Contents

I Introduction

1. Network Technology: from local to global

1.1 Network Hardware

依据传输技术 (transmission technology) 与规模 (scale) 分类.

传输技术分两种:

- Broadcast links (Multicasting 多播)
  链路被所有机器共享 (有线 + 无线, 长距离只能用有线)
  无线网络是典型的 broadcast links
  地址全部是 1 表示将数据包寻址到所有目的地地址字段 (向所有人发信息), 地址全是 0 表示初始/未分配地址.
- Point-to-point links
  有很多路径, 选择基于路由算法.

依据规模分类: 距离是重要指标.

Interprocessor distance	Processors located in same	Example
1 m	Square meter	Personal area network
10 m	Room	Local area network
100 m	Building	
1 km	Campus	
10 km	City	Metropolitan area network
100 km	Country	Wide area network
1000 km	Continent	
10,000 km	Planet	The Internet

Figure I.1: scale

Communication Links e.g.

- Coaxial cable (telephone)
- Copper wire (TV)
- Fiber optics (the backbone the PSTN, Public Switched Telephone Network)
- Radio spectrum (cellphone)

Different links can transmit data at different rates, with **the transmission rate** of a link measured in bits/second. 传输速率以 bits/second 衡量. 每个 bits 的持续时间 (duration, seconds/bit) 就是传输速率的倒数.

高的传输速率易被噪声影响, 比如持续 10ms 的噪声, 影响 100bits/s 的 2bits, 1kbits/s 的 11btis.

Personal Area Network(PAN) e.g. Bluetooth networks use the master-slave paradigm.

Loacl Area Network(LAN) 是 broadcast link, 有问题, 处理分两种:

- Static: time slot (TDM) and FDM, 时分 or 频分.
- Dynamic allocation methods for a common channel are either centralized and decentralized. 动态分配, 要么中心化, 要么去中心化.

Home Networks 以后可能会有, 家庭网络.

Metropolitan Area Network(MAN) e.g.

- Wired: cable television
- Wireless: IEEE 802.16 (WiMax), telephone network

有两种方式:

- VPN (Virtual Private Network) 自建
  优点: reuse of resource 资源重利用
  缺点: 调用底层资源时受限
- ISP (Internet Service Provider) 租用资源

2. Examples of Networks

2.1 The Internet

鼻祖: The ARPANET, 后面 NSFNET, 然后 Internet.
层级结构网络不安全, 鲁棒性低. 分布式网络更安全.

ARPANET 诞生于 1969, 分两块 subnet 与 host.

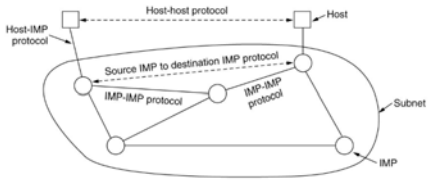

Figure I.2: ARPANET

NSFNET The path a packet takes through the internet depends on the peering choices of the ISPs.

1) 最大的 ISP: Tier 1 ISP
2) 下级的 ISP
3) 家庭之类的, DSL (Digital Subscriber Line) 宽带?

The Internet's Architecture 从层级到扁平

2.2 Mobile networks

- E-UTRAN (Evolved UMTS Terrestrial Radio Access Network)
- EPC (Evolved Packet Core)
  HSS (Home Subscriber Server)
  P-GW (Packet Data Network Gateway): 直连 internet
  S-GW (Serving Network Gateway)

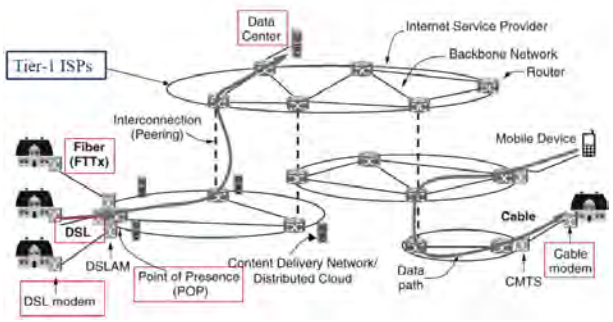

Figure I.3: the Internet’s Architecture

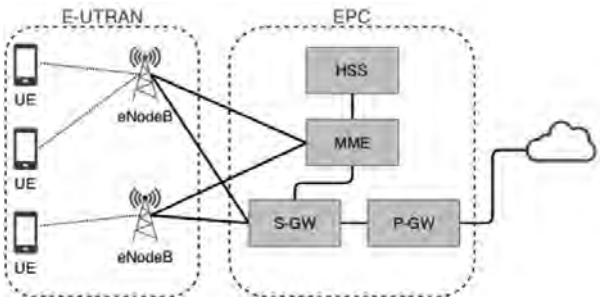

Figure I.4: Mobile Networks

Frequency reuse (频率重用) 无线电是稀缺资源, 需要政府 license 进行使用. 把通信区域分块, 设计为蜂窝网络. 理论依据: Cellular Concept

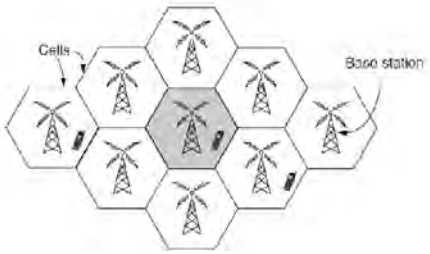

Figure I.5: Frequency reuse

但 3G 可以每一个 cell 用所有频率.

Handover 不同基站的转移.

SIM (Subscriber Identity Module) a removable chip

Evolution of Mobile Networks 主要区别: 依据频谱分类.

1) First-Generation (1G)
2) Second-Generation (2G)
3) Third-Generation (3G)
4) 4G
5) 5G

Packet Switching vs. Circuit Switching

- Packet switching: 数据包来选择路径
- Circuit switching: 建立路径传输数据

2.3 Wireless networks (802.11 WiFi)

为 LAN 制定的标准, 限制传输功率 (transmit power) 以允许不同的设备共存.

APs (Access Points) 链接有线网, 通信通过 APs 进行.

Multipath fading 传输的多个回波可能沿着不同的路径到达接收器. 回波可以相互抵消或增强, 导致接收信号大幅波动.

解决方法: path diversity, 使用多条独立路径传输信息, e.g.

- 在允许的频带内使用不同的频率
- 使用不同天线, 可以从不同的方向接受信号 (MIMO)
- 在不同的时间段重复信息

ALOHA CSMA (Carrier Sense Multiple Access) 为解决同时发送多个传输时的冲突 (collision) 问题, 也叫 ALOHA. 等待一个随机的时间段后发送.

3. Network Protocols

设计目标:

- Reliability 可靠性
- Resource allocation 资源分配
- Evolvability 可进化
- Security 安全

重要思想: layering (分层)

Connection-oriented vs. connectionless service

Specific service primitives (特殊服务功能)

3.1 Design Goals

Reliability 可靠, 有 Error detection 与 Error correction 机制. 可在各中状况下自动寻路.

Resource allocation 统计多路复用 (Statistical multiplexing). 在网络中每一层出现, 使用 flow control. 使用 congestion control.

Evolvability 解耦, 函数端口一直, 具体实现不同. 使用地址或命名进行区别. 不同技术各有优缺. Called internetworking.

Security 安全问题与解决.

3.2 Protocol Layering

每一层仅为上一层服务. 在第 1 层之下是实际通信发生的物理介质. Interface 位于每对相邻层之间.

Protocol stack a list of the protocols used by a certain system, one protocol per layer. (某个系统使用的协议列表，每层一个协议.)

Protocol A protocol defines the format and the order of messages exchanged between two or more communication entities, as well as the actions taken on the transmission and/or receipt of a message or other event. (协议定义了两个或多个通信实体之间交换的消息的格式和顺序, 以及在发送和/或接收消息或其他事件时采取的行动.)

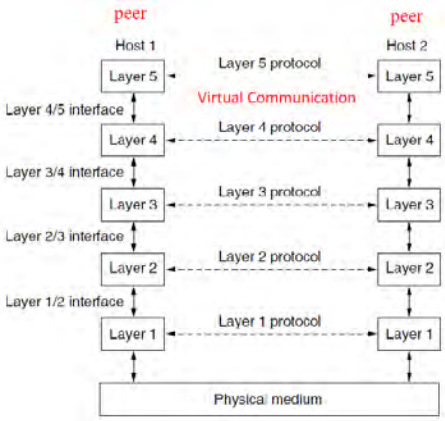

Figure I.6: Layers, protocols and interfaces

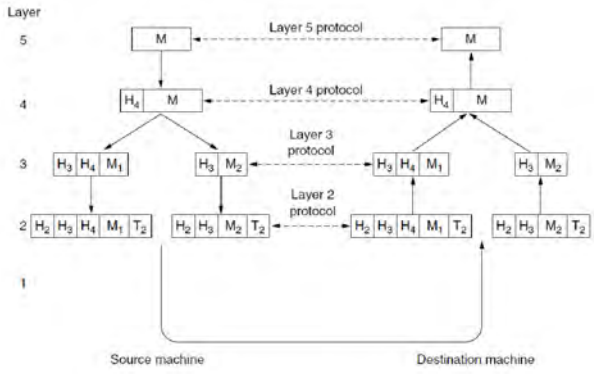

Figure I.7: Example information flow

Connection-Oriented vs. Connectionless Service

- Connection-oriented service: 建立链接, 使用链接, 释放链接.
- Connectionless service
  - 直发 Cut-through switching
  - 存储后转发 Store-and-forward switching

Service Primitives 等实验讲, 服务的函数.

Services vs. Protocols Services 是层与层, Protocols 是 host 与 host. 可以随意更改 Protocols, 但是不能更改用户可见的 Services.

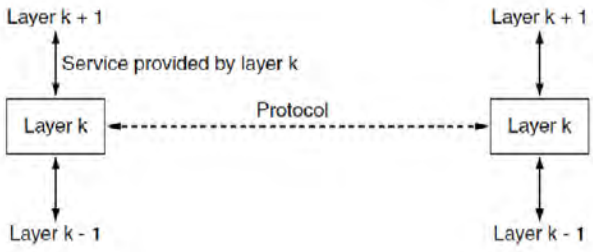

Figure I.8: Services vs. Protocols

4. Reference Models

4.1 The OSI reference model

OSI 是参考模型, 有七层:

1) The physical layer
2) The data link layer
3) The network layer
4) The transport layer
5) The session layer
6) The presentation layer
7) The application layer

OSI 并不是真正网络结构, 只是参考, 没有根基.

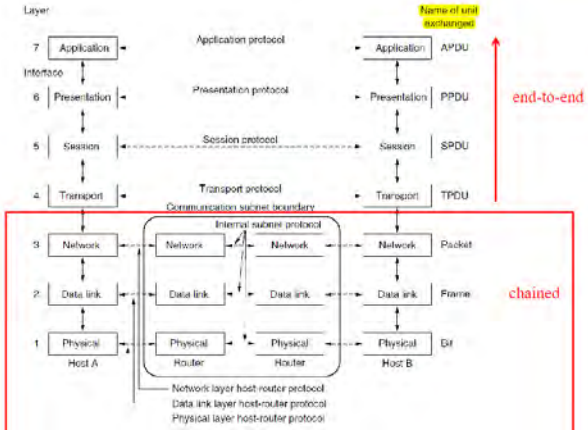

Figure I.9: The OSI reference model

最小传输单元:

- The physical layer 是 bit
- The data link layer 的是 frame
- The network layer 是 packet

1-3 层是链路 (chained), 可以被打开?

Physical Layer：最底层.

传输媒介: guided media or unguided media (有型与无型).

数字调制和多路复用: TDM, FDM, CDM

Data Link Layer 三个功能:

1) Framing 分帧
2) Flow Control
3) Error Control

涉及如何控制对共享频道的访问.

Network Layer：转发包. 涉及:

- 路由算法
- 网络协议: IPv4, IPv6
- Congestion control algorithms 与 Quality of services 自己看

Transportation Layer：端到端

- Connectionless Transport Protocol: UDP
- Connection-Oriented Transport Protocol: TCP

Session Layer and Presentation Layer 可放弃.jpg

Application Layer 协议:

- HTTP (Hyper Text Transfer Protocol) is basis of the World Wide Web.
- Email (SMTP)
- File transfer (FTP)

4.2 The TCP/IP Reference Model

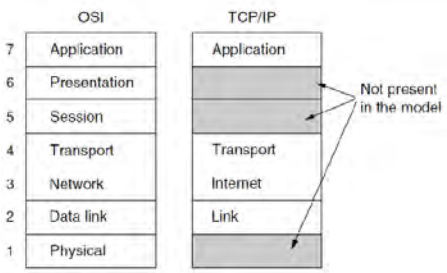

Figure I.10: The TCP/IP Reference Model

The link layer 包交换

The Internet layer IP, ICMP 协议

The transport layer TCP, UDP 协议

The application layer DNS (map host names onto their network address) 协议

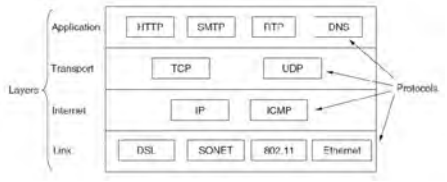

Figure I.11: The TCP/IP model with some protocols we will study

4.3 The OSI vs. TCP/IP

TCP/IP 模型在网络层仅支持一种模式 (无连接), 但在传输层同时支持这两种模式.

5. Standardization

Two categories: de facto and de jure. (先实践 or 先理论)

6. Metric Units

Exp.	Explicit	Prefix	Exp.	Explicit	Prefix
10^{-3}	0.001	milli	10^{3}	1,000	Kilo
10^{-6}	0.000001	micro	10^{6}	1,000,000	Mega
10^{-9}	0.000000001	nano	10^{9}	1,000,000,000	Giga
10^{-12}	0.000000000001	pico	10^{12}	1,000,000,000,000	Tera
10^{-15}	0.000000000000001	femto	10^{15}	1,000,000,000,000,000	Peta
10^{-18}	0.000000000000000001	atto	10^{18}	1,000,000,000,000,000,000	Exa
10^{-21}	0.000000000000000000001	zepto	10^{21}	1,000,000,000,000,000,000,000	Zetta
10^{-24}	0.000000000000000000000001	yocto	10^{24}	1,000,000,000,000,000,000,000,000	Yotta

Figure I.12: The principal metric prefixes

6.1 The Performance of Packet-Switching Networks

Quantitative Metrics

- Delay
  Processing delay: 检查数据包的标头并确定将数据包去向所需的时间
  Queuing delay: 计算排队时间
  Transmission delay: 包长度/速率
  Propagation delay: 路由间距/传输速度
- Loss: buffer 满了就丢包
- **Throughput**(吞吐量): 取决于瓶颈链路 (the bottleneck link) 的传输速率 (最慢的)

II The Physical Layer

最底层, network 基础. 最小单位: bit. 不同 physical channels 决定 throughput, latency (delay) and error rate.

1. The theoretical basis for data communication

1.1 Fourier Series

信息通过改变物理量来传输.

Any reasonably behaved **periodic** function $g(t) = g(t + nT_0)$ with period T can be represented as

$$\begin{aligned} g(t) &= \sum_{n=-\infty}^{+\infty} a_n e^{j2\pi f n t} \\ &= \sum_{n=-\infty}^{+\infty} a_n(\cos(2\pi f n t) + j\sin(2\pi f n t)) \end{aligned}$$

where $f_0 = \frac{1}{T_0}$ is **the fundamental frequency** of the periodic signal $g(t)$.

If the period T_0 is known and the amplitudes an are given, the original periodic signal $g(t)$ can be reconstructed. 也可反求 a_n.

$$a_n = \frac{1}{T}\int_T g(t)e^{-j2\pi f n t}dt$$

If $g(t)$ is a **real** signal, then the coefficient a_{-n} is **the conjugate** of a_n.

$$g(t) = C + \sum_{n=1}^{\infty} 2A_n \cos(2\pi n f t) + \sum_{n=1}^{\infty} 2B_n \sin(2\pi n f t) \quad \text{(II.1)}$$

$$A_n = \frac{1}{T}\int_0^T g(t)\cos(2\pi n f t)dt$$

$$B_n = \frac{1}{T}\int_0^T g(t)\sin(2\pi n f t)dt$$

$$C = \frac{1}{T}\int_0^T g(t)dt$$

但现实世界信号有限, 但可以假想为重复的无限信号, 如此可用 Fourier.

1.2 Bandwidth-limited Signals

信号在信道传输会衰减. 且信道对不同频率的影响不同. 但对某一根线, 振幅从 0 到 f_c(the cutoff frequency, 截止频率) 之前信号衰减很小.

The width of the frequency range transmitted without being strongly attenuated is called the **bandwidth**(带宽). (带宽是没有大量衰减的范围) 带宽是物理属性, 不依赖于环境.

Baseband Signals vs. Passband Signals

- Signals that run from 0 up to a maximum frequency are called basedband signals.
- Signals are shifted to occupy a higher range of frequencies, such as in the case of all wireless transmission, are called passband signals.

f_c carrier frequency

Bandwidth vs. Maximum Data Rate

- analogue bandwidth is a quantity measured in Hz.
- digital bandwidth is the maximum data rate of a channel, a quantity measured in bits/sec.

data rate 是使用 analogue bandwidth 最终传输结果.

1.3 The Maximum Data Rate of a Channel

Channel Capacity — the maximum rate at which data can be transmitted over a given communication path or channel under given conditions.

Four related concepts:

1) Data rate (bps)
2) Bandwidth (Hz)
3) Noise
4) Error rate

1.4 Nyquist Bandwidth

在理想情况下的 the maximum signaling rate

$$s(t) = \sum_n a_n g(t - nT)$$

where $g(t)$ represents a basic pulse shape and $\{a_n\}$ is the binary sequence of $\{\pm 1\}$ transmitted at a rate of $\frac{1}{T}$ bits/s.

The optimal pulse shape:

$$g(t) = \frac{\sin 2\pi B t}{2\pi B t}$$

- Binary: $C = 2B$ (V=2)
- Multilevel Signaling: $C = 2B\log_2 V$ (V is the number of discrete signal or voltage levels)

where C is maximum signaling rate, B is bandwidth.

提升信号种类的数量可以提升 data rate. 但提升有限度, 且会被在传输中的 noise and other impairment 限制.

