# BGP Policy Routing

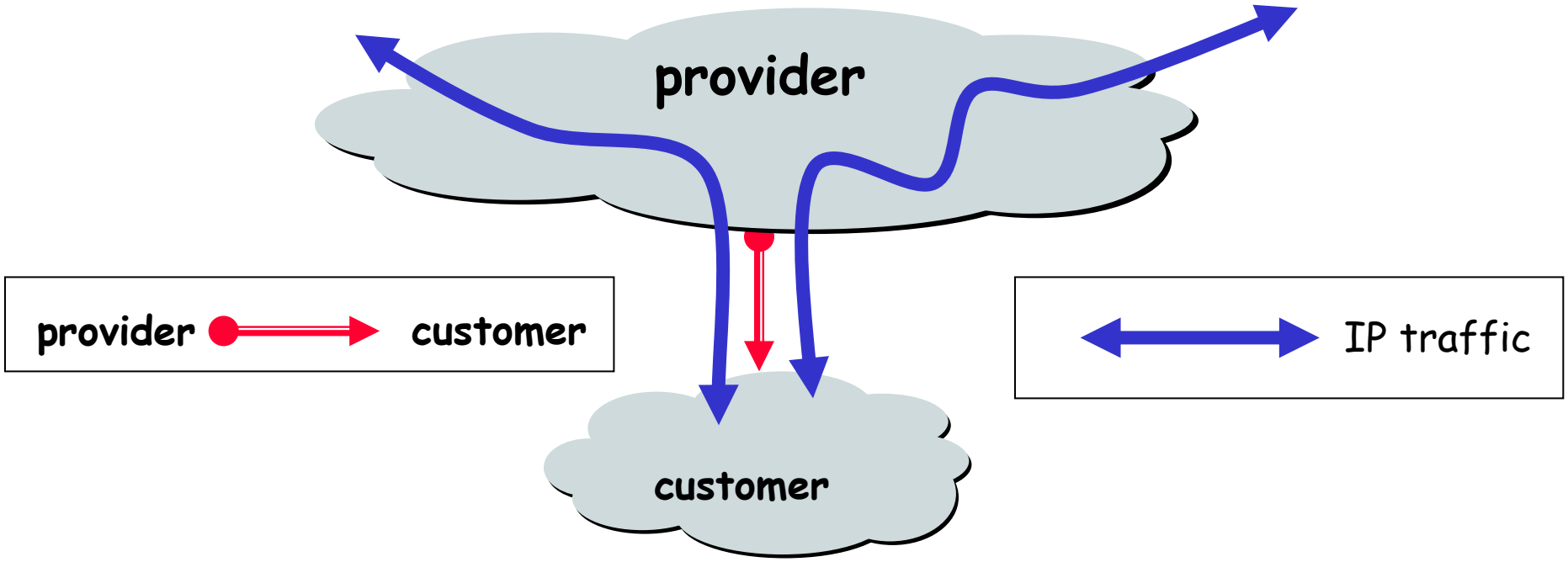
**Richard T. B. Ma**
School of Computing
National University of Singapore

CS 3103: Compute Networks and Protocols

# How is BGP used in practice?

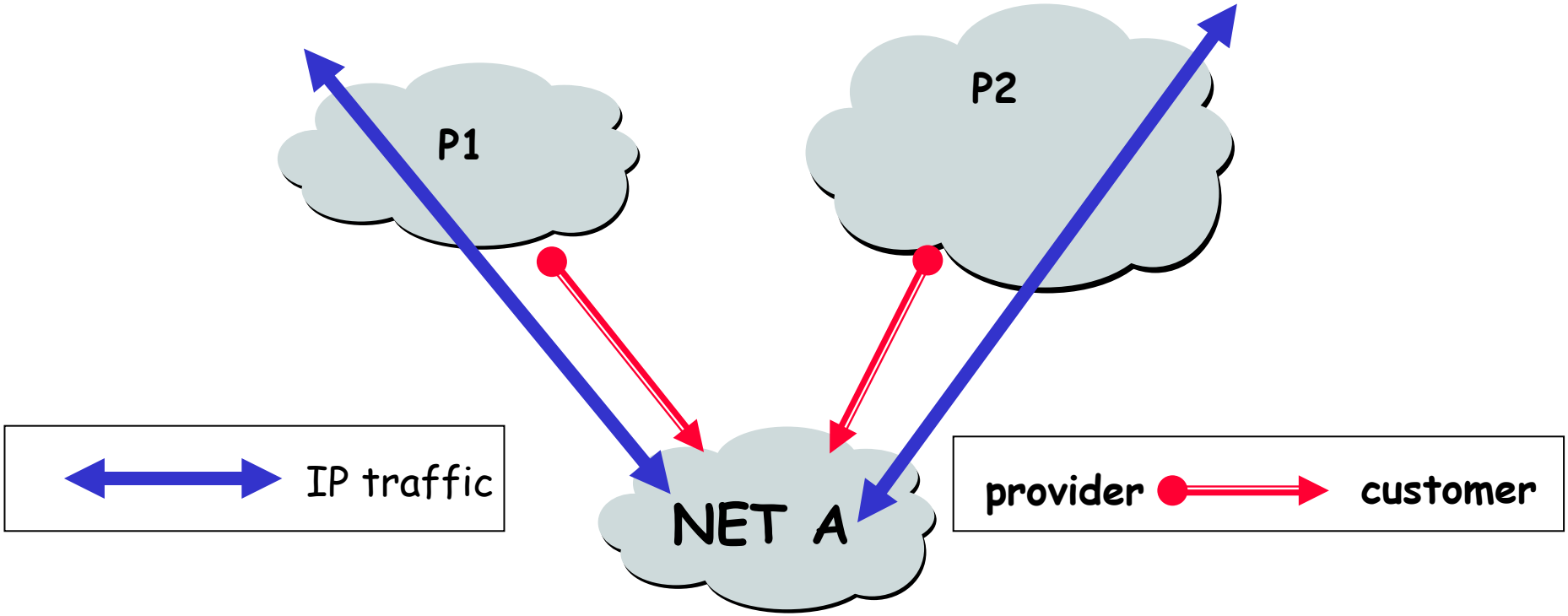
- objectives: used by commercial ISPs to
  - fulfill bilateral agreements with other ISPs
  - minimize monetary costs
  - ensure good performance for customers
- bilateral agreement (between neighboring ISPs)
  - defines who will provide transit for what
  - depends on business relationships
    - Customer-provider relationship
    - Peering relationship
    - Sibling (in an AS-topology) relationship

# Customers and Providers



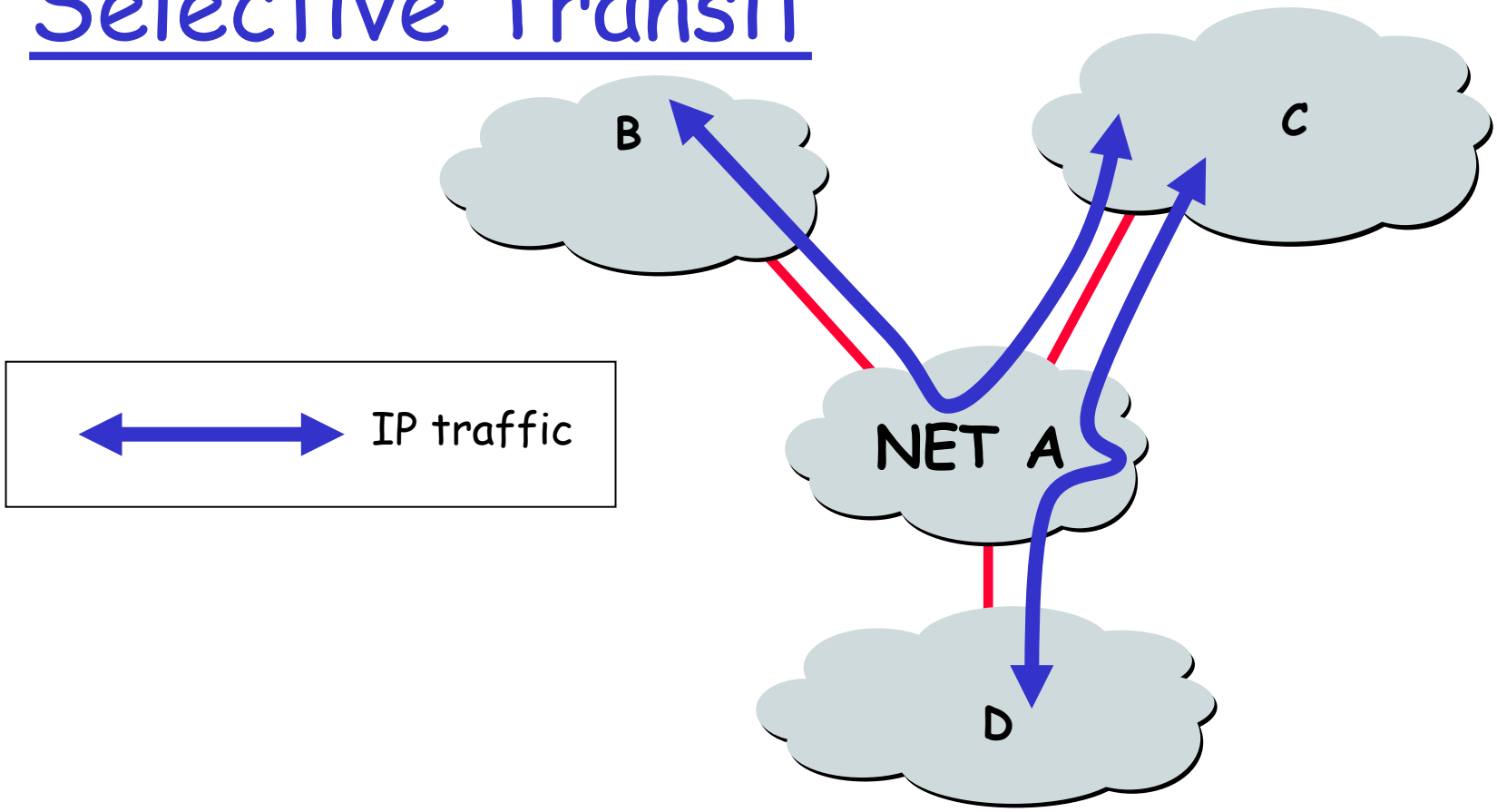

- Customer pays provider for
  - access to the Internet and reachable from anyone
- Provider provides *transit service* for the customer

