

IP编址和报文

理论课程

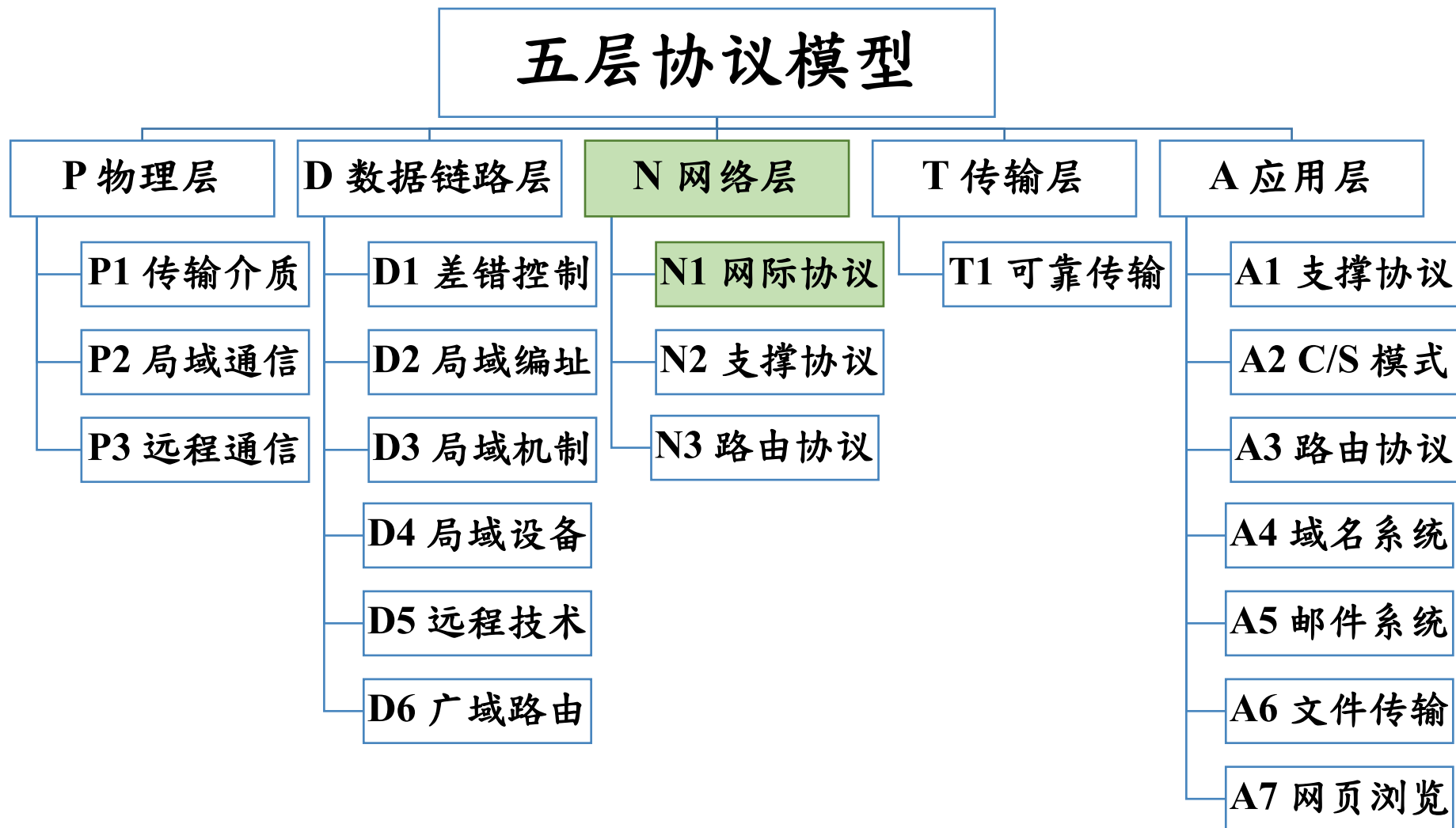


廈門大學
XIAMEN UNIVERSITY



信息学院 黄 焯
(特色化示范性软件学院) 博士, 副教授
School of Informatics Wei Huang

知识框架



主要内容

- 世界互联的概念和结构、虚拟网络
- IP（互联网协议）
 - IP编址方案、点分十进制表示法、CIDR表示法
 - 有类地址、无类地址
 - 子网划分（原则）、子网掩码
 - 特殊IP地址
- IP数据报的报头格式：组成和作用（不要求顺序）

主要内容

- IP数据报和数据报转发
 - IP路由表和路由转发的原理
- IP封装、分段与重组
 - IP封装：报文跨互联网传输时的数据链路层行为
 - MTU、IP分段与重组

对应课本章节

- **PART IV Internetworking**
 - **Chapter 20 Internetworking: Concepts, Architecture, and Protocols**
 - **Chapter 21 IP: Internet Addressing**
 - **Chapter 22 Datagram Forwarding**

