ISP, peering and inter-domain routing

IERG5090 – Jan 23, 2017

Outline

Two topics:

- ISPs, and their peering relationships
- Inter-domain routing between ISPs (BGP)

Acknowledgement: Some slides and drawings from Tim Griffin’s talk

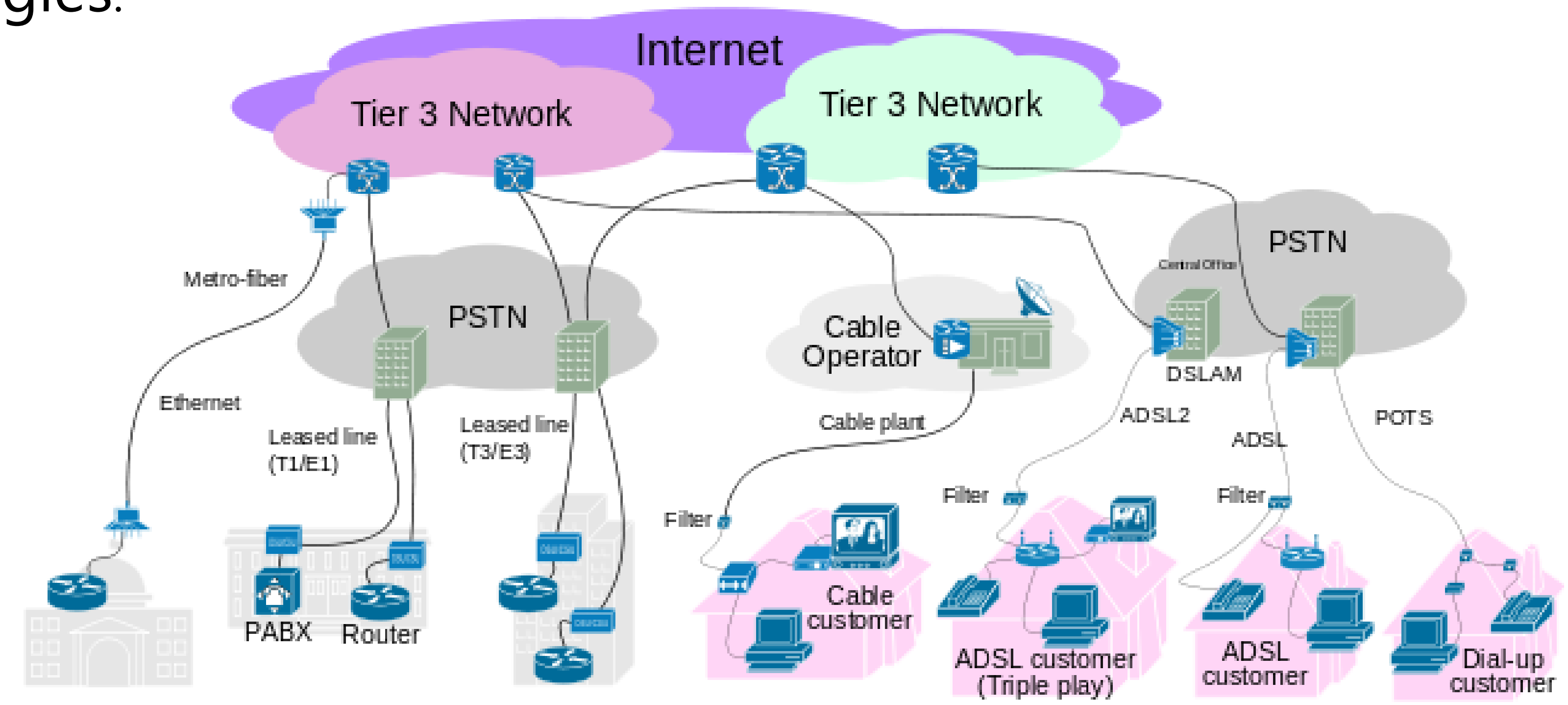
Internet Service Provider (ISP)

Different access technologies:

- Dialup
- ADSL
- Cable modem
- Leased lines
- Ethernet
- Wireless (Wifi, or Cellular)

Subscribers typically pay monthly fees:

- Bandwidth
- Traffic volume
- Flat rate



From Wikipedia: ISP

Autonomous System (AS)

Autonomous System

- A collection of physical networks that share the same policy and resources for external routing.
- Each AS has a AS Number, used to identify AS in routing: http://www.whatismyasn.org
- AS numbers assigned to ISP in a similar way as IP addresses
- A ISP needs one ASN, but may have more (due to merger etc)
- AS number used to be 16 bits, now 32 bits (written as x.y, 0.y are the old ASNs)

Examples:

Genuity: 1
MIT: 3
Harvard: 11
CUHK: 3661
AT&T: 7018, 6341, 5074, …
UUNET: 701, 702, 284, 12199, …
Sprint: 1239, 1240, 6211, 6242, …

Internet maps

- Maps by Tim B. Lee (2014): http://www.vox.com/a/internet-maps
- Internet activities (2013): https://www.wired.com/2015/06/mapping-the-internet/
- Various statistics: https://www.internetsociety.org/map/global-internet-report/
- CAIDA's Internet topology maps: https://www.caida.org/research/topology/
- Some example Tier 1 ISPs: https://en.wikipedia.org/wiki/Tier_1_network

Transit ISPs – Tier 1 and Tier 2

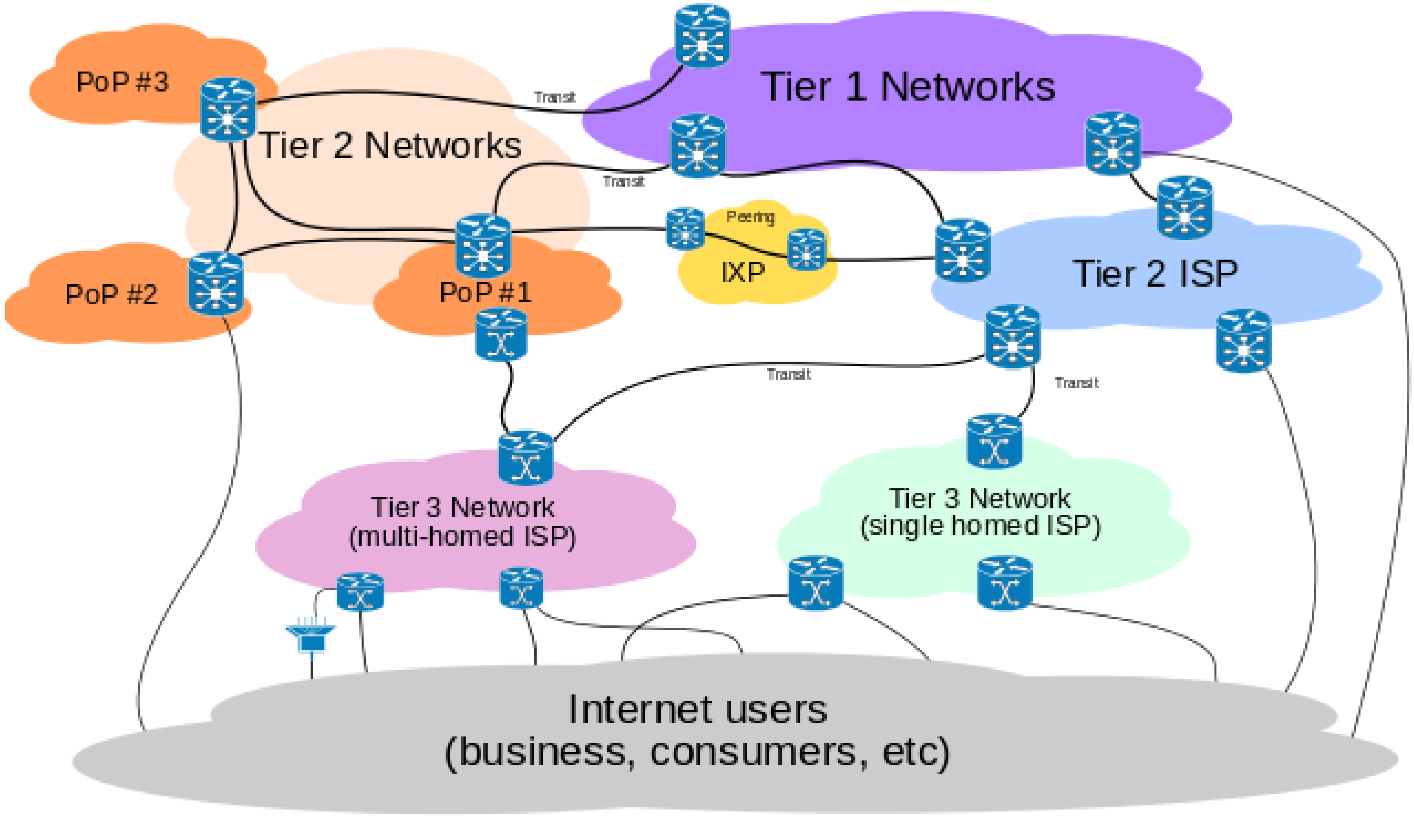
Transit ISPs – connecting distant ISPs together, provide transit service

- Tier 1: does not buy transit from others
- Tier 2: buys transit from Tier 1 ISPs for long distance transit
- Tier 3: buys transit from Tier 1 and Tier 2 ISPs

Who pays for transit service?

- Customer ISP pays Transit provider
- According to **peering** agreements

Multi-homing – serviced by multiple transit ISPs

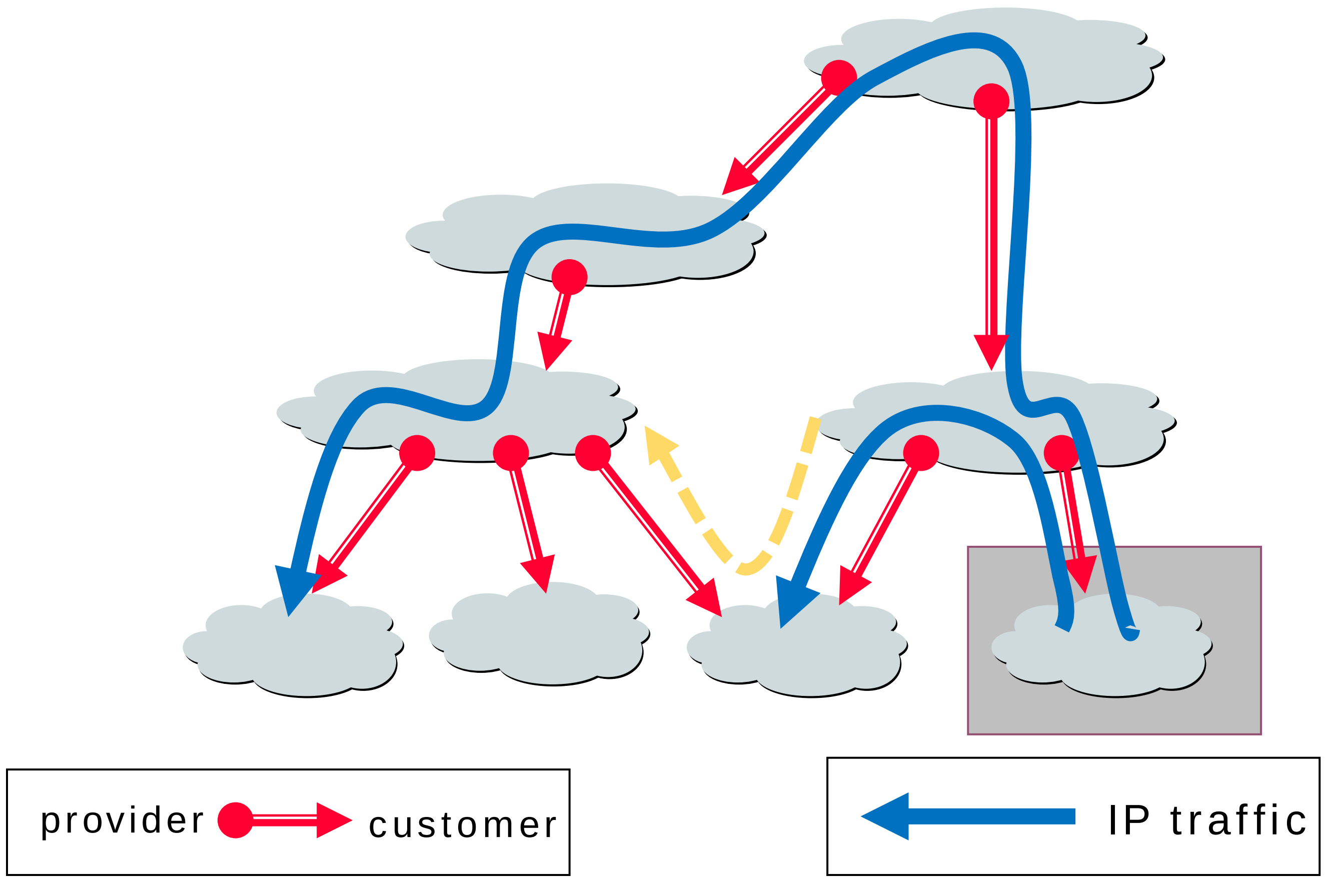


From Wikipedia: ISP

Customer Provider Hierarchy

Traffic can start from customer network, flow to provider networks, and destine in customer networks, e.g. blue flows.

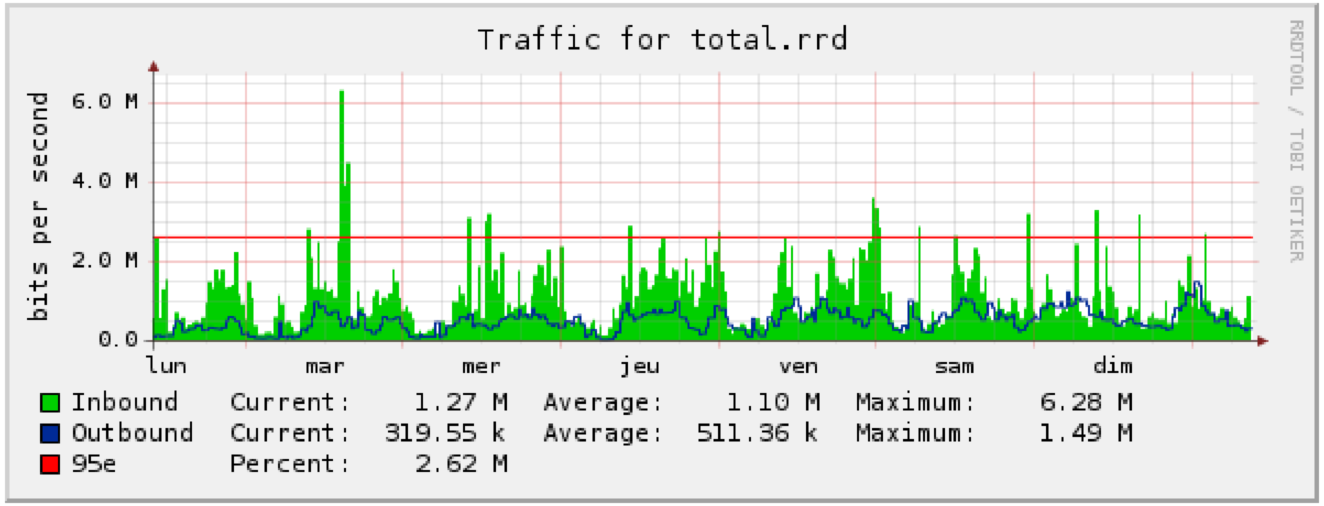
But traffic cannot start from provider network, through a customer network, end in another provider network, e.g. yellow flow.



