W7

(2)

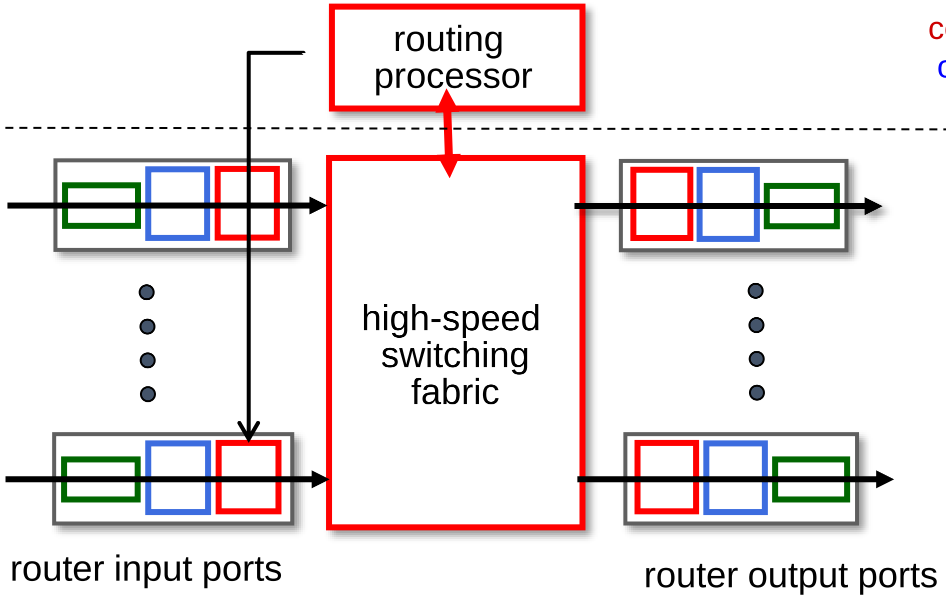
# Computer Networks

## Router Architecture and Scheduling

Amitangshu Pal
Computer Science and Engineering
IIT Kanpur

# Router Architecture

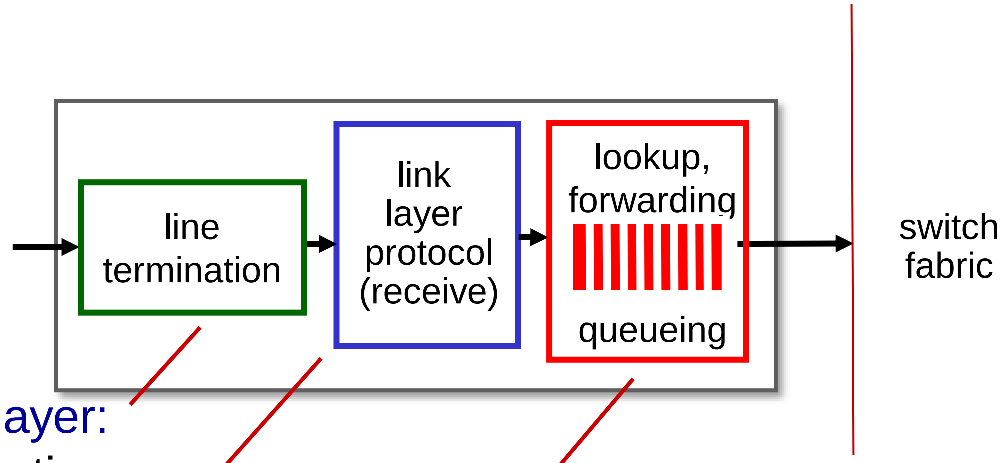
# Router architecture overview

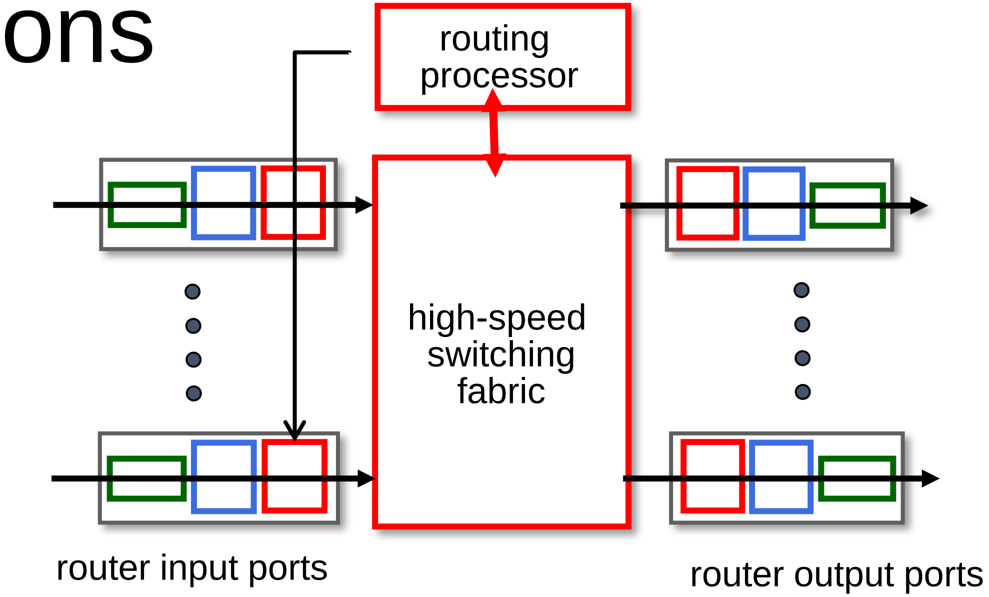


routing, management
control plane (software)
operates in millisecond
time frame

forwarding data
plane (hardware)
operates in
nanosecond
timeframe

# Input port functions





Physical layer:
Bit-level reception

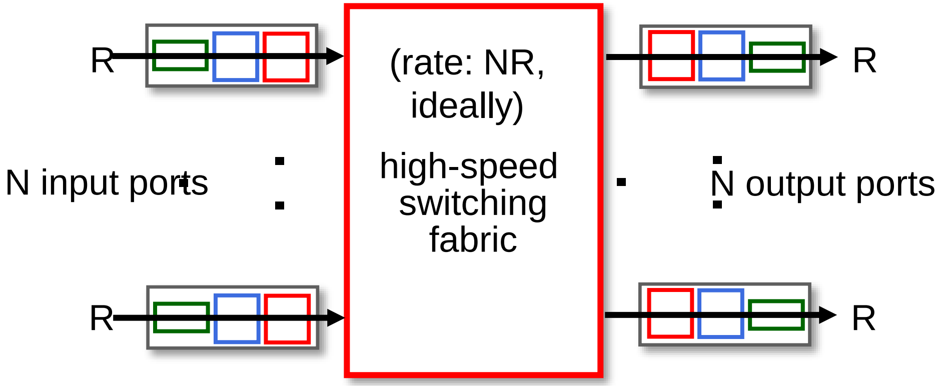
Link layer:
e.g., Ethernet

## Decentralized switching:

- Using header field values, lookup output port using forwarding table in input port memory ("match plus action")
- Goal: complete input port processing at 'line speed'
- Input port queuing: if datagrams arrive faster than forwarding rate into switch fabric

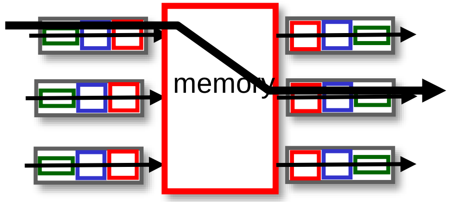
# Switching fabrics

- Transfer packet from input link to appropriate output link
- Switching rate: rate at which packets can be transfer from inputs to outputs
  - Often measured as multiple of input/output line rate
  - N inputs: switching rate N times line rate desirable

