网络互联 (I)

华中科技大学电子信息与通信学院
通信工程系
陈京文
Email: jwchen@hust.edu.cn
2020.10.16

内容提要

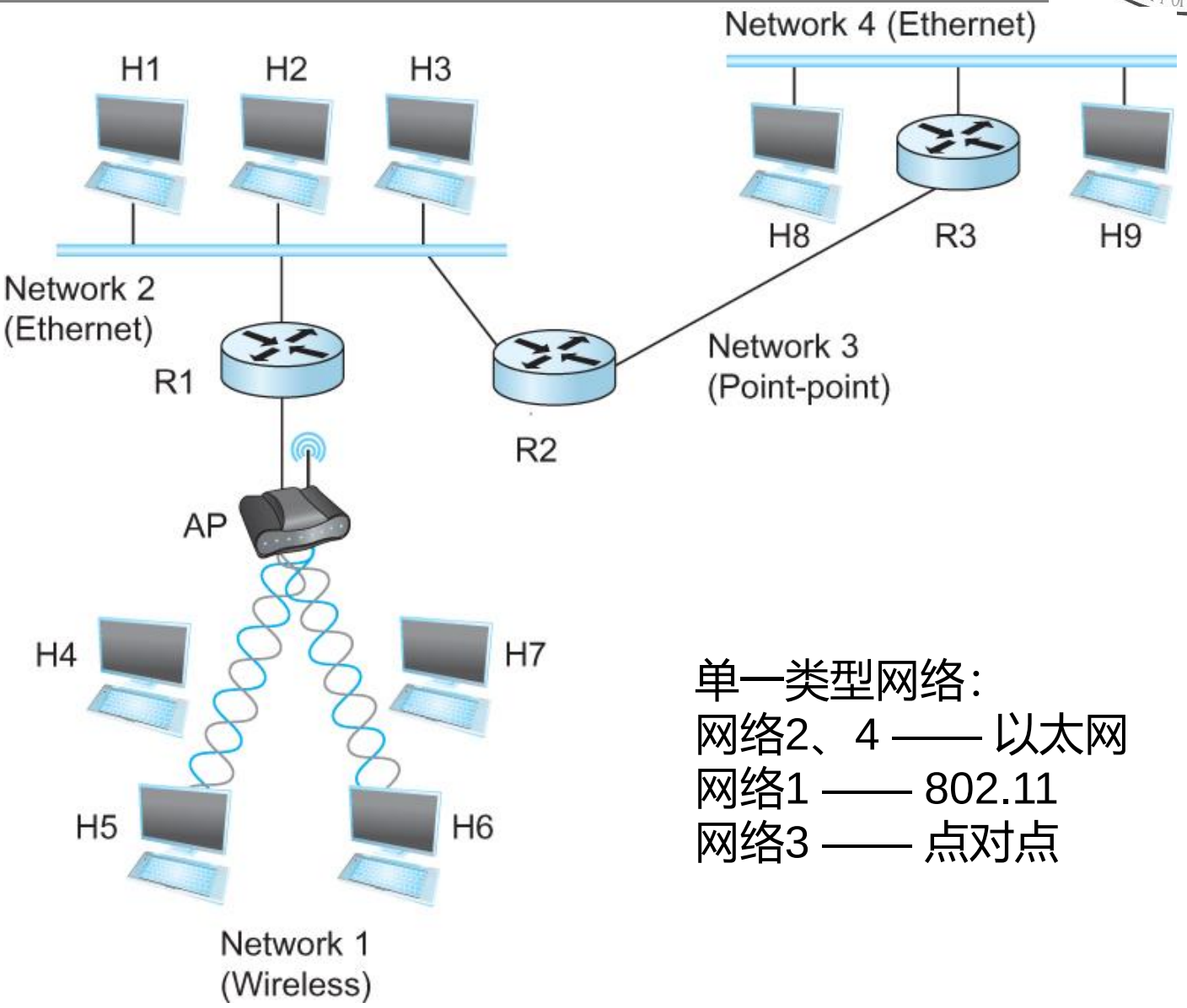
- 网络互联简介
- IP协议
  - 服务模型，分段与重组，原始编址方案
- 原始IP编址方案的改进
  - 子网划分，CIDR
- 辅助协议
  - ARP，DHCP，ICMP

网络互联

- “单一”类型网络
  - 采用单一类型网络技术，如点到点链路、多路接入网络、同一种链路层交换机等

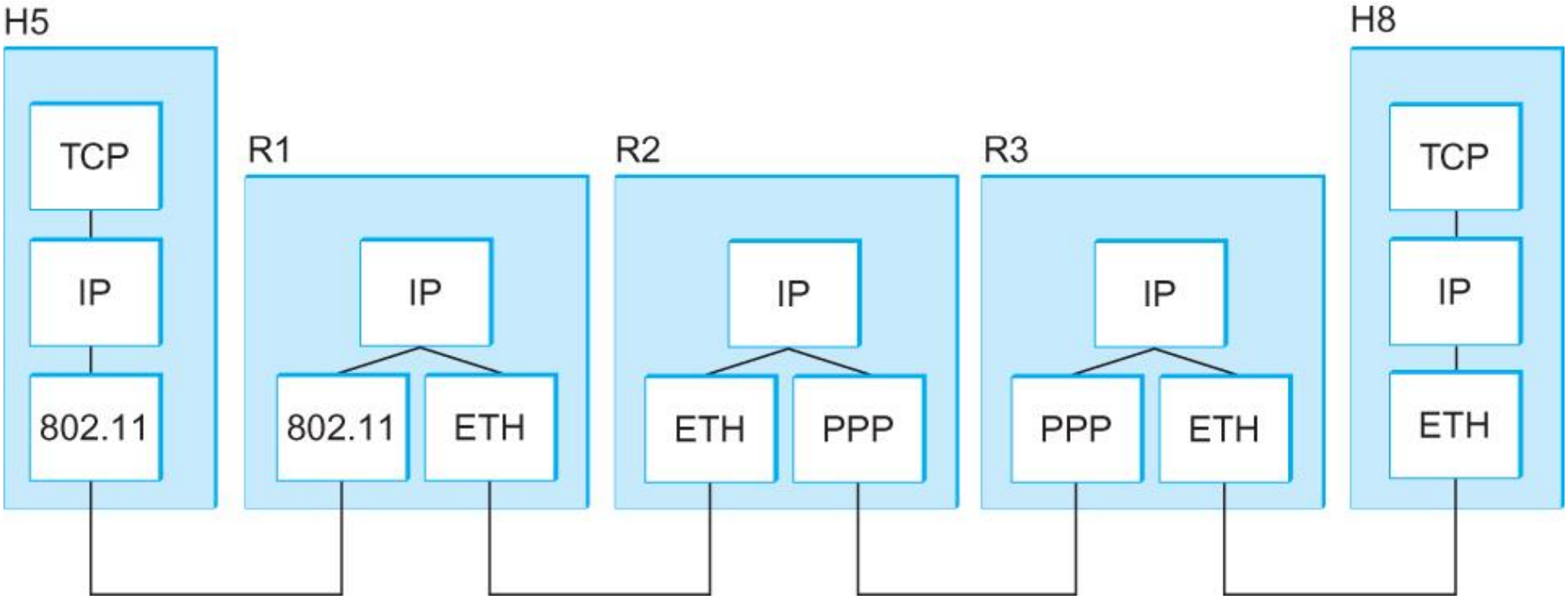
- 互联网络(internetwork)
  - 互联多个不同类型网络(如以太网、令牌环等)，使得这些不同类型网络中的主机之间能够相互通信
  - 需要独立(区别)于所互联的单一类型网络的协议，用于不同单一类型网络中主机之间的通信，即网络层协议
  - 同时需要一种中间节点连接单一类型网络，称为路由器(router)(早期称为网关gateway)
  - 示例：因特网(Internet)

