CS 352
Network: LPM, Protocols

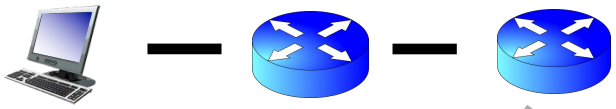
Lecture 21

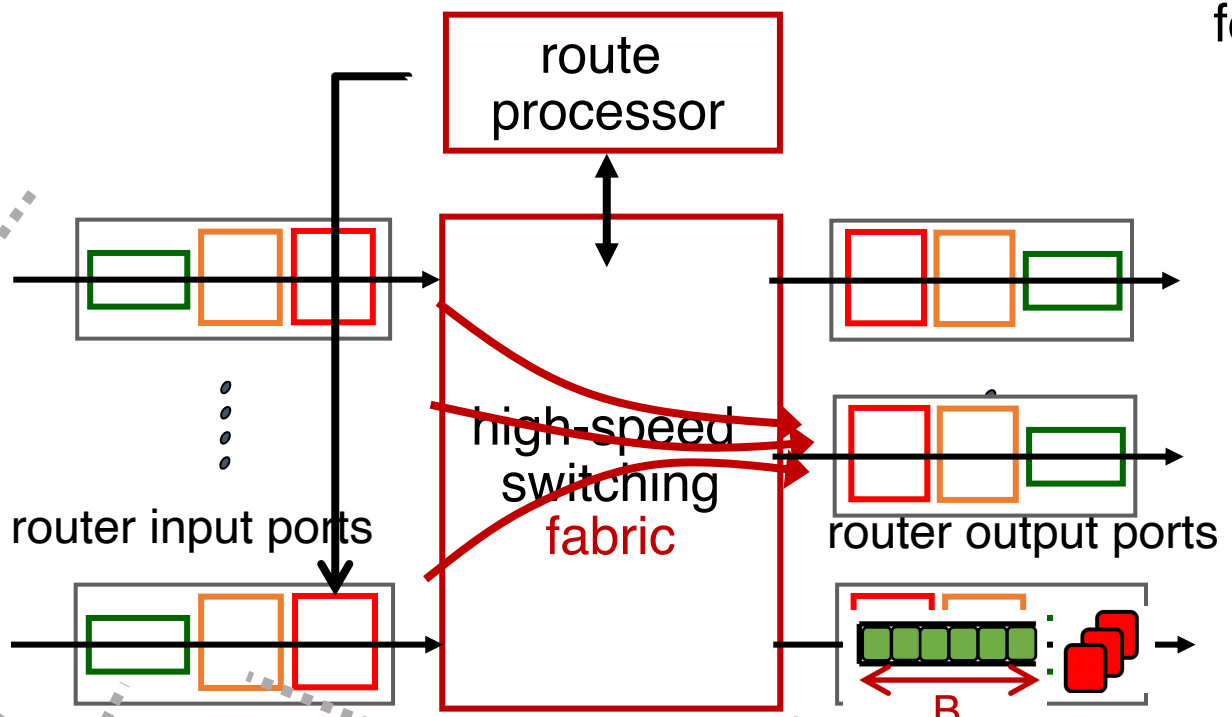
http://www.cs.rutgers.edu/~sn624/352-F22

Srinivas Narayana

RUTGERS
UNIVERSITY | NEW BRUNSWICK

Quick recap of concepts





In nonblocking fabrics, queues form only due to contention for output port.

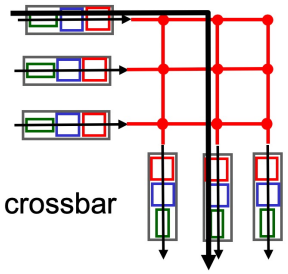
Output port contention is fundamental; unavoidable.

Scheduling and buffer management crucial.

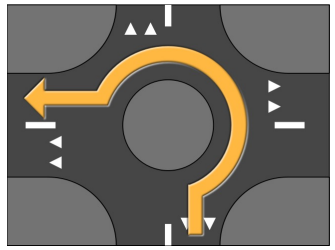
Forwarding Table

Dst-network	Port
65.0.0.0/8	3
128.9.0.0/16	1
⋮	⋮
149.12.0.0/19	7

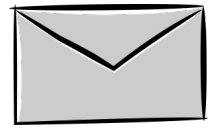
Nonblocking



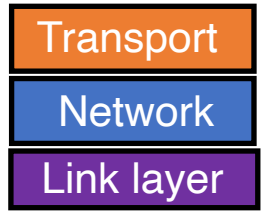
Blocking



The Internet uses destination IP based forwarding.



Parse

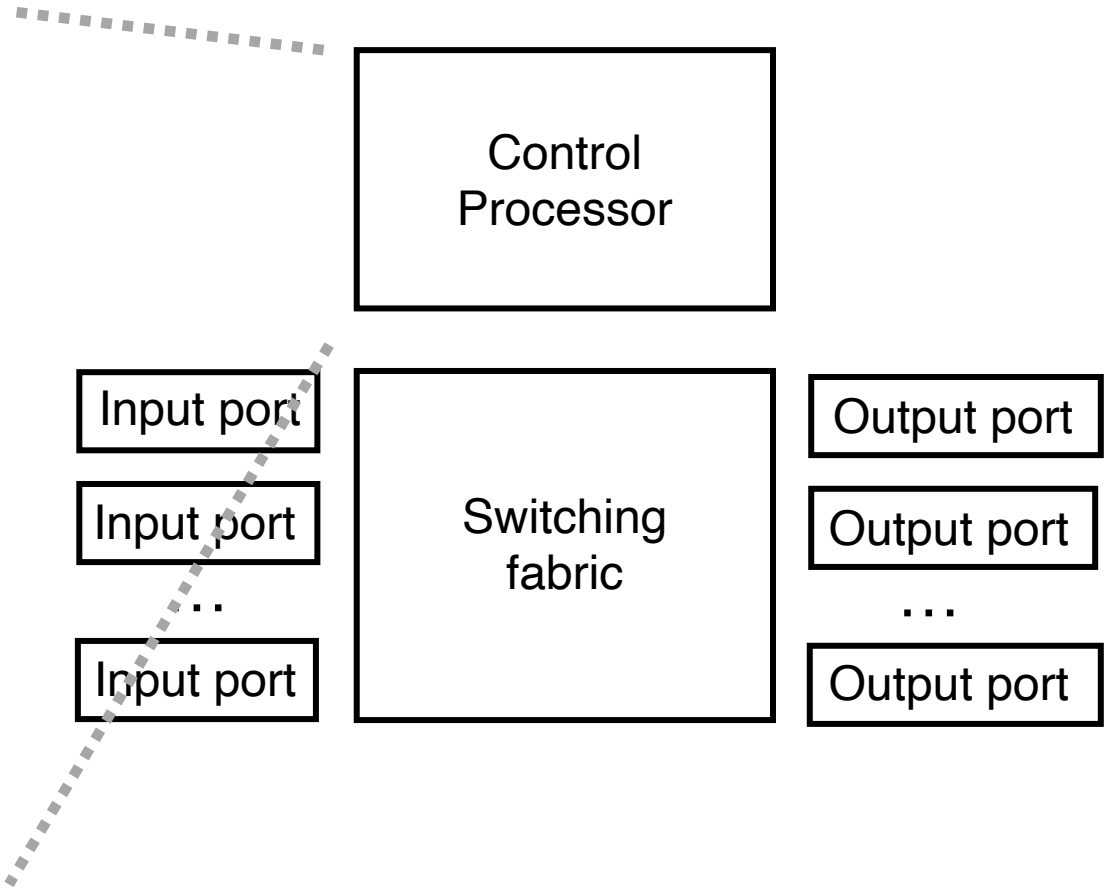


Extract destination IP address

Lookup forwarding table

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.