Financial settlement for customer-provider peering

- Negotiated Financial Terms
  - By volume
  - Packet based? Too fine-grained, drop, detour, difficult to implement
  - Flow based? Require complicated instrumentation
  - By other factors, unsymmetrical rates
  - Terms unrelated to costs may not be stable

- 95-Percentile monthly billing
  - Provider ISP measures traffic at 5-min intervals
  - On a monthly basis, consumer ISP pays at the 95-percentile value
  - Rate is up to negotiations, and changes over time



Examples of Local, Regional and Tier 1 ISPs

- Local ISP who needs to connection to Internet
  - E.g. Consumer ISPs in Hong Kong – PCCW, HKBB, iCable, Hutchison, SmartTone…
  - E.g. CUHK
- Regional ISPs need to connect to backbone ISPs
  - Backbone ISPs include AT&T, Sprint, Level3, Telefonica

- But how do backbone (Tier 1) ISPs peer?
  - Nobody wants to be customer!

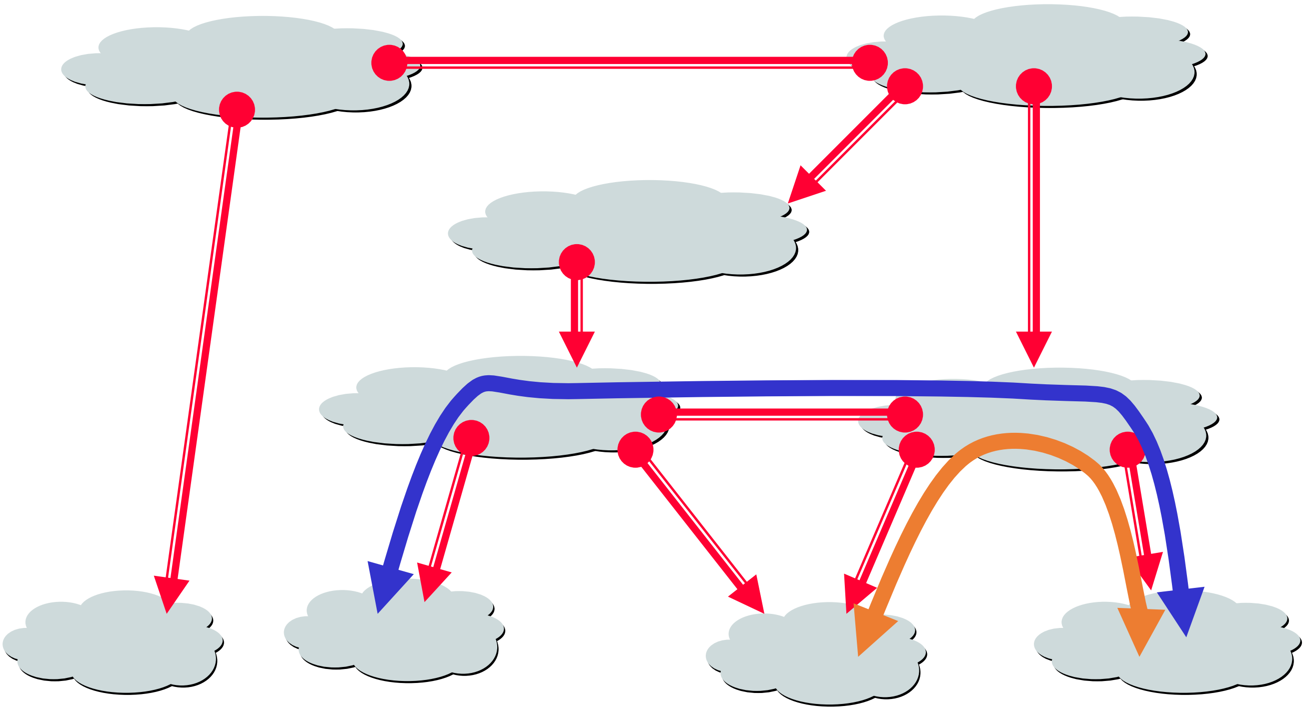
Equal peer relationship

Two ISPs may want to peer with each other, but neither is willing to be "customer"

They can agree to be "**equal peers**", connecting without paying each other; known as "**settlement free peering**"

This kind of peering needed for Tier 1 ISPs, and some regional ISPs of equal size.

- Provides shortcuts



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

peer ●═● peer
provider ●═▶ customer

Example equal peering requirement, from Verizon

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**.

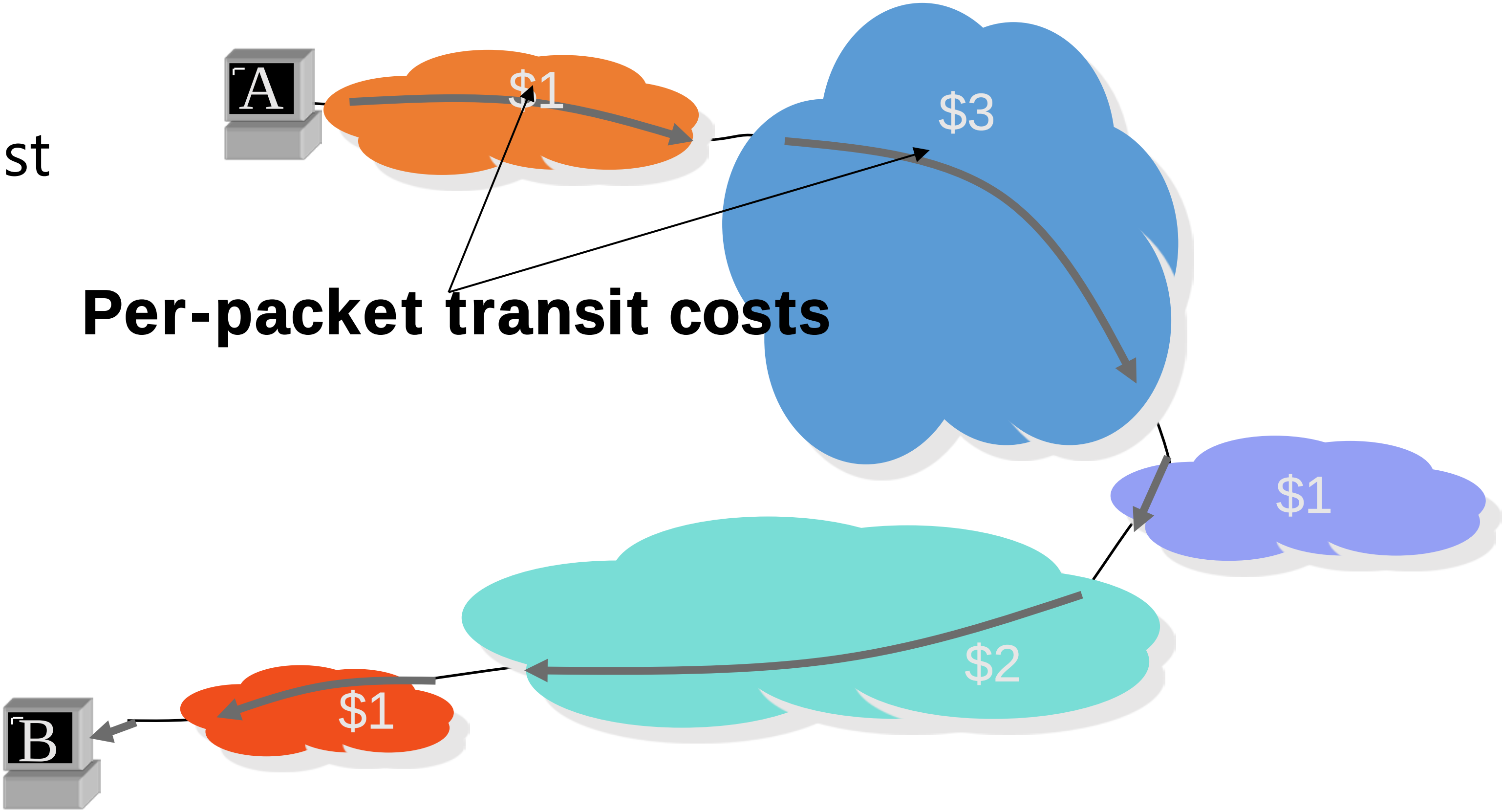
1.4 Traffic Volume. The **aggregate amount of traffic exchanged** in each direction over all interconnection links between the Requester and the Verizon Business Internet Network with which it desires to interconnect shall **equal or exceed 1500 Mbps** of traffic for Verizon Business-US, **150 Mbps**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

The costs of traffic through the network

Suppose the traffic transits through multiple ISP network, incurring some cost in each network

How are these costs paid?

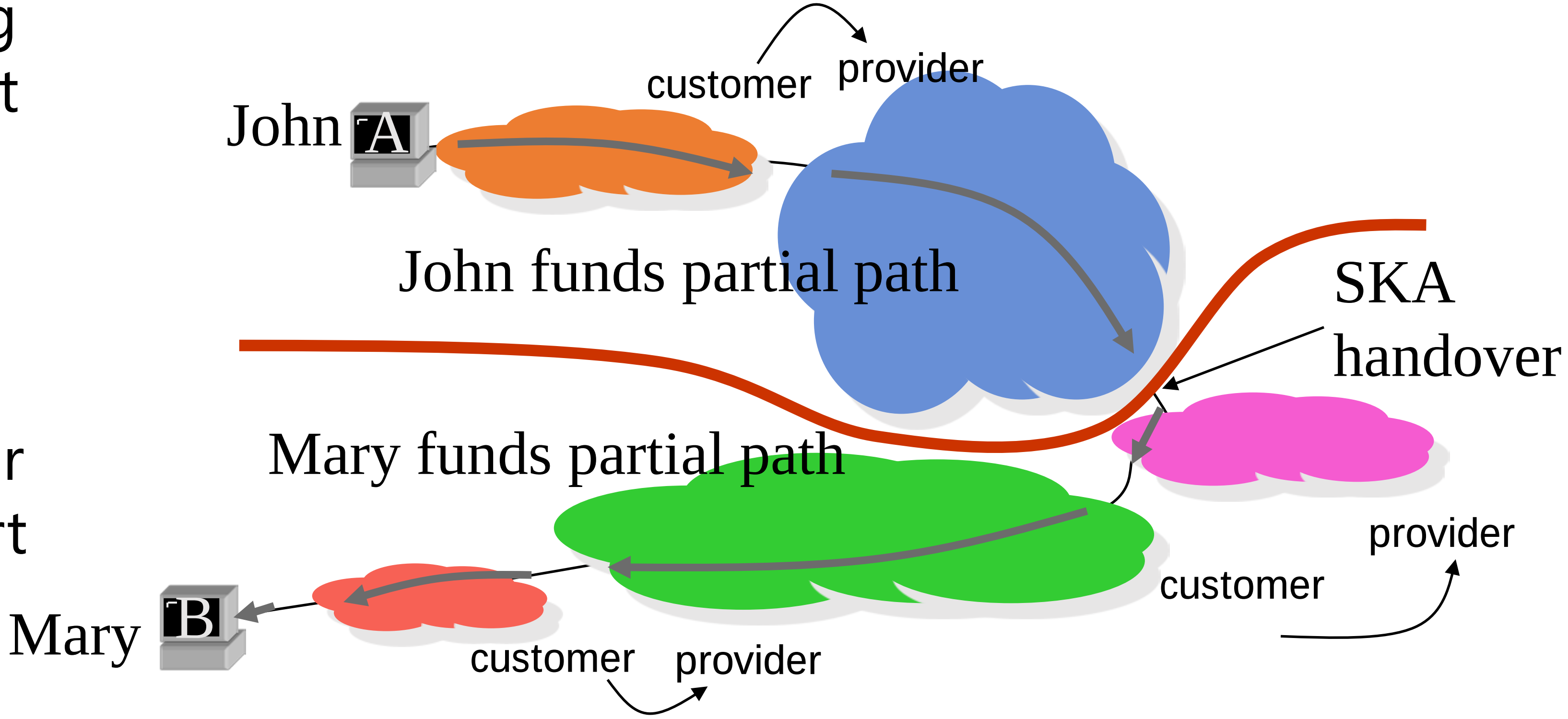


How transit service is paid

It depends on the peering relationships of the transit ISPs

Suppose there is a equal peer peering relationship in the middle, then sender and receiver each pay part of cost

What about other cases?



