第 五 章　路由协议

5.3　IP数据报的格式 (1-46):已讲授

5.3.1　IP 数据报的格式

5.3.2. IP 数据报首部的可变部分

5.3.3　IP 层转发分组的流程

5.4　网际控制报文协议 ICMP

5.4.1　ICMP 报文的种类

5.4.2　ICMP 的应用举例

5.5　因特网的路由选择协议

5.5.1　有关路由选择协议的几个基本概念

5.5.2　内部网关协议 RIP

5.5.3　内部网关协议 OSPF

5.5.4　外部网关协议 BGP

5.5.5　路由器的构成

第 5 讲　路由协议（续）

5.6 IP 多播

- 5.6.1　IP 多播的基本概念
- 5.6.2　在局域网上进行硬件多播
- 5.6.2　因特网组管理协议 IGMP 和多播路由选择协议

5.7 虚拟专用网 VPN 和网络地址转换 NAT

- 5.7.1　虚拟专用网 VPN
- 5.7.2　网络地址转换 NAT

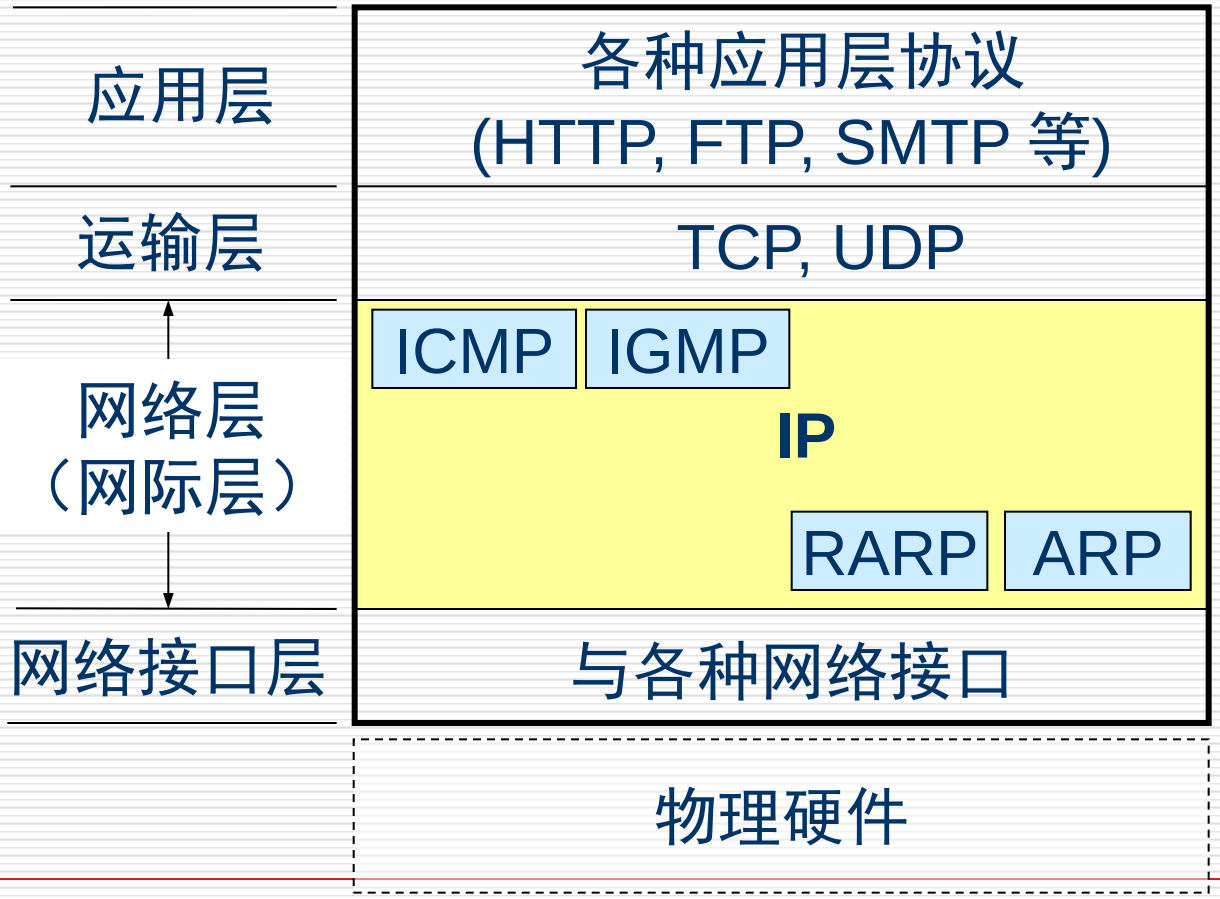
5.3.1 IP 协议与IP数据报的格式

- 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。

一. 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



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，如：
  - 不同的寻址方案
  - 不同的最大分组长度
  - 不同的网络接入机制
  - 不同的超时控制
  - 不同的差错恢复方法
  - 不同的状态报告方法
  - 不同的路由选择技术
  - 不同的用户接入控制
  - 不同的服务（面向连接服务和无连接服务）
  - 不同的管理与控制方式

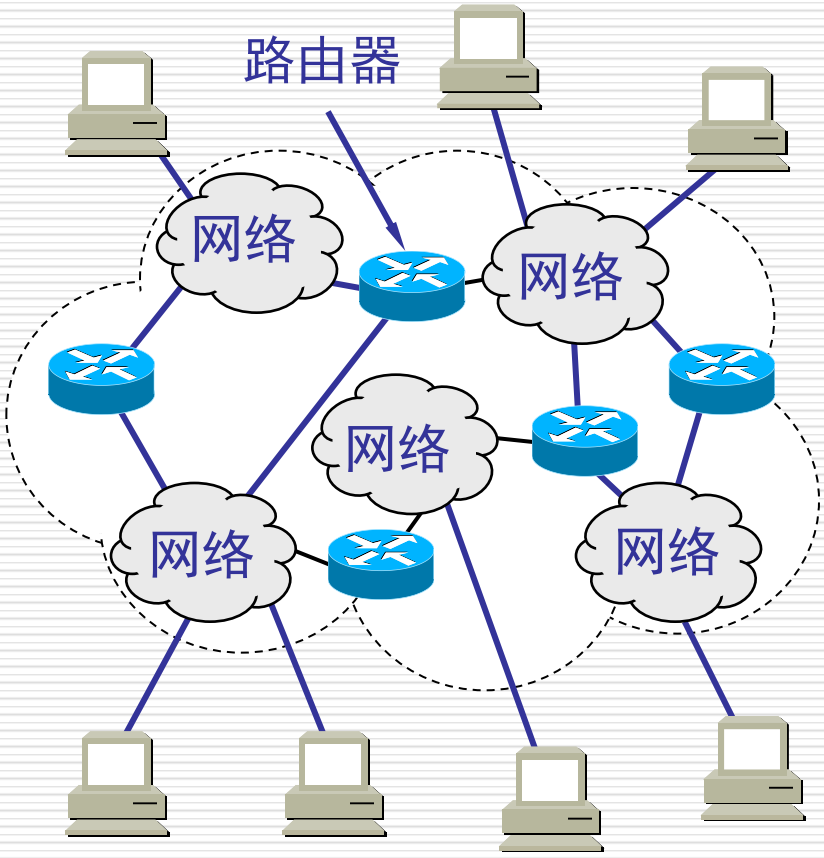
网络互相连接起来要使用一些中间设备

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

网络互连使用路由器

- 当中继系统是转发器或网桥时，一般并不称之为网络互连，因为这仅仅是把一个网络扩大了，而这仍然是一个网络。
- 网关由于比较复杂，目前使用得较少。
- 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。

互连网络与虚拟互连网络



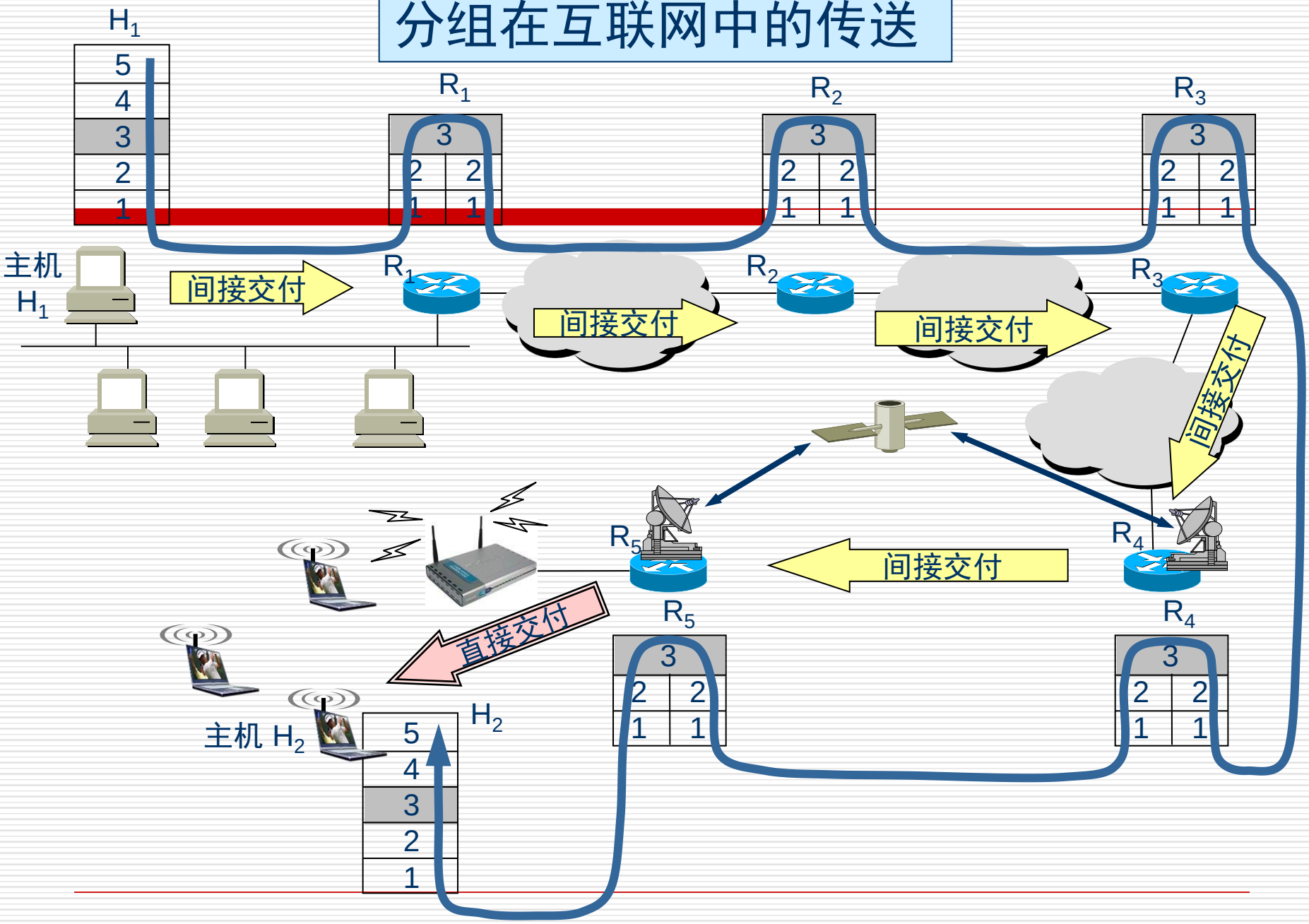
(a) 互连网络

(b) 虚拟互连网络

虚拟互连网络的意义

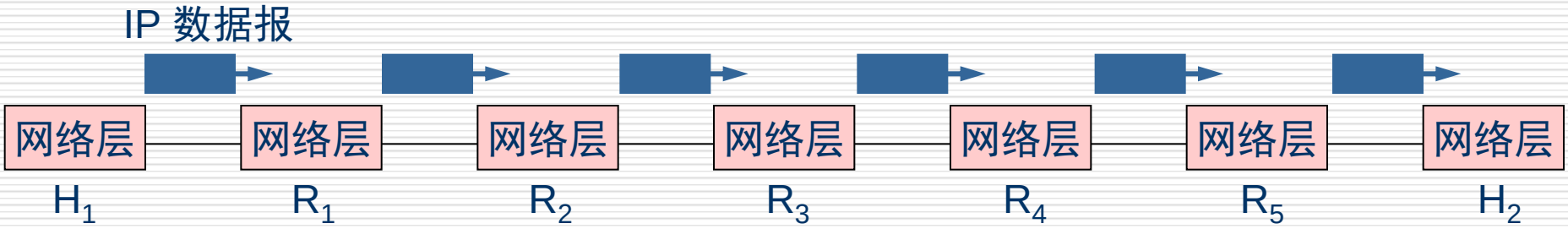
- 所谓虚拟互连网络也就是逻辑互连网络，它的意思就是互连起来的各种物理网络的异构性本来是客观存在的，但是我们利用 IP 协议就可以使这些性能各异的网络从用户看起来好像是一个统一的网络。
- 使用 IP 协议的虚拟互连网络可简称为 IP 网。
- 使用虚拟互连网络的好处是：当互联网上的主机进行通信时，就好像在一个网络上通信一样，而看不见互连的各具体的网络异构细节。

分组在互联网中的传送

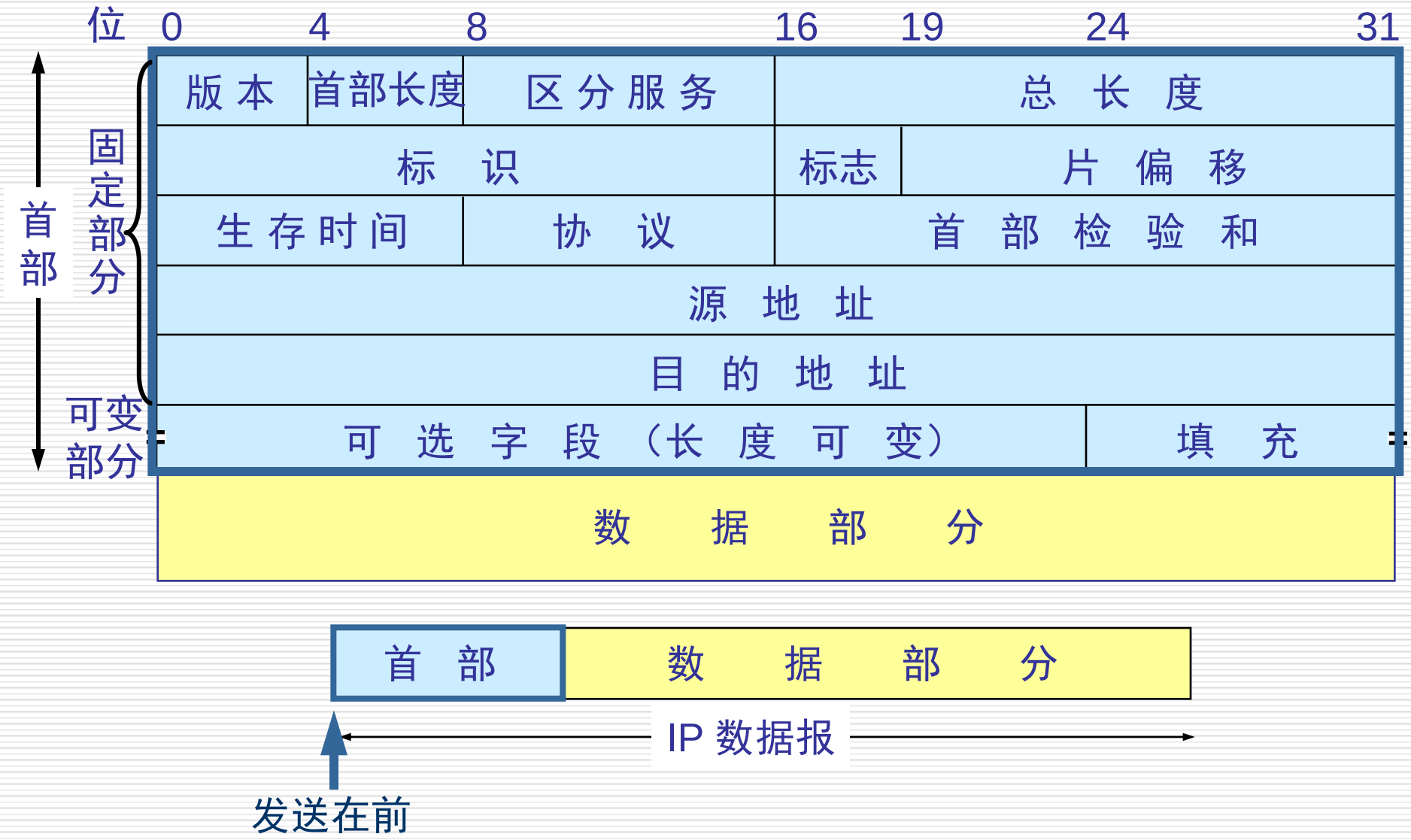


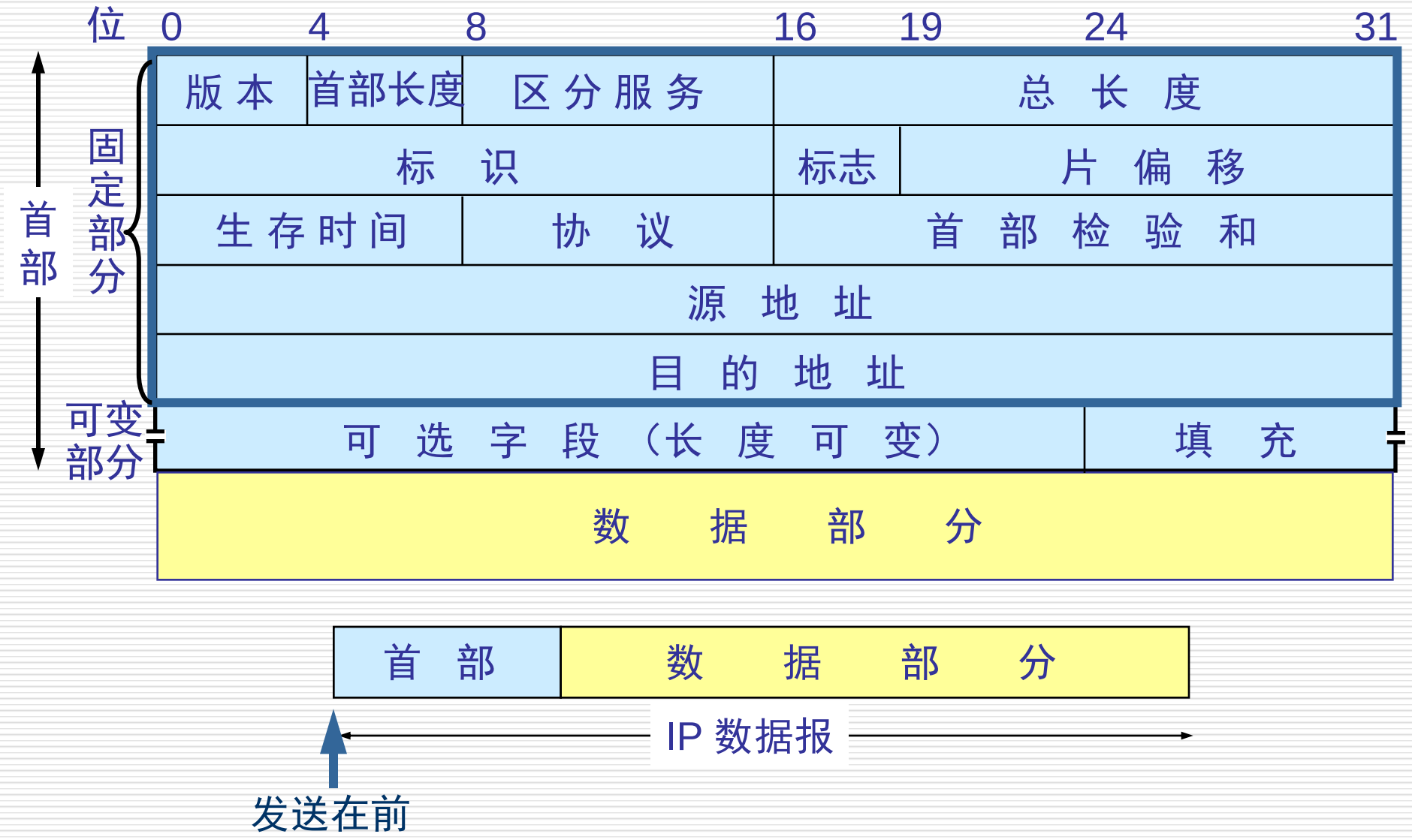
从网络层看 IP 数据报的传送

- 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。

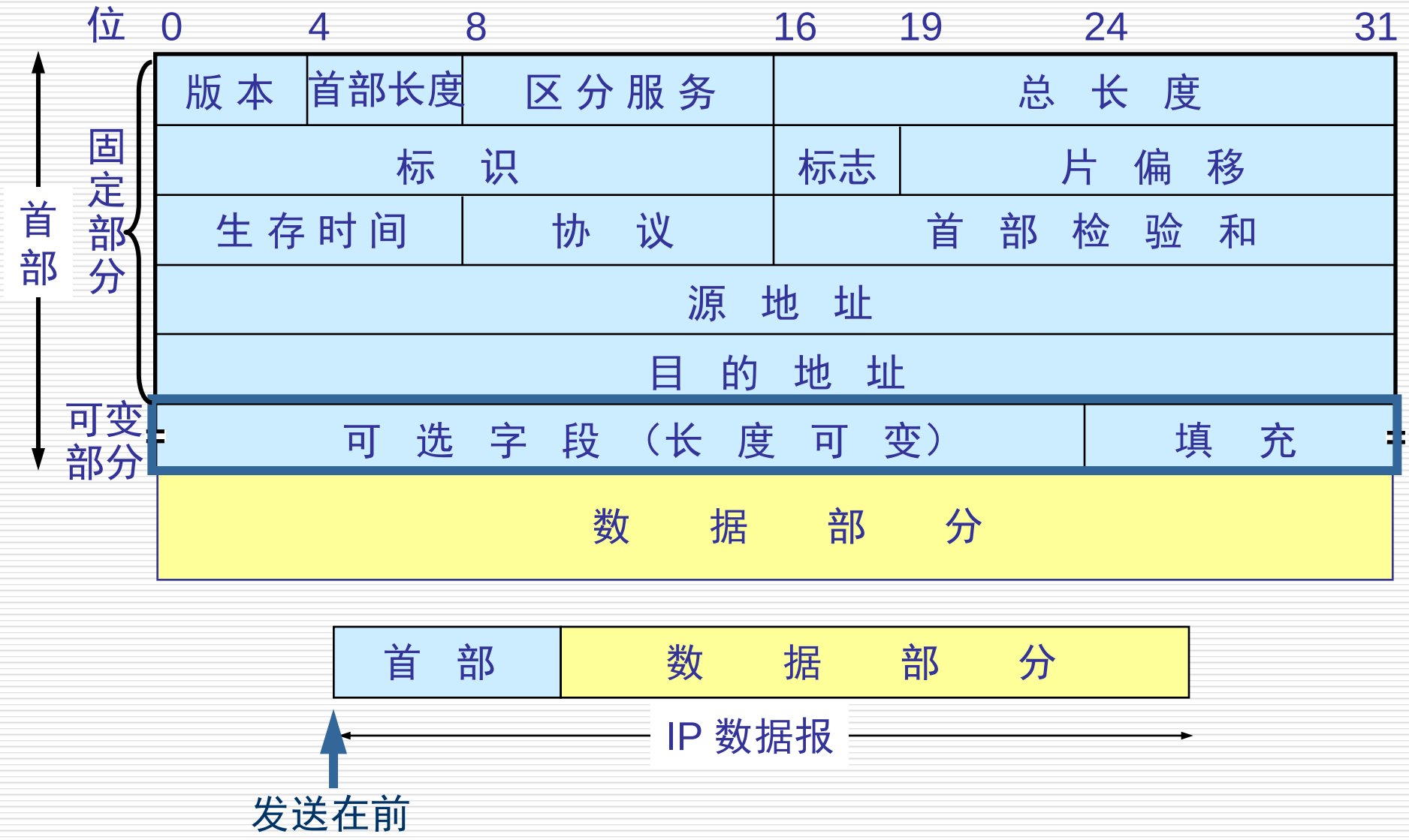


二. IP数据报的格式


位
0
4
8
16
19
24
31
首部
固定部分
可变部分
版本
首部长度
区分服务
总长度
标识
标志
片偏移
生存时间
协议
首部检验和
源地址
目的地址
可选字段（长度可变）
填充
数据部分
首部
数据部分
IP 数据报
发送在前