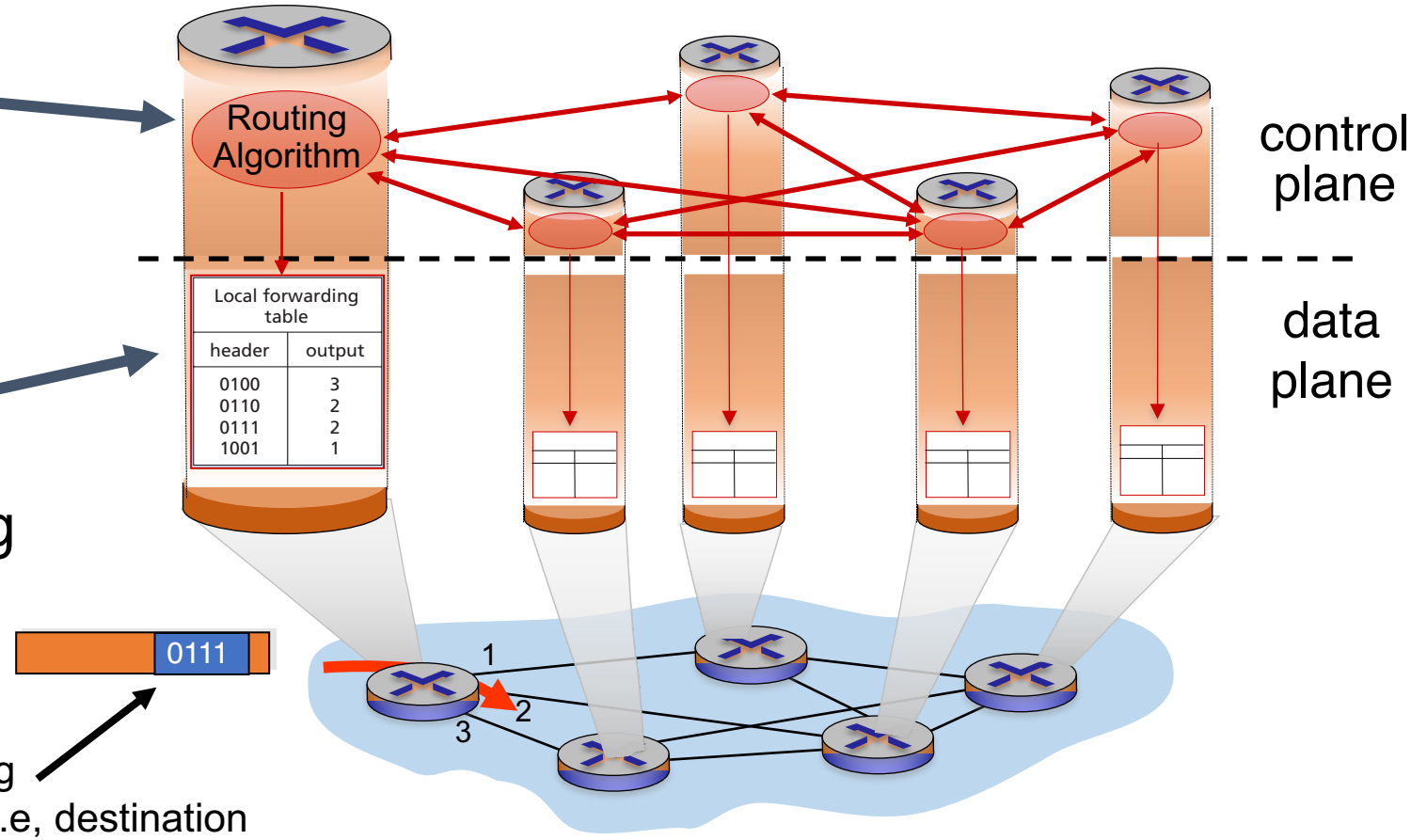
Router design: the bigger picture

Control plane

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

Data plane

per-packet processing (~ tens of nanoseconds)



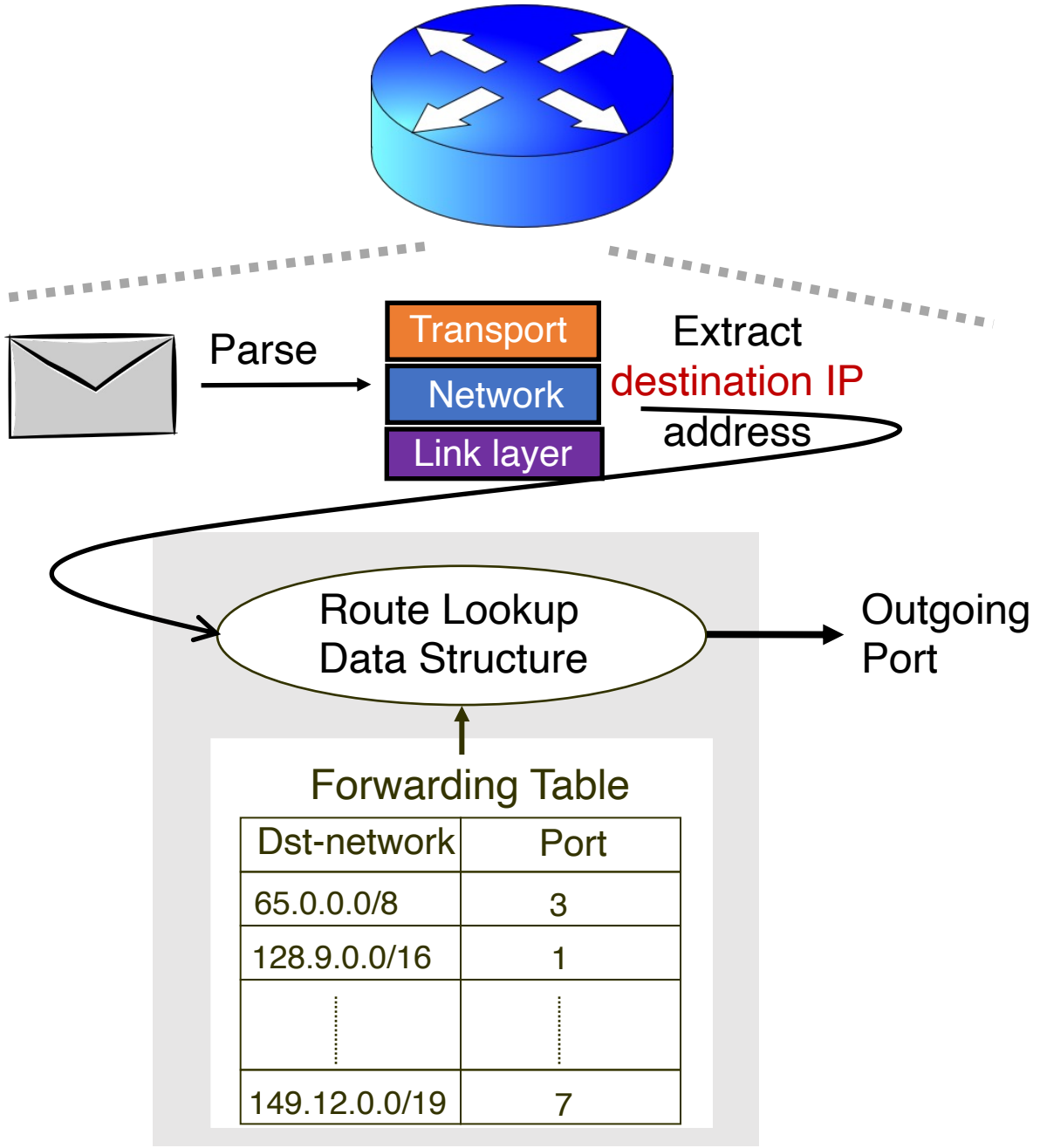
Longest Prefix Matching

Review: Route lookup

- Table lookup matches a packet against an IP prefix
  - Ex: 65.12.45.2 matches 65.0.0.0/8

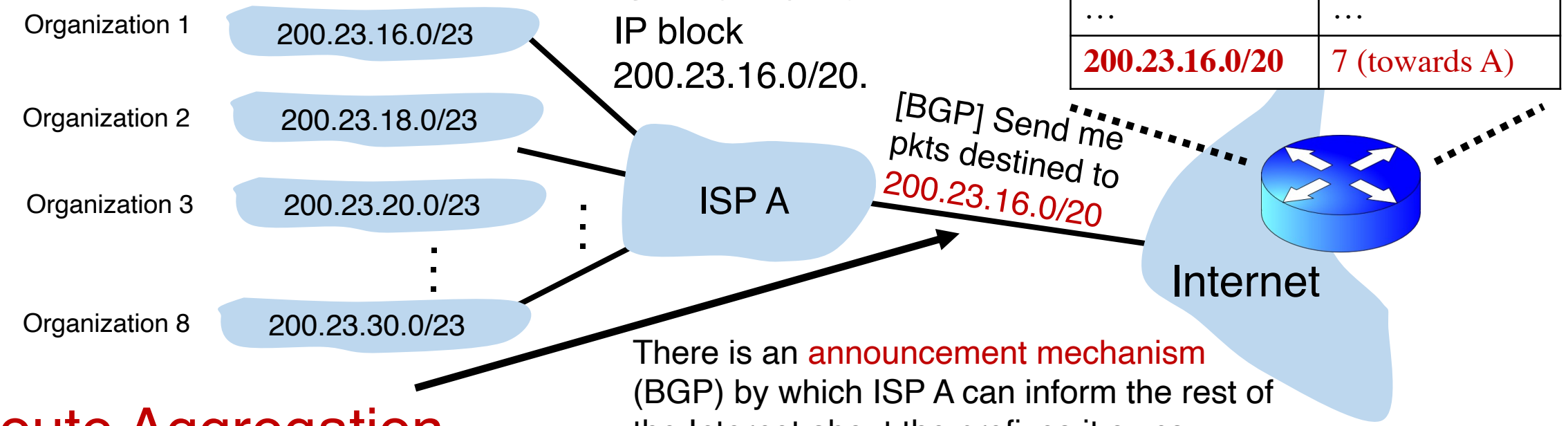
- Prefixes are allocated to organizations by Internet registries