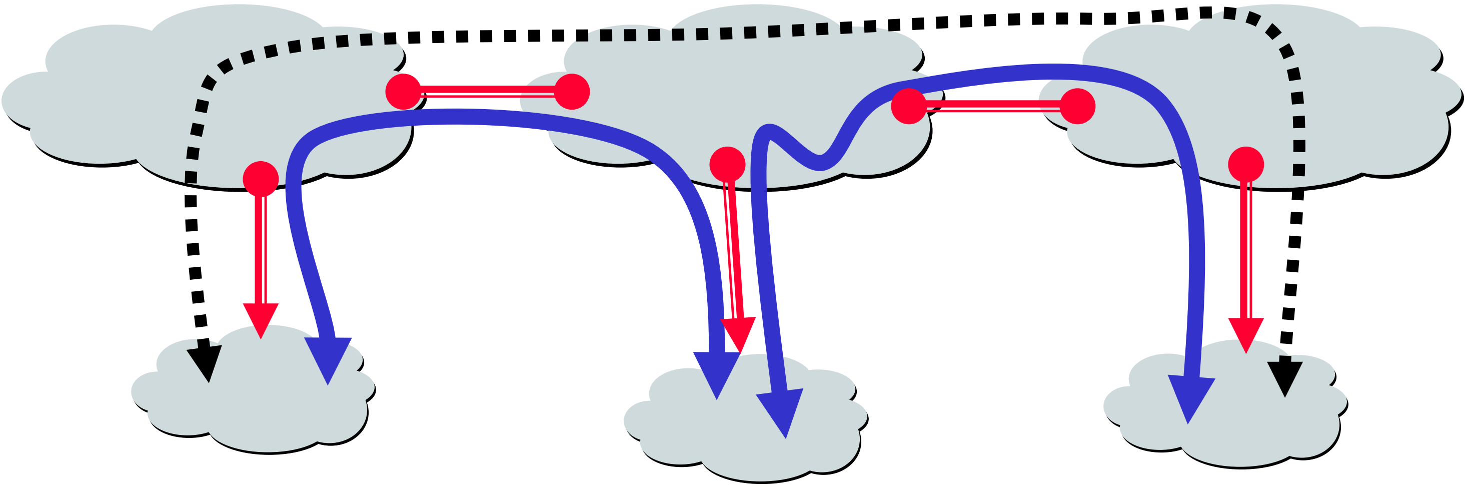
Contrast with telephone networks:

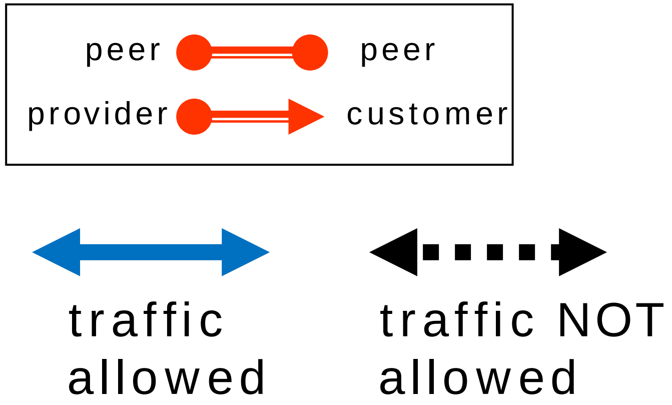
- Internet's peering and settlement model is based on (a) monthly aggregate customer payment to provider, and (b) no payment, or "Sender Keep All" (SKA)
  - It avoids peering ISPs having to keep track of each item of traffic

- In telephone networks, the initiator of a phone call pays for the call:
  - If it traverses multiple service providers, the initiator has to pay each
  - This requires a lot of book keeping

The transit service agreement for equal peering

- The source and destination of traffic can be either ISP, or its customer descendants
- But, the source of traffic cannot be from the **provider** or **peer** of either ISP





peer ●━━● peer

provider ●━━▶ customer

◀━━▶ traffic allowed

◀■■■▶ traffic NOT allowed

Peers provide transit between their respective customers

Peers do not provide transit between peers

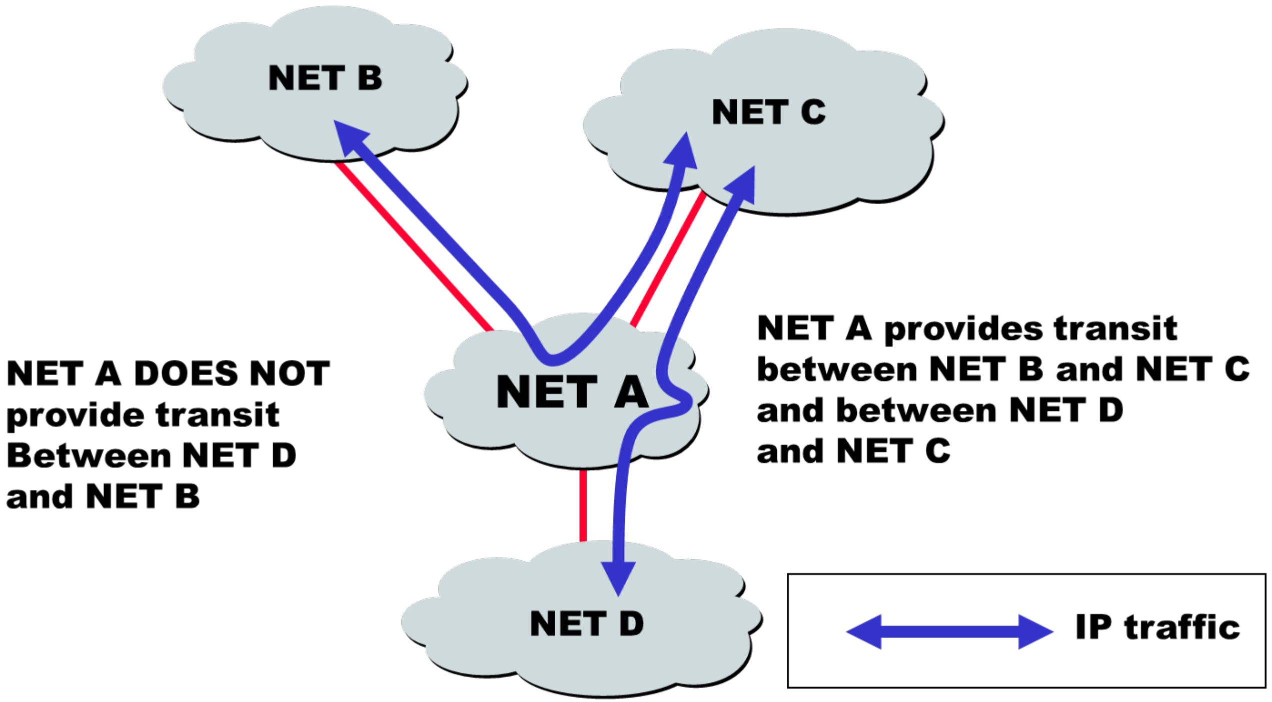
Peers (often) do not exchange $$$

Selective transit

- Perhaps the most significant property of inter-domain routing is **selective transit**

- Try give reasons for selective transit, based on peering relationships

- How will routing process implement 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...

Business implications for inter-domain routing

Customer ISPs:

- Each ISP tries to get traffic through least costly path
- Given two routes that both costs money (or both don't cost anything), then consider performance (minimize path length etc)

Provider ISPs:

- A backbone transit provider normally welcomes more traffic
- An ISP in the middle (has providers and customers) may not always welcome traffic – why?
- Two equal-basis peered ISPs try to get traffic to the other side as early as possible.

Summary of ISP peering

- AS
- Two types of peering
  - Customer-Provider (customer pays)
  - Equal-basis (no payment)
- ISP economy
- Implications to routing decisions

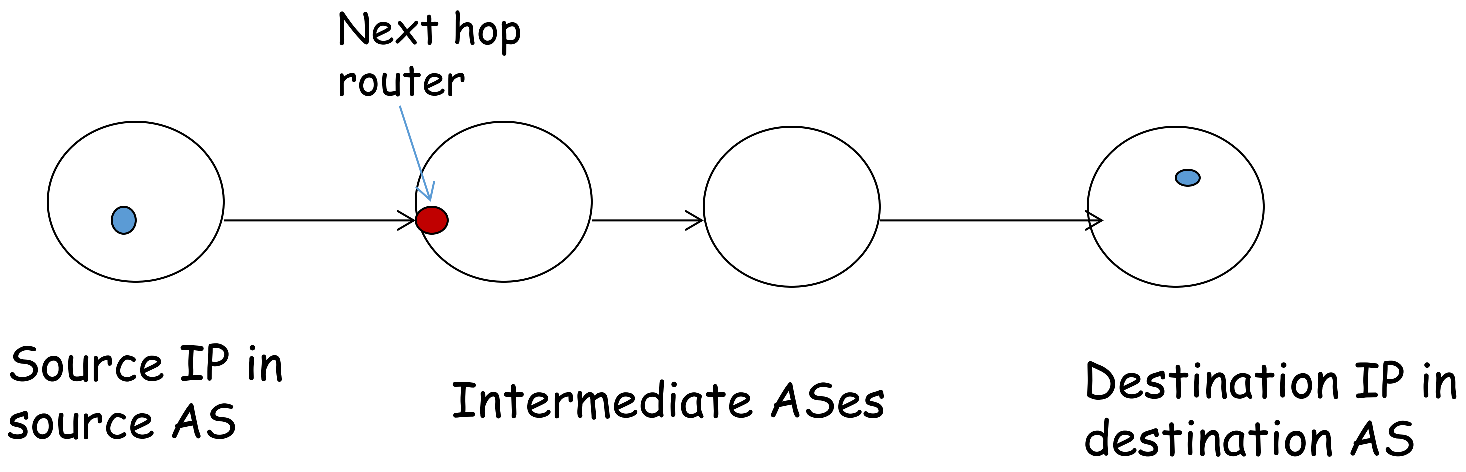
Inter-domain routing

The process to

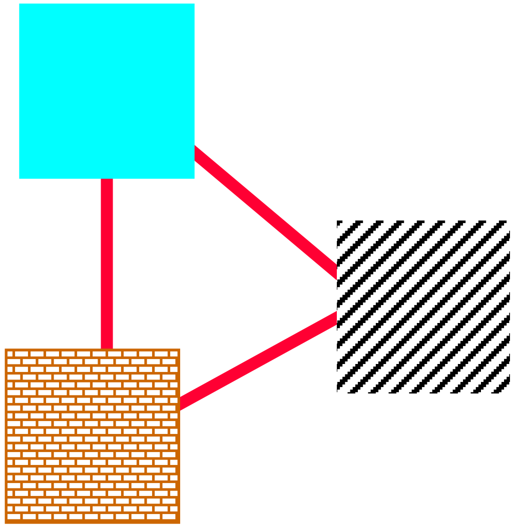
- Determine a AS Path for each destination IP prefix
- Determine the next hop router for the next AS

For internet, it is Border Gateway Protocol (BGP)

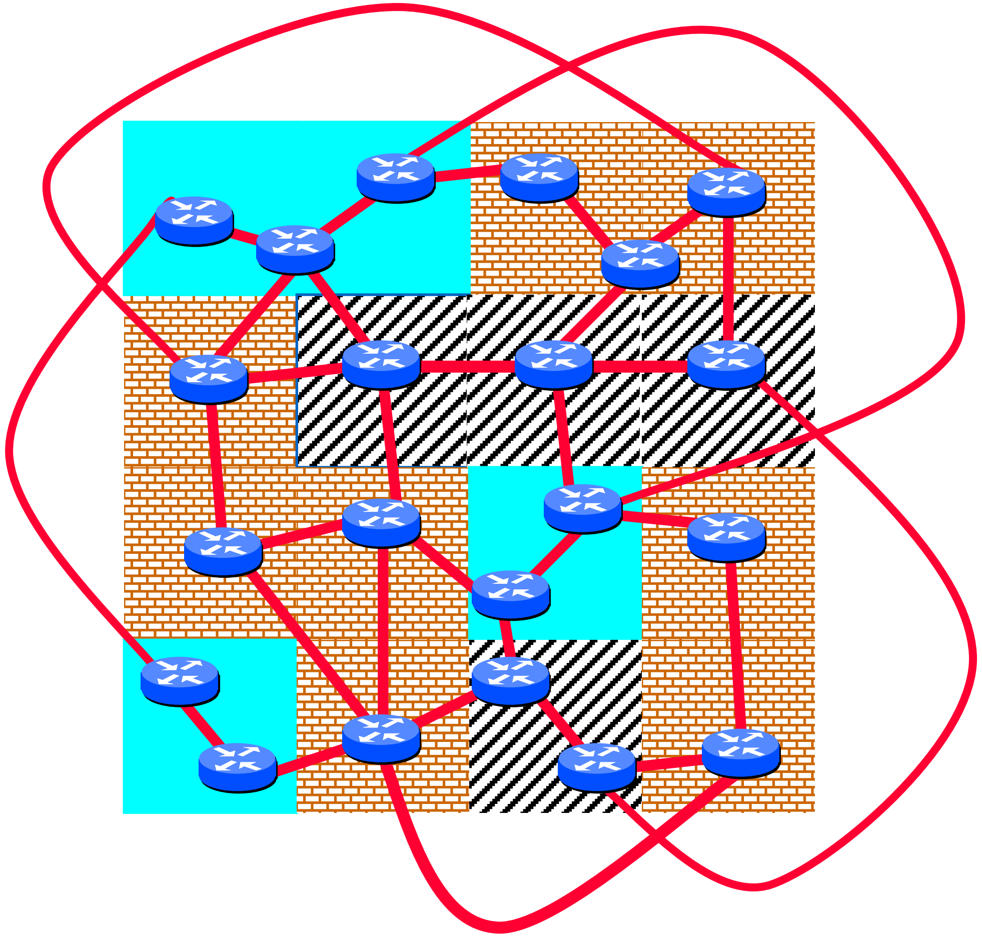
- Each AS may have multiple BGP routers, talking to different neighbor ASes



The AS graph



The AS graph
may look like this.



Reality may be closer to this…

BGP

Inter-domain Routing originally known as External Gateway Protocol (EGP)

- Router knows the domain topology of Internet, and every domain is a blackbox for EGP

- protocol (BGP) Border Gateway
  - BGP-1 developed in 1989 to address problems of EGP (RFC1105)
  - Current version BGP-4 (RFC1771), out in 1995
  - Views Internet as an arbitrarily interconnected set of ASs
  - Relatively simple protocol
    - OSPF RFC Size = 244 Pages (rfc2328, Apr. 1998)
    - BGP RFC Size = 57 Pages (rfc1771, Mar. 1995)
  - but configuration is complex, mistakes can impact entire Internet

Goals of BGP

Find loop-free paths that

- Support routing policy established as part of peering relationship
- Support traffic engineering to minimize (monetary) cost
- Optimizing performance is only another goal (not the only goal as in Intra-domain routing)
- BGP is known as a “policy-based” routing protocol

How BGP distributes routing information

It uses TCP as reliable transport

It uses Distance Vector algorithm

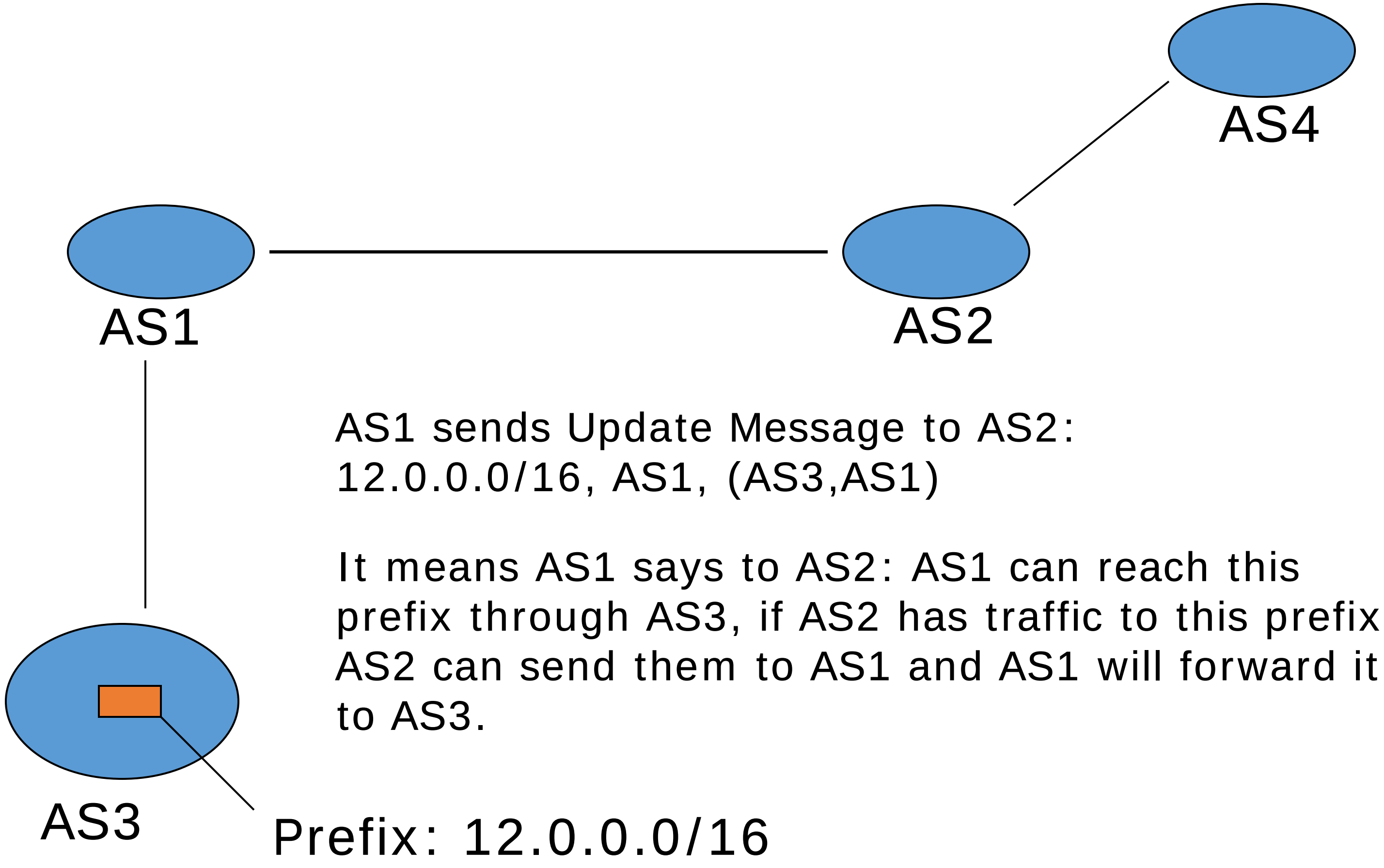
- BGP router advertises its best route to each neighbor (only if it will provide transit service for that neighbor)

