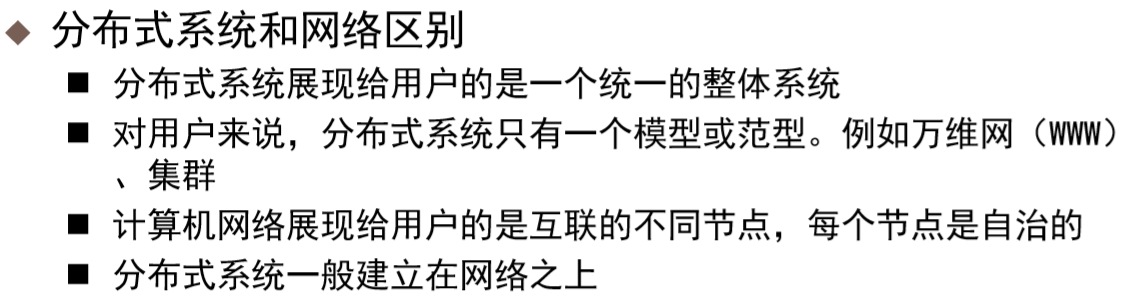
# Ch1 网络基础

## 体系结构

1. **网络基本概念**

计算机网络：互联的通用计算机的集合。



网络 = 节点 + 链路

节点（Node）：

* 端节点：自身拥有计算机资源的源宿用户设备；
* 转接节点：支持网络连通并对数据起交换和转接的节点，如交换机/路由器/集线器。

链路（Link）：

* 物理链路：物理层连接两个节点的物理介质，如电话线、同轴电缆 、光纤、无线电波。
* 逻辑链路：在两点间通过通信协议的作用建立起来的数据联结通路。

通路(Path): 从源点到宿点所经过的一串节点和链路的有序集，或端到端的通路。

协议（Protocol）：多个进程为完成一个任务而共同遵守的动作序列规范。三要素：语法、语义、规则(同步时序)。

网络云：区分通信子网交换交换结点（云内）和资源主机结点（云外）的分界线，可表示任何网络（单、多、交换等）。

IETF是各种工作组提出RFC，是互联网工程师主要活动组织。

IANA：早期互联网编号分配机构。

ICANN：互联网名称与数字地址分配机构，民间性的非盈利公司。

**网络体系结构**

网络体系结构的元素就是协议（Protocols），Protocols之间最重要的关系是分层（Layering）。

OSI（开放式系统互联）七层模型：物理层,数据链路层,网络层,传输层,会话层,表示层,应用层。

TCP/IP四层模型（细腰模型）：数据链路层,网络层,传输层,应用层。

应用编程接口API：

* 套接字（socket）是一个抽象层，允许应用程序将I/O插入到网络中，并与网络中的其他应用程序进行通信。
* 网络套接字是IP地址与端口的组合。
* 套接字接口定义的操作：创建套接字、将套接字连接到网络、通过套接字发送/接受消息及关闭套接字。

1. **性能参数（带宽和延迟）**
2. 带宽(Bandwidth)：Hz, KHz, MHz, GHz, bps

* 信号带宽：构成信号的各种不同频率成分所占据的频率范围。
* 媒体带宽：通信媒体允许通过的信号频带范围。
* 比特率：某时段内网络上可能传输的比特数，或传输每比特数据所需的时间。
* 比特率越高，高频分量越多，频率范围越大，信号带宽越高。

1. 吞吐率(Throughput)：bps

* 吞吐率：链路上实际每秒传输的比特数。
* 数字信号的发送速率，即发送带宽。

1. 延迟

* Delay：把一个报文从网络一端传输到另一端所需的时间。
* Round Trip Time (RTT)：发收来回时间（2次时延）
* 定义：延迟 = 处理＋排队+传输+传播（主要考虑后两个）：

