Routers; LPM; Protocols

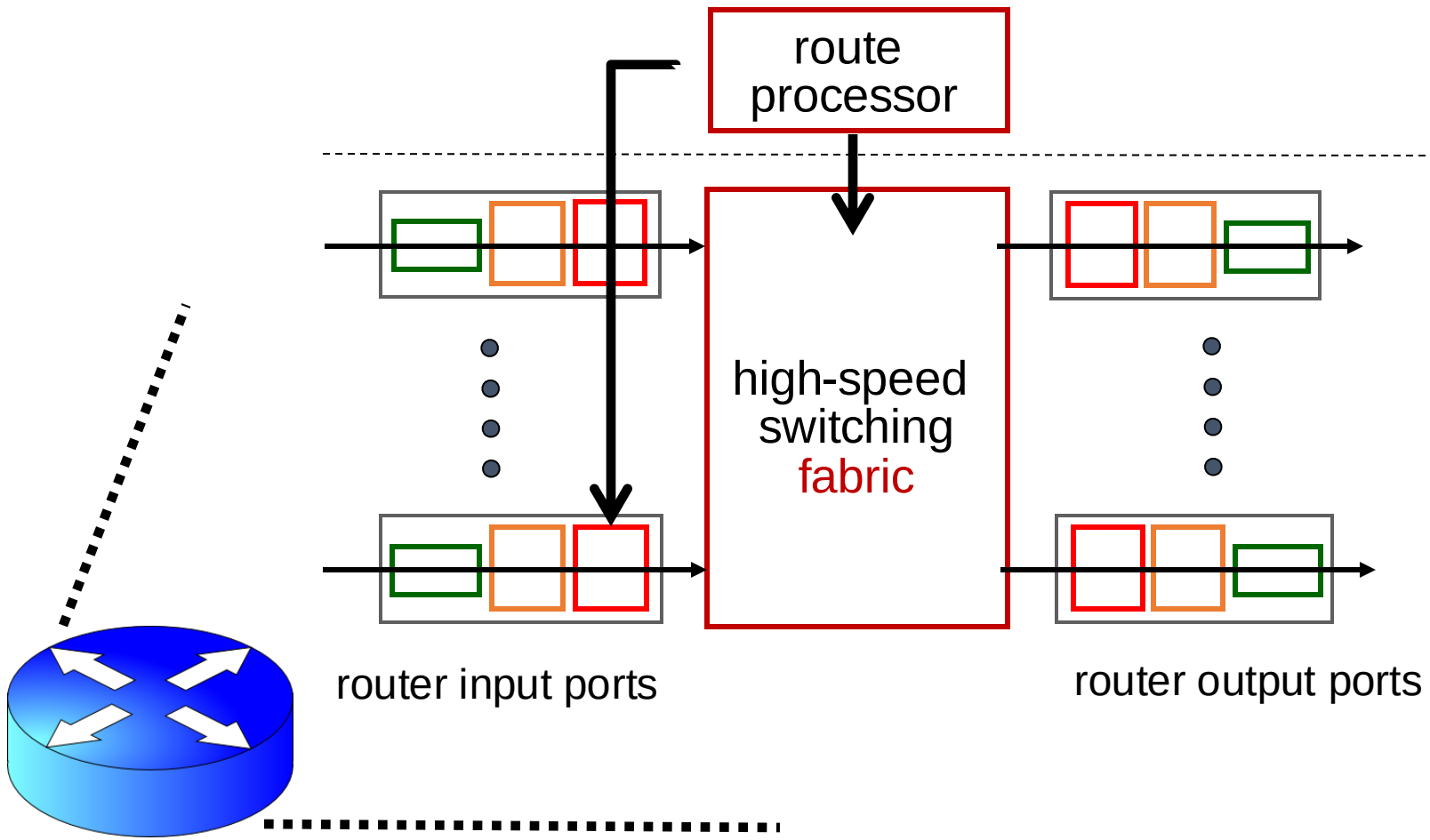
Lecture 22

http://www.cs.rutgers.edu/~sn624/352-F24

Srinivas Narayana

RUTGERS
UNIVERSITY | NEW BRUNSWICK

Review: Router architecture



assuming distributed routing,
routing function: decide which ports packets need to exit

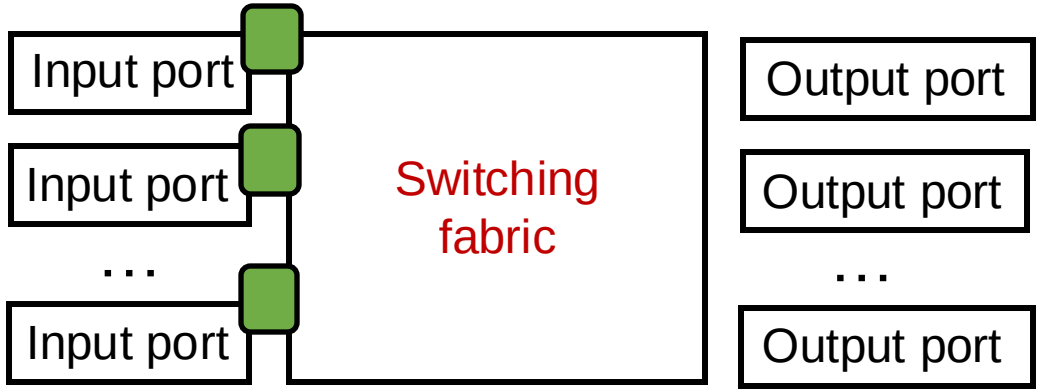
Control plane

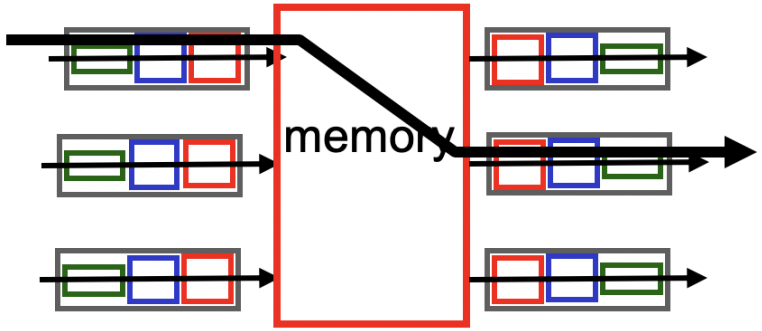
Data plane

Forwarding function: move packets from one input port to another.

Fabrics: Types

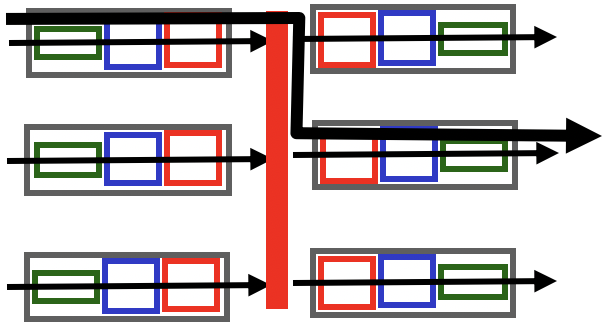
Fabric goal: Ferry as many packets as possible from input to output ports as quickly as possible.





