Network

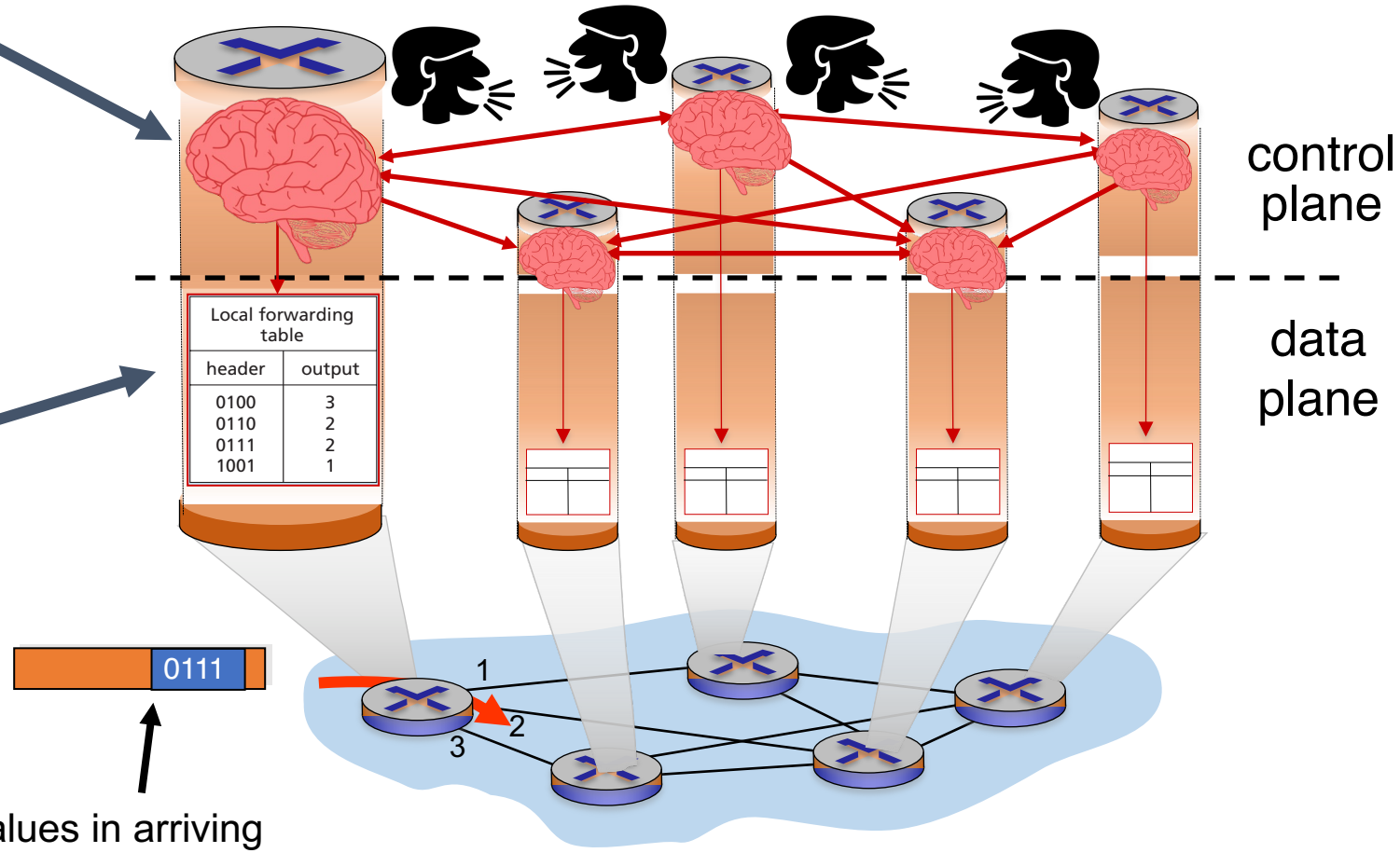
Ack: Slides heavily adapted from Jen Rexford and Michael Schapira

RUTGERS
UNIVERSITY | NEW BRUNSWICK

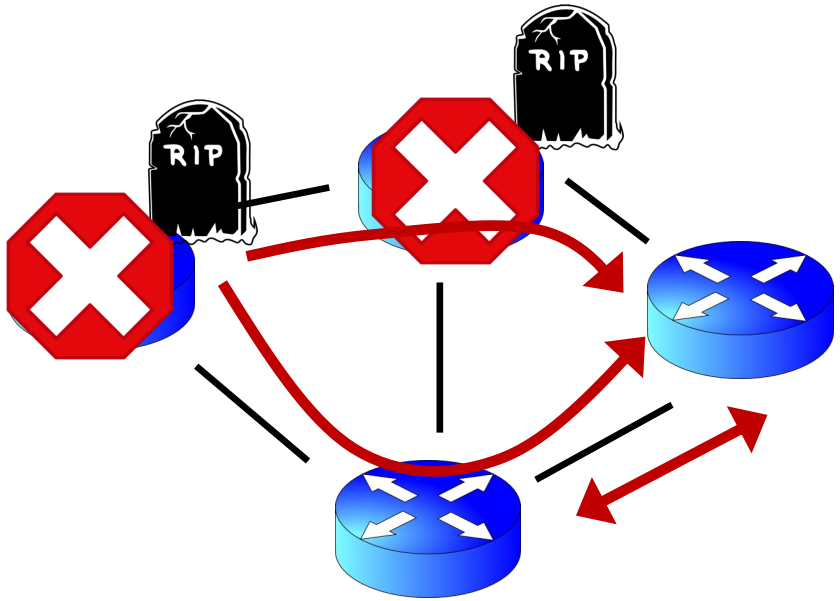
Per-router control plane

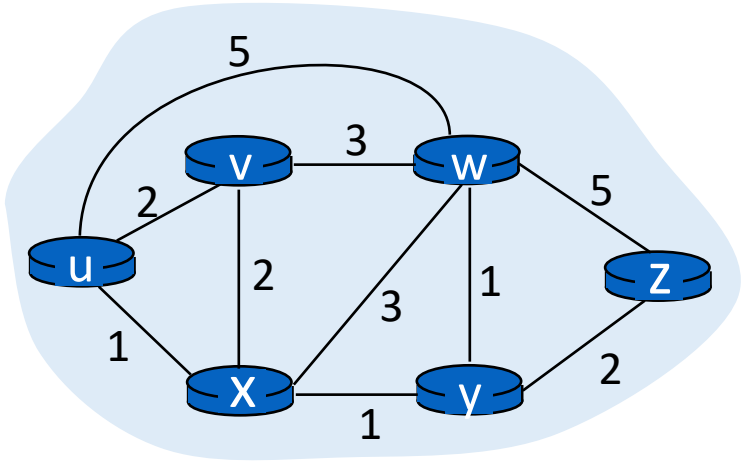
Distributed control plane: Components in every router interact with other components to produce a routing outcome.

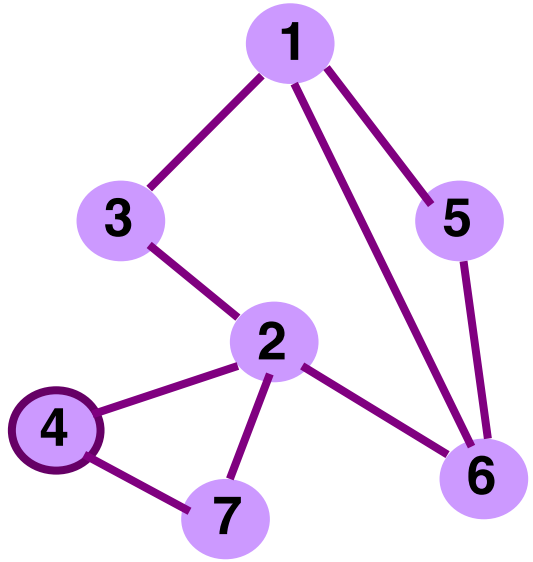
Data plane per-packet processing, moving packet from input port to output port



So far

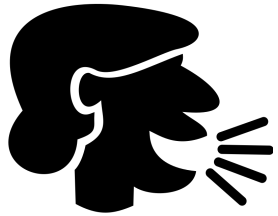




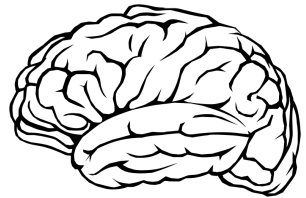


- What outcome is computed? Spanning tree, shortest paths, …
- What algorithm is used?

Routing protocol



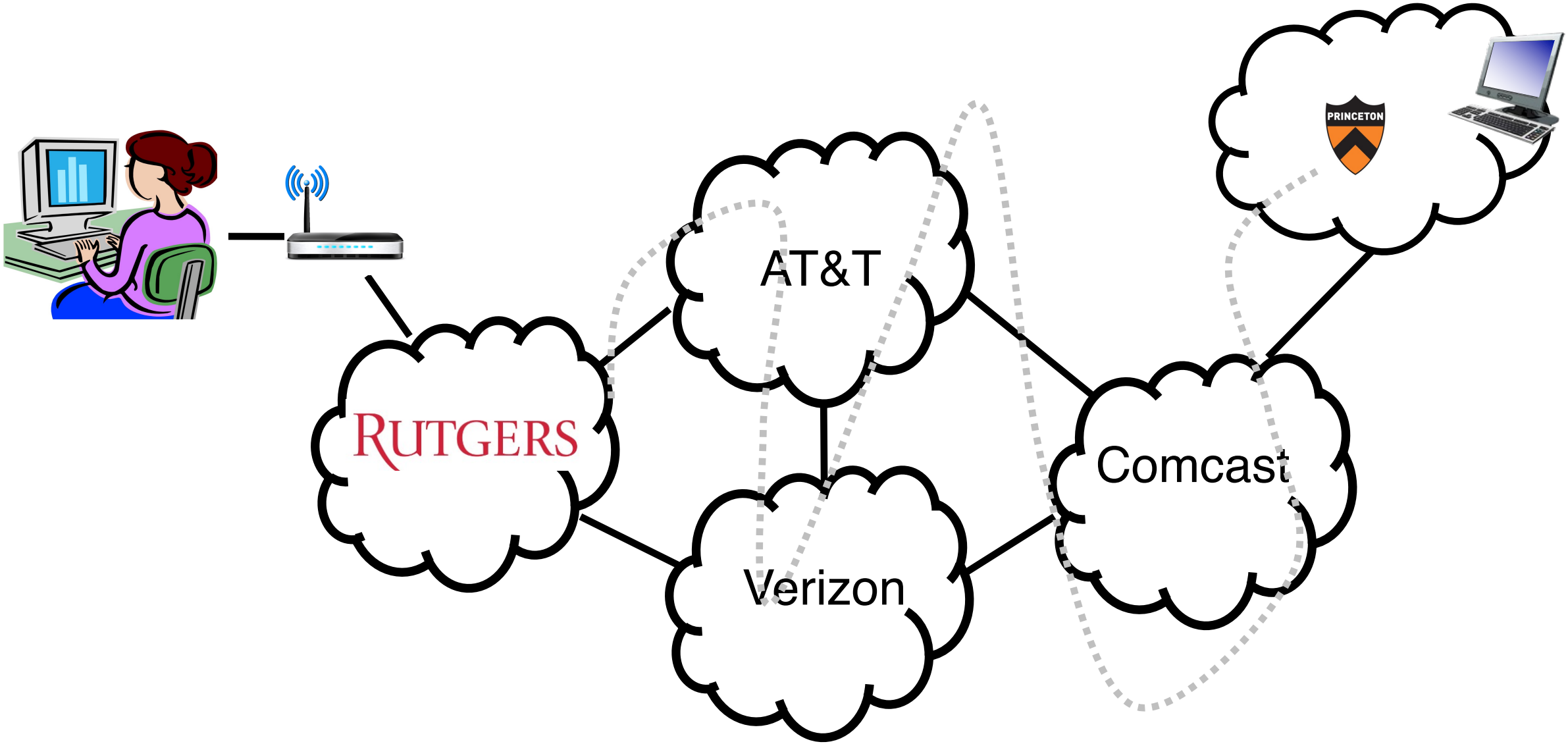
Q1. What info exchanged?



Q2. What computation?

