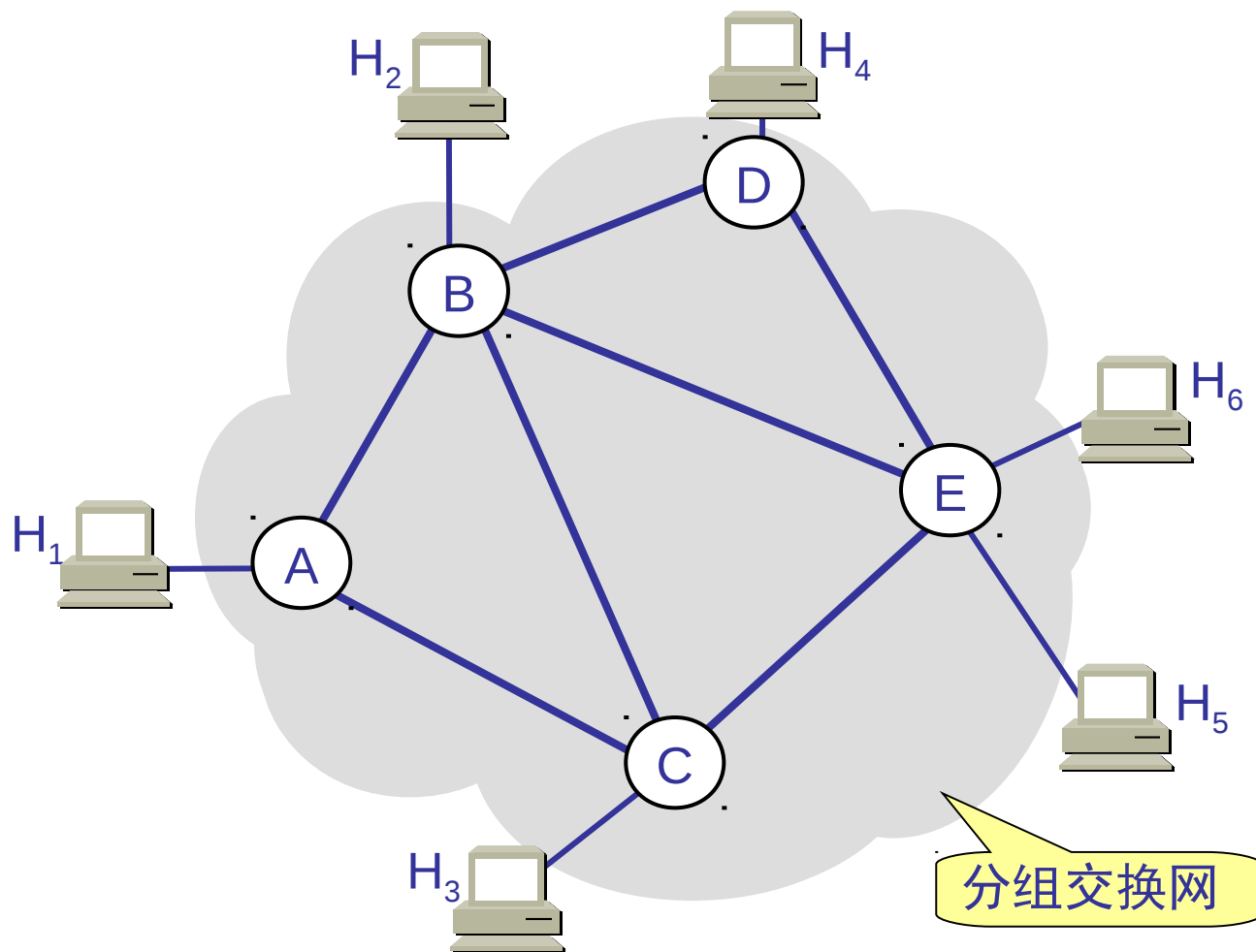




数据报提供的服务是不可靠的，  
它不能保证服务质量。  
实际上“尽最大努力交付”的服务  
就是没有质量保证的服务。



# 当网络发生拥塞时 网络中的结点可根据情况将一些分组丢弃





# 两种服务的思路来源不同

---

- **虚电路服务**的思路来源于传统的电信网。
  - 电信网负责保证可靠通信的一切措施，因此电信网的结点交换机复杂而昂贵。
- **数据报服务**力求使网络生存性好和使对网络的控制功能分散，因而只能要求网络提供尽最大努力的服务。
  - 可靠通信由用户终端中的软件（即 TCP ）来保证。



# 数据报服务和虚电路服务 优缺点的归纳

对比的方面	虚电路服务	数据报服务
思路	可靠通信应当由网络来保证	可靠通信由用户主
连接的建立	必须有	不
目的站地址	仅在连接建立阶段使用，每个分组使用短的虚电路号	每个分组都有目的站的



# 数据报服务和虚电路服务 优缺点的归纳

对比的方面	虚电路服务	数据报服务
分组的转发	属于同一条虚电路的分组均按照同一路由进行转发	每个分组独立选择路由
当结点出故障时	所有通过出故障的结点的虚电路均不能工作	故障结点可能丢失分组，一般可能



# 数据报服务和虚电路服务 优缺点的归纳

对比的方面	虚电路服务	数据报服务
分组的顺序	总是按发送顺序 到达目的站	到达目的站时不一 到达目的站
端到端的 差错处理和 流量控制	可以由分组交换网 负责也可以由用户 主机负责	由用户主机负责



# 网络层两种服务之争

---

- 从 20 世纪 70 年代起，关于网络层究竟应当采用数据报服务还是虚电路服务（“面向连接”还是“无连接”），曾引起了长期的争论。
- 争论焦点的实质就是：在计算机通信中，可靠交付应当由谁来负责？是网络还是端系统？



# 数据报服务与虚电路服务之争

---

- 网络只提供数据报服务就可大大简化网络层的结构。
- 且技术的进步使得网络出错的概率已越来越小，因而让主机负责端到端的可靠性不但不会给主机增加更多的负担，反而能够使更多的应用在这种简单的网络上运行。





# 数据报服务和虚电路服务 都各有一些优缺点

---

- 网络上传送的报文长度，在很多情况下都很短，用**数据报**既迅速又经济。
- 若用**虚电路**，为了传送一个分组而建立虚电路和释放虚电路就显得太浪费网络资源了。



# 数据报服务和虚电路服务 都各有一些优缺点

---

- 数据报服务对军事通信有其特殊的意义。当某个结点发生故障时，后续的分组就可另选路由，因而提高了可靠性。
- 但在使用虚电路时，结点发生故障就必须重新建立另一条虚电路。
- 数据报服务还很适合于将一个分组发送到多个地址（即广播或多播）。

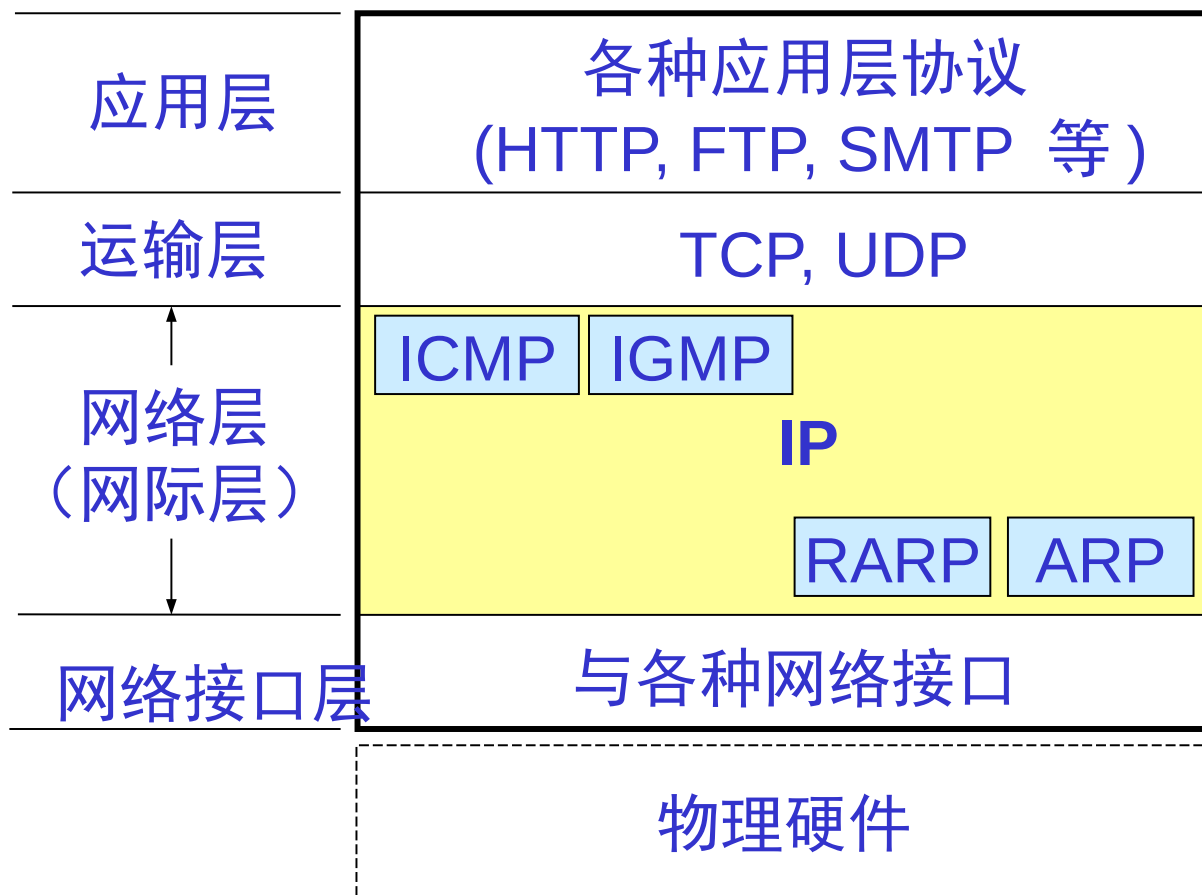


## 4.2 网际协议 IP 及配套协议

---

- 网际协议 IP 是 TCP/IP 体系中两个最主要的协议之一。与 IP 协议配套使用的还有四个协议：
- 地址解析协议 ARP  
(Address Resolution Protocol)
- 逆地址解析协议 RARP  
(Reverse 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)

# 网际层的 IP 协议及配套协议

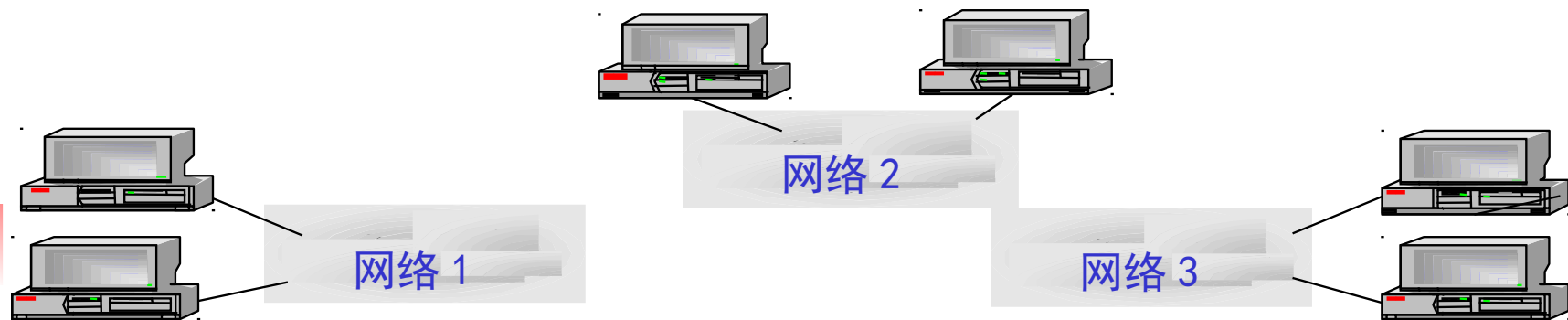




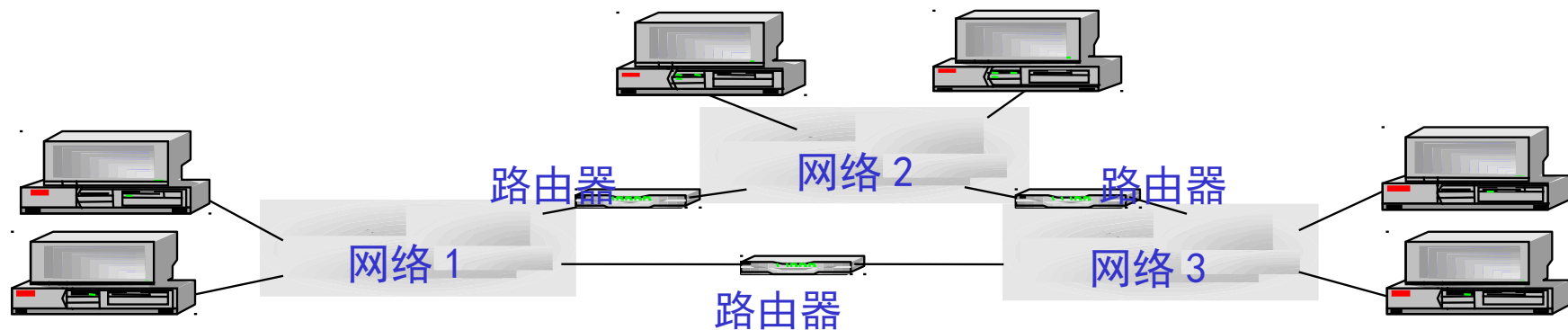
## 4.2.1 虚拟互连网络

---

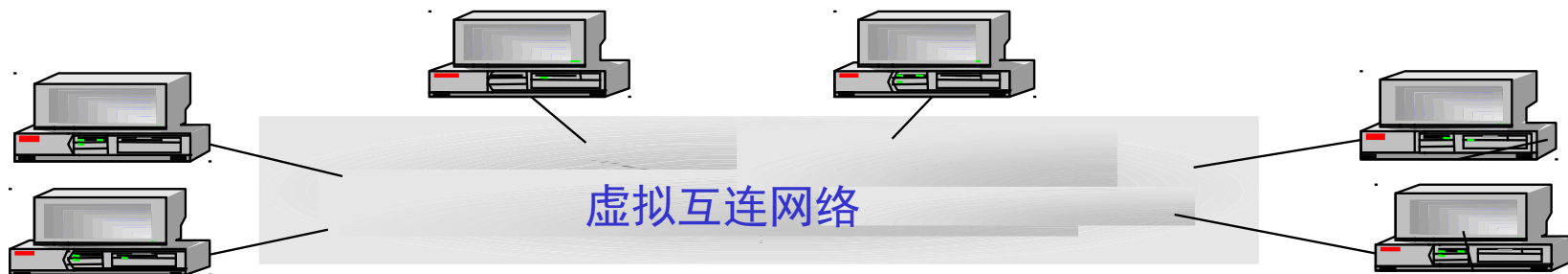
- 互连在一起的网络要进行通信，会遇到许多问题需要解决。
- 网络互相连接起来要使用一些中间设备。中间设备有网桥、交换机、路由器等，**交换机或网桥**只是扩大网络，并不能称为网络互连，**其仍是一个网络**。
- 互联网都是指用**路由器**进行互连的网络。由于历史的原因，许多有关 TCP/IP 的文献将网络层使用的路由器称为**网关**。



(a) 没有连接起来的物理网络形成了信息孤岛



(b) 用路由器将物理网络连接起来构成互连网络



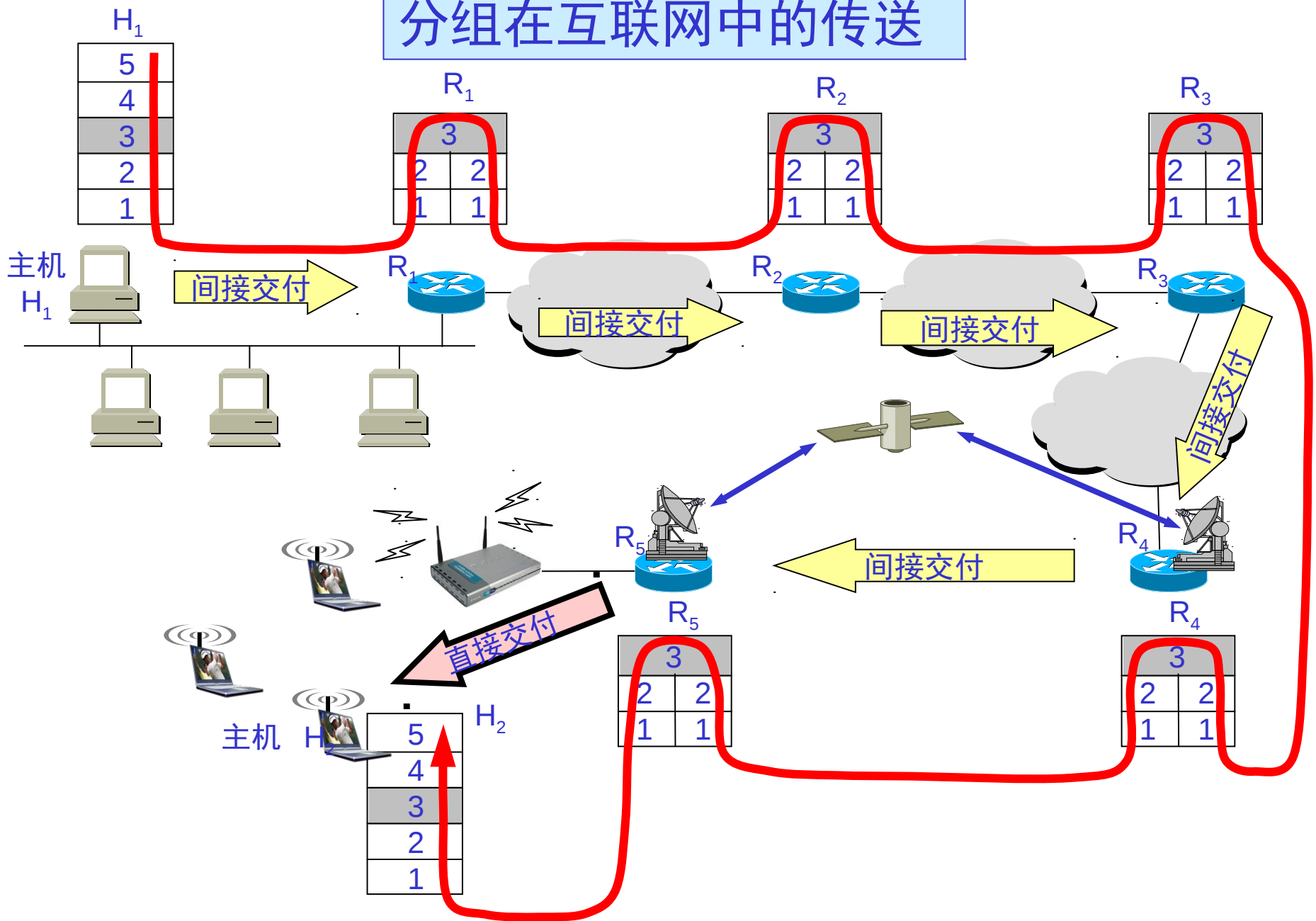
(c) 对互连网络结构抽象后的虚拟互连网络



# 虚拟互连网络的意义

- 所谓**虚拟互连网络**也就是逻辑互连网络，它的意思就是互连起来的各种物理网络的异构性本来是客观存在的，但是我们利用 IP 协议就可以使这些性能各异的网络从用户看起来好像是一个统一的网络。
- 使用 IP 协议的虚拟互连网络可简称为 **IP 网**。
- 使用**虚拟互连网络的好处**是：当互连网上的主机进行通信时，就好像在一个网络上通信一样，而看不见互连的各具体的网络异构细节。

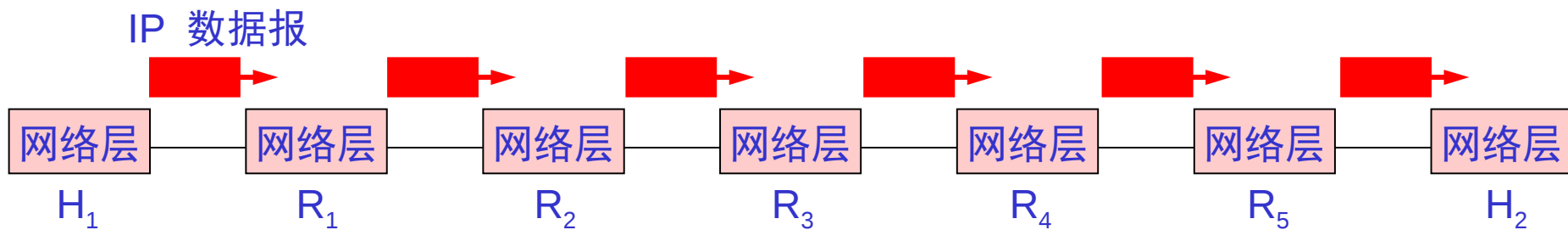
# 分组在互联网中的传送





# 从网络层看 IP 数据报的传送

- 如果我们只从网络层考虑问题，那么 IP 数据报就可以想象是在网络层中传送。





## 4.2.2 IP 地址

---

- ◆ 在互连网络系统中，为了给每一台主机或路由器的接口进行唯一的标识，常采用地址来识别；
- ◆ TCP/IP 协议的网络层使用的地址标识为 IP 地址；
- ◆ IP v.4 中 IP 地址是一个 32 位的二进制地址；
- ◆ 网络中每一个接口都需有唯一的 IP 地址；

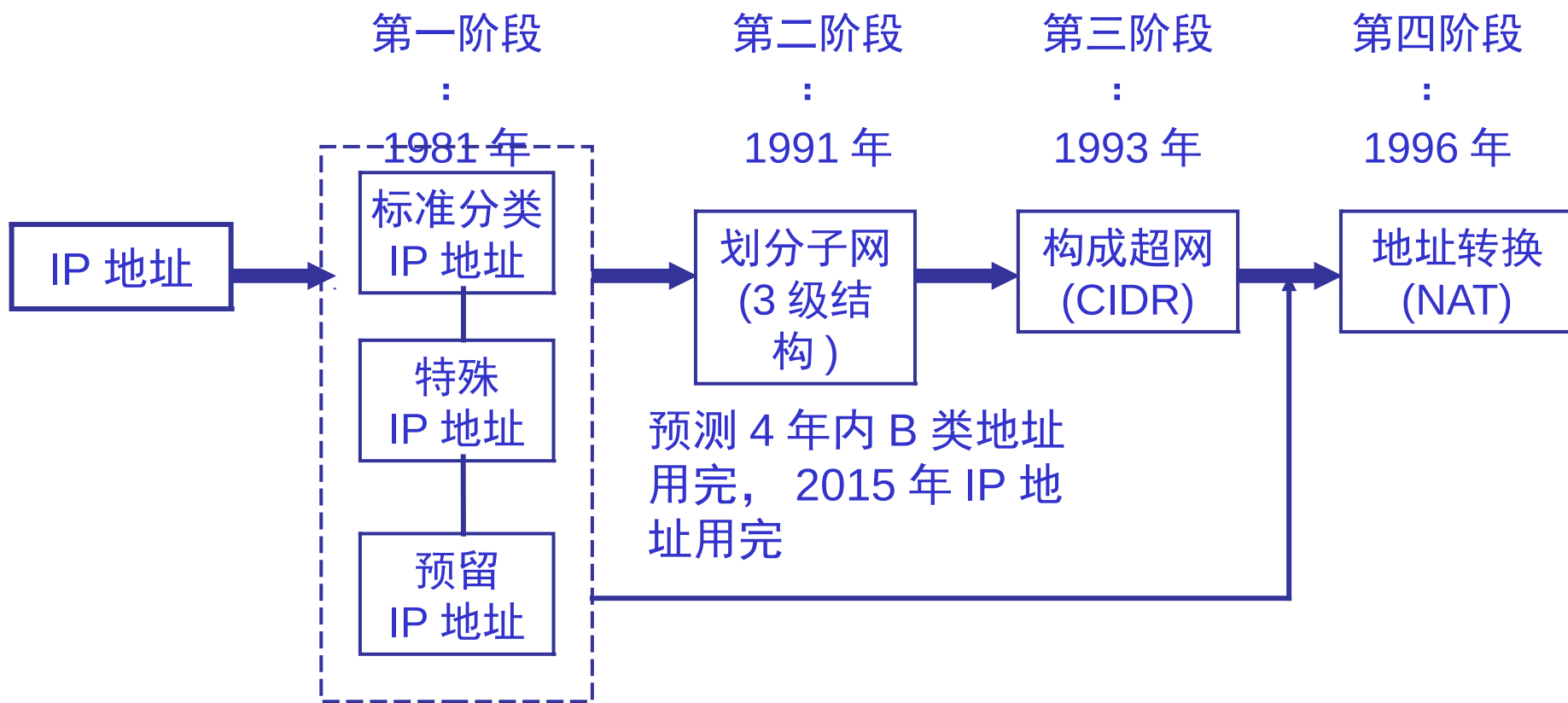


# 1. IP 地址及其表示方法

---

- IP 地址是因特网为每个连接在因特网上的主机（或路由器）分配一个在全世界范围是唯一的 32 位的标识符。
- IP 地址现在由因特网名字与号码指派公司 ICANN (Internet Corporation for Assigned Names and Numbers) 进行分配

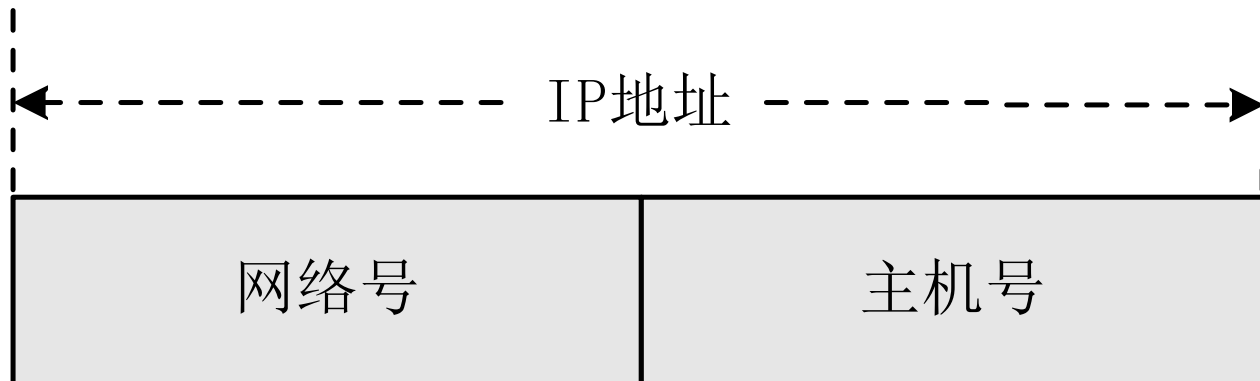
# IP 地址处理方法演变的过程





# IP 地址结构

- ◆ IP 地址采用**分层结构**，由网络号（net ID）与主机号（host ID）组成。
- ◆ IP 地址是一个 32 比特的二进制数据，每个字节用 1 个十进制数表示，格式为 W.X.Y.Z( 如 :210.32.81.10) 。



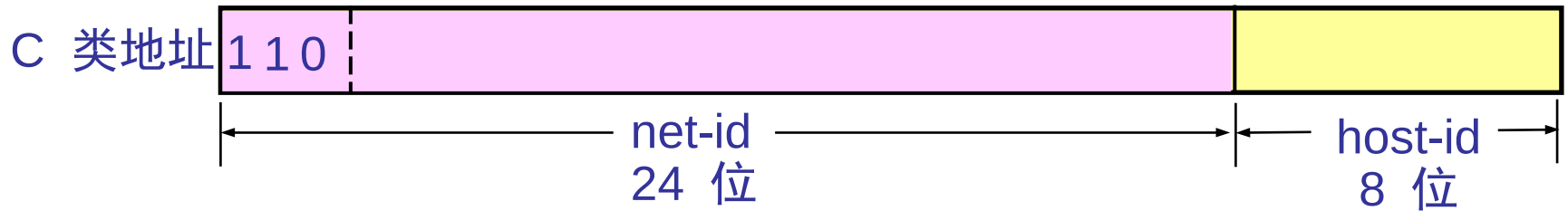
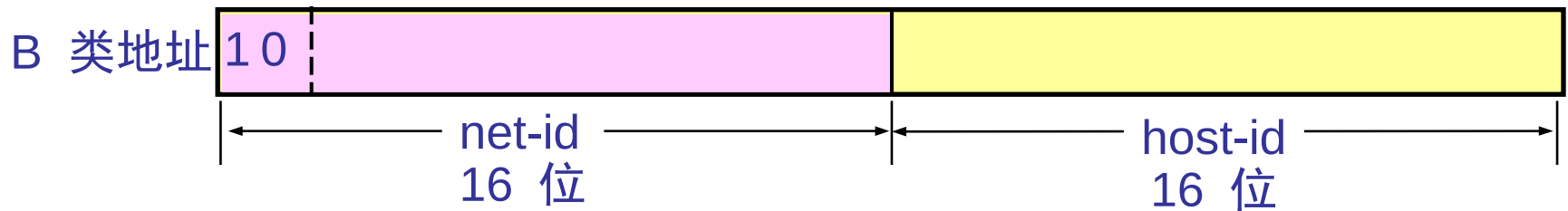
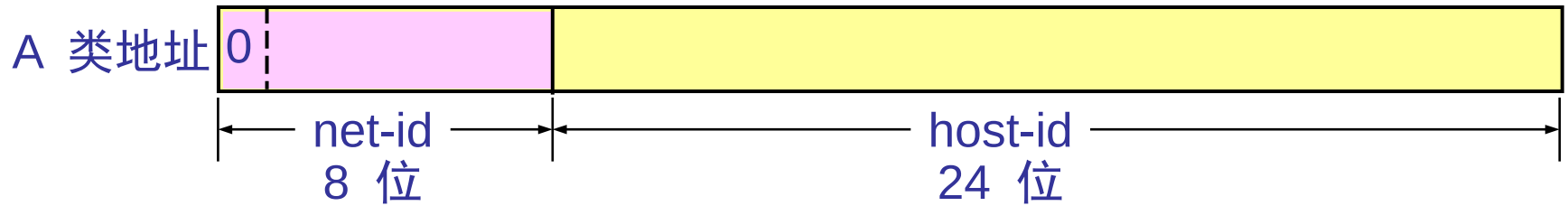


# IP 地址的分类

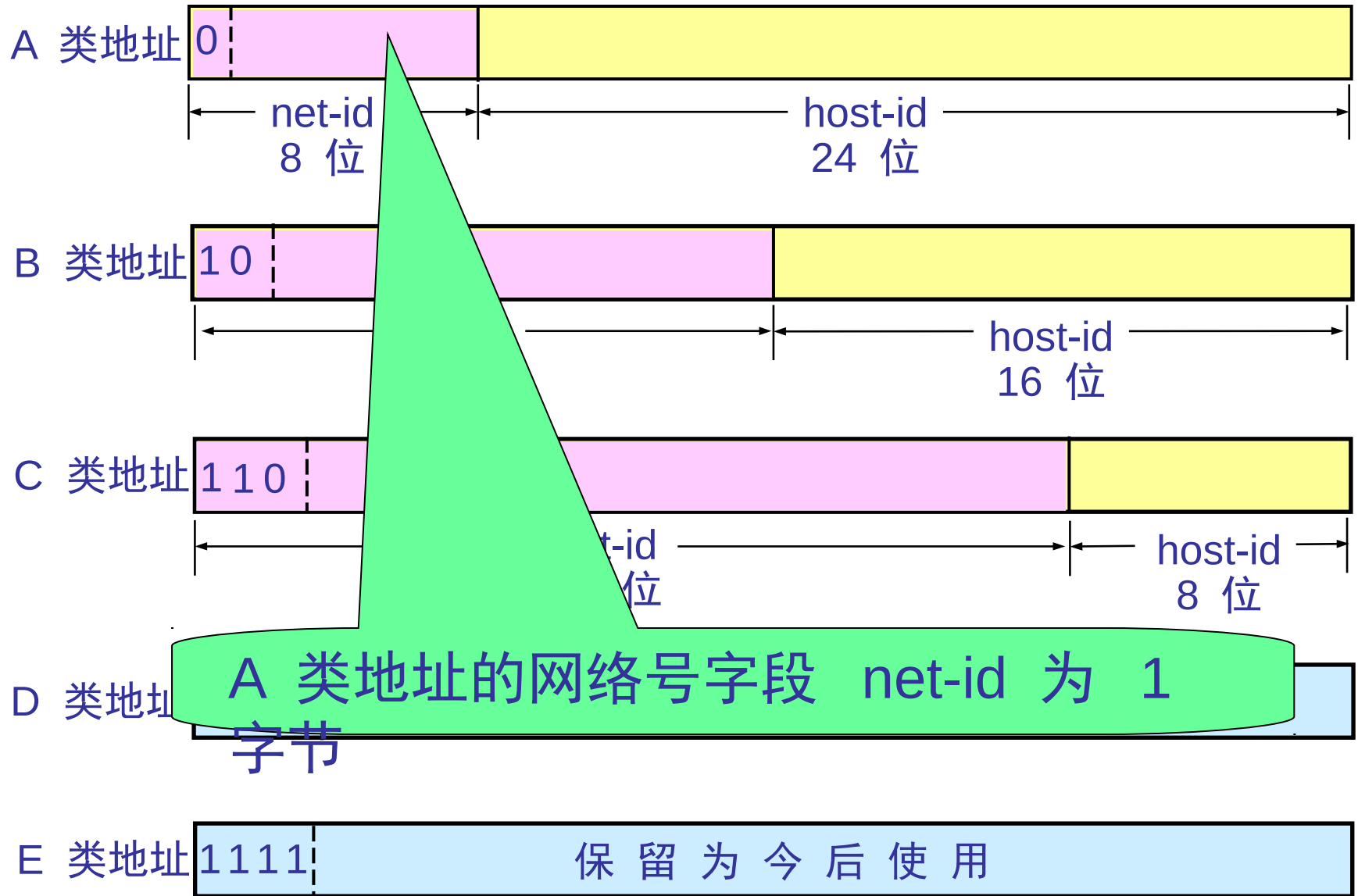
---

- ◆ 根据不同的取值范围， IP 地址可以分为五类；
- ◆ IP 地址中的第一个十进制的前几位来区分 IP 地址的类别
  - A 类地址的第一位为 0 ；
  - B 类地址的前两位为 10 ；
  - C 类地址的前三位为 110 ；
  - D 类地址的前四位为 1110 ；
  - E 类地址的前五位为 1111 。

# IP 地址中的网络号字段和主机号字段

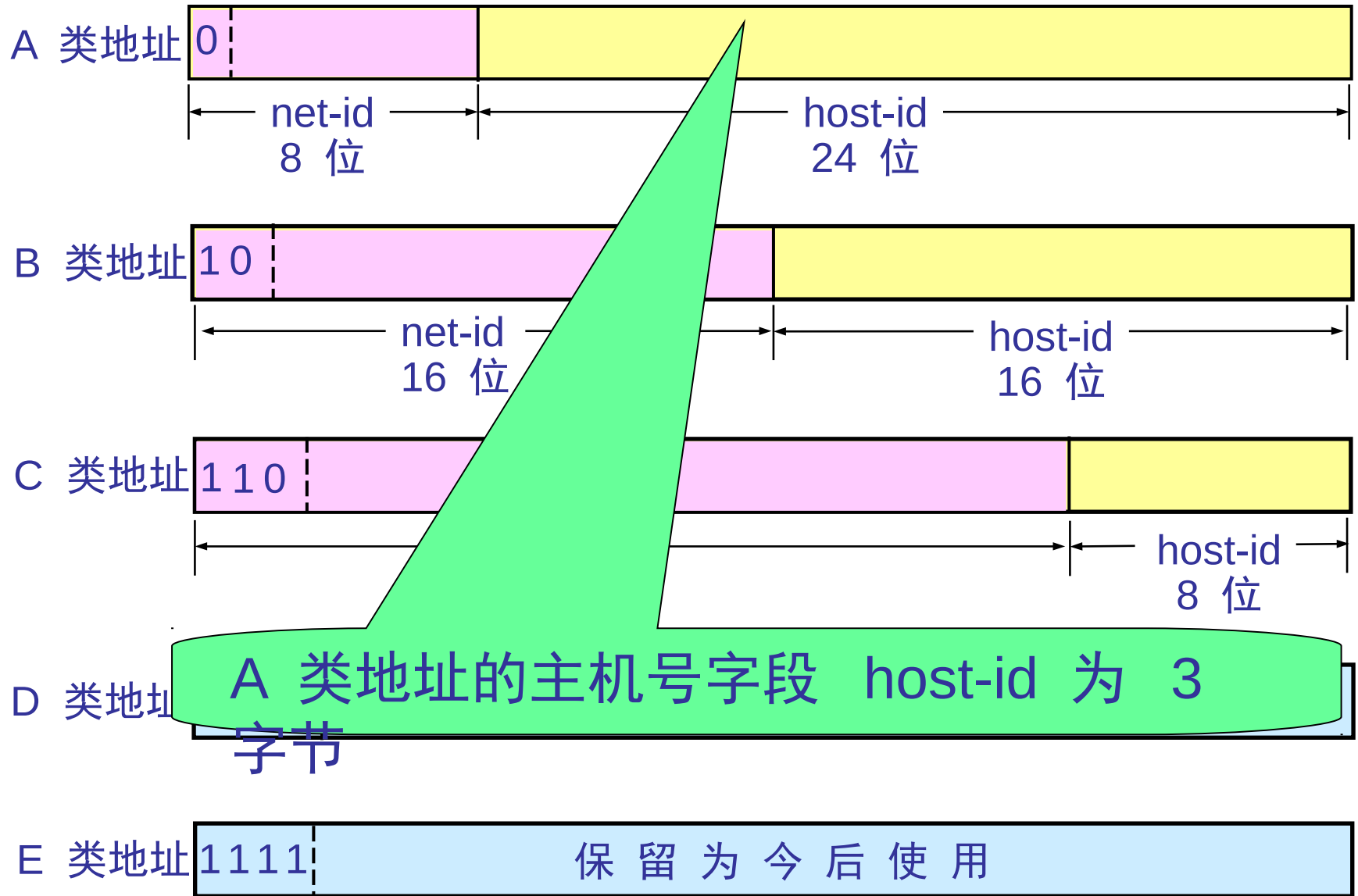


# IP 地址中的网络号字段和主机号字段

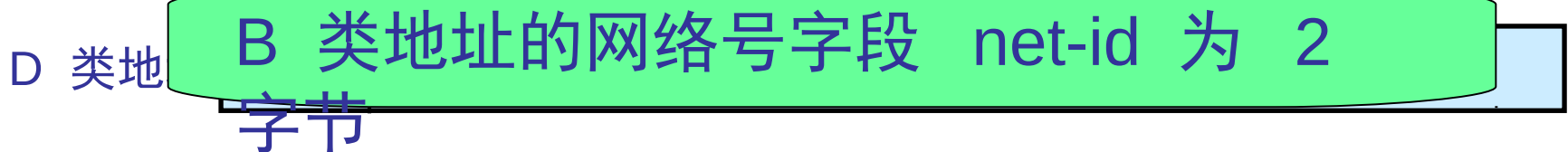
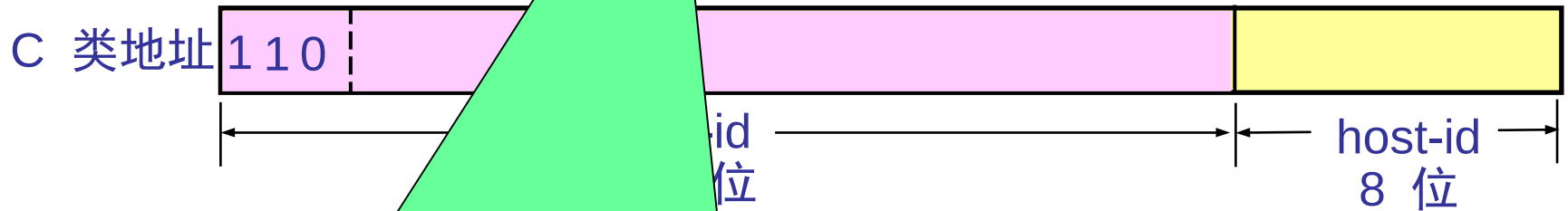
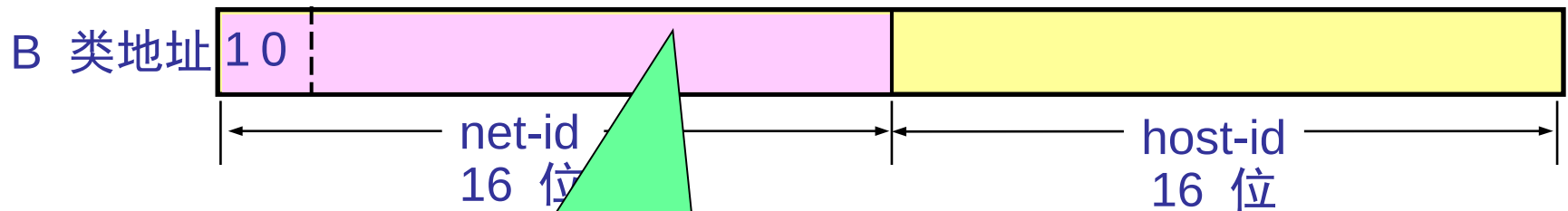
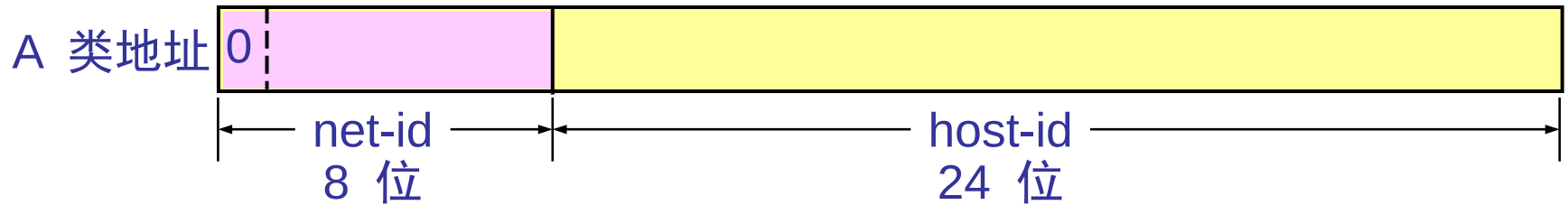




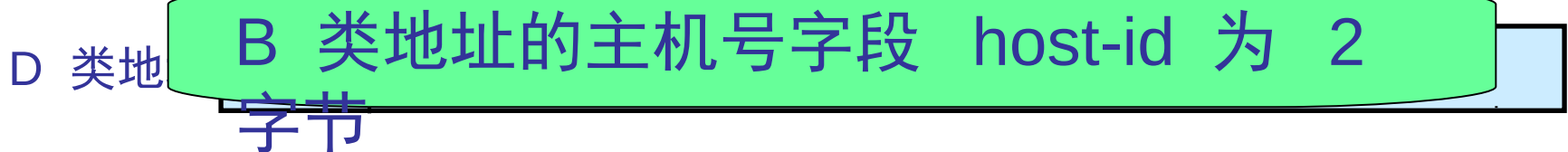
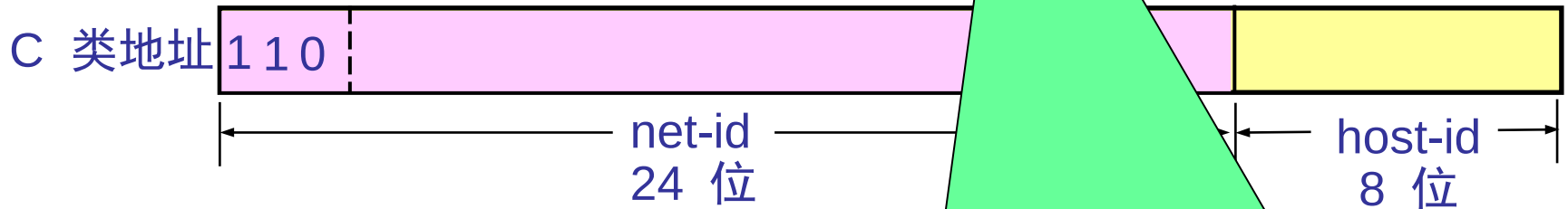
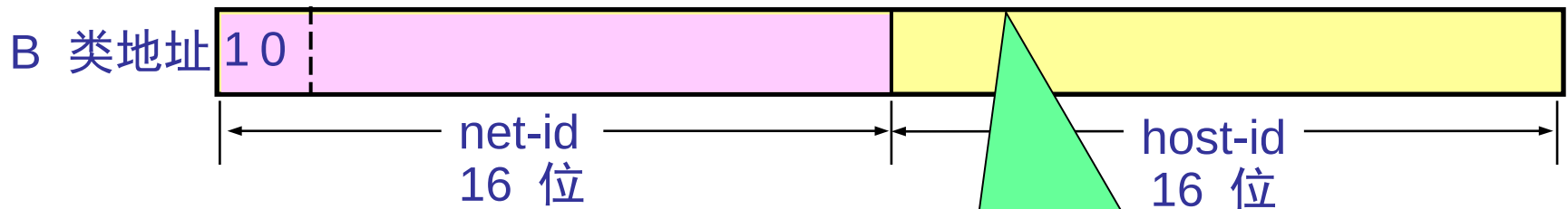
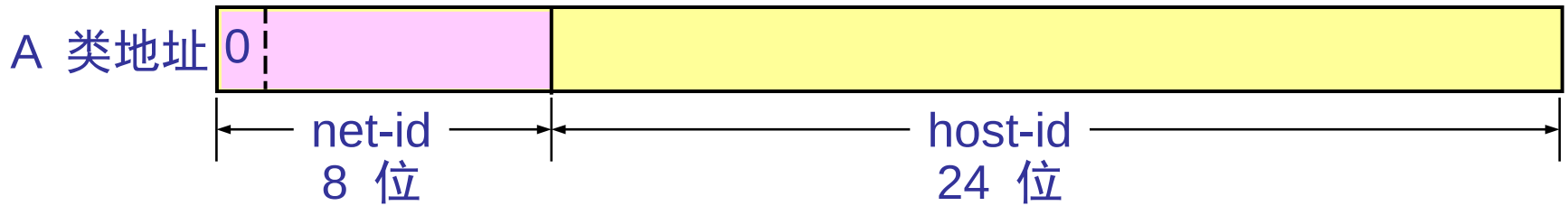
# IP 地址中的网络号字段和主机号字段



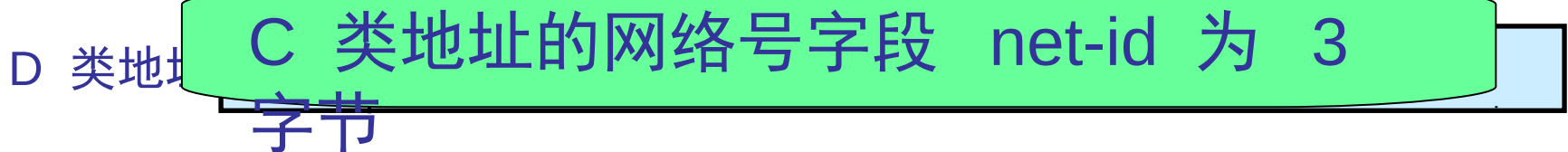
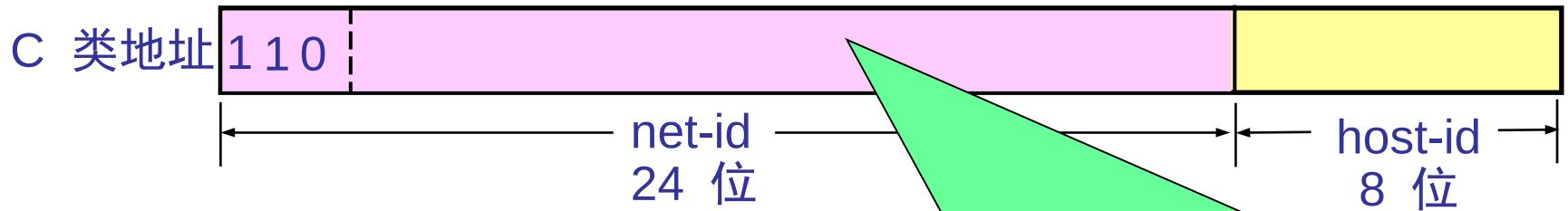
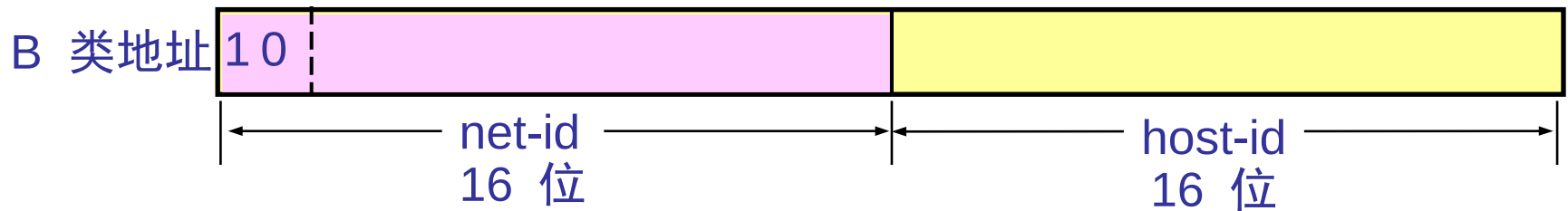
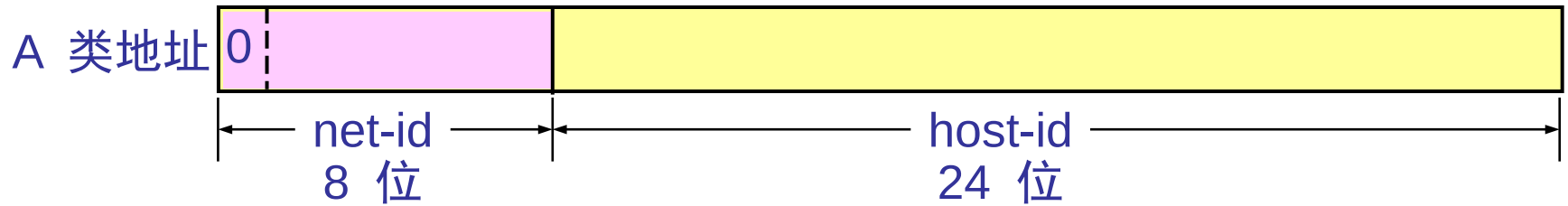
# IP 地址中的网络号字段和主机号字段



