过渡页

Transition Page

自治系统与路由选择协议分类

内部网关协议RIP

内部网关协议OSPF

外部网关协议BGP

网络层关注的是可达性问题。
其中的每个协议都解决了一个子问题。

终端如何获取IP地址？

DHCP

Debugging

ICMP

ping

traceroute

终端如何同外网的其他终端通信？

Routing protocols

如何根据IP地址获取MAC地址？

ARP

Broadcast

Gateways

NAT

OSPF, RIP, BGP

IPv6

南京郵電大學 Nanjing University of Posts and Telecommunications

PRINCETON

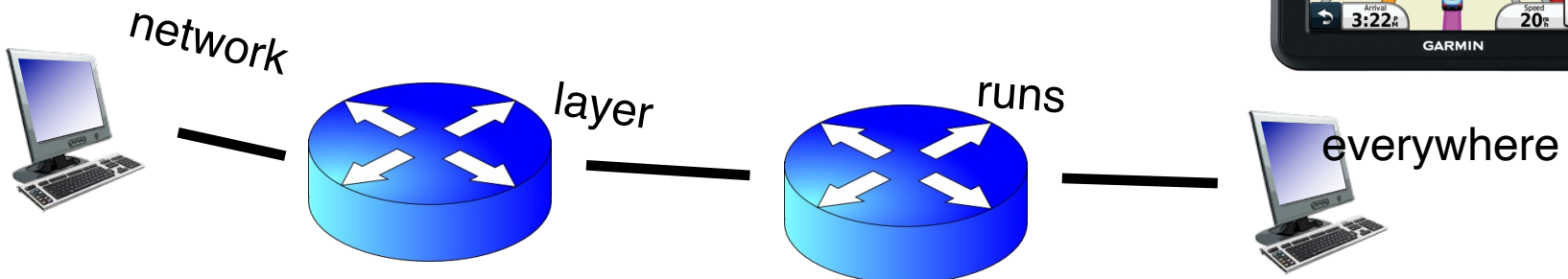
网络层关键功能

- 转发 (data plane): 将数据包从路由器的输入移动到相应的路由器输出
- 路由 (control plane):确定数据包从源到目的地所采取的路径

比如: 公路旅行

- 转发: 每一个岔路口该怎么走

- 路由: 从出发地到目的地整体规划



Routing is a fundamental problem in networking.

如何设计一个用于在互联网上导航的“高德地图”?

路由

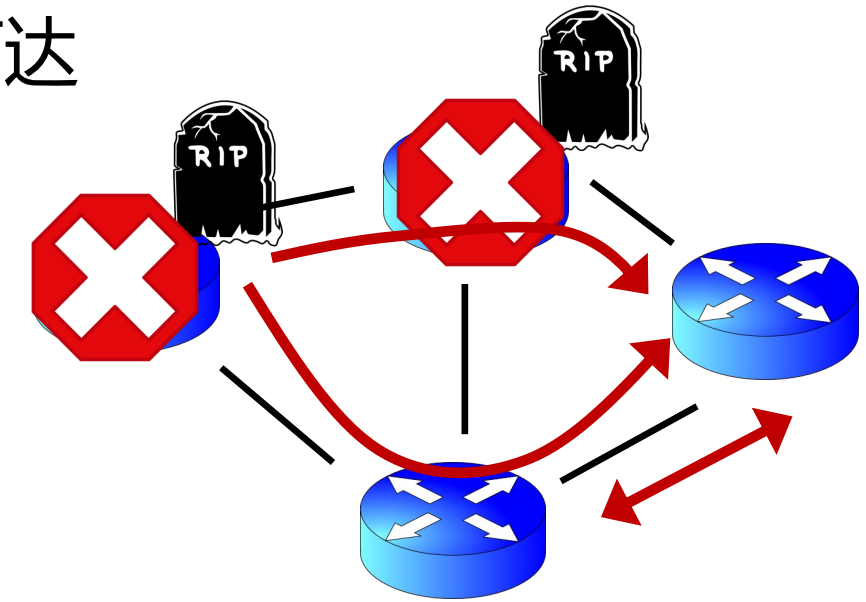
路由协议的目标#1

- 从源点到目的地的最优路径
- “最优” = 开销最小
  - 最小传播时延
  - 单位带宽最小开销 (e.g., $ per Gbit/s)
  - 最小拥塞(workload-driven)

- “路径” =一系列的路由器端口(links)

路由协议的目标#2

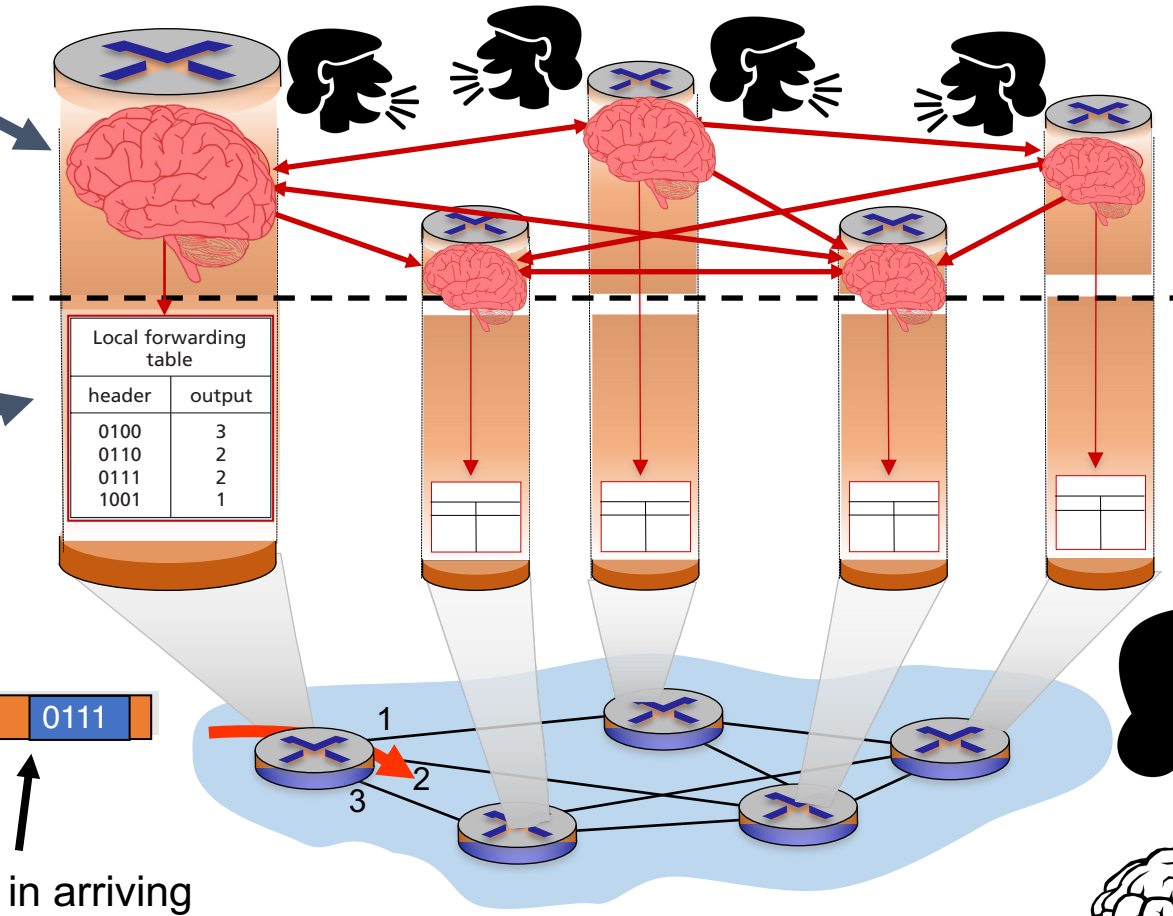
- 使网络能够抵抗网络故障
- 路由器和链路可以发生故障而不会使整个网络崩溃
- 整个子网可能无法访问，但其余网络仍然可达
- 因此，协议必须是分布式的



路由器控制面（control plane）

分布式

control plane:每个路由器中的组件相互交互，以产生路由结果。

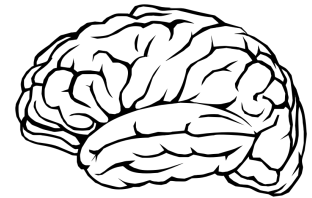


数据面（**Data plane**）

每个数据包的处理，将数据包从输入端口移动到输出端口。

路由协议

Q1. What info exchanged?



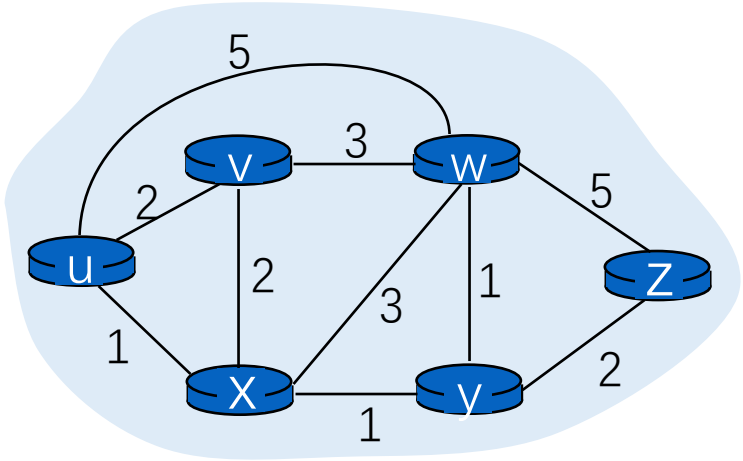
Q2. What computation?

图抽象

- 路由算法在网络的抽象表示上运作: 图抽象

Ex: 南邮校园

u: 通院
v: 人文院
…



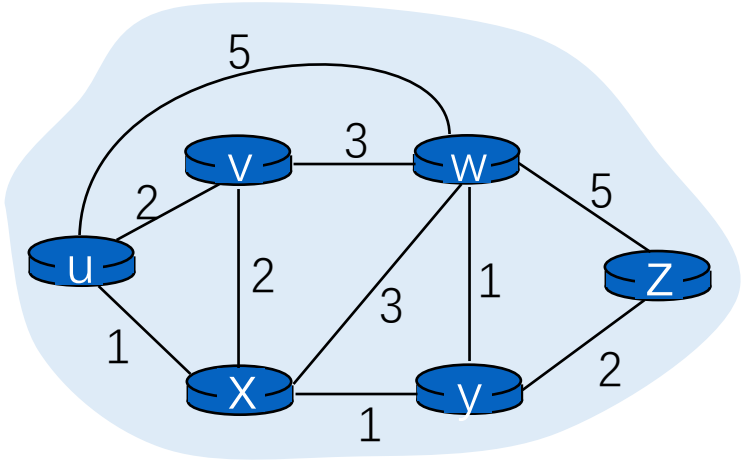
- 图中的每个节点都是一个路由器
- 图中的每条边都是一条链路
- 每条边有一个权重 (also called link metrics)。网络管理员来设置。

图抽象

- 路由算法在网络的抽象表示上运作: 图抽象

Ex: 南邮校园

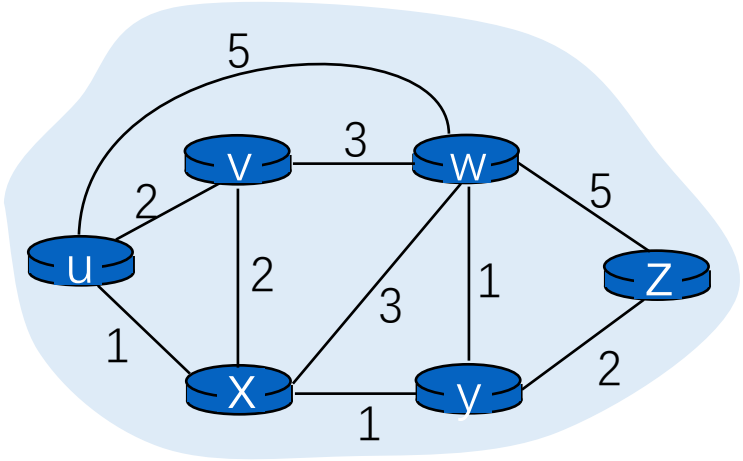
u: 通院
v: 人文院
…



- G = (N, E)
- N = {u, v, w, x, y, z}
- E = { (u,v), (u,x), (v,x), (v,w), (x,w), (x,y), (w,y), (w,z), (y,z) }

图抽象

- 每条边的开销: c(x, y)
  - Examples: c(u, v) = 2, c(u, w) = 5
- 每条路径的开销= 每条边开销之和
  - c(path x → w → y → z) = 3 + 1 + 2 = 6



- 路由结果:每个节点应确定到每个其他节点的最小开销路径。
- Q1:为了进行这个计算，节点之间应该交换什么信息呢?
- Q2:每个节点应该运行什么算法来计算到每个节点的最小成本路径呢?