- But organizations can reallocate a subset of their IP address allocation to other orgs



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)

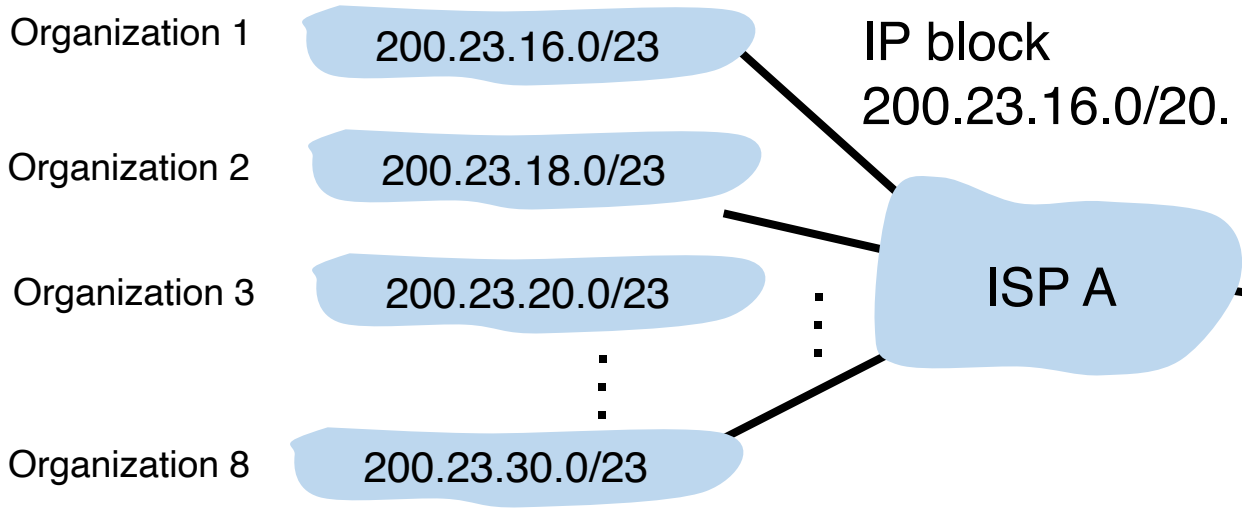
Route Aggregation

Save forwarding table memory
Fewer routing protocol msgs

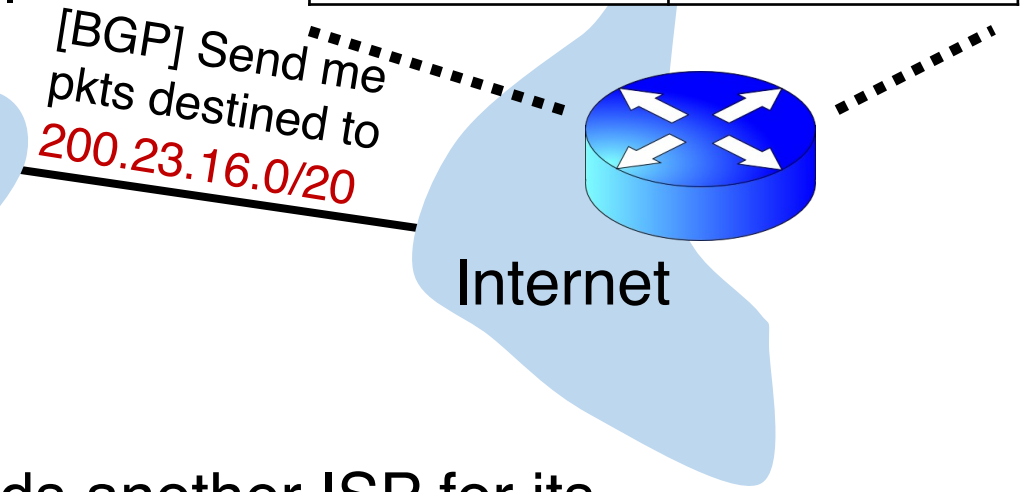
There is an announcement mechanism (BGP) by which ISP A can inform the rest of the Internet about the prefixes it owns.
It is enough to announce a coarse-grained prefix 200.23.16.0/20 rather than 8 separate sub-prefixes.

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)



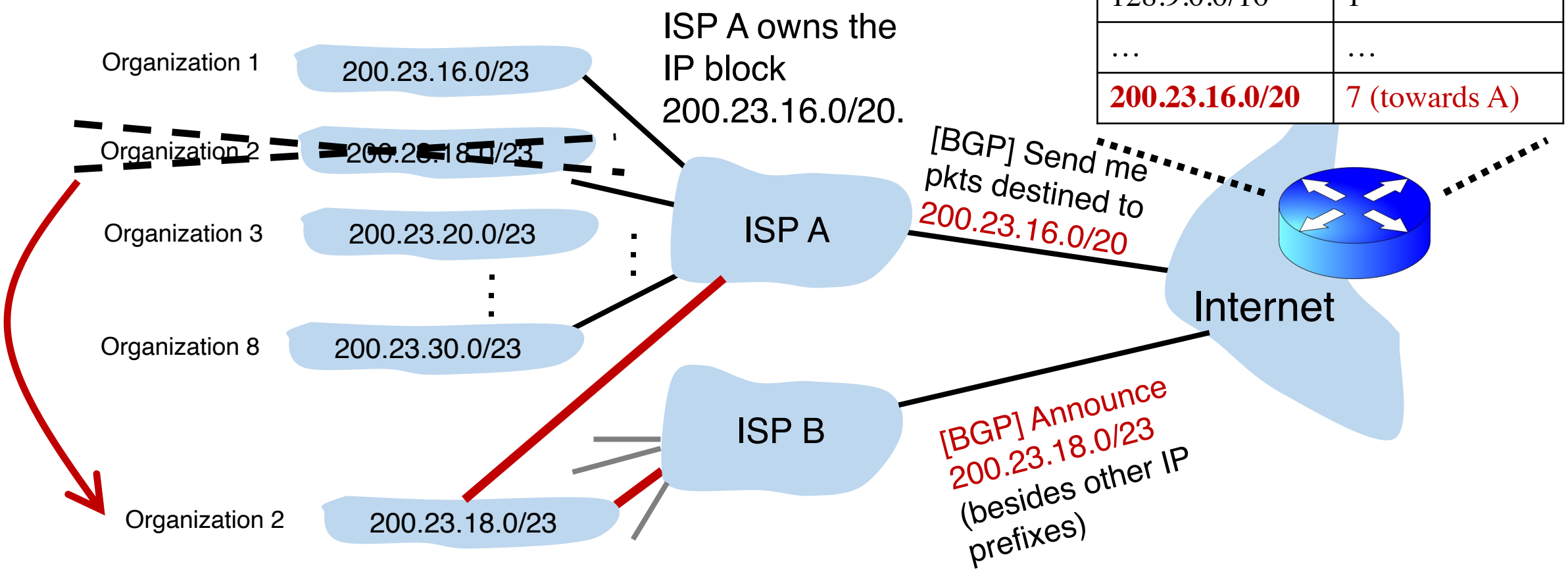
Now suppose one of these organizations adds another ISP for its Internet service and prefers using the new ISP.

