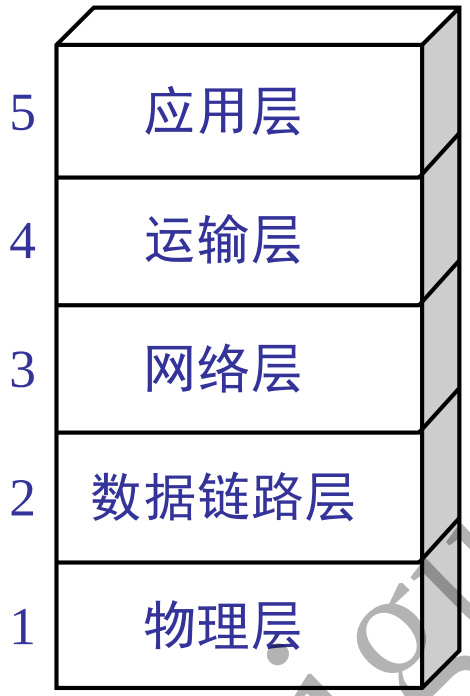
计算机网络

顾 军

计算机学院

jgu@cumt.edu.cn

专题4：数据包怎么在互联网中寻路和转发？



- 应用层(application layer)
- 运输层(transport layer)
- 网络层(network layer)
- 数据链路层(data link layer)
- 物理层(physical layer)

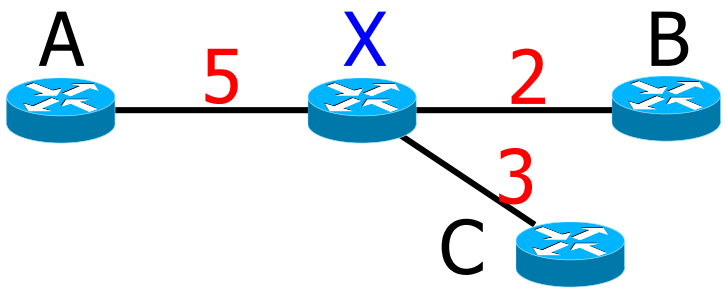
Q29：基于链路状态的OSPF协议？

- 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。
- OSPF 的原理很简单，但实现起来却较复杂。
- OSPF (Open Shortest Path First)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”。
  - OSPF是分布式的链路状态协议。

路由信息交换的三个要点

- ①交换什么：与本路由器相邻的所有路由器的链路状态，但只是路由器所知道的部分路由信息。
  - “链路状态”：
    - a.本路由器和哪些路由器相邻
    - b.与相邻路由器的链路的“度量”(metric)
      - “度量”可用来表示费用、距离、时延、带宽，等等，这些都可由管理人员决定。
- ②与谁交换：向本AS(区域)中所有路由器发送信息
- ③何时交换：只有当链路状态发生变化时

与相邻路由器的链路信息的组织



X（本节点的网络地址）	
SEQ（链路状态的序号）	
AGE（生存期）	
A(相邻节点)	5(链路状态)
B	2
C	3

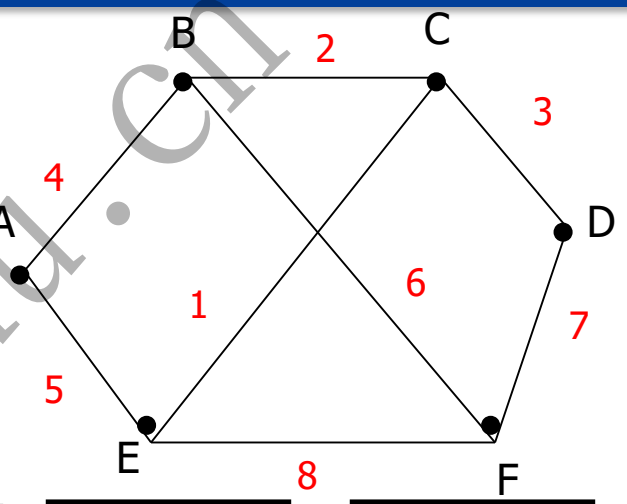
- 含到所有相邻节点的链路状态：
  - 路由器地址：A、B、C
  - 去往该节点的链路代价：5、2、3
- 序号（SEQ）：每次发送新的状态时加1。
  - 序号为32位，每10秒发一次，1370年不重复。
  - 序号越大，状态越新

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。

链路状态数据库LSDB (link-state database)

