Computer Networks
计算机网络

Classification of Networks

- **by scale**

  LAN, WAN and MAN

- **by topology**

  Bus, Star, Ring and Tree

- **by switching approach**

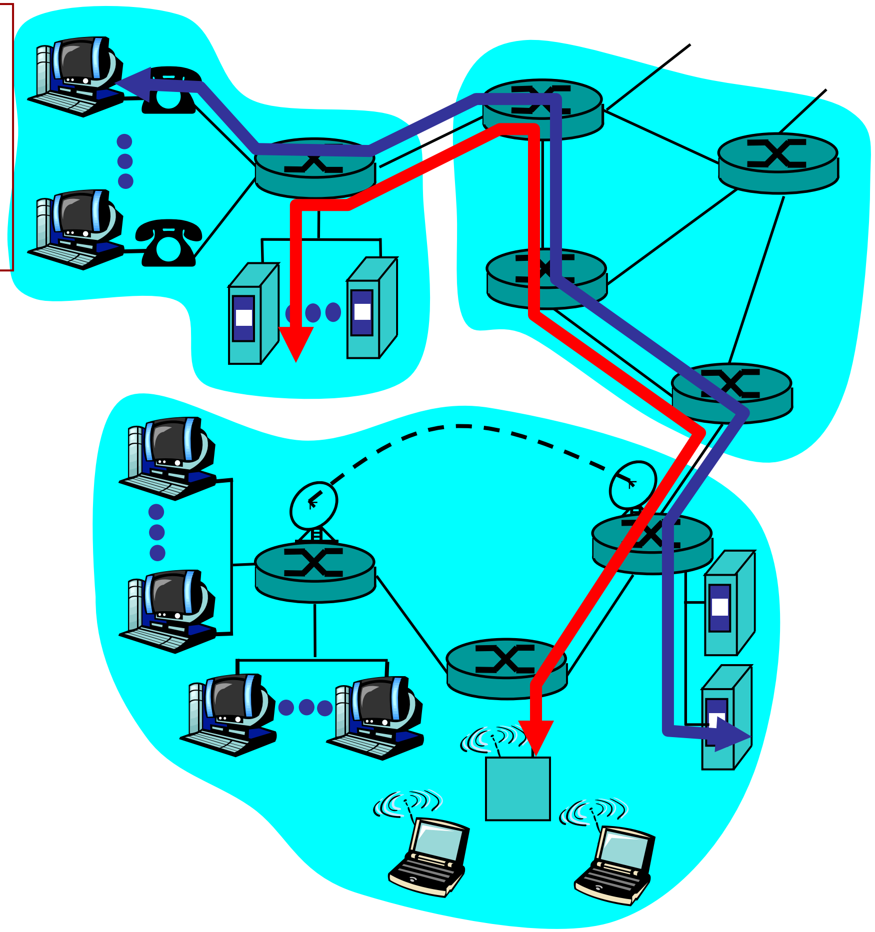
  Circuit switching and Packet switching

- **by transmission media**

  Wireless network and Wired network

Circuit Switching电路交换

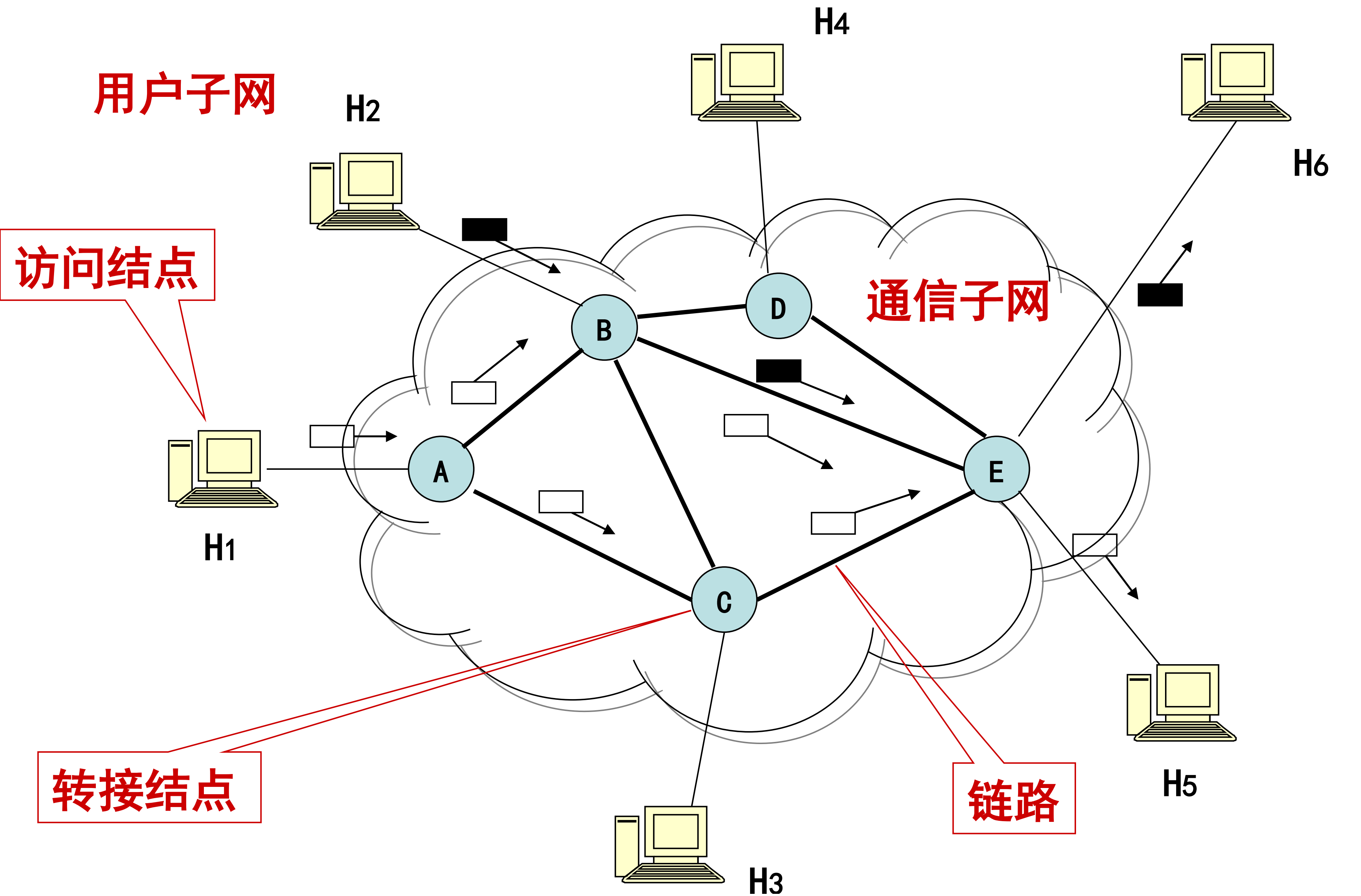
End-to-end resources reserved for "call"



- Link bandwidth, switch capacity
- Dedicated resources with no sharing
- Guaranteed transmission capacity
- Call setup required

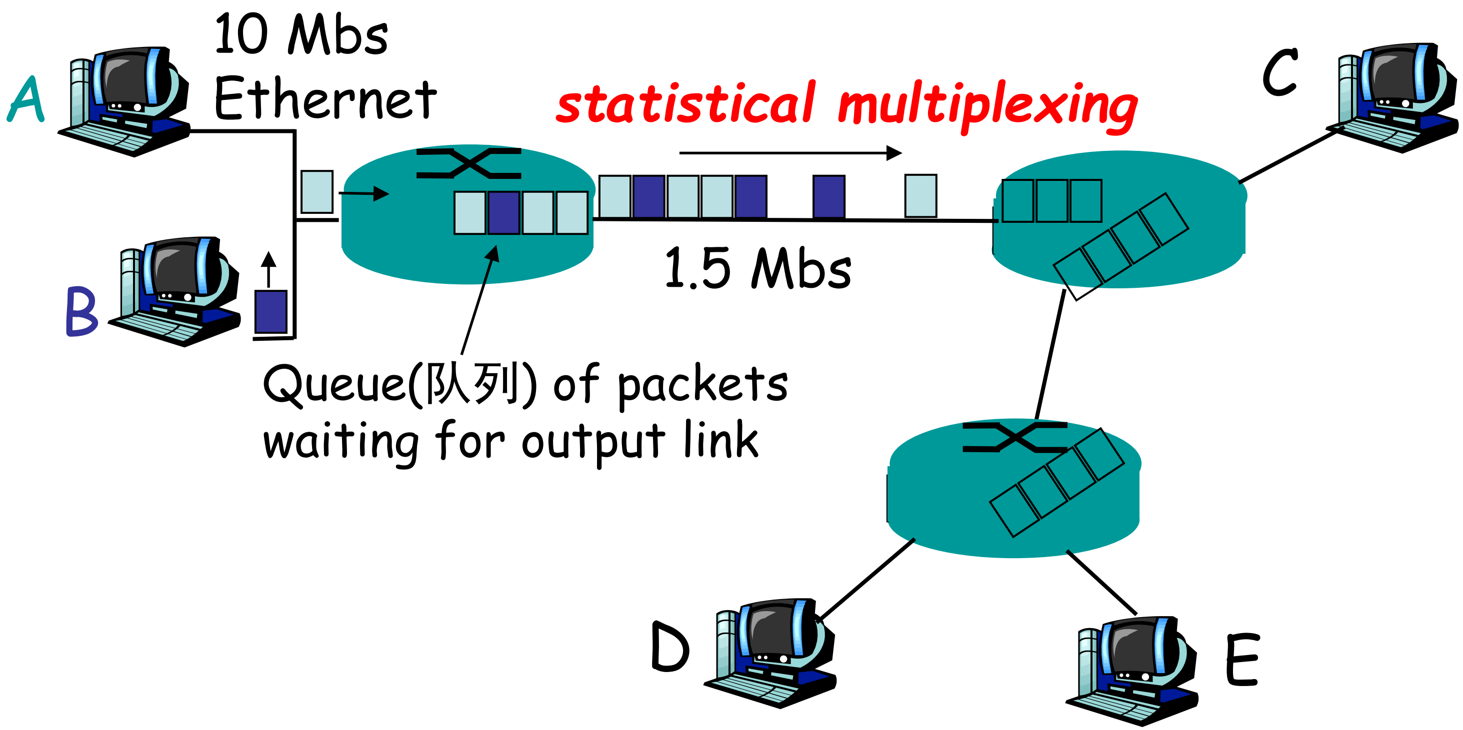
Packet Switching分组交换

- **source breaks long messages into smaller "packets(数据包，数据分组)"**
- **packets *share* network resources**
- **each packet uses full link bandwidth**
- **"store-and-forward(存储转发)" transmission**



分组交换网示意图

Packet Switching: Statistical Multiplexing



Sequence of A & B packets does not have fixed pattern ***statistical multiplexing统计多路复用***.

Four sources of packet delay

1. **d_{proc} = processing delay(处理时延)**
   typically a few microsecs or less
2. **d_{queue} = queuing delay(排队时延)**
   depends on congestion
3. **d_{trans} = transmission delay(发送时延)**
   = L/R, significant for low-speed links
4. **d_{prop} = propagation delay(传播时延)**
   a few microsecs to hundreds of msecs

Nodal delay

$$\boxed{d_{\text{nodal}} = d_{\text{proc}} + d_{\text{queue}} + d_{\text{trans}} + d_{\text{prop}}}$$

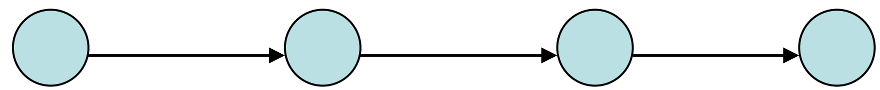
总时延=处理时延+排队时延+发送时延+传播时延

例题：Assignments 1 （作业1）

- 补充题：试在下列条件下比较电路交换和分组交换。要传送的报文共 x (bit)，从源站到目的站共经过 k 段链路，每段链路的传播时延为d 秒，数据率为 b (bit/s)。在电路交换时电路的建立时间为 s 秒。在分组交换时分组长度为 p (bit)，且各结点的排队等待时间可忽略不计。问在怎样的条件下，分组交换的时延比电路交换要小？

例题：Assignments 1 （作业1）

补充题：



- 电路交换时延：$s+x/b+kd$
- 分组交换时延：$x/b+kd+(k-1)p/b$

$$x/b+kd+(k-1)p/b < s+x/b+kd$$

$$\rightarrow s > (k-1)p/b$$

*但前提是：$x>>p$, 或分组数大于链路数.

例题：Assignments 1 （作业1）

- 补充题2：在上题的分组交换网中，设报文和分组长度分别为 x 和 $(p+h)$(bit)，其中 p 为分组的数据部分的长度，而 h 为每个分组所带的控制信息固定长度，与 p 的大小无关。通信的两端共经过 k 段链路。链路的数据率为b(bit/s)，但传播时延和结点的排队时延均可忽略不计。若打算使总的时延为最小，问分组的数据部分长度 p 应该取多大？

例题：Assignments 1 （作业1）

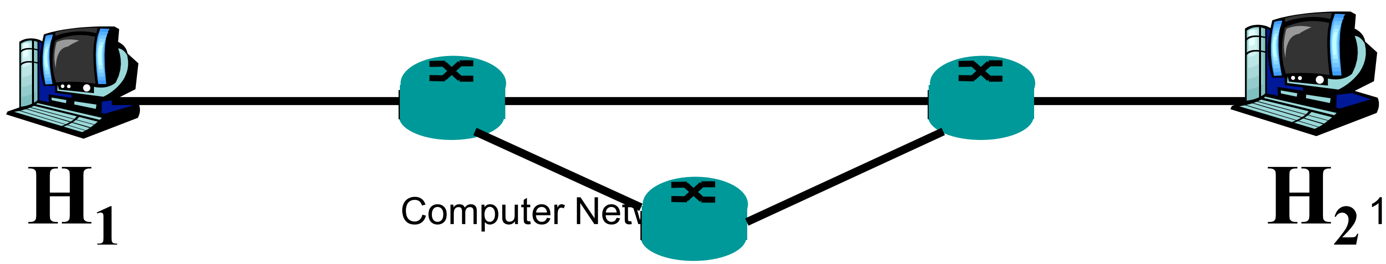
补充题（续）

- 分组数：$\dfrac{x}{p}$ ， 发送的数据量：$x+\dfrac{x}{p}\cdot h$
- 总时延 $D=\left(x+\dfrac{x}{p}\cdot h\right)\cdot\dfrac{1}{b}+(k-1)\cdot\dfrac{(p+h)}{b}$
- 求D对p的导数，令其为0： $D'_p=0$

$$\Longrightarrow \quad p=\sqrt{xh/(k-1)}$$

2010年全国硕士研究生入学统一考试
计算机学科专业基础综合

在下图所示的采用“存储-转发”方式的分组交换网络中，所有链路的数据传输速率为100Mbps，分组大小为1000 B，其中分组头大小为20 B。若主机H_1向主机H_2发送一个大小为980 000 B的文件，则在不考虑分组拆装时间和传播延迟的情况下，从H_1发送开始到H_2接收完为止，至少需要多少时间？



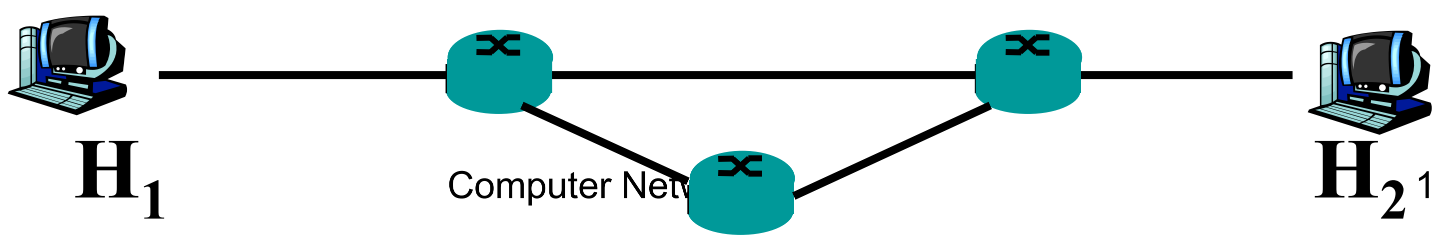
2010年全国硕士研究生入学统一考试
计算机学科专业基础综合

B=100Mbps;　　x=980 000 B

p=(1000-20) B;　　h=20B;　　k=2

$$D = \left(x + \frac{x}{p} \cdot h \right) \cdot \frac{1}{b} + (k-1) \cdot \frac{(p+h)}{b}$$

D=80.16 msec



2010年全国硕士研究生入学统一考试
计算机学科专业基础综合

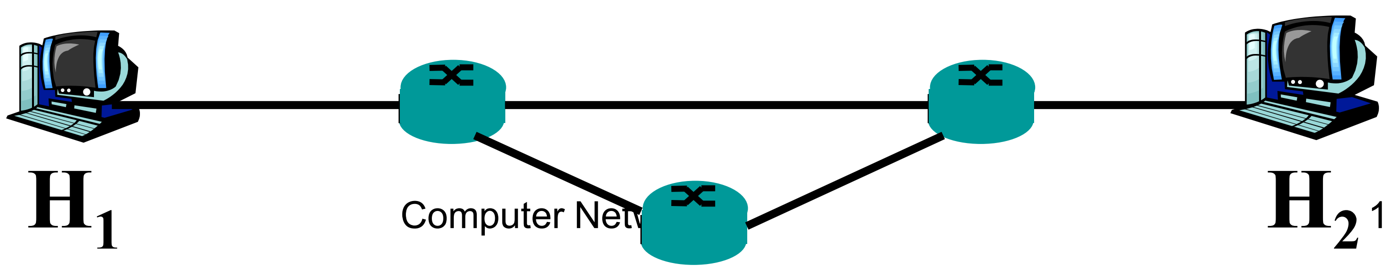
分组长度L=1000B=980B+20B

分组数N=980000/980=1000

发送1个分组的时间$T_{tran}=(1000x8)/(100x10^{6})$

$$=8x10^{-5} \text{ sec}$$

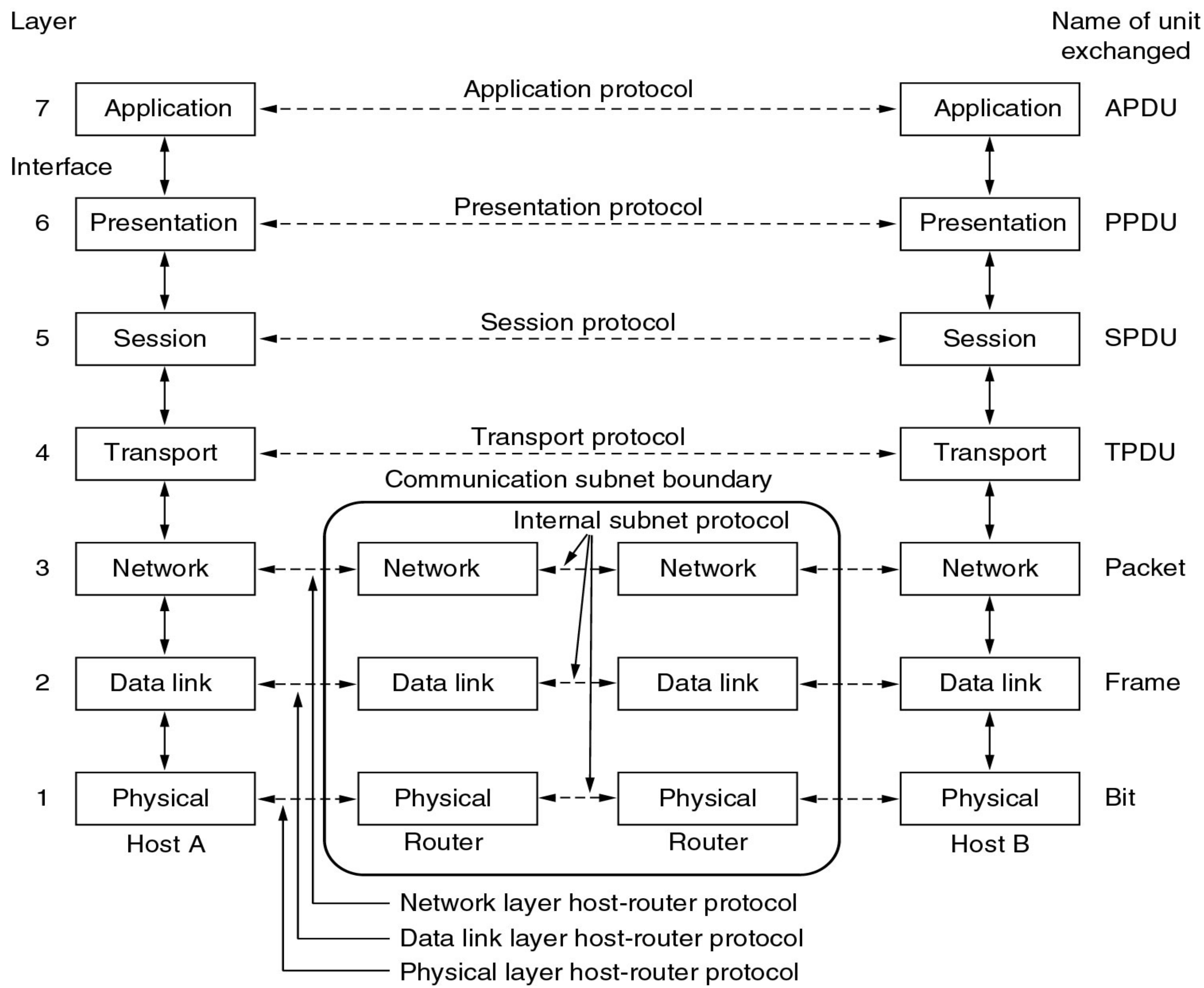
$$T_{total}=N \text{ x } T_{tran}+2 \text{ x } T_{tran}= 80.16 \text{ msec}$$



Network Architecture 网络体系结构

- **OSI (Open System Interconnection) is an ISO standard for worldwide communications**
- **The OSI Reference Model defines a networking framework for implementing protocols in seven layers.**

The ISO/OSI Reference Models



Network Architecture 网络体系结构

① **Physical Layer** → how to transmit bits to the channel;

② **Data Link Layer** → how to transmit frames to adjacent node (neighbour), over a single link ；

③ **Network Layer** → how to route packets to a host on the other side, across network(s)

Network Architecture 网络体系结构

④ **Transport Layer** → how to send data segments to a process running on another host , across network(s)

⑤ Session Layer → manage connections

⑥ Presentation Layer → encode/decode messages, security, encryption

⑦ **Application Layer** → everything else!

Network Architecture 网络体系结构

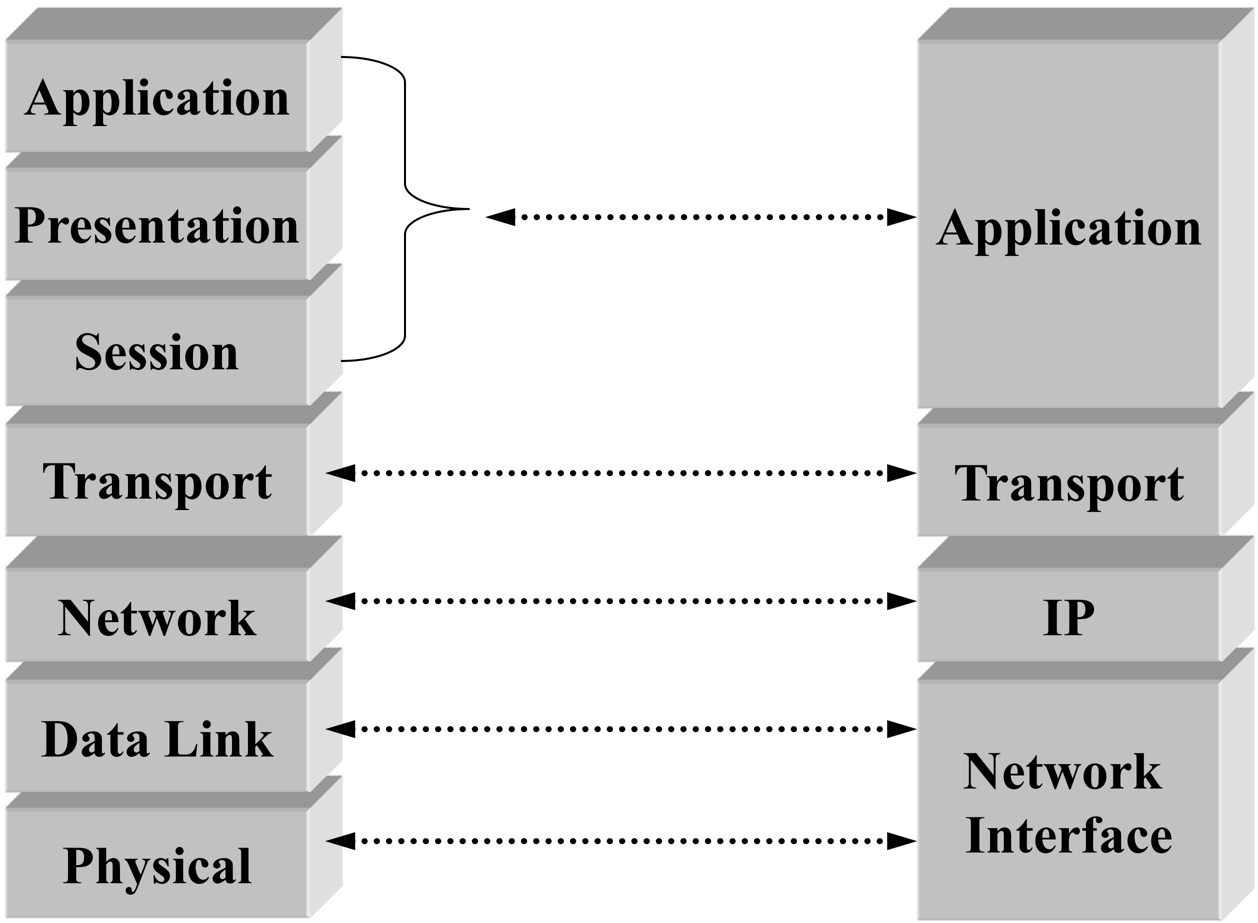
④ **Transport Layer** → how to send data segments to a process running on another host , across network(s)

⑤ Session Layer → manage connections

⑥ Presentation Layer → encode/decode messages, security, encryption

⑦ **Application Layer** → everything else!

