计算机网络

Computer Network

11

因特网路由

理论教程

厦門大學

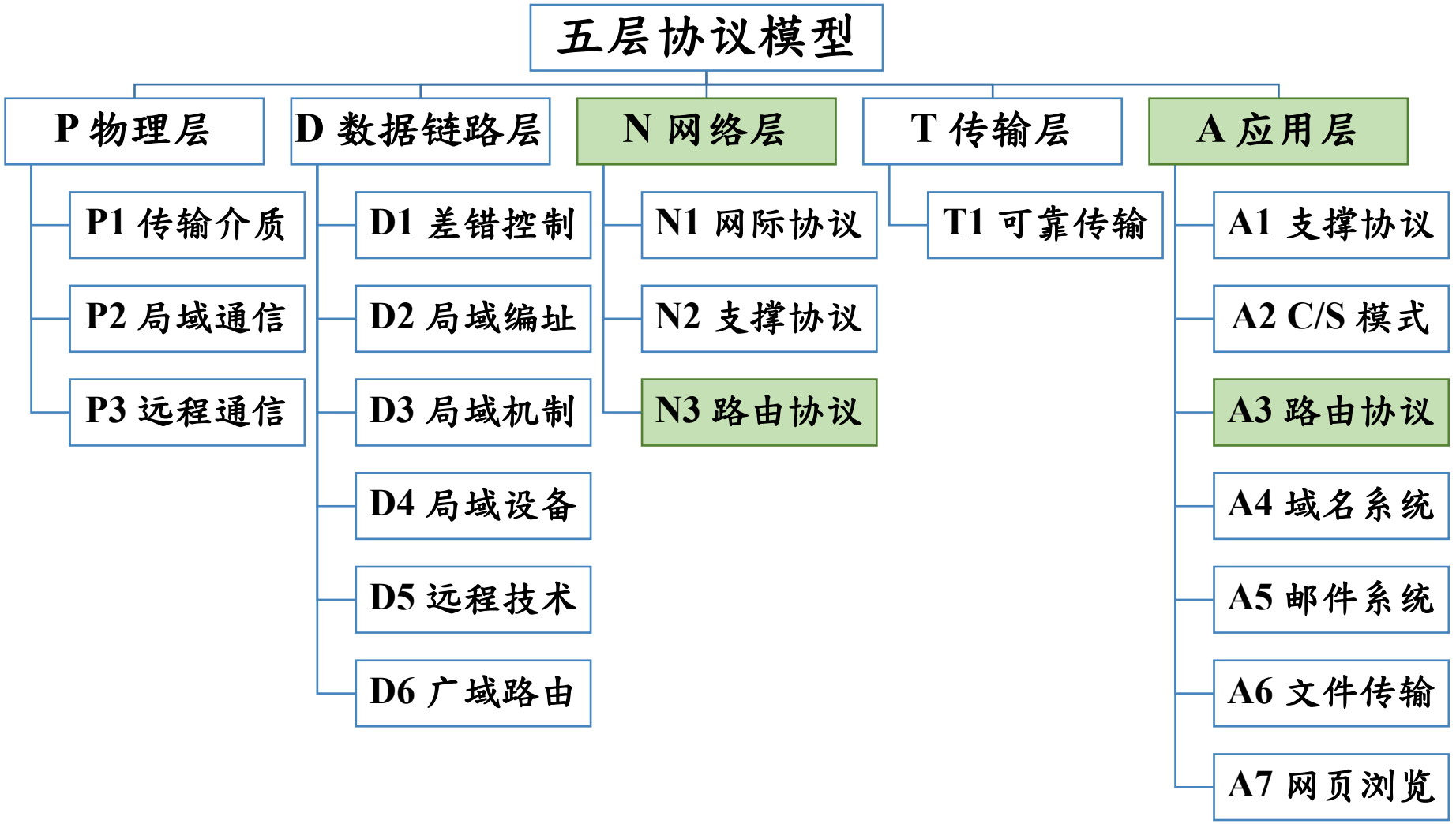
XIAMEN UNIVERSITY

信息学院 黄 炜

(特色化示范性软件学院) 博士,副教授

School of Informatics Wei Huang

知识框架


五层协议模型
P 物理层
D 数据链路层
N 网络层
T 传输层
A 应用层
P1 传输介质
P2 局域通信
P3 远程通信
D1 差错控制
D2 局域编址
D3 局域机制
D4 局域设备
D5 远程技术
D6 广域路由
N1 网际协议
N2 支撑协议
N3 路由协议
T1 可靠传输
A1 支撑协议
A2 C/S 模式
A3 路由协议
A4 域名系统
A5 邮件系统
A6 文件传输
A7 网页浏览

主要内容

- 静态路由与动态路由
- 自治系统（AS）的概念
- 内部网关协议（IGP）
  - RIP协议的工作原理和特点
  - OSPF协议的工作原理和特点
- 外部网关协议（EGP）
  - BGP协议

对应课本章节

- **PART IV Internetworking**
  - **Chapter 27 Internet Routing And Routing Protocols**

内容纲要

路由协议

- 路由协议（Routing Protocol）
- 定义
  - 一种指定数据报文转送方式的网络协议。
- 原理
  - 路由协议创建了路由表，描述了网络拓扑结构。
  - 路由协议与路由器协同工作，执行路由选择和数据包转发功能。

路由选择的复杂性

- 路由选择的复杂性
  - 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。
- 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。

路由算法的自适应性分类

- 静态路由选择策略（非自适应路由选择）
  - 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（自适应路由选择）
  - 能较好地适应网络状态的变化，但实现较复杂，开销较大。

内容纲要

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。
- 不应该让所有的路由器记录所有的网络的到达方式
  - 路由表非常大
    - 处理花时间，路由器间交换路由信息所需的带宽大。
  - 隐私考虑
    - 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因特网上。

自治系统 AS (Autonomous System)

- 自治系统（Autonomous System）
  - 定义：自治系统是在单一的技术管理下的一组路由器
  - 域内路由选择：使用一种 AS 内部的路由选择协议和共同度量以确定分组在该 AS 内的路由
    - 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。
  - 域间路由选择：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

路由选择协议分层

- 内部网关协议（Interior Gateway Protocol，IGP）
  - 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
  - 示例： RIP 和 OSPF 协议。
- 外部网关协议（External Gateway Protocol，EGP）
  - 自治系统边界网关使用的路由选择协议
    - 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传到另一个自治系统中。
  - 示例：BGP-4。

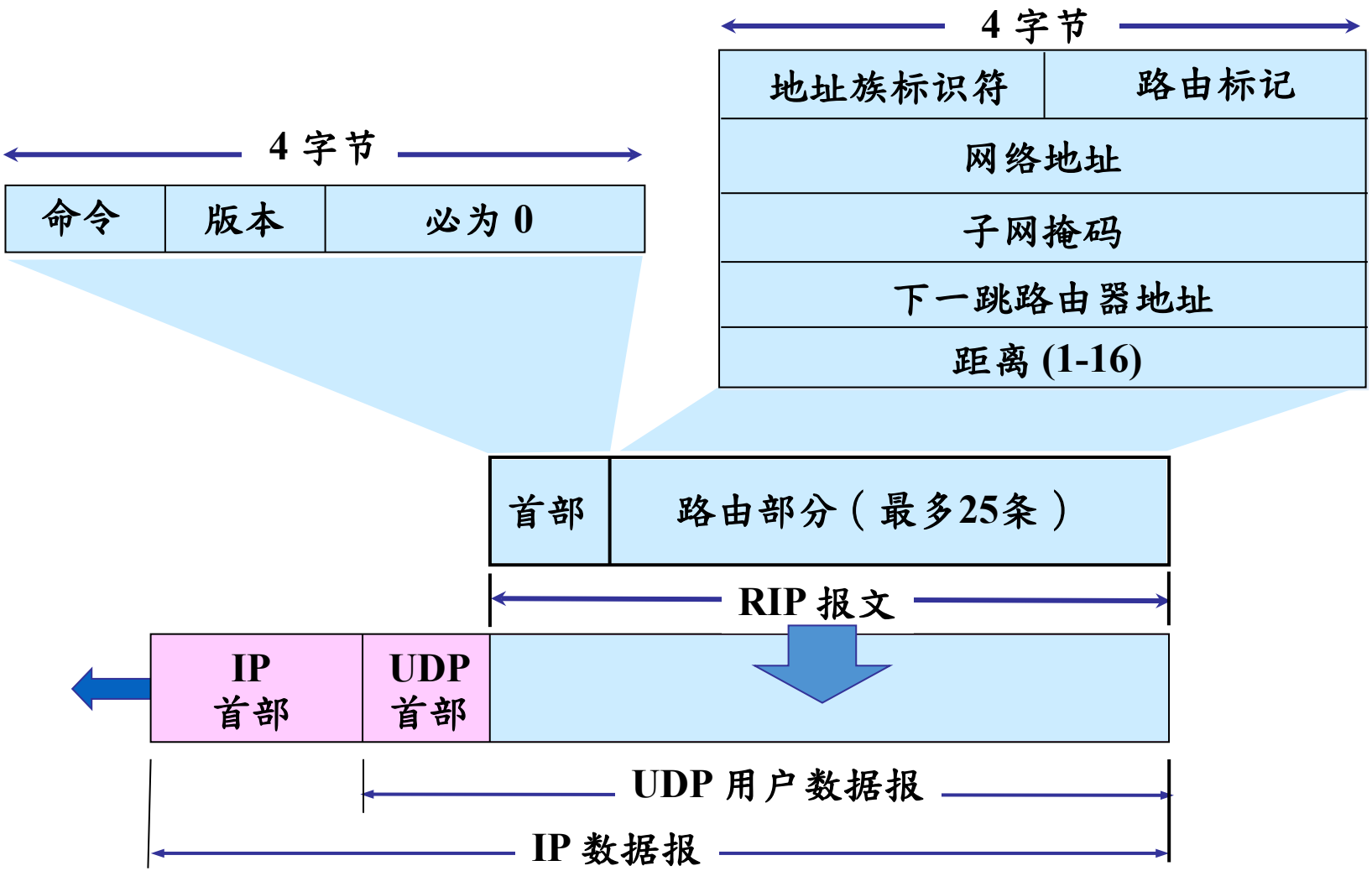
RFC对路由和网关应视为同义词

内容纲要

内部网关协议 RIP

- 路由信息协议（Routing Information Protocol，RIP）
  - 最先得到广泛使用的内部网关协议。
- 工作原理：分布式的基于距离向量的路由选择协议
  - 网络中每个路由器维护从自己到其它目的网络的距离记录。
- 要点
  - 仅和相邻路由器交换信息。
  - 交换的信息是自己的路由表。
  - 按固定间隔时间交换路由信息。