Internet Routing

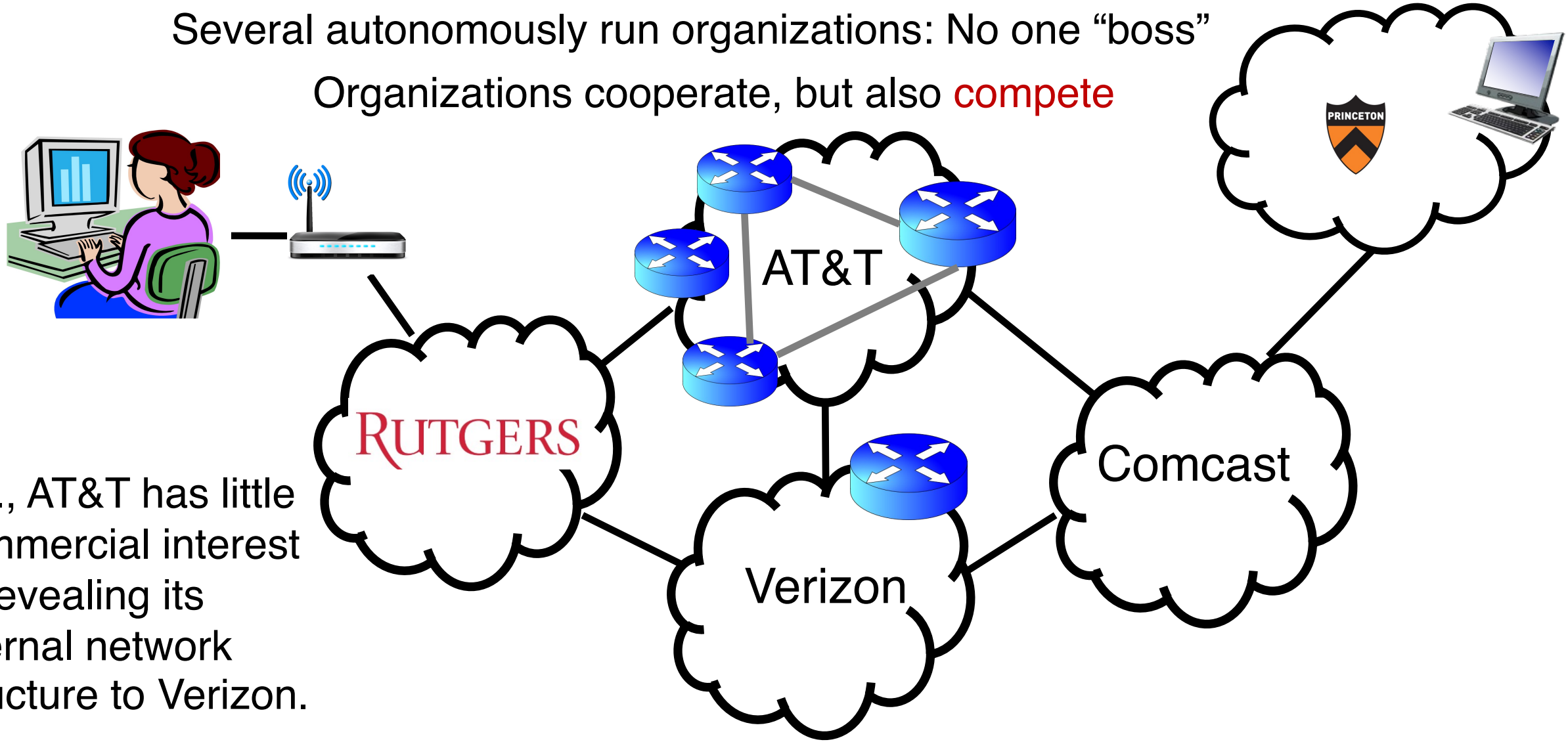
The Internet is a large federated network



The Internet is a large federated network

Several autonomously run organizations: No one “boss”

Organizations cooperate, but also compete

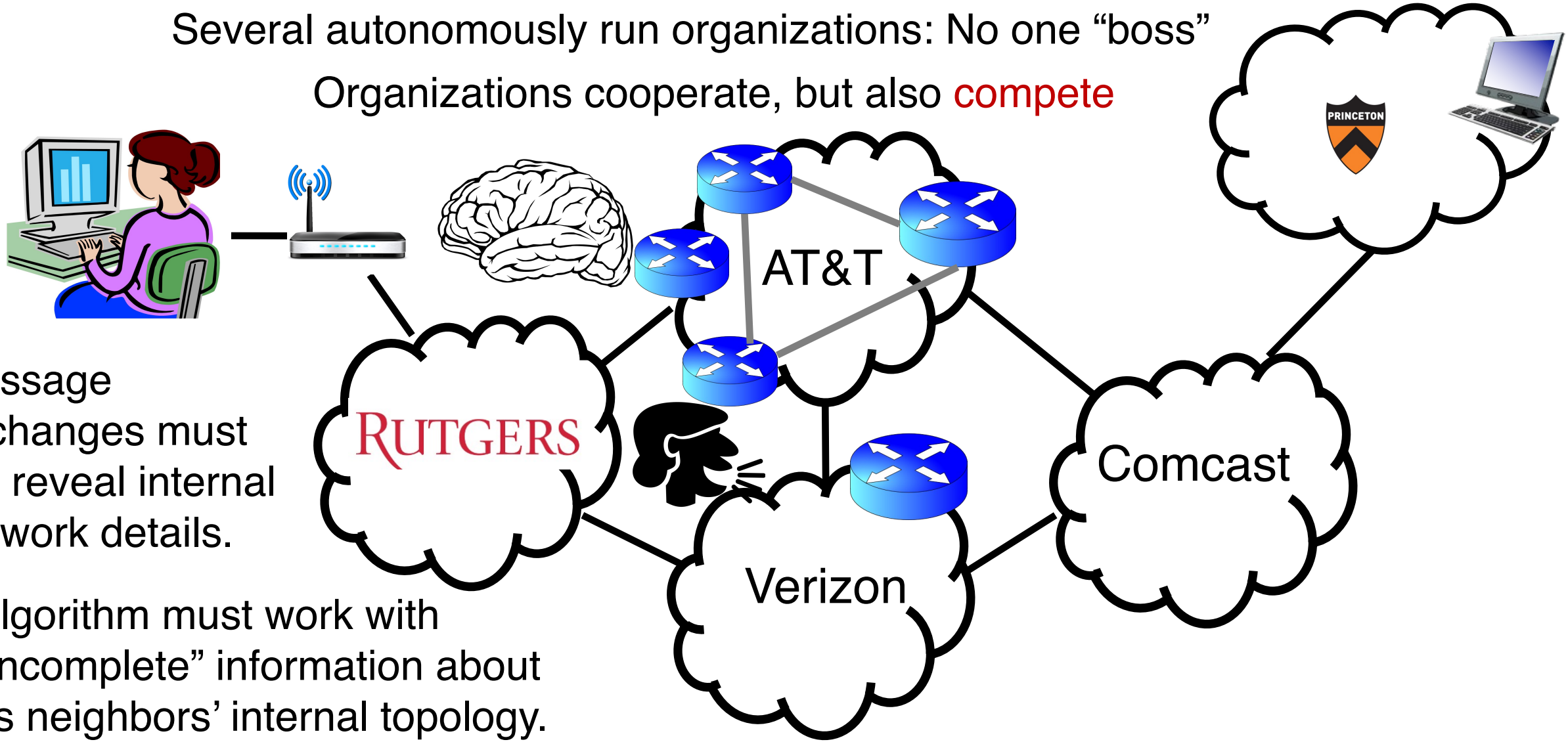


e.g., AT&T has little commercial interest in revealing its internal network structure to Verizon.

The Internet is a large federated network

Several autonomously run organizations: No one “boss”

Organizations cooperate, but also compete



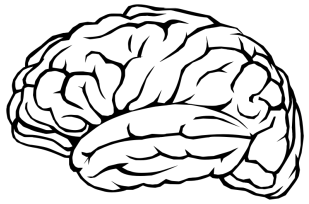
Message exchanges must not reveal internal network details.

Algorithm must work with “incomplete” information about its neighbors’ internal topology.

The Internet is a large federated network

Internet today: > 70,000 unique autonomous networks

Internet routers: > 800,000 forwarding table entries



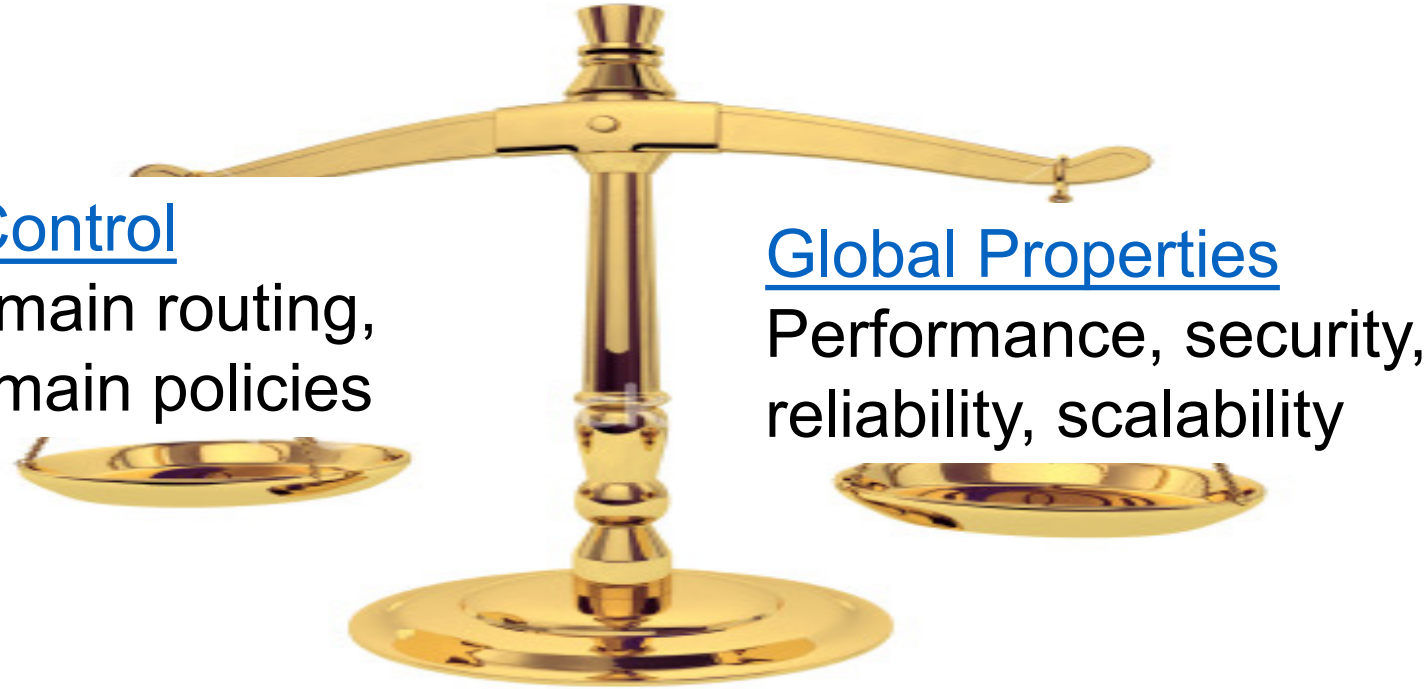
Keep messages & tables as small as possible. Don't flood

Algorithm must be incremental: don't recompute the whole table on every message exchanged.

Local Control vs. Global Properties

The Internet is a “network of networks”

- ~35,000 separately administered networks
- Competitive cooperation for e2e reachability



Local Control
Intradomain routing, interdomain policies

Global Properties
Performance, security, reliability, scalability

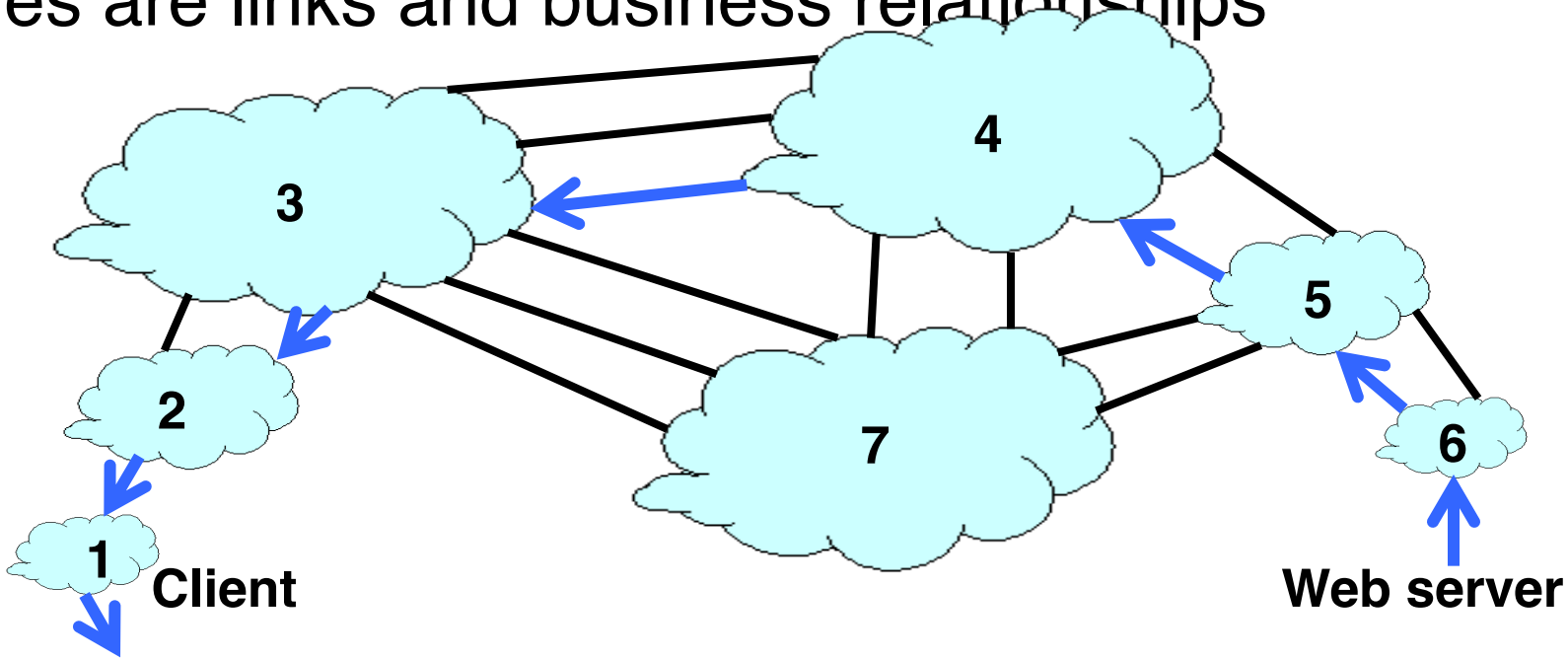
Two-Tiered Routing Architecture

- Goal: distributed management of resources
  - Internetworking of multiple networks
  - Networks under separate administrative control
- *Intra*domain: inside a region of control
  - Routers configured to achieve a common goal
  - Okay for routers to share *topology* information
  - Different ASes can run different protocols
- *Inter*domain: between regions of control
  - ASes have different (maybe conflicting) goals
  - Routers only share *reachability* information

Internet Structure

Autonomous Systems (ASes)

- AS-level topology
  - Nodes are Autonomous Systems (ASes)
  - Destinations are prefixes (e.g., 12.0.0.0/8)
  - Edges are links and business relationships



AS Numbers (ASNs)

ASNs are 16 bit values (or 32-bit).