Comparison: OSI and TCP/IP



The Physical Layer

- **Mechanical and electrical specifications**
- **Encoding/Dencoding techniques**
- **Propagation Effects:**
  - **Attenuation衰减,**
  - **Distortion失真,**
  - **Noise噪音,**
  - **Interference冲突, …**

The Physical Layer

- **Bandwidth带宽: Capacity of a media to carry information**
- **A channel is a portion of the total bandwidth used for a specific purpose.**
  - **Simplex channel 单工信道,**
  - **Half-duplex channel 半双工信道**
  - **Full-duplex channel 全双工信道.**

The Maximum Data Rate of a Channel: Nyquist's theorem尼奎斯特定理

Harry Nyquist (1889--1976), an important contributor to information theory.

- 如果一个任意信号通过带宽为H的低通滤波器，那么只需要每秒采样2H次就能完全地重现被滤波的信号。

The Maximum Data Rate of a Channel: Nyquist's theorem尼奎斯特定理

- **Suppose we know the bandwidth (*H*) of a channel and the number of signal levels (*V*) being used. What is the maximum number of bit we could transmit?**

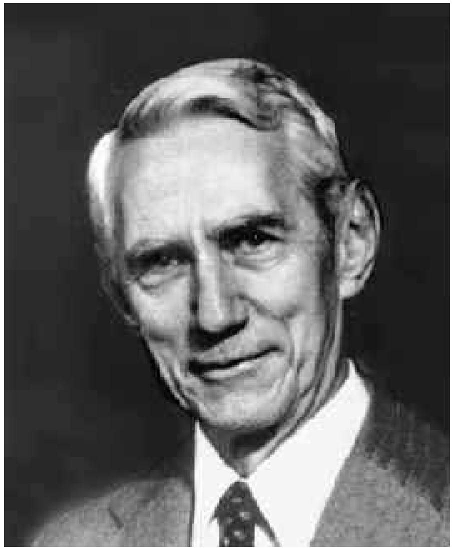
***Nyquist's theorem* says:**

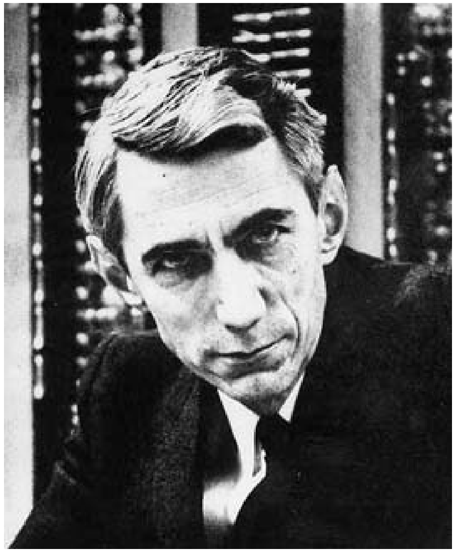
Max data rate $R_{\max} = 2H \log_2 V$ bits/sec

For example, if the bandwidth is 3100Hz and we are using 16 level modulation then the maximum number of bits per second is:

max data rate = $2 \times 3100 \times \log_2(16)$

The Maximum Data Rate of a Channel: Nyquist's theorem尼奎斯特定理





Claude Elwood Shannon

1916 - 2001

In 1948, he published a research paper at Bell System Technical Journal： "通信中的数学原理".

The Maximum Data Rate of a Channel: Nyquist's theorem尼奎斯特定理

- Shannon's major result is that the maximum data rate of a noisy channel whose bandwidth is ***H*** Hz, and whose signal-to-noise ratio is ***S/N***, is given by

$$R_{\max} = H \log_2 (1 + S / N) \quad \text{bits/sec}$$

- Shannon's result applies to any channel subject to thermal noise香农的结论适用于任何带有热噪声信道.

The Data Link Layer

- **The data link layer is responsible for efficient reliable communication across a physical link.**
  - **LOGICAL LINK sublayer (LLC)**
  - **MEDIA ACCESS sublayer (MAC)**
    - **Ethernet (IEEE802.3)**
    - **Wireless LAN (IEEE802.11)**

The Data Link Layer

OSI

- Application
- Presentation
- Session
- Transport
- Network
- Data Link
- Physical

LOGICAL LINK sublayer

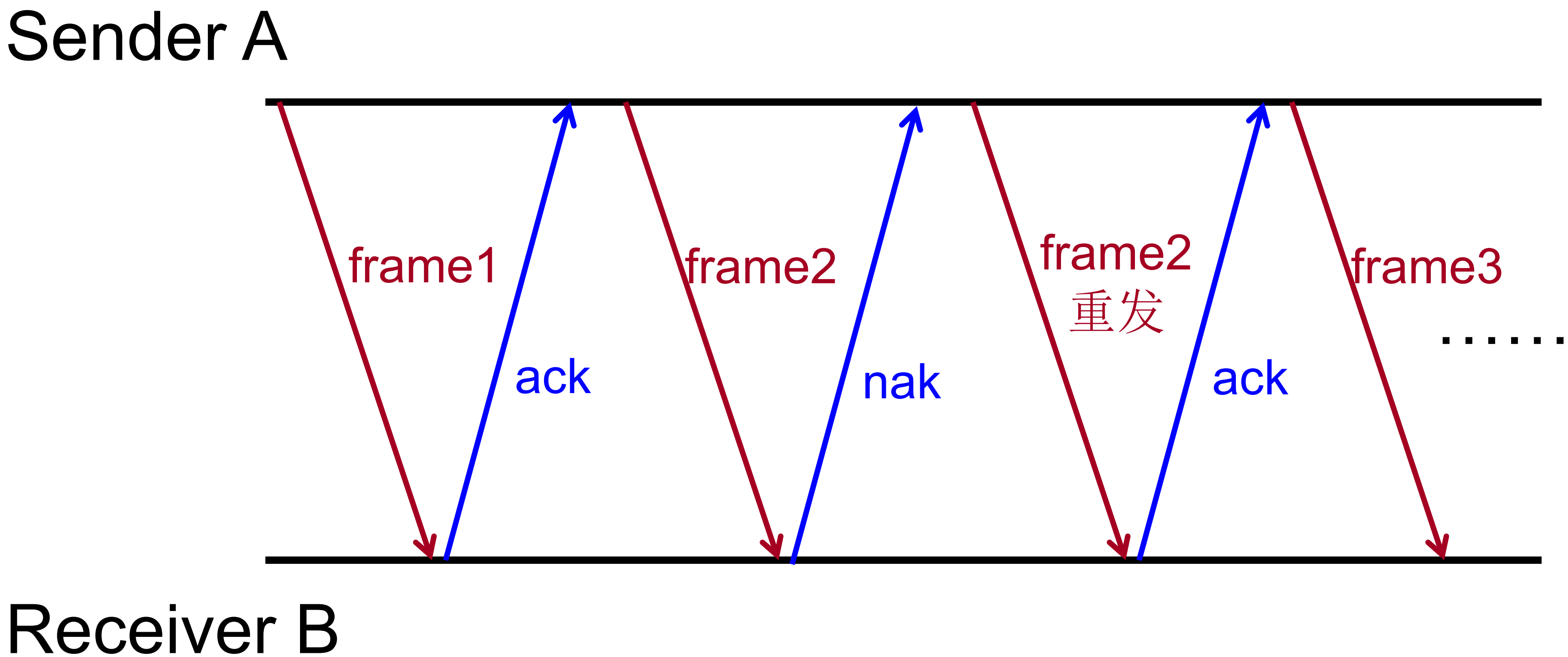
- Framing
- Error control
- Flow control

MEDIA ACCESS sublayer

- Transmission/reception of frames

Data Link Protocols

Protocols in which the sender sends one frame and then waits for an acknowledgement before proceeding are called stop-and-wait.



Stop and Wait Protocols
停止等待协议

Problem: B sends a NAK frame back to A, after having received a data frame with errors. What happens if A always gets NAK frames.

Solution: set a max. number for retransmission times, *e.g.* 8. If not successful, gives an error report to the above layer.

Stop and Wait Protocols
停止等待协议

Problem: **Due to poor link conditions, the frame sent by A doesn't get B at all. It gets lost ! In this case, A will never get any response from the peer.**

Solution: **schedule a timeout timer to expire at some time after the ACK should have been returned. If the timer goes off, retransmit the frame.**

Stop and Wait Protocols
停止等待协议

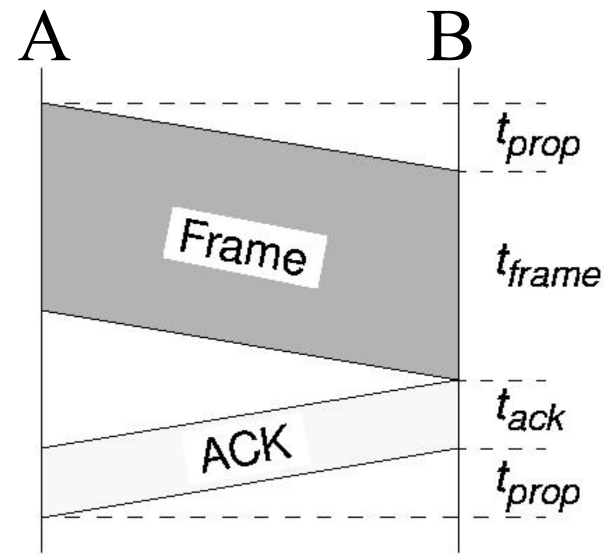
Problem: Retransmissions may introduce duplicate frames received by B

Solution: assign sequence numbers序号 for every frame, so that B can distinguish between new frames and old copies. However, an ACK for the duplicated frame is still necessary!

Stop and Wait Protocols
停止等待协议

- **A protocol, in which the sender waits for a positive acknowledgement before advancing to the next data item, are often called ARQ (Automatic Repeat reQuest).**
- **The ARQ protocol is very simple;**
- **Unfortunately, it gives poor link utilization.**

Stop-and-Wait Protocol Performance



t_{prop} → Propagation delay:

- defined as the delay between transmission and receipt of frames between hosts
- can be used to estimate timeout period

t_{frame} → Frame transmission time

t_{ack} → Acknowledgment transmission time

T_D → Total delay (ignoring ACK transmission time):

$$T_D = 2t_{prop} + t_{frame}$$

Stop-and-Wait Protocol Performance

Of this time, only t_{frame} is actually spent transmitting data. Therefore, the efficiency or utilization is:

$$U = \frac{t_{frame}}{T_D} = \frac{t_{frame}}{2t_{prop} + t_{frame}}$$

If we define *a*:

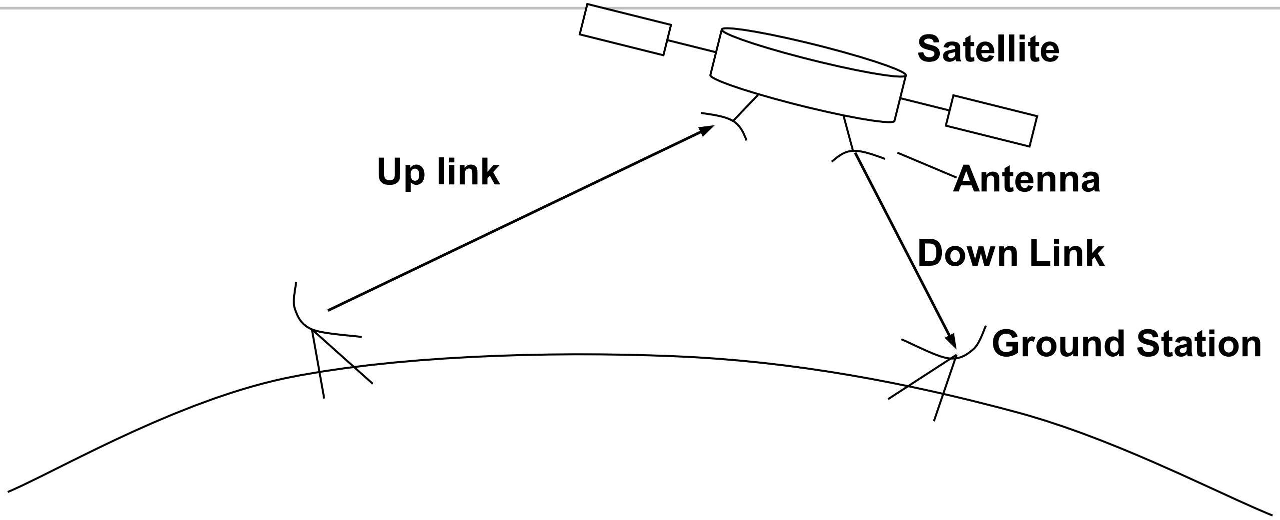
$$a = \frac{t_{prop}}{t_{frame}}$$

then

$$U = \frac{1}{1 + 2a}$$

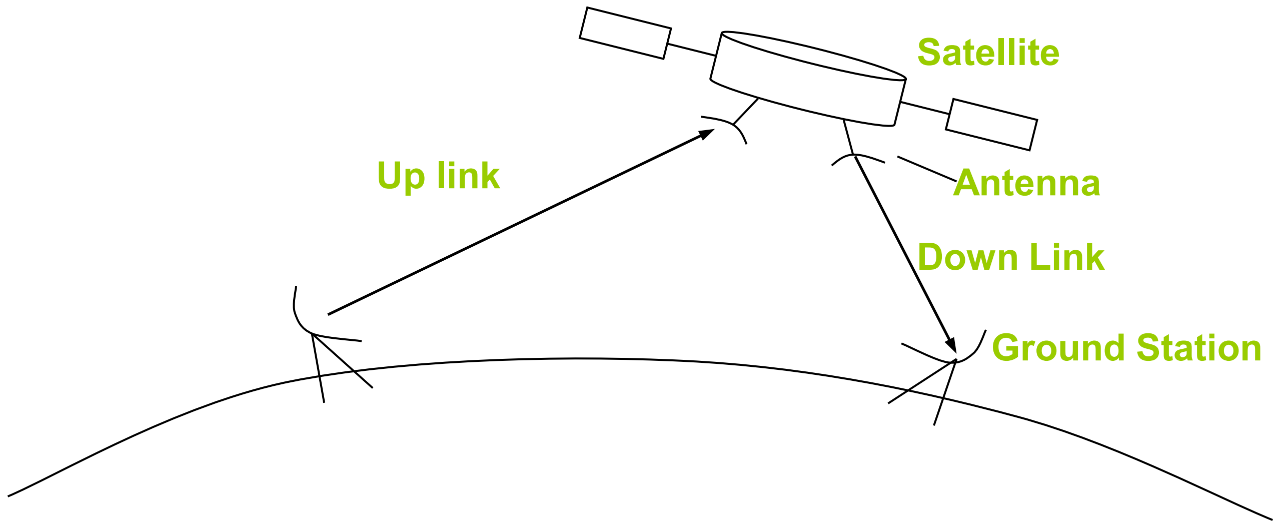
Quiz

Imagine a link that uses a geo-stationary satellite地球同步卫星:



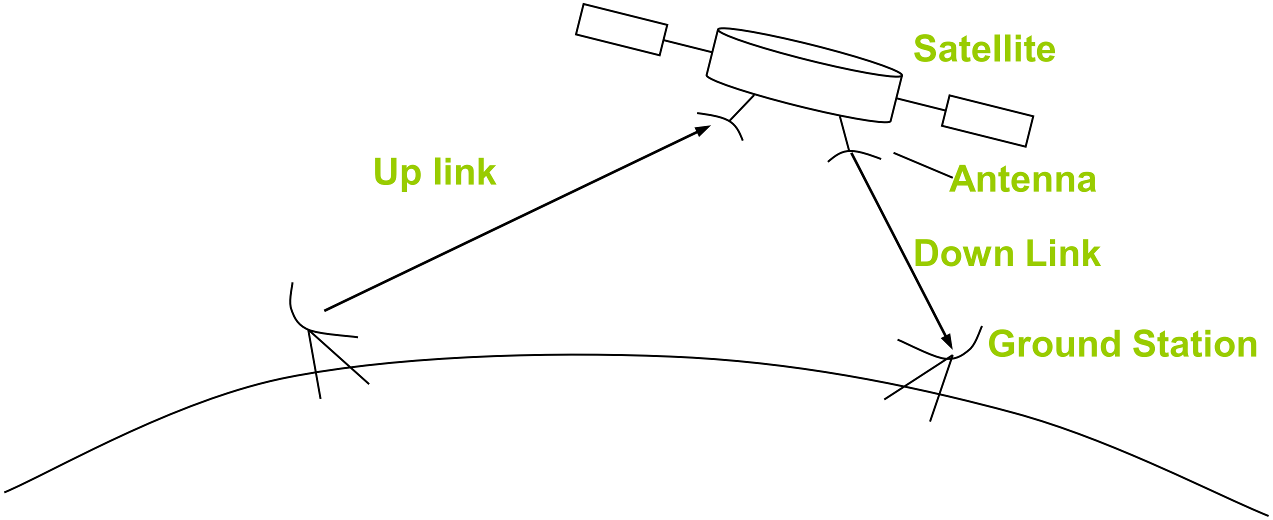
Quiz

Imagine a link that uses a geo-stationary satellite地球同步卫星：



- **The data rate is 50-kbps.**
- **The round-trip delay is 500*ms*.**
- **What is the link utilization, if you use stop-and-wait protocol to send 1000-bit frames?**

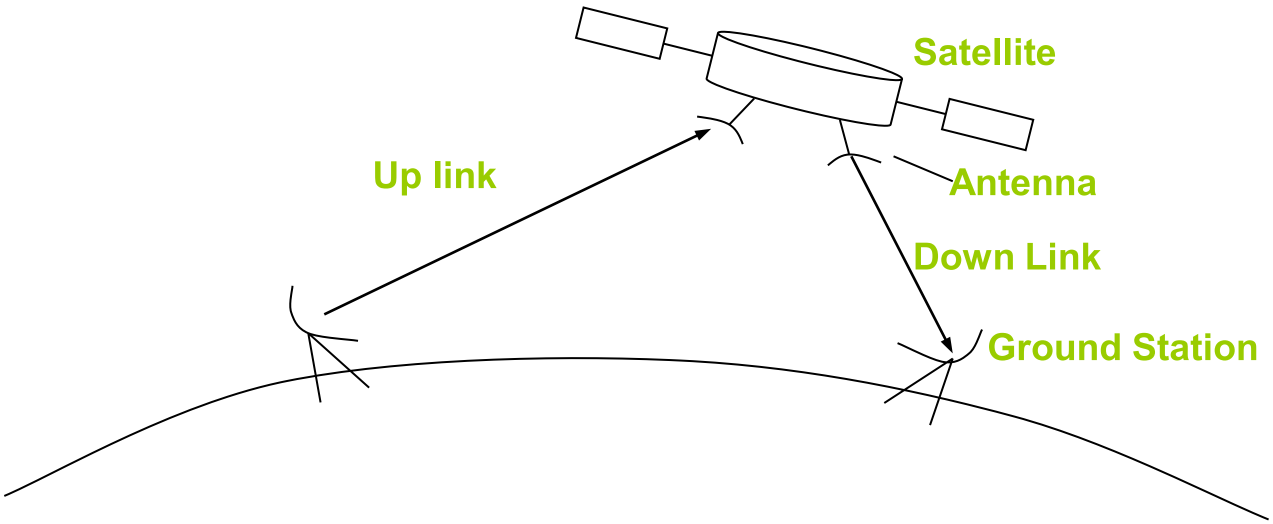
Imagine a link that uses a geo-stationary satellite地球同步卫星:



- The data rate is 50-kbps.
- The round-trip delay is 500 msec.
- **If we use stop-and-wait protocol to send a 1000-bit frame, the receiver will get the whole frame 270*msec* later.**

Long Transit Delays

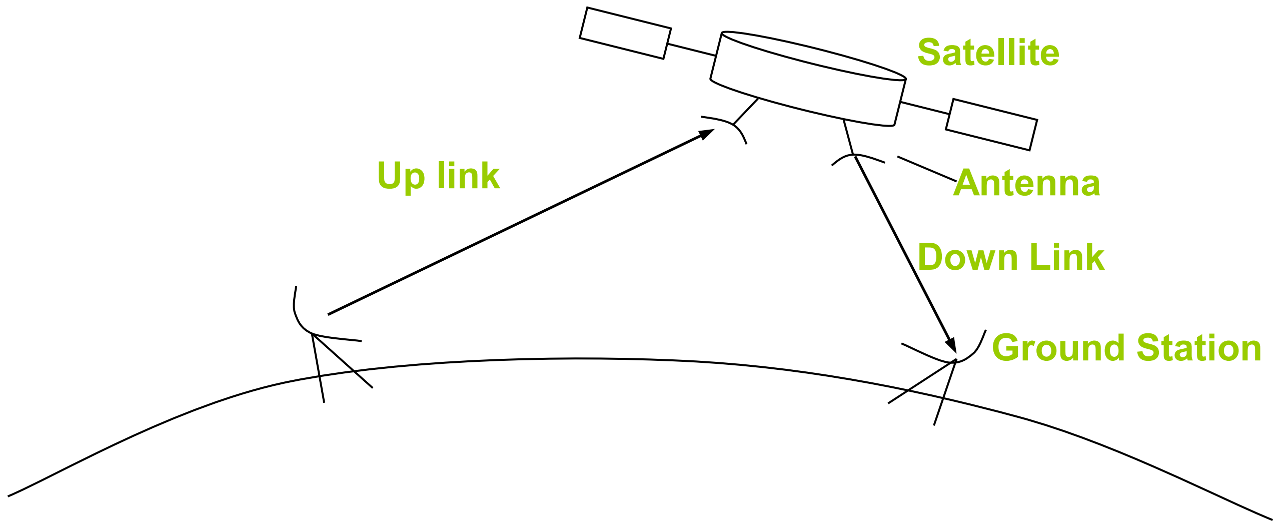
Imagine a link that uses a geo-stationary satellite地球同步卫星:



- **The acknowledgement will take a further 250*msec* to get back.**

Long Transit Delays

Imagine a link that uses a geo-stationary satellite地球同步卫星:

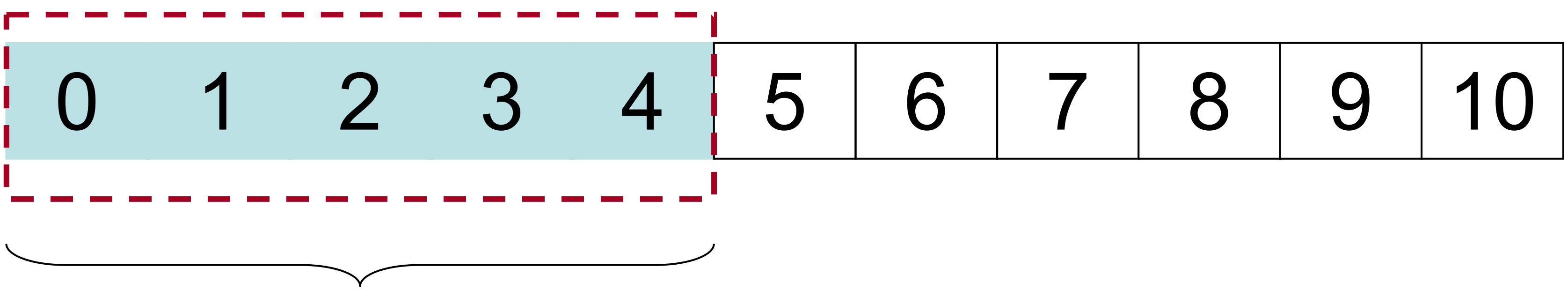


- **Out of 520*msec*, data is only being sent for 20*msec*.**
- **Only 20/520 ≈ 4% of the link's capacity is being utilised.**

滑动窗口协议
(Sliding Window Protocols)

- 目的：限制发送方已经发出，但未被确认的数据帧的数目。

- 发送窗口用来控制发送方的流量。发送窗口内的帧是允许发送的帧，而不考虑有没有收到接收方的确认。



数据链路层协议

- ***Go Back n*后退n帧协议**：**to discard the damage frame and all the frames that follow it, then retransmit all of them.**
- ***Selective Repeat*选择性重传**：**With selective repeat, only those frames that are damaged are re-sent.**

Ethernet以太网

- **The 802.3 standard describes the operation of the MAC sub-layer in a bus LAN that uses CSMA/CD**
- **Mechanism for Channel Access信道访问机制 CSMA/CD：** ***C**arrier **S**ense **M**ultiple **A**ccess with **C**ollision **D**etection载波侦听多路存取/碰撞检测*

CSMA载波侦听、多路存取

- **Carrier Sense**: With carrier sensing, A host will only transmit its own frames when it cannot hear any data being transmitted by other hosts.
- **Multiple Access**: Multiple hosts share a single channel

