2. 路由器组成

2.1 路由器结构异况

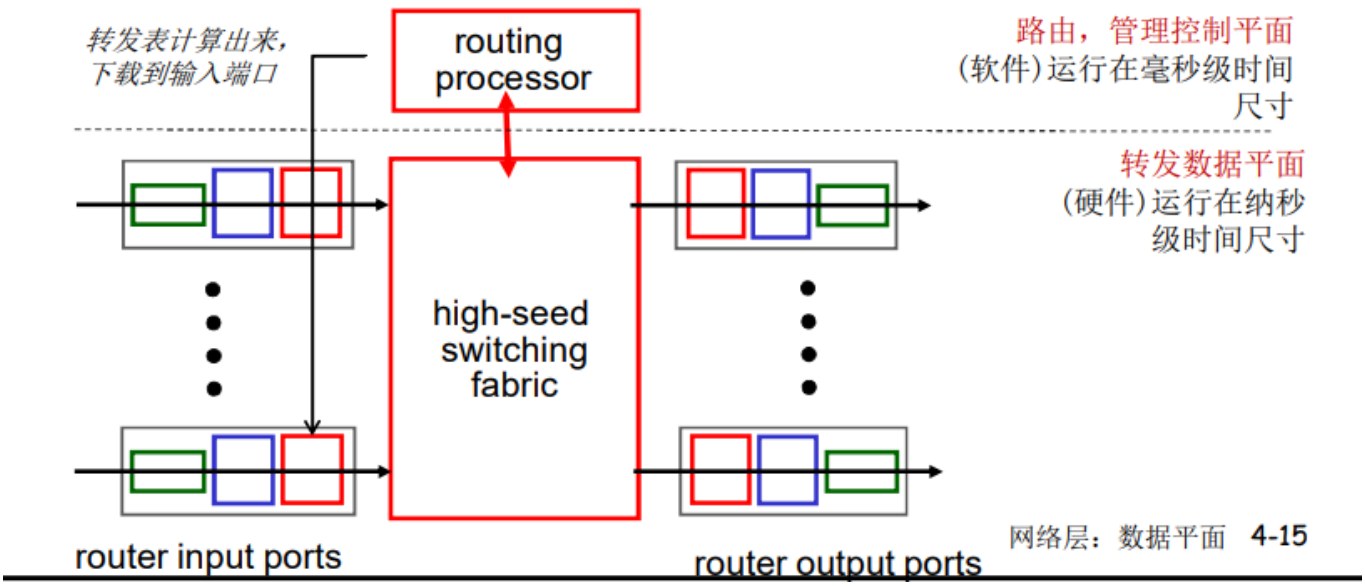
高层面（非常简化的）通用路由器体系架构
    路由：运行路由选择算法/协议（RIP,OSPF,GBP）-生成路由表
    转发：从输入到输出链路交换数据报-根据路由表进行分组的转发

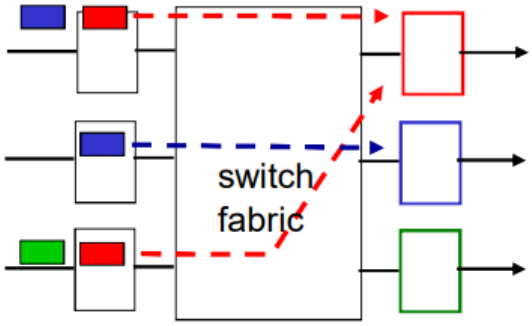
router input ports -->fabric --> router output ports
输入端口　　交换机（局部转发）　　输出端口

2.2 输入端口

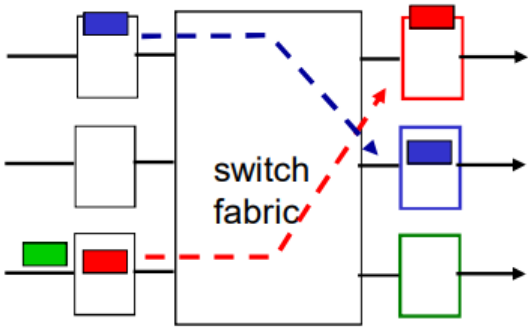
    输入端口功能
        输入输出端口是整合在一起的，没有独立的输入输出端口
        物理层：
            Bit级的接收
            物理信号转换成数字信号
        数据链路层：
            链路层协议动作、解封装　e.g. Ethernet
        分布式交换（网络层）：
            根据数据报头部的信息如：目标地址，在输入端口内存中的路由表中查找合适的输出端口（匹配+行动）
            基于目标的转发:仅仅依赖于IP数据报的目标IP地址(传统方法)
            通用转发：基于头部字段的任意集合进行转发