路由协议

链路状态（Link state）protocols

Each router has complete information of the graph

Messages exchanged by flooding all over the network

Communication expensive, but complete

距离向量（Distance vector） protocols

Each router only maintains distances & next hop to others

Messages are exchanged over each link and stay within the link

Communication cheap, but incomplete

距离向量和链路状态

使用距离向量路由选择协议时，路由器依赖于邻居的路由选择策略。路由器并不完全了解网络拓扑。

- **E.g.， Bellman-Ford algorithm，RIP**

使用链路状态路由选择协议时，每台路由器都完全了解网络拓扑，能够根据网络拓扑信息独立地做出决策。

- **E.g.， Dijkstra’s algorithm**

RIP

5.5.2 内部网关协议RIP

路由信息协议 RIP (Routing Information Protocol)是一种**分布式的基于距离向量（Distance Vector）算法的路由选择协议**。

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。

5.5.2 内部网关协议RIP

距离

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。

从一个路由器到直接连接的网络的距离定义为 1 or 0。

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。

距离向量算法特点

5.5.2 内部网关协议RIP

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协议三要素

5.5.2 内部网关协议RIP

运行RIP的路由器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(例如每隔 30 秒)交换路由信息，包括每个目的地以及到这些目的地的距离。

当网络拓扑发生变化时，路由器及时向相邻路由器通告网络拓扑变化后的路由信息。

互联网是一个庞大的联邦网络。

有多个自治运行的组织：没有一个“BOSS”。

组织之间存在合作，但也存在竞争。

尽量保持消息和表的大小最小化，避免泛洪。

算法必须是增量的：不要在每次交换消息时重新计算整个表。

自治系统

5.5.1 AS与路由选择协议分类

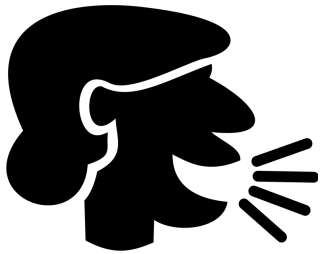
因特网将整个互联网划分为许多较小的自治系统(Autonomous System，AS)。

传统定义的AS 是在单一技术管理下的一组路由器，使用一个内部网关协议（IGP）和共同的测度确定如何在AS内路由分组，并使用一个AS间路由选择协议决定如何将分组发送到其他自治系统。

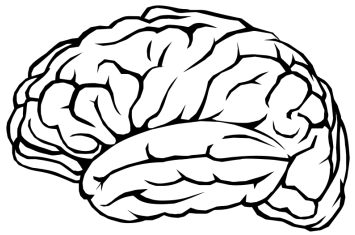
现在使用AS术语强调的是即使使用了多个IGPs和几组测度，一个AS的管理在其他AS看来具有单个一致的内部路由选择规划，并对通过该AS可到达的目的地提供一致的描述。

Inter-domain Routing（跨域路由）

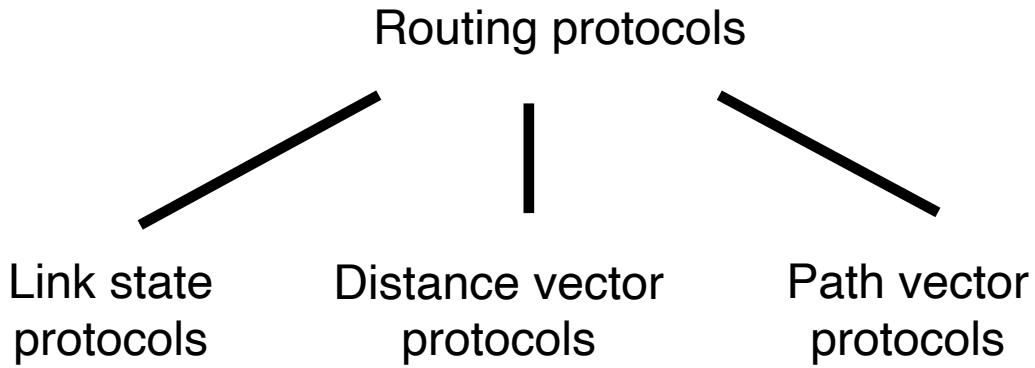
- 目前所讲到的路由算法(LS + DV) 都可以在同一自治域内(autonomous system, AS)使用, e.g., 南邮校园内。
  - 称为 intra-domain routing protocols
- 而互联网使用边界网关协议(Border Gateway Protocol, BGP)
- 所有的AS之间使用BGP。它是连接互联网的纽带。
- BGP是path vector protocol



Messages?



Algorithm?



IGP和EGP

5.5.1 AS与路由选择协议分类

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。

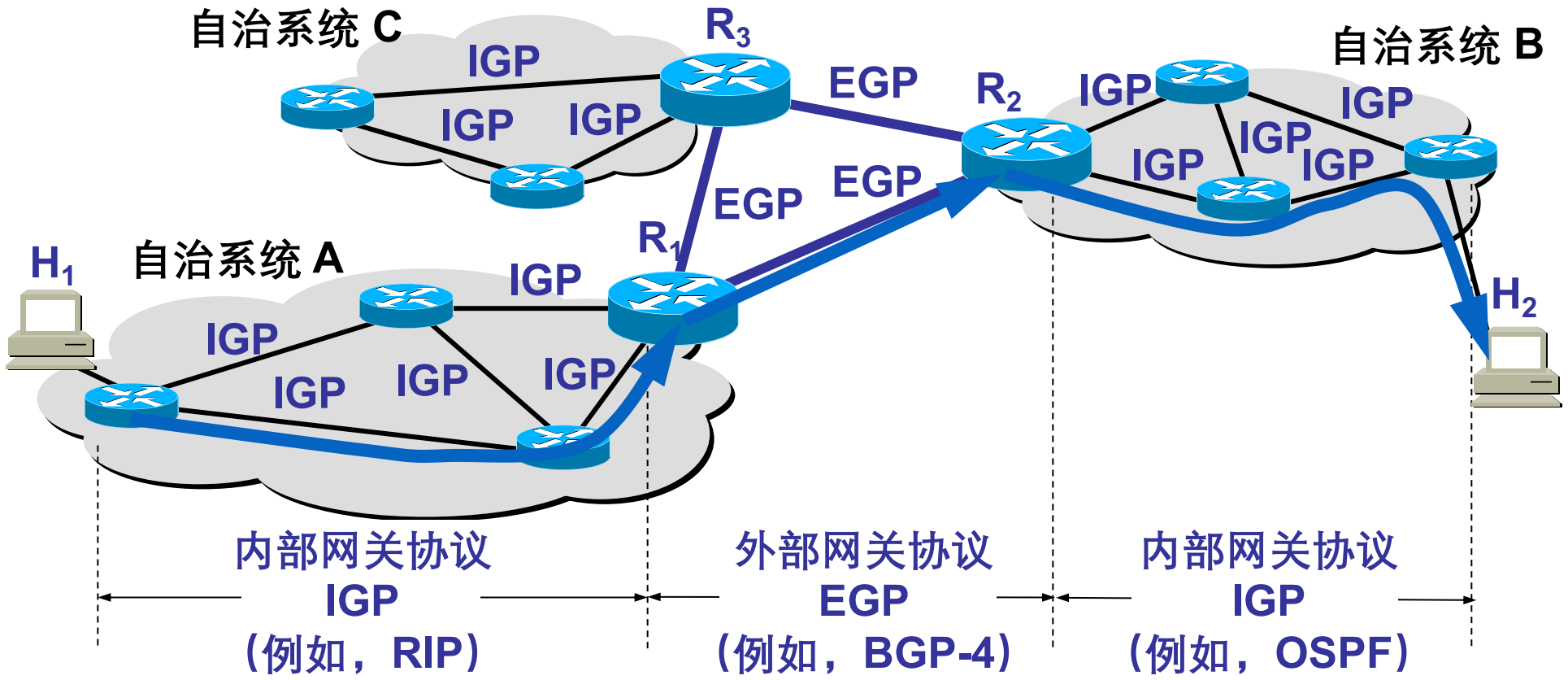
外部网关协议（EGP, External Gateway Protocol）即在自治系统之间交换网络可达性信息所用的路由选择协议。在外部网关协议中目前使用最多的是 BGP-4。

5.5.1 AS与路由选择协议分类

因特网采用分层次的路由选择协议,原因:

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。

AS、IGP和EGP　　5.5.1 AS与路由选择协议分类



5.5.1 AS与路由选择协议分类

从路由算法的自适应性(对网络变化的适应能力)考虑，路由算法可分为两类

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。

5.5.1 AS与路由选择协议分类

适合使用静态路由的情况：

链路带宽较低,不希望传输动态路由选择更新

管理员想完全控制路由器使用的路由

需要为动态路由提供一条备用路由

前往只有一条路径可以到达的网络时

静态路由的缺点：不能动态地适应网络变化

5.5.1 AS与路由选择协议分类

动态路由选择

管理员在每台路由器上配置路由选择协议。这样运行相同路由选择协议的路由器之间会交换有关可达的网络的信息和每个网络的状态。

网络拓扑发生变化后，新信息将动态地传遍整个网络，每台路由器都将更新其路由选择表，以反映变化后的拓扑。

路由器在**刚开始工作时只知道到直接连接的网络的距离**（可定义为1）。以后，**每个路由器只和相邻路由器交换并更新路由信息**。

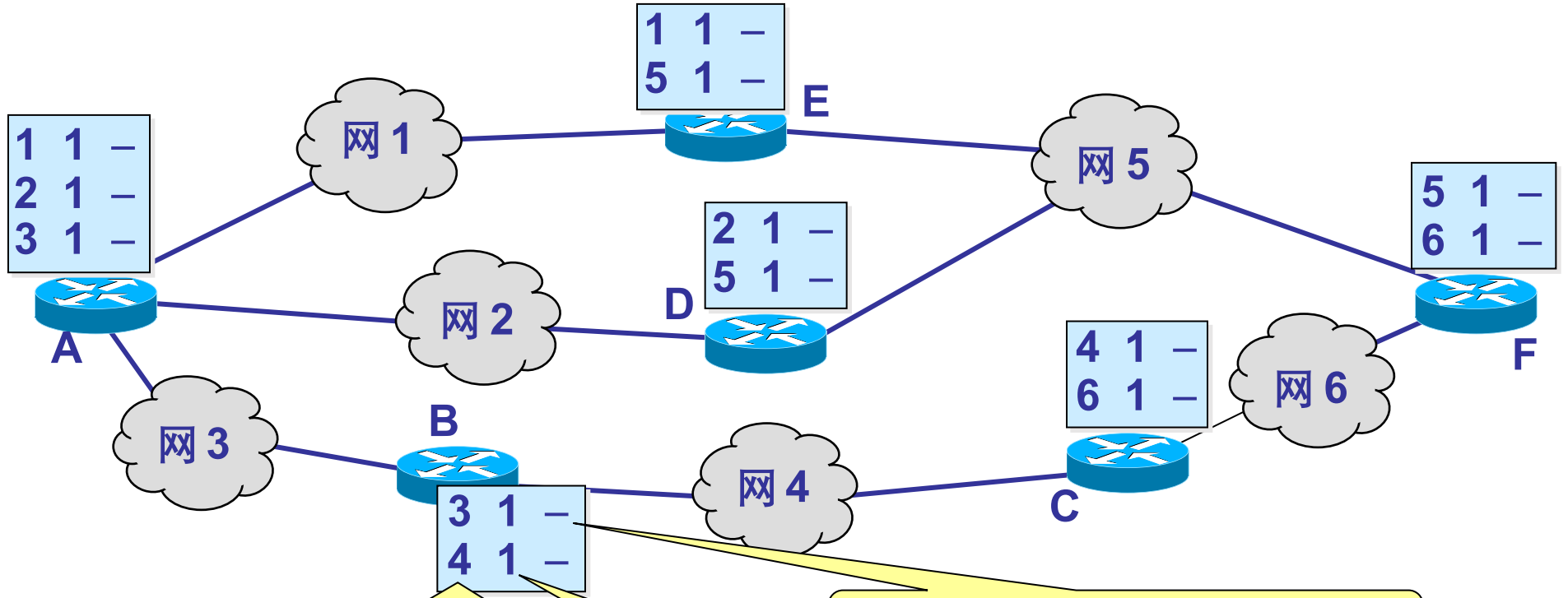
经过若干次更新后，所有的路由器最终都会知道到达本自治系统中任何一个网络的最短距离和下一跳路由器的地址。

