

Routing (part 3)

Lecture 25

<http://www.cs.rutgers.edu/~sn624/352-F24>

Srinivas Narayana

The network layer enables **reachability**. We'll see protocols that solve subproblems.

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

How does an endpoint
talk to another *within*
the same network?

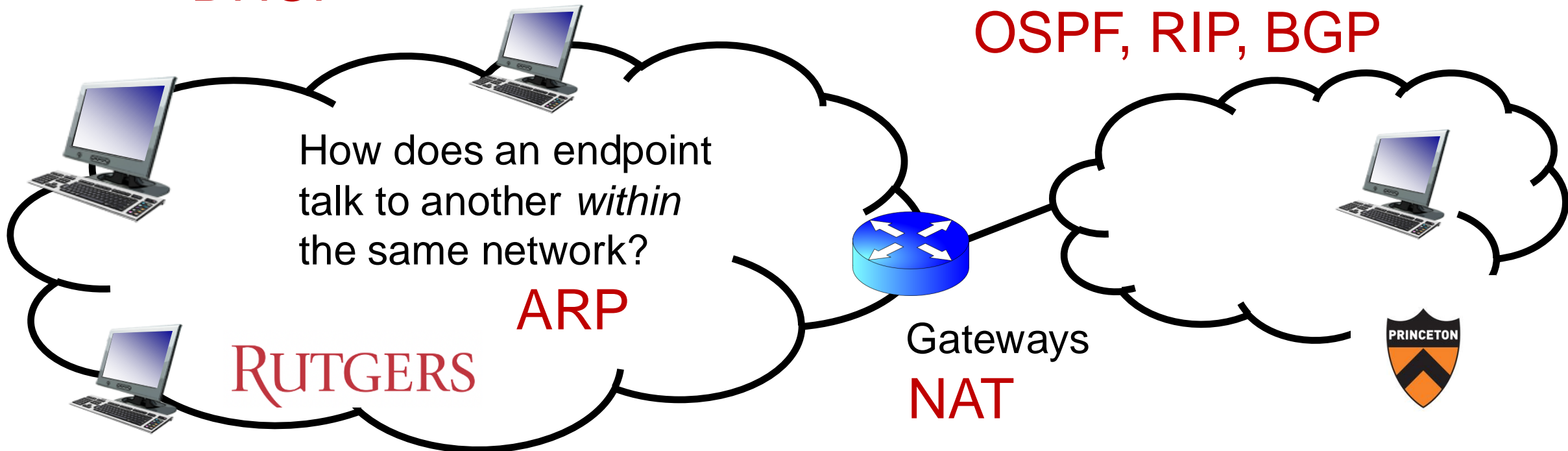
ARP

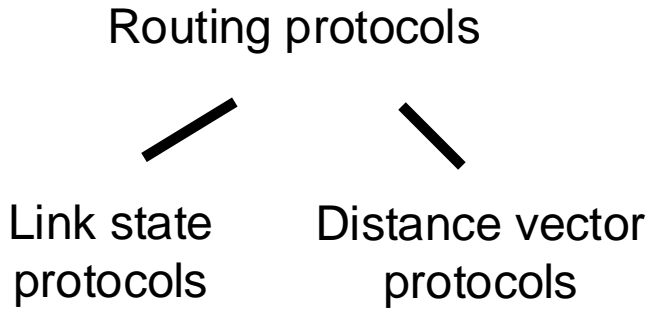
RUTGERS



Gateways

NAT





Every router is aware of the existence of every other router.

Messages reveal information on the full network (graph) structure.

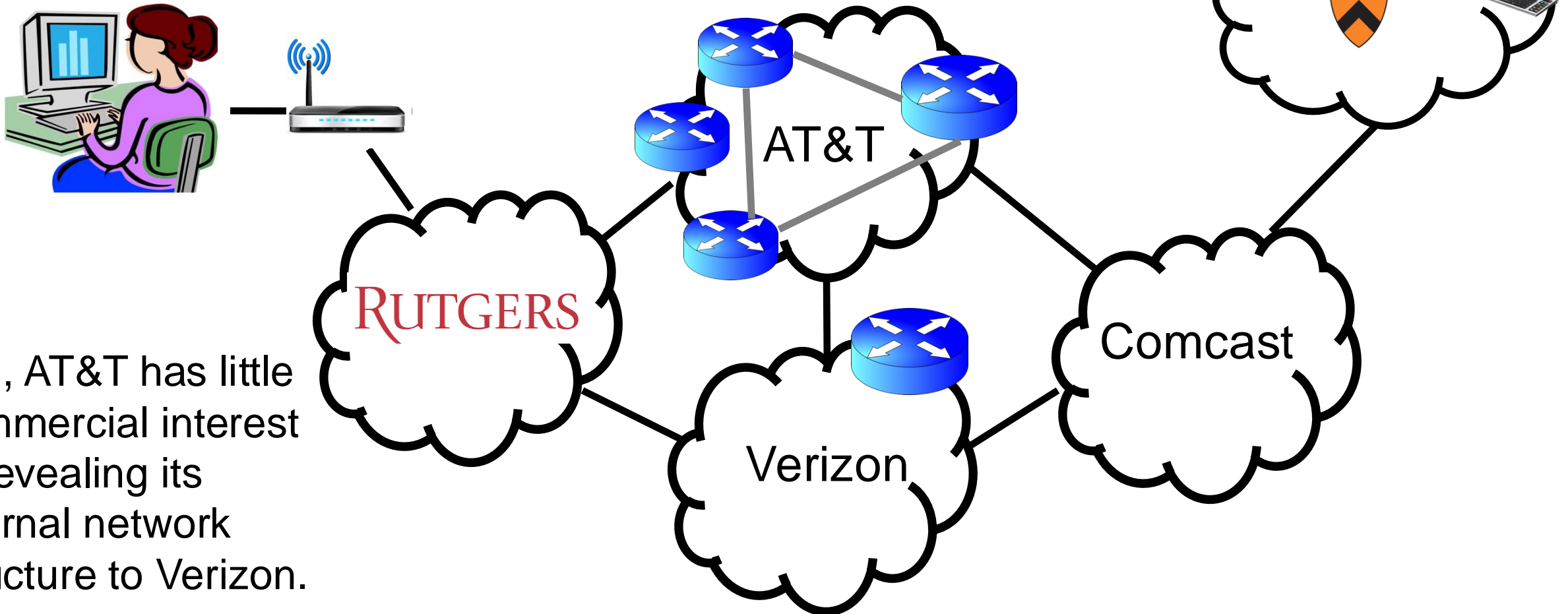
Message exchange and forwarding tables scale with network size.

These assumptions/settings cannot work on the Internet.

The Internet is a large **federated** network

Several autonomously run organizations: No one “boss”

Organizations cooperate, but also **compete**

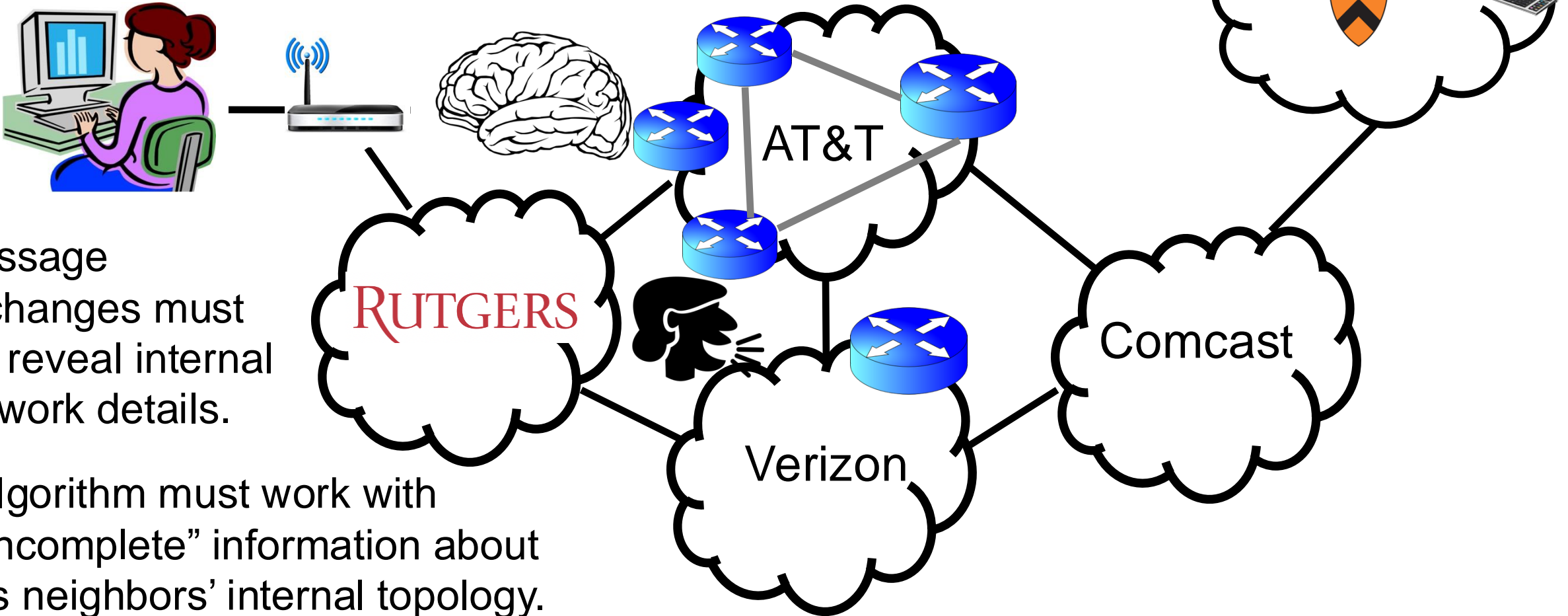


e.g., AT&T has little commercial interest in revealing its internal network structure to Verizon.

