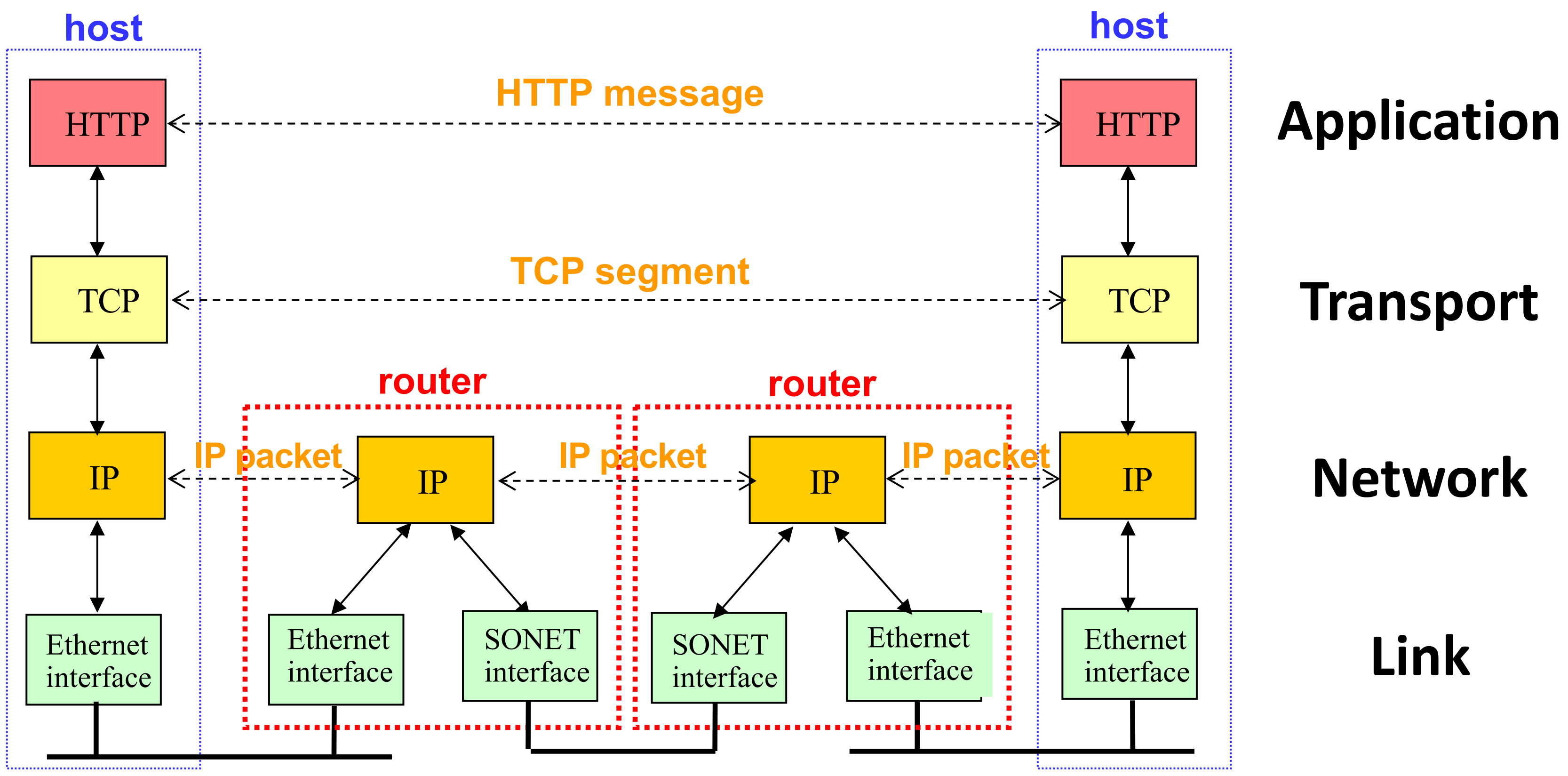
COS 461 *Computer Networks*

Lecture 4: Hubs, Switches, and Routers

Kyle Jamieson

Today: Hubs, Switches, and Routers, Oh My!

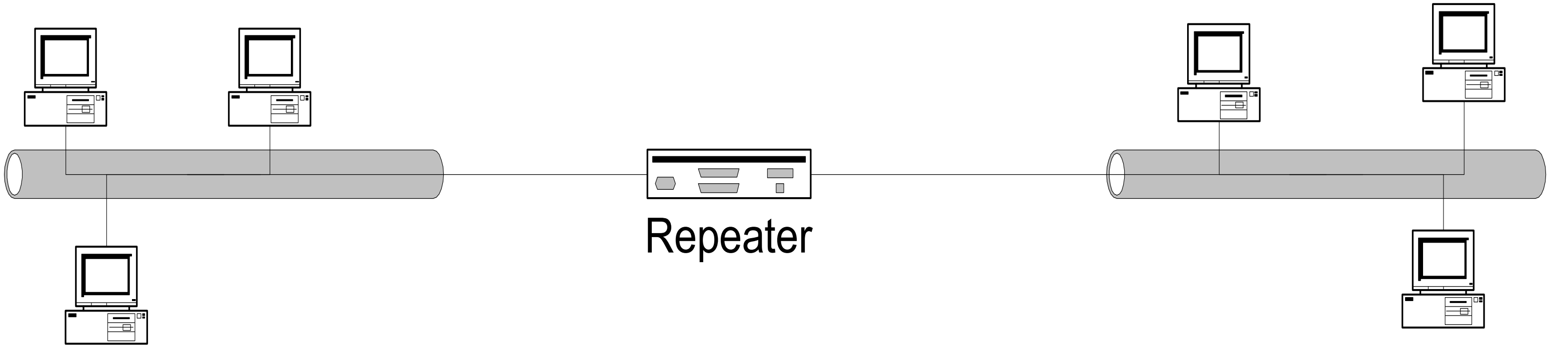


Terminology

- Hubs and Repeaters
  - Physical Layer devices; connect machines on same LAN
  - Broadcast: All frames are sent out all physical ports
- Switches and Bridges
  - Link Layer devices, connect machines on same LAN
  - Only send frames to selected physical port based on destination MAC address
- Routers
  - Connect between LANs at Network Layer, i.e., wide area
  - Only send packet to selected physical port based on destination IP address

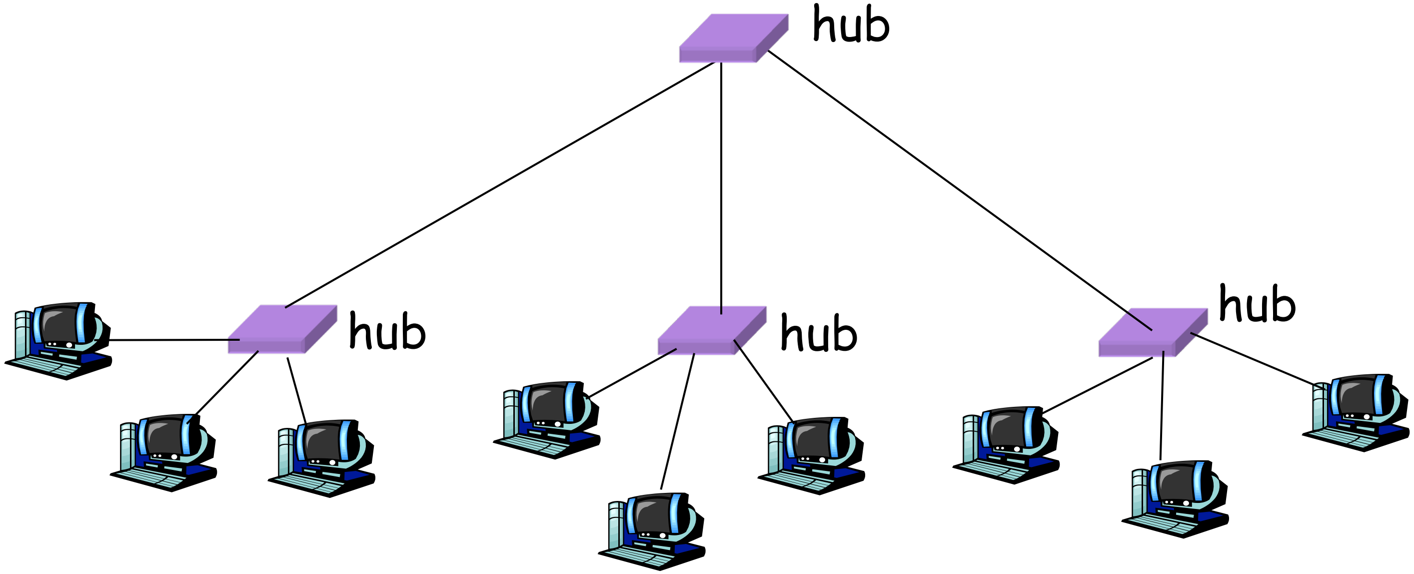
Physical Layer: Repeaters

- Distance limitation in local-area networks
  - Electrical signal becomes weaker as it travels
  - Imposes a limit on the length of a LAN
- Repeaters join LANs together
  - Analog electronic device
  - Continuously monitors electrical signals
  - Transmits an amplified copy



