计算机网络笔记

陈鸿峥

2019.04*

目录

本课程使用的教材为James F. Kurose和Keith W. Ross的《计算机网络—自顶向下方法（第七版）》。

1 计算机网络概述

计算机网络将终端设备连接起来并可以传输数据。

*Build 20190417

1.1 网络连接方式

1.1.1 直接连接的网络

- 点对点(point-to-point)网络：包括专用介质(dedicated medium)、节点/主机
  - 单向(simplex)：如广播、电视
  - 半双工(half duplex)：异步双向，如对讲机
  - 全双工(full duplex)：同步双向，如电话
- 多路访问(multiple access)网络：共享介质(shared medium)、广播、碰撞(collision)
  - 单播(unicast)：一对一
  - 多播(multicast)：一对多
  - 广播(broadcast)：一对所有

1.1.2 间接连接的网络

- 中间节点、路由器(router)
- 包(packet)
- 存储转发(store and forward)
- 路由选择(routing)
- 路由表(routing table)
- 目的地(destination)、下一跳(next hop)

1.2 因特网

网络互连：用路由器（或网关gateway）连接起来构成的网络称为互**连**网络(internetwork)。用实际的物理通信介质及相应的设备把两个或两个以上的网络连接起来的一种网络，如LAN和WAN都可看作是互连网络。

因特网/互联网(Internet)是一种互连网络，可以看作是把世界各地的广域网互连的网络，是世界上最大的特定计算机网络，采用TCP/IP协议簇作为通信规则

- 系统域网(System Area Network, SAN)：电脑、鼠标、USB
- 局域网(Local Area Network, LAN)：某一区域内由多台计算机互联成的计算机组，一般是方圆几千米以内，如小型实验室
- 城域网(Metropolitan Area Network, MAN)
- 广域网(Wide Area Network, WAN)

因特网：

- 终端系统/主机(end system)：运行网络应用程序
- 通信链路(communication link)：光纤、铜线、无线电、卫星等

- 路由器(router)

因特网的组成：

- 网络边界(Network edge)：主机
- 接入网络(Access network)：有线或无线接入，连接订阅者和服务提供商
- 网络核心(Core network)：连接局部提供商

因特网的结构：ISP(Internet Service Provider)

- 顶层ISP：主干网（中国电信、中国移动、中国网通）
- 区域ISP：可以私自互联
- 本地ISP

1.3 网络服务

通信服务类型：

- 可靠/不可靠：会不会丢包/收发是否完全相同，如文件（可靠）/视频（不可靠）
- 面向连接/无连接：需不需要建立通信线路，如电话（连接，双方都要在）/寄信、因特网（无连接，对方可能不在）
- 有确认/无确认：需不需要确认对方是否收包，因特网不需要
- 请求响应/消息流服务：有请求才有响应/一直发消息，如电视

因特网是数据报服务，无连接无确认

1.4 因特网体系结构

因特网体系结构包括以下这五层，而ISO/OSI(open system interconnection)网络包括七层协议：

- 应用层：提供对某些专门应用的支持，如FTP、SMTP、HTTP

OSI 表示层(presentationn)：提供数据转换服务, 例如，加密解密，压缩解压缩，数据格式变换

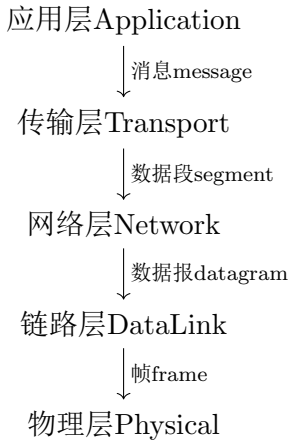
OSI 会话层(session)：简化会话实现机制，例如，数据流的检查点设置和回滚以及多数据流同步

- 传输层：将网络层获得的包在**进程之间**数据传送（端到端），如TCP、UDP
- 网络层：**路由选择**，实现在互联网中的数据传送（主机到主机），IP、路由协议
- 数据链路层：在物理网络中传送**包**（跳到跳，节点到节点），PPP、Ethernet
- 物理层：线上的**比特**（传送原始比特流）

