现代计算机网络

李伟明 lwm@hust.edu.cn

Ch.3 拥塞控制

- 3.1拥塞控制相关基础
  - 拥塞控制的基本原理
  - 拥塞控制方法分类
  - 网络模型及特点
  - 资源分配机制
  - 资源分配评价标准
- 3.2排队规则与流量整形
  - FIFO
  - 优先排队
  - 公平排队FQ
  - 加权公平排队WFQ
  - 流量整形
  - 令牌桶算法
- 3.3 TCP拥塞控制
  - 加乘式
  - 慢启动
- 3.4 TCP拥塞避免
  - DECbit
  - RED和源拥塞避免

3.1 拥塞控制相关基础

3.1.1 拥塞控制的基本原理

- 问题：因资源竞争的用户
  1. 怎样有效和公平地分配有限资源;
  2. 怎样共享资源：包括链路带宽、路由器、交换机中的缓冲区（包在此排队等待传输）和处理机时间等
- 拥塞（congestion）
  - 当过多的包在网络缓冲区中竞争某个相同链路时，队列会溢出丢包，当这种丢包成为普通事件时，则称网络发生拥塞

$$\sum_{i=1}^{n} \text{被请求资源}_i > \sum_{j=1}^{m} \text{能可用资源}_j$$

  - 对聚合带宽的需求超过了链路的可用容量

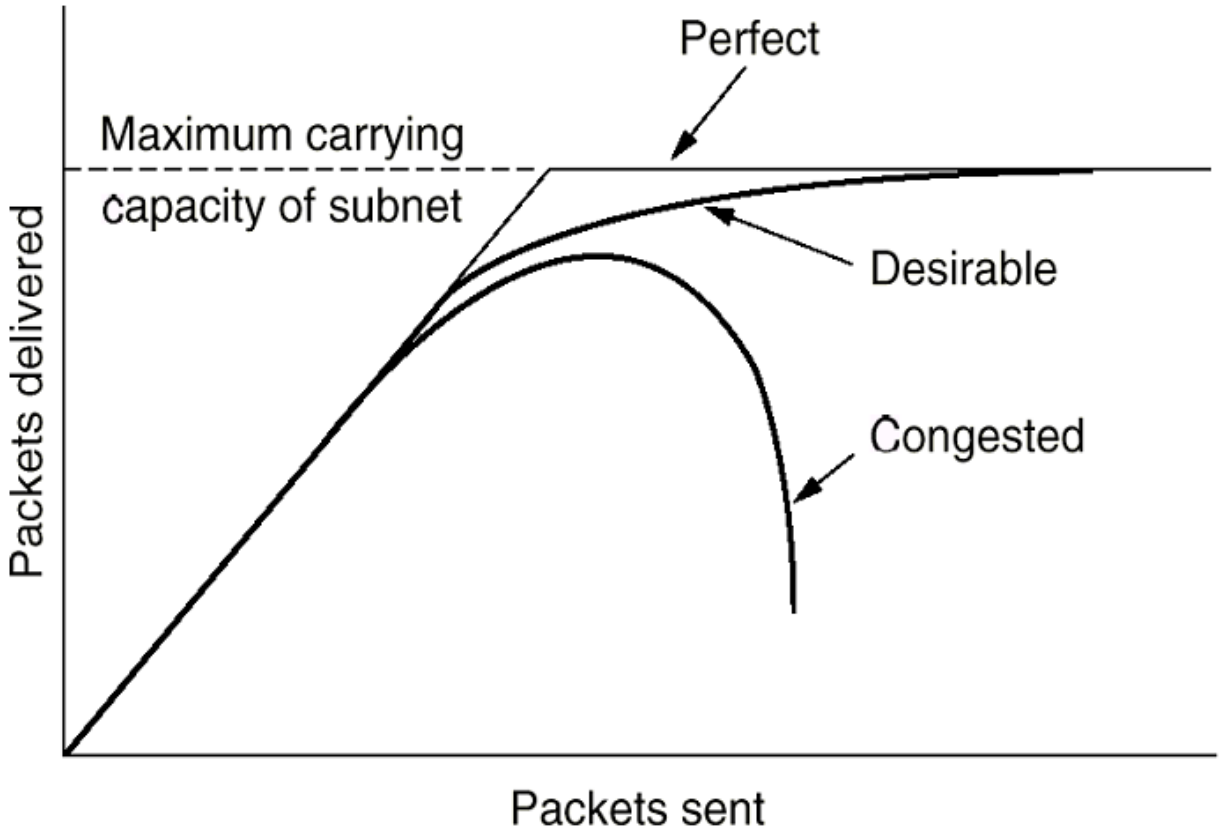


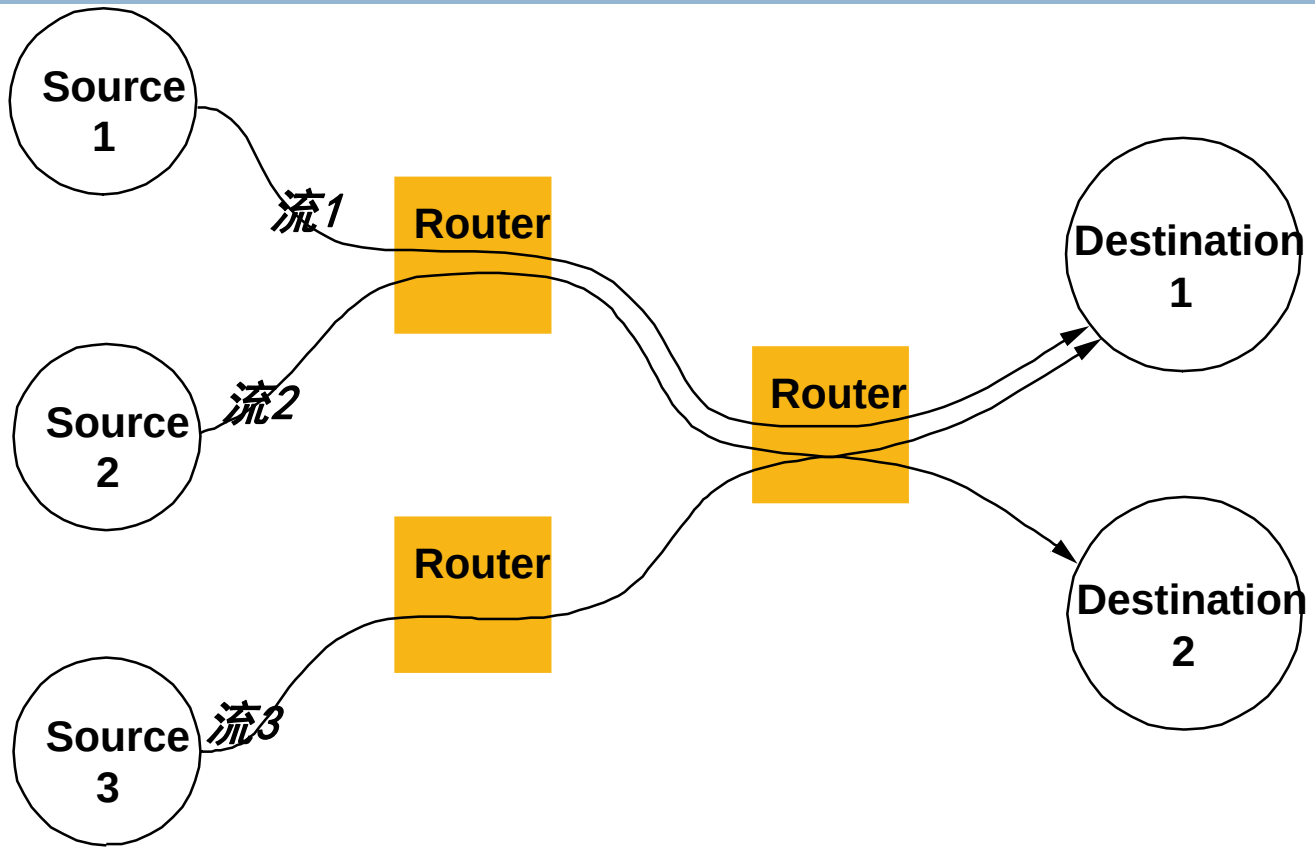
Fig. 3.1 超载下的拥塞与性能下降

- 拥塞产生的原因（需求大于供给，无准入控制）
  - 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
- 拥塞的后果：网络性能下降
  - 多个包丢失，链路利用率低下
  - 全网同步振荡，吞吐量下降
  - 高排队延迟，拥塞崩溃

- 拥塞控制和资源分配是一个硬币的两面(教科书)：
  - 资源分配就是尽量合理的满足用户对网络资源的需求，但是总会有满足不了的情况
  - 拥塞控制就是网络主机和路由器防止网络过载的情况采取的措施。这种措施可以是阻止几个用户使用网络，更好的方式是让全网都来减少流量，所以拥塞控制经常包含资源分配方法。

