Switch and Router Architectures

EECS 489 Computer Networks

http://www.eecs.umich.edu/~zmao/eecs489

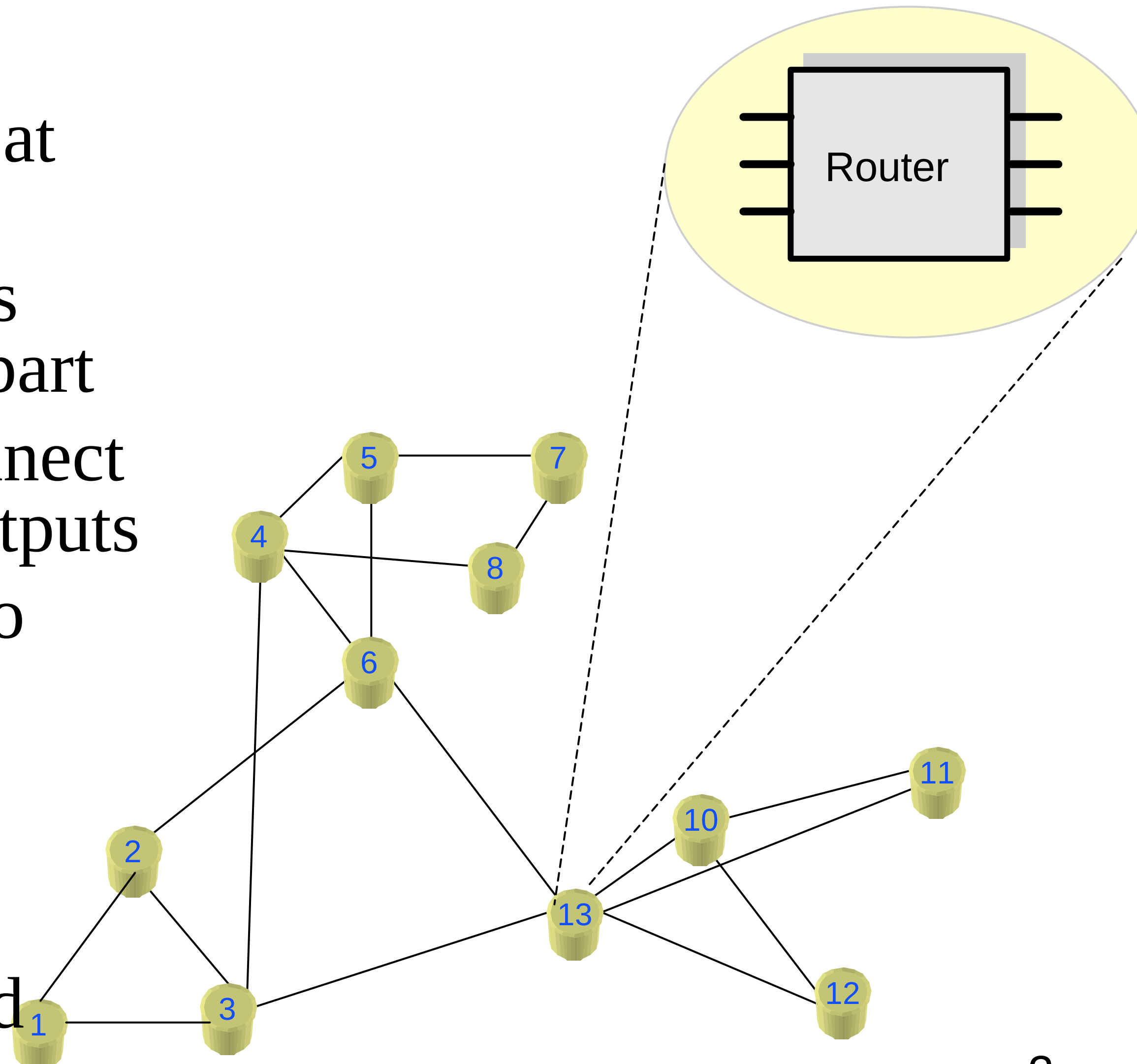
Z. Morley Mao

Tuesday Sept 28, 2004

Acknowledgement: Some slides adapted from Kurose&Ross and Katz&Stoica

IP Routers

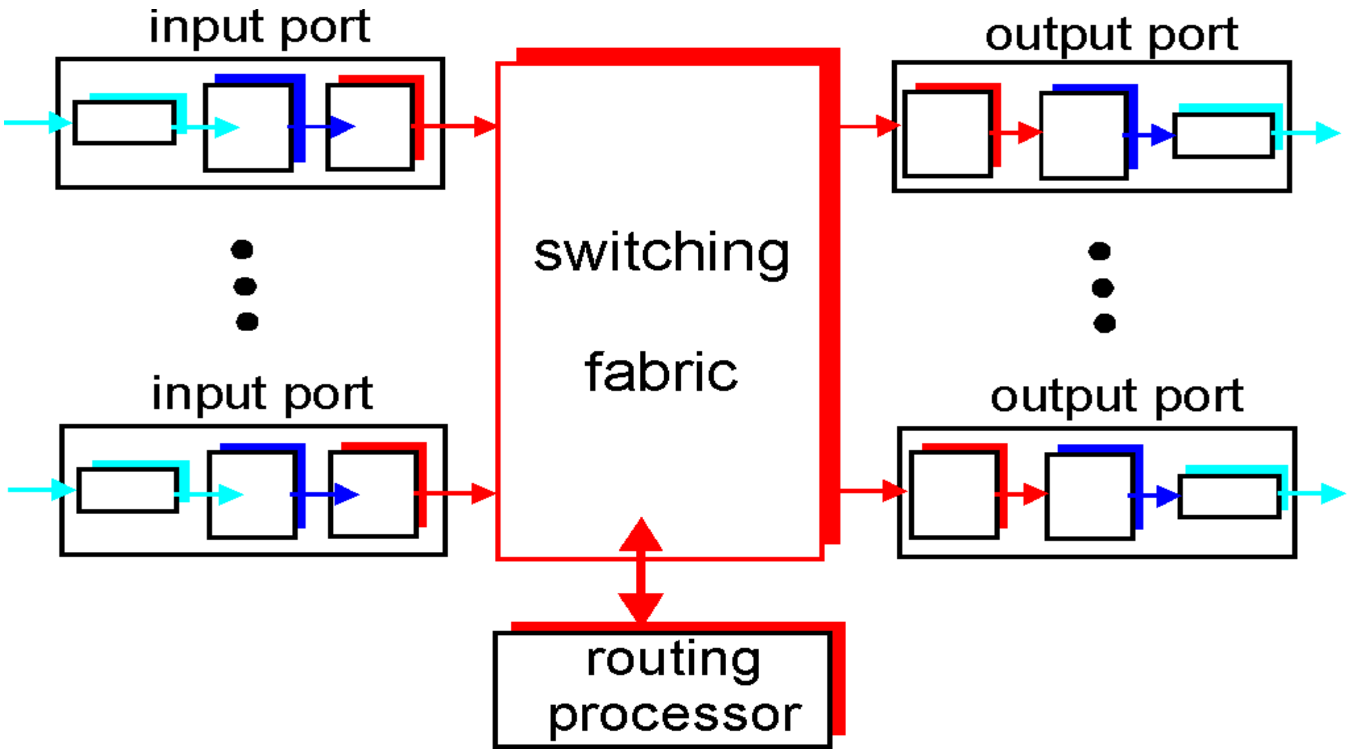
- Router consists
- Set of input interfaces at which packets arrive
- Set of output interfaces from which packets depart
- Some form of interconnect connecting inputs to outputs
- Router implements two main functions
- Forward packet to corresponding output interface
- Manage bandwidth and buffer space resources



Router Architecture Overview

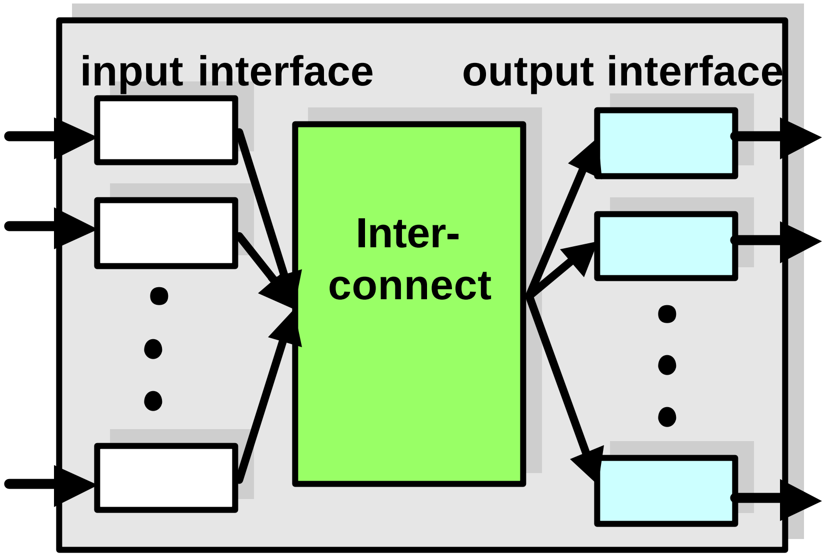
Two key router functions:

- run routing algorithms/protocol (RIP, OSPF, BGP)
- *forwarding* datagrams from incoming to outgoing link

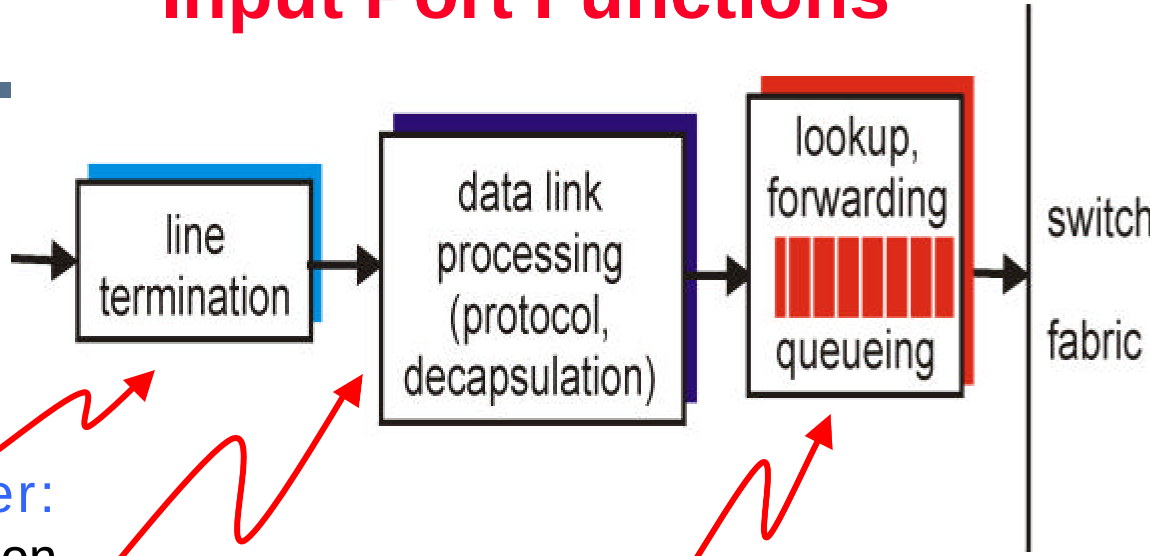


Generic Architecture

- Input and output interfaces are connected through an interconnect
- Interconnect can be implemented by
  - Shared memory
    - Low capacity routers (e.g., PC-based routers)
  - Shared bus
    - Medium capacity routers
  - Point-to-point (switched) bus
    - High capacity routers



