

计算机网络原理与实践（第2版）配套课件
机械工业出版社 2013年

第5章 网络层

第5章 网络层

5.1 网络层的基本概念

5.2 IPv4协议

5.3 因特网上的地址机制

5.4 因特网上的路由机制

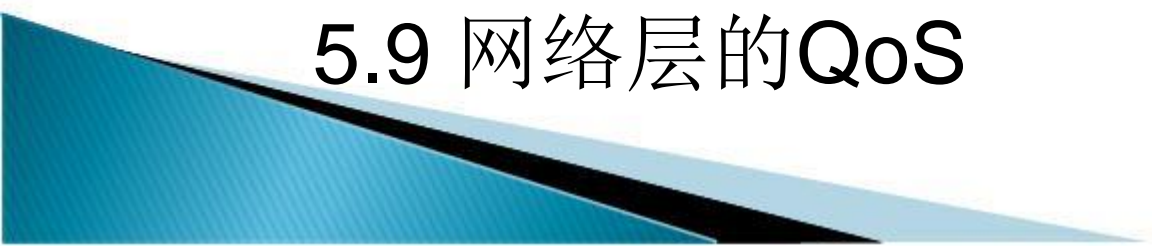
5.5 因特网上的控制协议ICMP

5.6 因特网上的多播

5.7 下一代因特网协议IPv6

5.8 移动IP

5.9 网络层的QoS



- 网络层主要解决什么问题？
- 数据链路层寻址和网络层寻址有什么不同？
- 目前网络层的主流协议是什么？




5.1 网络层的基本概念


- 网络层的主要功能
- 网络层向上提供的两种服务




5.1.1 网络层的主要功能

- 主要解决网络互连问题，完成将数据分组从一个网络中的源主机发送到另一个网络中的目的主机的任务。
 - 网络层的功能包括：网络层的编址、寻址和转发、跨网络的路由选择，报文长度的控制等。
 - 网络层设备：路由器。
- 

5.1.2 网络层向上提供的两种服务

- 网络层应该向运输层提供怎样的服务？
 - 网络层能够提供的服务应该有传输质量上的衡量，例如：
 - ✓ 有无连接？
 - ✓ 是否确保交付？
 - ✓ 分组有无按序交付？
 - ✓ 有无确保的带宽等。
- 


5.1.2 网络层向上提供的两种服务

- 网络层向运输层提供的分组转发服务可分两大类：
 - 数据报(datagram)服务
 - 虚电路(virtual circuit)服务
 - 早期数据通信网络普遍提供虚电路服务，如X.25和帧中继网络。
 - 90年代因特网迅速普及，提供数据报服务的IP协议成为网络层的主流协议。
- 

数据报服务

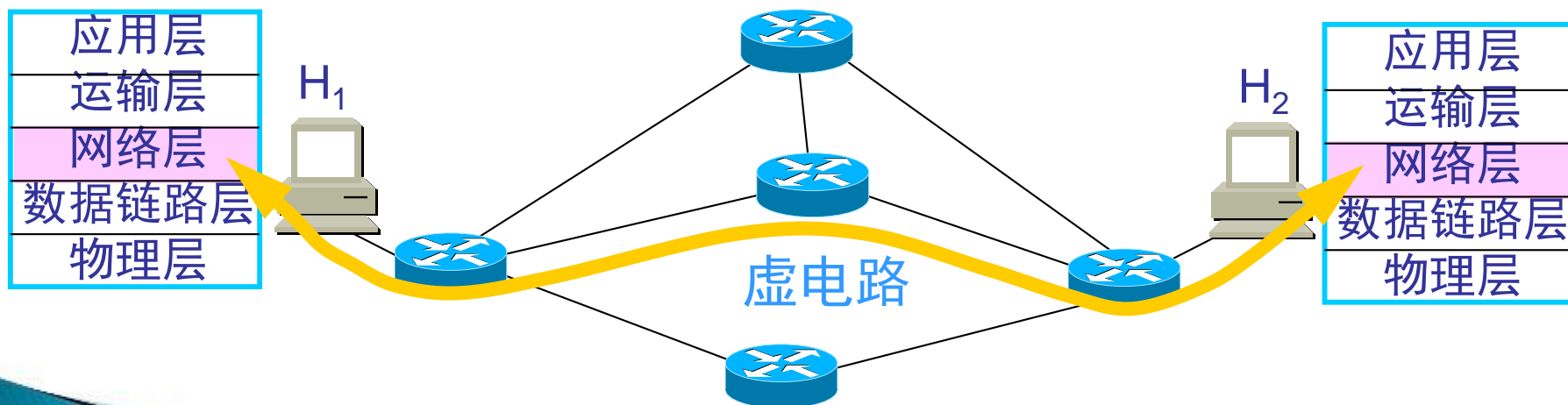
- 数据报服务：典型的分组交换技术、无连接、**尽最大努力交付的服务**。
- 工作方式：源结点为分组加上目的地址，发给相邻的下一站，沿途通过的每一个中间结点都利用该目的地址和自己的转发表来转发分组，每一个分组独立发送。

数据报服务

- 优点：转发功能的实现简单灵活，沿途结点不必为建立连接付出额外的消息传递和处理开销，网络生存性好。
 - 不足：可能会有分组丢失、来自同一应用的一组数据报可能沿不同的路径传输，导致失序的情况发生。
 - IP协议提供的是数据报服务。
- 

虚电路服务

- 面向连接的通信服务
- 在两个节点的应用进程之间建立起一个逻辑上的连接或虚电路后，就可以在两个节点之间按顺序发送分组，接受端无须重组、排序。



虚电路服务

- 虚电路通信的三个阶段：
 - 建立虚电路连接
 - 数据传输
 - 释放虚电路连接
- 虚电路一般采用**标记交换**（label switch）技术实现，在数据分组中携带标记来指示它的虚电路，标记通常由路由器结点分配。实用的标记交换协议有**ATM**、**MPLS**等

网络层两种服务的特点比较

比较项	数据报服务	虚电路服务
连接	无连接	面向连接
目的地址的使用	分组选路用	建立连接用
路由选择的时机	需要为每个分组独立选路	在建立连接时确定路由
转发方式	根据目的地址和路由表转发	根据虚电路号转发
分组到达顺序	可以不按顺序	按顺序
单点失效的影响	小，可以方便地改变路由	大，通过该点的所有虚电路连接将中断，需要重新建立
分组到达顺序	可以不按顺序	按顺序
额外的信令开销	不必要	必要
服务质量	尽力投递	有资源建立连接后，质量有保证

5.2 IPv4协议

- 数据报的格式
 - IP协议首部
 - 校验和运算的要点
 - 首部选项
- IP报文的分片

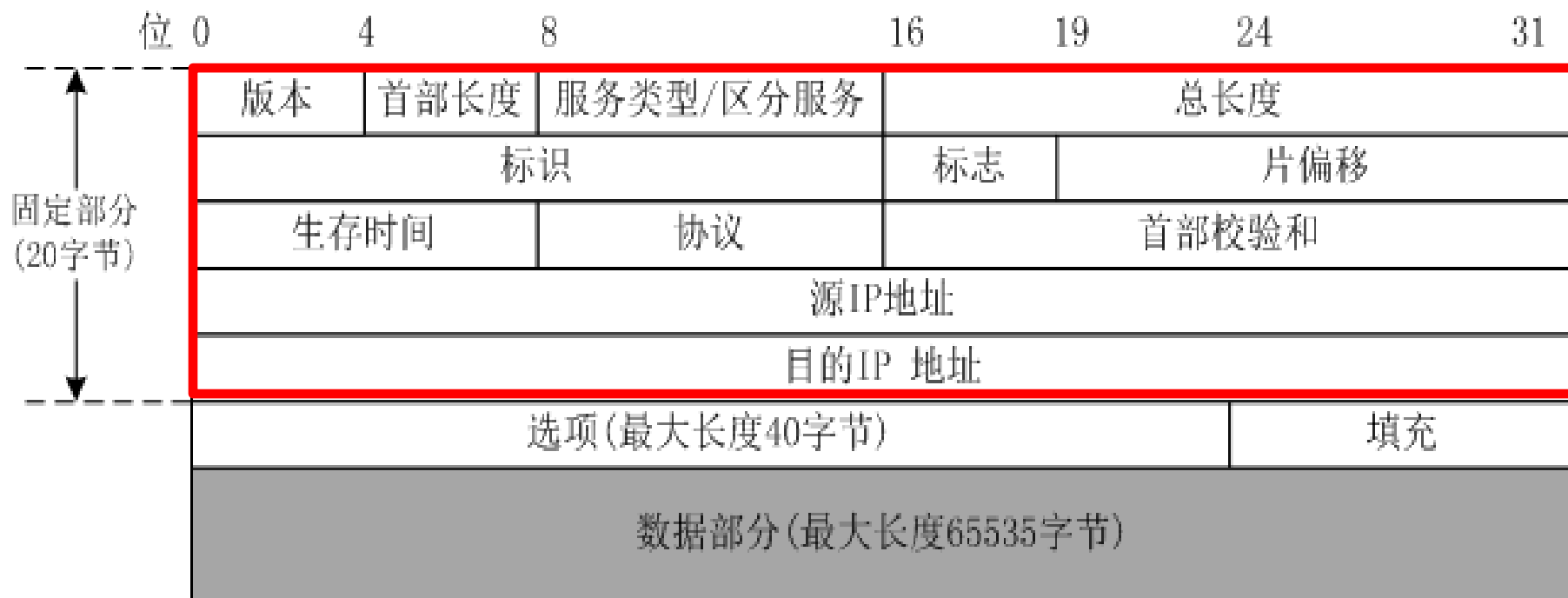


5.2.1 IP数据报的格式

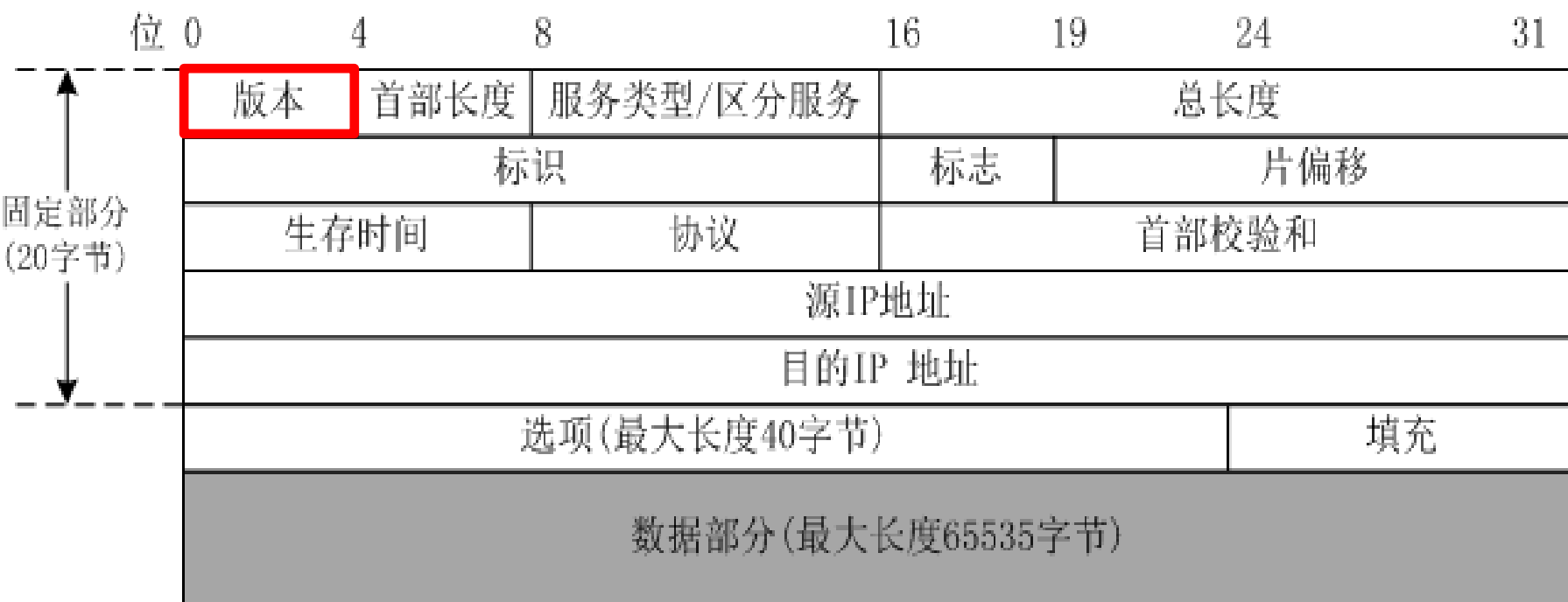
- 因特网的网络层采用IP协议
- IP数据报的封装包含首部（header）和数据部分
- IP数据报的首部包括20字节的固定部分和变长的可选项（option）。



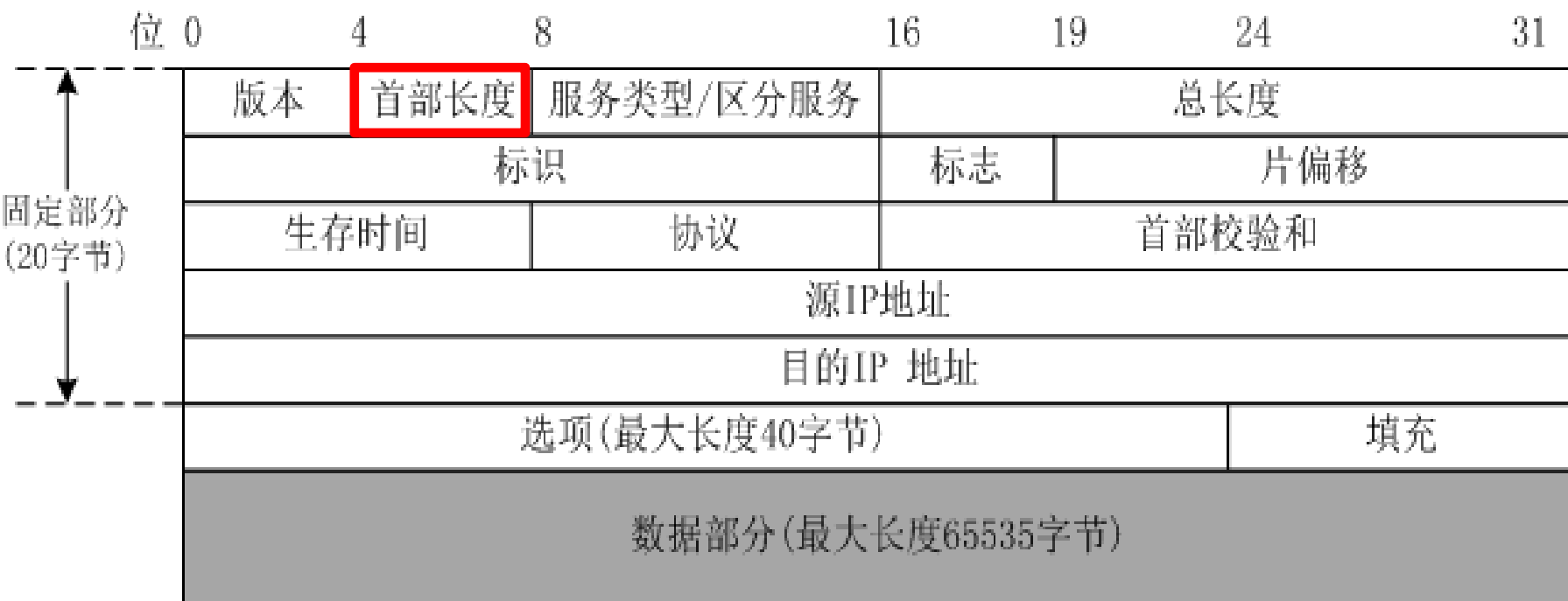
IP数据报的格式



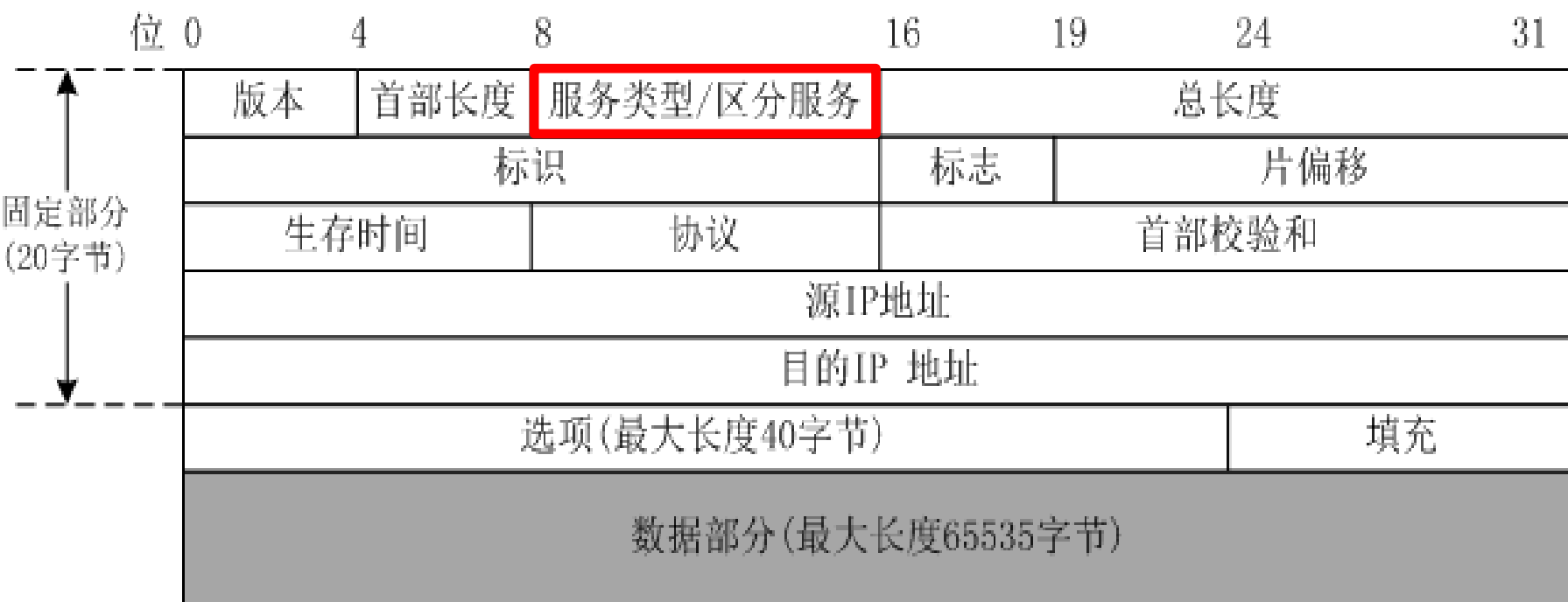
版本---占4位，IP 协议的版本号，
4（IPv4）、6（IPv6），目前版本为4。



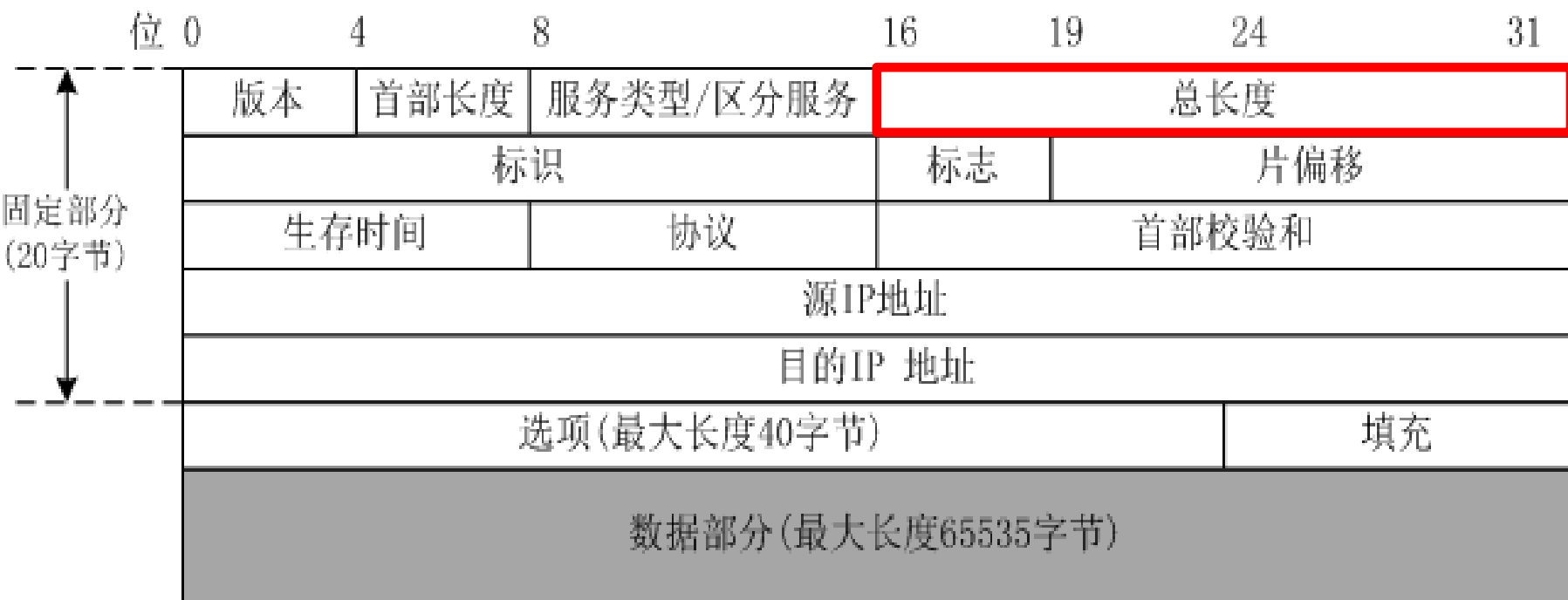
首部长度的占4位，4个字节为一个单位
常见值为5，Why?



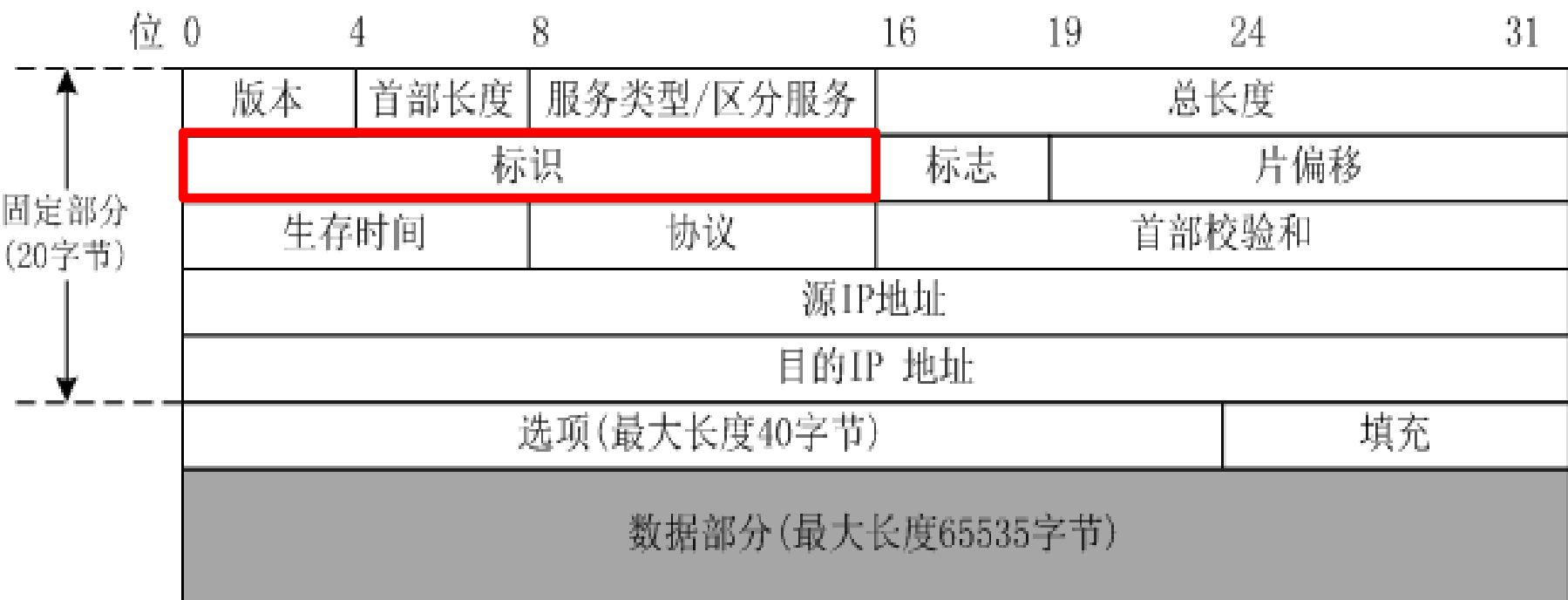
服务类型（TOS）/区分服务（DS）---占8位,用其中的若干位为IP 数据报分类。



总长度---占16位，首部和数据之和(字节数)，最大长度是 $2^{16}-1=65535$ 字节。实际上，IP数据报作为帧的数据部分被封装时，总长度不超过下层的MTU。



标识---占16位，在分片情况下，用于用于标识分组属于哪个数据报，以便接收端对数据报重组。



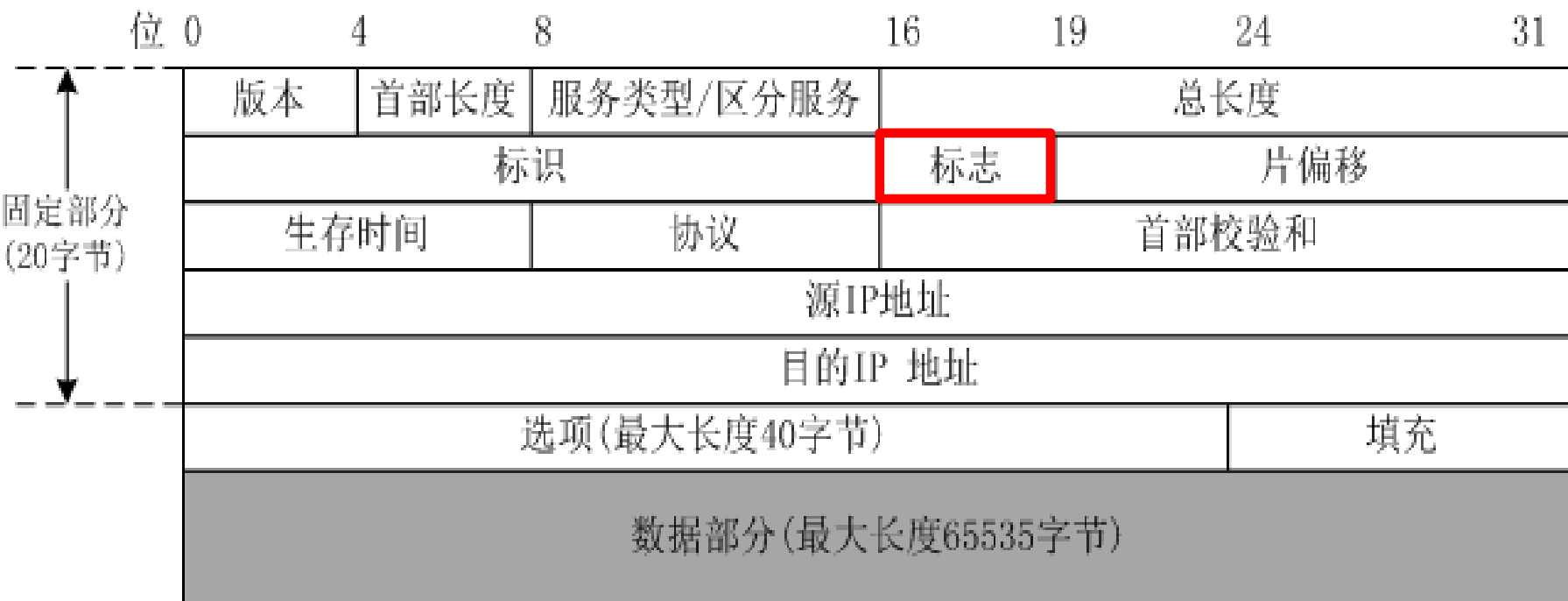
标志---占3位，目前只用前两位，与分片有关。

最低位： MF (More Fragment):

