COS 461 *Computer Networks*

Class Meeting, Lectures 3 & 4

Kyle Jamieson

Spring 2023

Today

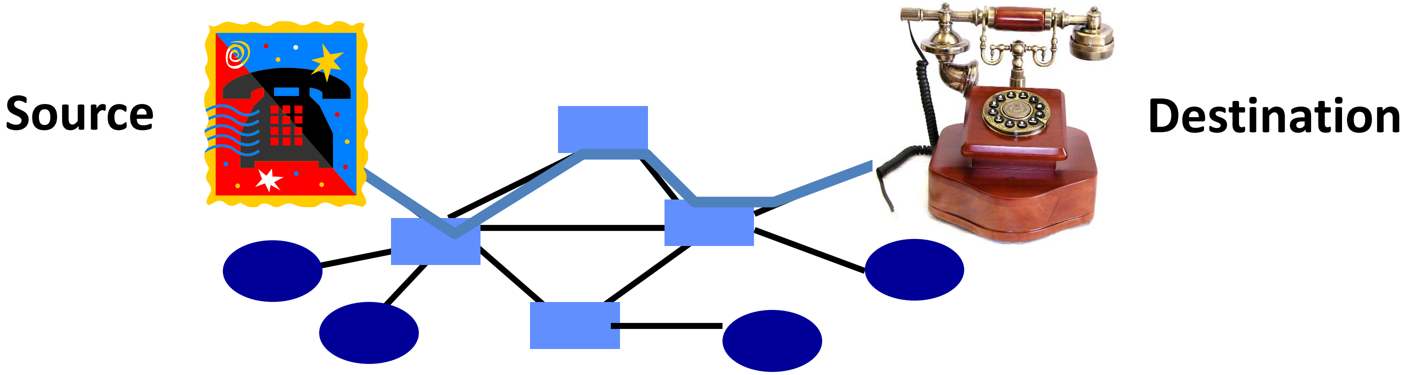
1. **Internet Protocol: Design Discussion**

2. Core Internet Routers

A Reliable Network: Circuit Switching

(*e.g.*, Phone Network)

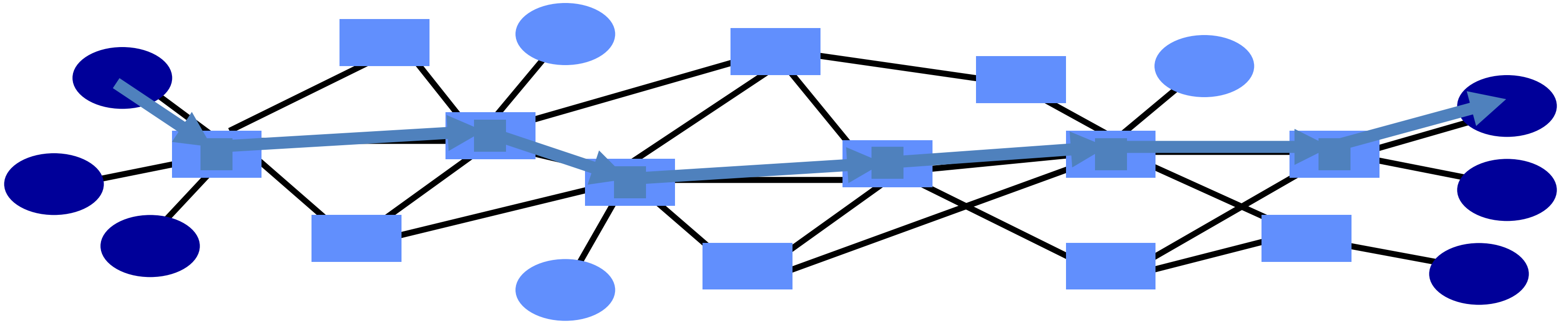
Network	*Global* ***reliable voice call***
Link	Best-effort *local* packet delivery