RIP2 协议的报文格式



RIP2 报文

- 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可能收到本自治系统以外的路由选择信息。
- 随后填入某个网络地址、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距离的定义

- 距离也称为跳数（hop count）
  - 从路由器到直接连接的网络的距离定义为 1。
    - 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。
  - 每经过一个路由器，跳数就加 1。
  - 距离为16 时即相当于不可达。（只适用于小型互联网）
- 好的路由就是它通过的路由器的数目少
  - 不能在两个网络之间同时使用多条路由。

路由表的建立

- 路由器刚开始工作时
  - 记录到直连网络的距离（记为1）。
- 路由器更新路由信息
  - 和有限的相邻路由器交换并更新路由信息。
- 收敛过程块
  - 经过若干次更新后，所有的路由器最终知道到达本自治系统中任何网络的最短距离和下一跳路由器的地址。
  - 在自治系统中所有的结点都得到正确的路由选择信息。

距离向量算法

- 算法基础：Bellman-Ford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到B 拆成两段路径 A到X 和 X到B，则每段路径 A到X 和 X到B 也都分别是结点 A到X 和结点X到B的最短路径。
- 收到相邻路由器（其地址为 X）的一个 RIP 报文
  - 第一步，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
    - 把所有的“距离”字段的值加 1。

距离向量算法

- 收到相邻路由器（其地址为 X）的一个 RIP 报文
  - 第二步，对修改后的 RIP 报文中的每个项目，重复：

目的网络是否在路由表中	下一跳地址和已有记录相同	距离比已有记录小	处理
否			添加
是	是		替换
	否	是	更新
		否	不更新

  - 第三步，若 3 分钟还没有收到相邻路由器的更新路由表，则把此相邻路由器距离置为16（不可达）。

路由器之间交换信息

- RIP协议工作过程
  - 让互联网的所有路由器都和相邻路由器不断交换路由信息
  - 并不断更新其路由表，使得每个路由器到每个目的网络的路由都是最短的（即跳数最少）。
- 结果
  - 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。

RIP 协议的优缺点

- 优点
  - 实现简单，开销较小。
- 缺点
  - 限制了网络的规模，最大距离为 15（16 表示不可达）。
  - 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。
  - 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。

RIP 协议的演示

正常情况

中间数字1表示
距离是 1

左边数字1表示
从本路由器到网 1

右侧为数字表示下一跳路由器号；
或为减号，表示直接交付

1 1 –

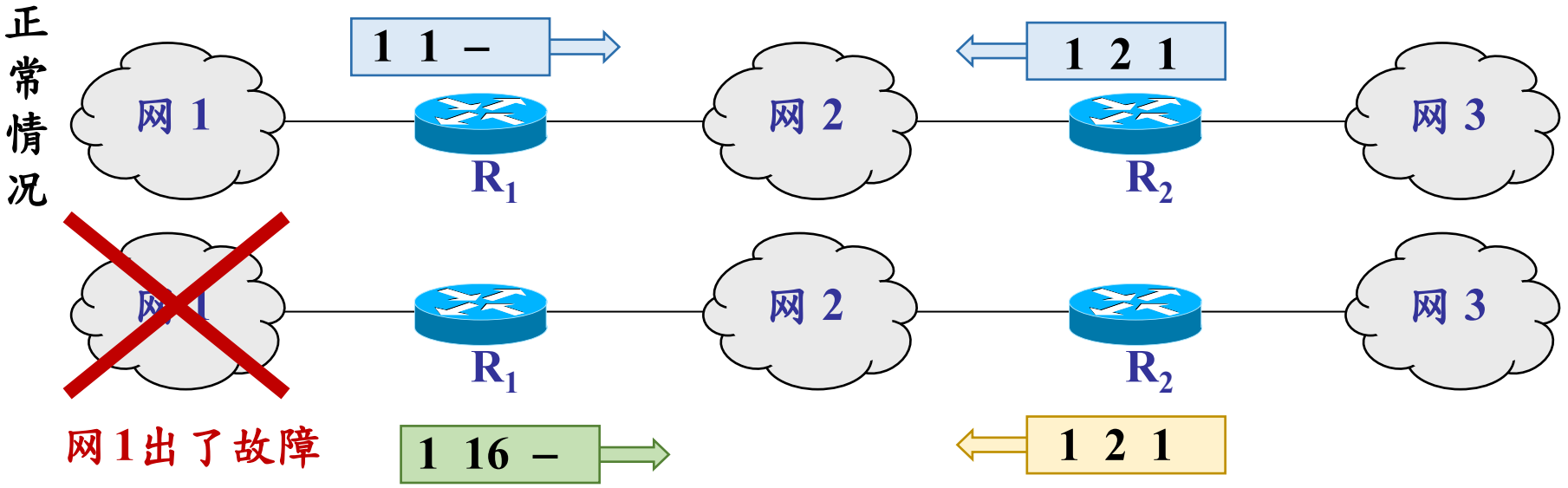
R_1 说：“我到网 1 的距离是 1，是直接交付。”

RIP 协议的演示

正常情况

R_2 说：“我到网 1 的距离是 2，是经过 R_1。”

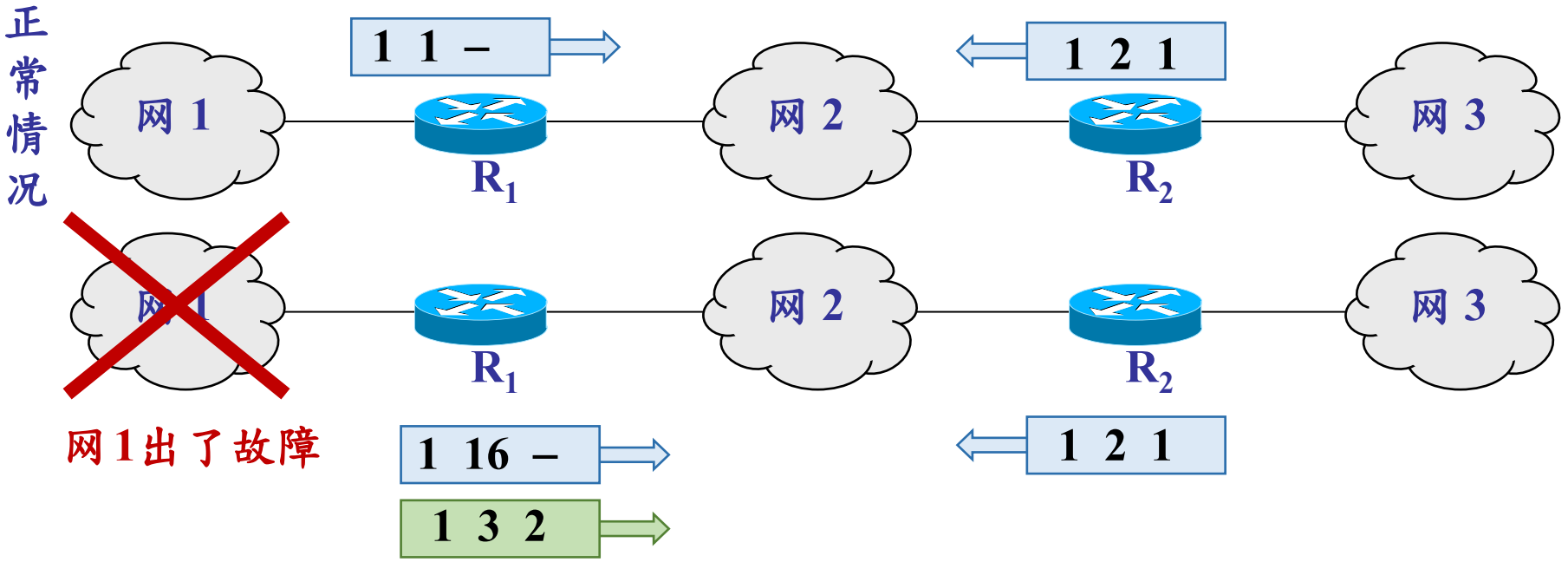
RIP 协议的演示



R_1 说：“我到网 1 的距离是 16（表示无法到达），是直接交付。”