Physical Layer: Hubs

- Joins multiple input lines electrically
  - Designed to hold multiple line cards
  - Do not necessarily amplify the signal
- Very similar to repeaters
  - Also operates at the physical layer

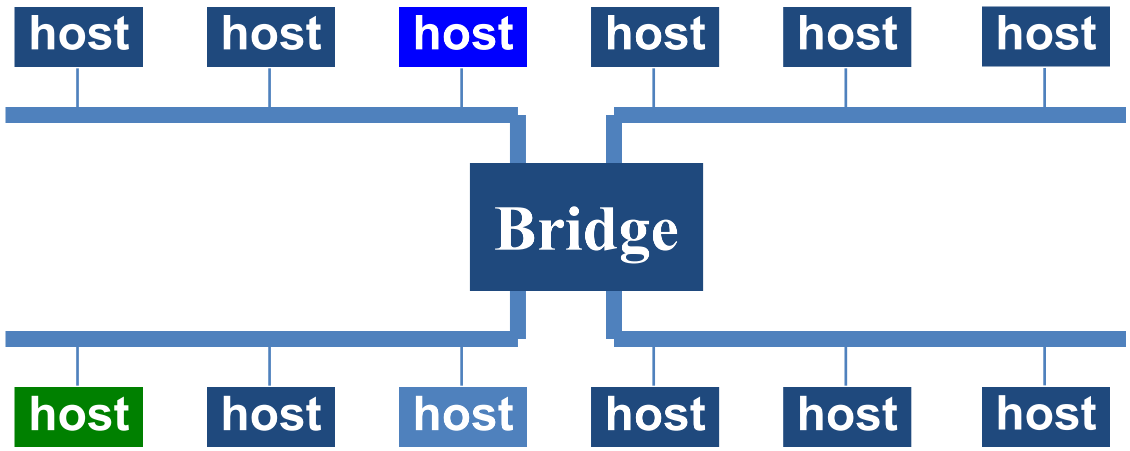


Limitations of Repeaters and Hubs

- One large shared link
  - Each bit is sent everywhere
  - So, aggregate throughput is limited
- Cannot support multiple LAN technologies
  - Does not buffer or interpret frames
  - Can't interconnect between different rates/formats
- Limitations on maximum nodes and distances
  - Shared medium imposes length limits
  - E.g., cannot go beyond 2500 meters on Ethernet

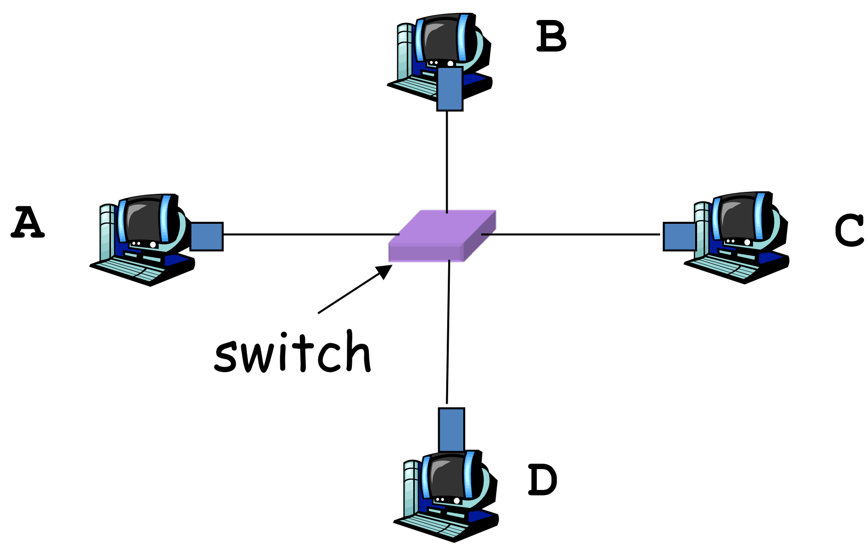
Link Layer: Bridges

- Connects two or more LANs at the link layer
  - Extracts destination address from the frame
  - Looks up the destination in a table
  - Forwards the frame to the appropriate segment
- Each segment can carry its own traffic



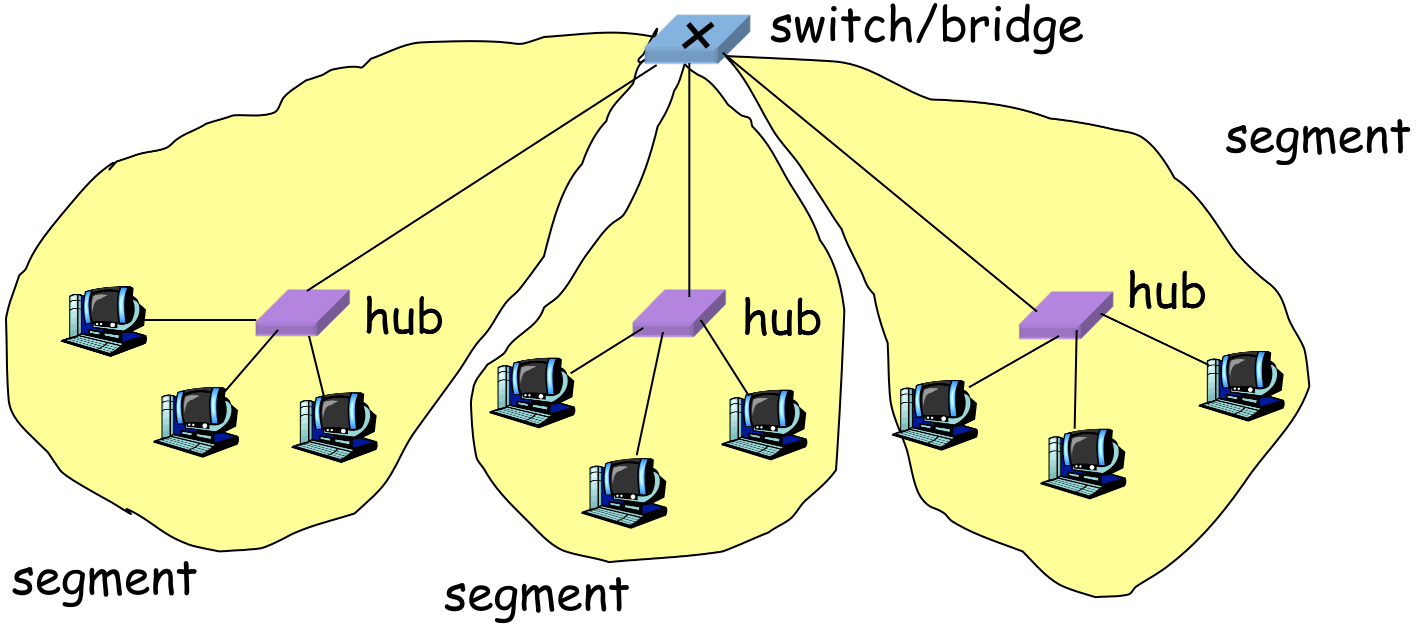
Link Layer: Switches

- Typically connects individual computers
  - A switch is essentially the same as a bridge
  - … though typically used to connect hosts
- Supports concurrent communication
  - Host A can talk to C, while B talks to D



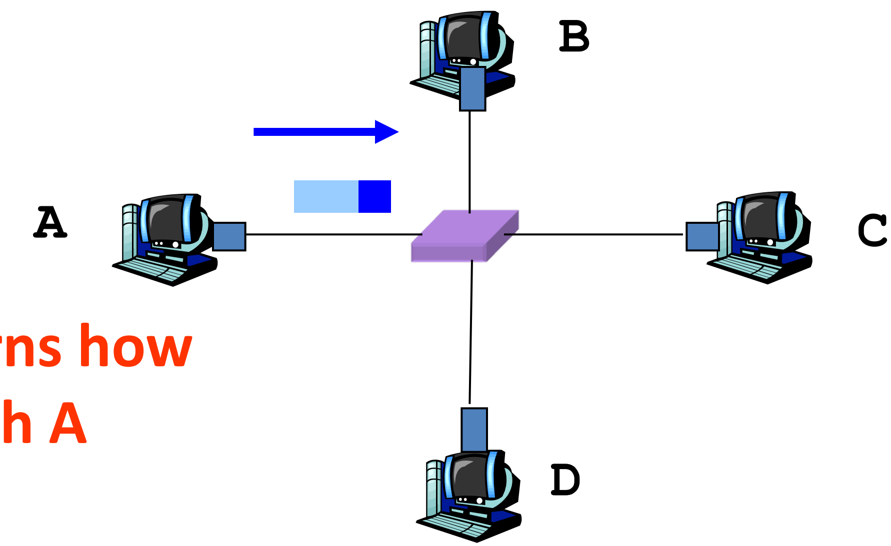
Bridges/Switches: Traffic Isolation

- Switch filters packets
  - Frame only forwarded to the necessary segments
  - Segments can support separate transmissions