1.5 Shannon Capacity

仅假设白噪声, 不考虑其他影响, 仅理论最大值. SNR (信噪比)

$$C = B\log_2(1 + SNR)$$

$$SNR_{dB} = 10\lg\frac{\text{signal power}}{\text{noise power}} = 10\lg(SNR)$$

2. Three kinds of transmission media

2.1 Guided Transmission Media

- Persistent storage (固态存储)
- Twisted pairs (双绞线)
- Coaxial cable (同轴线缆)
- Power lines (交流电的线)
- Fiber optics (光纤)

Persistent storage 卡车运硬盘仍是最高的 bandwidth. 但 delay 很大.

Twisted pairs 最古老但仍最常用的传输介质. 为什么要绞: 因为两根平行线会形成一个小天线, 纠缠会抵消其干扰. 一个信号是这两根线的电压差, 提升鲁棒性.

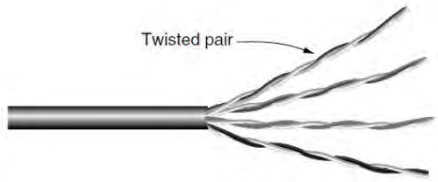

Figure II.1: Twisted pairs

Coaxial Cable 黄铜轴 x

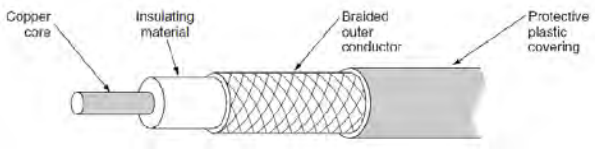

Figure II.2: Coaxial Cable

Power Lines 交流电的线. 中国民用交流电: 50Hz

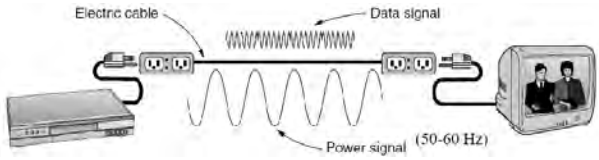

Figure II.3: Power Lines

Fiber Optics 理论上 bandwidth 可以无限大.

重要组成部分: Light source, transmission medium, and detector.

光源: LEDs, Semiconductor lasers.

2.2 Wireless

three variations:

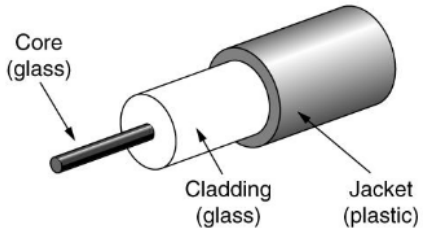

Figure II.4: Fiber Optics

- Frequency hopping spread spectrum
- Direct sequence spread spectrum e.g. CDMA
- UWB (Ultra WideBand)

Radio Transmission Radio frequency (RF) waves

Microwave Transmission 传输是直线. 有 multipath fading 现象. 需要批准, 但有 unlicensed 波段 e.g. ISM and U-NII.

Infrared/Light Transmission 不可穿过固体, 但不需要批准, 可随意使用. 但易受对流气流干扰.

2.3 Satellites

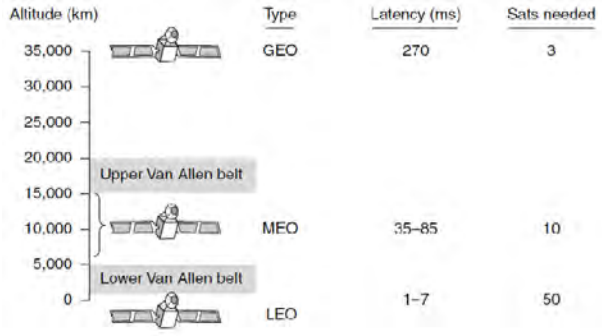

Figure II.5: communication satellites and some of their properties

e.g. 星链

Satellites vs. Fiber For Satellites

- Rapid deployment
- For communication in places where the terrestrial infrastructure is poorly developed.
- When broadcasting is essential

3. Digital modulation and multiplexing

Channels 所携带的是 analog signals (模拟信号), 人们使用 analog signals 表示 bits, 称为 digital modulation (数字调制).

有两种策略来将 bits 转换为 signal:

1) Baseband transmission (for wired channels): 依据信令速率 (the signaling rate) 占用频率.

2) Passband transmission (for wireless and optical channels): Amplitude, phase and frequency modulation (调幅, 调相, 调频), 使用频率携带信号. 即调整 $A\cos(2\pi f+\theta)$ 中的 A, f, θ 来携带信息.

3.1 Baseband Transmission

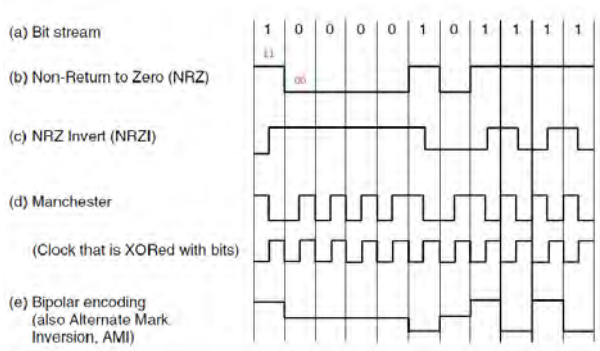

Figure II.6: Line codes (重点)

Non-Return to Zero (NRZ) positive voltage 表示 1, negative voltage 表示 0. 对于光纤, 有光为 1, 无光为 0.

但会有衰减和失真 (attenuated and distorted), receiver 会将信号采样映射到最近的表示. 传输速率为 B bits/sec 时, 至少需要 B/2 Hz 带宽.

可以提升电压种类来提升一次传输的 bit.

The bit rate = the symbol rate × the number of bits per symbol

Manchester 但 NRZ 定位错误信号解析也会错误, 所以衍生策略:

1) 额外传输时钟
2) 仅传输信号与时钟的异或 (时钟的频率一致).

Many Ethernet technologies use Manchester encoding. 但其带宽高.

NRZ Invert (NRZI) 1 时电平翻转, 0 时保持不变.

但易受攻击, 所以使用加密. 异或一个双方已知的 pseudorandom, 以保证 01 数量相近.

Bipolar encoding Balanced signals 无直流分量, 因为直流分量会衰减, 所以这是个优点. 用两个 voltage 表示 1, 0 voltage 表示 0.

3.2 Passband Transmission

传输 $x(t)$, 发出的信号是 $x(t)\cos(2\pi f_c t)$, 接受时调到 f_c 将信号处理成 $x(t)\cos^2(2\pi f_c t)=x(t)\frac{1+\cos(4\pi f_c t)}{2}$, 然后过滤 $4\pi f_c t$ 高频的信号, 就会留下 $\frac{x(t)}{2}$.

$$\lambda f = c = 3\times 10^8 m/s$$

将 f 调大, λ 就会较小, 便于发送.

Can modulate the amplitude, frequency or phase of the carrier signal:

1) ASK (Amplitude Shift Keying):
2) FSK (Frequency Shift Keying):
3) PSK (Phase Shift Keying):

结合上者来传输更多信号.

编码规则使用 Gray Code, 如此一位出错的影响不大.

FDM (Frequency Division Multiplexing) FDMA: Even for wires, placing a signal in a given frequency band is useful to let different kinds of signals coexist on the channel. 不同信号用不同频率.

OFDM (Orthogonal FDM) 信号可重叠.

TDM (Time Division Multiplexing) 一个时间一个信号. 但同传输的信号需要同步.

CDM (Code Division Multiplexing) 需要两路时钟. 将每个 bit time 分为多个间隔称 chips, 一般 64/128 chips per bit, 然后每个 station 使用唯一的 chip sequence (Walsh codes), sequence 要求正交.

4. Three examples of communication examples

4.1 Public Switch Telephone Network (PSTN)

The telephone system consists of three major components:

1) Local loops: telephone modem, ADSL, fiber
2) Trunks (digital fiber optic links connecting switching offices) — main consideration problem is multiplexing (FDM and TDM)
3) Switching offices (where calls are moved from one trunk to another) — two switching ways

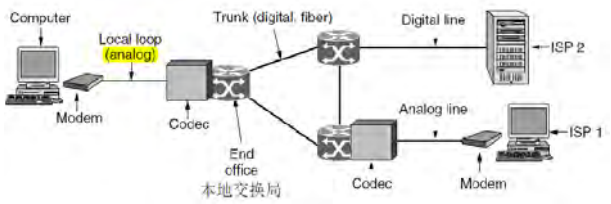

Figure II.7: PSTN

Local Loop Telephone modems: 转换二进制流与模拟信号 (modulator demodulator). 对于 4000Hz 需要每秒采样 8000 次 (Nyquist sampling). 下载一般比上传高.

Trunks 传输的是 digital information(e.g. bits). Accomplished with :

- SONET: the TDM system used for fiber optics
- FDM is applied to fiber optics (wavelength division multiplexing)

Switching Circuit Switching: a dedicated physical path. Packet Switching: store-and-forward.

Circuit Switching vs. Packet Switching

4.2 Cellular Networks

4.3 Cable Networks

III Data Link Layer

1. Overview of Data link layer

1.1 The Design Principles of the Data Link Layer

To achieve reliable and efficient communication between two adjacent machines

Three main functions of the data link layer:

- Framing: Break stream of bits up into discrete frames
- Error control: Error Detection and Error Correction
- Flow control: to prevent a sender to overload the receiver with packets

1.2 Position of the Data Link Layer

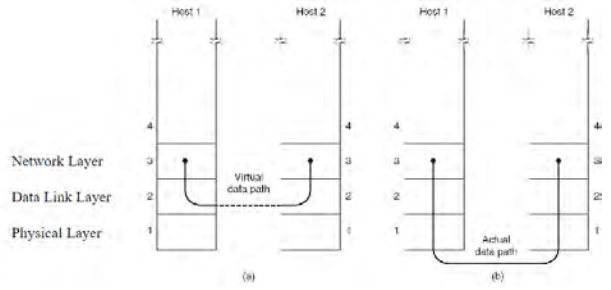

Figure III.1: The function of the data link layer is to provide services to the network layer.

1.3 The Services of Data Link Layer

1) Unacknowledged connectionless service
   for very low error-rate and real-time applications
   e.g. Ethernet
2) Acknowledged connectionless service
   for unreliable channels
   e.g. 802.11 WiFi
3) Acknowledged connection-oriented service
   for over long, unreliable links
   e.g. a satellite channel or a long-distance telephone circuit

2. Data link layer: Framing

General Frame Format: a header, a payload field, and a trailer

会加一些冗余的 bits 来降低 bit error rate. There are about four methods:

1) Byte count: 若刚好 count 出错了就会很寄.

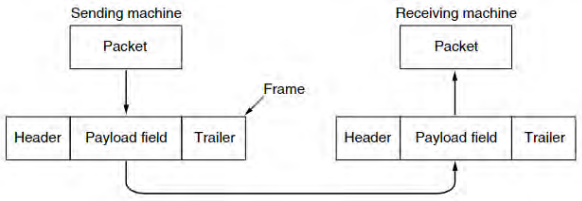

Figure III.2: Relationship between packets and frames.

2) Flag bytes with byte stuffing: 因为 flag 可能会在传输数据中出现, 所以在其前加上 a special escape byte (ESC) 以区别, 若 ESC 也在数据中, 则同样其前加 ESC.

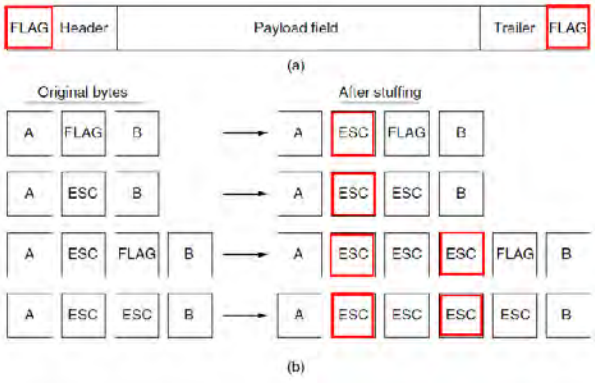

Figure III.3: (a) A frame delimited by flag bytes. (b) Four examples of byte sequences before and after byte stuffing.

3) Flag bits with bit stuffing: Whenever the sender's data link layer encounters five consecutive 1s in the data, it automatically stuffs a 0 bit into the outgoing bit stream.

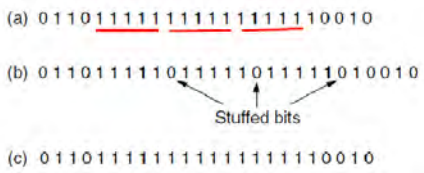

Figure III.4: Bit stuffing. (a) The original data. (b) The data as they appear on the line. (c) The data as they are stored in the receiver's memory after destuffing.

4) Physical layer coding violations: We can use some reserved signals to indicate the start and the end frames.

3. Data link layer: Error Control

- Error-Correcting Codes: For unreliable channels
- Error-Detecting Codes: For highly reliable channels

the redundant bits that offer protection are as likely to be received in error as the data bits.

3.1 Error Models of Communication Channels

3.2 Error Correction

two different error-correcting codes:

- Hamming codes
- Binary convolutional codes

对码本 (一对码字), from this list to find the two codewords with the smallest Hamming distance. This distance is the Hamming distance of the complete code. (码本中码字间两两最小的 Hamming distance 是码本的 Hamming distance)

The code rate is m/n. (n=m+r, nis the total length of a block, r is redundance)

- To reliably detect d errors, you need a distance d + 1 code
- To correct d errors, you need a distance 2d + 1 code.

we expect only single- or double-bit errors

For all single errors to be corrected, this requirement becomes

$$(m + r + 1) \leq 2^r$$

Given m, this puts a lower limit on the number of check bits needed to correct single errors.

Hamming Code k data bits $+(n-k)$ check bits. value 1 are XORed together to calculate the check bits.

Convolutional Code: the original m bits do not appear

3.3 Error Detection

three different error-detecting codes:

- Parity: the number of 1 bits in the codeword is even or odd.

  Advantage: bit error rates 非常低时, 花费比 error correction 少

  Difficulty: for a long burst error (一大段错误), 仅能检测一般的错误. To against burst errors:

  - Interleaving
  - Two-dimensional parity

- Checksums:

  - One's complement has two representations of zero, all 0s and all 1s
  - two's complement arithmetic

  Sending:

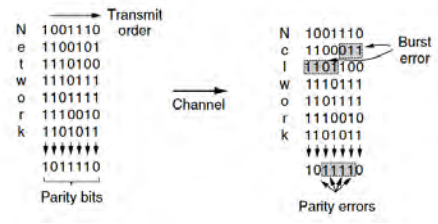

Figure III.5: Interleaving of parity bits to detect a burst error

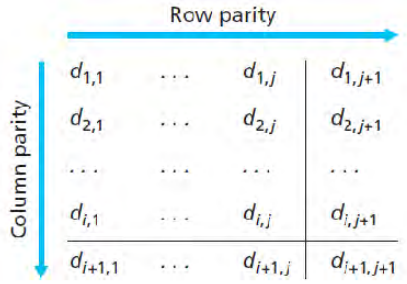

Figure III.6: Two-dimensional parity

1) Arrange data in 16-bit words
2) Put zero in checksum position, add
3) Add any carryover back to get 16 bits
4) Negative (complement) to get sum

Receiving

1) Arrange data in 16-bit words
2) Checksum will be non-zero, add
3) Add any carryover back to get 16 bits
4) Negate the result and check it is 0.

- Cyclic Redundancy Checks (CRCs): predetermined bit pattern - key (or the generator polynomial) (Both the high- and low-order bits of the generator must be 1)

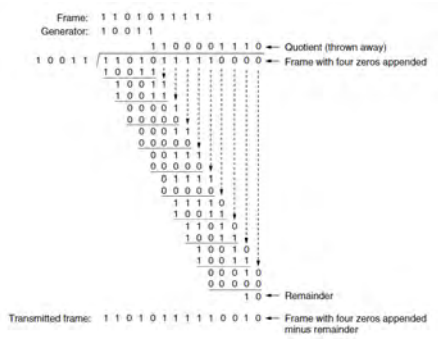

Figure III.7: Example calculation of the CRC

4. Elementary data link protocols

4.1 A Utopian Simplex Protocol

Unrealistic assumptions:

1) Data are transmitted in one direction only.
2) Both the transmitting and receiving network layers are always ready.
3) Processing time can be ignored.
4) Infinite buffer space is available.
5) The communication channel between the data link layer never damages or loses frames.

4.2 A Simplex Stop-and-Wait Protocol for an Error-Free Channel

This protocol is to tackle the problem of preventing the sender from flooding the receiver with frames faster than the latter is able to process them.

receiver provide a little dummy frame (an acknowledge frame) as feedback to sender (目的是告诉 sender, receiver 已完成接受).

4.3 A Simplex Stop-and-Wait Protocol for a Noisy Channel

To add a timer, the receiver would only send an acknowledgement frame if the data were correctly received. timer 触发但没收到 ack, sender 会重传.

A fatal flaw: duplicate. 若 ack 完全丢失, 会接受一个 frame 两次, 不可接受.

put a sequence number to avoid duplicate. the only ambiguity is between a frame and its immediate predecessor or successor

- After transmitting a frame and starting the timer, the sender waits for something exciting to happen.
- Only three possibilities exist: an acknowledgement frame arrives undamaged, a damaged acknowledgement frame staggers in, or the timer expires.

5. Sliding window protocols

5.1 Piggybacking(驮运)

ack 可以搭顺风车, 与下一个数据包一起发送. The technique of temporarily delaying outgoing acknowledgements so that they can be hooked onto the next outgoing data frame. In effect, the acknowledgement gets a free ride on the next outgoing data frame.

使用一个计时器决定 ack 是否搭顺风车, 在时间内有其他包就顺风车, 没有就单独发.

Sliding window bidirectional protocol(注意其不同点):

- A One-Bit Sliding Window Protocol (stop-and-wait)
- A Protocol Using Go Back N
- A Protocol Using Selective Repeat

5.2 The Essence of All Sliding Window Protocols

the sender : maintains **a set of sequence numbers** corresponding to frames it is permitted to send. These frames are said to fall within the **sending window**.

The sequence numbers within the sender's window represents frames that

1) have been sent but are yet not acknowledged
2) can be sent

(If some frames are lost or damaged in transit, the sender should prepare for possible retransmission)

If the window ever grows to its maximum size, the sending data link layer must forcibly shut off the network layer until another buffer becomes free. — **Flow Control**

the receiver : also maintains **a receiving window** corresponding to the set of frames it is permitted to accept.