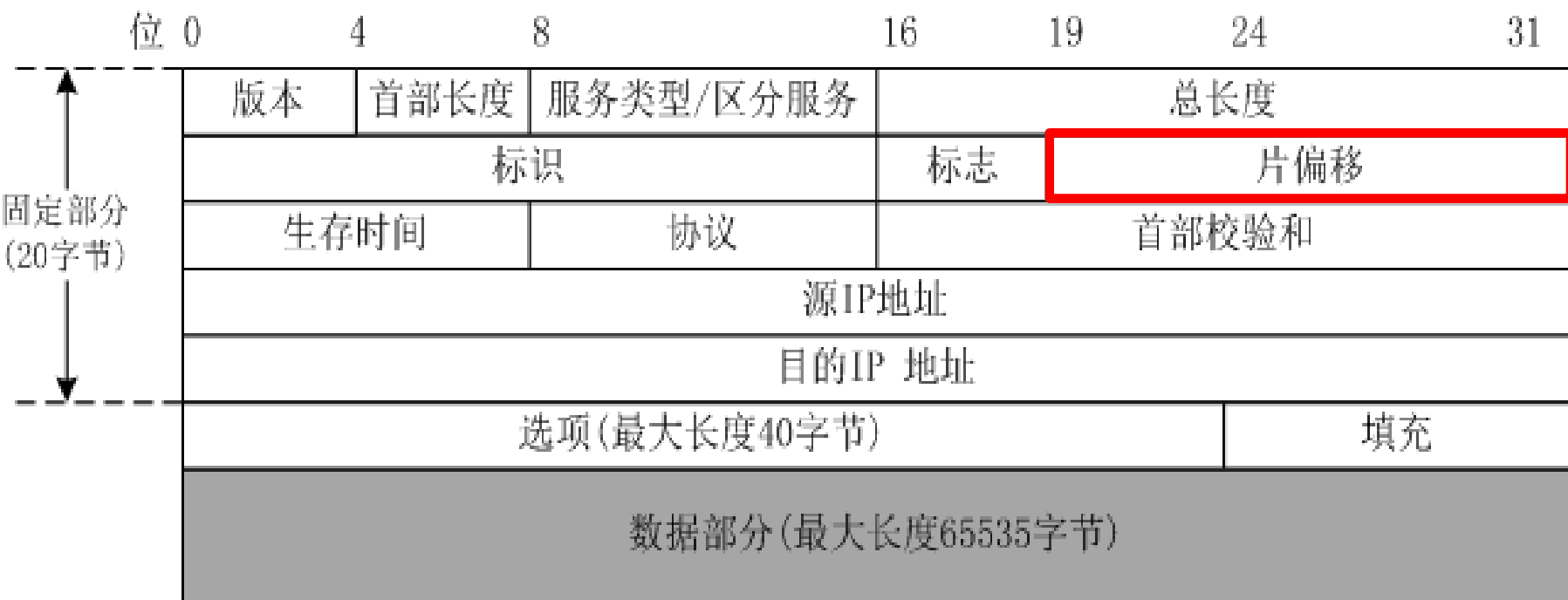
MF = 1 （后面还有分片）

MF = 0 表示最后一个分片。

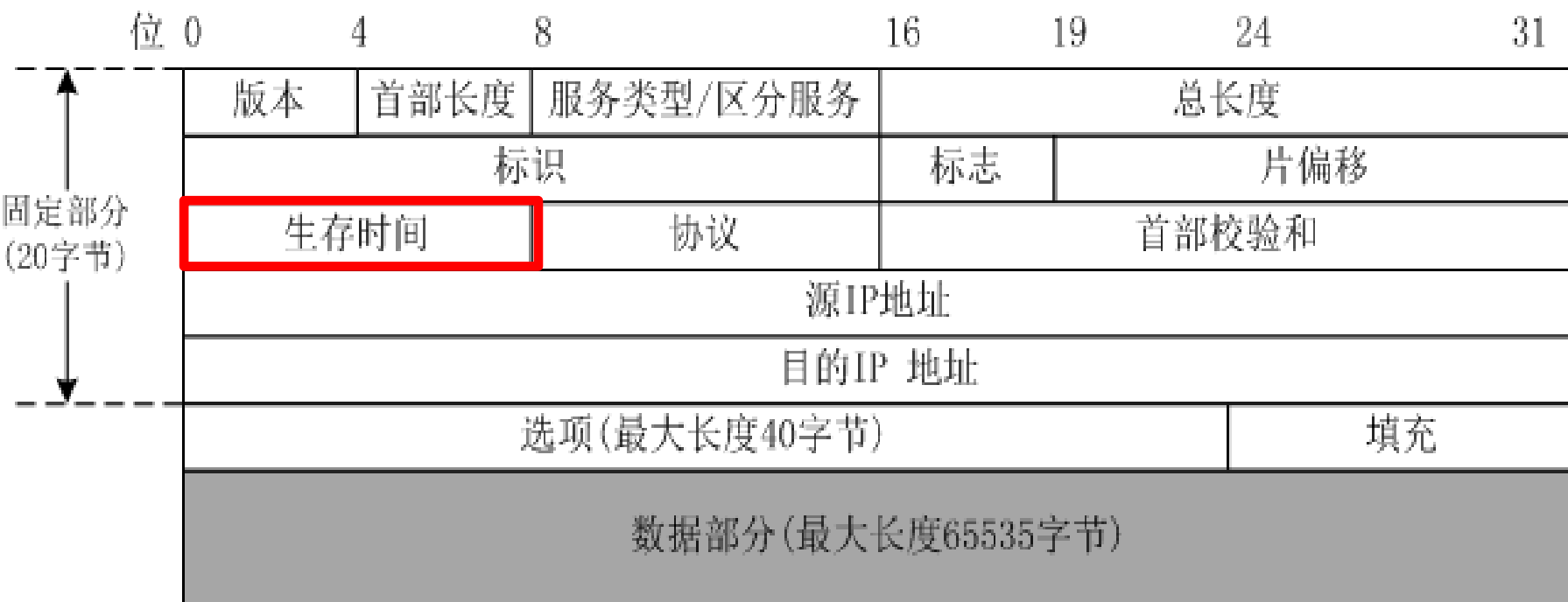
中间位： DF (Don't Fragment) : DF = 0 时允许分片。



片偏移---占13位，与分片有关。指示本片数据的第一个字节在原数据报数据区中的偏移量，偏移量以8个字节为单位。

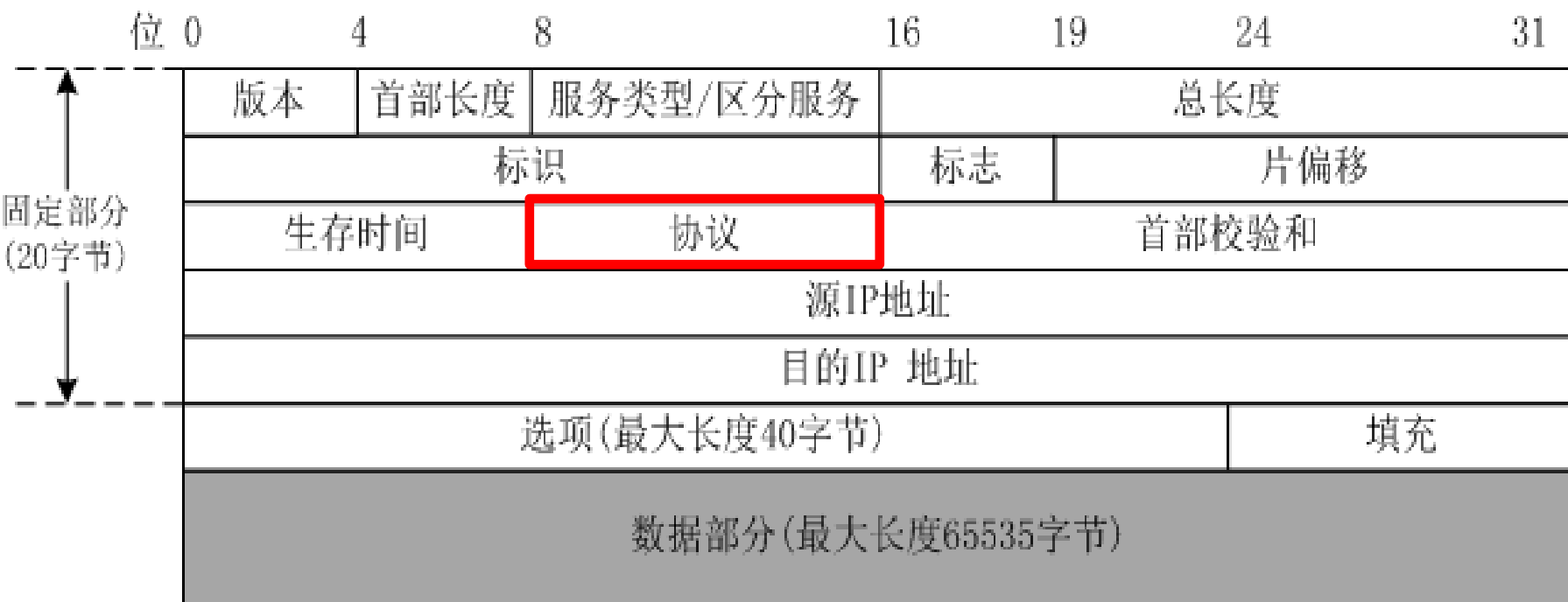


生存时间（TTL）——占8bit，IP分组在网络中允许通过的最大路由器数（跳数）。
不同操作系统会有自己默认的初始TTL值。

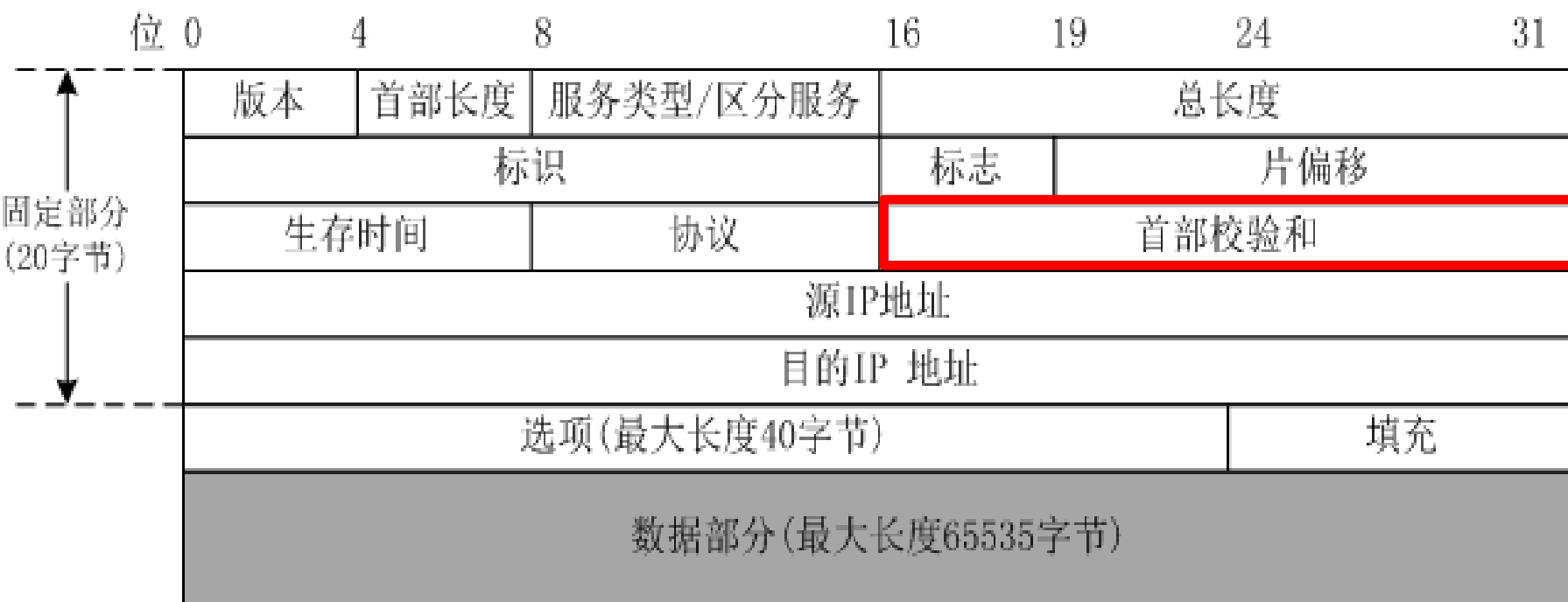


协议——占8bit。指示该数据报封装的数据部分所采用的上层协议类型。 例如：

UDP: 17, TCP: 6, ICMP: 1, OSPF: 89等。



首部校验和—— 16比特，只对首部校验，不采用CRC，采用反码运算的方法。



校验和运算的要点


发送时：

1. 先将校验和字段清零；
2. 每两个字节为一个运算数，所有的数反码求和（运算数不取反码，只是最高位有进位时循环加到最低位）；
3. 最后的和取反码，填入校验和字段，发送数据报。


在接收时：

1. 将收到的IP数据报首部做同样运算；
 2. 和取反码，结果为0时认可该数据报， 否则认为出错丢弃
- 注意：每个路由器都需要对转发的数据报计算首部校验和，因为在传输过程中，首部一些字段值会发生变化（如TTL）。

首部选项

- 安全选项： 表明数据报的安全级别。
 - 源路由选项（严格）： 给定数据报转发路径上的每一跳。
 - 源路由选项（松散）： 只给出必须经过的路由器。
 - 记录路由选项： 要求沿途每个路由器附上自己的IP地址。
 - 时间戳选项： 要求沿途每个路由器附上自己的IP地址和时间戳。
- 

5.2.2 IP报文的分片（fragment）

- 链路层协议的最大传输单元（MTU）限制IP数据报的长度，当MTU小于当前数据报长度时，就需要分片。
 - 分片时把原IP首部拷贝到每一片
 - 有些字段需重新计算，如：总长度，校验和，标志和片偏移。
- 

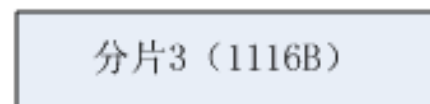
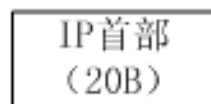
举例：IP报文分片

- 路由器收到了一个总长度为4096字节（4KB）IP数据报，首部20个字节。需要转发到一条MTU为1500字节的链路上（如以太网）
- 对原始的IP数据报4076（4096-20）字节的数据部分进行分片。分片后每一片携带20字节的IP首部，数据部分将分为3片依次为：1480字节、1480字节、1116字节。

数据报/片	总长度	标识	DF	MF	偏移
原始数据报	4096	789			0
数据报第1片	1500	789	0	1	0
数据报第2片	1500	789	0	1	185
数据报第3片	1136	789	0	0	370




拷贝IP 首部




5.3 因特网上的地址机制

- IP地址及IP报文的寻址
 - 子网编址
 - 无分类的域间编址CIDR
 - 特殊用途的CIDR地址块
 - 地址解析协议ARP
 - 网络地址转换NAT
- 

5.3.1 IP地址及IP报文的寻址

- 以IP地址为基础的网络寻址机制
 - IP地址由因特网编号分配机构(Internet Assigned Numbers Authority, IANA) 分配
 - IP地址的编址方案经历了：固定的分类编址、子网、无分类的域间选路（Classless Inter Domain Routing, CIDR），IP地址空间的管理越来越趋于合理、灵活和高效。
- 

IPv4地址的格式及分类

- IPv4地址长度为32比特（4字节）
 - 建立在两级地址结构的基础上
{<网络号>, <主机号>}
 - IP地址分类为：
 - ✓ A类地址
 - ✓ B类地址
 - ✓ C类地址
 - ✓ D类地址
 - ✓ E类地址
- 

IP地址的格式



地址空间

地址类别	覆盖的地址空间	网络号	该类网络的个数	该类网络最大主机数
A	1.0.0.0 ~ 126.255.255.255	1~126	126 (2^7-2)	16777214 ($2^{24}-2$)
B	128.0.0.0 ~ 191.255.255.255	128.0~191.255	13384 (2^{14})	65534 ($2^{16}-2$)
C	192.0.0.0 ~ 223.255.255.255	192.0.0~223.255.255 (其中192.0.0 保留)	2097152 (2^{21})	254 (2^8-2)
D	224.0.0.0~239.255.255.255			
E	240.0.0.0~255.255.255.255			

特殊含义的IP地址

网络号	主机号	含义	用途及举例
全0	全0	未知本机IP地址时，用来指代本网本主机	动态分配IP地址情况下，尚未获得IP地址时用全0代表自己的IP地址
全0	给定的主机号	本网络中的某主机	0.0.128.64，本网络中主机号为128.64的那台主机
全0	全1	本网广播地址（路由器不转发）	0.0.0.255，在本网络中的所有主机
给定网络号	全1	对给定网络上所有主机的广播地址	202.204.74.255在网络202.204.74中的所有主机
全1	全1	受限广播（Limited Broadcast），路由器不转发	在不知道本网络号及掩码时，向本网广播（如寻找IP配置服务）。

IP数据报的寻址原理

- 路由器的不同端口连接不同的网络，将收到的数据报转发到正确的下一站，这就是数据报的寻址问题。
- 路由器的寻址算法依据：
 - ✓ 数据报携带的目的IP地址
 - ✓ 路由器自己的路由表

路由表

路由表的每条记录代表一条路由，通常包括：

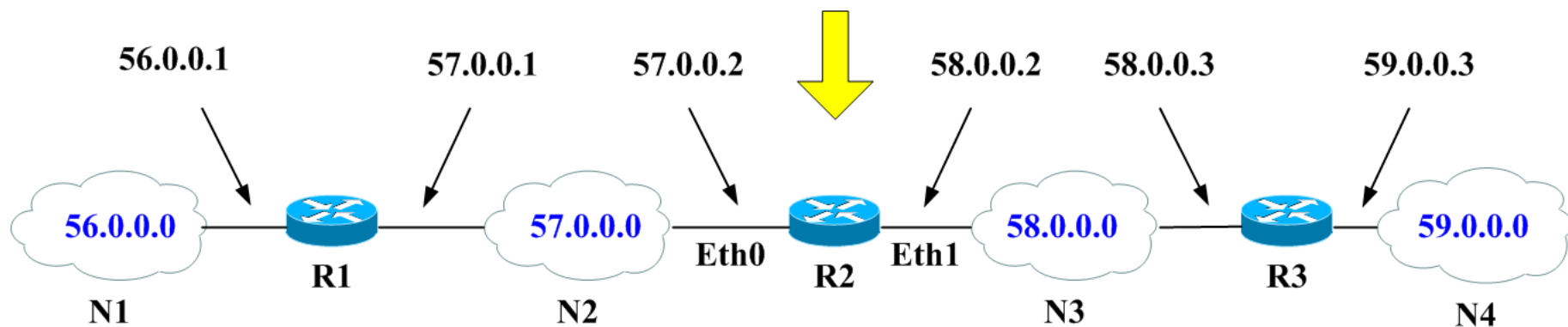
- 目的网络地址
- 下一跳（**next hop**）地址
- 该路由的开销（**cost**）或度量值（**metric**）
- 接口（**interface**）标记及对该条路由的描述



路由表举例

R2路由表

目的网络地址	下一跳
57.0.0.0	Eth0
58.0.0.0	Eth1
56.0.0.0	57.0.0.1
59.0.0.0	58.0.0.3



考虑：R1、R3的路由表是怎样的？

5.3.2 子网编址 (subnet addressing)

- 用于一个IP网络地址覆盖多个物理网络的场合
- 划分子网的必要性和好处：
 - ✓ 有效隔离网络的广播流量
 - ✓ 有效利用地址空间
 - ✓ 对外的路由汇聚：多个内部子网对外仍呈现为一个网络（一条路由）。既提高了内部网络管理的方便性，又提高了外部网络路由的效率。

子网的划分


- 划分子网的规定已经成为因特网的正式标准（**RFC950**）
- 保留网络号不动，根据需要的子网数借用主机号若干位作为子网号
- 原来两级的网络地址变为三级的地址结构。划分子网后**IP**地址结构包括三部分：
<网络地址，子网地址，主机地址>

子网掩码

- 子网掩码(netmask): 32比特的1和0的序列, 对应网络号和子网号的部分为连续的1, 对应主机号的位为全0。
- 作用: 在寻址中提取目的网络号。



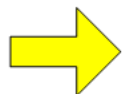
默认的子网掩码

- 为了保持算法上的一致，对不划分子网的A、B、C类地址也设置子网掩码，叫做默认的子网掩码。
 - A类地址的默认子网掩码：255.0.0.0
 - B类地址的默认子网掩码：255.255.0.0
 - C类地址的默认子网掩码：255.255.255.0
- 

子网掩码与主机网络地址

有了子网机制后，只有同时知道子网掩码，才能确定主机所在网络的网络号。

主机IP地址: 132.16.17.1

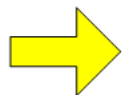


1 0 0 0 0 1 0 0 . 0 0 0 1 0 0 0 0

0 0 0 1 0 0 0 1 . 0 0 0 0 0 0 0 1

(相与)

子网掩码: 255.255.240.0

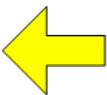


1 1 1 1 1 1 1 1 . 1 1 1 1 1 1 1 1

1 1 1 1 0 0 0 0 . 0 0 0 0 0 0 0 0



网络号: 132.16.16.0



1 0 0 0 0 1 0 0 . 0 0 0 1 0 0 0 0

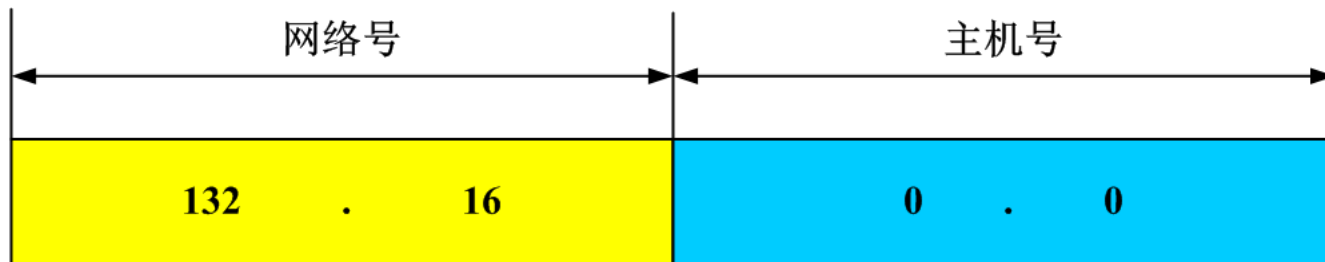
0 0 0 1 0 0 0 0 . 0 0 0 0 0 0 0 0

子网划分举例

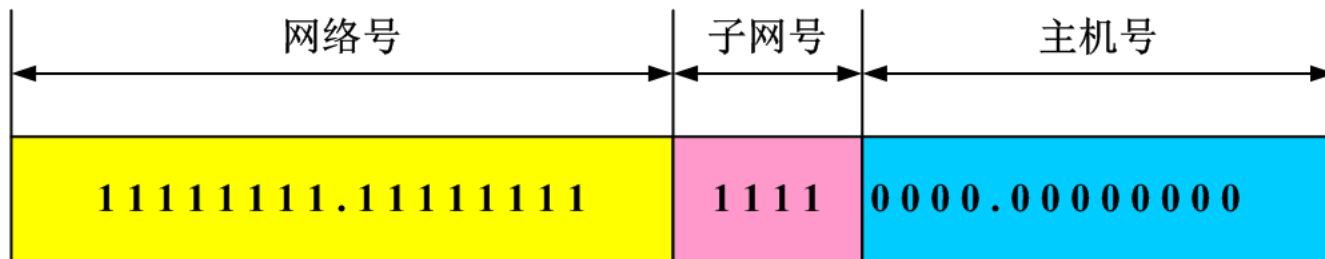
- 一个大学有10个学院，拥有一个B类的IP网络地址：132.16.0.0，需要划分子网。
- 从主机位中取4比特，共可支持 $2^4-2=14$ 个子网，剩余12比特做主机位，每个子网可容纳 $2^{12}-2=4094$ 台主机。
- 为保证子网掩码的连续，通常从主机位的高位开始选取，其三级的IP地址结构见下图所示。

子网划分举例

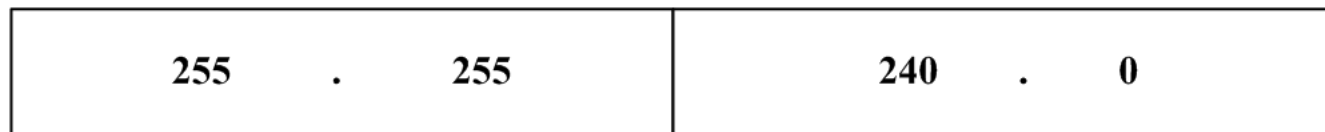
B类网络地址



子网掩码




子网掩码




推导：请按序给出10个子网划分子网后的网络号（该网络地址包含原来的网络号和子网号）

划分子网后数据报的寻址

- 标准要求所有的网络标示子网掩码，路由表也增加了子网掩码这一栏：
<目的网络地址，子网掩码，下一跳地址>
 - 网络地址的提取：将目的地址和子网掩码相“与”；
 - 对路由表的每一条记录进行操作，寻找匹配
- 