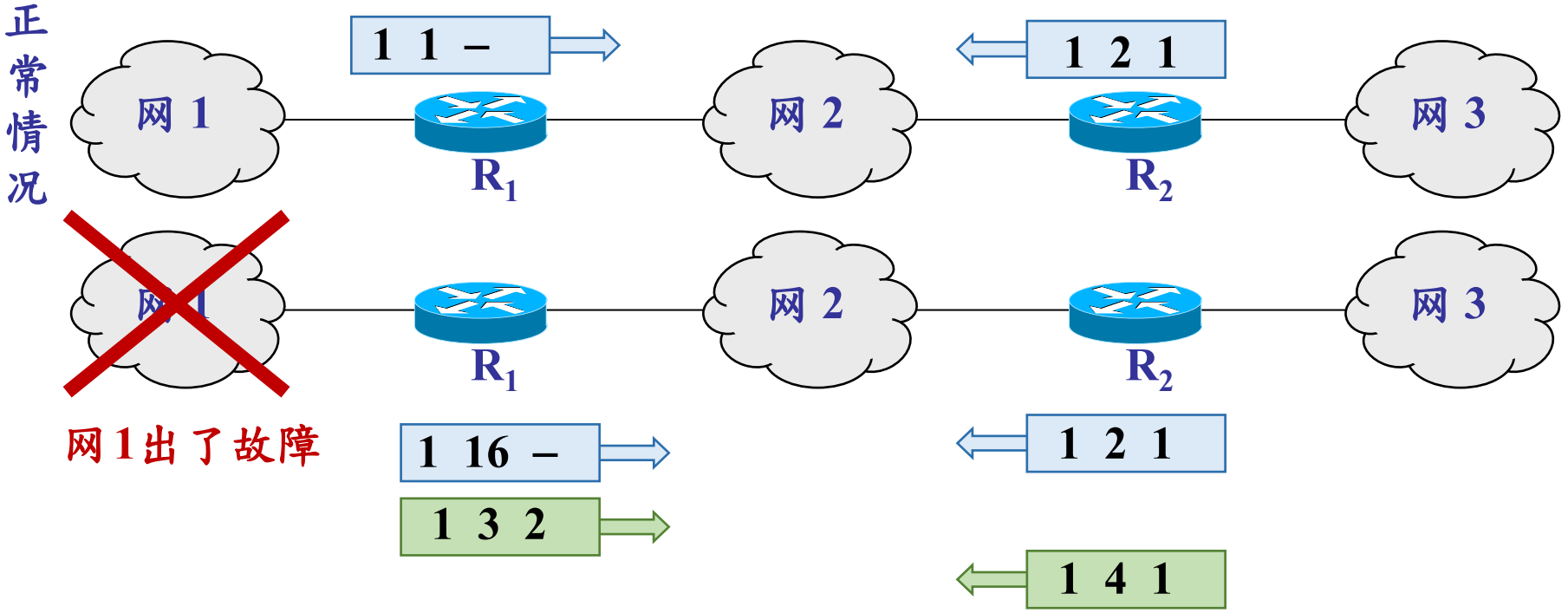
但 R2 在收到 R1 的更新报文之前，还发送原来的报文，因为这时 R2 并不知道 R1 出了故障。

RIP 协议的演示



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。

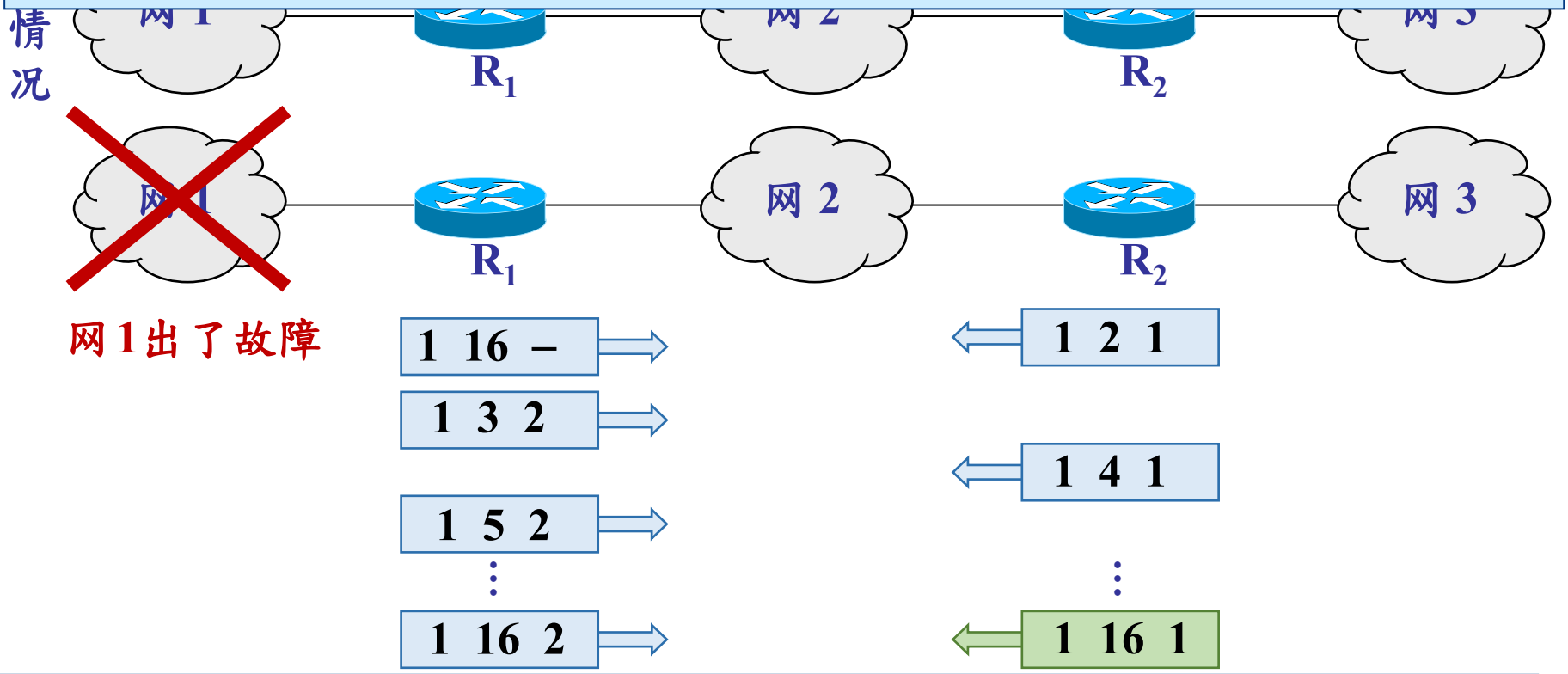
RIP 协议的演示



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”。

RIP 协议的演示

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。

内容纲要

内部网关协议 OSPF

- 开放最短路径优先（Open Shortest Path First，OSPF）
  - 开放：不受某一家厂商控制，而是公开发表的。
  - 最短路径优先：因为使用了 Dijkstra 提出的最短路径算法
  - 协议名不表示其他的路由选择协议不是“最短路径优先”。
  - 采用分布式的链路状态协议。

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。

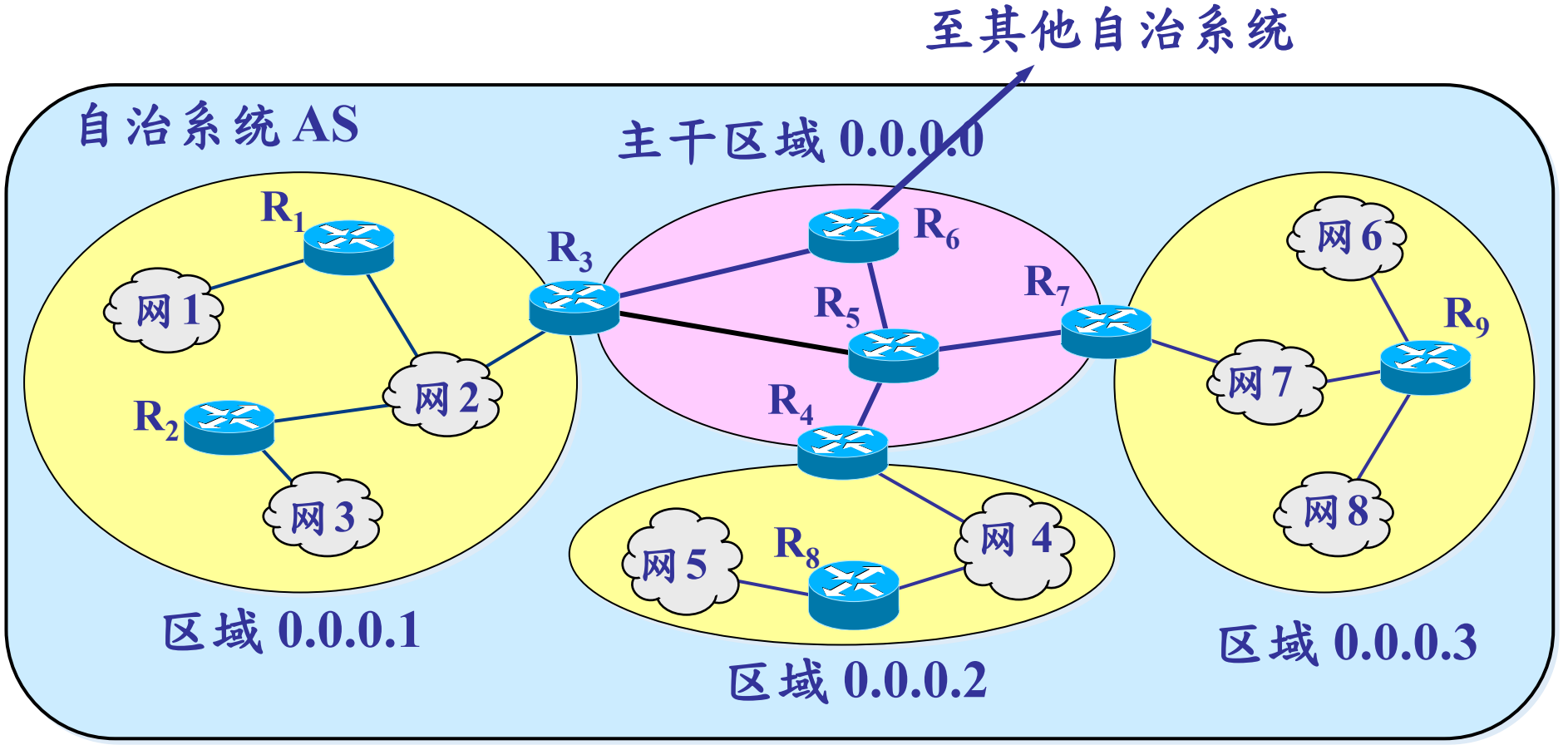
链路状态数据库

- 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。
- 链路状态数据库的同步
  - 所建立的数据库是全网的拓扑结构图，在全网范围内一致。
- 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)

