Root bridge的选取

BID（桥ID）：BID由2字节桥优先级和6字节的桥MAC地址组成

1. 先根据BID的优先级，小项先选

2. 优先级一致时，根据MAC地址，小项先选

在**所有非根网桥**上选举根端口（root port）

Root port确保非根网桥与根网桥报文传递路径最优

PID（端口ID）：由端口优先级和端口编号组成（注意是优先对端比较）

a. 非根桥交换机上，到根桥的根路径开销最小的端口，即为该非根交换机的根端口

b. 如果根路径开销相同，则比较对端交换机的BID，越小越优

c. 如果对端交换机的BID相同，则比较对端的PID，越小越优

d. 如果对端的PID相同，则比较本端的PID，越小越优。

在**所有链路**上选举指定端口（Designated Port）

确保每一网段到根桥的路径最优

a. 在各个链路上，到根桥的根路径开销最小的端口，即为指定端口

b. 如果根路径开销相同，则比较两端交换机的BID，越小越优

c. 如果对端交换机的BID相同，则比较两端交换机的PID，越小越优

阻塞端口

既不是根端口也不是指定端口的都是阻塞端口。

Blocked state

端口被标记为阻塞状态，不转发数据包。可以防止环路的形成，确保生成树中不存在冗余的路径。

Learning state

端口处于学习状态，它会学习网络中的MAC地址，但仍然不会转发数据包。这有助于构建MAC地址表。

Forwarding state

端口在活动状态下，可以转发数据包。这是生成树协议认为是生成树上最佳路径的端口状态。

Retrieve 取回、恢复

Ubiquitous computing 普世计算（ubiquitous 无处不在的）

•Web applications: Server generates Web pages in response to client requests

•Peer-to-peer communication: Individuals form a loose group to communicate with others in the group

P to P

Instant message（即时聊天）

Twitter messaging service

Social network application

Online auction（在线拍卖）

Client-server

Online shopping

financial institution transactions（金融机构交易）

Metcalfe’s law（梅特卡夫定律）：一个网络的价值等于该网络内的[节点](https://baike.baidu.com/item/%E8%8A%82%E7%82%B9/0?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/%E6%A2%85%E7%89%B9%E5%8D%A1%E5%A4%AB%E5%AE%9A%E5%BE%8B/_blank)数的平方。

Fluctuate 波动，起伏不定

Wireless hotspots（无线热点）：基于802.11标准

两种模式Ad hoc自组网和Access Point（AP）接入点

Ad hoc没有中央控制，设备间直接相连

AP设备通过一个中央接入点连接到网络

多径衰落（multipath fading）可能引起接受信号强度的剧烈波动（不同路径信号间发生干扰）

802.11b 扩频技术 -- 11M

802.11a/g OFDM（正交频分复用）-- 54M

801.11ad限制于一个房间内使用

CSMA（carrier sense multiple access）控制冲突问题

Security 从WEP转变为WPA再到WPA2/802.1x

Ethernet network（以太网）需要switch的有线直连

Annotation 注释

Wireless与mobile不同

Wireless 无线的连接通信

Mobile 移动式连接（改变网络连接点）

NFC near field communication

CDN content delivery network（只能加速静态网页）

CDN origin server --- CDN node（CDN镜像服务器）

把资源缓存到CDN node（副本文件），提升资源获取速度

在client DNS查询时，智能DNS服务器，返回最近的CDN node

Mobile

间接路由

移动所在network会向home agent报备

信息通过home agent的转发到达IP，IP则可以直接向外发送

直接路由

外部请求访问通过request从home network处获得foreign agent的IP

FA间可以直接主动交换信息

Transit network（中继网络）

在content service与internet service provider（ISP）之间传递流量

双向收费

Transit network也被称为backbone network（骨干网络）

Peering network 直连网络

通常对等，直接实现流量交换，没有第三方收费

VPN virtual private networks

Connect individual networks at different sites into one logical network

PAN personal area networks（bluetooth）

Local area network

Diverse 多样的

Metropolitan Area Networks（MAN城域网）

Television signals and internet 被连接至head-end

Wide area network（WAN 广域网）

Illustrate 说明

Network 由subnet（子网）和host（主机）构成

Subnet也可以被称为（ISP network）

Internetwork可能被称为WAN network

Internet 可以用于连接LAN与WAN

不同网络间的连接需要gateway（网关）

Packet switching（分组交换），数据包发送，发出地与目的地重装

Circuit switching（电路交换），建立连接，一直占用通信资源

Switching office 交换局，用于处理本地电话电话通信

Toll office长途局，处理长途电话通信

ARPANET中的站点 -- IMP（interface message processor）连接至subnet

ARPANET将host与subnet split，两种类型三种协议

---> NSFNET

Infrastructure 基础设施

依赖internet的television

HFC （Hybrid Fiber-Coaxial）混合光纤同轴 = 光纤 + 同轴电缆

DOCSIS是一种基于数据包的传输协议，用于在HFC网络上传输多种数据服务，包括电视信号、高速互联网数据和语音通信。

Home end called cable modem

Head end called CMTS（cable modem termination system）

DSL digital Subscriber Line（数字用户线）

DSL 技术通过普通的电话线传输数字信号，允许用户在使用电话线进行语音通信的同时，通过同一根电话线实现高速数据传输

一小部分给voice，一部分给upstream，绝大部分给downstream（高HZ，高带宽）

ADSL（Asymmetric Digital Subscriber Line）非对称数字用户线，用于在普通电话线上提供高速互联网接入。ADSL的特点是其非对称性，即上行速率和下行速率可以不同，通常下行速率高于上行速率

Network结构

Backbone provider

Regional ISP

ISP

Customer IP network

Mobile network

EPC 核心网络架构

MME（mobility management entity）移动性管理单元

SGW（serving gateway）处理流量转发，设备切换

PGW（PDN gateway）与外部数据网络连接的关键点

PCRF（policy and charging rules function）处理策略与计费规则部分

蜂窝网络：相同频率的信道一般不会在相邻的单元格中使用

相邻小区越来越小的趋势 ---> 对映于更大的用户需求

4G known as LTE（long term evolution）

1G使用模拟信号

2G使用数字信号，可加密

GSM（for mobile communication）：使用124 frequency channels --- 8 slot每一TDM frame

分mobile to base and base to mobile（各124 channels）

3G支持digital voice and data

4G（又称为 IMT advanced）

only based packet switching

语音流量与数据网络分离

5G MIMO技术：通过使用多个天线进行信号的发送和接收，以提高通信系统的性能

Cable network（有线网络）

通过headend统一接收信号

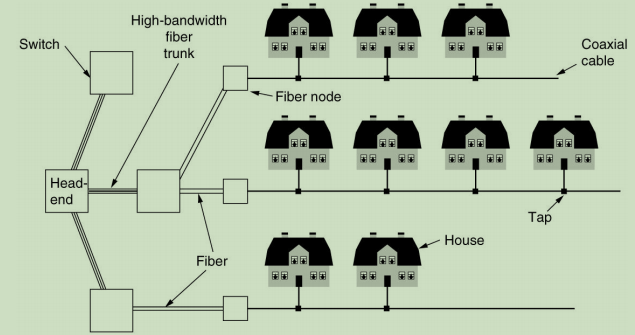
Coaxial cable（总线）传递

Tap形成分支

Drop cable进行分支信号（数据）的传输

HFC 有线宽带互联网接入（使用DOCSIS标椎）

Headend使用fiber trunk进行传递，至fiber node后再转换为coaxial cable



Fixed phone system使用local loop + end office + trunk结构

Reconcile 调和，妥协

Negotiation：协商

Network protocol

Reliability：检查error/纠正error，find work path using routing

Resource allocation：flow control/congestion problem

Evolvability：扩展性

Security：保密性（confidentiality），身份验证（authentication），完整性（integrity）

Network architecture: a set of layers and protocols

Protocol stack: a list of the protocols used by a certain system, one protocol per layer

Protocol：对等层际（peer）之间

Interface：上下层之间接口

Connection-oriented service

Message stream/byte stream，unreliable connection（VoIP）

Connectionless service

Cut-through switching（直通交换）未完全接收就开始转发

Unreliable datagram（junk mail）

Acknowledge datagram（text message）

Request-reply（database query）

Service primitive

Listen connect accept receive send disconnect

OSI model

Physical、 data link/network/transport/session（建立同步）、presentation（数据压缩，加解密）、application

TCP/IP model

Link --- SONET/SDH (Synchronous Optical Network)同步光纤网络：整个同步网络的各级时钟都来自一个非常精确的主时钟

同步发送 --- 没有数据则发送“空”

