



# 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
- 4.3 划分子网和构造超网
- 4.4 网际控制报文协议 ICMP
- 4.5 互联网的路由选择协议



## 4.1 网络层提供的两种服务

---

- 虚电路服务
- 数据报服务

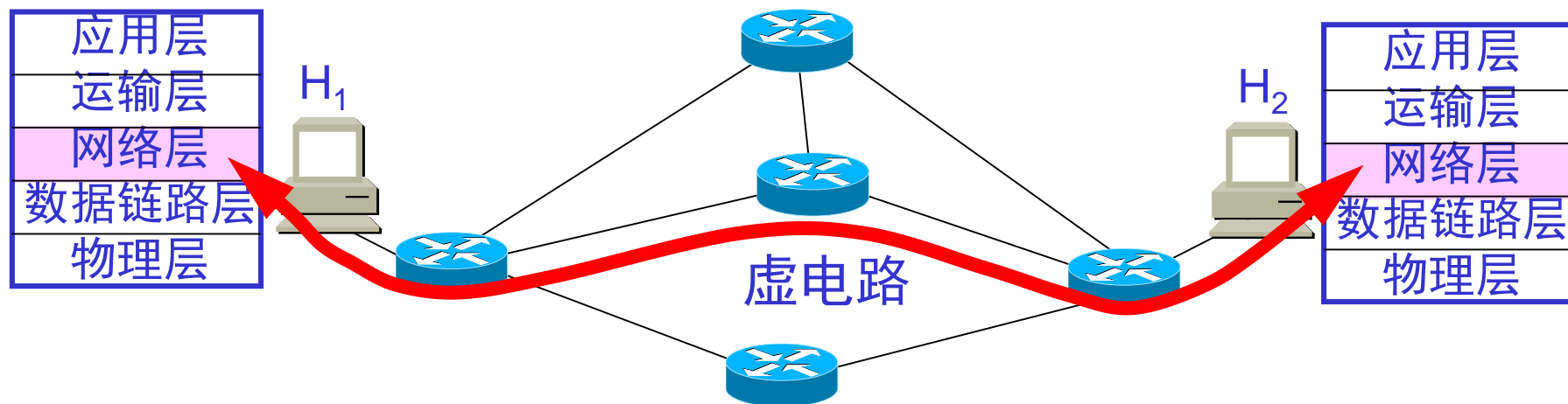


# 虚电路服务

---

- 面向连接的通信方式
- 建立虚电路(Virtual Circuit)，以保证双方通信所需的一切网络资源。
- 如果再使用可靠传输的网络协议，就可使所发送的分组无差错按序到达终点。

# 虚电路服务



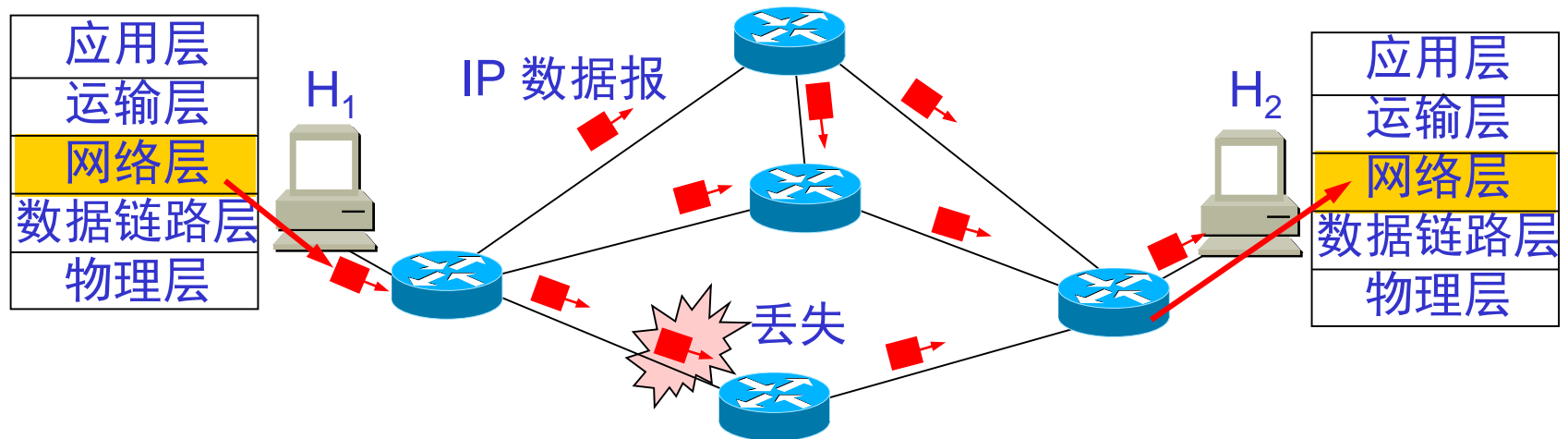
$H_1$  发送给  $H_2$  的所有分组都沿着同一条虚电路传送



# 互联网采用的设计思路

- 网络层向上只提供简单灵活的、**无连接的、尽最大努力交付的数据报服务**。
- 网络在发送分组时不需要先建立连接。每一个分组（即 IP 数据报）独立发送，与其前后的分组无关（不进行编号）。
- 网络层不提供服务质量的承诺。即所传送的分组可能出错、丢失、重复和失序（不按序到达终点），当然也不保证分组传送的时限。

# 数据报服务



$H_1$  发送给  $H_2$  的分组可能沿着不同路径传送

# 虚电路服务与数据报服务的对比

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	可靠通信应当由用户主机来保证
连接的建立	必须有	不需要
终点地址	仅在连接建立阶段使用，每个分组使用短的虚电路号	每个分组都有终点的完整地址
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进行转发
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	出故障的结点可能会丢失分组，一些路由可能会发生变化
分组的顺序	总是按发送顺序到达终点	到达终点时不一定按发送顺序
端到端的差错处理和流量控制	可以由网络负责，也可以由用户主机负责	由用户主机负责

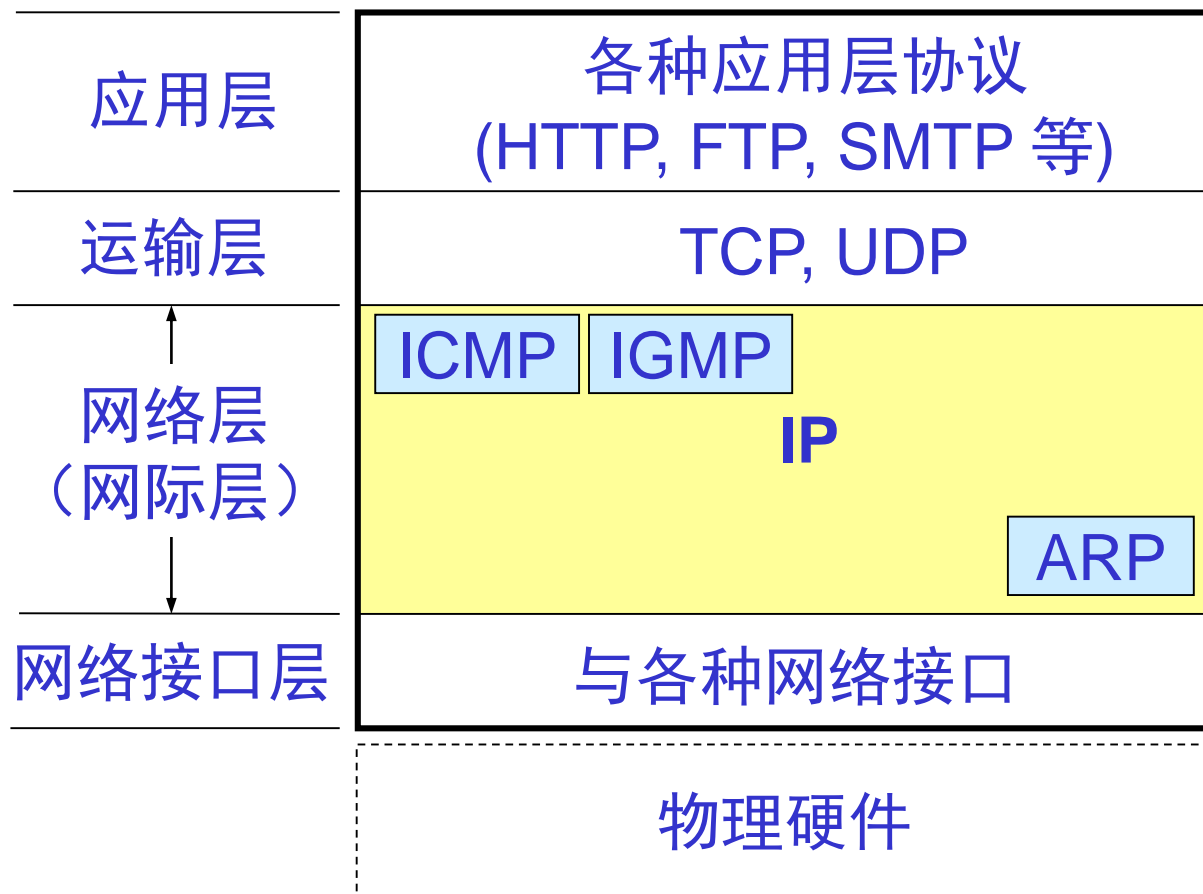




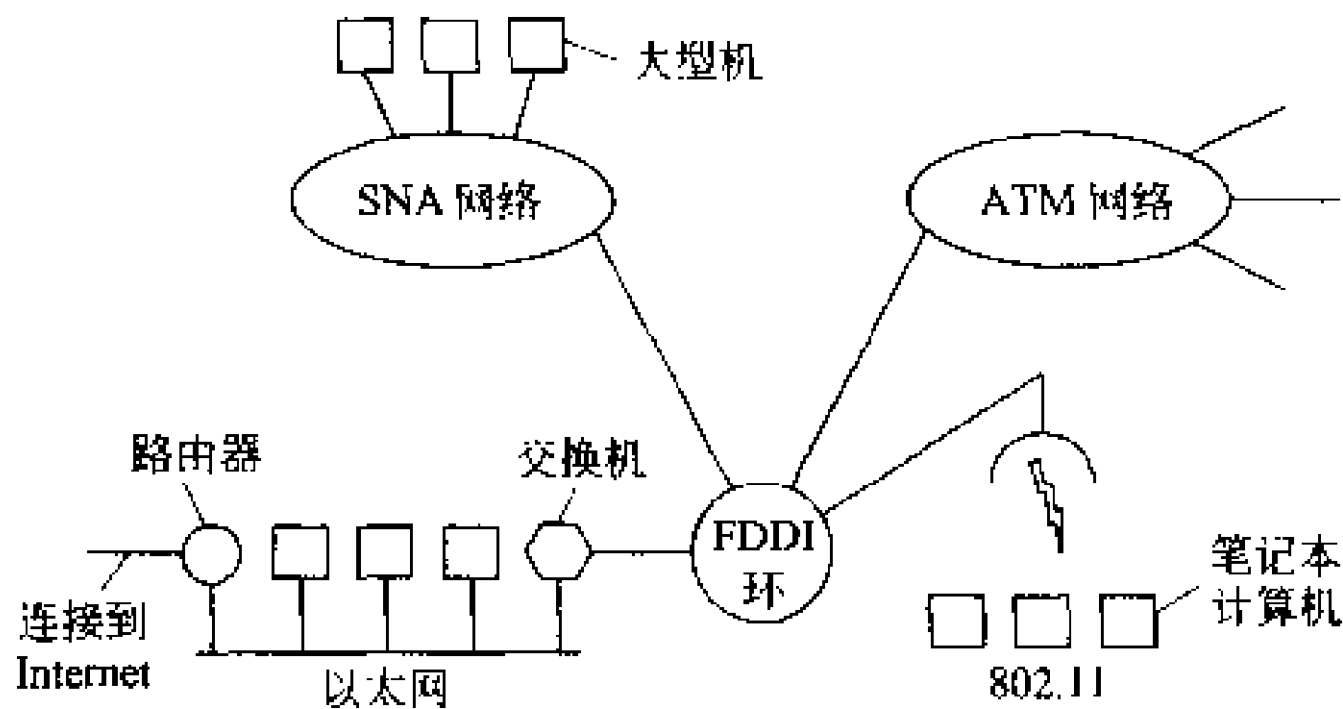
## 4.2 网际协议IP

- 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



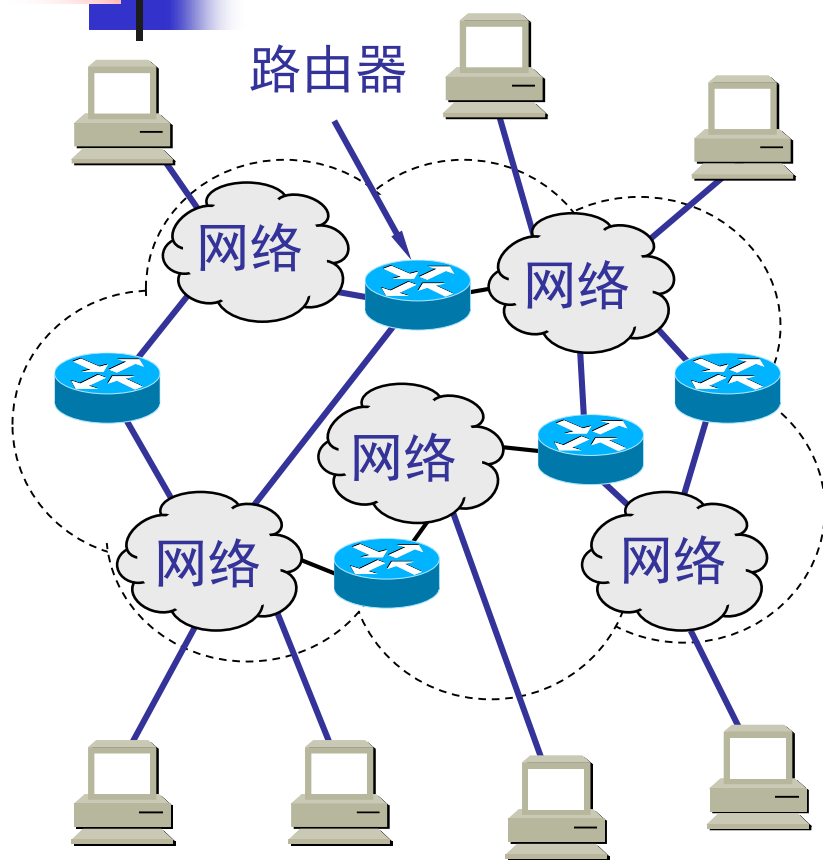


# 网络互相连接起来 要使用一些中间设备

---

- 中间设备又称为中间系统或中继(relay)系统。
  - 物理层中间系统：转发器(repeater)。
  - 数据链路层中间系统：网桥或桥接器(bridge)。
  - 网络层中间系统：路由器(router)。
  - 网络层以上的中间系统：网关(gateway)。

# 互连网络与虚拟互连网络



(a) 互连网络



(b) 虚拟互连网络

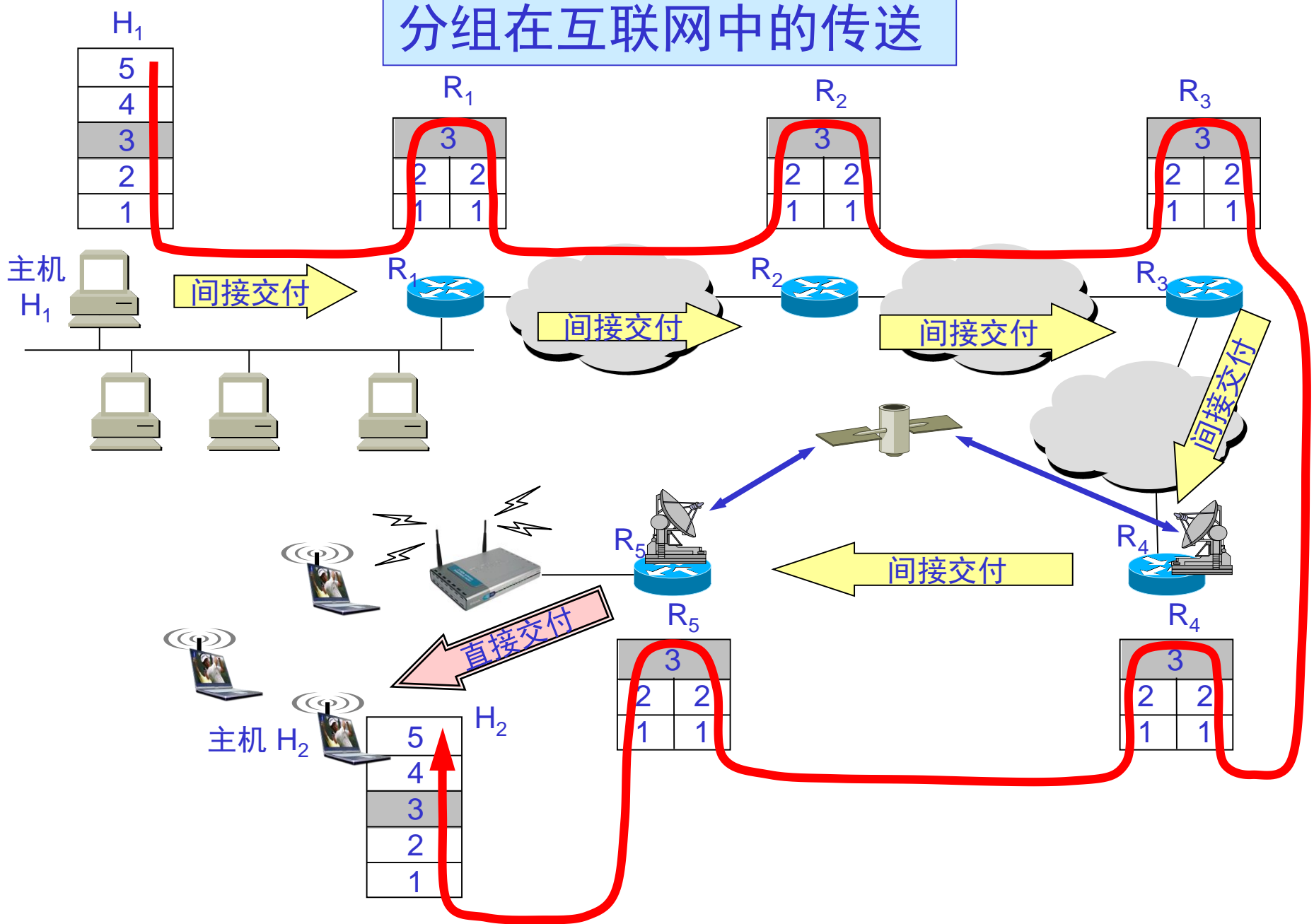


# 虚拟互连网络的意义

---

- 所谓虚拟互连网络(简称为 IP 网)，意思就是互连起来的各种物理网络存在异构性，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。
- 当互联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



## 4.2.2 分类的 IP 地址

### 1. IP 地址及其表示方法

- IP 地址就是给每个连接在互联网上的主机（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地址现在由互联网名字与号码指派公司ICANN (Internet Corporation for Assigned Names and Numbers)进行分配





# IP 地址的编址方法

---

- **分类的 IP 地址**。最基本的编址方法，在 1981 年就通过了相应的标准协议。
- **子网的划分**。对最基本的编址方法的改进，其标准[RFC 950]在1985 年通过。
- **无分类编址CIDR**。比较新的无分类编址方法CIDR。1993 年提出后很快就得到推广应用。



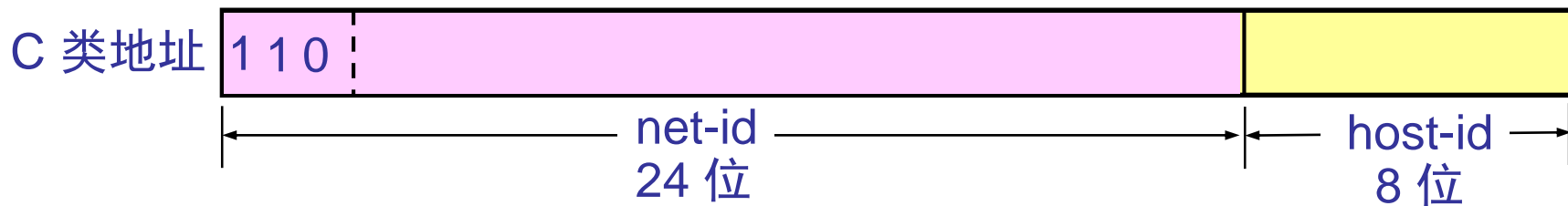
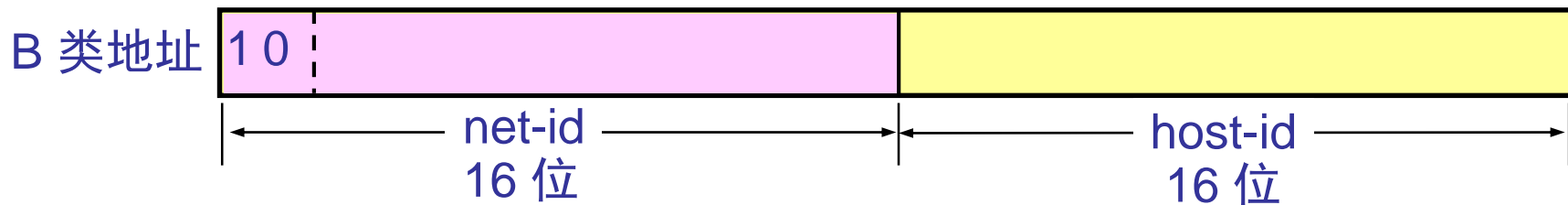
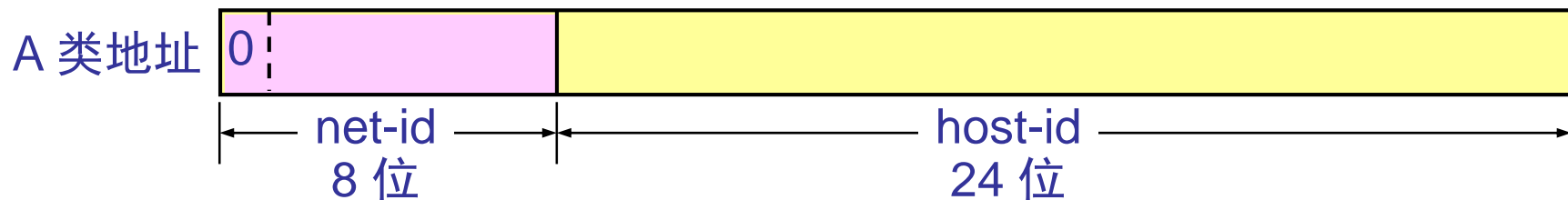
# 分类 IP 地址

- 分类IP地址中一个字段是网络号 net-id, 它标志主机（或路由器）所连接到的网络, 而另一个字段则是主机号 host-id, 它标志该主机（或路由器）。
- 两级的 IP 地址可以记为:

IP 地址 ::= { <网络号>, <主机号> }

::= 代表 “定义为”

# IP 地址中的网络号字段和主机号字段



# 点分十进制记法

机器中存放的 IP 地址  
是 32 位 二进制代码

10000000000010110000001100011111

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



## 2. 常用的三种类别的 IP 地址

### IP 地址的使用范围

网络类别	最大网络数	第一个可用的网络号	最后一个可用的网络号	每个网络中最大的主机数
A	126 ( $2^7 - 2$ )	1	126	16,777,214
B	16,384 ( $2^{14}$ )	128	191.255	65,534
C	2,097,152 ( $2^{21}$ )	192	223.255.255	254

# 无分类编址 CIDR

## 1. 网络前缀

在 1992 年互联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。
- 互联网主干网上的路由表中的项目数急剧增长。
- 整个IPv4的地址空间最终将全部耗尽。



# IP 编址问题的演进

---

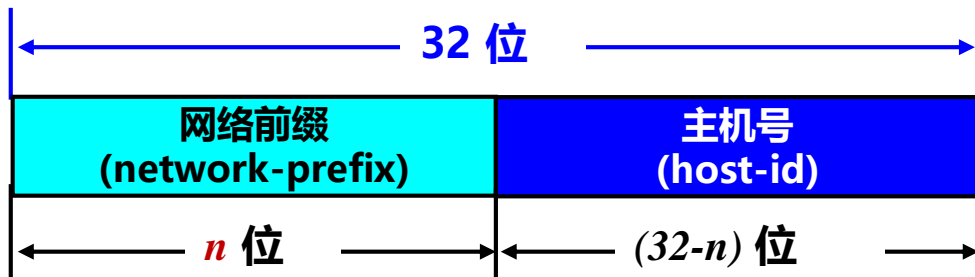
- 1993年,开始应用无分类编址方法**无分类域间路由选择 CIDR** (Classless Inter-Domain Routing)。

# CIDR 最主要的特点 -1

- 无分类的两级编址的记法是：

IP地址 ::= {<网络前缀>, <主机号>}

最大的区别：前缀的位数  $n$   
不固定，可以在 0 ~ 32 之间  
选取任意值。



- CIDR 记法：斜线记法 a.b.c.d/n：二进制 IP 地址的前  $n$  位是网络前缀。例如：  
128.14.35.7/20：前 20 位是网络前缀。





## CIDR 最主要的特点 -2

- CIDR 把网络前缀都相同的连续的 IP 地址组成“**CIDR 地址块**”。
- 使用地址块中的最小地址和网络前缀的位数指明一个地址块。
- 128.14.35.7/20 是 CIDR 地址块 128.14.32.0/20 中的一个地址。
- 128.14.32.0/20 表示的地址块共有  $2^{12}$  个地址，所以这个地址的主机号是 12 位。
-

# 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

...

...

