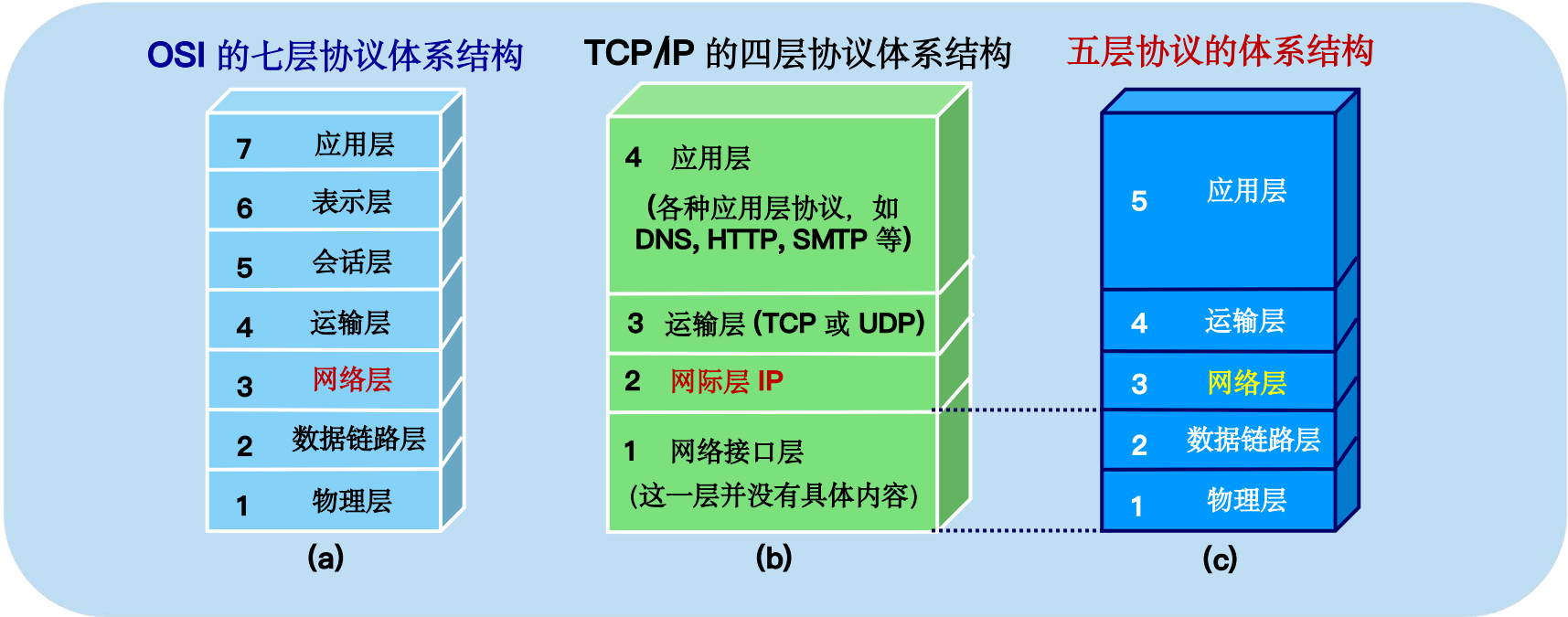
第4章

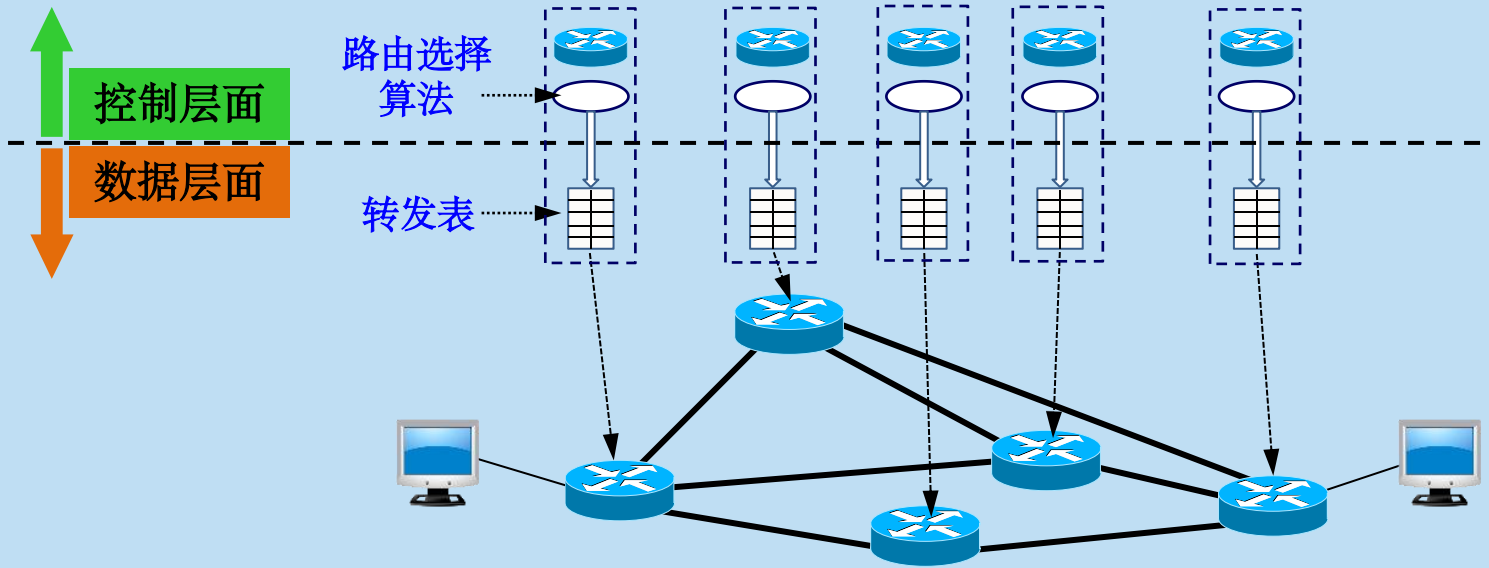
网络层

计算机网络体系结构


OSI 的七层协议体系结构
7 应用层
6 表示层
5 会话层
4 运输层
3 网络层
2 数据链路层
1 物理层
(a)
TCP/IP 的四层协议体系结构
4 应用层
(各种应用层协议，如 DNS, HTTP, SMTP 等)
3 运输层 (TCP 或 UDP)
2 网际层 IP
1 网络接口层
(这一层并没有具体内容)
(b)
五层协议的体系结构
5 应用层
4 运输层
3 网络层
2 数据链路层
1 物理层
(c)

4.6 互连网的路由选择协议

4.6.1 有关路由选择协议的几个基本概念



路由选择协议属于网络层控制层面的内容

路由算法分类（自适应）

静态路由选择策略

- 非自适应路由选择；
- 不能及时适应网络状态的变化；
- 简单，开销较小。

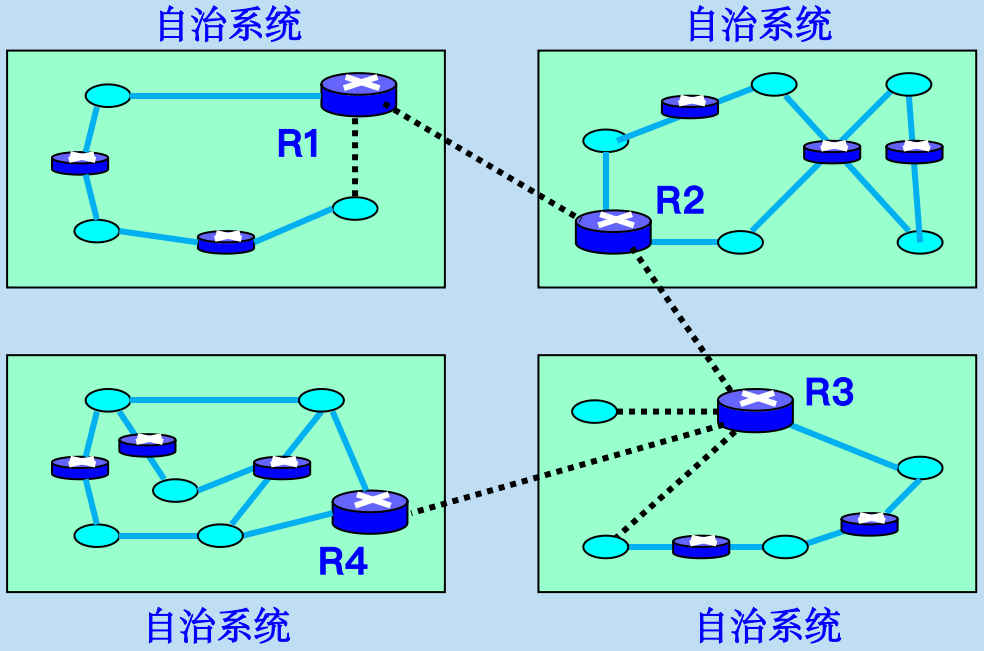
动态路由选择策略

- 自适应路由选择；
- 能较好地适应网络状态的变化；
- 实现较为复杂，开销较大。

2. 分层次的路由选择协议

- 互联网：
  - 采用自适应的（即动态的）、分布式路由选择协议。
  - 把整个互联网划分为许多较小的自治系统 AS，采用分层次的路由选择协议。
- 分为 2 个层次：
  - 自治系统之间的路由选择 或 域间路由选择 (interdomain routing)；
  - 自治系统内部的路由选择 或 域内路由选择 (intradomain routing)；

自治系统 AS (Autonomous System)



自治系统 AS：

是在单一技术管理下的许多网络、IP地址以及路由器，而这些路由器使用一种自治系统内部的路由选择协议和共同的度量。每一个 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。

