中南大學
CENTRAL SOUTH UNIVERSITY

第五章 网络层

本章目标

1. 学习网络层服务：无连接和面向连接服务，数据报和虚电路网络
2. 学习Internet网络层协议：IPv4/IPv6, ICMP, DHCP, NAT, ARP
3. 掌握链路状态、距离矢量等路由算法，了解层次路由结构
4. 掌握Internet路由协议：OSPF、RIP、BGP，了解MPLS
5. 了解路由器工作原理：控制层和数据层，报文转发机制，交换结构
6. 了解网络拥塞及拥塞控制思想
7. 了解网络服务质量及设计思想
8. 了解其它重要的网络层相关机制：三层交换、VPN
9. 了解IPv6相关机制

本章内容

5.1 网络层服务

1. 网络层服务概述
2. 无连接服务的实现
3. 面向连接服务的实现
4. 虚电路与数据报网络的比较

5.2 Internet网际协议

5.3 路由算法

5.4 Internet路由协议

5.5 路由器工作原理

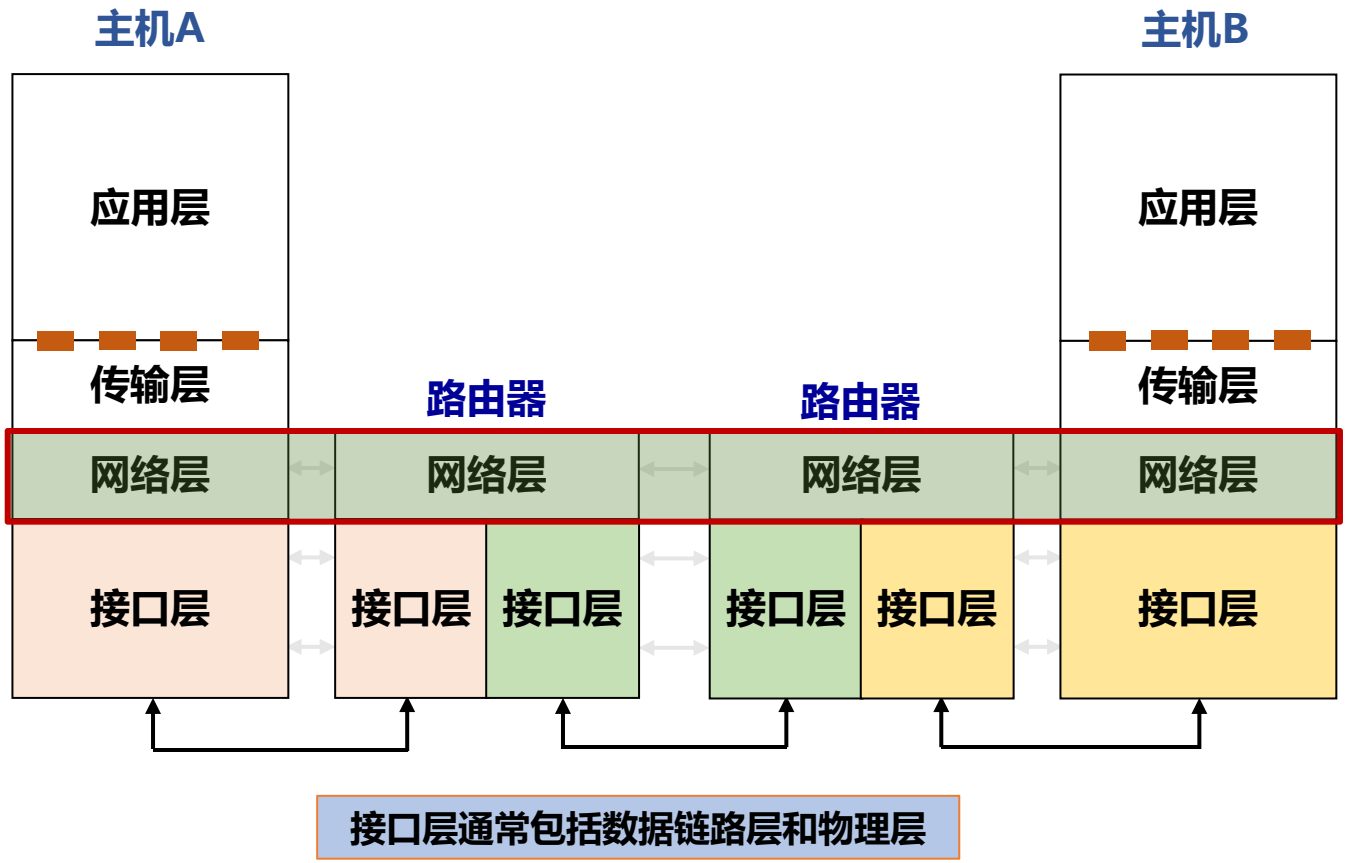
5.6 拥塞控制算法

5.7 服务质量

5.8 三层交换与VPN

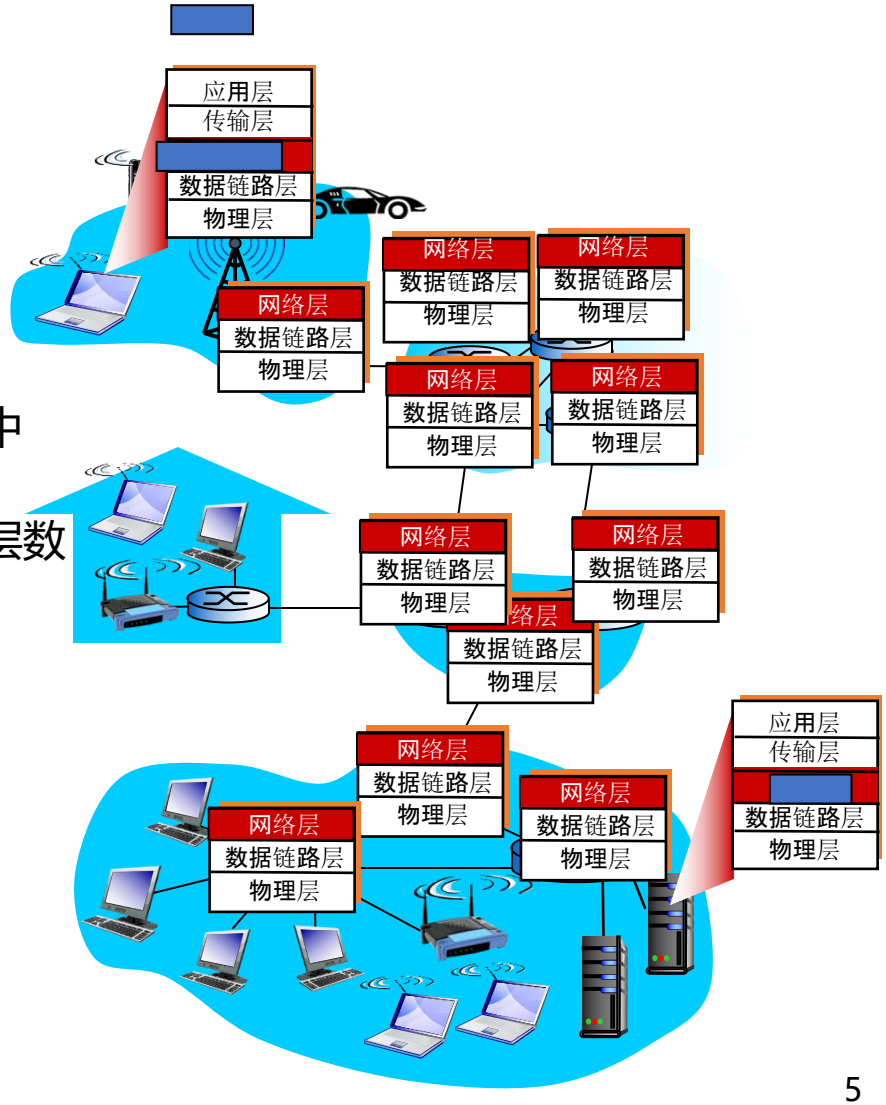
5.9 IPv6技术

网络层服务概述


主机A
主机B
应用层
应用层
传输层
传输层
路由器
路由器
网络层
网络层
网络层
网络层
接口层
接口层
接口层
接口层
接口层
接口层通常包括数据链路层和物理层

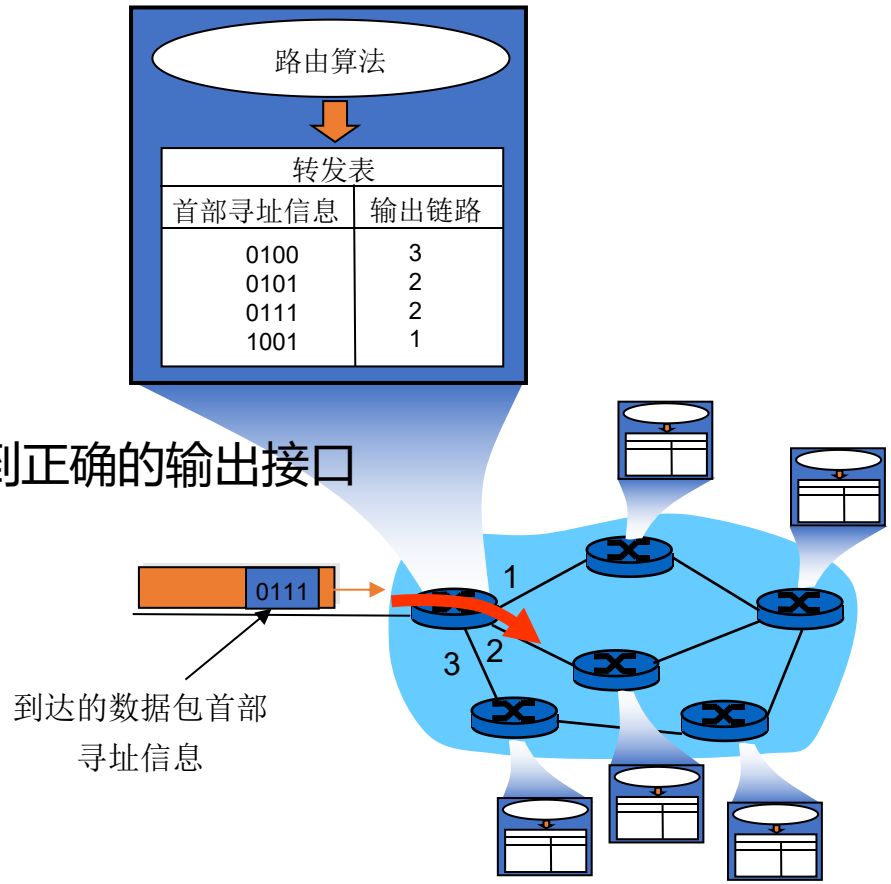
网络层服务的实现

- 网络层实现端系统间多跳传输可达
- 网络层功能存在每台主机和路由器中
  - 发送端：将传输层数据单元封装在数据包中
  - 接收端：解析接收的数据包中，取出传输层数据单元，交付给传输层
  - 路由器：检查数据包首部，转发数据包



网络层关键功能

- 路由（控制面）
  - 选择数据报从源端到目的端的路径
  - 核心：路由算法与协议
- 转发（数据面）
  - 将数据报从路由器的输入接口传送到正确的输出接口



提供给传输层的服务

- 网络通信的可靠交付服务，谁来负责？

  “网络”OR“端系统”

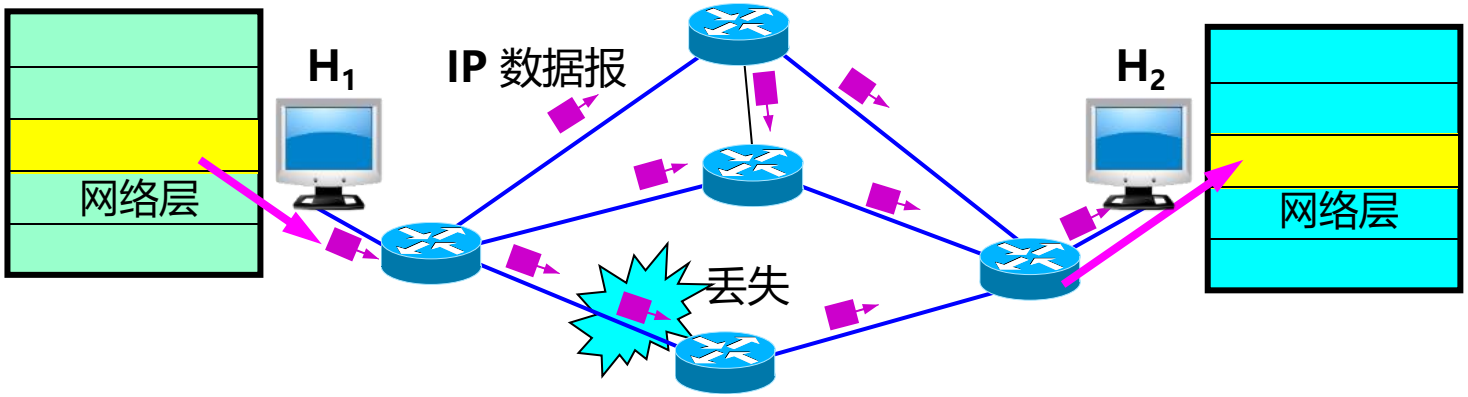
- 网络层应该向运输层提供怎样的服务？

  “面向连接”（虚电路）OR“无连接“（数据报）

无连接服务的实现

- **无连接服务：如寄信**
  - 不需要提前建立连接
- **数据报服务**
  - 网络层向上只提供简单灵活无连接的、尽最大努力交付的数据报服务
  - 发送分组时不需要先建立连接，每个分组独立发送
  - 数据报独立转发，相同源-目的的数据报可能经过不同的路径
  - 网络层不提供服务质量的承诺
- **尽力而为交付**
  - 传输网络不提供端到端的可靠传输服务：丢包、乱序、错误
  - 优点：网络的造价大大降低，运行方式灵活，能够适应多种应用