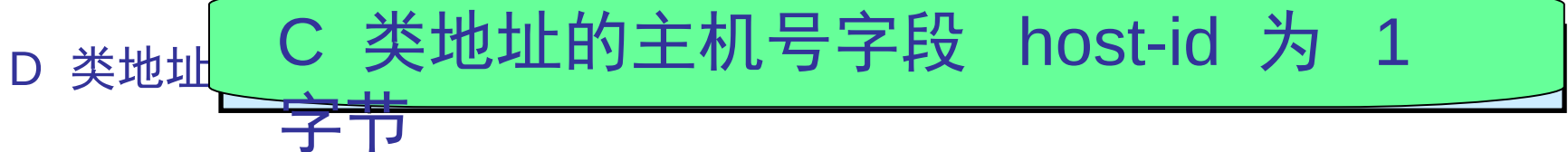
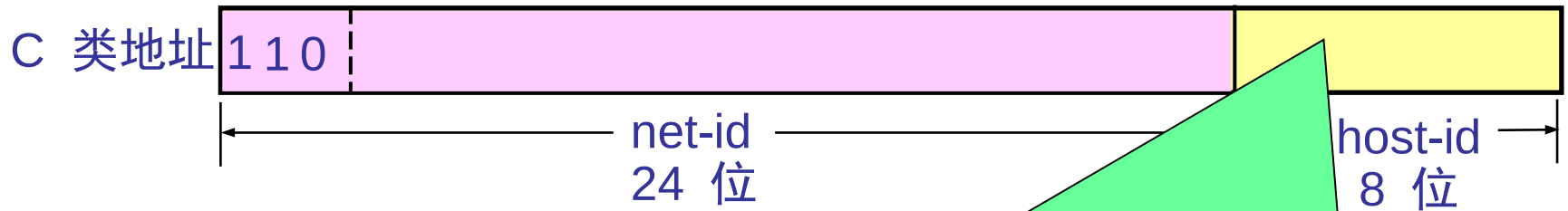
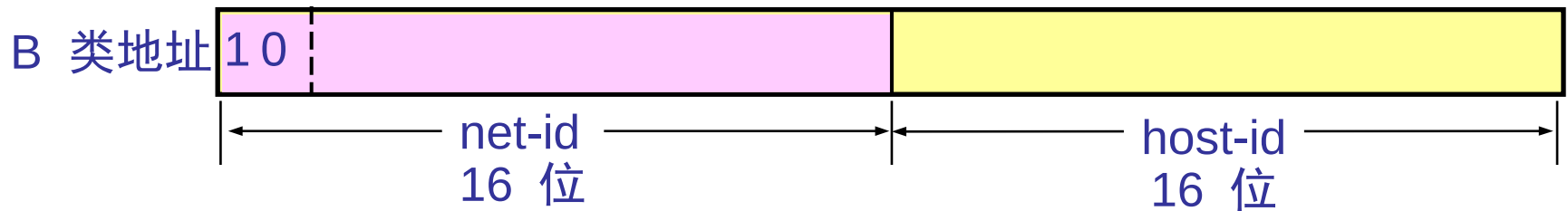
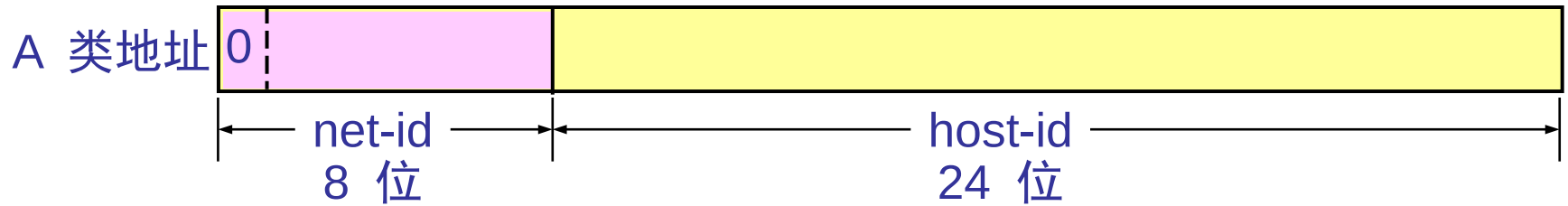
# IP 地址中的网络号字段和主机号字段



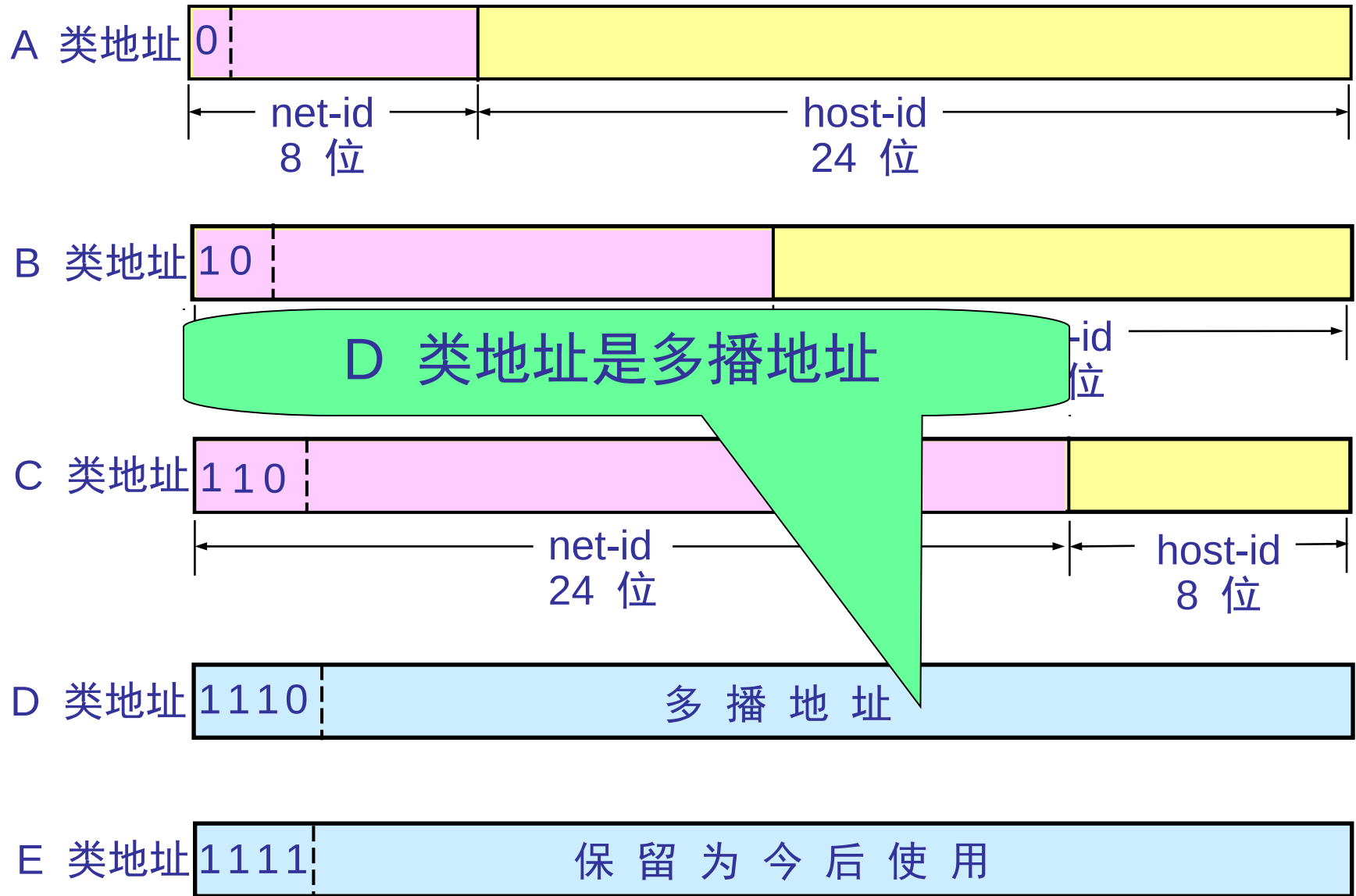
# IP 地址中的网络号字段和主机号字段



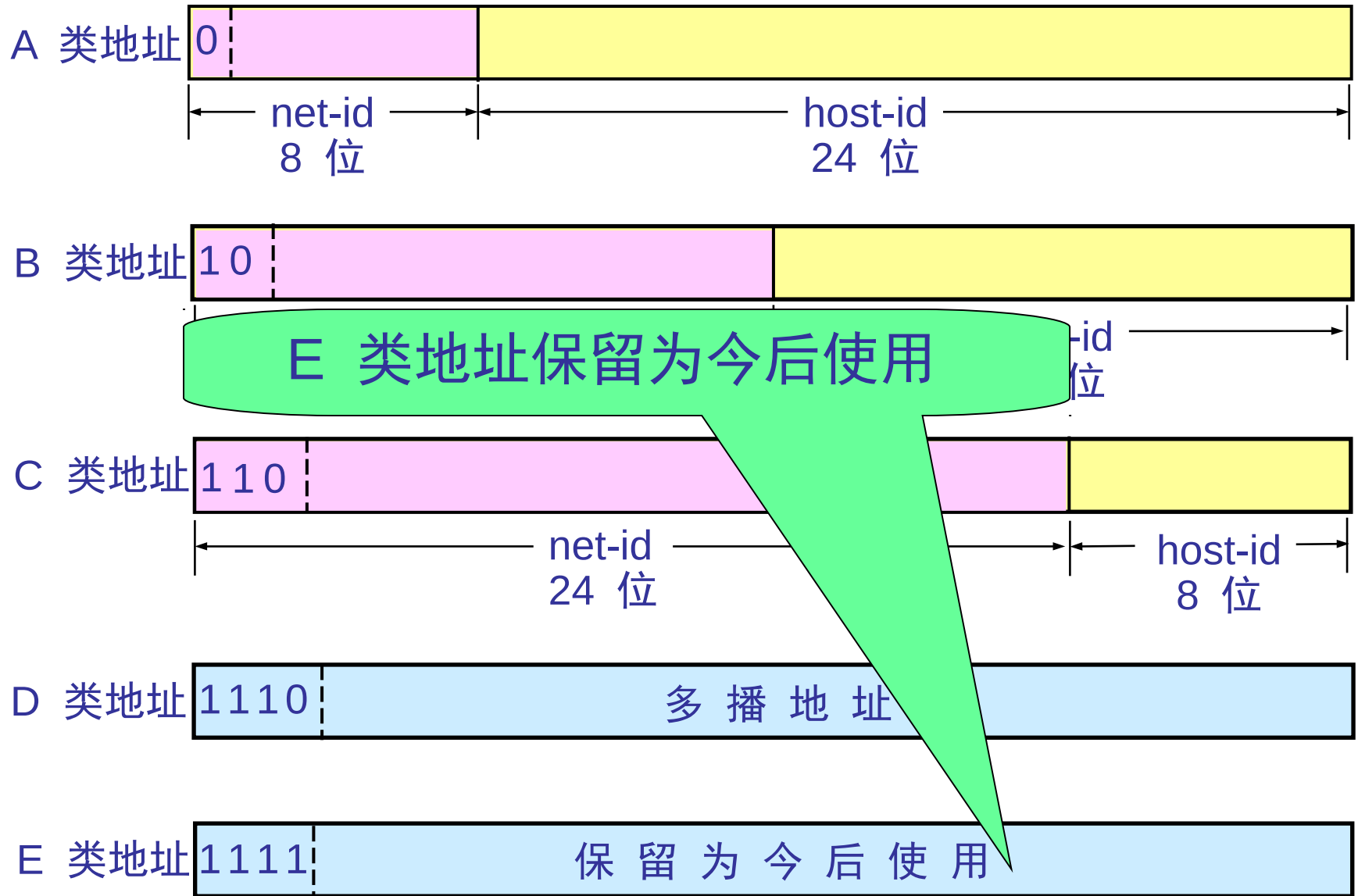
# IP 地址中的网络号字段和主机号字段



# IP 地址中的网络号字段和主机号字段



# IP 地址中的网络号字段和主机号字段





# A 类 IP 地址

---

- ◆ 网络号：其长度为 7 位，从理论上可以有  $2^7=128$  个网络，但网络号全 0 和全 1（即 0 和 127）的两个地址保留用于特殊目的，实际允许有 126 个不同的 A 类网络；
- ◆ 主机号：其长度为 24 位，从理论上每个 A 类网络的主机数为  $2^{24}=16777216$ ，但全 0 和全 1 的两个地址保留用于特殊目的，实际允许连接  $2^{24}-2=16777214$  个主机；
- ◆ A 类 IP 地址结构适用于有大量主机的大型网络。





## B 类 IP 地址

---

- ◆ 网络号：其长度为 14 位，允许有  $2^{14}-1=16383$  (128. 0. 0. 0 不用) 个不同的 B 类网络；
- ◆ 主机号：其长度为 16 位，理论上每个 B 类网络可以有  $2^{16}=65536$  个主机数，而**实际为  $2^{16}-2= 65534$  个**；
- ◆ B 类 IP 地址适用于一些国际性大公司与政府机构等中等大小的组织使用。



# C 类 IP 地址

---

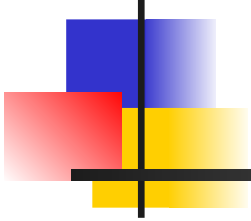
- ◆ 网络号：其长度为 21 位，因此允许有  $2^{21} - 1 = 2097152$  个不同的 C 类网络；
- ◆ 主机号：其长度为 8 位，每个 C 类网络的主机地址数最多为  $2^8 = 256$  个，实际允许连接  $2^8 - 2 = 254$  个主机；
- ◆ C 类 IP 地址适用于一些小公司与普通的研究机构。



## D 类和 E 类 IP 地址

---

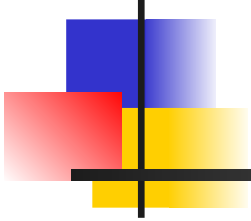
- ◆ D 类 IP 地址不标识网络，地址范围：224.0.0.0 ~ 239.255.255.255，用于其他特殊的用途，如多播地址；
- ◆ E 类 IP 地址暂时保留，地址范围：240.0.0.0 ~ 255.255.255.255，用于某些实验和将来使用。



# IP 地址分类

---

- A 类地址： 1.0.0.0~126.0.0.0 ； 主机号 24 位
- B 类地址： 128.1.0.0~191.255.0.0 ； 主机号 16 位
- C 类地址： 192.0.1.0~223.255.255.0 ； 主机号 8 位
- D 类地址： 224.0.0.0~239.255.255.255
- E 类地址： 240.0.0.0~255.255.255.254



# 局域网地址

---

- 在 IP 地址的 A、B、C 类地址中，各保留了 3 个区域作为局域网地址，其地址范围如下：

A类地址： 10.0.0.0 ~ 10.255.255.255

B类地址： 172.16.0.0 ~ 172.31.255.255

C类地址： 192.168.0.0 ~ 192.168.255.255



# 特殊的 IP 地址

以下这些 **IP** 地址具有特殊的含义：

00...00	0000	...	0000
---------	------	-----	------

本网络中的本机地址，仅在启动时用

00...00	主 机 号		
---------	-------	--	--

本网中的特定主机

11...11	1111	...	1111
---------	------	-----	------

本网络中的进行广播地址，各路由器均不转发。

网络号	1111	...	1111
-----	------	-----	------

指定网络中的所有主机进行广播。

网络号	0000	...	0000
-----	------	-----	------

指定网络的网络地址

127	任 意 值		
-----	-------	--	--

回路，在本地软件测试时用

# 点分十进制记法

机器中存放的 IP 地址  
是 32 位 二进制代码

100000000000010110000001100011111

每隔 8 位插入一个空格  
能够提高可读性

10000000 00001011 00000011 00011111

将每 8 位的二进制数  
转换为十进制数

128 11 3 31

采用点分十进制记法  
则进一步提高可读性

128.11.3.31



# IP 地址的一些重要特点

---

(1) IP 地址是一种**分等级的地址结构**。分两个等级的好处是：

- 第一，IP 地址管理机构在分配 IP 地址时**只分配网络号**，而剩下的主机号则由拥有该网络号的单位自行分配。
- 第二，**路由器**仅根据目的主机所连接的**网络号**来转发分组（而不考虑目的主机号），这样就可以使路由表中的项目数大幅度减少，从而减小了路由表所占的存储空间。





# IP 地址的一些重要特点

- (2) 实际上 IP 地址是标志一个主机（或路由器）和一条链路的接口。
- 当一个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 net-id 必须是不同的。这种主机称为多归属主机。
  - 由于一个路由器至少应当连接到两个网络（这样它才能将 IP 数据报从一个网络转发到另一个网络），因此一个路由器至少应当有两个不同的 IP 地址。

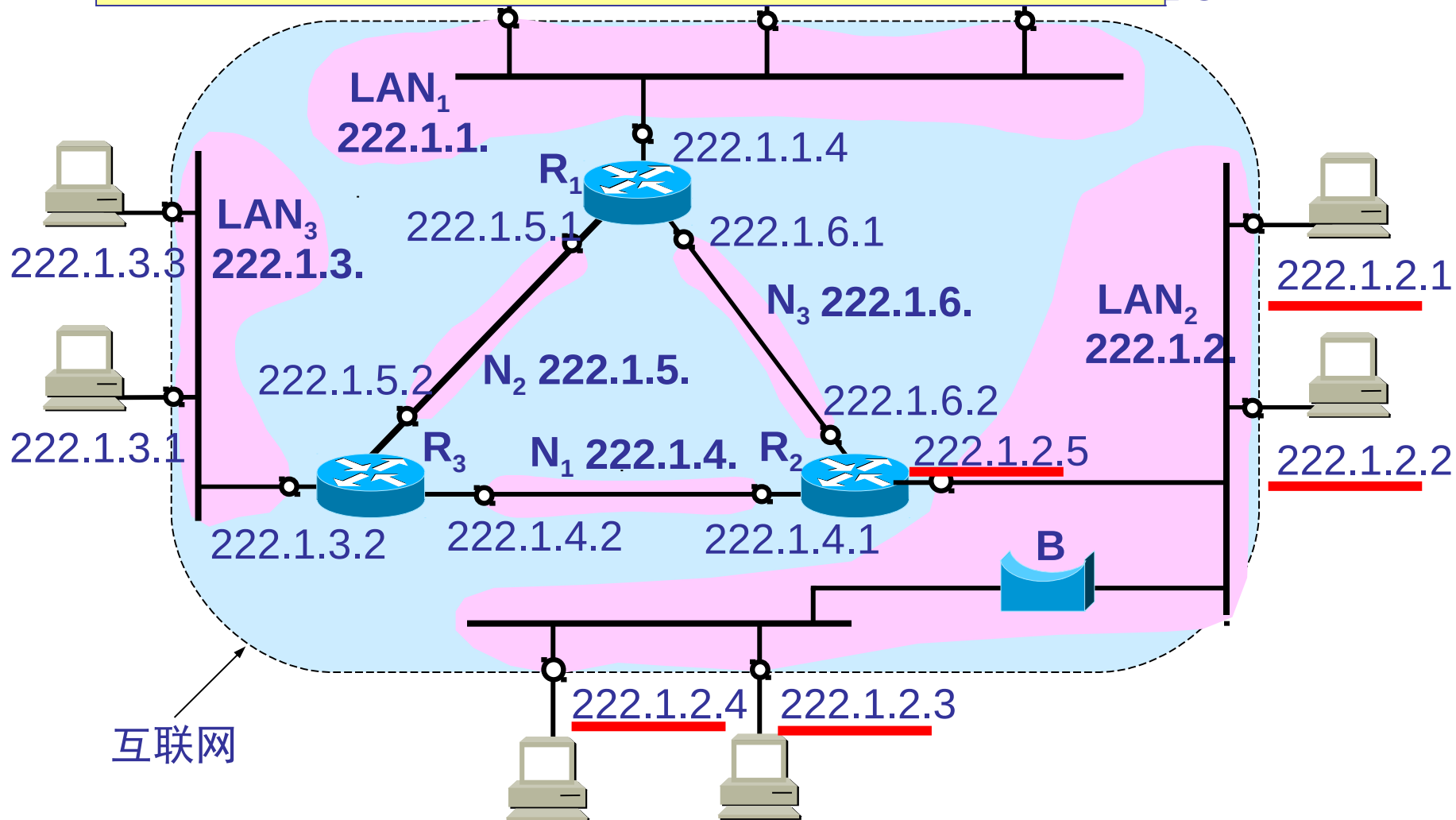


# IP 地址的一些重要特点

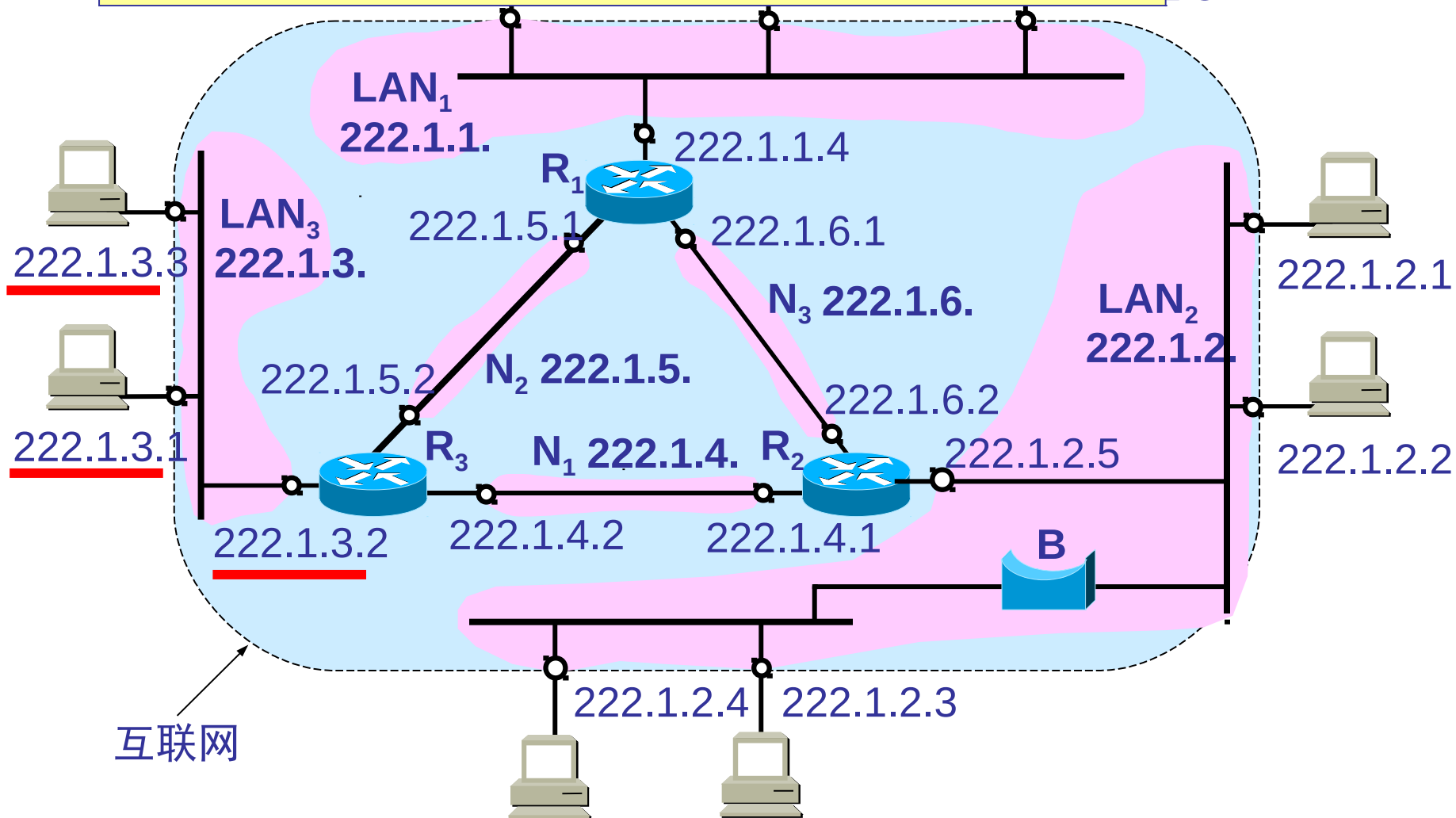
---

- (3) 用转发器或网桥连接起来的若干个局域网仍为一个网络，因此这些局域网都具有同样的网络号 net-id 。
- (4) 所有分配到网络号 net-id 的网络，范围很小的局域网，还是可能覆盖很大地理范围的广域网，都是平等的。

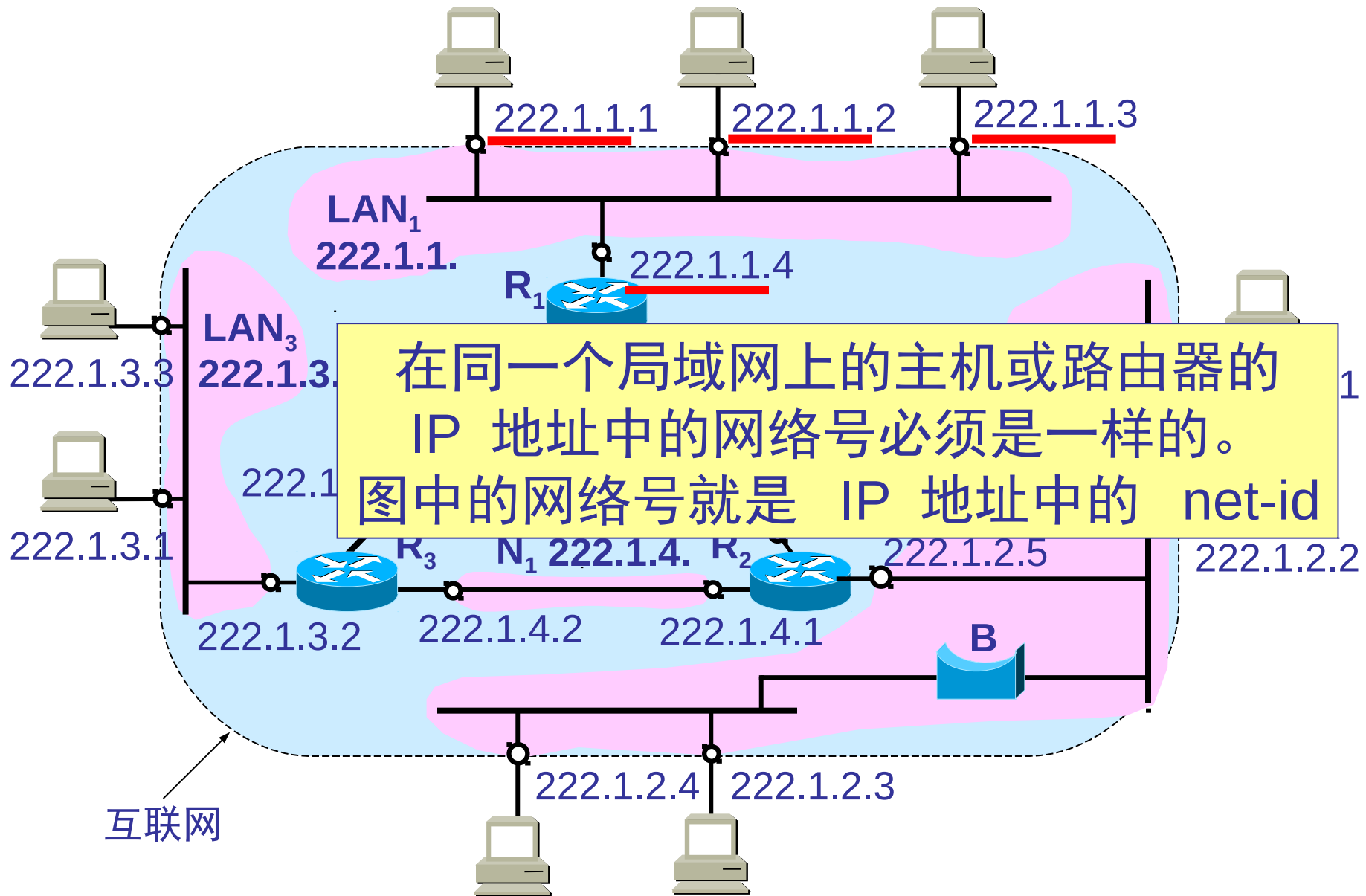
在同一个局域网上的主机或路由器的  
IP 地址中的网络号必须是一样的。  
图中的网络号就是 IP 地址中的 net-id



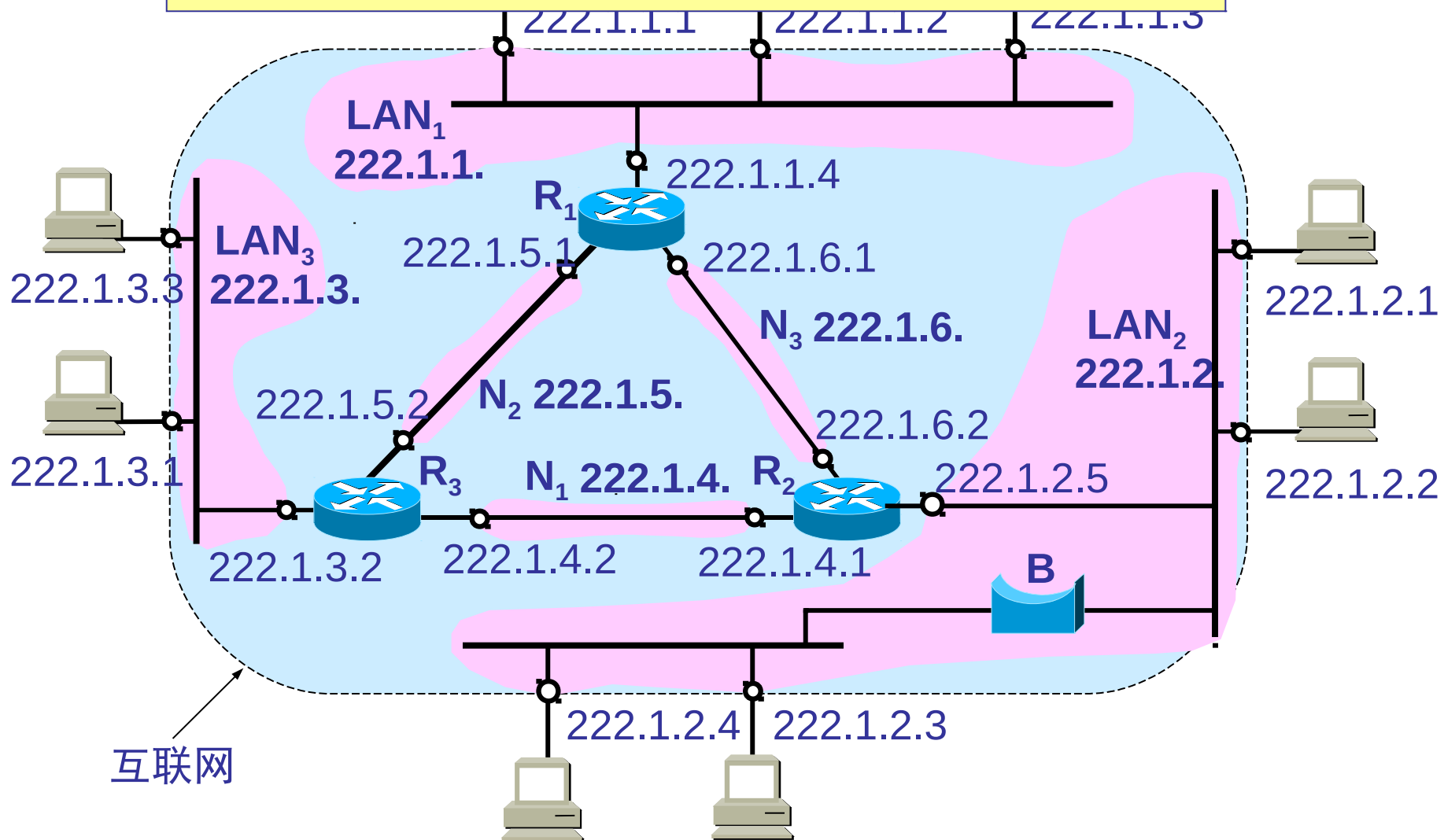
在同一个局域网上的主机或路由器的  
IP 地址中的网络号必须是一样的。  
图中的网络号就是 IP 地址中的 net-id



# 互联网中的 IP 地址



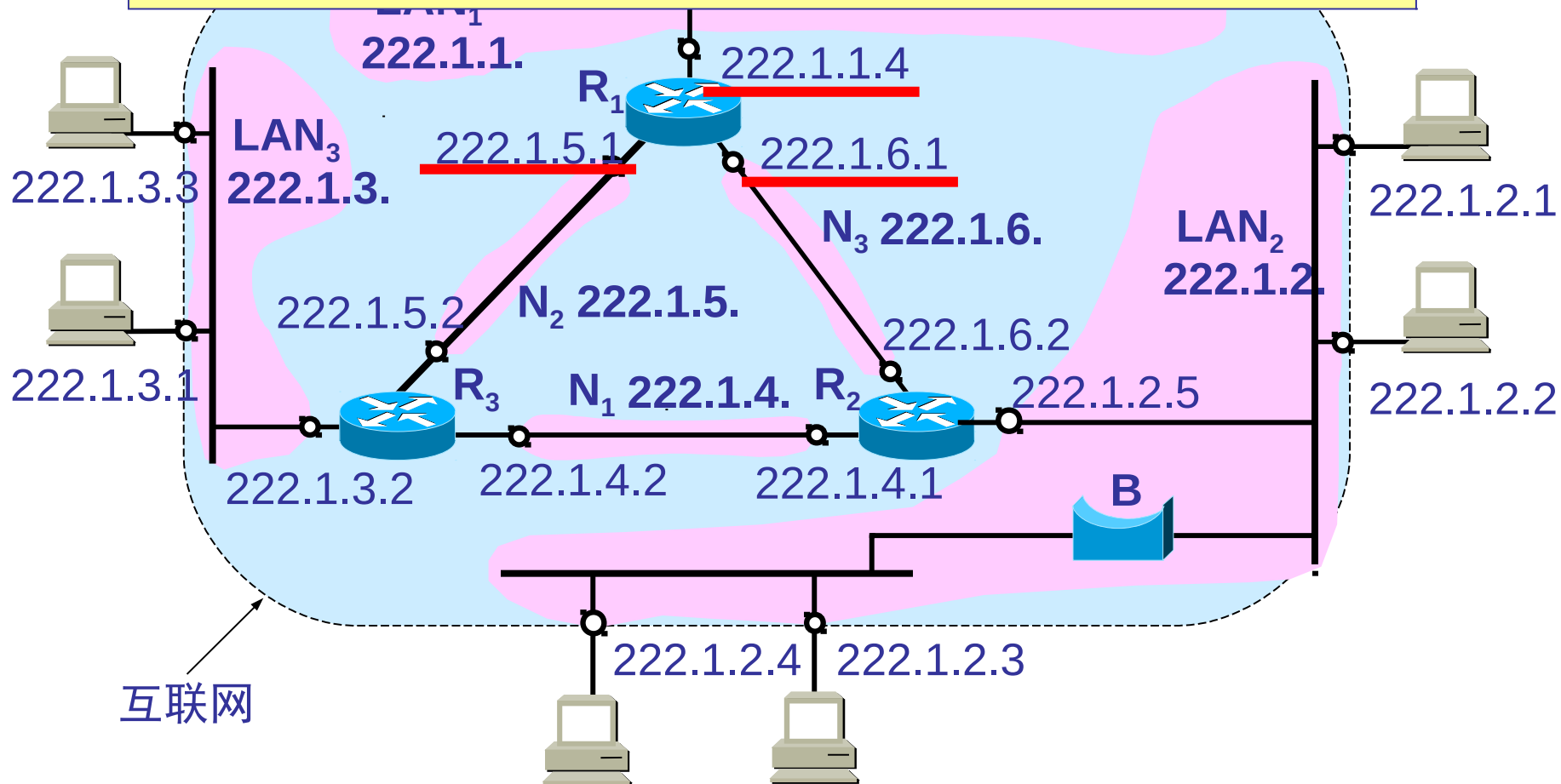
在同一个局域网上的主机或路由器的  
IP 地址中的网络号必须是一样的。  
图中的网络号就是 IP 地址中的 net-id



路由器总是具有两个或两个以上的 IP 地址

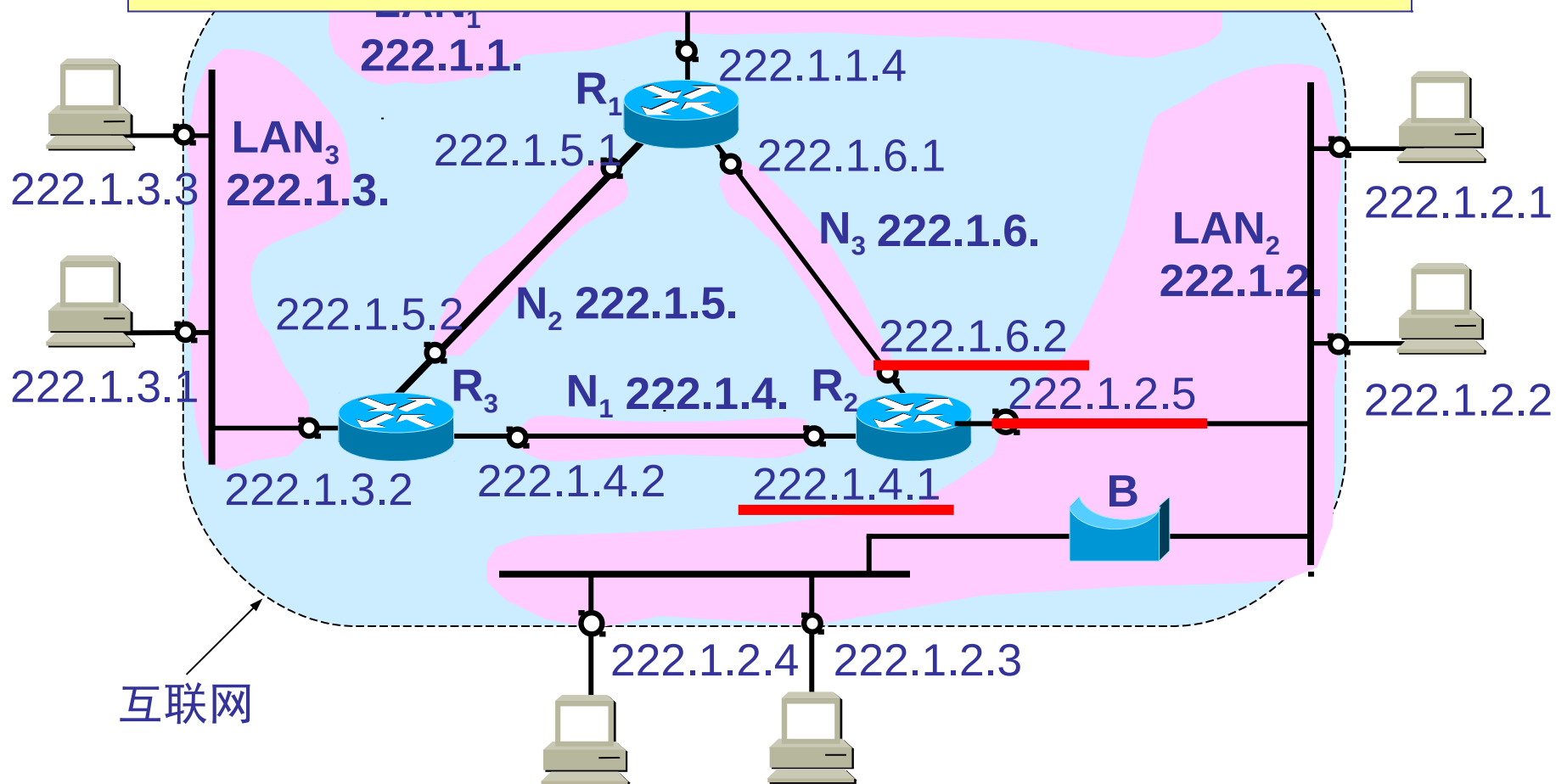
。

路由器的每一个接口都有一个  
不同网络号的 IP 地址。



路由器总是具有两个或两个以上的 IP 地址

。路由器的每一个接口都有一个不同网络号的 IP 地址。

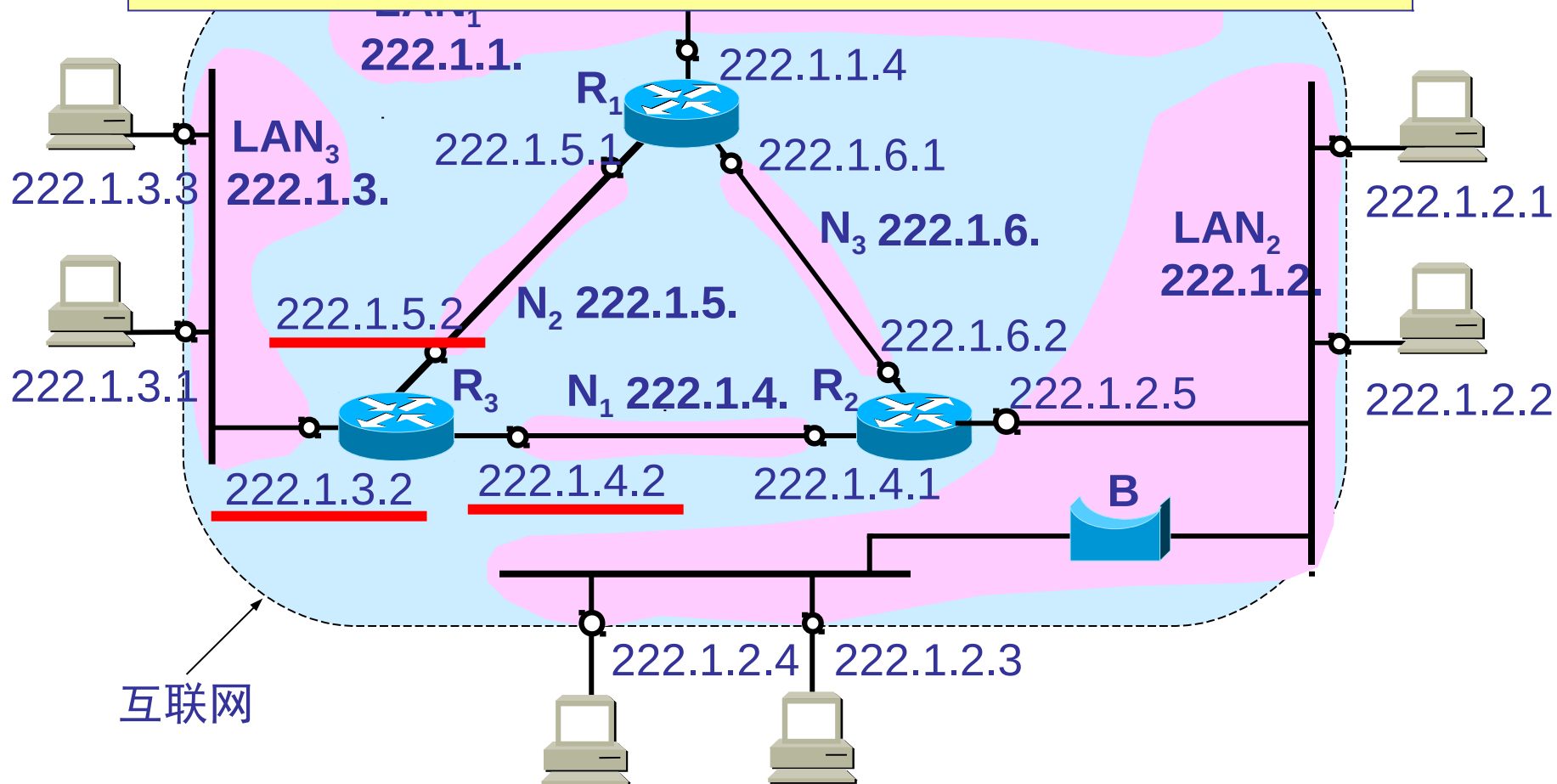




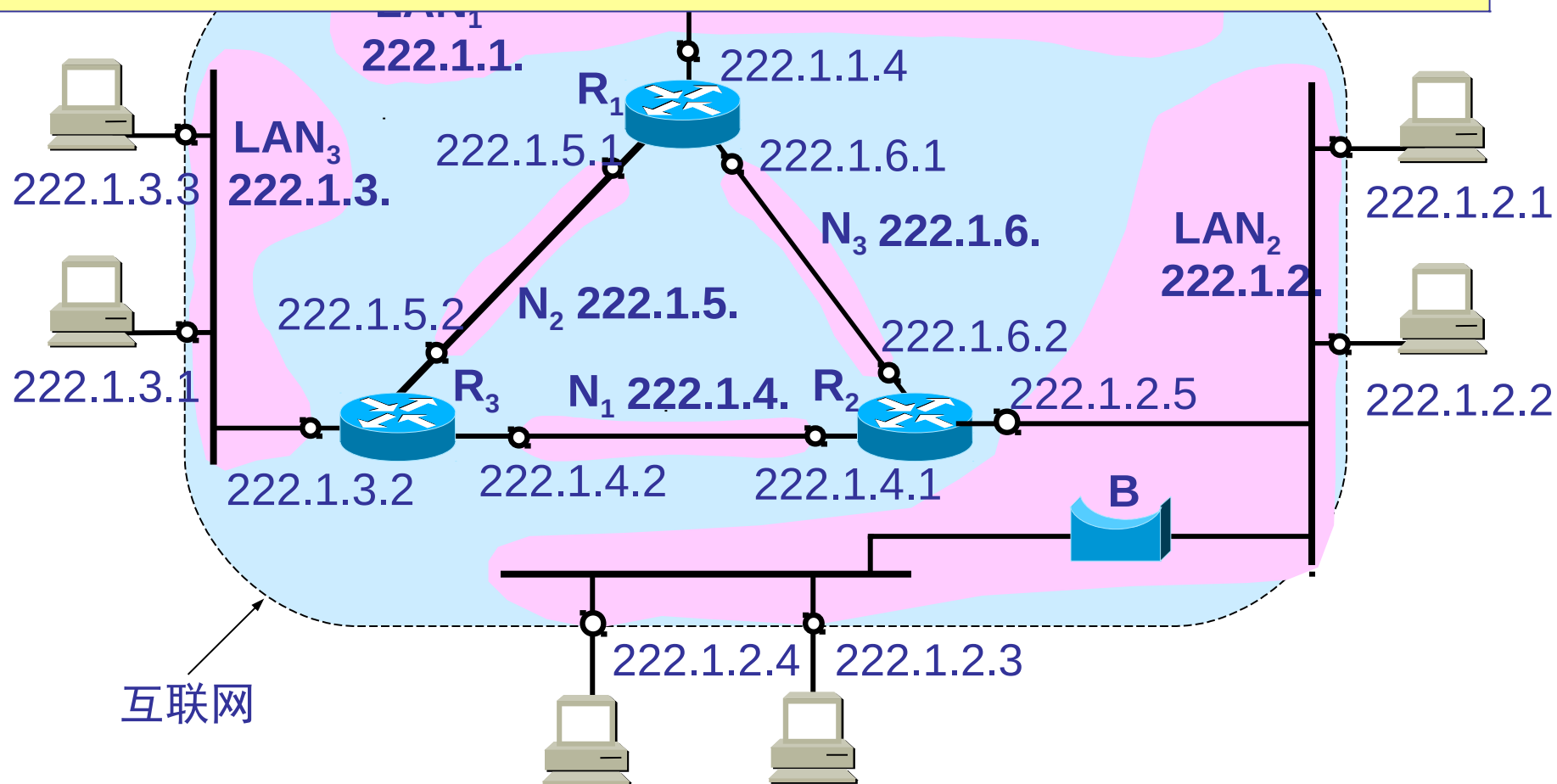
路由器总是具有两个或两个以上的 IP 地址

。

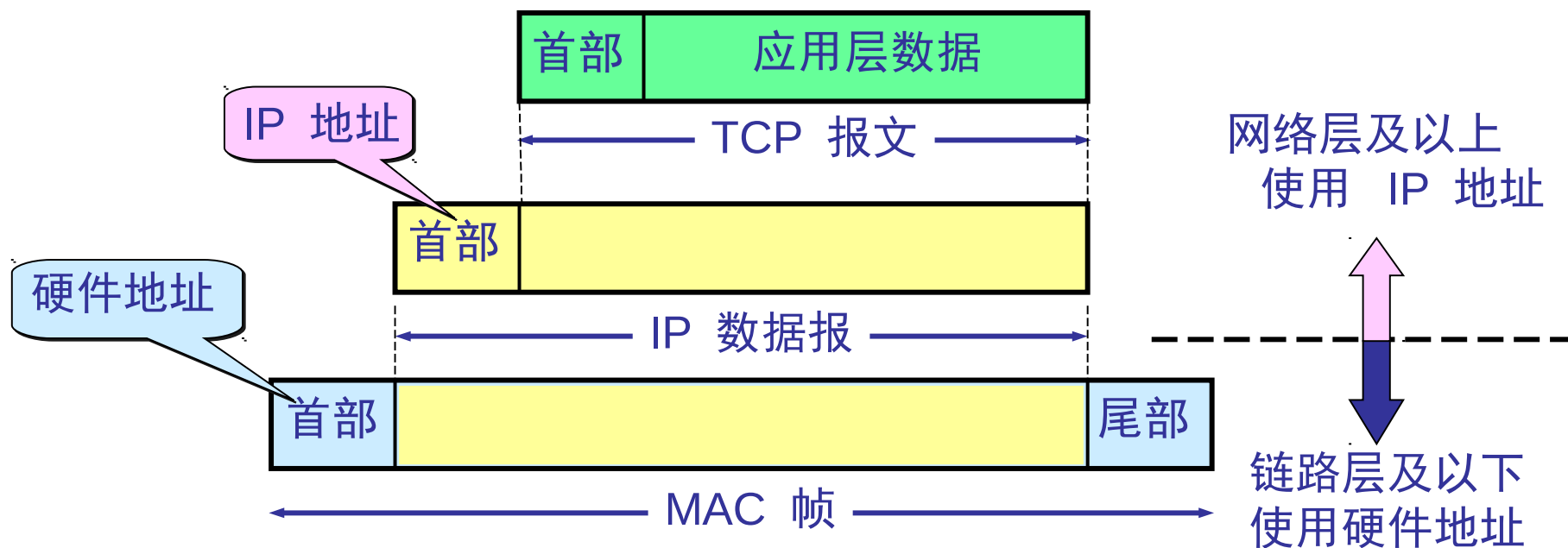
路由器的每一个接口都有一个  
不同网络号的 IP 地址。

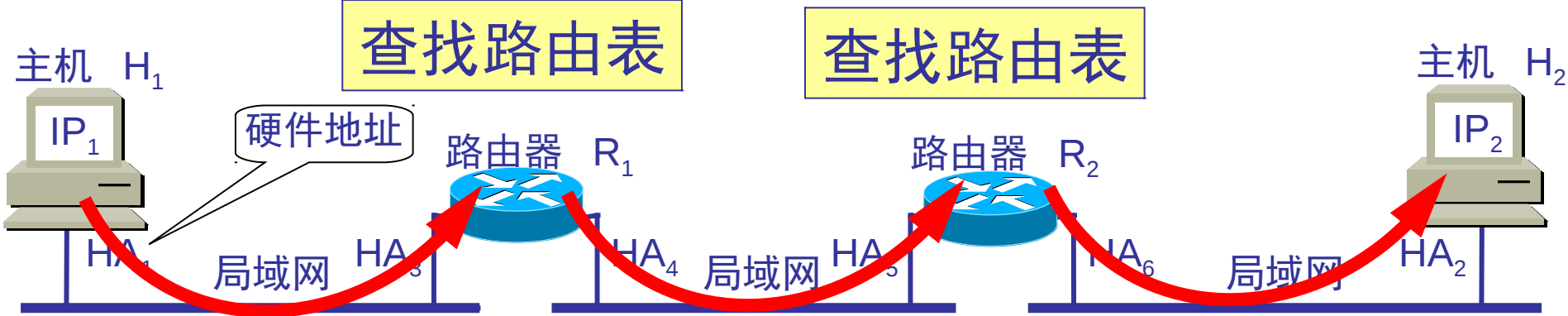


两个路由器直接相连的接口处，可指明也可不指明 IP 地址。如指明 IP 地址，则这一段连线就构成了一种只包含一段线路的特殊“网络”。现在常不指明 IP 地址。



## 4.2.3 IP 地址与硬件地址

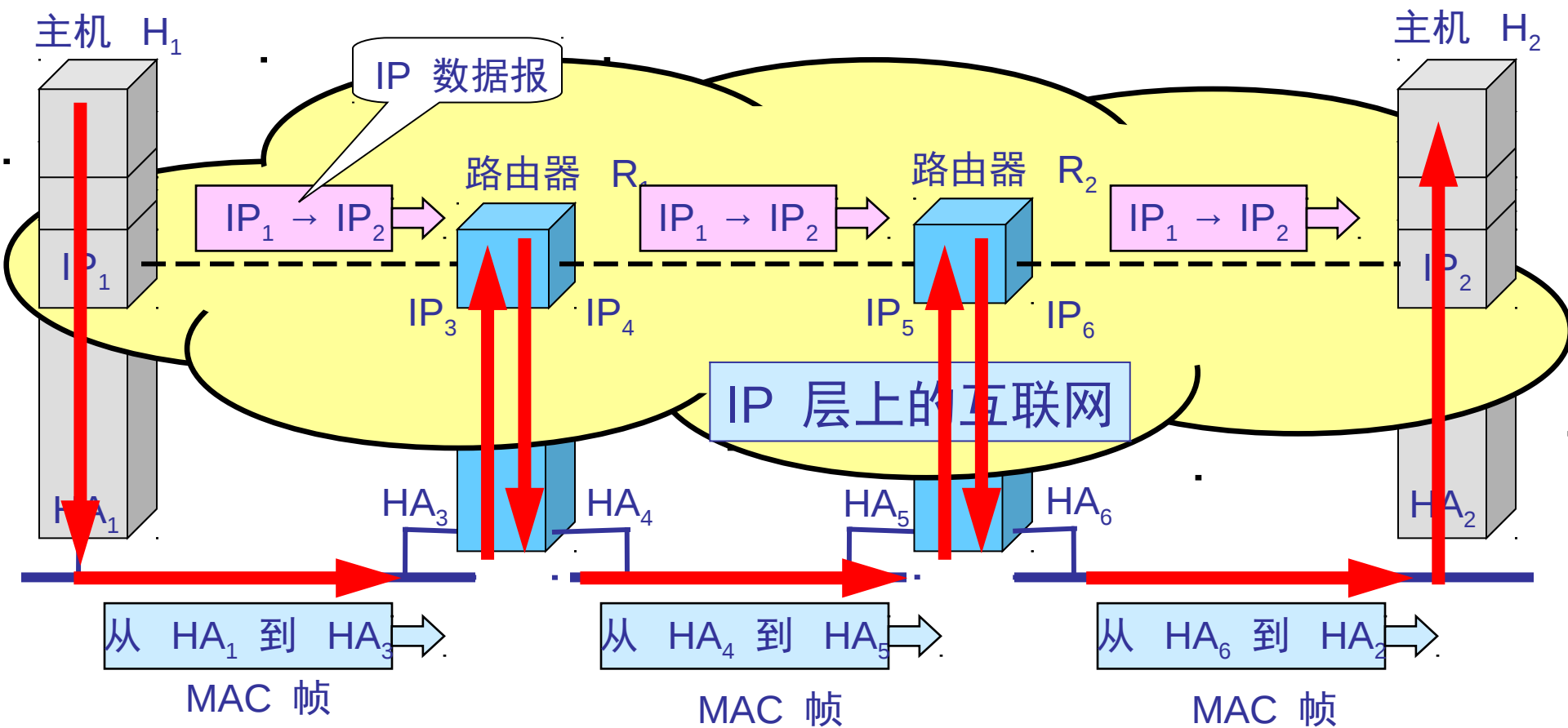
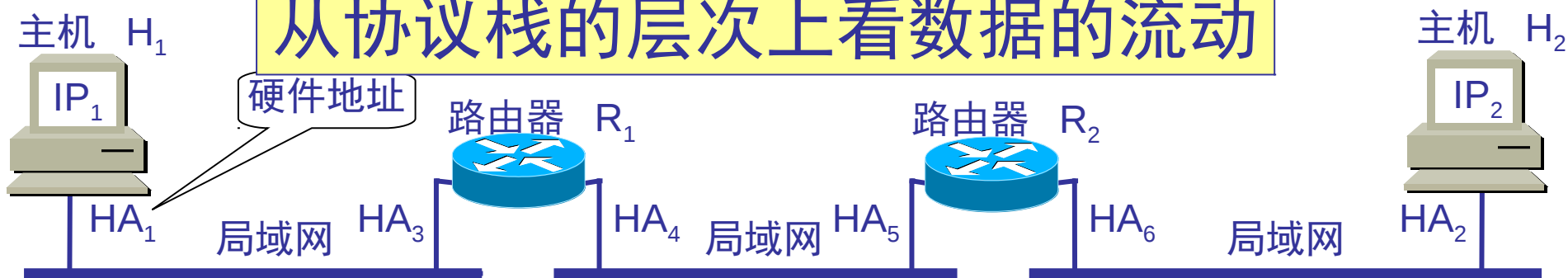




