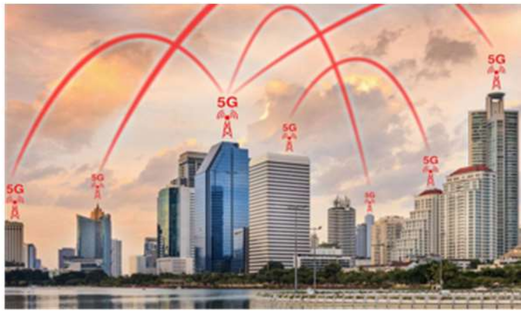


Computer Networks

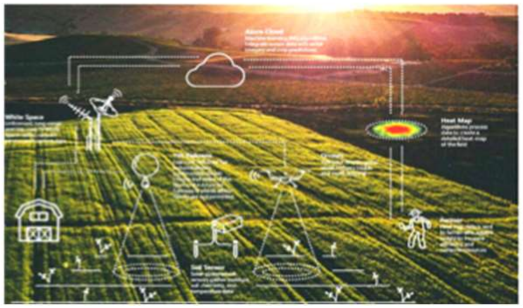
CMSC 417 : Spring 2024

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

Topic: BGP
(Textbook chapter 4)

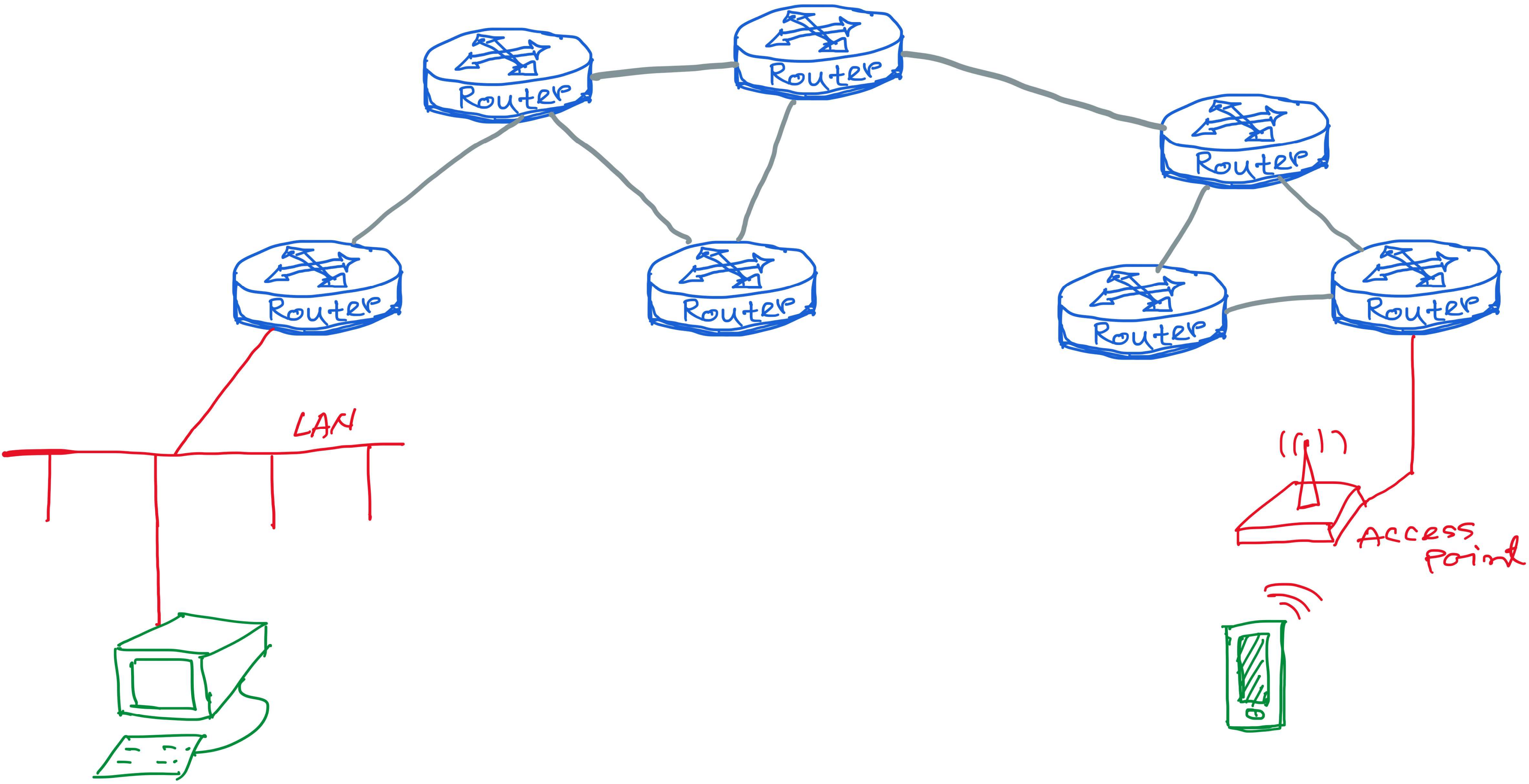
Nirupam Roy
Tu-Th 2:00-3:15pm
CSI 2117

April 30th, 2024



Inter-domain routing

Our (improved) view of the Internet



Making routing scalable

our routing study thus far - idealized

- **all routers identical**
- **network "flat"**

… ***not* true in practice**

scale: with billions of destinations:

- Can't store all destinations in routing tables!
- routing table exchange would swamp links!

administrative autonomy

- internet = network of networks
- each network admin may want to control routing in its own network

Autonomous Routing Domains

A collection of physical networks glued together using IP, that have a unified administrative routing policy.

- **Campus networks**
- **Corporate networks**
- **ISP Internal networks**
- **…**

Autonomous Systems (ASes)

An autonomous system is an autonomous routing domain that has been assigned an Autonomous System Number (ASN).

> … the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it.
>
> RFC 1930: Guidelines for creation, selection,
> and registration of an Autonomous System

AS Numbers (ASNs)

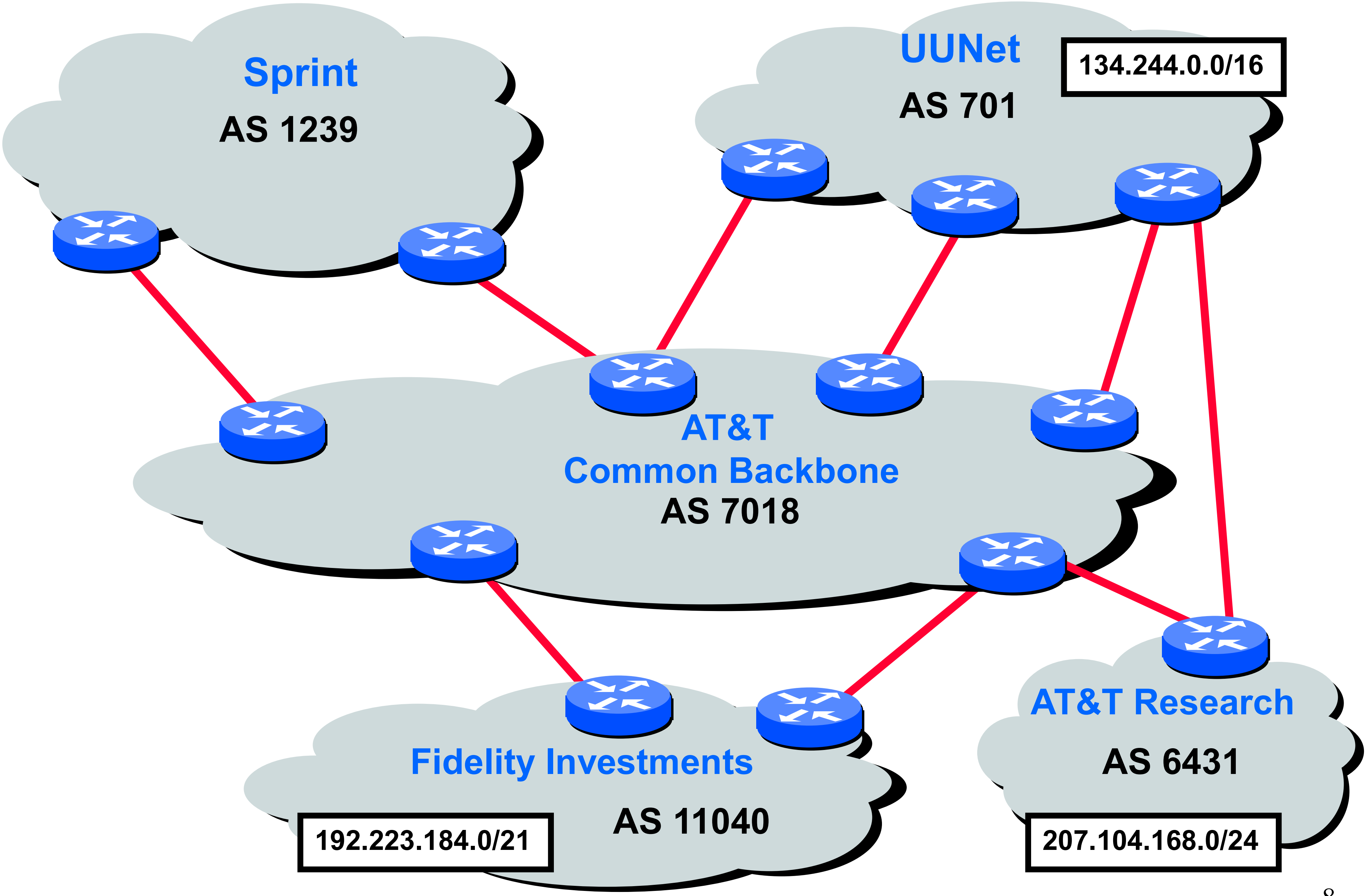
ASNs are 16 and 32 bit values.
64512 through 65535 are “private”

Currently over 11,000 in use.

- **Genuity: 1**
- **MIT: 3**
- **Harvard: 11**
- **UC San Diego: 7377**
- **AT&T: 7018, 6341, 5074, …**
- **UUNET: 701, 702, 284, 12199, …**
- **Sprint: 1239, 1240, 6211, 6242, …**
- **…**

ASNs represent units of routing policy

Interdomain routing = routing between autonomous systems



Internet approach to scalable routing

aggregate routers into regions known as "autonomous systems" (AS) (a.k.a. "domains")

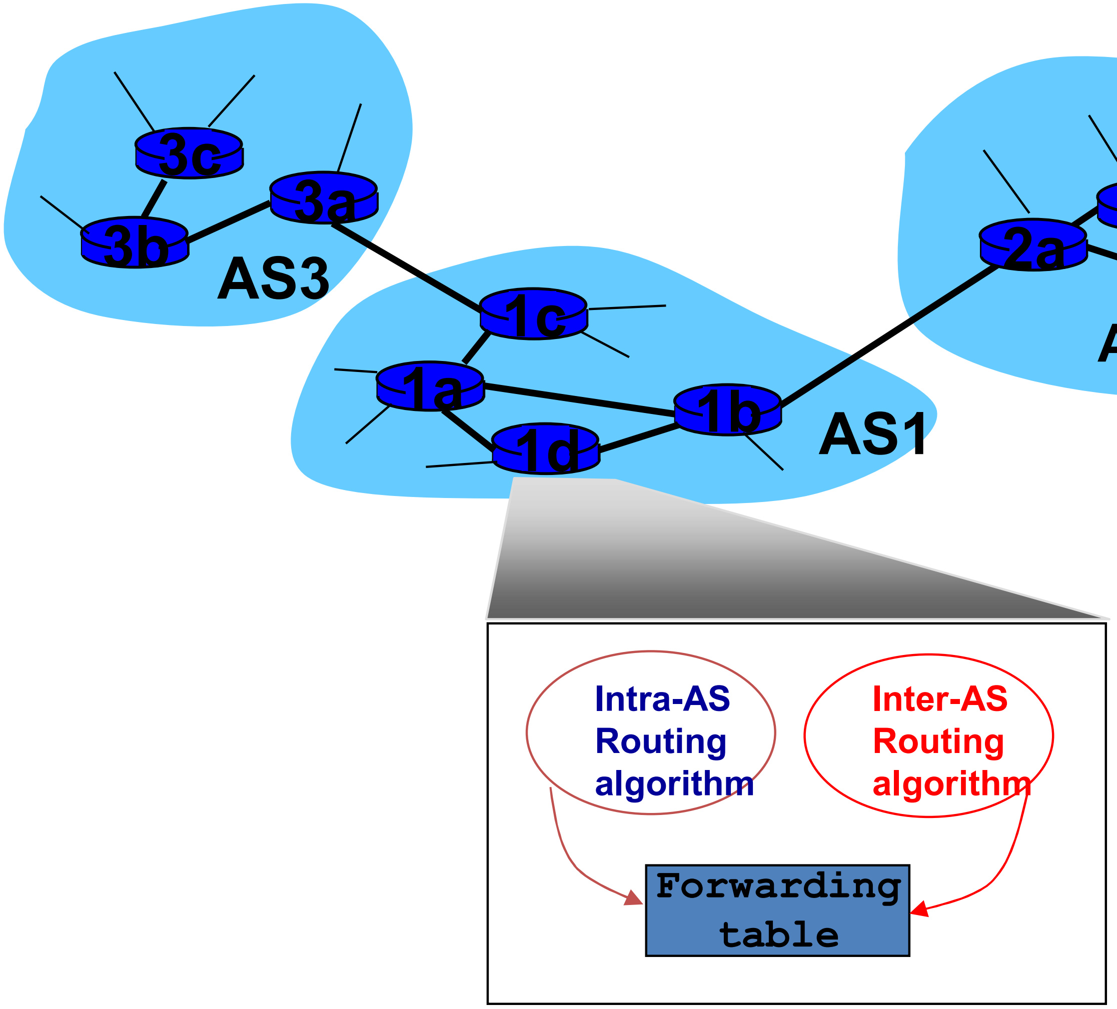
intra-AS routing

- routing among hosts, routers in same AS ("network")
- all routers in AS must run *same* intra-domain protocol
- routers in *different* AS can run *different* intra-domain routing protocol
- gateway router: at "edge" of its own AS, has link(s) to router(s) in other AS'es

inter-AS routing

- routing among AS'es
- gateways perform inter-domain routing (as well as intra-domain routing)

Interconnected ASes



- forwarding table configured by both intra- and inter-AS routing algorithm
  - intra-AS routing determine entries for destinations within AS
  - inter-AS & intra-AS determine entries for external destinations

Intra-AS Routing

Intra-AS Routing

- also known as *interior gateway protocols (IGP)*
- most common intra-AS routing protocols:
  - RIP: Routing Information Protocol
  - OSPF: Open Shortest Path First
  - IGRP: Interior Gateway Routing Protocol (Cisco proprietary for decades, until 2016)

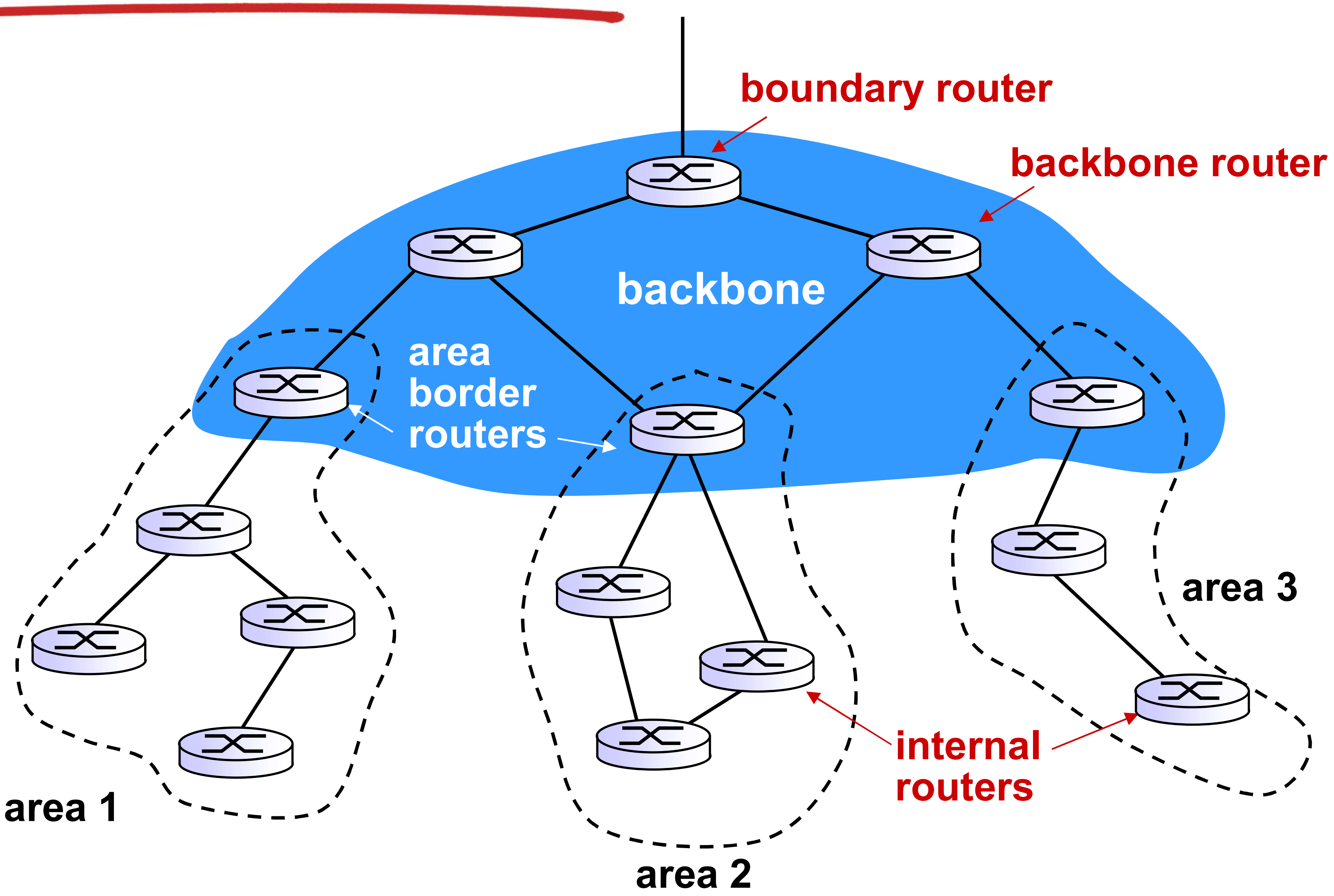
OSPF (Open Shortest Path First)

- “open”: publicly available
- uses link-state algorithm
  - – link state packet dissemination
  - – topology map at each node
  - – route computation using Dijkstra’s algorithm
- router floods OSPF link-state advertisements to all other routers in *entire* AS
  - – carried in OSPF messages directly over IP (rather than TCP or UDP
  - – link state: for each attached link

OSPF “advanced” features

- *security:* all OSPF messages authenticated (to prevent malicious intrusion)
- multiple same-cost paths allowed (only one path in RIP)
- for each link, multiple cost metrics for different ToS (e.g., satellite link cost set low for best effort ToS; high for real-time ToS)
- integrated uni- and multi-cast support:
  - Multicast OSPF (MOSPF) uses same topology data base as OSPF
- hierarchical OSPF in large domains.

Hierarchical OSPF



Hierarchical OSPF

- *two-level hierarchy:* local area, backbone.
  - link-state advertisements only in area
  - each nodes has detailed area topology; only know direction (shortest path) to nets in other areas.
- *area border routers:* “summarize” distances to nets in own area, advertise to other Area Border routers.
- *backbone routers:* run OSPF routing limited to backbone.
- *boundary routers:* connect to other AS’es.

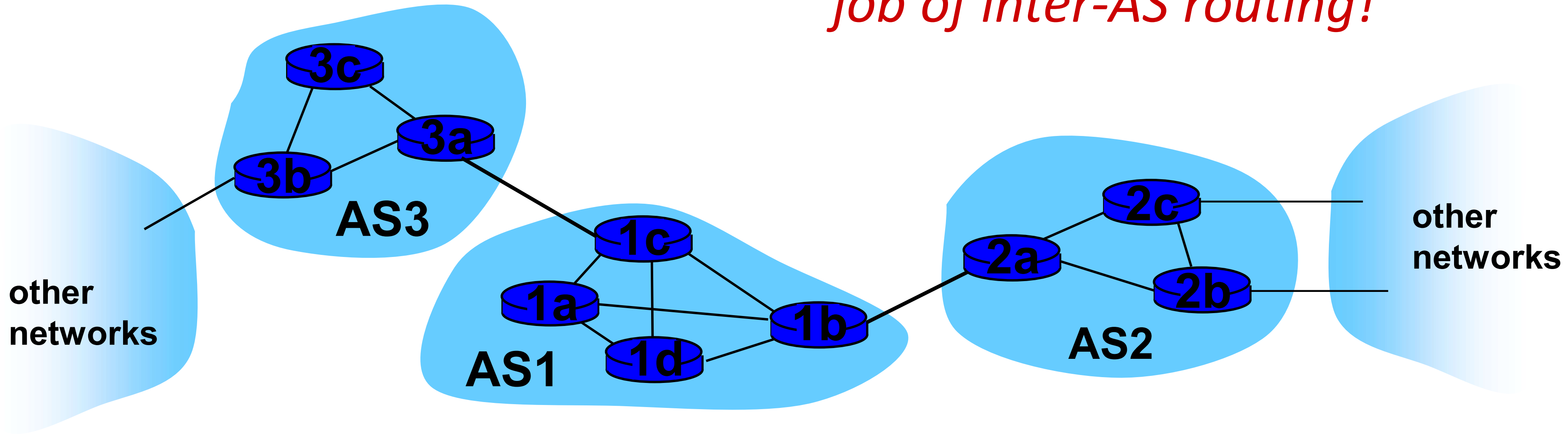
Inter-AS tasks

- suppose router in AS1 receives datagram destined outside of AS1:
  - router should forward packet to gateway router, but which one?

AS1 must:

1. learn which dests are reachable through AS2, which through AS3
2. propagate this reachability info to all routers in AS1

job of inter-AS routing!