Input Port Functions



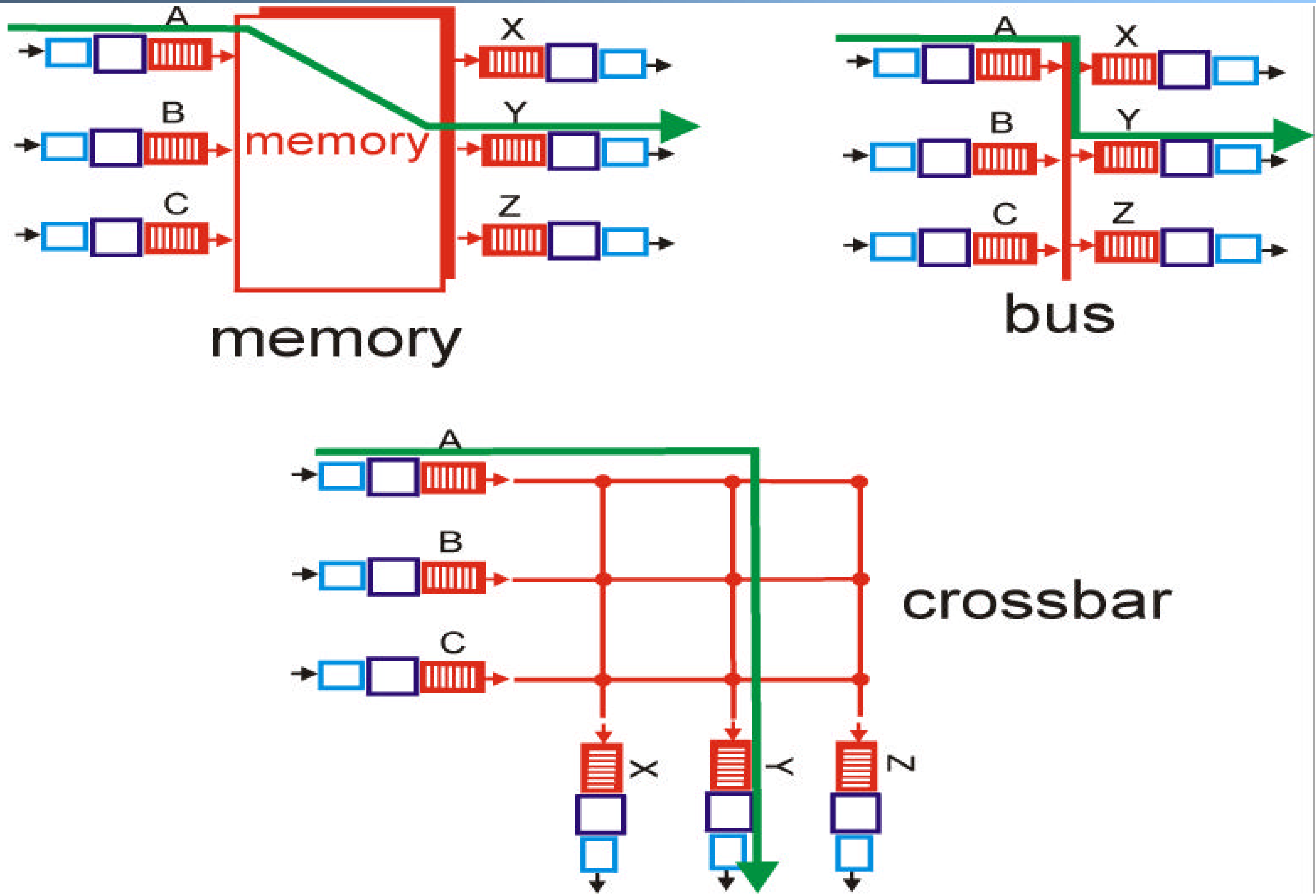
Physical layer:
bit-level reception

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

Decentralized switching:

- given datagram dest., lookup output port using forwarding table in input port memory
- goal: complete input port processing at 'line speed'
- queuing: if datagrams arrive faster than forwarding rate into switch fabric

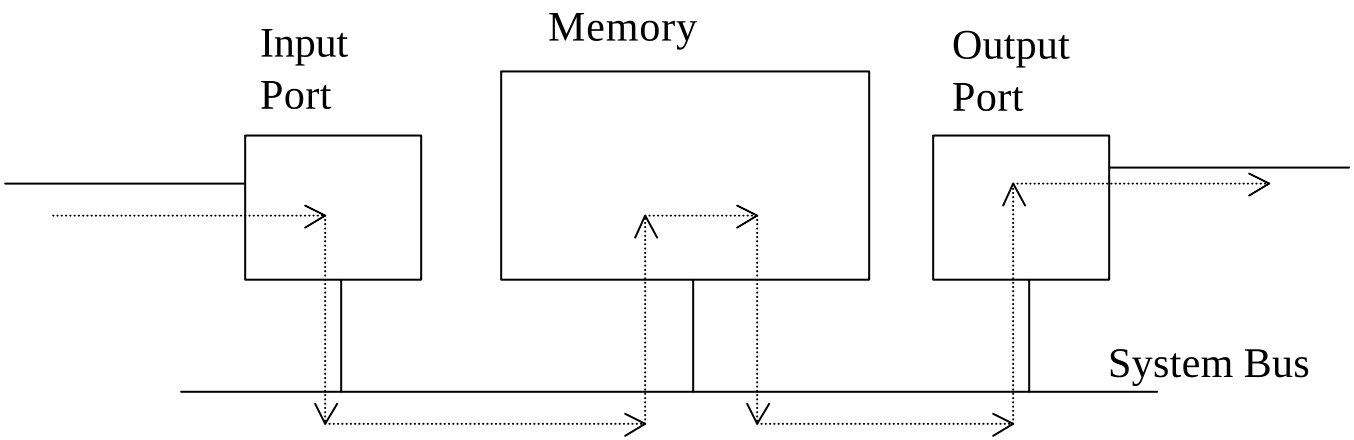
Three types of switching fabrics



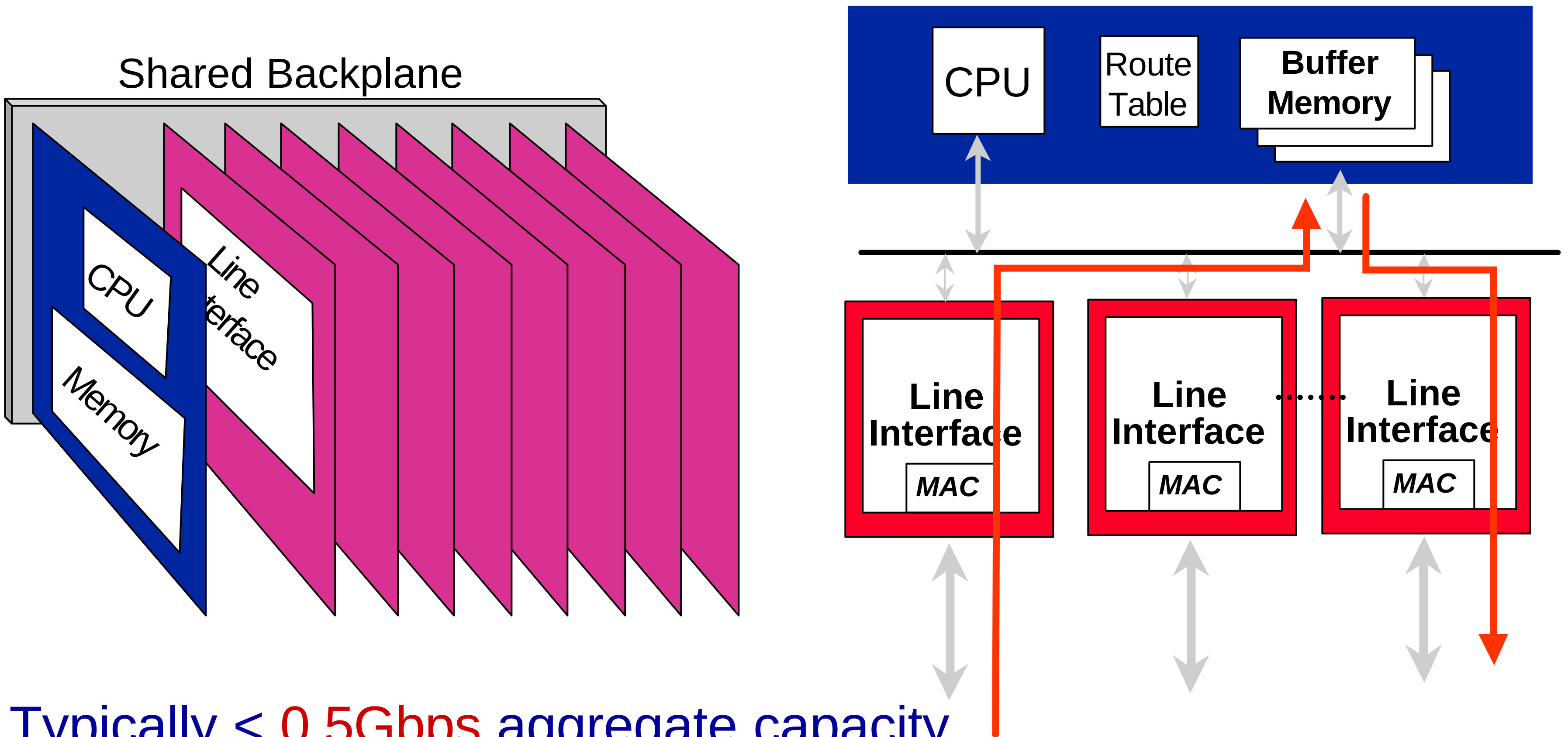
Switching Via Memory

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)



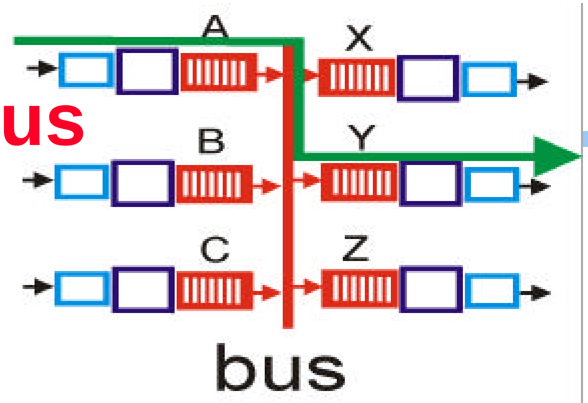
Shared Memory (1st Generation)