RIP协议的**收敛(convergence)过程**(即在AS中所有的结点都得到正确的路由选择信息的过程)较快

路由表构建举例

5.5.2 内部网关协议RIP

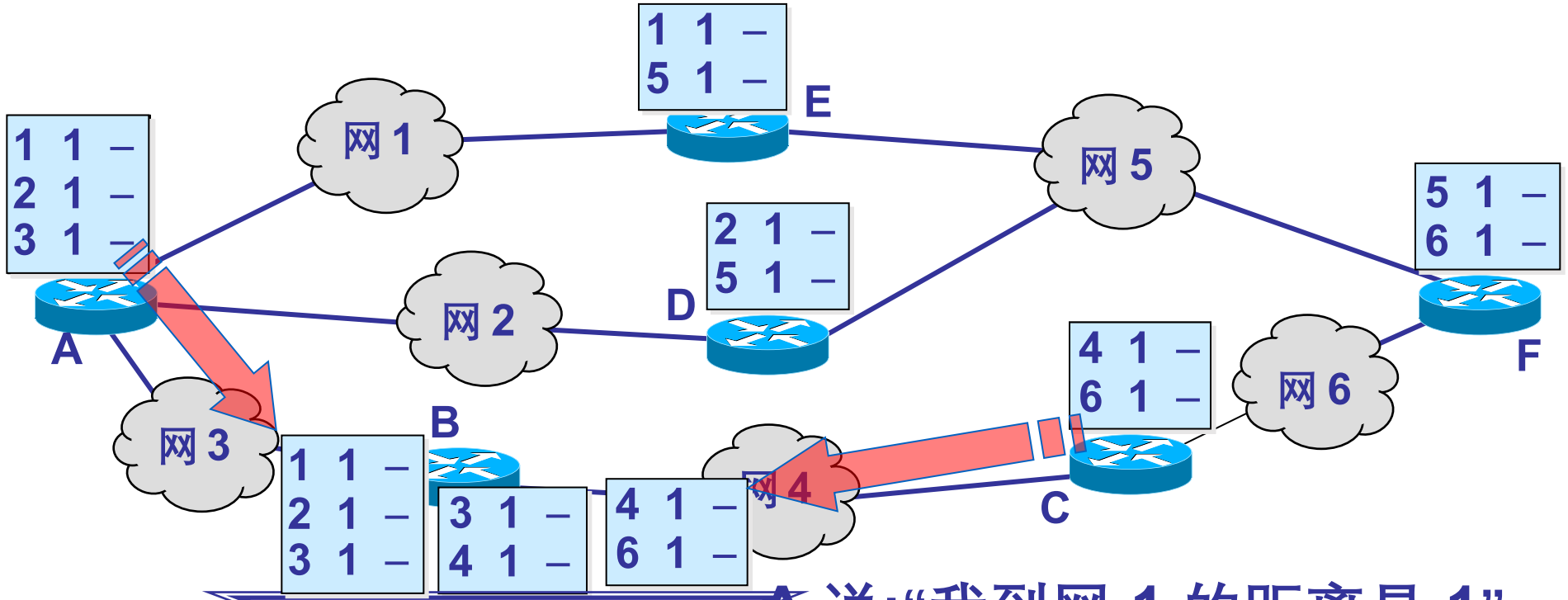
一开始，各路由表只有到相邻路由器的信息



路由表构建举例

5.5.2 内部网关协议RIP

路由器 B 收到相邻路由器 A 和 C 的路由表

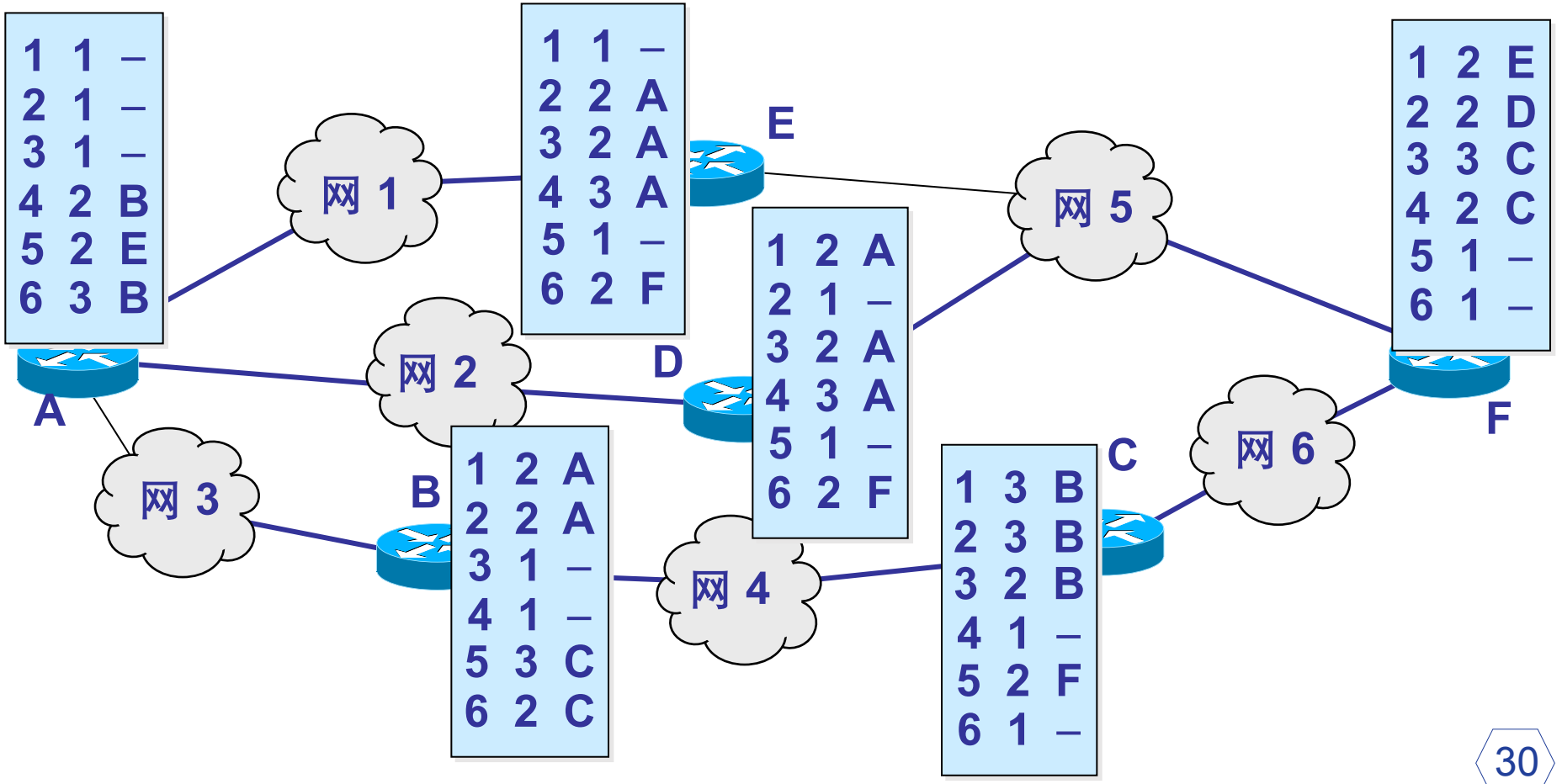


A 说:“我到网 1 的距离是 1”
因此 B 现在也可以到网 1，
距离是 2，经过 A。”

路由表构建举例

5.5.2 内部网关协议RIP

最终所有的路由器的路由表都更新了



RIP报文

5.5.2 内部网关协议RIP

RIP协议使用用户数据报(UDP)进行传送，熟知端口号为 520。RIP 是应用层协议；而转发IP数据报的过程是在网络层完成的。

RIP允许silent RIP进程，即只听不发报文.

RIP协议有2个版本：RIPv1仅使用有类路由，且无子网的概念；RIPv2提供子网掩码信息，使用无类路由。

RIPv1广播发送路由更新。RIPv2支持组播(224.0.0.9)发送，还具有简单鉴别功能。

RIP优缺点

5.5.2 内部网关协议RIP

RIP 存在一个问题：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。

5.5.3 内部网关协议OSPF

OSPF (Open Shortest Path First)是一个**链路状态路由选择协议(link-state routing protocol)**，设计运行于一个自治系统内部。

Open表示公开发表，使用Dijkstra 提出的最短路径算法SPF。

每个OSPF路由器保持一个相同的**描述AS拓扑的数据库(链路状态数据库,LSDB)**。通过从这个库构造一个**最短路径树**来计算得出**路由表**。

5.5.3 内部网关协议OSPF

当拓扑有改变时，OSPF使用最少的路由选择协议流量，快速重新计算路由。OSPF提供对**等价多路径(ECMP, equal-cost multipath)**的支持。

OSPF 提供**区域路由选择能力**(area routing capability)，使**更高水平的路由选择保护**和**路由选择协议流量减少**成为可能。此外，所有OSPF路由选择协议传输都是经过认证的。

链路状态路由选择协议 5.5.3 内部网关协议OSPF

快速适应网络变化，在网络拓扑发生变化时,发送**触发更新**

链路状态发生变化后,检测到变化的设备将生成一个针对该链路的链路状态通告(LSA),使用组播地址将LSA传播给其所有邻接设备.它们据此更新LSDB,再将LSA转发给其所有邻接设备．LSA报告路由器和链路的状态.

以较低的频率(如每隔30分钟)发送**定期更新**（链路状态刷新)

OSPF的三个要点

- 每个路由器向本自治系统中相邻路由器发送信息，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相邻路由器发送此信息。