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

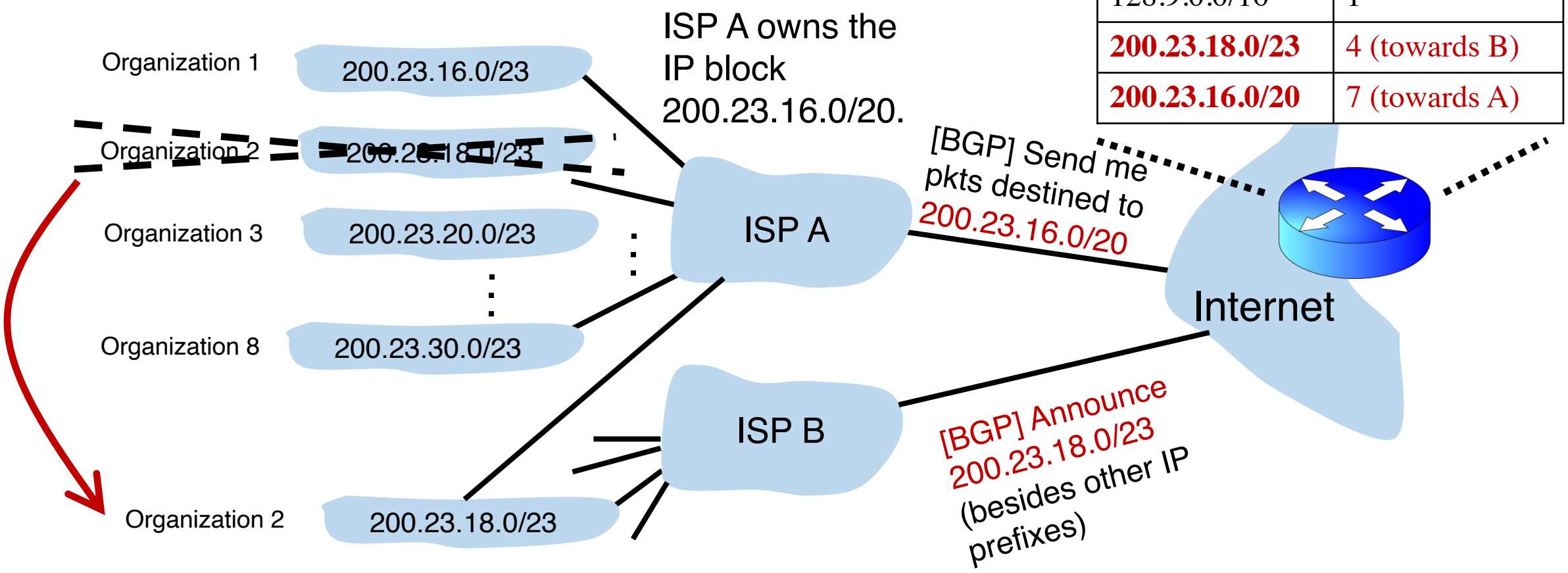
Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
…	…
200.23.16.0/20	7 (towards A)



Example of IP block reallocation

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

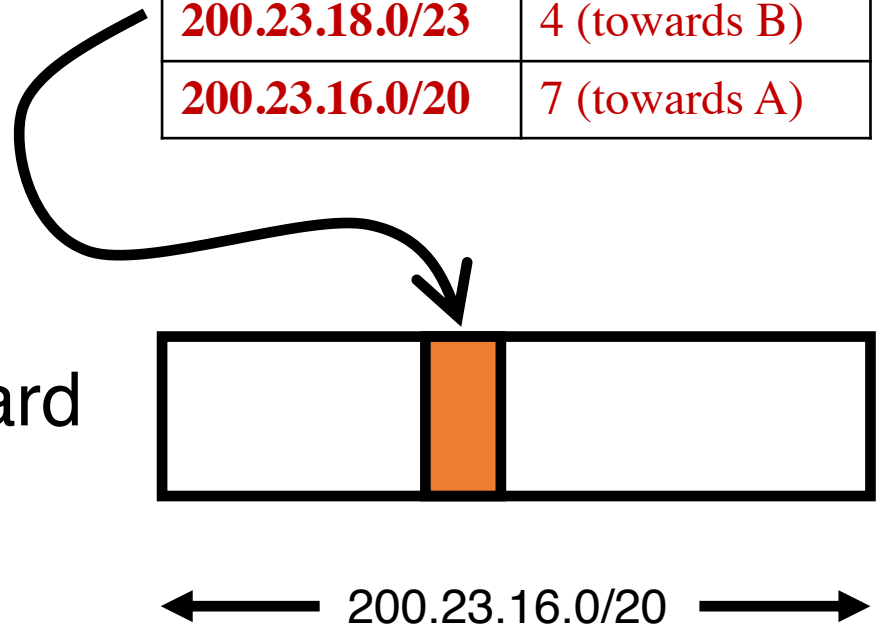
Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
200.23.18.0/23	4 (towards B)
200.23.16.0/20	7 (towards A)



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?
  - 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)

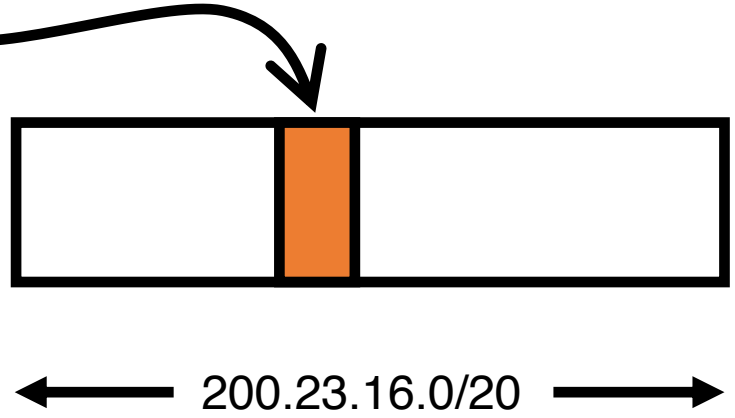


The Internet uses a policy to prioritize: Longest Prefix Matching

Longest Prefix Matching (LPM)

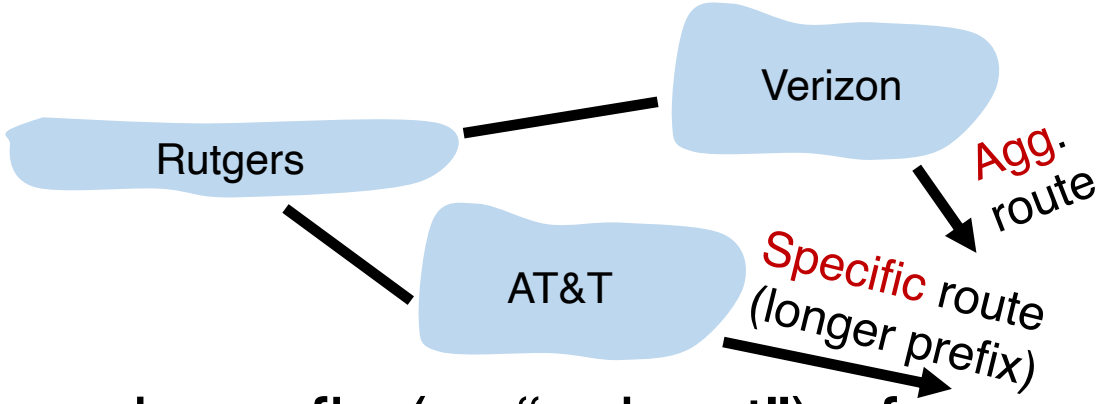
- Use the longest matching prefix, i.e., the most specific route, among all prefixes that match the packet.
- Policy borne out of the Internet's IP allocation model: prefixes and sub-prefixes are handed out
- Internet routers use longest prefix matching.
  - Very interesting algorithmic problems
  - Challenges in designing efficient software and hardware data structures

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)