Self Learning: Building the Table

- When a frame arrives
  - Inspect the *source* MAC address
  - Associate the address with the *incoming* interface
  - Store the mapping in the switch table
  - Use a timer to eventually forget the mapping

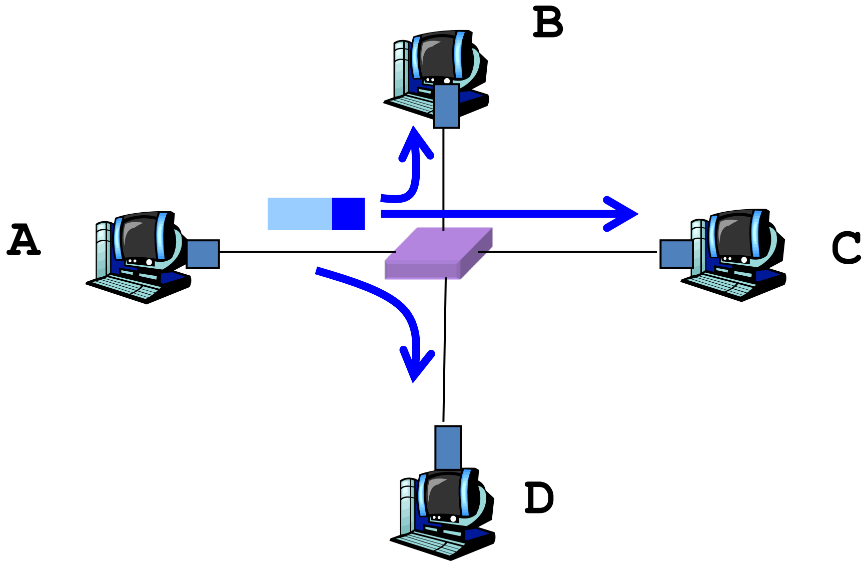


Switch learns how to reach A

Self Learning: Handling Misses

- When frame arrives with unfamiliar destination
  - Forward the frame out all of the interfaces
  - … except for the one where the frame arrived
  - Hopefully, this case won't happen very often!

When in doubt, shout!



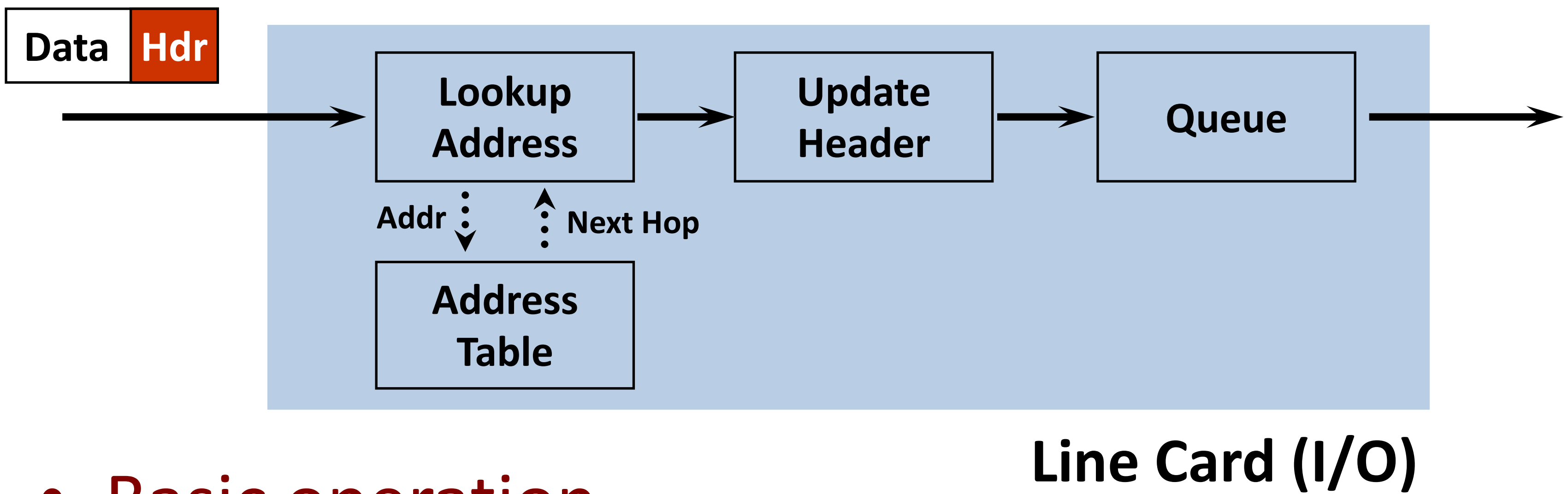
Routers: Looking closer…

Basic Router Architecture

- Each switch/router has a forwarding table
  - Maps destination address to outgoing interface

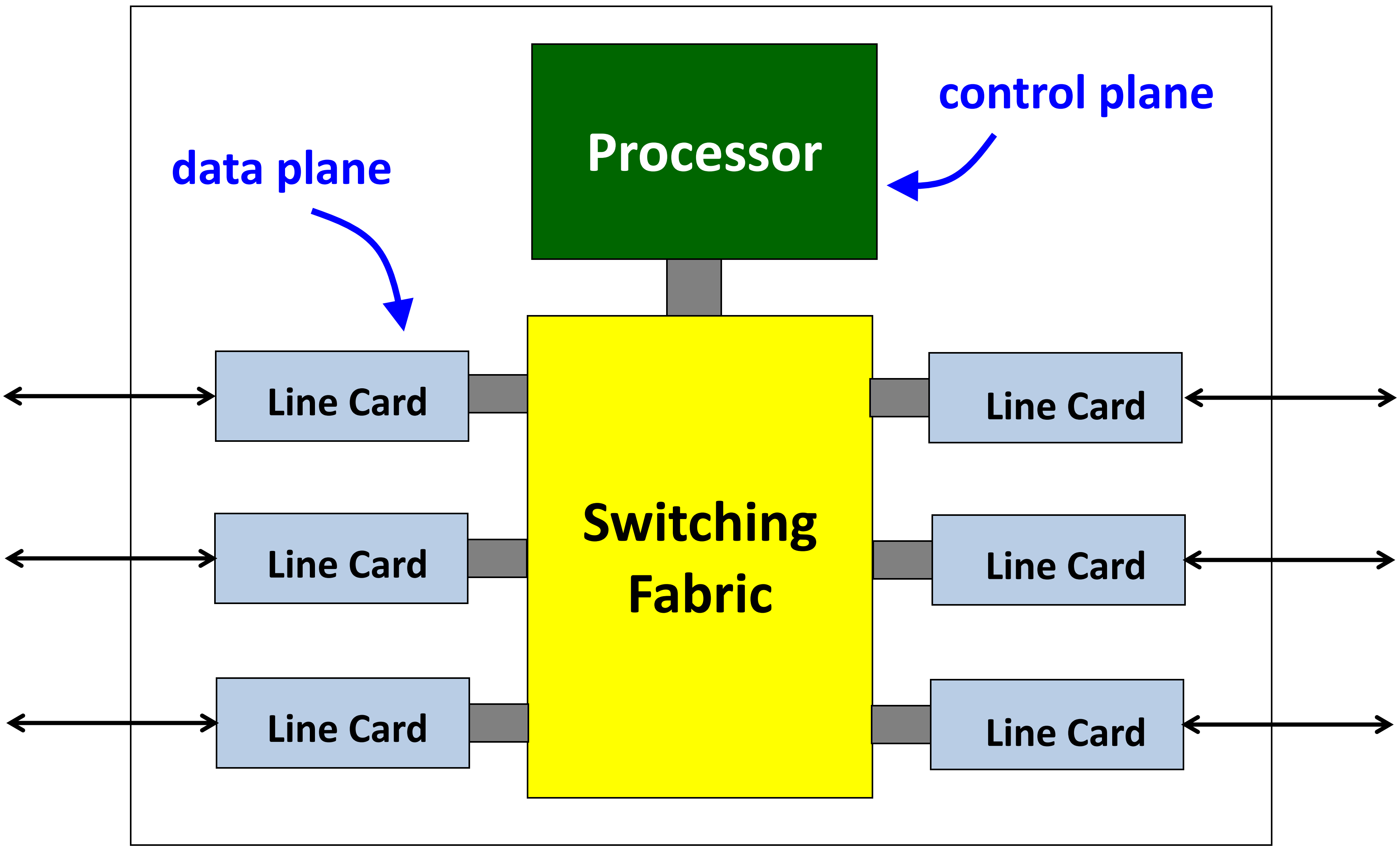
- Basic operation:
  1. Receive packet
  2. Look at header to determine destination address
  3. Look in forwarding table to determine output interface
  4. Modify packet header (e.g., decr TTL, update chksum)
  5. Send packet to output interface