内容纲要

1	互 联 网 络
2	IP地址
3	子网划分
4	IP数据报格式
5	IP数据报转发

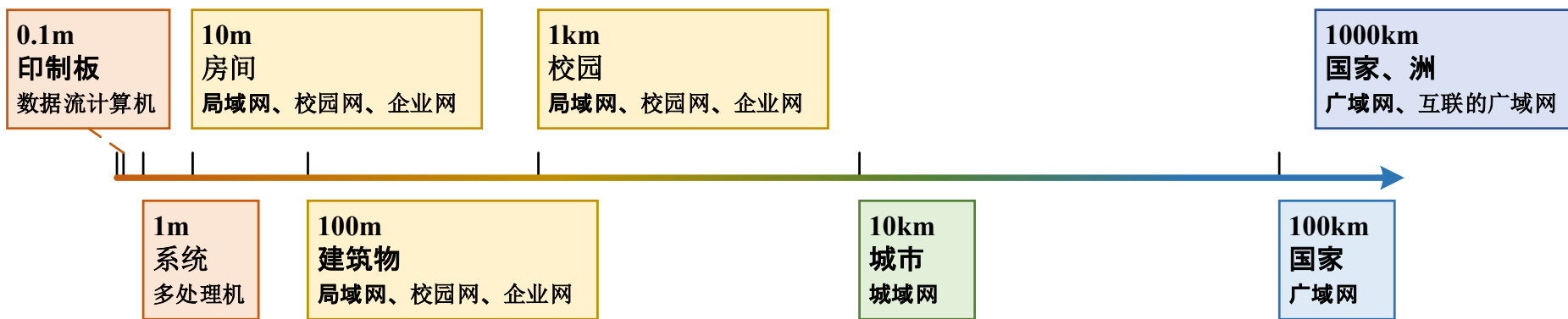
大网络和广域

- 网络技术分为三大类

- 局域网 (LAN) 、城域网 (MAN) 、广域网 (WAN)

- 广域网遇到的新问题

局域网	广域网
距离近，通信随机发生 主机少，设备性能接近	距离远，不确定因素多 主机多，设备性能各异



世界互联

- 困难

- 底层网络（机制、帧格式）、主机性能各异

- 多个网络技术、统一网络服务

- 已有网络互不兼容，不可能仅用导线连接不同类型的网络

- 从硬件上，使用路由器

- 转换不同协议格式的设备
 - 在复杂网络中维护连通性
信息、选择路径

- 从软件上，使用高层的协议格式和地址

英文	中文
internet/internetwork	互联网络
internetworking	网络互联
Internet	因特网

路由

- 路由 (Routing)

- 为单网络中、多网络间或跨多网络的流量选择路径的过程。
 - 两个局域网；局域网和广域网；或两个广域网
- 路由在许多类型的网络中执行，狭义上指IP路由。

- 路由器 (Router)

- 路由器是在计算机网络之间转发数据包的网络设备。
- RFC文档中，路由器和网关 (Gateway) 是同一个概念。
- 物理结构：处理器、内存，以及所连接网络的单独I/O接口

虚拟网络

- 互联网提供了表面上单一的无缝通信系统
- 硬件和软件的结合提供了一个统一的网络系统的错觉
 - 互联网软件隐藏细节物理网络连接物理地址路由信息
 - 应用不需要知道底层物理硬件或路由器的存在

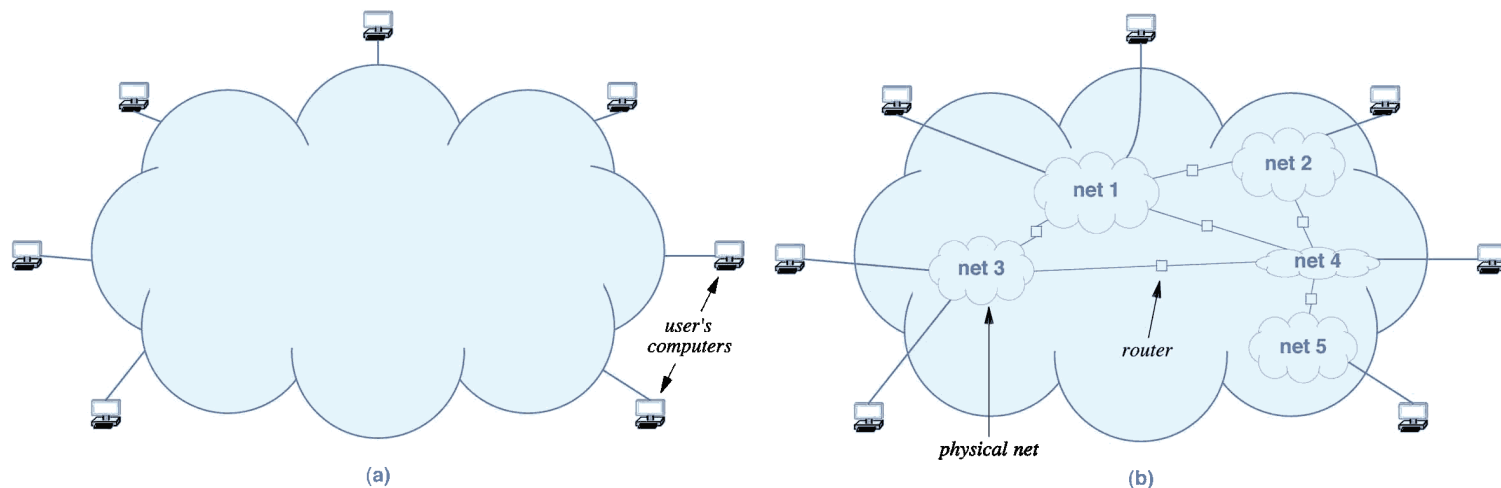


Figure 20.3 The Internet concept. (a) The illusion of a single network provided to users and applications, and (b) the underlying physical structure with routers interconnecting networks.

