Computer Networks

Jiaqi Zheng

Material with thanks Mosharaf Chowdhury, and many other colleagues.

Agenda

- IP routers

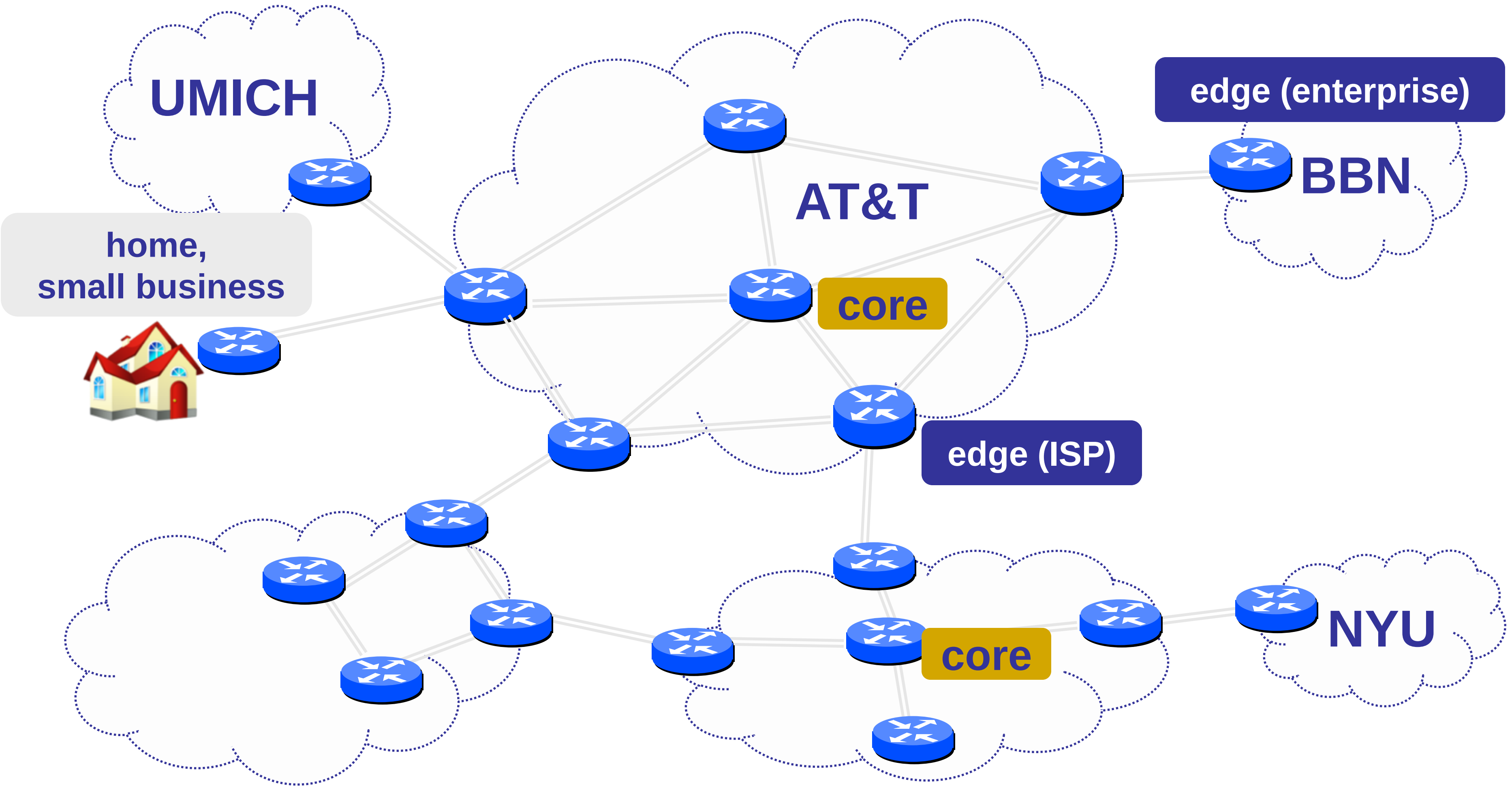
IP routers

- Core building block of the Internet infrastructure
- $120B+ industry
- Vendors: Cisco, Huawei, Juniper, Alcatel-Lucent (account for >90%)