- Advertisements are only sent when their routes change

Contrast with other routing protocols:

- RIP's distribution is based on UDP datagrams, so periodically you need to exchange information (a form of "soft state")
- BGP's distribution is based on a TCP connection - no need to refresh. But a consequence is that if the connection is broken, you have to assume everything is lost. more vulnerable to malicious (or natural) failures of connections
- OSPF uses its own transport to do in-sequence and intelligent flooding (which neither UDP or TCP supports).

Simple example BGP message



AS1 sends Update Message to AS2:
12.0.0.0/16, AS1, (AS3,AS1)

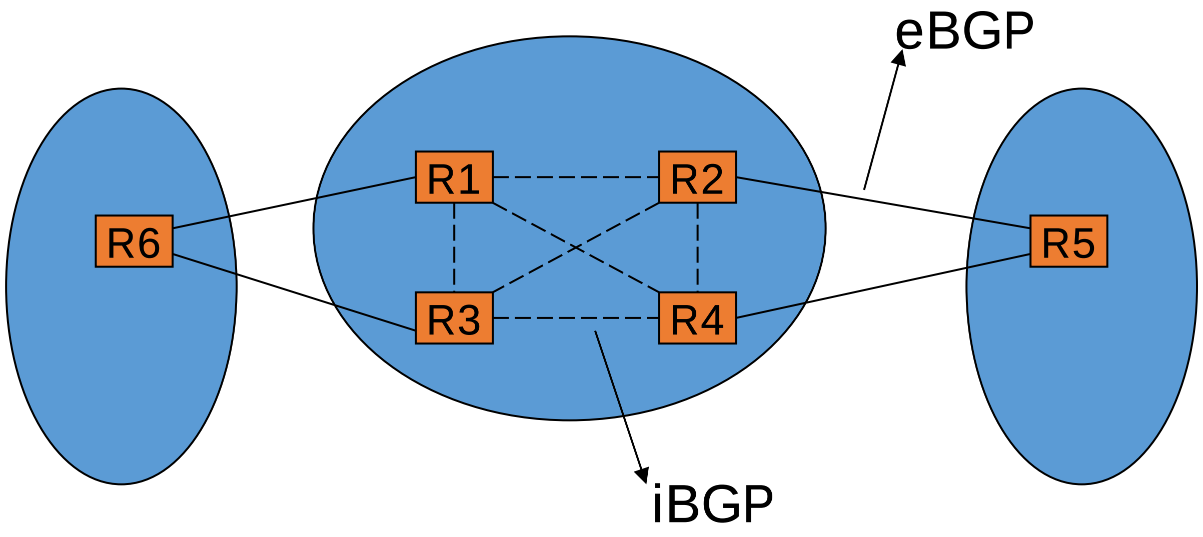
It means AS1 says to AS2: AS1 can reach this prefix through AS3, if AS2 has traffic to this prefix, AS2 can send them to AS1 and AS1 will forward it to AS3.

BGP Message format = [Prefix] [A list of attributes]

Same BGP protocol used for two purposes:

eBGP: between border routers of distinct AS

Used to distribute the inter-domain routes from one AS to another AS, according to their peering agreements.



iBGP: between BGP routers inside one AS

Used to distribute the inter-domain routes to all routers in the same AS, so they all have consistent view of external world

Rule: A route learned from an iBGP session cannot be forwarded to another iBGP router – to prevent loops (this problem does not exist for eBGP because of AS path info in each route)

Why is iBGP needed?

- Why not use IGP protocol to exchange external routes?
  - Too many external routes
  - If all IGP routers receive all these routes, it generates too much control traffic
  - Too much burden on all IGP routers

- Scalability of iBGP
  - For large ISPs, it is difficult to make BGP routers form a full mesh
  - Various solutions to avoid the problem, not covered in this course

BGP attributes

Major functions:

- Loop detection

  AS_Path

- Enforce transit agreement

  Local_Pref

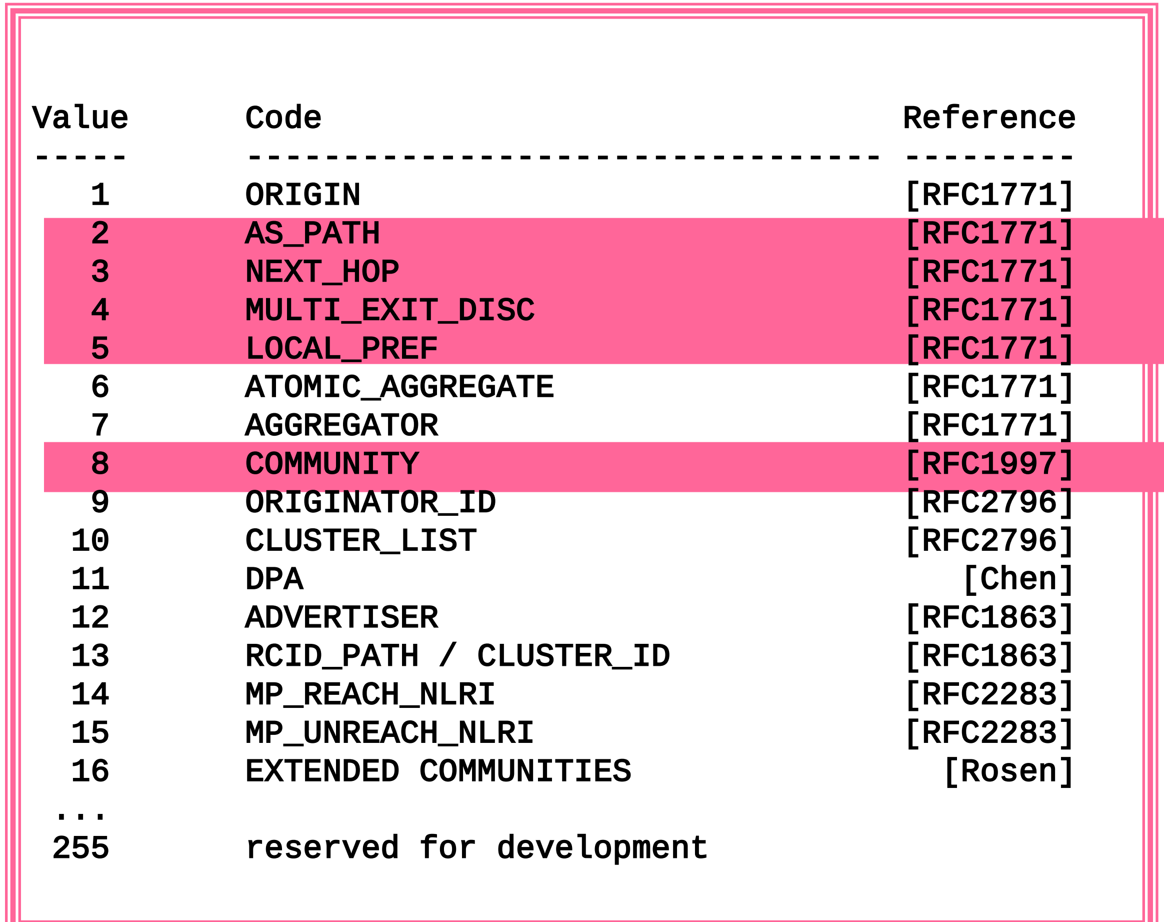
- Transfer load to peer

  MED

- Traffic engineering in multi-homing scenario

  AS_Path

  Community



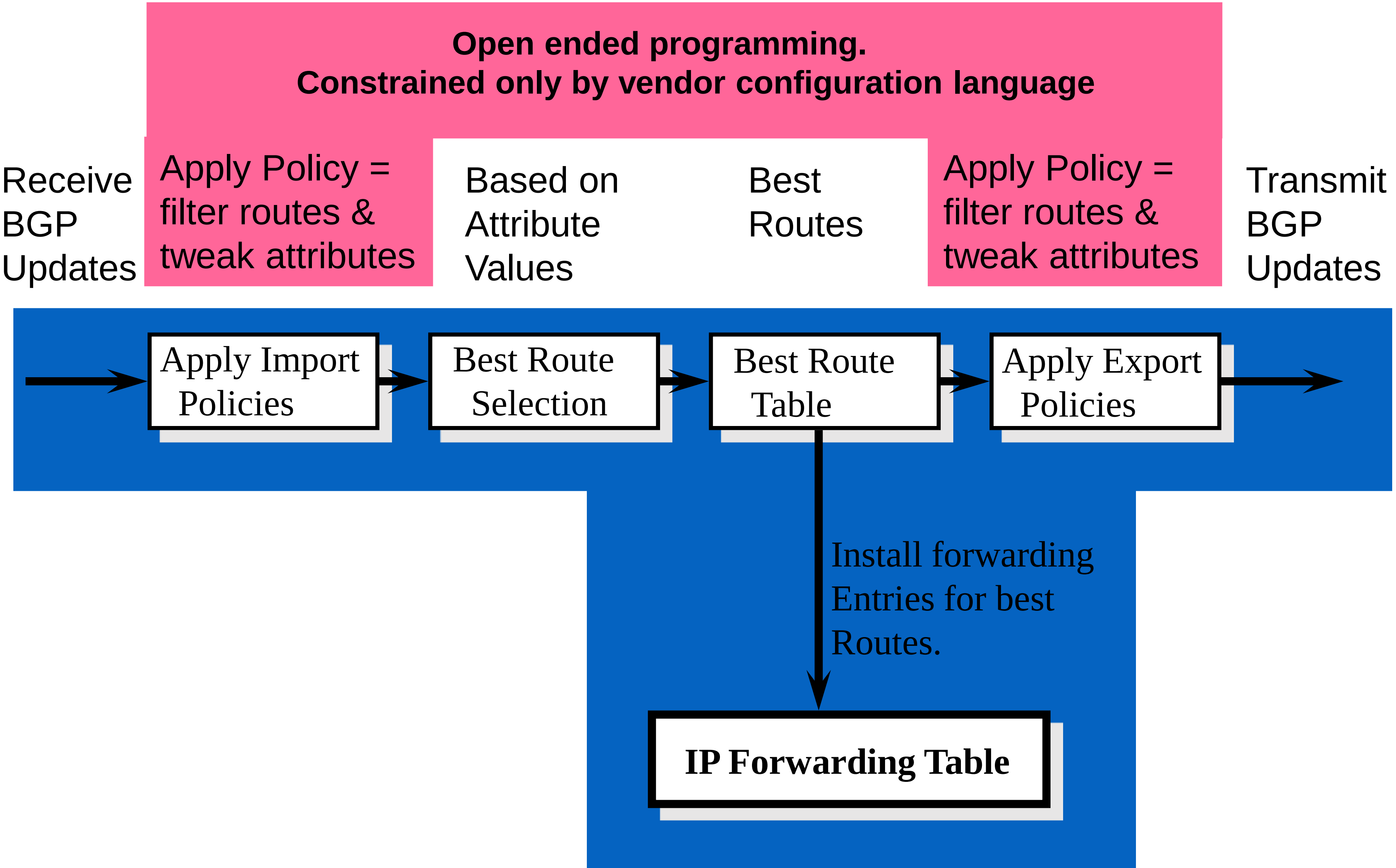
```
Value       Code                                Reference
-----       ----------------------------------  ---------
  1         ORIGIN                              [RFC1771]
  2         AS_PATH                             [RFC1771]
  3         NEXT_HOP                            [RFC1771]
  4         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

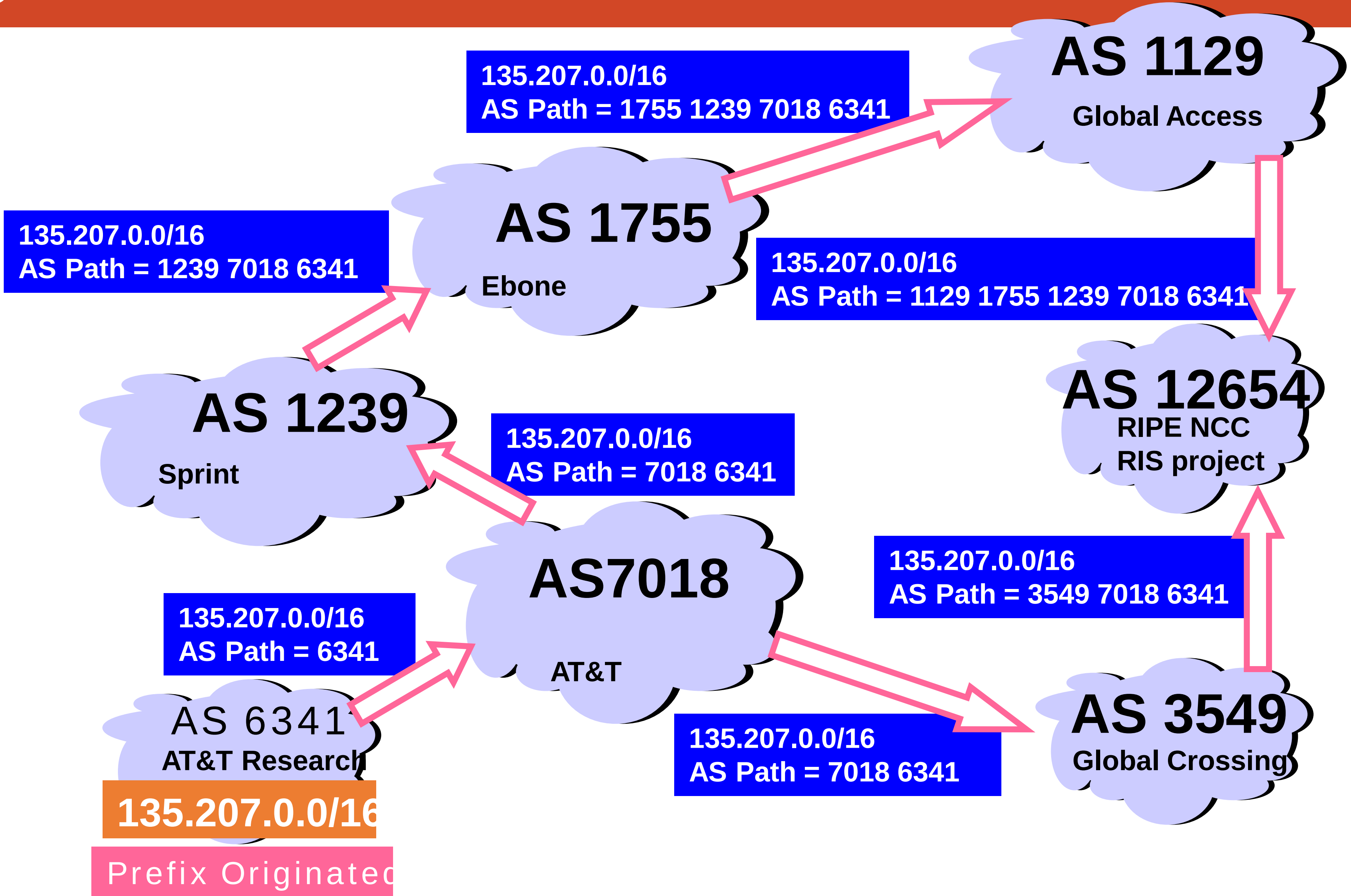
We only look at thesese attributes

Not all attributes need to be present in every announcement

Route processing by a BGP router

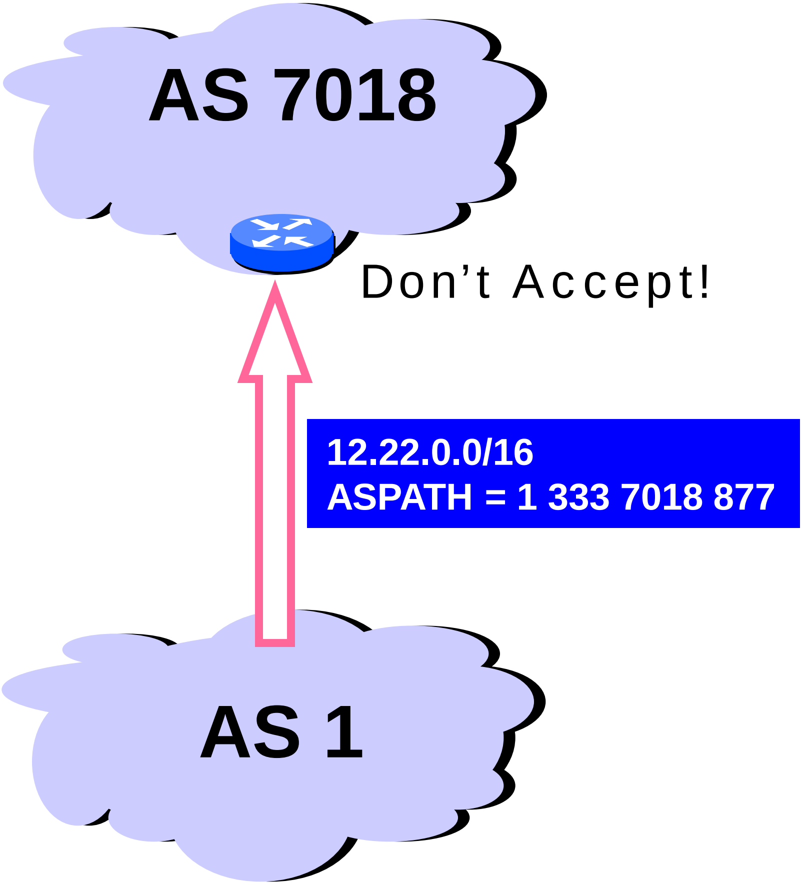

Open ended programming.
Constrained only by vendor configuration language
Receive BGP Updates
Apply Policy = filter routes & tweak attributes
Based on Attribute Values
Best Routes
Apply Policy = filter routes & tweak attributes
Transmit BGP Updates
Apply Import Policies
Best Route Selection
Best Route Table
Apply Export Policies
Install forwarding Entries for best Routes.
IP Forwarding Table

AS_Path usage

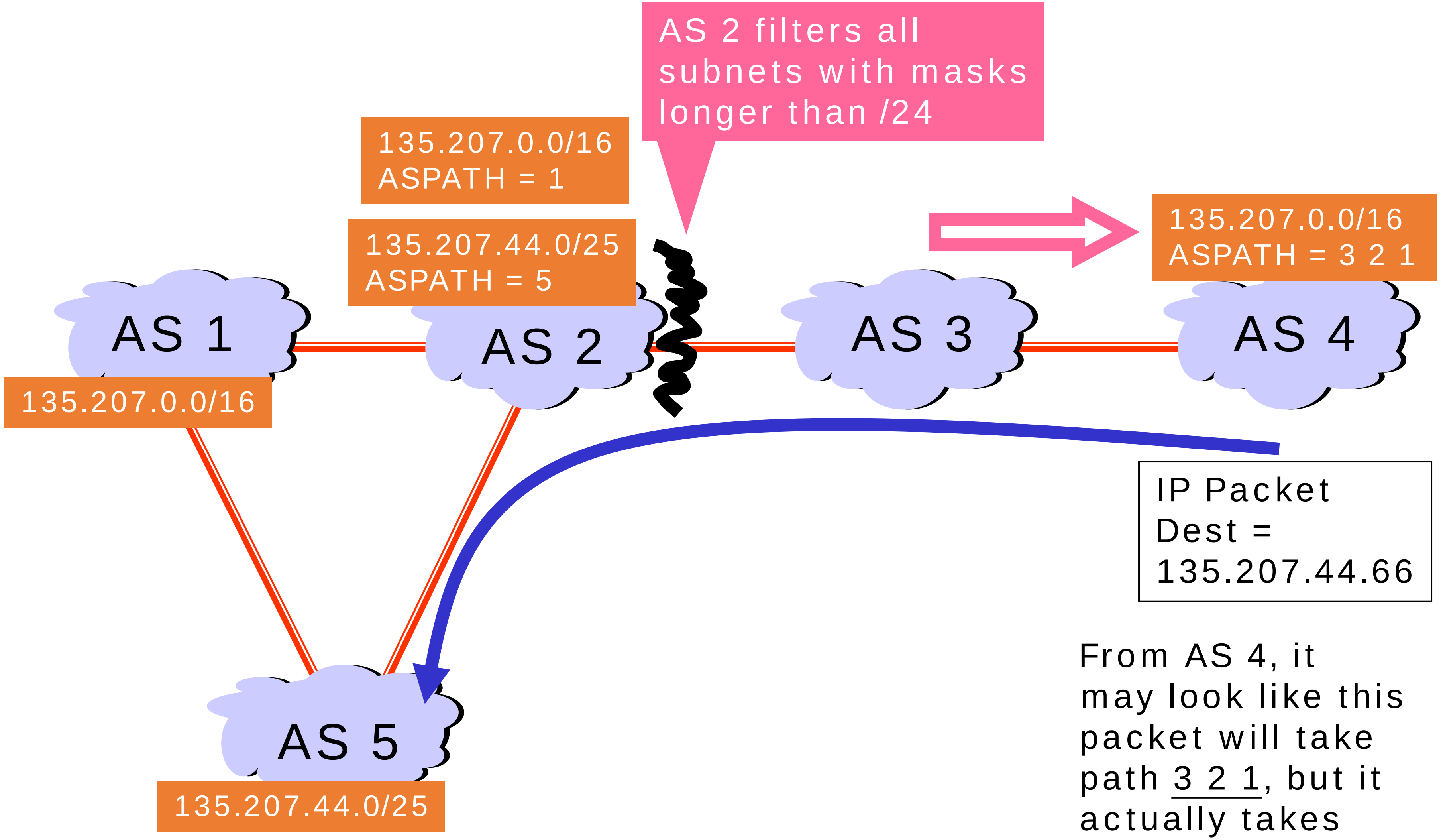


Inter-domain loop prevention