无连接服务的实现

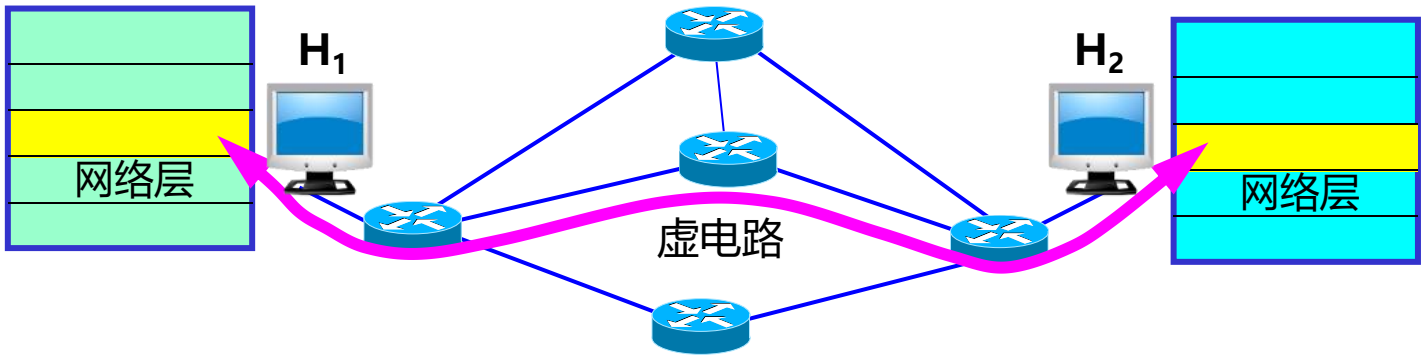


H1 发送给 H2 的分组可能沿着不同路径传送
在数据包分片的情况下，尽量还是沿相同路径

面向连接服务的实现

- **面向连接服务：如打电话**
  - 通信之间先建立逻辑连接：在此过程中，如有需要，可以预留网络资源
  - 结合使用可靠传输的网络协议，保证所发送的分组无差错按序到达终点
- **虚电路是逻辑连接**
  - 虚电路表示这只是一条逻辑上的连接，分组都沿着这条逻辑连接按照存储转发方式传送，而并不是真正建立了一条物理连接
  - 注意，电路交换的电话通信是先建立了一条真正的连接
  - 因此分组交换的虚连接和电路交换的连接只是类似，但并不完全相同