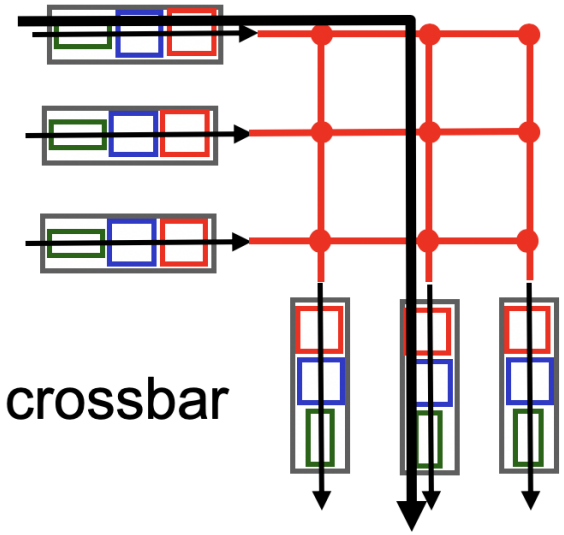
memory

Input port writes packets into shared memory.
Output port reads the packet when output link ready to transmit.



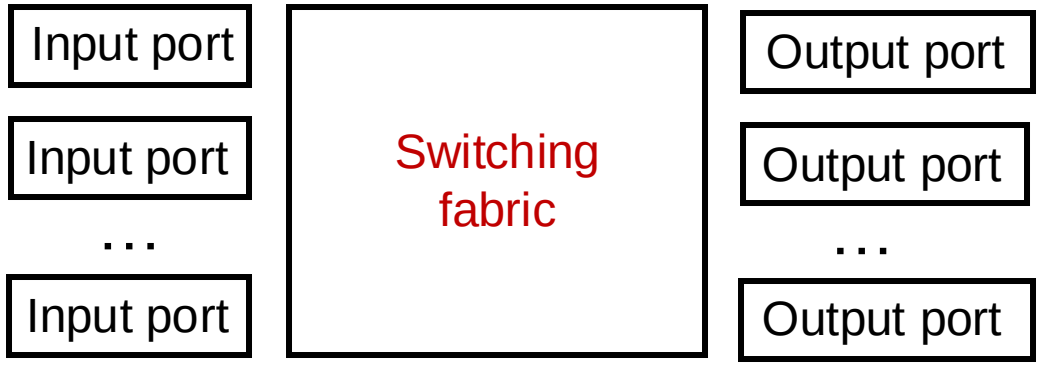
bus

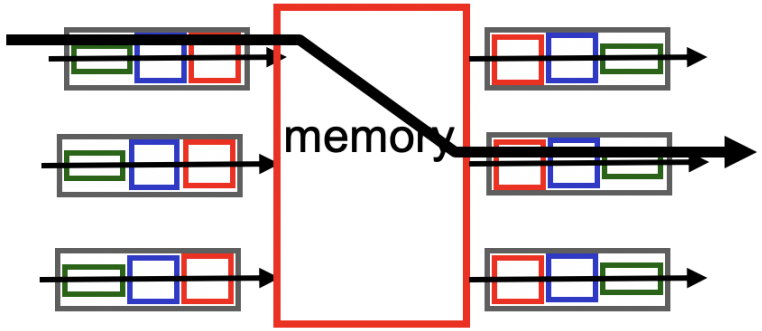
Single shared channel to move data from input to output port.
Easy to build buses; technology is quite mature.



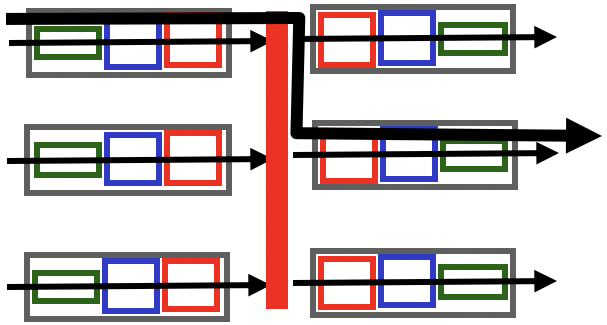
Each input port has a physical data path to every output port.
Switch at the cross-over points turns on to connect pairs of ports.

Fabrics: Types

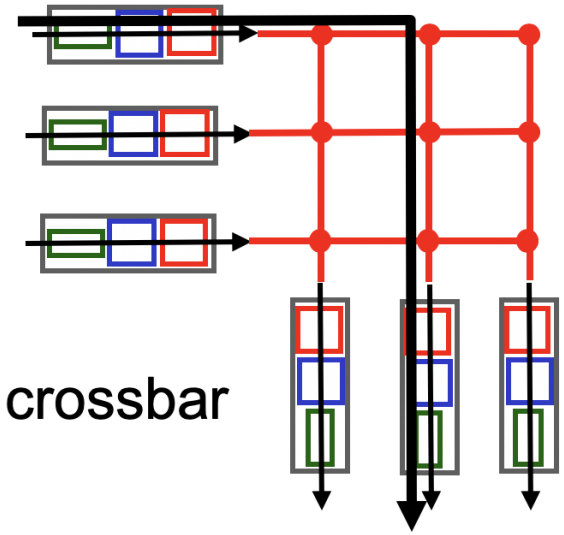




memory



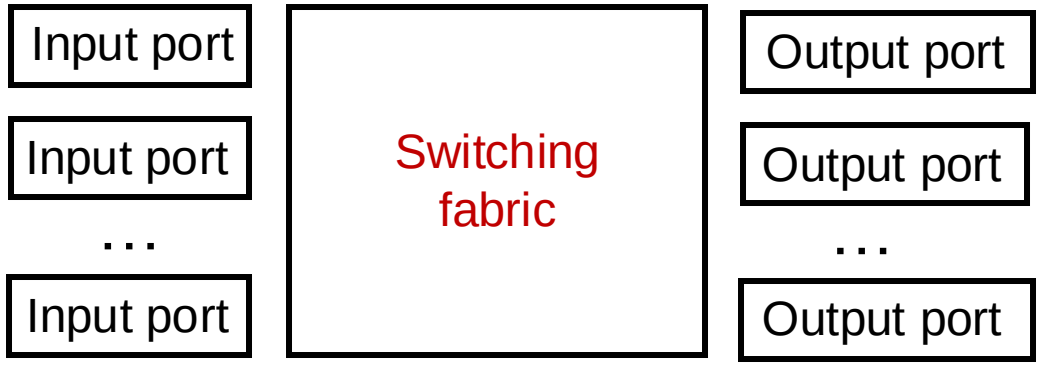
bus



Modern high-speed routers use highly optimized shared-memory-based interconnects.

Crossbars can get expensive as the number of ports grows (N^2 connections for N ports)
MGR uses a crossbar and schedules (in,out) port pairs.