资源分配和拥塞控制基本单位：数据流

- 流是在源目主机对之间，通过相同路由而发送的一系列包，在路由器中的资源分配中这是一个重要的抽象
- 流过同一路由器，且具有若干相同头部特征的一系列数据包。这些特征是：源地址、目地址、源端口号、目端口号、协议类型、服务类型TOS、输入端逻辑接口等
- “通道”是端-端连通的一种抽象，“流”对网络路由器是可见的。


Source 1
流1
Router
Destination 1
Source 2
流2
Router
Router
Destination 2
Source 3
流3

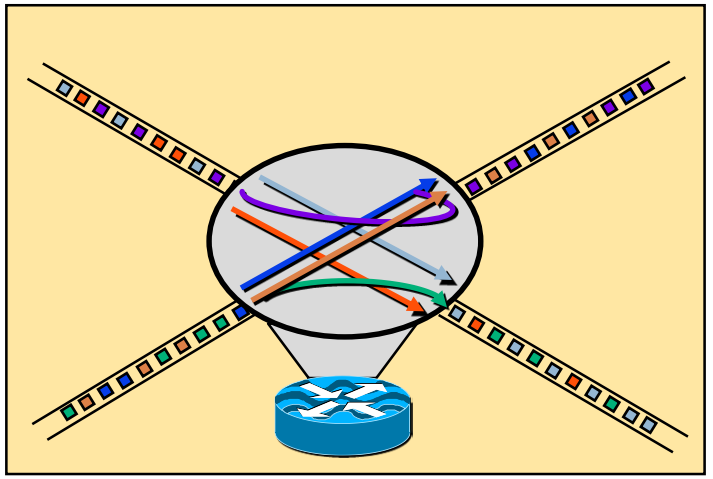
多个流通过一组路由器

什么是NetFlow

- 1996年，Cisco系统公司的Darren Kerr和Barry Bruins开发了一种流量监控技术，即NetFlow。
- 目前Cisco的绝大多数路由器集成了该特性，Juniper、Extreme以及其他厂商也有集成该特性的路由器和交换机

流记录（flow record）

- 记录一段时间内网络的某个观测点所通过的一系列分组（packets）的相关信息
- 通常的流记录分类依据由一个五元组（即源地址、目的地址、源端口、目的端口和协议类型）所组成。



源地址	目的地址	源端口	目的端口	协议类型	分组数	字节数	其它...

NetFlow Cache Example


```
#show ip cache flow
IP packet size distribution (78630M total packets):
   1-32   64   96  128  160  192  224  256  288  320  352  384  416  448  480
   .002 .448 .062 .027 .013 .011 .008 .011 .003 .003 .002 .006 .005 .003 .002
    512  544  576 1024 1536 2048 2560 3072 3584 4096 4608
   .002 .003 .015 .033 .331 .000 .000 .000 .000 .000 .000
IP Flow Switching Cache, 6553988 bytes
  32929 active, 32607 inactive, 524367786 added
  4111490554 ager polls, 0 flow alloc failures
  Active flows timeout in 30 minutes
  Inactive flows timeout in 15 seconds
IP Sub Flow Cache, 794824 bytes
  32895 active, 16257 Inactive, 519171584 added, 519168554 added to flow
  0 alloc failures, 12911870 force free
  3 chunks, 1155 chunks added
last clearing of statistics never
--More—
```

启发是提高小报文性能

Active：表示当前的流数量；inactive表示仅仅分配数据结构；added表示总的创建流数量

Never：表示没有清理过flow统计

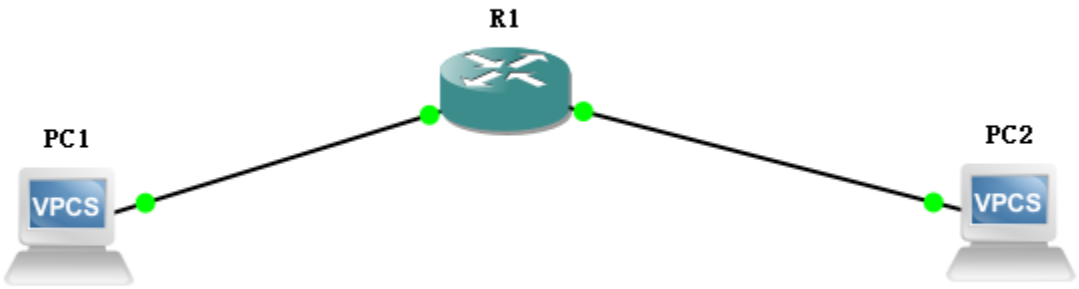
NetFlow Cache Example (cont.)

Protocol	Total Flows	Flows /Sec	Packets /Flow	Bytes /Pkt	Packets /Sec	Active(Sec) /Flow	Idle(Sec) /Flow
TCP-Telnet	3833510	0.8	10	179	9.2	9.0	26.8
TCP-FTP	12511306	2.9	6	132	19.7	6.3	16.5
TCP-FTPD	1194796	0.2	544	866	151.5	86.7	21.2
TCP-WWW	944754736	219.9	13	627	2871.0	3.2	23.7
TCP-SMTP	53320030	12.4	14	399	185.8	6.6	19.2
TCP-X	913841	0.2	41	631	8.9	19.2	24.5
TCP-BGP	1867	0.0	1	49	0.0	0.5	20.5
TCP-NNTP	1086658	0.2	252	874	63.8	15.2	26.8
TCP-Frag	228697	0.0	9	131	0.5	6.5	25.3
TCP-other	2264274585	527.1	23	568	12466.6	12.9	24.4
UDP-DNS	231113128	53.8	2	79	114.7	3.6	26.0
UDP-NTP	6394017	1.4	3	76	5.2	9.7	27.2
UDP-TFTP	13567	0.0	1	95	0.0	3.1	29.3
UDP-Frag	211973	0.0	3165	1266	156.2	116.2	27.5
UDP-other	1177902953	274.2	5	293	1385.8	6.1	25.8
ICMP	103453714	24.0	2	62	57.9	3.8	26.0
IGMP	726	0.0	2	300	0.0	3.5	29.1
IPINIP	1	0.0	1	40	0.0	0.0	27.2
GRE	10272668	2.3	309	774	740.5	35.3	23.1
IP-other	544060	0.1	533	484	67.5	123.7	22.6
Total:	4812026833	1120.3	16	567	18305.6	8.7	24.7

NetFlow Cache Example (cont.)

GNS3的一个优点就是可以模拟Cisco的路由器，在下面的拓扑中，可以观察PC1到PC2的Newflow：

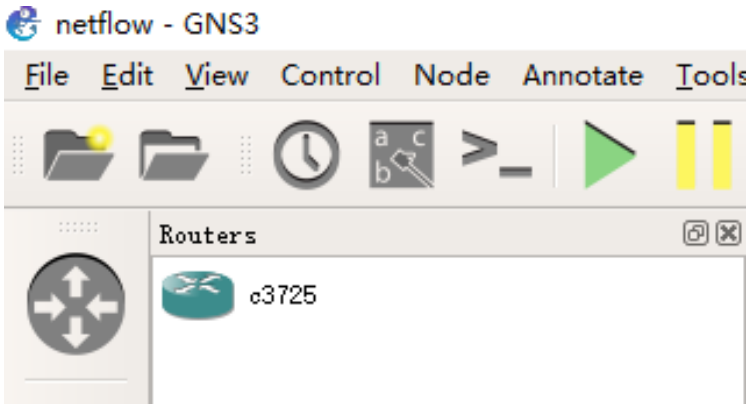
- r1(config)#ip cef
- r1(config)#int f0/0
- r1(config-if)#ip route-cache flow



NetFlow Cache Example (cont.)

GNS3的一个优点就是可以模拟Cisco的路由器：

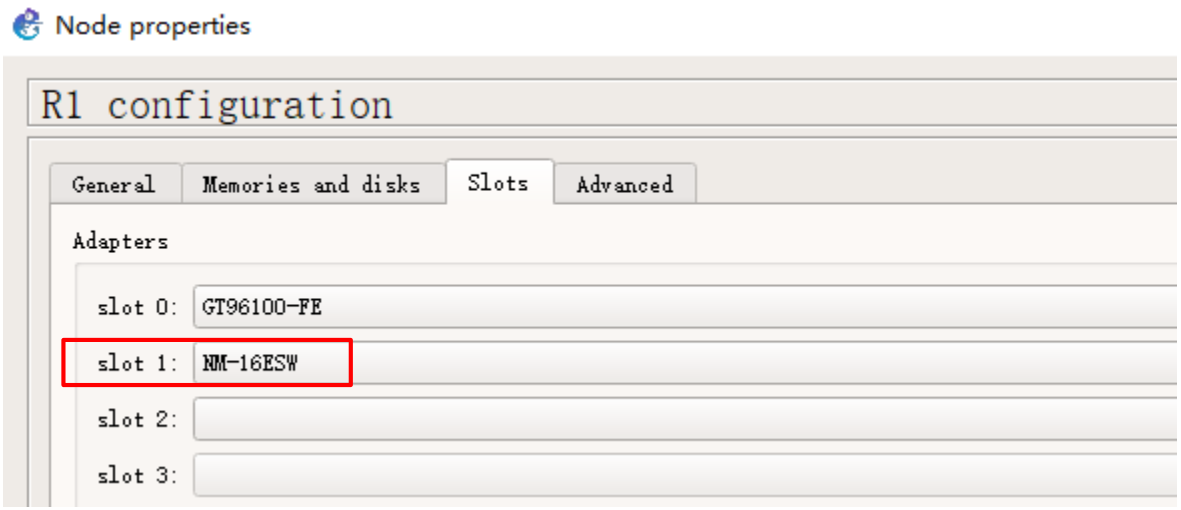
- GNS3可以用于体验Cisco网际操作系统IOS或者检验将要在真实的路由器上部署实施的相关配置。
- GNS3整合了Dynamips，这是一个可以让用户直接运行Cisco系统(IOS)的模拟器
- 简单说来是GN3是Dynamips的一个图形前端，相比直接使用Dynamips这样的虚拟软件要更容易上手和更具有可操作性。



NetFlow Cache Example (cont.)