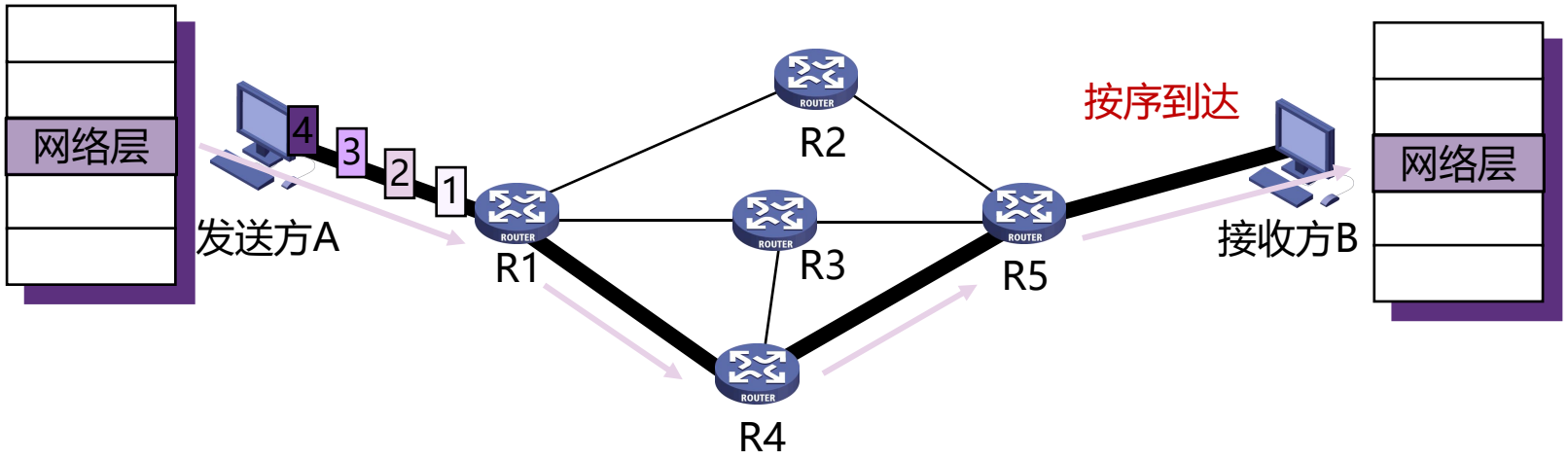
面向连接服务的实现



H_1 发送给 H_2 的所有分组都沿着同一条虚电路传送

面向连接的虚电路

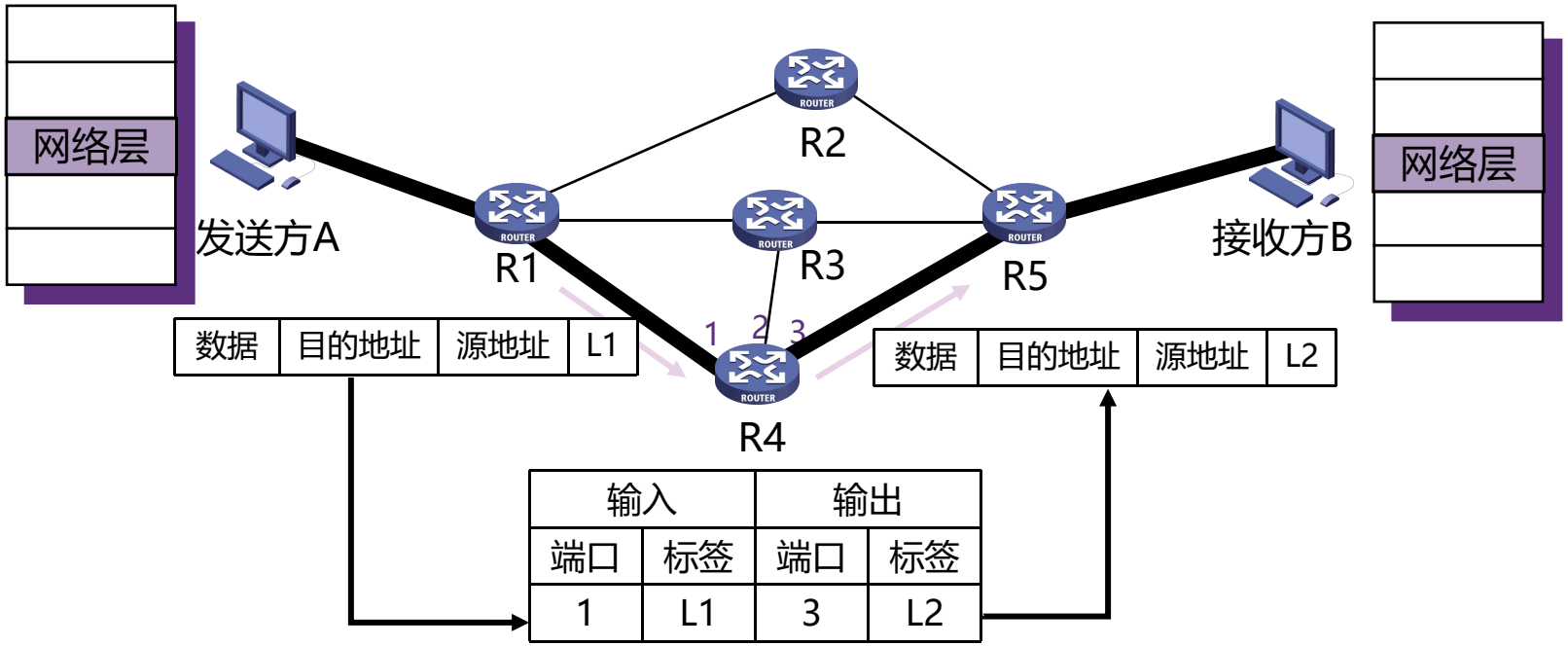
- **虚电路（virtual circuit）：面向连接的方法**



面向连接的方法也不一定能完全保证数据的可靠传输，链路中的任何一个组成环节仍有可能失效，而这种失效是严重的，可能导致所有数据丢失

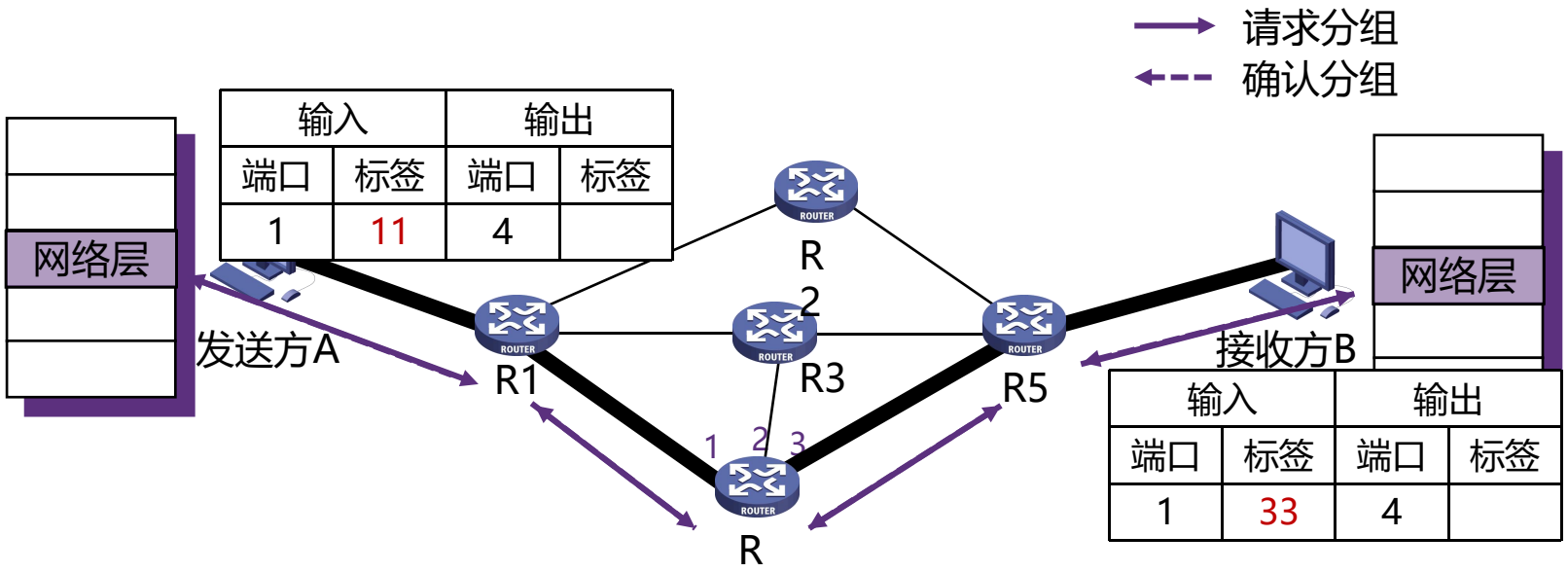
虚电路转发策略

- **虚电路的转发策略：虚电路转发决策基于分组标签，即虚电路号**



面向连接的虚电路

- **面向连接的服务第一阶段：建立连接**



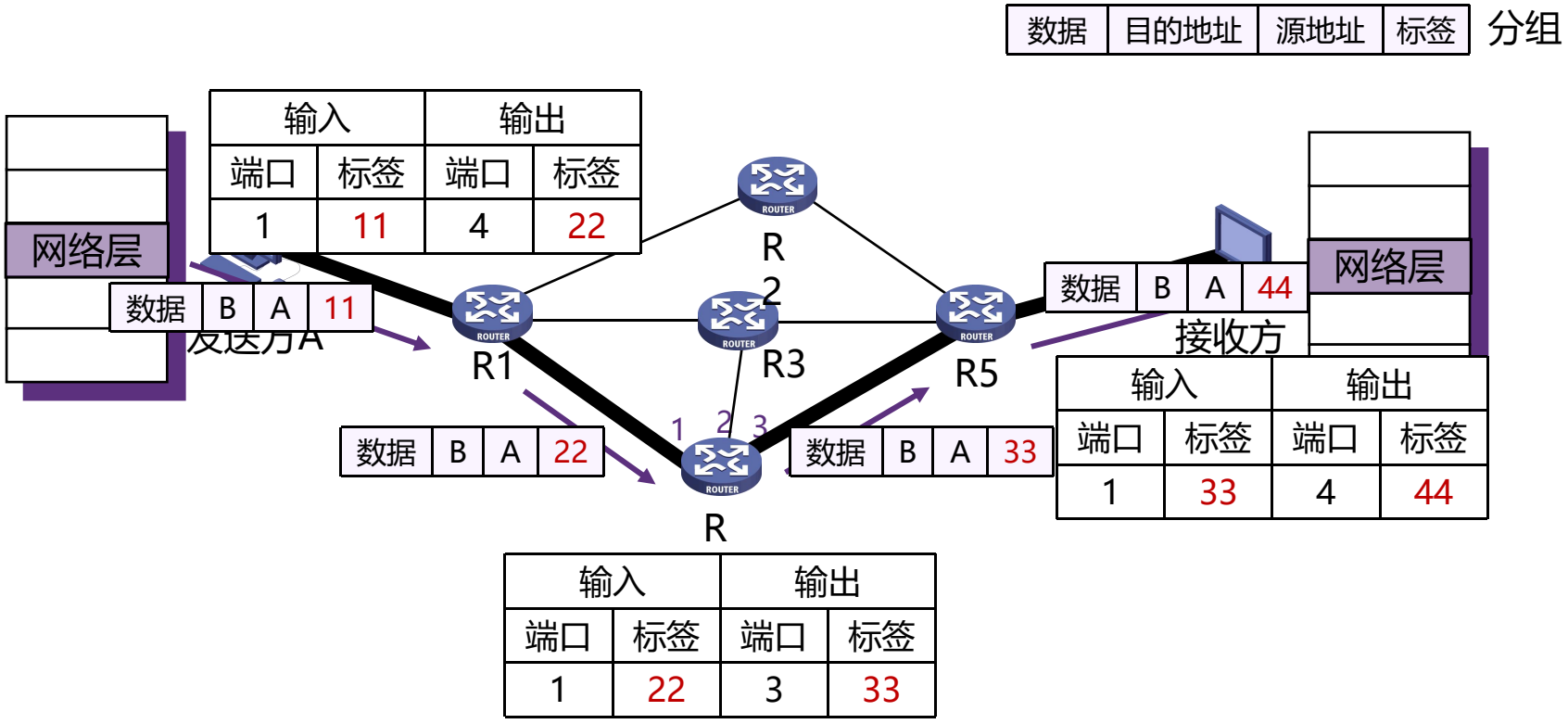
输入		输出	
端口	标签	端口	标签
1	11	4	

输入		输出	
端口	标签	端口	标签
1	33	4	

输入		输出	
端口	标签	端口	标签
1	22	3	

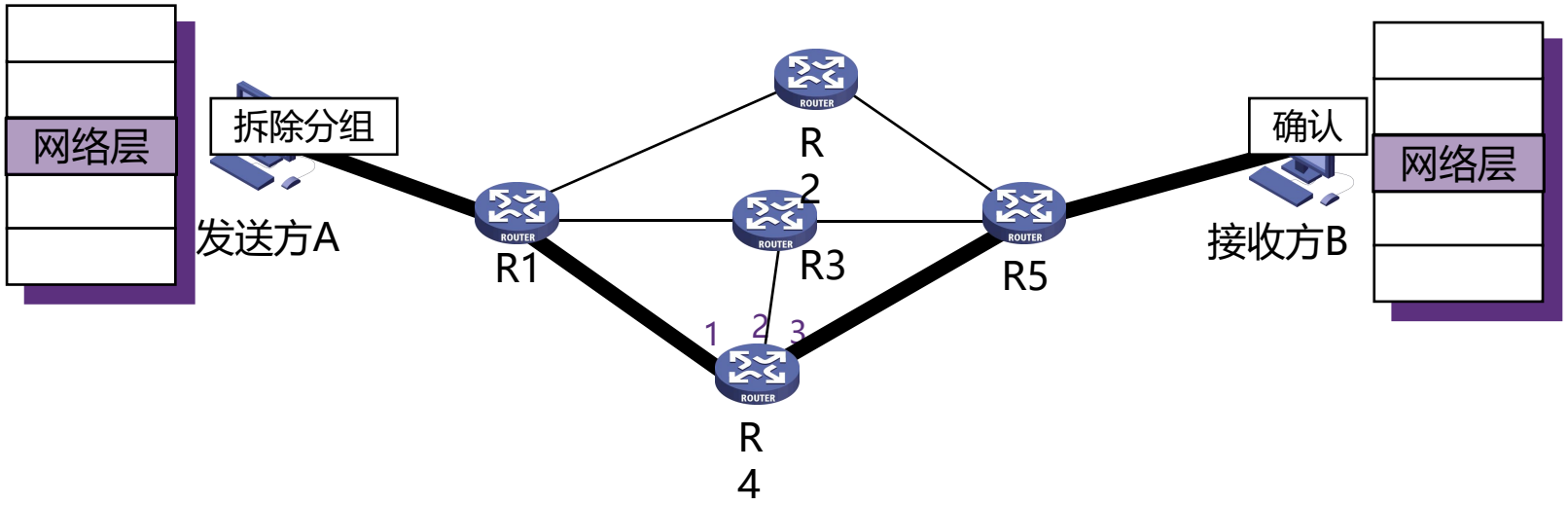
面向连接的虚电路

- **面向连接的服务第二阶段：发送数据**



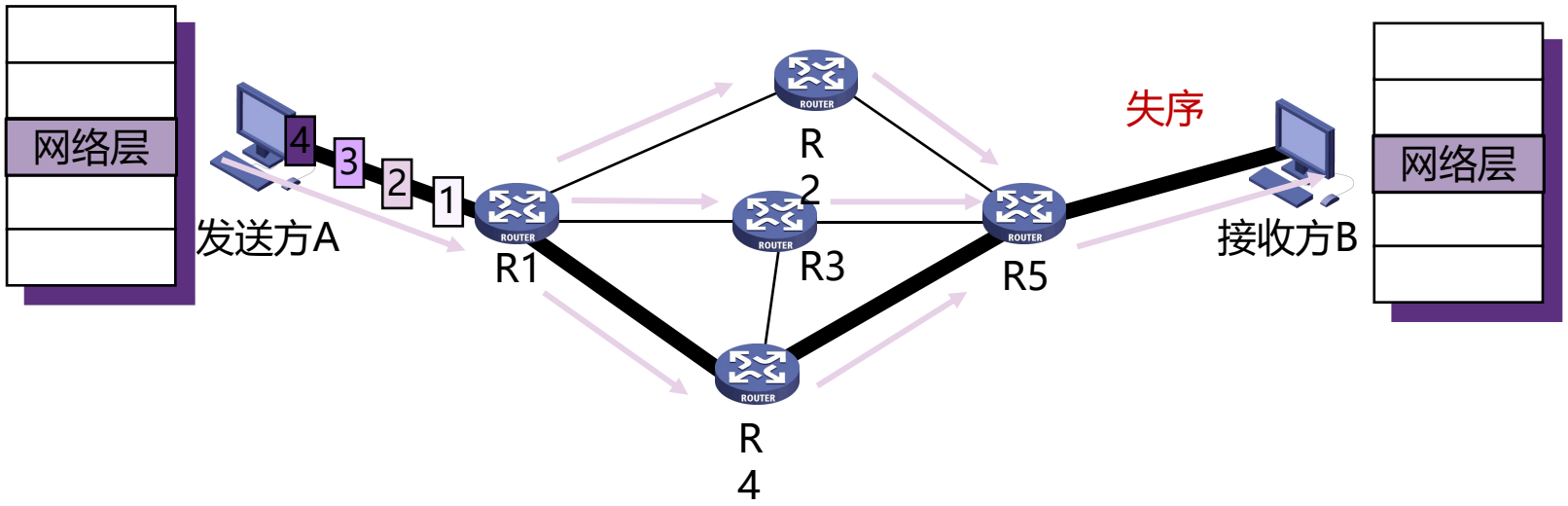
面向连接的虚电路

- **面向连接的服务第三阶段：释放连接**



无连接的数据报

➢ **数据报（datagram）：无连接的方法**

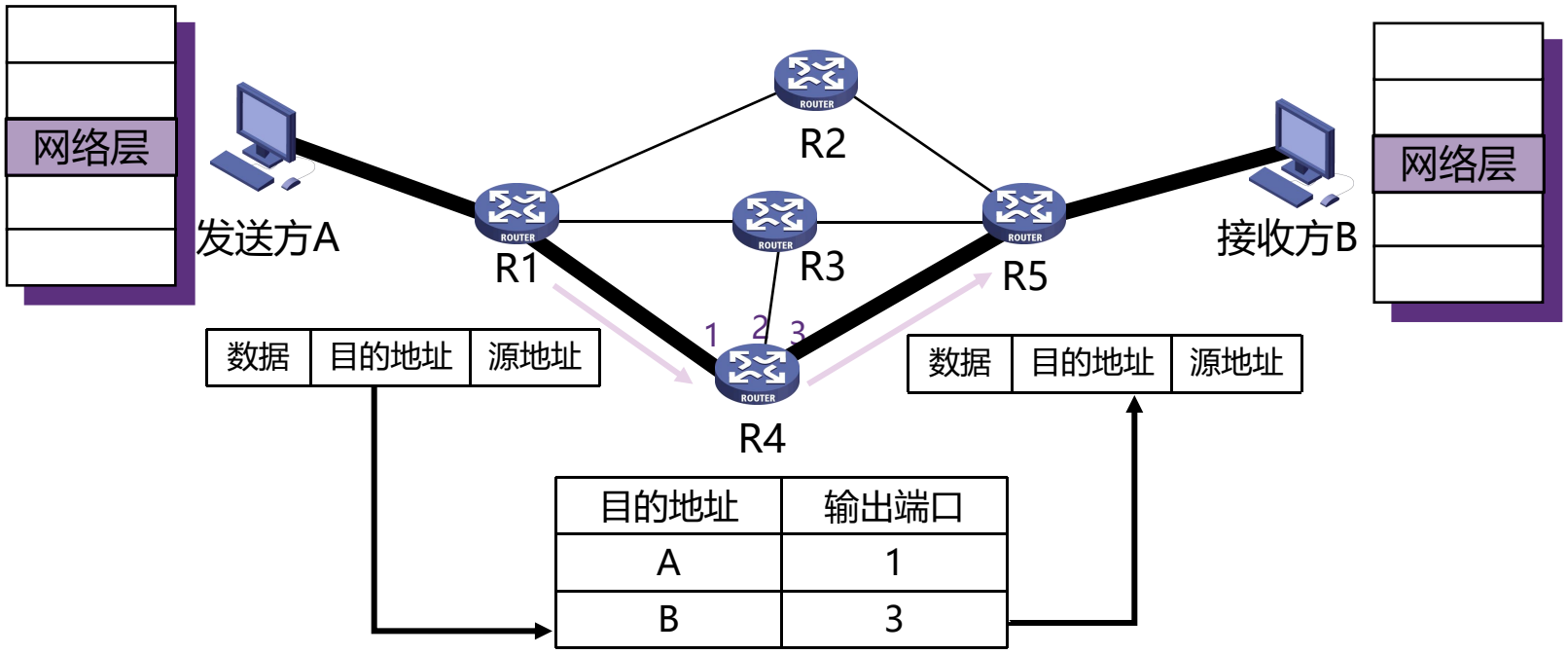


无连接的方法允许分组有选择不同路径的可能性，但这样可能会导致接收数据的失序；需要说明的是，为避免增加额外的开销进行数据排序，网络并不会完全随意地发送数据，在大多数情况下，仍然是会尽量沿着某一条路径发送。

5.1.4 虚电路与数据报网络的比较

数据报的转发策略

- **数据报转发策略：数据报转发决策基于分组的目的地址**



虚电路与数据报网络的比较

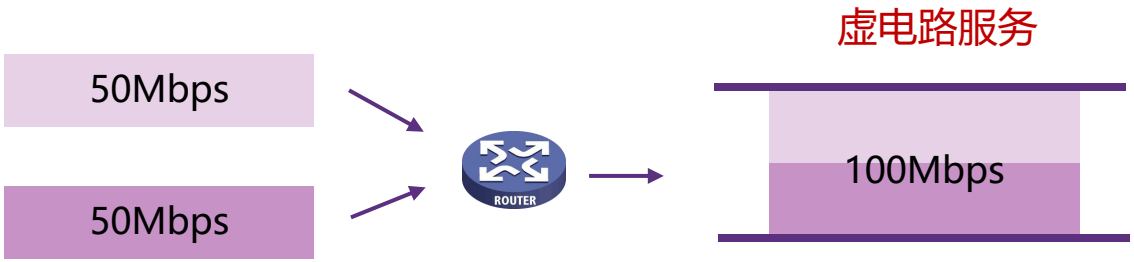
对比内容	虚电路服务	数据报服务
可靠传输的保证	可靠通信由网络保证	可靠通信由主机保证
连接的建立	必须要	不需要
地址	每个分组含有一个短的虚电路号	每个分组需要有源地址和目的地址
状态信息	建立好的虚电路要占用子网表空间	子网不存储状态信息
路由选择	分组必须经过建立好的路由发送	每个分组独立选择路由
分组顺序	总是按序到达	可能乱序
路由器失效	所有经过失效路由器的虚电路都要终止	失效结点可能丢失分组
差错处理和流量控制	网络或用户主机负责	用户主机负责
拥塞控制	容易控制	难控制

虚电路与数据报网络的比较

- 虚电路 *vs.* 数据报：哪种方法更好?
- 取决于从什么角度进行比较
  - 性能
  - 效率
  - 对失败的控制
  - 实现的复杂性
  - ……

虚电路与数据报网络的性能比较

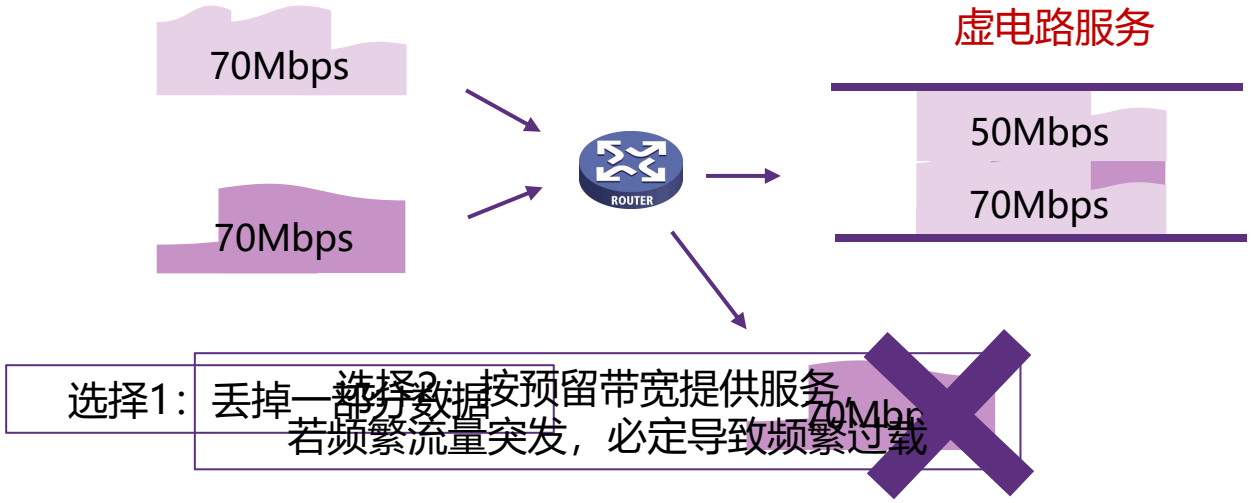
- 例1：从性能角度比较
  - 假设总带宽100Mbps，有2个数据源共享带宽
  - 如果每个数据源按50Mbps的**恒定速率**发送数据，使用虚电路服务，结果如何?



带宽不浪费
每个数据源发送数据的带宽都可被保证

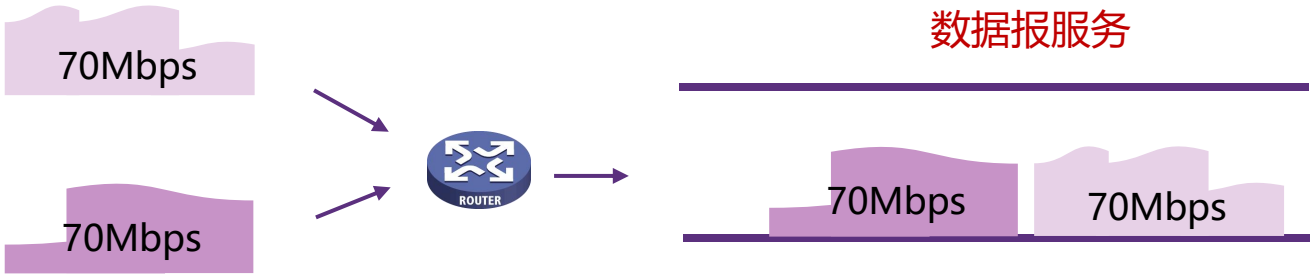
虚电路与数据报网络的性能比较

- 例1：从性能角度比较
  - 假设总带宽100Mbps，有2个数据源共享带宽
  - 如果每个数据源都是**突发流量**，且最高可达70Mbps，使用虚电路服务，结果如何?



