CS 352
Router Design

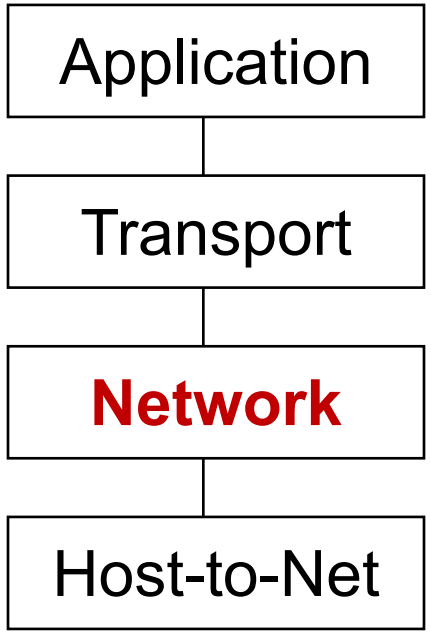
CS 352, Lecture 15.1

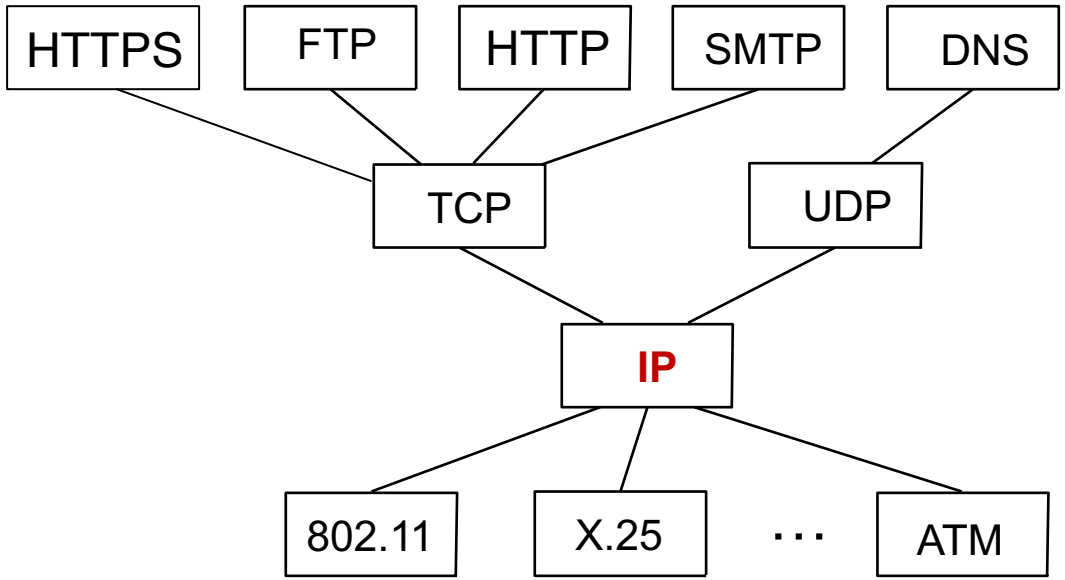
http://www.cs.rutgers.edu/~sn624/352

Srinivas Narayana

RUTGERS
UNIVERSITY | NEW BRUNSWICK

Network



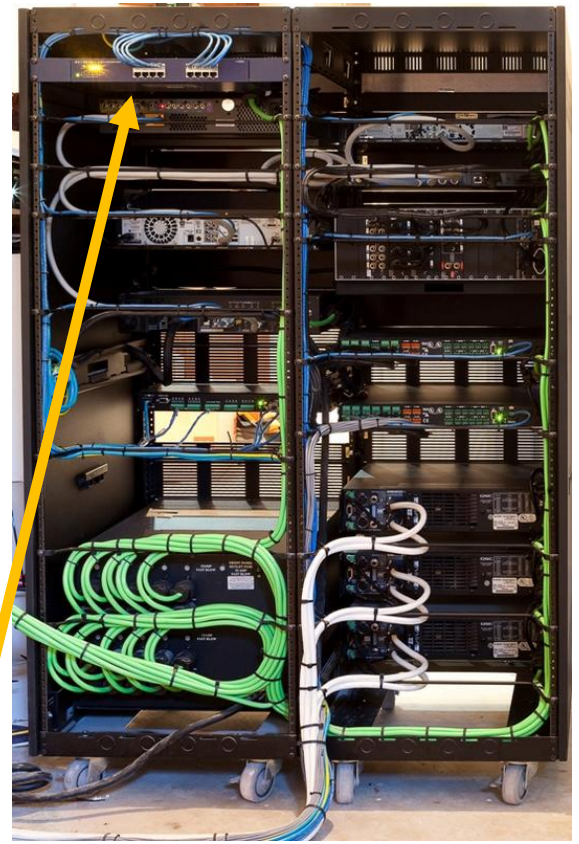


The main function of the network layer is to
move packets from one endpoint to another.
The network will make its best effort to deliver packets
but doesn't guarantee anything.

This lecture is about routers

Access routers

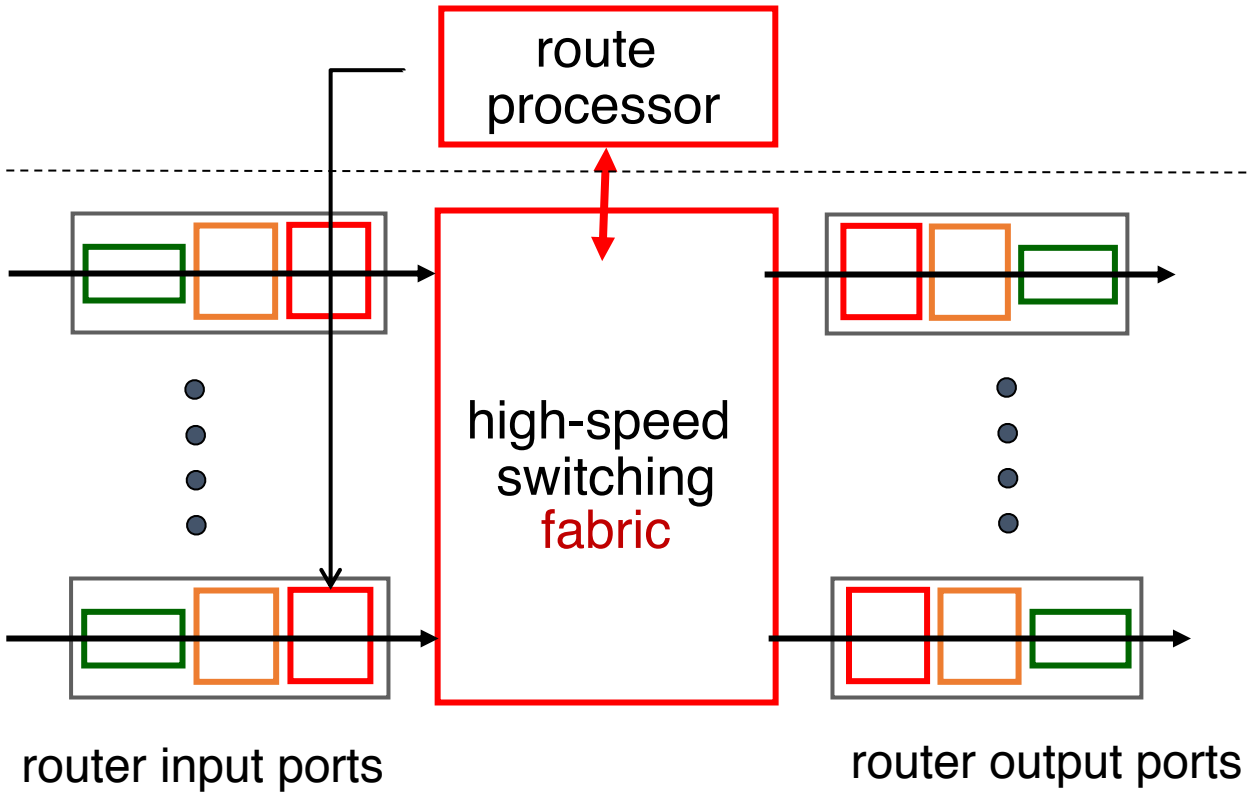
Core router



Data center top-of-rack switch

What’s inside a router?