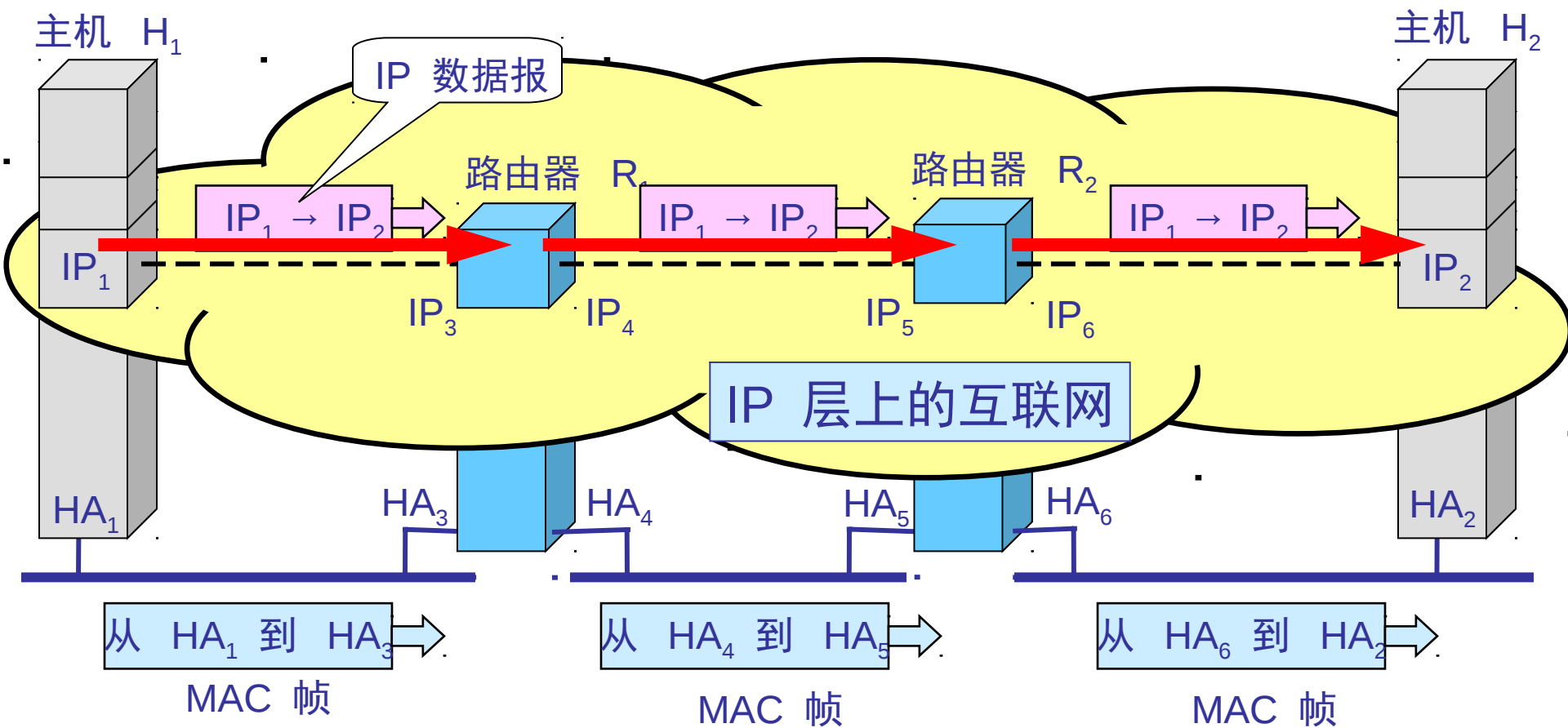
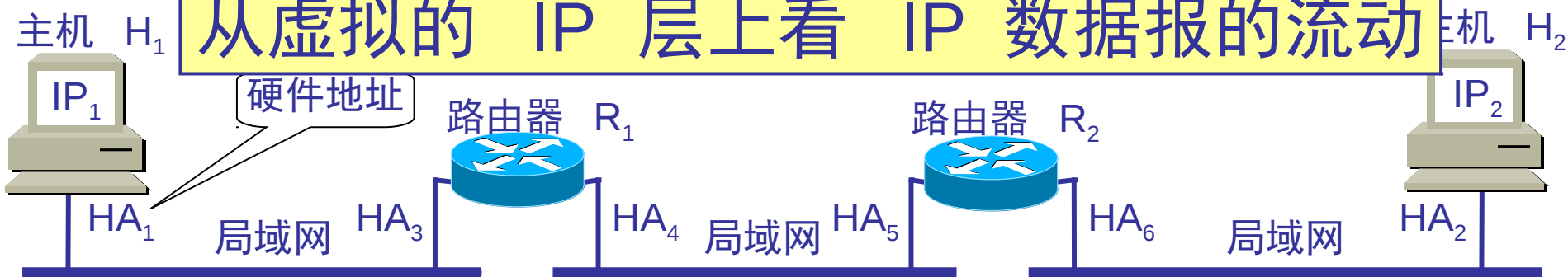
### 通信的路径

$H_1 \rightarrow$  经过  $R_1$  转发  $\rightarrow$  再经过  $R_2$  转发  $\rightarrow H_2$

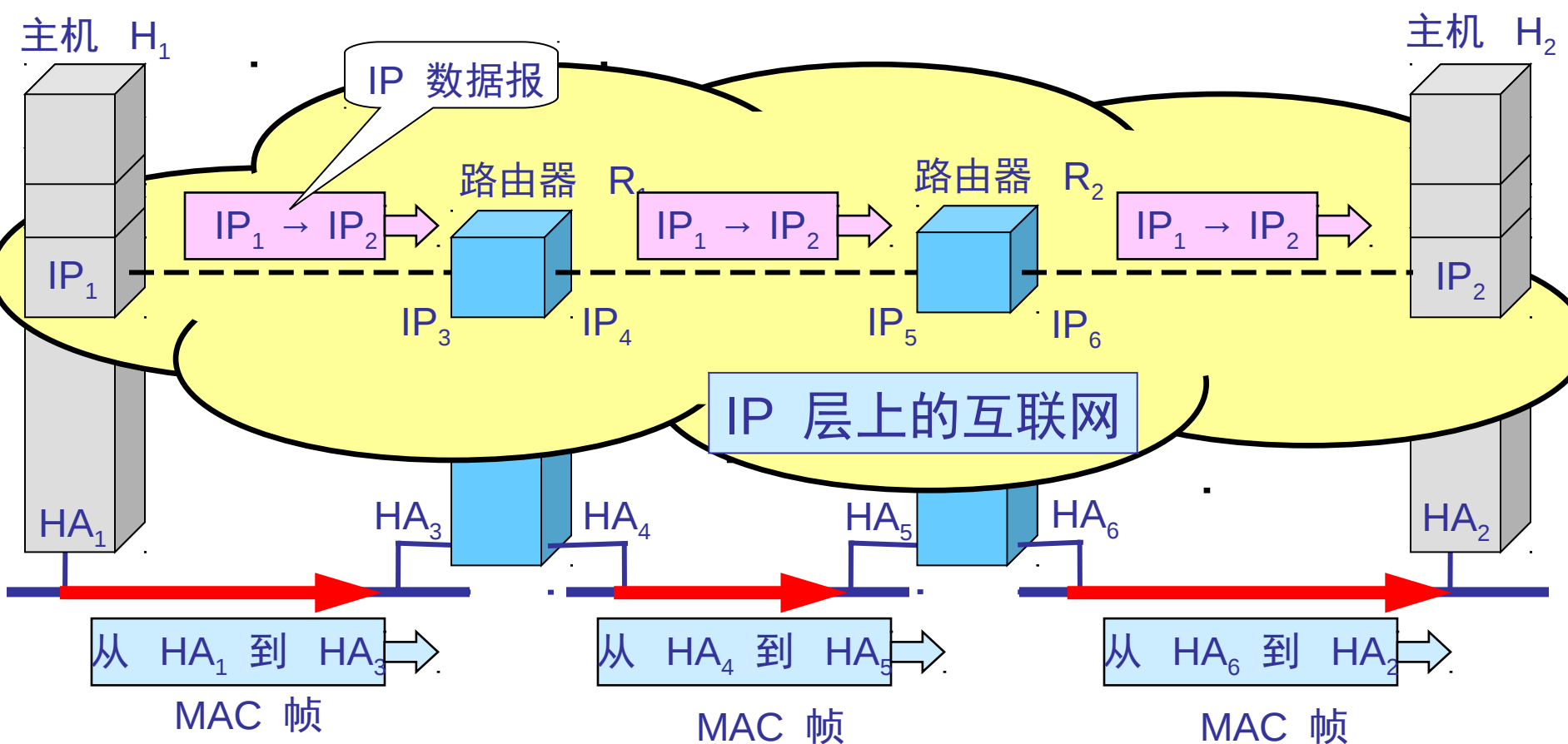
# 从协议栈的层次上看数据的流动



# 从虚拟的 IP 层上看 IP 数据报的流动

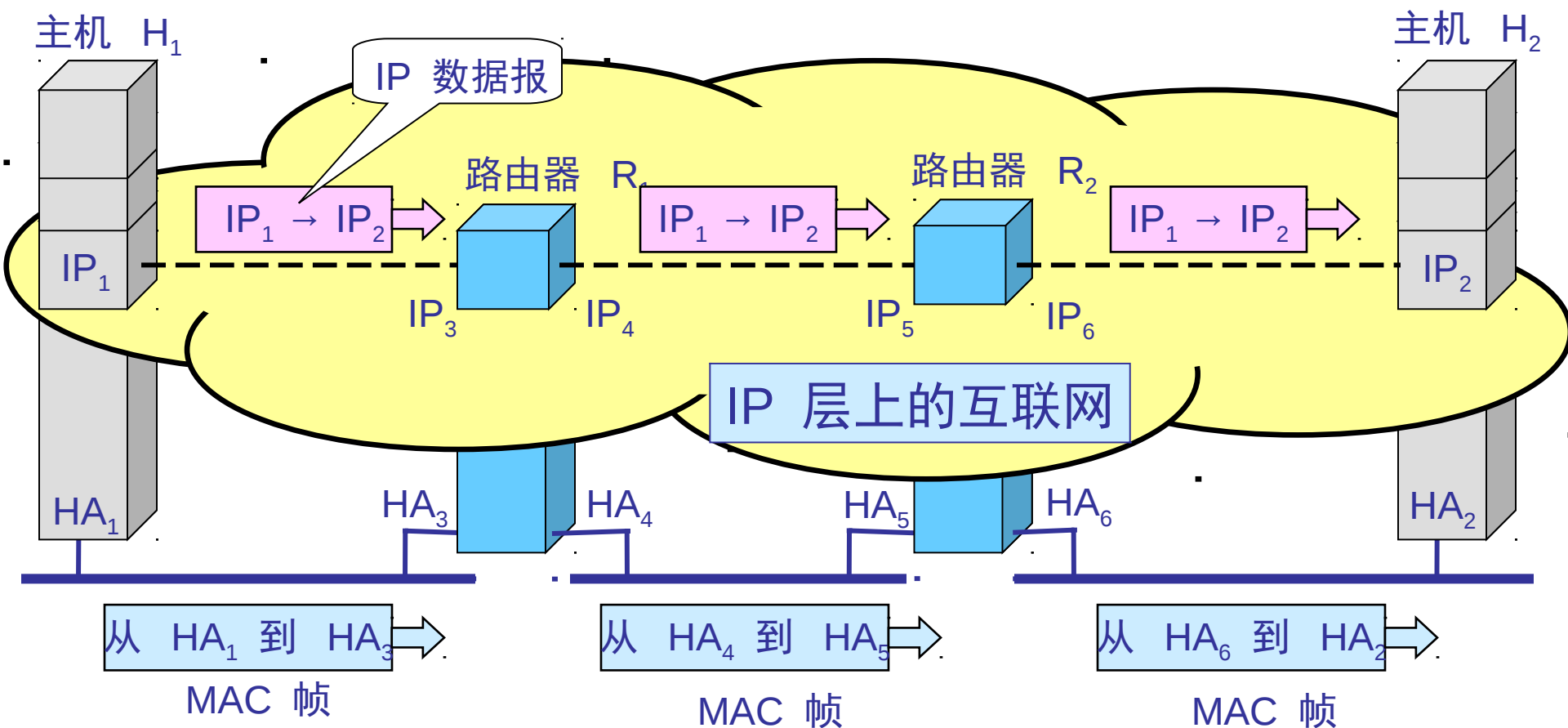


# 在链路上看 MAC 帧的流动



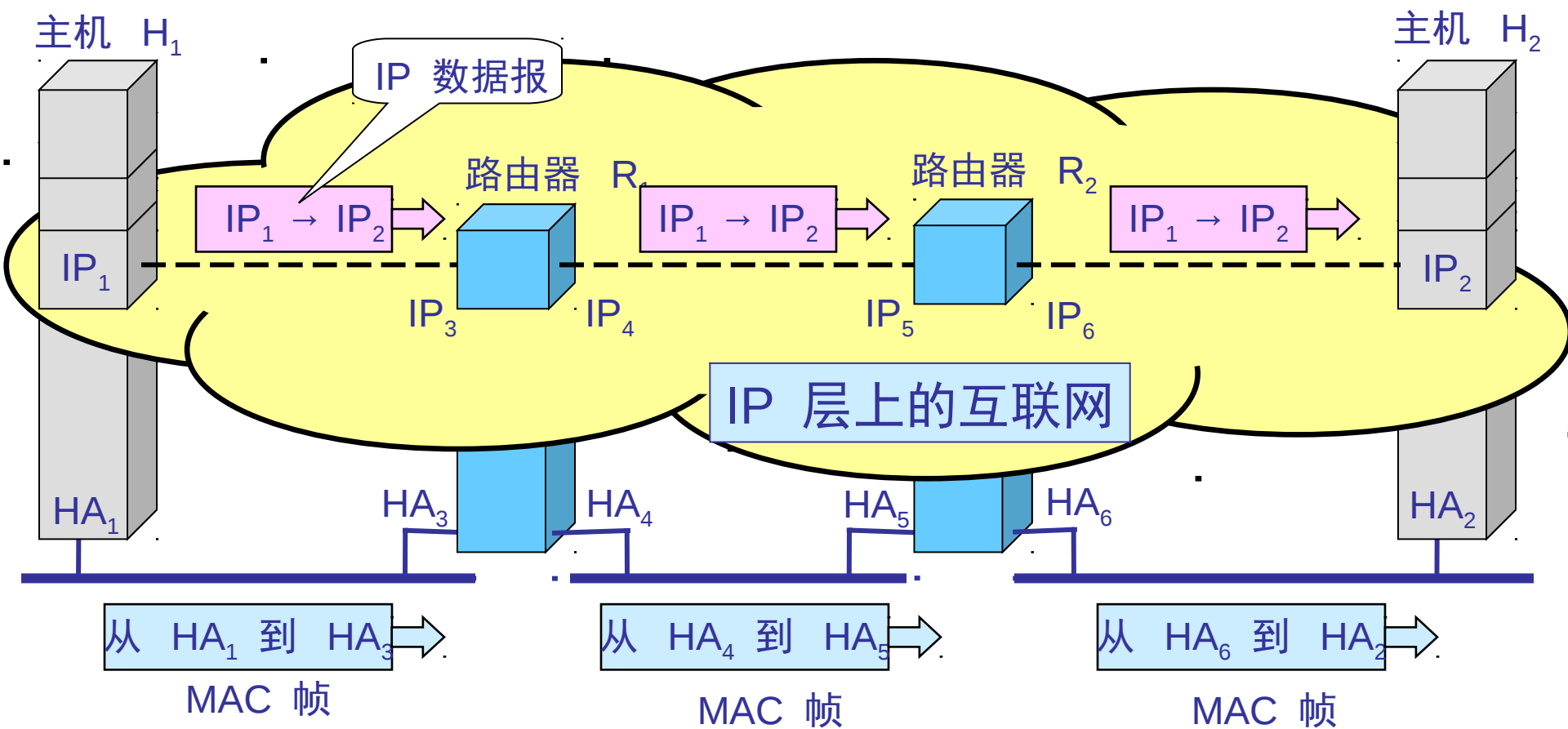
在 IP 层抽象的互联网上只能看到 IP 数据报

图中的  $IP_1 \rightarrow IP_2$  表示从源地址  $IP_1$  到目的地址  $IP_2$   
两个路由器的 IP 地址并不出现在 IP 数据报的首部中

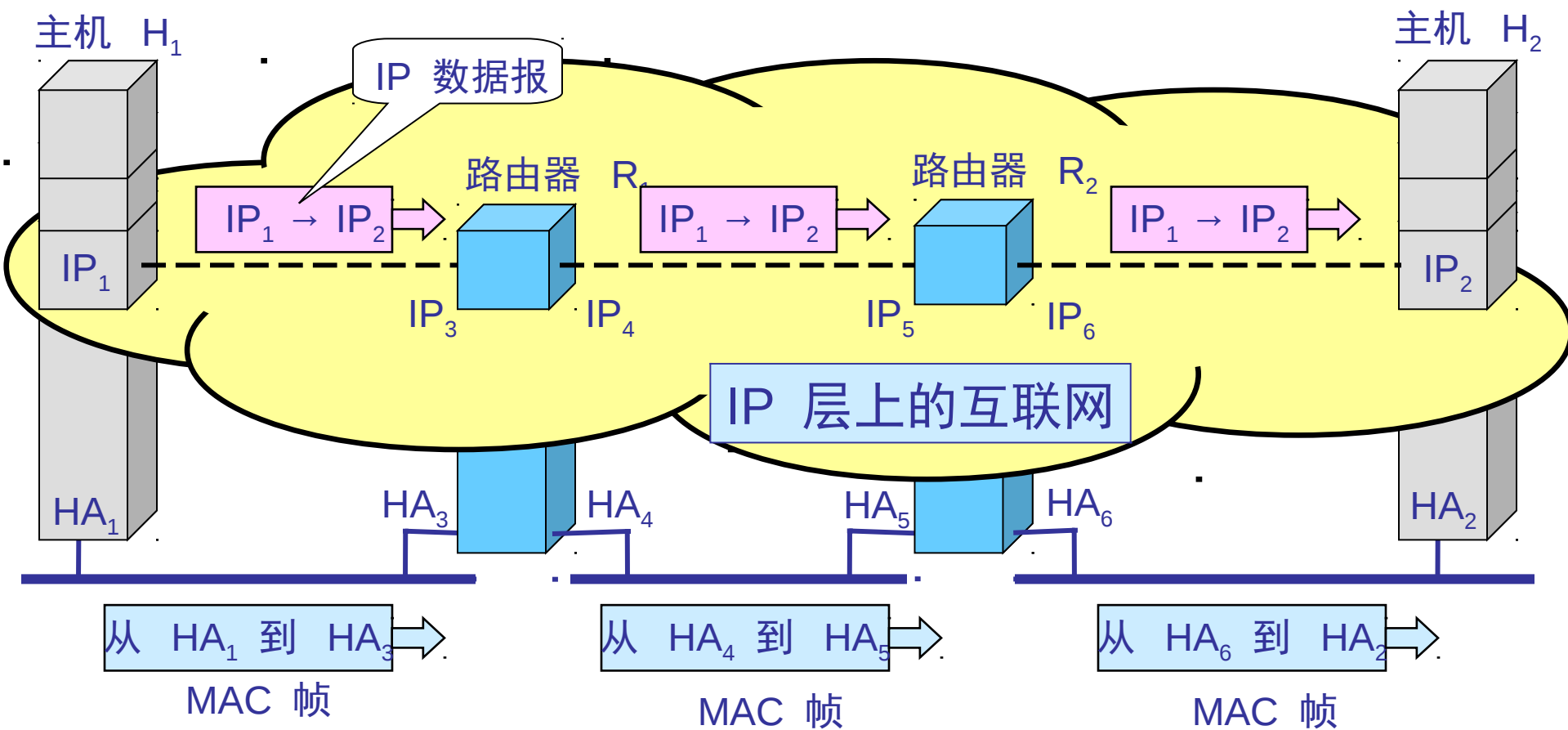




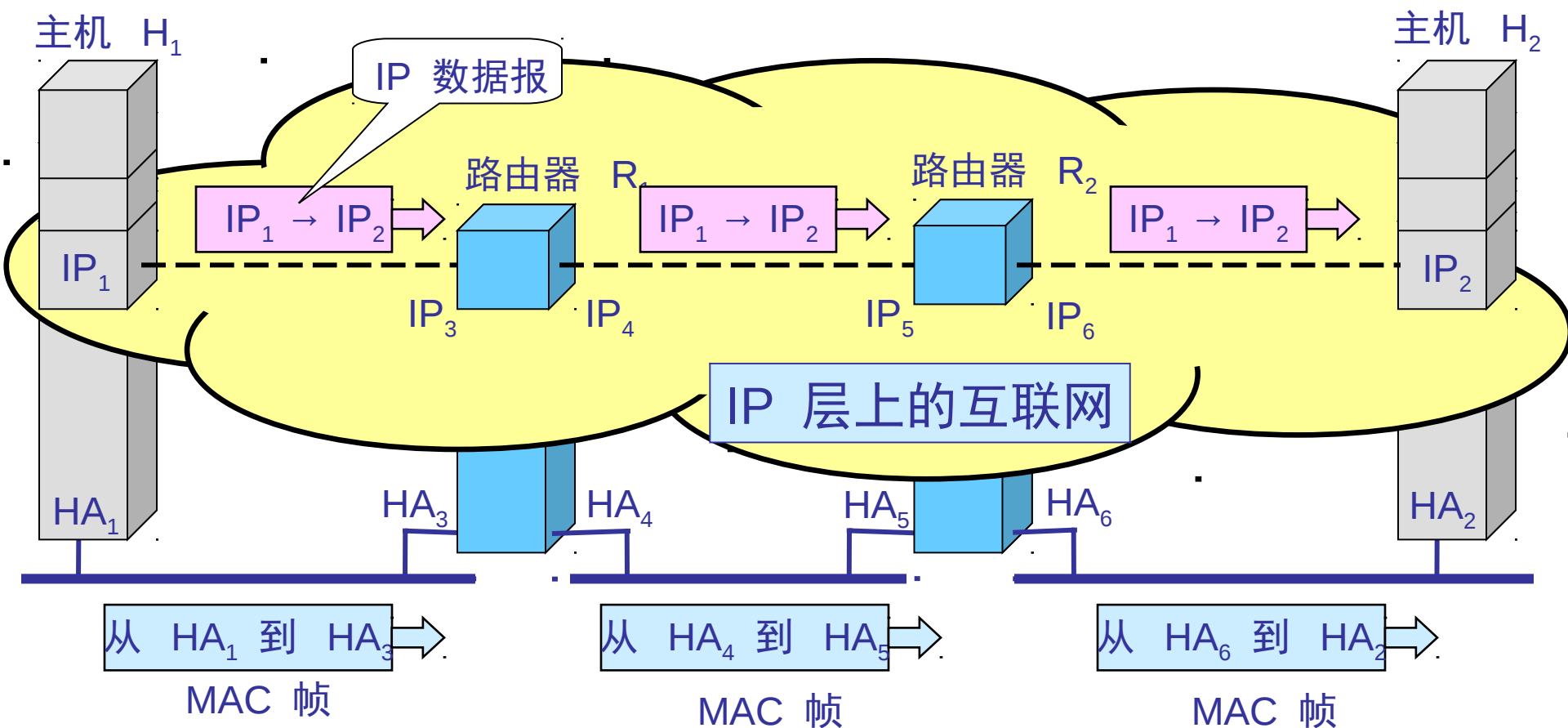
路由器只根据目的站的 IP 地址的网络号进行路由选择



在具体的物理网络的链路层  
只能看见 MAC 帧而看不见 IP 数据报



IP 层抽象的互联网屏蔽了下层很复杂的细节  
在抽象的网络层上讨论问题，就能够使用  
统一的、抽象的 IP 地址  
研究主机和主机或主机和路由器之间的通信



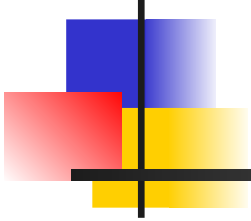


## 4.2.5 IP 协议

---

### IP 协议的特点：

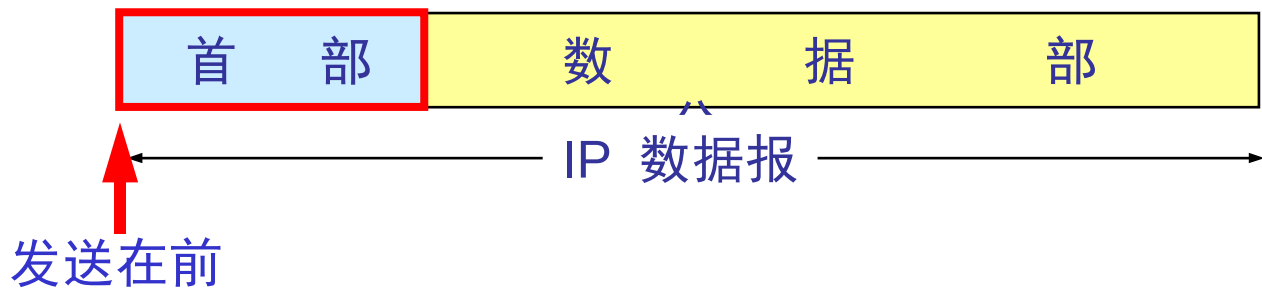
- IP 协议是一种不可靠、无连接的数据报传送服务协议 ；
- IP 协议是点 - 点的网络层通信协议 ；
- IP 协议向传输层屏蔽了物理网络的差异 ；

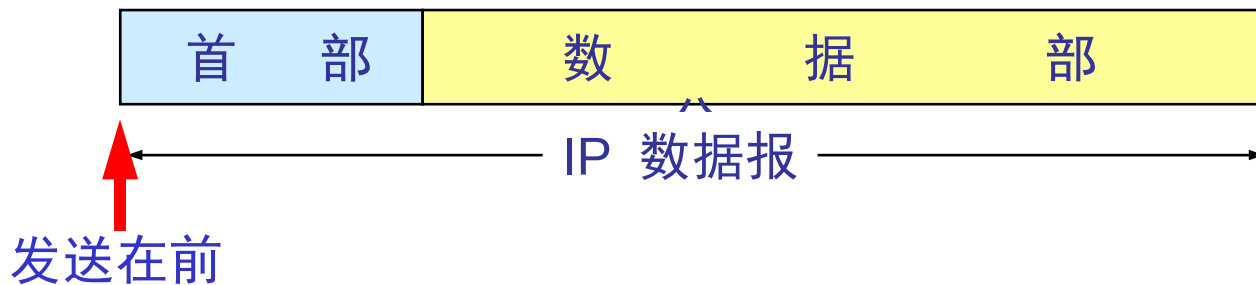


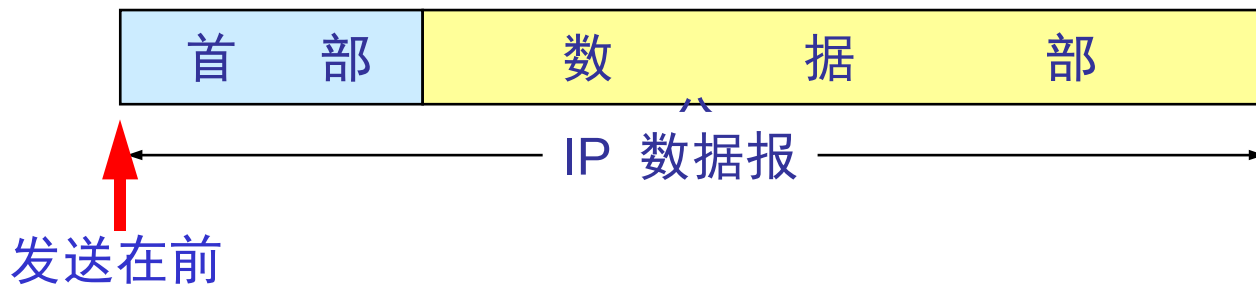
# IP 数据报的格式

---

- 一个 IP 数据报由首部和数据两部分组成。
- 首部的前一部分是固定长度，共 20 字节，是所有 IP 数据报必须具有的。
- 在首部的固定部分的后面是一些可选字段，其长度是可变的。

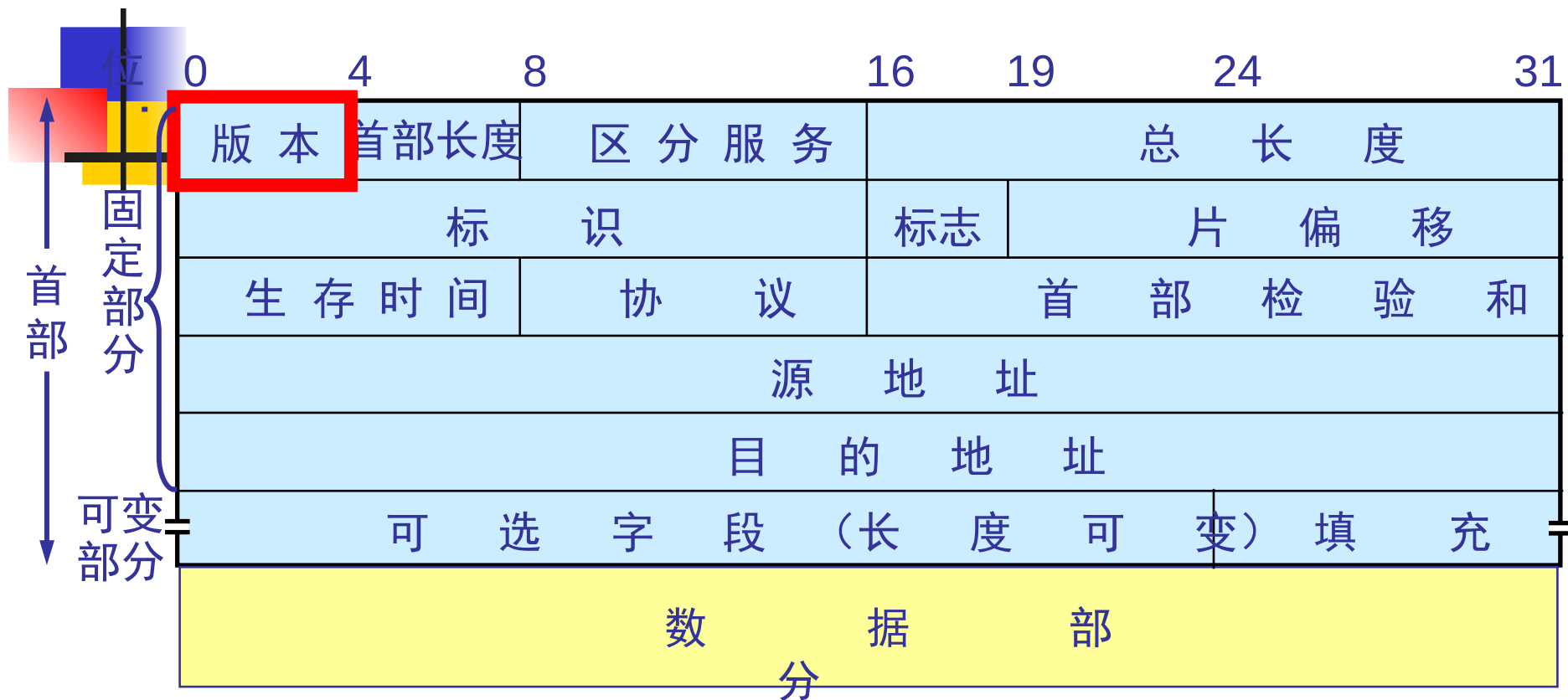








# 1. IP 数据报首部的固定部分中的各字段



版本——占 4 位，指 IP 协议的版本  
目前的 IP 协议版本号为 4 (即 IPv4)



首部长度的——占 4 bit，可表示的最小数值为 5，最大是 15 个单位（以 4 字节为单位，以 32 位字为单位的报头长度）因此 IP 的首部长度的最大值是 60 字节。



区分服务——占 8 位，用来获得更好的服务，在旧标准中叫做服务类型，但实际上一直未被使用过。



# 服务类型

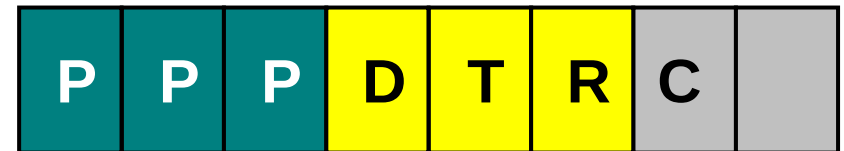
**D(Delay)** : 时延。 0- 常规时延, 1- 低时延。

**T(Throughput)** : 吞吐量。 0- 常规吞吐量, 1- 高吞吐量。

**R(Reliability)** : 可靠性。 0- 常规可靠性, 1- 高可靠性。

**C**: 选择代价最小的路由。

优先级





总长度——占 16 位，指**首部**和**数据**之和的长度，单位为字节，因此数据报的最大长度为 65535 ( $2^{16}$ ) 字节。

总长度必须不超过最大传送单元 MTU 。



标识 (identification) 占 16 位，  
它是一个计数器，用来产生数据报的标识。



标志 (flag) 占 3 位，目前只有前两位有意义。标志字段的最低位是 **MF**。MF = 1 表示后面“还有分片”。MF = 0 表示最后一个分片。标志字段中间的一位是 **DF**。只有当 DF = 0 时才允许分片。



片偏移 (13 位) 指出：较长的分组在分片后，某片在原分组中的相对位置。

片偏移以 8 个字节为偏移单位。  $2^{13}=8192$ ，分组最多 8192 个分段，  $8192*8=65536$



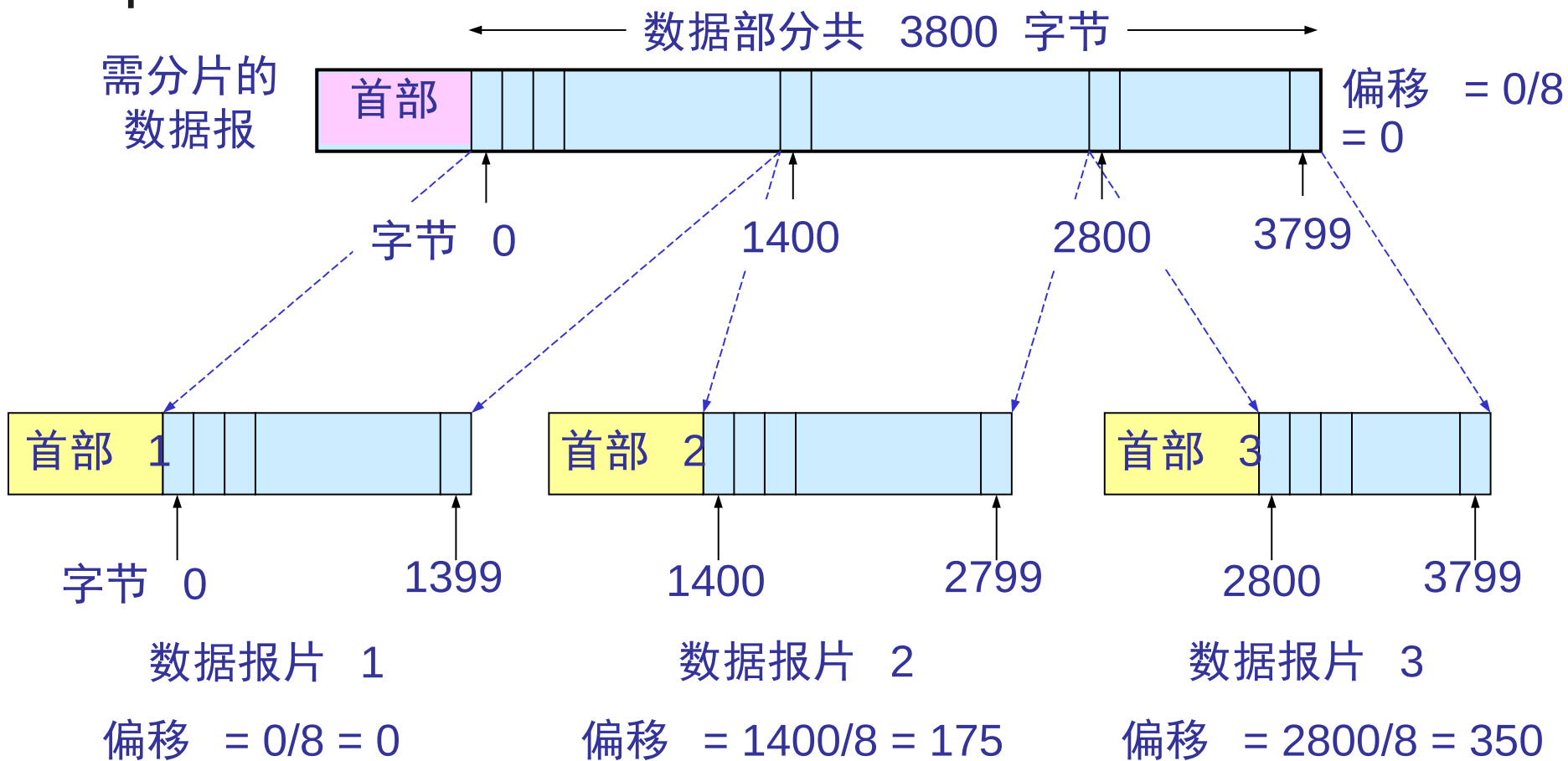


# 最大传输单元与 IP 数据报分片

---

- 每一种物理网络都规定了各自帧的数据域最大字节长度的最大传输单元；
- IP 数据报作为网络层数据必然要通过帧来传输；一个数据报可能要通过多个不同的物理网络；

# 【例 4-1】 IP 数据报分片



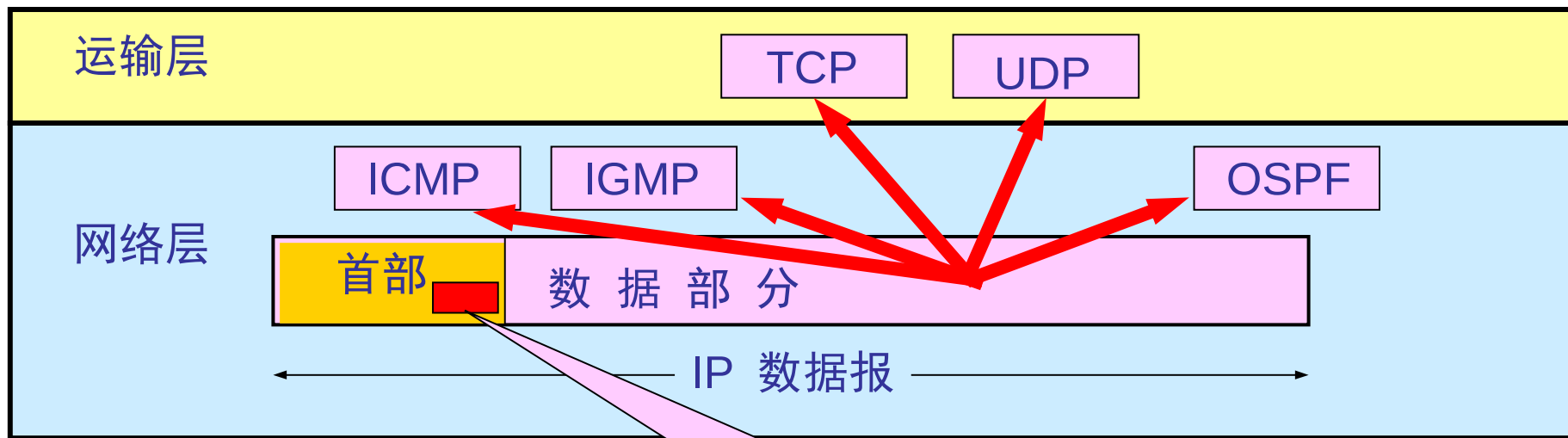
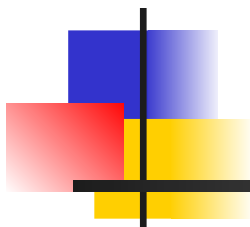


