CS305 Final Review

Transport Layer(Continue...)

TCP

1. TCP views data as **an unstructured, but ordered, stream of bytes rather than pkt in rdt**
2. ACK # is the next byte expected from other side
3. ★这里有可能考计算题！！！如何计算给出比较好的超时时间呢？太短了会造成没必要的重传，太长了会造成对segment loss的slow reaction。所以总体来说就根据上一次的时间加权算出estimatedRTT，然后再结合DevRTT(safety margin)根据公式算出Timeout Interval
4. $EstimatedRTT = (1-\alpha) * EstimatedRTT + \alpha * SampleRTT$, where α typically equals 0.125, 然后通过他们来计算真正的interval
5. $TimeoutInterval = EstimatedRTT + 4 \times DevRTT$ where $DevRTT$ 计算公式是 $DevRTT = (1-\beta) * DevRTT + \beta * |SampleRTT - EstimatedRTT| (typically, \beta = 0.25)$
6. fast retransmission: 如果发送方连续收到 **3 个重复的 ACK**（即总计 4 个相同的 ACK），它会认为某个数据包丢失了，立即进行快速重传，而不需要等待超时。

Flow Control 流量控制

Receiver $rwnd = RcvBuffer - [LastByteRcvd - LastByteRead]$ 相当于就是receiver告诉发送端他还剩多少buffer空间，要不然，以此来控制发送速度，所有发送端有 $LastByteSent - LastByteAcked \leq rwnd$

TCP的三次握手

1. 客户端发送一个SYN,以及seq = x
2. 服务器发送SYN+ACK, seq = y, 确认收到了客户端的SYN，所以ACKnum = x + 1(32 bits), ACKbit = 1
3. 客户端发送ACK, ACK = y + 1, SYN = 0

TCP closing connection(2个RTT)

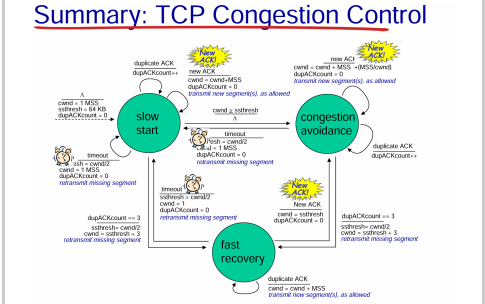
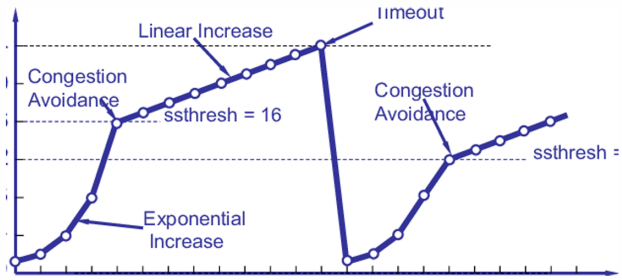
1. client : FIN bit = 1, seq = x
2. server: ACK bit = 1, ACK num = x+1(这时候还能发送消息)
3. server: FIN bit = 1, seq = y(这时候就不能发送了)
4. client: ACK bit = 1, ACK num = y+1

   Reset segment: when source IP address do not match with any of the ongoing sockets. Then the host will send a special reset segment to the source. RST flag bit is set to 1.

Congestion Control 拥塞控制

1. $\lambda_{in}^{'} \geq \lambda_{in}$, original data + retransmission data.
2. Cause :Shared link; limited link capacity ;Sending at a high rate
3. Cost of Congestion:• Delay • Packet lost and retransmission • Unneeded retransmission: waste • “upstream” transmission capacity was wasted