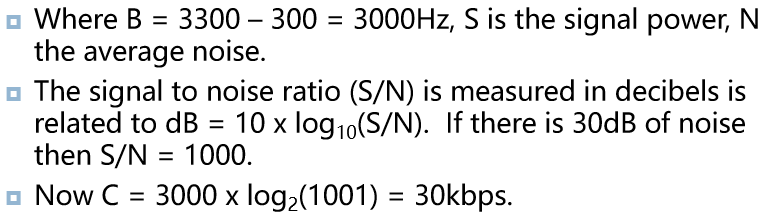
传输时延：路由器把分组发送出去所需时间，数据量/带宽；

传播时延：一个比特从发送路由器到接收路由器传播所需的时间，距离/光速。

* 延迟带宽积（Delay×BW）：信道管道的体积 = 链路上所容纳的比特数

## 1.2 直连网络

香农定理决定了一个链路传输能力上限：



**1.2.2 编码技术**

模拟→数字：PAM（脉冲幅度调制）、PCM（脉冲码调制）

数字→模拟：ASK（幅移键控）、FSK（频移键控）、PSK（相移键控）、QAM（正交幅度编码）

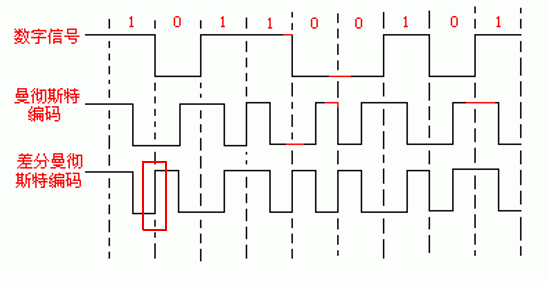
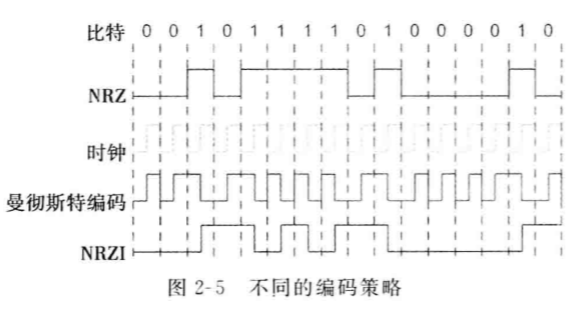
* 波德率（Baud Rate）：每秒内为表示某些比特而需要的信号单元数（或码元数）。
* 



模拟→模拟：AM（幅度调制）、FM（频率调制）、PM（相位调制）

数字→数字：逻辑层编码

* 原因：信号中连续0和连续1造成 ①信号平均值的漂移、②信号时钟的不同步。



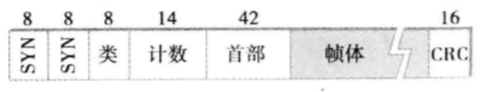
4B/5B编码：先4位变成5位编码，再NRZ-I编码。效率80%＞曼码50%。

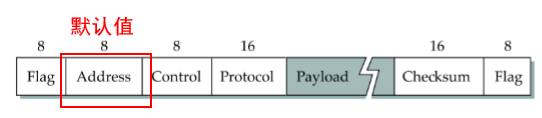
**1.2.3 成帧技术**

Frame是一个在具体网络第二层上实现的、与硬件有关的特殊分组。是网上传输的最小数据 单元。Frame＝数据部分＋发送和接收站点的物理地址＋处理控制部分。

1. 面向字节的协议

BISYNC：IBM，起止标记法、字符填充法

DDCMP：DECNET网络，字节计数法

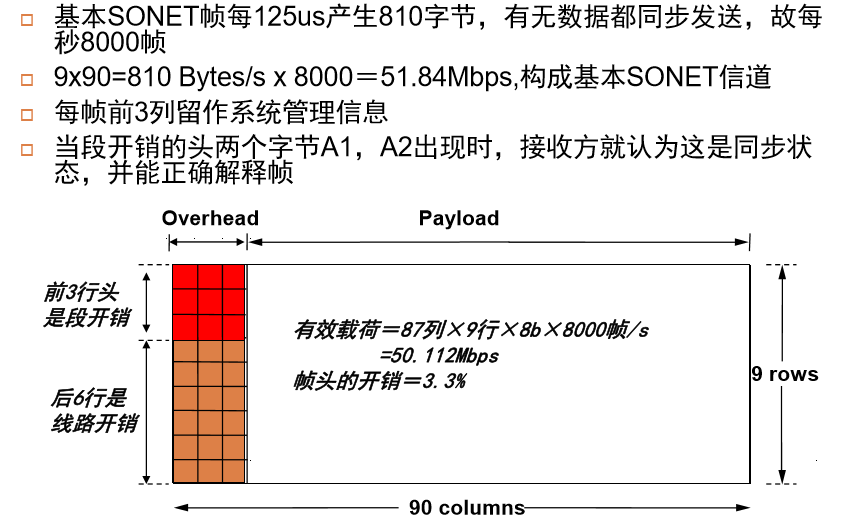
PPP：点对点链路传输IP分组

1. 面向比特的协议

HDLC：比特填充法，5个连续‘1’插‘0’ 

