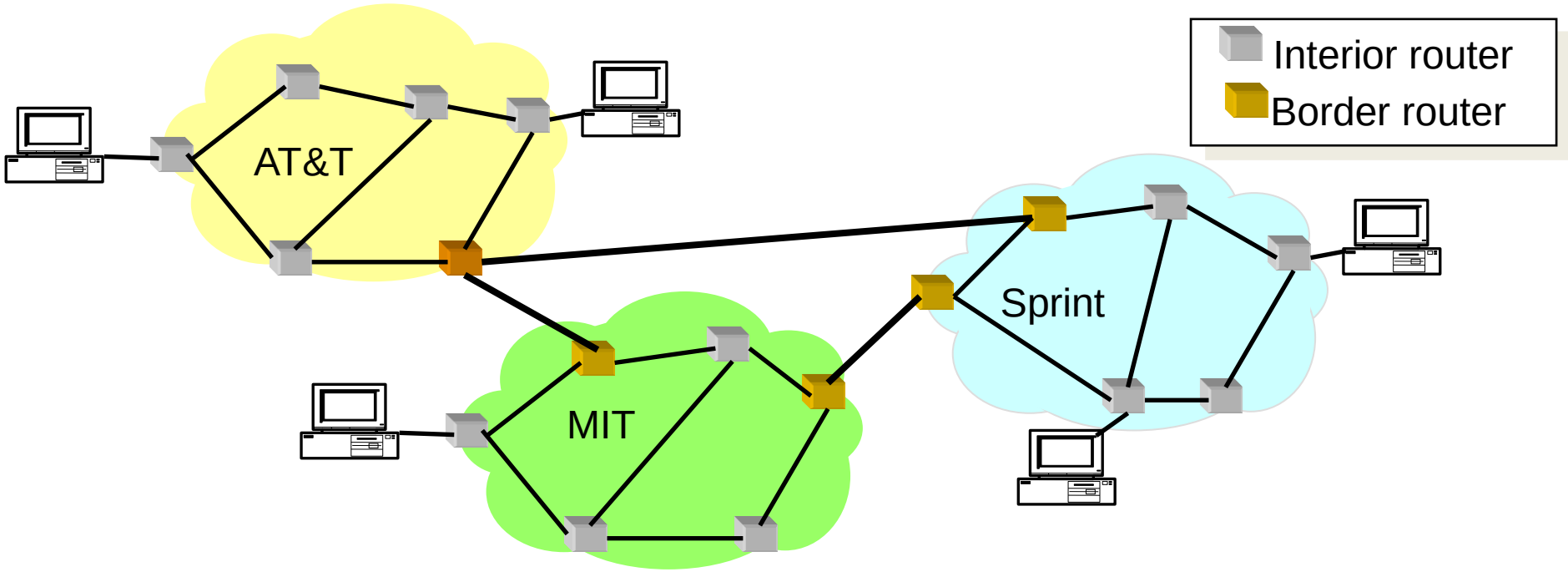
Network Layer: Internet-Wide Routing & BGP

Dina Katabi & Sam Madden

Inter-Domain Routing



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

Requirements of Internet-Wide 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