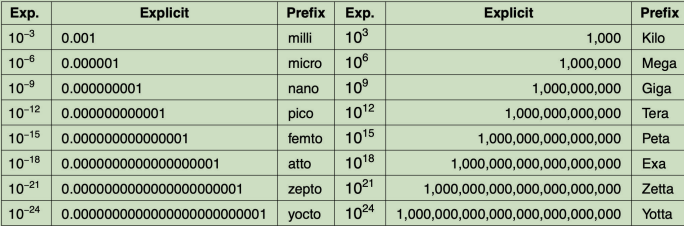
拆解为link和physical层更好

Internet --- IP/ICMP

Transport --- end to end TCP/UDP

Application --- HTTP/SMTP/RTP/DNS/FTP（控制连接21端口，数据连接20端口）

Critique：评论/批判



Persistent storage：磁带存储，物理传递

Twisted pair 双绞线

非屏蔽双绞线 UTP

屏蔽双绞线 STP

Coaxial cable同轴光缆（传输容量较双绞线更大）抗干扰能力更强

Fiber optic光纤

存在Attenuation：衰减，衰减作用 --- 与wavelength有关

Defined as input / output（signal power）

Light source：LED（只支持multi mode） / 激光（更快，更远，存在时间较短，昂贵，支持multi mode和single mode）

Frequency hopping spread spectrum 跳频扩频（FHSS）

载波频率不断跳变（发送方与接收方同步的固定时间间隔跳变）

抗干扰（离散频点通信）

兼容性强（频点多）

Direct sequence spread spectrum 直接序列扩频（DSSS）

频谱拉宽

抗衰落（因为带宽较宽的好处）

抗多路径干扰

Ultra-wideband communication 超宽带通信

Radio transmission（无线电传播）

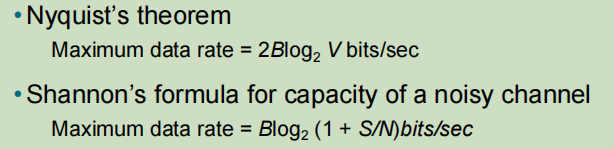
可在电离层反射

传播距离长

Light transmission（光传输）

对流会对激光通信产生干扰

The maximum Data Rate of a Channel



Baseband transmission 基带传输（数值信道传输 -- 矩形脉冲）

RZ 高代表1，低代表0，每个时钟周期均归零（同步实现）

Non-return to zero（高表示1，低表示0） --- 难以同步

NRZ invert（中间变化为1，不变为0）

Manchester（曼切斯特编码），（中间）高变低为1，低变高为0，频带宽度是基带宽度的两倍

差分曼切斯特，有高变低，低变高两种信号（同前就是0，改变就是1）

Bipolar encoding（双极性编码）进制数字0被表示为零电平，而二进制数字1则交替表示为正电平和负电平（一次增一次减）

4B/5B mapping

Passband transmission（频带传输）

Amplitude shift keying 幅移键控 AFK（幅度变化）

Frequency shift keying 频移键控 FSK（频率变化）

Phase shift keying 相移键控 PSK（相位变化）

QPSK 旋转90 -- 对映四个状态（看首字母，DPSK对映2个状态）

QAM-16 16个状态（相位 + 振幅结合）

QAM-64 64个状态（相位 + 振幅结合）

Frequency division multiplexing（FDM）频分多路复用

不同信号使用不同频率

OFDM 使用正交信号（不同频率信号可以融合分离，同时使用信道）

Time Division Multiplexing（TDM）时分复用

不同信号轮转使用，可使用全频率全时间

Code Division Multiplexing（支持动态）

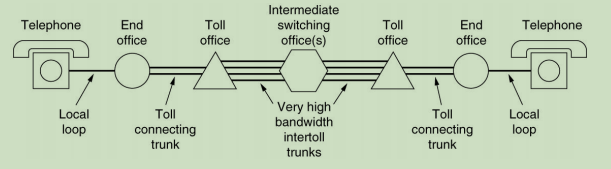
单频道信号相加（发1乘1，发0乘-1，不发乘0） ---> 总信号

拆解时，总信号乘以单频道信号即可（正发1，负发-1，不发为0）

Wavelength Division Multiplexing（WDM）波分复用

波分复用 一根光纤上使用不同的波长来传输多路光信号

Structure of the telephone system



Local loop：telephone modems（电话调制解调器），ADSL（非对称数字用户线），fiber（光纤）

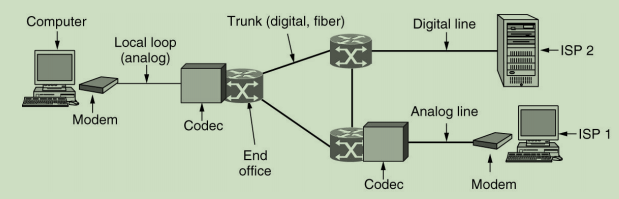
computer-to-computer call use digital transmission