1. 基于时钟的组帧

SONET所有级别都使用字节交叉的多路复用，线路速率都是STS-1=51.84Mbps的整数倍，帧头开销3.3%。



**1.2.4 信道共享技术**

时分复用TDM、统计时分复用STDM、频分复用FDM、波分复用WDM、码分复用CDM

**1.2.5 差错控制技术**

纠错码：奇偶校验、循环冗余校验（CRC）、校验和（Check Sum）

## 1.3 报文交换

**以太网**

优势：可扩展（10M—10G），灵活（多种媒介、全/半双工、共享/交换），便宜、易于安装使用、稳健性好。

接入协议：CSMA/CD（多点接入、载波侦听多路访问/碰撞检测）

网卡功能：数据封装与解封、链路管理（CSMA/CD）、bit的编码与解码

MAC地址：Unicast（单播、网卡）、Broadcast（广播、子网）、Multicast（多播、组地址）

集线器（HUB）：物理层互联设备，冲突域=广播域。

交换机（Switch）：链路层互联设备（一般指的是二层交换机）。

二层交换机：每个端口是一个冲突域，所有端口是一个广播域；

三层交换机：每个端口是一个广播域。

广播域：广播报文可以达到的范围。冲突域：可以发生报文冲突的范围。

虚拟局域网：VLAN划分广播域，每个VLAN一个广播域。

**交换机生成树协议（STP）**

环路：导致转发表的不正确与不稳定，还会导致重复的数据包在网络中传递，引起广播风暴。

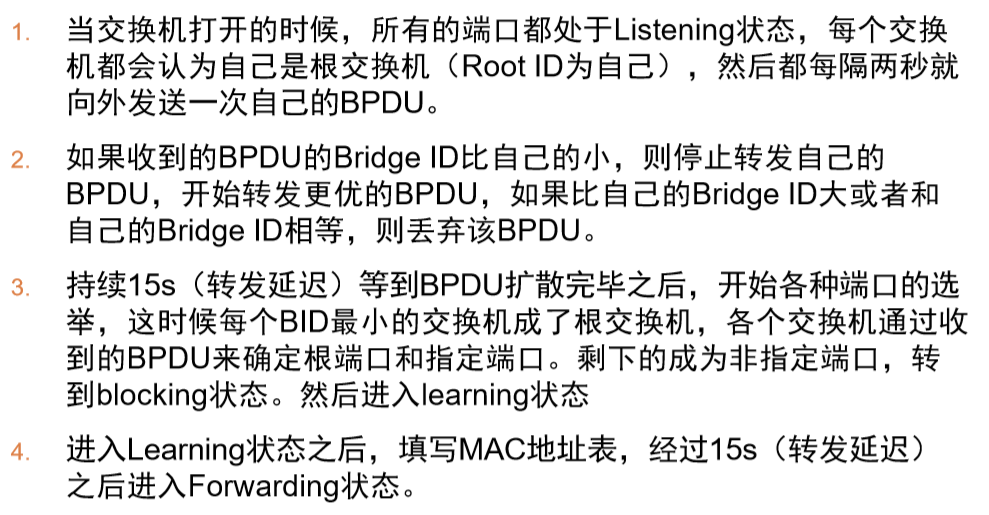
思想：网络中选出一台交换机作为核心交换机（Root），其它交换机则需要留出一条活动链路去往根交换机，因为只要普通交换机到根是通的，到其它交换机也就是通的。

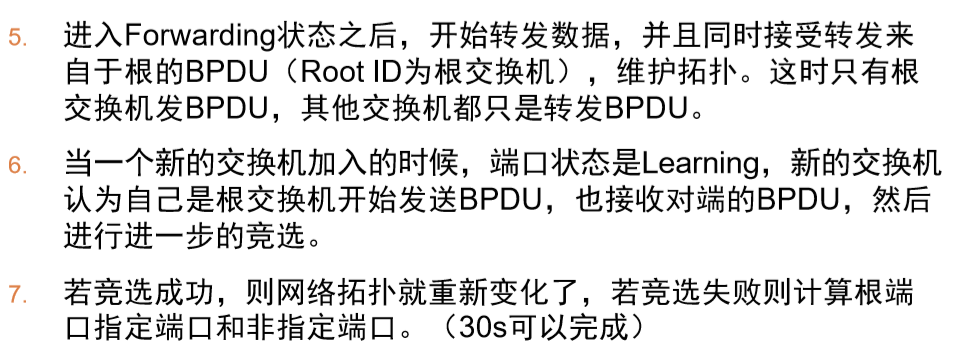
范围：一个广播域内独立选举STP。

步骤：

* 选举根交换机（Root）：Bridge-ID小的→MAC地址小的；
* 非根交换机选举根端口（Root Port）：到根交换机的Path Cost小的→上一跳交换机Bridge-ID小的→对端端口优先级Port ID小的；
* 每条链路选举指定端口(Designated Port)：根交换机上所有的端口都是指定端口，非根交换机同选举根端口；
* 剩余端口状态为Blocking。