GNS3可以模拟Cisco的路由器：

- 但是GNS3仅仅支持Cisco路由器的模拟，不支持交换机模拟
- 但是可以在路由器插入一个交换板，实现交换机功能：



路由器中面向TCP流分类

- TCP-friendly流
  - 采用了类似TCP拥塞控制机制的流，即同质或大体上在相似环境下按可比条件（丢包率、RTT、MTU）不会占用比TCP流更多带宽的流
- 非适应（unresponsive）流
  - 不采用拥塞控制机制，不能对拥塞作出反应的流，如多媒体应用和组播应用、UDP、AH、ESP 等
- 有适应性但非TCP-friendly流
  - 使用拥塞控制算法，但非TCP友好。如：接收端驱动分层组播（Receiver-driven Layered Multicast RLM）采用的就是这种算法。
  - 使用non-TCP的拥塞控制算法。如修改TCP，使窗口的初始值很大并且保持不变，即所谓的"快速TCP"。

拥控与流控的区别

- 路由器中拥塞控制的必要性
  - 不遵守TCP拥塞控制机制的应用进一步加剧因特网范围内拥塞崩溃的可能性
  - TCP拥塞控制还存在着自相似、效率、公平性等方面的问题
- 拥控：
  - 问题：某些点上存在资源瓶颈
  - 防止发送者把太多的数据发送到网路中，适应瓶颈链路和路由器有限buffer，保护网络
- 流控：
  - 问题：接收方可能存在缓存不足、进程等待
  - 防止发送方的发送速度比接收方的接收速度快，适应收发双方的buffer＋cpu能力，保护端点。

拥塞控制与流量控制的区别

- 拥塞控制（congestion control）与全网有关，涉及多个端到端、主机、路由器等很多网络元素；目的是确保通信子网能够承载用户提交的通信量，是一个全局问题，
- 流量控制（flow control）只与一对端到端的通信量有关，只涉及快速发送方与慢速接收方的问题，是局部问题，一般都是基于反馈进行控制的
- 要防止这两者的混淆，但它们有些机制是相同的

影响拥塞的网络策略

Layer	Policies
Transport	• Retransmission policy • Out-of-order caching policy • Acknowledgement policy • Flow control policy • Timeout determination
Network	• Virtual circuits versus datagram inside the subnet • Packet queueing and service policy • Packet discard policy • Routing algorithm • Packet lifetime management
Data link	• Retransmission policy • Out-of-order caching policy • Acknowledgement policy • Flow control policy

3.1.2 解决拥塞：资源分配机制

解决拥塞本质上是更合理的资源分配：

- 以路由器为中心与以主机为中心
- 基于预留与基于反馈
- 基于窗口与基于速率

1）R/H为中心的分配

- 以R为中心，链路算法(Link Algorithm). 在网络设备中(如路由器和交换机) 执行,作用是检测网络拥塞的发生,产生拥塞反馈信息
  - R决定何时转发包，决定丢掉哪个包，
  - 通知正产生网络流量的那些主机允许它们发送包的数目
- 以H为中心，源算法(Source Algorithm).源算法在主机和网络边缘设备中执行,作用是根据反馈信息调整发送速率
  - 端主机观察网络状态（多少个包成功通过了网络）
  - 并因而调整其行为
- 二者并不互斥，
  - R中心的拥塞管理仍需要H执行R发送的建议消息
  - H中心的拥塞管理业需要R丢弃R溢出时那些包

链路拥塞控制算法

- 链路算法的研究目前主要集中在主动队列管理AQM算法方面.
- 下表给出主动队列管理算法的简单描述

主动队列管理算法	基本内容描述
RED	比例控制器＋低通滤波器
ARED	根据网络负载的情况调整标记/丢失概率.
SRED	通过估计网络中流的个数来调整报文标记/丢失概率.
BLUE	以“分组丢失率”和“链路有效利用率”作为拥塞是否发生的标准.
REM	利用了网络流量优化理论中“代价”的概念来探测和控制网络拥塞.
AVQ	使用PI控制器,控制的是队列的入口速率.
PI	使用PI控制器,控制的是队列长度

源拥塞控制算法-TCP层

- 源算法中使用最广泛的是TCP协议中的拥塞控制算法。TCP是目前在Internet中使用最广泛的传输协议.
- 下表给出拥塞控制源算法的简单描述

拥塞控制源算法	描述
Tahoe-TCP	慢启动、拥塞避免、快速重传-早期较为普遍采用版本
Reno-TCP	多一个：快速恢复
NewReno-TCP	引入了部分确认和全部确认的概念.
SACK-TCP	规范了TCP中带选择的确认消息.
Vegas-TCP	采用带宽估计,缩短了慢启动阶段的时间

- 源算法测量网络是否拥塞的参数
  - 缓冲区缺乏造成的丢包率；
  - 平均队列长度；
  - 超时重传包数目；
  - 平均包延迟（**RTT是重要指标**）；
  - 包延迟变化（Jitter）。
- 路由器向源反馈方法
  - 向负载发生源发送一个告警包；
  - 包结构中保留一个位或域用来表示发生拥塞，一旦发生拥塞，路由器将所有的输出包置位，向邻居告警；
  - 主机或路由器主动地、周期性地发送探测（probe），查询是否发生拥塞。

2）基于预留与基于反馈

- 基于预留
  - 建立流时，端主机请求网络给予一定的容量，每个R分配足够的资源（缓冲区或链路带宽的比例）来满足这一请求。问题：利用率？
  - 如果由于资源的过量则R不能满足该请求而拒绝该流。这同打电话时碰到忙音一样？
- 基于反馈
  - 端主机首先没有预留任何容量，而按照它们接收到的反馈来调整其发送。
  - 调整或者是显式的，如拥塞R发送一个“请慢下来”消息到主机
  - 或隐式的：如端主机根据外部可观察的网络行为，如包丢失率，来调整其发送率

两种机制的比较

- 基于预留的系统总是意味着
  - 以路由器为中心的资源分配机制：因每个R负责保持跟踪现在有多少容量被预留，并保障在其所做的预留内每个主机是活着的。
  - 如果在作了预留后某主机发送数据快于它曾经所要求的，则该主机的包将是准备丢弃。R也将会拥塞
- 基于反馈的系统
  - 可以既是以R为中心也可是以H为中心的机制。
  - 一般如果反馈是显式的，则R至少某种程度就被包括在资源分配模式里；
  - 如果反馈是隐式的，则几乎所有的重担都落在端主机上，当R变成拥塞后就默默地丢包。

3）基于窗口与基于速率

- 不论流控还是拥控，对发送者，都需要一个方法来表达：多少数据要传输？
  - 用窗口大小来描述：如TCP：流控也可用窗口通告机制来预留缓冲空间，来支持资源分配
  - 用速率大小来描述：用速率控制发送者行为，如接收方可说它可处理1Mbps的视频，而发送方则坚持这一速率