可以使用4B / 5B编码策略

Modem可以将（发送端）数字信号转化为模拟信号，（接收端）将模拟信号转化为数字信号

Codec也可以双向转变

Analog line（传输模拟信号的线路）只能传输模拟信号



T1

1.544Mbps =（1 +（7 + 1）\*24 ）--- 193bit / 125μs

24个信道 56k/channel

1 --- 成帧码、7 --- data传输/channel、第8bit --- signaling

4:1 T2、7:1 T3、6:1 T4（层层向上）

E1

32个信道，0号信道供同步使用，另外31个信道为有效信道

2.048Mbps --- 64k/channel

Store and forward transmission 存储转发传输

Cellular network蜂窝网络

Geostationary satellites（同步卫星）

Bandwidth从小到大：L、S、C、Ku、Ka

L、S会被crowd遮挡

C会受到地面干扰（Terrestrial interference）

Ku、Ka会受到rain的干扰

VSAT使用hub两次通过卫星介导通信

Data link layer design issue

需要为packet添加header and trailer

Unacknowledged connectionless service

Acknowledged connectionless service

Acknowledged connection-oriented service

Frame帧定界的方式

Byte count（帧一开始时指明字节的个数）

Flag bytes with byte stuffing（字节填充标志）

特定字节标志表示帧的开始与结尾，文中正常出现时需要转义

在header与trailer前/后增添

FLAG标志开始/结束，ESC表示转义（本身使用也要转义）

Flag bits with bit stuffing（比特填充标志）

类似于字节标志，文中出现时插入额外bit信息（阻止其出现）

不能出现六个连续的1

Physical layer coding violations（物理层编码违规）

编码违规是指在数据中插入一些特殊的比特模式，这些模式在正常数据传输中是不允许的。通过检测这些编码违规，可以确定帧的开始和结束。

Error control

Ensure all frames be delivered

To the network layer destination

In proper order

Ensure reliable/connection-oriented service

Require acknowledge frames and timers

Error correcting

FEC（forward error correction）向前纠错

Hamming codes

海明距

检错d bit需要 d + 1（还是错误的，没有转变为正确的）

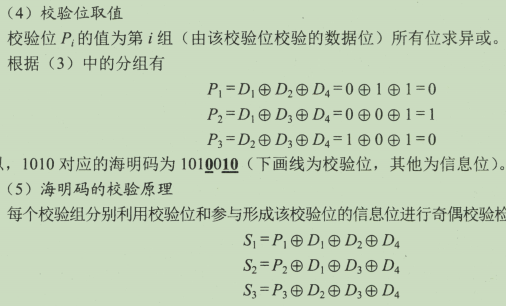
纠错d bit需要 2d + 1（离当前还是最近的，不会被错误纠错）

海明码

（位数确定）n bit信息位 + k bit海明码 + 1 bit（全部正确）<= 2^k（总）

海明码放置于2 ^ (i-1) 位次上

校验位的计算与校验

Di二进制位次占位

Binary convolutional codes（感觉有点复杂，编码过程见知乎）

Reed-solomon codes（里德-所罗门码） --- 线性代数关联，怪（知乎）

Low-density parity check codes（低密度奇偶校验码）

Error detecting

Parity check（奇偶校验）一般而言额外1 bit补至偶数个1

Checksum 通过“求和”操作进行校验

CRC（Cyclic Redundancy Code）循环冗余码

G(x) 多项式 --- 除数（阶数对映多项式最高次）

FCS （帧检验序列） --- r bit（对映阶数）

帧/报文 m bit + 最低位补计算得余数（r bit），生成计算时补r bit 0

异或除获得余数（校验时需要为0，生成时为拼接目标）

Flow control

Link-layer protocols

Utopia：no flow control or error correction

Stop-and -wait：adding flow control

Sequence number and ARQ（自动请求重传）：adding error correction

Improve efficiency

Bidirectional transmission：piggybacking（确认信息、确认帧可以与数据传递帧合并）

双向传递数据

Sliding window protocols

Send window and receive window

One-bit sliding window（send/receive均只有one window）

Go-back-n（GBN），send window可以连续发帧，receive window size只有1

1 < Send window <= 2^n - 1（n对映sequence的bit数）

Selective repeat（SR）send window and receive window均可不为1

Send window and receive window应该均小于等于 2^(n-1)，大于1

Data link protocol

Packet over SONET（POS）

同步光网络中传输分组化数据

IP数据报 --- PPP frame --- SONET帧（POS是对映的数据包）

PPP（point-to-point）面向字节，长度为字节的整数倍

数据链路层协议，直接的，点对点数据通信链接

Error detection，separate packets

Link control protocol（LCP）

两个通信设备之间建立、配置和测试数据链路连接

Network control protocol（NCP）

在已建立的数据链路上协商并激活网络层协议，PPP可以同时支持多个NCP，每个NCP负责一个特定的网络层协议

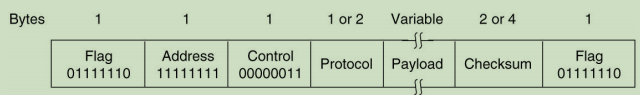
提供身份验证机制，可检错**不能纠错**，可在不同的物理介质上传输（CRC检验）

PPP只支持全双工，点对点，且两端可以运行不同的网络协议（默认异步）

Frame结构，variable部分不超1500 Byte（最小0 byte）

不采用CSMA/CD协议，不需要最短帧限制

Protocol 0x0021表示IP数据报

另外有HDLC协议，面向bit，使用编号和确认机制

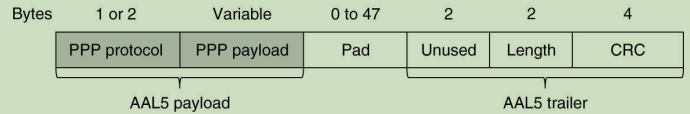
ADSL（asymmetric digital subscriber loop）

借助与local loop + ISP’s office link internet，（PPP用于在用户与ISP之间建立连接）

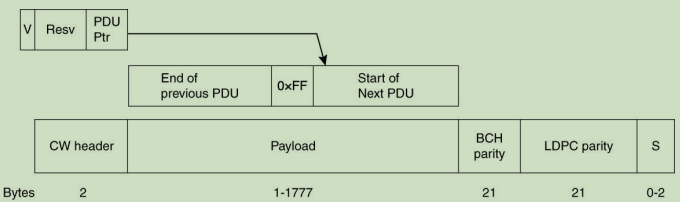
AAL5将PPP帧进行封装，使其适配ATM网络

ATM是面向连接的网络技术，以小而固定大小的细胞来传输数据

PP 通常用于在用户端和ISP之间建立连接，而AAL5和ATM技术则用于在电话线上传输数据



DOCSIS frame 映射到物理层



Static channel allocation：存在带宽浪费的情况

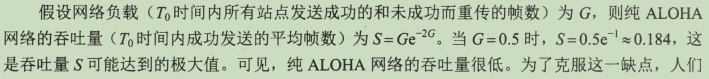
Dynamical channel allocation

ALOHA（Additive Link On-line HAwaii system）

Pure ALOHA（可以应用至总线型网络和无线信道网络）

网络中站点可以任意发送数据，如果在一段时间内未收到确认，那么该站点就认为传输过程中发生了冲突；未有确认则需要重发数据，直至发送成功

使用随机等待重传



Slotted ALOH 时隙ALOHA协议

将time划分为等长的时隙，只有在时隙的start处，才可以发出数据包

使用随机重传

f516316c1c5530faf5ef4980857ac9a

CSMA（carrier sense multiple access）

每个站点在发送前都先监听一下共用信道，发现信道空闲后再发送，

1. persistent CSMA

如果信道空闲，则立即发送数据；如果信道忙，则等待，同时继续监听直至信道空闲；如果发生冲突，则随机等待一段时间后，再重新开始监听信道

Non-persistent CSMA

监听信道；如果信道空闲，则立即发送数据；如果信道忙，则放弃监听，等待一个随机的时间后再重复上述过程

p-persistent CSMA

如果信道忙，则持续监听；如果信道空闲，则以概率p发送数据，以概率1 - p推迟到下一时隙；下一时隙仍重复上述过程

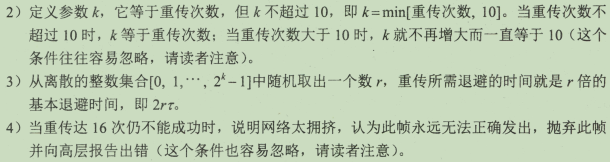
CSMA with collision detection（CSMA/CD）

以太网使用，作用于半双工网络

Collision Detection：边发送边监听，如果监听到了碰撞，则立即停止数据发送，等待一段随机时间后，重新开始尝试发送数据

到端往返时间称为“争用期”（冲突窗口、碰撞窗口） --- 最短帧长

截断二进制退避算法



Collision-free protocol（碰撞避免协议）

Bit-map protocol

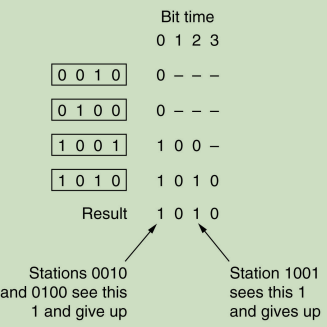
想发送数据的单元置位示意，数据包按从小至大依次发送，每次单元发送占用全部的带宽

Token passing

令牌传递，拥有令牌的才能传送数据

只有当data回收后，才释放令牌，令牌依次传递，需要发送数据者持有+发送数据

Binary countdown

最左侧可以认为是二进制地址分配，在时钟同步的前提下，每个（需要发送数据的）地址依次发出当前位bit，若发射为0接收为1，则该地址停止后续位发送；重复直至所有位发送完毕，仅result与地址一致者可发送数据 --- 等价于高位优先措施

Limited-contention protocol（有限争用协议）

Use contention at low load to provide low delay

Use collision-free technique at high to provide good channel efficiency

是一种非对称协议

自适应树竞争协议（Adaptive Tree Walk Protocol）

在一次数据包成功传输后的第一个竞争时隙，所有站点同时竞争。

如果只有一个站点申请，则直接获得信道；否则在下一竞争时隙，有一半站点参与竞争，再下一时隙由另一半站点参与竞争，即所有站点构成一棵完全二叉树

CSMA/CA协议（无线通信）

Collision avoidance（CA）碰撞避免

SIFS：短间隔，分隔一次对话的帧

PIFS：在PCF操作中使用（一般不涉及）

DIFS：异步帧竞争访问的时延

检测信道空闲，等待DIFS后，发送整个数据帧；否则，使用退避算法，随机回退，若信道忙则计时器不变，若信道空闲则计时器倒计时；计时器为0时发送整个帧并等待确认，若发送第二帧，则重复步骤（使用随机间隔）

隐蔽站问题（基于AP的管理）：

RTS（request to send）：请求发送数据

CTS（clear to send）：请求准许与阻抑其他单元发送数据

可以引入NAV的概念（从RTS --- ACK、CTS --- ACK）

不同站点保持静默时长不同

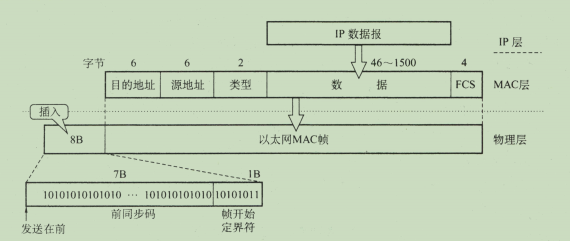
Ethernet

通常将802.3局域网简称为“以太网”