Any frame falling within the window is put in the receiver's buffer. Any frame falling outside the window is discarded.

When a frame whose sequence number is equal to the lower edge of the window is received, it is passed to the network layer and the window is rotated by one.

Note that a window size of 1 means that the data link layer only accepts frames in order. (For large window, it should answer how to keep the correct order)

The sender's window and the receiver's window need not have the same size.

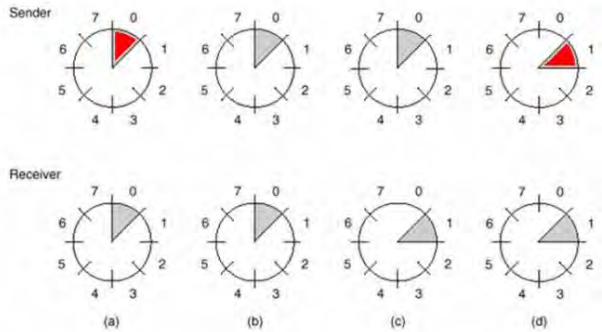

Figure III.8: A sliding window of size 1, with a 3-bit sequence number: (a) Initially. (b) After the first frame has been sent. (c) After the first frame has been received. (d) After the first acknowledgement has been received.

5.3 One-Bit Sliding Window Protocol

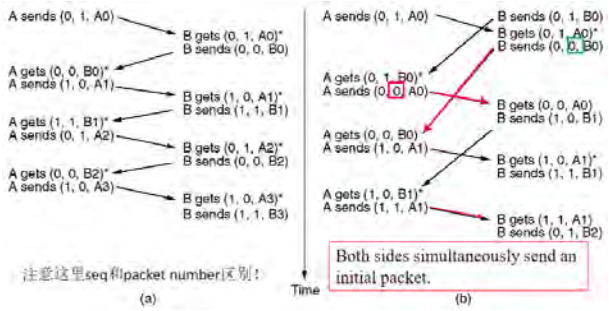

Figure III.9: Two scenarios for protocol 4 (One-bit Sliding Window). (a) Normal case. (b) Abnormal case. The notation is (seq, ack, packet number). An asterisk indicates where a network layer accepts a packet. (Here the red arrows denotes the retransmissions because of timeout)

e.g. The rule requiring a sender to wait for an acknowledgement before sending another frame, 导致 bandwidth 利用率极低. The protocol 4 is disastrous in terms of efficiency under the situation of a long transit time, high bandwidth, and short frame length.

To find an appropriate value for w

1) the bandwidth-delay product
2) We can divide this quantity by the number of bits in a frame to express it as a number of frames. — BD
3) w should be set to $2BD + 1$

5.4 A Protocol Using Go Back N

Pipelining frames over an unreliable channel raises some serious problem.

- What happens if a frame in the middle of a long stream is damaged or lost?
- What should the receiver do with all the correct frames following a damaged frame?

Remember that the receiving data link layer is obligated to hand packets to the network layer in sequence.

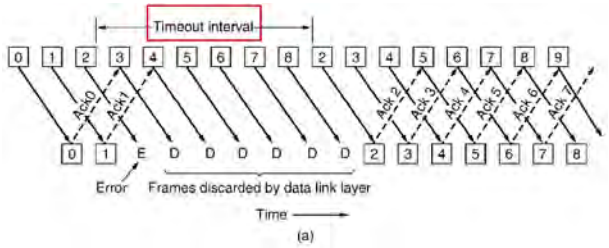

Figure III.10: (a) Receiver's window size is 1. — Go Back N

In **Figure** III.10, the receiver simply discard all subsequent frames, sending no acknowledgements for the discarded frames.

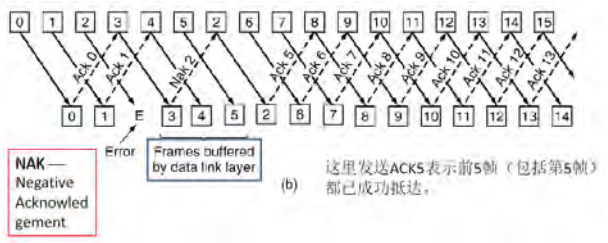

Figure III.11: Receiver's window size is large. — Selective Repeat

In **Figure** III.11, a bad frame is received and discarded, but any good frames received after it are accepted and buffered. When the sender times out, only the oldest unacknowledged frame is retransmitted.

Assumptions :

- Data in both directions: piggybacking of ACK. (No separate ACK packets)
- Noisy channel
- Limited stream of data from network layer. (a "network_layer_ready" event)

Protocol : Tanenbaum 5th edition: Fig. 3.19

The sender window of size n. The receive window of size 1.

MAX_SEQ+1 distinct sequence numbers $(0, 1, 2, \ldots, MAX\_SEQ)$, no more than MAX_SEQ unacknowledged frames, the sender $window \leq MAX\_SEQ$.

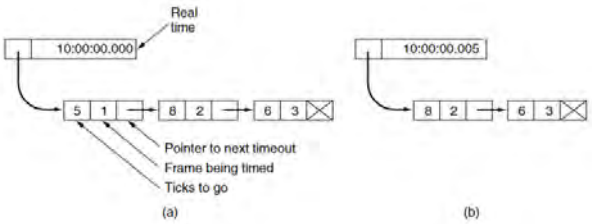

Figure III.12: Simulation of multiple timers in software. (a) The queued timeouts. (b) The situation after the first timeout has expired.

5.5 A Sliding Window Protocol Using Selective Repeat

The selective repeat protocol is to allow the receiver to accept and buffer the frames following a damaged or lost one.

Assumptions :

- Data in both directions: piggybacking of ACK. (Separate ACK packets (requires ACK timer), accelerate ack)
- Noisy channel
- Limited stream of data from network layer
- NACK packets: receiver detected problem

Protocol : Tanenbaum 5th Edition Fig. 3.21

Nonsequential receive introduces further constraints on frame sequence number compared to protocols in which frames are only accepted in order!

6. Examples of data link protocols

point-to-point lines in the internet in two common situations:

- SONET
- ADSL (Asymmetric Digital Subscriber Loop)

6.1 Point-to-Point Protocol

A standard protocol called PPP (Point-to-Point Protocol) is used to send packets over these links.

PPP provides three main features:

1) A framing method that unambiguously delineates the end of one frame and the start of the next one. The frame format also handles error detection.

2) A link control protocol for bringing lines up, testing them, negotiating options, and bringing them down again gracefully when they are no longer needed. This protocol is called LCP (Link Control Protocol).

3) A way to negotiate network-layer options in a way that is independent of the network layer protocol to be used. The method chosen is to have a different NCP (Network Control Protocol) for each network layer supported.

6.2 Packet over SONET

SONET is the physical layer protocol. It provides a bitstream that runs at a well-defined rate. This bitstream is organized as fixed-size byte payloads that recur every 125 μsec, whether or not there is user data to send.

PPP runs on IP router to provide the transmission mechanism

The PPP frame format

6.3 The PPP Frame Format

All PPP frames begin with the standard HDLC flag byte of 0x7E (01111110).

- The Address field
- The Control field
- The Protocol field
- The Payload field (1500 bytes)
- The Checksum field

6.4 PPP Link Up & Down

6.5 ADSL (Asymmetric Digital Subscriber Loop)

An AAL5 frame

IV The Medium Access Control(MAC) Sublayer

网络连接分两种: P2P or Broadcasting
MAC 是 Broadcasting.

1. The Channel Allocation Problem

Static channel allocation: FDMA (bandwidth 分 N 等大的块, 每个人一块)

但网络有 bursty (突发性), 所以使用 Dynamic channel allocation, also called Statistical Multiplexing.

1.1 Preliminary Queueing Theory

$$T_N = \frac{1}{\mu(C/N) - (\lambda/N)} = \frac{N}{\mu C - \lambda} = NT$$

- Consider any system that has a capacity C, the maximum rate at which it can perform work
- Assume that R represents the average rate at which work is demanded from this system.

If $R < C$, 系统可以处理工作, 但因为有 the irregularity of the demands, 所以还是不行. If $R > C$, 系统寄.

Two unpredictable sources lead to the irregularity of the demands:

1) the arrival times
2) the service times

Queueing Theory: the characterization of the arrival times and the service times and the evaluation of their effect on queueing phenomena form the essence of queueing theory.

Let C_n denote the n-th customer to arrive at a queueing facility, and

- τ_n = arrival time for C_n
- $t_n = \tau_n - \tau_{n-1}$ = interarrival time between C_n and C_{n-1}
- x_n = service time for C_n.

$\{t_n\}$, $\{\tau_n\}$ 是随机变量且独立分布. 定义两个一般随机变量:

- $\tilde{t}$ = interarrival time
- $\tilde{x}$ = service time

probability distribution function (PDF)

- $A(t) = P[\tilde{t} \le t]$
- $B(x) = P[\tilde{x} \le x]$

probability density function (pdf)

- $a(t) = \frac{dA(t)}{dt}$

- $b(x) = \frac{dB(x)}{dx}$

important moments 联系这些随机变量

- $E[\tilde{t}]$ = mean interarrival time = $\frac{1}{\lambda}$, where λ is the average arrival rate.
- $E[\tilde{x}]$ = mean service time = $\frac{1}{\mu}$.

The study of queues naturally breaks into three cases:

1) elementary queueing theory
2) intermediate queueing theory
3) advanced queueing theory

根据 $a(t)$ 与 $b(x)$ 区分.

A three-component description $A/B/m$, which denotes

- m-server queueing system
- A the interarrival time distribution
- B the service time distribution

And distributions symbols:

- M = exponential (i.e. Markovian)
- Er = r-stage Erlangian
- Hr = r-stage Hyperexponential
- D = Deterministic
- G = General

The simplest interesting system is the $M/M/1$ queue.

General Results The most important system parameter for $G/G/1$ is the utilization factor ρ,

$$\rho = \lambda\tilde{x}$$

单个 server 单位时间处于忙碌状态的时间. ρ is the expected fraction of the system's capacity that is in use.

For multiple-server system $G/G/m$

$$\rho = \frac{\lambda\tilde{x}}{m}$$

In all cases a stable system is one for which $0 \geq \rho < 1$.

The closer ρ approaches unity, the larger are the queues and the waiting time.

The average time in system

$$T = \tilde{x} + W$$

where W is waiting time in queue.

Little's Result 系统中平均客户数

$$\bar{N} = \lambda T$$

The average queue size is

$$\bar{N}_q = \lambda W$$

In $G/G/m$, these quantities are related by

$$\bar{N}_q = \bar{N} - m\rho$$

可以理解为系统中平均客户数减去正在被服务的客户数就是正在等待服务的客户数了

动态客流:

- E_k k customers 在系统中的状态
- $P_k(t) = P[N(t) = k]$ 在 t 时, 系统处于 E_k 的概率

$$\frac{dP_k(t)}{dt} =$$

Consider a stable system, $p_k = \lim_{t\to 0} P_k(t)$.

The $M/M/1$ queue:

- This system has a Poisson input.
- The probability $P_k(t)$ of k arrivals in an interval whose duration is t sec is given by

$$P_k(t) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}$$

- The average number of arrivals during this interval is

$$\bar{N}(t) = \sum_{k=0}^{\infty} kP_k(t) = \lambda t$$

So the average arrival rate in a unit time is λ.

- The average service time is $\tilde{x} = \frac{1}{\mu}$
- The probability of having k customers in the system is given by

$$p_k = (1-\rho)\rho^k$$

- The average number in the system is

$$\bar{N} = \sum_{k=0}^{\infty} kp_k = \frac{\rho}{1-\rho}$$

- Using Little's result and the average time in the system T is

$$T = \frac{\bar{N}}{\lambda} = \frac{\frac{1}{\mu}}{1-\rho}$$

- W is

$$W = \frac{\bar{N}_q}{\lambda} = \frac{\frac{\rho}{\mu}}{1-\rho}$$

$$\bar{N}_q = \bar{N} - m\rho = \frac{\rho^2}{1-\rho}$$

1.2 The Performance of Static FDM

The mean time delay T to send a frame onto a channel of capacity C bps (Note only one channel here). Assume

- the average arrival rate λ
- the average length $\frac{1}{\mu}$
- the service rate $\tilde{x} = \frac{1}{\mu C}$
- The utilization factor $\rho = \lambda \tilde{x}$
- According to the results of the $M/M/1$ queue

$$T = \frac{\frac{1}{\mu C}}{1-\rho} = \frac{1}{\mu C - \lambda}$$

Let us divide the single channel into N independent subchannels, each with capacity C/N bps. For subchannels:

- mean input rate $\frac{\lambda}{N}$
- According to the results of an $M/M/m$ queue, the mean time delay

$$T_N = \frac{1}{\mu(C/N) - (\lambda/N)} = \frac{N}{\mu C - \lambda} = NT$$

分别排队等待时间更长.

1.3 Key Assumptions of Dynamic Channel Allocation

Five key assumptions:

1) Independent Traffic: The model consists of N independent stations. 每个人收发信号随机且独立.

2) Single Channel: A single channel is available. 所以可能有冲突.

3) Observable Collisions: All stations can detect the collision when it occurred.

4) Continuous or Slotted Time: Frame transmission can begin at any instant, or must begin at the start of a slot. (连续 or 有时钟)

5) Carrier Sense or No Carrier Sense: (发送前是否检测信道状态)

With the carrier sense assumption, stations can tell if the channel is in use before trying to use it.

If there is no carrier sense, stations cannot sense the channel before trying to use it.

2. Multiple Access Protocols

2.1 ALOHA

Each user terminal sharing the same upstream frequency to send frames to the central computer.

Two versions of ALOHA: pure and slotted.

2.2 Pure ALOHA

sender 向 central computer 发 frame, 然后 central computer rebroadcasts the frame to all of the stations, sender 监听这个转播来确定是否发送成功. (收到的 frame 需要和发送的一致) 如果发送失败, sender 等待随机时间后再次发送.

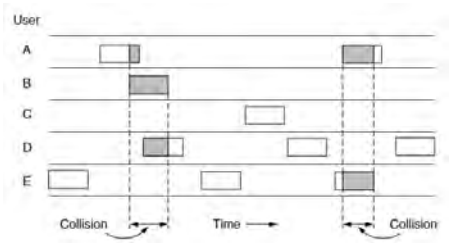

Figure IV.1: In pure ALOHA, frames are transmitted at completely arbitrary times.

If the first bit of a new frame overlaps with just the last bit of a frame that has almost finished, both frames will be totally destroyed.

An interesting question is: what is the efficiency of an ALOHA channel?

- a Poisson distribution
- with mean of N new frames per frame time.
  If $N > 1$, 超过信道容量, 寄.
  expect $0 < N < 1$
- with mean of G old frames per frame time.
  $G \geq N$
  At load low ($N \approx 0$), $G \approx N$
  At high low, $G > N$
  the throughput $S = GP_0$, P_0 is the probabilityof a transmission succeeding

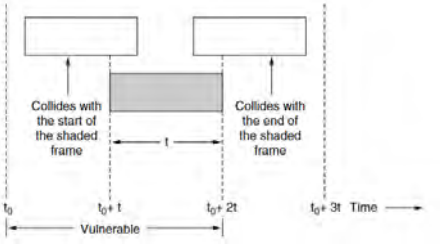

Figure IV.2: Vulnerable period for the shaded frame

对一帧要保证 $2t$ 时间内没有其他帧发送.

- The probability that k frames are generated during a given frame time

$$P[k] = \frac{G^k e^{-G}}{k!},\ G \sim \lambda t$$

The probability of zero frames is e^{-G}
the entire vulnerable period is $P_0 = e^{-2G}$

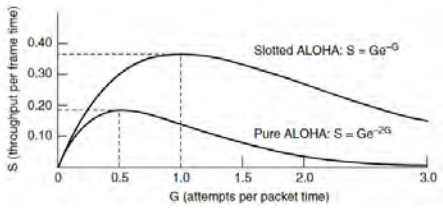

Figure IV.3: Throughput versus offered traffic for ALOHA systems.

pure ALOHA 的最值在 $G = 0.5$ 处. Slotted ALOHA 的在 $G = 1$ 处.

2.3 Slotted ALOHA

divide time into discrete intervals called slots, agree on slot boundaries: 帧只能在 slot 的起点发.

The probability of no other traffic is e^{-G}, so $S = Ge^{-G}$. The probability of a collision is $1 - e^{-G}$.

推导:

- The probability of a transmission requiring exactly k attempts is

$$P_k = e^{-G}(1 - e^{-G})^{k-1}$$

- The expected number of transmissions E is

$$E = \sum_{k=1}^{\infty} kP_k = e^G$$

As a result of the exponential dependence of E upon G, small increases in the channel load can drastically reduce its performance.

2.4 Carrier Sense Multiple Access(CSMA) Protocols

With slotted ALOHA, the best channel utilization that can be achieved is $1/e$. 因为任意发送太容易导致冲突了.

1-persistent CSMA 先监听信道看是否空闲, 空闲就发送, 不空闲一直侦听直到信道空闲然后发送. 遇到冲突仍然是等待随机时间再执行. The protocol is called 1-persistent because the station transmits with a probability of 1 when it finds the channel idle.

Nonpersistent CSMA 先监听信道看是否空闲, 空闲就发送, 不空闲等待随机时间然后再重复算法.

p-persistent CSMA applies to slotted channels. 先监听信道看是否空闲, 空闲有 p 概率发送, $1 - p$ 概率不发送, 等待下一个 slot. 不空闲就等待下一个 slot

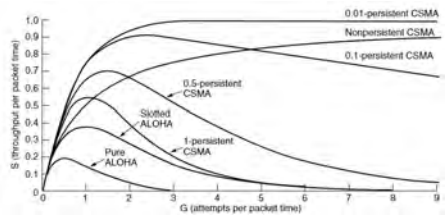

Figure IV.4: Comparison of the channel utilization versus load for various random access protocols.

2.5 CSMA with Collision Detection

Collision detection 边发送边读取. 若读的数据与发的一样就继续, 否则就是产生冲突, 中止发送.

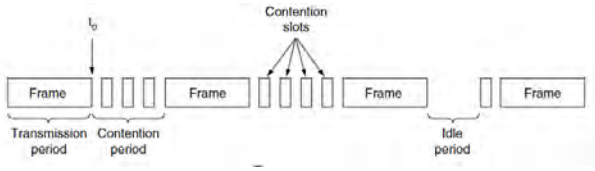

Figure IV.5: CSMA/CD can be in contention, transmission, or idle state

Suppose that two stations both begin transmitting at exactly time t_0. How long will it take them to realize that they have collided?