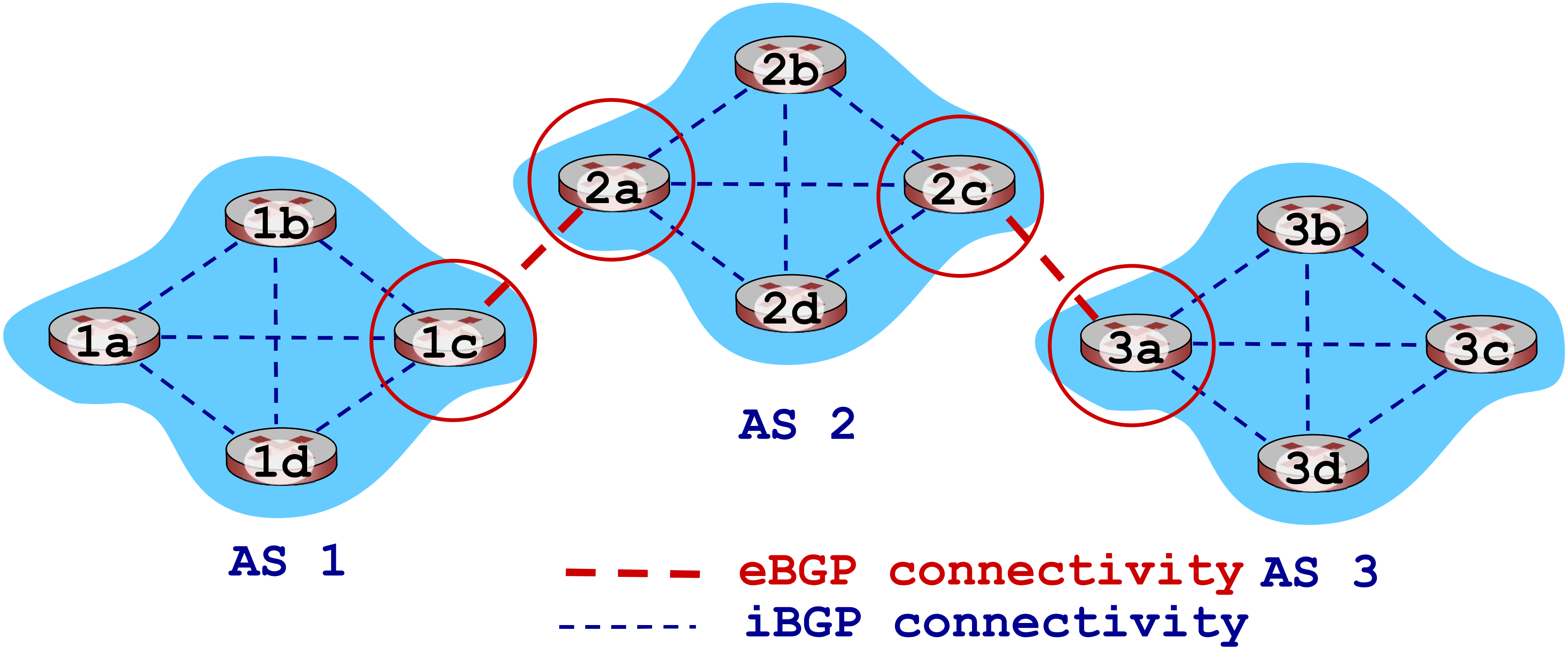


Inter-AS Routing

Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): *the* de facto inter-domain routing protocol
  - "glue that holds the Internet together"
- BGP provides each AS a means to:
  - eBGP: obtain subnet reachability information from neighboring ASes
  - iBGP: propagate reachability information to all AS-internal routers.
  - determine "good" routes to other networks based on reachability information and *policy*
- allows subnet to advertise its existence to rest of Internet: *"I am here"*

eBGP, iBGP connections

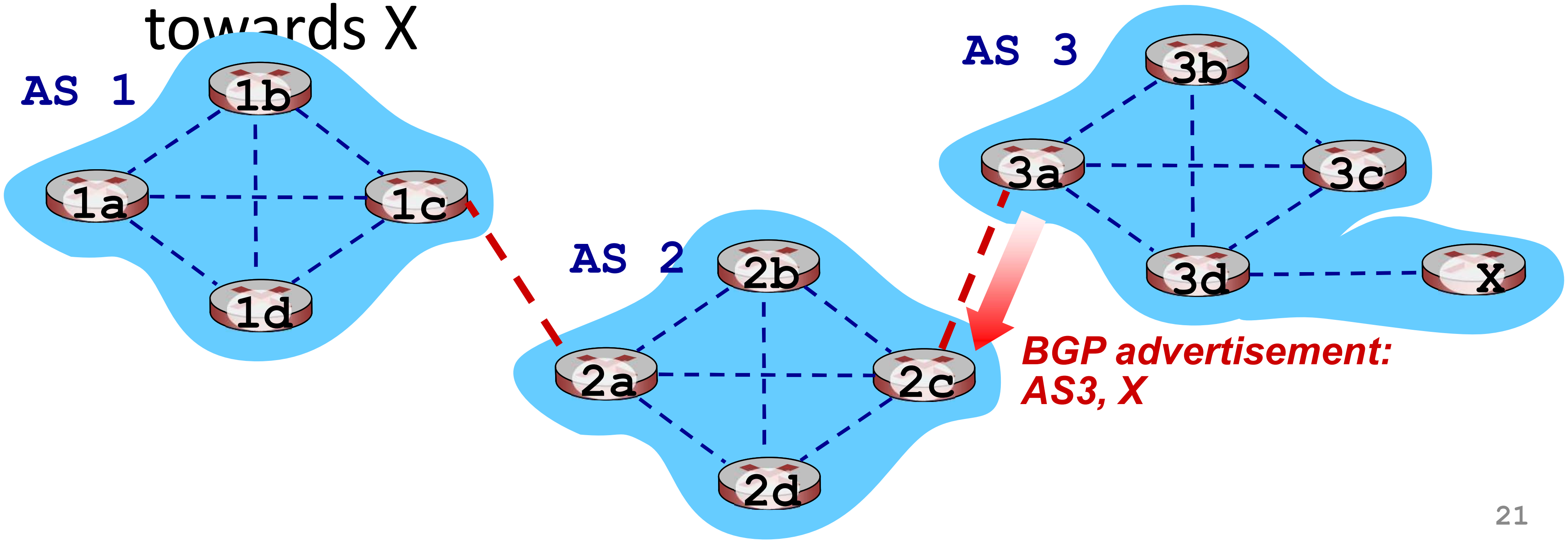


gateway routers run both eBGP and iBGP protoco

BGP basics

- **BGP session:** two BGP routers ("peers") exchange BGP messages over semi-permanent TCP connection:
  - advertising *paths* to different destination network prefixes (BGP is a "path vector" protocol)

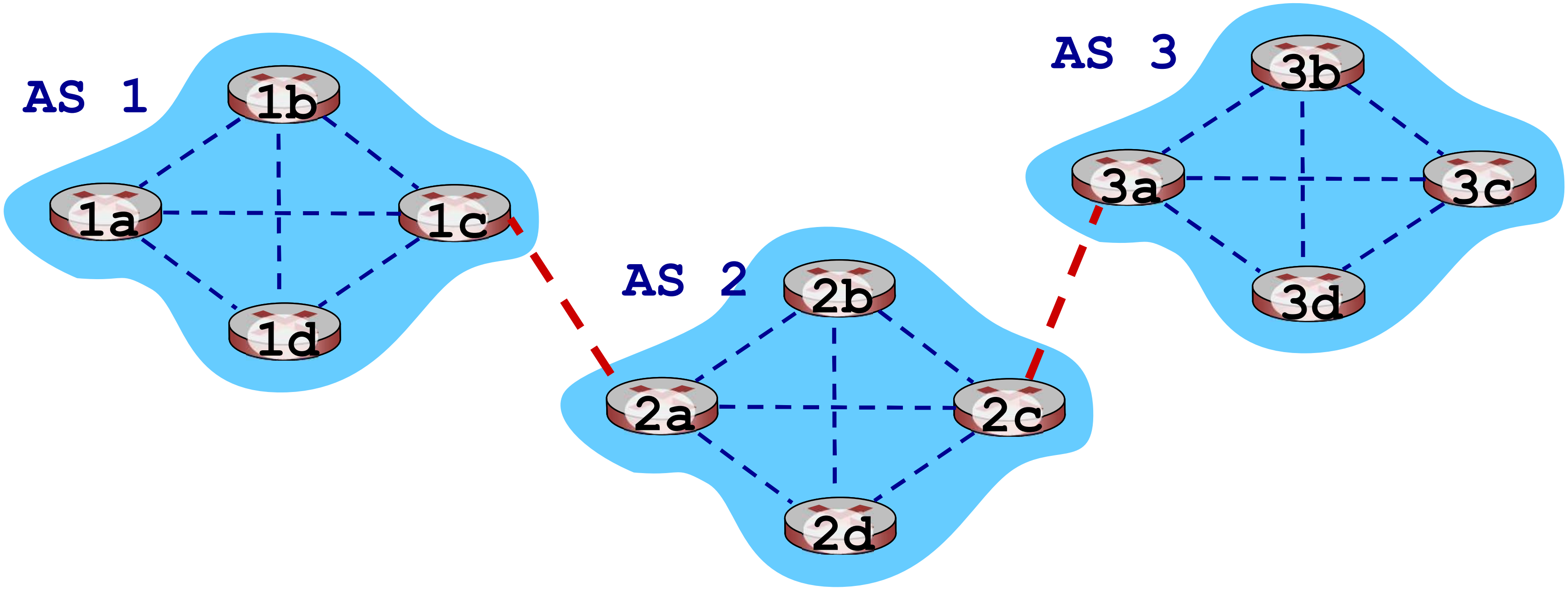
- when AS3 gateway router 3a advertises path AS3,X to AS2 gateway router 2c:
  - AS3 *promises* to AS2 it will forward datagrams towards X



BGP advertisement: AS3, X

BGP basics

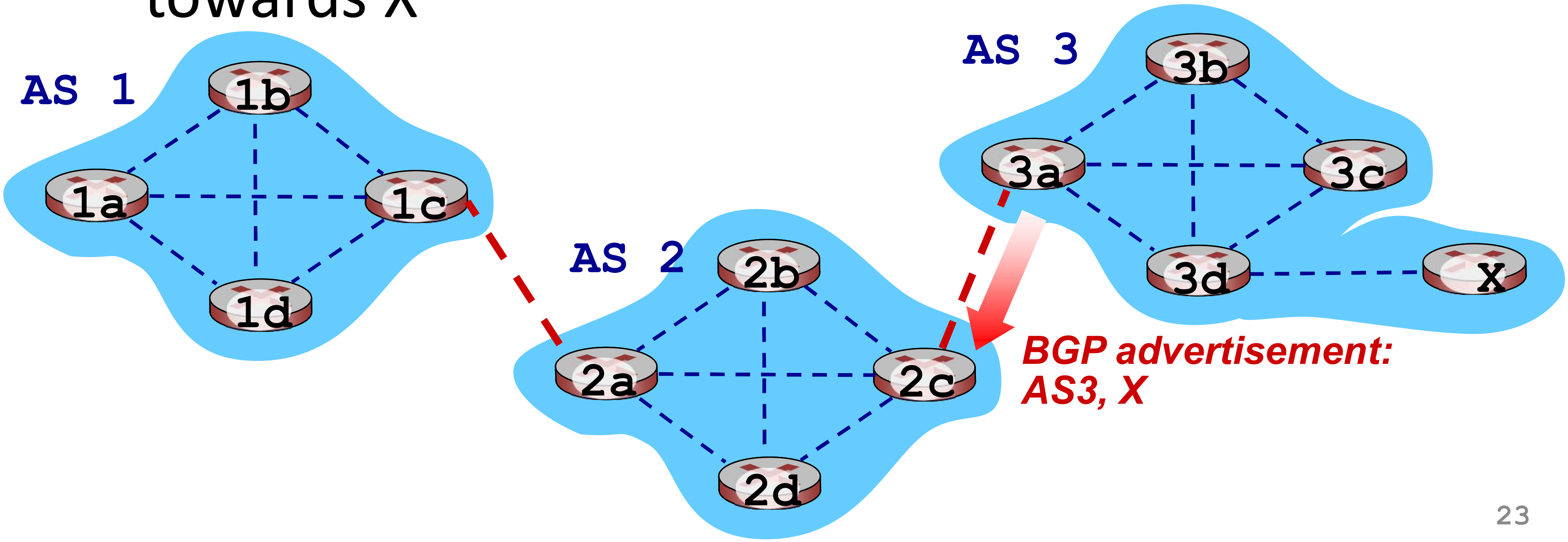
- **BGP session:** two BGP routers ("peers") exchange BGP messages over semi-permanent TCP connection:
  - advertising ***paths*** to different destination network prefixes (BGP is a "path vector" protocol)



BGP basics

- **BGP session:** two BGP routers ("peers") exchange BGP messages over semi-permanent TCP connection.

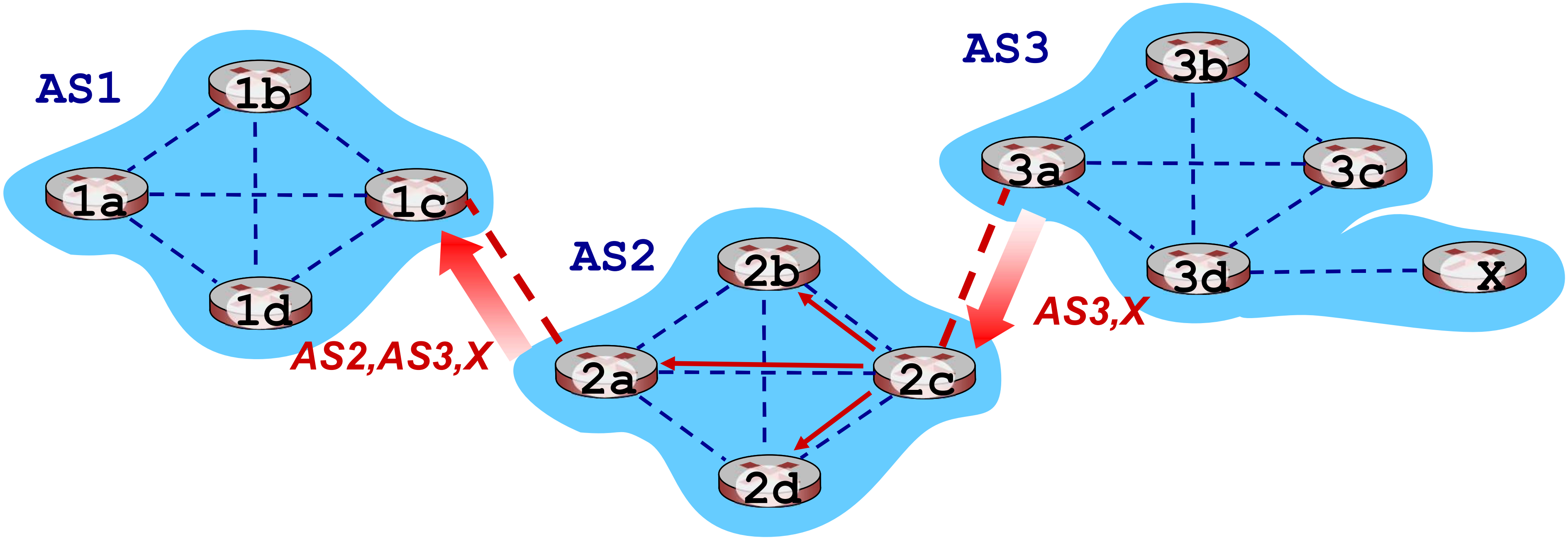
- when AS3 gateway router 3a advertises path AS3,X to AS2 gateway router 2c:
  - AS3 *promises* to AS2 it will forward datagrams towards X



Path attributes and BGP routes

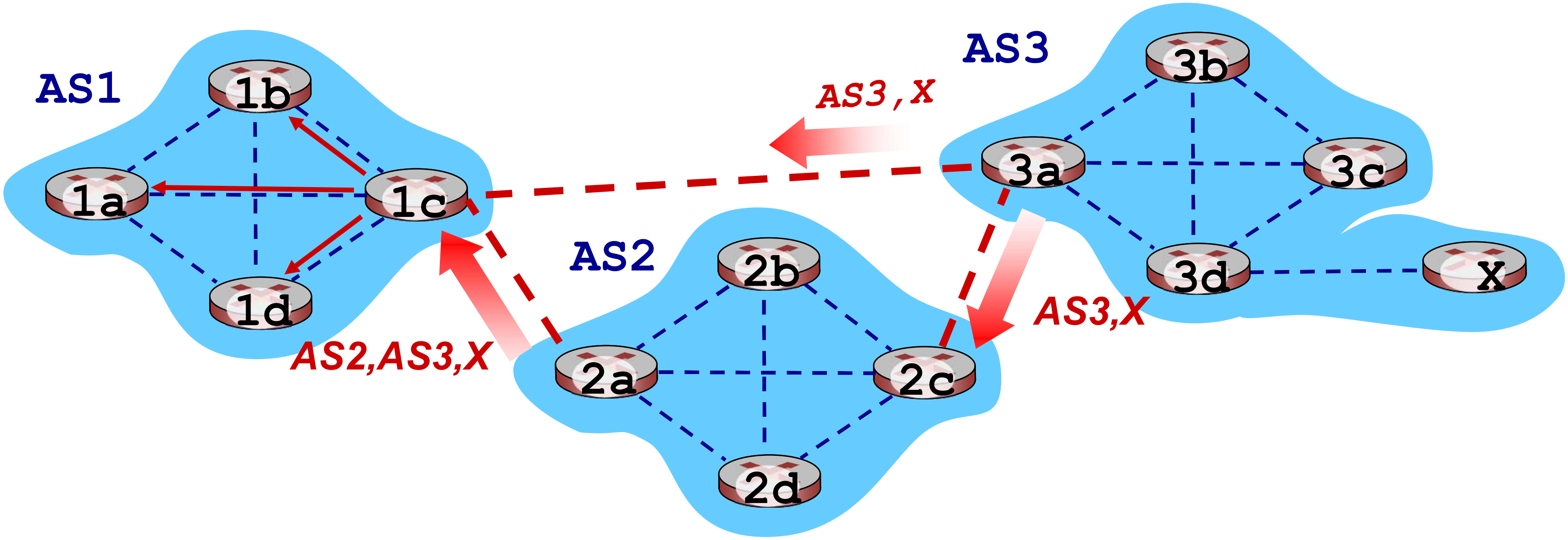
- advertised prefix includes BGP attributes
  - prefix + attributes = “route”
- two important attributes:
  - AS-PATH: list of ASes through which prefix advertisement has passed
  - NEXT-HOP: indicates specific internal-AS router to next-hop AS
- *Policy-based routing:*
  - gateway receiving route advertisement uses *import policy* to accept/decline path (e.g., never route through AS Y).
  - AS policy also determines whether to *advertise* path to other other neighboring ASes

BGP path advertisement



- **AS2 router 2c receives path advertisement AS3,X (via eBGP) from AS3 router 3a**
- Based on AS2 policy, AS2 router 2c accepts path AS3,X, propagates (via iBGP) to all AS2 routers
- **Based on AS2 policy, AS2 router 2a advertises (via eBGP) path AS2, AS3, X to AS1 router 1c**

BGP path advertisement



gateway router may learn about multiple paths to destination:

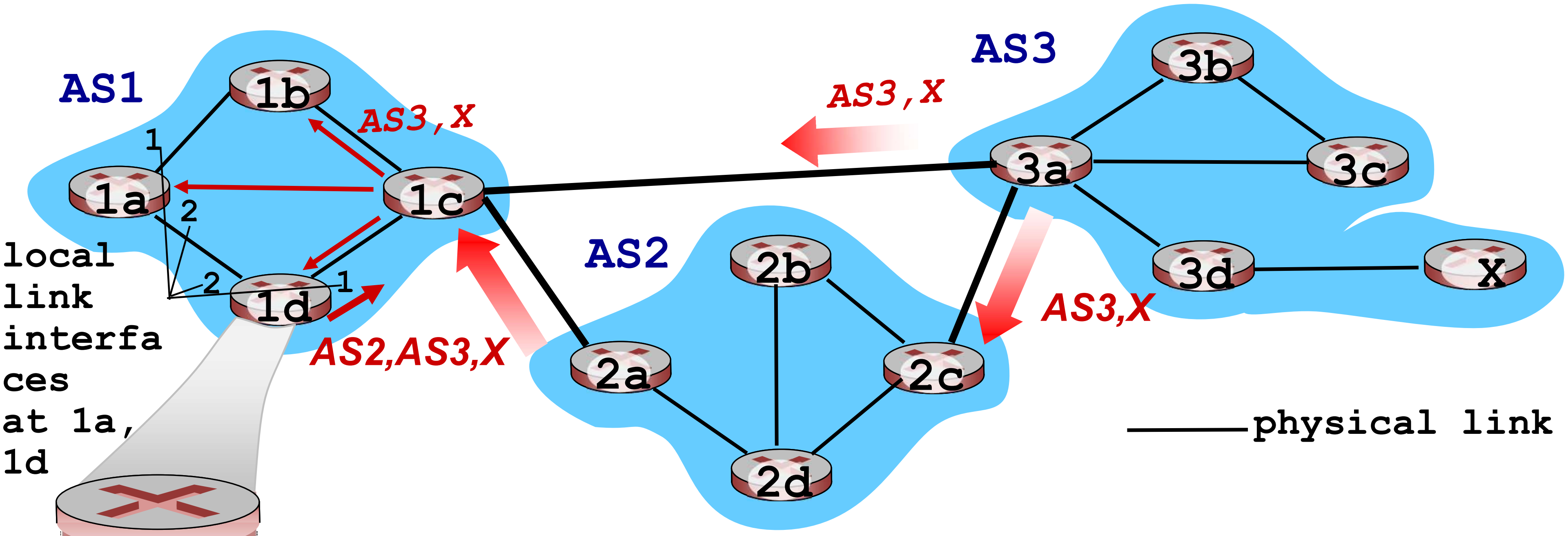
- AS1 gateway router 1c learns path *AS2,AS3,X* from 2a
- **AS1 gateway router 1c learns path *AS3,X* from 3a**
- **Based on policy, AS1 gateway router 1c chooses path *AS3,X, and advertises path within AS1 via iBGP***

BGP messages

- BGP messages exchanged between peers over TCP connection
- BGP messages:
  - OPEN: opens TCP connection to remote BGP peer and authenticates sending BGP peer
  - UPDATE: advertises new path (or withdraws old)
  - KEEPALIVE: keeps connection alive in absence of UPDATES; also ACKs OPEN request
  - NOTIFICATION: reports errors in previous msg; also used to close connection

BGP, OSPF, forwarding table entries

Q: how does router set forwarding table entry to distant



local link interfaces at 1a, 1d

physical link

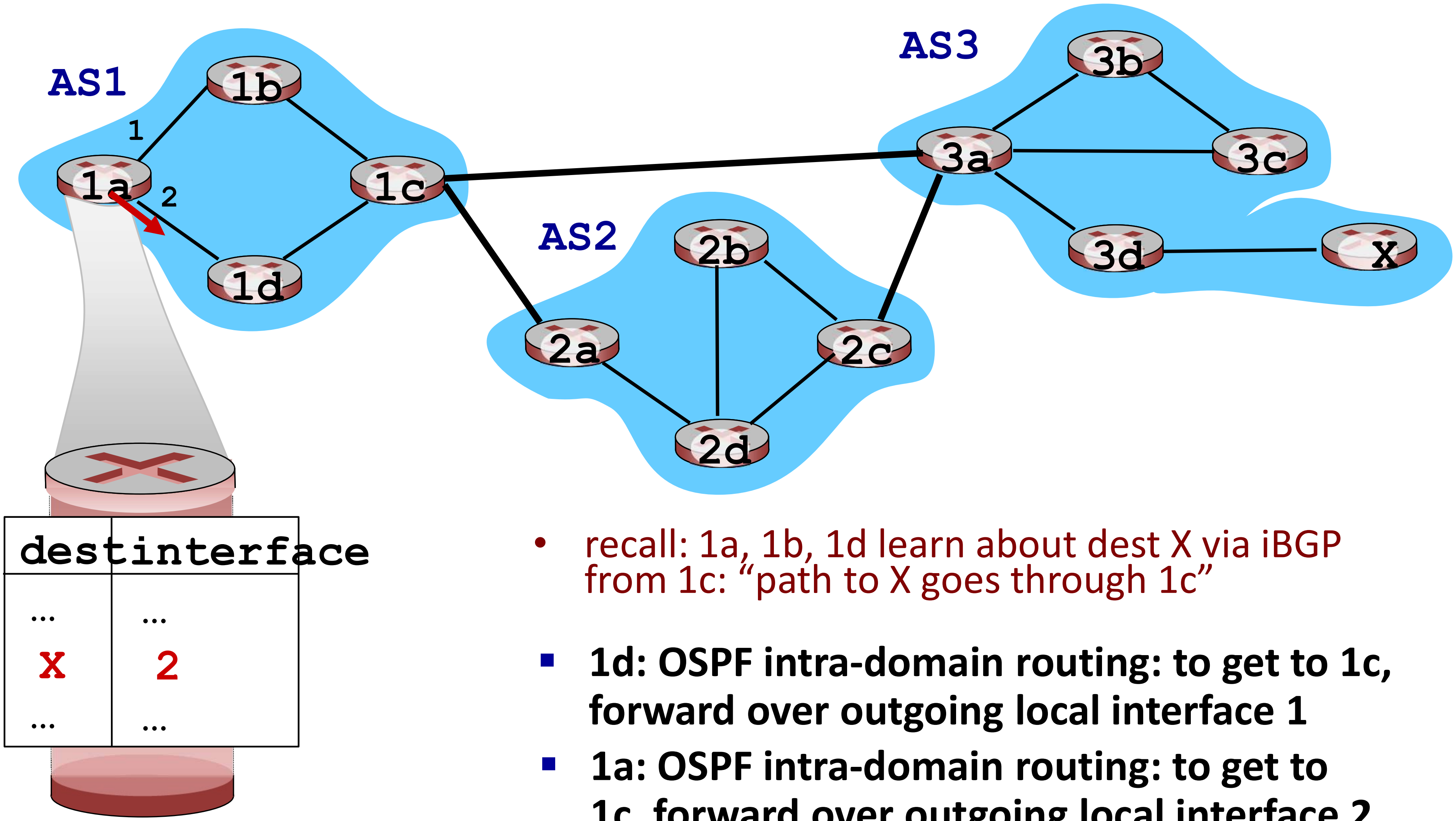
dest	interface
…	…
X	1
…	…

- recall: 1a, 1b, 1d learn about dest X via iBGP from 1c: "path to X goes through 1c"

- **1d: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 1**

BGP, OSPF, forwarding table entries

Q: how does router set forwarding table entry to distant



dest	interface
…	…
X	2
…	…

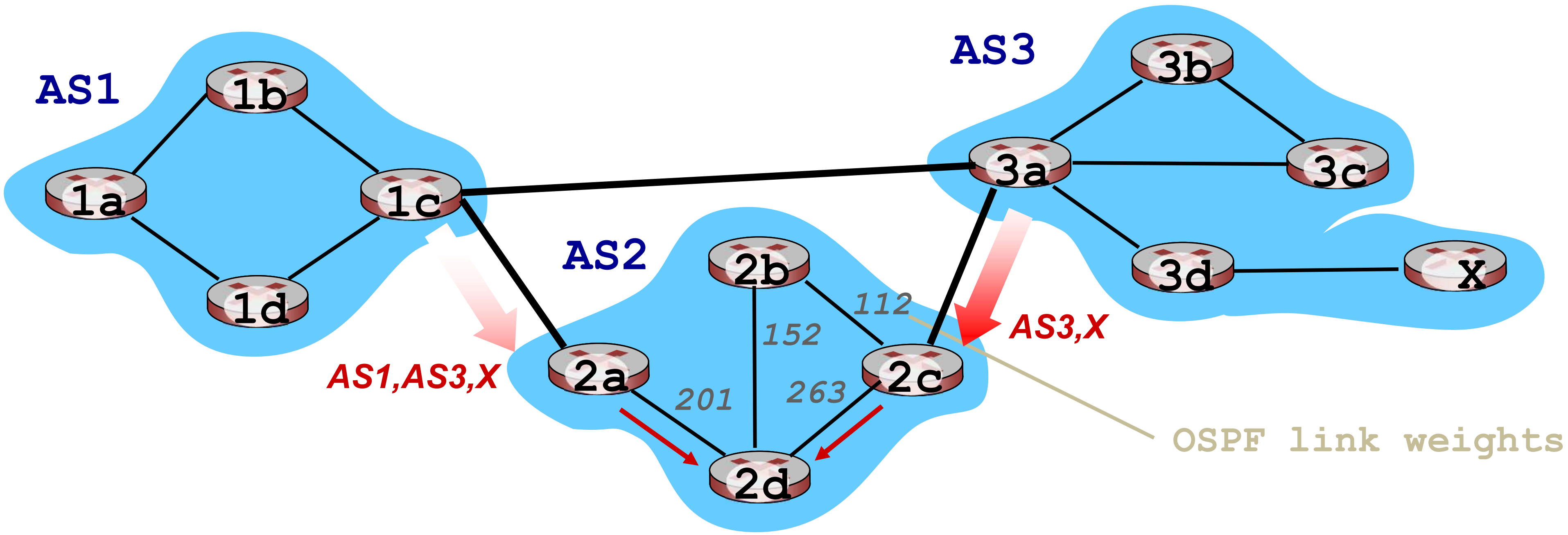
- recall: 1a, 1b, 1d learn about dest X via iBGP from 1c: "path to X goes through 1c"