小结

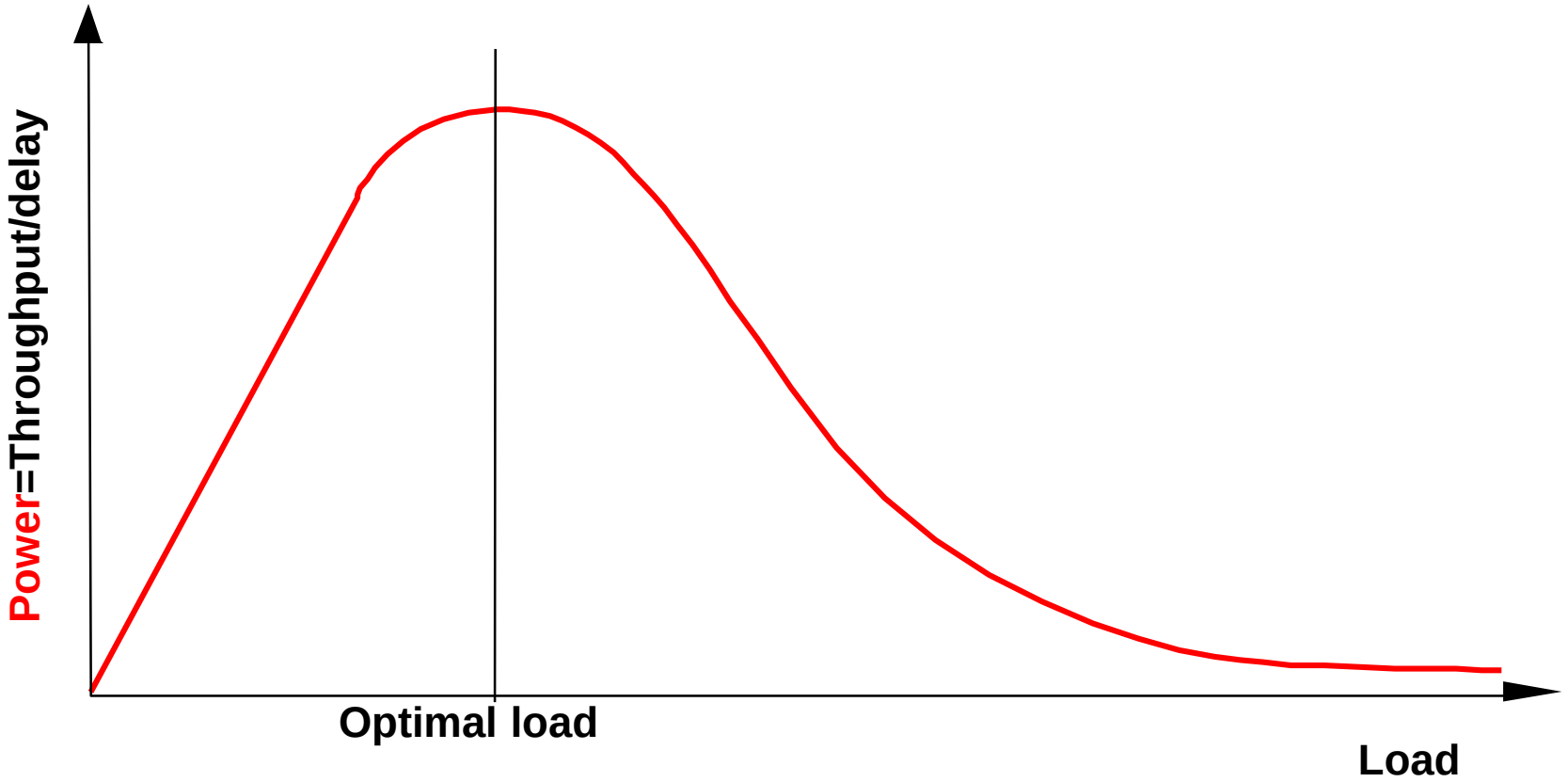
- 3个分类，每类有2种，故共有8种不同策略，而实际上看来只有2个是常用的，并与网络的现行服务模式连在一起
- Best-effort服务模式意味着
  - 是反馈机制，
  - 最终由端用户来解决拥塞，
  - 实际上这样的网络使用基于窗口的信息
- QoS服务模式意味着
  - 资源预留，
  - 需要包含路由器，如排队包的不同取决于它所需预留资源的级别，
  - 这时用速率来表示这种预留是很自然的，因为窗口是针对某带宽而言的

3.1.3 资源分配评价标准

- 怎样评价资源分配机制的好坏
  - 有效性
  - 公平性
- 资源分配的有效性（提高网络效率）
  - 2个主要测量指标：
    - 网络吞吐量（实际传输bps）
    - 延迟
  - 如果有方法，增加吞吐量，提高链路利用率，但是增加了R内队列长度，包平均延迟也变长了，并不是个很好的方法。

网络能力

- 用吞吐率和延迟之比来描述资源分配模式的有效性，该比称为网络能力
  - Power = 吞吐率/延迟
  - Power是加速度量纲 bit/s^2



能力曲线—负载与吞吐/延迟比

目标是最大化吞吐/延迟比

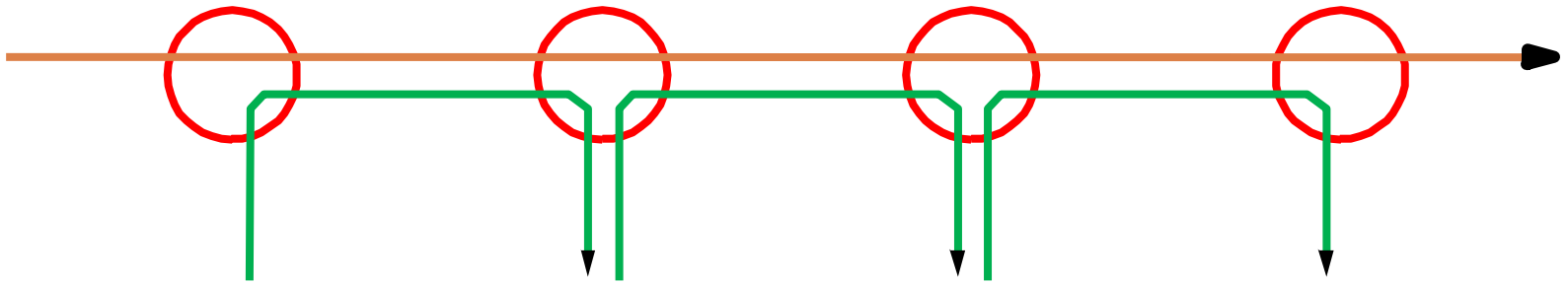
- 网络上有多少负载，这将由资源分配机制来设置。
- 理想情况下应在该曲线峰顶上运行。
- 峰值左边太保守，允许包太少、链路利用率低
- 峰值右边太过分，过多报文容许发送到网络，队列的延迟增加超过了吞吐率大增加
- 一般实行粗略控制

资源分配的公平性

- 网络资源分配还要考虑公平性
- 定义公平性时又陷入什么是公平的困难
  - 基于预约的资源分配机制显然导致了可控制的不公平
  - 该机制可能通过预约使接收视频流在1Mbps，而相同链路中的FTP却只有10Kbps
- 一般公平原则（若没有明确要求）
  - 当某条链路中存在几个流时，对接收每个流都应该有相等的带宽（平均主义）

公平的定义

- 假设公平分享带宽就是带宽相等
  - 即使不预约情况下，平均分配也不等于公平分配，
  - 是否要考虑流的路径长度？如下图，1个4跳流同3个1跳流竞争时，公平性是什么？



Raj Jain 公平指数

- 假设公平性是指带宽相等且所有通道长度相等，给定一组流的吞吐量（$x_1,x_2,\ldots,x_n$）[用单位bps测量]
- 赋予0－1的公平性指数的函数（n数和之平方/n数平方和之n倍）

$$f(x_1,x_2,\ldots,x_n)=\frac{(\sum_{i=1}^{n}x_i)^2}{n\sum_{i=1}^{n}x_i^2}$$

  - 当n个流都获得1个单位吞吐量时，公平指数

$$f(x_1,x_2,\ldots,x_n)=\frac{n^2}{n\times n}=1$$

  - 当某1个流获得1＋△的吞吐量时，公平指数是

$$f(x_1,x_2,\ldots,x_n)=\frac{((n-1)+1+\Delta)^2}{n(n-1+(1+\Delta)^2)}=\frac{n^2+2n\Delta+\Delta^2}{n^2+2n\Delta+n\Delta^2}<1$$

  - N个流中只有k个接收相等等吞吐量，其余为0，公平指数

$$f(x_1,x_2,\ldots,x_n)=\frac{k}{n}$$

3.2 排队规则与流量整形

- 排队规则
  - 不管其它资源分配机制是多么简单或复杂
  - 每个路由器都必须实现某些排队或调度规则，以便管理等待发送包的缓冲区（分配缓冲区资源）
- 排队算法
  - 是带宽（包得到发送）和缓冲区的分配方法（包被丢弃），它还直接影响包经历的延迟
- 3个排队算法：FIFO、公平排队、加权公平排队

3.2.1 FIFO

- First In First Out:
  - 首先到达的包首先发送
  - 当R的缓冲空间满时，尾部的包就丢弃（tail drop）,不考虑包是否重要
  - FIFO和tail drop是不同的概念，前者是发送调度策略，后者是丢弃策略，但二者常捆绑在一起叫做FIFO