64512 through 65535 are "private"

Currently > 70,000 in use.

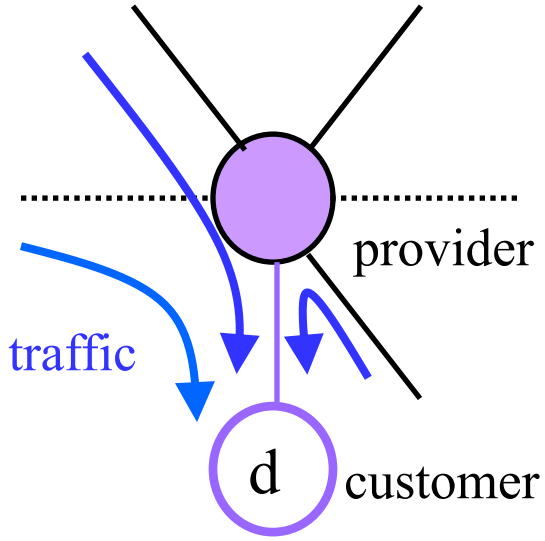
- **Level 3: 1**
- **MIT: 3**
- **Harvard: 11**
- **Rutgers:**
- **Princeton: 88**
- **AT&T: 7018, 6341, 5074, …**
- **Verizon: 701, 702, 284, 12199, …**
- **Sprint: 1239, 1240, 6211, 6242, …**
- **…**

Business Relationships Between ASes

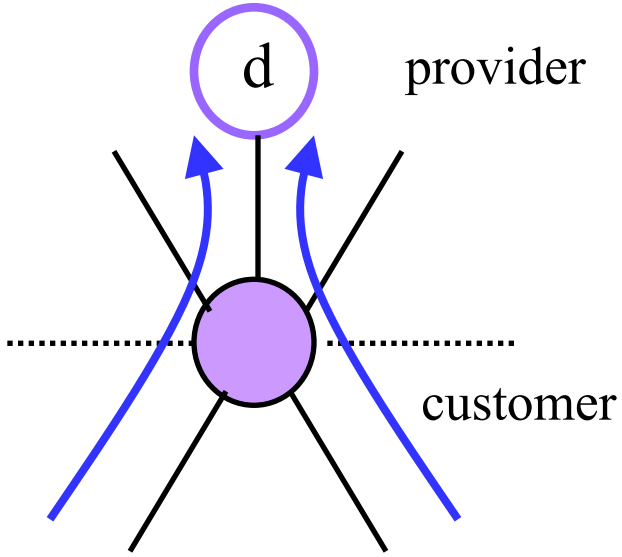
- Neighboring ASes have business contracts
  - How much traffic to carry
  - Which destinations to reach
  - How much money to pay
- Common business relationships
  - Customer-provider
  - Peer-peer

Customer-Provider Relationship

- Customer needs to be reachable from everyone
  - Provider ensures all neighbors can reach the customer
  - Customer “default-routes” to provider
- Customer does not want to provide transit service
  - Customer does not let its providers send traffic through it



Traffic **to** the customer

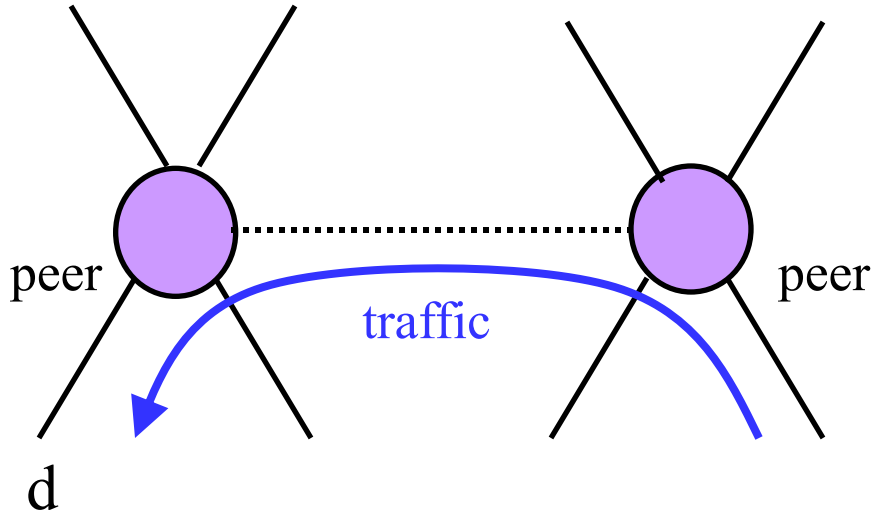


Traffic **from** the customer

Peer-Peer Relationship

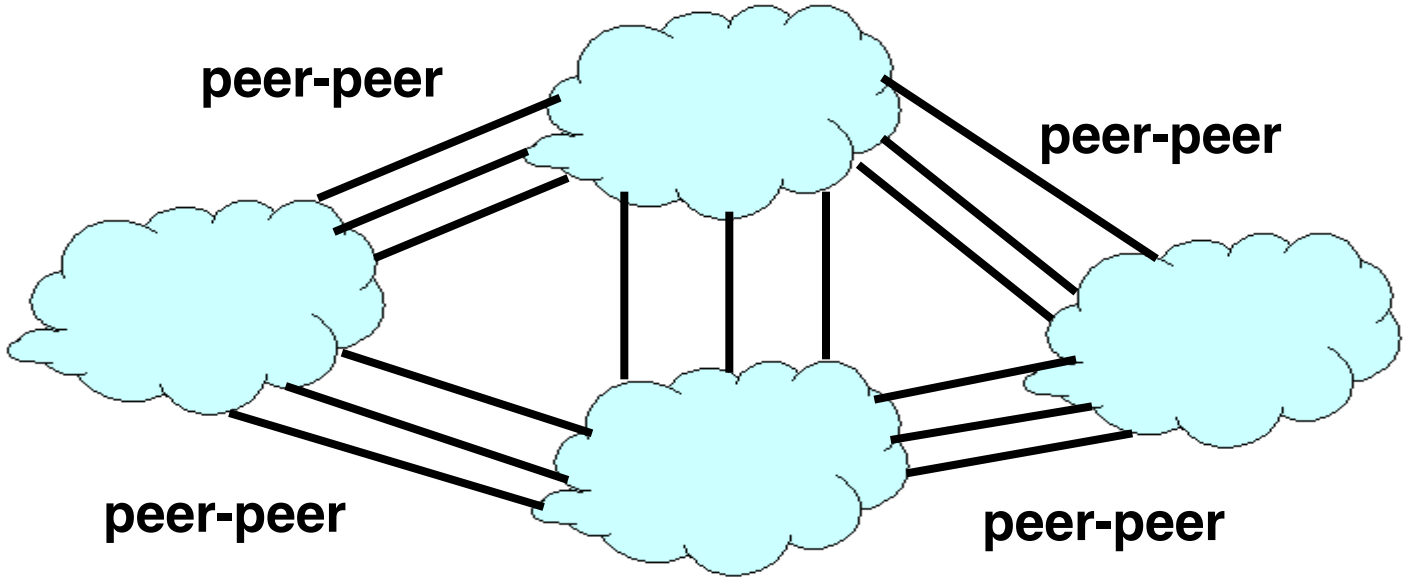
- Peers exchange traffic between customers
  - AS lets its peer reach (only) its customers
  - AS can reach its peer's customers
  - Often the relationship is settlement-free (i.e., no $$$)

Traffic to/from the peer and its customers



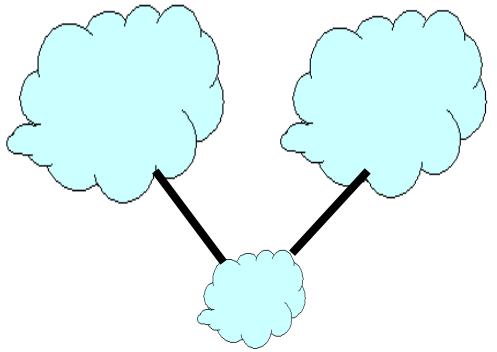
AS Structure: Tier-1 Providers

- Top of the Internet hierarchy
  - Has no upstream provider of its own, no default routes
  - Typically has a large (inter)national backbone
  - Around 10-12 ASes: AT&T, Sprint, Level 3, …



AS Structure: Other ASes

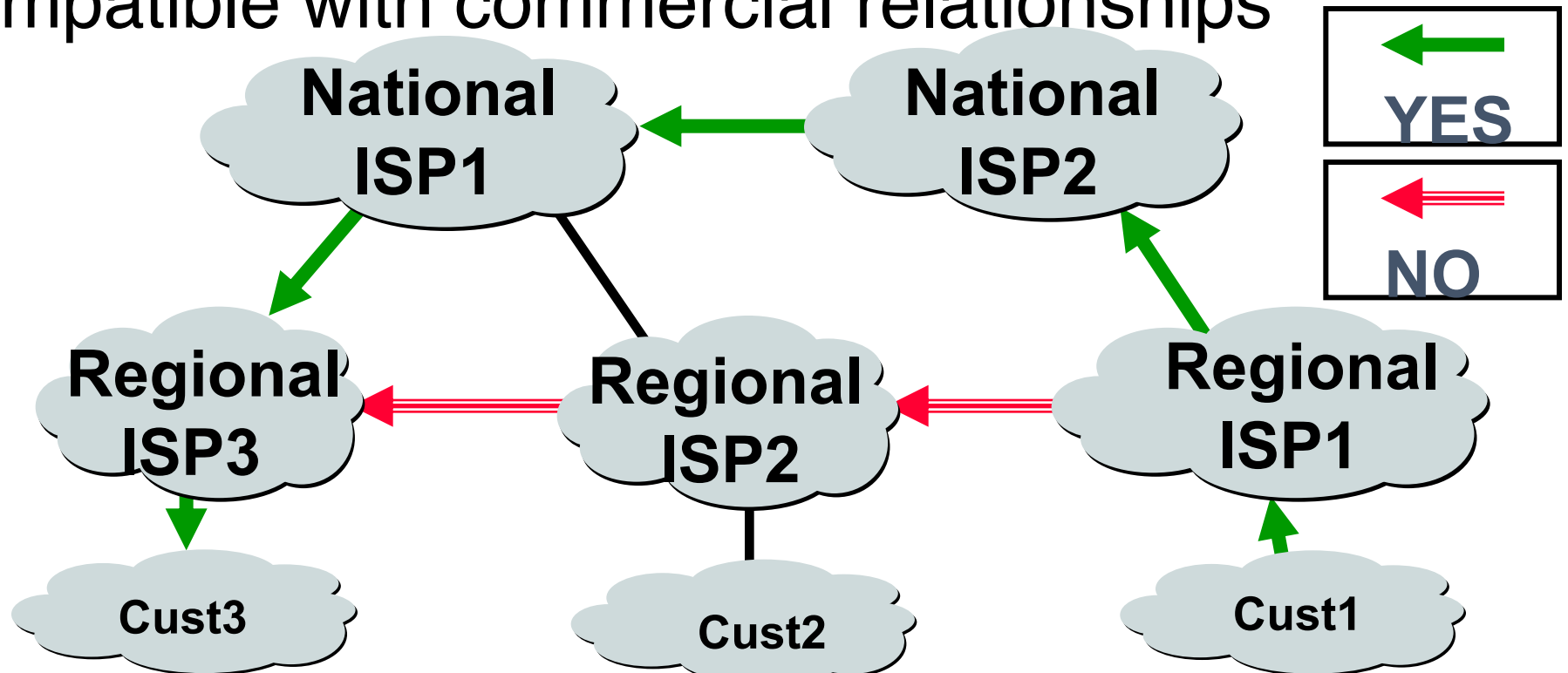
- Lower-layer providers (tier-2, …)
  - Provide transit service to downstream customers
    - But need at least one provider of their own
  - Typically have national or regional scope
    - E.g., Minnesota Regional Network
- Stub ASes
  - Do not provide transit service
  - Connect to upstream provider(s)
  - Most ASes (e.g., 85-90%)
  - E.g., Princeton, Rutgers, …