    输入端口缓存
        为什么fabric与数据链路层之间有queue？
            因为输入端口的汇聚速率与转发的速度不一致，需要queue来匹配瞬间速度的不一致　fabric > 输入端口的汇聚速率
            简化：需要queue来匹配瞬时的输入速率与输出速率的不一致性
        当交换机构的速率小于输入端口的汇聚速率时，在输入端口可能要排队
            排队延迟以及由于输入缓存溢出造成丢失
        Head-of-the-Line（HOL）blocking 头端阻塞：排在队头的数据报阻止了队列中其他数据报的向前移动





输出端口竞争：
只能有一个红色分组被传递，交换到一个输出端口.
下面红色的分组被阻塞



一个分组时间：
绿色分组经历了头端阻塞

2.3 交换结构

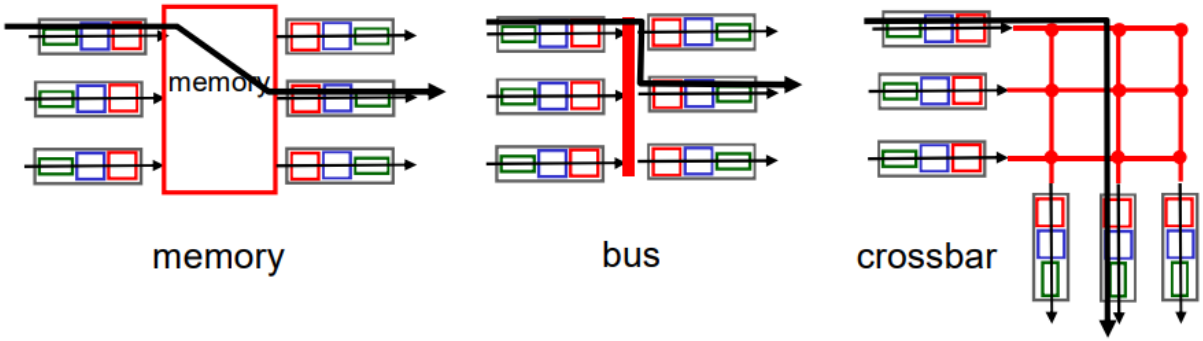
将分组从输入缓冲区传输到合适的输出端口
交换速率：分组可以按照该速率从输入传输到输出
运行速度经常是输入/输出链路速率的若干倍
N个输入端口：交换机构的交换速度是输入线路速度的N倍比较理想，才不会成为瓶颈

3种典型的交换机构：
memory 内存交换
通过2次总线
总线-内存-总线
第一代路由器：
在CPU直接控制下的交换，采用传统的计算机
分组被拷贝到系统内存，CPU从分组的头部提取出目标地址，查找转发表，找到对应的输出端口，拷贝到输出端口

bus 总线交换
通过1次总线
输入-总线-输出
数据报通过共享总线，从输入端口转发到输出端口
总线竞争：交换速度受限于总线带宽
1次处理1个分组
1Gbps bus，Cisco 1900
32Gbps bus，Cisco 5600
对于接入或企业级路由器，速度足够，但是不适合区域或者骨干网络

crossbar 互联网络交换：
同时并发转发多个分组，克服总线带宽限制
Banyan（榕树）网络，crossbar（纵横）和其他的互联网络被开发，将多个处理器连接成多处理器

当分组从端口A到达，转给端口Y；控制器短接相应的两个总线
高级设计：将数据报分片为固定长度的信元，通过交换网络交换
Cisco12000:以60Gbps的交换速率通过互联网络



2.4 输出端口

物理层：
数字信号转换成物理信号
链路层：
链路层协议动作、封装（加入帧头）
网络层：
排队，转发

当数据报从交换机构的 到达速率 比 传输速率 快，就需要输出端口缓存
由于拥塞，缓冲区没有空间，数据报（分组）可能被丢弃
由调度规则选择排队的数据报进行传输
优先权调度-谁会获得最佳性能，网络中立？