其中网络层以下不可靠，以上可靠；防止丢包的机制：重发

协议(protocol)：在网络实体(entities)之间传送消息的规则，如消息的格式、收发消息的次序等

协议栈：发送时封装(encaptulation)，接收时拆封。每层传输的数据单元都称为包(packets)，都属于某个协议，又被称为协议数据单元(protocol data unit, PDU)=协议控制信息(potocal control data, PCI)+服务数据单元(SDU)



不同协议则添加不同头部。路由做的事情是拆一层封装，然后重新加一层。同一个互联网络中网络层协议需要相同，链路层协议可以不同。对等实体(peer entity)即实现相同协议的实体。

协议簇(protocol family)：应用层FTP、HTTP、DNS，传输层TCP/UDP，网络层IP

1.5 网络性能分析

当一个包到达时如果有空闲缓存则排队等待转发，产生延迟(delay)；如果没有空闲缓存，则丢弃该包，造成丢失(loss)。包交换网络中的延迟

- 处理(processing)延迟：查路由，存储转发(store-and-forward)技术则延迟很大
- 排队(queueing)延迟：依赖于路由器的拥塞程度
- 发送/传输(transmission)延迟：包长(bits)/链路带宽(bps, bit per second)；指从发送第一个包到发送最后一个包的间隔
- 传播(propagation)延迟：指对于一个包来说从发送到接收所需的时间，物理链路长度/信号传播速度

接收延迟与传播延迟重合。故

$$\text{总延迟（从第一个包被发送到最后一个包被接收的时间）} = \text{传播延迟} + \text{发送延迟}$$

往返时间(round trip time, RTT)：从源主机到目的主机再返回源主机所花的时间

- 带宽(bandwidth)：一条链路或通道可达到的最大数据传输速率(bps)
- 吞吐量(thoughput)：一条链路或通路实际数据传输速率

2 物理层

直连网，不管包。需要做的事情：（调制解调为模拟信号，编码解码为数字信号）

$$\text{信息源} \rightarrow \text{调制/编码} \rightarrow \text{信道传输} \rightarrow \text{解调/解码} \rightarrow \text{目的地}$$

信息能够被解释为数据(msg/data)，用符号(sign)记录，用信号(signal)（光、电）传递(transmit)，用熵(entropy)测量

- 模拟信号-传输：连续取值，放大器(amplifier)
- 数字信号/跳变信号-传输：离散取值，中继器(repeater)

2.1 编码方式

2.1.1 模拟信号

载波信号(carrier)一般采用正弦波信号：角频率ω、频率f、周期T、振幅A、相位(φ)

- 频移键控(frequency-shift keying, FSK)：通过不同频率表示不同信息0/1
- 幅移键控(amplitude-shift keying, ASK)
- 相移键控(phase-shift keying, PSK)
- 正交调幅(quadrature amplitude modulation, QAM)：用不同的振幅/相位表示不同的多位信息$000 \sim 111$

2.1.2 数字信号

1. 单极编码(unipolar)：0V即0，$+E$V为1，但是会产生
   - 时钟漂移：不同的时钟会有差别，一定要有跳变
   - 基线漂移：线很长会有（积累很多电荷，以为是1），一定要有变化/正负
2. 不归零编码/双极编码(non-return-to-zero/bipolar, NRZ)：$-E$为0，$+E$为1，解决基线漂移问题（平衡01）；全是0或全是1，还是没法区分
3. 不归零反转编码(Inverted, NRZI)：**差分**码波形，相邻码元的电位改变表示1，而电位不改变表示0；也可以反过来。该表示方法与码元本身电位或极性无关，而仅与相邻码元的电位变化有关
4. 曼彻斯特(Manxhester)编码：从相邻时刻的中间起$M-E \sim +E$，$0 \to 10, 1 \to 01$，可克服时钟漂移和基线漂移；频率高，传输有问题，对传输介质要求高
5. 差分曼彻斯特编码：在每一位开始时间如果跳变则为0，否则为1
6. 4B/5B编码：用5比特代表4比特，多一位冗余；每个编码没有多于1个前导零和多于2个末端零

2.2 物理介质

2.2.1 分类

有线介质

- 双绞线：
  - 非屏蔽双绞线(unshielded twisted pair, UTP)：四对线（绿绿白、橙橙白、蓝蓝白、棕棕白），cat6千兆以太网

  - 屏蔽双绞线(STP)