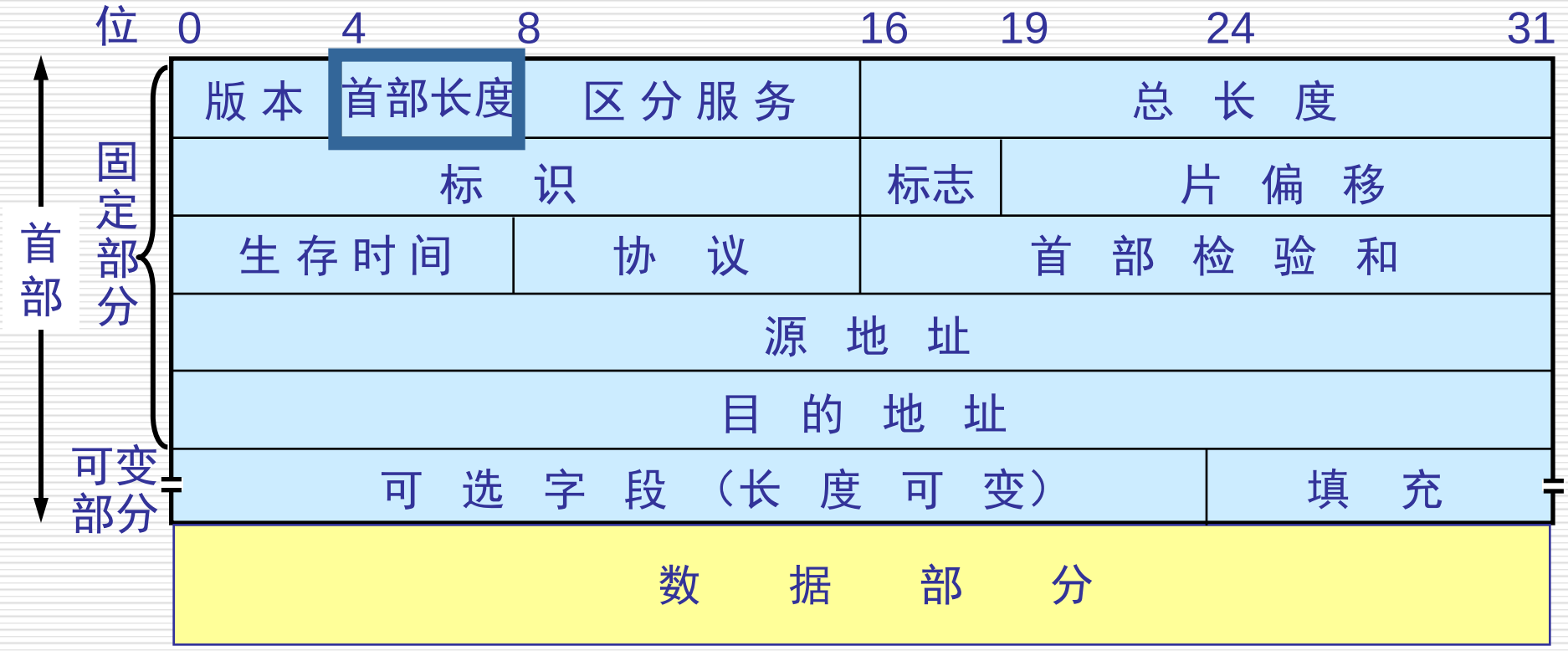

位
0
4
8
16
19
24
31
首部
固定部分
版本
首部长度
区分服务
总长度
标识
标志
片偏移
生存时间
协议
首部检验和
源地址
目的地址
可变部分
可选字段（长度可变）
填充
数据部分

首部
数据部分
IP 数据报
发送在前


位
0
4
8
16
19
24
31
首部
固定部分
版本
首部长度
区分服务
总长度
标识
标志
片偏移
生存时间
协议
首部检验和
源地址
目的地址
可变部分
可选字段（长度可变）
填充
数据部分
首部
数据部分
IP 数据报
发送在前