Basic Router Architecture



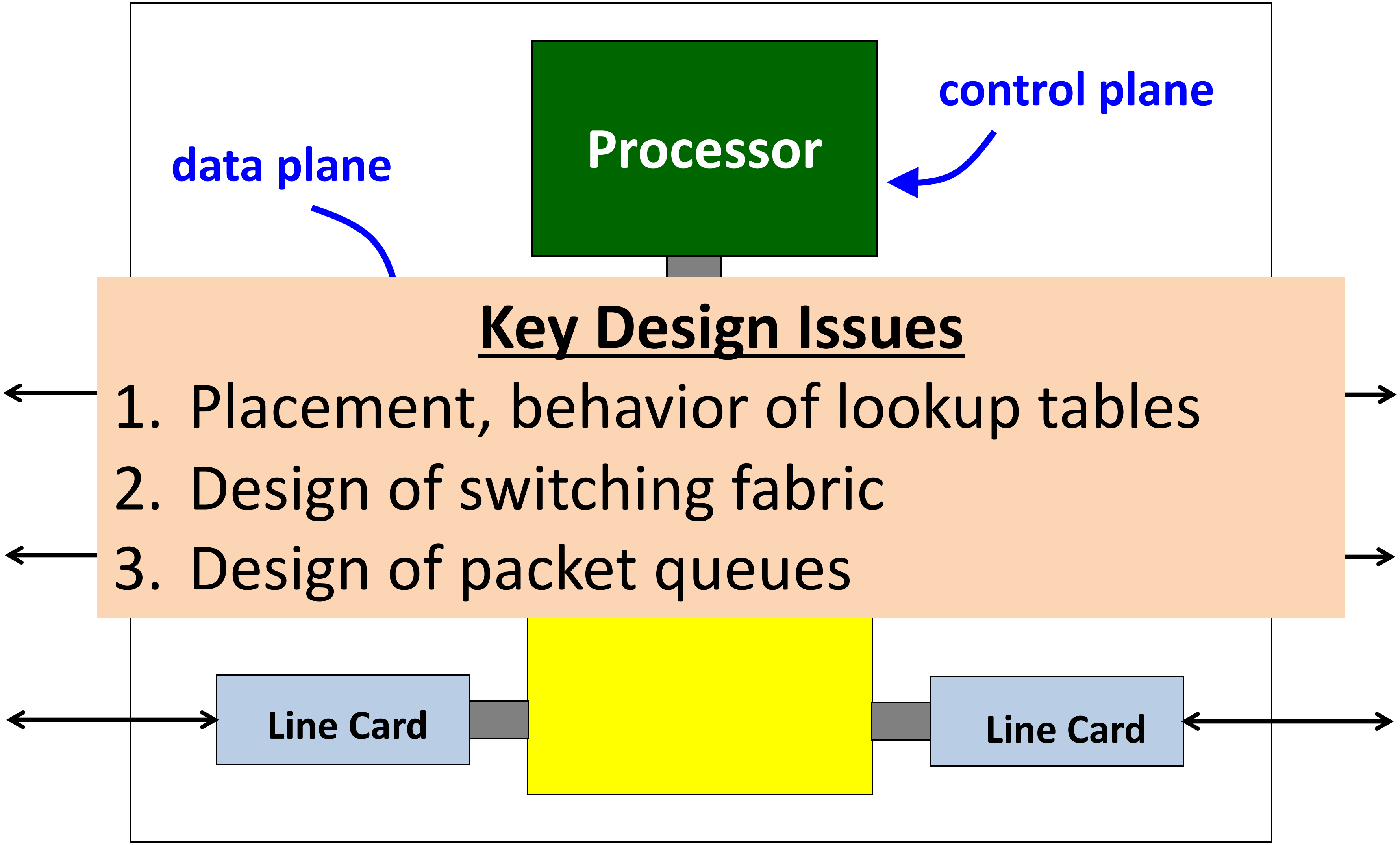
- Basic operation
  1. Receive packet
  2. Look at header to determine destination address
  3. Look in forwarding table to determine output interface
  4. Modify packet header (e.g., decr TTL, update chksum)
  5. Send packet to output interface

Router



Processor
control plane
data plane
Switching Fabric
Line Card
Line Card
Line Card
Line Card
Line Card
Line Card

Router



Key Design Issues

1. Placement, behavior of lookup tables
2. Design of switching fabric
3. Design of packet queues

Lookup algorithm: L2 vs L3

Protocol	Mechanism	Techniques
Ethernet (48 bits) MPLS ATM	Exact Match	• Direct lookup • Associative lookup • Hashing • Binary tree
IPv4 (32 bits) **IPv6 (128 bits)**	**Longest-Prefix Match**	• **Radix trie** • **Compressed trie** • **TCAM**

Longest Prefix Match (LPM)

- Each packet has destination IP address
- Router finds <u>longest</u> table prefix that matches address

destIP = 68.211.6.120 →

	Prefix	Output
✓Match	68.208.0.0/12	1
✓Match	68.211.0.0/17	1
	68.211.128.0/19	2
	68.211.160.0/19	2
	68.211.192.0/18	1

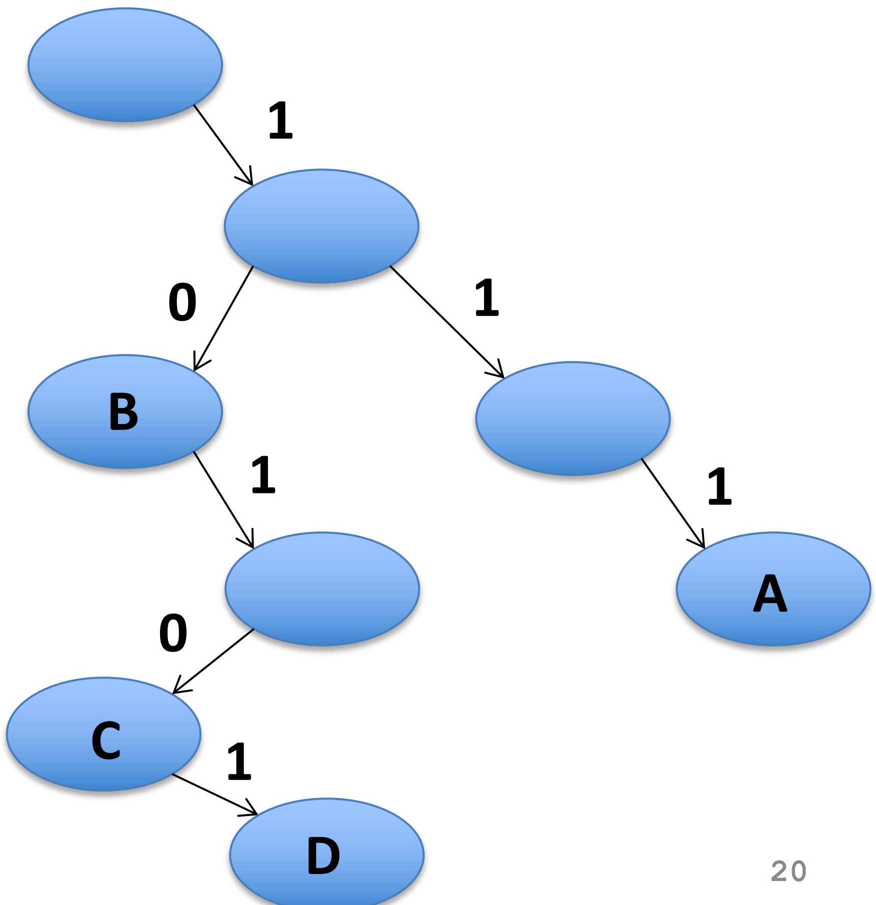
LPM: Benefits

- Benefits of CIDR allocation and LPM
  - **Efficiency:** Prefixes can be allocated at much finer granularity
  - **Hierarchical aggregation:** Upstream ISP can aggregate 2 contiguous prefixes from downstream ISPs to shorter prefix