生存时间 (8 位) 记为 TTL (Time To Live)  
数据报在网络中可通过的路由器数的最大值。

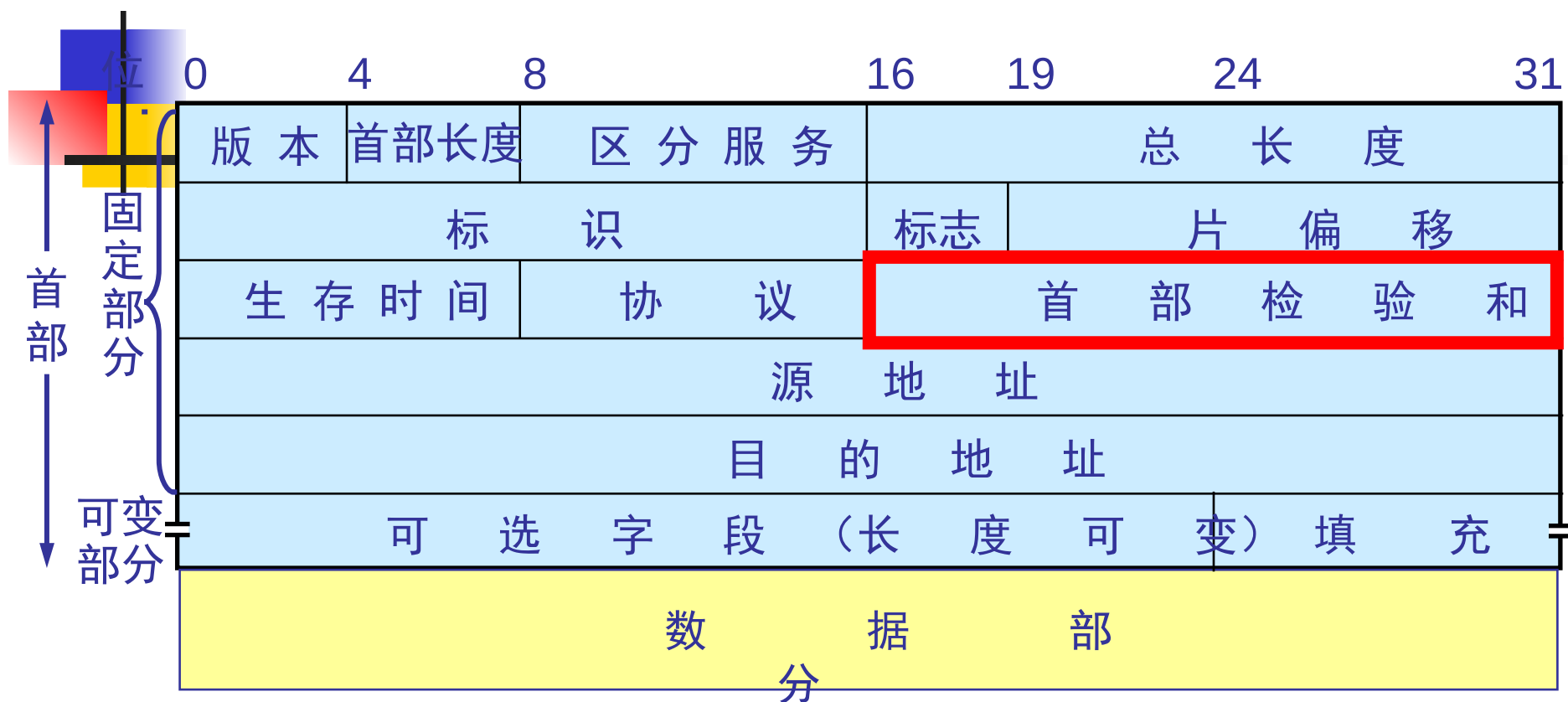


协议 (8 位) 字段指出此数据报携带的数据使用何种协议

以便目的主机的 IP 层将数据部分上交给哪个处理过程。



协议字段指出应将数据部分交给哪一个进程

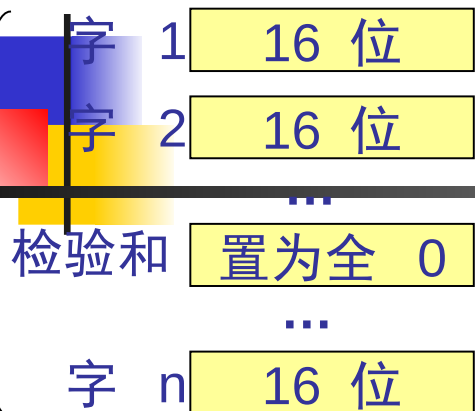


首部检验和 (16 位) 字段只检验数据报的首部，不检验数据部分。这里不采用 CRC 检验码而采用简单的计算方法。

发送端

接收端

数据报首部



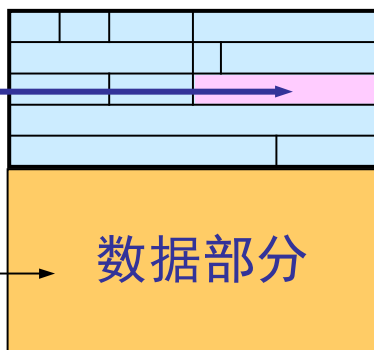
反码算术  
运算求和 16 位

取反码

校验和 16 位

数据部分  
不参与校验和的计算

IP 数据报



反码算术  
运算求和 16 位

取反码

结果 16 位

若结果为 0, 则保留;  
否则, 丢弃该数据报



源地址和目的地址都各占 4 字节



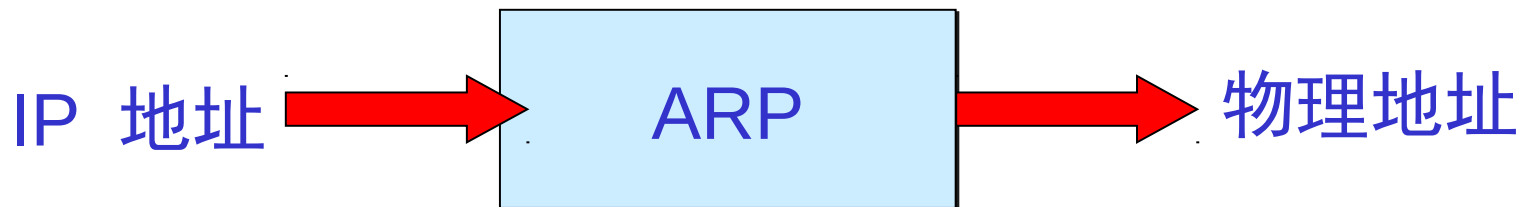


## 2. IP 数据报首部的可变部分

---

- IP 首部的可变部分就是一个选项字段，用来支持排错、测量以及安全等措施，内容很丰富。
- 选项字段的长度可变，从 1 个字节到 40 个字节不等，取决于所选择的项目。
- 增加首部的可变部分是为了增加 IP 数据报的功能，但这同时也使得 IP 数据报的首部长度成为可变的。这就增加了每一个路由器处理数据报的开销。
- 实际上这些选项很少被使用。

## 4.2.4 地址解析协议 ARP 和 逆地址解析协议 RARP



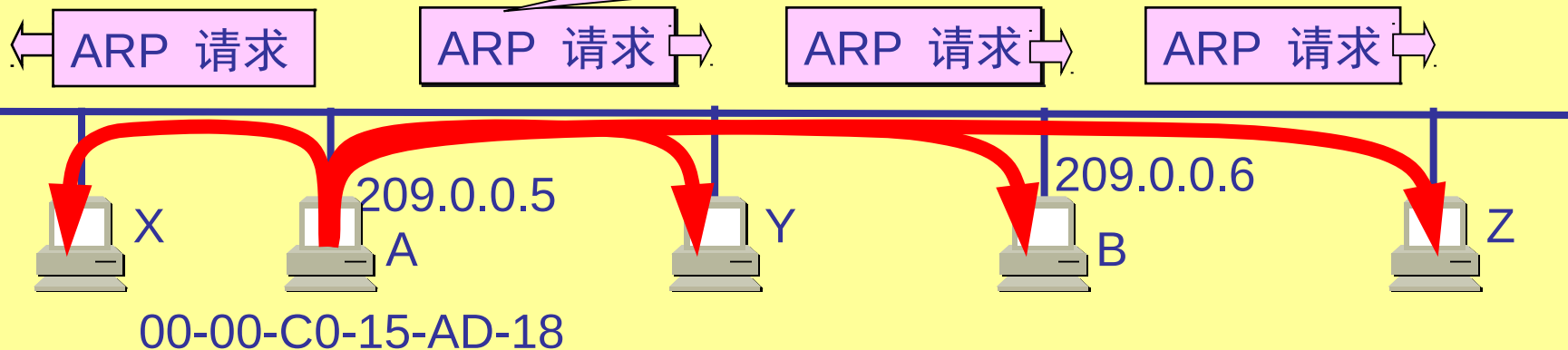


# 地址解析协议 ARP

- 不管网络层使用的是什麼协议，在实际网络的链路上传送数据帧时，最终还是必须使用硬件地址。
- 每一个主机都设有一个 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。
- 当主机 A 欲向本局域网上的某个主机 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。如有，就可查出其对应的硬件地址，再将此硬件地址写入 MAC 帧，然后通过局域网将该 MAC 帧发往此硬件地址。

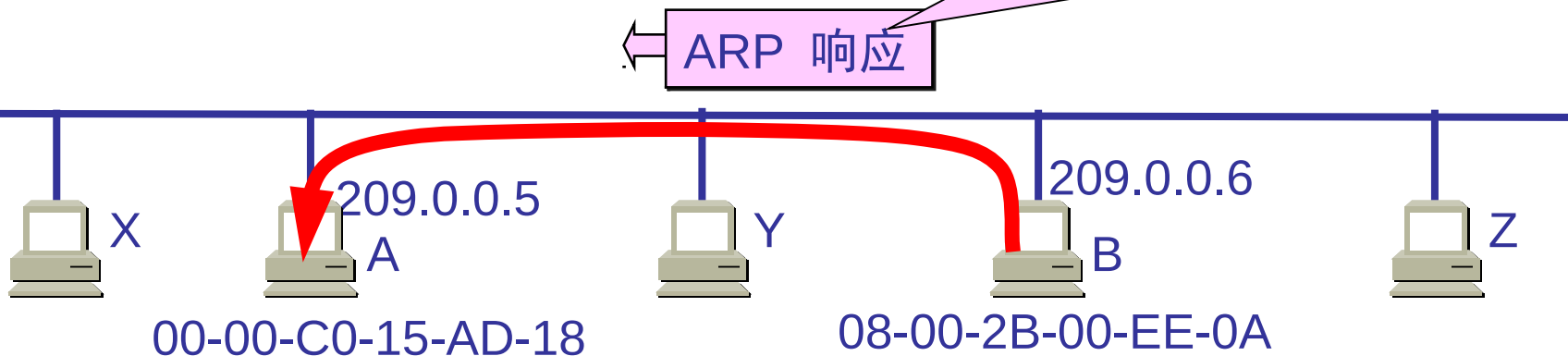
主机 A 广播发送  
ARP 请求分组

我是 209.0.0.5，硬件地址是 00-00-C0-15-AD-18  
我想知道主机 209.0.0.6 的硬件地址



主机 B 向 A 发送  
ARP 响应分组

我是 209.0.0.6  
硬件地址是 08-00-2B-00-EE-0A





# ARP 高速缓存的作用

---

- 为了减少网络上的通信量，主机 A 在发送其 ARP 请求分组时，就将自己的 IP 地址到硬件地址的映射写入 ARP 请求分组。
- 当主机 B 收到 A 的 ARP 请求分组时，就将主机 A 的这一地址映射写入主机 B 自己的 ARP 高速缓存中。这对主机 B 以后向 A 发送数据报时就更方便了。



# 应当注意的问题

---

- 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。
- 如果所要找的主机和源主机不在同一个局域网，那么就要通过 ARP 找到一个位于本局域网上的某个路由器的硬件地址，然后把分组发送给这个路由器，让这个路由器把分组转发给下一个网络。剩下的工作就由下一个网络来做。



## 应当注意的问题（续）

---

- 从 IP 地址到硬件地址的解析是自动进行的，主机的用户对这种地址解析过程是不知道的。
- 只要主机或路由器要和本网络上的另一个已知 IP 地址的主机或路由器进行通信，ARP 协议就会自动地将该 IP 地址解析为链路层所需要的硬件地址。



# 使用 ARP 的四种典型情况

- 发送方是主机，要把 IP 数据报发送到**本网络**上的另一个主机。这时用 ARP 找到目的主机的硬件地址。
- 发送方是主机，要把 IP 数据报发送到**另一个网络**上的一个主机。这时用 ARP 找到本网络上的一个路由器的硬件地址。剩下的工作由这个路由器来完成。
- 发送方是路由器，要把 IP 数据报转发到**本网络**上的一个主机。这时用 ARP 找到目的主机的硬件地址。
- 发送方是路由器，要把 IP 数据报转发到**另一个网络**上的一个主机。这时用 ARP 找到本网络上的一个路由器的硬件地址。剩下的工作由这个路由器来完成。





# 为什么我们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**非常复杂的硬件地址转换工作**，因此几乎是不可能的事。
- 连接到因特网的主机都拥有统一的 IP 地址，它们之间的通信就像连接在同一个网络上那样简单方便，因为调用 ARP 来寻找某个路由器或主机的硬件地址都是由**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。



## 注意：

---

- 1)ARP 分层的位置是 TCP/IP 的网络层
- 2)ARP 报文是由以太网帧进行封装传输的。没有封装进 IP 包。
- 3 ) 实际上，对网络接口层的以太网帧来讲，它们同样是帧的上层协议，当收到以太帧时，根据帧的协议字段判断是送到 ARP 还是 IP 。
- 4 ) 之所以不把它放在数据链路层，是因为它并不具备数据链路层的功能，它的作用是为数据链路层提供接收方的 MAC 地址。



# 逆地址解析协议 RARP

---

- 逆地址解析协议 RARP 使只知道自己硬件地址的主机能够知道其 IP 地址。
- 这种主机往往是无盘工作站。因此 RARP 协议目前已很少使用。

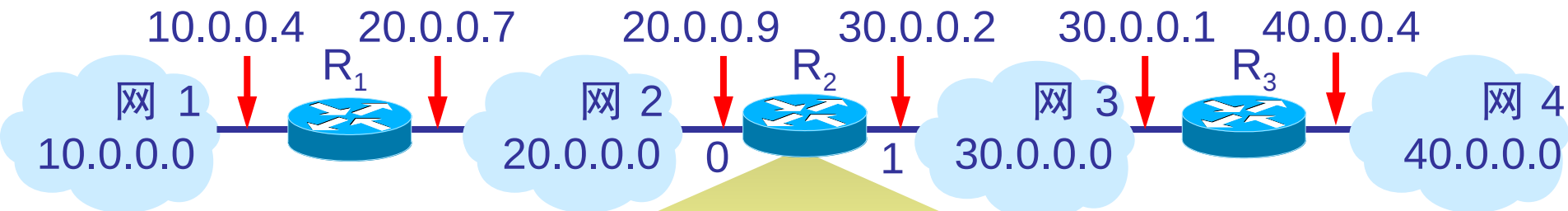


## 4.2.6 IP 层转发分组的流程

---

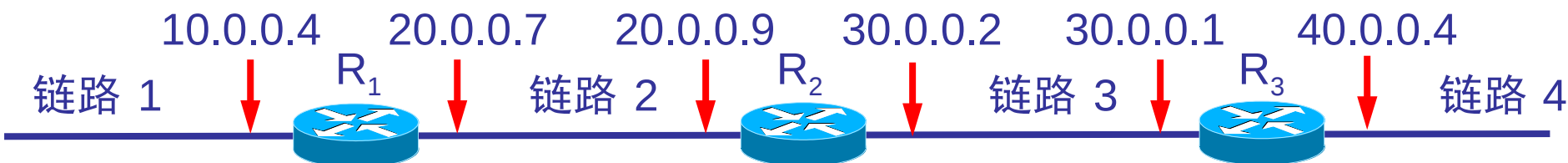
- 有四个 A 类网络通过三个路由器连接在一起。每一个网络上都可能有成千上万个主机。
- 可以想像，若按目的主机号来制作路由表，则所得出的路由表就会过于庞大。
- 但若按主机所在的网络地址来制作路由表，那么每一个路由器中的路由表就只包含 4 个项目。这样就可使路由表大大简化。

在路由表中，对每一条路由，最主要的是  
(目的网络地址，下一跳地址)



路由器 R<sub>2</sub> 的路由表

目的主机所在的网络	下一跳地址
20.0.0.0	直接交付, 接口 0
30.0.0.0	直接交付, 接口 1
10.0.0.0	20.0.0.7
40.0.0.0	30.0.0.1





# 查找路由表

---

根据目的网络地址就能确定下一跳路由器，  
这样做的结果是：

- IP 数据报最终一定可以找到目的主机所在目的网络上的路由器（可能要通过多次的间接交付）。
- 只有到达最后一个路由器时，才试图向目的主机进行直接交付。



# 特定主机路由

---

- 这种路由是为特定的目的主机指明一个路由。
- 采用特定主机路由可使网络管理人员能更方便地控制网络 and 测试网络，同时也可在需要考虑某种安全问题时采用这种特定主机路由。



## 默认路由 (default route)

---

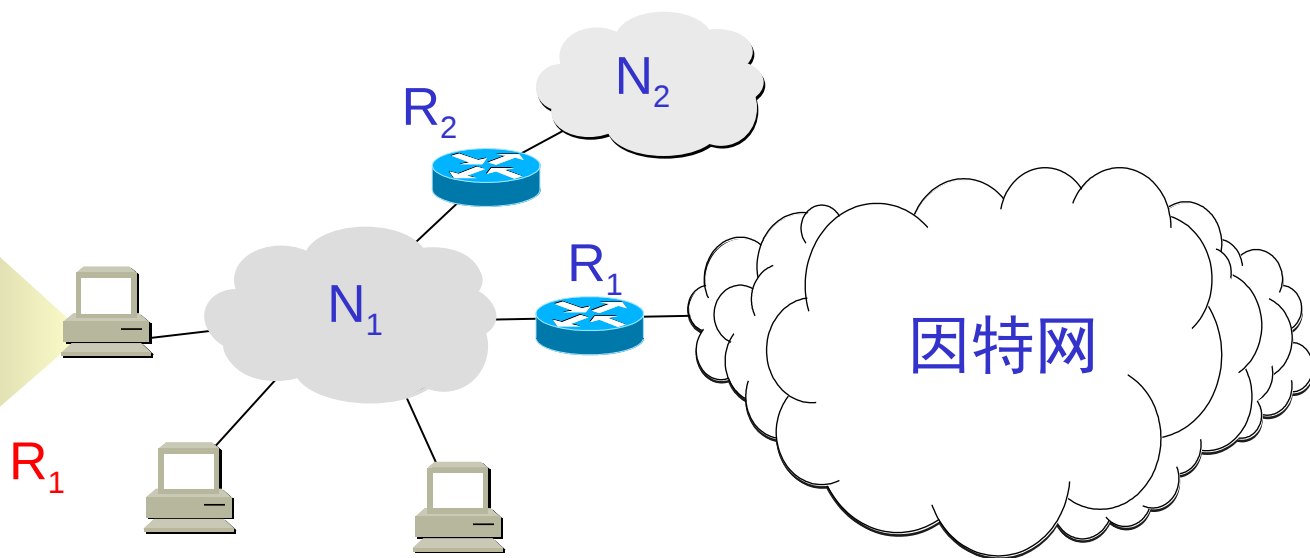
- 路由器还可采用**默认路由**以减少路由表所占用的空间和搜索路由表所用的时间。
- 这种转发方式在一个网络只有很少的对外连接时是很有用的。
- 默认路由在主机发送 IP 数据报时往往更能显示出它的好处。
- 如果一个主机连接在一个小网络上，而这个网络只用一个路由器和因特网连接，那么在这种情况下使用默认路由是非常合适的。



只要目的网络不是  $N_1$  和  $N_2$ ，  
就一律选择默认路由，  
把数据报先间接交付路由器  $R_1$ ，  
让  $R_1$  再转发给下一个路由器。

路由表

目的网络	下一跳
$N_1$	直接
$N_2$	$R_2$
默认	





# 必须强调指出

---

- IP 数据报的首部中没有地方可以用来指明“下一跳路由器的 IP 地址”。
- 当路由器收到待转发的数据报，不是将下一跳路由器的 IP 地址填入 IP 数据报，而是送交下层的网络接口软件。
- 网络接口软件使用 ARP 负责将下一跳路由器的 IP 地址转换成硬件地址，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。

# 分组转发算法

- (1) 从数据报的首部提取目的主机的 IP 地址  $D$ ，得出目的网络地址为  $N$ 。
- (2) 若网络  $N$  与此路由器直接相连，则把数据报直接交付目的主机  $D$ ；否则是间接交付，执行 (3)。
- (3) 若路由表中有目的地址为  $D$  的特定主机路由，则把数据报传送给路由表中所指明的下一跳路由器；否则，执行 (4)。
- (4) 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指明的下一跳路由器；否则，执行 (5)。
- (5) 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；否则，执行 (6)。
- (6) 报告转发分组出错。



## 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

- 在 ARPANET 的早期，IP 地址的设计确实不够合理。
  - IP 地址空间的利用率有时很低。
  - 给每一个物理网络分配一个网络号会使路由表变得太大因而使网络性能变坏。
  - 两级的 IP 地址不够灵活。



## 三级的 IP 地址

---

- 从 1985 年起在 IP 地址中又增加了一个“子网号字段”，使两级的 IP 地址变为三级的 IP 地址。
- 这种做法叫作划分子网 (subnetting)。划分子网已成为因特网的正式标准协议。



# 划分子网的基本思路

- 划分子网纯属一个单位内部的事情。单位对外仍然表现为没有划分子网的网络。
- 从主机号借用若干个位作为子网号 subnet-id，而主机号 host-id 也就相应减少了若干个位。

IP 地址 ::= {< 网络号 >, < 子网号 >, < 主机号 >}

(4-2)

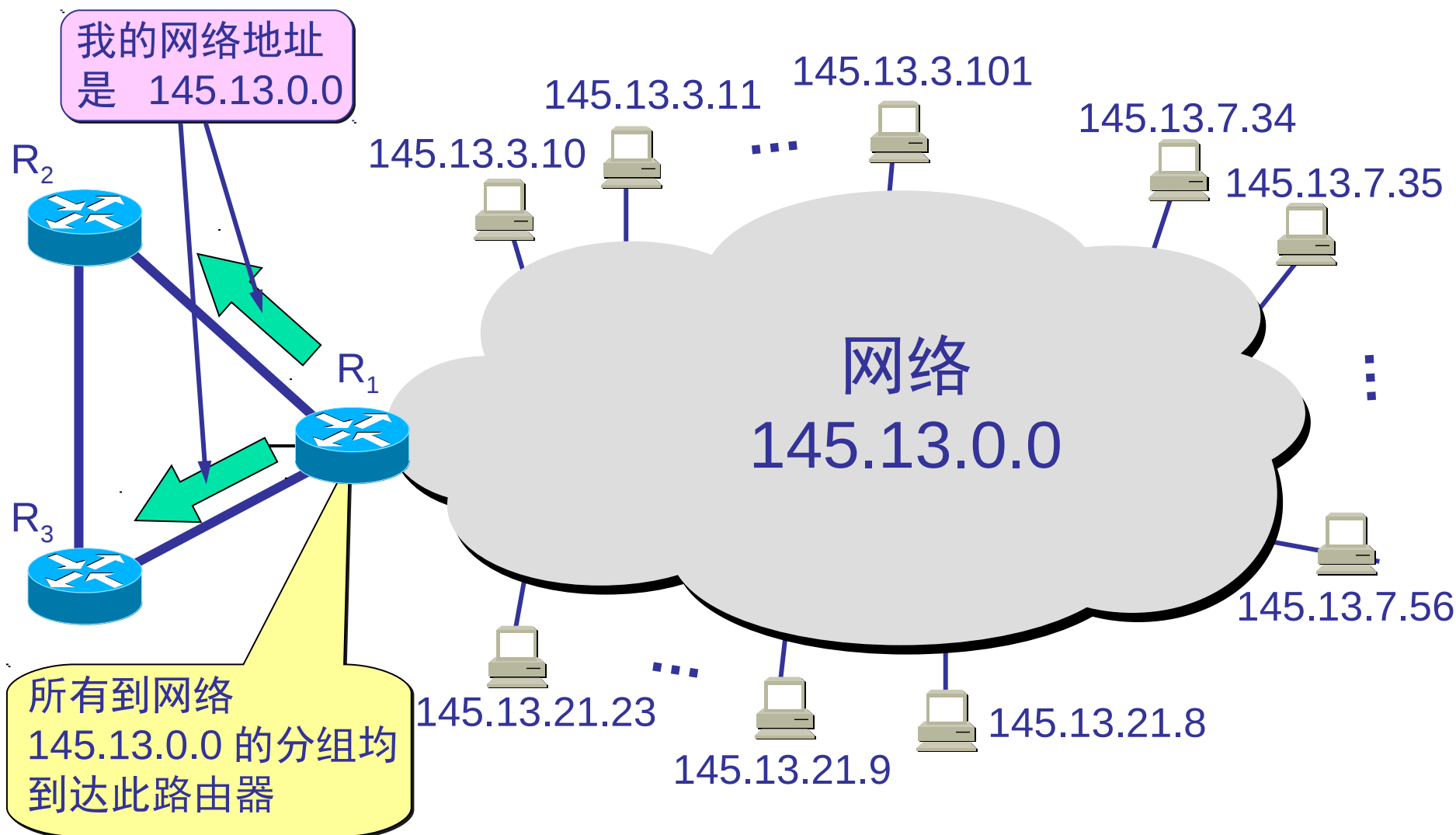


## 划分子网的基本思路（续）

---

- 凡是从其他网络发送给本单位某个主机的 IP 数据报，仍然是根据 IP 数据报的**目的网络号** net-id，先找到连接在**本单位网络上的路由器**。
- 然后**此路由器**在收到 IP 数据报后，再按目的网络号 net-id 和子网号 subnet-id 找到目的子网。
- 最后就将 IP 数据报直接交付目的主机。

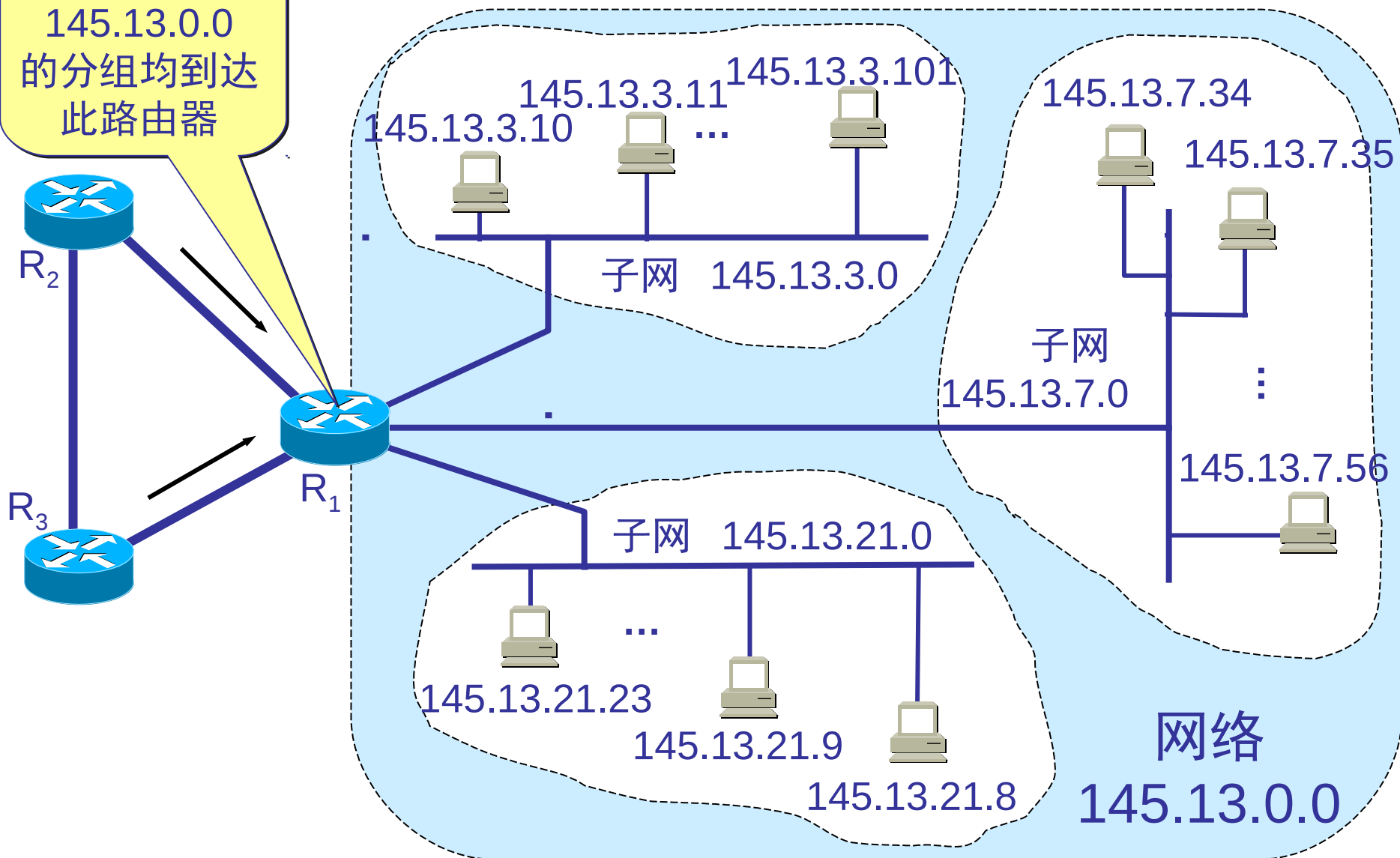
# 一个未划分子网的 B 类网络 145.13.0.0





# 划分为三个子网后对外仍是一个网络

所有到达网络  
145.13.0.0  
的分组均到达  
此路由器





# 划分子网后变成了三级结构

---

- 当没有划分子网时，IP 地址是**两级结构**。
- 划分子网后 IP 地址就变成了**三级结构**。
- 划分子网只是把 IP 地址的主机号 host-id 这部分进行再划分，而不改变 IP 地址原来的网络号 net-id 。

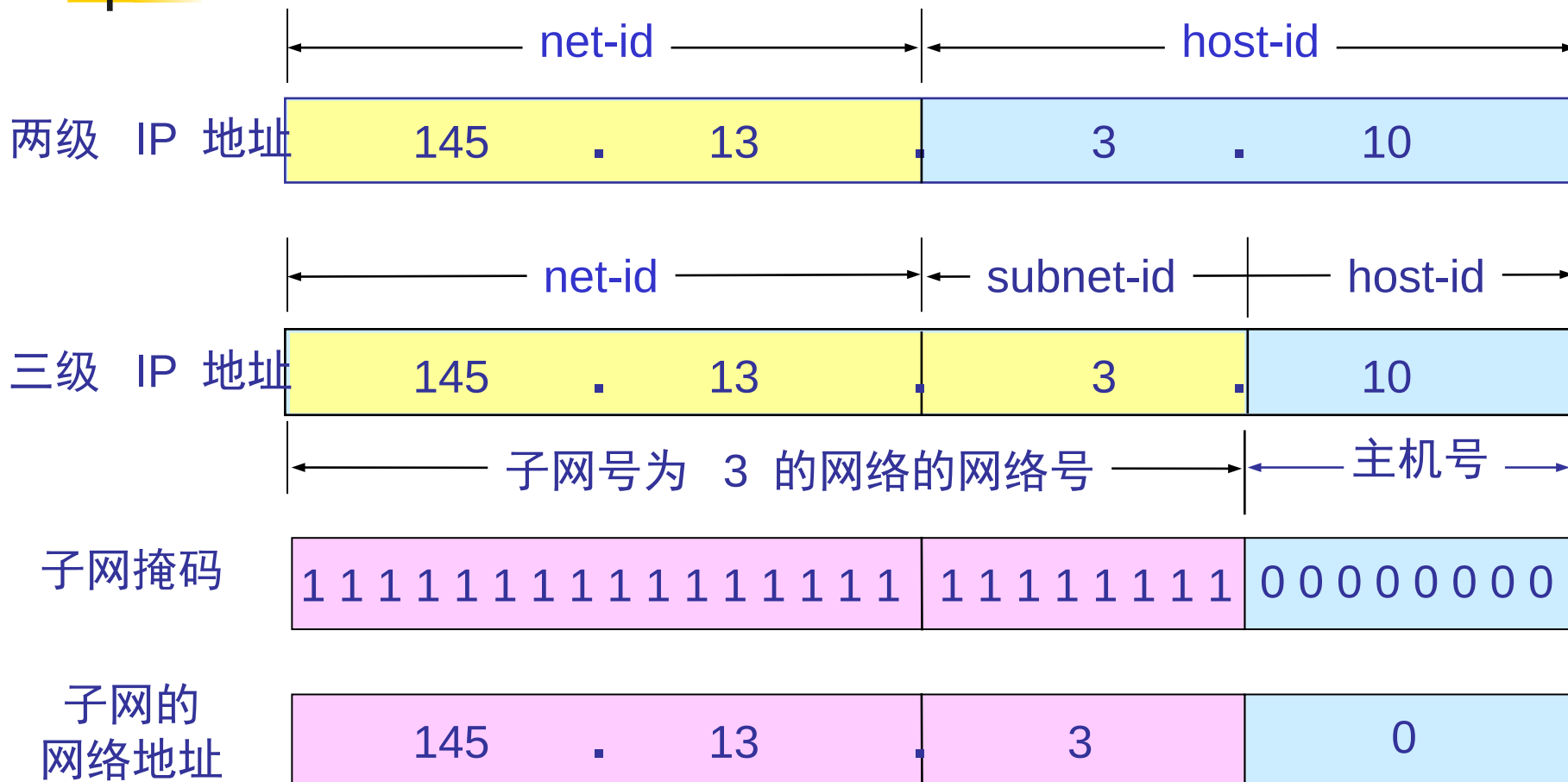


## 2. 子网掩码

---

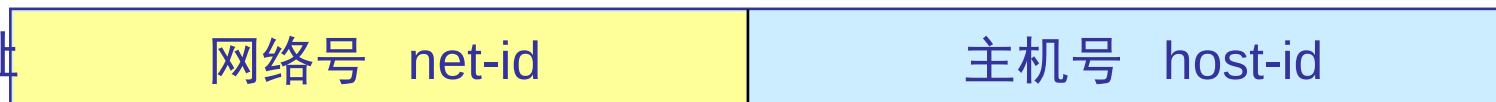
- 从一个 IP 数据报的首部并**无法判断**源主机或目的主机所连接的**网络**是否进行了子网划分。
- 使用**子网掩码** (subnet mask) 可以找出 IP 地址中的子网部分。

# IP 地址的各字段和子网掩码

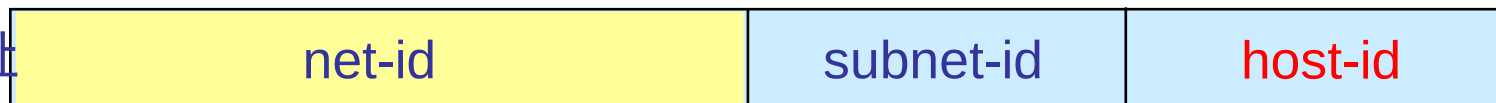


# (IP 地址) AND (子网掩码) = 网络地址

两级 IP 地址



三级 IP 地址

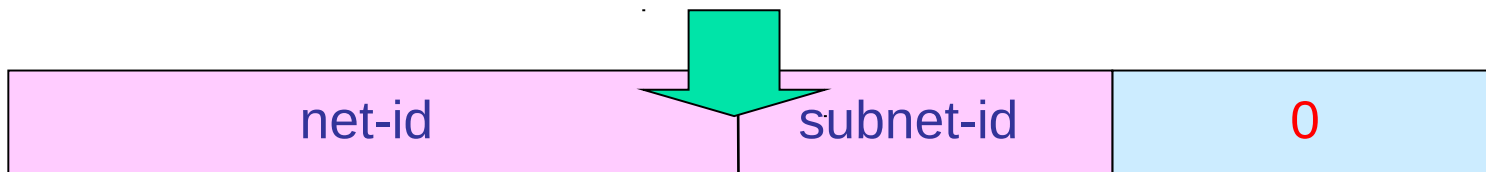


逐位进行 AND 运算

子网掩码



子网的  
网络地址







# 子网掩码是一个重要属性

---

- 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。
- 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。

**【例 4-2】已知 IP 地址是 141.14.72.24，子网掩码是 255.255.192.0。试求网络地址。**

(a) 点分十进制表示的 IP 地址

141	.	14	.	72	.	24
-----	---	----	---	----	---	----

(b) IP 地址的第 3 字节是二进制

141	.	14	.	01001000	.	24
-----	---	----	---	----------	---	----

(c) 子网掩码是 255.255.192.0

11111111	11111111	11000000	00000000
----------	----------	----------	----------

(d) IP 地址与子网掩码逐位相与

141	.	14	.	01000000	.	0
-----	---	----	---	----------	---	---

(e) 网络地址（点分十进制表示）

141	.	14	.	64	.	0
-----	---	----	---	----	---	---



**【例 4-3】** 在上例中，若子网掩码改为 255.255.224.0。试求网络地址，讨论所得结果。

(a) 点分十进制表示的 IP 地址

141	.	14	.	72	.	24
-----	---	----	---	----	---	----

(b) IP 地址的第 3 字节是二进制

141	.	14	.	01001000	.	24
-----	---	----	---	----------	---	----

(c) 子网掩码是 255.255.224.0

11111111	11111111	11100000	00000000
----------	----------	----------	----------

(d) IP 地址与子网掩码逐位相与

141	.	14	.	01000000	.	0
-----	---	----	---	----------	---	---

(e) 网络地址（点分十进制表示）

141	.	14	.	64	.	0
-----	---	----	---	----	---	---

不同的子网掩码得出**相同**的网络地址。  
但不同的掩码的效果是不同的。



## 4.3.2 使用子网掩码的分组转发过程

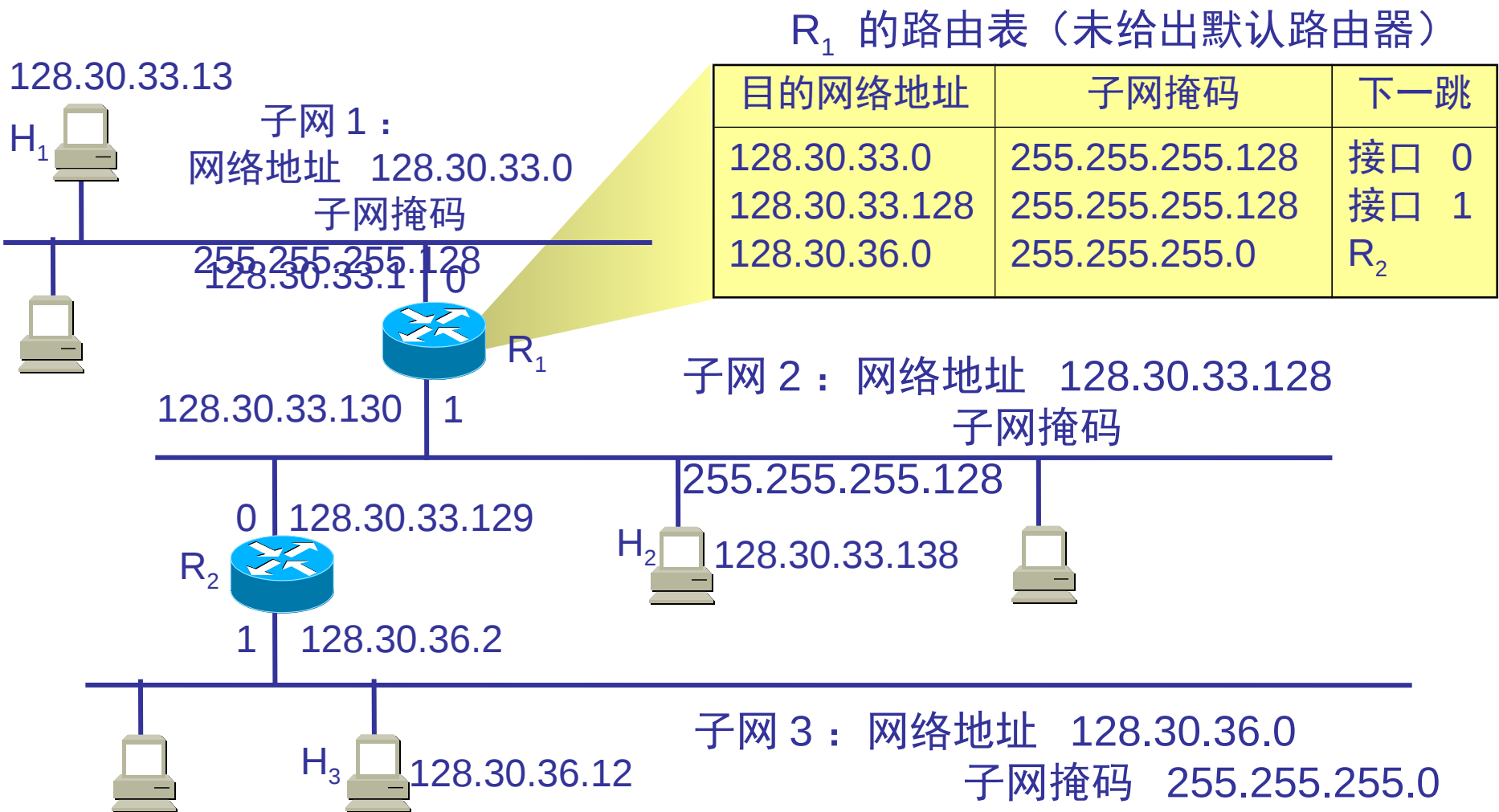
---

- 在不划分子网的两级 IP 地址下，从 IP 地址得出网络地址是个很简单的事。
- 但在划分子网的情况下，从 IP 地址却不能唯一地得出网络地址来，这是因为网络地址取决于那个网络所采用的子网掩码，但数据报的首部并没有提供子网掩码的信息。
- 因此分组转发的算法也必须做相应的改动。

# 在划分子网的情况下路由器转发分组的算法

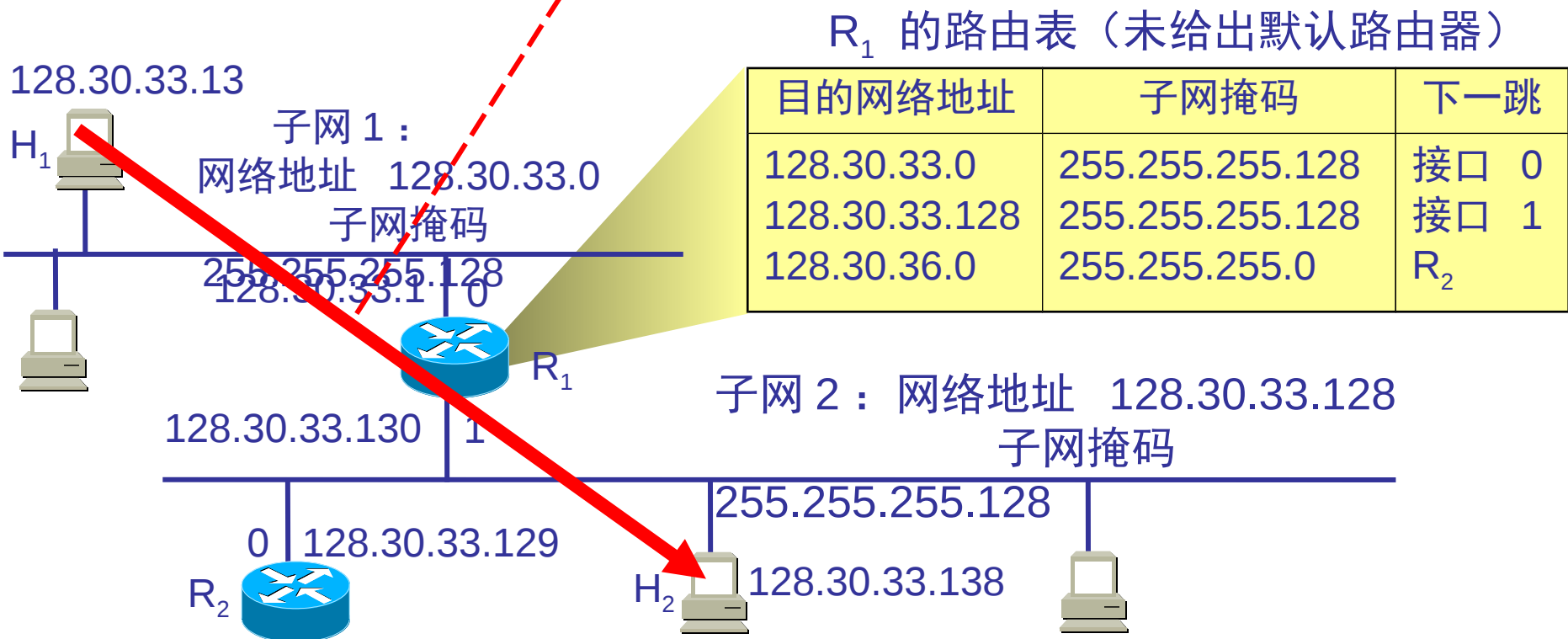
- (1) 从收到的分组的首部提取目的 IP 地址  $D$ 。
- (2) 先用各网络的子网掩码和  $D$  逐位相“与”，看是否和相应的网络地址匹配。若匹配，则将分组直接交付。  
否则就是间接交付，执行 (3)。
- (3) 若路由表中有目的地址为  $D$  的特定主机路由，则将分组传送给指明的下一跳路由器；否则，执行 (4)。
- (4) 对路由表中的每一行的子网掩码和  $D$  逐位相“与”，  
若其结果与该行的目的网络地址匹配，则将分组传送给该行指明的下一跳路由器；否则，执行 (5)。

**【例 4-4】** 已知互联网和路由器  $R_1$  中的路由表。主机  $H_1$  向  $H_2$  发送分组。试讨论  $R_1$  收到  $H_1$  向  $H_2$  发送的分组后查找路由表的过程。



# 主机 $H_1$ 要发送分组给 $H_2$

要发送的分组的目的地 IP 地址：128.30.33.138



因此  $H_1$  首先检查主机 128.30.33.138 是否连接在本网络上

如果是，则直接交付；

否则，就送交路由器  $R_1$ ，并逐项查找路由表。

主机  $H_1$  首先将  
本子网的子网掩码 255.255.255.128  
与分组的 IP 地址 128.30.33.138 逐比特相“与” (AND)

255.255.255.128 AND 128.30.33.138 的计算

255 就是二进制的全 1，因此 255 AND xyz = xyz，  
这里只需计算最后的 128 AND 138 即可。

128 → 10000000

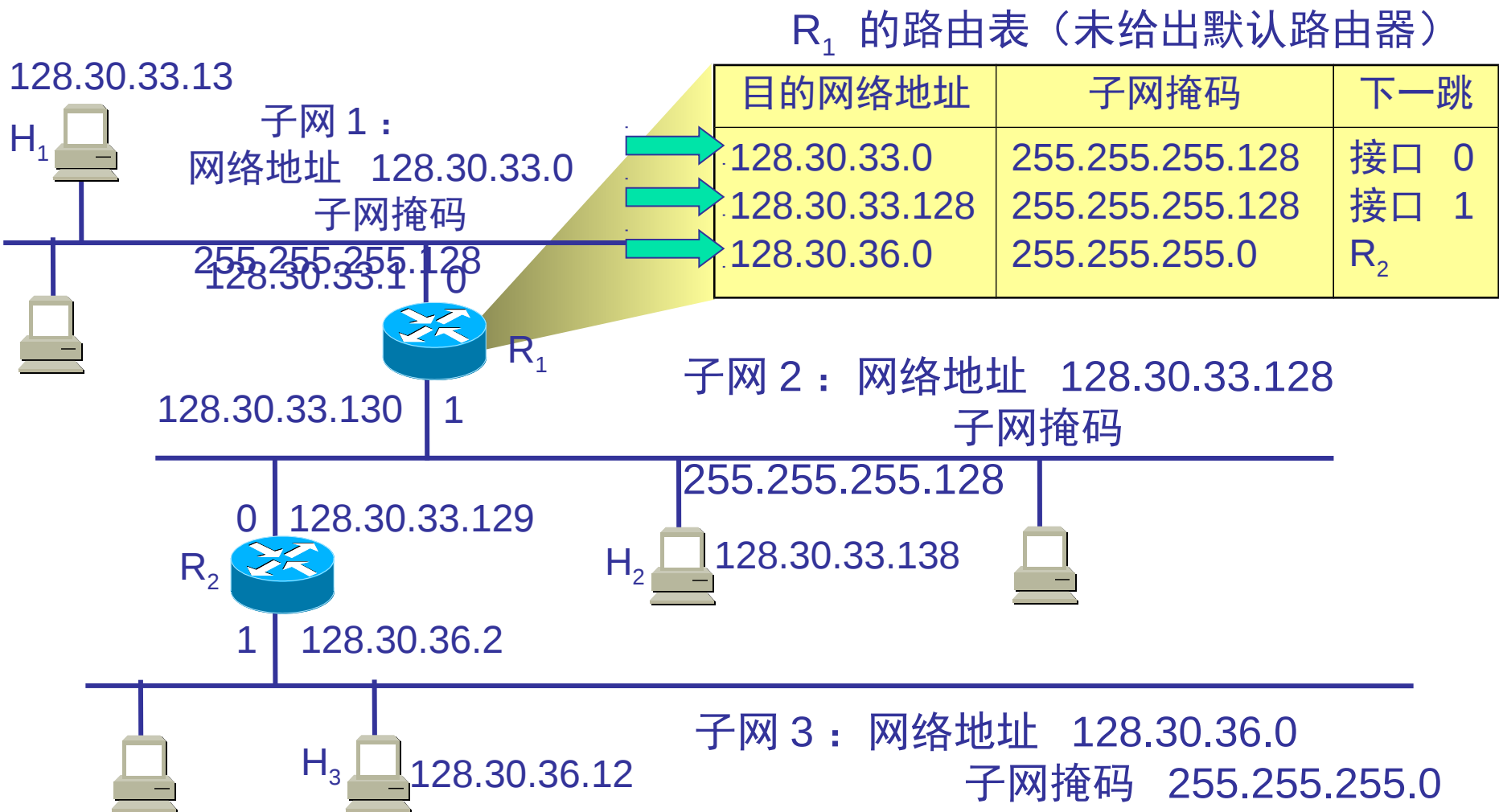
138 → 10001010

逐比特 AND 操作后: 10000000 → 128

255.255.255.128
128. 30. 33.138
128. 30. 33.128

128. 30. 33.128  $\neq H_1$  的网络地址

因此  $H_1$  必须把分组传送到路由器  $R_1$   
然后逐项查找路由表

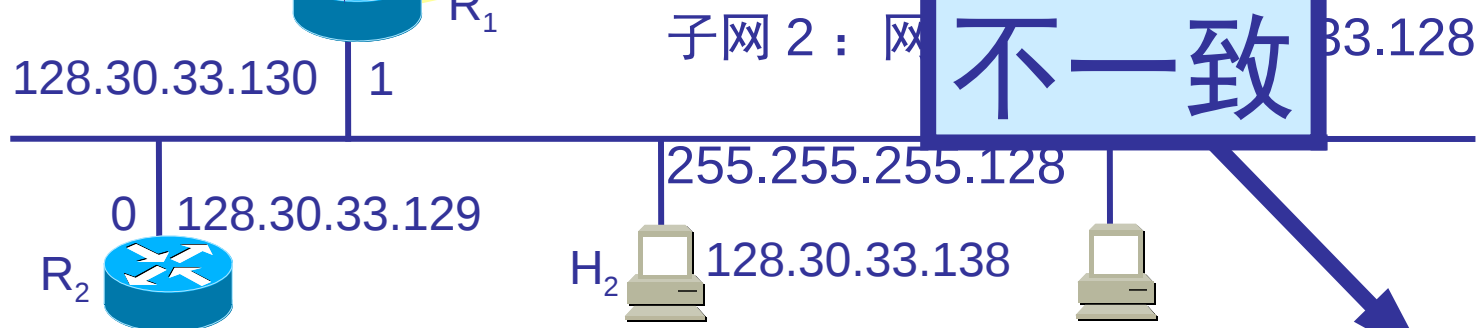
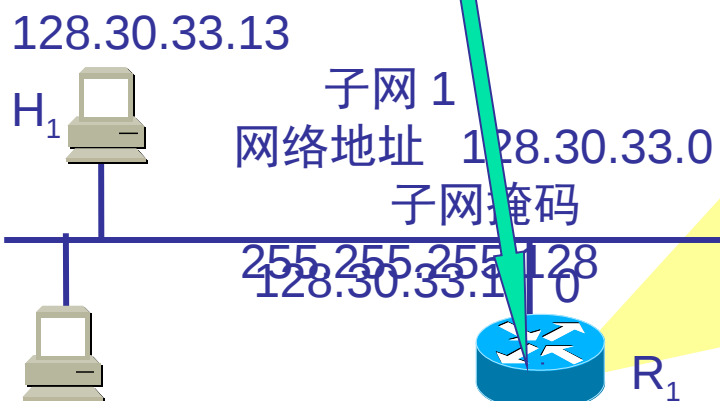


路由器  $R_1$  收到分组后就用路由表中第 1 个项目的子网掩码和 128.30.33.138 逐比特 **AND** 操作

$R_1$  收到的分组的目的 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	$R_2$



不一致

255.255.255.128 **AND** 128.30.33.138 = 128.30.33.128

不匹配!

