Chapter 4: outline

4.1 Overview of Network layer
- data plane
- control plane

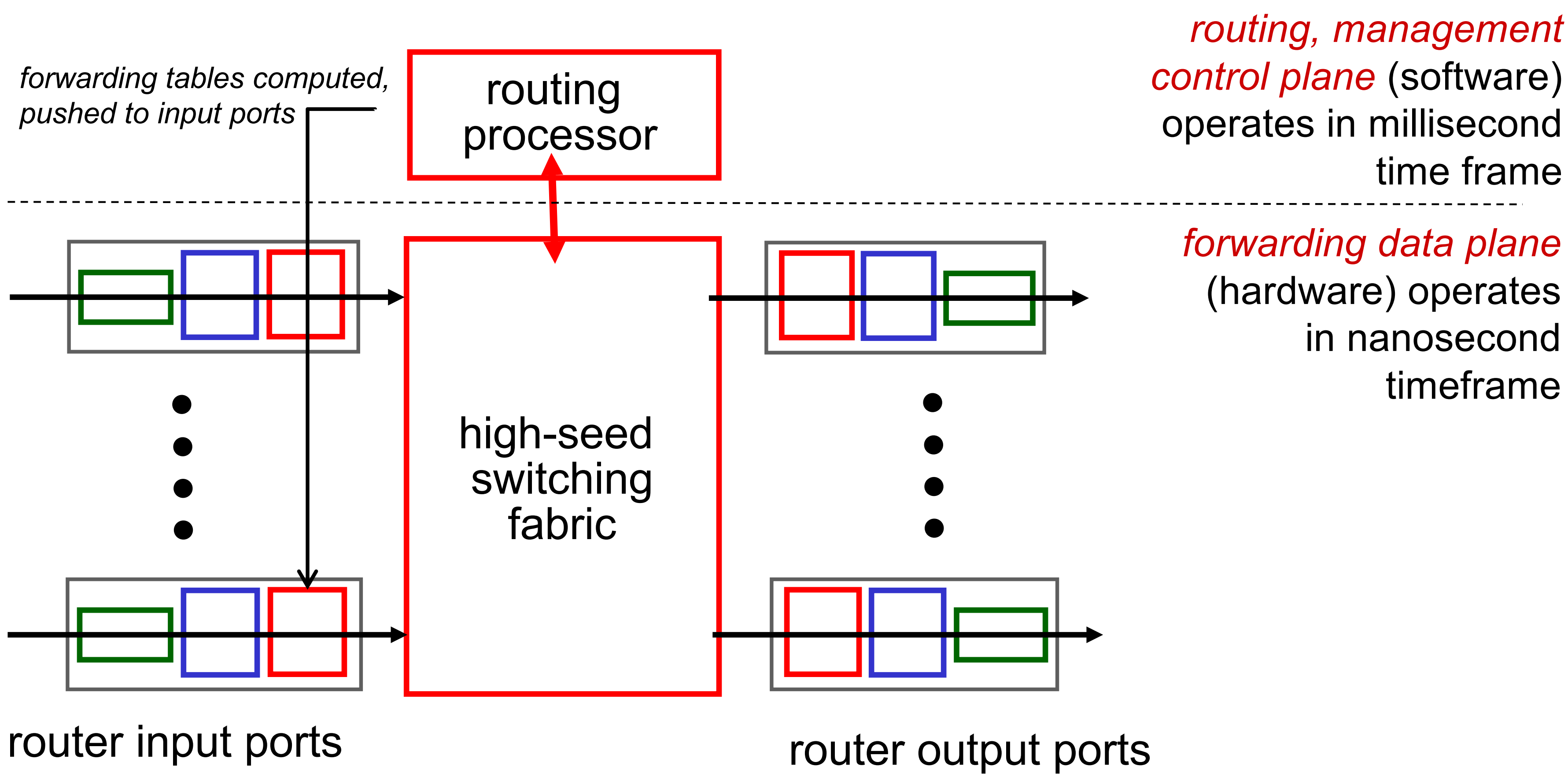
4.2 What's inside a router

4.3 IP: Internet Protocol
- datagram format
- fragmentation
- IPv4 addressing
- network address translation
- IPv6

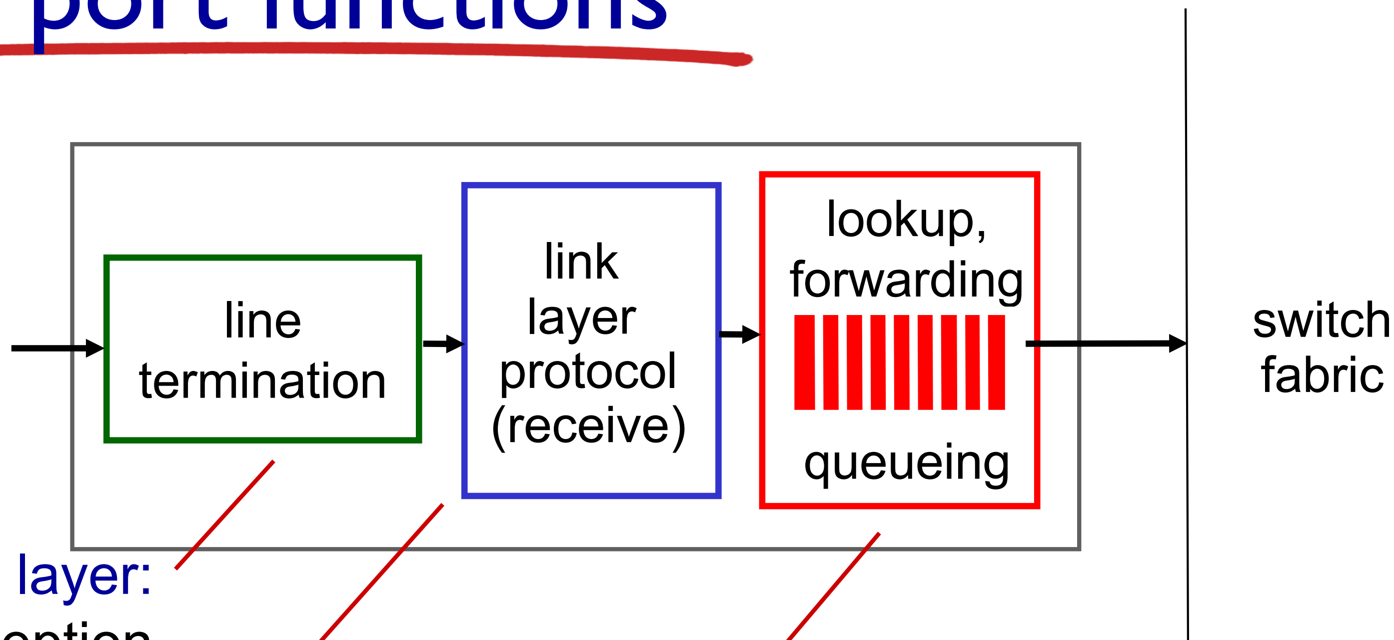
4.4 Generalized Forward and SDN
- match
- action
- OpenFlow examples of match-plus-action in action