Typically < 0.5Gbps aggregate capacity
Limited by rate of shared memory

(* Slide by Nick McKeown)

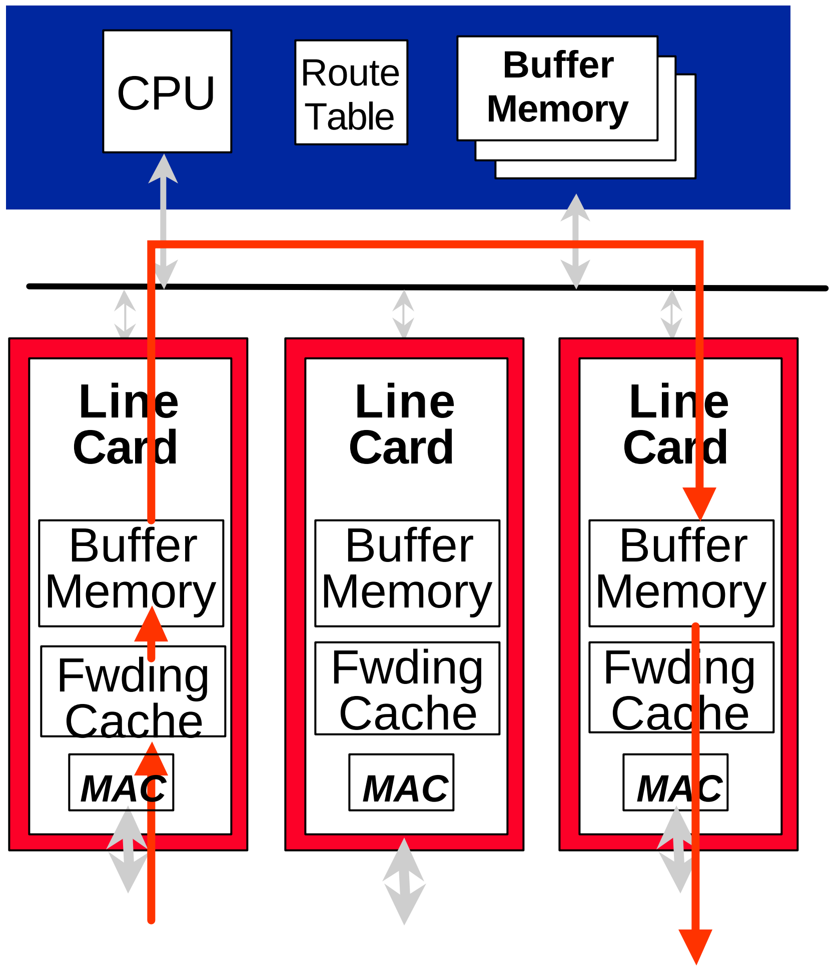
Switching Via a Bus



- datagram from input port memory to output port memory via a shared bus
- bus contention: switching speed limited by bus bandwidth
- 1 Gbps bus, Cisco 1900: sufficient speed for access and enterprise routers (not regional or backbone)

Shared Bus (2nd Generation)

Typically < 5Gb/s aggregate capacity; Limited by shared bus



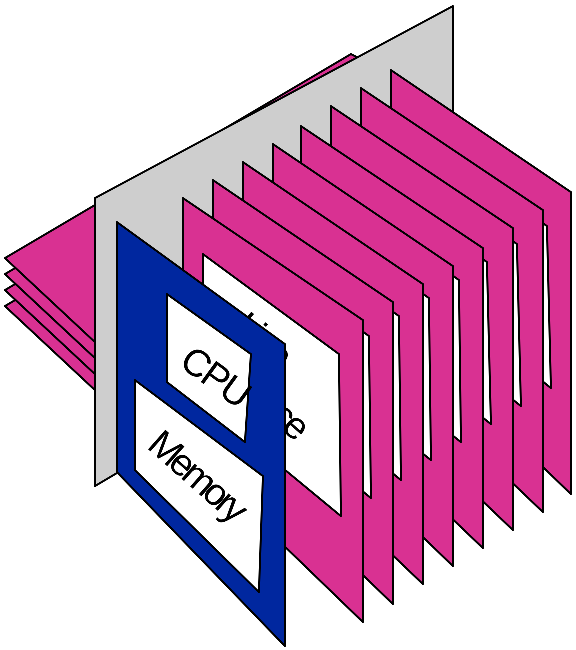
(* Slide by Nick McKeown)

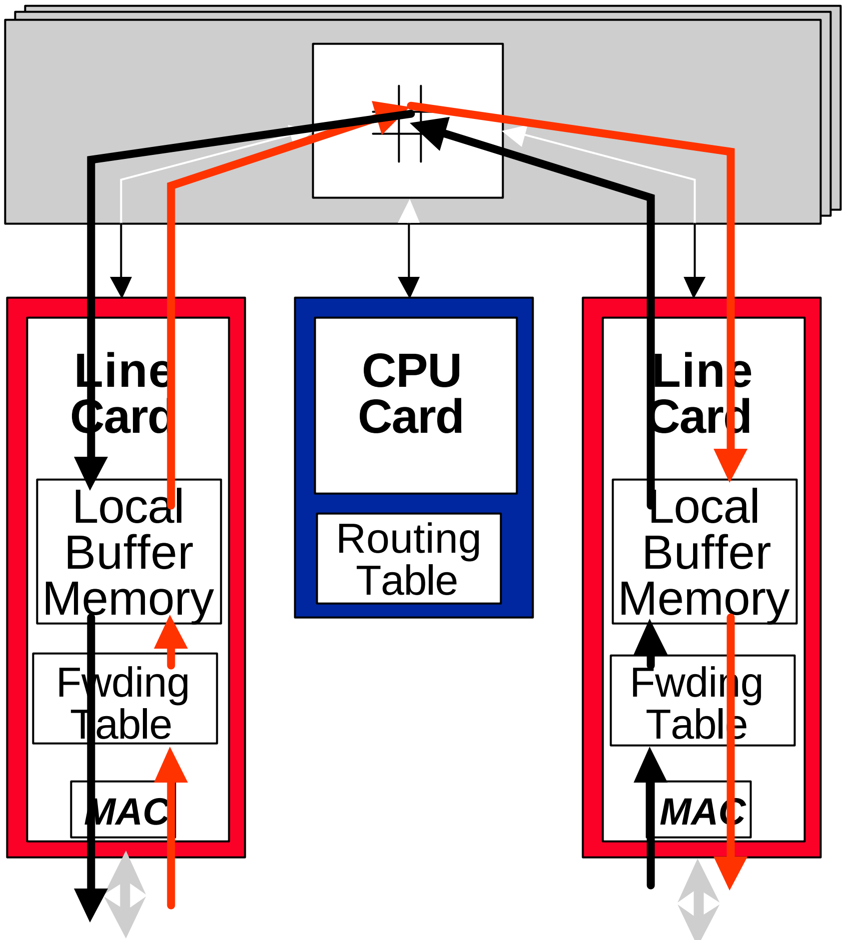
Switching Via An Interconnection Network

- overcome bus bandwidth limitations
- Banyan networks, 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 Gbps through the interconnection network

Point-to-Point Switch (3rd Generation)

Switched Backplane



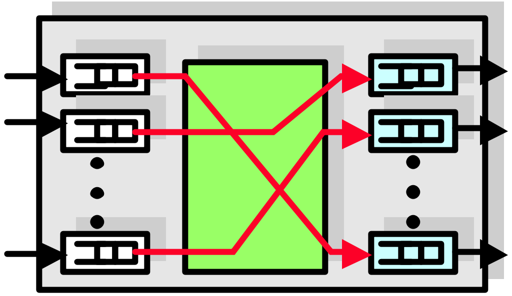


Typically < 50Gbps aggregate capacity

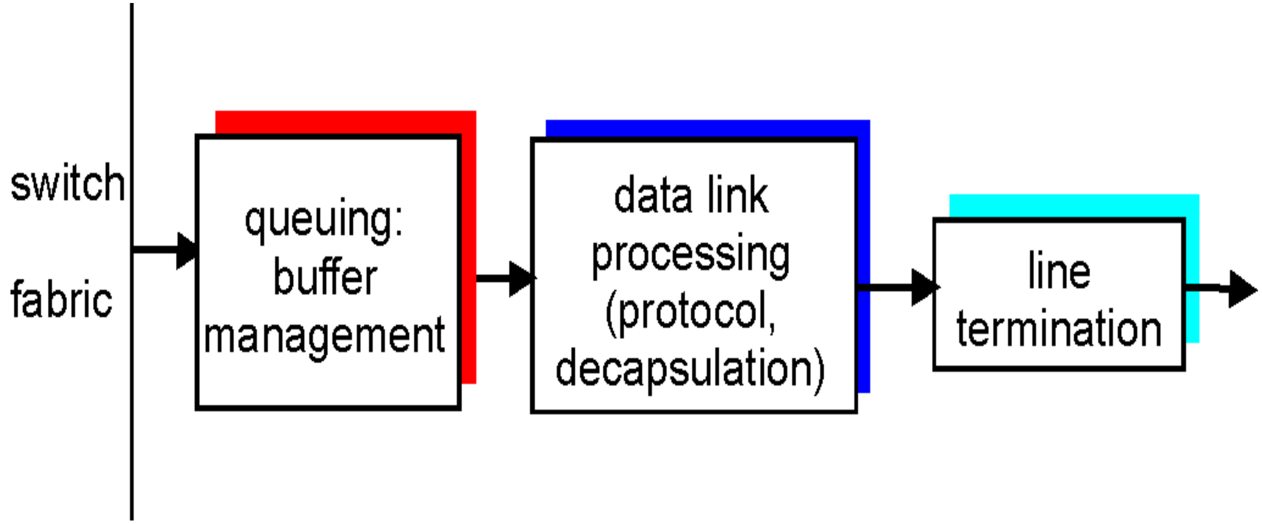
(*Slide by Nick McKeown)

Interconnect

- Point-to-point switch allows to simultaneously transfer a packet between any two disjoint pairs of input-output interfaces
- Goal: come-up with a schedule that
  - Provide Quality of Service
  - Maximize router throughput
- Challenges:
  - Address head-of-line blocking at inputs
  - Resolve input/output speedups contention
  - Avoid packet dropping at output if possible
- Note: packets are fragmented in fix sized cells at inputs and reassembled at outputs