Set up circuit (allocate resources), transfer data, tear down circuit

Review: Internet Best Effort **Datagram Switching**

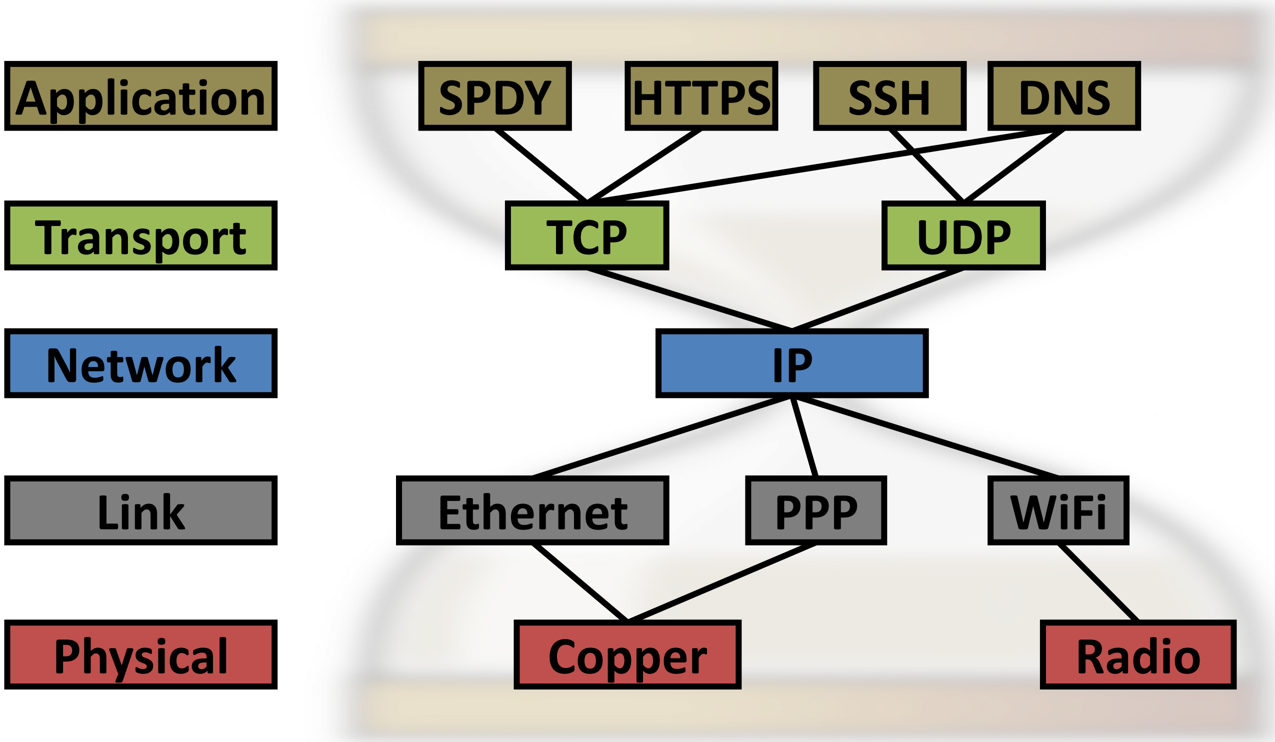
- Message divided into packets (***datagrams***)
  - – Header identifies the destination address
- Datagrams travel **separately** through network
  - – Forwarding based on the destination address
  - – Packets may be buffered temporarily
- Destination reconstructs the message



Packet (Y) vs. Circuit Switching (A)?

• Predictable performance	Circuit
• Network never blocks senders	Packet
• Reliable, in-order delivery	Circuit
• Low delay to send data	Packet
• Simple forwarding	Circuit
• No overhead for packet headers	Circuit
• High utilization under most workloads	Packet
• No per-connection network state	Packet

The Internet hourglass



- Only one network-layer protocol: Internet Protocol (IP)
- The narrow IP layer facilitates **interoperability**

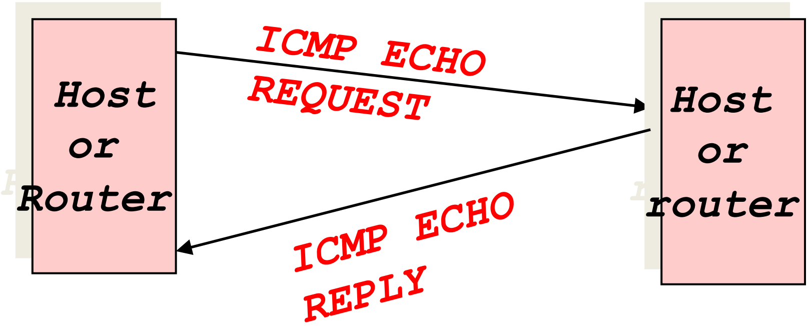
Problem for the Internet: How to cope with different MTUs?