10000000 00001110 00101111 11111011

10000000 00001110 00101111 11111100

10000000 00001110 00101111 11111101

10000000 00001110 00101111 11111110

最大地址



10000000 00001110 00101111 11111111

所有地址  
的 20 位  
前缀都是  
一样的



# CIDR 记法的其他形式 1

---

- 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。
- 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。

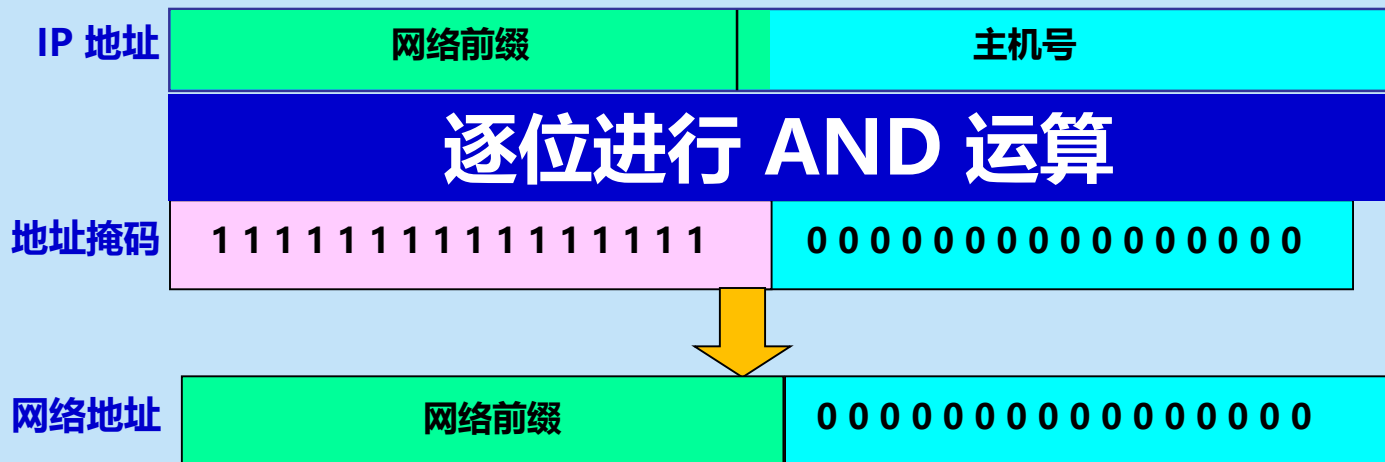


# 地址掩码(子网掩码)

- 使用子网掩码可以找出 IP 地址所在的网络。
- 由一连串 1 和接着的一连串 0 组成，而 1 的个数就是网络前缀的长度。
- /20 地址块的地址掩码：11111111  
11111111 11110000 00000000  
点分十进制记法：255.255.240.0  
CIDR 记法：255.255.240.0/20。



# (IP 地址) AND (子网掩码) = 网络地址



# 【例】已知 IP 地址是 141.14.72.24/18，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



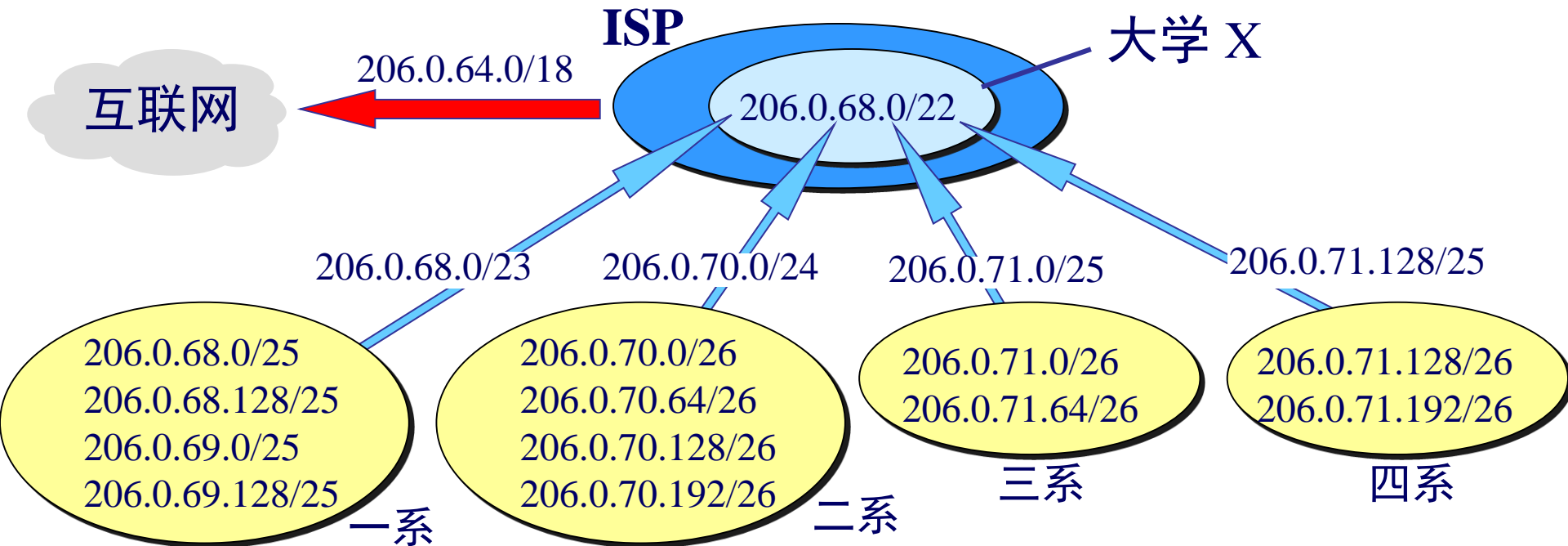
# 路由聚合

---

- CIDR地址块有很多地址,路由表中利用CIDR地址块查找目的网络,这种聚合称为**路由聚合**。
- 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**。

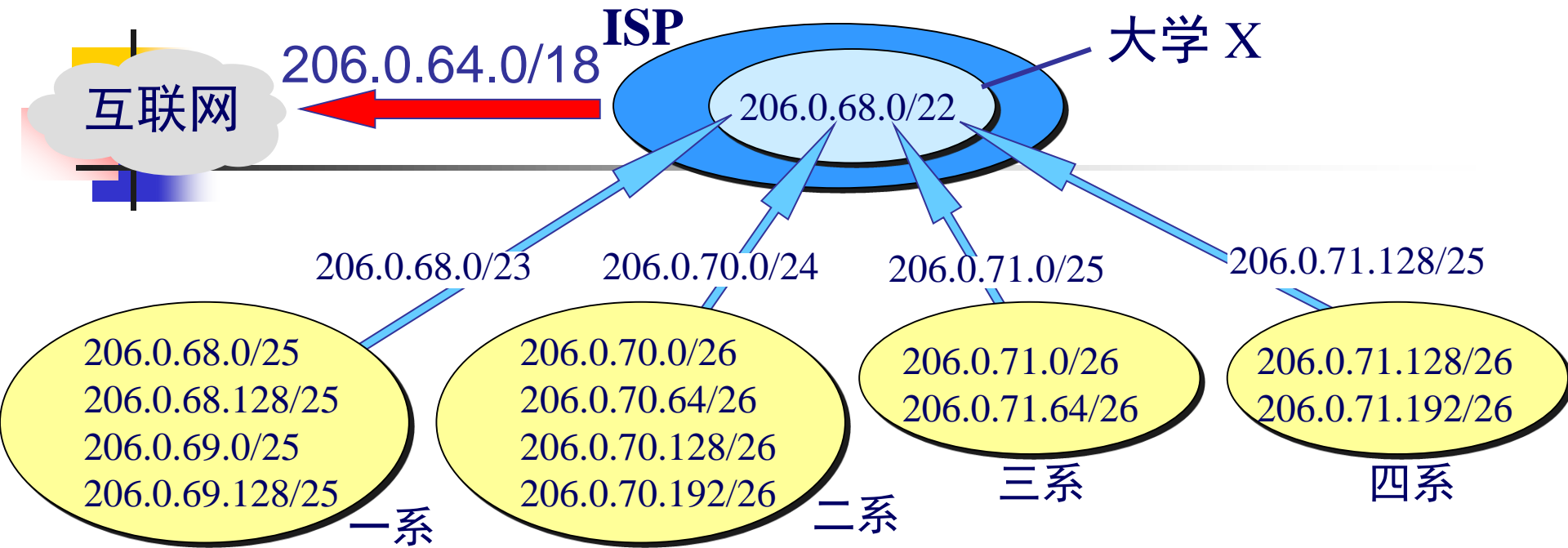


# CIDR 地址块划分举例



单位	地址块	二进制表示	地址数
ISP	206.0.64.0/18	11001110.00000000.01*	16384
大学	206.0.68.0/22	11001110.00000000.010001*	1024
一系	206.0.68.0/23	11001110.00000000.0100010*	512
二系	206.0.70.0/24	11001110.00000000.01000110.*	256
三系	206.0.71.0/25	11001110.00000000.01000111.0*	128
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。(206.0.68.0/22)



# 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，方便了 IP 地址的管理。
- 第二，路由器仅根据目的主机所连接的网络前缀来转发分组（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，其网络前缀必须是不同的。
- 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。

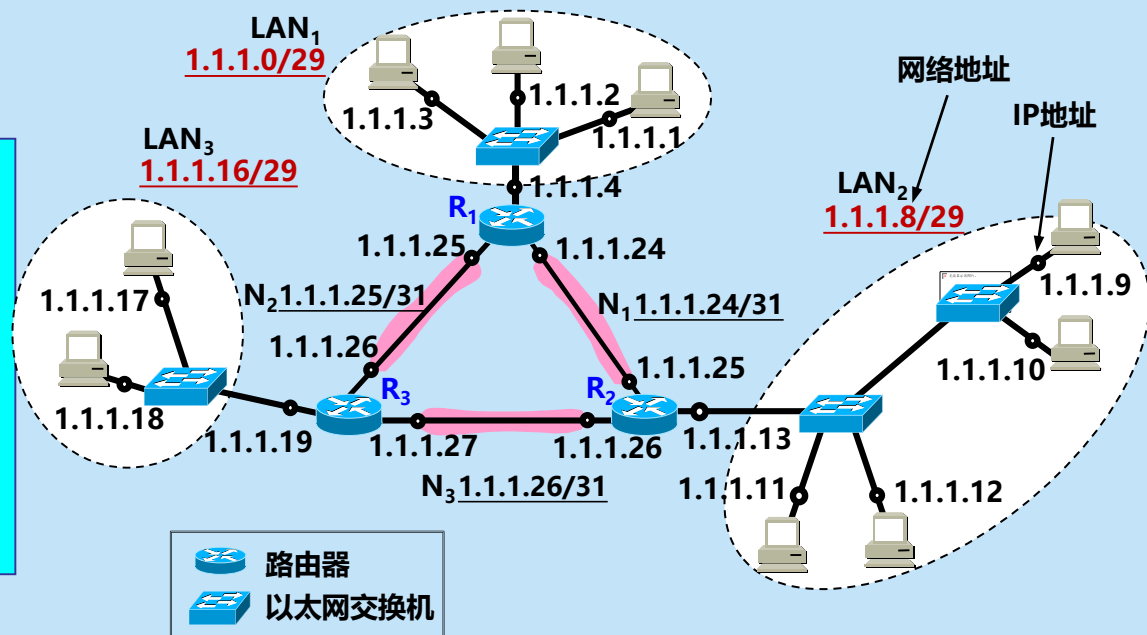


# IP 地址的一些重要特点

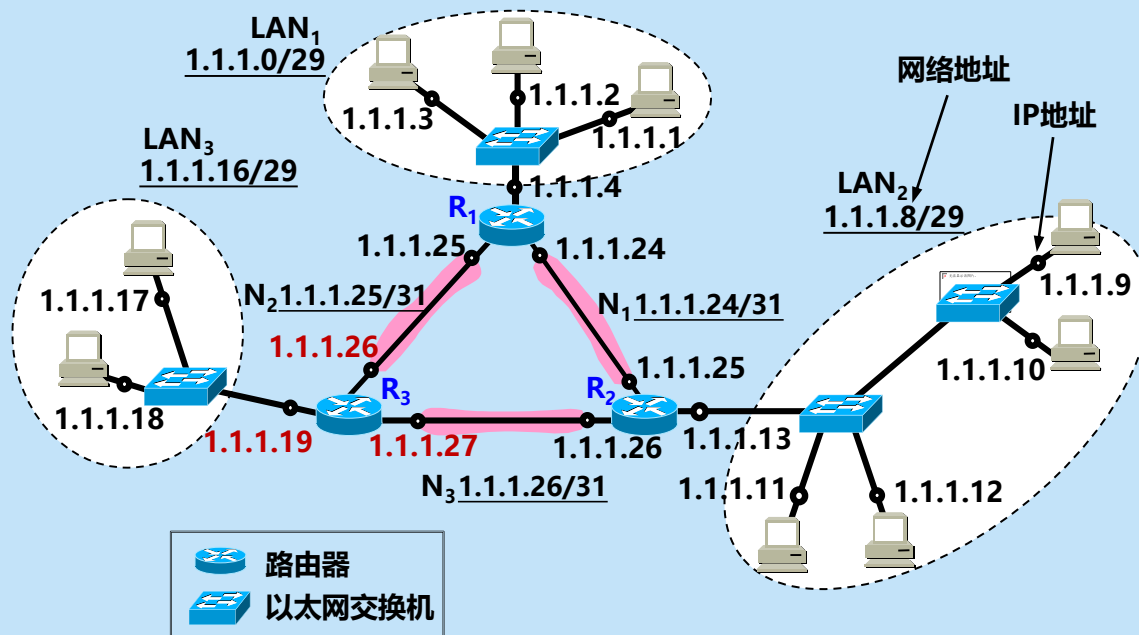
---

- (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前缀。
- (4) 所有分配到网络前缀的网络，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。

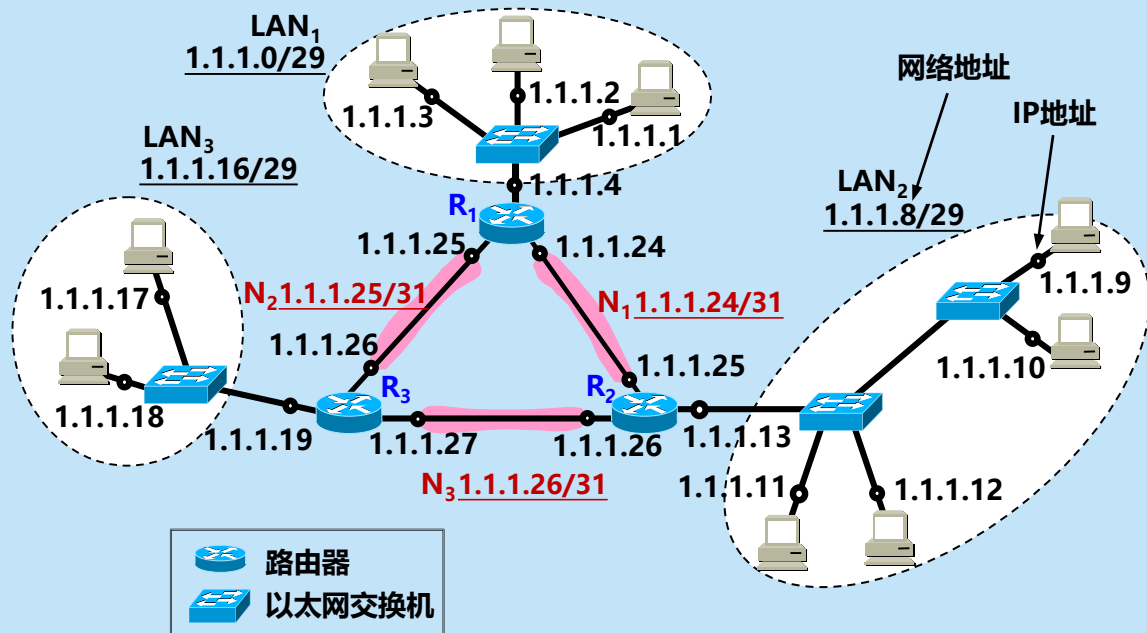
同一个局域网  
上的主机或路  
由器的IP地  
址中的**网络号**  
必须一样。



路由器的每一个接口都有一个不同网络号的IP地址。

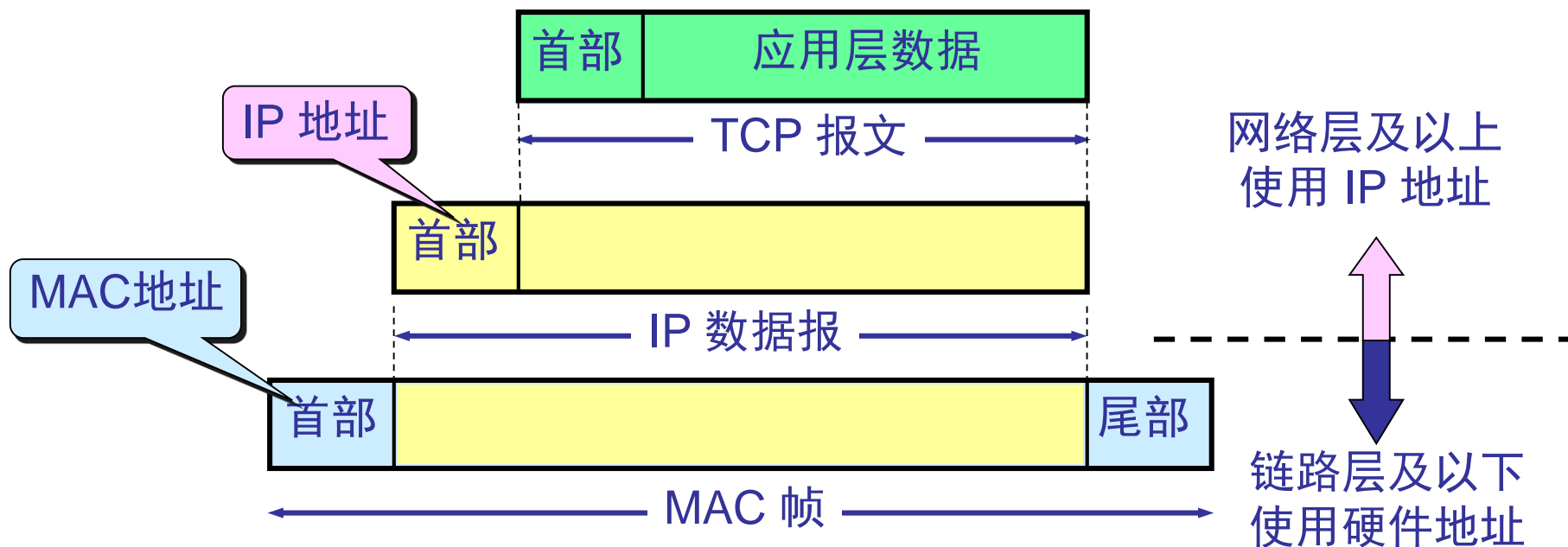


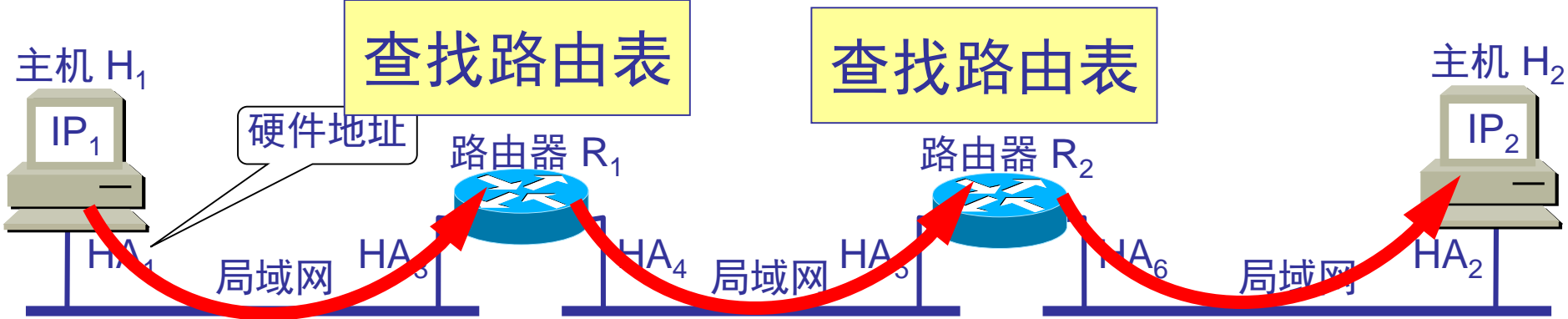
两个路由器直接相连的接口处，可指明也可不指明 IP 地址。如指明 IP 地址，则这一段连线就构成了一种只包含一段线路的特殊“网络”。这种网络**仅需**两个 IP 地址，可以使用 **/31** 地址块。主机号可以是 0 或 1。





## 4.2.3 IP 地址与MAC地址

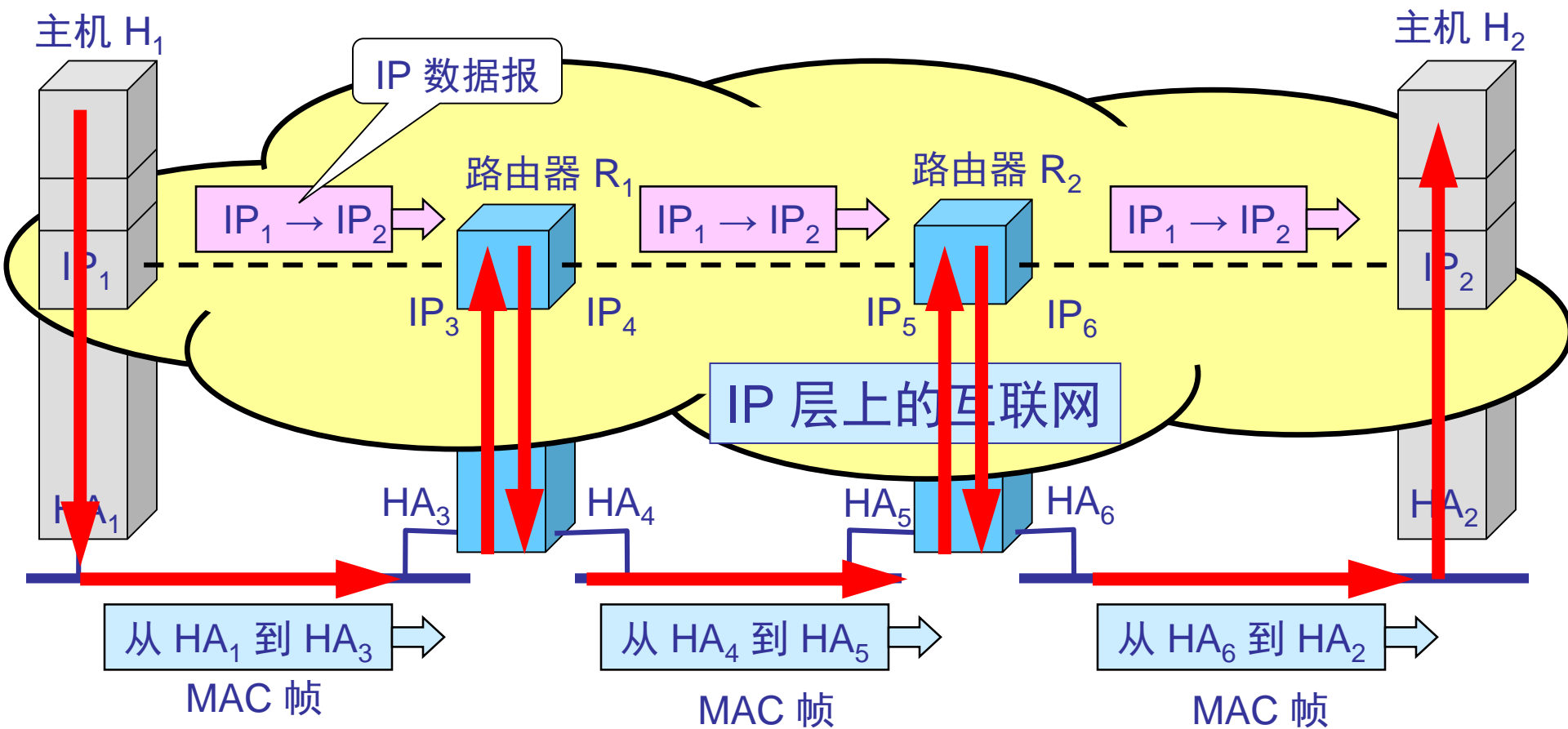
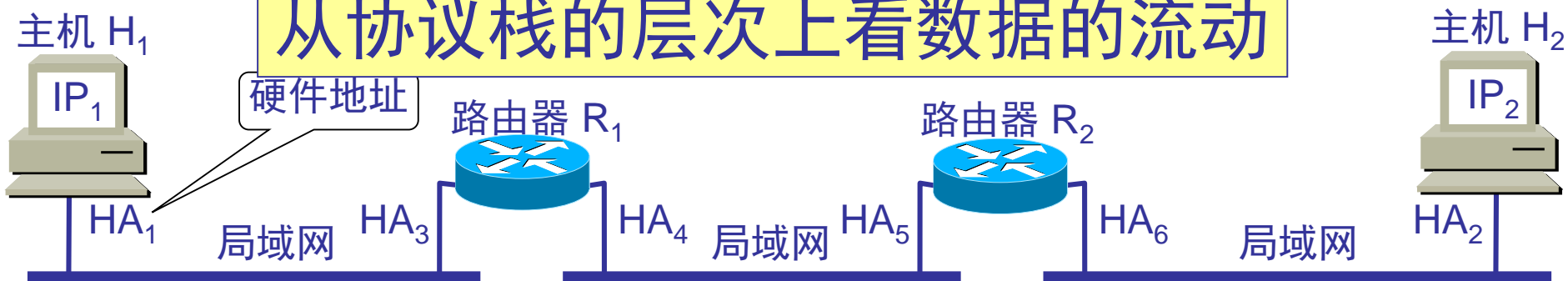