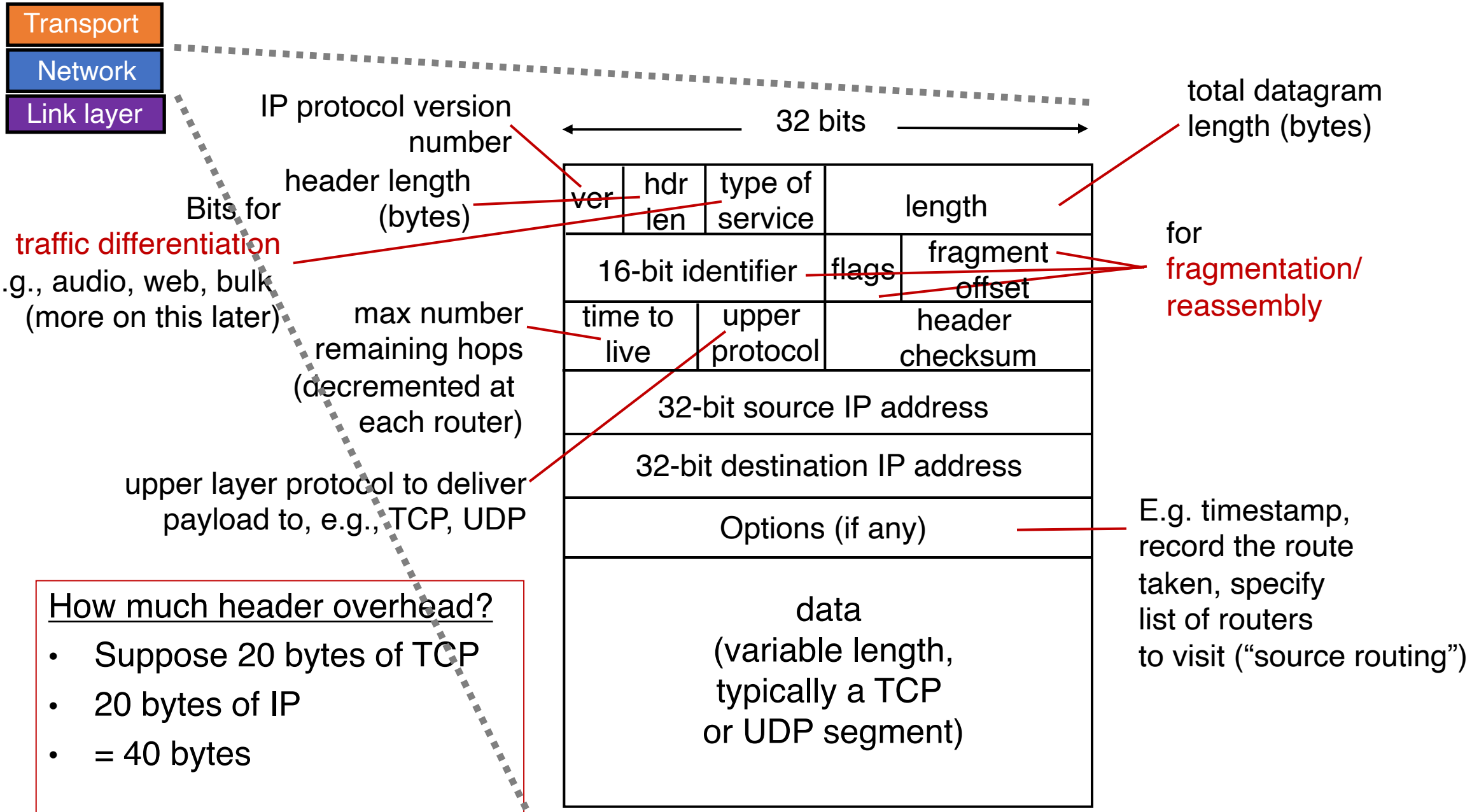
Internet routers perform longest-prefix matching on destination IP addresses of packets.

Why is LPM prevalent?



- 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?

- Suppose 20 bytes of TCP
- 20 bytes of IP
- = 40 bytes

Rest of this lecture

- Discuss 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

The network layer is all about reachability. Every protocol we'll see solves a sub-problem.

How does an endpoint get an address?

DHCP

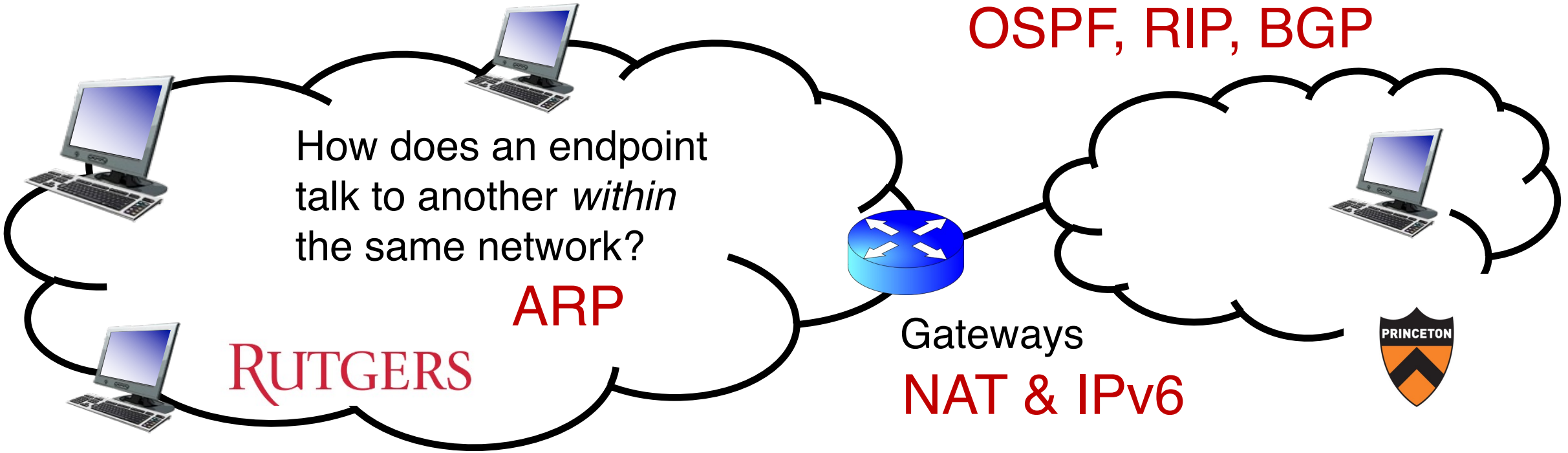
Debugging?

ICMP

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

Routing protocols

OSPF, RIP, BGP

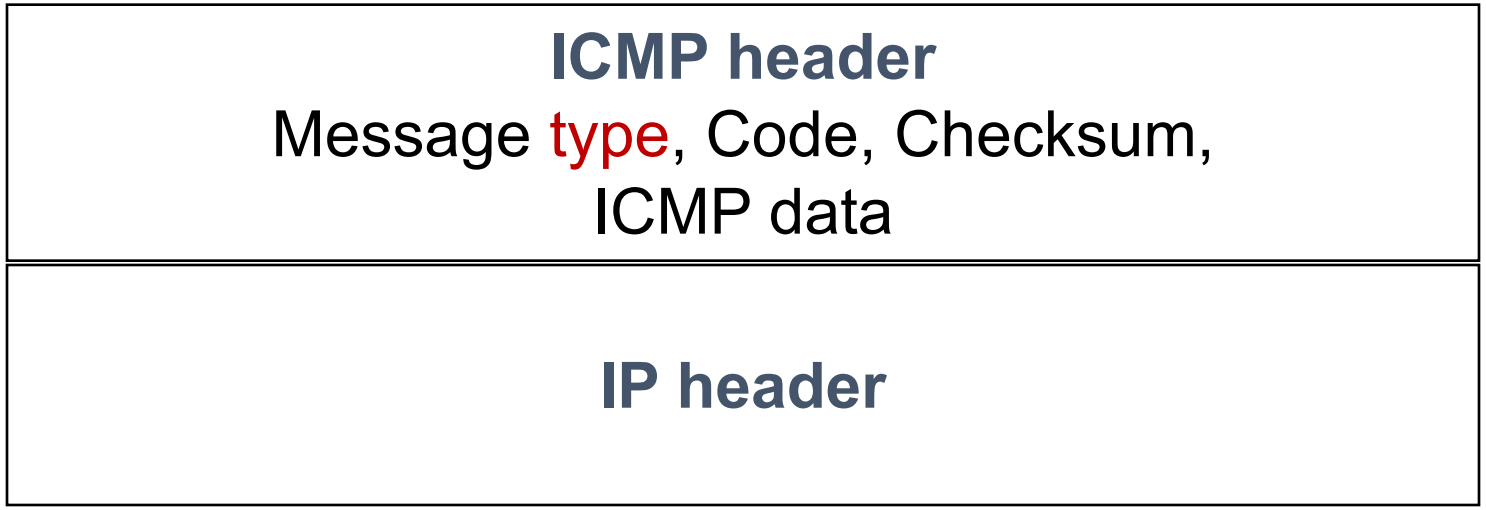


Internet Control Message Protocol (ICMP)

