计算机网络

- 主讲：肖林

内容

- 第1章 概述
- 第2章 物理层
- 第3章 数据链路层
- 第4章 介质访问控制子层
- 第5章 网络层
- **第6章 传输层**
- 第7章 应用层

第6章　传输层

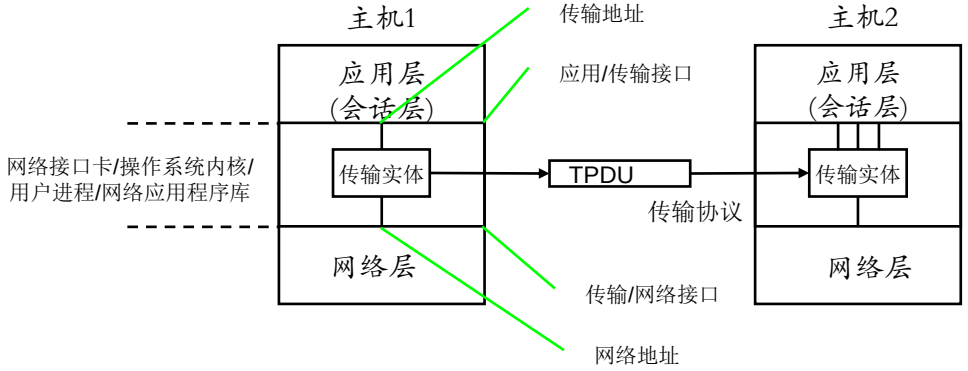
- 传输服务
- 传输协议的要素
- 拥塞控制
- 因特网传输协议UDP
- 因特网传输协议TCP
- 传输协议与拥塞控制
- 性能问题

6.1 传输服务

- 提供给上层的服务
- 传输服务原语
- Berkeley 套接字

提供给上层的服务

- 提供服务
  - 传输实体
    - 面向连接服务/无连接服务，传输层将子网的技术、设计和缺陷与上层隔离
  - 1~4传输服务提供者，5~7传输服务用户
    - 传输层将增强网络层的服务质量



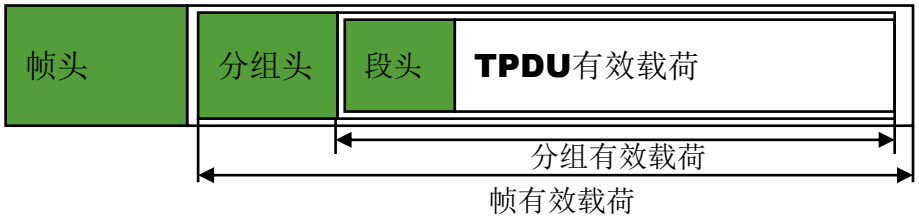
传输服务原语

- 传输层与网络层
  - 传输服务屏蔽不同的网络
    - 面向连接服务：不可靠的网络服务上建立可靠服务
    - 无连接服务
  - 服务对象：程序只能看到传输服务，无法控制网络
- 传输服务原语

Primitive	Packet sent	Meaning
LISTEN	(none)	阻塞，直到某进程与之连接
CONNECT	CONNECTION REQ.	主动建立一个连接
SEND	DATA	发送数据
RECEIVE	(none)	阻塞，直到收到数据
DISCONNECT	DISCONNECTION REQ.	释放连接

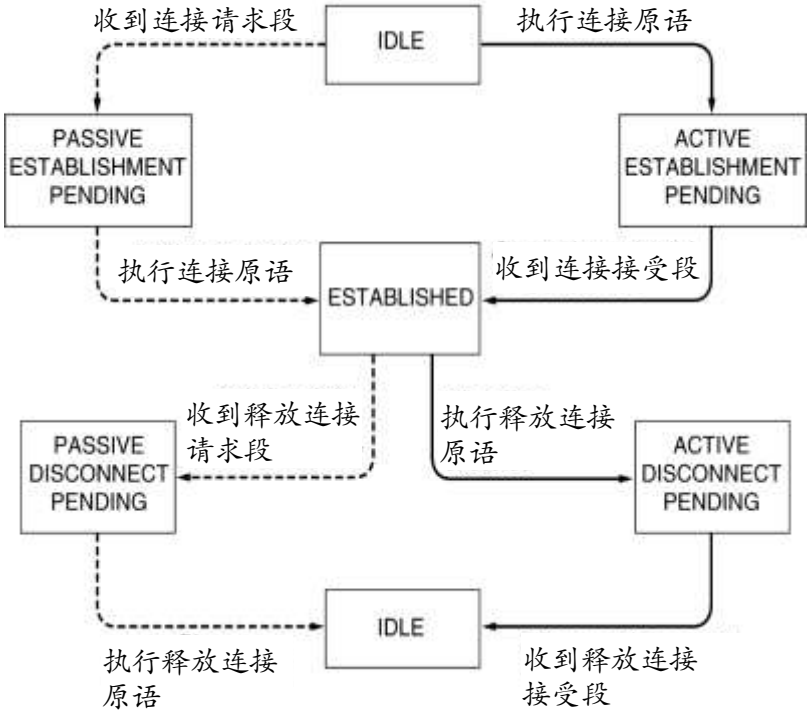
传输服务原语(2)

- 原语
  - LISTEN
  - CONNECT
  - SEND
  - RECEIVE
  - DISCONNECT
- 段（传输协议数据单元TPDU）



  - 释放连接：非对称方式，对称方式

简单连接管理的状态图


收到连接请求段
IDLE
执行连接原语
PASSIVE ESTABLISHMENT PENDING
ACTIVE ESTABLISHMENT PENDING
执行连接原语
ESTABLISHED
收到连接接受段
收到释放连接请求段
执行释放连接原语
PASSIVE DISCONNECT PENDING
ACTIVE DISCONNECT PENDING
IDLE
执行释放连接原语
收到释放连接接受段

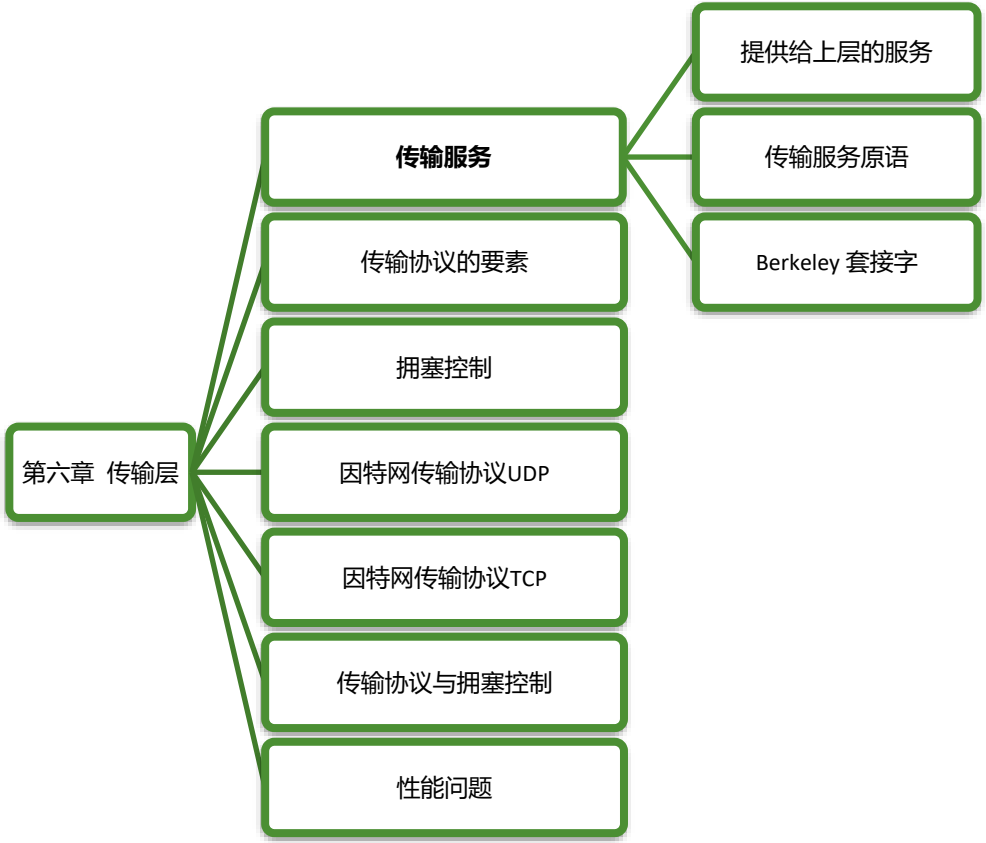
Berkeley 套接字

- TCP所用套接字
  - UNIX，Windows(winsock)

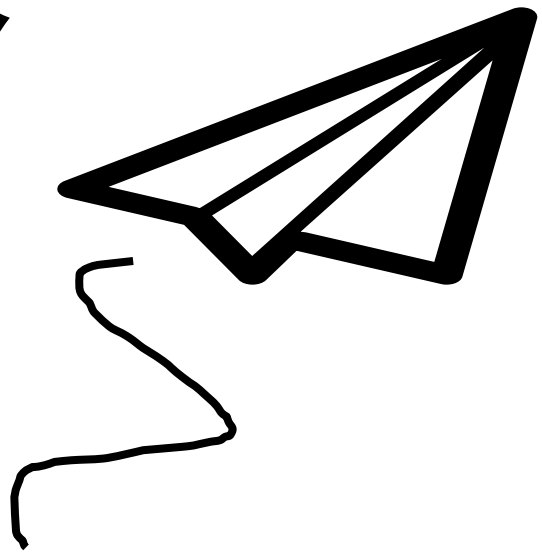
原语	含义
SOCKET	创建一个新通信端点（传输实体中分配相应表空间）
BIND	将套接字与一个本地地址关联（控制使用知名端口，或者使用任意端口）
LISTEN	声明愿意接受连接，给出队列长度（为多客户并发入境呼叫准备）
ACCEPT	阻塞调用者，直到入境连接到来（连接到达，创建新的套接字并返回关联文件描述符）
CONNECT	创建一个链接
SEND	通过连接发送一些数据
RECEIVE	从连接接收一些数据
CLOSE	释放连接

套接字编程实例：Internet文件服务器

本章导航与要点


第六章 传输层
传输服务
传输协议的要素
拥塞控制
因特网传输协议UDP
因特网传输协议TCP
传输协议与拥塞控制
性能问题
提供给上层的服务
传输服务原语
Berkeley 套接字

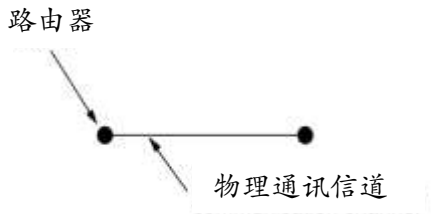
本节课程结束



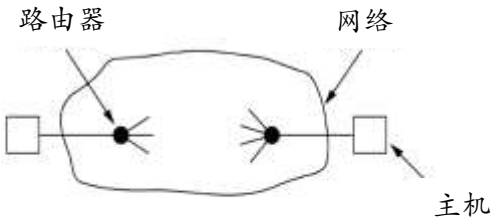
6.2 传输协议的要素

- 寻址
- 建立连接
- 释放连接
- 流量控制和缓冲策略
- 多路复用
- 崩溃恢复

与数据链路层的区别：
目的地址、初始连接的复杂度、子网存储能力、缓冲区数量



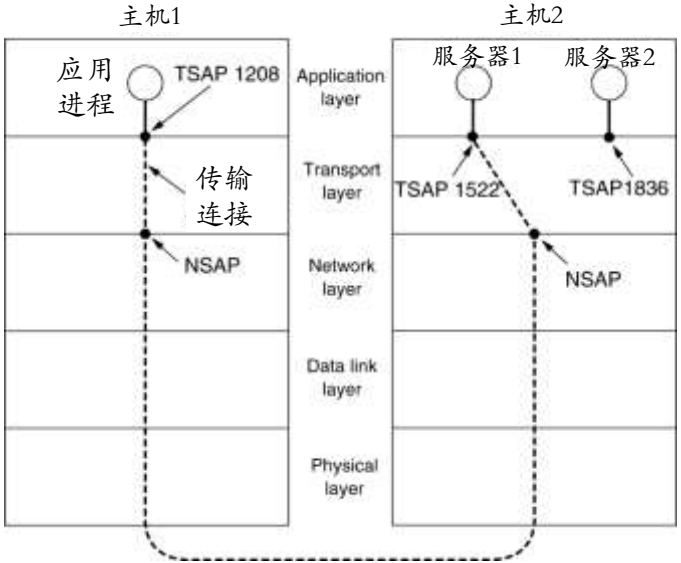
(a) 数据链路层环境

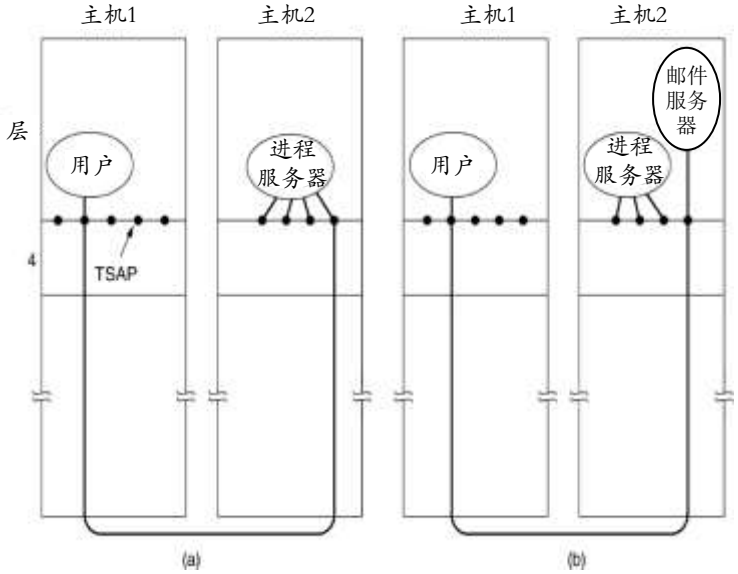


(b) 传输层环境

寻址

- 因特网：<IP地址，本地端口>
  - 传输服务访问点TSAP
  - 网络服务访问点NSAP

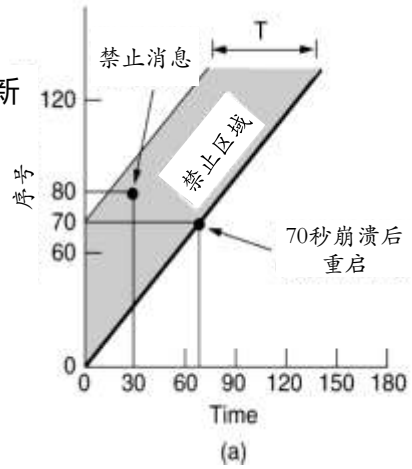


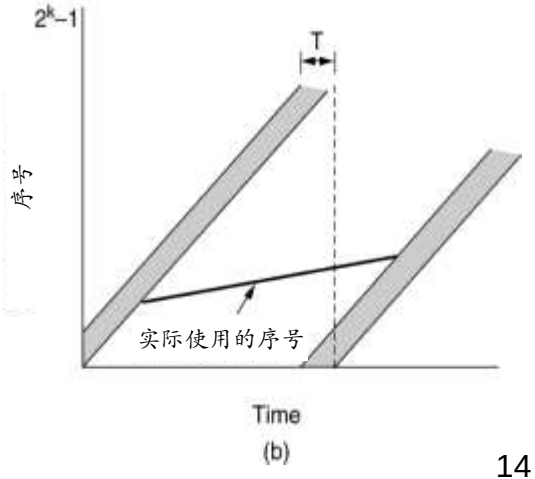


建立连接

- 网络问题：分组可能丢失、存储、重复出现
  - 一次性：废弃使用过的传输地址（首次连接，地址确定困难）
  - 分配一个连接表示符（每次连接加1，控制数据包生存期）
    - 生存期T：限制网络设计，确定最大延迟；包含一个跳计数器，每次减1；保存创建时间，所有路由器同步时钟。 时钟速率C，序号空间S，S/C>T
    - 段标签：序号T秒内不重复，老不干扰新
    - 日时钟：永远运行，