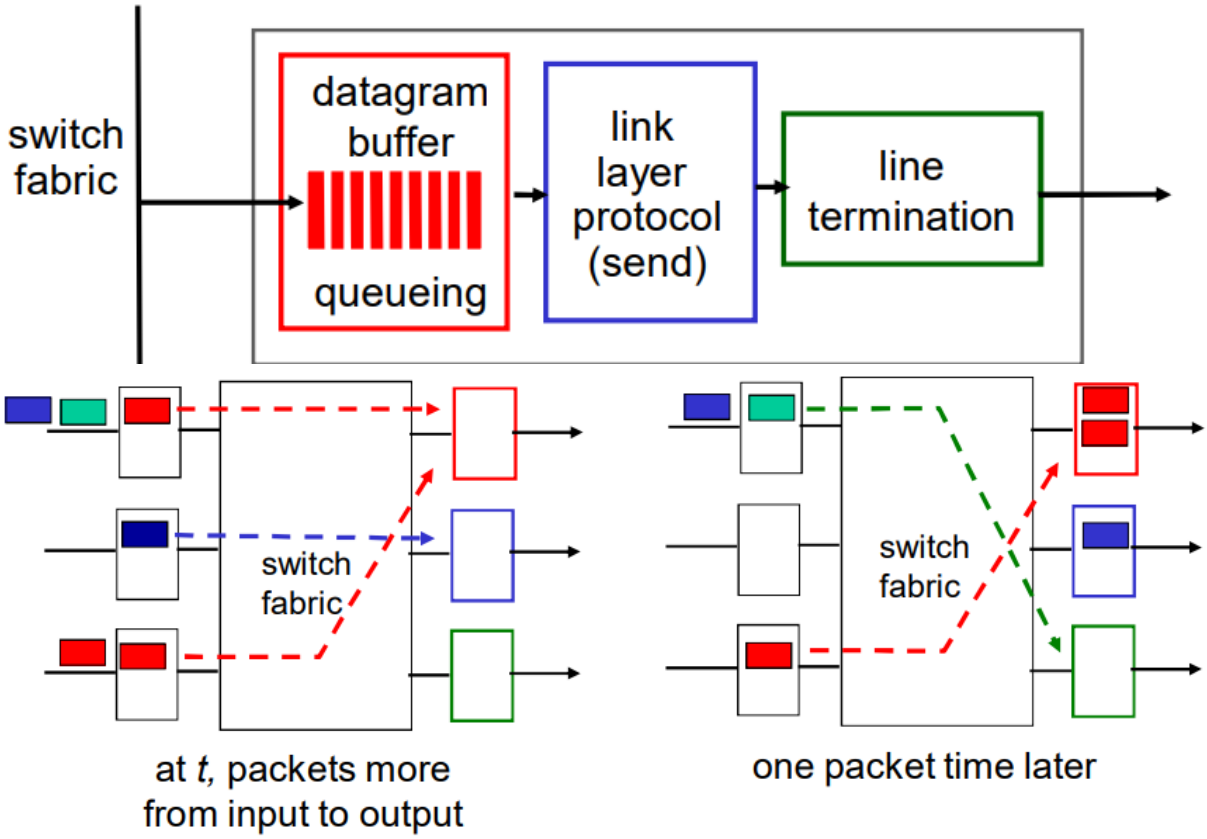
输出端口排队
假设交换速率是 传输速率的N倍
当多个输入端口同时向输出端口发送时，缓冲该分组（当通过交换网络到达的速率超过输出速率则缓存）
排队带来延迟，由于输出端口缓存溢出则丢弃数据报

需要多少缓存：
RFC 3439 拇指规则（经验性规则）：平均缓存大小=典型的RTT（例如250ms）倍于链路容量
e.g., C=10 Gbps Link
250ms*10Gbps=2.5Gbit buffer
最近的一些推荐：
有N（非常大）个流，缓存大小等于=RTT*C/sqrt(N)