虚电路与数据报网络的性能比较

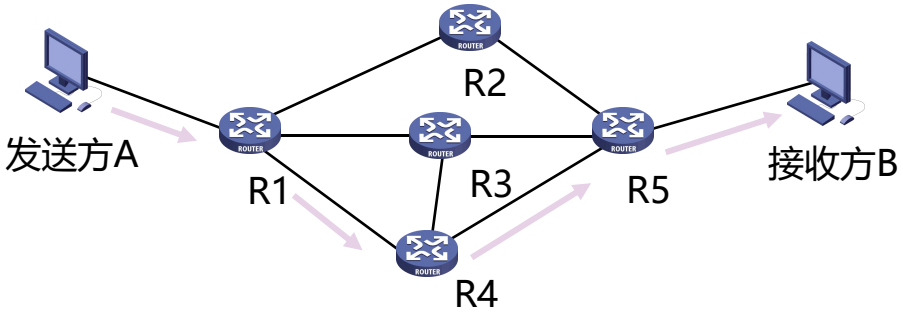
- 例1：从性能角度比较
  - 假设总带宽100Mbps，有2个数据源共享带宽
  - 如果每个数据源都是**突发流量**，且最高可达70Mbps，使用数据报服务，结果如何?



性能完全不受影响
也不会过载

虚电路与数据报网络的效率比较

- 例2：从效率角度比较
  - 假设不考虑过载，发送同样多的数据，消耗的时间比较



➢ 假设不考虑A的发送时延和链路传播时延，在上图3个转接节点的情况下，链路上的数据传输速率B bps，每个分组的长度P bit，每个分组的开销$H_v$bit（虚电路） 和$H_d$bit（数据报），虚电路分组交换的呼叫建立时间S s，虚电路每个转接点的转接延迟时间D s，数据报在每个转接点的排队时延为X s，则：

- 虚电路分组交换总时延$T=S+3[D+（P+H_v）/B]$
- 数据报分组交换总时延 $T=3[X+（P+H_d）/B]$

虚电路与数据报网络的发展历史

- 70-80年代：分组交换
  - X.25，帧中继
- 80年代末-90年代：研究人员和工业应用认为电路交换更好
  - 认为语音/电视直播将成为互联网真正的杀手级应用
- 分组交换已经成为互联网的实际服务方式，电路交换最终没有广泛应用于互联网...Why?
  - 人们重新编写应用程序以适应网络（应用程序并不需要保证带宽）
  - Email和Web广泛应用（突发流量）
- 虚电路仍有使用（MPLS，租用专线等）
  - 企业分支结构之间，昂贵，通常是静态设置（而非最初所希望的动态预留资源）

思考：
到底哪种方法更好？

本章内容

5.1 网络层服务

5.2 Internet网际协议

5.3 路由算法

5.4 Internet路由协议

5.5 路由器工作原理

5.6 拥塞控制算法

5.7 服务质量

5.8 三层交换与VPN

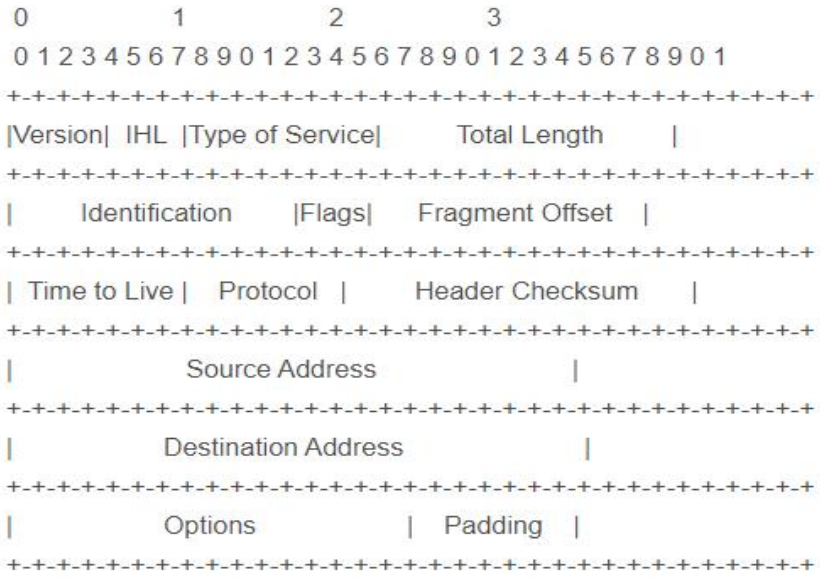
5.9 IPv6技术

1. IPv4协议
2. IP地址
3. DHCP
4. ARP
5. NAT
6. Internet控制报文协议

IPv4协议

- IPv4协议，网际协议版本4，一种无连接的协议，是互联网的核心，也是使用最广泛的网际协议版本，其后继版本为IPv6

- internet协议执行两个基本功能
  - 寻址(addressing)
  - 分片(fragmentation)



RFC 791

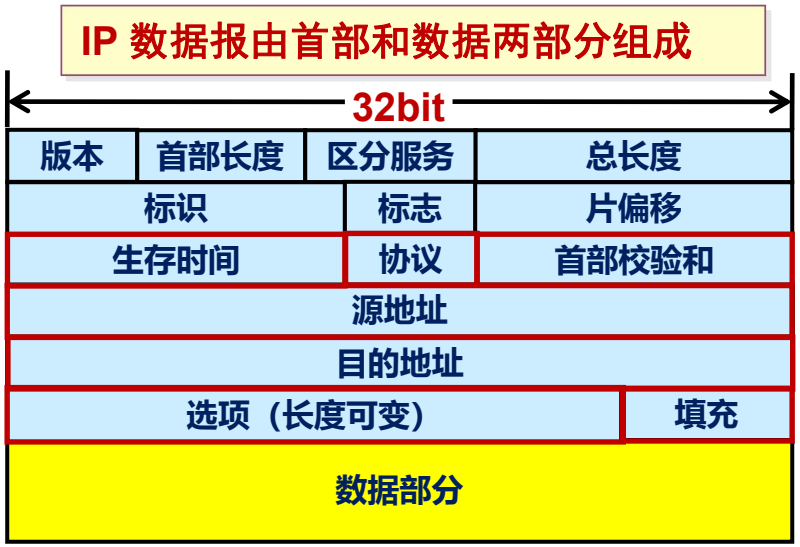
IPv4数据报格式

- 版本： 4bit，表示采用的IP协议版本
- 首部长度： 4bit，表示整个IP数据报首部的长度
- 区分服务： 8bit，该字段一般情况下不使用
- 总长度： 16bit，表示整个IP报文的长度,能表示的最大字节为2^16-1=65535字节
- 标识： 16bit， IP软件通过计数器自动产生，每产生1个数据报计数器加1；在ip分片以后，用来标识同一片分片
- 标志： 3bit，目前只有两位有意义； MF，置1表示后面还有分片，置0表示这是数据报片的最后1个；DF，不能分片标志，置0时表示允许分片
- 片偏移： 13bit，表示IP分片后，相应的IP片在总的IP片的相对位置

IP 数据报由首部和数据两部分组成

IPv4数据报格式

- 生存时间TTL(Time To Live)：8bit,表示数据报在网络中的生命周期，用通过路由器的数量来计量，即跳数（每经过一个路由器会减1）
- 协议：8bit，标识上层协议（TCP/UDP/ICMP...）
- 首部校验和：16bit，对数据报首部进行校验，不包括数据部分
- 源地址：32bit，标识IP片的发送源IP地址
- 目的地址：32bit，标识IP片的目的地IP地址
- 选项：可扩充部分，具有可变长度，定义了安全性、严格源路由、松散源路由、记录路由、时间戳等选项
- 填充：用全0的填充字段补齐为4字节的整数倍

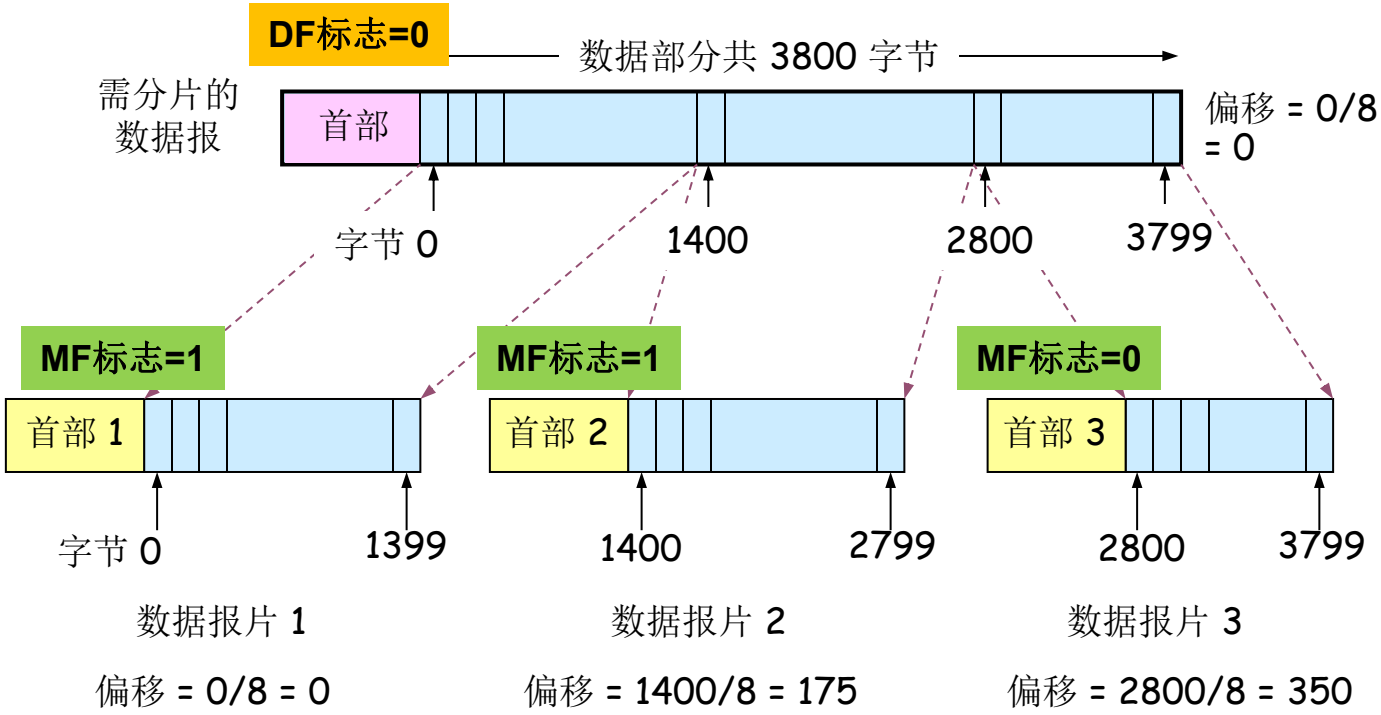


5.2.1 IPv4协议

数据报分片

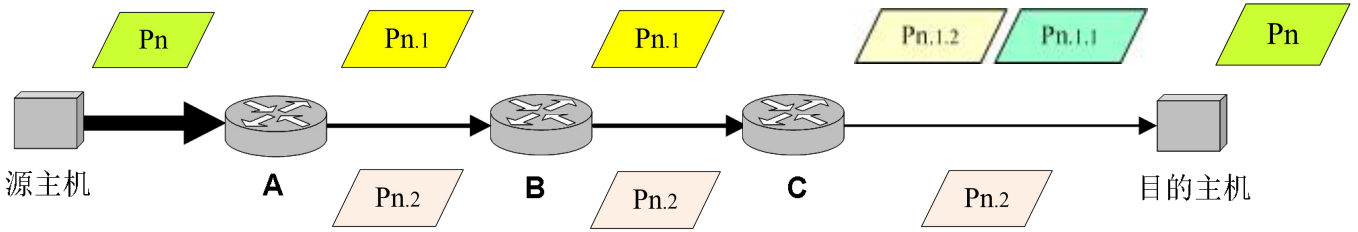
- MTU（Maximum Transmission Unit），最大传输单元
  - 链路MTU
  - 路径MTU (Path MTU)
- 分片策略
  - 允许途中分片：根据下一跳链路的MTU实施分片
  - 不允许途中分片：发出的数据报长度小于路径MTU（路径MTU发现机制）
- 重组策略
  - 途中重组，实施难度太大
  - 目的端重组（互联网采用的策略）
  - 重组所需信息：原始数据报编号、分片偏移量、是否收集所有分片

数据报分片



原始报文和分片报文具有相同的IP标识（IP头部字段）

数据报分片



- IPv4分组在传输途中可以多次分片
  - 源端系统，中间路由器（可通过标志位设定是否允许路由器分片）
- IPv4分片只在目的IP对应的目的端系统进行重组
- IPv4分片、重组字段在基本IP头部
  - 标识、标志、片偏移
- IPv6分片机制有较大变化（见IPv6部分的介绍）

IP协议功能及报头字段总结

中南大學 CENTRAL SOUTH UNIVERSITY

网络层基本功能

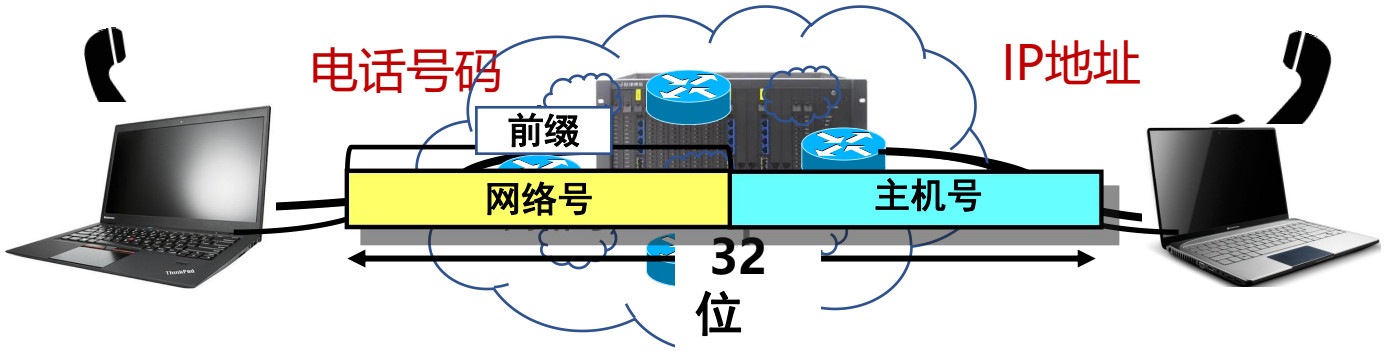
- 支持多跳寻路将IP数据报送达目的端：目的IP地址
- 表明发送端身份：源IP地址
- 根据IP头部协议类型，提交给不同上层协议处理：协议