Software LPM lookup using trie

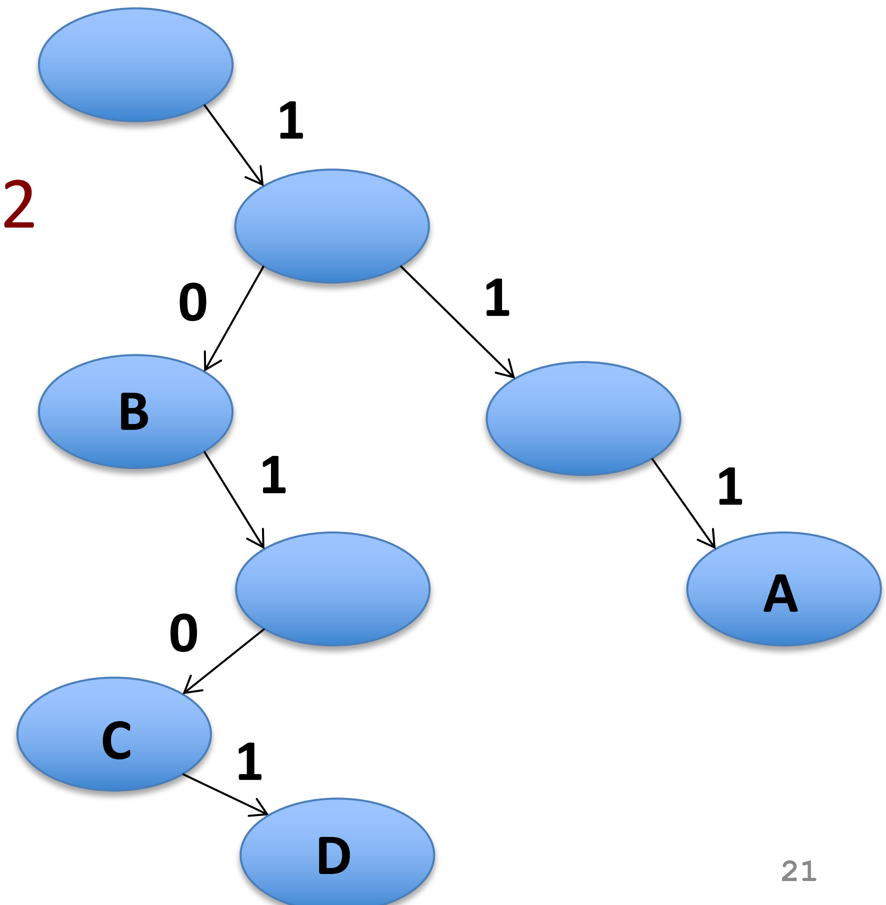
- Prefixes "spelled out" by following path from root
- To find the best prefix, spell out packet address in trie

	Prefixes
A	111*
B	10*
C	1010*
D	10101



Software LPM lookup using trie

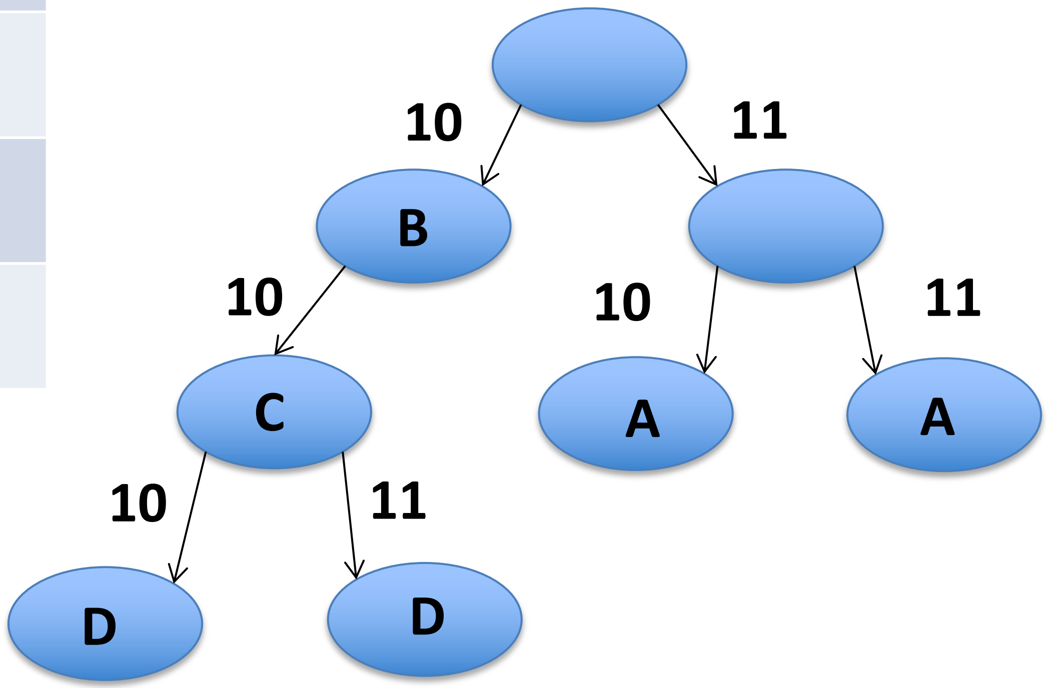
- Prefixes "spelled out" by following path from root
- To find the best prefix spell out address in trie



- 1 lookup per level → max 32 lookups/address!
- Too slow:
  - E.g., "Optical Carrier 48" line (2.5 Gbps) requires 160ns lookup … or 4 memory accesses

Software LPM lookup: k-ary trie (k=2)

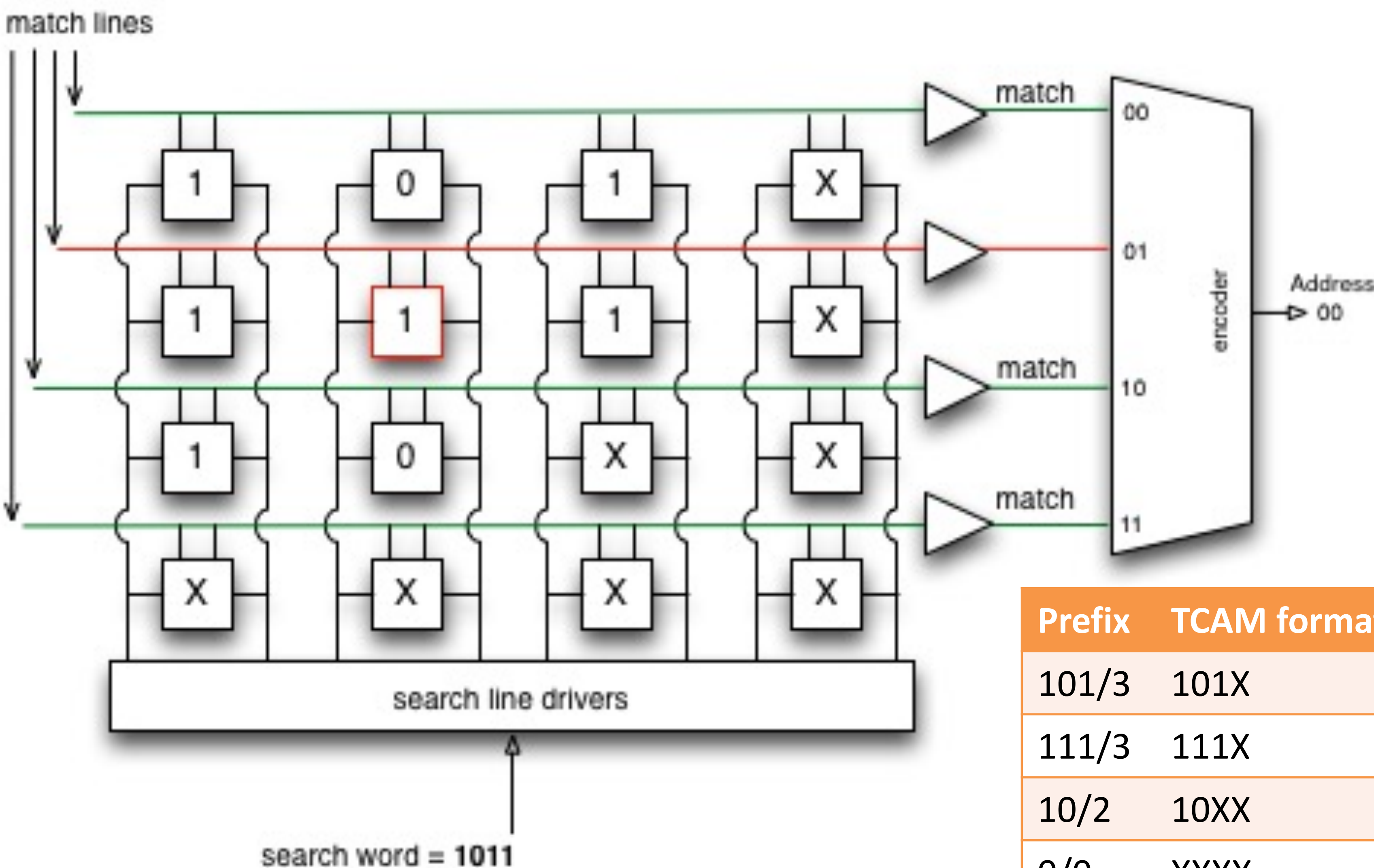
	Prefixes
A	111*
B	10*
C	1010*
D	10101



TCAM: Hardware for LPM lookup

- Content-Address Memory (CAM)
  - Input: address
  - Output: port
  - Exact match, but O(1) in hardware
- ***Ternary*** CAM (TCAM)
  - *i.e.*, can have wildcards: 0, 1, *
  - "value" memory cell and "mask" (care / don't care) cell
- LPM via TCAM
  - In parallel, search all prefixes for all matches
  - Then choose longest match
    - Trick: choose first match, but already sorted by prefix length