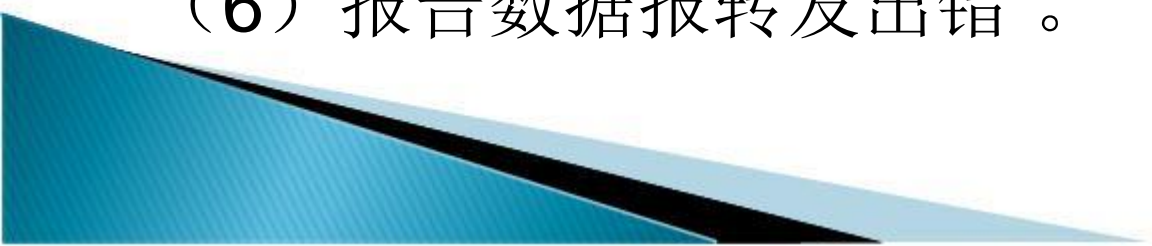
特殊路由——特定主机路由

- 以某个主机的**IP**地址为目的地址设定的路由
 - 通常是出于网络管理的需要，如：通过人为的设定，让访问某台服务器的流量必须经由某个结点转发。
 - 特定主机路由的子网掩码一般设为：
255.255.255.255。
- 

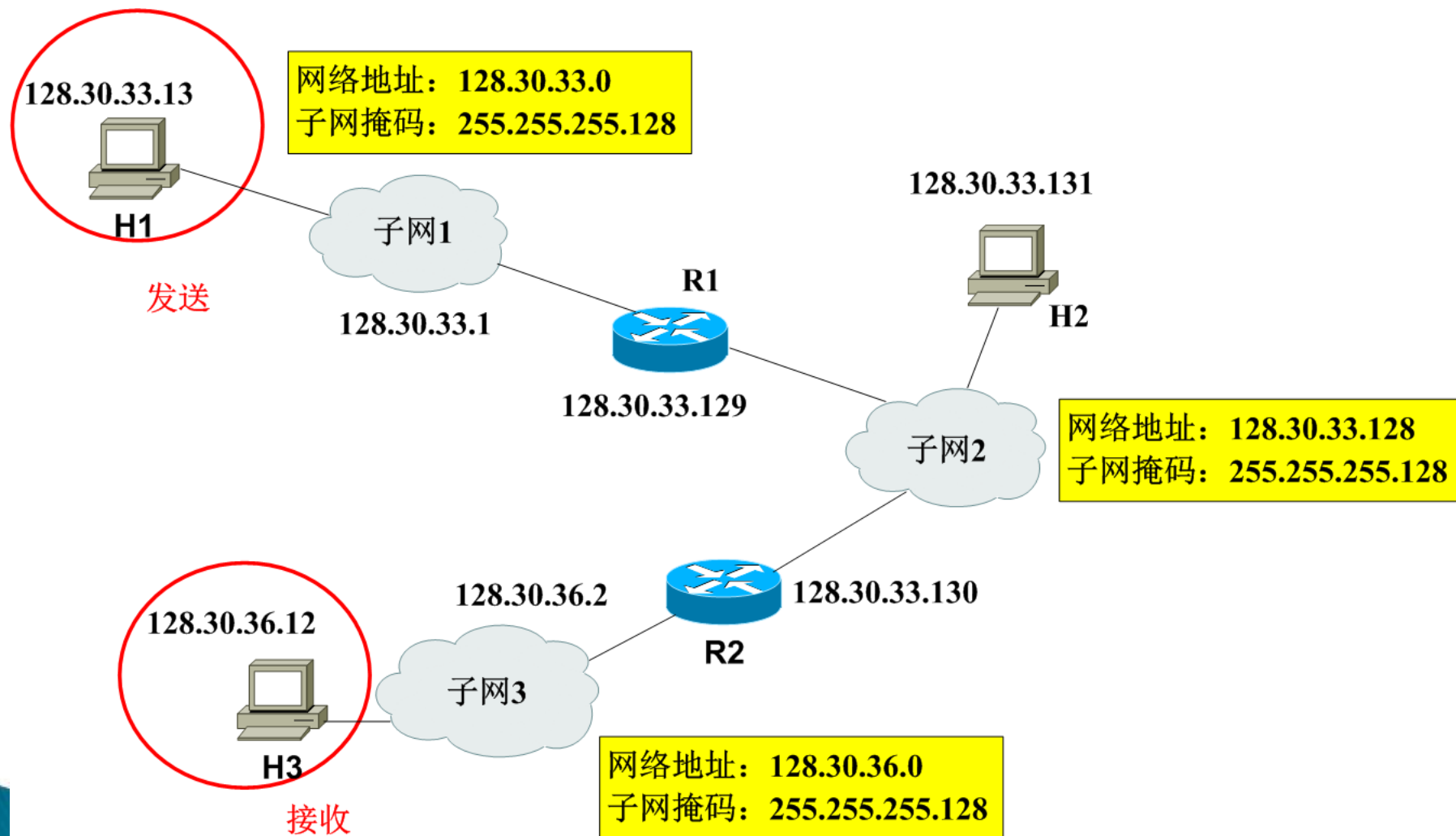
特殊路由——默认（缺省）路由

- 当数据报的目的网络地址与路由表的所有条目都不匹配时，则按默认路由转发。
- 默认路由通常是合并了路由表中下一跳相同的记录（压缩路由表），例如，把到因特网的出口作为默认路由的下一跳。
- 默认路由的目的网络地址和子网掩码通常设为0.0.0.0。
- 特定主机路由优先级最高，基于目的网络寻址的路由优先级次之，默认路由的优先级最低。

数据报寻址和转发算法

- (1) 提取IP数据报目的地址DA。
 - (2) 依次将DA与某个直连网络的子网掩码相与，若获得的网络地址与其匹配，则直接交付(通过数据链路层)；否则，执行(3)。
 - (3) 若DA与路由表中某条特定主机路由匹配，则将数据报转发到该条记录所标示的下一跳；否则，执行(4)。
 - (4) 用每条路由的子网掩码和DA相与，提取网络地址，若与该条记录的目的地网络匹配，则按该路由转发；否则，执行操作(5)。
 - (5) 按默认路由转发；若无默认路由，执行(6)。
 - (6) 报告数据报转发出错。
- 


跨IP子网的路由器寻址举例



5.3.3 无分类的域间编址CIDR

- 无分类域间路由（Classless Inter-Domain Routing, CIDR）
- 划分子网仍然没有打破分类的地址界限。
存在两个问题：
 - ✓ 因特网地址空间的分配仍然是基于A、B、C类，不够灵活，存在浪费，导致地址空间消耗过快
 - ✓ 随着网络规模的增大，因特网主干路由器的路由表增大，导致网络的性能降低

CIDR

- CIDR的基本思想体现为两点：
 1. 突破了**IP**地址分类的限制：采用**连续可变**的网络前缀，支持大小可变地址块，既可支持子网，又能够支持**超网**。
 2. 支持**路由聚合**（route aggregation）：通过地址聚合（把连续的**C**类地址合并为一个大的地址块，又被称为**超网——supernet**），将原来多条路由合并为一个，减少了**Internet**中路由表条目的数量。
- 

CIDR地址表示法

- 用一个可变长的网络前缀（prefix）来标示网络地址部分，其长度连续可变。
- IP地址的CIDR表示法：
 - 地址块的起始IP地址 / 网络前缀长度
 - 斜线前面表示该地址空间的第一个可用IP地址
 - 斜线后面表示网络地址的位数

- **CIDR** 虽然不使用子网，但仍然采用“掩码”这一名词。
- 对于 **/20** 地址块，它的掩码是 **20** 个连续的 **1**。
- 所以，从**CIDR**地址表示法可以推导出对应的网络地址和网络掩码。



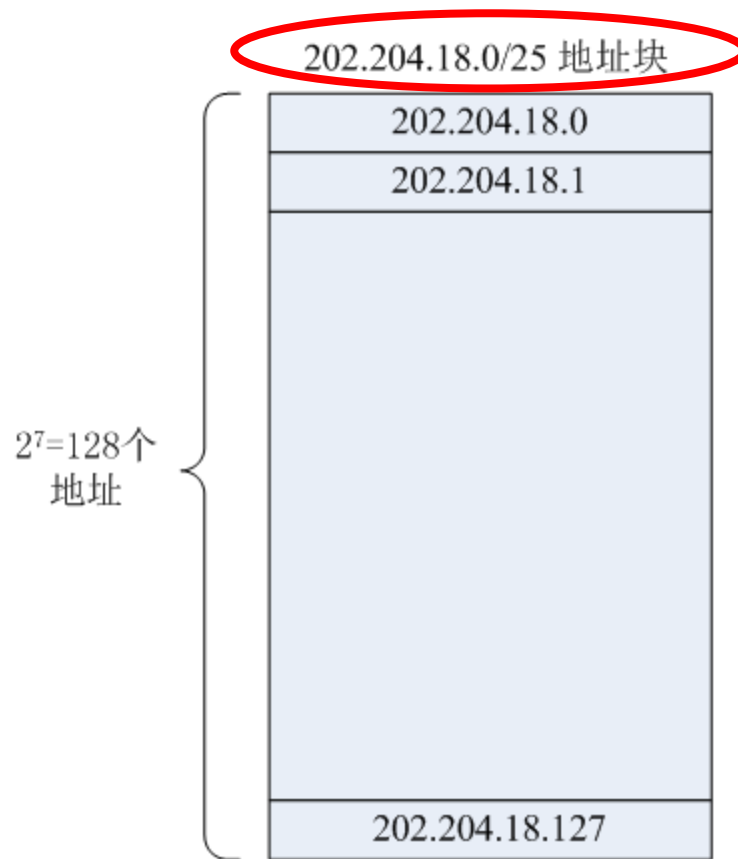
CIDR地址举例

202.204.18.0/25:

- 该地址块以202.204.18.0为第一个地址
- 所有主机具有相同的**25**位前缀，即：网络地址**25**位，主机地址**7**位。
- 网络掩码：255.255.255.128，推导：
11111111.11111111.11111111.10000000
- 网络地址：202.80.18.0，推导：将起始地址与网络掩码相与，得到网络地址。

CIDR地址举例

——地址空间



思考：202.204.18.128/25 的网络掩码和网络地址是什么？

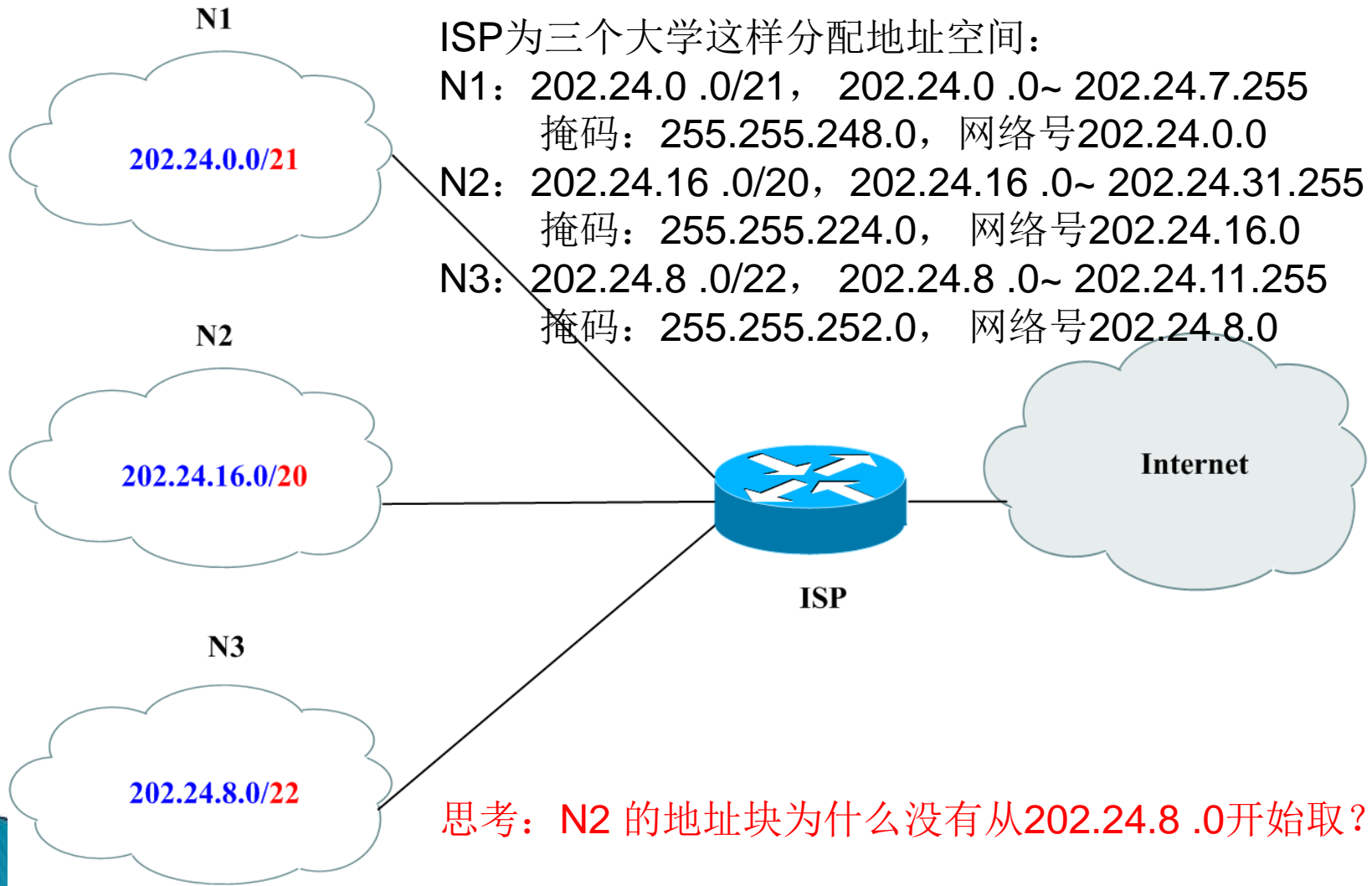
基于CIDR的地址空间分配和路由聚合

- 因特网的地址空间管理方式经历了从集中分配到分级管理的演变。
- IANA→五个RIR（Regional Internet Registry）→国家级注册机构(NIR)和本地注册机构(LIR)，我国RIR为CNNIC (Network Information Center of China,)。
- CIDR 地址块中的地址数一定是 2^n 个。
- CIDR地址块及其对应的掩码（见表5-6）

CIDR地址空间的分配举例

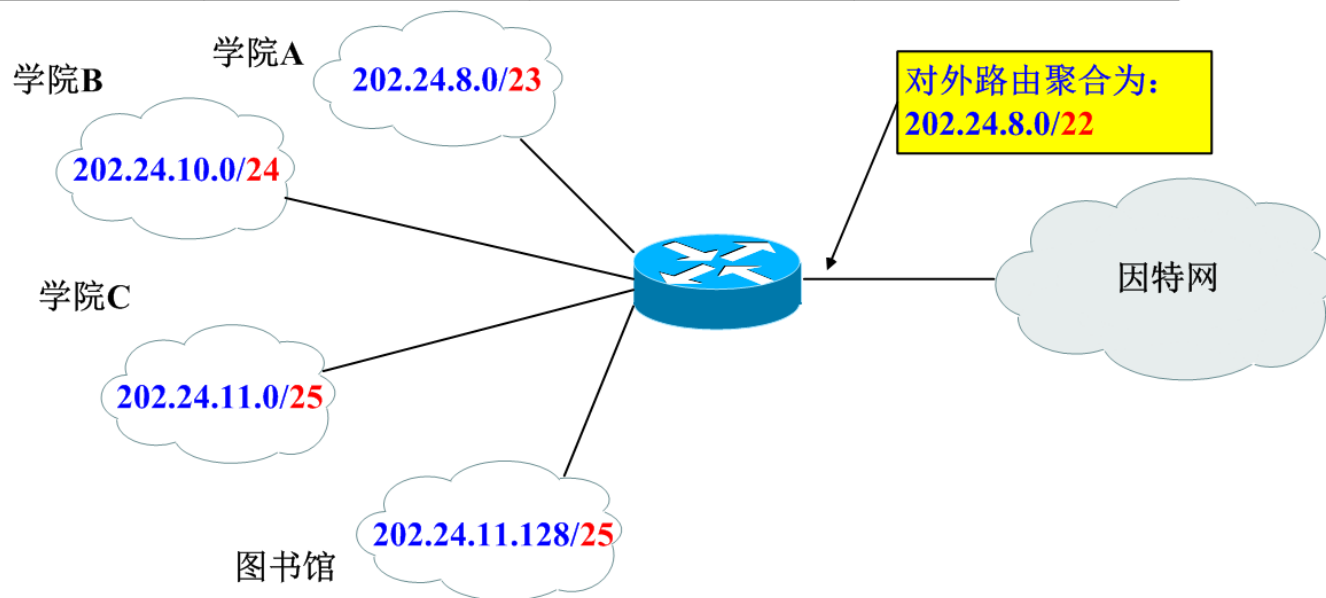
- 某ISP拥有可分配的地址块202.24.0.0/13
 - 网络号：202.24.0.0
 - 网络掩码：255.248.0.0
 - 超网：2048个C类， 202.24.0.0~202.31.255.0
- 三个大学的网络N1、N2、N3要接入该ISP
 - N1：需要2048个IP地址，即8个C类网络
 - N2：需要4096个IP地址，即16个C类网络
 - N3：需要1024个IP地址，即4个C类网络

CIDR地址空间的分配举例



路由聚合举例


单位	CIDR地址块	掩码	地址空间
大学对外	202.24.8.0/22	255.255.252.0	共4个C类网络
学院A	202.24.8.0/23	255.255.254.0	1个C类网络
学院B	202.24.10.0/24	255.255.255.0	1个C类网络
学院C	202.24.11.0/25	255.255.255.128	1/2个C类网络
图书馆	202.24.11.128/25	255.255.255.128	1/2个C类网络



5.3.4 特殊用途的CIDR地址块

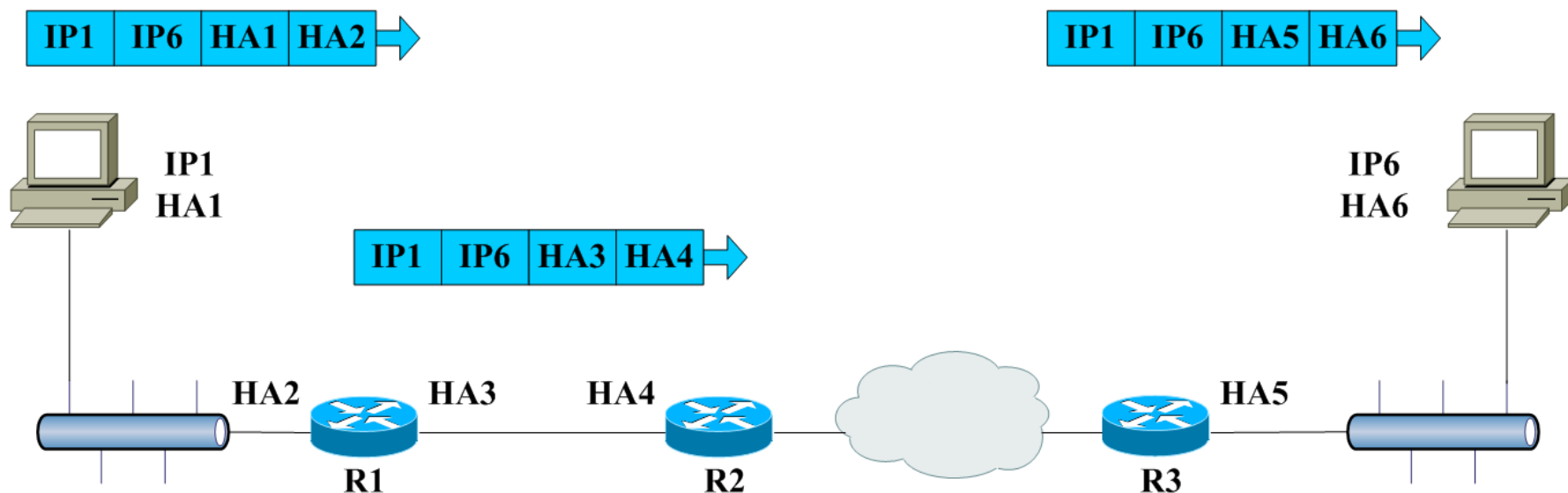
地址	特殊用途	备注
0.0.0.0/8	指示本网本主机	不会作为合法地址在因特网中出现
10.0.0.0/8	内部网络使用	
172.16.0.0/12	内部网络使用	
192.168.0.0/16	内部网络使用	
127.0.0.0/8	环回地址（Loopback）	
169.254.0.0/16	链接本地（link local）	
192.0.0.0/24	IETF 保留地址，不分配	
192.0.2.0/24	文档举例使用	
198.51.100.0/24	文档举例使用	
203.0.113.0/24	文档举例使用	
198.18.0.0/15	网络互联设备测试使用	
192.88.99.0/24	IPv6到IPv4中继的任意播	可以出现在因特网中

5.3.5 地址解析协议ARP

- ARP（Address Resolution Protocol）
 - 互连网络中，网络层的IP地址和数据链路层的MAC地址是并存的。
 - IP地址被封装在在IP数据报中，用于跨网络的端到端寻址。
 - MAC地址被封装在数据链路层的MAC帧中，用于在同一个网络中的点到点寻址。
 - 思考：分组在转发过程中，它所携带的IP地址是否发生变化？MAC地址是否发生变化？
- 

MAC地址与IP地址

- 在IP分组的传输过程中，IP地址保持不变，MAC地址则是变化的，MAC帧总是把当前发送结点的网络接口硬件地址作为源地址，把下一站的网络接口硬件地址作为目的地址。



ARP协议的作用

- ARP协议：已知IP地址查找对应的MAC地址。
- 因为：IP数据报在转发时总要封装为MAC帧，这时就需要确定目的MAC地址（下一站的MAC地址）。
- 什么时候需要ARP协议？
 - 主机发送时
 - 路由器转发时
- 主机和路由器都会用到ARP协议。

ARP协议的工作原理

1. 同一子网中的结点**A**要向结点**B**发送分组，**A**已知**B**的**IP**地址，但不知道其**MAC**地址；
2. **A**向本网络中所有结点发送以太网广播的**ARP**请求帧，询问**B**的**MAC**地址；
3. **B**收到请求后，在响应报文中加入自己的**MAC**地址，而其他主机不予响应。
4. 为了减少广播流量，主机维持一个动态更新的**ARP**高速缓存，超过生存期的记录将被自动删除。

- ARP报文长度为28字节，直接采用MAC帧（如：以太网帧）封装
- ARP请求采用MAC广播帧发送
- ARP协议支持多种网络层协议。
- ARP报文主要有两种：
 - ARP请求（操作字段值=3）
 - ARP应答（操作字段值=4）

ARP报文格式

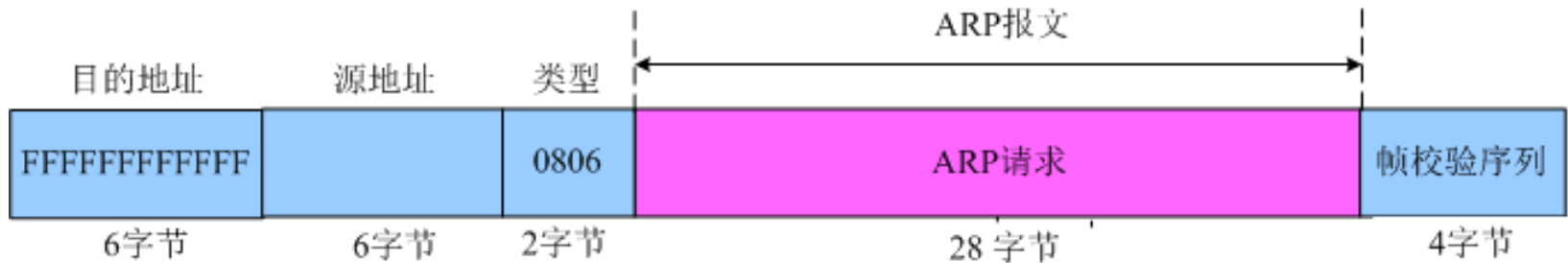
网络层协议

位0 8 16 24 31

硬件类型		协议类型	
硬件地址长度	协议地址长度	操作	可顺便保 对方的地
发送结点的硬件地址			
发送结点的硬件地址		发送结点的协议地址	
发送结点的协议地址		目的结点的硬件地址	
目的结点的硬件地址		目的结点的硬件地址	
目的结点的协议地址			

可顺便保存
对方的地址

ARP报文的封装



5.3.6 网络地址转换NAT

- 私有(内部)IP地址到合法（公开）IP地址的转换技术。
- 不仅能解决了IP地址不足的问题，还能够隐藏并保护内部网络的主机。
- NAT通常在网络的边界路由器上部署。
- 保留的内部IP地址段。

网段	地址范围	主机数
10.0.0.0/8	10.0.0.0 ~ 10.255.255.255	16 777 214个主机地址
172.16.0.0/12	172.16.0.0 ~ 172.31. 255.255	1 048 574个主机地址
192.168.0.0/16	192.168.0.0 ~ 192.168. 255.255	65 534个主机地址

网络地址转换NAT

- 基本NAT：一对一的地址转换被称为基本NAT，
- 网络地址端口转换（Network Address Port Translation, NAPT）



基本NAT

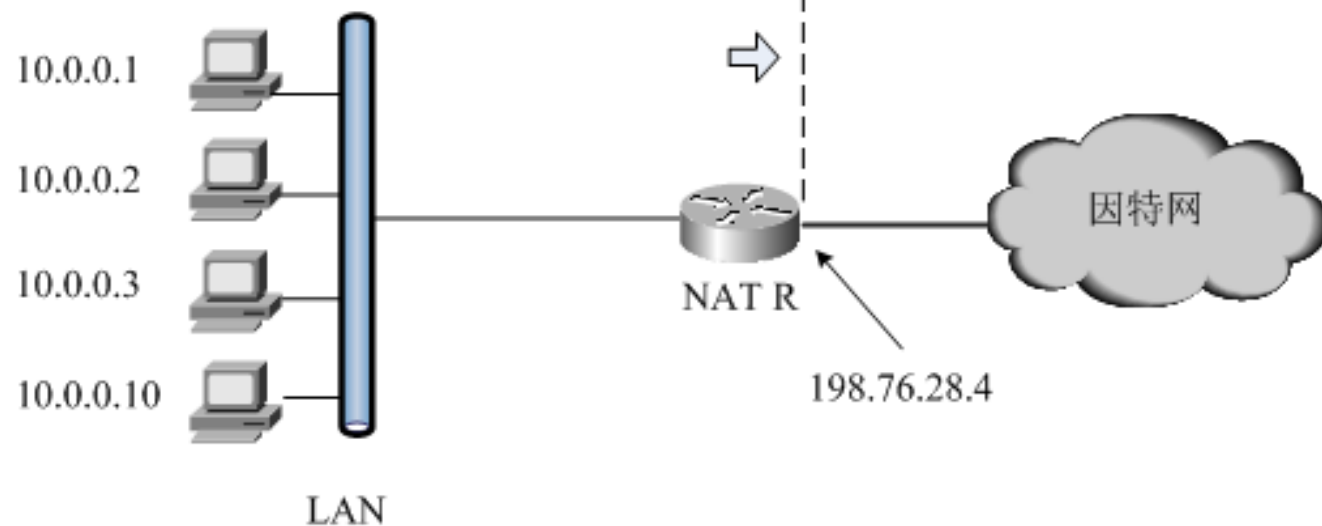
- 只做**IP**地址转换，将一组内部**IP**地址动态地与一组公开**IP**地址绑定，为每一个内部的**IP**地址分配一个临时的外部**IP**地址。
- 若有 n 个公开**IP**地址，则允许与外部进行通信的内部**IP**地址数必须小于或等于 n ，以保证每一个本地地址都能映射到一个公开**IP**地址。