载波侦听：每个主机在发送数据之前首先监听信道，只有当信道空闲时才发送数据帧；如果信道忙，暂不发送（退避一个随机的时间），以免发生碰撞；

多路存取：多个主机共享信道。

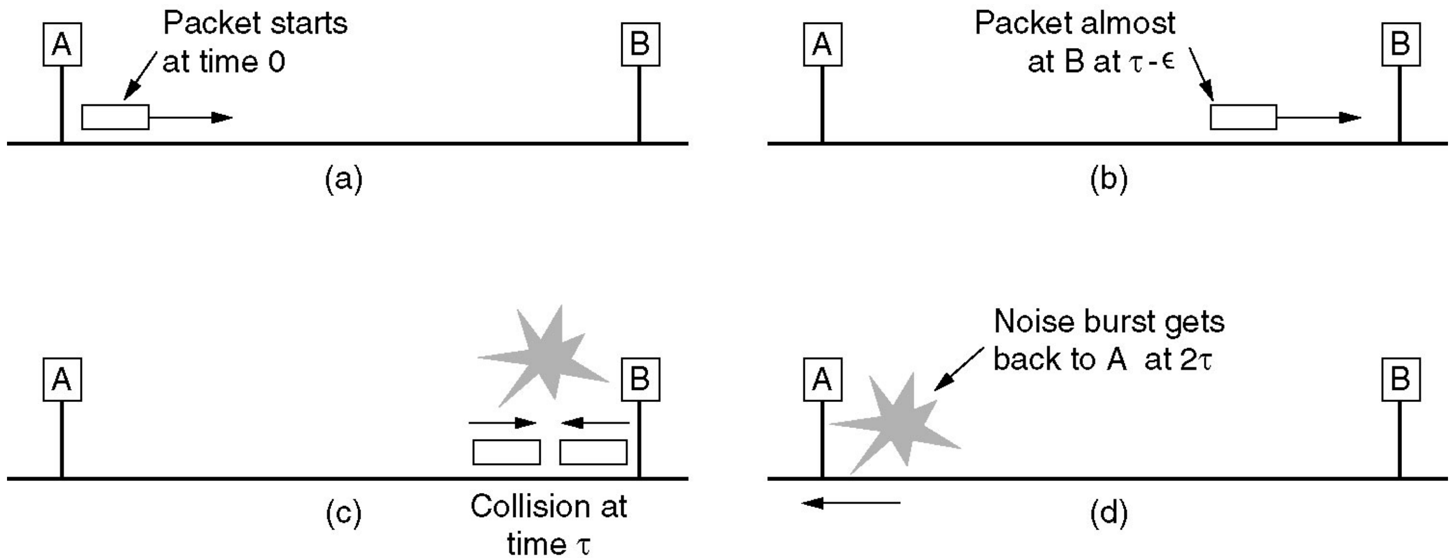
CSMA/CD
载波侦听多路存取 / 冲突检测

- CSMA with Collision Detection
  - The host always listens to the cable while it is transmitting data.
  - It aborts transmission as soon as it detects a collision
  - It tries later again, with a "*binary exponential back-off algorithm*"

冲突检测:在发送数据的过程中始终监听信道，一旦冲突立即中止发送，退避一个随机时间(二进制指数退避算法)，再重新发送。

Ethernet以太网

- When a host transmits a frame, there is a small chance that a collision will occur, i.e. non-deterministic不确定性
- The frame should be longer enough for sender to detect the collision



Ethernet以太网

- **This means that the frame must be of a minimum length.**
- **The minimum frame size is related to**
  - **the distance which the network spans;**
  - **the type of media being used;**
  - **the number of repeaters中继器 which the signal may have to pass through to reach the furthest part of the LAN.**

Ethernet以太网

- **Ethernet defines a minimum frame size, i.e. no frame may have less than 46 bytes of payload.**

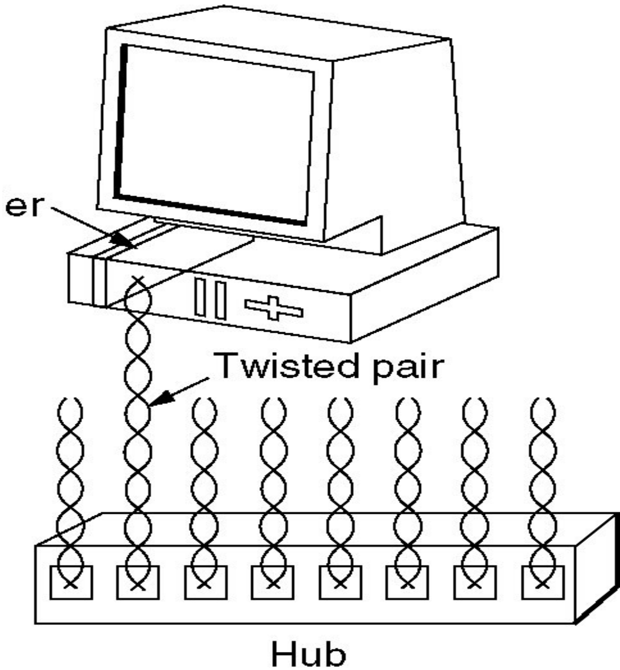
Ethernet以太网

Minimum Frame Length

- Two nodes are communicating using CSMA/CD protocol. Transmission rate is 100 Mbits/sec and frame size is 1500 bytes. The propagation speed is $3*10^8$ m/sec.
- Calculate the distance between the nodes such that the time to transmit the frame = time to recognize that the collision have occurred.

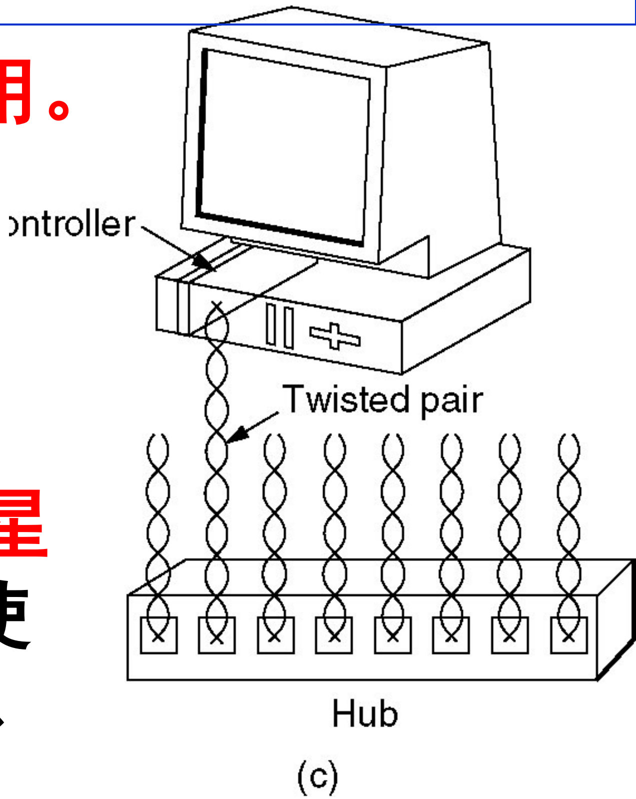
Ethernet Cabling

- In option of 10Base-T Ethernet Cabling, all stations have a cable running to a central hub 集线器 in which they are all connected electrically.
- In this case, the physical topology of the LAN is star, however, the logical topology is still bus.



(c)

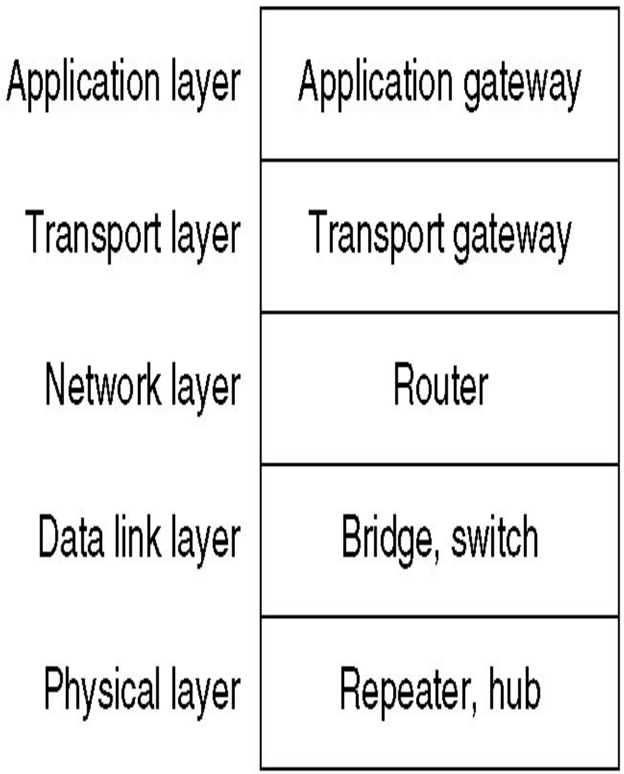
Ethernet Cabling



- 双绞线Ethernet总是和集线器配合使用。
- 术语：10Base-T Ethernet
  - ➢"10"代表10Mbit/s的数据率。
  - ➢ "T"代表双绞线星型网。
- 用集线器Hub来连接站点。物理上是星型网，但逻辑上仍是总线网。各站仍使用CSMA/CD协议，并共享逻辑上的总线。

- 各站必须竞争公共信道，并在任何时刻只有一个站可以发送数据。否则发生碰撞！！

Ethernet Extension



(a)

- **At the Physical Layer**
  - **Hub集线器**
  - **Repeaters中继器**
- **At the Data Link Layer**
  - **Bridge(网桥)**
  - **Switch(交换机)**
- **At the Network Layer**
  - **Router(路由器)**
- **……**

Wireless LAN无线局域网

- **IEEE 802.11 defines CSMA/CA protocol.** CSMA part is the same as in 802.3 Ethernet, CA stands for Collision Avoidance 冲突避免
- **How CSMA/CA works:**
  - Device wanting to transmit senses the medium (Air)
  - If medium is busy – defers
  - If medium is free for certain period, transmits frame

CSMA/CA 载波监听多路存取/冲突避免

- avoid collisions: two or more nodes transmitting at same time
- 802.11: CSMA – sense(监听) before transmitting
  - don't collide with ongoing transmission by other node

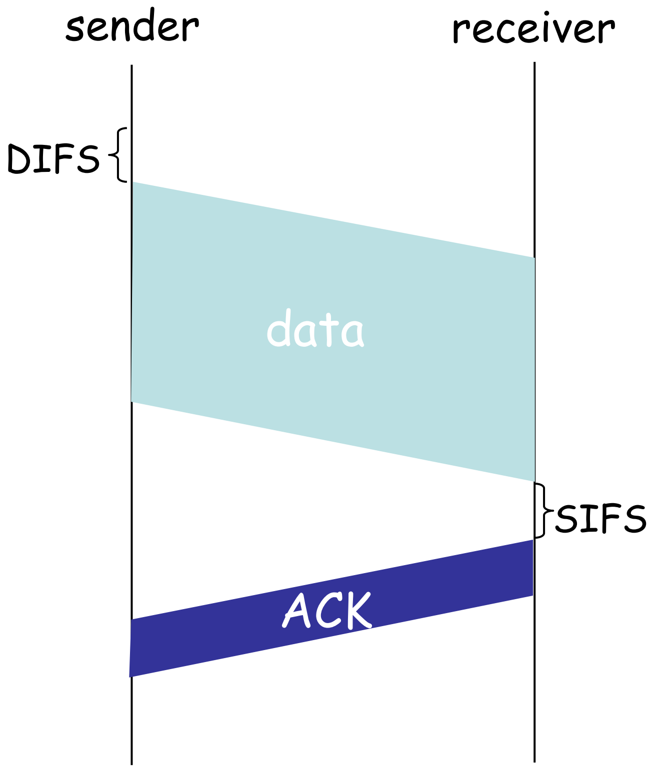
CSMA/CA 载波监听多路存取/冲突避免

<u>802.11 sender</u>

if sense channel idle for DIFS
then transmit entire frame (no CD)

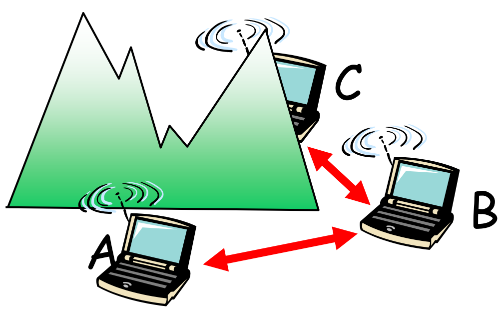
<u>802.11 receiver</u>

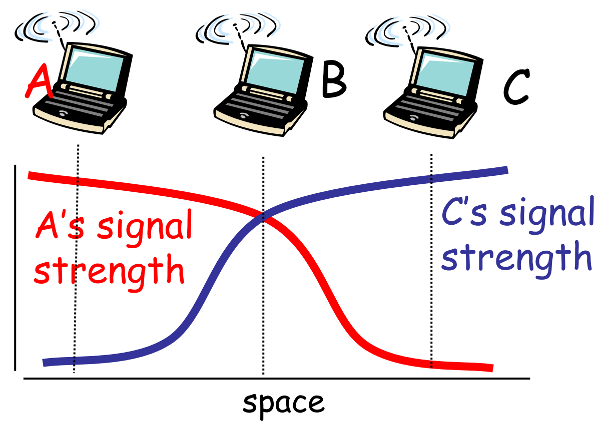
- if frame received OK
  return ACK after SIFS (ACK needed due to hidden terminal problem)



CSMA/CA 载波监听多路存取/冲突避免

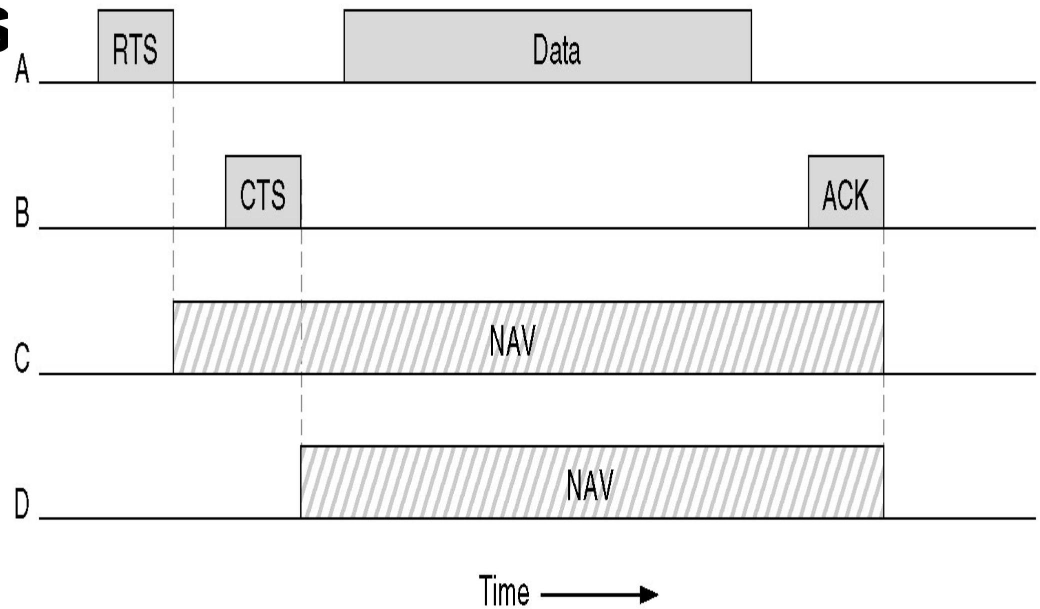
- **802.11: *no* collision detection(没有冲突检测)!**
  - can't sense all collisions in any case: hidden terminal隐蔽终端, fading信号衰减
  - goal: *avoid collisions:* CSMA/C(ollision)A(voidance)



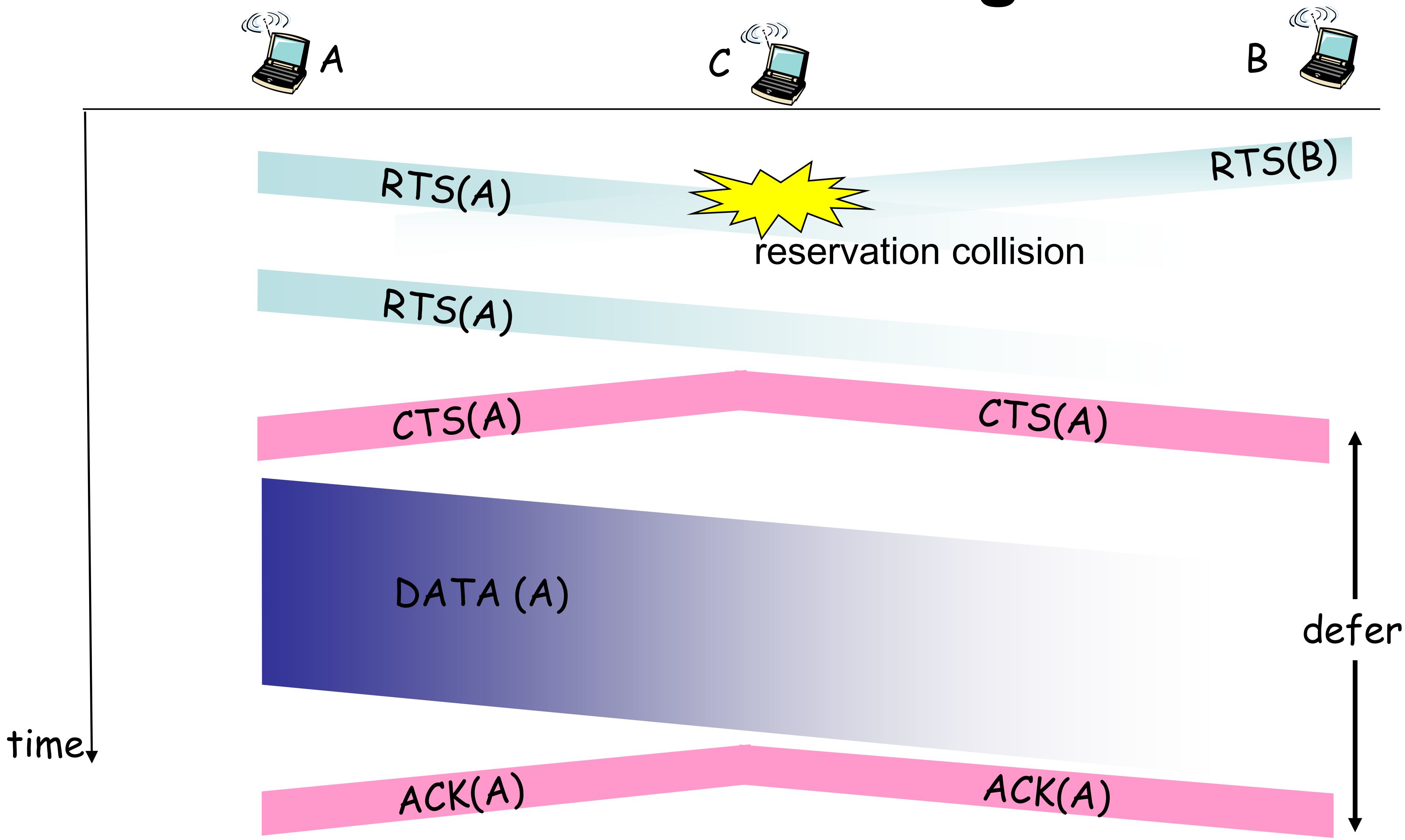


Wireless LAN无线局域网

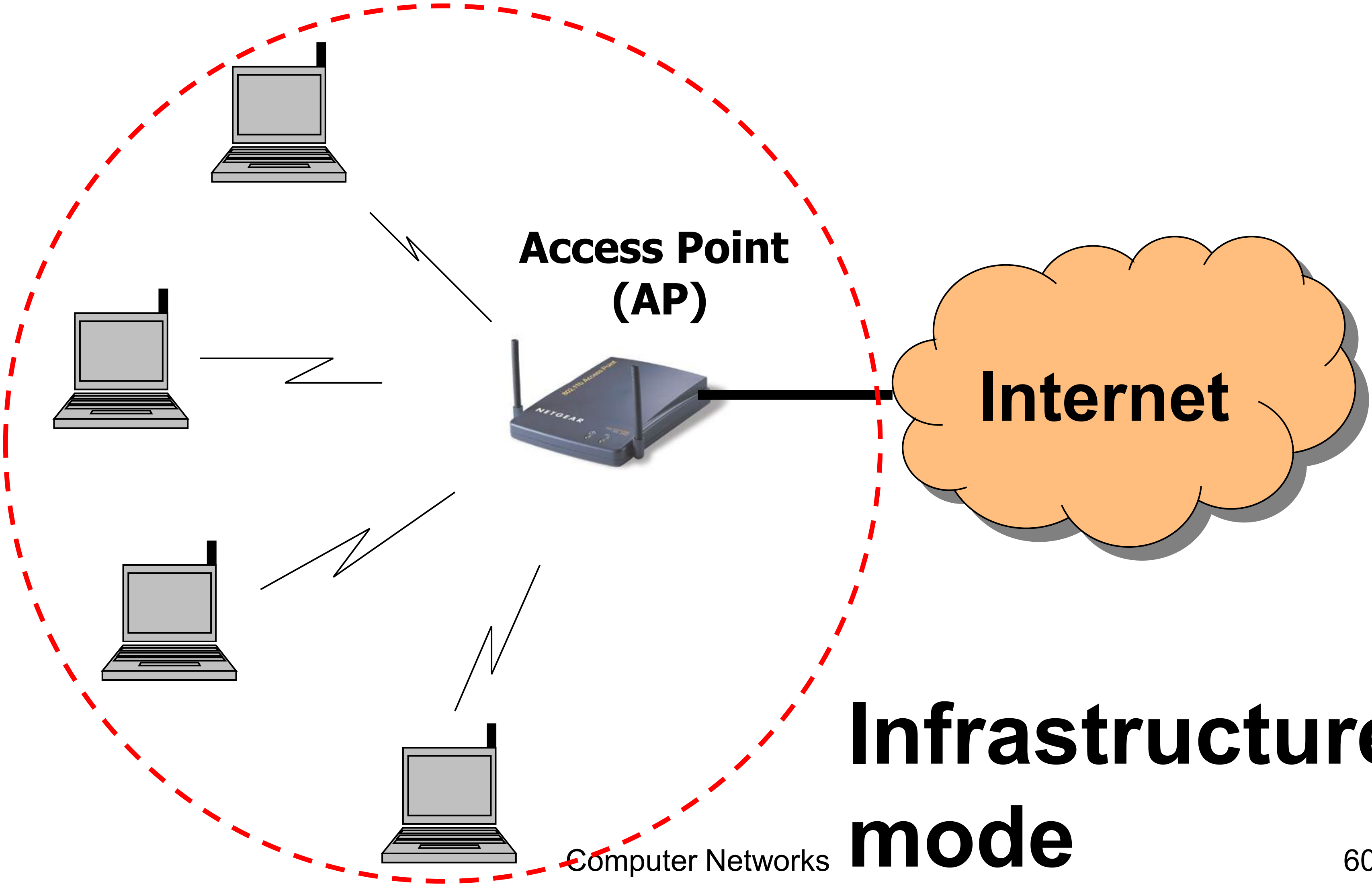
- **By sending Request to Send (RTS), sender is allowed to "reserve" channel rather than random access of data frames: avoid collisions of long data frames**



Collision Avoidance(冲突避免): RTS/CTS exchange

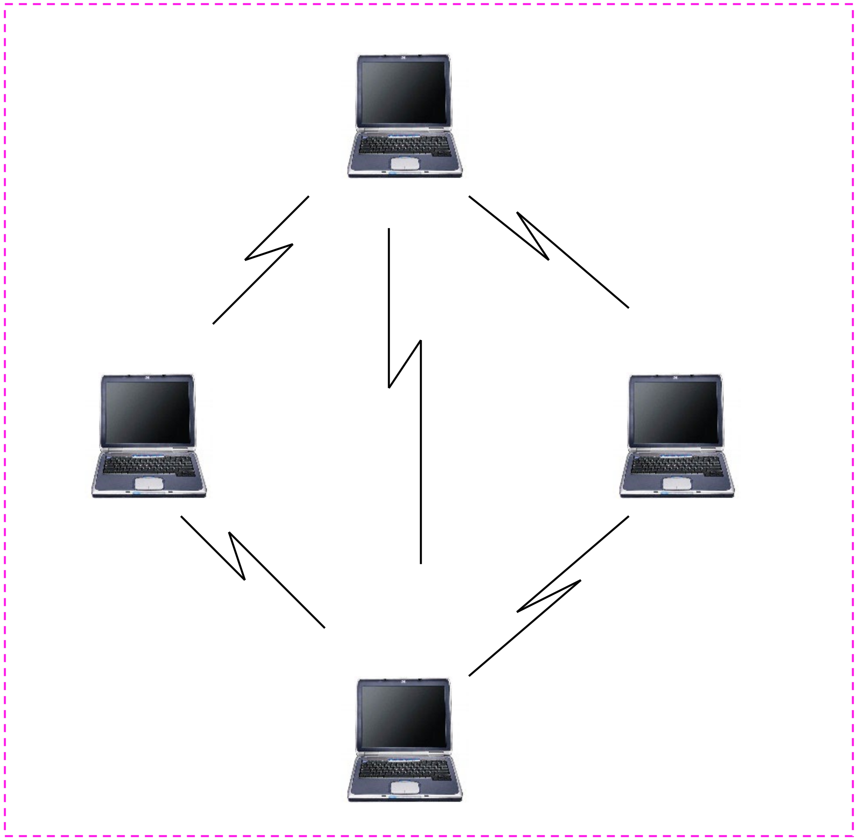


IEEE802.11 WLAN



Infrastructure mode

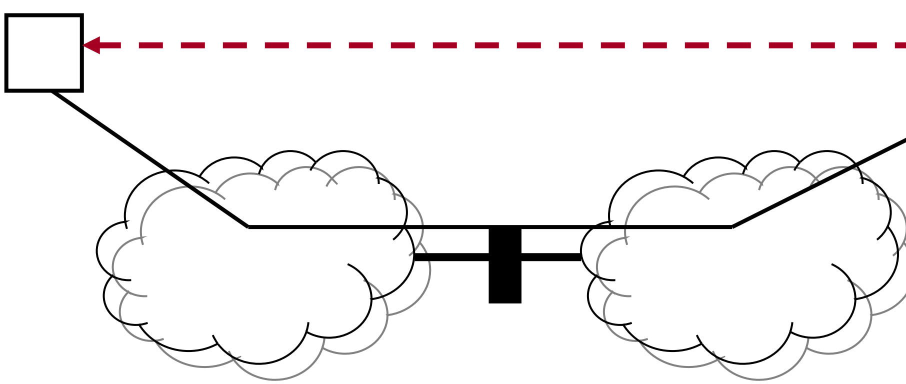
IEEE802.11 WLAN



Ad Hoc mode

The Network Layer

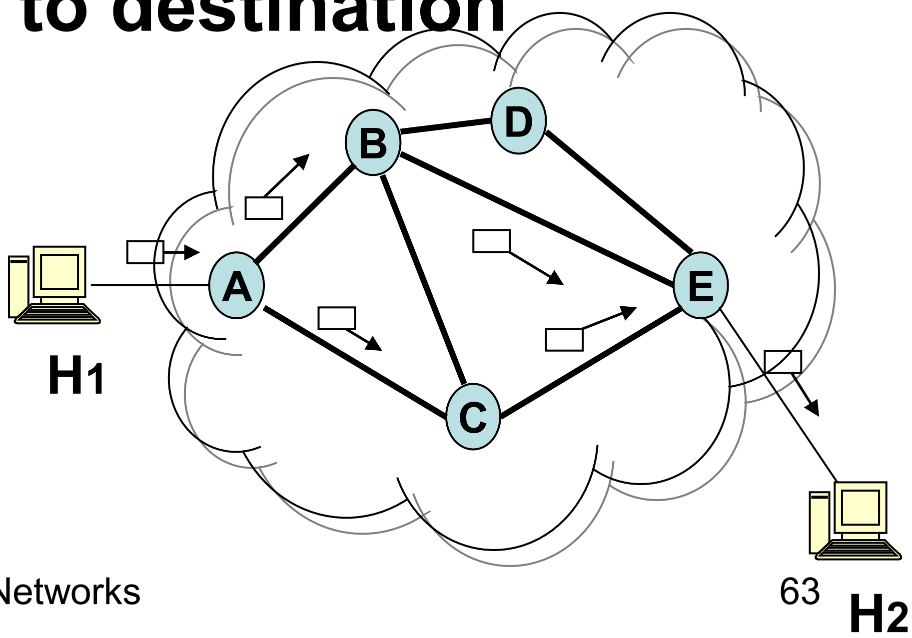
- **The network layer, based on services provided by the data link layer, provides a end-to-end transparent path for end-to-end transparent data transmission across networks.**



The functions of network layer：

Routing路由选择

- packets are injected into the subnet individually and routed independently of each other.
- Routing involves the selection of the best paths for packets from source to destination
- The routing algorithm路由选择算法 is responsible for deciding which output line an incoming packet should be transmitted on



IP Addresses

- Every node on the Internet has IP address(es)
- IP address is used to identify the network and the host on a given network
- Each IP address is 32 bits long, e.g.