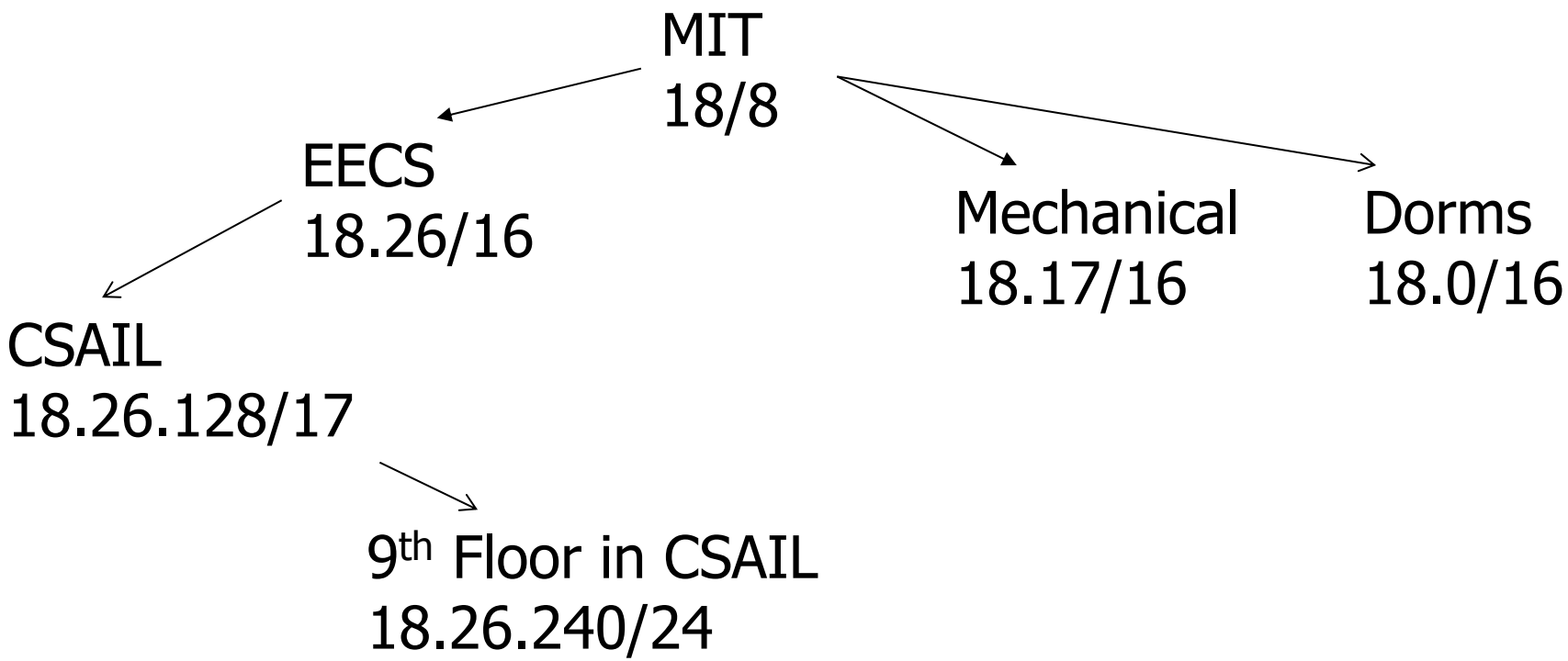
- Need less information with increasing distance to destination

→ Hierarchical Routing and Addressing

Hierarchical Addressing

- The IP address space is divided into segments of contiguous chunk of addresses; each such 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
- 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