The Internet is a large **federated** network

Several autonomously run organizations: No one “boss”

Organizations cooperate, but also **compete**



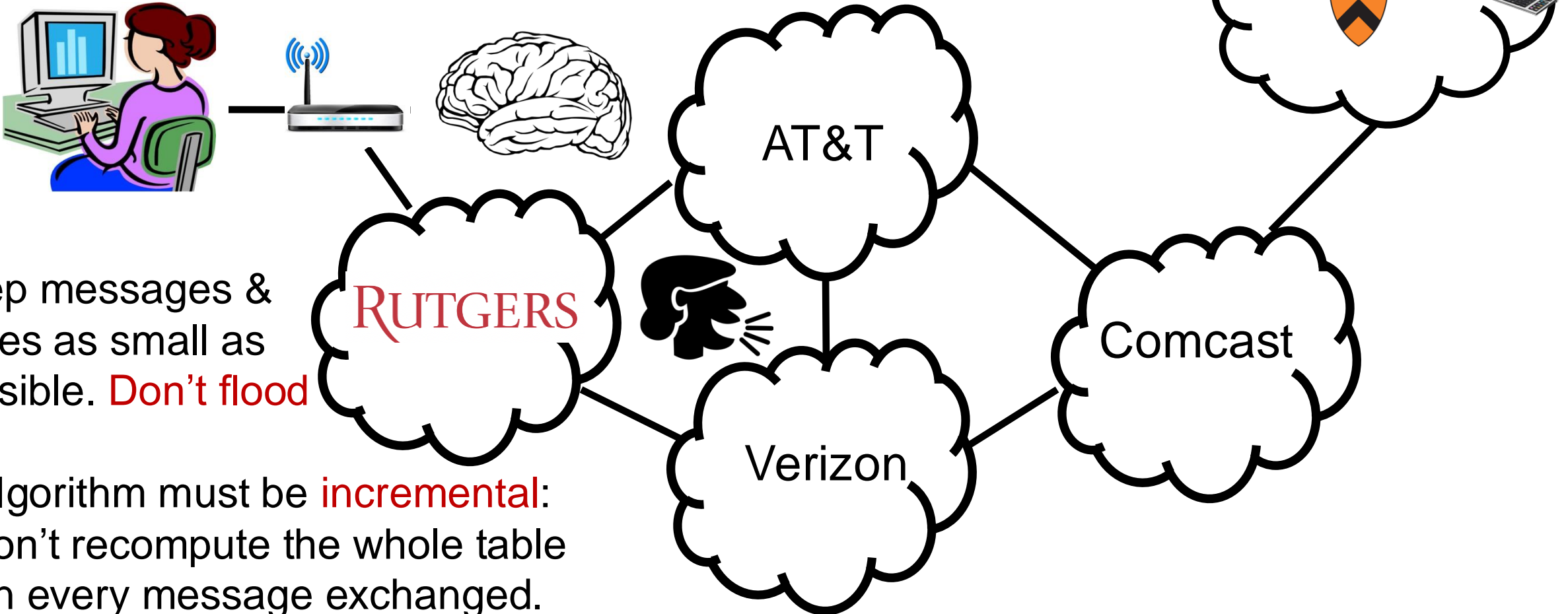
Message exchanges must not reveal internal network details.

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

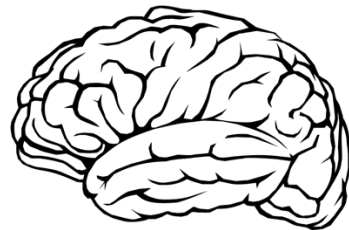


Inter-domain Routing

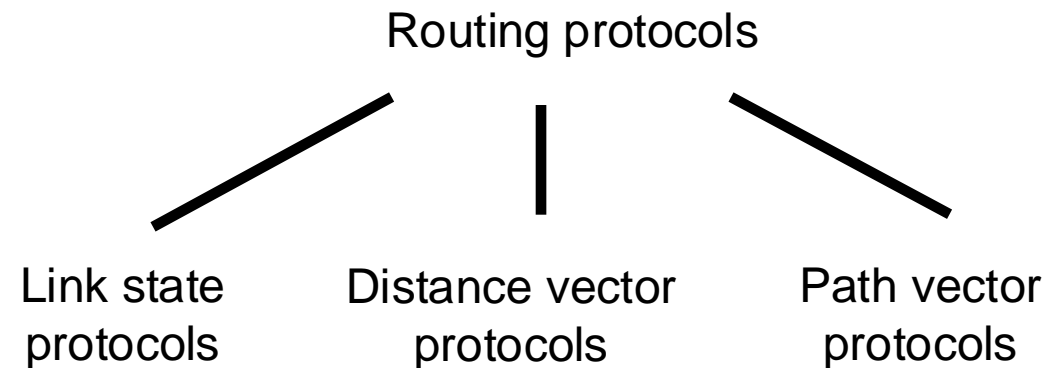
- Routing approaches so far (LS + DV) are applicable within one **autonomous system (AS)**, e.g., Rutgers
 - Called **intra-domain** routing protocols
- The Internet uses **Border Gateway Protocol (BGP)**
- **All AS'es speak BGP**. It is the glue that holds the Internet together
- BGP is a **path vector protocol**



Messages?



Algorithm?



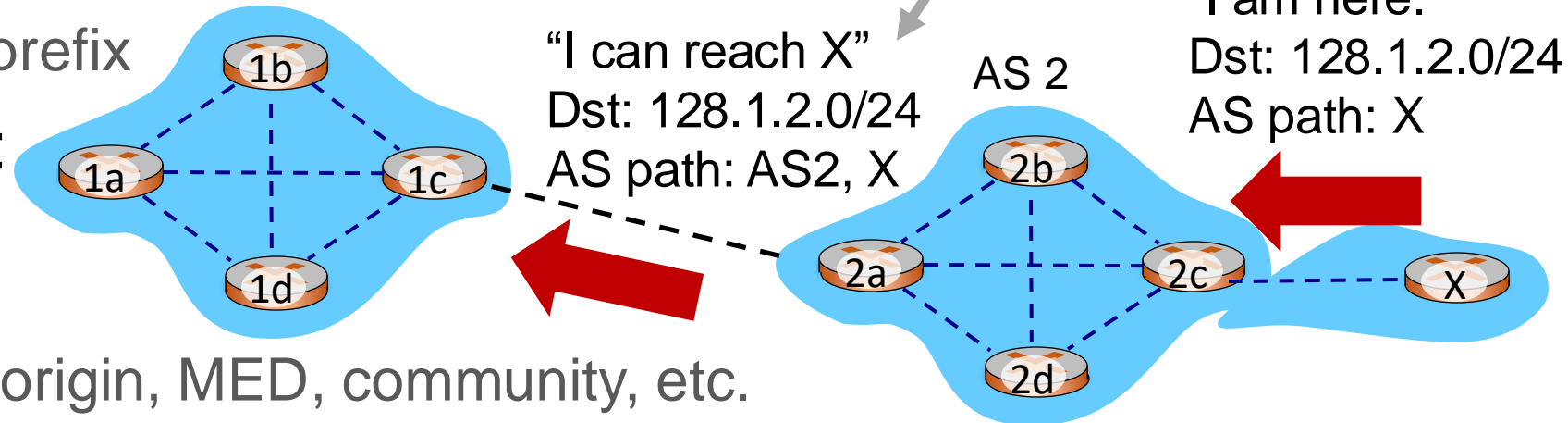
Q1. BGP Messages



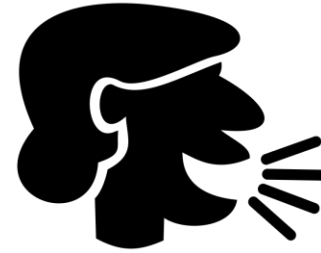
Loop detection is easy
(no “count to infinity”)

Exchange paths: **path vector**

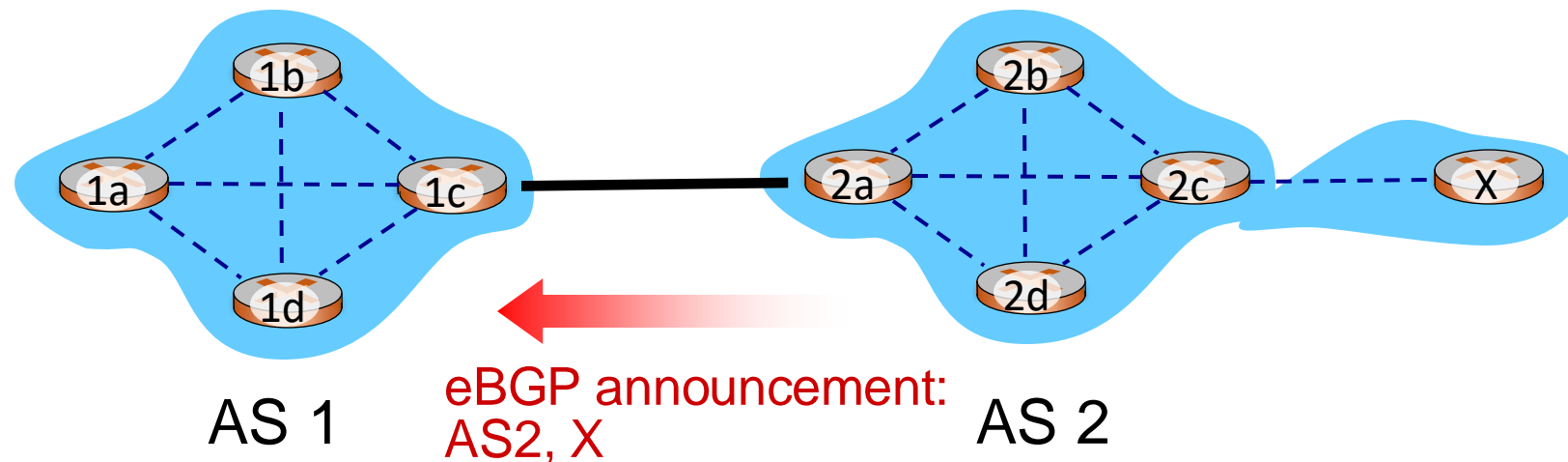
- Routing **Announcements** or **Advertisements** No link metrics, distances!
 - “I am here” or “I can reach here”
 - Occur over a TCP connection (**BGP session**) between routers
- Route announcement = destination + attributes
 - Destination: IP prefix
- Route Attributes:
 - **AS-level path**
 - Next hop
 - Several others: origin, MED, community, etc.
- An AS promises to use advertised path to reach destination
- Only route changes are advertised after BGP session established