OSPF的区域

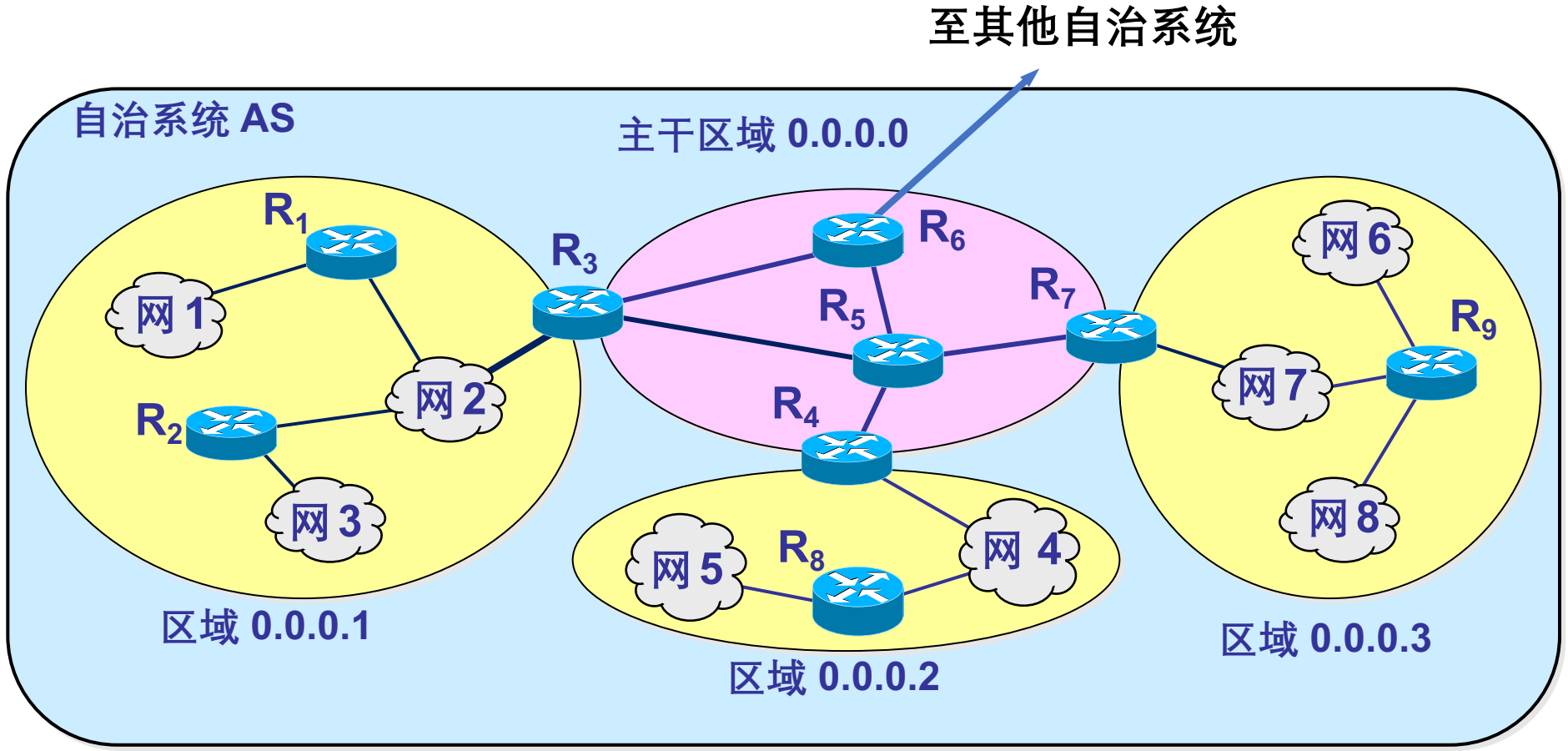
5.5.3 内部网关协议OSPF

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(area)。

每一个区域都有一个 32 bit 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。

主干区域和常规区域　　5.5.3 内部网关协议OSPF



5.3 因特网的路由协议

4、外部网关协议BGP

- 外部网关协议主要解决两个自治系统之间路由器交换路由信息的问题。
- 边界网关协议BGP是为TCP/IP互联网设计的外部网关协议。
- BGP 是不同自治系统的路由器之间交换路由信息的协议。
- BGP 的较新版本是 1995 年发表的 BGP-4。
- 可以将 BGP-4 简写为 BGP。

BGP 使用的环境不同

- 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。
- 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。

过渡页

Transition Page

5.6.1 虚拟专用网 VPN

Internet 由通过路由器相互连接的网络组成，这种网络的缺点是缺乏保密性。

如果一个机构由分散的多个网点构成，为了保证私密性，较容易的办法是建立一个“专用网络”，但成本较高。

虚拟专用网技术保证了VPN中任何一对计算机之间的通信对外界是隐藏的。

VPN的编址

5.6.1 虚拟专用网 VPN

VPN所提供的编址选择与专用网络所提供的是一样的，可以根据需要选择

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——全球惟一的IP地址，必须向因特网的管理机构申请

专用地址(RFC1918)

5.6.1 虚拟专用网 VPN

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一般对目的地址是专用地址的数据报不进行转发。

VPN的实现

5.6.1 虚拟专用网 VPN

VPN的实现主要使用了两种基本技术：隧道传输 和加密技术。

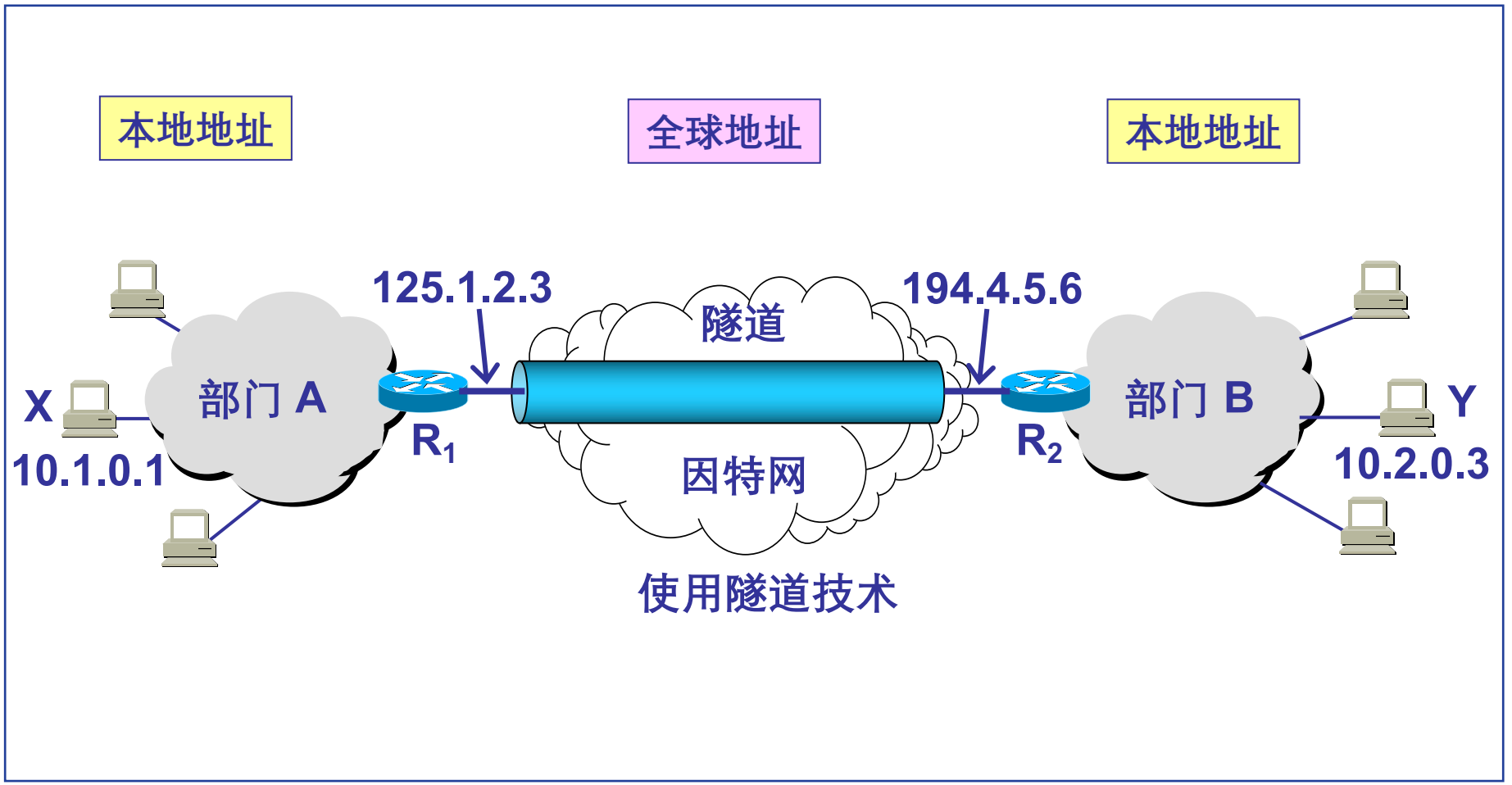
VPN定义了两个网络的路由器之间通过Internet的一个隧道，并使用IP-in-IP封装通过隧道转发数据报。

为了保证保密性，VPN把外发的数据报加密，然后封装在另一个数据报中传输。

隧道接收路由器将数据报解密，还原出内层数据报，然后转发该数据报。

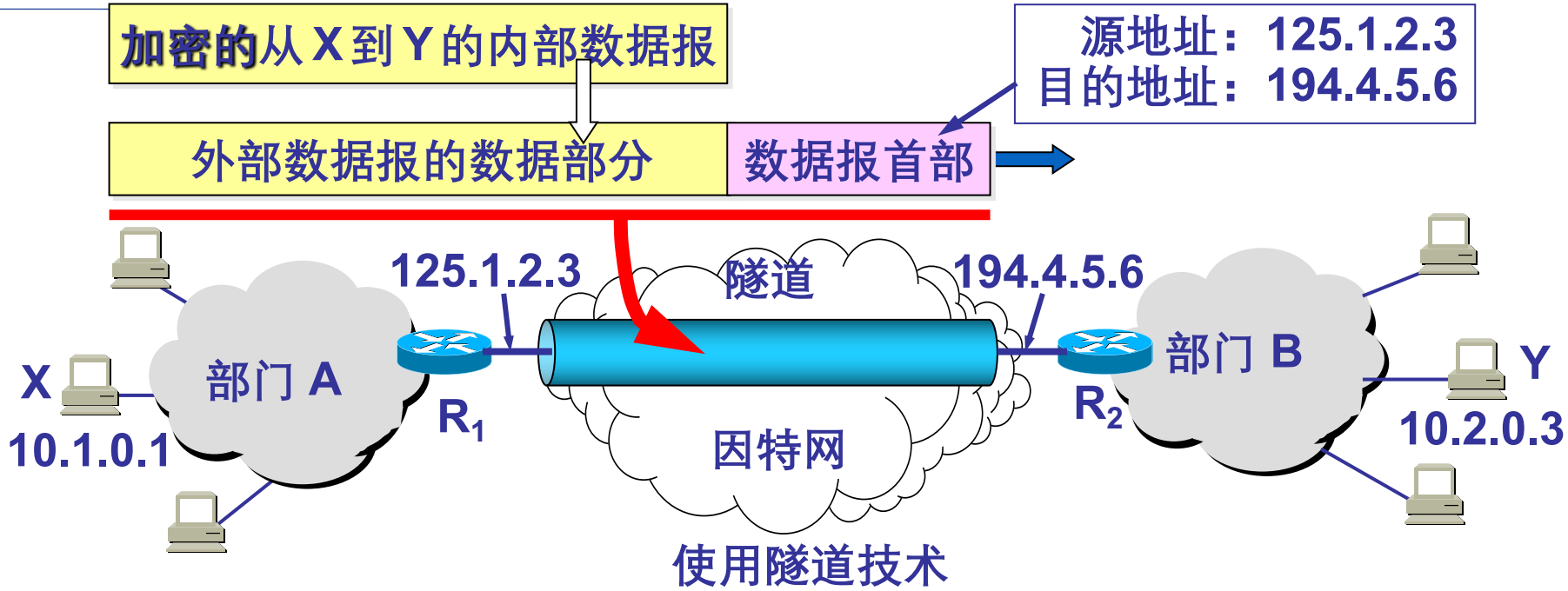
VPN的实现

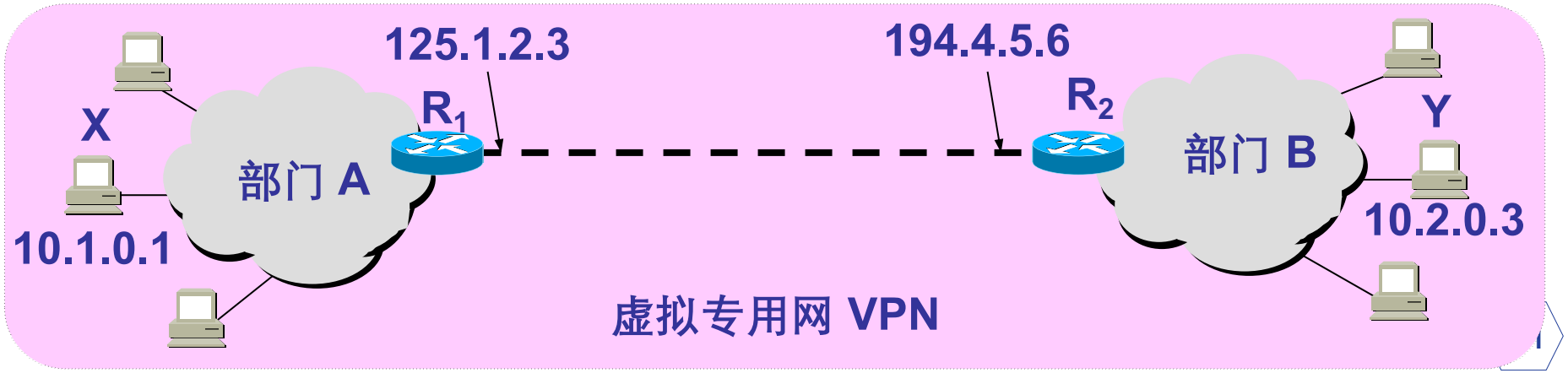
5.6.1 虚拟专用网 VPN



VPN的实现

5.6.1 虚拟专用网 VPN





5.6.2 网络地址转换 NAT

网络地址转换 NAT (Network Address Translation) 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。