Internet协议 (IP)

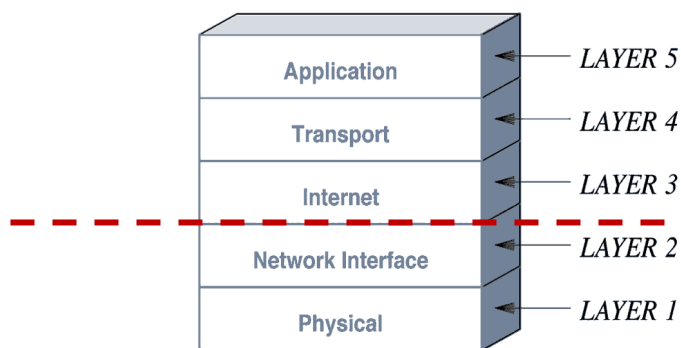
- Internet协议 (Internet Protocol)

- 协议族中用于跨网络边界中继数据报的主要通信协议。

- 版本

- 主要版本：IPv4 (默认)

- 继任版本：IPv6



0 实验版本	1 实验版本	2 实验版本	3 实验版本
4 主要版本	5 历史版本	6 继任版本	7
8 历史版本	9 愚人节笑话	10	11
12	13	14	15 保留版本

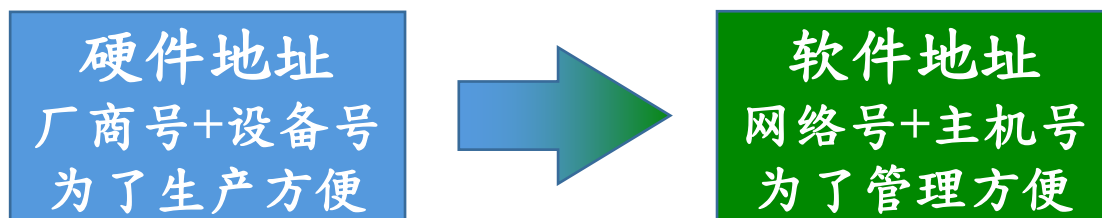
Figure 1.1 The layering model used with the Internet protocols (TCP/IP).

内容纲要

1	互联网络
2	IP地址
3	子网划分
4	IP数据报格式
5	IP数据报转发

虚拟互联网的地址：从硬件到软件

- 因特网是由设计师想象的完全由协议软件实现的
 - 物理层：异构网络的编址各异、非结构化
 - 软件层：需要一个新的编址来隐藏异构的物理细节
- 主机太多使得广域网需要分层寻址 (*net, host_in_net*)
 - 网络号：标识主机附加的物理网络，全球协调
 - 主机号：标识网络上的特定计算机，局域网内协调



IP地址（IPv4地址）

- Internet协议地址（IP地址）
 - 分配给连接到使用IP协议进行通信的每个设备的数字标签。
- 主要功能：主机或网络接口标识；位置寻址
- 所有权：ICANN是处理地址分配和裁决争端的机构
 - ICANN：互联网名称与数字地址分配公司
 - ICANN授权注册商分配个人前缀，ISP向用户提供地址
- 组成：网络号+主机号
- 表示：点分十进制记数法
 - 每8位作为无符号十进制值（0~255）并用点分隔

二进制：11000000 00000101 00110000 00000011

十进制：192 . 5 . 48 . 3

内容纲要

1	互联网络
2	互联网协议地址
3	子网划分
4	IP数据报格式
5	IP数据报转发

地址掩码 (Address Masks)

- 地址掩码指示IP地址中主机所在子网地址的位掩码
- 组成：连续的 N 位1在前，紧接着连续的 $32-N$ 位0
- 作用：求取网络地址（网络号）

$$N = D \& M$$

名称	变量	二进制表示				点分十进制
目标地址	D	10000000	00001010	00000010	00000011	128.10.2.3
网络掩码	M	11111111	11111111	00000000	00000000	255.255.0.0
网络地址	N	10000000	00001010	00000000	00000000	128.10.0.0

特殊的IP 地址

- 主机号全0代表网络，主机号全1代表广播
- 一些IP地址是保留的，不分配给主机

名词	英文名词	规则	示例
网络地址	Network address	主机号全0	59.0.0.0
直接广播地址	Directed Broadcast Address	主机号全1	59.255.255.255
有限广播地址	Limited Broadcast Address	地址全1	255.255.255.255
本机地址	This Computer address	全0	0.0.0.0
回送地址	Loopback address	127.0.0.0/8	127.0.0.1

IP地址编址方法

- 对于网络号与主机号比例关系，先后出现3种方法
 - 分类IP地址（1981~1993）
 - 根据最前的位内容分5类确定网络号长度
 - 子网划分（1985~）
 - 将主机号的部分位用于表示子网号，对分类地址方法的改进。
 - 无分类IP地址（1993~）
 - 灵活调整网络大小。

传统分类地址：ABCDE类

- 机制：根据IP地址的最前位确定网络掩码
- 存续时间：1981-1993

类	位前缀	网络位	主机位	网络容量	网络中主机容量	起始地址	结束地址	子网掩码
A	0	8	24	128 (2^7)	16,777,216 (2^{24})	0.0.0.0	127.255.255.255	255.0.0.0
B	10	16	16	16,384 (2^{14})	65,536 (2^{16})	128.0.0.0	191.255.255.255	255.255.0.0
C	110	24	8	2,097,152 (2^{21})	256 (2^8)	192.0.0.0	223.255.255.255	255.255.255.0
D	1110			多播地址		224.0.0.0	239.255.255.255	无
E	1111			保留地址		240.0.0.0	255.255.255.255	无