其他单元的Back-off避让需要在接收到ACK后再进行

以太网MAC帧结构

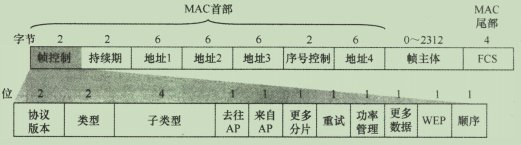


802.11的frame structure

地址1是直接接收数据帧的结点地址

地址2是实际发送数据帧的结点地址

地址3：源地址（来自AP时），目的地址（去往AP时）



Hub 同一冲突域

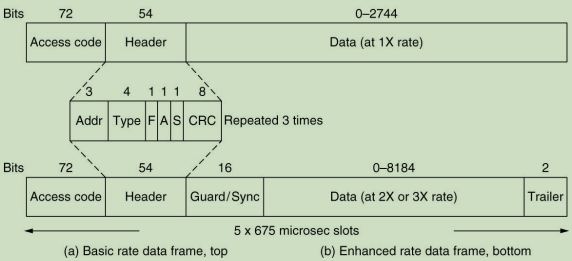
Switch 分离冲突域

Hub、Switch属于同一广播域

Bluetooth architecture

与一般的协议不同，Bluetooth的protocol可能跨越layer产生功能（称为profile）

Bluetooth 的frame structure



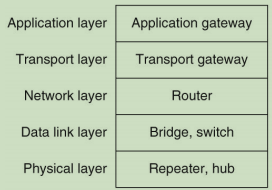
不同transport之间使用gateway

Bridge与switch均在data link层使用，其中switch多应用于MAC层，而bridge多应用于LLC层；bridge常用语连接不同的局域网

LLC与MAC层的区别

LLC子层主要功能是传输可靠性保障和控制，数据包的分段与重组，数据包的顺序传输（802系列定义）

MAC子层主要功能是数据帧的封装/卸装，帧的寻址和识别，帧的接收与发送，链路的管理，帧的差错控制



Leaning bridge

使用反向学习

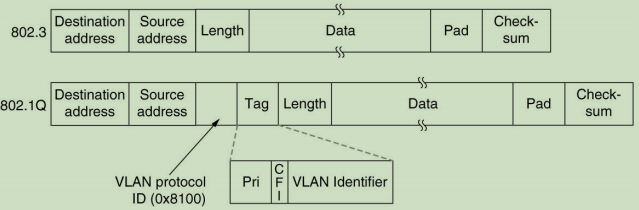
分割冲突域，广播信息则自动全局转发

需要使用“生成树协议”防止网络风暴的出现（root bridge等详细内容见文档开始时的分析）

Virtual LANs（虚拟局域网）

划分多个广播域，二层交换机需要trunk口才能介导通信

802.1Q协议，可以携带VLAN头部，使不同VLAN间数据包可以通过bridge传输

其frame structure差别：

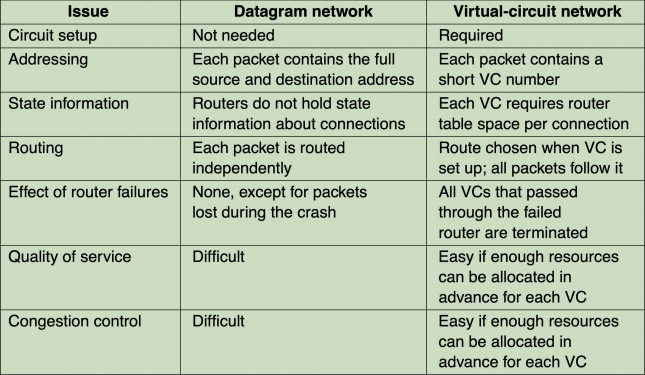
Network layer

Store and forward packet switching

Connectionless service（datagram network）

Connection-oriented service（virtual-circuit network）

两者的比较：



Optimality principle

从所有的源节点到一个指定目标节点的最优路径的集合构成了一棵以目标节点为根的树，这样一棵树被称汇集树（SinkTree）

Distance vector routing

所有结点都定期地将它们的**整个路由选择表**传送给所有与之直接**相邻的结点**

采用“跳数”作为距离的度量，维护routing table，记录到每个node的最小跳数

Count-to-infinity problem（无穷计算问题）

坏消息传递较慢（一段网络断开后的无限扩大）

改进：消息单向传递，消息来源依赖不回发消息（水平分裂法）

RIP（routing information protocol）路由信息协议（应用层协议使用UDP传输数据，端口：520）

规定从一个路由器到直接连接网络的距离（跳数）为1；而每经过一个路由器，距离（跳数）加1

RIP允许一条路径最多只能包含15个路由器（即最多允许15跳）。因此距离等于16时，它表示网络不可达

RIP默认在任意两个使用RIP的路由器之间每30秒广播一次RIP路由更新信息

在RIP中不支持子网掩码的RIP广播

RIP默认超时时间为180秒，如果180秒还没有收到相邻路由器的更新路由表，那么把此相邻路由器记为不可达路由器

OSPF（开放最短路径优先协议）

向自治系统中所有路由器发送消息，本路由器相邻的所有路由器的链路状态

OSPF是网络层协议，它不使用UDP或TCP,而**直接用IP数据报**传送

如果到同一个目的网络有多条相同代价的路径，那么可以将通信量分配给这几条路径（多路径间的负载平衡）

每个链路状态都带上一个32位的序号，序号越大，状态就越新

每个路由器根据自己建立的全网拓扑结构图，使用Dijkstra最短路径算法计算从自己到各目的网络的最优路径

五种分组类型：

1） 问候分组，用来发现和维持邻站的可达性（通常每隔10秒，每两个相邻路由器要交换一次问候分组）

2） 数据库描述分组，向邻站给出自己的链路状态数据库中的所有链路状态项目的摘要信息

3） 链路状态请求分组，向对方请求发送某些链路状态项目的详细信息

4） 链路状态更新分组，用洪泛法对全网更新链路状态

5） 链路状态确认分组，对链路更新分组的确认

2/3步用于建立全网同步的链路数据库，对于“状态更新”需要返回“状态确认”消息

IS-IS (Intermediate-System to Intermediate-System)

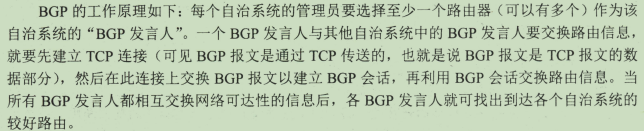
*（详细见CSDN收藏）*

BGP（边界网关协议）

不同自治系统的路由器之间交换路由信息的协议，是一种外部网关协议。边界网关协议常用于互联网的网关之间

其只力求寻找一条能够到达目的网络且比较好的路由，而并非寻找一条最佳路由

BGP是应用层协议，其是基于TCP的

不同自治系统间消息传递需要（不一定）依赖于主干网

在BGP刚运行时，BGP的邻站交换整个BGP路由表

Link state routing

当前route将自己与相邻router的cost发布给所有自治系统中的router

多播发送，运行相同协议router接收

多播（也称组播）：给一组特定的主机发送数据（需要D类IP地址）

一个group最多可以有28 bit available element

Hierarchical routing within a network

Hierarchical routing具有多个region

Router并不记录full table，而是可以将每一个region进行整合当做同一个目的地（仅记录最小跳数）

Broadcast routing

Reverse path forwarding（RPF逆向路径转发）

为了防止组播报文在转发过程中出现重复报文及环路的情况，路由器必须执行RPF检查，防止环路的产生

Multicast routing（组播，见前）

Multicast tree往往为broadcast tree的子图

Anycast routing（任播，IPv6类型）

在任播通信中，数据包被发送到一组设备中的任何一个（最近的一个，仅一个）

流量控制（traffic management）

Congestion（拥塞控制）

Traffic-aware routing 流量感知网络，根据时刻的流量大小进行路径选择

Admission control 准入控制

Traffic shaping（流量整形）：控制网络上的数据流量，以便使其在网络上的传输更加有序

“漏桶”类比，声称进水量 -- 实际出水量间的不匹配

数据包流量不丢失，只是先进行缓存

流量调节的三种方法

Choke pocket（抑制包）：路由器选择一个被拥塞的数据包，给其源主机返回choke packet（抑制主机）

显式阻塞通知（ECN）：拥塞的路由，加上ECN标识，主机返回告知host 拥塞的发生，以便于控制发送方的速率（抑制主机对映端口）

逐跳后压（hop-by-hop backpressure）：抑制包影响途经的每一跳路由器（抑制每一个路由器）

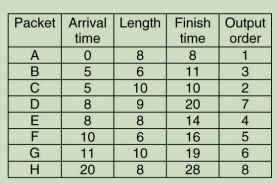
Packet scheduling

Fair queueing：在不同queue之间使用RR（round-robin）进行调度

分类进入不同的queue

Weighted fair queueing：为每个数据流分配一个加权值来控制该数据流在队列中获得带宽资源的比例（比例越大，带宽占比越大）

每个queue中数据包依次发，不同queue可以同时发（带宽占比等价于数据发送倍速）



Integrated service（个性化定制）：为每个网络流量流提供个性化的服务，使用 Resource Reservation Protocol (RSVP) 协议，应用程序可以向网络发送请求，要求特定的服务质量