- Each network has a maximum datagram size: ***maximum transmission unit*** (***MTU***)

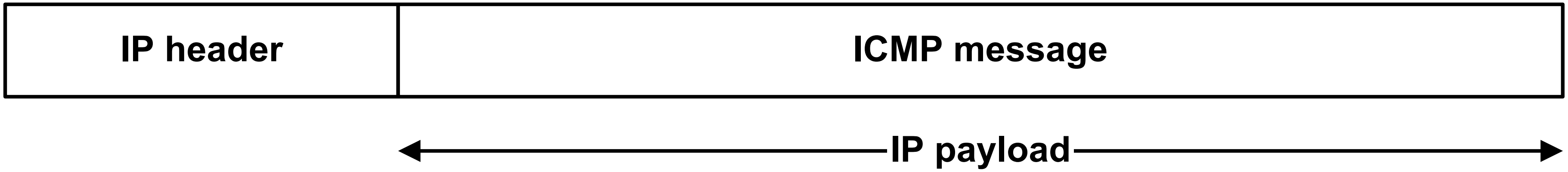
- Don't want to send all datagrams sized with the lowest MTU of any link layer in existence
  - Inefficient, MTU is unknown, and changes depending on route

ICMP: A Helper

- The **Internet Control Message Protocol (ICMP)** is a helper protocol that supports IP with facility for
  - Error reporting
  - Simple queries
  - “ping!”:

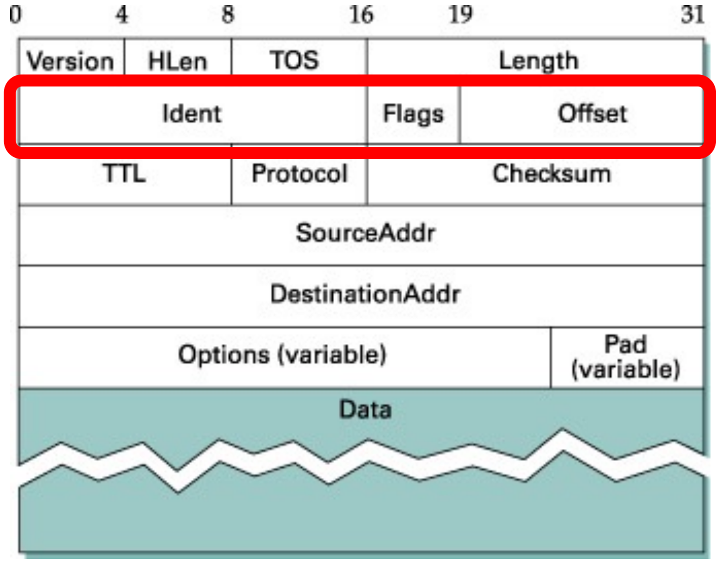


- ICMP messages are encapsulated as IP datagrams:



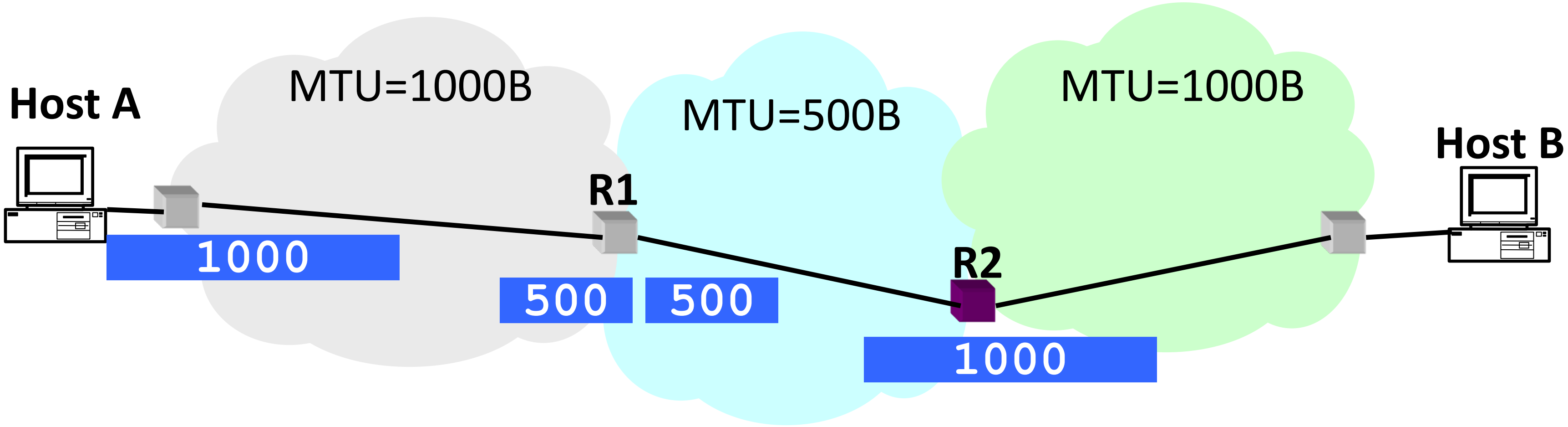
IP’s datagram fragmentation

- **Ident** (16 bits): identifies which fragments belong together
- **Flags**:
  - More (**M**): =1 if this fragment is not the last one, else =0
  - Don’t Fragment (**D**) even if packet too big
    - Instead, routers drop & send back a “Too Large” ICMP control message
- **Offset** (13 bits): part of the original datagram this fragment covers (eight-byte units)



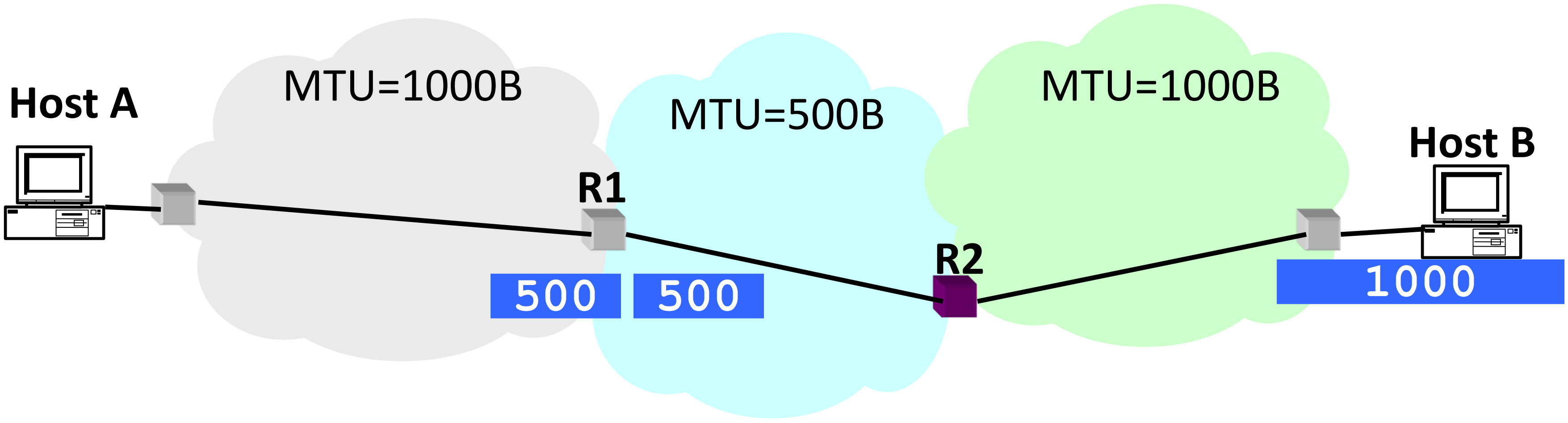
Where should reassembly happen?

- **Answer #1:** within the network, with no help from end-host *B* (receiver)