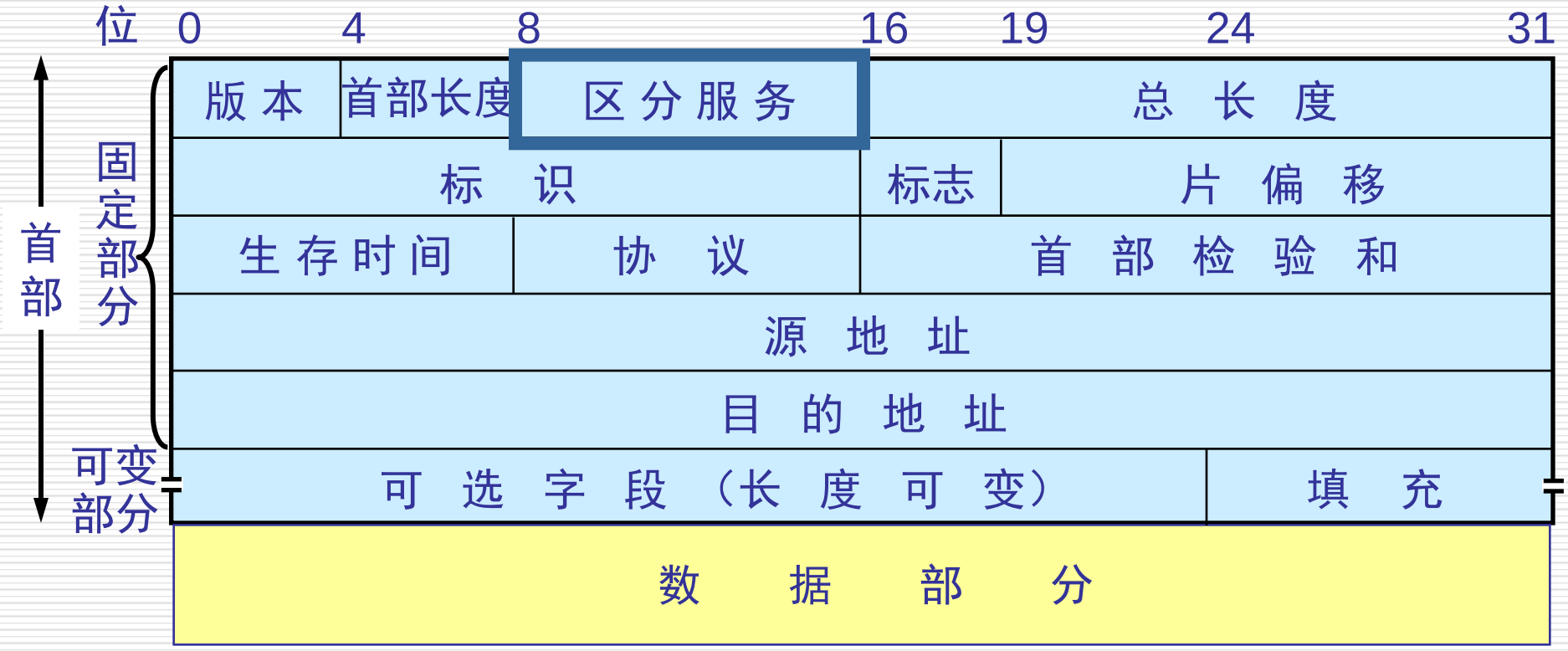
1. IP 数据报首部的固定部分中的各字段

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。



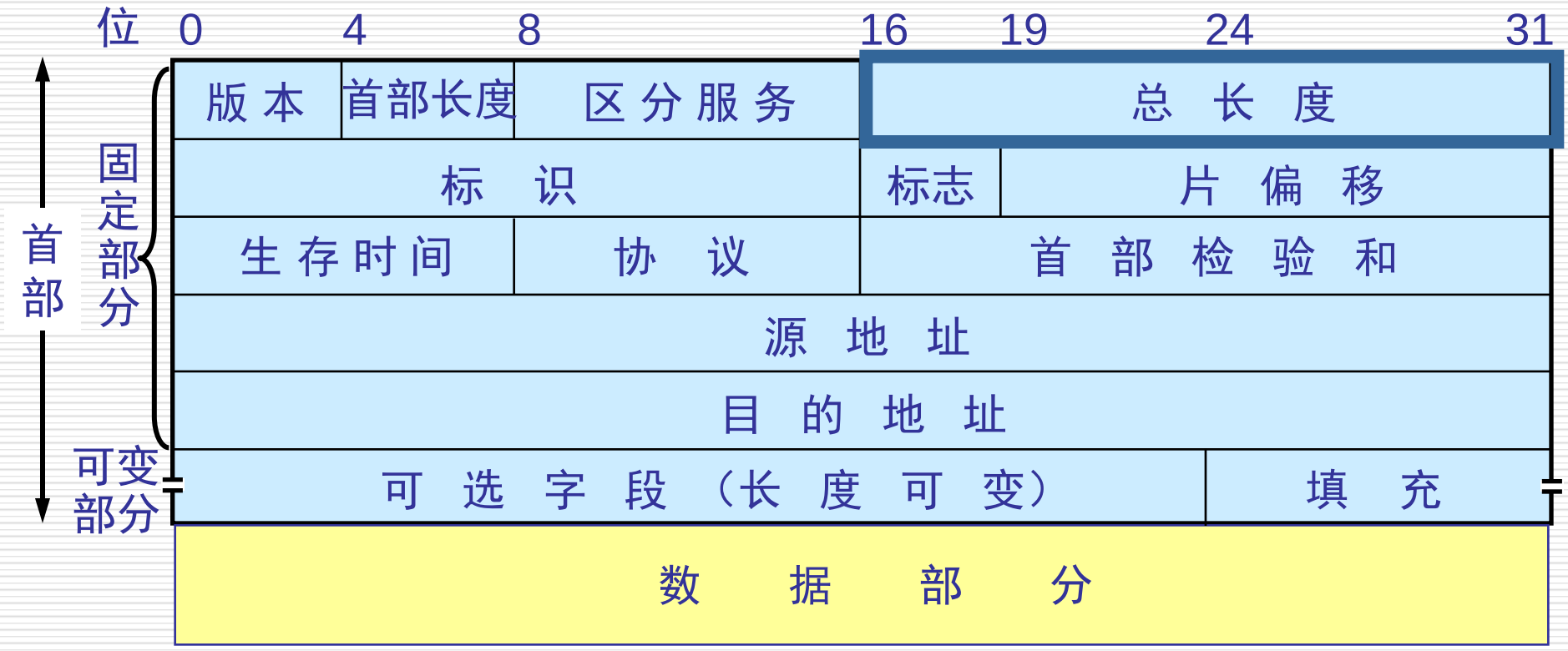
区分服务——占 8 位，用来获得更好的服务

在旧标准中叫做服务类型，但实际上一直未被使用过。

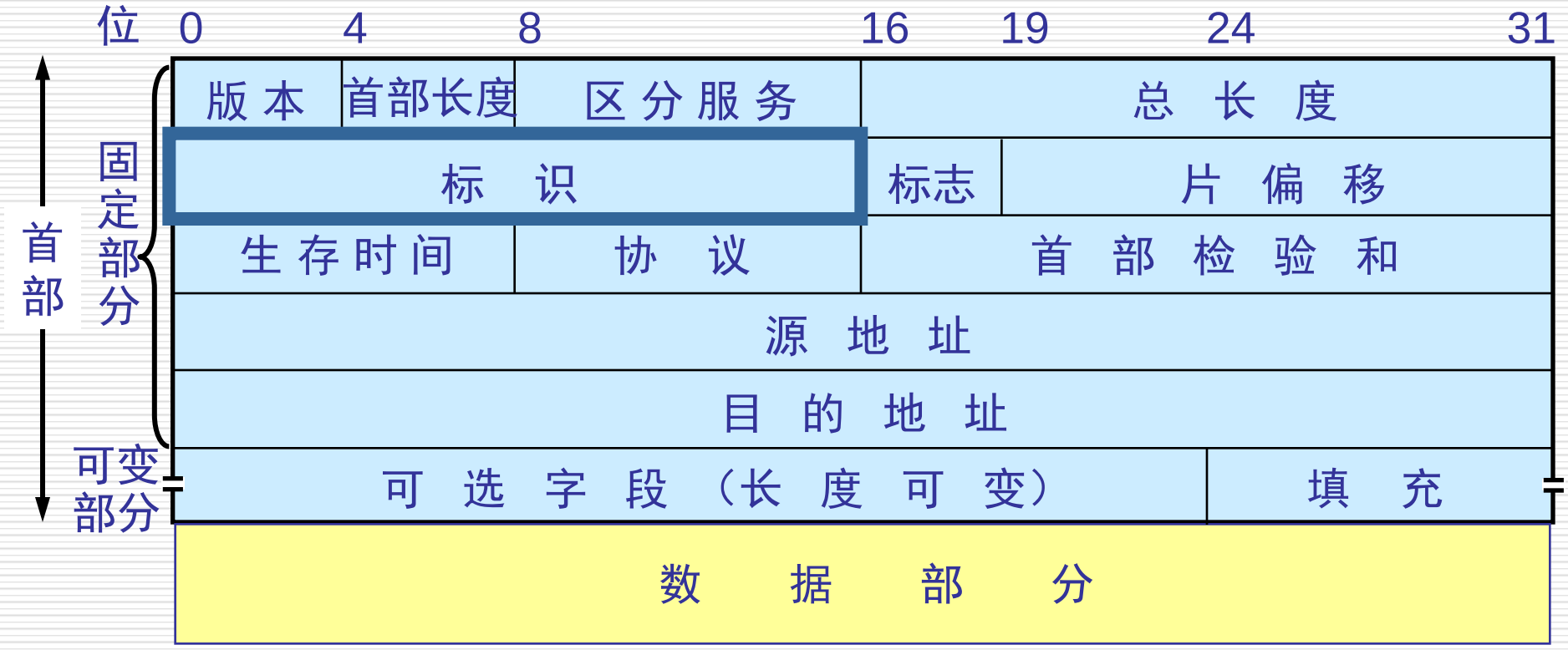
1998 年这个字段改名为区分服务。

只有在使用区分服务（DiffServ）时，这个字段才起作用。

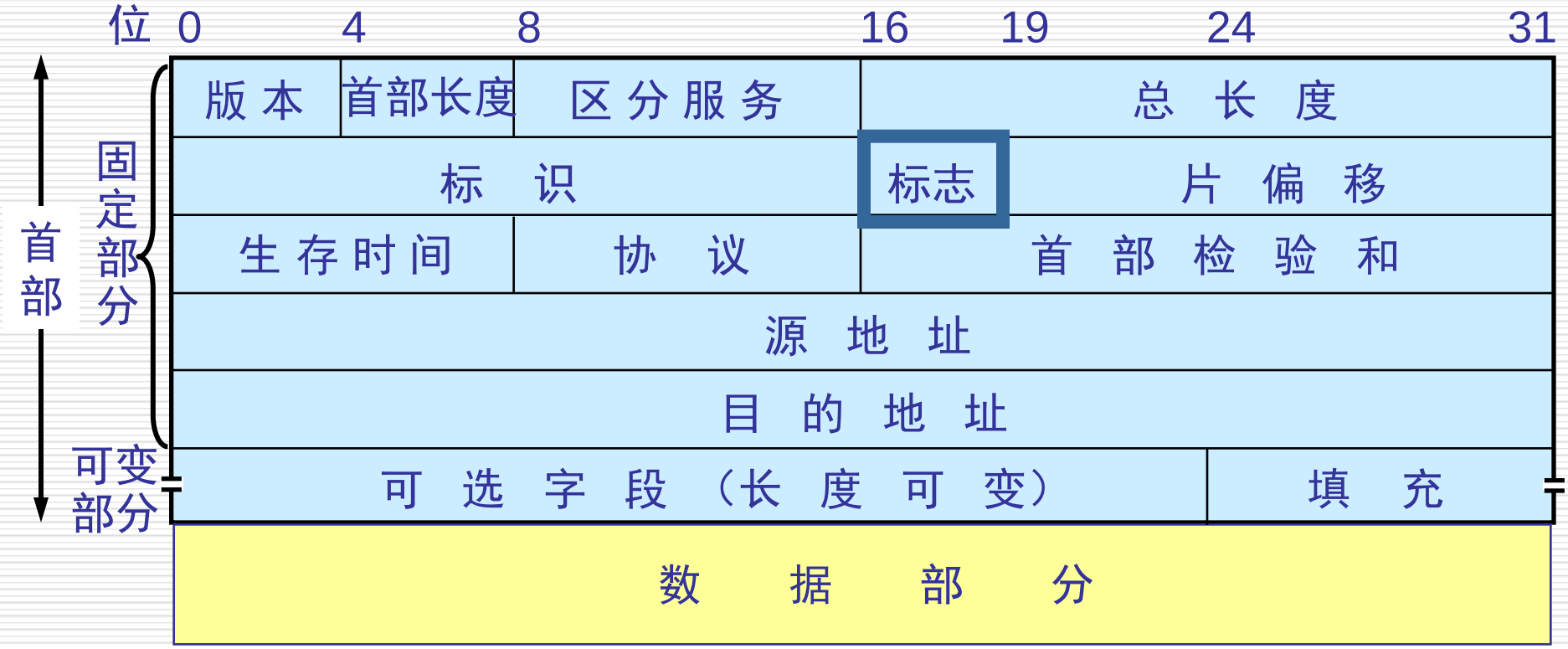
在一般的情况下都不使用这个字段



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。



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



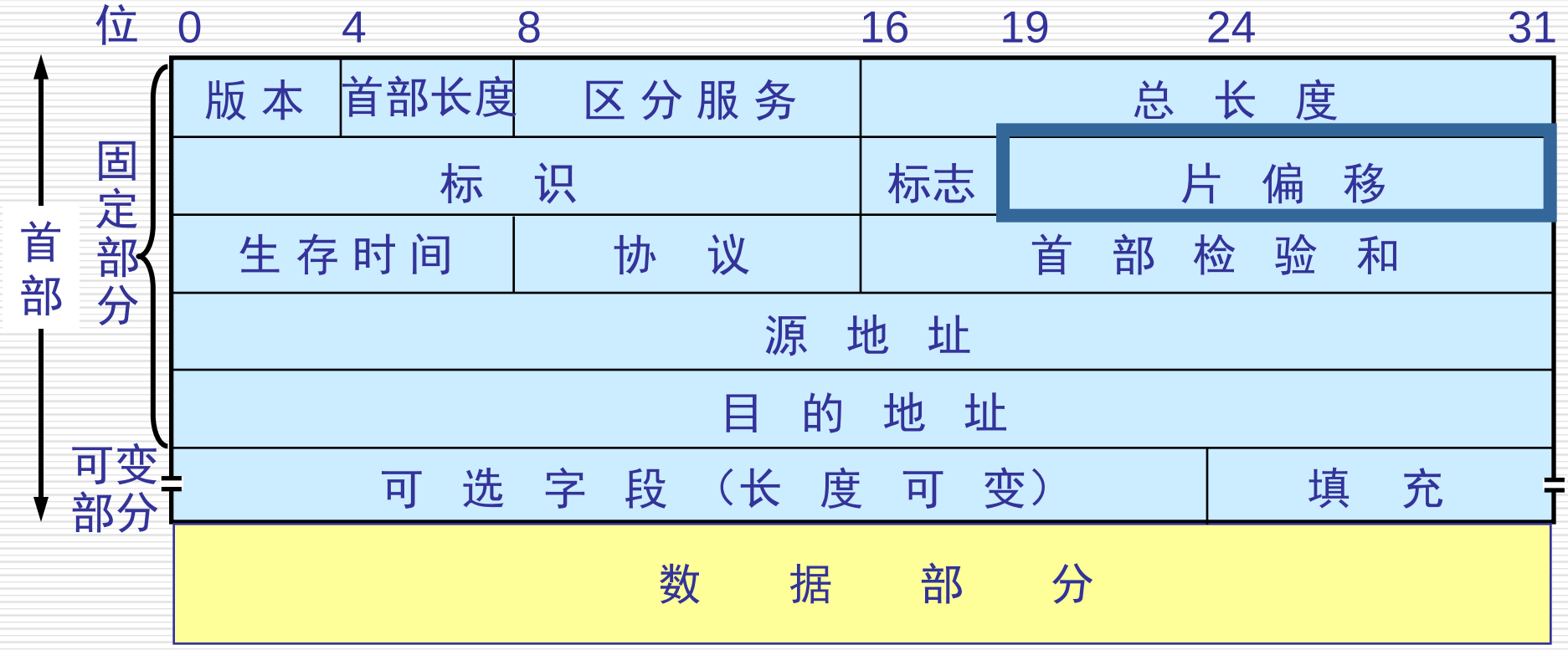
标志(flag)　占 3 位，目前只有前两位有意义。

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) 。

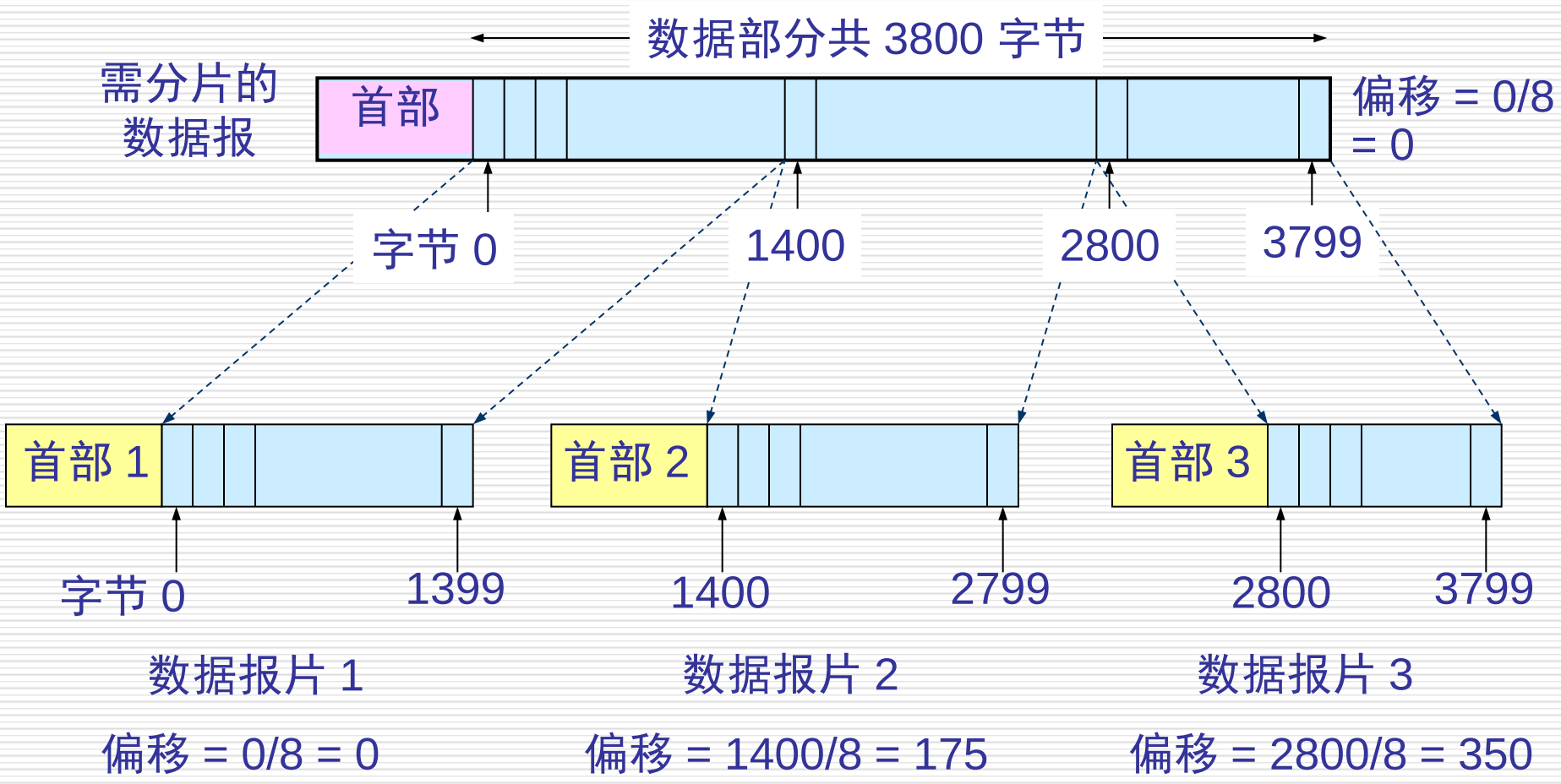
只有当 DF = 0 时才允许分片。

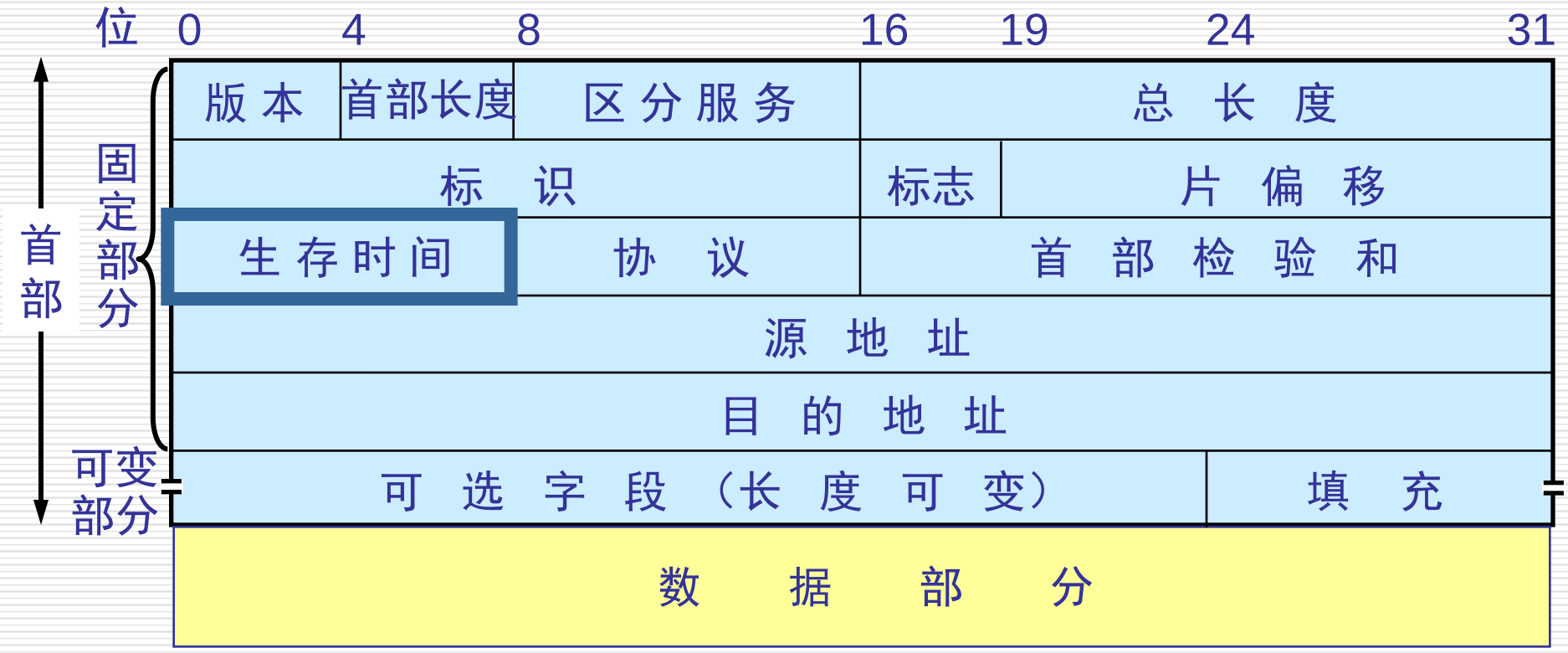


片偏移(12 位)指出：较长的分组在分片后某片在原分组中的相对位置。

片偏移以 8 个字节为偏移单位。

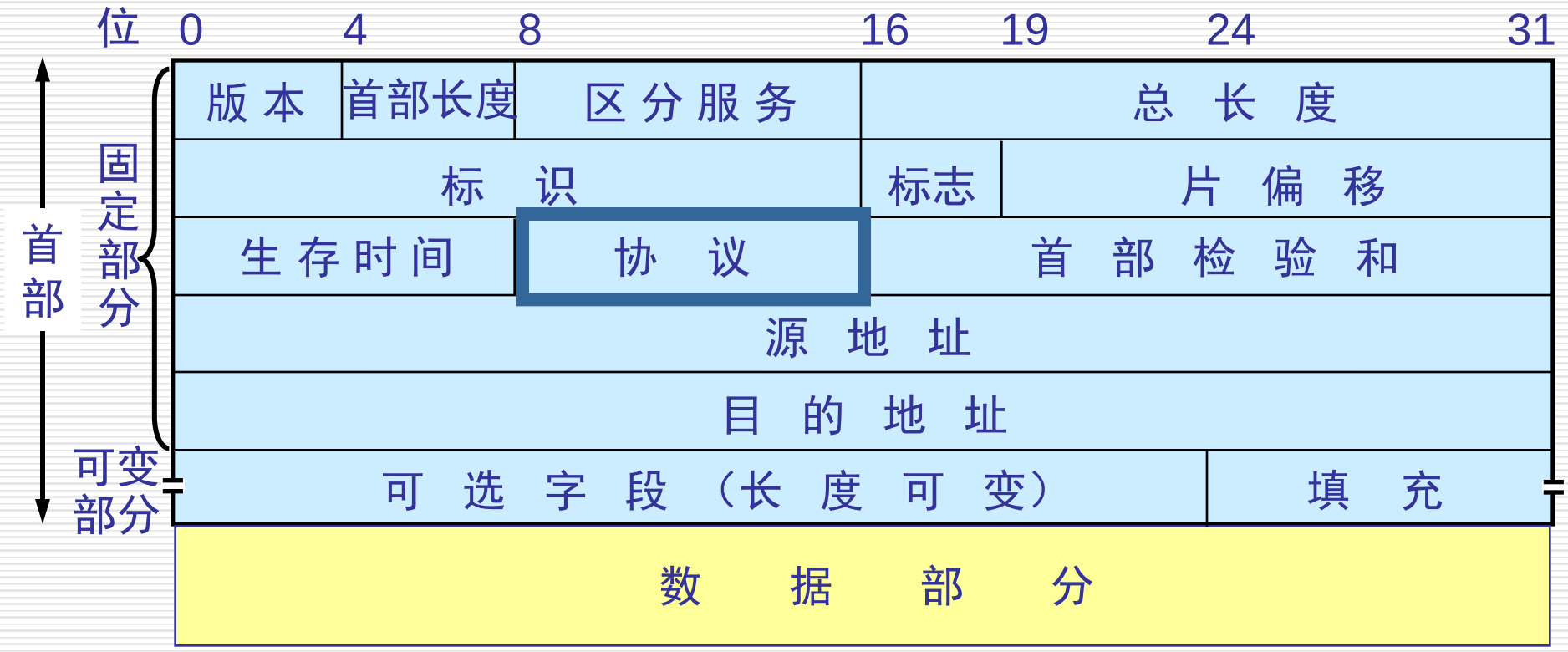
【例5-1】 IP 数据报分片


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

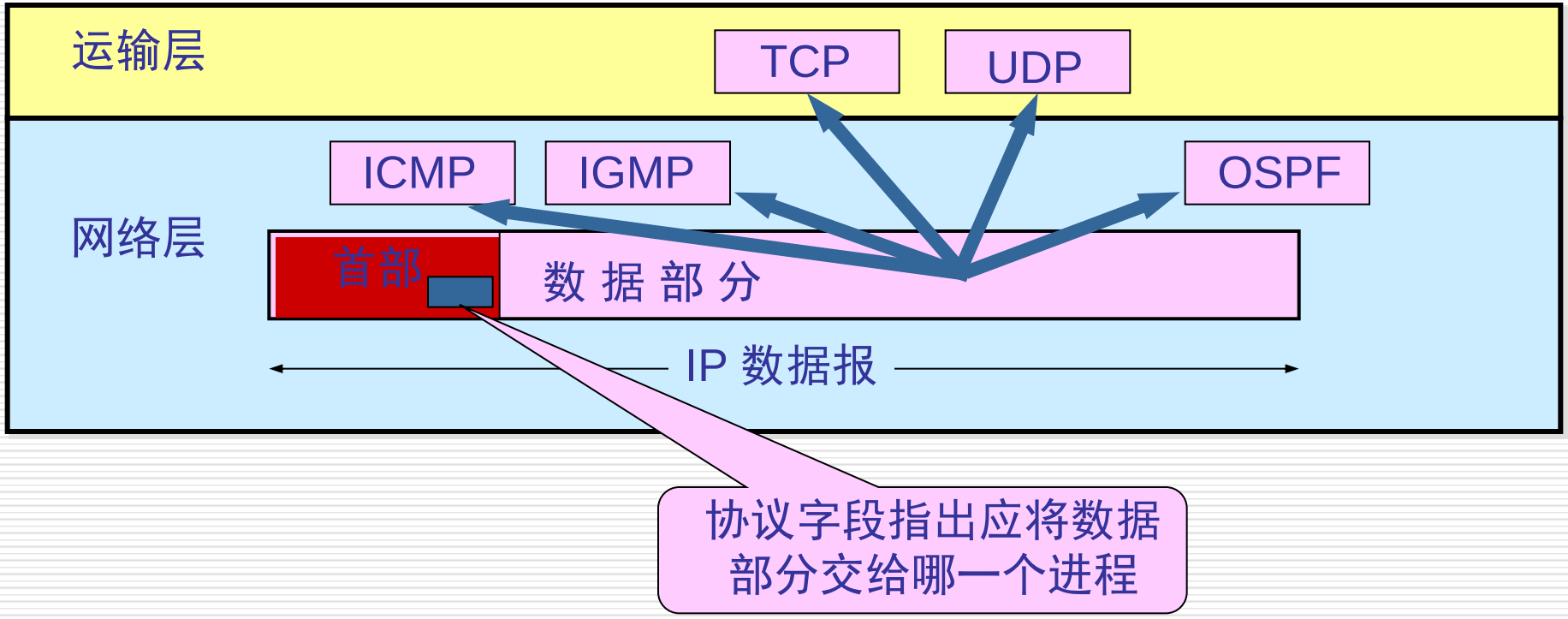


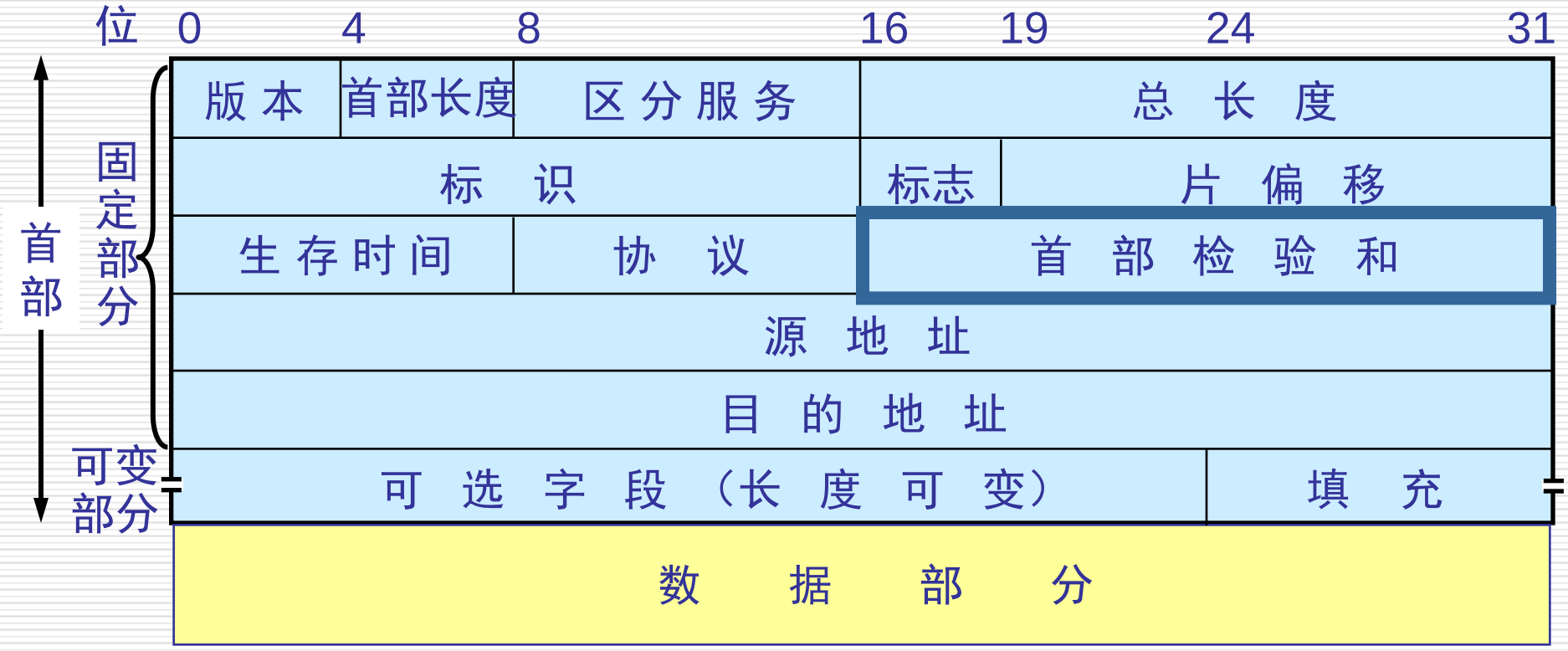
生存时间(8 位)记为 TTL (Time To Live)

数据报在网络中可通过的路由器数的最大值。



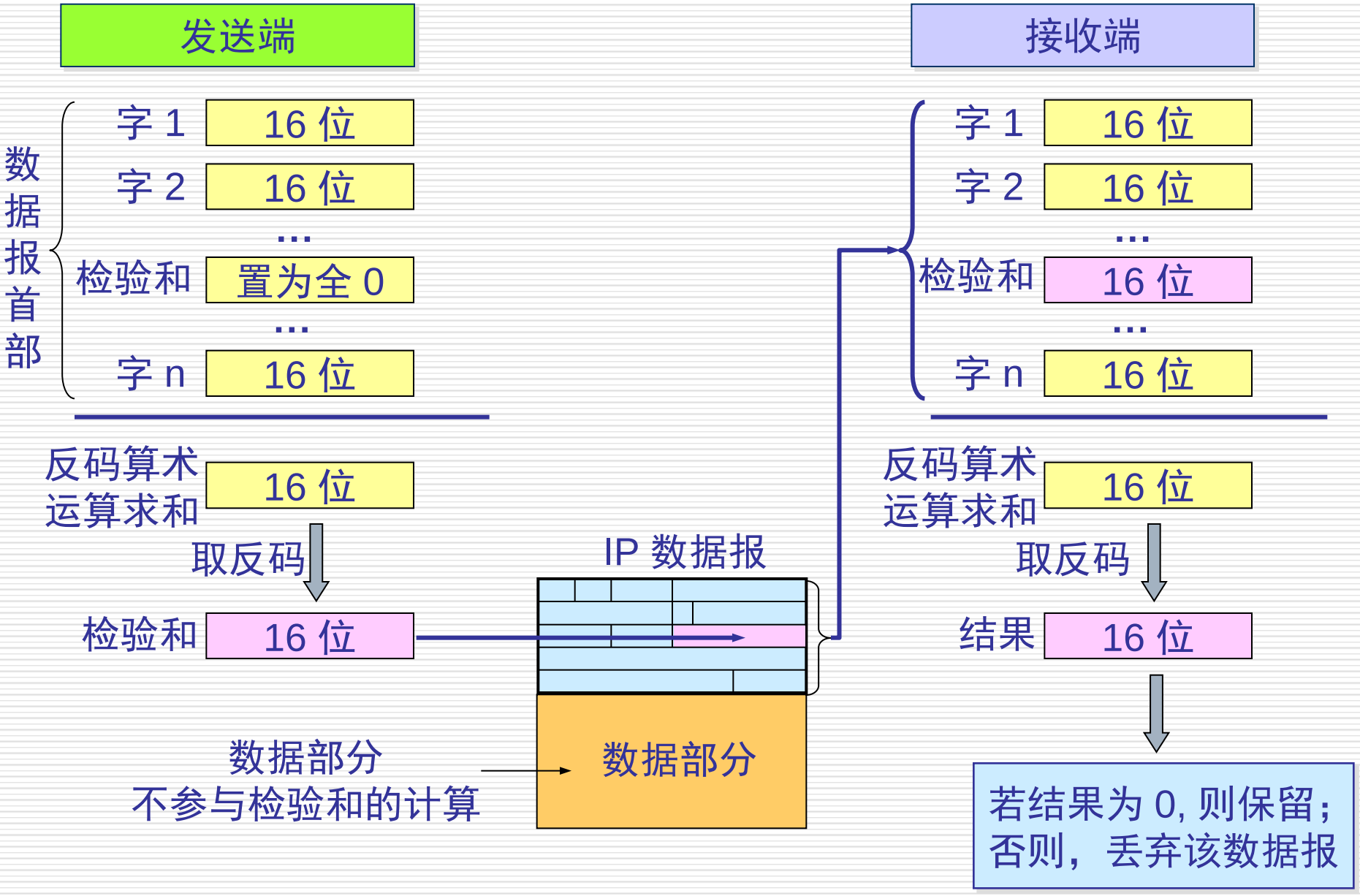
协议(8 位)字段指出此数据报携带的数据使用何种协议以便目的主机的 IP 层将数据部分上交给哪个处理过程


运输层
TCP
UDP
网络层
ICMP
IGMP
OSPF
首部
数 据 部 分
IP 数据报
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
字 1
16 位
字 2
16 位
...
检验和
置为全 0
...
字 n
16 位
反码算术运算求和
16 位
取反码
检验和
16 位
IP 数据报
数据部分
数据部分
不参与检验和的计算
字 1
16 位
字 2
16 位
...
检验和
16 位
...
字 n
16 位
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

5.3.2. IP 数据报首部的可变部分

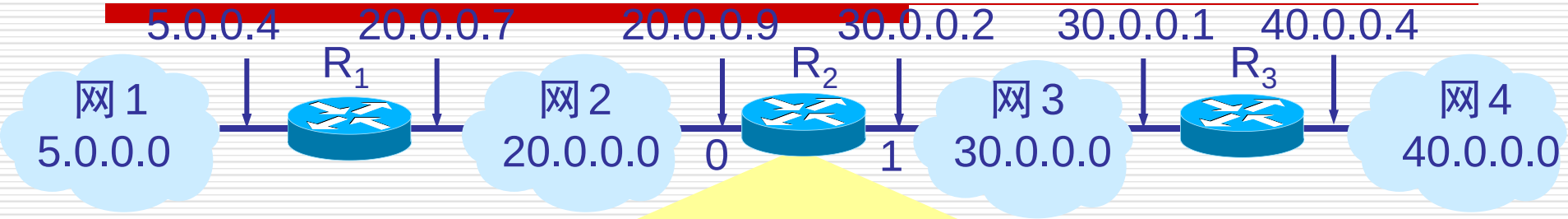
- IP 首部的可变部分就是一个选项字段，用来支持排错、测量以及安全等措施，内容很丰富。
- 选项字段的长度可变，从 1 个字节到 40 个字节不等，取决于所选择的项目。
- 增加首部的可变部分是为了增加 IP 数据报的功能，但这同时也使得 IP 数据报的首部长度成为可变的。这就增加了每一个路由器处理数据报的开销。
- 实际上这些选项很少被使用。

5.3.3 IP 层转发分组的流程

路由器和结点交换机有些区别：

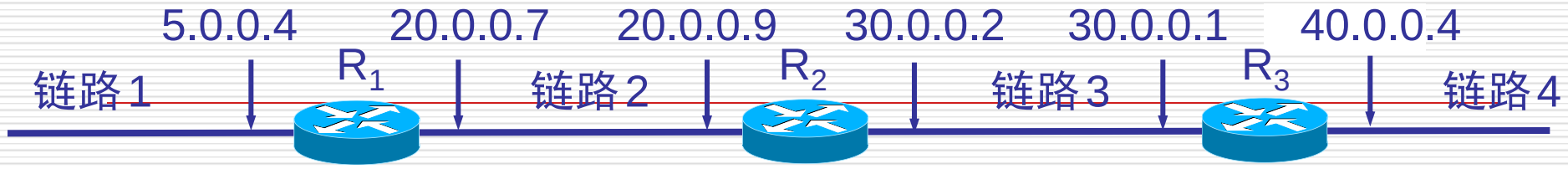
- 路由器是用来连接不同的网络，而结点交换机只是在一个特定的网络中工作。
- 路由器是专门用来转发分组的，而结点交换机还可接上许多个主机。
- 路由器使用统一的 IP 协议，而结点交换机使用所在广域网的特定协议。
- 路由器根据目的网络地址找出下一个路由器，而结点交换机则根据目的站所接入的交换机号找出下一跳（即下一个结点交换机）。

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_2 的路由表

目的主机所在的网络	下一跳路由器的地址
20.0.0.0	直接交付，接口 0
30.0.0.0	直接交付，接口 1
5.0.0.0	20.0.0.7
40.0.0.0	30.0.0.1



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。

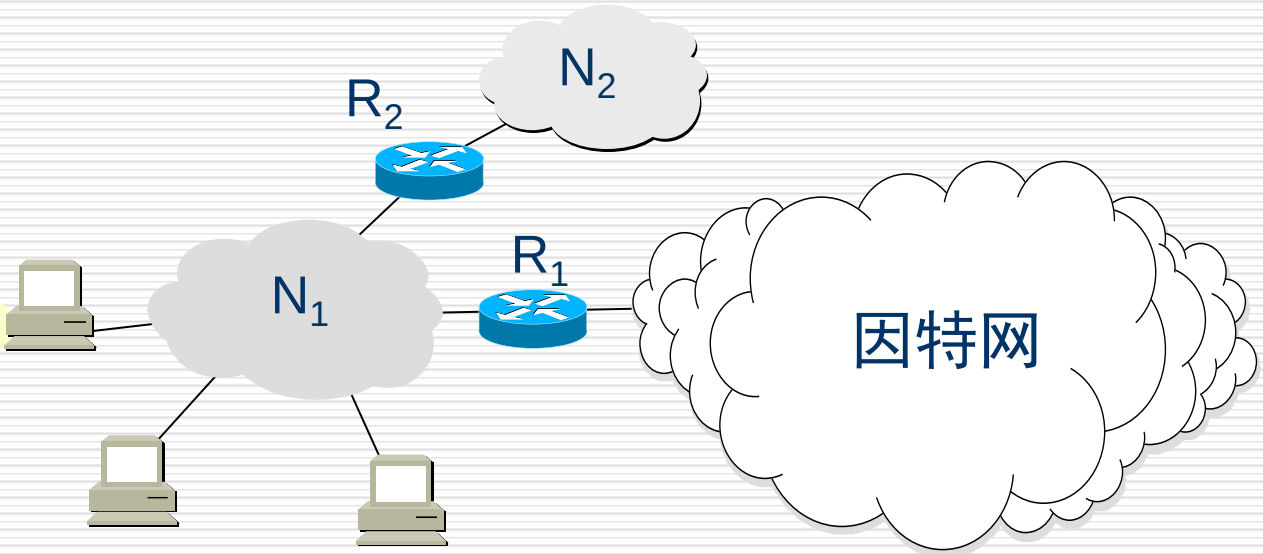
默认路由(default route)

- 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	R_1



分组转发算法

(1) 从数据报的首部提取目的站的 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，则直接将数据报交付给目的站 *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，则将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，则将数据报传送给路由表指明的下一跳路由器；否则，执行(5)。

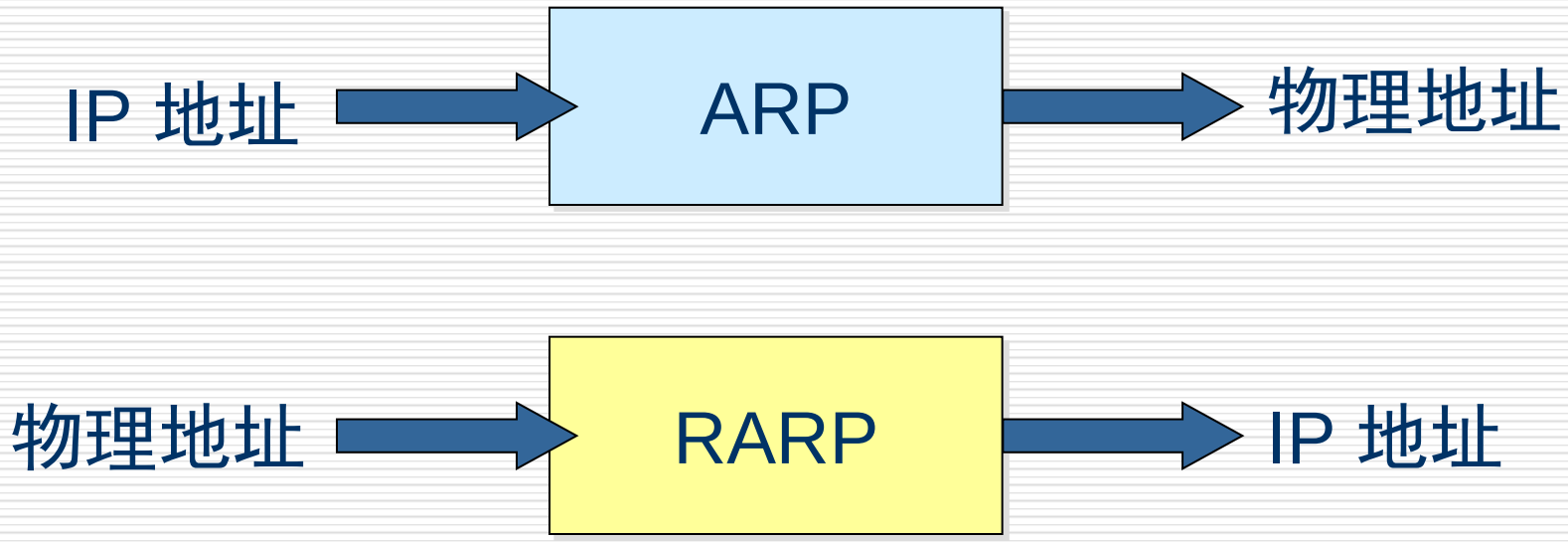
(5) 若路由表中有一个默认路由，则将数据报传送给路由表中所指明的默认路由器；否则，执行(6)。