- 各路由器交换链路状态后：
  - 所有路由器都建立一个整个AS的LSDB
  - LSDB描述了整个AS的拓扑结构



A	
Seq	
Age	
B	4
E	5

B	
Seq	
Age	
A	4
C	2
F	6

C	
Seq	
Age	
B	2
D	3
E	1

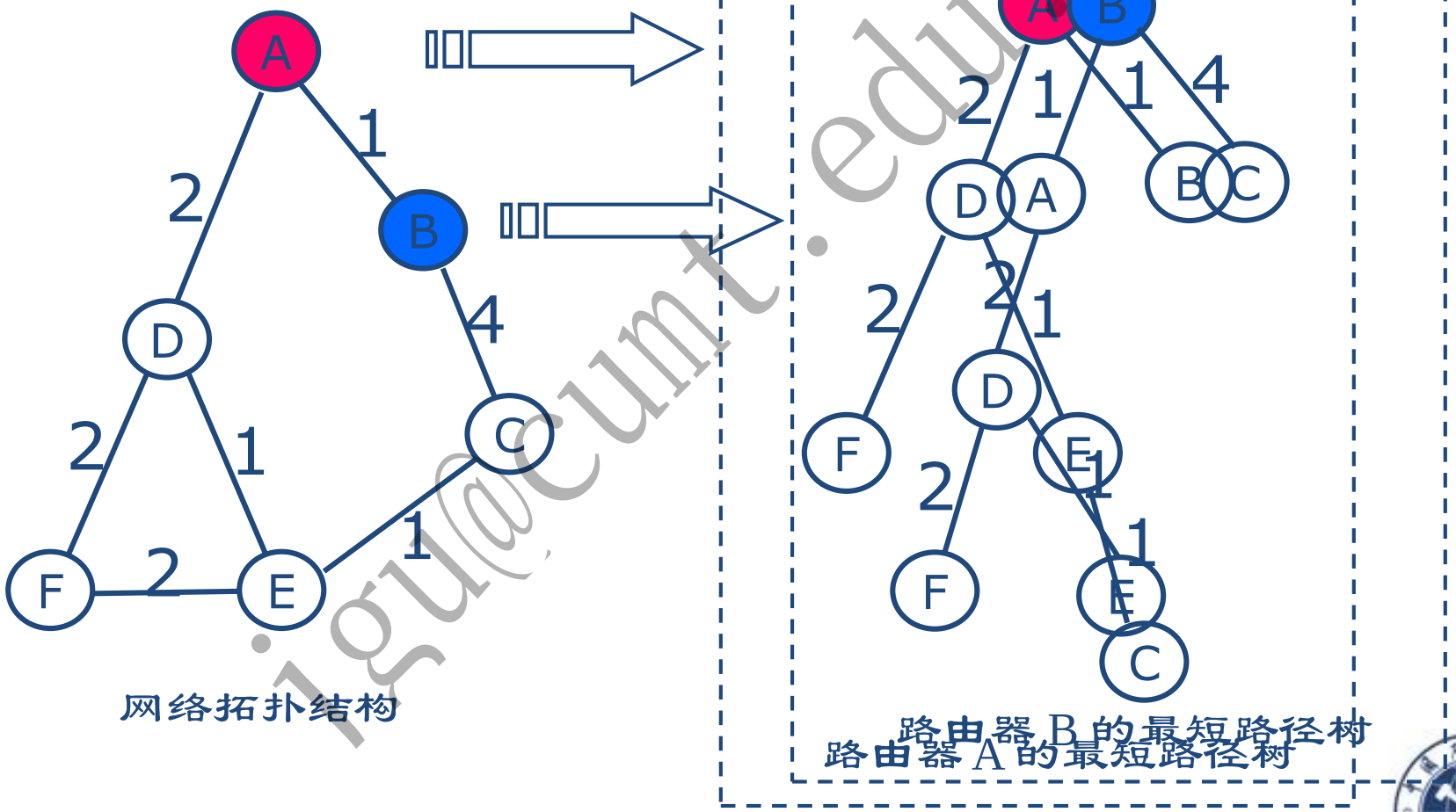
D	
Seq	
Age	
C	3
F	7

E	
Seq	
Age	
A	5
C	1
F	8

F	
Seq	
Age	
B	6
D	7
E	8

链路状态路由选择算法的步骤

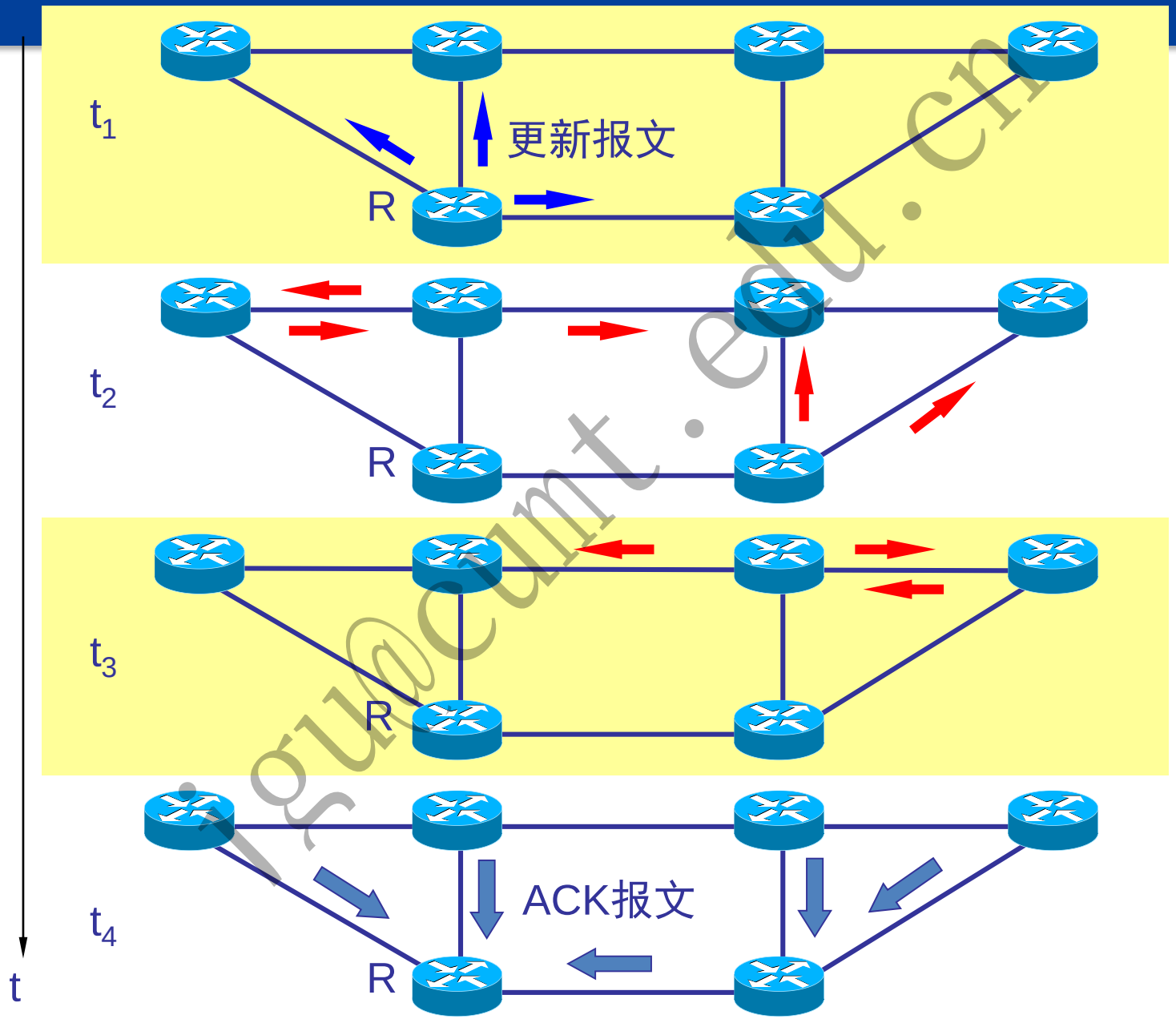
1. 找出所有可达的相邻结点及它们的网络地址；
2. 测定到这些相邻结点的代价（度量）；
3. 将以上信息构成链路状态分组（link state packet）；
4. 向网上所有结点发送链路状态分组；
5. 利用收到的链路状态分组计算到各目的结点的最短通路。



发送链路状态更新分组的方法

- 考虑到
  - OSPF 分组封装在 IP 数据报中传送。
    IP协议是不可靠的， OSPF要提供可靠机制
  - OSPF：向本AS中所有路由器发送路由信息
- 方法：使用扩散法发送链路状态更新分组
  （扩散= flooding=洪泛）
  - 向所有端口发送
  - 相邻的路由器继续转发。

OSPF 使用的是可靠的洪泛法



Flooding(洪泛,扩散法)

- 源站：向各个端口发送更新报文
- 收到更新报文的路由器

if 序号<＝已收到的相同源地址的LSP的最大序号
then　丢弃　//过时的、或重复的报文
else　//新的状态报文

{　1.接收，更新自己的数据库，
将新的链路状态信息发送给其他路由器。
2.记录关键字：源地址、序号(以后比较用)
3.向除输入端口外的各端口发送
4.向源站发送链路状态确认报文应答
}

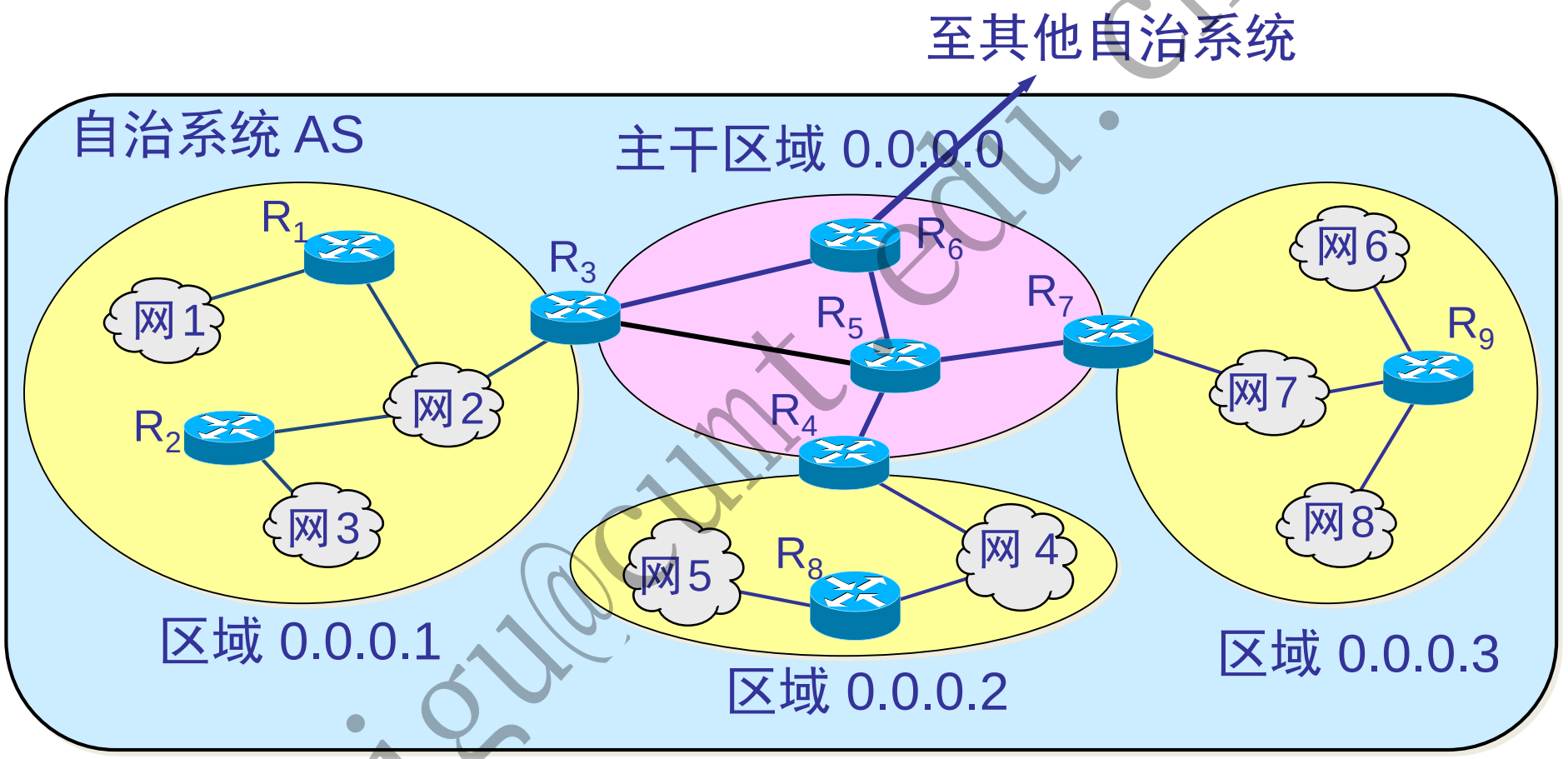
OSPF 的区域(area)