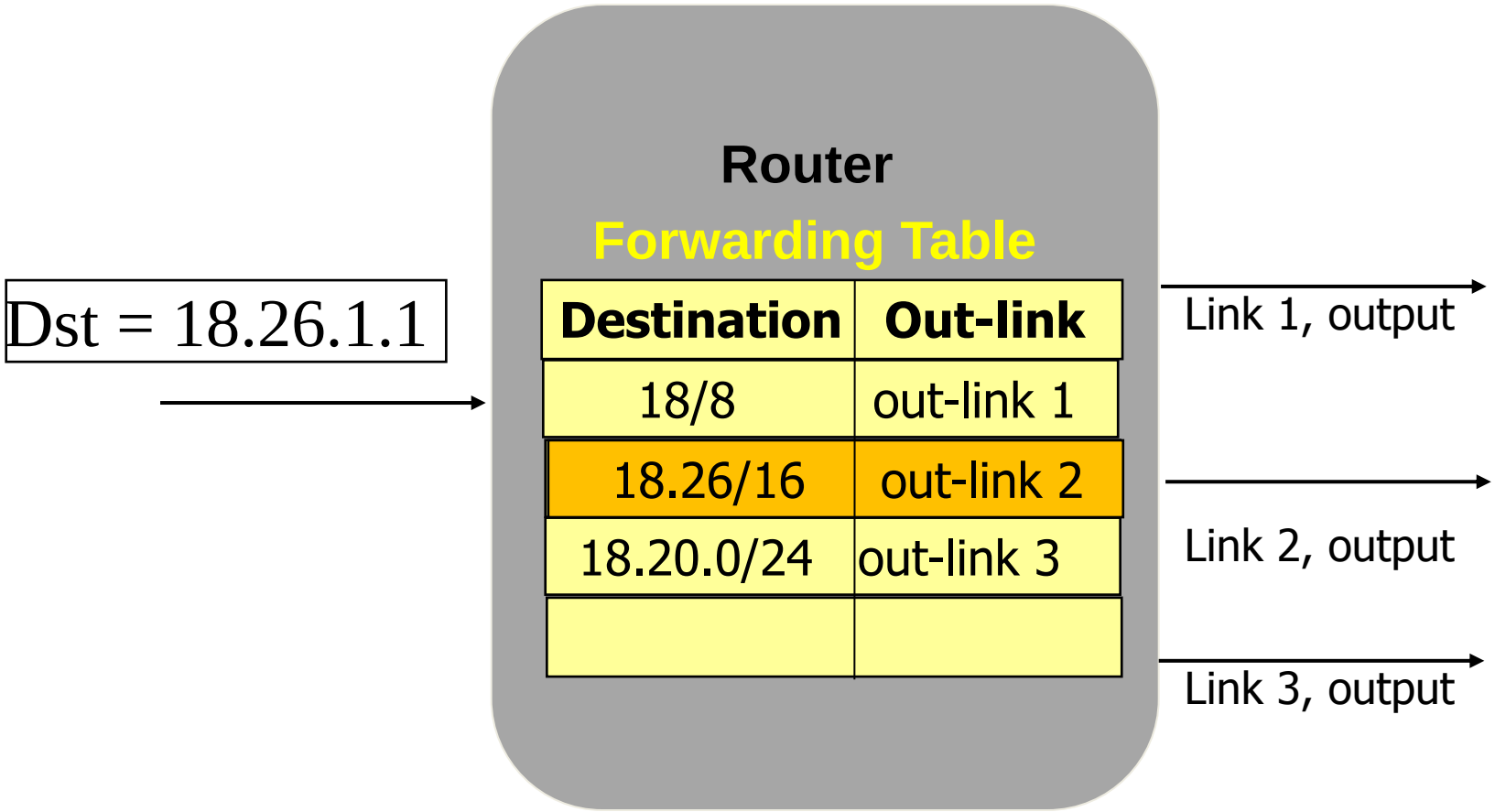
Hierarchical Addressing



- 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

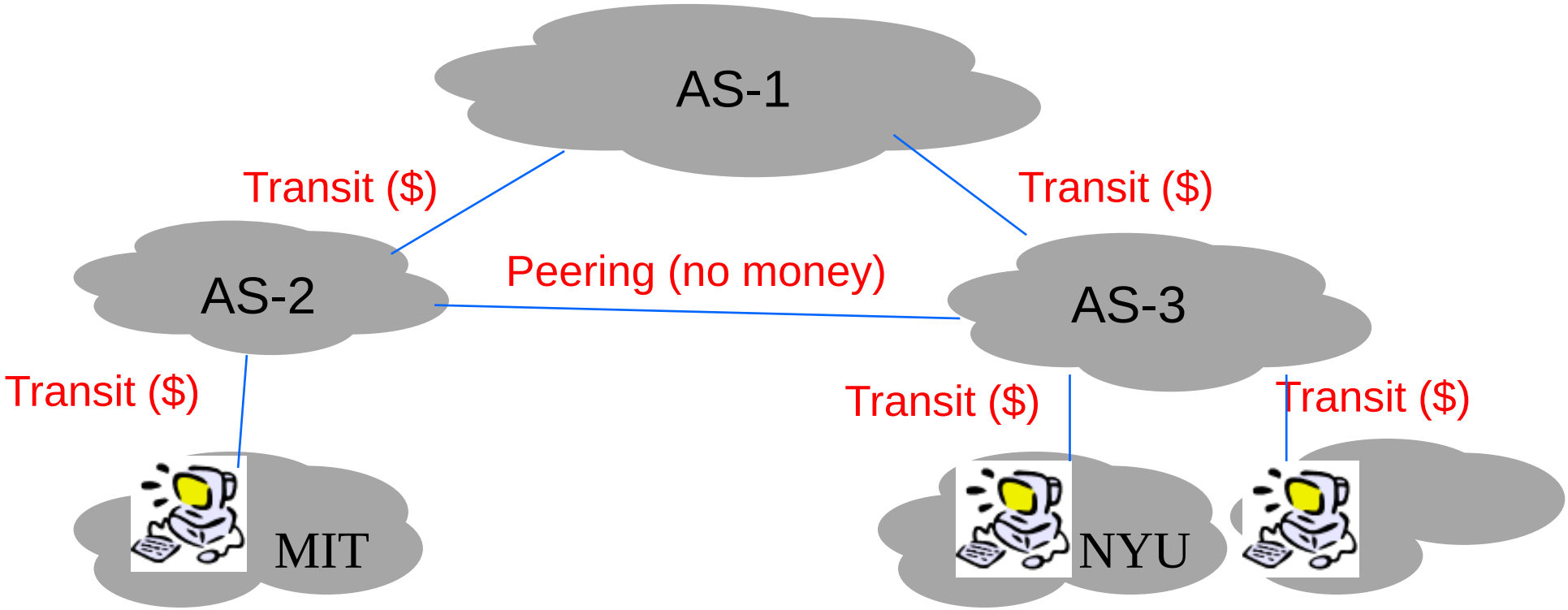
Longest Prefix Match

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