- Consider the worst-case scenario. propagate between the two farthest stations is τ
- 在 t_0 A 发送, 在 $t_0 + \tau - \epsilon$ 时 B 发送, 直到 $2\tau - \epsilon$ 冲突才被 A 探测.

We can think of CSMA/CD contention as a slotted ALOHA system with a slot width of 2τ.

2.6 Collision Free Protocols

若 the bandwidth-delay product 大, 冲突影响大. 所以来些没有冲突的算法.

A Bit-Map Protocol — Reservation Protocol N slots, 当作 bit map, bit 1 是发, 仅一个 1 时才发. 但不支持过大的网络.

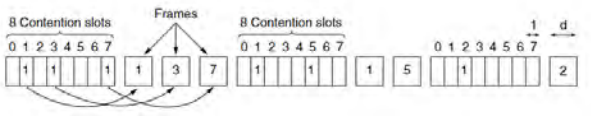

Figure IV.6: The basic bit-map protocol

Token Passing 令牌轮询, 只有拿到令牌才能发送数据. 但令牌产生与销毁会导致不公平.

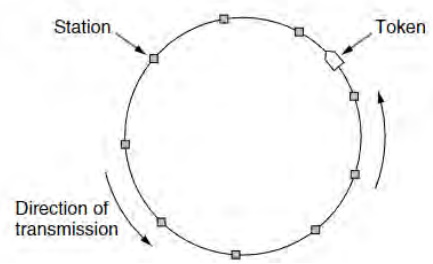

Figure IV.7: Token ring

Binary Countdown 结合前两方法的优点. 每个站点给个编号, 更大的站点发.

2.7 *Limited-Contention Protocols*

Two important performance measures broadcast network:

- delay at low load
- channel efficiency at high load.

Limited-contention Protocols combine the best properties of the contention and collision-free protocols.

- 每个站点有 p 概率获得信道
- 但一个站点获得信道, 其他所有站点有 $1-p$ 不获得信道. The value is $kp(1-p)^{k-1}$

when $p = \frac{1}{k}$, the value is max.

$$Pr[\text{success with optimal}p] = \left(\frac{k-1}{k}\right)^{k-1}$$

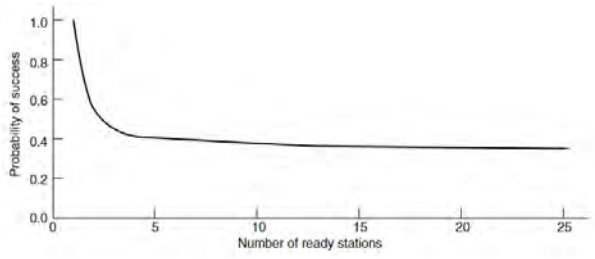

Figure IV.8: Acquisition probability for a symmetric contention channel

As soon as the number of stations reaches 5, the probability has dropped close to its asymptotic value of $1/e$.

所以将用户分组, group x 竞争 slot x, 成功发送, 失败不发. 合适分组可以降低竞争程度. low load with big group, high load with small group.

The Adaptive Tree Walk Protocol 结点 1 是所有用户, 冲突后会分结点, 进行 DFS, 成功会进行 BFS.

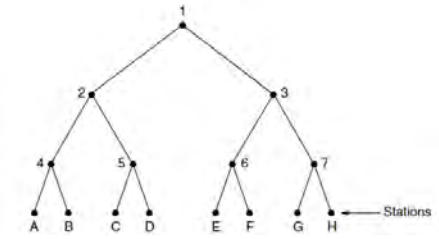

Figure IV.9: The tree for eight stations

2.8 *Wireless LAN Protocols*

Wireless is more complicated than the wired case

- Nodes may have different coverage areas.
  不同结点覆盖区域不同, 要避免之间的冲突.

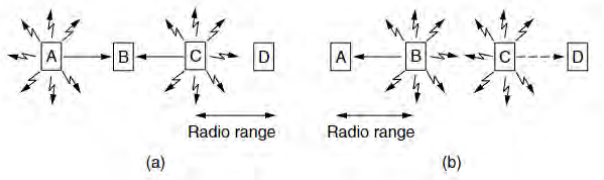

Figure IV.10: A wireless LAN. (a) A and C are **hidden terminals** when transmitting to B. (b) B and C are **exposed terminals** when transmitting to A and D.

- Nodes cannot hear while sending
  A node can hardly hear anything while sending messages.

Possible Solution: MACA (Multiple Access with Collision Avoidance) A short handshake before sending a frame

Protocol rules:

1) A sender node transmits a RTS (Request-To-Send, with frame length)

2) The receiver replies with a CTS (Clear-To-Send, with frame length)

3) Sender transmits the frame while nodes hearing the CTS stay silent

- 任意收到 RTS 的站点需要等待一个时间来保证 A 接受了 CTS.
- 任意收到 CTS 的站点需要保持静默直到数据传输完成, 静默时间可以由 CTS 中信息决定.

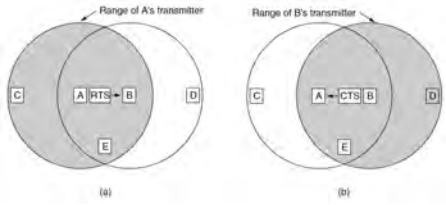

Figure IV.11: The MACA protocol. (a) A sending an RTS to B. (b) B responding with a CTS to A.

3. Ethernet(IEEE 802.3)

Two kinds of Ethernet exist

- Classical Ethernet: 3 to 10 Mbps
- Switched Ethernet: runs at 100, 1000 and 10,000 Mbps

Over the cables, information was sent using **the Manchester encoding**.

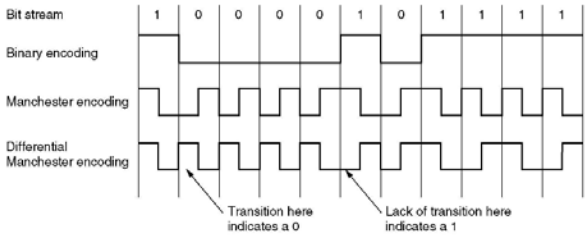

Figure IV.12: the Manchester encoding

Ethernet brief history:

1) In the mid-1970s, a bus topology. (A broadcast LAN)

2) By the late 1990s, a hub-based star topology. A hub is a physical layer device. (A broadcast LAN)

3) In the early 2000s, a switch — a switch-based star topology (switched Ethernet). A switch is a link layer device. Modern switches are full-duplex. In a switch-base Ethernet LAN there are no collisions.

Ethernet is by far the most prevalent wired LAN technology. All of the Ethernet technologies provide **connectionless service** to the network layer.

3.1 Ethernet Frame Structure

- **Preamble**: 8 bytes, each containing the bit pattern 10101010 (In 802.3, is 10101011). The last byte is called **the Start of Frame** delimiter for 802.3. It produces a 10-MHz square wave, allow the receiver's clock to **synchronize**.

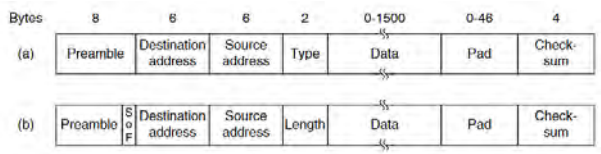

Figure IV.13: Ethernet Frame formats. (a) Ethernet (DIX). (b) IEEE 802.3.

- **Two addresses** (MAC addresses): 6*2 bytes.

The first transmitted bit of the destination address is a 0 for ordinary addresses and a 1 for group addresses.

The special address consisting of all 1 bits is reserved for broadcasting.

The 48-bit number of source address. The first 3 bytes of the address field are assigned by IEEE(IEEE 给网卡制造商分配的). The last 3 bytes of the address and programs the complete address into the NIC(网卡制造商烧录进网卡的 id).

- The **Type or Length** field: 2 bytes

Any number there less than or equal to 0x600 (1536) can be interpreted as Length, and any number greater than 0x600 can be interpreted as Type.

a type code of 0x0800 means that the data contains an IPv4 packet.

- The Data field: 46 to 1500 bytes.

This field carries the IP datagram. The maximum transmission unit (MTU) of Ethernet is 1500 bytes.

The minimum size of the data is 46 bytes. Ethernet 要求 from destination address to checksum 的 valid frame 长度必须至少为 64 bytes.

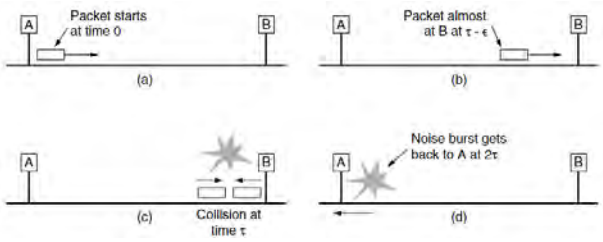

Figure IV.14: Collision detection can take as long as 2τ

- **Checksum**: a 32-bit CRC

3.2 CSMA/CD with Binary Exponential Backoff

Classic Ethernet uses the 1-persistent CSMA/CD algorithm.

How the random interval is determined when a collision occurs?

1) After a collision
2) After i collision
3) After 10 collision
4) After 16 collision, the controller *throws in the towel*(不干了) and reports failure back to the computer.

3.3 Ethernet Performance

Ethernet under conditions of heavy and constant load with k stations always ready to transmit.

If each station transmits during a contention slot with probability p, the probability A that some station acquires the channel in that slot is

$$A = kp(1-p)^{k-1}$$

A is maximized when $p = \frac{1}{k}$, with $A \to \frac{1}{e}$ as $k \to \infty$.

The probability that the contention interval has exactly j slots in it is $A(1-A)^{j-1}$, so the mean number of slots per contention is given by

$$\sum_{j=0}^{\infty} jA(1-A)^{j-1} = \frac{1}{A}$$

Since each slot has a duration 2τ, the mean contention interval $w = \frac{2\tau}{A}$.

If the mean frame takes P sec to transmit, transmit, when many stations have frames to send

$$\text{Channel efficiency} = \frac{P}{P + \frac{2\tau}{A}}$$

The longer the cable, the longer the contention interval.

In terms of the frame length, F, the network bandwidth, B, the cable length, L, and the speed of signal propagation, c, for the optimal case of e contention slots per frame. With $P = \frac{F}{B}$,

$$\text{Channel efficiency} = \frac{1}{1 + \frac{2BLe}{cF}}$$

3.4 Switched Ethernet

In a hub, all stations are in the same collision domain. But in a switch, each port is its own independent collision domain.

3.5 Ethernet Technologies

There are many other Ethernet technologies have been standardized over the years by the IEEE 802.3 CSMA/CD (Ethernet) working group.

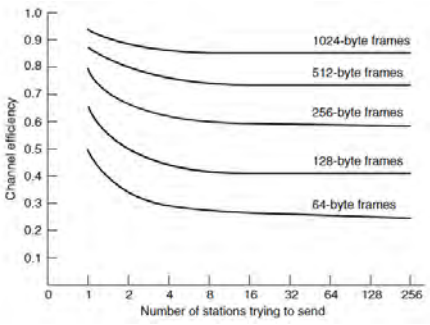

Figure IV.15: Efficiency of Ethernet at 10 Mbps with 512-bit slot times.

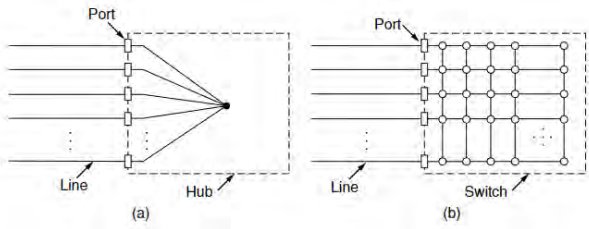

Figure IV.16: (a) Hub. (b) Switch.

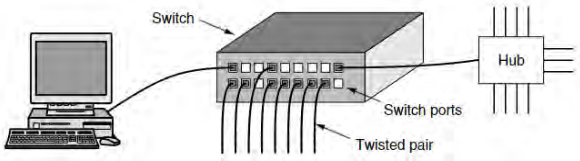

Figure IV.17: An Ethernet switch

e.g. 10BASE-T, 10BASE-2, 100BASE-T, 1000BASE-LX and 10GBASE-T.

Ethernet is both a link-layer and a physical-layer specification.

3.6 Gigabit Ethernet

All configurations of gigabit Ethernet use point-to-point links.

supports two different modes of operation:

- full-duplex mode: there is a central switch connected to computers.
- half-duplex mode: the computers are connected to a hub

Carrier extension: at least 512 bytes

Frame bursting

4. Wireless LANS (802.11 WiFi)

802.11 networks can be used in two modes

- In infrastructure mode, each client is associated with an AP (Access Point), and APs may be connected together by a wired network
- Ad hoc network, there is no access point, and each computer can sent frames to each other directly.

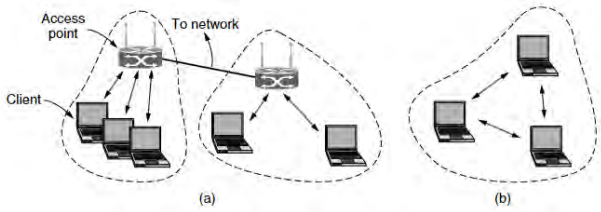

Figure IV.18: 802.11 architecture. (a) Infrastructure mode. (b) Ad-hoc mode.

The data link layer in all the 802.11 protocols is split into two or more sublayers:

- The MAC (Medium Access Control)
- The LLC (Logical Link Layer)

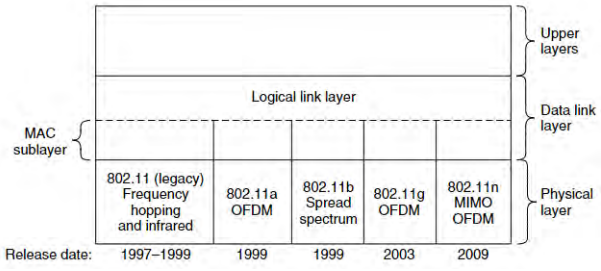

Figure IV.19: Part of the 802.11 protocol stack

4.1 802.11 Physical Layer

Either the 2.4 GHz or the 5GHz ISM (Industrial, Scientific and Medical) frequency bands.

4.2 The 802.11 MAC Sublayer Protocol

Problems with wireless:

- Half duplex: cannot transmit and listen for noise bursts at the same time on a single frequency
- Limited Range: The hidden station problem and the exposed station problem

The 802.11 tries to avoid collision with a protocol called CSMA/CA (CSMA with Collision Avoidance).

- Channel sensing before sending
- Exponential backoff after collision

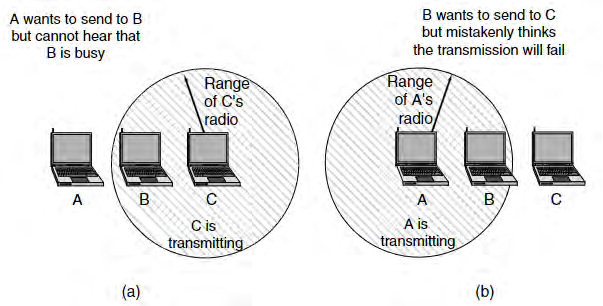

Figure IV.20: (a) The hidden terminal problem. (b) The exposed terminal problem.

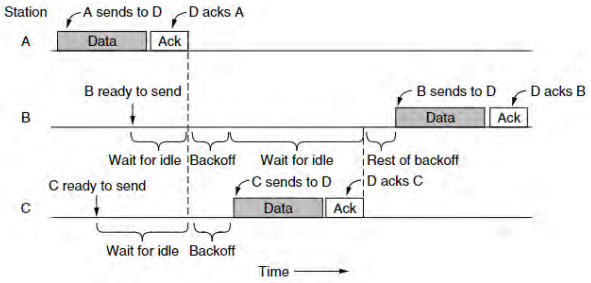

Figure IV.21: Sending a frame with CSMA/CA. (A, B and C want to send frames to D.)

Compared to Ethernet, there are main two differences:

1) starting backoffs early helps to avoid collision

2) acknowledgements are used to infer collision(collision cannot be detected.)

Two modes of operations (backoff):

1) DCF (Distributed Coordination Function) because each station acts independently.

2) PCF (Point Coordinate Function) in which the AP may act as the BS in the cellular network.

To reduce ambiguities about which station is sending, 802.11 defines channel sensing to consists of physical sensing and virtual sensing.

Method 1 Physical sensing

- If idle, just starts transmitting. Does not sense channel while transmitting.
- If collision occurs, wait random time, using Ethernet binary exponential backoff algorithm

Be used to transmit RTS frame.

Method 2 Virtual Sensing

Scenario: A wants to send to B. C is a station within

range of A. D is a station within range of B but not within range of A. (C A B D)

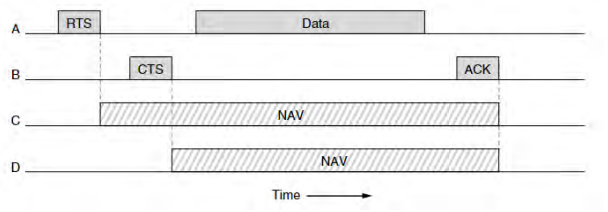

Figure IV.22: Virtual channel sensing using CSMA/CA

Note: the NAV signals are NOT transmitted; they are just internal reminder to keep quiet for a certain period of time.

There are several other mechanisms to improve its performance:

1) Reliability

To lower the transmission rate. send shorter frames, allow frames to be split into smaller pieces, called fragments.

2) Power: the basic mechanism for saving power builds on beacon frames

3) Quality of Services: After a frame has been sent, a certain amount of idle time is required before any station may send a frame to check that the channel is no longer in use.

4.3 The 802.11 Frame Structure

The 802.11 standard defines three different classes of frames in the air: data, control, and management.