Router architecture overview

- high-level view of generic router architecture:



Input port functions



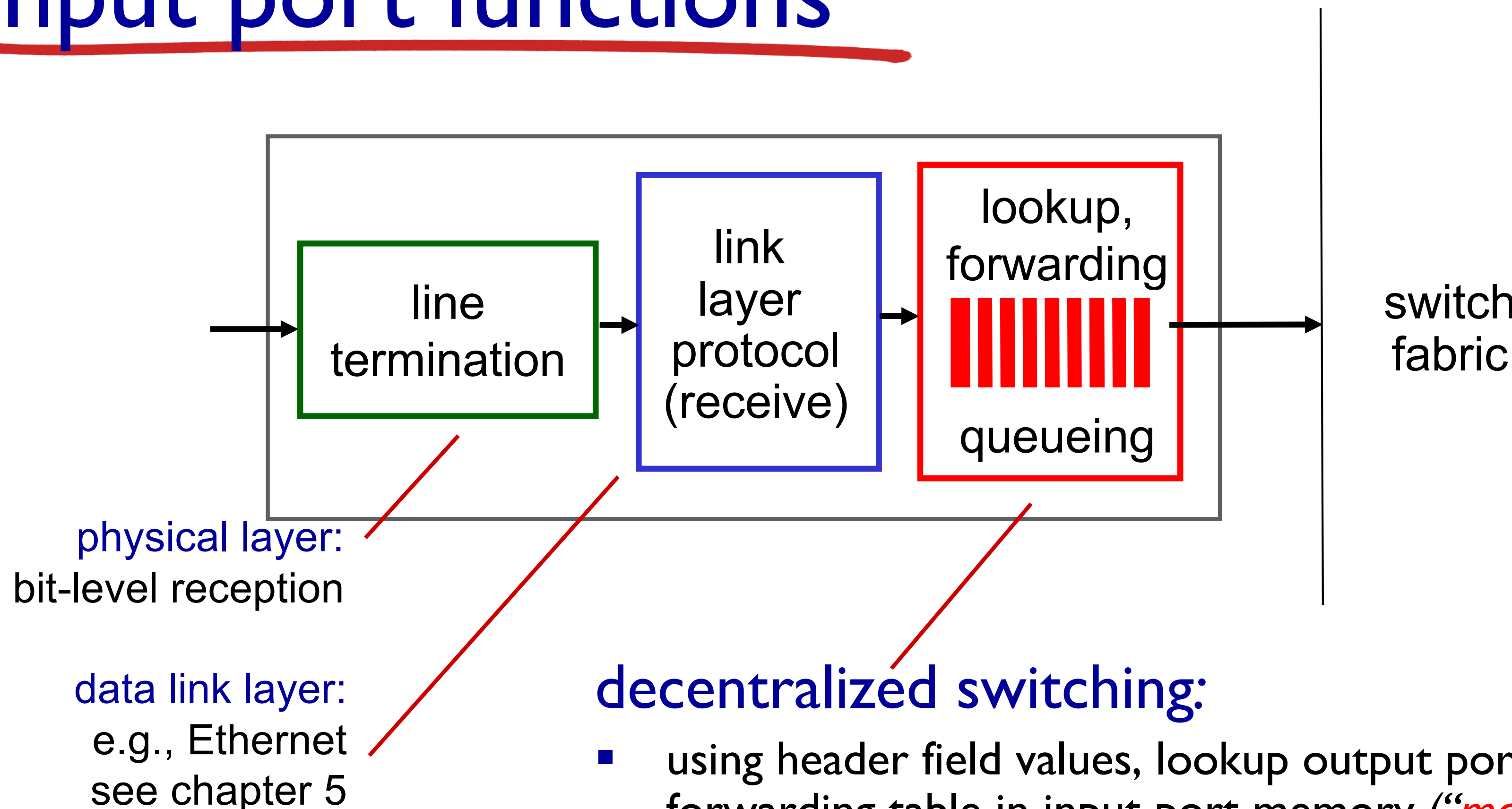
physical layer:
bit-level reception

data link layer:
e.g., Ethernet
see chapter 5

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'
- queuing: if datagrams arrive faster than forwarding rate into switch fabric

Input port functions



decentralized switching:

- using header field values, lookup output port using forwarding table in input port memory (*"match plus action"*)
  - *destination-based forwarding:* forward based only on destination IP address (traditional)
  - *generalized forwarding:* forward based on any set of header field values
- goal: complete input port processing at 'line speed'
- queuing: if datagrams arrive faster than forwarding rate into switch fabric

Destination-based forwarding

forwarding table

Destination Address Range	Link Interface
11001000 00010111 00010000 00000000 through 11001000 00010111 00010111 11111111	0
11001000 00010111 00011000 00000000 through 11001000 00010111 00011000 11111111	1
11001000 00010111 00011001 00000000 through 11001000 00010111 00011111 11111111	2
otherwise	3

Q: but what happens if ranges don't divide up so nicely?

