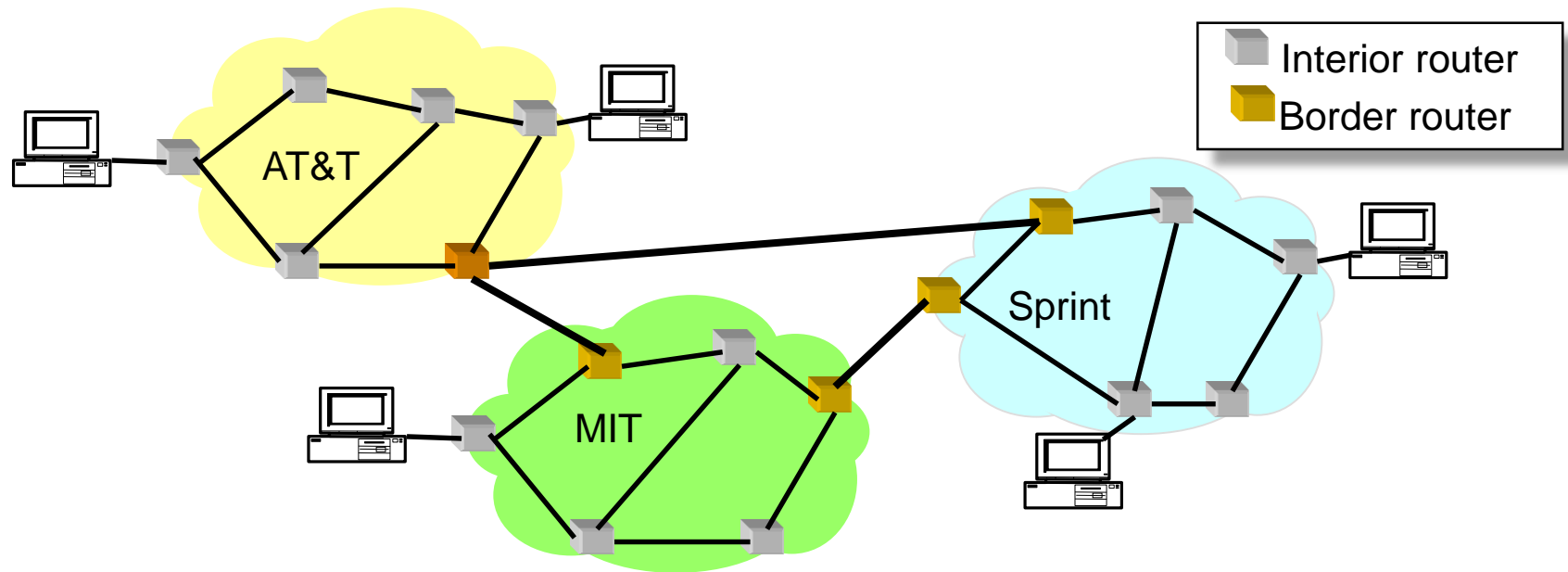


6.829 Computer Networks

Inter-Domain Routing -- BGP


Dina Katabi - MIT

The Internet is a Network of Networks

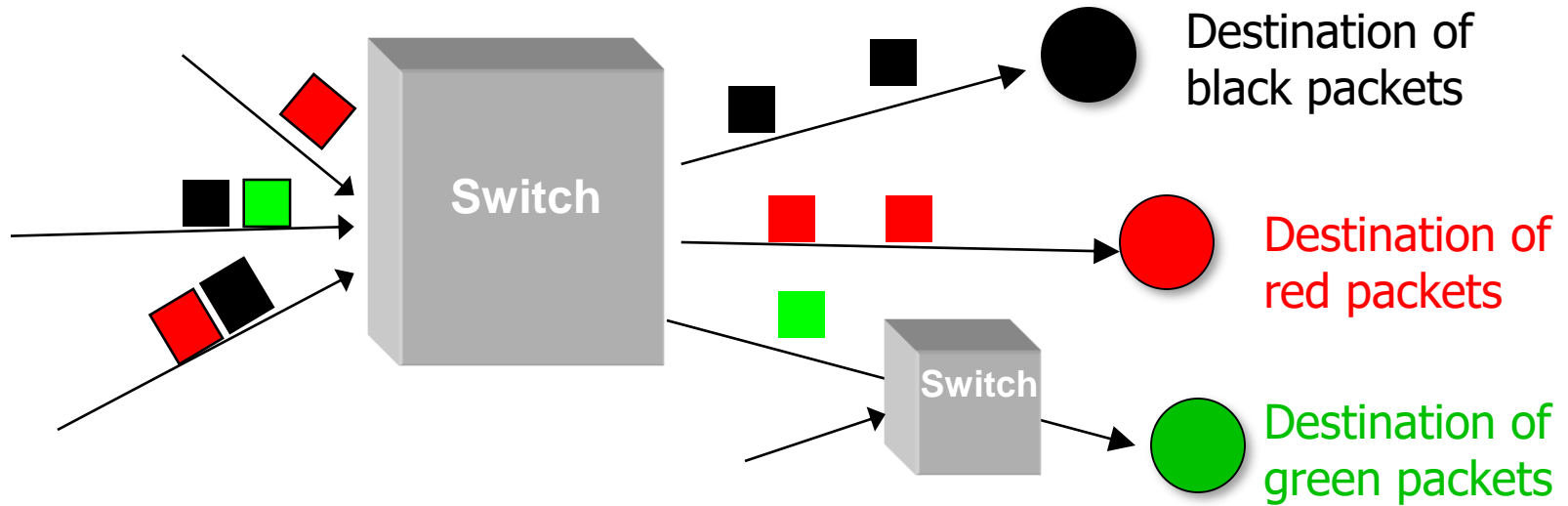


- The Internet is a network of Autonomous Systems (ASs)
 - E.g., MIT, AT&T, Stanford, ...
- Internally, each AS runs its own routing protocol → Autonomy
- Across ASs, we run a different routing protocol (called BGP)