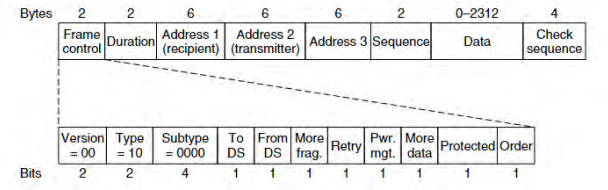

Figure IV.23: Format of the 802.11 data frame

- data frame
  - Address fields: data frames sent to or from an AP, but the AP may be a relay point, so the Address 3 gives a distant destination.
  - Protocol Version: set to 00
  - Type: data, control, or management; Subtype: RTS, CTS, or ACK of control frame; for a regular data frame (without quality of service), they are set to 10 and 0000 in binary.
  - To DS and From DS: to indicate whether the fame is going to or coming from the networks connected to the APs.
  - More Fragments: more fragments will follow;
  - Retry: retransmission;
  - Pwr: sleep state;
  - More data: sender has additional frames;
  - The Protected Frame bit: encrypted usign WEP;
  - The Order bit: processed strictly in order;
  - Duration: time the frame and its acknowledgement takes; it also exists in control frame.
  - Sequence: 16 bits, 12 for frame, 4 for fragment.
- Management frames: similar to data frames, except without one of the base station addresses, because management frames are restricted to a single cell.
- Control frames: shorter

Each 802.11 LAN must provide nine services

5. Data Link Layer Switching

We can treat one physical LAN as multiple logical LANs, called VLANs (Virtual LANs)

5.1 Learning Bridge

All of the stations attached to the same port on a bridge belong to the same collision domain.

have a big (hash) table. This table can list each possible destination and which output port it belongs to. use a flooding algorithm:

- Every incoming frame for an unknown destination is output on all the ports to which the bridge is connected except the one it arrived on
- Once a destination is known, frames destined for it are put only on the proper port

By looking at the source addresses, they can tell which machines are accessible on which ports.

But the network topology is dynamic. To handle it, the arrival time of the frame is noted in the entry.

Periodically, a process in the bridge scans the hash table and purges all entries more than a few minutes old.

The routing:

1) If the port for the destination address is the same as the source port, discard the frame.

2) If the port for the destination address and the source port are different, forward the frame on to the destination port.

3) If the destination port is unknown, use flooding and send the frame on all ports except the source port.

5.2 Protocol Processing at a Bridge

In the general case, relays at a given layer can rewrite the headers for that layer.

e.g.

5.3 Spanning Tree Bridges

Spanning Tree Algorithm

6. Repeaters, Hubs, Bridges, Switches, Routers, and Gateways

The layer matters because different devices use different pieces of information to decide how to switch

6.1 Repeaters and Hubs (Physical Layer)

6.2 Bridges and Switches (Data Link Layer)

6.3 Routers (Network Layer)

6.4 Gateways

6.5 Virtual LANs

V Newwork Layer

1. Overview of network layer

1.1 Two Important Network Layer Functions

The role of the network layer is to move packets from a sending host to a receiving host.

Each router has a forwarding (internal, routing) table. The forwarding table is indexed by either the destination address in the packet header or an indication of connection to which the packet belongs.

1) Forwarding (the main function of a router)

2) Routing (to build the forwarding table for each router): via routing protocols

The issues include the service provided to the transport layer and the internal design of the network.

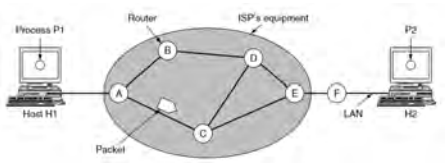

Figure V.1: The environment of the network layer protocols

1.2 Services Provided to the Transport Layer

Design goals:

- The services should be independent of the router technology
- The transport layer should be shielded from the number, type, and topology of the routers present
- The network addresses should use a uniform numbering plan.

Two warring factions, Connection-oriented & Connectionless service. Network layer should provide connection-oriented service or connectionless service.

The packets are frequently called datagrams.

Implementation of Connectionless Service Packets are injected into the network individually and routed independently of each other.

Every router has a forwarding table. Each table entry is a pair consisting of a destination and the outgoing line to use for that destination.

Implementation of Connection-Oriented Service Each packet carries an identifier telling which virtual circuit it belongs to.

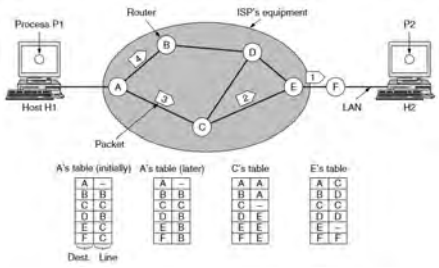

Figure V.2: Routing within a datagram network

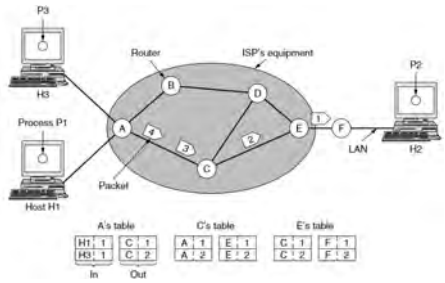

Figure V.3: Routing within a virtual-circuit network

Table V.1: Virtual-Circuit vs. Datagram Networks

—

—

2. Routing algorithms

Two functions of a router: 转发 + 建立路由表.
Difference in Datagram and Virtual Circuit:

- In datagram, route 实时更新
- In virtual circuit, route 仅在初始化时设定

2.1 Classification of Routing Algorithms

Classify routing algorithms into global routing algorithms and decentralized algorithms.

- Global routing algorithms: 适用于小网络. e.g. Link-state (LS) algorithms
- Decentralized routing algorithms: 适用于大网络. e.g. Distance-vector (DV) algorithms

A second broad way to classify routing algorithm:

- Static routing algorithms (non-adaptive)
- Dynamic routing algorithms (adaptive): 可能会摇摆, 有回路.

2.2 Link-state (LS) algorithms

In a link-state algorithm, the network topology and all link costs are known. 首先要 broadcast 以获得这些信息.

All nodes have an identical and complete view of the network.

The well-known LS algorithm is Dijkstra's algorithm.

The Shortest Path Algorithm The shortest path is one that has the least cost.

Measure path cost (length)

- Number of hops
- Delay
- distance
- Bandwidth
- Communication cost
- Average traffic

The optimality principle: If router J is on the optimal path from router I to router K, then the optimal path from J to K also falls along the same route.

Sink Tree Sink tree for a destination is the union of all shortest paths towards the destination. Similarly source tree.

Implications:

1) Only need to use destination to follow shortest paths
2) Each node only need to send to the next hop

Forwarding table at a node List next hop for each destination.

Dijkstra Algorithm For single source shortest path problem. Approach: Greedy.

The Complexity of Dijkstra Algorithm: worst-case complexity is $O(n^2)$.

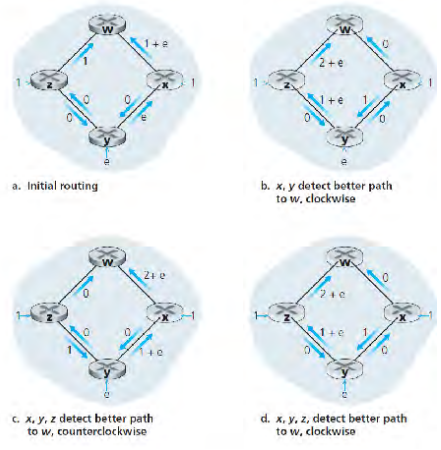

Figure V.4: The Oscillation Problem with the LS Algorithm

Flooding(洪放) Each router must make decisions based on local knowledge.In flooding, every incoming packet is sent out

on every outgoing line except the one it arrived on. 但会有 node 接收到重复的拷贝.

e.g. Consider a flood from A: first reaches B via AB, E via AE.

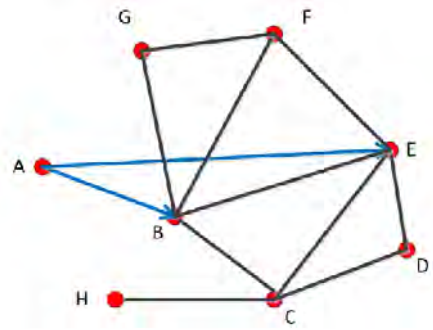

Figure V.5: Flooding Example

There are some solutions to avoid flooding duplicate packets.

1) Solution 1: A hop counter contained in the header of each packet, packet is discarded when the counter reaches zero.
2) Solution 2: Every node keeps track of which packets have been flooded, to avoid sending them out a second time.

Link State Routing :

1) Learning about the neighbors
2) Setting link costs
3) Building link state packets

   The hard part is determining when to build them.
4) Distributing the link state packets

   The fundamental idea is to use flooding to distribute the link state packets to all routers.

   The solution to all these problems is to include the age of each packet

   To guard against errors on the links, all link state packets are acknowledged.
5) Computing the new routes

2.3 Distance-vector (DV) algorithms

The distance vector routing algorithm is iterative, asynchronous, and distributed.

- Distributed: 从邻居获取信息, 并给邻居信息.
- Iterative: 迭代直到邻居没有信息变更. It is self-terminating.
- Asynchronous: 不要求同步.

DV-like algorithms are used in many routing protocols in practice, including the Internet's RIP, BGP, and so on.

Bellman-Ford Equation Let $d_x(y)$ be the cost of the least-cost path from node x to node y. Then the least costs are related by the celebrated Bellman-Ford equation, namely,

$$d_x(y) = \min_v\{c(x,v) + d_d(y)\}$$

where the $\min_v$ in the equation is taken over all of x's neighbors.

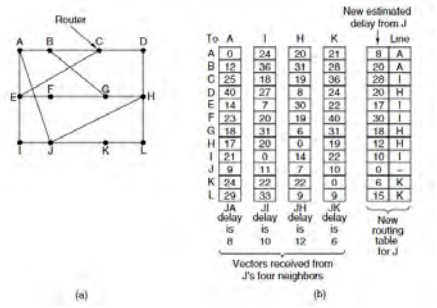

Figure V.6: (a) A network. (b) Input from A, I, H, K, and the new routing table for J

The Distance Vector Routing Algorithm The basic idea: each node x begins with $D_x(y)$, an estimate of the cost of the least-cost path from itself to node y, for all nodes in N (the number of nodes in the network). Let $D_x(y) = [D_x(y) : y \in N]$ be node x's distance vector, which is the vector of cost estimates from x to all other nodes, $y \in N$.

With the DV algorithm, each node x maintains the following information:

- $c(x,v)$: The cost, for each neighbor node v.
- $D_x(y) = [D_x(y) : y \in N]$: Node x's distance vector.
- $D_v(y) = [D_v(y) : y \in N]$: The distance vectors of each of its neighbors.

1) 接受新的 distance vector 后进行更新.
2) 有更新就广播给邻居
3) 直到不变, $D_x(y)$ 收敛于 $d_x(y)$.

最后达到 A quiescent state.

The Count-to-Infinity Problem 收敛可能所需时间过长.

Suppose node A is down initially and all other routers know this. In other words, they have all recorded the delay to A as infinity.

2.4 Hierarchical Routing

the routers are divided into what we will call regions.

Table V.2: Distance Vector Routing vs. Link State Routing

Goal	Distance Vector	Link State
Correctness	Distributed Bellman-Ford	Replicated Dijkstra's Algorithm
Efficient Paths	Approx. with shortest paths	Approx. with shortest paths
Fair Paths	Approx. with shortest paths	Approx. with shortest paths
Convergence	Slow (many exchanges)	Fast (Flood and compute)
Scalability	Excellent (storage/compute)	Moderate (storage/compute)

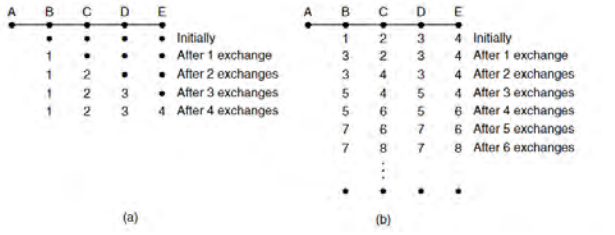

Figure V.7: The count-to-infinity problem

When routing is done hierarchically, there are entries for all the local routers, but all other regions are condensed into a single router.

2.5 Broadcast Routing

the network layer provides a service of delivering a packet sent from a source node to all other nodes in the network.

1) One of the most simple and straightforward way is to send a separate copy of packet to each destination.
2) Multidestination routing: Each packet contains either a list of destinations or a bit map indicating the desired destinations
3) Uncontrolled flooding
the source node sends a copy of the packet to all of its neighbors
Sequence-number-controlled flooding
Reverse path forwarding (RPF)
4) Spanning-Tree Broadcast
core: the creation and maintenance of the spanning tree. The center-based approach to building a spanning tree:

a. A center node is defined.
b. Nodes then unicast tree-join messages addressed to the center node. A tree-join message is forwarded using unicast routing toward the center until it either arrives at a node that already belongs to the spanning tree or arrives at the center.
c. In either case, the path that the tree-join message has followed defines the branch of the spanning tree between the edge node that initiated the tree-join message and the center. Graft.

2.6 Multicast Routing

2.7 Anycast Routing

- Unicast — a single destination
- Broadcast — to all destinations
- Multicast — to a group of destinations

In anycast, a packet is delivered to the nearest member of a group. Anycast is used in the Internet as part of DNS.

3. The network layer in the Internet

There are two basic choices for connecting different networks:

1) build devices that translate or convert packets
2) building a common layer on top of the different networks.

IP is the foundation of the modern Internet.

The network layer of the internet has three main components:

- The IP protocol
- The Internet control protocols (including ICMP, DHCP, ARP)
- The Internet routing protocols (including RIP, OSPF and BGP)

4. IP Protocol

4.1 The IPv4 Datagram

The header has a 20-byte fixed part and a variable-length optional part. The bits are transmitted from left to right and top to bottom. This is "big-endian" network byte order.

1) The Version field (4 bits)
2) IHL field (Header Length) (4 bits)

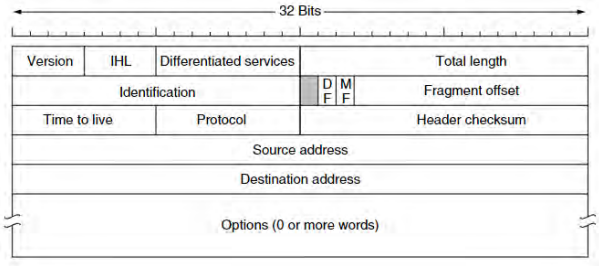

Figure V.8: The IPv4 (Internet Protocol) header

tell how long the header is in 32-bit words, limits the header to 60 bytes, thus the options field to 40 bytes.

the typical IP datagram has a 20-bytes header.

3) The Different Service field (Type of Service) (8 bits)

The top 6 bits are used to mark the packet with its service class

The bottom 2 bits are used to signal **explicit congestion indications**

4) The Total Length field (16 bits)

The total length of header and data.

The theoretical maximum length is 65,535 bytes.

Datagrams are rarely larger than 1,500 bytes. (由以太网的 MTU 决定)

5) The Identification field (16 bits), flags (DF, MF), and fragment offset (13 bits)

All the fragments of a packet contain the same identification field.

- The Unused bit
- DF – Don't Fragment
- MF – More Fragments

  (MF = 0 means this the last fragment)
- The Fragment Offset field (13 bits)

  There is a maximum of 8192 fragments per datagram

6) The TtL (Time to live) field (8 bit)

This field is decremented by one each time the datagram is processed by a router.

it just counts hops. When it hits zero, the packet is discarded and a warning packet is sent back to the source host.

7) The Protocol field (8 bit)

tells is which transport process to give the packet to. (6 is TCP, 17 is UDP)

The protocol number connect the network and transport layer

the port number connect that binds the transport and application layers

8) The header checksum

in the header and using one's complement arithmetic.

the checksum must be recomputed and stored again at each router, as the TTL field, and possibly the options fields as well, may change.

9) The Source address and Destination address (each with 32 bit)

10) The Options field 用得少

11) Data (payload)

Packet Fragmentation The maximum payloads of different networks

- Ethernet – 1500 bytes
- 802.11 – 2272 bytes
- IP – 65,515 bytes

Two Solutions:

1) To make sure the packet fragmentation does not occur in the 1st place

2) To break up packets into fragments, sending each fragment as a separate network layer packet.

Two opposing strategies exist for recombining the fragments back into the original packet

- Transparent fragmentation is straightforward but has some problems
- Nontransparent fragmentation is to refrain from recombining fragments at any intermediate routers.

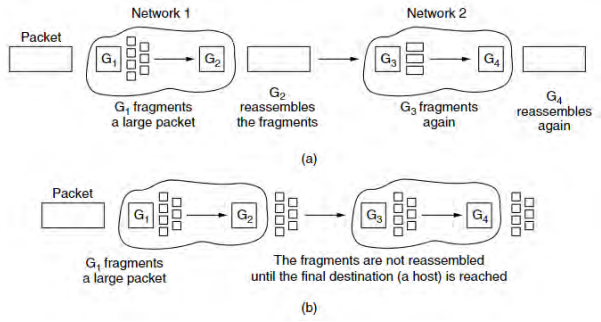

Figure V.9: (a) Transparent fragmentation. (b) Nontransparent fragmentation.

e.g.

Path MTU discovery

The disadvantage of path MTU discovery is that there may be added startup delays simply to send a packet, more than one round-trip delay may be needed to probe the path

4.2 IPv4 Addressing

IP 地址有 32 位. 其指向的是一个网络界面, 而不是一台主机. 其是分级的. 其由顶 network portion 与底 a host portion 组成. IP 也可用 16 进制表示.

Addresses are allocated in blocks called prefixes.

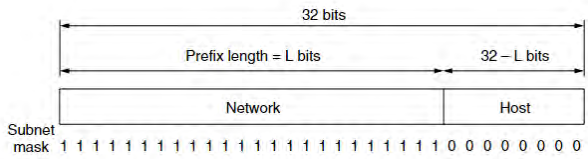

Figure V.10: An IP prefix and a subnet mask

IP Prefixes (Network portion) IP address/length. The /N sometimes known as a subnet mask.

The key advantage of prefixes is that routers can forward packets based on only the network portion of the address.

Subnet 子网个数等于网络中断开路由器后孤岛的个数. 子网 network portion 的地址一致. 将地址块分割成若干部分以供内部用作多个网络.

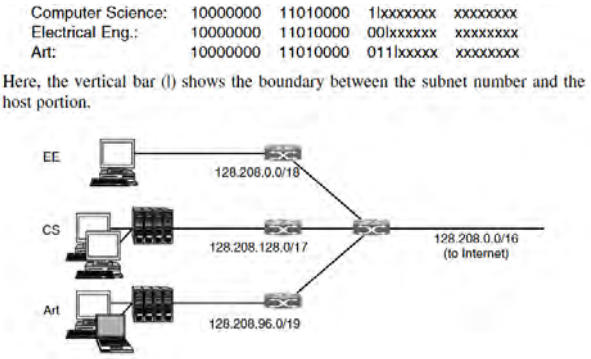

Figure V.11: Splitting an IP prefix into separate networks with subnetting

When a packet comes into the main router, how does the router know which subnet to give it to?