- 区域：OSPF 将自治系统再划分为若干个更小的范围。
  - 目的：使得 OSPF 能够用于规模很大的网络
- 区域标识符： 32 位（用点分十进制表示）。
- 区域不能太大
  - 在一个区域内的路由器最好不超过 200 个。

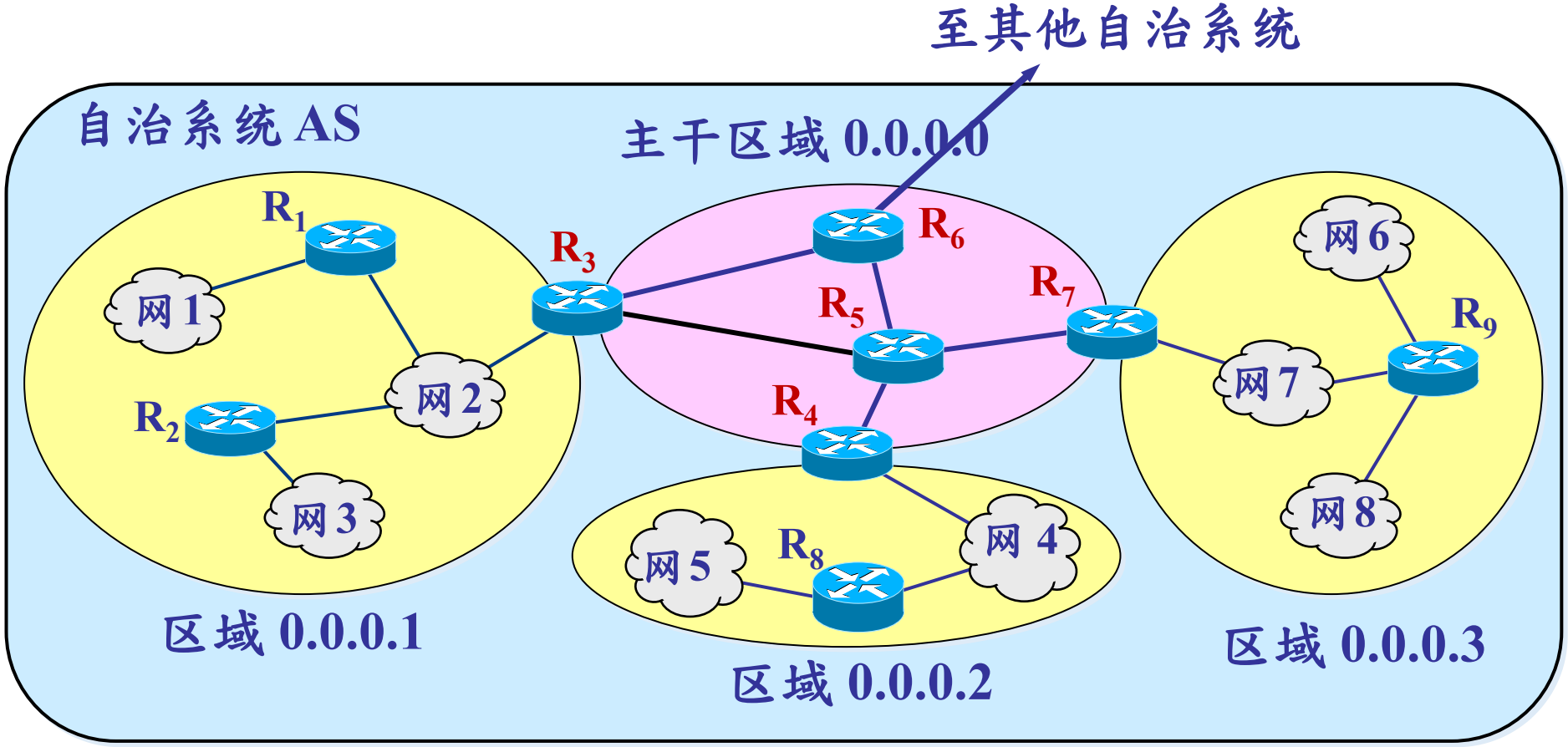
OSPF 划分为两种不同的区域



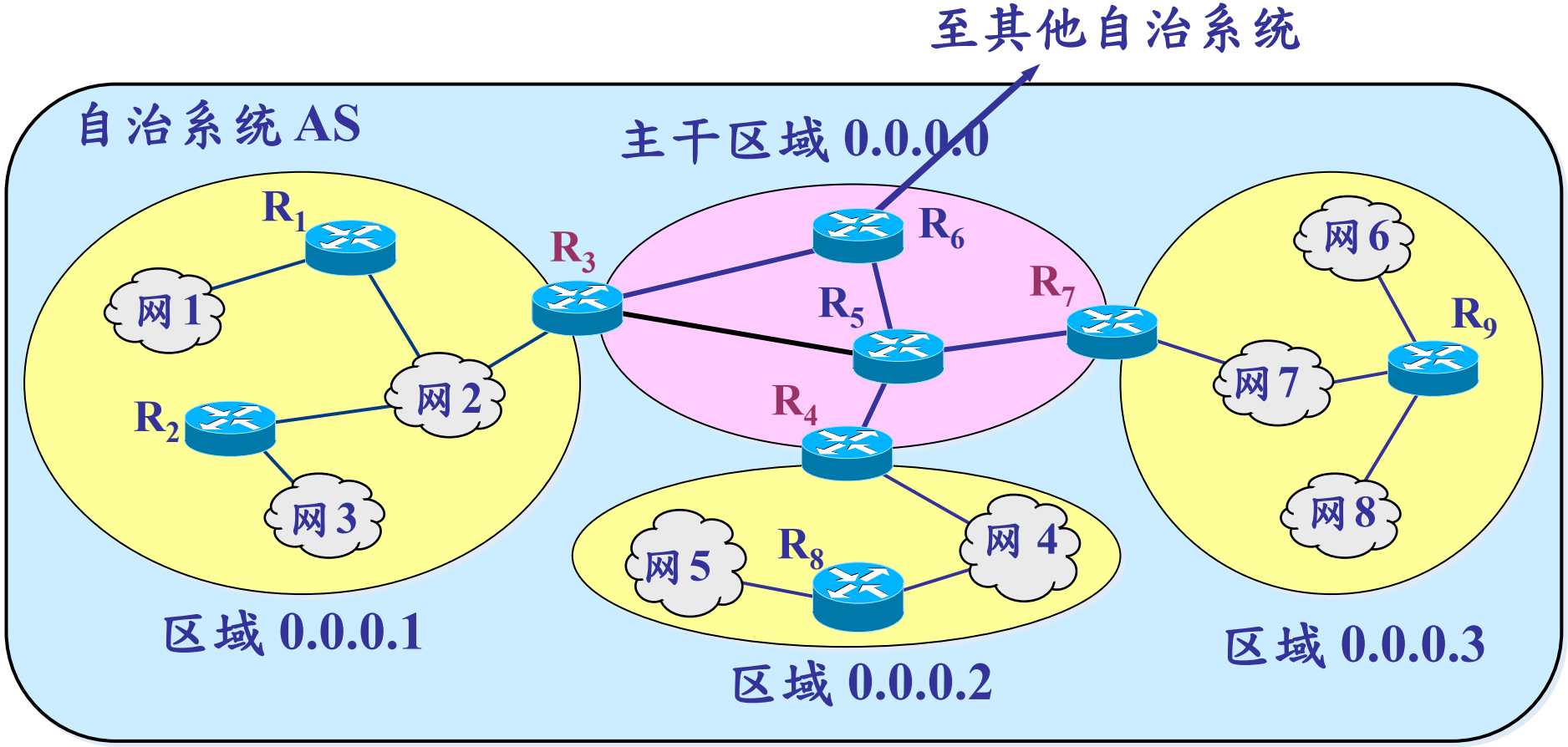
划分区域

- 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
R2
R3
R4
R5
R6
R7
R8
R9
网 1
网 2
网 3
网 4
网 5
网 6
网 7
网 8
区域 0.0.0.1
区域 0.0.0.2
区域 0.0.0.3