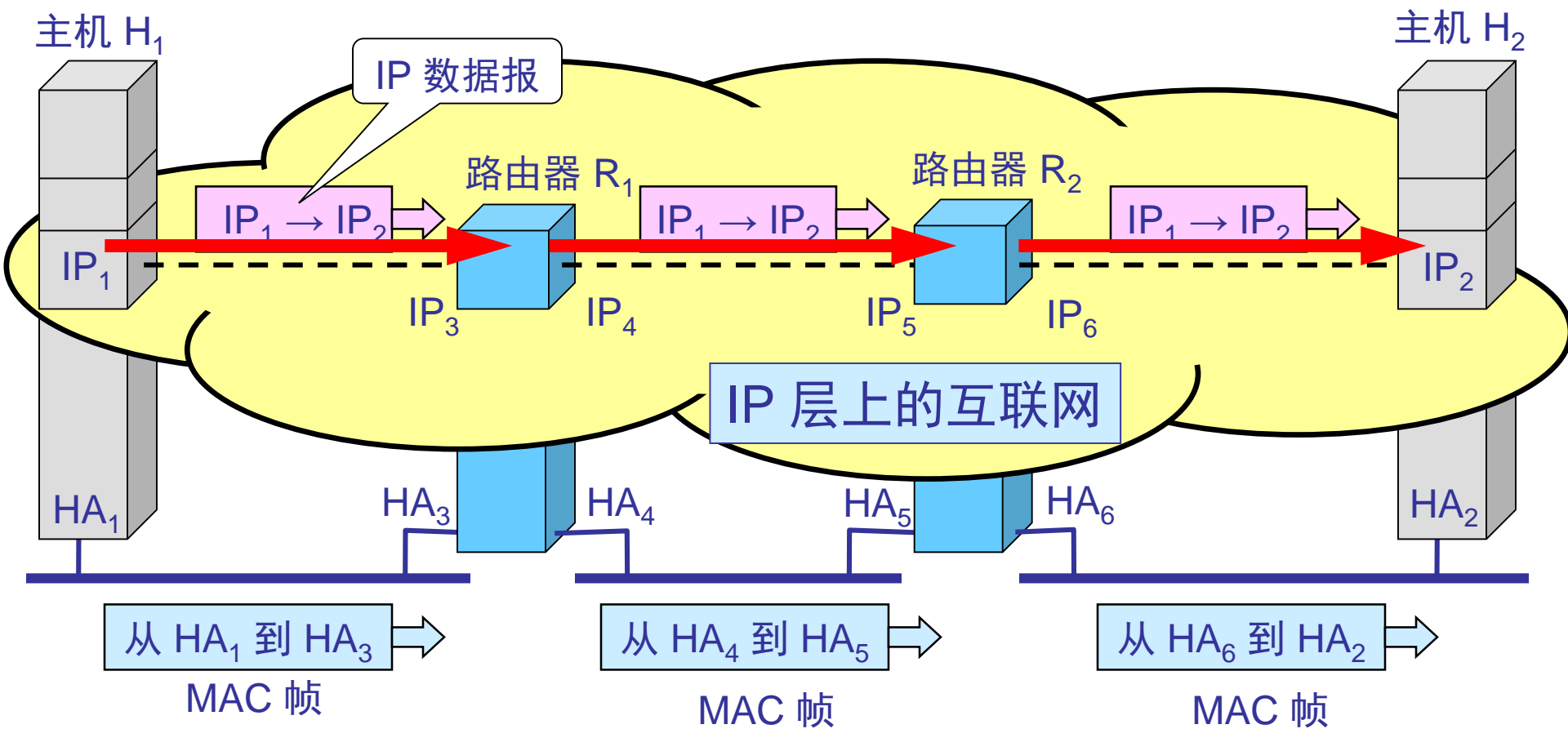
### 通信的路径

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

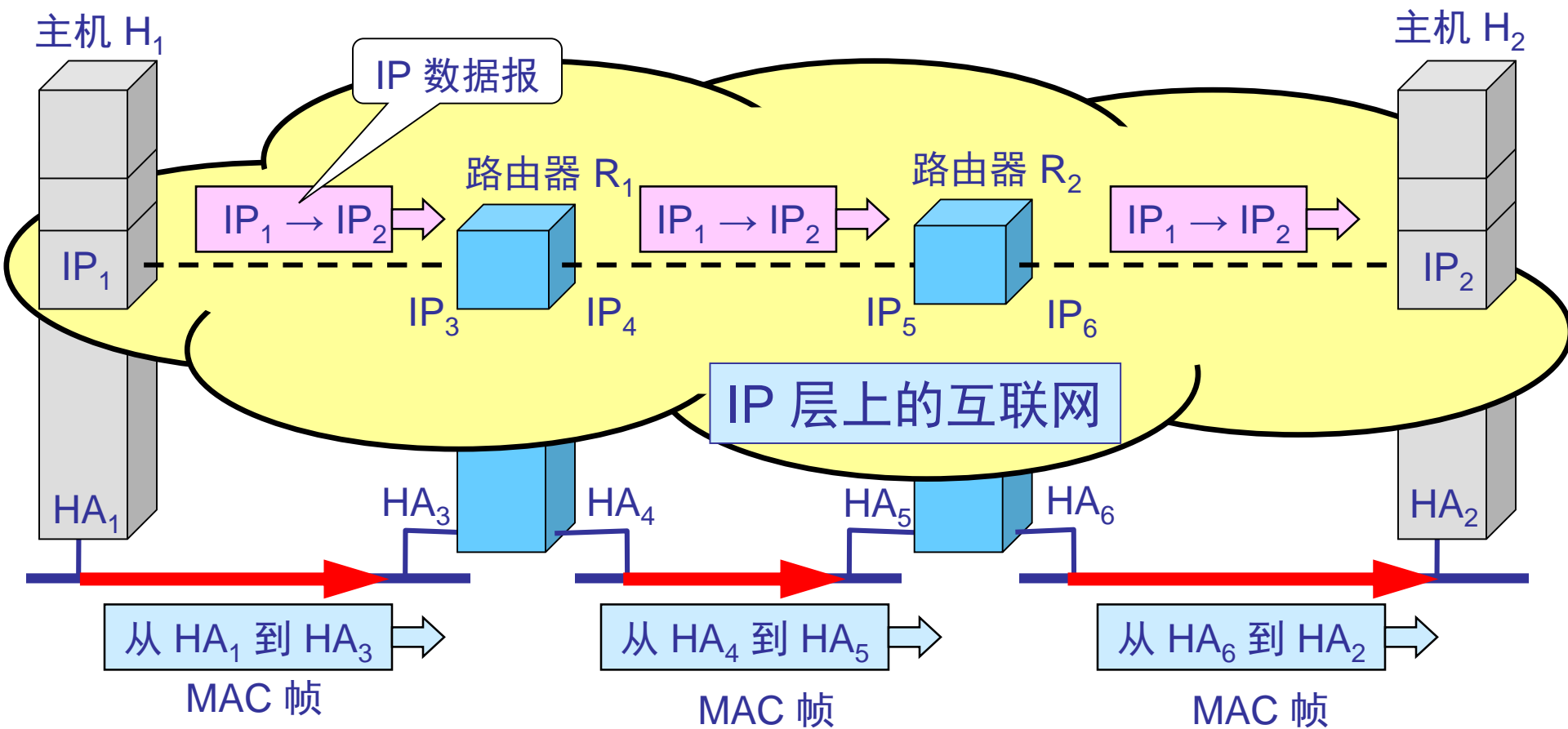
# 从协议栈的层次上看数据的流动



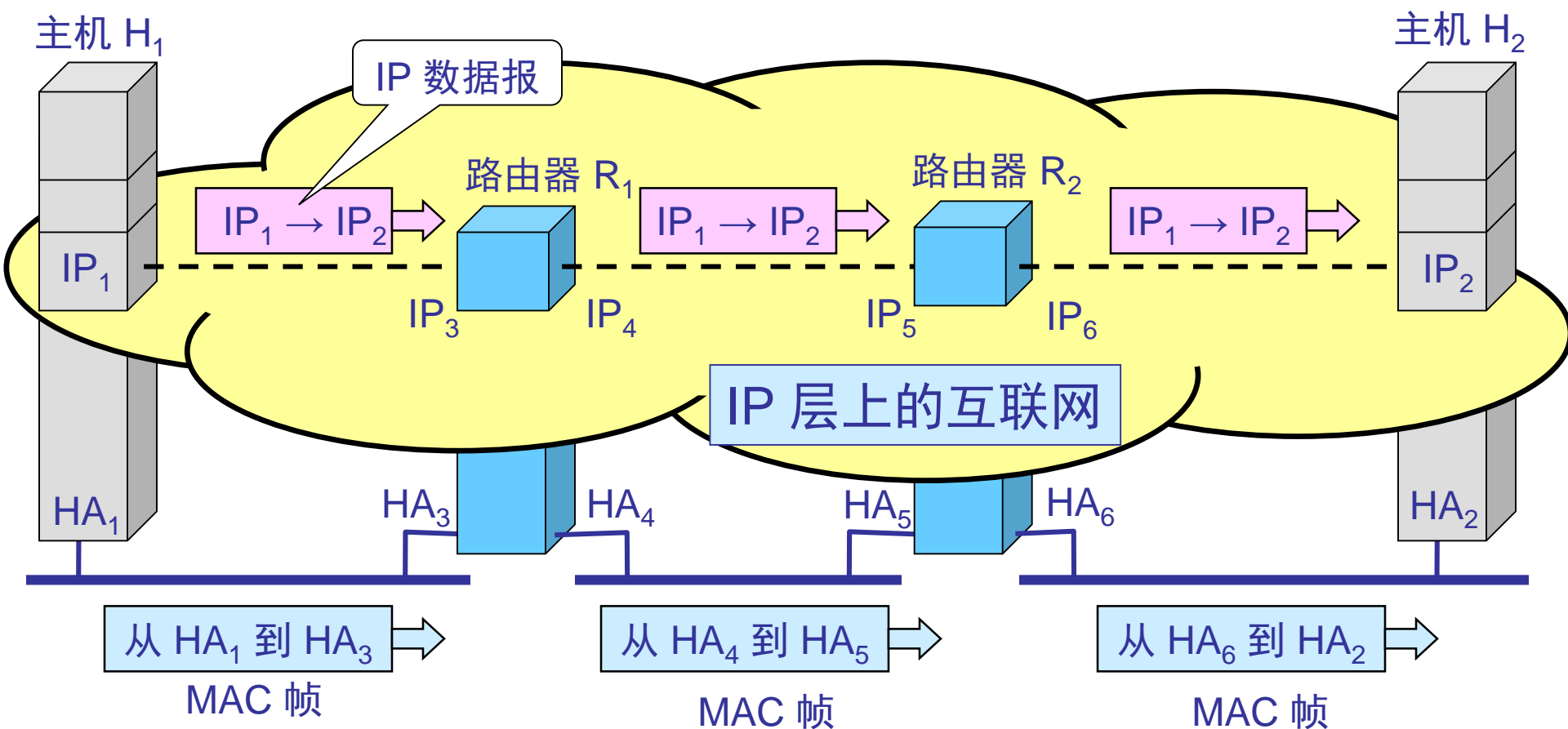
# 从虚拟的 IP 层上看 IP 数据报的流动



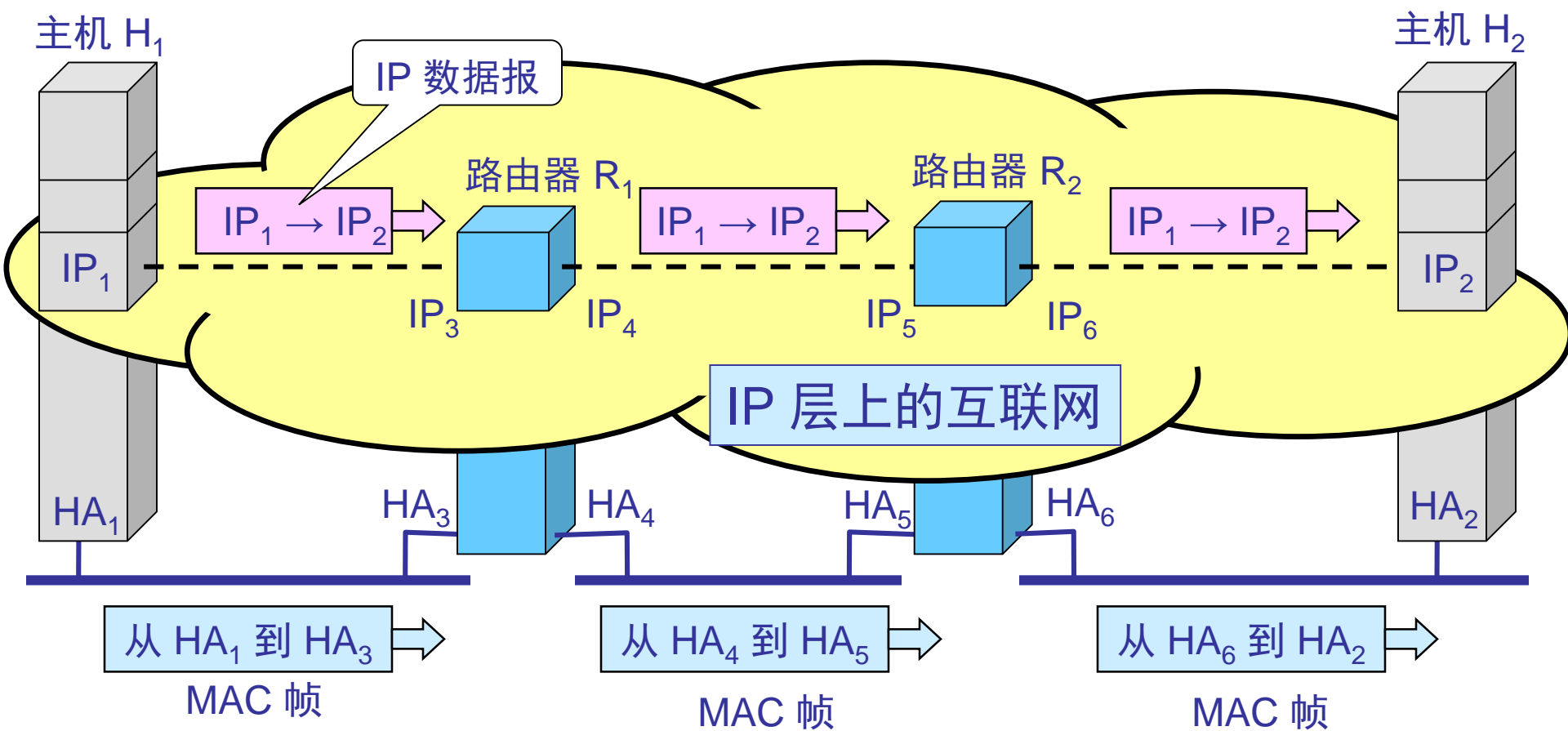
# 在链路上看 MAC 帧的流动



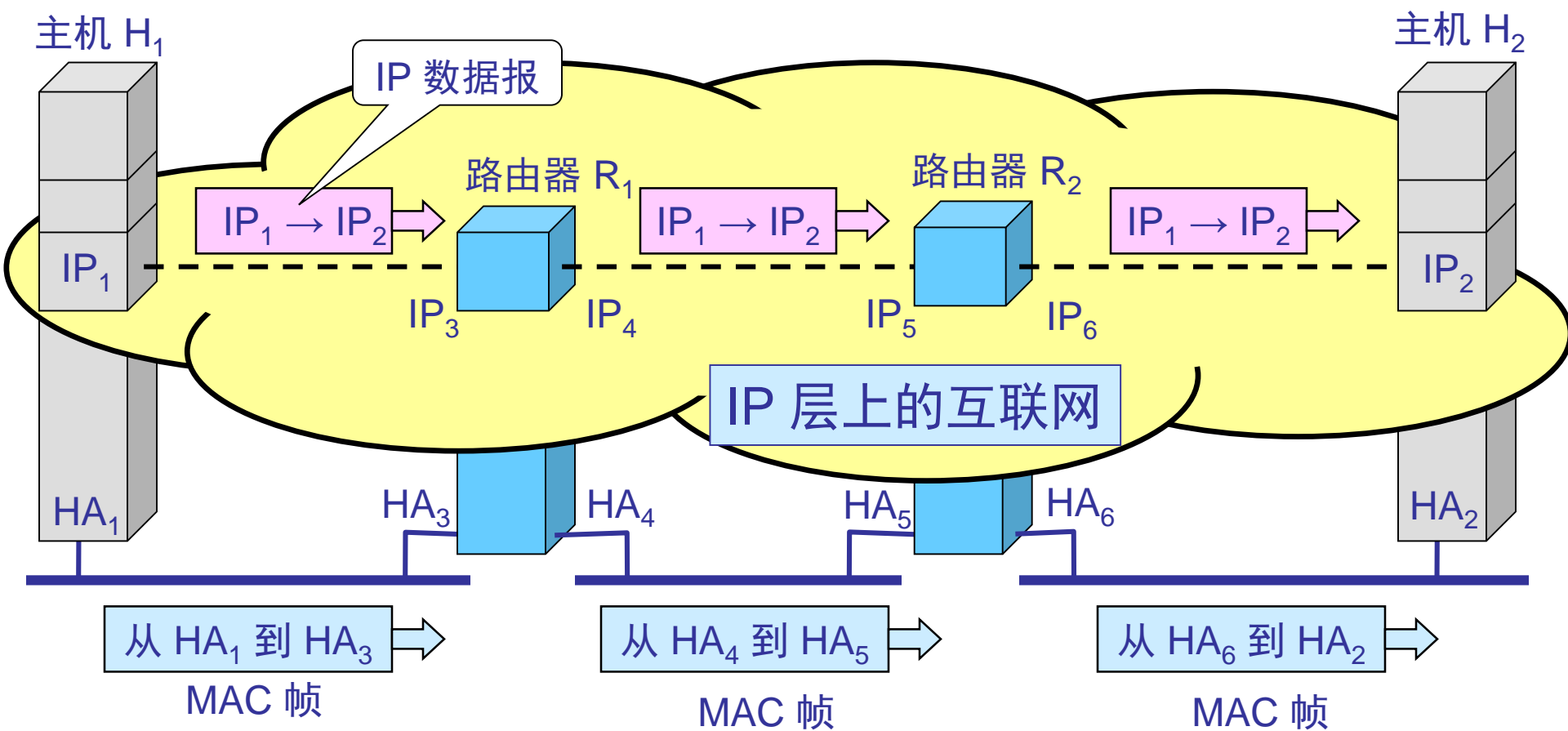
在 IP 层抽象的互联网上只能看到 IP 数据报  
图中的  $IP_1 \rightarrow IP_2$  表示从源地址  $IP_1$  到目的地址  $IP_2$   
两个路由器的 IP 地址并不出现在 IP 数据报的首部中



路由器只根据目的站的 IP 地址的网络号进行路由选择

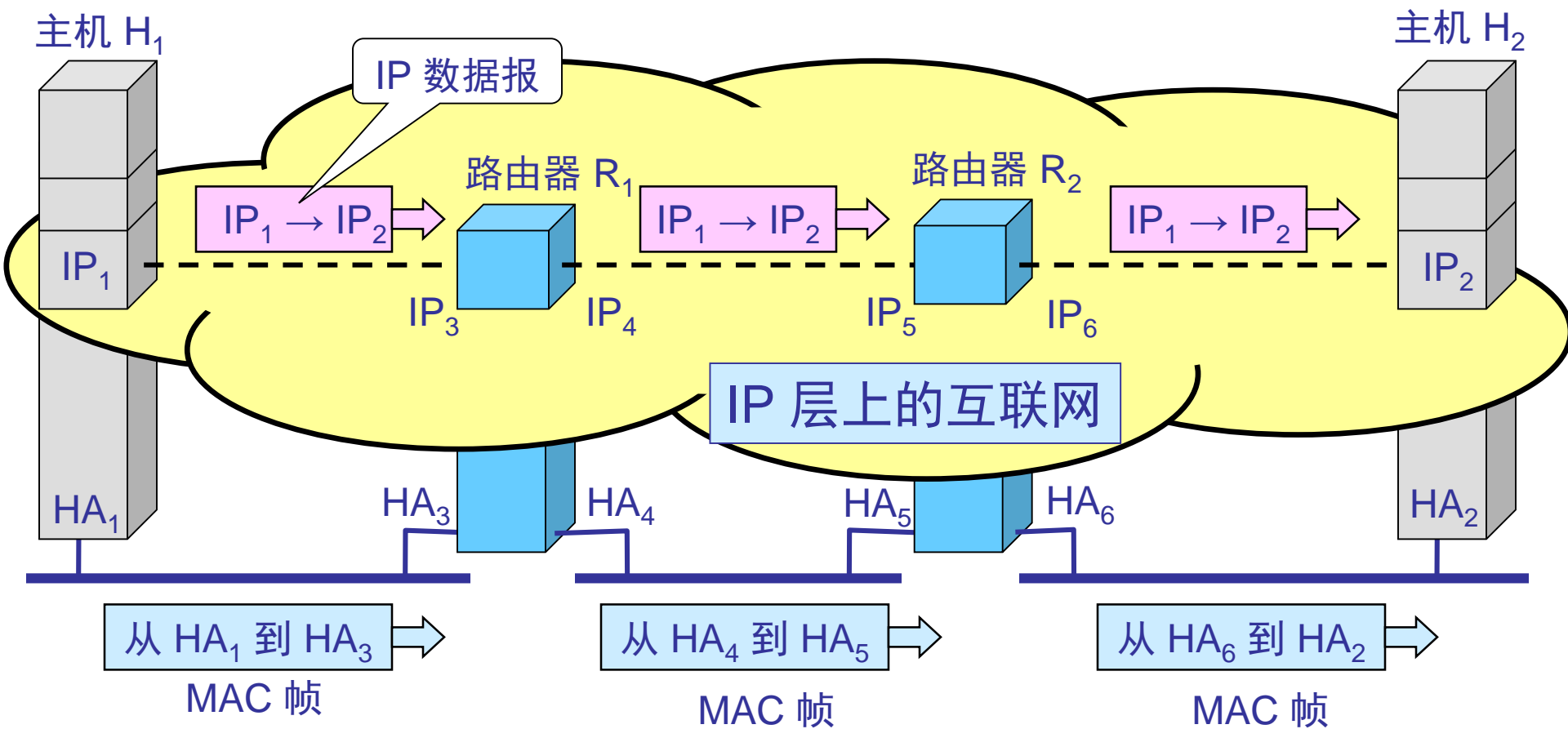


在具体的物理网络的链路层  
只能看见 MAC 帧而看不见 IP 数据报

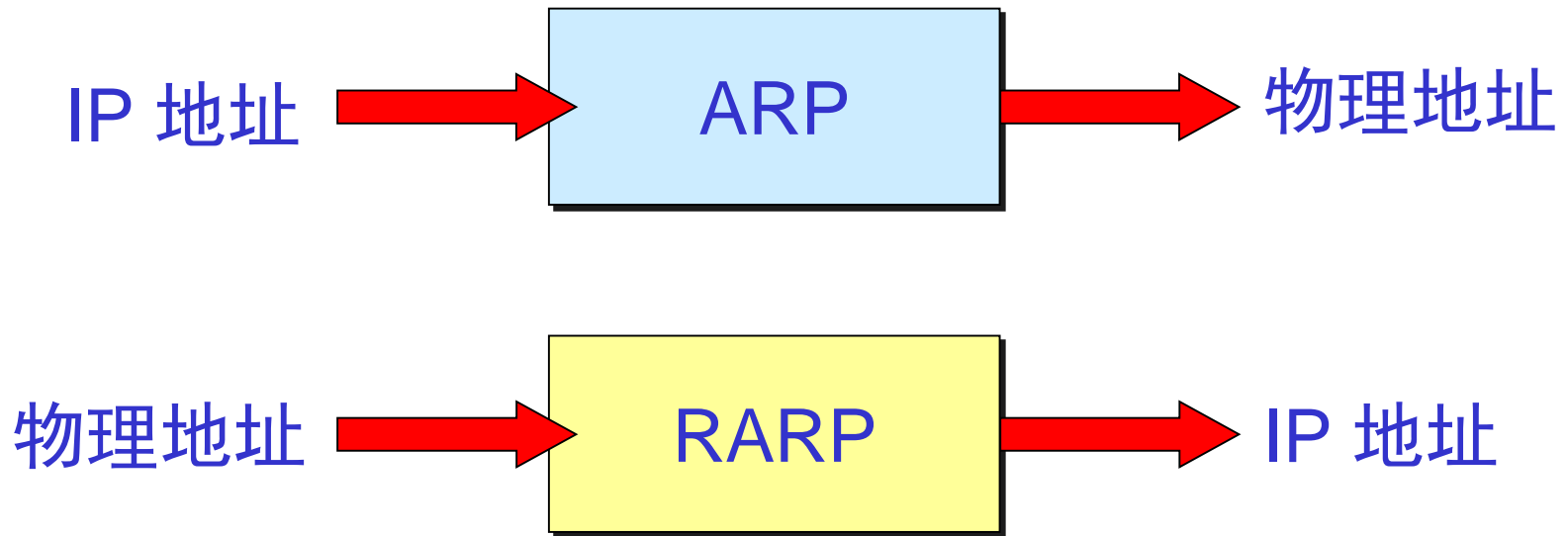




IP层抽象的互联网屏蔽了下层很复杂的细节  
在抽象的网络层上讨论问题，就能够使用  
统一的、抽象的 IP 地址  
研究主机和主机或主机和路由器之间的通信

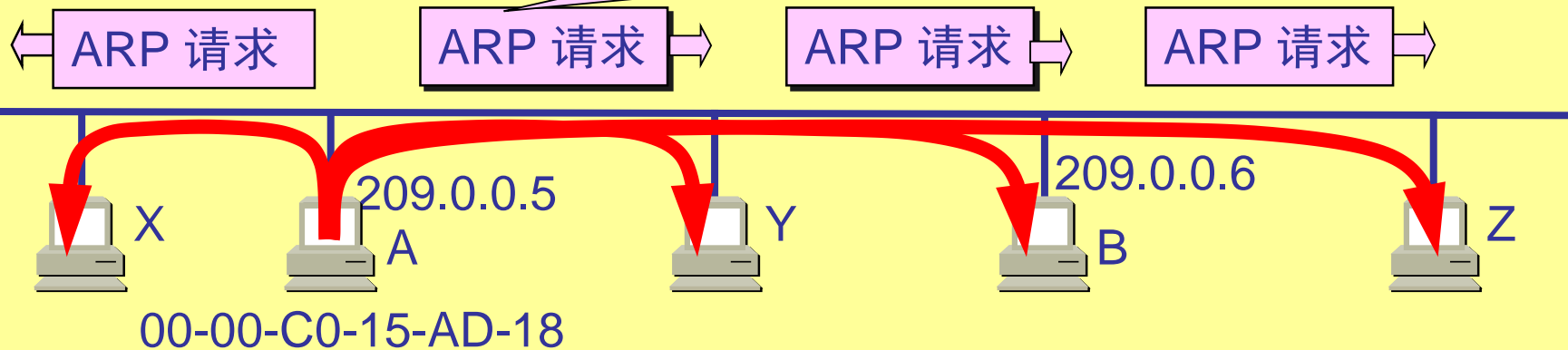


## 4.2.4 地址解析协议 ARP



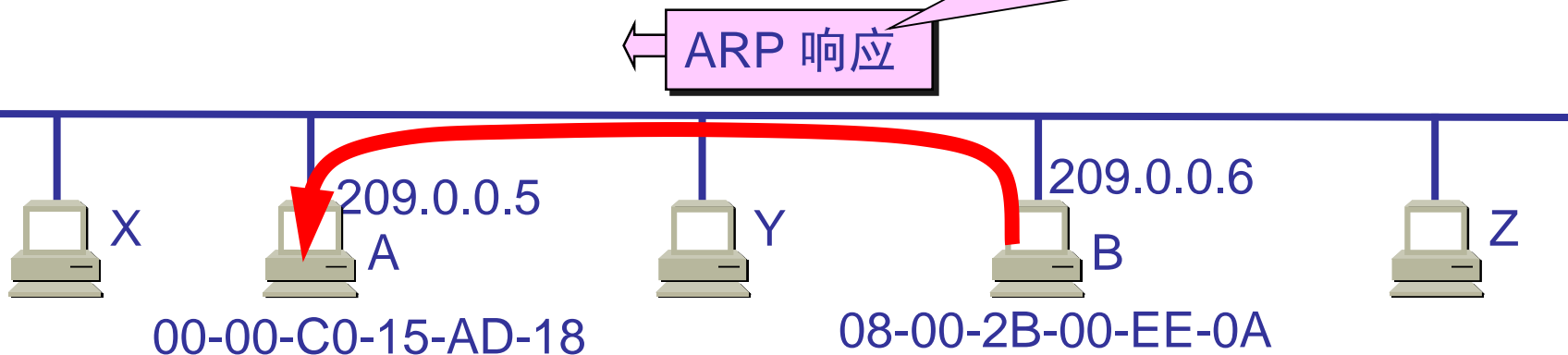
主机 A 广播发送  
ARP 请求分组

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

---

- 主机A要将B的地址映射写入高速缓存。
- 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。



# ARP 高速缓存

---

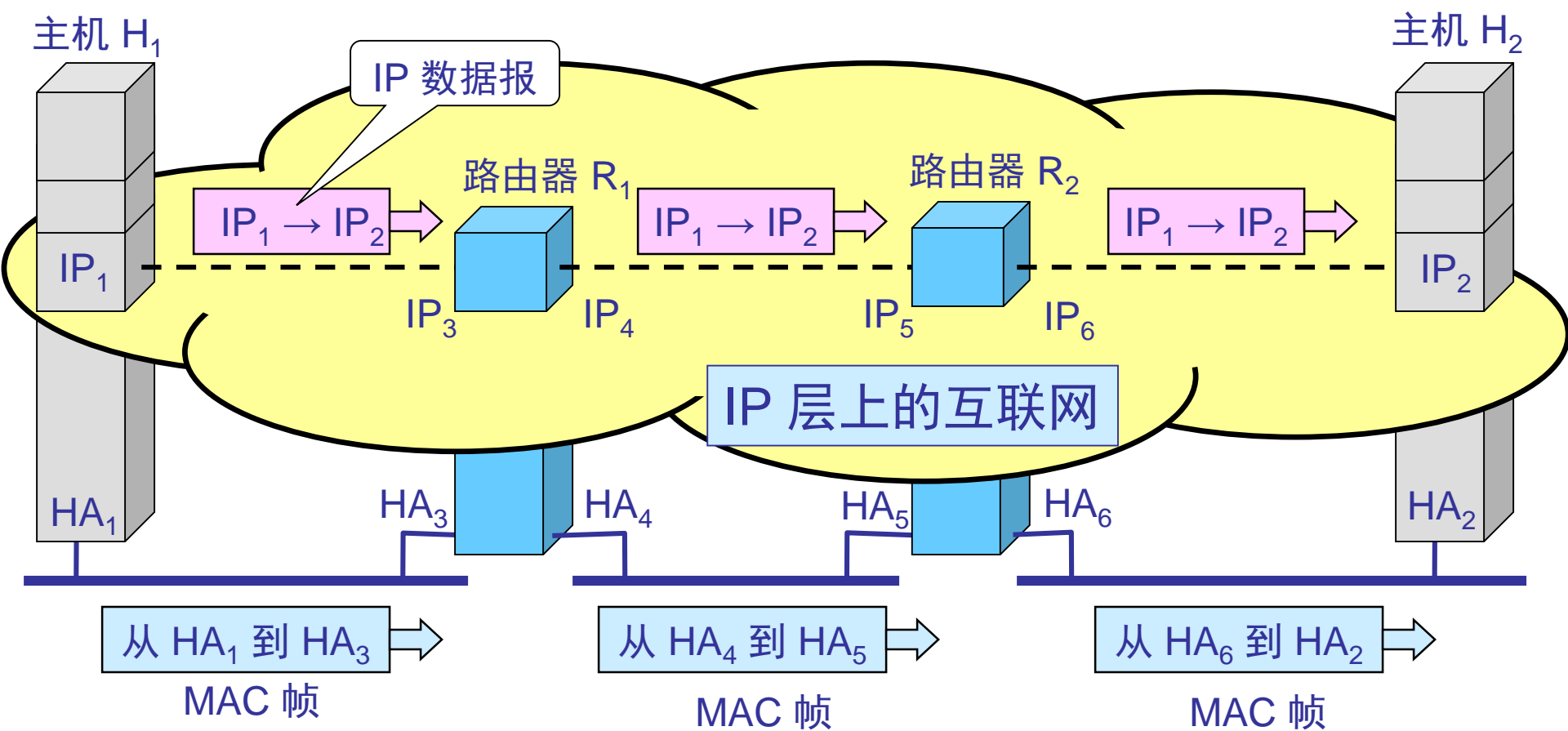
- 当主机 A 欲向本局域网上的某个主机 B 发送 IP 数据报时，就先在其 ARP 高速缓存中查看有无主机 B 的 IP 地址。
- 如有，就可查出其对应的硬件地址，再将此硬件地址写入 MAC 帧，然后通过局域网将该 MAC 帧发往此硬件地址。
- 如没有，则发送 ARP 请求。



# 应当注意的问题

---

- ARP缓存中的映射地址都有生存时间。
- 如果所要找的主机和源主机不在同一个局域网上，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IP数据报发送给这个路由器，让这个路由器把IP数据报转发给下一个网络。剩下的工作就由下一个网络来做。





# ARP欺骗

- 对路由器ARP表的欺骗：通知路由器一系列错误的内网MAC地址，并按照一定的频率不断进行，使真实的地址信息无法通过更新保存在路由器中，结果路由器的所有数据只能发送给错误的MAC地址，造成正常PC无法收到信息。
- 对内网PC的网关欺骗：建立假网关，让被它欺骗的PC向假网关发数据，而不是通过正常的路由器途径上网。在PC看来，就是上不了网了，“网络掉线了”。



# 指引

- 网络层提供的两种服务
- 网际协议IP
  - 虚拟互联网络
  - IP地址
  - IP地址与MAC地址
  - 地址解析协议ARP
  - IP数据包的格式
- IP层转发分组的过程
- 网际控制报文协议 ICMP
- 互联网的路由选择协议
- IPV6