网络地址端口转换（NAPT）

- NAPT需要同时转换IP地址和传输层的端口。
- 将一组内部IP地址与一个全球有效的IP地址绑定。
- 为了唯一性地标识发送进程，避免两台主机的发送进程碰巧使用了同一个源端口号，还需要进行端口号转换。



内网 IP	内网端口	外网 IP	外网端口
10.0.0.1	3017	198.76.28.4	4444
10.0.0.2	5200	198.76.28.4	5555




5.4 因特网上的路由机制

- 路由协议的基本概念
- RIP协议
- OSPF协议
- BGP协议



5.4.1 路由协议的基本概念

- 路由建立的方式：静态路由和动态路由
 - 静态路由：路由表中的每一条路由是由网络管理人工配置的，路由表中的路由是固定不变，只有当网络管理员进行配置时，路由才会发生变化。
 - 动态路由：通过在路由器之间交换彼此的网络连接信息，每个路由结点根据收到的路由信息和具体的选路算法自动建立和更新路由器表。
- 

两种路由方式的优缺点

● 静态路由

优点：无需额外的路由协议，无路由信息传递，网络资源开销小、简单、高效。

缺点：不能及时反映网络拓扑结构的变化，不具有自动调整路由的能力。

适用于小规模、网络拓扑结构简单固定的网络


● 动态路由

优点：能自动建立路由表，适应网络流量负载和拓扑结构的变化。

缺点：工作机制复杂，网络带宽开销、处理开销，算法稳定性问题。

动态路由适用于规模大、网络拓扑复杂的网络


路由算法

- 实现路由算法是路由协议的两个任务之一：
 1. 在路由器之间交换信息
 - 2. 实现路由算法。**
 - 选路算法应该兼顾：正确性、简洁性、健壮性、稳定性、公平性、最优化。
 - 两种主流路由算法：
 1. 距离向量路由（distance vector routing）算法
 2. 链路状态路由算法(link state routing)
- 

距离向量路由算法

- 采用**Bellman-Ford**算法计算路由，计算到达网络中所有目的网络的方向和距离：
 - 方向：指数据报发送的方向，表示为下一跳的地址和接口标示。
 - 距离：到达目的结点的开销度量，如**RIP**协议中用跳数、**IGRP**协议中用延时、可用带宽等，以此为依据确定最佳路径。
- 使用距离矢量路由算法的路由器需要周期性地向相邻的路由器发送自己的路由表信息。
- 典型的距离矢量路由协议：**RIP**和**IGRP**。

链路状态路由算法

- 主要采用**Dijkstra**的最短路径算法计算路由，因此，也叫最短路径优先算法。
 - 是全局算法，路由器结点向其它路由器广播自己的链路状态信息，每个路由器建立起拓扑数据库，并通过此数据库建立网络拓扑的完整信息。
 - 在拓扑数据库基础上，运行**Dijkstra**最短路径算法，计算通往各目标网络的最佳路径，构成本结点的路由表。
 - 典型的链路状态路由协议有**IS-IS**，**OSPF**协议。
- 

链路状态路由选择协议 的工作过程

1. 通告链接信息，建立完整的网络连接图：
 - (1) 确立邻接关系：路由器与它的邻居之间建立联系。
 - (2) 广播链路状态信息：路由器向每个邻居发送链路状态通告（**Link State Advertisement, LSA**）消息。在**LSA**中标识本结点、邻居结点、本结点的链路状态、链路开销度量值，使用序列号表示该通告的版本。该通告在全网的传播是采用洪泛（**Flooding**）法，即：邻居结点收到后依次向它的邻居广播。
 - (3) 建立链路状态数据库：每台路由器在数据库中保存它所收到的**LSA**的备份，如果所有路由器工作正常，那么全网结点的链路状态数据库是一致的。

链路状态路由选择协议 的工作过程

2. 执行路由算法，建立路由表

- (4) 计算最短路径：**Dijkstra**算法利用链路状态数据库对网络图进行计算得出到每个路由器的最短路径。
- (5) 填写路由表：对链路状态数据库进行查询找到每台路由器所连接的子网，并把这些信息输入到路由表中。



自治系统及分级的路由

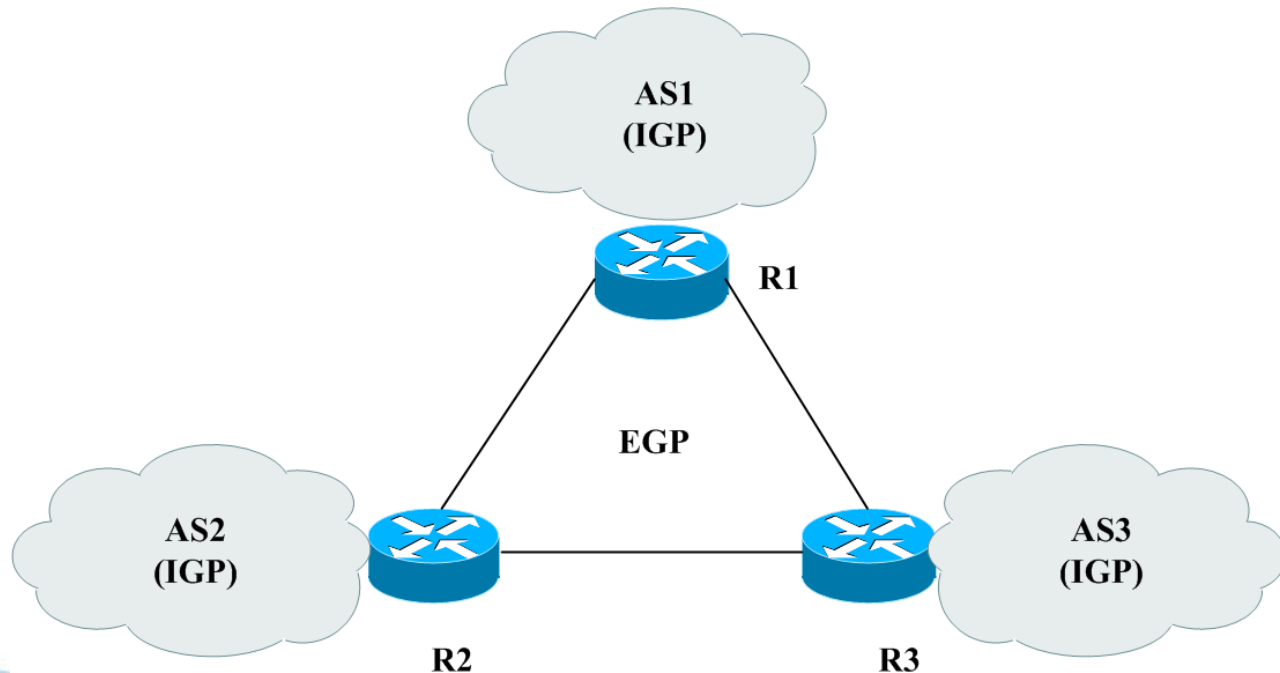
- 当网络规模巨大时，要在所有的结点之间交换海量的路由信息是不现实的。
 - 消耗过多的带宽、存储空间和处理时间。
 - 网络的管理者和运营商也希望对自己的网络管理域有某种程度的自治，如自由地选择路由协议、不向外发布自己的内部网络拓扑细节等。
- 因特网采用了分层的路由机制，整个因特网被划分为多个自治系统。

自治系统

- 自治系统（Autonomous System, AS），是一个具有统一管理机构、统一路由策略的网络区域。
- 采用自治系统这种分区域的管理方式后，路由的管理便分为了两级：AS内部的选路和跨AS的选路。
- 两类路由协议：
 - ✓ 内部网关协议(Interior Gateway Protocol, IGP)：自治系统内部运行的路由协议，常用的有RIP、OSPF。
 - ✓ 外部网关协议(Exterior Gateway Protocol, EGP)：是自治系统之间运行的路由协议，主要用于域间的路由选择，常用的是BGP和BGP-4。

分级的选路

- 由3个自治系统AS1、AS2、AS3，自治系统内部的路由器运行内部路由协议，自治系统边界的路由器同时运行内部网关协议和外部网关协议



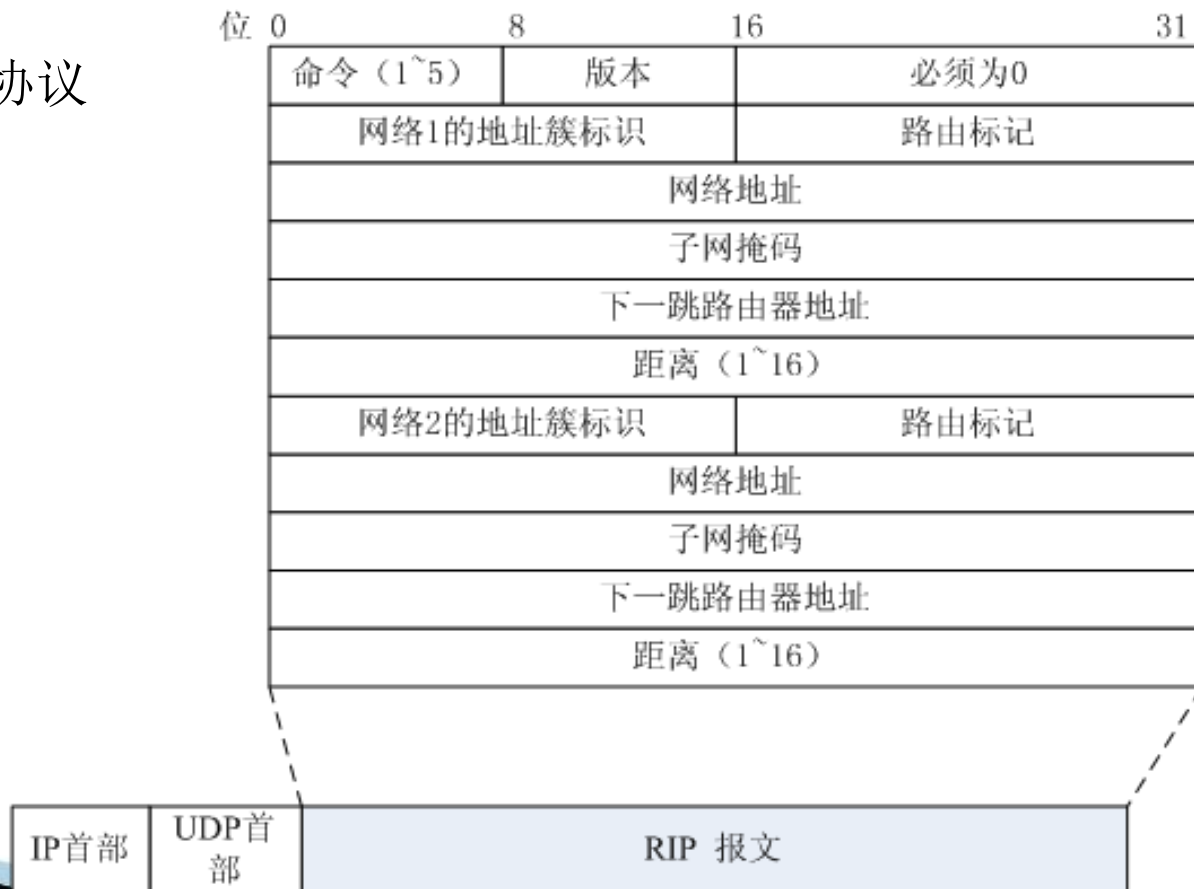
5.4.2 RIP协议

- Internet中最早的内部网关协议之一，起源于Xerox网络系统（XNS）的Xerox parc通用协议，早在1982年BSD UNIX中就包含了RIP的实现（routed, gated），RIP得到广泛应用，目前仍是常用的路由协议。
- RIP采用距离向量算法。使用跳数（hop count）作为距离的度量值。跳数是从源路由器到目的网络（包括目的网络）的最短路径所经过的网络的数量。
- RIP协议规定，一条路由的最大距离不可超过15跳，即：一条路径最多只能包含15个路由器，距离为16跳的路由被认为是不可达的（距离无穷大）。

RIP报文

- RIP报文封装在UDP协议的数据报中，使用UDP 520号端口接收路由器的路由通告信息

右图为RIPv2协议的
的报文格式



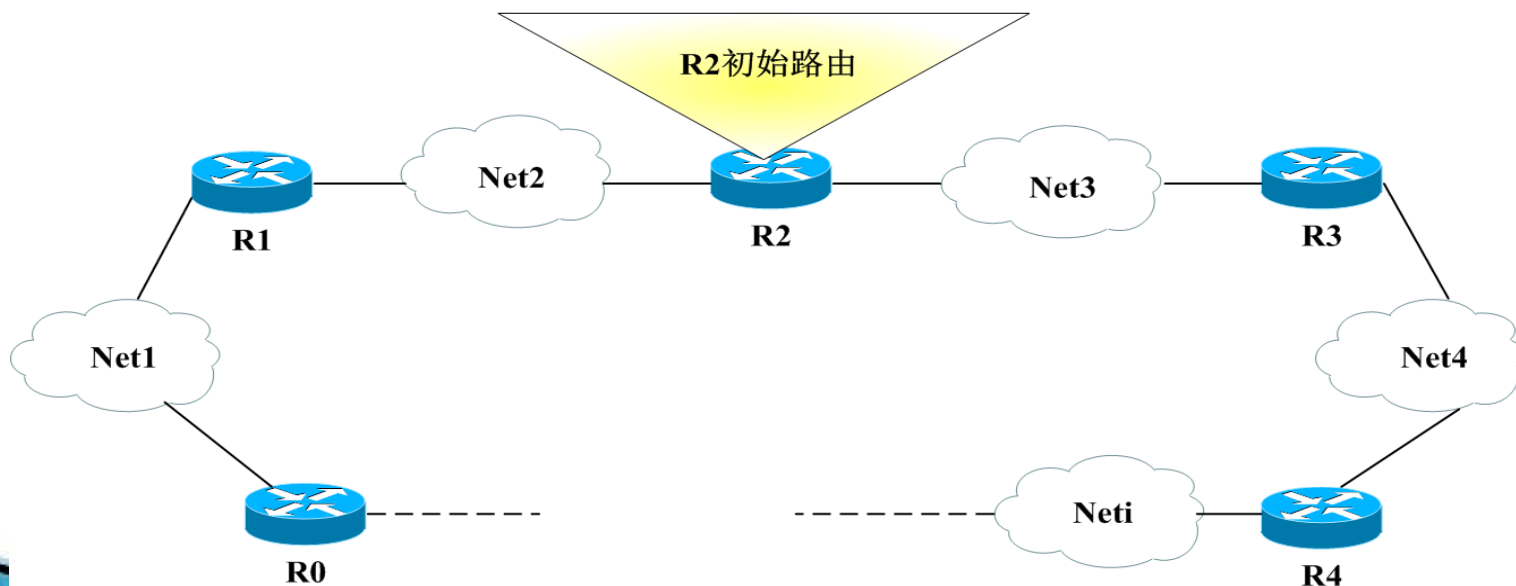
RIP协议工作原理

- 相邻路由器间每隔30秒发送一次路由信息
- Rm收到邻居Rn发来的路由信息后，首先对每条路由的下一跳都改为Rn，把距离加1。
- 路由更新分三种情况
 - 添加自己没有的路由：若某条路由的目的网络没有出现在自己路由表中，加入该条路由。
 - 替换已有的距离较长的路由：同一目的网络的路由，但下一跳不是Rn，且距离比通过Rn转发的路径长。
 - 更新已有的过时的路由：有到达同一目的网络的路由，下一跳也是Rn，则无论新路由距离长短如何，都做替换，因为Rn所连接的网络拓扑可能有变化。

RIP协议的路由更新过程例（一）

- 某网络自治系统，其部分网络结构及路由器R2的初始路由表如下图：

目的网络	下一跳路由器	距离
Net2	直连	1
Net3	直连	1



RIP协议的路由更新过程例（二）

- 30秒后，R2收到来自R1的路由通告，内容如表a，根据RIP协议的算法，R2的路由表中添加了Net1、Neti两条路由，如表b。

目的网络	下一跳路由器	距离
Net1	直连	1
Net2	直连	1
Neti	R0	4

表a R1的路由通告

目的网络	下一跳路由器	距离
Net1	R1	2
Net2	直连	1
Net3	直连	1
Neti	R1	5

表b R2的路由表

RIP协议的路由更新过程例（三）

- 随后，R2又收到来自R3的路由通告（表c），增加了到达Net4的路由、替换了到Neti的路由（因为经过R3距离为3，而原来走R1的路由距离为5），见表d。

目的网络	下一跳路由器	距离
Net3	直连	1
Net4	直连	1
Neti	R4	2

表C R3的路由通告

目的网络	下一跳路由器	距离
Net1	R1	2
Net2	直连	1
Net3	直连	1
Net4	R3	2
Neti	R3	3

表d R2的路由表

RIP协议的“慢收敛”问题

- 所谓“收敛”：当网络拓扑发生变化时，自治系统中所有的结点都建立起正确的路由选择信息的过程。
- 采用距离向量算法的路由协议存在慢收敛问题，“好消息传播的快、坏消息传播的慢”，当发现更短的路由时，结点能很快地更新路由，但当网络中某处出现故障时，要使这个网络不可达信息在所有结点的路由表中得到正确体现，需要一个较长的过程。

解决慢收敛问题的方法

- 水平分割 (**split horizon**) 的方法: 为了避免路由震荡, 任何一个结点不把从邻居学来的路由再反馈给邻居。
- 毒性反转 (**poison reverse**) 的方法: 允许把从邻居学来的路由再反馈给相邻结点, 但把该路由的距离设为无穷大, 确保邻居结点不会采用这条路由。
- 带触发更新的毒性反转 (**triggered update poison reverse**), 触发更新是指当结点发现网络不可达故障时, 马上发送路由更新报告, 不必等待30秒的更新周期。

5.4.3 OSPF协议

- 开放式最短路径优先(Open Shortest Path First, OSPF)
- OSPF是链路状态路由协议
- OSPF通过路由器之间通告网络接口的链路状态来建立链路状态数据库，生成最短路径树，每个OSPF路由器使用这些最短路径构造路由表。在一个自治系统中，所有的OSPF路由器都维护一个相同的链路状态数据库，路由器正是利用这个数据库计算出其路由表。

RIP协议和OSPF协议的比较

- OSPF协议的路由器
 - 发送什么信息？
 - 发送给谁？
 - 什么时候发送？

（RIP协议：周期性地向相邻结点发布路由信息）




RIP协议和OSPF协议的比较

- **OSPF**路由器结点需要向本区域内的所有**OSPF**结点发送信息。
- 发送链路状态信息：本结点每个接口直连网络的链接信息、链路的量度（**metric**）等，度量值可以是费用、距离、时延、带宽和其他网络管理人员设定的参数值。
- **OSPF**链路状态信息发送的频度：链路状态信息只在链路状态发生变化时才发送。


洪泛法

- 链路状态信息需要发送给本区域内的所有OSPF路由器结点，采用洪泛法（flooding）
 - 一个结点向所有相邻的结点发送链路状态信息
 - 每个相邻结点再把收到的信息发给自己的所有相邻结点（但不发给信息的来源结点），该结点的链路状态信息便逐步扩散到整个区域。
 - 所有的结点都如此，最终在整个网络路由管理区域中建立起一个统一的链路状态数据库（称为链路状态数据库的同步）。
- OSPF结点采用链路状态通告（Link State Advertisement, LSA）传送给自己的链路状态。

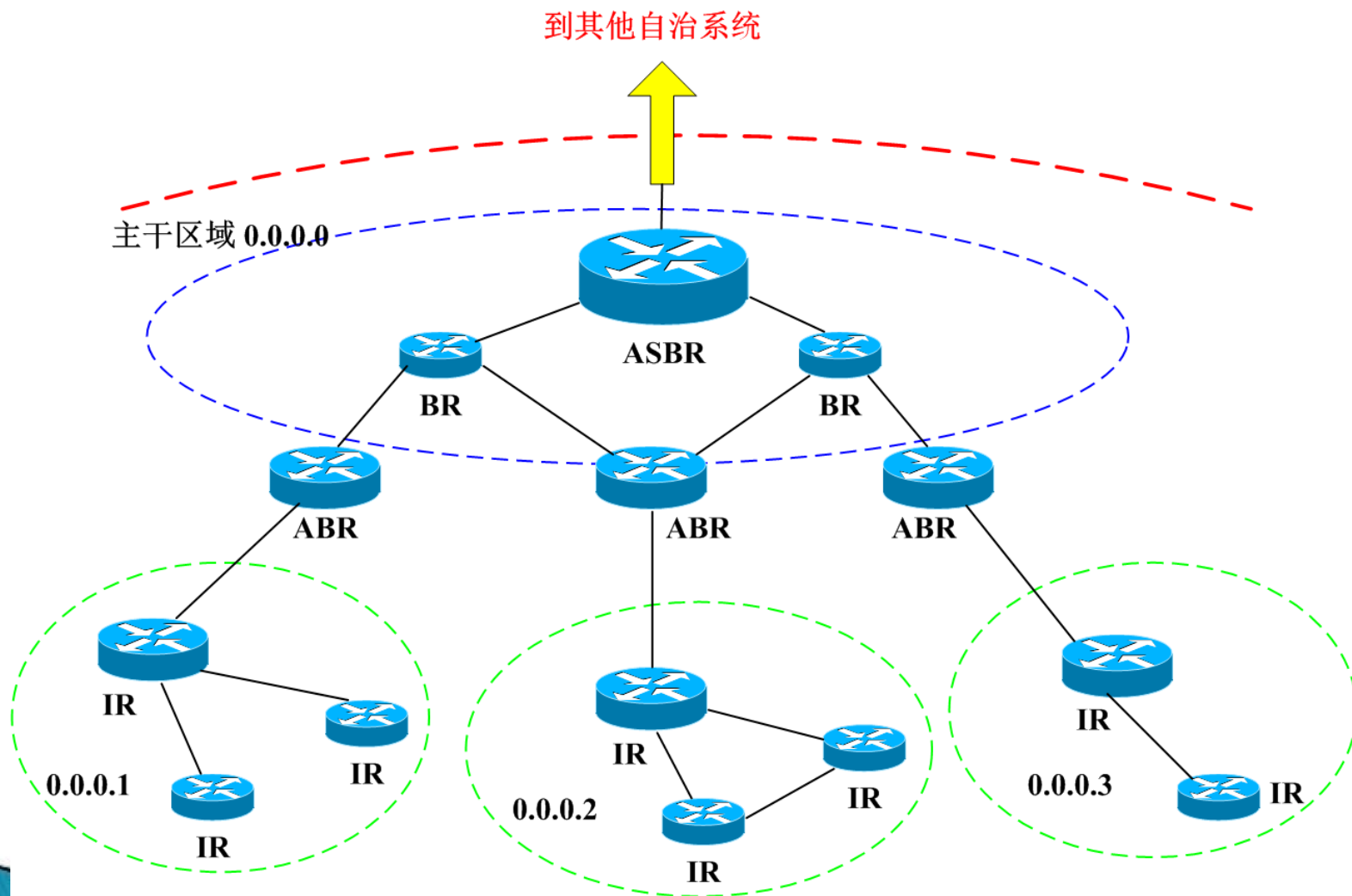
OSPF的分区

- 为了限制链路状态信息在整个网络带来大范围的洪泛流量，OSPF协议将一个自治系统（AS）划分为区（area）。
 - 区分为两级：主干区（称为0区或0.0.0.0区）和非主干区，所有区必须和主干区相连。
 - 洪泛法的链路状态信息扩散只在本区内进行，区内的网络拓扑只对区内的结点可见；跨区域之间的路由信息传递通过主干区的路由器。
- 

OSPF协议中的四种路由器

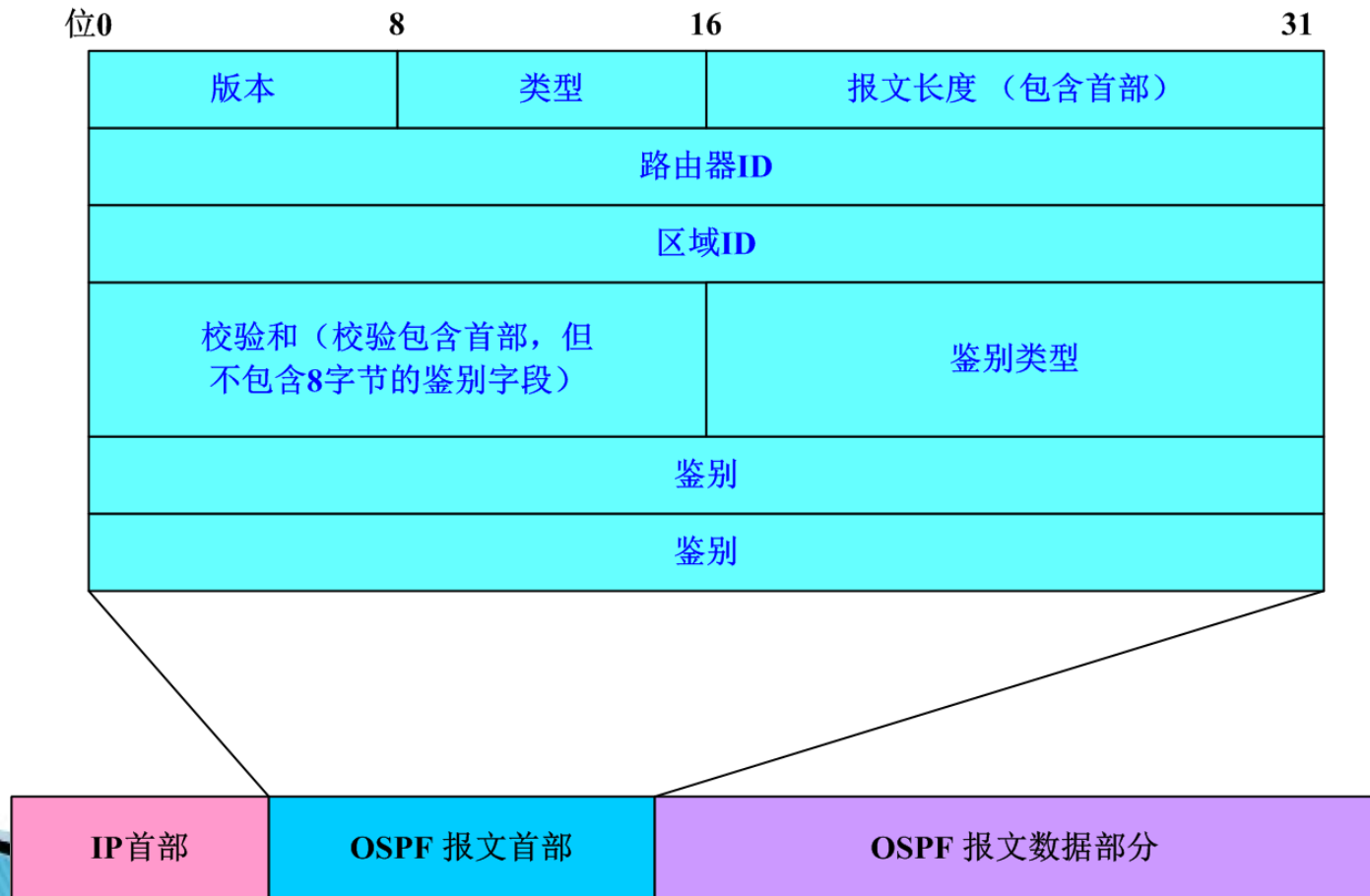
- 分区后：
 - ✓ 内部路由器（Internal Router, IR）：非主干区内路由器
 - ✓ 主干路由器（Backbone Router, BR）：主干区内路由器
 - ✓ 区边界路由器（Area Border Router, ABR）：连接不同区域的路由器
 - ✓ 自治系统边界路由器（Autonomous System Boundary Router, ASBR）：连接多个AS的路由器
- 