Longest prefix matching

longest prefix matching

when looking for forwarding table entry for given destination address, use *longest* address prefix that matches destination address.

Destination Address Range	Link interface
11001000 00010111 00010*** *********	0
11001000 00010111 00011000 *********	1
11001000 00010111 00011*** *********	2
otherwise	3

examples:

DA: 11001000 00010111 00010110 10100001 which interface?

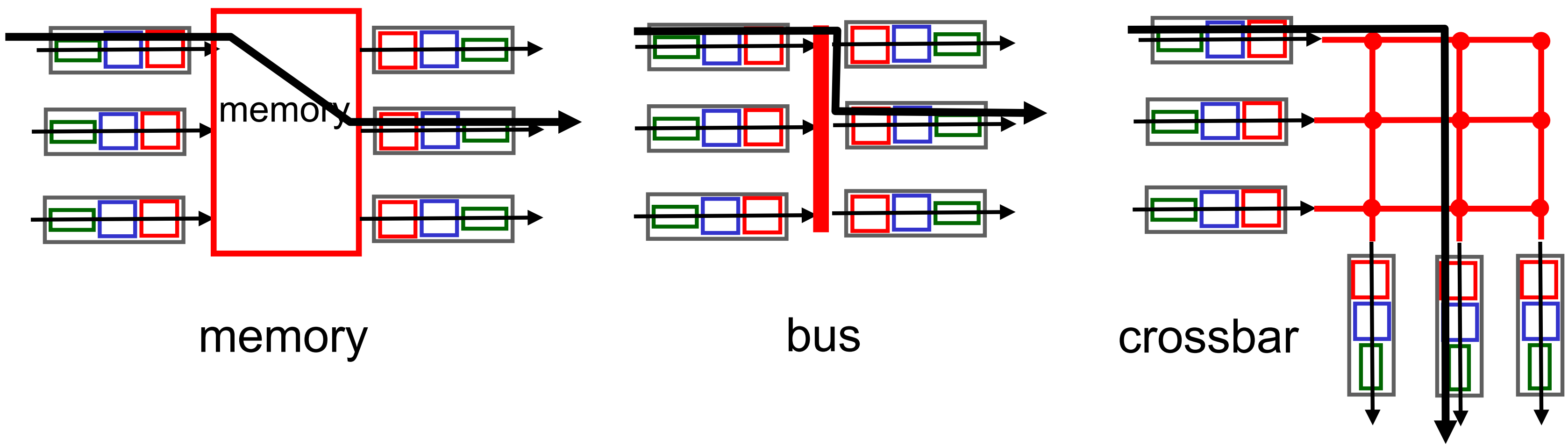
DA: 11001000 00010111 00011000 10101010 which interface?

Longest prefix matching

- we'll see *why* longest prefix matching is used shortly, when we study addressing
- longest prefix matching: often performed using ternary content addressable memories (TCAMs)
  - *content addressable:* present address to TCAM: retrieve address in one clock cycle, regardless of table size
  - Cisco Catalyst: can up ~1M routing table entries in TCAM

Switching fabrics

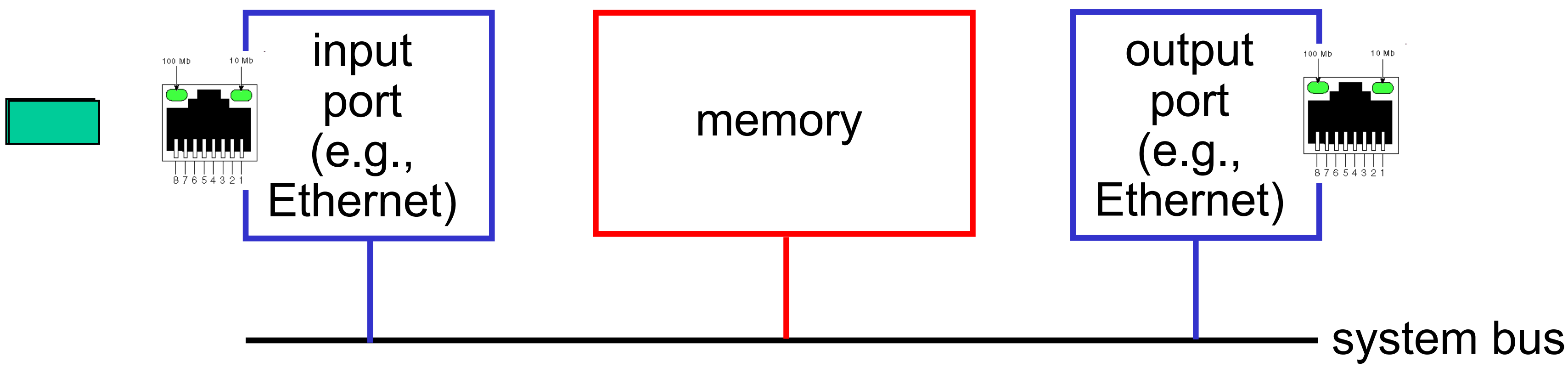
- transfer packet from input buffer to appropriate output buffer
- switching rate: rate at which packets can be transfer from inputs to outputs
  - often measured as multiple of input/output line rate ("speedup")
  - N inputs: switching rate N times line rate desirable
- three types of switching fabrics



Switching via memory

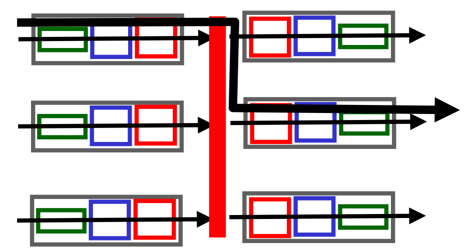
first generation routers:

- 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)