Q1. Next Hop



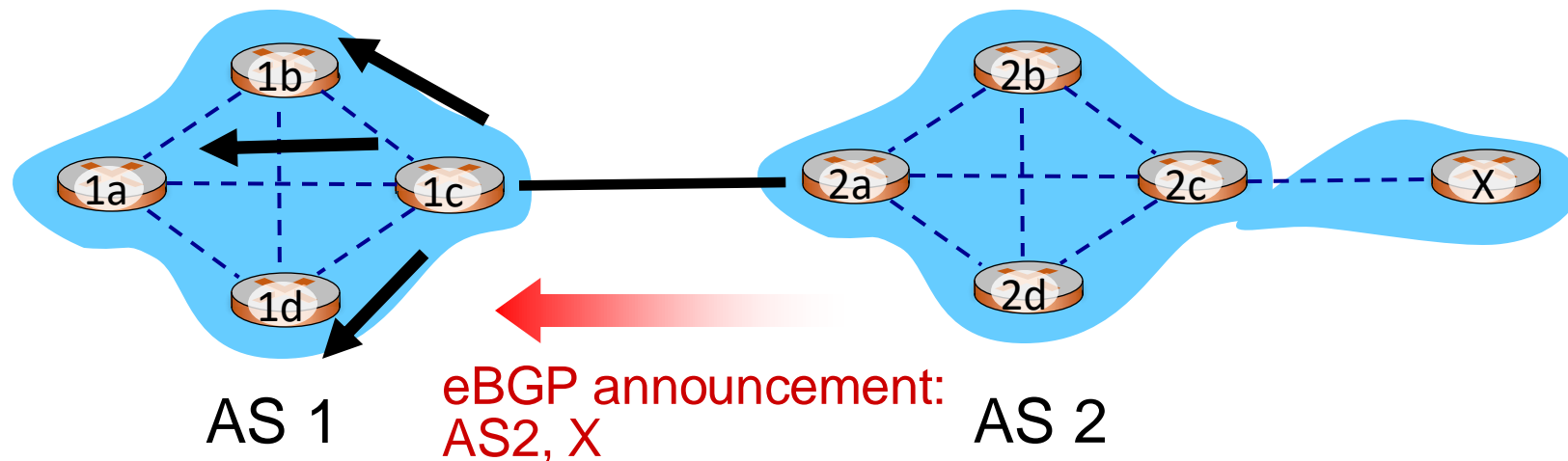
- **Next hop** conceptually denotes the first router interface that begins the AS-level path
 - The meaning of this attribute is context-dependent
- In an announcement arriving from a different AS (**eBGP**), next hop is the router **in the next AS** which sent the announcement
 - Example: Next Hop of the eBGP announcement reaching 1c is **2a**



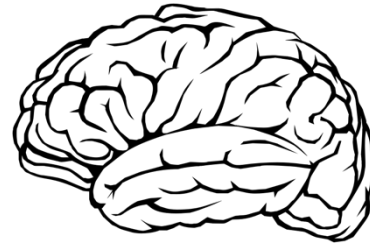
Q1. Next Hop



- Suppose router 1c **imports** the path (more on this soon)
- Router 1c will propagate the announcement **inside the AS** using **iBGP**
- The next hop of this (iBGP) announcement is set to 1c
 - In particular, the next hop is an AS1 **internal** address



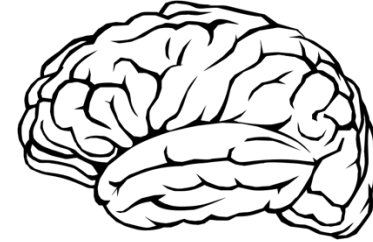
Q2. The algorithm



- A BGP router does *not* consider every routing advertisement it receives by default to make routing decisions!
 - An **import policy** determines whether a route is even considered a candidate
 - Once imported, the router performs **route selection**
 - A BGP router does *not* propagate its chosen path to a destination to all other AS'es by default!
 - An **export policy** determines whether a (chosen) path can be advertised to other AS'es and routers
- Programmed by network operator

Policy considerations make BGP very different from intra-domain (LS / DV) protocols

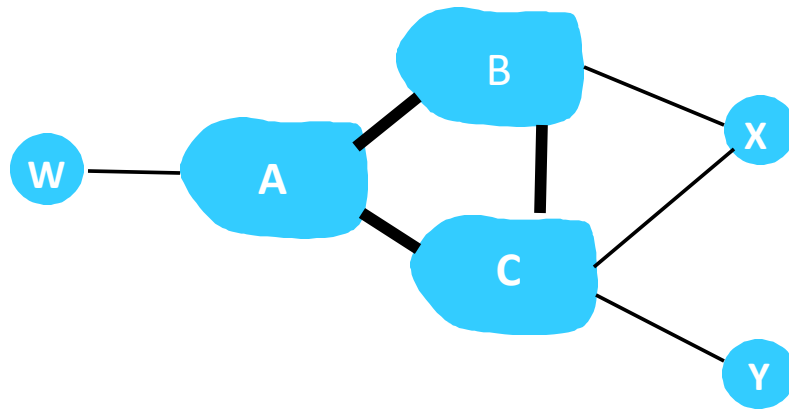
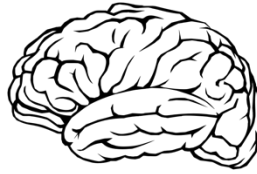
Policies in BGP





Policy arises from business relationships

- Customer-provider relationships:
 - E.g., Rutgers is a customer of AT&T
- Peer-peer relationships:
 - E.g., Verizon is a peer of AT&T
- Business relationships depend on **where** connectivity occurs
 - “Where”, also called a “point of presence” (PoP)
 - e.g., customers at one PoP but peers at another
 - Internet-eXchange Points (IXPs) are large PoPs where ISPs come together to connect with each other (often for free)

BGP Export Policy

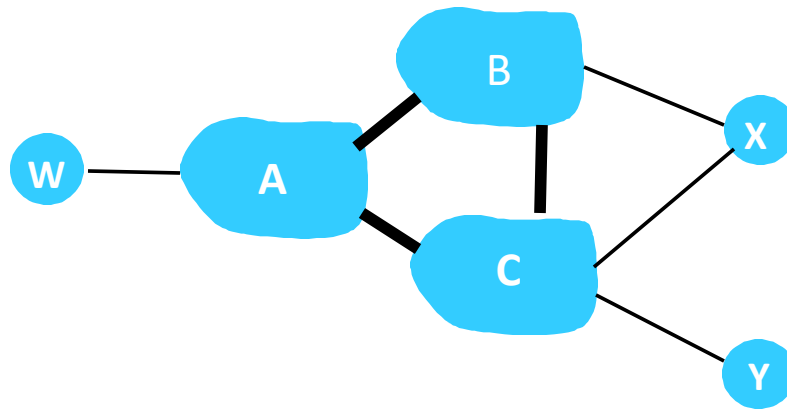
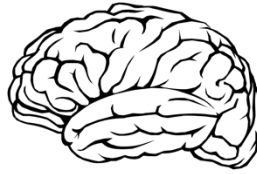




legend:  provider network
 customer network:

Suppose an ISP only wants to route traffic to/from its customer networks (does not want to carry **transit traffic** between other ISPs)

- A,B,C are **provider networks**
- X,W,Y are customers (of provider networks)
- X is **dual-homed**: attached to two networks
- policy to enforce: X does not want to route from B to C via X
 - So, X **will not announce** to B a route to C

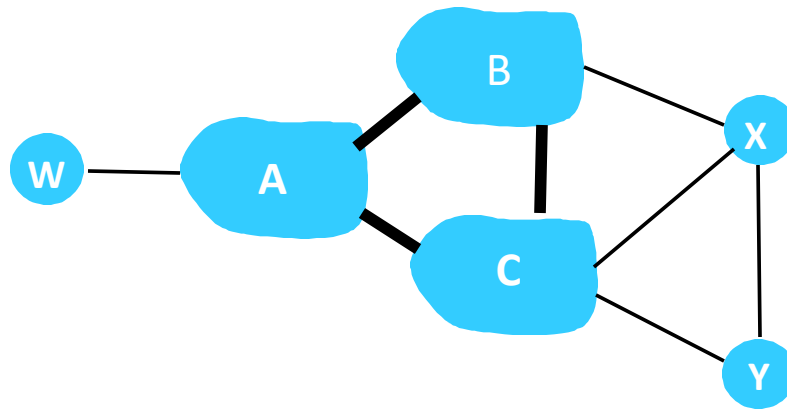
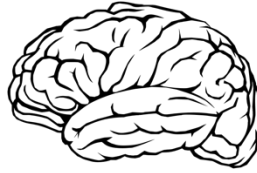
BGP Export Policy





legend:  provider network
 customer network:

- Suppose an ISP only wants to route traffic to/from its customer networks (does not want to carry **transit traffic** between other ISPs)
- A announces path Aw to B and to C
 - B **will not announce** BAw to C:
 - B gets no “revenue” for routing CBAw, since none of C, A, w are B’s customers
 - C will route CAw (not using B) to get to w

BGP Import Policy

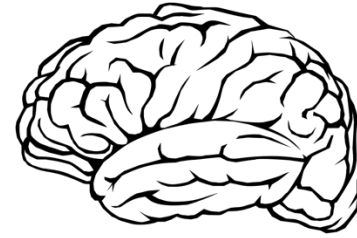


legend:  provider network
 customer network:

Suppose an ISP wants to **minimize costs** by avoiding routing through its providers when possible.