Router architecture overview



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

Control plane

Data plane

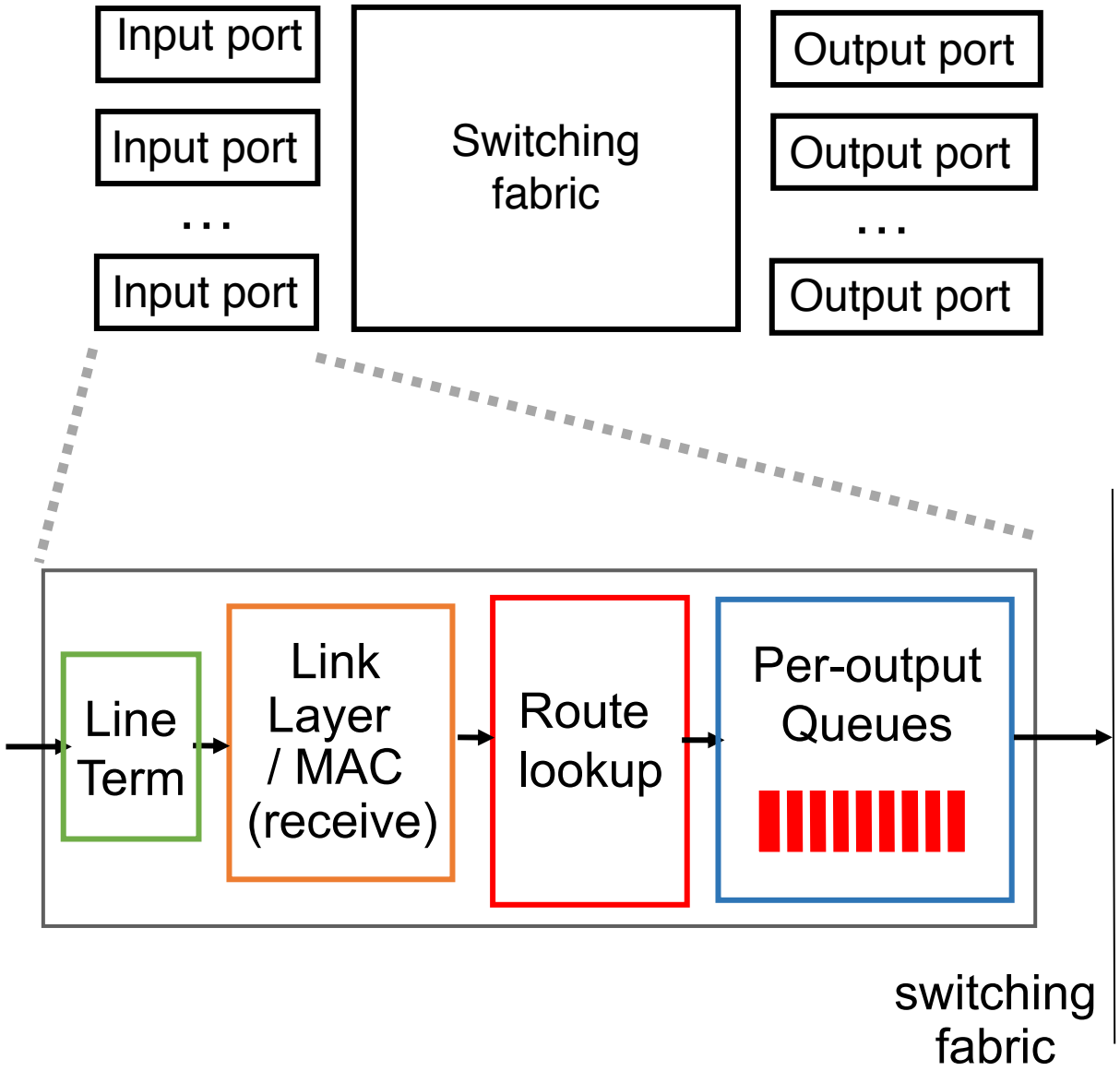
Review: Forwarding function: move packets from one input port to another.

Different and evolving designs

- There are different kinds of routers, with their own designs
  - Access routers (e.g., home WiFi), chassis/core routers, top-of-rack switches
- Router designs have also evolved significantly over time
- For simplicity and concreteness, we will learn about one high-speed router design from the early 2000s.
- Called the MGR (multi-gigabit router). It could support an aggregate rate of 50 Gbit/s (1 G = 10^9)
  - Today's single-chip routers can support aggregate rates of ~10 Tbit/s (1 T = 10^{12})

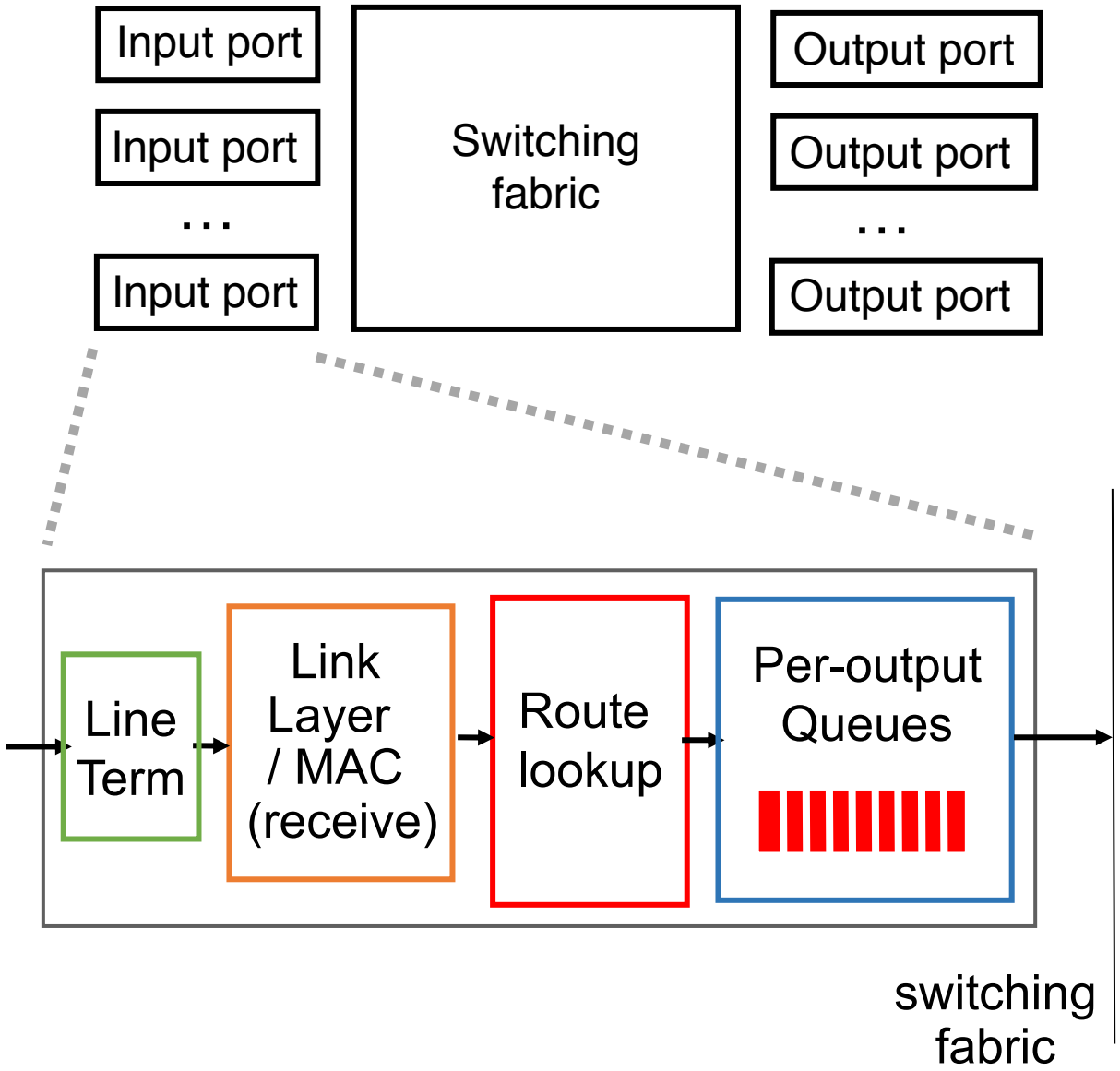
Input port functions

- Line termination: receives physical (analog) signals and turns them into digital signals.
- Speed of data on a single network interface termed line speed or line rate (e.g., 100 Mbit/s)
- Link layer: performs medium access control functions (e.g., Ethernet): more on this much later.



Input port functions

- Route lookup: high-speed lookup of which output port the packet is destined to
- Goal: must complete this processing at the line rate
- Queueing: packets may wait in per-output-port queues if packets are arriving too fast for the switching fabric to send them to the output port

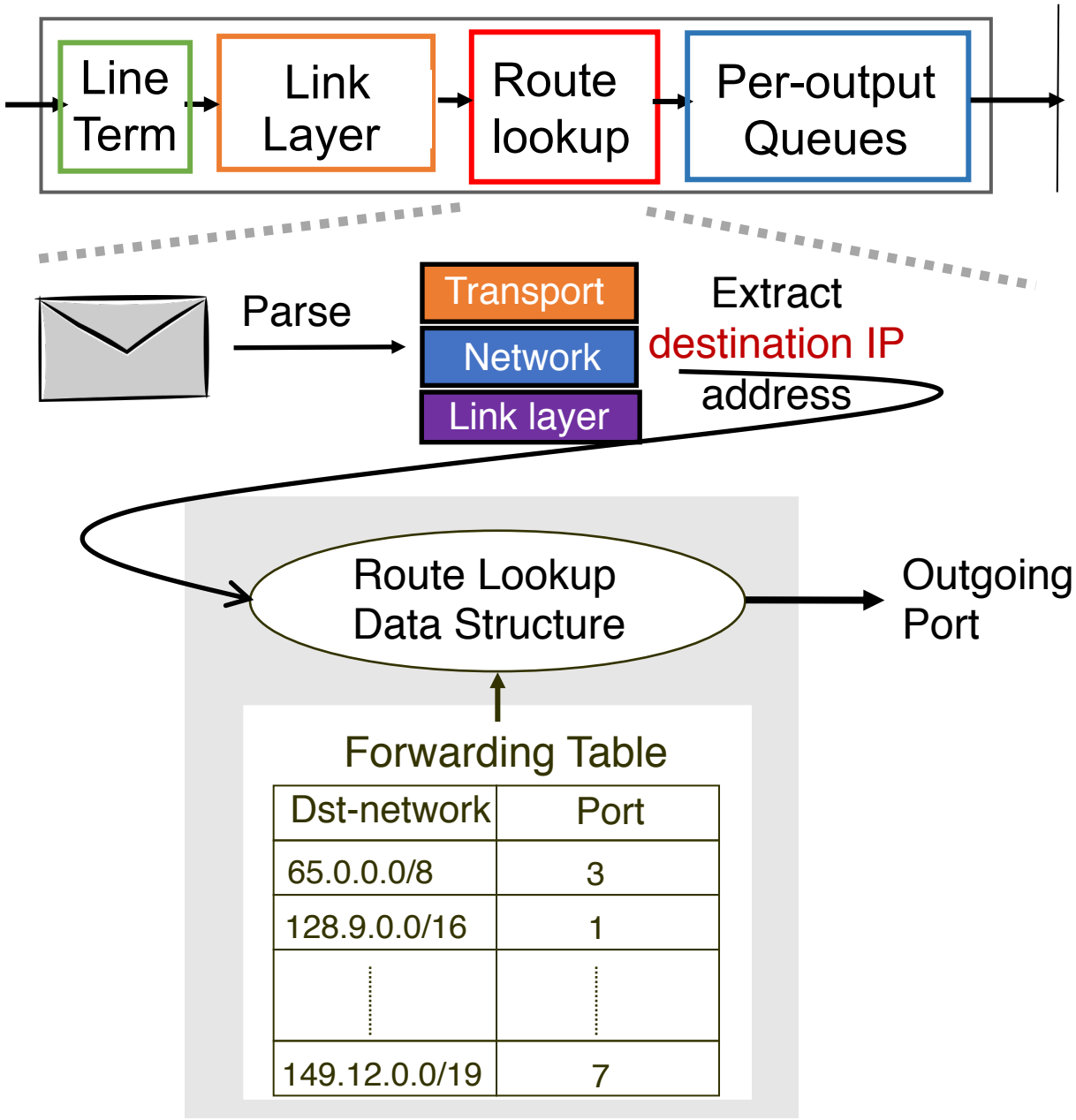


Route lookups

Packet forwarding in the Internet is based on the destination IP address on the packet.