Differentiated Services（分类）：通过对数据流进行分类，为不同的流量提供不同的服务，这是通过在数据包头部的 Differentiated Services Code Point (DSCP) 字段中设置不同的值来实现的

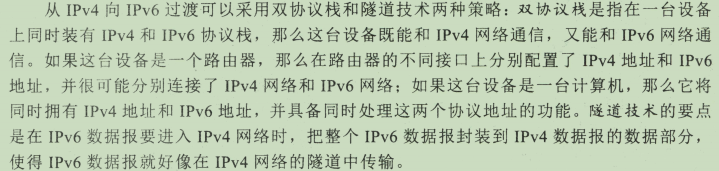
Connecting heterogeneous networks（连接异构网络）

MPLS（Multi-protocol Label Switching）是一种在网络中进行数据包转发的技术。它是一种基于标签的多协议交换技术，通常用于提高数据包在网络中的传输效率和可管理性

可以连接不同标准的network，将不同标准的header信息替换为MPLS header信息，进行异构network的传播

转发速度极快，对映于label edge router硬件

IPv4和IPv6的过渡策略：双协议栈和隧道（tunneling）技术：

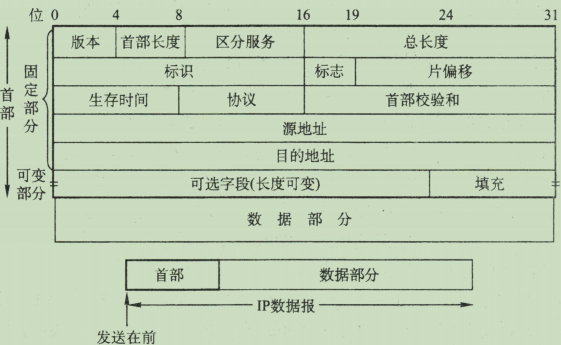


Software defined networking（软件定义网络）

采用集中式的控制平面和分布式的数据平面，路由器之间不再相互交换路由信息；在网络的控制平面有一个逻辑上的远程控制器（可以由多个服务器组成）。远程控制器掌握各主机和整个网络的状态，为每个分组计算出最佳路由，通过Openflow协议（也可以通过其他途径）将**转发表**（在SDN中称为**流表**）下发给路由器

对上层应用的开发者，SDN提供的**编程接口称为北向接口；**SDN**控制器和转发设备**建立双向会话的接口称为**南向接口**，通过不同的南向接口协议（如Openflow）, SDN控制器就可兼容不同的硬件设备；SDN**控制器集群**内部控制器之间的通信接口称为**东西向接口**，用于增强整个控制平面的可靠性和可拓展性

IPv4数据包结构

首部长度：4bit，以4B为单位

总长度：首部与数据之和长度（B为单位）

标识：唯一区分（每产生一个 + 1）

标志：最低位为MF，MF = 1表示后面还有分片；中间位为DF，DF = 0才允许分片

片位移：每个分片的长度一定是8B的倍数

首部校验和：类似于UDP累加计算（四个十六进制数一组），计算溢出位加回原式，最后求反码（加上校验位结果为FFFF --- 未求反码）

子网掩码

CIDR（classless interdomain routing）

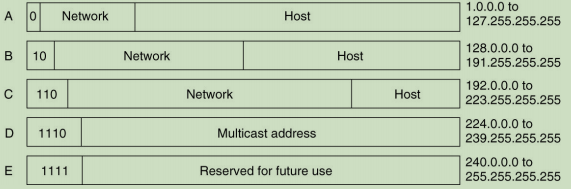
CIDR使用“斜线记法”（或称CIDR记法），即IP地址/网络前缀所占比特数

Aggregation of IP prefixes（前缀聚合）

将网络前缀都相同的连续IP地址组成“CIDR地址块”。一个CIDR地址块可以表示很多地址，这种地址的聚合称为路由聚合，或称构成超网

最长前缀匹配：使用CIDR时，路由表中的每个项目由“网络前缀”和“下一跳地址”组成。在查找路由表时可能会得到不止一个匹配结果。此时，应当从匹配结果中选择具有最长网络前缀的路由

IPv4地址



主机号全为0表示本网络本身

主机号全为1表示本网络的广播地址（直接广播地址）

1. x.x.x保留为环回自检

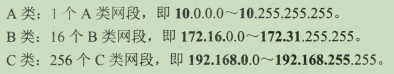
32位全为0，表示本网络的本主机

32 bit全为1，表示整个TCP/IP广播地址（受限广播地址）

NAT（网络地址转换，network address translation）--- 传输层

私有IP地址（可重用地址）

多个私有IP地址映射到一个全球IP地址（端口对映不同）



IPv6

地址空间由32 bit转换为128 bit

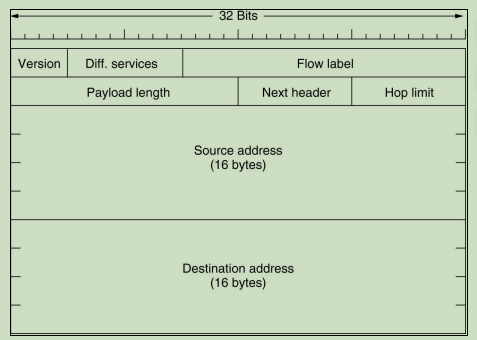
包的源结点才能分片，是端到端的，传输路径中的路由器不能分片

IPv6首部长度必须是8B的整数倍，而IPv4首部是4B的整数倍

IPv6的表示

每四个十六进制数之间使用“:”分隔，若有连续的0，可将任意连续的0（需要为4的倍数个）简化为“::”的形式（::只能含有一个）

Structure



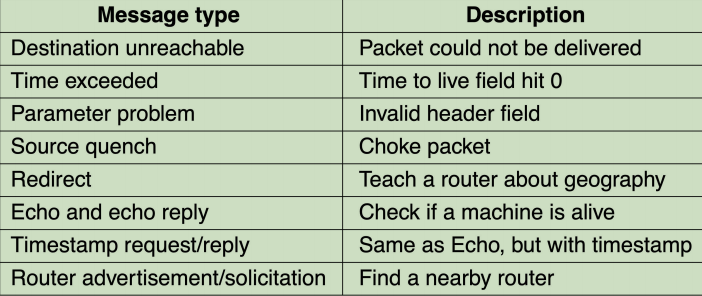
Flow label（流标签）：同的流标签+源地址可以唯一确定一条数据流

Payload Length（数据包的有效载荷）：以B为单位，指IPv6基本报头后的长度和扩展报头长度

Hop Limit（跳数限制）：类似TTL字段

Next Header（下个报头）：长度8bit，指明跟在基本头后面是哪种扩展头或者上层协议中的协议类型，如果只有基本报头，没有扩展报头，该字段标识上层使用协议，如UDP为6，TCP为17

Source Address（源[IPv6地址](https://so.csdn.net/so/search?q=IPv6%E5%9C%B0%E5%9D%80&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/qq_42606357/article/details/_blank)）：必须是IPv6单播地址  
 Destination Address（目的IPv6地址）：必须是IPv6单播地址或组播地址

ICMP—The Internet Control Message Protocol（网络层协议）

差错报告报文：

终点不可达：不能交付数据报时使用

源点抑制：拥塞而丢弃数据报时，使源点降低发送速率

时间超时：TTL为0，且不是目的接收主机；或目的主机在规定时间内不能收到一个数据报的全部数据信息

参数问题：当路由器或目的主机收到的数据报的首部中有的字段值不正确时，就丢弃该数据报，并向源点发送参数问题报文

改变路由（重定向）：路由器把改变路由报文发送给主机，让主机知道下次应将数据报发送给另外的路由器（可通过更好的路由）

对127.x.x.x和0.0.0.0不发送ICMP差错报告报文

对ICMP差错报告报文不能再发送ICMP差错报告报文

对具有组播地址的数据报都不发送ICMP差错报告报文

对第一个分片的数据报片的所有后续数据报片都不发送ICMP差错报告报文

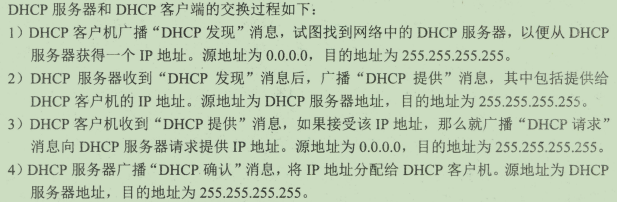
ARP—The Address Resolution Protocol（地址解析协议）--- 网络层协议