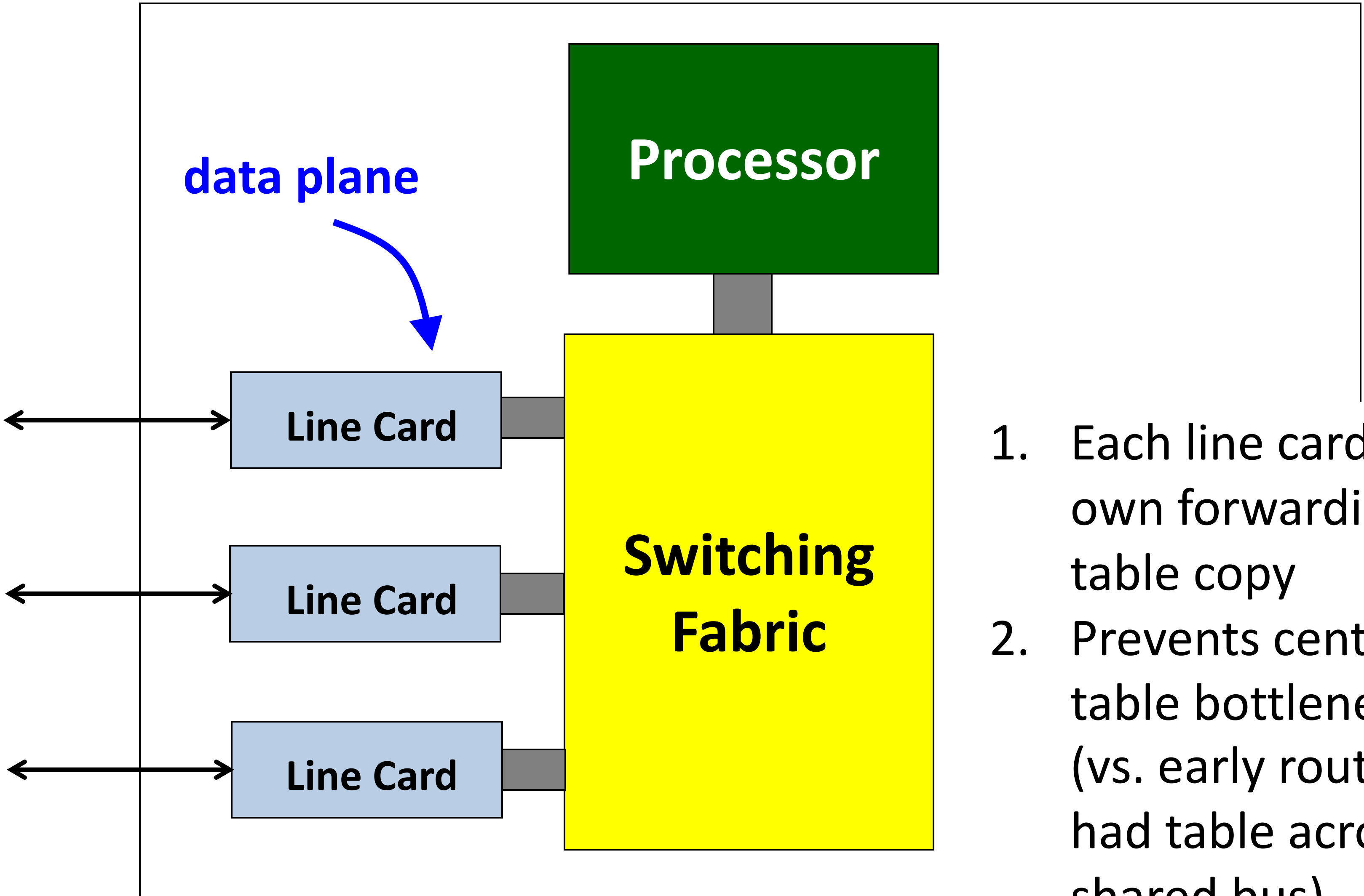
Example: LPM with a TCAM



Prefix	TCAM format
101/3	101X
111/3	111X
10/2	10XX
0/0	XXXX

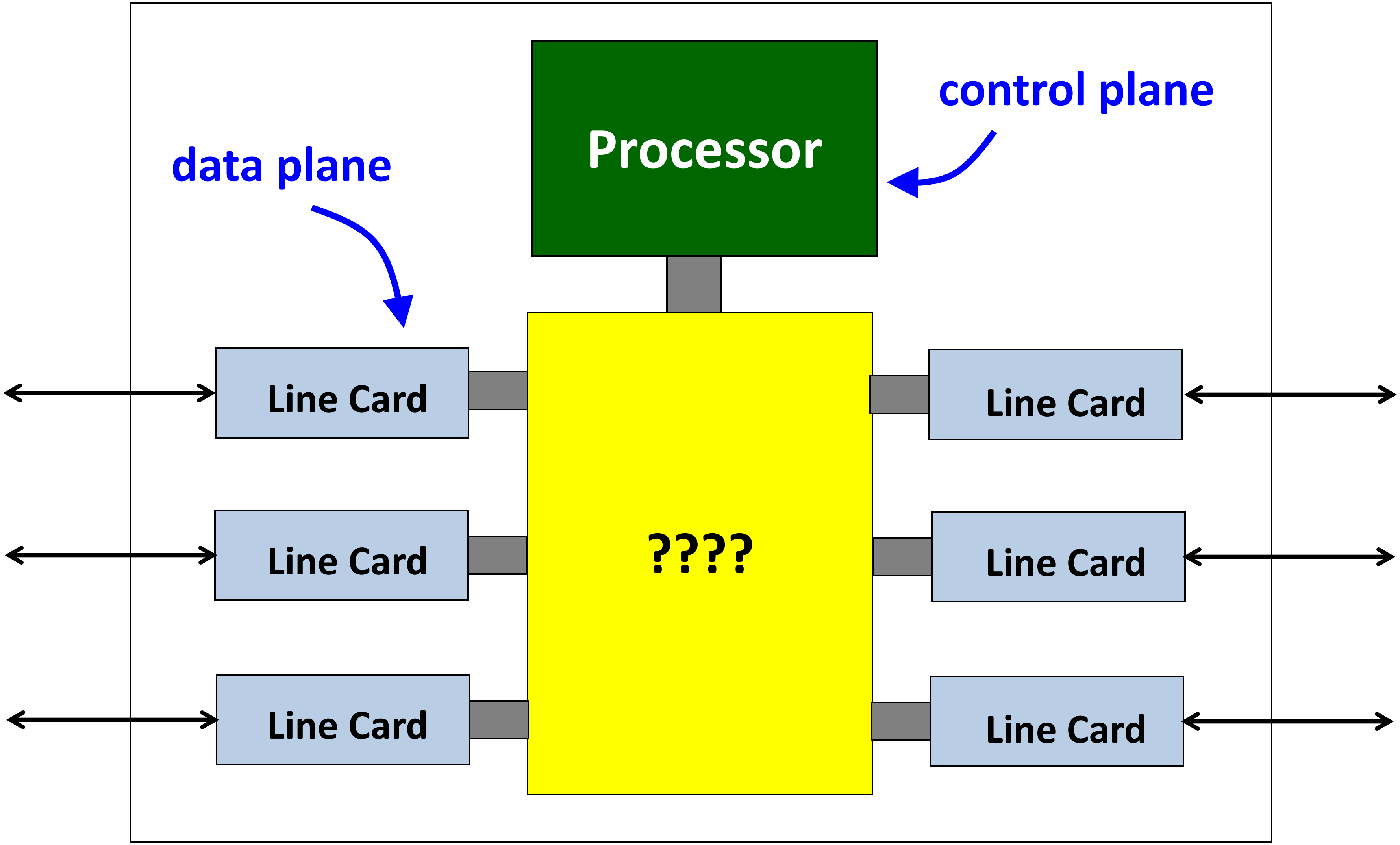
[thenetworksherpa.com/tcam-in-the-forwarding-engine]

Decision: Forwarding table per line card



1. Each line card has own forwarding table copy
2. Prevents central table bottleneck (vs. early routers had table across shared bus)