四个分区的 OSPF 网络结构



OSPF报文首部的格式（一）

- 目前OSPF协议为版本2，首部长度固定为24字节



OSPF报文首部的格式（二）

- 根据类型字段的值，OSPF报文有5种类型


类型	报文内容	报文描述
1	Hello报文	以多播的方式定期发送，用于建立和维护与邻居结点的联系
2	数据库描述	刚开始的邻接关系的结点，向邻居发送自己链路状态数据库的摘要
3	链路状态请求	收到数据库描述报文的结点发现自己的某个记录过时，请求发送更新信息
4	链路状态更新	该报文可携带多个链路状态通告（LSA）采用洪泛法更新整个区域的链路状态
5	链路状态确认	对收到的更新报文进行确认

5.4.4 BGP协议

- **BGP**（**Border Gateway Protocol**）是因特网中广泛采用的外部网关协议，用于多个自治系统（**AS**）之间的路由选择。
- 路径向量协议（**path vector protocol**）：**BGP**的选路算法既不属于链路状态算法，也不属于距离向量算法，**BGP**不采用内部网关协议常用的一些度量值（**metrics**）去衡量某条路径的距离或成本，而是基于路径属性去考虑。



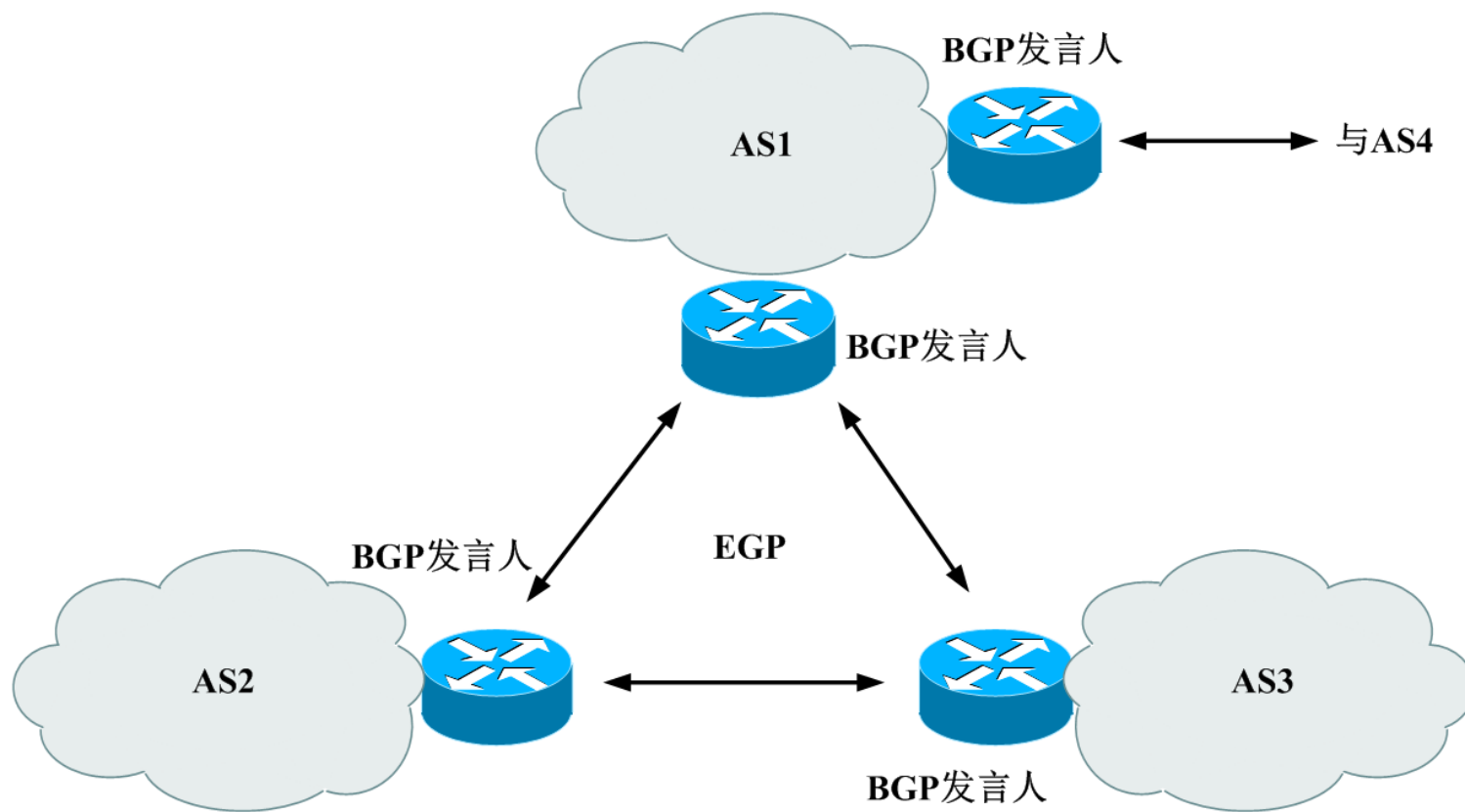
外部网关协议BGP选路的特点

- 不同的**AS**代表不同网络管理域，每个**AS**拥有对自己网络的自主管理权。
 - 跨**AS**的网络规模太大，在庞大的数据集合上去进行路由信息的交换和计算，其处理、存储和带宽的开销太大。
 - 主要任务是通过在**AS**之间交换网络可达性信息来建立跨越不同**AS**的网络层路由。
 - 每个**AS**只需要有自己的代表路由器参与信息交换就可以了。
- 

BGP的工作原理

- 在AS之间交换网络可达性信息（network reachability information）来建立跨不同AS的路由。
- 建立起BGP发言人（BGP speaker）系统，BGP发言人是自治系统中实施BGP协议的路由器，用来同其它BGP系统交换网络的可达性信息。
- 一个AS至少应该有一个BGP发言人（通常是AS的边界路由器）。BGP发言人维护一个路由信息库（RIB），RIB包含三部分：
 - 邻居输入的路由信息（Adj-RIBs-In）
 - 本地路由信息库（Loc-RIB）
 - 向邻居输出的路由信息（Adj-RIBs-Out）

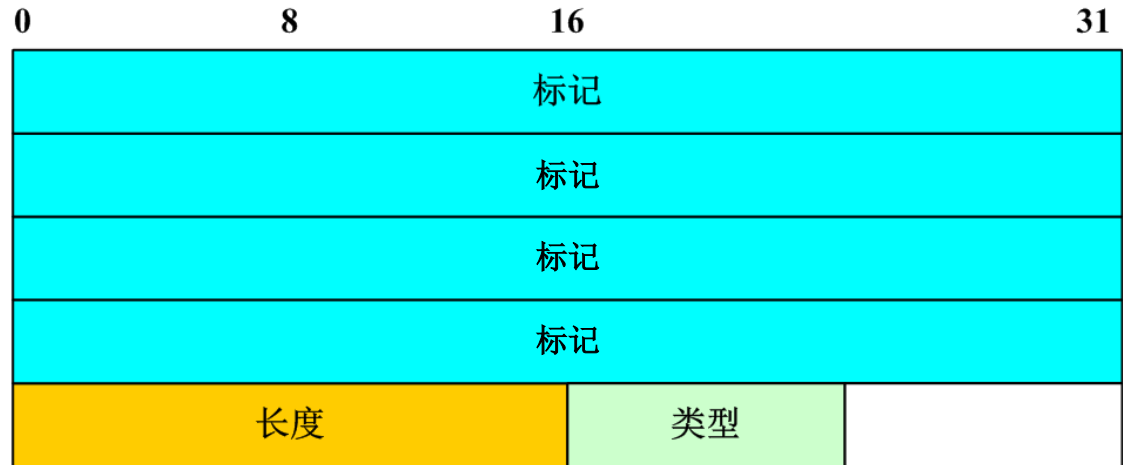
BGP发言人系统



BGPv4报文首部

- BGP采用TCP 封装
- 报文首部采用固定的格式，长度为19字节，

16字节的标记字段
必须设为全1




BGPv4报文首部

- 标记字段：规定标记字段设为全1
- 长度字段：整个报文的总长度（19~4096字节）
- 类型字段：1字节无符号整型数，四种消息报文：
 - 类型1 - OPEN报文：用来创建BGP的邻接关系。
 - 类型2 - UPDATE报文：用来交换路由信息。
 - 类型3 - NOTIFICATION报文：用于通告错误的发生，当检测到出错时，发送该报文，然后立即关闭BGP连接。
 - 类型4 - KEEPALIVE报文：用于定期检测BGP邻接路由器是否还可达。OPEN消息也需要回复一个KEEPALIVE报文进行确认。

BGP的路由更新

- 用UPDATE报文更新路由信息，UPDATE报文可以通告一条BGP发言人提供连接的目的网络路由，也可以通告多条撤销的路由。
- UPDATE报文包含5个字段：
 - 撤销路由的长度字段：2字节整数；撤销路由字段的长度
 - 撤销路由：可变长度；列出不再可行的路由的列表
 - 全部的路径属性长度：2字节；路径属性字段的长度
 - 路径属性：可变长度；列出与网络层可达性信息有关的属性
 - 网络层可达性信息（NLRI）：可变长度，可达性信息包含在一系列AS列表中。

5.5 因特网上的控制协议ICMP

- ICMP（Internet Control Message Protocol）Internet控制报文协议。
 - ICMP用于在IP主机、路由器之间传递控制消息（网络通不通、主机是否可达、路由是否可用等网络本身的消息）。
 - ICMP是TCP/IP协议族的一个子协议，是IP层的有机组成部分，每一个IP模块都必须包含ICMP的实现。
- 

5.5.1 ICMP报文

- 报文直接封装在IP数据报中，ICMP报头紧接IP报头之后。
- 一般并不把ICMP它作为高层协议看待，因为它不用于在端点间传输数据。

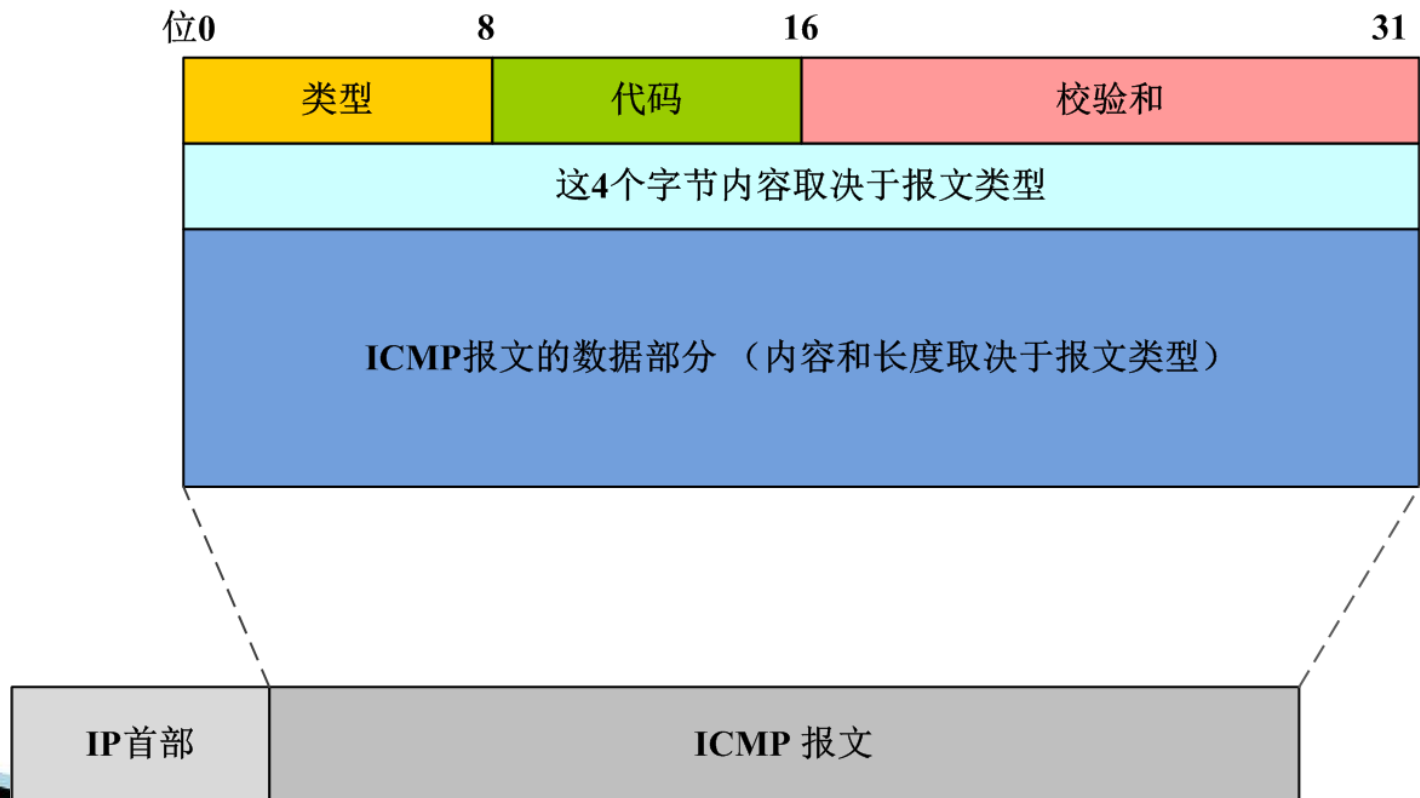


ICMP报文的格式

类型字段： ICMP消息的类型； 代码字段： 划分ICMP消息的子类型；

校验和字段： 对ICMP报头和数据进行检查；

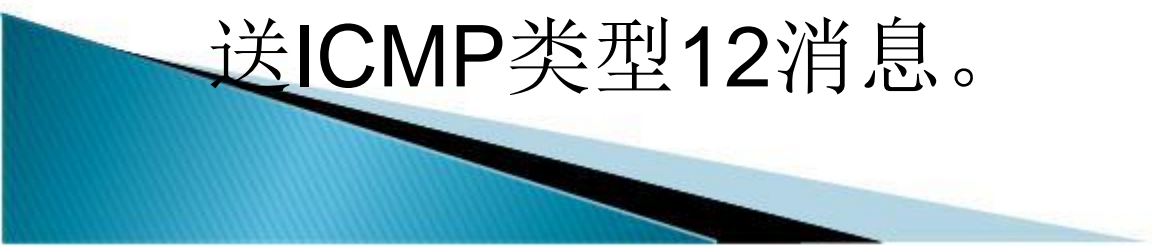
第4个字段： 内容取决于ICMP报文的类型



常用的ICMP报文类型

1. **echo请求和应答**：探测目标结点是否可达并了解有关参数。
2. **目的地不可达** 当中间结点不能交付数据报时向源结点发送**ICMP类型3**消息。
3. **源抑制** 当结点由于网络拥塞开始丢弃数据报时，就向源结点发送**ICMP类型4**消息，通知源结点放慢数据报发送速率。

常用的ICMP报文类型

4. 重定向： 类型**5**消息包含路由器对源主机的路由建议。
 5. 超时： 分两种情况，一种是**TTL**达到零；另一种是目的结点无法在给定的时间内收到一个数据包的所有分片进行重组。向源结点发送**ICMP**类型**11**消息。
 6. 首部参数问题： 收到的**IP**数据报首部字段值不正确时，丢弃该数据包，向源结点发送**ICMP**类型**12**消息。
- 

5.5.2 典型的ICMP应用实例

- ICMP的主要作用是在传输和处理IP数据报的过程中报告差错，这个过程通常是由协议栈自动启动的。
- 用于网络测试的应用程序：
 - ping
 - traceroute




5.6 因特网上的多播

- 多播的概念
- IP多播地址与硬件多播地址
- Internet上的组管理协议IGMP
- 多播的路由选择 *




5.6.1 多播的概念

- 单播（**unicast**）：以单一主机为目的的报文发送
 - 广播（**broadcast**）：以网络中所有主机为目的的报文发送
 - 多播（**multicast**）：介于单播和广播之间的一种点到多点的报文发送方式
 - 多播以**D类IP**地址为报文的目的地地址，所有加入到该多播组的主机都可以收到该**IP**报文
- 

多播的应用

源主机只用向多播组发送一份报文，多播组成员都可收到， 可用于：

- 网络视频会议
 - 流媒体（音频、视频）数据的传输
 - 网络交互式游戏
 - 分布式数据库的更新、备份数据等
- 

多播的实现

IP多播技术并不仅仅是一个D类地址问题，实现多播还需要解决：

- 多播地址的映射
- 多播组的管理
- 多播路由的选择



5.6.2 IP多播地址与硬件多播地址

- IP多播地址
- 硬件多播地址



IP多播地址

- **D类IP地址**（前4位为“1110”）：
224.0.0.0~239.255.255.255）
- **D类地址**标示一个组，需要路由器把报文拷贝并转发到通往组成员主机的链路上，因此，多播地址又被称为间接地址。
- 多播地址只能作为目的地址。
- **D类地址空间的分段管理。**
 - 224.0.0.0 ~ 224.0.0.255**：保留多播地址，可以直接使用，无需事先建立组。用于本地网络中的控制流量。
 - 224.0.1.0 ~ 238.255.255.255**：全球多播地址。
 - 239.0.0.0 ~ 239.255.255.255**：内部多播地址。

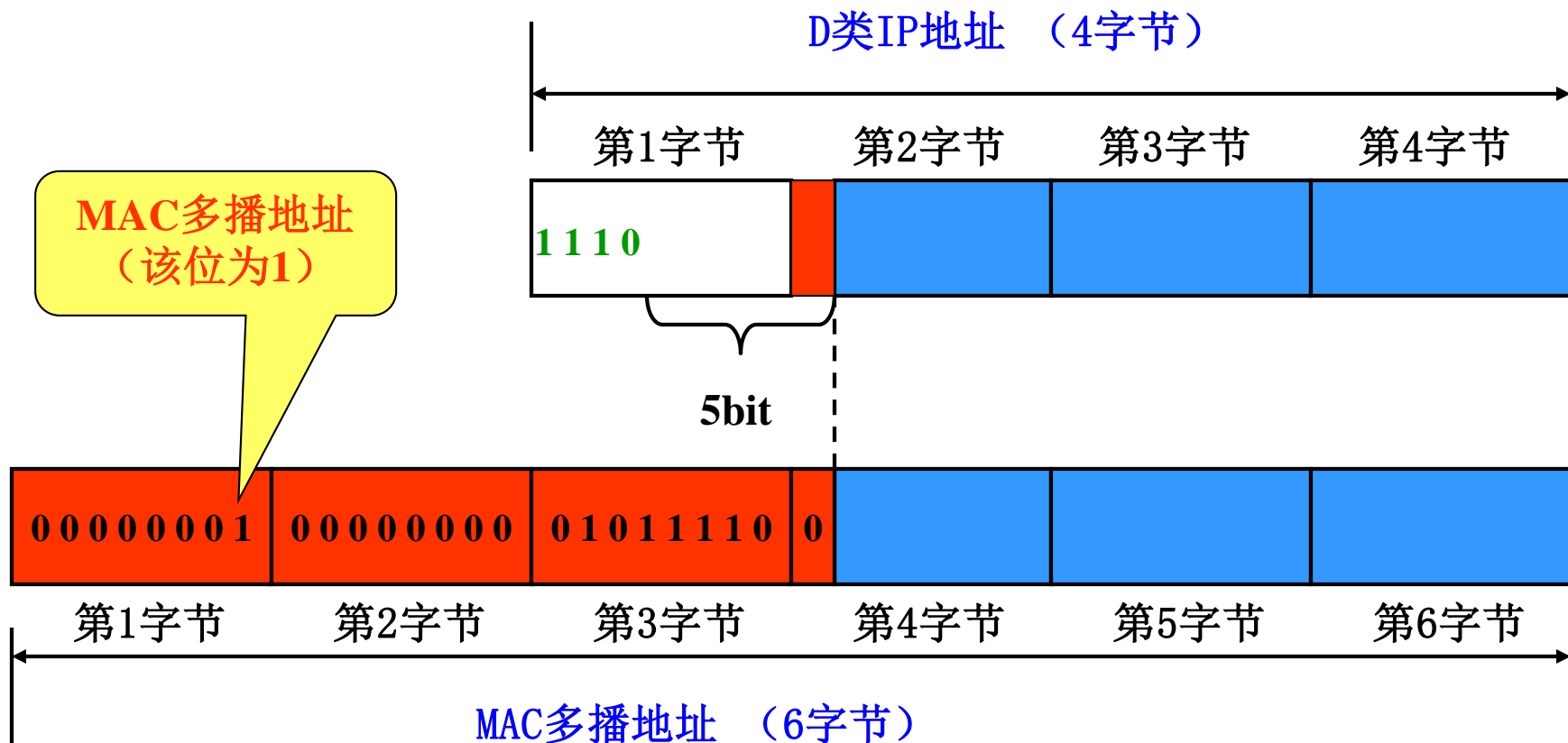
硬件多播地址

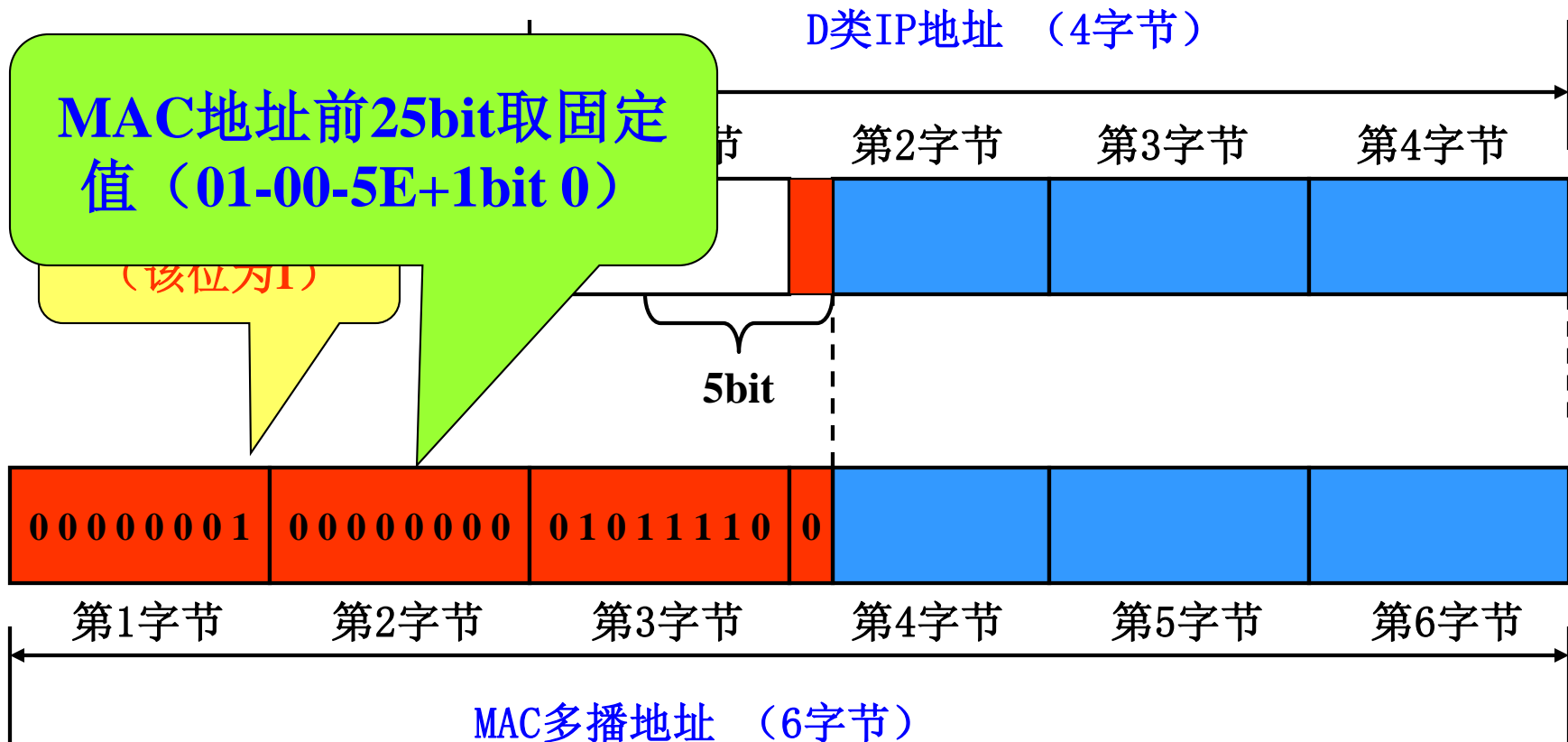
- MAC地址第1字节的最低位为1时表示这是一个多播MAC地址
- 以太网网络MAC多播地址（ IANA 分配）：
01-00-5E-00-00-00~01-00-5E-7F-FF-FF
- 规律：前**25** bit 为固定值