(6) 报告转发分组出错。

细节

- 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。

5.3.4 地址解析协议 ARP /逆地址解析协议 RARP

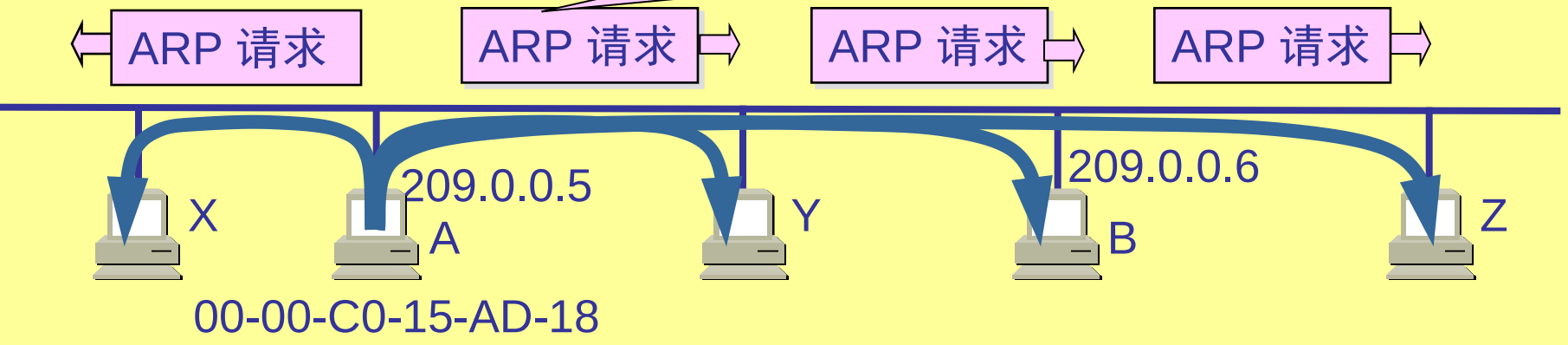


地址解析协议 ARP

- 不管网络层使用的是什么协议，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。

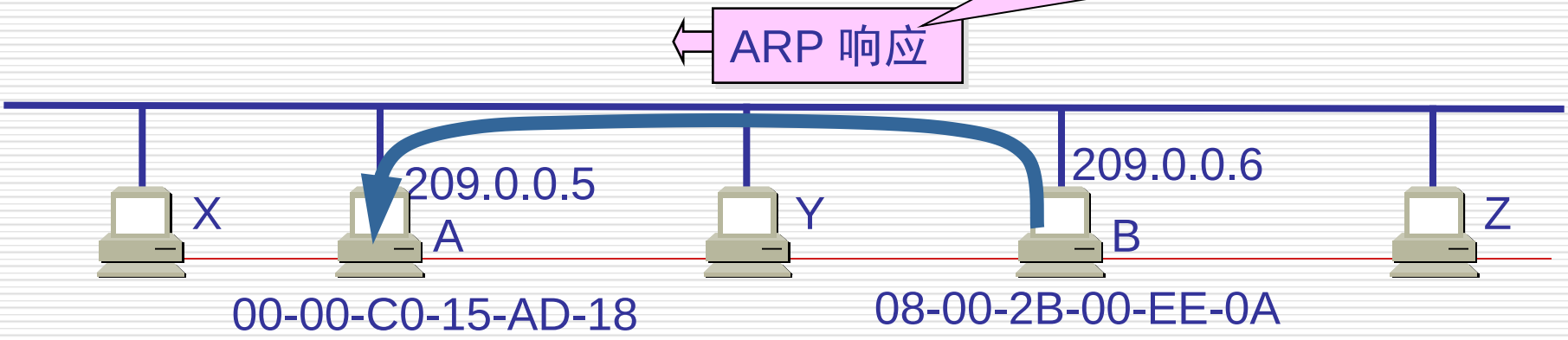
主机 A 广播发送
ARP 请求分组

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。

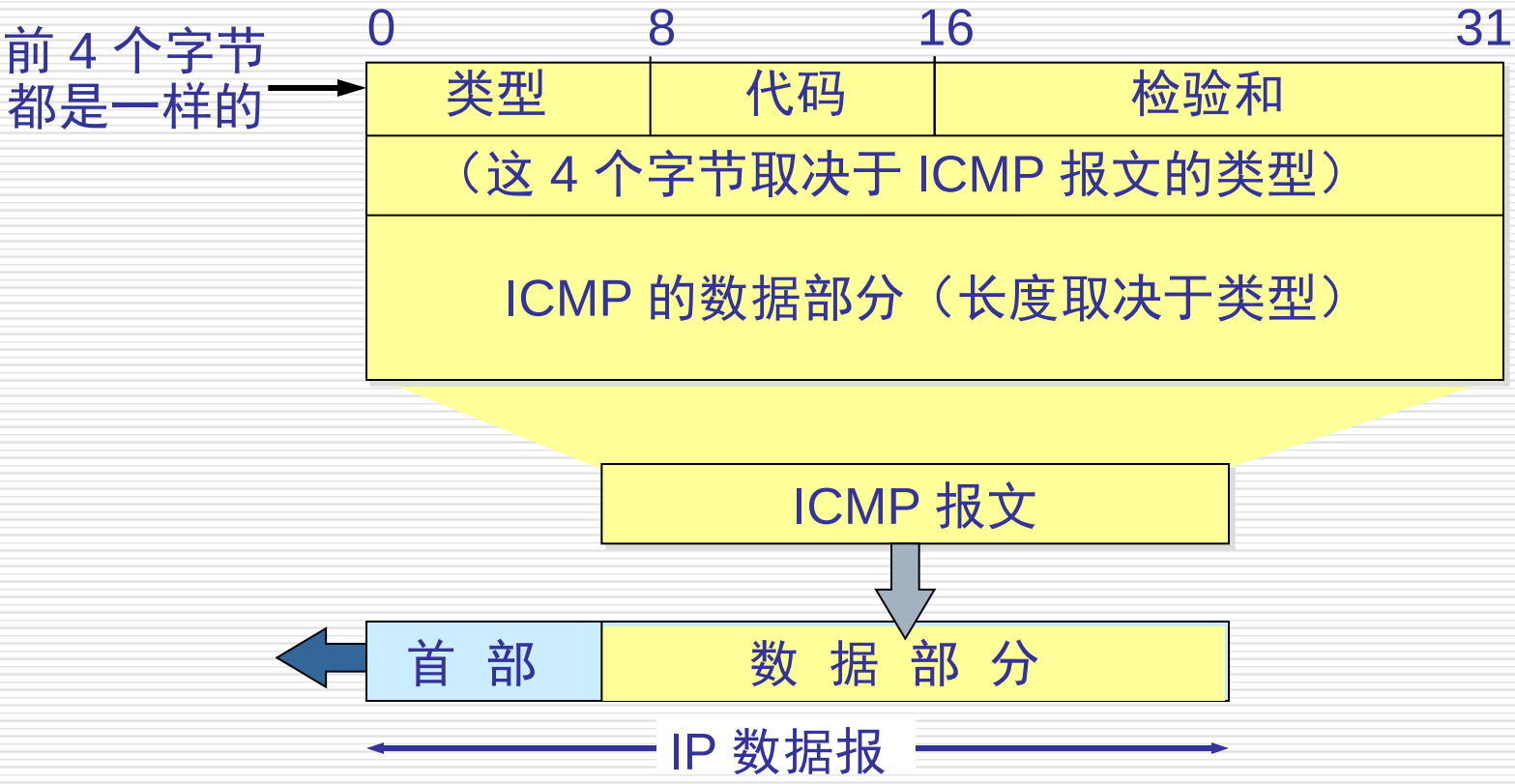
逆地址解析协议 RARP

- 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。

5.4 网际控制报文协议 ICMP

- 为了提高 IP 数据报交付成功的机会，在网际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


前 4 个字节
都是一样的
0
8
16
31
类型
代码
检验和
（这 4 个字节取决于 ICMP 报文的类型）
ICMP 的数据部分（长度取决于类型）
ICMP 报文
首 部
数 据 部 分
IP 数据报