4. Approaches dealing with congestion control
    - TCP segment loss or round-trip segment delay & TCP decreases its window size accordingly(End-to-End, which TCP use)
    - routers provide feedback to the sender and/or receiver & a single bit indicating congestion at a link; the maximum host sending rate the router can support(Network-assisted)
    - $LastByteSent - LastByteAcked \leq cwnd, rate \approx \frac{cwnd}{RTT} bytes/sec$
5. Congestion control algorithm has three major components:
    - **Slow start**: exponentially increase (initiate CWND, and double it every RTT)
    - **Congestion avoidance**: linearly increase
    - **Fast recovery**
    - 对Loss的反应：如果是timeout 重置，开始SS阶段；如果是tri-duplicate ACK , TCP Tahoe和前面一样, TCP RENO 就进入快速恢复阶段
    - (SS阶段)刚开始指数级增长到threshold(每个RTT*2，相当于每个ACK增加一个MSS)，然后开始linear增长(when CWND gets to 1/2 of its value before timeout Thus，**when timeout occurs，ssthresh = cwnd/2**)
    - 当cwnd>=ssthresh时候，开始以(MSS/cwnd) MSS的速度增长**每个新的ACK**（进入CA阶段）相当于**每个RTT增加一个MSS**，ssthresh是动态调整的，当发生丢包时候一般他会设为1/2 cwnd
    - 在慢启动、CA阶段如果收到tri-dup ACK, 就Trigger (RENO)：Fast Recovery, triple duplicate ACK ; ssthresh = cwnd/2, cwnd = ssthresh + 3MSS（相当于每个duplicate增加一个MSS）
    - 当CA和FR阶段出现了TIMEOUT，那么就cwnd = 1MSS, ssthresh = cwnd/2重置到SS阶段
    - RENO算法于Tahoe的区别就是对timeout一个是直接归零，一个是快速恢复
    - 这里应该会考一个计算★！！

<table>
<tr>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</table>

6. AIMD(additive increase multiplicative decrease) increase 1个MSS per RTT，detect loss就减半
7. Explicit Congestion Notification (ECN): Two bits in IP header marked by network router to indicate congestion

Network Layer (Data plane)

Network Layer 分为 Data plane & Control plane

网络层两个关键的函数: forwarding(data plane), routing(control plane)

Data plane: local, per-route function, hardware

Control plane: network-wide logic, software

Router

1. Input ports

   physical layer -> link layer protocol -> lookup forwarding, queueing -> switch fabric...

   用最长前缀longest prefix来决定目标地址，有冲突的话一般把冲突的最后一位改成1

2. Switch fabric

   Three types of switching: via memory(一次一个packet,不能同时forward); via bus(broadcast); via crossbar(multiple pkt can forward in parallel)

3. Output ports

   buffering -> link layer protocol -> physical layer

   scheduling discipline chooses among queued datagrams for transmission(using Priority scheduling – who gets best performance, network neutrality)

4. Queueing

   - buffering when arrival rate via switch exceeds output line speed AND **queueing (delay) and loss due to output port buffer overflow!**
   - recommend buffer is $\frac{RTT*C}{\sqrt{N}}$
   - FIFO & Discard policy(tail drop, priority, random), priority queuing are as following:
     - RR(Round Robin scheduling): multiple classes AND cyclically scan class queues, sending one complete packet from each class (if available) 平等对待每一个数据流
     - Weighted Fair Queuing (WFQ): generalized RR 加权公平性，考虑流的权重和数据量来决定发送顺序

IP(Internet Protocol)

1. IP datagram format

   20 bytes TCP + 20 bytes IP = 40 bytes + app layer overhead

   MTU (max transmission unit) divides large byte diagram. (fragmented)

   IP分片注意IP本身头部的20 bytes开销

2. IPv4 addressing

   IP 头部20-60 bytes可变

   32-bits identifier for interface of hosts and routers

   Subnet part - host part. Hosts and routers within the subnet can reach each other **without the help of intervening router**.

3. **CIDR (Classless InterDomain Routing)**

   a.b.c.d/x, where x is # bits in subnet portion of address(地址分配应该会考一个计算题★！！)，注意地址不能有冲突

ICANN: Internet Corporation for Assigned Names and Numbers

4. **DHCP: Dynamic Host Configuration Protocol** dynamically get address from as server E.X. host send **discover** msg -> DHCP send server **offer** msg -> host send **request** msg -> server send **ack** msg

   what DHCP also returns is as following:

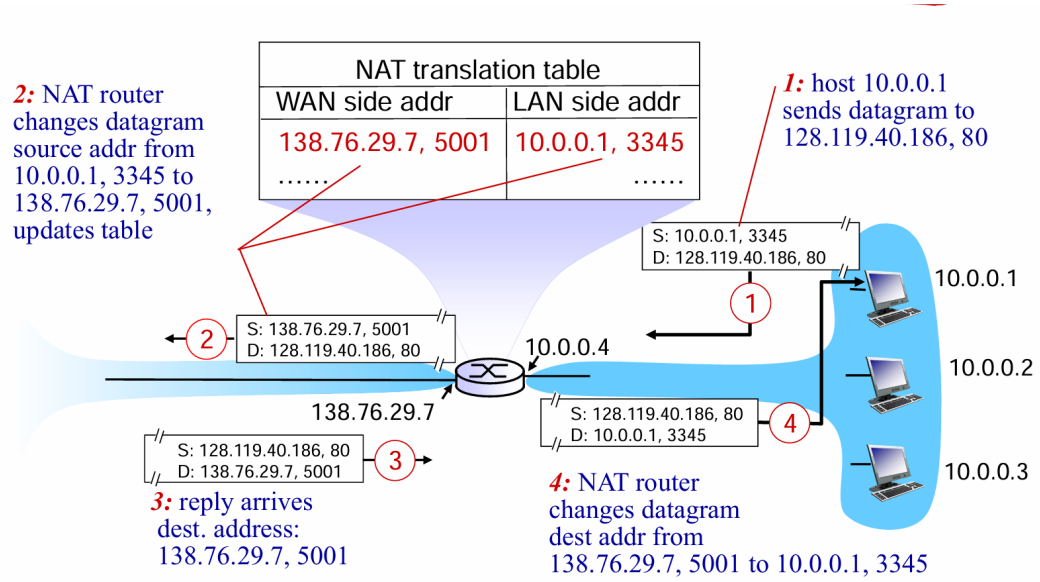
   address of first-hop router for client AND name and IP address of DNS sever AND network mask (indicating network versus host portion of address)

5. **Network address translation(NAT)**

   NAT 的motivation是一个LAN仅使用一个IP，便于更改本地的IP(不用告诉外面世界)，更换ISP不影响本地网络设备，本地网络设备对外部不可见提高安全性。

   outgoing datagrams(replace info), remember(in NAT translation table) 记录的是WAN方的ip, port mapping的 LAN方的ip, port(如下图所示)

   在翻译过程中 NAT router 接收到局域网中用户的IP，在translate table过程中把本身的src IP改成自己的IP地址以及端口，收到最终dst消息后再按照pair返回给局域网内对应的client IP 和 port



6. IPv6 addressing

   128 address space (motivation 是 32 bits 很快就completely allocated); fix-length 40 byte header; no fragmentation allowed(超过了就会发送"packets too big" 的error，motivation是减轻路由器负担，因为分片一般是路由器做的事，但是分片的任务转移到了发送端了); hop limit(每次经过一个router就-1, reach zero datagram->discard)

   相对于IPv4 没有checksum

   tunneling: IPv6 datagram carried as payload in IPv4 datagram among IPv4 routers; 即IPv6在IPv4中作为payload存在, ipv4的tunnel连接ipv6的routers

Generalized forward and **SDN(software-defined networking)**

1. Generalized forwarding: pattern + action(forward packet to port, encapsulate and forward to controller, drop, modify fields) + priority + counter(#packets, time last updated)
2. Firewall : match IP addresses and TCP/UDP port numbers AND action is permit or deny.
3. Load balancing: 采用分流，同样的dst子网，但是从不同的inference出去

- *match+action:* unifies different kinds of devices

- Router
  - *match:* longest destination IP prefix
  - *action:* forward out a link
- Switch
  - *match:* destination MAC address
  - *action:* forward
- Firewall
  - *match*: IP addresses and TCP/UDP port numbers
  - *action:* permit or deny
- NAT
  - *match:* IP address and port
  - *action:* rewrite address and port

Network Layer (Control plane)

1. Two approaches to structuring network control plane
   - per-route control (traditional)
   - logically centralized control (software defined networking)
2. routing protocols

   link state(global): Dij algorithm to find shortest path. 然而如果考虑congestion就会有oscillation

   distance vector(decentralized): distributed & iterative & asynchronous

   1. BF algorithm 公式就是$d_x(y) = min\{c(x,v) + d_v(y)\}$
   2. DV algorithm: 分步执行，当表有变化时候通知相邻节点并且更新然后重新计算，记住他节点只能更新它对应的那一行，然后如果出现变化要更新完整！！

      Good news travel fast & Bad news travel slow.