Outline

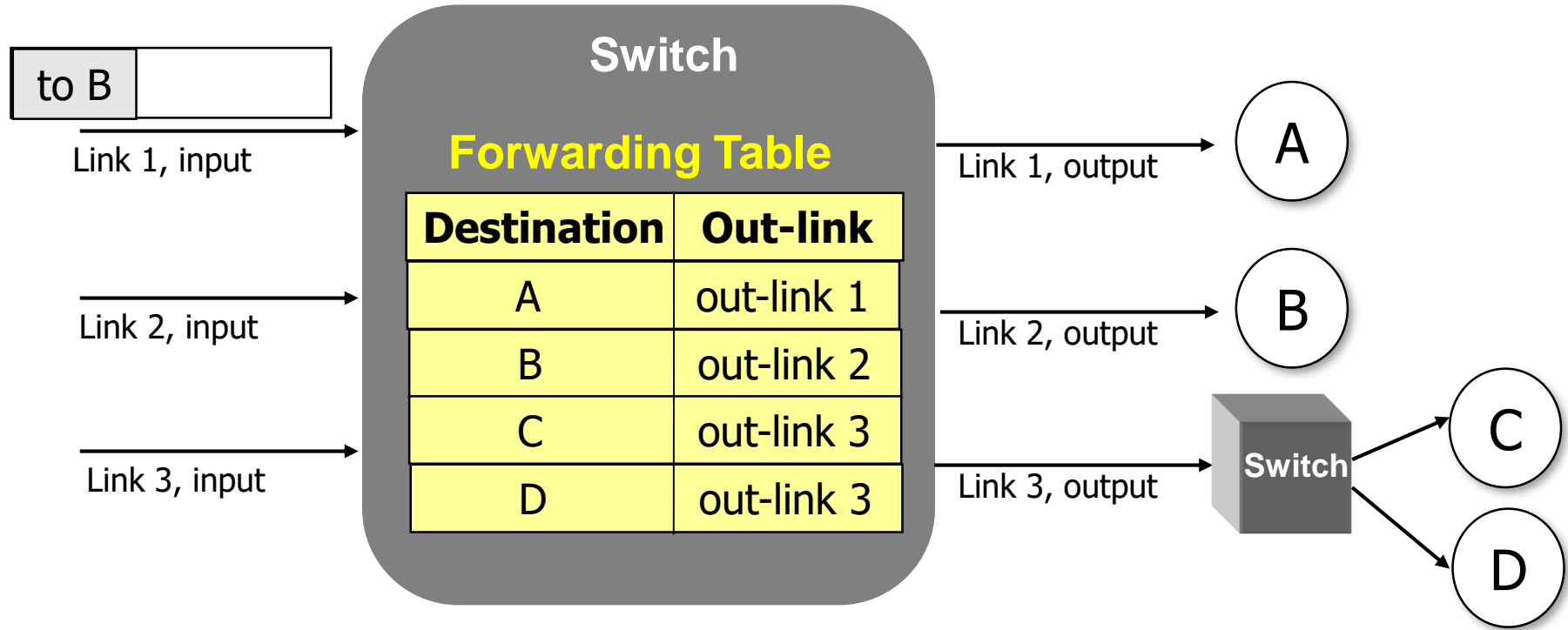
- 
- Review of intra-Domain Routing
 - Inter-Domain Routing
 - BGP

The Job of a Switch (or Router)



- A switch has input links and output links
- A switch **sends** an input **packet** on the output link leading **toward the packet's destination node**
- A switch does not care of who generated the packet

How does the switch know which output link leads to a packet destination?

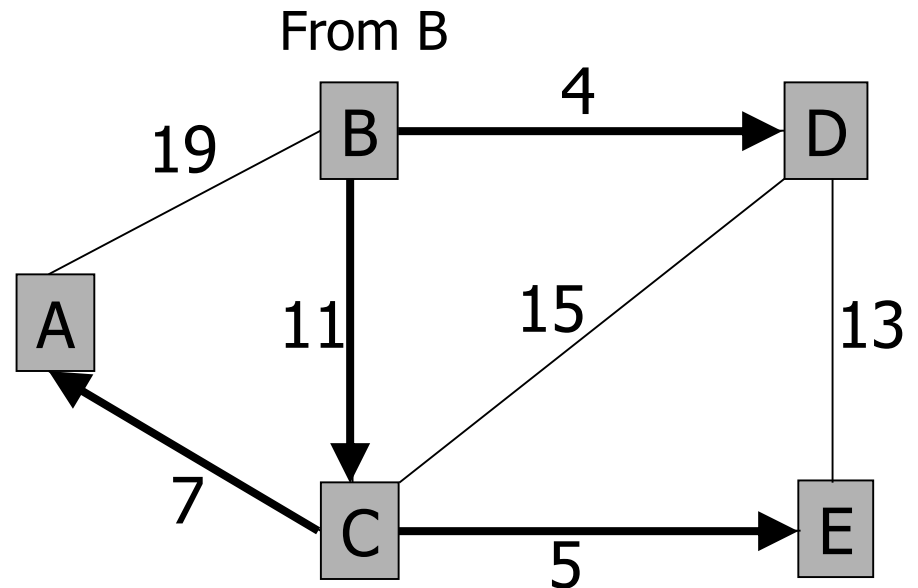


- Packet **header** has the destination
- Switch **looks up the destination in its table** to find output link
- Table is built using a **routing protocol**

Intra-Domain Routing

- Objective: Allows each node in the network (i.e., in the AS) to find the **shortest path** to all other reachable nodes

- Network is a graph
- Links have costs (may refer to delay, 1/BW, congestion, etc.)
- “Shortest path” means the path with the minimum total link cost



- Note paths from a node, e.g., B, form a **shortest path tree** rooted at B

Requirement from a Routing Protocol

- Correctness
 - Each route must lead to the correct destination
- Completeness
 - If destination is reachable, the protocol should find a route
- Convergence
 - If network graph does not change, the routes must eventually converge (i.e., stop changing)
- Loop-freedom
 - After convergence, the routes have no loops