选举过程：BPDU报文





第二种类型的BPDU包：TCN BPDU。

## 1.4 TCP/IP网络技术

因特网的三地址：用户识别地址（域名）、网络地址（IP）、物理地址（MAC地址）

DNS（域名解析服务）：资源记录（Name, Value, Type, Class, TTL）

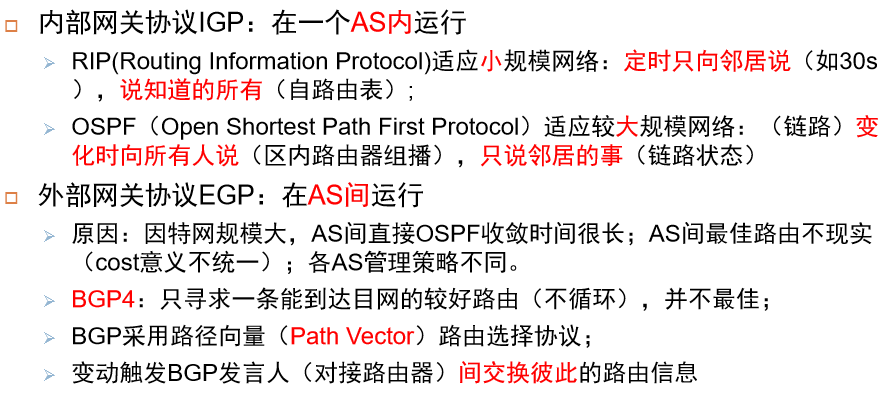
IP报文分片机制：IP头部三个域（ID、MF、Offset）；

不分片：DF域，默认不分片。不断增大报文并设置DF位，直到中间路由器发出错误ICMP（类型3不可达）消息，告诉链路实际的MTU，就可以推断PMTU。

**ICMP协议（差错与控制报文协议）**

* 报告IP传输中发生的（差错报文+控制报文+测试报文），封装在IP数据包中进行传输。
* 典型应用：Ping，traceroute

**路由与寻址**



**距离矢量算法RIP**

**链路状态算法OSPF**

LSP包含内容：创建LSP节点的ID、邻居列表及开销；LSP序号、TTL。

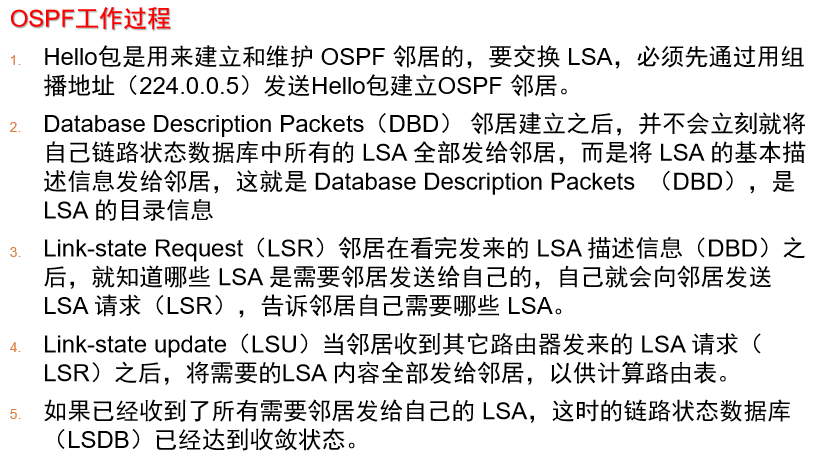
LSP产生的原因：

1）路由器定期产生，由于flooding会影响性能，所以这个周期很长；

2）路由器定期向邻居发hello报文，探测邻居可达性，一旦发现拓扑变化 ，就发出LSP。

路由区（Area）：OSPF区域基于路由器接口划分，IR、ABR。

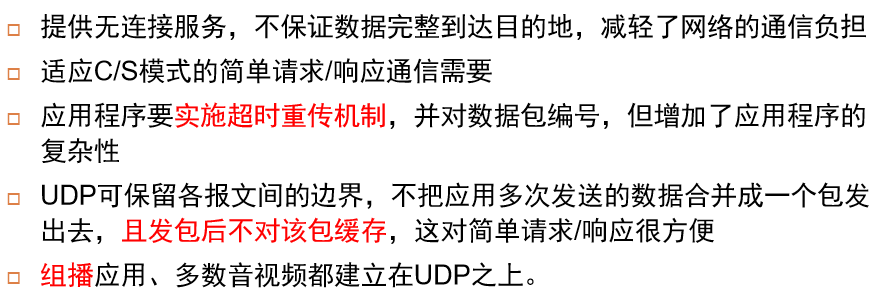
OSPF末节区域（Stub Area）：末梢网络的路由器只要知道去往ABR即可。