(因为 128.30.33.128 与路由表中的 128.30.33.0 不一致)

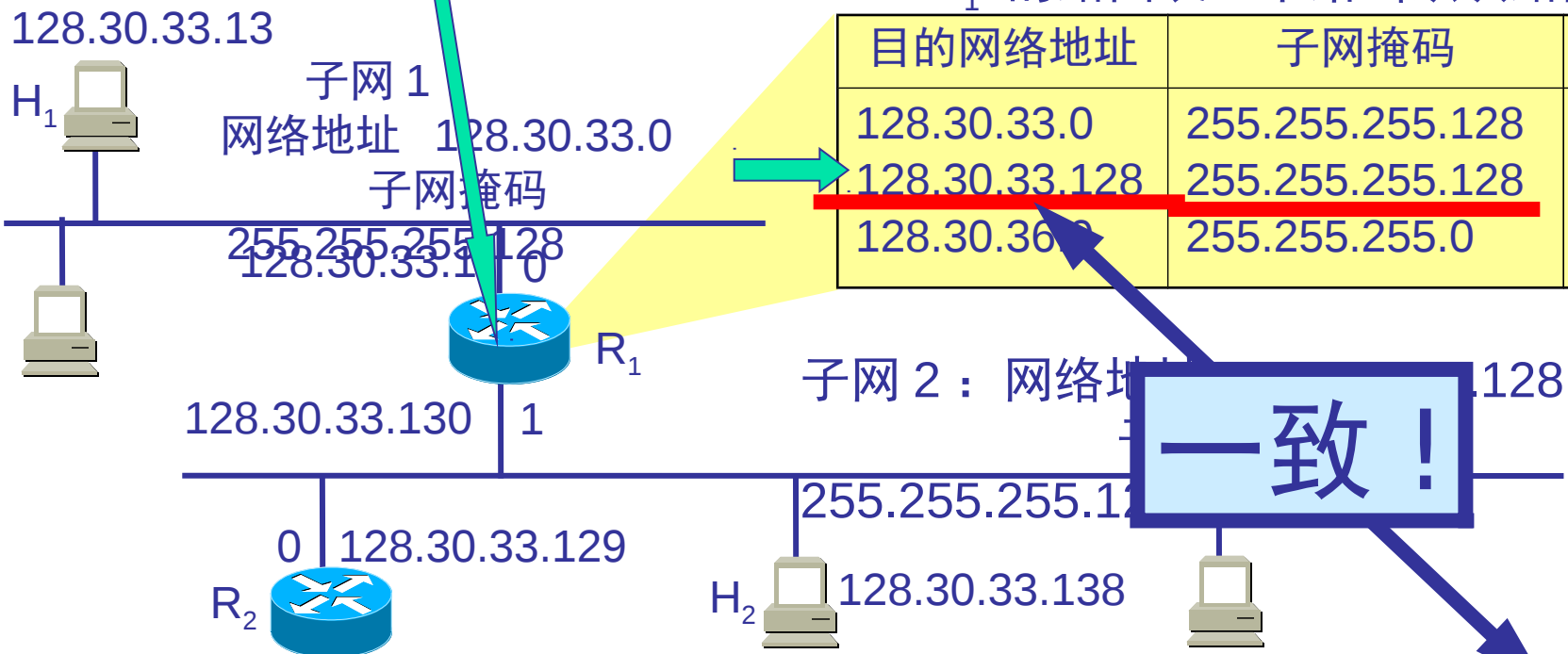


路由器  $R_1$  再用路由表中第 2 个项目的  
子网掩码和 128.30.33.138 逐比特 **AND** 操作

$R_1$  收到的分组的目的 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	$R_2$



255.255.255.128 **AND** 128.30.33.138 = 128.30.33.128

**匹配!**

这表明子网 2 就是收到的分组所要寻找的目的网络

## 4.3.3 无分类编址 CIDR

### 1. 网络前缀

划分子网在一定程度上缓解了因特网在发展中遇到的困难。然而在 1992 年因特网仍然面临三个必须尽早解决的问题，这就是：

- B 类地址在 1992 年已分配了近一半，眼看就要在 1994 年 3 月全部分配完毕！
- 因特网主干网上的路由表中的项目数急剧增长（从几千个增长到几万个）。
- 整个 IPv4 的地址空间最终将全部耗尽。



# IP 编址问题的演进

---

- 1987 年，RFC 1009 就指明了在一个划分子网的网络中可同时使用几个不同的子网掩码。使用**变长子网掩码 VLSM** (Variable Length Subnet Mask) 可进一步提高 IP 地址资源的利用率。
- 在 VLSM 的基础上又进一步研究出无分类编址方法，它的正式名字是**无分类域间路由选择 CIDR** (Classless Inter-Domain Routing)。



# CIDR 最主要的特点

---

- 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，因而可以更加有效地分配 IPv4 的地址空间。
- CIDR 使用各种长度的“网络前缀” (network-prefix) 来代替分类地址中的网络号和子网号。
- IP 地址从三级编址（使用子网掩码）又回到了两级编址。



# 无分类的两级编址

- 无分类的两级编址的记法是：

IP 地址 ::= {< 网络前缀 >, < 主机号 >} (4-3)

- CIDR 还使用“斜线记法” (slash notation)，它又称为 CIDR 记法，即在 IP 地址后加上一个斜线“/”，然后写上网络前缀所占的位数（这个数值对应于三级编址中子网掩码中 1 的个数）。
- CIDR 把网络前缀都相同的连续的 IP 地址组成“CIDR 地址块”。



# CIDR 地址块

---

- 128.14.32.0/20 表示的地址块共有  $2^{12}$  个地址（因为斜线后面的 20 是网络前缀的位数，所以这个地址的主机号是 12 位）。
- 这个地址块的起始地址是 128.14.32.0。
- 在不需要指出地址块的起始地址时，也可将这样的地址块简称为“/20 地址块”。
- 128.14.32.0/20 地址块的最小地址：128.14.32.0
- 128.14.32.0/20 地址块的最大地址：128.14.47.255
- 全 0 和全 1 的主机号地址一般不使用。

128.14.32.0/20 表示的地址 (  $2^{12}$  个地址 )

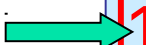
最小地址



10000000	00001110	00100000	00000000
10000000	00001110	00100000	00000001
10000000	00001110	00100000	00000010
10000000	00001110	00100000	00000011
10000000	00001110	00100000	00000100
10000000	00001110	00100000	00000101
...			

所有地址  
的 20 位  
前缀都是  
一样的

最大地址



10000000	00001110	00101111	11111011
10000000	00001110	00101111	11111100
10000000	00001110	00101111	11111101
10000000	00001110	00101111	11111110
10000000	00001110	00101111	11111111



# 路由聚合 (route aggregation)

- 一个 CIDR 地址块可以表示很多地址，这种地址的聚合常称为**路由聚合**，它使得路由表中的一个项目可以表示很多个（例如上千个）原来传统分类地址的路由。
- 路由聚合也称为**构成超网** (supernetting)。
- CIDR 虽然不使用子网了，但仍然使用“**掩码**”这一名词（但不叫子网掩码）。
- 对于 /20 地址块，它的掩码是 20 个连续的 1。  
斜线记法中的数字就是掩码中 1 的个数。





# CIDR 记法的其他形式

- 10.0.0.0/10 可简写为 10/10，也就是把点分十进制中低位连续的 0 省略。
- 10.0.0.0/10 隐含地指出 IP 地址 10.0.0.0 的掩码是 255.192.0.0。此掩码可表示为

11111111 11000000 00000000 00000000  
└───┘ └───┘ └───┘ └───┘  
255 192 0 0

掩码中有 10 个连续的 1



# CIDR 记法的其他形式

---

- 网络前缀的后面加一个星号 \* 的表示方法  
如 00001010 00\*，在星号 \* 之前是网络前缀，而星号 \* 表示 IP 地址中的主机号，可以是任意值。
- 前缀长度不超过 23 位的 CIDR 地址块都包含了多个 C 类地址。
- 这些 C 类地址合起来就构成了超网。

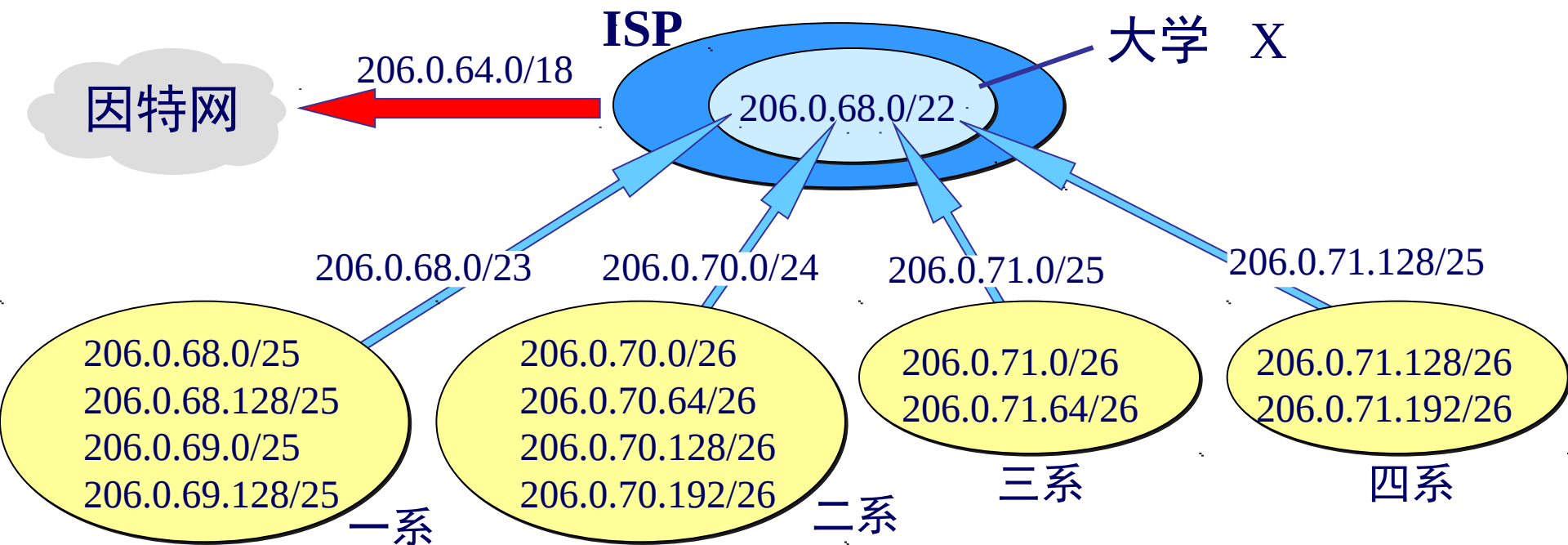


# 构成超网

---

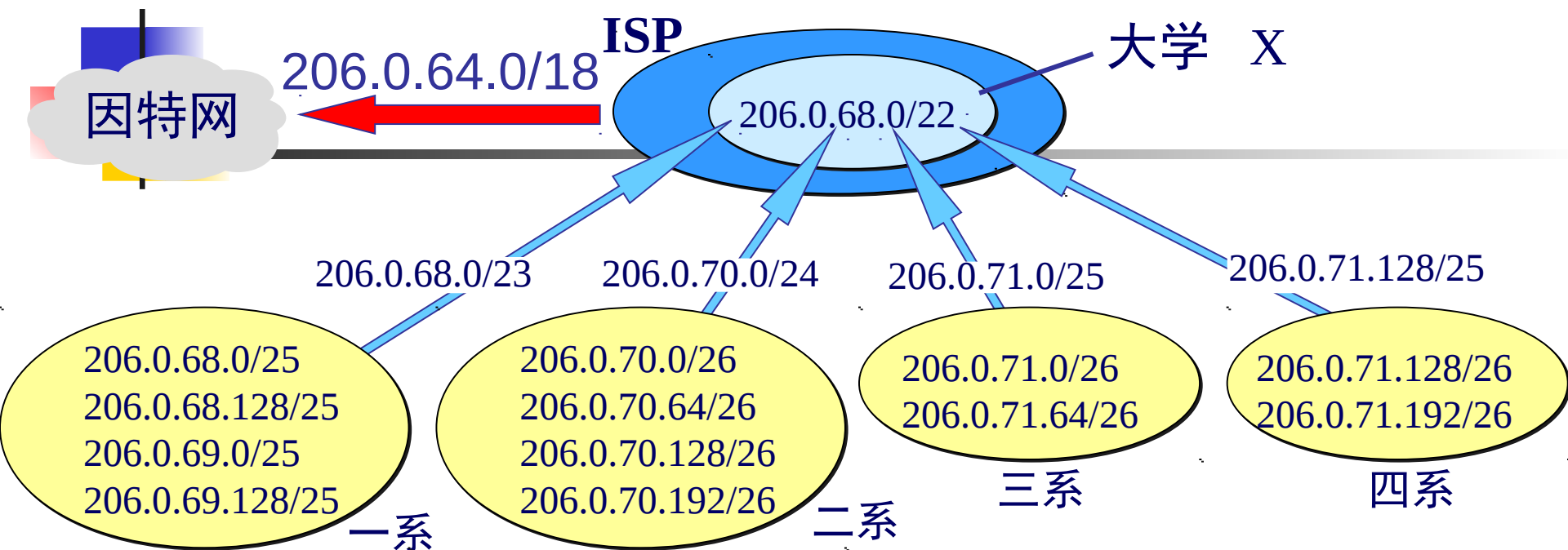
- CIDR 地址块中的地址数一定是 2 的整数次幂。
- 网络前缀越短，其地址块所包含的地址数就越多。而在三级结构的 IP 地址中，划分子网是使网络前缀变长。

# CIDR 地址块划分举例



单位	地址块	二进制表示	
ISP	206.0.64.0/18	地址数	
大学	206.0.68.0/22	11001110.00000000.010001*	16384
一系	206.0.68.0/23	11001110.00000000.0100010*	1024
二系	206.0.70.0/24	11001110.00000000.01000110.*	512
三系	206.0.71.0/25	11001110.00000000.01000111.0*	256
四系	206.0.71.128/25	11001110.00000000.01000111.1*	128

# CIDR 地址块划分举例



这个 ISP 共有 64 个 C 类网络。如果不采用 CIDR 技术，则在与该 ISP 的路由器交换路由信息的每一个路由器的路由表中，就需要有 64 个项目。但采用地址聚合后，只需用路由聚合后的 1 个项目  $206.0.64.0/18$  就能找到该 ISP。



## 2. 最长前缀匹配

- 使用 CIDR 时，路由表中的每个项目由“网络前缀”和“下一跳地址”组成。在查找路由表时可能会得到不止一个匹配结果。
- 应当从匹配结果中选择具有最长网络前缀的路由：最长前缀匹配 (longest-prefix matching)。
- 网络前缀越长，其地址块就越小，因而路由就越具体 (more specific)。
- 最长前缀匹配又称为最长匹配或最佳匹配。

# 最长前缀匹配举例

收到的分组的地址  $D = 206.0.71.128$

路由表中的项目:  $206.0.68.0/22$  (ISP)

$206.0.71.128/25$  (四系)

查找路由表中的第 1 个项目

第 1 个项目  $206.0.68.0/22$  的掩码  $M$  有 22 个连续的 1

$M = 11111111\ 11111111\ 11111100\ 00000000$

因此只需把  $D$  的第 3 个字节转换成二进制。

$M = 11111111\ 11111111\ 11111100\ 00000000$

AND	$D =$	206.	0.	01000100.	0
-----	-------	------	----	-----------	---

206.	0.	01000100.	0
------	----	-----------	---

与  $206.0.68.0/22$  匹配

# 最长前缀匹配举例

收到的分组的地址  $D = 206.0.71.128$

路由表中的项目:  $206.0.68.0/22$  (ISP)

$206.0.71.128/25$  (四系)

再查找路由表中的第 2 个项目

第 2 个项目  $206.0.71.128/25$  的掩码  $M$  有 25 个连续

$M = 11111111\ 11111111\ 11111111\ 10000000$

因此只需把  $D$  的第 4 个字节转换成二进制。

$M = 11111111\ 11111111\ 11111111\ 10000000$

AND	$D =$	206.	0.	71.	10000000
-----	-------	------	----	-----	----------

206.	0.	71.	10000000
------	----	-----	----------

与  $206.0.71.128/25$  匹配





# 最长前缀匹配

---

$D \text{ AND } (11111111 \ 11111111 \ 11111100 \ 00000000)$   
 $= 206.0.68.0/22$  匹配

$D \text{ AND } (11111111 \ 11111111 \ 11111111 \ 10000000)$   
 $= \underline{206.0.71.128/25}$  匹配

- 选择两个匹配的地址中更具体的一个，即选择最长前缀的地址。



### 3. 使用二叉线索查找路由表

---

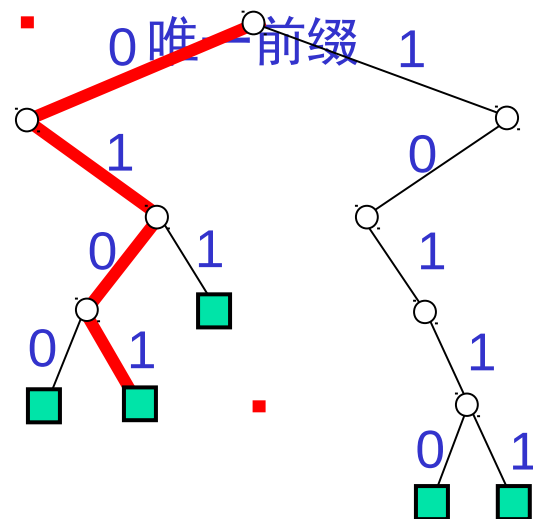
- 当路由表的项目数很大时，怎样设法减小路由表的查找时间就成为一个非常重要的问题。
- 为了进行更加有效的查找，通常是将**无分类编址的路由表**存放在一种层次的数据结构中，然后自上而下地按层次进行查找。这里最常用的就是**二叉线索** (binary trie) 。
- IP 地址中从左到右的比特值决定了从根结点逐层向下层延伸的路径，而二叉线索中的各个路径就代表路由表中存放的各个地址。
- 为了提高二叉线索的查找速度，广泛使用了各种压缩技术。

# 用 5 个前缀构成的二叉线索

32 位的 IP 地址

01000110 00000000 00000000 00000000  
01010110 00000000 00000000 00000000  
01100001 00000000 00000000 00000000  
10110000 00000010 00000000 00000000  
10111011 00001010 00000000 00000000

0100  
0101  
011  
10110  
10111



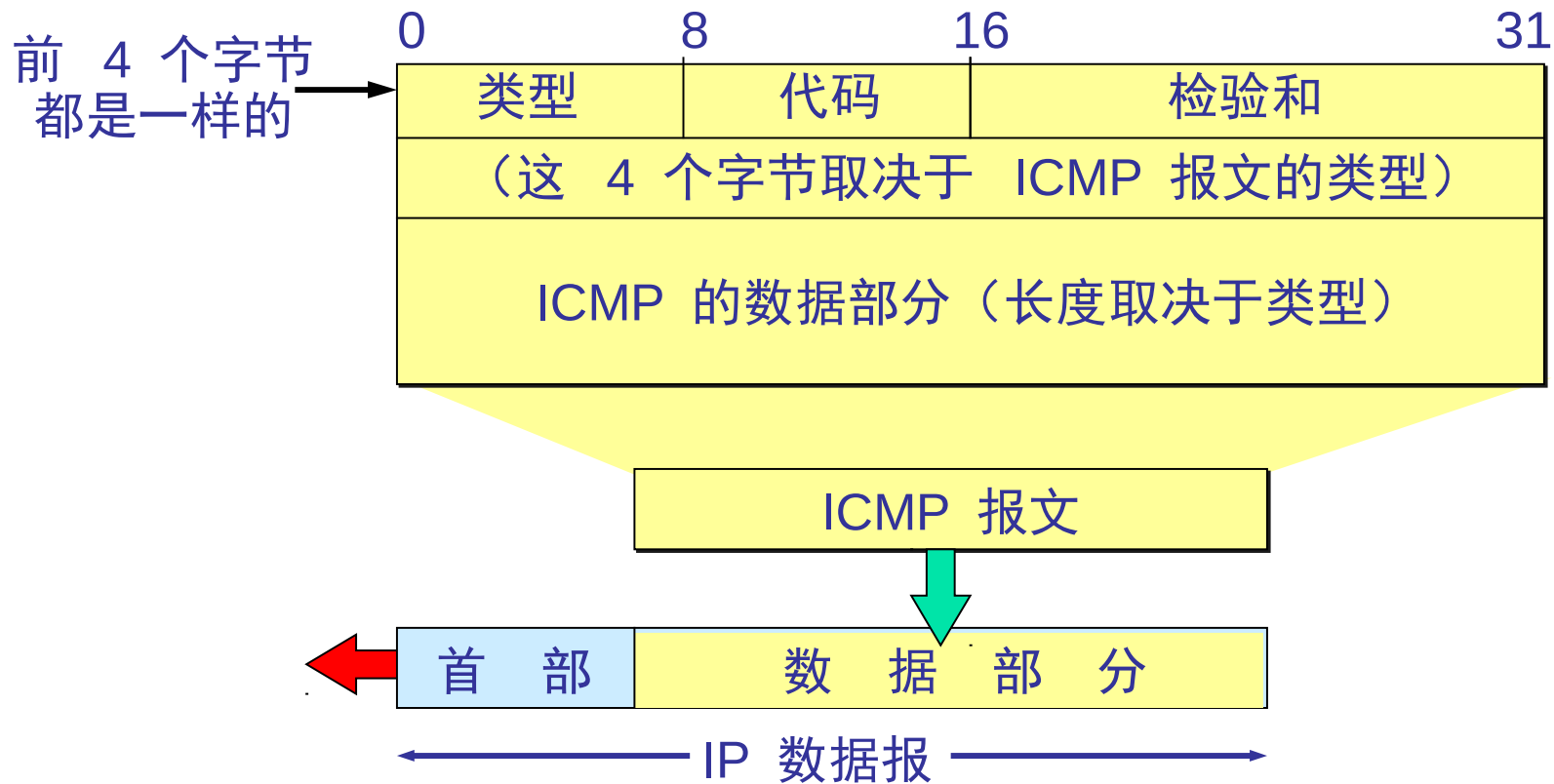


## 4.4 网际控制报文协议 ICMP

---

- 为了提高 IP 数据报交付成功的机会，在网络层使用了网际控制报文协议 ICMP (Internet Control Message Protocol) 。
- 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 不是高层协议，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。

# ICMP 报文的格式





## 4.4.1 ICMP 报文的种类

---

- ICMP 报文的种类有两种，即 ICMP 差错报告报文和 ICMP 询问报文。
- ICMP 报文的前 4 个字节是统一的格式，共有三个字段：即类型、代码和检验和。接着的 4 个字节的内容与 ICMP 的类型有关。

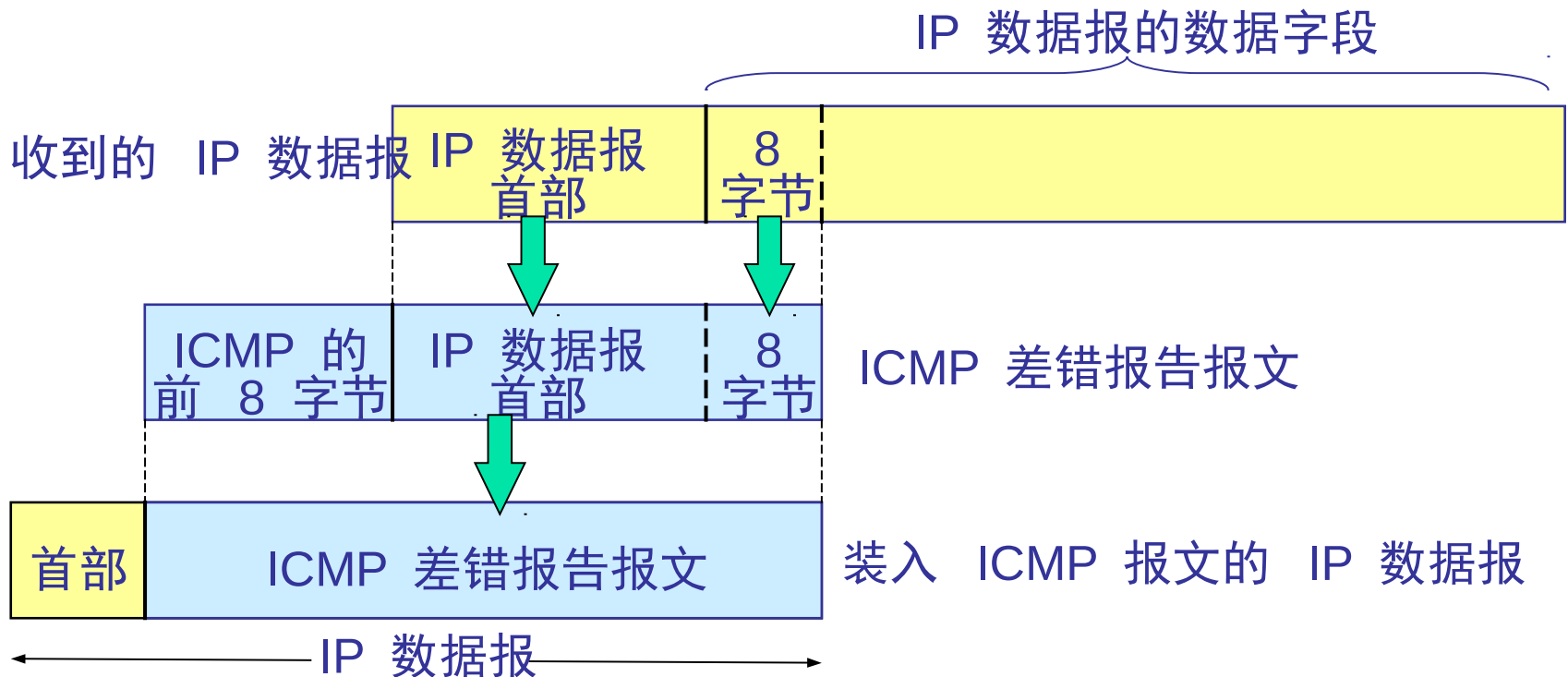


# ICMP 差错报告报文共有 5 种

---

- 终点不可达
- 源点抑制 (Source quench)
- 时间超过
- 参数问题
- 改变路由（重定向） (Redirect)

# ICMP 差错报告报文的数据字段的内容



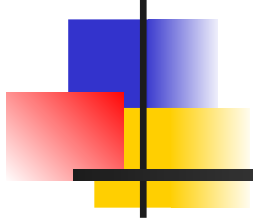




# 不应发送 ICMP 差错报告报文的几种情况

---

- 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（如 127.0.0.0 或 0.0.0.0）的数据报不发送 ICMP 差错报告报文。



# ICMP 询问报文有两种

---

- 回送请求和回答报文
- 时间戳请求和回答报文

下面的几种 ICMP 报文不再使用

- 信息请求与回答报文
- 掩码地址请求和回答报文
- 路由器询问和通告报文

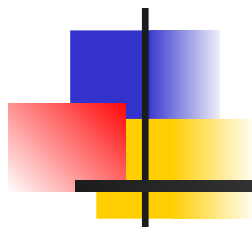


## 4.4.2 ICMP 的应用举例

### PING (Packet InterNet Groper)

---

- PING 用来测试两个主机之间的连通性。
- 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 的例子，它没有通过运输层的 TCP 或 UDP。



# PING 的应用举例

```
C:\Documents and Settings\XXR>ping mail.sina.com.cn

Pinging mail.sina.com.cn [202.108.43.230] with 32 bytes of data:

Reply from 202.108.43.230: bytes=32 time=368ms TTL=242
Reply from 202.108.43.230: bytes=32 time=374ms TTL=242
Request timed out.
Reply from 202.108.43.230: bytes=32 time=374ms TTL=242

Ping statistics for 202.108.43.230:
    Packets: Sent = 4, Received = 3, Lost = 1 (25% loss),
Approximate round trip times in milli-seconds:
    Minimum = 368ms, Maximum = 374ms, Average = 372ms
```



# Traceroute 的应用举例

```
C:\Documents and Settings\XXR>tracert mail.sina.com.cn
```

```
Tracing route to mail.sina.com.cn [202.108.43.230]  
over a maximum of 30 hops:
```

1	24 ms	24 ms	23 ms	222.95.172.1
2	23 ms	24 ms	22 ms	221.231.204.129
3	23 ms	22 ms	23 ms	221.231.206.9
4	24 ms	23 ms	24 ms	202.97.27.37
5	22 ms	23 ms	24 ms	202.97.41.226
6	28 ms	28 ms	28 ms	202.97.35.25
7	50 ms	50 ms	51 ms	202.97.36.86
8	308 ms	311 ms	310 ms	219.158.32.1
9	307 ms	305 ms	305 ms	219.158.13.17
10	164 ms	164 ms	165 ms	202.96.12.154
11	322 ms	320 ms	2988 ms	61.135.148.50
12	321 ms	322 ms	320 ms	freemail43-230.sina.com [202.108.43.230]

```
Trace complete.
```



## 4.5 因特网的路由选择协议

### 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

- 算法必须是正确的和完整的。
- 算法在计算上应简单。
- 算法应能适应通信量和网络拓扑的变化，这就是说，要有自适应性。
- 算法应具有稳定性。
- 算法应是公平的。
- 算法应是最佳的。



# 关于“最佳路由”

---

- 不存在一种绝对的最佳路由算法。
- 所谓“最佳”只能是相对于某一种特定要求下得出的较为合理的选择而已。
- 实际的路由选择算法，应尽可能接近于理想的算法。
- 路由选择是个非常复杂的问题
  - 它是网络中的所有结点共同协调工作的结果。
  - 路由选择的环境往往是不不断变化的，而这种变化有时无法事先知道。



# 从路由算法的自适应性考虑

---

- **静态**路由选择策略——即非自适应路由选择，其特点是简单和开销较小，但不能及时适应网络状态的变化。
- **动态**路由选择策略——即自适应路由选择，其特点是能较好地适应网络状态的变化，但实现起来较为复杂，开销也比较大。





## 2. 分层次的路由选择协议

---

- 因特网采用分层次的路由选择协议。
- 因特网的规模非常大。如果让所有的路由器知道所有的网络应怎样到达，则这种路由表将非常大，处理起来也太花时间。而所有这些路由器之间交换路由信息所需的带宽就会使因特网的通信链路饱和。
- 许多单位不愿意外界了解自己单位网络的布局细节和本部门所采用的路由选择协议（这属于本部门内部的事情），但同时还希望连接到因特网上。



# 自治系统 AS (Autonomous System)

---

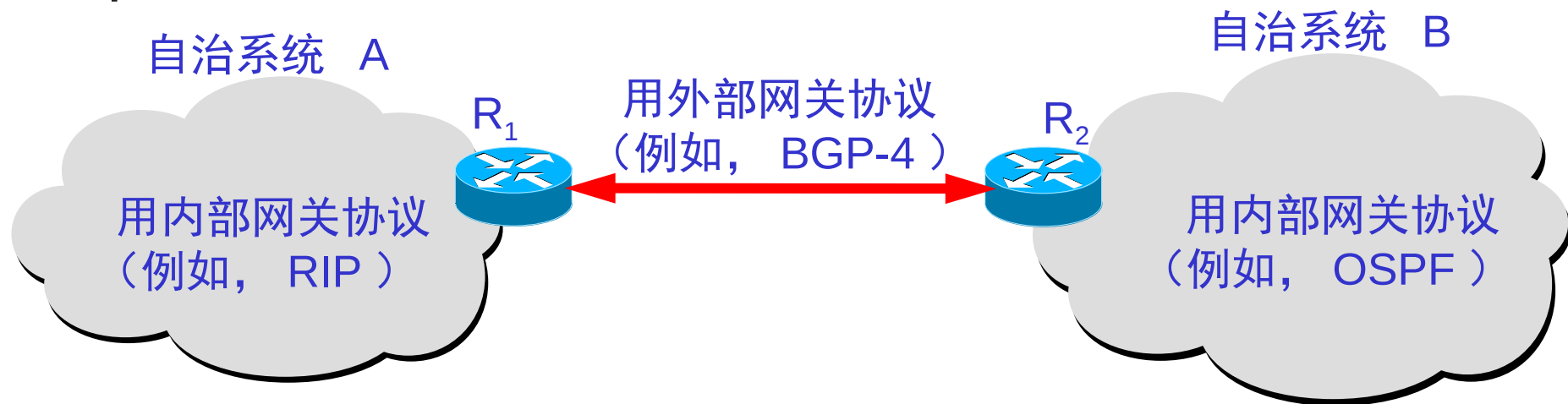
- 自治系统 AS 的定义：在单一的技术管理下的一组路由器，而这些路由器使用一种 AS 内部的路由选择协议和共同的度量以确定分组在该 AS 内的路由，同时还使用一种 AS 之间的路由选择协议用以确定分组在 AS 之间的路由。
- 对自治系统 AS 的定义是强调：尽管一个 AS 使用了多种内部路由选择协议和度量，但重要的是一个 AS 对其他 AS 表现出的是一个**单一的和一致的路由选择策略**。



# 因特网有两大类路由选择协议

- **内部网关协议 IGP** (Interior Gateway Protocol)  
即在一个自治系统内部使用的路由选择协议。  
目前这类路由选择协议使用得最多，如 RIP 和 OSPF 协议。
- **外部网关协议 EGP** (External Gateway Protocol)  
若源站和目的站处在不同的自治系统中，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 EGP。在外部网关协议中目前使用最多的是 BGP-4。

# 自治系统和 内部网关协议、外部网关协议



自治系统之间的路由选择也叫做  
**域间路由选择** (interdomain routing) ,  
在自治系统内部的路由选择叫做  
**域内路由选择** (intradomain routing)



# 这里要指出两点

---

- “网关”与“路由器”

早期 RFC 文档中用“网关”这一名词，不使用“路由器”。但在新的 RFC 文档中又使用了“路由器”这一名词。应当把这两个属于当作同义词。

- RFC 中“EGP”有点混乱，

最早的一个外部网关协议的协议名字叫 EGP。



## 4.5.2 内部网关协议 RIP

### (Routing Information Protocol)

---

#### 1. 工作原理

- 路由信息协议 RIP 是内部网关协议 IGP 中最先得到广泛使用的协议。
- RIP 是一种分布式的基于距离向量的路由选择协议。
- RIP 协议要求网络中的每一个路由器都要维护从它自己到其他每一个目的网络的距离记录。



# “距离”的定义

- RIP 协议中的“距离”也称为“跳数” (hop count)，因为每经过一个路由器，跳数就加 1。“距离”指的是“最短距离”
- 从一路由器到直接连接的网络的距离定义为 1。
- 从一个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。



# “距离”的定义

---

- RIP 不能在两个网络之间同时使用多条路由。RIP 选择一个具有最少路由器的路由（即最短路由），哪怕还存在另一条高速（低时延）但路由器较多的路由。
- RIP 允许一条路径最多只能包含 15 个路由器。
- “距离”的最大值为 16 时即相当于不可达。可见 RIP 只适用于小型互联网。





# RIP 协议的三个要点

---

- 仅和**相邻路由器**交换信息。
- 交换的信息是当前本路由器所知道的**全部信息**，即自己的路由表。
- 按固定的时间间隔**交换路由信息**，例如，每隔 30 秒。

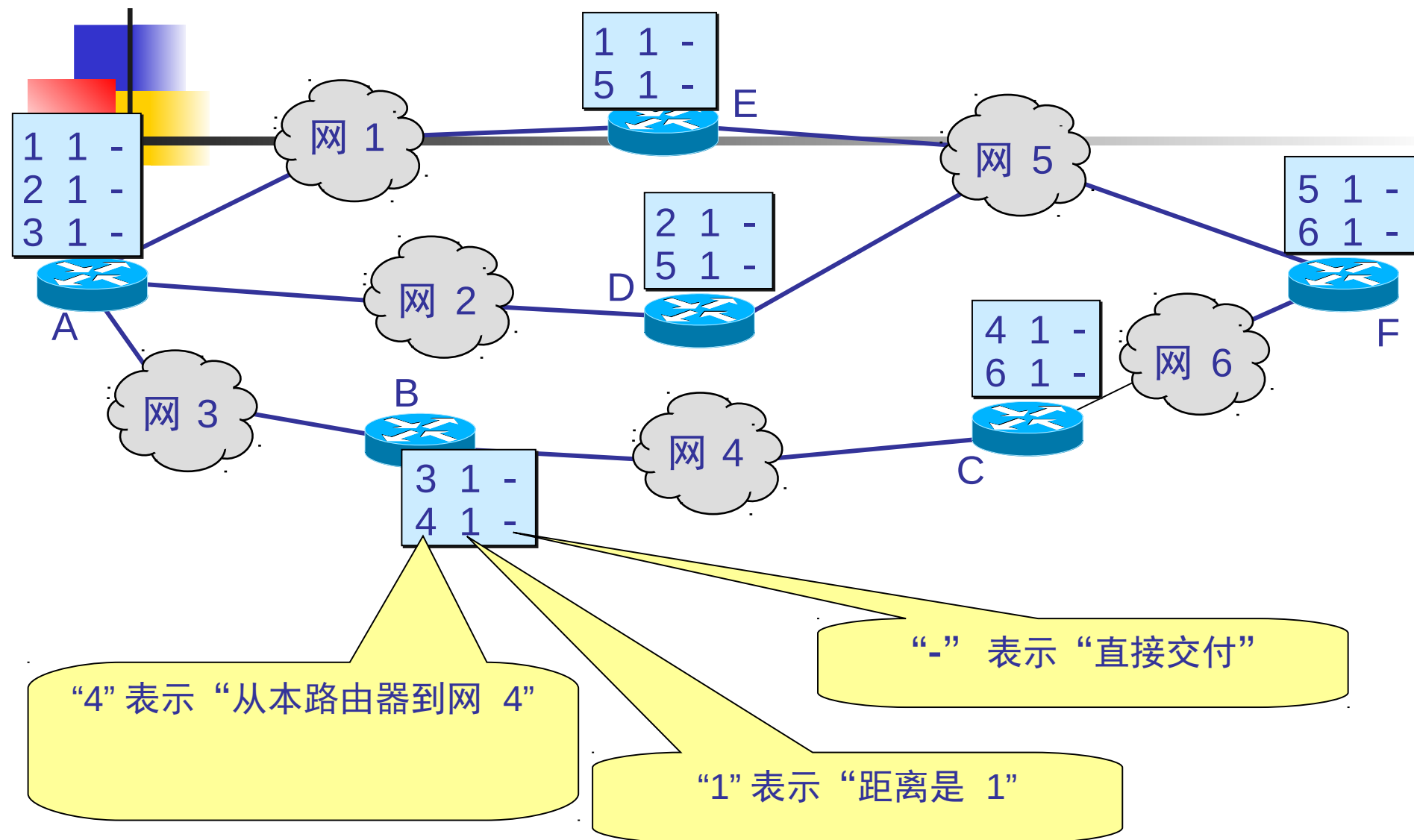


# 路由表的建立

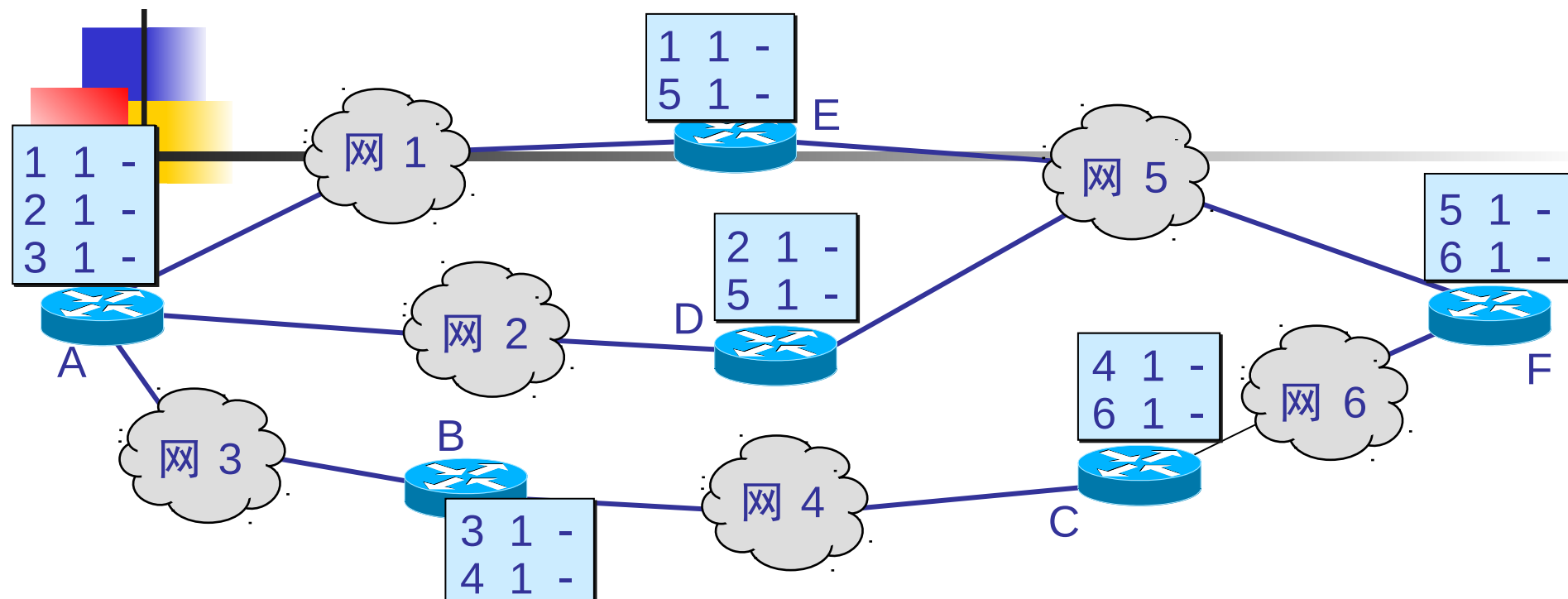
---

- 路由器在刚刚开始工作时，只知道到直接连接的网络的距离（此距离定义为 1）。
- 以后，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。
- 经过若干次更新后，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。
- RIP 协议的**收敛** (convergence) 过程较快，即在自治系统中所有的结点都得到正确的路由选择信息的过程。

一开始，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

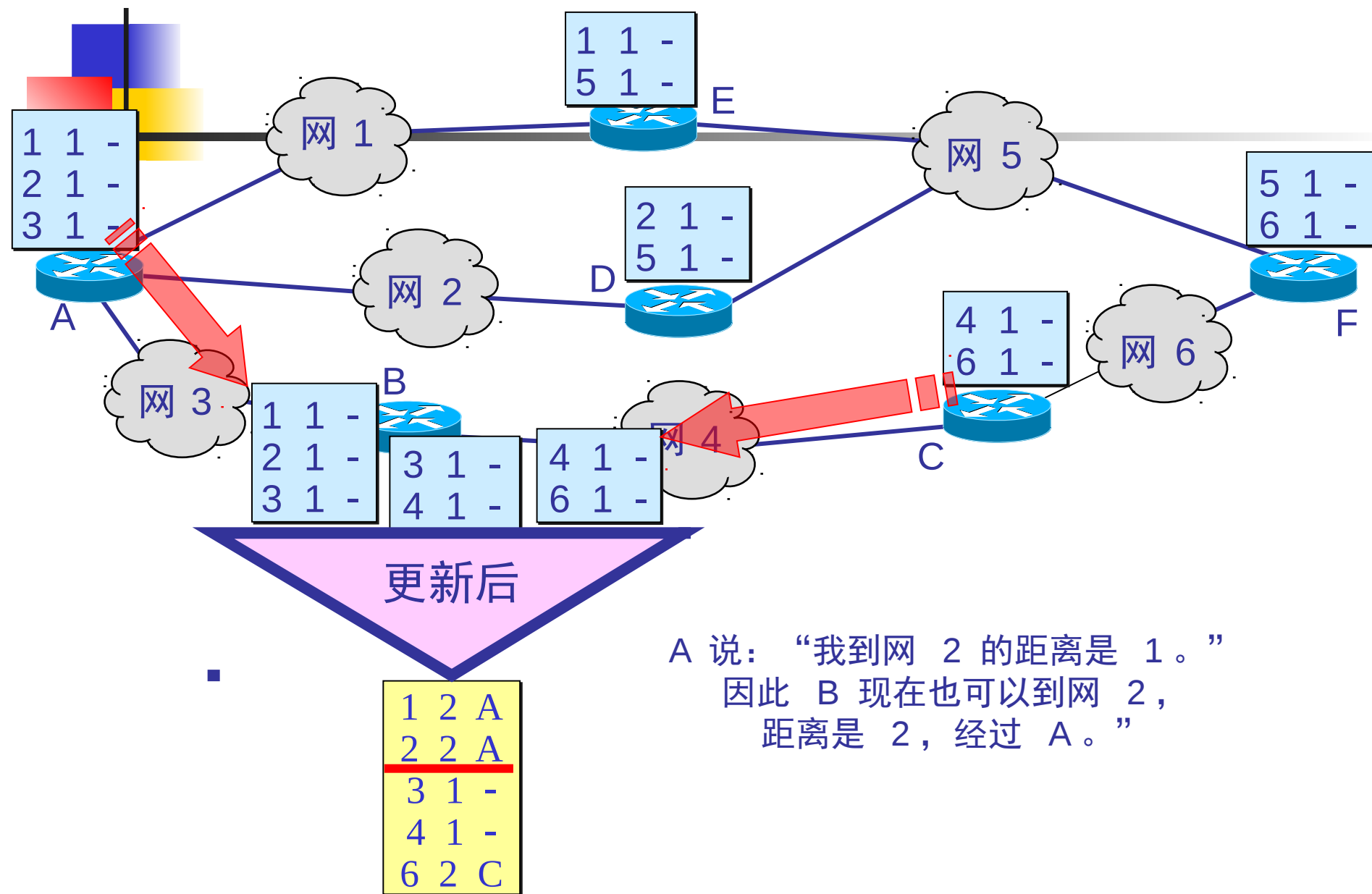


更新后

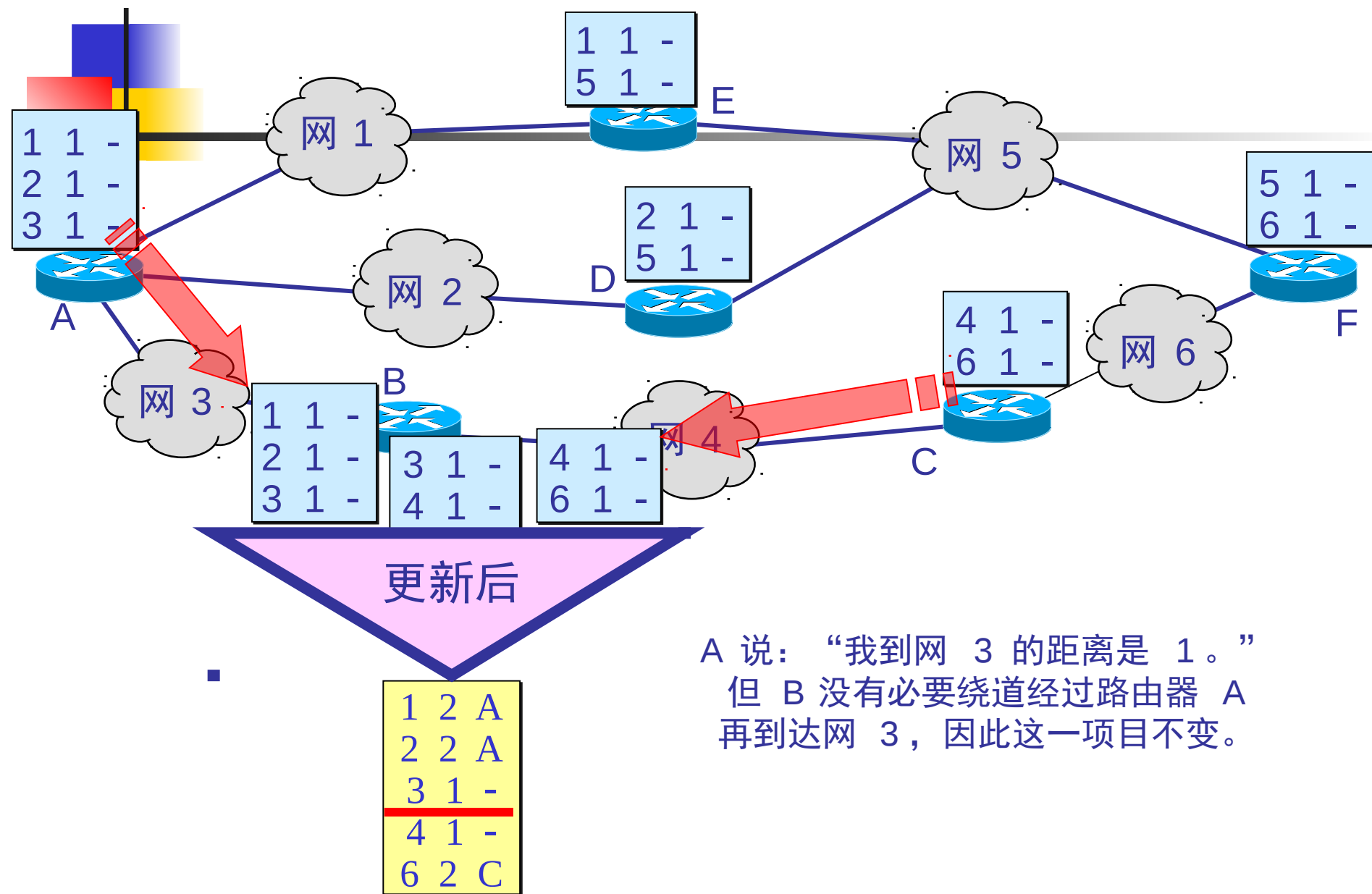
1	2	A
2	2	A
3	1	-
4	1	-
6	2	C

A 说：“我到网 1 的距离是 1。”  
因此 B 现在也可以到网 1，  
距离是 2，经过 A。”