      时钟低k位作为开始序号





- 示例（a）

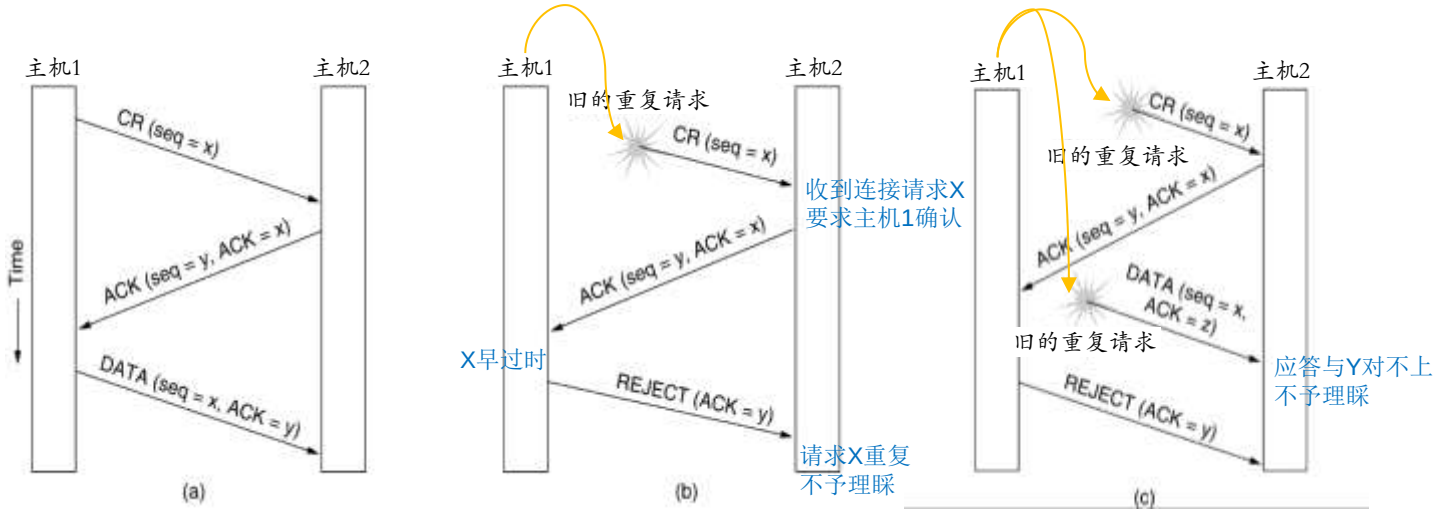
  70秒后崩溃，不会使用前面时钟低k位序号

- 示例（b）

  实际使用序号增长不能过慢或过快，

建立连接(2)

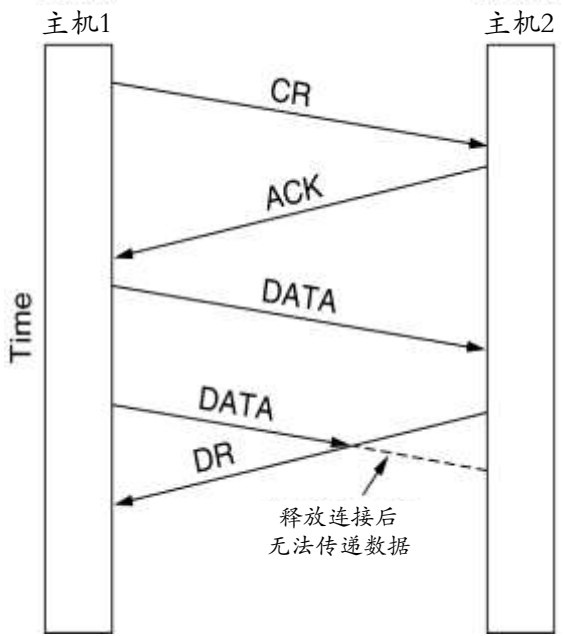
- 三次握手（避免新连接请求与接收方当前连接序号重复问题）
  - 时间戳：辅助扩展32位序号，确保生存期内不会回绕
  - 防止序号回绕(PASW)机制：伪随机初始序号
    - 防止攻击者预测到下一个初始序号



释放连接

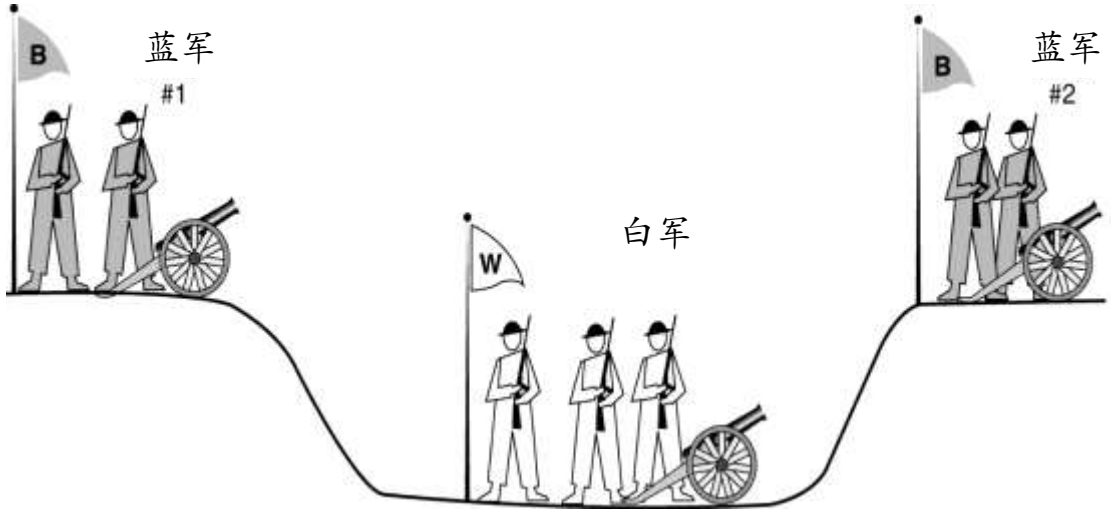
- 释放
  - 非对称
    - 任何一方可DISCONNECT，

      其TPDU到达对方-连接释放
    - 右图导致数据丢失
  - 对称：
    - 各方单独DISCONNECT，停止发送但继续接收
    - 双方均DISCONNECT后释放连接



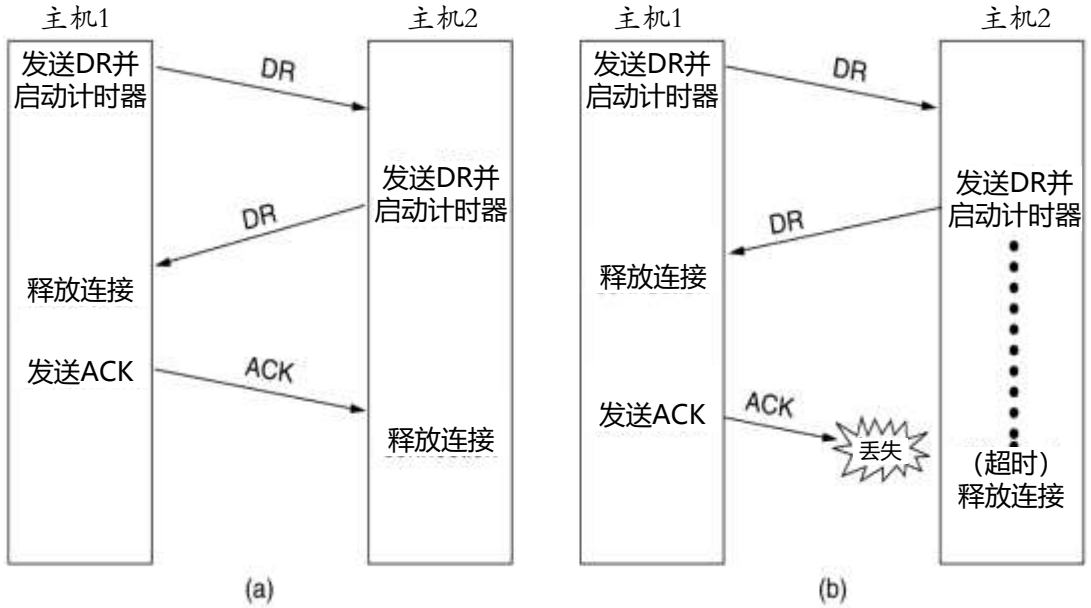
释放连接(2)

- 两军问题



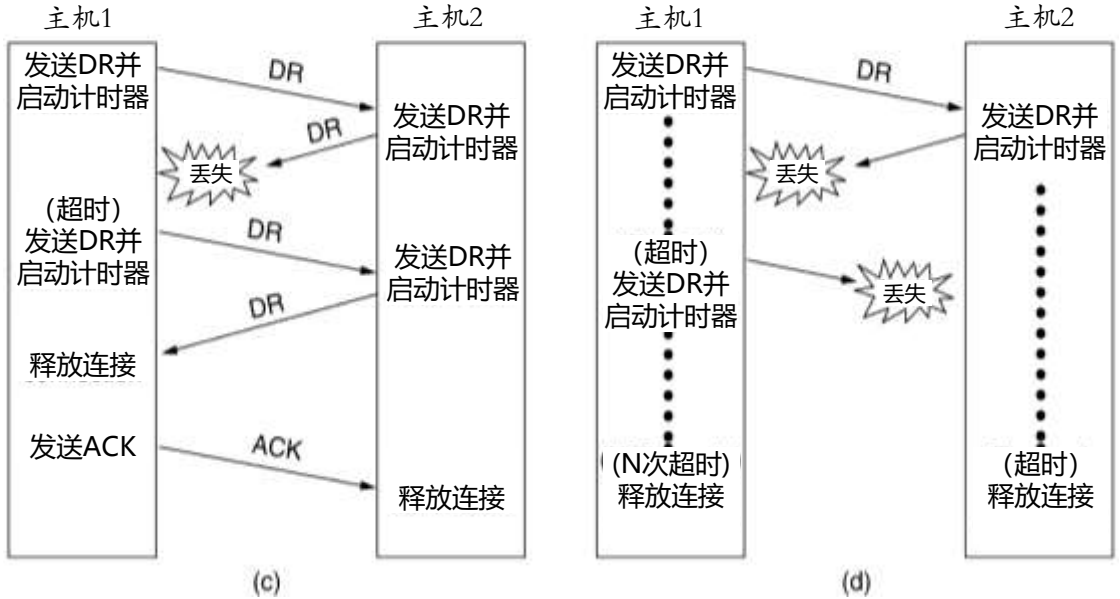
释放连接(3)

- 释放连接的协议场景
  - 1、正常情况
  - 2、超时释放



释放连接(4)

- 释放连接的协议场景
  - 3、第二次成功
  - 4、重传全部失败

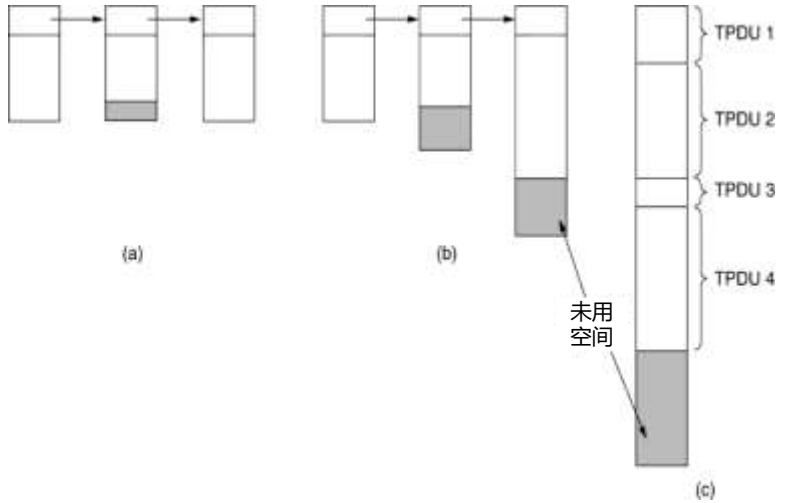


差错控制和流量控制

- 与数据链路层协议不同
  - 数据链路层
    - CRC；自动重发请求（ARQ）；停等；滑动窗口（双向）
  - 传输层
    - 错误检测：
      - 传输层校验不可替代，路由器内错误在链路层校验之外；
      - 链路层校验能提高性能（局部重传）
    - 滑动窗口协议与重传：
      - 无线线路（卫星除外）：链路的带宽延迟积很小，如802.11使用停等协议；
      - 有线与光纤：链路误码率极低，可忽略重传，错误由端到端重传解决；
      - 传输层如TCP的连接，带宽延迟积远大于单个段，必须使用较大的滑动窗口