两种NAT方法

5.6.2 网络地址转换 NAT

基本网络地址转换（NAT）

并发访问特定外部地址存在限制

网络地址和端口转换（NAPT）

通过对TCP和UDP端口以及IP地址的转换允许并发访问

NAT网络地址转换过程　5.6.2 网络地址转换 NAT

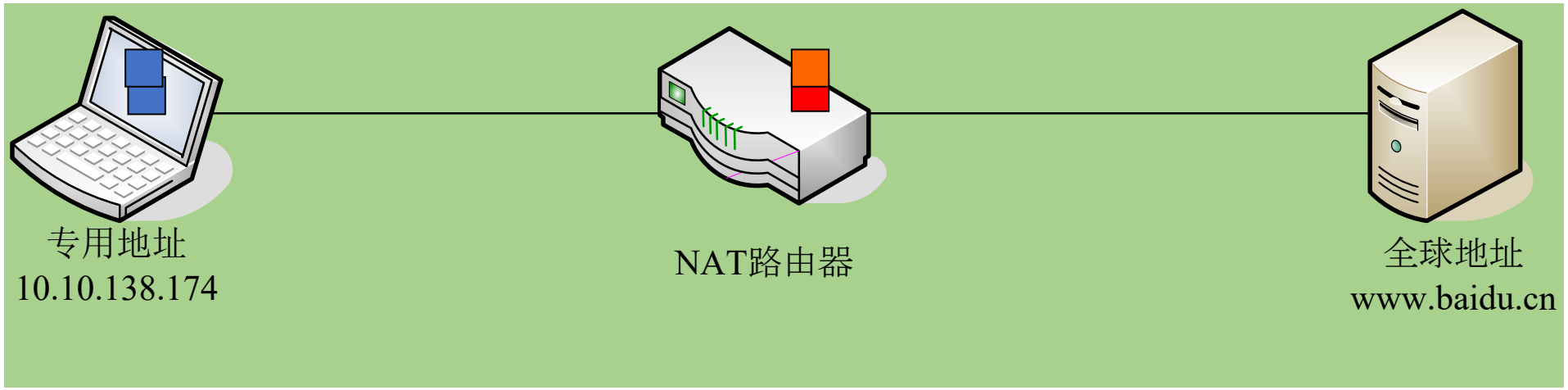
内部主机 X 用本地地址 IP_X 和因特网上主机 Y 通信所发送的数据报必须经过 NAT 路由器。

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。

网络地址转换的过程



源端口：80
目端口：21043
源IP:115.239.211.110
目IP:10.10.138.174

源端口：21043
目端口：80
源IP:10.10.138.174
目IP:115.239.211.110

源端口：80
目端口：14013
源IP:115.239.211.110
目IP:218.2.216.24

NAPT转换表举例

5.8.2 网络地址转换 NAT

内部地址	内部端口	NAPT端口	外部地址	外部端口	协议
10.10.8.27	21043	14007	211.23.33.12	80	tcp
10.10.9.23	43572	14012	211.23.33.12	80	tcp
10.10.9.12	21043	14013	211.23.33.12	80	tcp
10.10.12.124	9542	14015	130.126.13.45	21	tcp
10.10.1.10	5112	14018	202.115.232.57	6919	udp

过渡页

Transition Page

5.5 下一代网际协议IPv6

- 从计算机本身发展以及从因特网规模和网络传输速率来看，现在 IPv4 已很不适用。
- 最主要的问题就是 32 bit 的 IP 地址不够用。
- 要解决 IP 地址耗尽的问题的措施：
  - 采用无类别编址 CIDR，使 IP 地址的分配更加合理。
  - 采用网络地址转换 NAT方法以节省全球 IP 地址。
  - 采用具有更大地址空间的新版本的 IP 协议 IPv6。

5.5 下一代网际协议IPv6

1、IPv6的主要特点

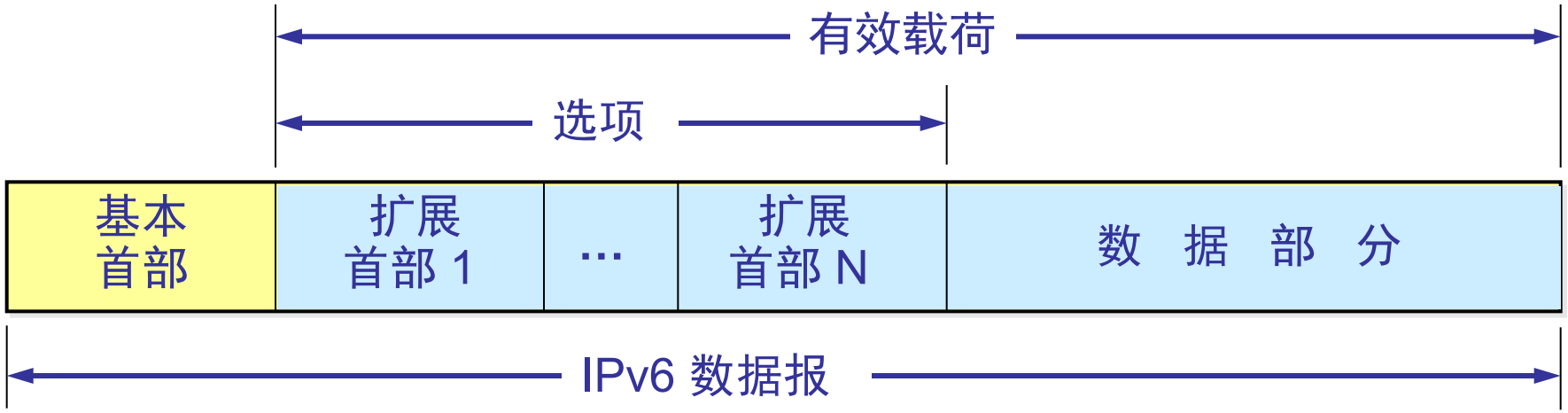
IPv6 所引进的主要变化如下:

- 更大的地址空间。IPv6 将地址从 IPv4 的 32 bit 增大到了 128 bit
- 扩展的地址层次结构
- 灵活的首部格式
- 改进的选项
- 允许协议继续扩充
- 支持即插即用（即自动配置）
- 支持资源的预分配

IPv6 数据报的格式

- IPv6 将首部长度变为固定的 40 字节，称为基本首部(base header)。
- 将不必要的功能取消了，首部的字段数减少到只有 8 个。
- 取消了首部的检验和字段，加快了路由器处理数据报的速度。
- 在基本首部的后面允许有零个或多个扩展首部。
- 所有的扩展首部和数据合起来叫做数据报的有效载荷(payload)或净负荷。

IPv6 数据报的一般形式



5.5 下一代网际协议IPv6

2、IPv6基本首部格式

- 与IPv4首部的固定部分相比,IPv6基本首部主要有下列变化:

  (1) 取消了v4的首部长度字段，v4中的数据报总长度字段被有效载荷长度字段所取代

  (2) 源、目的地址由4字节增大到16字节

  (3) 分片有关字段被转移到了“分片扩展首部”