Example: if dst IP on packet is 65.45.145.34, it matches the forwarding table prefix 65.0.0.0/8.

The packet is forwarded out port 3.

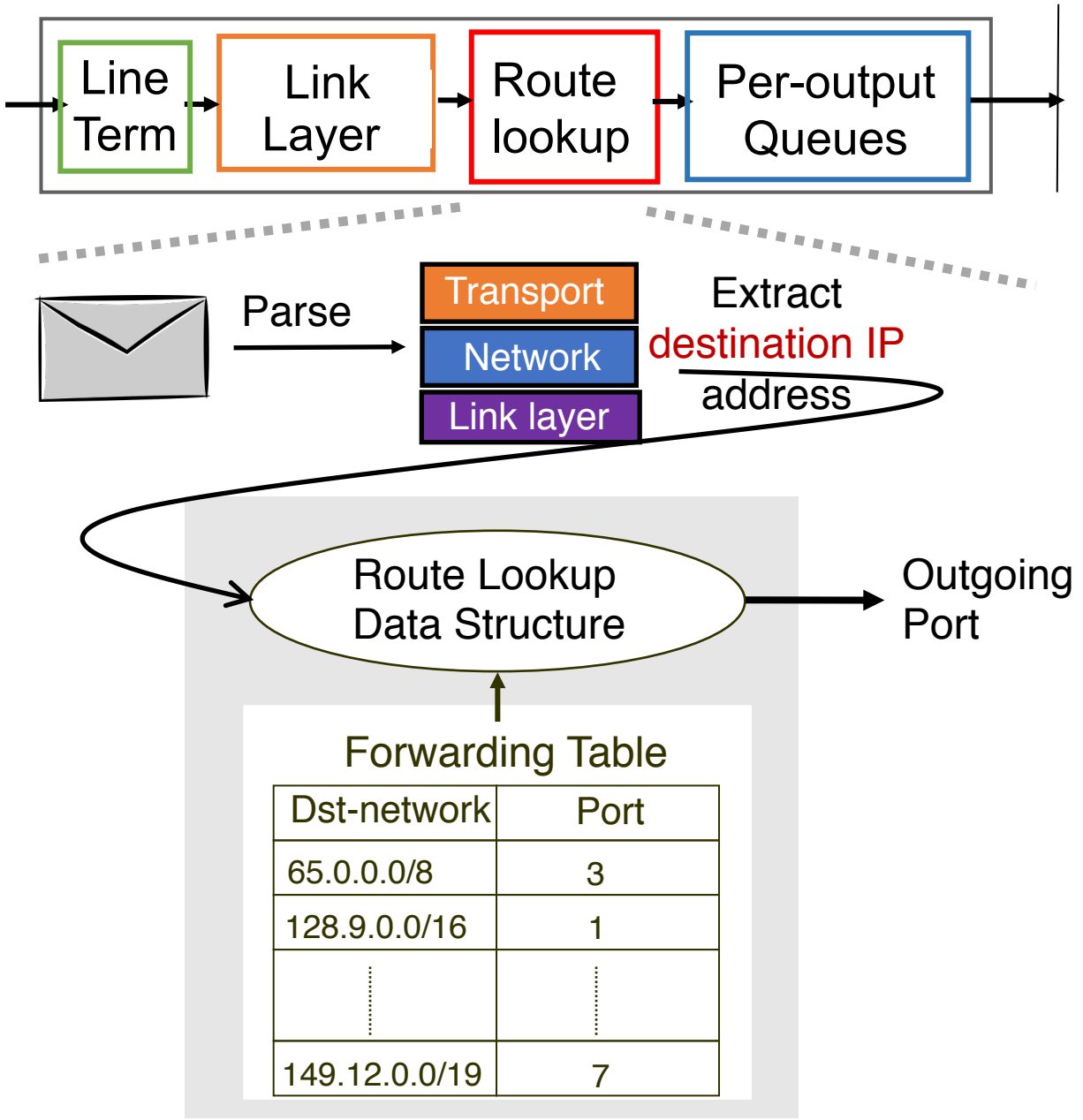


Route lookups

Number of entries in the forwarding table matters.

Fitting into router memory

Designing hardware and software for fast lookups

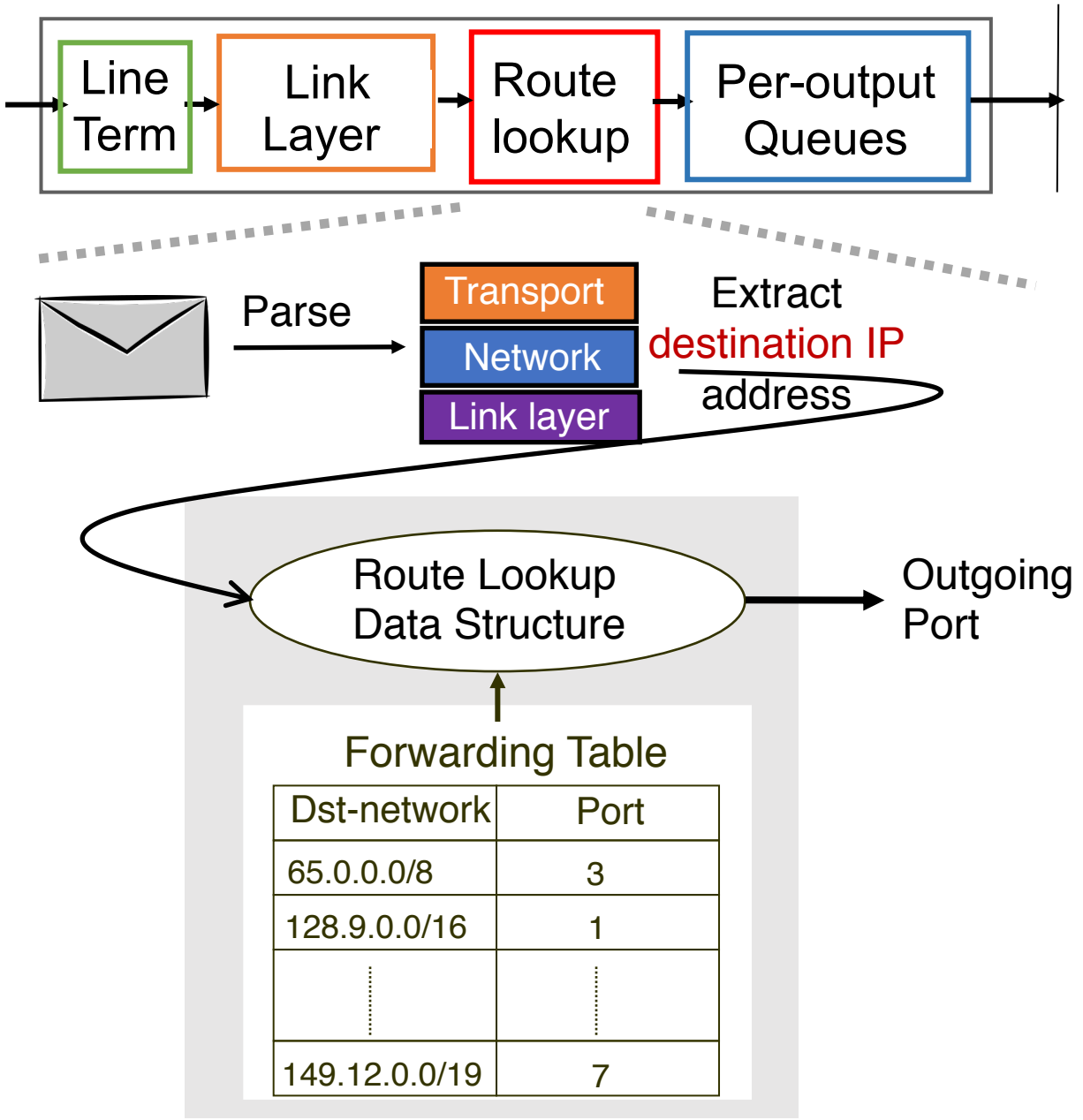


Route lookups

Recall: IP addresses can be aggregated based on shared prefixes.

The number of table entries in a router is proportional to the number of prefixes, NOT the number of endpoints.