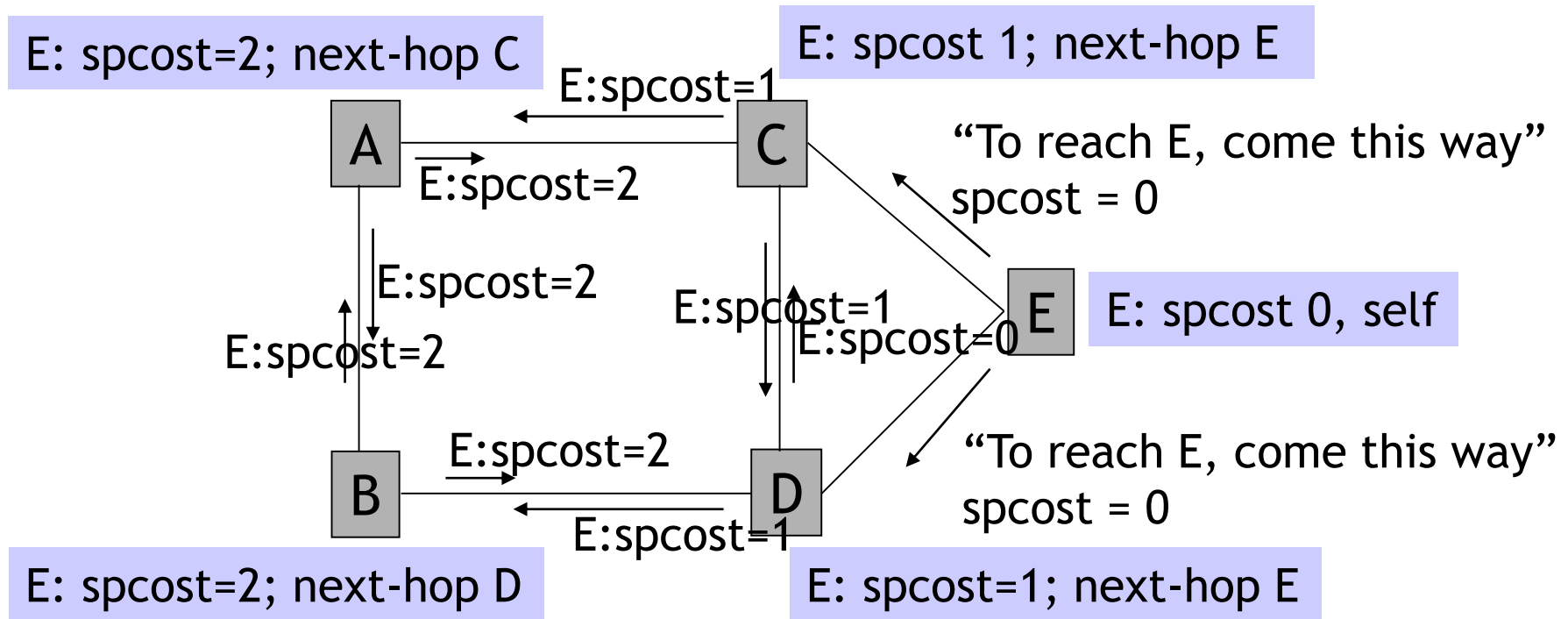
Two Common Intra-Domain Routing Protocols

(expected to know from undergrad.)

- **Link State (Based on Dijkstra shortest path alg.)**
 - Each node tells everyone about its neighbors and link costs
 - Each node obtains the network graph
 - Each node locally computes paths from itself to everyone
- »
- **Distance Vector (Based on Bellman-Ford shortest path alg.)**
 - A node tells only its neighbors its shortest path cost to every node in network
 - Each node updates its shortest path based on the shortest path of its neighbors
 - No one has the full graph

Example Distance Vector Routing

Find routes from all nodes to E; assume all links have a cost of 1



- Each node periodically sends a vector of <destination:spcost> pairs to all its neighbors
- On hearing announcement,
if ($\text{my spcost to dst} > \text{spcost in announcement} + \text{link_cost}$),
 - My spcost == spcost in announcement + link cost
 - Next-hop == node that sent me the announcement

Outline

- Review of intra-Domain Routing
- • Inter-Domain Routing
- BGP

Requirements of Inter-Domain Routing

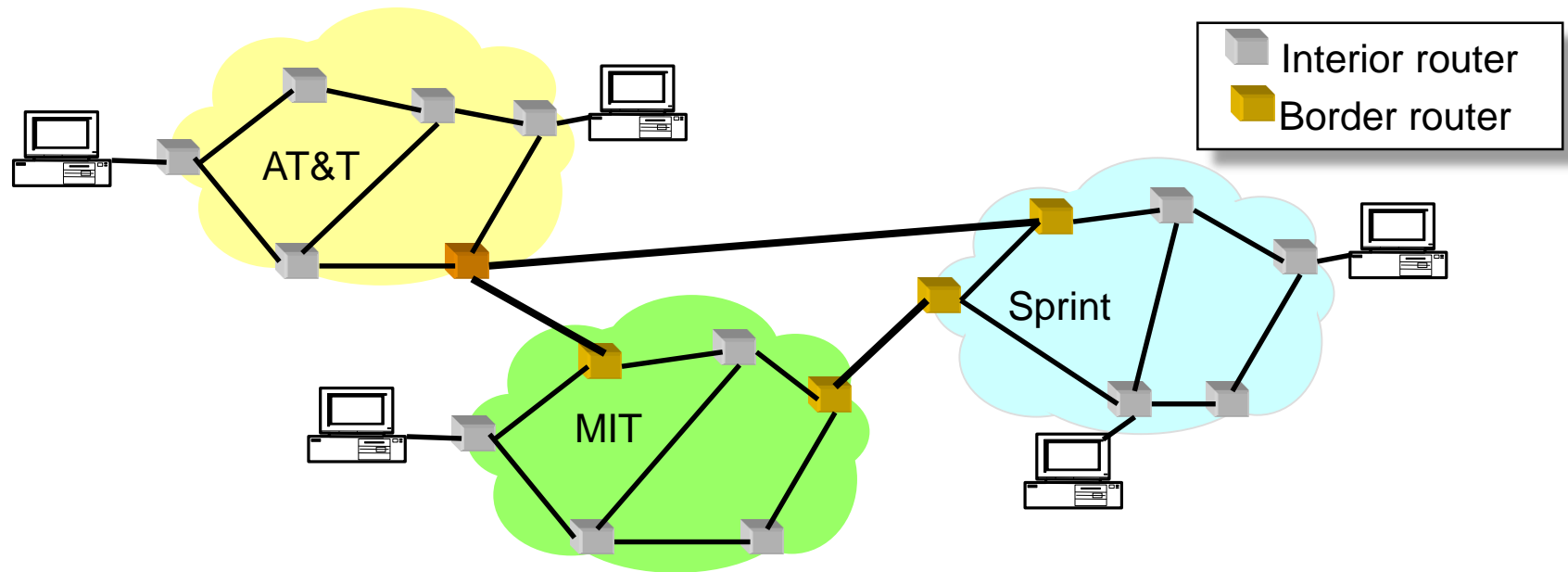
- Scalability
 - **Small routing tables:** Cannot have an entry per machine
→ causes large look up delay
 - **Small message overhead and fast convergence:** A link going up or down should not cause routing messages to spread to the whole Internet
- Policy-compliant
 - Shortest path is not the only metric; Internet Service Providers (ISPs) want to maximize revenues!