- Suppose C announces path Cy to x
- Further, y announces a direct path (“y”) to x
- Then x may **choose not to import** the path Cy to y since it has a peer path (“y”) towards y

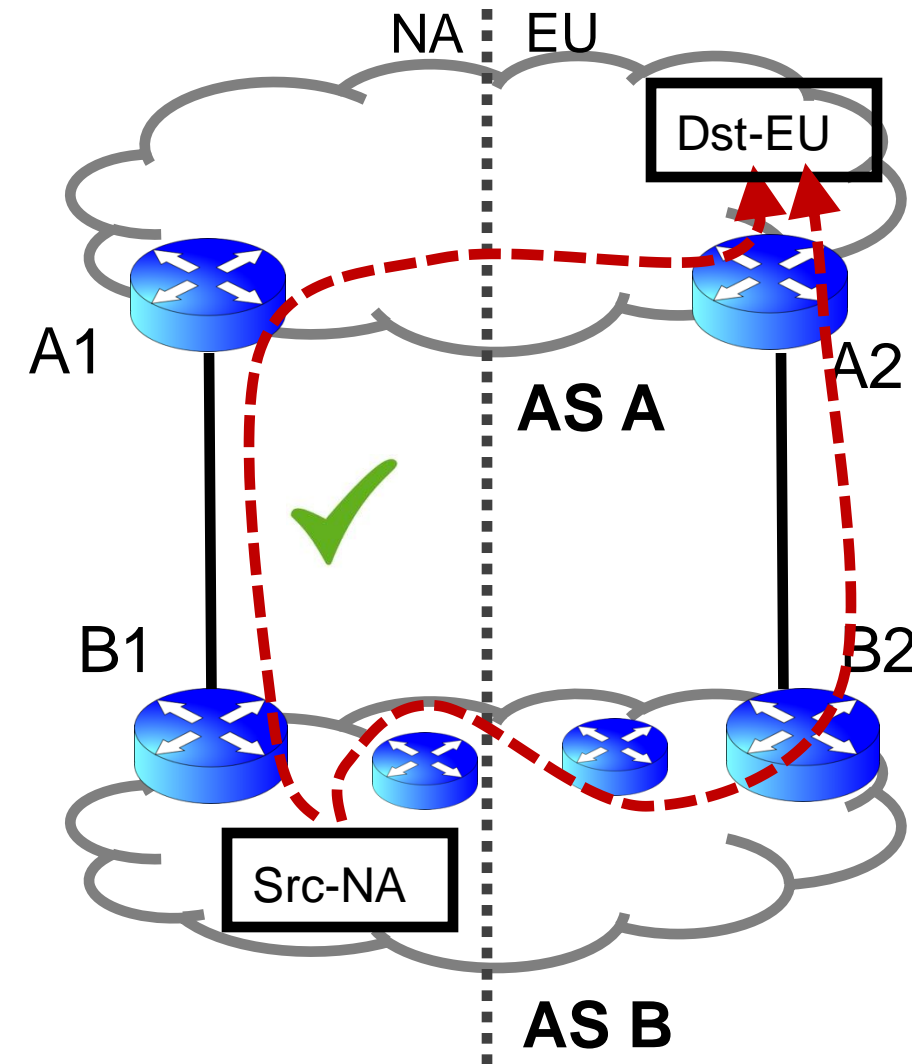
Q2. BGP Route Selection



- When a router imports more than one route to a destination IP prefix, it selects route based on:
 1. **local preference value** attribute (import policy decision -- set by network admin)
 2. shortest AS-PATH
 3. closest NEXT-HOP router
 4. Several additional criteria: You can read up on the full, complex, list of criteria, e.g., at <https://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html>

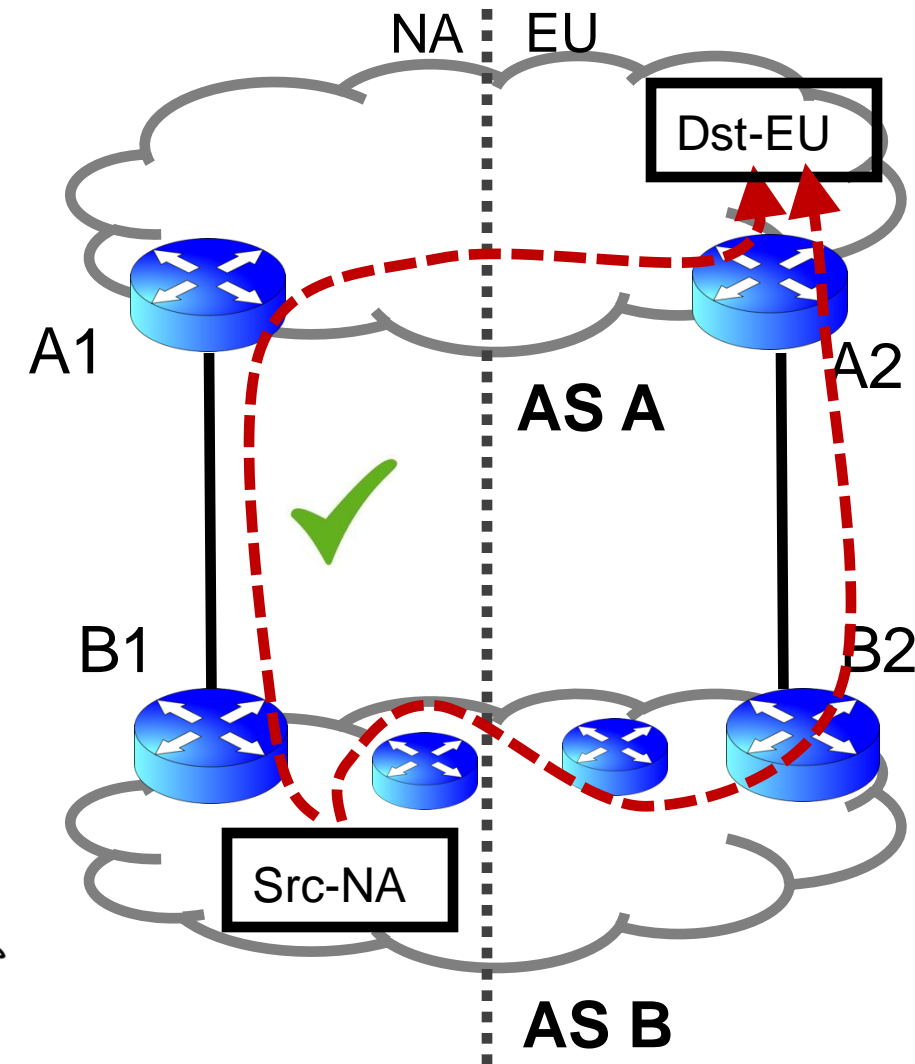
Example of route selection

- Suppose AS A and B are connected to each other both in North America (NA) and in Europe (EU)
- A source in NA wants to reach a destination in EU
- There are two paths available
 - *Assume* same local preference
 - Same AS path length
- **Closest next hop-router:** choose path via B1 rather than B2

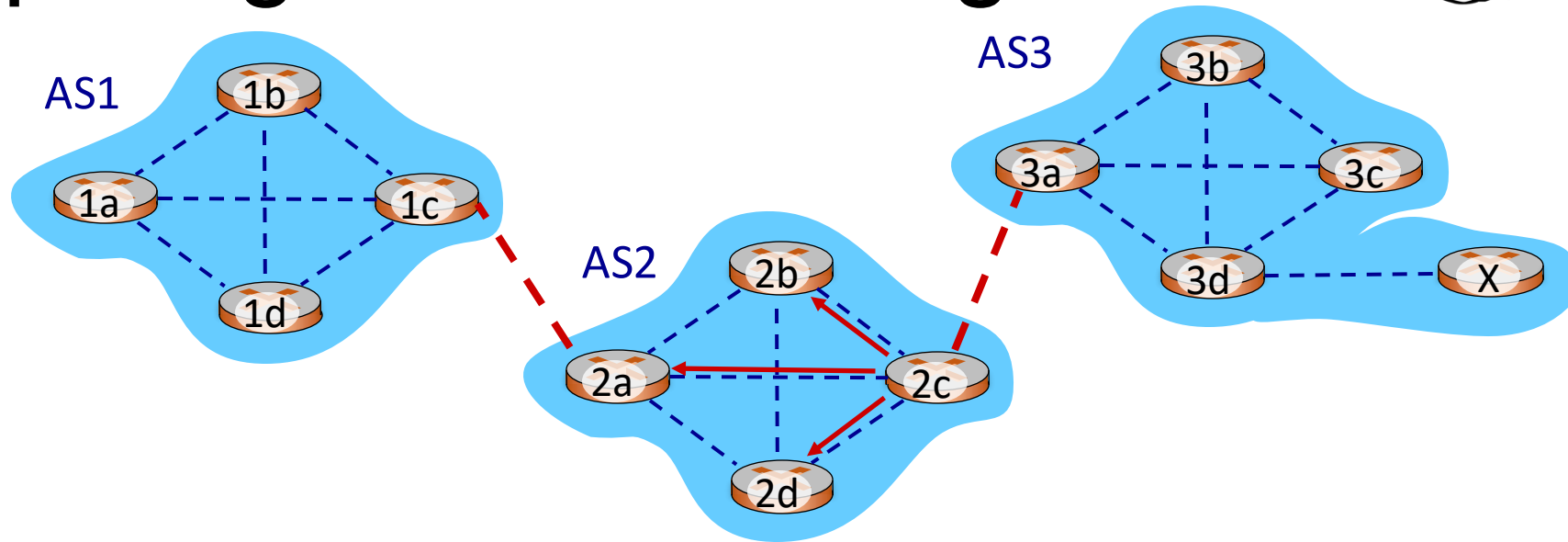
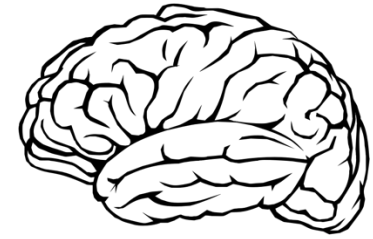