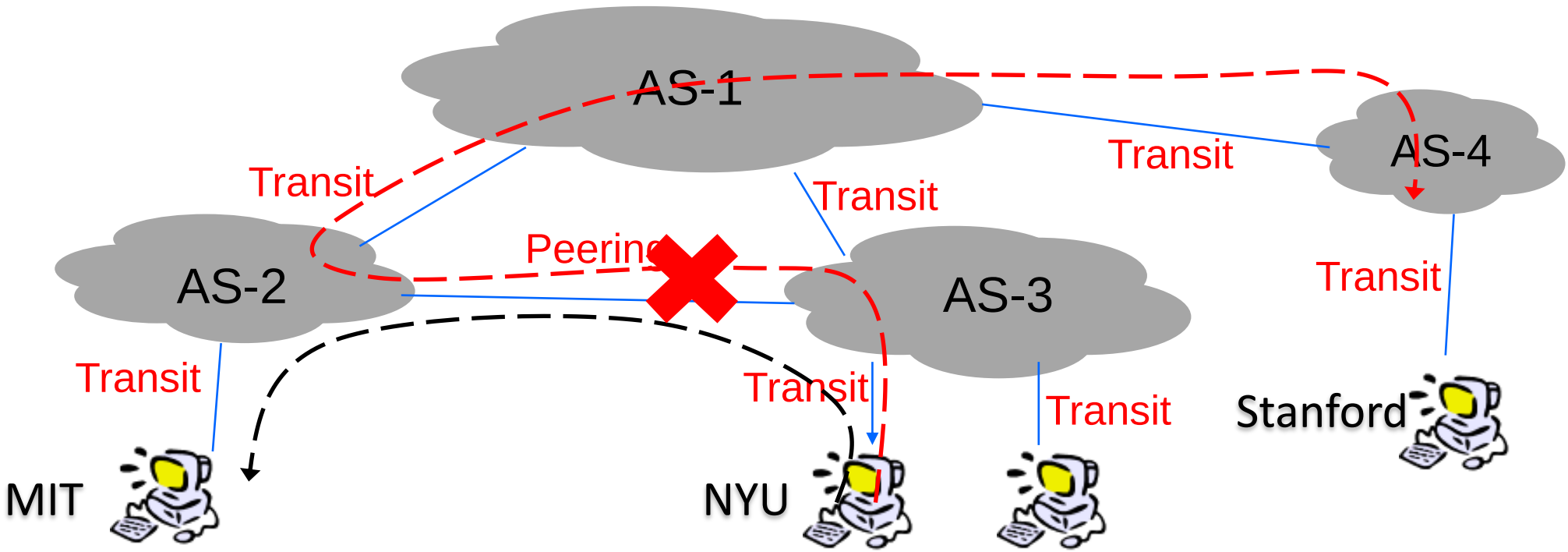


- Transit relationship
  - One AS is a customer of the other AS, who is the provider; customer pays provider both for sending and receiving packets
- Peering relationship
  - Two ASs forward packets for each other without exchanging money