2 大类路由选择协议

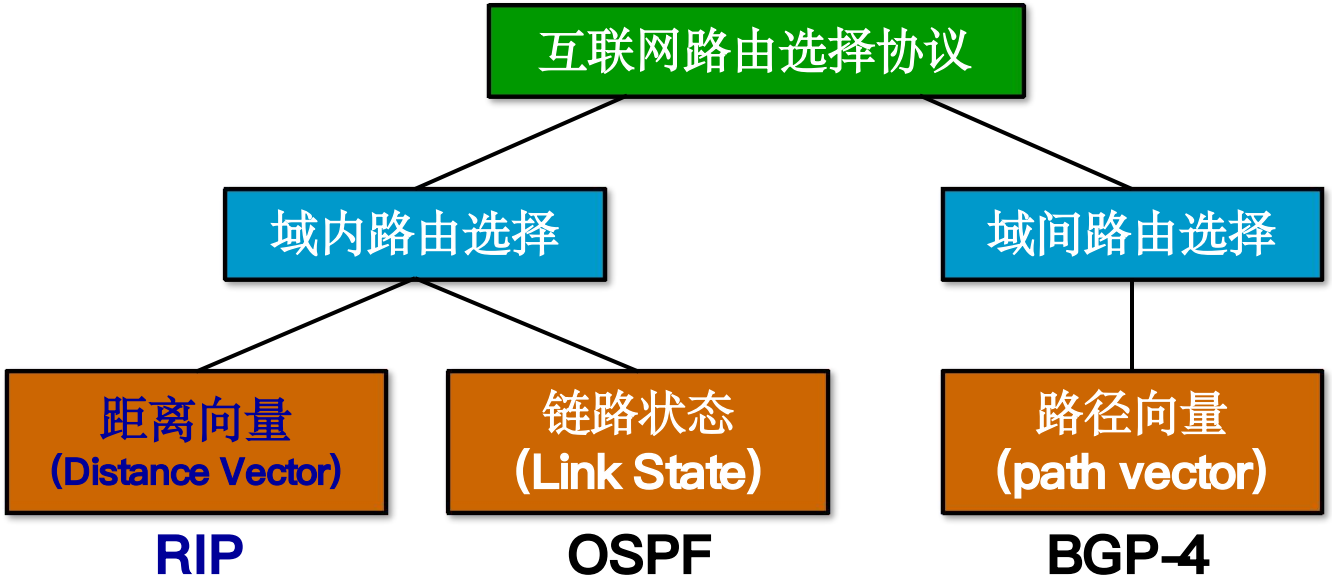
内部网关协议 IGP

- Interior Gateway Protocol
- 在一个自治系统内部使用的路由选择协议
- 常用：RIP，OSPF

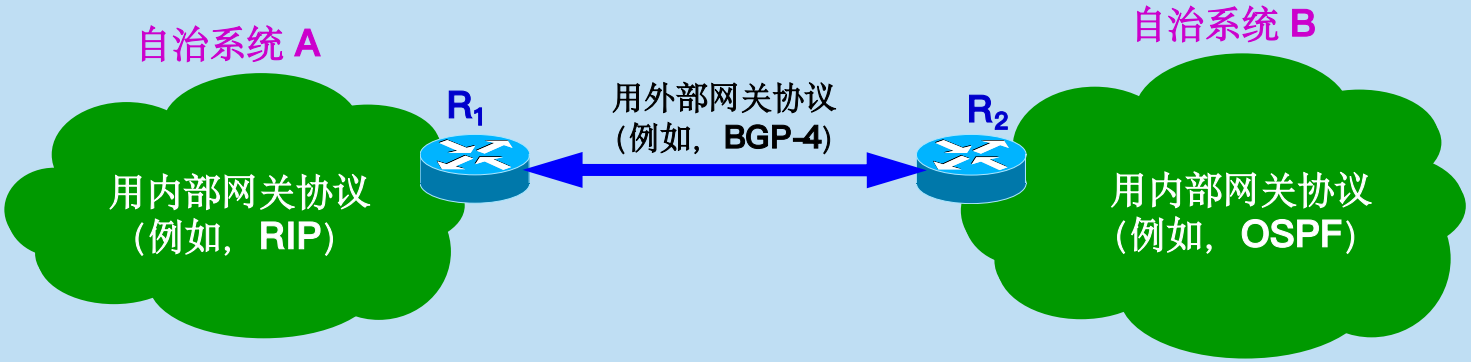
外部网关协议 EGP

- External Gateway Protocol
- 在不同自治系统之间进行路由选择时使用的协议
- 使用最多：BGP-4

4.6.2 内部网关协议 RIP



自治系统和内部网关协议、外部网关协议



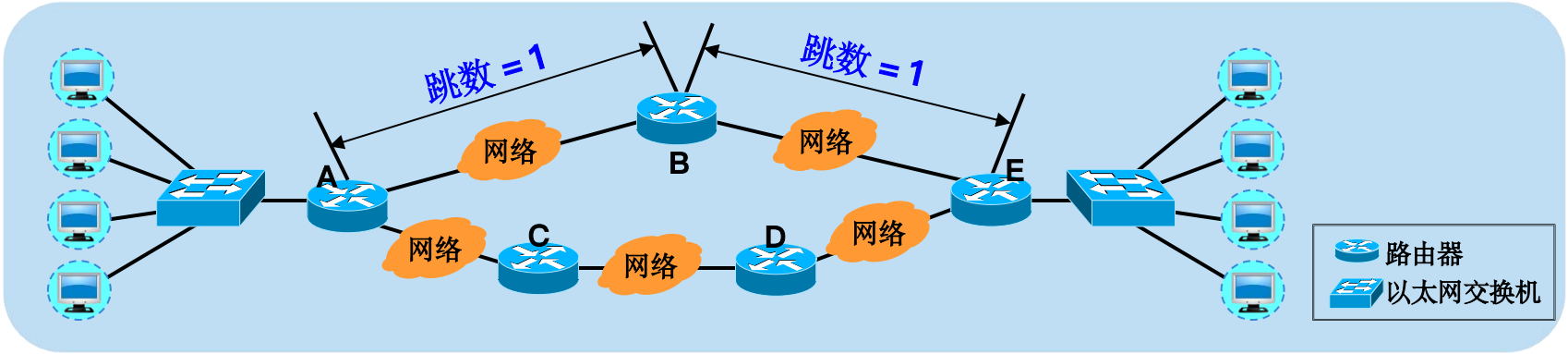
自治系统之间的路由选择也叫做域间路由选择（interdomain routing）。
自治系统内部的路由选择叫做域内路由选择（intradomain routing）。

1. 协议 RIP 的工作原理

- 路由信息协议 RIP (Routing Information Protocol) 是一种分布式的、基于距离向量的路由选择协议。
- 互联网的标准协议。
- 最大优点：简单。
- 要求网络中的每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。

RIP“距离”的定义

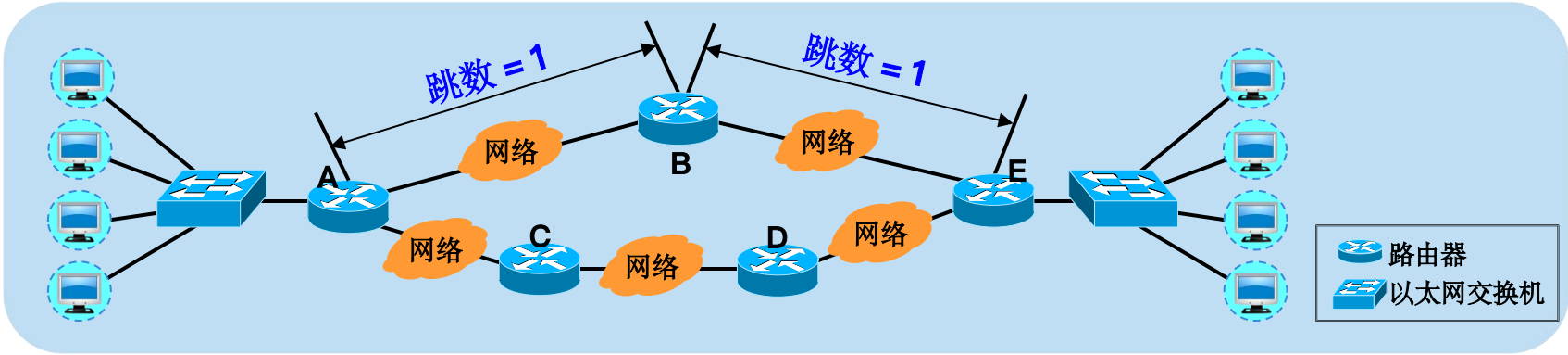
- 路由器到直接连接的网络的距离 = 1。
- 路由器到非直接连接的网络的距离 = 所经过的路由器数 + 1。
- RIP 协议中的“距离”也称为“跳数”(hop count)，每经过一个路由器，跳数就加 1。



路由 A-B-E 的距离 = 2，路由 A-C-D-E 的距离 = 3。

RIP“距离”的定义

- 好路由 = “距离短”的路由。最佳路由 = “距离最短”的路由。
- 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。
- RIP 不能在两个网络之间同时使用多条路由，只选择距离最短”的路由。



路由 A-B-E 的距离 = 2，路由 A-C-D-E 的距离 = 3。
最佳路由为 A-B-E。

RIP 协议的三个特点

1. 仅和相邻路由器交换信息。
2. 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。
3. 按固定时间间隔交换路由信息，例如，每隔 30 秒。当网络拓扑发生变化时，路由器也及时向相邻路由器通告拓扑变化后的路由信息。

路由表的建立

- 路由器在刚刚开始工作时，路由表是空的。
- 然后，得到直接连接的网络的距离（此距离定义为 1）。
- 之后，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。“收敛”就是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。

路由表主要信息和更新规则

- 路由表主要信息：

目的网络	距离（最短）	下一跳地址

- 路由表更新规则：

  使用距离向量算法找出到达每个目的网络的最短距离。

2. 距离向量算法

对每个相邻路由器（假设其地址为 X）发送过来的 RIP 报文，路由器：

(1) 修改 RIP 报文中的所有项目（即路由）：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：

若路由表中没有目的网络N，则把该项目添加到路由表中。否则

若路由表中网络 N 的下一跳路由器为 X，则用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，则用收到项目更新原路由表中的项目。否则

什么也不做。

(3) 若 3 分钟还未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（表示不可达）。

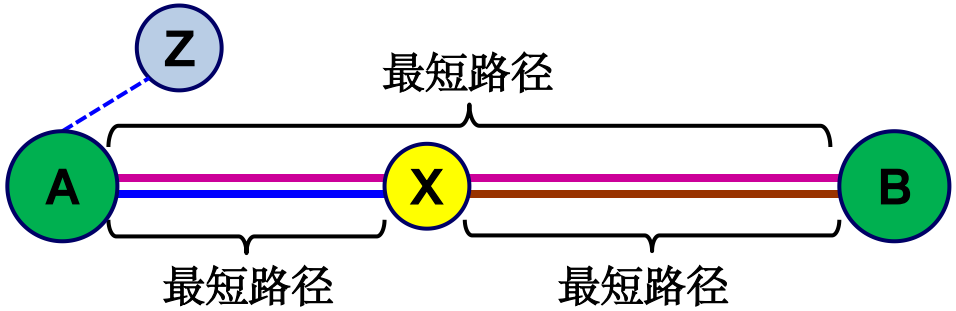
(4) 返回。

2. 距离向量算法

- 算法基础：Bellman-Ford 算法（或 Ford-Fulkerson 算法）。
- 算法要点：

  设 X 是结点 A 到 B 的最短路径上的一个结点。

  若把路径 A→B 拆成两段路径 A→X 和 X→B，则每一段路径 A→X 和 X→B 也都分别是结点 A 到 X 和结点 X 到 B 的最短路径。



【例】已知路由器 R_6 有表 4-8(a) 所示的路由表。现在收到相邻路由器 R_4 发来的路由更新信息，如表 4-8(b) 所示。试更新路由器 R_6 的路由表。

表 4-8(a) 路由器 R_6 的路由表

目的网络	距离	下一跳路由器
Net2	3	R_4
Net3	4	R_5
…	…	…

表 4-8(b) R4 发来的路由更新信息

目的网络	距离	下一跳路由器
Net1	3	R_1
Net2	4	R_2
Net3	1	直接交付

① 距离+1，修改下一跳地址

表 4-8(c) 修改后的表 4-9(b)

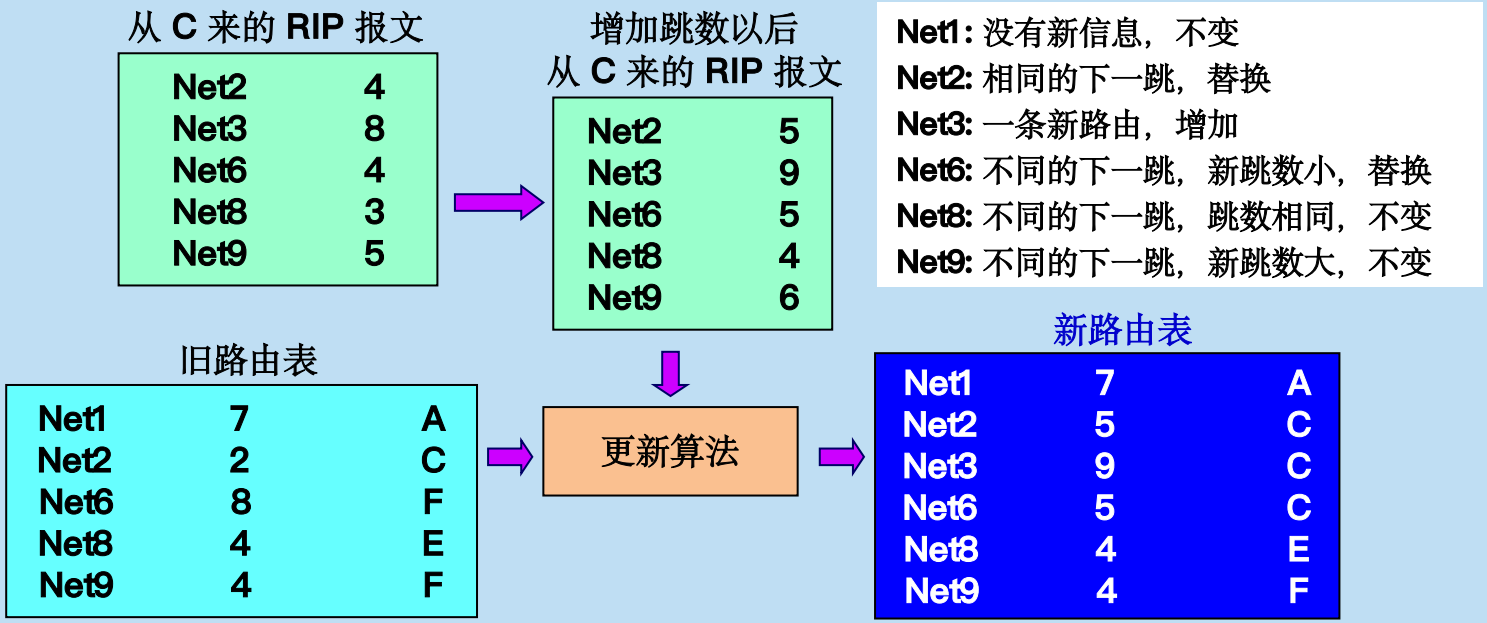
目的网络	距离	下一跳路由器
Net1	4	R_4
Net2	5	R_4
Net3	2	R_4