- **1d: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 1**
- **1a: OSPF intra-domain routing: to get to 1c, forward over outgoing local interface 2**

BGP route selection

- router may learn about more than one route to destination AS, selects route based on:
    1. local preference value attribute: policy decision
    2. shortest AS-PATH
    3. closest NEXT-HOP router: hot potato routing
    4. additional criteria

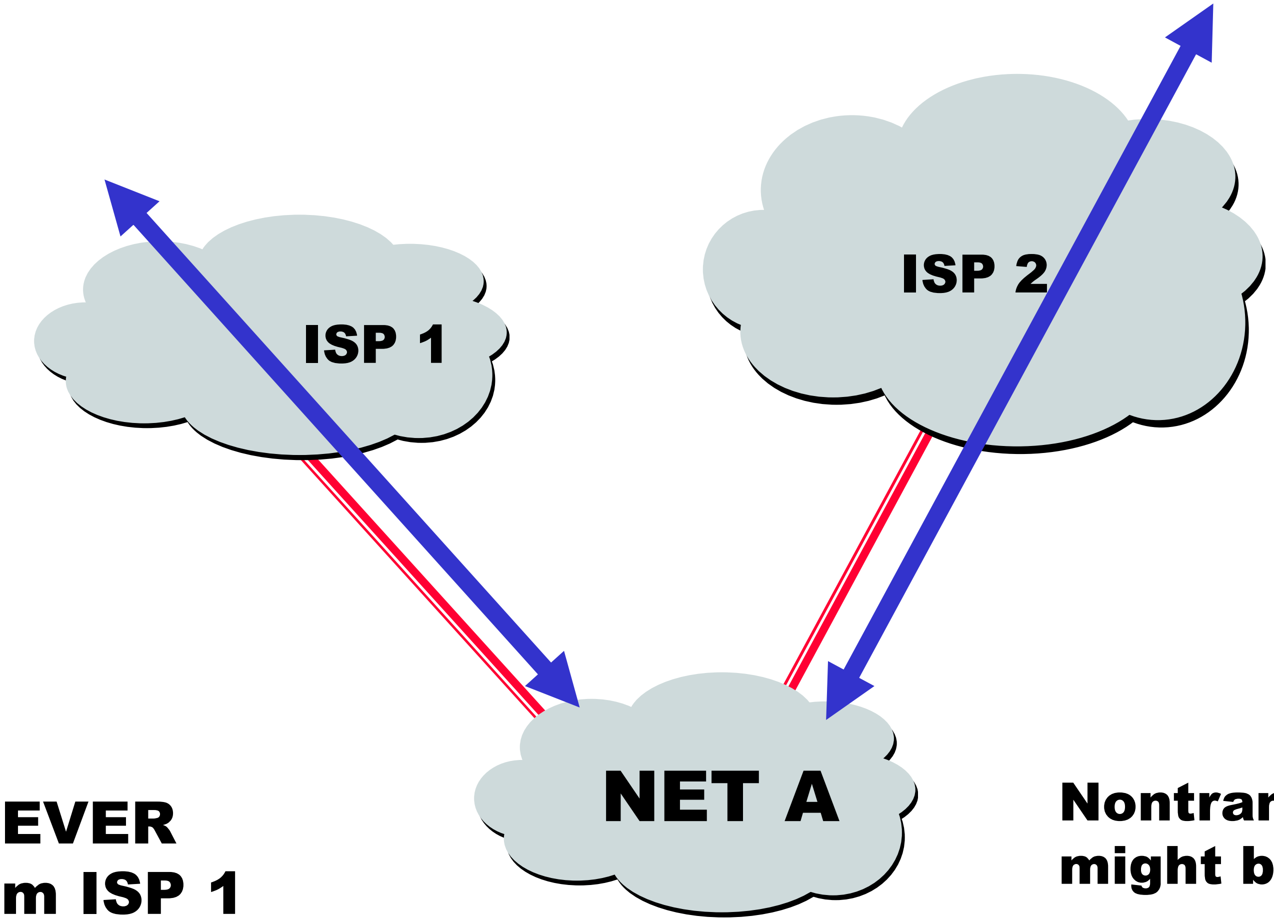
Hot Potato Routing



- 2d learns (via iBGP) it can route to X via 2a or 2c
- *hot potato routing:* choose local gateway that has least intra-domain cost (e.g., 2d chooses 2a, even though more AS hops to *X*): don't worry about inter-domain cost!

A dive into the BGP policies

Nontransit vs. Transit ASes



Internet Service providers (often) have transit networks

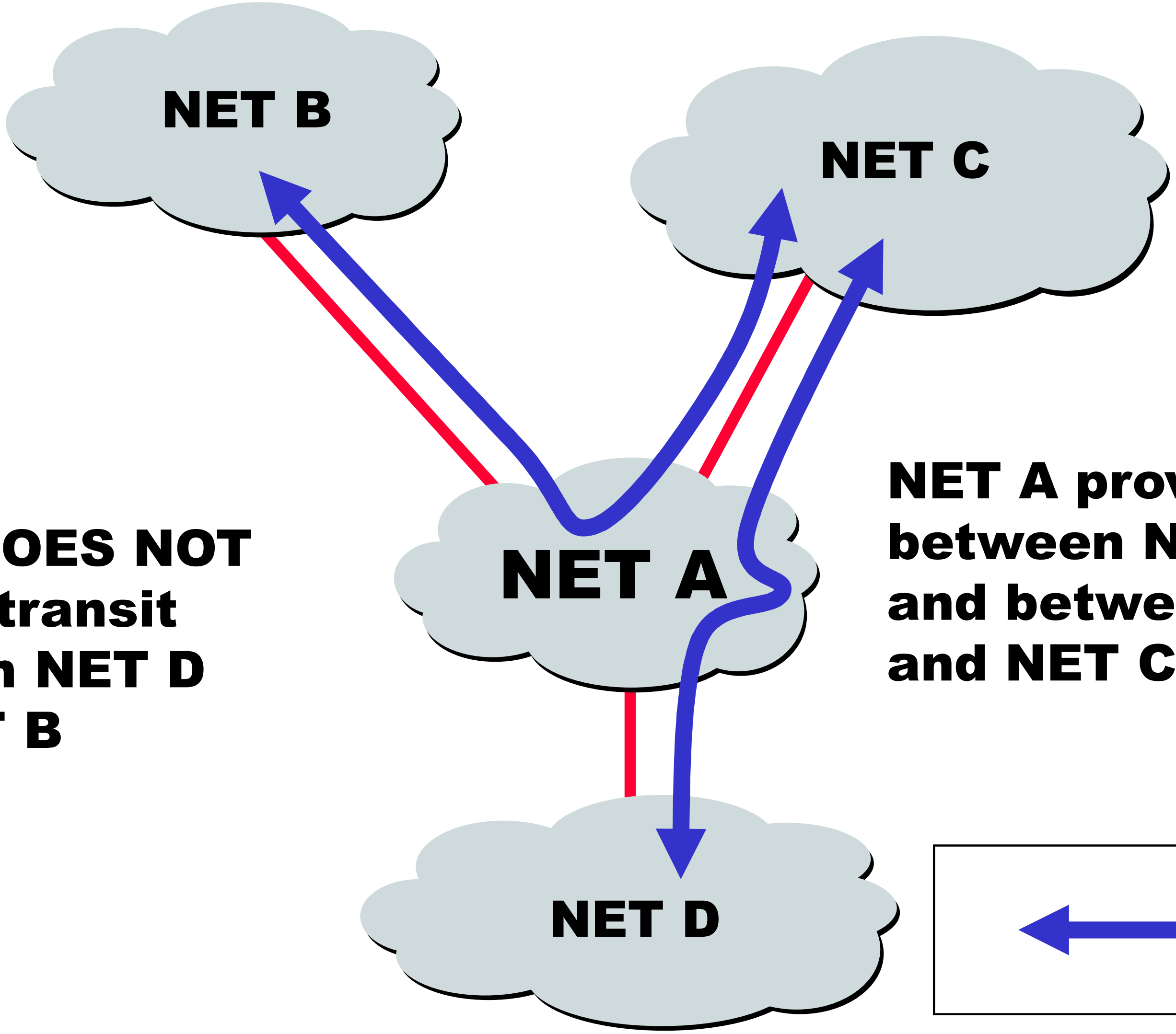
Traffic NEVER flows from ISP 1 through NET A to ISP 2 (At least not intentionally!)

Nontransit AS might be a corporate or campus network. Could be a “content provider”

↔ IP traffic

Selective Transit

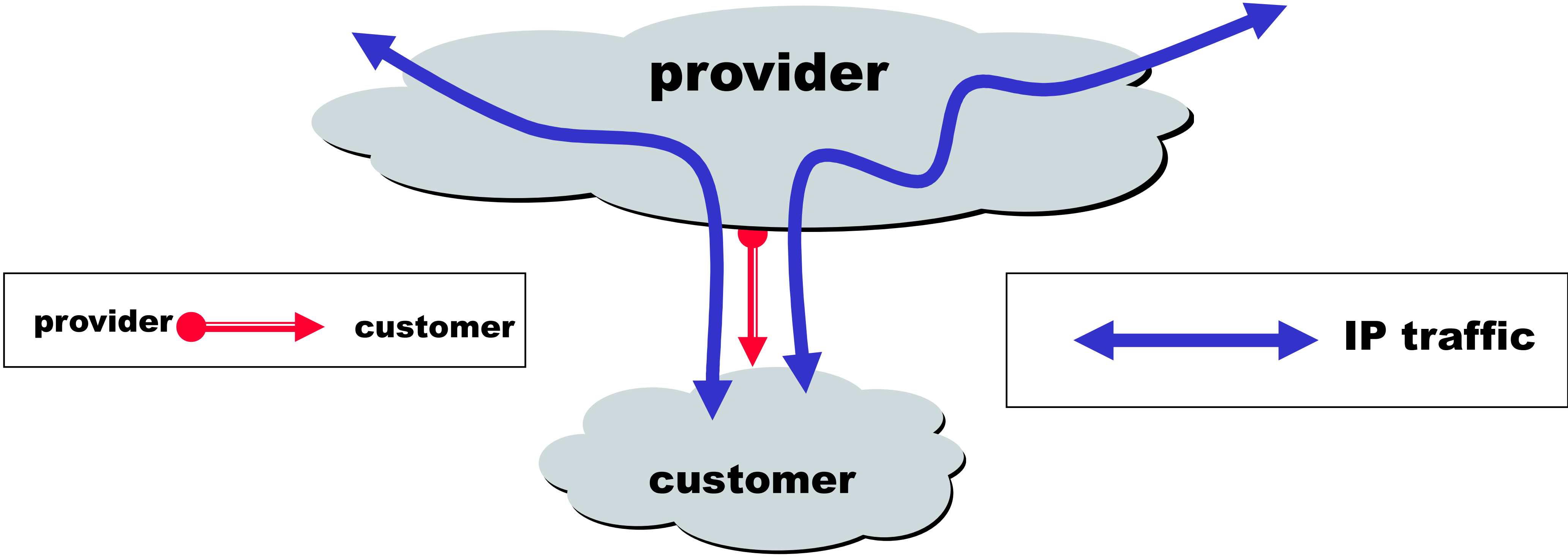
NET A DOES NOT provide transit Between NET D and NET B



NET A provides transit between NET B and NET C and between NET D and NET C

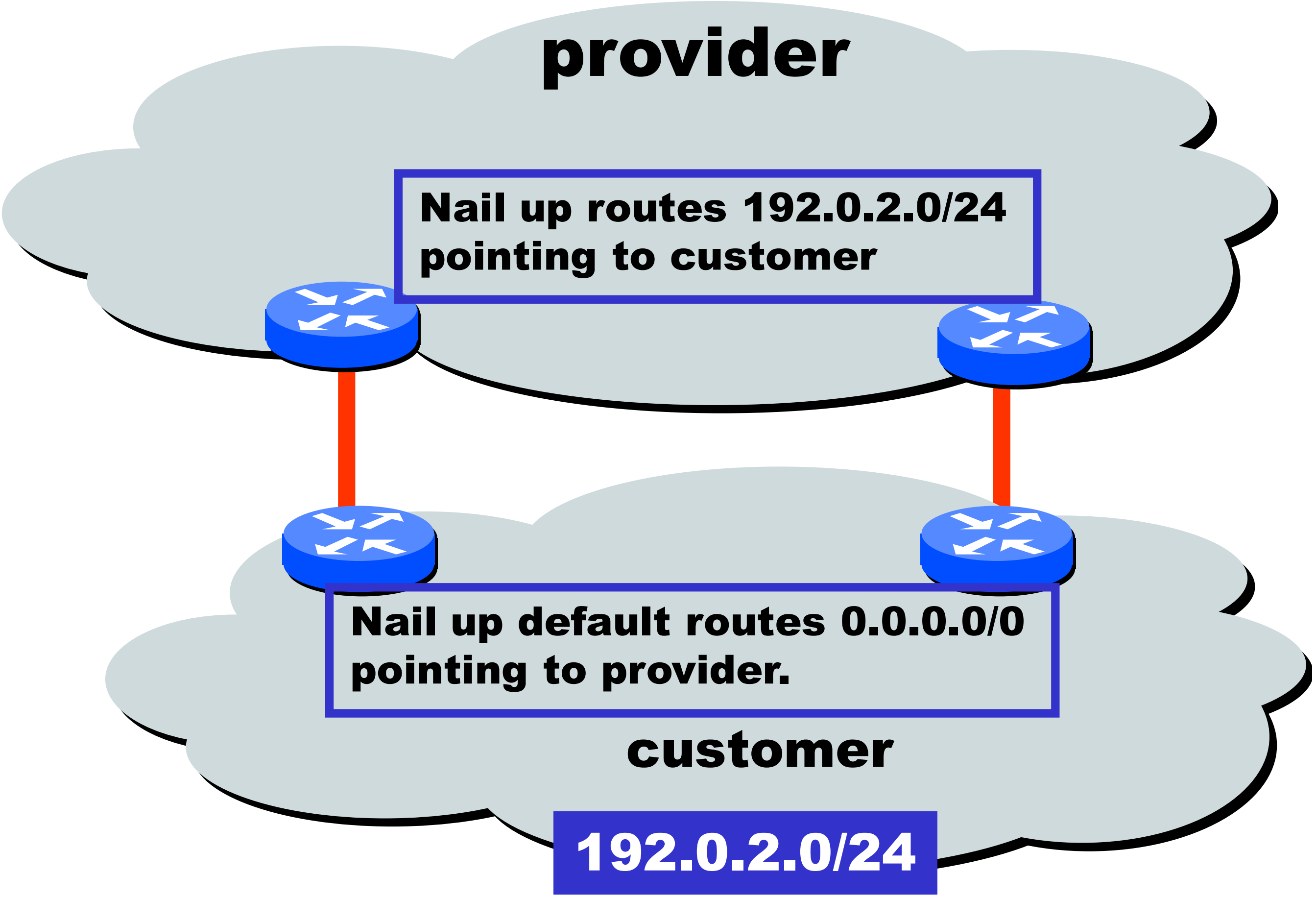
Most transit networks transit in a selective manner…

Customers and Providers



Customer pays provider for access to the Internet

Customers Don’t Always Need BGP

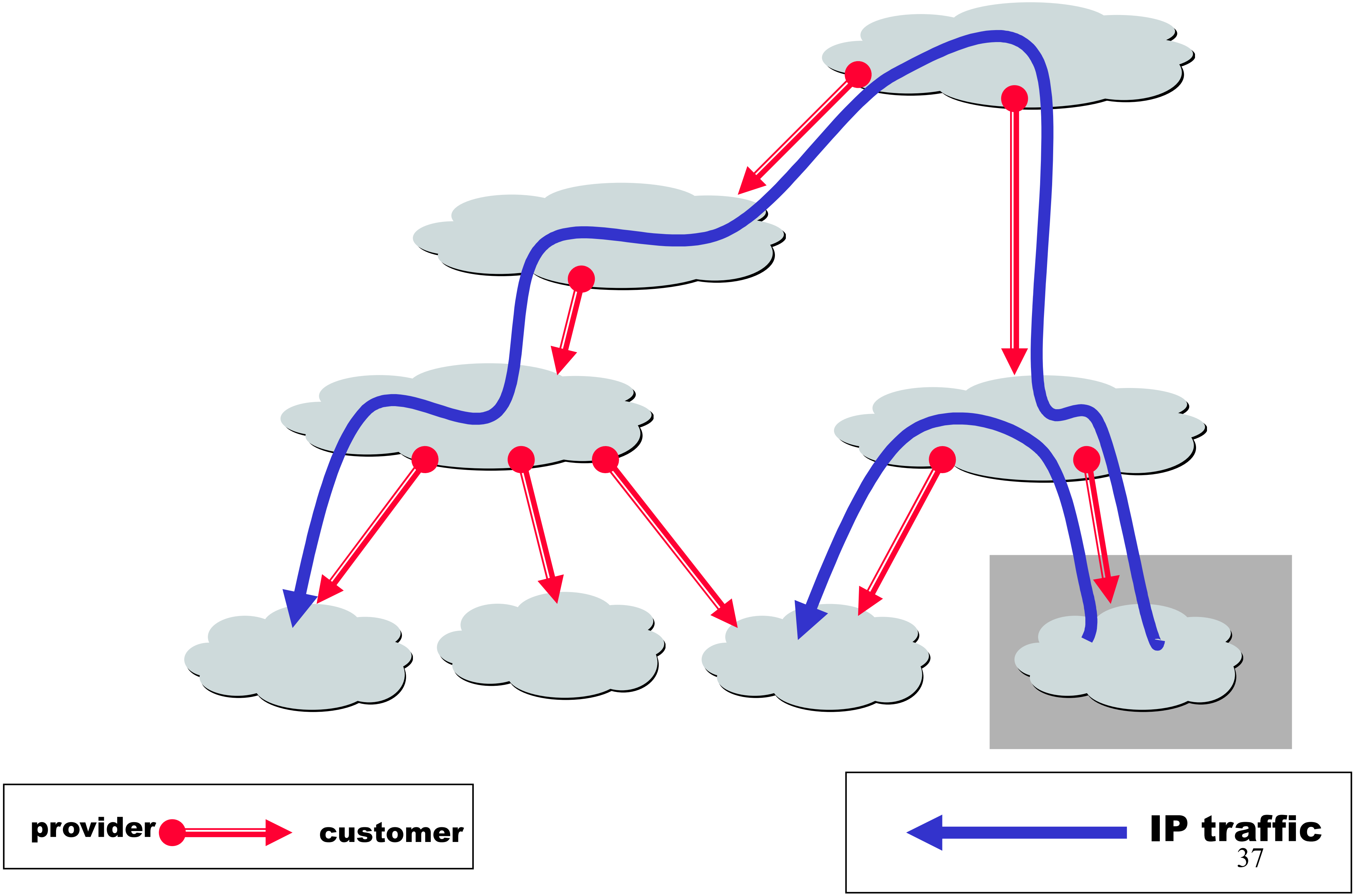


Nail up routes 192.0.2.0/24 pointing to customer

Nail up default routes 0.0.0.0/0 pointing to provider.

Static routing is the most common way of connecting an autonomous routing domain to the Internet.
This helps explain why BGP is a mystery to many …