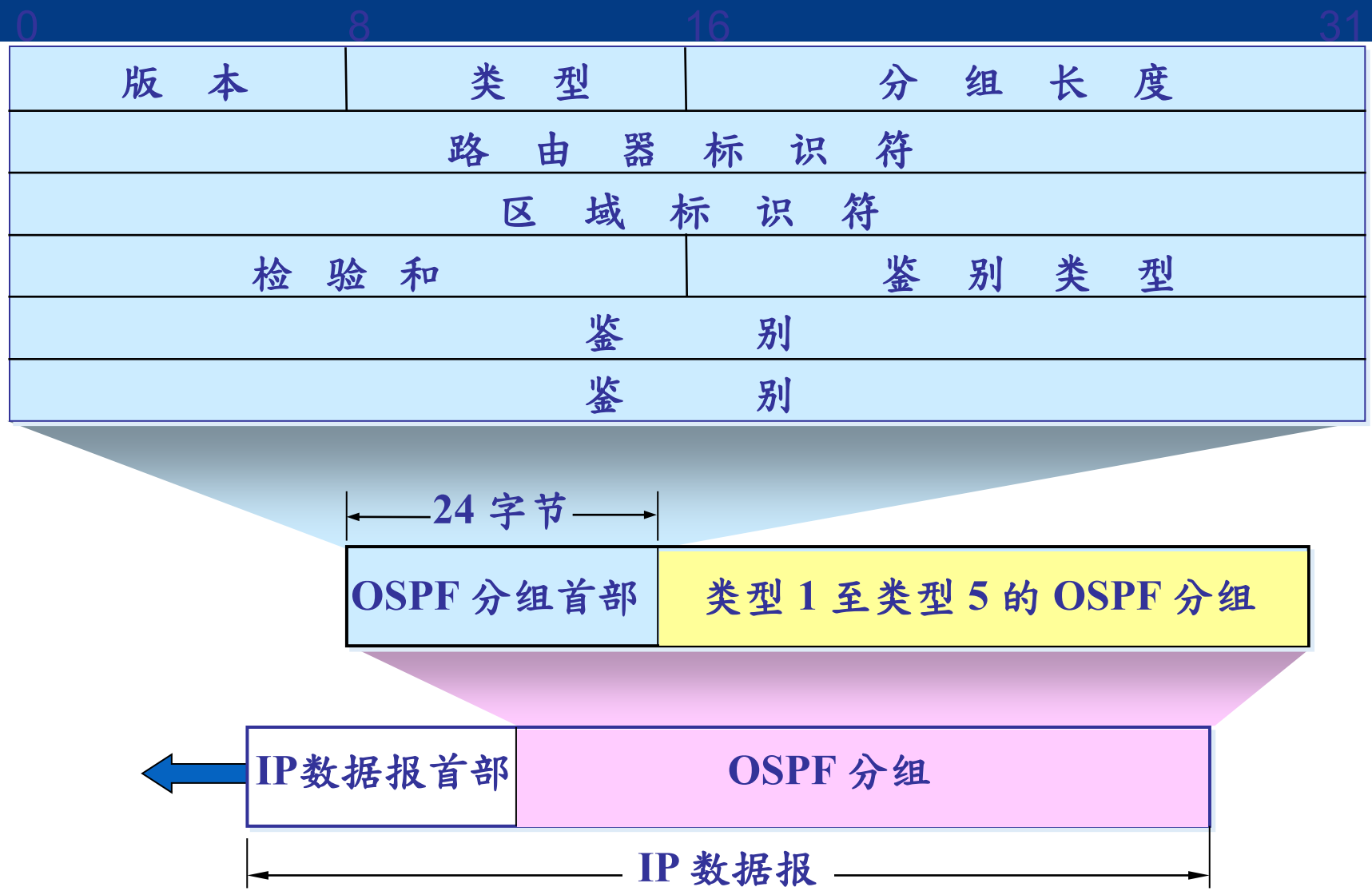
OSPF 直接用 IP 数据报传送

- 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。
- 如果到同一目的网络有多条相同代价的路径，则可将通信量分配给这几条路径，称：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

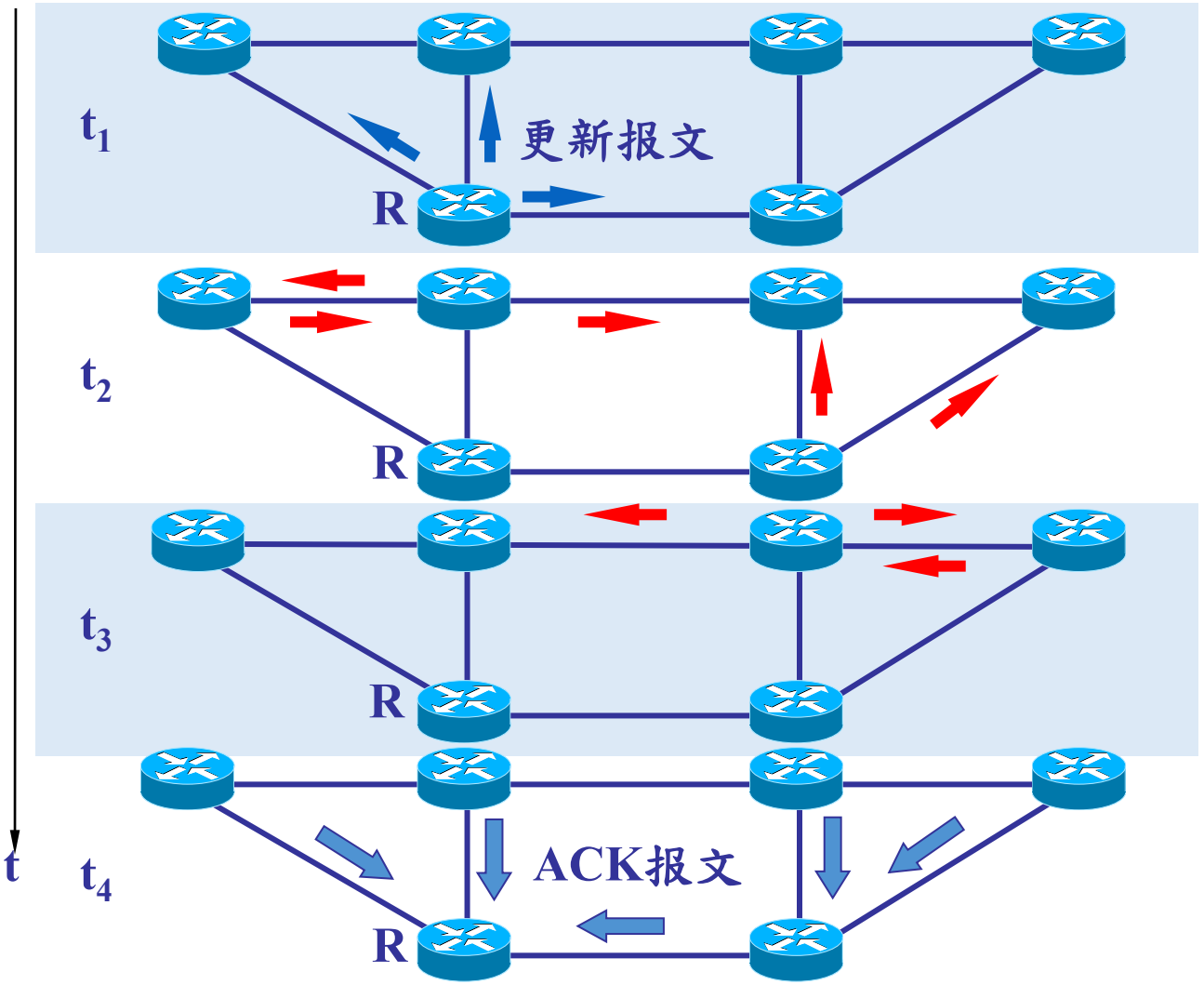
OSPF 的五种分组类型

- 问候(Hello)分组
- 数据库描述(Database Description)分组
- 链路状态请求(Link State Request)分组
- 链路状态更新(Link State Update)分组
  - 用洪泛法对全网更新链路状态
- 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 100 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。

内容纲要

1	路由协议的基本概念
2	自治系统
3	路由信息协议
4	开放最短路径优先协议
5	边界网关协议

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

- 因特网规模太大，使自治系统之间路由选择非常困难。
- 自治系统之间路由选择，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。
- 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。

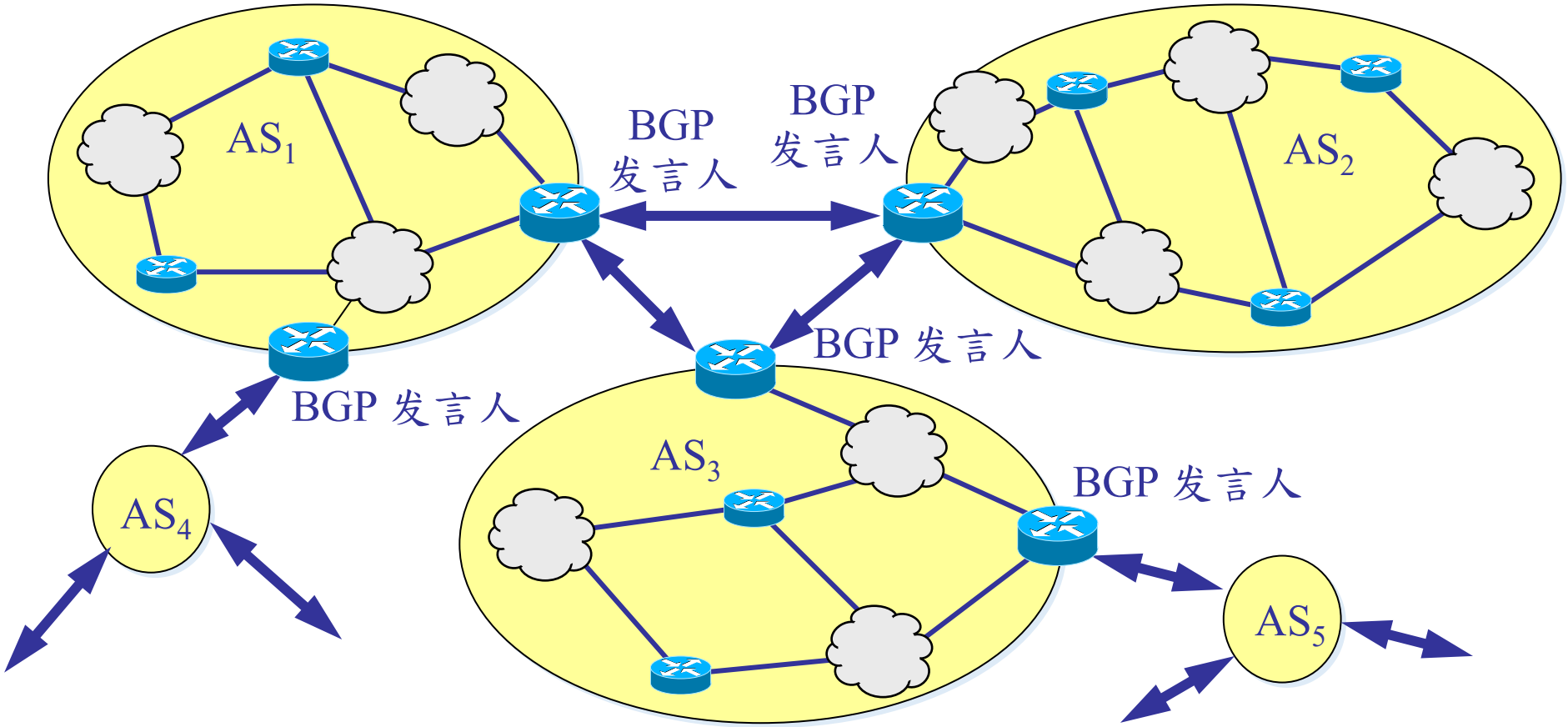
BGP 发言人 (BGP speaker)