Router definitions

- Router capacity = N x R
- N = Number of external router “ports”
- R = Speed (“line rate”) of a port

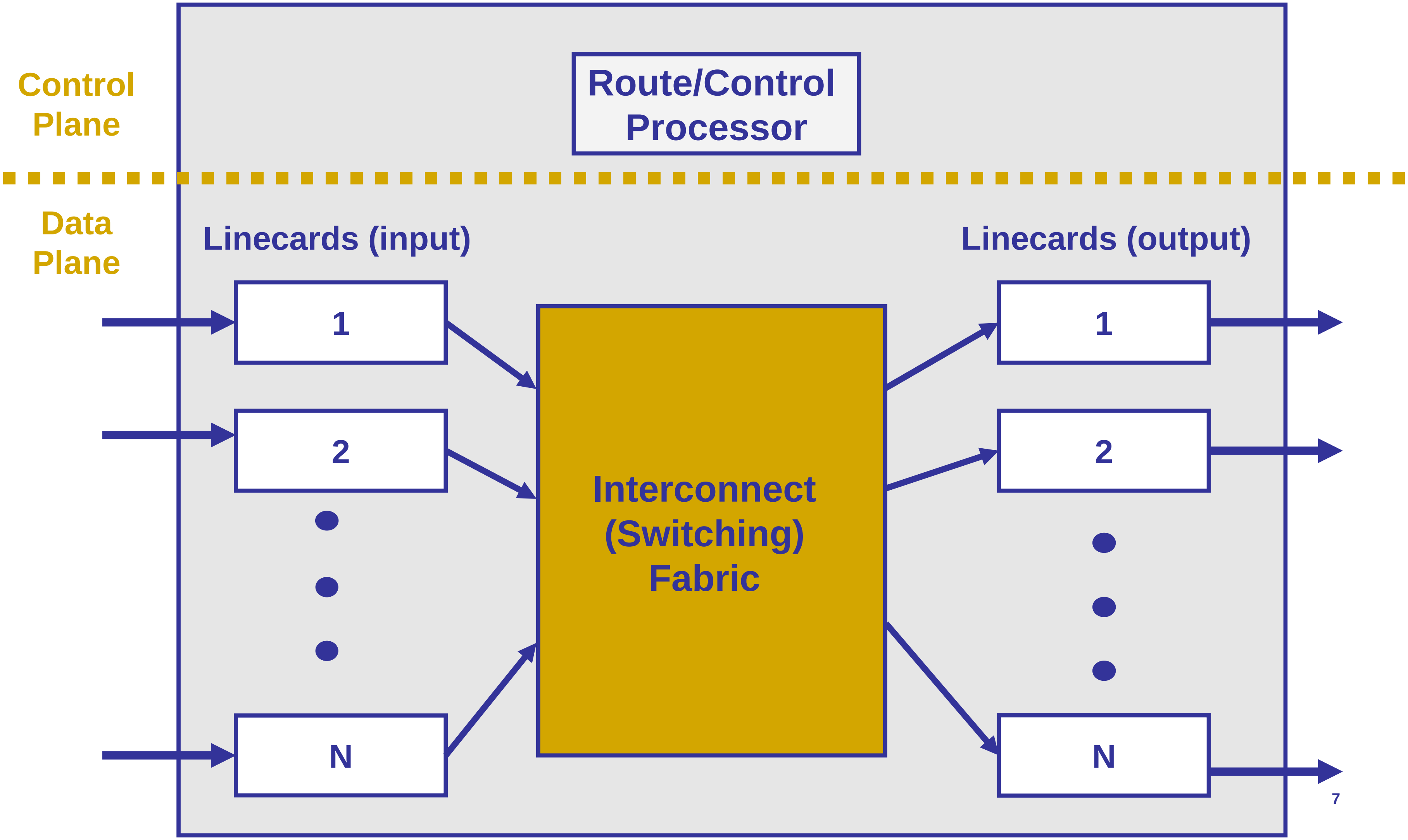
Networks and routers



Many types of routers

- Core
  - R = 10/40/100 Gbps
  - NR = O(100) Tbps (Aggregated)
- Edge
  - R = 1/10/40
  - NR = O(100) Gbps
- Small business
  - R = 10/100/1000 Mbps
  - NR < 10 Gbps

What's inside a router?