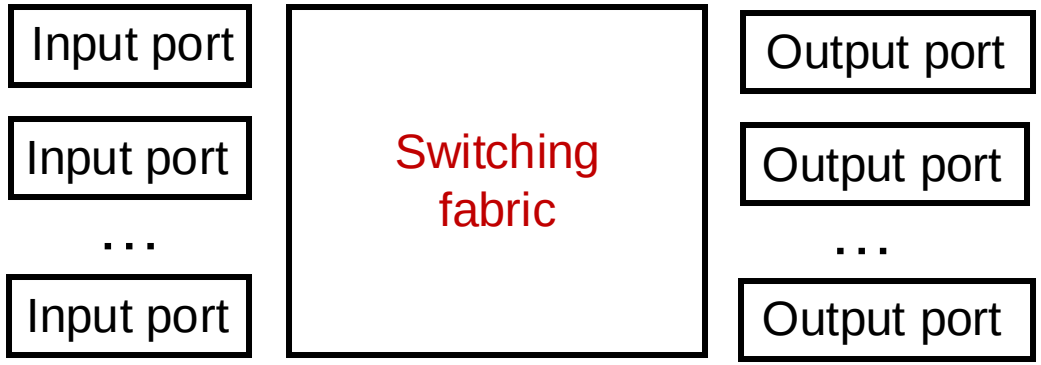
Nonblocking fabrics



- High-speed switching fabrics designed to be nonblocking:
  - If an output port is “available”, an input port can always transmit to it without being blocked by the switching fabric itself
  - Nontrivial to achieve
- Crossbars are nonblocking by design

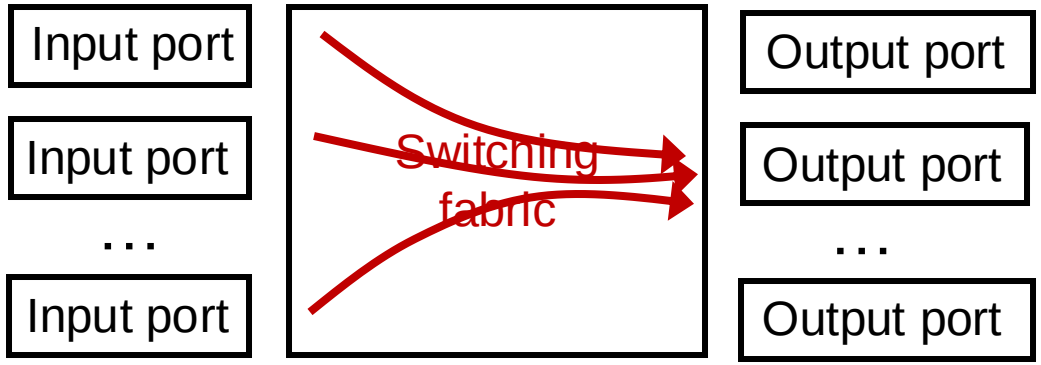
- Shared memory can be designed to be nonblocking if memory accesses can be made fast enough

Nonblocking fabrics



- With a nonblocking fabric, queues aren't formed due to the switching fabric.
  - With a nonblocking fabric, there are no queues due to inefficiencies at the input port or the switching fabric
- Queues only form due to contention for the output port
  - Fundamental, unavoidable, given the route

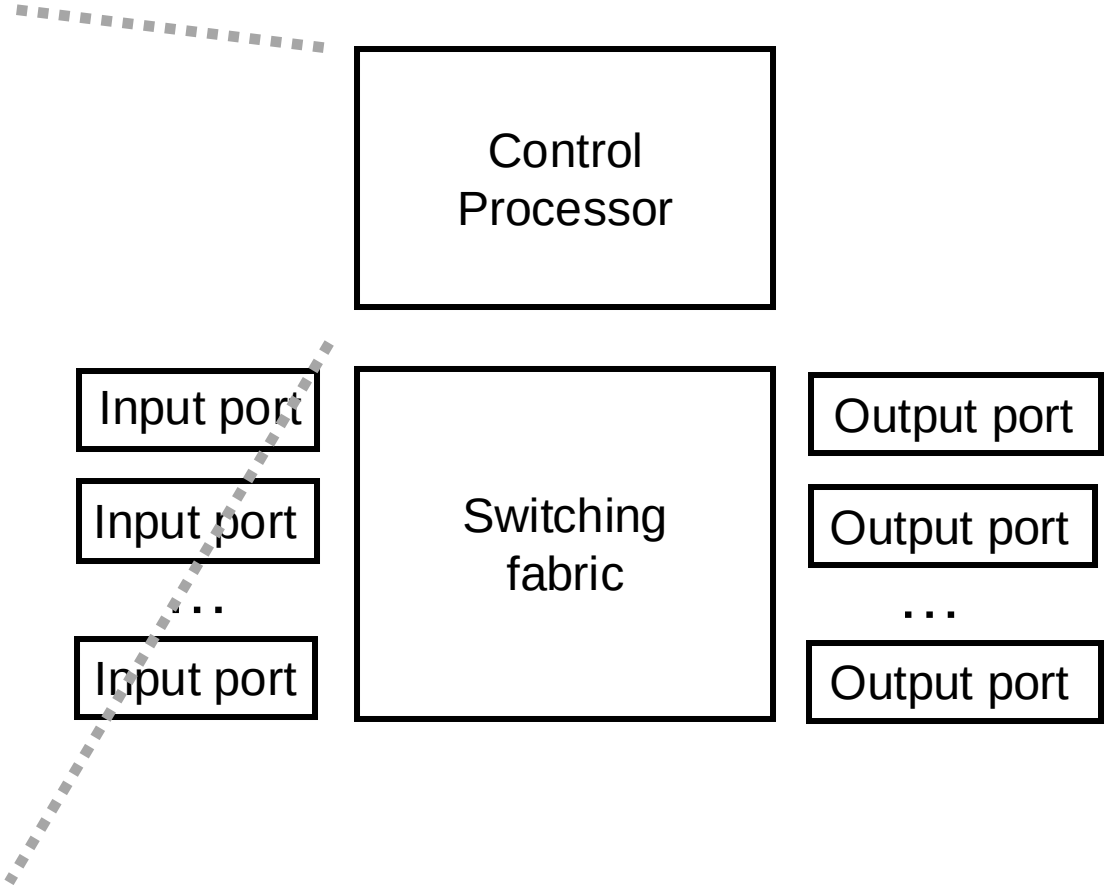
Nonblocking fabrics



- With a nonblocking fabric, queues aren't formed due to the switching fabric.
  - With a nonblocking fabric, there are no queues due to inefficiencies at the input port or the switching fabric
- Queues only form due to contention for the output port
  - Fundamental, unavoidable, given the route
- Typically, these queues form on the output side
  - But can also "backpressure" to the input side if there is high contention for the output port
  - i.e.: can't move pkts to output Qs since buffers full, so buffer @ input

Control (plane) processor

- A general-purpose processor that “programs” the data plane:
  - Forwarding table
  - Scheduling and buffer management policy
- Implements the routing algorithm by processing routing protocol messages
  - Mechanism by which routers collectively solve the Internet routing problem
  - More on this soon.