完成从IP地址到MAC地址的转变，每一台主机会设有一个ARP高速缓存（ARP表）

通过使用目的MAC地址为FF-FF-FF-FF-FF-FF的帧来封装并广播ARP请求分组（广播发送），目标主机B收到该ARP请求后，向主机A发出ARP响应分组（单播发送）

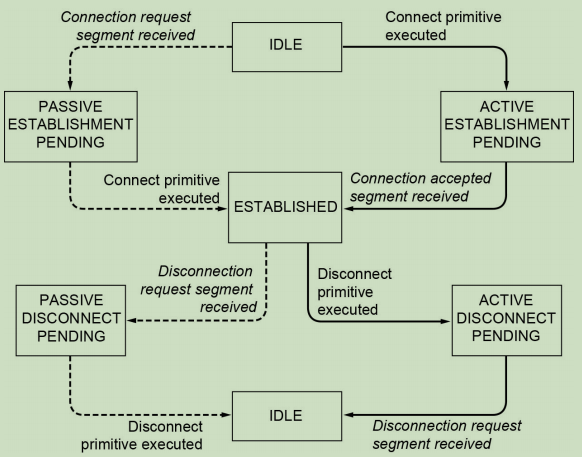
DHCP—The Dynamic Host Configuration Protocol（动态主机配置协议）---应用层协议（基于UDP）

DHCP服务器给主机动态地分配IP地址（DHCP服务器回答报文称为**提供报文**）

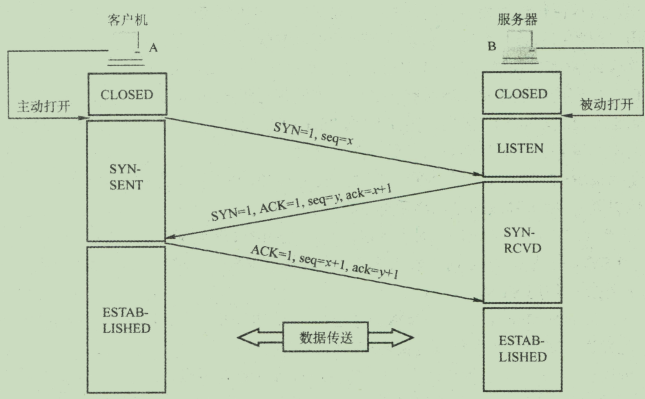
对于多个DHCP请求，DHCP服务器采用先到达先服务原则

Transport service

状态变换如下（实线表示客户端，虚线表示服务器）



TCP的connection establishment（注意其中状态）



TCP规定，SYN报文段不能携带数据，但要消耗掉一个序号

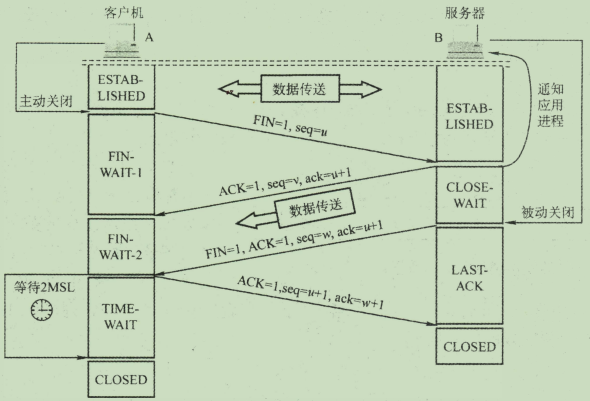
第一次、第二次握手都不能携带数据，第三次握手时可以携带数据（若不携带数据则不消耗序号）

TCP连接释放（特别注意状态描述）

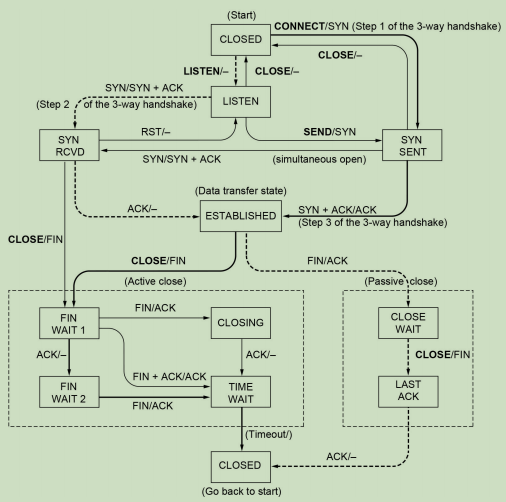
四次握手释放，FIN与ACK信号的交互

FIN = 1（去），ACK = 1（来），FIN = 1 ACK = 1（来），ACK = 1（去）

注意2MSL （最长报文段寿命）等待后才可真实释放连接



合体状态图



三种buffer类型（可能会存在unused space）

Chained fixed-size buffers

Chained variable-sized buffers

One large circular buffer per connection.

Multiplexing

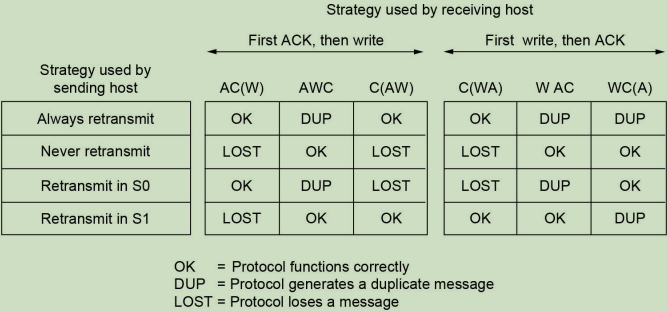
发送端（复用）：发送方不同的应用进程可使用同一个传输层协议传送数据

接收端（分用）：指接收方的传输层在剥去报文的首部后能够把这些（不同）数据正确交付到目的应用进程（一个）

Crash recovery（灾难恢复）

接收方：先发送ACK还是先write

发送方：always retransmit（总是重发），never retransmit（总不重发）

A： ACK确认；C：crash产生；W：write进入buffer（顺利接收与提交）

Congestion control

Max-min bandwidth allocation：最小最大公平性分配原则

假定每两个router之间的bandwidth都是1，将1平均分配给这两个router之间的所有link，一步步推算（非所有link均相等，但每两个router之间不能超过最大bandwidth）所有router之间的bandwidth

Regulating the sending rate

网络层可以知道数据流量大小

传输层只能通过发包与回包的间隔判断流量及运输情况

TCP在感知网络拥塞时会主动降低传输速率，UDP不会

Additive Increase Multiplicative Decrease (AIMD) control law（加增乘减）

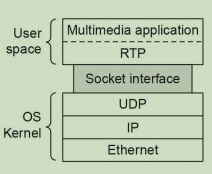
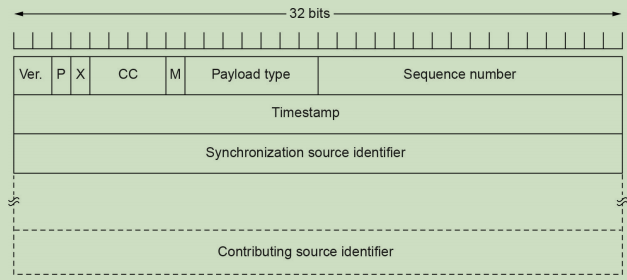
UDP

Real-time transport

RTP：the Real-time Transport Protocol（应用层）

RTCP：the Real-time Transport Control Protocol

RTP与UDP关系如下：

RTP header如下：

RTP允许少量的丢包

RTP建立连接流通信，声明开放端口的流传递（sequence number意味着连接）

Timestamp用于同步播放，identifier对映唯一连接（随机产生）

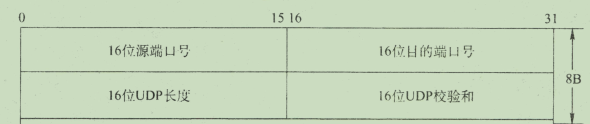
因为网络抖动，packets arrive at buffer的时间可能断断续续，因而需要缓冲一定的packet数量再进行播放（smooth the output stream）

抖动：不同packet数据包发送延迟存在的差别（差别较大 --- 抖动较大）

使用UDP的应用主要包括小文件传送协议（TFTP）、DNS、SNMP和实时传输协议（RTP）

UDP仅有8B的header开销，且UDP面向报文，发送方UDP对应用层交下来的报文，在添加首部后就向下交付给IP层，一次发送一个报文，既不合并，也不拆分

Header structure

长度包括首部和数据部分

校验和（可选）：不想计算就全写0

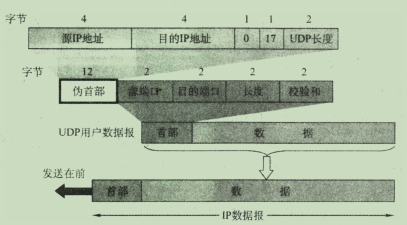
在计算校验和时，要在UDP数据报之前增加12B的伪首部，伪首部并不是UDP的真正首部，只是在计算校验和时，临时添加在UDP数据报的前面