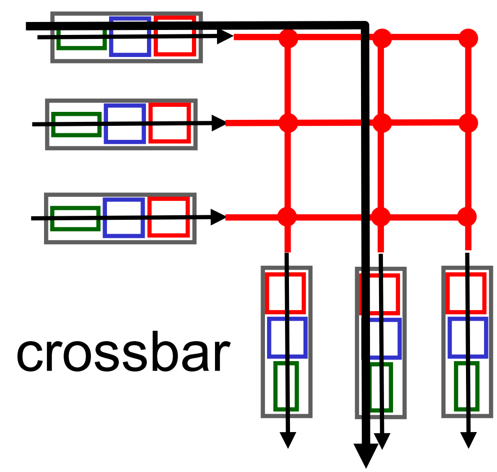
Switching via a bus

- 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 and enterprise routers



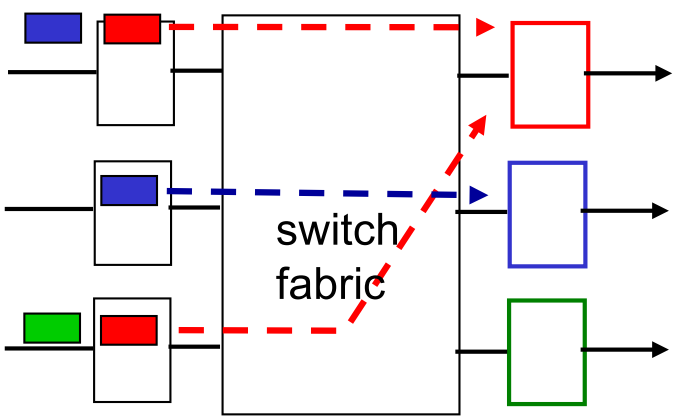
Switching via interconnection network

- overcome bus bandwidth limitations
- banyan networks, crossbar, other interconnection nets initially developed to connect processors in multiprocessor
- advanced design: fragmenting datagram into fixed length cells, switch cells through the fabric.
- Cisco 12000: switches 60 Gbps through the interconnection network

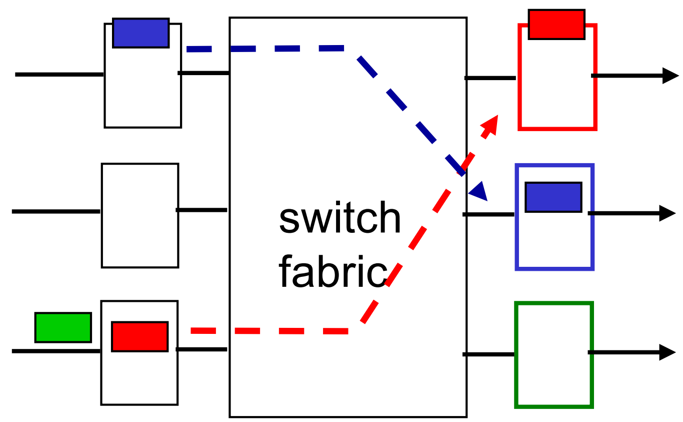


Input port queuing

- fabric slower than input ports combined (speedup<N) →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

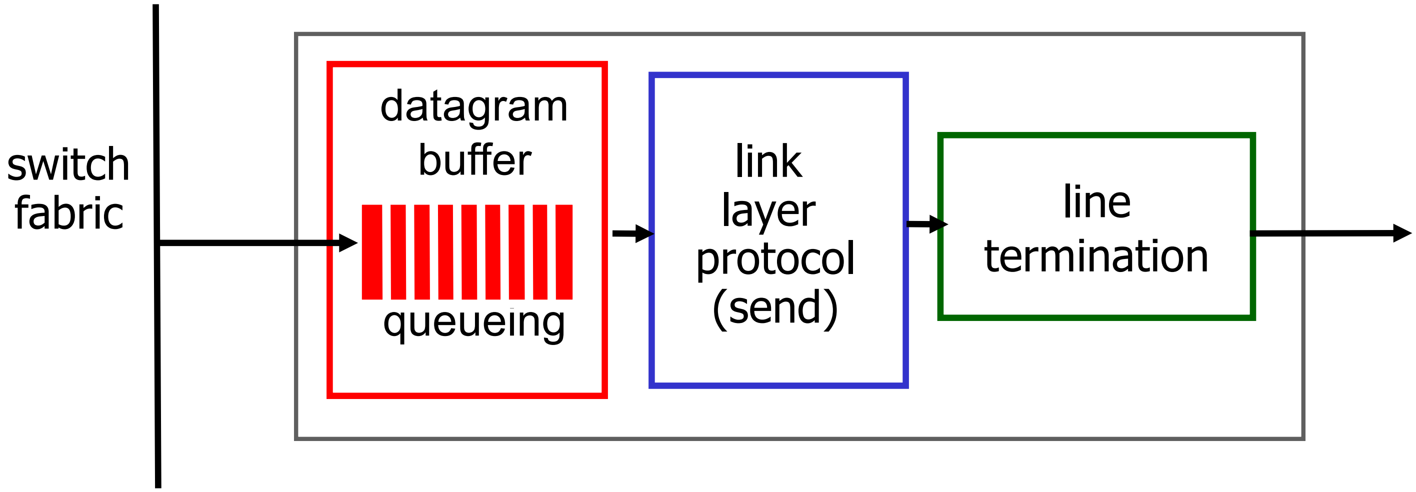


output port contention:
only one red datagram can be transferred.
lower red packet is blocked