      (这里应该会有一个计算题★！！) 用poisoned reverse方式来解决，即善意欺骗相邻节点(THAT IS If Z routes through Y to get to X : Z tells Y its (Z's) distance to X is infinite (so Y won't route to X via Z))

      Poison reverse 并不能完全解决count-to-infinity问题，当loop增加到3个或者更多node时候会失效
   3. 一个比较

Comparison of LS and DV algorithms

message complexity

- ***LS:*** with n nodes, E links, O(nE) msgs sent
- ***DV:*** exchange between neighbors only
  - convergence time varies

speed of convergence

- ***LS:*** $O(n^2)$ algorithm requires O(nE) msgs
- ***DV:*** convergence time varies
  - may be routing loops
  - count-to-infinity problem

robustness: what happens if router malfunctions?

LS:

- node can advertise incorrect *link* cost
- each node computes only its *own* table

DV:

- DV node can advertise incorrect *path* cost
- each node's table used by others
  - error propagate thru network

3. intra-AS(autonomous system) routing in the Internet: OSPF

   特性：scale, autonomy

   AS also known as domains divided into **Gateway router** and **Interior router**

   Intra-AS(domain 内部自治系统路由) and Inter-AS(自治系统之间的路由协议) 共同贡献routing table，不同的destination用不同协议传输

   ★一些intra-AS protocol IGP(interior gateway protocols):

   1. RIP(Routing information Protocol): DV algorithm
   2. OSPF(open shortest path first): link-state based
   3. IS-IS same as OSPF
   4. IGRP(interior gateway routing protocols)

   OSPF Router floods OSPF link-state advertisements to all other routers in entire AS, 直接通过IP传输而不是UDP,TCP，消息传输更加可靠，some features are as following

   1. security
   2. multiple same-cost paths (only one path in RIP)
   3. support **multi-cast** and integrated uni-cast
   4. hierarchical (backbone and local area), routers(area-border, backbone, gateway)

4. Inter-AS routing: BGP(border gateway protocol)

   Distance-vector

   向外宣传内部信息以及获得别人信息，eBGP获得别的AS的访问信息，iBGP告诉内部router reachability信息

   BGP-peer基于TCP传输

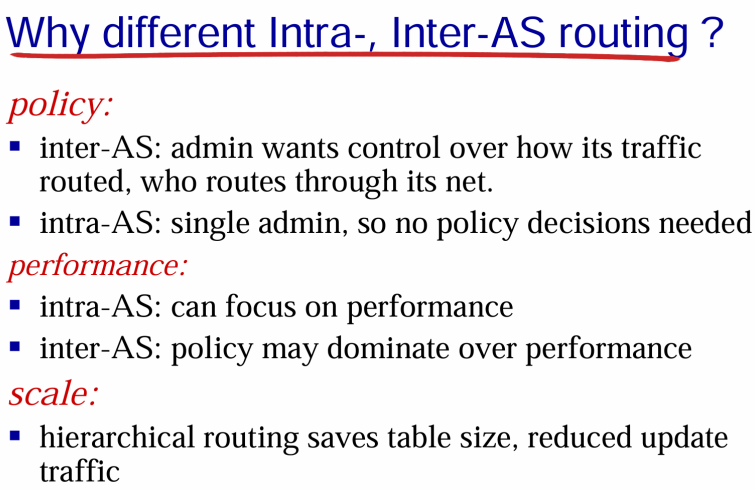
   Prefix(destination) + attributes = "router" 两个重要的attribute (AS-PATH and NEXT-HOP)，这里可能会有一些填空题，比如下一跳的信息从哪里来用什么协议

   Route-selection，如果收到了去同一个地点有很多条路径，应该怎么选择 (Hot-potato Routing) 即选择**least intra-domain cost**，不用care inter-domain-cost.

   Next-HOP是收到宣传路线的起点路由器

IP-Anycast CDN公司给每个server分配同样的IP然后宣传，当另一个BGP-router收到了两个线路的选择时候locally用route-selection algorithm来选择最好的路线

Routing-policy, 当一个ISP连接了两个backbone，一个backbone不会选择将他的路线给另一个，这样一个ISP就不用carry别的route-path过境的流量



5. SDN (software defined networking)

   远程控制器来管理计算routing table(原因？easier management, open implementation of control plane, programmable)

   1. network-control application(control plane), 一些可以通过外部implement的应用
   2. control-plane OS(SDN controller)通过southbound API与switches interact，通过northbound API与 application interact. (三个layer，interface layer, network-wide state management layer, communication layer)