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

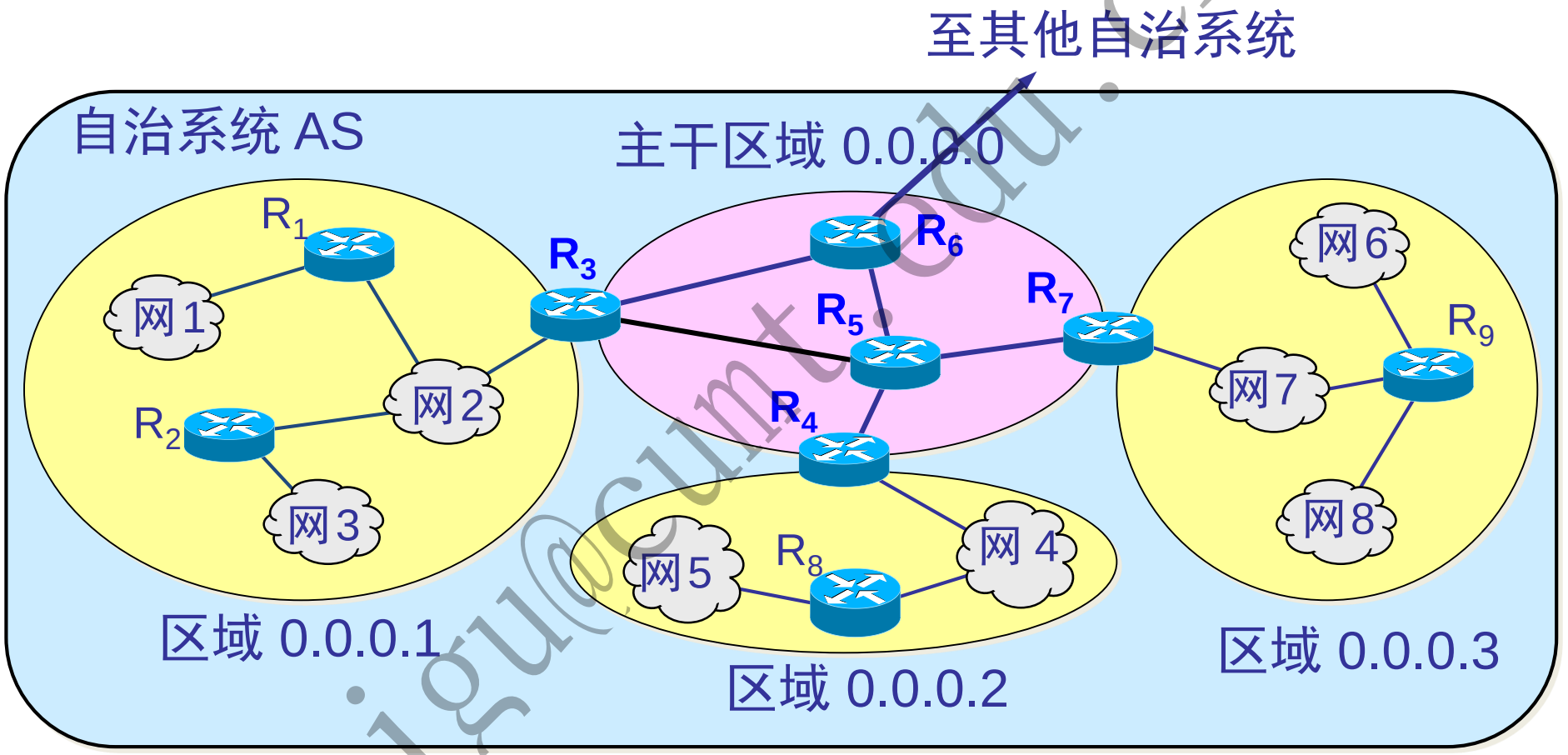
划分区域

- 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。
  - 一个区域可以是一个网络或一组相邻网络；
  - 一个网络只属于一个区域；
  - 在一个区域内的路由器最好不超过 200 个。
- OSPF 使用层次结构的区域划分。在上层的区域叫作主干区域(backbone area)。主干区域的标识符规定为0.0.0.0。主干区域的作用是用来连通其他在下层的区域。