What's inside a router?

- Linecards
  - Input linecards process packets on their way in
  - Output linecards process packets on way out
  - Input and output for the same port are on the same physical linecard
- Interconnect/switching fabric
  - Transfers packets from input to output ports

Input linecards

- Tasks
  - Receive incoming packets (physical layer stuff)
  - Update the IP header
    - »TTL, Checksum, Options and Fragment (maybe)
  - Lookup the output port for the destination IP address
  - Queue the packet at the switch fabric
- Challenge: speed!
  - 100B packets @ 40Gbps → new packet every 20 nano secs!
  - Typically implemented with specialized ASICs (network processors)

Looking up the output port

- One entry for each address $\rightarrow$ 4 billion entries!
- For scalability, addresses are aggregated

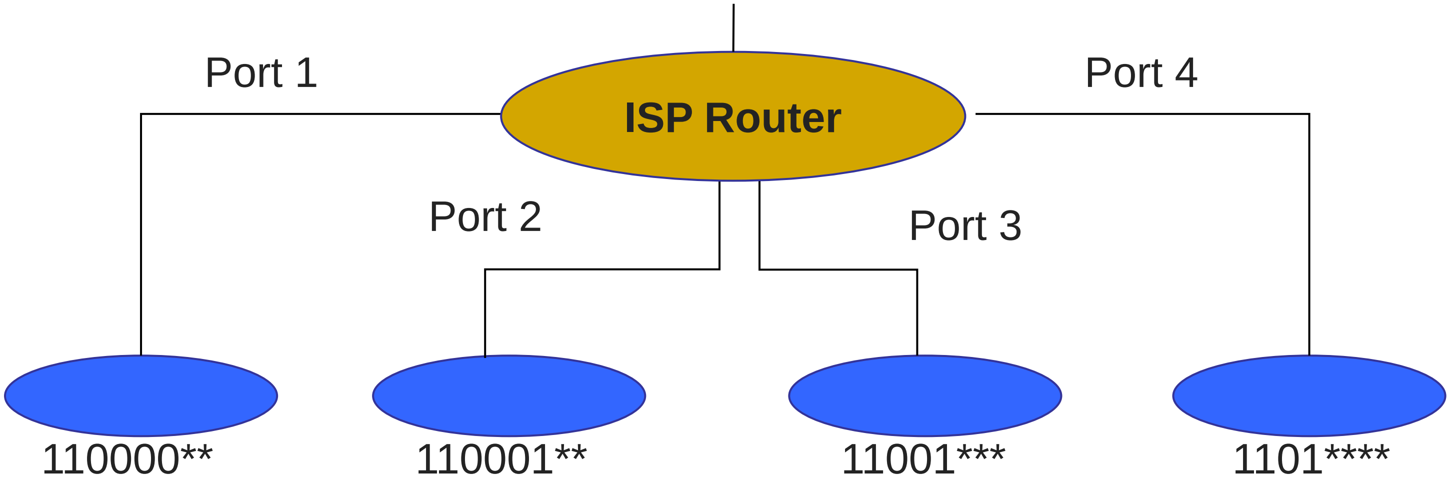
Example

- Router with 4 ports
- Destination address range mapping
  - 11 00 00 00 to 11 00 00 11: Port 1
  - 11 00 01 00 to 11 00 01 11: Port 2
  - 11 00 10 00 to 11 00 11 11: Port 3
  - 11 01 00 00 to 11 01 11 11: Port 4

Example

- Router with 4 ports
- Destination address range mapping
  - 11 00 00 00 to 11 00 00 11: Port 1
  - 11 00 01 00 to 11 00 01 11: Port 2
  - 11 00 10 00 to 11 00 11 11: Port 3
  - 11 01 00 00 to 11 01 11 11: Port 4