# Nontransit vs. Transit ASes



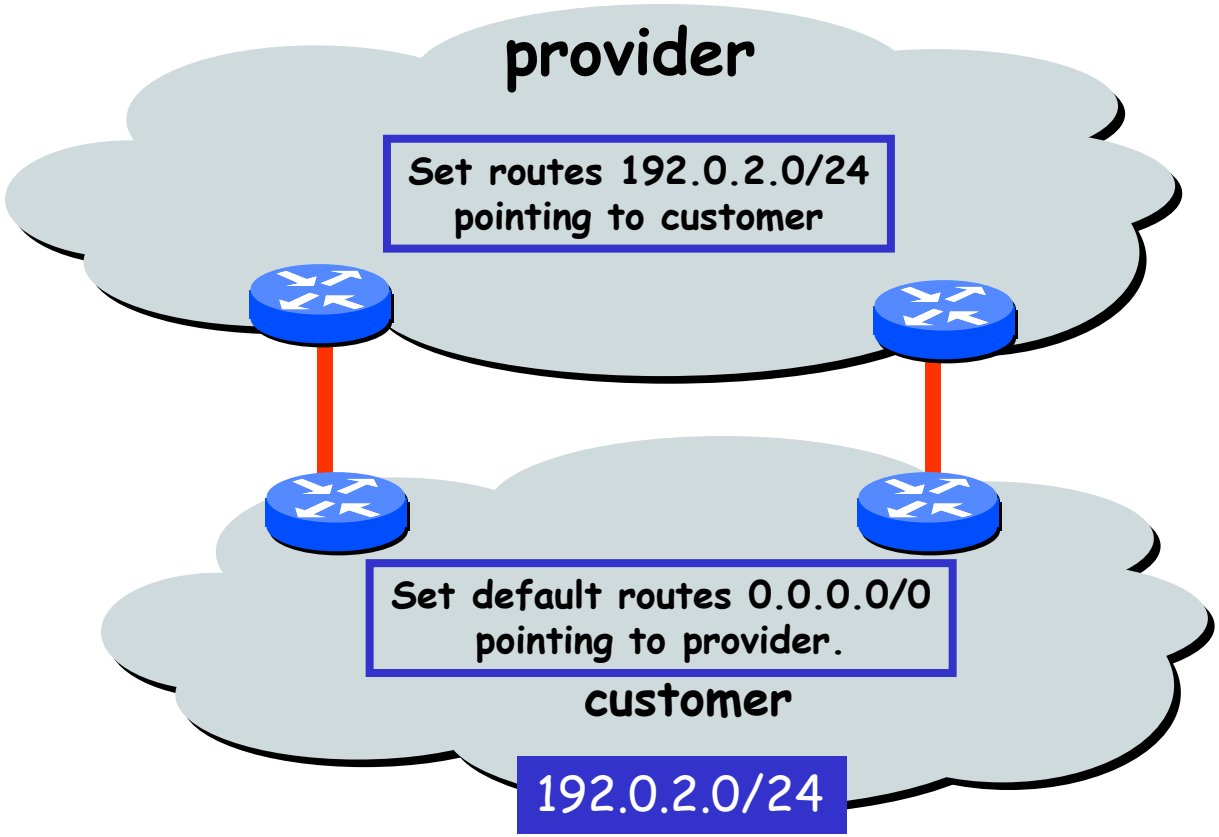

- however, customer doesn't allow traffic go through it
- NET A has two providers, called multi-homing
- traffic should NEVER flows from P1 through NET A to P2
- nontransit AS might be a corporate or campus network, or a "content provider"

# Selective Transit



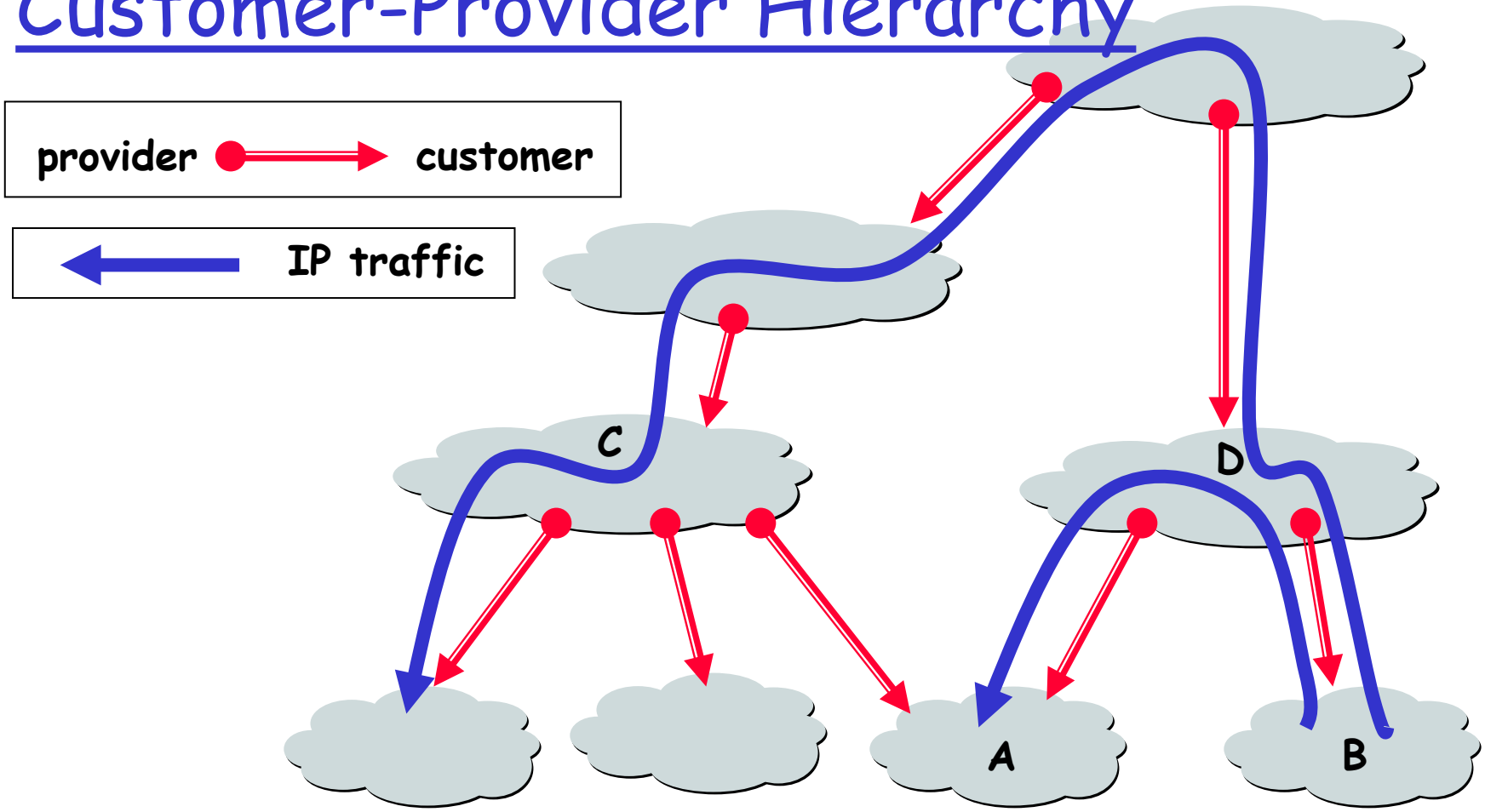

- NET A provides transit between B & C and between C & D
- NET A DOES NOT provide transit Between D & B
- Most transit networks transit in a selective manner…

# Customers Don't Always Need BGP



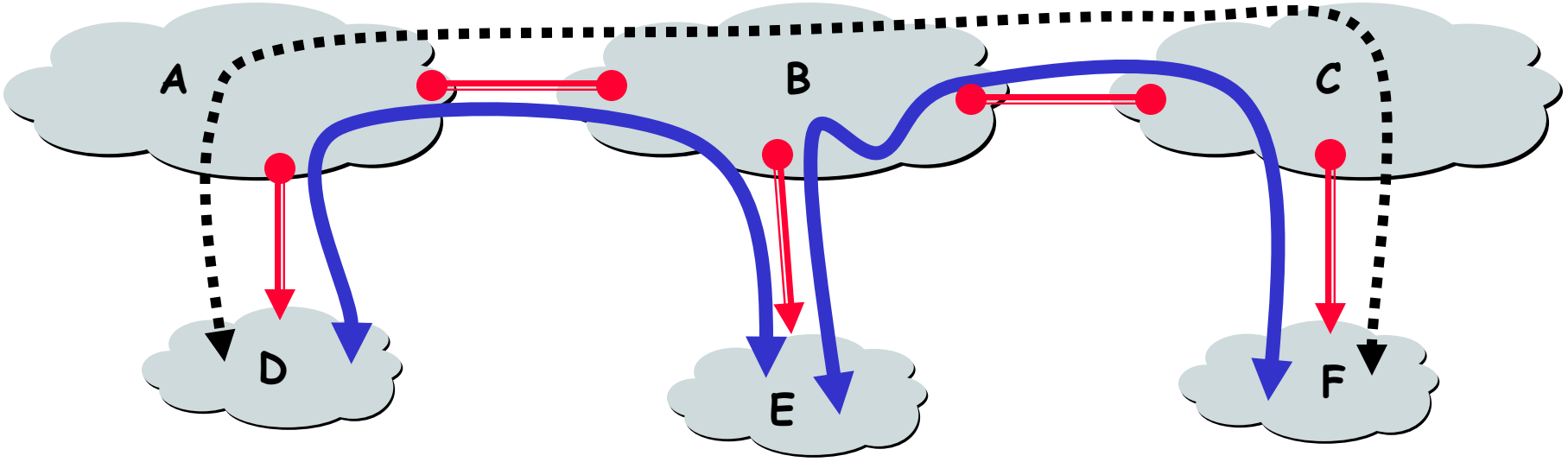

- Static routing is the most common way of connecting an autonomous routing domain to the Internet.