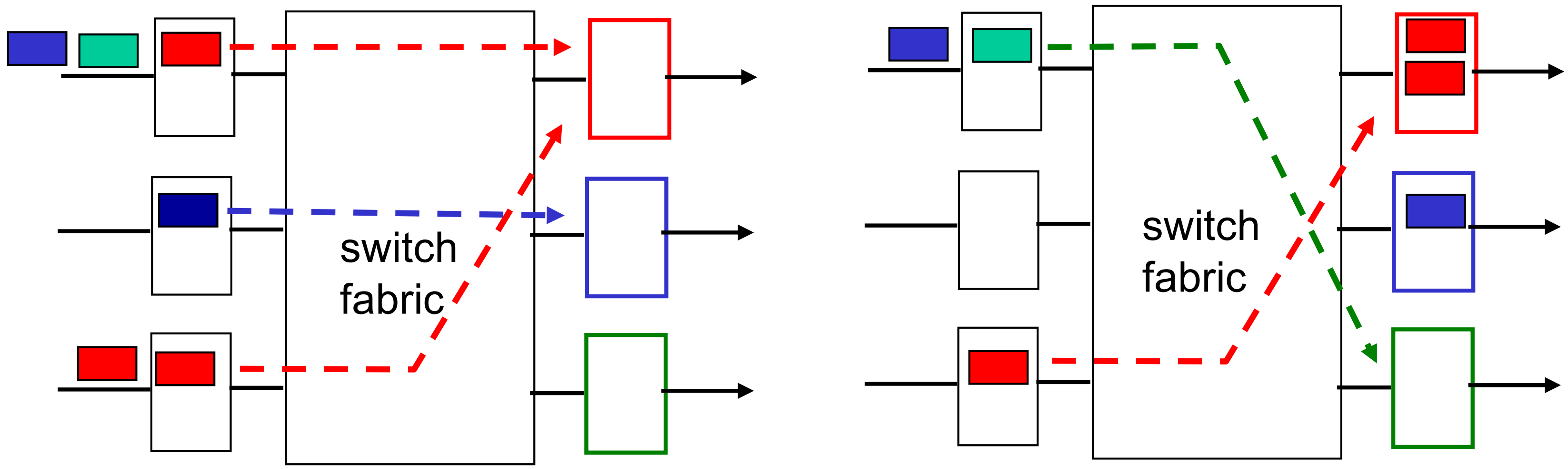
one packet time later:
green packet experiences HOL blocking

Output ports



- *buffering* required when datagrams arrive from fabric faster than the transmission rate
  - Datagrams can be dropped from buffers, due to congestion
- *scheduling discipline* chooses among queued datagrams for transmission
  - Determines who gets best performance, network neutrality

Output port queueing



at *t*, packets more from input to output

one packet time later

- buffering when arrival rate via switch exceeds output line speed
- *queueing (delay) and loss due to output port buffer overflow!*

How much buffering?

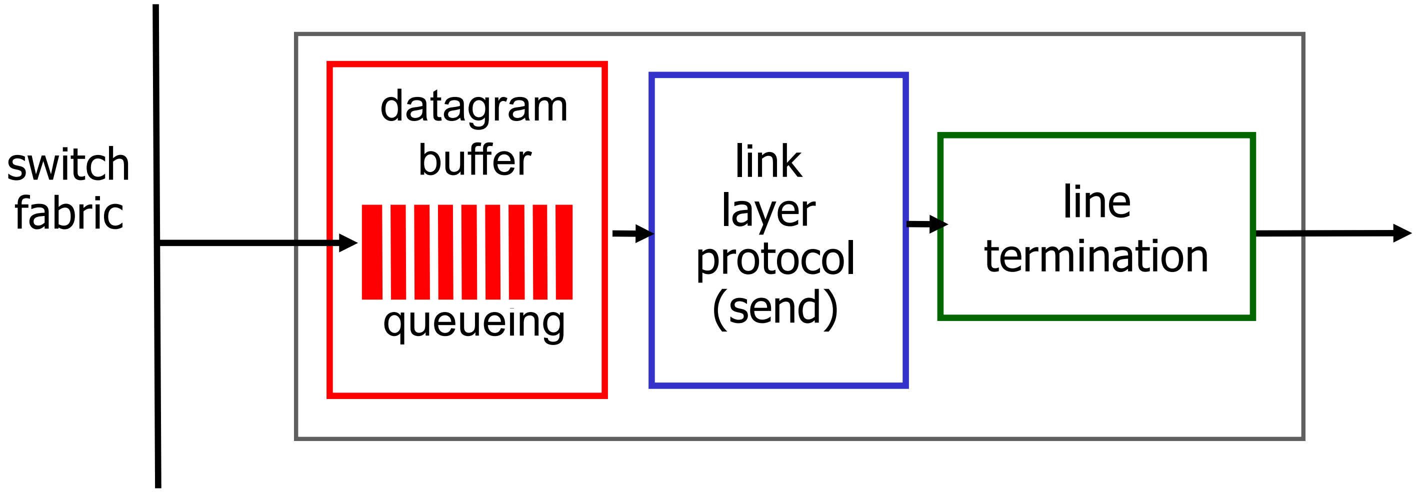
- RFC 3439 rule of thumb: average buffering = "typical" RTT (say 250 msec) * link capacity C
  - e.g., C = 10 Gpbs link → 2.5 Gbit buffer

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

$$\frac{\mathrm{RTT} \cdot \mathrm{C}}{\sqrt{\mathrm{N}}}$$

  - *Appenzeller et al.*, "Sizing Routing buffers", SIGCOMM 2004, http://guido.appenzeller.net/publications.html