无类域间路由表示法

- 作用

- 分配地址与路由汇总时表示地址块

- 形式：IP地址/网络号长度 $ddd.ddd.ddd.ddd/m$

- 根据IP地址和网络掩码（网络号长度）确定网络地址

- m 为掩码中1的个数（不一定是8的倍数）

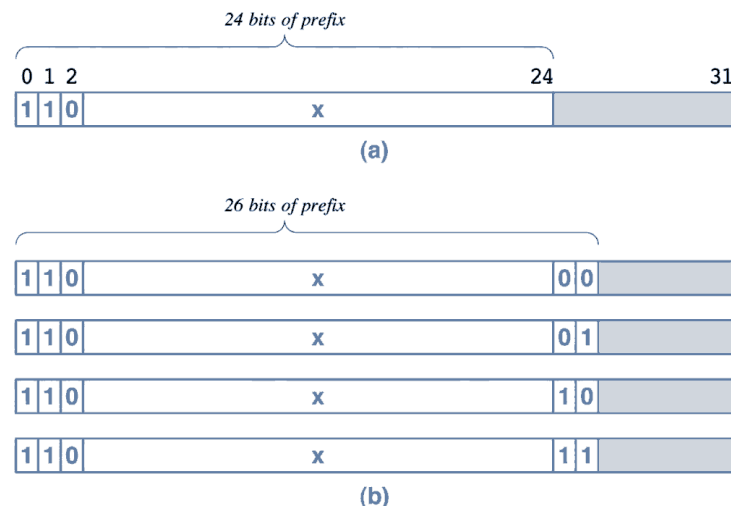
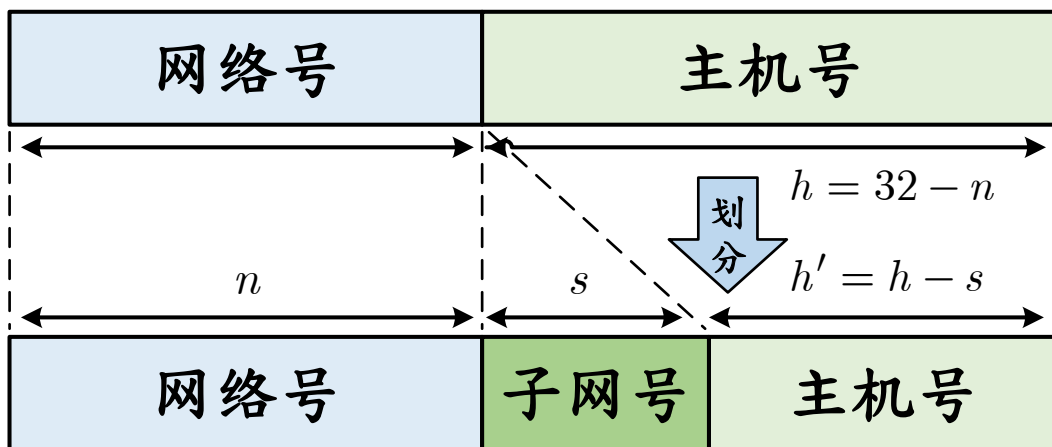
无类域间路由	CIDR
Classless Inter-Domain Routing	

0	Network Prefix 128.211.0.16 /28																								28	31	
1 0 0 0 0 0 0 0								1 1 0 1 0 0 1 1								0 0 0 0 0 0 0 0								0 0 0 1			
0	Address Mask 255.255.255.240																								28	31	
1 1 1 1 1 1 1 1								1 1 1 1 1 1 1 1								1 1 1 1 1 1 1 1								1 1 1 1			
0	Lowest Host Address 128.211.0.17																								28	31	
1 0 0 0 0 0 0 0								1 1 0 1 0 0 1 1								0 0 0 0 0 0 0 0								0 0 0 1			
0	Highest Host Address 128.211.0.30																								28	31	
1 0 0 0 0 0 0 0								1 1 0 1 0 0 1 1								0 0 0 0 0 0 0 0								0 0 0 1			
																								1 1 1 0			

Figure 21.6 Illustration of CIDR addressing for an example /28 prefix.