② 计算更新

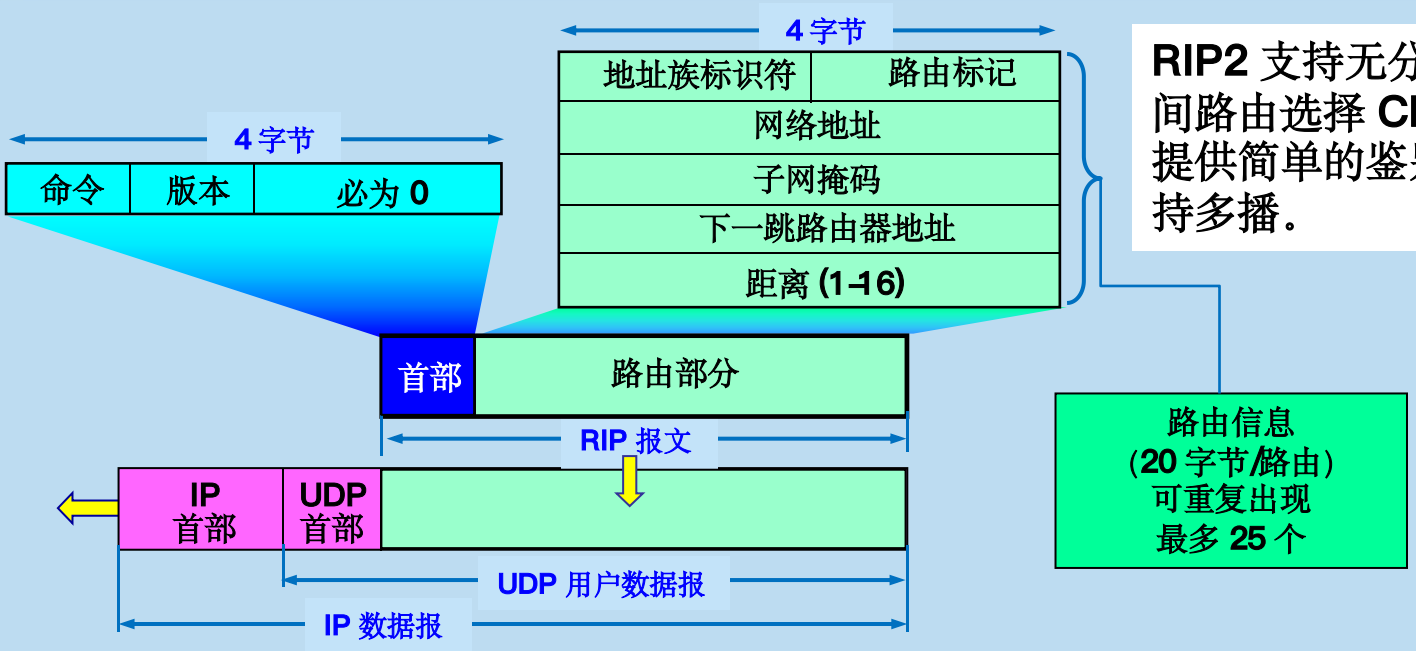
表 4-8(d) 路由器 R_6 更新后的路由表

目的网络	距离	下一跳路由器
Net1	4	R_4
Net2	5	R_4
Net3	2	R_4
…	…	…

【例】路由表更新。


从 C 来的 RIP 报文
Net2 4
Net3 8
Net6 4
Net8 3
Net9 5
增加跳数以后
从 C 来的 RIP 报文
Net2 5
Net3 9
Net6 5
Net8 4
Net9 6
Net1：没有新信息，不变
Net2：相同的下一跳，替换
Net3：一条新路由，增加
Net6：不同的下一跳，新跳数小，替换
Net8：不同的下一跳，跳数相同，不变
Net9：不同的下一跳，新跳数大，不变
旧路由表
Net1 7 A
Net2 2 C
Net6 8 F
Net8 4 E
Net9 4 F
更新算法
新路由表
Net1 7 A
Net2 5 C
Net3 9 C
Net6 5 C
Net8 4 E
Net9 4 F

RIP2 报文



RIP2 的报文用使用 UDP 传送（使用 UDP 端口 520）。

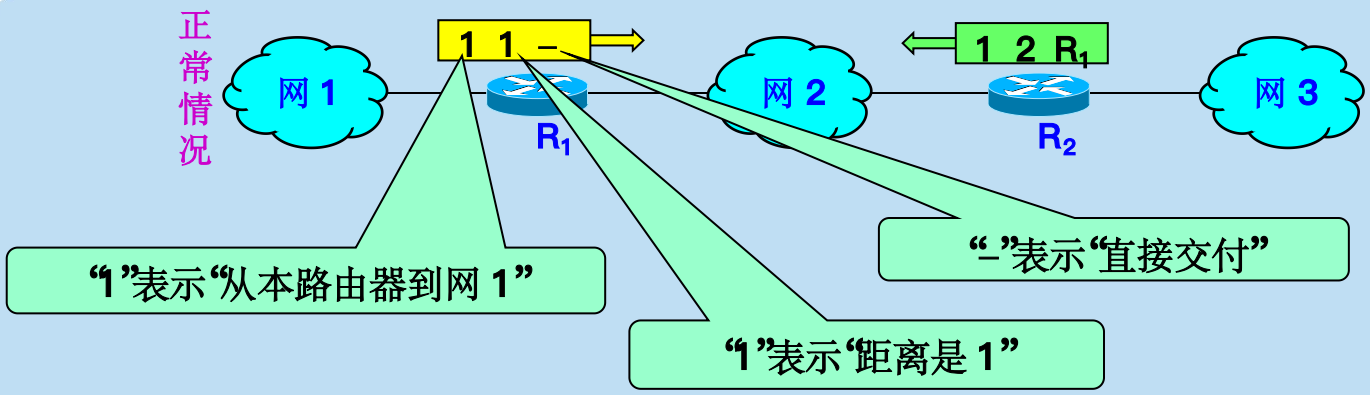
RIP2 报文

- 组成：首部和路由 2 个部分。
- 路由部分：由若干个路由信息组成。每个路由信息共 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。
  - 后面为具体路由，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。
- 一个 RIP 报文最多可包括 25 个路由，因而 RIP 报文的最大长度是 4+20 x25=504 字节。如超过，必须再用一个 RIP 报文来传送。
- RIP2 具有简单的鉴别功能。

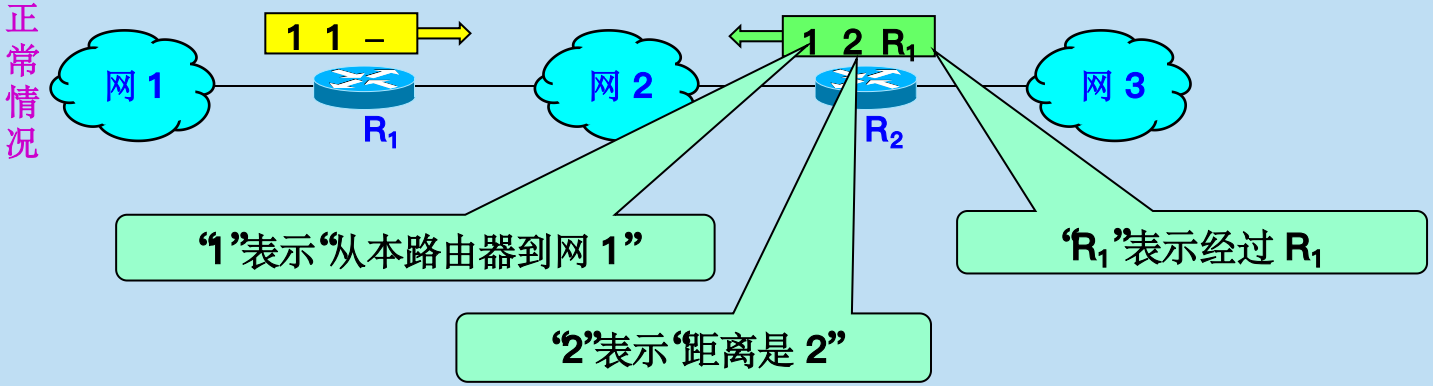
3. 坏消息传播得慢

- RIP 协议特点：好消息传播得快，坏消息传播得慢。
- 问题：坏消息传播得慢（慢收敛）。

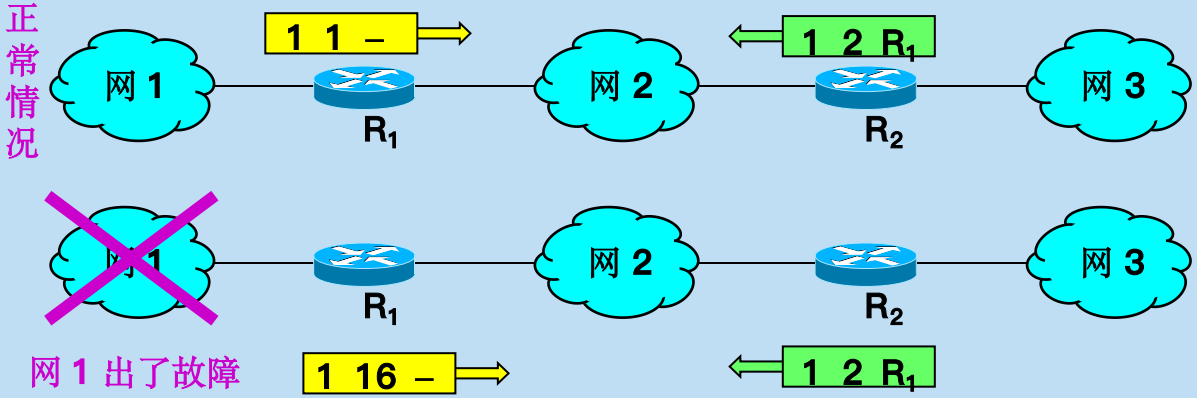
当网络出现故障时，要经过比较长的时间才能将此信息（坏消息）传送到所有的路由器。



R_1 说：“我到网 1 的距离是 1，是直接交付。”

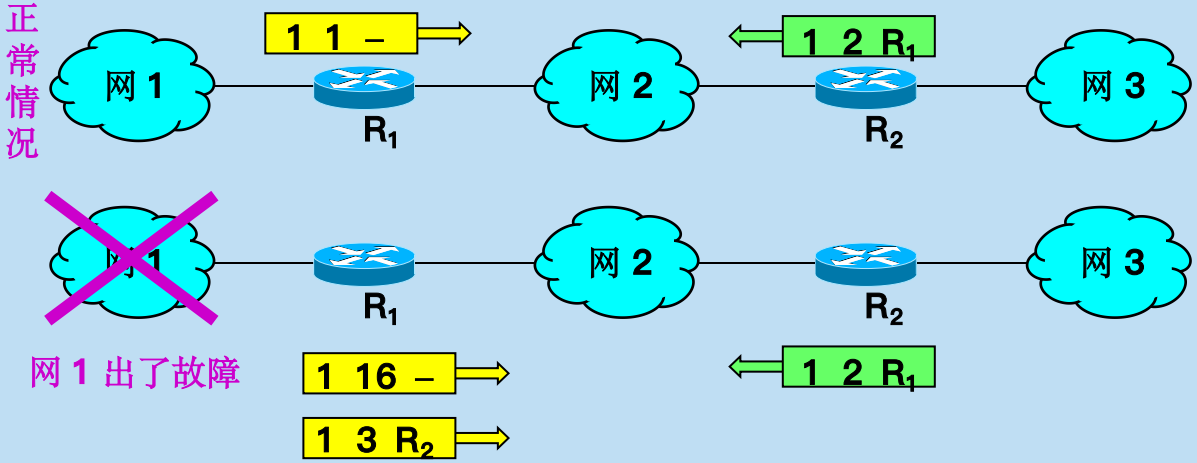


R_2 说：“我到网 1 的距离是 2，是经过 R_1。”