Policy-Based Routing

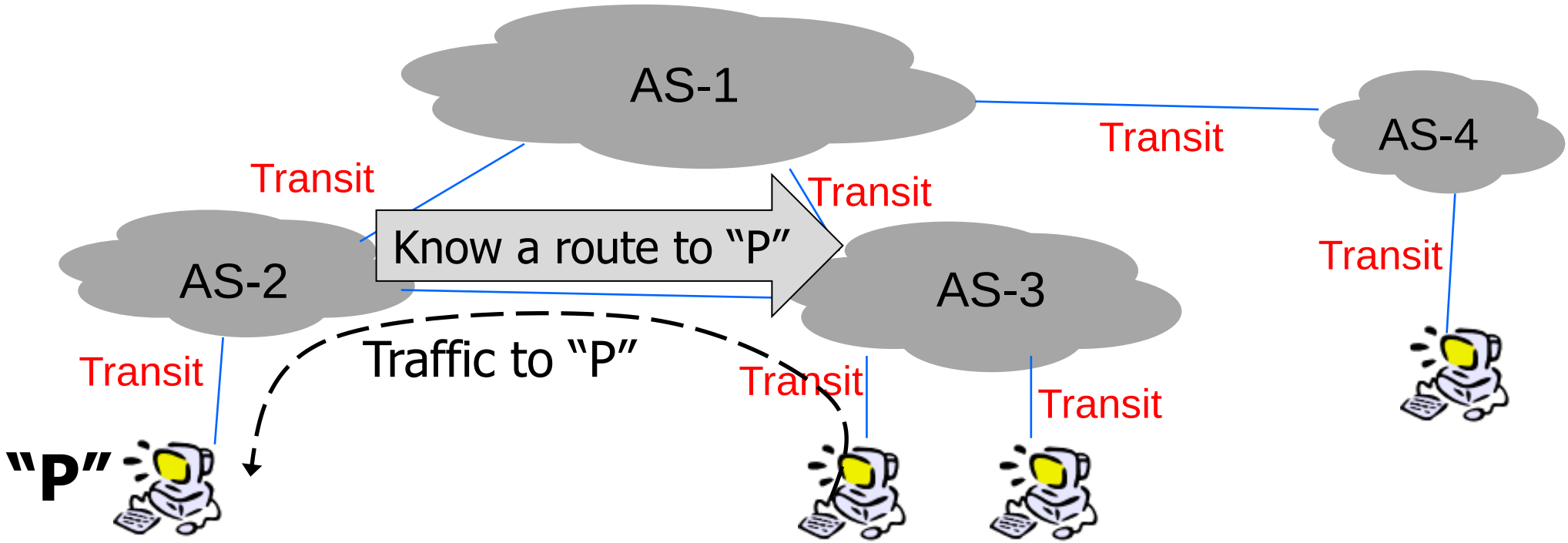
- Main Rule:
  - An AS does not transit traffic unless it makes money of it
- Note Customer pays for both incoming and outgoing traffic

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?