其它相关问题

- 数据报长度大于传输链路的MTU的问题，通过分片机制解决：标识、标志、片偏移
- 防止循环转发浪费网络资源（路由错误、设备故障...），通过跳数限制解决：生存时间TTL
- IP报头错误导致无效传输，通过头部机校验解决：首部校验和

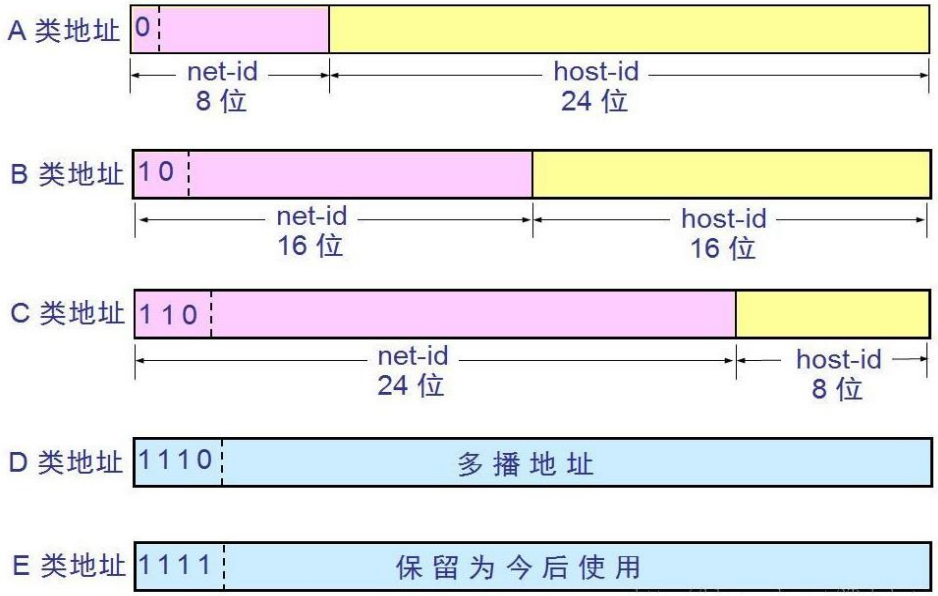
IP地址

- IP地址，网络上的每一台主机（或路由器）的每一个接口都会分配一个全球唯一的32位的标识符
- 将IP地址划分为固定的类，每一类都由两个字段组成
- 网络号相同的这块连续IP地址空间称为地址的前缀，或网络前缀



分类的IP地址

- IP地址共分为A、B、C、D、E五类，A类、B类、C类为单播地址
- IP地址的书写采用点分十进制记法，其中每一段取值范围为0到255



请判断下列地址的类型

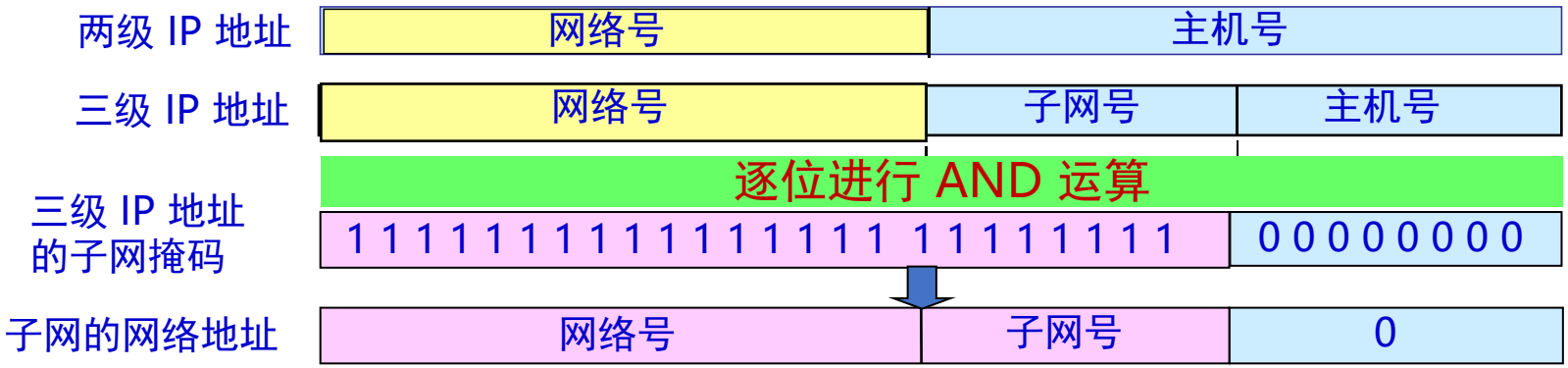
10.2.1.1	**A类**
128.63.2.100	**B类**
201.222.5.64	**C类**
256.241.201.10	**不存在，超出范围**

IP特殊地址

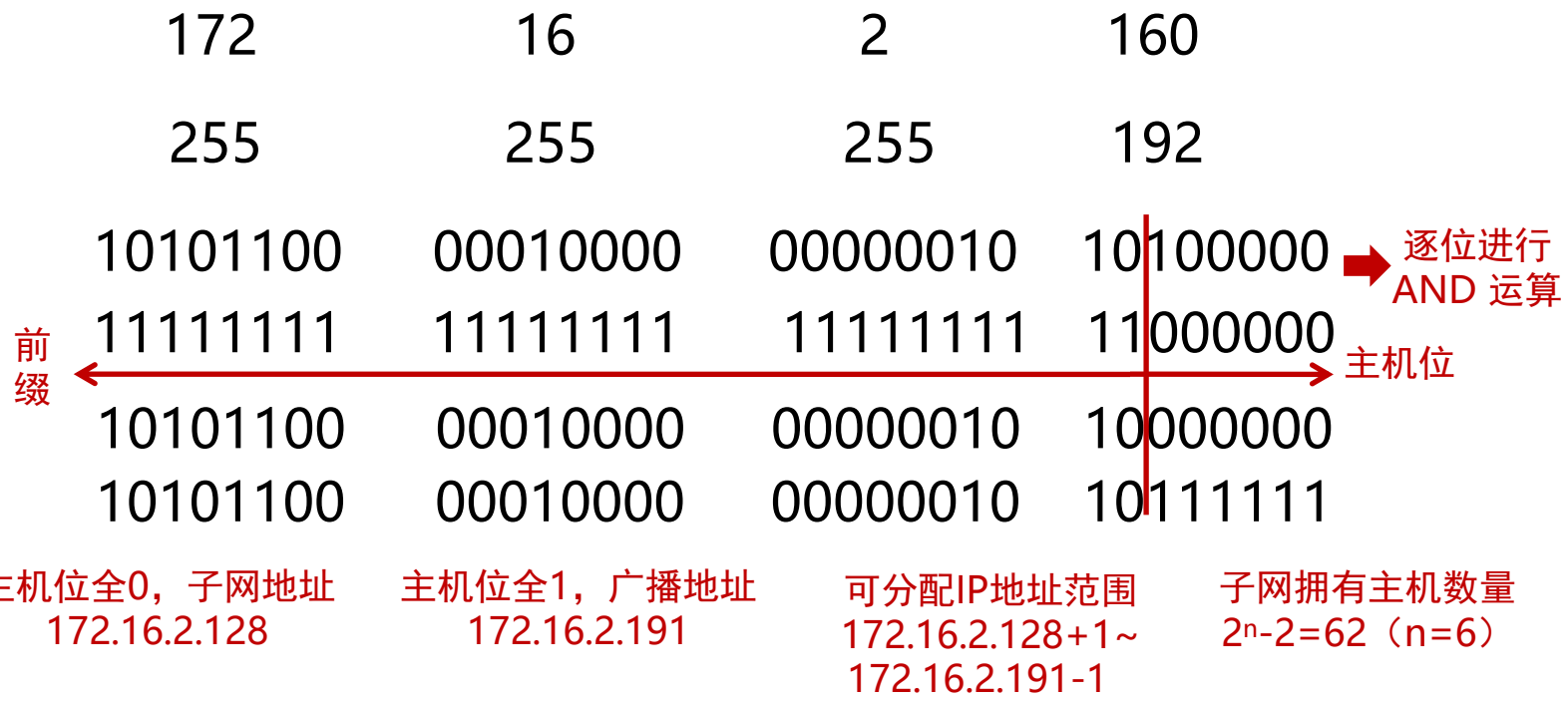
地址	用途
全0网络地址	只在系统启动时有效，用于启动时临时通信，又叫主机地址
网络127.0.0.0	指本地节点(一般为127.0.0.1)，用于测试网卡及TCP/IP软件，这样浪费了1700万个地址
全0主机地址	用于指定网络本身，称之为网络地址或者网络号
全1主机地址	用于广播，也称定向广播，需要指定目标网络
0.0.0.0	指任意地址
255.255.255.255	用于本地广播，也称有限/受限广播，无须知道本地网络地址

子网划分

- 子网划分(subnetting)，在网络内部将一个网络块进行划分以供多个内部网络使用，对外仍是一个网络
- 子网(subnet)，一个网络进行子网划分后得到的一系列结果网络称为子网
- 子网掩码(subnet mask)，与 IP 地址一一对应，是32 bit 的二进制数，置1表示网络位，置0表示主机位
- 子网划分减少了 IP 地址的浪费、网络的组织更加灵活、便于维护和管理



子网划分


172 16 2 160
255 255 255 192
10101100 00010000 00000010 10100000
11111111 11111111 11111111 11000000
逐位进行
AND 运算
前缀
主机位
10101100 00010000 00000010 10000000
10101100 00010000 00000010 10111111
主机位全0，子网地址
172.16.2.128
主机位全1，广播地址
172.16.2.191
可分配IP地址范围
172.16.2.128+1~
172.16.2.191-1
子网拥有主机数量
2n-2=62（n=6）

无类域间路由

- CIDR（Classless Inter-Domain Routing）
  - 基于可变长子网掩码实现。通过把多个地址块组合到一个路由表表项而使路由更加方便。
  - 将32位的IP地址划分为前后两个部分，并采用斜线记法，即在IP地址后加上“/”，然后再写上网络前缀所占位数

  - 一个 CIDR 地址块可以表示很多地址，这种地址的聚合常称为路由聚合（route aggregation），也称为构成超网 (supernet)
  - 聚合技术在Internet中大量使用，它允许前缀重叠，数据包按具体路由的方向发送，即具有最少IP地址的最长匹配前缀

最长前缀匹配

最长前缀匹配（Longest prefix match）

- CIDR可变长子网掩码以及路由聚合，需要最长前缀匹配来实现最精确匹配
- IP地址与IP前缀匹配时，总是选取子网掩码最长的匹配项
- 主要用于路由器转发表项的匹配，也应用于ACL规则匹配等

IP前缀（2种描述方式）		出接口号
200.23.16.0/21	11001000 00010111 00010	0
200.23.24.0/24	11001000 00010111 00011000	1
200.23.24.0/21	11001000 00010111 00011	2
Otherwise 0.0.0.0/0	--	3

IP地址：200.23.22.161 （11001000 00010111 00010110 10100001），接口0

IP地址：200.23.24.170 （11001000 00010111 00011000 10101010），接口2

最长前缀匹配

根据最长前缀匹配，下述目的IP将匹配哪个表项（出接口）？

2.128.0.0/9	interface 1
2.192.0.0/10	interface 2
2.0.0.0/8	interface 3
2.2.3.0/24	interface 4
0.0.0.0/0	interface 5