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



AS_Path may not be the real path

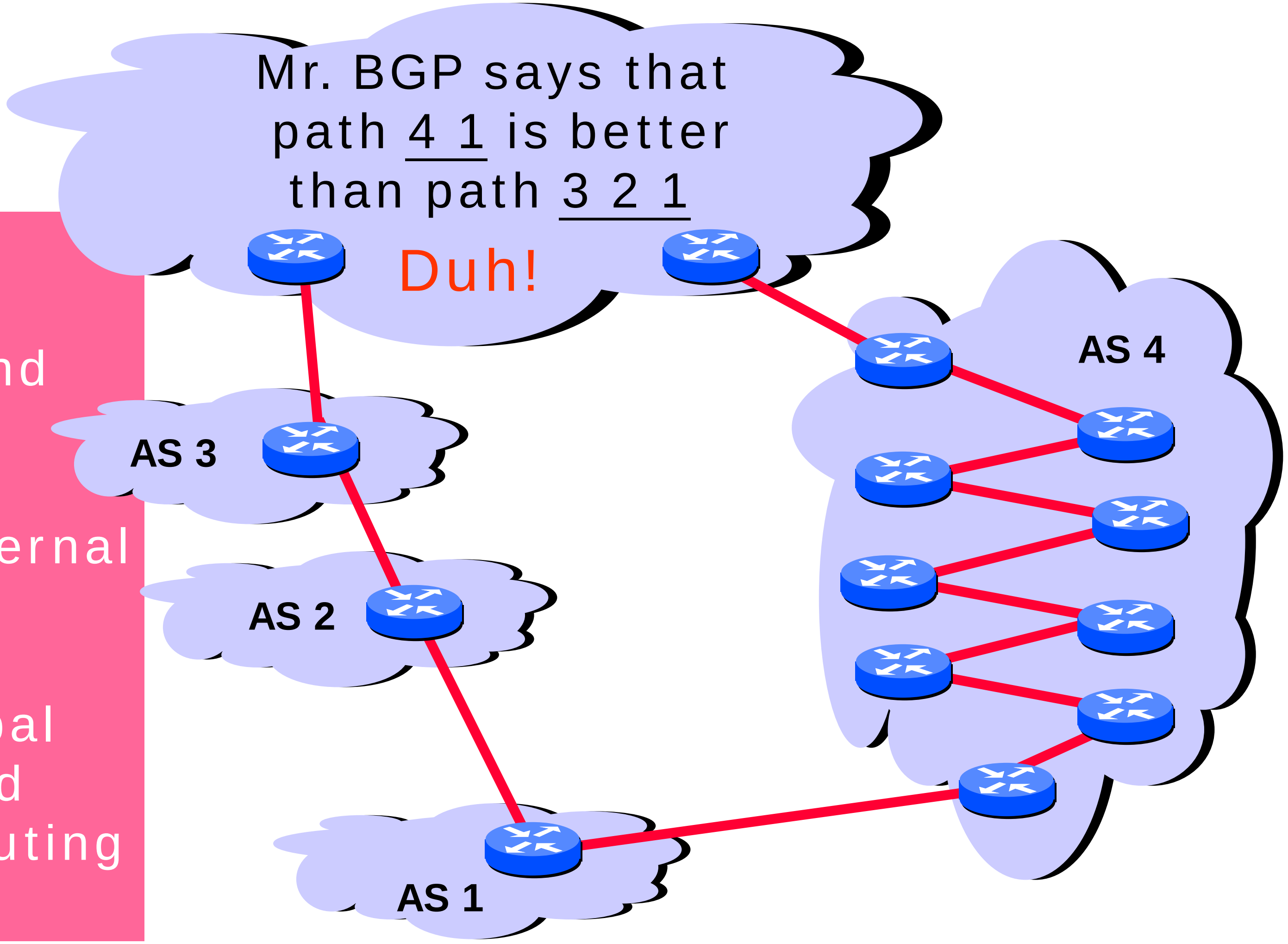


From AS 4, it may look like this packet will take path 3 2 1, but it actually takes path 3 2 5

AS_Path length may not indicate “distance”

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

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



Implementing transit policy

When sending route advertisement, do **outbound filtering**:

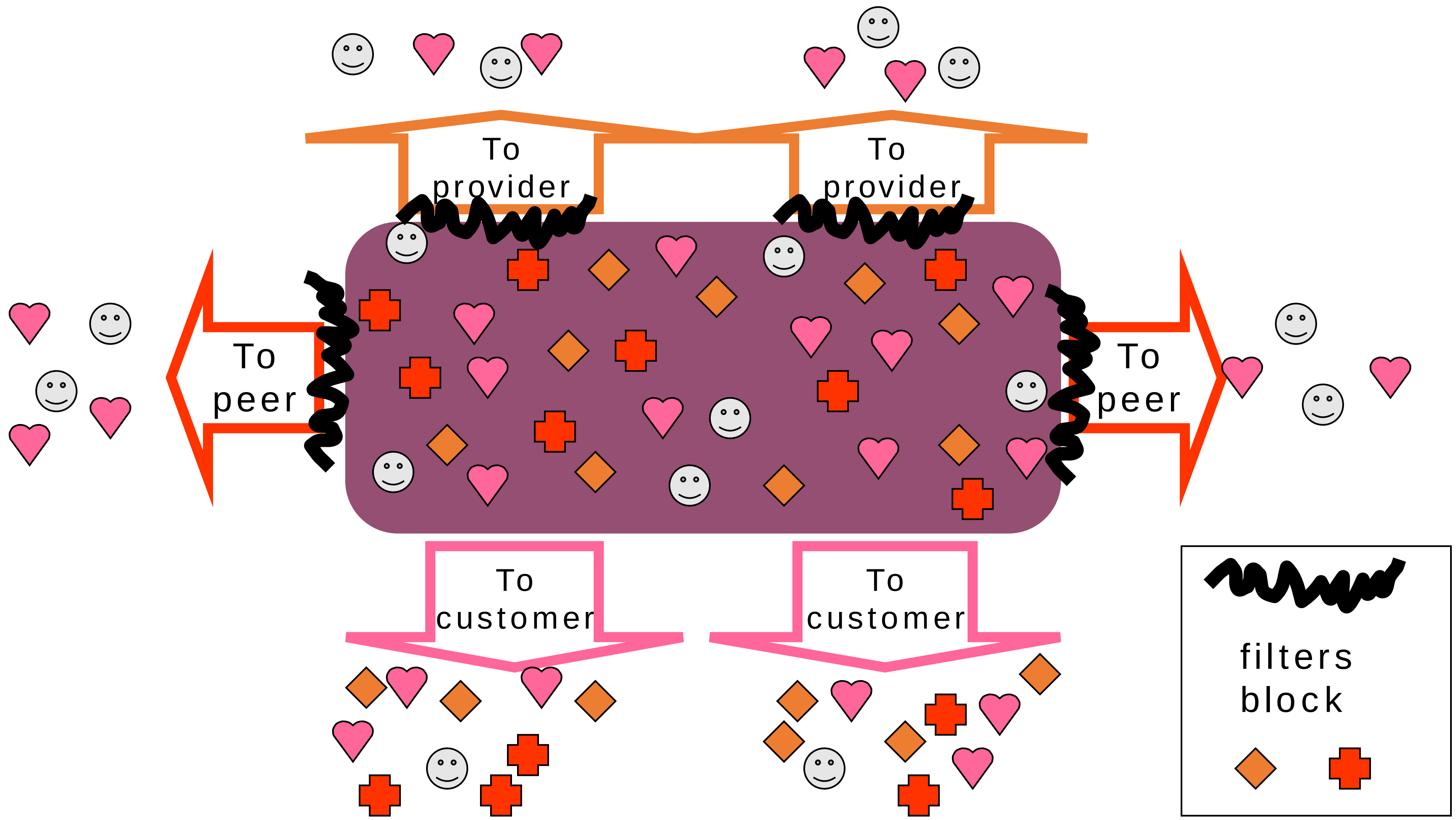
- Don’t advertise those routes if you don’t want to provide transit

When receiving route advertisements, set Local_Pref according to service type, during **inbound processing**

- If a **customer route** is available, it is always preferred since customer pays you for it
- Else if a (**free) peer route** is available, preferred that next, since it is free
- Use a **provider route** as last resort since it costs you

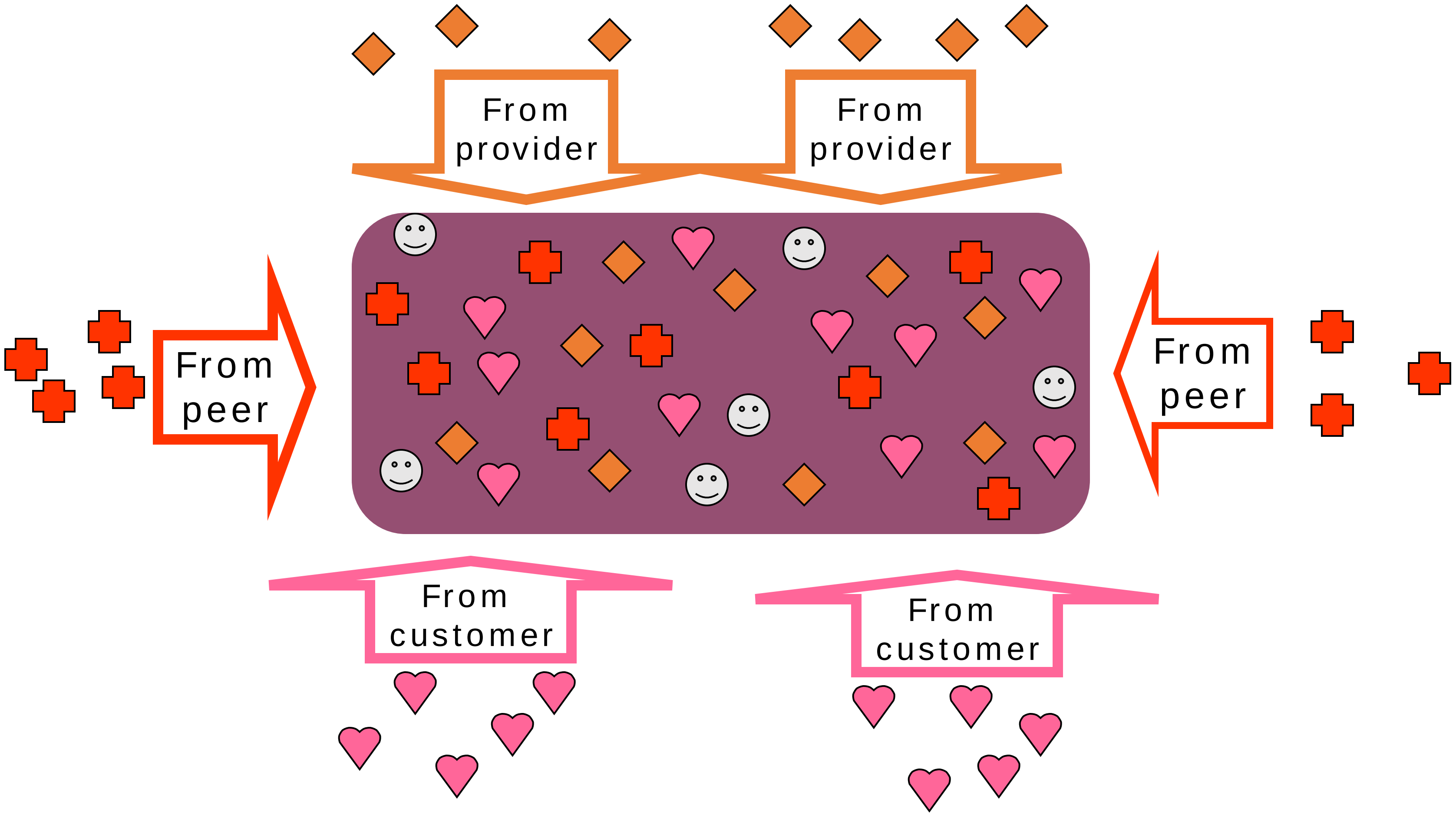
Outbound filtering illustrated

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

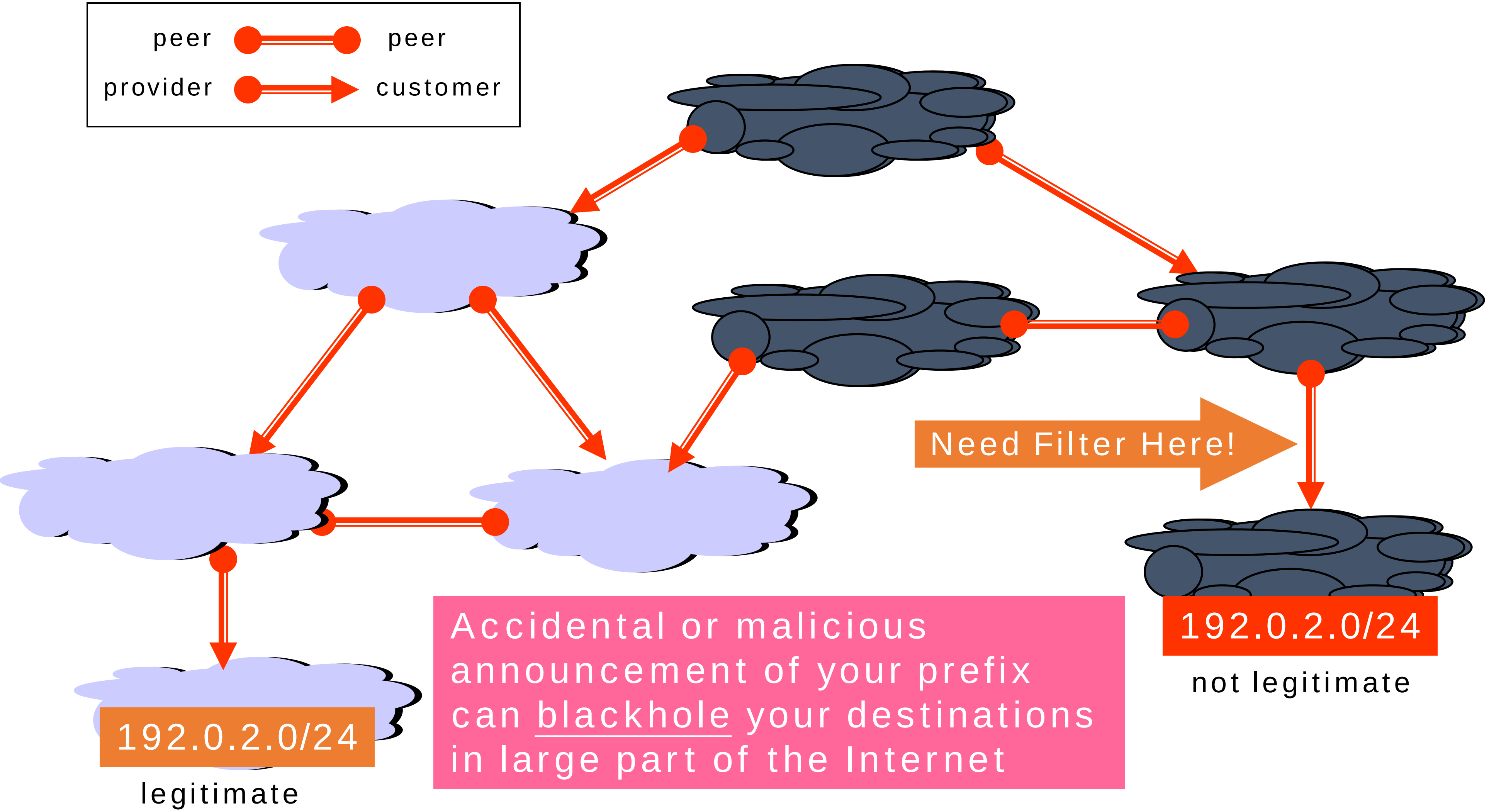


Inbound situation illustrated

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



Potential of blackholes

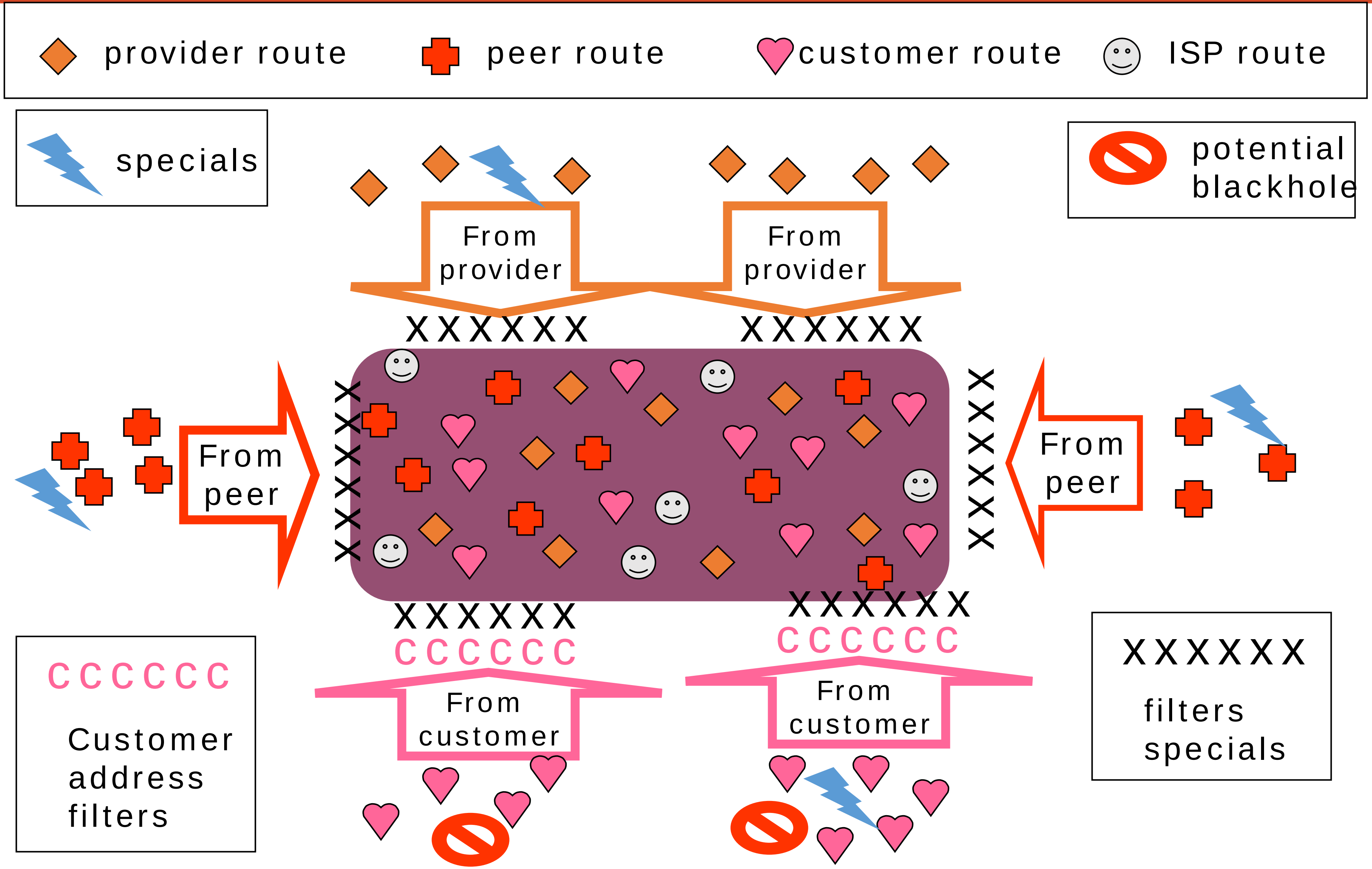