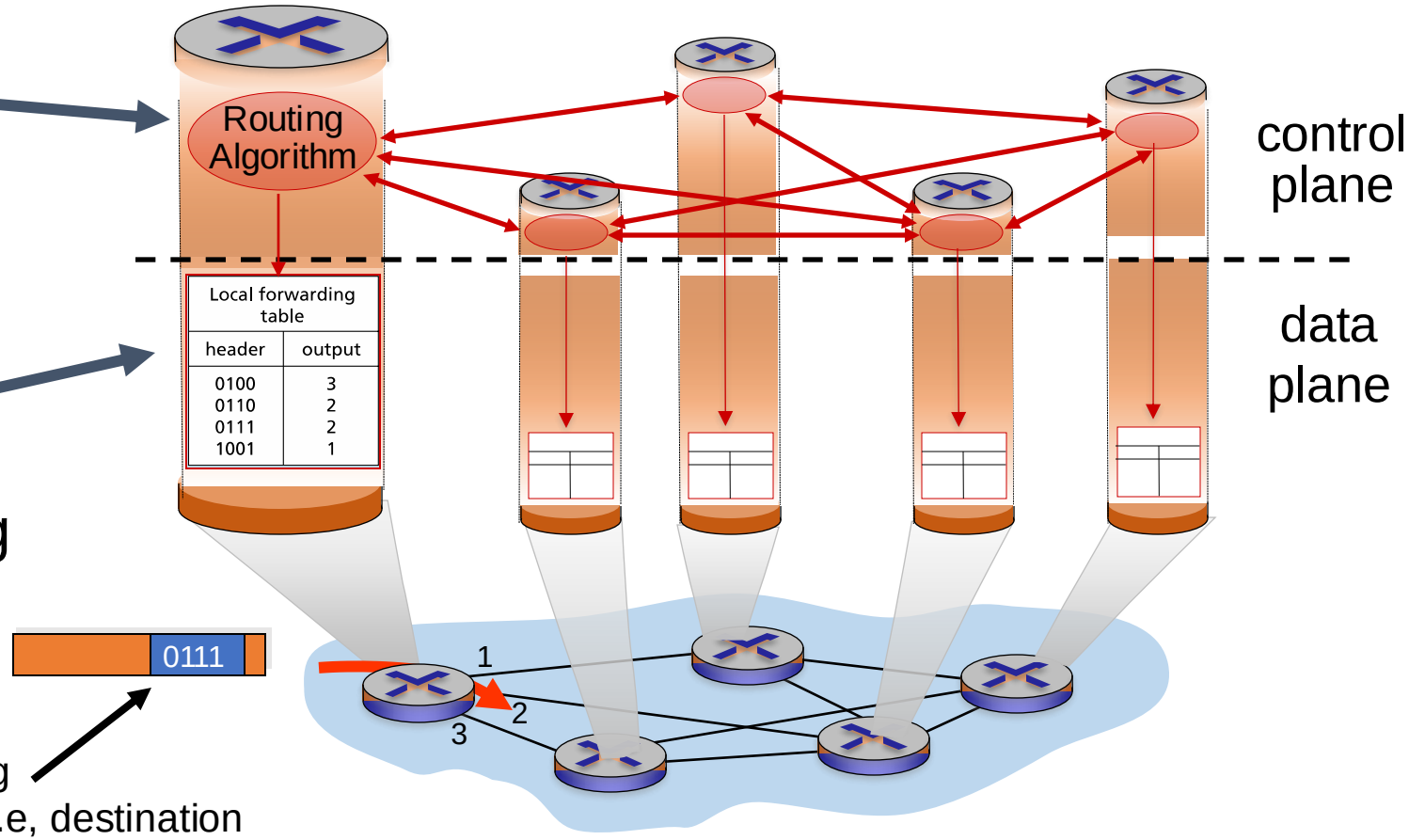
Router design: the bigger picture

Control plane

Traditional distributed routing: per route-change processing (~ a few tens of seconds)

Data plane

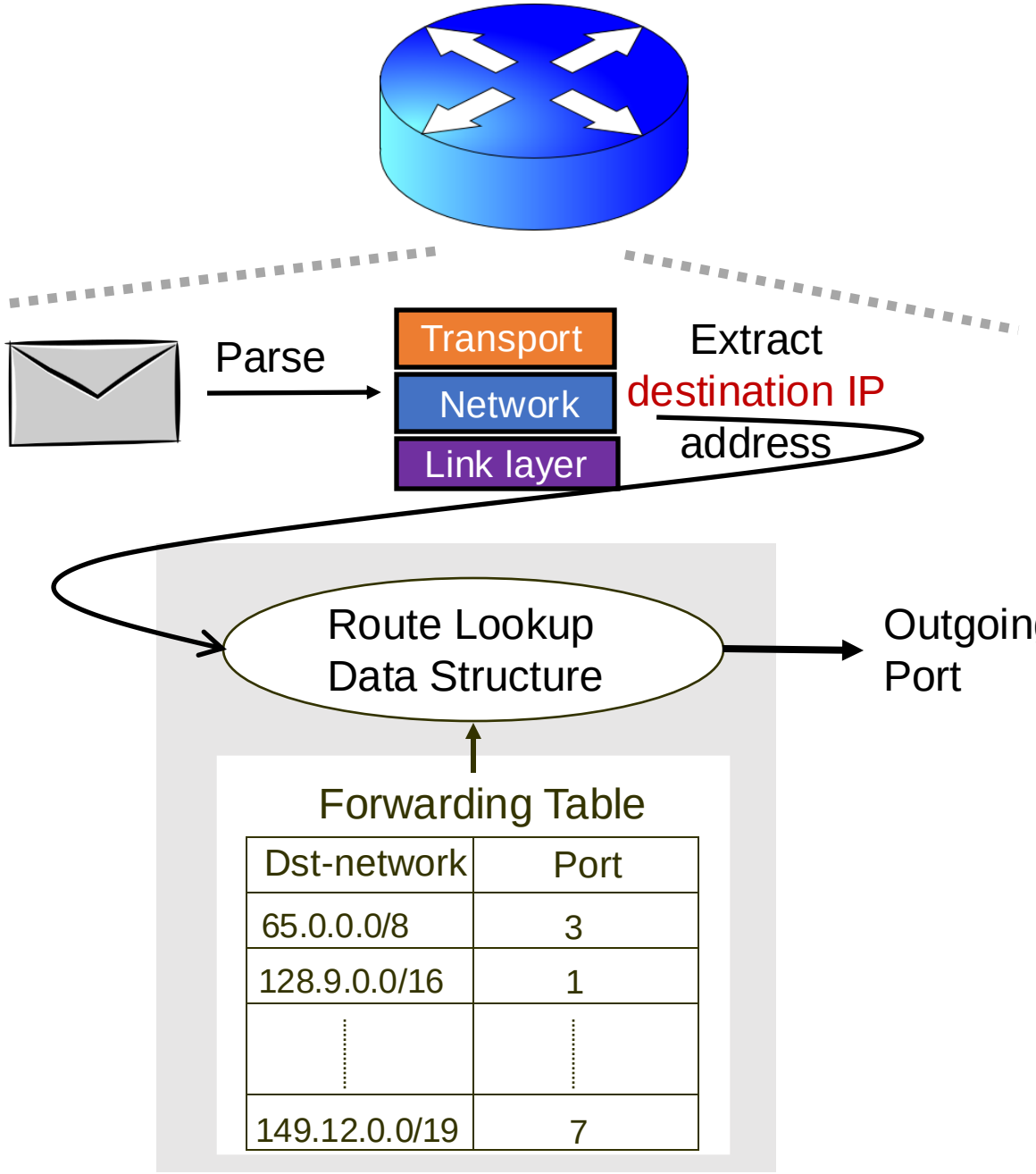
per-packet processing (~ tens of nanoseconds)