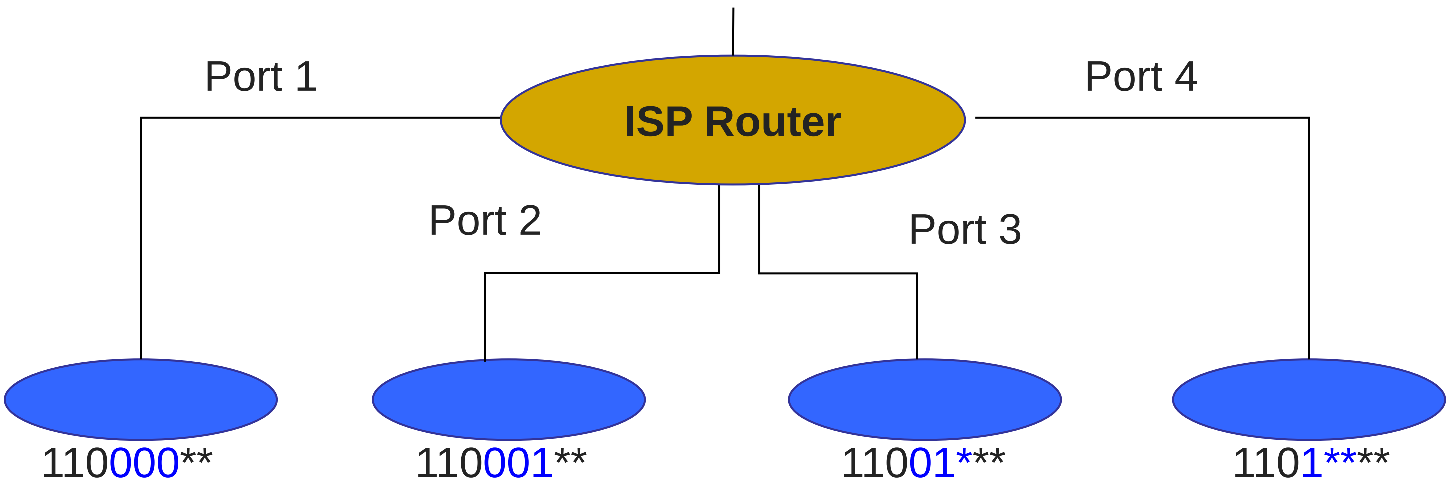
Longest prefix matching



Finding match efficiently

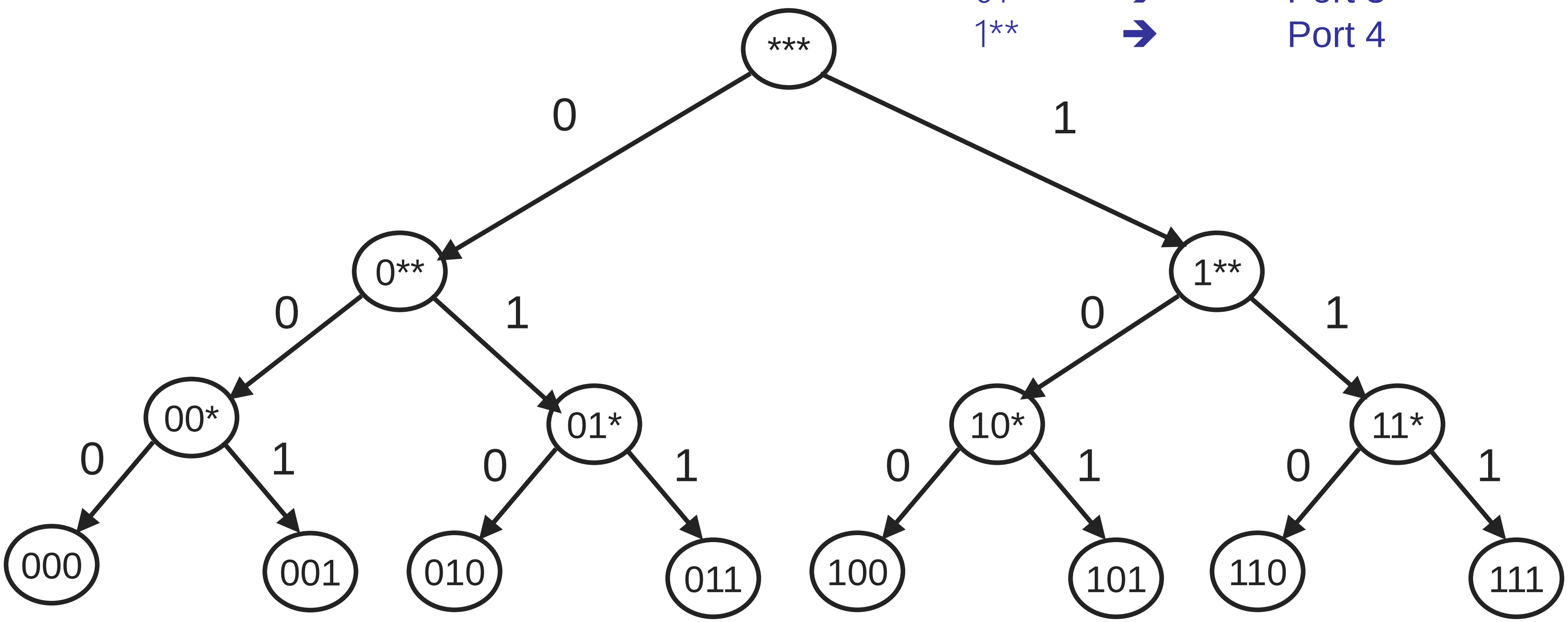
- Testing each entry to find a match scales poorly
  - On average: O(number of entries)
- Leverage tree structure of binary strings
  - Set up tree-like data structure