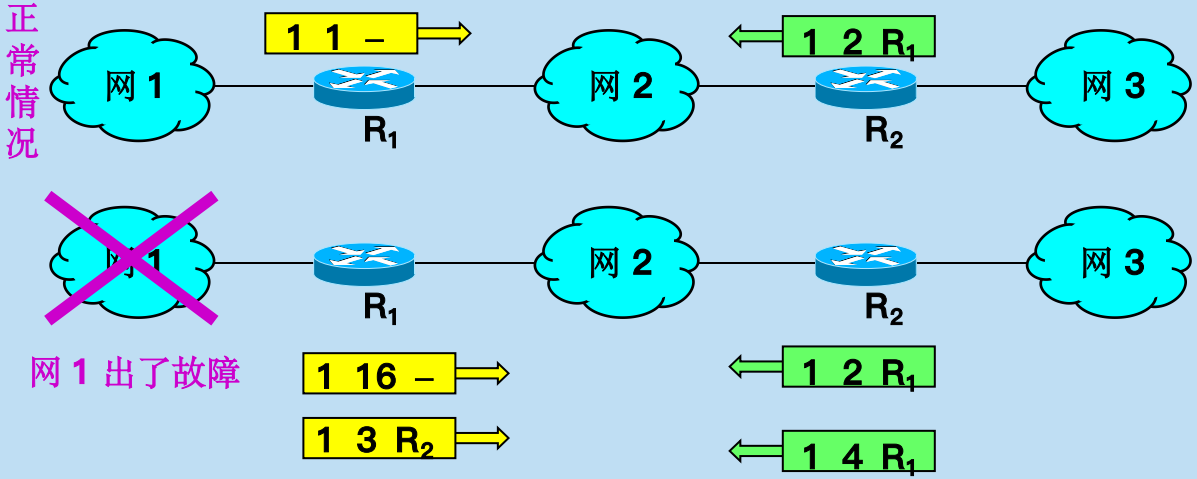


R_1 说：“我到网 1 的距离是 16 （表示无法到达），是直接交付。”

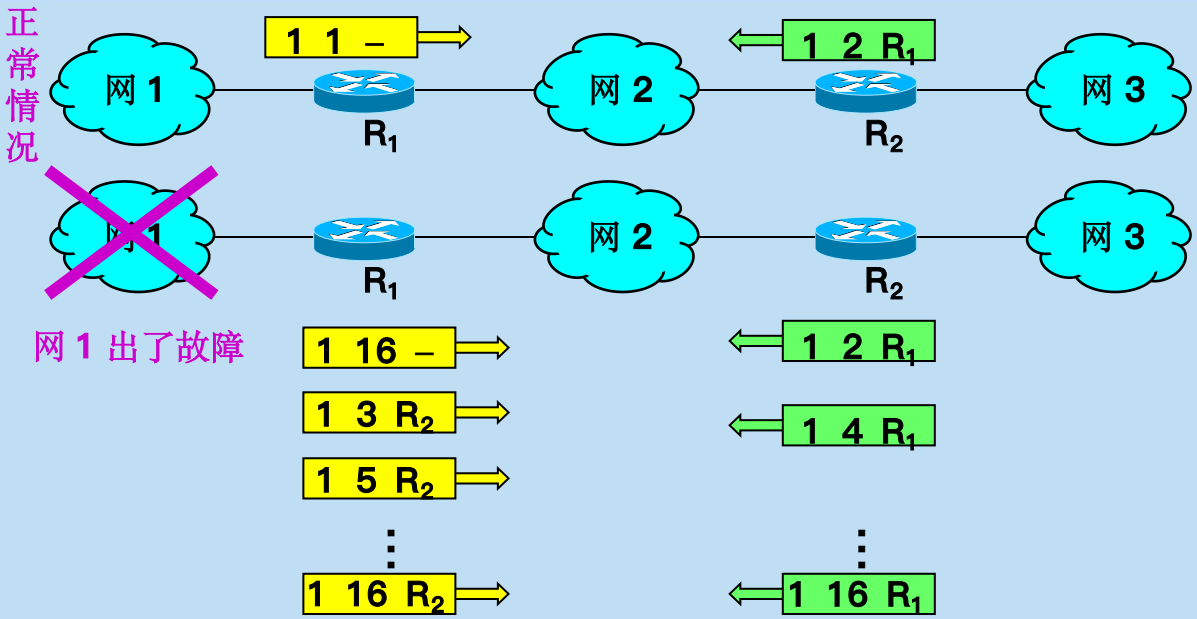
但 R_2 在收到 R_1 的更新报文之前，还发送原来的报文，
因为这时 R_2 并不知道 R_1 出了故障。



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”。



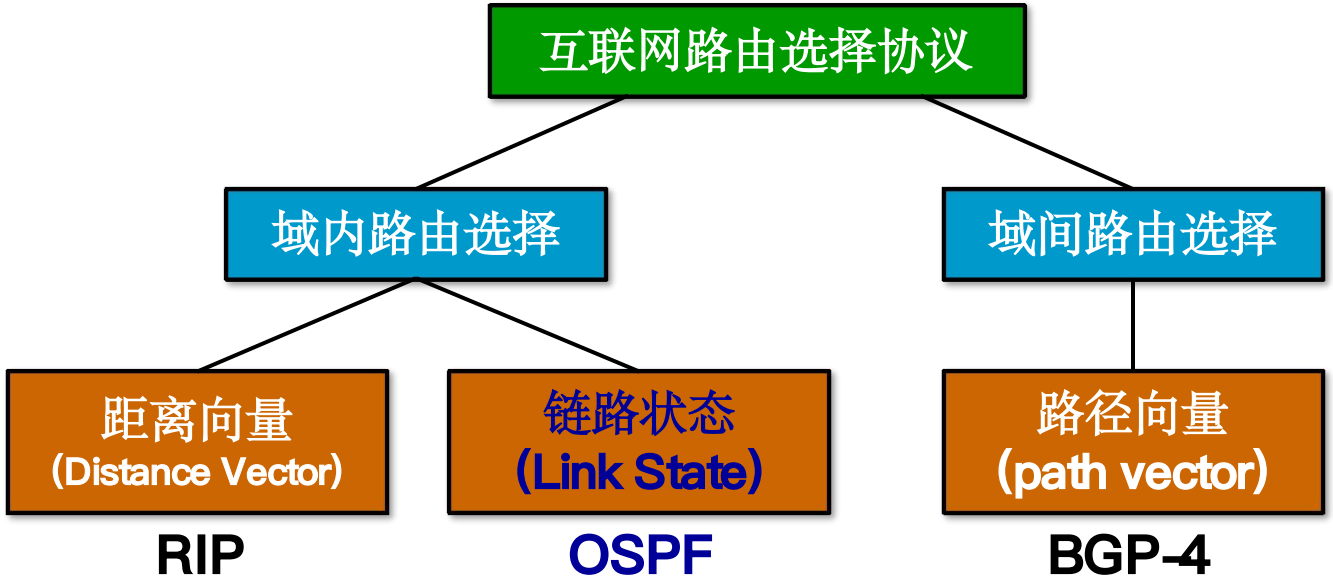
这样不断更新下去，直到 R_1 和 R_2 到网 1 的距离都增大到 16 时，R_1 和 R_2 才知道网 1 是不可达的。

这就是好消息传播得快，而坏消息传播得慢。这是 RIP 的一个主要缺点。

RIP 协议的优缺点

- 优点：
  1. 实现简单，开销较小。
- 缺点：
  1. 网络规模有限。最大距离为 15（16 表示不可达）。
  2. 交换的路由信息为完整路由表，开销较大。
  3. 坏消息传播得慢，收敛时间过长。

4.6.3 内部网关协议 OSPF



4.6.3 内部网关协议 OSPF

- 开放最短路径优先 OSPF (Open Shortest Path First)是为克服 RIP 的缺点在 1989 年开发出来的。
- 原理很简单，但实现很复杂。
- 使用了 Dijkstra 提出的最短路径算法 SPF。
- 采用分布式的链路状态协议 (link state protocol)。
- 现在使用 OSPFv2。

三个主要特点

- 采用洪泛法 (flooding)，向本自治系统中所有路由器发送信息。
- 发送的信息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：说明本路由器都和哪些路由器相邻，以及该链路的度量 (metric)。
- 当链路状态发生变化或每隔一段时间（如30分钟），路由器才用洪泛法向所有路由器发送此信息。

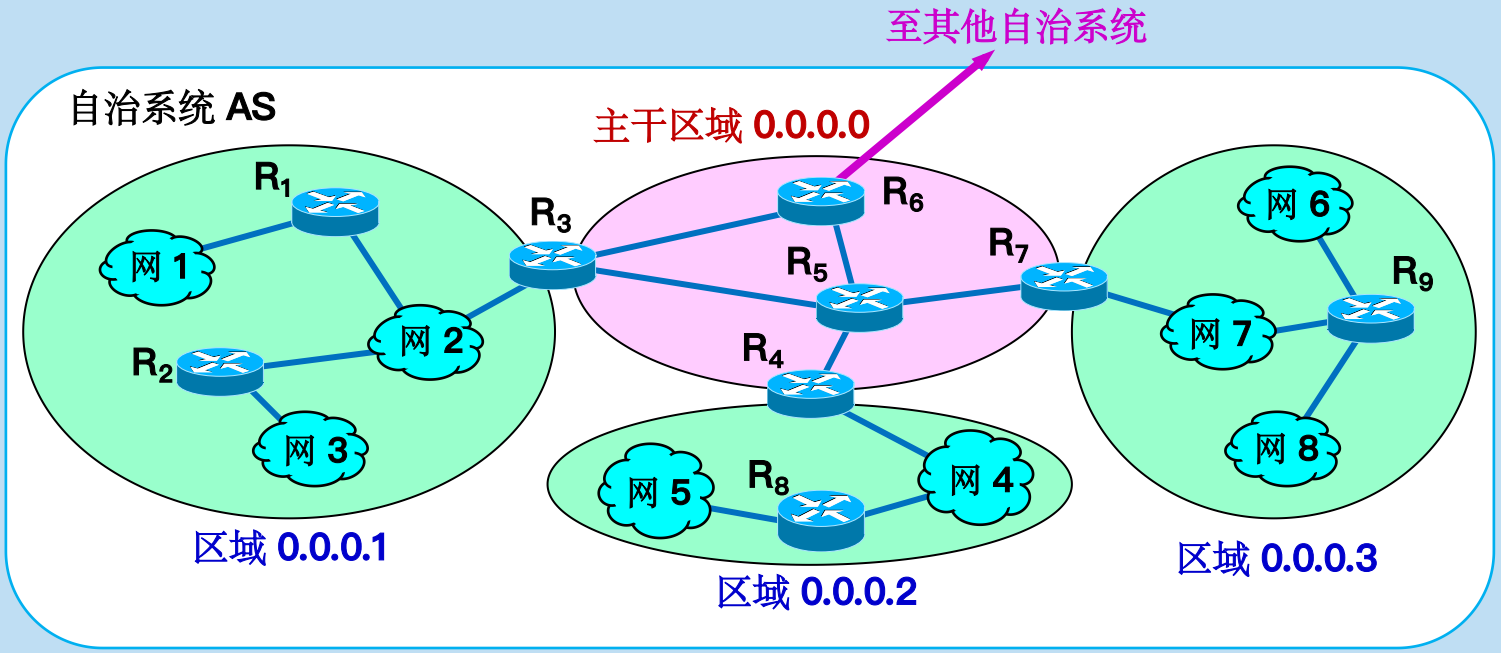
链路状态数据库 (link-state database)