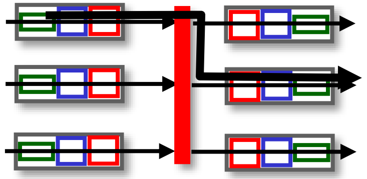


# Switching fabrics

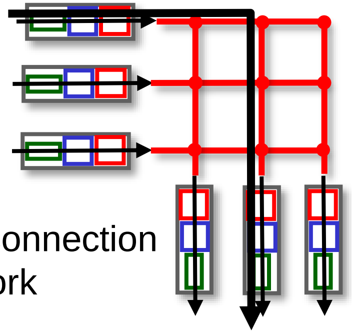
- Transfer packet from input link to appropriate output link
- Switching rate: rate at which packets can be transfer from inputs to outputs
  - Often measured as multiple of input/output line rate
  - N inputs: switching rate N times line rate desirable
- Three major types of switching fabrics:



Memory



Bus

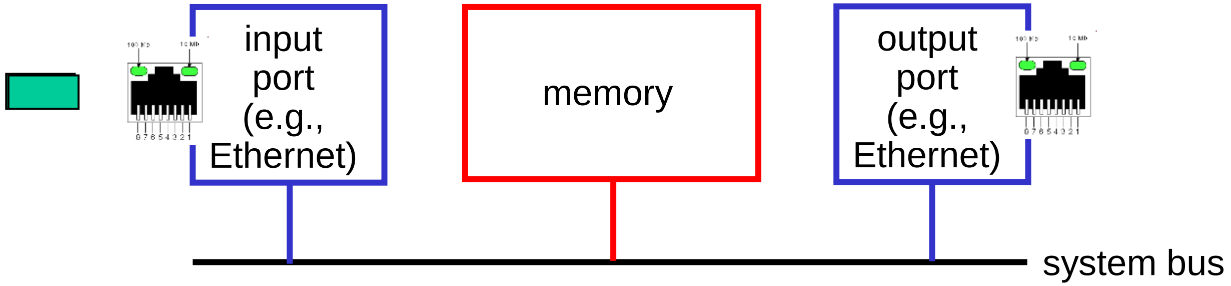


Interconnection network

# Switching via memory

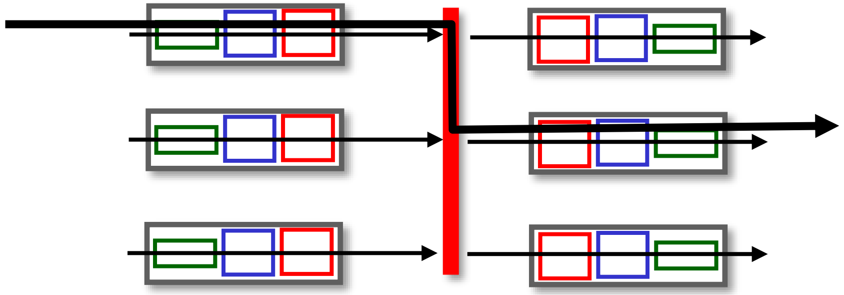
## First generation routers:

- Traditional computers with switching under direct control of CPU
- Packet copied to system's memory
- Speed limited by memory bandwidth (2 bus crossings per datagram)



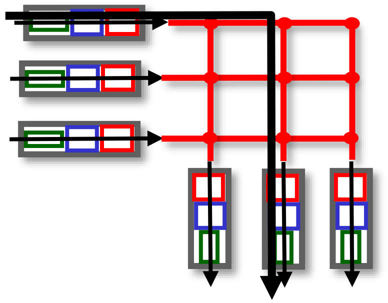
# Switching via a bus

- Datagram from input port memory to output port memory via a shared bus
- Bus contention: switching speed limited by bus bandwidth
- 32 Gbps bus, Cisco 5600: sufficient speed for access routers