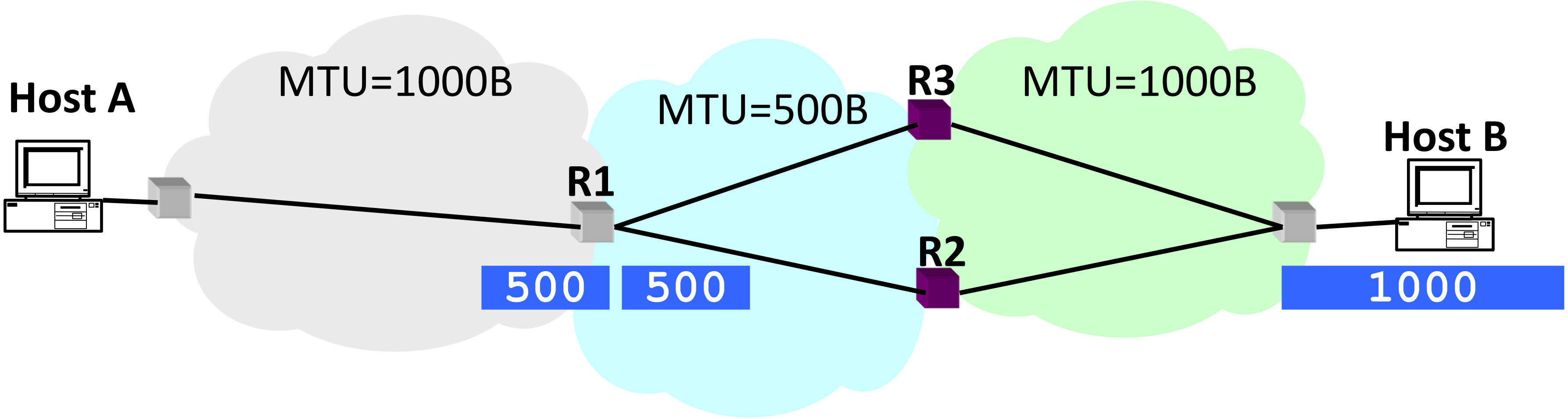
Where should reassembly happen?

- **Answer #1:** within the network, with no help from end-host *B* (receiver)
- **Answer #2:** at end-host *B* (receiver) with no help from the network



Where should reassembly happen?

- **Answer #1:** within the network, with no help from end-host *B* (receiver) ✗
- **Answer #2:** at end-host *B* (receiver) with no help from the network ✓
- Fragments can travel across different paths!



Problem for the Internet: How to cope with different MTUs?

- **Goal:**
  - Send datagrams of size = minimum MTU over all networks on the path they take (***path MTU***)
    - This would minimize header overheads

Path MTU discovery

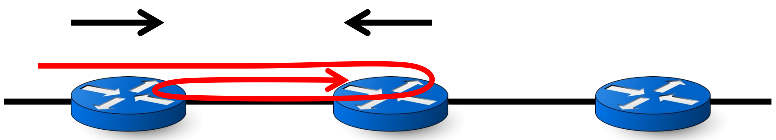
- Source initially sets path MTU (PMTU) estimate = MTU of first hop

- Send datagrams with **Don’t Fragment** (DF) bit set in Flags field

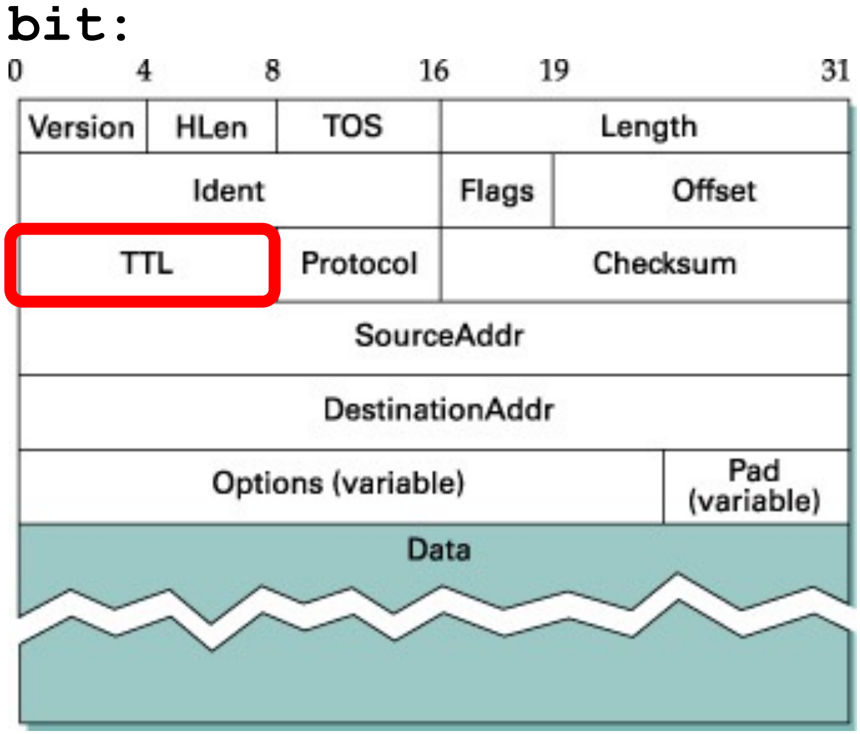
- If any datagrams are too big to be forwarded
  - Intermediate router will discard them & send "too large" ICMP message
  - Source reduces its PMTU estimate