Longest prefix matching



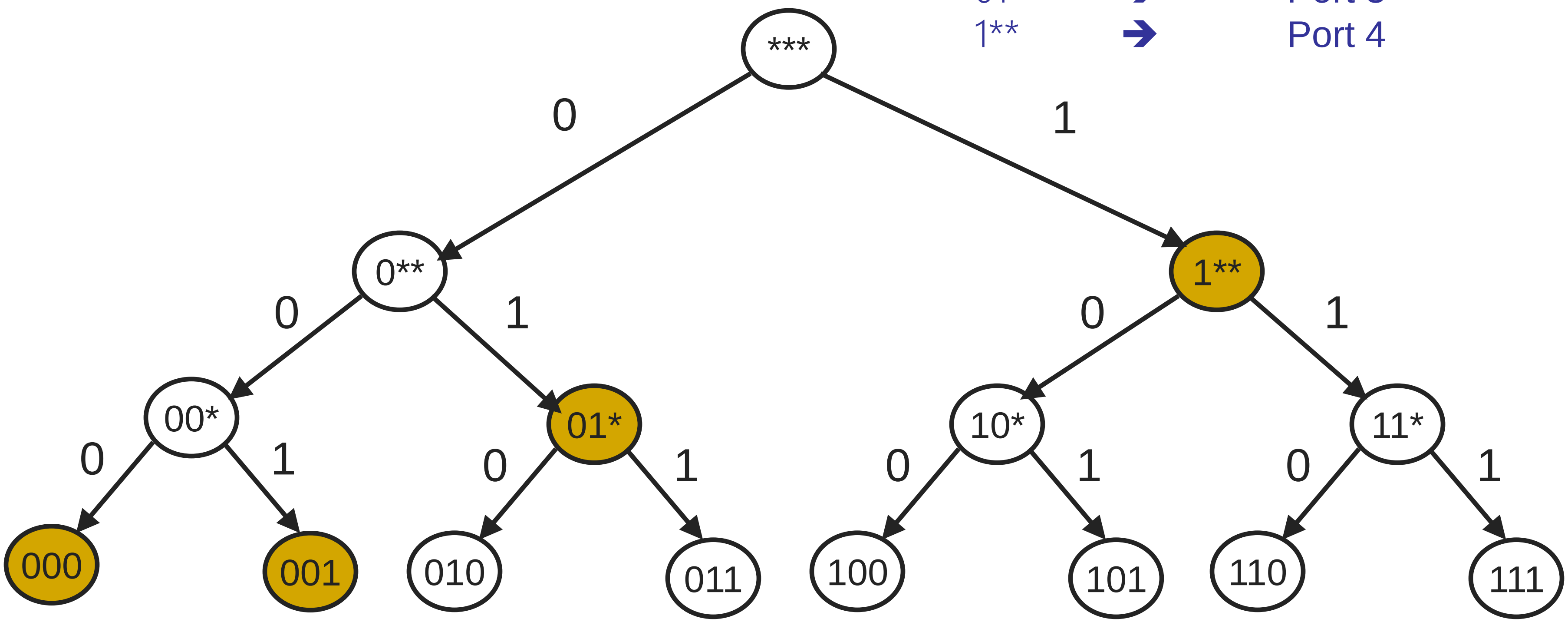
Tree structure

000	➔	Port 1
001	➔	Port 2
01*	➔	Port 3
1**	➔	Port 4



Tree structure

000	➔	Port 1
001	➔	Port 2
01*	➔	Port 3
1**	➔	Port 4



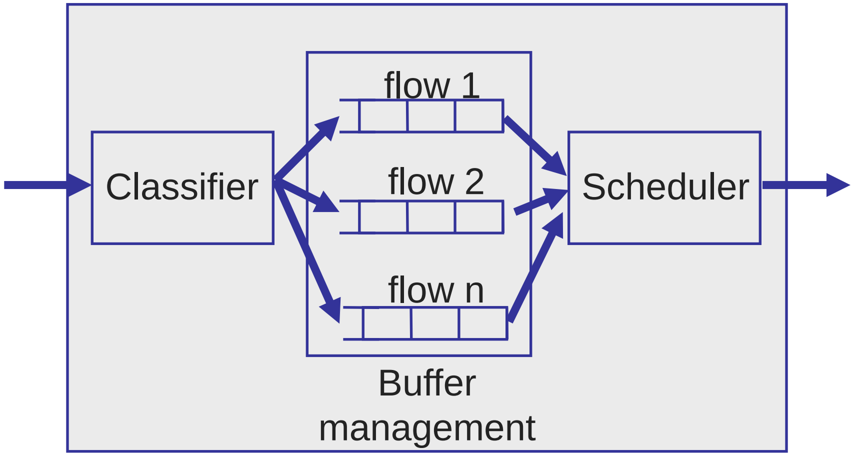
Record port associated with latest match, and only override when it matches another prefix during walk down tree

Input linecards

- Main challenge is processing speeds
- Tasks involved:
  - Update packet header (easy)
  - LPM lookup on destination address (harder)
- Mostly implemented with specialized hardware

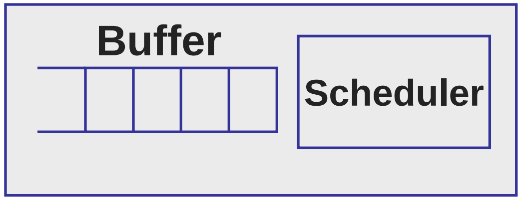
Output linecards

- Packet classification: map packets to flows
- Buffer management: decide when and which packet to drop
- Scheduler: decide when and which packet to transmit