- 同轴电缆(coaxial cable)
- 光导纤维(optical fiber)
  - 单模光纤(single mode)：最大传输速率
  - 多模光纤：阶跃(step-index)光纤、渐变(graded-index)光纤

无线介质：地面微波、WiFi、3G网络、卫星

2.2.2 多路复用

- 时分多路复用(time division multiplexing, TDM)
- 频分多路复用(frequency, FDM)：无线电台常用
- 波分多路复用(wavelength, WDM)：利用多个激光器在单条光纤上同时发送多束不同波长激光的技术
- 码分多路复用(code, CDM)
- 统计多路复用(static, SDM)：动态分配方法共享通信链路，比如FIFO；对于多个可变速率的数据流，SDM可以提高链路利用率

3 数据链路层

数据链路层把数据包，即**帧(frame)**，从一个节点通过链路（直连网络或**物理网络**）传给相邻另一个节点（主机和路由器）

数据链路层的功能如下：

- 成帧(framing)
- 差错检测(error detect)：比特错，纠错
- 差错控制(error control)：丢包、重复、错序、流控制(flow control)
- 介质访问控制(medium access control)：多路访问，碰撞(collision)

针对点对点和多路访问网络分别制定了两个子层：

- 逻辑链路控制(Logic Link Control, LLC)子层：提供可靠数据传输
  - LLC1提供无确认无连接服务
  - LLC2提供有确认面向连接的服务，实现滑动窗口协议
  - LLC3提供有确认无连接的服务
- 介质访问控制(Media Access Control, MAC)子层：专门用来处理多路访问网络中的冲突（点对点网络没有冲突就不用）

注意数据链路层、网络层错了就错了，不提供纠正服务，由上层纠正

3.1 逻辑链路控制子层

3.1.1 差错检测

在数据报后加校验码（头部加序号），通过链路传输看是否有数据报/校验码错误

1. 奇偶校验：若接收方收到奇数个1，则有出错
   - 一维偶校验：只能检错；最后补一位使得全部为偶数个1，如010补为010—1，而101补为101—0
   - 二维偶校验：检错+纠错一位；横纵同时偶校验
2. 校验和(checksum)：将所有数据加起来
   由于需要使用加法器，校验和一般不用于数据链路层，而是更高层，例如IP层和传输层
3. 循环冗余校验码(Cyclic Redundancy Check, CRC)：补充n位后除以一个n+1位的除数，模2除法（按位异或，做减法时没有借位）
   接收方连带校验码一起除，余数为0则没错
   链路层常用CRC，因为检错率很高，且容易实现（触发器+异或门）

3.1.2 可靠数据传输

超时则自动重发请求(Automatic Repeat reQuest, ARQ)：每发送一帧都启动一个超时定时器，如果它的确认帧(ACK)在其超时时间内到达就删除该定时器，否则，重传该帧并重启定时器

主要的协议如下：

- 停等协议(stop-and-wait)：ARQ协议，只有收到前一个数据帧的确认帧才可以发送下一个数据帧
  三种出错情况：
  - 数据帧丢失
  - 确认帧丢失
  - 超时：收到ACK表明接收方一定收到，可以发送新的数据帧，重传的也一定要发ACK

  效率/吞吐量十分低，信道空闲时间长
- 滑动窗口协议(sliding window)：ARQ协议，不需等待前面发送的帧的确认帧返回，就可以连续发送下一个，其个数不能超过发送窗口大小(sending window size, SWS)（连续发送数据帧可用序号范围，用于流控制：控制发送速度，否则会发生溢出(overlow)，后面覆盖前面的）
  这里的确认帧是指在此之前的帧都已收到（直连网中间没有节点，后面收到前面一定收到；只要出错纠正不了直接丢弃）
  - 回退N协议(go back N)：同滑动窗口连续发送，但某个ACK没收到则重传在此ACK之后的所有帧（超时重传），丢3则ACK4发2
    发送窗口需要缓存SWS个帧，以便重传；发送窗口中序号最小的为sendbase
  - 选择性重传(selective repeat)：通过发送否定性确认帧(negative acknowledgement, NAK)要求重传该帧；如3丢失，4发送NAK=3，5发送ACK=2，重传3
    接收窗口(receiving window size, RWS)表示接收缓冲区大小(RWS≤SWS，最好是等于，尽量