Idea for Scaling

- Need less information with increasing distance to destination

→ Hierarchical Routing

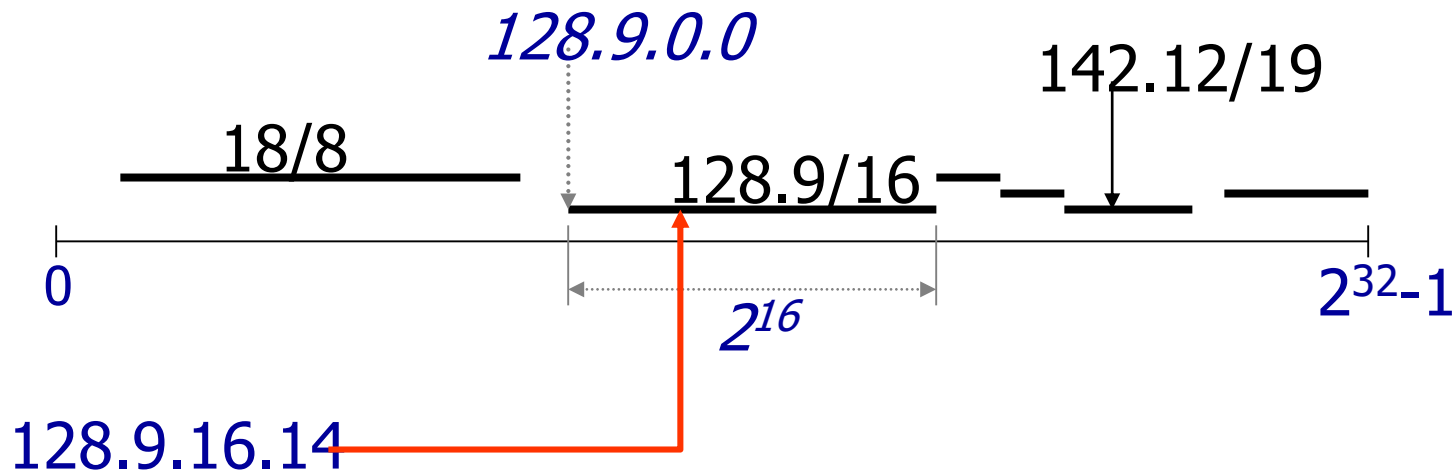
The Internet Hierarchy



- Internally, each AS runs its own routing protocol (link state or distance vector) → Autonomy
- Across ASs, we run a different routing protocol (called BGP) → Hierarchy → More scalable

Hierarchical Addressing

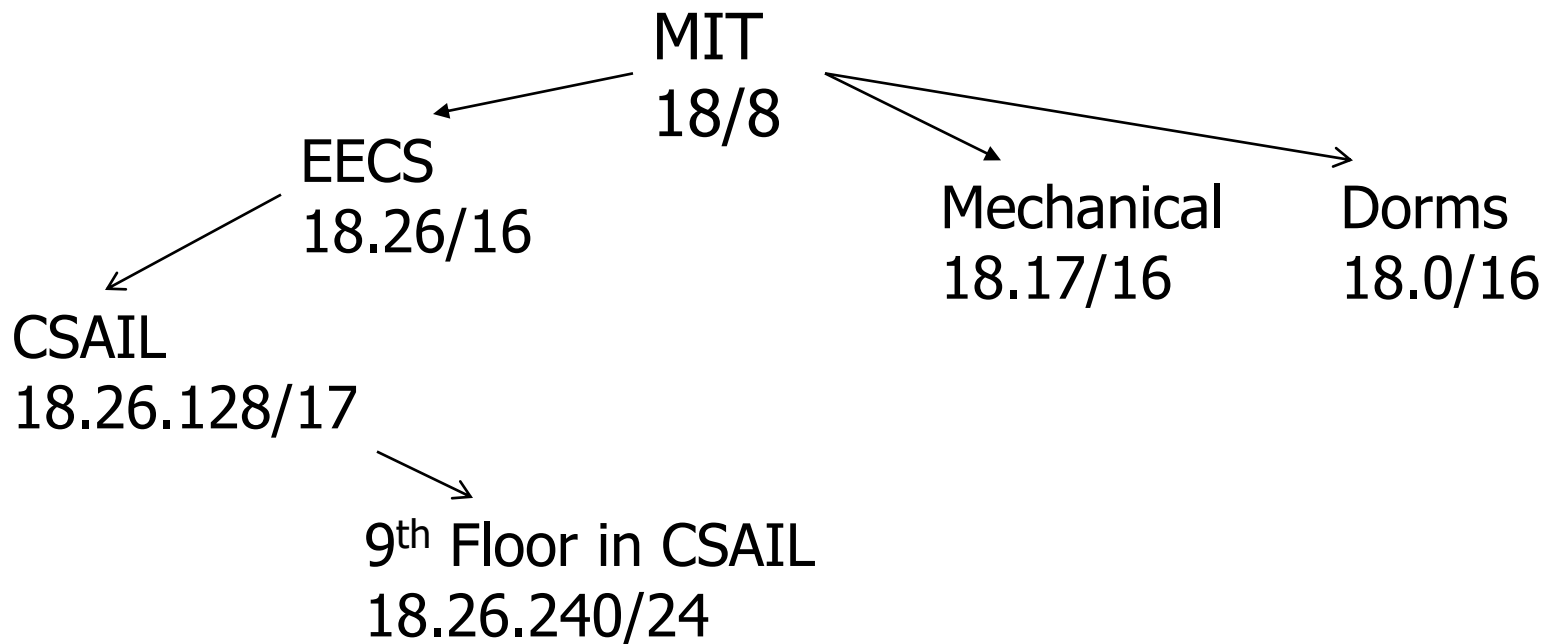
- ❖ Each IP address is 4 bytes, e.g., 18.0.1.2
- ❖ The IP address space is divided into line segments (i.e., contiguous chunk of addresses)
- ❖ Each segment is described by a *prefix*.
- ❖ A prefix is of the form x/y where x is the prefix of all addresses in the segment, and y is the length of the segment in bits
- ❖ e.g. The prefix 128.9/16 represents the segment containing addresses in the range: 128.9.0.0 ... 128.9.255.255.