差错控制和流量控制(2)

- 缓存数据问题
  - 发送端的源缓冲区：保存已发送未确认的段；
  - 接收端的目标缓冲区：丢弃段不会造成永久伤害；
  - 连接承载的流量类型决定了缓冲区设立与否，低带宽突发流量（如Telnet）可以不设立
- 建立缓冲池
  - A. 链接固定大小的缓冲区
  - B. 链接可变大小的缓冲区
  - c. 每条连结使用一个缓冲区



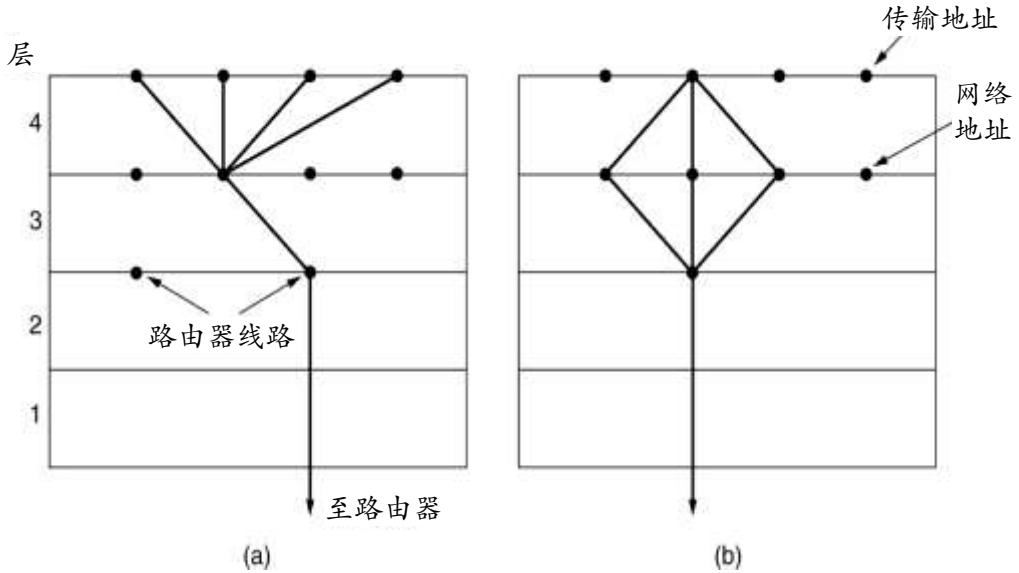
差错控制和流量控制(3)

- 动态管理窗口
  - 16 潜在的死锁；定期在连接上发送控制段

	主机A	消息	主机B	注释
1	→	< request 8 buffers>	→	A需要8个缓冲区
2	←	<ack = 15, buf = 4>	←	B只能提供4个，可接收消息0-3
3	→	<seq = 0, data = m0>	→	A还有3个缓冲区可用
4	→	<seq = 1, data = m1>	→	A还剩2个缓冲区
5	→	<seq = 2, data = m2>	•••	消息丢失，A认为还剩1个缓冲区
6	←	<ack = 1, buf = 3>	←	B确认0和1，允许2-4
7	→	<seq = 3, data = m3>	→	A还剩1个缓冲区
8	→	<seq = 4, data = m4>	→	A还剩0个缓冲区，必须停止
9	→	<seq = 2, data = m2>	→	A超时并重传
10	←	<ack = 4, buf = 0>	←	全部确认，但无缓存区，A仍然阻塞
11	←	<ack = 4, buf = 1>	←	A可以发送m5了
12	←	<ack = 4, buf = 2>	←	B又找到一个新的缓冲区
13	→	<seq = 5, data = m5>	→	A还剩1个缓冲区
14	→	<seq = 6, data = m6>	→	A被再次阻塞
15	←	<ack = 6, buf = 0>	←	A仍然阻塞
16	•••	<ack = 6, buf = 4>	←	潜在的死锁

多路复用

- 向上多路复用：多个传输连接使用同一网络连接
- 向下多路复用：一个传输连接使用多个网络连接



崩溃恢复

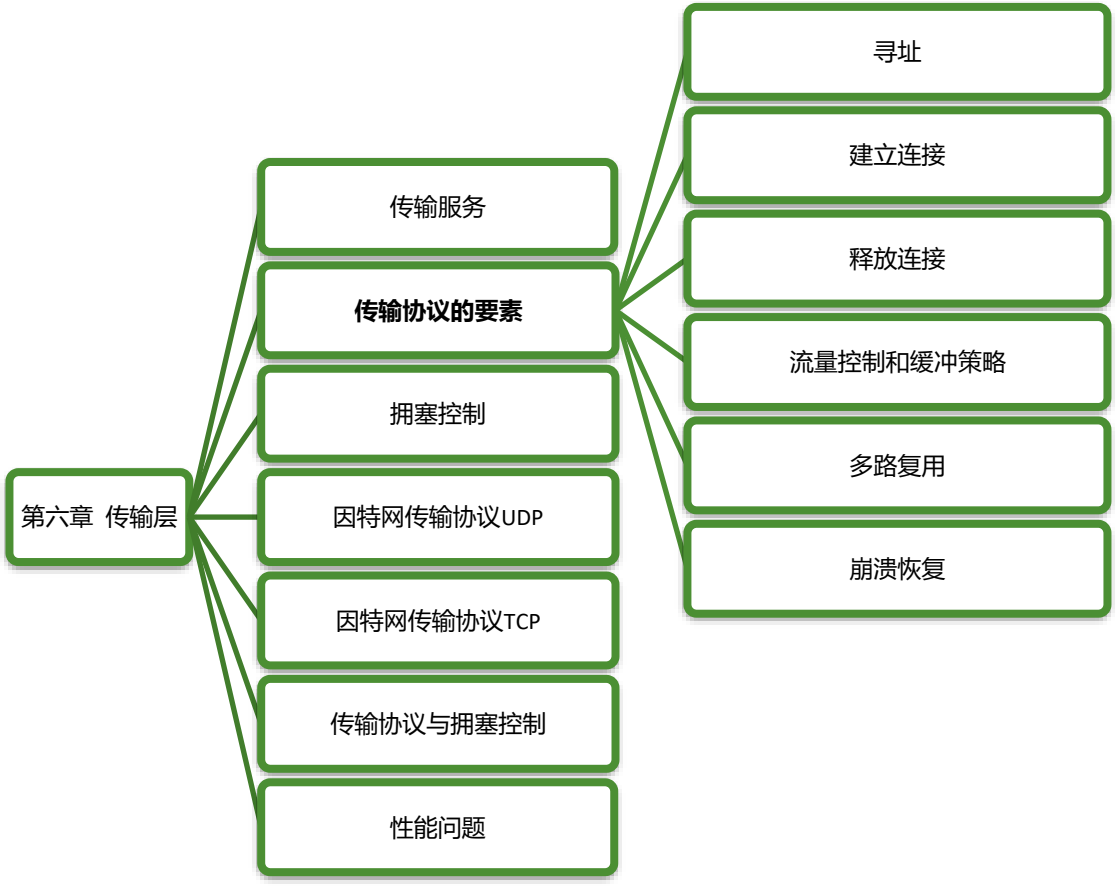
- 发送主机状态
  - S0：没有未被确认的TPUD
  - S1：有一个未被确认的TPUD
- 接收主机策略
  - 先确认后写：AC(W)、AWC、C(AW)
  - 先写后确认：WC(A)、WAC、C(WA)
- 发送主机策略
  - 总是重传
  - 从不重传
  - S0状态下重传
  - S1状态下重传
- 结论
  - 图6-18P408
  - 崩溃恢复只能由上一层完成

接收主机的策略

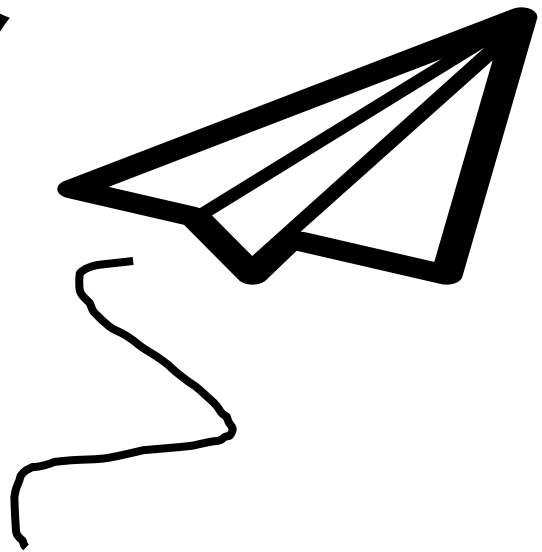
发送主机的策略	先ACK，再写			先写，再ACK		
	AC(W)	AWC	C(AW)	C(WA)	W AC	WC(A)
总是重传	OK	DUP	OK	OK	DUP	DUP
永不重传	LOST	OK	LOST	LOST	OK	OK
S0状态下重传	OK	DUP	LOST	LOST	DUP	OK
S1状态下重传	LOST	OK	OK	OK	OK	DUP

OK = 协议功能正常
DUP = 协议产生重复消息
LOST = 协议丢失一个消息

本章导航与要点


第六章 传输层
传输服务
传输协议的要素
拥塞控制
因特网传输协议UDP
因特网传输协议TCP
传输协议与拥塞控制
性能问题
寻址
建立连接
释放连接
流量控制和缓冲策略
多路复用
崩溃恢复