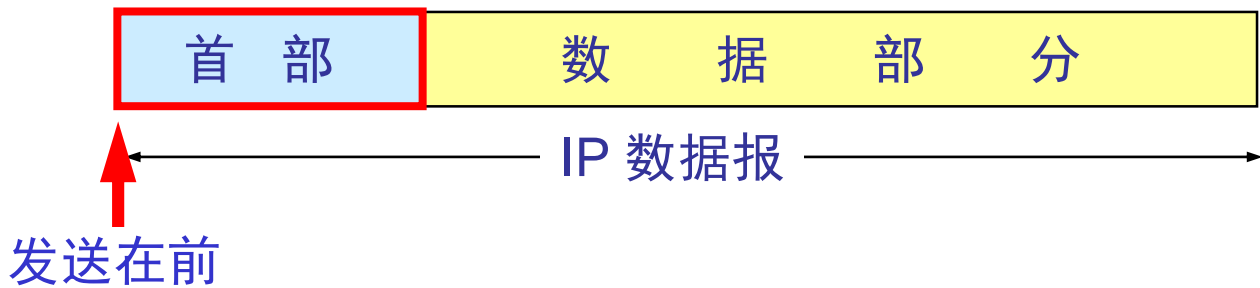




## 4.2.5 IP 数据报的格式

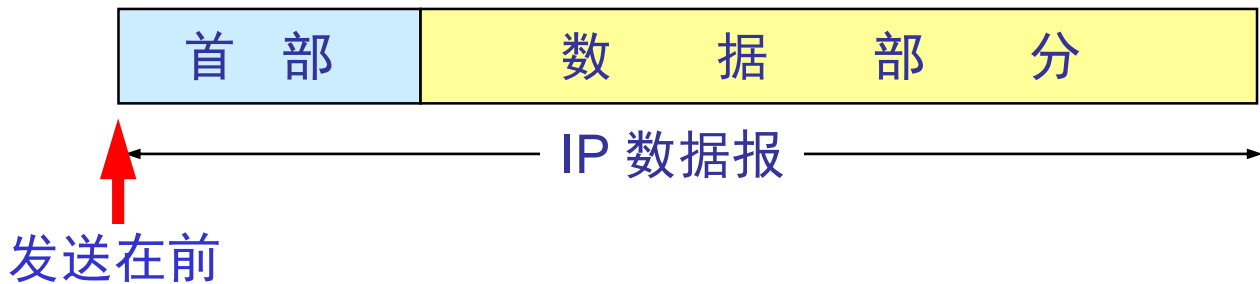
---

- 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。

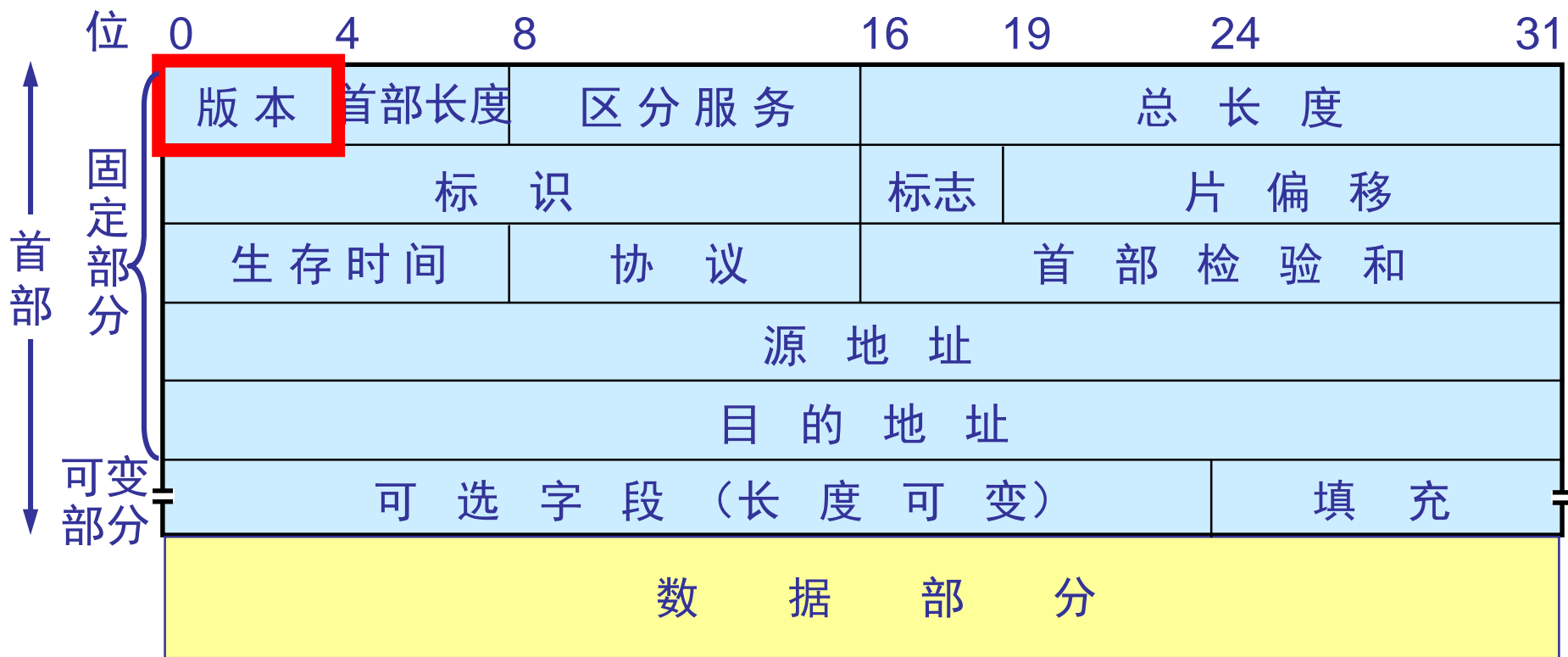




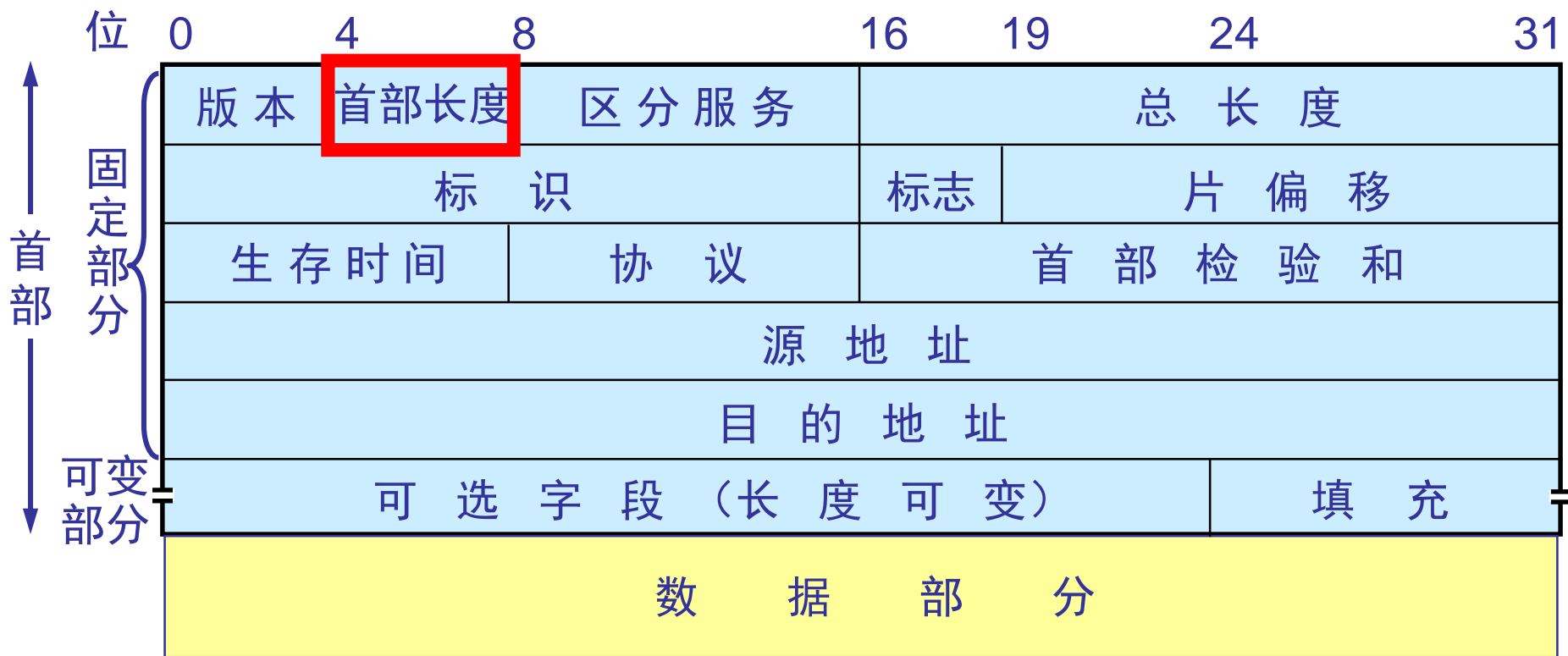
发送在前



# 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——占 4 位，可表示的最大数值  
是 15 个单位(一个单位为 4 字节)  
因此 IP 的首部长度的最大值是 60 字节。



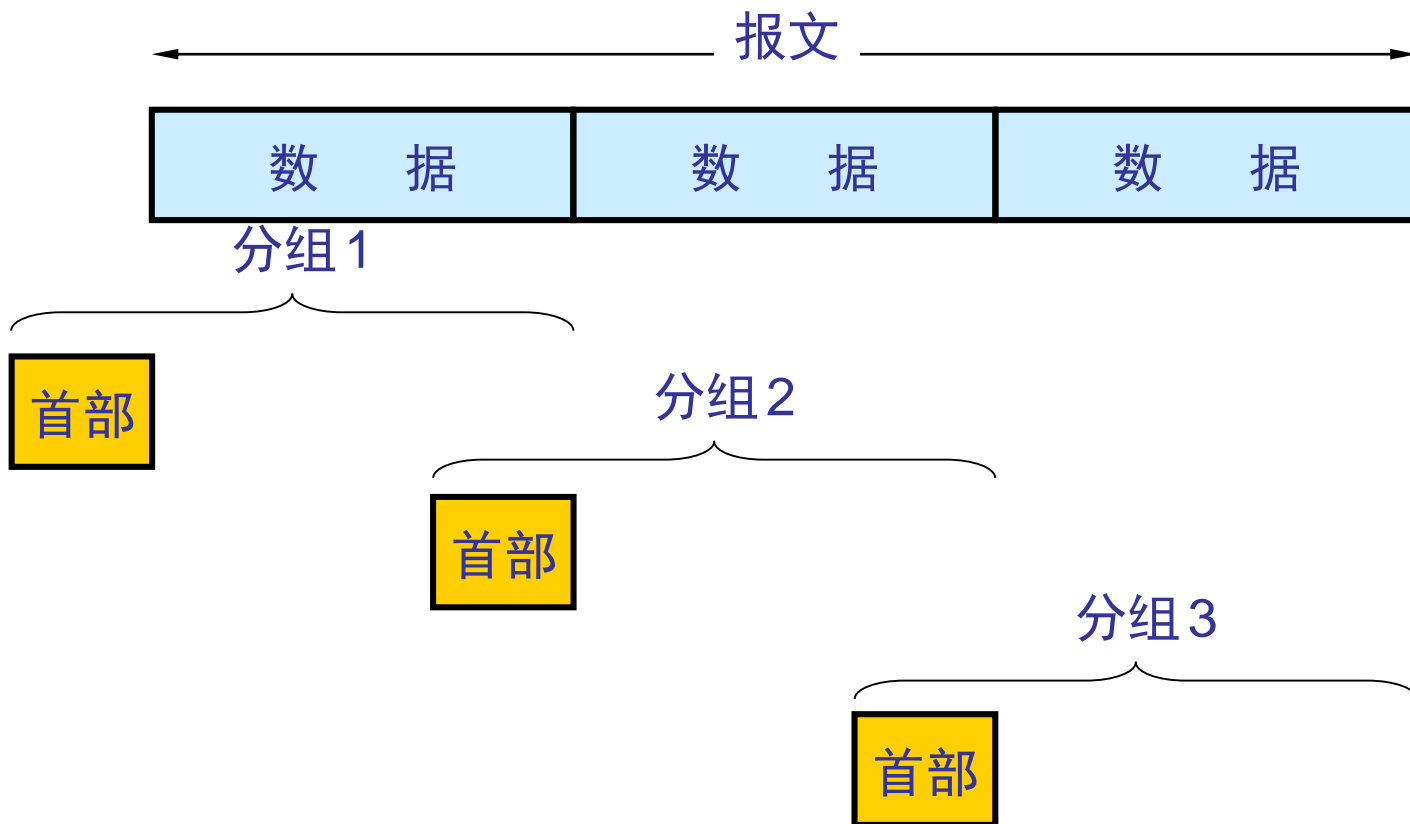
区分服务——占 8 位，用来获得更好的服务  
在旧标准中叫做服务类型，但实际上一直未被使用过。  
1998 年这个字段改名为区分服务。  
只有在使用区分服务（DiffServ）时，这个字段才起作用。  
在一般的情况下都不使用这个字段





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 字节。  
总长度必须不超过最大传送单元 MTU。

# IP数据报超过最大传送单元





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

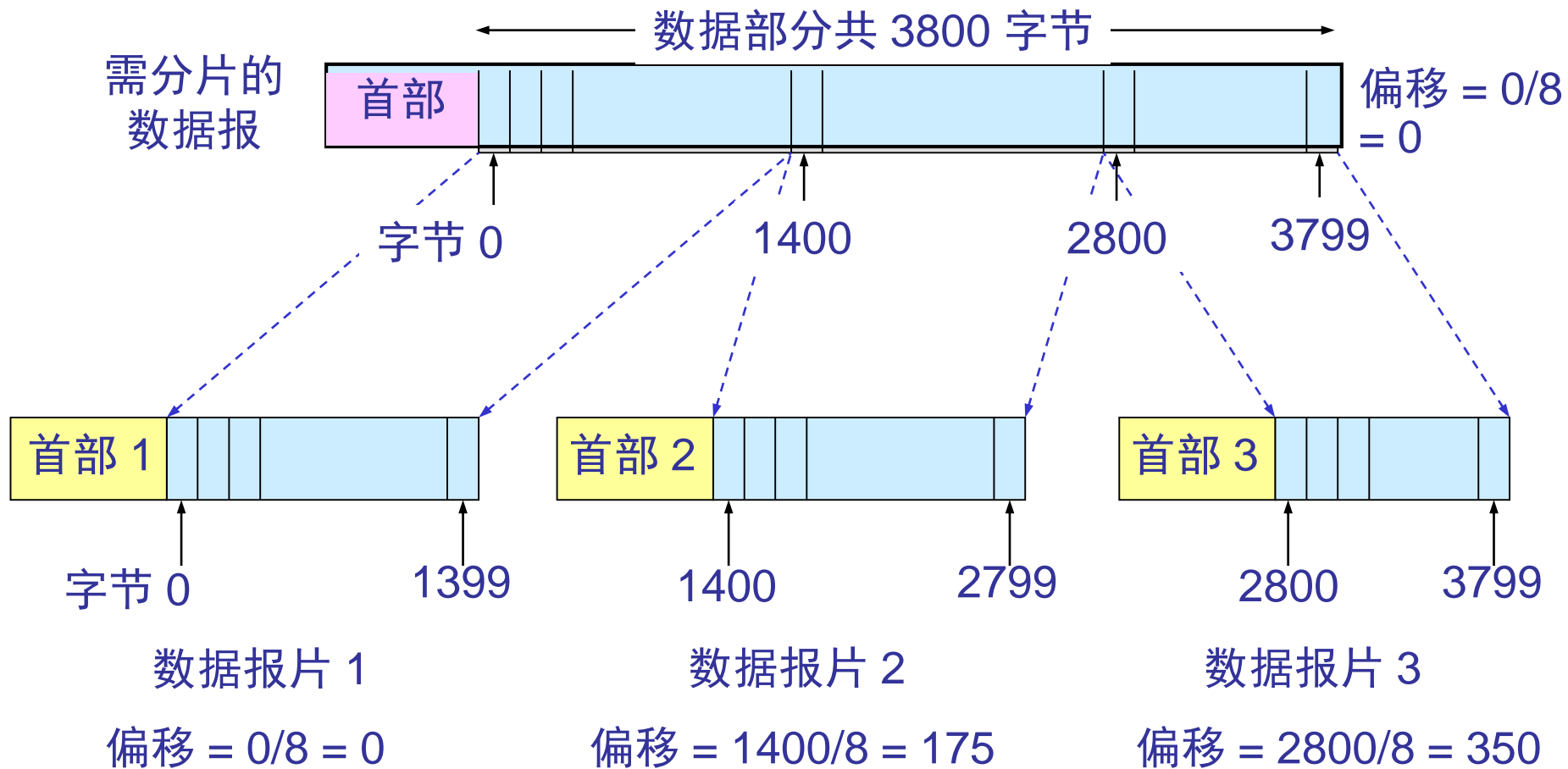


标志(flag) 占 3 位，目前只有前两位有意义。  
标志字段的最低位是 **MF** (More Fragment)。  
MF = 1 表示后面“还有分片”。MF = 0 表示最后一个分片。  
标志字段中间的一位是 **DF** (Don't Fragment)。  
只有当 DF = 0 时才允许分片。



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。

# 【例4-1】 IP 数据报分片



## 【例4-1】 IP 数据报分片

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

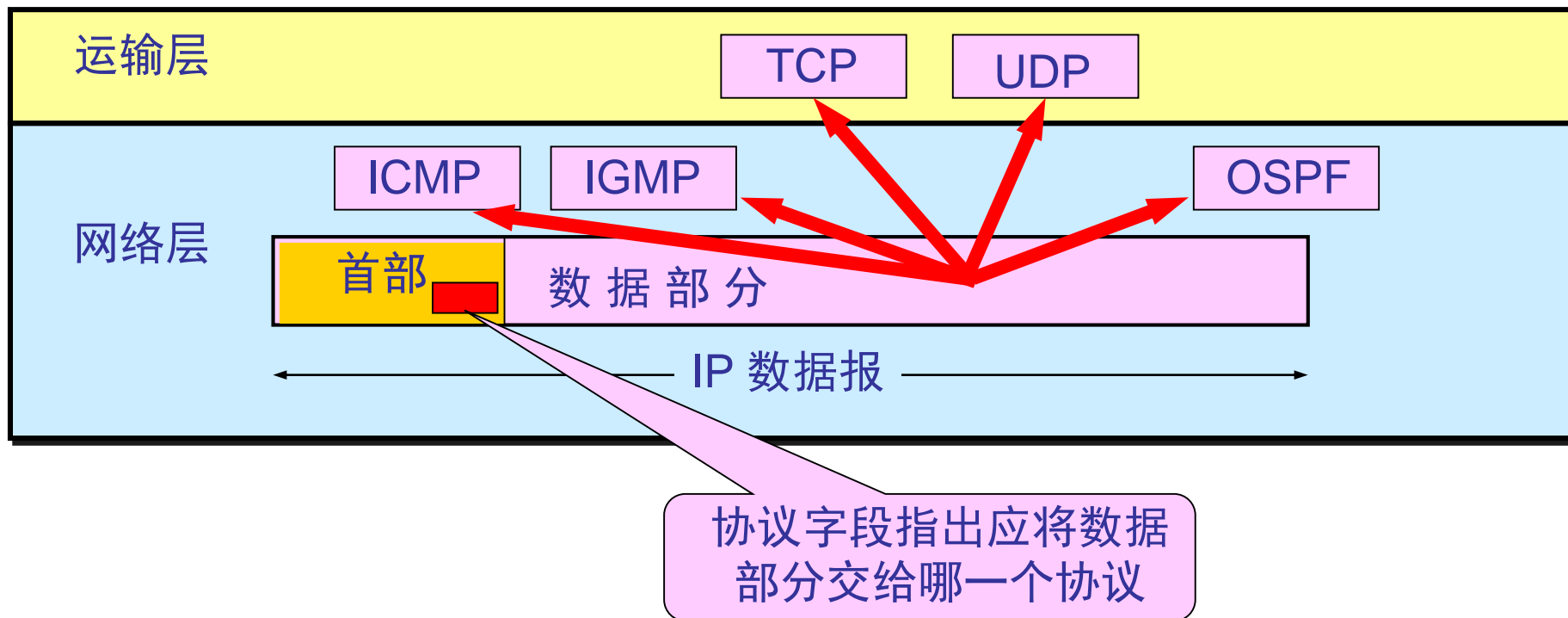


生存时间(8 位)记为 TTL (Time To Live)  
数据报在网络中可通过的路由器数的最大值。





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





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。

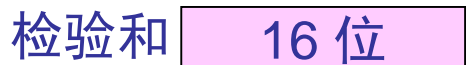
发送端

接收端

数据报首部

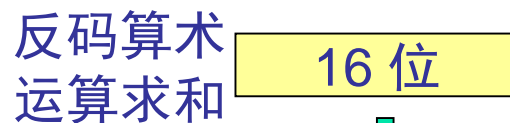
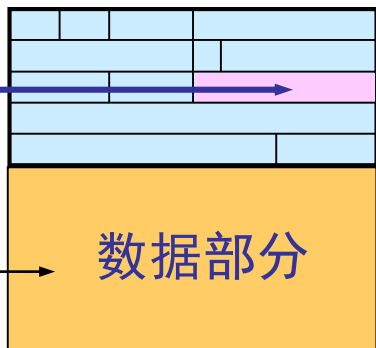


取反码

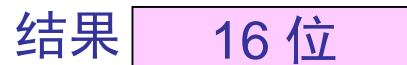


数据部分  
不参与检验和的计算

IP 数据报



取反码



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。
- 实际上这些选项很少被使用。

# 首部格式总结



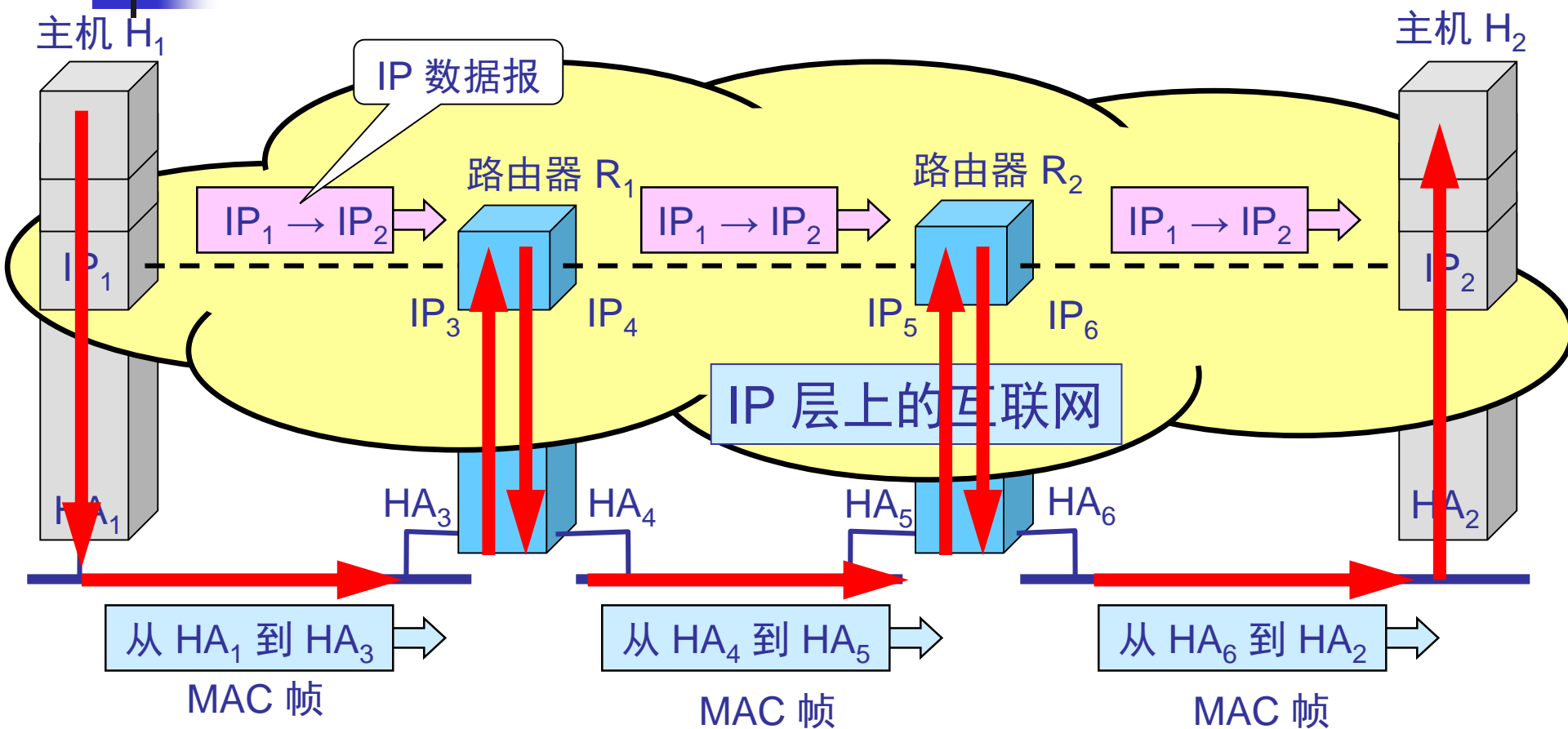
# 指引

- 网络层提供的两种服务
- 网际协议IP
  - 虚拟互联网络
  - IP地址
  - IP地址与MAC地址
  - 地址解析协议ARP
  - IP数据包的格式
- IP层转发分组的过程
- 网际控制报文协议 ICMP
- 互联网的路由选择协议
- IPV6





# IP 层转发分组的过程



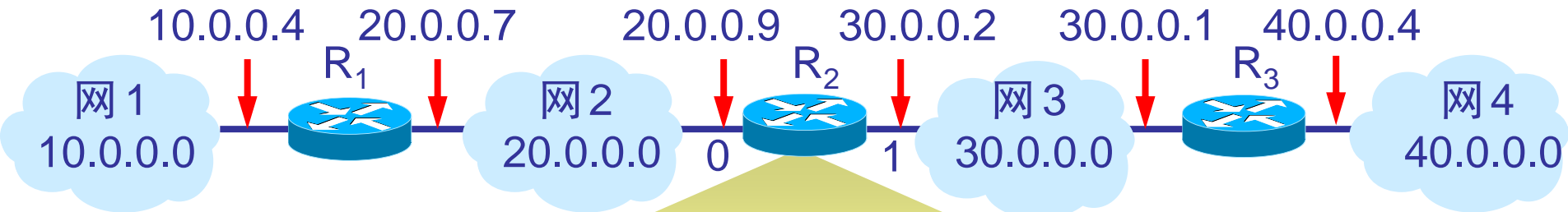


## 4.3 IP 层转发分组的过程

---

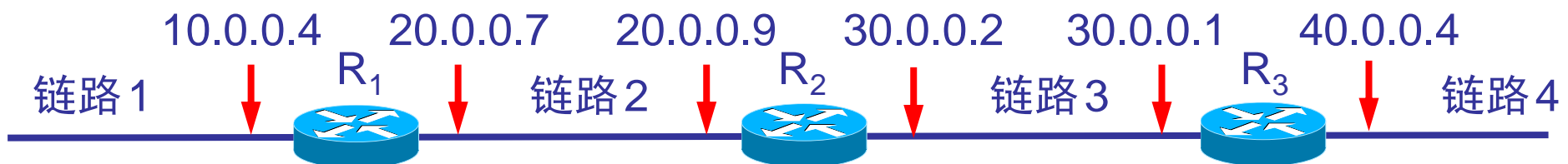
- 若按目的主机号来制作路由表，则所得出的路由表就会过于庞大。
- 但若按主机所在的**网络地址**来制作路由表，即是基于目的主机的所在的网络,就可使路由表大大简化。

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)