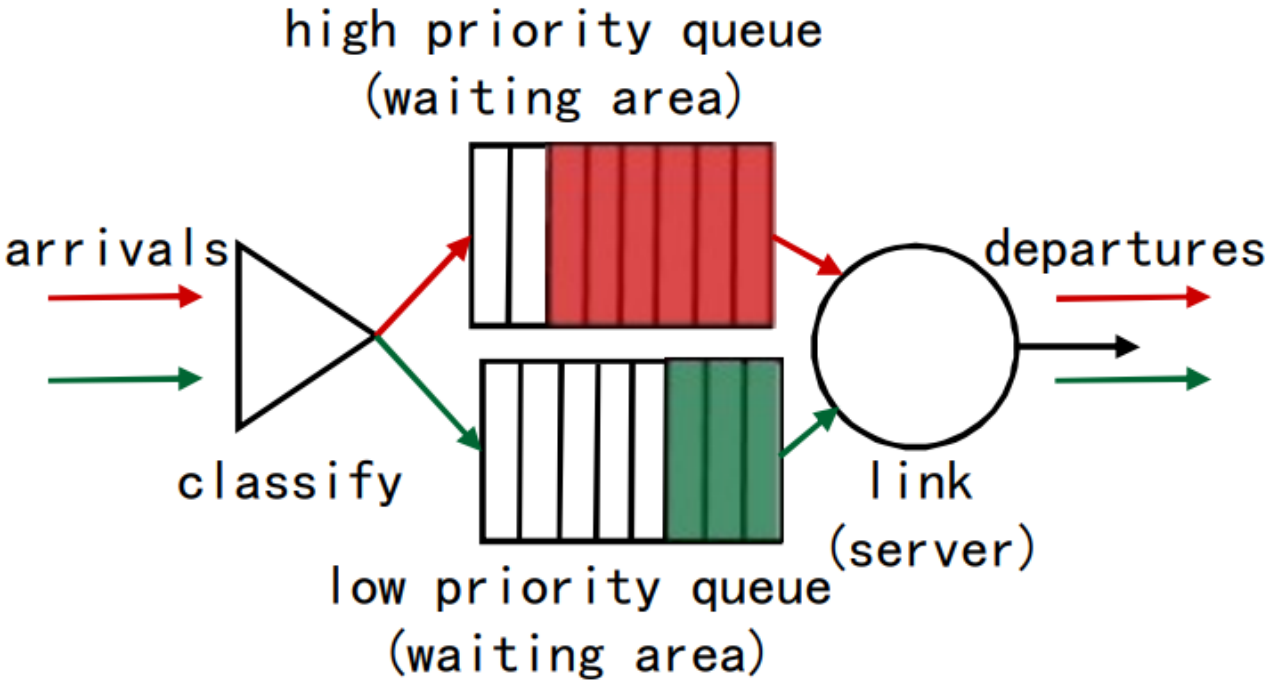
调度机制：
调度：选择下一个要通过链路传输的分组
FIFO scheduling：（first in first out）按照分组的到来的次序发送
丢弃策略：如果分组到达一个满的队列，哪个分组被抛弃？
tail drop：丢弃刚到达的
priority：根据优先权丢弃/移除分组
random：随机地丢弃/移除