# Customer-Provider Hierarchy

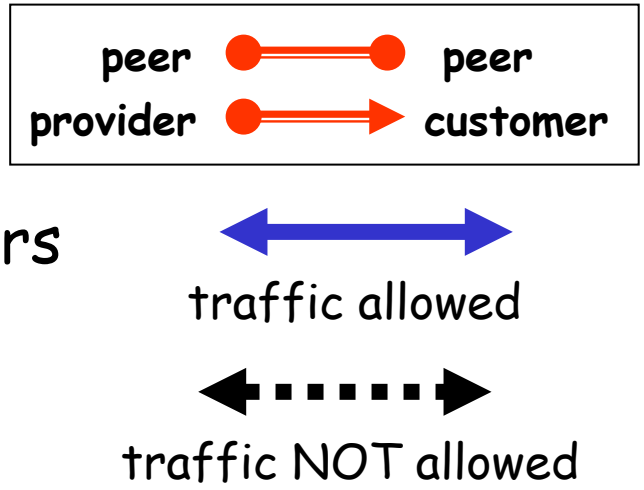



- A multi-home with C and D, one of which is a backup
- A and B are siblings in the AS-level topology

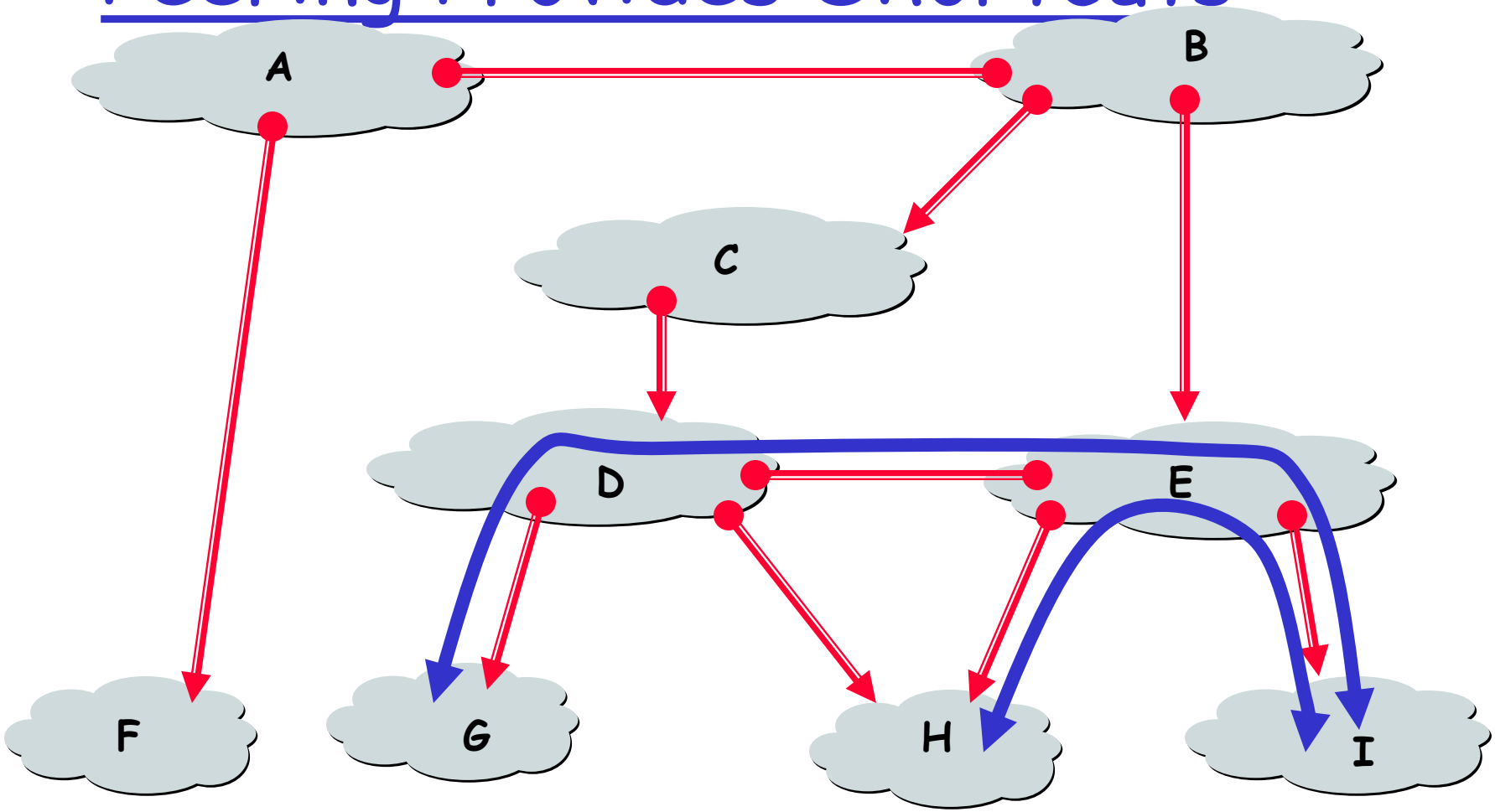
# The Peering Relationship




- Peers provide transit between their respective customers
- don't provide transit between peers
- often don't pay each other (the relationship is settlement-free)



# Peering Provides Shortcuts




Peering also allows connectivity between the customers of "Tier 1" providers.

peer — peer

provider — customer

# Peering Dilemma

| To Peer | Not To Peer |
| --- | --- |
| ❖ reduce upstream transit costs | ❖ you would rather have customers |
| ❖ improve end-to-end performance | ❖ peers are usually your competition |
| ❖ be the only way to connect customers to some part of the Internet (tier-1) | ❖ peering relationships may require periodic renegotiation |

- Peering struggles are by far the most contentious issues in the ISP world!
- Peering agreements are often confidential.

# MCI/Verizon free-peering requirements

**Interconnection Requirements**

1.1 **Geographic Scope**. The Requester shall operate facilities capable of terminating IP customer leased line connections onto a device in at least 50% of the geographic region in which the Verizon Business Internet Network with which it desires to interconnect operates such facilities. This currently equates to **25 states in the United States**, **9 countries in Europe, or 3 countries in the Asia-Pacific region**. The Requester also must have a geographically-dispersed network. In the United States, at a minimum, the Requester **must have a backbone node in each of the following eight geographic regions: Northeast; Mid-Atlantic; Southeast; North Central; South Central; Northwest; Mid-Pacific; and Southwest**.

1.2 **Traffic Exchange Ratio**. The ratio of the aggregate amount of traffic exchanged between the Requester and the Verizon Business Internet Network with which it seeks to interconnect shall be **roughly balanced and shall not exceed 1.8:1.**

1.3 Backbone Capacity. The Requester shall have a fully redundant backbone network, in which the **majority of its inter-hub trunking links** shall have **a capacity of at least 9953 Mbps (OC-192)** for interconnection with Verizon Business-US, **2488 Mbps** (STM-16) for interconnection with Verizon Business-**Europe**, and **622 Mbps (OC-12)** for interconnection with Verizon Business-**ASPAC**.

1.4 **Traffic Volume.** The **aggregate amount of traffic exchanged** in each direction over all interconnection links between the Requester and the Verizon Business Internet Network with which it desires to interconnect shall **equal or exceed 1500 Mbps** of traffic for Verizon Business-US, **150 Mbps** of traffic for Verizon Business-Europe, and **30 Mbps** of traffic for Verizon Business-ASPAC.