减少重传帧；但序号少的话导致重复)，用于确定应该保存哪些帧，用序号范围表示
超时时间应该设长，确保帧的2次来回；没有后续帧也会超时重传；无论窗口内窗口外收到都要发确认

注意不管哪一个重传机制，序号可以重复使用，因此有最小序号问题。

例 1. *序号8个，SWS=RWS=4，345670123456，5丢失*

分析. *回退N：346705670123456*
选择性重传：346705123456

提高滑动窗口协议的效率：

- 选择性确认(selective acknowledgement)：接受方把已收到的帧的序号告诉发送方（收到不用重传告诉发送方哪些已收到，则只用重传某帧）
- 捎带确认(piggybacking)：通信双方**全双工**方式工作，接收方在发数据给对方时顺便把确认号也告诉对方（两个滑动窗口，两边都要发数据），需要结合下面一起使用
- 延迟确认(delayed acknowledgement)：接收方收到一帧后并不立即发送确认帧，而是等待一段时间再发送

标志位成帧，区分包的起始和终结；同时需要转义

链路层的实现：在网络接口卡(network interface card, NIC)及其驱动程序上实现，路由器在接口模块上实现

3.2 介质访问控制子层

3.2.1 简介

1. **PPP协议(point-to-point)：点到点网络**
   - 根据HDLC(high-level data link control)协议进行设计，主要用于串行电缆、电话线(MODEM)等串行链路
   - 提供连接认证、传输加密和压缩功能，为网络层协议提供服务
   - 没有纠错功能，也没有流控制和确保有序的功能
2. **以太网：多路访问网络**
   - 纯ALOHA：想发送就发送，超时未收到确认则发生冲突
   - 分槽ALOHA：将时间分为长度相同的时槽，每个站点只在时槽开始时发送。
     信道空，立即以概率p发送，以概率$1-p$延迟一个时间槽；信道忙，延迟一个时间槽。
   - 载波监听CSMA(Carrier Sense Multiple Access)：发送前先监听信道
     - 信道空，立即发送；信道忙，持续监听(1-persistent CSMA，以太网)
     - 信道空，发送；信道忙，延迟一段随机长度时间(non-persistent CSMA)
     - 信道空，立即以概率p发送，以概率$1-p$延迟一个时间槽；信道忙，延迟一个时间槽(p-persistent CSMA，分槽ALOHA)

3.2.2 以太网MAC层协议

载波监听CSMA/CD(with collision detection)

1. 发送数据帧之前先监听信道。如果信道空闲，立即发送。如果信道忙，则持续监听，直到信道空闲，立即发送。
2. 边发送边检测冲突。如果发送完毕都没有检测到冲突，则发送成功。
3. 如果检测到冲突，则停止发送，并发送32位干扰位(jamming signal)以加强冲突信号。采用二进制指数退避算法**随机延迟**一段时间后，转(1)。

二进制指数退避算法(binary exponential backoff)

- 时间片τ的长度为512b时间，10Mbps的以太网为51.2μs
- 帧间空隙(interframe gap)为96b
- 每次从2^j-1个时间片随机选择一个

802.3的MAC帧格式

前导字符(8B)+目的地址(6B)+源地址(6B)+类型/长度(2B)+有效载荷/填充位+帧校验序列(4B)

- 源地址：一般为发送者的单播地址
- 目的地址(6B)：一般为接收者的单播地址
- 单播/网卡/烧录地址：全球唯一，每个网卡/接口一个
- 多播地址：字节0第0位为1，地址非全1
- 广播地址：48位全为1

3.2.3 透明网桥

网桥/交换机：二层（从下往上数）交换机数据链路、物理层
校园网大多交换机，路由少
扩展/桥接局域网：每一个局域网(LAN)都是一个网段
MAC地址表，如

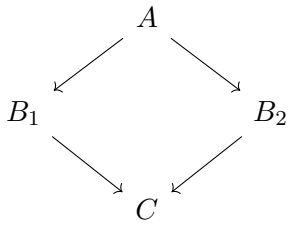
B	P1
D	P2

- 表里查到则转发(forward)
- 表里没有则扩散/泛洪(flood)：由端口P1，扩散到其他所有端口P2，P3（多播、广播一定扩散）扩散不回传收到的部分不会往回转
- 从某一条路发来则不能传回去，过滤/丢弃(filter)