- 每个路由器最终都能建立。
- 全网的拓扑结构图。
- 在全网范围内是一致的（这称为链路状态数据库的同步）。
- 每个路由器使用链路状态数据库中的数据构造自己的路由表（例如，使用Dijkstra的最短路径路由算法）。

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。

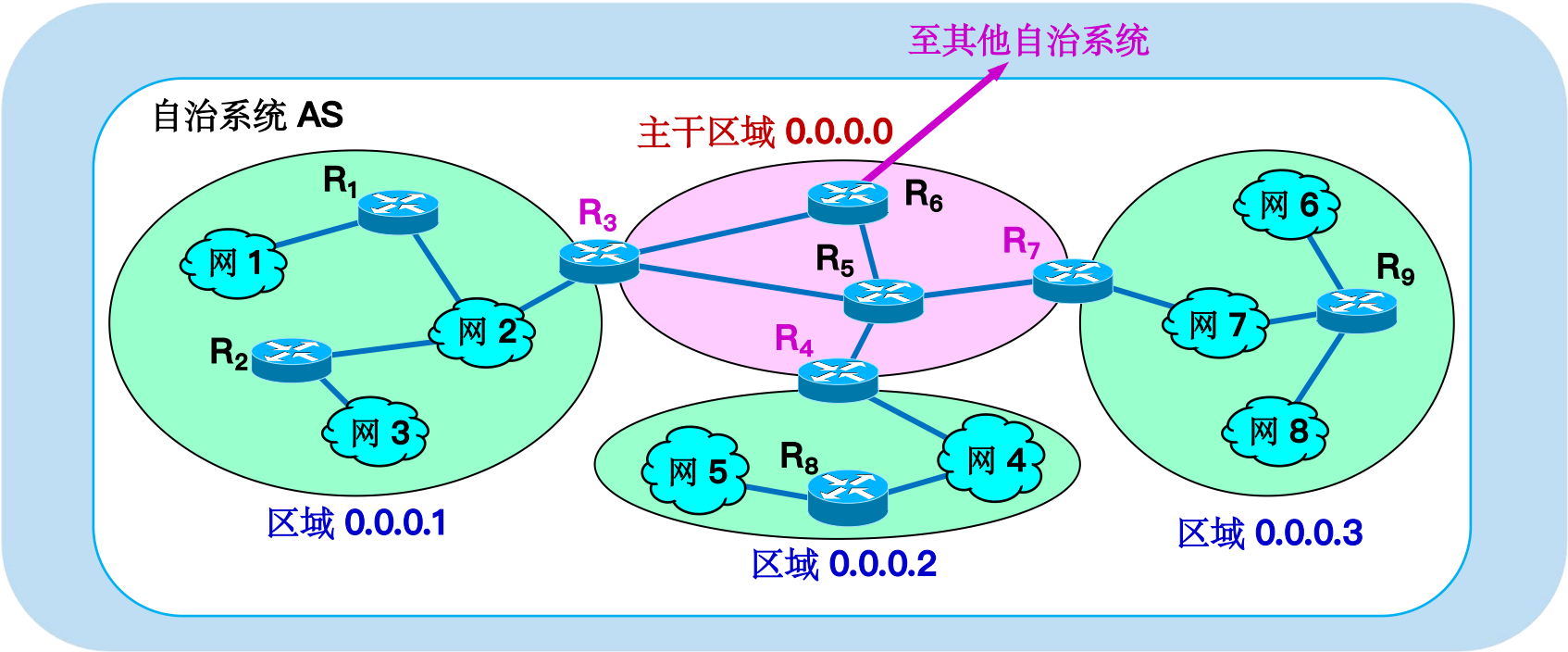
重要优点：OSPF 更新过程收敛速度快。

OSPF 将自治系统划分为两种不同的区域 (area)

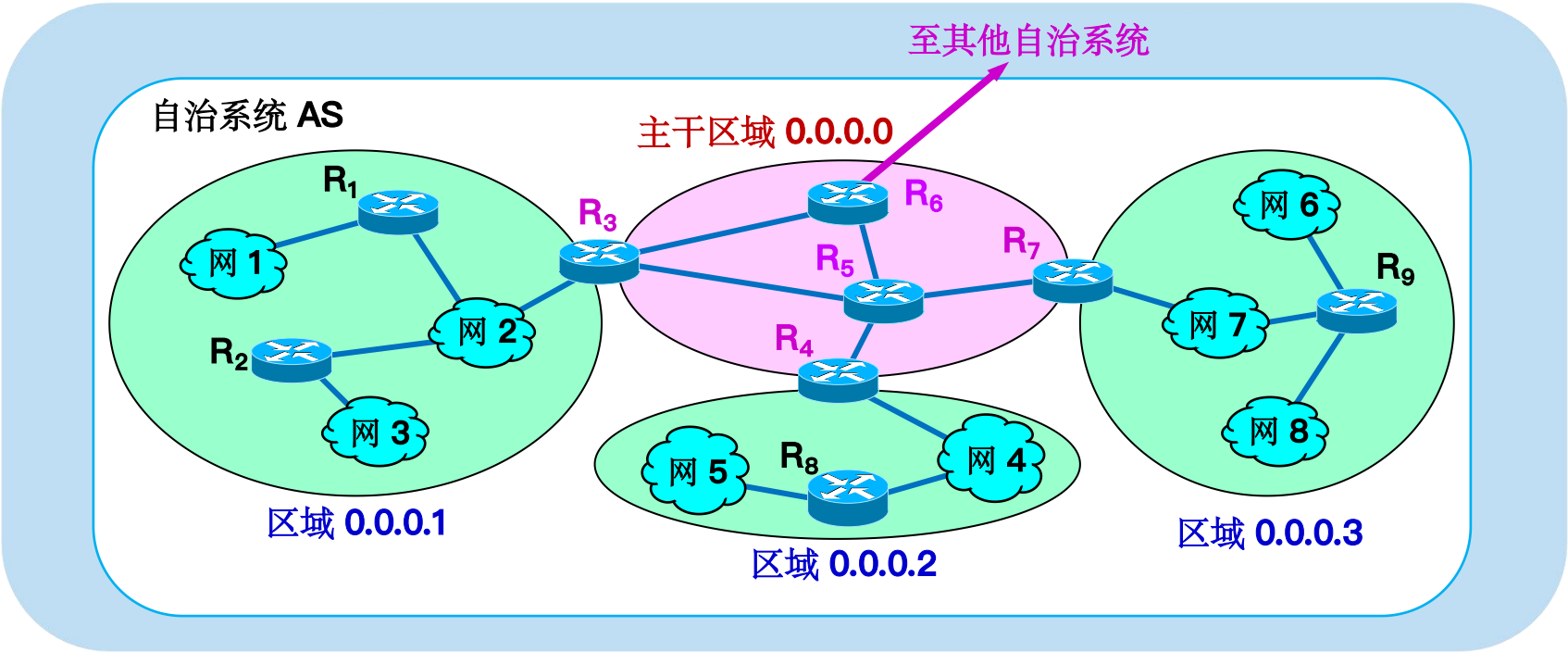


主干区域 (backbone area) 标识符= 0.0.0.0， 作用=用来连通其他下层区域。

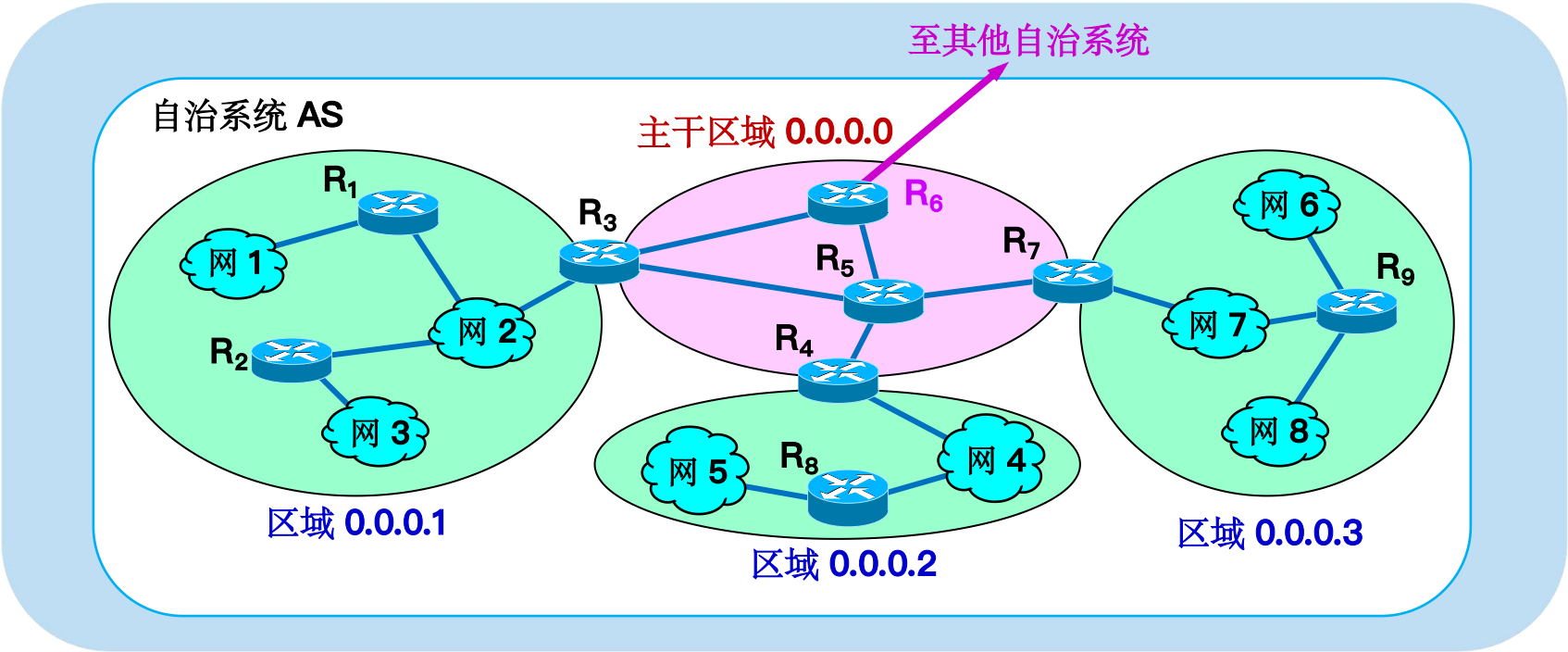
OSPF 中的路由器：区域边界路由器 ABR (area border router)



OSPF 中的路由器：主干路由器 BR (backbone router)



OSPF 中的路由器：自治系统边界路由器 ASBR (AS border router)