Customer-Provider Hierarchy

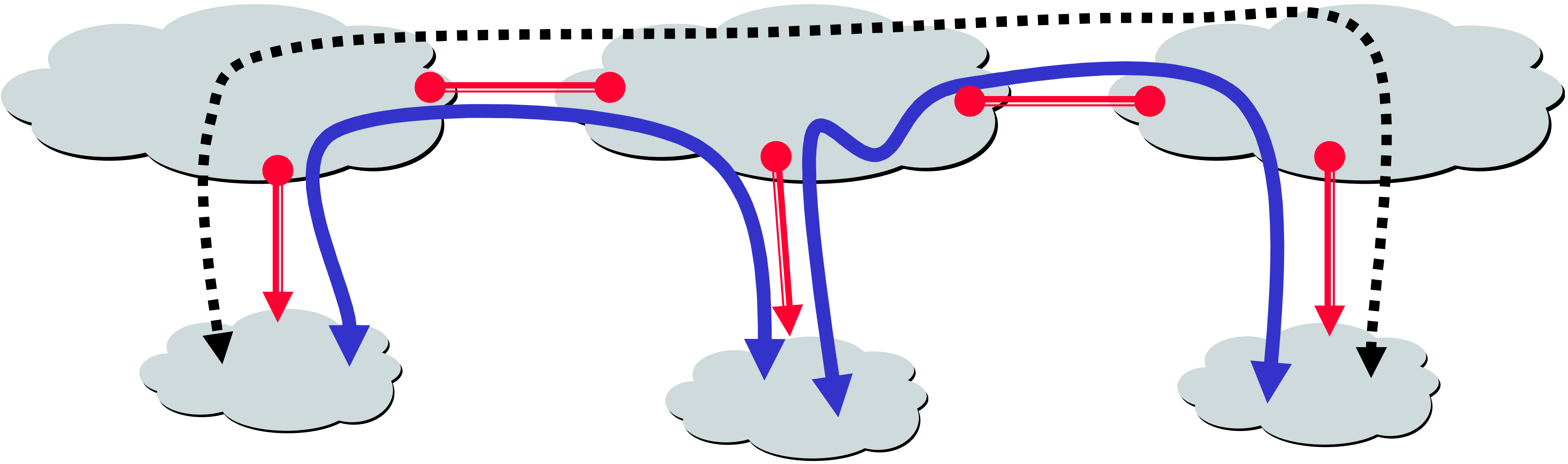


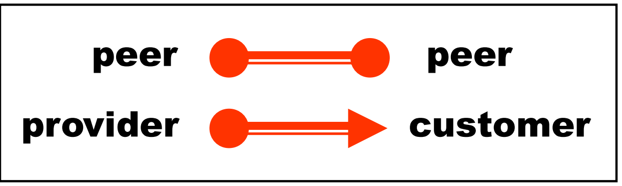
provider

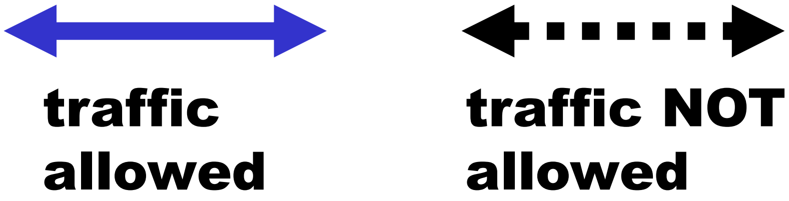
customer

IP traffic

The Peering Relationship





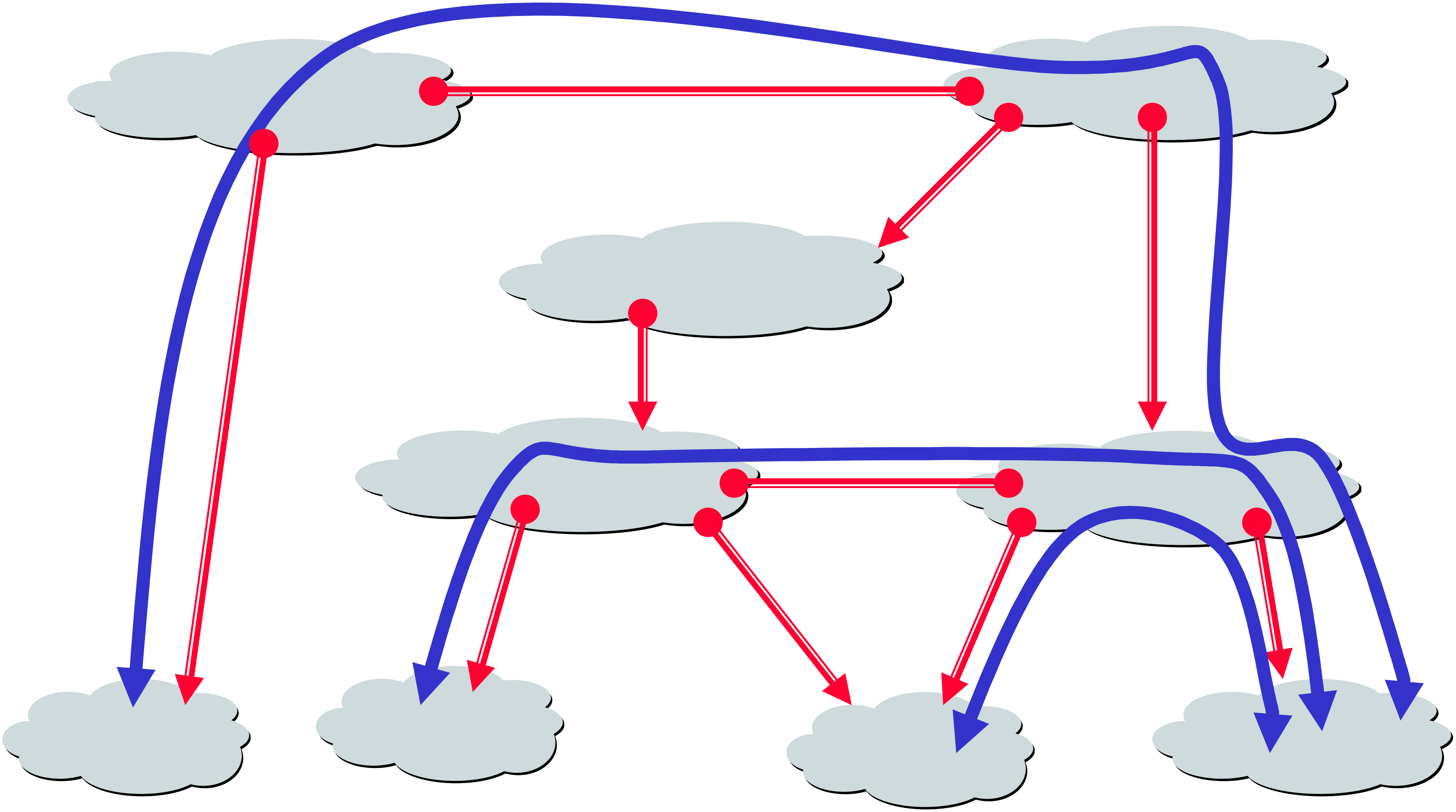


Peers provide transit between their respective customers

Peers do not provide transit between peers

Peers (often) do not exchange $$$

Peering Provides Shortcuts



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

peer ●—● peer

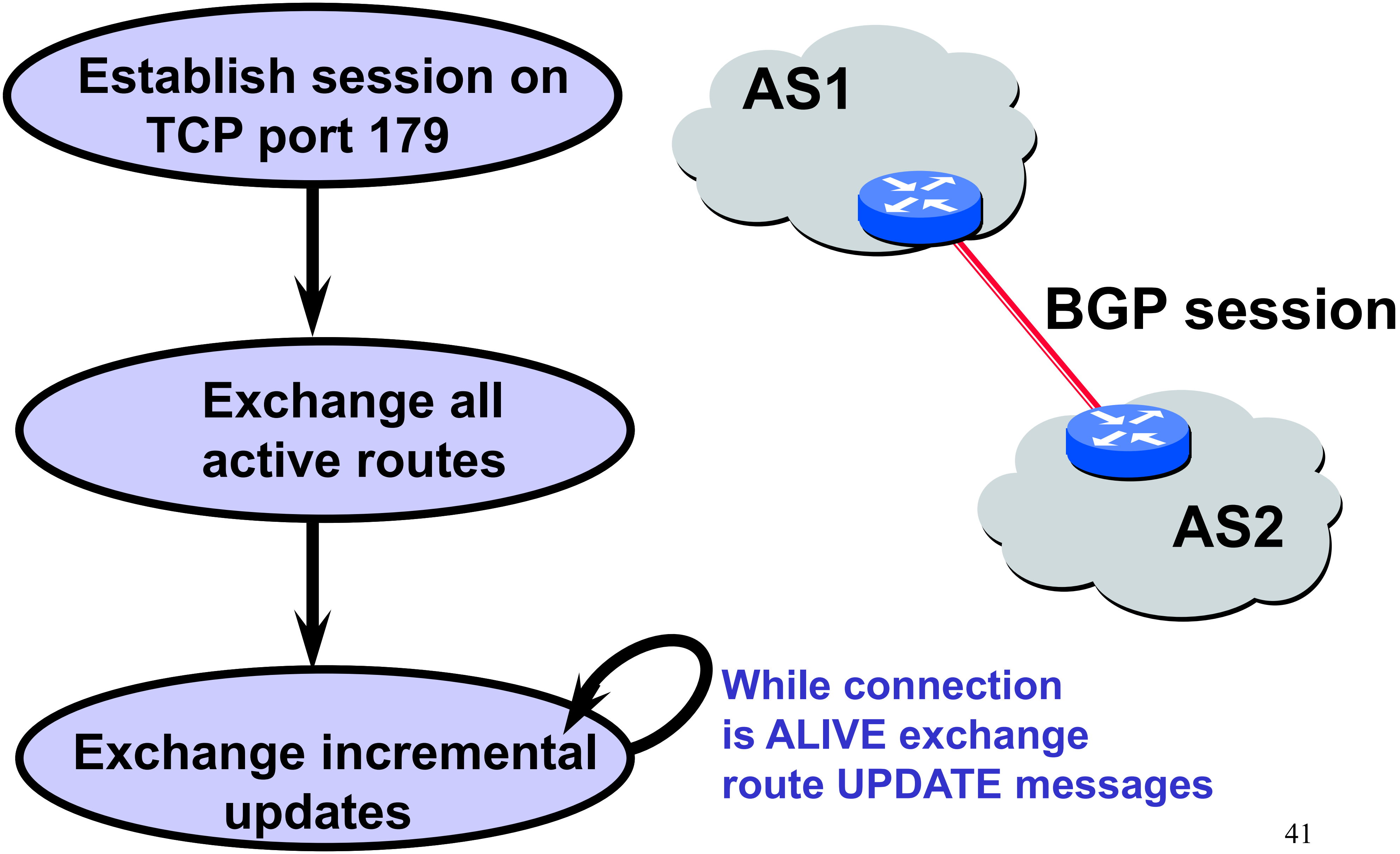
provider ●⟹ customer

BGP-4

- **BGP** = **B**order **G**ateway **P**rotocol
- Is a **Policy-Based** routing protocol
- Is the **de facto EGP** of today's global Internet
- Relatively simple protocol, but configuration is complex and the entire world can see, and be impacted by, your mistakes.

- **1989 : BGP-1 [RFC 1105]**
  - **Replacement for EGP (1984, RFC 904)**
- **1990 : BGP-2 [RFC 1163]**
- **1991 : BGP-3 [RFC 1267]**
- **1995 : BGP-4 [RFC 1771]**
  - **Support for Classless Interdomain Routing (CIDR)**

BGP Operations (Simplified)



Establish session on TCP port 179

Exchange all active routes

Exchange incremental updates

While connection is ALIVE exchange route UPDATE messages

Four Types of BGP Messages

- **Open : Establish a peering session.**
- **Keep Alive : Handshake at regular intervals.**
- **Notification : Shuts down a peering session.**
- **Update : Announcing new routes or withdrawing previously announced routes.**

announcement

=

prefix + attributes values

BGP Attributes

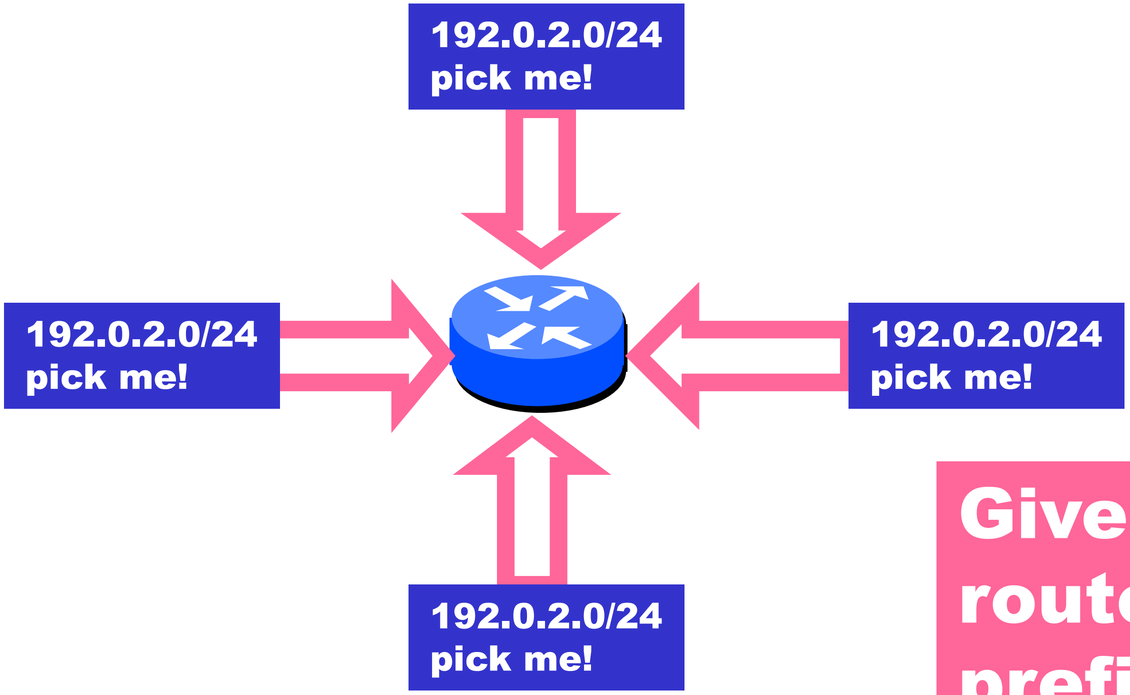
```
Value      Code                                 Reference
-----      ---------------------------------    ---------
   1       ORIGIN                               [RFC1771]
   2       AS_PATH                              [RFC1771]
   3       NEXT_HOP                             [RFC1771]
   4       MULTI_EXIT_DISC                      [RFC1771]
   5       LOCAL_PREF                           [RFC1771]
   6       ATOMIC_AGGREGATE                     [RFC1771]
   7       AGGREGATOR                           [RFC1771]
   8       COMMUNITY                            [RFC1997]
   9       ORIGINATOR_ID                        [RFC2796]
  10       CLUSTER_LIST                         [RFC2796]
  11       DPA                                     [Chen]
  12       ADVERTISER                           [RFC1863]
  13       RCID_PATH / CLUSTER_ID               [RFC1863]
  14       MP_REACH_NLRI                        [RFC2283]
  15       MP_UNREACH_NLRI                      [RFC2283]
  16       EXTENDED COMMUNITIES                   [Rosen]
 ...
 255       reserved for development
```

From IANA: http://www.iana.org/assignments/bgp-parameters

Most important attributes

Not all attributes need to be present in every announcement

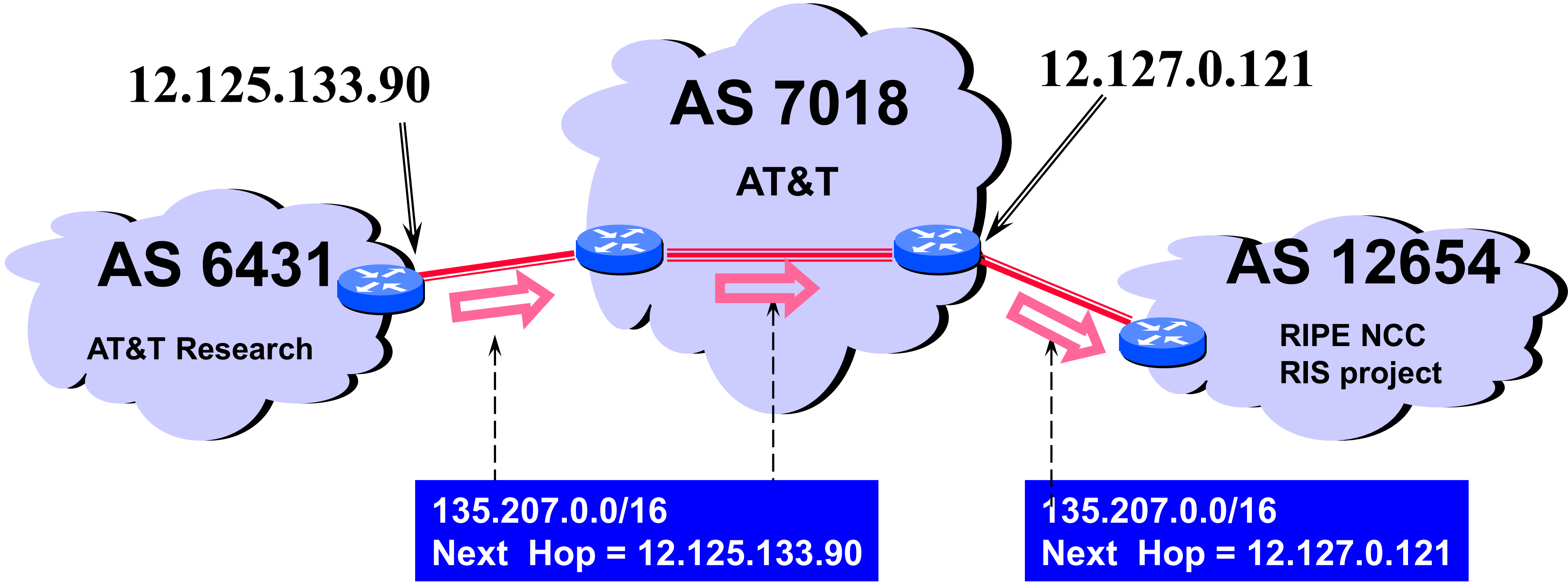
Attributes are Used to Select Best Routes



Given multiple routes to the same prefix, a BGP speaker must pick at most <u>one</u> best route

(Note: it could reject them all!)

BGP Next Hop Attribute



12.125.133.90

AS 6431

AT&T Research

AS 7018

AT&T

12.127.0.121

AS 12654

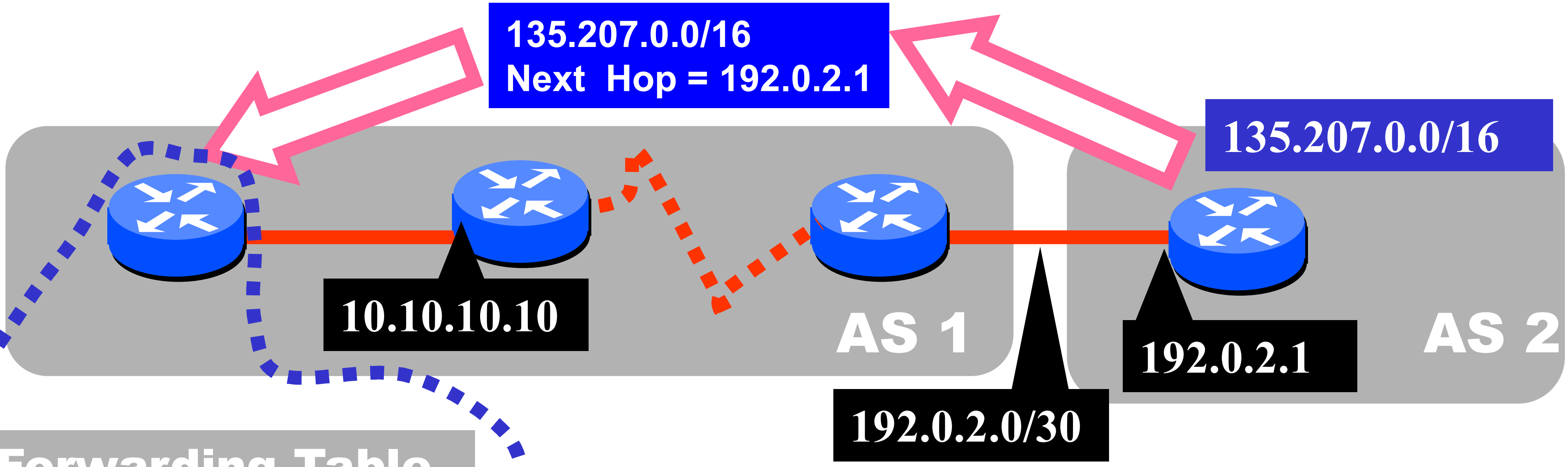
RIPE NCC
RIS project

135.207.0.0/16
Next Hop = 12.125.133.90

135.207.0.0/16
Next Hop = 12.127.0.121

Every time a route announcement crosses an AS boundary, the Next Hop attribute is changed to the IP address of the border router that announced the route.

Join EGP with IGP For Connectivity



Forwarding Table

destination	next hop
192.0.2.0/30	10.10.10.10

+

EGP

destination	next hop
135.207.0.0/16	192.0.2.1

Forwarding Table

destination	next hop
135.207.0.0/16	10.10.10.10
192.0.2.0/30	10.10.10.10

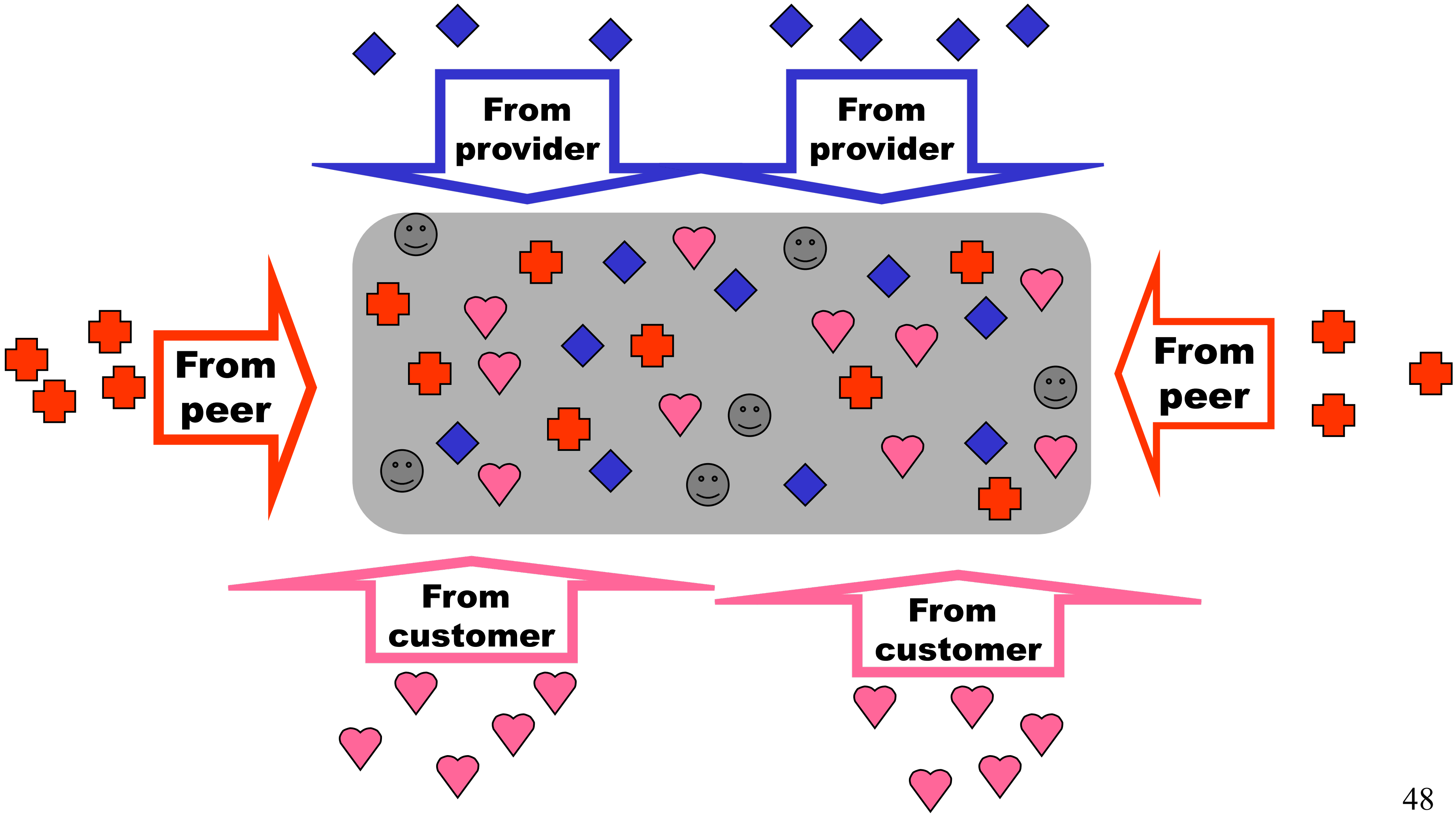
Implementing Customer/Provider and Peer/Peer relationships

Two parts:

- Enforce transit relationships
  - Outbound route filtering
- Enforce order of route preference
  - provider < peer < customer

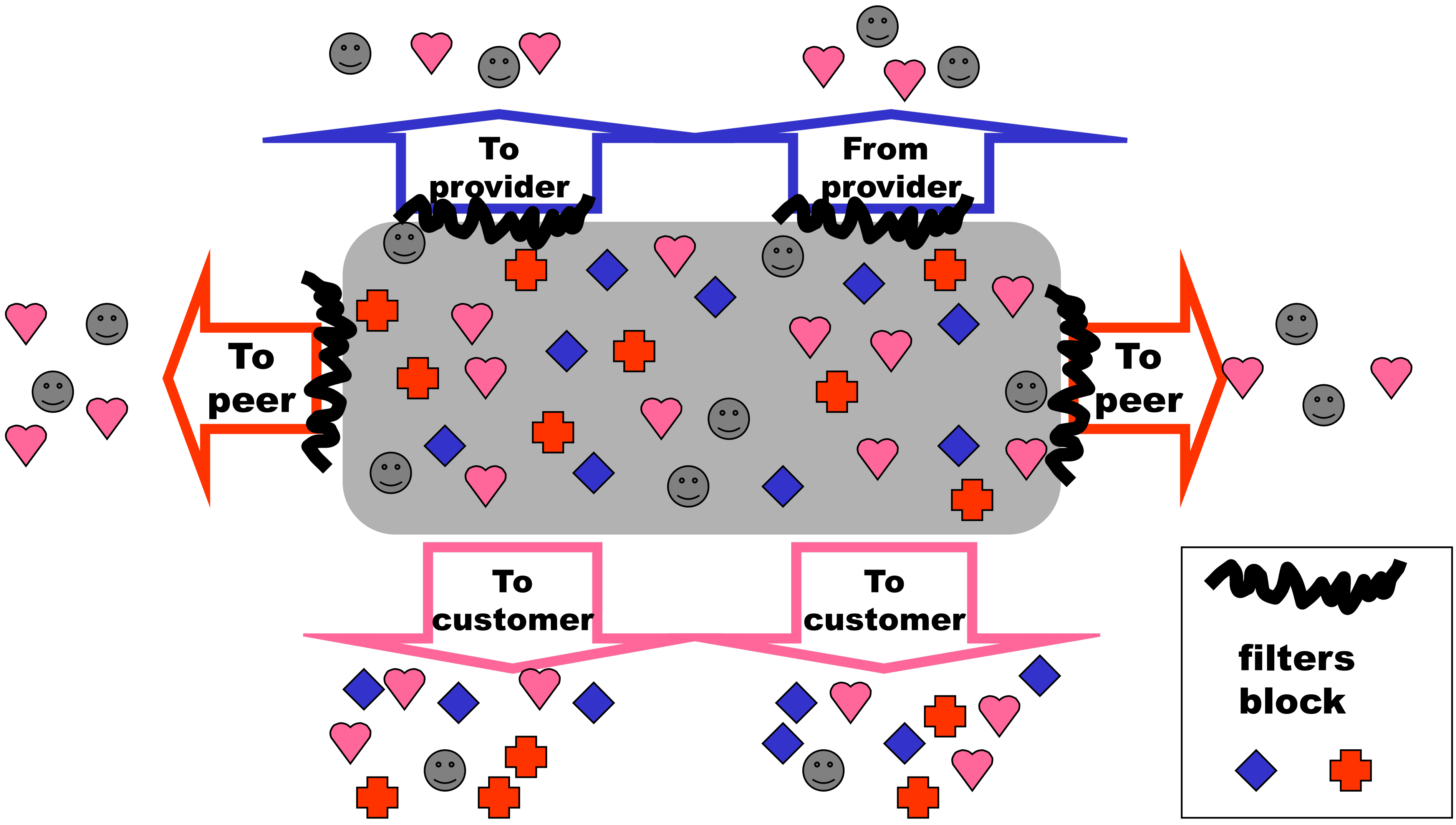
Import Routes

◆ **provider route** ✚ **peer route** ♥ **customer route** ☺ **ISP route**



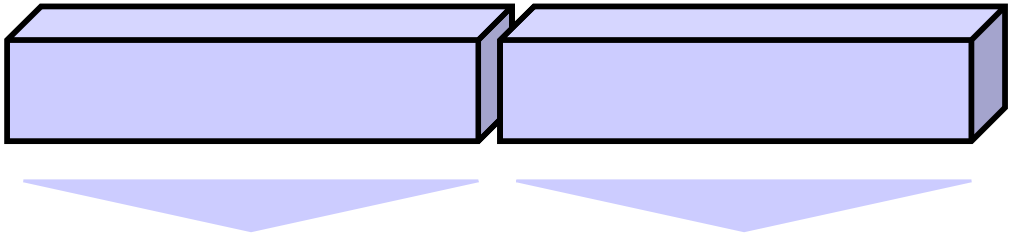
Export Routes

◆ provider route ✚ peer route ♥ customer route ☺ ISP route



How Can Routes be Colored? BGP Communities!