OSPF 划分为两种不同的区域

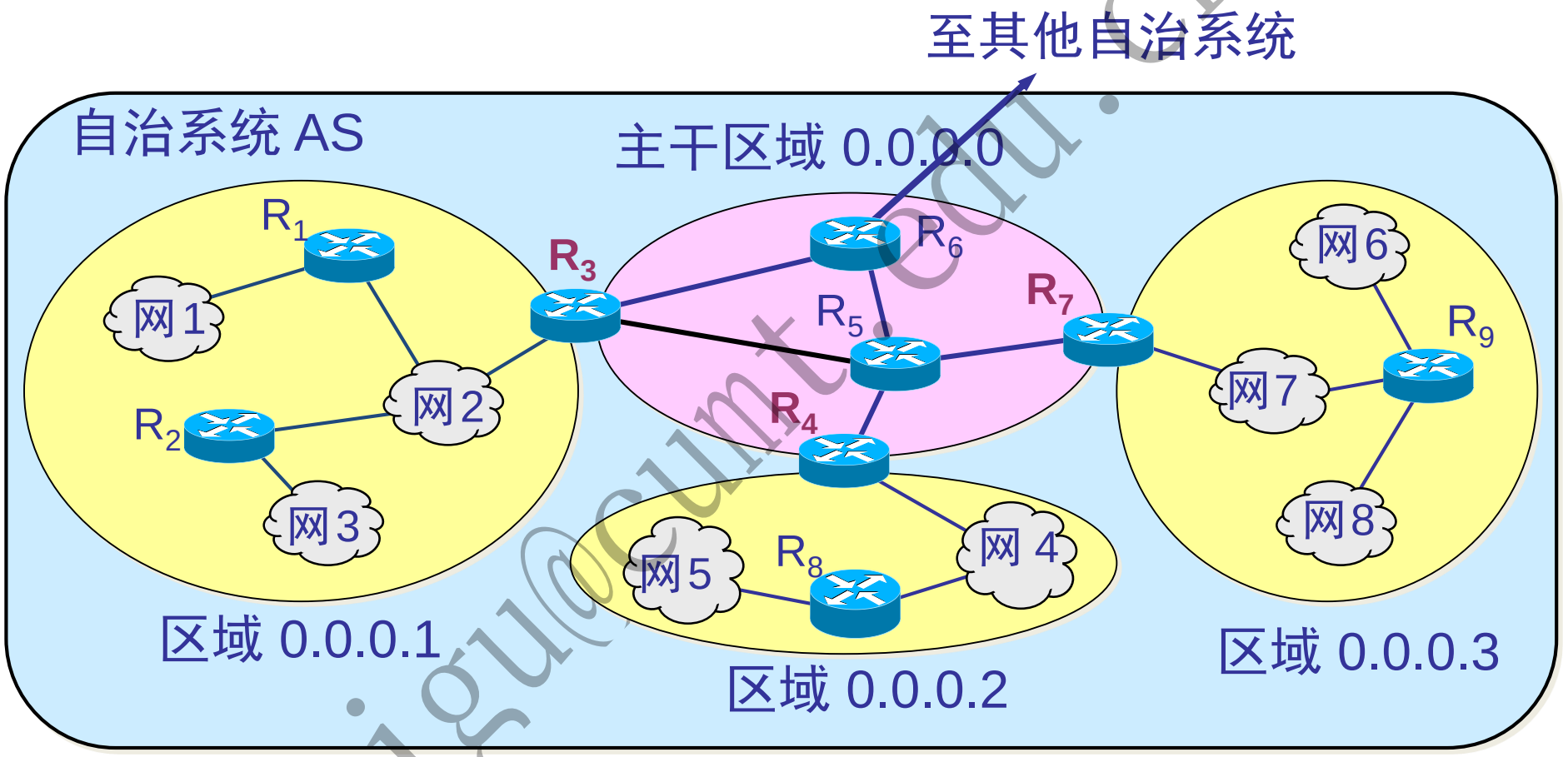


主干路由器



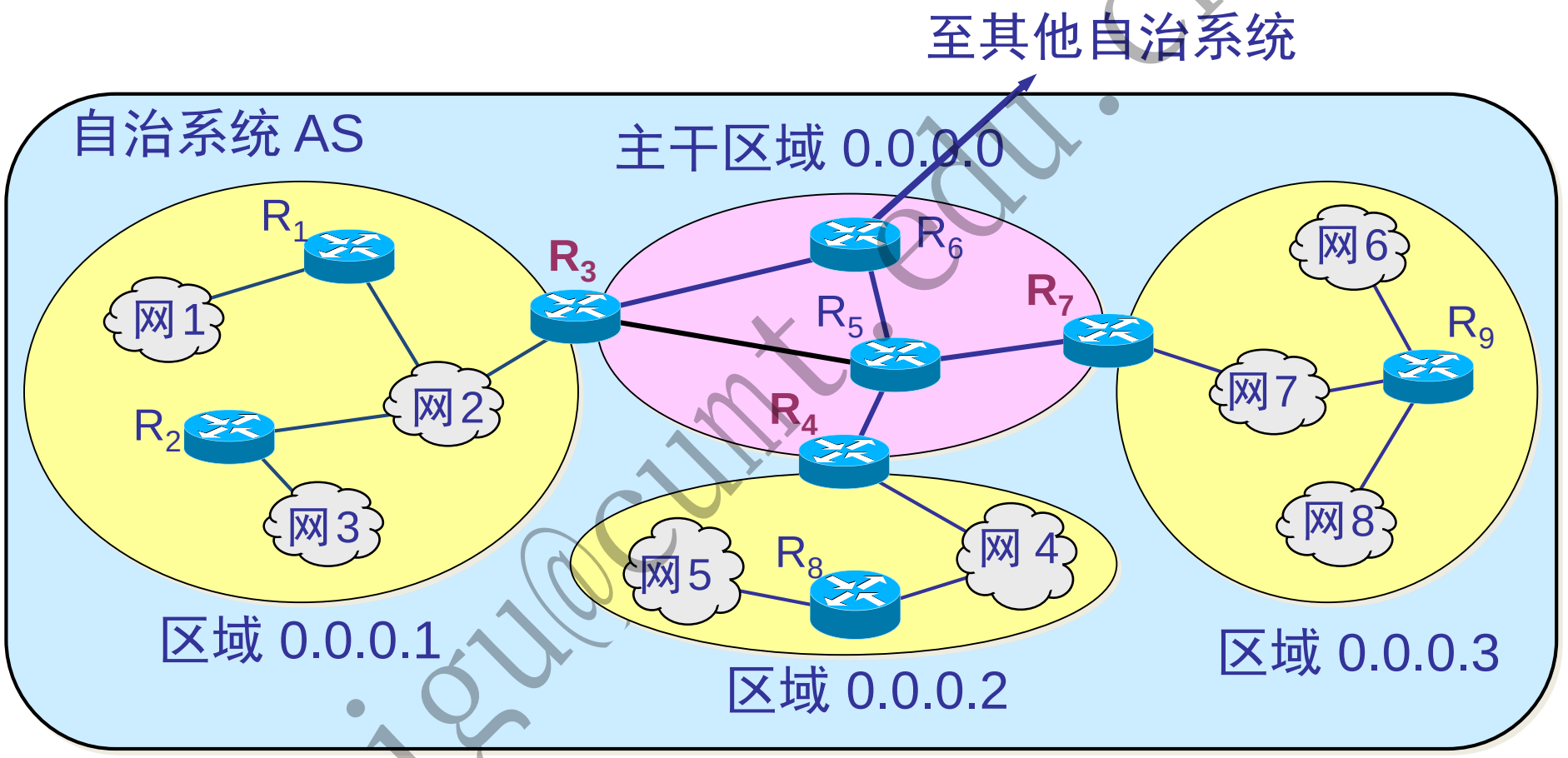
- 主干路由器：R_3，R_4，R_5，R_6，R_7

区域边界路由器



- 区域边界路由器： R_3， R_4， R_7

自治系统边界路由器



- 自治系统边界路由器： R_6

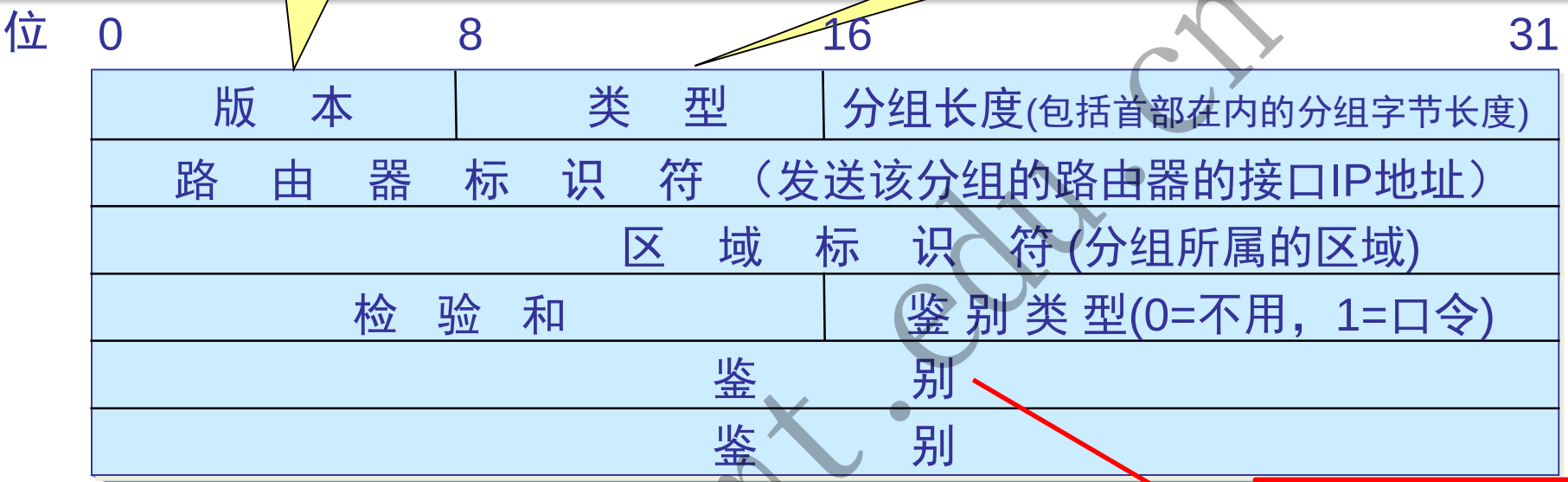
划分区域的优点

- 区域内部路由器
  - 仅与同区域的路由器交换链路状态
  - 只需知道本区域的完整网络拓扑
- 优点：
  - 减少了链路状态的通信流量（只在区域内广播）
  - 减少了链路状态信息库的表项；
  - 提高最短路径算法SPF的计算速度；
  - 简化管理；
  - 扩展能力远远超过RIP协议；
  - 使路由器的运行更快速、更经济、占用的资源更少。

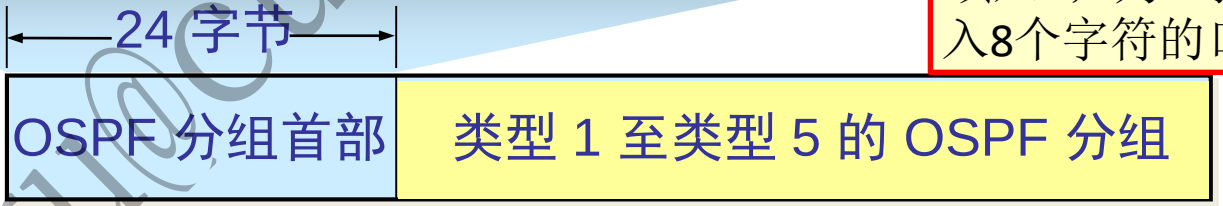
OSPF 分组

1 或 2

5种OSPF分组类型之一



鉴别类型为0时，填入0；为1时，填入8个字符的口令。



OSPF 的五种分组类型及其作用

- 类型1，问候(Hello)分组
  - 确认哪些相邻路由器是可达的；
    - 作用：建立和维护连接和邻居关系；
- 类型2、3、4、5：用于数据库同步
  保持整个区域内各路由器链路状态数据库一致。
  - 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。
  - 类型5，链路状态确认(Link State Acknowledgment) 分组。

(1) Hello分组：获取链路状态

- 路由器启动时的初始化：
  发现相邻结点，并获取其网络地址：
  - 向所有输出端口发Hello分组；
  - 对方给予响应。
- 维护相邻路由器的可达性：
  相邻路由器每隔10 秒交换一次Hello分组
  - 若40 秒没有收到某个相邻路由器的Hello分组，则认为不可达，do：
    - 修改链路状态数据库
    - 重新计算路由
    - 发送链路状态更新分组。

(2) 链路状态的同步（使用四种分组）

- 同步：
  使不同路由器的链路状态数据库的内容相同。
- 为减少路由器之间交换链路状态的数量，采取了以下方法：
  - 通过数据库描述分组定期向邻站发送“摘要信息”：
    - 本站数据库中有哪些链路、链路状态序号等；
  - 路由器收到的数据库描述分组，若发现本地数据库中缺少某些链路，或本地数据库中该链路状态的序号小于数据库描述分组中的序号，就向该邻站发送链路状态请求分组；
  - 收到链路状态请求分组，用链路状态更新分组广播发送这些链路的详细状态信息；
  - 收到链路状态更新分组，返回链路状态确认分组

发送链路状态更新分组的几种情况

- 本地链路状态发生变化

  （链路状态序号+1,再发送更新分组）

- 收到链路状态请求分组

  （发送的更新分组的链路状态序号不变）

- 30分钟定时刷新一次数据库中的链路状态

  （发送的更新分组的链路状态序号不变）

OSPF的基本操作

确定可达性

问候

问候

每10秒交换一次问候分组；若40秒没收到，则不可达。

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

同步：不同路由器的链路状态数据库的内容是一样的。

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

指定的路由器DR (designated router)

- 多点接入的局域网中，若N个路由器连接在一个以太网上，则每个路由器要向其他(N-1)个路由器发送那个链路状态信息，因而共有(N-1)^{2}个链路状态要在这个以太网上传送。
- 为了避免路由器之间建立的完全的邻接关系而引起大量的开销，OSPF要求在多路访问的网络中要选举一个DR，每个路由器都要与这个DR路由器建立邻接关系。
- 在这个网络中，每个路由器都要与这个DR路由器交换路由信息，由这个DR路由器负责通知其它的路由器，告知整个网络的拓扑。