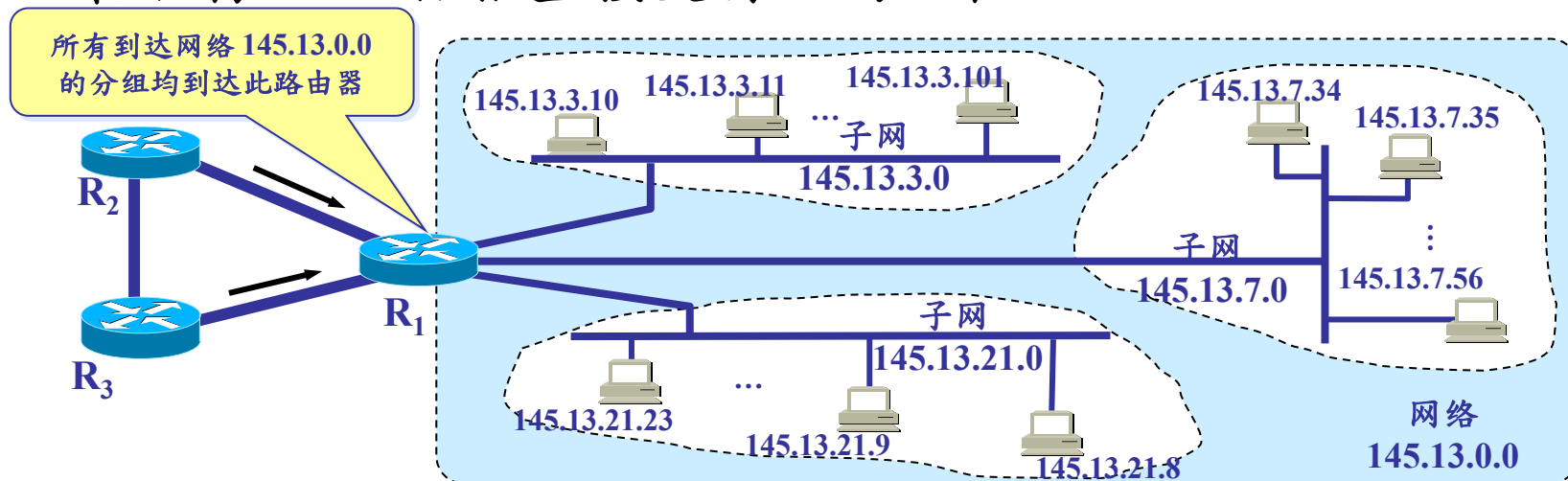
子网划分

- 动机：ARPANET早期，IP地址的设计确实不够合理。
 - IP地址空间利用率有时很低。
 - 每个物理网络分配一个网络号，会使路由表太大性能变坏
- 划分方法
 - 从主机号借若干位作为子网号，两级IP地址变为三级地址



划分子网的基本思路

- 单位对内划分子网，对外仍表现为统一的网络。
 - 从IP数据报无法判断源或目的主机所在网络是否子网划分。
 - 外网发来的数据报，根据目的网络号找到本单位的路由器。
 - 本单位路由器收到数据报后，提取子网号找到目的子网。
 - 最后将IP数据报直接交付目的主机。



子网掩码 (subnet mask)

- 分组转发算法必须做相应的改动
 - 路由器在和相邻路由器交换路由信息时，必须把自己所在网络（或子网）的子网掩码告诉相邻路由器。
 - 路由器的路由表中的每一个项目，除了要给出目的网络地址外，还必须同时给出该网络的子网掩码。
 - 连接两个子网的路由器拥有两个网络地址和两个子网掩码。
- 解决方案：使用子网掩码替代网络掩码。
- 子网划分应剔除主机号或网络号全0全1两种情形
 - RFC 1878 废止：现代软件将能够利用所有可定义的网络。

多穴主机 (Multi-Homed Hosts)

- IP地址标识计算机和网络之间的连接
 - IP地址不标识特定计算机
- 多穴主机
 - 多个网络连接的计算机，它为每个连接分配一个IP地址
 - 作用：用于提高可靠性
 - 如果一个网络失效，主机仍可以通过第二连接到达互联网
 - 连接到多个网络可以直接发送流量，避开拥塞的路由器

内容纲要

1	互联网络
2	IP地址
3	子网划分
4	IP数据报格式
5	IP数据报转发

虚拟分组 (Virtual Packet)

- 网络互联协议定义独立于底层硬件的“分组”格式
 - 为解决异构问题，其结果是一个通用的，虚拟的数据包。
- 底层硬件不理解，路由器和主机（软件）理解
 - 底层硬件不理解或不识别Internet包格式。
 - 因特网中每个主机或路由器都包含理解因特网数据包的协议软件。

IP数据报 (IP Datagram)

- IP数据报 (datagram) 是Internet数据分组的术语。
 - 通用的虚拟的分组独立于底层硬件
- IPv4数据报总长最多64KB，头部至少20B。

0	4	8	16	19	24	31
VERS	H. LEN	SERVICE TYPE	TOTAL LENGTH			
IDENTIFICATION			FLAGS	FRAGMENT OFFSET		
TIME TO LIVE		TYPE	HEADER CHECKSUM			
SOURCE IP ADDRESS						
DESTINATION IP ADDRESS						
IP OPTIONS (MAY BE OMITTED)					PADDING	
BEGINNING OF PAYLOAD (DATA BEING SENT)						
⋮						

IP报文头格式

- IP报文头格式的组成（基本长度：20B）
 - 版本：4bits，取值：4或6
 - 报头长度：4bits，单位为4Bytes
 - 服务类型：8bits，未实际使用
 - 报文总长度：16bits，单位为字节
 - 标识：16bits，IP软件在存储器中的计数器在产生一个数据报后自增1，并将值赋给标识字段。标识在分片时复制。
 - 分片标志：3bits，高到低位：无意义、不分片、还有分片
 - 片偏移：13bits，分片在原始报文的位置，单位为8Bytes

IP报文头格式

- IP报文头格式的组成（基本长度：20B）
 - 生存时间（TTL）：8bits，单位为秒，路由器减去在其环节所消耗时间，直至零丢弃。
 - 协议类型：8bits，可能的取值有：ICMP、IGMP、TCP、UDP、OSPF等，用于将数据交给第四层的哪个软件。
 - 报头校验和：16bits，检验报头的完整性，不含数据部分。
 - 源IP地址：32bits
 - 目标IP地址：32bits
 - 选项内容：1~40bits，用来支持排错、测量和安全等措施。
 - 填充部分：长度可变，为了使报文头部是4Bytes的整数倍。