… for rest of it see **http://www.verizonbusiness.com/uunet/peering/**

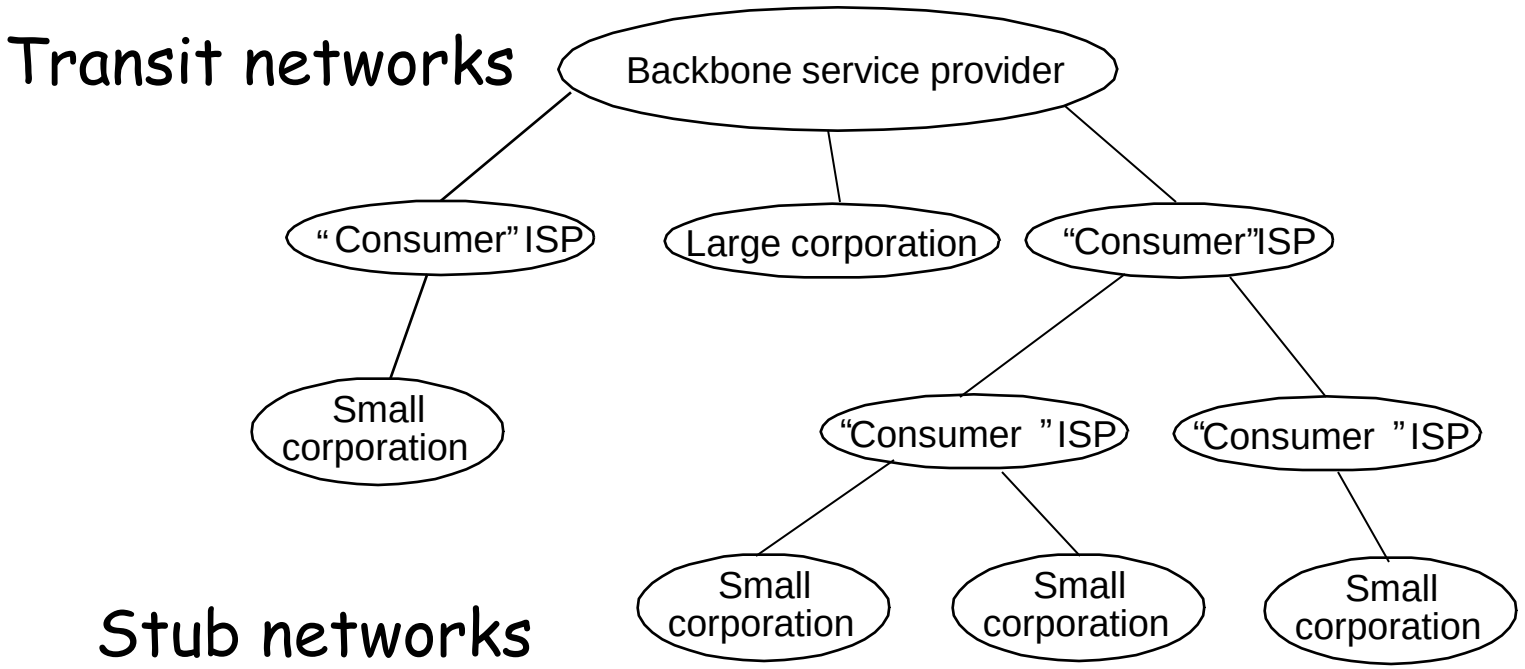
# Tier 1 Ases/ISPs

- have access to the entire Internet only through its settlement-free peering links
- top of the customer-provider hierarchy
- typically large (inter)national backbones
- have no upstream provider
- peer with each other to form a full-mesh
- around 10-12 Ases: AT&T, Sprint, Level 3

# Other ASes

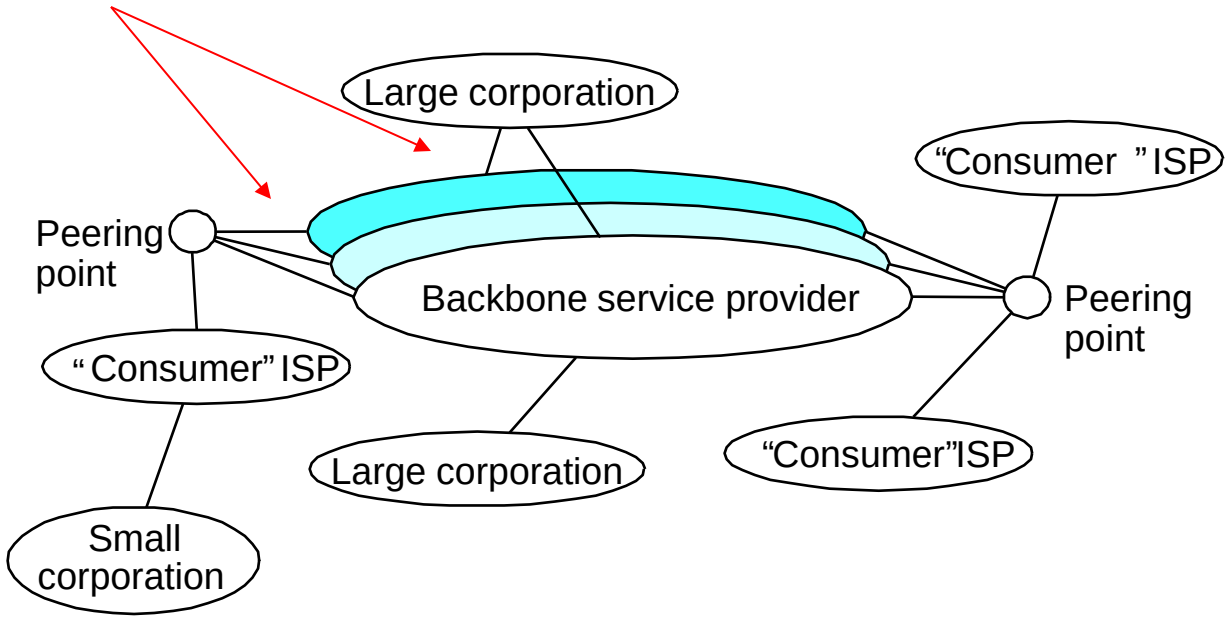
- Lower layer providers (tier-2, …)
  - provide transit to downstream customers
    - but need at least one provider of their own
  - typically have national or regional scope
  - include a few thousand of ASes
- Stub Ases
  - Do not provide transit service
  - Connect to upstream provider(s)
  - Most Ases (e.g., 85-90%)
  - E.g., NUS