  10000000 00001011 00000011 00011111

- The IP address is divided into two parts:

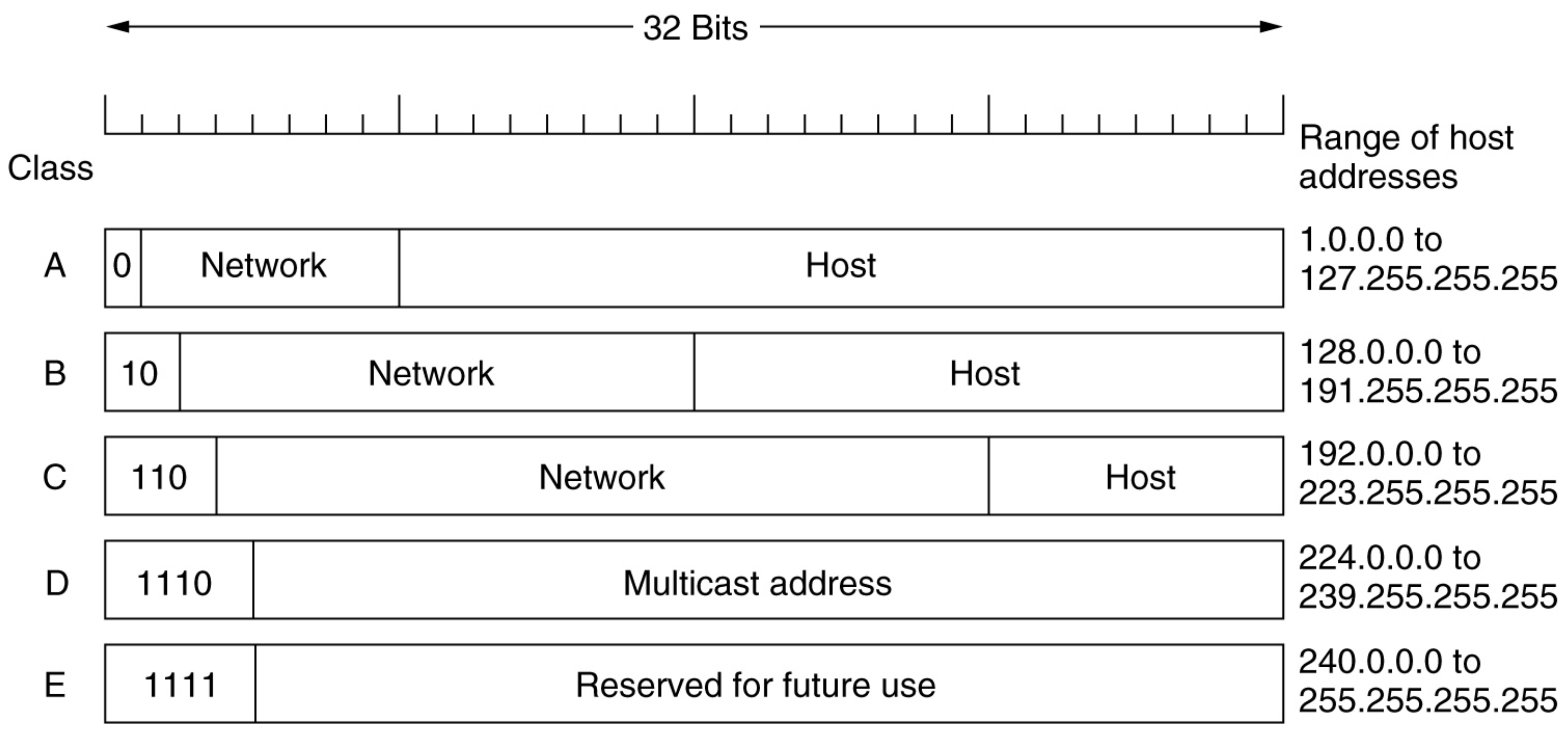
  IPaddr ::= { <net-id>, <host-id>}

IP Addresses

- **IP addresses are usually written in dotted decimal notation(点分十进制记法)， e.g.**

  10000000 00001011 00000011 00011111 □ →128.11.3.31

Classes of IP Addresses IP地址分类



Subnetting子网划分

- **Basic idea:**
  - **A organization with a large network can be divided to many smaller networks, i.e. subnets. Subnetting is an intramural matter.**
  - **take some bits from the host number part to create a “subnet” number.**

IPaddr ::= { <***net-id***>, <***host-id***>} →

IPaddr ::= {<***net-id***>, <***subnet-id***>, <***host-id***>}

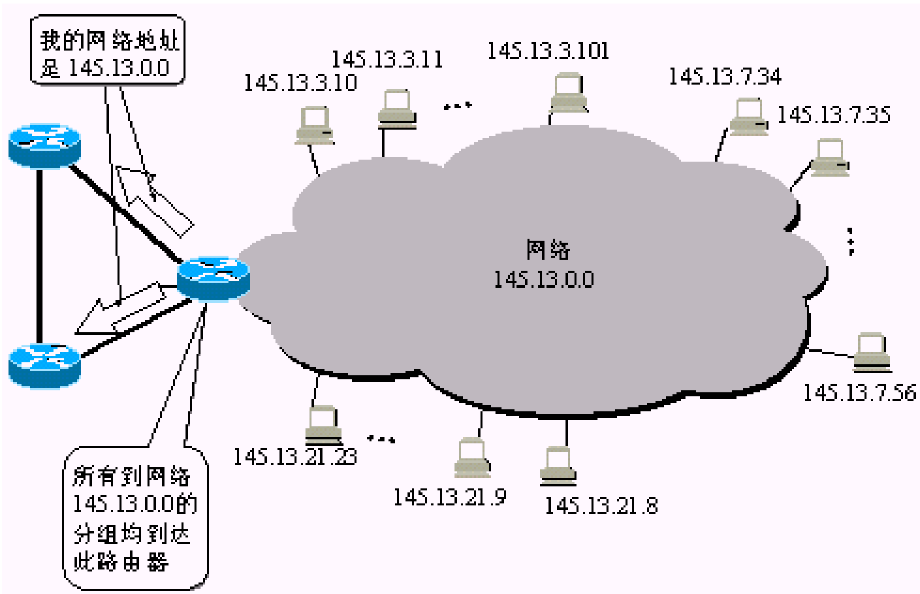


图 7-17 一个 B 类网络 145.13.0.0

A organization with a large network can be divided to many smaller networks, i.e. **subnets**. Subnetting is an intramural matter.

Subnetting:
take some bits from the host number part to create a “subnet” number.

IPaddr ::= {<net-id>, <subnet-id>, <host-id>}

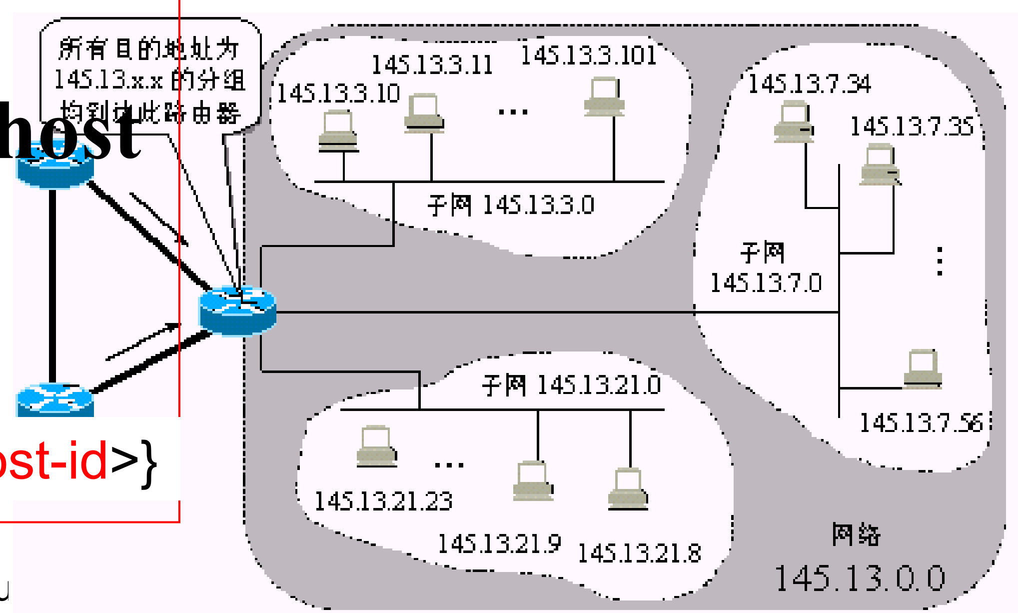


图 7-18 将图 7-17 的网络划分为三个子网，但对外仍是一个网络

Subnetting子网划分

- **Packet routing from the source to the destination across the network:**

  ……→ Destination Network

  → Destination Subnet

  →Destination Host

- **Subnet masks子网掩码 indicates which part of a 32-bit IP address represents *net-id* and *subnet-id***

Subnetting子网划分

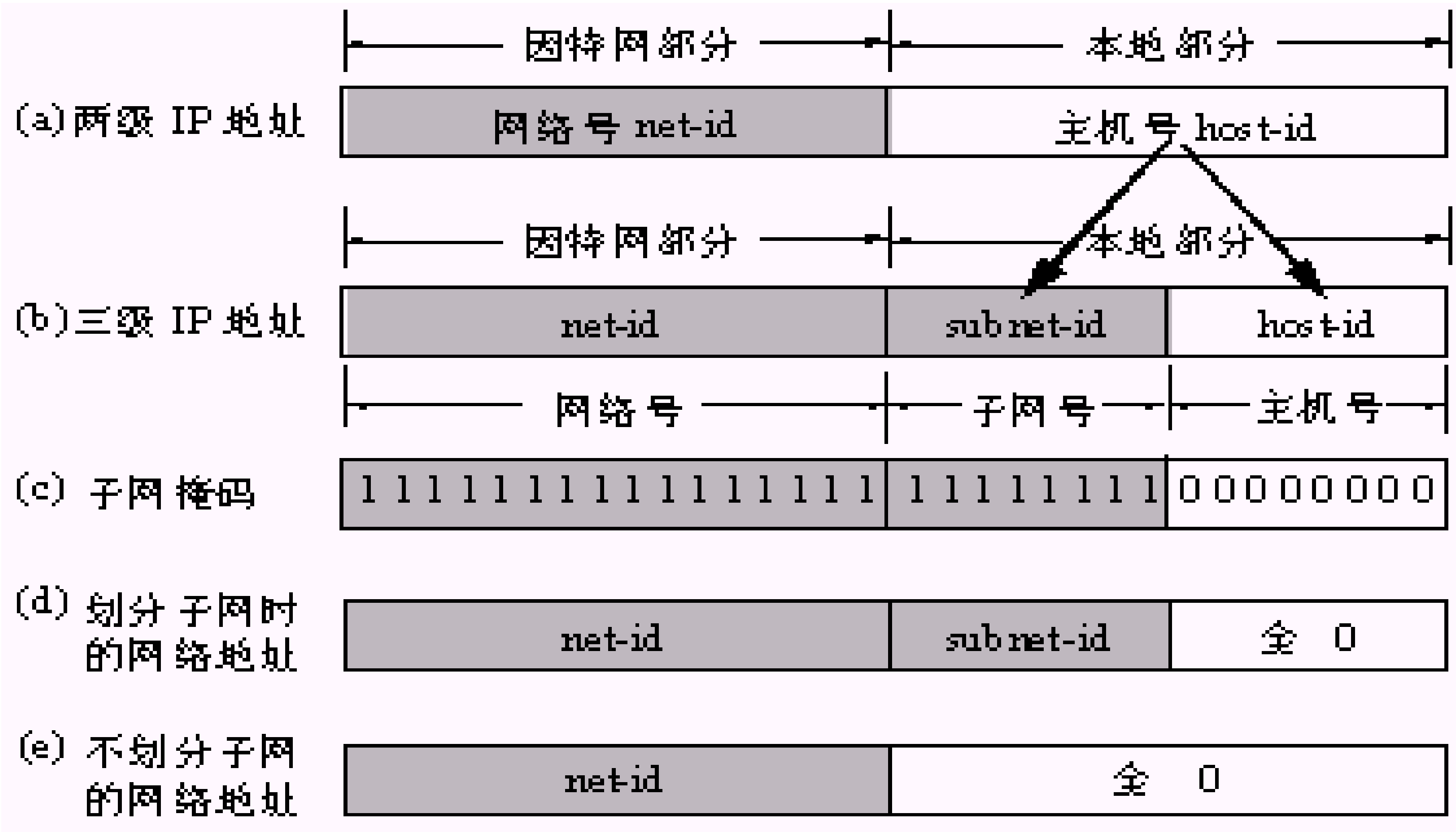


图 7-19 IP 地址的各字段和子网掩码

Subnet Examples

- IP Address: 130.97.16.132

  Subnet Mask: 255.255.255.192
  - Net-id=?
  - Host-id=?
- IP Address: 130.97.17.132

  Subnet Mask: 255.255.254.0
  - Net-id=?
  - Host-id=?

CIDR – Classless InterDomain Routing 无类域间路由选择

- *net-id* and *host-id* are replaced by *network-prefix网络前缀*

i.e. IPaddr ::= {<net-prefix>, <host-id>}.

e.g. **128.14.46.34/20**

$$\rightarrow \underbrace{10000000\quad 00001110\quad 0010}_{\text{net-prefix}}\underbrace{1110\quad 00100010}_{\text{host-id}}$$

常用的 CIDR 地址块

CIDR 前缀长度	点分十进制	包含的地址数	包含的分类的网络数
/13	255.248.0.0	512 K	8 个 B 类或 2048 个 C 类
/14	255.252.0.0	256 K	4 个 B 类或 1024 个 C 类
/15	255.254.0.0	128 K	2 个 B 类或 512 个 C 类
/16	255.255.0.0	64 K	1 个 B 类或 256 个 C 类
/17	255.255.128.0	32 K	128 个 C 类
/18	255.255.192.0	16 K	64 个 C 类
/19	255.255.224.0	8 K	32 个 C 类
/20	255.255.240.0	4 K	16 个 C 类
/21	255.255.248.0	2 K	8 个 C 类
/22	255.255.252.0	1 K	4 个 C 类
/23	255.255.254.0	512	2 个 C 类
/24	255.255.255.0	256	1 个 C 类
/25	255.255.255.128	128	1/2 个 C 类
/26	255.255.255.192	64	1/4 个 C 类
/27	255.255.255.224	32	1/8 个 C 类

Chapter 5, Problem 27

A large number of consecutive IP address are available starting at 198.16.0.0. Suppose that four organizations, A, B, C, and D, request 4000, 2000, 4000, and 8000 addresses, respectively, and in that order. For each of these, give the first IP address assigned, the last IP address assigned, and the mask in the w.x.y.z/s notation.

Chapter 5, Solution 27

To start with, all the requests are rounded up to a power of two.

Organizations A, B, C and D want to have 4000, 2000, 4000, and 8000 addresses, respectively,

so the address for them must have a host-id of 12, 11, 12 and 13 bits long,

the net-prefix is 20, 21, 20, 19 bits long, respectively.

- **单位A需要4000个地址，因此主机地址应为12位（2^{12}=4096）；由于4000/256=15.625，因此，末尾地址为198.16.15.255，即地址范围为198.16.0.0—198.16.15.255**
- **单位B需要2000个地址，主机地址11位，2000/256=7.8125，则地址范围为186.16.16.0—198.16.23.255**
- **以此类推，可以得到单位C和D的地址范围**
- **如果单位B需要8000个地址，则主机地址需要13位，8000/256=31.25，则地址范围变成198.16.32.0—198.16.63.255**

Chapter 5, Solution 27

Therefore, The starting address, ending address, and mask are as follows:

A: **198.16.0.0 –198.16.15.255**

written as 198.16.0.0/20

B: **198.16.16. 0--198.16.23.255**

written as 198.16.16. 0/21

C: **198.16.32. 0--198.16.47.255**

written as 198.16.32. 0/20;

D: **198.16.64. 0--198.16.95.255**

written as 198.16.64. 0/19

37. 某网络的IP地址空间为192.168.5.0/24，采用定长子网划分，子网掩码为255.255.255.248，则该网络中的最大子网个数、每个子网内的最大可分配地址个数分别是多少？