Longest Prefix Matching

Review: Route lookup

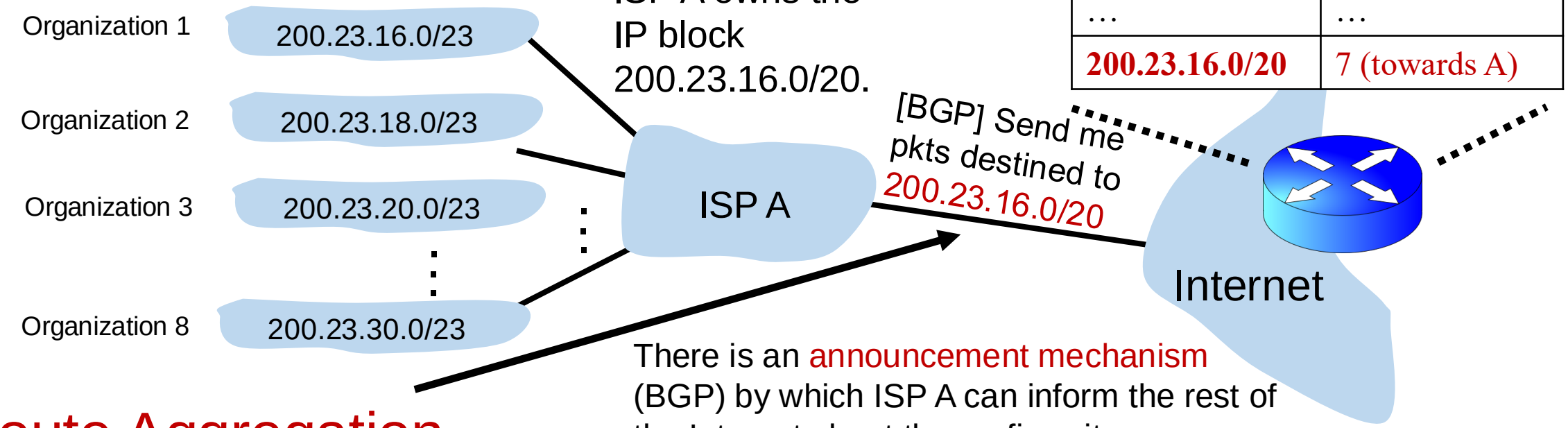
- Table lookup matches a packet against an IP prefix
  - Ex: 65.12.45.2 matches 65.0.0.0/8
- Prefixes are allocated to organizations by Internet registries
- But organizations can reallocate a subset of their IP address allocation to other orgs



ISP = Internet Service Provider

Example of IP block reallocation

Suppose ISP A reallocates a part of its IP block to orgs 1… 8



Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
…	…
200.23.16.0/20	7 (towards A)

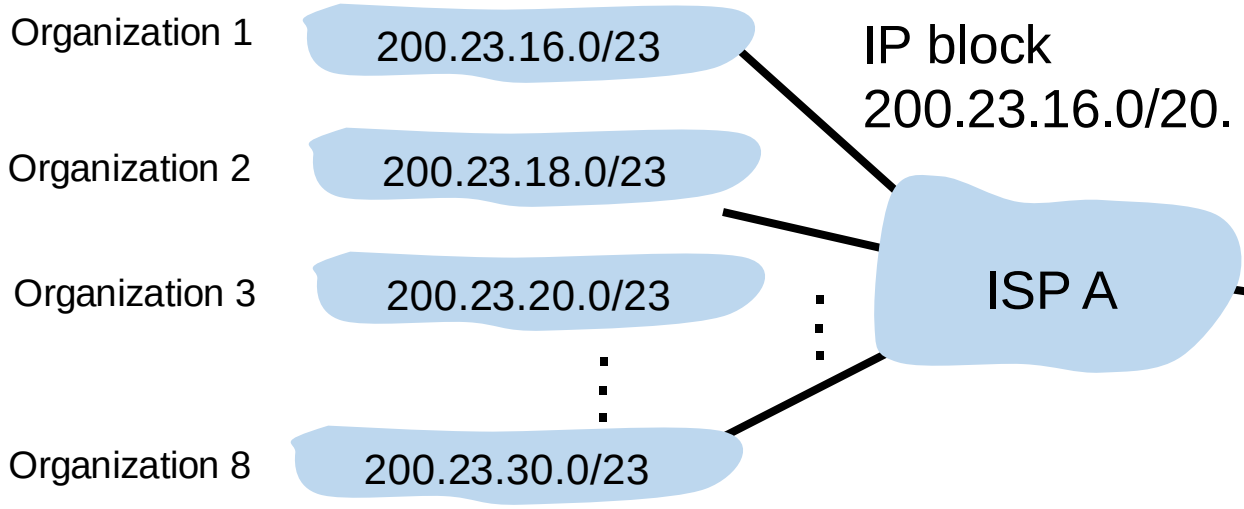
There is an announcement mechanism (BGP) by which ISP A can inform the rest of the Internet about the prefixes it owns.
It is enough to announce a coarse-grained prefix 200.23.16.0/20 rather than 8 separate sub-prefixes.

Route Aggregation

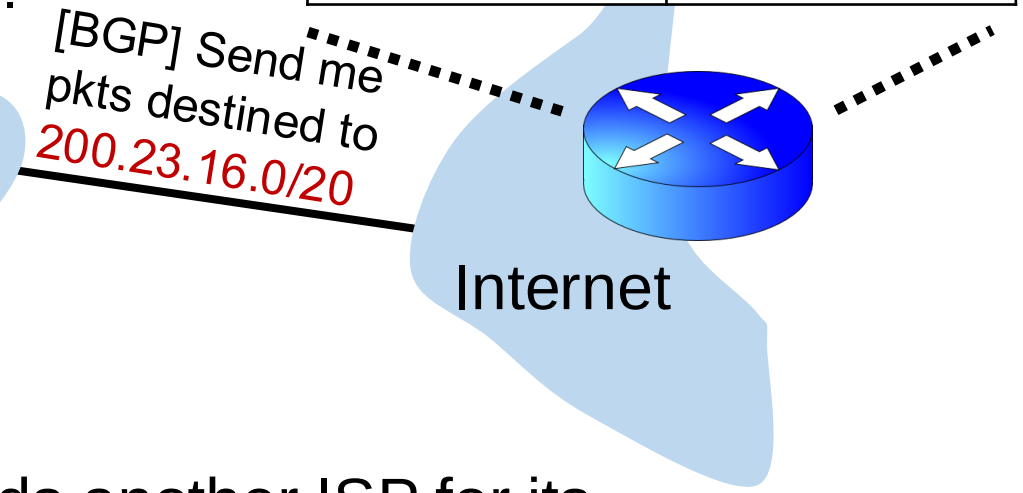
Save forwarding table memory
Fewer routing protocol msgs

Example of IP block reallocation

Suppose ISP A reallocates a part of its IP block to orgs 1… 8



Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
…	…
200.23.16.0/20	7 (towards A)