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

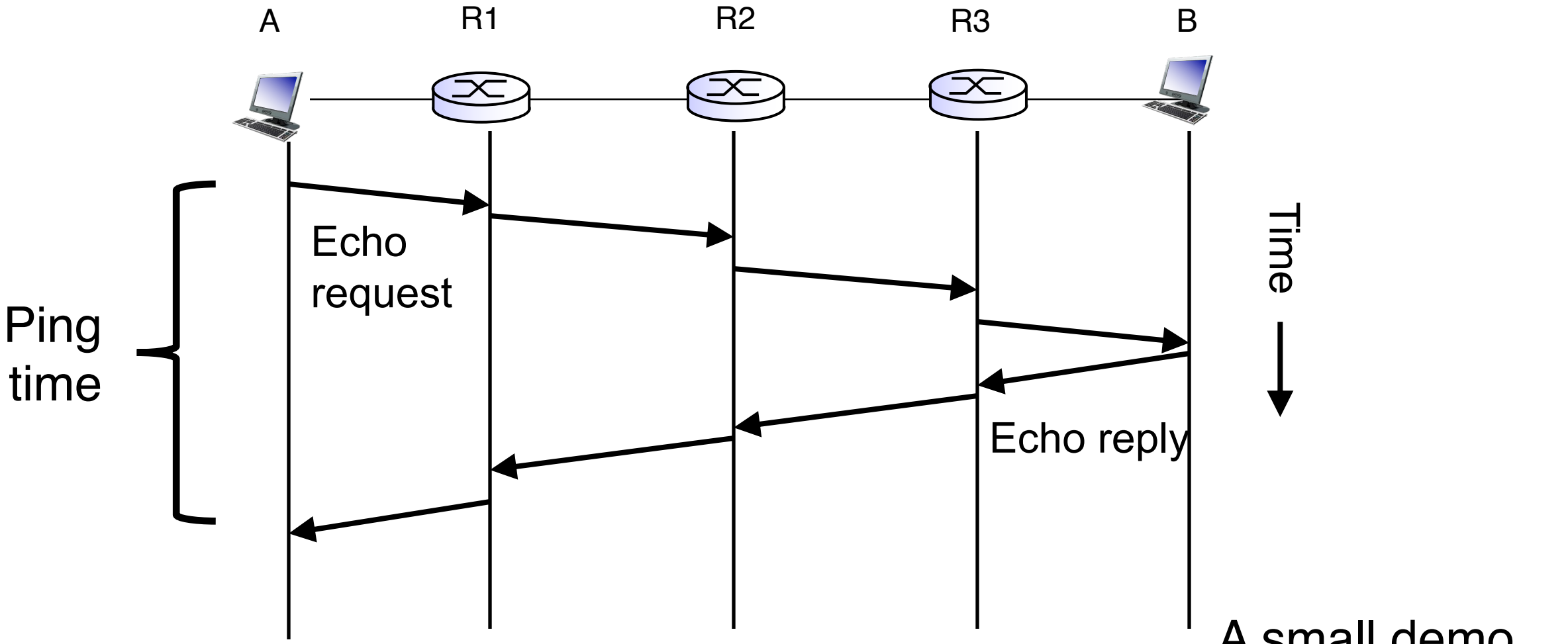
- Echo request reply
  - Check remotely if an endpoint is alive and connected
  - *Without* running an app remotely or controlling that endpoint
- An unreachable destination
  - Invalid address and/or port
- Knowing if packet's IP time-to-live expired
  - Example, due to routing loops

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

Ping

- Uses ICMP echo request (type=8, code=0) and reply (type=0, code=0)
- Source sends ICMP echo request message to dst address
- Destination network stack replies with an ICMP echo reply message
- Source can calculate round trip time (RTT) of packets
- If no echo reply comes back, then the destination is unreachable
- Don’t need to have a server program running on the other side
  - In general, the remote endpoint can be completely outside your control

Ping



A small demo…

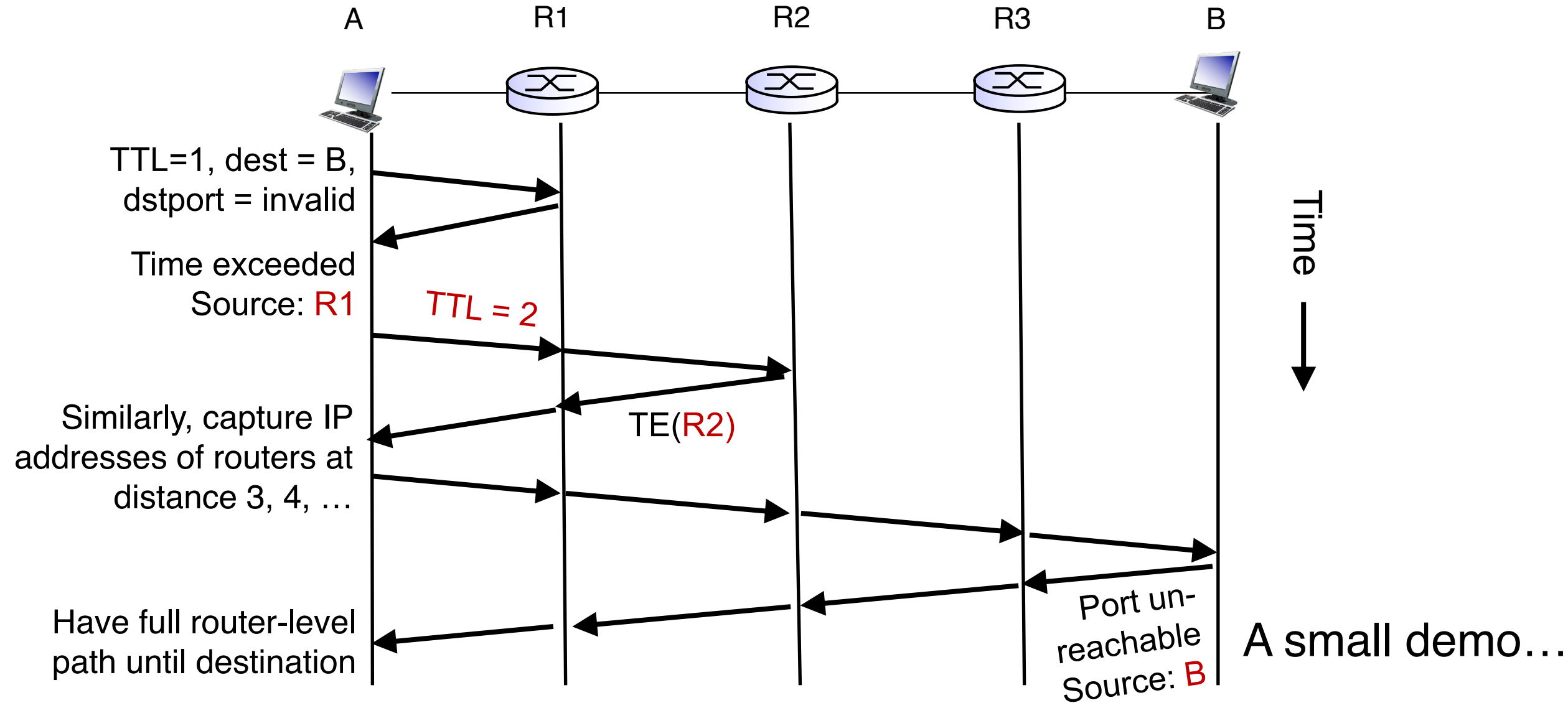
Traceroute

- A tool that can record the router-level path taken by packets
- A clever use of the IP time-to-live (TTL) field
- In general, when a router receives an IP packet, it decrements the TTL field on the packet
  - A failsafe mechanism to ensure packets don't keep taking up network resources for too long
- If a router receives a packet with TTL=0, it sends an ICMP time exceeded message (type=11, code=0) to the source endpoint

Traceroute

- Traceroute sends multiple packets to a destination endpoint
- But it progressively increases the TTL on those packets: 1, 2, ...
- Every time a time exceeded message is received, record the router's IP address
- Process repeated until the destination endpoint is reached
- If the packet reaches the destination endpoint (i.e.: TTL is high enough), then the endpoint sends a port unreachable message (type=3, code=3)

Traceroute



A small demo…

Summary of ICMP

- A protocol for network diagnostics and troubleshooting
- Two useful tools: ping and traceroute
- Ping: test connectivity to a machine totally outside your control
  - Use ICMP echo request and reply
- Traceroute: determine router-level path to a remote endpoint
  - A smart use of the TTL field in the IP header

Network Address Translation (NAT)