1) One solution is that for each router to have a table with 65536 entries telling it which outgoing line to use for each host on campus.

2) The other way is that the router can do this by ANDing the destination address with the mask for each subnet and checking to see if the result is the corresponding prefix.

IP Address Classes - Historical Before CIDR, the network portions of an IP address were constrained to be 8, 16, or 24 bits in length, and addressing scheme known as classful addressing.

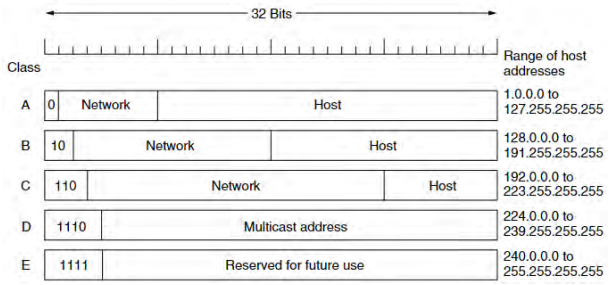

Figure V.12: IP address formats

分配 IP 地址的机构: the Internet Corporation for Assigned Names and Numbers (ICANN) following a hierarchical process

Classless InterDomain Routing(CIDR) There is still a problem: routing table explosion.

CIDR solution:

- Subnetting: split IP prefixes
- Route aggregation: combine multiple small prefixes into a single larger prefix.

The Longest Matching Prefix The rule is that packets are sent in the direction of the most specific route or the longest matching prefix that has the fewest IP address

Special IP Addresses

- 0.0.0.0: means “this network” or “this host”.
- 255.255.255.255: mean all hosts on the indicated network, allows broadcasting
- 127.0.0.1 (本机地址)

4.3 Protocol

Network Address Translation(NAT) IP addresses are scarce.

Solution:

1) DHCP
2) NAT box (Network Address Translation box)

The NAT translation table includes port numbers as well as IP addresses in the table entries. 通过端口号区分内部的电脑.

Ports are effectively an extra 16 bits of addressing that

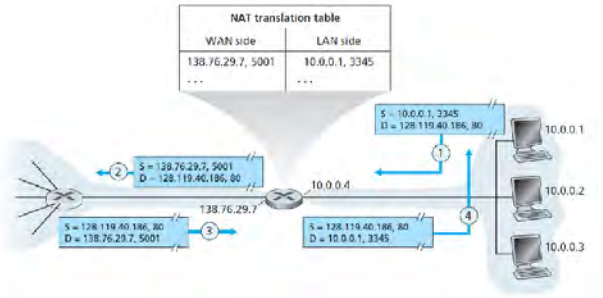

Figure V.13: Network Address Translation

identify which process gets which incoming packet. Ports 0-1023 are reserved for well-known services

Problems with NAT:

- Port numbers are meant to be used for addressing processes.
- changes the Internet from a connectionless network to a peculiar kind of connection-oriented network.
- Routers are supposed to process packets only up to layer 3.
- violates the most fundamental rule of protocol layering

Dynamic Host Configuration Protocol DHCP is often referred to as a plug-and-play protocol. DHCP is a client-server protocol, each subnet will have a DHCP server.

In addition to host IP address assignment, DHCP also allows a host to learn additional information, such as its subnet mask, the address of its first-hop router (often called the default gateway), and the address of its local DNS server.

the DHCP protocol is a four-step process:

1) DHCP server discovery

using a DHCP discover message, which a client sends within a UDP packet to port 67.

broadcast destination IP address of 255.255.255.255 and a "this host" source IP address of 0.0.0.0.

2) DHCP server offer(s)

A DHCP server receiving a DHCP discovery message responds to the client with a DHCP offer message that is broadcast to all nodes on the subnet

Several DHCP servers can be present on the subnet, the client may choose from among several offers.

3) DHCP request

The newly arriving client will choose from among one or more server offers and respond to its selected offer with a DHCP request message echoing back the configuration parameters.

4) DHCP ACK

The server responds to the DHCP request message with a DHCP ACK message, confirming the requested parameters.

Once the client receives the DHCP ACK, the interaction is complete and the client can use the DHCP-allocated IP address for the lease duration.

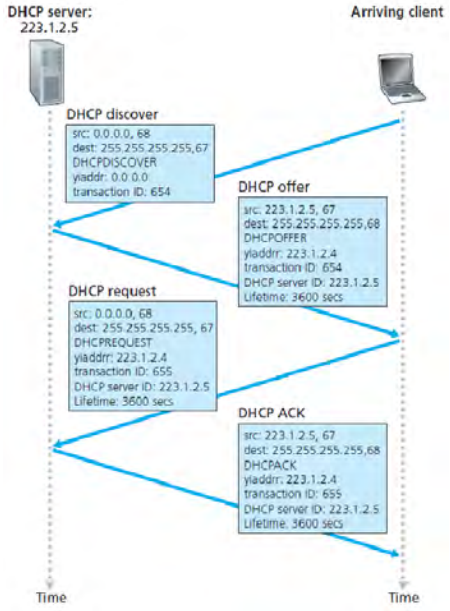

Figure V.14: DHCP

4.4 IPv6 datagram

IPv6 uses 128-bit addresses

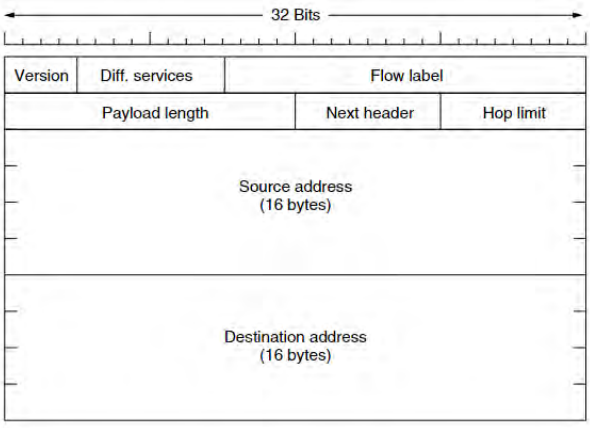

Figure V.15: The IPv6 fixed header

1) the Version field (4 bits)
2) the Difference Services field (8 bits)
3) the Flow Label field (20 bits)

4) the Payload length field (16 bits)

5) the Next Header field (8 bits)

6) the Hop Limit field (8 bits)

7) the Source address and Destination address fields (each with 128 bits or 16 bytes)

4.5 The IPv6 Address

Since many addresses will have many zeros inside them, three optimization have been authorized:

1)

Several Fields in IPv4 are no longer present in the IPv6 datagram

- Fragmentation/reassembly
- Header checksum
- Options

Transitioning from IPv4 to IPv6: two approaches

1) IPv6-capable nodes is a dual-stack approach, where IPv6 nodes also have a complete IPv4 implementation

2) Tunneling. To take the entire IPv6 datagram into the data (payload) field of an IPv4 datagram.

Tunneling

5. Control Protocols

5.1 Internet Control Message Protocol (ICMP)

The most typical use of ICMP is for error reporting. ICMP is often considered part of IP but architecturally it lies just above IP, as ICMP messages are carried inside IP datagrams.

Use of ICMP — "ping"

- ping sends an ICMP type 8 code 0 message (echo request) to the specified host.
- The destination host, seeing the echo request, sends back a type 0 code 0 ICMP echo reply.

Use of ICMP — "Tracert" Tracert is implemented with ICMP messages, to determine the names and addresses of the routers between source and destination

1) Tracert in the source sends a series of ordinary IP datagrams to the destination.

2) When the nth datagram arrives at the nth router, the nth router observes that the TTL of the datagram has just expired.

3) When this ICMP message arrives back at the source, the source obtains the round-trip time from the timer and the name and IP address of the nth router from the ICMP message

4) How does a Tracert source know when to stop sending UDP segments?

5.2 ARP (The Address Resolution Protocol)

The purpose of the ARP query packet is to query all the other nodes on the subnet to determine the MAC address corresponding to the IP address that is being resolved.

e.g. How a user on host 1 sends a packet to a user on host 2 on the CS network?

1)

Various Optimizations of ARP

1) Cache
2) The default gateway
   Through the Default Gateway

   a.

3) Proxy ARP

ARP vs. DNS

- ARP resolves an IP address to a MAC address only for nodes on the same subnet
- DNS resolves host names to IP addresses for hosts anywhere in the Internet.

ARP is probably best considered a protocol that straddles the boundary between the link and network layers

6. Routing Protocols

A two-level routing algorithm:

- Distance vector routing
- Link state routing

The networks may all use different intradomain protocols, but they must use the same interdomain protocol.

6.1 The Internet

In the network layer, the Internet can be viewed as a collection of networks or ASes (Autonomous Systems) that are interconnected. The glue that holds the whole Internet together is the network layer protocol, IP (Internet Protocol).

6.2 Autonomous System(AS)

we will examine two intra-AS routing protocols (RIP and OSPF) and the inter-AS routing protocol (BGP) that are used in today's Internet

Intra-AS Routing in the Internet: RIP RIP (Routing Information Protocol) is a distance-vector protocol.

RIP uses hop count as a cost metric.

•

1) RIP 从已配置接口发送整个路由表，该表以广播或组播 (Multicast, 目的地址 224.0.0.9) 的形式周期性向所有节点发送。

2) “路由更新”定时器规定了周期性广播或组播的频率，缺省值定为 30 秒。

3) 如果达到无效 (invalid) 超时时间 (缺省为 180 秒) 时，路由器仍未收到某路由器的路由更新信息，将把该条路由标记为无效，认为不可达，标记为无效的方式就是将跳数设定为 16。

4) 被标记为不可达的路由仍将保存在路由表里，直到清除定时器超时后，才被清除掉，清除定时器的缺省值为 240 秒。

5) 当某一路由项 RI 包含的目的网络变成不可达到后，该路由项对应的抑制定时器开始计时，在此定时器超时前，即使路由器收到的路由更新指明路由项 RI 又可达到了，路由项 RI 仍被本路由器标识为不可到达，只有在抑制定时器超时后，指明路由项 RI 可到达的路由更新才起作用。抑制定时器的目的在于使路由稳定，不要发生路径摇摆。

OSPF: An Interior Gateway Routing Protocol OSPF (Open Shortest Path First, 开放式最短路径优先协议). Link state routing.

The higher data rate, the lower cost of the channel.

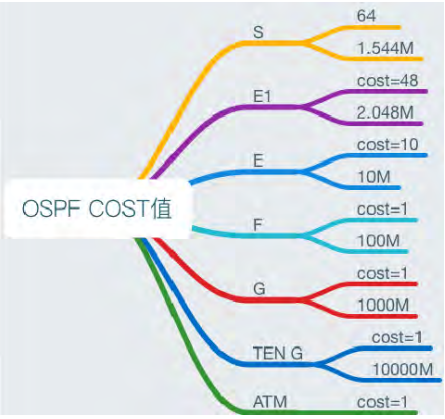

Figure V.16: OSPF Cost Table

OSPF allows an AS to be divided into numbered areas. An area is a generalization of an individual network. Routers that lie wholly within an area are called internal routers.

Every AS has a backbone area, called area 0. The routers in this area are called backbone routers. All areas are connected to the backbone, possibly by tunnels.

Each router that is connected to two or more areas is called an area border router (ABR). It must also be part of the backbone. The job of an area border router is to summarize the destinations in one area and to inject this summary into the other areas to which it is connected. This summary includes cost information but not all the details of the topology within an area. If there is only one border router out of an area, even the summary does not need to be passed. This kind of area is called a stub area.

Each router within an area has the same link state database and runs the same shortest path algorithm.

the designed router and A backup designed router

7. BGP: The Exterior Gateway Routing Protocol

BGP is an absolutely critical protocol for the Internet — in essence, it is the protocol that glues the whole thing together.

In BGP, pairs of routers exchange routing information over semipermanent TCP connections using port 179. In BGP, destinations are not hosts but instead are CIDRized prefixes. In BGP, an autonomous system is identified by its globally unique autonomous system number (ASN). In BGP jargon, *a prefix along with its attributes* is called **a route**.

7.1 BGP Route Advertising

Route advertisements contain **an IP prefix, AS-path, next hop**. Route advertisements move in the opposite direction to traffic.

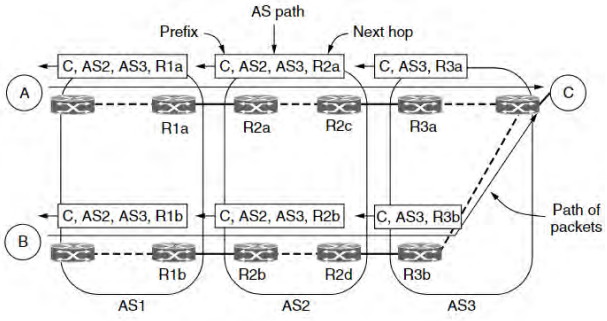

Figure V.17: Propagation of BGP route advertisements

TRANSIT service(TR) and PEER service(PE). Note that peering is not transit

7.2 BGP Policy

Policy is implemented in two ways:

1) Border routers of ISP announce paths only to other parties who may use those paths

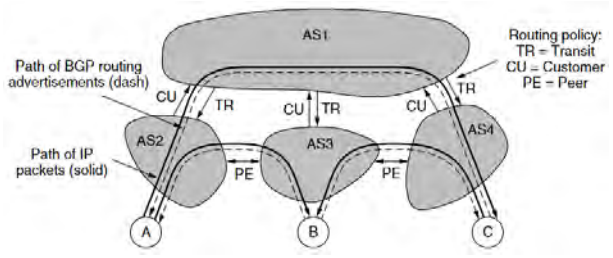

Figure V.18: Routing policies between four autonomous systems

2) Border routers of ISP select the best path of the ones they hear in any, non-shortest way

8. MPLS

MPLS (Multiprotocol Label Switching) is perilously close to circuit switching.

- To improve the forwarding speed of IP routers by adopting a key concept from the world of virtual-circuit networks: a fixed-length label
- allowing routers to forward datagrams based on fixed-length labels (rather than destination IP addresses)

1) The 1st question to ask is where does the label go?

MPLS falls between the network layer protocol and the data link layer protocol. MPLS is sometimes described as a layer 2.5 protocol.

2) the 2nd question is to ask when and how the labels are attached to packets?

Within the MPLS network, the label is used to forward the packet.

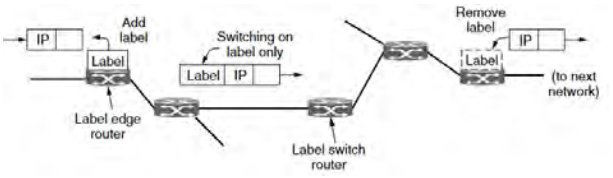

Figure V.19: Forwarding an IP packet through an MPLS network

3) The final question we will ask is how the label forwarding tables are set up so that packets follow them

a setup packet is launched into the network to create the path and make the forwarding table entries. allocates a label for each one and passes the labels to its neighbors.

9. Internet Multicasting

IP supports one-to-many communication, or multicasting, using class D IP addresses. 28 bits are available for identifying groups, so over 250 million groups can exist at the same time.

Some examples of local multicast addresses are:

-

VI Transport Layer

1. Overview of the transport layer

Together with the network layer, the transport layer is the heart of the protocol hierarchy. The transport layer provides end-to-end connectivity across the network.

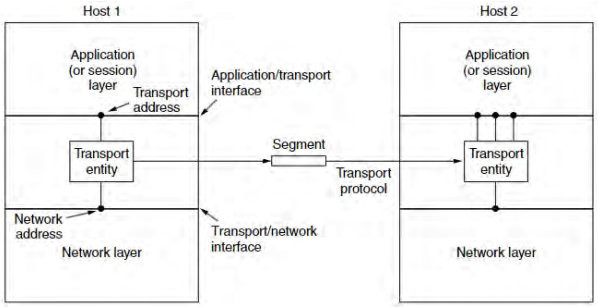

Figure VI.1: The network, transport, and application layers

1.1 The Transport Service

Two types of transport service

- Connection-oriented transport service
- Connectionless transport service

The transport code runs entirely on the users'machines, but the network layer mostly runs on the routers. The network service is generally unreliable.

1.2 Transit Units of Different Layers

- Transport layer: segment or TPDU (Transport Protocol Data Unit)
- Network layer: packet
- Data link layer: frame
- Physical layer: bit

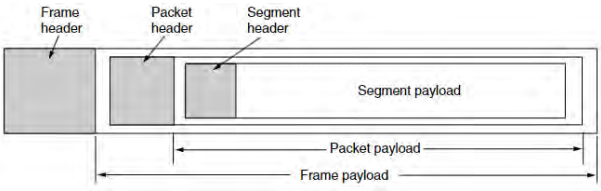

Figure VI.2: Nesting of segments, packets, and frames

The Internet has two main protocols in the transport layer:

- UDP (User Datagram Protocol, connectionless protocol): It does nothing beyond sending packets between applications. It typically runs in the operating system.
- TCP (connection-oriented protocol): It does almost everything. It makes connections and adds reliability with retransmission, along with flow control and congestion control.

2. The internet transport protocols: UDP

UDP: connectionless transport protocol.

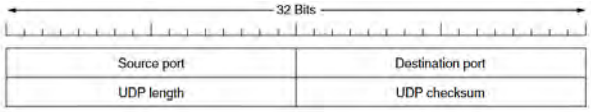

Figure VI.3: The UDP header

The two ports serve to identify the endpoints within the source and destination machines

- UDP header (8 bytes)
  - The UDP length field includes the 8-byte header and the data

    The minimum length is 8 bytes. The maximum length is 65,515 bytes.
  - The UDP checksum field (optional) is to provide extra reliability

    It checksums the header, data, and a conceptual IP pseudoheader

    The checksum algorithm is simply to add up all the 16-bit words (note here a word = 16 bits = 2 bytes) in one's complement and to take the one's complement of the sum.
- The IPv4 pseudoheader

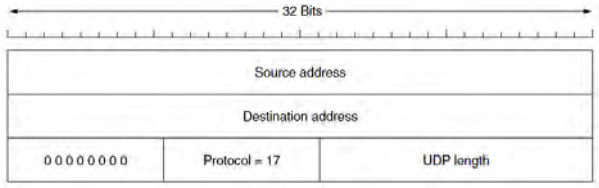

Figure VI.4: The IPv4 pseudoheader included in the UDP checksum.