Policy-Based Path-Vector Routing

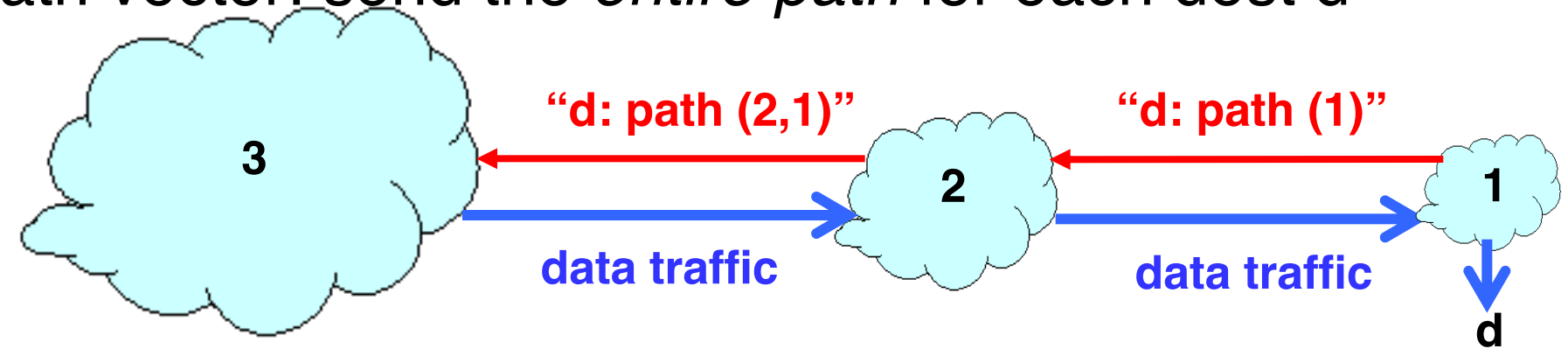
Shortest-Path Routing is Restrictive

- All traffic must travel on shortest paths
- All nodes need common notion of link costs
- Incompatible with commercial relationships



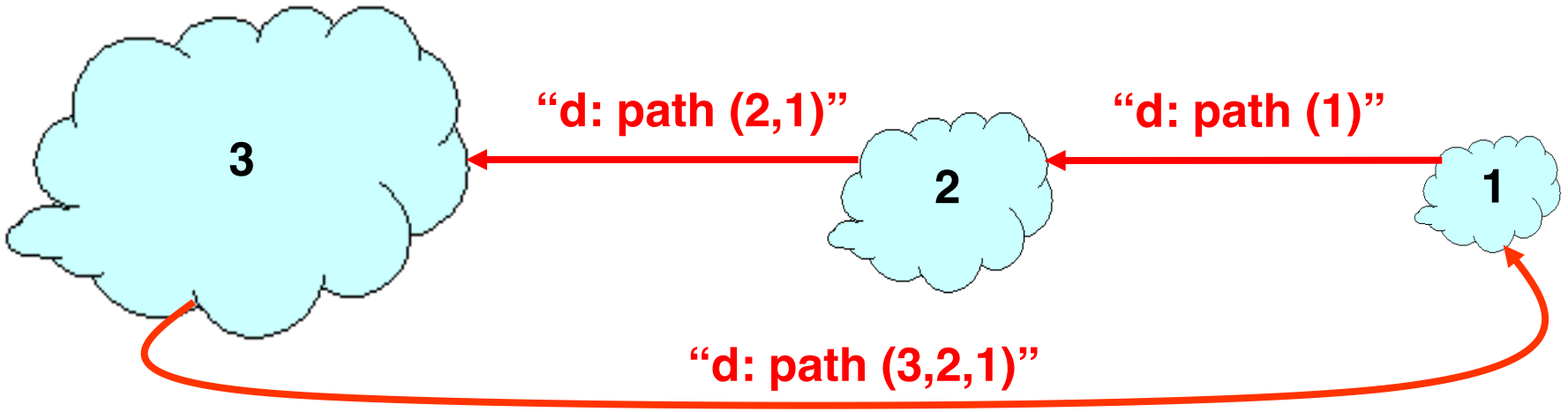
Path-Vector Routing

- Extension of distance-vector routing
  - Support flexible routing policies
  - Faster convergence(avoid count-to-infinity)
- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per dest d
  - Path vector: send the *entire path* for each dest d



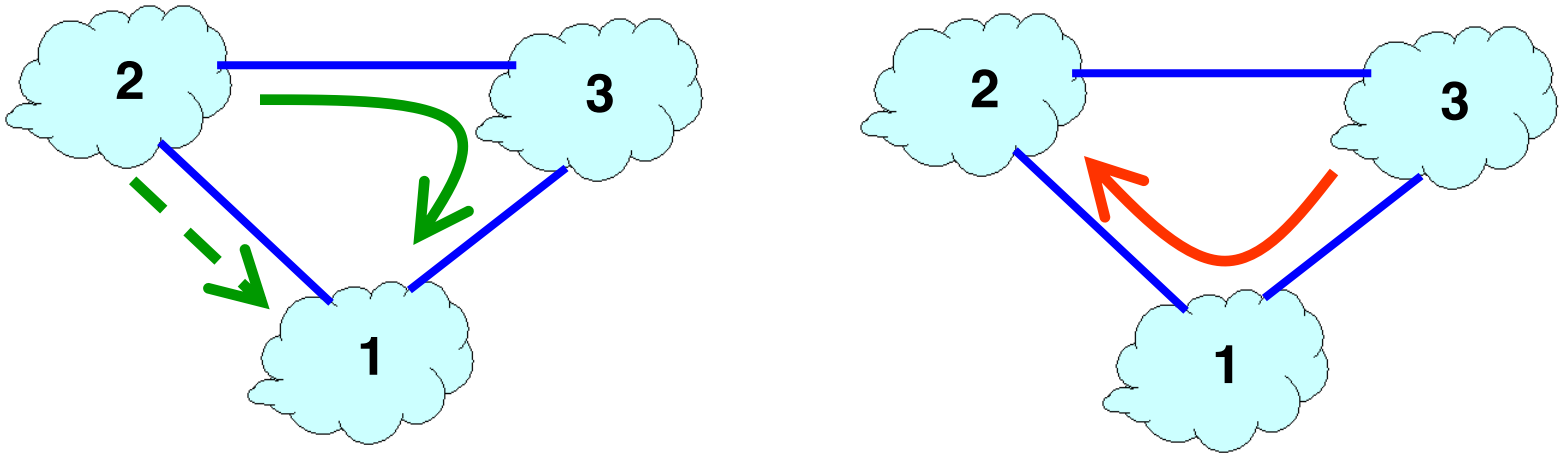
Faster Loop Detection

- Node can easily detect a loop
  - Look for its own node identifier in the path
  - E.g., node 1 sees itself in the path “3, 2, 1”
- Node can simply discard paths with loops
  - E.g., node 1 simply discards the advertisement



Flexible Policies

- Each node can apply local policies
  - Path selection: Which path to use?
  - Path export: Whether to advertise the path?
- Examples
  - Node 2 may prefer the path “2, 3, 1” over “2, 1”
  - Node 1 may not let node 3 hear the path “1, 2”



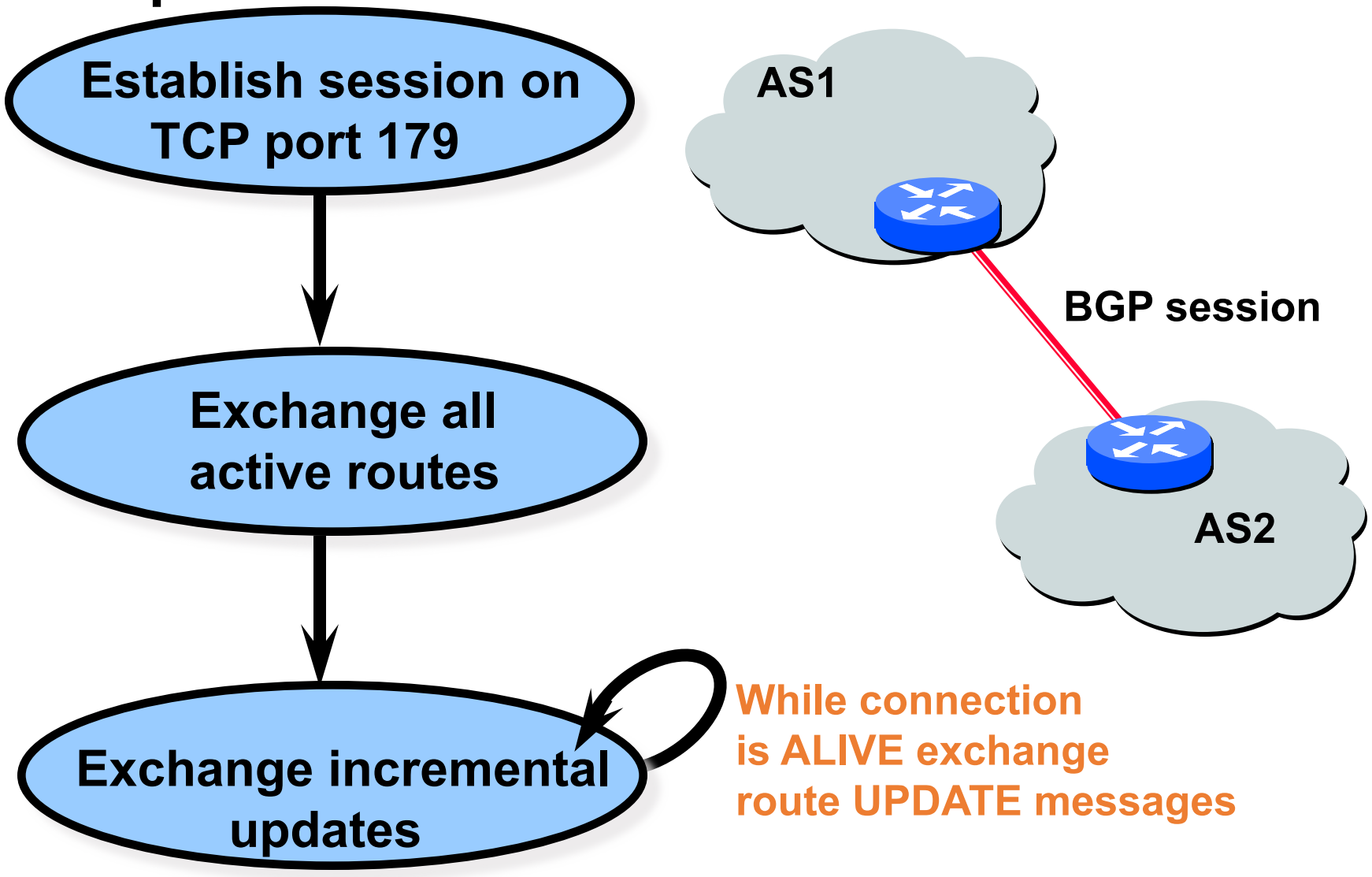
Border Gateway Protocol

Border Gateway Protocol

- IP-Prefix-based path-vector protocol
- Policy-based routing based on AS Paths
- Evolved during the past 40 years
  - **1989 : BGP-1 [RFC 1105], replacement for EGP**
  - **1990 : BGP-2 [RFC 1163]**
  - **1991 : BGP-3 [RFC 1267]**
  - **1995 : BGP-4 [RFC 1771], support for CIDR**
  - **2006 : BGP-4 [RFC 4271], update**