Decision: Switching Fabric



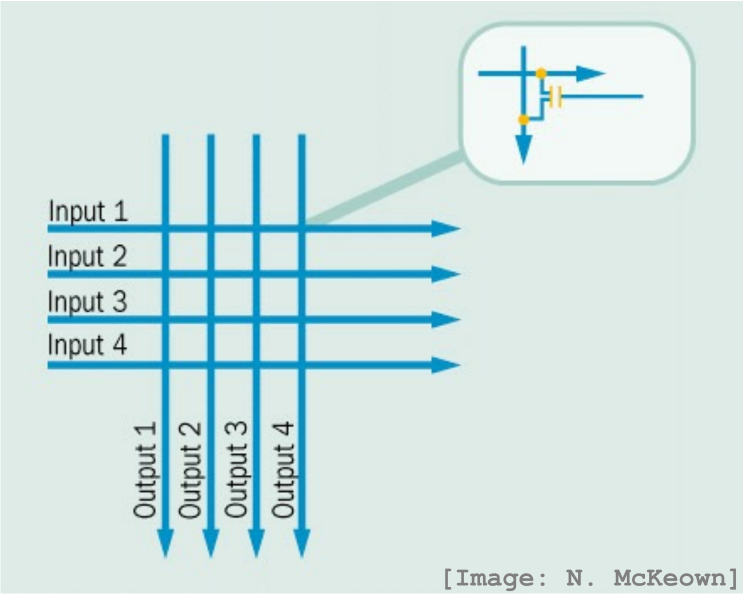
Decision: Switching Fabric

- Shared bus
  - Only one input can speak to one output at a time
  - Shared buses **divide bandwidth** among contenders
  - Electrical reason: speed of bus limited by # connectors

- → **Crossbar interconnect...**

Crossbar interconnect

- Replaces shared bus
- Up to n^2 connects join n inputs to n outputs
- **Multiple input ports can then communicate simultaneously w/multiple output ports**

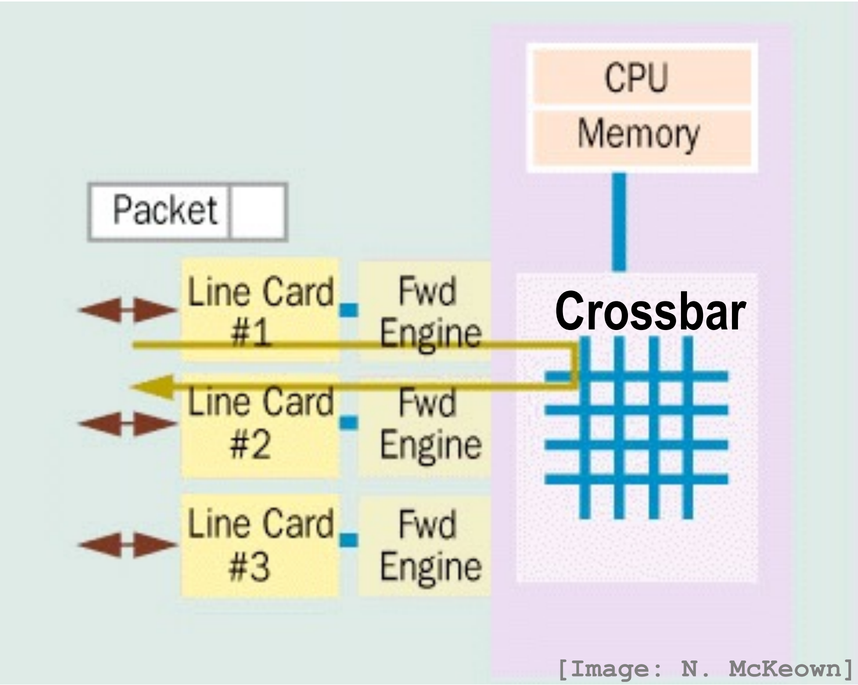


Switching via crossbar

- Datagram moves from input port memory to output port memory via the crossbar
  - e.g. Cisco 12000 family: 60 Gbit/s; fast for core router

✓ Eliminates bus bottleneck

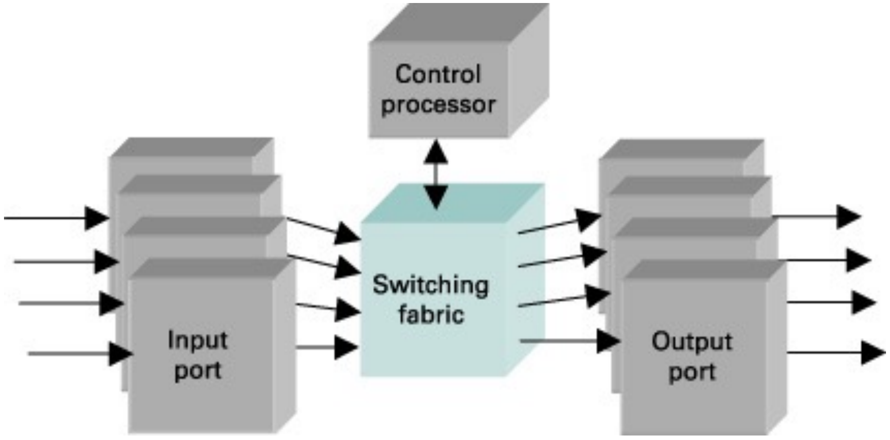
– **Requires algorithm to determine crossbar configuration**



Where does queuing occur?

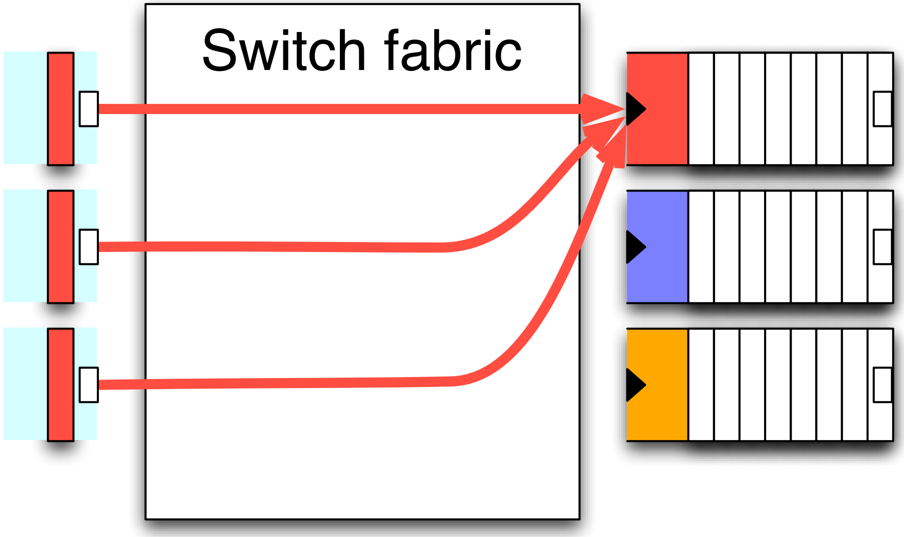
- Central issue in switch design: three choices
  – At input ports (**input queuing**)
  – At output ports (**output queuing**)
  – Some combination of the above

- *n* = max(# input ports, # output ports)



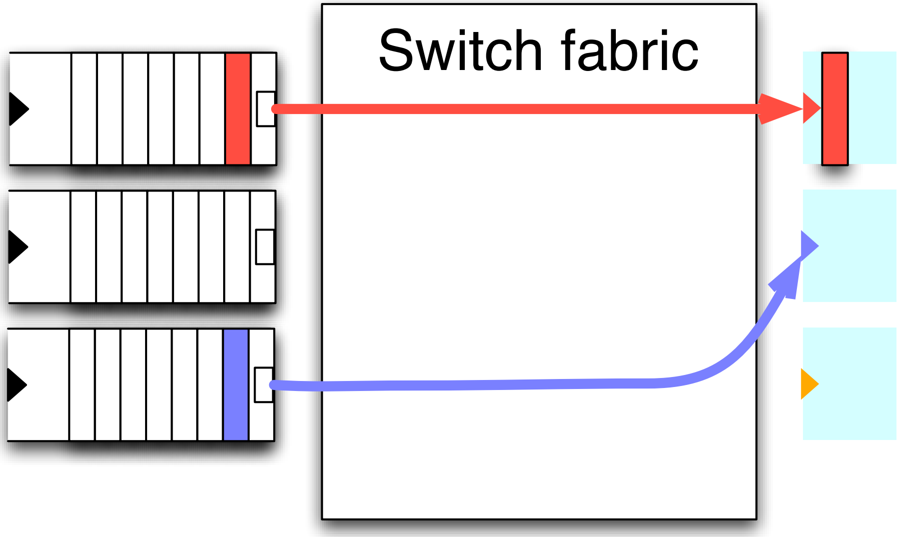
Output queuing

- No buffering at input ports, therefore:
  - **Multiple packets may arrive to an output port** in one cycle; **requires switch fabric speedup of *n***
  - Output port buffers all packets
- **Drawback: Output port speedup required: *n***