Example of route selection

- Choosing closest next-hop results in **early exit routing**
 - Try to exit the local AS as early as possible
 - Also called **hot potato routing**
- Reduce resource use within local AS
 - potentially at the expense of another AS

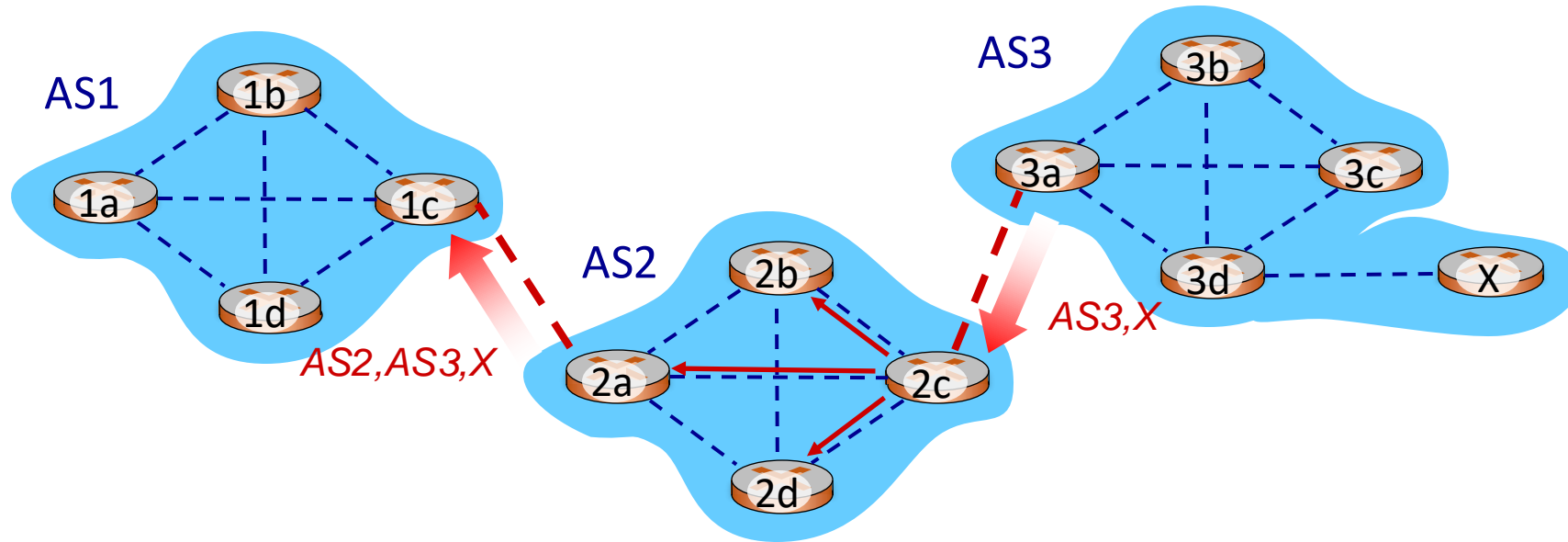


Computing the forwarding table



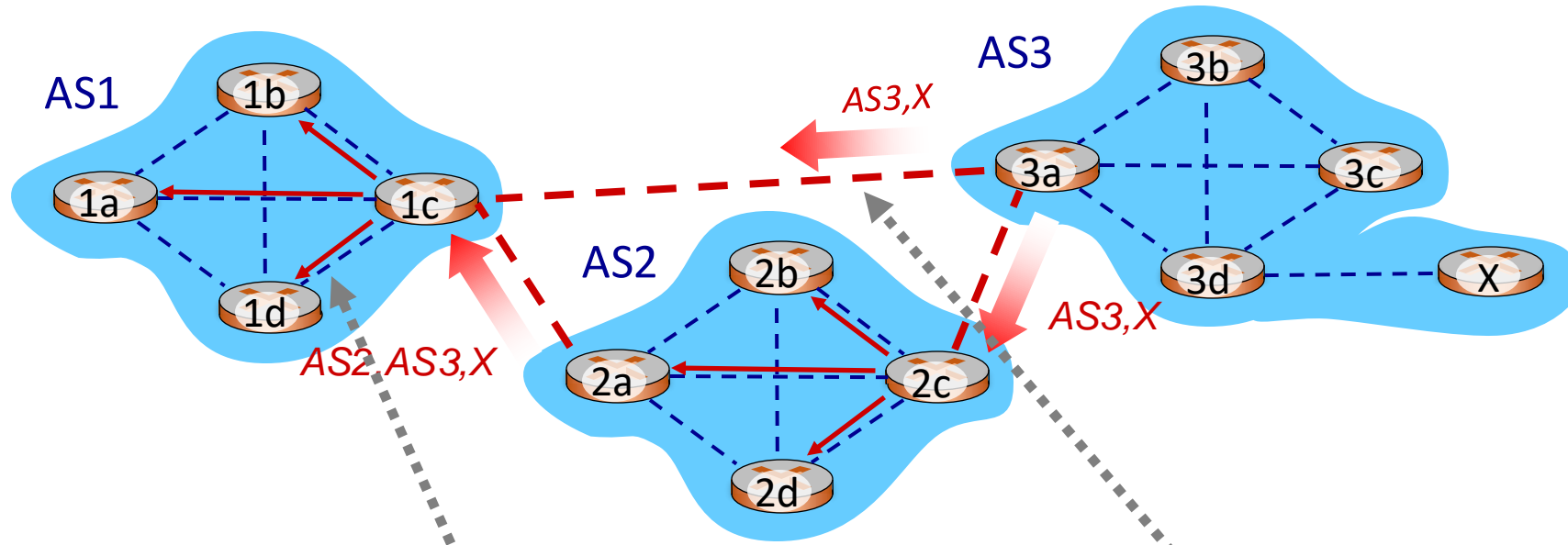
- Suppose a router in AS1 wants to forward a packet destined to external prefix X.
- How is the forwarding table entry for X at 1d computed?
- How is the forwarding table entry for X at 1c computed?

eBGP and iBGP announcements



- AS2 router 2c receives path announcement **AS3,X** (via **eBGP**) from AS3 router 3a
- Based on AS2 import policy, AS2 router 2c imports and selects path AS3,X, propagates (via **iBGP**) to all AS2 routers
- Based on AS2 export policy, AS2 router 2a announces (via eBGP) path **AS2, AS3, X** to AS1 router 1c