若UDP数据报的数据部分不是偶数个字节，则要在数据部分末尾填入一个全零字节（但此字节不发送）

伪首部的UDP长度，也记录“头部 + 数据”的总长度

校验和计算：

4个十六进制数一组，相加，溢出位末尾加回，最后求反（只需要计算结果，一般要求取反后全为0）



RPC协议（远程过程调用协议） --- 基于UDP

通过库入口调用服务器上库代码 --- （瘦客户端）

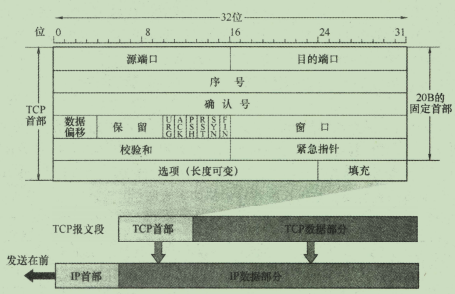
The internet transport Protocol（TCP）

TCP提供面向连接的服务，在传送数据之前必须先建立连接，数据传送结束后要释放连接。TCP不提供广播或组播服务

TCP提供可靠交付的服务，保证传送的数据无差错、不丢失、不重复且有序

TCP提供全双工通信，允许通信双方的应用进程在任何时候都能发送数据，为此TCP连接的两端都设有发送缓存和接收缓存

TCP面向字节流，TCP报文长度根据接收方给出的窗口值和当前网络拥塞程度来决定

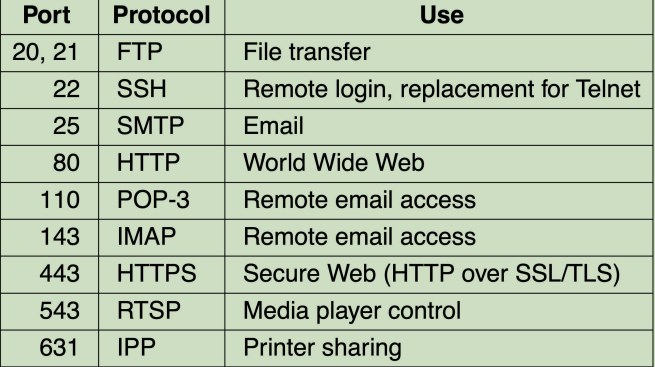
Header Structure

TCB头部最小为20B，头部长度（数据偏移）需要为4B的整数倍

窗口：现在允许对方发送的数据量

校验和：检验首部和数据两个部分，计算方法基本同UDP，需要加上12B的伪首部（只需将UDP伪首部的协议字段的17改成6, UDP长度字段改成TCP长度，其他的和UDP一致）

TCP支持的model



TCP sliding window

在数据包回包中，Receiver应该告知接收区buffer的大小

Round-trip time 往返时间

Congestion control

cwnd拥塞窗口：接收方根据目前接收缓存大小所许诺的最新窗口值

rwnd接收窗口：发送方根据自己估算的网络拥塞程度而设置的窗口值

慢开始算法：成功发送则成倍递增cwnd

拥塞避免算法：到达cwnd阈值（ssthresh）后，每次发送成功cwnd仅增加1

网络拥塞后，把慢开始门限ssthresh设置为出现拥塞时的发送方的cwnd值的一半（但不能小于2）

把拥塞窗口 cwnd重新设置为1（Tahoe）

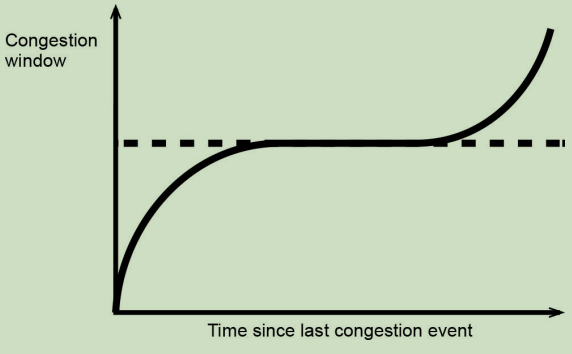
把拥塞窗口 cwnd设置为阻塞前cwnd的一半（快恢复）（Reno）

快重传：当发送方连续收到三个重复的ACK报文时，直接重传对方尚未收到的报文段，而不必等待那个报文段设置的重传计时器超时。

冗余ACK：TCP规定每当比期望序号大的失序报文段到达时，就发送一个冗余ACK

CUBIC：算法，三次函数介导窗口调整算法

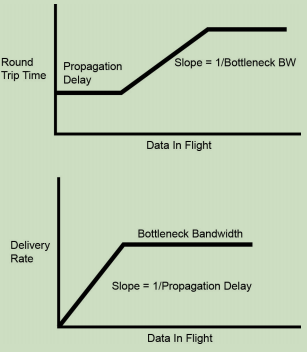
注意time从last congestion event开始，并非一直增长（对TCP而言增长速度过快）



BBR（congestion control based on bottleneck bandwidth）

BBR算法的核心是找到最大带宽（Max BW）和最小延时（Min RTT）这两个参数，最大带宽和最小延时的乘积可以得到BDP（Bandwidth Delay Product）, 而BDP就是网络链路中可以存放数据的最大容量

RTT先保持不变，增大data in flight（尝试增加数据的传输），寻找BDP --- 链路可存放的数据最大容量



Long fat network（长肥管道）

TCP独有，UDP没有

时延带宽积很大的网络

传输时延（发送时延）很小  
 收发包速度很快，非常短的时间就能把大量的数据发送到网络上。

传播时延很大  
 数据包从发送到网络上开始，要给过很长的时才能传送到接收端（距离长）

QUIC

Quick UDP Internet Connection)

基于UDP的传输协议，实现了TCP + HTTPS + HTTP/2的功能

以一个64位的随机数作为ID来标识，即使IP或者端口发生变化时，只要ID不变，这条连接依然维持着，也就不需要重连

DNS（domain name system） 域名系统（数据包使用UDP形式）

根域名服务器（知晓所有必要的IP域名对映）、顶级域名服务器、授权域名服务器、本地域名服务器

域名解析

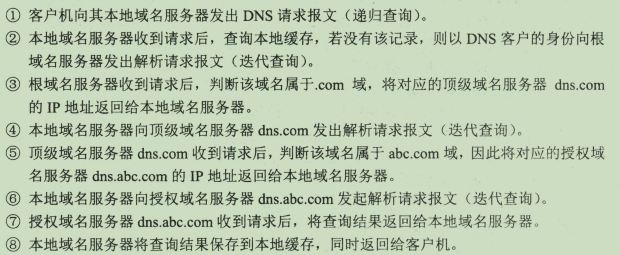
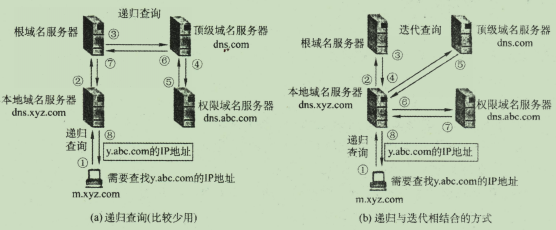
递归查询

本地域名服务器只需向根域名服务器查询一次，后面的几次查询都是递归地在其他几个域名服务器之间进行

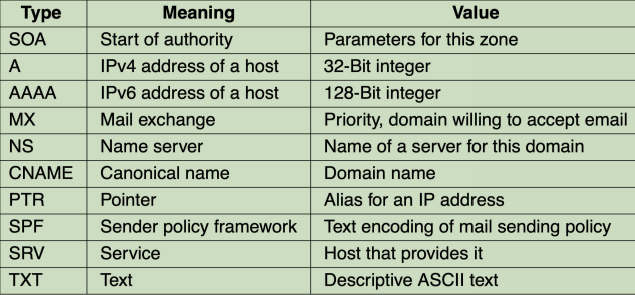
递归与迭代相结合

主机向本地域名服务器的查询采用的是递归查询

本地域名服务器向根域名服务器的查询采用迭代查询



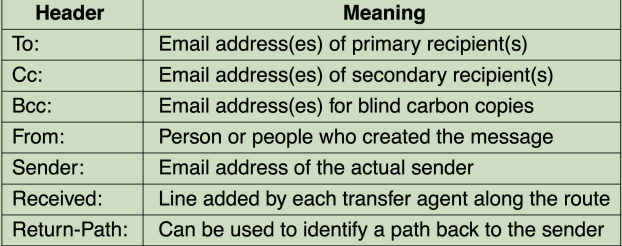
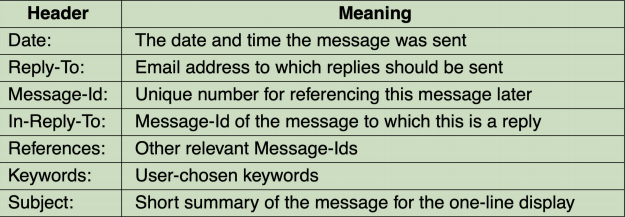
DNS的请求与响应