      一个例子，OpenFlow controller 用TCP 实现controller-to-switch以及switch-to-controller通信

      controller-to-switch: configure, modify-state(OpenFlow tables), read-state, packet-out

      switch-to-controller: packet-in, flow-removed, port-status

      Two important differences from the earlier per-router-control scenario:

      1. Dijkstra’s algorithm is executed as a separate application, outside of the packet switches.
      2. Packet switches send link updates to the SDN controller and not to each other.
   3. Data-plane交换更加迅速

6. ICMP (Internet Control Message Protocol)

   Network-layer above IP ICMP消息 carry IP消息

   Type + code + header + first 8 bytes of IP datagram causing error

   停止条件：达到dst, 当回传type3 code3时候 , source停止传输

Link Layer

1. Introduction

   transportation analogy

Link-layer services: framing (encapsulate datagram), reliable delivery, error detection, error correction, half and full-duplex

Link-layer implemented in "adaptor" (network interface card, NIC)

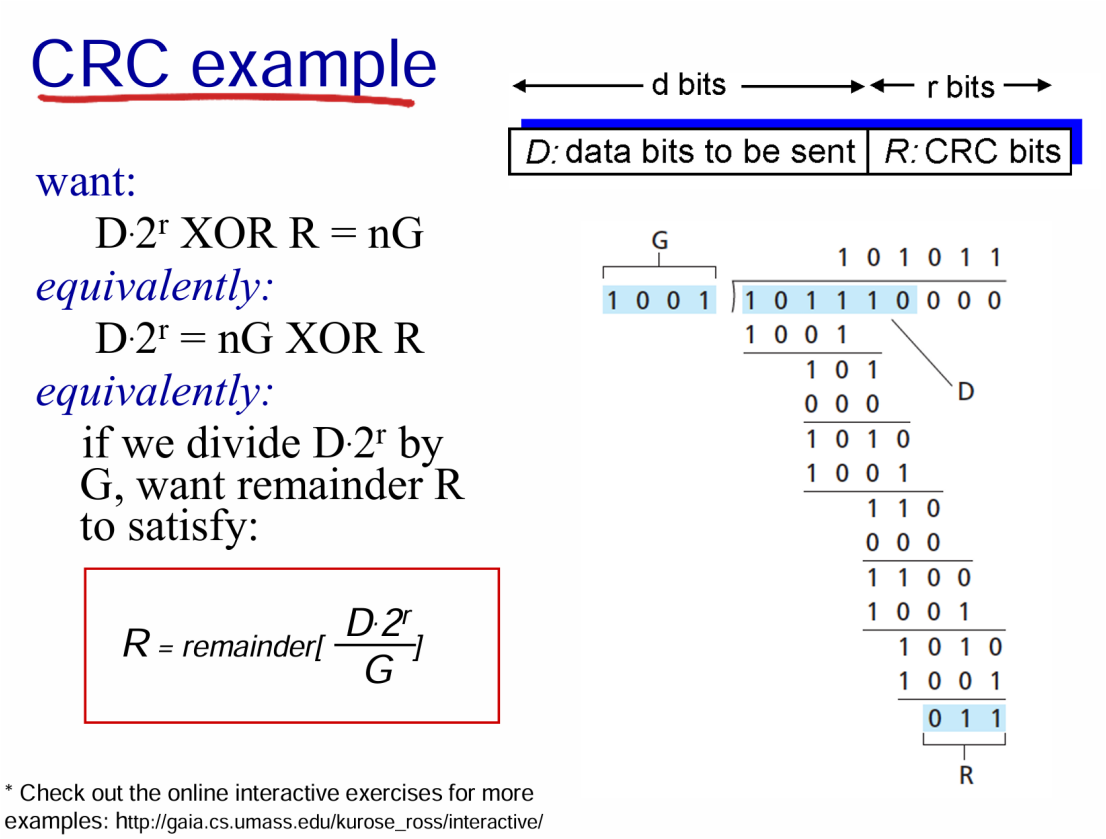
sending host(encapsulate datagram and add error checking bits) and receiving host(look for error and extract datagram)

2. Error detection and correction (EDC)

   Parity checking(single, two-dimensional 相当于有column parity bit 和 row parity bit 通过行列的交叉可以得到wrong bit)

   Cyclic redundancy check $D * 2^r XORR = nG$ with <D,R> d bits data bits and r bits CRC bits.