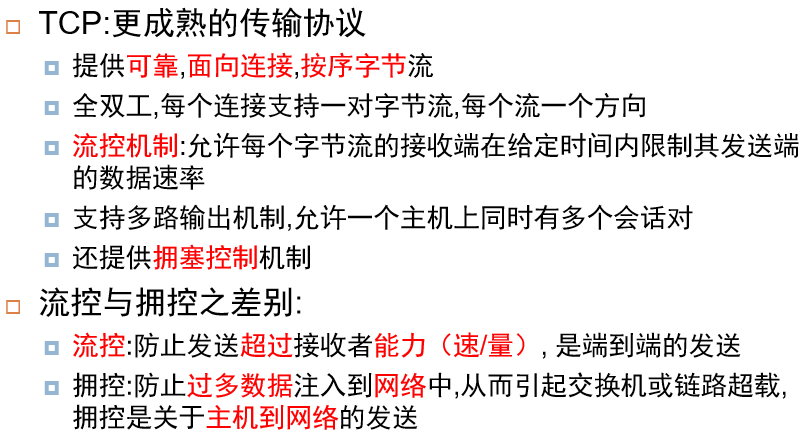
**1.4.4 传输层**

**UDP协议（简单多路分解协议）**

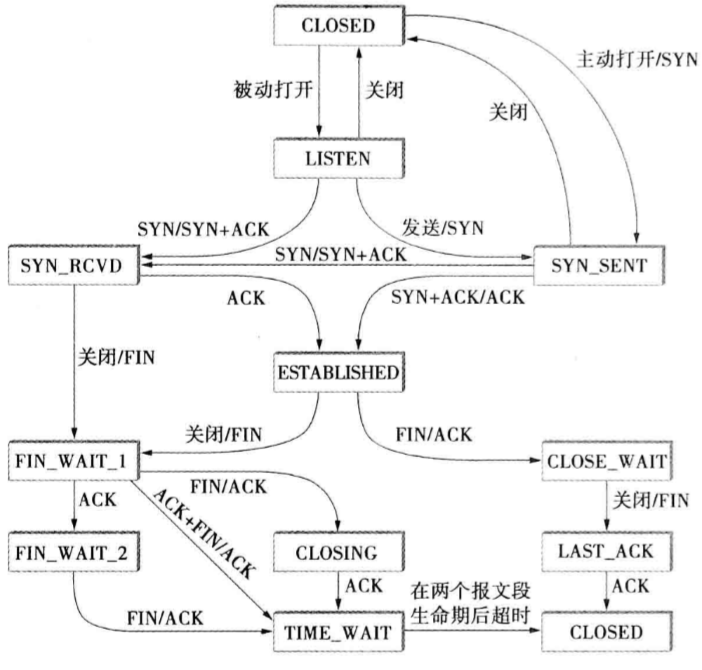
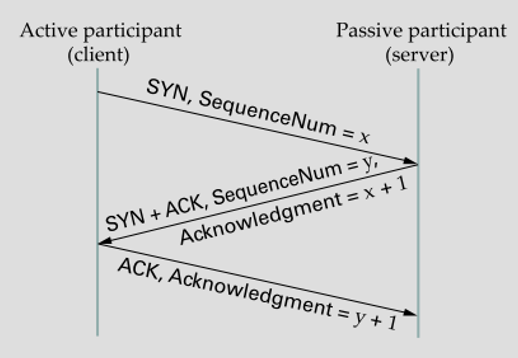
UDP消息队列，使用端口区分不同队列，在应用层保证可靠性。



**TCP（可靠字节流协议）**



三次握手 状态转换 & 四次挥手



# Ch2 IPv6

**IPv4和IPv6差异：**

* 地址空间：32 vs 128
* 全局unicast地址等价于IPv4公开地址
* 直接使用CIDR，网络前缀取代掩码，前缀表示子网号
* 报头格式大简化，方便硬件处理
* 网络管理更加简单：自动发现和自动配置功能
* 安全性支持：IPSec、实现认证头、安全载荷封装
* Qos能力、多播寻址、可移动性

**IPv6地址分类和特点：**

①Unicast（单播） ②Anycast（任播） ③Multicast（组播，广播由组播代替） ④Reserved

IPv6地址是分配给interface而不是node的，所有接口都必须有至少一个link-local unicast，一个单接口可分配任何一种类型的多重地址（unicast/any/multicast）或地址范围。

地址类型由前缀FP决定：

* 可聚类全局unicast地址：2000::/3（001）
* Link-local Unicast地址：FE80::/10（1111 1110 10）
* Multicast地址：FF00::/8
* 保留地址：全零地址、Loopback地址、嵌入到IPv4的IPv6地址。