互联网络示例



单一类型网络：
网络2、4 —— 以太网
网络1 —— 802.11
网络3 —— 点对点

互联网络示例(续)



数据报传输及相关的协议

网络基本技术问题

- 数据交换
  - 交换方式：电路、数据报、虚电路
  - 编址：与交换方式相关
- 路由
  - 获取网络拓扑信息，计算传输路径(即路由计算)
  - 将路由计算结果配置到交换/路由节点
- 传输服务
  - 传输质量：可靠性，速率，时延等
- 其它
  - 速率控制：发送端发送速率自适应于接收端速率、网络当前容量
  - 网络资源分配
  - …

网络互联(即网络层)同样存在相应的问题

网络互联的技术挑战

- 互联的网络的异质性(Heterogeneity)
  - 寻址(addressing)
  - 媒质接入控制(media access control)
  - 路由(routing)
  - 服务模型(service model)

- 可扩展性(Scalability)
  - 寻址：地址空间 —— 分级(hierarchical)或平铺式
  - 路由：计算路径，转发表(forwarding table)及其规模

内容提要

- 网络互联简介
- IP协议
  - 服务模型，分段与重组，原始编址方案
- 原始IP编址方案的改进
  - 子网划分，CIDR
- 辅助协议
  - ARP，DHCP，ICMP

IP协议

- Internet Protocol (IP)
  - 一种连接不同单一类型网络的网络层协议
  - 包含分组格式、分组转发、控制协议、与上层协议接口等的规范
  - TCP/IP协议栈的重要组成部分
- 互联单一类型网络的中间节点：IP路由器(router)
- 版本
  - IPv4：用于Internet和许多私有网络
  - IPv6：用于少数网络，如校园网和少数大型Internet应用服务提供商，目前还未广泛部署
- 最成功的网络互联协议！

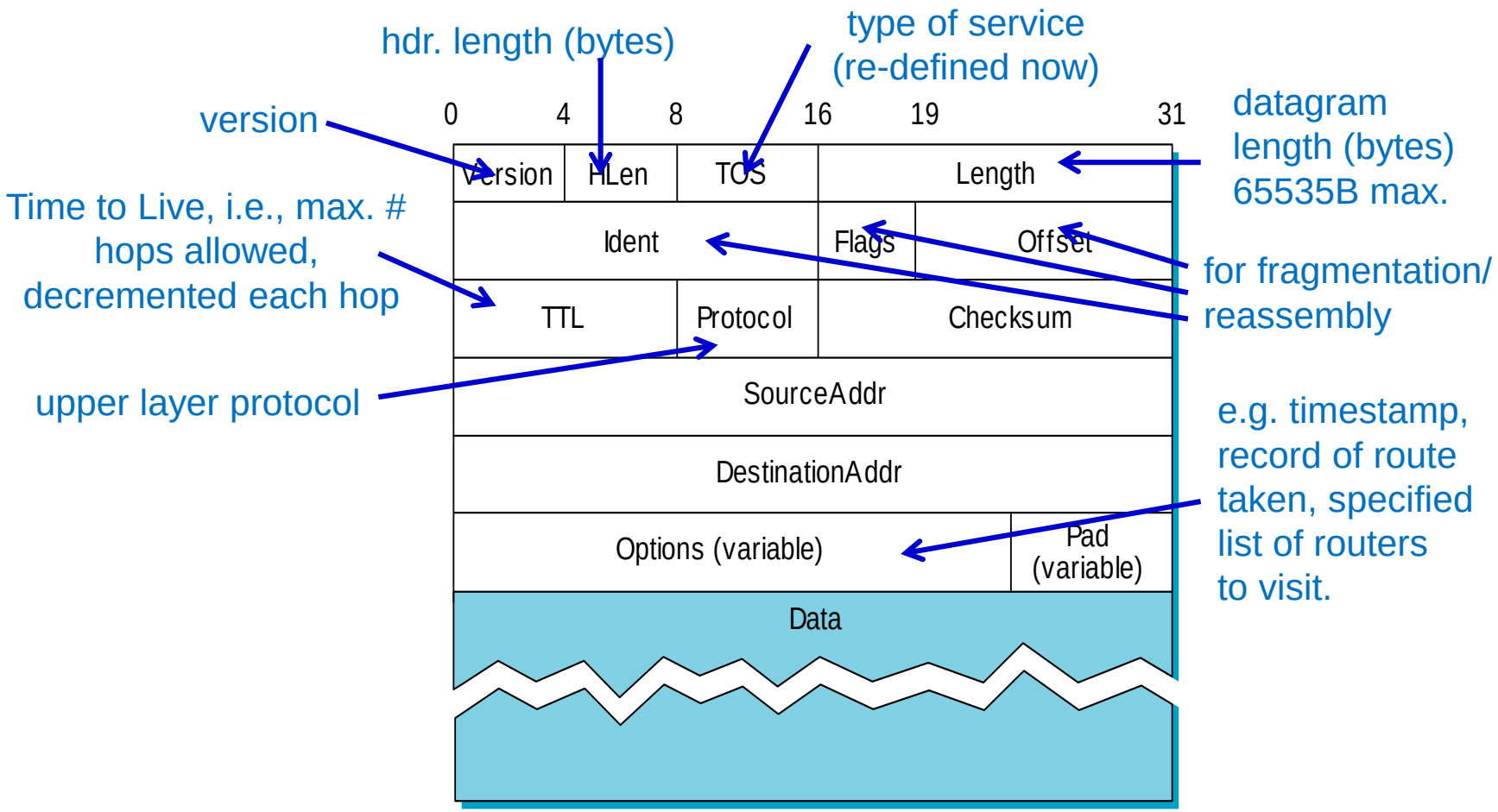
IP服务模型

- 基本思路：**最小服务(minimal service)**
  - 复杂的操作(如可靠传输)留给主机执行
  - 能够与任何类型网络技术协同，即使是还未推出的网络技术
  - 将新网络技术对于IP带来的麻烦降到最低
  - 从而使得IP可运行于任何底层网络(running over anything)