划分区域优点和缺点

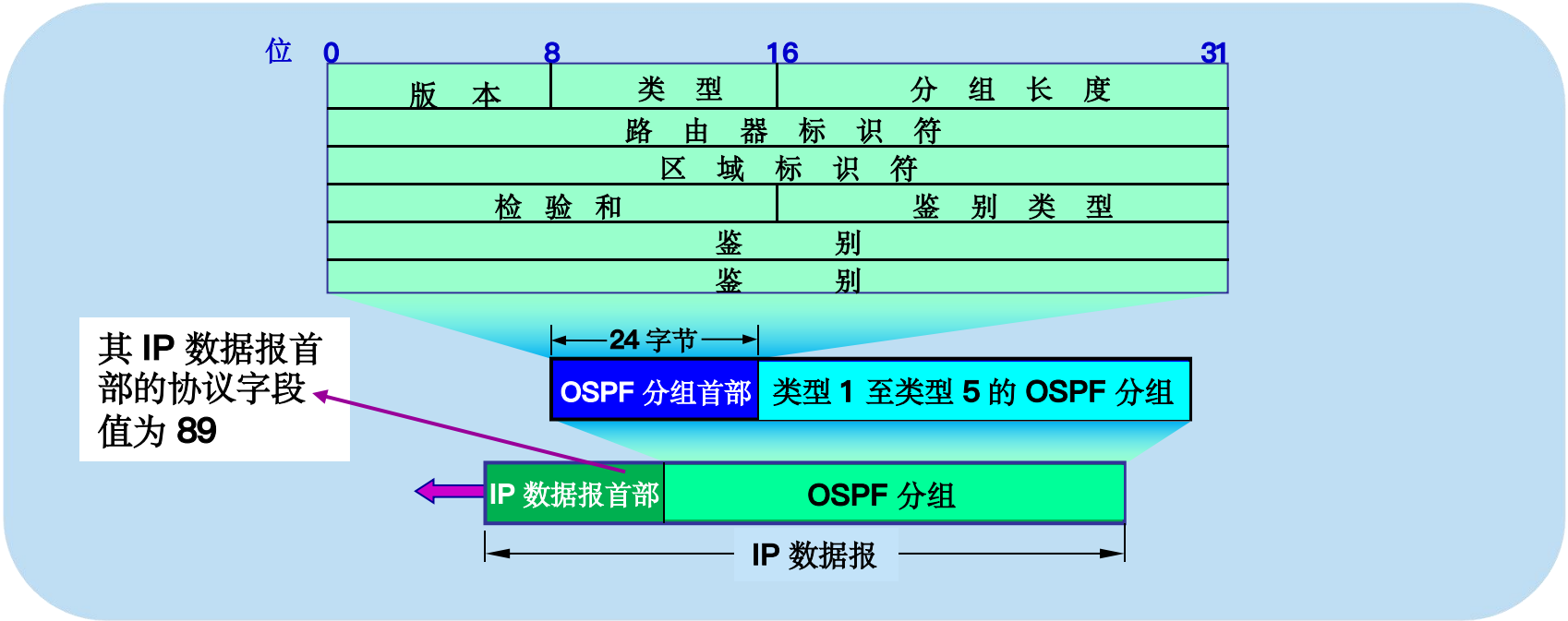
- 优点：
  - 减少了整个网络上的通信量。
  - 减少了需要维护的状态数量。
- 缺点：
  - 交换信息的种类增多了。
  - 使 OSPF 协议更加复杂了。

分层次划分区域的好处：
使每一个区域内部交换路由信息的通信量大大减小，因而使 OSPF 协议能够用于规模很大的自治系统中。

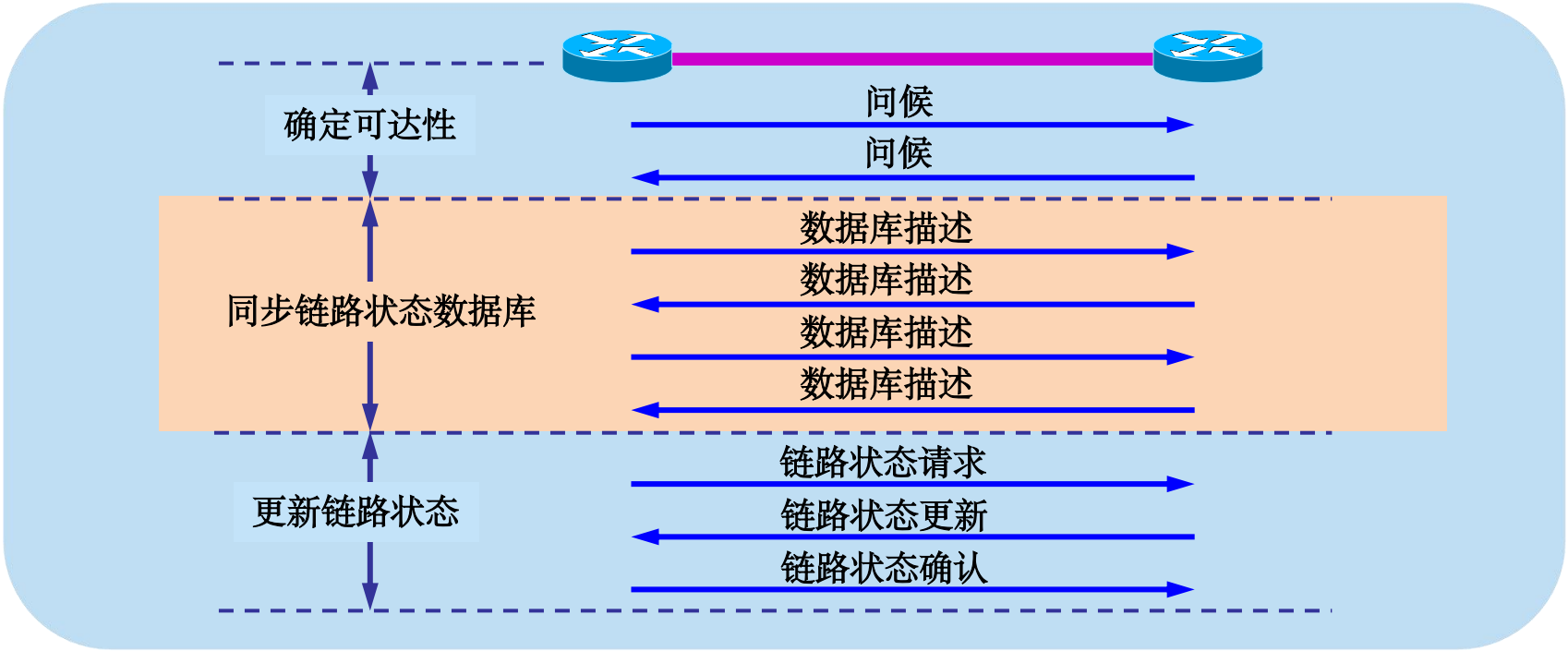
2. OSPF 的五种分组类型

1. 问候 (Hello) 分组。
2. 数据库描述 (Database Description) 分组。
3. 链路状态请求 (Link State Request) 分组。
4. 链路状态更新 (Link State Update) 分组。
5. 链路状态确认 (Link State Acknowledgment)分组。