   按照图示红圈的公式即可得到是否出错(G是提前约定好的) 计算时候相当于作XOR



* Check out the online interactive exercises for more examples: http://gaia.cs.umass.edu/kurose_ross/interactive/

3. Multiple access protocols

   two types of link(point to point(PPP) and broadcast)

   Medium Access Control(MAC) protocols:

   1. channel partitioning 基于time, frequency...进行分割
      - TDMA: time division multiple access, each station get fix length slot (length = packet trans time)
      - FDMA: frequency division multiple access, each station assigned fixed frequency band
      - Limitations: 如果当只有一个node在传输时候他不能用整个带宽R，只能用fixed R/M
   2. random access(dynamic)

      相对于上一个，如果只有一个node传输，是可以利用整个带宽的。

      - Slotted ALOHA, synchronization 当node到达一个新的frame如果没有冲突就发送，如果有冲突就以p的概率在接下来的slot进行发送，any nodes that can success $Np(1-p)^{N-1}$, max efficiency is 37%$(1/e)$

- Pure(unslotted) ALOHA, non-synchronization, frame第一次到达立刻传输 collision pro increases. $p*(1-p)^{2(N-1)}$. max efficiency is 18%($1/2e$). worse than slotted ALOHA
- **CSMA(carrier sense multiple access)**: 思想是先listen看是不是idle，如果是直接传输，不是则defer传输. 潜在的collision就是由于propagation delay 两个node可能听不到对方
- CSMA+CD(collision detection) in **Ethernet** collision detected within short time. Colliding transmission are aborted, reducing channel wastage 相当于在传输过程中也在检测冲突，如果有冲突就发一个jam signal， 通过**Binary Exponential Backoff** 算法可以减少多次冲突后重复发生的概率
- CSMA+CA(collision avoidance) used in **IEEE802.11**, CSMA/CA 侧重于通过提前的协调和避免冲突来减少冲突的概率，而不是像 CSMA/CD 那样在发生冲突后进行检测和恢复。
- Limitations: high load: collision overhead

3. taking turns

   target to实现一个node可以send at rate R, 同时传输也可以实现send at rate R/M

   - polling. 引入master node "invites" slave nodes to transmit in turn.
   - Token passing. control token passed from one node to next sequentially

4. Case study

   DOCSIS: data over cable service interface specification

4. LANs

1. Mac addresses

   - 与adapter相关，而不是host或者router
   - link-layer switches没有mac地址
   - each adapter在LAN下有唯一的mac address
   - MAC 地址是唯一的，就像ID
   - 一个adapter发送一个包含destination address 的 datagram到LAN，如果有match就读取，没有就删除

2. ARP(address resolution protocol)

   IP -> MAC(Resolve addresses only for interfaces on the **same subnet**) each IP node has a table mapping <IP, MAC, TTL>

   当一个A想要和B通信时候发出ARP报文包含B的IP，B 收到之后返回他的MAC，然后A存储在他的ARP table，相当于cache(有一个TTL)

   Q：How to send a datagram from one host to another？

   - Same subnet: framing with destination MAC; send it
   - Different subnets:

     假设A,B处于不同subnet，中间R连接。A创建一个link-layer frame包含A->B datagram, 当A发给R他的src和dst(这里应该是R的mac)，R会remove datagram,上升到IP层,然后forward. R创建link-layer frame包含A->B ip datagram, 但是在mac地址部分改成自己(R)的mac和dst mac,

3. Ethernet

   widely use LAN technology

   bus, star(active switch in center)

   frame structure

   - addresses, 如果收到的mac address匹配就发送给IP层处理

- type, indicate higher layer protocol
- CRC cyclic redundancy check
- connectionless, unreliable, protocol is unslotted CSMA/CD with binary backoff

4. Ethernet switch
    1. Link-layer device: store and forward Ethernet frames; selectively forward

       switch 有 switch table存储 <HOST MAC, interface, time stamp>
    2. transparent: host are unaware of presence of switches
    3. plug-and-play, self-learning

       switch 学习 sender location
    4. 流程大概是，switch 接到表中没有的dst, 那么他就给除了src interface的所有接口都发送一个copied消息，然后收到某一个dst回复之后就可以记录他的dst对应的interface，然后以后就可以针对性只发给一个link
    5. 一些特点: elimination of collision(never trans more than one frame at a time), heterogenous link, management
    6. switches VS routers