- IP服务模型(service model)
  - 数据报传送(datagram delivery)
    - 数据报交换：无连接，每分组处理
  - 尽力交付服务(best effort service)
    - 尽可能交付数据报
    - 但没有保证，即不可靠传输服务(注：并不表明分组丢失)

—— Cerf和Kahn提出的网络互联原则的一部分

IP数据报格式


hdr. length (bytes)
type of service
(re-defined now)
version
datagram
length (bytes)
65535B max.
Time to Live, i.e., max. #
hops allowed,
decremented each hop
for fragmentation/
reassembly
upper layer protocol
e.g. timestamp,
record of route
taken, specified
list of routers
to visit.
0
4
8
16
19
31
Version
HLen
TOS
Length
Ident
Flags
Offset
TTL
Protocol
Checksum
SourceAddr
DestinationAddr
Options (variable)
Pad
(variable)
Data

MTU的多样性及影响

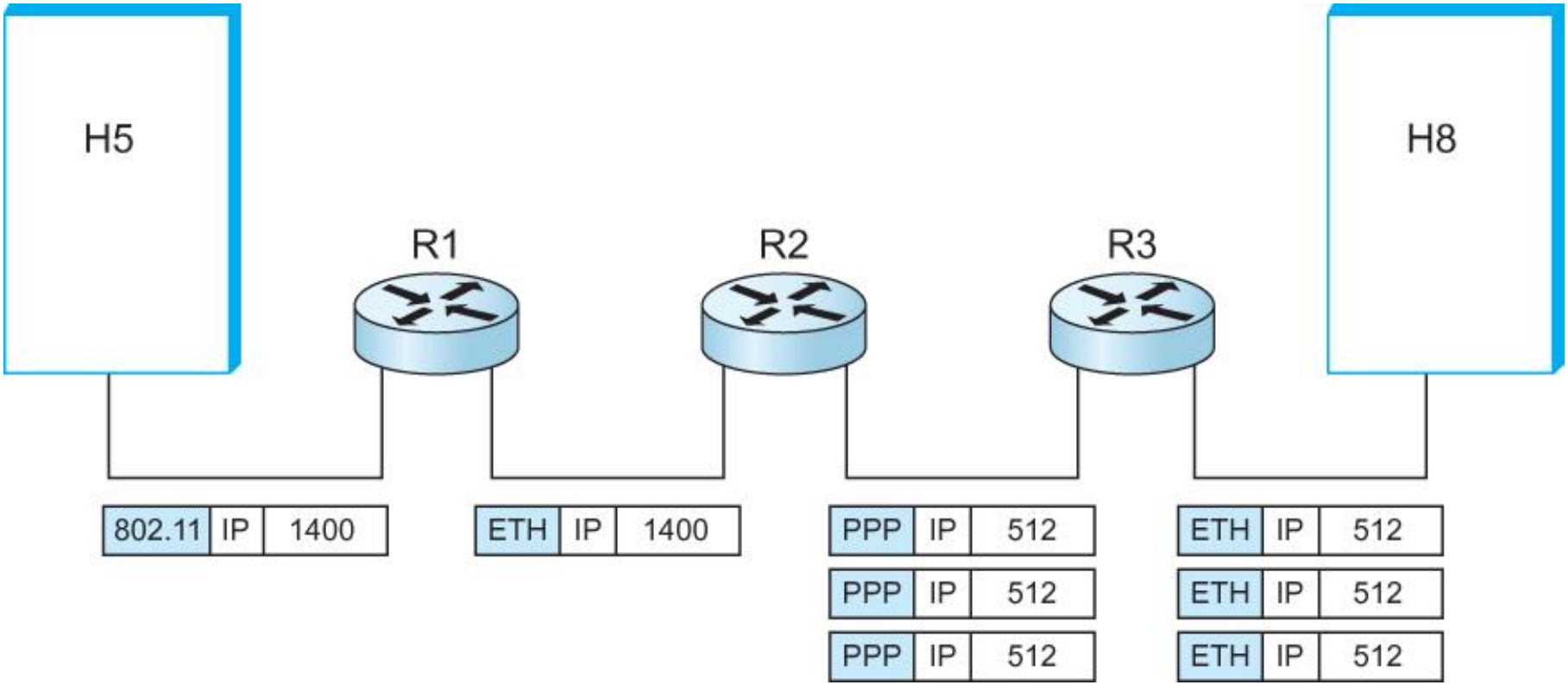
- 最大传输单元(Maximum Transmission Unit, MTU)：可传送的最大分组长度
  - 例如，以太网MTU — 1500 bytes, PPP MTU — 532 bytes
- 由于底层网络MTU各不相同，有时需要将大数据包分割成较小的数据包，然后重组恢复原来的大数据包
- 为何不采用互联的所有网络的最小MTU?
  - 新的网络技术会出现，其MTU未知
  - 效率方面的考虑：如数据包首部比例较高，相同长度的数据传输需要更多数据包承载和处理

IP分段与重组

- 基本策略
  - 主机采用其所附着的网络的MTU作为其发送数据报标准长度
  - 随后的路由节点中，如数据报长度 > 输出链路MTU，则进行分段
  - 如有需要，可以再次分段
  - 每个分段都是自包含(self-contained)的数据报，独立传输
  - 数据报重组只在目的地主机进行
  - 发生分段丢失时，IP本身并不进行恢复处理