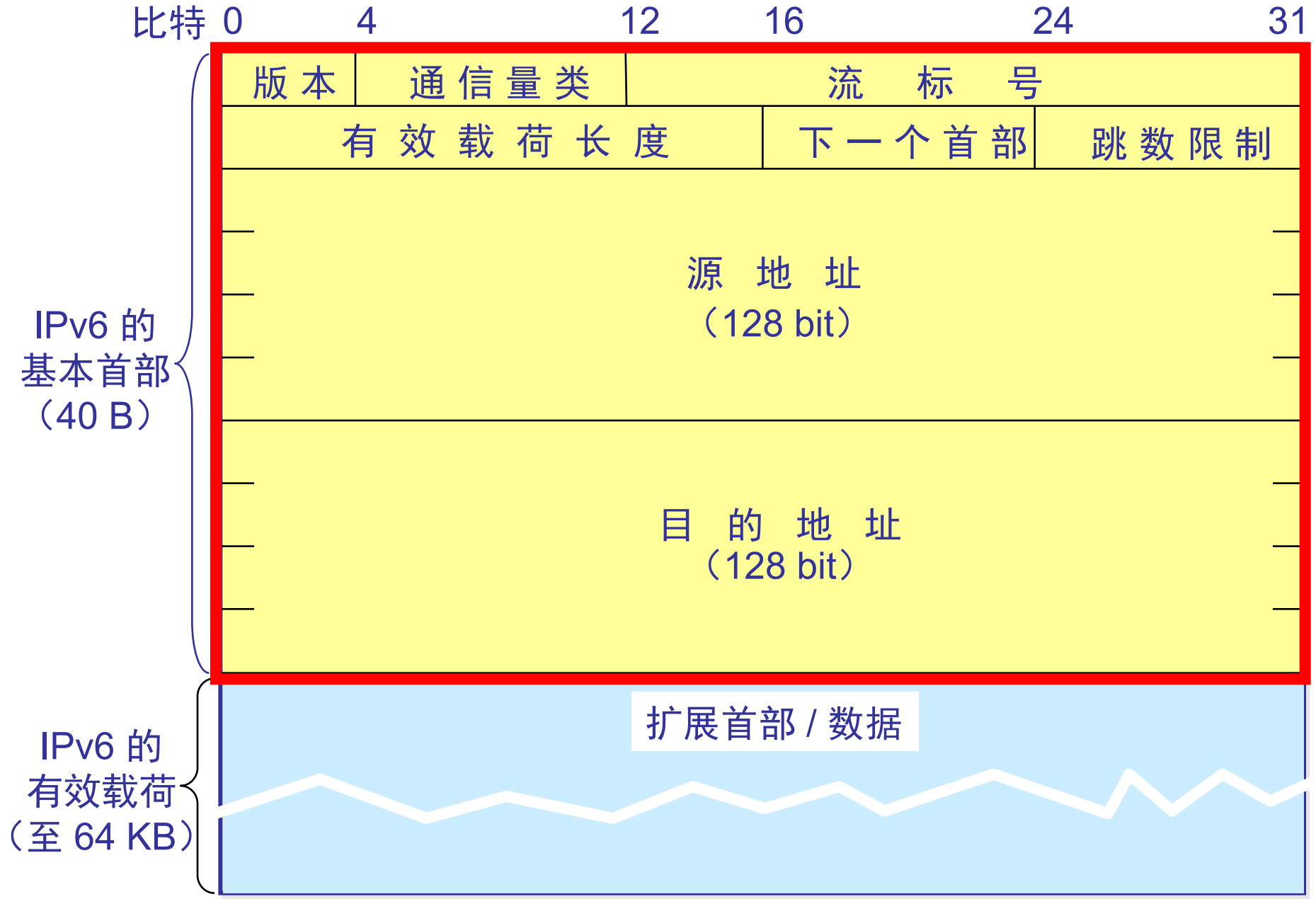
  (4) 生存时间字段改名为跳数限制（hop limit）字段。

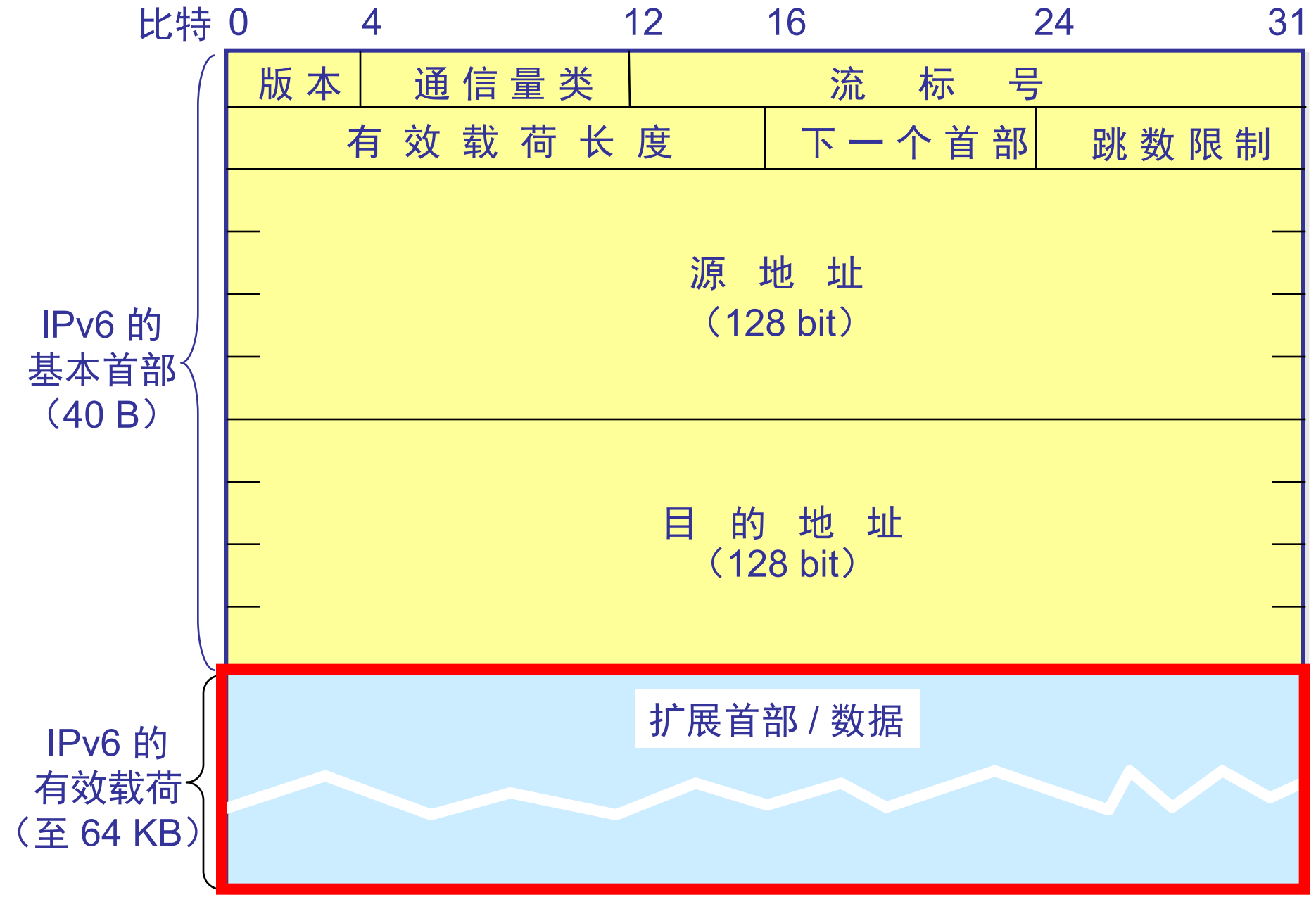
  (5) 服务类型字段改名为通信量类别字段，并增加了流标签字段，一并用于支持资源的预分配。

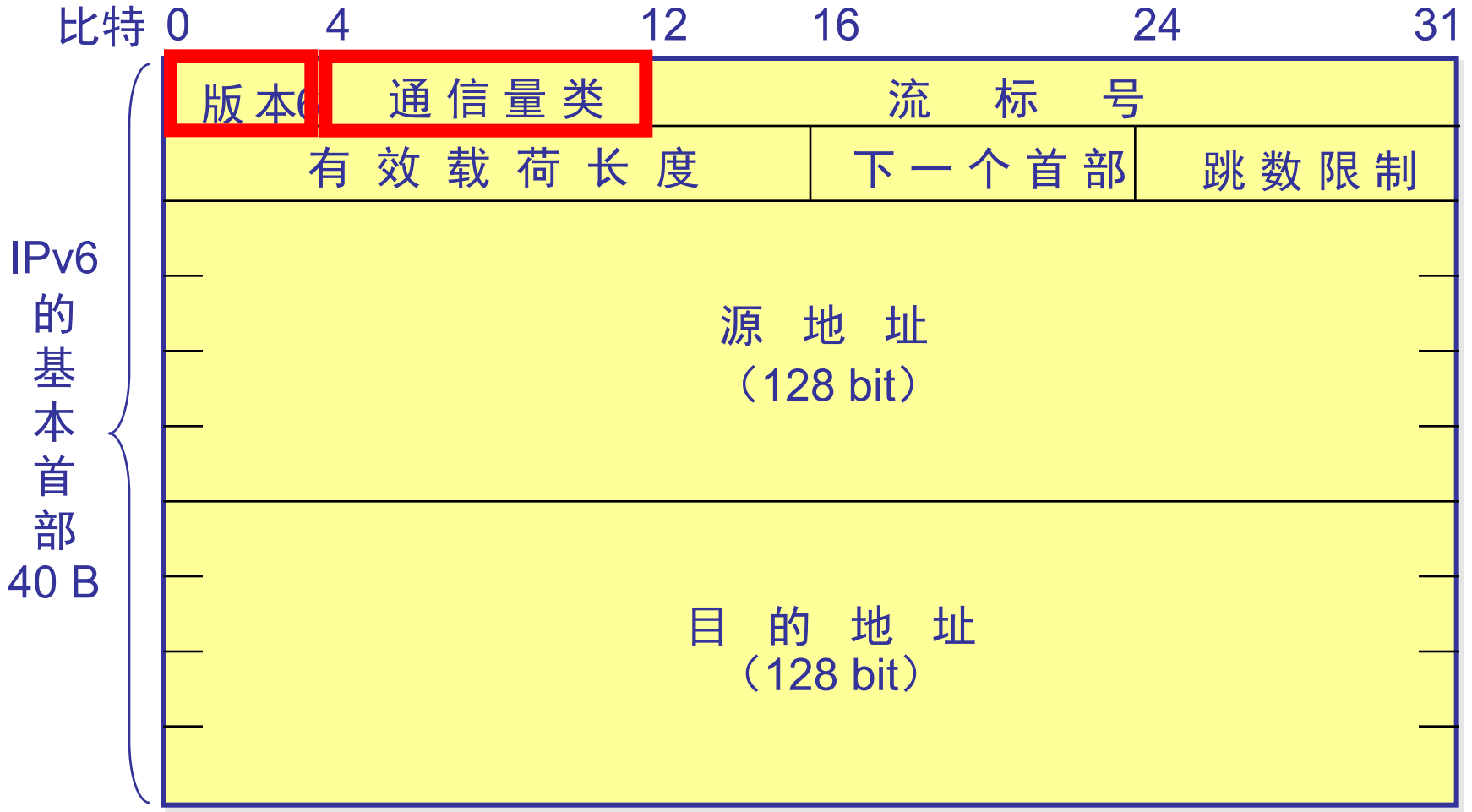
  (6) 协议字段由指明后续内容格式的下一首部字段替代，注意下一首部可能是IPv6数据报的扩展首部，也可能是ICMP、TCP、UDP、IGMP、OSPF等首部。

IPv6 数据报基本首部与 IPv4 数据报首部的对比

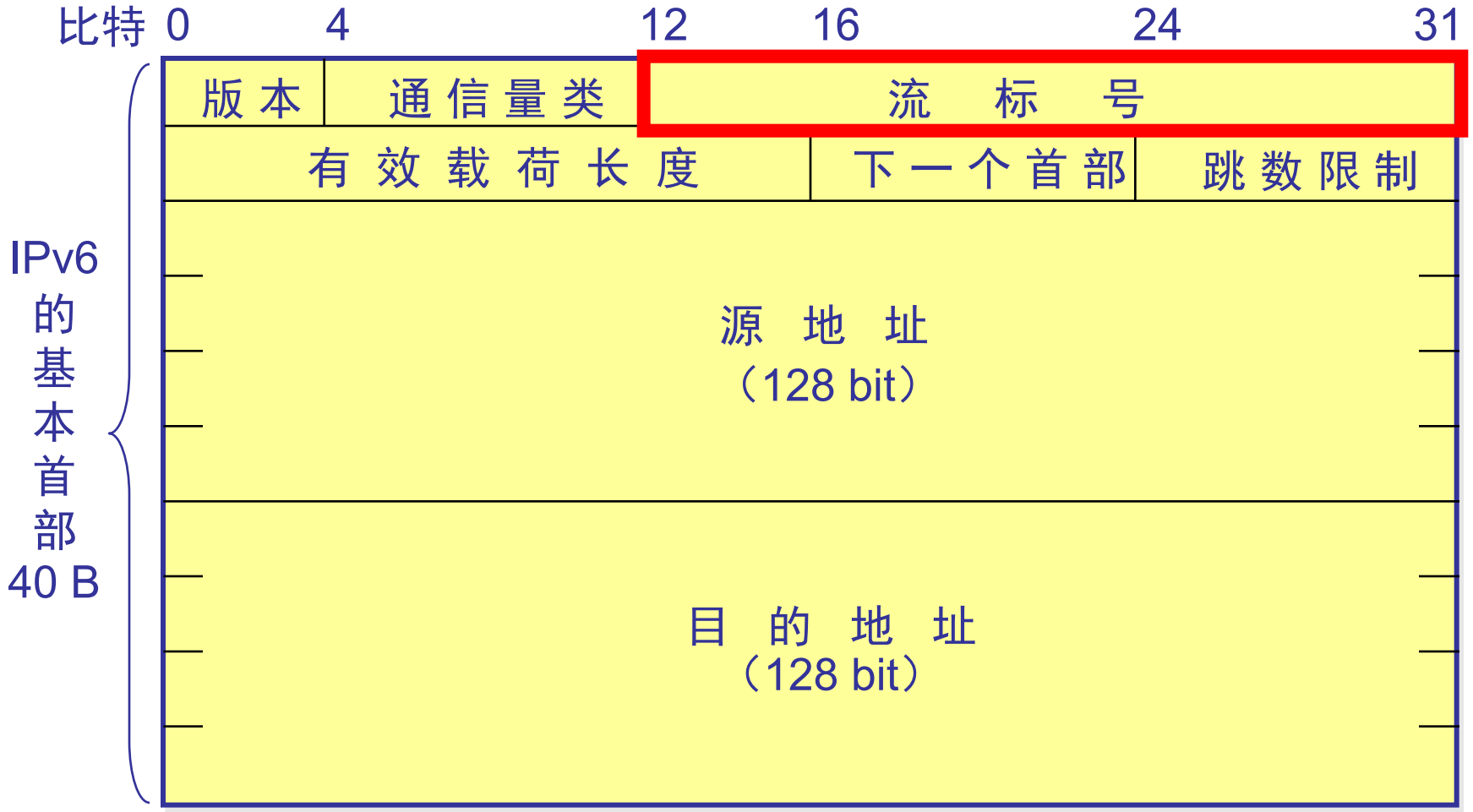
上面是 IPv4 数据报的首部


比特 0
4
12
16
24
31
版 本
通 信 量 类
流 标 号
有 效 载 荷 长 度
下 一 个 首 部
跳 数 限 制
源 地 址
（128 bit）
目 的 地 址
（128 bit）
IPv6 的
基本首部
（40 B）
扩展首部 / 数据
IPv6 的
有效载荷
（至 64 KB）


比特 0
4
12
16
24
31
版 本
通 信 量 类
流 标 号
有 效 载 荷 长 度
下 一 个 首 部
跳 数 限 制
源 地 址
（128 bit）
目 的 地 址
（128 bit）
IPv6 的
基本首部
（40 B）
扩展首部 / 数据
IPv6 的
有效载荷
（至 64 KB）