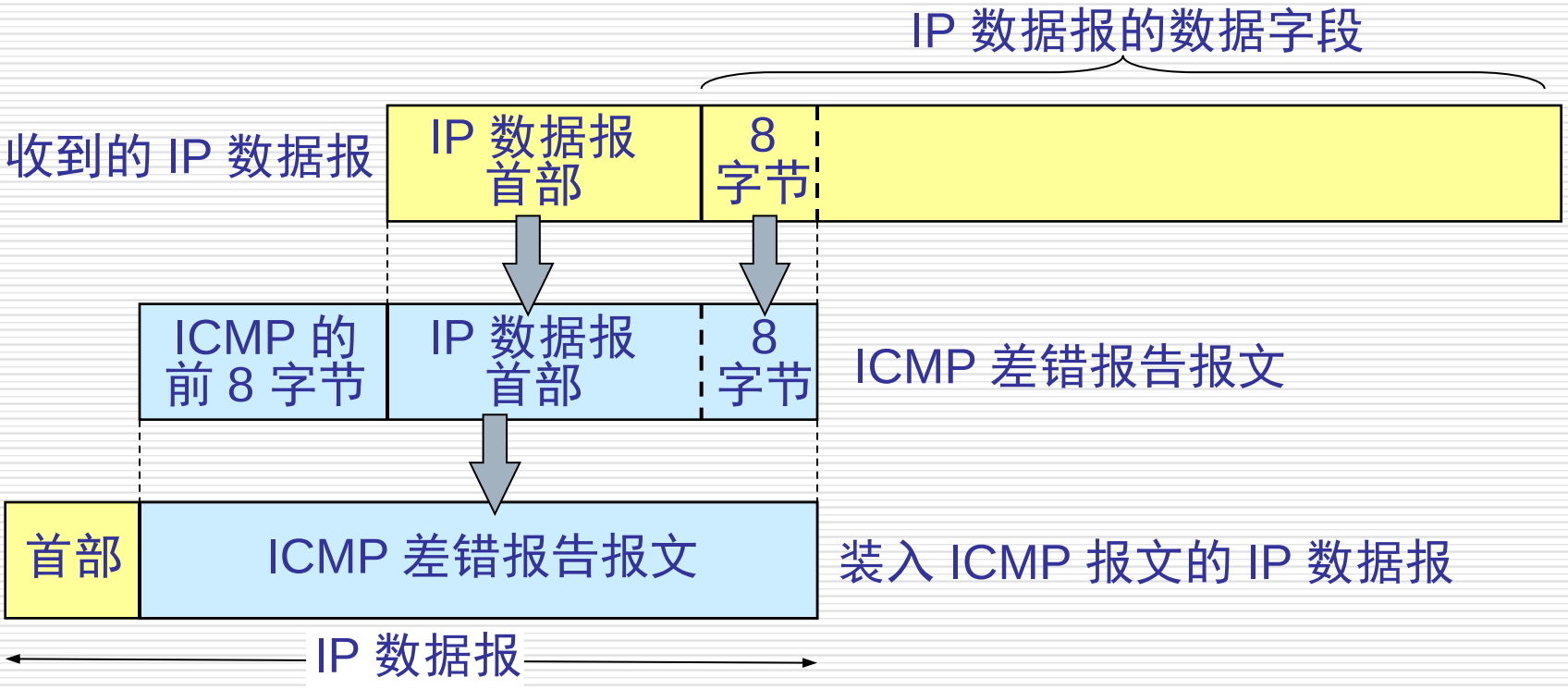
5.4.1 ICMP 报文的种类

- 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

5.4.2 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.
```

5.5 因特网的路由选择协议

5.5.1 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。
- 现在对自治系统 AS 的定义是强调下面的事实：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

- 因特网的早期 RFC 文档中未使用“路由器”而是使用“网关”这一名词。但是在新的 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。
- IGP 和 EGP 是协议类别的名称。但 RFC 在使用 EGP 这个名词时出现了一点混乱，因为最早的一个外部网关协议的协议名字正好也是 EGP。因此在遇到名词 EGP 时，应弄清它是指旧的协议 EGP 还是指外部网关协议 EGP 这个类别。

因特网的路由选择协议

- 内部网关协议 IGP：具体的协议有多种，如 RIP 和 OSPF 等。

- 外部网关协议 EGP：目前使用的协议就是 BGP。

5.5.2 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

- 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。
- 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。
- 这里的“距离”实际上指的是“最短距离”，

“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 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 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。

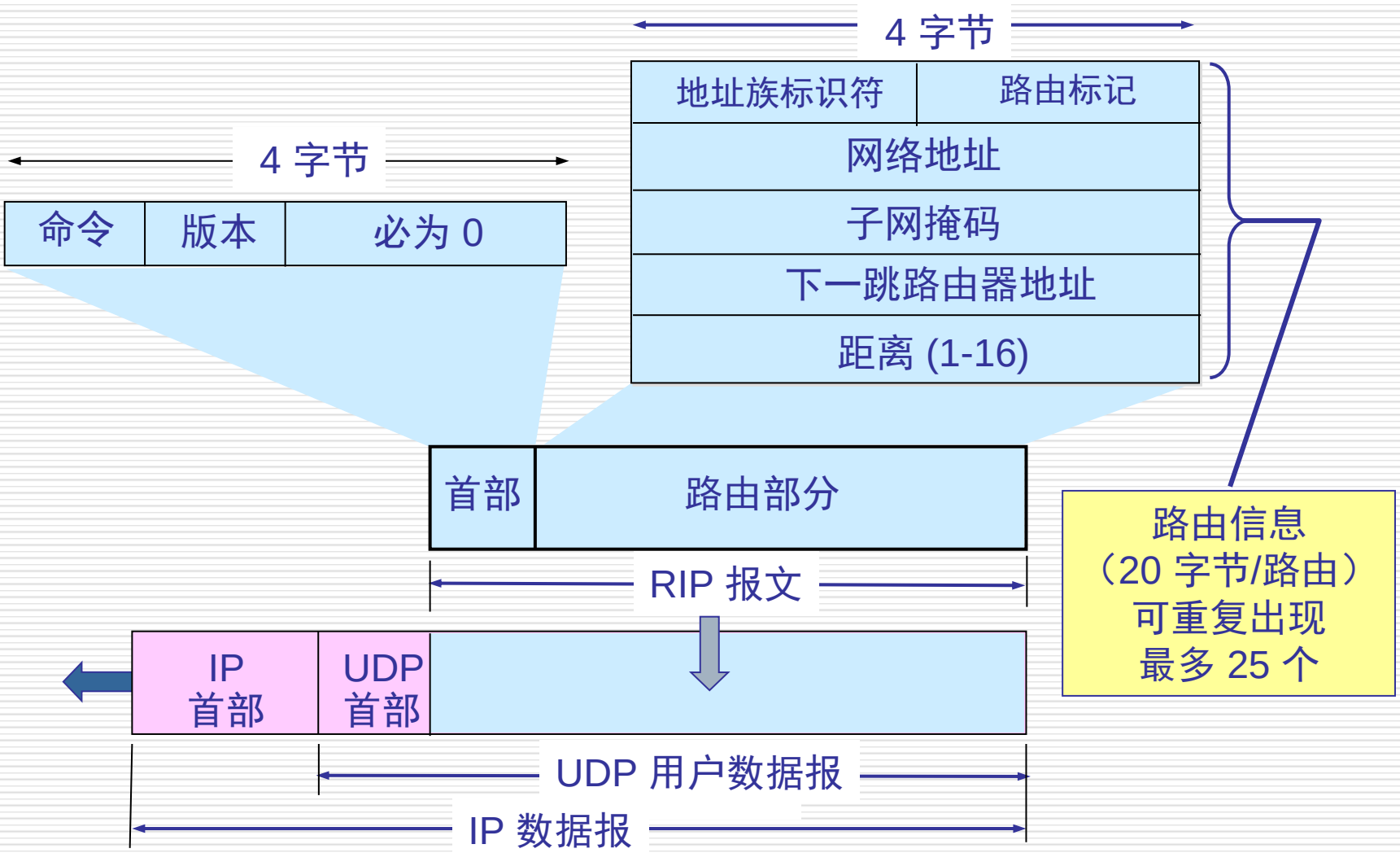
2. 距离向量算法(重述)

收到相邻路由器（其地址为 X）的一个 RIP 报文：
(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不在路由表中，则把该项目加到路由表中。
　否则
　　若下一跳字段给出的路由器地址是同样的，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的更新路由表，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。

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局路由信息，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

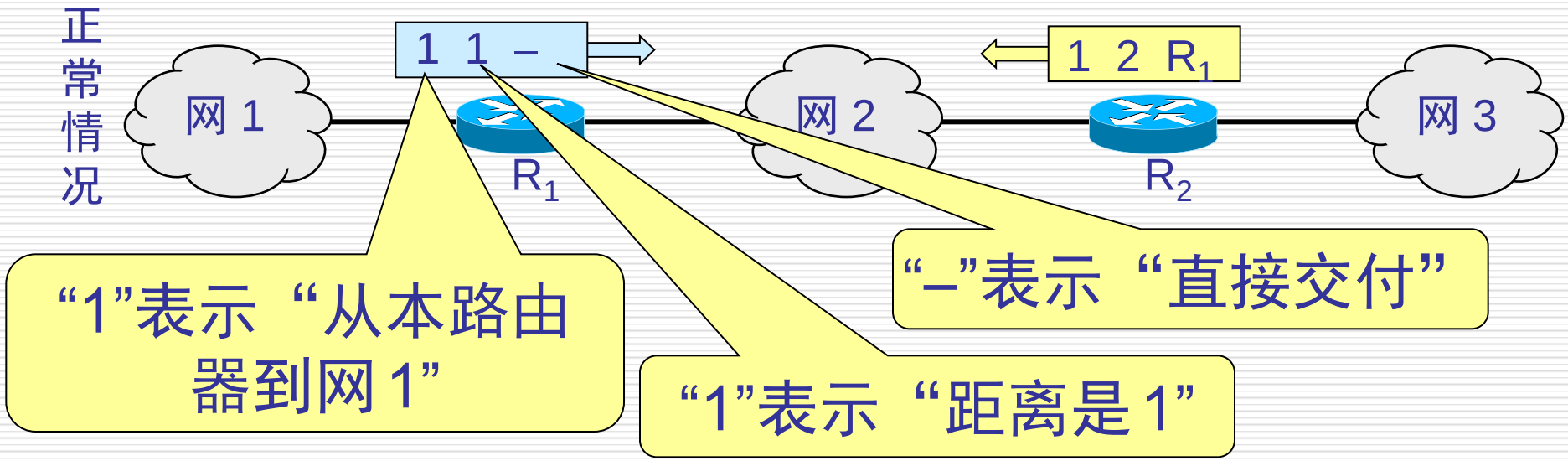


RIP2 的报文由首部和路由部分组成

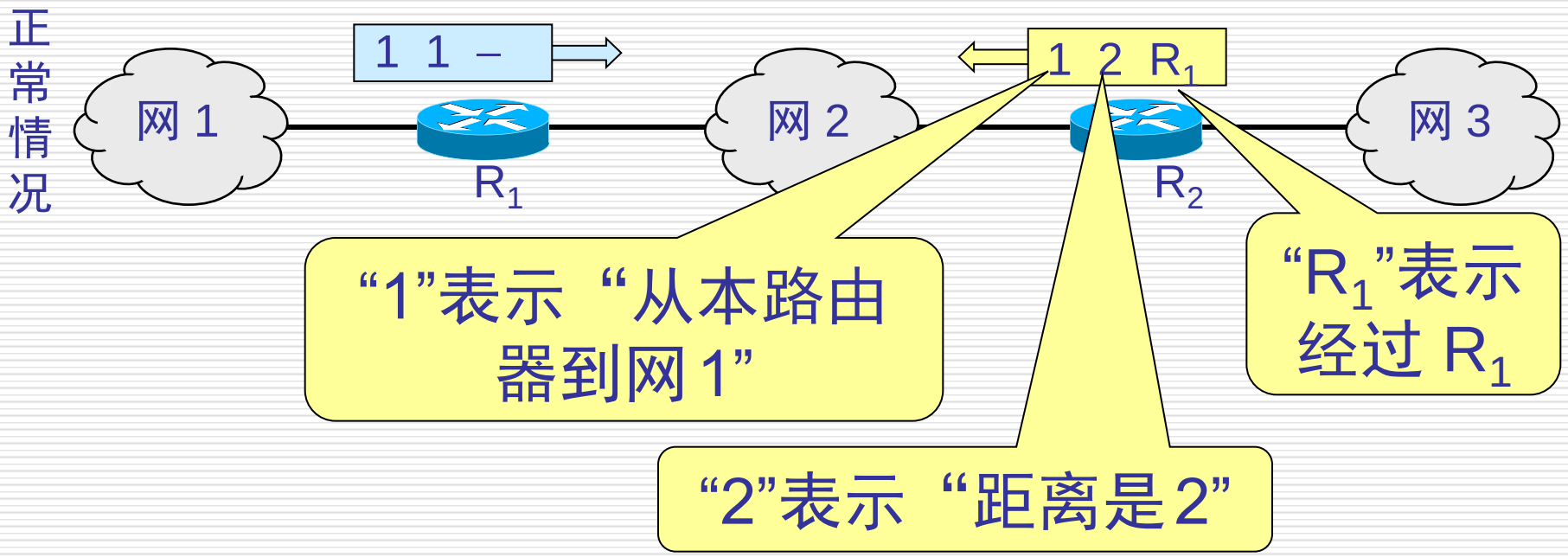
- RIP2 报文中的路由部分由若干个路由信息组成。每个路由信息需要用 20 个字节。地址族标识符（又称为地址类别）字段用来标志所使用的地址协议。
- 路由标记填入自治系统的号码，这是考虑使RIP 有可能收到本自治系统以外的路由选择信息。再后面指出某个网络地址、该网络的子网掩码、下一跳路由器地址以及到此网络的距离。

RIP 协议的优缺点

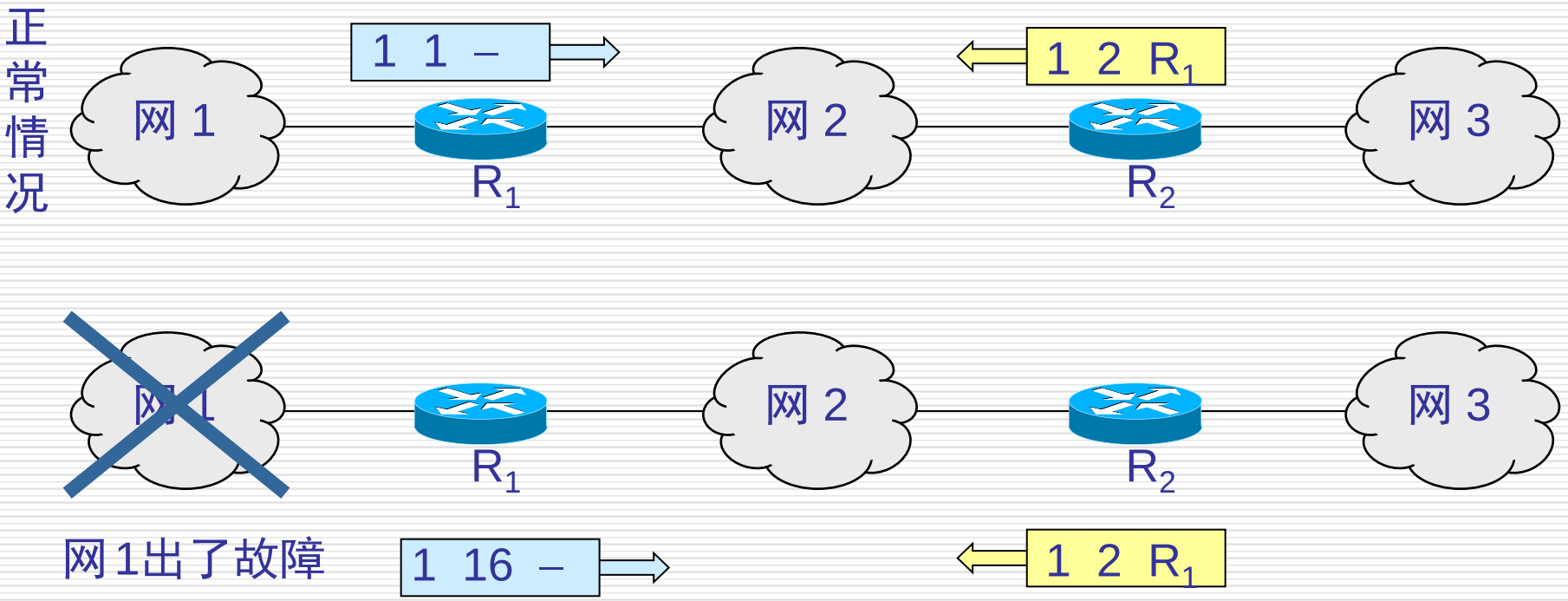
- RIP 存在的一个问题是当网络出现故障时，要经过比较长的时间才能将此信息传送到所有的路由器。
- RIP 协议最大的优点就是实现简单，开销较小。
- RIP 限制了网络的规模，它能使用的最大距离为 15（16 表示不可达）。
- 路由器之间交换的路由信息是路由器中的完整路由表，因而随着网络规模的扩大，开销也就增加。



R_1 说：“我到网 1 的距离是 1，是直接交付。”

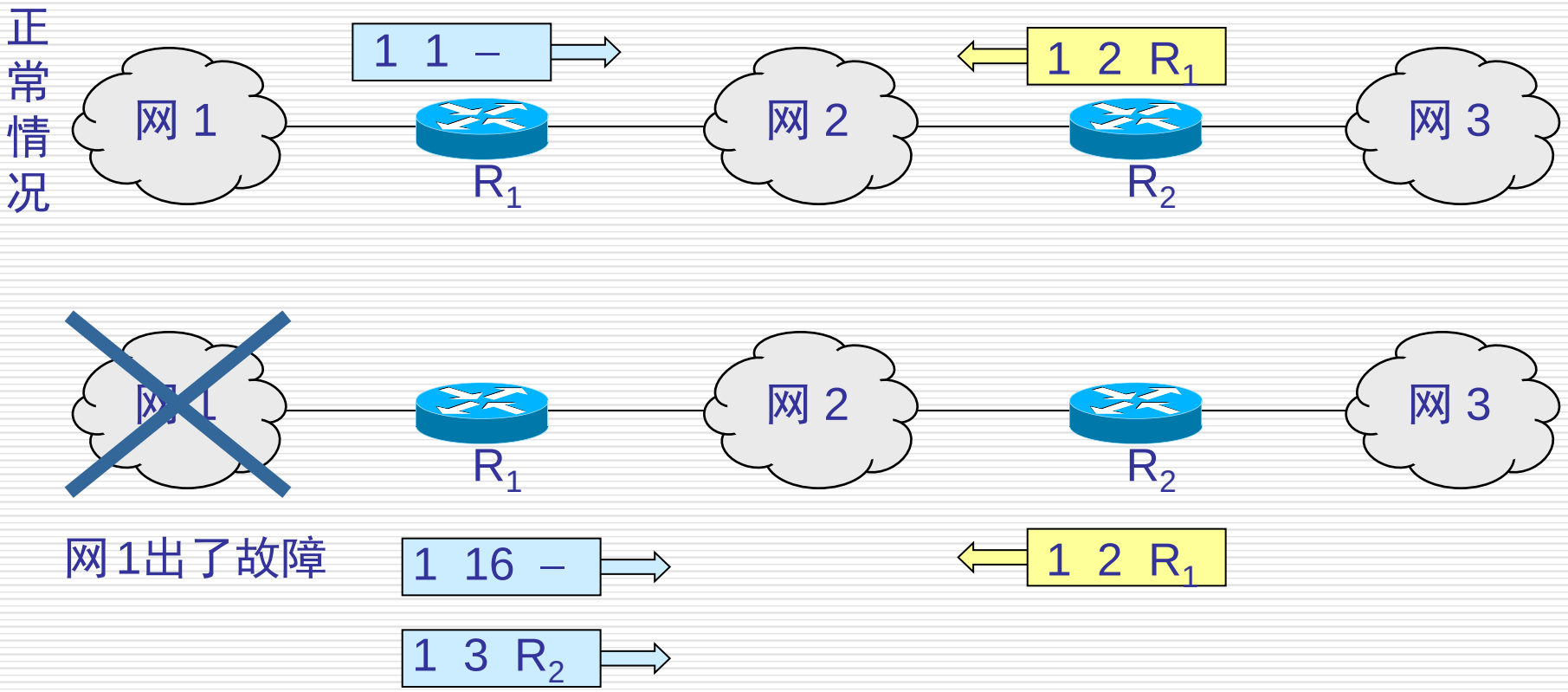


R_2 说：“我到网 1 的距离是 2，是经过 R_1。”

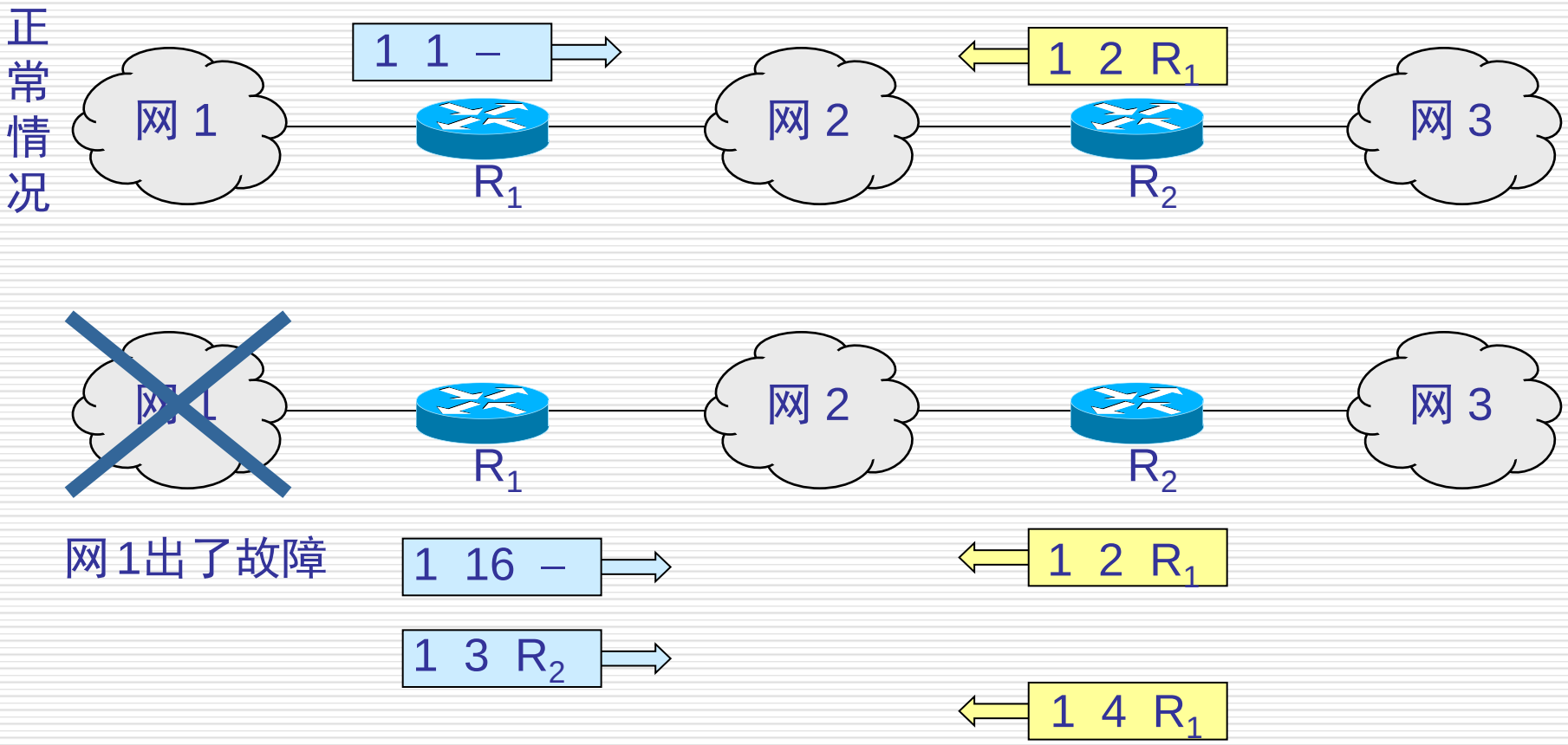


R_1 说：“我到网 1 的距离是 16（表示无法到达），是直接交付。”