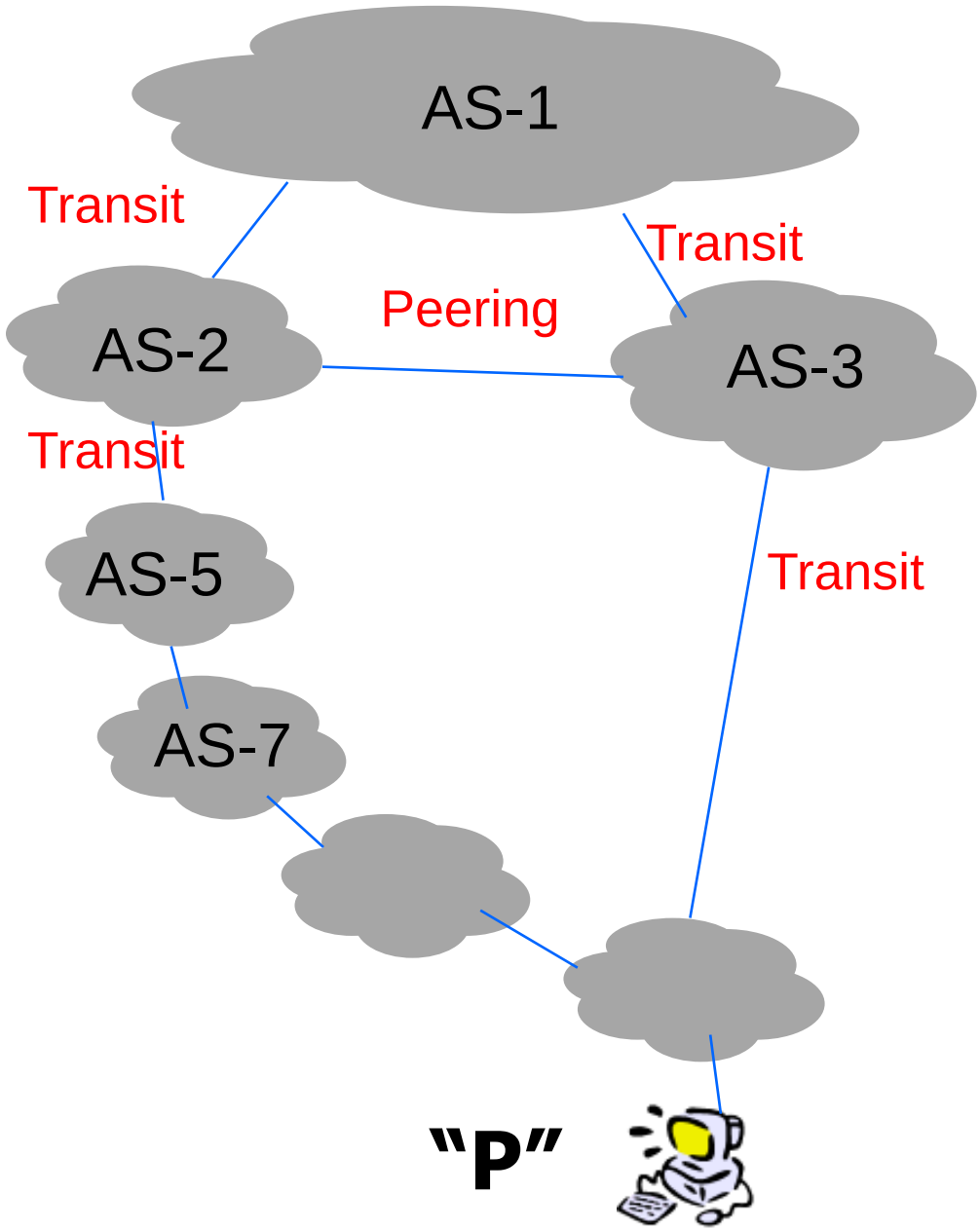


- AS-2 advertises to AS-3 a route to its customer's IP prefix "P"
- AS-2 does not tell AS-3 that it has a route to AS-4, i.e., it does not tell AS-3 routes to non-customers IP-prefixes