Background: The Internet's growing pains

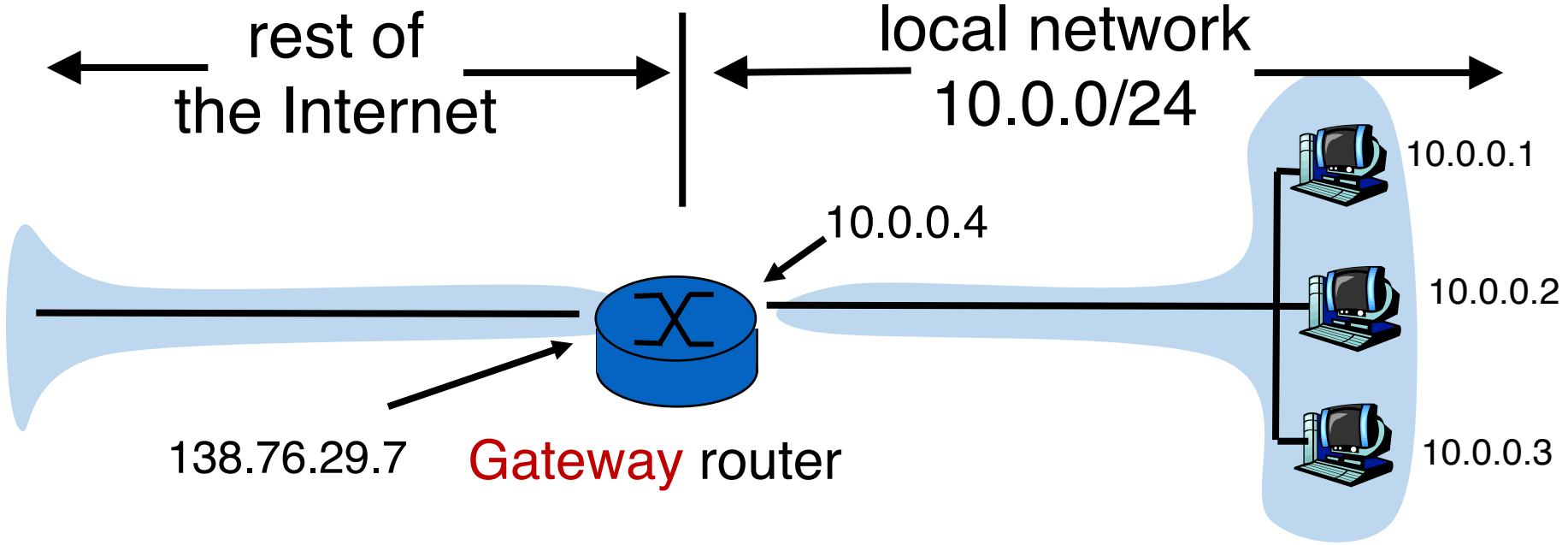
- Networks had incompatible addressing
  - IPv4 versus other network-layer protocols (X.25)
  - Routable address ranges different across networks
- Entire networks were changing their Internet Service Providers
  - ISPs don't want to route directly to internal endpoints, just to the gateway
- IPv4 address exhaustion
  - Insufficient large IP blocks even for large networks
  - Rutgers (AS46) has > 130,000 publicly routable IP addresses
  - IIT Madras (a well-known public university in India, AS141340) has 512

(Source: ipinfo.io)

Network Address Translation

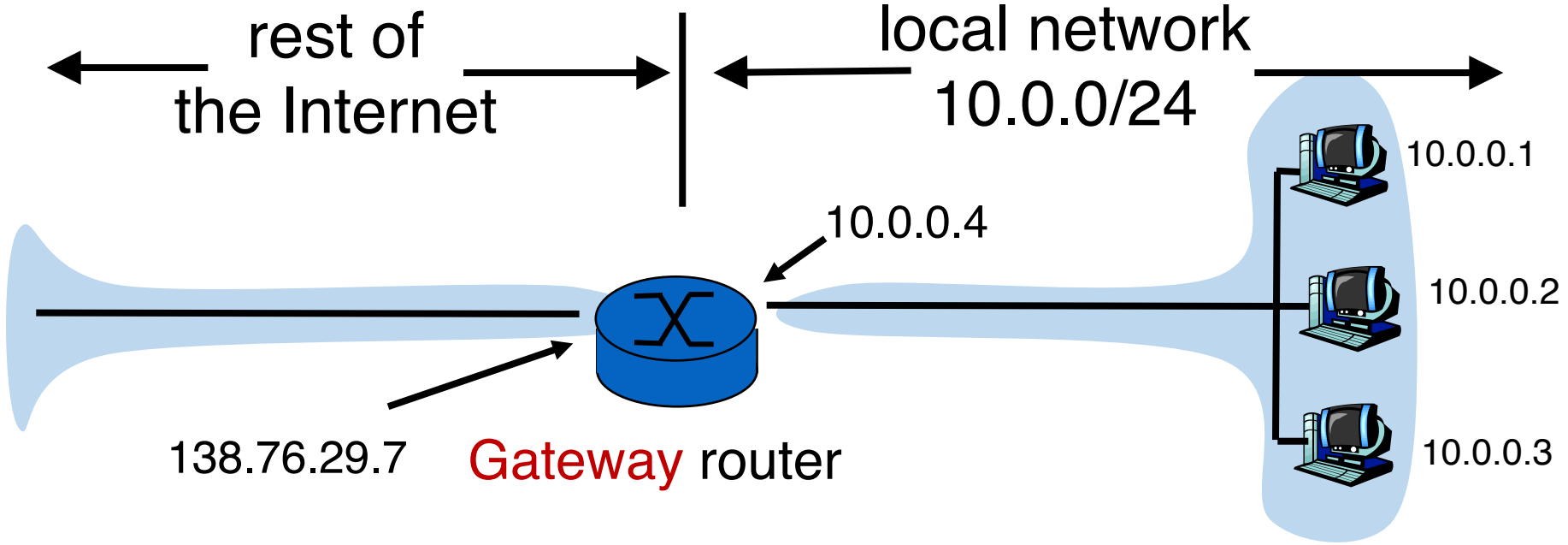
- When a router modifies fields in an IP packet to:
- Enable communication across networks with different (network-layer) addressing formats and address ranges
- Allow a network to change its connectivity to the Internet en masse by modifying the source IP to a (publicly-visible) gateway IP address
- Masquerade as an entire network of endpoints using (say) one publicly visible IP address
  - Effect: use fewer IP addresses for more endpoints!
- We'll see a standard design: "Network address and port translation" (NAPT)

Typical NAT setup (NAPT)



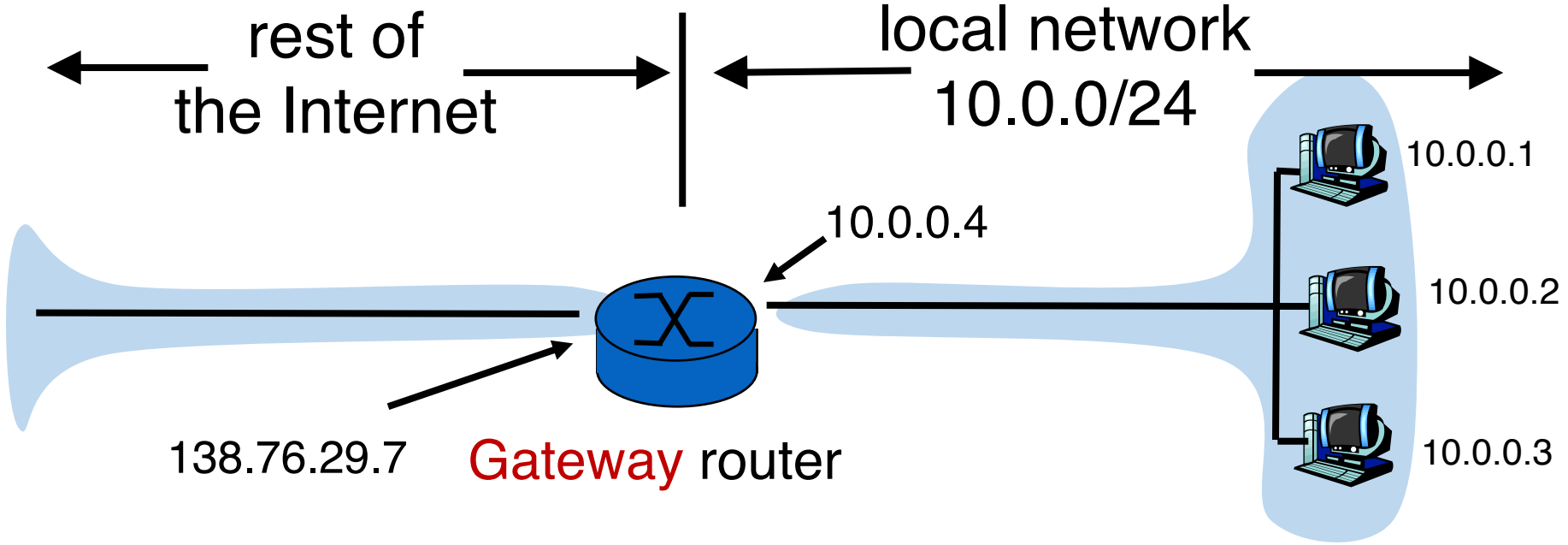
- The gateway's IP, 138.76.29.7 is publicly visible
- The local endpoint IP addresses in 10.0.0/24 are private
- All datagrams leaving local network have the same source IP as the gateway