本节课程结束



6.3 拥塞控制

- 理想的带宽分配
- 调整发送速率
- 无线问题

带宽分配

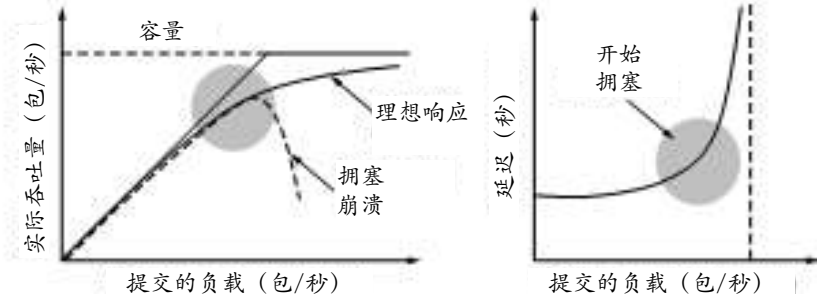
- 效率与功率

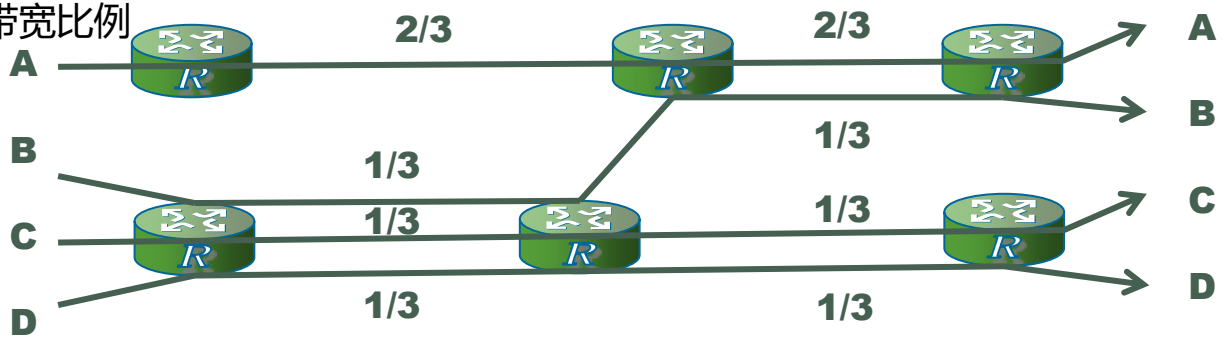
  功率 = 负载/延迟

- 最大-最小公平性

  增加带宽不影响他人

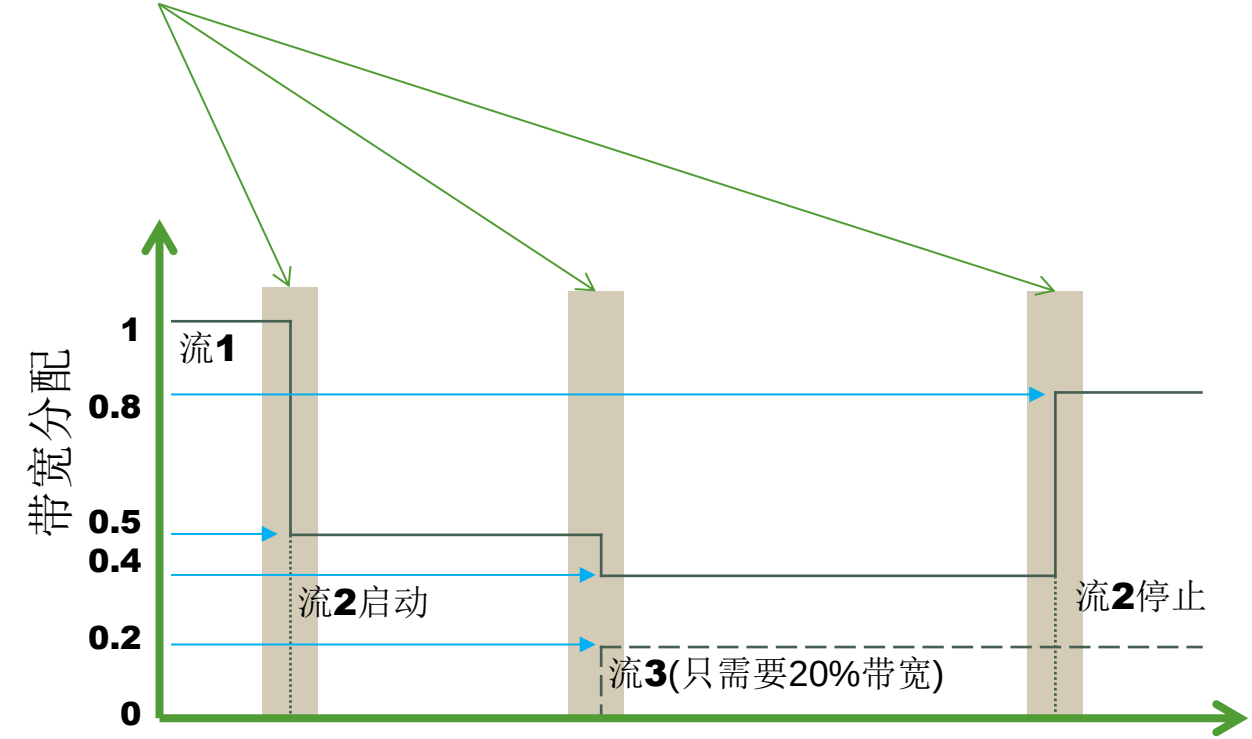
  - 均等带宽比例





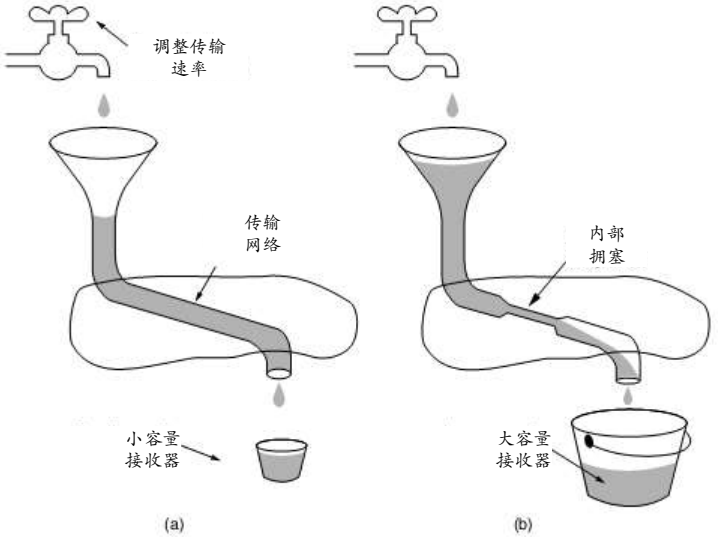
- 收敛

收敛



调整发送速率

- 发送速率与带宽分配
  - 接收端缓冲区不够（a）
  - 拥塞控制（b）
- 显式拥塞协议



协议	信号	显式否	精确否
XCP	使用速率	是	是
TCP	拥塞警告	是	否
FAST TCP	端到端延迟	否	是
Compound TCP	数据包丢失&端到端延迟	否	是
CUBIC TCP	数据包丢失	否	否
TCP	数据包丢失	否	否

调整发送速率（2）

- 加法递增乘法低减(AIMD)

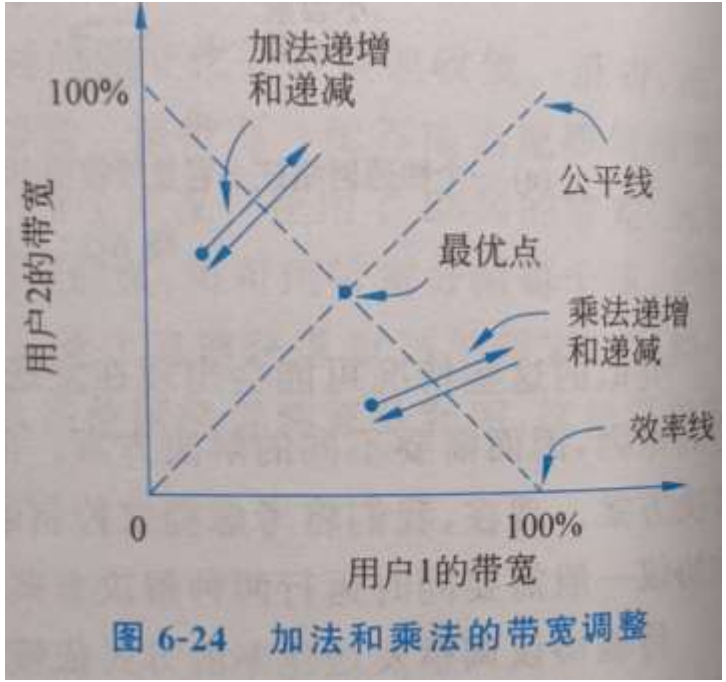


图 6-24　加法和乘法的带宽调整

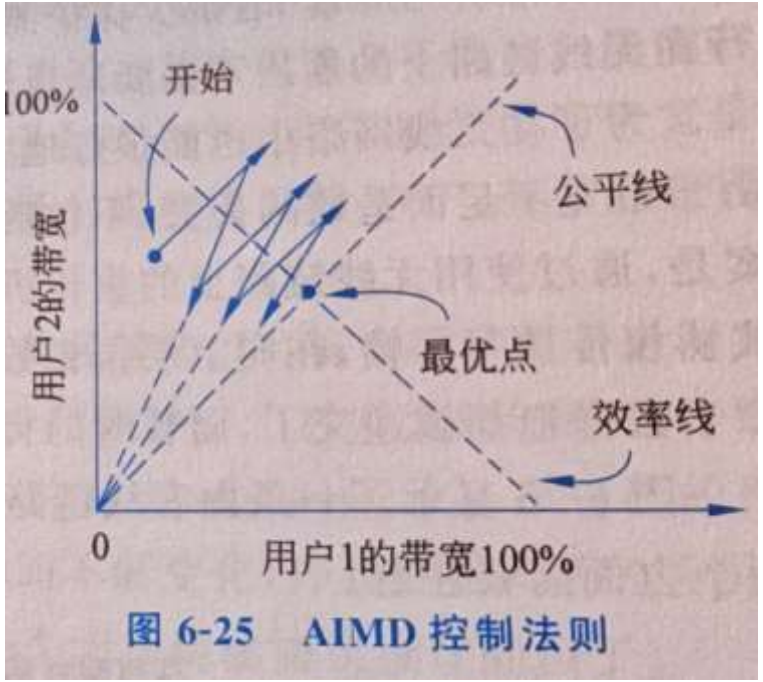
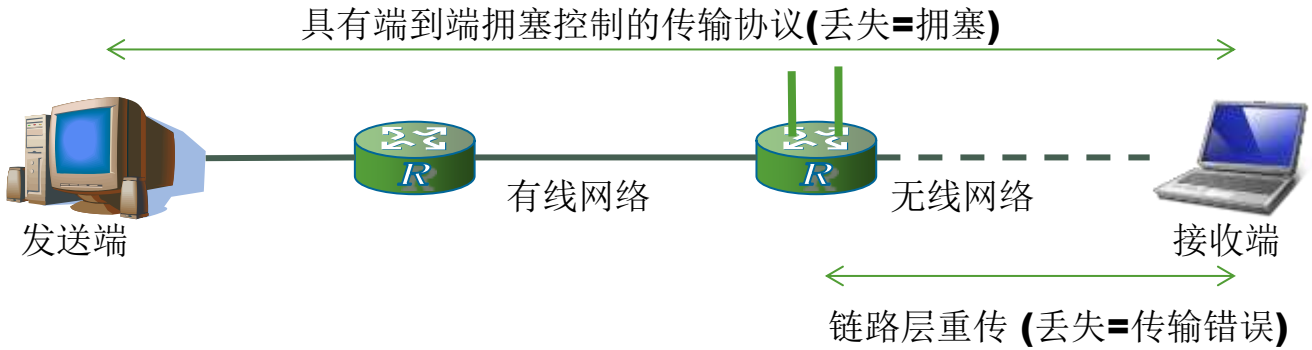


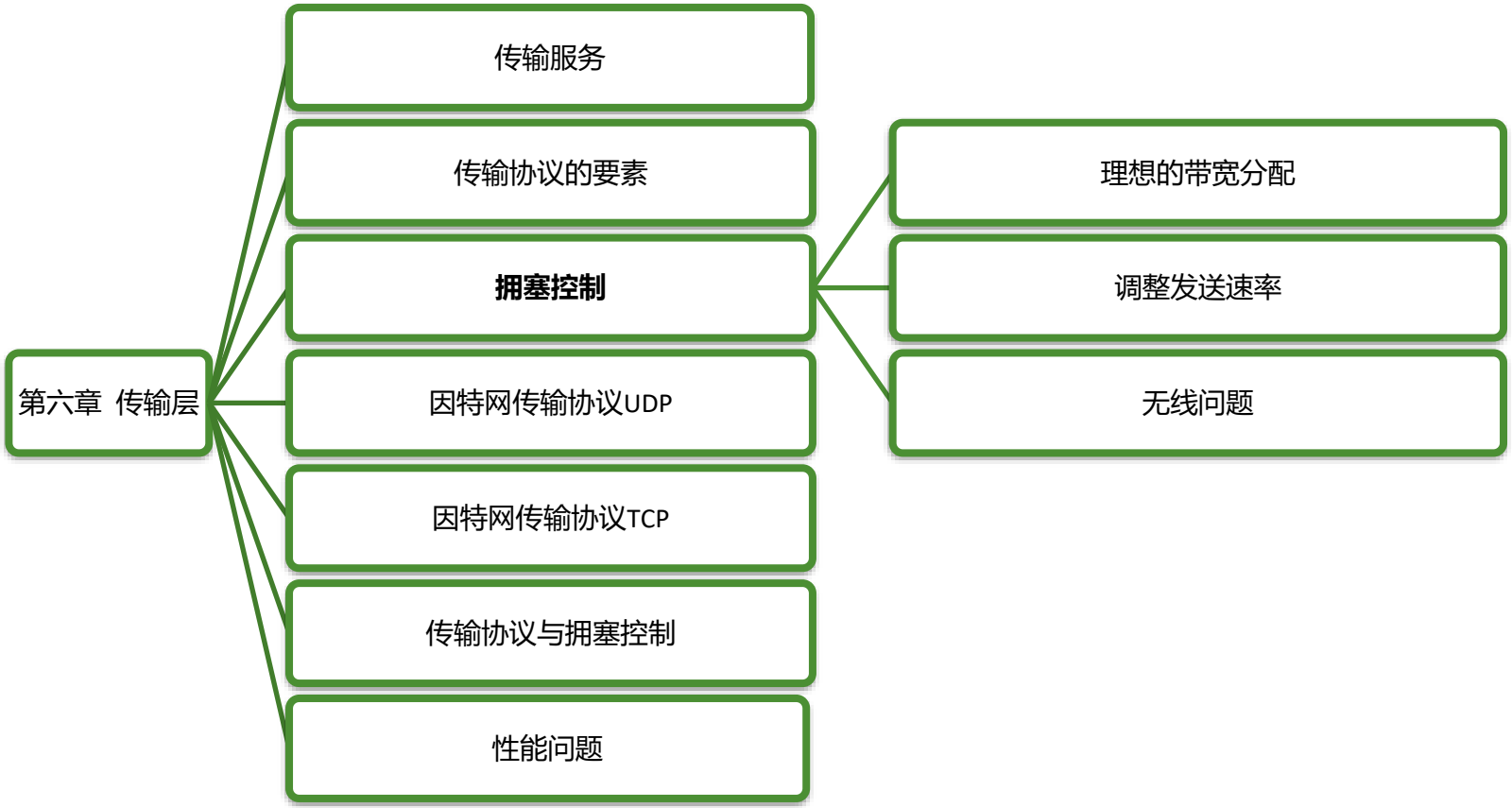
图 6-25　AIMD 控制法则

无线问题

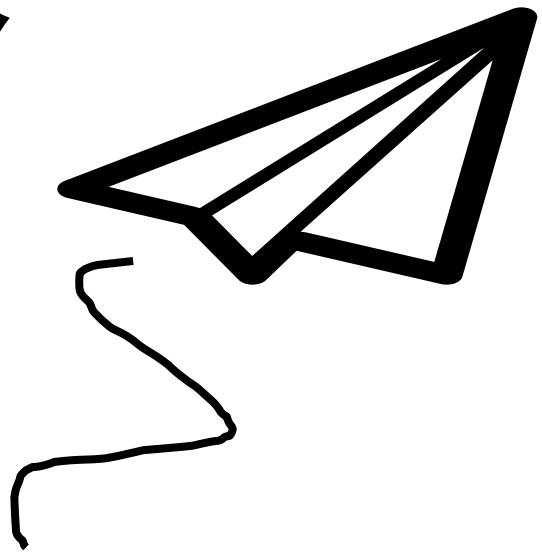
- 拥塞与传输错误
  - TCP连接(AIMD)丢失率：0%~1%~10%
  - 802.11正常丢失率：10%
- 屏蔽策略
  - 链路层重传微秒到毫秒；传输层丢失毫秒到秒



本章导航与要点


第六章 传输层
传输服务
传输协议的要素
拥塞控制
因特网传输协议UDP
因特网传输协议TCP
传输协议与拥塞控制
性能问题
理想的带宽分配
调整发送速率
无线问题