- 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

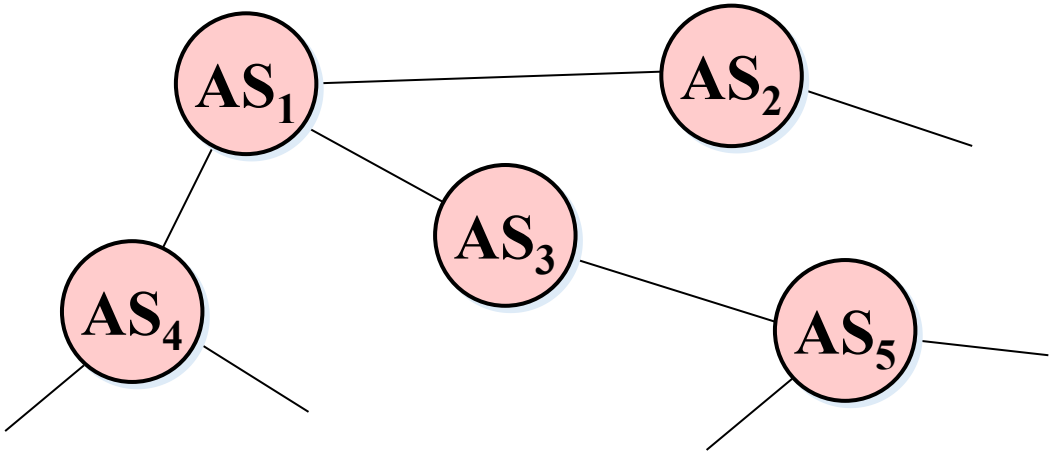
- 一个 BGP 发言人与其他自治系统中的 BGP 发言人要交换路由信息，就要先建立 TCP 连接，然后在此连接上交换 BGP 报文以建立 BGP 会话(session)，利用 BGP 会话交换路由信息。
- 使用 TCP 连接能提供可靠的服务，也简化了路由选择协议。
- 使用 TCP 连接交换路由信息的两个 BGP 发言人，彼此成为对方的邻站或对等站。