Hierarchical Address Allocation

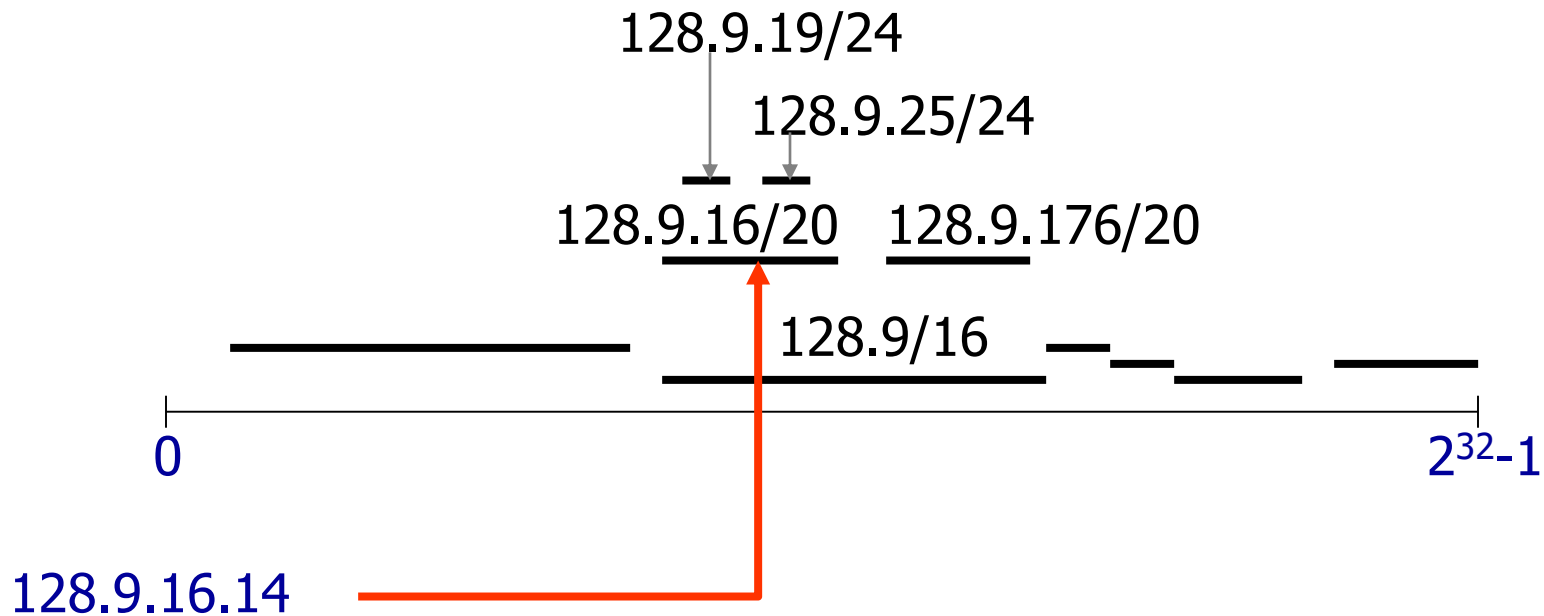
- Addresses that start with same prefix are co-located
 - E.g., all addresses that start with prefix 18/8 are in MIT
- Entries in the routing/forwarding table are for IP prefixes → shorter routing tables

Address Aggregation



- Forwarding tables in Berkeley can have one entry for all MIT's machines. E.g., (18/8, output-link)
- Forwarding tables in Mechanical Engineering have one entry for all machines in EECS
- But, a switch on the 9th floor subnet knows about all machines on its subnet

Longest Prefix Match

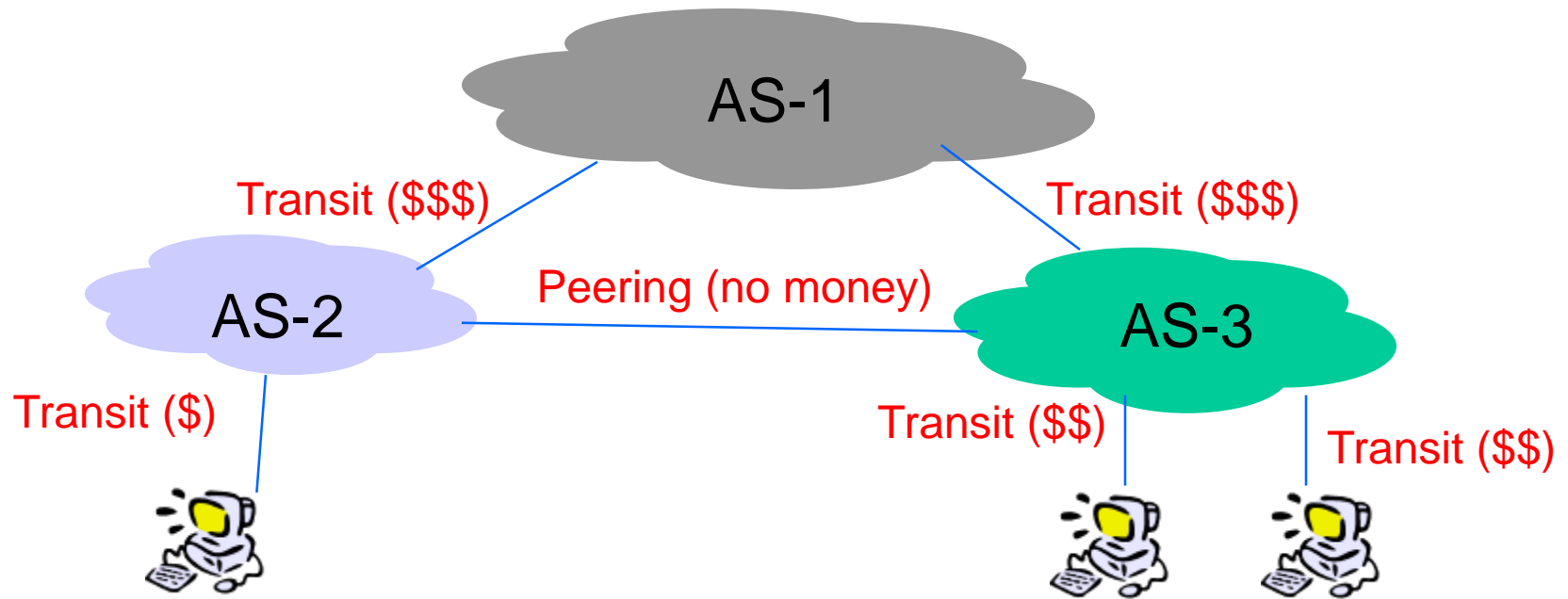


Most specific route = “longest matching prefix”