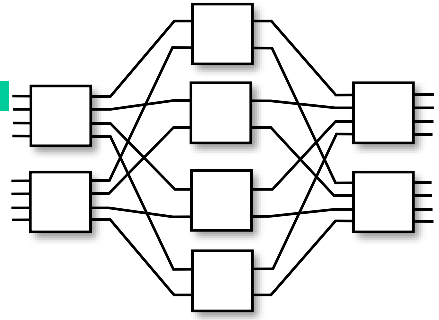


# Switching via interconnection network

- Crossbar, Clos networks, other interconnection nets initially developed to connect processors in multiprocessor computer architecture
- Multistage switch: nxn switch from multiple stages of smaller switches
- Exploiting parallelism:
  - Fragment datagram into fixed length cells on entry
  - Switch cells through the fabric, reassemble datagram at exit



3x3 crossbar

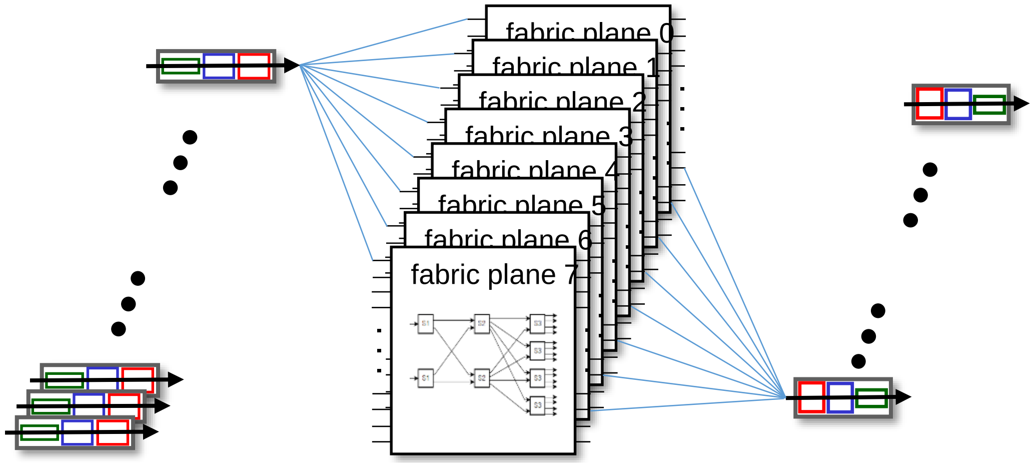


8x8 multistage switch
built from smaller-sized switches

# Switching via interconnection network

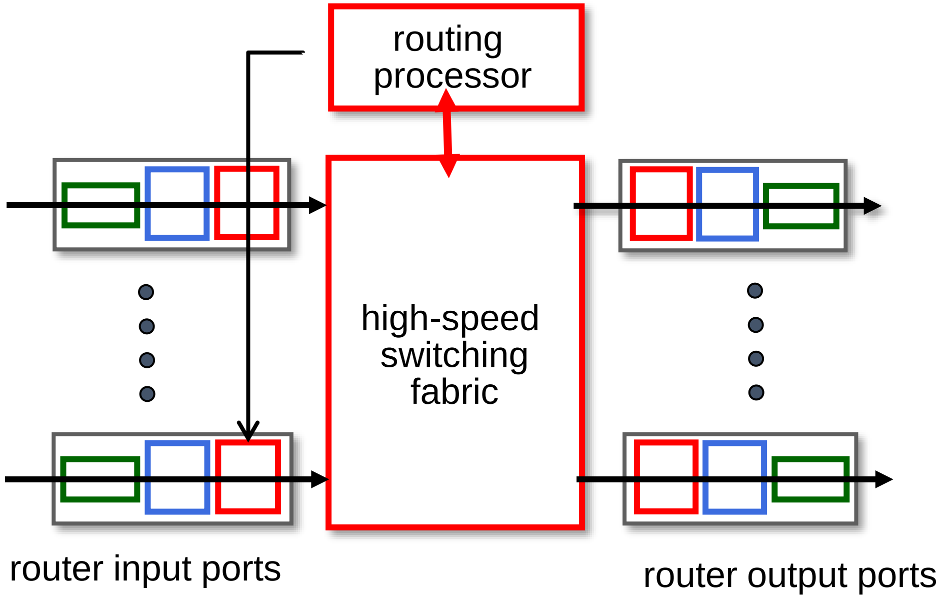
- Scaling, using multiple switching "planes" in parallel:
  - Speedup, scaleup via parallelism