IP地址空间为192.168.5.0/24

$\underbrace{\text{xxxxxxxx xxxxxxxx xxxxxxxx}}$ xxxxxxxx

子网掩码为255.255.255.248

11111111 11111111 11111111 11111000

37. 单位网络的IP地址空间为192.168.5.0/24，

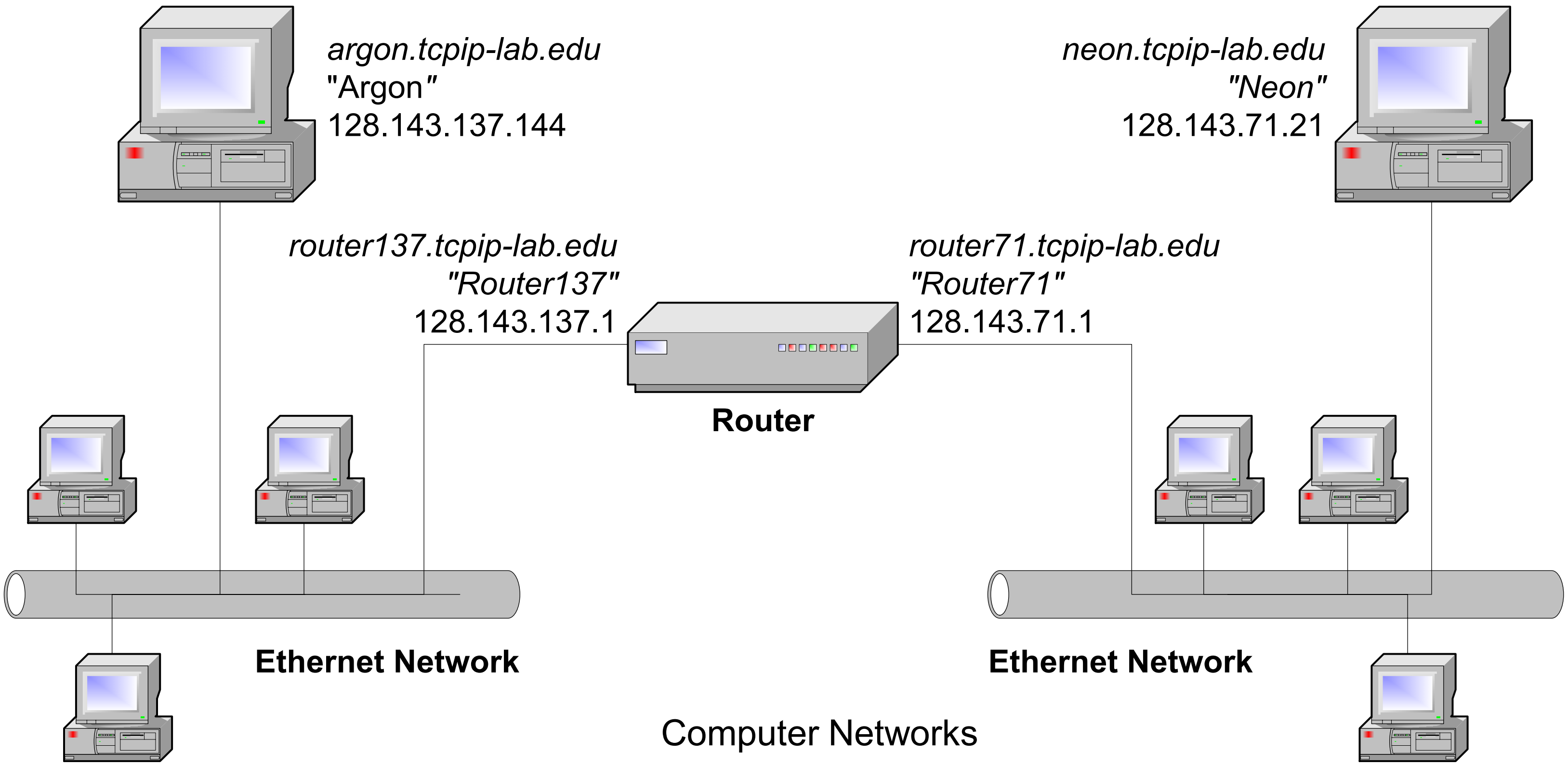
单位内部采用定长子网划分，

子网掩码为255.255.255.248，则

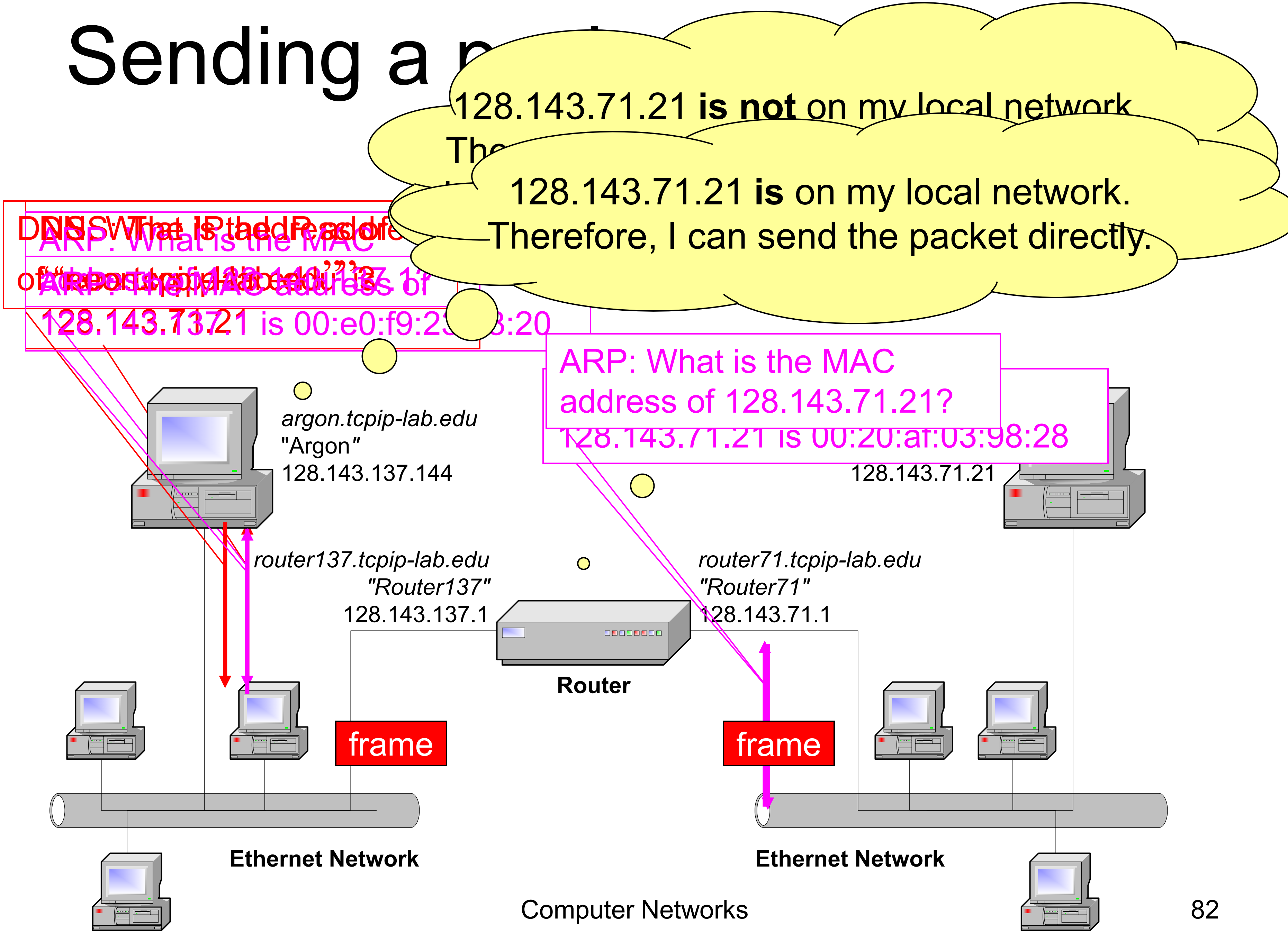
该网络中的最大子网个数：32

每个子网内的最大可分配地址个数是：6

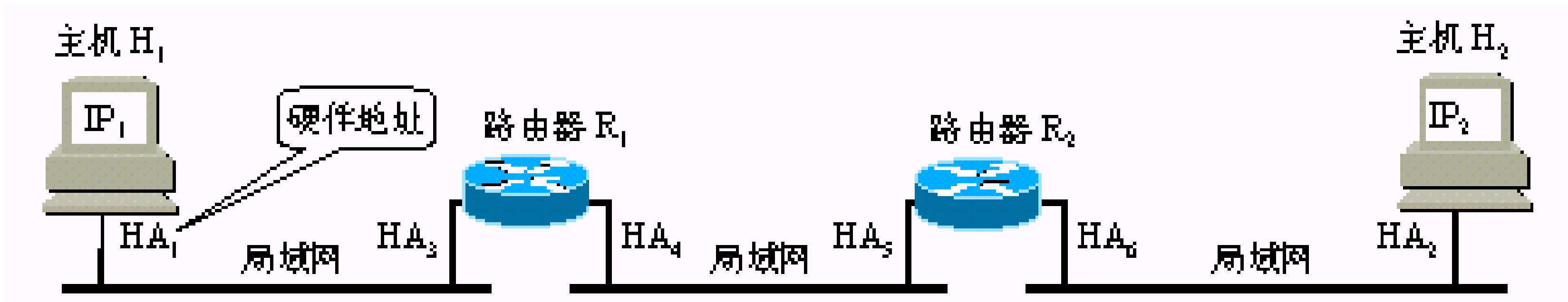
Sending a packet from Argon to Neon



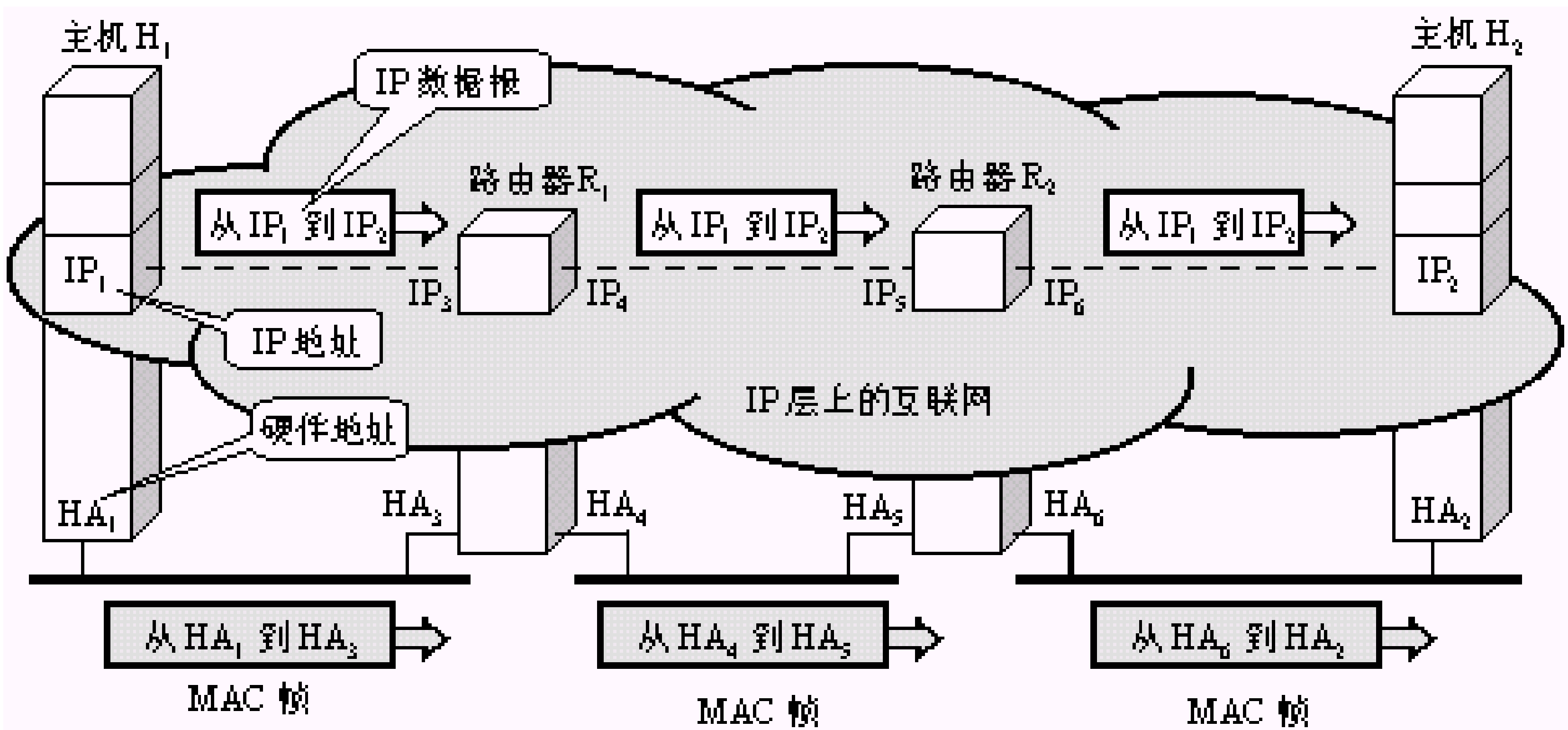
Sending a p



128.143.71.21 is not on my local network
128.143.71.21 is on my local network.
Therefore, I can send the packet directly.
ARP: What is the MAC address of 128.143.71.21?
128.143.71.21 is 00:20:af:03:98:28
argon.tcpip-lab.edu
"Argon"
128.143.137.144
128.143.71.21
router137.tcpip-lab.edu
"Router137"
128.143.137.1
router71.tcpip-lab.edu
"Router71"
128.143.71.1
Router
frame
frame
Ethernet Network
Ethernet Network