- 相关问题
  - 如果丢失了一个分段，目的地主机将试图重组整个数据报，直至其放弃，即丢弃该数据报中已收到的其它分段

IP分段与重组示例


H5
R1
R2
R3
H8
802.11 IP 1400
ETH IP 1400
PPP IP 512
PPP IP 512
PPP IP 512
ETH IP 512
ETH IP 512
ETH IP 512

IP编址

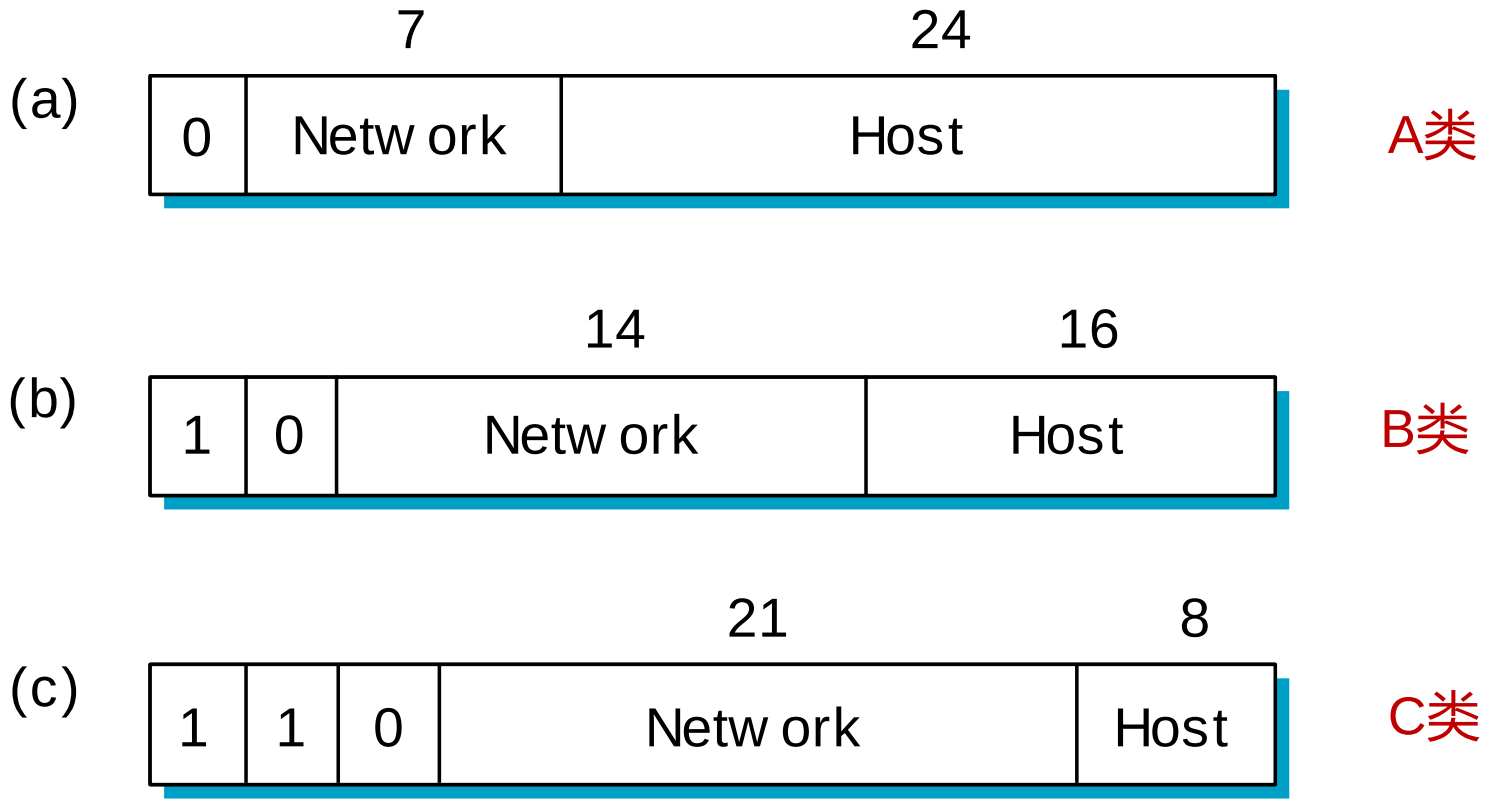
- 低层网络寻址机制的多样性
  - 互联的每一种底层网络都有着自己的寻址机制

- IP编址：无论底层网络采用何种寻址机制，采用统一的编址机制
  - 长度为4字节，通常记为点分十进制形式，如202.114.0.242 (校园网主要域名服务器的IPv4地址)
  - 分层式编址：应对Internet的大规模

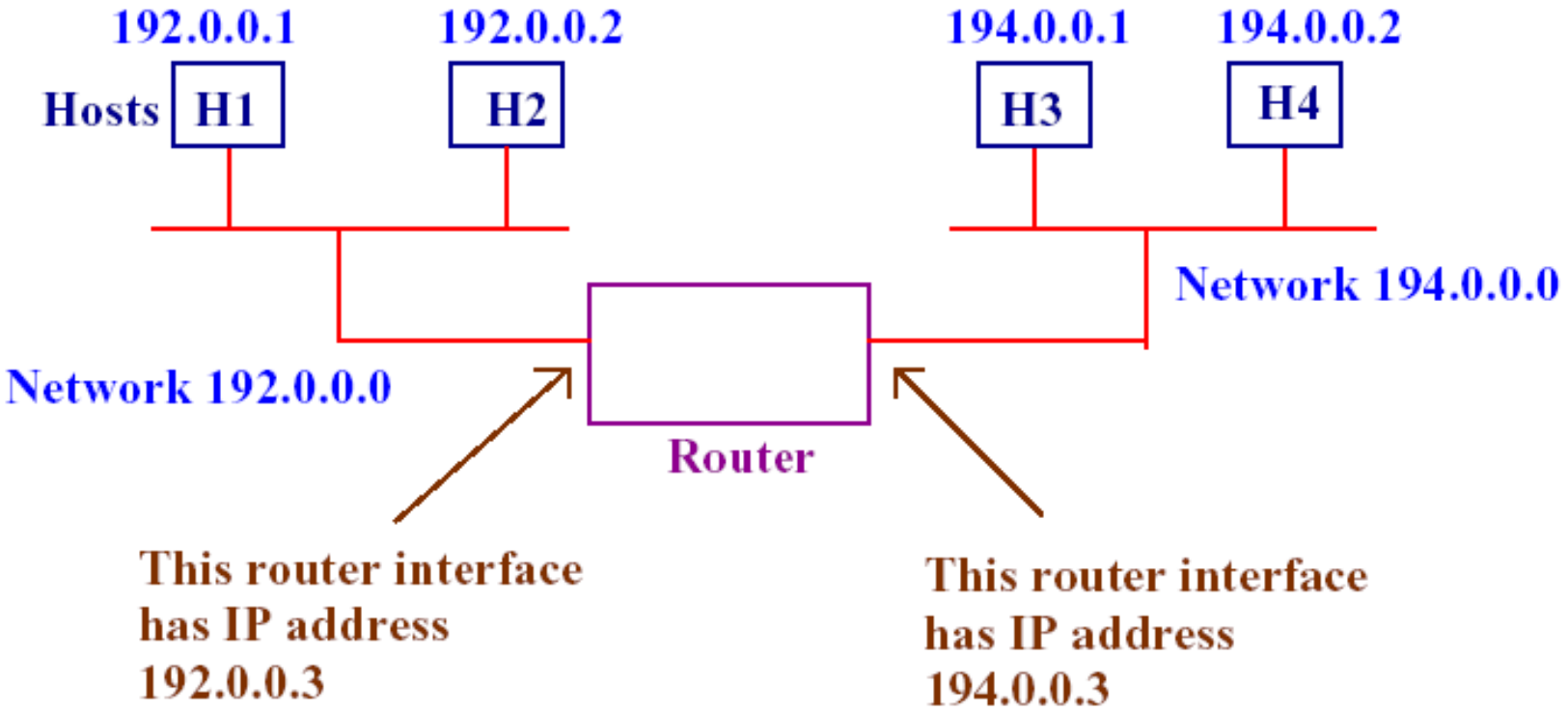
- 考虑到互联的低层网络已有的编址以及IP编址，两个地址空间并存意味着重复处理，是否合理？

IP的分层式编址

- IP地址分为两个部分
  - 网络(network) —— 也称为网络号
  - 主机(host) —— 也称为主机号
- 原始IP编址采用类别：A, B, C三类



IP地址分配对象为网络接口



Assumption: Network part is given by the first 24 bits of the IP addresses in both networks

IP编址基本原则

- IP地址中，网络号唯一地标识一个网络
  - 所有连接至一个网络的主机和路由节点的网络接口IP地址有着相同的网络号

- 给定任何IP地址，可以确定这一地址对应IP网络(号)
  - 如何确定?