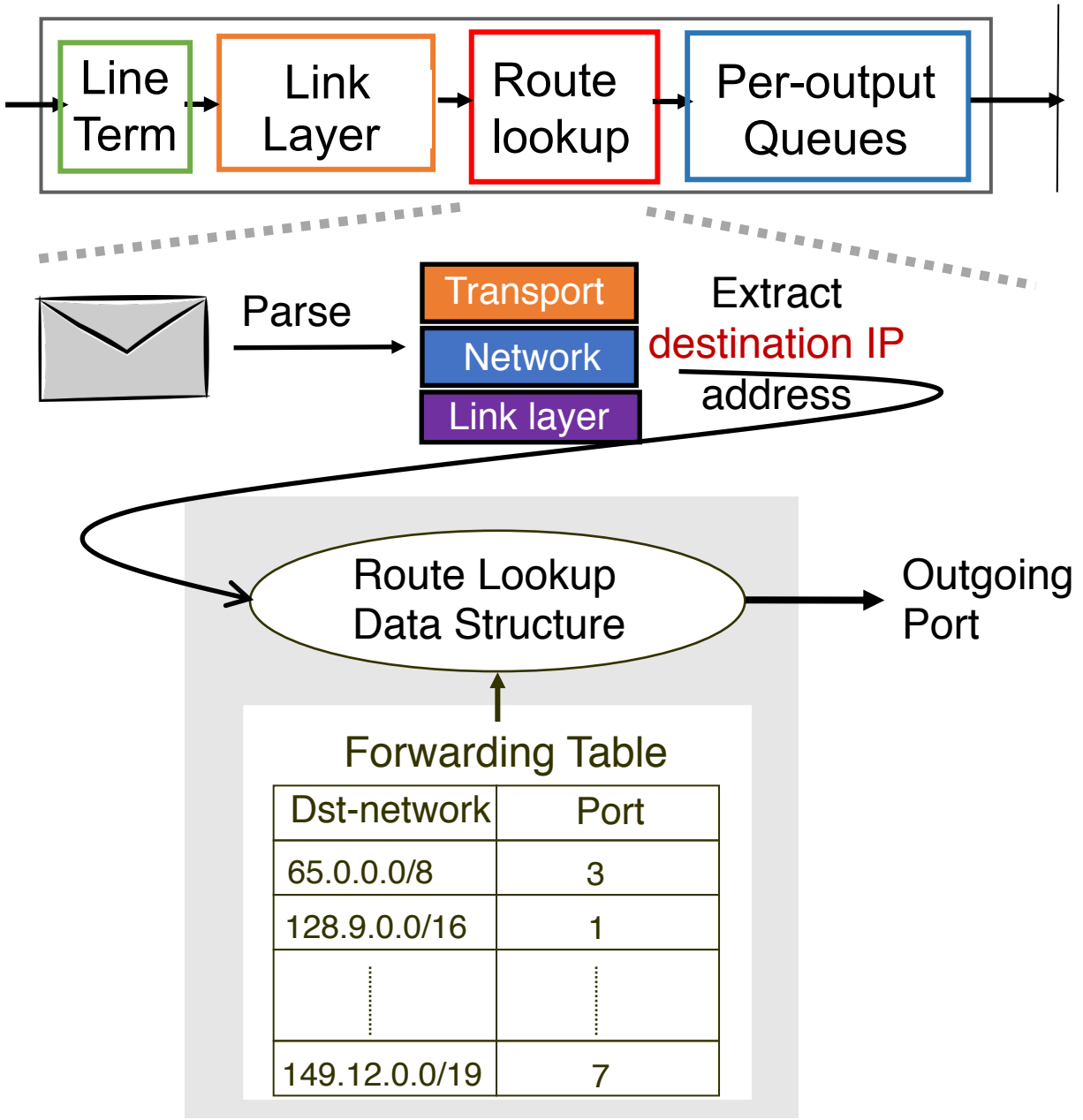
Today: ~ 1 million prefixes.



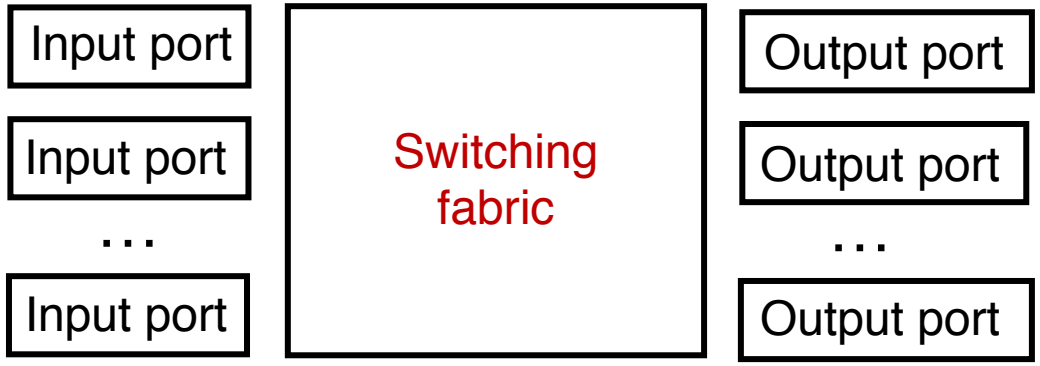
Route lookups

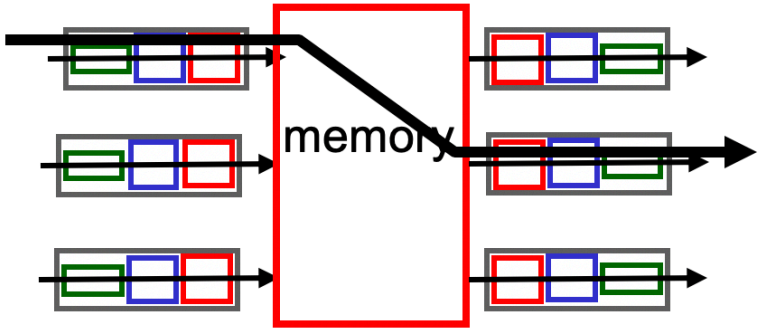
Destination-IP-based forwarding has consequences.

- Forwarding behavior is independent of the source: legitimate source vs. malicious attack traffic
- Forwarding behavior is independent of the application: web traffic vs. file download vs. video
- IP-based packet processing is “baked into” router hardware: evolving the IP protocol faces tall deployment hurdles



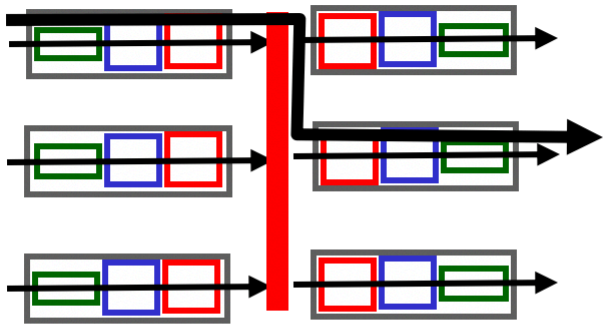
Types of fabrics





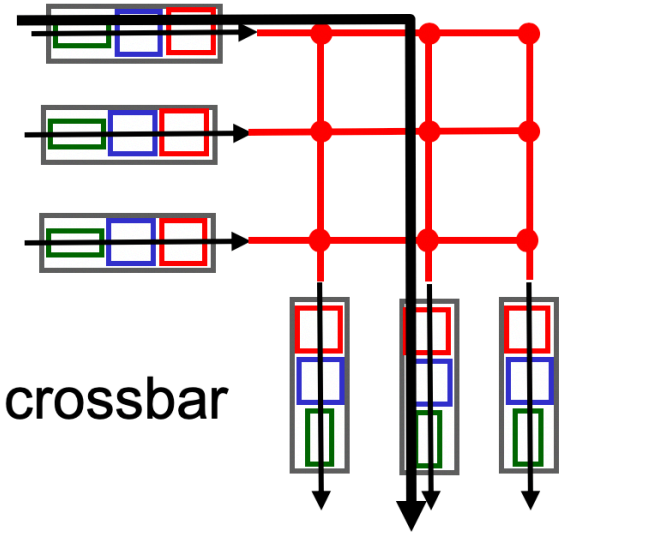
memory

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



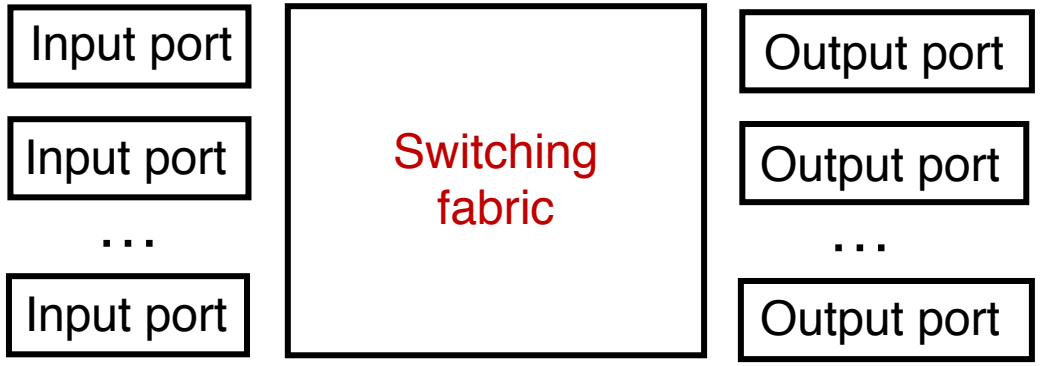
bus

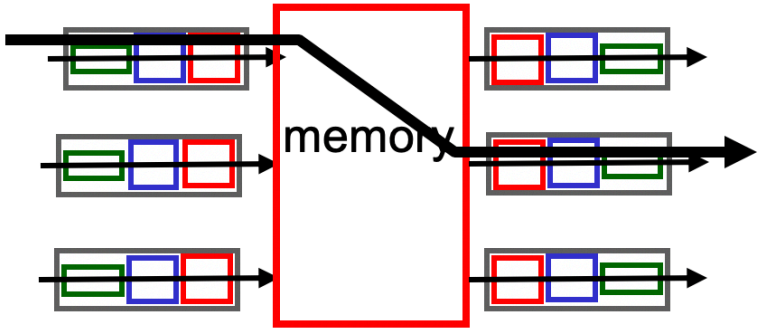
Single shared channel to move data from input to output port.
Easy to build buses; technology is quite mature.



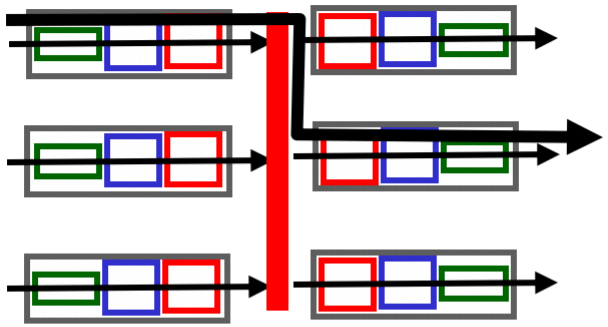
Each input port has a physical data path to every output port.
Switch at the cross-over points turns on to connect pairs of ports.