2.5.1.2 Interface 3

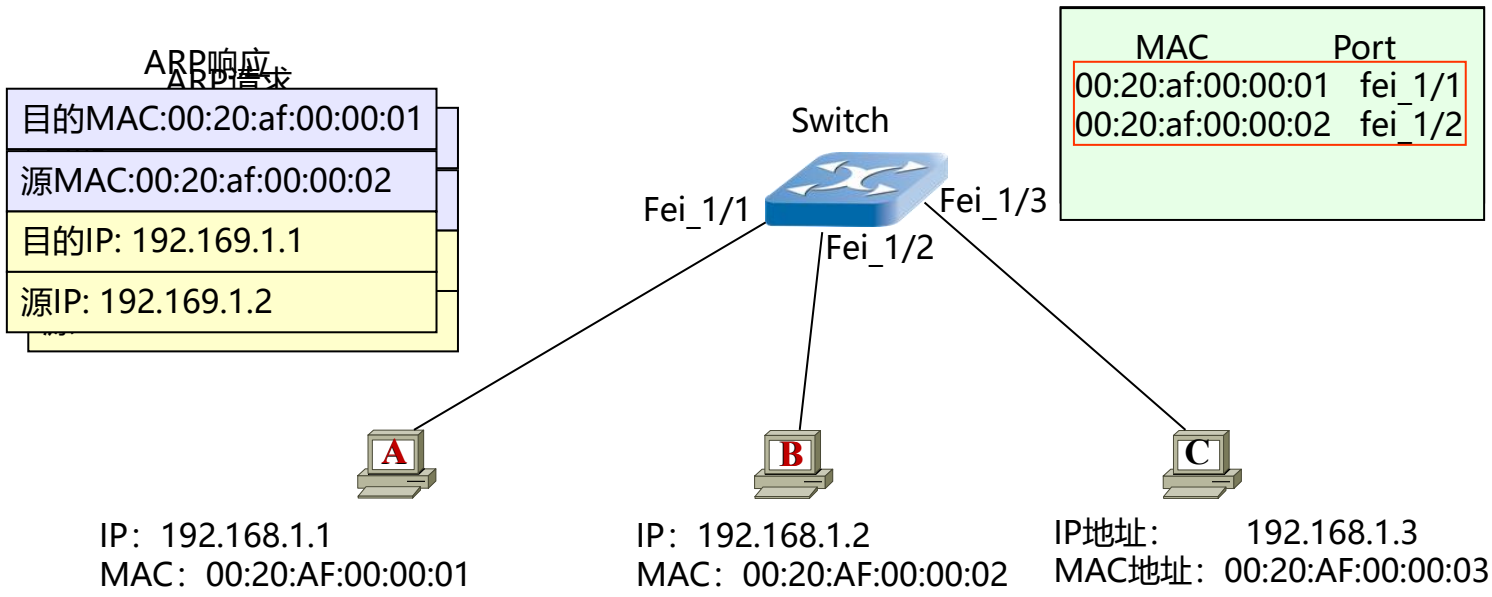
2.200.1.2 Interface 2

2.150.1.2 Interface 1

3.150.1.2 Interface 5

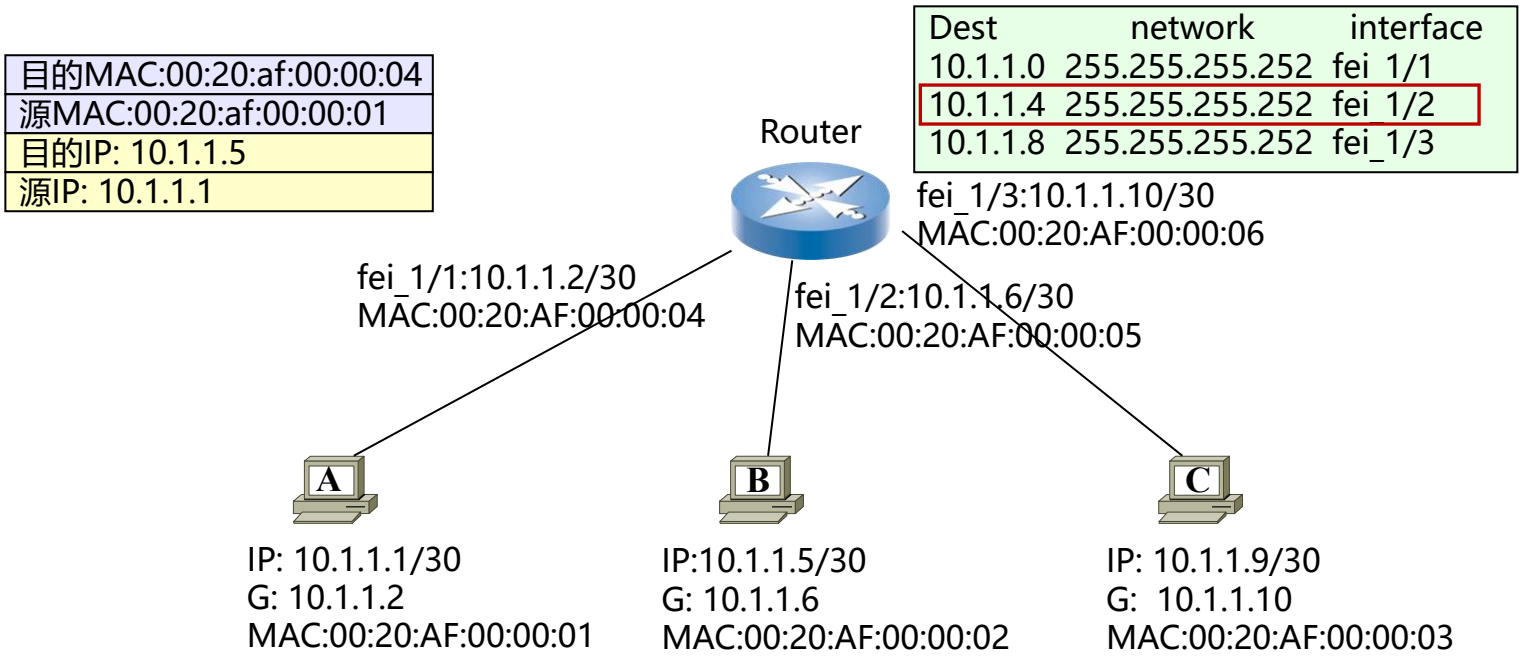
IP包转发

- 直接交付：与目的主机在同一个IP子网内



IP包转发

- 间接交付：与目的主机不在同一个IP子网内



IPv4地址如何获取

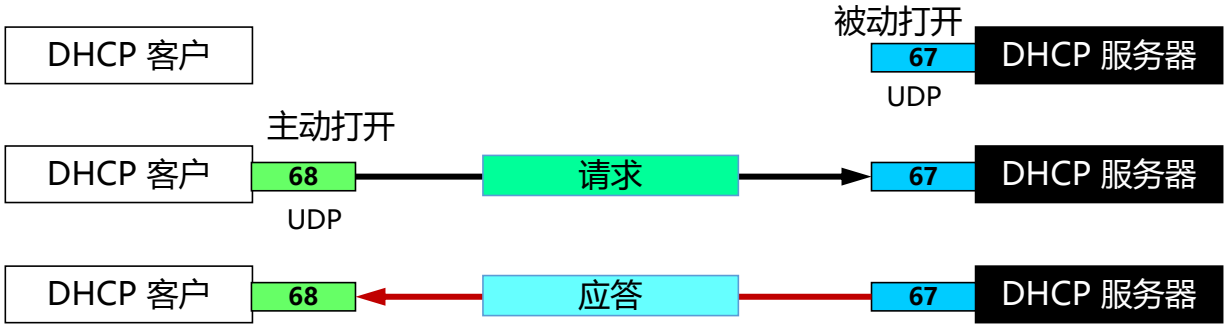
- **公有IP地址要求全球唯一**
  - ICANN（Internet Corporation for Assigned Names and Numbers）即互联网名字与编号分配机构向ISP分配，ISP再向所属机构或组织逐级分配
- **静态设定**
  - 申请固定IP地址，手工设定，如路由器、服务器
- **动态获取**
  - 使用DHCP协议或其他动态配置协议
  - 当主机加入IP网络，允许主机从DHCP服务器动态获取IP地址
  - 可以有效利用IP地址，方便移动主机的地址获取

DHCP动态主机配置协议

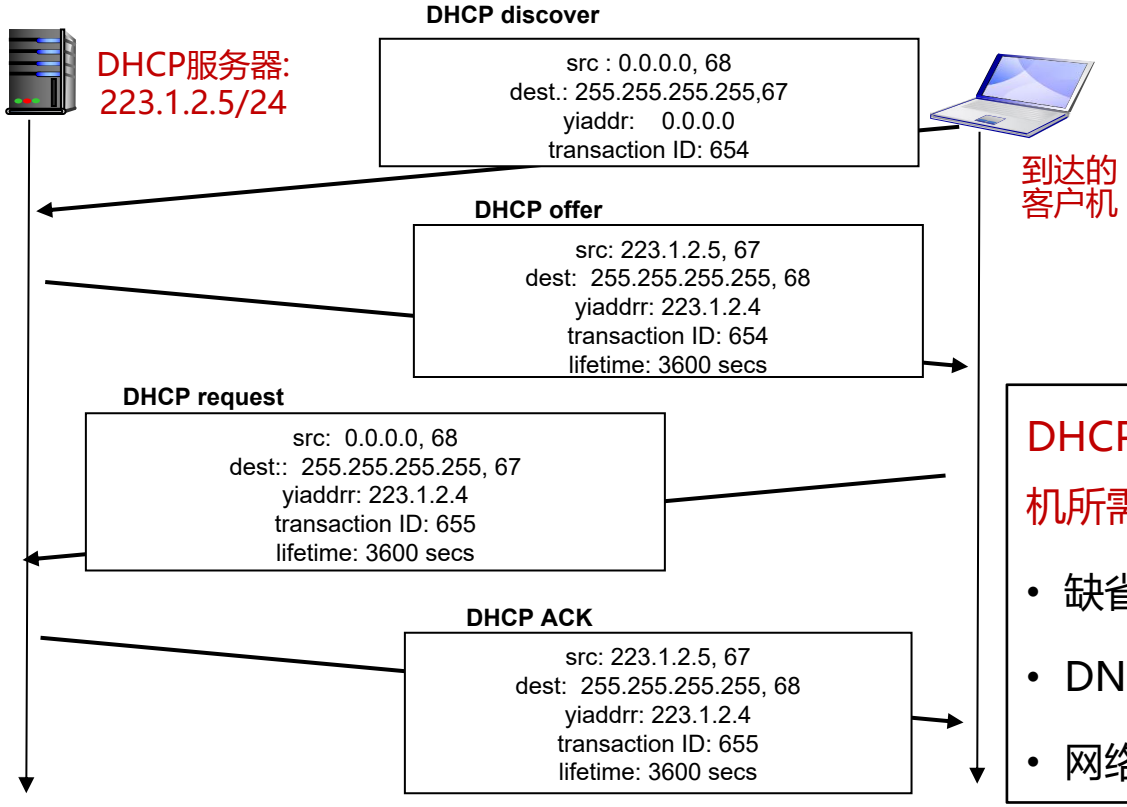
- DHCP：动态主机配置协议
  - 当主机加入IP网络，允许主机从DHCP服务器动态获取IP地址
  - 可以有效利用IP地址，方便移动主机的地址获取

➢ 工作模式：客服/服务器模式（C/S）

- 基于 UDP 工作，服务器运行在 67 号端口， 客户端运行在 68 号端口



DHCP 工作过程



DHCP服务不只返回客户机所需的IP地址，还包括:

- 缺省路由器IP地址
- DNS服务器IP地址
- 网络掩码

DHCP 工作过程

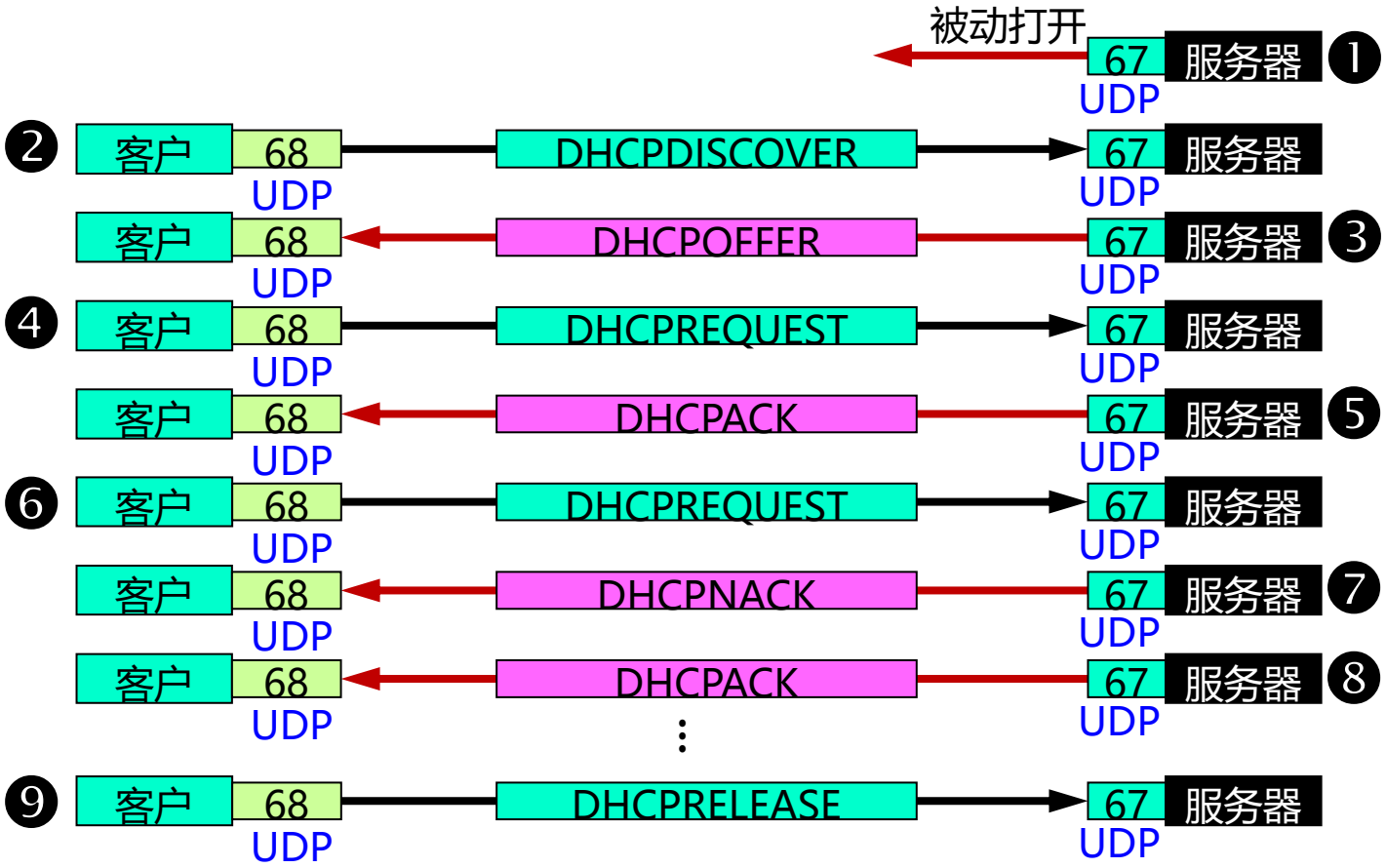
- DHCP 客户从UDP端口68以广播形式向服务器发送发现报文（DHCPDISCOVER）
- DHCP 服务器单播发出提供报文（DHCPOFFER）
- DHCP 客户从多个DHCP服务器中选择一个，并向其以广播形式发送DHCP请求报文（DHCPREQUEST）
- 被选择的DHCP服务器单播发送确认报文（DHCPACK）