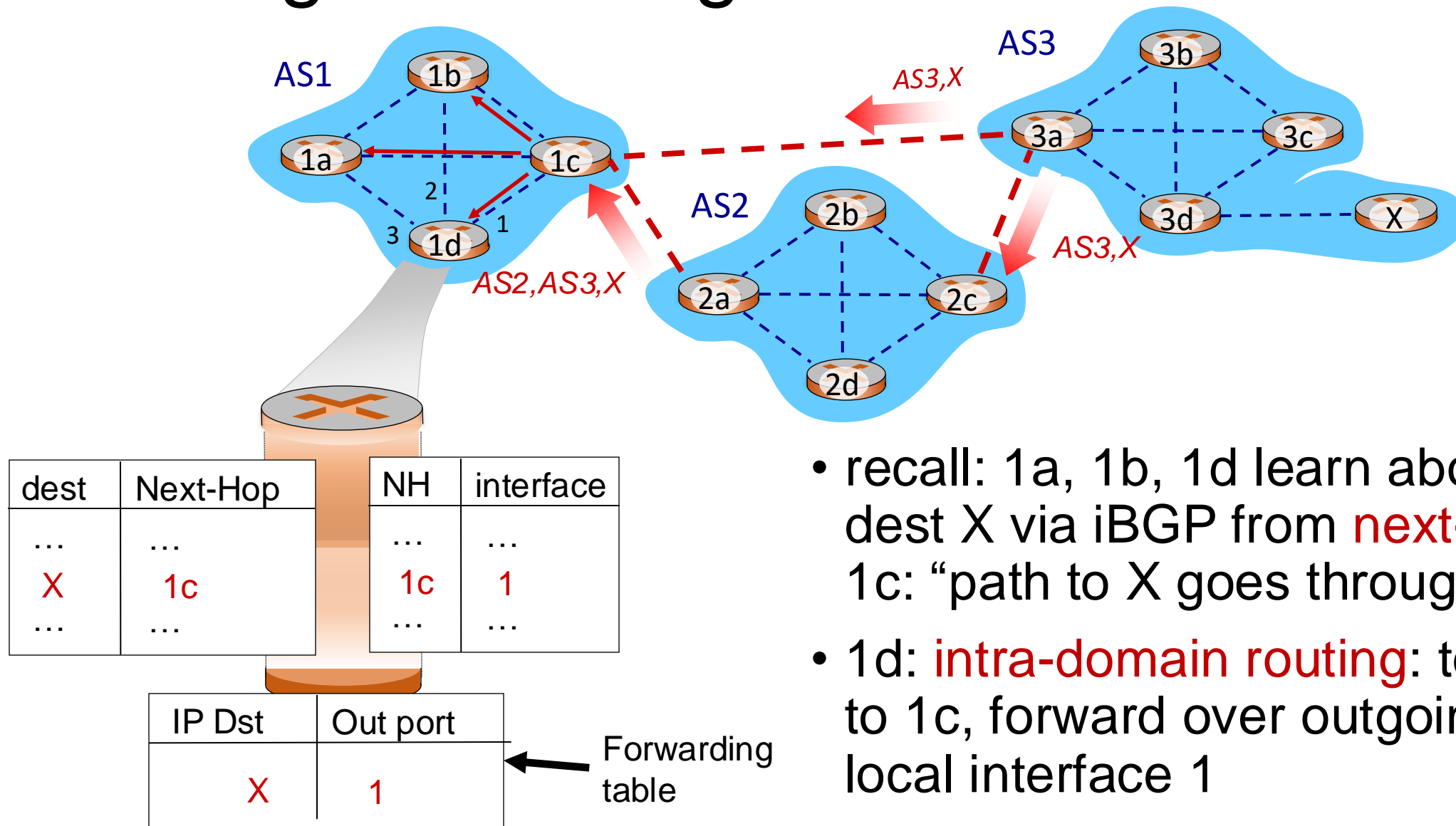
eBGP and iBGP announcements



A given router may learn about **multiple** paths to destination:

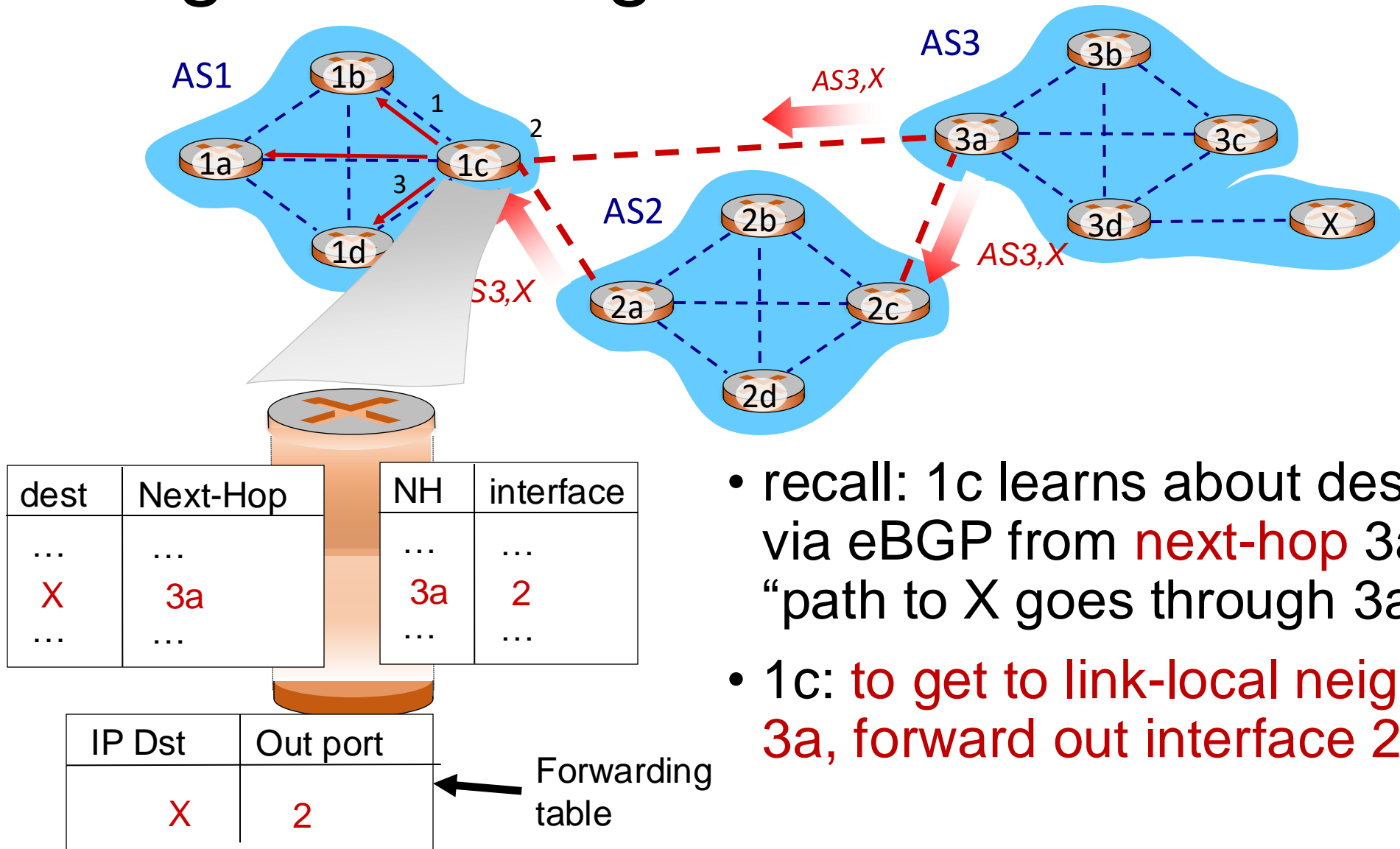
- AS1 gateway router 1c learns path **AS2,AS3,X** from 2a (next hop 2a)
- AS1 gateway router 1c learns path **AS3,X** from 3a (**next hop 3a**)
- Through BGP route selection process, AS1 gateway router 1c chooses path **AS3,X**, and announces path within AS1 via iBGP (**next hop 1c**)

Setting forwarding table entries



- recall: 1a, 1b, 1d learn about dest X via iBGP from **next-hop** 1c: “path to X goes through 1c”
- 1d: **intra-domain routing**: to get to 1c, forward over outgoing local interface 1

Setting forwarding table entries

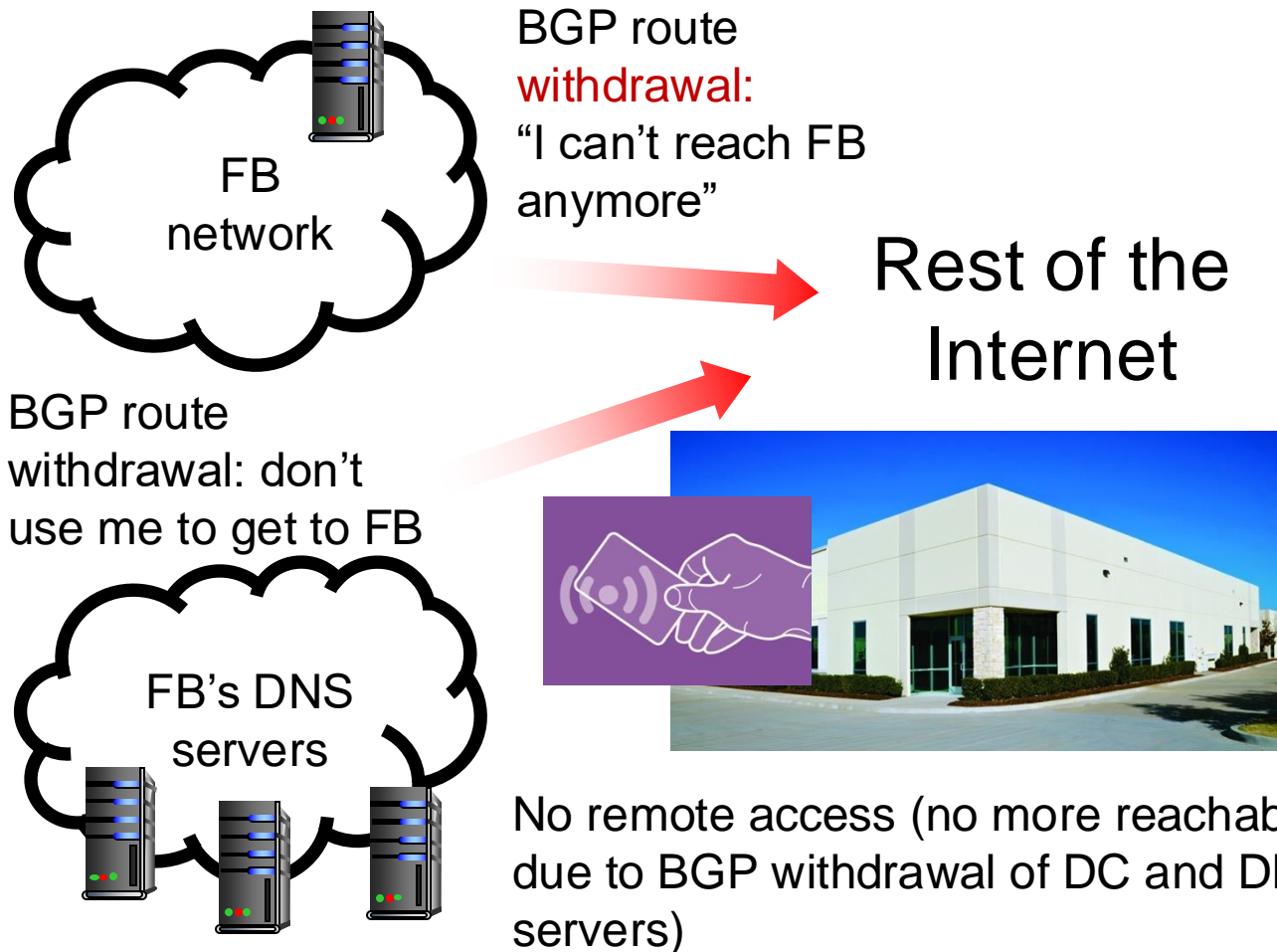


- recall: 1c learns about dest X via eBGP from **next-hop** 3a: “path to X goes through 3a”
- 1c: **to get to link-local neighbor 3a, forward out interface 2**