# Simplified logical model

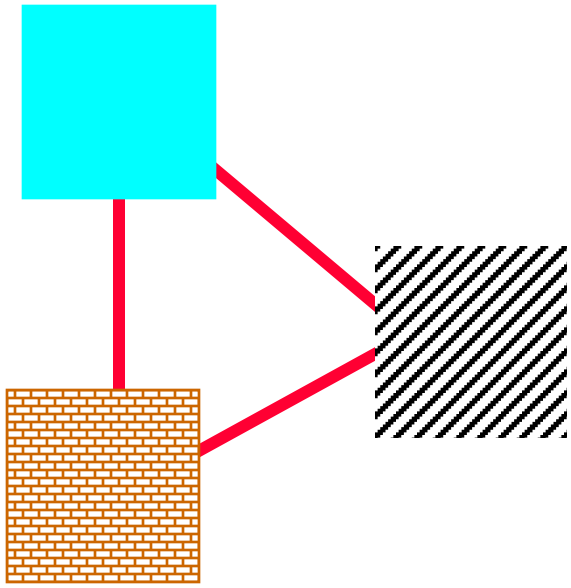


# More realistic competitive view

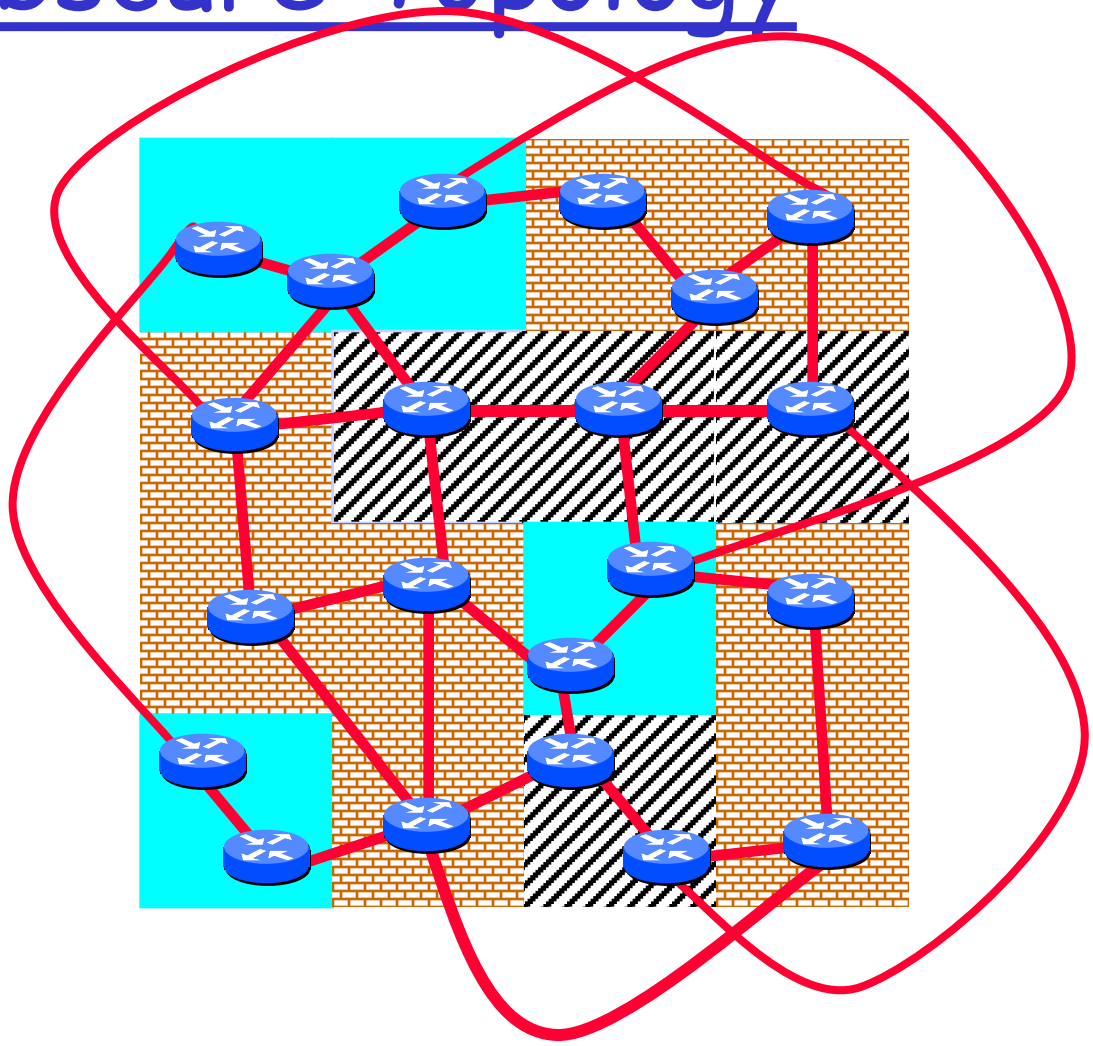
Multi-homing



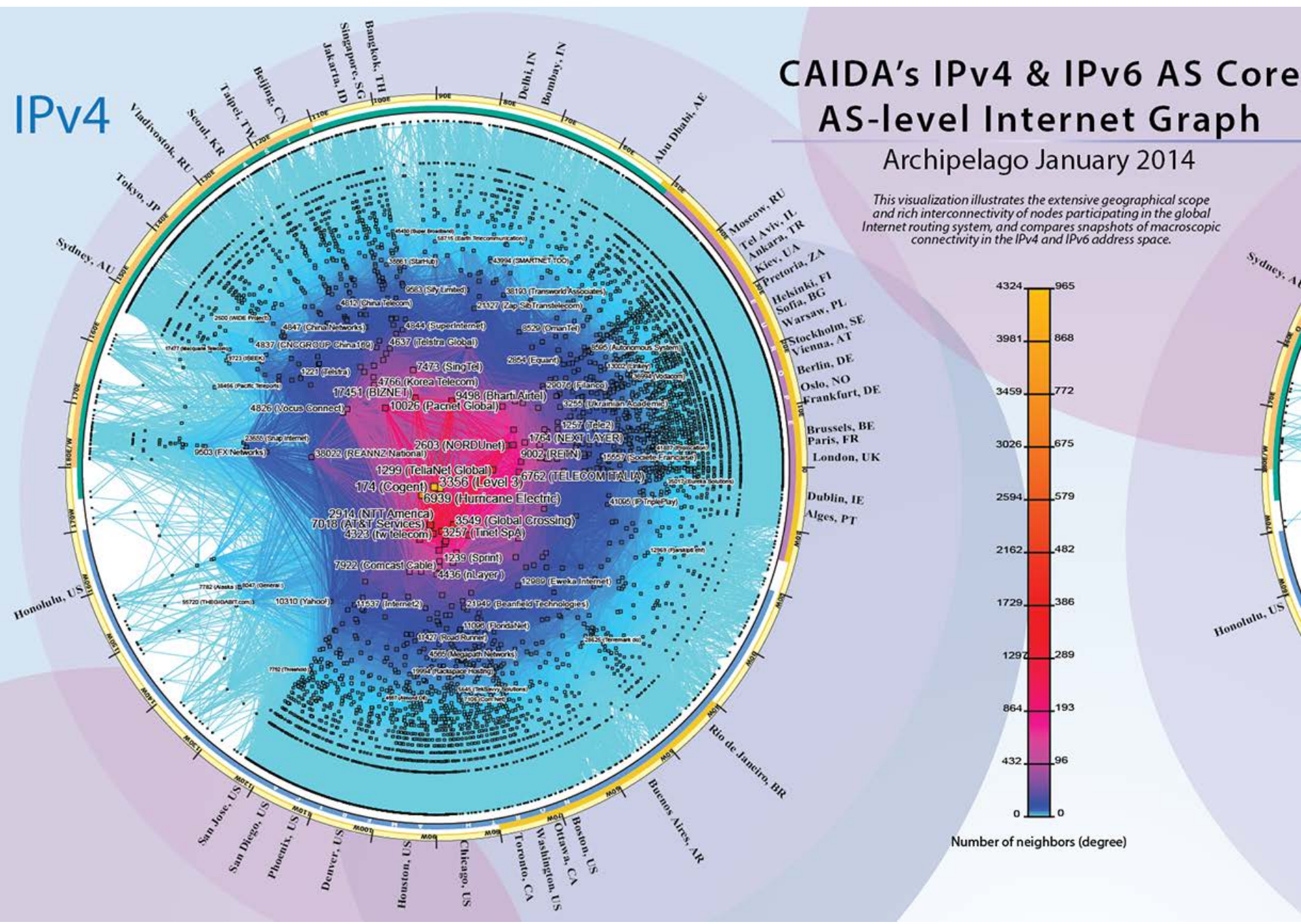
# AS Graphs Obscure Topology



The AS graph
may look like this.



Reality may be closer to this…




http://www.caida.org/research/topology/as_core_network/pics/2014/ascore-2014-jan-ipv4v6-poster-2000x1294.png

# At The Core


7473 (SingTel)
4766 (Korea Telecom)
17451 (BIZNET)
9498 (Bharti Airtel)
10026 (Pacnet Global)
1257 (Tele2)
1764 (NEXT LAYER)
2603 (NORDUnet)
38022 (REANNZ National)
9002 (RETN)
1299 (TeliaNet Global)
6762 (TELECOM ITAL
174 (Cogent)
3356 (Level 3)
6939 (Hurricane Electric)
2914 (NTT America)
7018 (AT&T Services)
3549 (Global Crossing)
4323 (tw telecom)
3257 (Tinet SpA)
1239 (Sprint)
7922 (Comcast Cable)
4436 (nLayer )

AS Ranking | Org Ranking | Information for a single AS | Information for a single Org | Background | Da

**The top ASes ranked by customer cone size are displayed below.**
For information about a specific AS, enter its AS name, its AS number, or the name of the Org of which the AS is a membe

Look up an AS by number or name | Search

Table **shows** 10 of 45658 ASes, **sorted by** number of ASes in customer cone | update view

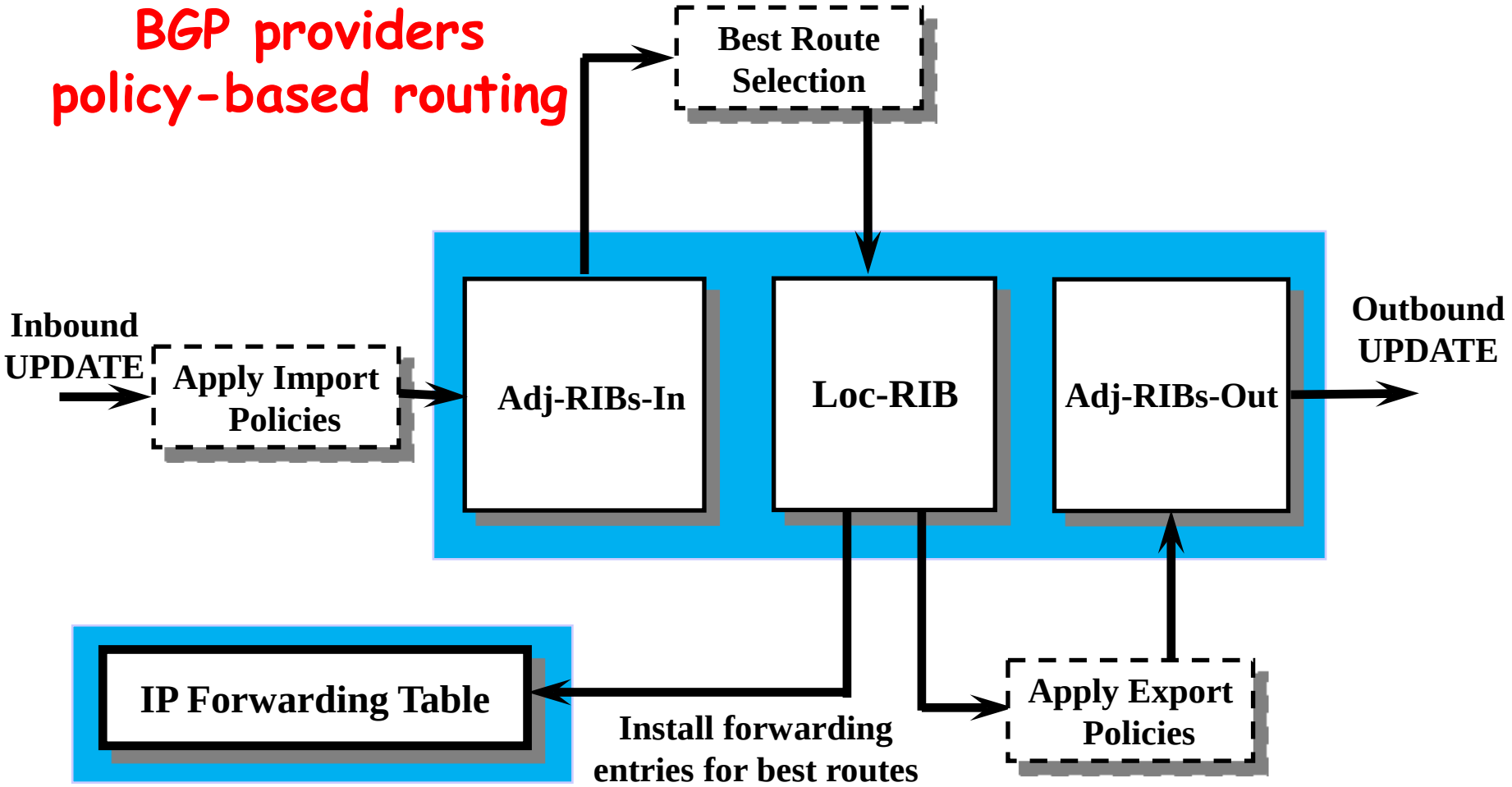
| AS rank | AS number | AS name | Org name | AS Type(s) | Number of ASes | Number of IPv4 Prefixes | Number of IPv4 Addresses |
|---|---|---|---|---|---|---|---|
| 1 | 3356 | LEVEL3 | Level 3 Communications, Inc. | Tr Co | 25,318 | 322,403 | 1,562,430,335 |
| 2 | 174 | COGENT-174 | Cogent Communications | Tr | 17,484 | 208,063 | 744,220,957 |
| 3 | 3257 | TINET-BACK... | Tinet SpA | Tr Co | 15,623 | 222,392 | 846,663,937 |
| 4 | 1299 | TELIANET | TeliaSonera International Carrier | Tr Co | 15,178 | 228,540 | 785,632,128 |
| 5 | 2914 | NTT-COMMUN... | NTT America, Inc. | Tr Co | 14,876 | 224,278 | 929,277,565 |
| 6 | 3549 | LVLT-3549 | Level 3 Communications, Inc. | Tr Co | 10,586 | 172,217 | 560,436,792 |
| 7 | 6453 | AS6453 | Tata Communications | Tr Co | 10,229 | 167,716 | 610,754,120 |
| 8 | 6762 | SEABONE-NET | TELECOM ITALIA SPARKLE S.p.A. | Tr Ac | 9,904 | 129,816 | 405,609,356 |
| 9 | 6939 | HURRICANE | Hurricane Electric, Inc. | Tr Co | 6,240 | 73,271 | 288,745,110 |
| 10 | 1273 | CW | Cable&Wireless Worldwide | Tr | 5,945 | 69,712 | 250,224,888 |

http://as-rank.caida.org/

# BGP Routing Information Bases

- What is a route in a BGP speaker?
  - route = prefix + attributes = NLRI + Path Attributes
- How about all the routes in a BGP speaker?
  - Routing Information Bases (RIBs)
  - RIBs = Adj-RIBs-In + Loc-RIB + Adj-RIBs-Out
    - Adj-RIBs-In: unprocessed routes from peers via inbound UPDATE; input for decision making
    - Loc-RIB: selected local routes used by the router
    - Adj-RIBs-Out: selected for advertisement to peers