Types of fabrics

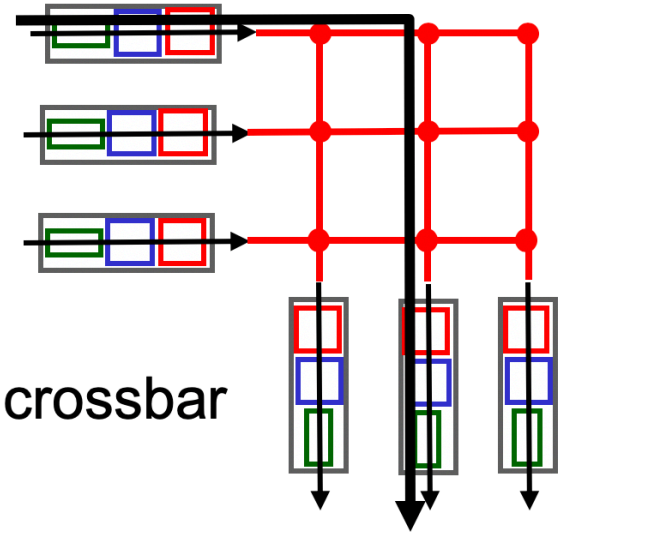




memory



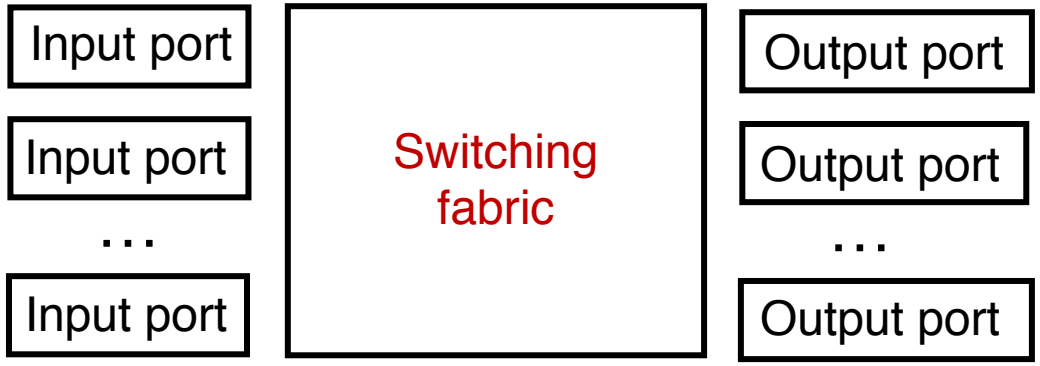
bus



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

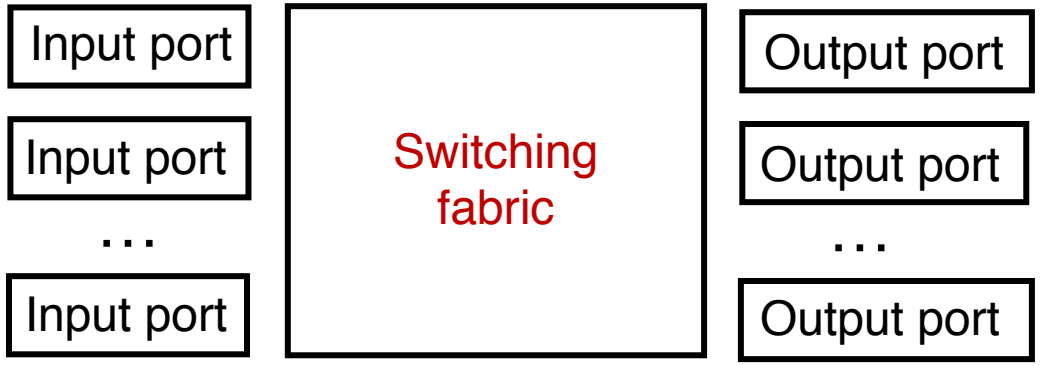
Crossbars can get expensive as the number of ports grows (N^2 connections for N ports)
MGR uses a crossbar and schedules (in,out) port pairs.

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
- Crossbars are nonblocking by design
- Shared memory can be designed to be nonblocking if memory is optimized to be fast enough for the access pattern

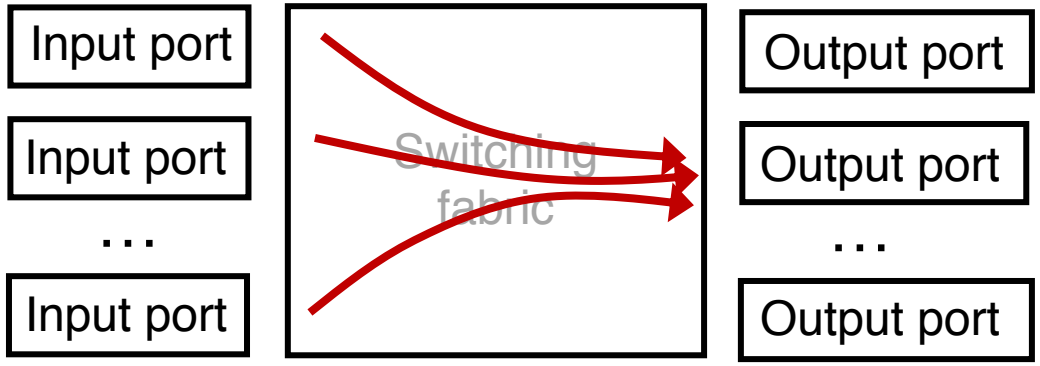
Fabrics



- With a nonblocking fabric, queues aren't formed due to the switching fabric.
  - i.e., 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

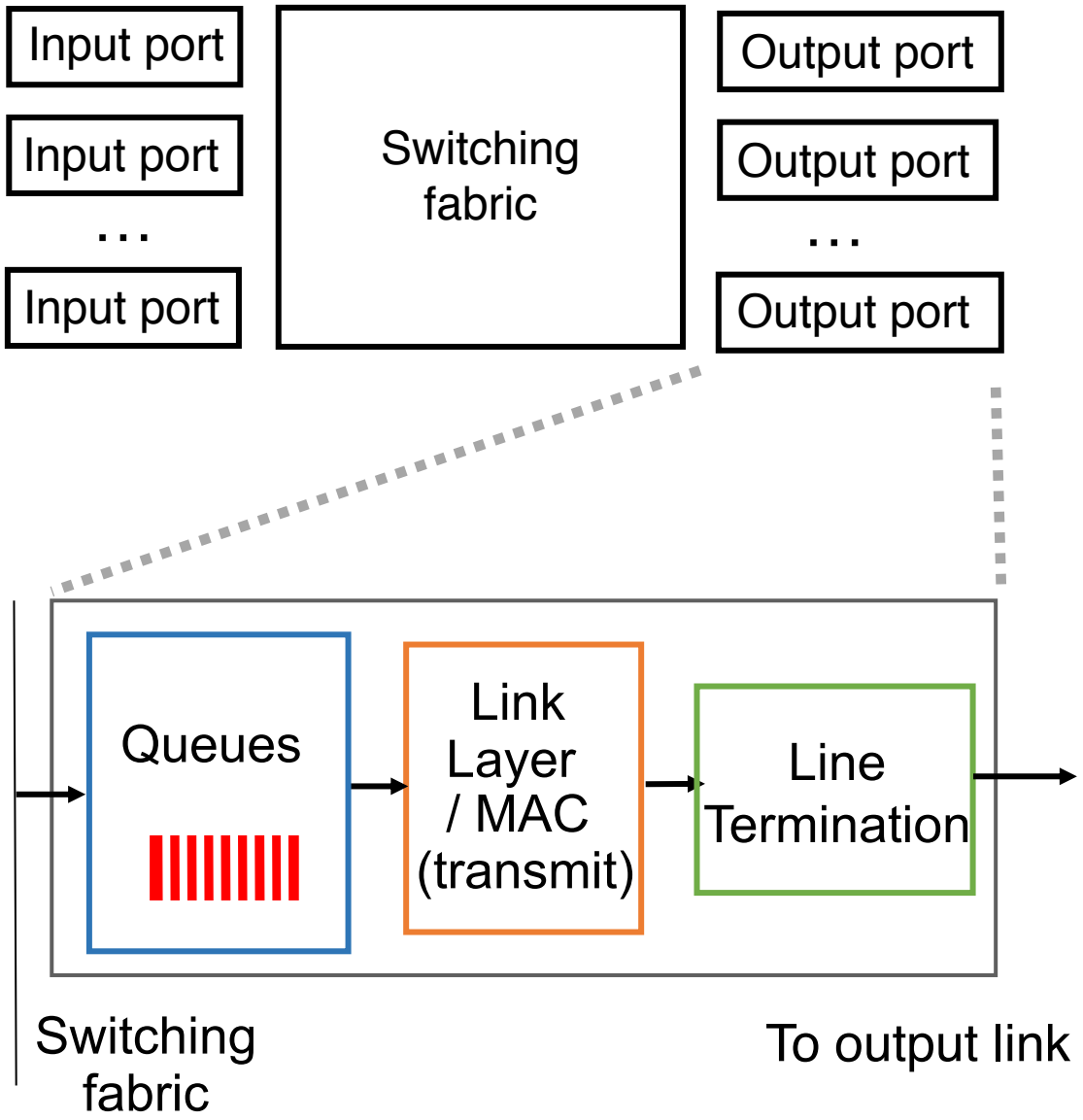
Fabrics



- With a nonblocking fabric, queues aren't formed due to the switching fabric.
  - i.e., 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
- Typically, these (per-output-port) queues form on the output side
  - But queues can also form on the input side if the fabric can't even move packets to the output port's buffer (ie: if output-side buffer filled up)