内容纲要

	2	IP地址
	3	子网划分
	4	IP数据报格式
	5	IP数据报转发
	6	IP封装、分片与重组

路由表

- 作用

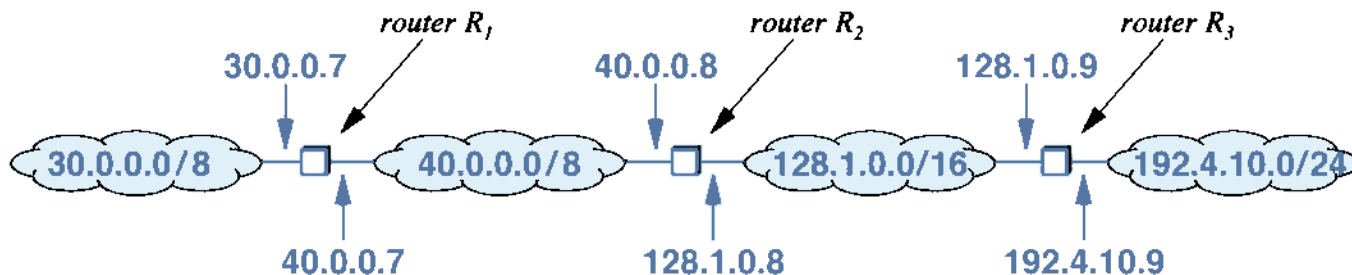
- 通过查表，将数据报转发到下一跳，从而实现主机互联
- 同一网络中不同路由器因视角不同，路由表不同

- 组成：目标网络号、子网掩码、下一跳

- 目标网络号：主机位全0；
 - 默认路由：网络号和子网号全0，只有1项
- 子网掩码：网络位全1，主机位全0；
- 下一跳
 - 如果与目标网络号在同一子网，则直接发送
 - 如果不在同一子网，则填写与该路由器同一子网的路由器IP地址

路由表

- 请画出该网络各路由器的路由表。



R₁

目的	掩码	下一跳
30.0.0.0	255.0.0.0	直接转发
40.0.0.0	255.0.0.0	直接转发
0.0.0.0	0.0.0.0	40.0.0.8

R₃

目的	掩码	下一跳
128.1.0.0	255.255.0.0	直接转发
192.4.10.0	255.255.255.0	直接转发
0.0.0.0	0.0.0.0	128.1.0.8

R₂

目的	掩码	下一跳
30.0.0.0	255.0.0.0	40.0.0.7
40.0.0.0	255.0.0.0	直接转发
128.1.0.0	255.255.0.0	直接转发
192.4.10.0	255.255.255.0	128.1.0.9

路由转发

- 通过子网掩码进行计算的路由表匹配
 - 获得IP报文的目标IP地址D
 - 用D顺序逐条匹配路由表各个条目 $T[0]$, $T[1]$, $T[2]$
 - 如果 $D \& T[i].m == T[i].d$, 则下一跳为 $T[i].n$
 - D : 目标端地址
 - $T[i]$: 路由表中第 i 条目标
 - d 子网网络号 ; m : 子网掩码 ; n : 下一跳IP地址
- 最长前缀匹配 (Longest Prefix Match)
 - 设路由表有以下两个网络前缀 : 128.10.0.0/16; 128.10.2.0/24
 - 对于报文的IP地址128.10.2.3 , 选择128.10.2.0/24

内容纲要

	2	IP地址
	3	子网划分
	4	IP数据报格式
	5	IP数据报转发
	6	IP封装、分片与重组

尽力而为的传输

- IP要适应不同硬件的需要，但底层硬件可能不起作用
- IP提供一种尽力而为的传输 (Best-Effort Delivery)
 - IP数据报可能会丢失、重复、延迟、乱序，或数据损坏。
 - 需要高层协议软件来处理上述每一个错误。

数据报传输与帧

- 封装 (Encapsulation)

- IP软件通过路由表选择下一跳，并通过物理网络发送

- 底层网络硬件不理解数据报文格式和因特网地址
 - 底层通过自己的方式发送。

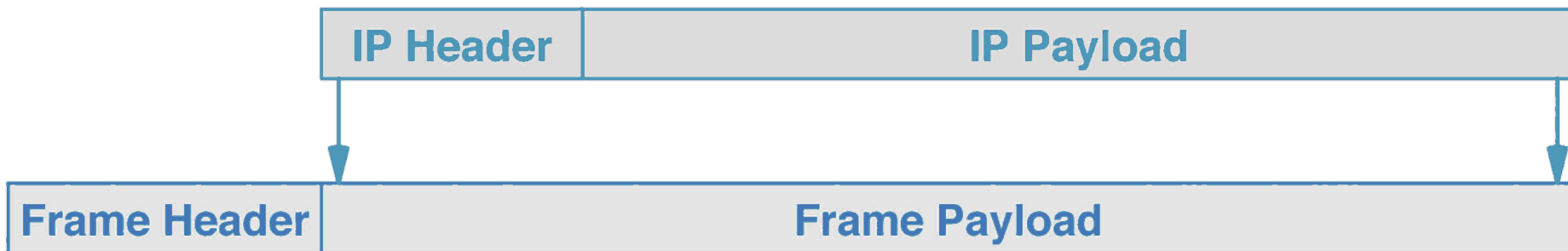
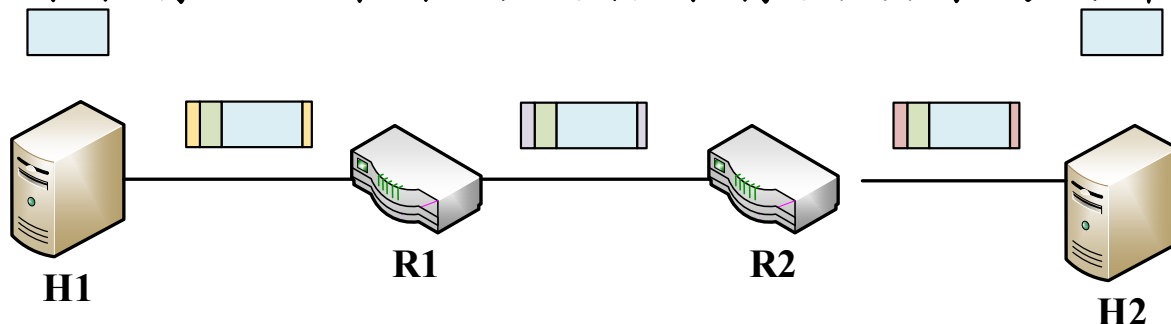


Figure 22.4 Illustration of an IP datagram encapsulated in a frame.