A community value is 32 bits



By convention, first 16 bits is ASN indicating who is giving it an interpretation

community number

Used for signalling within and between ASes

Very powerful BECAUSE it has no (predefined) meaning

Community Attribute = a list of community values. (So one route can belong to multiple communities)

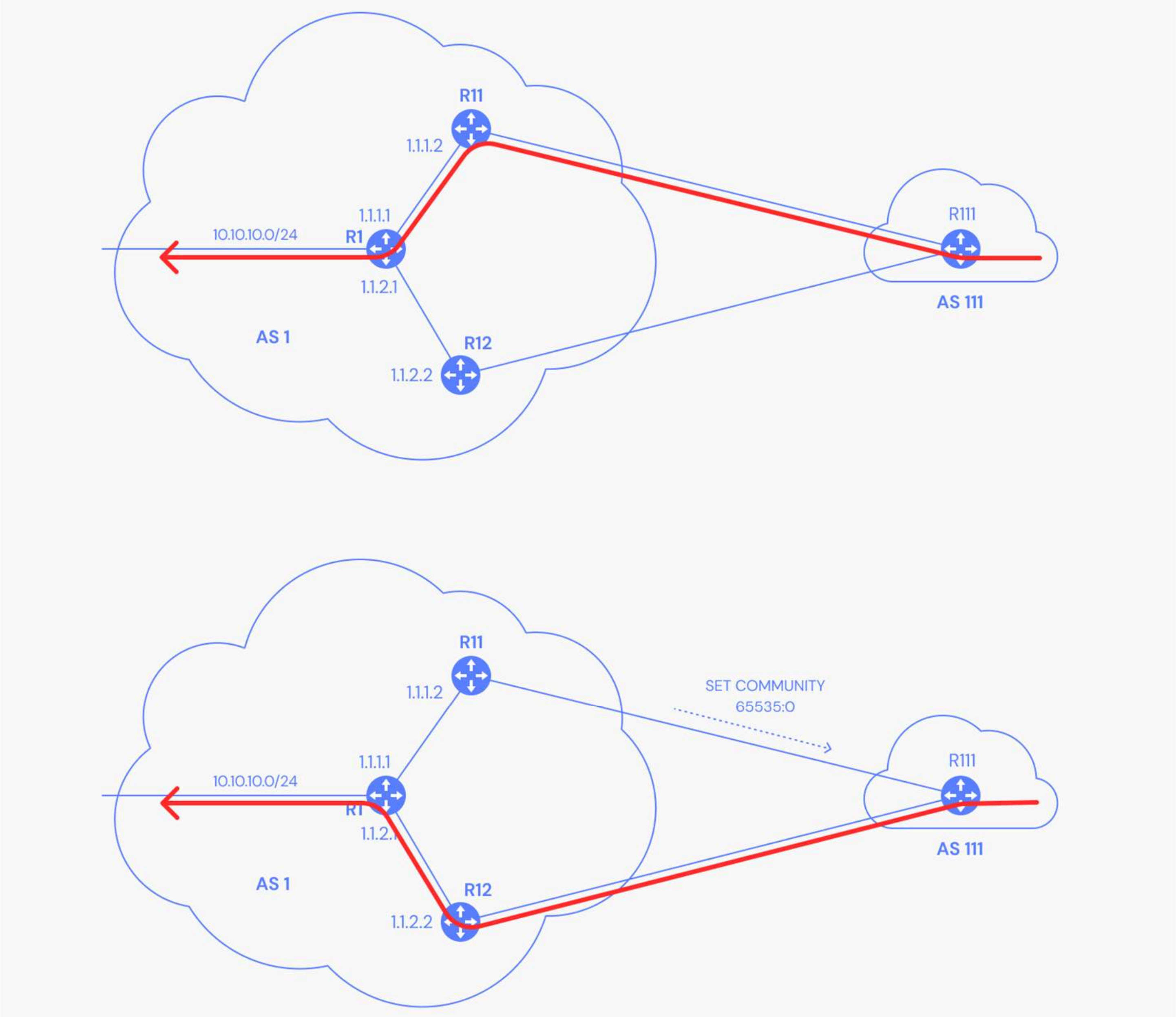
RFC 1997 (August 1996)

Two reserved communities

no_export = 0xFFFFFF01: don’t export out of AS

no_advertise 0xFFFFFF02: don’t pass to BGP neighbors

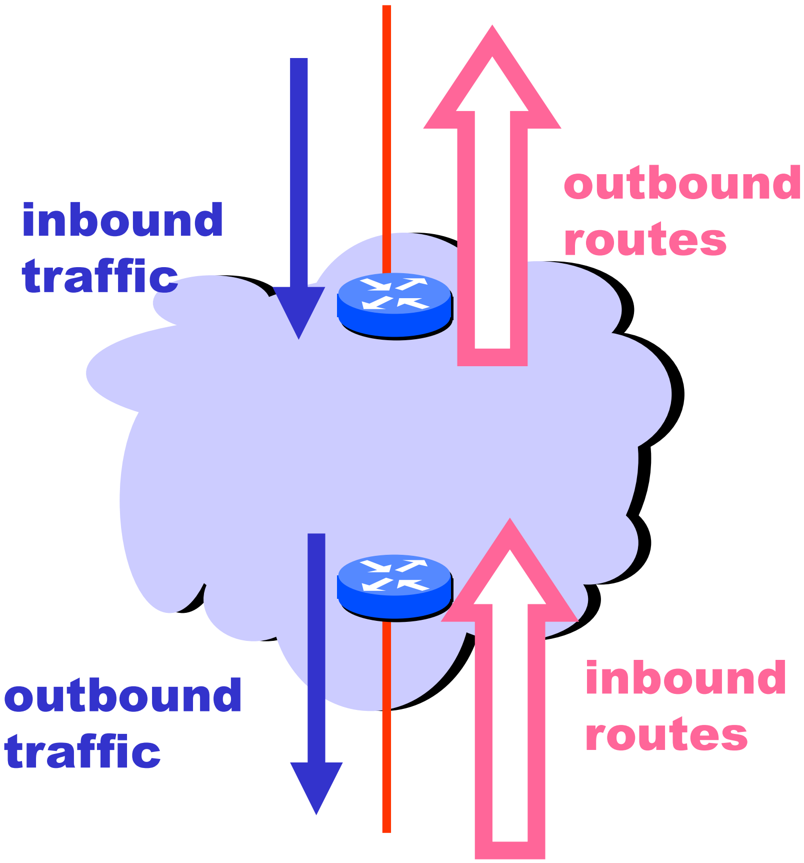
BGP Community attribute: Example



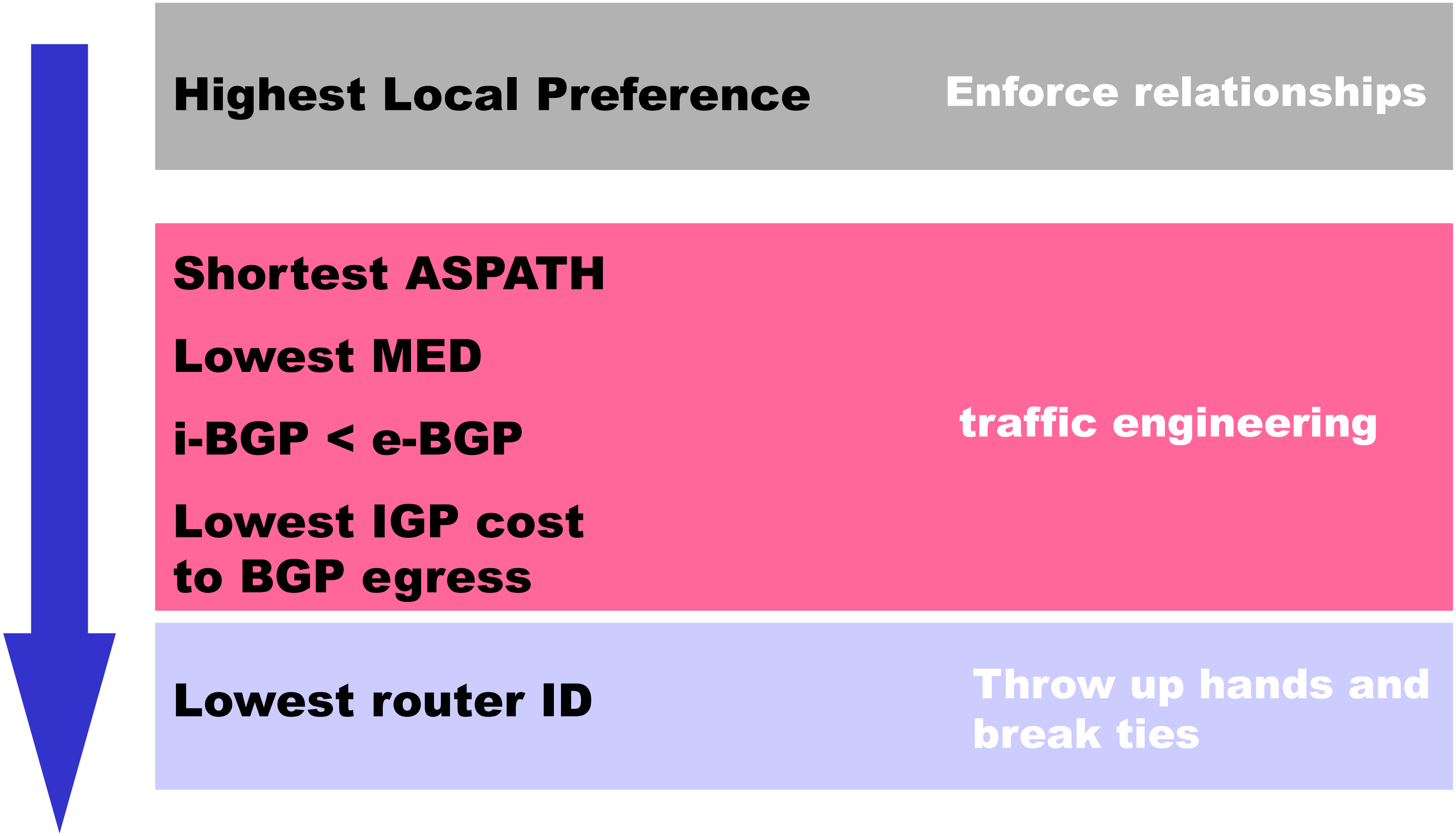
Tweak Tweak Tweak

- For inbound traffic
  - Filter outbound routes
  - Tweak attributes on outbound routes in the hope of influencing your neighbor’s best route selection
- For outbound traffic
  - Filter inbound routes
  - Tweak attributes on inbound routes to influence best route selection

In general, an AS has more control over outbound traffic



Route Selection Summary



Highest Local Preference

Enforce relationships

Shortest ASPATH

Lowest MED

i-BGP < e-BGP

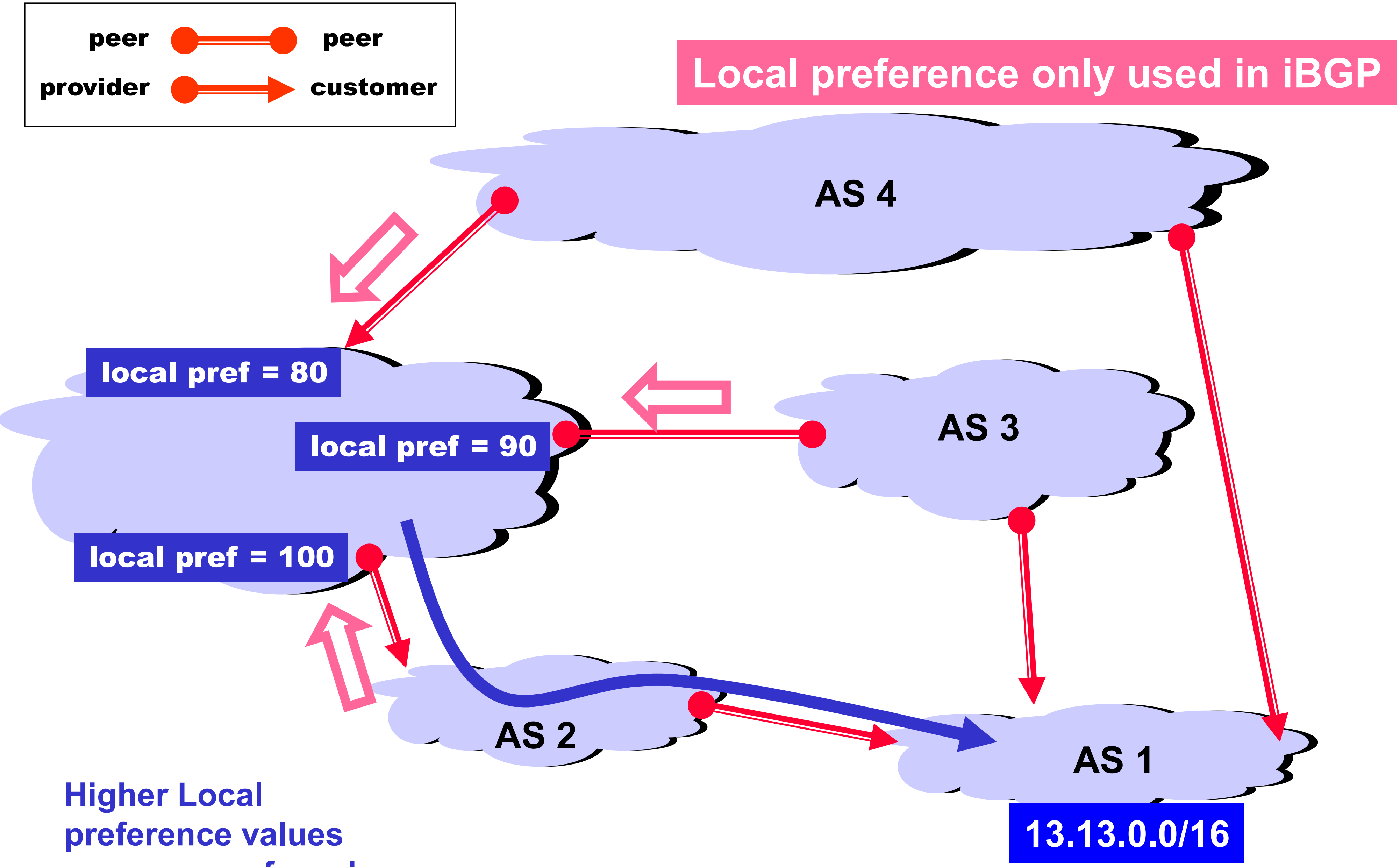
Lowest IGP cost
to BGP egress

traffic engineering

Lowest router ID

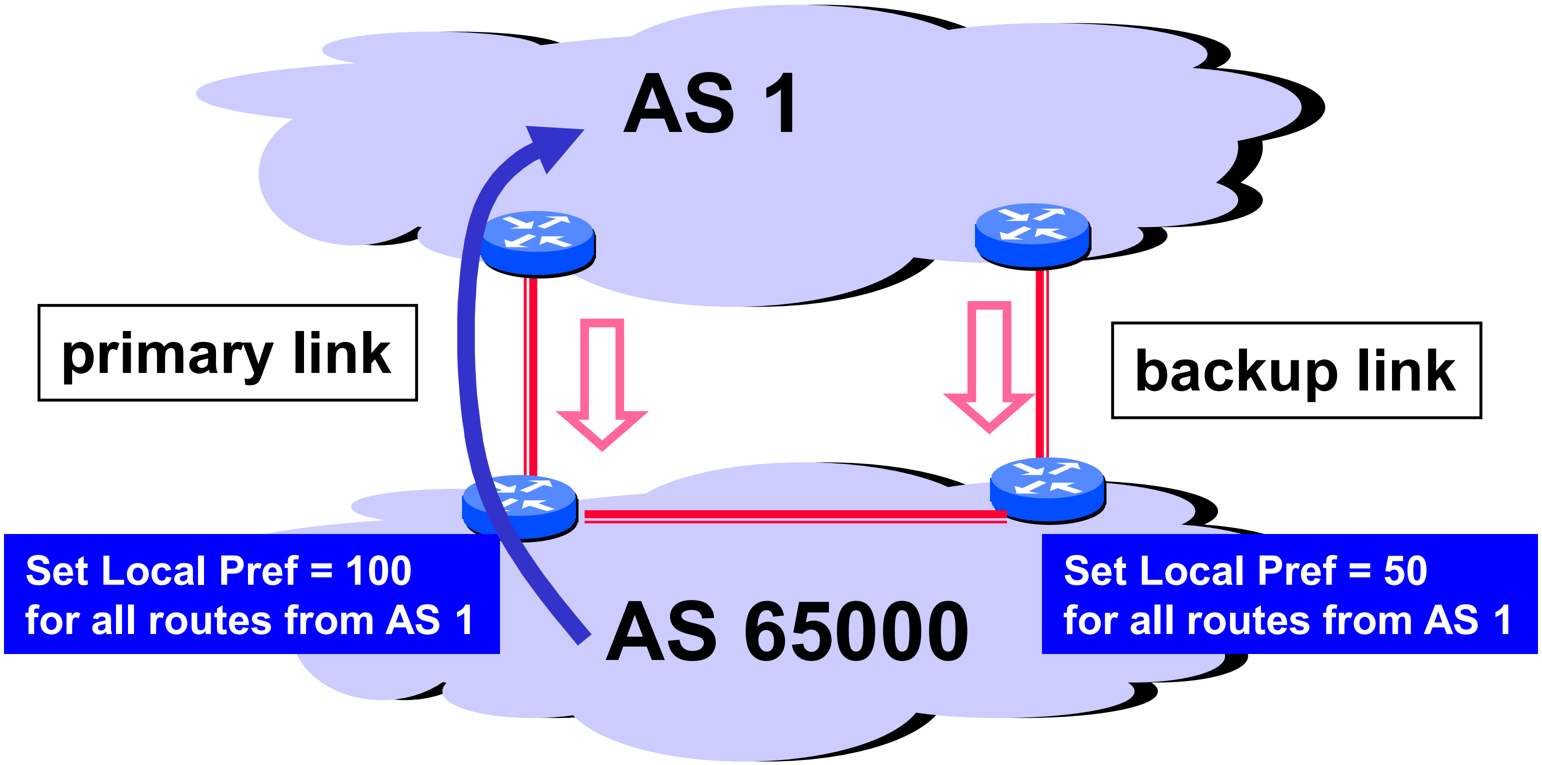
Throw up hands and
break ties

Back to Frank ...



Higher Local preference values are more preferred

Implementing Backup Links with Local Preference (Outbound Traffic)



AS 1

primary link

backup link

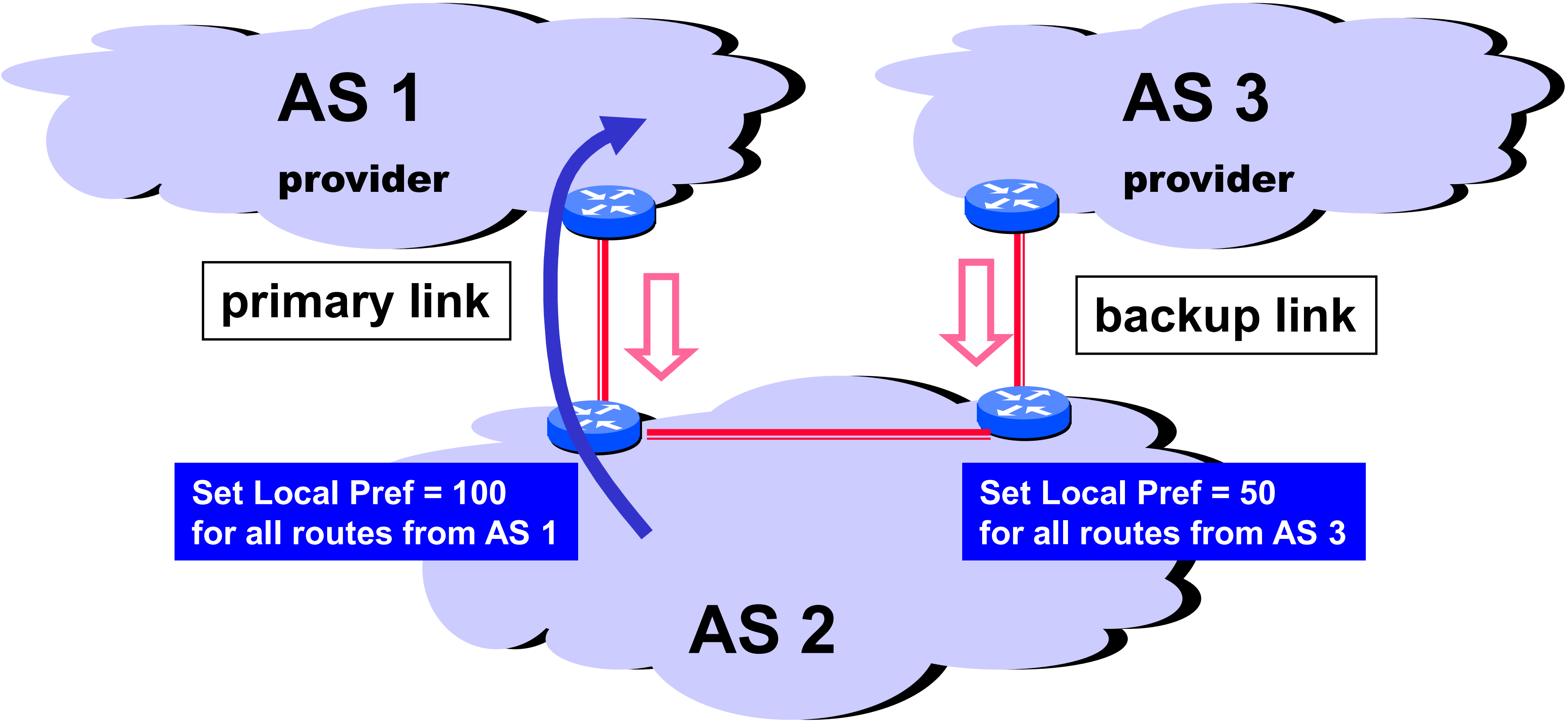
Set Local Pref = 100
for all routes from AS 1

AS 65000

Set Local Pref = 50
for all routes from AS 1

Forces outbound traffic to take primary link, unless link is down.

Multihomed Backups (Outbound Traffic)

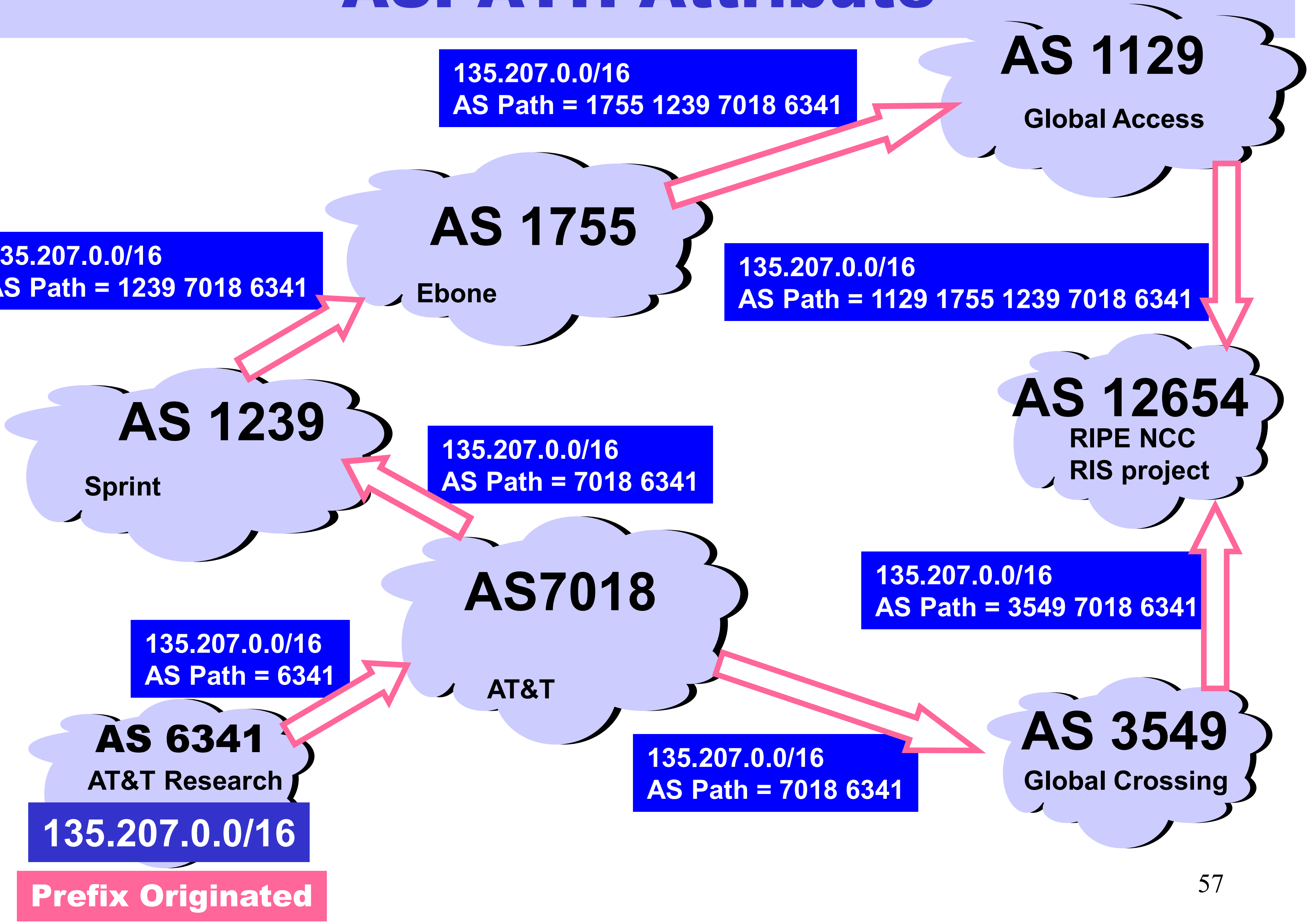


Set Local Pref = 100 for all routes from AS 1

Set Local Pref = 50 for all routes from AS 3

Forces outbound traffic to take primary link, unless link is down.