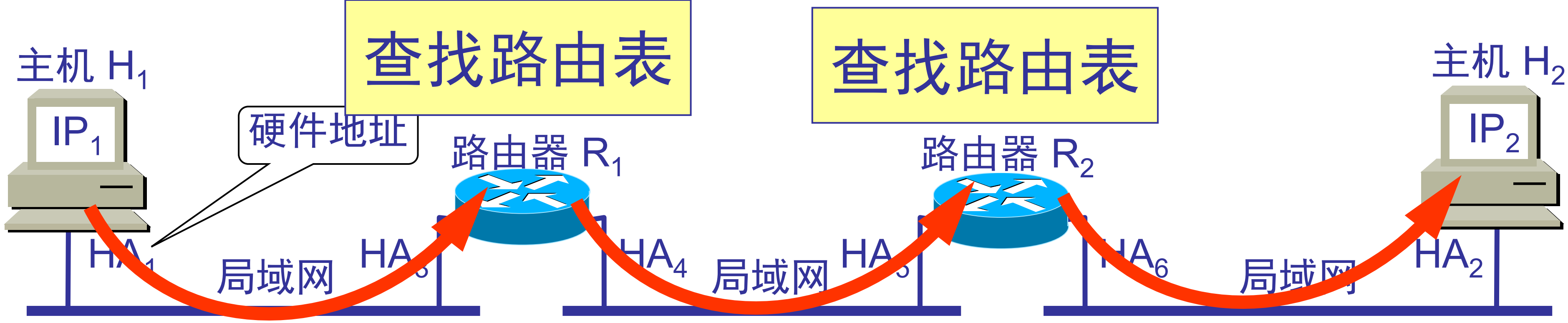


(a)网络配置



(b)不同层次、不同区间的源地址和目的地址

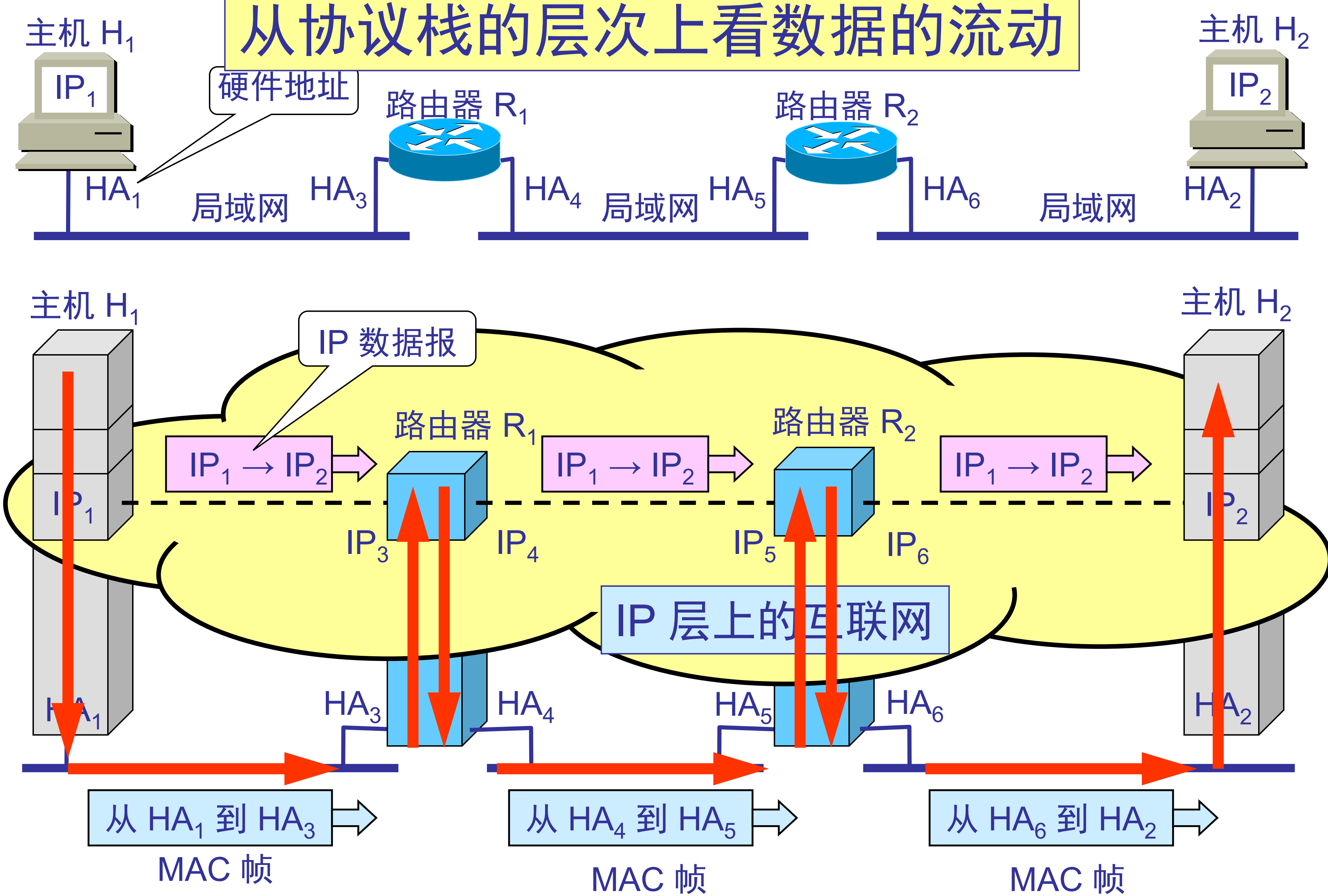
从不同层次上看 IP 地址和硬件地址



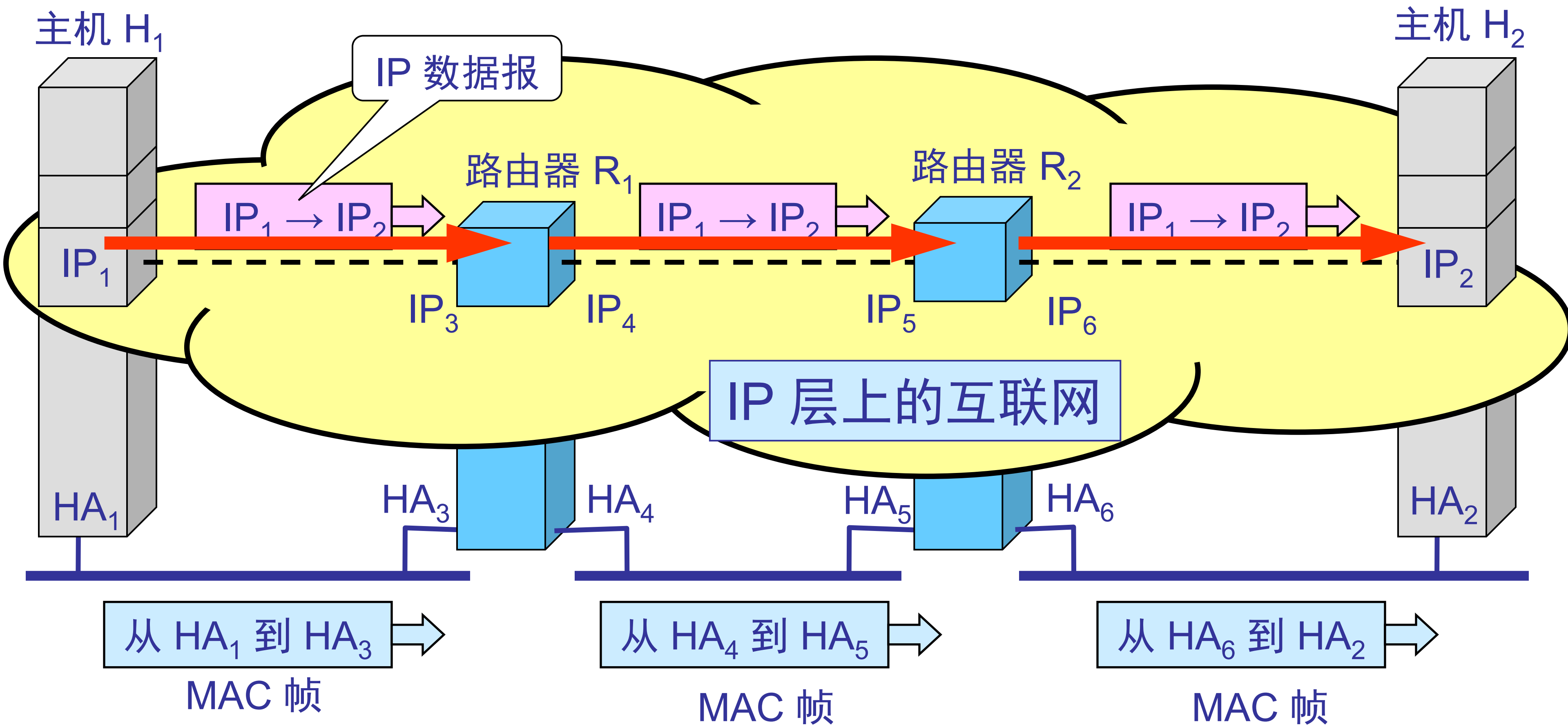
通信的路径

H_1→经过 R_1 转发→再经过 R_2 转发→H_2

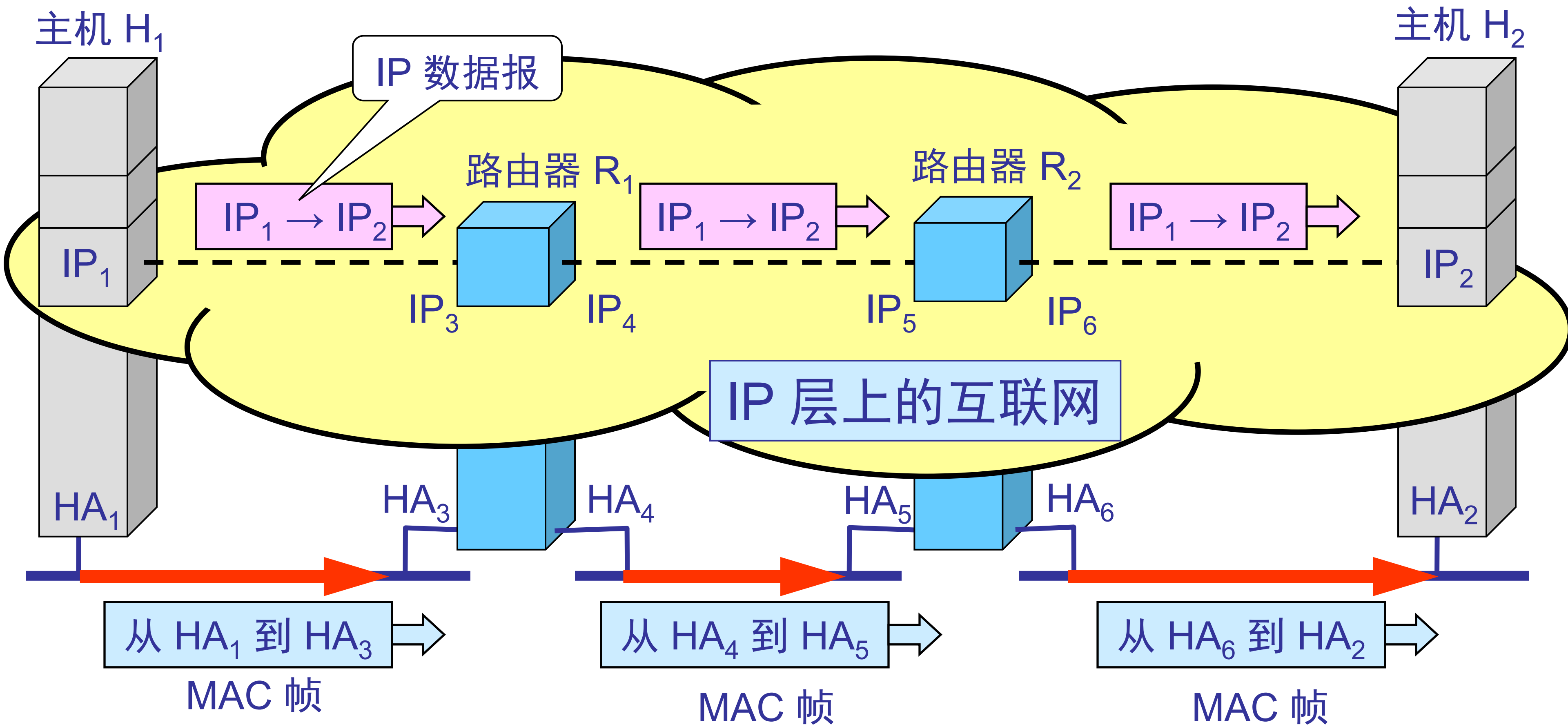
从协议栈的层次上看数据的流动



从虚拟的 IP 层上看 IP 数据报的流动



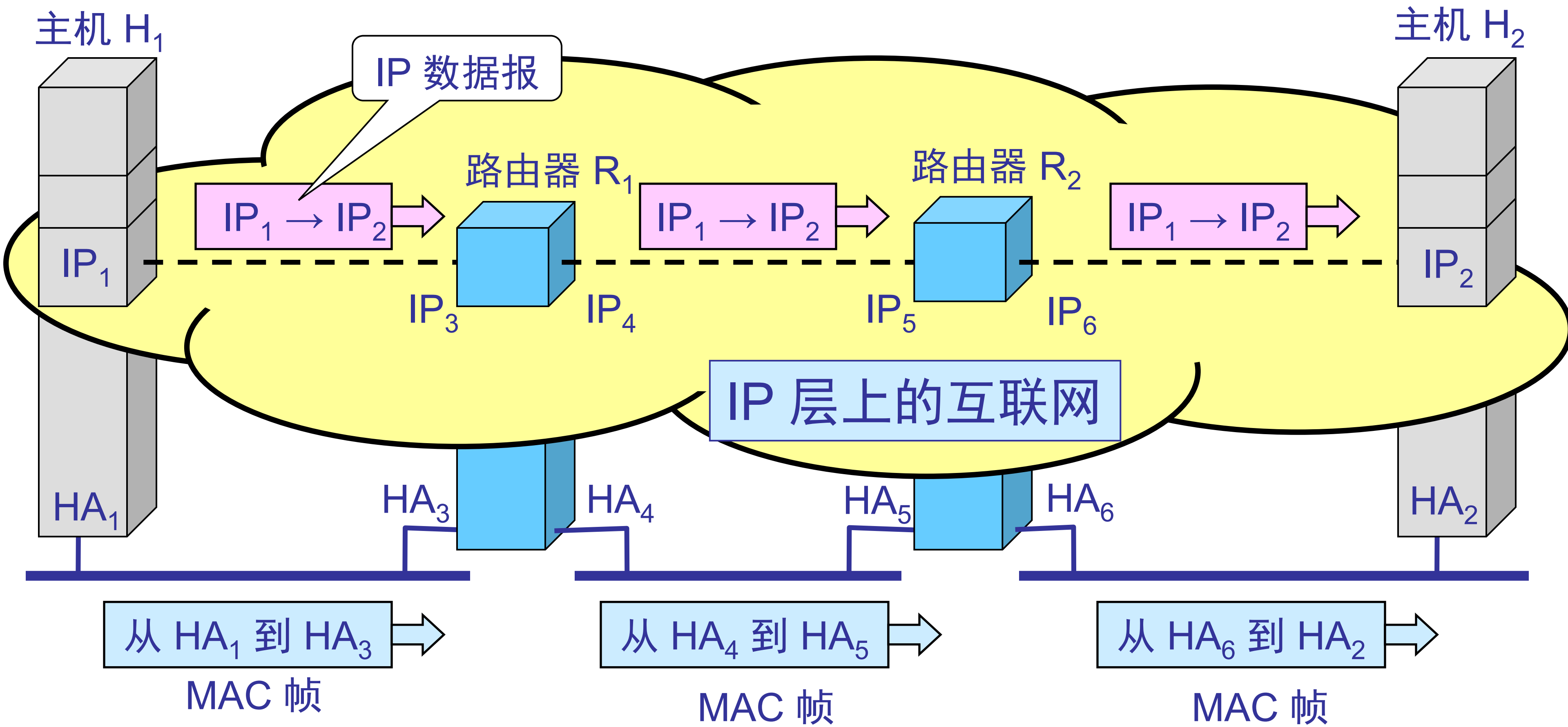
在链路上看 MAC 帧的流动



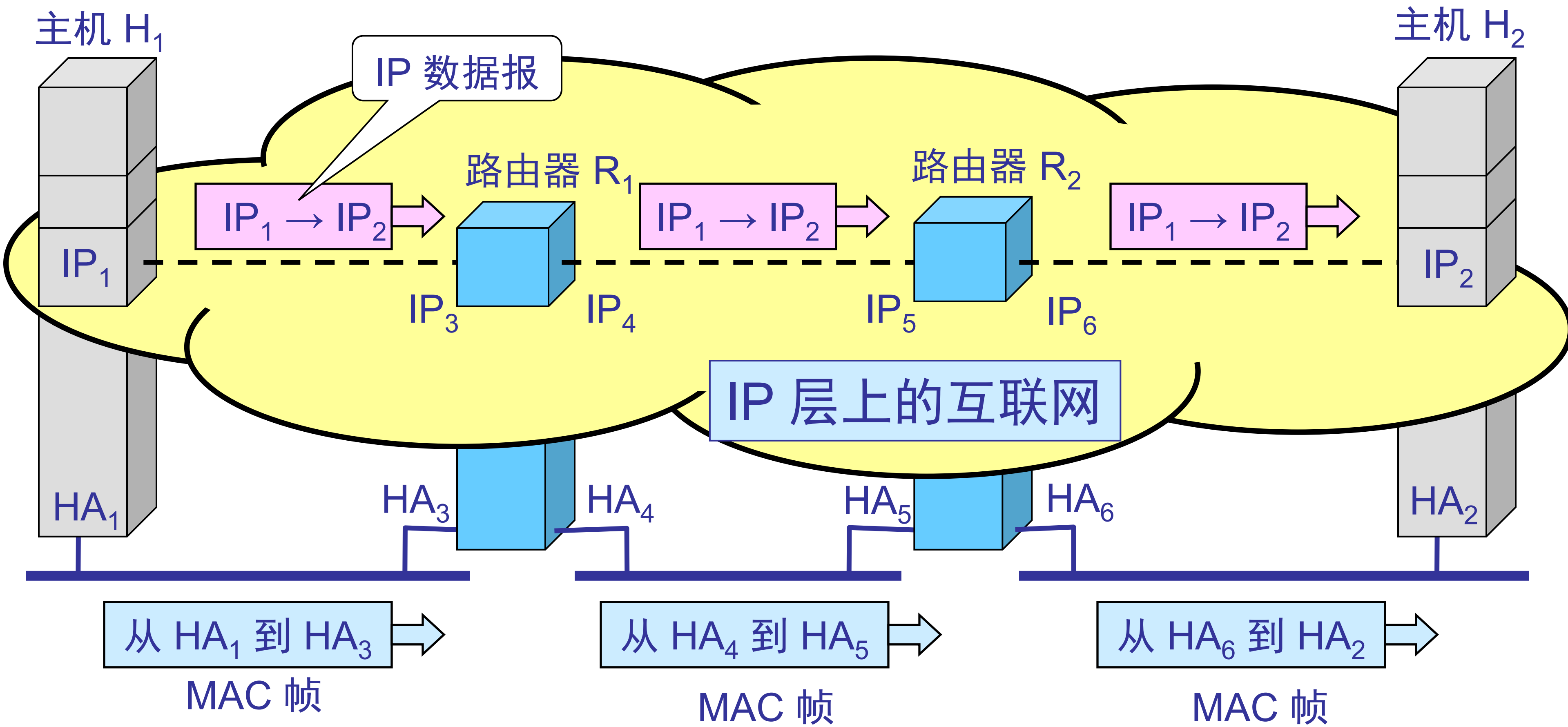
在 IP 层抽象的互联网上只能看到 IP 数据报

图中的 $IP_1 \rightarrow IP_2$ 表示从源地址 IP_1 到目的地址 IP_2

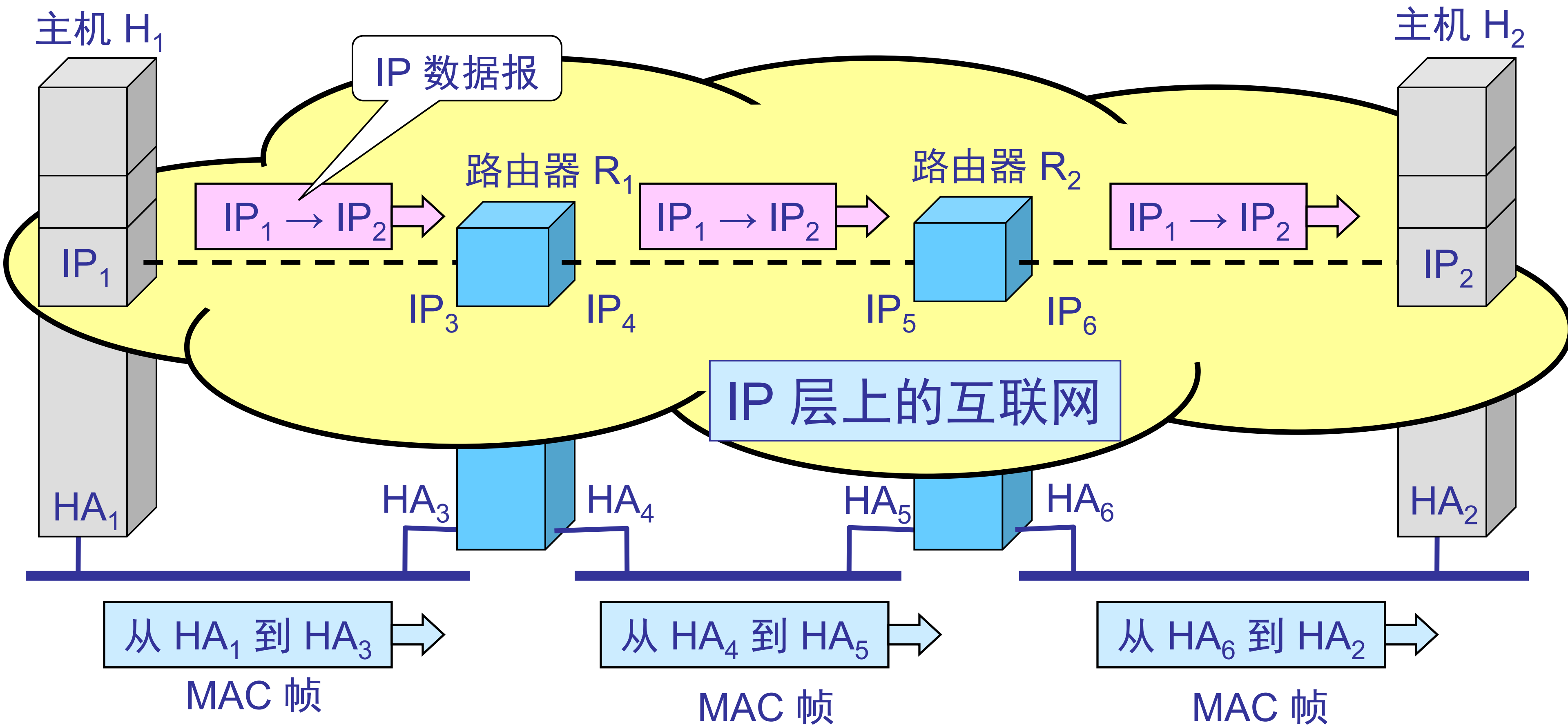
两个路由器的 IP 地址并不出现在 IP 数据报的首部中



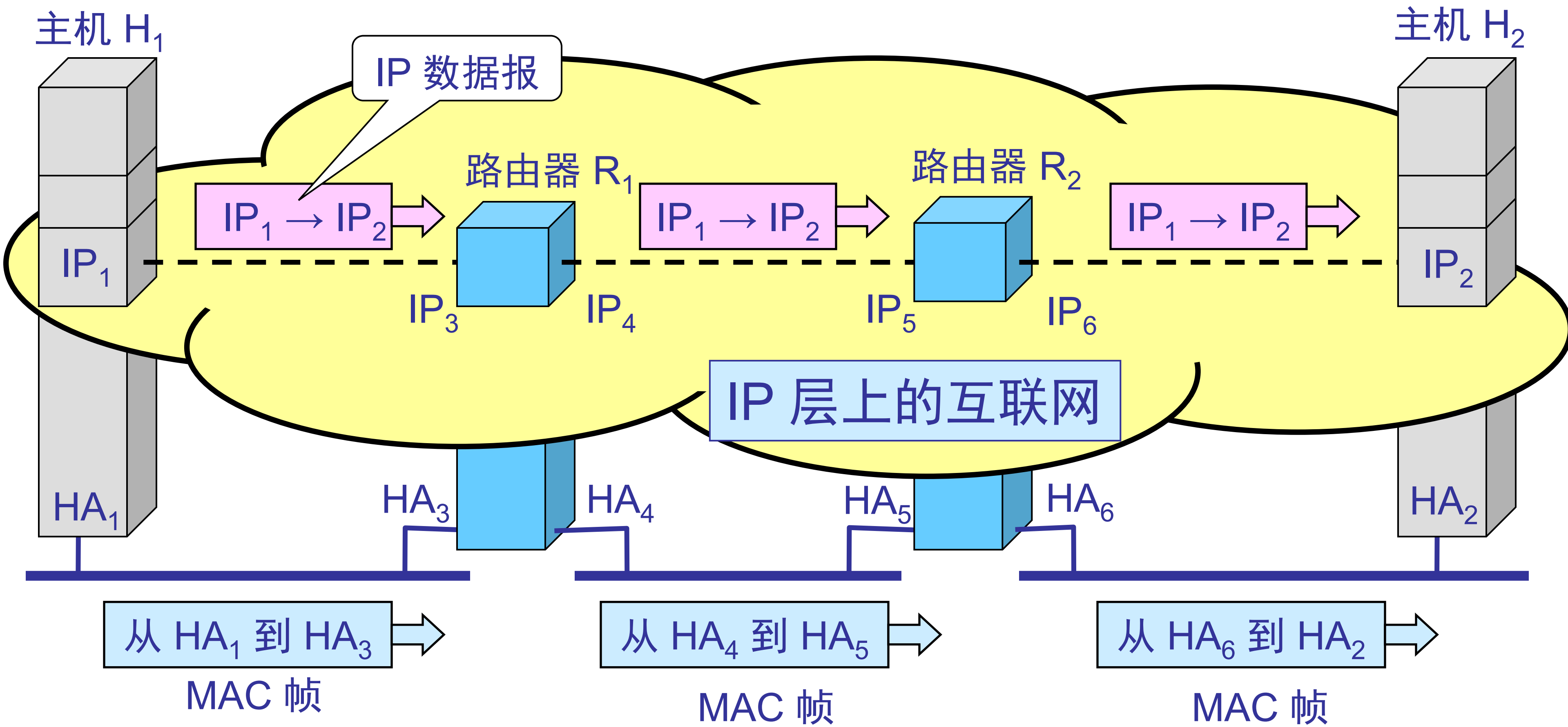
路由器只根据目的站的 IP 地址的网络号进行路由选择



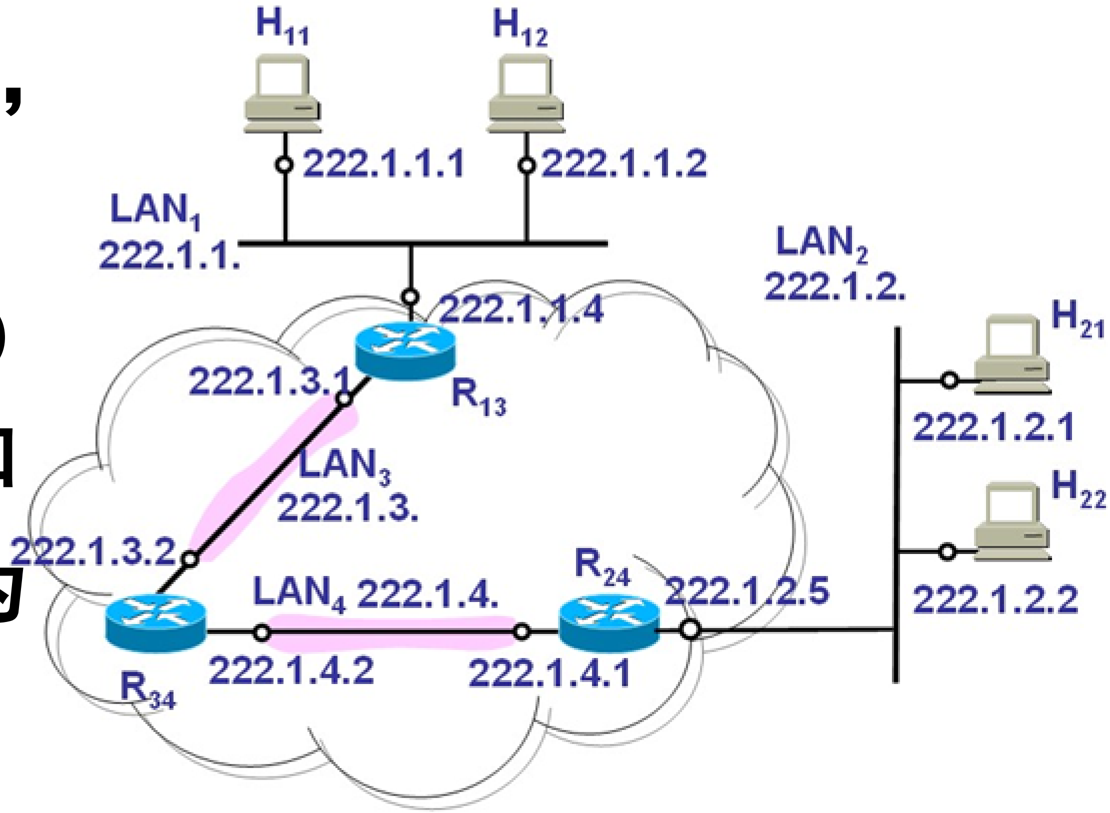
在具体的物理网络的链路层
只能看见 MAC 帧而看不见 IP 数据报



IP层抽象的互联网屏蔽了下层很复杂的细节
在抽象的网络层上讨论问题，就能够使用统一的、抽象的 IP 地址
研究主机和主机或主机和路由器之间的通信

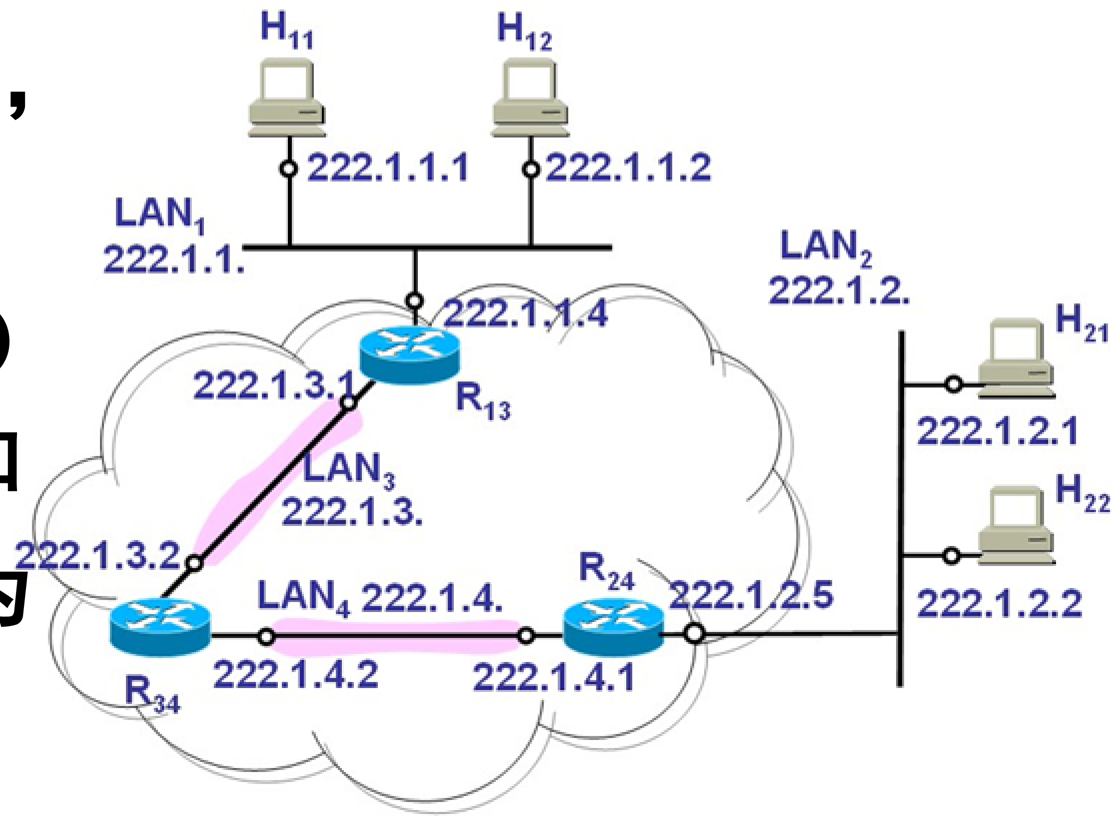


课堂练习：如图所示，四个局域网(LAN_1, LAN_2, LAN_3和LAN_4)通过路由器R_{13}, R_{24}和R_{34}互联。子网掩码为255.255.255.0



(1)试问主机H_{21}和H_{22}的默认网关地址是多少？

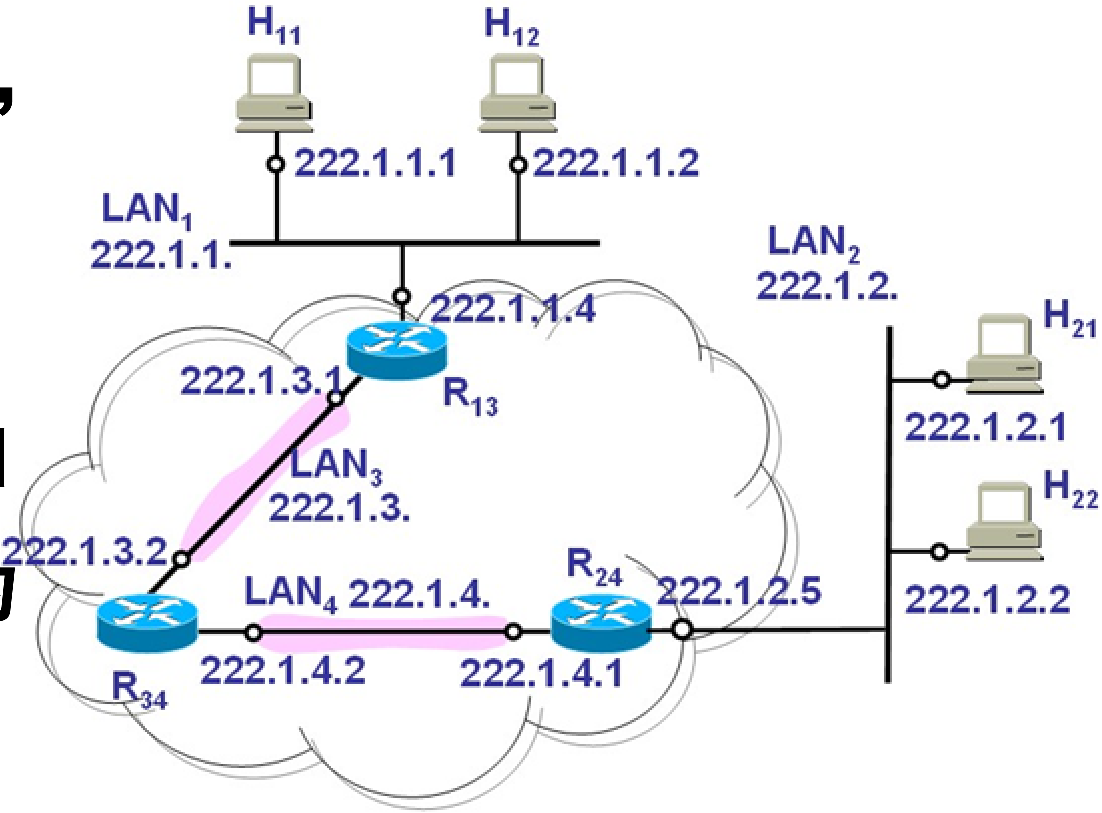
课堂练习：如图所示，四个局域网(LAN_1, LAN_2, LAN_3和LAN_4)通过路由器R_{13}, R_{24}和R_{34}互联。子网掩码为255.255.255.0



(1)试问主机H_{21}和H_{22}的默认网关地址是多少？

答：222.1.2.5

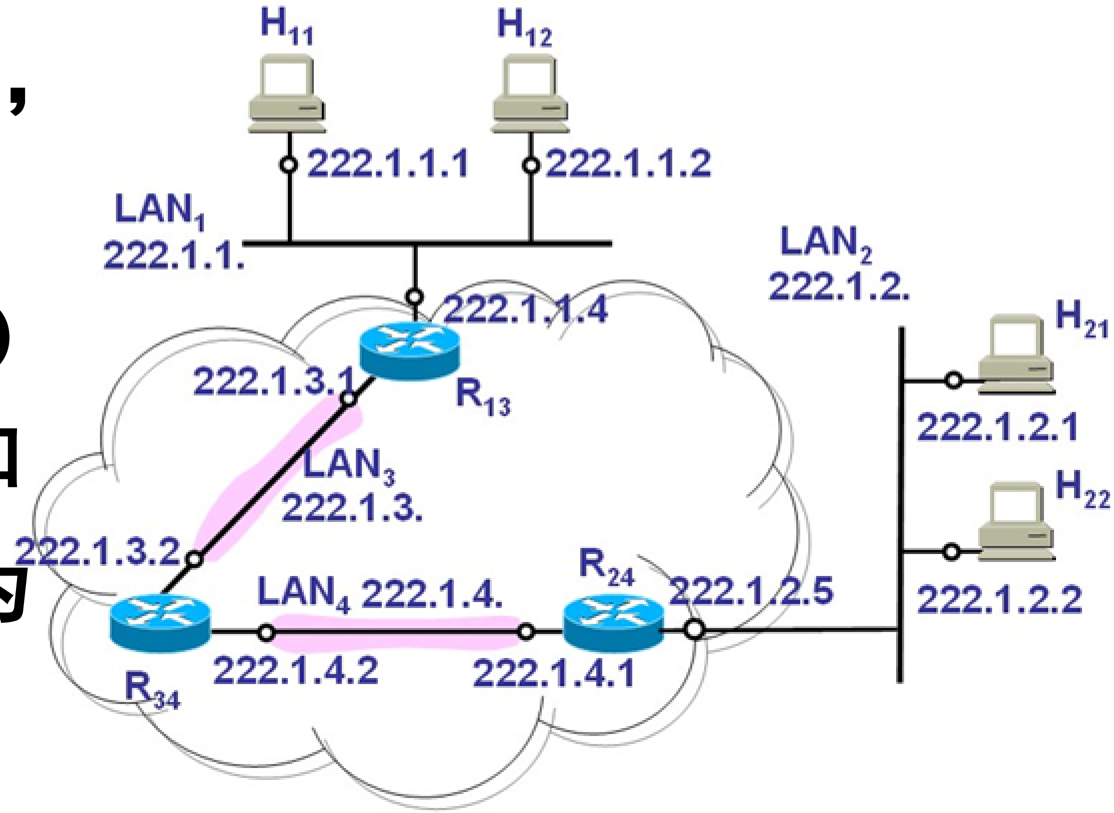
课堂练习：如图所示，四个局域网(LAN_1, LAN_2, LAN_3和LAN_4)通过路由器R_{13}, R_{24}和R_{34}互联。子网掩码为255.255.255.0



(2)请完成路由器R_{34}的路由表配置

目的网络	子网掩码	下一跳地址

课堂练习：如图所示，四个局域网(LAN_1, LAN_2, LAN_3和LAN_4)通过路由器R_{13}, R_{24}和R_{34}互联。子网掩码为255.255.255.0



(2)请完成路由器R_{34}的路由表配置

目的网络	子网掩码	下一跳地址
222.1.1.0	255.255.255.0	222.1.3.1
222.1.2.0	255.255.255.0	222.1.4.1

Network Layer Protocols in the Internet

- **IP (Internet Protocol)网际互连协议**
- **ICMP (Internet Control Message Protocol控制报文协议)**
- **ARP (Address Resolution Protocol)地址解析协议**
- **RARP (Reverse Address Resolution Protocol)逆地址解析协议**

The Transport Layer

- **The service of the transport layer is to provide a virtual end-to-end “message-pipe” for applications:**
  - ***connection-oriented***
  - ***connectionless***
  - ***reliable***
  - ***unreliable***

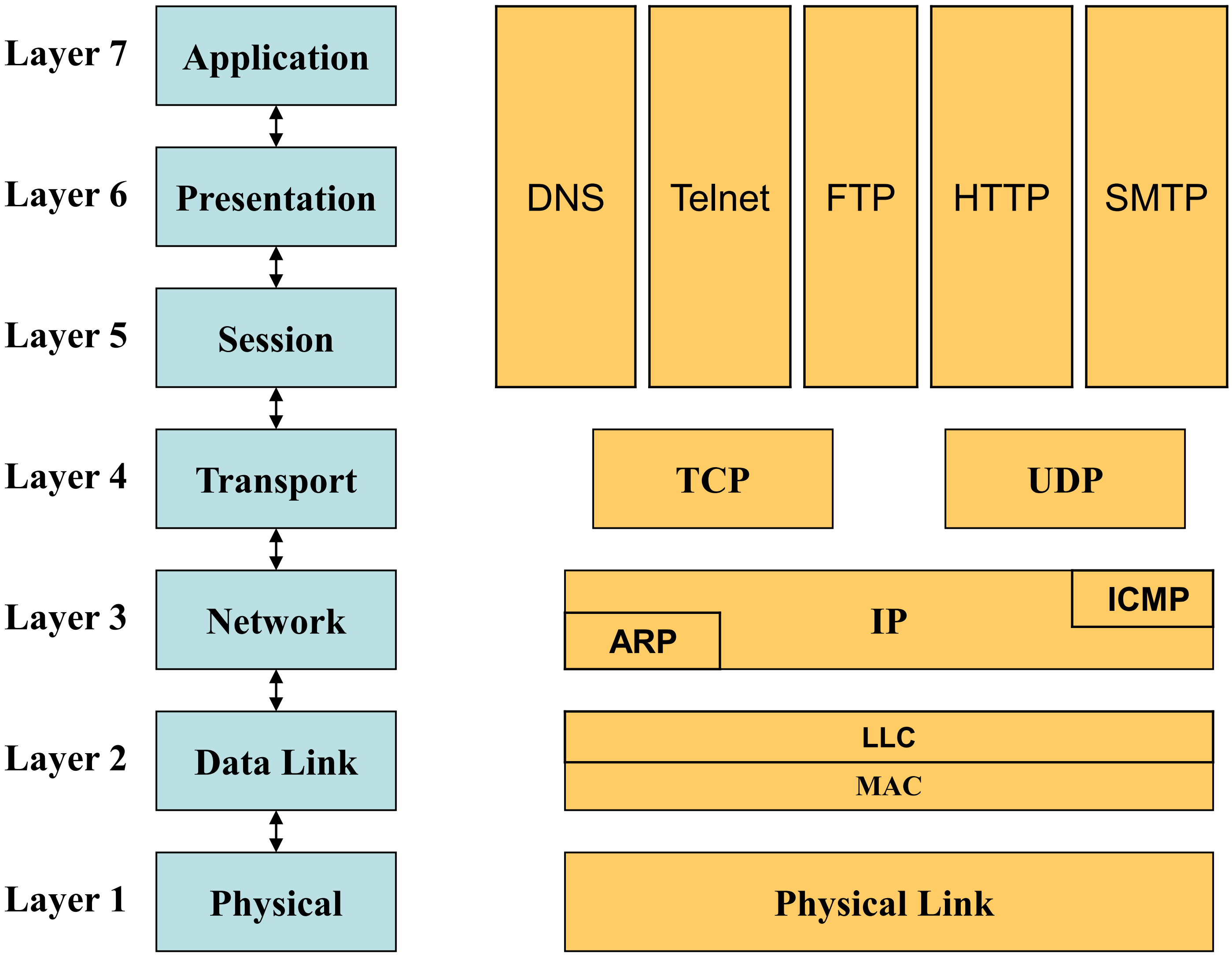
Transport Layer vs. Network Layer

- **The network layer provides communication between two hosts.**
- **The transport layer provides communication between two processes running on different hosts.**
- **Addressing: IP address vs. Port number**

Transport Layer vs. Data Link Layer

- **At the data link layer, two adjacent nodes相邻结点 (host-router or router-router) communicate directly via a physical link,**
- **At the transport layer, two transport entities 传输实体 within two different hosts communicate across the entire subnet.**

OSI Model versus TCP/IP



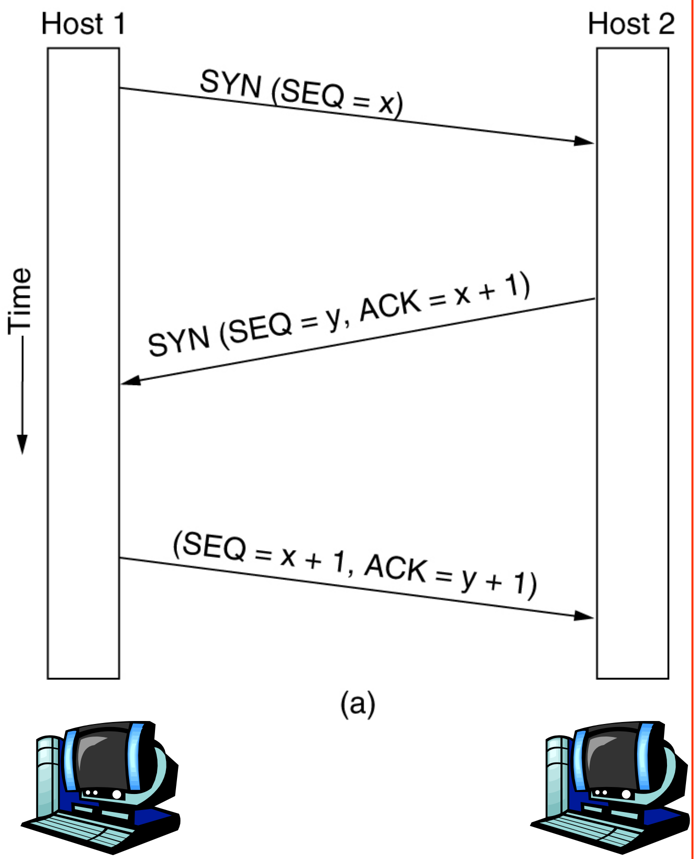
TCP (Transmission Control Protocol) 传输控制协议

- TCP provides point-to-point communi.
- TCP provides a reliable end-to-end byte stream over an unreliable IP network.
- TCP ensures that each segment is delivered correctly, only once, and in order.
- TCP is connection-oriented protocol

TCP (Transmission Control Protocol) 传输控制协议

- TCP provides full duplex communication, with Flow control and Congestion control
- TCP uses sliding window protocol滑动窗口协议 for flow control
- The Round-Trip Time (RTT往返时间) for TCP will change with different routes from the source host to the destination

TCP Connection Establishment



Three way handshake:

Step 1: client host sends TCP SYN segment to server with initial seq number, but no data

Step 2: server host receives SYN, replies with SYN/ACK segment

- server allocates buffers
- specifies server initial seq. no.