本节课程结束

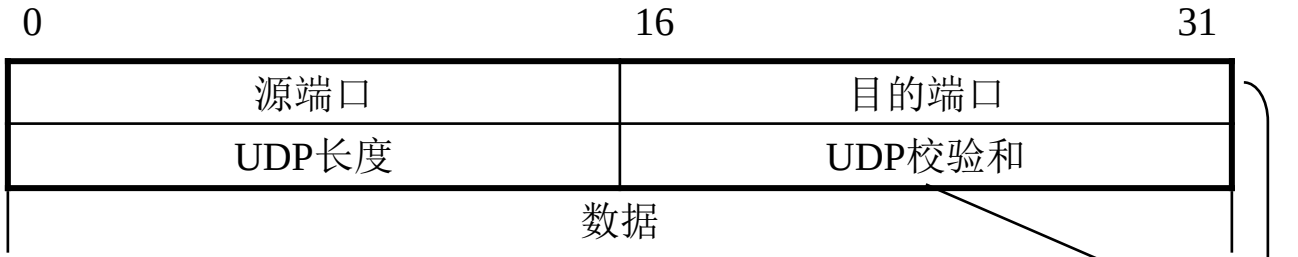


6.4 因特网传输协议UDP

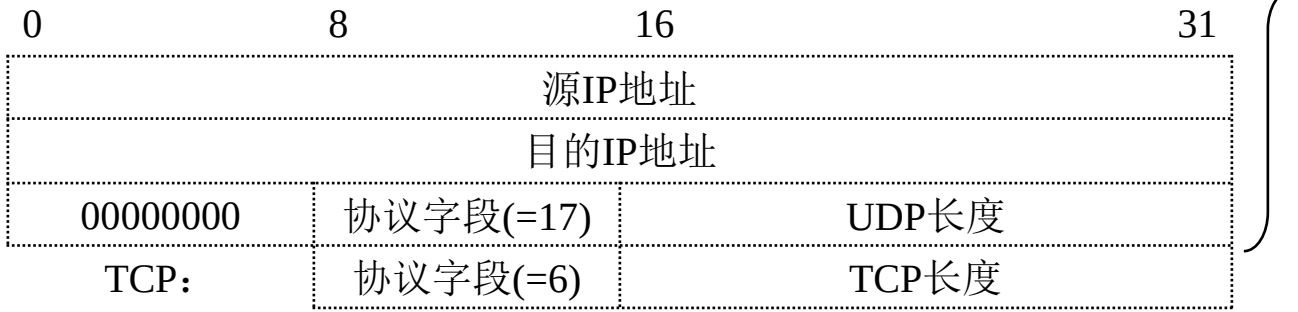
- UDP介绍
- 远程过程调用
- 实时传输协议

UDP介绍

- 用户数据报协议
- UDP头部

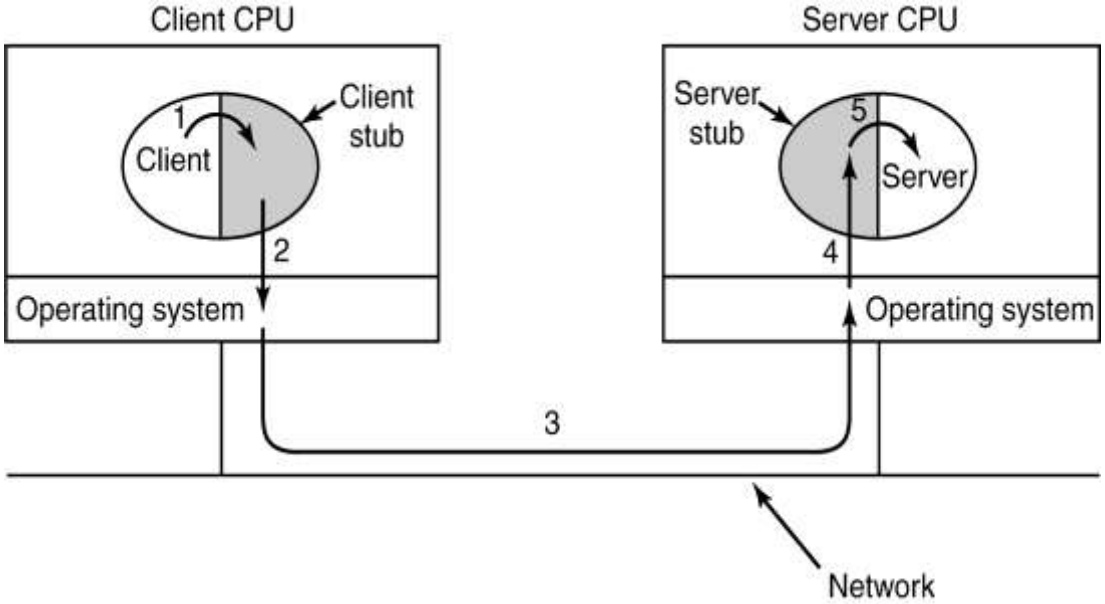


- IPv4伪头部



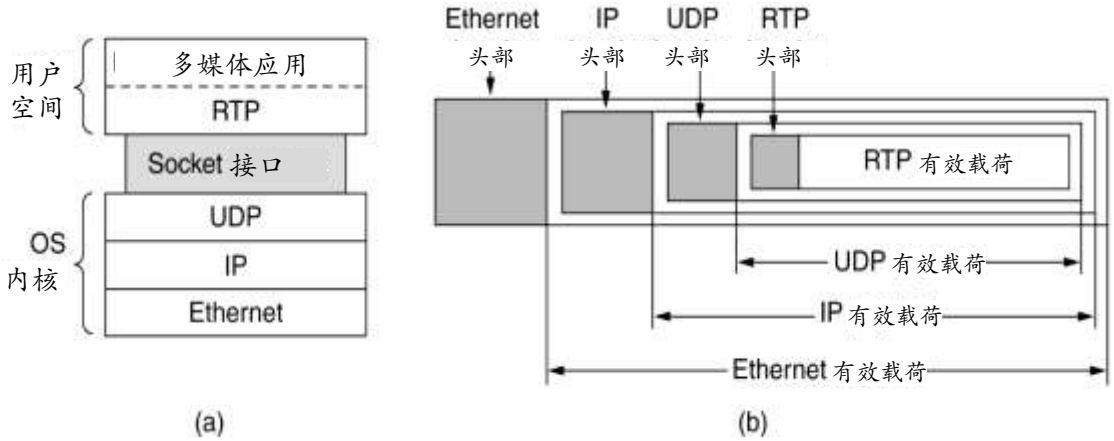
远过程调用

- RPC
  - 客户存根：封装参数（列集）
  - 服务器存根：解封参数（散集）



实时传输协议

- RTP(RFC3550)

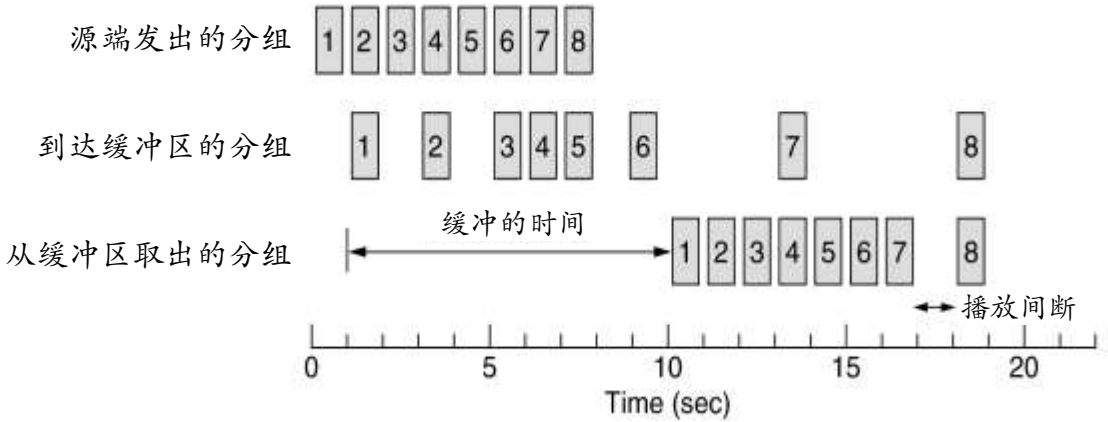


实时传输协议(2)

- RTP——实时传输协议
  - 将多个实时数据流复用到一个UDP数据段中
  - 时间戳：网络延迟变化，多个流之间同步
- RTP头

实时传输协议(3)

- RTCP——实时传输控制协议
  - 向源端提供网络特征反馈信息
  - 延迟、抖动、带宽、拥塞等
- 带有缓冲和抖动控制的播放

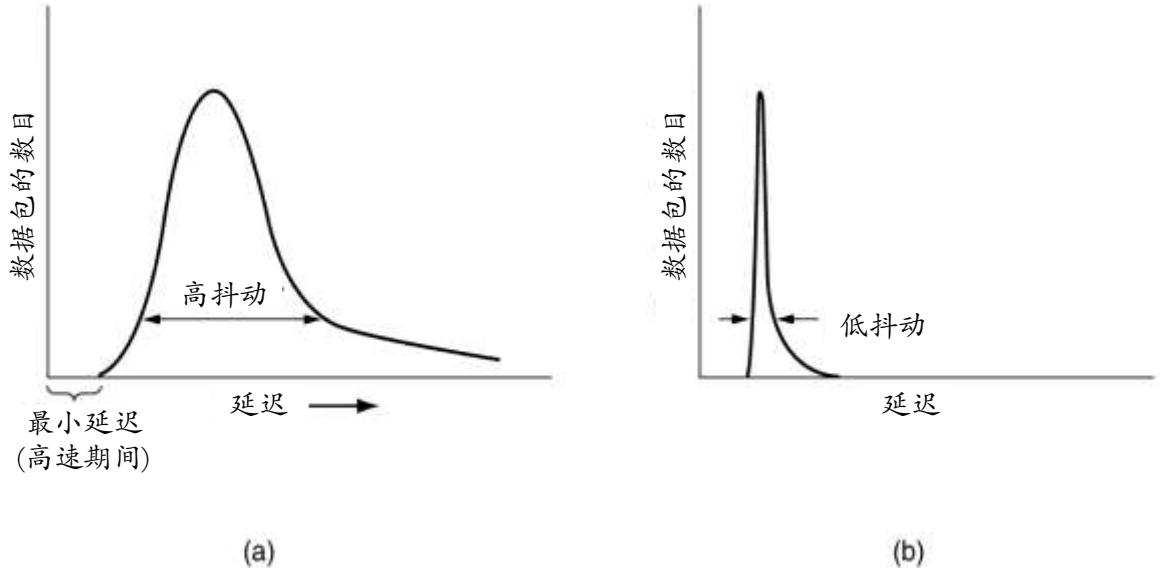


抖动控制

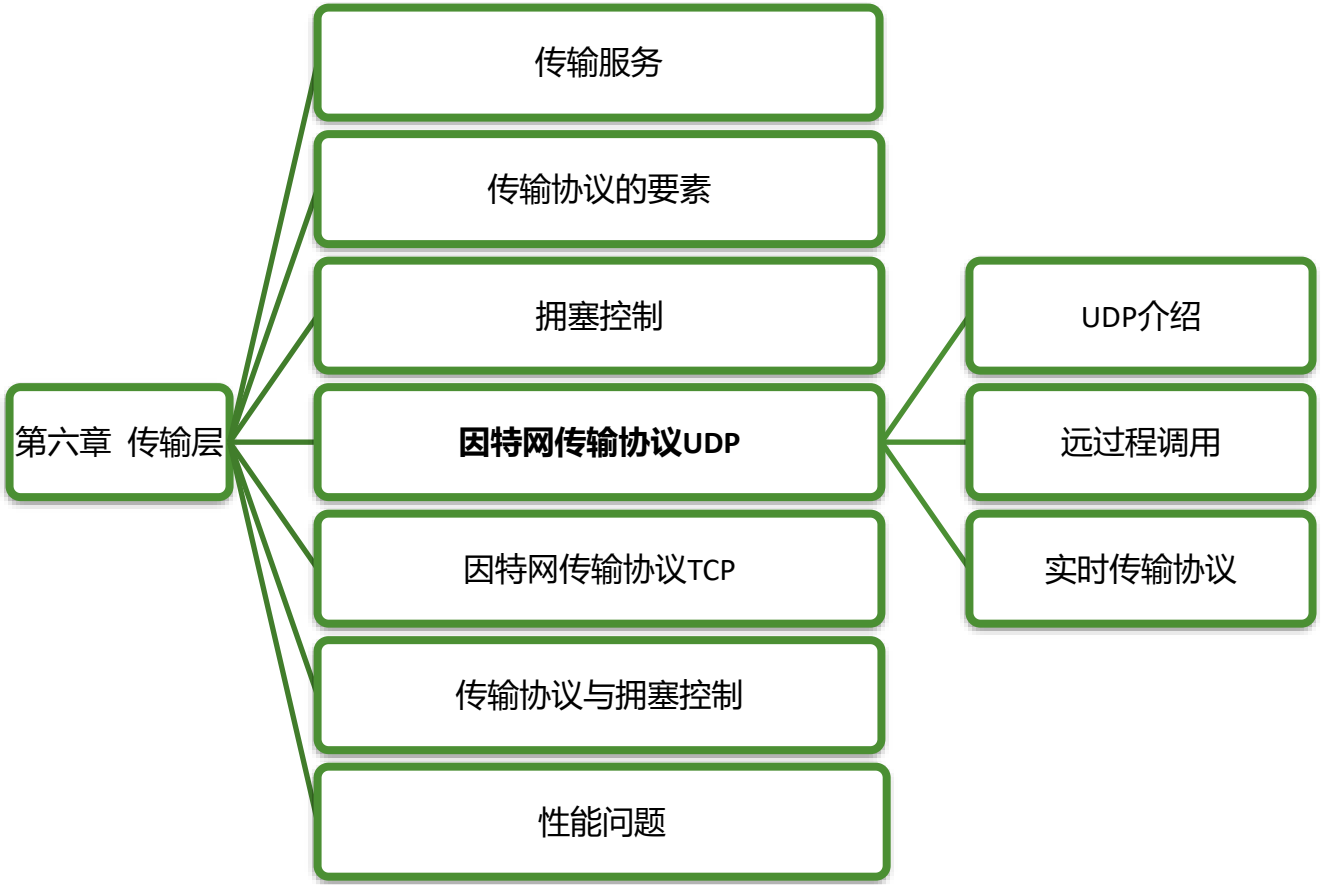
- 播放点
  - 最小延迟、平均延迟
  - 等待时间：

    a、高抖动

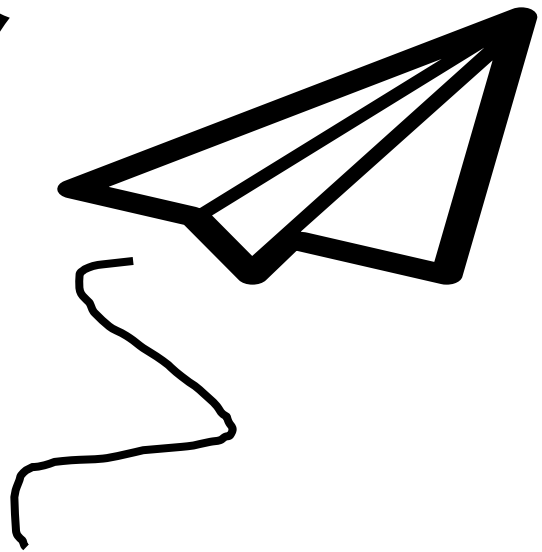
    b、低抖动



本章导航与要点


第六章 传输层
传输服务
传输协议的要素
拥塞控制
因特网传输协议UDP
因特网传输协议TCP
传输协议与拥塞控制
性能问题
UDP介绍
远过程调用
实时传输协议