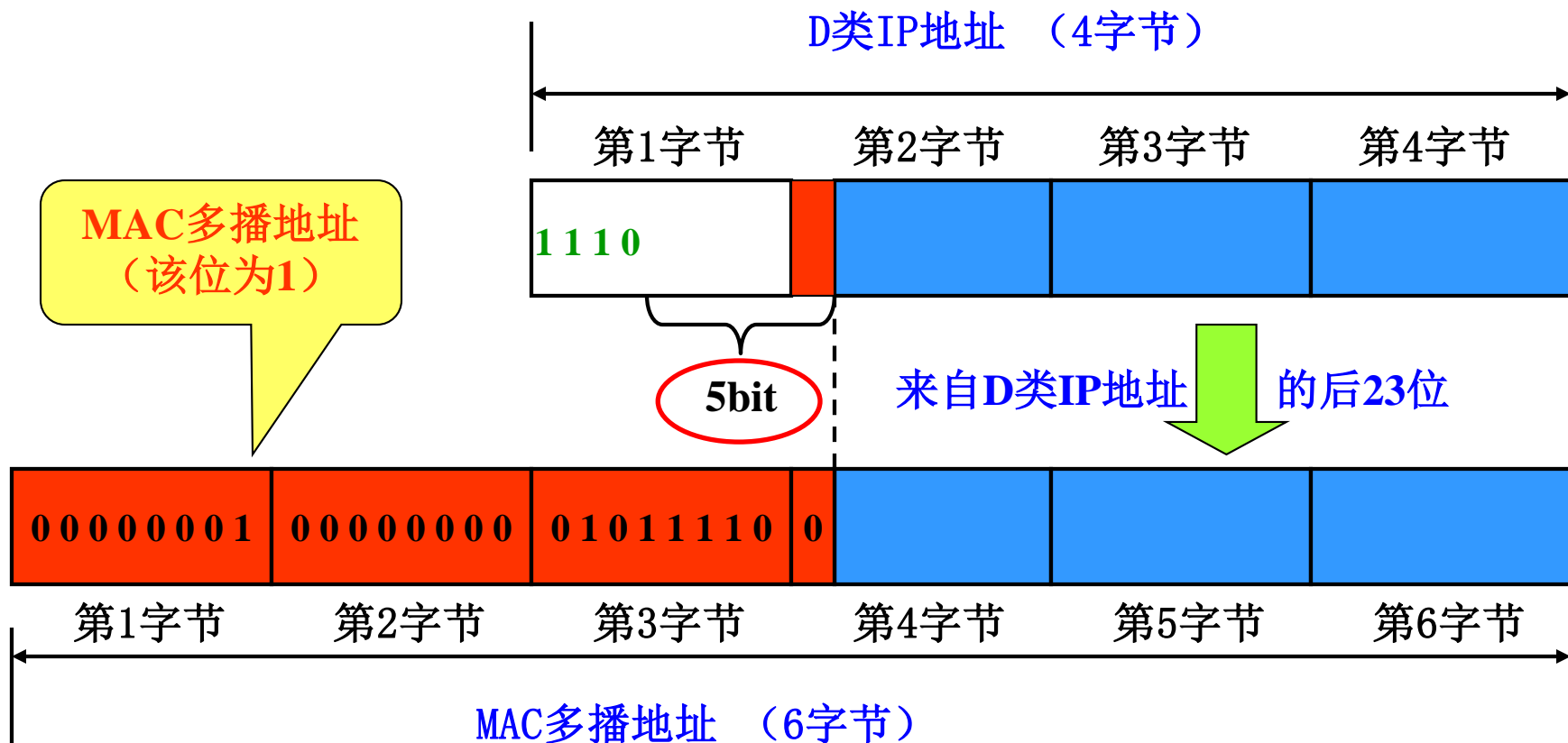


IP多播地址到MAC多播地址的映射

- IP多播地址与MAC多播地址的映射关系不是一一对应的（见下页示意图）。
 - IP多播地址有 **28**位（32-4）要映射；
 - 而MAC多播地址前25bit取固定值，只剩**23**位（48-25）
 - IP地址有**5bit**没有映射到MAC地址。
- 多个IP多播组地址被映射到同一个MAC多播地址上，需接收端在 IP层进行过滤。







5.6.3 Internet上的组管理协议IGMP

- 因特网组管理协议（Internet Group Management Protocol, **IGMP**）组播协议，用于 IP 主机向任一个直接相邻的路由器报告组成员情况。
- 让本地路由器掌握其直连网络上的多播组情况。
- 用于建立、维护组播组成员关系，所有参与组播的主机必须实现**IGMP**。
- **IGMP**有三个版本：**IGMPv1**、**IGMPv2**、**IGMPv3**

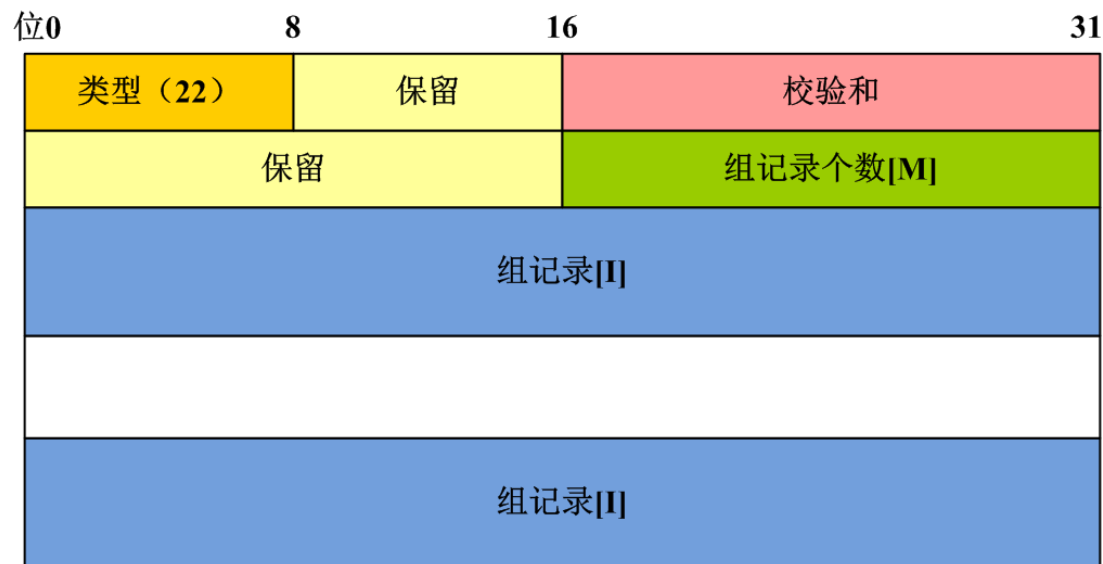


IGMPv3

- 两种消息报文：组成员的查询和报告
 - （1）成员关系探测（membership query）：路由器周期性地向自己的所有接口发送一般查询报文，发给接口上所有主机系统（224.0.0.1），以了解多播组的存在。
 - （2）成员关系报告（membership report）：主机用该消息响应路由器的查询，报告自己加入的组。此外，当一台主机新加入一个组时，主动发送成员关系报告
- 通过IGMP消息的交互，在组播路由器中建立一张组-接口对照表，记录路由器的每个接口所接入的网络中存在哪些组（成员）。

IGMP报文格式

- IP协议封装（协议字段值=2、TTL字段值=1）
- IGMP的成员关系报告报文格式及其封装：



(TTL=1, Why?)



5.6.4 多播的路由选择

- **IGMP**只用于本地网络中的动态组成员管理
- 多播路由器之间建立路由需要多播路由协议
- 多播报文的路由选择比单播复杂，主要问题：
 - 一是如何发现是哪些路由器连接了多播主机，然后建立起连接这些路由器的多播路由；
 - 二是由于多播组成员的加入和退出是变化的，所以需要管理动态的多播路由



距离向量多播路由协议（DVMRP）

- DVMRP（Distance Vector Multicast Routing Protocol），因特网中的第一个多播协议，在MBONE(multicast backbone)中得到普遍使用，是基于距离向量算法的路由协议，DVMRP是在RIP协议的基础上发展而来的。
- DVMRP协议的要点如下：
 - ①DVMRP是一个AS内的多播路由算法；
 - ②DVMRP采用IP数据包封装，协议字段值为2；
 - ③DVMRP只支持多播报文的选路；
 - ④DVMRP采用剪除的反向路径广播（Truncated Reverse Path Broadcasting algorithm, TRPB）的算法，建立基于源的多播树(Source-based multicast tree)

TRPB算法

- TRPB算法的工作原理包含两个方面：

(1) 源结点发送多播报文时，路由器首先采用广播的方法在组播范围中扩散多播报文。为了防止洪泛传播带来的重复报文，路由器只在源结点发来的报文是经最短路径到达时，才向外转发报文。这样路由器就只扩散第一次收到的报文，而不会再次转发绕道而来的重复报文

(2) 采用剪枝（**pruning**）的方法确定多播树，剪除不需要接受多播报文的树枝结点。当某个路由器不再转发多播报文时（它所连接网络中的组成员都退出了该组），便向上游结点发送剪枝消息，该结点就脱离了多播树，不再收到多播报文。其上游结点在满足两个条件的情况下进一步向上发送剪枝消息：一是从连接的所有链路上都收到剪枝消息，二是本身不连接组成员。

其它多播路由协议

- (1) OSPF的多播扩展MOSPF: 包含区域内路由、区域间路由、跨AS的路由
- (2) 协议无关的多播协议-稀疏方式 (Protocol Independent Multicast-Sparse Mode, PIM_DM)
- (3) 协议无关的多播协议-密集方式 (Protocol Independent Multicast -Dense Mode, PIM_DM)



5.7 下一代因特网协议IPv6

- IPv6的背景及主要特点
- IPv6的报文格式
- IPv6地址
- ICMPv6
- 向IPv6的过渡



5.7.1 IPv6的背景及主要特点

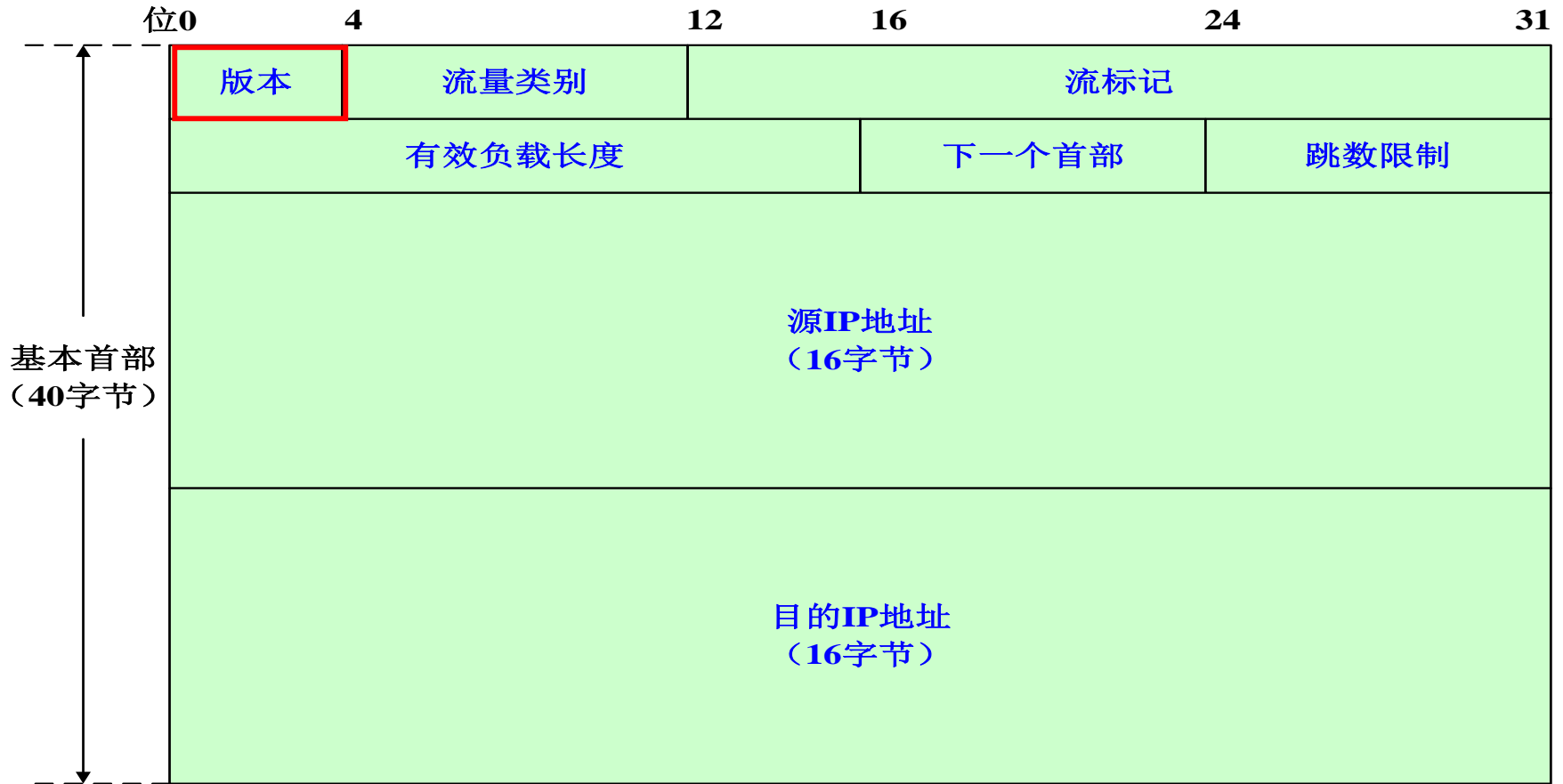
- IPv6又称为下一代因特网协议（IP next generation, IPng）。
- 解决IPv4地址资源紧缺的问题。
- 解决IPv4协议的处理效率、网络性能、安全等问题。



IPv6的主要特点

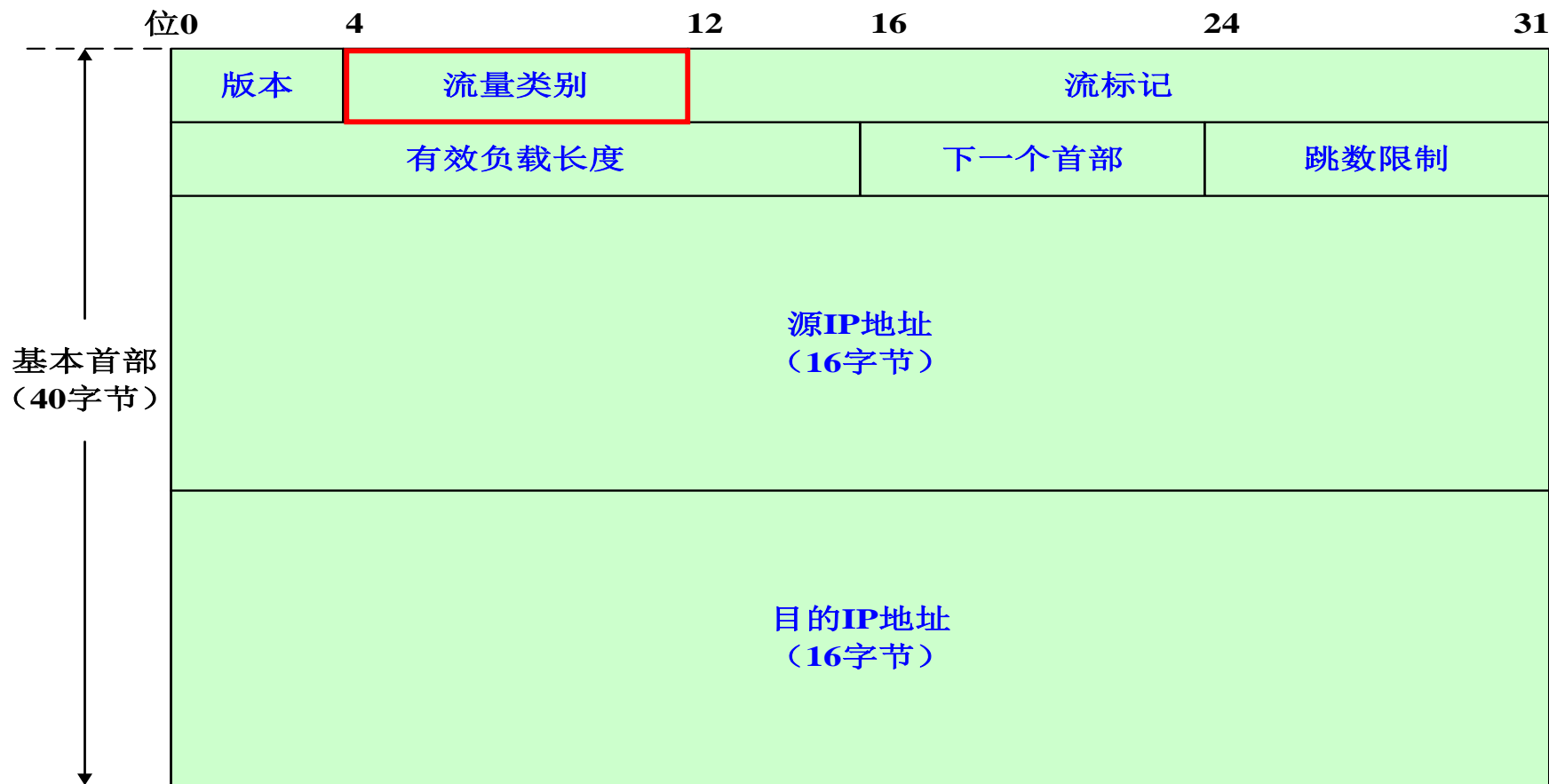
- 扩展的地址空间：IPv6地址长度为16字节
- 简化了首部：基本首部只有8个字段，取消了IPv4的首部长度的、标识、标志、片偏移、首部校验和。
- 基本首部长度的固定：基本首部长度的固定为40字节
- 选项格式更灵活高效：扩展首部用来实现选项的功能，扩展首部不作为首部看待，而是和数据部分一起作为有效载荷（pay load）。
- 提高中间结点的处理效率：除了逐跳选项，其它扩展首部不再由中间路由器处理。
- 增加了对流量的分类、安全认证等功能。

5.7.2 IPv6的报文格式



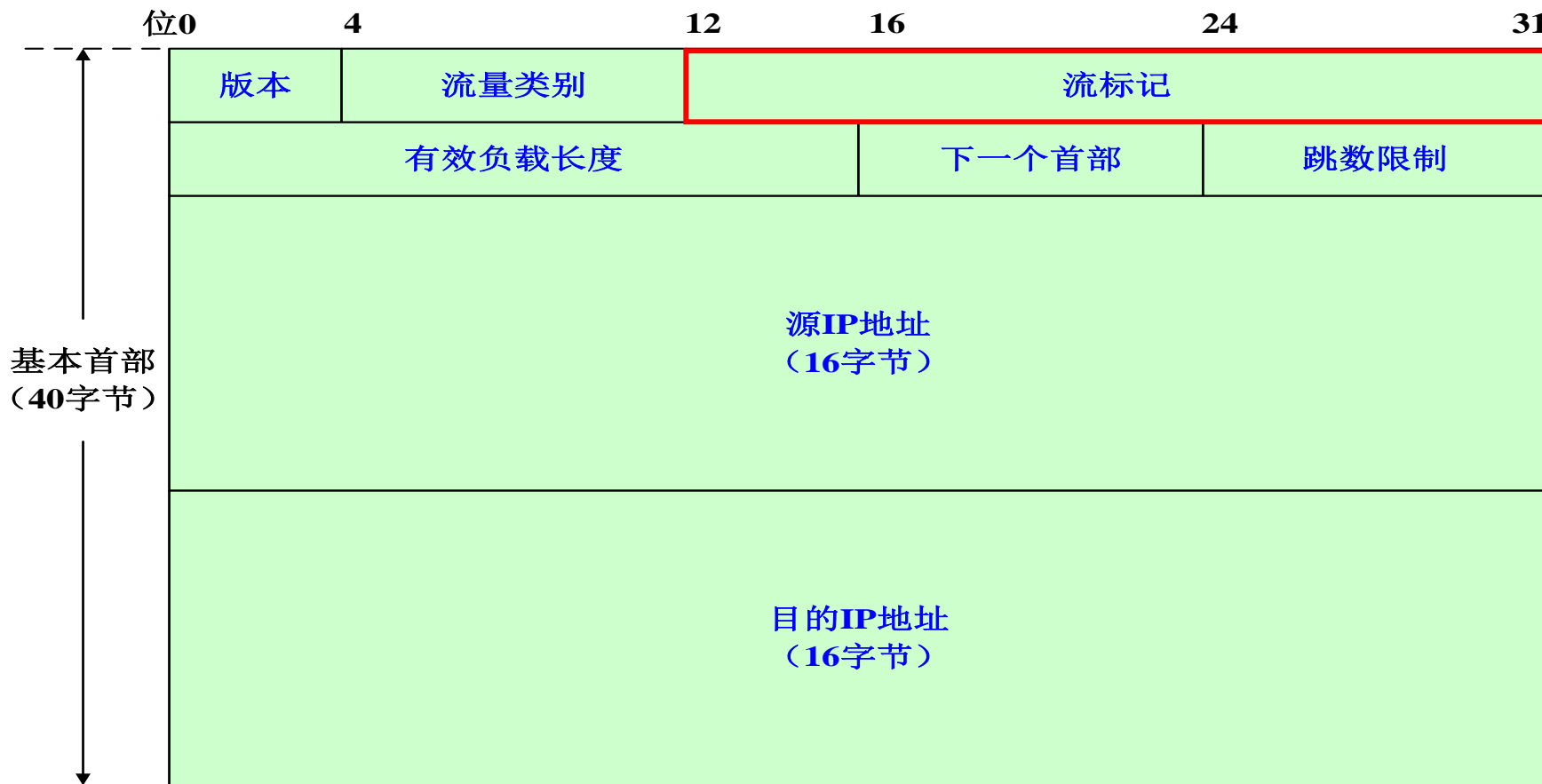
版本号：4比特。IPv6该字段为6

5.7.2 IPv6的报文格式



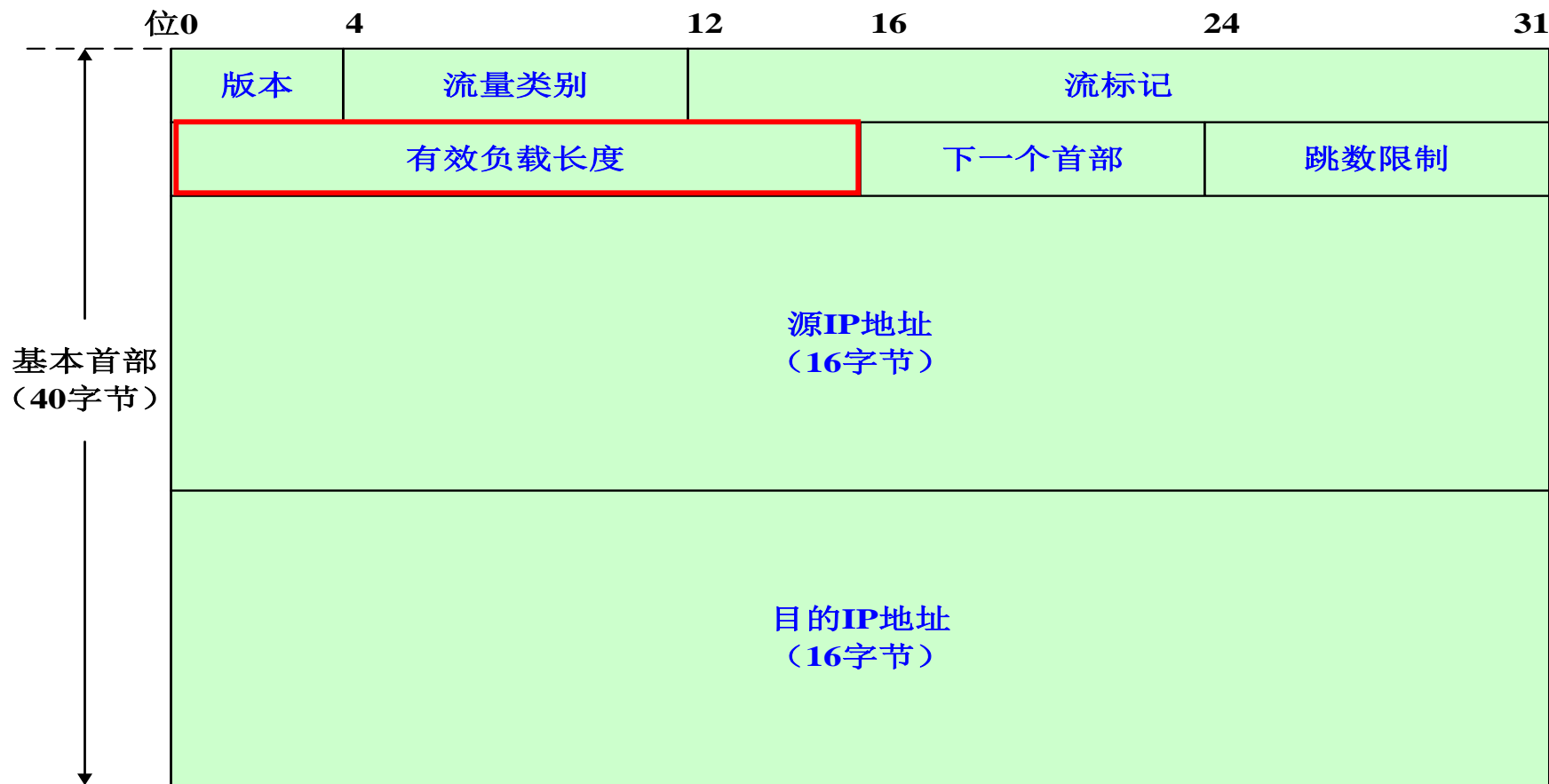
流量类型(Traffic Class): 8比特。对IPv6分组区分不同的类别和优先级