BGP 发言人和自治系统 AS 的关系



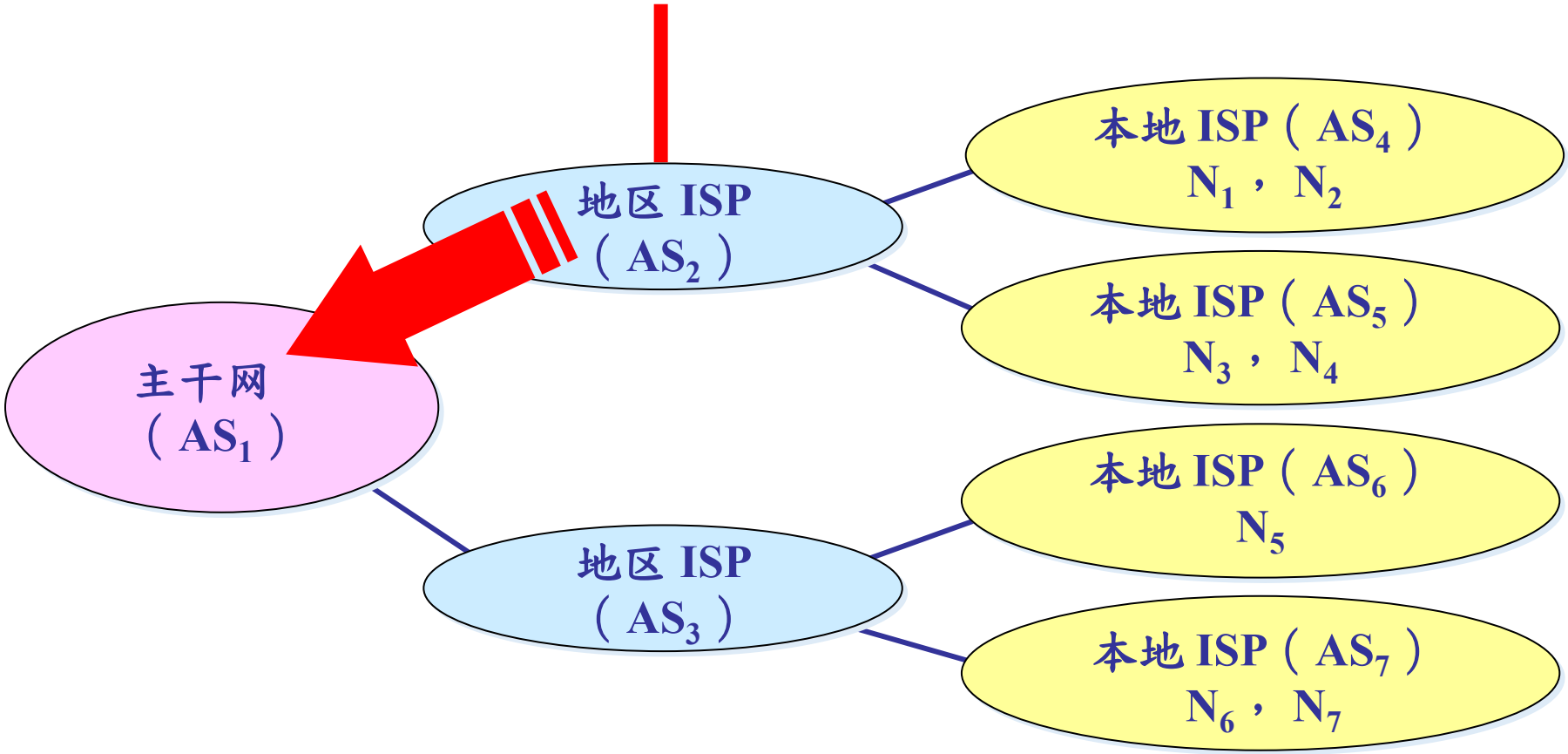
AS 的连通图举例

- 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。



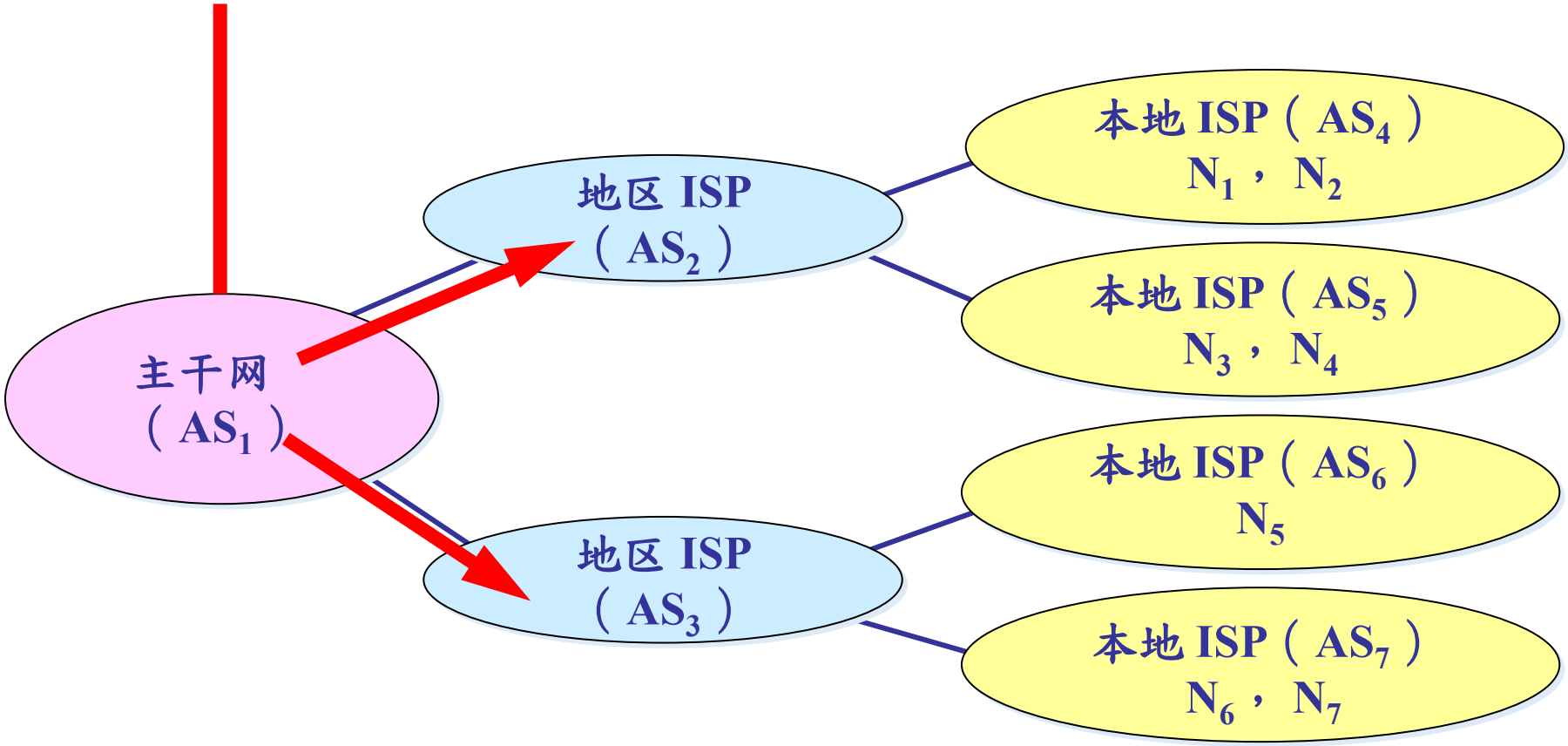
BGP 发言人交换路径向量

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。

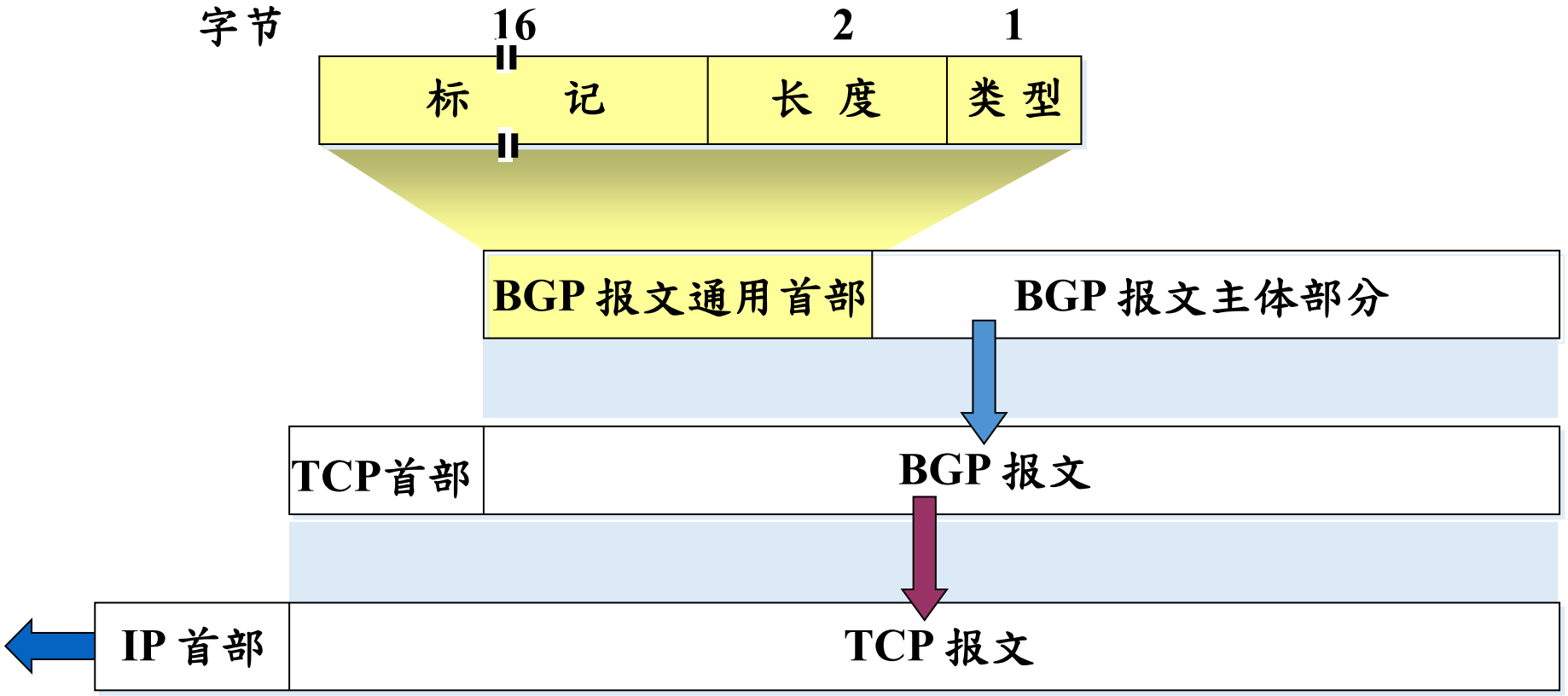
BGP 协议的特点

- 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

- 打开(open)：与相邻的另一个BGP发言人建立关系。
- 更新(update)：发送某一路由的信息，以及列出要撤消的多条路由。
- 保活(keepalive)报文：确认打开报文和周期性地证实邻站关系。
- 通知(notification)报文：发送检测到的差错。
- RFC 2918 中增加了 Route-refresh 报文：请求对等端重新通告。

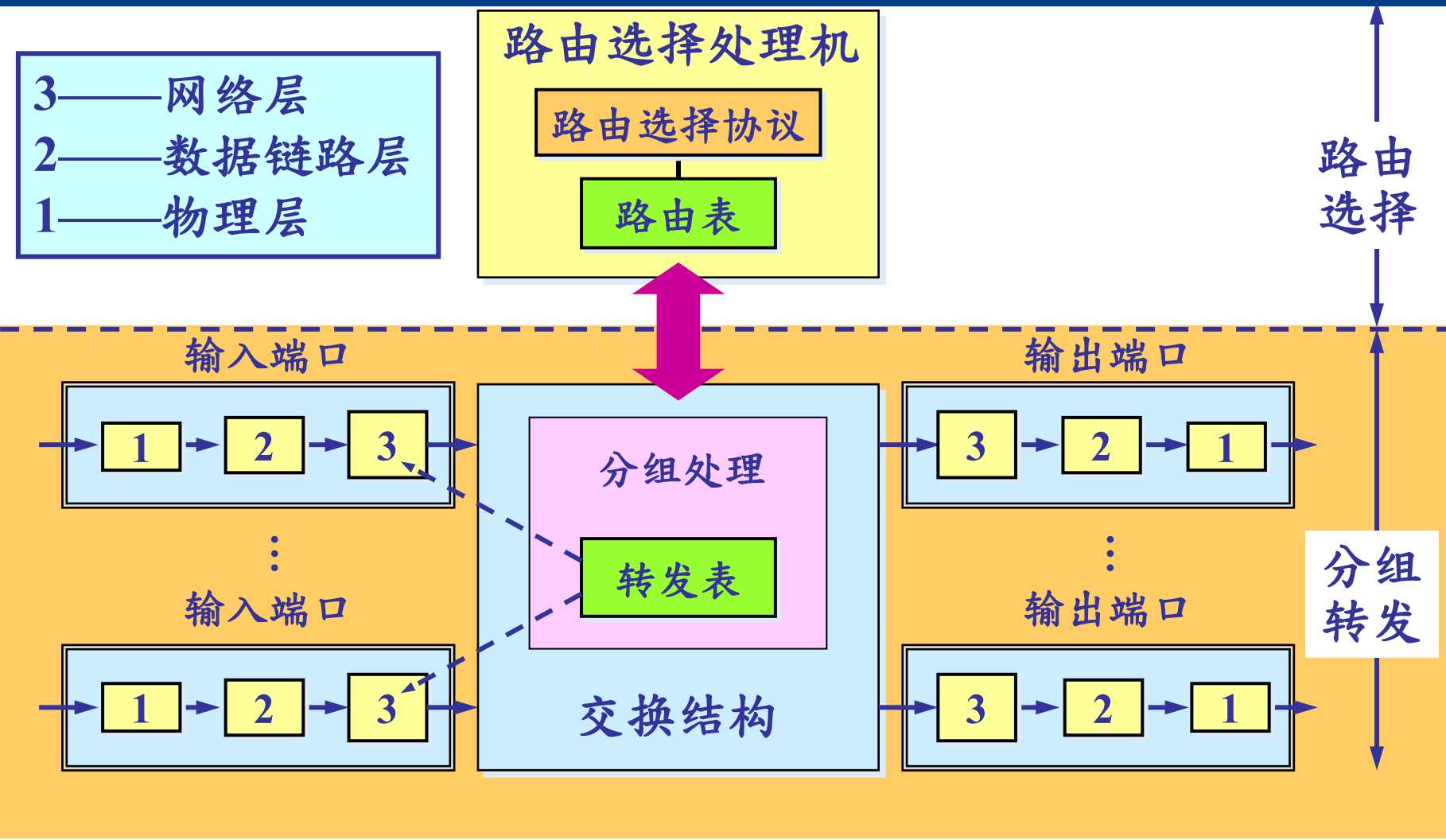
BGP 报文具有通用的首部



路由器在网际互连中的作用

- 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

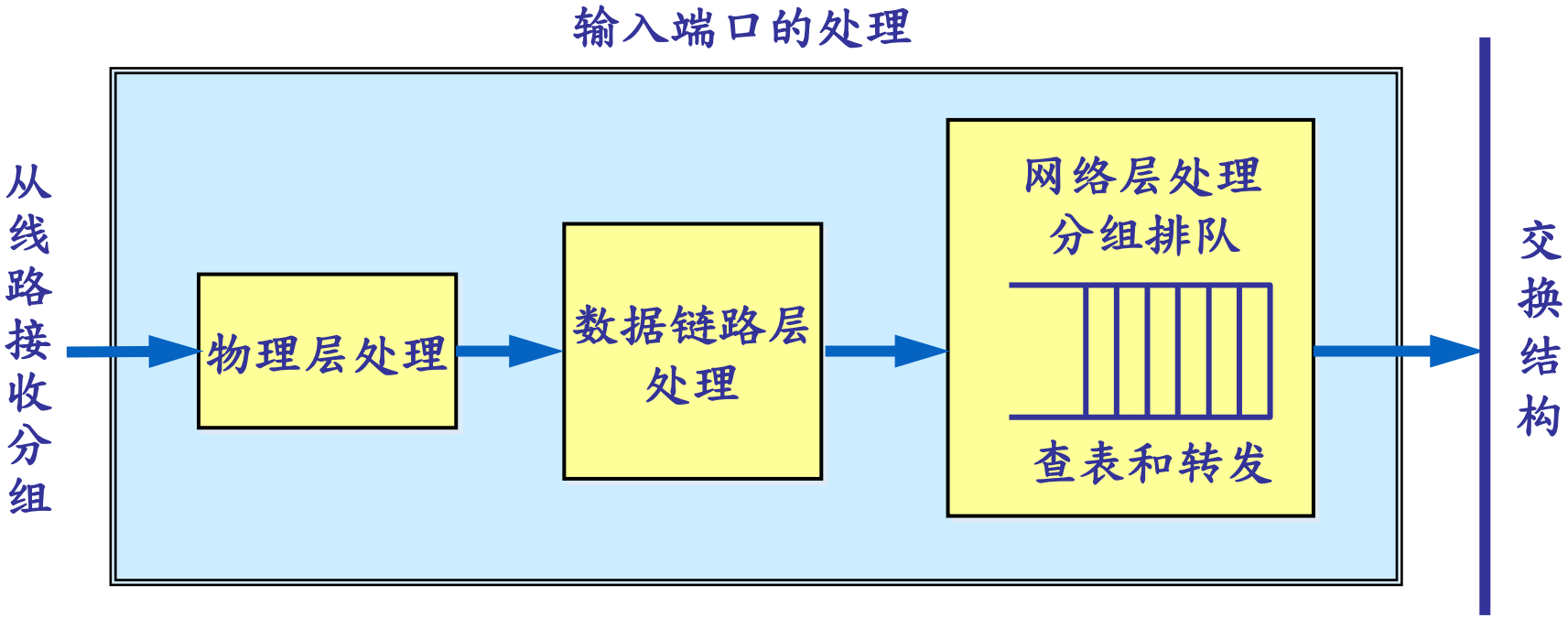

3——网络层
2——数据链路层
1——物理层
路由选择处理机
路由选择协议
路由表
路由选择
输入端口
输出端口
1
2
3
分组处理
转发表
3
2
1
输入端口
输出端口
1
2
3
交换结构
3
2
1
分组转发