本节课程结束



6.5 因特网传输协议TCP

- TCP服务模型
- TCP协议
- TCP段的头
- TCP连接建立
- TCP连接释放
- TCP连接管理模型
- TCP滑动窗口
- TCP计时器管理
- TCP拥塞控制

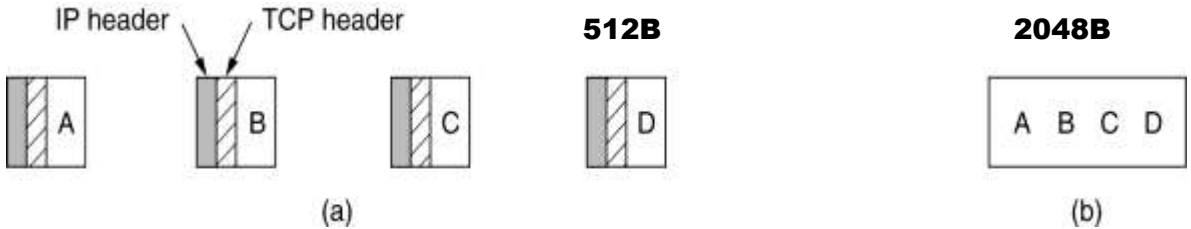
TCP服务模型

- TCP：可靠的端到端的字节(非消息)流
  - 套接字
    - IP + <端口>
  - 端到端之间不保留消息边界

    a、发送端4个512字节数据块

    b、接收端1个2048字节数据块

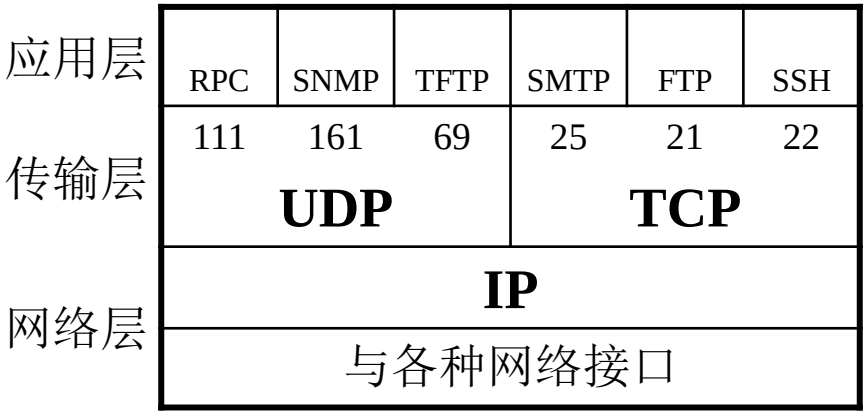
端口	协议	用途
20、21	FTP	文件传输
22	SSH	远程控制台（替代远程登录Telnet）
25	SMTP	电子邮件
80	HTTP	万维网
110	POP-3	远程邮件访问
143	IMAP	因特网邮件访问
443	HTTPS	安全web（SSL/TLS上的HTTP）
543	RTSP	媒体播放控制
631	IPP	打印共享

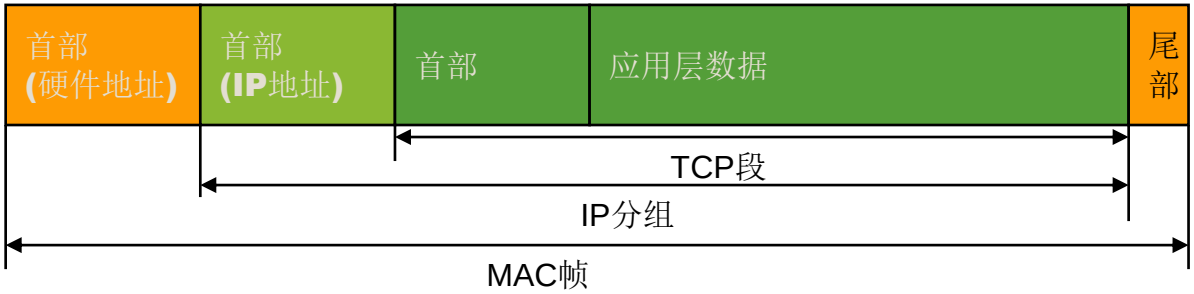


- 紧急数据（如Ctrl+C）

TCP服务模型(2)

- TCP与UDP
- 端口：16b
- 熟知端口：<256





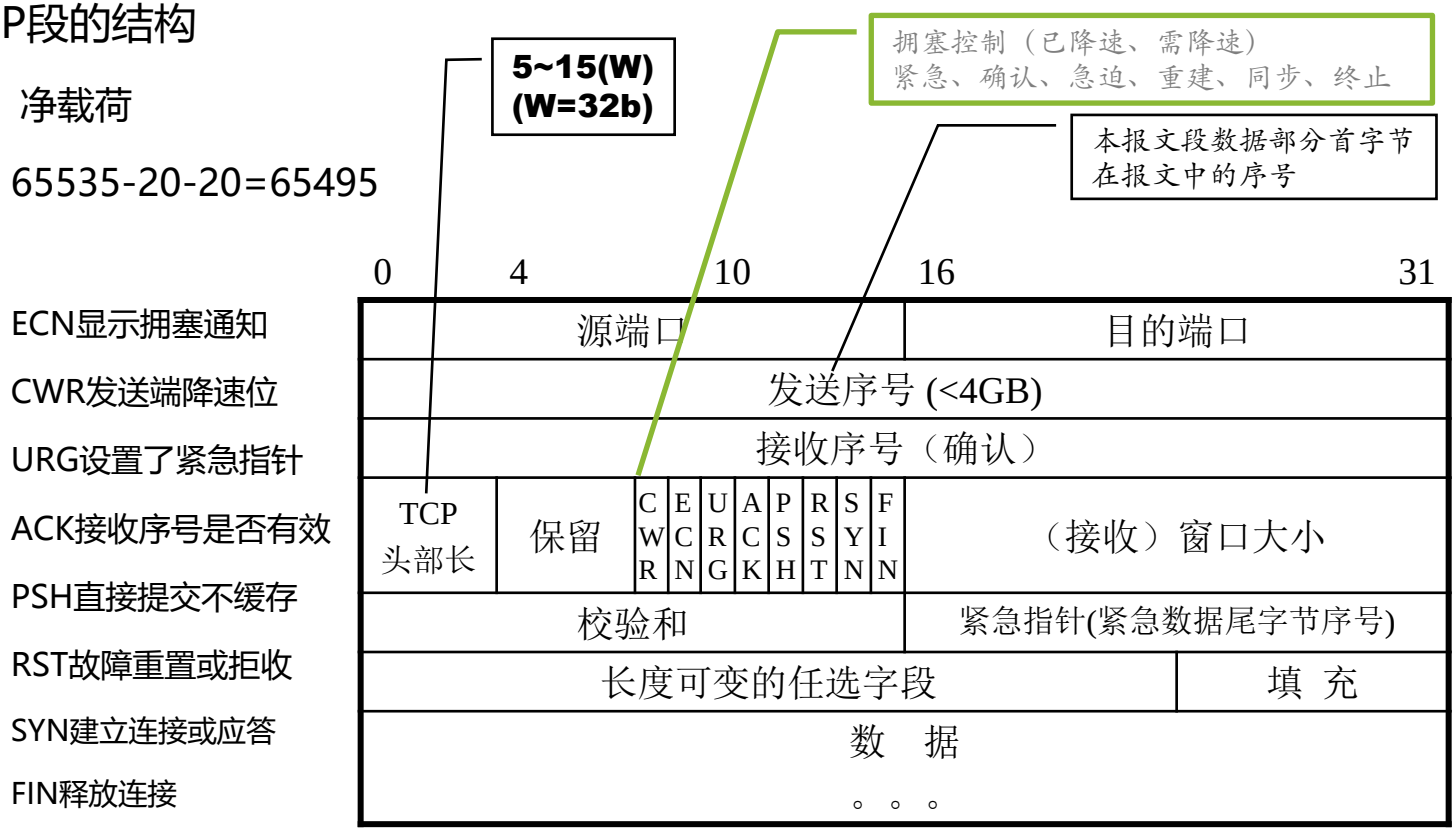
TCP协议

- TCP段
  - 20字节的头
  - IP有效载荷：65515B
  - 网络最大传输单元（MTU）
  - 路径MTU发现

TCP段的头

- TCP段的结构
  - 净载荷

    65535-20-20=65495



ECN显示拥塞通知

CWR发送端降速位

URG设置了紧急指针

ACK接收序号是否有效

PSH直接提交不缓存

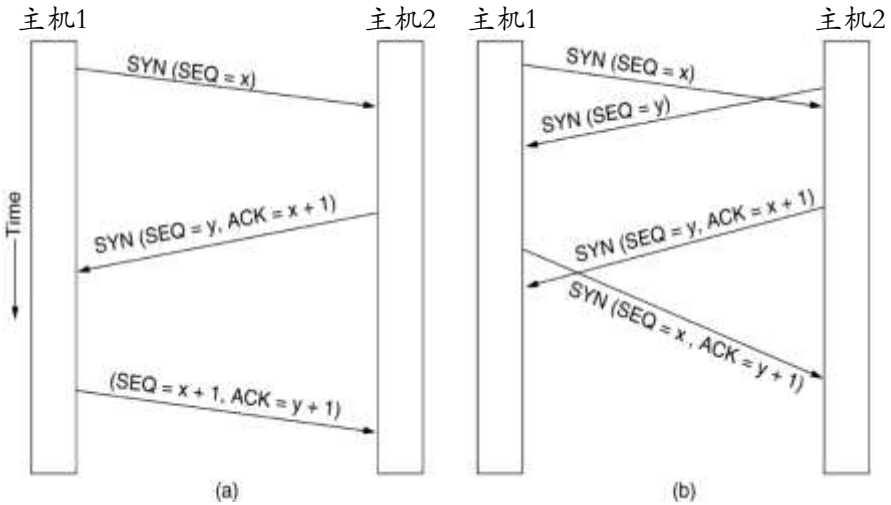
RST故障重置或拒收

SYN建立连接或应答

FIN释放连接

TCP连接建立

- 建立连接采用三次握手方法
  - a.建立（x，y）连接
  - b.还是建立（x，y）



- SYN泛洪攻击
  - 类似DDOS拒绝服务攻击，因为SYN_timeout只能应付少量半连接
  - 应答方必须记忆响应SYN的序号，因而占用主机资源
  - 解决方法：SYN Cookie

    由DDOS防护系统在SYN_ACK中放入加密的y序号，对方ACK=y+1则响应

TCP连接释放

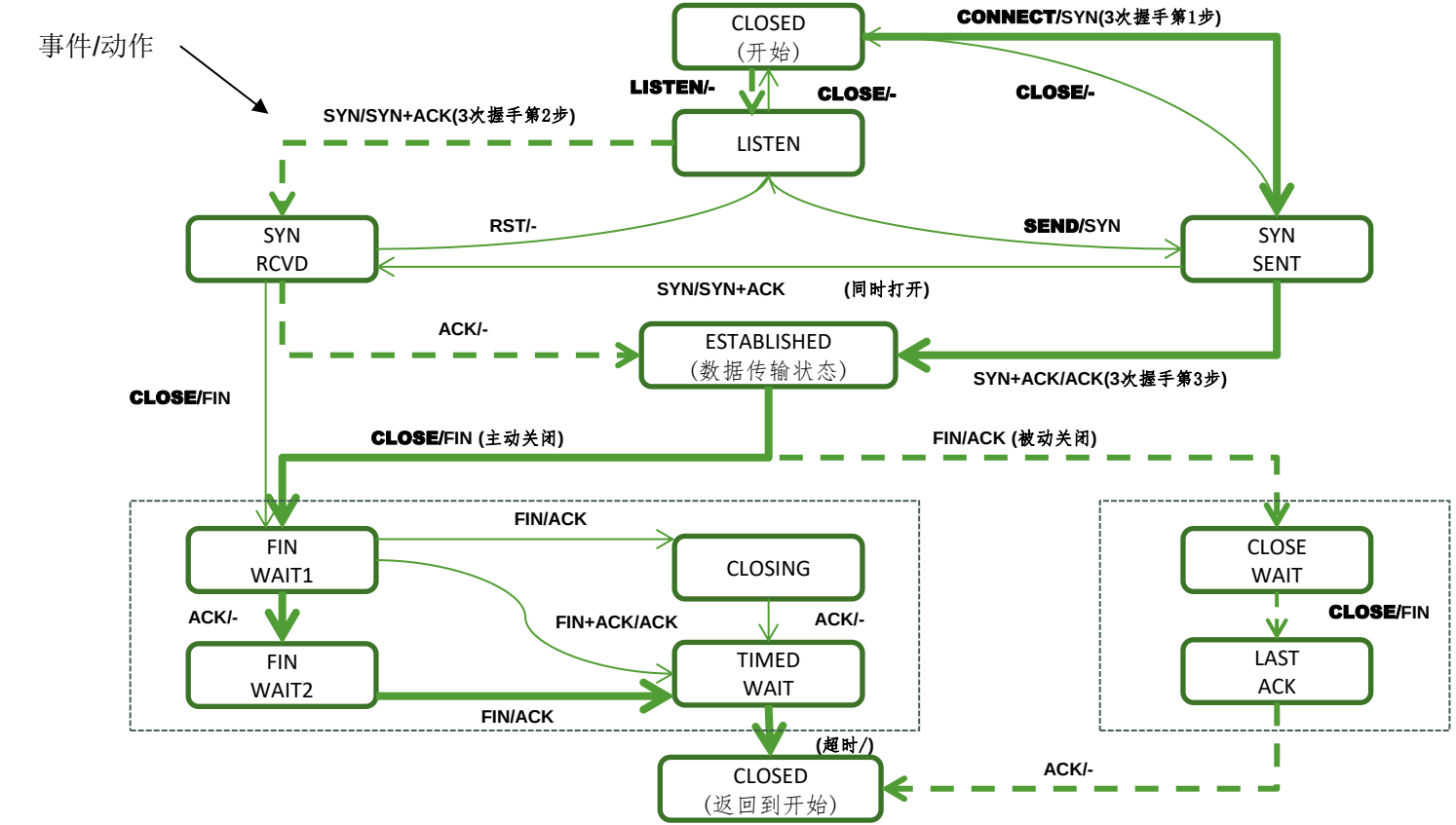
- 释放连接
  - 发送设置了FIN标志位的TCP段
  - 三个TCP段
    - FIN
    - ACK+FIN
    - ACK