指定路由器(DR)和非指定路由器(DR OTHER)

- 在一个OSPF的网络中，所有的路由器将被分为两类：指定路由器(DR/BDR)和非指定路由器(DR OTHER)。
- 所有的非指定路由器都要和指定路由器建立邻居关系，并且把自己的LAS发送给DR，而其他的OSPF路由器将不会相互之间建立邻居关系。
- 也就是说，OSPF网络中，DR/BDR的LSDB(链路状态数据库)将会包含有整个网络的完整拓扑。
- DR根据Router-ID和优先级Priority来进行选举，如果优先级相同，则根据ID的大小来选择。

OSPF的特点-1

- OSPF允许管理员给每条路由指派不同的代价，因此对于不同类型的业务可计算出不同的路由。
  - 例如：高带宽的卫星链路对于非实时的业务可设置为较低的代价，但对于时延敏感的业务就可设置为非常高的代价。
  - 链路的代价可以是1至65535中的任何一个无量纲的数。
  - 商用的网络在使用OSPF时，通常根据链路带宽来计算链路的代价
- 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，叫作多路径间的负载平衡。
  （RIP：到一个目的网络只有一条的路径）

OSPF的特点-2

- 所有在OSFP路由器之间交换的分组（例如，链路状态更新分组）都具有鉴别的功能，因而保证了仅在可信赖的路由器之间交换链路状态信息。
- 支持可变长度的子网划分和无分类编址 CIDR。
- 由于网络中的链路状态可能经常发生变化，因此OSPF让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规定，链路状态序号增长的速率不得超过每5秒钟1次。这样，全部序号空间在600年内不会产生重复号。

OSPF与RIP的比较

- OSPF是分布式、动态、基于链路状态的IGP。
  - “链路状态”
    - 本路由器和哪些路由器相邻
    - 该链路的“度量”(metric)
- 当自治系统很大时，OSPF 比 RIP更新收敛得快
  - OSPF 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。使各个路由器能及时更新其路由表。
- 没有“坏消息传播得慢”的问题
  - 据统计，其响应网络变化的时间小于 100 ms
- OSPF采用IP作为通信协议，RIP基于UDP转发

Q30：AS之间如何交换路由信息？

- 内部网关协议（RIP或OSPF）主要是设法使数据报在一个AS中尽可能有效地从源站传送到目的站。
- 在一个AS内部也不需要考虑其他方面的策略。
- 但是，OSPF或RIP这样的域内路由协议(IGP)不能用于AS之间的路由选择。

IGP协议不能用于AS间路由

- 第一，因特网的规模太大，使得自治系统之间路由选择非常困难，要寻找最佳路由是不现实的。
  - RIP只适合小规模网络
  - 主干网络路由器中路由表的项目数早已超过5万个网络前缀。如果使用OSPF，则每个路由器都要维持一个很大的链路状态数据库，而且用Dijkstra算法计算最短路径要花费太长时间

技术上做不到

IGP协议不能用于AS间路由

- 第二，自治系统之间的路由选择必须考虑有关策略，即人为因素。
  - AS各自运行自己选定的内部路由选择协议，并使用本AS指明的路径度量。因此，当一条路径通过几个不同 AS (路由协议不同，路径度量方式不同)时，要想对这样的路径计算出有意义的代价是不太可能的。
    - 例如，对某AS来说，代价为1000可能表示一条比较长的路由，但是对另一AS代价为1000却可能表示不可接受的坏路由。

事实上做不了

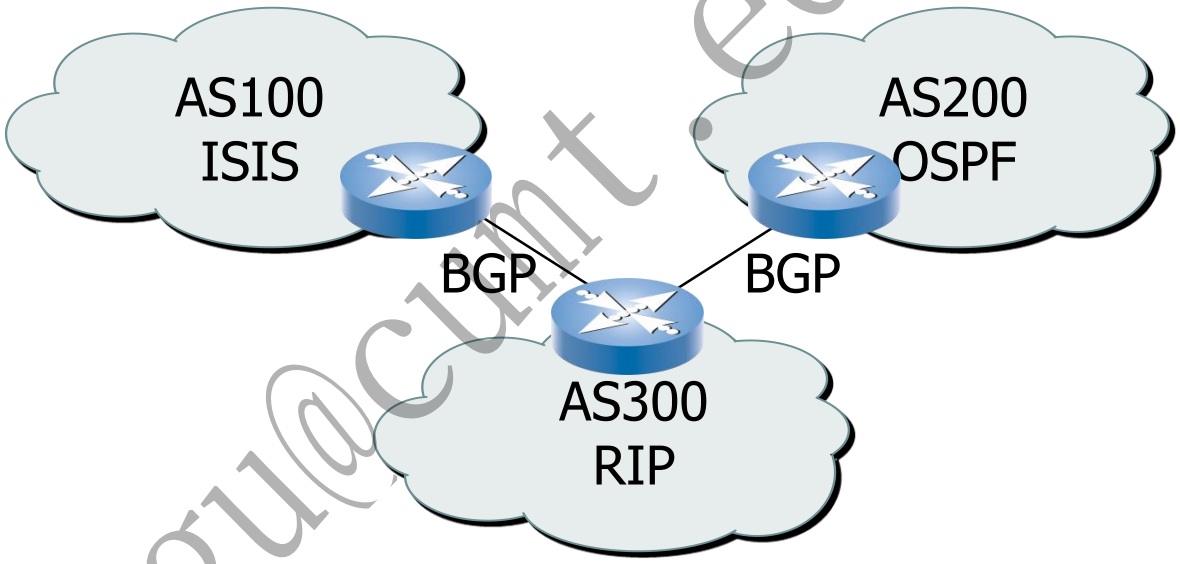
IGP协议不能用于AS间路由

- 最短距离（即最少跳数）未必一定合适
- 有的路径的使用代价很高或很不安全
- 有些AS可能不愿意同意借路
- 自治系统之间的路由选择协议应当允许使用多种路由选择策略，这些策略包括政治、安全或经济方面的考虑。

- 比较合理的做法是在 AS 之间交换“可达性”信息（即“可到达”或“不可到达”）。
  - 例如，告诉相邻路由器：到达目的网络N可经过AS_X。

BGP协议

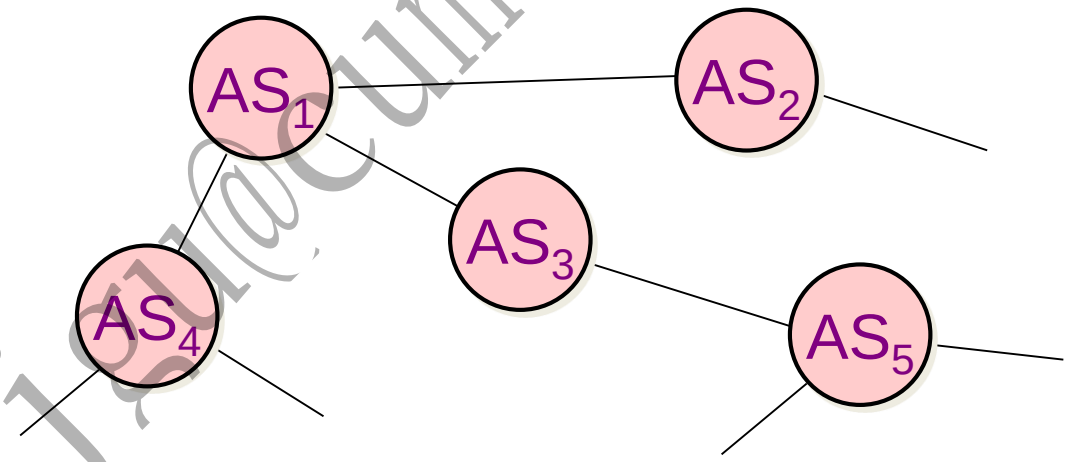
- 边界网关协议 BGP是唯一一个用来处理像因特网大小的网络的协议，也是唯一能够妥善处理好不相关路由域间的多路连接的协议。