路由器 B 收到相邻路由器 A 和 C 的路由表

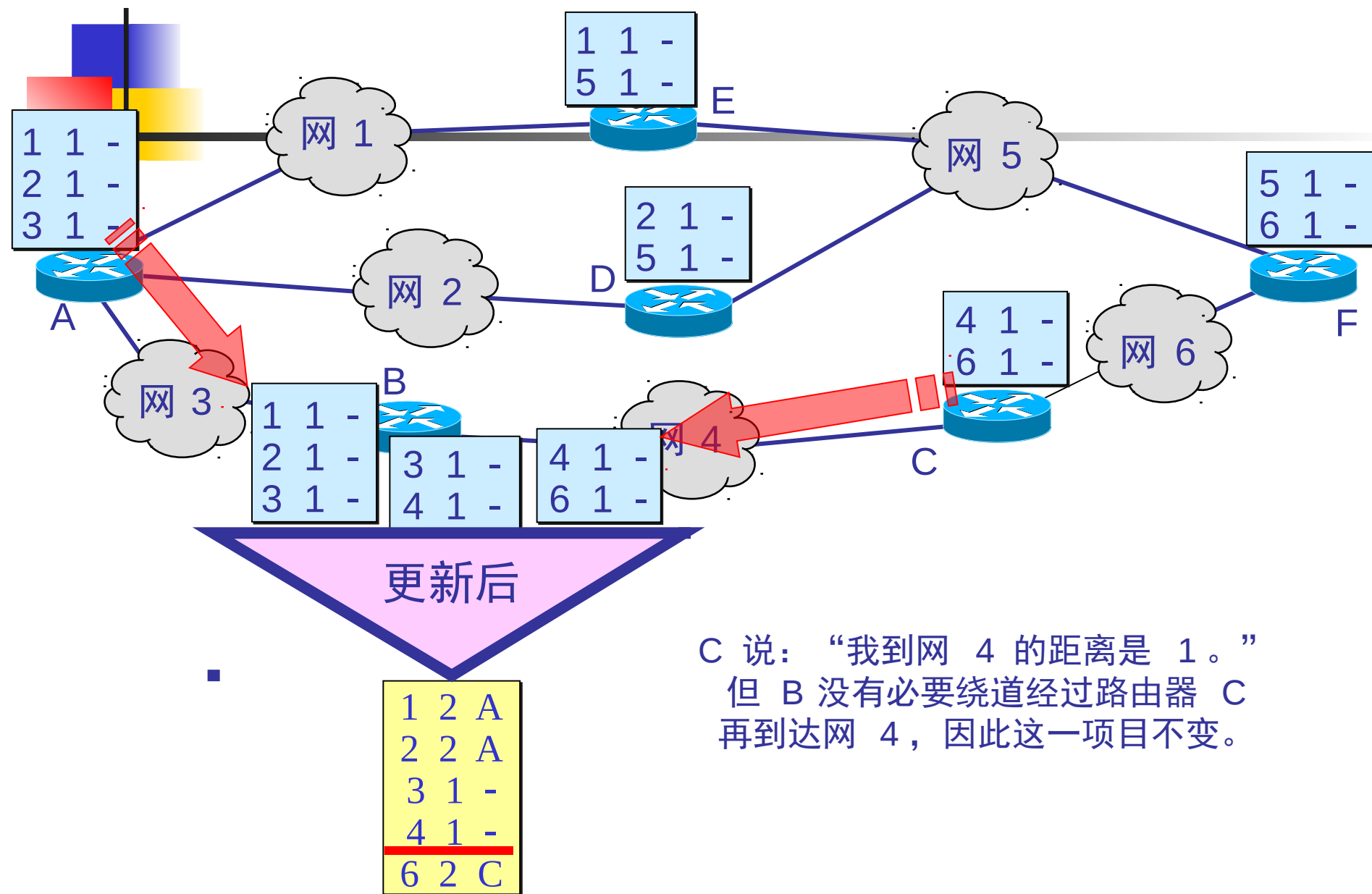


路由器 B 收到相邻路由器 A 和 C 的路由表



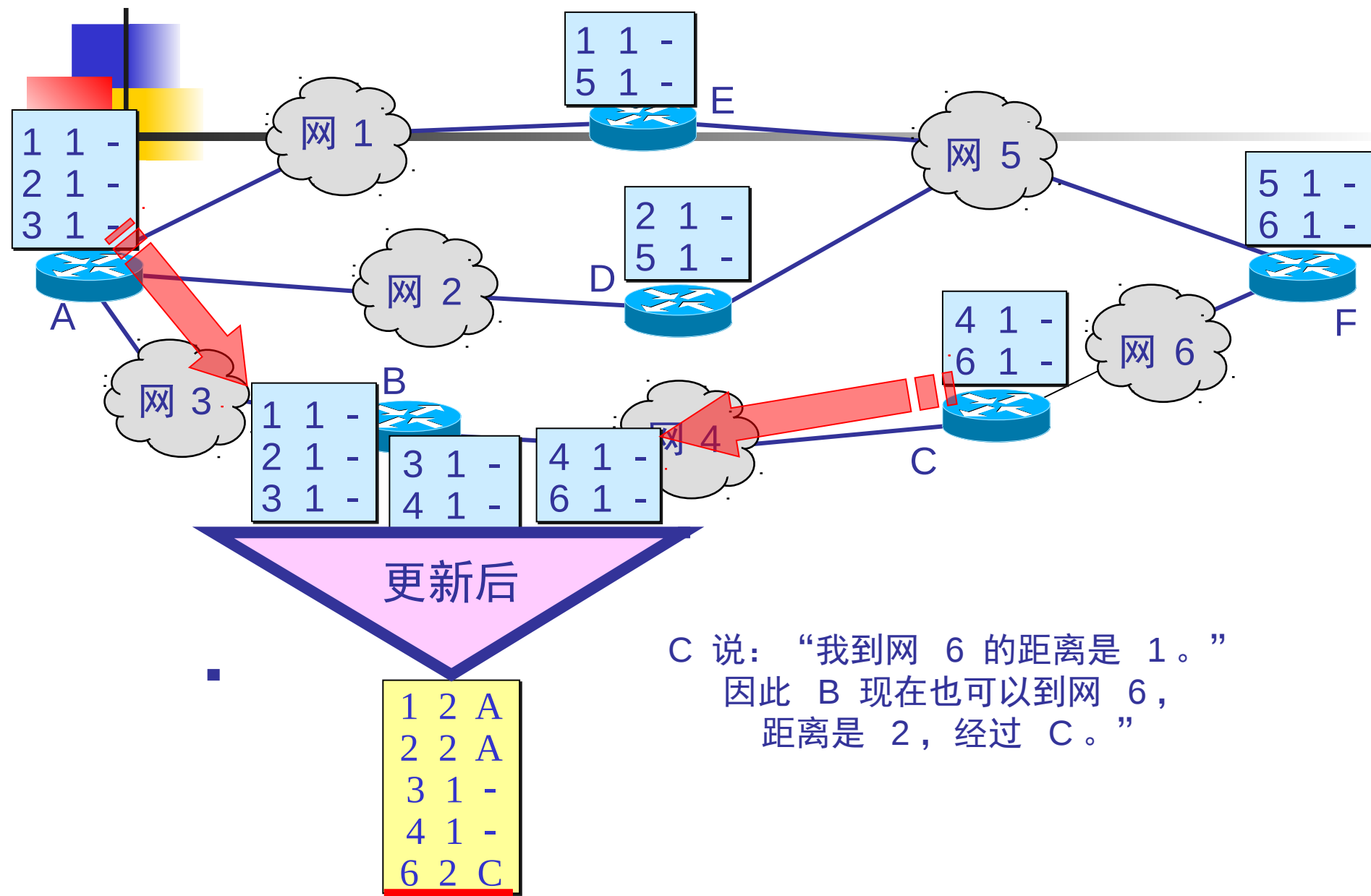
A 说：“我到网 3 的距离是 1。”  
但 B 没有必要绕道经过路由器 A 再到达网 3，因此这一项目不变。

路由器 B 收到相邻路由器 A 和 C 的路由表



C 说：“我到网 4 的距离是 1。”  
但 B 没有必要绕道经过路由器 C 再到达网 4，因此这一项目不变。

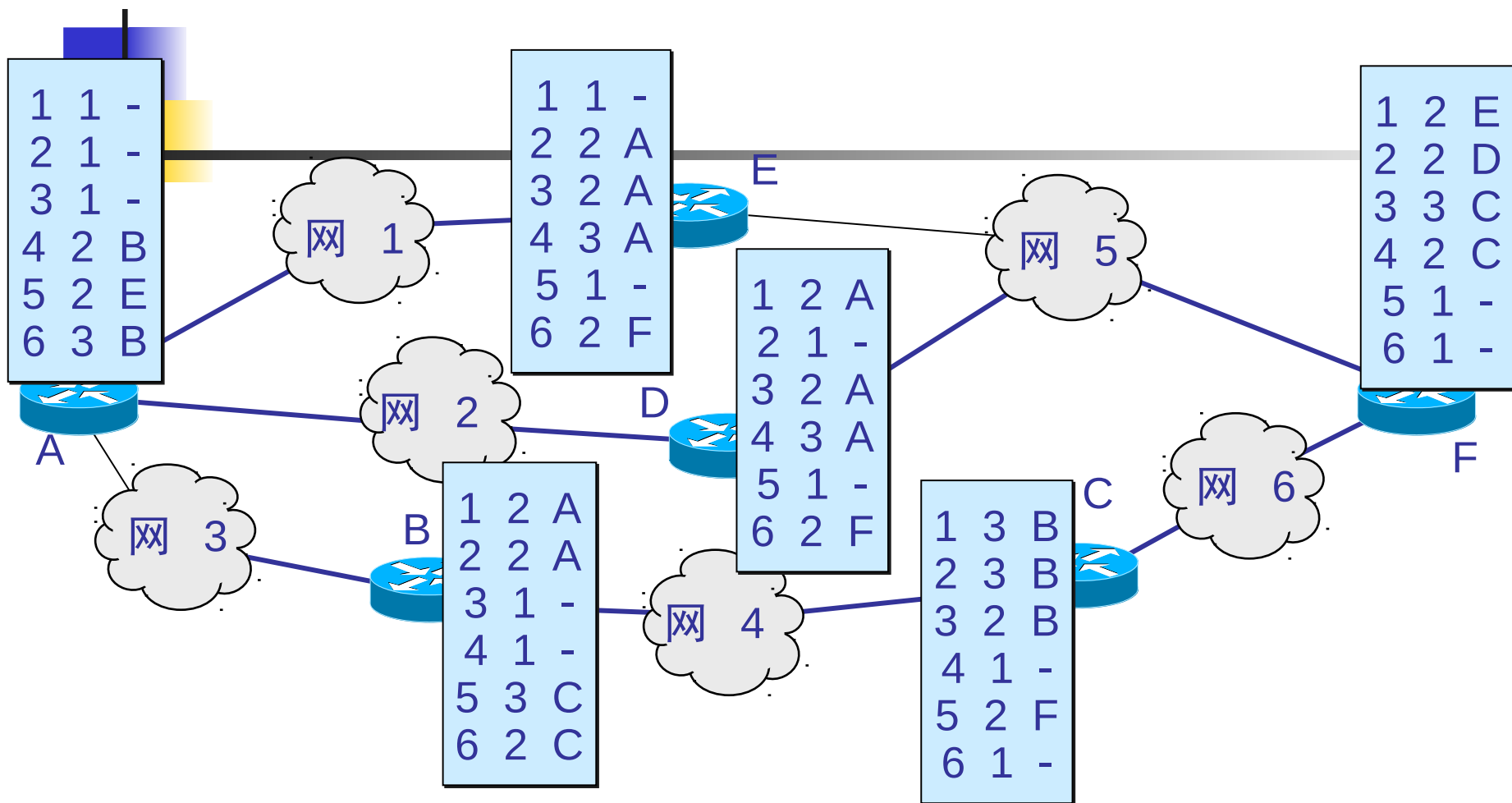
路由器 B 收到相邻路由器 A 和 C 的路由表



C 说：“我到网 6 的距离是 1。”  
因此 B 现在也可以到网 6，  
距离是 2，经过 C。”



最终所有的路由器的路由表都更新了



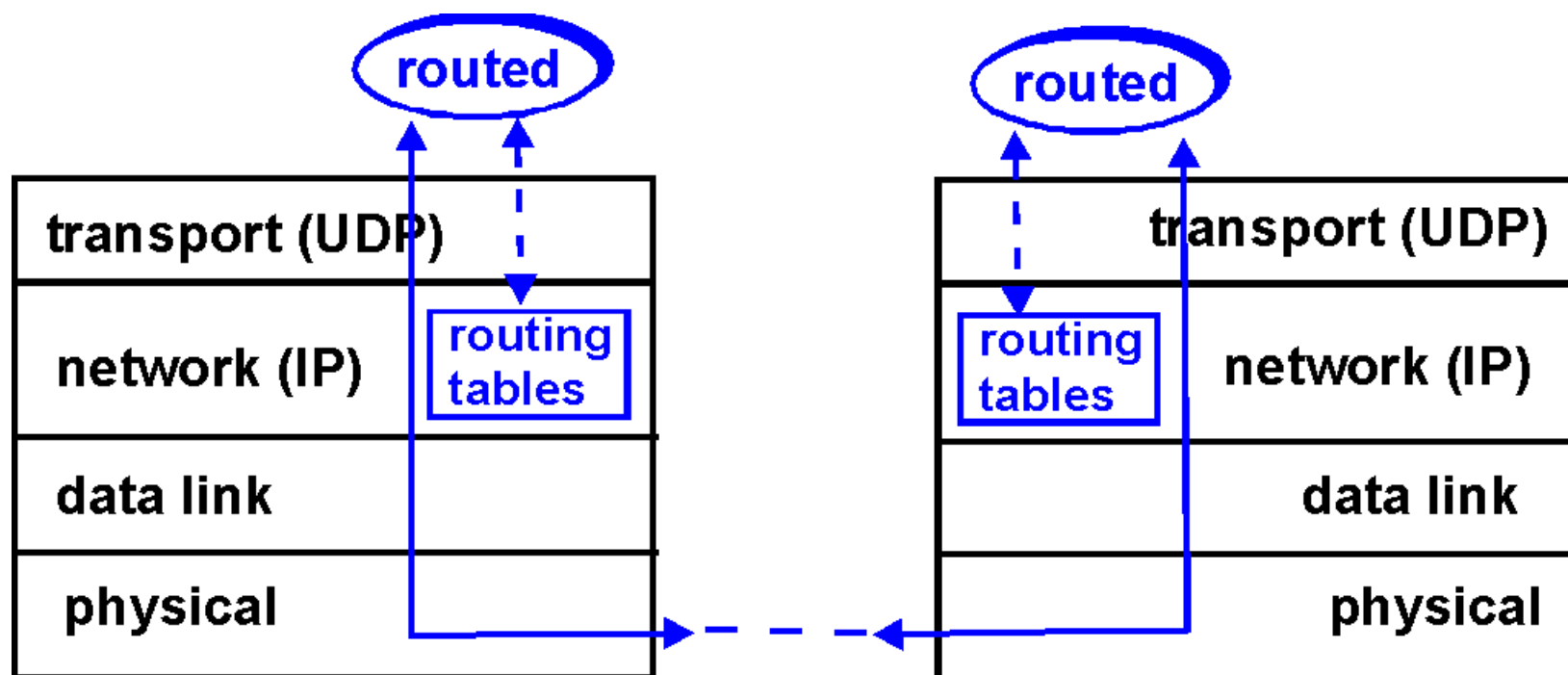


# RIP 协议的位置

---

- RIP 协议 由 **routed 进程** 执行，即对路由表进行维护并和相邻路由器上运行的 routed 进程进行消息的交换。
- 所以 RIP 协议应当在应用层。但转发 IP 数据报的过程是在网络层完成的。
- RIP 是运行在 UDP 之上的应用层协议，（使用 UDP 的端口 520 ）。

# RIP 协议



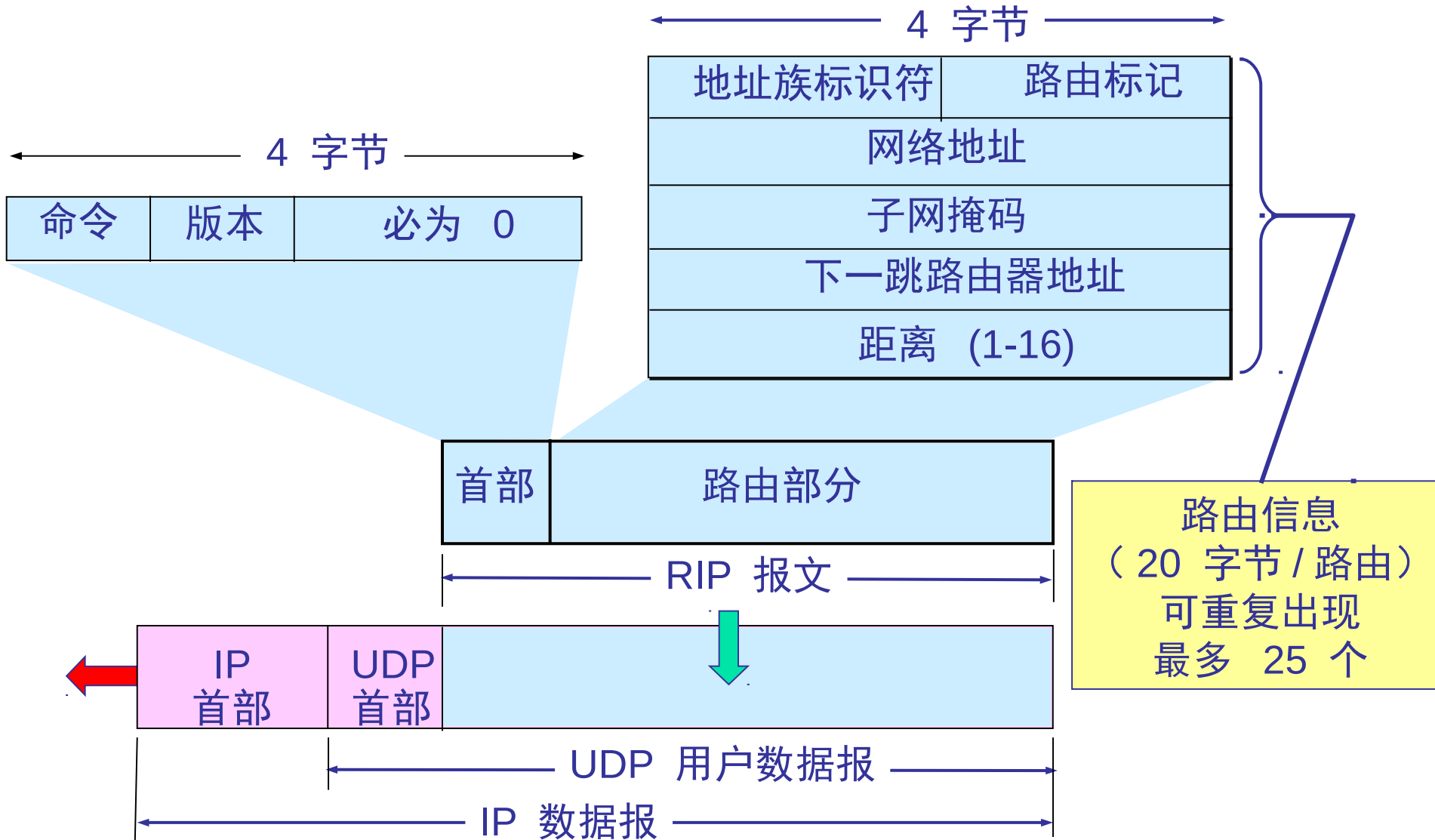


# 路由器之间交换信息

---

- RIP 协议让互联网中的所有路由器都和自己的相邻路由器不断交换路由信息，并不断更新其路由表，使得从每一个路由器到每一个目的网络的路由都是最短的（即跳数最少）。
- 虽然所有的路由器最终都拥有了整个自治系统的全部路由信息，但由于每一个路由器的位置不同，它们的路由表当然也应当是不同的。

# 3. RIP2 协议的报文格式





# RIP2 的报文

## 由首部和路由部分组成。

---

- RIP2 报文中的路由部分由若干个路由信息组成。每个路由信息需要用 20 个字节。**地址族标识符**（又称为地址类别）字段用来标志所使用的地址协议。
- **路由标记**填入自治系统的号码，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。再后面指出某个网络地址、该网络的子网掩码、下一跳路由器地址以及到此网络的距离。

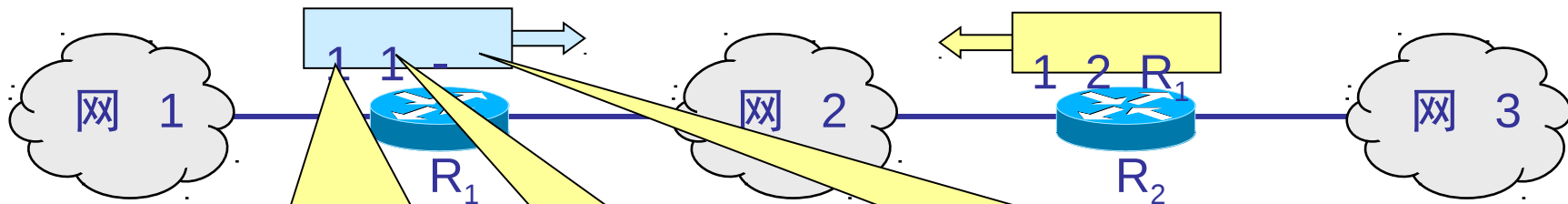


# RIP 协议的优缺点

---

- 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。
- RIP 协议最大的优点就是实现简单，开销较小。
- RIP 限制了网络的规模，它能使用的最大距离为 15（16 表示不可达）。
- 路由器之间交换的路由信息是路由器中的完整路由表，因而随着网络规模的扩大，开销也就增加。

正常情况



“1”表示“从本路由器到网 1”

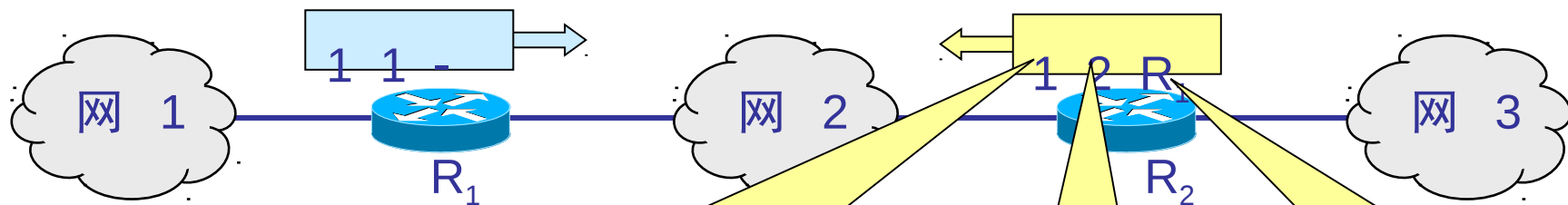
“-”表示“直接交付”

“1”表示“距离是

R<sub>1</sub> 说：“我到网 1 的距离是 1，是直接交付。”



正常情况



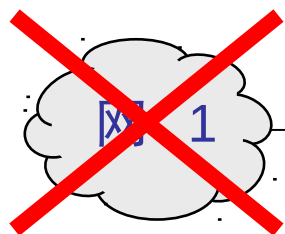
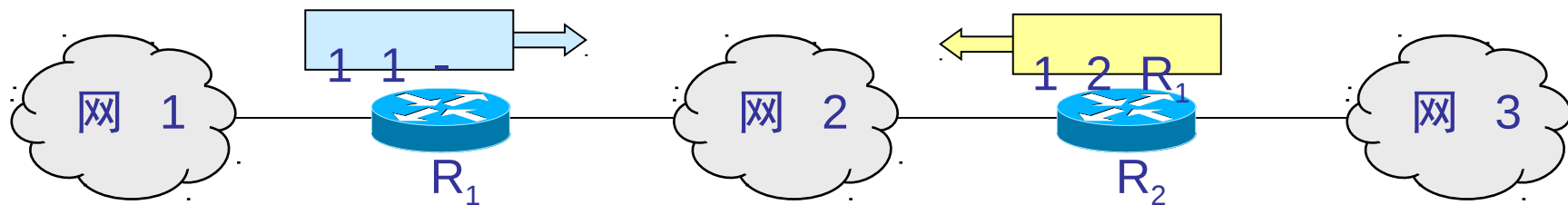
“1”表示“从本路由器到网 1”

“R<sub>1</sub>”表示经过 R<sub>1</sub>

“2”表示“距离是

R<sub>2</sub> 说：“我到网 1 的距离是 2，是经过 R<sub>1</sub>。”

正常情况



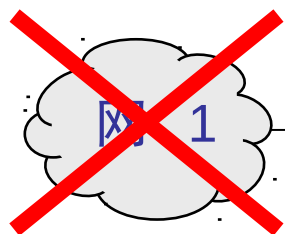
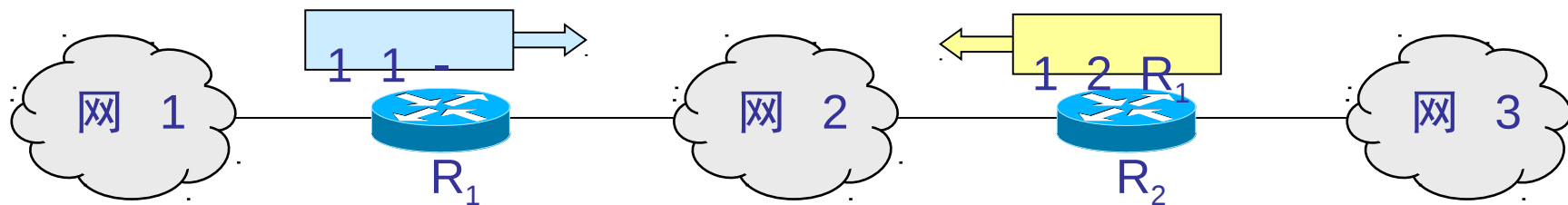
网 1 出了故障



$R_1$  说：“我到网 1 的距离是 16（表示无法到达），是直接交付。”

但  $R_2$  在收到  $R_1$  的更新报文之前，还发送原来的报文，因为这时  $R_2$  并不知道  $R_1$  出了故障。

正常情况

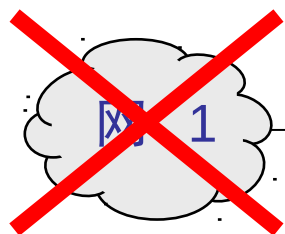
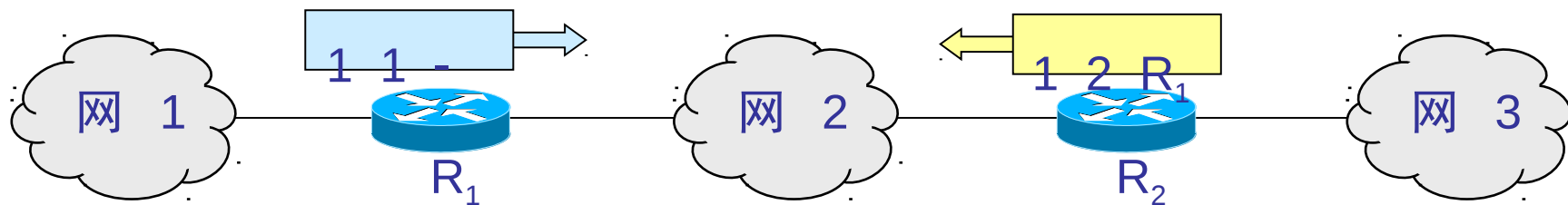


网 1 出了故障



$R_1$  收到  $R_2$  的更新报文后，误认为可经过  $R_2$  到达网 1，于是更新自己的路由表，说：“我到网 1 的距离是 3，下一跳经过  $R_2$ ”。然后将此更新信息发送给  $R_2$ 。

正常情况

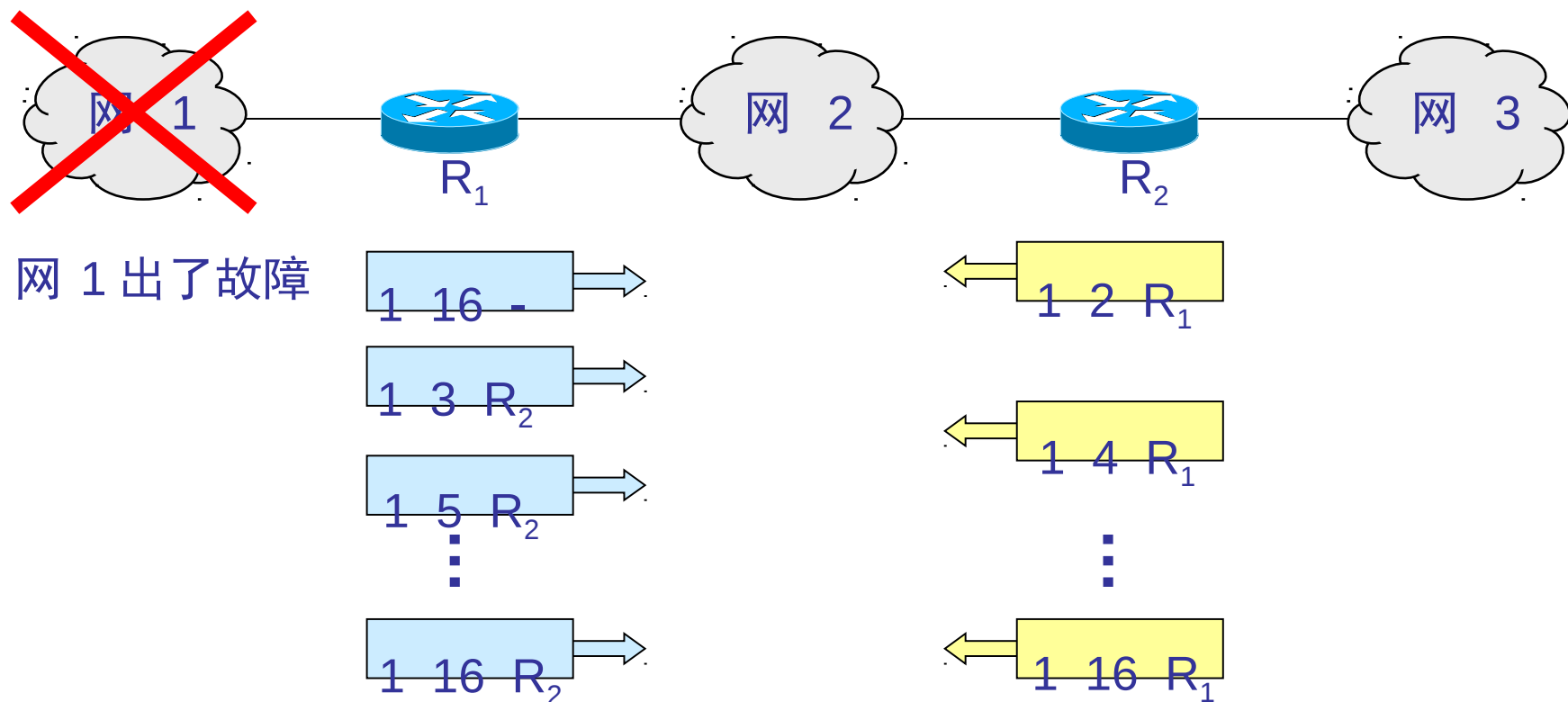


网 1 出了故障



$R_2$  以后又更新自己的路由表为 “1, 4,  $R_1$ ”，表明  
“我到网 1 距离是 4，下一跳经过  $R_1$ ”。

这就是好消息传播得快，而坏消息传播得慢。网络出现故障的传播时间往往需要较长的时间（例如数分钟）。这是 RIP 的一个主要缺点。



这样不断更新下去，直到  $R_1$  和  $R_2$  到网 1 的距离都增大到 16 时， $R_1$  和  $R_2$  才知道网 1 是不可达



## 4.5.3 内部网关协议 OSPF

### (Open Shortest Path First)

---

- Open Shortest Path First-- 开放式最短路径优先
  - I E T F 于 1 9 8 9 年开发的一种链路状态路由技术。作为一个内部网关协议，专门为 TCP/IP Internet 环境设计，它可以在那些 R I P 不能处理的、大型的网络上使用。



# 内部网关协议 OSPF

---

- OSPF 协议的基本特点
  - “**开放**”表明 OSPF 协议不是受某一家厂商控制，而是公开发表的。
  - “**最短路径优先**”是因为使用了 Dijkstra 提出的最短路径算法 SPF
  - OSPF 是分布式的链路状态协议。
  - OSPF 只是一个协议的名字，它并不表示其他的路由选择协议不是“最短路径优先”。



# OSPF 协议的主要特点

---

- 使用分布式的链路状态协议；
- 路由器发送的信息是本路由器与哪些路由器相邻，以及链路状态（距离、时延、带宽等）信息；
- 当链路状态发生变化时用洪泛法向所有路由器发送；
- 所有的路由器最终都能建立一个链路状态数据库 ；
  - 通过各个节点之间的路由信息交换
  - 每个节点可获得关于全网的拓扑信息，得知网中所有的节点、各节点间的链路连接和各条链路的代价。
  - 将这些拓扑信息抽象成一张带权无向图，然后利用最短通路路由选择算法计算出到各个目的节点的最短通路。
- 将一个自治系统再划分为若干个更小的区域，一个区域内的路由器数不超过 200 个。



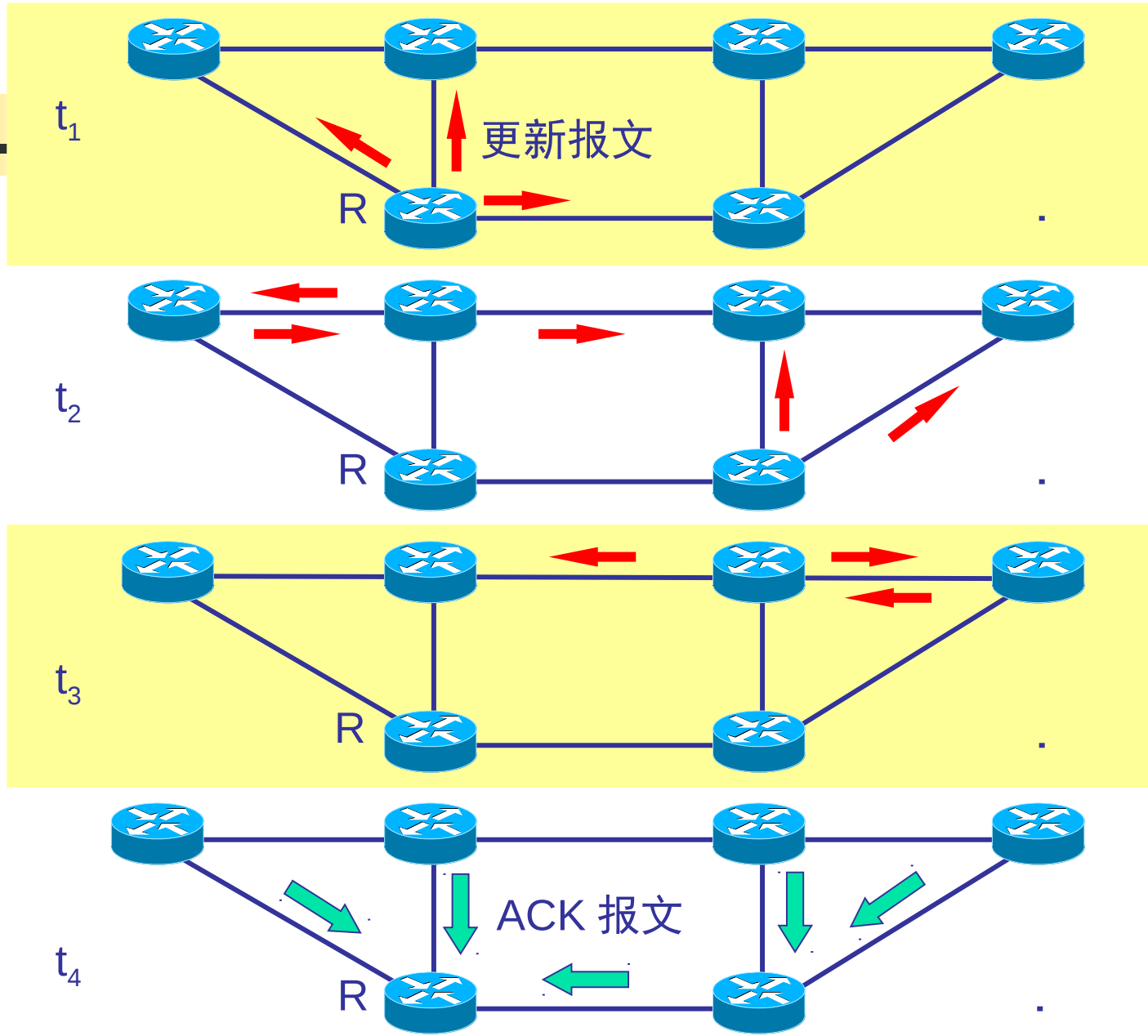
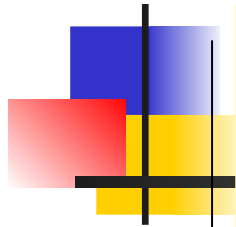


# 三个要点

---

- 向本自治系统中所有路由器发送信息，这里使用的方法是洪泛法。
- 发送的信息就是与本路由器相邻的所有路由器的链路状态，但这只是路由器所知道的部分信息。
  - “链路状态”就是说明本路由器都和哪些路由器相邻，以及该链路的“度量” (metric)。
- 只有当链路状态发生变化时，路由器才用洪泛法向所有路由器发送此信息。

# OSPF 使用的是可靠的洪泛法





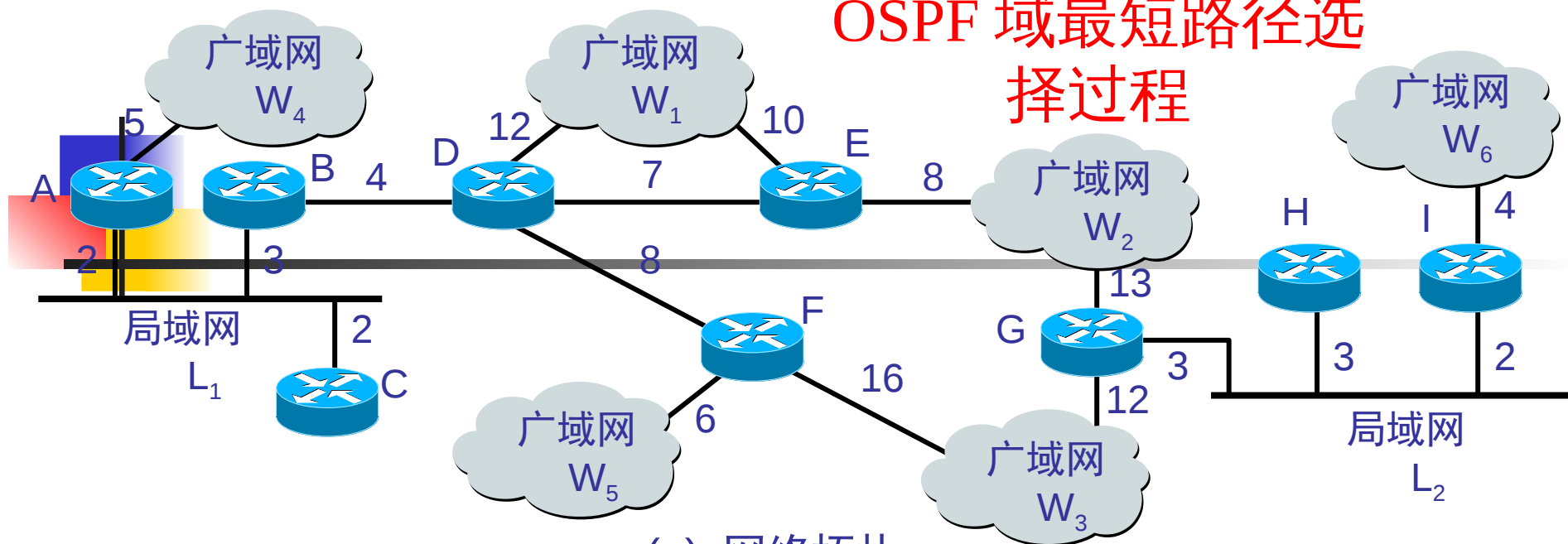
# 链路状态数据库

## (link-state database)

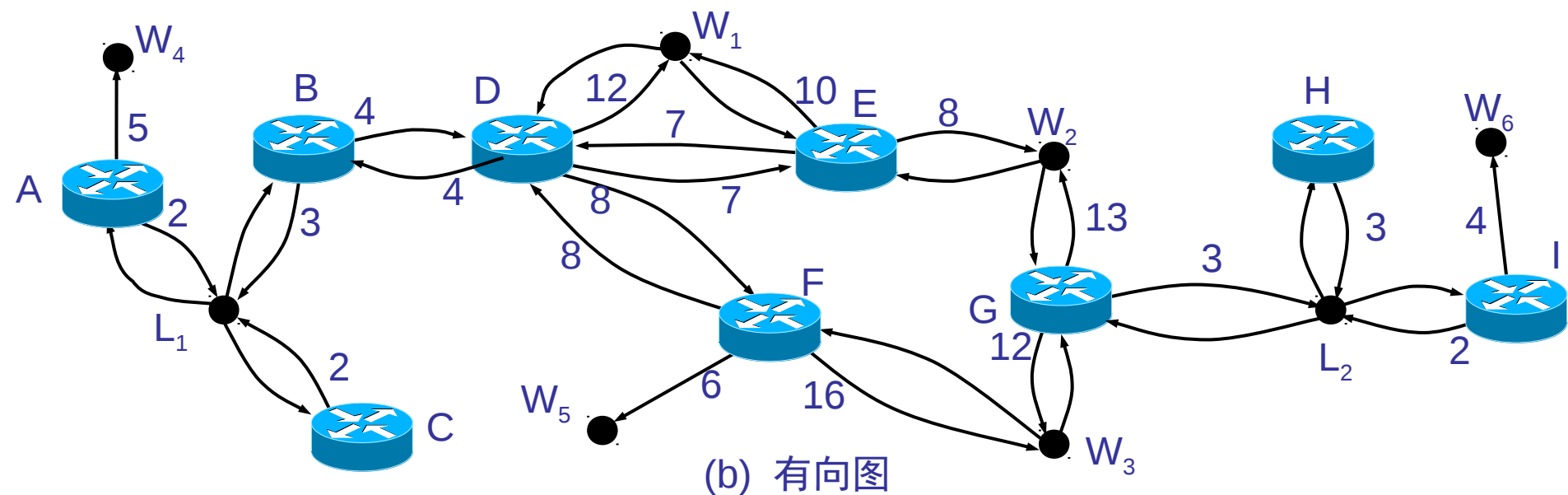
---

- 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**，它在全网范围内是一致的（这称为链路状态数据库的同步）。
- OSPF 的链路状态数据库能较快地进行更新，使各个路由器能及时更新其路由表。OSPF 的更新过程收敛得快是其重要优点。

# OSPF 域最短路径选择过程

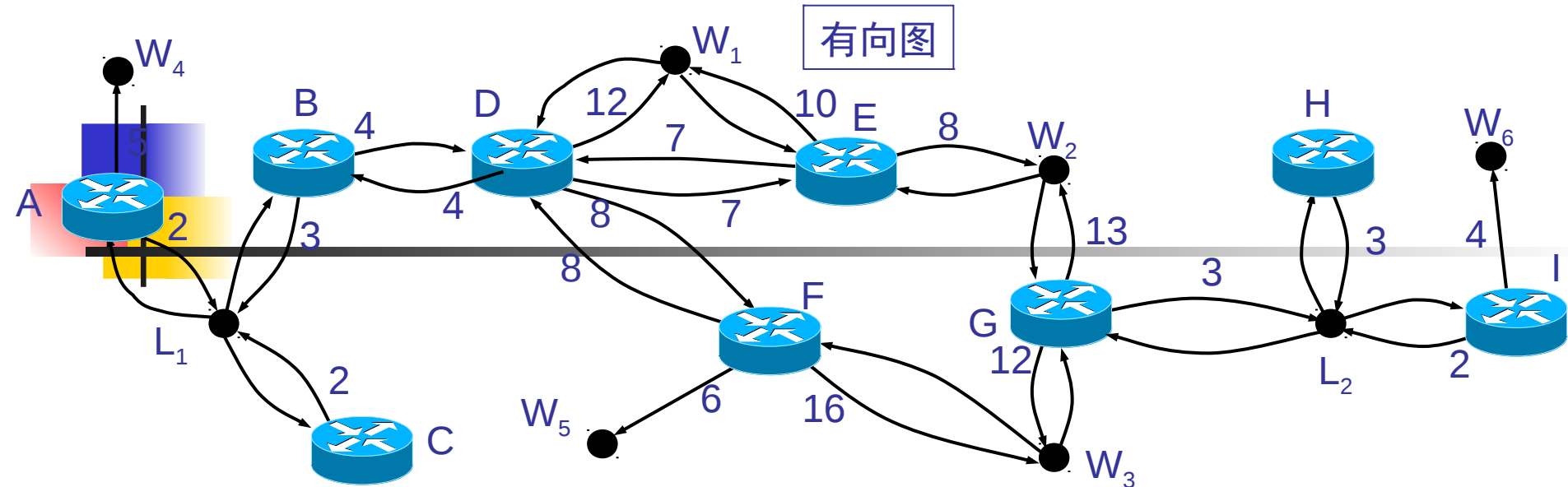


(a) 网络拓扑

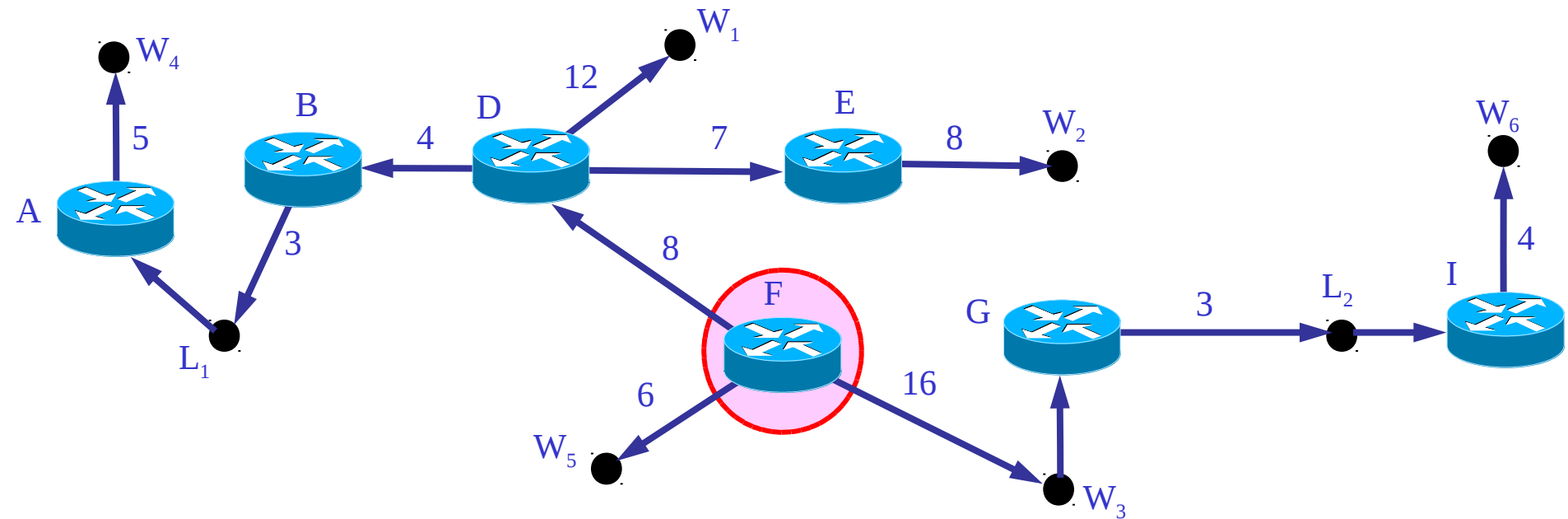


(b) 有向图

有向图



以路由器 F 为根的最短路径树





# OSPF 的区域 (area)

---

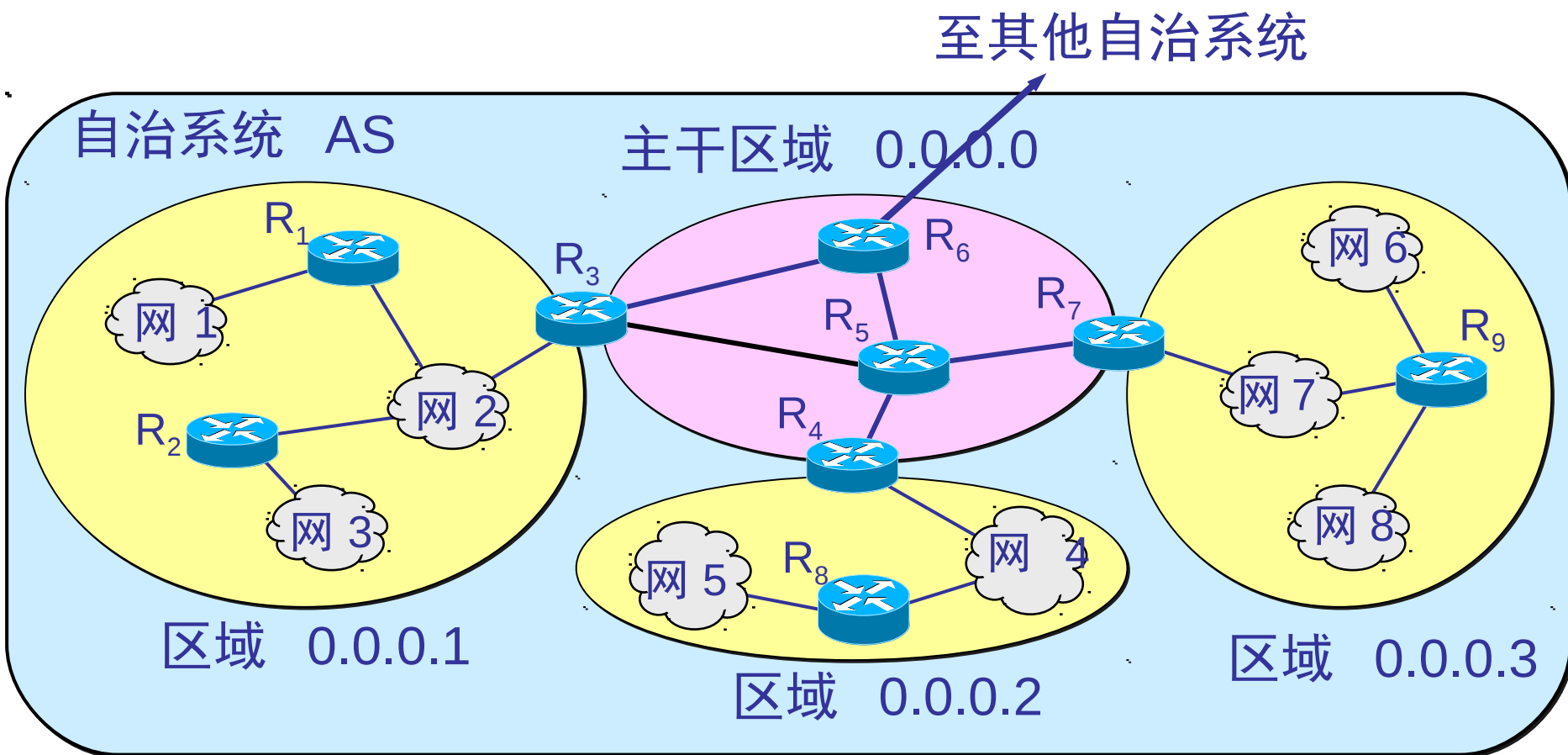
- 为了使 OSPF 能够用于规模很大的网络，OSPF 将一个自治系统再划分为若干个更小的范围，叫作区域。
- 每一个区域都有一个 32 位的区域标识符（用点分十进制表示）。
- 区域也不能太大，在一个区域内的路由器最好不超过 200 个。