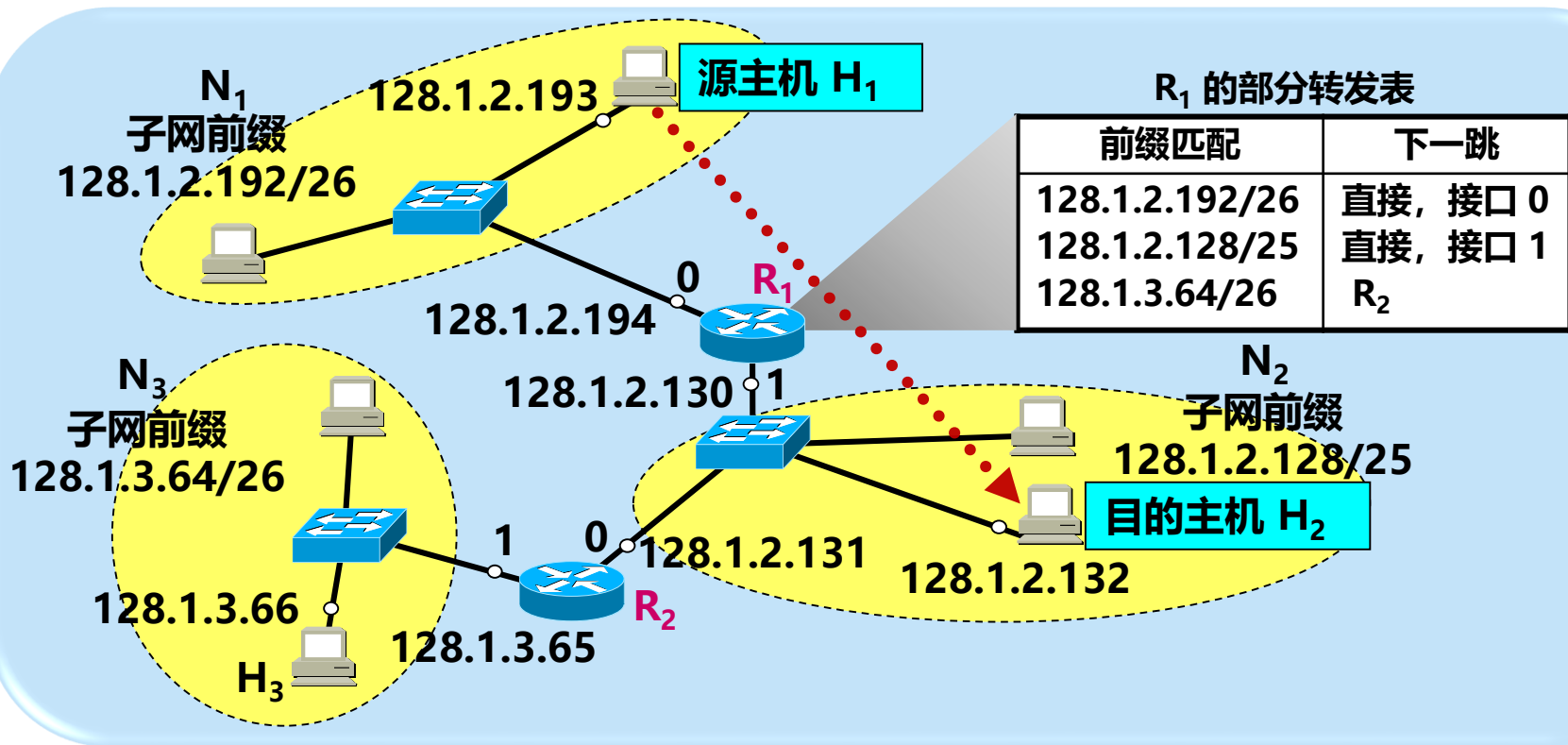


路由器 R<sub>2</sub> 的路由表

目的主机所在的网络	下一跳地址
20.0.0.0/20	直接交付, 接口 0
30.0.0.0/20	直接交付, 接口 1
10.0.0.0/20	20.0.0.7
40.0.0.0/20	30.0.0.1



## 4.3 IP 层转发分组的过程



主机 H<sub>1</sub> 发送出的、目的地址是 128.1.2.132 的分组是如何转发的？

## 4.3 IP 层转发分组的过程

H<sub>1</sub> 首先检查 128.1.2.132 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<sub>1</sub>。

N<sub>1</sub> 的网络地址为 128.1.2.192

N<sub>1</sub> 的网络掩码为 /26 = 255.255.255.192

目的地址与网络掩码 255.255.255.192

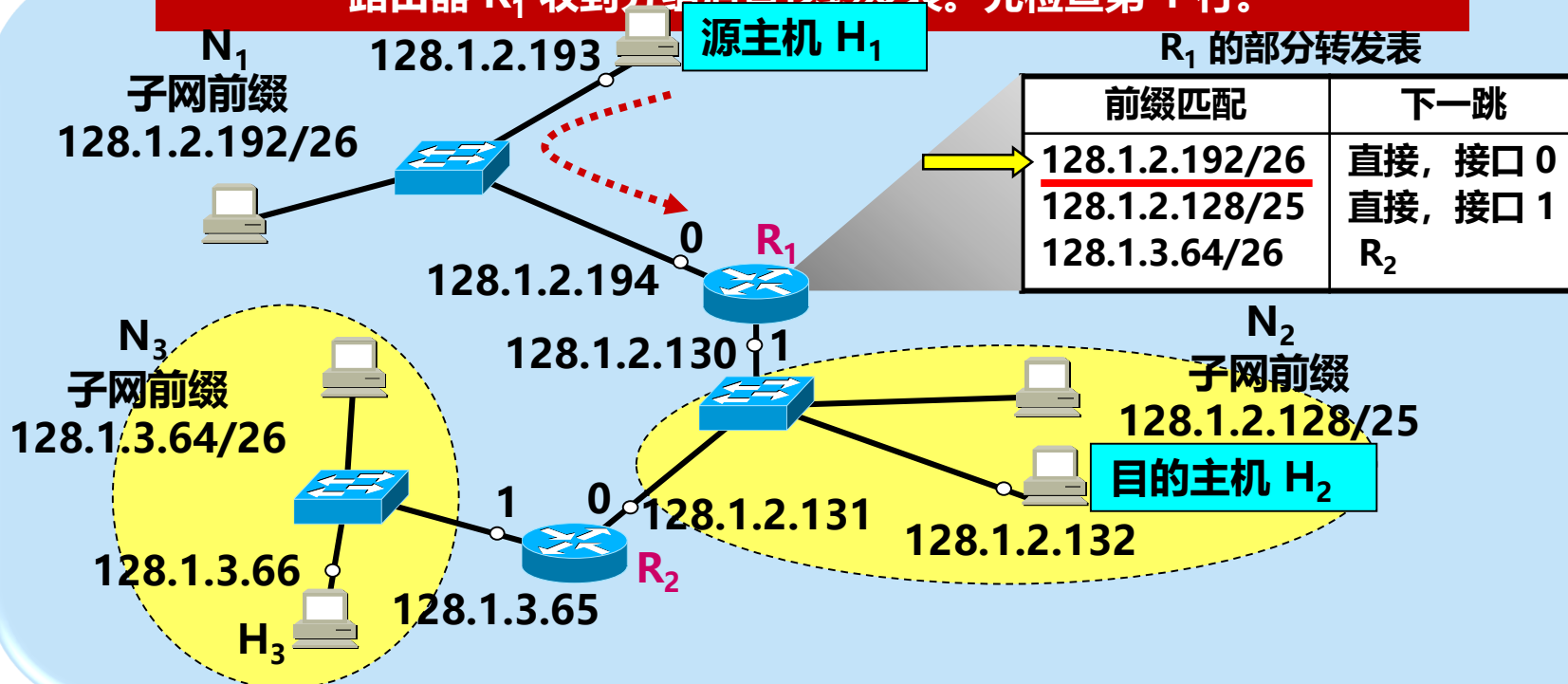
逐比特 AND 128. 1 . 2 .138

128. 1 . 2 .128 ≠ H<sub>1</sub> 的网络地址

源主机 H<sub>1</sub> 必须把分组发送给路由器 R<sub>1</sub>。

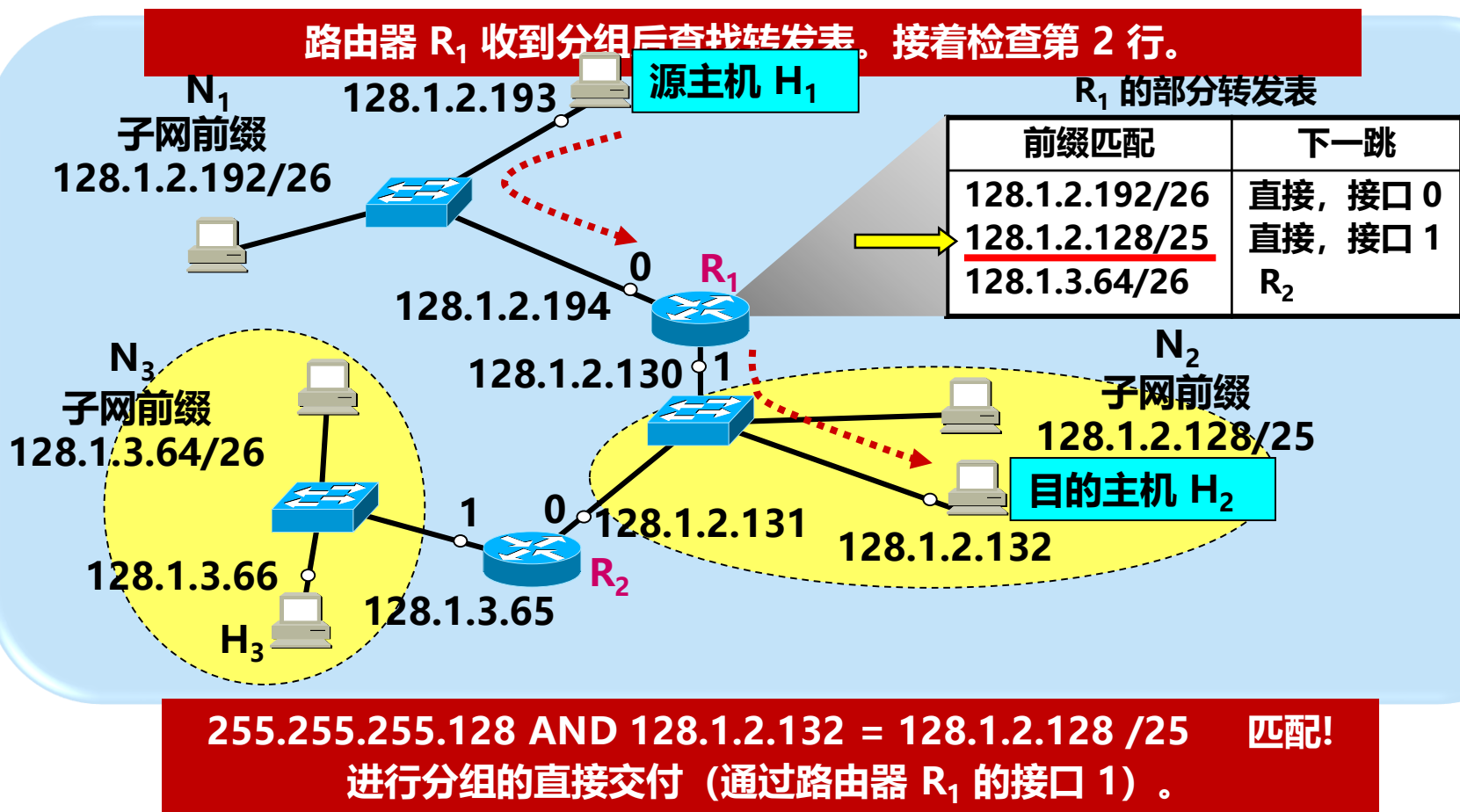
## 4.3 IP 层转发分组的过程

路由器 R<sub>1</sub> 收到分组后查找转发表。先检查第 1 行。

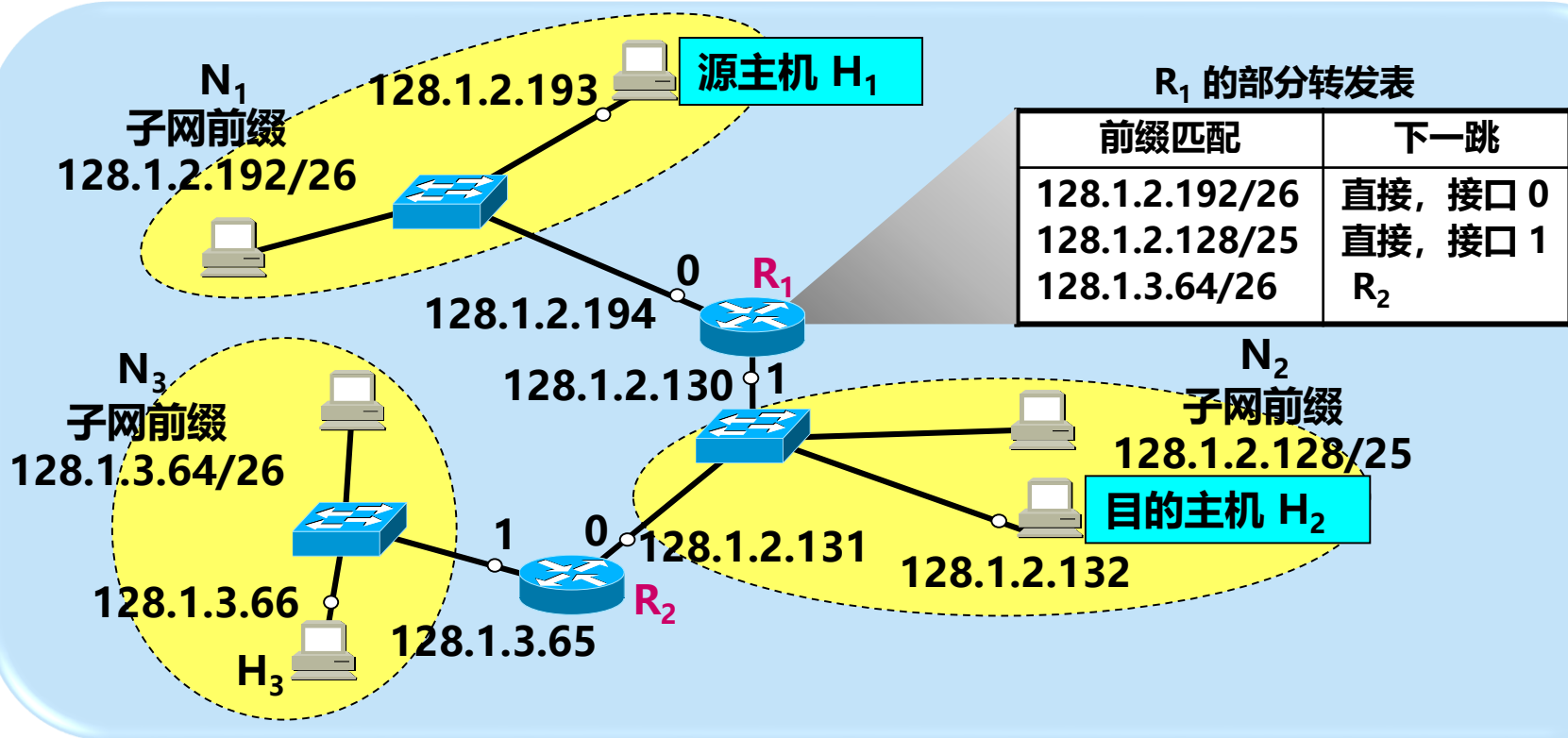


255.255.255.192 AND 128.1.2.132 = 128.1.2.128 /26 不匹配!

## 4.3 IP 层转发分组的过程



# 最长前缀匹配

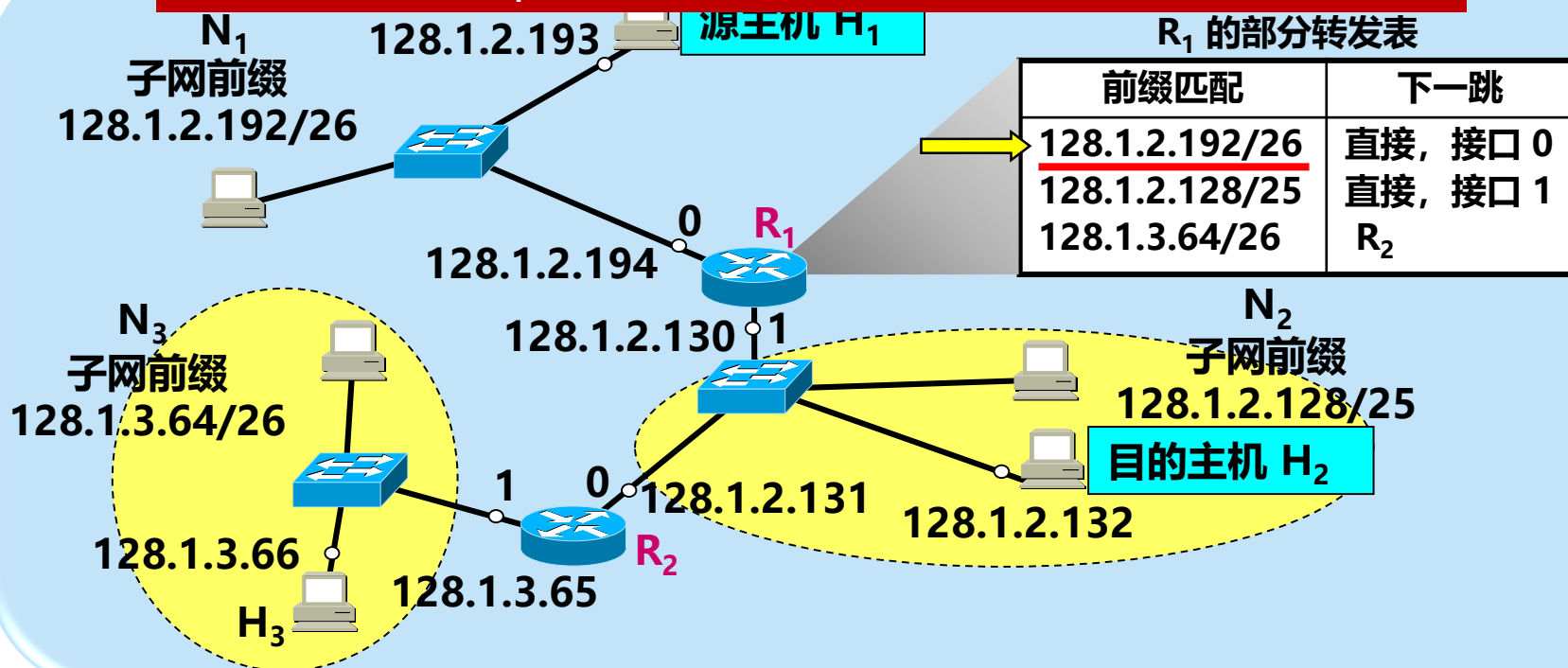


路由器  $R_1$  如何转发目的地址是 128.1.2.194 的分组?



# 最长前缀匹配

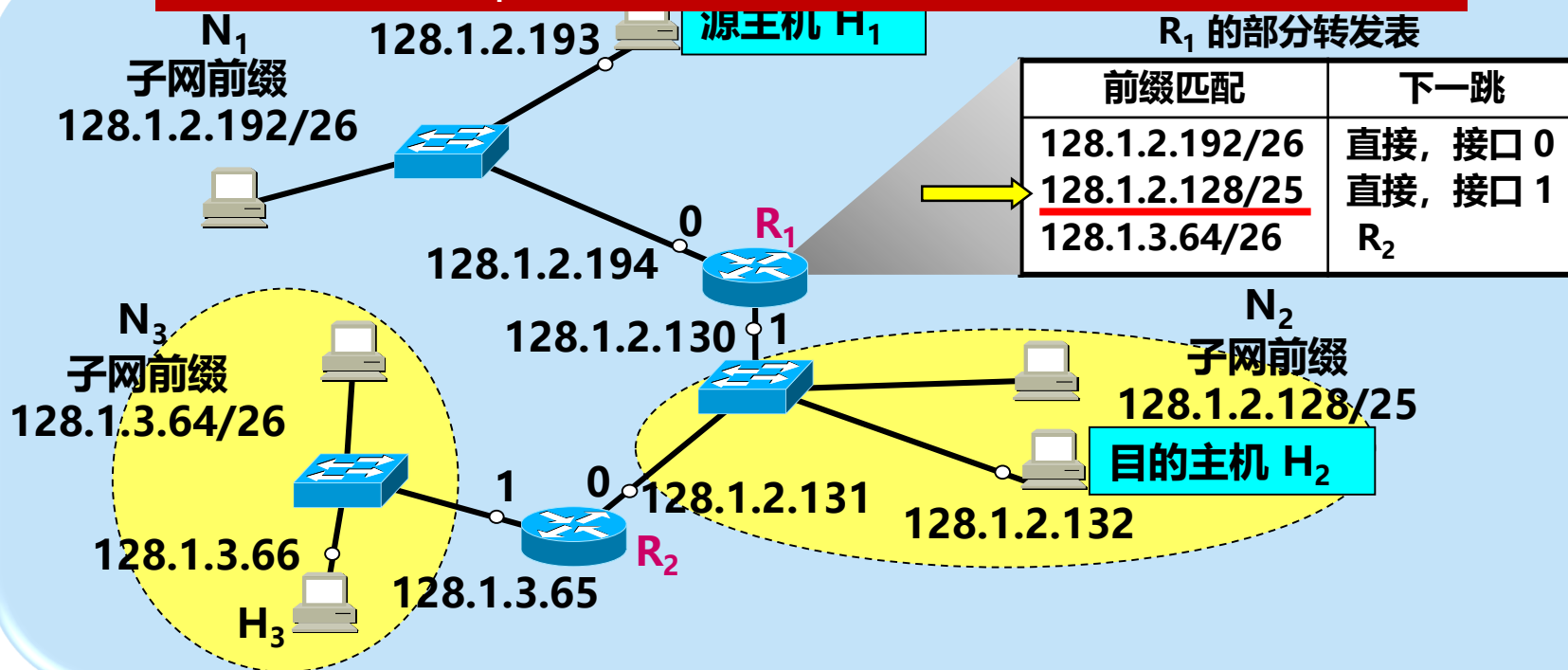
路由器 R<sub>1</sub> 收到分组后查找转发表。先检查第 1 行。



255.255.255.192 AND 128.1.2.194 = 128.1.2.192 /26 匹配!

# 最长前缀匹配

路由器 R<sub>1</sub> 收到分组后查找转发表。接着检查第 2 行。



255.255.255.128 AND 128.1.2.194 = 128.1.2.128 /25 匹配!

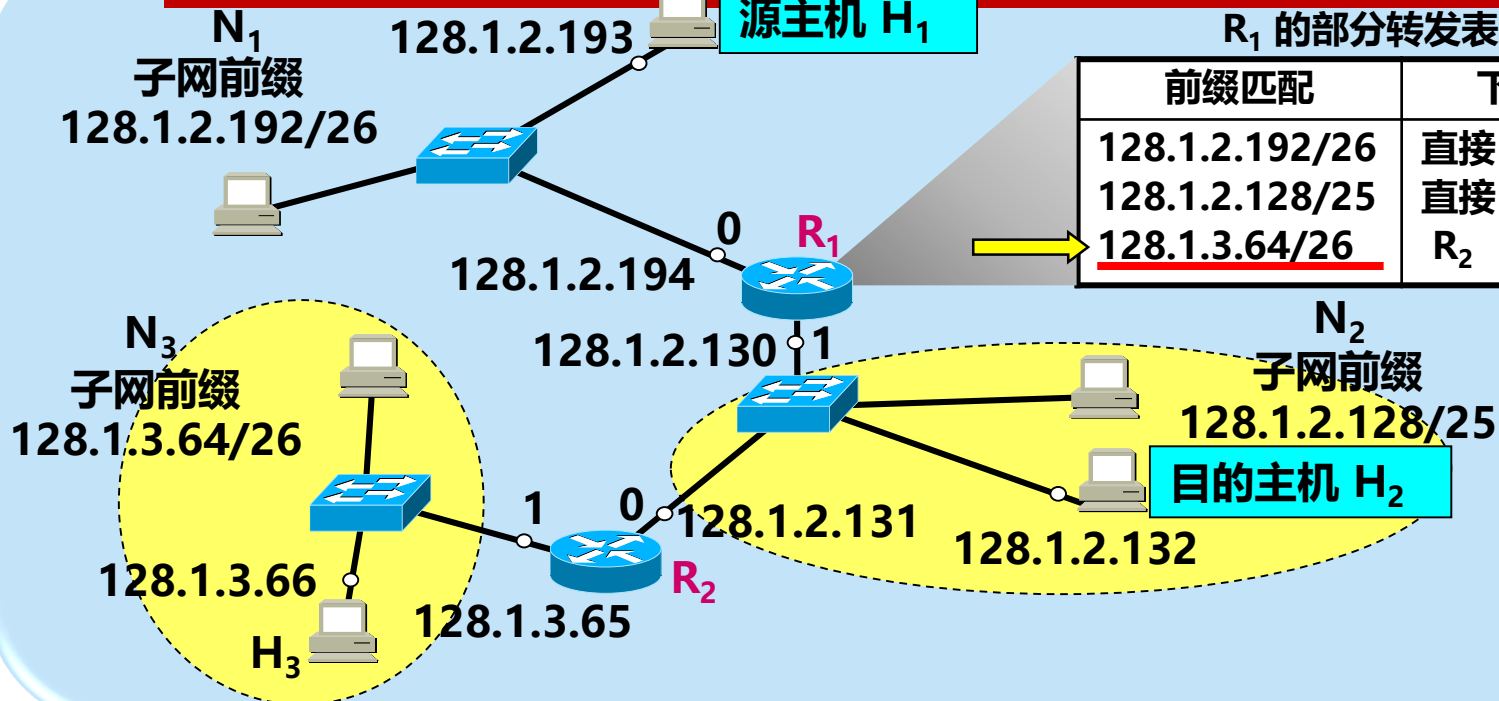
# 最长前缀匹配

路由器 R<sub>1</sub> 收到分组后查找转发表。接着检查第 3 行。

源主机 H<sub>1</sub>

R<sub>1</sub> 的部分转发表

前缀匹配	下一跳
128.1.2.192/26	直接, 接口 0
128.1.2.128/25	直接, 接口 1
<u>128.1.3.64/26</u>	R <sub>2</sub>





# 最长前缀匹配

- 路由表中的每个项目主要由“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), 又称为最长匹配或最佳匹配。
- 网络前缀越长，其地址块就越小，因而路由就越具体(more specific)。



# 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和测试网络，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。 (a.b.c.d/32)

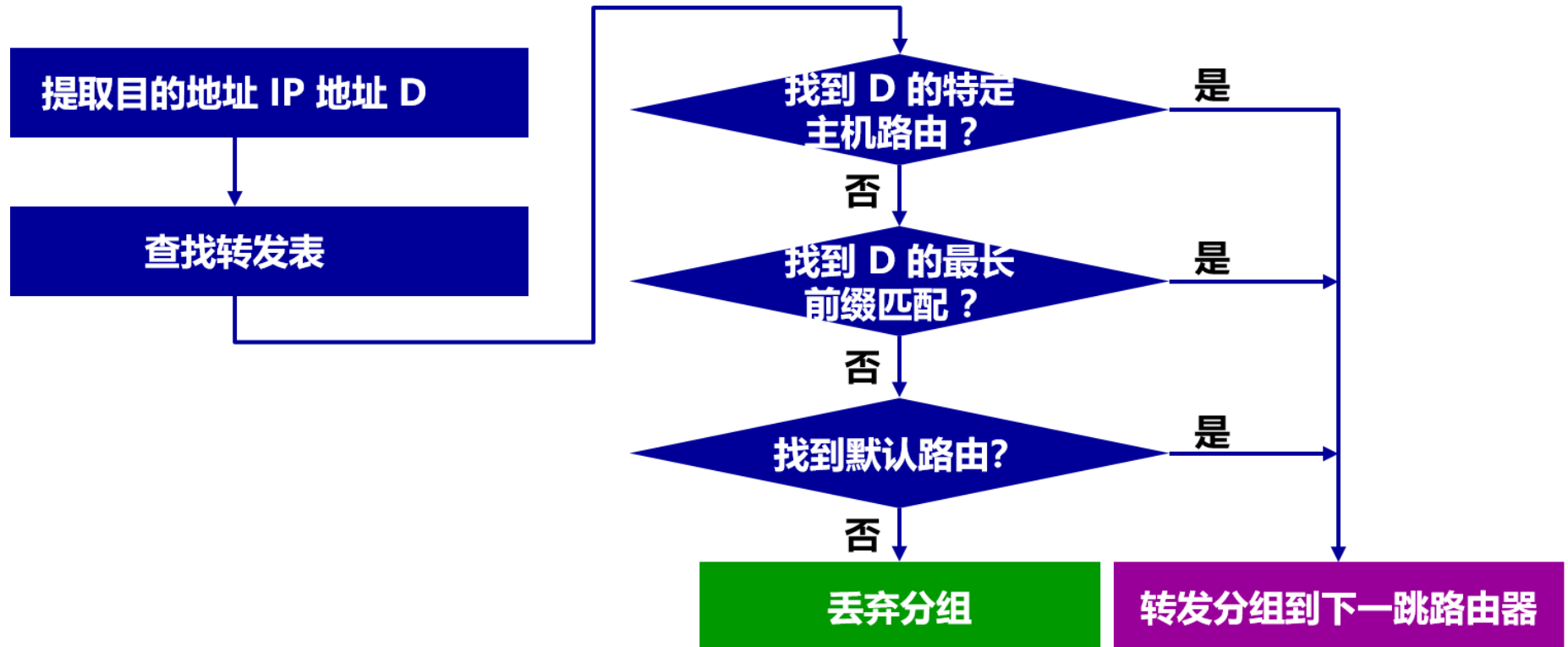


# 默认路由(default route)

---

- 默认路由:一个当别的路由在路由表中未被找到的时候使用的路由。  
(0.0.0.0/0)