DHCP 工作过程

- 小结

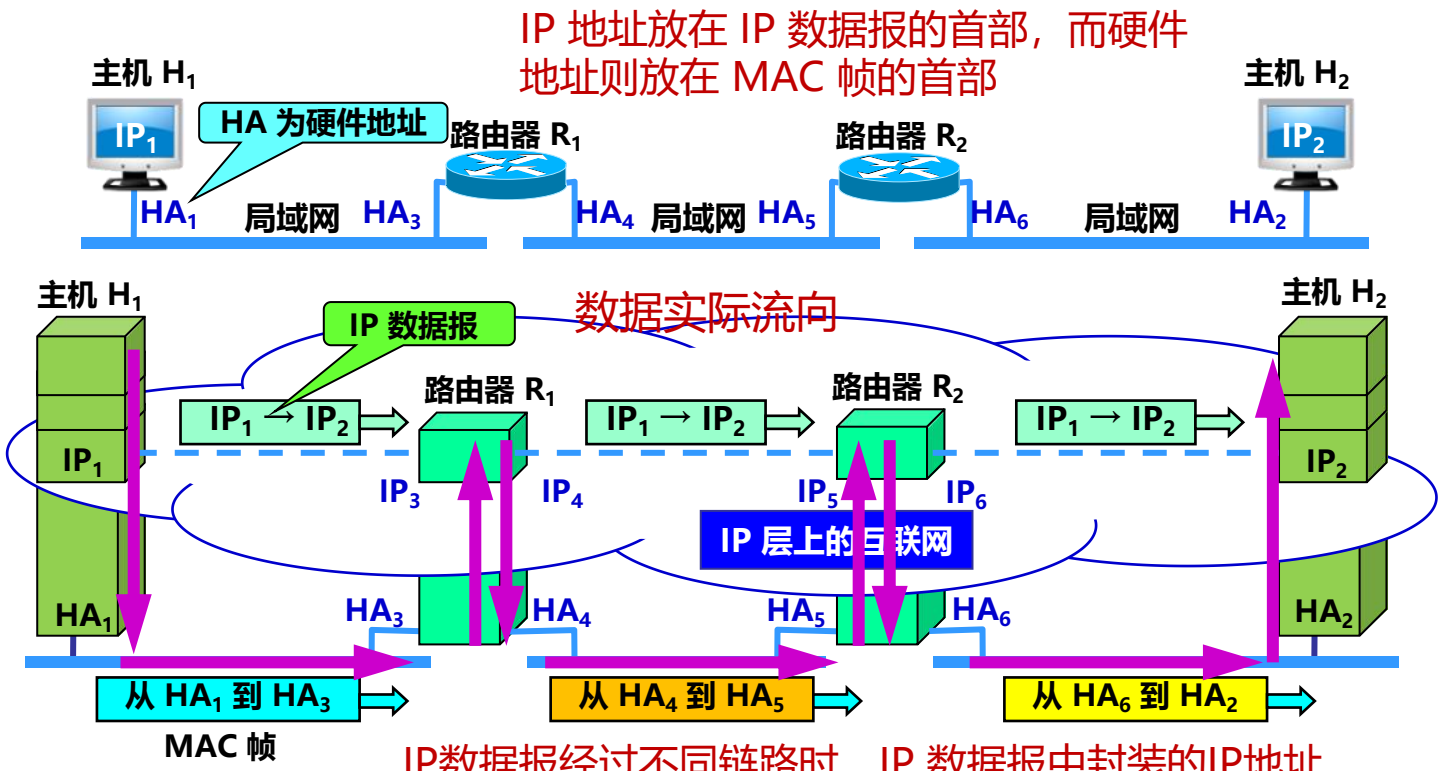
阶段	源MAC	目标MAC	源IP	目标IP	传输形式
Discover	PC机的MAC	全FF	0.0.0.0	255.255.255.255	广播
Offer	DHCP服务器或者中继器路由的MAC	DHCP客户机的MAC	DHCP服务器或中继路由器的IP地址	准备分配的IP地址	单播
Request	PC机的MAC	全FF	0.0.0.0	255.255.255.255	广播
Ack	DHCP服务器或者中继器路由的MAC	DHCP客户机的MAC	DHCP服务器或中继路由器的IP地址	准备分配的IP地址	单播

DHCP 完整工作过程



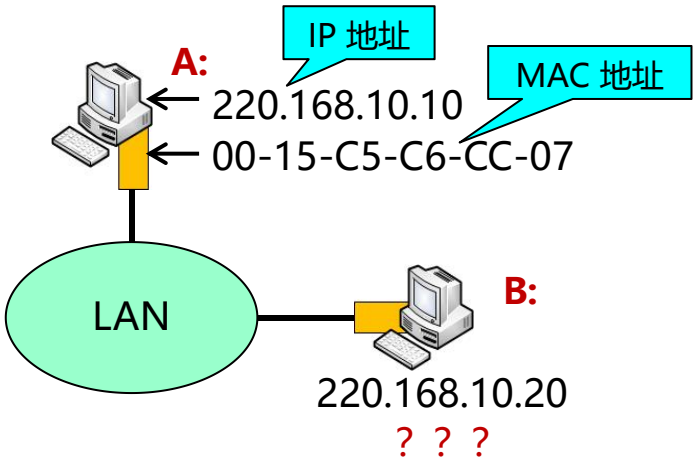
注：说明见备注

IP 与 MAC地址

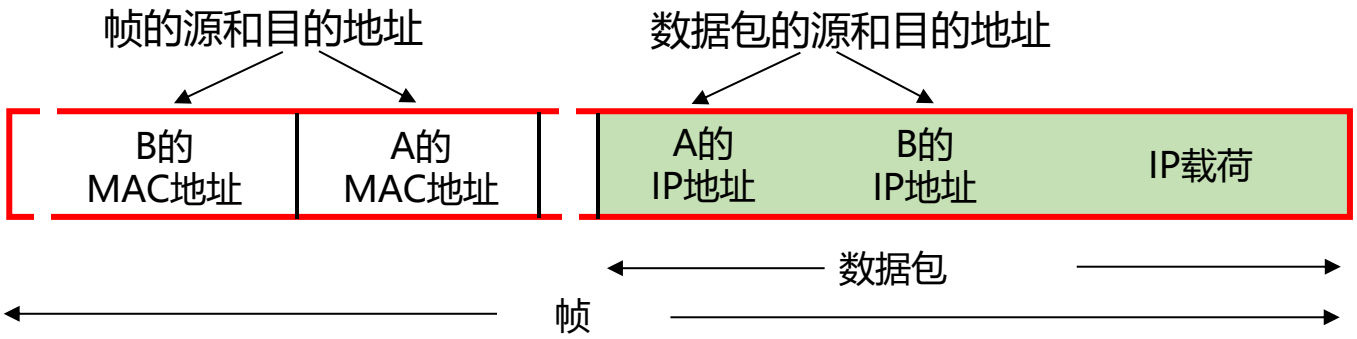

IP 地址放在 IP 数据报的首部，而硬件地址则放在 MAC 帧的首部
主机 H1
主机 H2
IP1
IP2
HA 为硬件地址
路由器 R1
路由器 R2
HA1
局域网
HA3
HA4
局域网
HA5
HA6
局域网
HA2
数据实际流向
IP 数据报
IP1 → IP2
IP1 → IP2
IP1 → IP2
IP3
IP4
IP5
IP6
IP 层上的互联网
从 HA1 到 HA3
从 HA4 到 HA5
从 HA6 到 HA2
MAC 帧
IP数据报经过不同链路时，IP 数据报中封装的IP地址不发生改变，而Mac帧中的硬件地址是发生改变的

ARP地址解析协议

- IP数据包转发：从主机A到主机B
  - 检查目的IP地址的网络号部分
  - 确定主机B与主机A属相同IP网络
  - 将IP数据包封装到链路层帧中，直接发送给主机B



问题：给定B的IP地址，如何获取B的MAC地址?

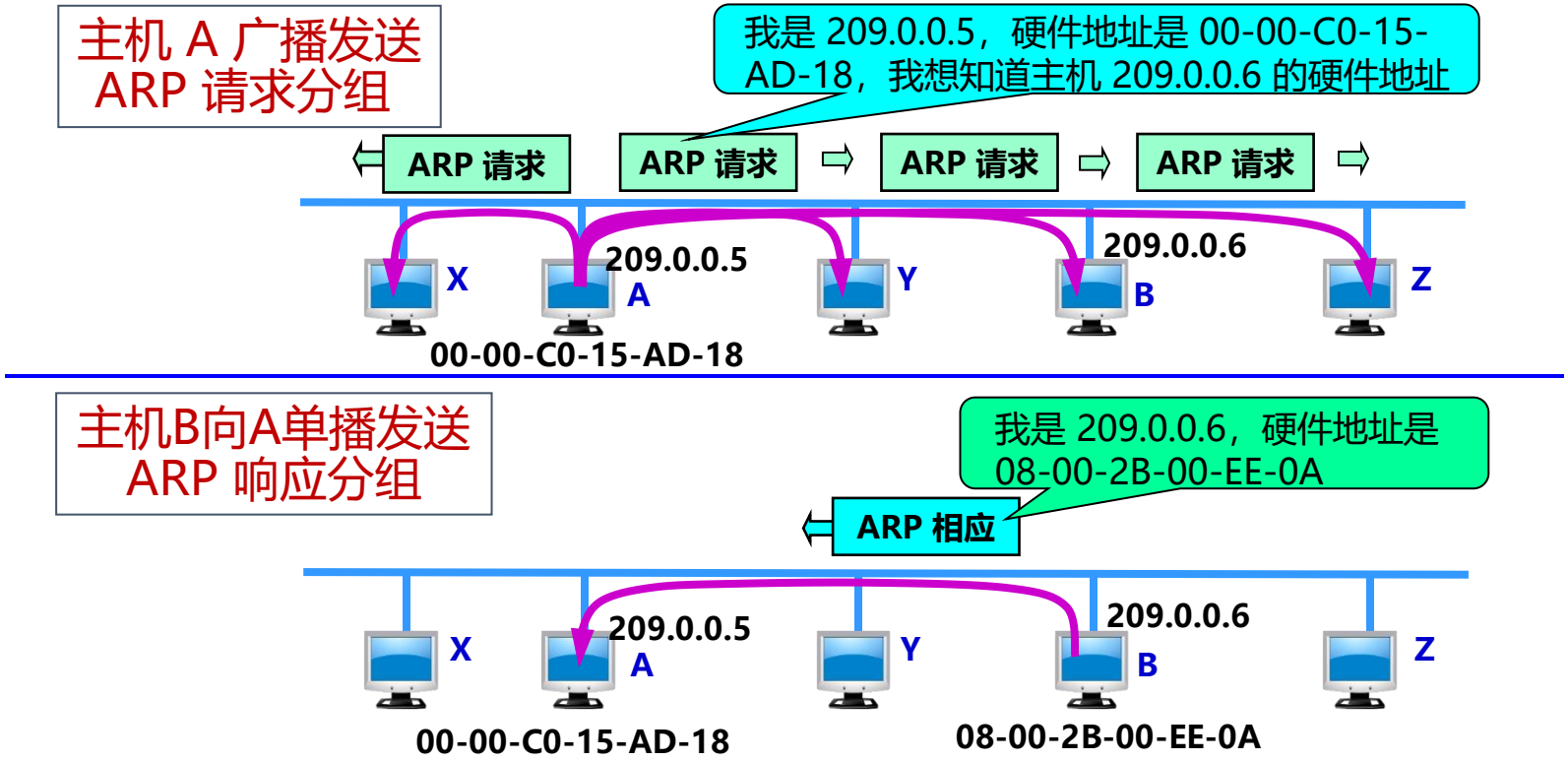


ARP地址解析协议

- A已知B的IP地址，需要获得B的MAC地址（物理地址）
- 如果A的ARP表中缓存有B的IP地址与MAC地址的映射关系，则直接从ARP表获取
- 如果A的ARP表中未缓存有B的IP地址与MAC地址的映射关系，则A广播包含B的IP地址的ARP query分组
  - 在局域网上的所有节点都可以接收到ARP query
- B接收到ARP query分组后，将自己的MAC地址发送给A
- A在ARP表中缓存B的IP地址和MAC地址的映射关系
  - 超时时删除

思考：ARP的优化策略?

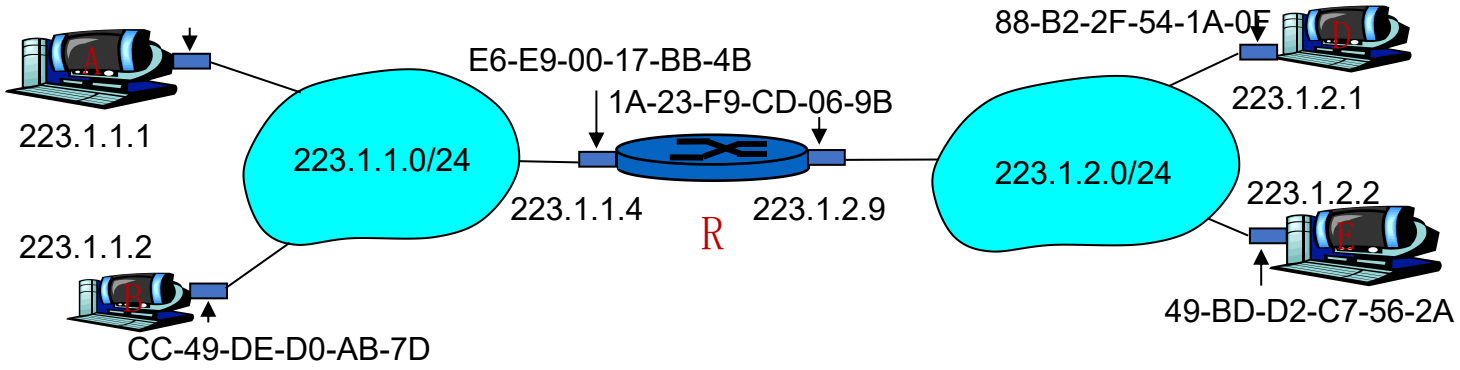
ARP协议工作过程


主机 A 广播发送
ARP 请求分组
我是 209.0.0.5，硬件地址是 00-00-C0-15-AD-18，我想知道主机 209.0.0.6 的硬件地址
ARP 请求
ARP 请求
ARP 请求
ARP 请求
X
209.0.0.5
A
Y
209.0.0.6
B
Z
00-00-C0-15-AD-18
主机B向A单播发送
ARP 响应分组
我是 209.0.0.6，硬件地址是 08-00-2B-00-EE-0A
ARP 相应
X
209.0.0.5
A
Y
209.0.0.6
B
Z
00-00-C0-15-AD-18
08-00-2B-00-EE-0A