Accidental or malicious announcement of your prefix can blackhole your destinations in large part of the Internet

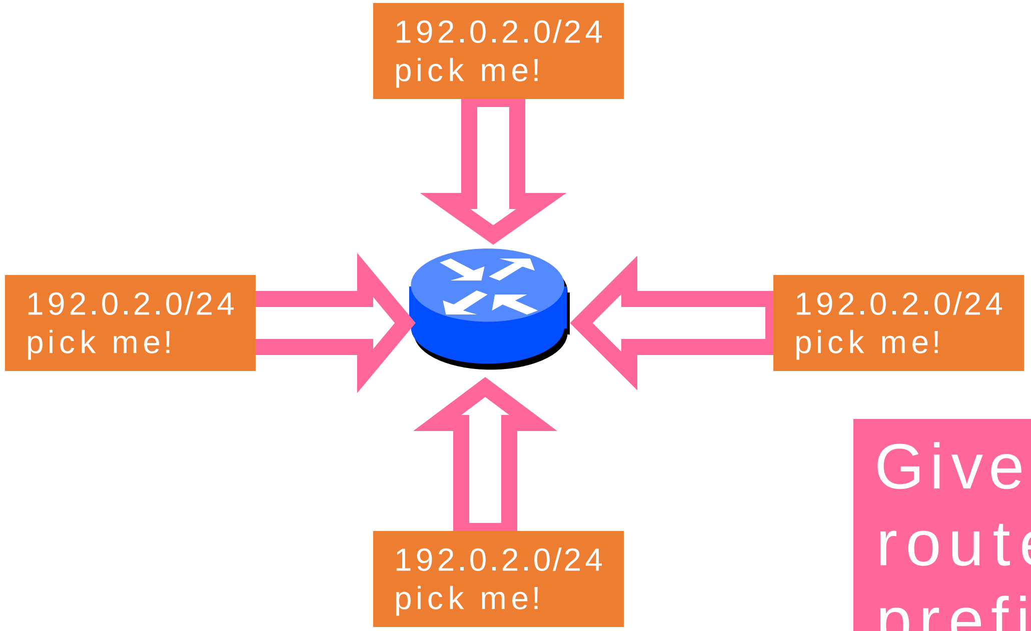
Addresses with special meanings

- 0.0.0.0/0: default
- 10.0.0.0/8: private
- 172.16.0.0/12: private
- 192.168.0.0/16: private
- 128.0.0.0/16: IANA reserved
- 192.0.2.0/24: test networks
- 224.0.0.0/3: classes D and E
- etc

Imported address get filtered, and preference indicated



How do you choose?



192.0.2.0/24
pick me!

192.0.2.0/24
pick me!

192.0.2.0/24
pick me!

192.0.2.0/24
pick me!

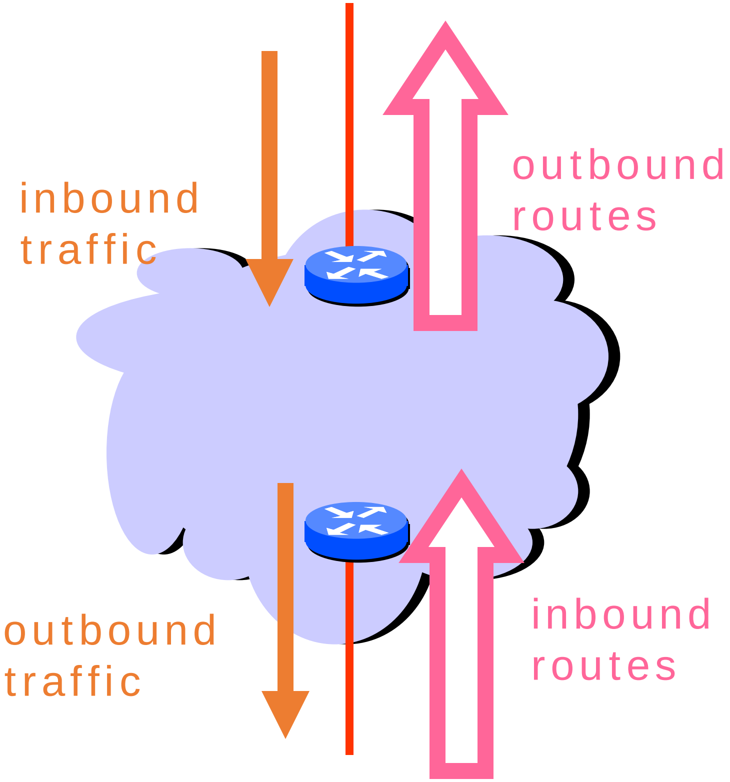
Given multiple routes to the same prefix, a BGP speaker must pick at most one best route

(Note: it could reject them all!)

The process of policy routing

- 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



Local_pref

A preference set depending on where route comes from

Propagated by iBGP to AS routers in same AS

Only used inside a single AS

AS_path

List of AS traversed by path

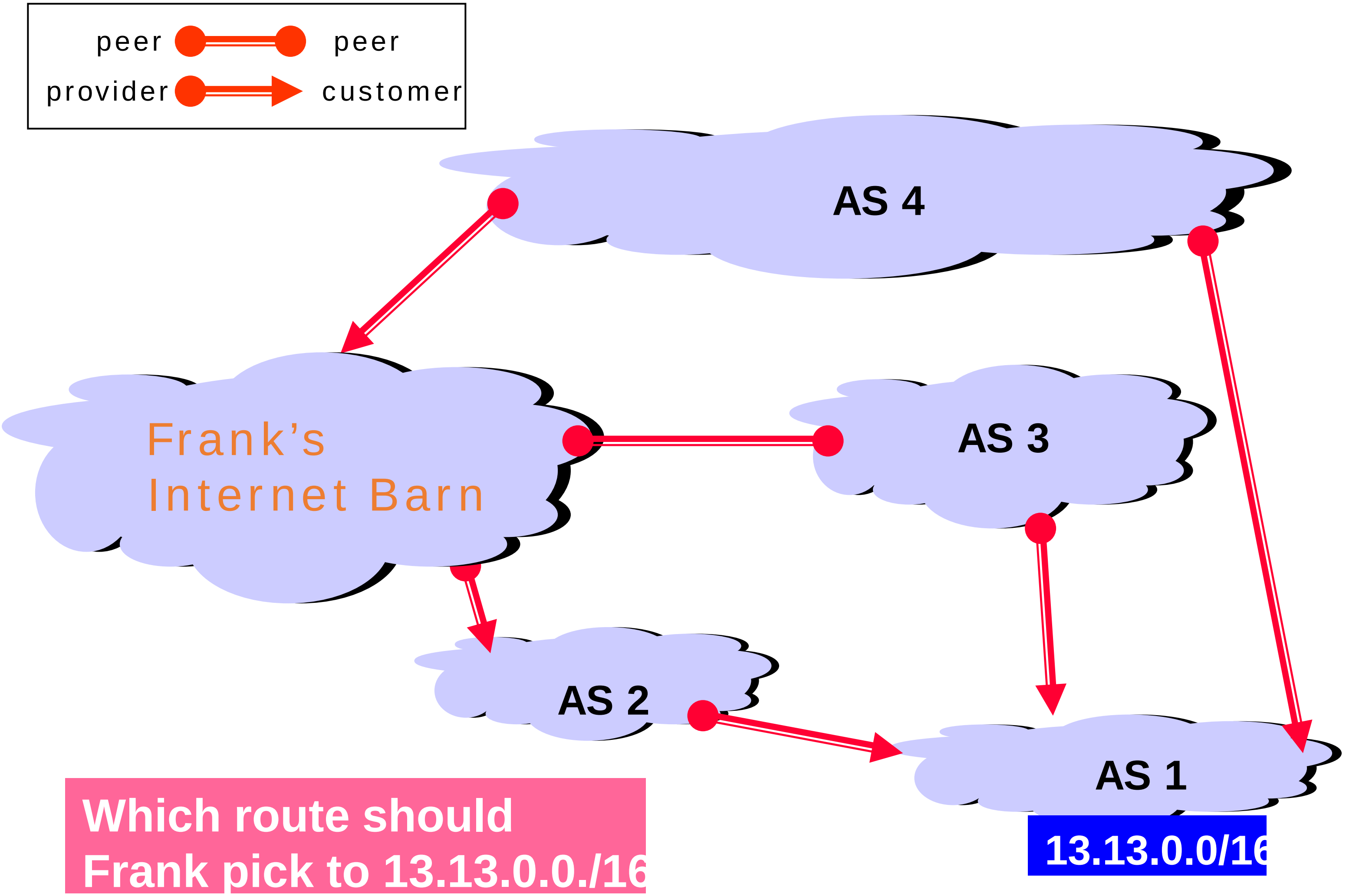
Community

An id used to tell neighbor AS how to set local pref

Multi_exit_discriminator (MED)

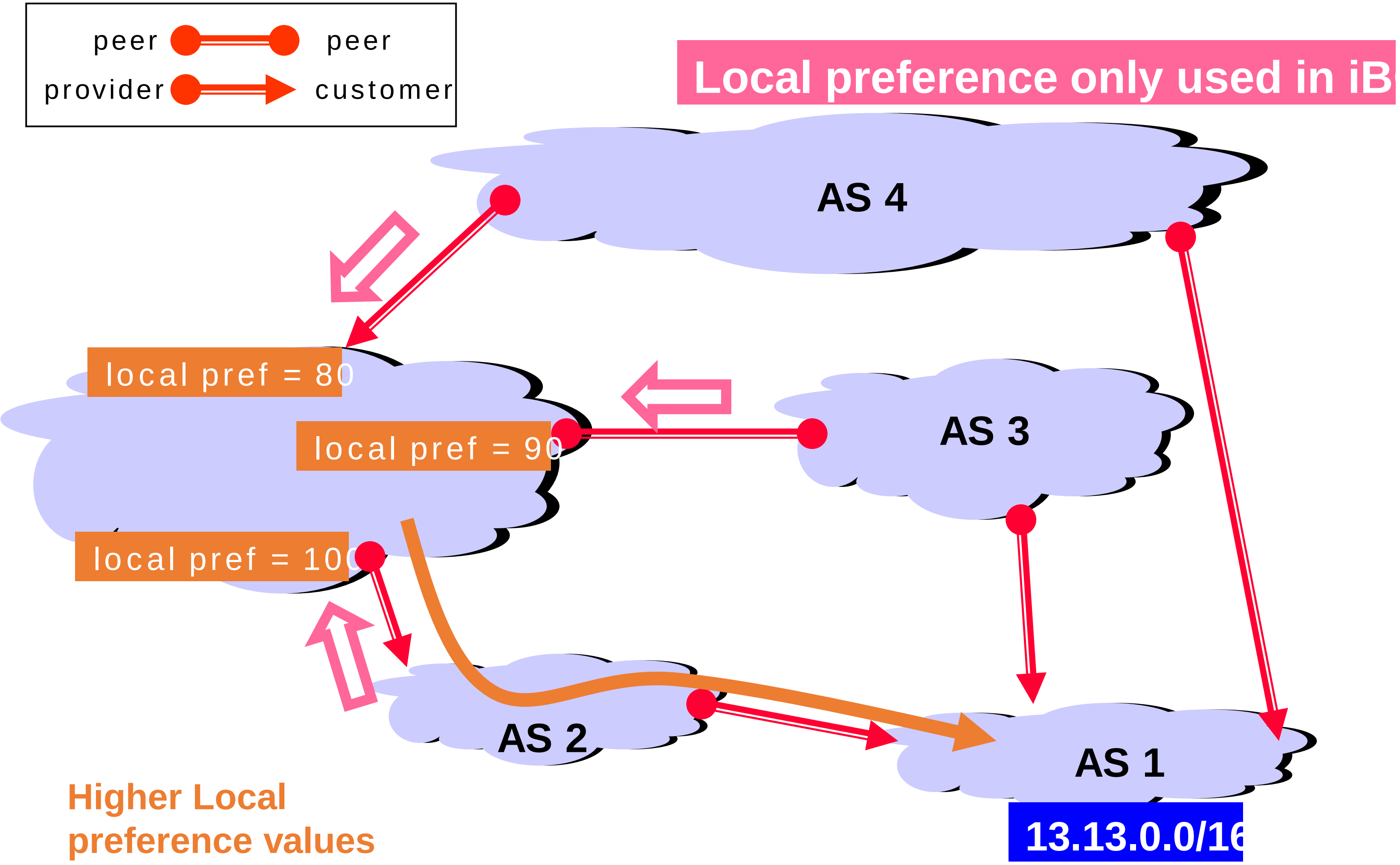
A cost indicator to discriminate between multiple exit paths from one AS to another

Multiple route choices



Which route should Frank pick to 13.13.0.0./16