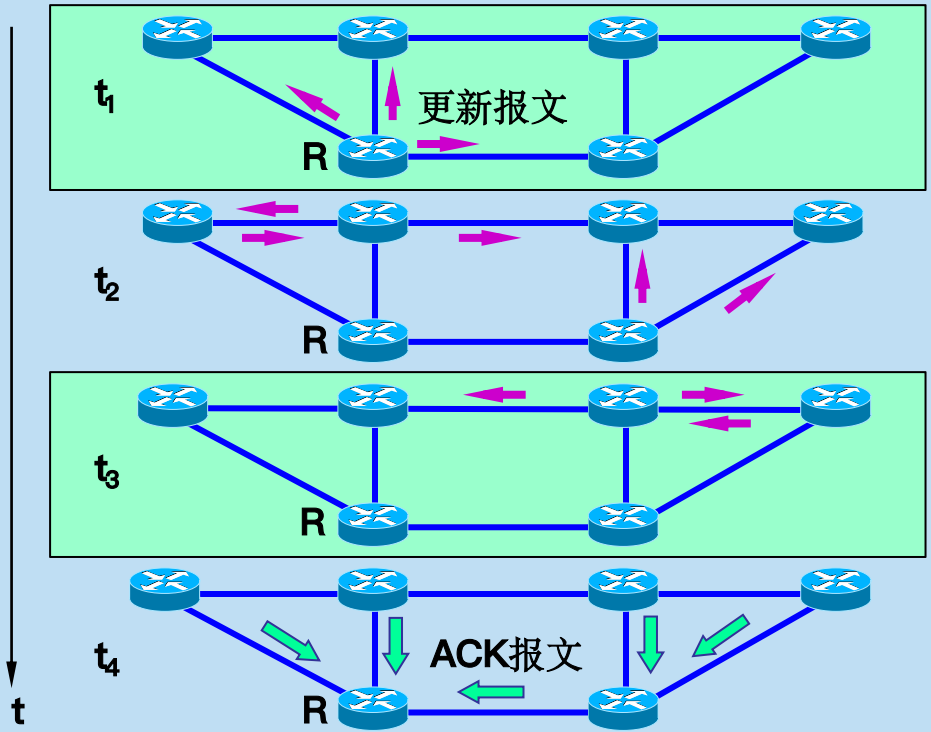
OSPF 分组用 IP 数据报传送



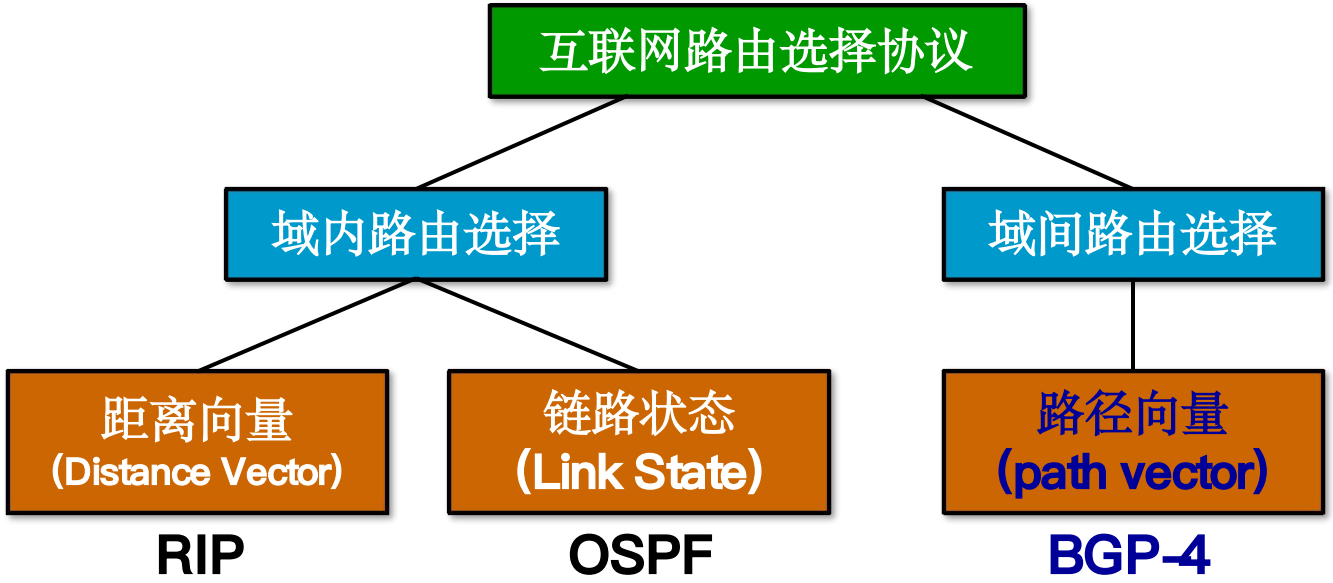
OSPF 工作过程


确定可达性
同步链路状态数据库
更新链路状态
问候
问候
数据库描述
数据库描述
数据库描述
数据库描述
链路状态请求
链路状态更新
链路状态确认

OSPF 使用
可靠的洪泛法
发送更新分组



4.6.4 外部网关协议 BGP



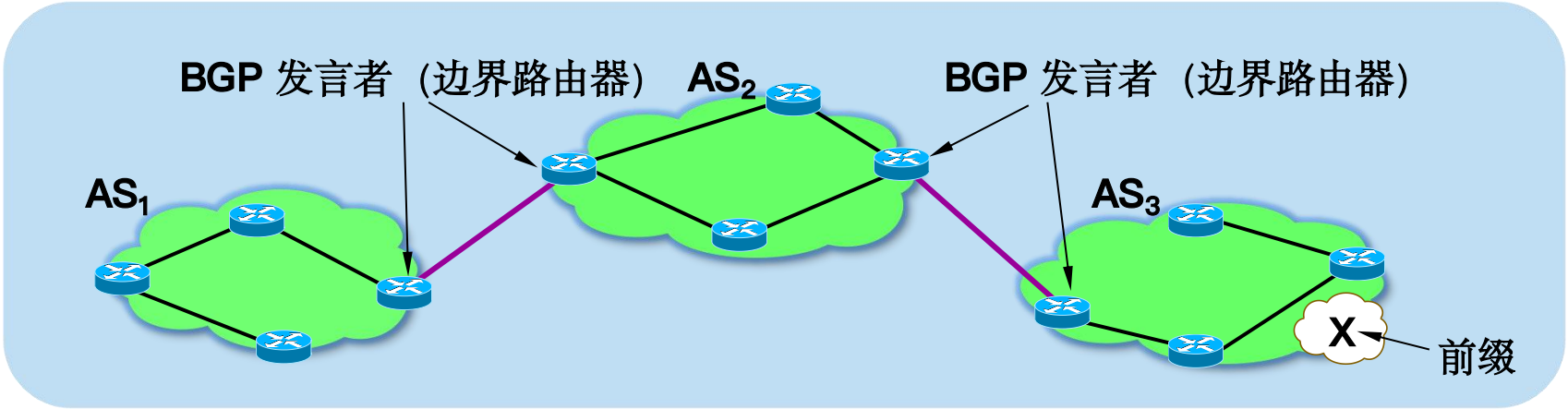
4.6.4 外部网关协议 BGP

- 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。

1. 协议 BGP 的主要特点

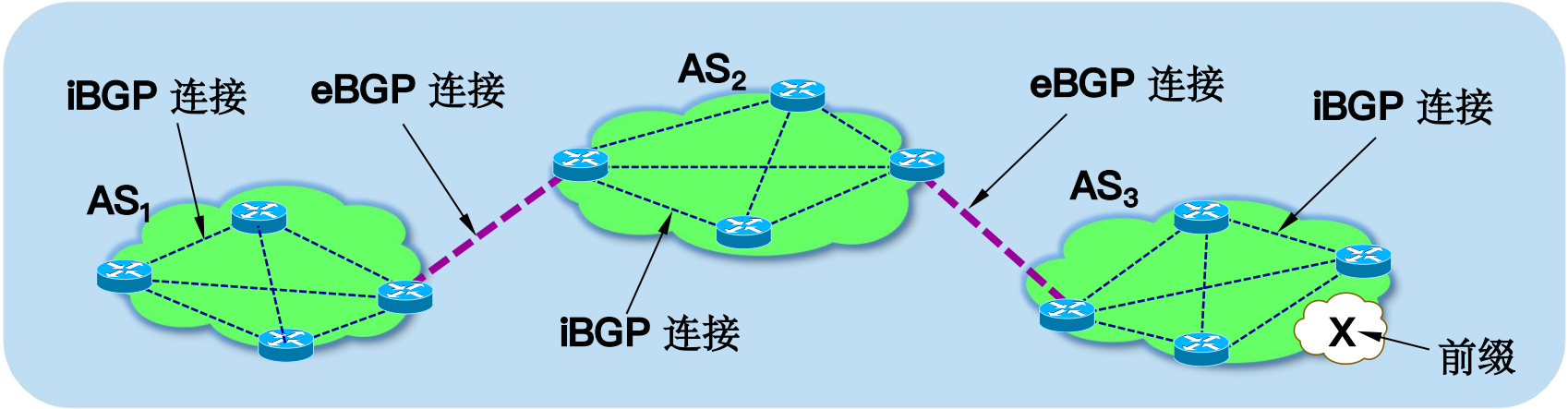
- 用于自治系统 AS 之间的路由选择。
- 只能是力求选择出一条能够到达目的网络且比较好的路由（不能兜圈子），而并非要计算出一条最佳路由。
  1. 互联网的规模太大，使得自治系统AS之间路由选择非常困难。
  2. 自治系统AS之间的路由选择必须考虑有关策略。
- 采用了路径向量 (path vector) 路由选择协议。

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



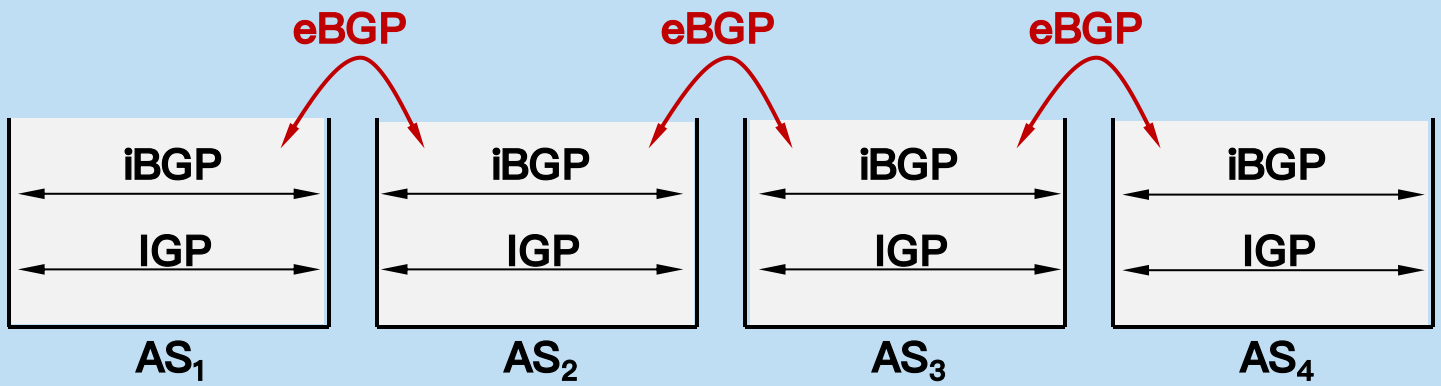
在 AS 之间， BGP 发言者在半永久性 TCP 连接（端口号为179）上建立 BGP 会话(session)。这种连接又称为 eBGP 连接。

在 AS 内部，任何相互通信的两个路由器之间必须有一个逻辑连接（也使用 TCP 连接）。AS 内部所有的路由器之间的通信是全连通的。这种连接常称为 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。

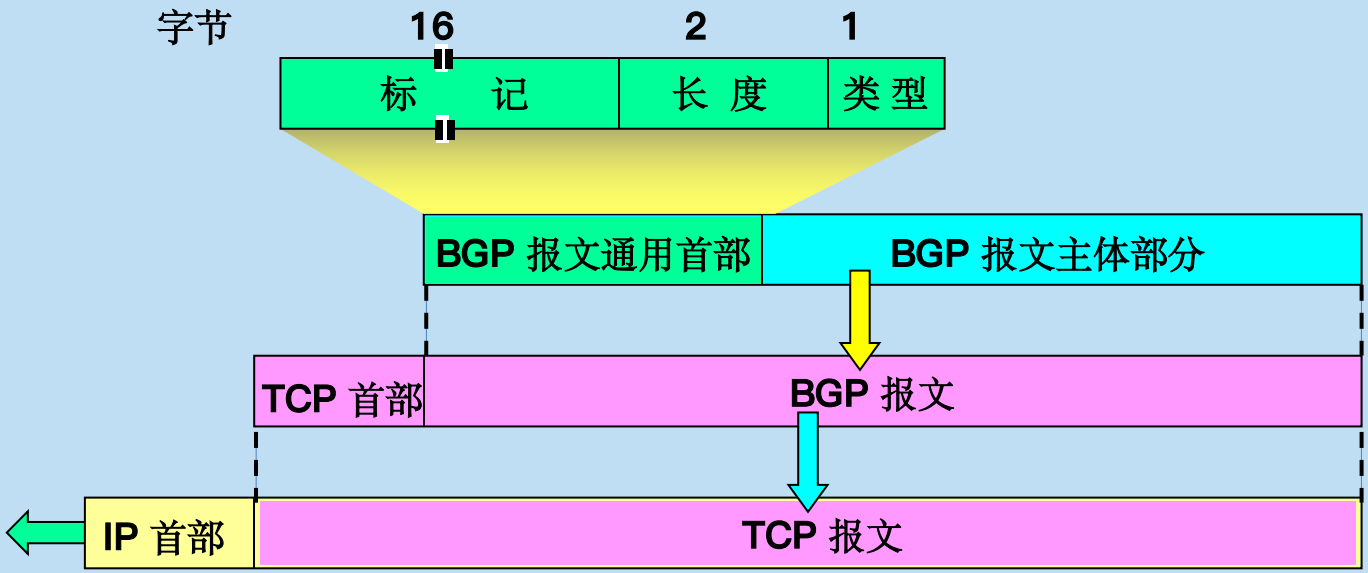
iBGP (internal BGP) 连接：运行 iBGP 协议， 在 AS 内部的路由器之间交换 BGP 路由信息。

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。

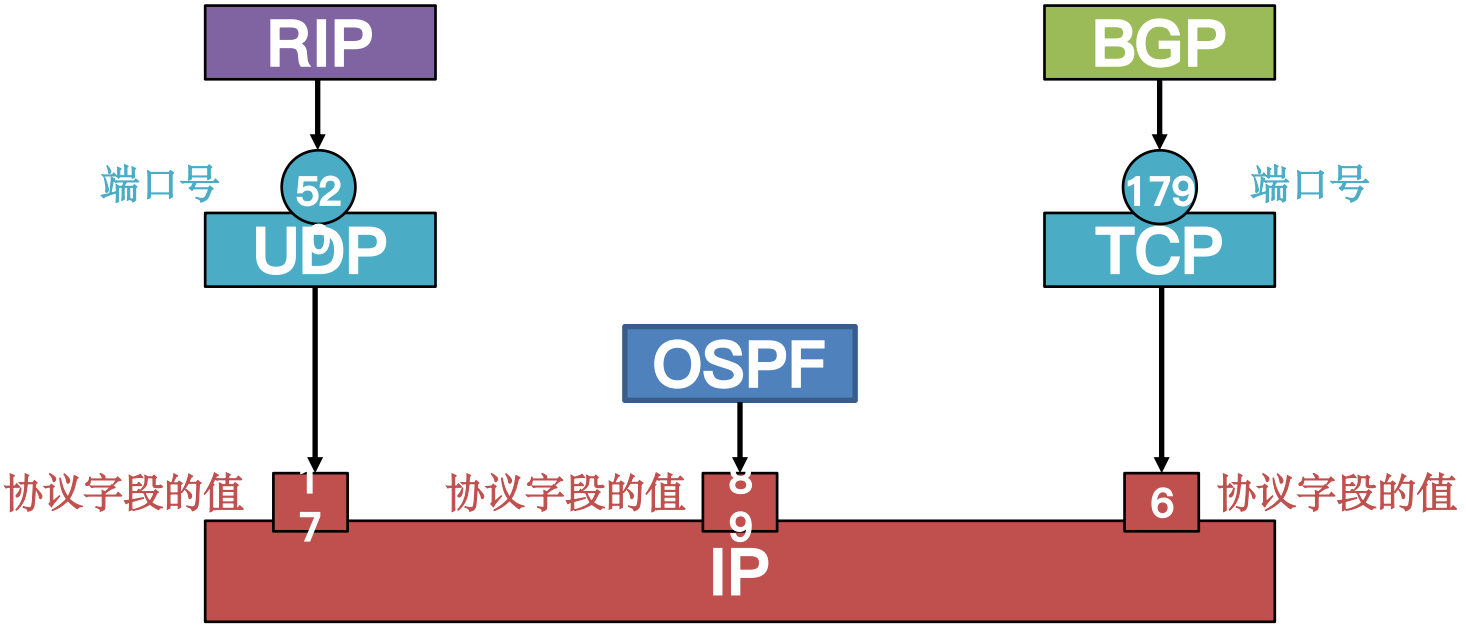
BGP 报文具有通用首部



【2017年 题37】直接封装RIP、OSPF、BGP报文的协议分别是（ D）。

A. TCP、UDP、IP　　B. TCP、IP、UDP　　C. UDP、TCP、IP　　D. UDP、IP、TCP

解析

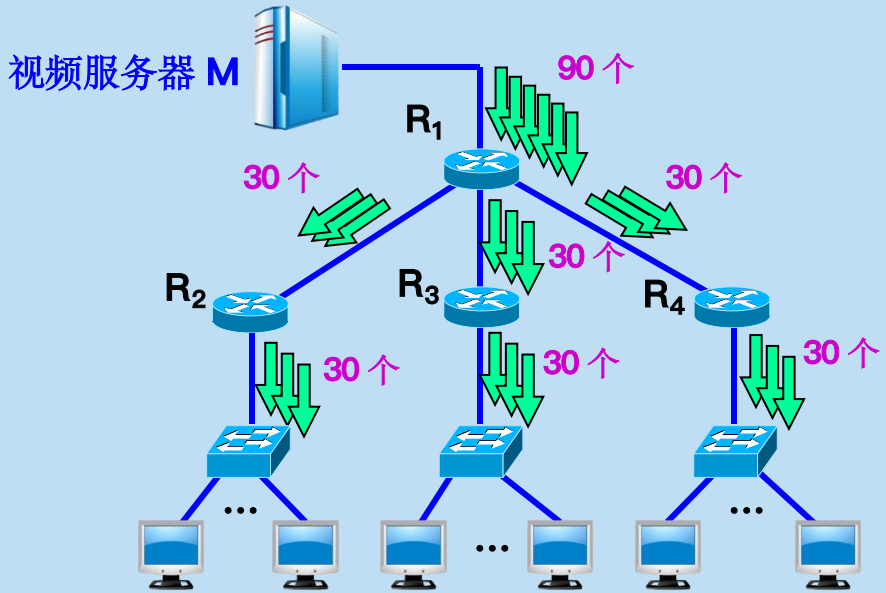


4.7 IP 多播

4.7.1 IP 多播的基本概念

- 1988 年，Steve Deering 首次提出 IP 多播的概念。
- 多播 (multicast)：以前曾译为组播。
- 目的：更好地支持一对多通信。
- 一对多通信：一个源点发送到许多个终点。