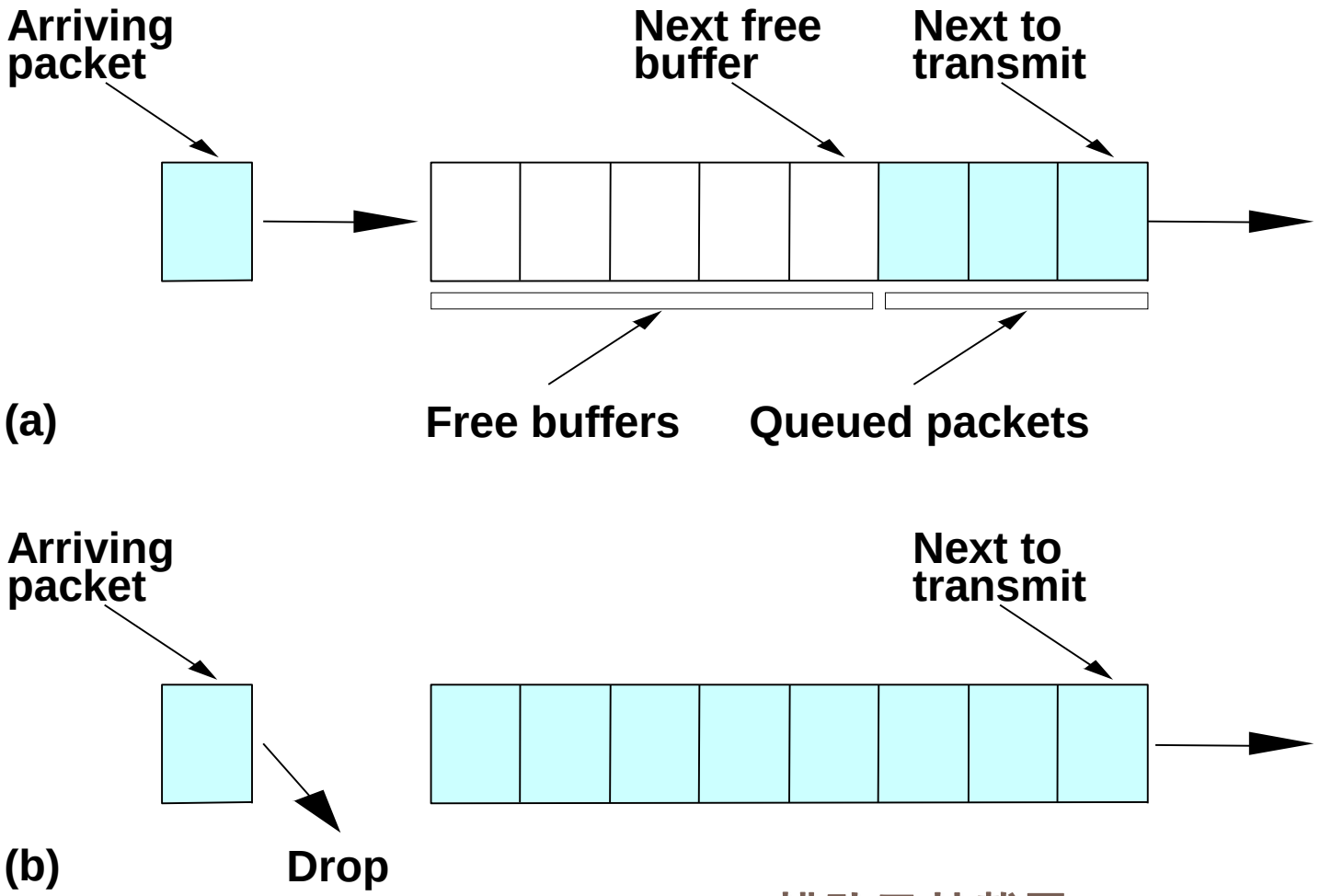


Fig. 2.4 FIFO排队及其截尾

3.2.2 优先排队

- 对基本FIFO的简单变化是优先排队策略：
  - 给每个包打上可携带的优先权标记，如IP服务类型TOS（DiffServ）。
  - R先发送其优先权高的包
  - 这一方式离Best-effort delivery并不远

- TOS（DiffServ）
  - 在IPv4报文头部有个TOS域（8bit），IPv6有个Class域（8bit）
  - RFC791，RFC1122，RFC1349都定义过每个bit的含义； RFC1349废除了之前两个RFC的定义
  - RFC 2474又重新定义了TOS的前6个bit，称为DiffServ Code Point (DSCP)
  - 后两个bit留给Explicit Congestion Notification (ECN)

优先排队策略的应用

RFC 2474中的DSCP，目前在IPv4和IPv6中使用，Class Selector和Drop Precedence组合使用：

低到高

DSCP	Binary	Hex	Decimal	Typical application	Examples
CS0 (Default)	000	0x00	0		
CS1	001	0x08	8	Scavenger	YouTube, Gaming, P2P
CS2	010	0x10	16	OAM	SNMP, SSH, Syslog
CS3	011	0x18	24	Signaling	SCCP, SIP, H.323
CS4	100	0x20	32	Realtime	TelePresence
CS5	101	0x28	40	Broadcast video	Cisco IPVS
CS6	110	0x30	48	Network control	EIGRP, OSPF, HSRP, IKE
CS7	111	0x38	56		

Drop Precedence	Binary	Hex	Decimal
Undefined	000	0x00	0
Low	010	0x02	2
Medium	100	0x04	4
Heigh	110	0x06	6

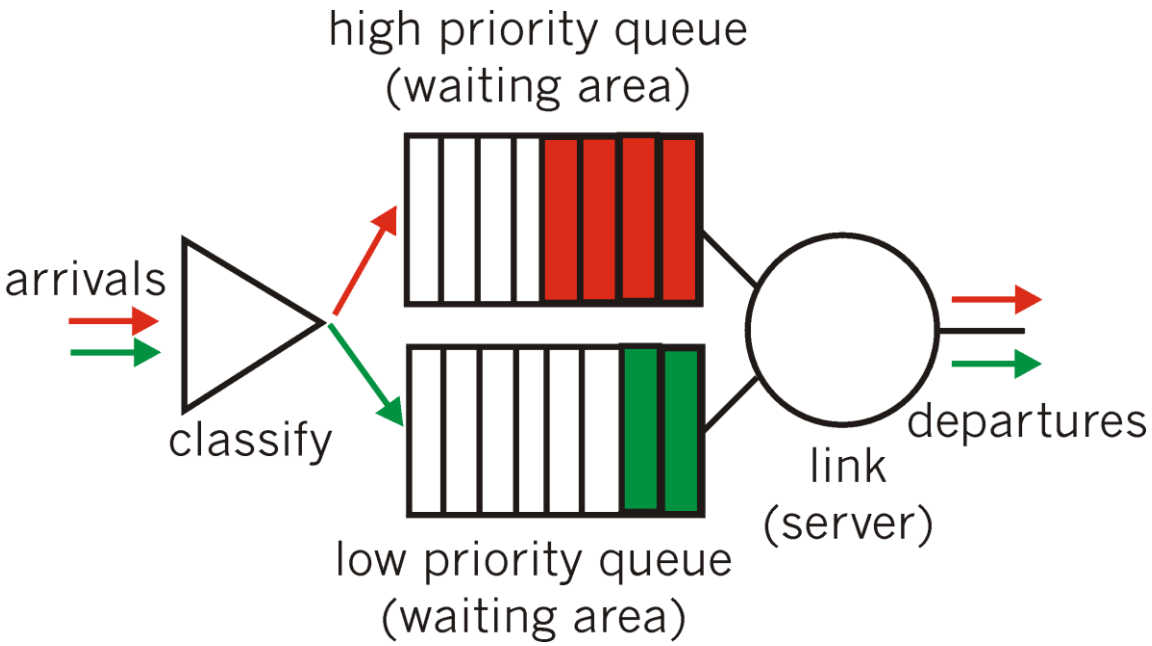


Fig 2.5 优先级排队模型

优先排队策略的应用

- 一个显然的方式是运用经济杠杆，包的有限权越高，则付费越高。
- 然而，富有挑战的是，在因特网这样的分布式环境下如何实现？
- 优先权排队常用在因特网中以保护重要包的传递：
  - 如RIP/OSPF/BGP等路由更新包，以保障在网络拓扑改变后路由表的稳定。
  - 根据TOS或者DiffServ使用特别队列