- ✓ 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。

AS 连通图

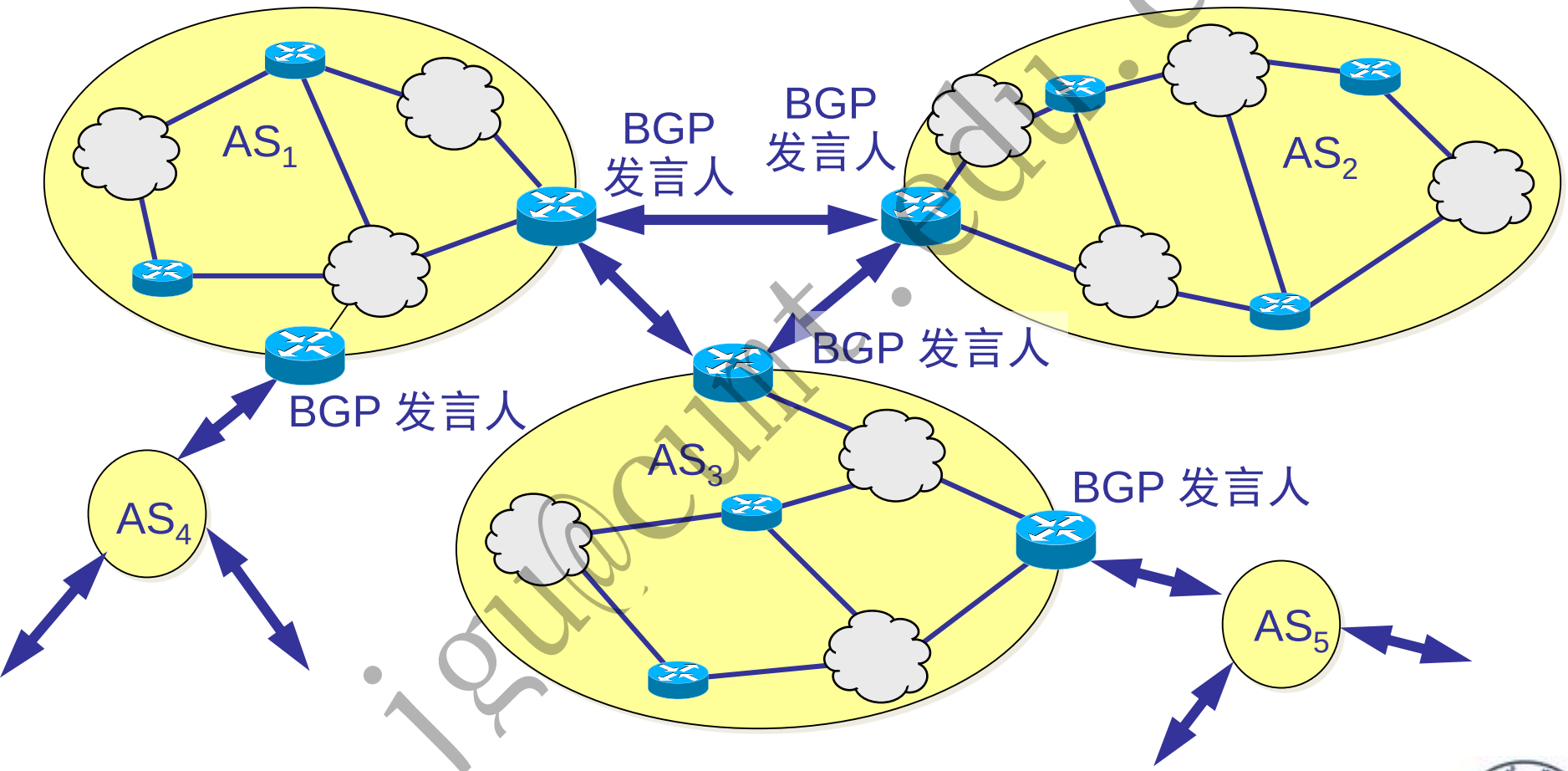
- BGP 是不同自治系统的路由器之间交换路由信息的协议，主要功能是和其他的 BGP 系统交换网络可达信息。
- BGP 所交换的网络可达性的信息就是要到达某个网络(用网络前缀标识)所要经过的一系列 AS。



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 发言人和自治系统 AS 的关系



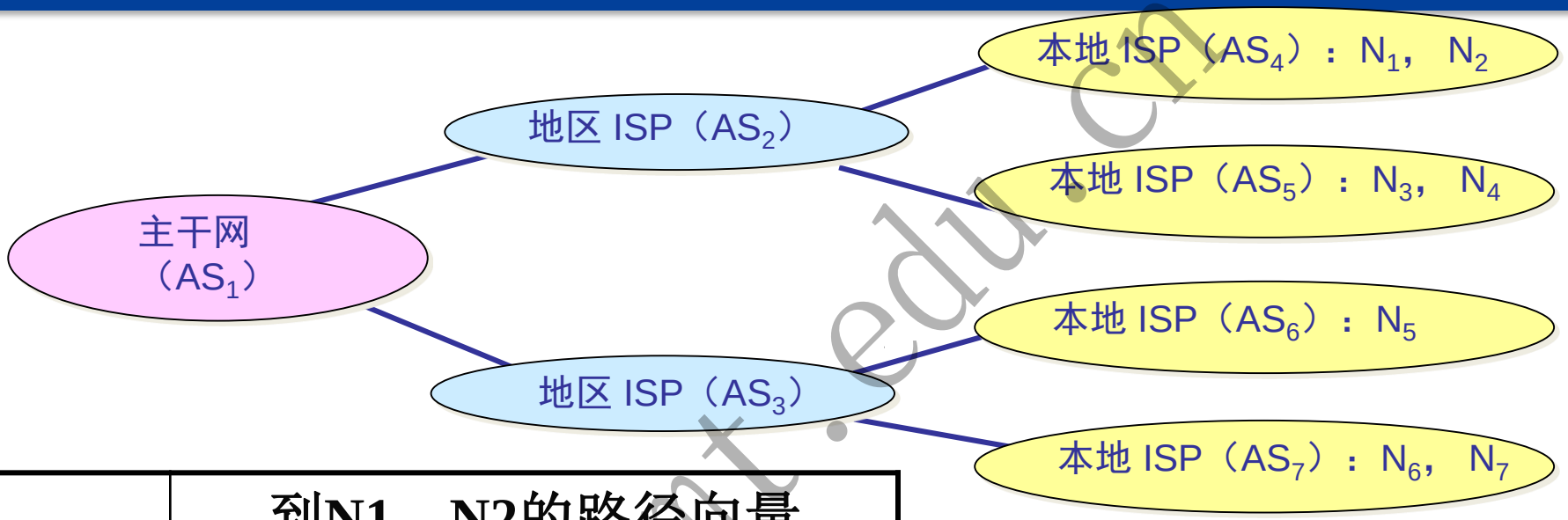
三个自治系统中的五个BGP发言人

路径向量(path vector)路由选择协议

- BGP采用了路径向量(path vector)路由选择协议，它与距离向量协议(如RIP)和链路状态协议(如OSPF)都有很大区别。
- 路径向量（路由所经过的各个AS序列）