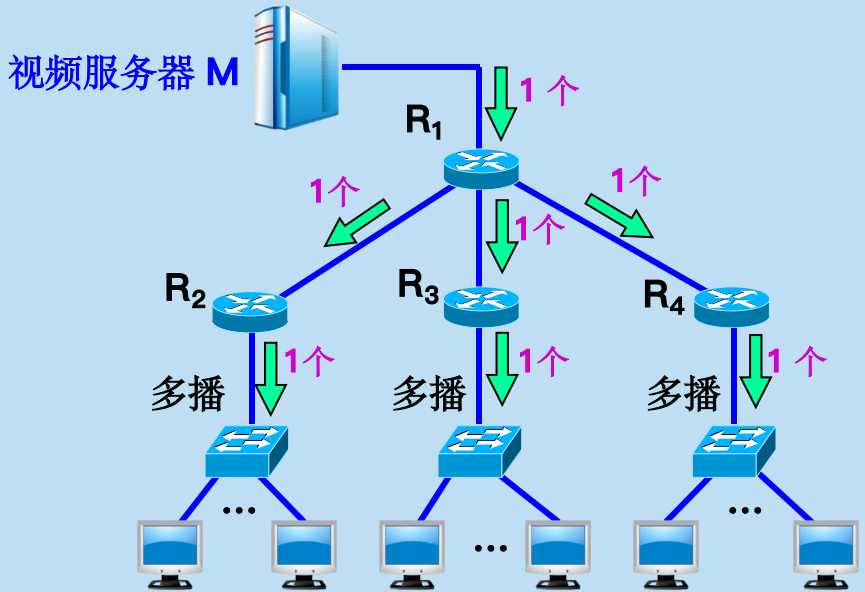
多播可大大节约网络资源



共有 90 个主机接收视频节目

采用单播方式，
向 90 台主机传送
同样的视频节目
需要发送 90 个单播

多播可大大节约网络资源



采用多播方式，
只需发送一次到多播组。
路由器复制分组。
局域网具有硬件多播功能，
不需要复制分组。

当多播组的主机数很大时（如成千上万个），采用多播方式就可明显地减轻网络中各种资源的消耗。

多播 IP 地址

- 在 IP 多播数据报的目的地址需要写入多播组的标识符。
- 多播组的标识符就是 IP 地址中的 D 类地址（多播地址）。

  地址范围：224.0.0.0 ~ 239.255.255.255
- 每一个 D 类地址标志一个多播组。

多播地址只能用于目的地址，不能用于源地址。

4.8 虚拟专用网 VPN 和网络地址转换 NAT

4.8.1 虚拟专用网 VPN

- 由于 IP 地址的紧缺，一个机构能够申请到的IP地址数往往远小于本机构所拥有的主机数。
- 考虑到互联网并不很安全，一个机构内也并不需要把所有的主机接入到外部的互联网。
- 如果一个机构内部的计算机通信也是采用 TCP/IP 协议，那么这些仅在机构内部使用的计算机就可以由本机构自行分配其 IP 地址。

本地地址与全球地址

- 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，而不需要向互联网的管理机构申请。
- 全球地址：全球唯一的 IP 地址，必须向互联网的管理机构申请。
- 问题：如何区分本地地址和全球地址？
- 解决：RFC 1918 指明了一些专用地址 (private address)。

  - 专用地址只能用作本地地址，而不能用作全球地址。
  - 互联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。

RFC 1918 指明的专用 IP 地址

三个专用 IP 地址块：

(1) 10.0.0.0/8

A类，从 10.0.0.0 到 10.255.255.255。1 个。

(2) 172.16.0.0/12

B类，从 172.16.0.0 到 172.31.255.255。连续 16 个。

(3) 192.168.0.0/16

C类，从 192.168.0.0 到 192.168.255.255。连续 256 个。

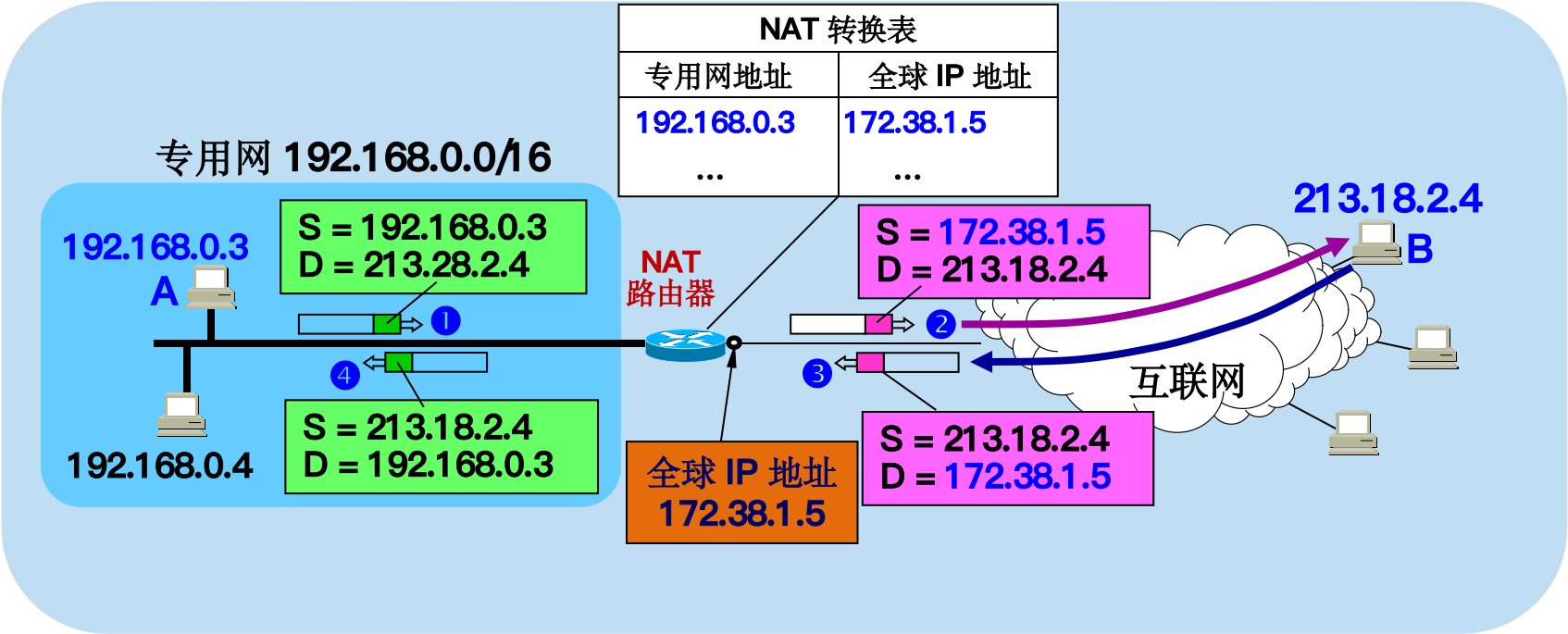
4.8.2 网络地址转换 NAT

- 问题：在专用网上使用专用地址的主机如何与互联网上的主机通信（并不需要加密）？
- 解决：
  1. 再申请一些全球 IP 地址。但这在很多情况下是不容易做到的。
  2. 采用网络地址转换 NAT。这是目前使用得最多的方法。

网络地址转换 NAT (Network Address Translation)

- 1994 年提出。
- 需要在专用网连接到互联网的路由器上安装 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 IP 地址。
- 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全球 IP 地址，才能和互联网连接。

网络地址转换的过程


NAT 转换表
专用网地址
全球 IP 地址
192.168.0.3
172.38.1.5
...
...
专用网 192.168.0.0/16
192.168.0.3
A
S = 192.168.0.3
D = 213.28.2.4
192.168.0.4
S = 213.18.2.4
D = 192.168.0.3
NAT
路由器
全球 IP 地址
172.38.1.5
S = 172.38.1.5
D = 213.18.2.4
S = 213.18.2.4
D = 172.38.1.5
互联网
213.18.2.4
B
1
2
3
4

网络地址转换的过程

- 在内部主机与外部主机通信时，在 NAT 路由器上发生了两次地址转换：

1. 离开专用网时：替换源地址，将内部地址替换为全球地址。
2. 进入专用网时：替换目的地址，将全球地址替换为内部地址。

NAT 地址转换表举例

方向	字段	旧的 IP 地址	新的 IP 地址
出（发往互联网）	源 IP 地址	192.168.0.3	172.38.1.5
入（进入专用网）	目的 IP 地址	172.38.1.5	192.168.0.3

网络地址转换 NAT

- 当 NAT 路由器具有 *n* 个全球 IP 地址时，专用网内最多可以同时有 *n* 台主机接入到互联网。
- 可以使专用网内较多数量的主机轮流使用 NAT 路由器有限数量的全球 IP 地址。

通过 NAT 路由器的通信必须由专用网内的主机发起，因此，专用网内部的主机不能充当服务器用。

网络地址与端口号转换 NAPT

- NAT 并不能节省 IP 地址。
- NAPT 可以使多台拥有本地地址的主机，共用一个 全球 IP 地址，同时和互联网上的不同主机进行通信。
- 使用运输层端口号的 NAT 叫做网络地址与端口号转换 NAPT (Network Address and Port Translation)，而不使用端口号的 NAT 就叫做传统的 NAT (traditional NAT)。

NAPT 地址转换表

NAPT 地址转换表举例

方向	字段	旧的 IP 地址和端口号	新的 IP 地址和端口号
出	源 IP 地址 : TCP 源端口	192.168.0.3 : 30000	172.38.1.5 : 40001
出	源 IP 地址 : TCP 源端口	192.168.0.4 : 30000	172.38.1.5 : 40002
入	目的 IP 地址 : TCP 目的端口	172.38.1.5 : 40001	192.168.0.3 : 30000
入	目的 IP 地址 : TCP 目的端口	172.38.1.5 : 40002	192.168.0.4 : 30000

- NAPT 把专用网内不同的源 IP 地址都转换为相同的全球 IP 地址，将 TCP 源端口号转换为新的 TCP 端口号（互不相同）。
- 收到从互联网发来的应答时，从 IP 数据报的数据部分找出运输层端口号，从 NAPT 转换表中找到正确的目的主机。

本讲小结

掌握：

- RIP内部网关路由协议计算过程
- 内部网关协议RIP与OSPF的区别

了解：

- 多播
- 网络地址映射NAT实现过程