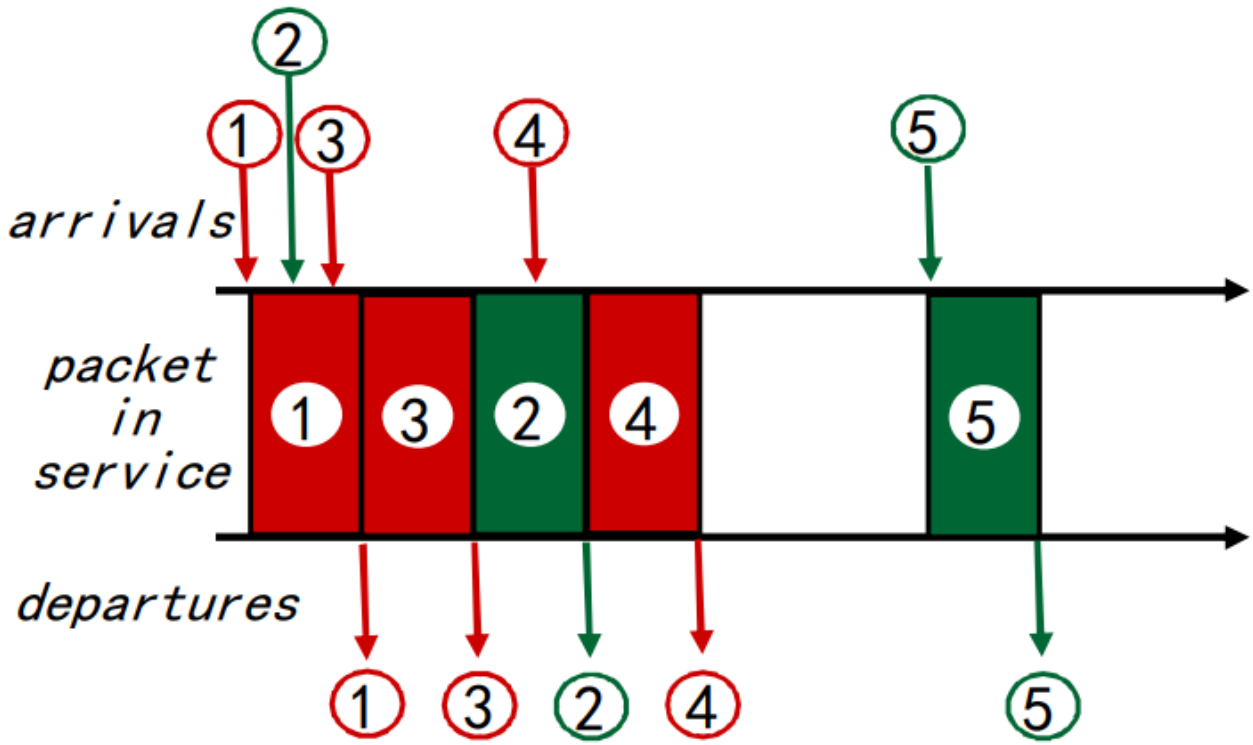
调度策略：
优先权调度：
发送最高优先权的分组
多类，不同类别有不同的优先权
类别可能依赖于标记或者其他头部字段，e.g. IP source/dest, port numer,ds,etc.
先传高优先级的队列种的分组，除非没有
高（低）优先权中的分组传输次序：FIFO
其他：
Round Robin （RR）scheduling：
多类
循环扫描不同类型的队列，发送完一类的一个分组，再发送下一个类的一个分组，循环所有类
类似打印机色带，打完这个色打下个色
Weighted Fair Queuing（WFQ）： 加权公平队列