ASPATH Attribute



AS 6341

AT&T Research

135.207.0.0/16

Prefix Originated

135.207.0.0/16
AS Path = 6341

AS7018

AT&T

135.207.0.0/16
AS Path = 7018 6341

AS 1239

Sprint

135.207.0.0/16
AS Path = 1239 7018 6341

AS 1755

Ebone

135.207.0.0/16
AS Path = 1755 1239 7018 6341

AS 1129

Global Access

135.207.0.0/16
AS Path = 1129 1755 1239 7018 6341

135.207.0.0/16
AS Path = 7018 6341

AS 3549

Global Crossing

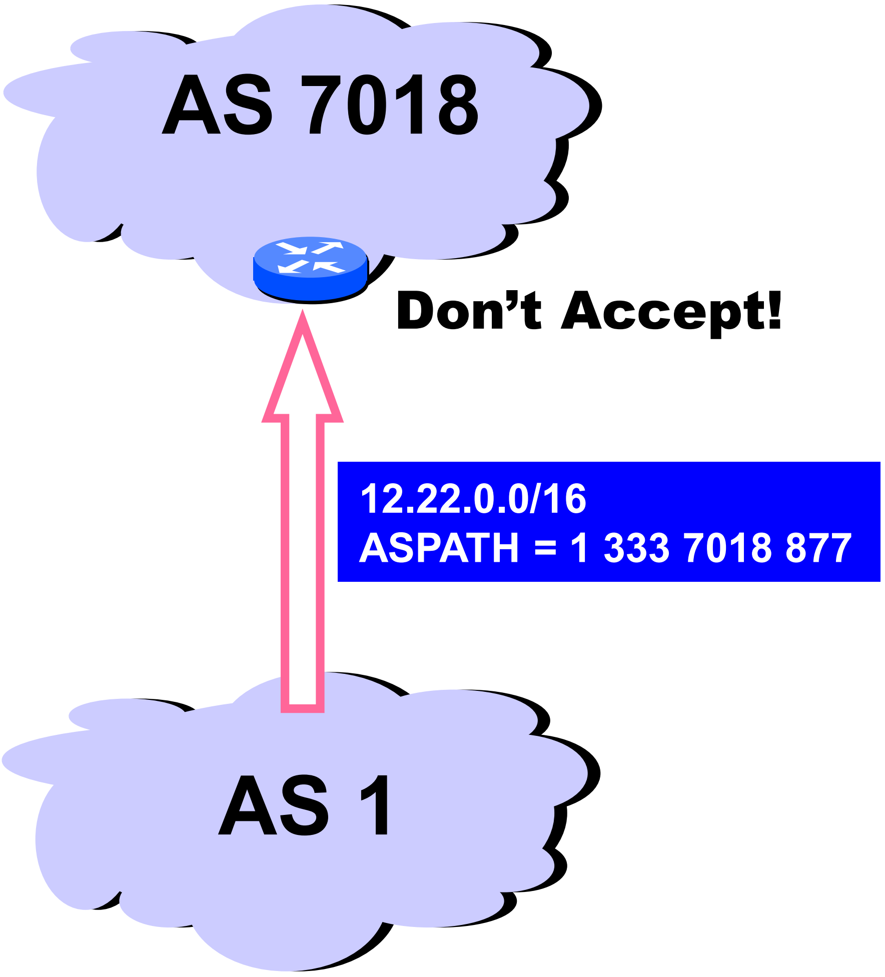
135.207.0.0/16
AS Path = 3549 7018 6341

AS 12654

RIPE NCC
RIS project

Interdomain Loop Prevention

BGP at AS YYY will never accept a route with ASPATH containing YYY.



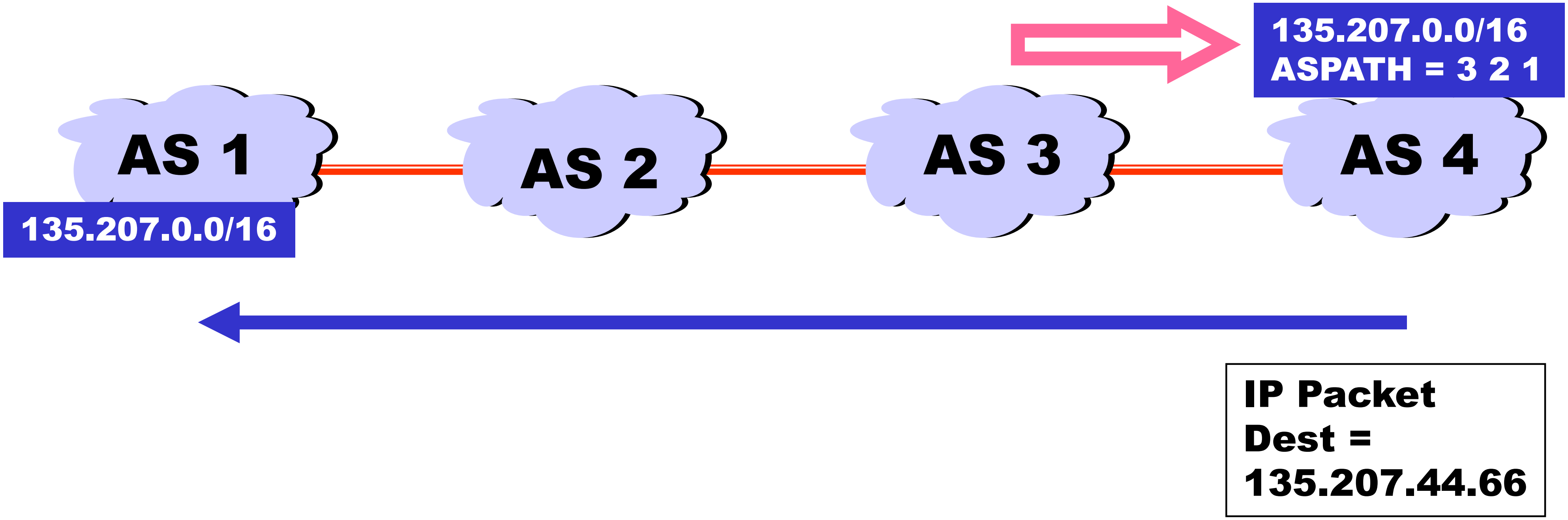
AS 7018

Don’t Accept!

12.22.0.0/16
ASPATH = 1 333 7018 877

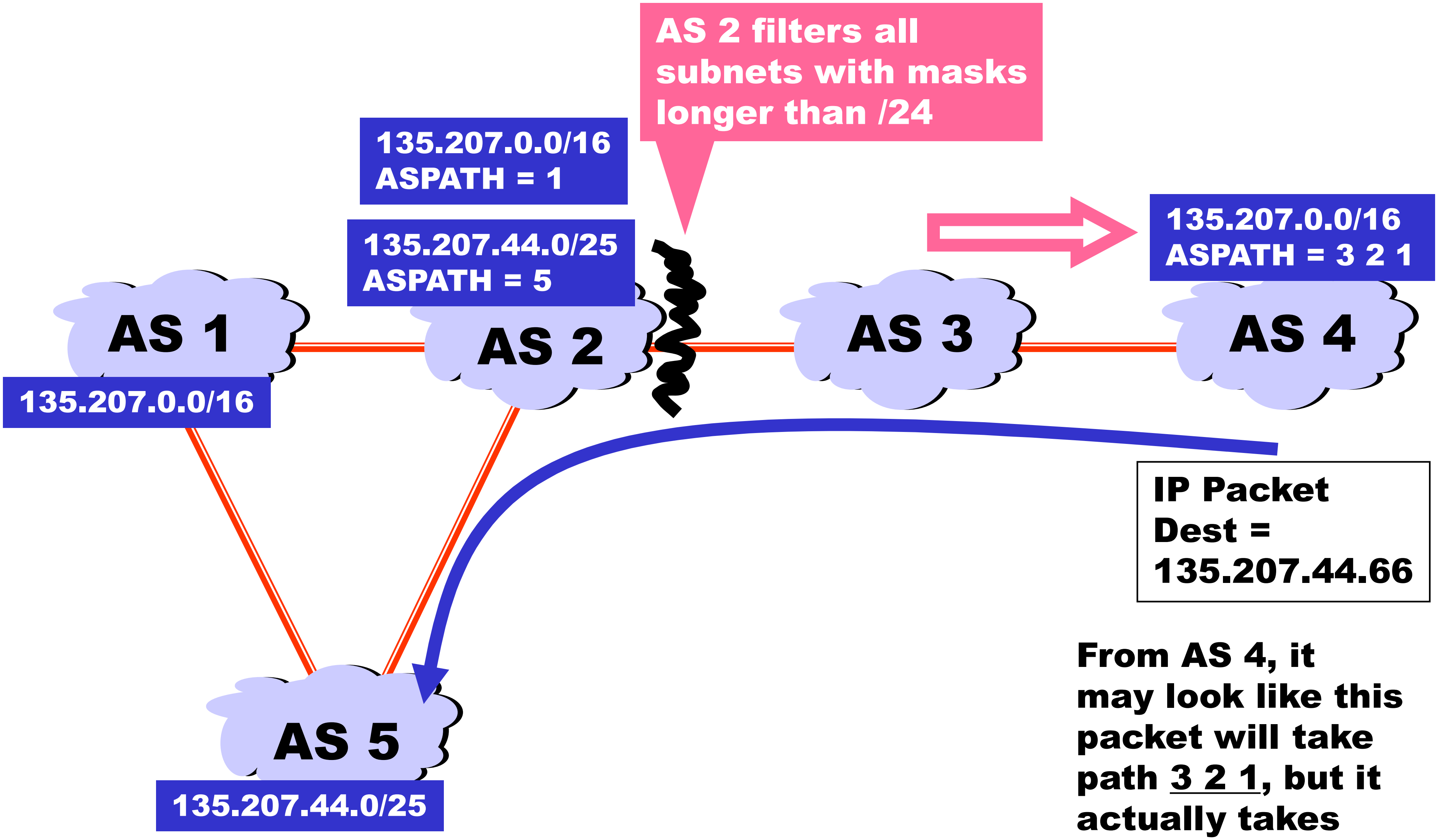
AS 1

Traffic Often Follows ASPATH



IP Packet
Dest =
135.207.44.66

... But It Might Not

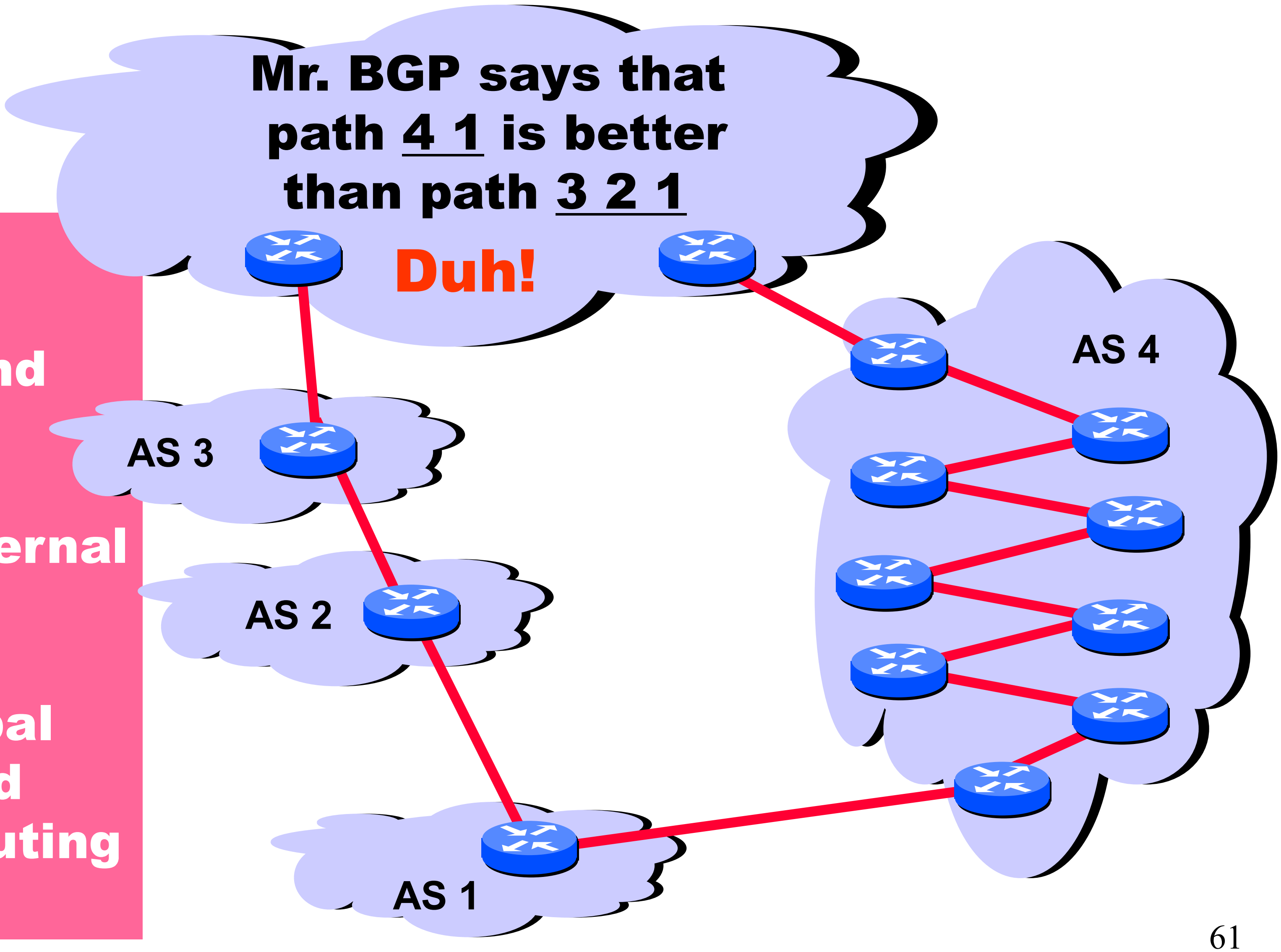


From AS 4, it may look like this packet will take path 3 2 1, but it actually takes path 3 2 5

Shorter Doesn’t Always Mean Shorter

In fairness: could you do this “right” and still scale?

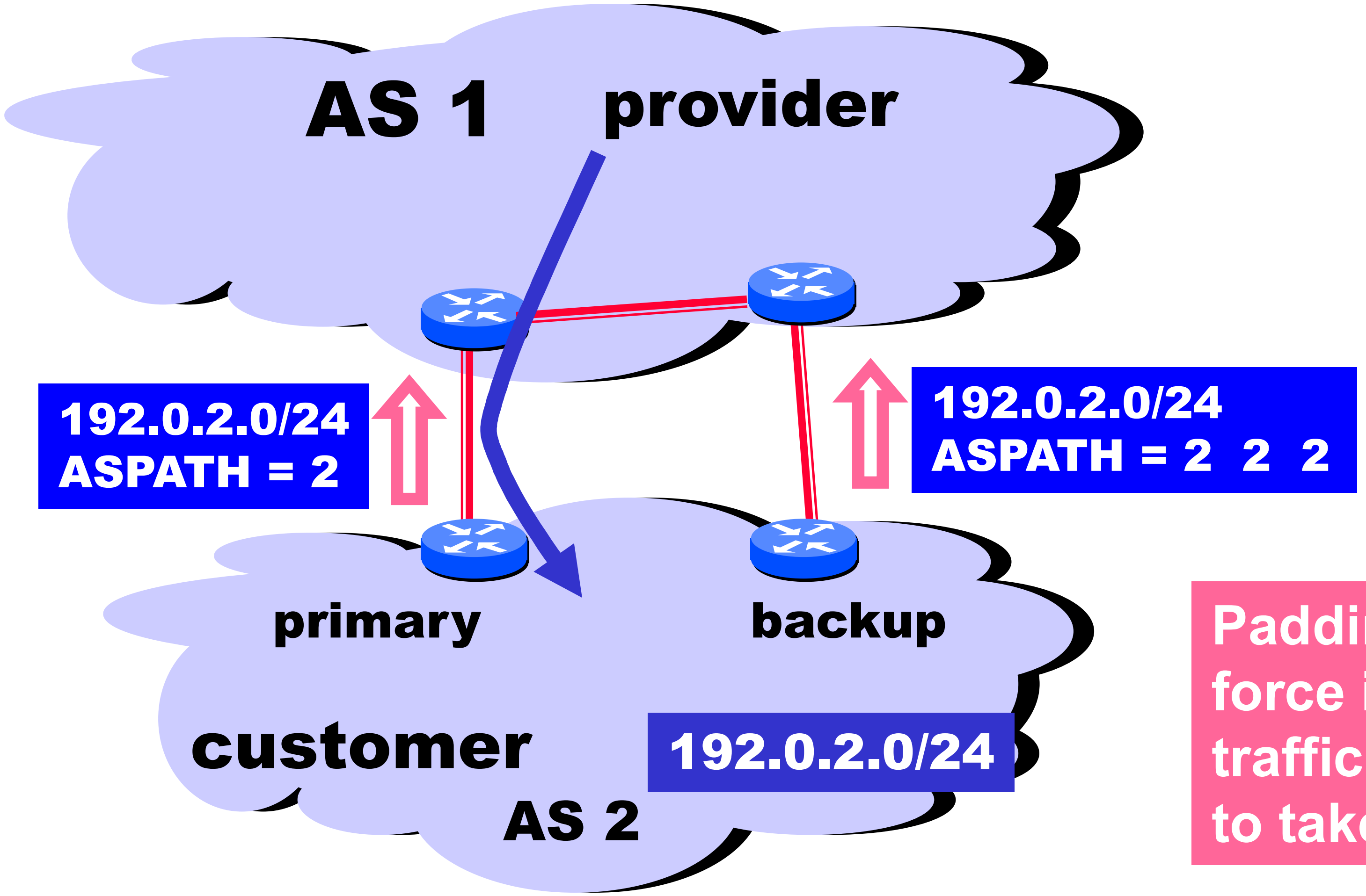
Exporting internal state would dramatically increase global instability and amount of routing state



Mr. BGP says that path 4 1 is better than path 3 2 1

Duh!

Shedding Inbound Traffic with ASPATH Padding Hack



AS 1 provider

192.0.2.0/24
ASPATH = 2

192.0.2.0/24
ASPATH = 2 2 2

primary

backup

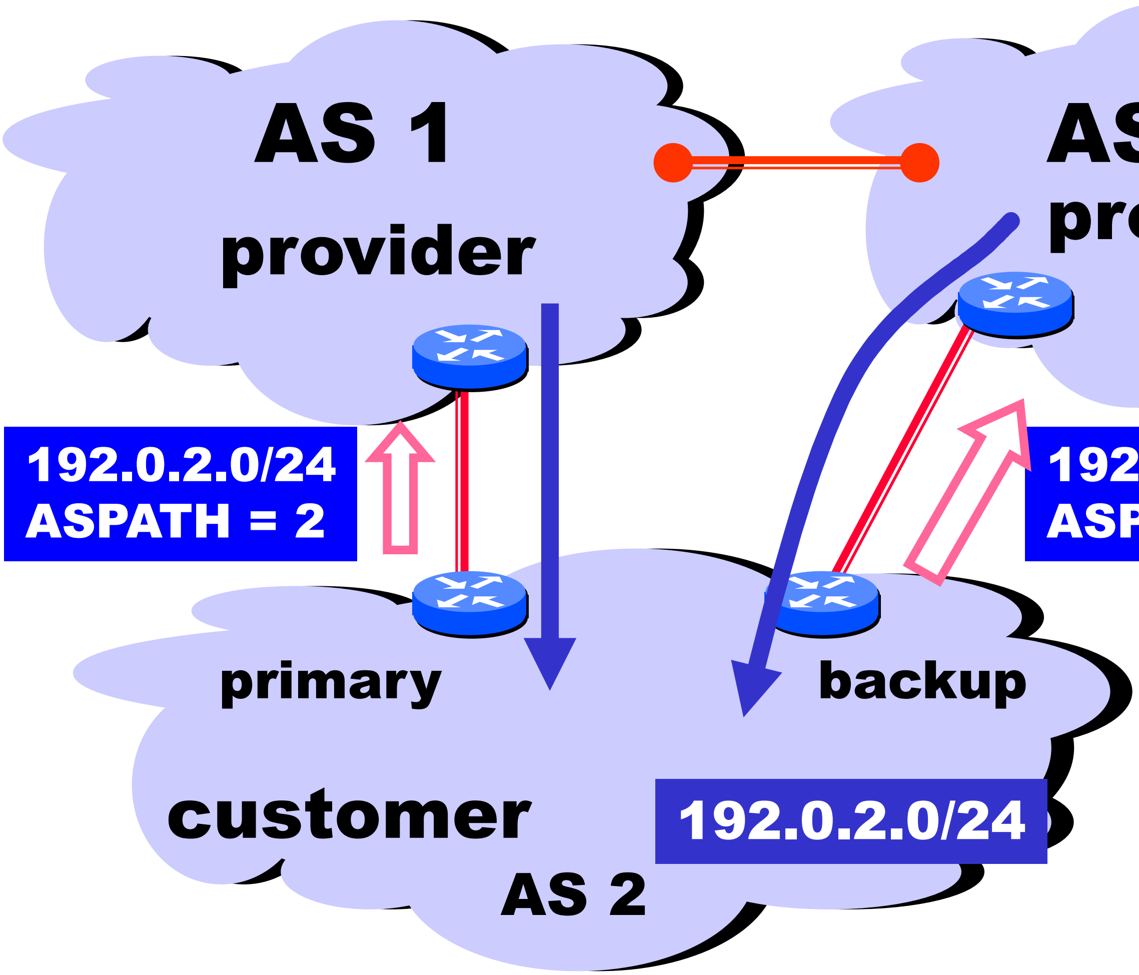
customer

AS 2

192.0.2.0/24

Padding will (usually) force inbound traffic from AS 1 to take primary link

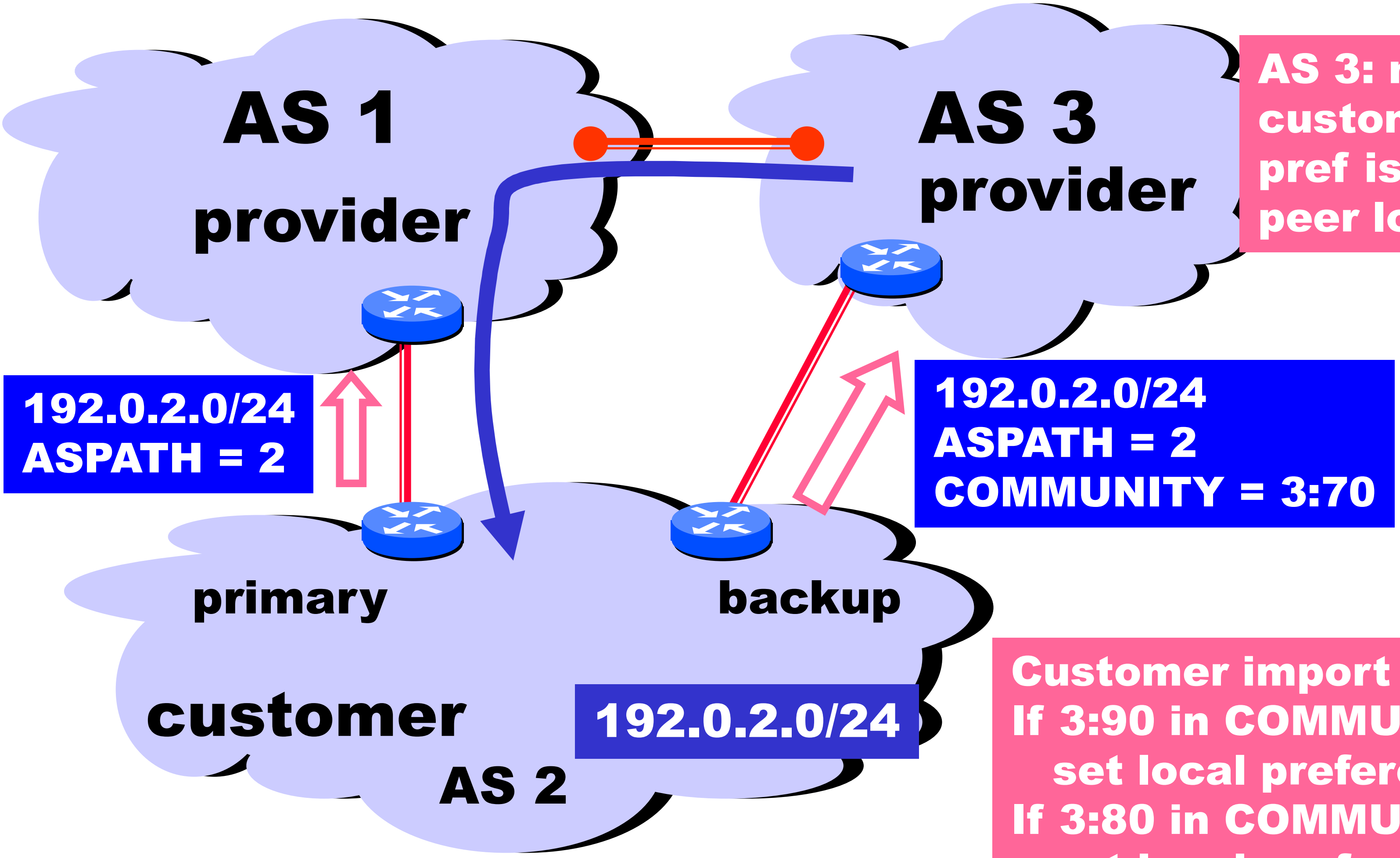
Padding May Not Shut Off All Traffic



AS 3 will send traffic on “backup” link because it prefers customer routes and local preference is considered before ASPATH length!