# 划分区域的好处

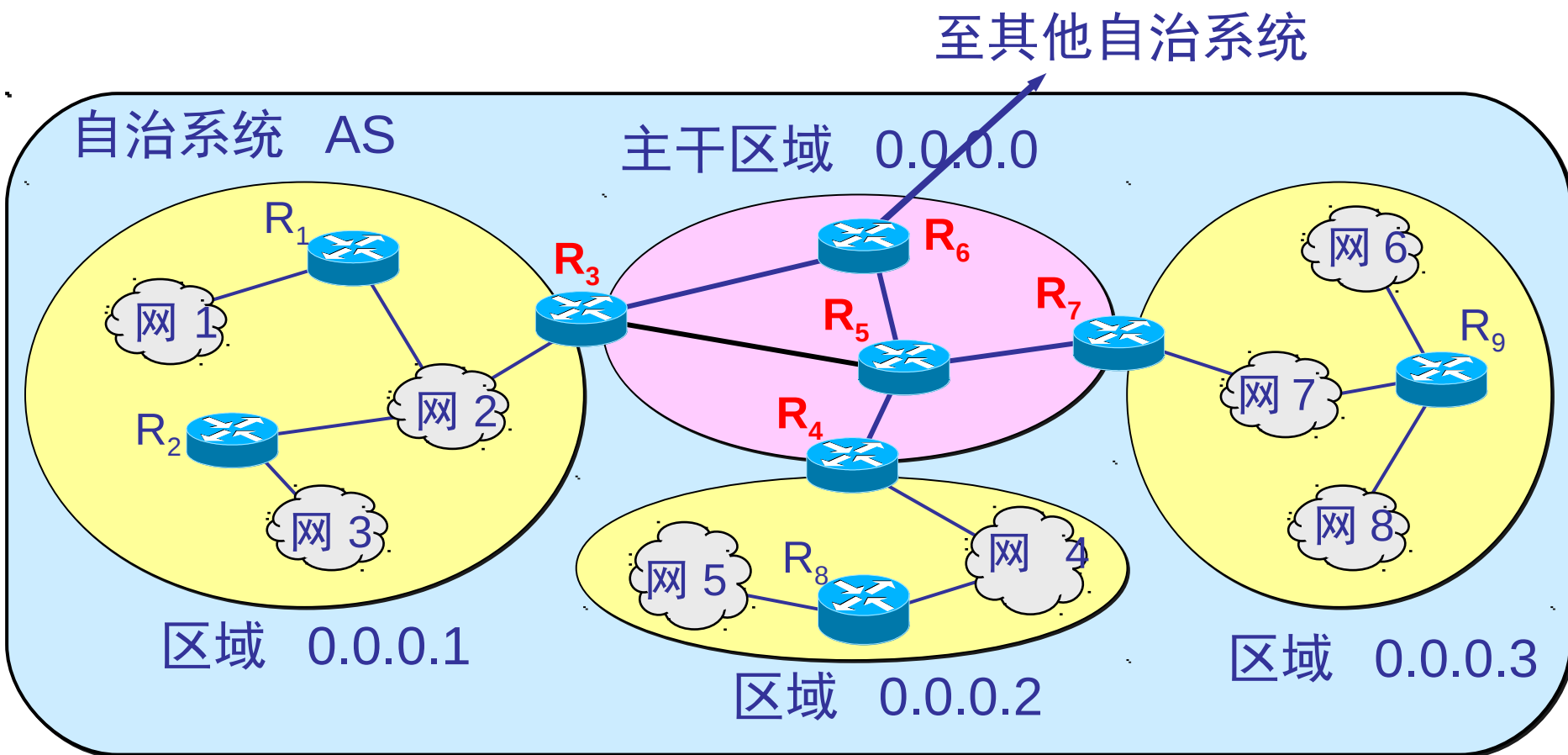
- **划分区域的好处**就是将利用洪泛法交换链路状态信息的范围局限于每一个区域而不是整个的自治系统，这就**减少了整个网络上的通信量**。
- 在一个区域内部的路由器只知道本区域的完整网络拓扑，而不知道其他区域的网络拓扑的情况。
- OSPF 使用**层次结构的区域划分**。在上层的区域叫作**主干区域** (backbone area)。主干区域的标识符规定为 0.0.0.0。**主干区域的作用**是用来连通其他在下层的区域。

# OSPF 划分为两种不同的区域



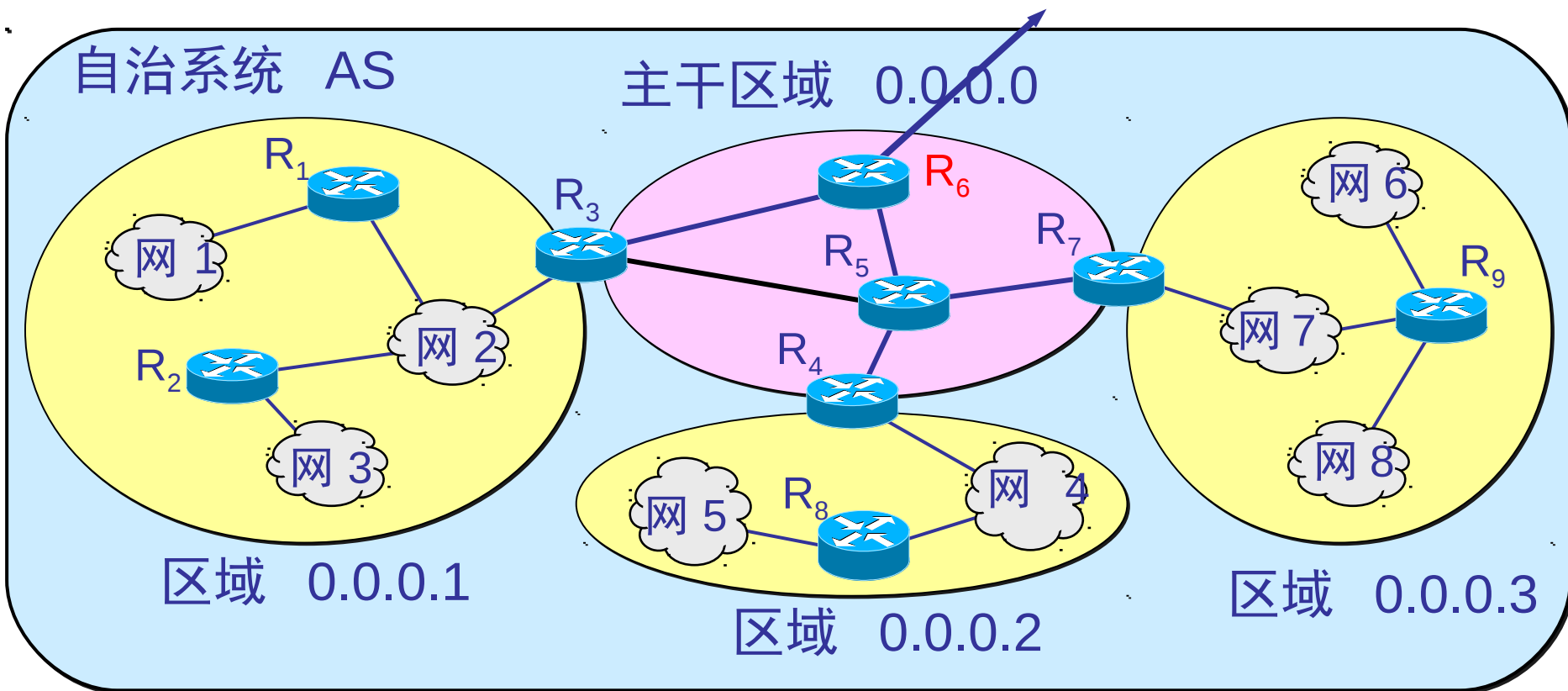


主干区域内部的路由器叫做**主干路由器**（如图中 R3、R4、R5、R6、R7），它连接各个区域的边界路由器。

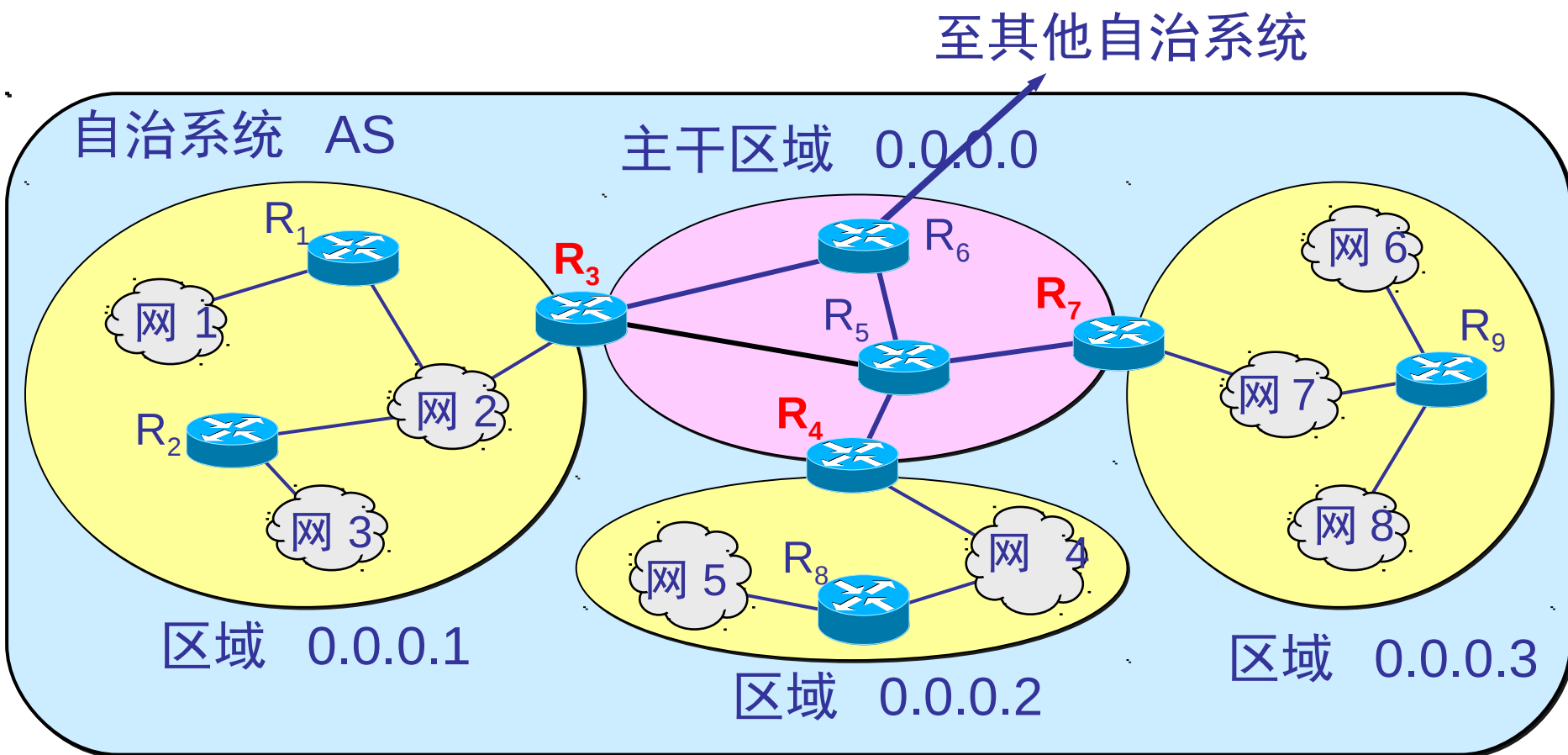


在主干区域内还要有一个路由器专门和该自治系统之外的其他自治系统交换路由信息。这样的路由器叫做  
**自治系统边界路由器（如 R6）**。

至其他自治系统



# 区域边界路由器





# OSPF 的其他特点

---

- (1) 直接用 IP 数据报传送。 OSPF 不用 UDP 而是**直接用 IP 数据报传送**，可见 OSPF 的位置在**网络层**。
- (2) OSPF 构成的数据报很短。这样做的好处：
  - 可减少路由信息的通信量。
  - 可以不必将长的数据报分片传送。分片传送的数据报只要丢失一个，就无法组装成原来的数据报，而整个数据报就必须重传。



# OSPF 的其他特点

- (3) OSPF 对不同的链路可根据 IP 分组的不同服务类型 而设置成不同的代价。因此，OSPF 对于不同类型的业务可计算出不同的路由。
- (4) 如果到同一个目的网络有多条相同代价的路径，那么可以将通信量分配给这几条路径。这叫作**多路径间的负载平衡**。
- (5) 所有在 OSPF 路由器之间交换的分组（例如，链路状态更新分组）都具有**鉴别的功能**。因而保证了仅在可信赖的路由器之间交换链路状态信息。
- (6) 支持可变长度的子网划分和无分类编址 CIDR。
- (7) 每一个链路状态都带上一个 32 bit 的序号，序号越大状态就越新。OSPF 规定，链路状态序号增长的速率不得超过每 5 秒钟 1 次。这样，全部序号空间在 600 年内不会产生重复号。

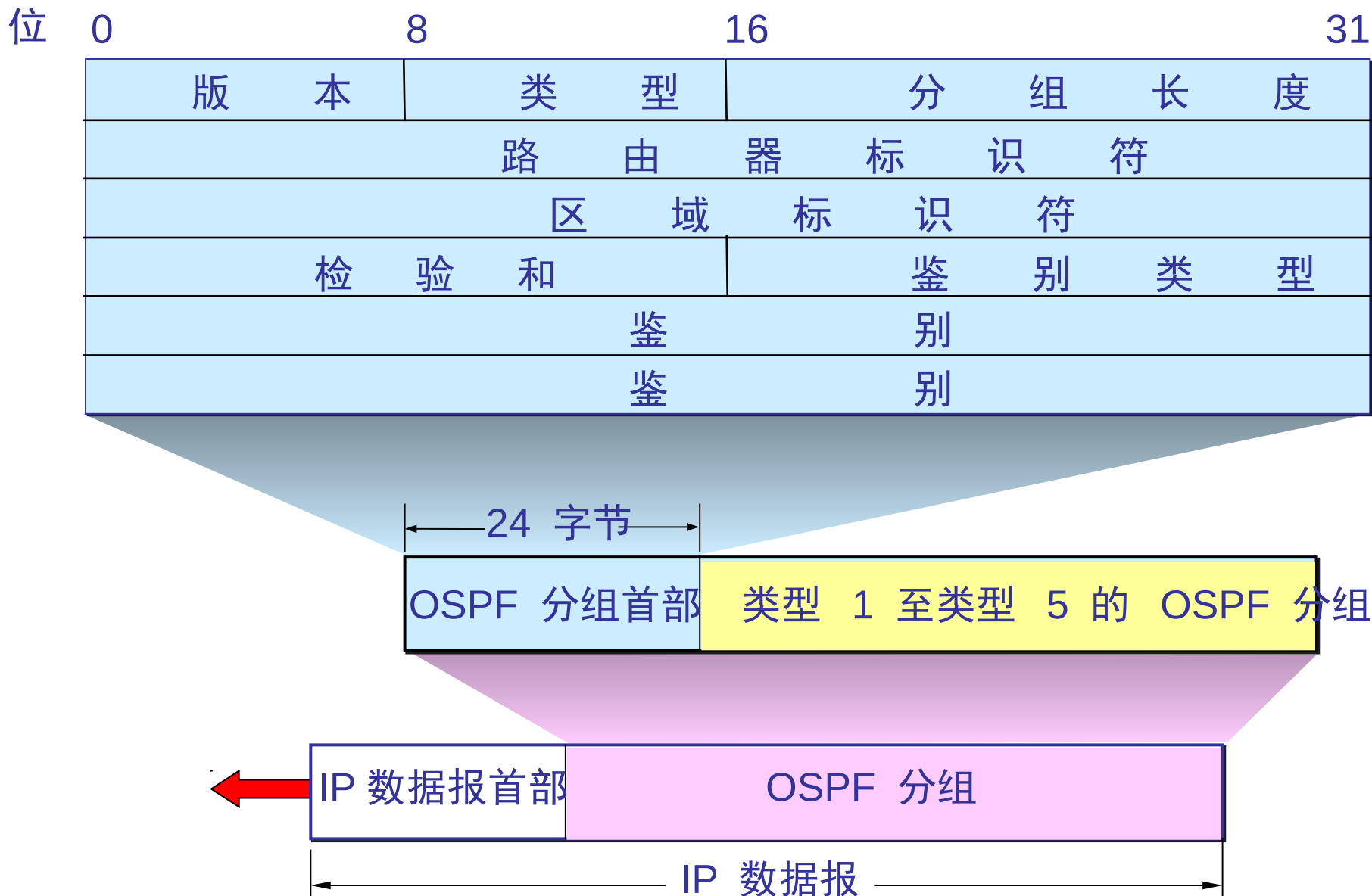


# OSPF 的其他特点

---

- OSPF 还规定每隔一段时间，如 30 分钟，要刷新一次数据库中的链路状态。
- 由于一个路由器的链路状态只涉及到与相邻路由器的连通状态，因而与整个互联网的规模并无直接关系。因此当互联网规模很大时，OSPF 协议要比距离向量协议 RIP 好得多。
- OSPF 没有“坏消息传播得慢”的问题，据统计，其响应网络变化的时间小于 100 ms。

# OSPF 分组





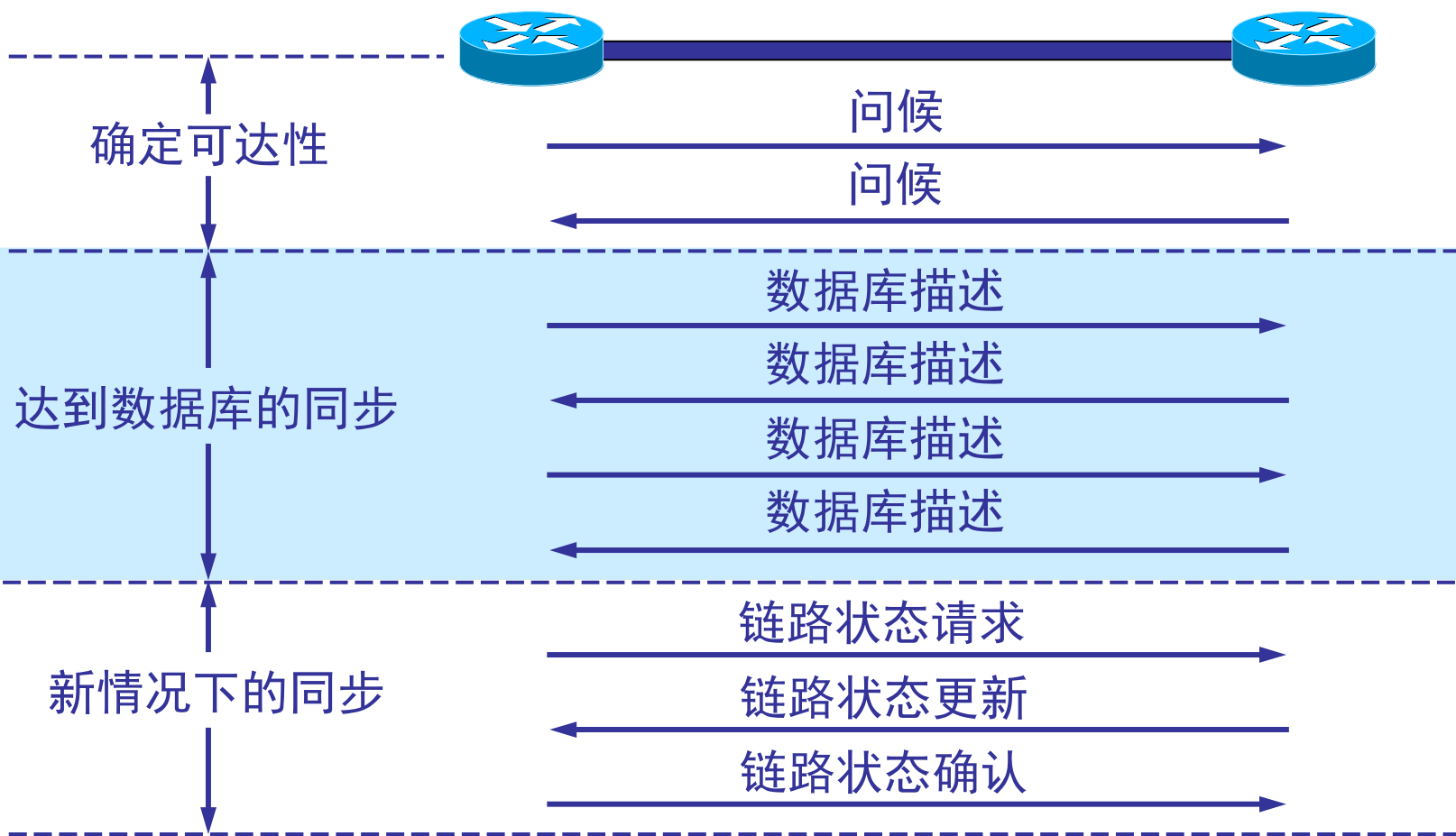
## 2. OSPF 的五种分组类型

---

- 类型 1 问候 (Hello) 分组。
- 类型 2 数据库描述 (Database Description) 分组。
- 类型 3 链路状态请求 (Link State Request) 分组。
- 类型 4 链路状态更新 (Link State Update) 分组，  
用洪泛法对全网更新链路状态。
- 类型 5 链路状态确认 (Link State Acknowledgment) 分组。



# OSPF 的基本操作





# OSPF 的其他特点

---

- OSPF 还规定每隔一段时间，如 30 分钟，要刷新一次数据库中的链路状态。
- 由于一个路由器的链路状态只涉及到与相邻路由器的连通状态，因而与整个互联网的规模并无直接关系。因此当互联网规模很大时，OSPF 协议要比距离向量协议 RIP 好得多。
- OSPF 没有“坏消息传播得慢”的问题，据统计，其响应网络变化的时间小于 100 ms。



# OSPF 和 RIP 的区别

---

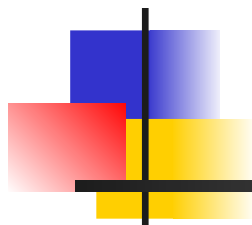
- OSPF 把路由选择信息广播给本自治系统中其他路由器而不仅仅是其邻接路由器，该路由信息是路由器到每个邻接点的链路状态。而 RIP 的发送到邻接点的公告包含了自治系统中所有网络节点的信息。
- OSPF 是一种动态的路由选择算法，即能自动而快速的适应拓扑结构的变化。而 RIP 不能采用动态方法（如延迟或负载）来选择路由。
- 只有当链路状态发生变化时，路由器才向所有路由器发送此信息。RIP 无论链路状态是否发生变化，定时交换路由表的信息。



## 4.5.4 外部网关协议 BGP

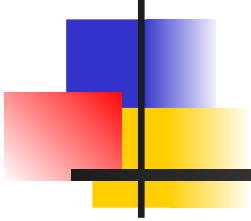
---

- BGP 是不同自治系统的路由器之间交换路由信息的协议。
- BGP 较新版本是 2006 年 1 月发表的 BGP-4（BGP 第 4 个版本），即 RFC 4271 ~ 4278。
- 可以将 BGP-4 简写为 BGP。



# BGP 使用的环境却不同

- (1) 因特网的规模太大，使得自治系统之间路由选择非常困难。
  - (2) 对于自治系统之间的路由选择，要寻找最佳路由是很不现实的。
  - (3) 自治系统之间的路由选择必须考虑有关策略
- 因此，边界网关协议 BGP 只能是力求寻找一条能够到达目的网络且**比较好的路由**（不能兜圈子），而**并非要寻找一条最佳路由**。



# BGP 发言人

---

- 每一个自治系统的管理员要选择至少一个路由器作为该自治系统的“BGP 发言人”。
- 一般说来，两个 BGP 发言人都是通过一个共享网络连接在一起的，而 BGP 发言人往往就是 BGP 边界路由器，但也可以不是 BGP 边界路由器。

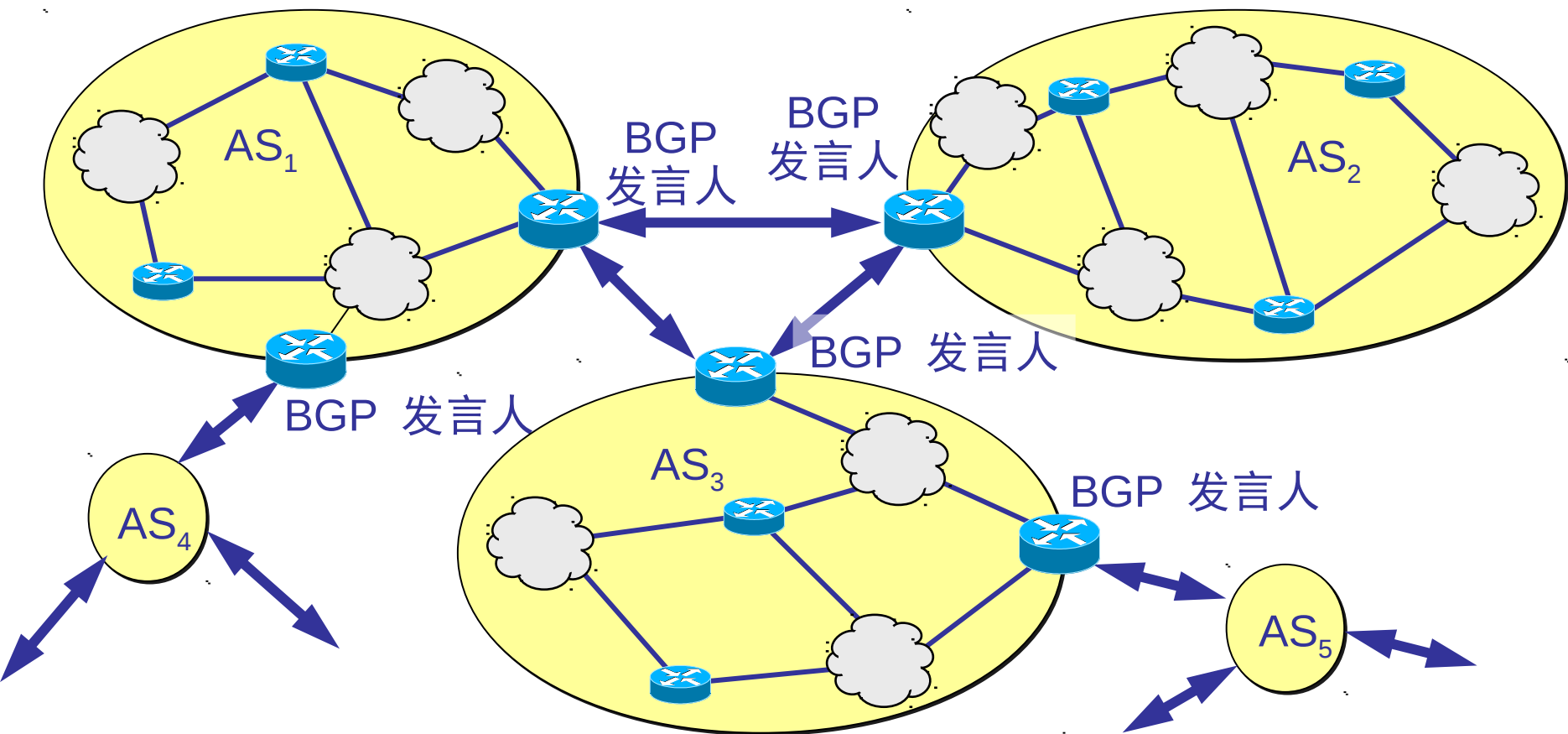


# BGP 交换路由信息

---

- 一个 BGP 发言人与其他自治系统中的 BGP 发言人要交换路由信息，就要**先建立 TCP 连接**，然后在此连接上交换 BGP 报文以建立 BGP **会话** (session)，利用 BGP 会话交换路由信息。
- 使用 TCP 连接能提供可靠的服务，也简化了路由选择协议。
- 使用 TCP 连接交换路由信息的两个 BGP 发言人，彼此成为对方的邻站或对等站。

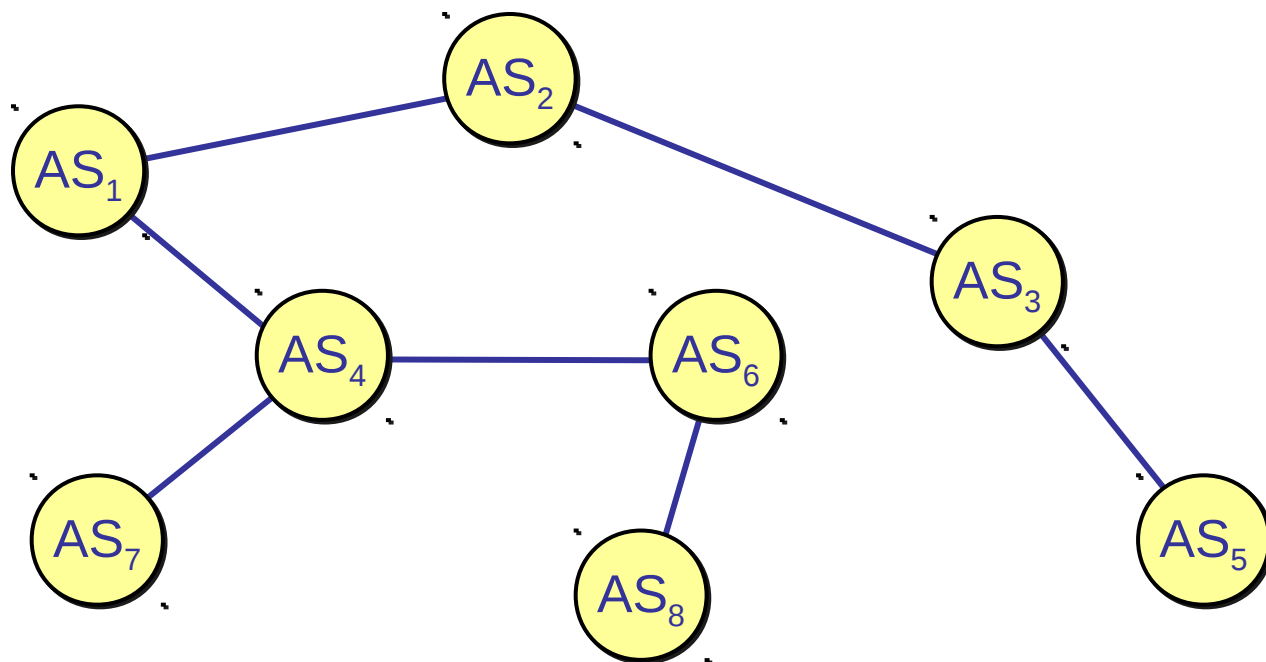
# BGP 发言人和 自治系统 AS 的关系





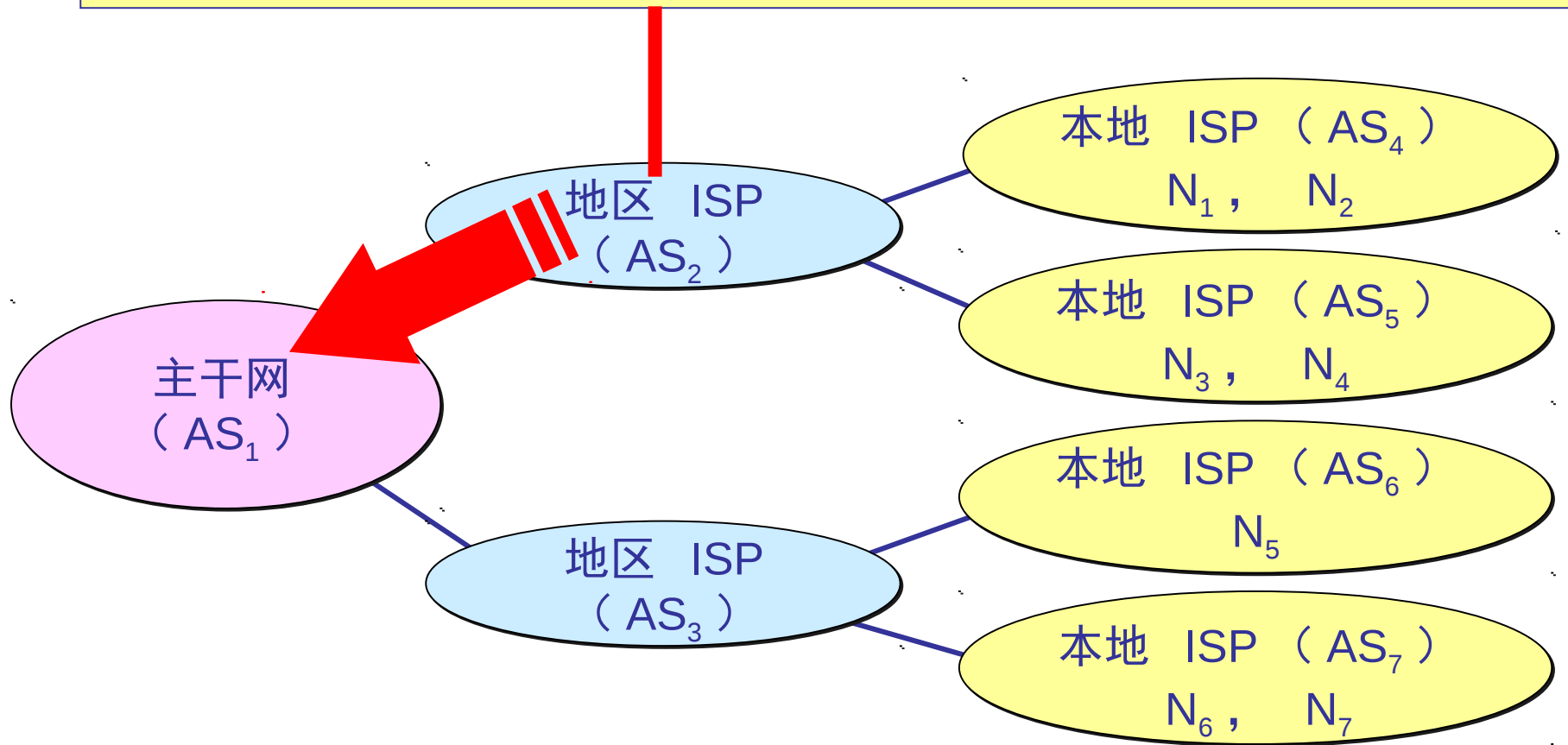
# 自治系统连通图

- BGP 发言人互相交换网络可达性的信息后，各 BGP 发言人就可找出到达各自自治系统的比较好的路由。



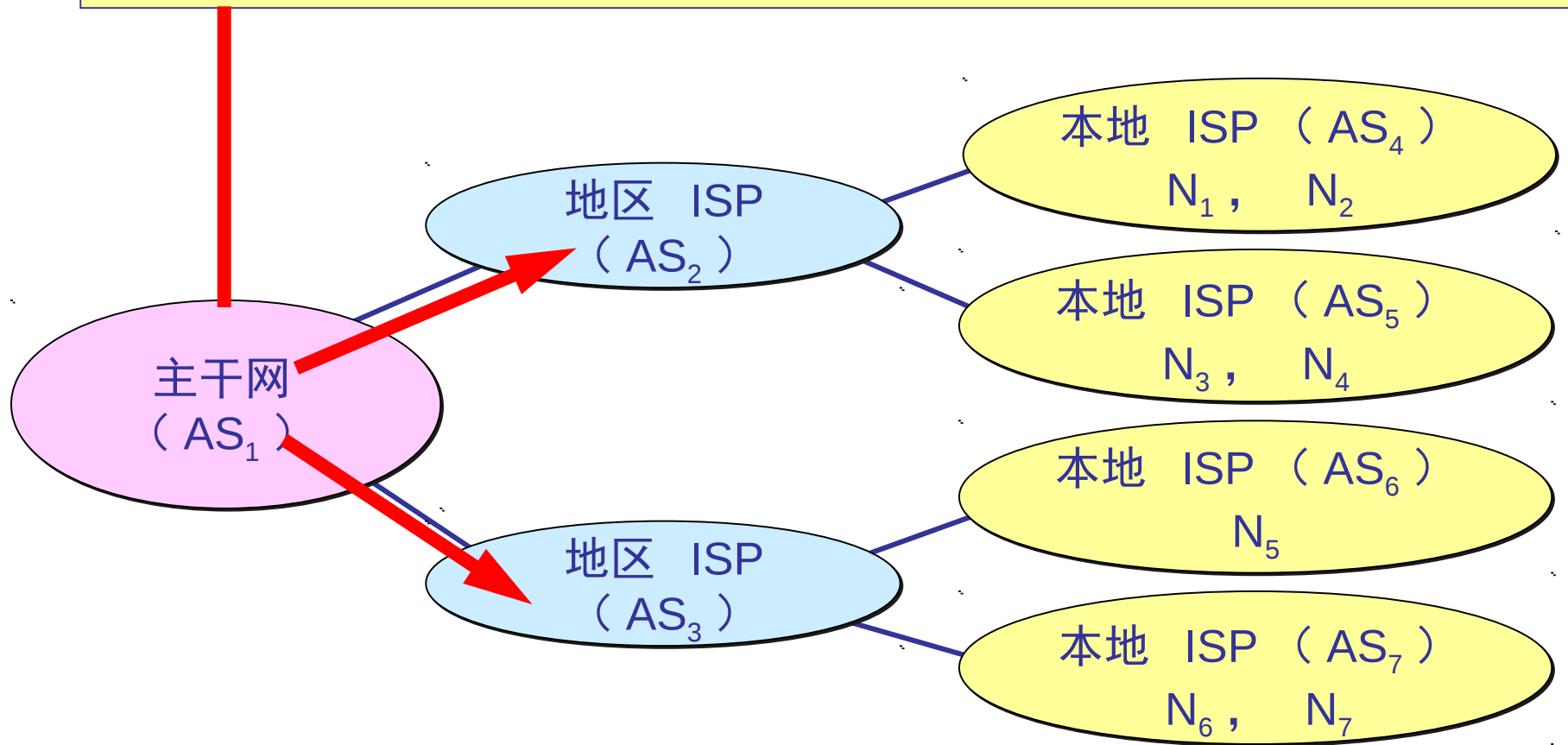
# BGP 发言人交换路径向量

自治系统  $AS_2$  的 BGP 发言人通知主干网的 BGP 发言人：  
“要到达网络  $N_1, N_2, N_3$  和  $N_4$  可经过  $AS_2$ 。”



# BGP 发言人交换路径向量

主干网还可发出通知：“要到达网络  $N_5, N_6$  和  $N_7$  可沿路径  $(AS_1, AS_3)$ 。”





# BGP 协议的特点 1

---

- BGP 协议交换路由信息的**结点数量级**是自治系统数的量级，这要比这些自治系统中的网络数少很多。
- 每一个自治系统中 BGP 发言人（或边界路由器）的数目是很少的。这样就使得自治系统之间的路由选择不致过分复杂。



# BGP 协议的特点 2

## ■ 算法

- BGP 以距离向量算法为基础，但它与 RIP 有很大的区别。
- BGP 不是保留到每一个目的站的费用（距离），而是保留到每一个目的站的完整路由。
- BGP 的路由表应当包括目的网络前缀、下一跳路由器，以及到达该目的网络所要经过的各个自治系统序列。
- 在 BGP 刚刚运行时，BGP 的邻站是交换整个的 BGP 路由表。但以后只需要在发生变化时更新有变化的部分



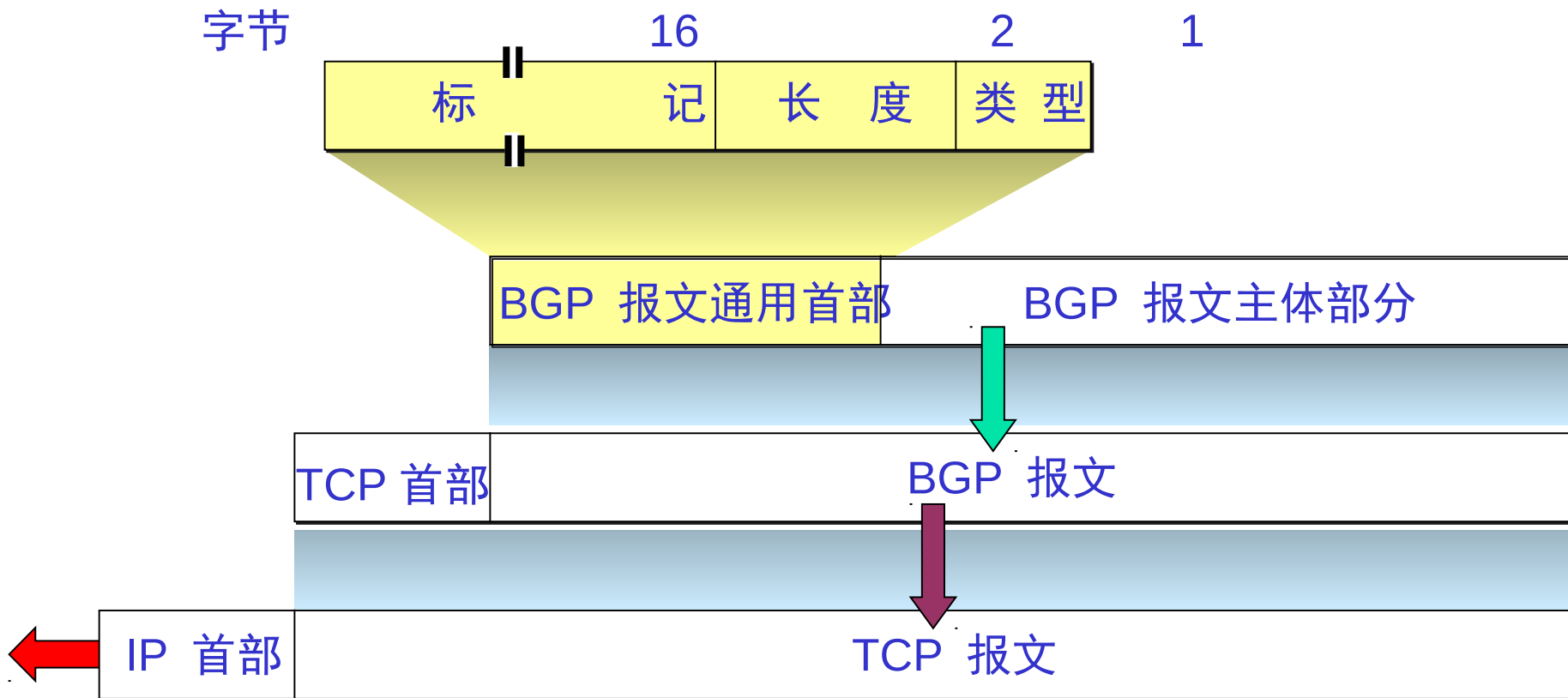
# BGP-4 共使用四种报文

---

- (1) 打开 (Open) 报文，用来与相邻的另一个 BGP 发言人建立关系。
- (2) 更新 (Update) 报文，用来发送某一路由的信息，以及列出要撤消的多条路由。
- (3) 保活 (Keepalive) 报文，用来确认打开报文和周期性地证实邻站关系。
- (3) 通知 (Notificaton) 报文，用来发送检测到的差错。

# BGP 报文具有通用的首部

字节





## 4.5.6 路由器在网际互连中的作用

### 1. 路由器的结构

- 路由器是一种具有多个输入端口和多个输出端口的**专用计算机**，其任务是转发分组。也就是说，将路由器某个输入端口收到的分组，按照分组要去的目的地（即目的网络），把该分组从路由器的某个合适的输出端口转发给下一跳路由器。
- 下一跳路由器也按照这种方法处理分组，直到该分组到达终点为止。





# 典型的路由器的结构

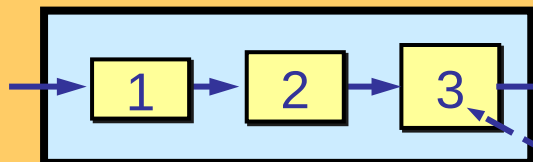
3——网络层  
2——数据链路层  
1——物理层

路由选择处理机

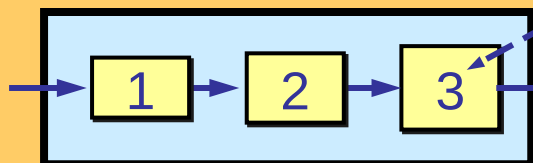
路由选择协议

路由表

输入端口



输入端口

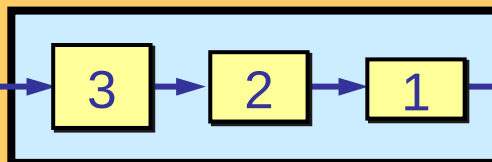


分组处理

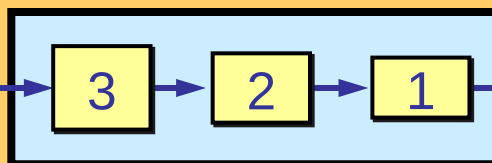
转发表

交换结构

输出端口



输出端口



路由选择

分组转发



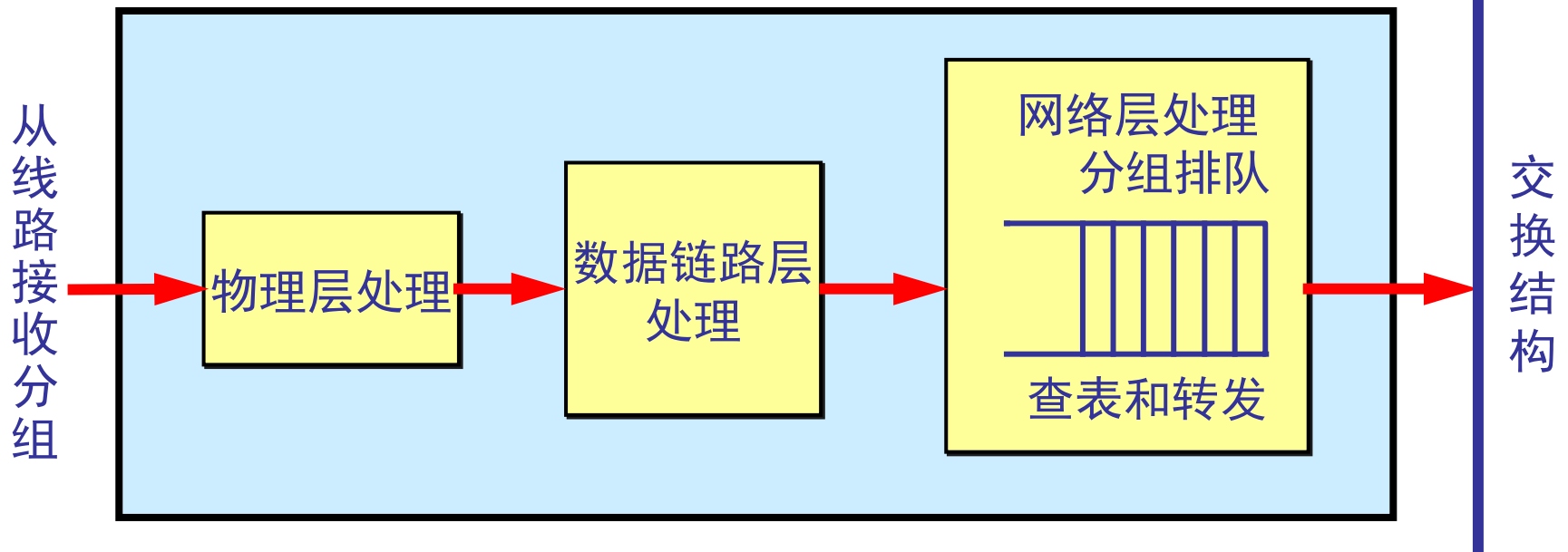
# “转发”和“路由选择” 的区别

- “**转发**” (forwarding) 就是路由器根据转发表将用户的 IP 数据报从合适的端口转发出去。
- “**路由选择**” (routing) 则是按照分布式算法，根据从各相邻路由器得到的关于网络拓扑的变化情况，动态地改变所选择的路由。
- **路由表**是根据路由选择算法得出的。而**转发表**是从路由表得出的。
- 在讨论路由选择的原理时，往往不去区分转发表和路由表的区别。