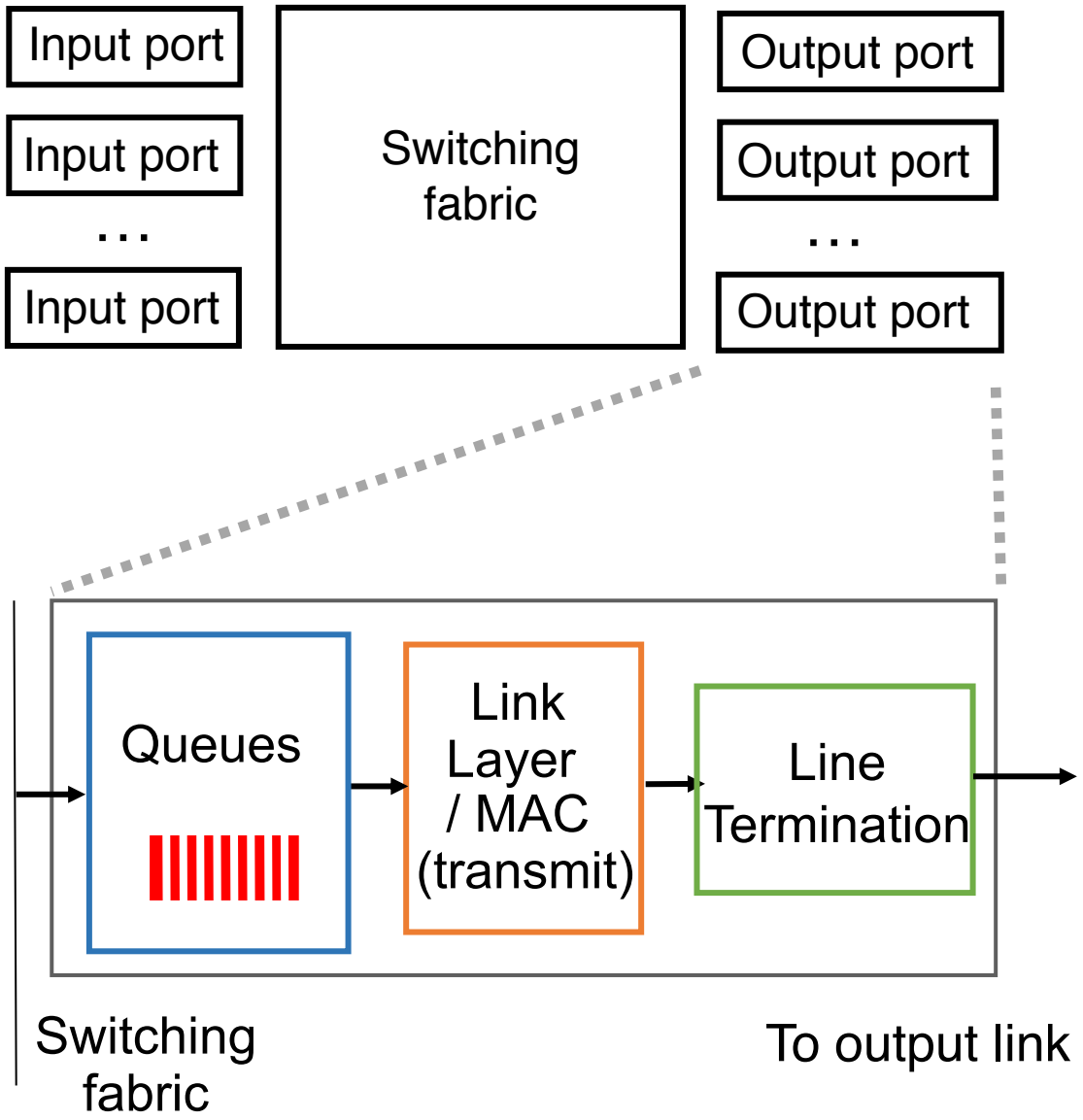
Output port functions

- Components in reverse order of those in the input port
- This is where most routers have the bulk of their packet buffers
  - Recall our discussions regarding router buffer sizes in the transport layer
- MGR uses per-port output buffers, but modern routers have shared memory buffers
  - More efficient use of memory under varying demands



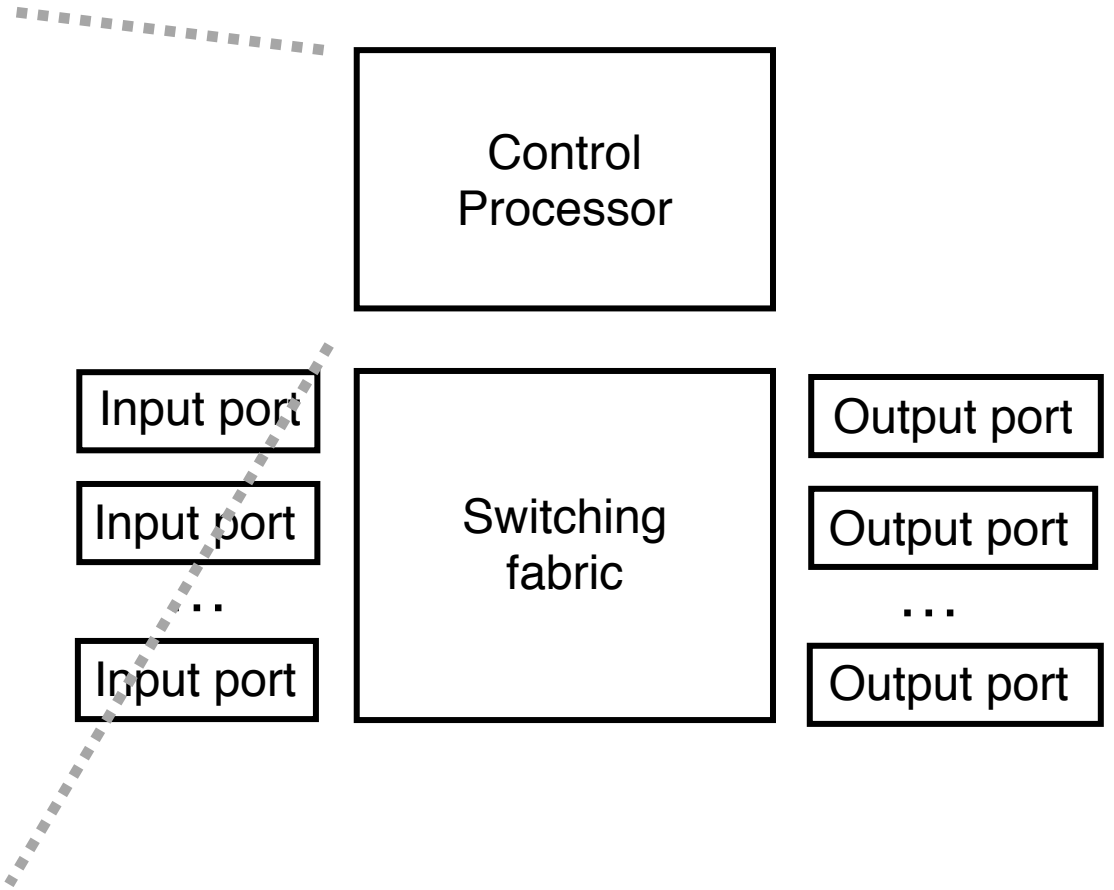
Output port functions

- Two important policy decisions
- Scheduling: which among the waiting packets gets to be transmitted out the link?
  - Ex: First-In-First-Out (FIFO)
- Buffer management: which among the packets arriving from the fabric get space in the packet buffer?
  - Ex: Tail drop: later packets dropped first



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.