Link-Local地址作用：本地链路地址分配、自动地址配置、邻站发现

**接口标识符**

MAC Address（IEEE 802、48bits）→EUI-64 Address（64bits）→IPv6 Interface ID（64bits）

**两种地址自动分配方式：**

stateless：通过邻居发现完成：

* 1. 路由器发现：主机定位链路上的路由器；
  2. 前缀发现：主机发现链路上的地址前缀；
  3. 参数发现：e.g. MTU；
  4. 地址自动配置：网络接口地址可选无状态配置；
  5. 地址解析：IP地址和MAC地址之间映射；
  6. 找到目标的下一跳路由器；
  7. 邻居不可访问性检测（NUD）：确定连接上的邻居不可达；
  8. 重复地址检测（DAD）：节点可以检查地址是否已在使用中。

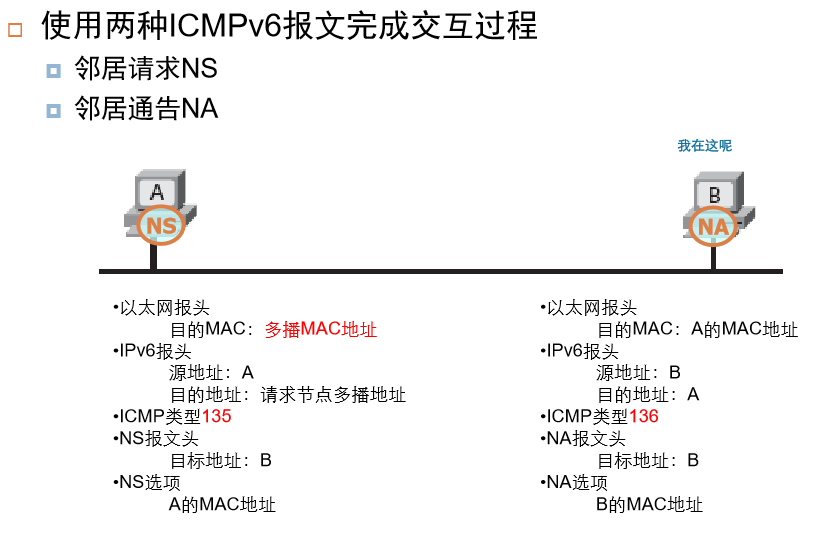
stateful：DHCP，需要服务器

1. 路由器要求新主机进入DHCP服务器（通过设置managed configuration位）；
2. 新主机多播到所有DHCP服务器；
3. DHCP服务器为新主机分配一个IPv6地址。

**邻居发现协议——地址解析**

优势：

* 地址解析在三层完成，不同的二层介质可以采用相同的地址解析协议；
* 可以使用三层的安全机制（例如IPSec）避免地址解析攻击；
* 使用组播方式发送请求报文，减少了二层网络的性能压力。



**邻居发现协议——重复地址检测DAD**

* 重复地址检测确保网络中无两个相同的单播地址；
* 所有Unicast地址都需要做DAD；
* 使用NS和NA完成DAD交互过程；
* 每次系统默认一个主机在应用新的IP地址之前会发送3次DAD，如果均没有收到任何回应，那么该地址被认为是可以配置在接口上的。
* 发现地址重复：全局单播地址，不安排给接口；链路本地地址将接口置于不可用状态。
* 源地址为全零地址(::)；目的地址为请求节点组播地址；Target Address为需要检测的IPv6。
* IPv4自身是不具备重复地址检测的，依赖免费ARP。

**IPSec**

IETF建立的一组IP安全协议集。定义了在网际层使用的安全服务，其功能包括数据加密、对网络单元的访问控制、数据源地址验证、数据完整性检查和防止重放攻击。

# Ch3 拥塞控制

**3.1.1 拥塞控制基本概念**

原因：

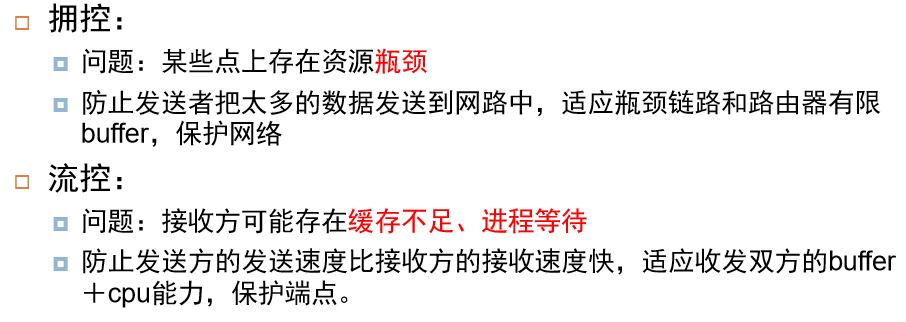
* 宏观原因：网络资源分布不均匀，流量分布不均匀；
* 微观原因：报文聚合到达率大于路由器输出链路的带宽。