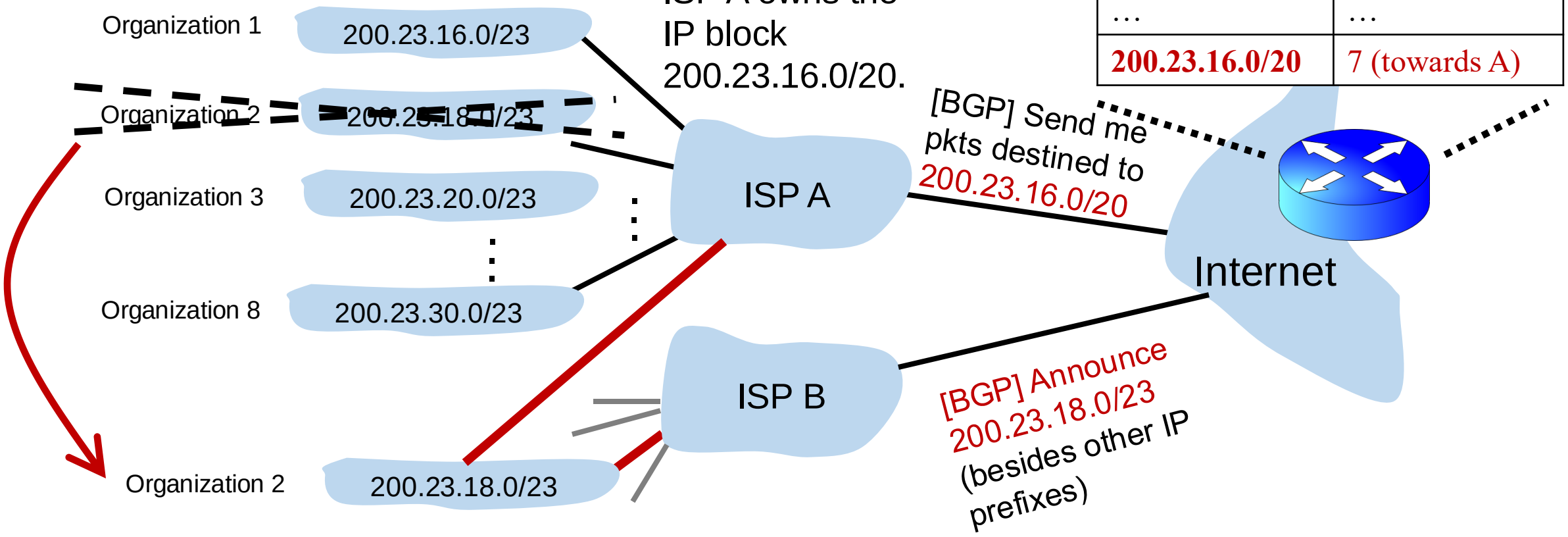


Now suppose one of these organizations adds another ISP for its Internet service and prefers using the new ISP.
Note: it's possible for the organization to retain its assigned IP block.

Example of IP block reallocation

Suppose ISP A reallocates a part of its IP block to orgs 1… 8

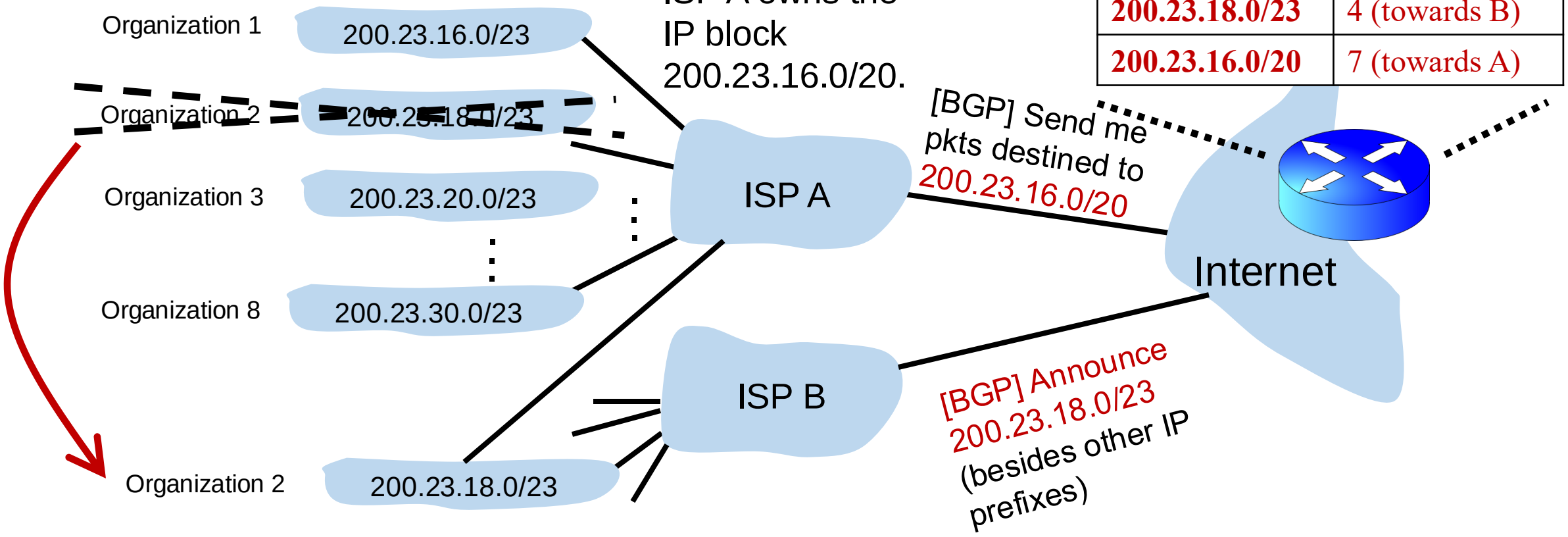
Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
…	…
200.23.16.0/20	7 (towards A)



Example of IP block reallocation

Suppose ISP A reallocates a part of its IP block to orgs 1… 8

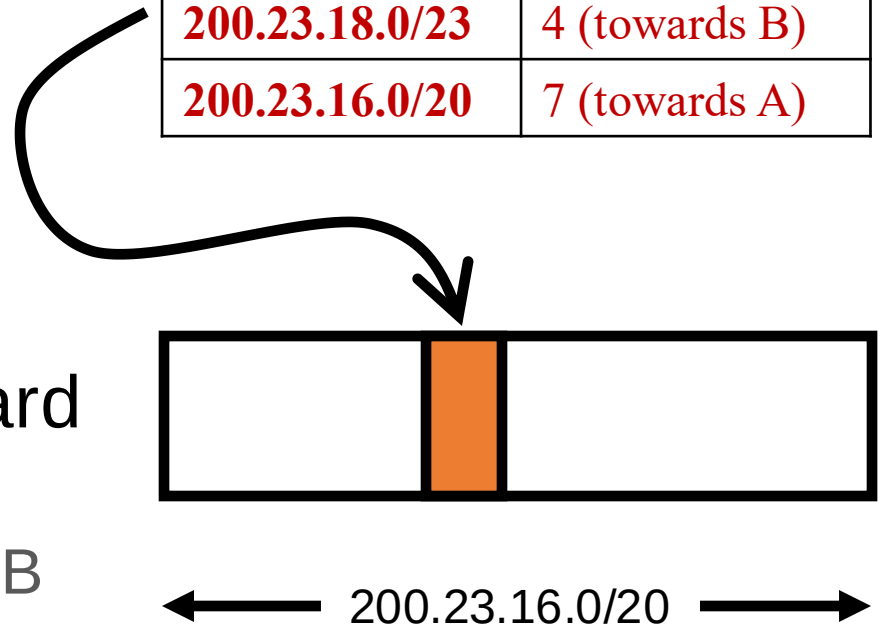
Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
200.23.18.0/23	4 (towards B)
200.23.16.0/20	7 (towards A)