# BGP Decision Process: Overview



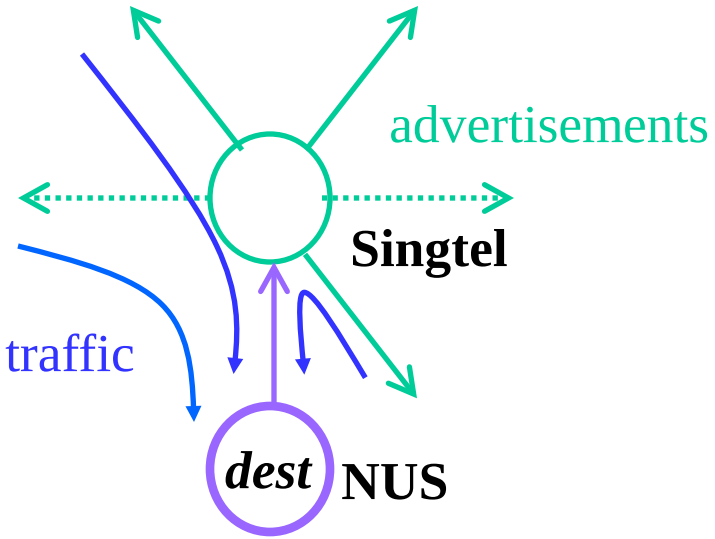
# BGP: applying policy to route

- **Import policy**
  - filter unwanted routes from neighbor
    - e.g., prefix that your customer does not own
  - used to rank customer routes over peer routes
  - 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., export only customer routes to peers & providers
  - manipulate attribute to control what they see
    - e.g., make paths look artificially longer (AS prepending)

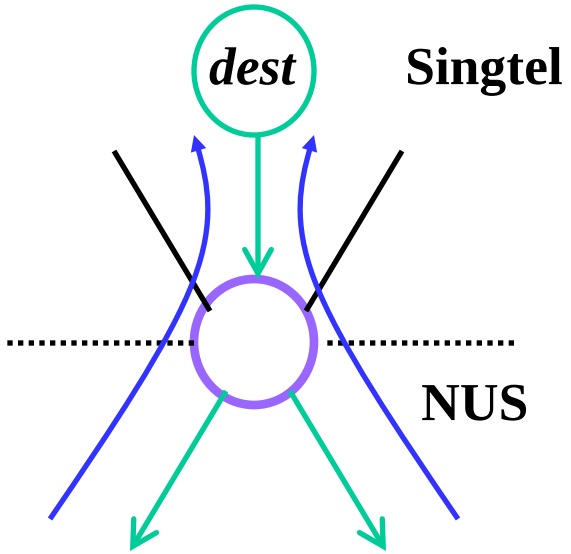
# Customer-Provider Relationship

- Customer pays provider for access to Internet
  - Provider exports customer's routes to everybody
  - Customer exports provider's routes to customers

Traffic **to** the customer

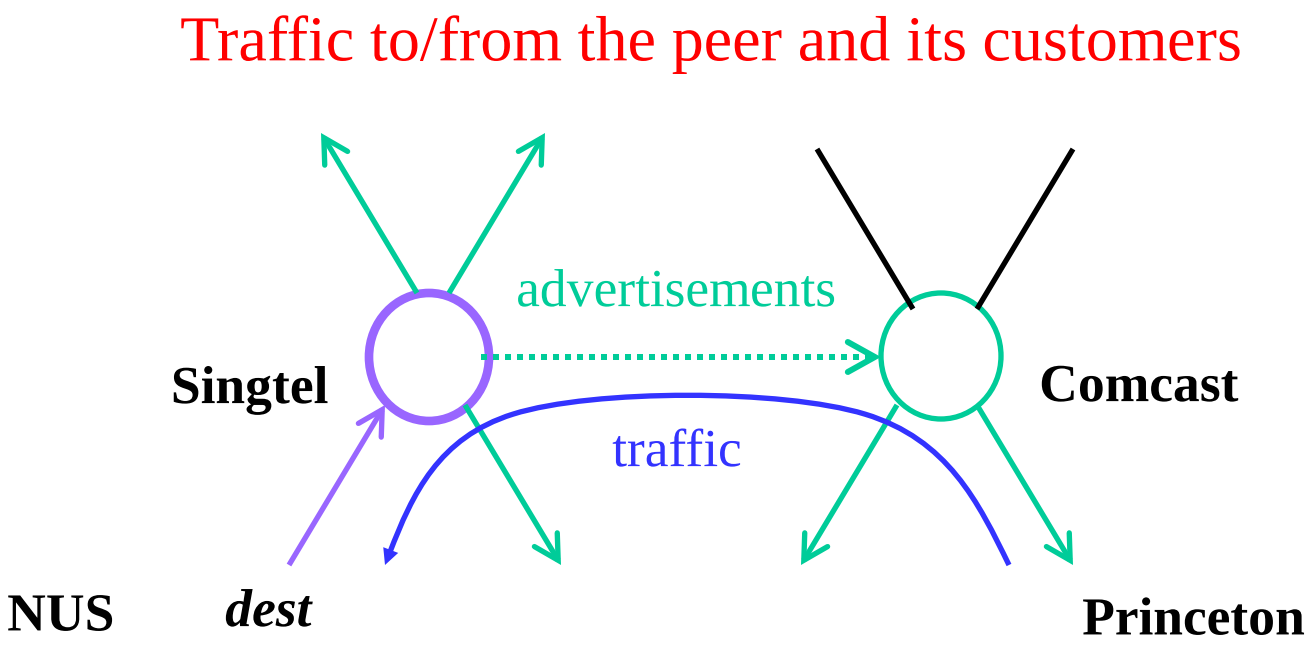



Traffic **from** the customer

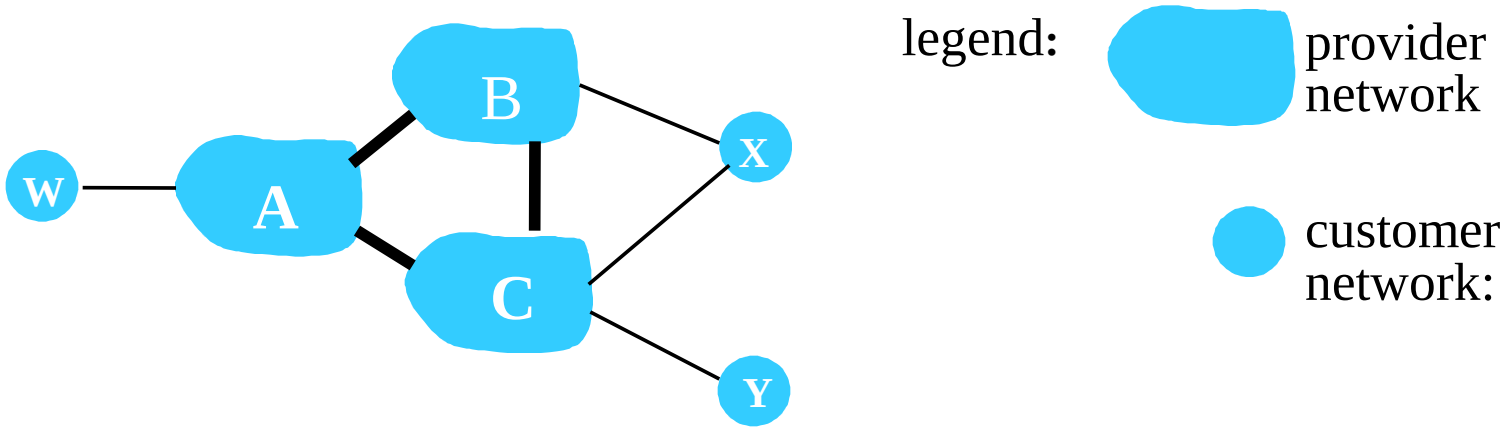


# Peer-Peer Relationship

- Peers exchange traffic between customers
  - AS exports *only* customer routes to a peer
  - AS exports a peer's routes *only* to its customers