Simplest: FIFO router

- No classification
- 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



Packet classification

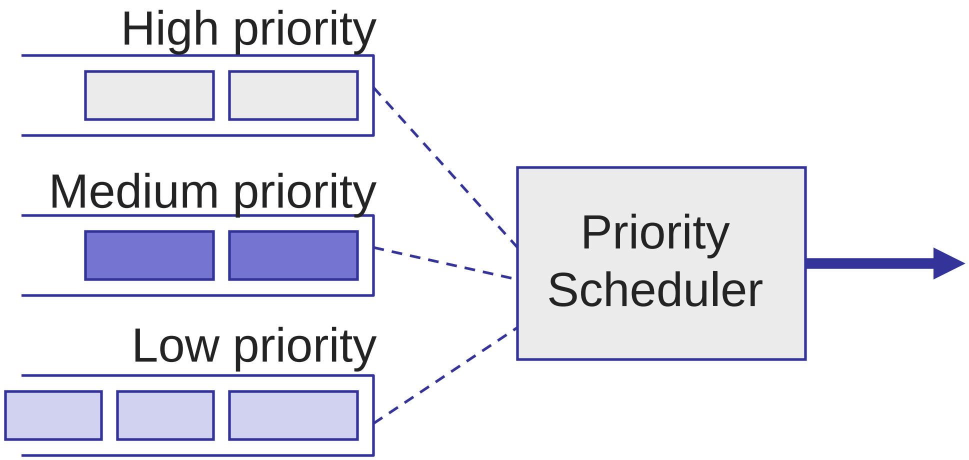
- Classify an IP packet based on a number of fields in the packet header, e.g.,
  - Source/destination IP address (32 bits)
  - Source/destination TCP port number (16 bits)
  - Type of service (TOS) byte (8 bits)
  - Type of protocol (8 bits)
- In general fields are specified by range
  - Classification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

Scheduler

- One queue per “flow”
- Scheduler decides when and from which queue to send a packet
- Goals of a scheduling algorithm
  - Fast!
  - Depends on the policy being implemented (fairness, priority, etc.)