“BGP at 18”: http://www.youtube.com/watch?v=HAOVNYSnL7k

BGP Operations

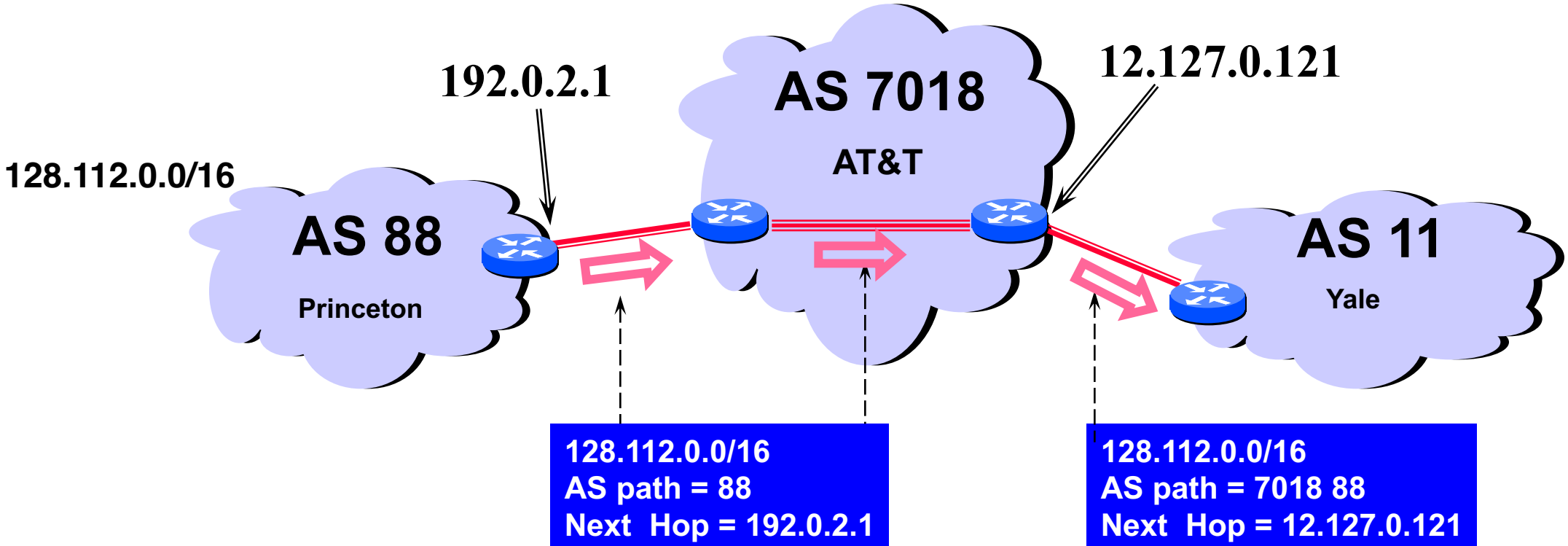


Incremental Protocol

- A node learns multiple paths to destination
  - Stores all of the routes in a routing table (RIB)
  - Applies policy to select a single active route (FIB)
  - … and may advertise the route to its neighbors
- Incremental updates
  - Announcement
    - Upon selecting a new active route, add node id to path
    - … and (optionally) advertise to each neighbor
  - Withdrawal
    - If the active route is no longer available
    - … send a withdrawal message to the neighbors

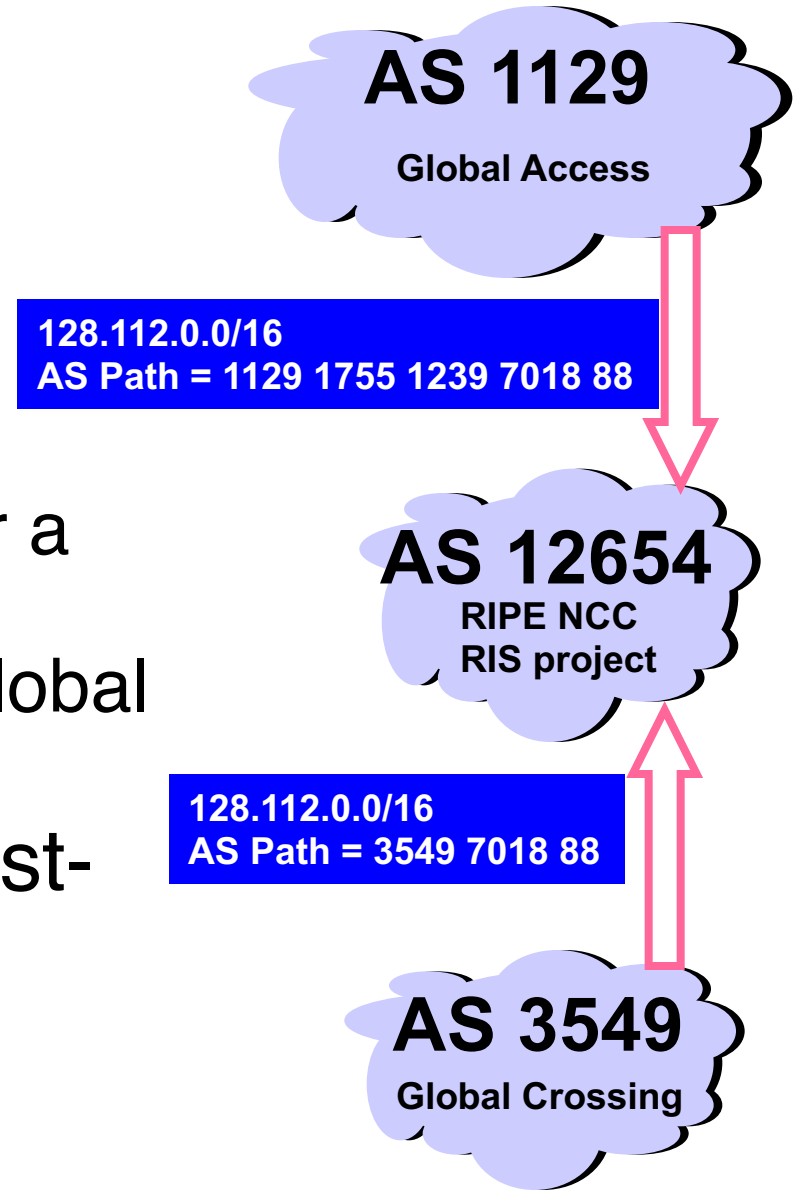
BGP Route

- Destination prefix (e.g., 128.112.0.0/16)
- Route attributes, including
  - AS path (e.g., “7018 88”)
  - Next-hop IP address (e.g., 12.127.0.121)



BGP Path Selection

- Simplest case
  - Shortest AS path
  - Arbitrary tie break
- Example
  - Three-hop AS path preferred over a five-hop AS path
  - AS 12654 prefers path through Global Crossing
- But, BGP is not limited to shortest-path routing
  - Policy-based routing



BGP Policy: Influencing Decisions

Open ended programming.
Constrained only by vendor configuration language

Receive BGP Updates

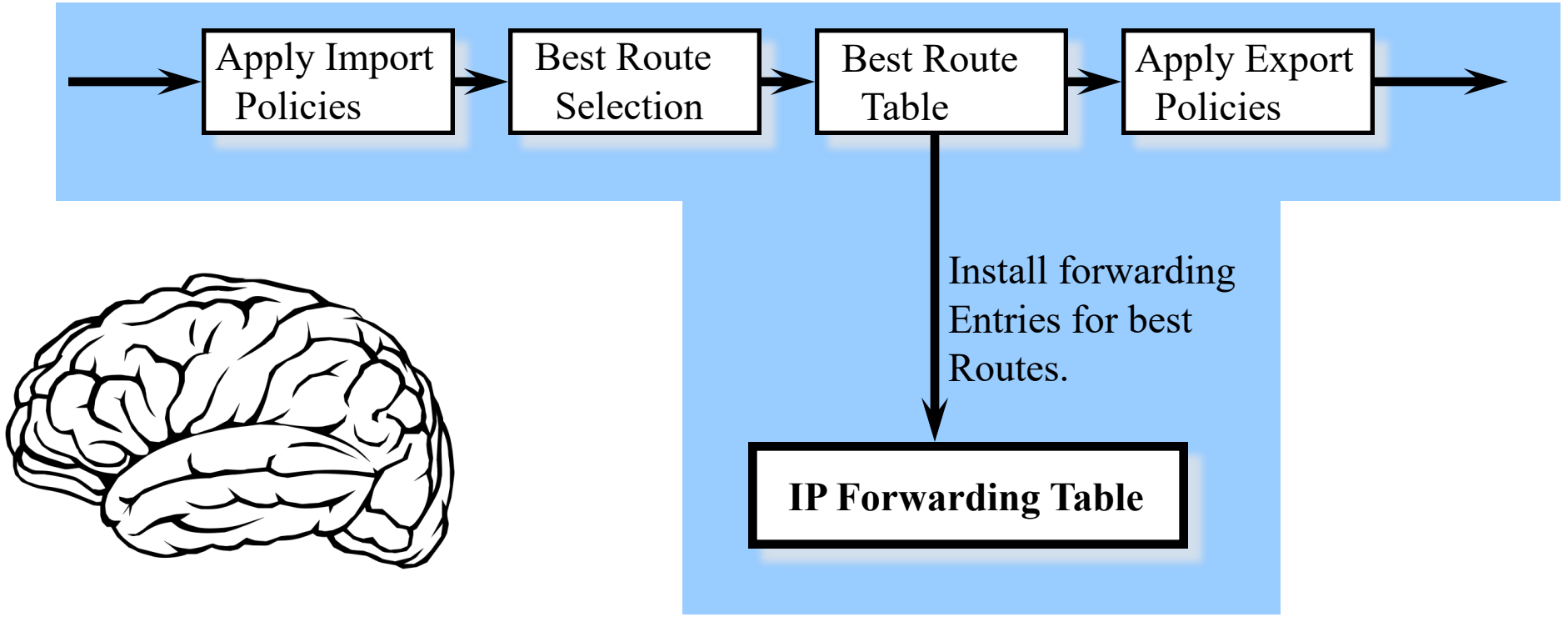
Apply Policy = filter routes & tweak attributes

Based on Attribute Values

Best Routes

Apply Policy = filter routes & tweak attributes

Transmit BGP Updates



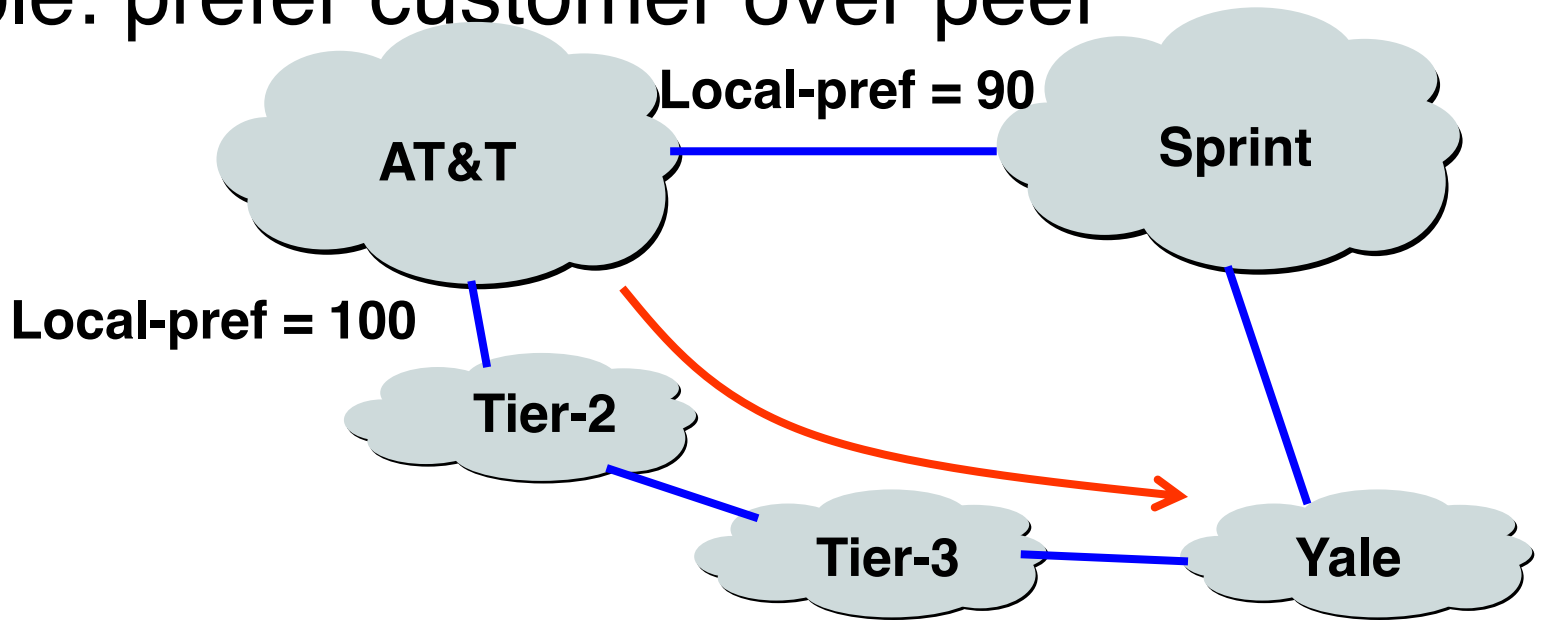
BGP Policy: Applying Policy to Routes

- Import policy
  - Filter unwanted routes from neighbor
    - E.g. prefix that your customer doesn't own
  - Manipulate attributes to influence path selection
    - E.g., assign local preference to favored routes
- Export policy
  - Filter routes you don't want to tell your neighbor
    - E.g., don't tell a peer a route learned from other peer
  - Manipulate attributes to control what they see
    - E.g., make a path look artificially longer than it is