       都有table以及store-and-forward

       switches link-layer, plug-and-play, broadcast storm, spanning tree, self study

       routers network-layer, not plug-and-play, choose best path, firewall protection, compute tables using routing algorithms
5. VLANS

   Port-based VLANs

    1. traffic isolation
    2. dynamic membership
    3. forwarding between VLANs

   spanning multiple switches

   trunk port 交换不同VLANs之间的信息，而且必须carry VLAN ID info

5. Data center networking

   好像没讲什么？

6. a day in the life of a web request

   ★一个复习！

Wireless and Mobile Networks

1. Wireless network
    1. wireless hosts(laptop, ...)
    2. base station
    3. wireless link
    4. infrastructure mode 有基站设施
    5. ad hoc mode 无线设备直接通信

2. Wireless link characteristics

   import differences to wired link: decreased signal length, interference from other sources, multi-path propagation

   SNR(signal-to-noise ratio) larger SNR – easier to extract signal from noise (a “good thing”)

   SNR versus **BER(bit error rate)** 负相关 带宽越高，达到同样的BER，需要更高的SNR

   有一个hidden terminal problem，两个无法直接沟通的host无法意识到他们有一个可以监听到他们两个的中间点

   CDMA(code division multiple access) unique code assigned to each user, encoded signal = original data * chipping sequence, decoding, 通过这个方式允许多个用户在同频上传输信号而不干扰，相当于是用特殊的code进行encode然后收到之后解码

3. IEEE 802.11 wireless LAN

   802.11 a/b/g/n均使用CSMA/CA，但不同的工作频率以及数据速率都不相同

   architecture:

   1. wireless host
   2. Basic Service Set(BSS, cell) 基础设施mode比如AP(access point, base station)

   Channels, association

   802.11: passive scanning, 被动接受来自beacon帧其中关于AP的信息 active scanning, 主动发送查询.

   CSMA/CA

   1. 如果idle，直接发送(没有CD)
   2. 如果busy, 那么random backoff time
   3. CA思想是预留channel，他会先发送很短的RTS(request-to-send) to BS using CSMA,然后BS会广播给所有节点他的回复**clear-to-send CTS**，别的node就会取消发送，他就可以发送更长的without collision

   IEEE802.11 addressing

   1. 802.11是receiver AP(host) MAC + transmitting AP(host) MAC + attached AP MAC 相当于AP+H1+R1
   2. Ethernet frame是R1 MAC + H1 MAC
   3. switch也可以通过self-learning来获得AP对应的H1

   Advanced capability

   1. rate adaptation
   2. power management

4. Cellular Internet access(CIA)

   cell (base station BS, mobile users, air-interface)

   MSC(mobile switching center): connect, manage cell, handle mobility

   two techniques: combined FDMA/TDMA, CDMA

   2G(voice) base transceiver station(BTS), Base station controller (BSC), Mobile Switching Center (MSC), Mobile subscribers

   3G(voice+data data operates in parallel) Add Serving GPRS Support Node (SGSN)+Gateway GPRS Support Node (GGSN) AND **GPRS**（General Packet Radio Service）

4G-LTE difference to 3G: all IP core(tunneled), no separation of voice and data

5. Mobility

   home network, **permanent address**(address in home network), home agent(entity that will perform mobility functions on behalf of mobile, when mobile is remote), visited network(network in which mobile currently resides), **care-of-address**(address in visited network), correspondent(wants to communicate with mobile), foreign agent(entity that in visited network that performs mobility function on behalf of mobile)

   Approaches: routing(not scalable to mills of mobiles), 下面是end-system的处理方式:

   1. Indirect-routing mobile使用care-of-address和permanent-address两个地址，care-of-address是连接访客网络时候foreign agent分配的，home agent可以通过这个care-of-address给设备发送数据因为即使当前他在别的network下面

      所以mobile到别的网络的步骤是：register to foreign agent, foreign agents register to home agent, home agent update care-of-address of device, packets can forward

   2. Direct-routing 解决了triangle routing problem. non-transparent to correspondent相当于直接和对应network通信，但是也必须询问home agent 要care-of-address

   Anchor FA(FA in first visited network), data总是先流向AFA, 所以new FA总是获得来自old FA的数据