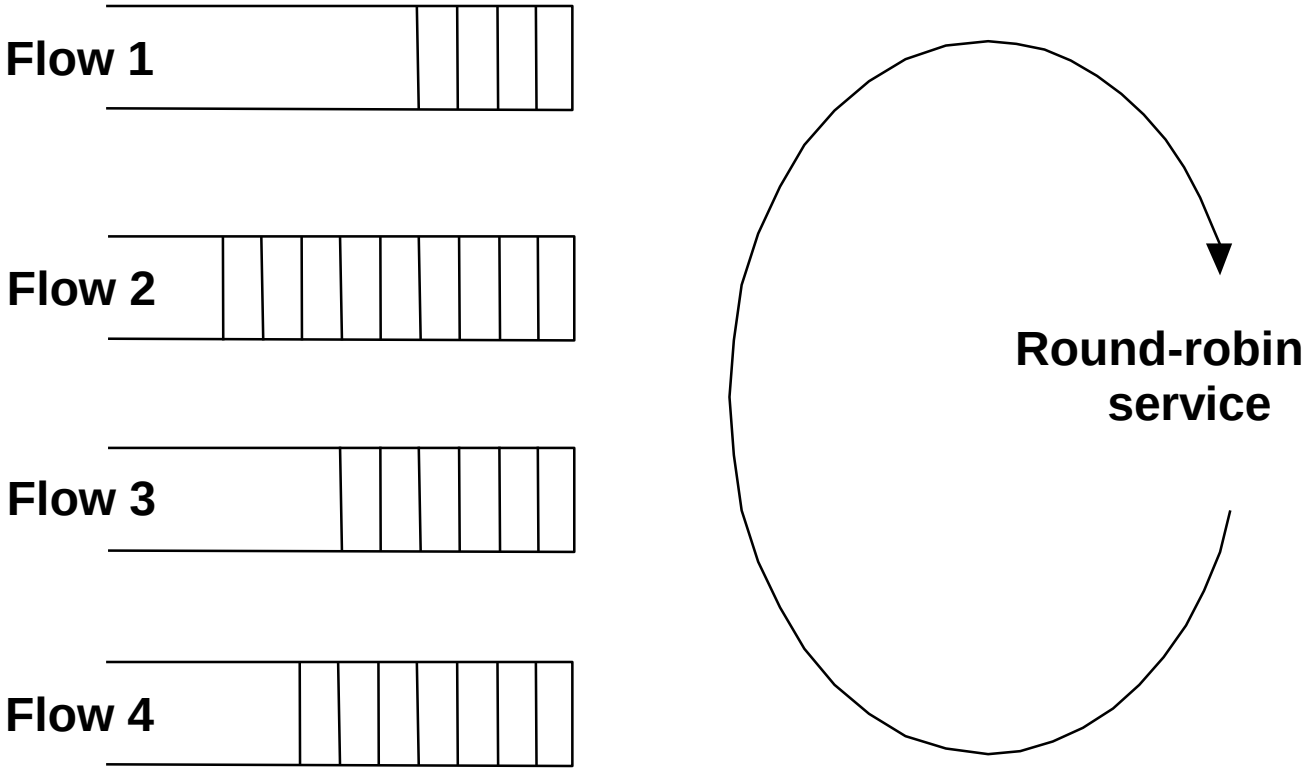
3.2.3 公平排队（6.2.2）

- FIFO主要问题
  - 不能区别不同的流，即不能按报文所属的流来区分。问题：
    - 对在源端实现的拥塞控制算法没有帮助
  - 由于存在非TCP应用，FIFO可能把端-端拥塞机制旁路掉
    - 例如IP电话，这类应用能把自己的包洪泛到网上路由器，引起丢弃其他应用的包
    - 大多数基于UDP对多媒体应用也会产生同样问题

公平排队算法FQ

- FQ主要思想：
  - 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。
  - 当一个流发送包太快，则其队列填满；当一个队列到达一个特定的长度，属于该队列的后继包就被丢掉，这样，任何源就不可能多占用其他流的网络容量
  - FQ并不告诉源端任何有关路由器的状态，或并不限制源端发送怎样快

路由器中的公平排队



FQ机制（6.2.2）

- 主要的复杂性是路由器中处理的包长度不相同。公平须考虑包的长度
  - 如路由器管理2个流，一个是1000Byte/包，另一个是500Byte/包。
  - 对每个流队列的简单轮循服务将给第一个流以2/3链路带宽，第二个流仅给1/3的带宽

- 显然R从不同的包来交叉比特Bit-by-bit轮循是不可行的。
  - 模拟FQ比特轮循机制
  - 先判断如果该包按位轮流方式发送,何时可发完;然后根据完成时间对要发送的包排序

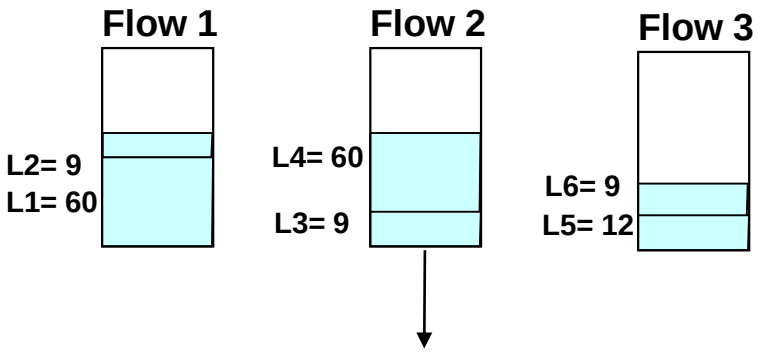
FQ发送算法：

- 先考虑单个流行为,并假设有个Wall Clock,每当传了1位时,时钟滴答一次
  - P_i是包i的长度， S_i是R开始发送包i的时间，F_i是R传输包i的结束时间
  - 故有$F_i = S_i + P_i$
  - 何时开始传输第i个包？
    - 取决于R何时完成第i-1包的最后1比特
    - 或第i-1包的最后1比特传输完成后很长时间，第i个包才到达。令A_i 表示分组i到达路由器的时间。则$S_i = \max(F_{i-1}, A_i)$
    - 故 $F_i = \max(F_{i-1}, A_i) + P_i$

FQ发送算法：

- 扩展到n个active flow，这个时候计算Ai应该使用FQ-clock，FQ-clock当n个比特传输后时钟就记录1次所有到达报文A_i，所以FQ-clock比Wall clock慢n倍
- 而Fi是Fi= max(Fi-1 ,Ai)+ Pi，实际是对每个flow是local的，主要通过Ai实现轮询每个flow
- 对n个active flow，得到每个包的Fi，然后选择最小Fi的包发送

- 6个报文（L1-L6）几乎同时到达，分为三个流，那么计算每个flow中报文发送的时间：F1=60，F2=69，F3=9，F5=12，F4=69，F6=21。注意这个时候Ai是FQ-clock，1个tick=传输3个bit时间，但是从0开始没有太大意义了
- 发送顺序：L3，L5，L6，L1，L2，L4 ，基本上比较公平

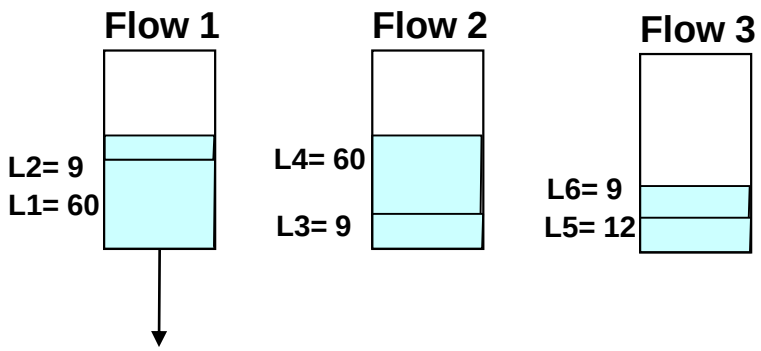


- 教科书第6章第11题（答案在教科书上）

Packet	Size	Flow
1	200	1
2	200	1
3	160	2
4	120	2
5	160	2
6	210	3
7	150	3
8	90	3

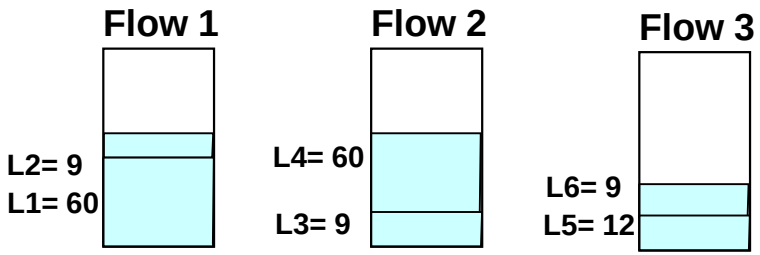
Table 6.2 Packets for Exercise 11.

- 如果三个流的报文不是同时到达，那么active flow的数量会发生变化，那么FQ-clock每个tick时间会变化
- 假设L1到L6报文从Wall Clock为1开始按照顺序到达，然后Wall Clock每tick一下就到达一个报文。那么肯定是先传输L1。



Wall Clock	Ai	arrivals	Fi	Active flows
1	1.0	L1	F1=61	1
2	2.0	L2	F2=70	1
3	3.0	L3	F3=12	1，2
4	3.5	L4	F4=72	1，2
5	4	L5	F5=16	1，2，3
6	4.333	L6	F6=25	1，2，3

- 发送顺序：L1，L3，L5，L6，L2，L4



- 教科书第6章第15题（答案在教科书上）

15 Consider a router that is managing three flows, on which packets of constant size arrive at the following wall clock times:

Flow A: 1, 3, 5, 6, 8, 9, 11

Flow B: 1, 4, 7, 8, 9, 13, 15

Flow C: 1, 2, 4, 6, 7, 12

All three flows share the same outbound link, on which the router can transmit one packet per time unit. Assume that there is an infinite amount of buffer space.