5.7.2 IPv6的报文格式



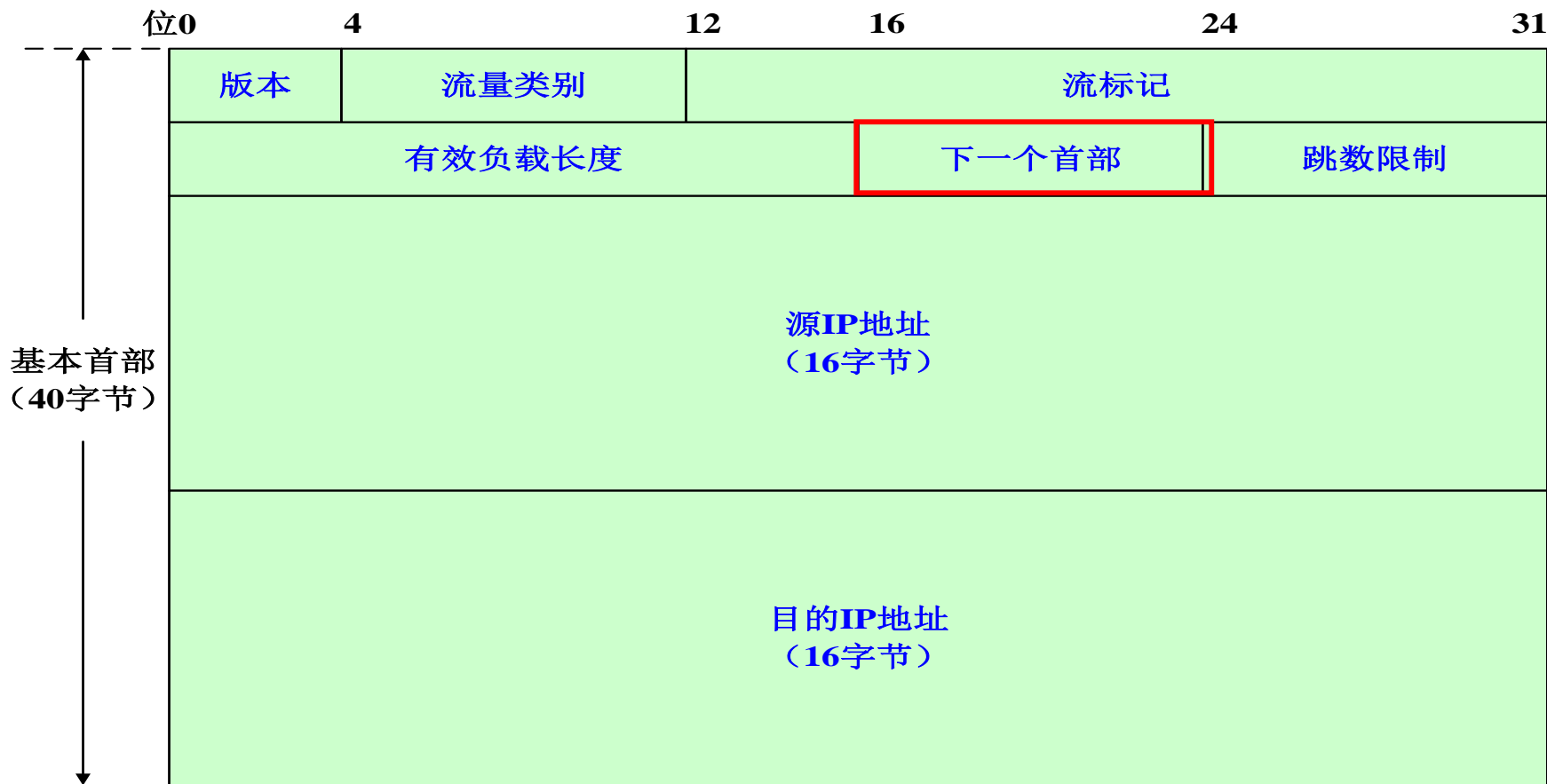
流标记 (Flow Labels): 20比特。源节点可用该字段标记IP分组，让中间的路由器提供特别的处理，以保证数据传输的QoS

5.7.2 IPv6的报文格式



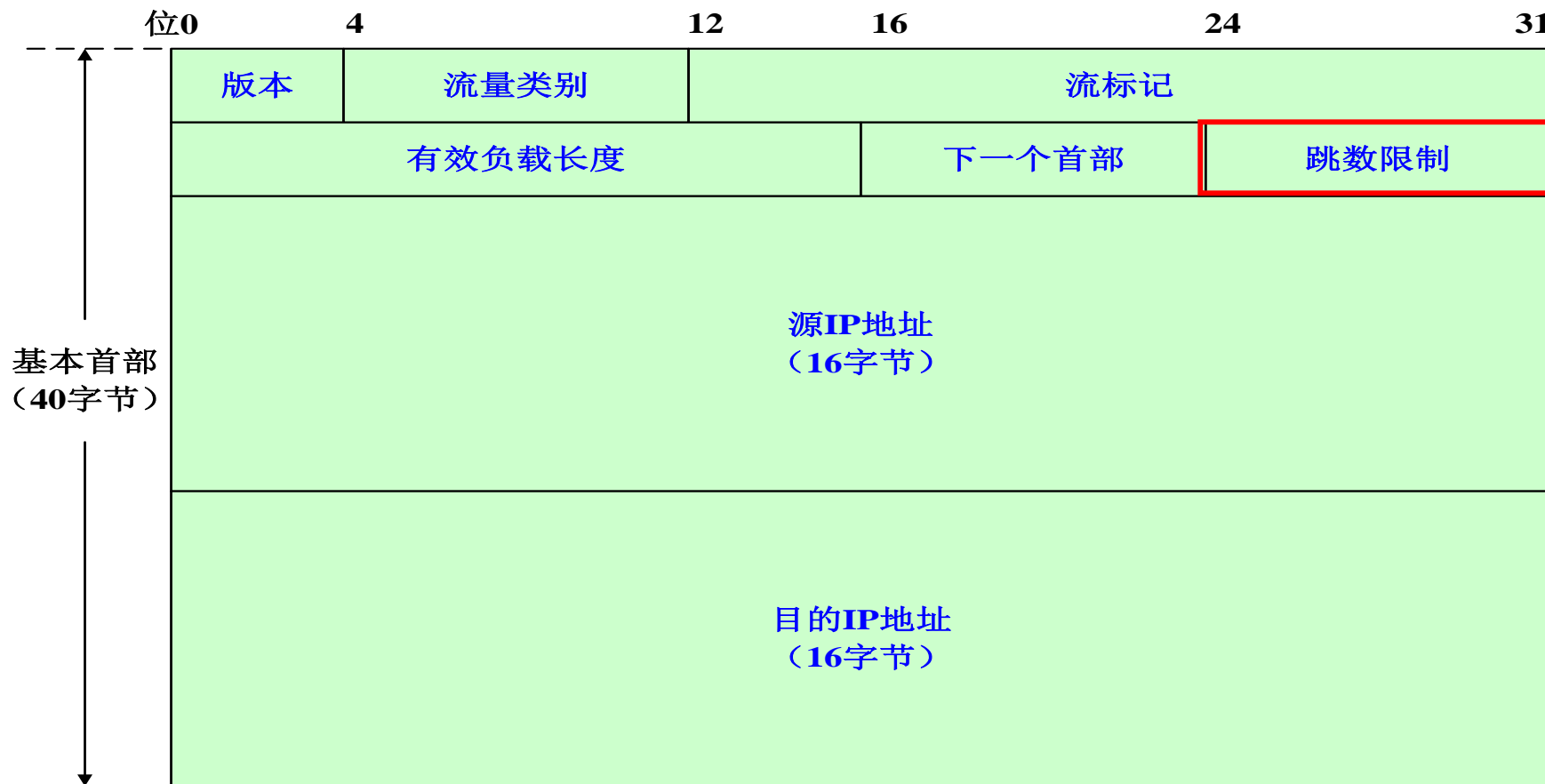
有效负载长度(Payload Length): 16比特。指明基本首部后面的有效负载的字节数（包括数据和扩展首部的长度）

5.7.2 IPv6的报文格式



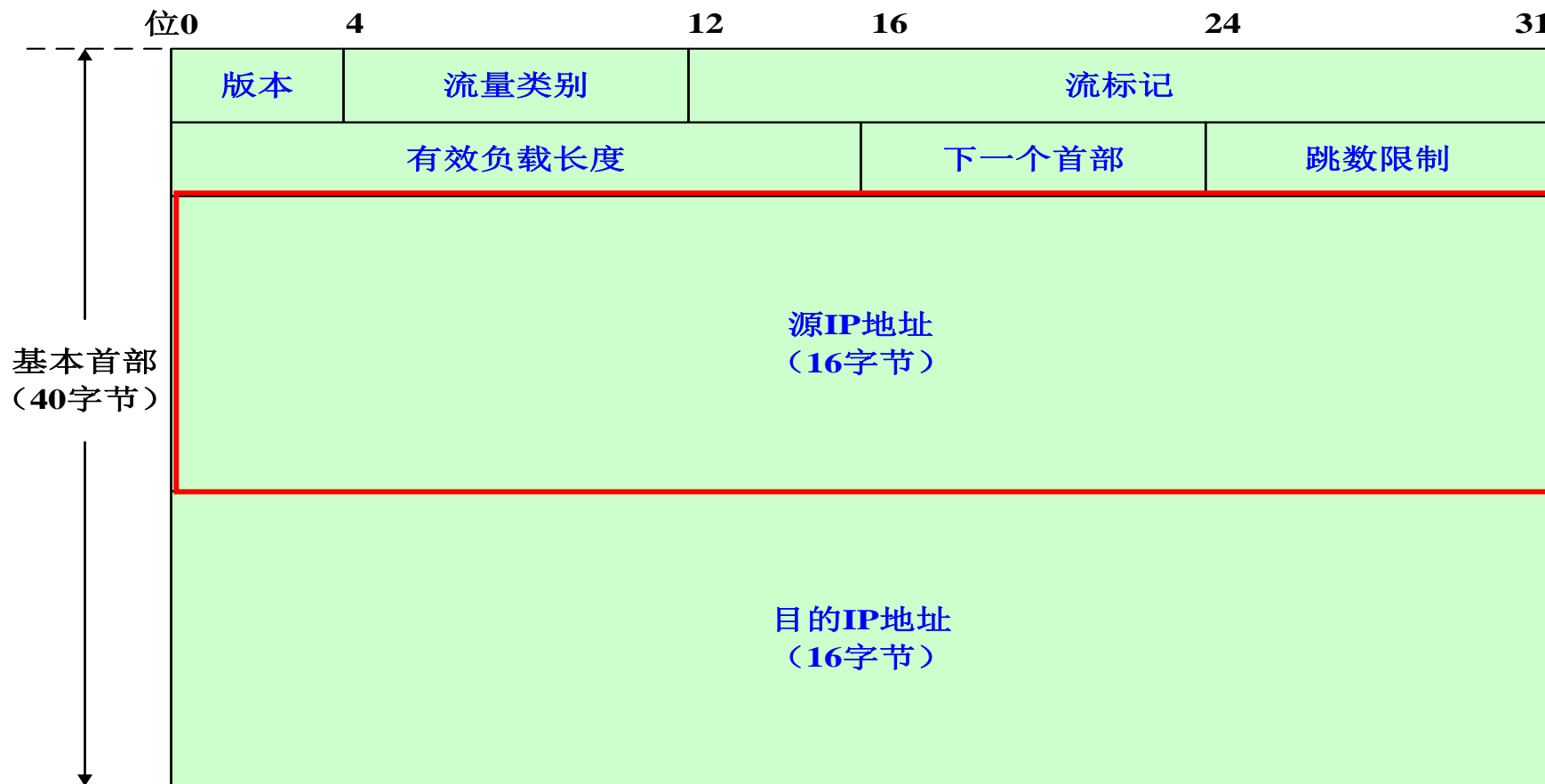
下一个首部(Payload Length): 8比特。指示紧随后面的扩展首部的类型, 扩展首部位于基本首部和高层协议(如UDP/TCP)首部之间, 因此当后面没有扩展首部时, 该字段的取值和含义与IPv4首部的“协议”字段相同

5.7.2 IPv6的报文格式



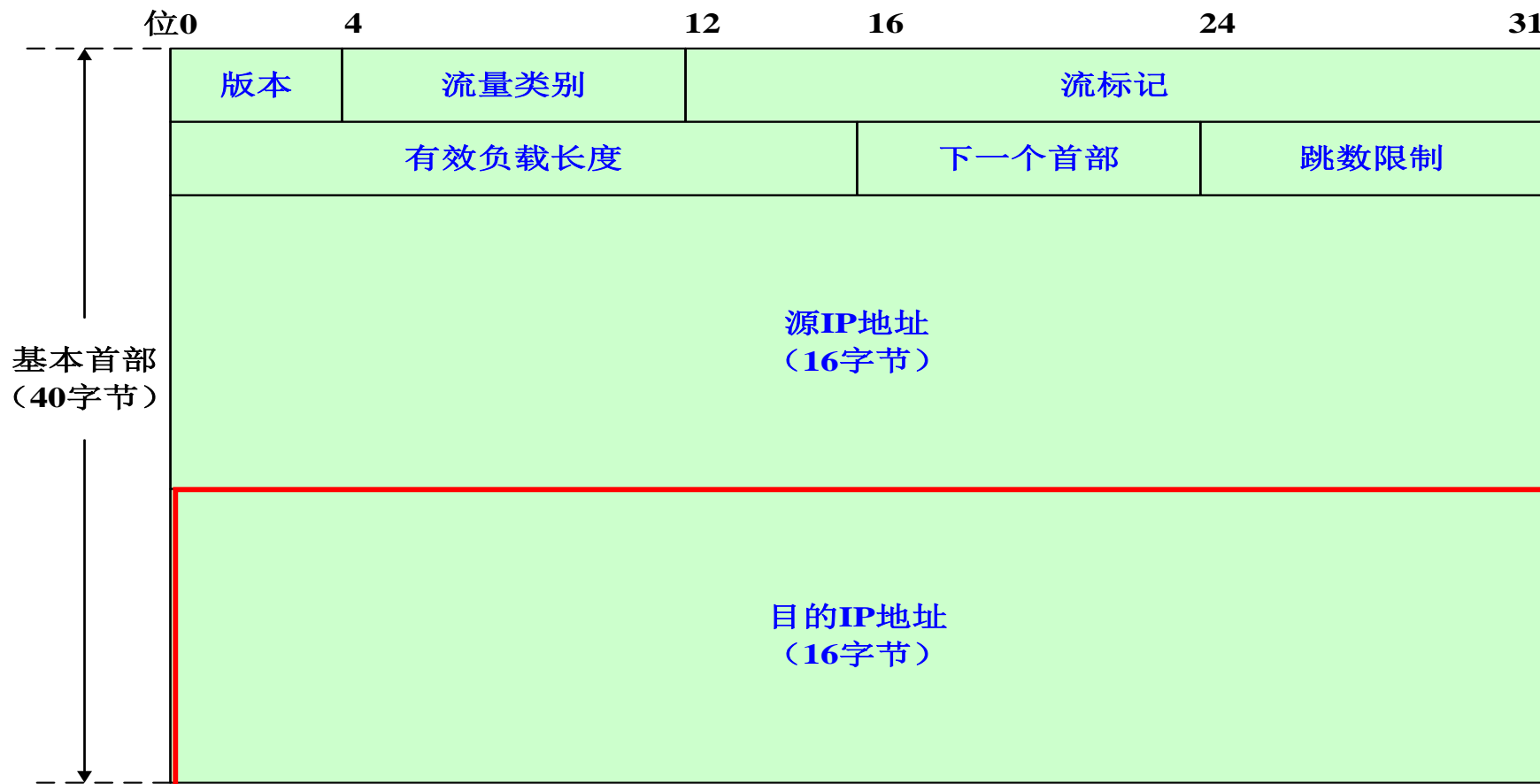
跳数限制(Hop Limit): 8比特的无符号整数。作用相当于IPv4中的TTL，该值减到0时，分组被丢弃

5.7.2 IPv6的报文格式



源IP地址字段：128比特。表示IPv6分组的源结点(如发送主机)

5.7.2 IPv6的报文格式



目的IP地址字段：128比特。表示IPv6分组的目的结点(如接收主机)

基本首部各字段含义

- 版本号：4比特。IPv6该字段值为6。
- 流量类型（traffic class）：8比特。区分IPv6分组的类别。
- 流标记（Flow Labels）：20比特。源结点标记IP分组所属的流，让中间路由器提供特别处理，保证数据传输的QoS。
- 有效负载长度（Payload Length）：16比特。有效负载的字节数（包括数据和扩展首部的长度）
- 下一个首部（Next Header）：8比特。指示紧随其后的扩展首部的类型，当后面没有扩展首部时，该字段的取值和含义与IPv4首部的“协议”字段相同。
- 跳数限制（Hop Limit）：8比特。相当于IPv4中的TTL。
- 源和目的IP地址：各128比特。标识源结点和目的结点网络接口。

IPv6的扩展首部

扩展首部位于基本首部和高层协议首部之间，通过基本首部中的“下一个首部”字段的值来判断扩展首部类型。

扩展首部类型	下一首部 字段值	长度	说明
逐跳选项	0	可变长度	携带沿途路由器结点都要处理的信息
路由	43	可变长度	指出沿途要必须经过的路由器结点
分片	44	64bits	包含分片信息，分片由源主机进行
封装安全有效载荷	50	可变长度	提供对数据的加密(参见IPsec).
鉴别首部(AH)	51	可变长度	提供对IPv6分组的认证功能
目的选项	60	可变长度	携带只是目的结点处理的信息

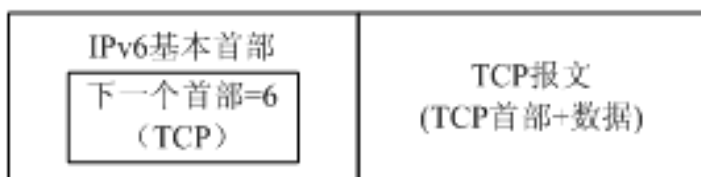
IPv6的扩展首部

- 扩展首部的第一个字段为“**Next Header**”字段（8比特），后面扩展首部的类型需要前面一个首部的“**Next Header**”字段的值来指明。
- 紧接数据的首部必须指出后面携带的高层数据的协议类型
- 多个扩展首部时，按以下顺序封装：
①逐跳首部 ②路由首部 ③ 分片首部 ④认证首部



多个扩展首部的连接

- IPv6数据包携带0个、1个和多个扩展首部的封装形式：



5.7.3 IPv6地址

- IPv6地址长度：16字节（128比特），为IPv4的4倍，支持多达 2^{128} 个（约为 3.4×10^{38} 个）IP地址。
- IPv6地址采用更合理的层次结构
- 为了增加IPv6地址的可读性，RFC4291 中规定了IPv6地址的数字字符串形式的常规表示法



IPv6地址的表示法

- **冒号十六进制表示法：** 每一个16进制的数字表示4比特，如：
FEDC : BA98 : 7654 : 0000 : 0000 : BA98 : 0001 : 3210
- **零压缩法：** 连续的全零段可以用“::”代替（每个地址只能用一次双冒号）。上述地址的0压缩形式为：
FEDC : BA98 : 7654 :: BA98 : 1 : 3210
- **IPv4和IPv6混合表示法：** 用 x表示16进制，d表示10进制，方便IPv4 和IPv6混合的网络环境，如：
x : x : x : x : x : x : d . d . d . d
- **CIDR表示法：** IPv6地址/前缀长度，如：
2001:0DB8:0000:CD30:0000:0000:0000:0000/60， 或
2001:0DB8:0:CD30::/60

IPv6地址的三种基本类型


- **单播（unicast）**：单播分组发给该地址指定的一个网络接口，即点到点通信。两种单播地址：
 - 本地链路单播地址：在本地网络中使用
 - 全球的单播地址：在因特网中使用（公开地址）
- **多播（multicast）**：多播分组发给多个网络接口（通常不在一个结点上），即点到多点通信。
IPv6中把广播视为多播的特例。
- **任意播(anycast)**：新类型，指向一组网络接口，但分组只需要发送给其中一个即可，通常是路由距离最近的一个。

IPv6地址空间的规划

- RFC 4291中对IPv6地址空间的大致规划如下表

二进制前缀	地址类型	IPv6地址表示
00...0 (128 bits)	未指定地址	::/128
00...1 (128 bits)	环回地址	::1/128
11111111 (8比特)	多播地址	FF00::/8
1111111010(10 比特)	本地链路单播地址	FE80::/10
剩余的其它地址	全球单播地址	

IPv6单播地址的结构

- IPv6地址结构更合理和更系统。
 - 当一个结点拥有多个网络接口时，需要多个IPv6地址。
 - 单播地址的最后一级是指向网络接口，称作“接口标识（**Interface ID**）”，也可以通俗地称为接口地址。
 - 本地链路单播地址和全球单播地址拥有不同的结构。
- 

本地链路单播地址结构



- 本地链路单播地址（**Link-Local unicast**）只用于本地网络，路由器不会转发以本地链路单播地址为源地址或目的地址的**IP**分组。
- 可以利用硬件地址生成本地链路单播地址。
- 在邻居发现、**IPv6**地址自动配置的过程中会用到本地链路单播地址。

全球单播地址结构

全球路由选择前缀 (48比特)

子网标识 (16比特)

接口标识 (64比特)

一般采用三级结构：

1. 全球路由选择前缀：通常分配给一个组织机构的网络（site）
相当于IPv4的网络号，用于因特网中路由器选路。
2. 子网标识：用于划分内部子网。不划分子网时该字段置0。
3. 接口标识：标识结点的网络接口，相当于IPv4中的主机号。

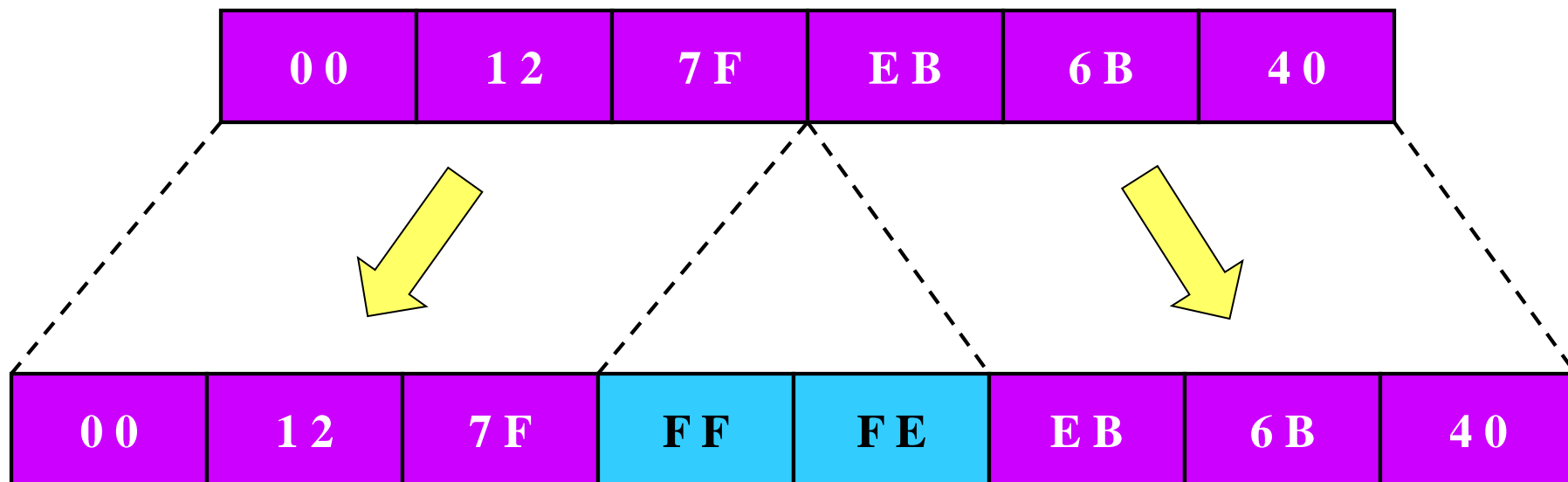
接口标识的生成

可以从MAC地址（EUI-48）生成IPv6接口地址，需进行两步转换：

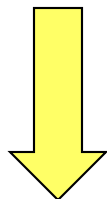
- ①地址扩展，把48比特的地址扩展为64比特的EUI-64地址；
- ②修改EUI-64的U/L位，值取1（Universal，可全球寻址，若设0则Local）。

(见下图示意)





00000000




U/L位

00000010




5.7.4 ICMPv6

- ICMPv6继承了差错报告和信息查询的功能。
 - ICMPv6使用IPv6分组封装，下一个首部字段得值为58，代表ICMPv6协议。
 - ICMPv6报文支持多播听众发现（Multicast Listener Discovery, MLD)协议，因此，IPv6不再采用IGMP 协议进行多播组管理。
- 

ICMPv6报文类型

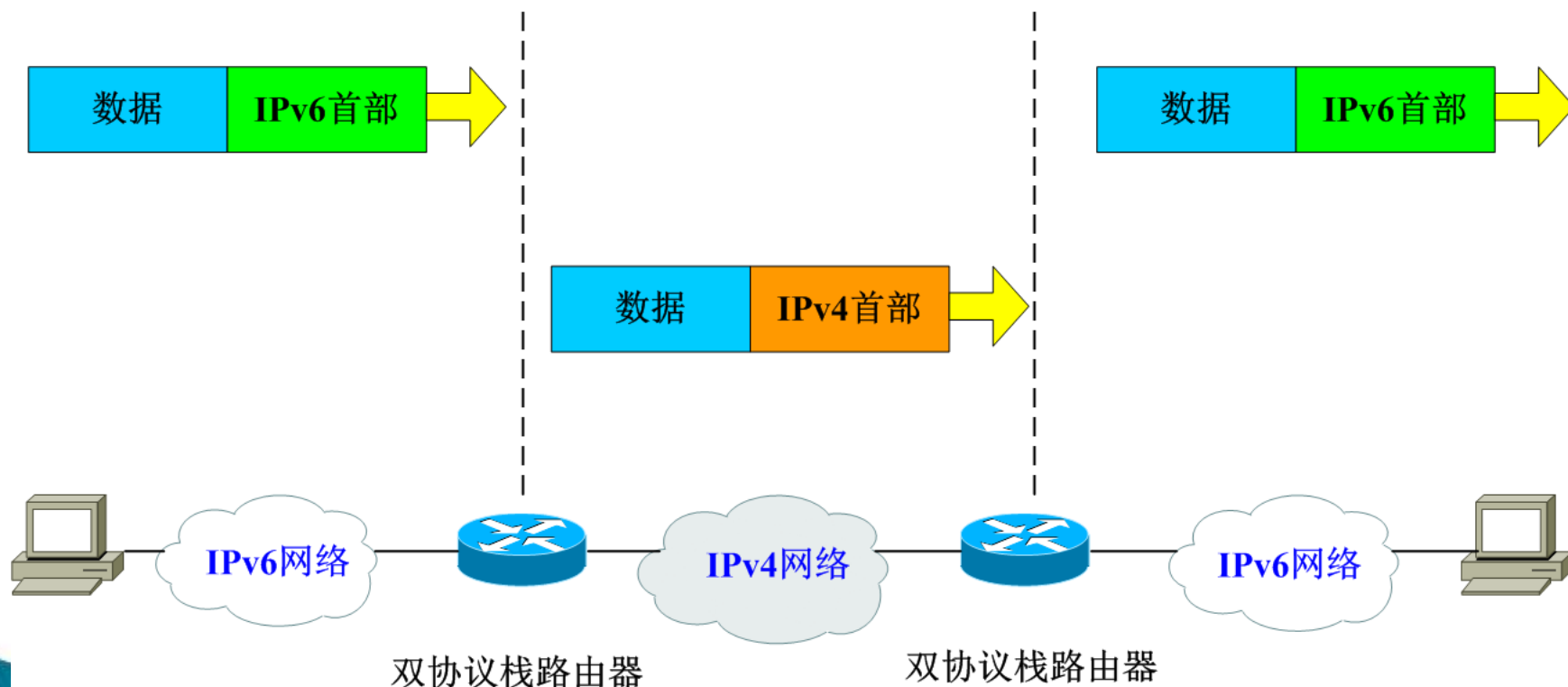
类型值	含义	用途
1	目的地不可达	差错报告
2	分组太长	
3	超时	
4	参数问题	
127	保留用于以后扩展的ICMPv6差错消息	
128	Echo 请求	探测和应答报文
129	Echo应答	
130	多播听众探测	多播
143	多播听众报告	
255	保留用于以后扩展的ICMPv6信息报文	信息报文