- 数据包传送的关键在于传输数据包至其目的网络，剩下的任务(即在目的IP网络中传输至目的主机)容易完成
  - 如何完成?

内容提要

- 网络互联简介
- IP协议
  - 服务模型，分段与重组，原始编址方案
- 原始IP编址方案的改进
  - 子网划分，CIDR
- 辅助协议
  - ARP，DHCP，ICMP

原始IP编址方案的问题

- 问题1：地址空间利用的效率低
- 问题2：路由/转发表过长(下次课介绍路由/转发)
- 两个改进
  - 子网划分(Subnetting)：解决问题1
  - 无类别域间路由(Classless InterDomain Routing, CIDR)：解决问题1、2

IP地址耗尽问题

- 32位IP地址空间 —— 2^{32}(约4G)个有效地址
  - 并非所有地址均为主机及其网络接口所用
  - 部分空间用于多播(multicast)或预留其它用途，即D、E类地址

- IP地址不够用的原因
  - Internet的快速增长需要更多的地址
  - 原始IP编址方案的低效率地址分配
    - 例1：10台主机的网络，也需要一个C类地址块
      - 地址利用效率：10/256 ≈ 4%
    - 例2：300台主机的网络，则需要一个B类地址块
      - 地址利用效率：300/65536 ≈ 0.5%

子网划分

- 子网划分(subnetting) —— 将一个网络(地址块)分成多个较小的子网(地址块)

- 例：包括3个组成部分的IP网络，其中2个各包含50台主机，1个包含100台主机
  - 原始IP编址：需要3个C类地址块
    - 效率为200/(3x256) ≈ 26%
  - 采用子网划分：只需1个C类地址块
    - 将给定C类地址块分为3个子网地址块：2个包含64个地址，用于2个有着50台主机的网络；1个包含128个地址，用于1个有着100台主机的网络
    - 效率为200/256 ≈ 78%

如何划分子网

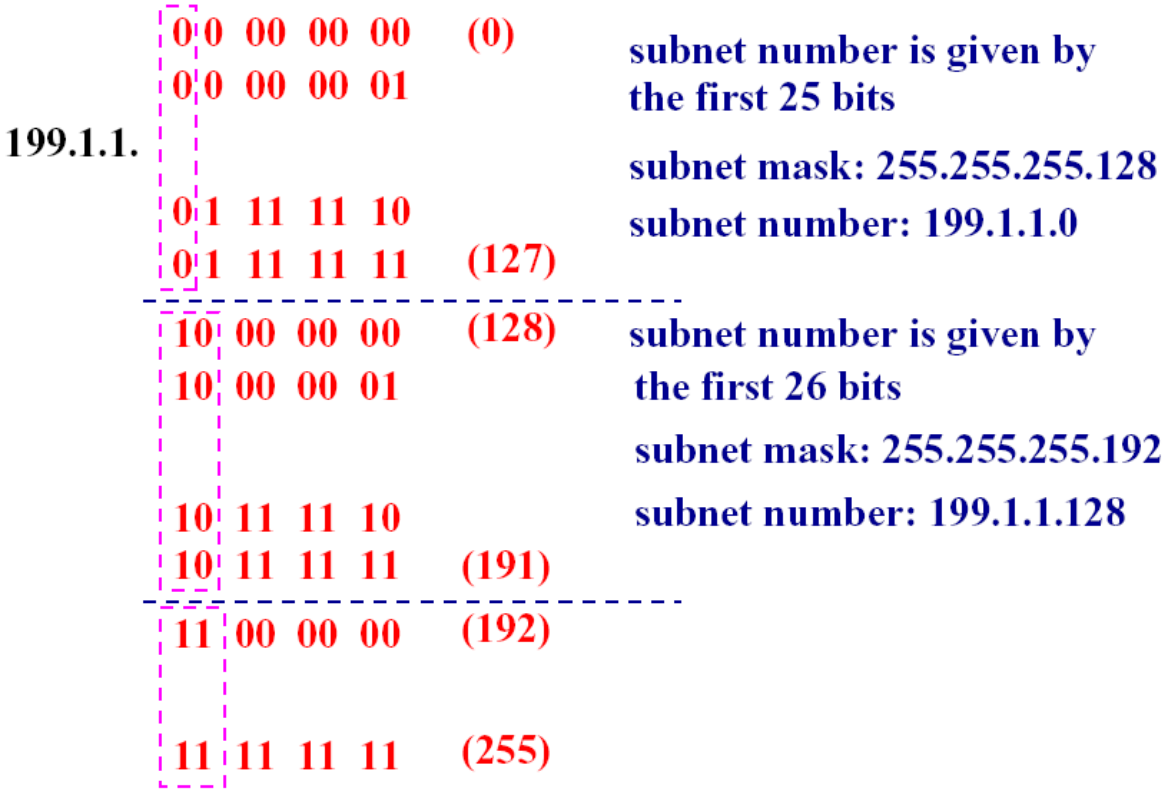
- 示例：给定C类地址块199.1.1.0，将其分为3个子网地址块，其中2个包括64个地址，1个包含128个地址

$$256 = 1 \times 128 + 2 \times 64$$

子网号 (Subnet Number)	子网掩码 (Subnet Mask)	地址范围
192.1.1.0	255.255.255.128	192.1.1.0 ~ 192.1.1.127
192.1.1.128	255.255.255.192	192.1.1.128 ~ 192.1.1.191
192.1.1.192	255.255.255.192	192.1.1.192 ~ 192.1.1.255

- 子网掩码(subnet mask)：确定网络号与主机号的边界
  - 特征：一串连续比特1 + 随后一串连续比特0，如255.255.255.192
- 给定k比特，可用于主机地址的数量为2^k-2，保留下列2个特殊用途地址
  - 主机号为全0：表示整个子网
  - 主机号为全1：表示该子网全体主机，即该网络的广播地址

子网地址划分图解



简捷表达形式
(网络号/网络号长度)