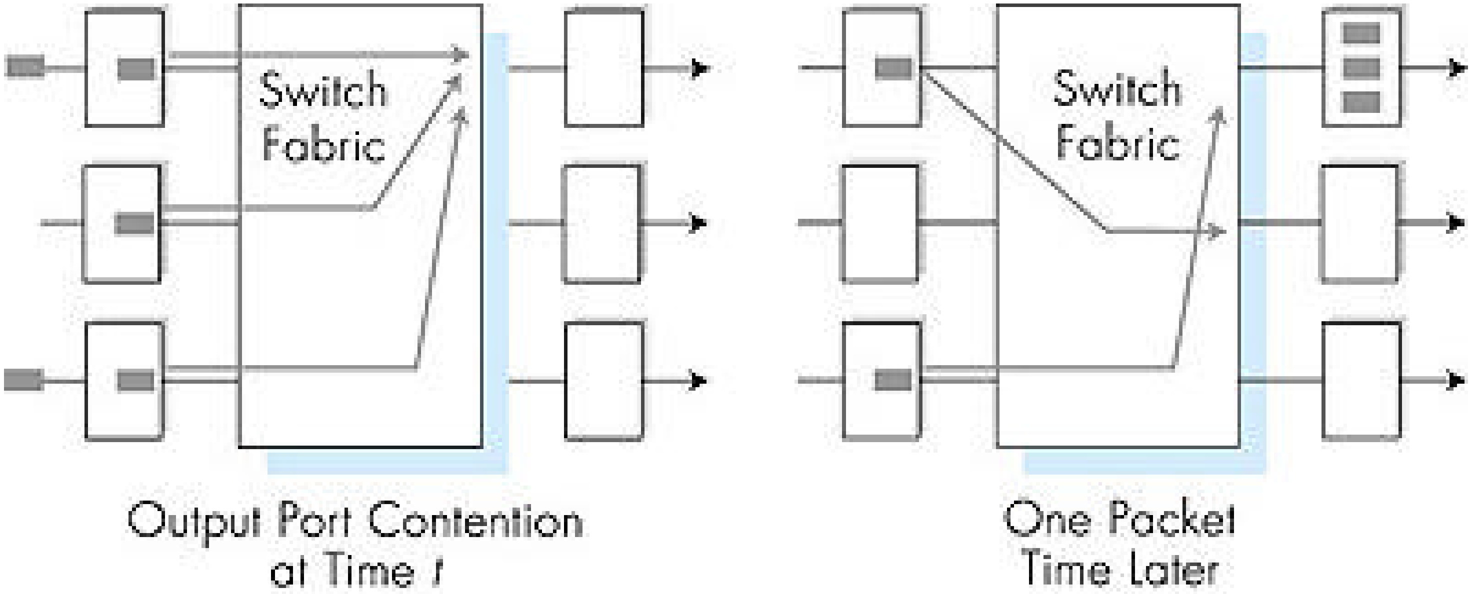


Output Ports



- *Buffering* required when datagrams arrive from fabric faster than the transmission rate
- *Scheduling discipline* chooses among queued datagrams for transmission

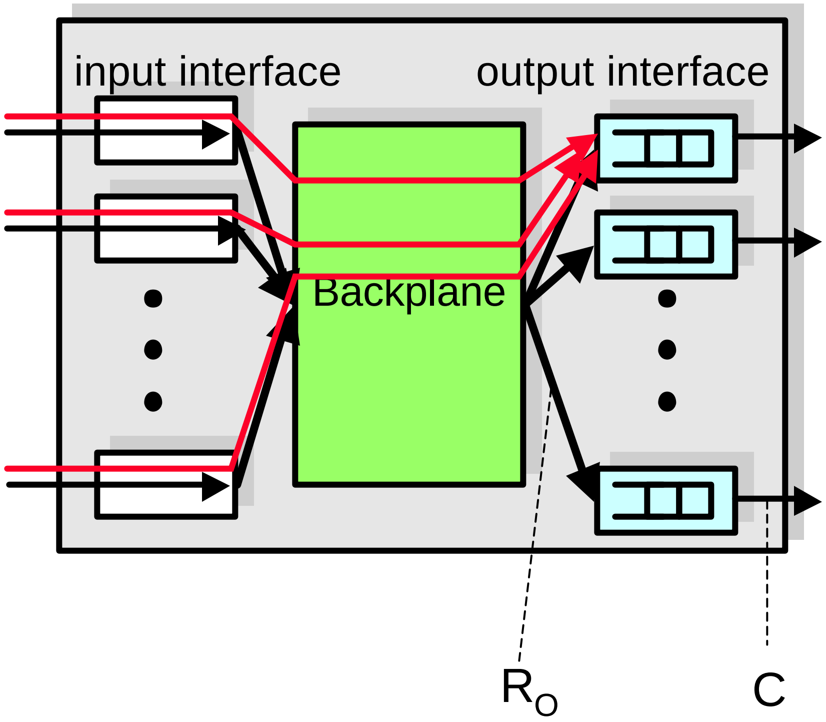
Output port queueing



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

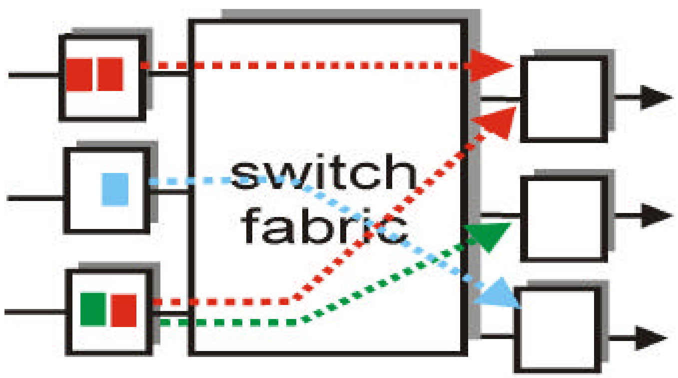
Output Queued Routers

- Only output interfaces store packets
- Advantages
  - Easy to design algorithms: only one congestion point
- Disadvantages
  - Requires an output speedup of N, where N is the number of interfaces → not feasible

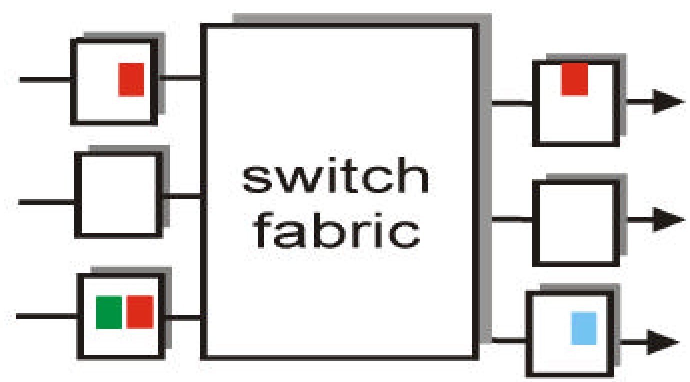


Input Port Queuing

- Fabric slower than input ports combined -> queueing may occur at input queues
- Head-of-the-Line (HOL) blocking: queued datagram at front of queue prevents others in queue from moving forward
- *queueing delay and loss due to input buffer overflow!*



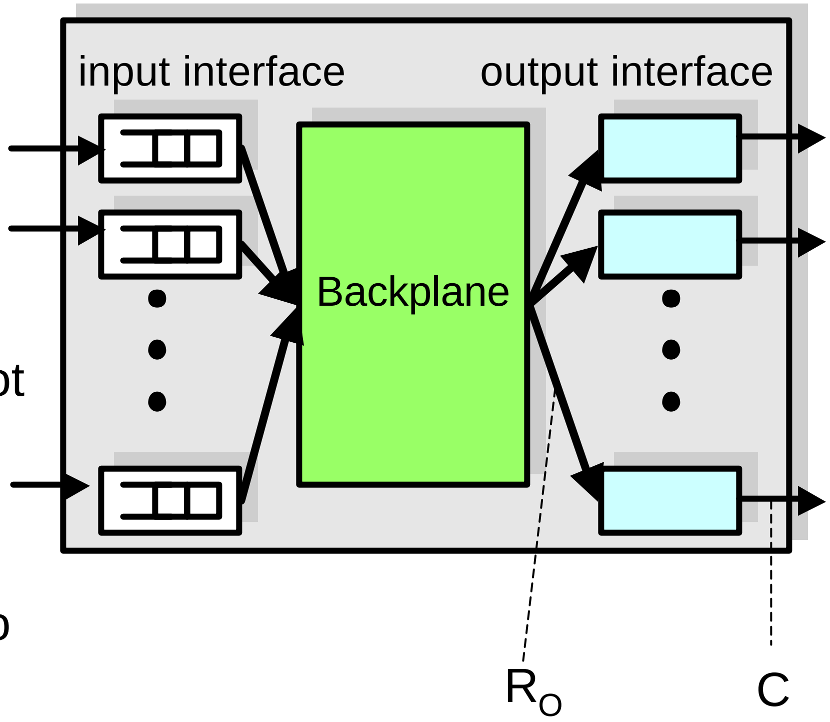
output port contention at time t - only one red packet can be transferred



green packet experiences HOL blocking

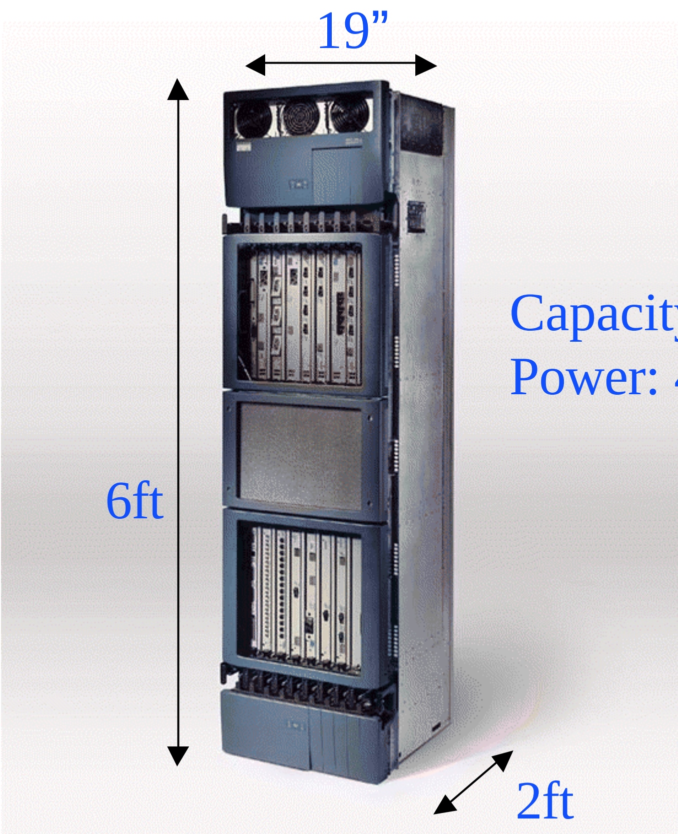
Input Queued Routers

- Only input interfaces store packets
- Advantages
  - Easy to built
    - Store packets at inputs if contention at outputs
  - Relatively easy to design algorithms
    - Only one congestion point, but not output…
    - Need to implement backpressure
- Disadvantages
  - Hard to achieve utilization → 1 (due to output contention, head-of-line blocking)
    - However, theoretical and simulation results show that for realistic traffic an input/output speedup of 2 is enough to achieve utilizations close to 1