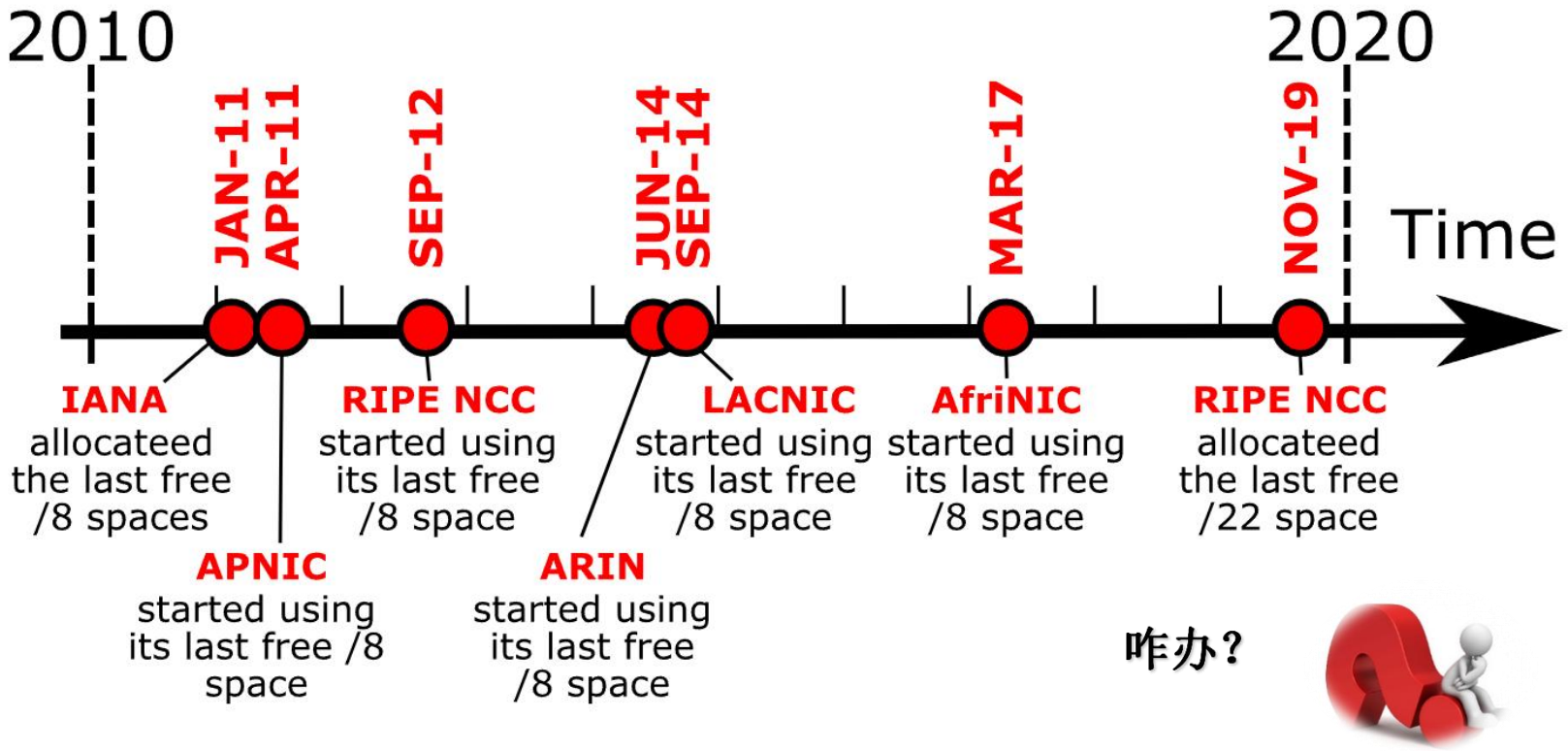
路由到另一个局域网

- A创建IP数据包（源为A、目的为E）
- 在源主机A的路由表中找到路由器R的IP地址223.1.1.4
- A根据R的IP地址223.1.1.4，使用ARP协议获得R的MAC地址
- A创建数据帧（目的地址为R的MAC地址）
- 数据帧中封装A到E的IP数据包
- A发送数据帧，R接收数据帧

例：从A经过R到E



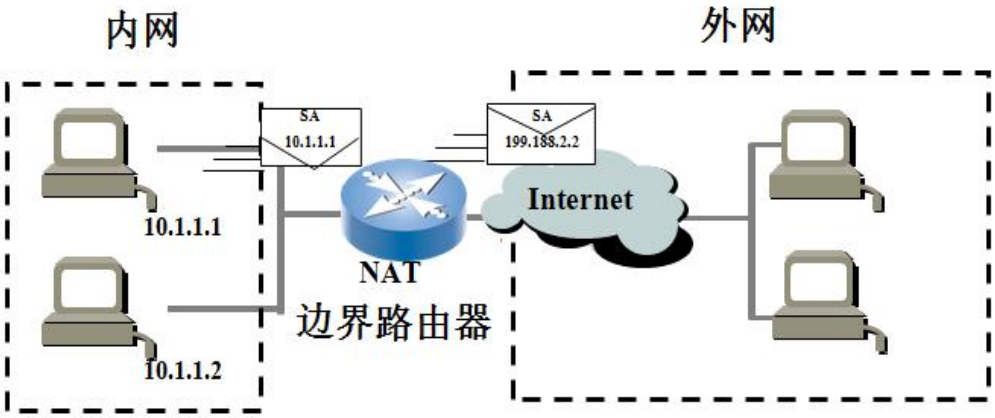
IPv4地址池耗尽



来源：IPv4 address exhaustion
https://en.wikipedia.org/wiki/IPv4_address_exhaustion

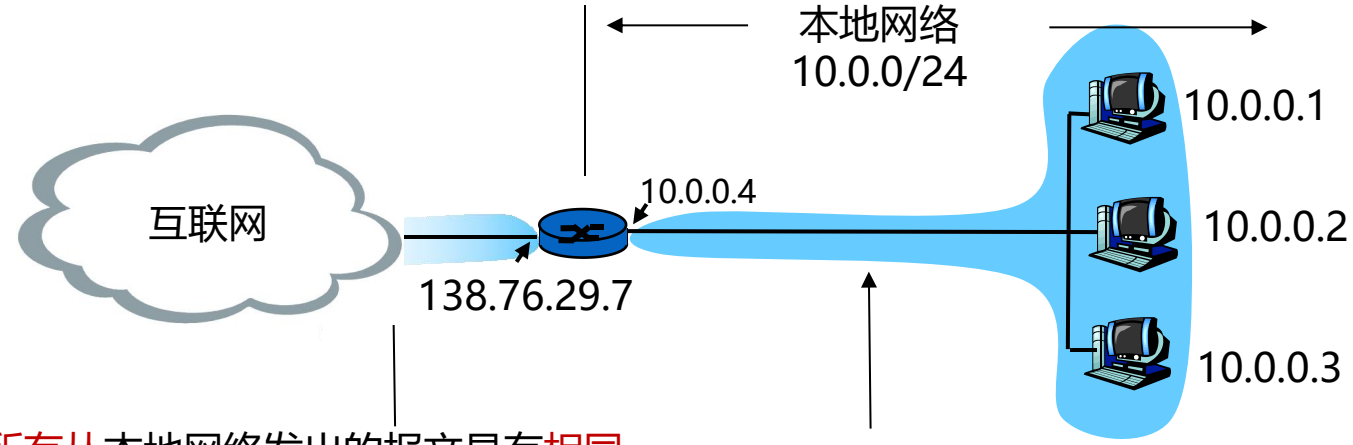
网络地址转换

- 网络地址转换(NAT)用于解决IPv4地址不足的问题，是一种将私有（保留）地址转化为公有IP地址的转换技术
- 私有IP地址：
  - A类地址：10.0.0.0--10.255.255.255
  - B类地址：172.16.0.0--172.31.255.555
  - C类地址：192.168.0.0--192.168.255.255

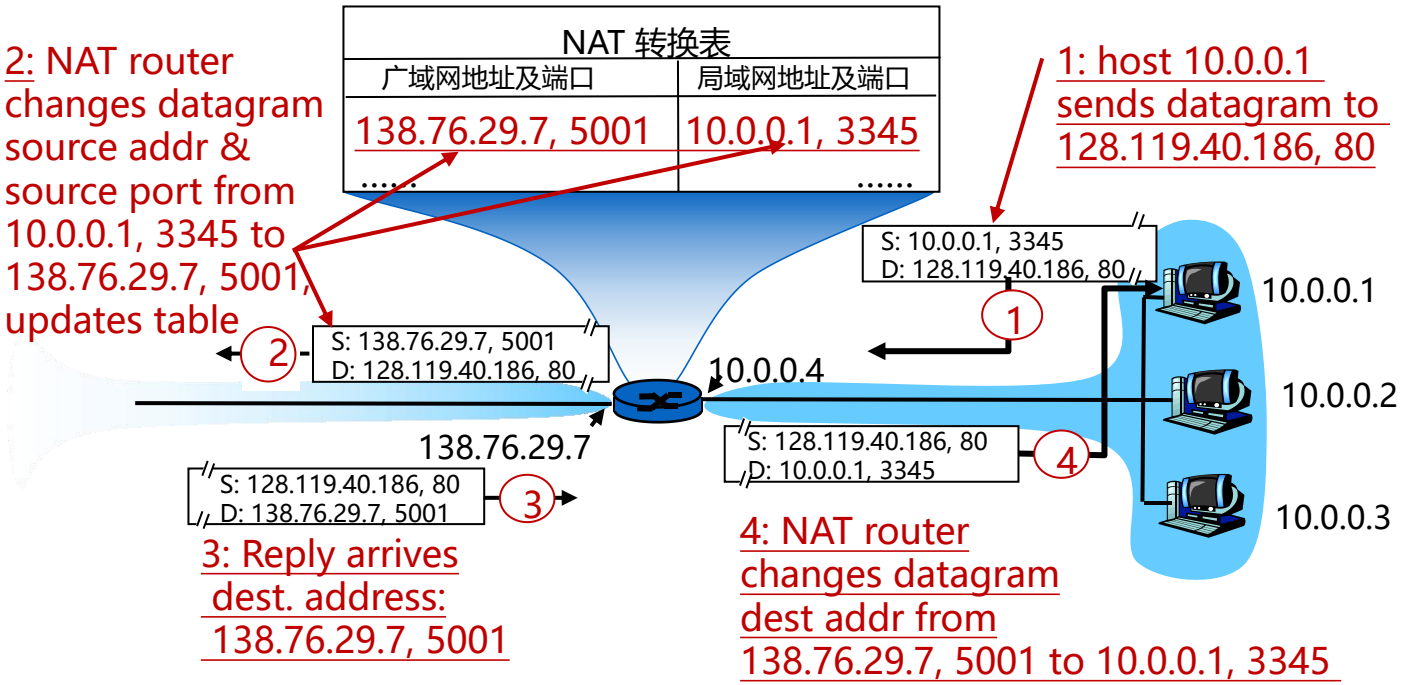


5.2.5 NAT

NAT工作机制


本地网络
10.0.0/24
10.0.0.1
10.0.0.2
10.0.0.3
10.0.0.4
互联网
138.76.29.7
所有从本地网络发出的报文具有相同的单一源IP地址: 138.76.29.7
报文的源地址(发出时)或目的地址(接收时)都属于10.0.0.0/24的地址
不同的源端口号

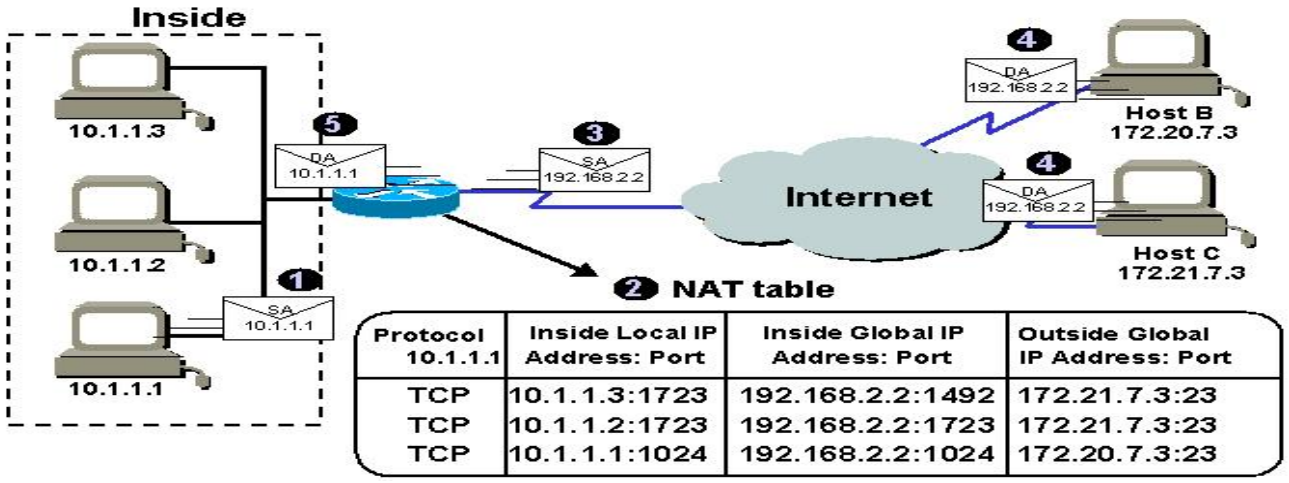
NAT工作机制



思考：同一主机不同应用，或者不同主机的同一端口，NAT转换如何处理？

NAT工作机制

- 出数据报：外出数据报用 NAT IP地址(全局), 新port # 替代 源IP地址(私有), port #
- NAT转换表：每个 (源IP地址, port #)到(NAT IP地址, 新port #) 映射项
- 入数据报：对每个入数据报的地址字段用存储在NAT表中的(源IP地址, port #)替代对应的 (NAT IP地址, 新port #)



Protocol 10.1.1.1	Inside Local IP Address: Port	Inside Global IP Address: Port	Outside Global IP Address: Port
TCP	10.1.1.3:1723	192.168.2.2:1492	172.21.7.3:23
TCP	10.1.1.2:1723	192.168.2.2:1723	172.21.7.3:23
TCP	10.1.1.1:1024	192.168.2.2:1024	172.20.7.3:23

网络地址转换

- NAT根据不同的IP上层协议进行NAT表项管理
  - TCP，UDP，ICMP
- 传输层TCP/UDP拥有16-bit 端口号字段
  - 所以一个WAN侧地址可支持60,000个并行连接

➢ NAT的优势
- 节省合法地址，减少地址冲突
- 灵活连接Internet
- 保护局域网的私密性

➢ 问题或缺点
- 违反了IP的结构模型，路由器处理传输层协议
- 违反了端到端的原则
- 违反了最基本的协议分层规则
- 不能处理IP报头加密
- 新型网络应用的设计者必须要考虑 NAT场景，如 P2P应用程序