生存期(Time to live, TTL)：单位为秒，每次发送都会重置，对于不活跃的表项自动删掉（减少表的大小，查找速度更快）

自学习：利用**源地址**学习，如信息从A-P1来，则记为A-P1，同时设好TTL 如果收到的帧有错则直接丢弃，根本不会学习更新记录，重置超时计时器

广播风暴：回路，容错，防止一个网桥崩溃就全局崩溃



冲突域：中间可以加集线器

3.3 生成树协议

IEEE 802.1w RSTP(Rapid Spanning Tree Protocol)

将所有的LAN和网桥都抽象为结点，避免冲突即构造一棵生成树（注意不是最小生成树）

先确定根网桥，即BID(Bridge ID)最小的

每个网段（需要集线器）依赖于连通的网桥，每个网桥都把自己到根的距离发出去（竞选/配置消息）

网桥之间的开销为1，选一条最短路径

网桥只在根端口和指定端口之间转发**数据**帧

主机端口出去不加帧

扩散自己BID，最后只剩下根网桥认为自己是根取得优胜的，作为制定网桥；相同距离时，BID小的优胜；端口号小的优胜

只有从根端口过来的才扩散配置消息其他端口来的不扩散这样不会形成回路

只能在根端口和指定端口转发，不可通过阻塞端口

断了/失效了则变成无穷大，其他网桥可成为指定网桥

防止广播风暴，又能自动修复损害网桥（通过冗余方式），增加可靠性

3.4 虚拟局域网

虚拟局域网(Virtual LAN, VLAN)将原来的局域网分割成多个相互隔离的局域网，只在具有相同颜色的端口间转发。

扩散到帧内制定的端口或**干道端口**

查MAC地址表转发

如果从干道收到的帧没有VLAN ID，则认为是本征(native)VLAN，默认为VLAN1。

多生成树协议：管理员规定哪些VLAN为一组，构成多生成树，其余的用公公生成树

- 公共生成树(common spanning tree, CST)
- 多生成树(Multiple spanning tree protocol, MSTP)

3.5 交换机

交换机是一个把多个网段连接起来的设备，也称为**多端口网桥**(switch=bridge)。注意输入输出端口一样。

交换结构(fabrics)

- 共享总线式交换机：同样有冲突问题
- 纵横式(crossbar)

交换机转发方法

- 存储转发(store and forward)：交换机收到整个帧后转发
  目的MAC地址（6B），转也要依照CSMA/CD来转发
- 直通(cut through)：收到一个转一个，出现碎片

- 无碎片(fragment free)：都没冲突才开始转发，**最小帧保证**
  交换机不用收到整个帧而是收到64B（冲突窗口）后自动转发
- 适应性交换(adaptive switching)：自动在上面三种方式中选择

全双工模式：因为没有冲突，CSMA/CD算法可以被关闭自动翻转(auto-MDIX)：大部分交换机可以自动选择连接方式，交叉线或直通线

令牌环网(token ring)：通过在站点之间传递令牌防止冲突，具有优先权，take turns protocol 以太网没有确认机制，没有优先权必然有特殊的站点（监控站点）产生令牌帧，选举出监控站点（MAC地址最小）保有令牌帧的时间是相同的

源路由网桥：用来连令牌环网，将路径记到头部，下一次就不用查

4 网络层

4.1 IP数据报

两种数据传输技术：

- 电路交换(circuit switching)：实际接通一条物理线路，时分多路复用，电话；频分多路复用，电视；一直占用，不管有无数据交互
- 包交换/分组交换(packet switching)：统计多路复用，按需分配
  - 虚电路：需建立连接才可以传输数据（仿照电话系统，因特网之前），好处在于保留带宽
    - 交换式（要交换才建立连接）：建立虚电路(VC)表，虚电路标识符(VCI)，类似于电话
    - 永久式（建立后一直保持）：由管理员维护
  - 数据报(datagram)：不需建立连接，因特网，不预留带宽

IP协议是因特网的网络层协议

- 可路由的(routable)：全局地址，按层分配
- 尽力服务(best effort)：无连接无确认的数据报服务
- IP协议可以运行在**任何**网络上，不仅仅是因特网