→ 199.1.1.0/25

→ 199.1.1.128/26

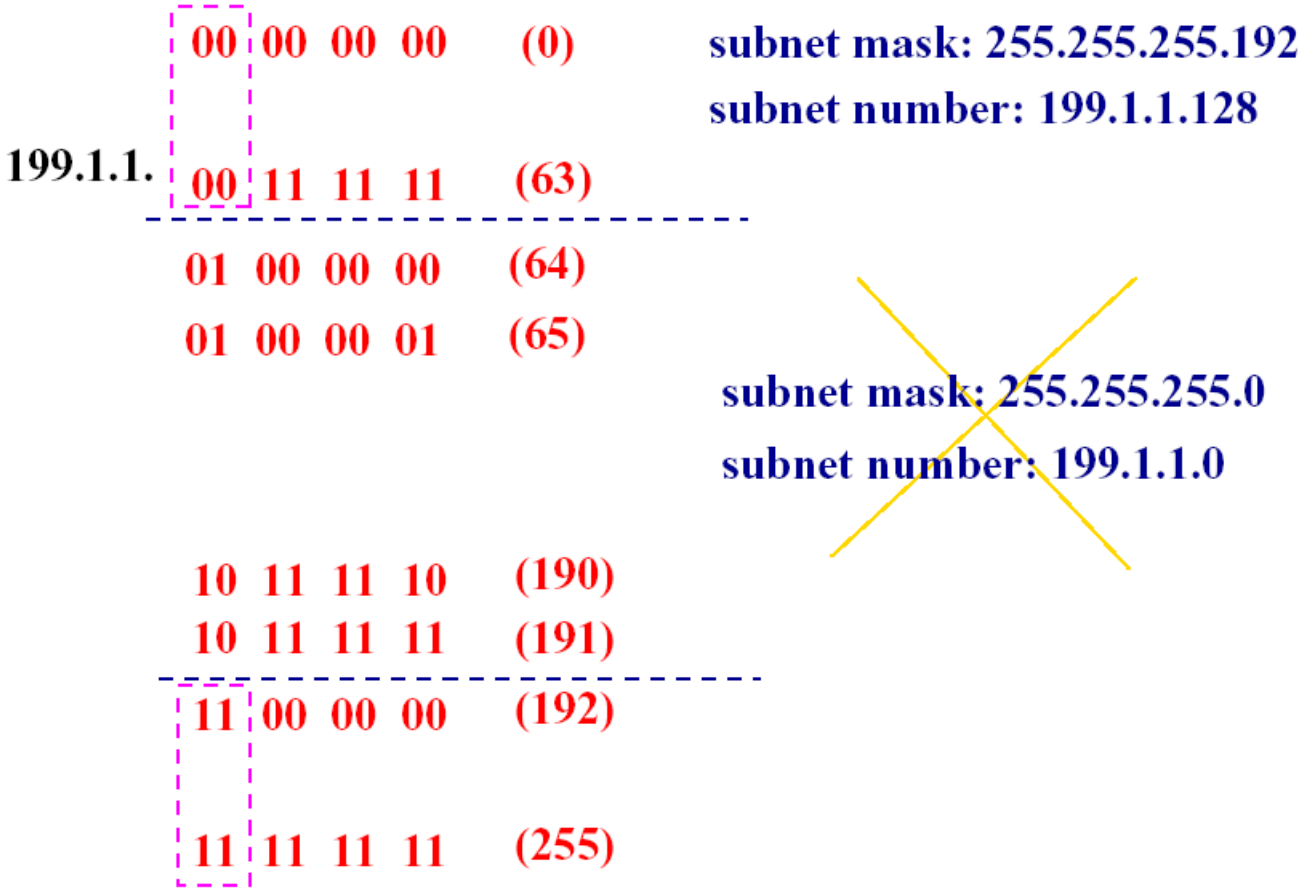
→ 199.1.1.192/26

子网划分原则与方法

- 基本原则
  - <子网号, 子网掩码>唯一标定一个子网的地址空间
  - 不同子网的地址空间不得重叠

- 基本方法：根据子网的大小，从大到小依次划分
  - 给定子网大小，确定能够区别主机号的比特数***x***
    - 例如，14个地址需要4比特
    - 16个地址需要多少比特?
  - 该子网中所有地址的前32-*x*比特必须相同
  - 网络掩码前32-*x*比特全为1，剩下*x*比特全为0

子网划分错误示例


00 00 00 00 (0)
199.1.1.
00 11 11 11 (63)
01 00 00 00 (64)
01 00 00 01 (65)
10 11 11 10 (190)
10 11 11 11 (191)
11 00 00 00 (192)
11 11 11 11 (255)
subnet mask: 255.255.255.192
subnet number: 199.1.1.128
subnet mask: 255.255.255.0
subnet number: 199.1.1.0

CIDR

- 原始IP编址方案为有类别编址(classful addressing)
  - 网络大小只能为256、65k、或17M台主机，颗粒度太粗
    - 例如，包含1000台主机的网络，采用原始IP编址，需要一个B类地址块，效率为1000/65536 ≈ 0.15%!
  - 路由缺乏可扩展性
    - 路由器的路由表中每一条目对应于一个网络
    - 由于A、B、C类地址块的总数量，路由器需要支持很大规模路由表，维护和存储开销都难以承受

- 解决方案：无类别域间路由(Classless Inter-Domain Routing, CIDR)
  - 网络大小颗粒度更精细
    - 给定符合一定条件(见后续)的多个C类地址块用于网络地址，网络大小可以为2^n ($n \geq 8$)，即256, 512, 1024, … (对应于n = 8, 9, 10, …)
    - 比较：有类别编址方式中，C、B、A类地址块分别对应于n = 8, 16, 24
  - 路由聚合(即将路由表多个条目合并为一条) —— 减小路由表规模

CIDR示例

- 1000台主机的网络，分配有4个连续的C类地址块
  199.199.0.0 ~ 199.199.3.0
  - 这些地址块的前22位比特相同

- 采用CIDR，上述4个C类地址块的集合，成为前22比特标定的1个地址块，记为199.199.0.0/22
  - 即网络号为199.199.0.0，网络掩码为255.255.252.0
    (类似于子网划分中的网络号、子网掩码的形式)