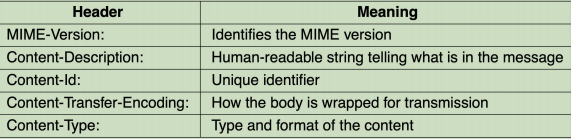
Message format

RFC5322格式header

用户自由撰写



MIME header

MIME支持非ASCII码的写入，会将内容转换为ASCII码



SMTP（基于TCP连接，25端口）simple mail transfer protocol（简单邮件传输协议）

用户代理与邮件服务器之间的交互

只能传送长度一定的ASCII码

POP3（post office protocol邮局协议）

使用TCP协议，端口为110

可选“下载并保留”与“下载并删除”两种方式删除邮件

IMAP（the Internet Message Access Protocol因特网报文存取协议）

使用端口143，默认基于TCP

允用户代理只获取报文的某些部分

用户浏览器与Hotmail或Gmail的邮件服务器之间的邮件发送或接收使用的是HTTP，而仅在不同邮件服务器之间传送邮件时才使用SMTP

The world wide web（万维网）

并由一个全域“统一资源定位符”（URL）标识

b066bb3b394b0df4748cdae30a7af57超文本传输协议（HTTP）应用层协议，它使用TCP连接进行可靠的传输

使用基于TCP协议的80端口

HTTP本身是无连接的（不考虑数据传输时丢失的重传）

Cookie，记录浏览状态

非持久连接：一个元素需要单独建立一个TCP连接

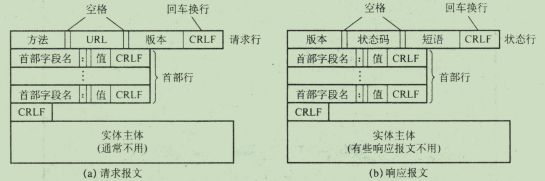
持久连接：万维网服务器在发送响应后仍然保持这条连接（可多次发送请求与响应报文）

流水线/非流水线：连续发送/在收到前一个响应后才能发出下一个请求

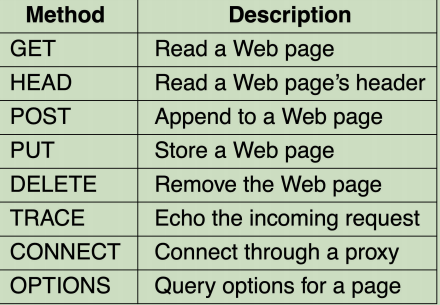
默认使用流水线的持久连接

A multi-threaded Web server with **a front end** and processing modules

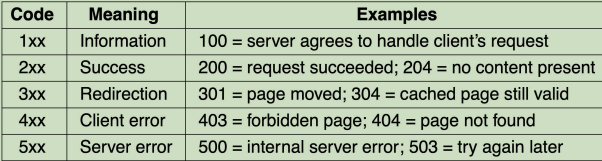
Structure

HTTP/1.1 每次请求一个元素（page确认与style sheet信息均需要单独询问）

HTTP/2 可以同时请求多个元素（同时返回page确认 + style sheet信息）

HTTP method类型

状态码



Static web object（静态对象）

适合cache（每次取用都不会发生变化）

Dynamic pages

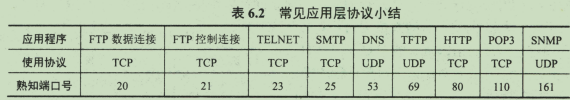
Server-side scripting with PHP（服务器端计算）

JSP（java server page）、c#、python、node.js、ASP

Client-side scripting with JavaScript（客户端端计算）

AAivx、Applex、vb

超文本标记语言（HTML）文档结构的标记语言，它使用一些约定的标记对页面上的各种信息（包括文字、声音、图像、视频等）、格式进行描述

常见的应用层协议小结

Digital video

MPEG frames三种类型图像

I帧（Intra-Frame）是帧内压缩，不使用运动补偿，由于I帧不依赖于其他帧，是解码中的基准帧。

P帧（Predicated-Frame）根据前面的I帧或P帧进行预测，使用运动补偿算法进行压缩，压缩比比I帧高，P帧是对前后的B帧和后继的P帧进行解码的基准帧。P帧本身是有误差的，如果P帧的前一个基准帧也是P帧，就会造成误差传播。

B帧（Bidirectinal-Frame）是基于内插重建的帧，它基于前后的两个I、P帧或P、P帧，它使用双向预测，B帧本身不作为基准，因此可以在提供更高的压缩比的情况下不传播误差。

流文件，在web上点击时只是返回meta file文件（含有URL），仅描述如何获得相应的流文件；web browser也不会处理这个meta file文件，而是直接丢个本地的player media，让其与media server建立连接（本地建立buffer），获取资源（片段传输，要求视频流音频流同步，不能整体先后存放）

好像在request每一segment前，会measure bandwidth，根据当前bandwidth决定request segment的大小

A media player has five major jobs to do:

Manage the user interface（控制用户界面）

Handle transmission errors（处理传输错误）

Decompress the content（解压内容）

Eliminate jitter（消除抖动）

Decrypt the file（解密文件）

Voice over IP

目录服务的查找与连接（可以借助代理）

H.323

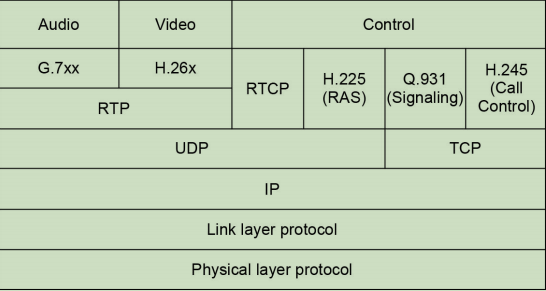
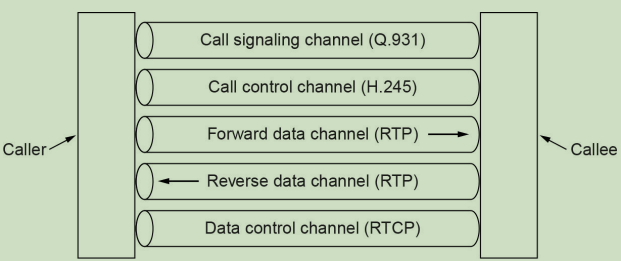
严谨，官方，使用二进制

Protocol stack，控制只能使用TCP

G：音频协议

H：视频协议

建立不同的管道

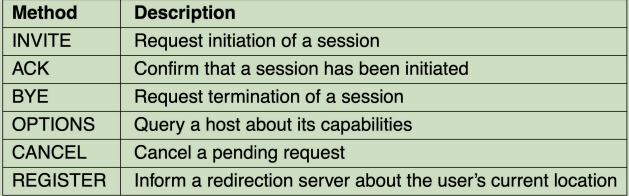


SIP（the session initiation protocol）

简单，分头部与体部，头部类似于HTTP协议头部，基本使用文本

区别于H.323，仅修改了控制部分，兼容TCP与UDP

含有method（类似于HTTP使用）

Callee需要向location server注册（register）

Invite/ACK要协商开放端口以及使用的压缩格式等等

代理只介导连接，实际传输data是点到点的

齐夫定律（Zipf distribution）：反比关系

一小部分内容吸引绝大部分流量（？）

Server farm

Front end需要balance load，一个router去介导多台server（分担load）运行不同client的请求（可能查询同一个back-end database）

Clients可能只知道一个相同的IP，所以需要在serve内部进行load分担（连接的router需要进行IP转换）

Peer-to-peer network（对等网络）

Interested，别的peer有我需要的数据

not interested，别的peer有的数据我也有

Choked，a pair of peer之间不能交换数据

Unchoked peer，未处于“惩罚”状态