Output ports



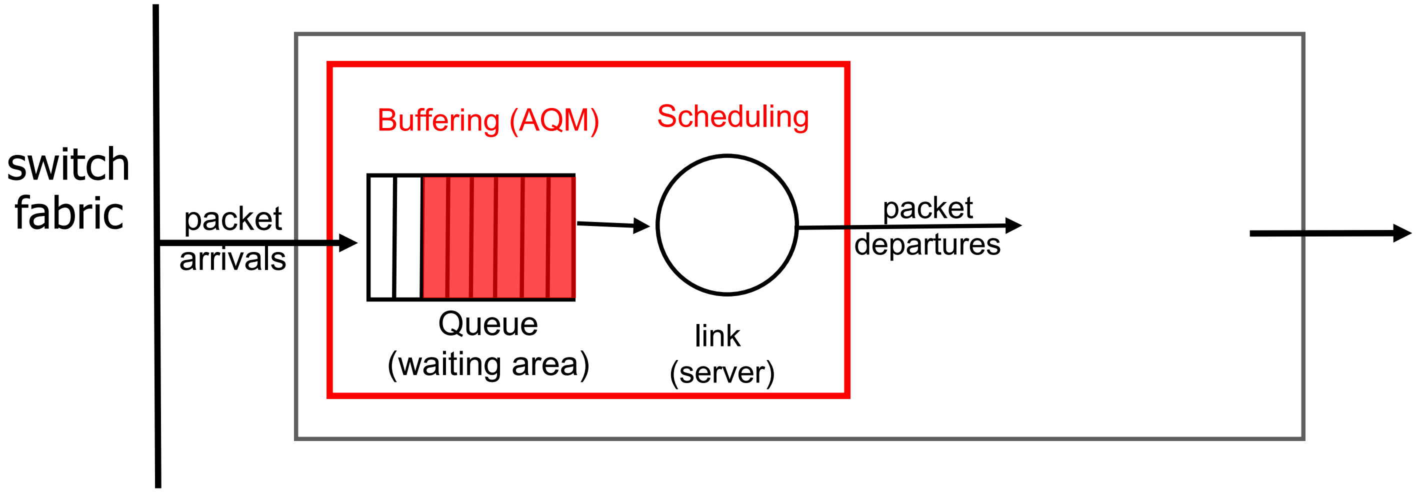
1. Buffering required when datagrams arrive from fabric faster than the transmission rate

- Datagrams can be dropped from buffers, due to congestion

2. Scheduling discipline chooses among queued datagrams for transmission

- Determines who gets best performance, network neutrality

Output ports



1. Buffering/Active Queue Management (AQM)

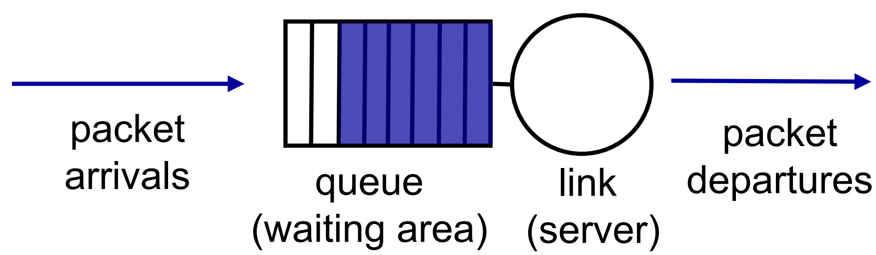
- *Decide how many queues? what packet to drop or mark?*
- *AQM –TCP. interaction*

2. Scheduling discipline

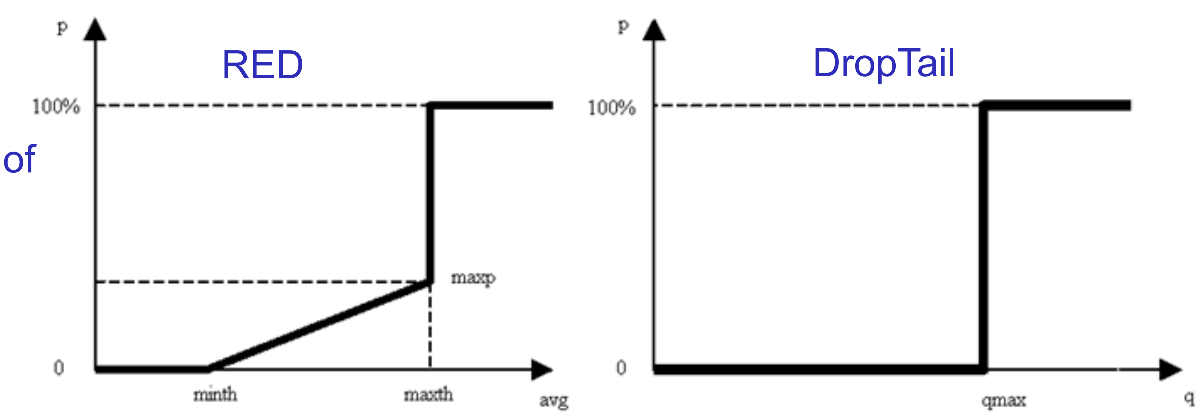
- chooses which queued datagrams to transmit

1. Active Queue Management

- *Buffer packets: needed* if packets arrive faster than depart
- *AQM:* what to drop or mark when queues build up
  - *discard policy:* if a packet arrives to full queue: what to discard?
    - *DropTail:* drop most recent arriving packet, if there is no room.
    - *RED:* random early drop: drop randomly
  - or *mark packets* when queues build up (e.g. ECN)