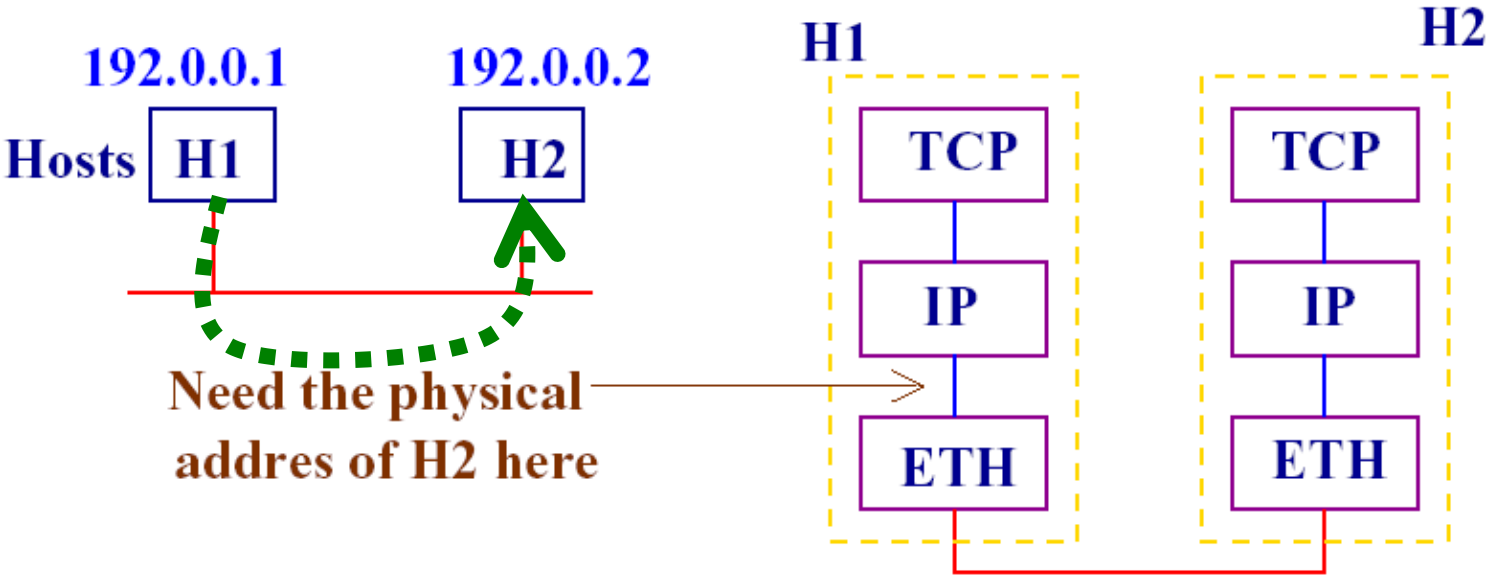
CIDR地址块聚合

- **可聚合**：地址块连续 + 相同前缀(高位比特)覆盖所有地址块，例如199.199.0.0 ~ 199.199.3.0
  - 第3字节二进制形式为000000**00** ~ 000000**11**
  - 等价于1个地址块，即199.199.0.0/22
  - 外部路由节点视其为单个网络，在路由表中仅需维护1条条目
- **不可聚合**：地址块连续 + 相同前缀(高位比特)不能覆盖所有子网，例如199.199.1.0 ~ 199.199.4.0
  - 第3字节二进制形式为00000**001**~ 00000**100**
  - 同时还存在有着相同前缀 00000的其它4个地址块，如00000**000**, 00000**101**，因此不能聚合(用单一地址块表征)!!!
- **不可聚合**：地址块不连续，例如199.199.1.0, 199.200.2.0, 199.210.3.0, 199.234.1.0
  - 不能用单一地址块表征
- 对于地址块不可聚合的多个IP网络，只能视其为多个单独的网络，路由表分别维护相应的条目

内容提要

- 网络互联简介
- IP协议
  - 服务模型，分段与重组，原始编址方案
- 原始IP编址方案的改进
  - 子网划分，CIDR
- 辅助协议
  - ARP，DHCP，ICMP

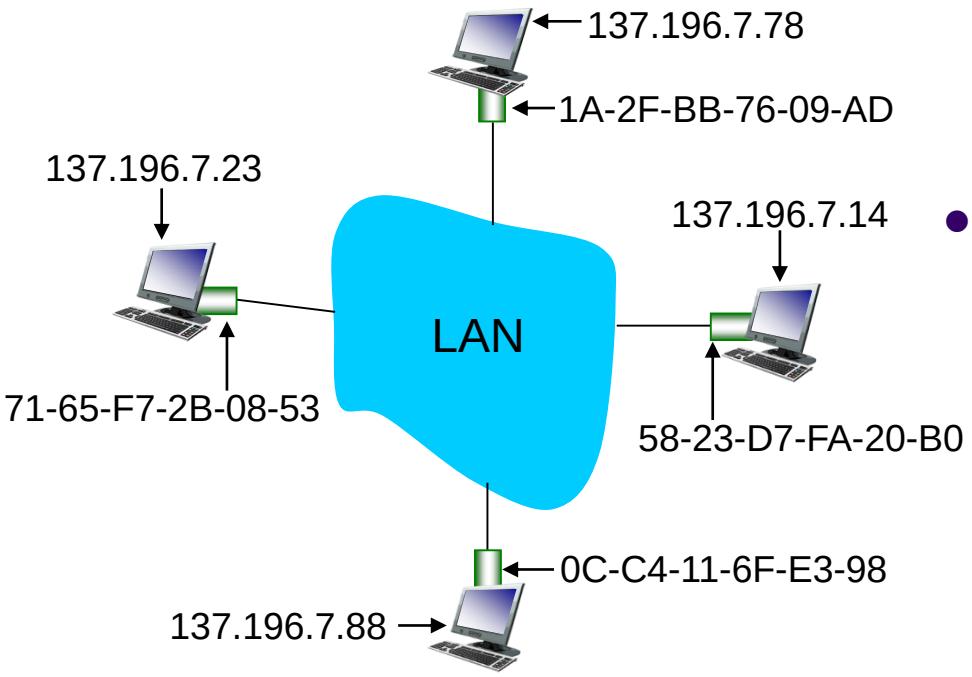
IP数据包传送



IP数据包传送，是通过数据链路层帧封装并传输，
需要知晓下一跳网络接口卡的数据链路层地址

IP与链路层地址的映射

Q: 给定B的IP地址，如何确定B的MAC地址?



- 对于同一个局域网的其它节点，每个IP节点(主机或路由器)维护一个IP地址与MAC地址的映射表
  - 结构：<IP地址; MAC地址; TTL>
  - TTL (Time To Live): 失效时间，典型值为15分钟
- IP地址与MAC地址映射表的维护
  - 初始为空
  - 对于给定IP地址，如映射表中没有对应条目，采用ARP (Address Resolution Protocol)在局域网询问，并将结果添加至映射表中