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。

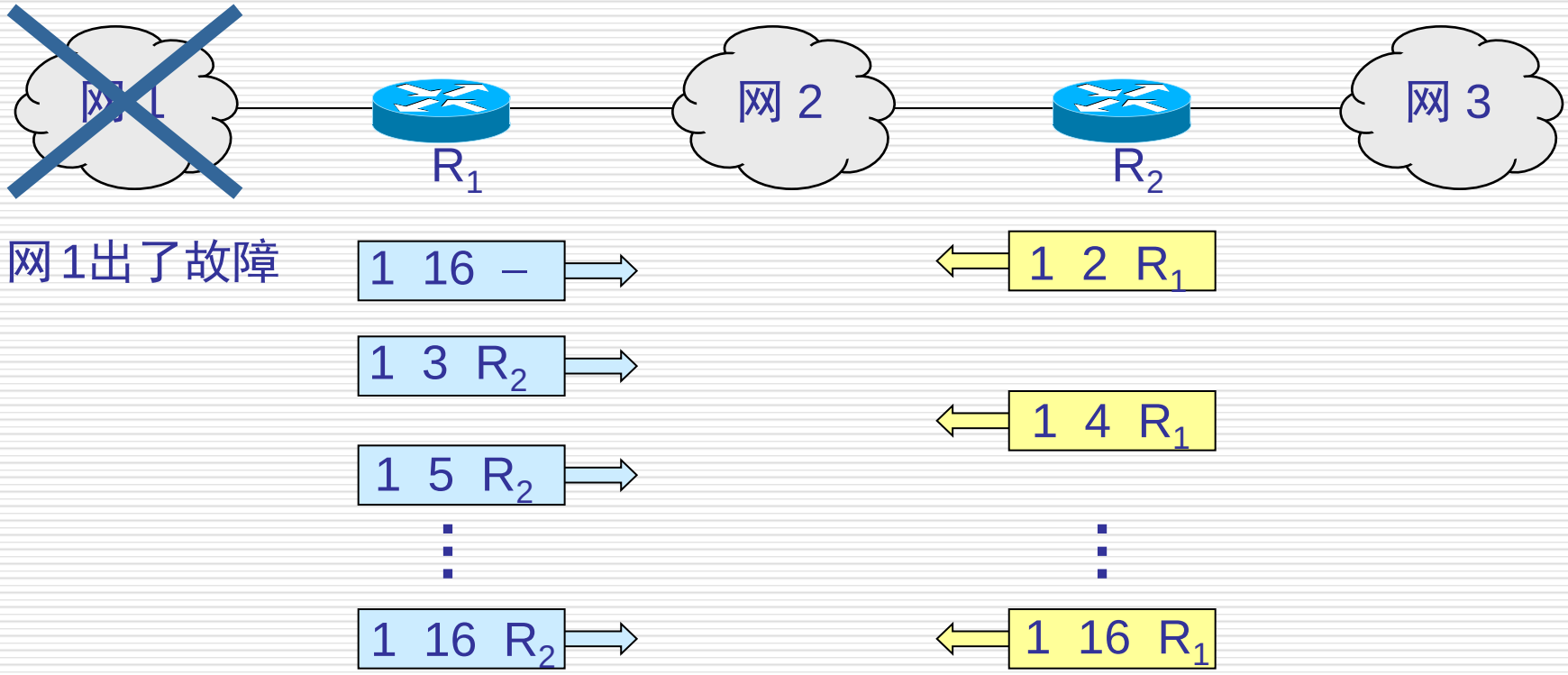


R_1 收到 R_2 的更新报文后，误认为可经过 R_2 到达网1，于是更新自己的路由表，说：“我到网 1 的距离是 3，下一跳经过 R_2”。然后将此更新信息发送给 R_2。



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 是不可达的。

5.5.3 内部网关协议 OSPF--(Open Shortest Path First)

1. 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 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”。
- 是分布式的**链路状态协议**。

OSPF协议三个要点

- 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。

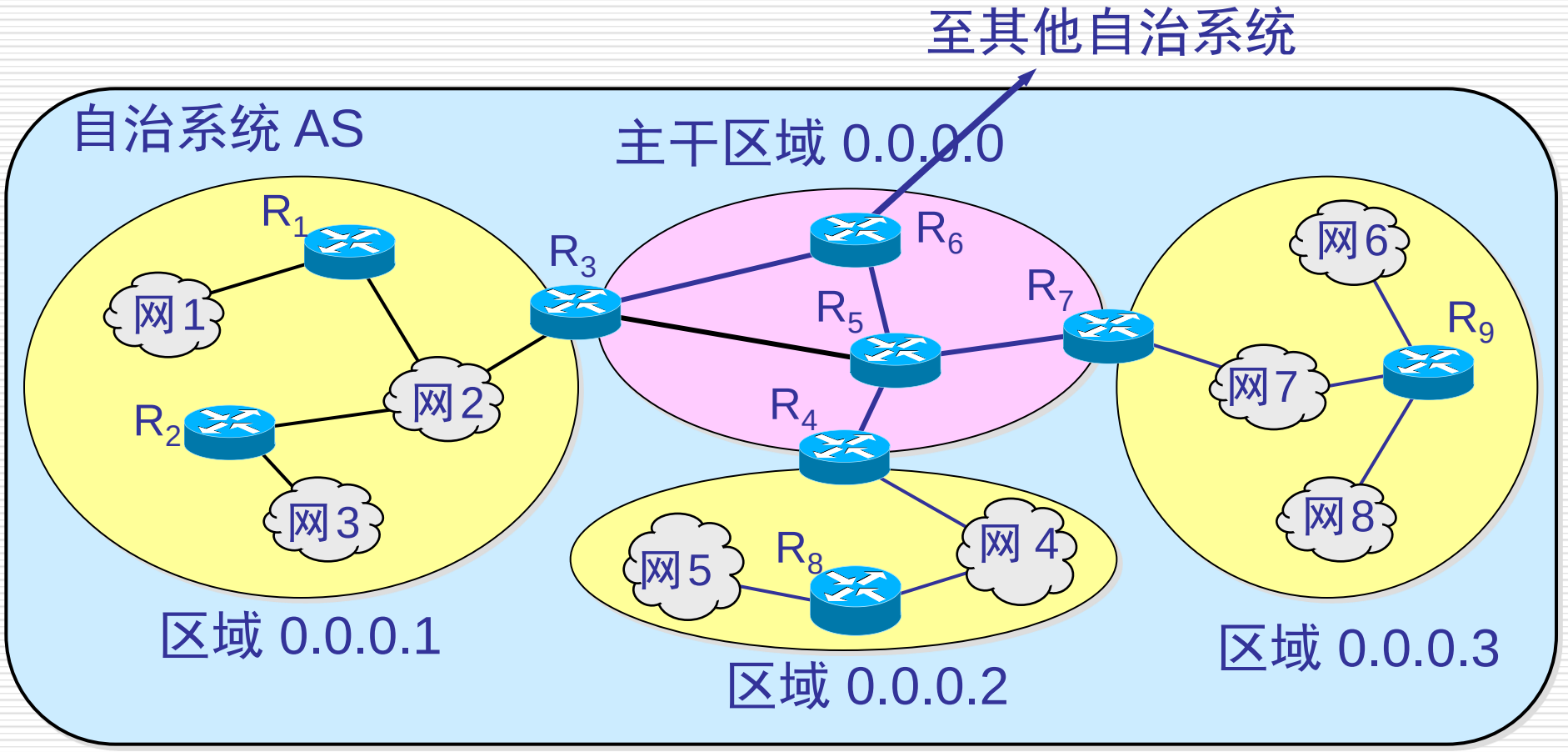
链路状态数据库--(link-state database)

- 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 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。

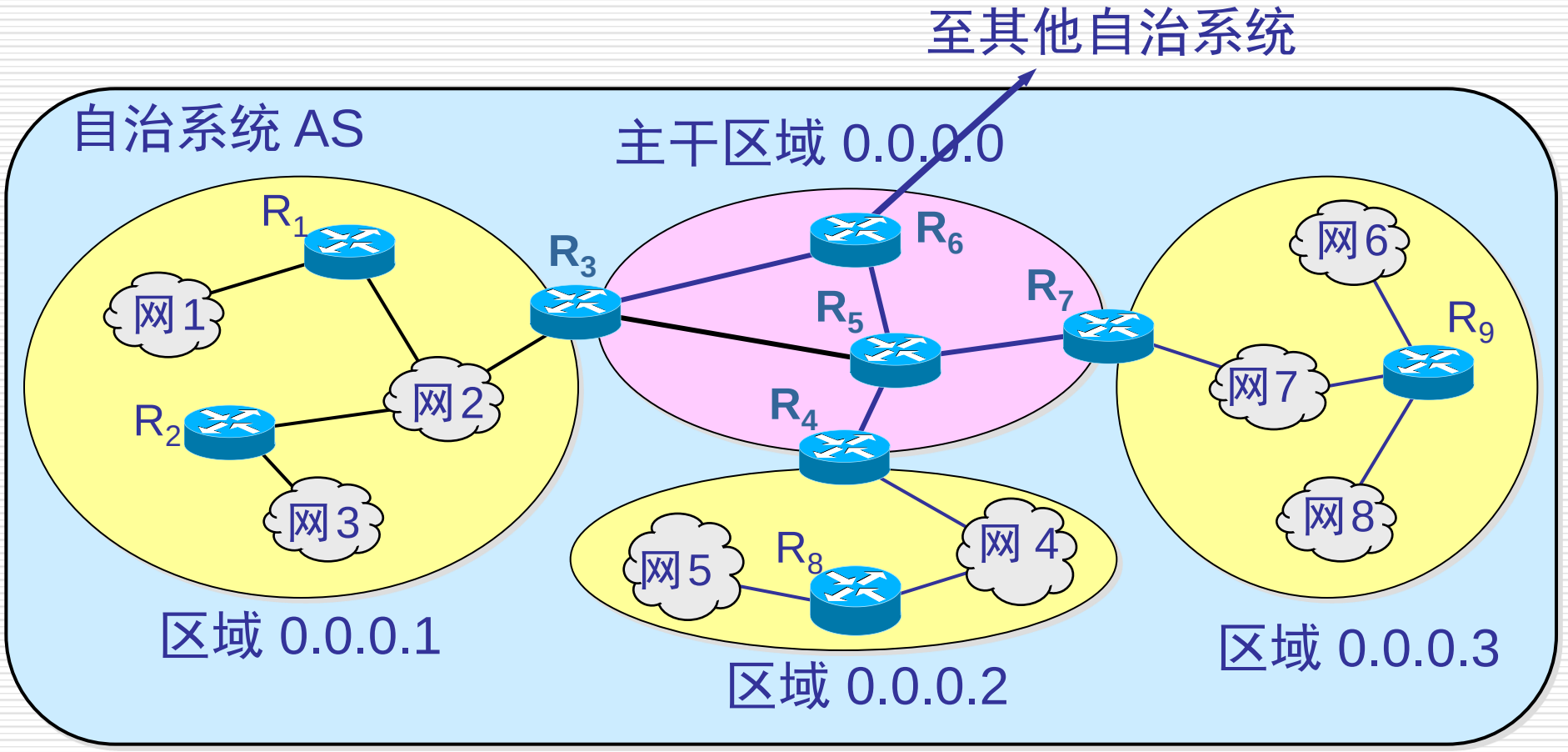
OSPF 划分为两种不同的区域



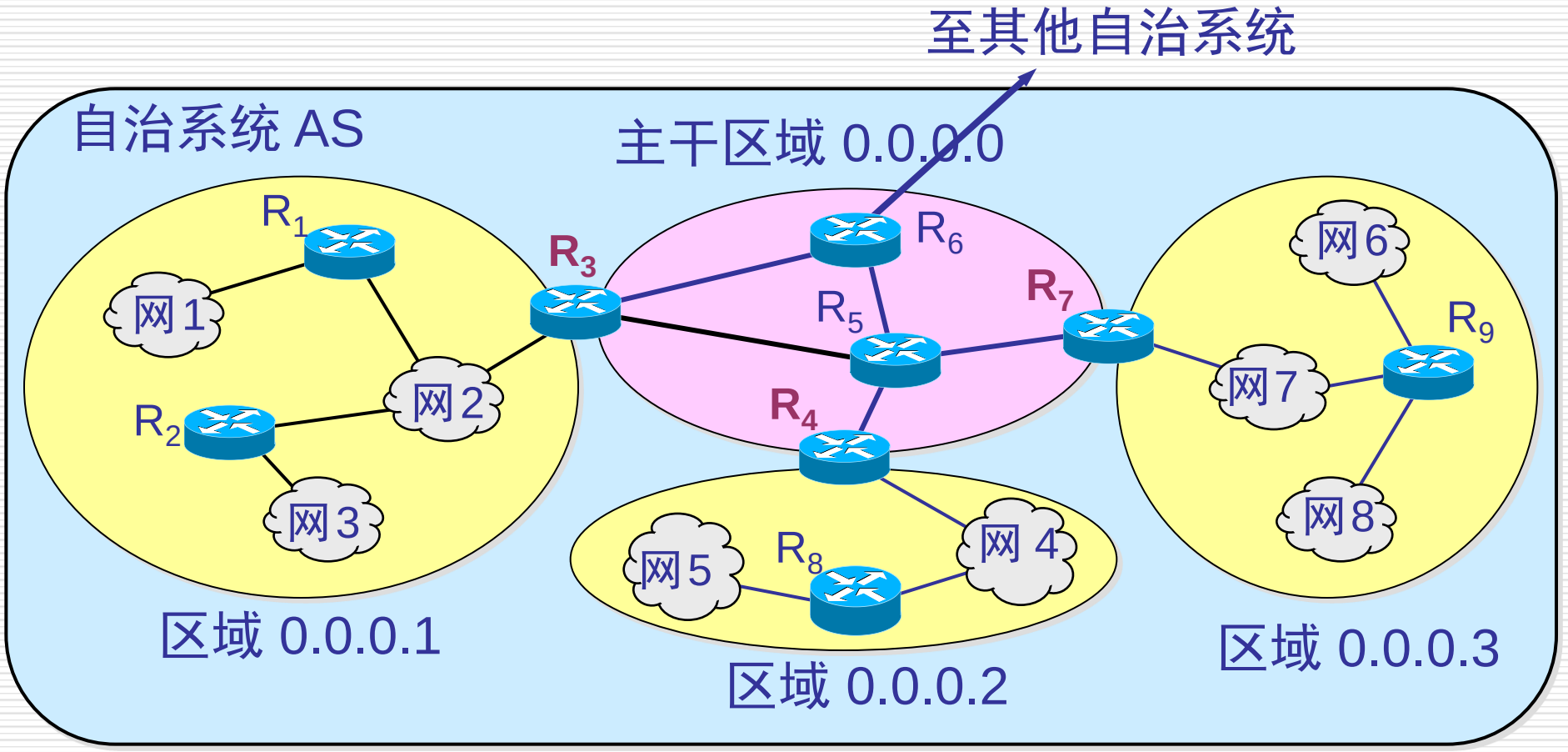
划分区域

- 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。
- 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。

主干路由器


至其他自治系统
自治系统 AS
主干区域 0.0.0.0
R1
网 1
网 2
R2
网 3
区域 0.0.0.1
R3
R6
R5
R7
R4
网 5
R8
网 4
区域 0.0.0.2
网 6
R9
网 7
网 8
区域 0.0.0.3

区域边界路由器


至其他自治系统
自治系统 AS
主干区域 0.0.0.0
R1
网 1
网 2
R2
网 3
区域 0.0.0.1
R3
R6
R5
R7
R4
网 5
R8
网 4
区域 0.0.0.2
网 6
R9
网 7
网 8
区域 0.0.0.3

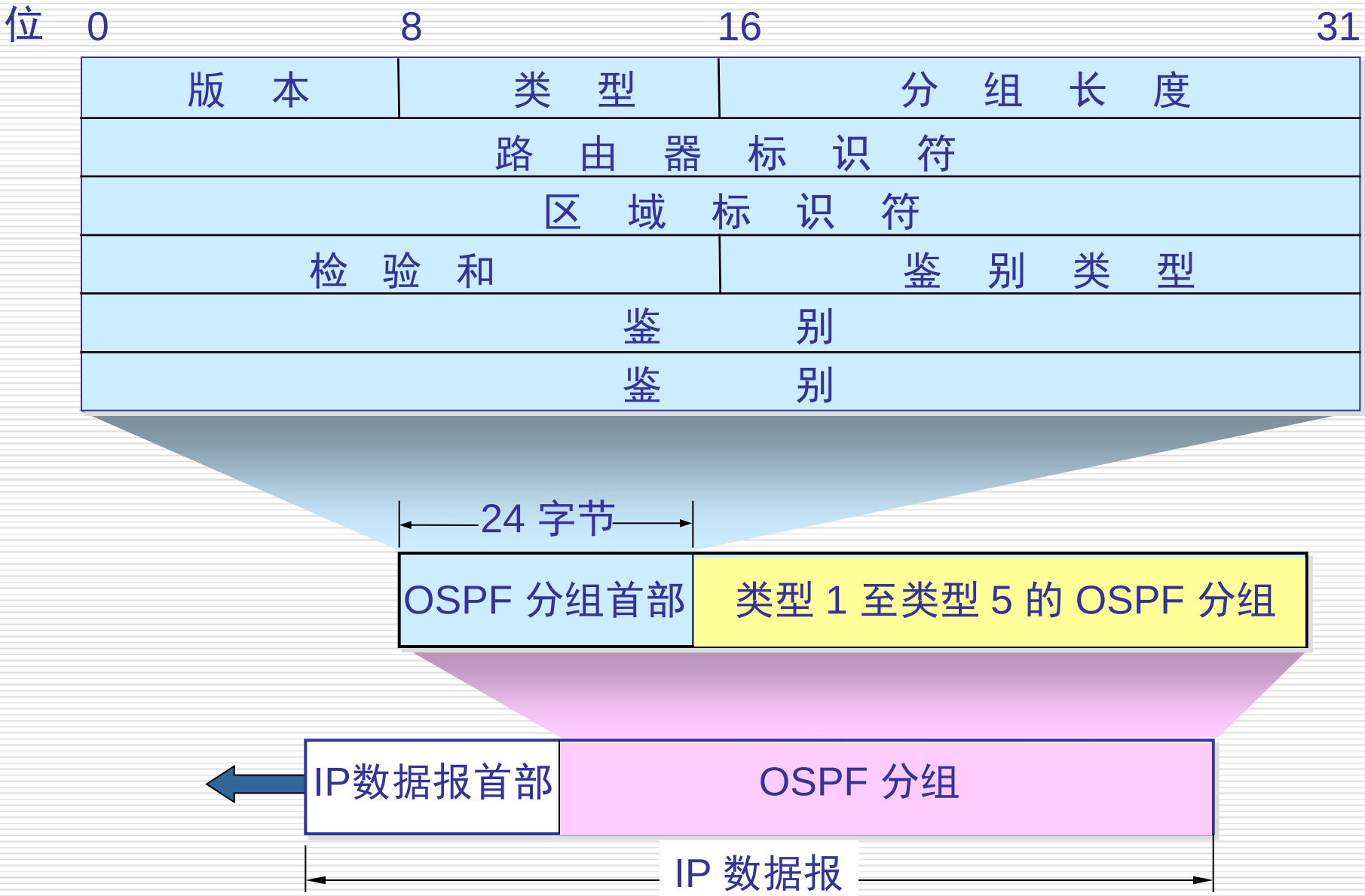
OSPF 直接用 IP 数据报传送

- OSPF 不用 UDP 而是直接用 IP 数据报传送。
- OSPF 构成的数据报很短。这样做可减少路由信息的通信量。
- 数据报很短的另一好处是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

- OSPF 对不同的链路可根据 IP 分组的不同服务类型 TOS 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。
- 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。
- 所有在 OSPF 路由器之间交换的分组都具有鉴别的功能。
- 支持可变长度的子网划分和无分类编址 CIDR。
- 每一个链路状态都带上一个 32 位的序号，序号越大状态就越新。

OSPF 分组


位
0
8
16
31
版 本
类 型
分 组 长 度
路 由 器 标 识 符
区 域 标 识 符
检 验 和
鉴 别 类 型
鉴 别
鉴 别
24 字节
OSPF 分组首部
类型 1 至类型 5 的 OSPF 分组
IP数据报首部
OSPF 分组
IP 数据报

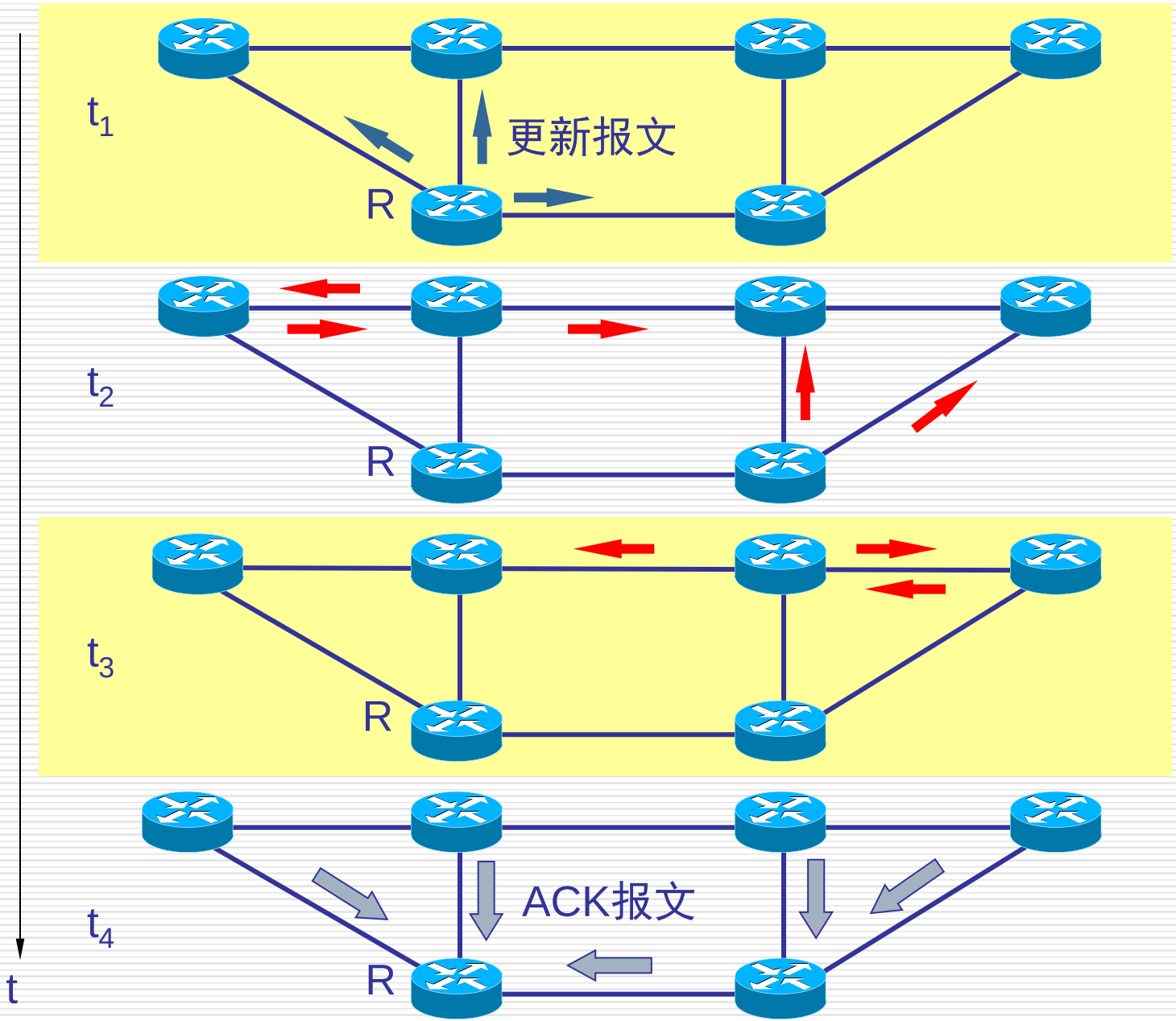
2. OSPF 的五种分组类型

- 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


确定可达性
问候
问候
达到数据库的同步
数据库描述
数据库描述
数据库描述
数据库描述
新情况下的同步
链路状态请求
链路状态更新
链路状态确认

OSPF 使用的是可靠的洪泛法



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 50 ms。

指定的路由器--(designated router)