The time-to-live field

- **TTL** (8 bits)
  - Potentially catastrophic problem
  - Forwarding loops can cause datagrams to cycle forever
  - As these accumulate, eventually consume all capacity



- Solution: Routers decrement TTL field at each hop, packet is discarded if TTL reaches zero
  - ICMP "time exceeded" message sent back to the source



bit:

0	4	8	16	19	31
Version	HLen	TOS	Length		
Ident			Flags	Offset	
TTL		Protocol	Checksum		
SourceAddr					
DestinationAddr					
Options (variable)				Pad (variable)	
Data					

Q's: MAC vs. IP Addressing

- Hierarchically allocated

  Y) MAC M) IP [✓ C) Both] A) Neither

- Organized topologically

  Y) MAC [✓ M) IP] C) Both A) Neither

- Forwarding via exact match on address

  [✓ Y) MAC] M) IP C) Both A) Neither

- Automatically calculate forwarding by observing data

  [✓ Y) Ethernet switches] M) IP routers C) Both A) Neither

- Per connection state in the network

  Y) MAC M) IP C) Both [✓ A) Neither]

- Per host state in the network

  [✓ Y) MAC] M) IP C) Both A) Neither

Core Internet Router Design

- e.g. Cisco 8000 Series Routers
  - Up to 648 400 GbE
  - 260 Tbps backplane

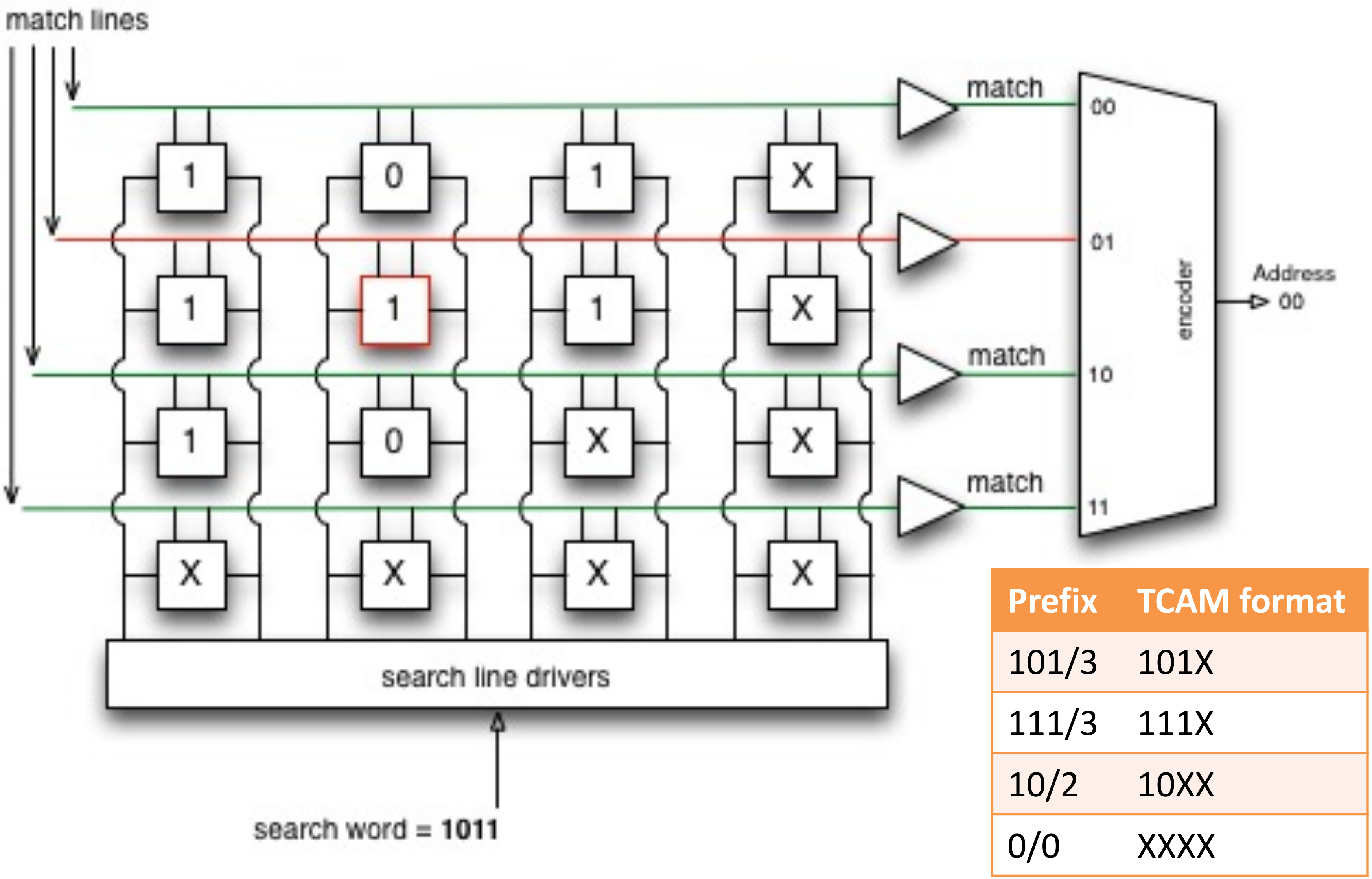
Longest Prefix Match (LPM)