跨互联网传输

- 帧到达下一跳，接收方软件提取IP数据报并丢弃帧头
 - 若还需转发，则再封装
 - 帧头部不积累，主机和路由器不存储额外的头部



项目	值 (H1-R1)	值 (R1-R2)	值 (R2-H2)
源MAC	MAC(H1)	MAC(R1,R)	MAC(R2,R)
目的MAC	MAC(R1,L)	MAC(R2,L)	MAC(H2)
源IP	IP(H1)	IP(H1)	IP(H1)
目的IP	IP(H2)	IP(H2)	IP(H2)

最大传输单元 (MTU)

- 最大传输单元：数据链路层帧支持的最大传输字节数
- 当前链路MTU小于IP报文长时，分成较小分片传输
 - 路由器将数据报分成更小的碎片称为分片。
 - 原始数据报首部被复制为各数据报片的首部，但必须修改有关字段的值。
 - 每个分片以IP数据报格式独立发送。

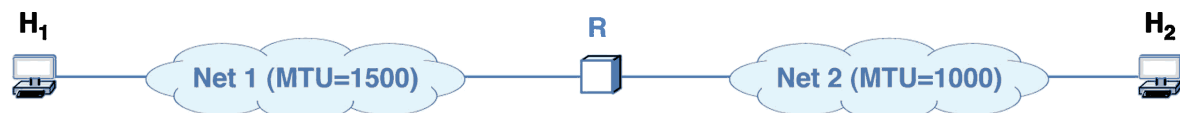


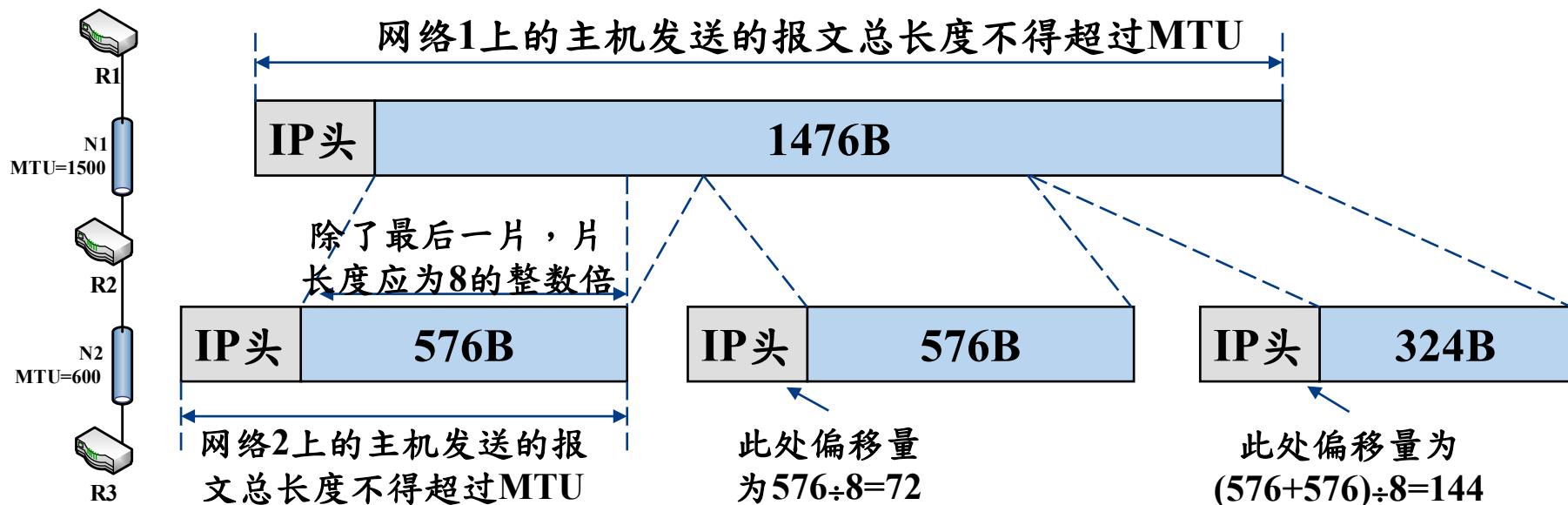
Figure 22.6 Illustration of a router that connects two networks with different MTUs.