Priority scheduler

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



Round-robin scheduler

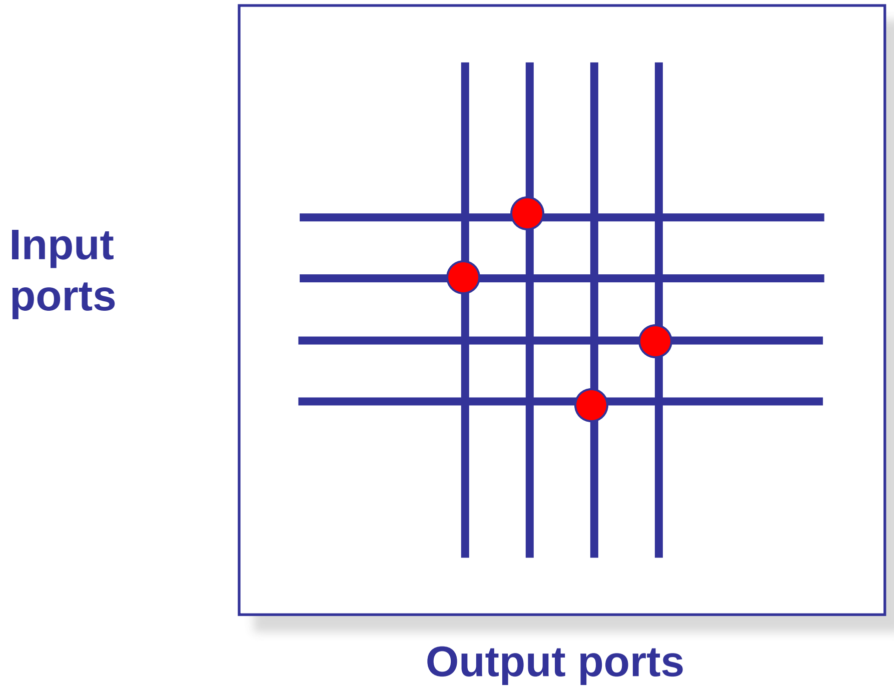
- Round robin: packets are served from each queue in turn
- Fair queuing (FQ): round-robin for packets of different size
- Weighted fair queueing (WFQ): serve proportional to weight
  - FQ gives equal weight to each flow

Connecting inputs to outputs: Switching fabric

- Mini-network
- Three primary ways to switch
  - Switching via shared memory
  - Switching via a bus
  - Switching via an inter-connection network
    - »For example, cross-bar

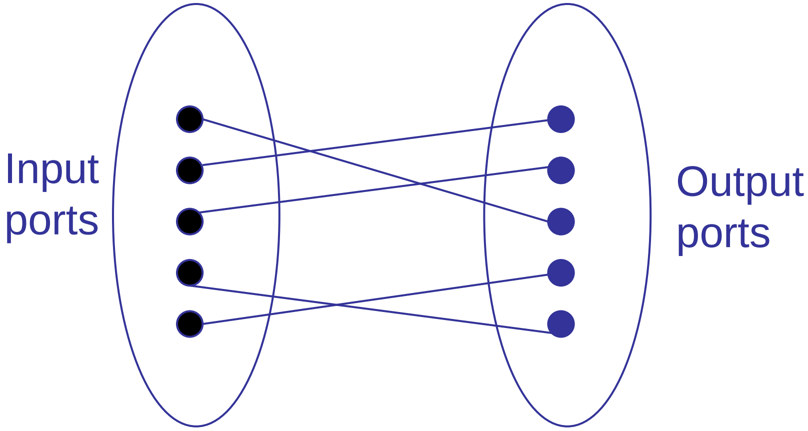
Context

- Crossbar fabric
- Centralized scheduler



Scheduling

- Run links at full capacity, fairness across inputs
- Scheduling formulated as finding a matching on a bipartite graph



- Practical solutions look for a good maximal matching (fast)

Summary

- IP routers form the backbone of the Internet
- Aims for speed while providing fairness