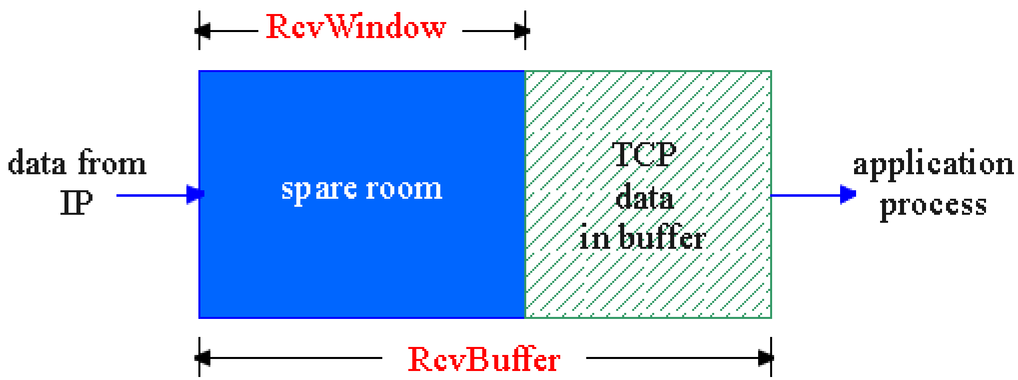
Step 3: client receives SYN/ACK, replies with ACK segment, which may contain data

Flow Control流量控制

- **Flow control is a technique for preventing the sender from overwhelming the receiver with "data"流量控制是一种技术用于防止发送方用"数据"淹没接收方。**
  - **A receiver reserves some buffer space for storing data from a sender, while the data is being processed.**
  - **If the sender sends data faster than the receiver can process it, then buffer overflow will occur**

Flow Control流量控制

The receive side of TCP connection has a receive buffer:

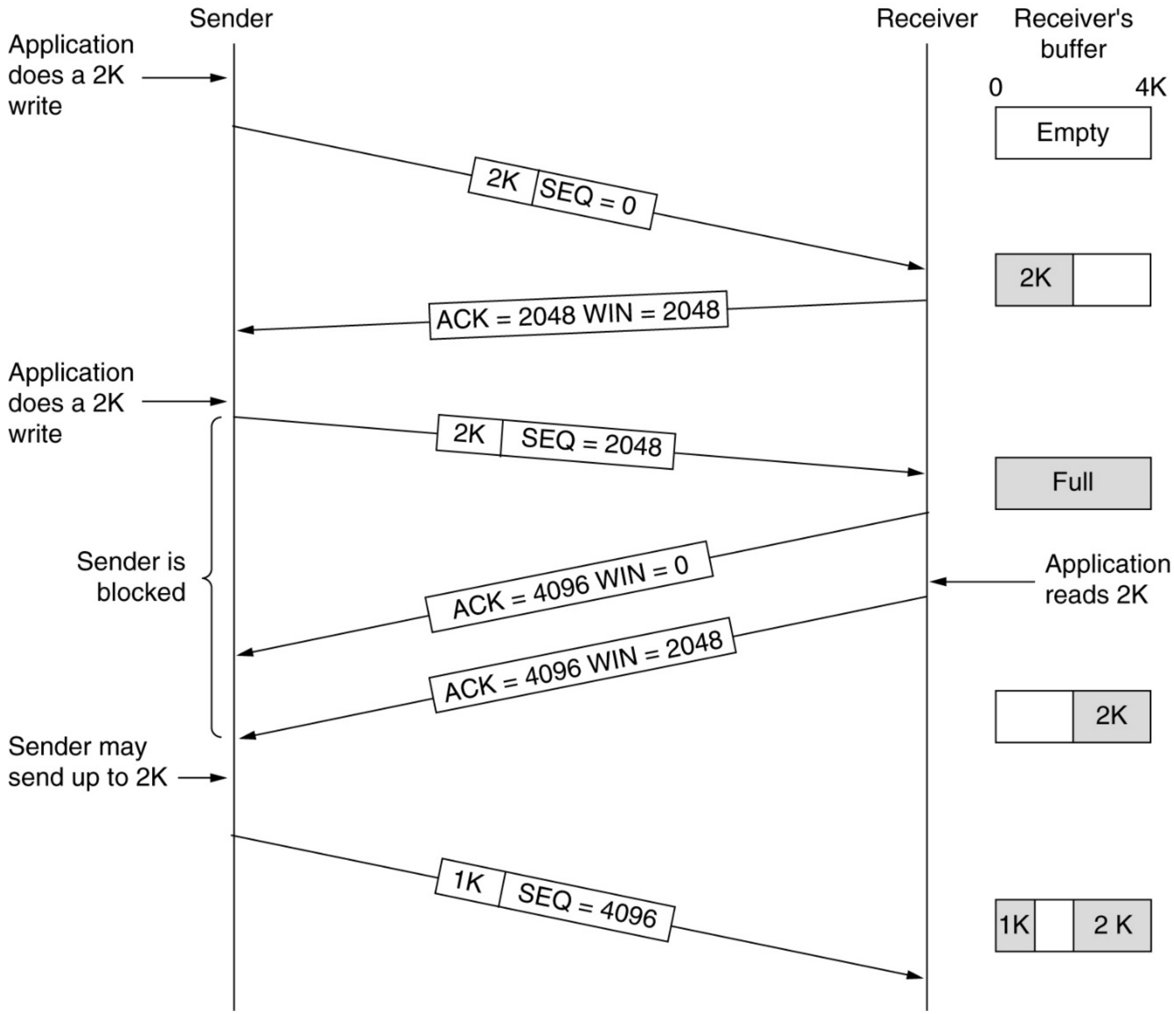


The application process may be slow at reading from the buffer

flow control
sender won't overflow receiver's buffer by transmitting too much, too fast

speed-matching service: matching the send rate to the receiving app's drain rate

TCP Transmission Policy传输策略



Window management in TCP.

例题1

- 主机甲与主机乙之间已建立一个TCP连接，双方持续有数据传输，且数据无差错与丢失。若甲收到1个来自乙TCP报文段，该段的序号为1913，确认序号为2046，有效载荷为100字节，则主机甲立即发送给主机乙的TCP报文段的序号和确认号分别是多少？

例题1,答案

解析：

- 若甲收到1个来自乙TCP报文段，该段的序号seq=1913，确认序号ack=2046，有效载荷为100字节，
- 则主机甲立即发送给主机乙的TCP报文段
  - 序号seq1=ack=2046
  - 确认号ack1=seq+100=2013

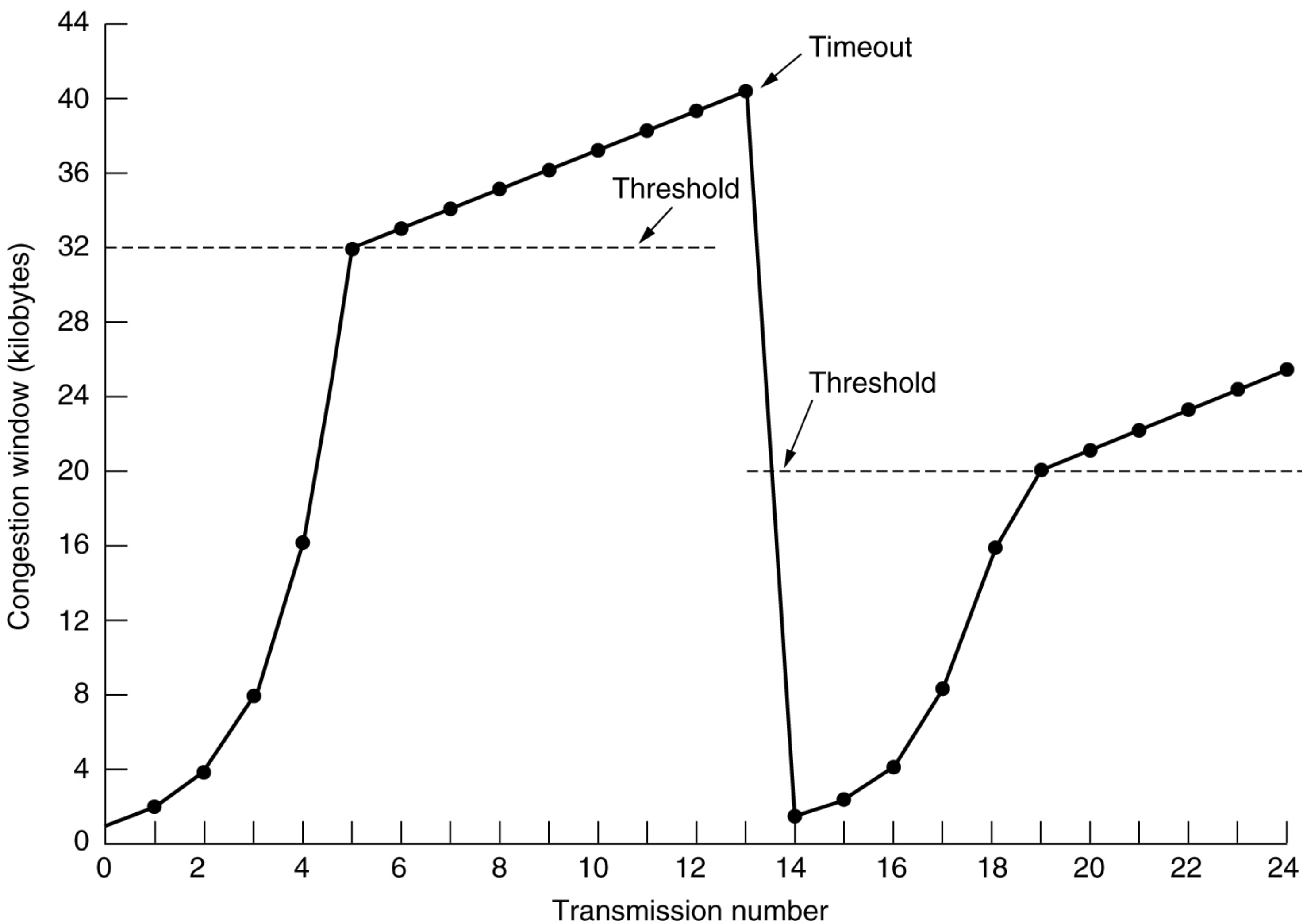
Flow Control流量控制

- **Flow control at both the Data Link Layer and the transport layer**
  - **Stop and Wait Protocol (停止等待协议)**
  - **ARQ(Automatic Repeat reQuest)**
  - **Sliding Window Protocols(滑动窗口协议)**

TCP Congestion Control 拥塞控制

- Congestion: “too many sources sending too much data too fast for network to handle”
- The purpose is to limit senders as needed to ensure load on the network is “reasonable”.
- Solution :
  - Slow Start (SS)
  - Congestion Avoidance (CA)
  - Fast Retransmit (FR)

TCP Congestion Control 拥塞控制



课堂练习

如果主机A向主机B发起一个TCP连接，最大段长MSS=1KB，RTT=5ms，主机B开辟的接收缓存为64KB。则主机A从连接建立成功至发送窗口达到32KB，至少需要经过多少时间？

课堂练习

答案：从TCP连接建立好开始，主机A的发送窗口初始值为1个MSS段。

在*slow start*阶段按照指数规律增长：

1、2、4、8、16、32，……

经过5个RTT后，发送窗口增长到32个MSS段，即32KB。因此 5xRTT=25ms

TCP Congestion Control 拥塞控制

- **Gently probe逐渐探测 network for spare capacity (SS+CA慢启动+拥塞避免)**
- **Drastically reduce rate on congestion冲突发生迅速降低速率**
- **Retransmission on timeout超时重传**
- **Detecting Packet Loss/Fast Retransmit丢包检测/快速重传**
- **Fast Recovery快速恢复**

TCP Congestion Control 拥塞控制

Fast Retransmit + Fast Recovery

- Wait for a timeout is quite long !
- using duplicate ACKs to signal lost packet.
- Upon receipt of three duplicate ACKs, the TCP Sender retransmits the lost packet right away!

TCP Transmission Policy传输策略

- When the TCP entity at sender always transmit data as soon as they come in from the application layer, the performance may be very poor!
- The worst case: telnet connection to an interactive editor(交互式编辑器), *a TCP segment=header (20 bytes) + data payload (1 byte for one character)*

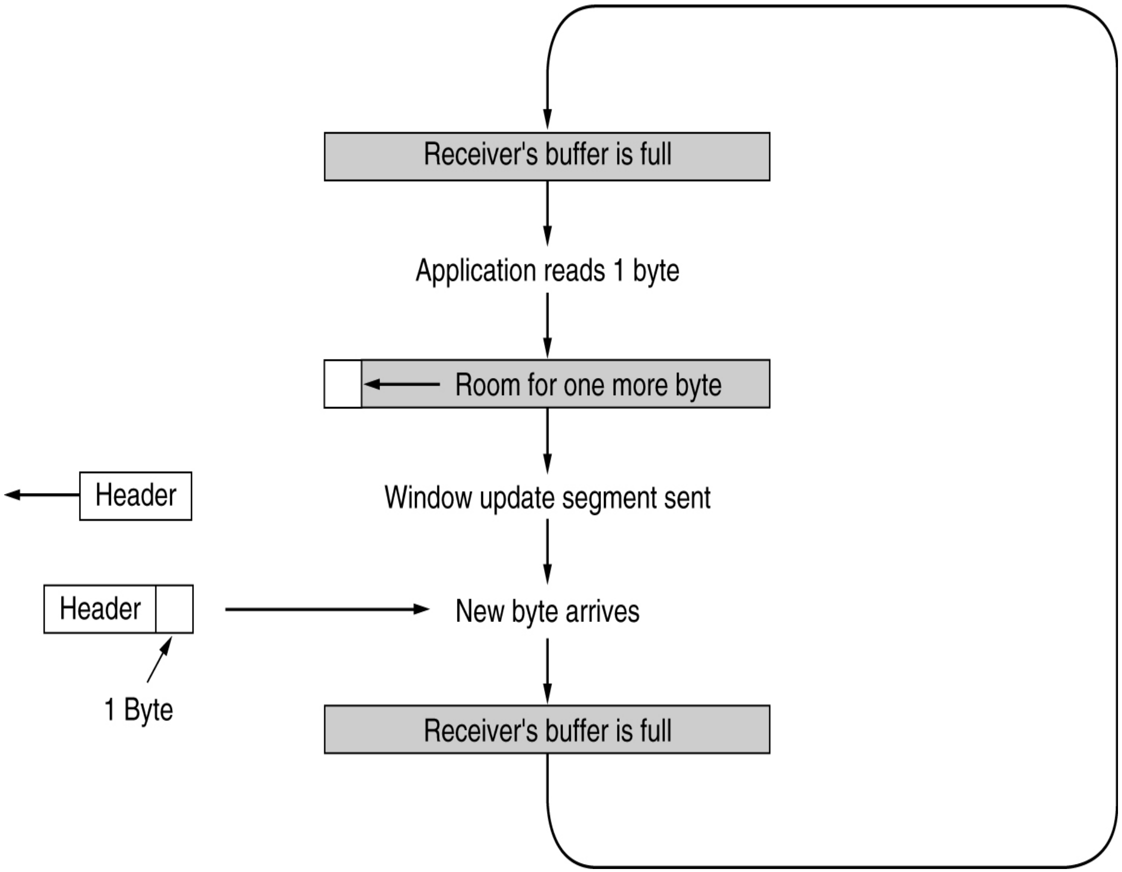
TCP Transmission Policy传输策略

- **In order to improve performance, the TCP buffers data for a moment first, then send out**
- **Nagle's algorithm (Nagle, 1984):** ***when data come into the sender one byte at a time, just send the first byte and buffer all the rest until the outstanding byte is acknowledged. Then send all the buffered characters in one TCP segment and start***

NAGLE, J.: "Congestion Control in TCP/IP Internetworks," Computer Communication Review, vol. 14, pp. 11-17, Oct. 1984

TCP Transmission Policy传输策略

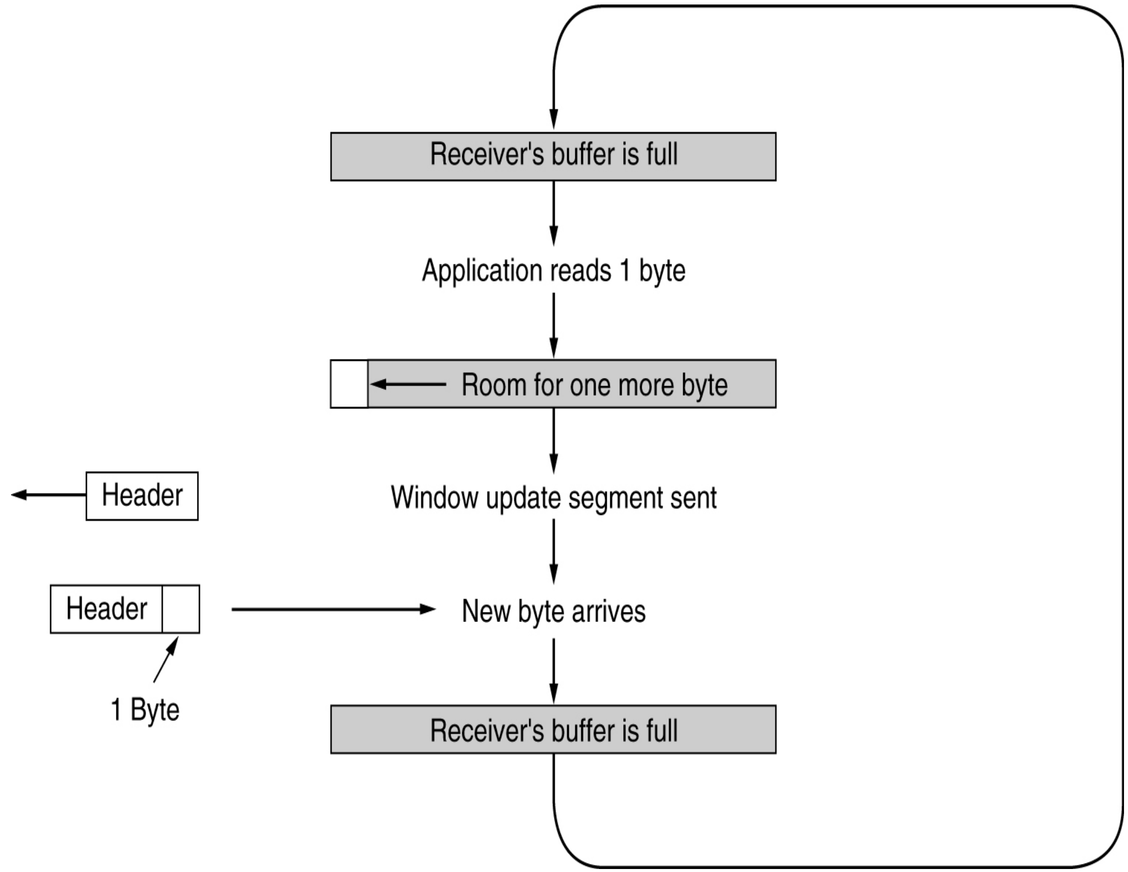
In 1982, Clark realized another problem that can degrade TCP performance, so called **Silly Window Syndrome(傻瓜窗口综合症)**



CLARK, D.D.: "Window and Acknowledgement Strategy in TCP," RFC 813, July 1982

TCP Transmission Policy传输策略

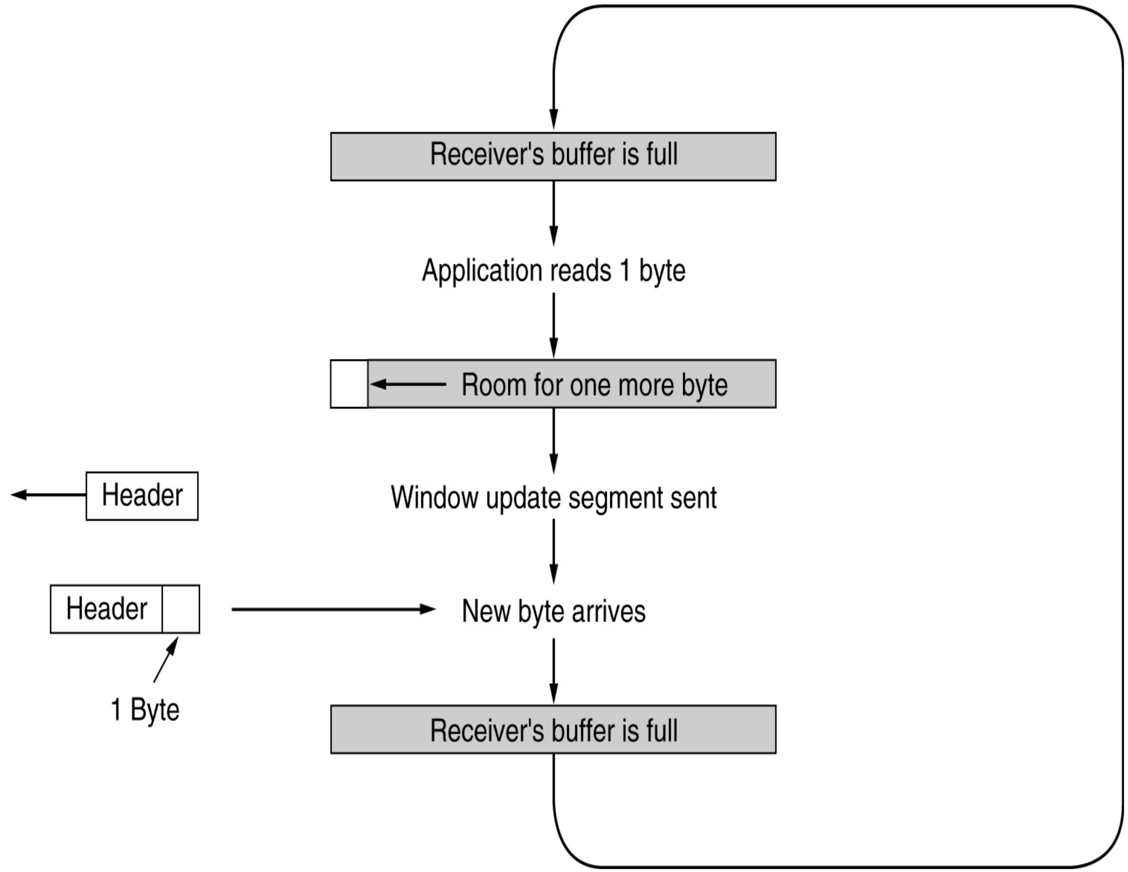
The "Silly Window" problem occurs when data are passed to the sending TCP entity in large blocks, but an interactive application on the receiving side reads data 1 byte at a time.