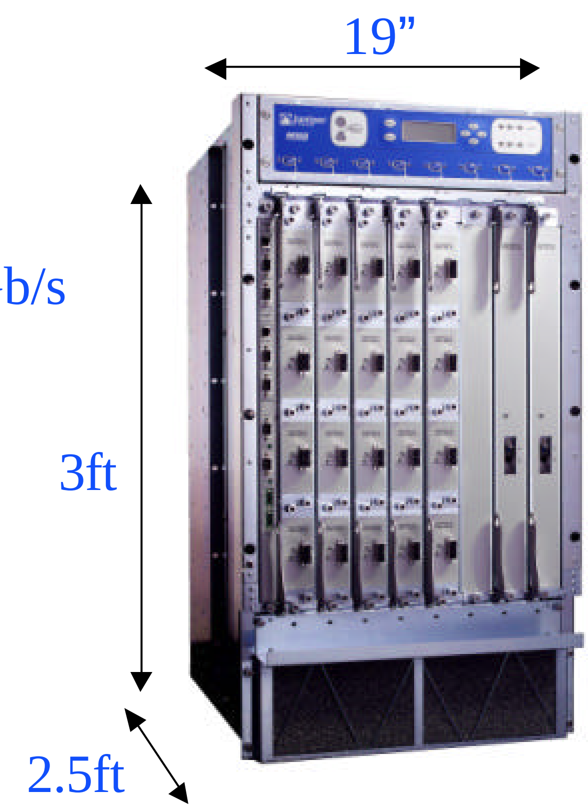
What a Router Looks Like

Cisco GSR 12416



Capacity: 160Gb/s
Power: 4.2kW

Juniper M160

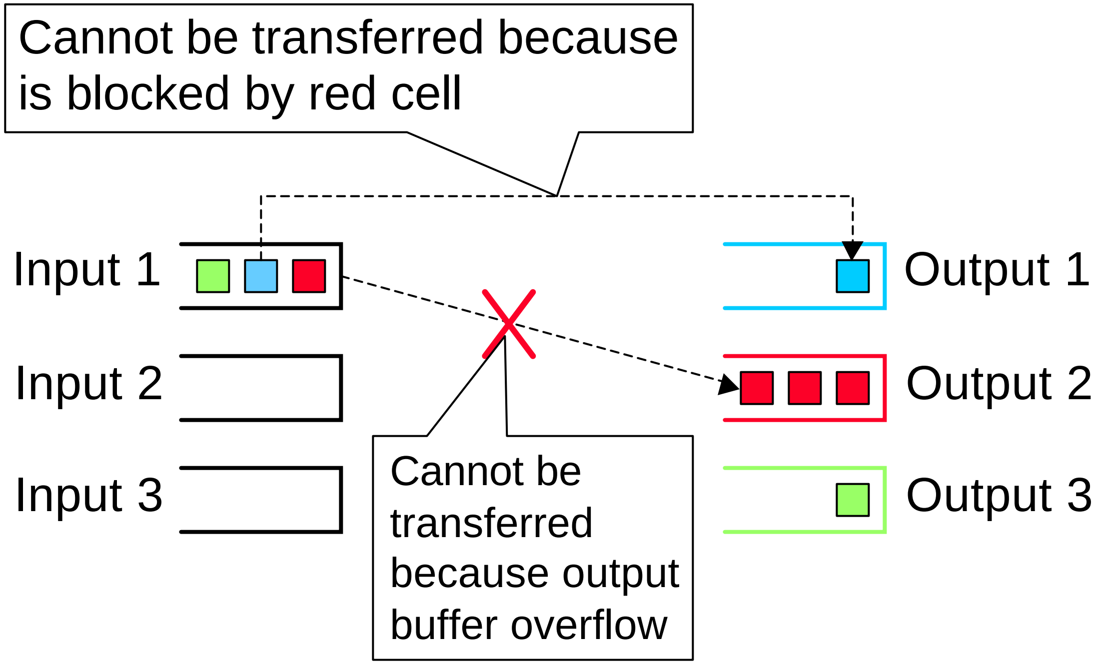


Capacity: 80Gb/s
Power: 2.6kW

Slide by Nick McKeown

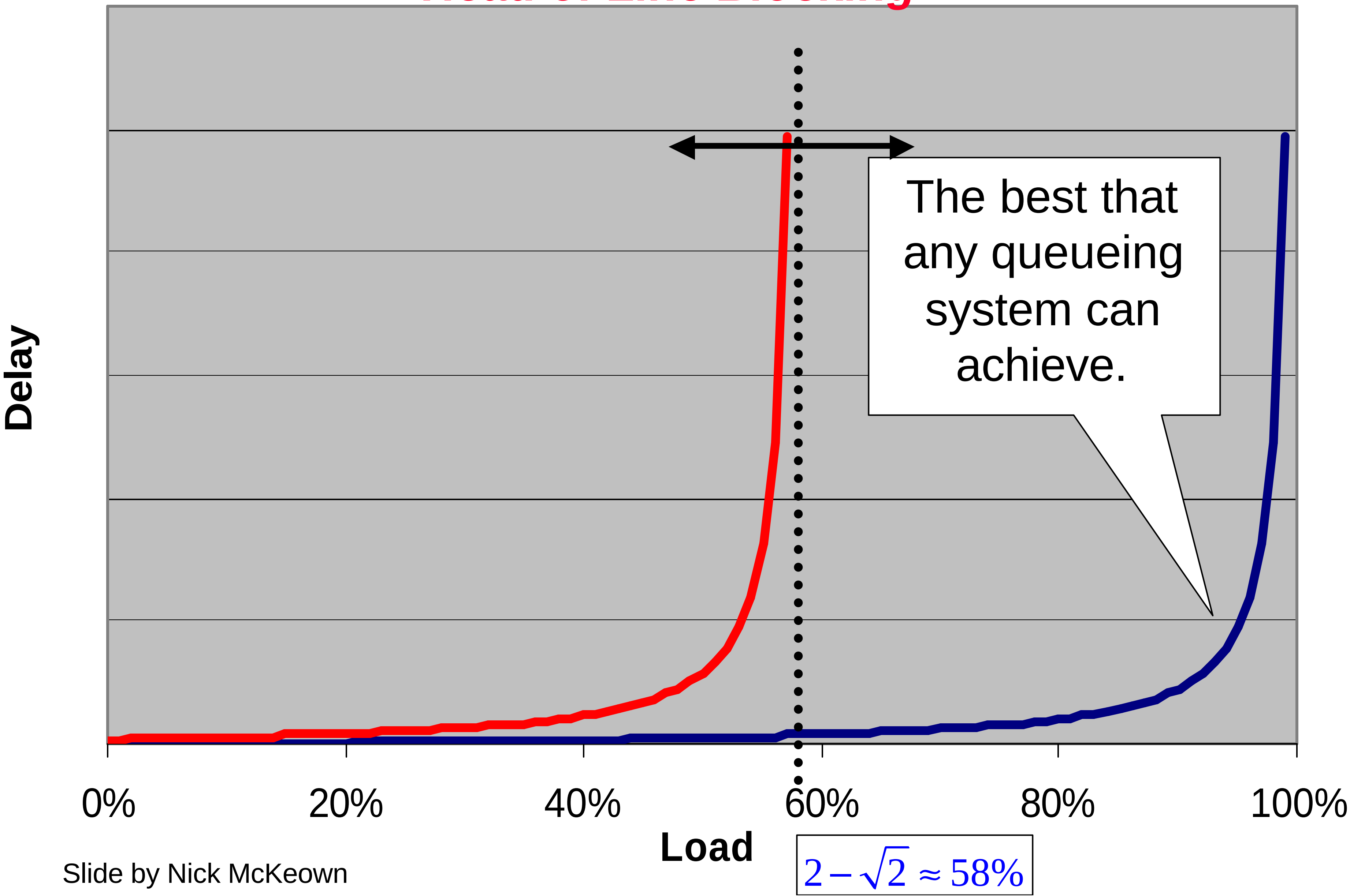
Head-of-line Blocking

- Cell at head of an input queue cannot be transferred, thus blocking the following cells