- 多点接入的局域网采用了指定的路由器的方法，使广播的信息量大大减少。
- 指定的路由器代表该局域网上所有的链路向连接到该网络上的各路由器发送状态信息。

5.5.4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

- 因特网的规模太大，使得自治系统之间路由选择非常困难。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。
  - 当一条路径通过几个不同 AS 时，要想对这样的路径计算出有意义的代价是不太可能的。
  - 比较合理的做法是在 AS 之间交换“可达性”信息。
- 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。

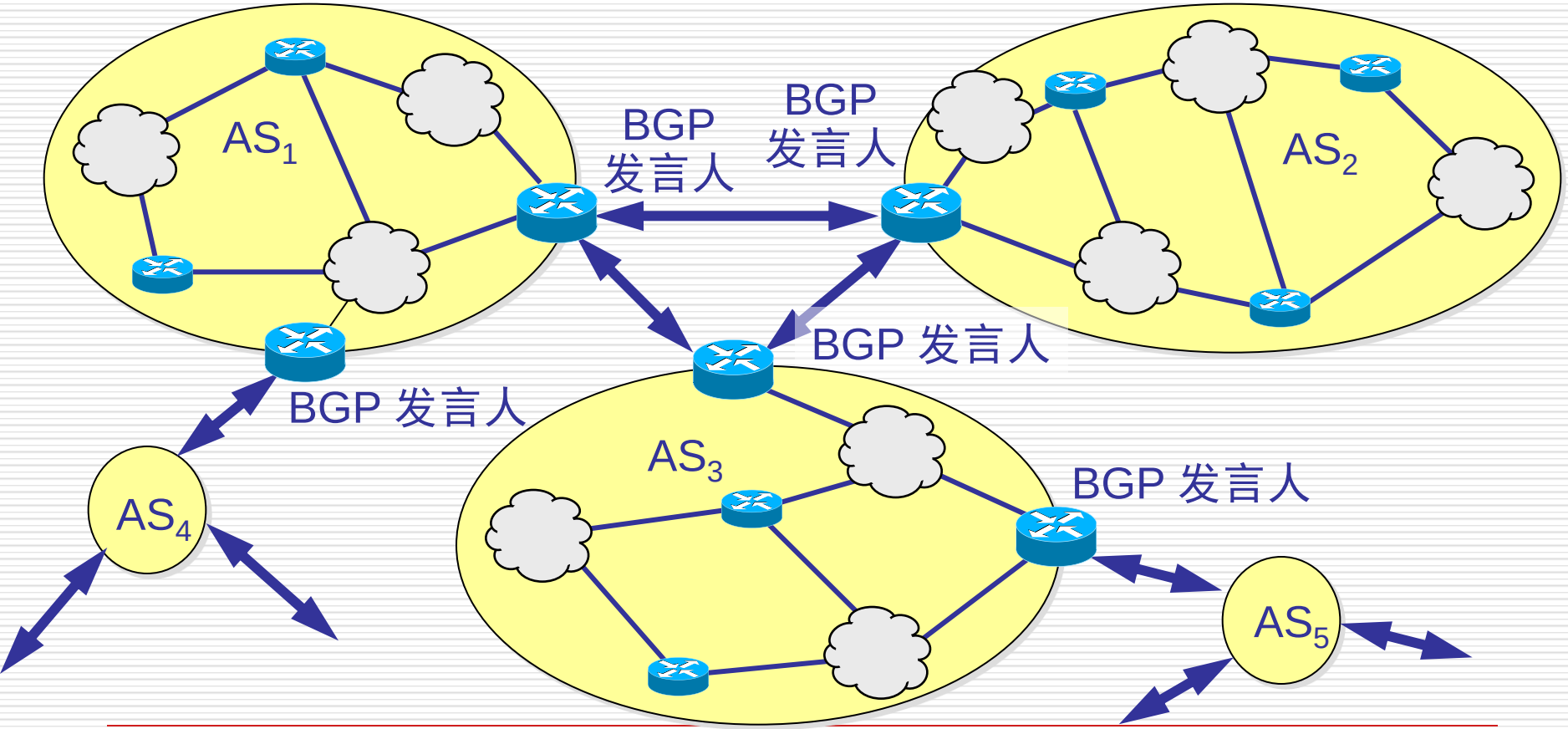
BGP 发言人--(BGP speaker)

- 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

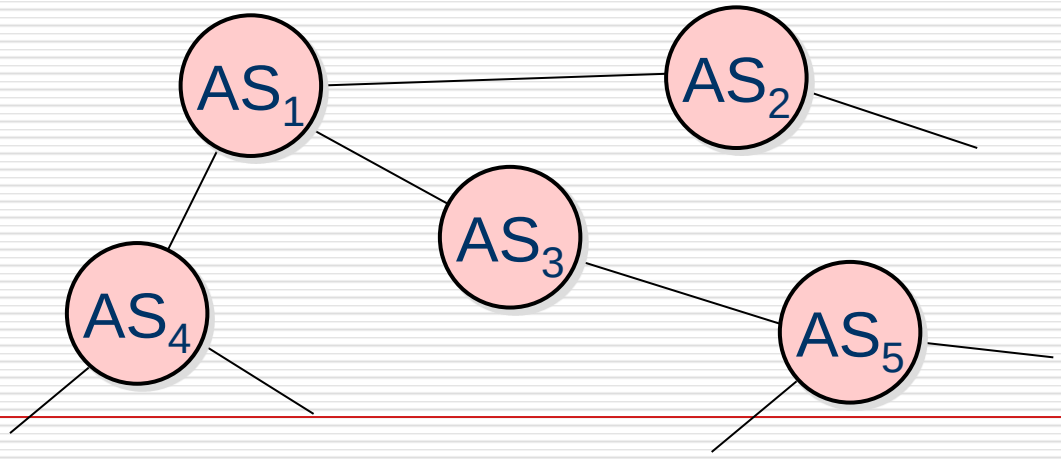
- 一个 BGP 发言人与其他自治系统中的 BGP 发言人要交换路由信息，就要先建立 TCP 连接，然后在此连接上交换 BGP 报文以建立 BGP 会话(session)，利用 BGP 会话交换路由信息。
- 使用 TCP 连接能提供可靠的服务，也简化了路由选择协议。
- 使用 TCP 连接交换路由信息的两个 BGP 发言人，彼此成为对方的邻站或对等站。

BGP 发言人和自治系统 AS 的关系



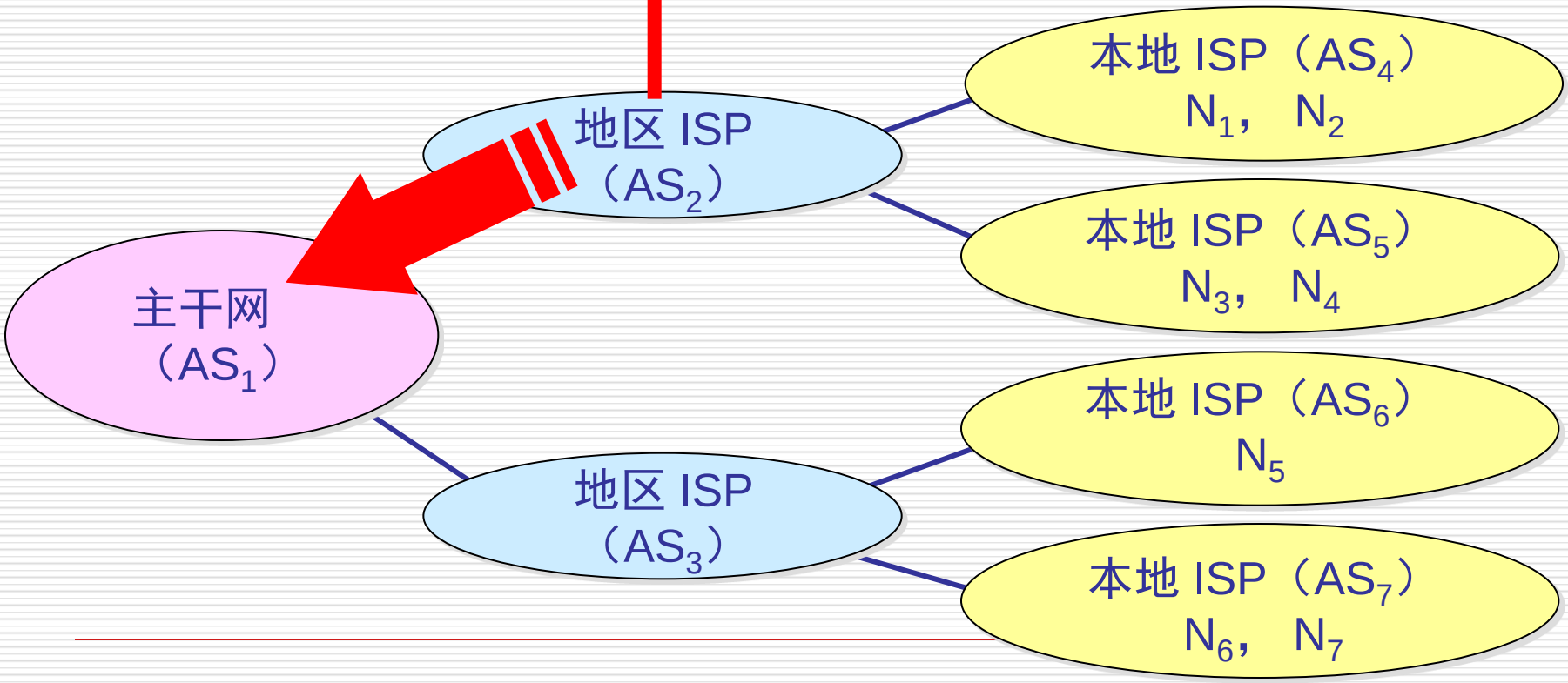
AS 的连通图举例

- BGP 所交换的网络可达性的信息就是要到达某个网络所要经过的一系列 AS。
- 当 BGP 发言人互相交换了网络可达性的信息后，各 BGP 发言人就根据所采用的策略从收到的路由信息中找出到达各 AS 的较好路由。



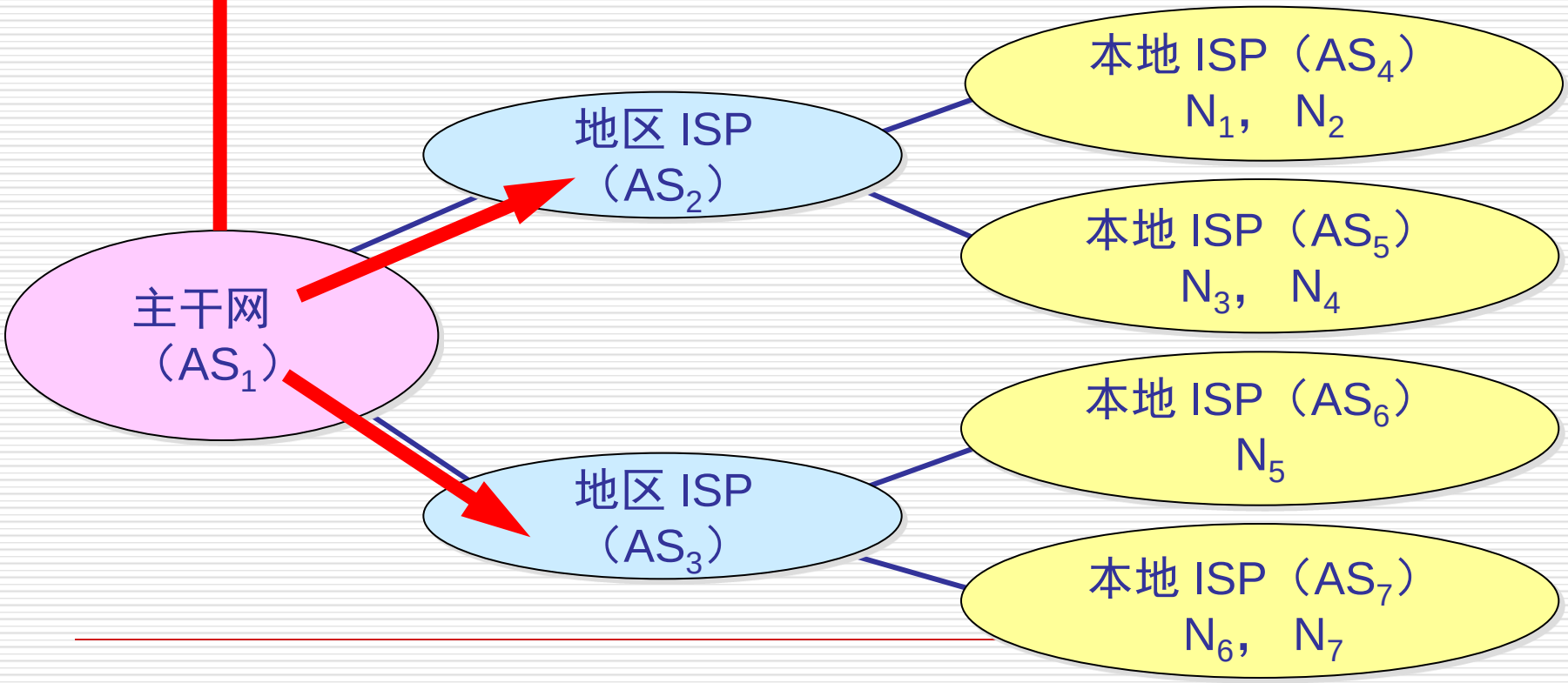
BGP 发言人交换路径向量

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

- 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

- BGP 支持 CIDR，因此 BGP 的路由表也就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。这样做对节省网络带宽和减少路由器的处理开销方面都有好处。

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。

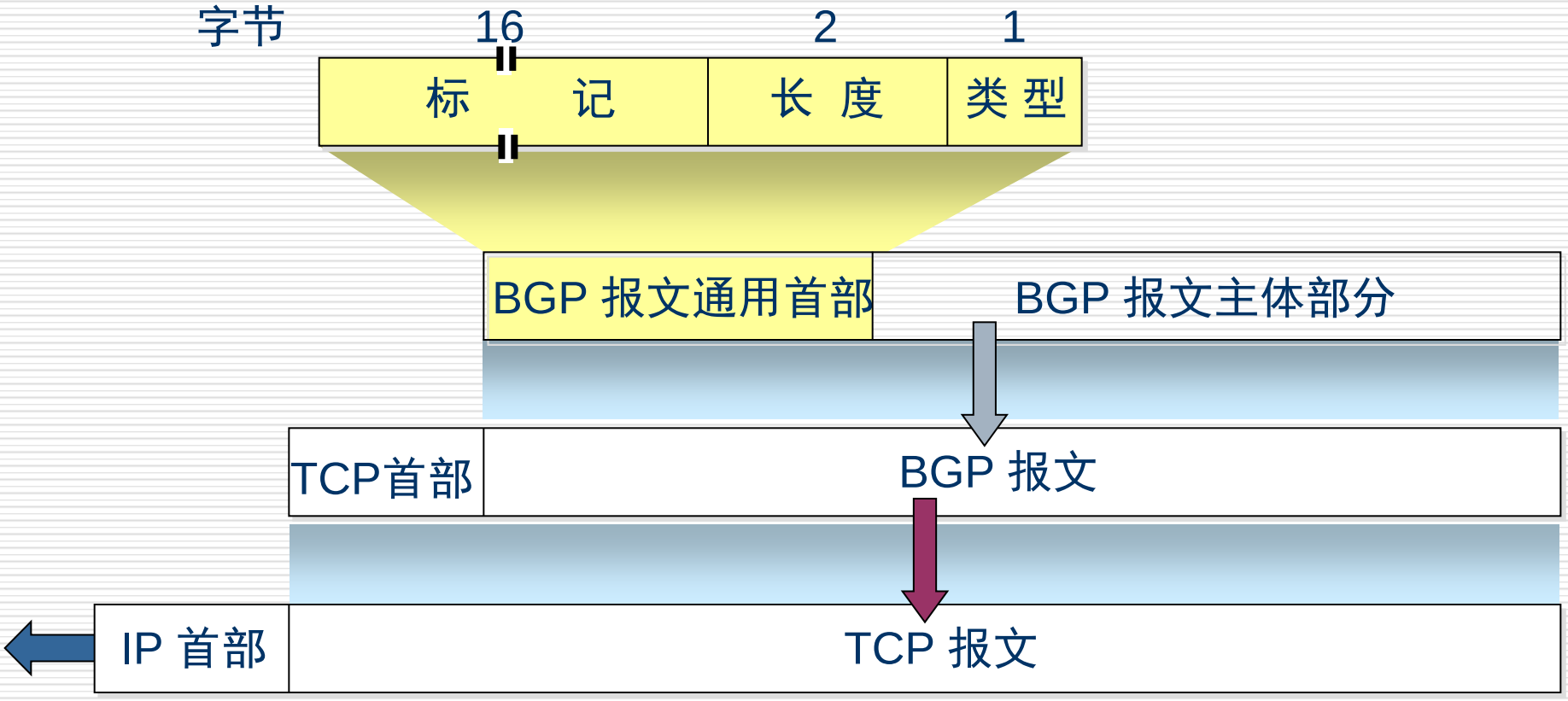
(2) 更新(UPDATE)报文，用来发送某一路由的信息，以及列出要撤消的多条路由。

(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。

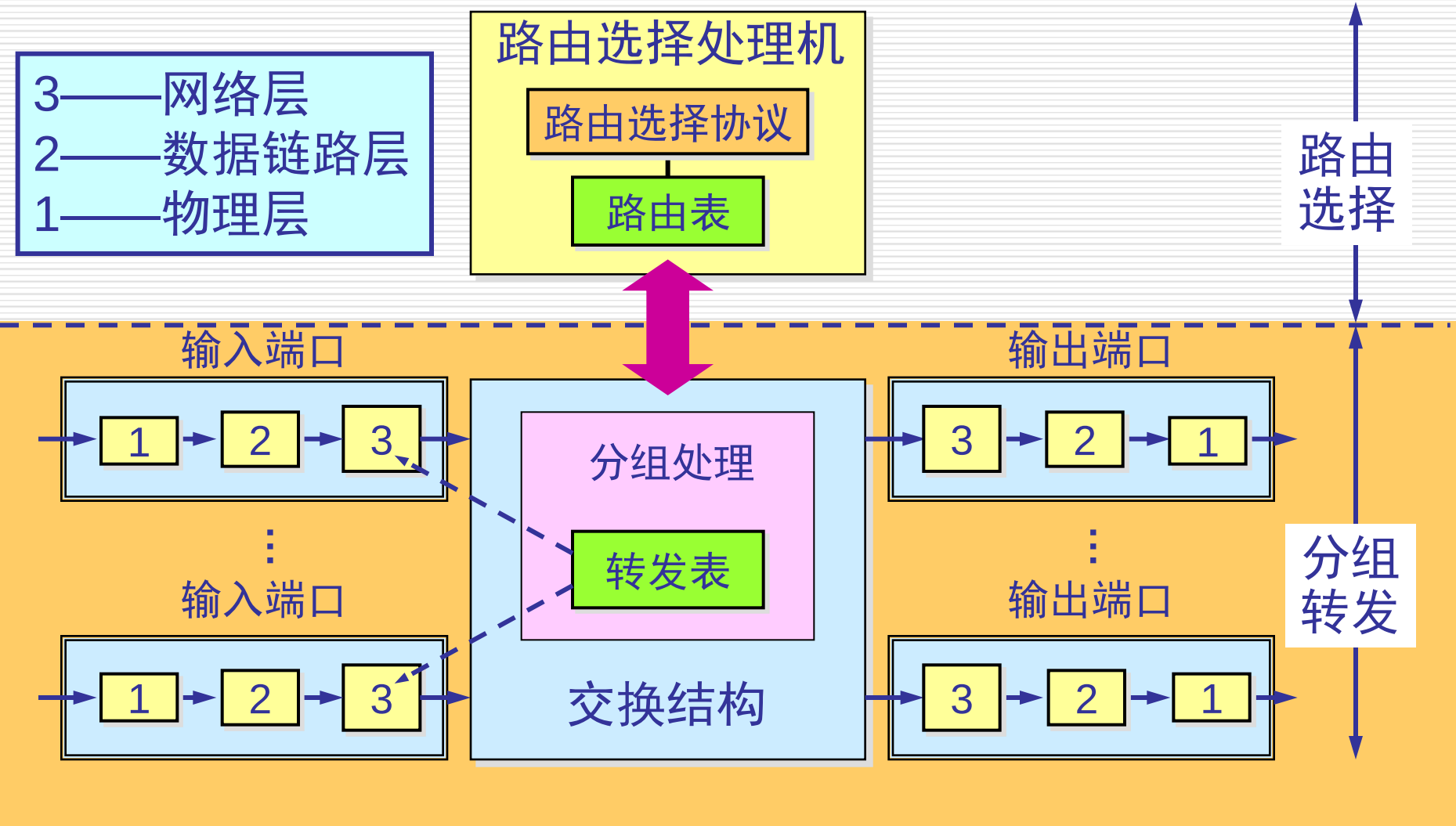
BGP 报文具有通用的首部



5.5.6 路由器在网际互连中的作用--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

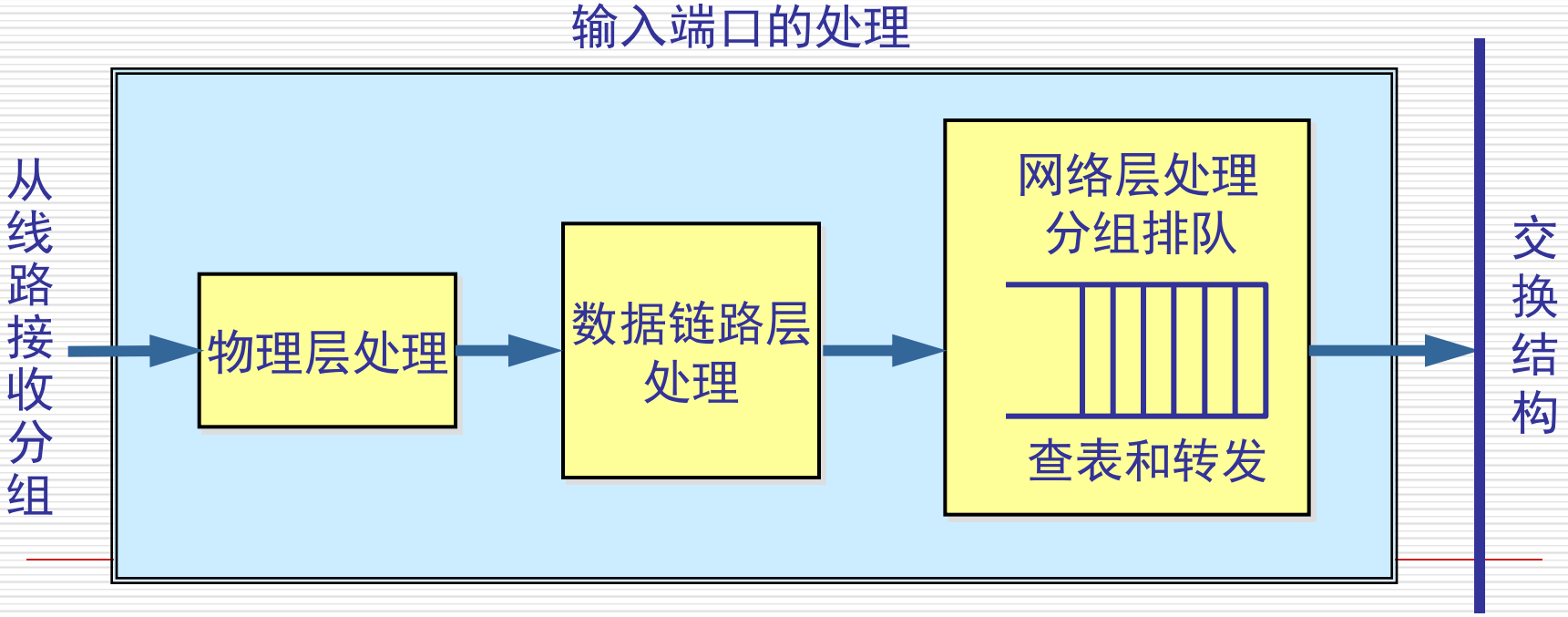

3——网络层
2——数据链路层
1——物理层
路由选择处理机
路由选择协议
路由表
路由选择
输入端口
1
2
3
输入端口
1
2
3
分组处理
转发表
交换结构
输出端口
3
2
1
输出端口
3
2
1
分组转发

“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，

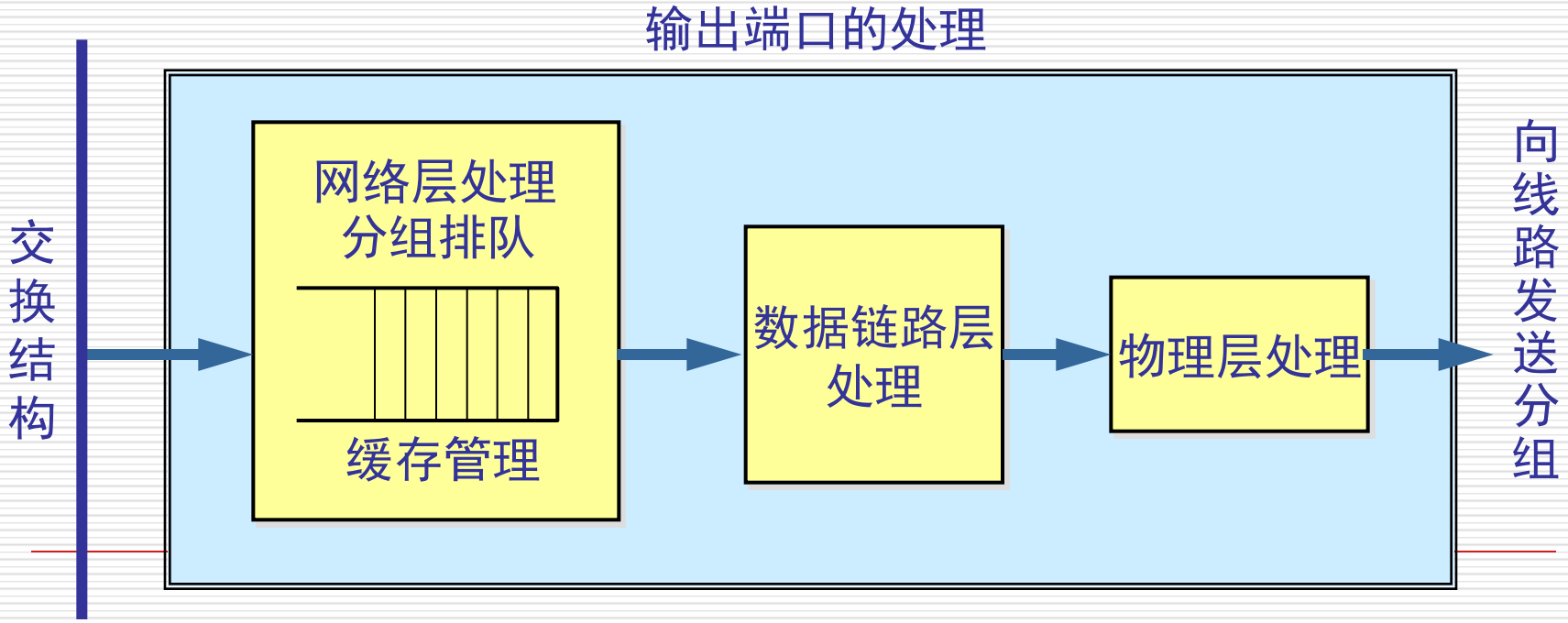
输入端口对线路上收到的分组的处理

- 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

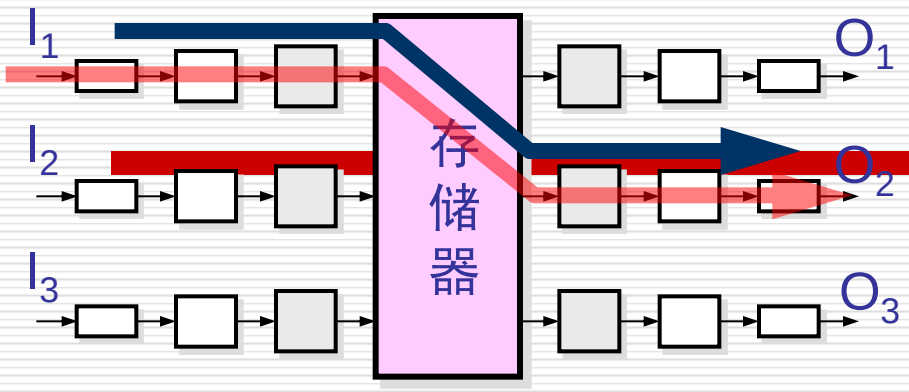
- 当交换结构传送过来的分组先进行缓存。数据链路层处理模块将分组加上链路层的首部和尾部，交给物理层后发送到外部线路。



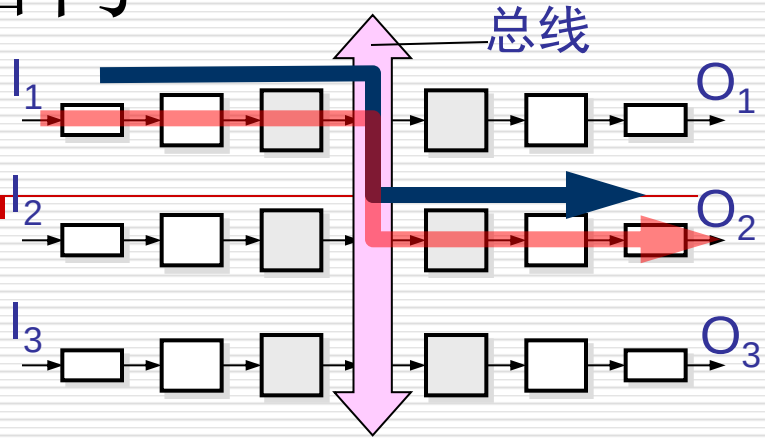
分组丢弃

- 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。

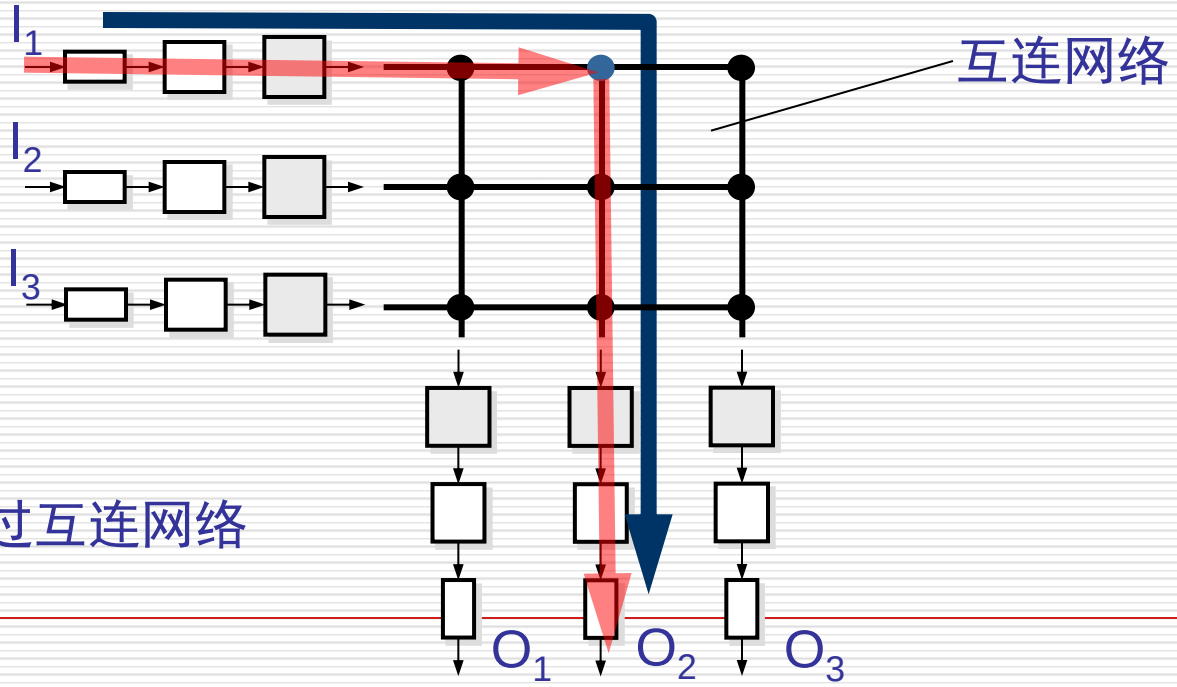
2 交换结构