- 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

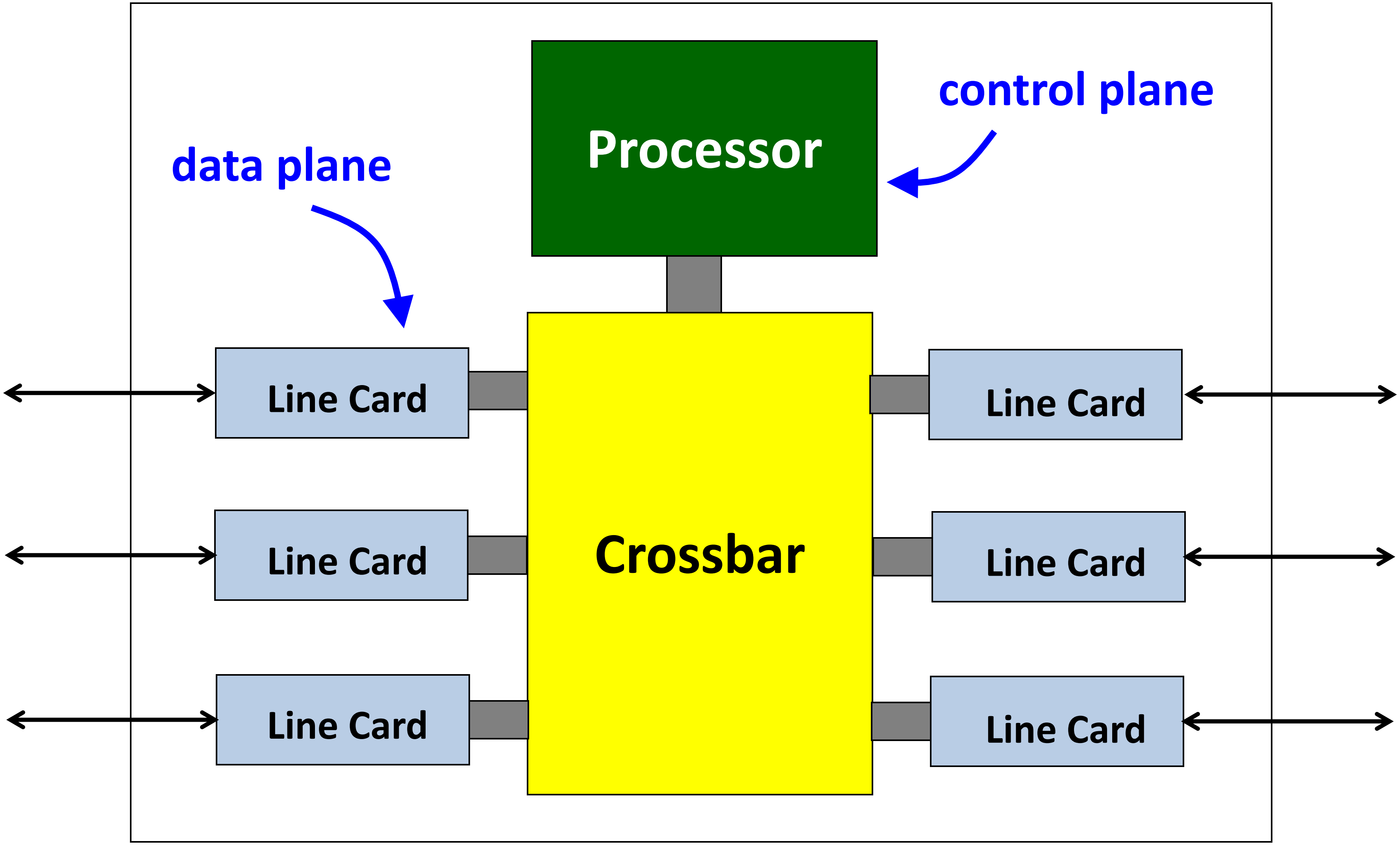
Example: LPM with a TCAM



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

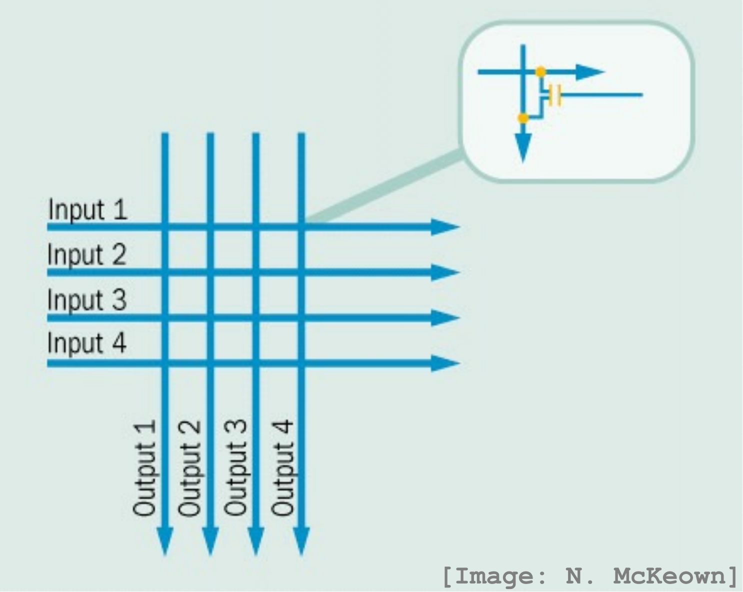
[thenetworksherpa.com/tcam-in-the-forwarding-engine]

Router Design: Overview



Crossbar interconnect

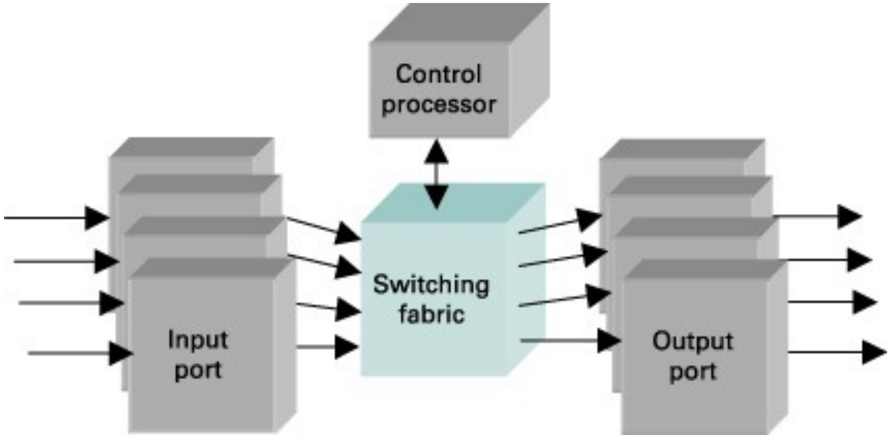
- Replaces shared bus
- Up to *n*^{2} connects join *n* inputs to *n* outputs
- **Multiple input ports can then communicate simultaneously w/multiple output ports**



[Image: N. McKeown]

Key Design Question:
Where does queuing occur?

- Central issue in router design: three choices
  - At input ports (**input queuing**)
  - At output ports (**output queuing**)
  - Some combination of the above
- *n* = max(# input ports, # output ports)



Coming Up in 461

Next Class Meeting

Lectures 5 (Transport Layer) and
6 (Congestion Control)

Precepts this Thursday and Friday:
Error Control Codes