TCP连接管理模型

- TCP连接管理有限自动机状态

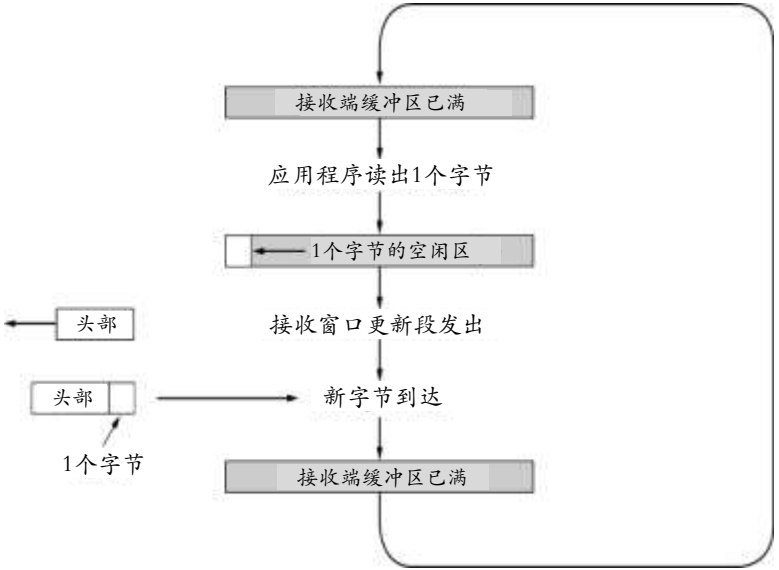
状态	描述
CLOSED	没有连接，挂起
LISTEN	服务器等待入境呼叫
SYN RCVD	连接请求已到达，等待ACK
SYN SENT	应用层开始打开一个链接
ESTABLISHED	正常数据传送状态
FIN WAIT 1	应用层已完成数据发送
FIN WAIT 2	另一端同意释放连接
TIMED WAIT	等待所有数据包消亡
CLOSING	两端同时关闭连接
CLOSE WAIT	另一端已发起关闭连接
LAST ACK	等待所有数据包消亡

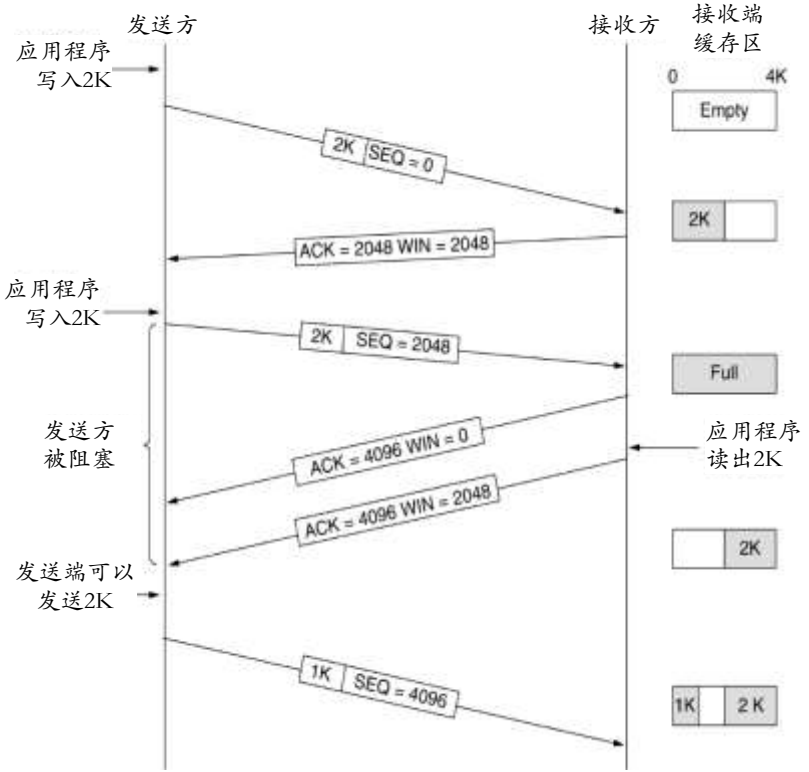
TCP连接管理有限自动机



TCP滑动窗口

- TCP滑动窗口
  - 接收端窗口4096B
- 低能窗口





TCP计时器管理

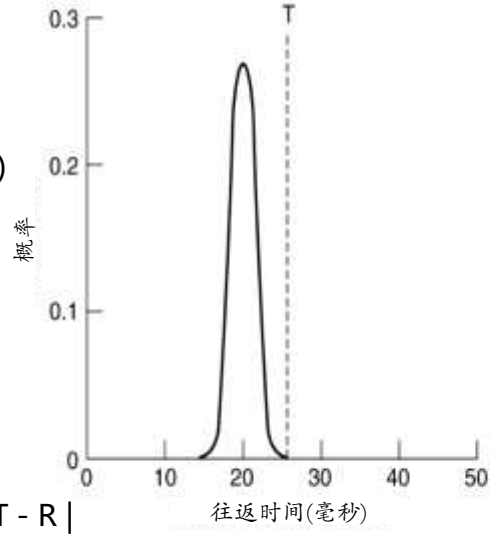
- 重传定时器
  - 平滑往返时间SRTT

    SRTT = αSRTT + (1-α)R （一般α=7/8）

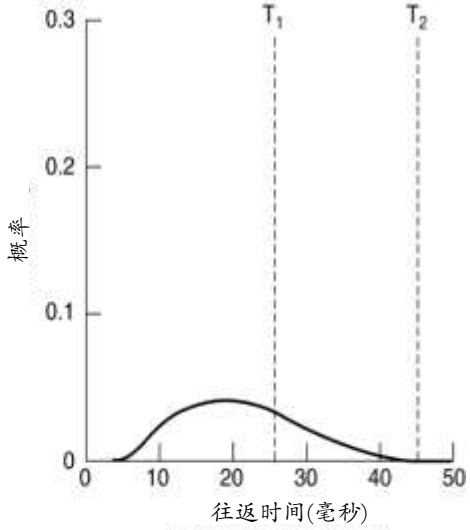
    本次测量确认所花时间R
  - 超时值
    - 原来 = 2×RTT
    - 现在：往返时间变化RTTVAR

      RTTVAR = β RTTVAR + (1- β) | SRTT - R |

      超时值: RTO = SRTT + 4* RTTVAR



(a)数据链路层的确认到达时间概率密度

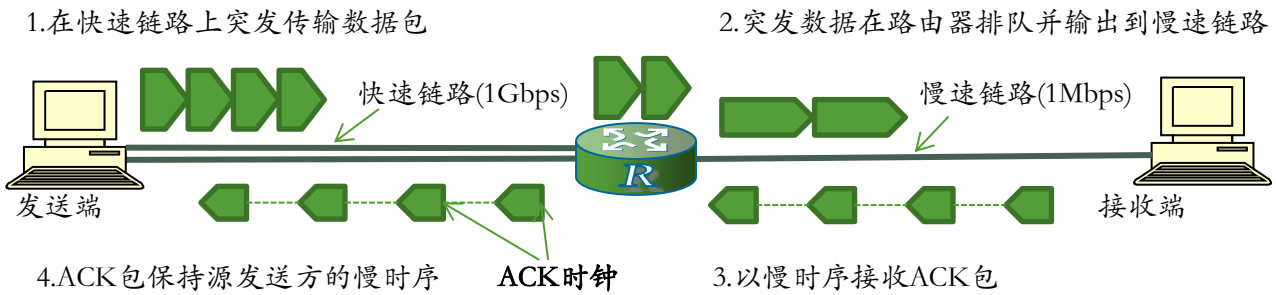


（b）TCP的确认到达时间的概率密度

- 持续计时器：探询接收方，返回窗口大小
- 保活计时器

TCP拥塞控制

- 拥塞控制
  - TCP在拥塞控制和可靠传输中发挥主要作用
- 拥塞窗口
  - 采用AIMD(加增乘减)规则调整窗口大小
  - 流量控制窗口 = min(拥塞窗口，接收窗口)
- 发送与网络带宽匹配（千兆以太网与ADSL）
  - 小的突发包
  - 确认时钟

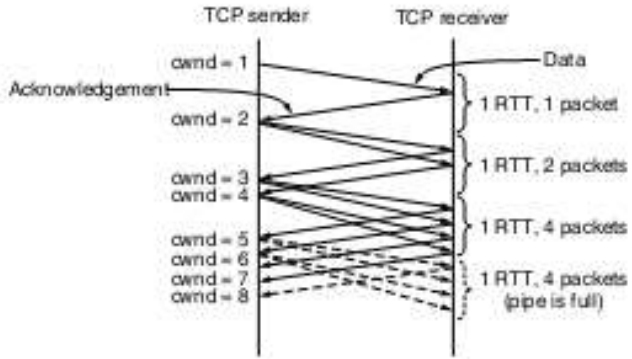


TCP拥塞控制(2)

- 拥塞窗口计算
  - 例：10Mbps网络，RTT=100ms
    - 拥塞窗口 = 带宽延迟积 = 1Mb = 100(个)*1250(字节)
    - 拥塞窗口从1开始，每RTT加1，100 * RTT = 10秒，才能达到100
- 慢速启动
  - 拥塞窗口(cwnd)初始值

    早期为1个段，后来根据经验改为4个段
  - 发送端:

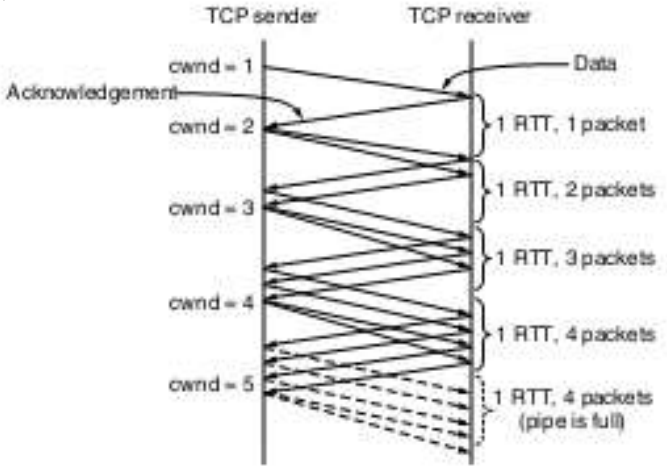
    按初始值发送段，每收到一个未超时确认，cwnd++
  - 拥塞窗口指数增长，慢速启动控制
    - 慢启动阈值(或阀值)



慢速启动增长

TCP拥塞控制(3)

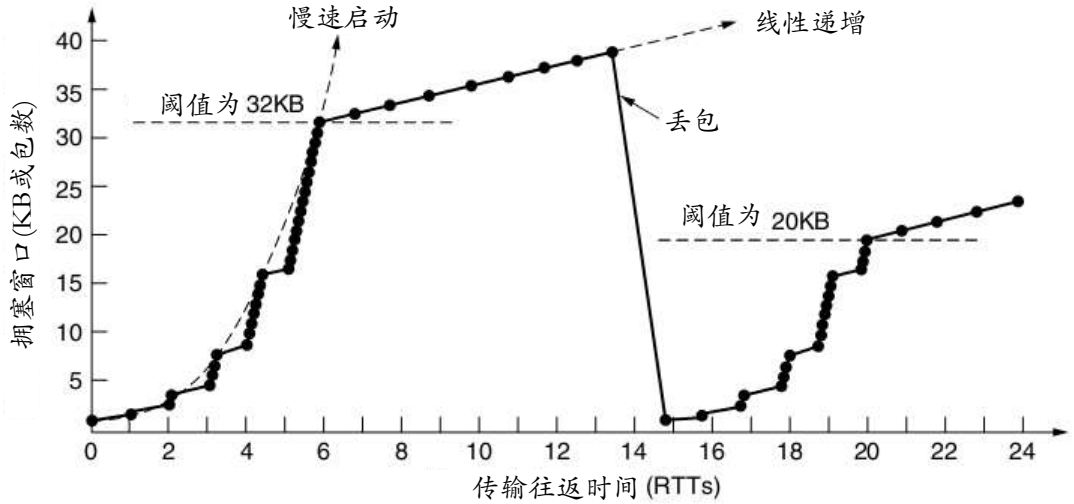
- 慢速启动阈值
  - 慢速启动阈值初始为流量控制窗口大小
  - 超时后，阈值变为当前拥塞窗口的一半，然后重新启动
  - 一旦慢速启动超过阈值，TCP从慢启动切换到线性增加
    - 每一RTT，cwnd++
  - 方案缺陷
    - 等待超时，超时时间太长，
    - 数据包丢失后，接收方确认号不变，
    - 发送端因拥塞窗口已满，无法发送新包
  - 重复确认
    - 三个重复确认意味着丢包
    - 称为快速重传，重传计时器超时前重发



线性增长

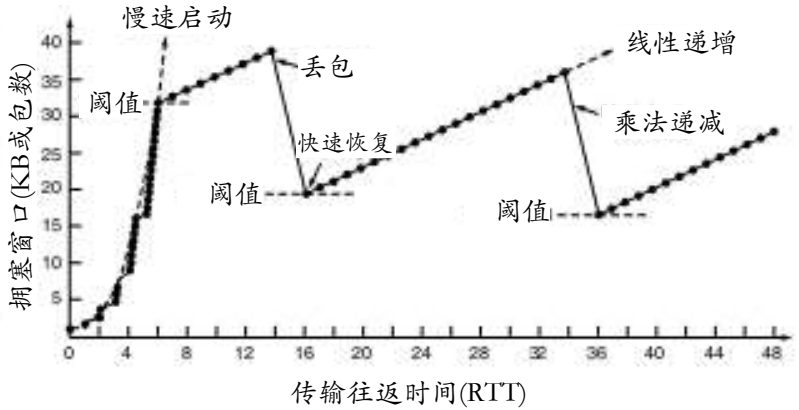
TCP拥塞控制(4)