A Router with Input Queues

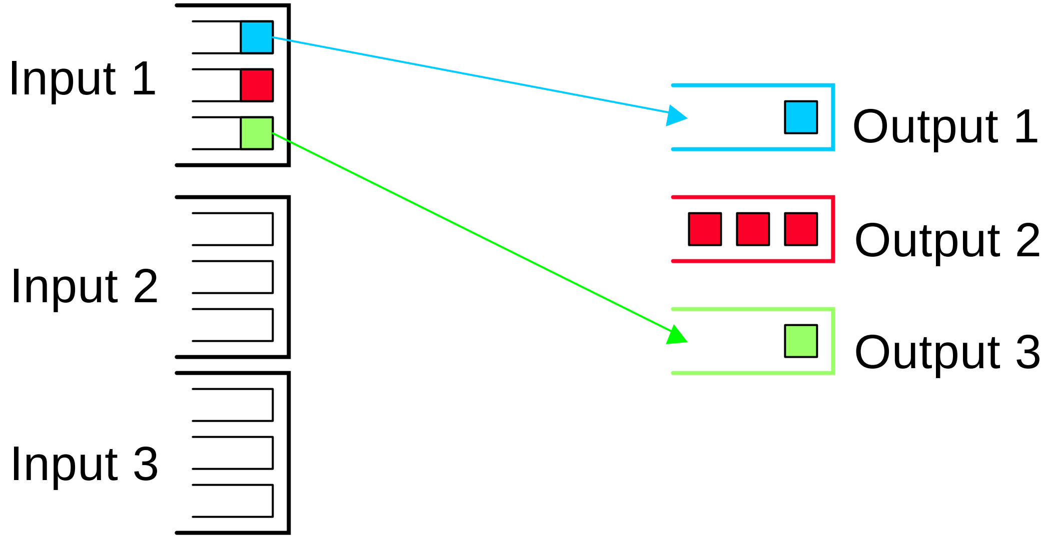
Head of Line Blocking



Slide by Nick McKeown

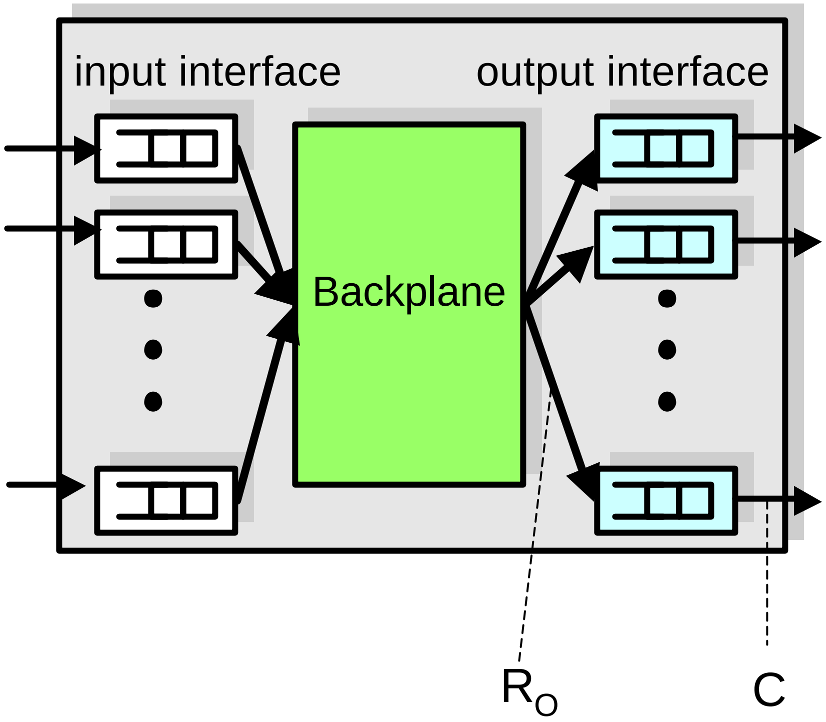
Solution to Avoid Head-of-line Blocking

- Maintain at each input N virtual queues, i.e., one per output



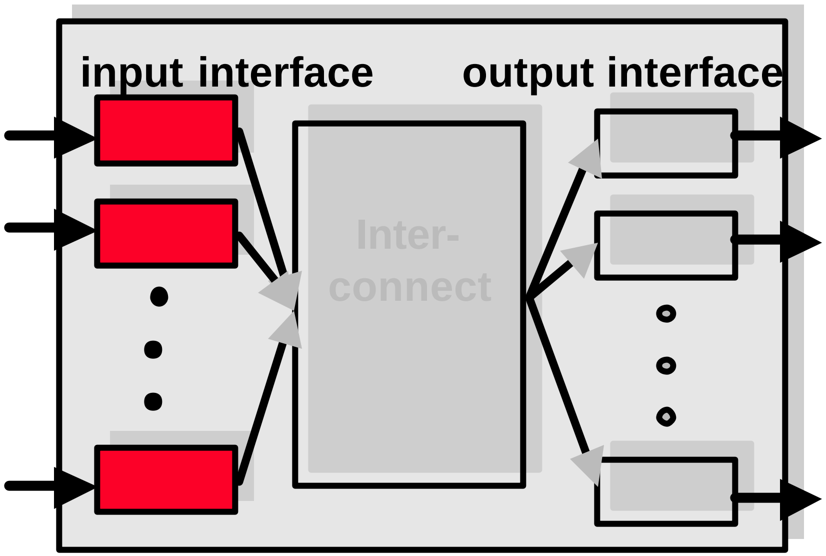
Combined Input-Output Queued (CIOQ) Routers

- Both input and output interfaces store packets
- Advantages
  - Easy to built
    - Utilization 1 can be achieved with limited input/output speedup (<= 2)
- Disadvantages
  - Harder to design algorithms
    - Two congestion points
    - Need to design flow control



Input Interface

- Packet forwarding: decide to which output interface to forward each packet based on the information in packet header
  - Examine packet header
  - Lookup in forwarding table
  - Update packet header



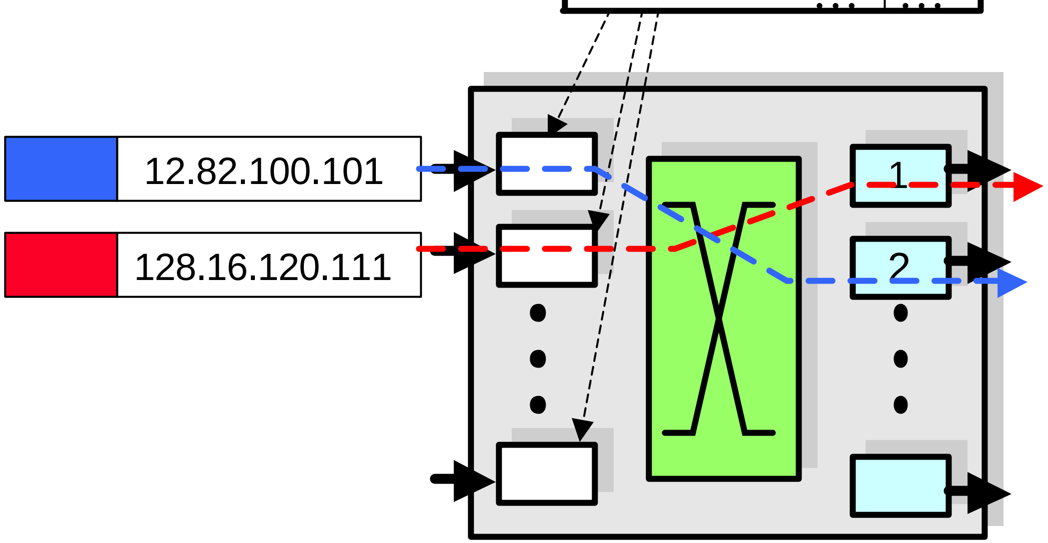
Lookup

- Identify the output interface to forward an incoming packet based on packet's destination address
- Routing tables summarize information by maintaining a mapping between IP address prefixes and output interfaces
  - How are routing tables computed?
- Route lookup → find the longest prefix in the table that matches the packet destination address

IP Routing

- Packet with destination address 12.82.100.101 is sent to interface 2, as 12.82.100.xxx is the longest prefix matching packet's destination address

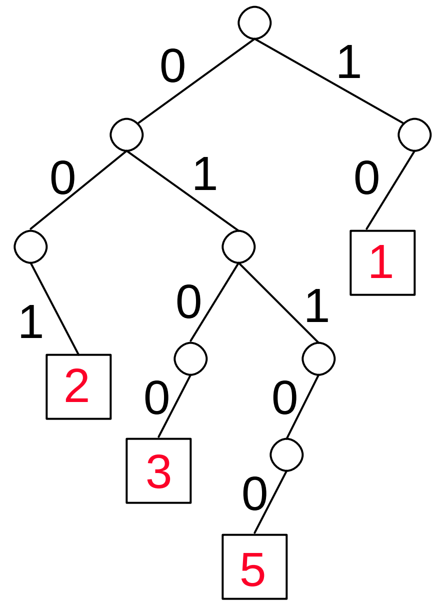
128.16.120.xx	1
12.82.xxx.xxx	3
12.82.100.xxx	2
…	…



Patricia Tries

- Use binary tree paths to encode prefixes

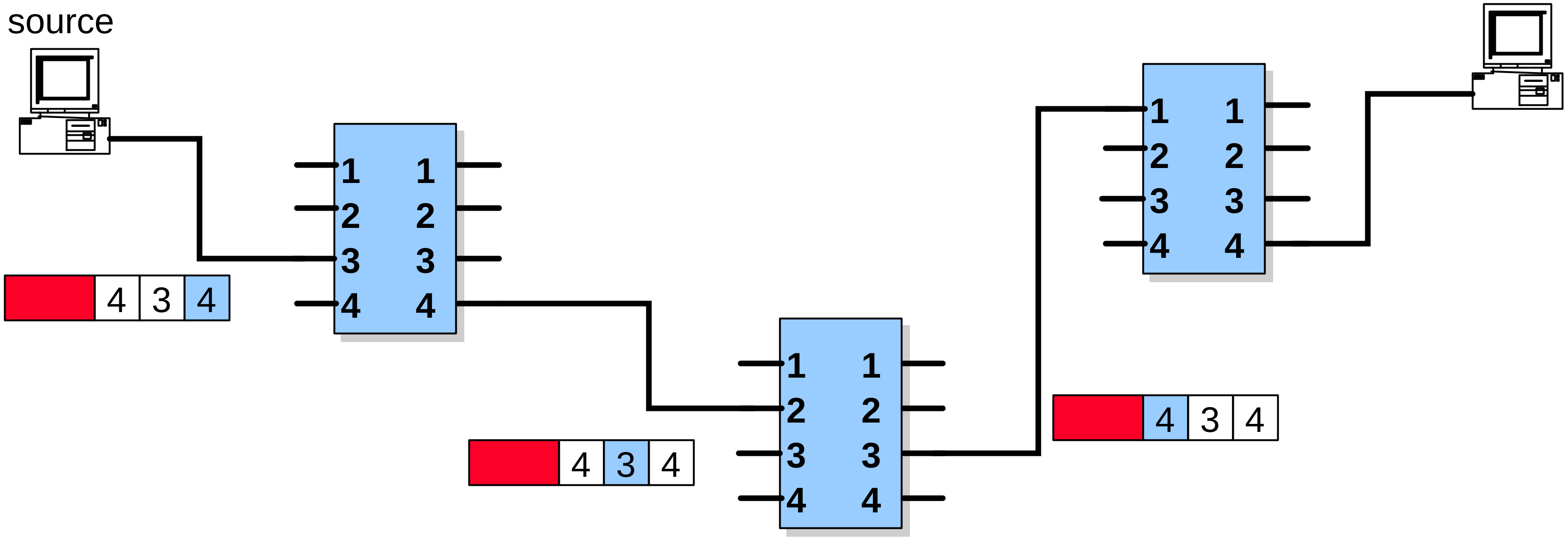
001xx	2
0100x	3
10xxx	1
01100	5



- Advantage: simple to implement
- Disadvantage: one lookup may take O(m), where m is number of bits (32 in the case of IPv4)

Another Forwarding Technique: Source Routing

- Each packet specifies the sequence of routers, or alternatively the sequence of output ports, from source to destination

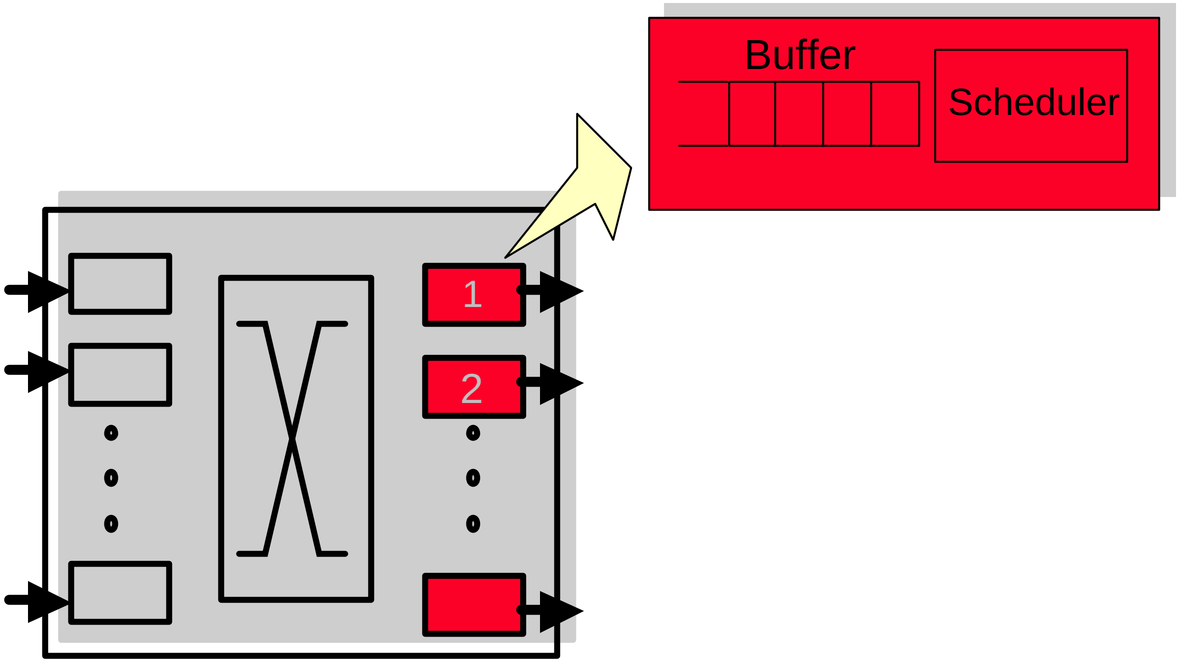


Source Routing (cont'd)