通信量类(traffic class)—— 8 bit。这是为了区分不同的 IPv6 数据报的类别或优先级。

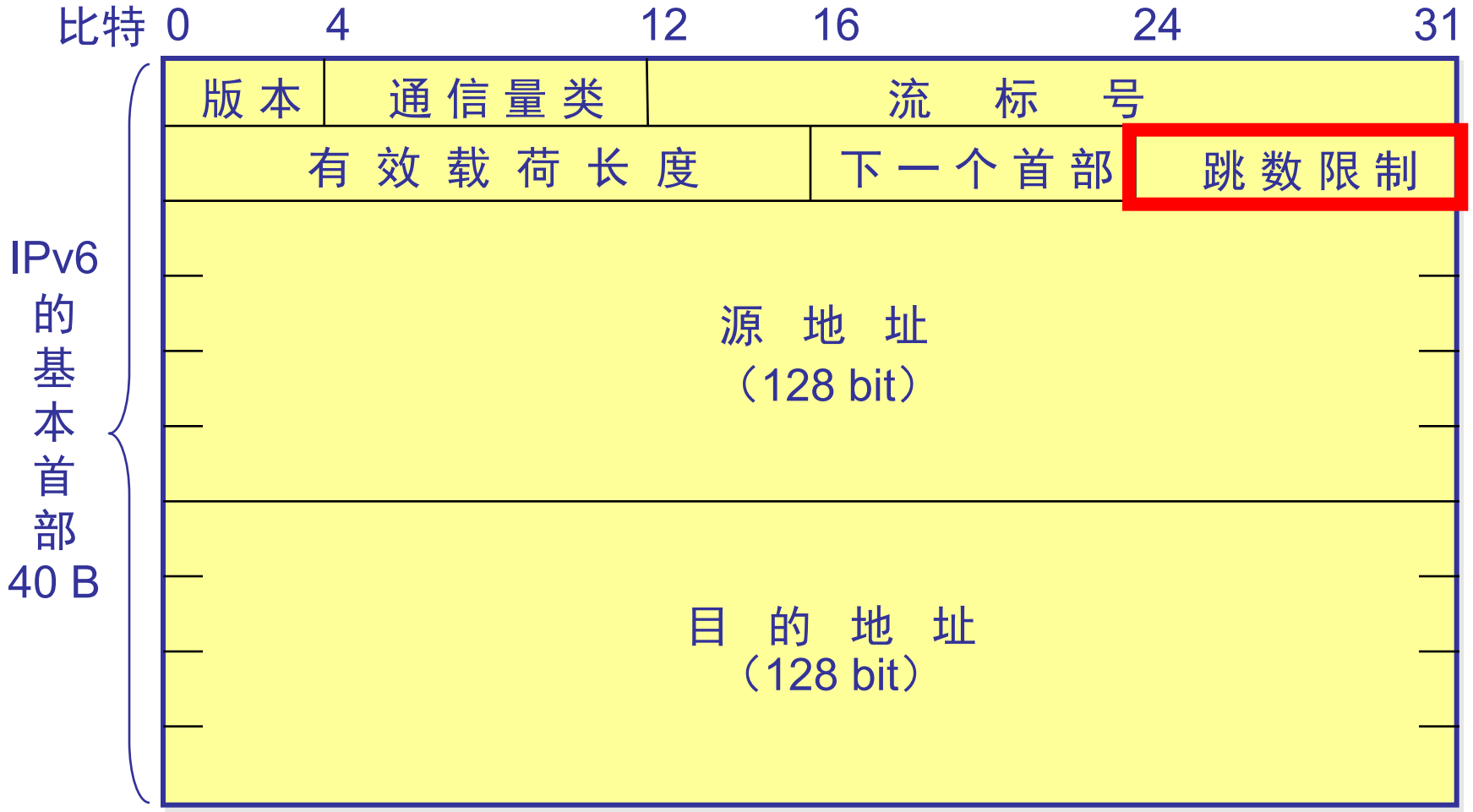


流标号(flow label)—— 20 bit。“流”是互联网络上从特定源点到特定终点的一系列数据报，“流”所经过的路径上的路由器都保证指明的服务质量。

所有属于同一个流的数据报都具有同样的流标号。

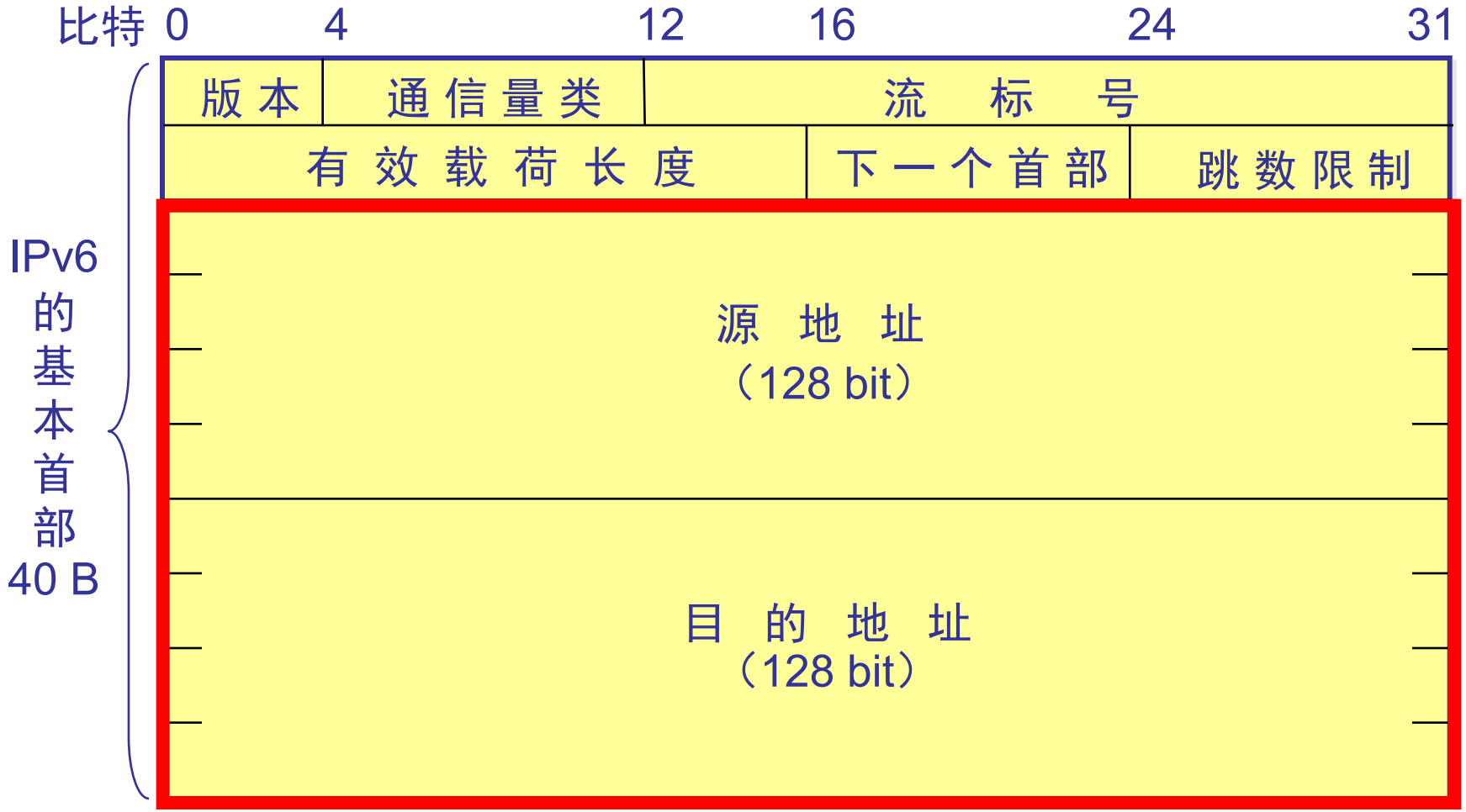
有效载荷长度(payload length)—— 16 bit。它指明 IPv6 数据报除基本首部以外的字节数（所有扩展首部都算在有效载荷之内），其最大值是 64 KB。

下一个首部(next header)—— 8 bit。它相当于 IPv4 的协议字段或可选字段。



跳数限制(hop limit)—— 8 bit。源站在数据报发出时即设定跳数限制。路由器在转发数据报时将跳数限制字段中的值减1。

当跳数限制的值为零时，就要将此数据报丢弃。



源地址—— 128 bit。是数据报的发送站的 IP 地址。

目的地址—— 128 bit。是数据报的接收站的 IP 地址。

5.5 下一代网际协议IPv6

3、IPv6扩展首部

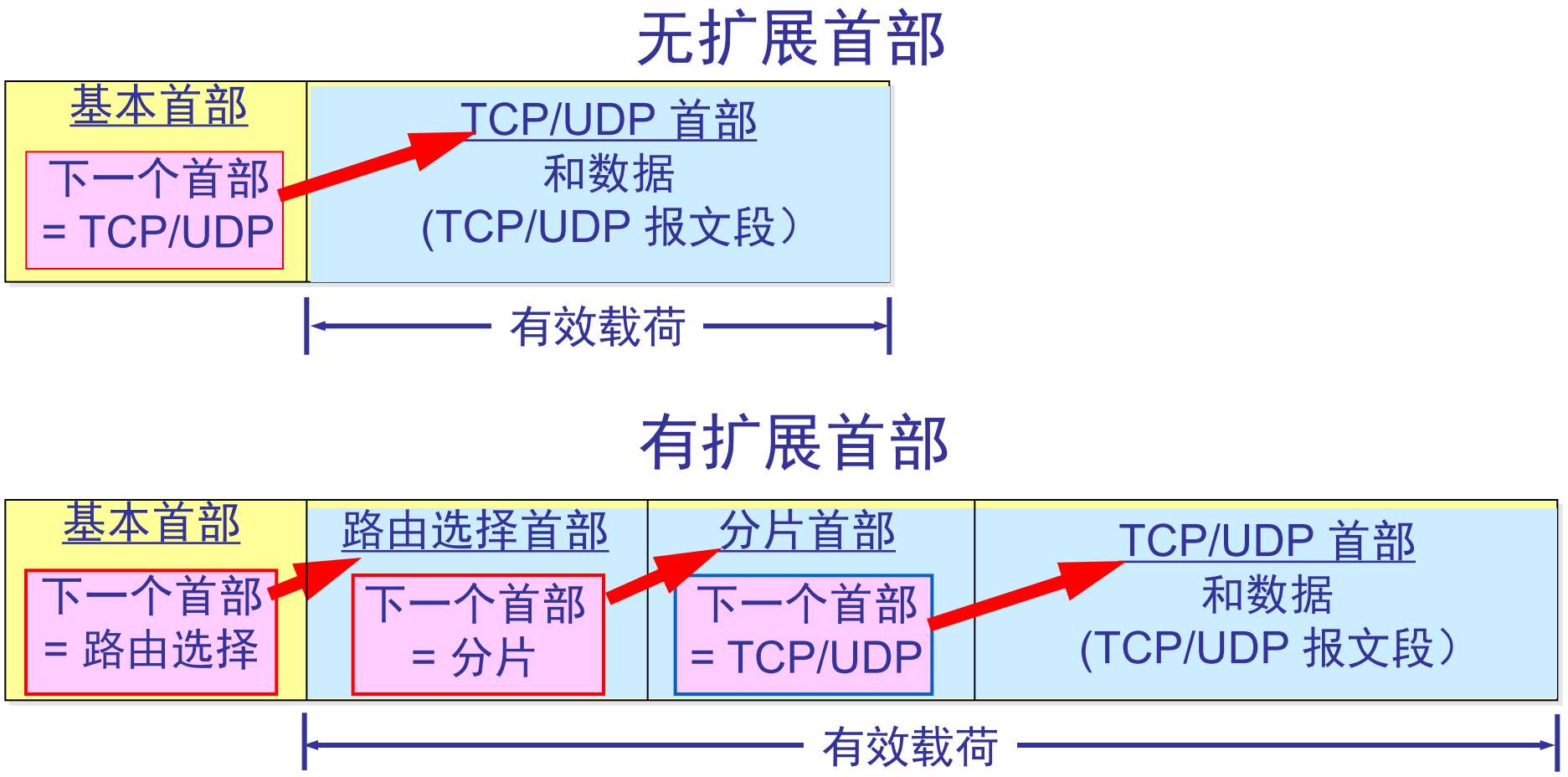
- IPv6 将原来 IPv4 首部中选项的功能都放在扩展首部中，并将扩展首部留给路径两端的源站和目的站的主机来处理。
- 数据报途中经过的路由器都不处理这些扩展首部（只有一个首部例外，即逐跳选项扩展首部）。
- 这样就大大提高了路由器的处理效率。

六种扩展首部

在[RFC 2460]中定义了六种扩展首部:

- 逐跳选项
- 路由选择
- 分片
- 鉴别
- 封装安全有效载荷
- 目的站选项

IPv6 的扩展首部


无扩展首部
基本首部
下一个首部
= TCP/UDP
TCP/UDP 首部
和数据
(TCP/UDP 报文段）
有效载荷
有扩展首部
基本首部
下一个首部
= 路由选择
路由选择首部
下一个首部
= 分片
分片首部
下一个首部
= TCP/UDP
TCP/UDP 首部
和数据
(TCP/UDP 报文段）
有效载荷

5.5 下一代网际协议IPv6

4、IPv6编址

IPv6 128 bit 的地址空间,IPv6 数据报的目的地址可以是以下3种基本类型地址之一:

(1) **单播**(unicast) 单播就是传统的点对点通信。

(2) **组播**(multicast) 组播是一点对多点的通信。

(3) **任播**(anycast) 这是 IPv6 增加的一种类型。任播的目的站是一组计算机，但数据报在交付时只交付给其中的一个，通常是距离最近的一个。

结点与接口

- IPv6 将实现 IPv6 的主机和路由器均称为结点。
- IPv6 地址是分配给结点上面的接口。
  - 一个接口可以有多个单播地址。
  - 一个结点接口的单播地址可用来惟一地标志该结点。

冒号十六进制记法 (colon hexadecimal notation)