Padding in this way is often used as a form of load balancing

COMMUNITY Attribute to the Rescue!



AS 1

provider

AS 3

provider

192.0.2.0/24
ASPATH = 2

192.0.2.0/24
ASPATH = 2
COMMUNITY = 3:70

primary

backup

customer

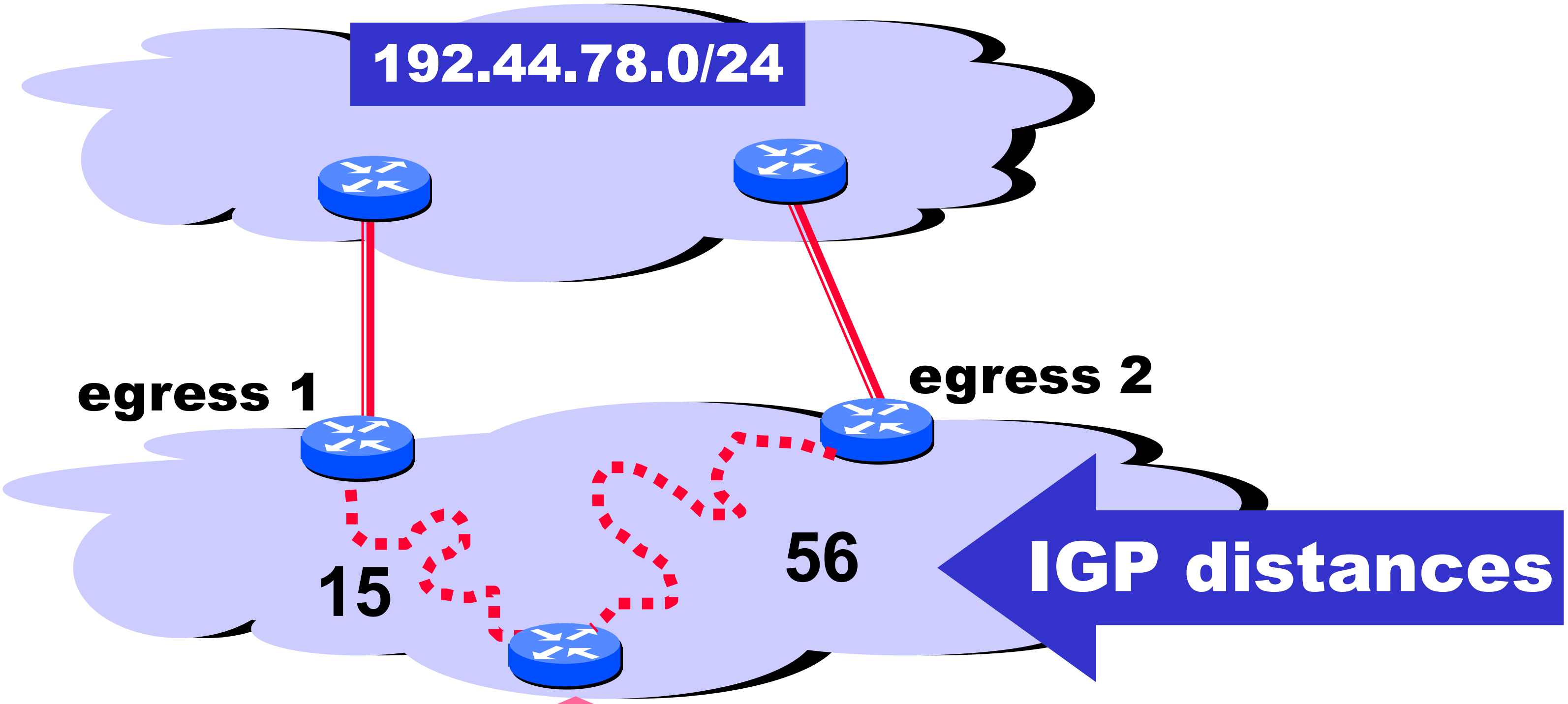
AS 2

192.0.2.0/24

AS 3: normal customer local pref is 100, peer local pref is 90

Customer import policy at AS 3:
If 3:90 in COMMUNITY then
  set local preference to 90
If 3:80 in COMMUNITY then
  set local preference to 80
If 3:70 in COMMUNITY then
  set local preference to 70

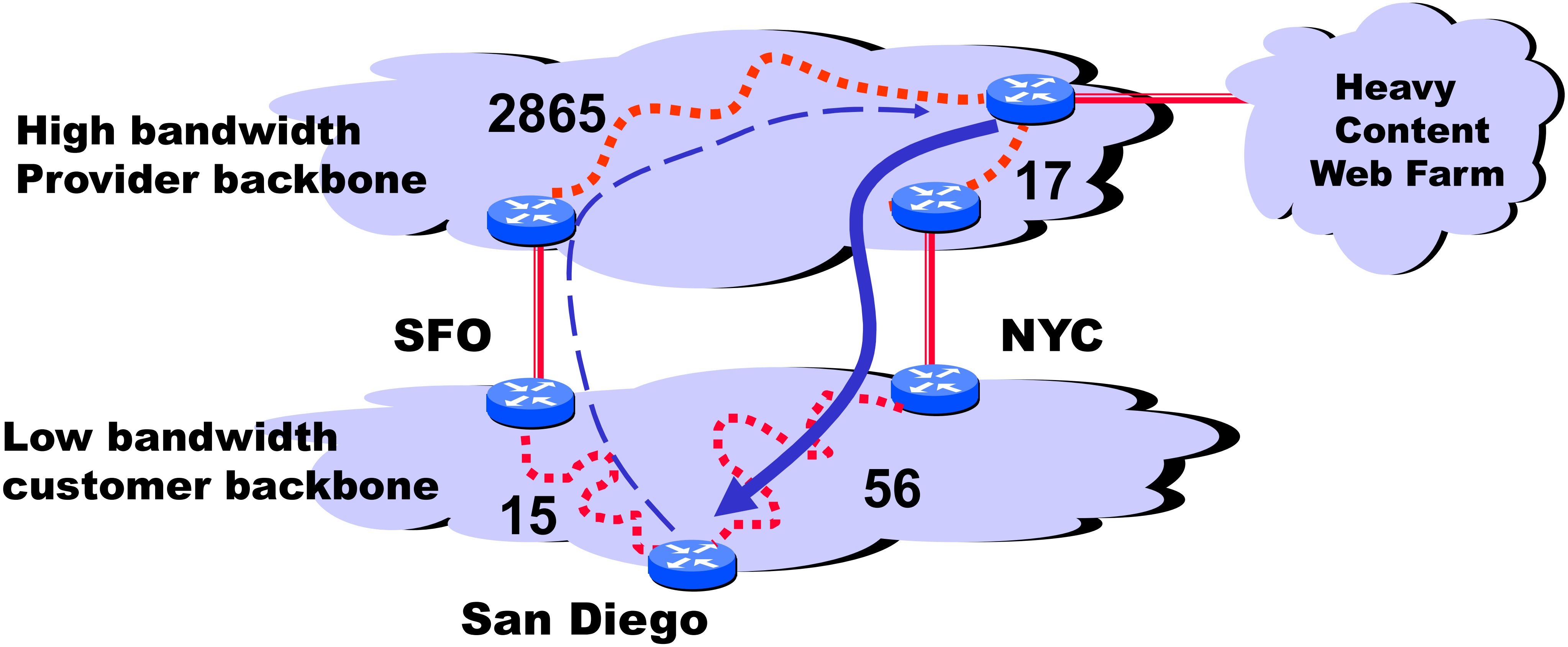
Hot Potato Routing: Go for the Closest Egress Point



This Router has two BGP routes to 192.44.78.0/24.

Hot potato: get traffic off of your network as Soon as possible. Go for egress 1!

Getting Burned by the Hot Potato



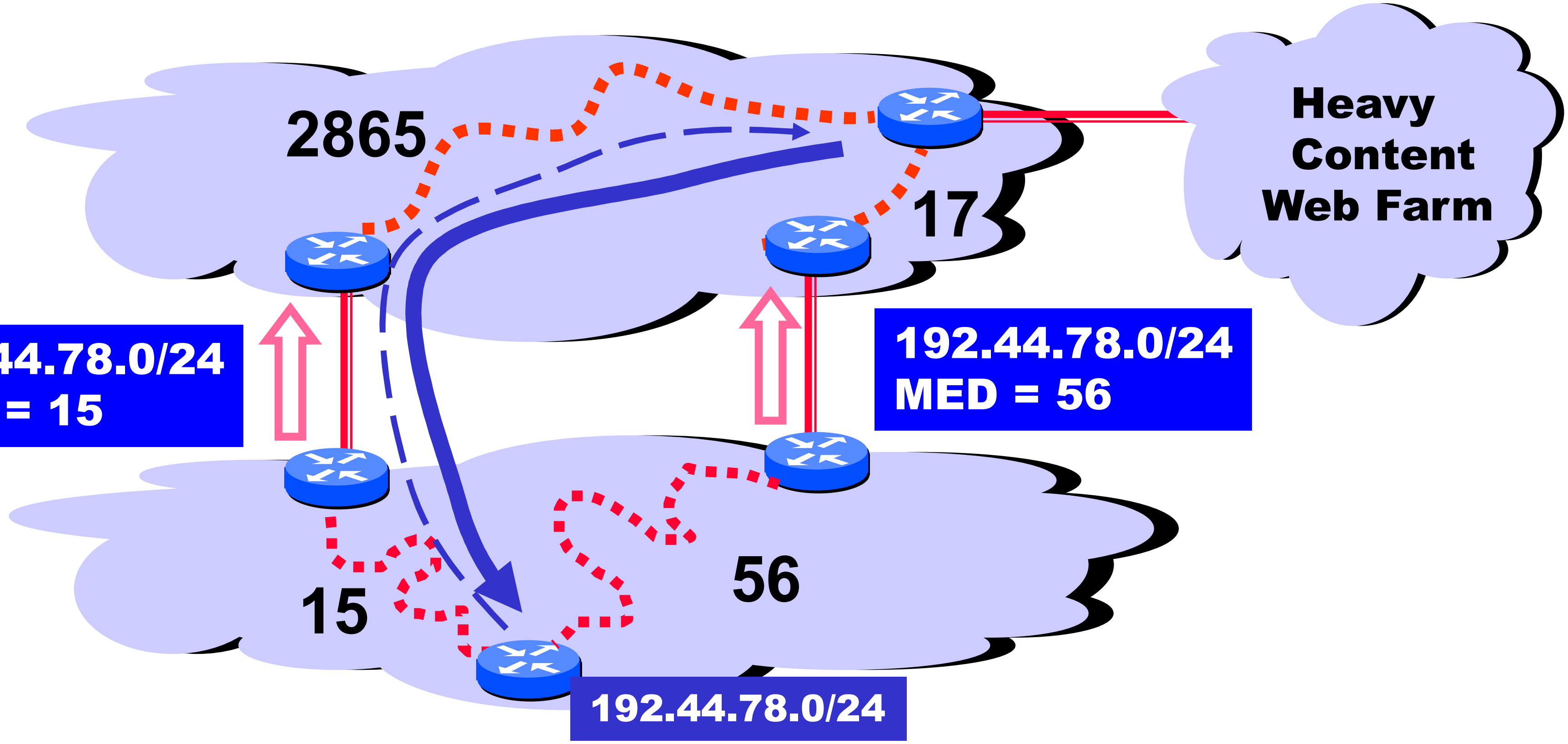
Many customers want their provider to carry the bits!

- - - → tiny http request

⟶ huge http reply

Cold Potato Routing with MEDs (Multi-Exit Discriminator Attribute)

Prefer lower MED values



2865

17

Heavy Content Web Farm

192.44.78.0/24
MED = 15

192.44.78.0/24
MED = 56

15

56

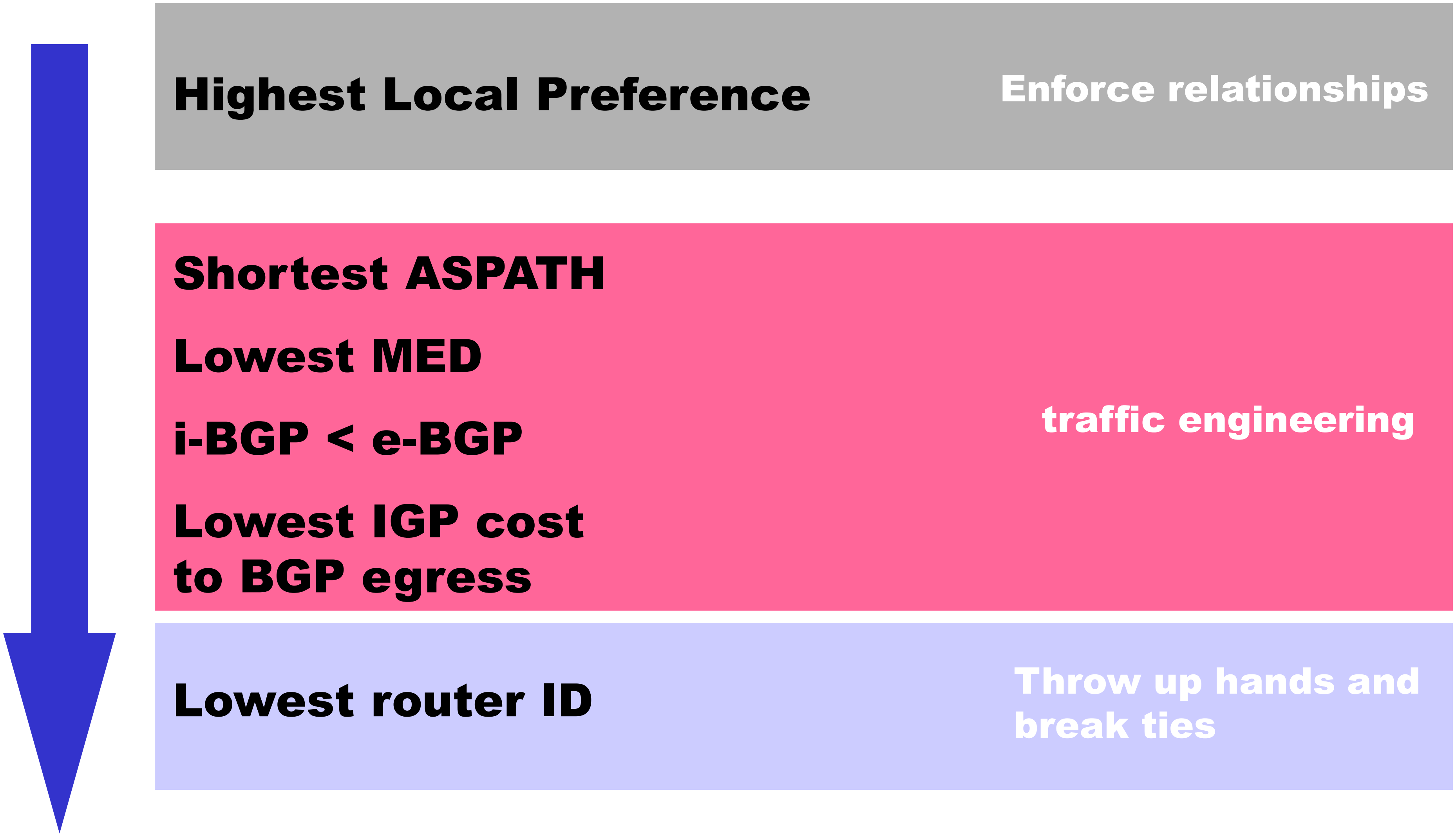
192.44.78.0/24

This means that MEDs must be considered BEFORE IGP distance!

Note1 : some providers will not listen to MEDs

Note2 : MEDs need not be tied to IGP distance

Route Selection Summary



Highest Local Preference

Enforce relationships

Shortest ASPATH

Lowest MED

i-BGP < e-BGP

Lowest IGP cost
to BGP egress

traffic engineering

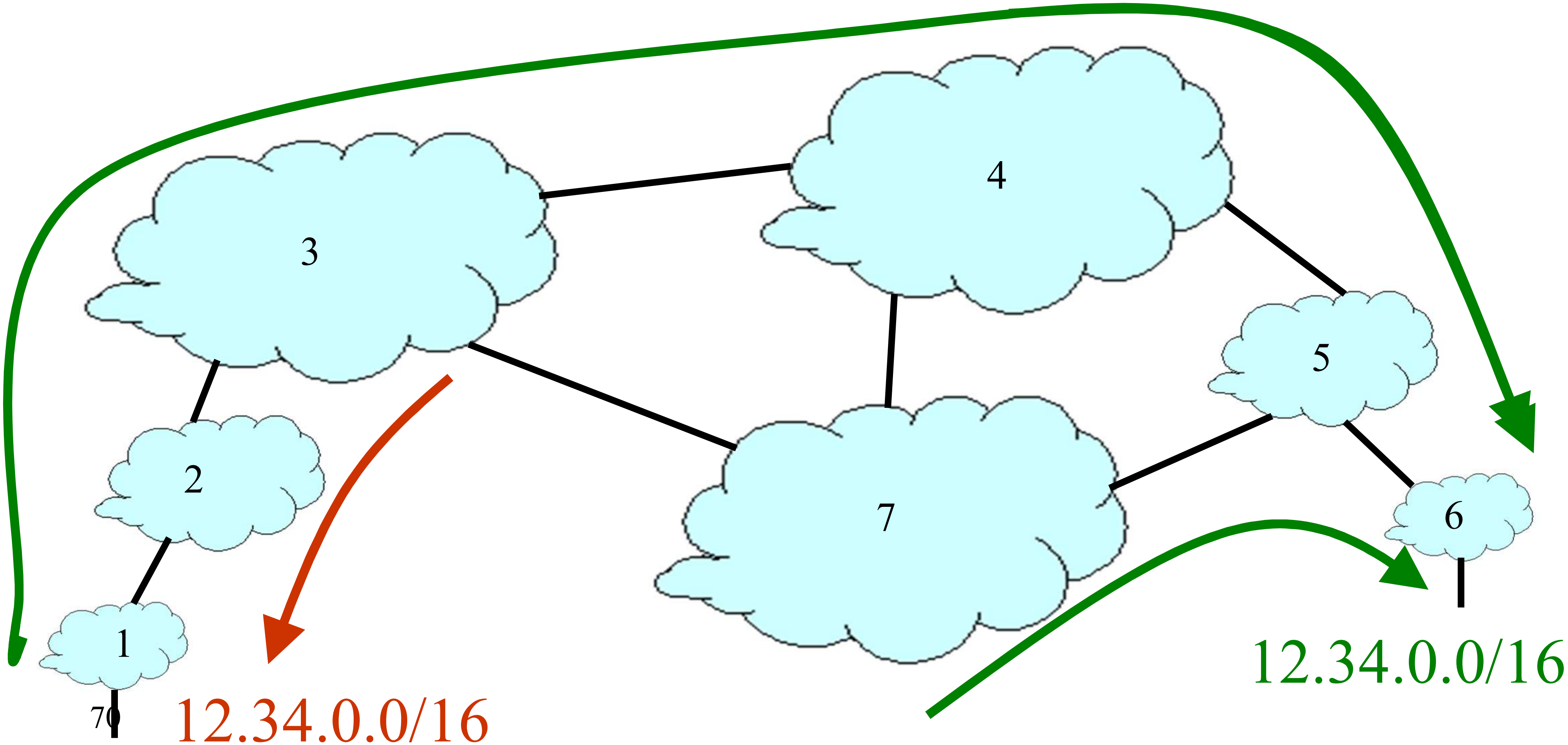
Lowest router ID

Throw up hands and
break ties

BGP Attacks

Prefix Hijacking

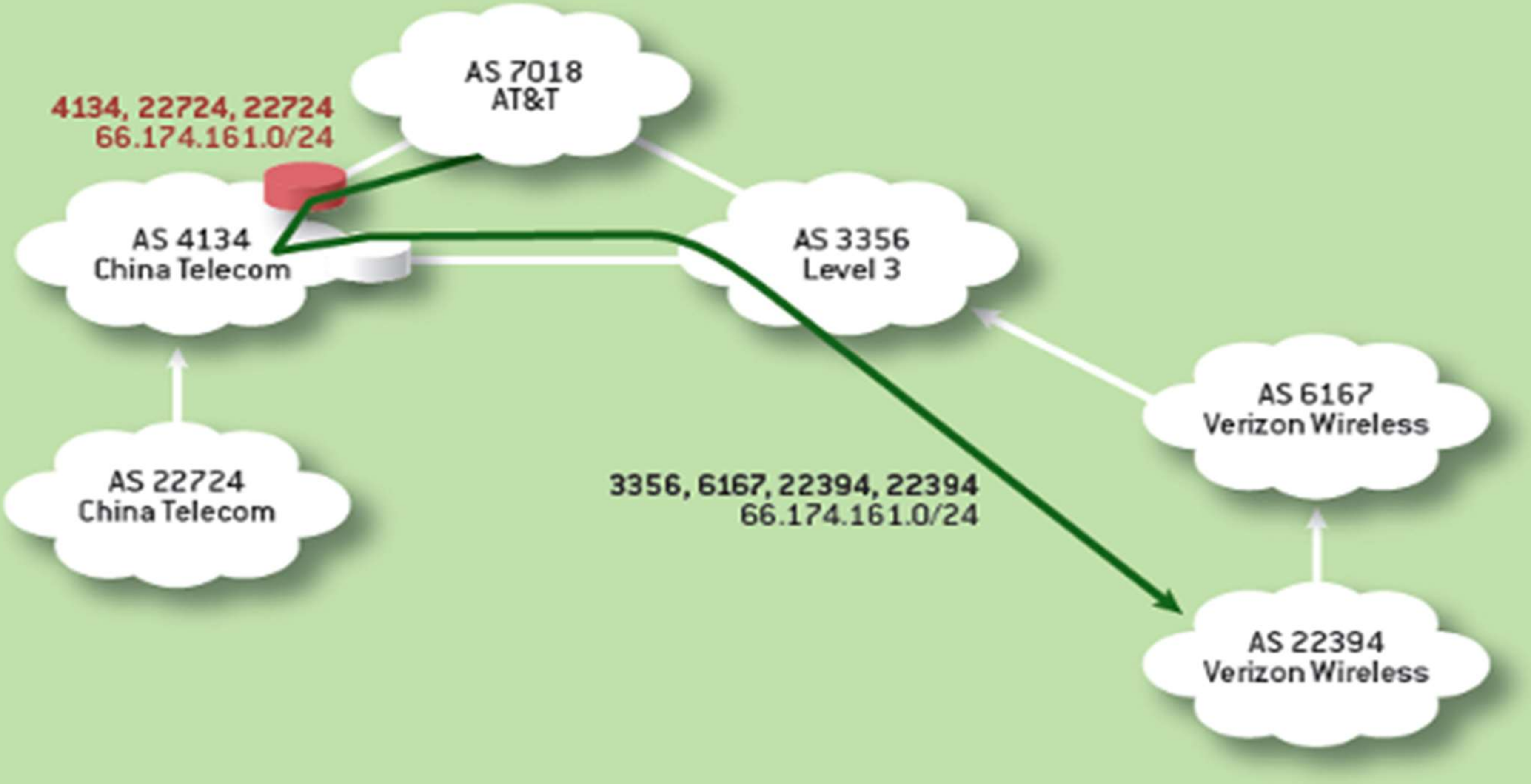
- Originating someone else's prefix
  - What fraction of the Internet believes it?