A Router forwards a packet according to the entry in the forwarding table that has the longest matching prefix

- Hierarchical addressing and routing give us scalability
- Still need to tackle policies

Inter-AS Relationship: Transit vs. Peering

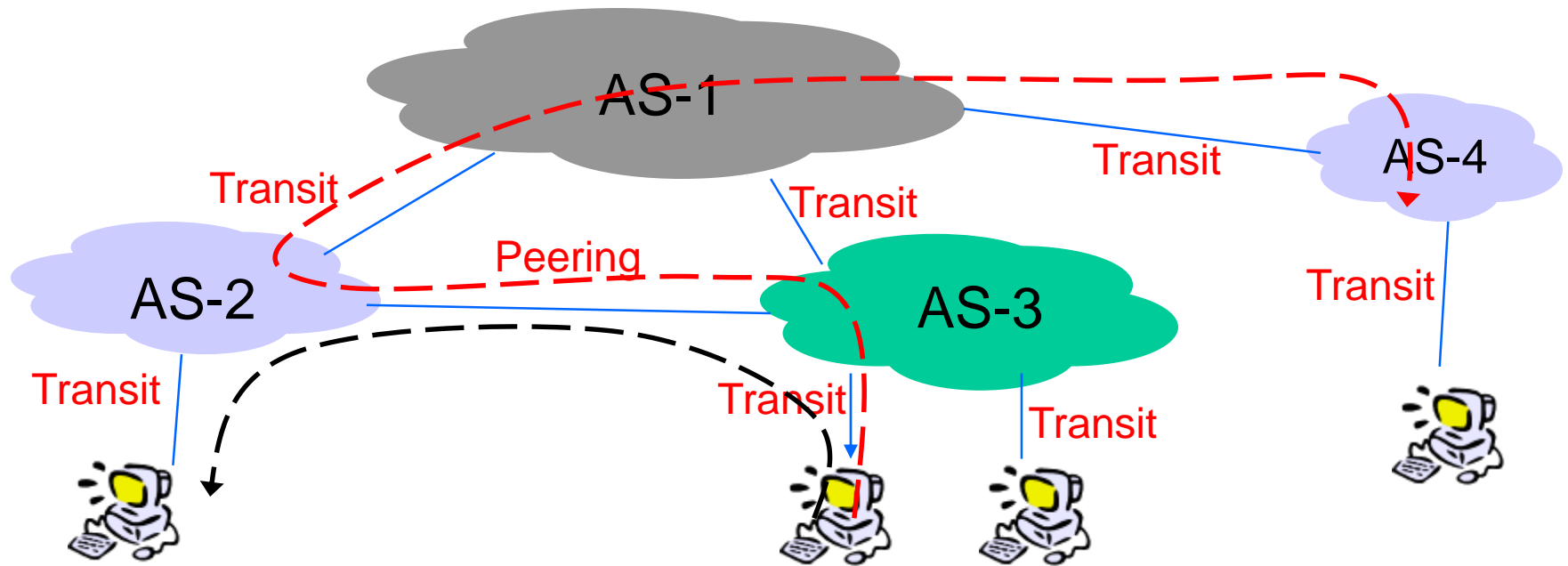


- Transit relationship
 - One AS is a customer of the other AS, who is the provider; customer pays provider both for sending and receiving packets
- Peering relationship
 - Two ASs forward packets for each other without exchanging money

Policy-Based Routing

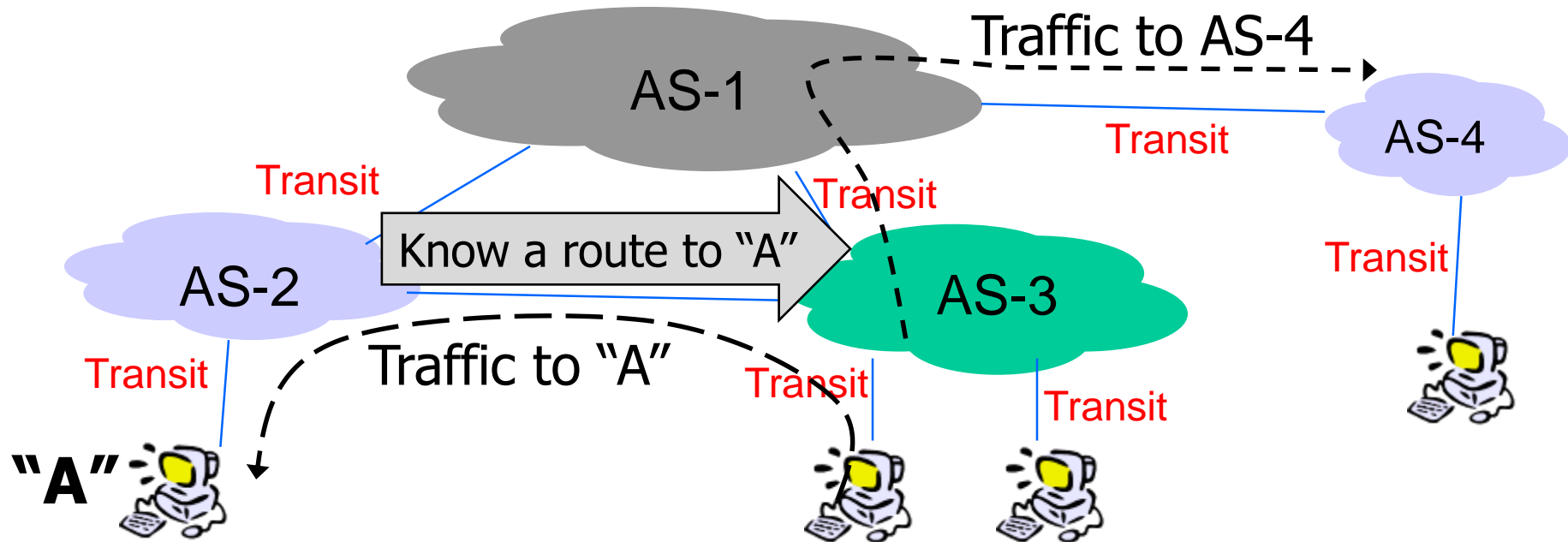
- Main Rule:
 - An AS does not accept transit traffic unless it makes money of it
- Rule translates into incoming and outgoing routing policies