# 输入端口对线路上收到的分组的处理

- 数据链路层剥去帧首部和尾部后，将分组送到网络层的队列中排队等待处理。这会产生一定的时延。

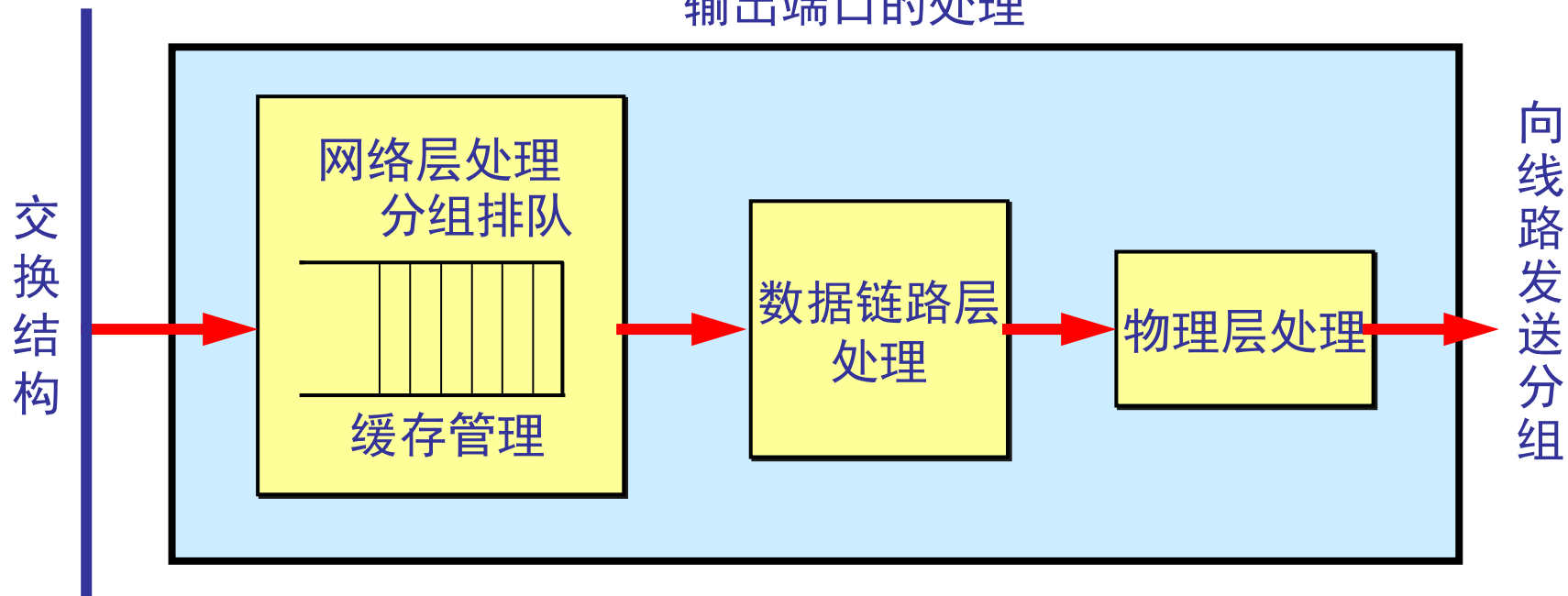
输入端口的处理



# 输出端口将交换结构传送来的分组发送到线路

- 当交换结构传送过来的分组先进行缓存。数据链路层处理模块将分组加上链路层的首部和尾部，交给物理层后发送到外部线路。

输出端口的处理



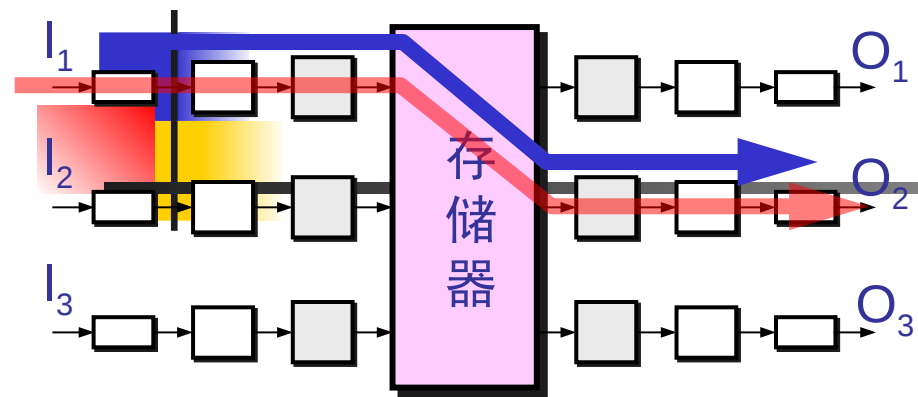


# 分组丢弃

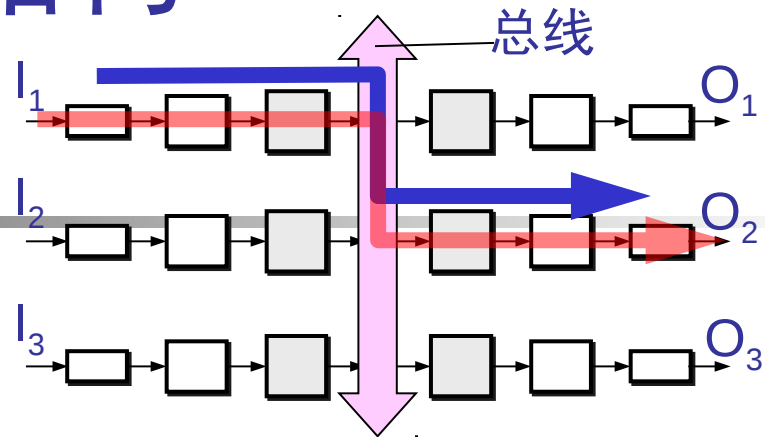
---

- 若路由器处理分组的速率赶不上分组进入队列的速率，则队列的存储空间最终必定减少到零，这就使后面再进入队列的分组由于**没有存储空间而只能被丢弃**。
- 路由器中的输入或输出队列产生**溢出**是造成分组丢失的重要原因。

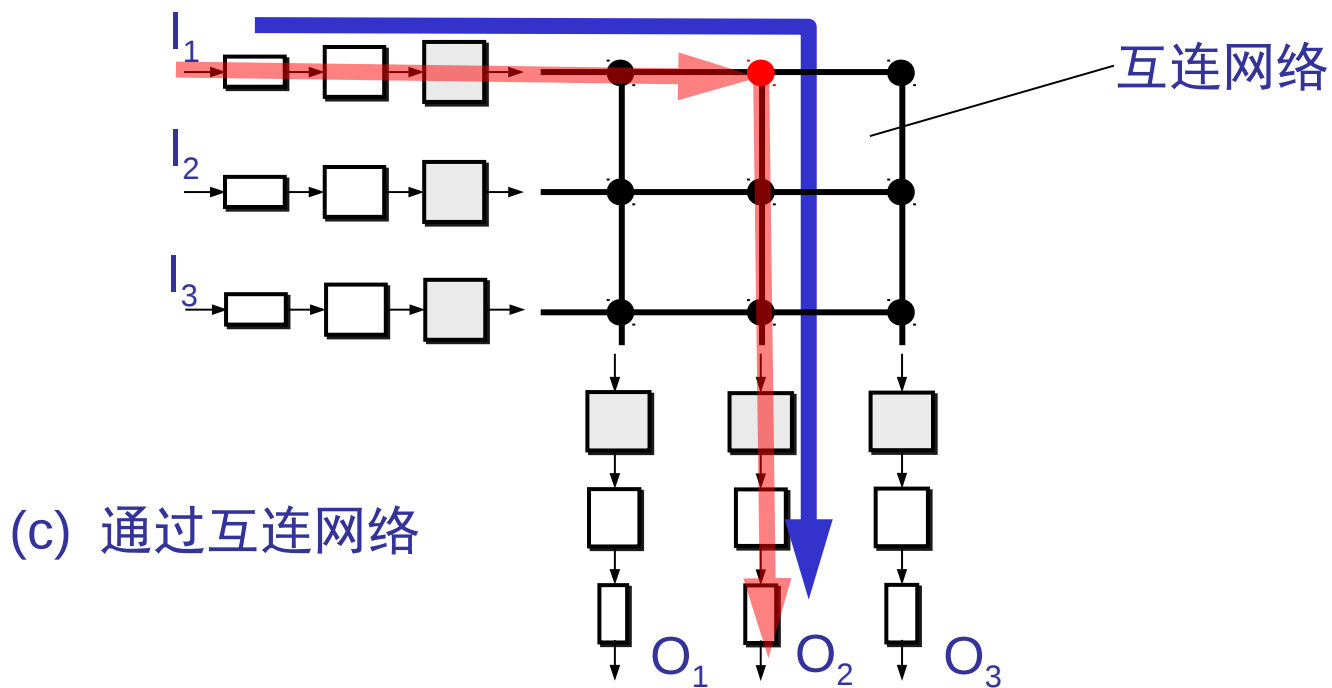
## 2 交换结构



(a) 通过存储器



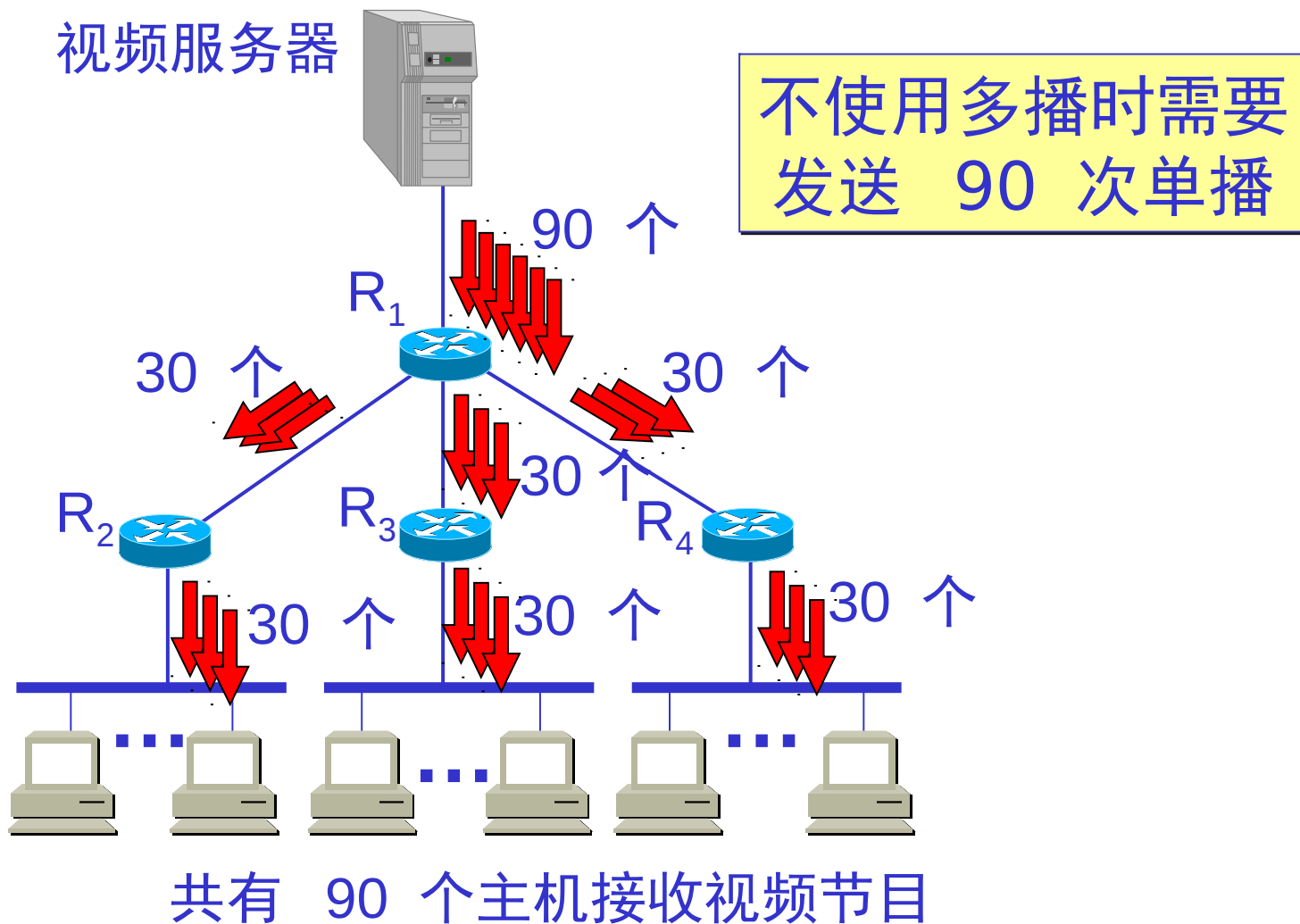
(b) 通过总线



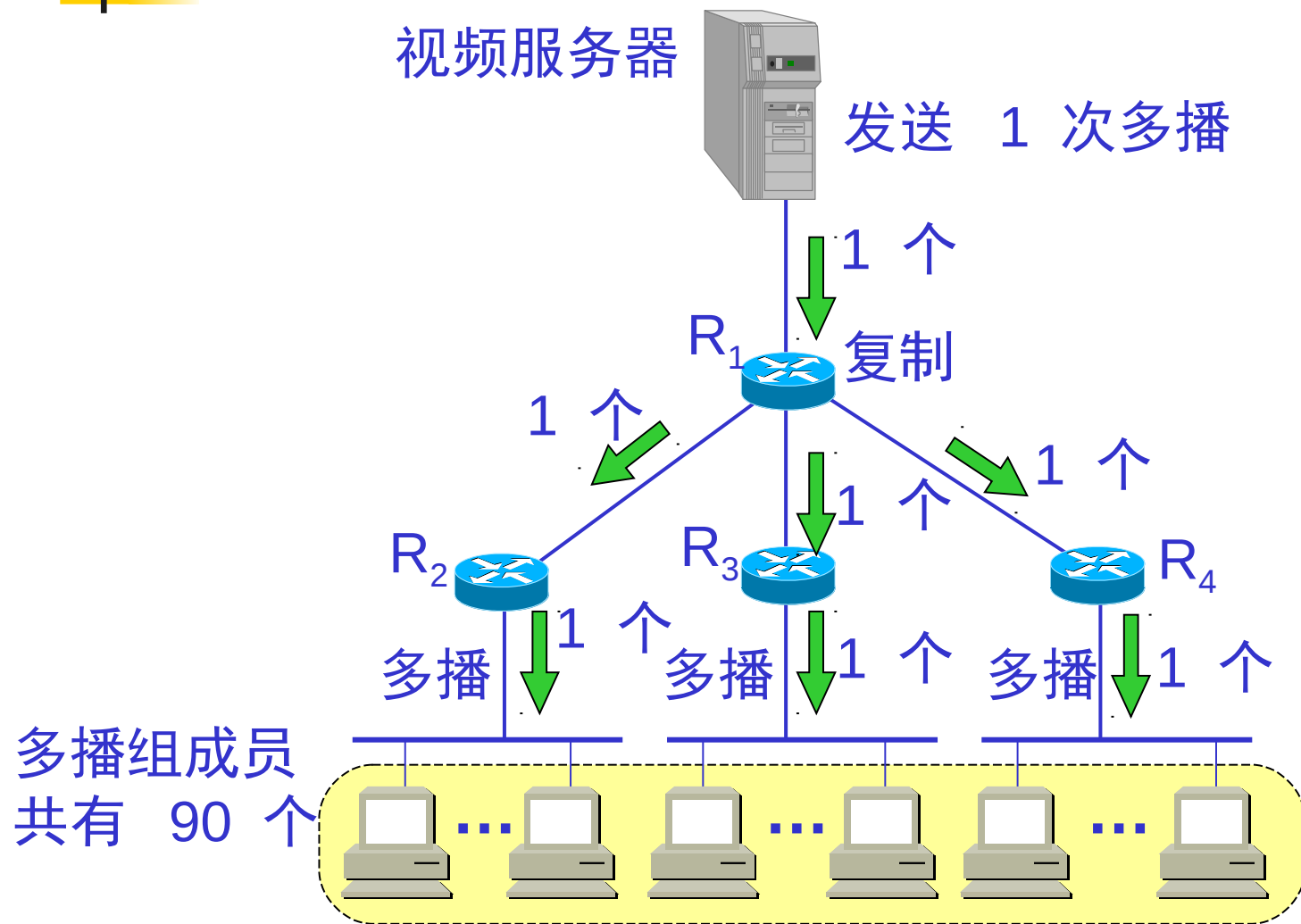
(c) 通过互连网络

## 4.6 IP 多播

### 4.6.1 IP 多播的基本概念



# 多播可明显地减少 网络中资源的消耗







# IP 多播的一些特点

---

- (1) 多播使用组地址—— IP 使用 D 类地址支持多播。多播地址只能用于目的地址，而不能用于源地址。
- (2) 永久组地址——由因特网号码指派管理局 IANA 负责指派。
- (3) 动态的组成员
- (4) 使用硬件进行多播

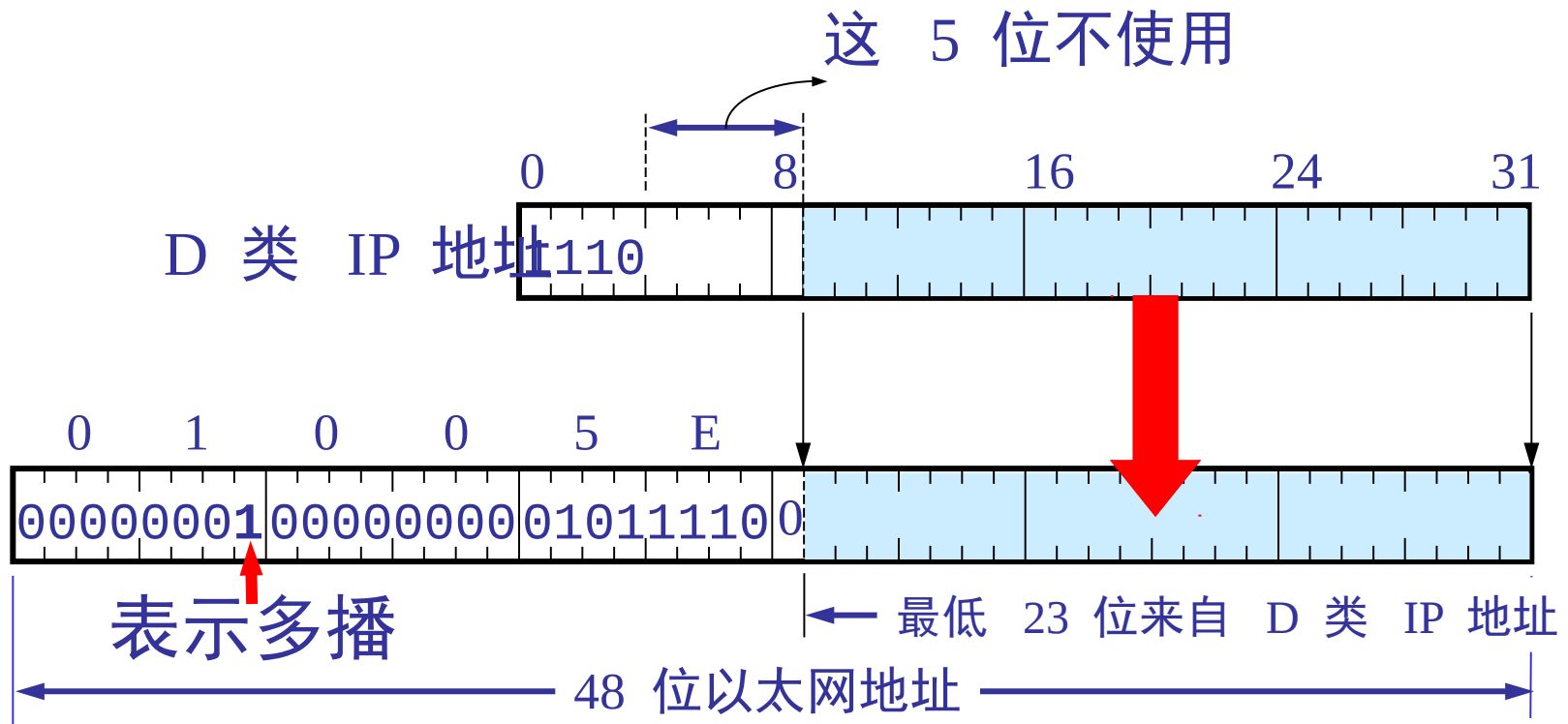


## 4.6.2 在局域网上进行硬件多播

---

- 因特网号码指派管理局 IANA 拥有的以太网地址块的高 24 位为 00-00-5E。
- 因此 TCP/IP 协议使用的以太网多播地址块的范围是：从 01-00-5E-00-00-00  
到 01-00-5E-FF-FF-FF
- D 类 IP 地址可供分配的有 28 位，在这 28 位中的前 5 位不能用来构成以太网硬件地址。

# D 类 IP 地址 与以太网多播地址的映射关系





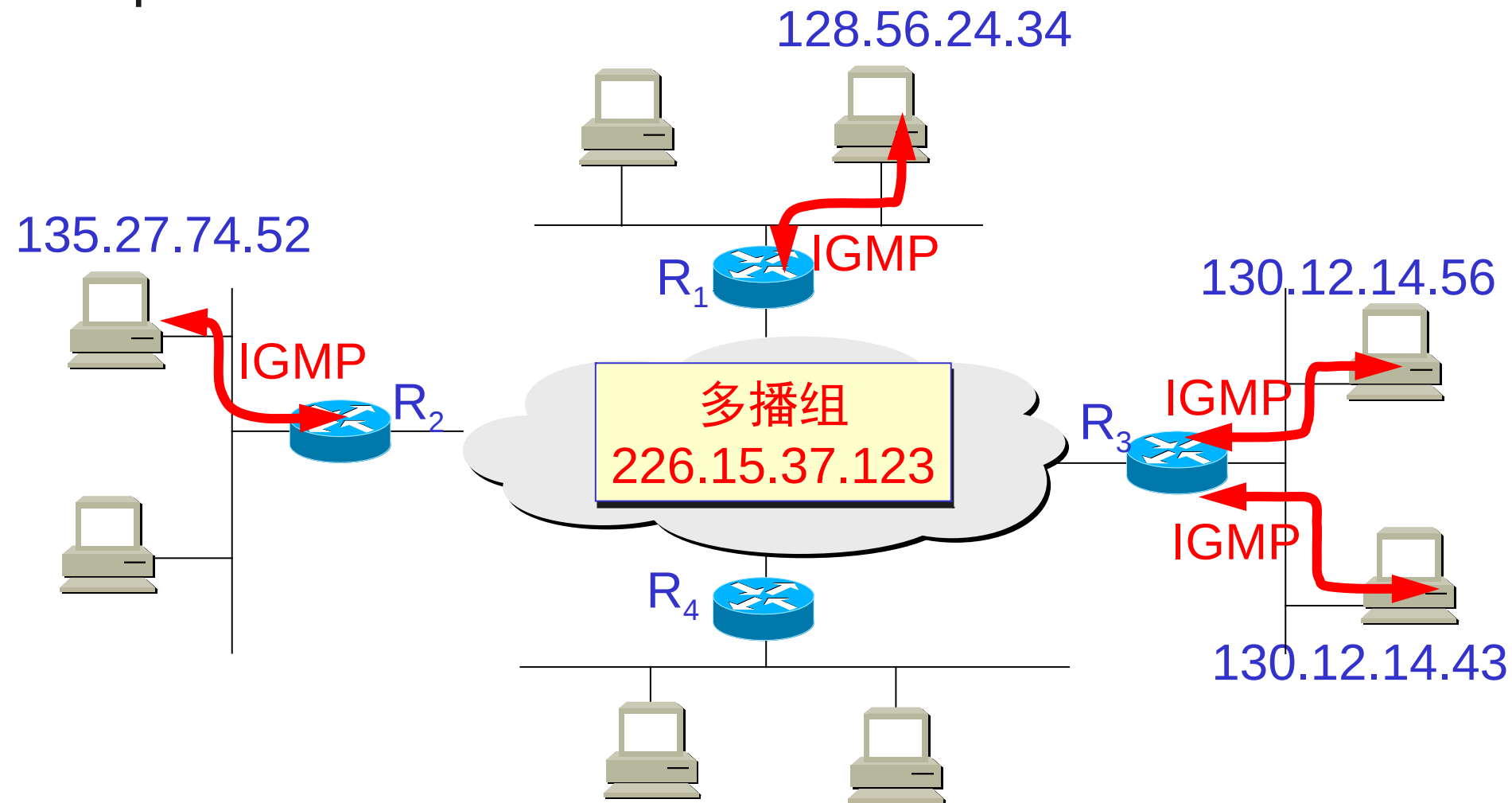
## 4.6.3 网际组管理协议 IGMP 和多播路由选择协议

---

### 1. IP 多播需要两种协议

- 为了使路由器知道多播组成员的信息，需要利用网际组管理协议 IGMP (Internet Group Management Protocol)。
- 连接在局域网上的多播路由器还必须和因特网上的其他多播路由器协同工作，以便把多播数据报用最小代价传送给所有的组成员。这就需要使用多播路由选择协议。

# IGMP 使多播路由器 知道多播组成员信息

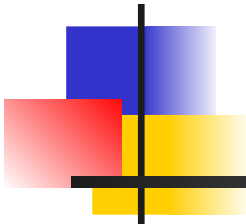




# IGMP 的本地使用范围

---

- IGMP 并非在因特网范围内对所有多播组成员进行管理的协议。
- IGMP 不知道 IP 多播组包含的成员数，也不知道这些成员都分布在哪些网络上。
- IGMP 协议是让连接在本地局域网上的多播路由器知道本局域网上是否有主机（严格讲，是主机上的某个进程）参加或退出了某个多播组。



# 多播路由选择协议

## 比单播路由选择协议复杂得多

- 多播转发必须**动态地**适应多播组成员的变化（这时网络拓扑并未发生变化）。请注意，单播路由选择通常是在网络拓扑发生变化时才需要更新路由。
- 多播路由器在转发多播数据报时，不能仅仅根据多播数据报中的目的地址，而是还要考虑这个多播数据报从什么地方来和要到什么地方去。
- 多播数据报可以由没有加入多播组的主机发出，也可以通过没有组成员接入的网络。



## 2. 网际组管理协议 IGMP

---

- 1989 年公布的 RFC 1112 （ IGMPv1 ） 早已成为了因特网的标准协议。
- 1997 年公布的 RFC 2236 （ IGMPv2 ， 建议标准 ） 对 IGMPv1 进行了更新。
- 2002 年 10 月公布了 RFC 3376 （ IGMPv3 ， 建议标准 ）， 宣布 RFC 2236 （ IGMPv2 ） 是陈旧的。





# IGMP 是整个网际协议 IP 的一个组成部分

---

- 和 ICMP 相似，IGMP 使用 IP 数据报传递其报文（即 IGMP 报文加上 IP 首部构成 IP 数据报），但它也向 IP 提供服务。
- 因此，我们不把 IGMP 看成是一个单独的协议，而是属于整个网际协议 IP 的一个组成部分。



# IGMP 可分为两个阶段

---

- 第一阶段：当某个主机加入新的多播组时，该主机应向多播组的多播地址发送 IGMP 报文，声明自己要成为该组的成员。本地的多播路由器收到 IGMP 报文后，将组成员关系转发给因特网上的其他多播路由器。



# IGMP 可分为两个阶段

---

- 第二阶段：因为组成员关系是**动态**的，因此本地多播路由器要**周期性地探询**本地局域网上的主机，以便知道这些主机是否还继续是组的成员。
- 只要对某个组有一个主机响应，那么多播路由器就认为这个组是活跃的。
- 但一个组在经过几次的探询后仍然没有一个主机响应，则不再将该组的成员关系转发给其他的多播路由器。



# IGMP 采用的一些具体措施

- ① 在主机和多播路由器之间的所有通信都是使用 IP 多播。
- ② 多播路由器在探询组成员关系时，只需要对所有的组发送一个请求信息的询问报文，而不需要对每一个组发送一个询问报文。默认的询问速率是每 125 秒发送一次。
- ③ 当同一个网络上连接有几个多播路由器时，它们能够迅速和有效地选择其中的一个来探询主机的成员关系。



# IGMP 采用的一些具体措施 (续)

---

- ④ 在 IGMP 的询问报文中有一个数值  $N$ ，它指明一个最长响应时间（默认值为 10 秒）。当收到询问时，主机在 0 到  $N$  之间随机选择发送响应所需经过的时延。对应于最小时延的响应最先发送。
- ⑤ 同一个组内的每一个主机都要监听响应，只要有本组的其他主机先发送了响应，自己就可以不再发送响应了。



### 3. 多播路由选择

---

- 多播路由选择协议尚未标准化。
- 一个多播组中的成员是动态变化的，随时会有主机加入或离开这个多播组。
- 多播路由选择实际上就是要找出以源主机为根结点的多播转发树。
- 在多播转发树上的路由器不会收到重复的多播数据报。
- 对不同的多播组对应于不同的多播转发树。同一个多播组，对不同的源点也会有不同的多播转发树。



# 转发多播数据报使用的方法

## (1) 洪泛与剪除

- 这种方法适合于较小的多播组，而所有的组成员接入的局域网也是相邻接的。
- 一开始，路由器转发多播数据报使用洪泛的方法（这就是广播）。为了避免兜圈子，采用了叫做**反向路径广播 RPB** (Reverse Path Broadcasting) 的策略。



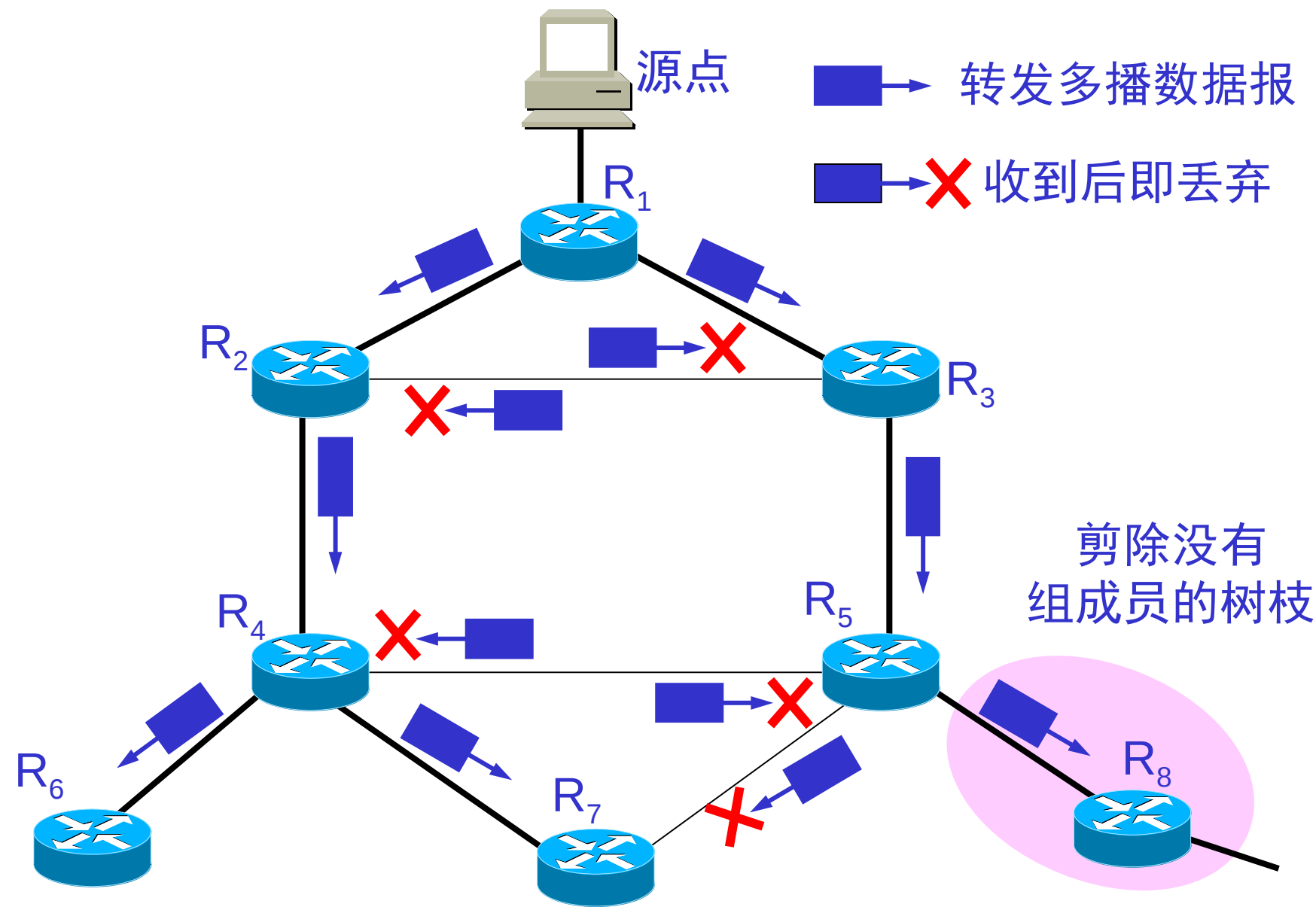
# RPB 的要点

---

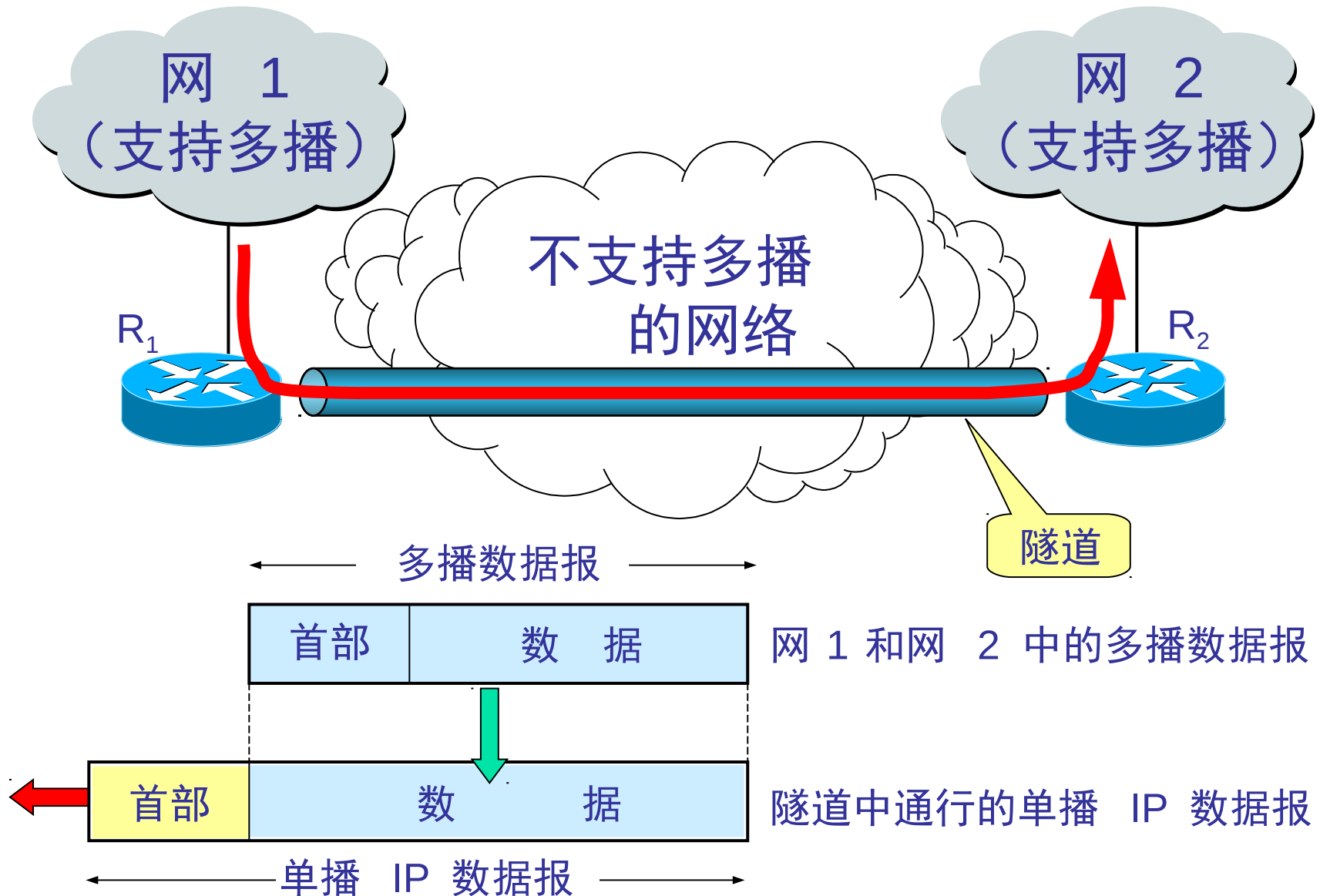
- 路由器收到多播数据报时，先检查是否从源点经最短路径传送来的。
- 若是，就向所有其他方向转发刚才收到的多播数据报（但进入的方向除外），否则就丢弃而不转发。
- 如果存在几条同样长度的最短路径），那么只能选择一条最短路径，选择的准则就是看这几条最短路径中的相邻路由器谁的IP 地址最小。



# 反向路径广播 RPB 和剪除



## (2) 隧道技术 (tunneling)





### (3) 基于核心的发现技术

---

- 这种方法对于多播组的大小在较大范围内变化时都适合。
- 这种方法是对每一个多播组  $G$  指定一个核心 (core) 路由器，给出它的 IP 单播地址。
- 核心路由器按照前面讲过的方法创建出对应于多播组  $G$  的转发树。



# 几种多播路由选择协议

---

- 距离向量多播路由选择协议 DVMRP (Distance Vector Multicast Routing Protocol)
- 基于核心的转发树 CBT (Core Based Tree)
- 开放最短通路优先的多播扩展 MOSPF (Multicast Extensions to OSPF)
- 协议无关多播 - 稀疏方式 PIM-SM (Protocol Independent Multicast-Sparse Mode)
- 协议无关多播 - 密集方式 PIM-DM (Protocol Independent Multicast-Dense Mode)

## 4.7 虚拟专用网 VPN 和网络地址转换 NAT

### 4.7.1 虚拟专用网 VPN

---

- **本地地址**——仅在机构内部使用的 IP 地址，可以由本机构自行分配，而不需要向因特网的管理机构申请。
- **全球地址**——全球唯一的 IP 地址，必须向因特网的管理机构申请。

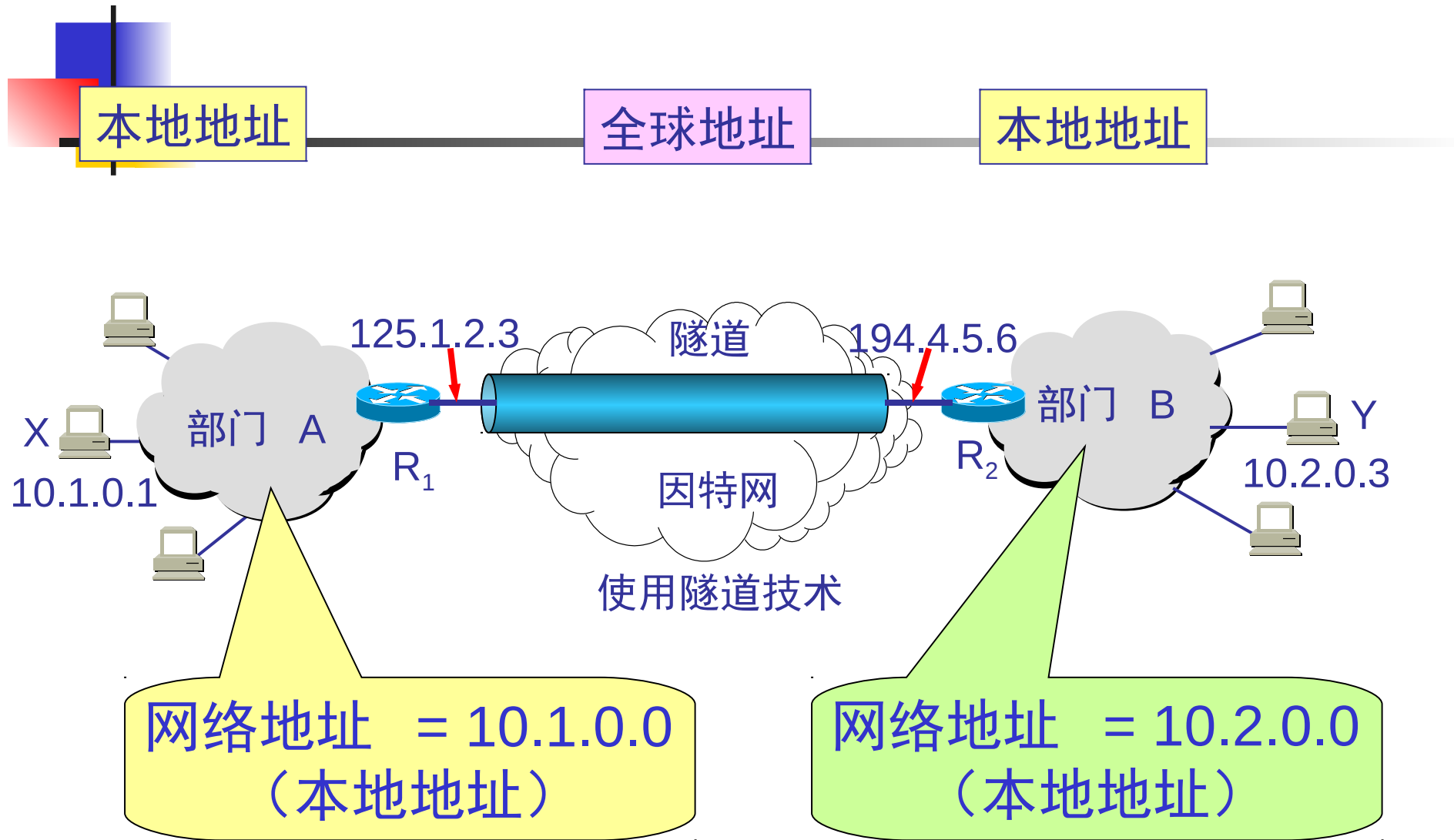


# RFC 1918 指明的专用地址 (private address)

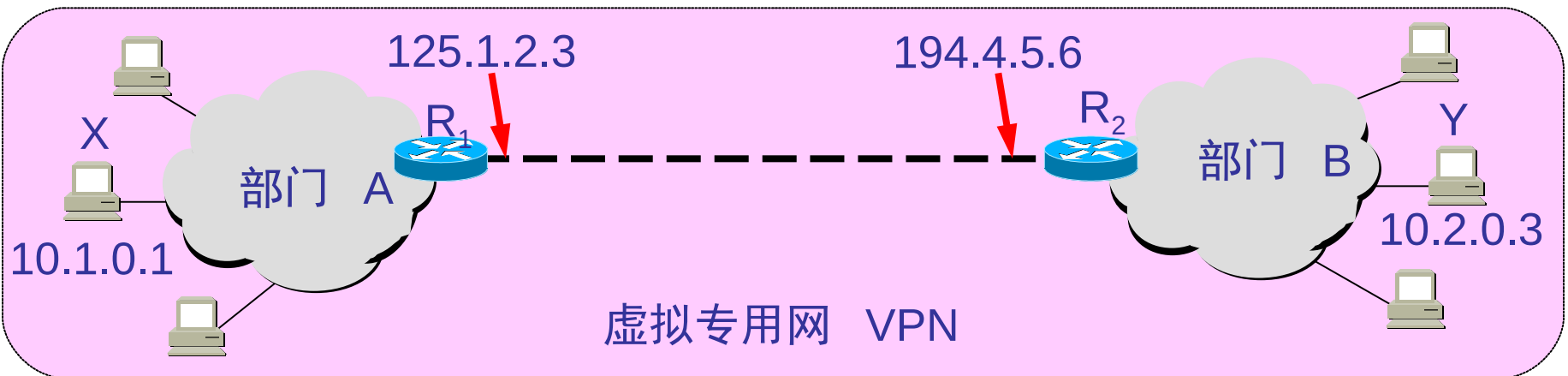
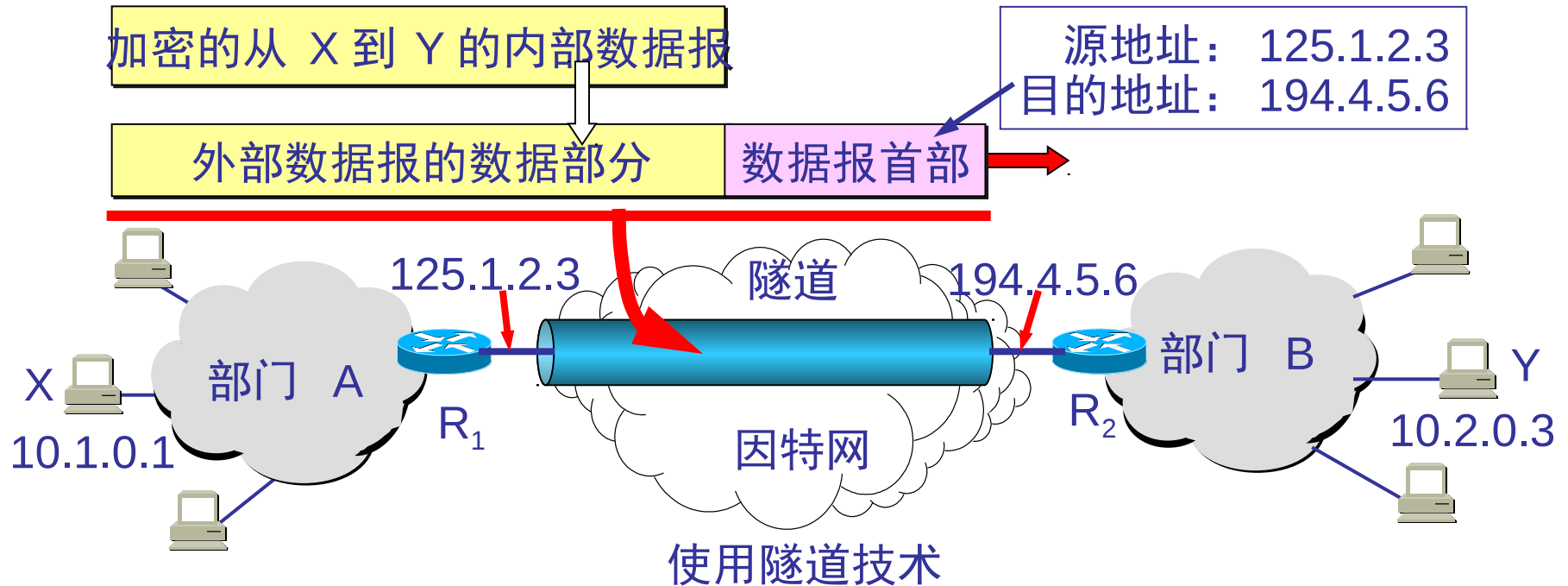
---

- 10.0.0.0 到 10.255.255.255
- 172.16.0.0 到 172.31.255.255
- 192.168.0.0 到 192.168.255.255
- 这些地址只能用于一个机构的内部通信，而不能用于和因特网上的主机通信。
- 专用地址只能用作本地地址而不能用作全球地址。在因特网中的所有路由器对目的地址是专用地址的数据报一律不进行转发。

# 用隧道技术实现虚拟专用网



# 用隧道技术实现虚拟专用网

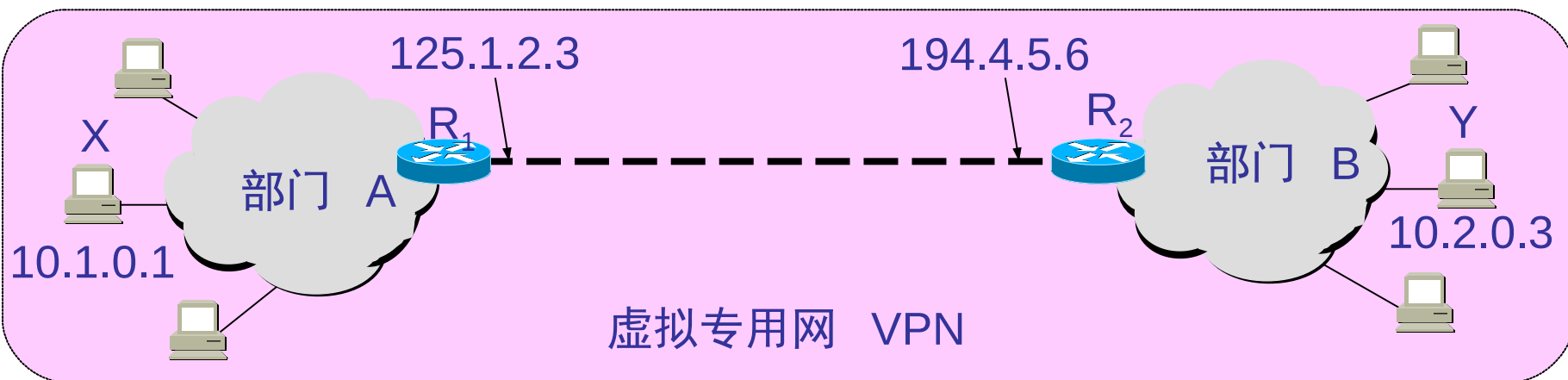




# 内联网 intranet 和外联网 extranet

(都是基于 TCP/IP 协议)

- 由部门 A 和 B 的内部网络所构成的虚拟专用网 VPN 又称为**内联网** (intranet)，表示部门 A 和 B 都是在**同一个**机构的内部。
- 一个机构和某些**外部机构**共同建立的虚拟专用网 VPN 又称为**外联网** (extranet)。





# 远程接入 VPN

## (remote access VPN)

---

- 有的公司可能没有分布在不同场所的部门，但有很多流动员工在外地工作。公司需要和他们保持联系，远程接入 VPN 可满足这种需求。
- 在外地工作的员工拨号接入因特网，而驻留在员工 PC 机中的 VPN 软件可在员工的 PC 机和公司的主机之间建立 VPN 隧道，因而外地员工与公司通信的内容是保密的，员工们感到好像就是使用公司内部的本地区域网络。



## 4.7.2 网络地址转换 NAT

### (Network Address Translation)

---

- 网络地址转换 NAT 方法于 1994 年提出。
- 需要在专用网连接到因特网的路由器上安装 NAT 软件。装有 NAT 软件的路由器叫做 NAT 路由器，它至少有一个有效的外部全球地址  $IP_G$ 。
- 所有使用本地地址的主机在和外界通信时都要在 NAT 路由器上将其本地地址转换成  $IP_G$  才能和因特网连接。



# 网络地址转换的过程

- 内部主机  $X$  用本地地址  $IP_X$  和因特网上主机  $Y$  通信所发送的数据报必须经过 NAT 路由器。
- NAT 路由器将数据报的源地址  $IP_X$  转换成全球地址  $IP_G$ ，但目的地址  $IP_Y$  保持不变，然后发送到因特网。
- NAT 路由器收到主机  $Y$  发回的数据报时，知道数据报中的源地址是  $IP_Y$  而目的地址是  $IP_G$ 。
- 根据 NAT 转换表，NAT 路由器将目的地址  $IP_G$  转换为  $IP_X$ ，转发给最终的内部主机  $X$ 。