# 分组转发算法



# 指引

- 网络层提供的两种服务
- 网际协议IP
  - 虚拟互联网络
  - IP地址
  - IP地址与MAC地址
  - 地址解析协议ARP
  - IP数据包的格式
- IP层转发分组的过程
- 网际控制报文协议 ICMP
- 互联网的路由选择协议
- IPV6







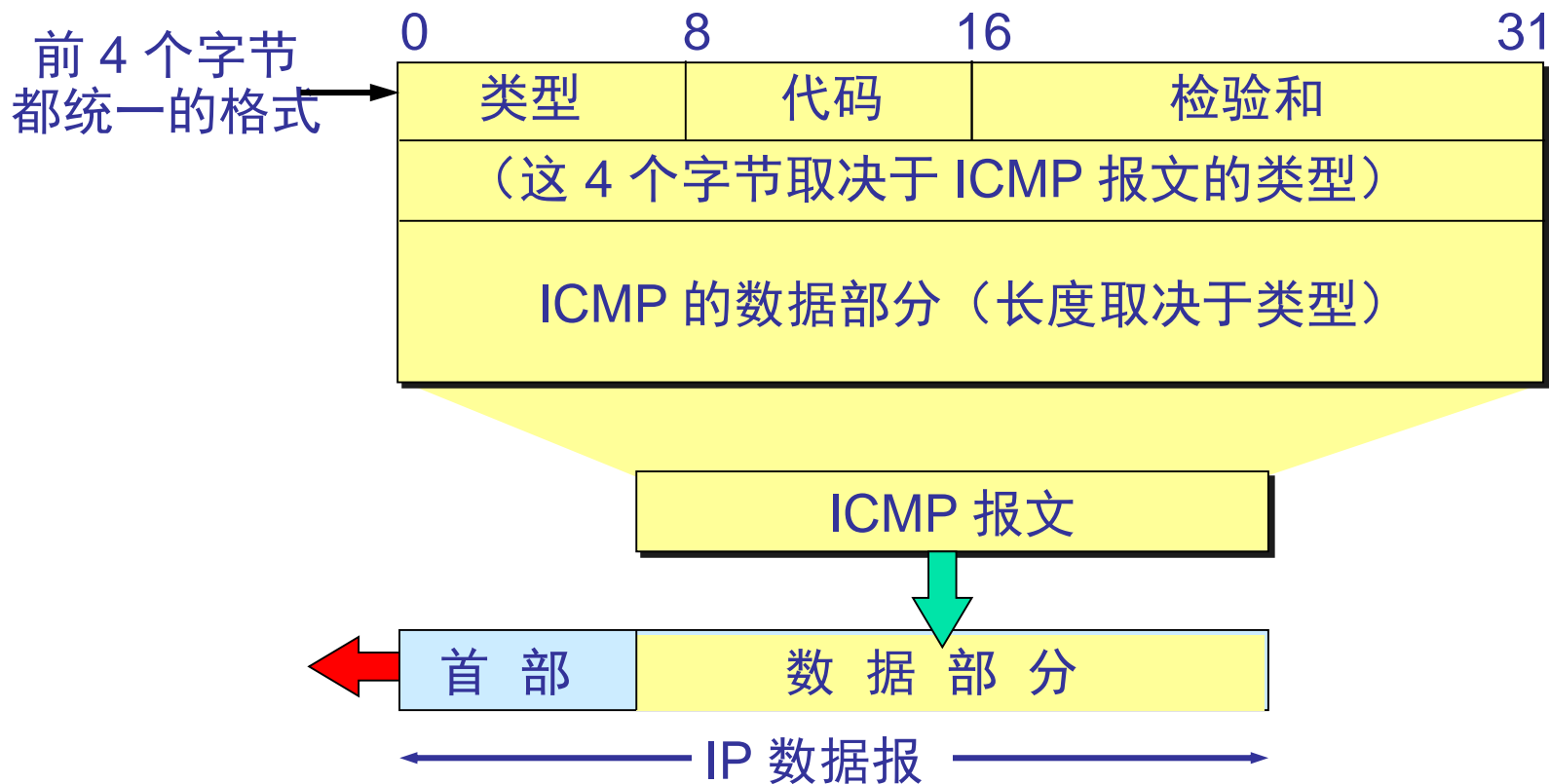
## 4.4 网际控制报文协议 ICMP

(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 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。

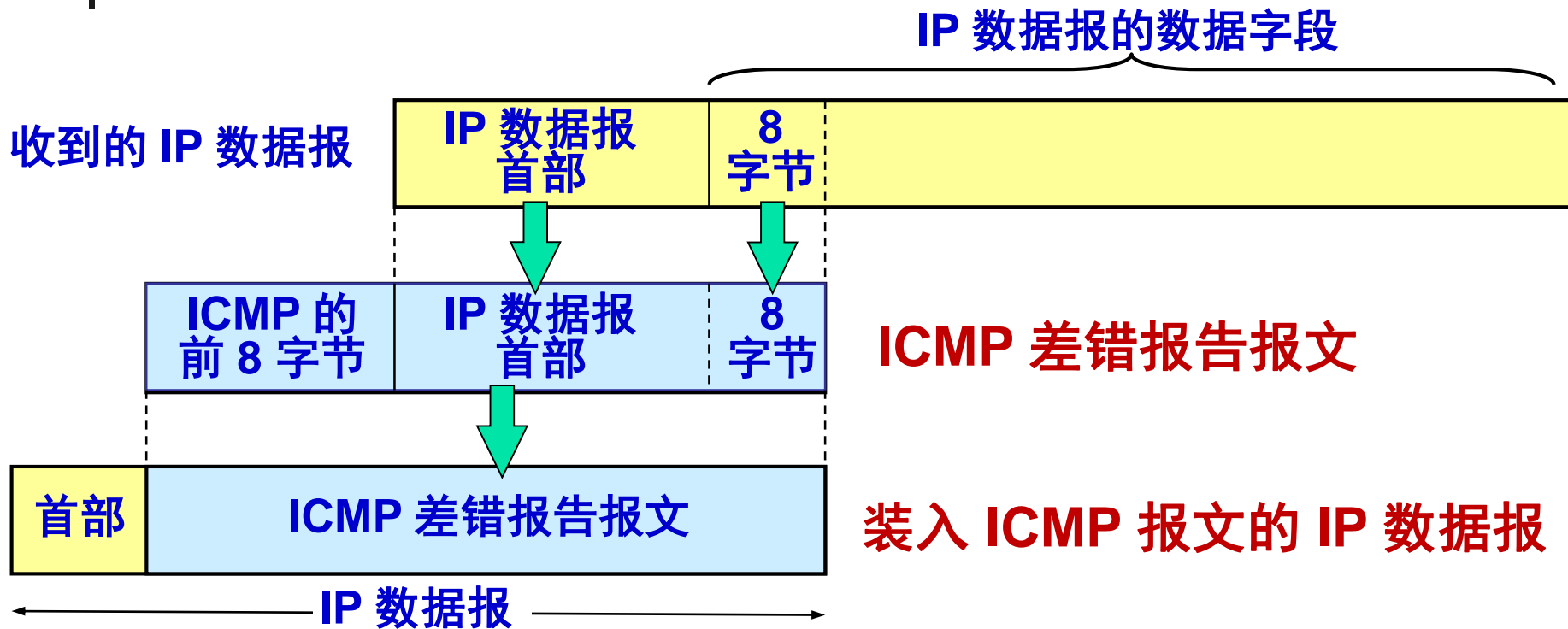


# ICMP 差错报告报文

---

- 终点不可达
- 时间超过
- 参数问题
- 改变路由（重定向）

# ICMP 差错报告报文的数据字段的内容





# ICMP 询问报文有两种

---

- 回送请求和回答报文:由主机或路由器向一个特定的目的主机发出的询问,收到此报文的机器必须给源主机发送ICMP回送回答报文。
  - 这种询问报文用来测试目的站是否可达以及了解其有关状态。
- 时间戳请求和回答报文:请某个主机或路由器回答当前的日期和时间。

## 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]

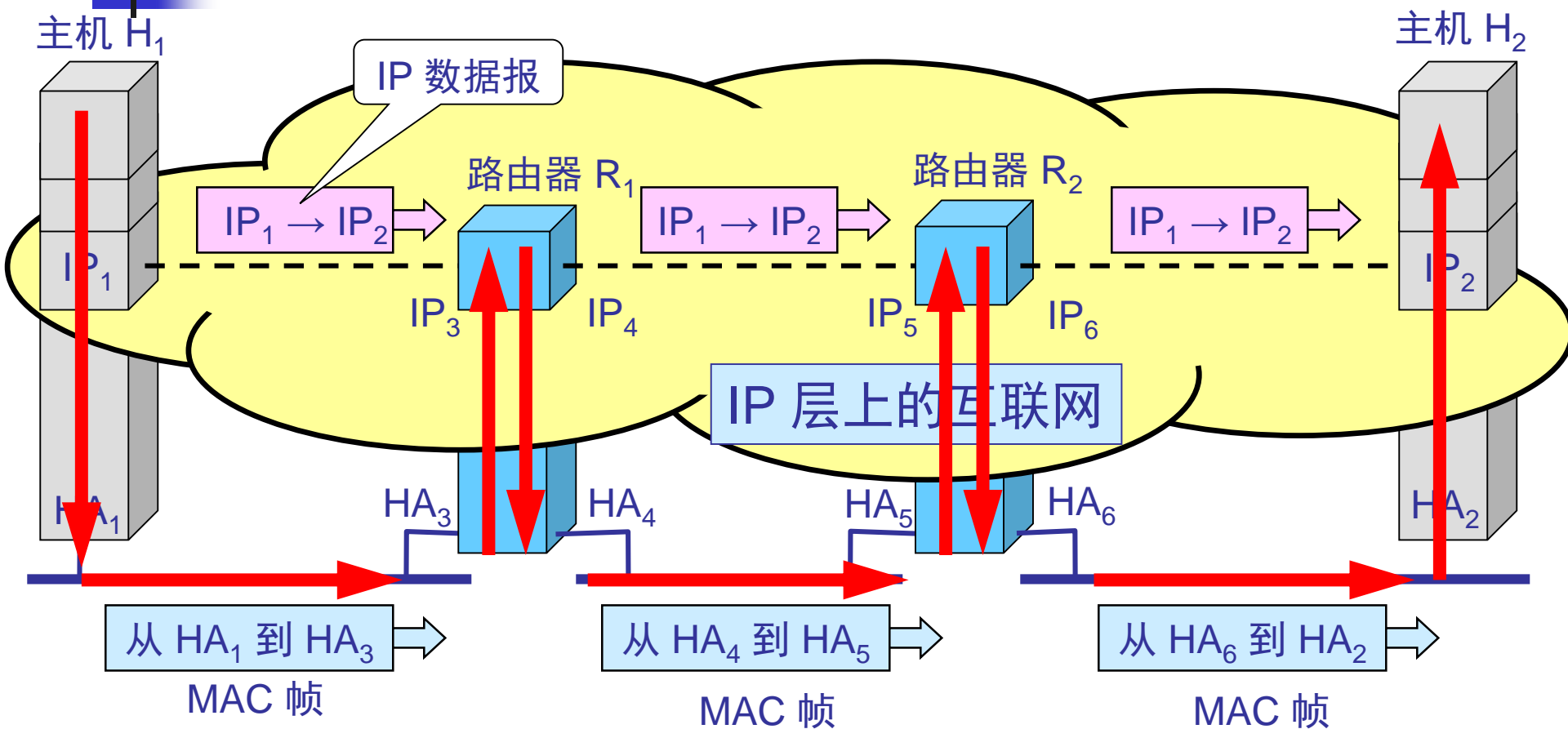
```
Trace complete.
```

# 指引

- 网络层提供的两种服务
- 网际协议IP
  - 虚拟互联网络
  - IP地址
  - IP地址与MAC地址
  - 地址解析协议ARP
  - IP数据包的格式
- IP层转发分组的过程
- 网际控制报文协议 ICMP
- 互联网的路由选择协议



# IP 层转发分组的过程





## 4.5 互联网的路由选择协议

### 4.5.1 有关路由选择协议的几个基本概念

---

#### 1. 理想的路由算法

- 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。



# 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

---

- 互联网采用动态分层次路由选择协议。
  - 互联网的规模非常大。
  - 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，但同时还希望连接到互联网上。



# 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。
- 一个 AS 对其他 AS 表现出的的是一个**单一的和一致的路由选择策略**。



# 互联网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。



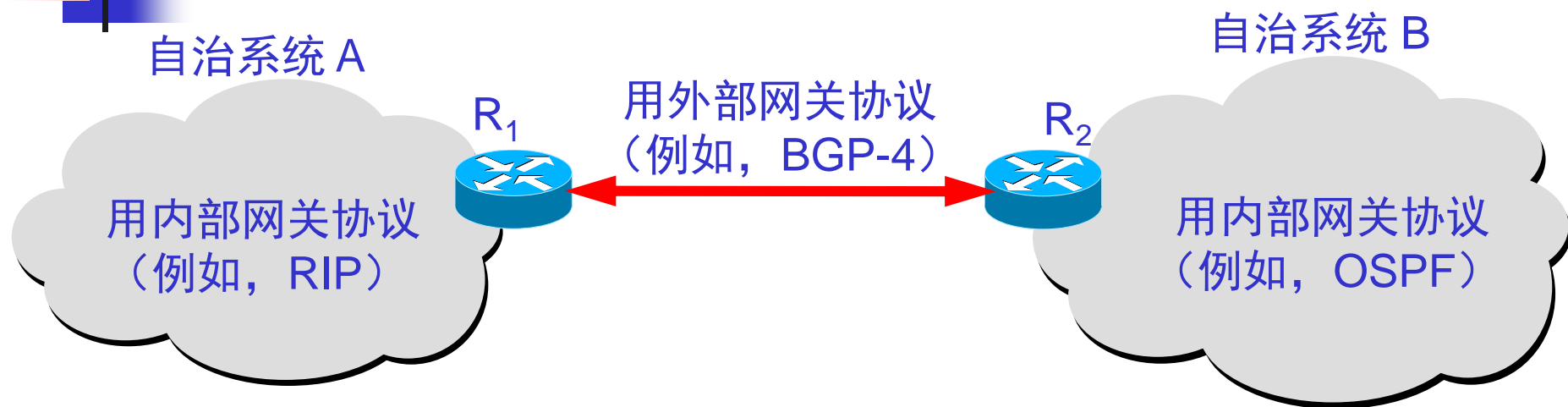


# 互联网两大类路由选择协议

---

- **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。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)



## 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

---

- 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” 也称为 “**跳数**” (hop count)。



# “距离”的定义

---

- RIP 认为一个好的路由就是它通过的路由器的数目少，即“距离短”。
- 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。

## 2. 距离向量算法

收到相邻路由器x的一个 RIP 报文：

①到目的网N ②距离d ③ 下一跳路由器

(1) 先修改此 RIP 报文中的所有项目：把“下一跳”字段中的地址都改为 X，并把所有的“距离”字段的值加 1。

(2) 对修改后的 RIP 报文中的每一个项目，重复以下步骤：

若项目中的目的网络N不在路由表中，则把该项目加到路由表中。

## 2. 距离向量算法

否则: 若下一跳字段给出的路由器地址是同样的 $x$ , 则把收到的项目替换原路由表中的项目。

否则: 若收到项目中的距离小于路由表中的距离, 则进行更新,

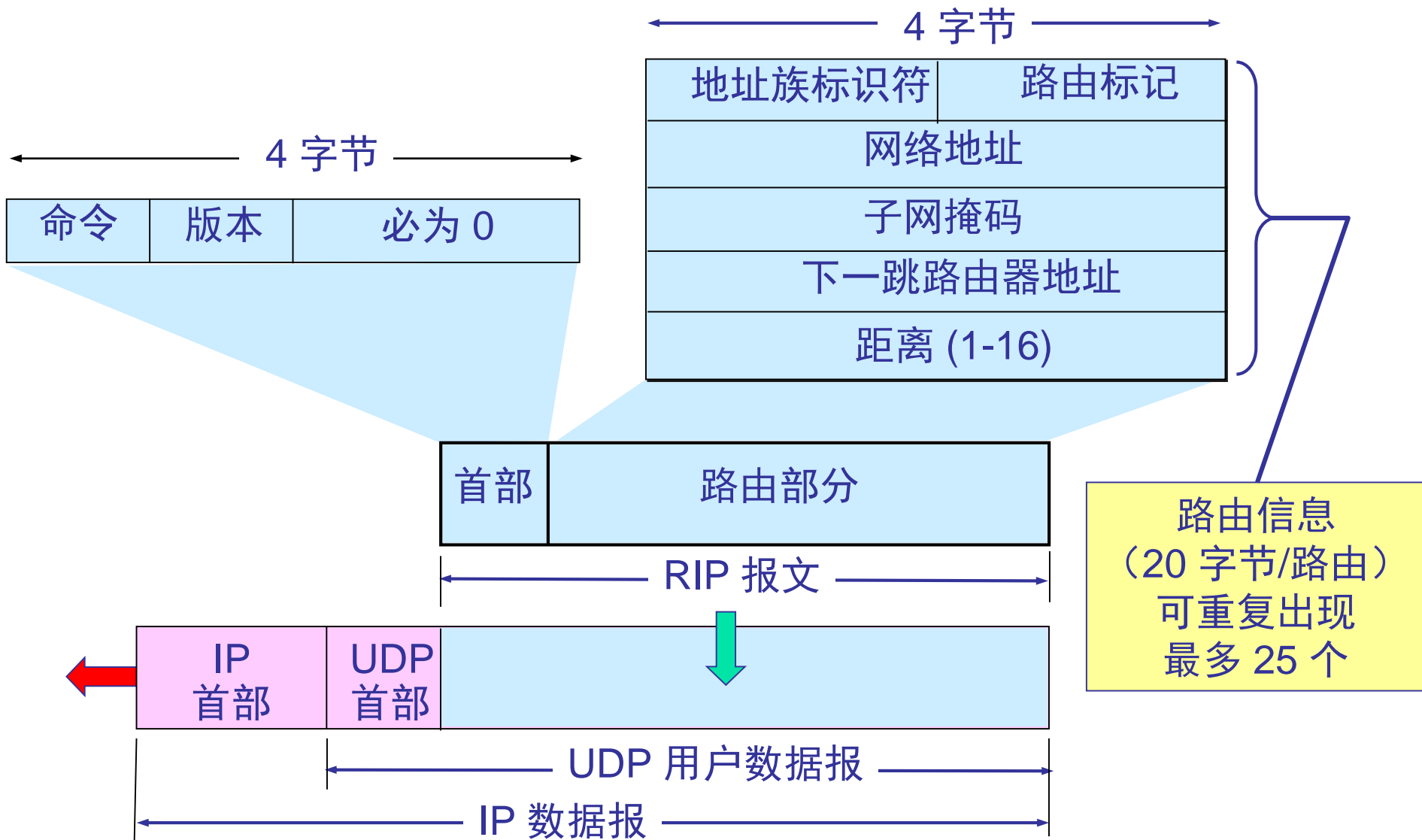
否则: 什么也不做。

(3) 若 3 分钟还没有收到相邻路由器更新路由表, 则把此相邻路由器记为不可达路由器, 即将距离置为16 (距离为16表示不可达)。

(4) 返回。



### 3. RIP2 协议的报文格式



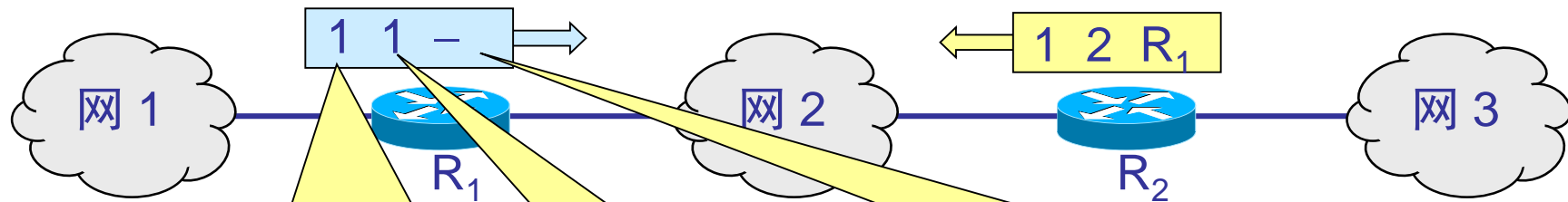


# RIP 协议的优缺点

---

- 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）。
- 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。

正常情况



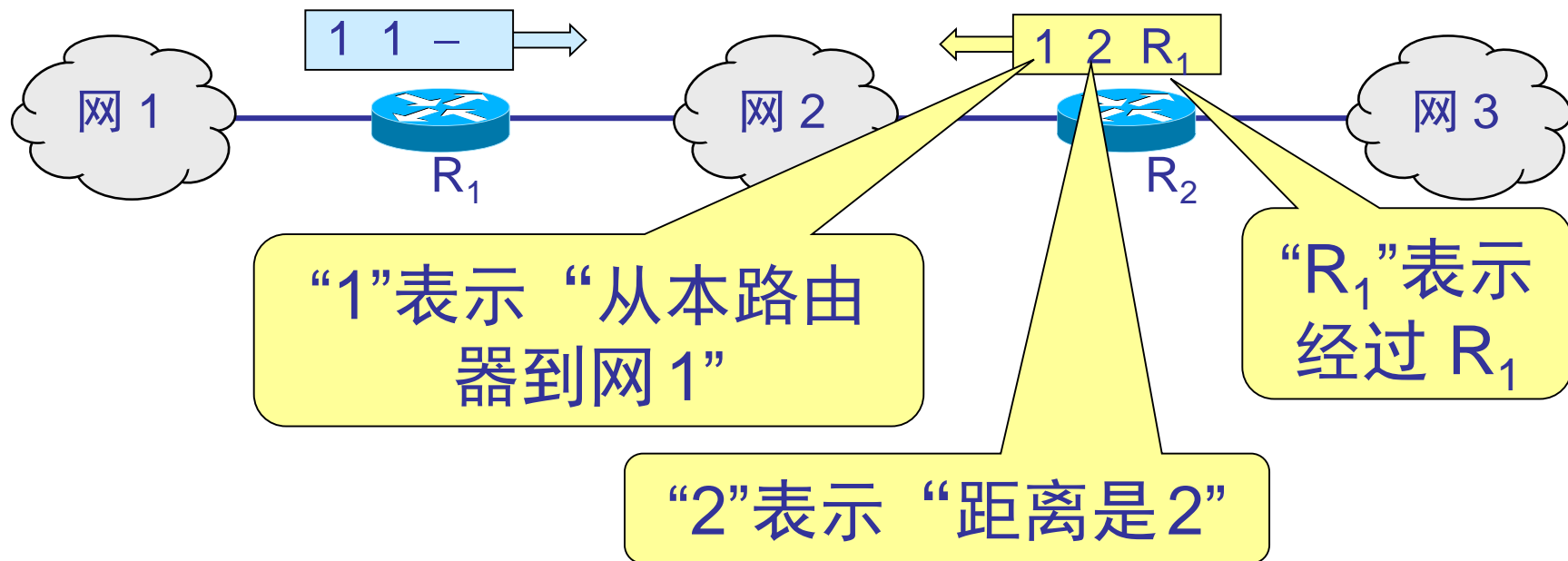
“1”表示“从本路由器到网 1”

“-”表示“直接交付”

“1”表示“距离是 1”

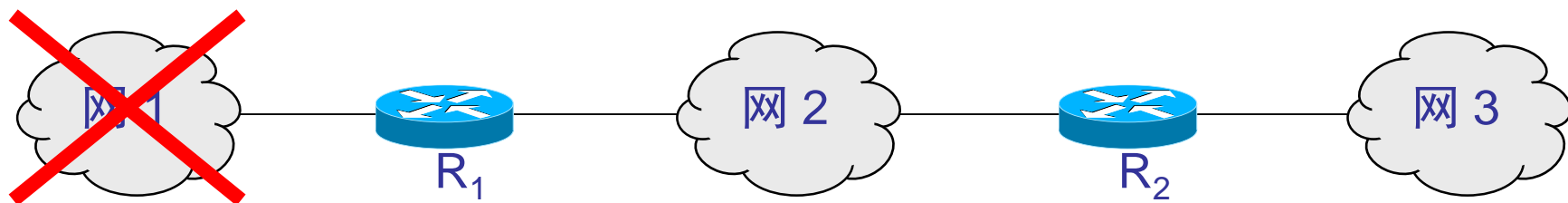
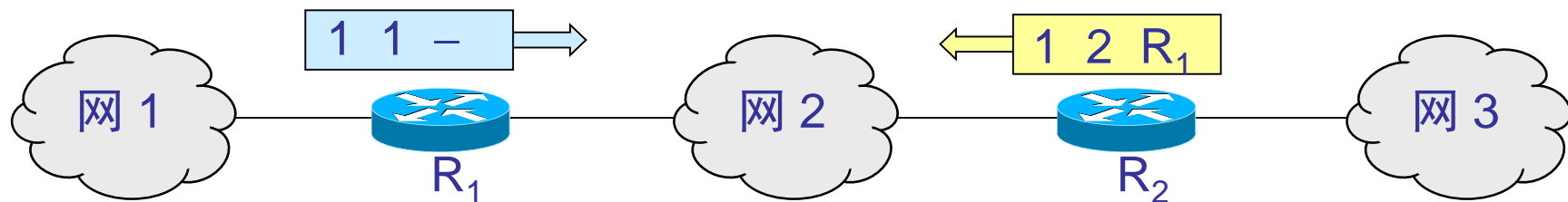
$R_1$  说：“我到网 1 的距离是 1，是直接交付。”

正常情况



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



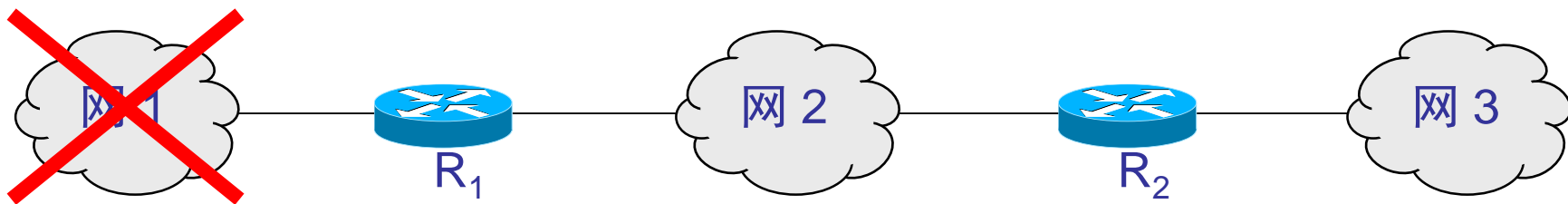
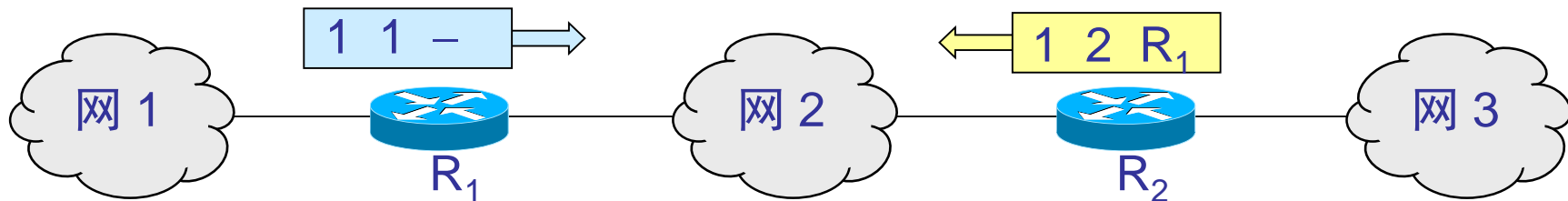
网 1 出了故障



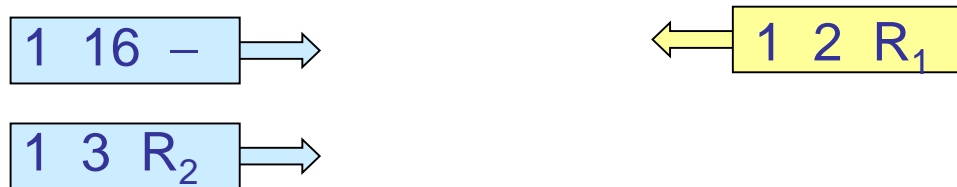
$R_1$  说：“我到网 1 的距离是 16（表示无法到达），是直接交付。”