Desirable Incoming Policies



- AS-2 likes AS-3 to use the peering link to exchange traffic between their customers → saves money because it bypasses AS-1
- But, AS-2 does not want to forward traffic between AS-3 and AS-4 because this makes AS-2 pay AS-1 for traffic that does not benefit its own customers

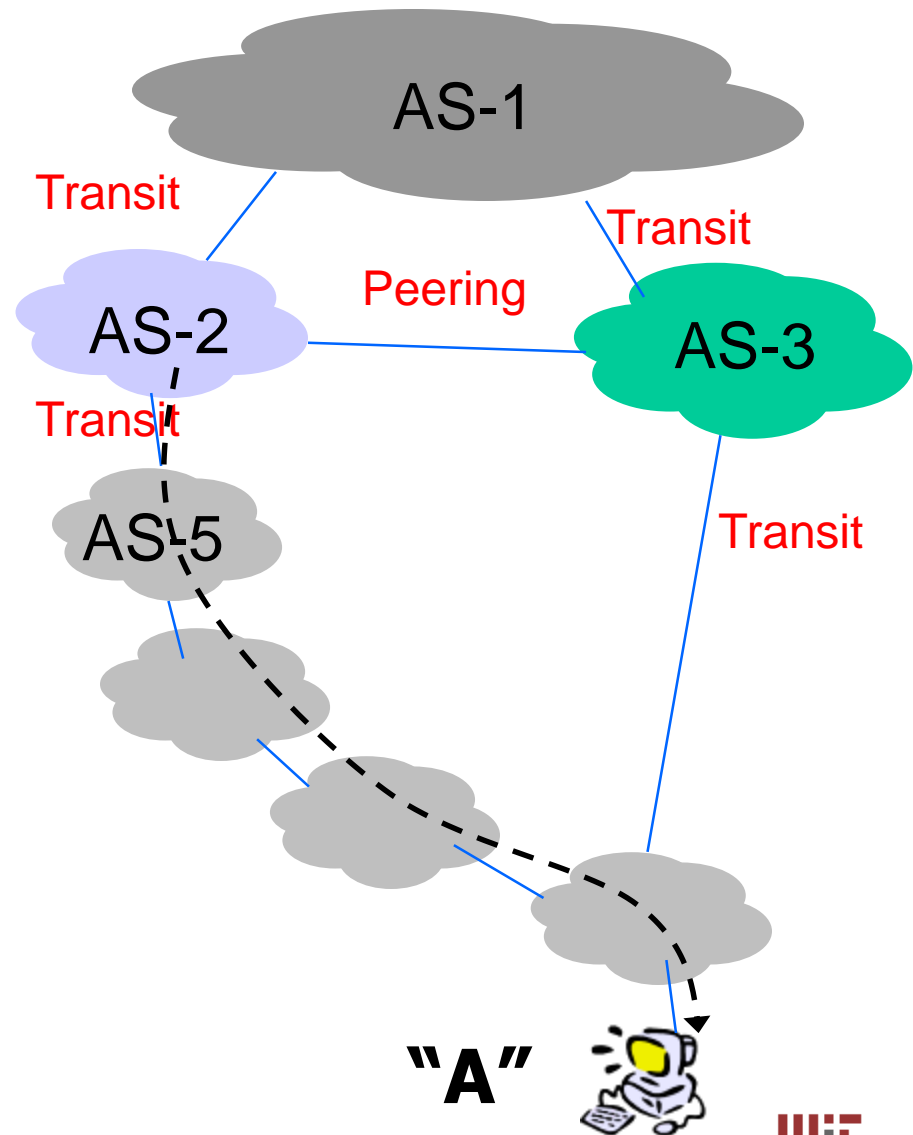
How Does AS-2 Control Incoming Traffic?



- AS-2 advertises to AS-3 a route to its customer's IP prefix
- AS-2 does not tell AS-3 that it has a route to AS-4, i.e., it does not tell AS-3 routes to non-customers IP-prefixes