TCP Transmission Policy传输策略

Clark's solution is to prevent the receiver from sending a window update for 1 byte,

i.e. the receiver should not be too sensitive接收方不要太敏感！



TCP Transmission Policy传输策略

- **Nagle‘s algorithm** and **Clark’s solution** to the silly window syndrome are complementary互补的.
- **The goal is for the sender not to transmit small segments and the receiver not to ask for them.**

TCP Transmission Policy传输策略

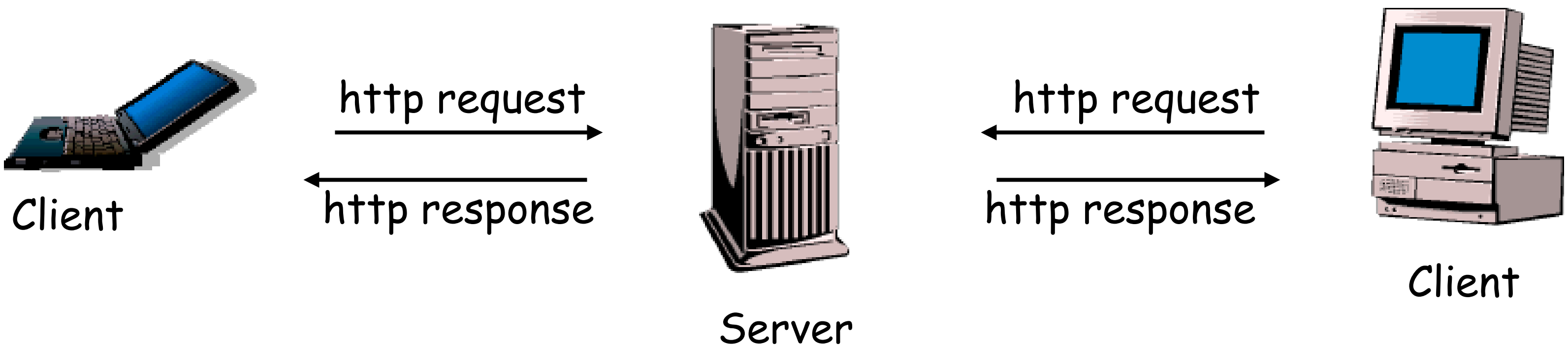
- Solution for sender side: wait until sender has enough data to transmit – “Nagle’s Algorithm”
- Clark was trying to solve the problem of the receiving application sucking the data up from TCP a byte at a time.

UDP (User Datagram Protocol) 用户数据报协议

- **As a connectionless transport protocol, it does not have to establish a connection to another process at the destination host before sending data.**
- **UDP has no flow control, no error control, no acknowledgements, and no mechanism to request retransmissions.**
- **There is no way to guarantee that the segment will reach its destination.**

The Application Layer

- **WWW and HTTP**
- **Client/Server (C/S) model**



port number: 80

Some "Web" Terminology术语

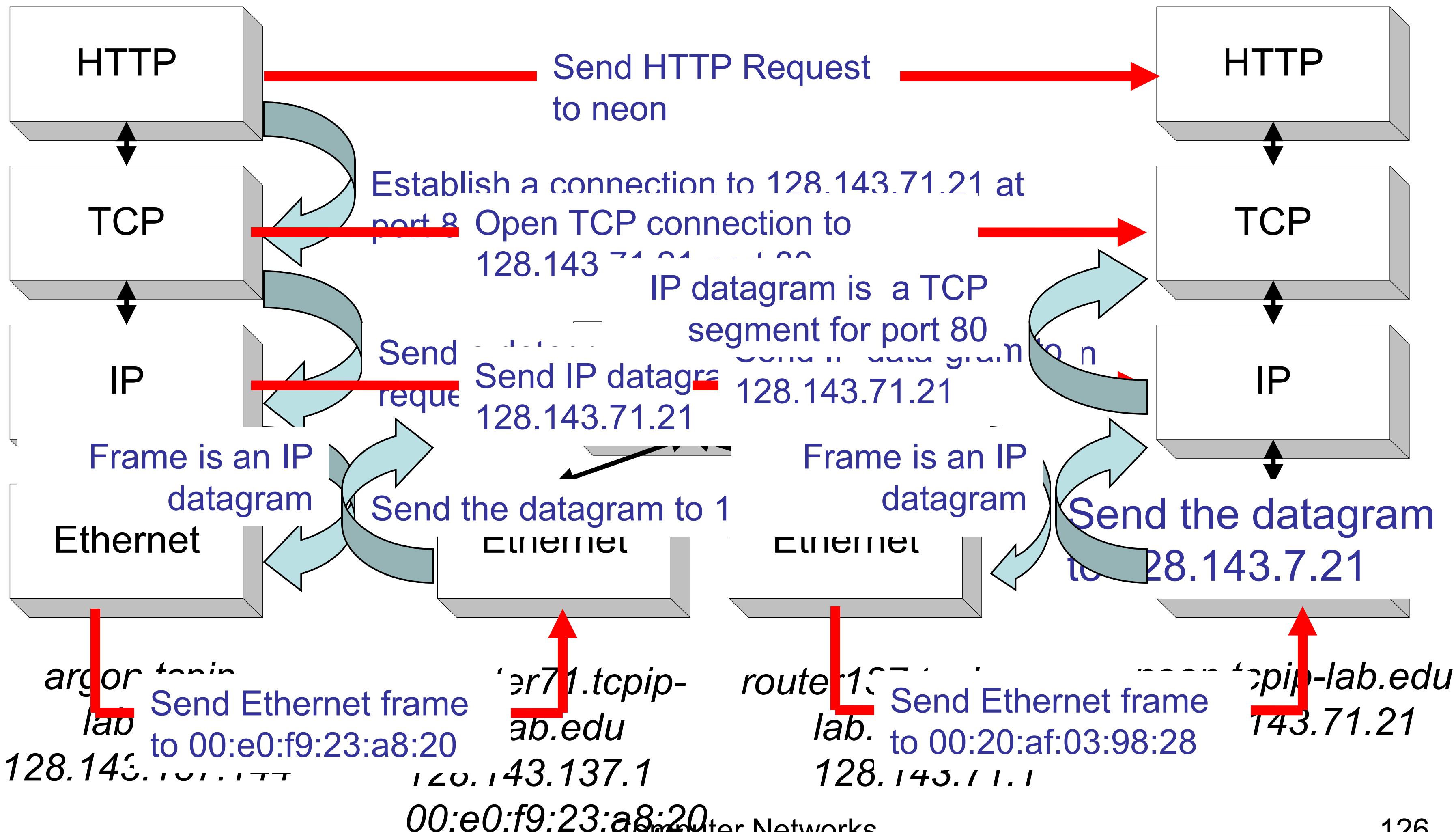
- Web page may contain links to other pages (sometimes also called Web Objects对象)
- Object can be HTML file, JPEG image, Java applet, audio file,…
- Each object is addressable by a URL (Uniform Resource Locaters通用资源定位器*):

```
http://www.someschool.edu/someDept/pic.gif
```

protocol　　host name　　path name

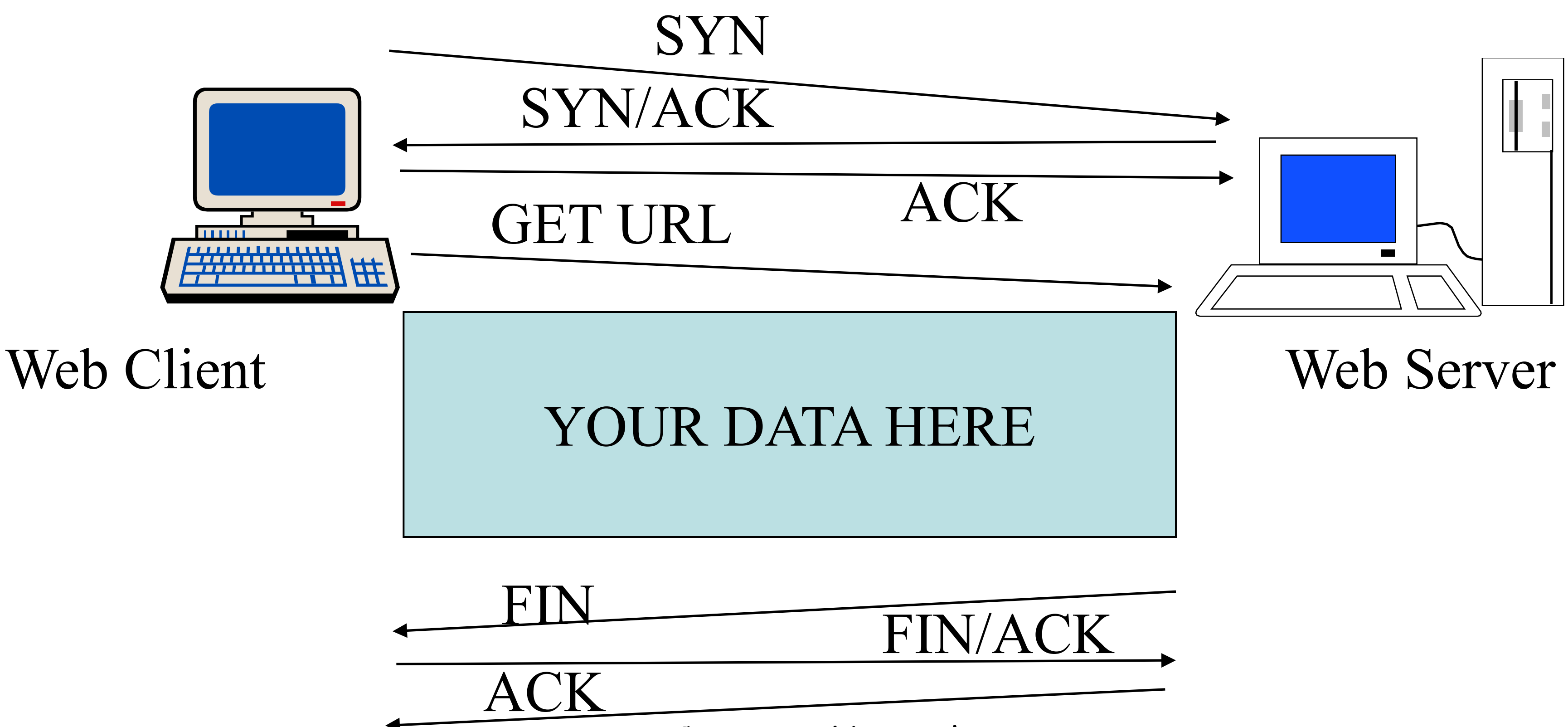
*URL：识别Internet上的文档或资源的一种标准化方法。

Layers in the Example



Network View: HTTP and TCP

- TCP is a connection-oriented protocol



HTTP connections

Non-persistent HTTP

- At most one object is sent over a TCP connection.
- HTTP/1.0 uses non-persistent HTTP

Persistent HTTP

- Multiple objects can be sent over single TCP connection between client and server.
- HTTP/1.1 uses persistent connections in default mode

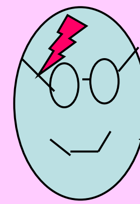
持久连接（**persistent connection**）

– Pipelined

Example Web Page

Harry Potter Movies

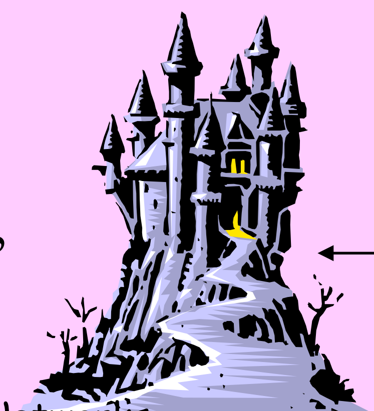
As you all know,
the new HP book
will be out in June
and then there will
be a new movie
shortly after that…



hpface.jpg

page.html

"Harry Potter and
the Bathtub Ring"

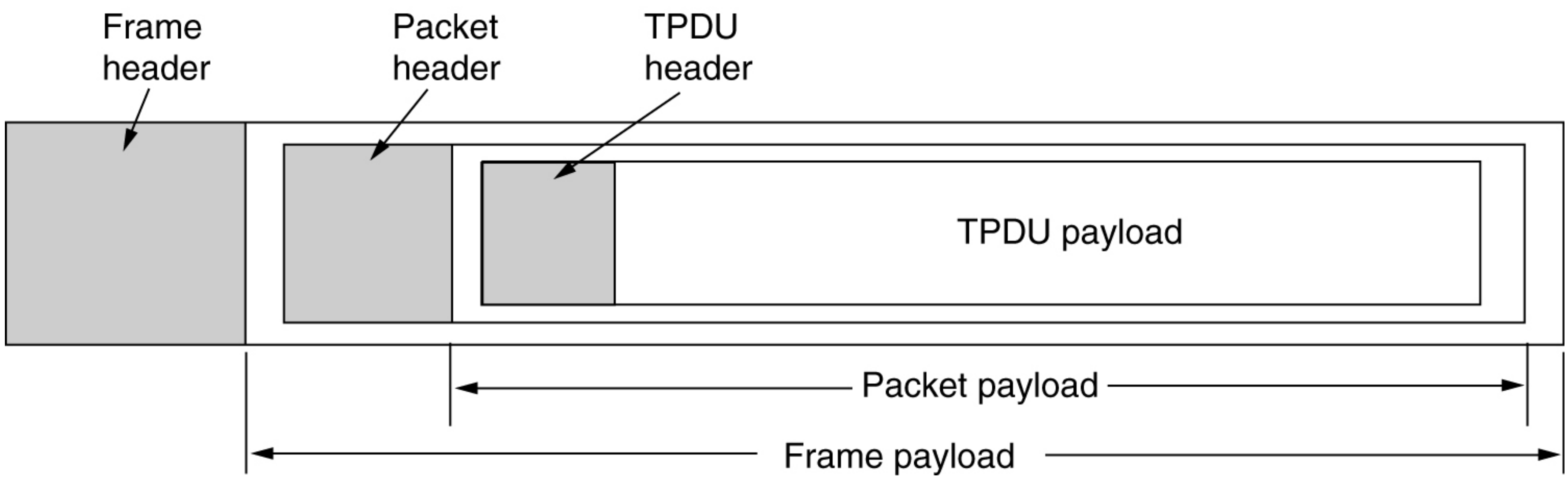


castle.gif

The Application Layer

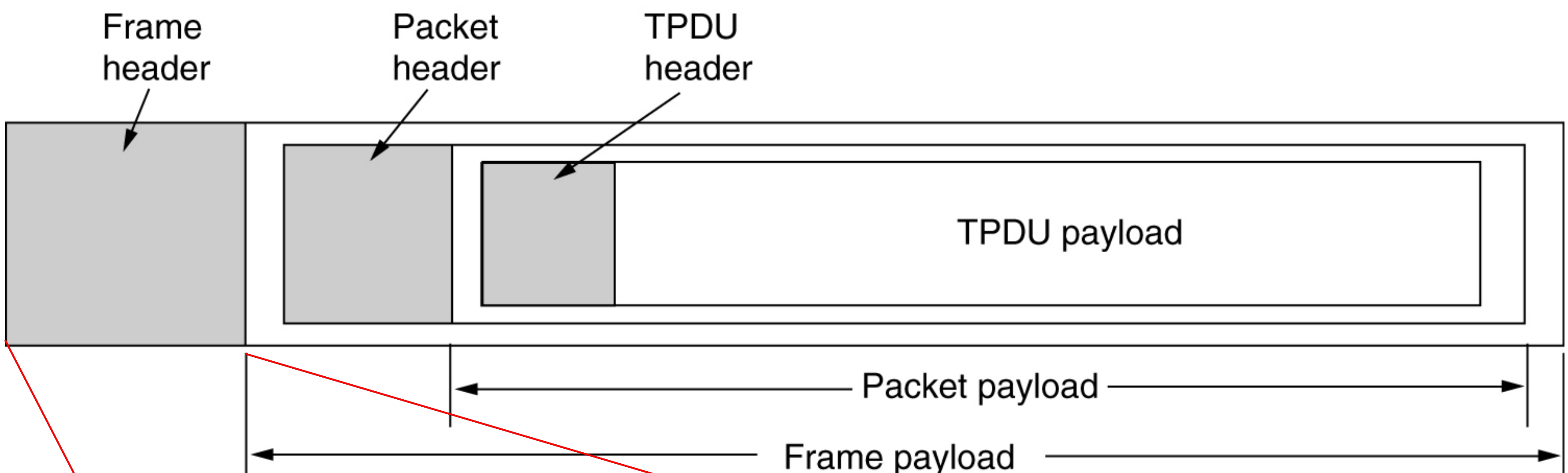
- **Name Servers名字服务器** is used to translate host names to IP Addresses: i.e. www.google.com → 192.168.11.11
- **Electronic Mail电子邮件**
  - SMTP: Simple Mail Transfer Protocol is a protocol used for sending mails
  - POP3: Post Office Protocol Version 3 is a protocol used for receiving mails
  - Port Number: 25

各层数据传输单元



Frame header
Packet header
TPDU header
TPDU payload
Packet payload
Frame payload

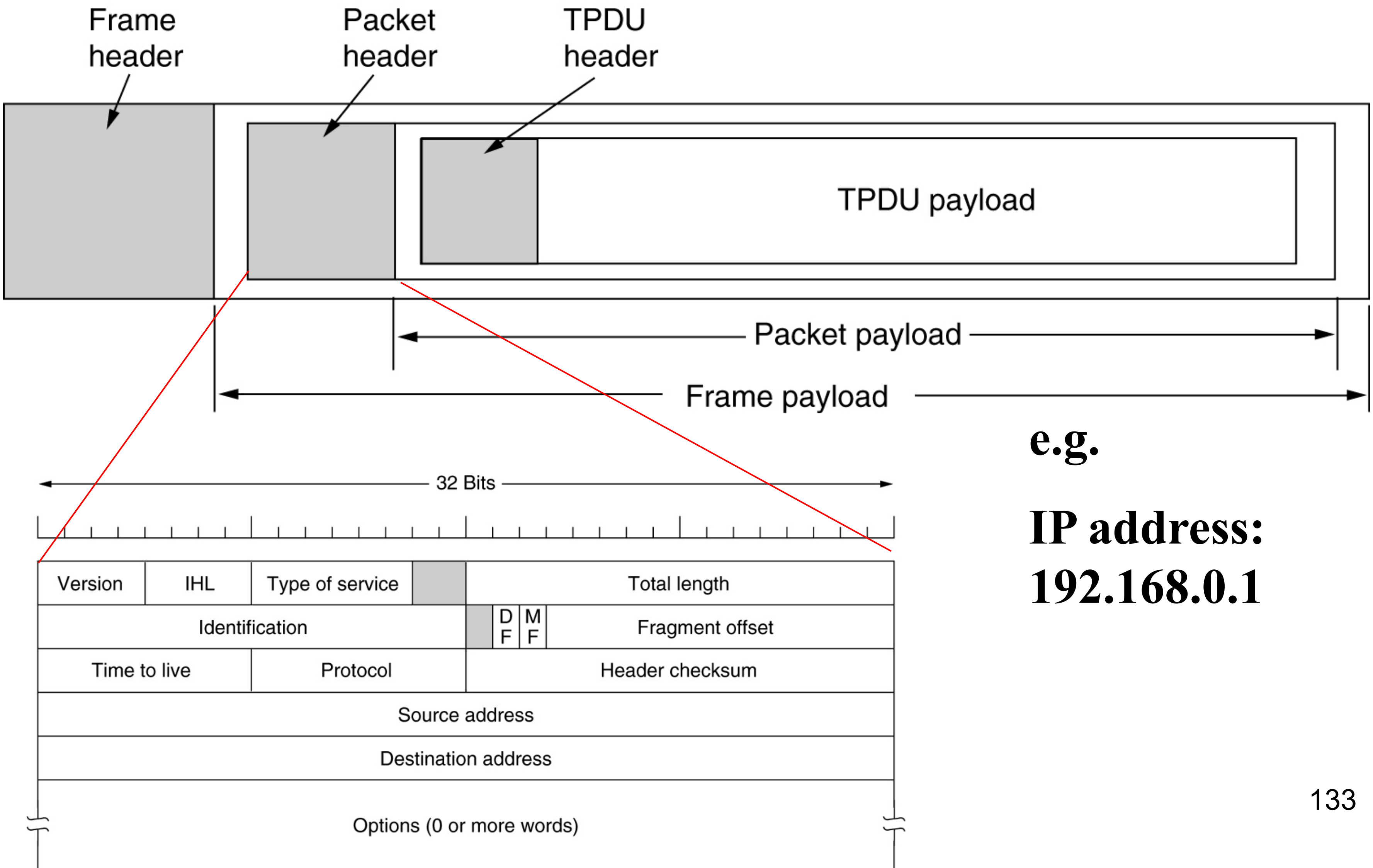
各层数据传输单元



e.g. MAC address: 00:D0:C9:35:56:11

Bytes	8		6	6	2	0-1500	0-46	4
(a)	Preamble		Destination address	Source address	Type	Data	Pad	Check-sum
(b)	Preamble	SoF	Destination address	Source address	Length	Data	Pad	Check-sum

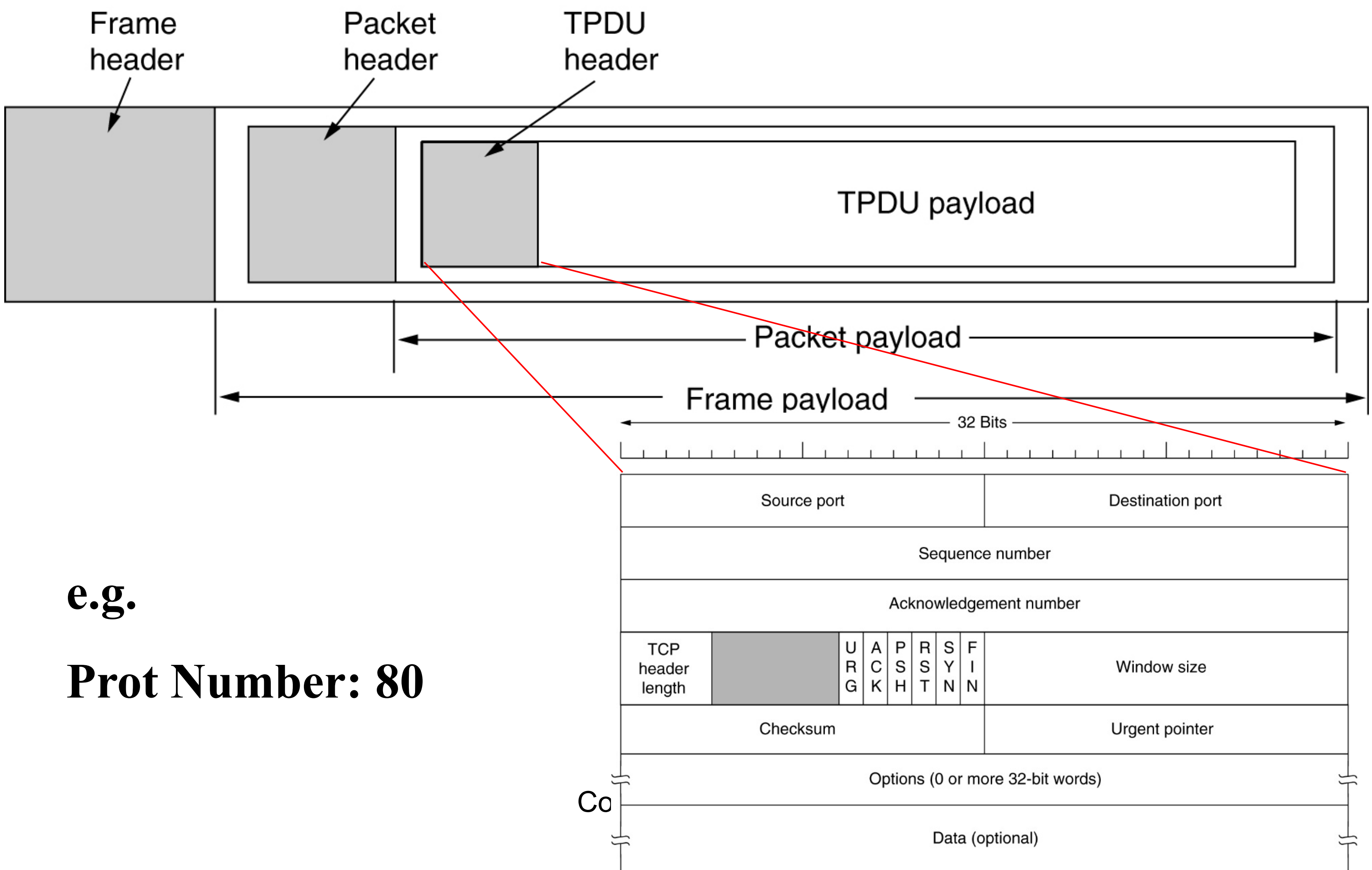
各层数据传输单元



e.g.

IP address: 192.168.0.1

各层数据传输单元



e.g.

Prot Number: 80

Co

The End!

- **Good Luck with your final examination !**
- **Let's keep in touch !**

hshen@njtech.edu.cn