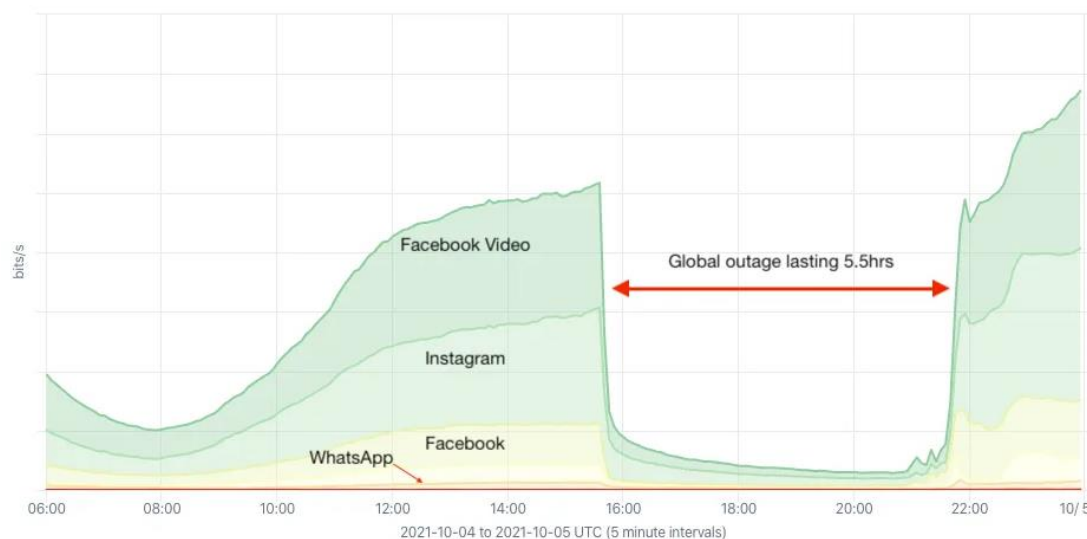
Summary: Inter-domain routing

- **Federation** and **scale** introduce new requirements for routing on the Internet
- **BGP** is *the* protocol that handles Internet routing
- **Path vector**: exchange paths to a destination with attributes
- **Policy-based** import of routes, route selection, and export

BGP's impact: October '21 FB++ outage



Top OTT Service by Average bits/s | Internet Traffic served by Facebook
Oct 04, 2021 06:00 to Oct 05, 2021 00:00 (18h) | Global outage 4-Oct-2021



Restricted physical access (prox can't verify, can't access prox server)

<https://engineering.fb.com/2021/10/05/networking-traffic/outage-details/>

By Doug Madory - <https://www.kentik.com/blog/facebook-historic-outage-explained/>, CC BY 4.0,
<https://commons.wikimedia.org/w/index.php?curid=110816752>

Network Address Translation (NAT)

Background: The Internet's growing pains

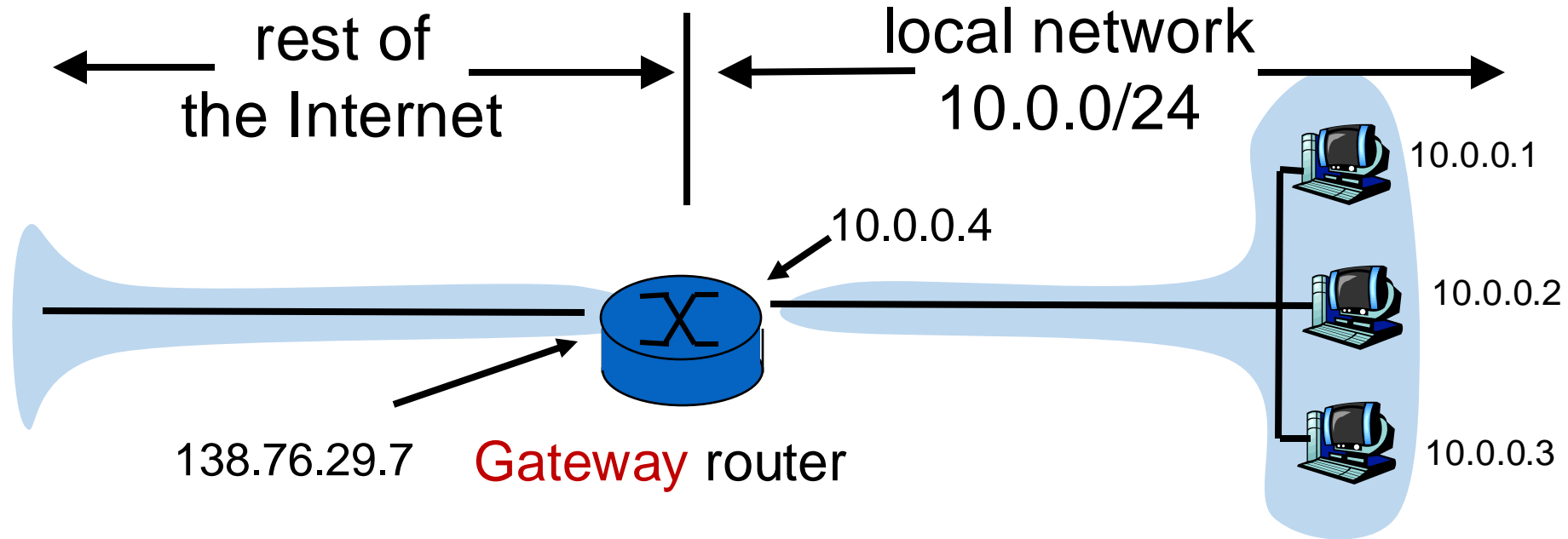
- Networks had incompatible addressing
 - IPv4 versus other network-layer protocols (X.25)
- Entire networks were changing their Internet Service Providers
 - ISPs don't want to route directly to internal endpoints
- **IPv4 address exhaustion**
 - Insufficient large IP blocks even for large networks
 - 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

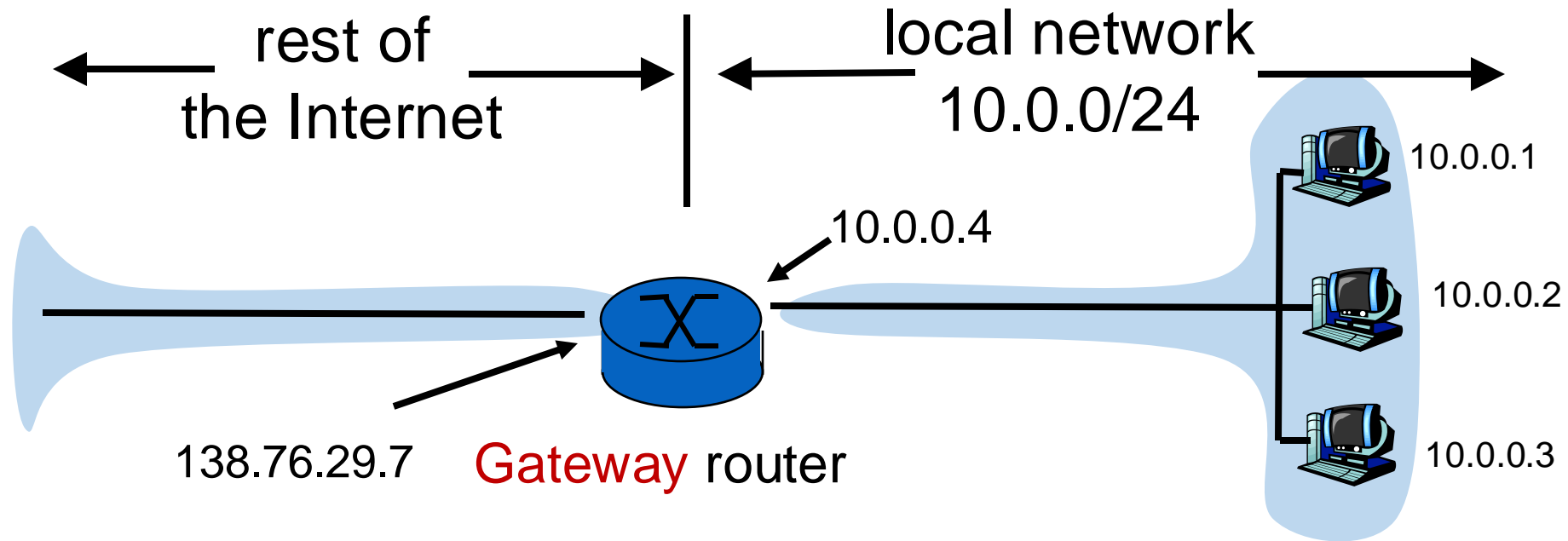
- When a router modifies fields in an IP packet to:
- Enable communication across networks with different (network-layer) addressing formats and address ranges
- Allow a network to change its connectivity to the Internet en masse by modifying the source IP to a (publicly-visible) gateway IP address
- **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, RFC 2663)

Typical NAT setup (NAPT)