What UDP does not do: Flow control, congestion control, or retransmission upon receipt of a bad segment.

What UDP does do:

- To provide an interface to the IP protocol with the added feature of demultiplexing multiple processes using the ports.
- Optional end-to-end error detection (checksum)

The application uses the UDP protocol:

- DNS (Domain Name System, Chapter 7)
- SSDP (Simple Service Discovery Protocol)

2.1 Real-Time Transport Protocol (RTP)

It is a transport protocol but just happens to be implemented in the application layer.

Two aspects of real-time transport:

- The RTP protocol for transporting audio and video data in packets
- How the receiver plays out the audio and video at the right time?

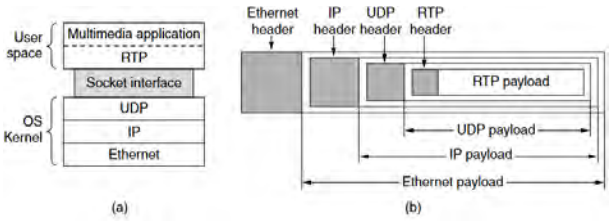

Figure VI.5: (a) The position of RTP in the protocol stack. (b) Packet nesting.

The basic function of RTP is to multiplex several real-time data streams onto a single stream of UDP packets

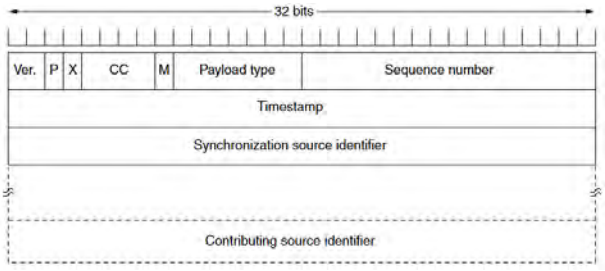

Figure VI.6: The RTP header

It consists of three 32- bit words and potentially some extensions

- The Version field: 2
- The P bit indicates that the packet has been padded to a multiple of 4 bytes. The last padding byte tells how many bytes were added.
- The X bit indicates that an extension header is present.
- The CC field tells how many contributing sources are present, from 0 to 15.
- The M bit field is an application-specific marker bit.
- The Payload type field tells which encoding algorithm has been used.
- The Sequence number is just a counter that is incremented on each RTP packet sent. It is used to detect lost packets.
- The Timestamp is produced by the stream's source to note when the 1st sample in the packet was made.
- The Synchronization source identifier tells which stream the packet belongs to.
- The Contributing source identifier, if any, are used when mixers are present in the studio.

2.2 RTCP

The RTCP (Real-time Transport Control Protocol) is a little sister protocol of RTP

Playout with Buffering and Jitter Control

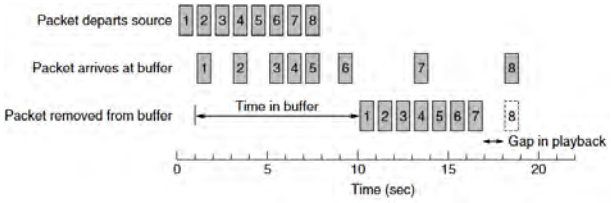

Figure VI.7: Smoothing the output stream by buffering packets

3. The internet transport protocols: TCP

TCP (Transmission Control Protocol) was designed to provide a reliable end-to-end byte stream over an unreliable internetwork.

An internetwork may have wildly different topologies, bandwidths, delays, packet sizes, and other parameters in different parts.

3.1 The TCP Service Model

TCP service is obtained by both the sender and the receiver creating end points, called sockets. Each socket has a socket number (address) consisting of the IP addressing of the host and a 16-bit number local to that host, called a port.

All TCP connections are full duplex and point-to-point. TCP does not support multicasting and broadcasting.

A TCP connection is a byte stream, not a message stream. (报文粘连问题)

When an application passes data to TCP, TCP may send it immediately or buffer it.

- PUSH flag: send immediately
- URGENT flag: high priority

The TCP Protocol includes: a fixed 20-byte header + <optional> + <0-N data bytes>.

Figure VI.8: (a) Four 512-byte segments sent as separate IP datagrams. (b) The 2048 bytes of data delivered to the application in a single READ call.

Port	Protocol	Use
20, 21	FTP	File transfer
22	SSH	Remote login, replacement for Telnet
25	SMTP	Email
80	HTTP	World Wide Web
110	POP-3	Remote email access
143	IMAP	Remote email access
443	HTTPS	Secure Web (HTTP over SSL/TLS)
543	RTSP	Media player control
631	IPP	Printer sharing

Figure VI.9: Some assigned ports

The form of data exchange: segment.

The basic TCP protocol: the sliding window protocol with dynamic window size.

There are many problems to solve:

1) Segment can arrive out of order
2) Segments can also be delayed

3.2 The TCP Segment Header

A key feature of TCP, and one that dominants the protocol design, is that every byte on a TCP connection has its own 32-bit sequence number. Every segment begins with a fixed-format, 20-byte header.

Segments without any data are legal and are commonly used for acknowledgements and control messages.

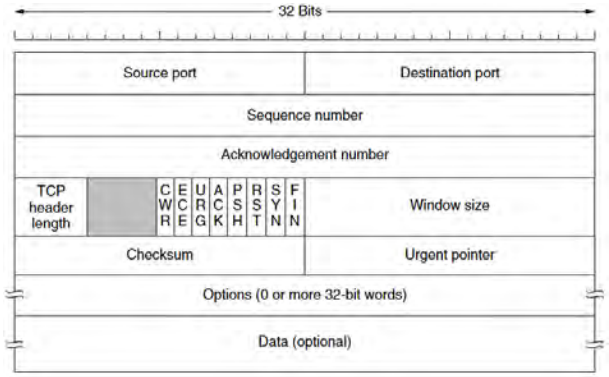

Figure VI.10: The TCP header

1) The Source port (16 bits) and Destination port (16 bits)

The connection identifier is a 5 tuple because it consists of five pieces of information: the protocol (TCP), source IP, and source port, and destination IP and destination port.

2) The Sequence number (32 bits) and Acknowledgement number (32 bits) fields

The Acknowledgement number specifies the next in-order byte expected, not the last byte correctly received.

It is a cumulative acknowledgement because it summarizes the received data with a single number.

3) The TCP header length (4 bits): because of the Options field.

4) The not used 4-bit field

5) eight 1-bit fields (ack, syn, fin is important)

- CWR and ECE are used to signal congestion when ECN (Explicit Congestion Notification) is used.
  ECE is set to signal an ECN-Echo to a TCP sender
  CWR is set to signal Congestion Window Reduced from the TCP sender
- URG is set to 1 if the Urgent pointer is in use.
  The Urgent pointer is used to indicate which urgent data are to be found
- The ACK bit is set to 1 to indicate that the Acknowledgement number is valid
- The PSH bit indicates PUSHed data
- The RST bit is used to abruptly reset a connection
- The SYN bit is used to establish connections
  - The connection request has SYN = 1 and ACK= 0
  - SYN = 1 and ACK = 1: the connection reply does bear an acknowledgement.
- The FIN bit is used to release a connection.
  Both SYN and FIN segments have sequence numbers and are thus guaranteed to be processed in the corrected order.

6) The Window size field (16 bits) tells how many bytes may be sent starting at the byte acknowledged

A window size field of 0 is legal, would like no more data for the moment.

7) Checksum (16 bits). It checksum the header, the data and a conceptual pseudoheader.

8) The Options field. may extended to 40 bytes to accommodate the longest TCP header

- MSS(Maximum Segment Size), it defaults to a 536-byte load.
- The window scale negotiate a window scale factor at the start of a connection

Figure VI.11: The pseudoheader of TCP

- The timestamp
- The SACK (Selective ACKnowledgement), is used after a packet has been lost but subsequent (or duplicate) data has arrived.

3.3 TCP Connection Establishment

Three Way Handshake

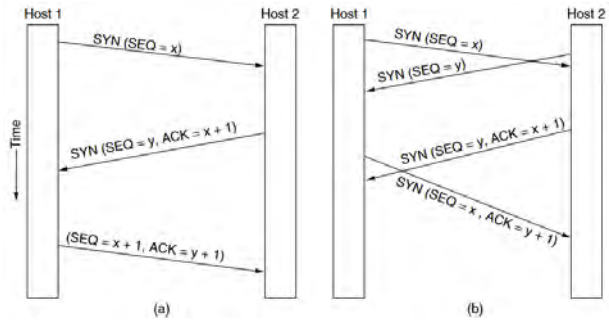

Figure VI.12: (a) TCP connection establishment in the normal case. (b) Simultaneous connection establishment on both sides.

1) Note that a SYN segment consumes 1 byte of sequence space so that it can be acknowledged unambiguously.

2) The initial sequence number chosen by each host should cycle slowly. This rule is to protect against delayed duplicated packets.

1) SYN — for establishing a connection

Request segment contains the following information in TCP header:

a. Initial sequence number (randomly chosen by the client)
b. SYN bit set to 1.
c. Maximum segment size
d. Receiving window size (the limit of unacknowledged data that can be sent to the client, contained in the window size field)

2) SYN + ACK — After receiving the request segment

Reply segment contains the following information in TCP header:

a. Initial sequence number (randomly chosen by the server)
b. SYN bit set to 1.
c. Maximum segment size
d. Receiving window size
e. Acknowledgement number
f. ACK bit set to 1.

3) ACK — After receiving the reply segment sending a pure acknowledgement. Not necessary.

Important Points

- Connection establishment phase consume 1 sequence number of both sides. But pure acknowledgement do not consume any sequence number.
- Pure acknowledgement for the reply segment is not necessary. Client sending the data packet immediately can be considered as an ack.
- For all the segments except the request segment, ACK bit is always set to 1.
- Certain parameters are negotiated during connection establishment.

  1) Window size
  2) Maximum segment size
  3) Timer values

- In any TCP segment
  - If SYN bit = 1 and ACK bit = 0, then it must be the request segment.
  - If SYN bit = 1 and ACK bit = 1, then it must be the reply segment.
  - If SYN bit = 0 and ACK bit = 1, then it can be the pure ACK or segment meant for data transfer.
  - If SYN bit = 0 and ACK bit = 0, then this combination is not possible.

SYN Flood Attack A malicious sender 发送大量 SYN segments, 但不完成链接.

用 SYN cookies 对抗. 即将 sequence number hash, 这样接受时只需要比对 hash 是否一致.

3.4 TCP Connection Release

Four TCP segments are needed to release a connection: one FIN and one ACK for each direction.

The Two-army Problem: 需要知道信号是否送达.

The following steps are followed in terminating the connection:

1) For terminating the connection
2) after receiving the FIN segment
3) After receiving the acknowledgement, client enters the state called FIN_WAIT_2. Now
4) Now suppose server wants to close the connection with the client. For terminating the connection
5) After receiving the FIN segment

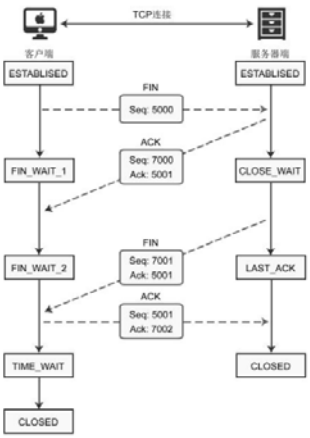

Figure VI.13: TCP Connection Release

State	Description
CLOSED	No connection is active or pending
LISTEN	The server is waiting for an incoming call
SYN RCVD	A connection request has arrived; wait for ACK
SYN SENT	The application has started to open a connection
ESTABLISHED	The normal data transfer state
FIN WAIT 1	The application has said it is finished
FIN WAIT 2	The other side has agreed to release
TIME WAIT	Wait for all packets to die off
CLOSING	Both sides have tried to close simultaneously
CLOSE WAIT	The other side has initiated a release
LAST ACK	Wait for all packets to die off

Figure VI.14: The states used in the TCP connection management finite state machine

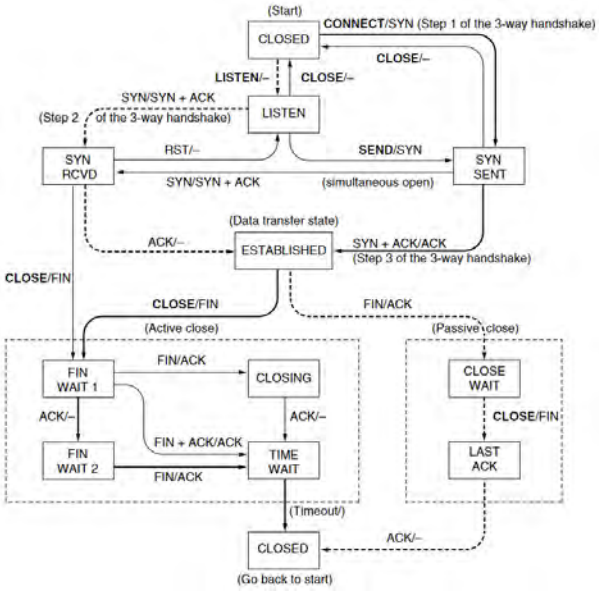

Figure VI.15: TCP connection management finite state machine

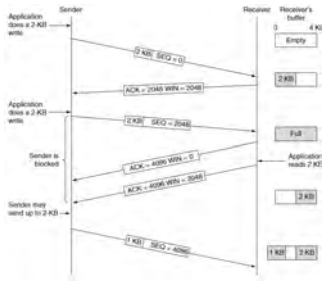

Figure VI.16: Window management in TCP

3.5 TCP Sliding Window

Example syn + ack 需要占用一个 sequence number, 纯 ack 不用占用. 每个 byte 都有一个对应的 sequence number, 这个 sequence number 也可以携带 ack.

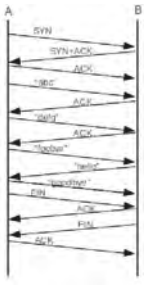

Figure VI.17: An Ladder Example

	A sends	B sends
1	SYN: **SEQ = 0**	
2		SYN+ACK: SEQ = 0, **ACK = 1** (expecting)
3	ACK: **SEQ = 1**, ACK = 1 (ACK of SYN)	
4	"abc": **SEQ = 1**, ACK = 1	
5		ACK of "abc": SEQ = 1, **ACK = 4** (no data)
6	"dfeg": **SEQ = 4**, ACK = 1	
7		ACK of "dfeg": SEQ = 1, **ACK = 8** (no data)
8	"foobar": **SEQ = 8**, ACK = 1	
9		"hello": SEQ = 1, **ACK = 14**
10	"goodbye": **SEQ = 14**, ACK = 6	
11	FIN: **SEQ = 21**, ACK = 6	ACK of "goodbye": SEQ = 6, **ACK = 21**
12		ACK of FIN: SEQ = 6, **ACK = 22**
13		FIN: SEQ = 6, **ACK = 22**
14	ACK of FIN: **SEQ = 22**, ACK = 7	

Figure VI.18: An Ladder Example

When the window is 0, the sender may not normally send segments, with two exceptions:

1) Urgent data may be sent, for example, to allow the user to kill the process running on the remote machine
2) The sender may send a 1-byte segment to force the receiver to re-announce the next byte expected and the window size. This packet is called a window probe

Nagle's Algorithm To reduce the bandwidth used by a sender that sends multiple short packets. 当有数据要发送, 先发送一个分片, 缓存其他分片, 受到 ack 后一次性发送缓存的分片. 但因为缓存, 高交互系统可能造成死锁.

Clark's Solution the silly window syndrome: an interactive application on the receiving side reads data only 1 byte at a time. 只有 receiver 有一定比例的 window 可用后再通知 sender.

The Silly Window Syndrome The goal is for the sender not to send small segments and the receiver not to ask for them.

Another issue that the receiver must handle is that segments may arrive out of order. Solution:

- The receiver will buffer the data until it can be passed up to the application in order
- Acknowledgements can be sent only when all the data up to byte acknowledged have been received. This is called a cumulative acknowledgement

3.6 TCP timer management

TCP uses multiple timers (at least conceptually) to do its work.

- The RTO (Retransmission TimeOut)
  How long should the RTO be ?
- The Persistence timer
- The Keepalive timer
- The one used in TIME WAIT state

The RTO (Retransmission TimeOut)

- set too short, unnecessary retransmissions will occur.
- set too long, performance will suffer due to the long retransmission delay whenever a packet is lost

Furthermore, the ack arrival can change rapidly within a few seconds as congestion builds up or is resolved

The solution is to use a dynamic algorithm:

- SRTT (Smoothed Round-Trip Time, Jacobson,1988)
  Exponentially Weighted Moving Average (EWMA, R is the current estimate of the RTT)

$$SRTT = \alpha SRTT + (1 - \alpha)R$$

where $\alpha = 7/8$.

- RTTVAR (Round-Trip Time VARiration)
  To make the timeout value sensitive to the variance in round-trip times as well as the smoothed round-trip time.

$$RTTVAR = \beta RTTVAR + (1 - \beta)|SRTT - R|$$
$$RTO = SRTT + 4 \times RTTVAR$$
$$RTO = \min(1sec, RTO)$$

where $\beta = 3/4$

One problem that occurs with gathering the samples, R, of the round-trip time is what to do when a segment times out and is sent again. 但当接受 ack 时无法确定是从发的 ack 还是先前的 ack.

Karn's algorithm: 重传的 rtt 不参与更新 rto. 每次失败的重传 timeout doubled.

the persistent timer It is designed to prevent the following deadlock

the keepalive timer 看看对方是否下线.

the TIME WAIT timer twice the maximum packet lifetime

3.7 TCP Congestion Control

The network layer 满了只能 dropping packets(丢包). 所以靠 transport layer 解决. In the Internet, TCP plays the main role in controlling congestion, as well as the main role in reliable transport.

TCP congestion control is based on a AIMD (Additive Increase Multiplicative Decrease) control law using a window and with packet loss(丢包) as the binary signal.

TCP maintains a congestion window(the sending window) and a flow control window(the receiving window):

- The congestion window size 是指每时 sender 最大发送的字节数.
  The corresponding rate = window size/RTT
  TCP 用 AIMD 调整 congestion window size
- The flow control window size 是指每时最大容纳的字节数. Receiver 通过 TCP Header 告知 sender 其 window size.
  发送字节数应该同时小于两 window size
  任意 window size full, TCP 停止传输数据.