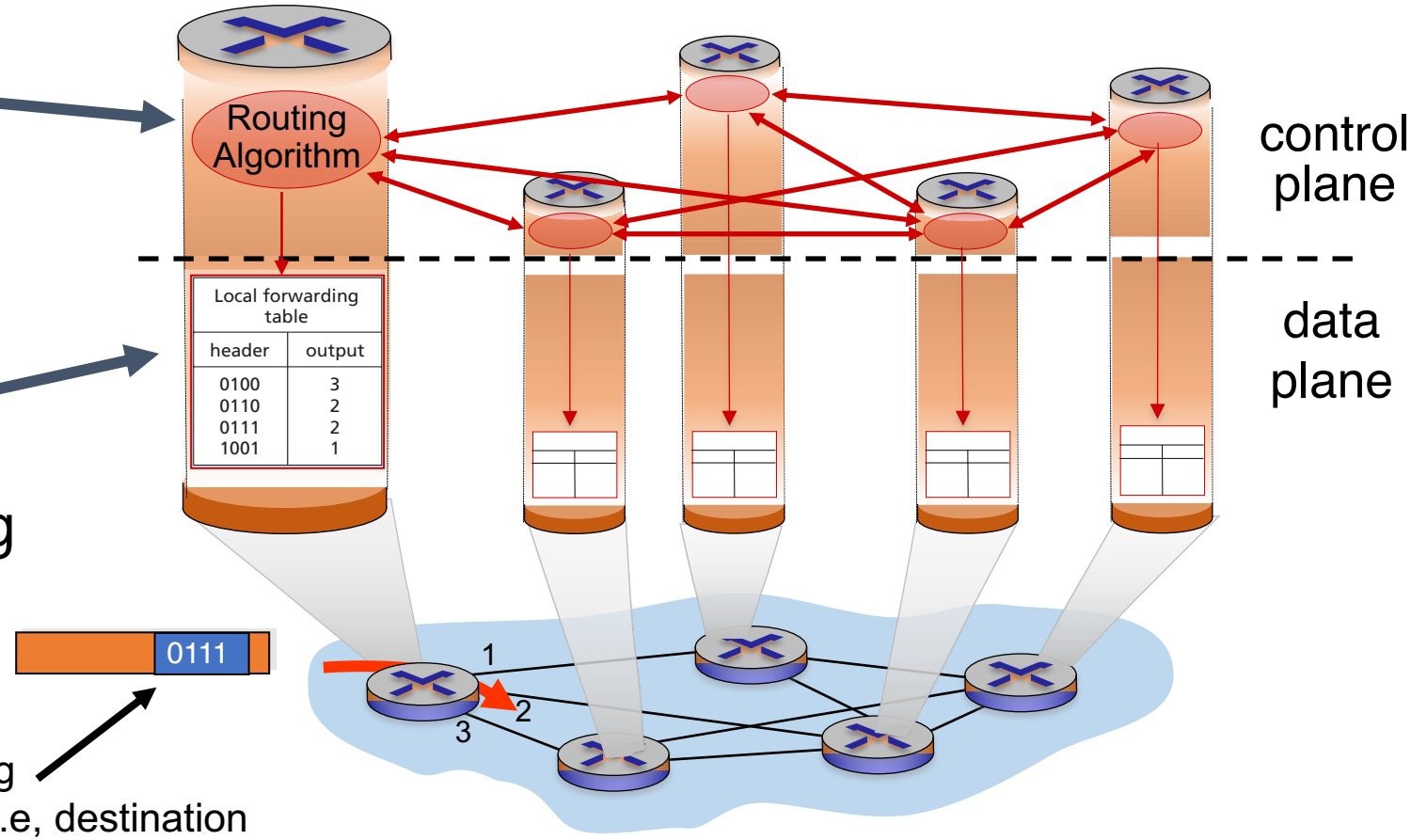
Router design: the bigger picture

Control plane

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

Data plane

per-packet processing (~ tens of nanoseconds)



CS 352
Longest Prefix Matching

CS 352, Lecture 15.2

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

Srinivas Narayana

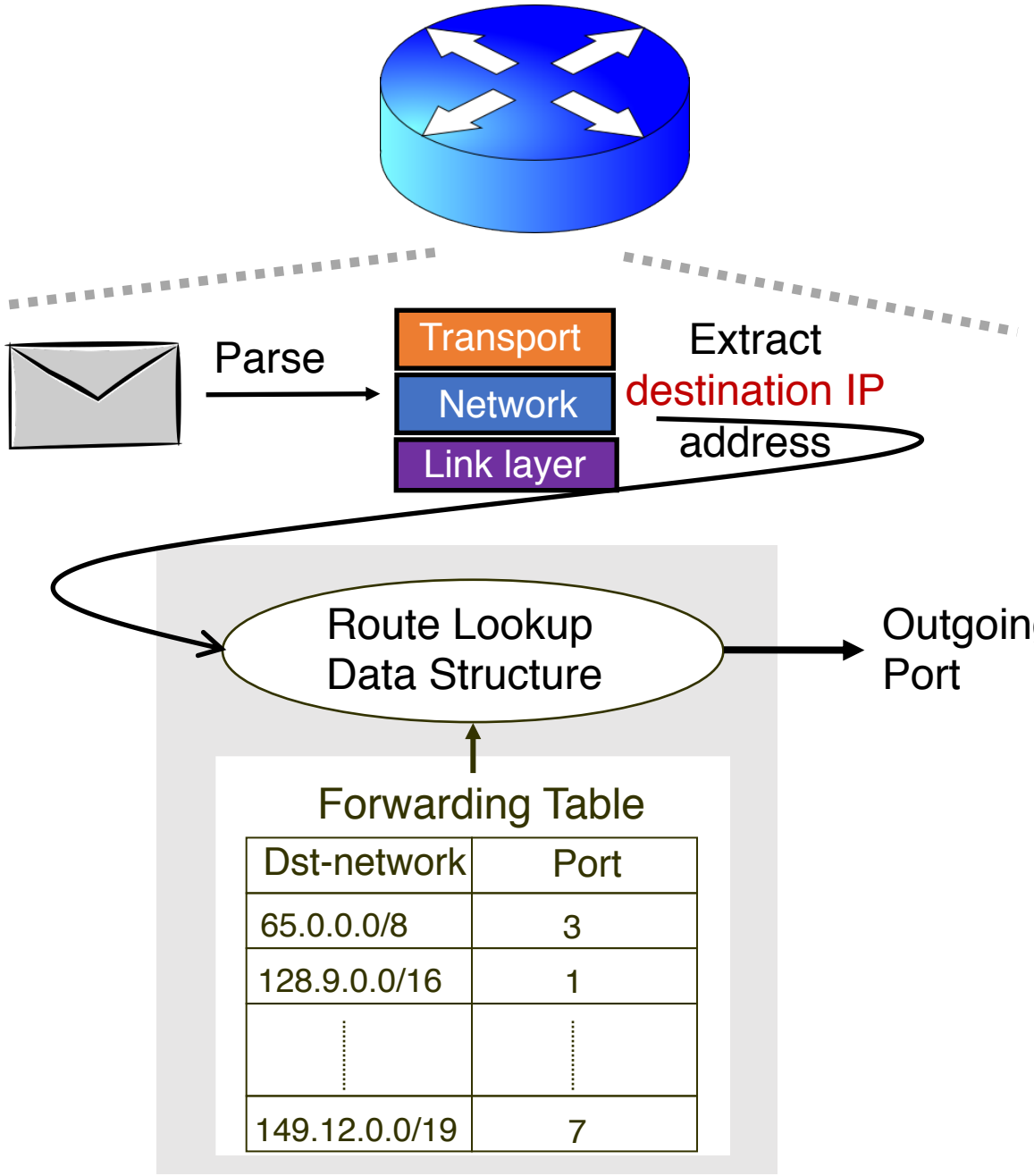
RUTGERS
UNIVERSITY | NEW BRUNSWICK

Review: Route lookup

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

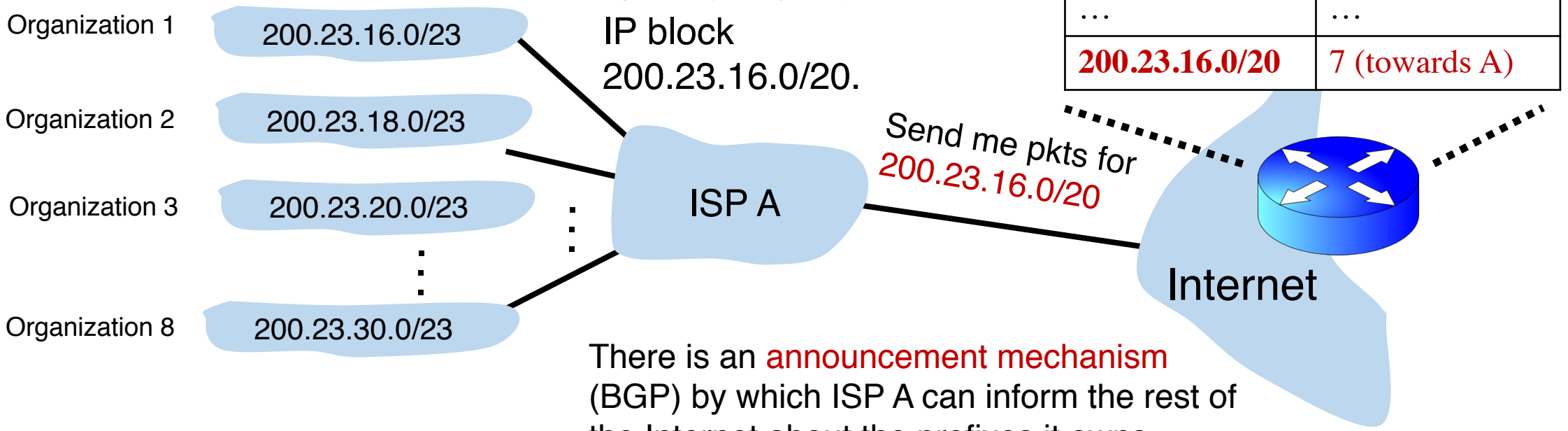
- Prefixes are allocated to organizations by Internet registries

- 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

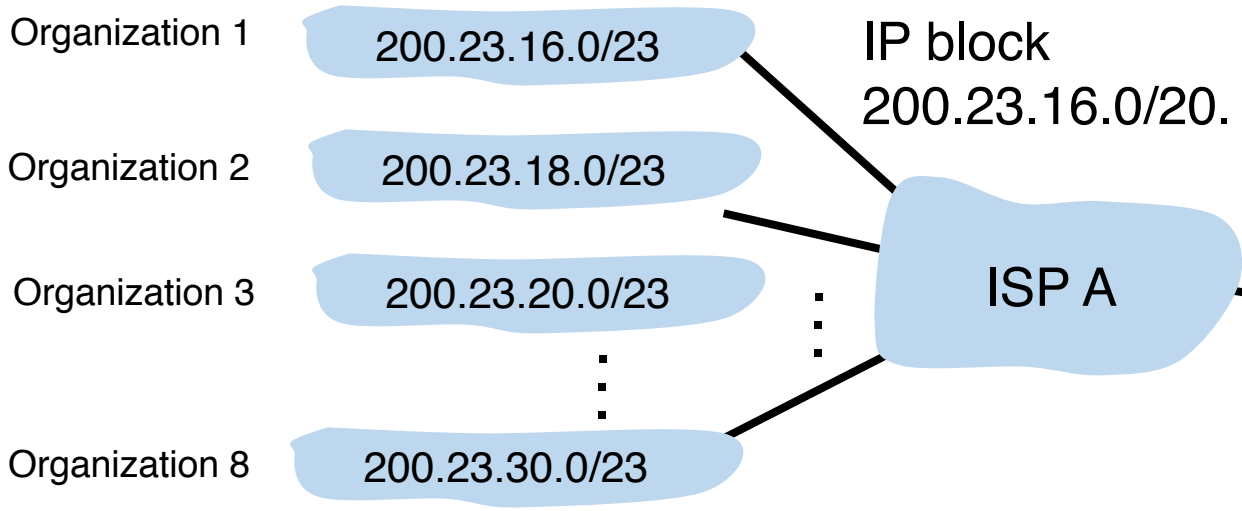


Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
…	…
200.23.16.0/20	7 (towards A)