Desirable Outgoing Policies

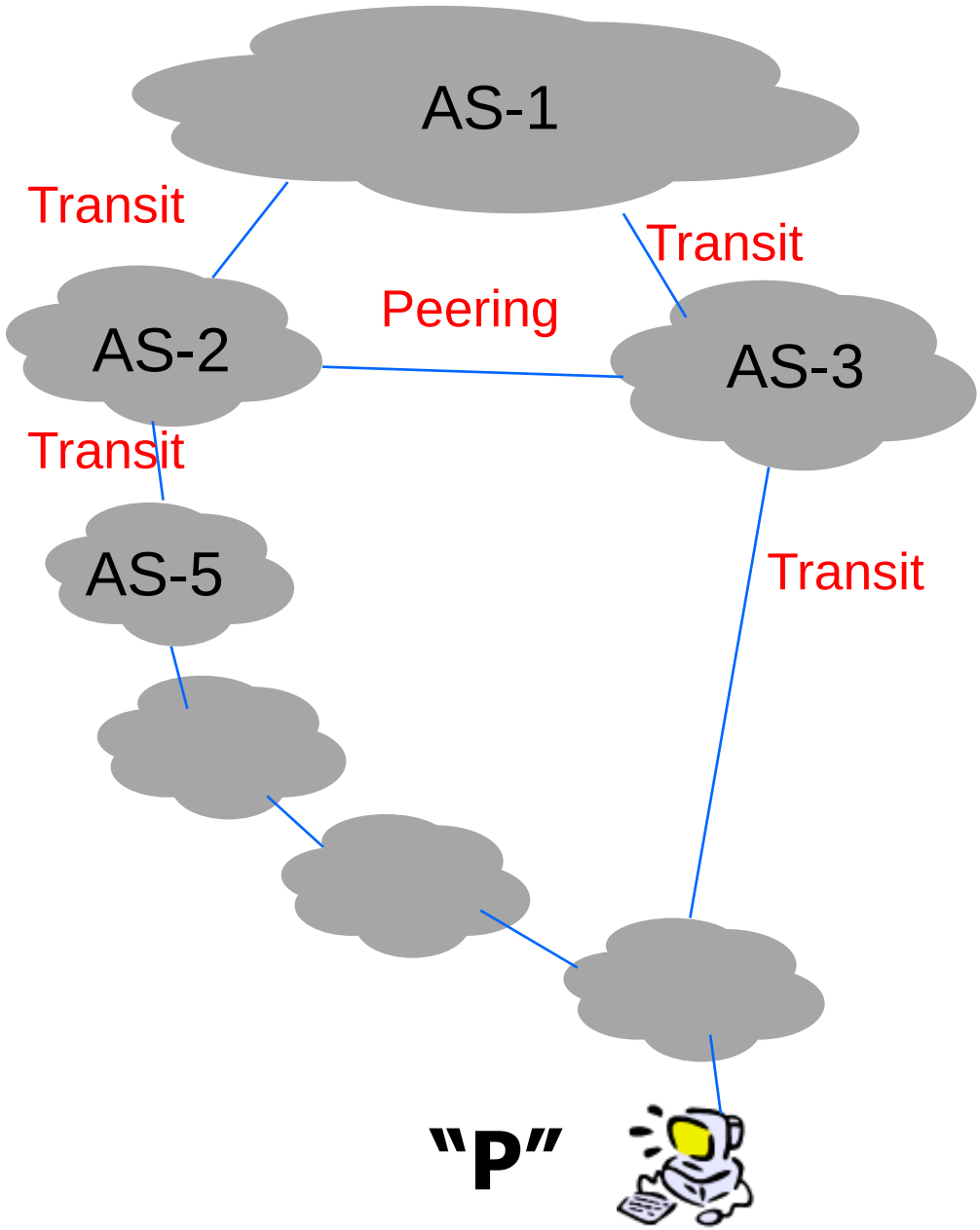
- AS-2 will hear 3 paths to “P” from its neighbors

- AS-2 prefers to send traffic to “P” 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-5 advertise their routes to "P" 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)



Enforcing Policies (i.e., making money)