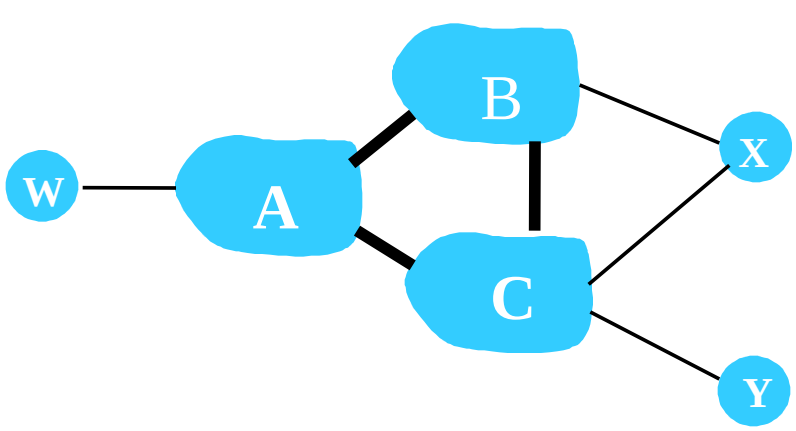


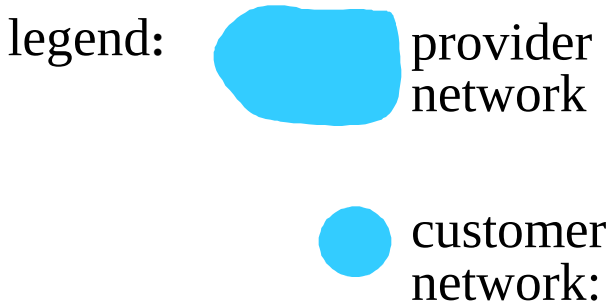
# BGP routing policy




- A,B,C are provider networks
- X,W,Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - X does not want to route from B via X to C
  - .. so X will not advertise to B a route to C

# BGP routing policy



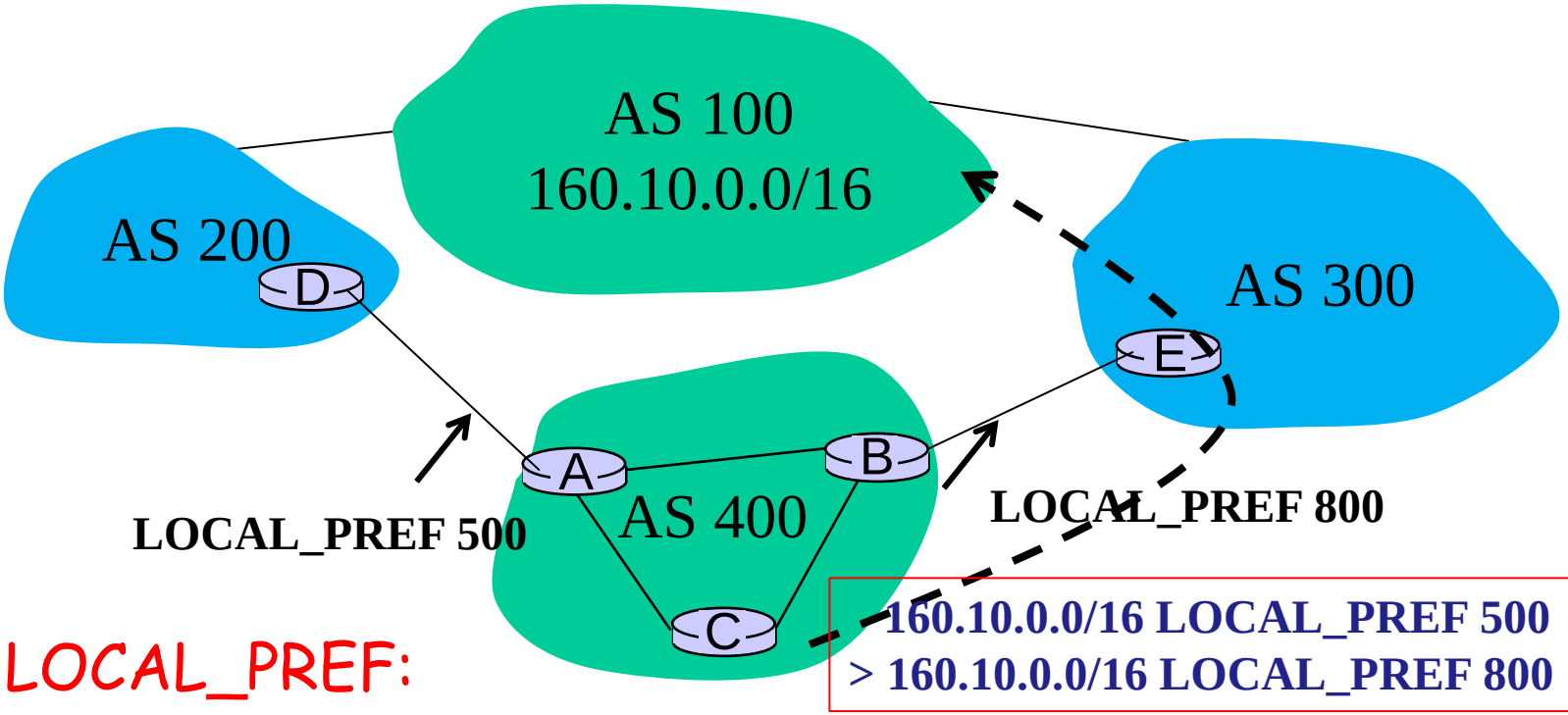




- A advertises path AW to B
- B advertises path BAW to X
- Should B advertise path BAW to C?
  - No way! B gets no "revenue" for routing CBAW since neither W nor C are B's customers
  - B wants to force C to route to w via A
  - B wants to route *only* to/from its customers!

# BGP best route selection

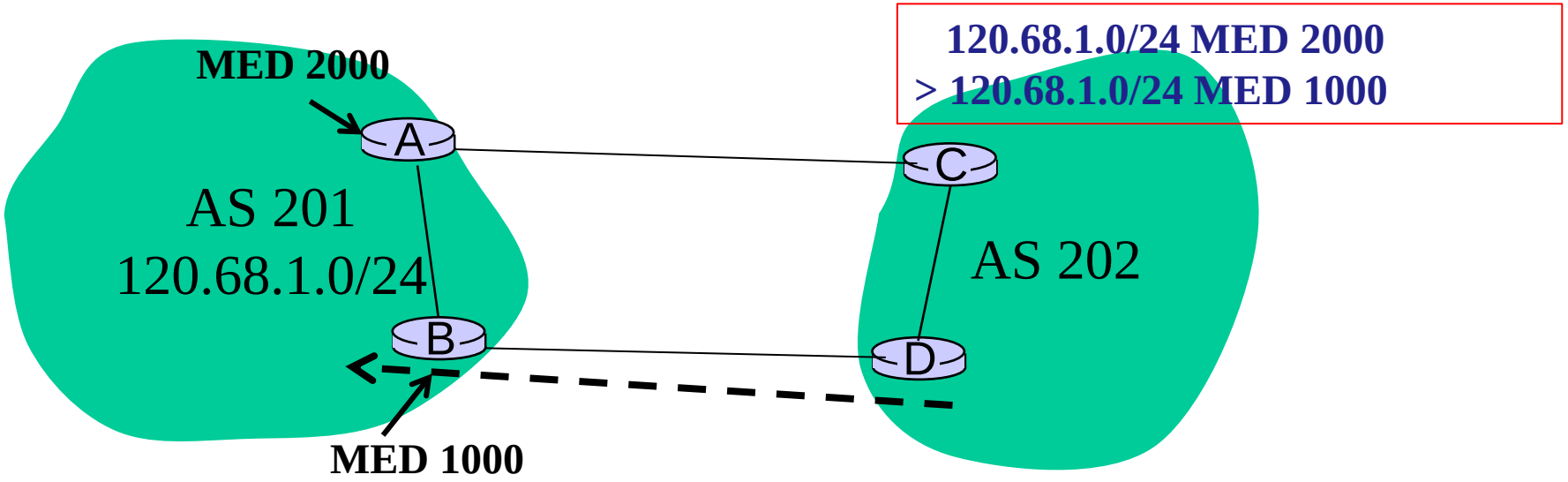
1. Calculation of degree of preference
   - If the route is learned from an internal peer, use LOCAL_PREF attribute or preconfigured policy
   - Otherwise, use preconfigured policy
2. Route selection (recommended process)
   - Highest degree of LOCAL_PREF (or the only route to the destination), and then tie breaking conditions on:
   - Smallest number of AS numbers in AS_PATH attribute
   - Lowest origin number in ORIGIN attribute
   - Most preferred MULTI_EXIT_DISC attribute
   - Routes from eBGP are preferred (over iBGP)
   - Lowest interior cost based on NEXT_HOP attribute

# LOCAL_PREF attribute




- LOCAL_PREF:
  - 4-byte unsigned integer (default value 100)
  - for a BGP speaker to inform its other internal peers of its degree of preference for a route
  - should include in UPDATE messages that are sent to internal peers; should not send to external peers