  AS1＝》AS2＝》AS4
- 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。

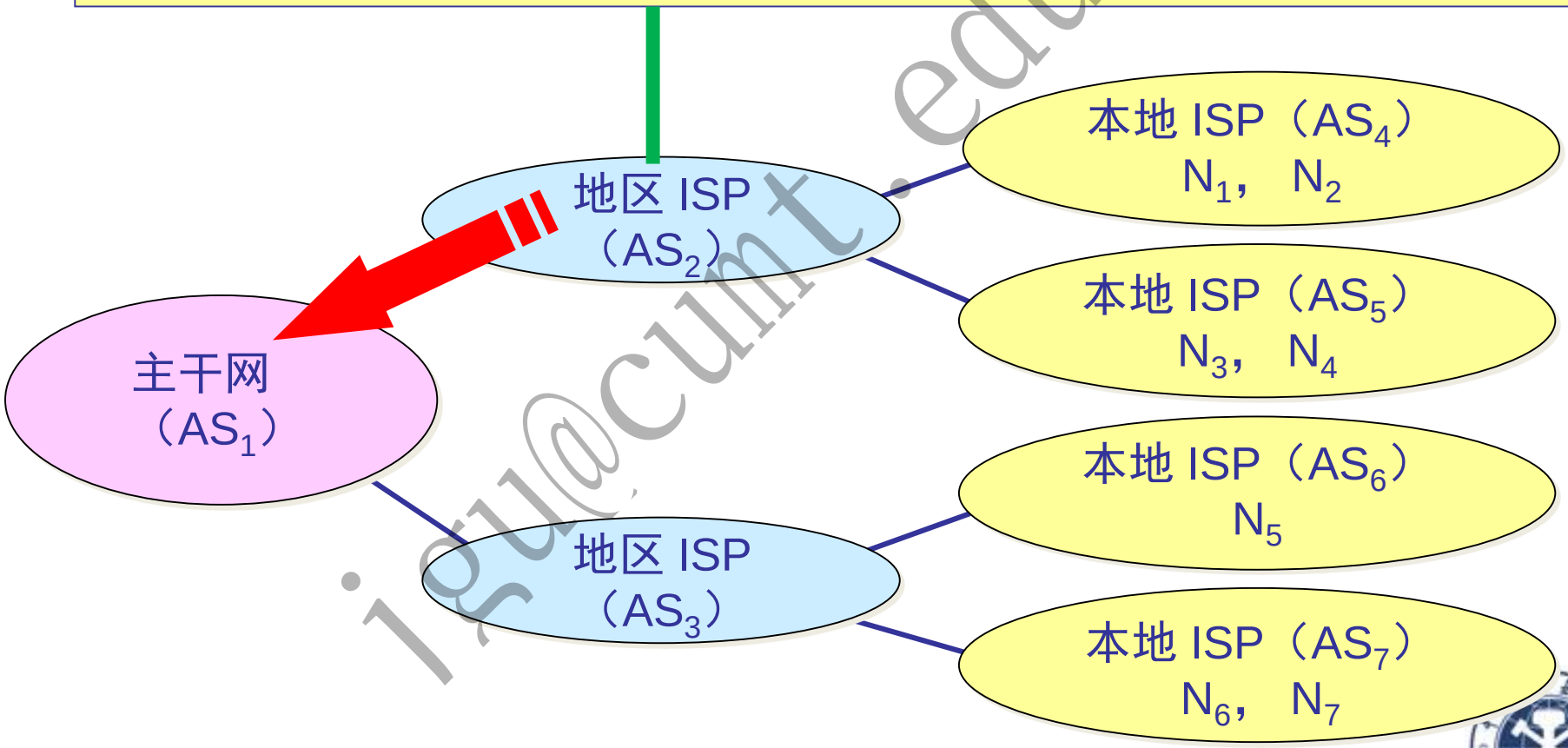
各个AS到网络N1、N2的路径向量



	到N1，N2的路径向量
AS1	AS2，AS4
AS2	AS4
AS3	AS1，AS2，AS4
AS5	AS2，AS4
AS6	AS3，AS1，AS2，AS4
AS7	AS3，AS1，AS2，AS4

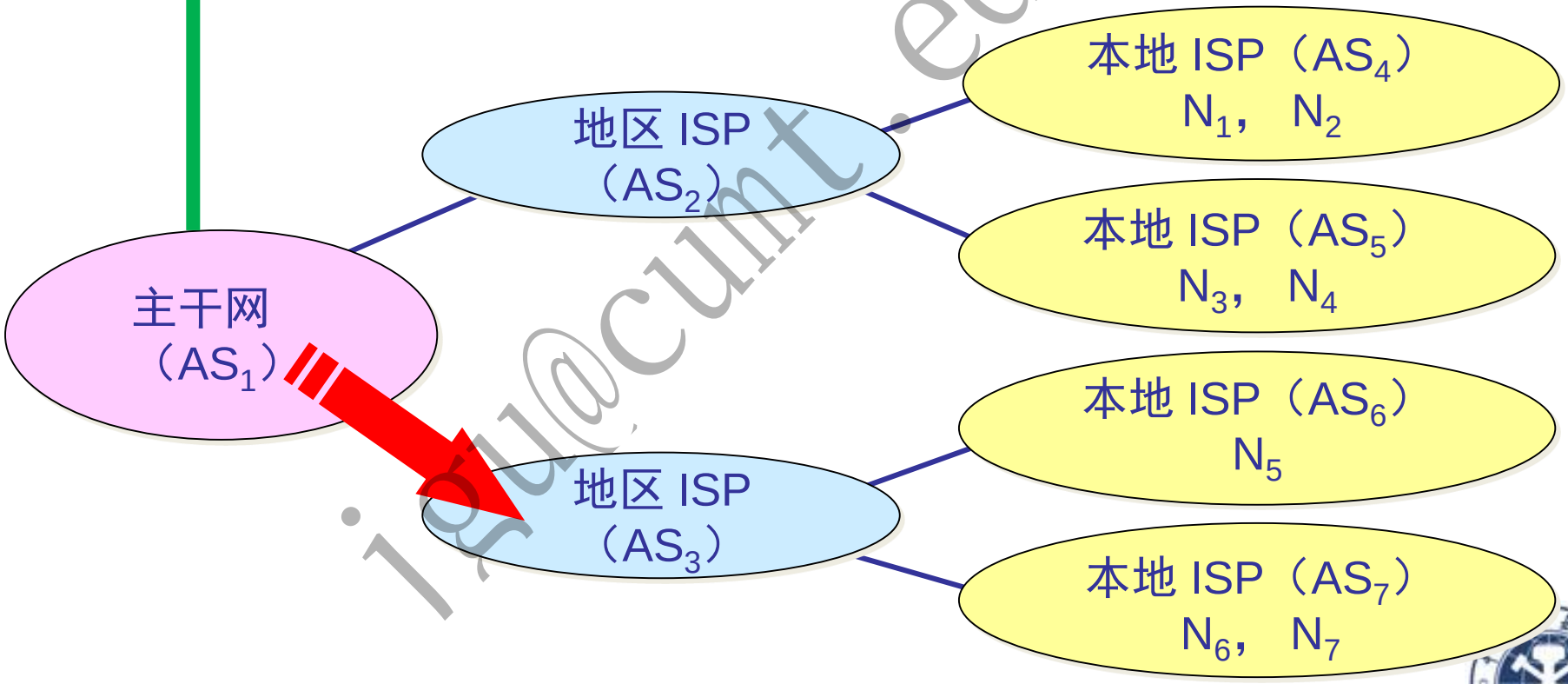
BGP 发言人交换路径向量

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。”



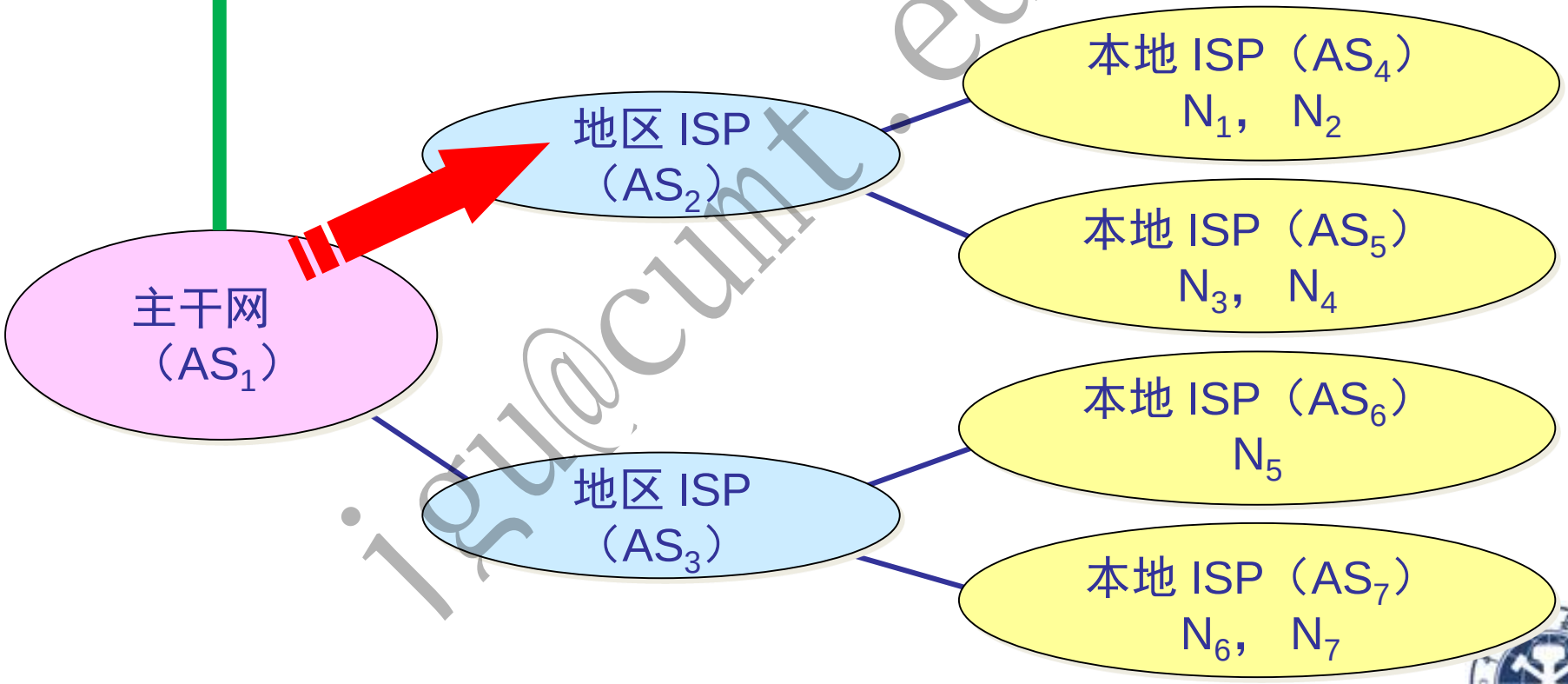
BGP 发言人交换路径向量

主干网通知AS3：
“经过 AS_1, AS_2可到达网络 N_1, N_2, N_3和 N_4”



BGP 发言人交换路径向量

主干网可通知AS_2：“要到达网络 N_5, N_6 和 N_7 可沿路径（AS_1, AS_3）。”