A closer look at the forwarding table

- 200.23.18.0/23 is inside 200.23.16.0/20
- A packet with destination IP address 200.23.18.xx is in both prefixes
  - i.e., both entries match
- Q: How should the router choose to forward the packet?
  - Ideally: The org prefers B, so should choose B

Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
200.23.18.0/23	4 (towards B)
200.23.16.0/20	7 (towards A)

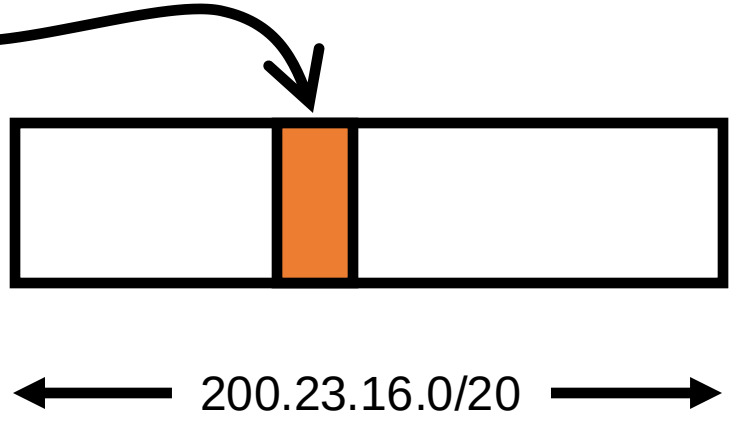


The Internet uses a policy to prioritize: Longest Prefix Matching

Longest Prefix Matching (LPM)

- Use the longest matching prefix, i.e., the most specific route, among all prefixes that match the packet.
- Policy borne out of the Internet's IP allocation model: prefixes and sub-prefixes are handed out
- Internet routers use longest prefix matching.
  - How would you implement this in software?
  - Interesting algorithmic and design challenges in developing software and hardware

Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
200.23.18.0/23	4 (towards B)
200.23.16.0/20	7 (towards A)

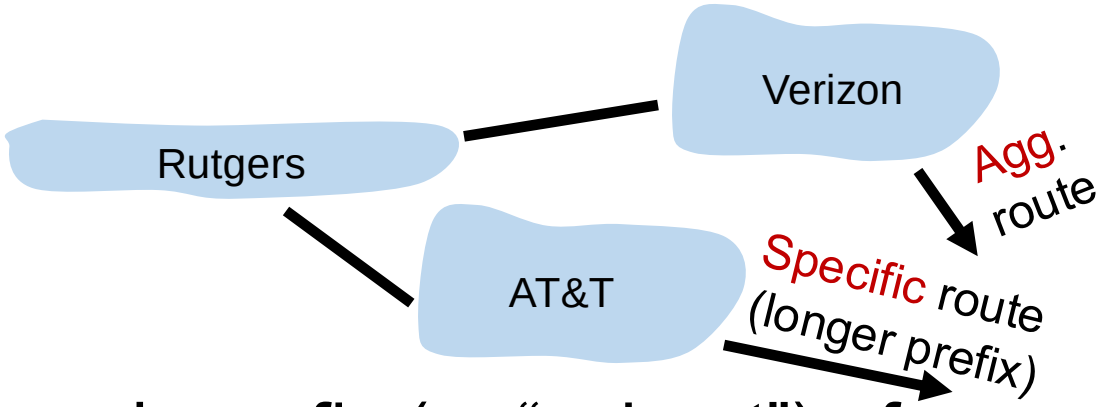


Internet routers perform longest-prefix matching on destination IP addresses of packets.

Why is LPM useful?

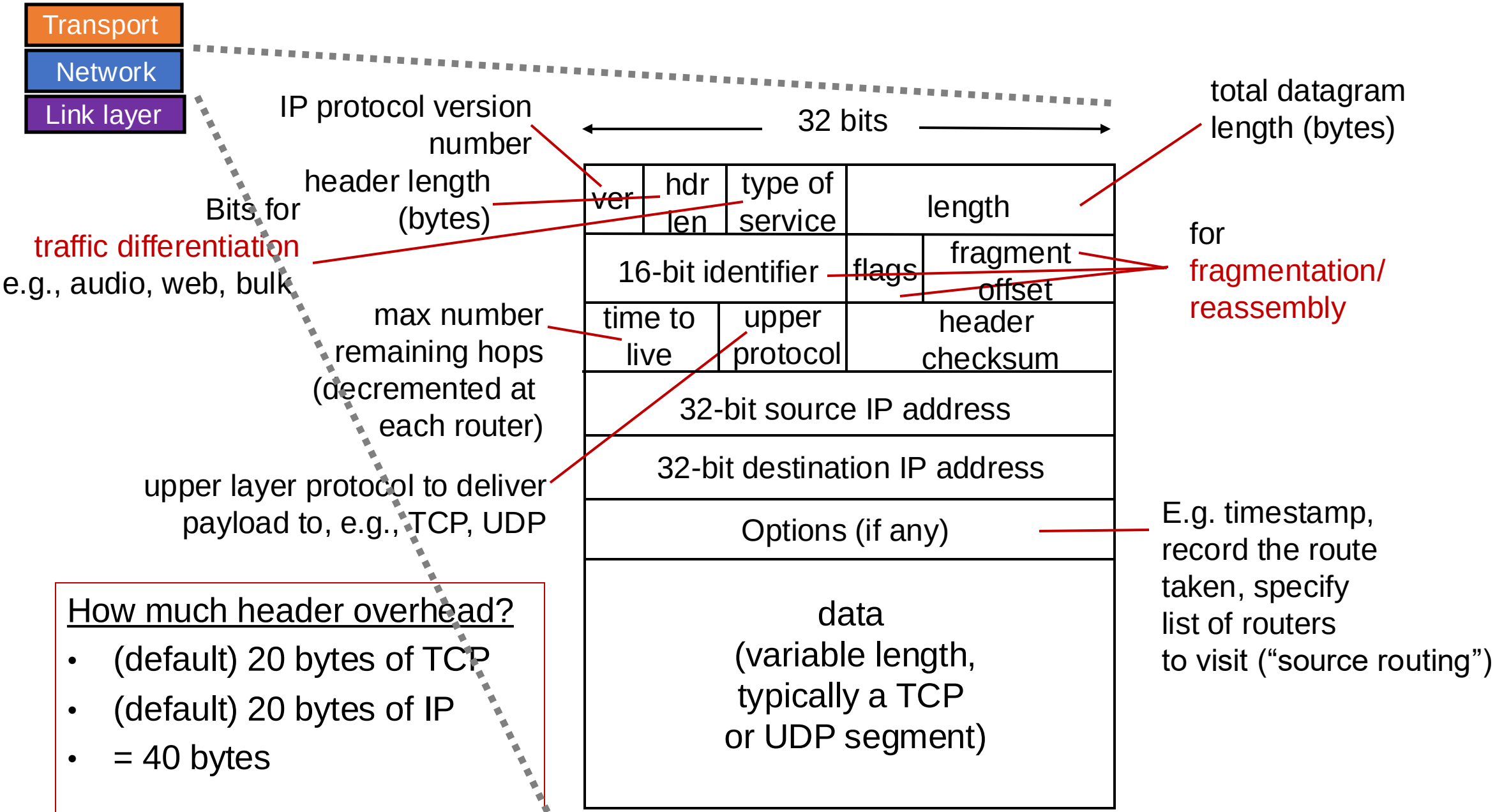
- Help organizations move in one block to a different ISP while retaining their IP prefix assignment.
  - IPs unchanged: e.g., don’t have to update DNS for services in the org
- Also enable an organization (e.g. Rutgers) to connect to two or more Internet Service Providers (ISPs) and express routing preferences
  - Announce longer prefixes to make the rest of the Internet prefer a certain path

Why is LPM useful?