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

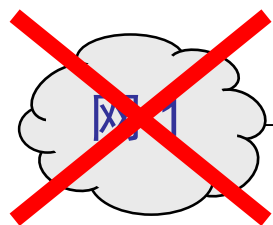
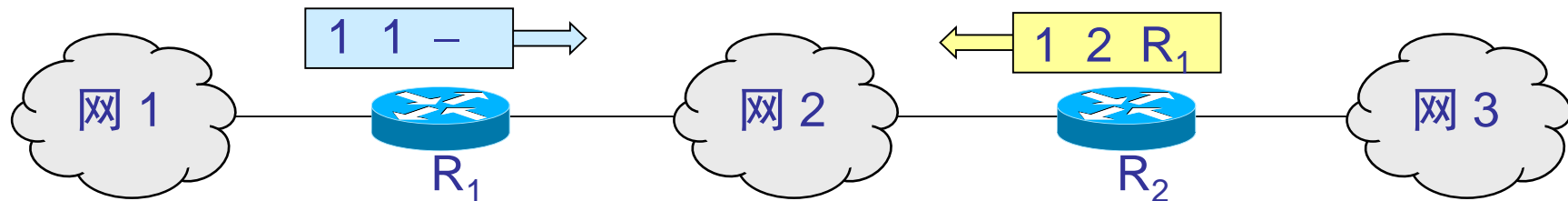


网1出了故障



R<sub>1</sub> 收到 R<sub>2</sub> 的更新报文后，误认为可经过 R<sub>2</sub> 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<sub>2</sub>”。然后将此更新信息发送给 R<sub>2</sub>。

正常情况

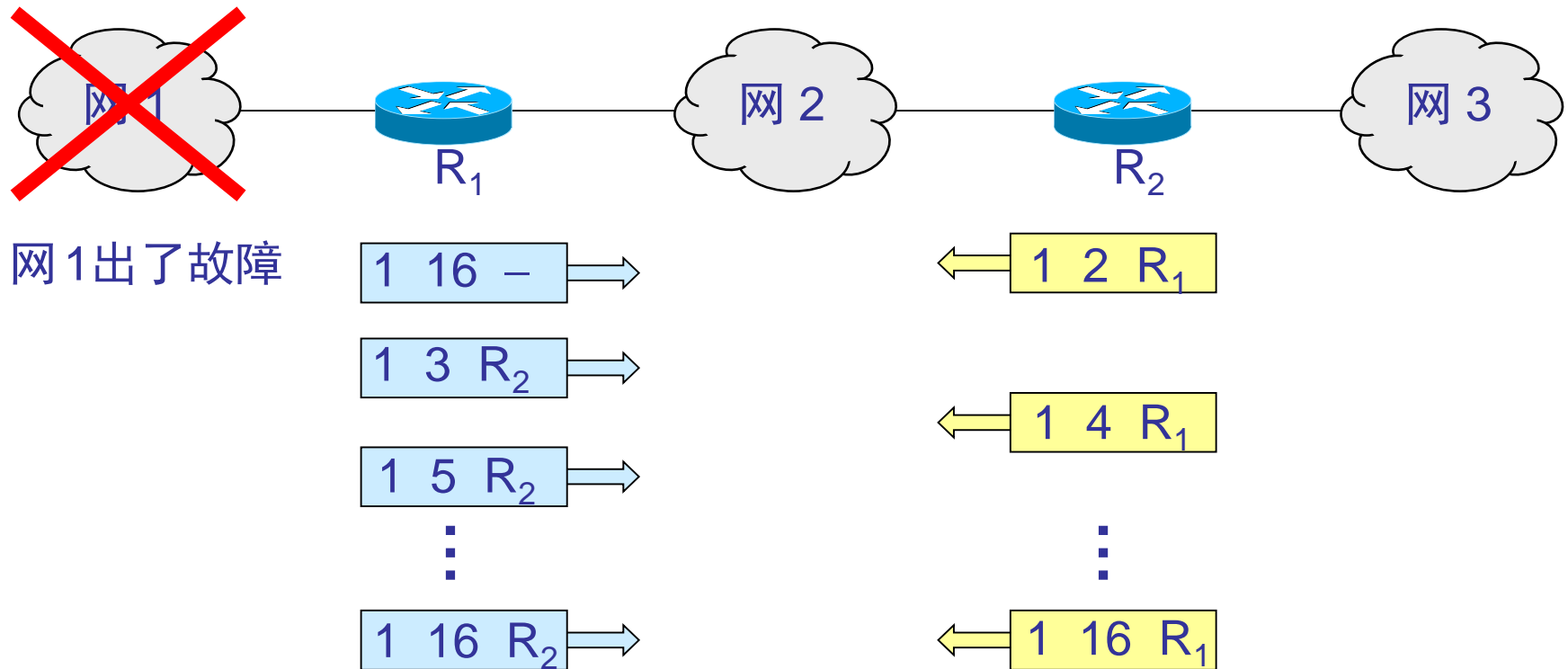


网 1 出了故障



R<sub>2</sub> 以后又更新自己的路由表为 “1, 4, R<sub>1</sub>”, 表明 “我到网 1 距离是 4, 下一跳经过 R<sub>1</sub>”。

这就是好消息传播得快，而坏消息传播得慢。网络出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<sub>1</sub> 和 R<sub>2</sub> 到网 1 的距离都增大到 16 时，R<sub>1</sub> 和 R<sub>2</sub> 才知道网 1 是不可达的。



## 4.5.3 内部网关协议 OSPF

### (Open Shortest Path First)

---

#### 1. OSPF 协议的基本特点

- 开放最短路径优先OSPF 协议最主要特征是使用分布式的链路状态协议。



# OSPF的三个要点

- 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(flooding)。
- 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或每隔一段时间，路由器才用洪泛法向所有路由器发送此信息。

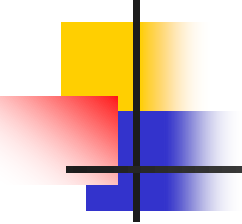


# 链路状态数据库

## (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）。
- 使用Dijkstra算法构造出路由表。

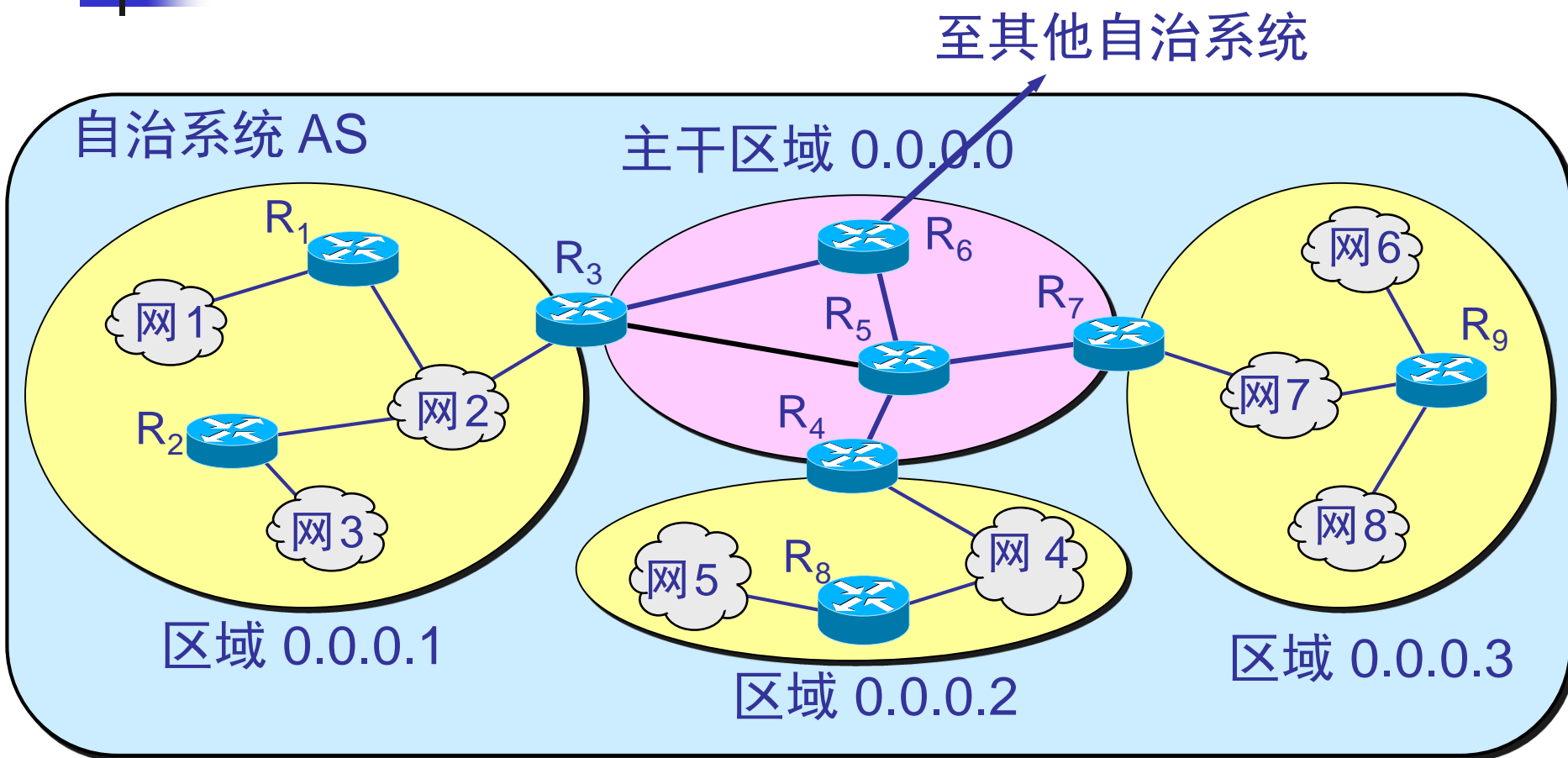


# OSPF 的区域(area)

---

- 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。

# OSPF 划分为两种不同的区域



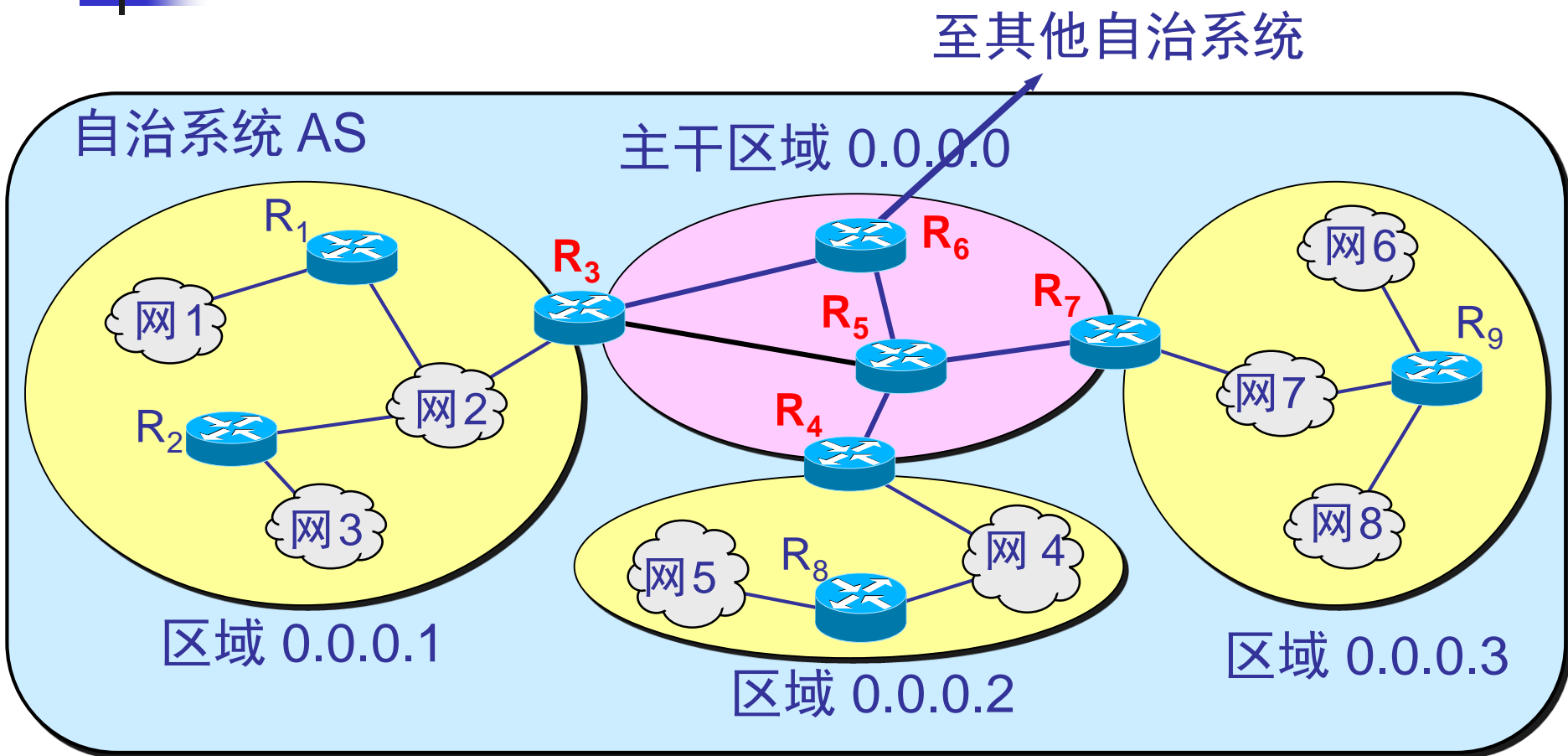


# 划分区域

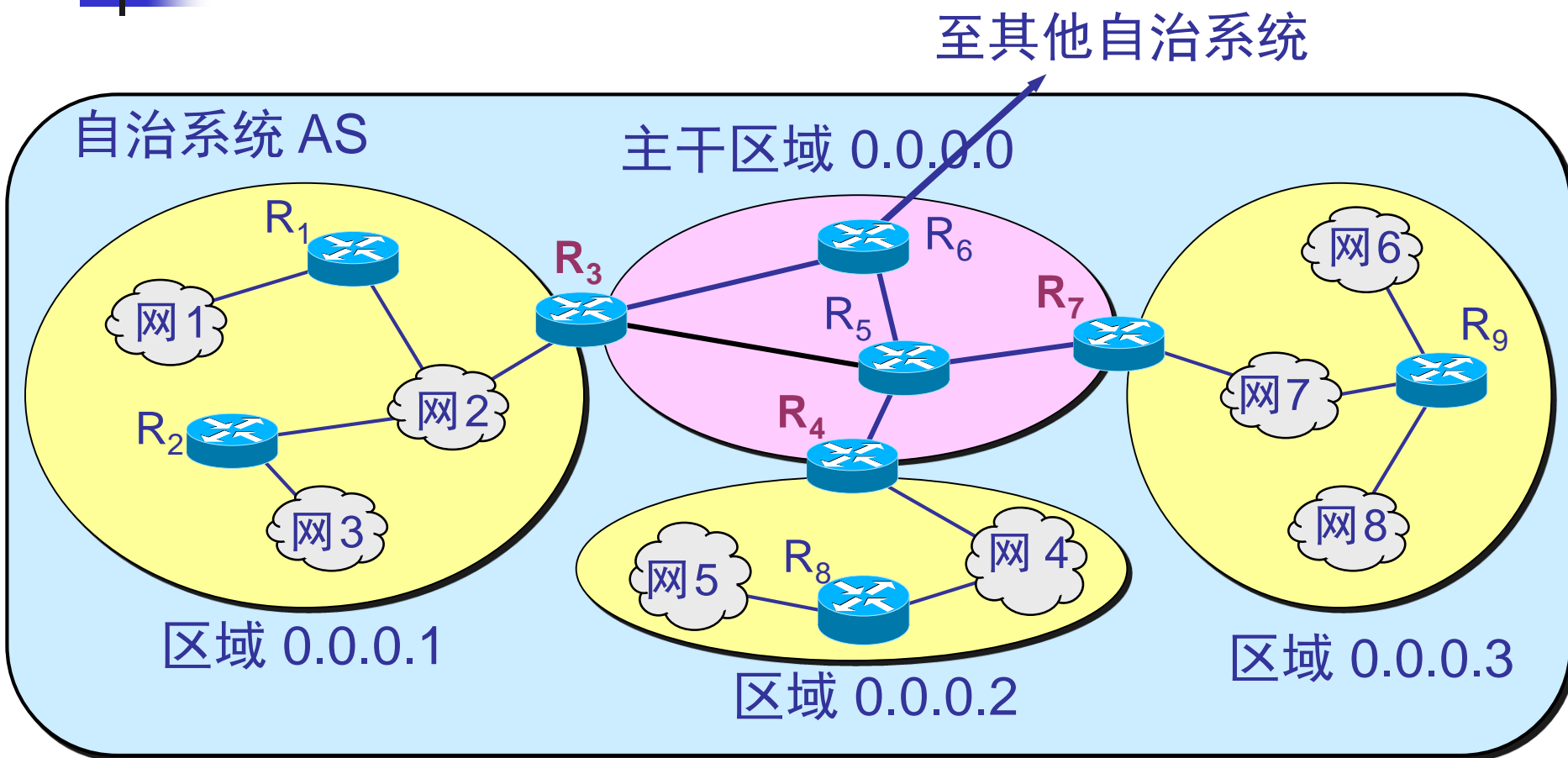
---

- 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。

# 主干路由器



# 区域边界路由器







## 2. OSPF 的五种分组类型

---

- 类型1， 问候(Hello)分组,用来发现和维持邻站的可达性。
- 类型2， 数据库描述(Database Description)分组,向邻站给出自己的链路状态数据库中的所有链路状态项目的摘要信息。

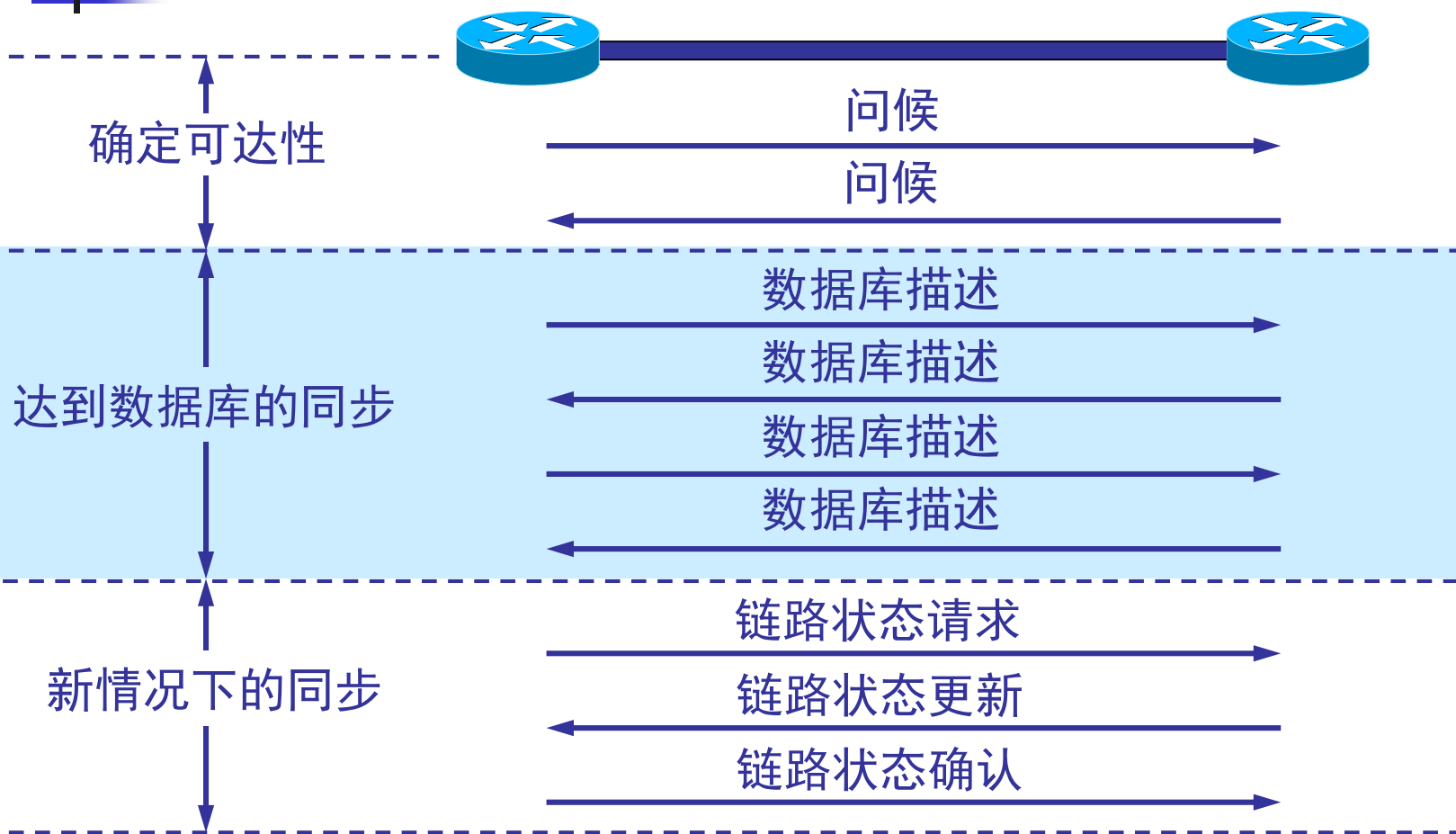


# OSPF 的五种分组类型

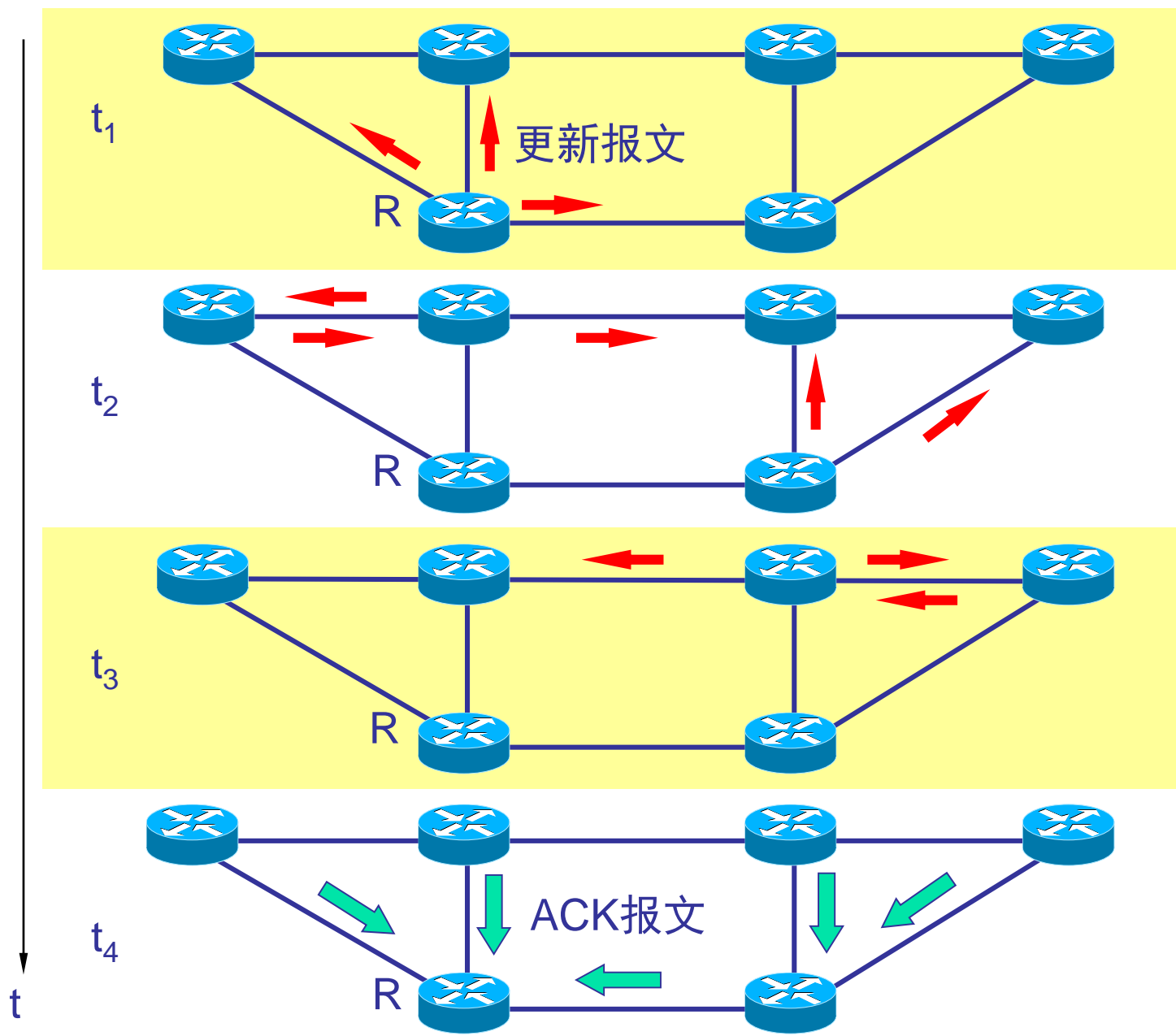
---

- 类型3，链路状态请求(Link State Request)分组,向对方请求发送某些链路状态项目的详细信息。
- 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,对链路更新分组的确认。

# OSPF的基本操作



# OSPF 使用的是可靠的洪泛法





## 4.5.4 外部网关协议 BGP

---

- 边界网关协议BGP 是不同自治系统的路由器之间交换路由信息的协议。
- 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，而并非要寻找一条最佳路由。BGP采用了路径向量路由选择协议。