BGP Policy Examples

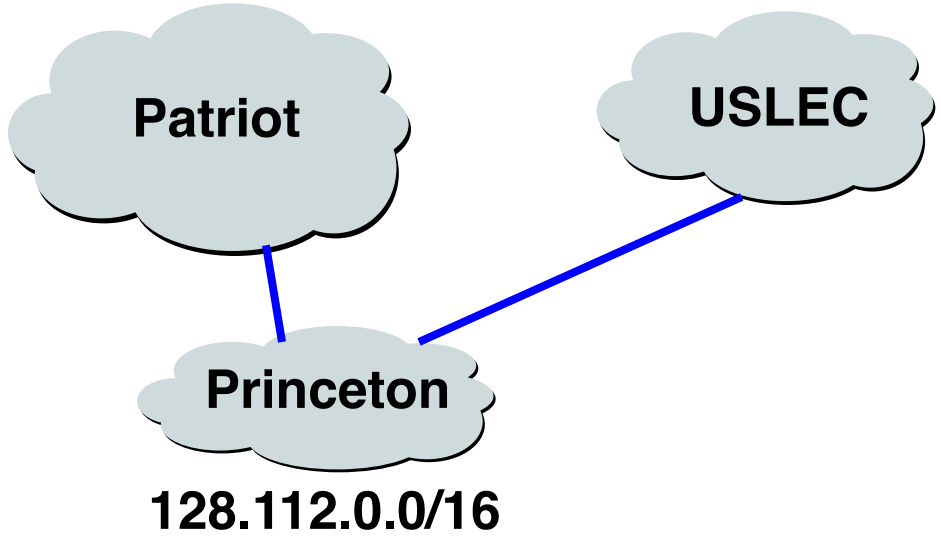
Import Policy: Local Preference

- Favor one path over another
  - Override the influence of AS path length
  - Apply local policies to prefer a path
- Example: prefer customer over peer



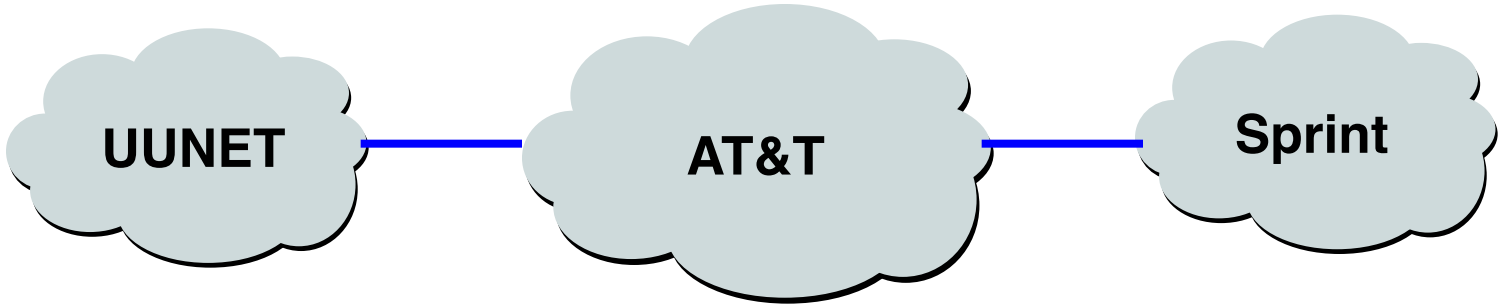
Import Policy: Filtering

- Discard some route announcements
  - Detect configuration mistakes and attacks
- Examples on session to a customer
  - Discard route if customer doesn't own the prefix
  - Discard route containing other large ISPs



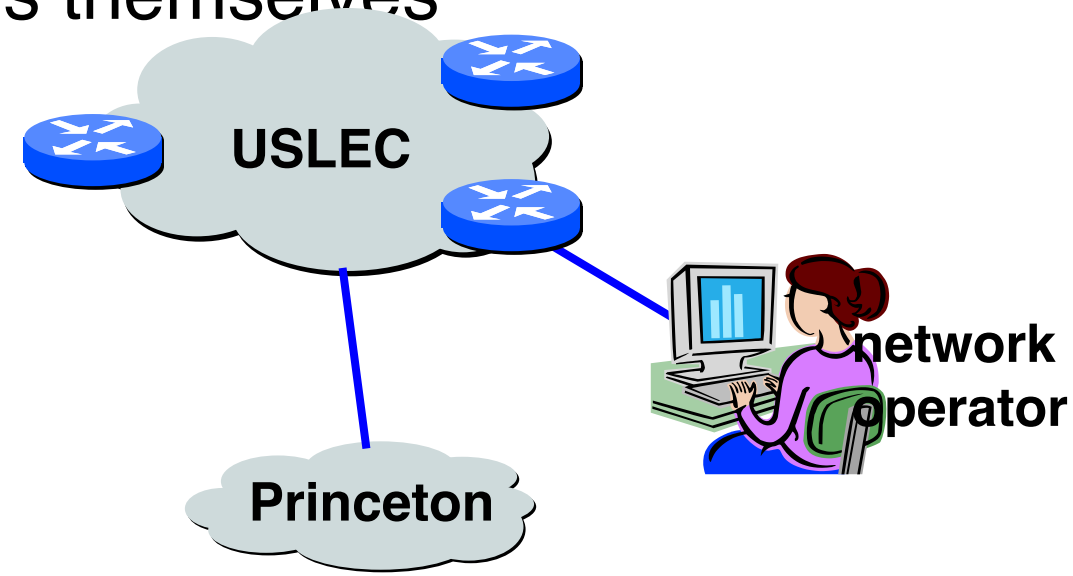
Export Policy: Filtering

- Discard some route announcements
  - Limit propagation of routing information
- Examples
  - Don’t announce routes from one peer to another



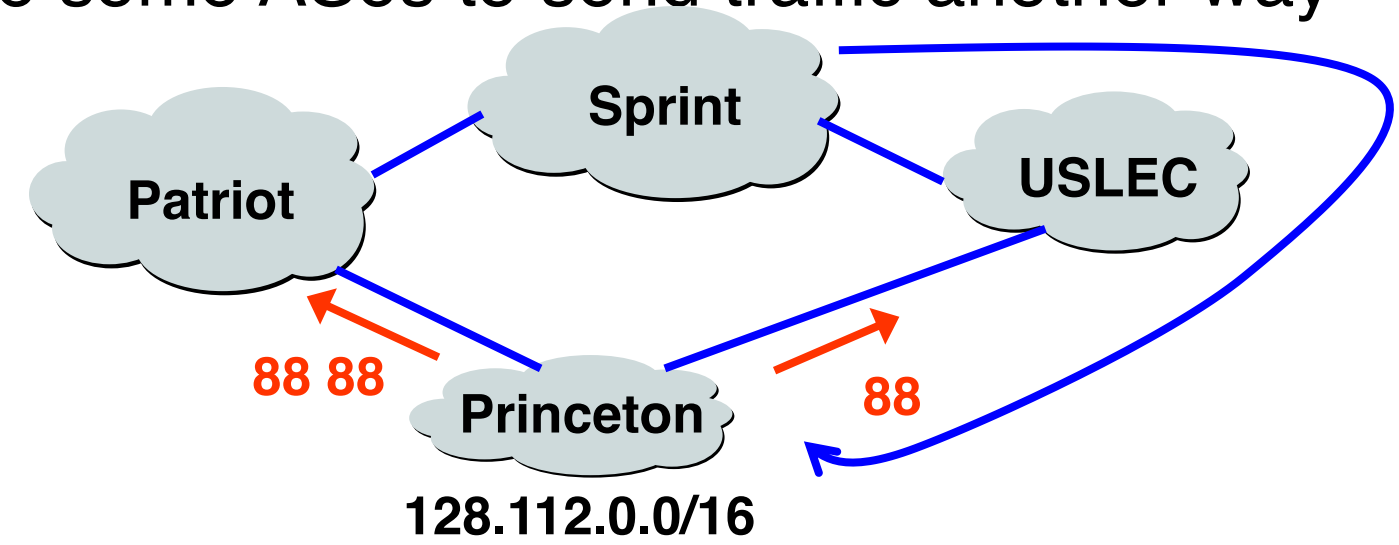
Export Policy: Filtering

- Discard some route announcements
  - Limit propagation of routing information
- Examples
  - Don’t announce routes for network-management hosts or the underlying routers themselves



Export Policy: Attribute Manipulation

- Modify attributes of the active route
  - To influence the way other ASes behave
- Example: AS prepending
  - Artificially inflate AS path length seen by others
  - Convince some ASes to send traffic another way



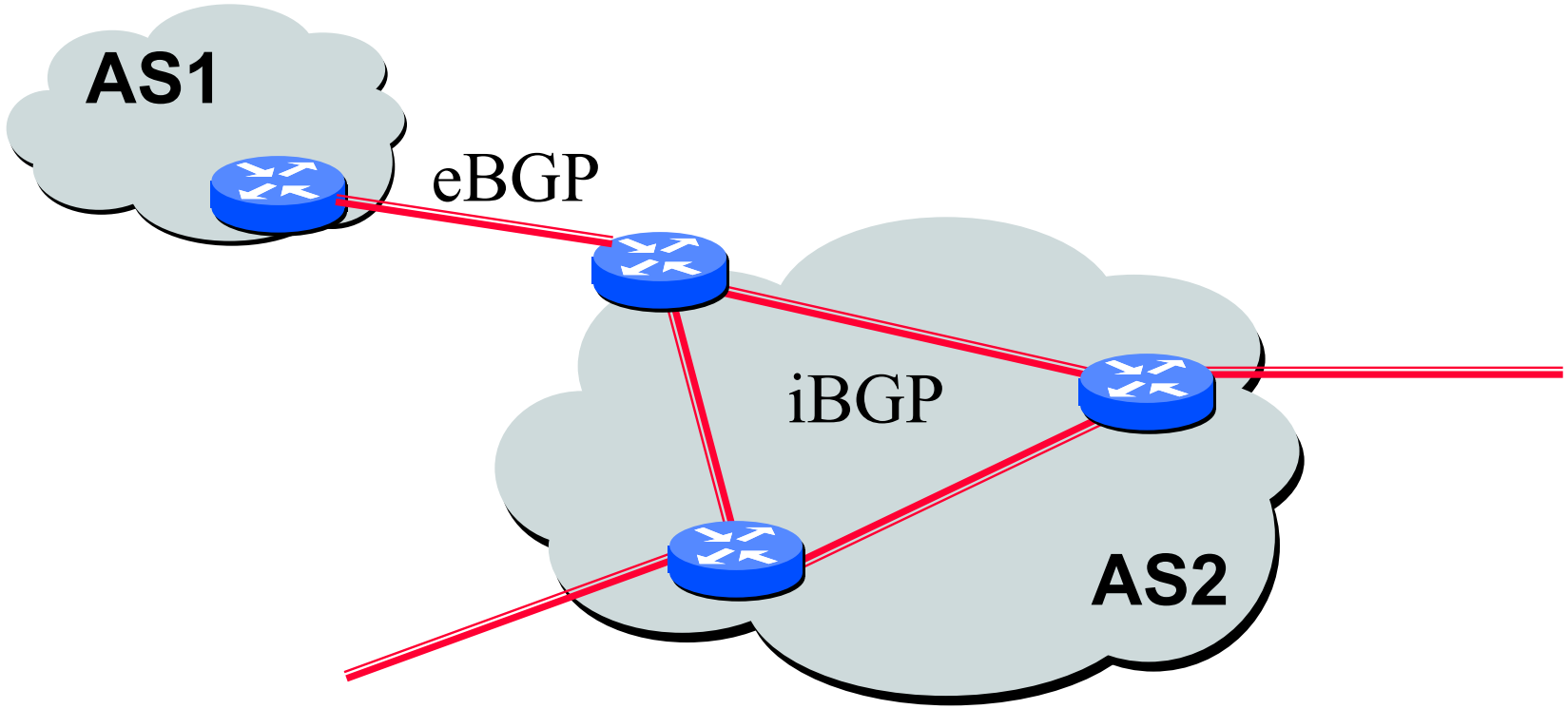
BGP Policy Configuration

- Policy languages are vendor-specific
  - Not part of the BGP protocol specification
  - Different languages for Cisco, Juniper, etc.
- Still, all languages have some key features
  - Policy as a list of clauses
  - Each clause matches on route attributes
  - … and discards or modifies the matching routes
- Configuration done by human operators
  - Implementing the policies of their AS
  - Biz relationships, traffic engineering, security, …