Route Export: controls incoming traffic

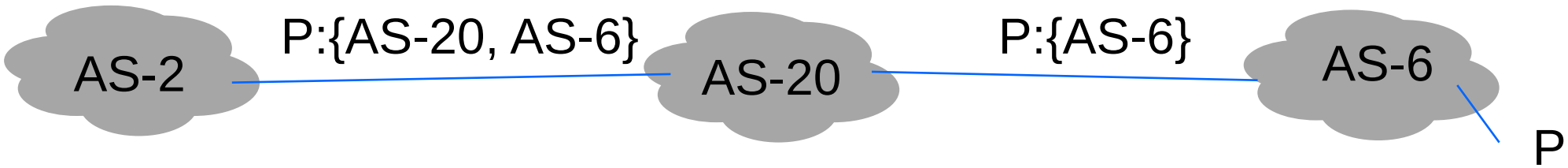
- AS advertises its customers (and internal prefixes) to all neighbors
- AS advertises all routes it uses to its customers (and internally)

Route Import: controls outgoing traffic

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

BGP: Border Gateway Protocol

1. Advertize whole path

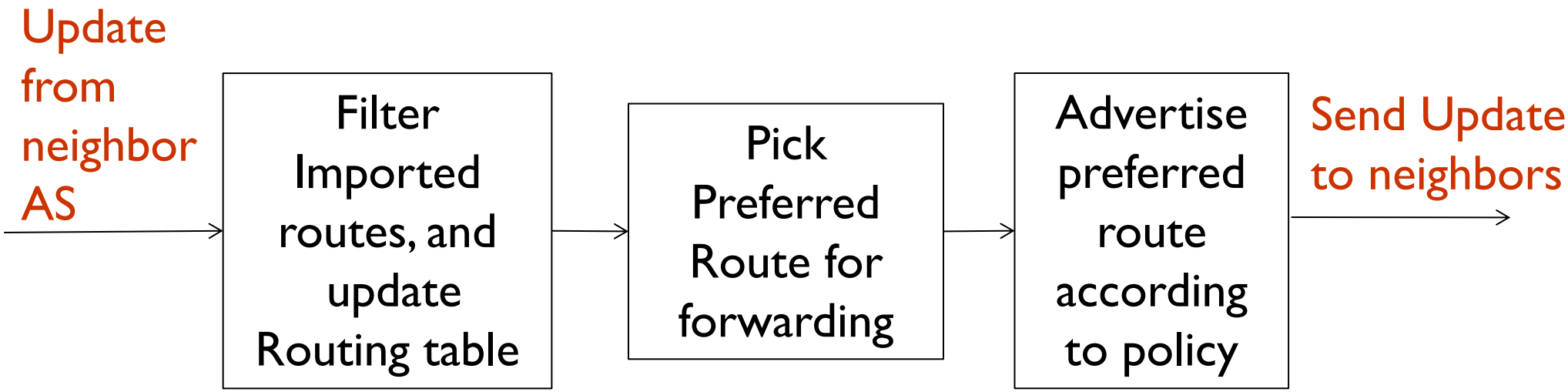


- Loop detection→ an AS checks for its own AS number in advertisement and rejects route if it has its own AS number

2. Incremental updates

- AS sends routing updates only when its preferred route changes (Messages are reliably delivered using TCP)
- Two types of update messages: advertisements, e.g., "P:{AS-20, AS-6}" and withdrawals "withdraw P"

BGP



Note: BGP Router can advertise only the preferred (i.e., currently used) route

BGP Update Message Processing

When AS receives an advertisement,

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
- Insert the most preferred path into your forwarding table
- Advertise the most preferred path to a neighbor according to policies

When AS receives a withdrawal

- If withdrawn path not used/preferred, remove from routing table
- If withdrawn path is used –i.e., preferred
  - Remove the path from forwarding table and routing table
  - insert the next preferred path from the routing table into forwarding table
  - For each neighbor decide whether to tell him about the new path based on policies
    - If yes, advertise the new path which implicitly withdraws the old path for the corresponding prefix
    - If no, withdraw old path

Summary

- Hierarchical addressing and hierarchical 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