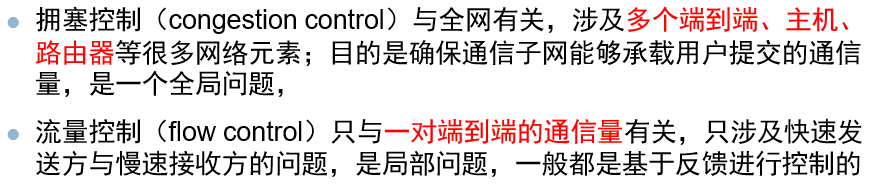
拥塞控制基本单位：数据流（在源目主机对之间，通过相同路由而发送的一系列包）。

路由器中面向TCP流分类：

* TCP-friendly流：
* 非适应性流：
* 有适应性但非TCP-friendly流：

**拥控 vs 流控**





**3.1.2 解决拥塞**

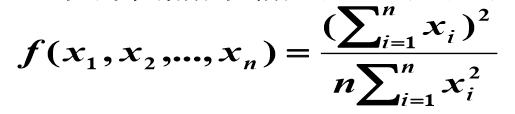
1. R/H为中心的分配

* R为中心：链路算法，路由器和交换机，主队队列管路AQM
* H为中心：源算法，主机和网络边缘设备，TCP拥塞控制算法

1. 基于预留与反馈
2. 基于窗口与速率

评价资源分配有效性指标：网络能力Power = 吞吐率/延迟（bit/s²）

公平的定义：

* 一般公平原则（每个流有相等的带宽）
* Raj Jain公平指数

**3.2 排队算法**

**FIFO：FIFO + tail drop**

* 变化：优先排队（TOS作优先权标记）
* 不能区别不同的流；由于存在非TCP应用，可能把端—端拥塞机制旁路掉，e.g. IP电话。

**FQ（公平排队）**

原理：为每个正在被路由器处理的流分别维护一个队列， 路由器以轮循方式服务每个队列。

发送算法：同时到达 & 不是同时到达（习题）

2点注意：

①工作守恒：只要有包留在队列中，就决不让链路空闲，尽量发送；

②若链路满负载且有n个流在发送数据，则任何一个队列不能长时间占用大于1/n的带宽。

**WFQ（加权公平排队）**

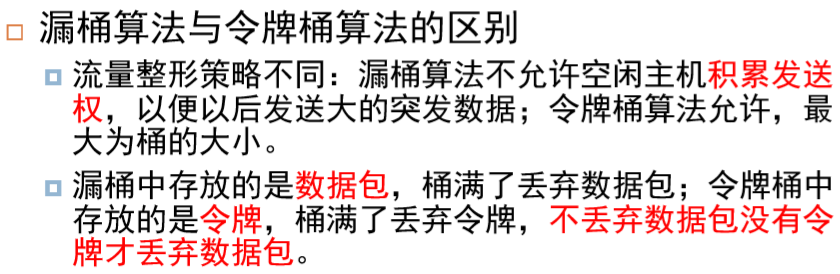
（习题）Fi = max(Ai, Fi-1) + Pi/权重

**流量整形**

漏桶算法：强迫输出按平均速率进行，不灵活，溢出时要丢包。

令牌桶算法：网络中最大包=rt+b（令牌产生率×时间间隔+桶容量）

串联：速率不变、峰值增加；并联：速率增加、峰值增加。



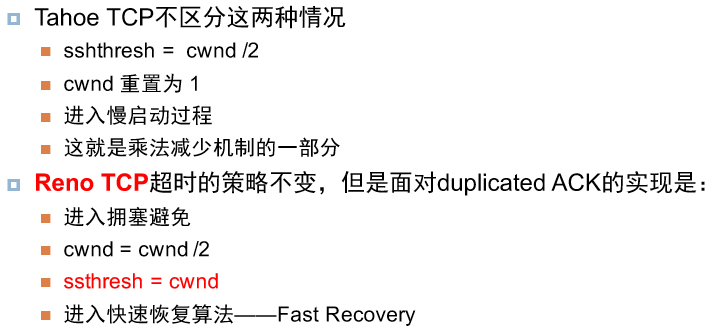
**3.3 TCP拥塞控制机制**

3.3.1 慢启动 & 拥塞避免

* 指数式增加CW；
* 当连接建立的时候慢启动一直到丢包才结束；
* 当超过一个发送方设置的阈值ssthresh，CW线性增加，这其实是一种拥塞避免。

TCP通过隐式反馈的方式学习CW：TCP源根据它所获得的网络中存在的拥塞级别来设定CW。慢启动之后，如果超过ssthresh，可能会导致当拥塞程度上升就减小CW；当拥塞程度下降（收到ACK）时，就增加CW，即加法式增加/乘法式减少。