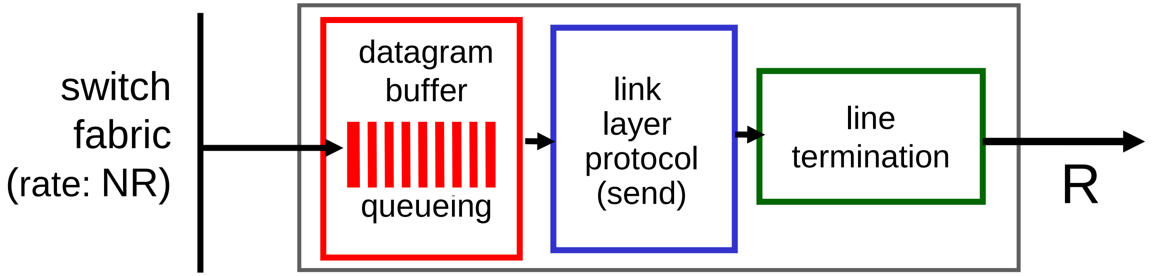
- Cisco CRS router:
  - Basic unit: 8 switching planes
  - Each plane: 3-stage interconnection network
  - up to 100's Tbps switching capacity



- Cisco CRS router: https://nexstor.com/wp-content/uploads/2018/05/cisco-crs-1-multishelf-system-datasheet.pdf

# Output port functions



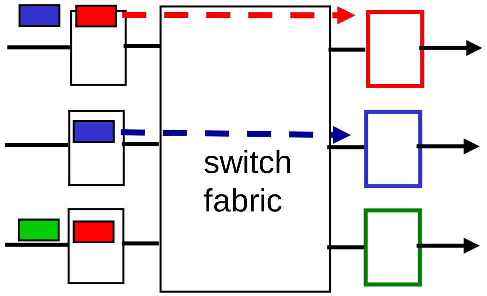


# Queuing, Buffer management and Scheduling

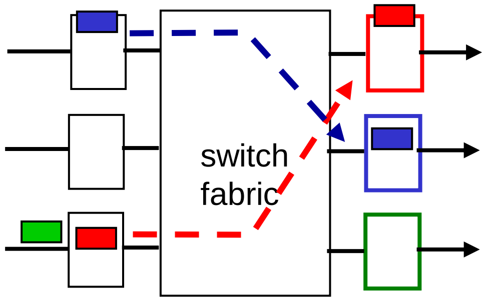
# Input port queuing

- If switch fabric slower than input ports combined → queueing may occur at input queues
  - Queueing delay and loss due to input buffer overflow!

- Head-of-the-Line (HOL) blocking: Queued datagram at front of queue prevents others in queue from moving forward

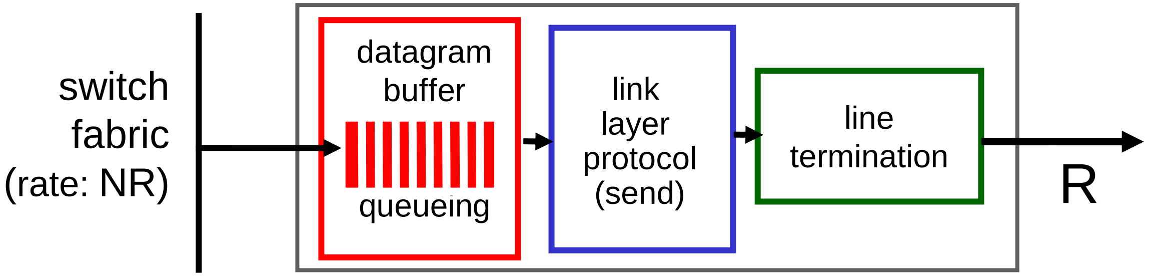


Output port contention: only one red datagram can be transferred. lower red packet is blocked



One packet time later: green packet experiences HOL blocking

# Output port queuing



- Buffering required when datagrams arrive from fabric faster than link transmission rate
- Drop policy: which datagrams to drop if no free buffers?