- TCP Tahoe
  - 1988年发布的4.2BSD TCP Tahoe
  - 最大段长：1KB
  - 拥塞窗口初值64KB，发生拥塞减半为32KB

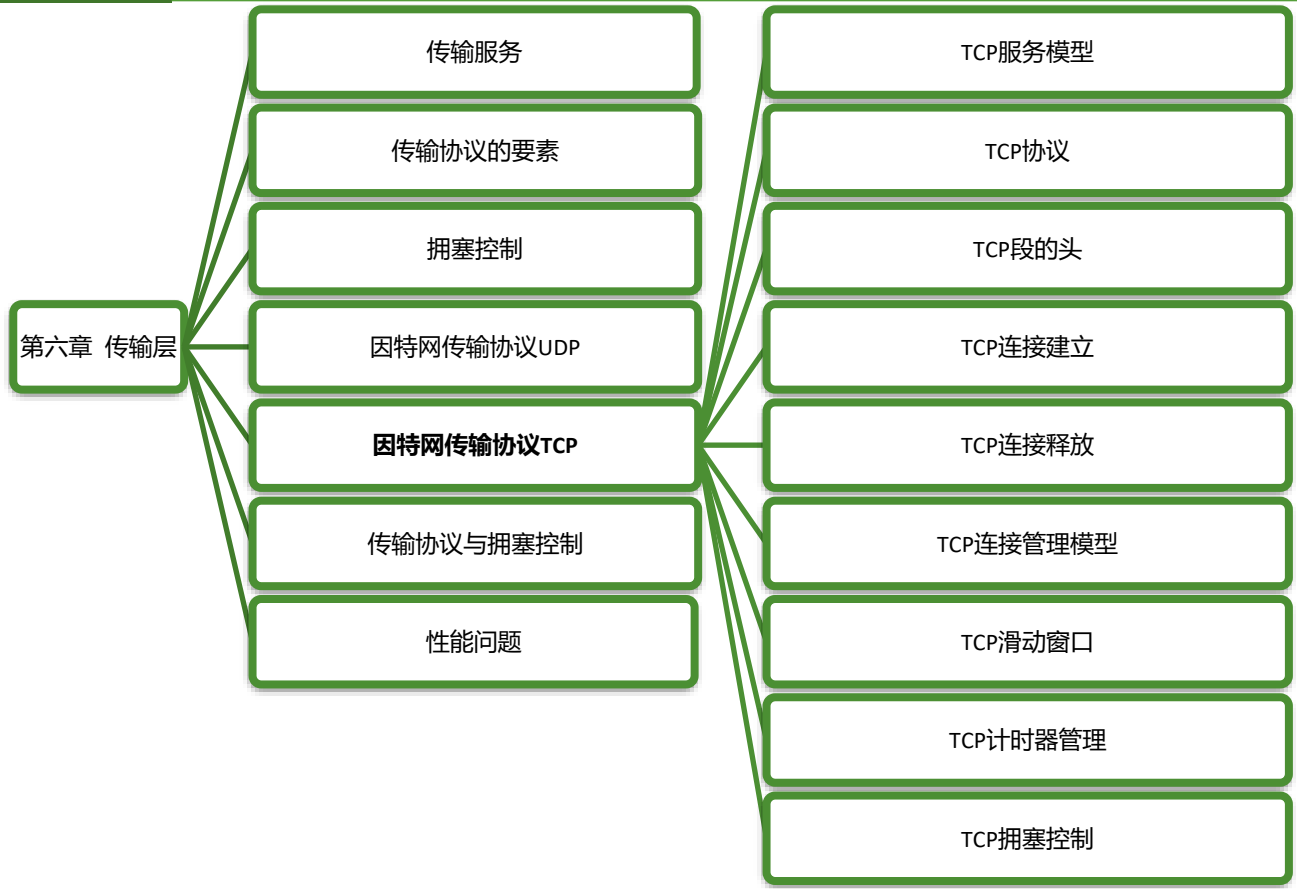


TCP拥塞控制(5)

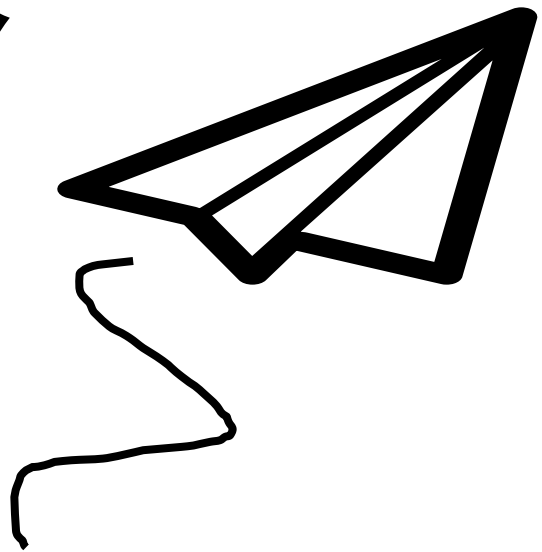
- TCP Reno
  - 快速恢复
    - 三个重复确认
  - 1990年发布
    - 4.3BSD Reno
- TCP的变化（TCP NewReno）
  - 选择确认SACK
  - 加入拥塞标志
    - ECE（显示拥塞通知ECN Echo），CWR（拥塞窗口减少）
- TCP CUBIC（对于大的带宽延迟积，拥塞窗口为自最后重复确认以来的时间的函数）



本章导航与要点


第六章 传输层
传输服务
传输协议的要素
拥塞控制
因特网传输协议UDP
因特网传输协议TCP
传输协议与拥塞控制
性能问题
TCP服务模型
TCP协议
TCP段的头
TCP连接建立
TCP连接释放
TCP连接管理模型
TCP滑动窗口
TCP计时器管理
TCP拥塞控制

本节课程结束



6.6 传输协议与拥塞控制

- QUIC：快速UDP互联网连接
- BBR：基于瓶颈带宽的拥塞控制
- TCP的未来

QUIC：快速UDP互联网连接

- QUIC用于改进TCP的某些吞吐量和延迟特征
- Chrome浏览器支持
  - QUIC运行在UDP之上，使得Web协议运行得更为快速

BBR：基于瓶颈带宽的拥塞控制

- 缓冲区膨胀：设备缓冲区太大，导致TCP（大拥塞窗口）发送方，超网络容量发送，导致后面的缓冲区填满
  - 针对发送太快的拥塞事件不能及时反馈
  - 数据包排列在大缓冲区后面的所有发送方，网络延迟大幅增加
- 解决方法：减小网络设备的缓冲区；替换基于数据包丢失的拥塞算法

BBR：基于瓶颈带宽的拥塞控制（2）

- BBR主要思路
  - 测量瓶颈带宽和往返传播延迟，按照推算出的操作点时速率发送数据
  - 带宽延迟积之前，往返时间不变，之后传递速率与往返时间仍然成反比
  - 超过带宽延迟积后，往返时间递增，传递速率增长停滞

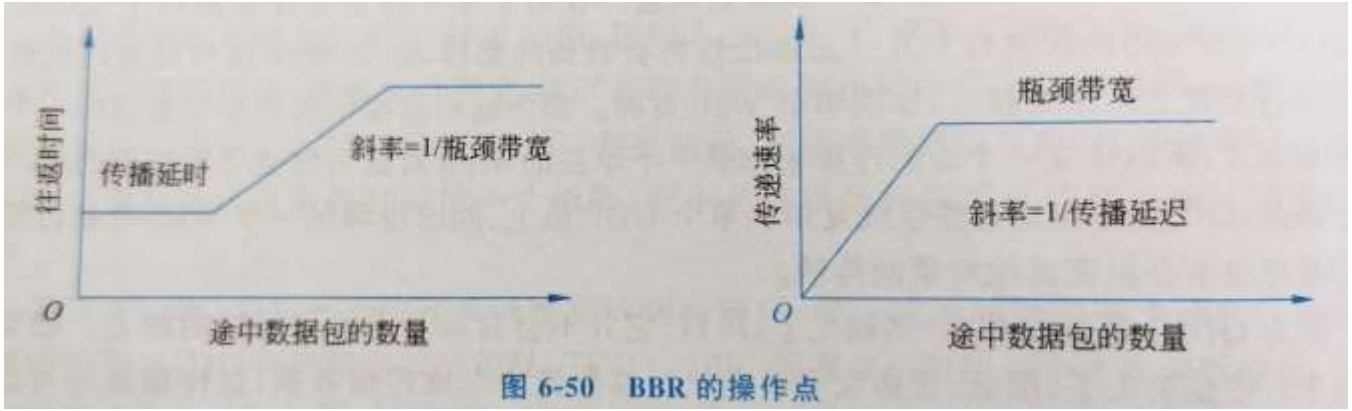


图 6-50　BBR 的操作点

TCP的未来

- 功能扩张
  - 需要提供传输语义：
    - 应用保留已发送的消息、记录的边界；Web从同一服务器传输多个对象等
    - SCTP流控制传输协议，SST结构化流传输
  - 拥塞控制：改变算法
    - 原TCP以数据包丢失为拥塞信号，随着速度加快，数据包丢失率急剧下降
    - 新算法采用其他拥塞信号，如往返时间增加

6.7 性能问题

- 计算机网络中的性能问题
- 网络性能测量
- 针对快速网络的主机设计
- 快速处理段
- 头压缩
- 长肥网络的协议

计算机网络中的性能问题

- 广播风暴
  - UDP广播风暴
- 同步触发过载
  - 同时访问某服务器，如DHCP、FTP等
- 缓冲区问题
  - 带宽延迟乘积 = 带宽(b/s) * 往返传输延迟(s)
- 超时间隔的设置
- 抖动问题

网络性能测量

- 确保样本空间足够大
- 确保样本具有代表性
- 缓存可以破坏测量结果
- 确保测试期间不会发生不可知事情
- 小心使用粗粒度时钟
- 小心推断结果

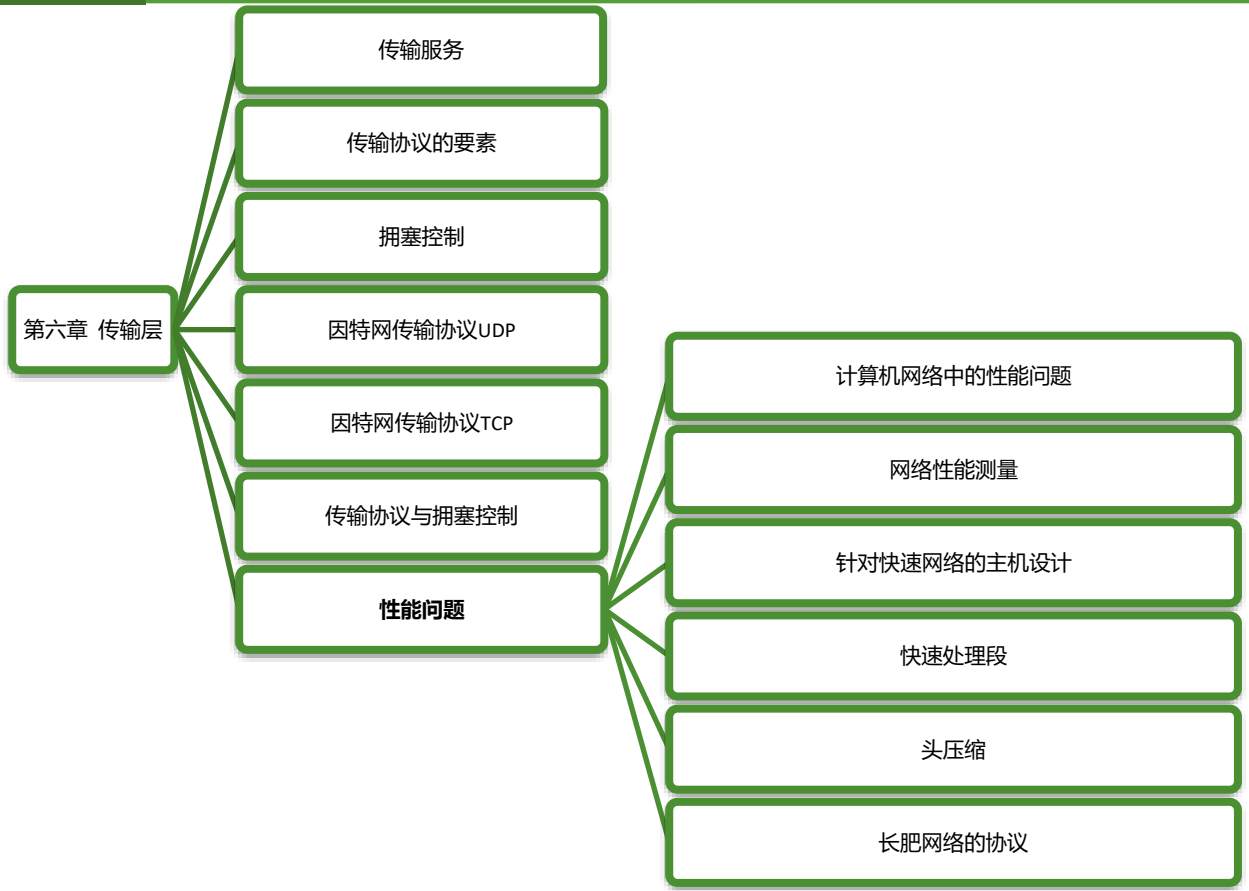
针对快速网络的主机设计

- 主机速度比网络速度重要
- 减少包技术来降低开销
  - 减少段数
- 最小化数据预取
  - 多个层次结合处理
- 最小化上下文切换
- 避免拥塞比从中恢复好
- 避免超时

其他性能问题

- 快速处理段
  - 连续段的头几乎相同
  - 头预测、计时轮
- 头压缩：专门协议减少高层协议头字节数
- 长肥网络的协议
  - 2^32：1周(56kbps)、57分钟(10Mbps)、34秒(1Gbps)
  - 流量控制窗口：带宽延迟乘积
  - 重传策略：选择重传协议
  - 千兆网：延迟的制约超过带宽
  - 通讯速度远超计算速度

本章导航与要点


第六章 传输层
传输服务
传输协议的要素
拥塞控制
因特网传输协议UDP
因特网传输协议TCP
传输协议与拥塞控制
性能问题
计算机网络中的性能问题
网络性能测量
针对快速网络的主机设计
快速处理段
头压缩
长肥网络的协议

本章课程结束