BGP Inside an AS

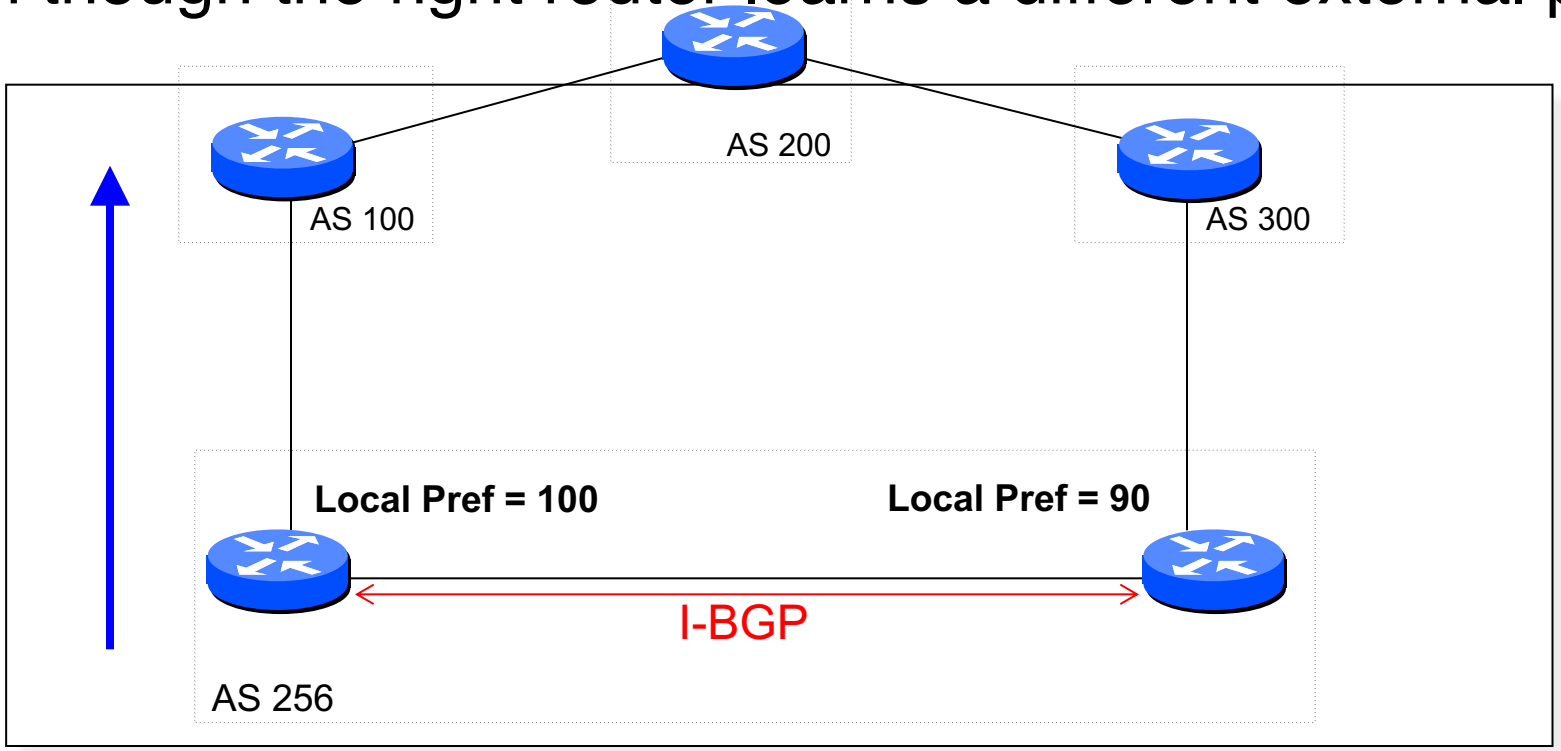
An AS is Not a Single Router

- Multiple routers in an AS
  - Need to distribute BGP information within the AS
  - Internal BGP (iBGP) sessions between routers



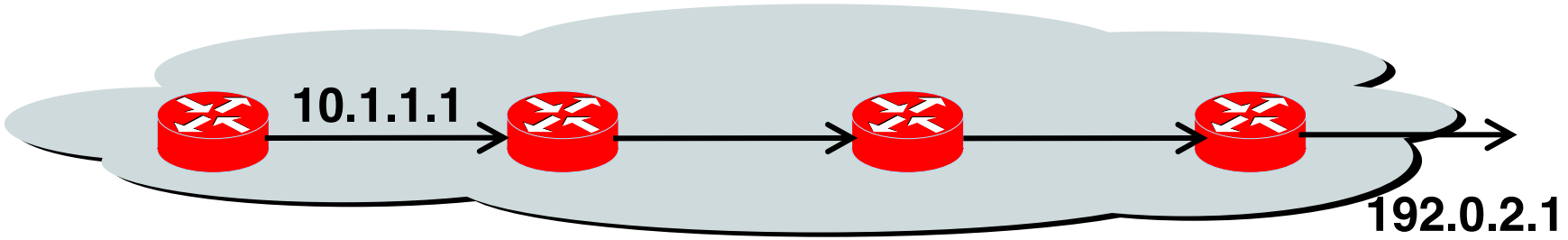
Internal BGP and Local Preference

- Example
  - Both routers prefer the path through AS 100 on the left
  - … even though the right router learns a different external path



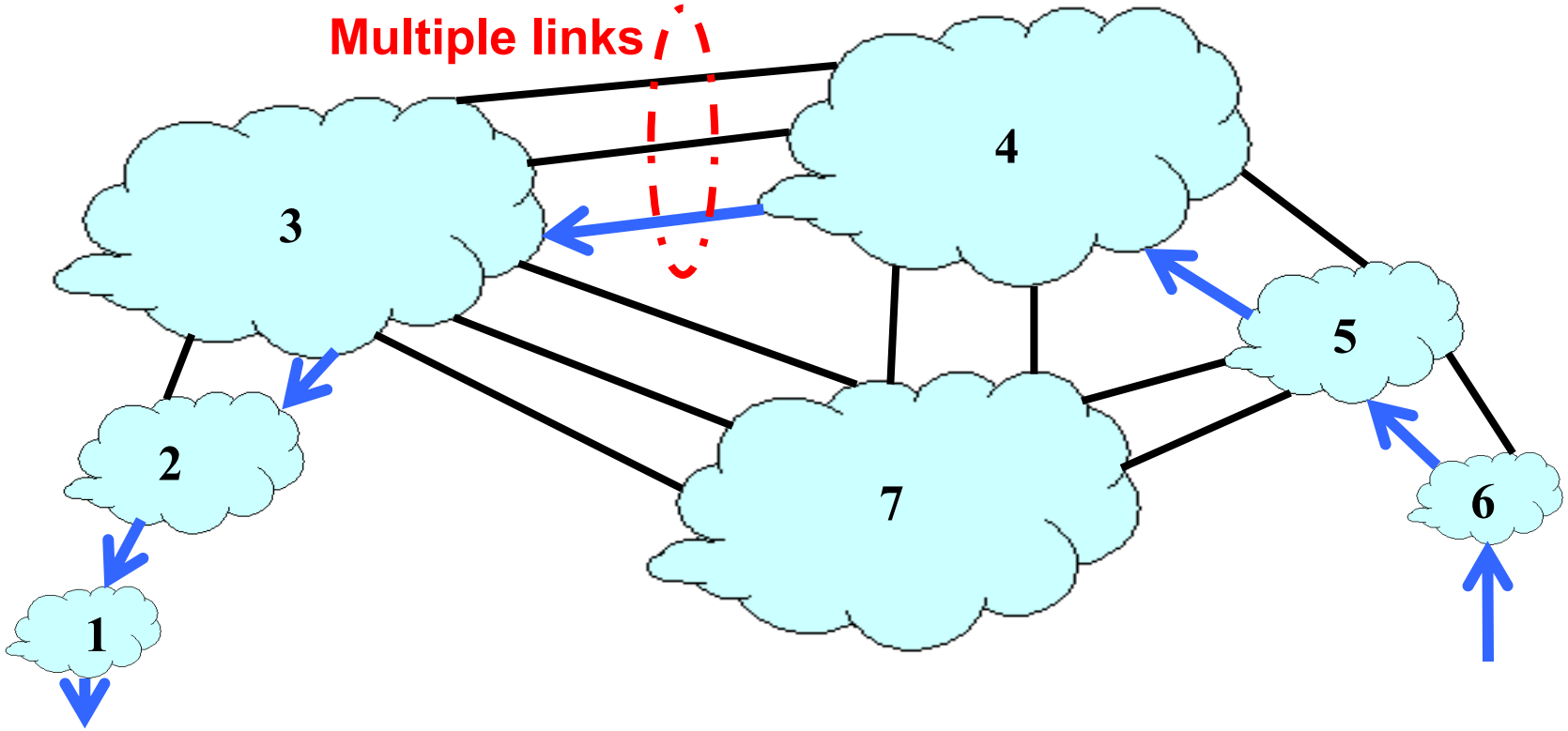
Joining BGP and IGP Information

- Border Gateway Protocol (BGP)
  - Announces reachability to external destinations
  - Maps a destination prefix to an egress point
    - 128.112.0.0/16 reached via 192.0.2.1
- Interior Gateway Protocol (IGP)
  - Used to compute paths within the AS
  - Maps an egress point to an outgoing link
    - 192.0.2.1 reached via 10.1.1.1



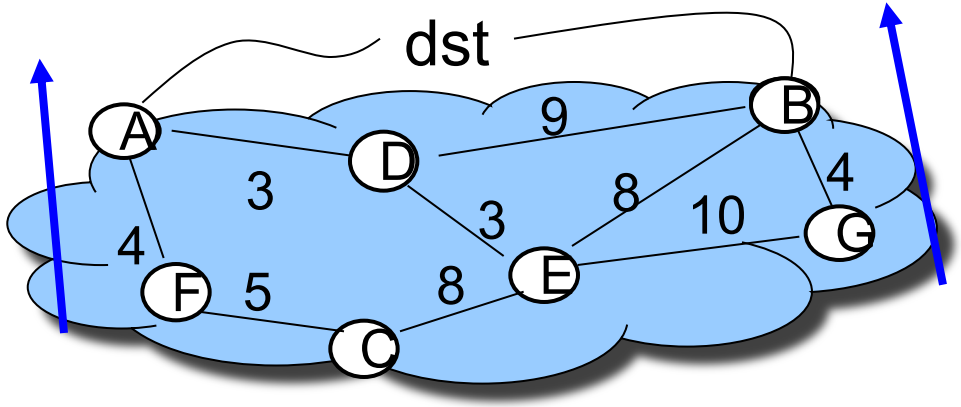
An AS May Learn Many Routes

- Multiple connections to neighboring ASes
  - Multiple border routers may learn good routes
  - … with the same local-pref and AS path length



Hot-Potato (Early-Exit) Routing

- Hot-potato routing
  - Each router selects the closest egress point
  - … based on the path cost in intradomain protocol
- BGP decision process
  - Highest local preference
  - Shortest AS path
  - **Closest egress point**
  - Arbitrary tie break

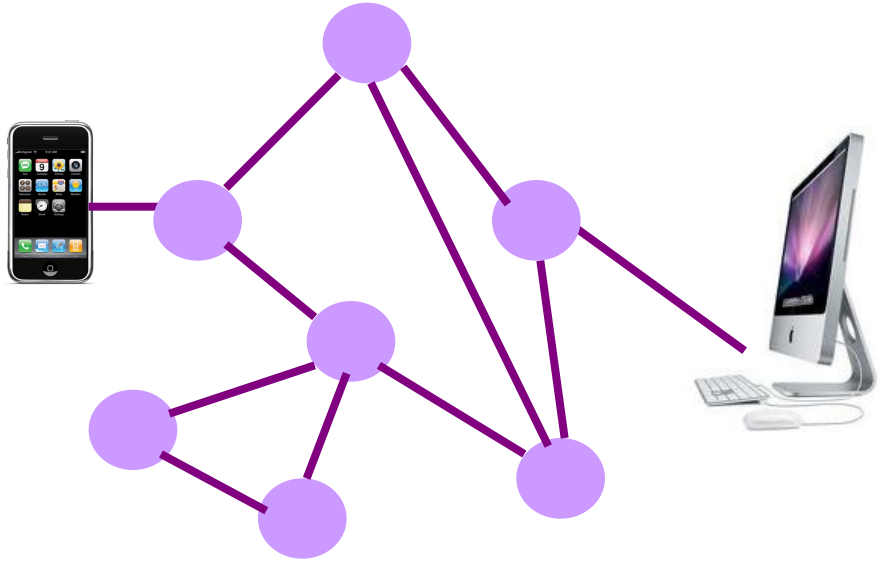


hot potato

Learning the locations of the endpoints

Finding the endpoints

- Computing the forwarding table
  - Still must figure out where the endpoints are