5.7.5 向IPv6的过渡

- 主流操作系统都已实现了IPv6协议。
 - IPv6试验床CERNET2
 - 中国下一代互联网(China's Next Generation Internet, CNGI)已建成，包括6个核心网络，22个城市的59个结点，2个交换中心，273个驻地网的IPv6示范网络。
 - IPv4和IPv6共存的局面，在两种协议的交界处，存在着协议共存和协议转换的问题。
 - 目前采用两种解决方法：双协议栈和隧道技术
- 

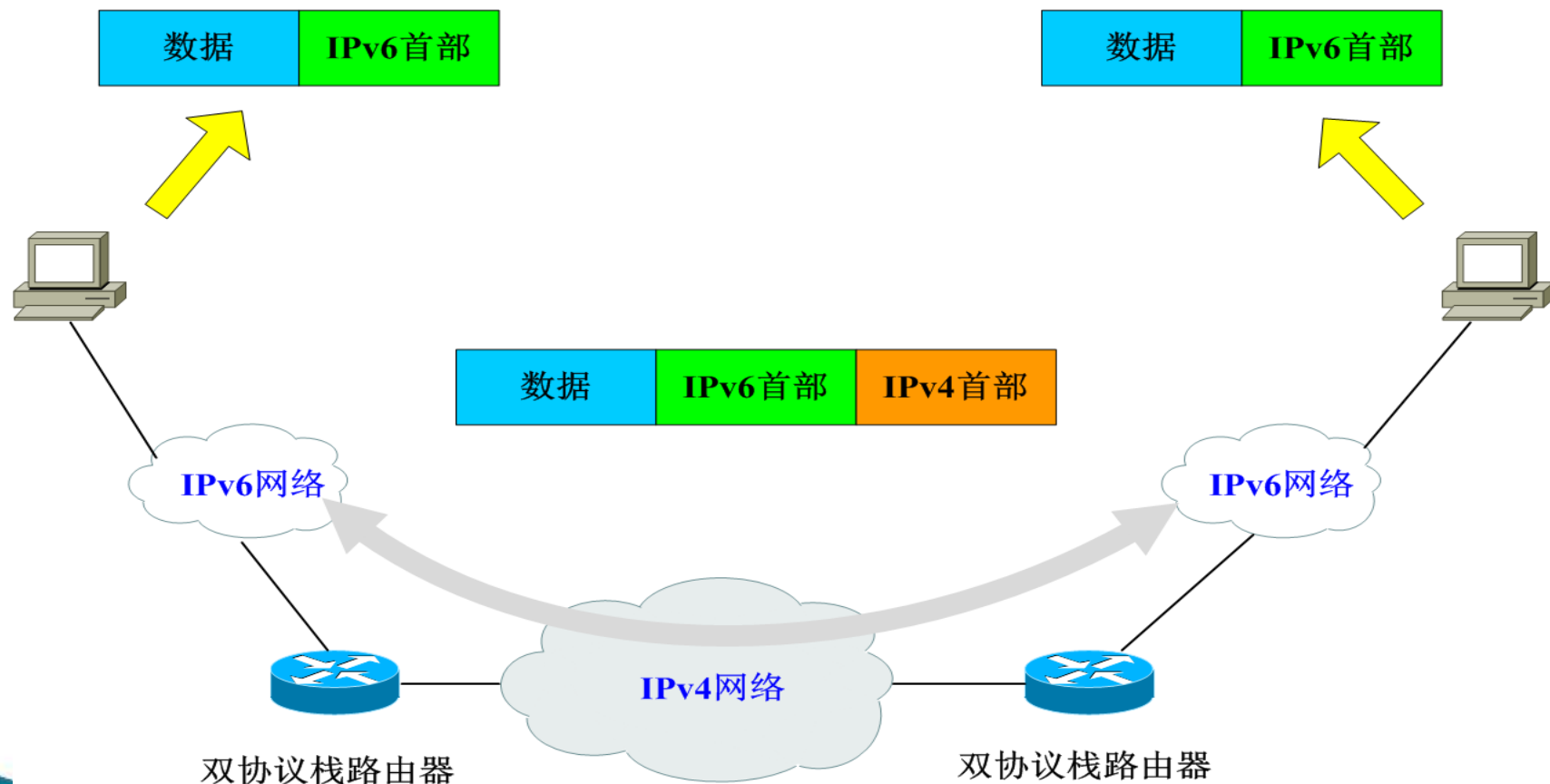
双协议栈

- 两个协议网络的边界结点中实现双协议栈（Dual stack），进行协议转换。




隧道技术


- IPv4首部的协议字段为41时，指明内部封装的是IPv6分组。



5.8 移动IP

- IPv4正确寻址的前提：主机的IP地址正确地指明了它所接入的网络。
 - 移动IP面临的寻址问题：
当移动结点跨IP网络漫游时，原来的网络地址就不匹配了。
 - 移动IP需要解决的核心问题是IP地址的适配问题。
- 

移动IP方案

- **目标：**在不改变移动终端的原IP地址配置的情况下，实现跨网络的移动接入。
 - **思路：**采用中间代理方法，代理通常由路由器担任。
 - **方案：**双代理、双地址的方法。为了，设外地代理是为了支持其它网络的移动结点接入到本网络。
 - **驻地代理（Home Agent）：**支持本网所属移动结点出去
 - **外地代理（Foreign Agents）：**支持外来结点接入到本网络
 - **驻地地址（Home Address）：**归属地址
 - **转交地址（Care-of Address）：**“临时通信地址”
- 

移动IP的工作原理（一）

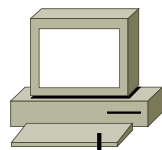
1. 代理发现：通过接收代理通告/响应
2. 代理判定：判断是驻地/外地代理，以确定是否在外地网络，是否需要移动服务
3. 获取转交地址和注册转交地址：若确定是
在外地网络，则：
 - ① 获取一个转交地址
 - ② 向驻地代理注册所获取的转发地址

移动IP的工作原理（二）

- 4. 经双代理转发的数据报接收过程
- 5. 不经代理的数据报发送：移动结点发送报文的传输可以不需要驻地代理转交。
- 6. 注销转交地址：返回驻地网络后，向驻地代理注销其转交地址



发送主机

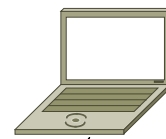


经双代理转发的数据报接收过程

目的网络



互联网络



移动结点

驻地网络

驻地代理

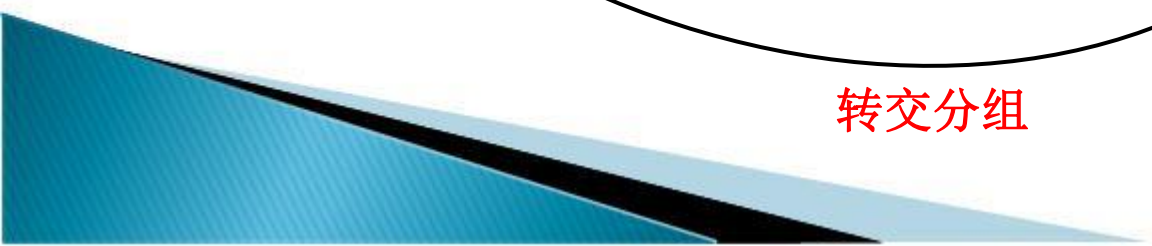
外地代理

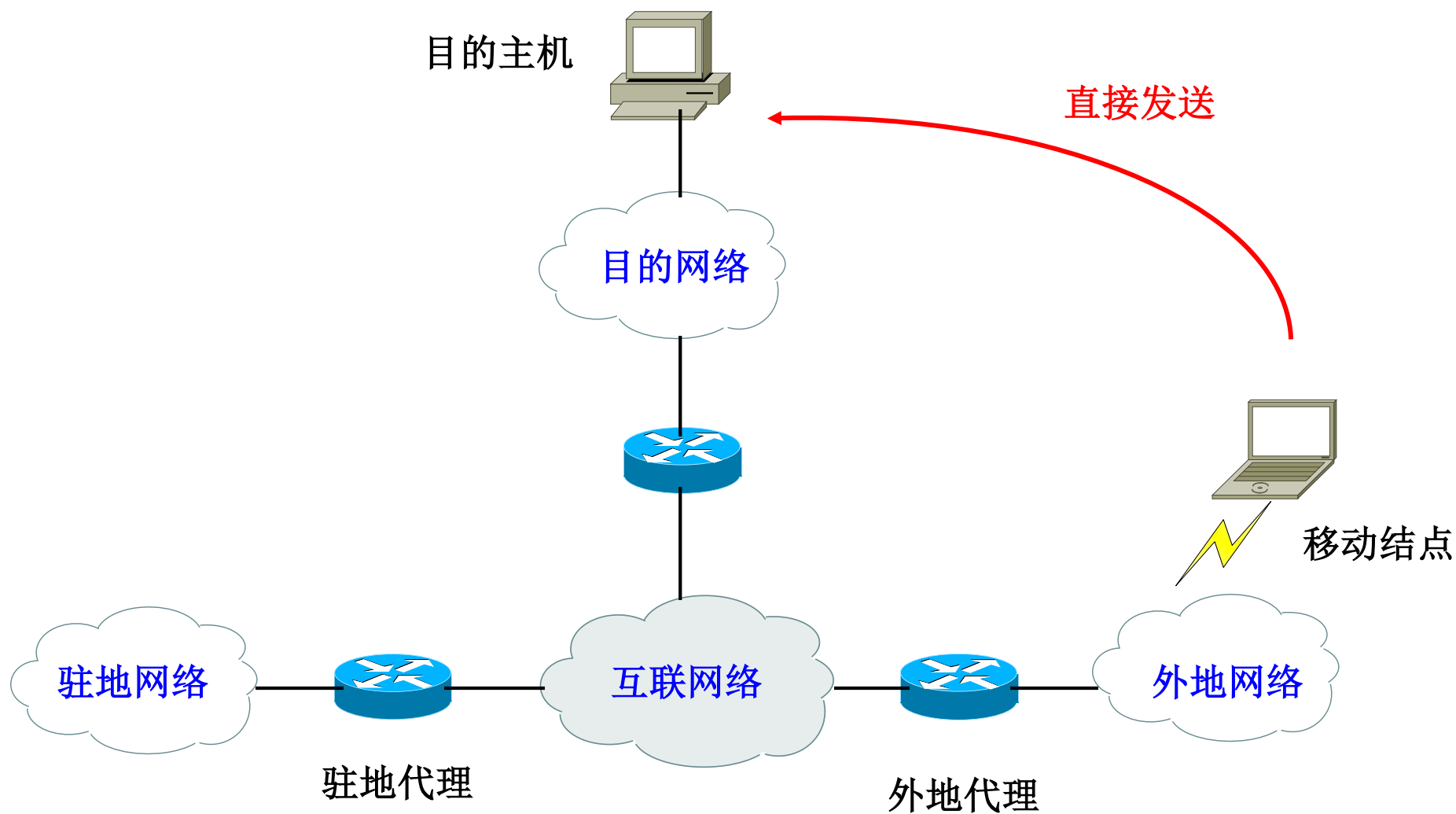
外地网络

截获分组

转交分组

转交分组






不经过代理转发的数据报发送过程

5.9 网络层的QoS

- QoS的一般概念
- 集成服务
- 区分服务



背景

- 因特网的**IP**协议：数据报服务是“尽力而为”服务。
 - 多媒体应用和其他实时应用的**QoS**需求。
 - 各种不同的针对**IP**网络的网络服务质量（Qualit of Service, **QoS**）体系结构。
 - **QoS**技术并不是**IP**网络才有的。**ISO**最早提出，**OSI**参考模型要求每层协议必须提供明确的服务质量指标。
 - **ATM**论坛较早提出了精确定义的**QoS**概念（恒定位速率服务**CBR**、实时可变位速率服务**rt-VBR**、非实时可变位速率服务**nrt-VBR**等等。
- 

5.9.1 QoS的一般概念


- QoS是网络在传输数据时应满足的一系列服务需求，这些需求可用网络性能指标来定量地衡量。
- RFC2216将QoS定义为：用带宽、分组延迟和分组丢失率等参数描述的关于分组传输的质量
- 常用以下的参数来衡量网络的QoS：
 - 带宽(bandwidth)，指网络提供的分组传输容量
 - 延迟(latency)，分组穿越网络所需要的时间
 - 抖动(jitter)，不同分组穿越网络的延迟的变化
 - 丢包率(loss rate)，分组传输过程中被结点丢弃的几率

两种有代表性的网络QoS体系


- 集成服务（InterServ, Integrated Service）模型
 - 为网络中的流量提供有保证的网络资源
 - 利用资源预留的方法
- 区分服务（DiffServ, Differentiated Services, DS）模型
 - 对数据分组进行分类，并区别对待
 - 为不同的网络应用提供可接受的服务质量



5.9.2 集成服务

- 会话的发起端先声明其QoS需求，路径上的各结点确定自己是否有足够的资源满足其需求。
 - 集成服务体系包含以下组成部分：
 - ①资源预留协议
(resource reservation protocol, RSVP)
 - ②流规范 (Flow Specification)
 - ③流量控制 (Flow Control)
 - ④服务类别
- 

RSVP

- 第一个支持QoS的Internet控制协议（信令协议）。
 - RSVP提供基于流（flow）的资源预留机制，InterServ的QoS也是基于流的。
 - 流的概念：具有相同目的地址、目的端口号和协议号，并具有相同的QoS请求的一系列数据报。
 - 将描述业务QoS需求的流规范传递给路径上的各结点。
 - 将网络资源预留给路由上的某个特定的流或者会话。
- 


RSVP协议的两种主要报文

路径（PATH）报文：从源结点向下游传递


预留（RESV）报文：从目的结点向上游传递




RSVP实现资源预留的过程

- (1) 源结点（发送端）的应用程序通过API向本地RSVP协议实体注册，将流规范包含在PATH分组中向接收端方向的下游传递。
 - (2) 路径上的路由器收到后，存储PATH报文中的路径状态以及QoS特性等信息，继续向下游转发
 - (3) 目的结点（接收端）最终收到的PATH报文包含了完整的路径状态信息，接收端的应用程序通过API将预留请求递交给本地RSVP协议实体，该实体生成一个RESV报文，并且沿PATH报文的路径，向上游传递（返回给发送端）。
- 

RSVP实现资源预留的过程

- (4) 沿途结点收到**RESV**报文后，如果有足够的资源满足流的**QoS**需求，则接受预留请求，预留带宽和缓冲区空间，然后向上游转发**RESV**报文；如果不能满足资源要求，则拒绝**RESV**报文请求，向接收端返回一个错误信息
 - (5) 当**RESV**协议报文到达发送端后，一条从发送端到接收端的、具有预留资源的固定通路就建立起来了。
 - (6) 开始传输数据流，当会话的数据流传输完毕，各结点释放预留的资源。
- 

RSVP特点和局限

- 能为单播的流与组播的流提供资源预留的支持。
 - 提供端到端的QoS，需要主机应用程序的支持。
 - 资源的预留是单向的（从发送端到接收端），预留请求由接收端发起。
 - 预留信息是软状态的，需要定时刷新，带来额外的信令开销。
 - 资源预留是基于会话流的，需要维护的状态信息的数量与流的数量成正比，这在大型主干网中将是可观的数量级，因此，其扩展性不够理想。
- 

5.9.3 区分服务

- 区分服务的基本工作原理：
 - (1) 在网络的边缘对所有数据报进行分类，并且标记每个分组所属的服务类型
 - (2) 网络中间结点（路由器）根据分组中的标记进行相应优先级别的处理
- 需要解决两个问题：
 - 如何分类和标记分组
 - 如何处理不同类别的分组



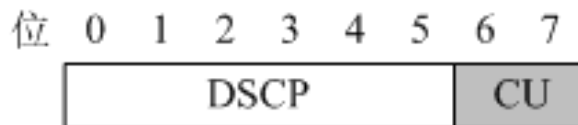
DiffServ技术的两个重要概念

- 如何分类和标记分组——区分服务码点（DSCP, Differentiated Service CodePoint）
- 如何处理不同类别的分组 ——每跳行为（PHB, Per Hop behavior）
- 在区分服务码点和每跳行为之间建立对应关系



DSCP

- 区分服务码点（DSCP），标记不同分类的分组。
- IPv4用TOS/DS字段标记，IPv6用流量类别字段。
- DSCP字段为1字节，使用其中的6比特来标记不同QoS需求的分组类别。



DSCP: 区分服务码点

CU: 当前未使用

DSCP字段

每跳行为（PHB）

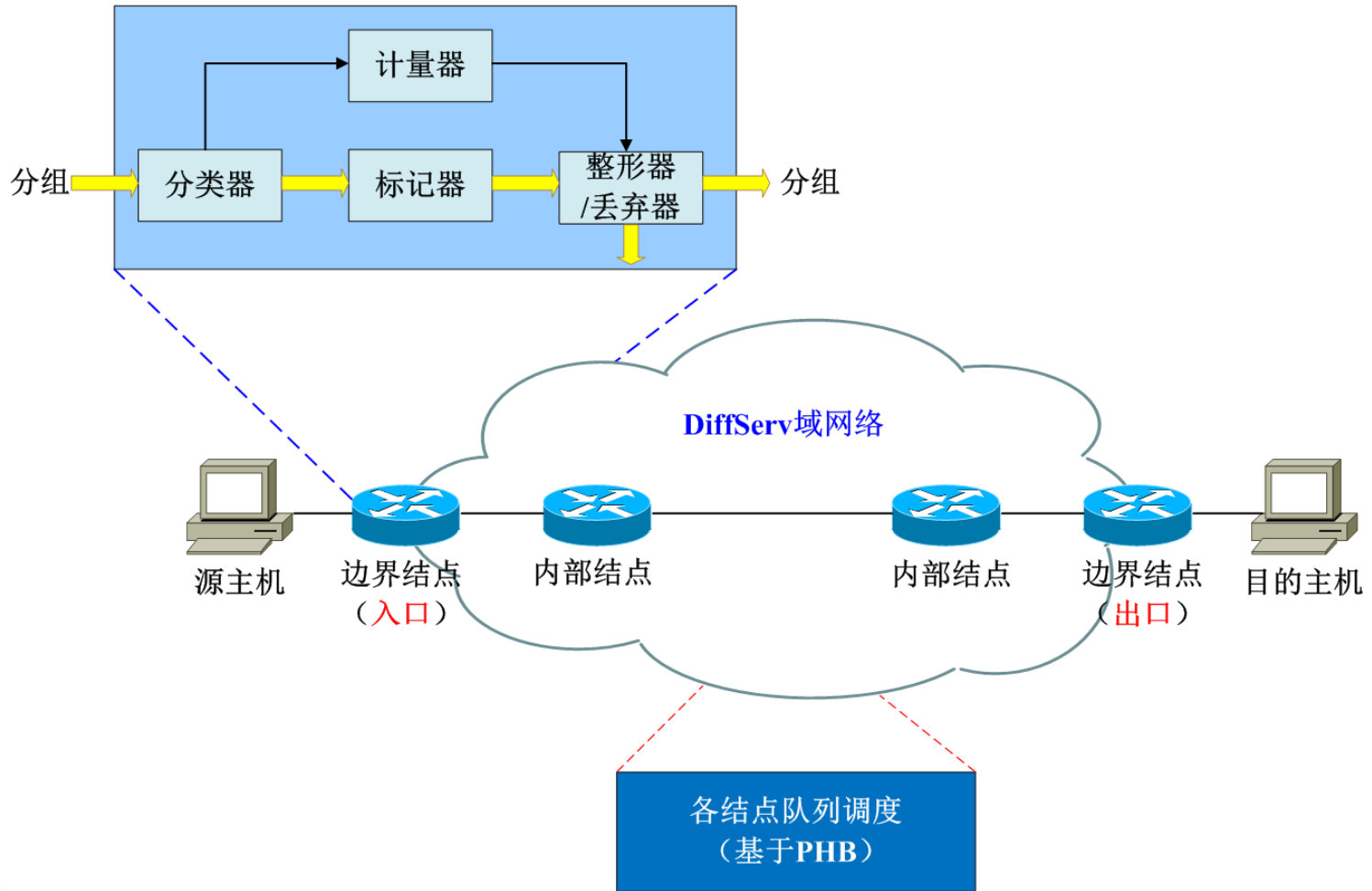
- 体现区分服务思想的关键概念
- 路由器对某类分组所采取的转发行为，指每个结点如何处理不同类型别的数据。
- 中间路由器根据一个分组**DSCP**字段的值来区别对待不同类的分组，对不同类的分组会采取不同资源调度优先级，表现为不同的转发行为。
- 每一跳结点通过**PHB**提供一致的**QoS**策略。



区分服务的三类PHB

- 快速转发EF（Expedited Forwarding）PHB，标记有EF的分组以最小的时延被转发，其丢包率应为最低。
- 有保证的转发AF（Assured Forwarding）PHB，AF的QoS性能参数低于EF类型。又分为4个子类：AF1、AF2、AF3和AF4，相同AF子类的分组进入同一个队列，每个子类又细分为3个丢弃级别。
- 默认的PHB，尽力而为的服务（Best Effort，BE）兼容。没有特别QoS要求的类、或者超出流量限制的AF类降级为该类。

DiffServ的工作原理示意图

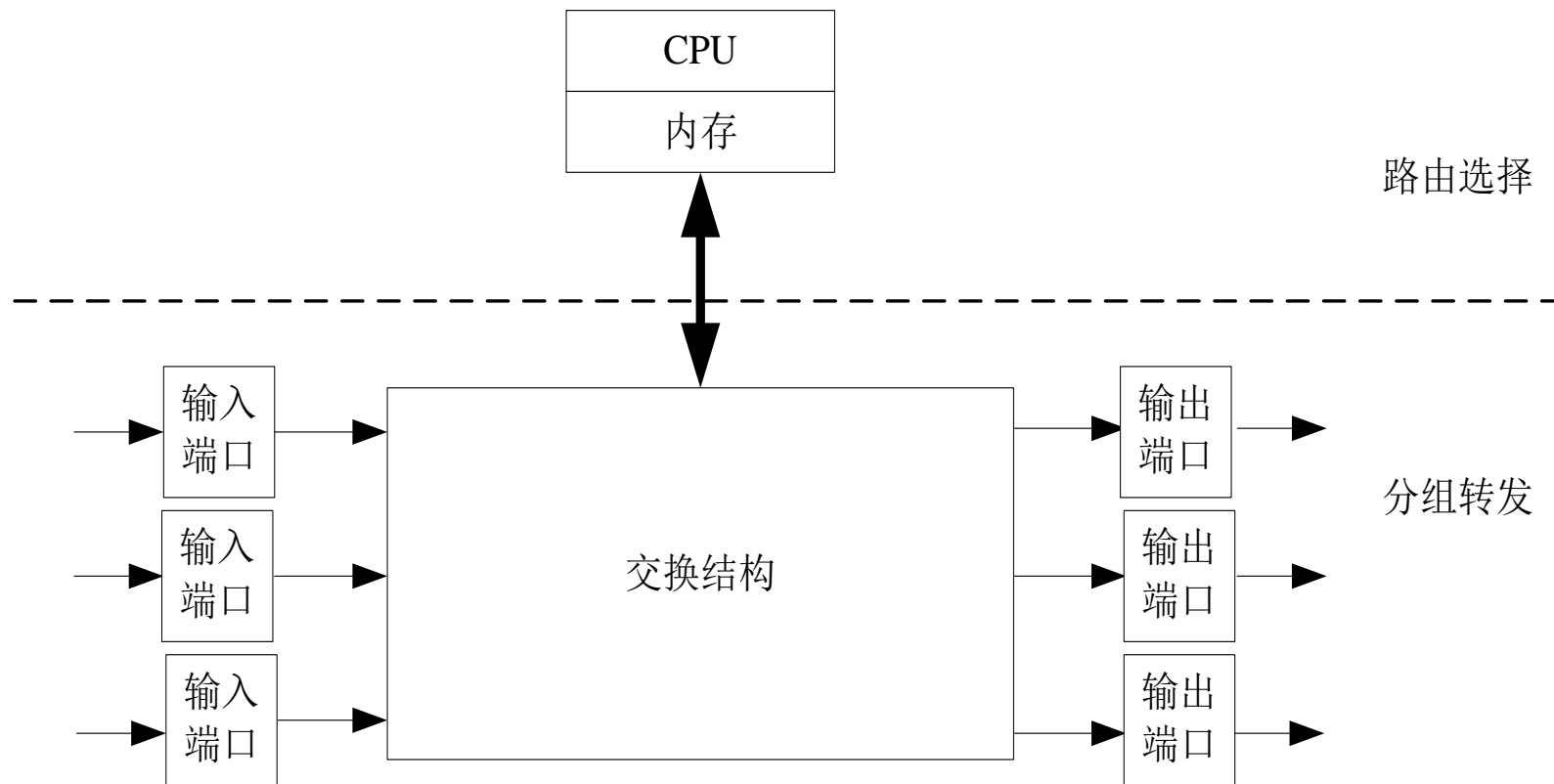


DiffServ模型的优势

- (1) 不需要维护每个应用流的信令与状态；通过将多个相同**QoS**需求的应用分组流汇聚成有限的服务等级，可以比基于每业务流的技术提供更好的扩展能力。
- (2) 由网络边界设备负责对进入网络的分组进行分类、标记、以及调整，主机及其应用程序可以不需要修改就能够得到具有**QoS**保障的网络服务
- (3) 目前大部分核心路由器都支持一些分类的队列管理机制，所以**DiffServ**技术可以得到较多的支持。



路由器的体系结构



路由器的体系结构

- **CPU:** 根据不同的路由协议和路由算法完成路由的计算，维护路由表，完成路由器的配置和管理等功能；
- **输入/输出端口:** 实现数据包的收发；
- **交换结构:** 实现分组的转发，是决定路由器性能的核心部件之一，不同的交换结构对于数据包的转发效率影响很大：
 - 采用共享内存：输入/输出端口共享存储器件
 - 交叉矩阵结构：实现无阻塞交换

课后思考题

1. 能否用一个8端口的交换机将8个C类的IP网络互连起来？为什么？
2. 路由器设备是否只要包含网络层功能就可以了？
3. 复习一下本章讲过的协议，在采用TCP/IP协议体系的网络中，哪些网络层协议会带来广播流量？其广播范围都是怎样的？