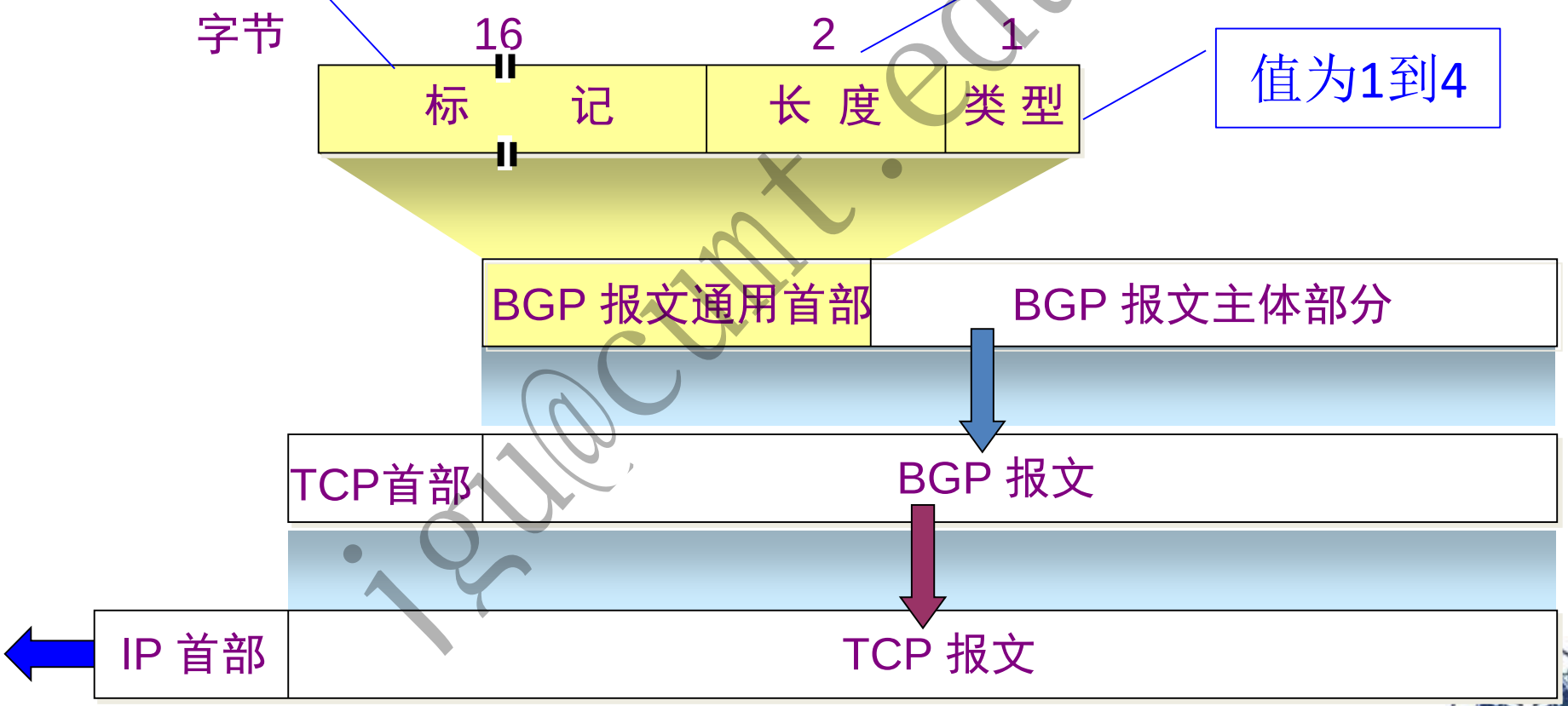


BGP 报文具有通用的首部

用来鉴别收到的BGP报文，不使用时置为全1

指出包括通用首部在内的整个BGP报文长度，以字节为单位，最小值为19，最大值是4096

值为1到4



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，使通信初始化。

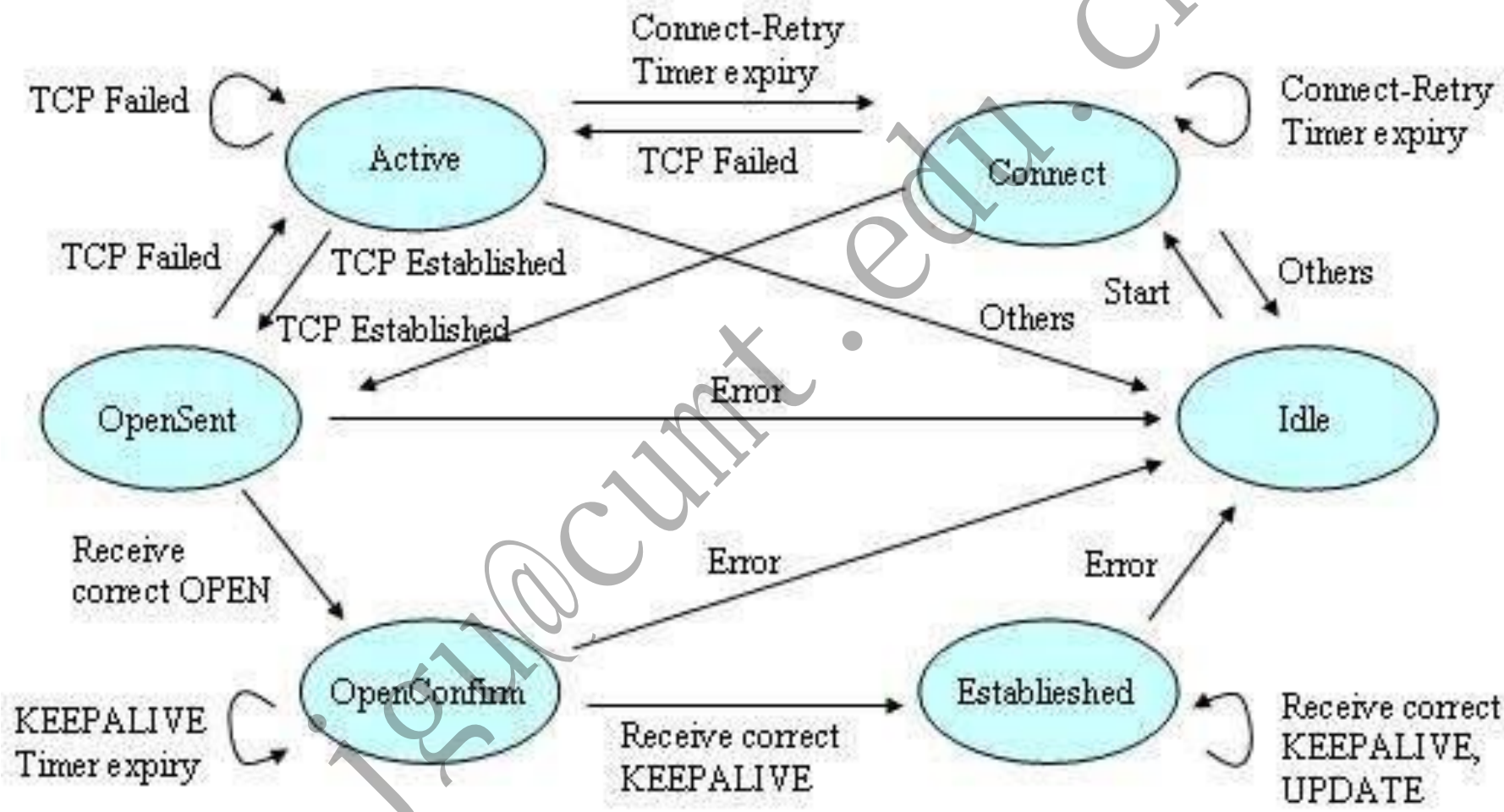
(2) 更新(UPDATE)报文，用来通告某一路由的信息，以及列出要撤消的多条路由。

(3) 保活(KEEPALIVE)报文，用来周期性地证实邻站的连通性。一般每隔30秒交换一次。只有19个字节（只用BGP报文的通用首部）。

(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基于TCP协议，端口号为179



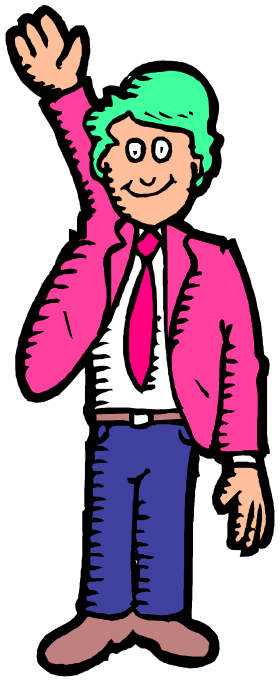
BGP、OSPF与RIP的比较

与RIP、OSPF比较，了解BGP特点

- 追求的目标不同：
  - RIP：跳数最少（最优）
  - OSPF：代价最小（最优）
  - BGP：能够到达
    只求寻找一条能够到达目的网络且比较好（不能兜圈子）的路由
- 策略不同：
  - RIP：　距离向量(到各个网络的跳数)
  - OSPF：链路状态(各链路的代价/时延等)
  - BGP：　路径向量(经过的AS序列)

BGP、OSPF与RIP的比较

协议	RIP	OSPF	BGP	
类型	内部	内部	外部	
路由算法	距离—向量	链路状态	路径-向量	
传递协议	UDP 端口号 520	IP 协议号 89	TCP 端口号 179	
路径选择	跳数最少	代价最低	较好，非最佳	
交换结点	和本结点相邻的路由器	网络中的所有路由器	和本结点相邻的路由器	
交换内容	当前本路由器知道的全部信息，即自己的路由表	与本路由器相邻的所有路由器的链路状态	首次	整个路由表
			非首次	有变化的部分



THANK YOU!