一般化的Round Robin
在一段时间内，每个队列得到的服务时间是：Wi/（XIGMA(Wi)）*t，和权重成正比
每个类在每一个循环中获得不同权重的服务量

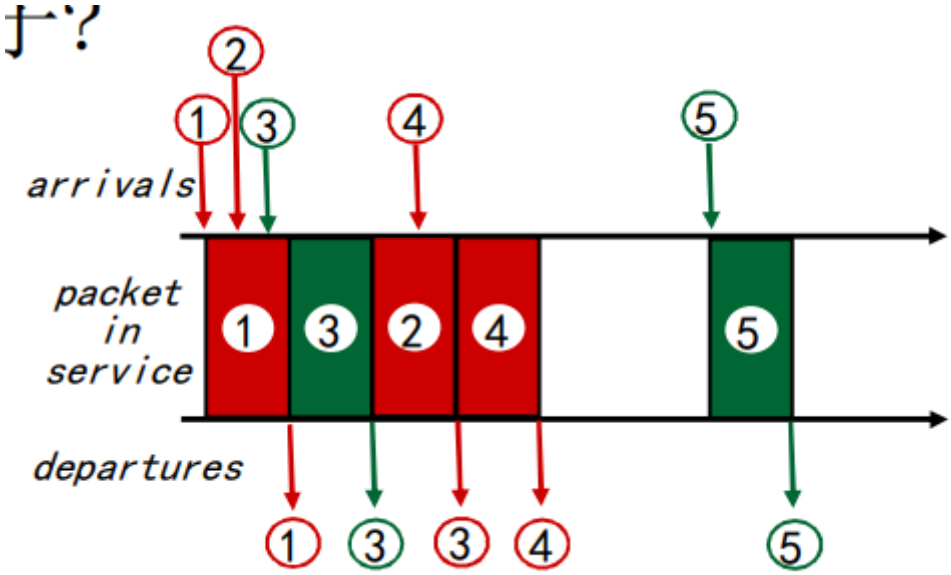


优先权调度

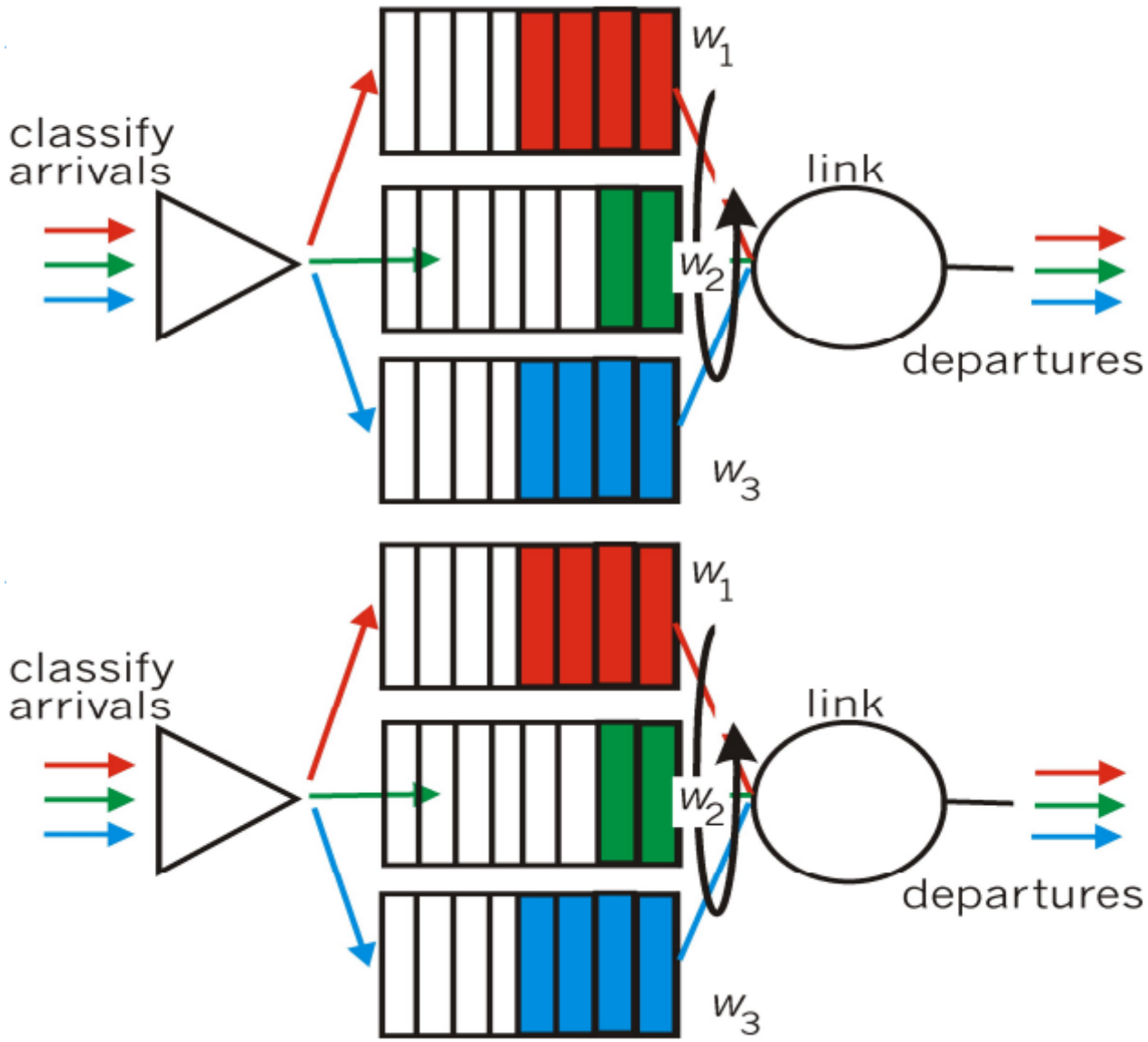

high priority queue
(waiting area)
arrivals
departures
classify
link
(server)
low priority queue
(waiting area)


2
1
3
4
5
arrivals
packet
in
service
1
3
2
4
5
departures
1
3
2
4
5

RR


arrivals
packet in service
departures

WFQ


classify arrivals
link
departures
w_1
w_2
w_3