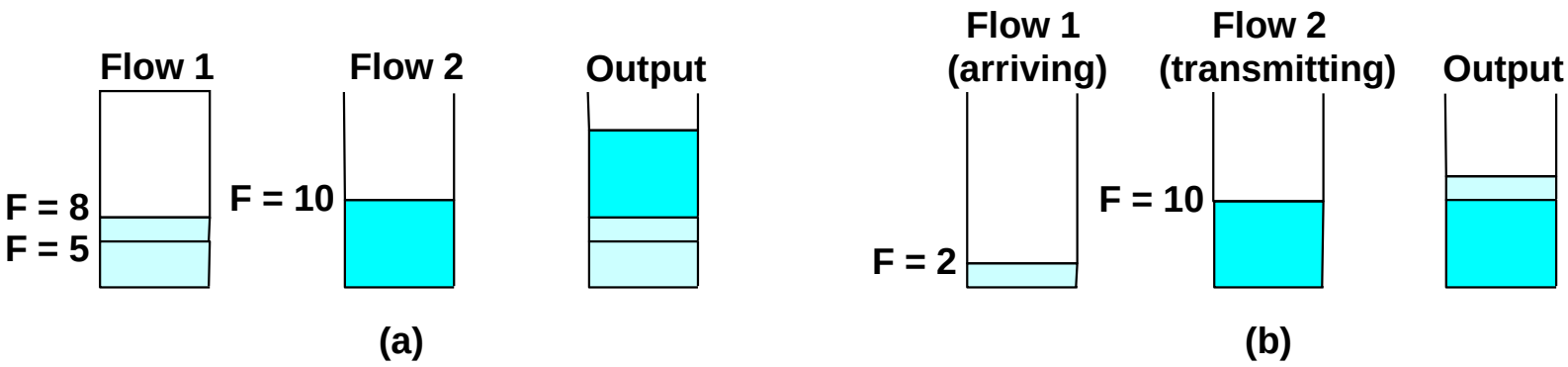
(a) Suppose the router implements fair queuing. For each packet, give the wall clock time when it is transmitted by the router. Arrival time ties are to be resolved in order A, B, C. Note that wall clock time $T = 2$ is FQ-clock time $A_i = 1.333$.

- 教科书第6章第15题（答案在教科书上）

Wallclock	A_i	Arrivals	F_i	Sent	A's Queue	B's Queue	C's Queue
1	1.0	A1,B1,C1	2.0	A1	A1	B1	C1
2	1.333	C2	*3.0*	B1		B1	C1,C2
3	1.833	A3	*3.0*	C1	A3		C1,C2
4	2.333	B4	3.333	A3	A3	B4	C2,C4
		C4	*4.0*				
5	2.666	A5	*4.0*	C2	A5	B4	C2,C4
6	3.0	A6	*5.0*	B4	A5,A6	B4	C4,C6
		C6	*5.0*				
7	3.333	B7	4.333	A5	A5,A6	B7	C4,C6,C7
		C7	*6.0*				
8	3.666	A8	*6.0*	C4	A6,A8	B7,B8	C4,C6,C7
		B8	*5.333*				
9	4	A9	*7.0*	B7	A6,A8,A9	B7,B8,B9	C6,C7
		B9	*6.333*				
10	4.333			A6	A6,A8,A9	B8,B9	C6,C7

2个流公平排队例子

- 图(a)：流1的两个包比流2先发；
- 图(b)：正在发送流2时，流1到达：包已在发送中，进行中优先。
  - 尽管发送中可能流1到达完毕，也不强占流2包

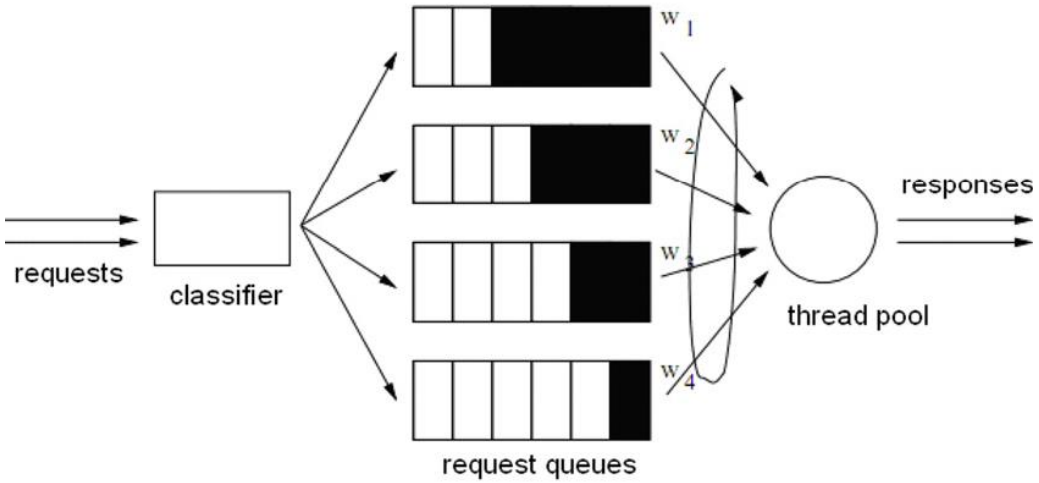


公平排队2点注意

- 只要有包留在队列中，就决不让链路空闲，尽量发送：具备这一特点的排队称为工作守恒work-conserving。其效果是：
  - 如果其他队列都没有报文，唯一的队列可以使用链路的全部容量
  - 一旦其他队列开始发送数据，将分享带宽
- 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的链路带宽。如果报文过多，包的时戳增加，故会在队列中久等，还有可能队列溢出，故FQ算法也应该有丢包的策略

3.2.4 加权公平排队WFQ

- 加权公平队列（Weighted Fair Queueing）
  - 给排队流加权，R服务每个队列时每次发送多少比特，它直接控制每个流将得到多少链路带宽
  - 简单FQ给每个队列权重是1，即每轮每个队列逻辑上仅1比特被传输->1/n
  - WFQ:可让3个队列分别权重2:1:3→则导致带宽分别是2/6:1/6:3/6
  - WFQ中的流在使用中可能是“流量类型”，由TOS或DiffServ.来定义



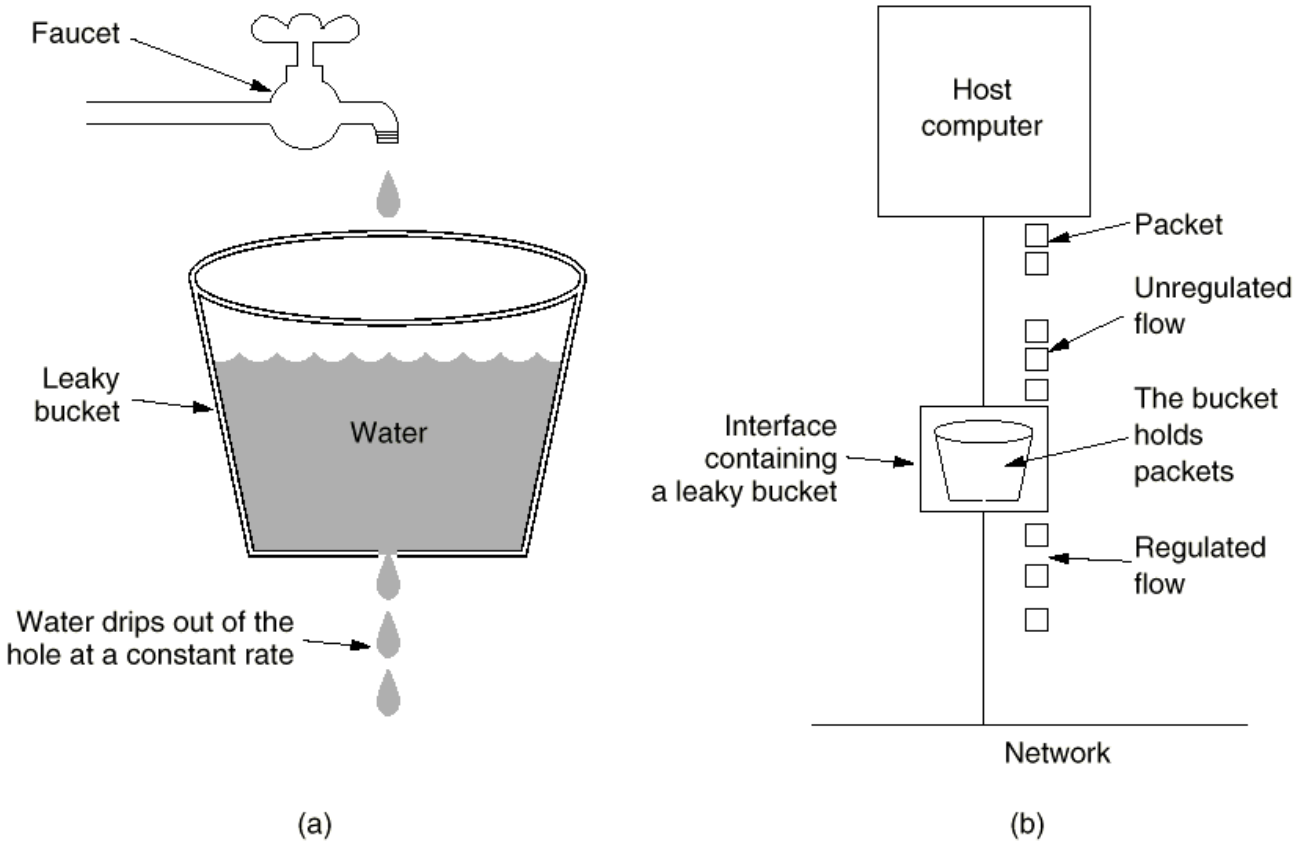
- 第 i 个类的权 = w_i
- 获得部分服务= $\frac{w_i}{\sum w_j}$
- 获得到吞吐量= $R\frac{w_i}{\sum w_j}$