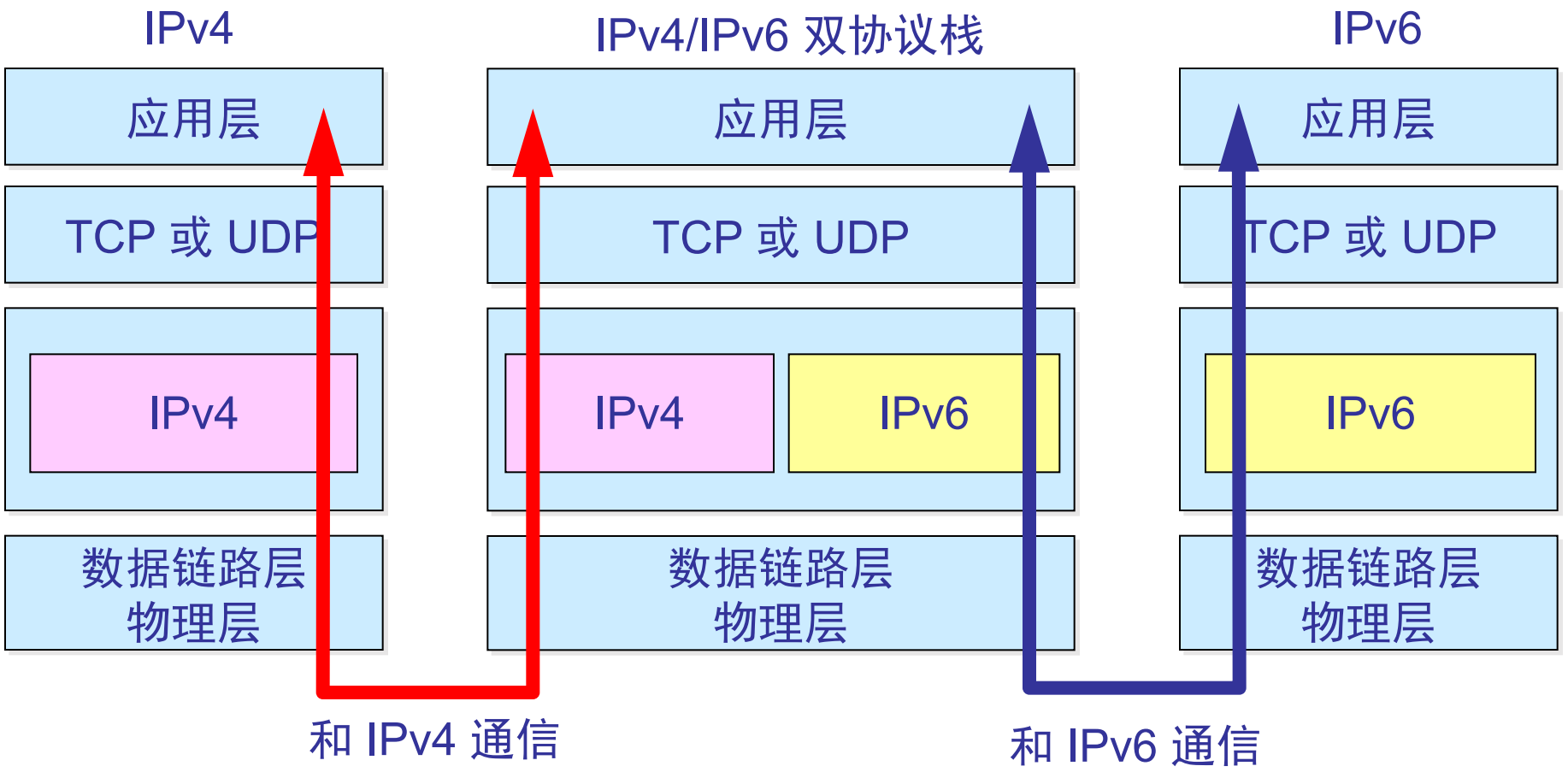
- 每个 16 bit 的值用十六进制值表示，各值之间用冒号分隔。
  68E6:8C64:FFFF:FFFF:0:1180:960A:FFFF
- 零压缩(zero compression)，即一连串连续的零可以为一对冒号所取代。
- FF05:0:0:0:0:0:0:B3 可以写成：FF05::B3
- 一个IPv6地址中，零压缩只能使用一次。

5.5 下一代网际协议IPv6

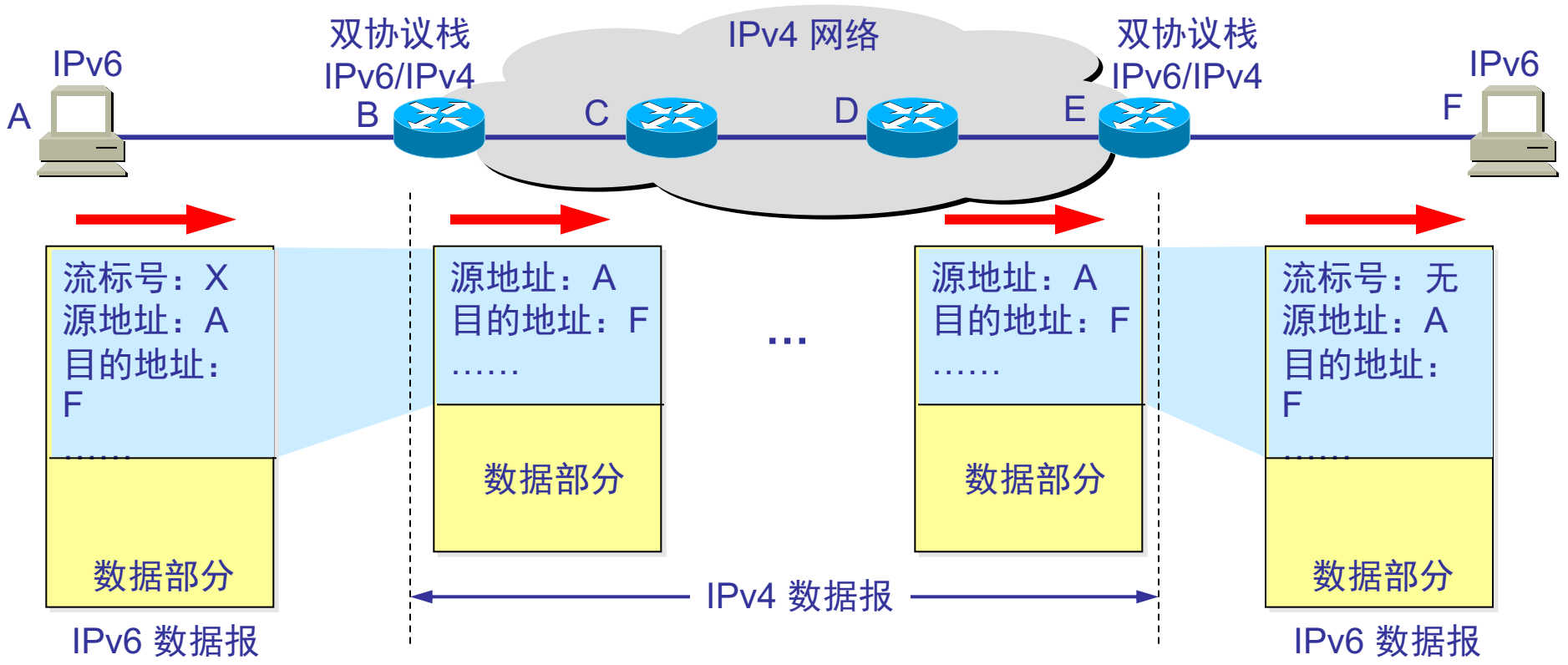
5、从 IPv4 向 IPv6 过渡

- 向 IPv6 过渡只能采用逐步演进的办法，同时，还必须使新安装的 IPv6 系统能够向后兼容。
- IPv6 系统必须能够接收和转发 IPv4 分组，并且能够为 IPv4 分组选择路由。
- 双协议栈(dual stack)是指在完全过渡到 IPv6 之前，使一部分主机（或路由器）装有两个协议栈，一个 IPv4 和一个 IPv6。
- 隧道技术

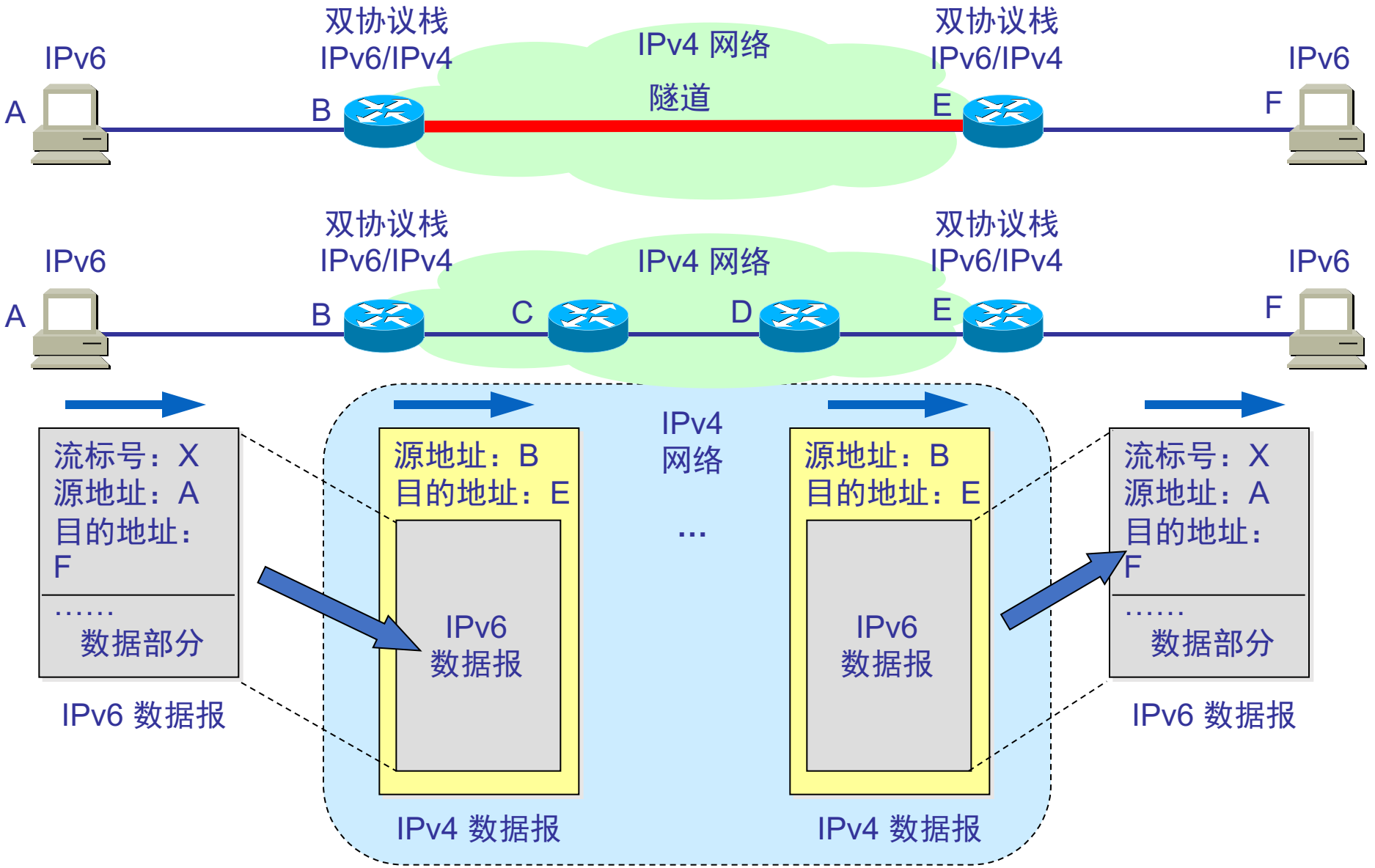
双协议栈


IPv4
IPv4/IPv6 双协议栈
IPv6
应用层
应用层
应用层
TCP 或 UDP
TCP 或 UDP
TCP 或 UDP
IPv4
IPv4
IPv6
IPv6
数据链路层
物理层
数据链路层
物理层
数据链路层
物理层
和 IPv4 通信
和 IPv6 通信

用双协议栈进行从 IPv4 到 IPv6 的过渡



使用隧道技术从 IPv4 到 IPv6 过渡



Have a Nice Day

南京邮电大学通信与信息工程学院

“计算机通信与网络” 国家精品课程组