Protocol	MTU
Token ring (16Mbps)	17914
Token ring (4Mbps)	4464
FDDI	4352
Ethernet	1500
X.25	576
PPP	296

IP报文的分片

• IP报文传输的分片原则

- 各片尽可能大，尽量少分片，但每片不能超过MTU
- IP头部固有长度（基本长度20字节），也在帧的载荷内
- 分片应使得后续片偏移量为8的整数倍



IP报文的分片

- IP报文中的信息表示

- 报文ID：分片时复制，始终保持初始ID不变
- 标志位：若第一次分片，则修改“是否分片”相应位
- 片偏移量：当前分片在初始IP报数据的偏移（单位：8字节）
 - 标志MF：1表示后面还有分片；0表示这是最后一个分片；
 - 标志DF：1表示不允许分片；0表示允许分片。

- 报文重组策略

- 源端到目标端数据传输过程中可能有多次分片
- 所有分片重组在目标端进行，中间路由可再分片但不重组
 - 减少中间节点的数据处理过程

IP报文的分片策略

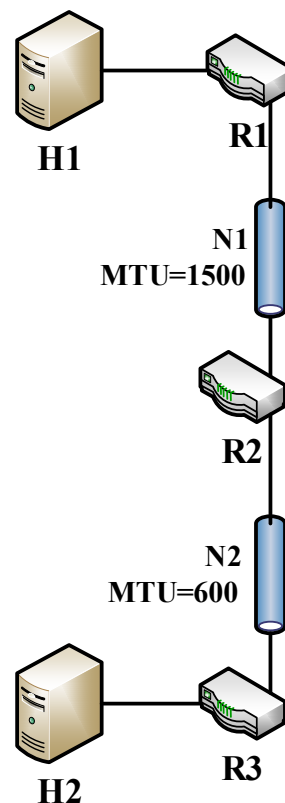
- 例题：某主机H1上的IP数据报文长度为3820字节（含固定报文头20字节），发送给主机H2。途经网络N1、N2的MTU如右图所示，请写出其在N1和N2上的分片情况，DF=0。

• 解：

— R1对报文的分片

MTU1=1500B		
IP头部 20B	IP数据 1480B	余数 0B

- 分片的数据总长度应为 $3820\text{B} - \text{头部}20\text{B} = 3800\text{B}$
- 每片数据长度不超过 $\text{MTU} - \text{头部}20\text{B} = 1480\text{B}$
- 除最后一片其余各片应为8B的整数倍



IP报文的分片策略

– R1对报文的分片结果

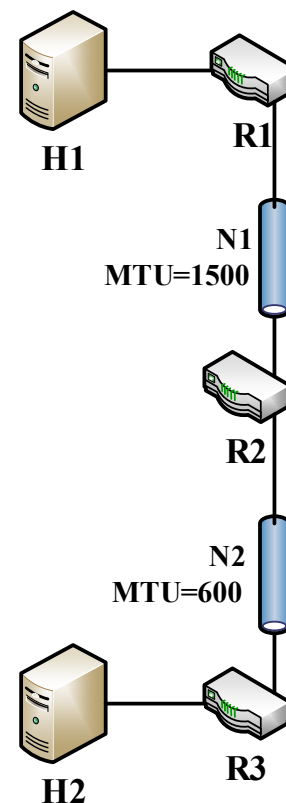
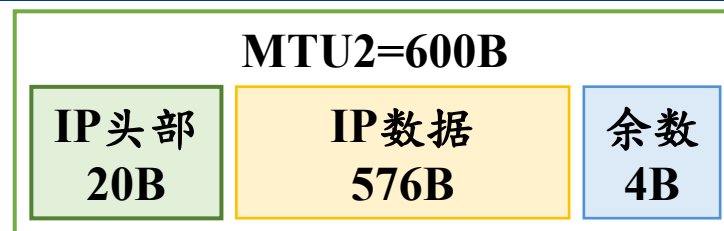
- 20+1480, 20+1480, 20+840

– R2对报文的分片情况

- 分片不重组, 分片再分片
- 每片数据长度不超过MTU - 头部20B = 580B
- 除最后一片其余各片应为8B的整数倍

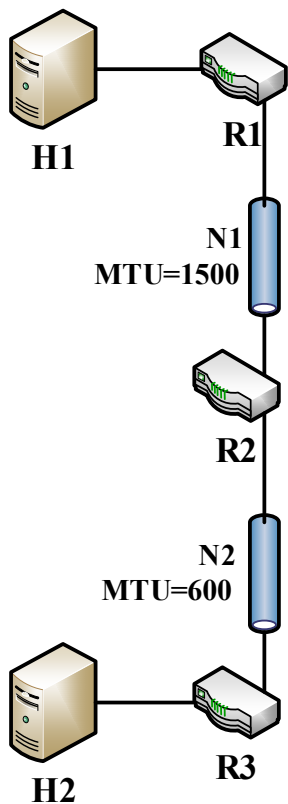
– R2对报文的分片结果

- 20+576, 20+576, 20+328, 20+576, 20+576, 20+328, 20+576, 20+264



IP报文的分片策略

• 答案



类别	序号	总长度	数据长度	MF	片偏移
原数据报	1	3820	3800	0	0
R1发给R2 (在N1) 的数据报	1	1500	1480	1	0
	2	1500	1480	1	185
	3	860	840	0	370
R2发给R3 (在N2) 的数据报	1	596	576	1	0
	2	596	576	1	72
	3	348	328	1	144
	4	596	576	1	185
	5	596	576	1	257
	6	348	328	1	329
	7	596	576	1	370
	8	284	264	0	442

IP报文丢失问题

• IP报文丢失判断

- 时机：目标主机在IP报文分片重组时进行丢失判断
- 对于源主机发出的每个报文，在收到第一个分片的时候，给出一个等待的有限时间T-out；
- 如果T-out之后还没有收到全部分片，则为超时。
- 任何一个分片丢失（超时）或出错，则丢弃整个报文。
 - 不能只丢弃分片，是因为报文重发不一定经过原路径

谢谢观看



廈門大學
XIAMEN UNIVERSITY



信息学院 黄 焯
(特色化示范性软件学院) 博士, 副教授
School of Informatics Wei Huang