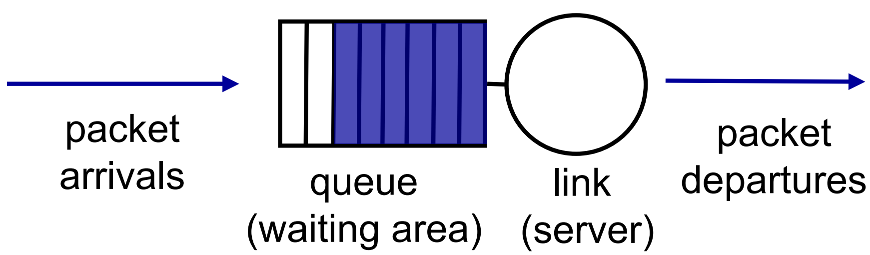
Probability of dropping a packet



Queue Size

2a. Scheduling Policy: FIFO

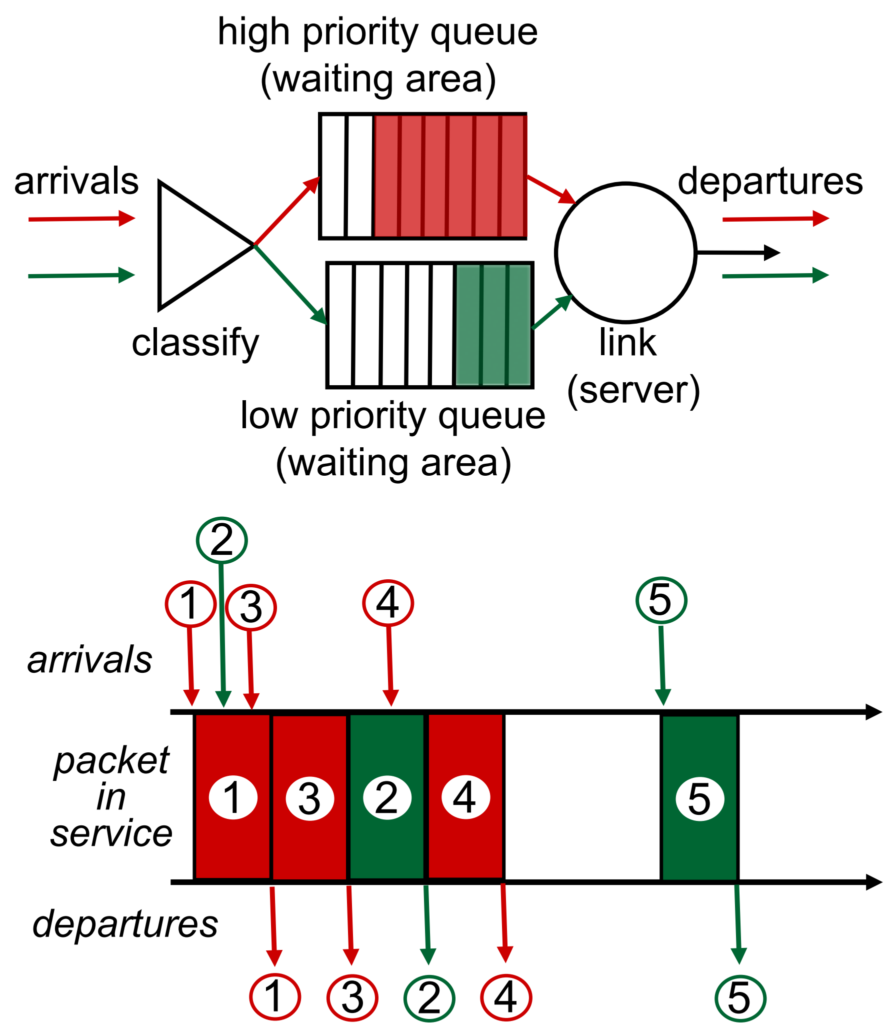
- *scheduling:* choose next packet to send on link
- *FIFO (first in first out) scheduling:* send in order of arrival to queue
  - The most commonly implemented one



2b. Scheduling policy: Priority

priority scheduling: send highest priority queued packet

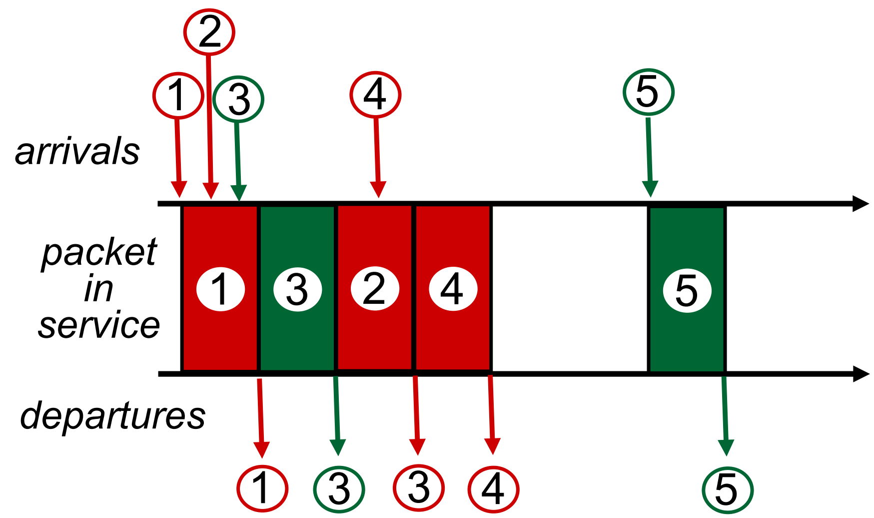
- multiple *classes*, with different priorities
  - class may depend on marking or other header info, e.g. IP source/dest, port numbers, etc.
  - real world example?

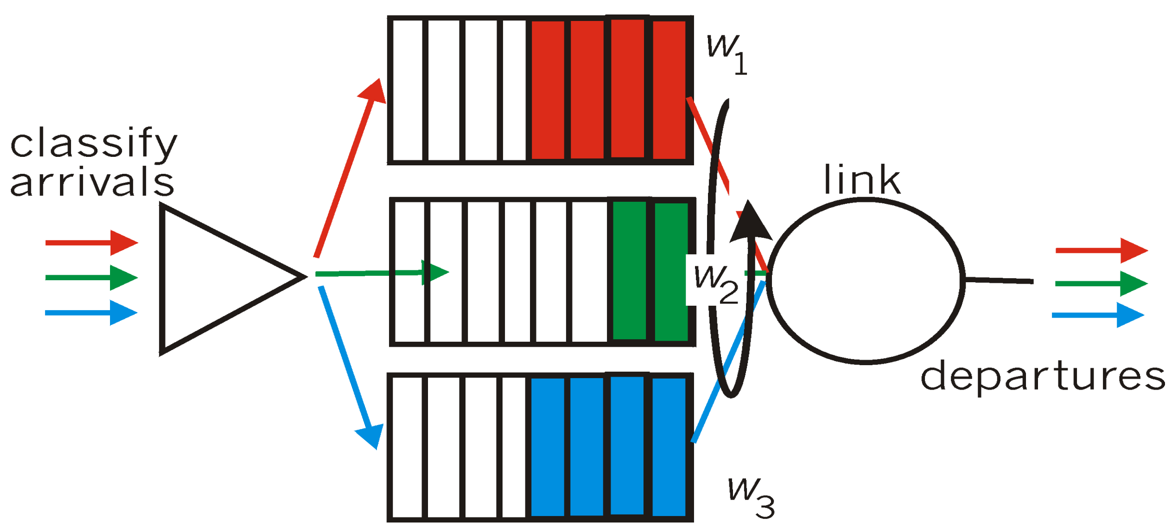


2c. Scheduling policy: RR

Round Robin (RR) scheduling:

- multiple classes
- cyclically scan class queues, sending one complete packet from each class (if available)
- $w_1=w_2=w_3$

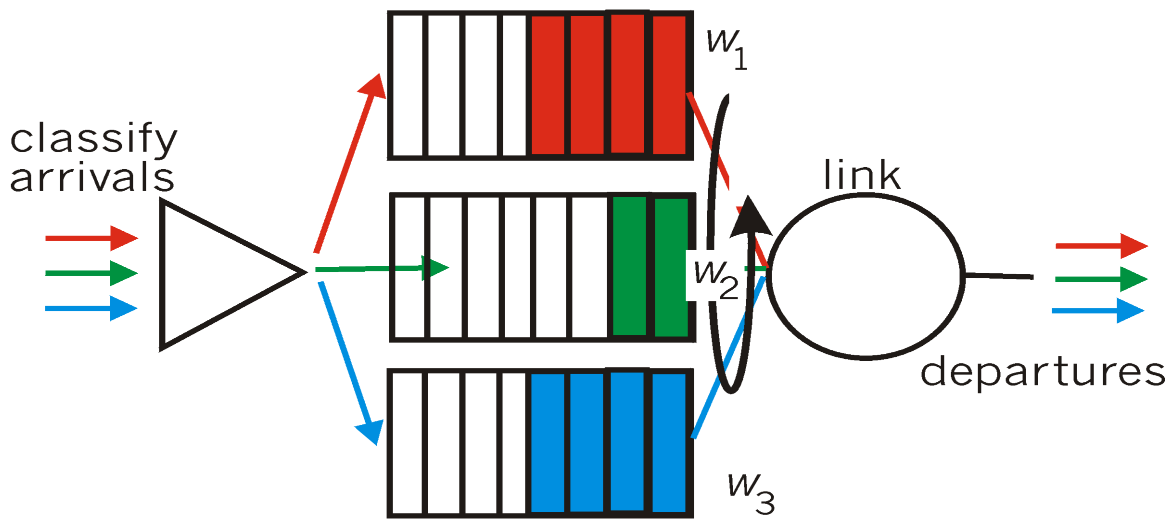




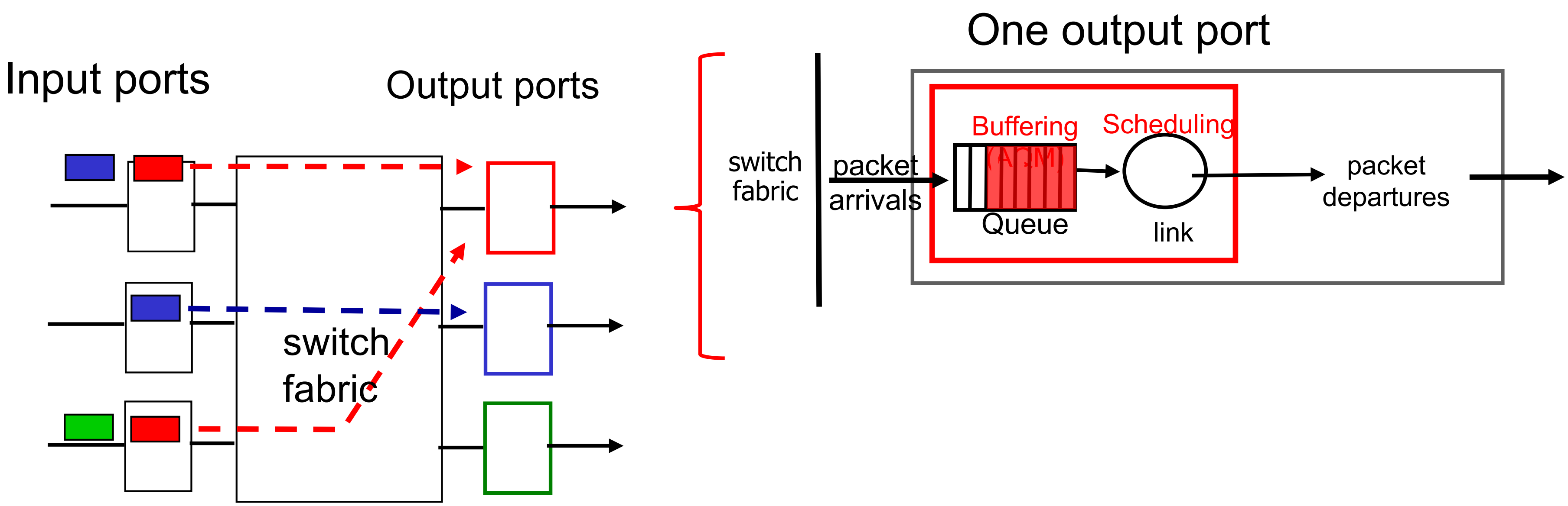
2c. Scheduling policy: WFQ

Weighted Fair Queuing (WFQ):

- generalized Round Robin
- each class gets weighted amount of service in each cycle



Summary: what is inside the router



I-clicker: Consider a NxN router: Each of the input and output links have the same speed W (bps). Which of the following is FALSE

- ☐ A: There are buffers at the input ports
- ☐ B: There are buffers at the output ports
- ☐ C: If the switch fabric has speedup N, there may or may not be input queues.
- ☐ D: If the switch fabric has speedup N, there are no output queues.
- ☐ E: Whether there are output queues, it depends on the traffic pattern.