(a) 通过存储器



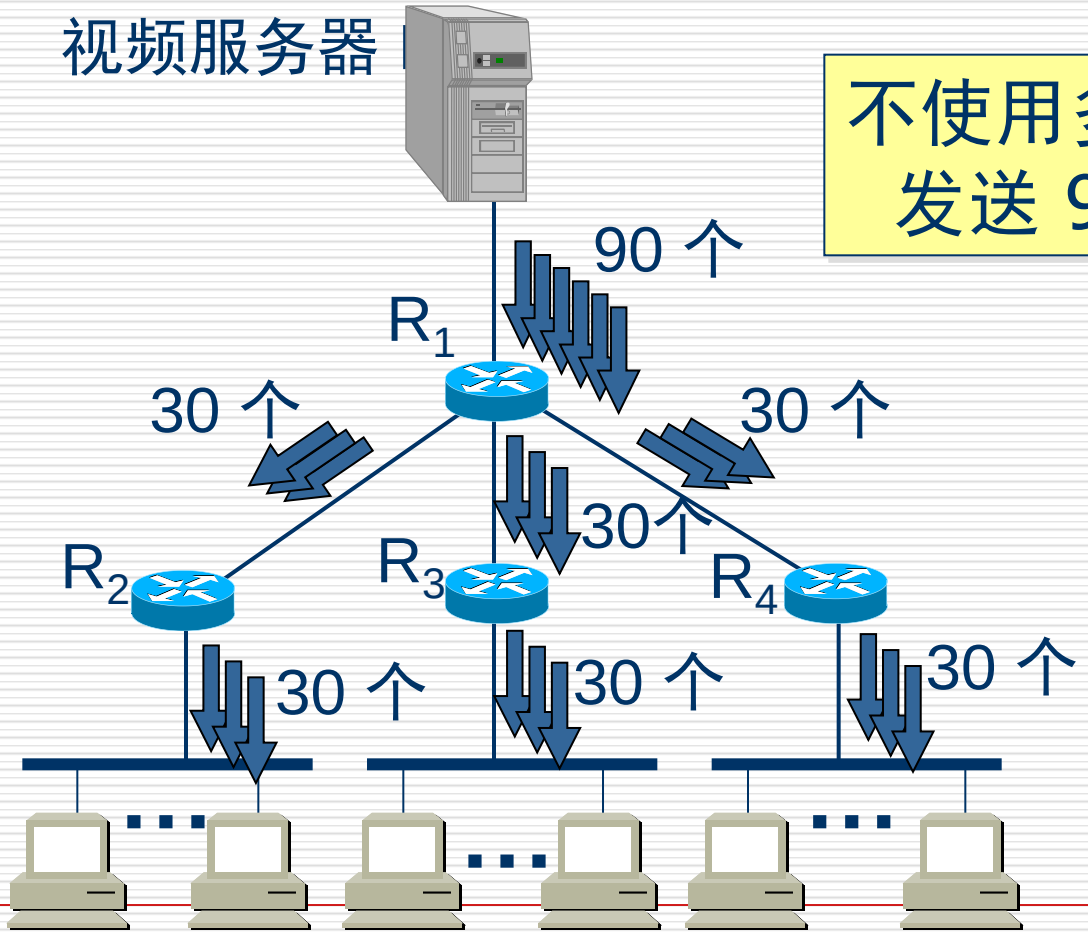
(b) 通过总线



(c) 通过互连网络

5.6 IP 多播

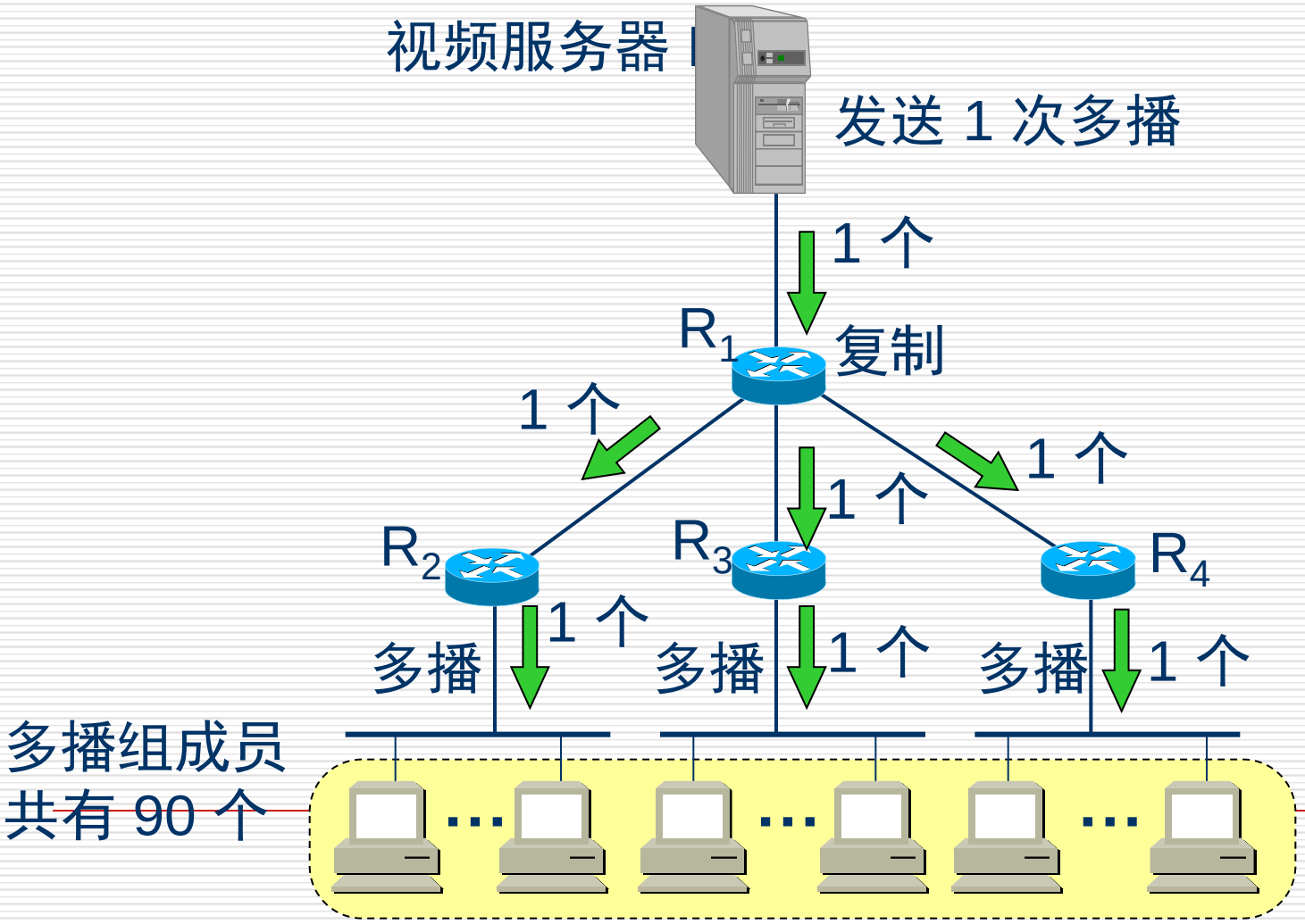
5.6.1 IP多播的基本概念



不使用多播时需要发送 90 次单播

共有 90 个主机接收视频节目

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

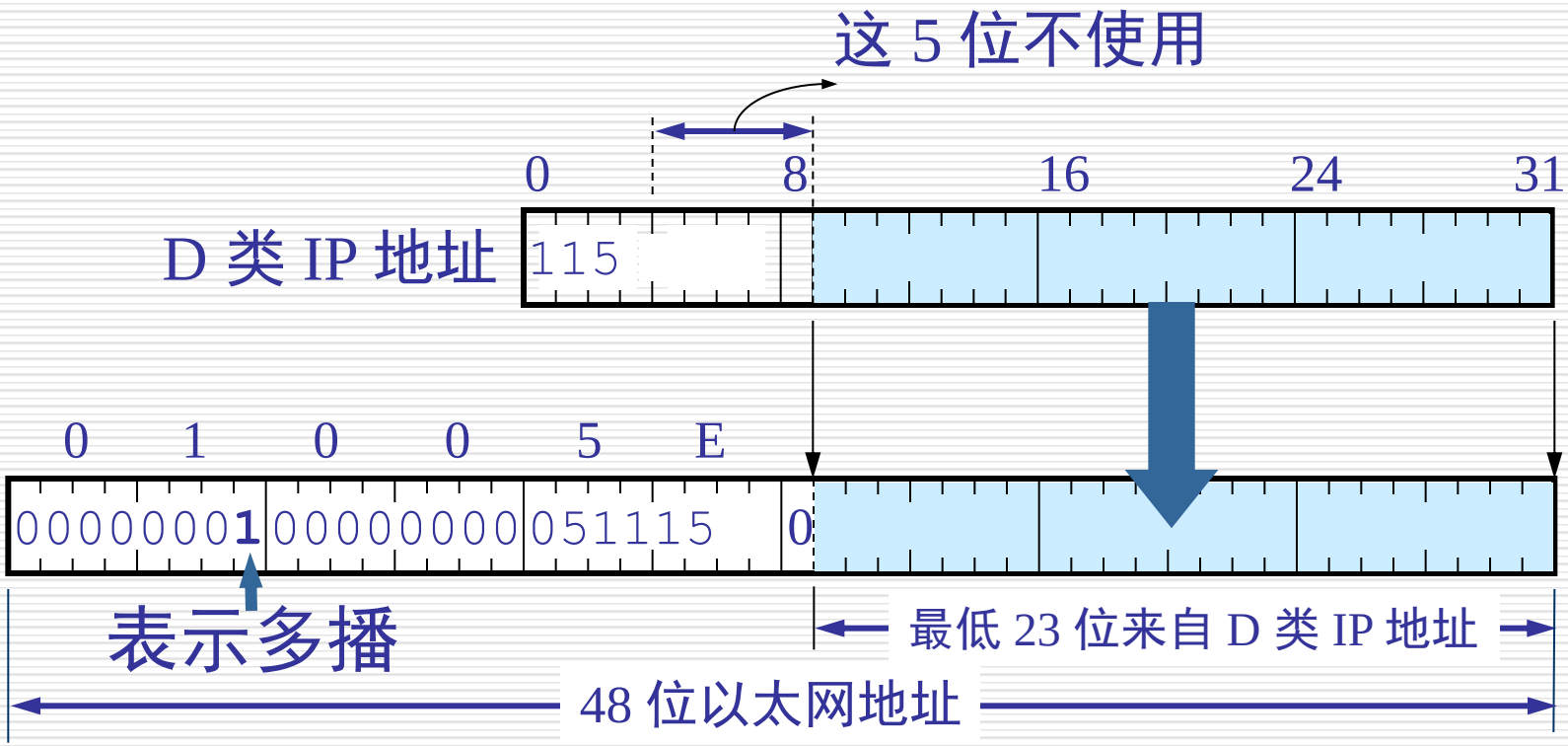
(4) 使用硬件进行多播

5.6.2 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 00-00-5E-00-00-00

  到 00-00-5E-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

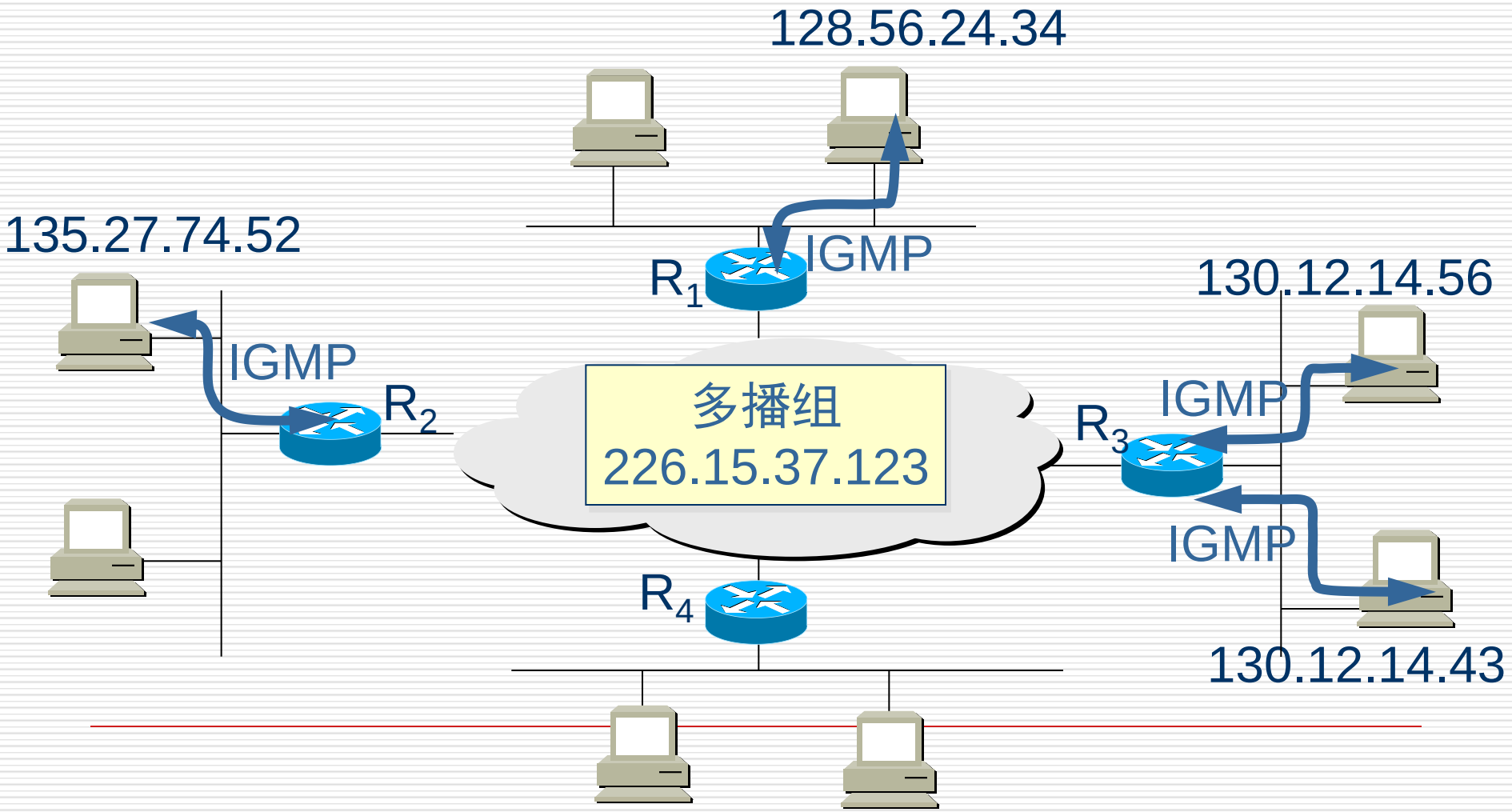


5.6.3 网际组管理协议 IGMP 和多播路由选择协议

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 5 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

- 在主机和多播路由器之间的所有通信都是使用 IP 多播。
- 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。
- 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）

- 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 5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。
- 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。

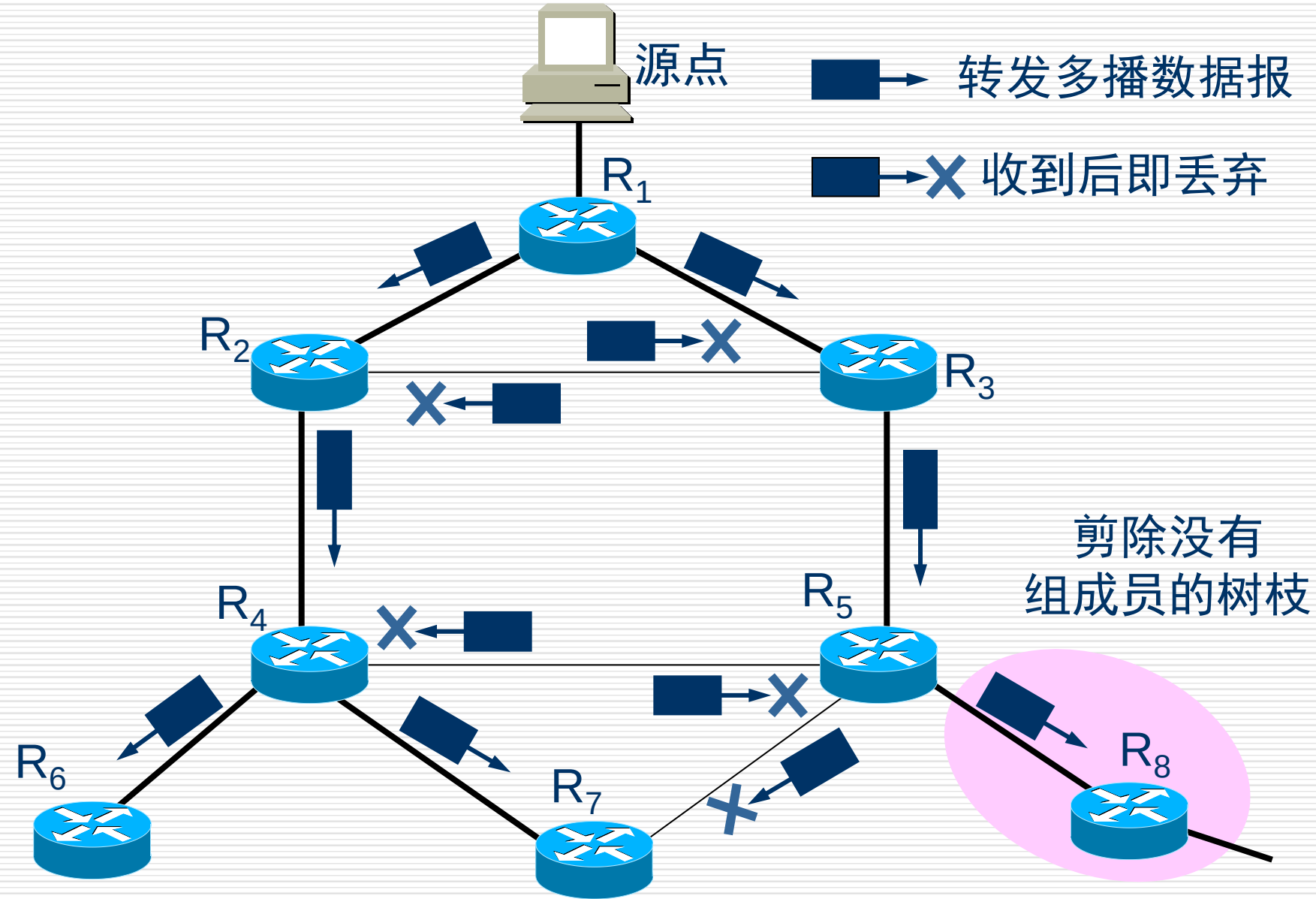
转发多播数据报使用的方法
(1) 洪泛与剪除

- 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。

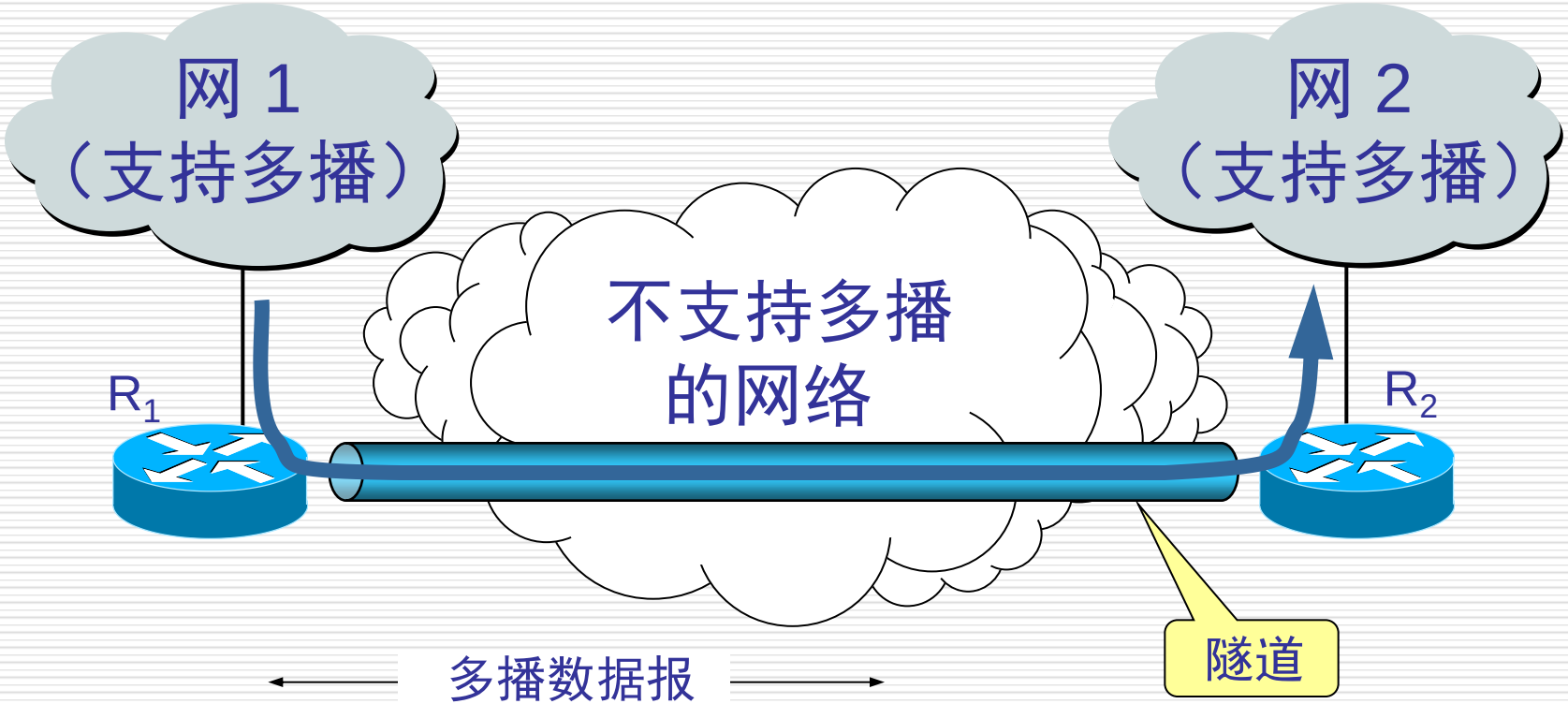
RPB 的要点

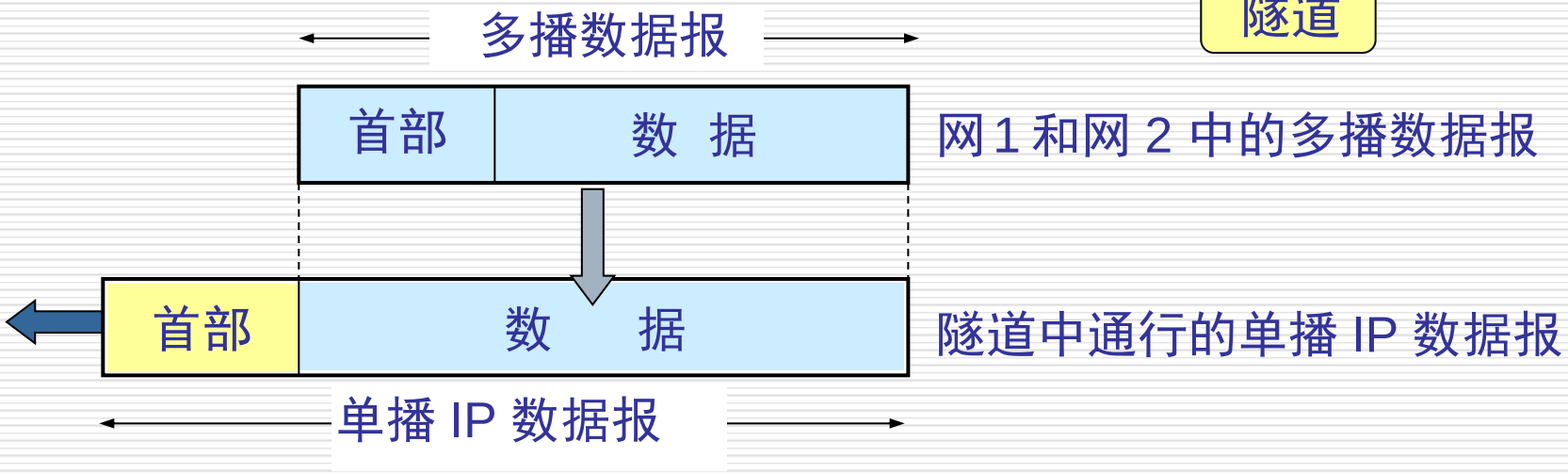
- 路由器收到多播数据报时，先检查是否从源点经最短路径传送来的。
- 若是，就向所有其他方向转发刚才收到的多播数据报（但进入的方向除外），否则就丢弃而不转发。
- 如果存在几条同样长度的最短路径），那么只能选择一条最短路径，选择的准则就是看这几条最短路径中的相邻路由器谁的 IP 地址最小。

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)

5.7 虚拟专用网 VPN 和网络地址转换 NAT

5.7.1 虚拟专用网 VPN

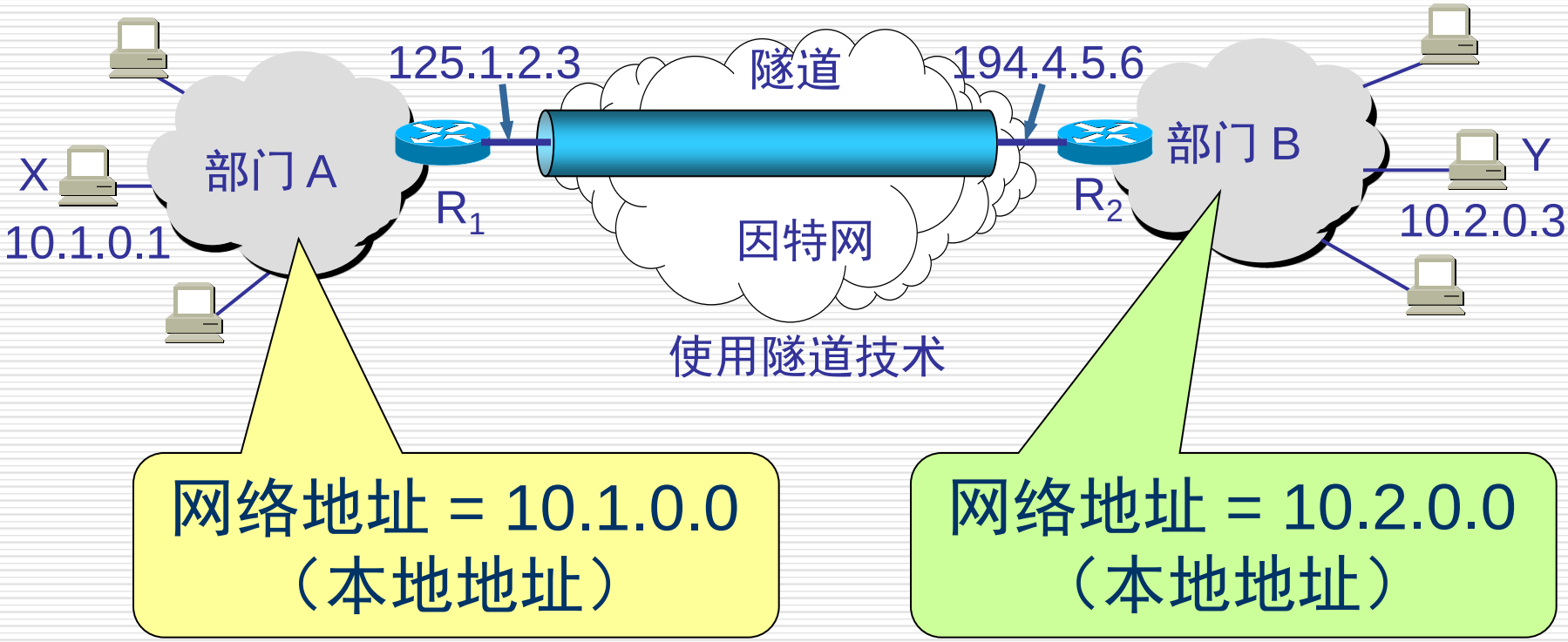
- 本地地址——仅在机构内部使用的 IP 地址，可以由本机构自行分配，而不需要向因特网的管理机构申请。
- 全球地址——全球唯一的IP地址，必须向因特网的管理机构申请。

RFC 1918 指明的专用地址(private address)

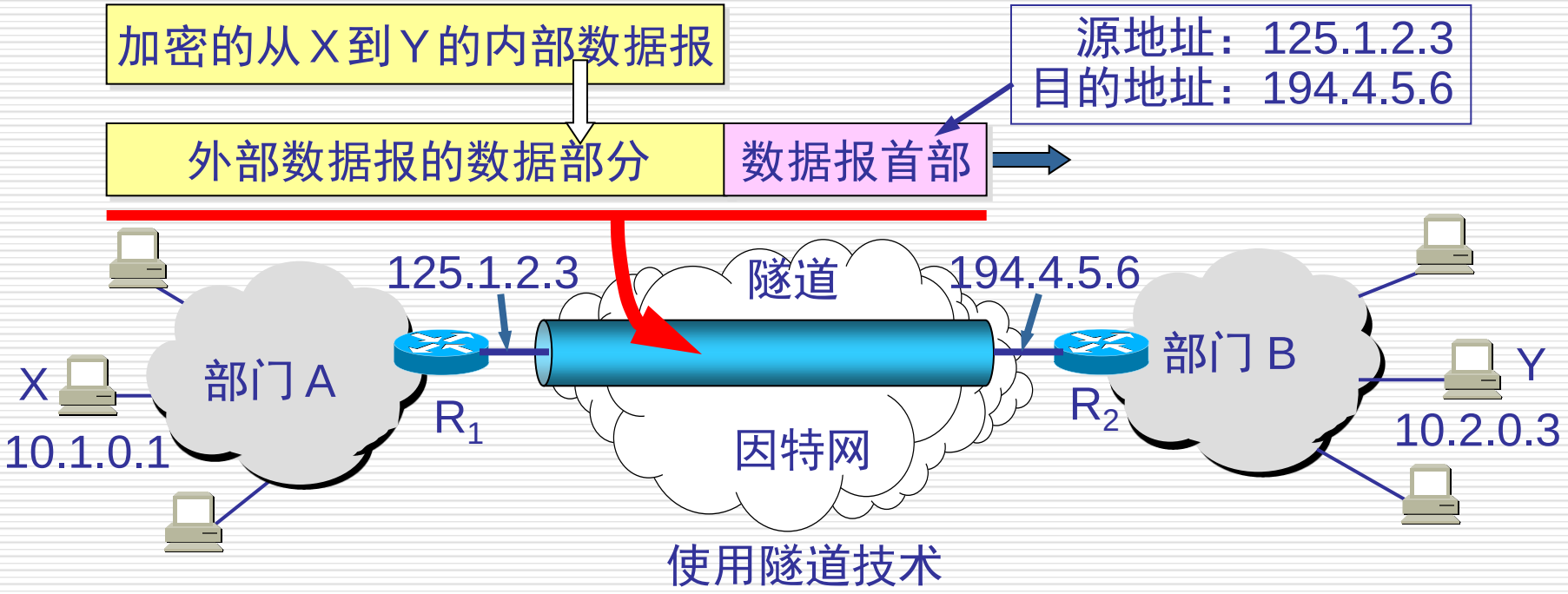
- 10.0.0.0 到 10.255.255.255
- 172.16.0.0 到 172.31.255.255
- 192.168.0.0 到 192.168.255.255
- 这些地址只能用于一个机构的内部通信，而不能用于和因特网上的主机通信。
- 专用地址只能用作本地地址而不能用作全球地址。在因特网中的所有路由器对目的地址是专用地址的数据报一律不进行转发。

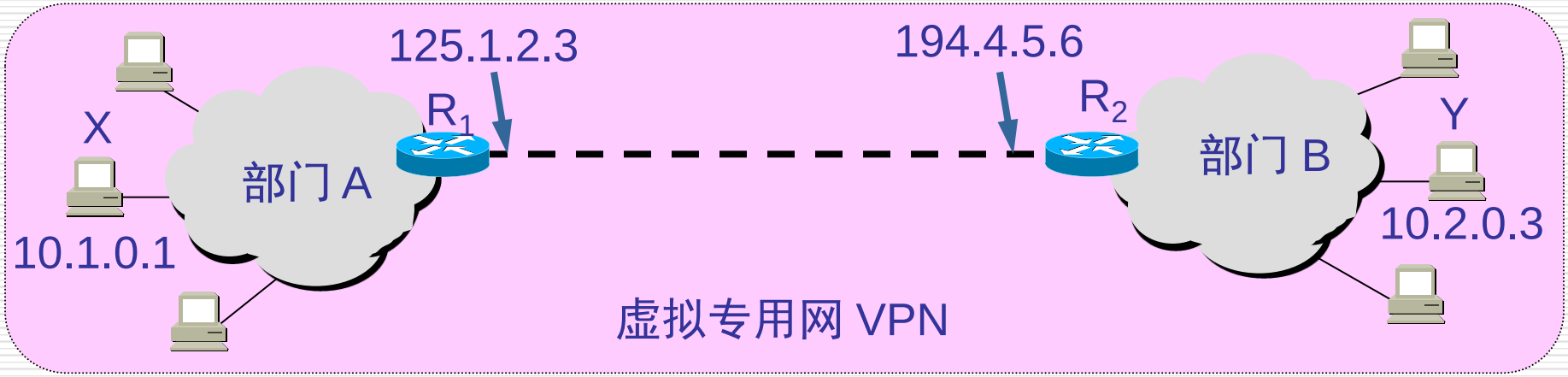
用隧道技术实现虚拟专用网

本地地址　　全球地址　　本地地址



用隧道技术实现虚拟专用网

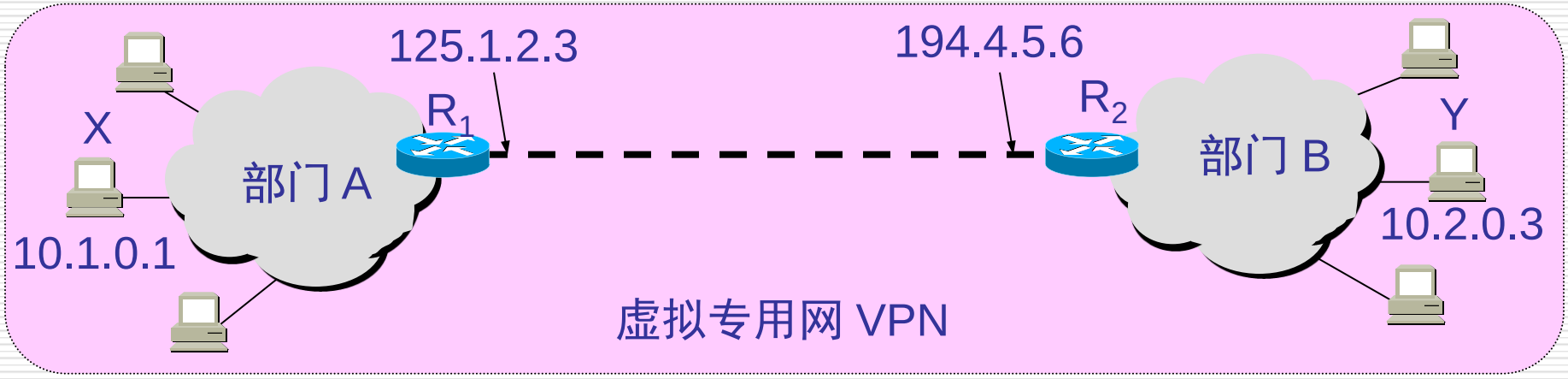




内联网 intranet 和外联网 extranet

（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网络。

5.7.2 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。