# MULTI_EXIT_DISC attribute

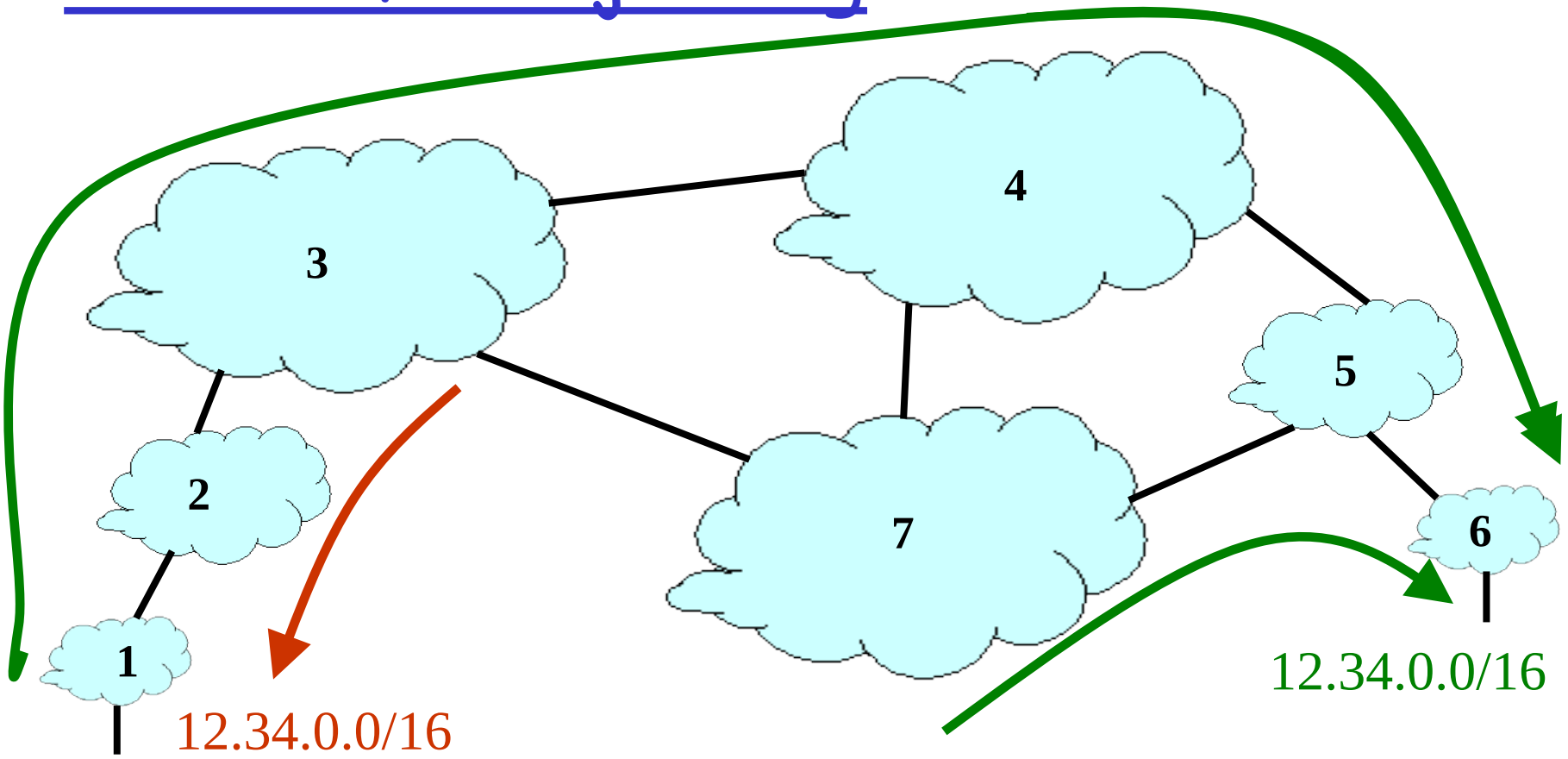



- MULTI_EXIT_DISC (MED):
  - 4-byte unsigned integer (default value 0)
  - for a BGP speaker to discriminate among multiple entry points to a neighboring AS to control inbound traffic
  - if received over eBGP, may be propagated over iBGP, but must not be further propagated to neighboring ASes

# COMMUNITY attribute

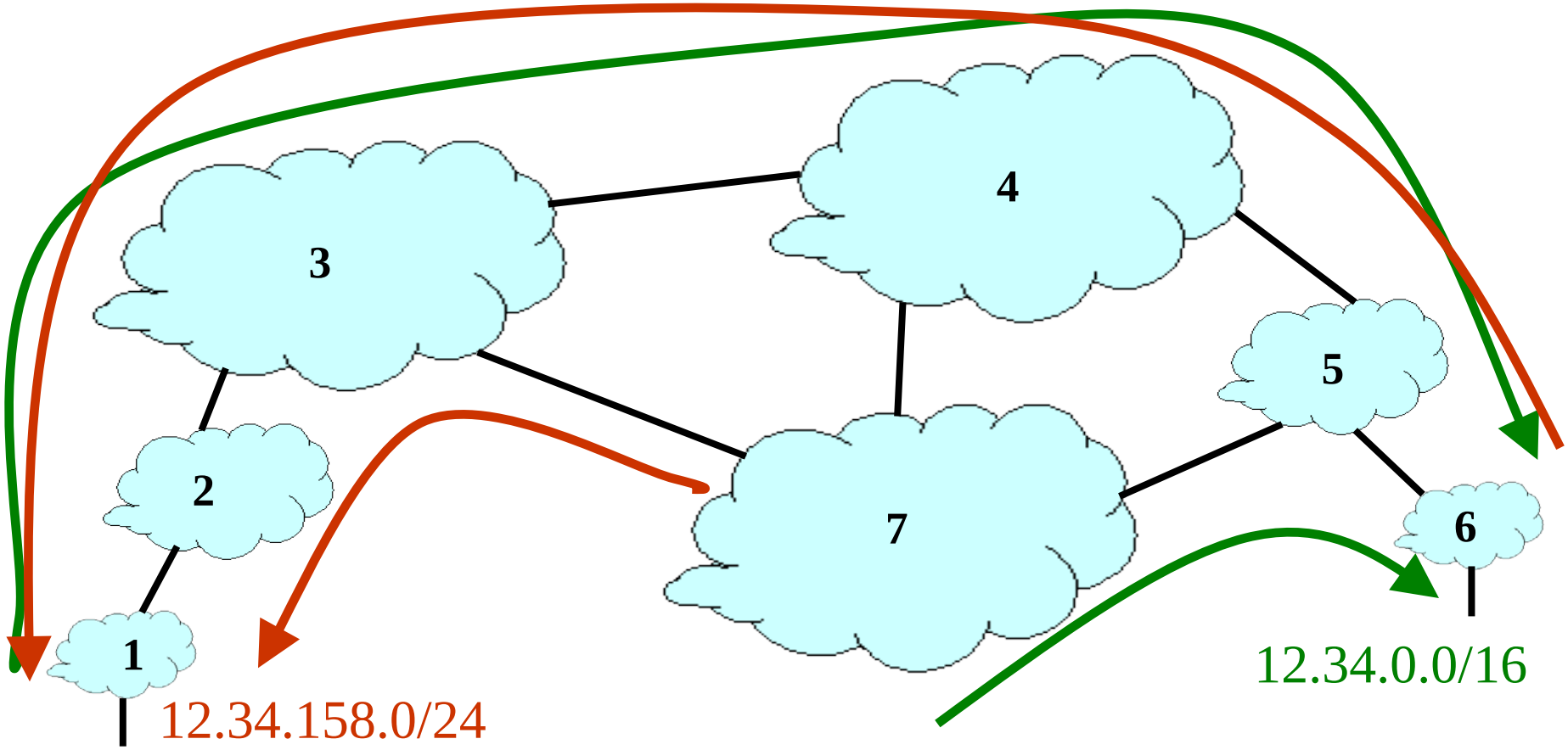
- Described in RFC 1997
- 4-byte integer value
- Used to group destinations
  - Each destination could be member of multiple communities
- Very useful in applying policies within and between Ases
  - import and export policies based on the COMMUNITY attributes

# BGP Prefix Hijacking




- Consequences for the affected ASes
  - Blackhole: data traffic is discarded
  - Snooping: data traffic is inspected, and then redirected
  - Impersonation: data traffic is sent to bogus destinations

# BGP Subprefix Hijacking




- Originating a more-specific prefix
  - Every AS picks the bogus route for that prefix
  - Traffic follows the longest matching prefix

# BGP prefix hijack example

- 18:47:00, 24 Feb 2008, Pakistan Telecom (AS 17557) began advertising 208.65.153.0/24, a more specific route of the prefix 208.65.152.0/22 used by YouTube (AS 36561)

- found 20 mins later and took ~2 hours to restore

http://research.dyn.com/2008/02/pakistan-hijacks-youtube-1/

- can be visualized by BGPlay

https://stat.ripe.net/special/bgplay

18:47:45 1st hijacked route propagated in Asia, AS path 3491 17557
18:48:00 9 big trans-Pacific providers carrying hijacked route
18:48:30 47 DFZ providers now carrying the bad route
18:49:00 most of the DFZ now carrying the bad route (93 ASNs)
18:49:30 all who will carry the hijacked route have it (97 ASNs)

20:07:25 AS 36561 advertises the hijacked /24 to its providers
20:07:30 several DFZ providers stop carrying the erroneous route
20:08:00 many downstream providers also drop the bad route
20:08:30 40 providers have stopped using the hijacked route
20:18:43 two more specific /25 routes are first seen from 36561
20:19:37 25 more providers prefer the /25 routes from 36561
20:28:12 peers of 36561 see the routes advertised at 20:07

20:50:59 attempted prepending, AS path was 3491 17557 17557
20:59:39 hijacked prefix is withdrawn by 3491, disconnected 17557

# Preventing (Sub)Prefix Hijacking

- Best common practice for route filtering
  - Each AS filters routes announced by customers
  - E.g., based on the prefixes the customer owns
- But not everyone applies these practices
  - Hard to filter routes initiated from far away
  - So, BGP remains very vulnerable to hijacks
- Other techniques
  - Secure extensions to BGP (e.g., S-BGP, soBGP)
  - Anomaly detection of suspected hijacks

# How is BGP used in practice?

- three classes of "knobs"
  - preference: add/delete/modify attributes
  - filtering: inbound/outbound filtering
  - tagging: e.g., COMMUNITY attribute
- applications
  - business relationships
    - Influencing the decision process (LOCAL_PREF)
    - Controlling route export (COMMUNITY)
  - traffic engineering
    - Inbound traffic control (MED, AS prepending)
    - Outbound traffic control (LOCAL_PREF, IGP cost)
    - Remote control (COMMUNITY)