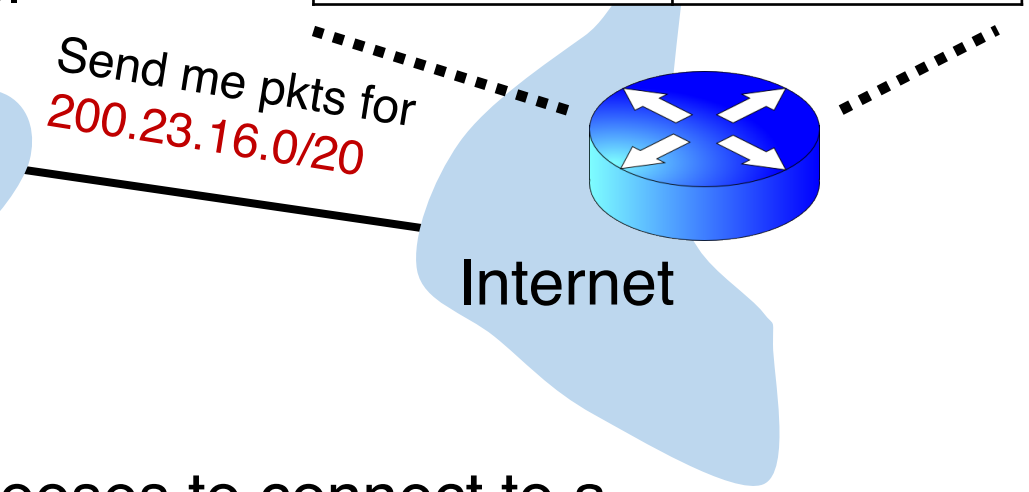
There is an announcement mechanism (BGP) by which ISP A can inform the rest of the Internet about the prefixes it owns.
It is enough to announce a coarse-grained prefix 200.23.16.0/20 rather than 8 separate sub-prefixes.

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)



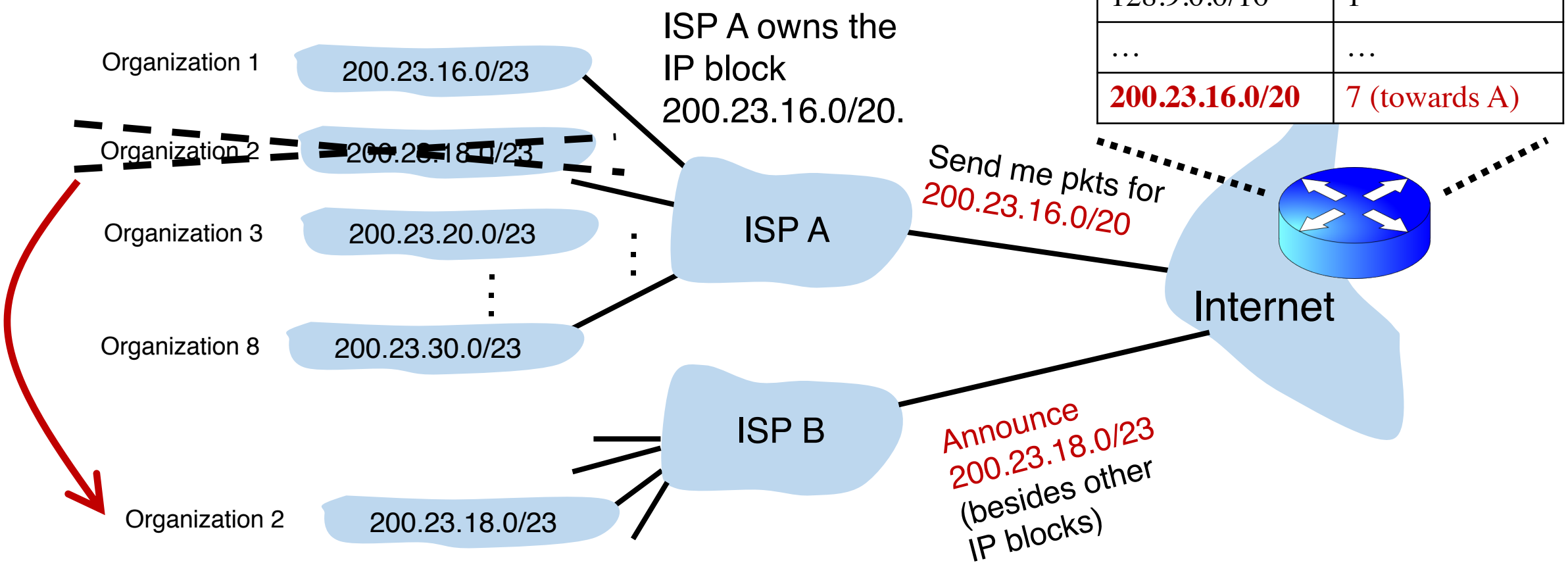
Now suppose one of these organizations chooses to connect to a different ISP for its Internet service.

It is possible for this 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

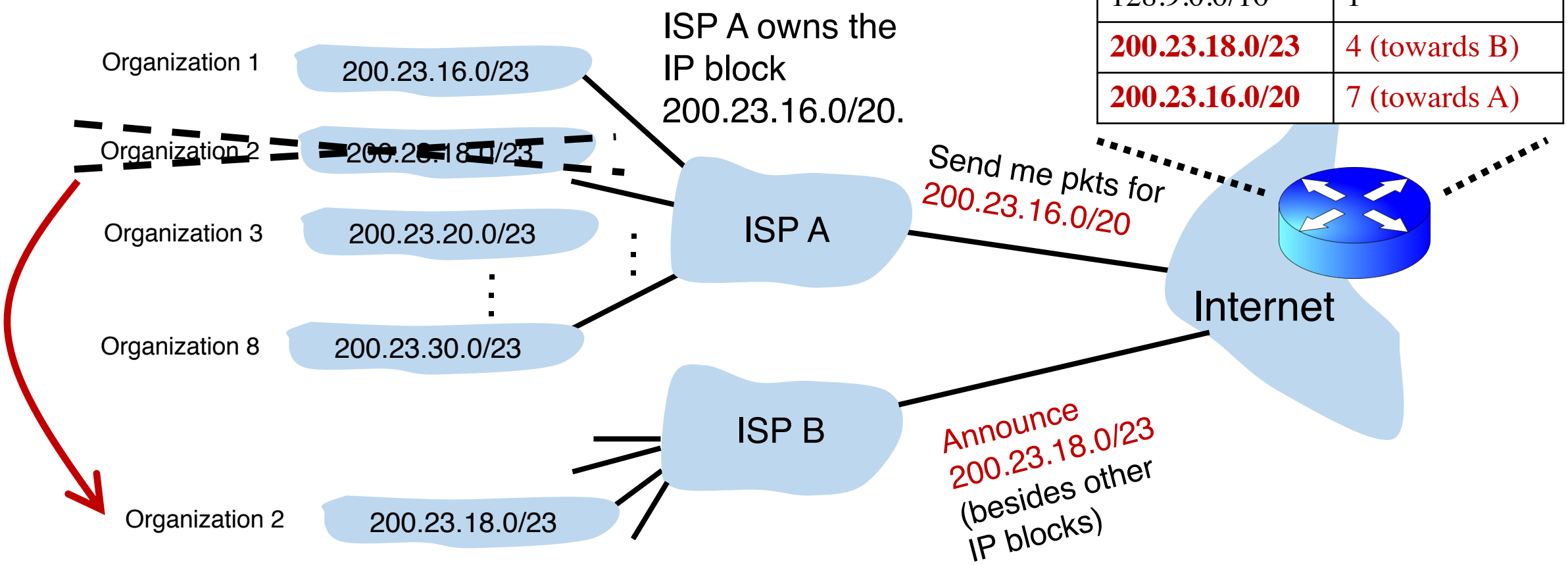
Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
…	…
200.23.16.0/20	7 (towards A)



Example of IP block reallocation

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

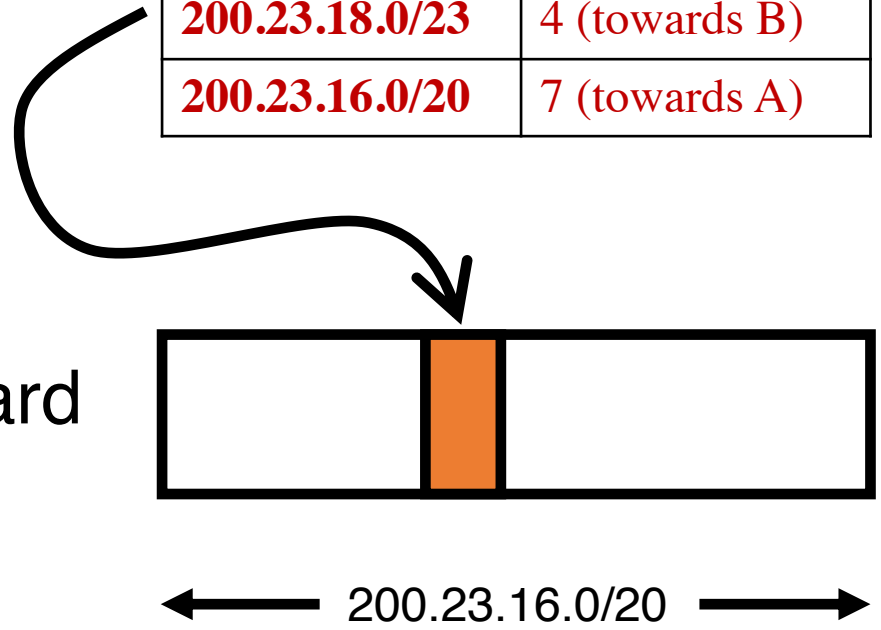
Dst IP Prefix	Output port
65.0.0.0/8	3
128.9.0.0/16	1
200.23.18.0/23	4 (towards B)
200.23.16.0/20	7 (towards A)



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?
  - Clearly, towards ISP 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)



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

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)