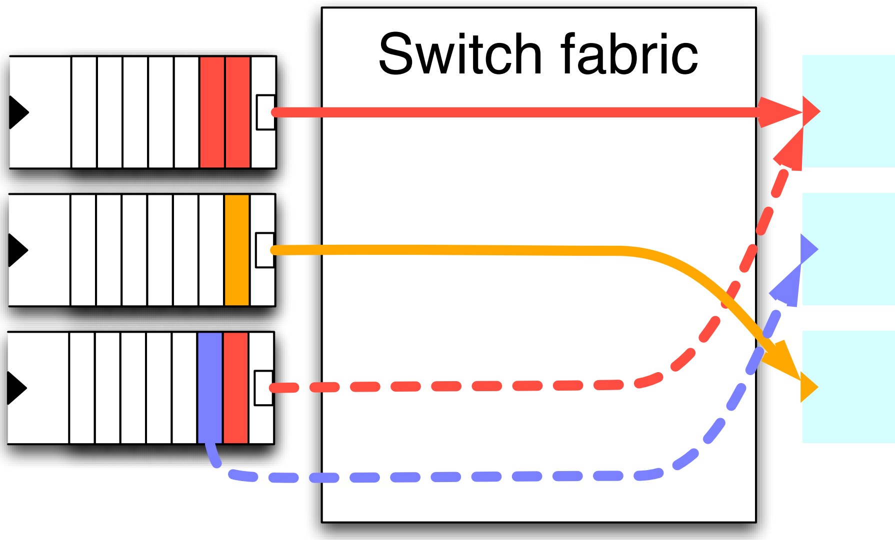
Input queuing

- Input ports buffer packets
- Send **at most one** packet per cycle to an output
  - **Switch fabric forwarding speedup required: *n***



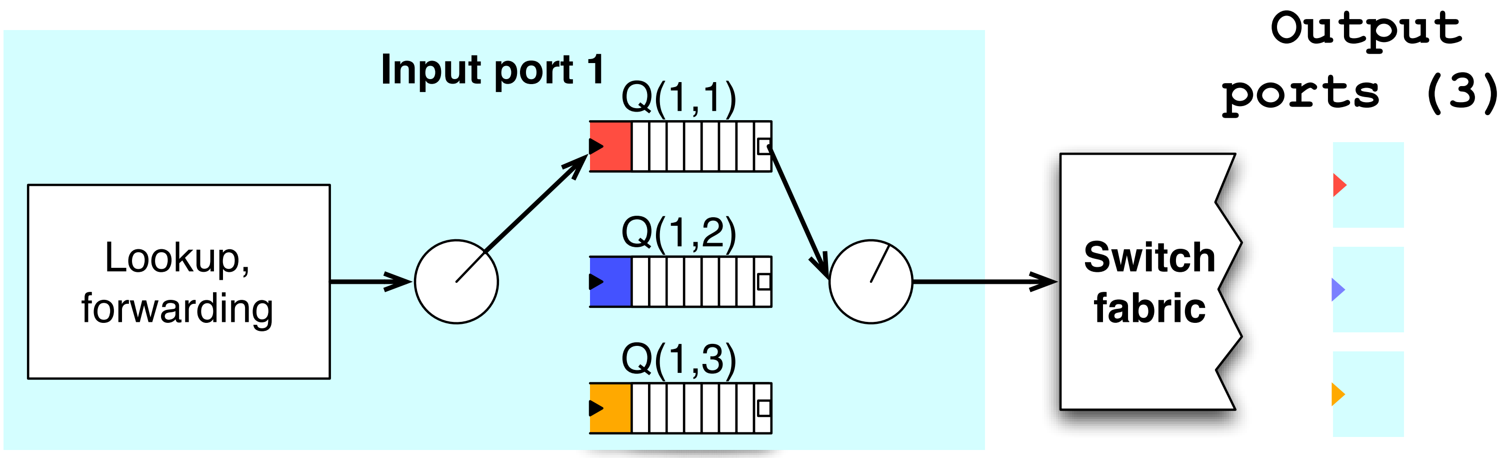
Input queuing: Head-of-line blocking

- One packet per cycle sent to any output
- Blue packet **blocked** despite available capacity at output ports and in switch fabric
- **Reduces throughput of the switch**

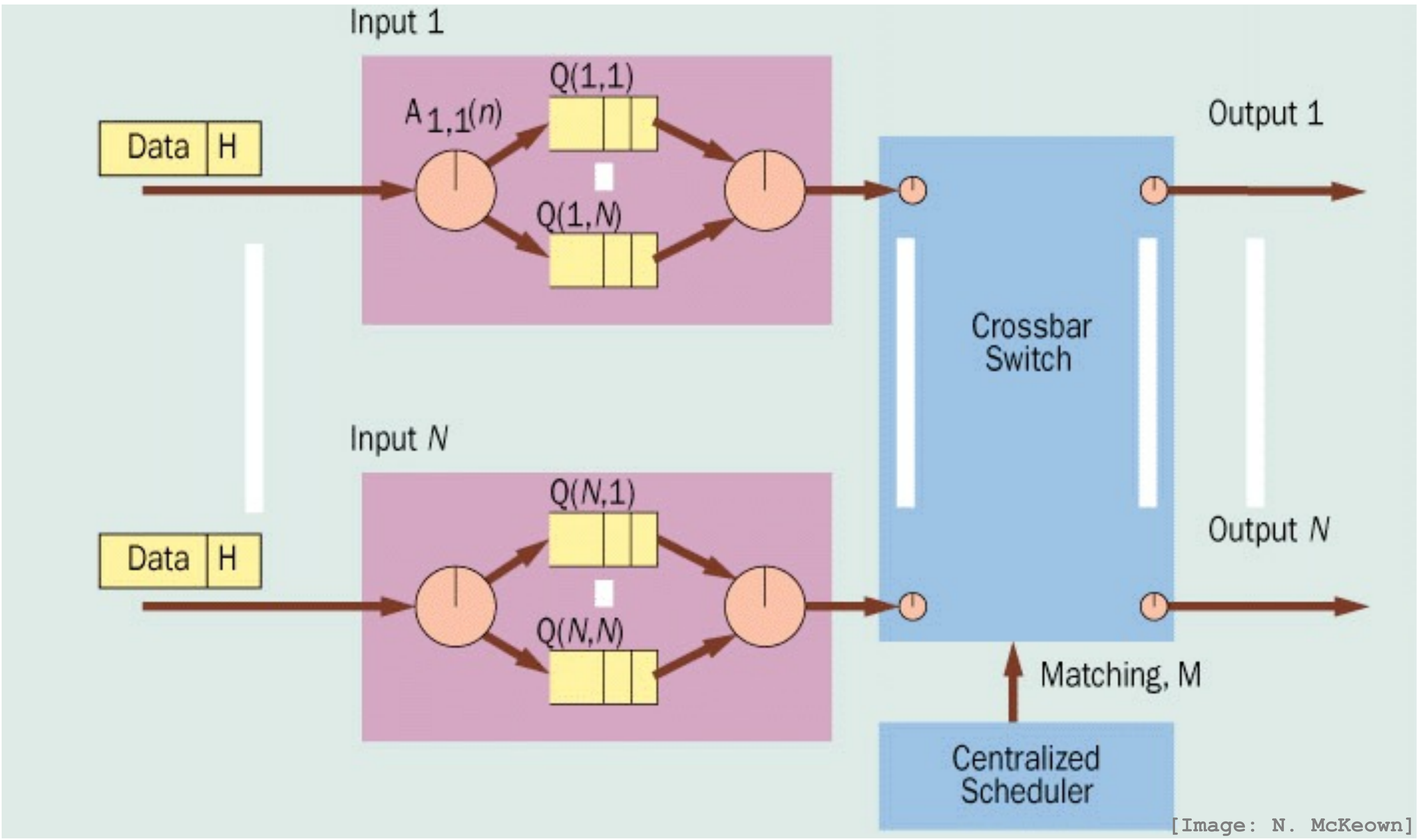


Solution: Virtual output queuing

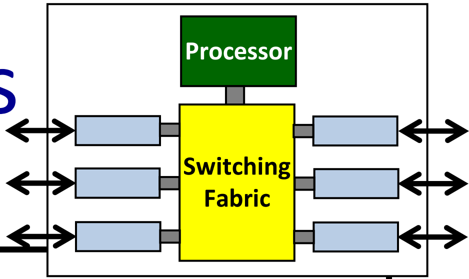
- On each input port, one input queue per output port
- Input port places packet in virtual output queue (VOQ) corresponding to output port of forwarding decision
  - ✓ No head-of-line blocking
  - ✓ All ports (input and output) operate at same rate
  - – **Need to schedule fabric, choosing which VOQs when**



Virtual output queuing



Data, Control, & Management Planes



	Data	Control	Management
Time-scale	Packet (ns)	Event (10 ms to sec)	Human (min to hours)
Tasks	Forwarding, buffering, filtering, scheduling	Routing, signaling	Analysis, configuration
Location	Line-card hardware	Router software	Humans or scripts

Cisco 8000 Series Routers

- Up to 648 400 GbE
- 260 Tbps backplane

Conclusions

- Physical devices sharing L2 & L3 networks have many common features
  - Forward table lookups
  - Queueing and backplane switching
  - Fast vs. slow paths
    - Switches and routers separate routing decisions (control plane) from forwarding actions (data plane)
- High speed necessitates innovation
  - Specialized hardware
  - Software algorithms