Datagrams can be lost due to congestion, lack of buffers

- Scheduling discipline chooses among queued datagrams for transmission

Priority scheduling – who gets best performance

# How much buffering?

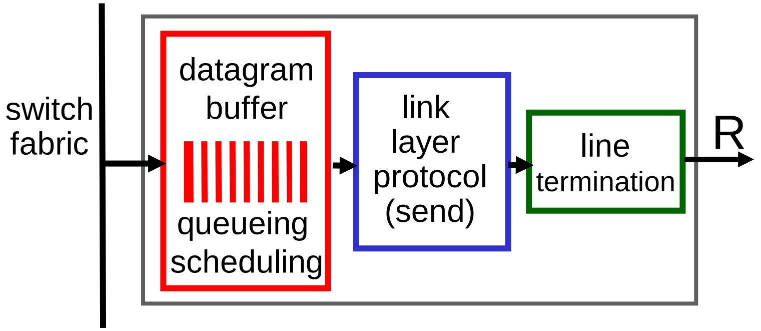
- Too much buffering will reduce packet loss, but can increase delays
  - Long RTTs: poor performance for real-time apps, sluggish TCP response

- RFC 3439 rule of thumb: average buffering equal to “typical” RTT times link capacity C
  - e.g., RTT = 250 msec, C = 10 Gbps link → 2.5 Gbit buffer
  - Delay-bandwidth product

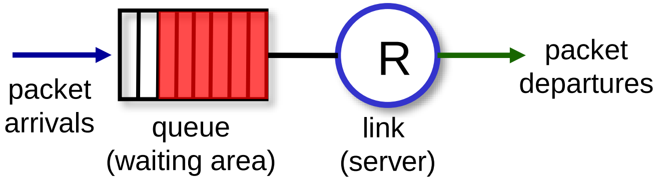
- More recent recommendation: with N flows, buffering equal to

$$\frac{\mathrm{RTT.\,C}}{\sqrt{\mathrm{N}}}$$

# Buffer Management



Abstraction: queue



## Buffer management:

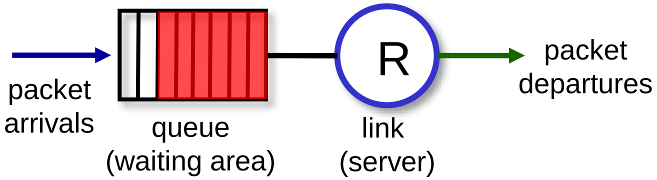
- Drop: which packet to add, drop when buffers are full
  - Tail drop: drop arriving packet
  - Priority: drop/remove on priority basis
- Marking: which packets to mark to signal congestion (i.e. ECN)

# Packet Scheduling: FCFS

Packet scheduling: deciding which packet to send next on link

- First come, first served
- Priority based
- Round robin
- Weighted fair queueing

Abstraction: queue



FCFS: packets transmitted in order of arrival to output port

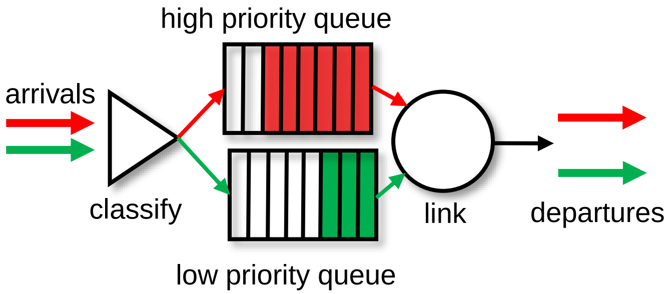
- also known as: First-in-first-out (FIFO)