感知拥塞：1）重复ACK；2）超时。



3.3.3 快速重传

发送方TCP等到3个Duplicate ACK后就重发丢失的包，比等待报文超时要快。

3.3.4 快速恢复

1）cwnd = cwnd/2；2）ssthresh = cwnd；3）重新进入拥塞避免阶段，CW线性增加。

慢启动仅在连接开始和超时发生。

**拥塞避免机制**

**基于路由器的拥塞避免（DECbit、RED、ECN）**

DECbit：

路由通过设置流过路由器包中的二进制拥塞位来实现；

路由器计算平均队列长度＞1 → 有可能拥塞 → 设置拥塞位 → 目的主机拷贝拥塞位到ACK报文 → 源主机调整发送速率，并设置标志位通知目的主机已经处理了拥塞。

ECN显式拥塞控制：

四个标志位：ECN、CE、CWR、ECE（IP、TCP头部各两个）

**基于源的拥塞避免**

TCP Vegas

**3.5 路由器主动队列管理**

RED（随机早期检测）：

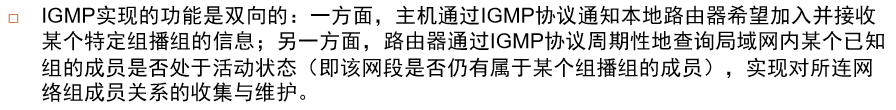
每台路由器监视自己的队列长度，通过丢弃一个分组隐式地通知源发生拥塞，源通过超时或重复确认得到通知。

# Ch3-2 组播

组成：

①组播地址；

②成员机制：IGMP查询和报告



IGMPv2：(1)选举IGMP查询器（IP地址最低的），查询一次所有路由器得到相同结果；

(2)主机离开报告机制。

③路由协议：PIM协议

组播树：发送者到接收者之间的接受路径。根（发送者）、组播出口

防止组播反向路径转发：

路由器发现、避免环路的能力：RPF检查：组播路由器上朝发送者方向的接口被称为RPF接口，只有从RPF接口来的数据才能被转发，从其它接口来的数据被认为是无效的。

PIM-DM（路由器递归查询）

* 记录方式：（S,G）+三类接口（forwarding、pruning、RPF），路由条目数 = |S|×|G|
* 组播树：SPT（最短路径树），也称源树。
* 五种数据包：Hello、Join/prune、Graft、Graft-Ack、Assert

PIM-SM（组成员主动报告）

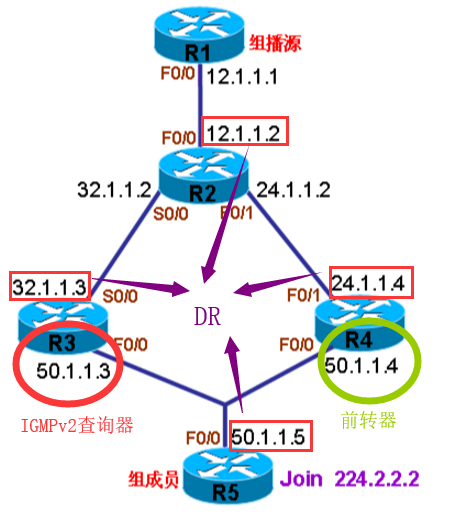
* 记录方式：（\*,G）+一类出口（forwarding），路由条目数 = |G|
* 组播树：RPT（共享树），选出组播汇聚点RP，认为RP是组播源。
* RP的确立：1）手工静态配置； 2）BSR（Bootstrap Router）通告（涉及两种数据包）
* 七种数据包：Hello、Bootstrap、Candidate-RP-Advertisement、Join/Prune、Assert、Register、Register-Stop

PIM DR（指定路由器）的作用：（每条链路上IP地址高的）

* IGMPv1选举DR充当查询IGMP查询器，避免重复查询；
* PIM-SM中RPF检测失败：组播源向RP注册，建立从DR到RP的源树，（S,G）记录。

PIM Forwarder（前转器）

选择规则：AD值小 → metric值小 → IP大的

  
Multicasting域间组播

MBGP（交换路由信息）+ MSDP（交换信源信息）+ PIM-SM（域内路由）

# Ch4-P2P