- An ISP (e.g., Verizon) has allocated a sub-prefix (or “subnet”) of a larger prefix that the ISP owns to an organization (e.g., Rutgers)
- Further, the ISP announces the aggregated prefix to the Internet to save on number of forwarding table memory and number of announcements
- The organization (e.g., Rutgers) is reachable over multiple paths (e.g., through another ISP like AT&T)
- The organization has a preference to use one path over another, and expresses this by announcing the longer (more specific) prefix
- Routers in the Internet must route based on the longer prefix

IPv4 Datagram Format



How much header overhead?

- (default) 20 bytes of TCP
- (default) 20 bytes of IP
- = 40 bytes

The network layer is all about reachability. We'll see protocols that solve subproblems.

How does an endpoint get an address?

DHCP

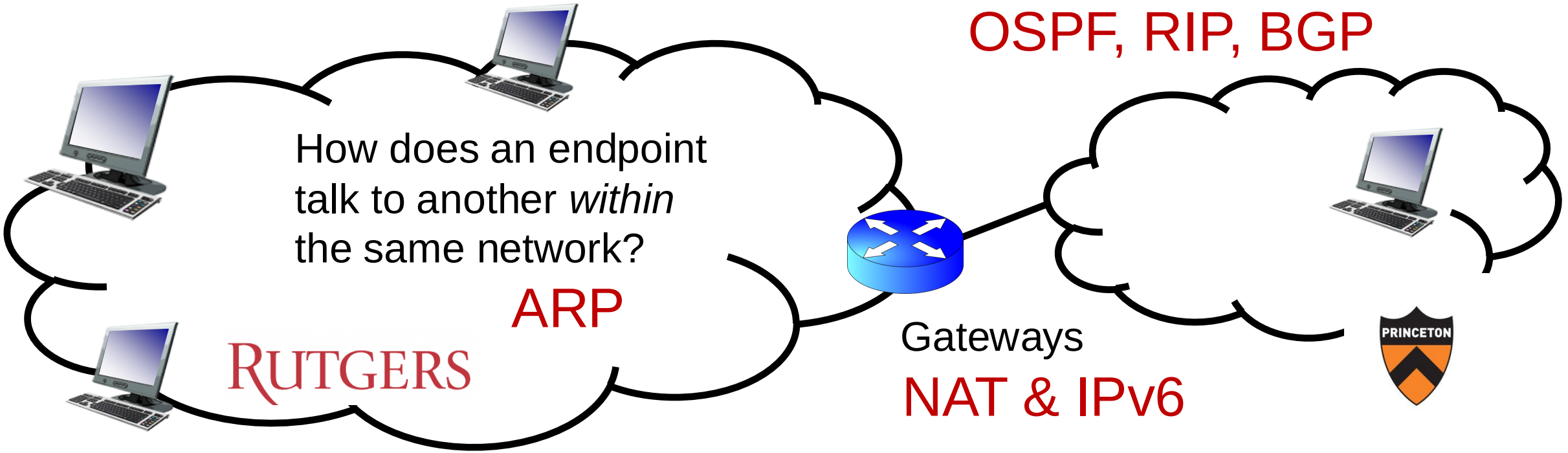
Debugging?

ICMP

How does an endpoint talk to another *outside* its network?

Routing protocols

OSPF, RIP, BGP



IP Support Protocols

- Many support protocols and mechanisms for the network layer
  - Protocols: DHCP, ICMP, ARP, IPv6, …
  - Mechanisms: NAT

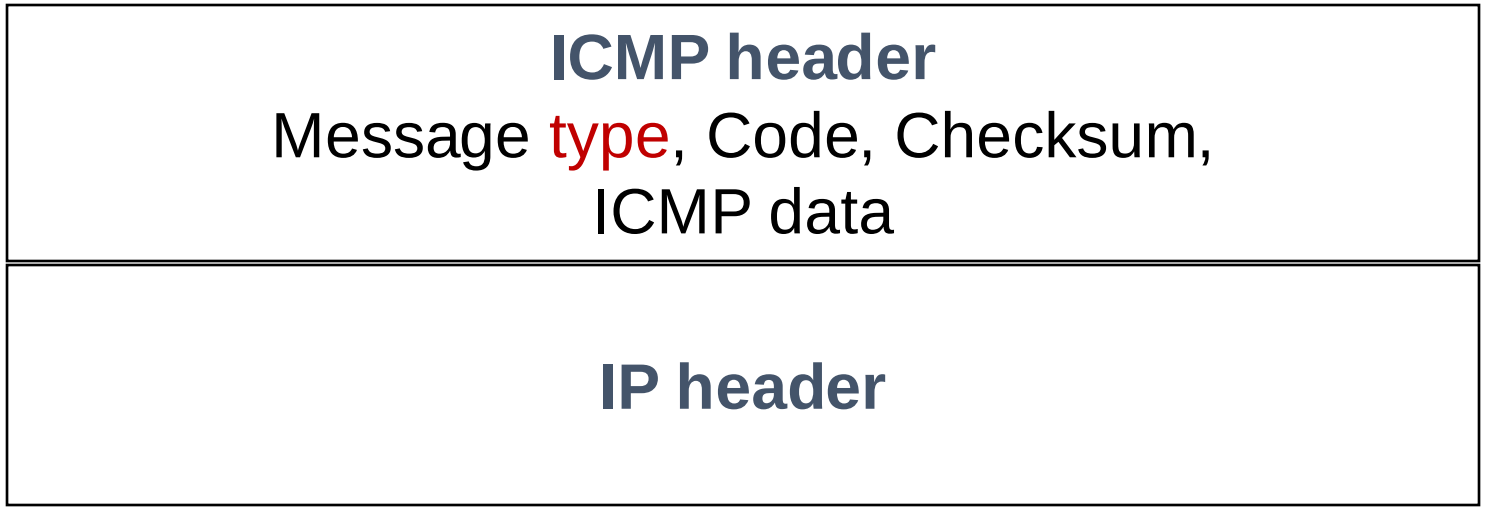
- Some of these protocols use an IP header underneath their own header (ICMP) or replace the IP header with their own (ARP)
  - But these shouldn’t be construed as transport/network protocols
  - They are fundamental to supporting IP/network layer functionality
  - More appropriately discussed as support protocols for the network layer

Internet Control Message Protocol (ICMP)

Internet Control Message Protocol

- A protocol for troubleshooting and diagnostics
- Works over IP: unreliable delivery of packets
- Some functions of ICMP:
  - Determine reachability and network errors
  - Specify that packets have been in the network for too long

ICMP message format (informal)



https://en.wikipedia.org/wiki/Internet_Control_Message_Protocol#Control_messages

Specific uses of ICMP

- Echo request reply
  - Check remotely if an endpoint is alive and connected
  - *Without* running an app remotely or controlling that endpoint
- An unreachable destination
  - Invalid address and/or port
- Knowing if packet's IP time-to-live expired
  - Example, due to routing loops

- Look at two tools built using ICMP: ping and traceroute