Prefix highjack

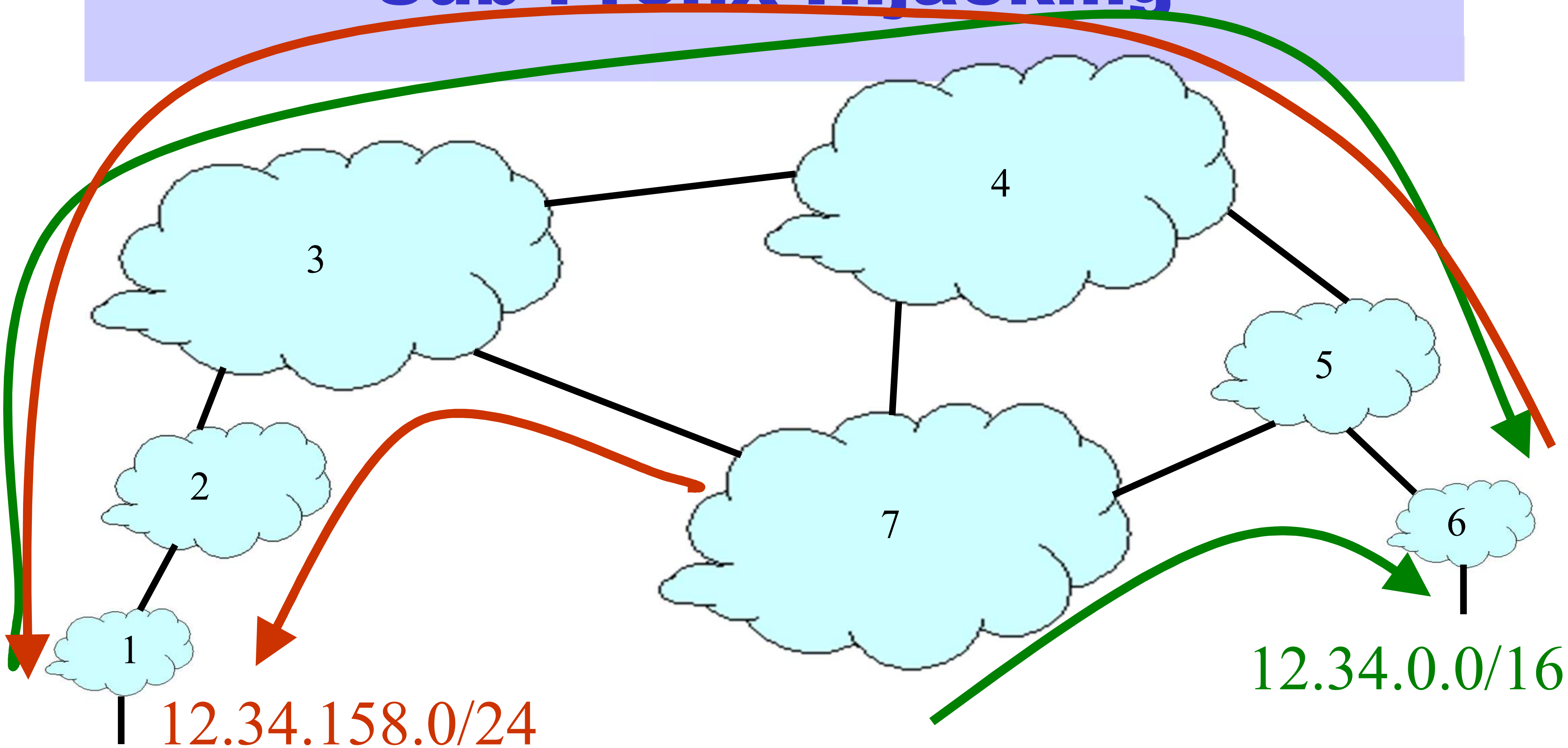
FIGURE 2

China Telecom Hijacks Verizon Wireless[17,41]



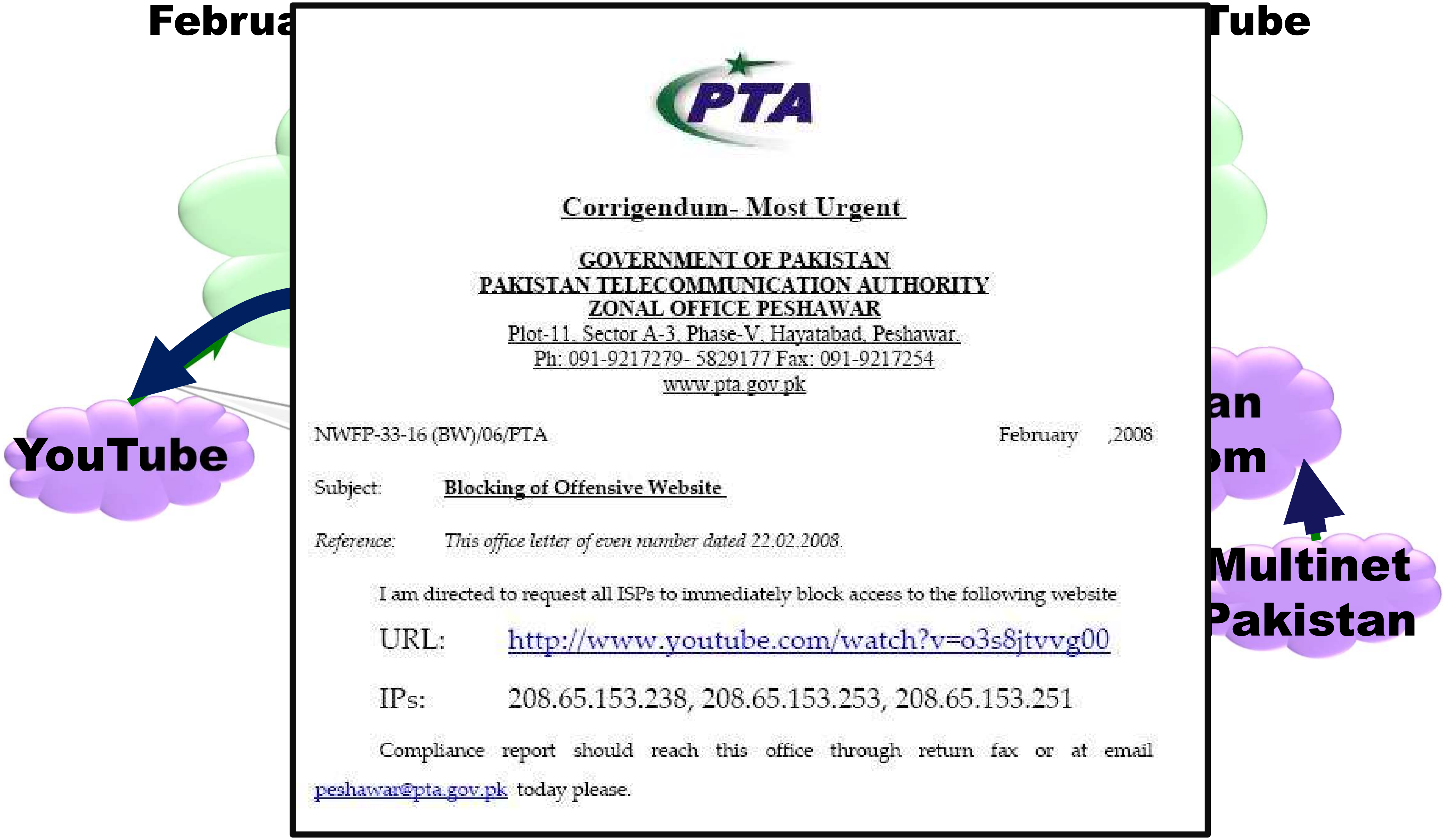
http://queue.acm.org/detail.cfm?id=2668966

Sub-Prefix Hijacking



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

Sub-prefix hijack


Februa
Tube
YouTube
an
om
Multinet
Pakistan

Corrigendum- Most Urgent

GOVERNMENT OF PAKISTAN
PAKISTAN TELECOMMUNICATION AUTHORITY
ZONAL OFFICE PESHAWAR
Plot-11, Sector A-3, Phase-V, Hayatabad, Peshawar.
Ph: 091-9217279- 5829177 Fax: 091-9217254
www.pta.gov.pk

NWFP-33-16 (BW)/06/PTA February ,2008

Subject: **Blocking of Offensive Website**

Reference: This office letter of even number dated 22.02.2008.

I am directed to request all ISPs to immediately block access to the following website

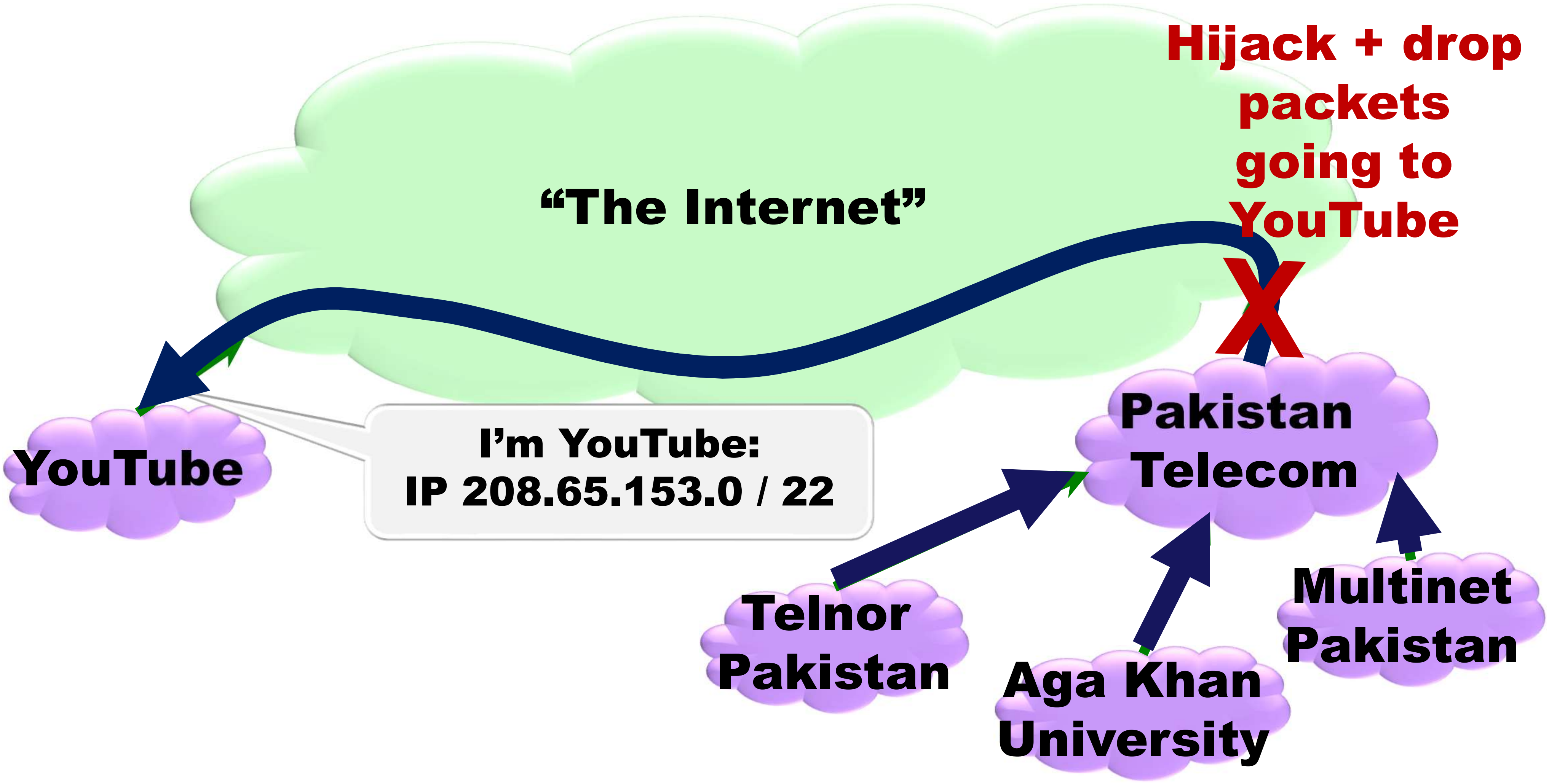
URL: http://www.youtube.com/watch?v=o3s8jtvvg00

IPs: 208.65.153.238, 208.65.153.253, 208.65.153.251

Compliance report should reach this office through return fax or at email peshawar@pta.gov.pk today please.

Pakistan Telecom: Sub-prefix hijack

Here’s what should have happened….



“The Internet”

Hijack + drop packets going to YouTube

X

YouTube

I’m YouTube:
IP 208.65.153.0 / 22

Pakistan Telecom

Telnor Pakistan

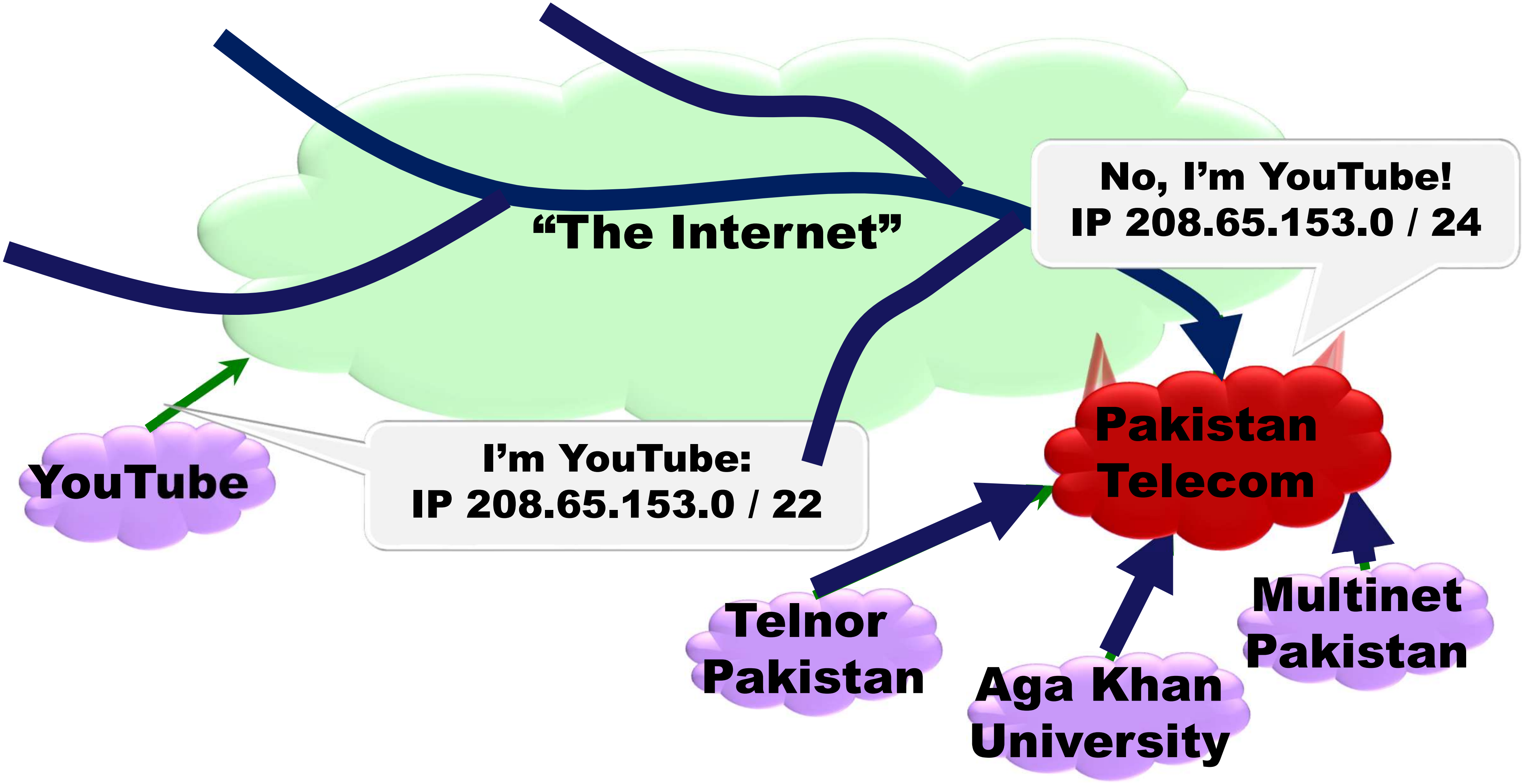
Aga Khan University

Multinet Pakistan

Block your own customers.

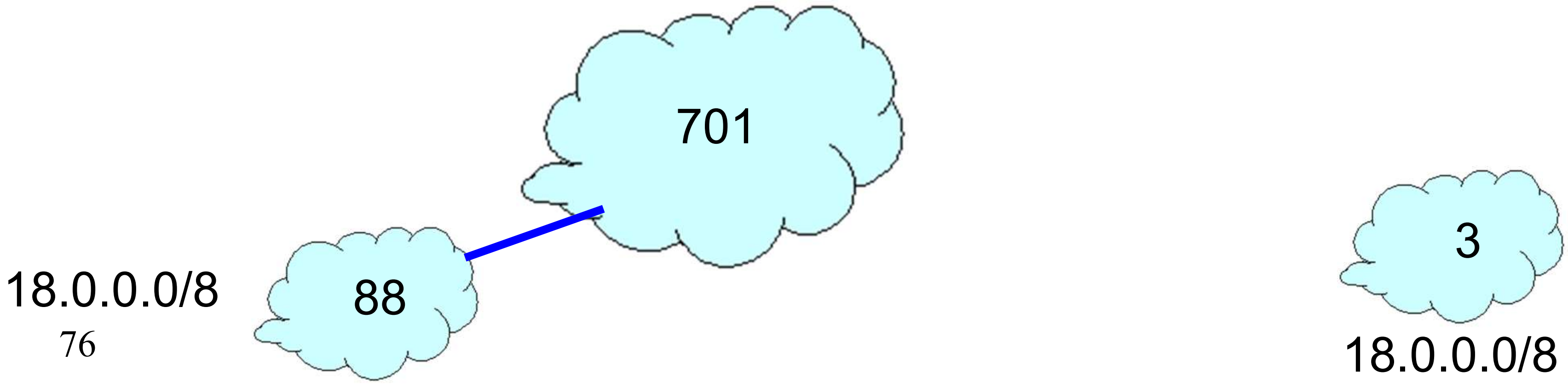
Pakistan Telecom: Sub-prefix hijack

But here's what Pakistan ended up doing…



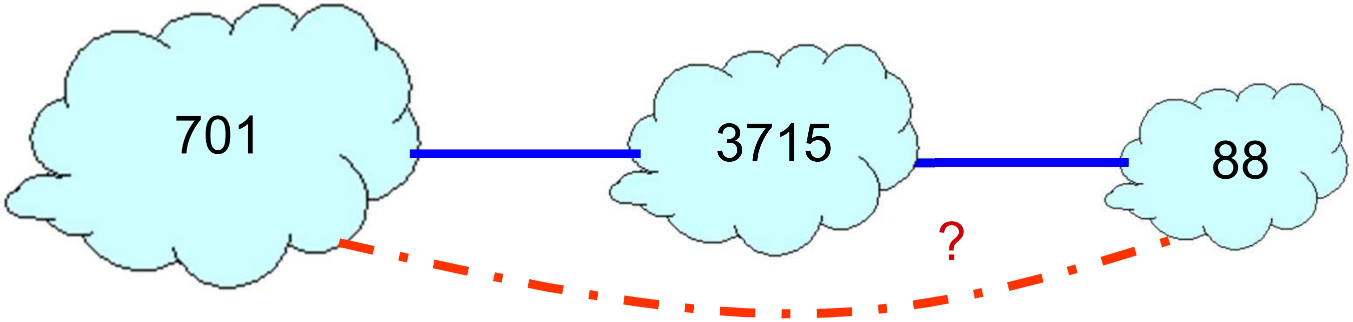
Bogus AS Paths to Hide Hijacking

- Adds AS hop(s) at the end of the path
  - E.g., turns “701 88” into “701 88 3”
- Motivations
  - Evade detection for a bogus route
  - E.g., by adding the legitimate AS to the end
- Hard to tell that the AS path is bogus…
  - Even if other ASes filter based on prefix ownership

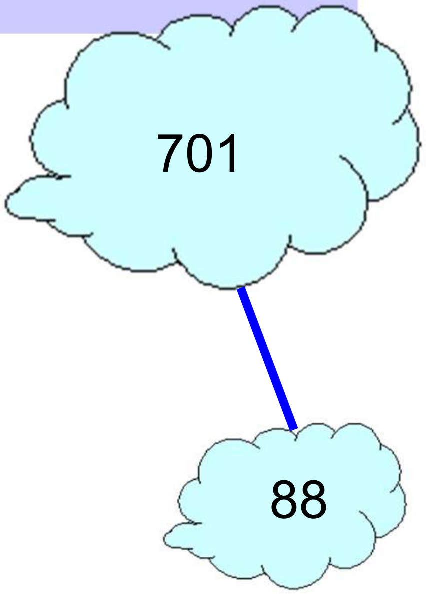


Path-Shortening Attacks

- **Remove ASes from the AS path**
  - **E.g., turn “701 3715 88” into “701 88”**
- **Motivations**
  - **Make the AS path look shorter than it is**
  - **Attract sources that normally try to avoid AS 3715**
  - **Help AS 88 look like it is closer to the Internet’s core**
- **Who can tell that this AS path is a lie?**
  - **Maybe AS 88 *does* connect to AS 701 directly**



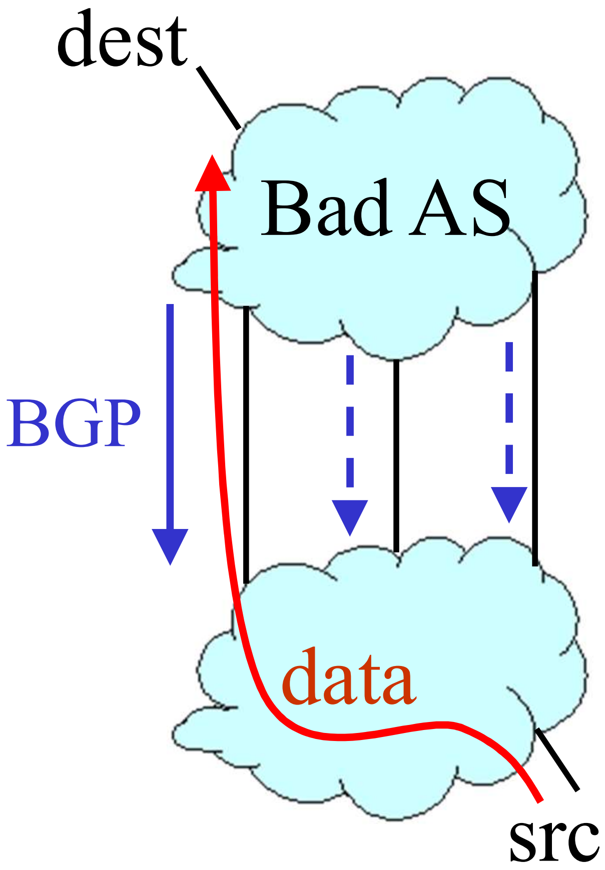
Attacks that Add a Bogus AS Hop



- **Add ASes to the path**
  - **E.g., turn "701 88" into "701 3715 88"**
- **Motivations**
  - **Trigger loop detection in AS 3715**
    - **Denial-of-service attack on AS 3715**
    - **Or, blocking unwanted traffic coming from AS 3715!**
  - **Make your AS look like is has richer connectivity**
- **Who can tell the AS path is a lie?**
  - **AS 3715 could, if it could see the route**
  - **AS 88 could, but would it really care as long as it received data traffic meant for it?**

Violating "Consistent Export" to Peers

- **Peers require consistent export**
  - **Prefix advertised at all peering points**
  - **Prefix advertised with same AS path length**
- **Reasons for violating the policy**
  - **Trick neighbor into "cold potato"**
  - **Configuration mistake**
- **Main defense**
  - **Analyzing BGP updates**
  - **… or data traffic**
  - **… for signs of inconsistency**



Other Attacks

- Attacks on BGP sessions
  - Confidentiality of BGP messages
  - Denial-of-service on BGP session
  - Inserting, deleting, modifying, or replaying messages
- Resource exhaustion attacks
  - Too many IP prefixes (e.g., BGP “512K Day”)
  - Too many BGP update messages
- Data-plane attacks
  - Announce one BGP routes, but use another

Solution Techniques

- **Protective filtering**
  - **Know your neighbors**
- **Anomaly detection**
  - **Suspect the unexpected**
- **Checking against registries**
  - **Establish ground truth for prefix origination**
- **Signing and verifying**
  - **Prevent bogus AS PATHs**
- **Data-plane verification**
  - **Ensure the path is actually followed**