加权公平排队，R表示总吞吐量

3.2.5 流量整形（Traffic Shaping）

- 开环控制
- 基本思想
  - 造成拥塞的主要原因是网络流量通常是突发性的
  - 强迫包以一种可预测的速率发送；
  - 在ATM网中广泛使用。
- 网络流监管的三准则
  - 平均速率:100bps比6000bps的源约束要严格多,希望限制某个流的长期平均速率
  - 峰值速率:网络可允许平均速率6000bps,希望限制其峰值速率<15000bps
  - 突发长度:希望限制极短时间间隔内所能发送到网络的最大包数

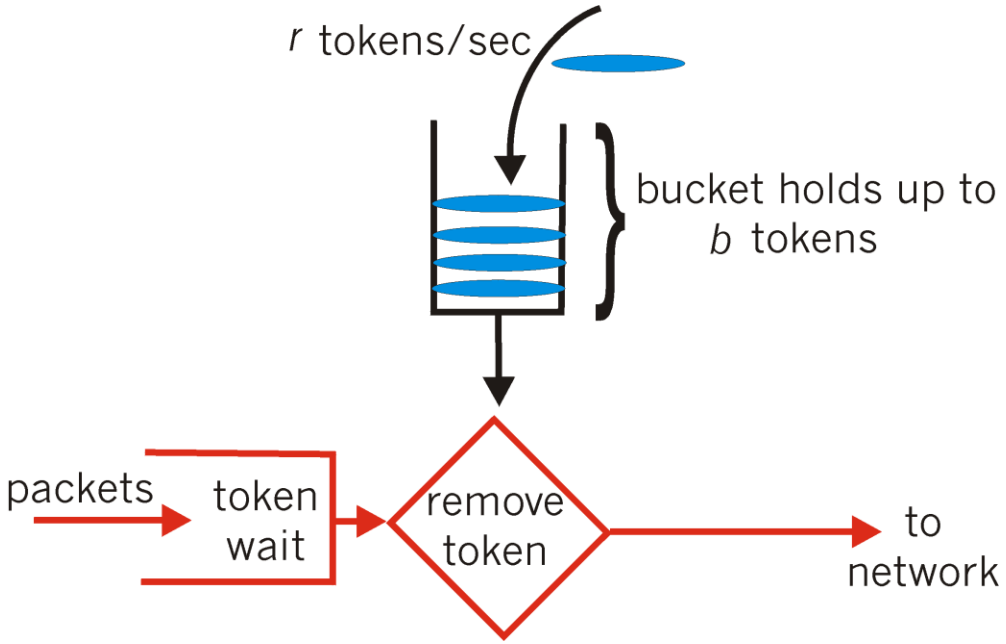
- 漏桶算法（The Leaky Bucket Algorithm）
  - 将用户发出的不平滑的数据包流转变成网络中平滑的数据包流；
  - 可用于固定包长的协议，如ATM；也可用于可变包长的协议，如IP，使用字节计数；
  - 无论负载突发性如何，漏桶算法强迫输出按平均速率进行，不灵活,溢出时要丢包。



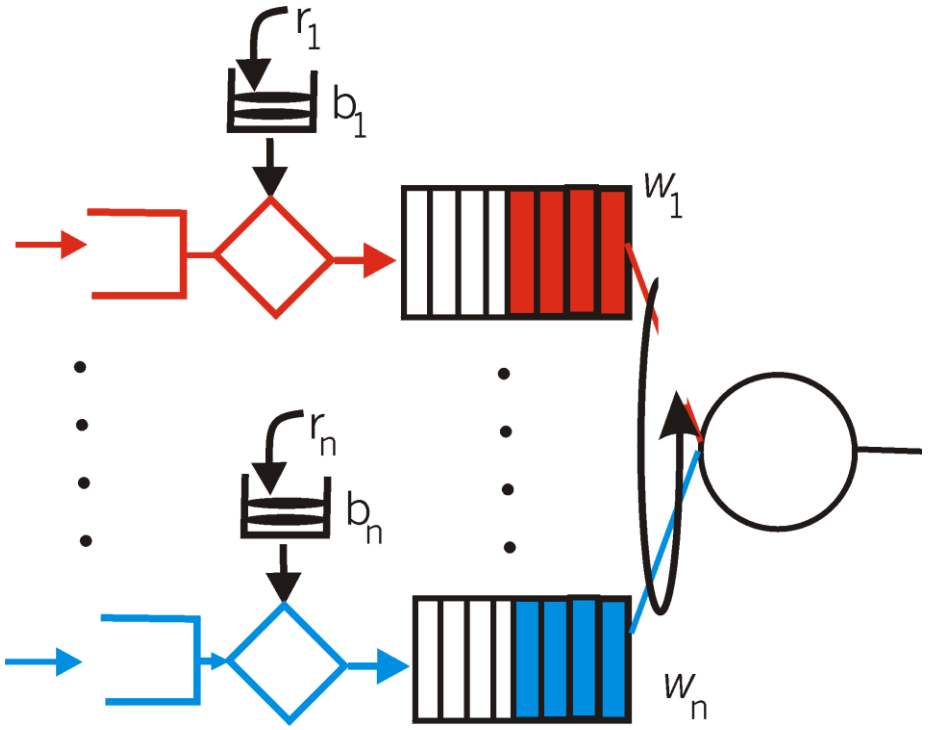
(a) A leaky bucket with water. (b) A leaky bucket with packets.

3.2.6 令牌桶算法 (The Token Bucket Algorithm)

- 基本思想
  - 漏桶算法不够灵活，因此加入令牌机制；
  - 基本思想：漏桶存放令牌，每ΔT秒产生一个令牌，令牌累积到超过漏桶上界时就不再增加。包传输之前必须获得一个令牌，传输之后删除该令牌；
- 漏桶算法与令牌桶算法的区别
  - 流量整形策略不同：漏桶算法不允许空闲主机积累发送权，以便以后发送大的突发数据；令牌桶算法允许，最大为桶的大小。
  - 漏桶中存放的是数据包，桶满了丢弃数据包；令牌桶中存放的是令牌，桶满了丢弃令牌，不丢弃数据包没有令牌才丢弃数据包。

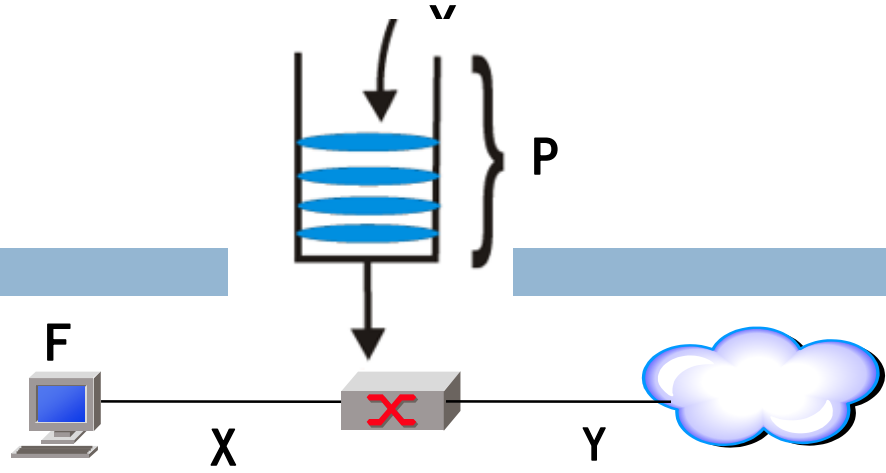


- 令牌漏桶整形算法
- 在时间间隔t内，网络中的最大包数=rt+b，b为桶中最大令牌数
- 故令牌长生率r成为包进入网络的平均速率
- 用2个串连漏桶可以整形网络峰值速率（容纳更多包，b增加了，但是速率不变）
- 并联：速率增加、峰值也增加了



- 采用WFQ的n路复用的漏桶流控
- 可在Diffserv中使用

问题分析与解答



- 问题：
  - 某台计算机有一输出文件长度F个Bits，该计算机的输出链路速率是X bps，但受到其接入网关的令牌桶交通管制；令牌桶始终保持Y bps 的填充速率；令牌桶大小是P Bits。假设发送开始时令牌桶为满，试求？
    a) X大于Y，网关缓冲区足够大情况下，计算机以最大速率发送的最长持续时间t之表达式？
    b) 若X＞Y，F＞P，求计算机把文件全部发送完毕所需要的最短时间T是多少
    c) 若X =6Mbps、Y=1Mbps、F=10.6MBits 、P=8MBits求t= ?、T=?
- 解答
  a) t时间内计算机输出数据 = 网关输入数据， tX=tY+P ， t=P/(X-Y)
  b) T = t+(F-Xt)/Y （即计算机先以X发送t秒，再用Y速率发送T-t秒）
  c) t=8MBits/(6Mbps-1Mbps)=1.6s; T=1.6s+(10.6MBits-1.6s*6Mbps)/1Mbps=2.6s