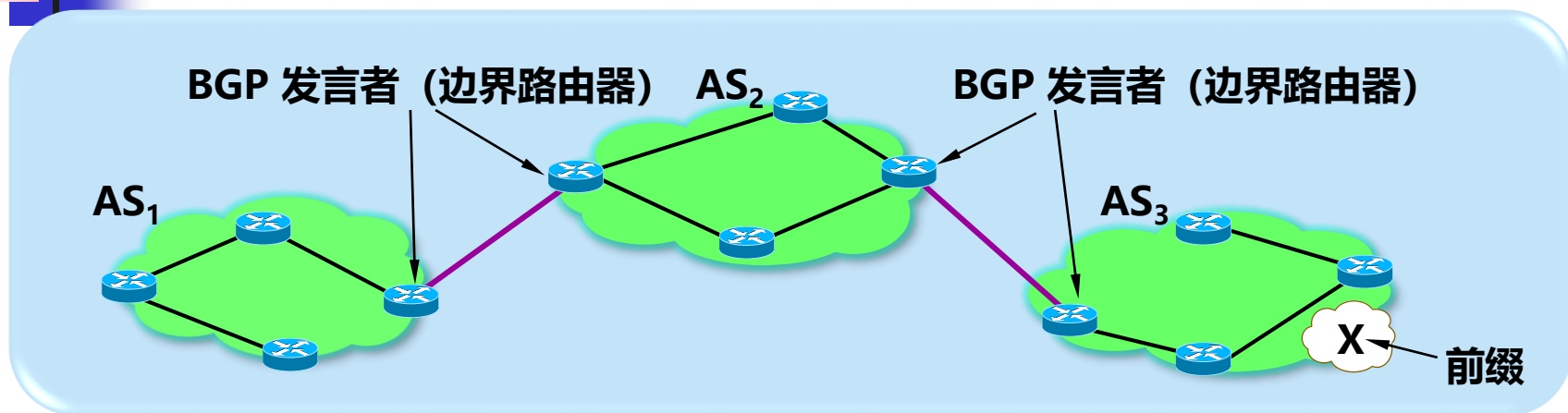


# 课后作业

---

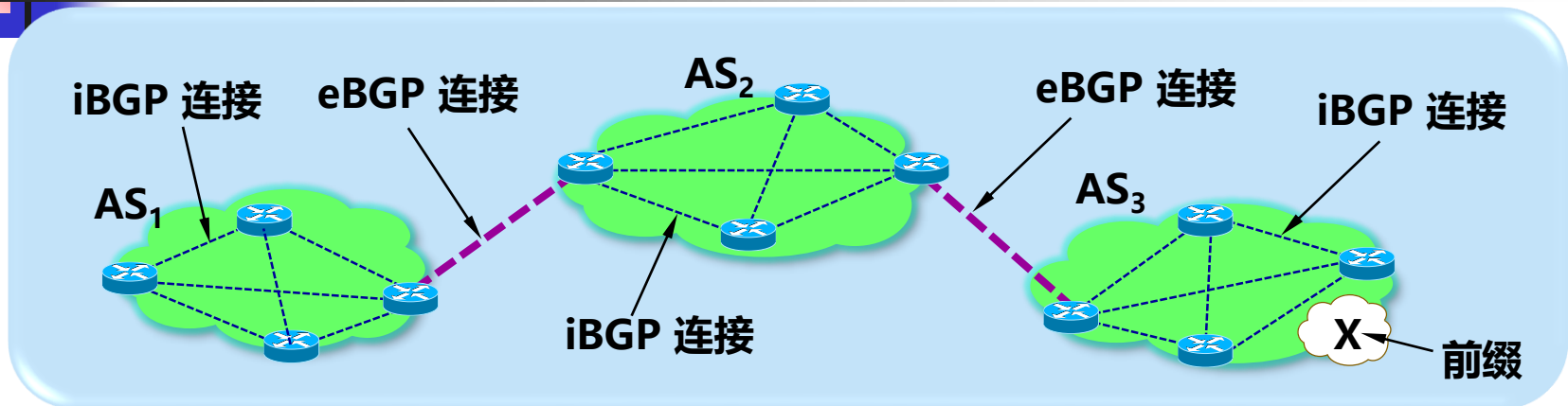
- 习题: 4-35, 4-37

# BGP 发言者 (BGP speaker)



- 对等 BGP 发言者（边界路由器）在 AS 之间交换信息

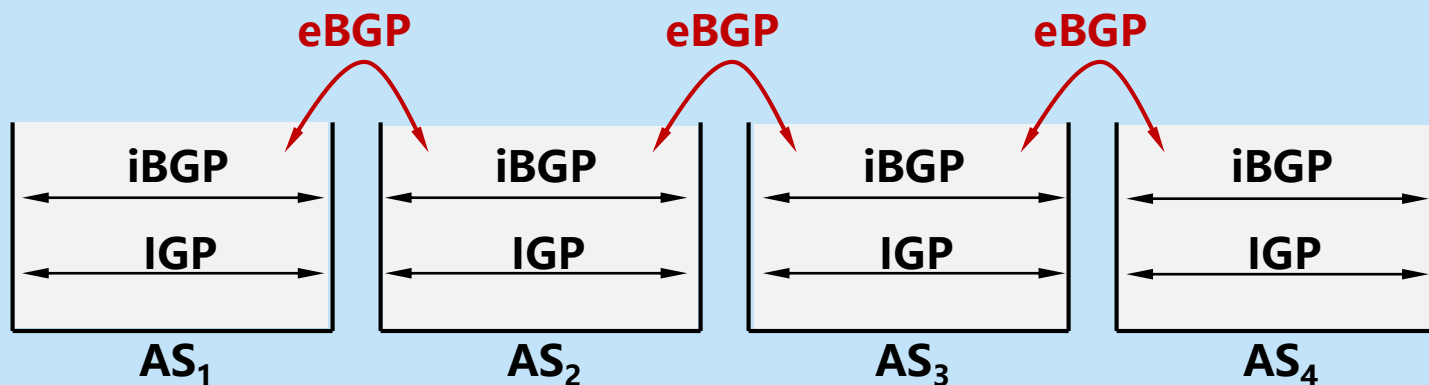
# eBGP 连接和 iBGP 连接



- eBGP (external BGP) 连接：运行 eBGP 协议，在不同 AS 之间交换路由信息。
- iBGP (internal BGP) 连接：运行 iBGP 协议，在 AS 内部的路由器之间交换 BGP 路由信息。



# IGP、iBGP 和 eBGP 的关系



- 在 AS 内部运行：内部网关协议 IGP（可以是协议 OSPF 或 RIP），协议 iBGP。
- 在 AS 之间运行：协议 eBGP。



# eBGP 和 iBGP

---

- 同一个协议 BGP，但它们在路由通告时采用的规则不同。
- 在 eBGP 连接的对等端得知的路由信息，可以通报给一个 iBGP 连接的对等端。反过来也是可以的。
- 但从 iBGP 连接的对等端得知的路由信息，则不能够通报给另一个 iBGP 连接的对等端。



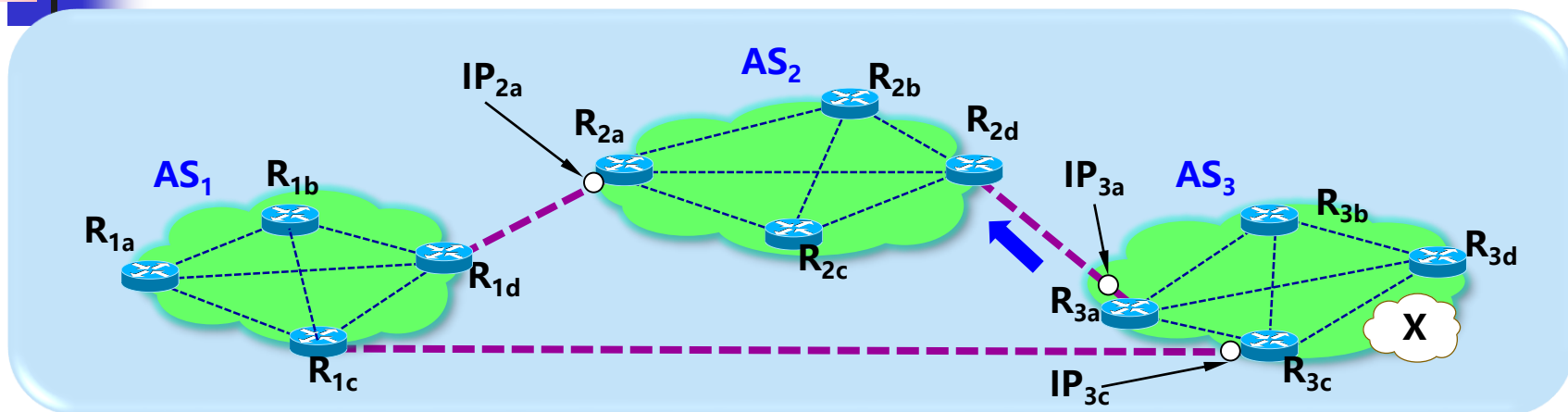
# BGP 路由信息

---

BGP 路由 = [ 前缀, BGP属性 ] = [ 前缀, AS-PATH, NEXT-HOP ]

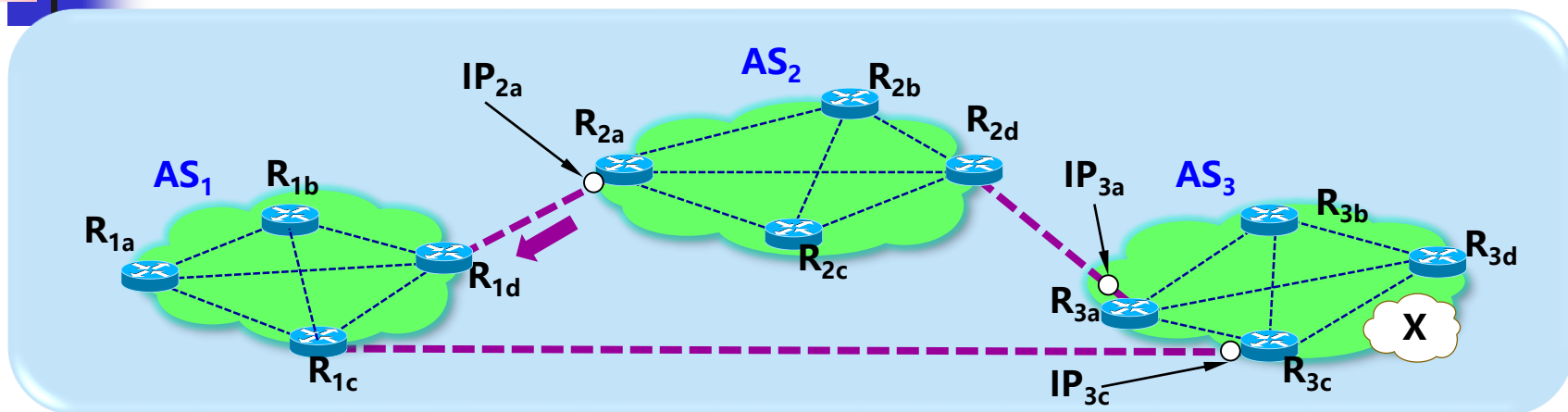
- 前缀：指明到哪一个子网（用 CIDR 记法表示）。
- BGP 属性：最重要的两个属性是
  - 自治系统路径 AS-PATH
  - 下一跳 NEXT-HOP。

# BGP 路由信息



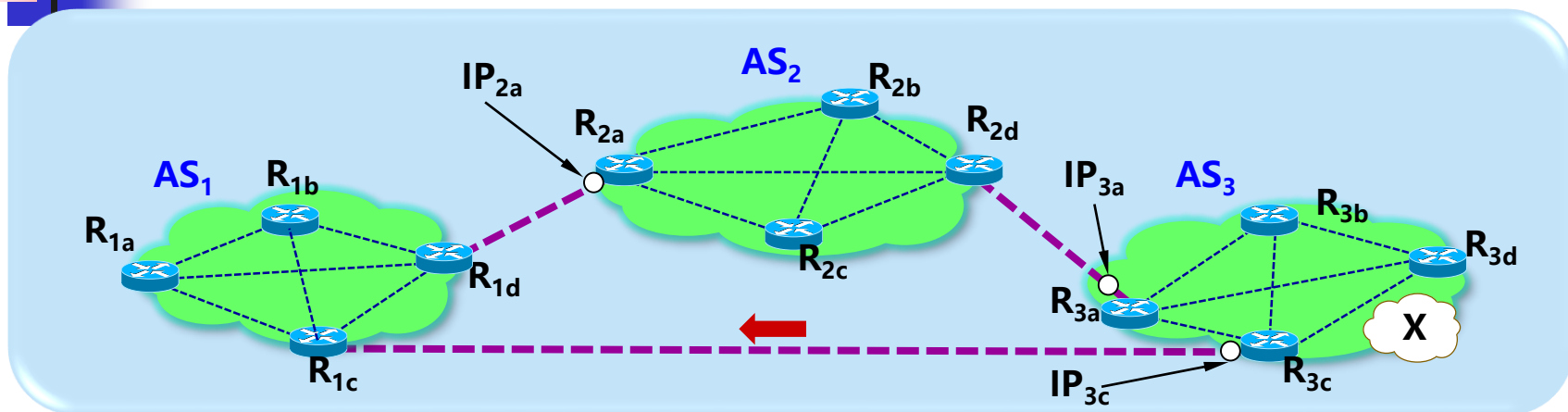
- $AS_2$  可经  $IP_{3a}$  到前缀 X 的路由 = [前缀, AS-PATH, NEXT-HOP] = [X,  $AS_3$ ,  $IP_{3a}$ ]

# BGP 路由信息



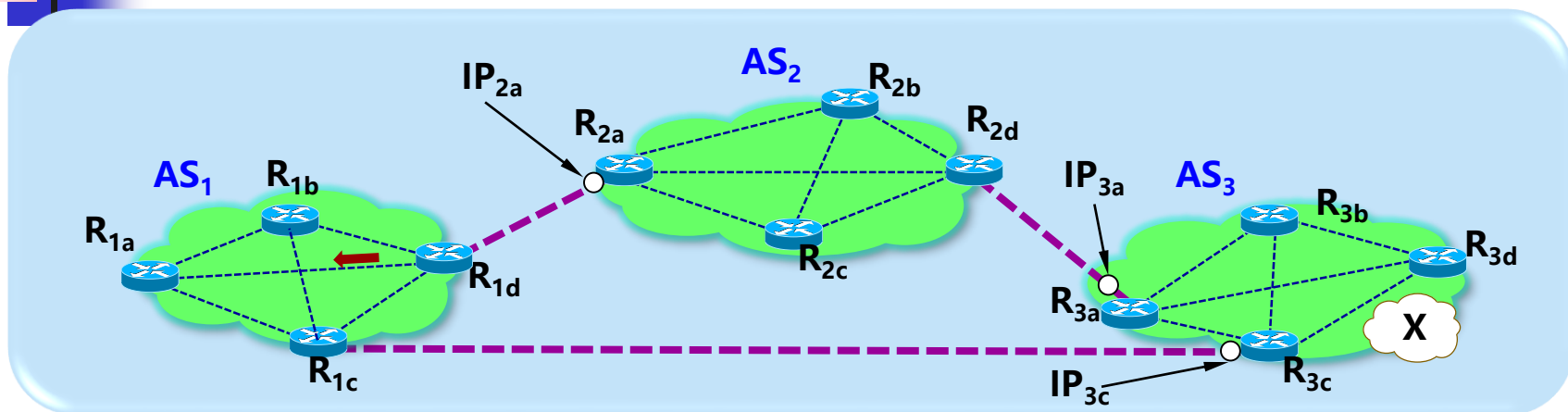
- 路由 1: AS<sub>1</sub> 可经 IP<sub>2a</sub> 到前缀 X 的路由  
= [前缀, AS-PATH, NEXT-HOP] = [X,  
AS<sub>2</sub> AS<sub>3</sub>, IP<sub>2a</sub>]

# BGP 路由信息



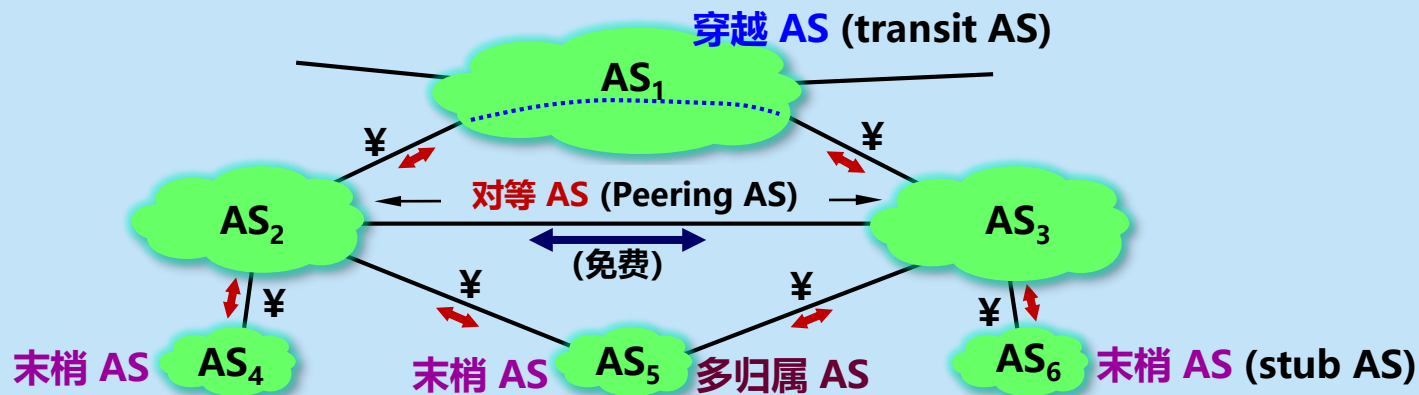
- 路由2:  $AS_1$  可经  $IP_{3c}$  到前缀 X 的路由 = [前缀, AS-PATH, NEXT-HOP] = [X,  $AS_3$ ,  $IP_{3c}$ ]

# BGP 路由信息



- 路由器  $R_{1a}$  的转发表中，沿 BGP 路由 1 到达前缀 X 的项目是：  
（匹配前缀 X，下一跳路由器  $R_{1b}$ ）或  
（匹配前缀 X，转发接口 0）。

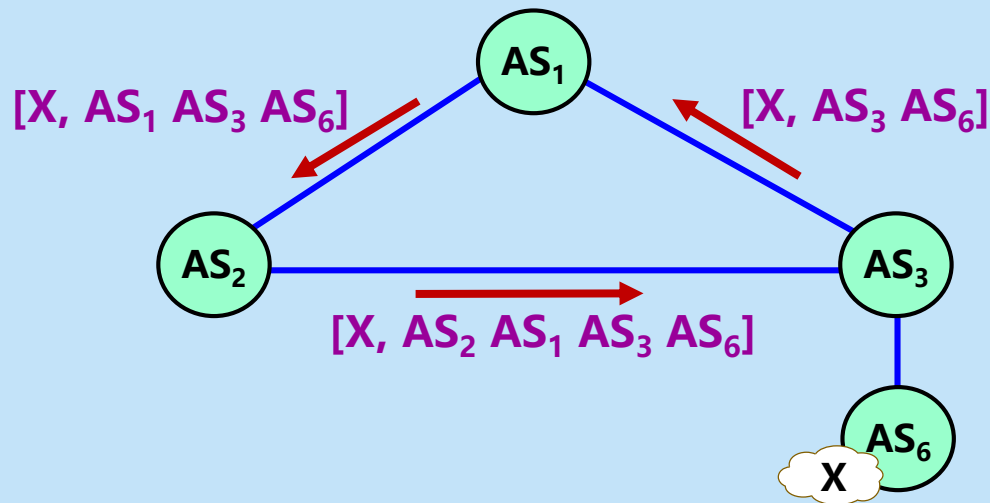
### 3. 三种不同的自治系统 AS



- 末梢 AS: 不会把来自其他 AS 的分组再转发到另一个 AS。必须向所连接的 AS 付费。
- 多归属 AS (multihomed AS): 同时连接到两个或两个以上的 AS。增加连接的可靠性。
- 穿越 AS: 为其他 AS 有偿转发分组。
- 对等 AS: 经过事先协商的两个 AS, 彼此之间的发送或接收分组都不收费。



# BGP 路由如何避免兜圈子？



AS<sub>3</sub> 检查收到的 BGP 路由的 AS-PATH 中已经有了自己，立即删除掉这条路由，从而避免兜圈子路由的出现。

- BGP采用了路径向量路由选择协议。请记住：在属性 AS-PATH 中，不允许出现相同的 AS 号。



## 4. BGP 的路由选择

1

- 本地偏好 (local preference) 值最高的路由

2

- AS 跳数最小的路由

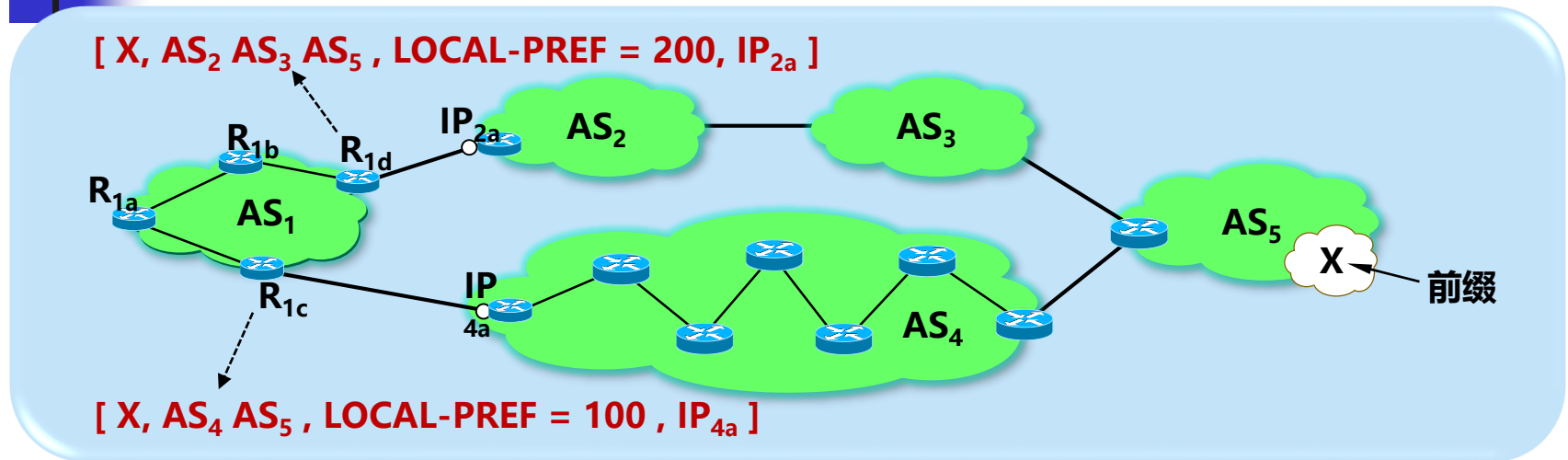
3

- 使用热土豆路由选择算法

4

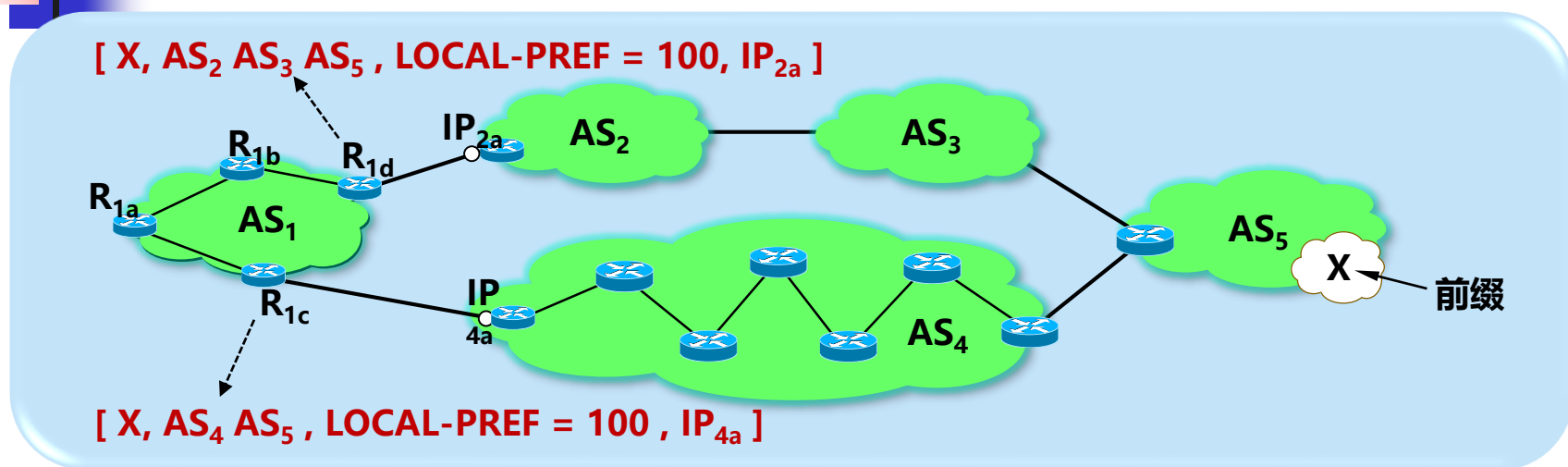
- 路由器BGP标识符数值最小的路由

# 本地偏好 (local preference) 值最高



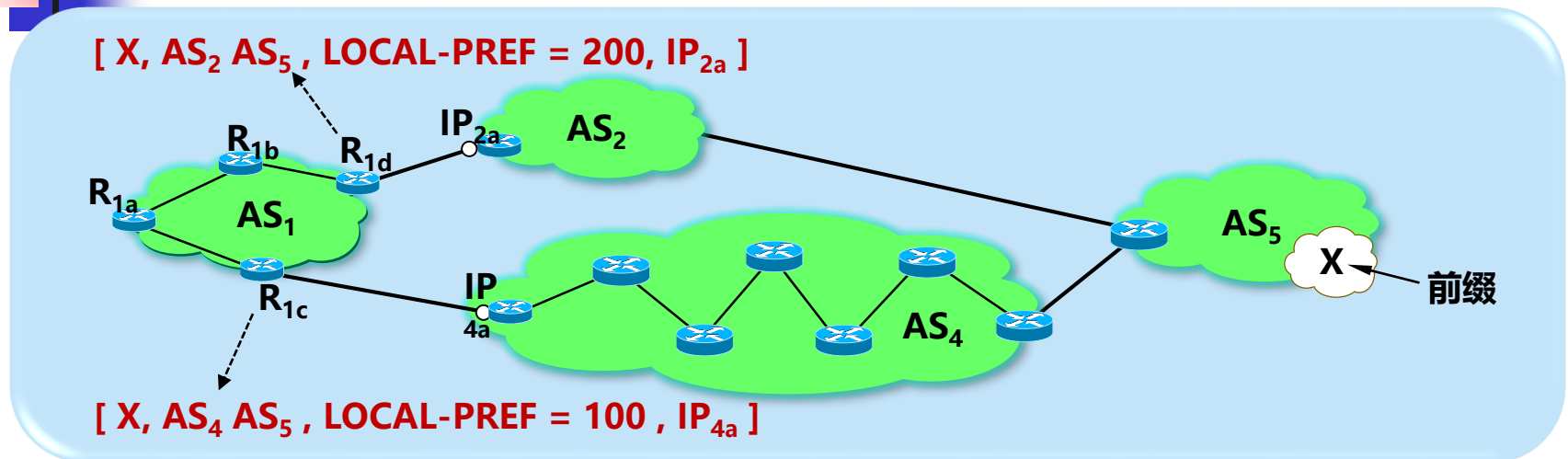
- AS<sub>1</sub> 选择通过路由器 R<sub>1d</sub> 到达 X 的 BGP 路由。

# AS 跳数最小



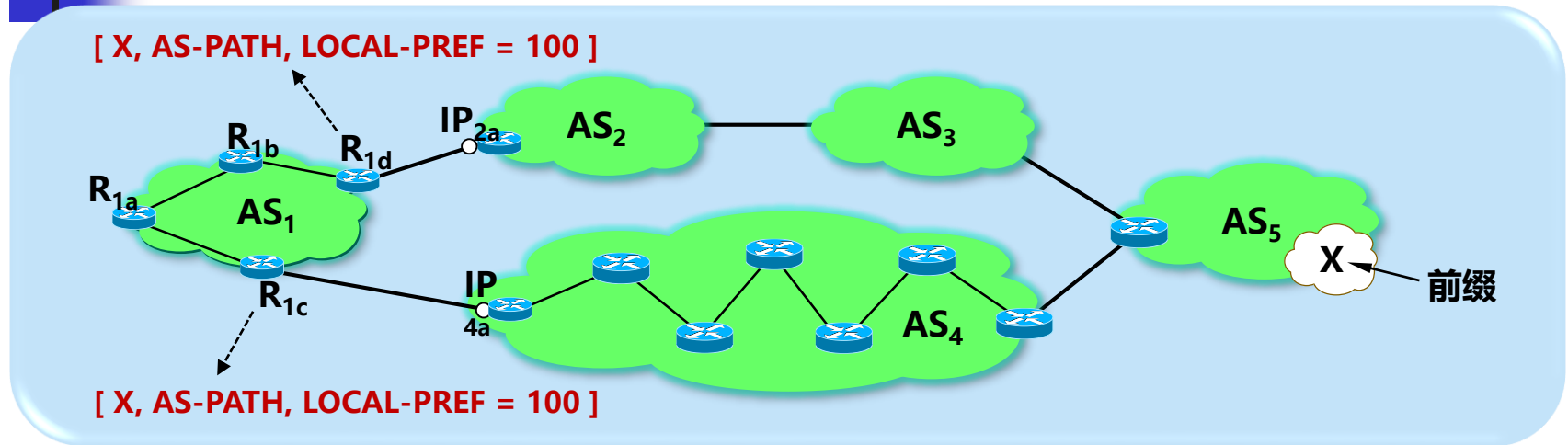
- AS<sub>1</sub> 选择通过路由器 R<sub>1c</sub> 到达 X 的 BGP 路由。
- 但事实上，分组在 AS<sub>4</sub> 中反而要经过更多次数的转发。
- 说明协议 BGP 不存在真正的最佳路由选择。。

# 热土豆路由选择算法



- $R_{1a}$  选择  $R_{1c}$  作为离开  $AS_1$  的最佳选择，其 BGP 转发表中对应的项目应当是：（匹配前缀 X，下一跳路由器  $R_{1c}$ ）。
- $R_{1b}$  选择  $R_{1d}$  作为离开  $AS_1$  的最佳选择，其 BGP 转发表中对应的项目应当是：（匹配前缀 X，下一跳路由器  $R_{1d}$ ）。

# 路由器BGP标识符数值最小的路由



- 具有多个接口的路由器有多个 IP 地址，BGP ID 就使用该路由器的 IP 地址中数值最大的一个。



# 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。
  - (4) 通知(**NOTIFICATION**)报文，用来发送检测到的差错。
- 在 RFC 2918 中增加了 ROUTE-REFRESH 报文，用来请求对等端重新通告。



# BGP-4 共使用四种报文

---

- 邻站进行商谈时就必须发送OPEN报文。如果邻站接受这种邻站关系，就用KEEPALIVE报文响应。这样，两个BGP发言人的邻站关系就建立了。
- 两个BGP发言人彼此要周期性地交换报文KEEPALIVE(一般每隔30秒)。
- 更新报文是BGP协议的核心内容。BGP发言人可以用更新报文撤消它以前曾经通知过的路由，也可以宣布增加新的路由。



# BGP 报文具有通用的首部

字节

16

2

1

标

记

长 度

类 型

BGP 报文通用首部

BGP 报文主体部分

TCP首部

BGP 报文

IP 首部

TCP 报文