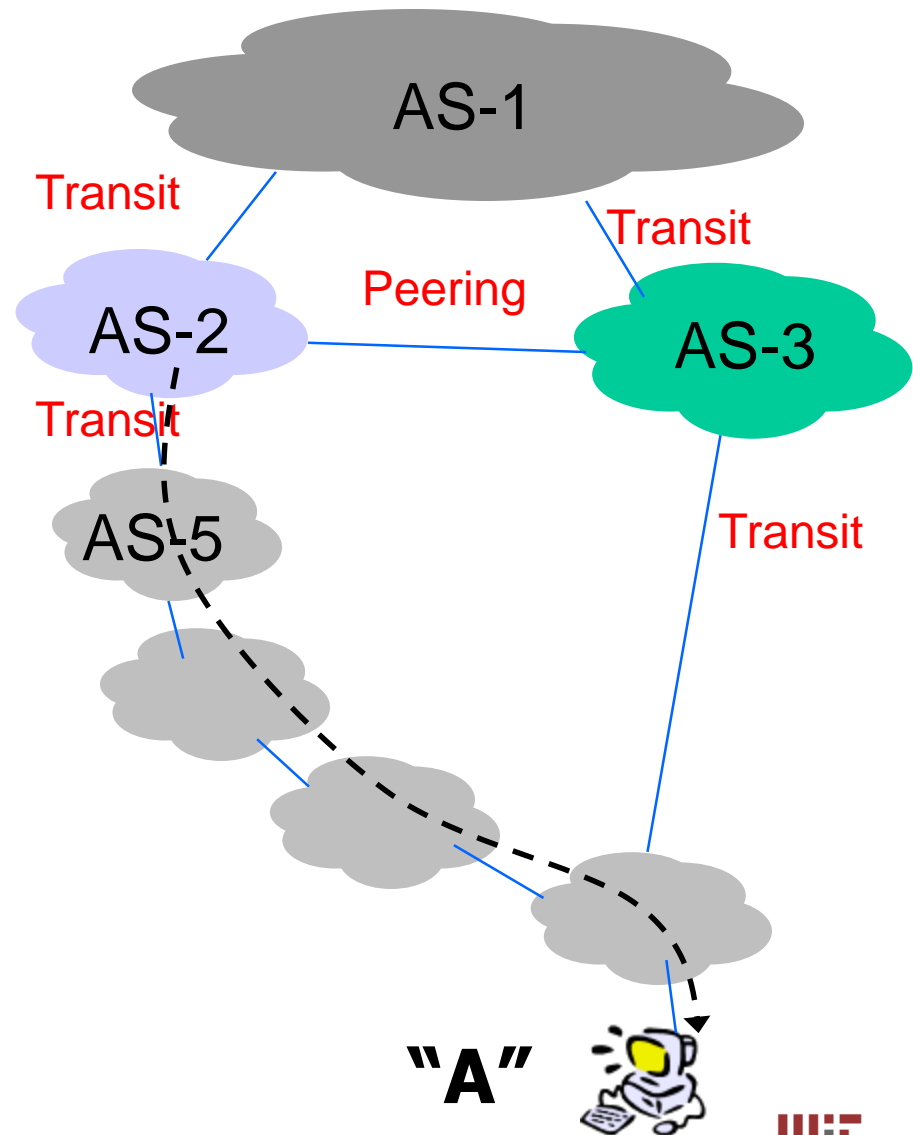
Desirable Outgoing Policies

- AS-2 prefers to send traffic to “A” via its customer AS-5 rather than its provider or peer despite path being longer



How Does AS-2 Control Outgoing Traffic?

- AS-1, AS-3, and AS-4 advertise their routes to “A” to AS-2
- But AS-2 uses only AS-5’s route (i.e., it inserts AS-5’s route and the corresponding output link into its forwarding table)

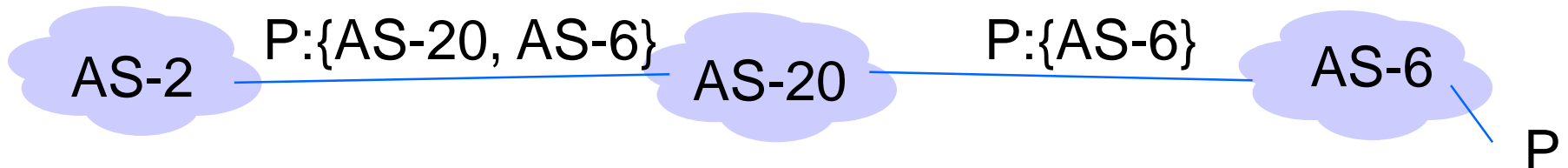


Outline

- Review of intra-Domain Routing
- Inter-Domain Routing
- • BGP

BGP: Border Gateway Protocol

1. Advertize whole path



- Faster loop detection → an AS checks for its own AS number in advertisement and rejects route if it has its AS number

2. Incremental updates

- AS sends routing updates only when its **best/current** route changes (Messages are reliably delivered using TCP)
- Two types of update messages: announcement, e.g., $P:\{AS-20, AS-6\}$ and withdrawals **“withdraw P”**

Enforcing Policies (i.e., making money) Using BGP

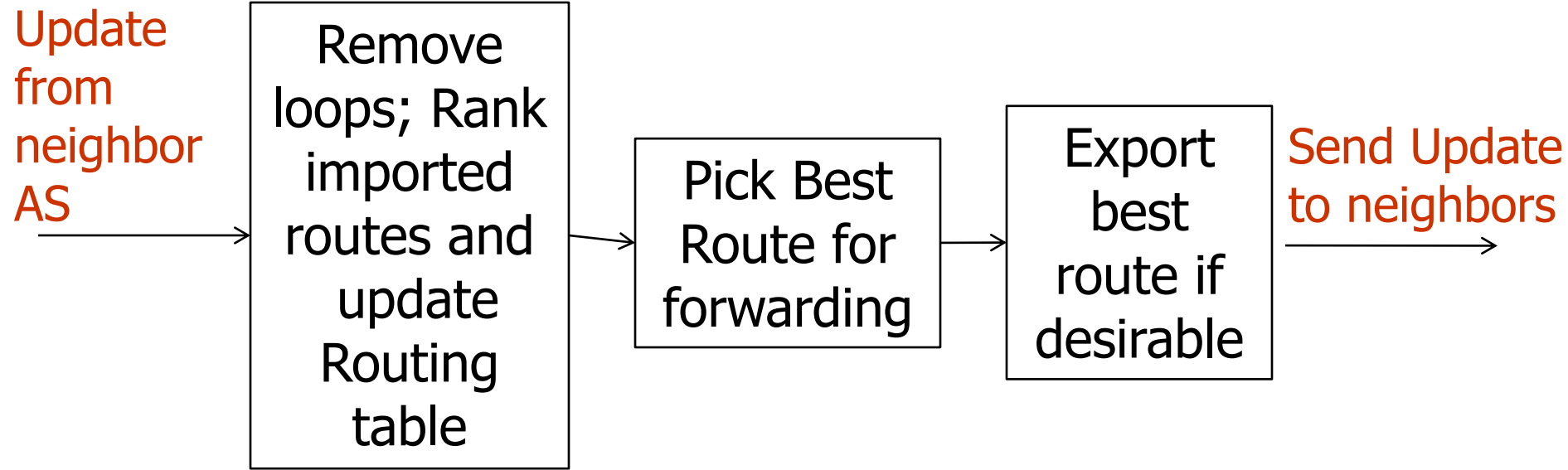
Route Export: controls incoming traffic

- AS advertises its customers (and internal prefixes) to all neighbors
- AS advertises routes learned from its peers or providers or customers to its customers (and internally)

Route Import: controls outgoing traffic

- For each dest. prefix, AS picks its best route from those in its routing table as follows:
 - Prefer route from **Customer > Peer > Provider**
 - Then, prefer route with shorter AS-Path

BGP



BGP Update Message Processing

For each destination prefix,

- Learn paths from neighbors
- Ignore loopy paths and keep the rest in your routing table
- Order paths according to AS preferences
 - Customers > peers > providers
 - Path with shorter AS hops are preferred to longer paths
 - Other metrics
- Insert the most preferred path into your forwarding table
- Announce the most preferred path to a neighbor according to policies

When you receive a withdrawal

- If path not used, remove from learned path
- If best/used path
 - Remove the path from your forwarding and routing tables and insert the next preferred path in routing table into forwarding table
 - For each neighbor decide whether to tell him about the new path based on policies
 - If yes, then announce the new path which implicitly withdraws old path
 - If no, withdraw old path

Summary

- Hierarchical addressing and routing improve scalability
- Inter-domain routing is policy-based not shortest path
 - An AS forwards transit traffic only if it makes money from it
- BGP is a path vector routing algorithm that implements policy-based routing