ARP工作过程

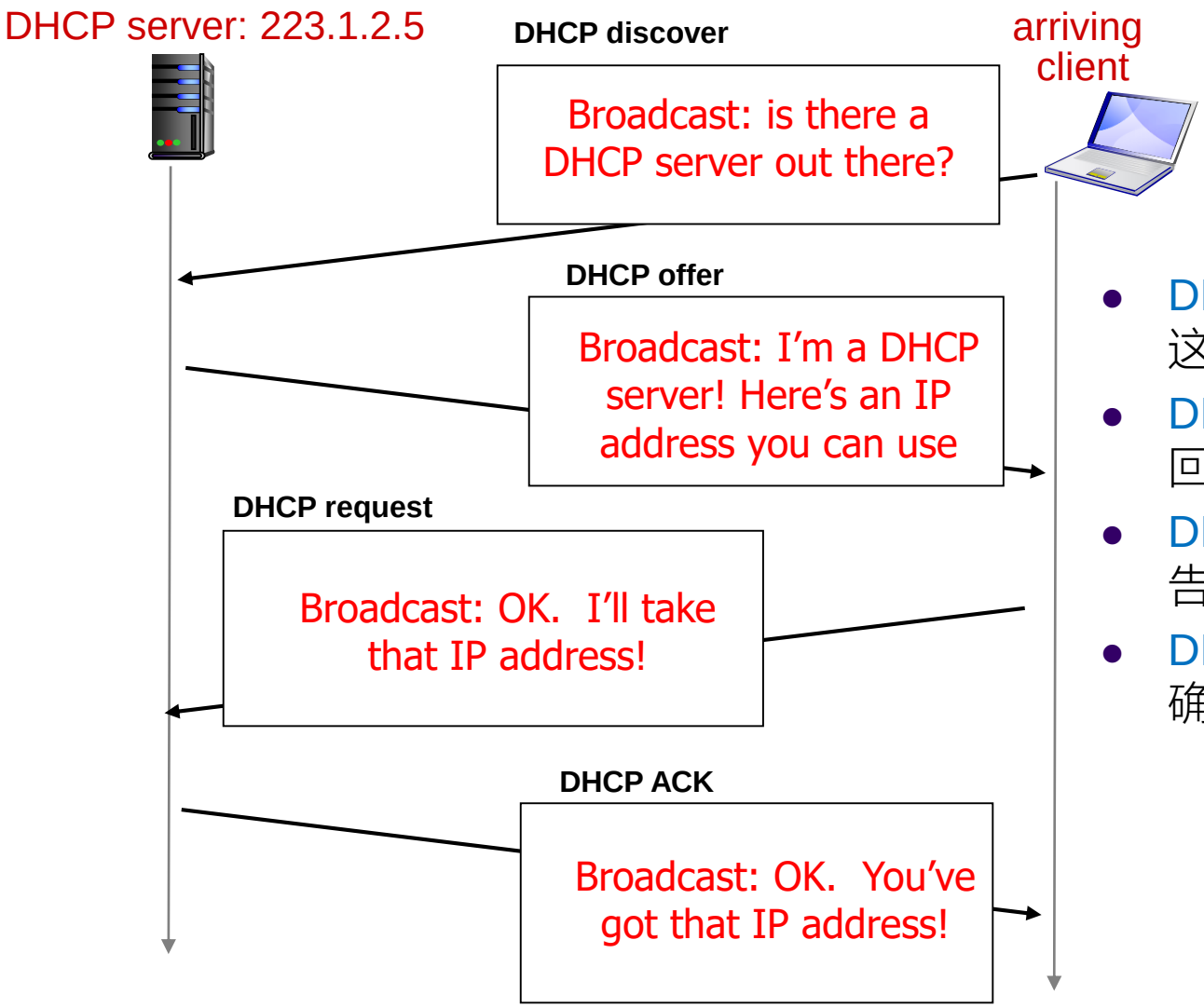
主机X向D (X, D均为IP地址) 发送数据报，但不知道其MAC地址，采用ARP进行相应的地址解析：

- X广播一条关于IP地址D的ARP查询消息
- 收到来自X的广播消息后，D回应告知其MAC地址
  - 借此机会，D同时添加或刷新其映射表中关于X的条目
- 基于D的回复消息，X在其映射表中添加一条关于D的条目
- 由于ARP查询消息发送是采用广播方式，局域网中其它主机也会收到
  - 如果其映射表中已有关于X的条目，刷新之
  - 否则，无操作

IP网络功能配置与DHCP

- 主机的IP网络功能运行需要下列参数
  - IP地址
  - 网络掩码(network mask)
  - 缺省路由器的IP地址
  - DNS服务器的IP地址
- 主机IP网络功能配置：如何获取上述必需信息
  - 静态：管理员手工配置
  - 动态：运行DHCP, BOOTP, RARP协议
- DHCP: Dynamic Host Configuration Protocol
  - 用于主机获取其IP网络功能配置信息
  - 动态地配置主机IP网络功能信息，包括IP地址等
  - 可以减少大型网络的管理工作量
  - 也是移动IP (Mobile IP)必需的补充

DHCP工作过程示例



- DHCPDISCOVER: 主机广播这一消息，以获取配置参数
- DHCPOFFER: DHCP服务器回应消息，提供配置参数
- DHCPREQUEST: 用于主机告知其选择的配置参数
- DHCPPACK: DHCP服务器确认主机的配置参数选择

ICMP

- ICMP (Internet Control Message Protocol)：用于主机/路由器之间的网络层部分信息交换
  - 错误报告：不可到达的主机、网络、协议
  - 回响(echo)请求/回应(用于ping)
- ICMP消息由IP数据包承载
- ICMP消息格式
  - Type
  - Code
  - 产生错误的IP数据包的前8个字节拷贝

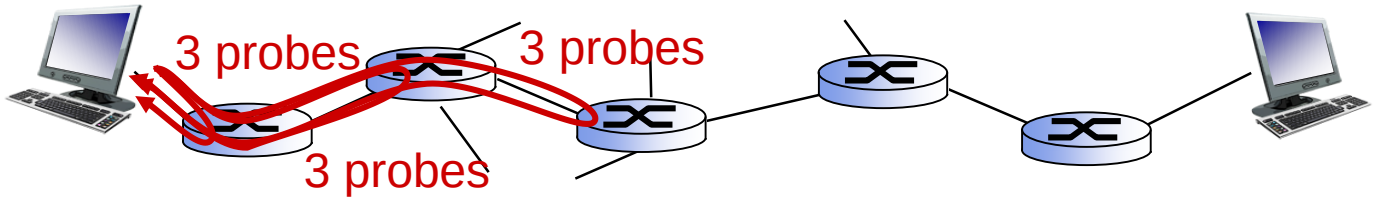
Type	Code	description
0	0	echo reply (ping)
3	0	dst. network unreachable
3	1	dst. host unreachable
3	2	dst. protocol unreachable
3	3	dst. port unreachable
3	6	dst. network unknown
3	7	dst. host unknown
4	0	source quench (congestion control - not used)
8	0	echo request (ping)
9	0	route advertisement
10	0	router discovery
11	0	TTL expired
12	0	bad IP header

ICMP的应用：traceroute

- 源主机发送一系列UDP报文段至目的主机
  - 1st TTL =1, 2nd TTL=2, …
- 当第*n*个数据报到达第*n*个路由器的时侯
  - 由于TTL为0，该路由器丢弃这一数据包，
  - 并回送一条ICMP消息(type 11, code 0)至源主机
  - 这一ICMP消息中包含有该路由器相关信息，包括IP地址等
- 当源主机收到ICMP消息时，计算相应的RTT
- Traceroute重复上述过程3次

停止条件

- 终究会有UDP报文段到达目的主机
- 目的主机返回以含义为“端口不可到达”的ICMP消息(type 3, code 3)
- 当源主机收到这一ICMP消息，停止发送



小结

- 网络互联的概念
- IP基础
  - 服务模型，分段与重组，原始编址，子网划分，CIDR
- IP辅助协议
  - ARP, DHCP, ICMP

- 参考文献
  - 教材3.2.1~3, 3.2.5~8节