- Gives the source control of the path
- Not scalable
  - Packet overhead proportional to the number of routers
  - Typically, require variable header length which is harder to implement
- Hard for source to have complete information
- Loose source routing → sender specifies only a subset of routers along the path

Output Functions

- Buffer management: decide when and which packet to drop
- Scheduler: decide when and which packet to transmit

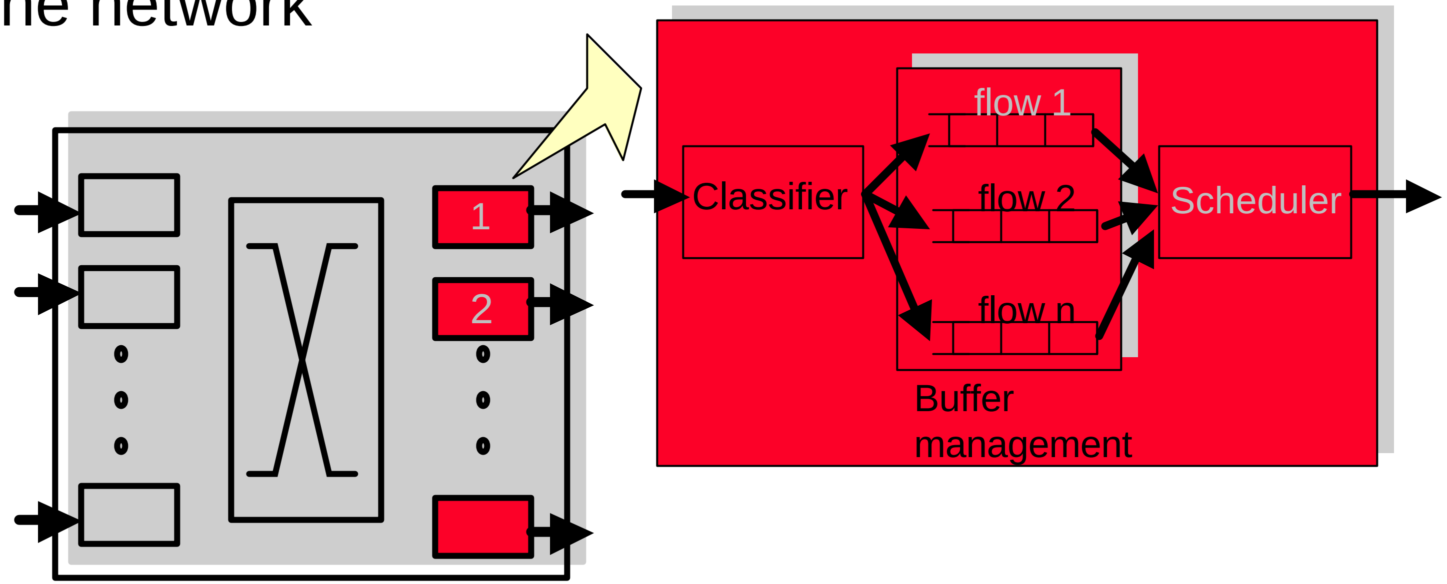


Example: FIFO router

- Most of today's routers
- Drop-tail buffer management: when buffer is full drop the incoming packet
- First-In-First-Out (FIFO) Scheduling: schedule packets in the same order they arrive

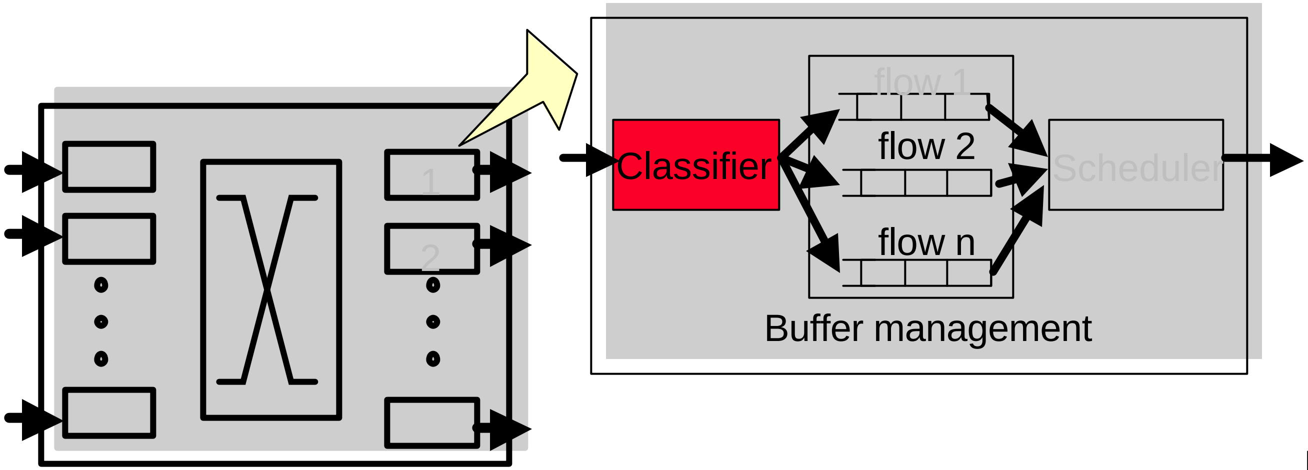
Output Functions (cont'd)

- Packet classification: map each packet to a predefined flow/connection (for datagram forwarding)
  - Use to implement more sophisticated services (e.g., QoS)
- Flow: a subset of packets between any two endpoints in the network



Packet Classification

- Classify an IP packet based on a number of fields in the packet header, e.g.,
  - source/destination IP address (32 bits)
  - source/destination port number (16 bits)
  - Type of service (TOS) byte (8 bits)
  - Type of protocol (8 bits)
- In general fields are specified by range

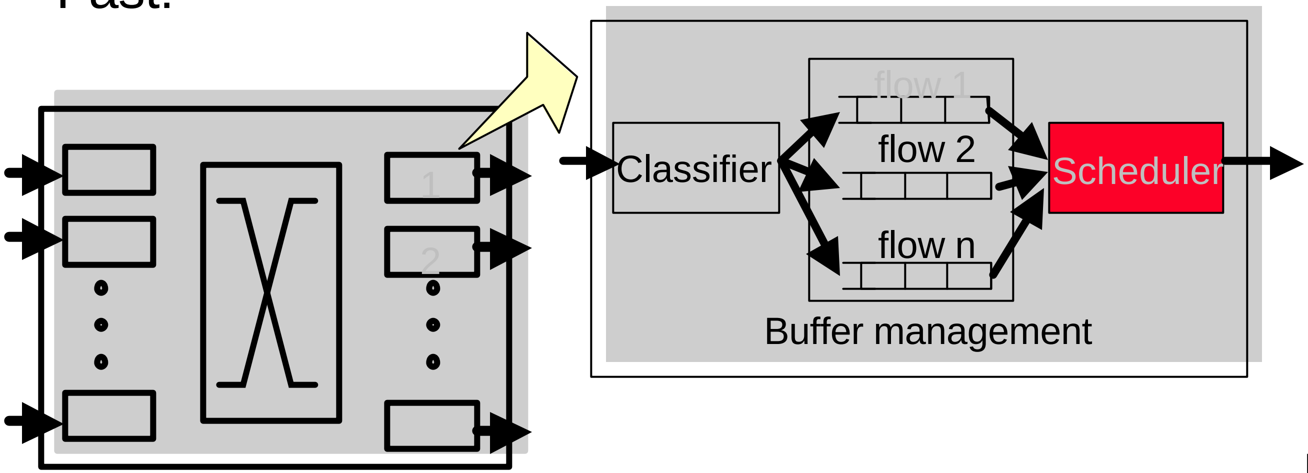


Example of Classification Rules

- Access-control in firewalls
  - Deny all e-mail traffic from ISP-X to Y
- Policy-based routing
  - Route IP telephony traffic from X to Y via ATM
- Differentiate quality of service
  - Ensure that no more than 50 Mbps are injected from ISP-X

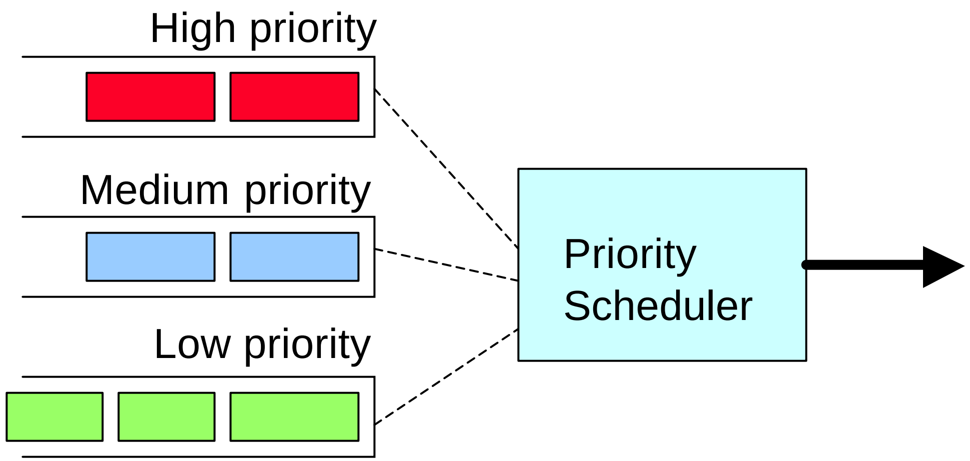
Scheduler

- One queue per flow
- Scheduler decides when and from which queue to send a packet
  - Each queue is FIFO
- Goals of a scheduler:
  - Quality of service
  - Protection (stop a flow from hogging the entire output link)
  - Fast!