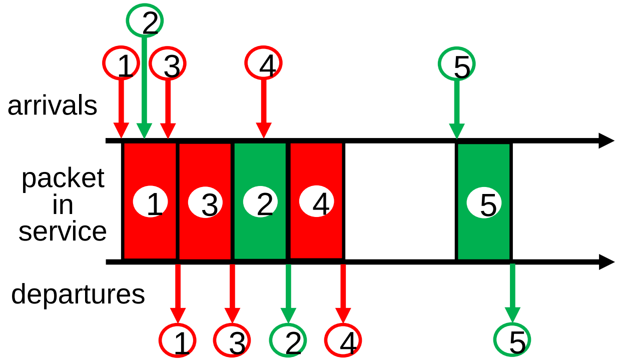
# Scheduling policies: Priority Based

Priority based scheduling:

- Arriving traffic classified, queued by class
  - Any header fields can be used for classification

- Send packet from highest priority queue that has buffered packets
  - FCFS within priority class

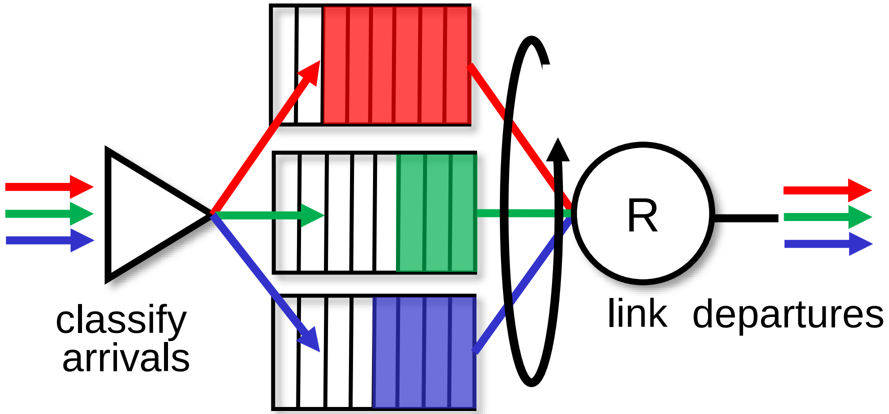




# Scheduling policies: round robin

Round Robin (RR) scheduling:

- Arriving traffic classified, queued by class
  - Any header fields can be used for classification
- Server cyclically, repeatedly scans class queues, sending one complete packet from each class (if available) in turn



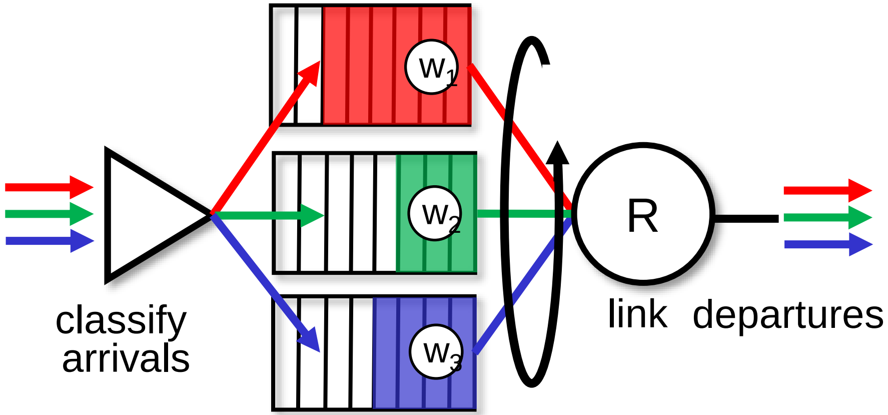
# Scheduling policies: Weighted Fair Queueing

Weighted Fair Queuing (WFQ):

- Generalized Round Robin
- Each class $i$, has weight, $w_i$ and gets weighted amount of service in each cycle:

$$\frac{w_i}{\sum_j w_j}$$

- Minimum bandwidth guarantee (per-traffic-class)



# Summary

❑Router architecture, queuing and packet scheduling:

- Router architecture
  - Input ports
  - High speed fabric
  - Out ports
- Packet scheduling
  - FCFS
  - Priority based
  - Round robin
  - Weighted fair queuing