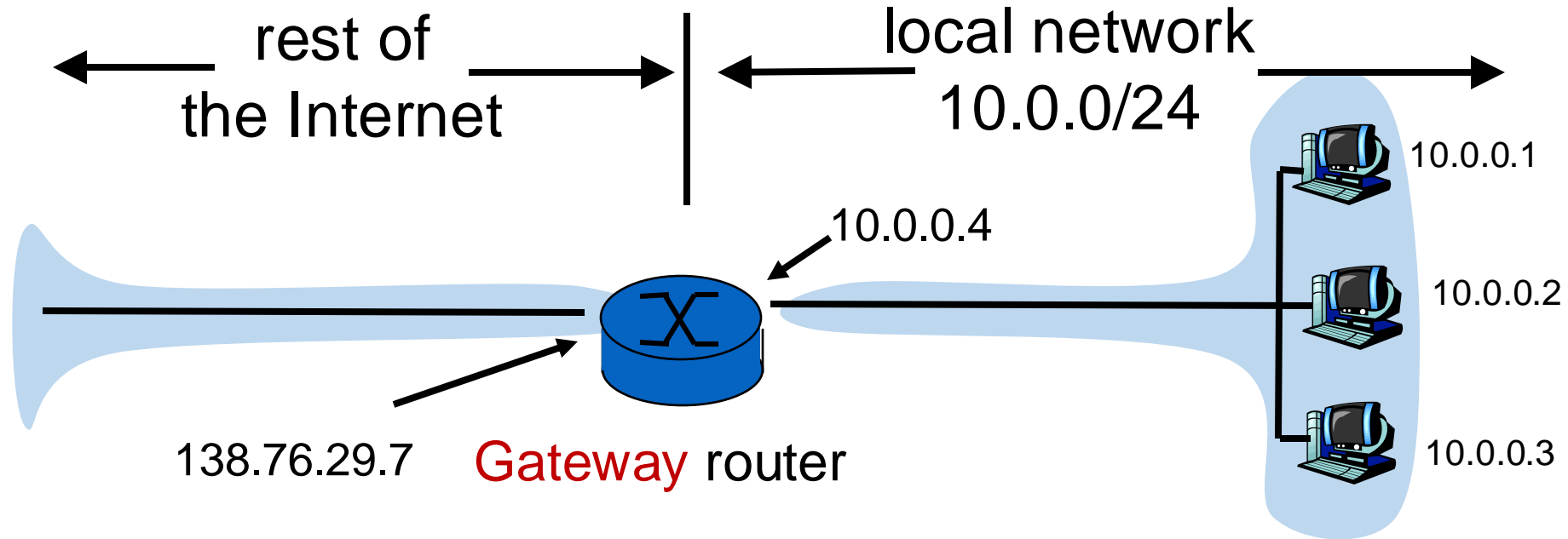
- The gateway's IP, 138.76.29.7 is publicly visible
- The local endpoint IP addresses in 10.0.0/24 are **private**
- **All** datagrams **leaving** local network have the **same source IP** as the **gateway**

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)



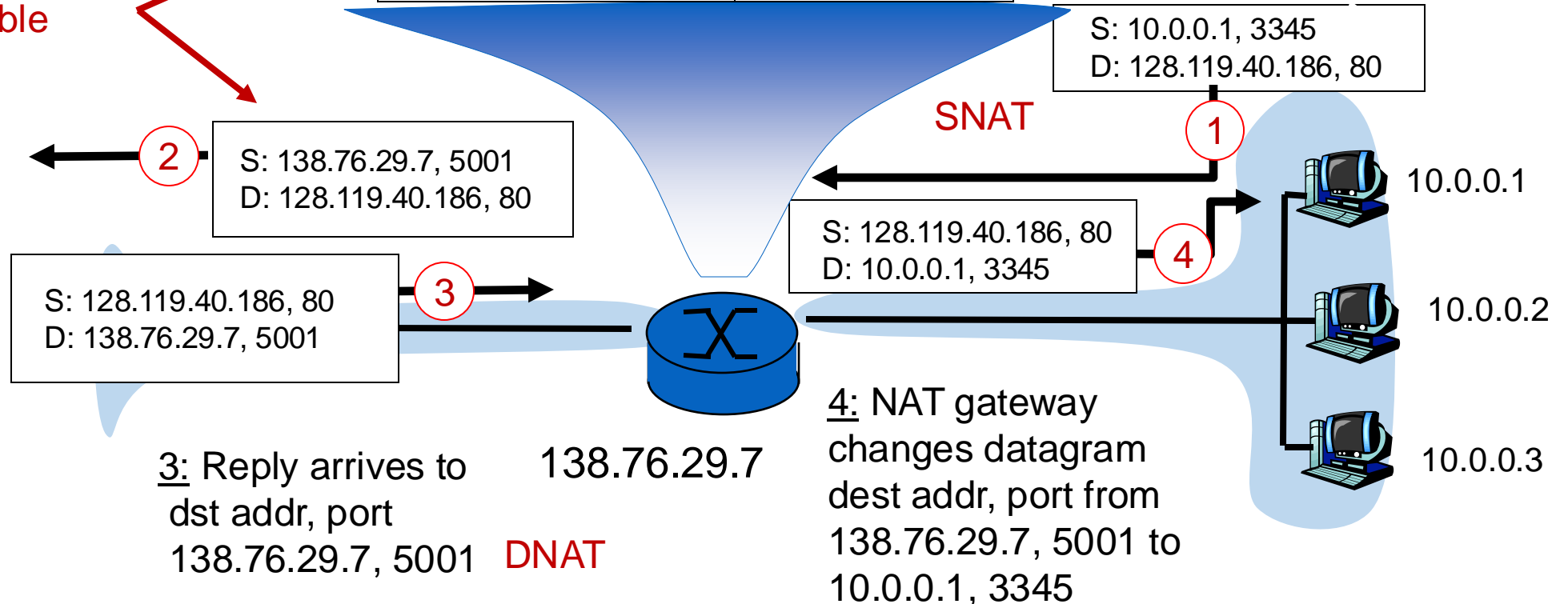
The NAT gateway router accomplishes this by using a **different transport port** for each distinct (transport-level) conversation between the local network and the Internet.

Typical NAT setup (NAPT)

2: NAT router
changes datagram
src addr, port from
10.0.0.1, 3345 to
138.76.29.7, 5001,
Updates table

Translation table	
Internet-side	Local side
138.76.29.7, 5001	10.0.0.1, 3345
..... 4: Map back

1: host 10.0.0.1
sends datagram to
an **external host**,
128.119.40.186, at port 80



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 0xffffffff broadcast 192.168.1.255
    nd6 options=201<PERFORMNUD,DAD>
    media: autoselect
    status: active
[flow:352-S20]$
```



what's my ip address

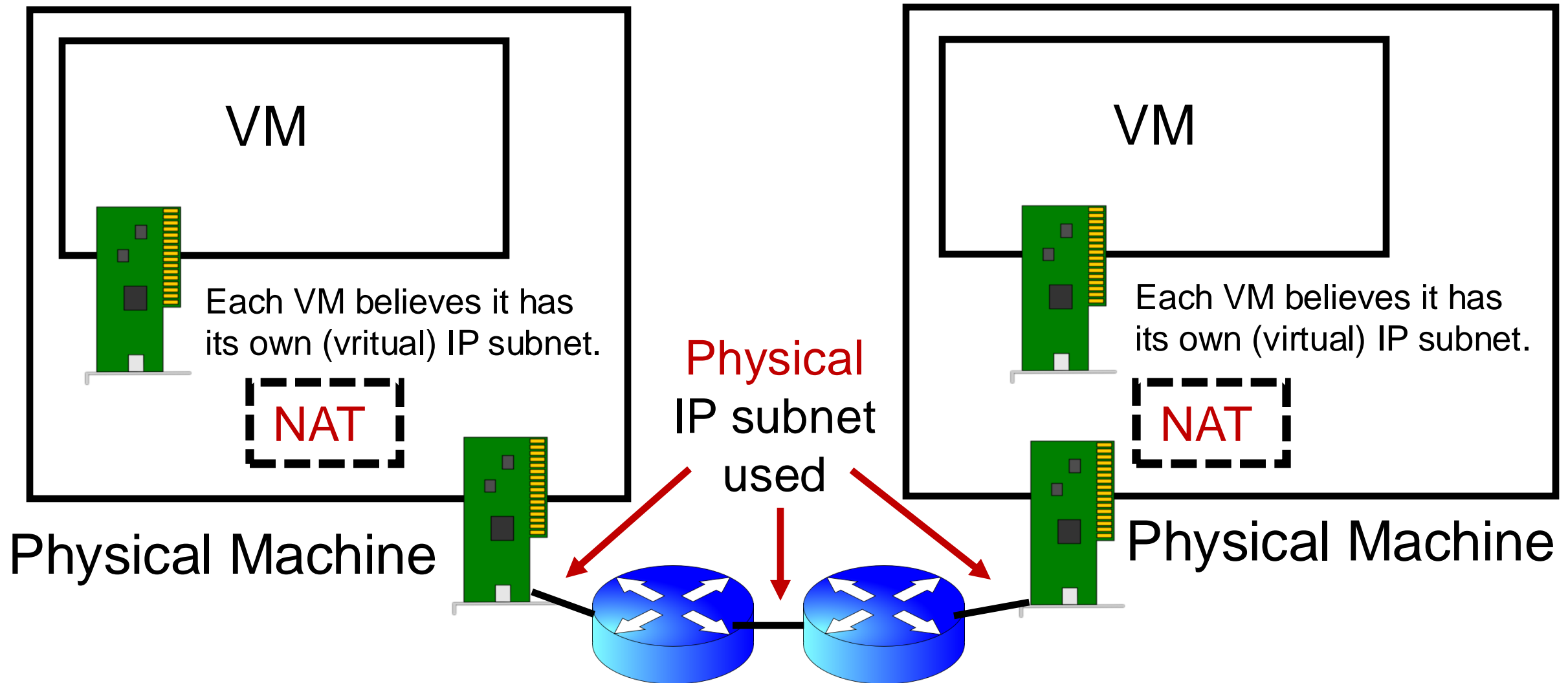


All Images Videos News Maps | Answer

Settings ▾

Your IP address is 74.102.79.209 in [New Brunswick, New Jersey, United States \(08901\)](#)

On public cloud, you're behind NAT



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
- 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