Typical NAT setup (NAPT)



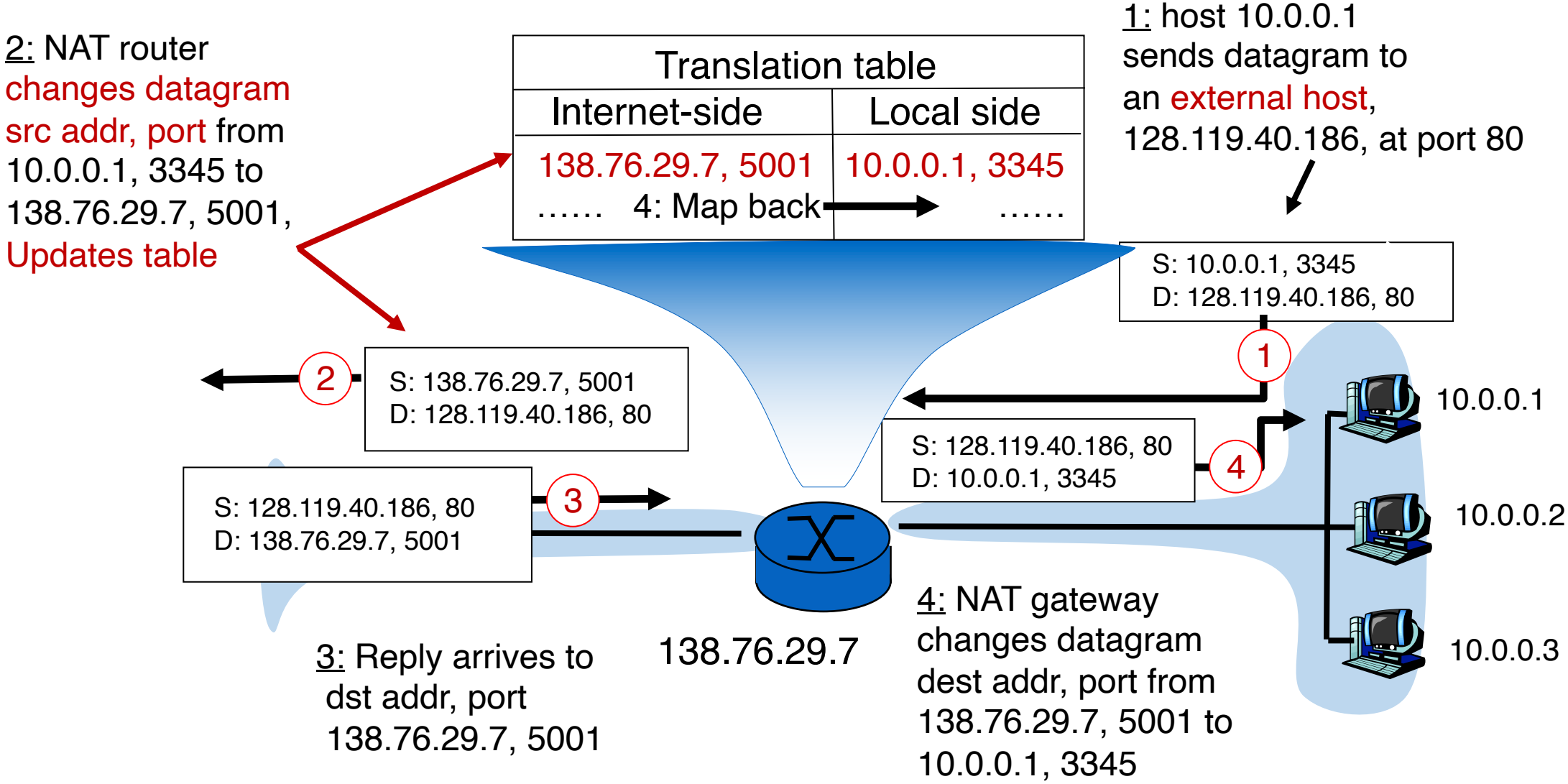
That is, for the rest of the Internet, the gateway masquerades as a single endpoint representing (hiding) all the private endpoints. The entire network just needs one (or a few) public IP addresses.

Typical NAT setup (NAPT)



The NAT gateway router accomplishes this by using a different transport port for each distinct (transport-level) conversation between the local network and the Internet.

Typical NAT setup (NAPT)



Features of IP-masquerading NAT

- Use one or a few public IPs: You don’t need a lot of addresses from your ISP
- Change addresses of devices inside the local network freely, without notifying the rest of the Internet
- Change the public IP address freely independent of network-local endpoints
- Devices inside the local network are not publicly visible, routable, or accessible
- Most IP masquerading NATs block incoming connections originating from the Internet
  - Only way to communicate is if the internal host initiates the conversation

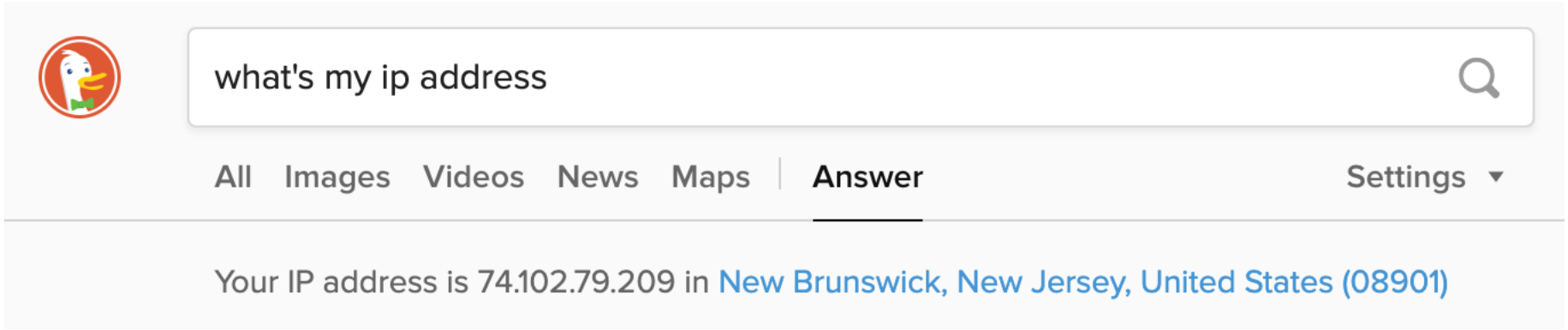
If you're home, you're likely behind NAT

- Most access routers (e.g., your home WiFi router) implement network address translation

- You can check this by comparing your local address (visible from `ifconfig`) and your externally-visible IP address (e.g., type "what's my IP address?" on your browser search bar)

If you're home, you're likely behind NAT

```
[flow:352-S20]$  ifconfig en0
en0: flags=8863<UP,BROADCAST,SMART,RUNNING,SIMPLEX,MULTICAST> mtu 1500
        ether f0:18:98:1c:fc:36
        inet6 fe80::1036:7dea:82ee:e868%en0 prefixlen 64 secured scopeid 0xa
        inet 192.168.1.151 netmask 0xffffff00 broadcast 192.168.1.255
        nd6 options=201<PERFORMNUD,DAD>
        media: autoselect
        status: active
[flow:352-S20]$
```



Limitations of IP-masquerading NATs

- Connection limit due to 16-bit port-number field
  - ~64K total simultaneous connections with a single public IP address
- NAT can be controversial
  - "Routers should only manipulate headers up to the network layer, not modify headers at the transport layer!"
- Application developers must take NAT into account
  - e.g., peer-to-peer applications like Skype
- Internet "purists": instead, solve address shortage with IPv6
  - 32-bit IP addresses are just not enough
  - Esp. with more devices (your watch, your fridge, …) coming online