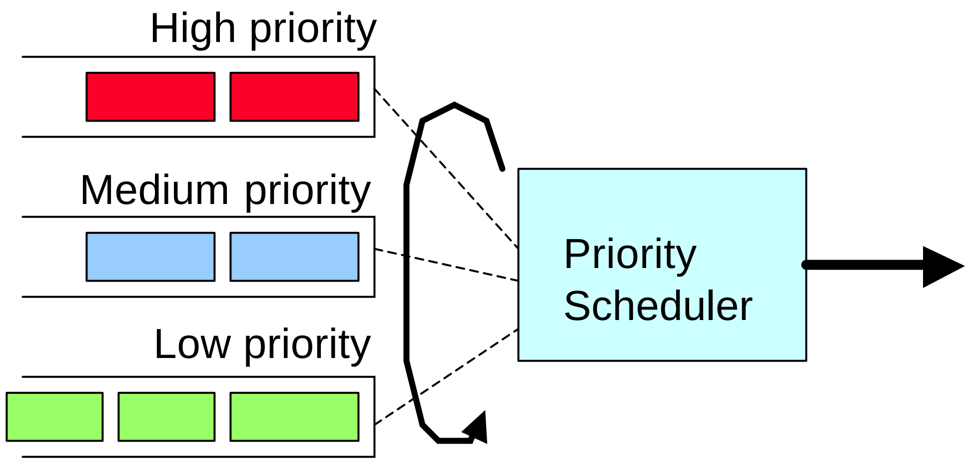
Example: Priority Scheduler

- Priority scheduler: packets in the highest priority queue are always served before the packets in lower priority queues



Example: Round Robin Scheduler

- Round robin: packets are served in a round-robin fashion

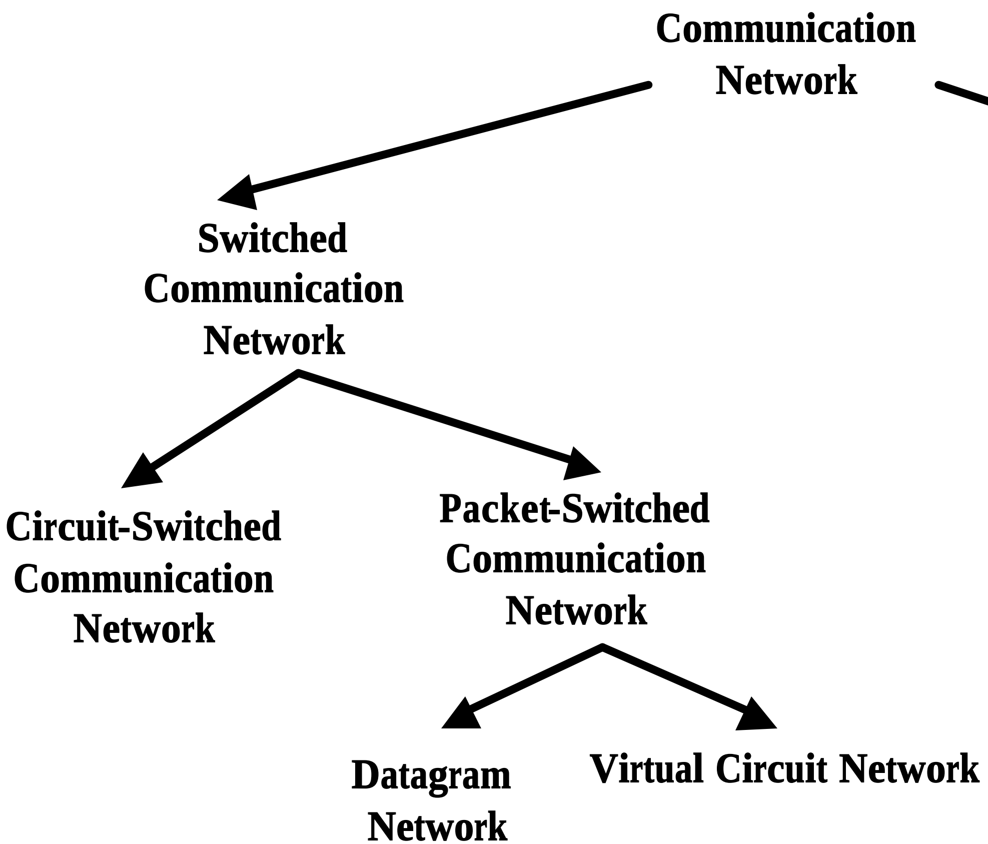


Discussion

- Priority scheduler vs. Round-robin scheduler
  - What are advantages disadvantages of each scheduler?

Big Picture

- Where do IP routers belong?



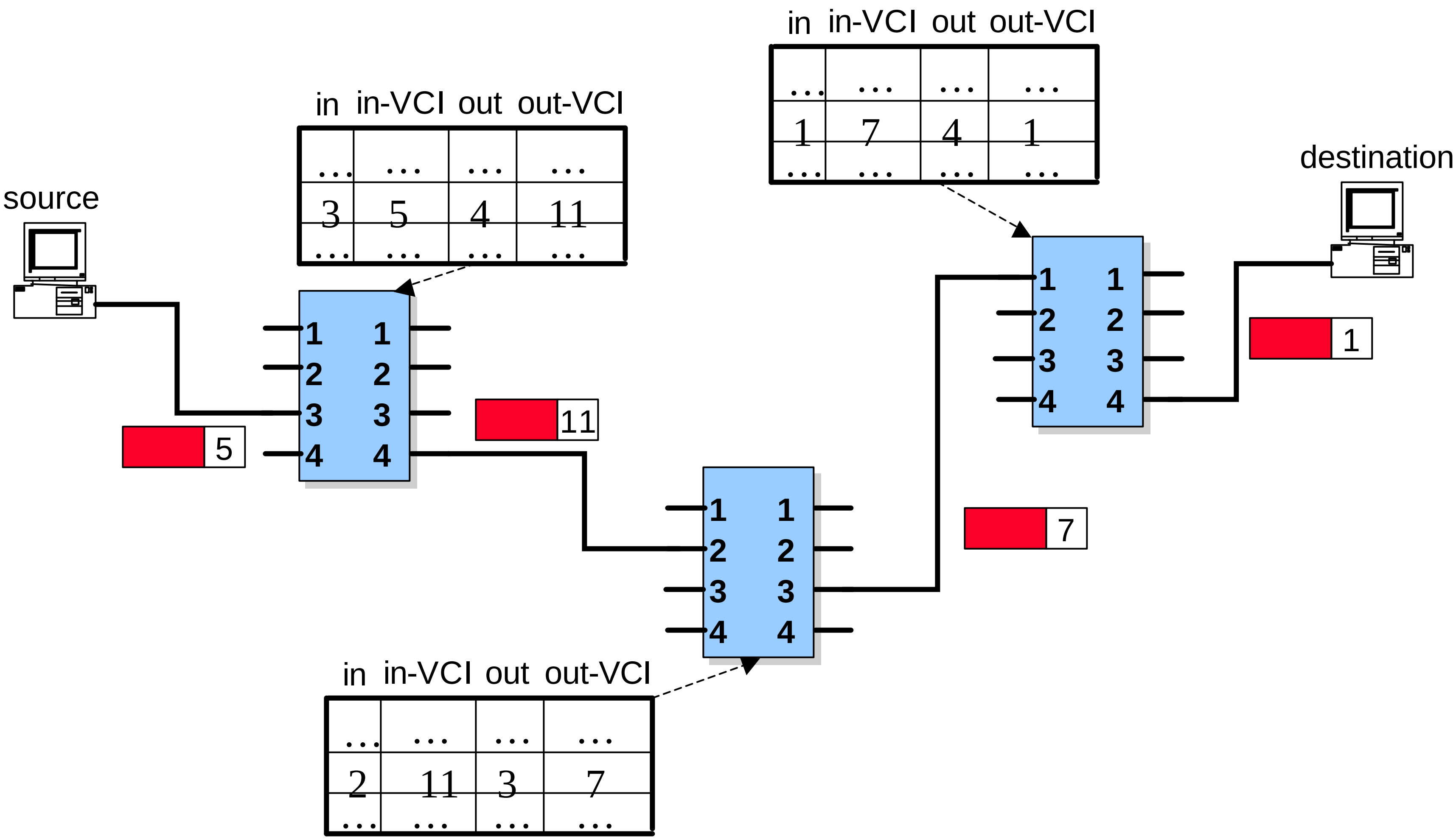
Packet (Datagram) Switching Properties

- Expensive forwarding
  - Forwarding table size depends on number of different destinations
  - Must lookup in forwarding table for every packet
- Robust
  - Link and router failure may be transparent for end-hosts
- High bandwidth utilization
  - Statistical multiplexing
- No service guarantees
  - Network allows hosts to send more packets than available bandwidth $\rightarrow$ congestion $\rightarrow$ dropped packets

Virtual Circuit (VC) Switching

- Packets not switched independently
  - Establish virtual circuit before sending data
- Forwarding table entry
  - (input port, input VCI, output port, output VCI)
  - VCI – Virtual Circuit Identifier
- Each packet carries a VCI in its header
- Upon a packet arrival at interface i
  - Input port uses i and the packet's VCI v to find the routing entry (i, v, i', v')
  - Replaces v with v' in the packet header
  - Forwards packet to output port i'

VC Forwarding: Example



VC Forwarding (cont'd)

- A signaling protocol is required to set up the state for each VC in the routing table
  - A source needs to wait for one RTT (round trip time) before sending the first data packet
- Can provide per-VC QoS
  - When we set the VC, we can also reserve bandwidth and buffer resources along the path

VC Switching Properties

- Less expensive forwarding
  - Forwarding table size depends on number of different circuits
  - Must lookup in forwarding table for every packet
- Much higher delay for short flows
  - 1 RTT delay for connection setup
- Less Robust
  - End host must spend 1 RTT to establish new connection after link and router failure
- Flexible service guarantees
  - Either statistical multiplexing or resource reservations

Circuit Switching

- Packets not switched independently
  - Establish circuit before sending data
- Circuit is a dedicated path from source to destination
  - E.g., old style telephone switchboard, where establishing circuit means connecting wires in all the switches along path
  - E.g., modern dense wave division multiplexing (DWDM) form of optical networking, where establishing circuit means reserving an optical wavelength in all switches along path
- No forwarding table

Circuit Switching Properties

- Cheap forwarding
  - No table lookup
- Much higher delay for short flows
  - 1 RTT delay for connection setup
- Less robust
  - End host must spend 1 RTT to establish new connection after link and router failure
- Must use resource reservations

Forwarding Comparison

	pure packet switching	virtual circuit switching	circuit switching
forwarding cost	high	low	none
bandwidth utilization	high	flexible	low
resource reservations	none	flexible	yes
robustness	high	low	low

Summary

- Routers
  - Key building blocks of today a network in general, and Internet in particular
- Main functionalities implemented by a router
  - Packet forwarding
  - Buffer management
  - Packet scheduling
  - Packet classification
- Forwarding techniques
  - Datagram (packet) switching
  - Virtual circuit switching
  - Circuit switching