IP数据报格式

- 4个字节一个字，头部最多$(2^4-1)*4=60$B，除选项20B，IPv4选项最多40B，太少了
- 生存期(TTL)限制在因特网上的停留时间，实际限制为经过的路由器数目，即跳数(hop count)，超过则自动清除，防止兜圈，每次经过路由器减1
  TTL初值默认设置为网络直径的两倍，Windows默认64
  长了就有捷径(cut-through)，因此发展到现在因特网的直径依然在32左右
- IP数据报一定要封装成帧，通过物理层传输，每次都要修改源和目的地址
- IP数据报服务类型(type of quality, ToQ)，但路由器都没有实现

IP数据报的分段和重组

- 一个物理网络的最大传输单元(maximum transmission unit, MTU)是该网络可以运载的最大有效载荷，即数据帧的数据部分的最大长度
  如：以太网(DIXv2)的MTU为1500, FDDI和令牌环的MTU分别为4353和4482
- 只要发出去一定会封装成帧（注意要加头部），帧最长就是MTU，因而要分成多段再分
- 如果一个数据报的大小大于要承载它的网络的MTU，路由器需要先对该数据报进行分段(fragment)
- 源主机每次发送IP数据报时都会把标识(Identification)字段加1。
- 分段时用标识的值保持不变，并且用偏移量字段(offset)指出该片段的数据部分相对原来数据报的偏移量(以8字节为单位)，给出原来片段的次序
- MF(More Fragment), DF(Don't Fragment)
- 小于MTU-20B，边界，一定要能被8整除，尽可能大（8字节，一定要除掉）
- IPv6中间不能分段
- 1400B=512B+512B+376B
- Path MTU discovery：找到路径上最小的MTU，发现路径上最小MTU
- 选项最后一定对齐到边界
- 生存期和头部校验（检验和）会变，其他不变

4.2 IP地址

48位的MAC地址和32位的IP地址都是全局的（全球分配），但是IP地址空间分层，是可路由的
IP地址可划分为两个部分：

- 网络号/网络前缀/网络标识：确定拥有该IP地址的主机位于哪个网络
- 主机号：确定属于该网络的哪台主机

有类网：ABC单播，D多播，E保留，地址范围如下（点分十进制）

- $0\sim127$
- $128\sim191$
- $192\sim223$
- $224\sim239$

- 240 ∼ 255

解决IPv4地址不够用的问题

- 将一个有类网可以划分为多个相同大小的子网(subnet)
  用子网掩码(subnet mask)划分边界：主机号全0，剩下的部分（网络号和子网号）全是1
  子网掩码与IP地址**相与**，若相等则在同个子网中
- 变长子网掩码(Variable-length subnet mask, VLSM)：允许把一个有类网划分为多个不同大小的子网，类似变长指令集
  解决主机数目不均匀的问题，如100、50、25、10，则不能等距划分子网
  用长度来表示子网掩码，如/26代表255.255.255.192
- 无类域间路由选择协议(classless inter-domain routing, CIDR)：将多个有类网合并为一个更大的网络，称为超网(supernet)
  可以显著减少路由表中路由的数量，称为路由聚合(route aggregation)
- 网络地址转换(network address translation, NAT)：将内部地址映射为外部地址的技术（可以扩展6w多倍），将私有地址映射为全局地址
  NAT将内部源地址转换为外部地址
  NAPT将端口号也加入NAT的映射中

地址解析协议(address resolution protocol, ARP)可以将IP地址映射为MAC地址

- 没有超时重传机制，超时没有收到响应则丢弃引发ARP查询的IP分组
- 源主机获得的映射结果缓存在ARP表中，TTL一般位2到20分钟
- 当收到ARP请求，目的主机会缓存源主机的映射，其他主机如果已缓存该映射，则会重置TTL
- 也可直接将映射加入ARP缓存，称为静态ARP映射，不会因超时而删除
- 源硬件地址和协议地址、目标协议地址都知道，但目的硬件地址不知

ARP请求广播帧，ARP响应单播帧

也有MAC映射为IP地址

动态主机配置协议(Dynamic Host Configuration Protocol, DHCP)用于主机在加入网络时动态租用IP地址，用UDP

- DHCP发现(discover)
- DHCP提供(offer)
- DHCP请求(request)
- DHCP确认(ACK)