All the Internet TCP algorithms assume that lost packets(丢包) are caused by congestion and monitor timeouts.

During the implementation, we have two important questions:

1) The transmission rate of packets which the sender will use ack clock
2) The size of the congestion window
   slow start

The key observation is that ack 到 sender 的 rate 受 slowest link 制约. The acknowledgements reflect the times at which the packets arrived at the receiver after crossing the slow link. This timing is known as an ack clock. It is an essential part of TCP

TCP Start Problem We want to quickly near the right rate, $cwnd_{IDEAL}$.

Slow-Start Solution: Start by doubling cwnd (the congestion window) every RTT. 每收到一个 ACK, 就增加一个数据包, 也就是一个变两个.

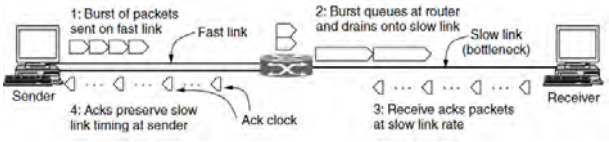

Figure VI.19: A burst of packets from a sender and the returning ack clock

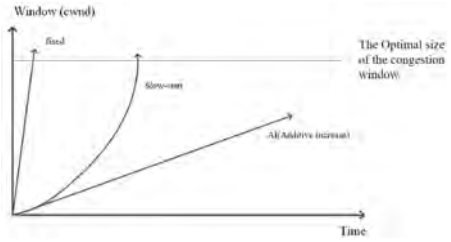

Figure VI.20: Slow-Start Solution

To keep slow start under control, the sender keeps a threshold for the connection called the slow start threshold (sst).

- 起始设置为 flow control window size.
- TCP keeps increasing the congestion window in slow start until

  1) a timeout occurs (packet loss): sst 降到 congestion window 的一半, 且重启 slow start.

  2) the congestion window exceeds the slow start threshold: 从 slow start 变为 additive increase (线性)

additive increase: 完成一个完整的 RTT, 才增加一个数据包.

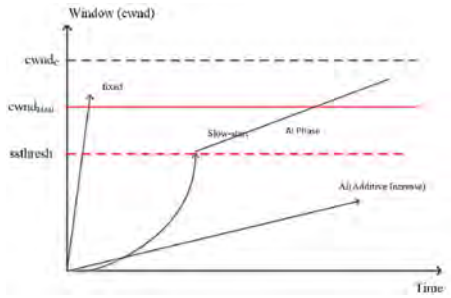

Figure VI.21: A mix of linear and the multiplication increase

Inferring Loss from ACKs

- TCP uses a cumulative ACK
- Duplicate ACKnowledgements gives us hints about what data hasn't arrived
  - 告知收到数据包, 但不是期望的
  - 将三次相同的 ack 当作 loss
  - 即使计时器没有 timeout, 但也做重传 (Fast Retransmission)
  - sst 设置为 congestion window/2, slow start 重启

Fast Retransmission It can repair single segment loss quickly, typically before a timeout

Fast Recovery Fast recovery is the heuristic that implements this behavior

- pretend further duplicate ACKs are the expected ACKs
- duplicate ACKs are counted (including the three that triggered fast retransmission) until the number of packets in the network has fallen to the new threshold
- Reconcile views when the ACK jumps

Two larger changes have also affect TCP implementations

1) SACK (Selective ACKnowledgements)

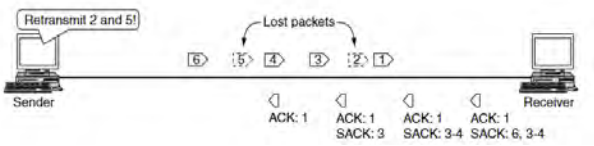

Figure VI.22: Selective acknowledgements

2) ECN (Explicit Congestion Notification): an IP layer mechanism

- ECE (ECN-Echo): tell the TCP sender to slow down
- CWR: The sender tells the receiver that it has heard the signal

TCP Reno, NewReno, and SACK. (笔记中没有前两个版本)

- Reno can repair one loss per RTT
  Multiple losses cause a timeout
- NewReno further refines ACK heuristics
  Repairs multiple losses without timeout
  SACK (Selective Acknowledgement) is a better idea

4. Introduce new transport layer protocols

4.1 QUIC

全称 (Quick UDP Internet Connection), 中文翻译成 "快速 UDP 互联网连接", 是由 Google 提出的使用 UDP 进行多路并发传输的协议.

1) Low latency to establish connection

2) Improved Congestion Control Scheme

3) reliable transmission based on monotonically increased packed number.

Stream offset

4) removal of the “Head-of-Line blocking” (HOL blocking) problem (队头阻塞问题)

5) 连接迁移

4.2 BBR

(Congestion Control based on Bottleneck Bandwidth)
现在丢包与拥塞不再等价了.

A full-duplex TCP connection has exactly one slowest link or bottleneck in each direction. The bottleneck is important because It determines the connection's maximum data-delivery rate.

Two physical constraints, RTprop (round-trip propagation time) and BtlBw (Bottleneck Bandwidth), bounds transport performance. If the network path were a physical pipe, Rtprop would be its length and BtlBW its minimum diameter.

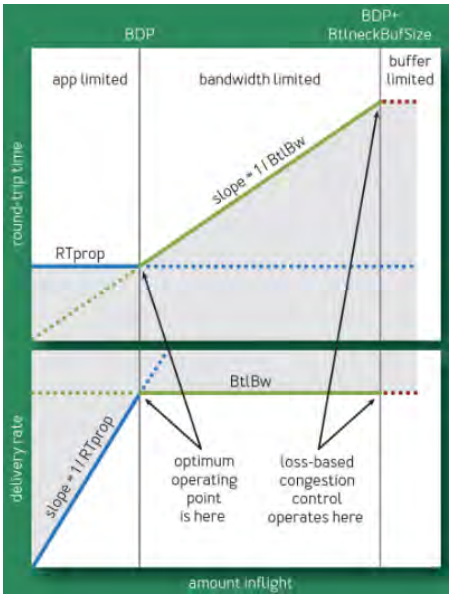

Figure VI.23: RTT and delivery rate variration with the amount of data inflight

data in flight (data sent but not yet acknowledged) = BtlBw × Rtprop

Bandwidth-delay product (BDP) is a measurement of how many bits can fill up a network link

Little's Result:

$$\bar{N} = \lambda T$$

- $\bar{N}$ is the data in flight in the network
- T RTT
- λ the delivery rate

$$T = \frac{\bar{N}}{\lambda} = \frac{\bar{N}}{BtlBw} = RTT$$
$$\lambda = \frac{\bar{N}}{T} = \frac{\bar{N}}{RTprop}$$

Rtprop and BtlBw obey an uncertainty principle: whenever one can measured, the other cannot.

Characterizing the bottleneck:

- Rate balance
- Full pipe
- BtlBw and RTprop vary over the life of a connection, so they must be continuously estimated

VII Application Layer

1. Overview of application layer

Builds distributed "network services" (DNS, Web) on transport services

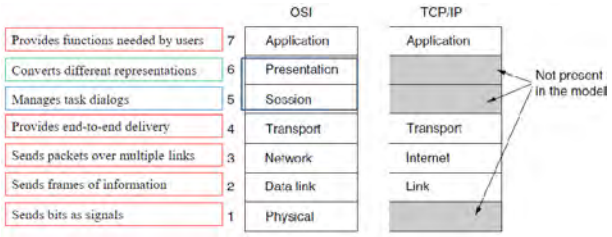

Figure VII.1: Presentation Layers

1.1 Applications and Application Level Protocols

The three concepts:

- Service model
- Protocol
- Interface

Network application contains Client site, Server site, Application level protocol

1.2 Parameters

- Data Loss
- Bandwidth
- Timing

2. Important application layer protocols

2.1 DNS (Domain Name System)

decouple machines names from machine addresses.

The essence of DNS is the invention of a hierarchical, domain-based naming scheme and a distributed database system for implementing converting the machine names to network addresses. an application program calls a library procedure called the resolver.

The query and response messages are sent as UDP packets. (现在不一定了, 不过书上是这样写的, 考试要这样答)

Why Not Centric:

- Single point of failure
- Traffic volume
- Distant name server means slow response
- Scalability

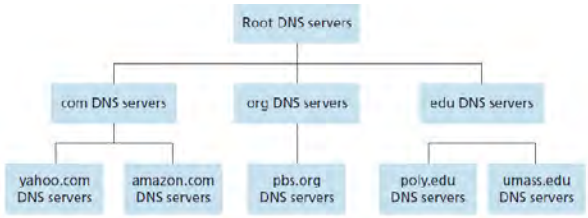

Figure VII.2: Portion of the hierarchy of DNS servers

Root Nameservers Root (dot) is served by 13 server names

- a.root-servers.net to m.root-servers.net
- Each "server" is actually a cluster of replicated servers, for both security and reliability purposes

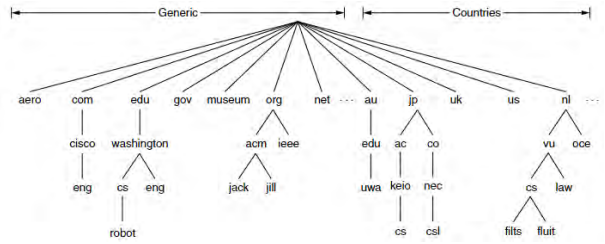

Figure VII.3: A portion of the Internet domain name space.

TLDs (Top-Level Domains) Run by ICANN.

DNS Zones Each zone has one or more nameservers to contact for information about it.

Domain Resource Records The primary function of DNS is to map domain names onto resource records.

A resource record is a five-tuple. The format is as follows:

- Domain name: This field is thus the primary search key used to satisfy queries
- Time to live: how stable the record is.
- Class: For Internet information, it is always IN
- Type

Type	Meaning	Value
SOA	Start of authority	Parameters for this zone
A	IPv4 address of a host	32-Bit integer
AAAA	IPv6 address of a host	128-Bit integer
MX	Mail exchange	Priority, domain willing to accept email
NS	Name server	Name of a server for this domain
CNAME	Canonical name	Domain name
PTR	Pointer	Alias for an IP address
SPF	Sender policy framework	Text encoding of mail sending policy
SRV	Service	Host that provides it
TXT	Text	Descriptive ASCII text

Figure VII.4: The principal DNS resource record types.

– An SOA record provides the name of the primary source of information about the name server's zone.
– A: 32-bit IPv4 address
– AAAA: 128-bit IPv6 address
– MX: it specifies the name of the host prepared to accept email for the specific domain
– NS: a name server
– CNAME: ppt 参考文献 [7]
– PTR: allow lookups of the IP address and return the name of the corresponding machine. – reverse lookups
– SRV: a newer type of record that allows a host to be identified for a given service in a domain.
– SPF: This helps receiving machines check that mail is valid.
– TXT:

- Value

DNS Resolution DNS protocol lets a host resolve any host name (domain) to IP address.

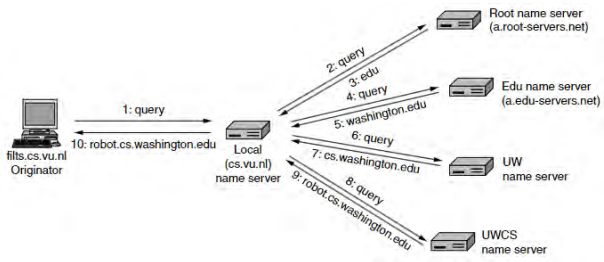

Figure VII.5: Example of a resolver looking up a remote name in 10 steps.

Iterative vs. Recursive Queries

- Recursive query: Nameserver completes resolution and returns the final answer (客户端 – 本地 DNS 服务器)
- Iterative query: Nameserver returns the answer or who to contact next for the answer (本地 DNS 服务器 – 外网)

DNS Caching vs. Freshness Information is cached between 5 minutes and 72 hours

Local Nameservers Local nameservers typically run by IT

Name Servers There are 13 root DNS servers. Most of the root servers are present in multiple geographical locations and reached using anycast routing, in which a packet is delivered to the nearest instance of a destination address.

DNS Protocol Query and response messages are Built on UDP messages, port 53. Service reliability via replicas. Security is a major issue.

Security of DNS Initially, the transaction ID was only 16 bits. This design choice left DNS vulnerable to a variety attacks including "cache poisoning attack". (伪造 DNS)

- Oblivious DNS
- DNSsec

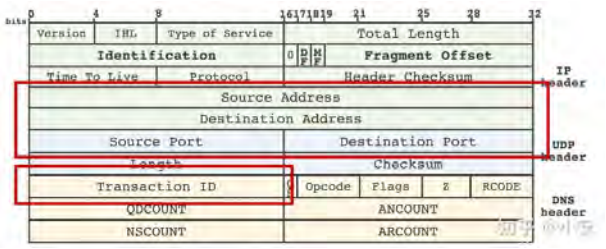

Figure VII.6: DNS Header

2.2 *FTP(File Transfer)*

FTP: to transfer files to and from a remote host.
Steps:

1) The user first provides the hostname of the remote host, causing the FTP client process in the local host to establish a TCP connection with the FTP server process in the remote host.
2) The user then provides the user identification and password, which are sent over the TCP connection as part of FTP commands.
3) Once the server has authorized the user, the user copies one or more files stored in the local file system into the remote file system (or vice versa).

FTP uses two parallel TCP connections to transfer a file, a control connection and a data connection.

2.3 *Email(Electronic Mail)*

Email is an asynchronous communication medium
Three major components:

1) User agents
2) Message transfer agents (mail servers)
3) Simple mail transfer protocol: SMTP

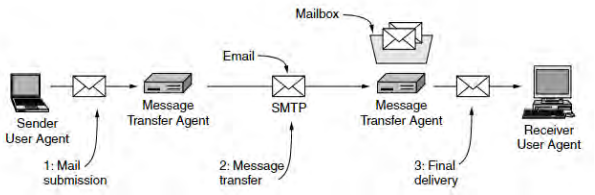

Figure VII.7: Architecture of the email system.

SMTP is the principal application layer protocol for Internet electronic mail

MIME The Multimedia Internet Mail Extension

Mail Access Protocols SMTP has been designed for pushing email form one host to another. obtaining the messages is a pull operation, whereas SMTP is a push protocol

Final Delivery:

- POP
- IMAP
- HTTP

2.4 HTTP(The HyperText Transfer Protocol)

The HTTP, the Web's application-layer protocol, is at the heart of the Web.

HTTP is a simple request-response protocol that normally runs over TCP. HTTP has nothing to do with how a Web page is interpreted by a client.

Web page as a set of related Web resources The content from these different servers is integrated for display by the browser.

HTTP is a request/response protocol for fetching Web resources

Fetching a Web page with HTTP the Client Side

Steps:

1) Resolve the server to IP address (DNS)
2) Set up TCP connection to the server
3) Send HTTP request for the page
4) (Await HTTP response for the page)
5) Execute / fetch embedded resources /render (不只是展示网页中内容，可能还要运行程序等。)
6) Clean up an idle TCP connections

The Server Side

Steps:

1) Accept a TCP connection from a client (a browser).
2) Get the path to the page, which is the name of the file requested.
3) Get the file (from disk).
4) Send the contents of the file to the client.
5) Release the TCP connection.

Cookies It remembers stateful information for the stateless HTTP protocol

A cookie may contain up to five fields:

1) The Domain tells where the cookie came from
2) The Path is a path in the server's directory structure that identifies which parts of the server's file tree may use the cookie.
3) The Content field is where the cookie's content is stored.
4) The Expires field specifies when the cookie expires
5) The Secure field can be sent to indicate that the browser may only return the cookie to a server using a secure transport. This feature is used for e-commerce.

HTTP Message Format

Request

Request Headers

Example

Response

Example

html 需要记 html 一些语法.

css

Dynamic Web Pages 也会考

VIII Network Security

Basic knowledge

1. Cryptography

1.1 Kerckhoff's principle

All algorithms must be public; only the keys are secret.

1.2 Substitution ciphers

In a substitution cipher, each letter or group of letters is replaced by another letter or group of letters to disguise it.

The basic attack takes advantage of the statistical properties of natural languages.

1.3 Transposition ciphers

Transposition ciphers, in contrast, reorder the letters but do not disguise them. The cipher is keyed by a word or phrase not containing any repeated letters.

To break a transposition cipher, the cryptanalyst must

1) first be aware that he is dealing with a transposition cipher.
2) The next step is to make the guess at the number of columns (the key length).
3) The remaining step is to order the columns.

1.4 One-time pads

First choose a random bit string as the key. Then convert the plaintext into a bit string, for example, by using its ASCII representation. Finally, compute the XOR of these two strings, bit by bit.

The reason derives from information theory: there is simply no information in the message because all possible plaintexts of the given length are equally likely.

1.5 Two fundamental cryptographic principles

1) All the encrypted messages must contain some redundancy.
2) Measures must be taken to ensure that each message received can be verified as being fresh.

2. Symmetric-key algorithms

In symmetric-key algorithms, they use the same key for encryption and decryption.

2.1 DES

要经历 19 个阶段.

2.2 Triple DES

whitening, 不确定性达到最大值.

2.3 Rijndael

很重要, 但不讲了.jpg

3. Public-key algorithms

These requirements can be stated simply as follows:

1) $D(E(P)) = P$
2) It is exceedingly difficult to deduce D from E.
3) E cannot be broken by a chosen plaintext attack.

Public-key cryptography requires each user to have two keys: a public key, used by the entire world for encrypting message to be sent to that user, and a private key, which the user needs for decrypting messages.

3.1 RSA

4. Digital Signatures

The authenticity of many legal, financial, and other documents is determined by the presence or absence of an authorized handwritten signature.

The basic requirements of digital signatures:

1) The receiver can verify the claimed identity of the sender.
2) The sender cannot later repudiate the contents of the message.
3) The receiver cannot possibly have concocted the message himself.

4.1 Symmetric-Key Signatures

Replay Attack

4.2 Public-Key Signatures

4.3 Message Digests

One criticism of signature methods is that they often couple two distinct functions: authentication and secrecy.

SHA-1 and SHA-2

4.4 The Birthday Attack

The birthday attack is that with two different plaintexts, but have the same message digests.

5. Management of public keys

5.1 Certificates

5.2 Public Key Infrastructures

IX 复习