“转发”和“路由选择”的区别

- 转发 (forwarding) ：从路由表得出的
  - 路由器根据转发表将 IP 数据报从合适的端口转发出去。
- 路由选择 (routing) ：根据路由选择算法得出
  - 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。
- 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

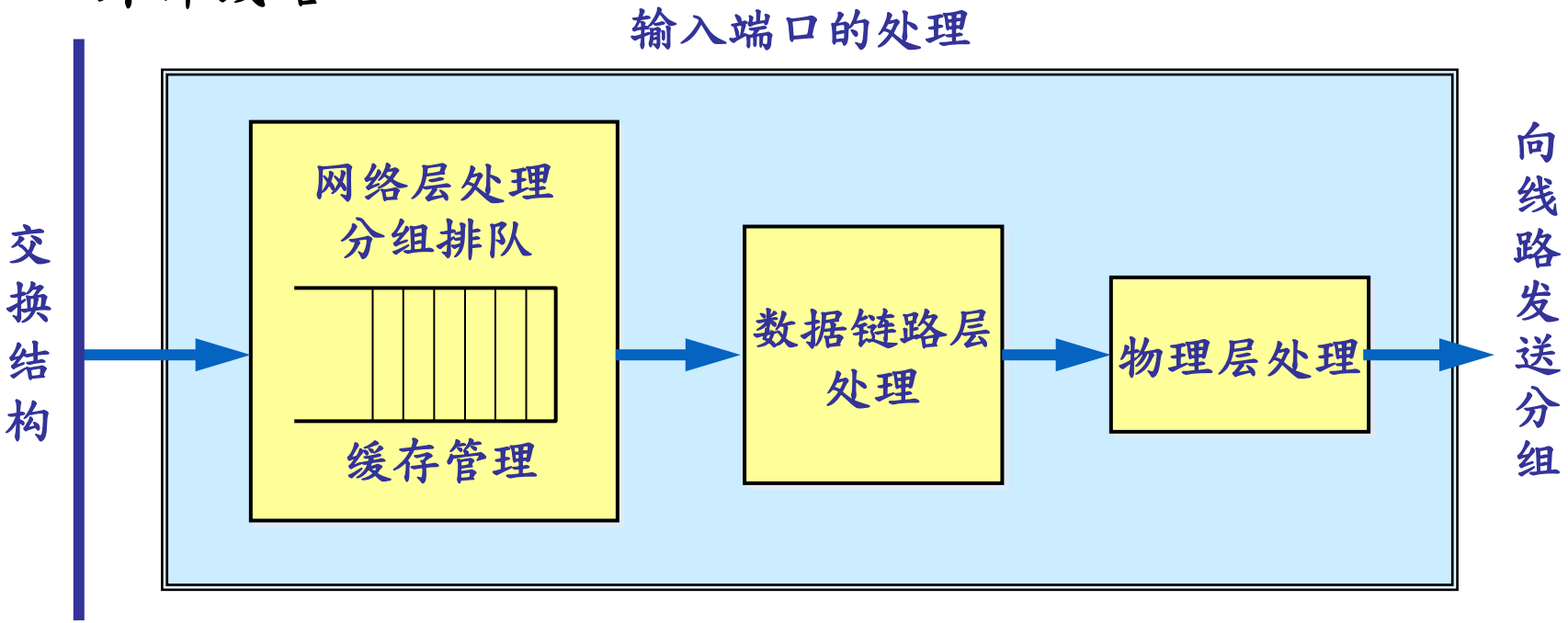
输入端口对分组的处理

- 输入端口对线路上收到分组的处理
  - 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。



输出端口将分组发送到线路

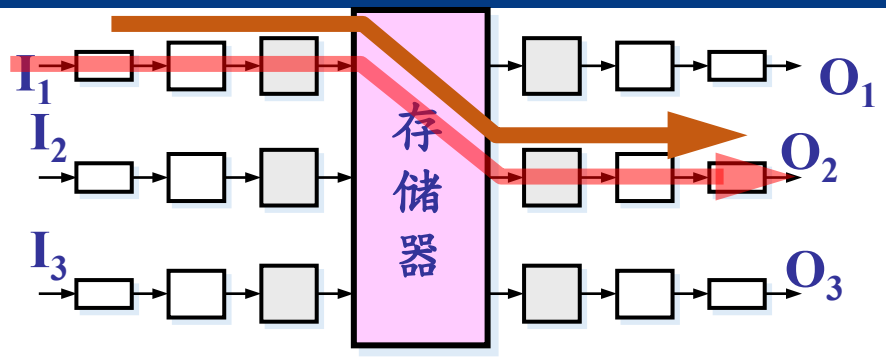
- 输出端口将交换结构传送来的分组发送到线路
  - 交换结构传送过来的分组先进行缓存。数据链路层处理模块将分组加上链路层的首部和尾部，交给物理层后发送到外部线路。



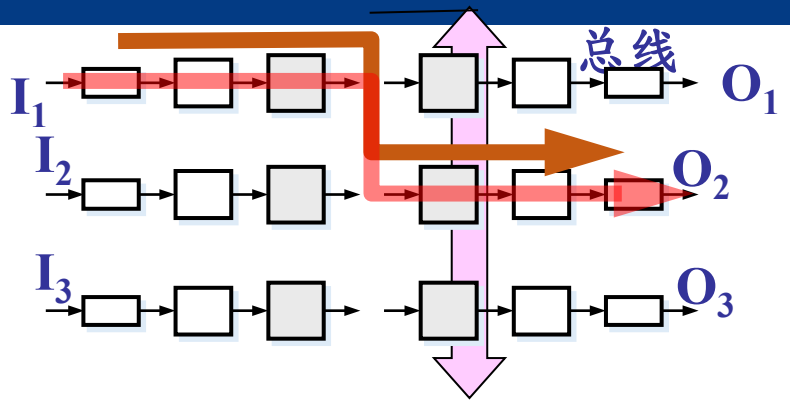
分组丢弃

- 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。

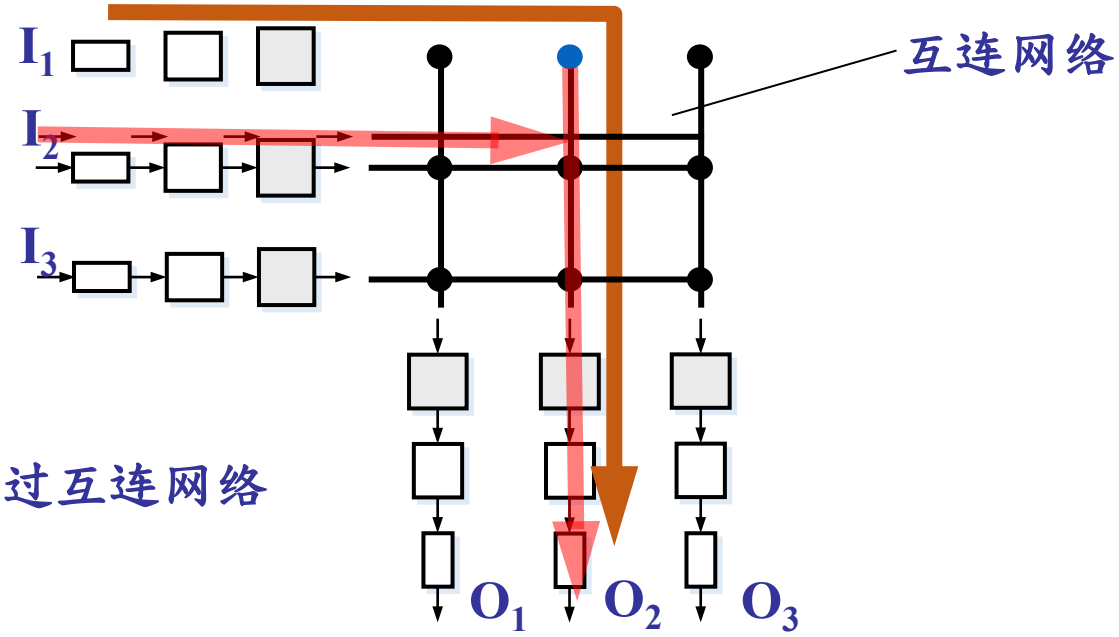
交换结构



(a) 通过存储器



(b) 通过总线



(c) 通过互连网络

计算机网络

Computer Network

13

谢谢观看

理论教程

厦門大學

XIAMEN UNIVERSITY

信息学院 黄 煒

(特色化示范性软件学院) 博士,副教授

School of Informatics Wei Huang