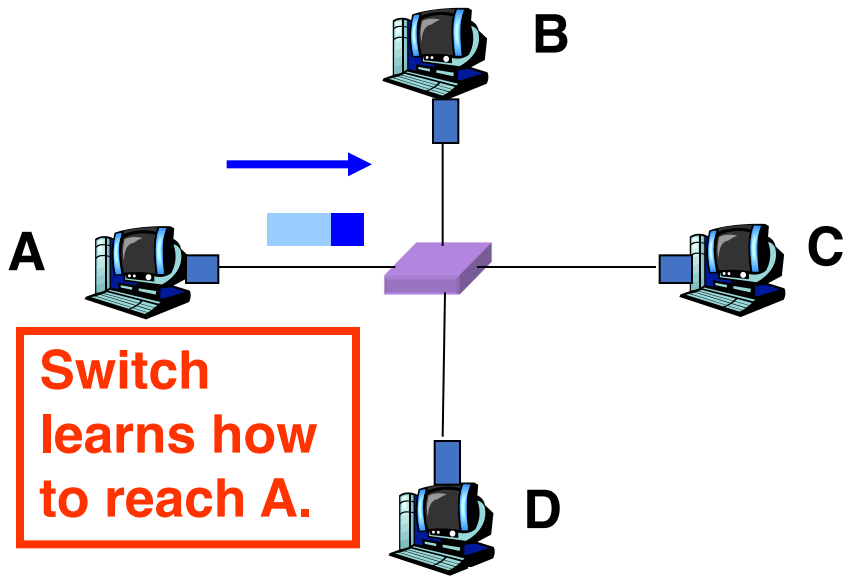


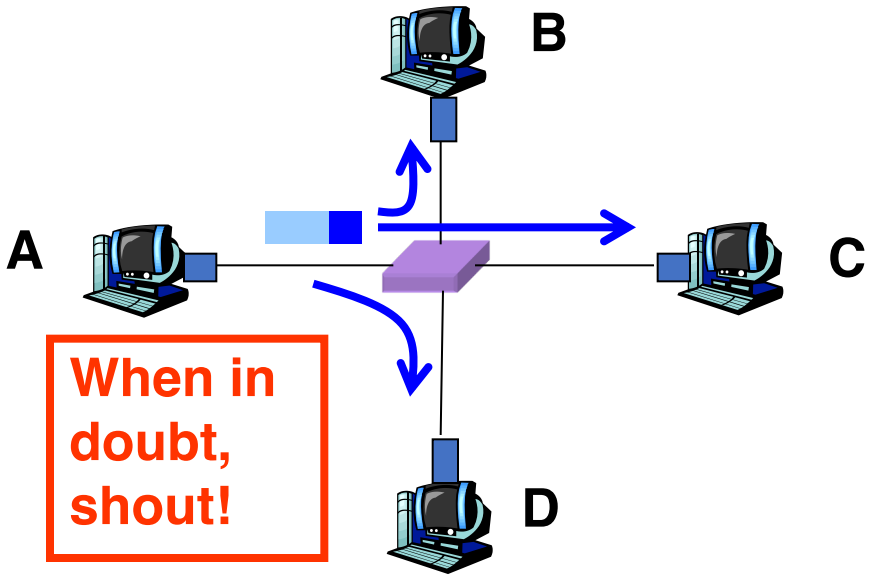
- How to find the endpoints?
  - Learning/flooding (Ethernet)
  - Injecting into the routing protocol
  - Dissemination using a different protocol
  - Central directory service

- Ways to curb scaling challenges
  - E.g., spanning tree per VLAN for endpoint flooding

Learning and Flooding

- When a frame arrives
  - Inspect the *source* address
  - Associate address with the *incoming* interface
- When the frame has an unfamiliar destination
  - Forward out all interfaces
  - … except for the one where the frame arrived

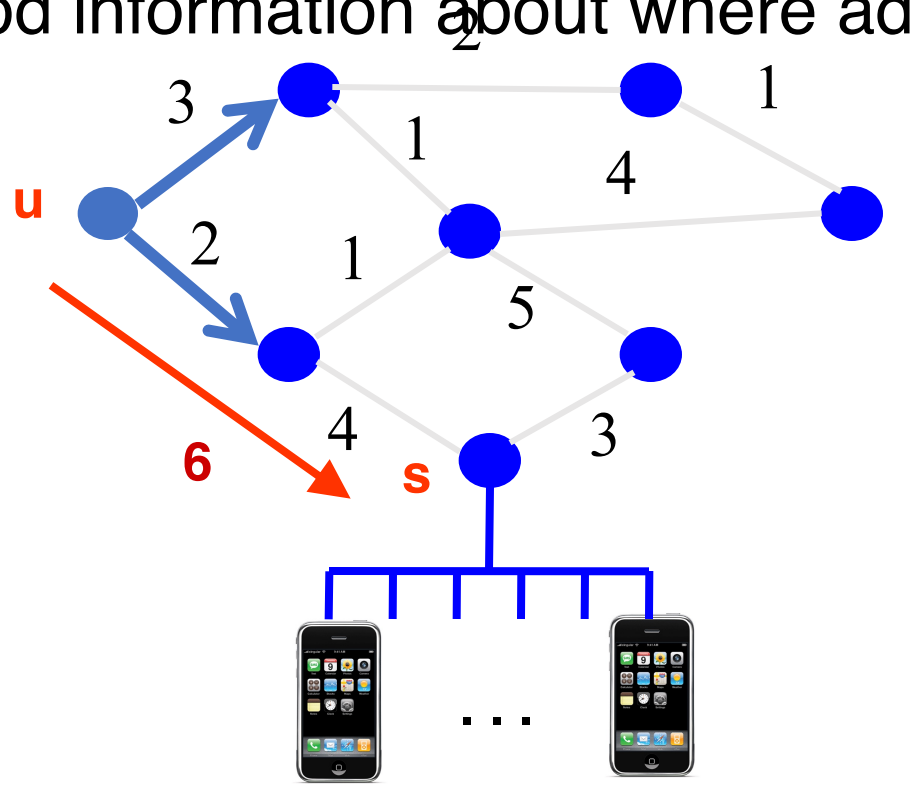




Used in Ethernet LANs

Inject into Routing Protocol

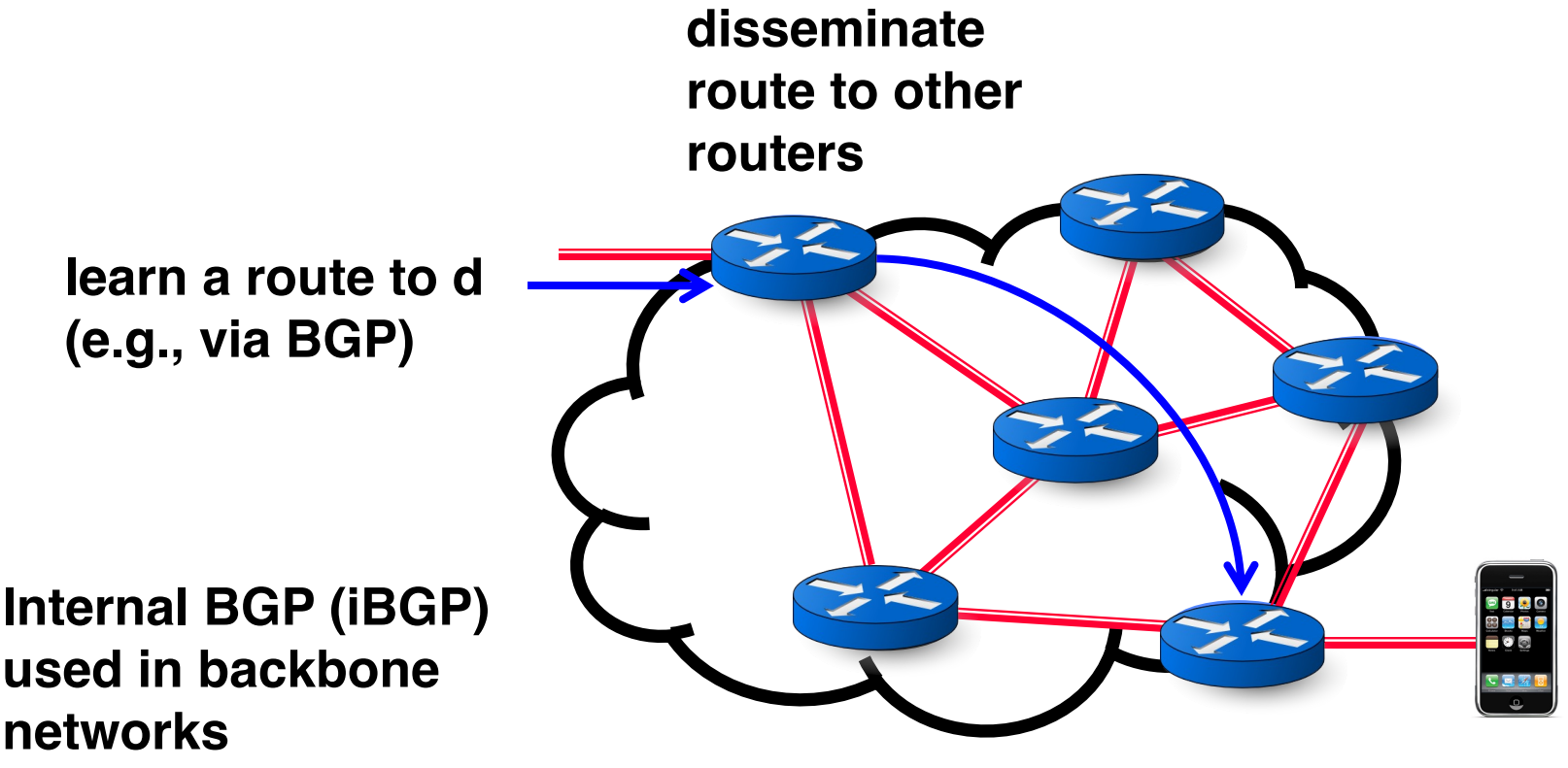
- Treat the end host (or subnet) as a node
  - And disseminate in the routing protocol
  - E.g., flood information about where addresses attach



Used in OSPF and IS-IS, especially in enterprise networks

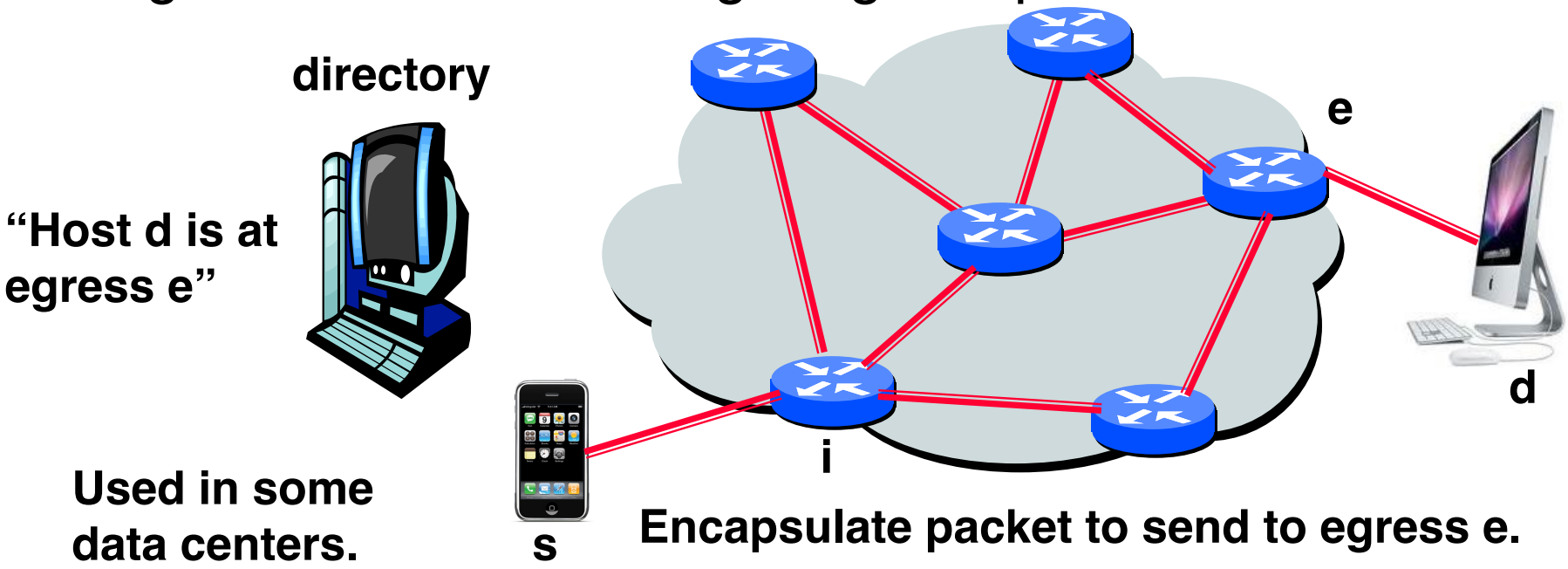
Disseminate with another protocol

- One router learns the route
- … and shares the information with other routers



Directory Service

- Contact a service to learn the location
  - Lookup the end-host or subnet address
  - … and learn the label to put on the packet
  - … to get the traffic to the right egress point



Conclusion

- Routing is a distributed computation
  - With challenges in scalability and handling dynamics

- Different solutions for different environments
  - Ethernet LAN: spanning tree, MAC learning, flooding
  - Enterprise: link-state routing, injecting subnet addresses
  - Backbone: link-state routing inside, path-vector routing with neighboring domains, and iBGP dissemination
  - Internet: BGP
  - Data centers: many different solutions, still in flux
    - An active research area…