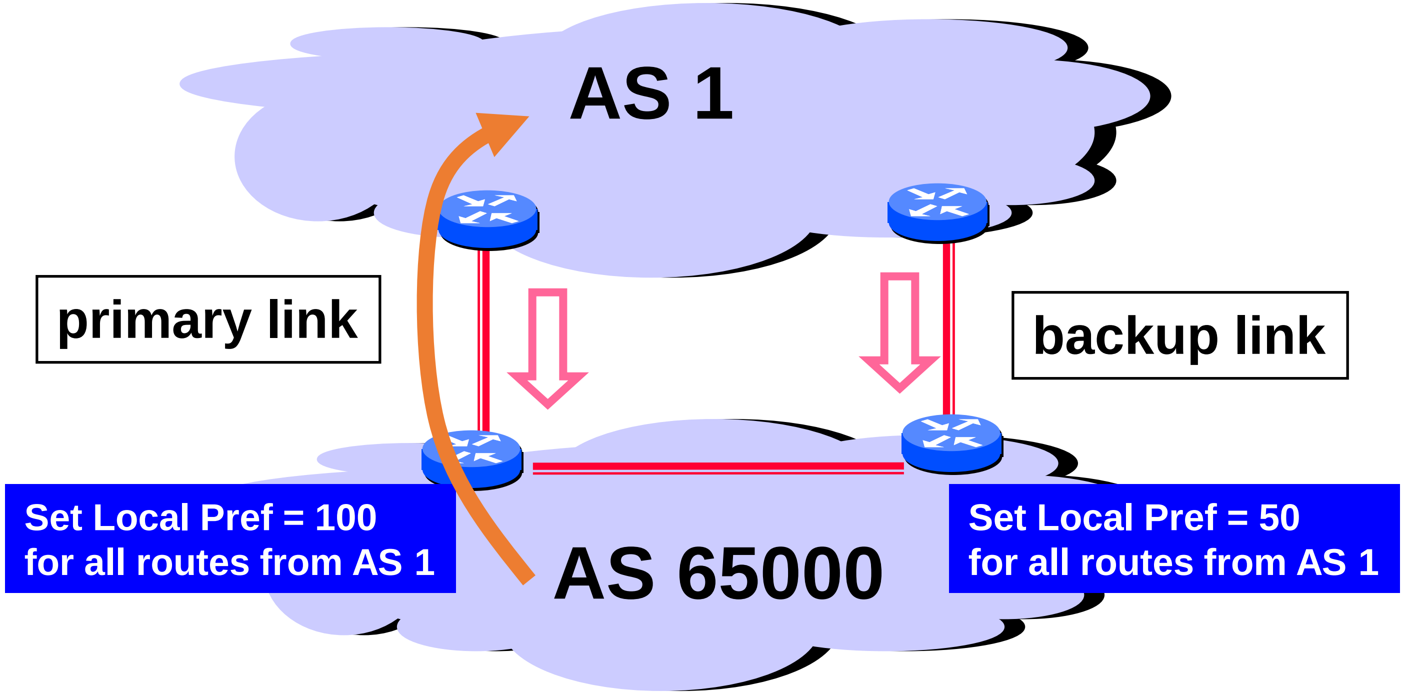
Use Local_Pref to select best route



This is done when processing inbound route advertisements

Higher Local preference values are more preferred

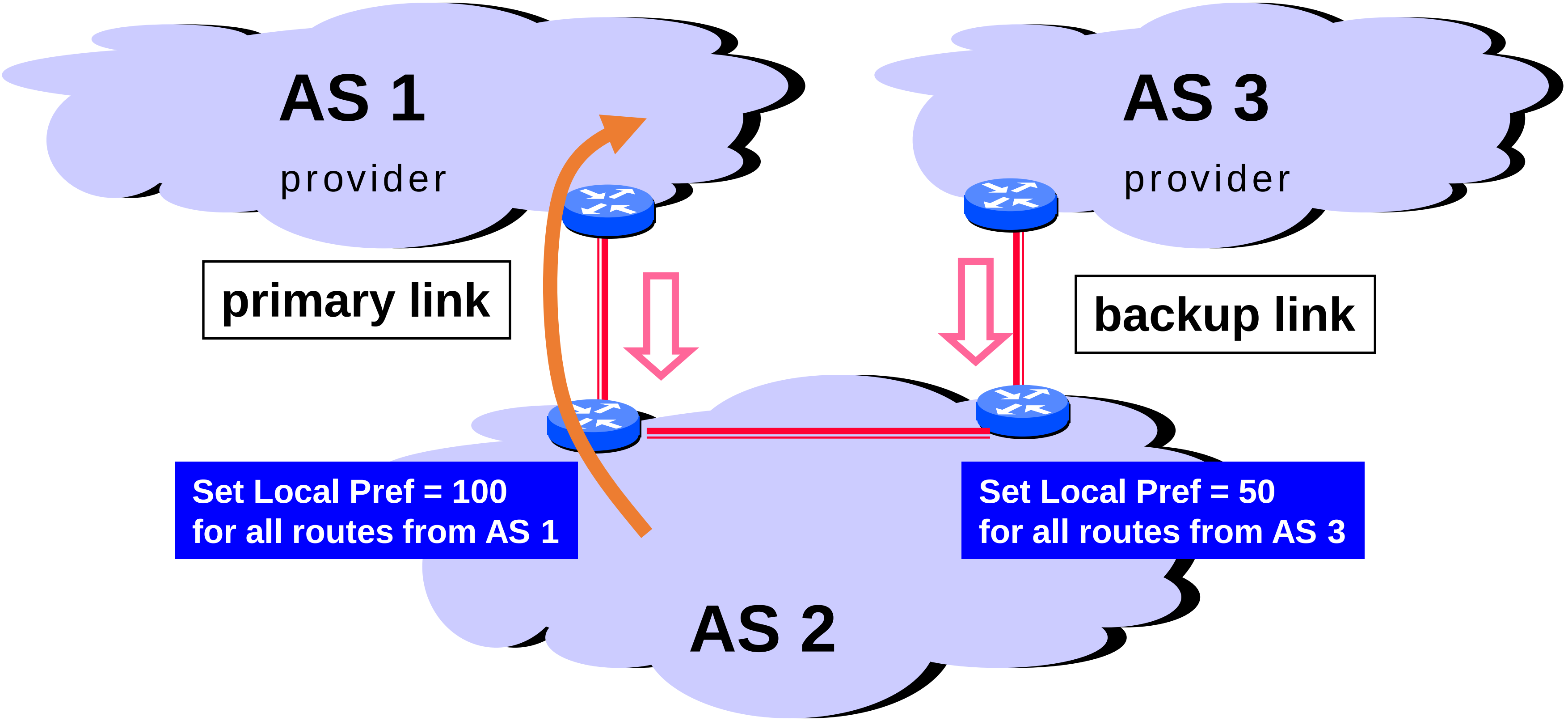
Implementing backup link using Local_Pref (for outbound traffic)



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

We’ll talk about inbound traffic soon …

Backup link in multi-homed case (for outbound traffic)

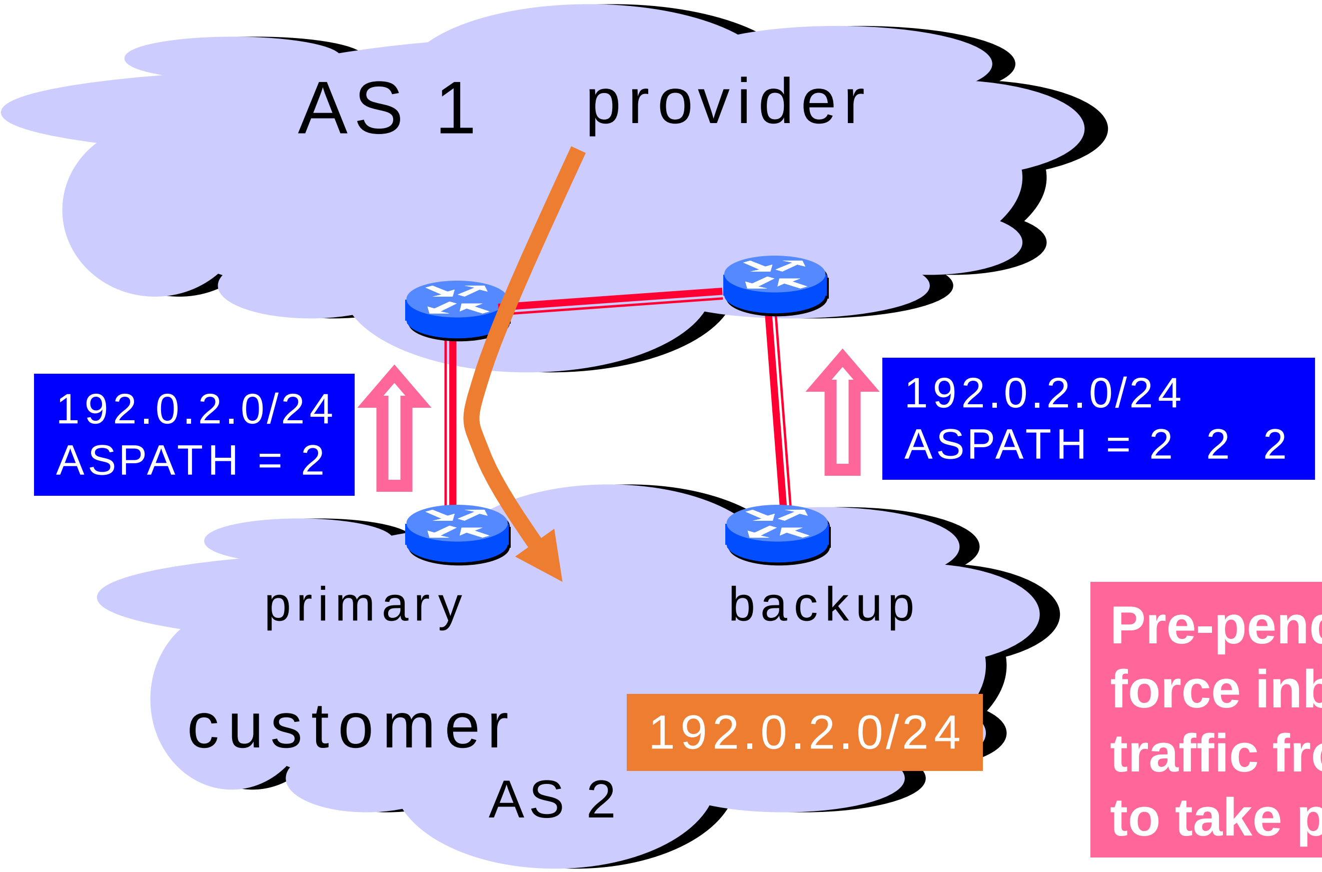


Set Local Pref = 100
for all routes from AS 1

Set Local Pref = 50
for all routes from AS 3

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

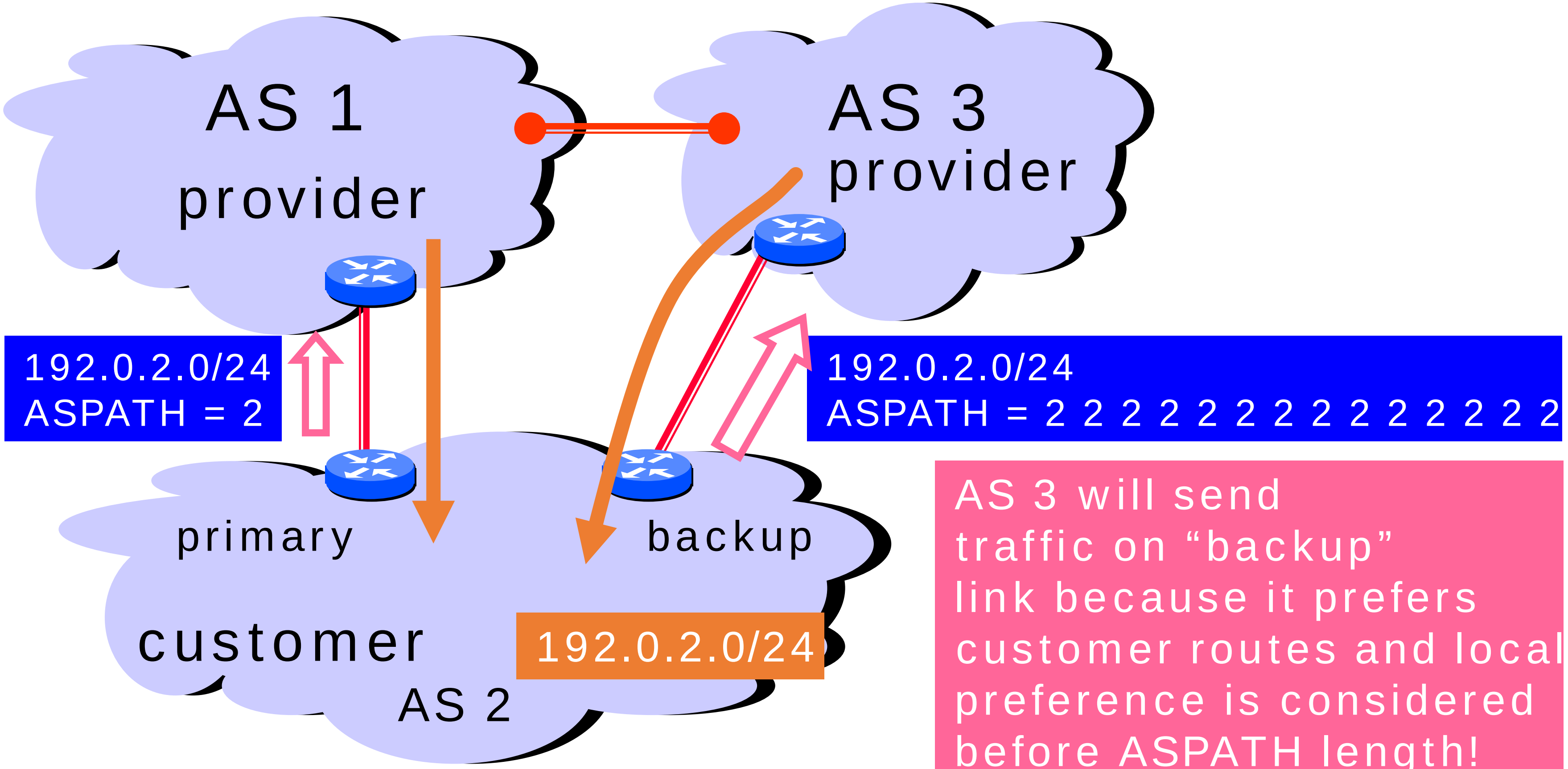
Implementing backlink using AS_Path (for inbound traffic)



This technique is referred to as "AS_Path Prepending"

Pre-pending will (usually) force inbound traffic from AS 1 to take primary link

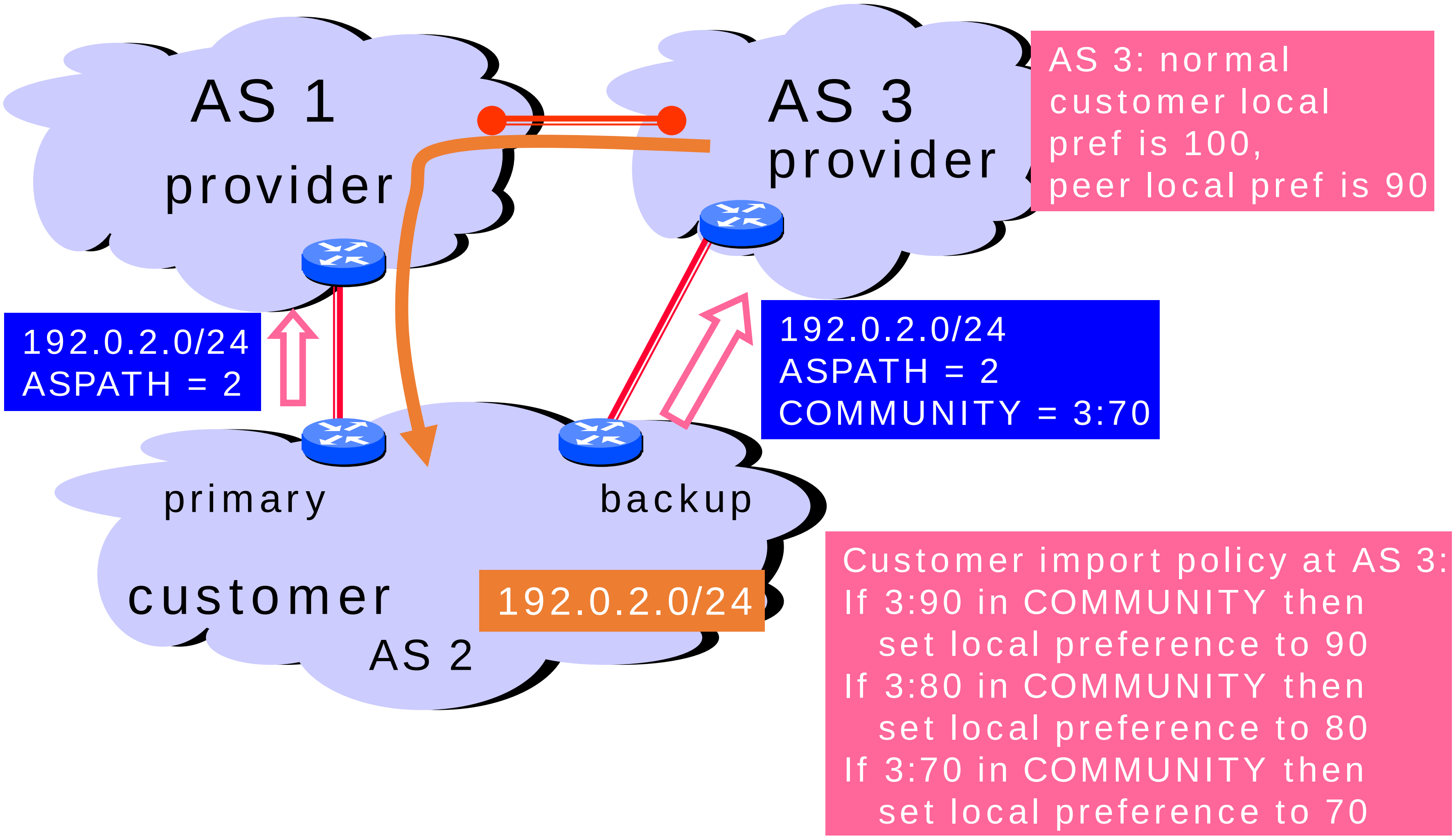
AS_Path prepending does not always work



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

Pre-pending in this way is often used as a form of load balancing

Use COMMUNITY to control parent's Local_Pref



AS 1

provider

AS 3

provider

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

192.0.2.0/24
ASPATH = 2

192.0.2.0/24
ASPATH = 2
COMMUNITY = 3:70

primary

backup

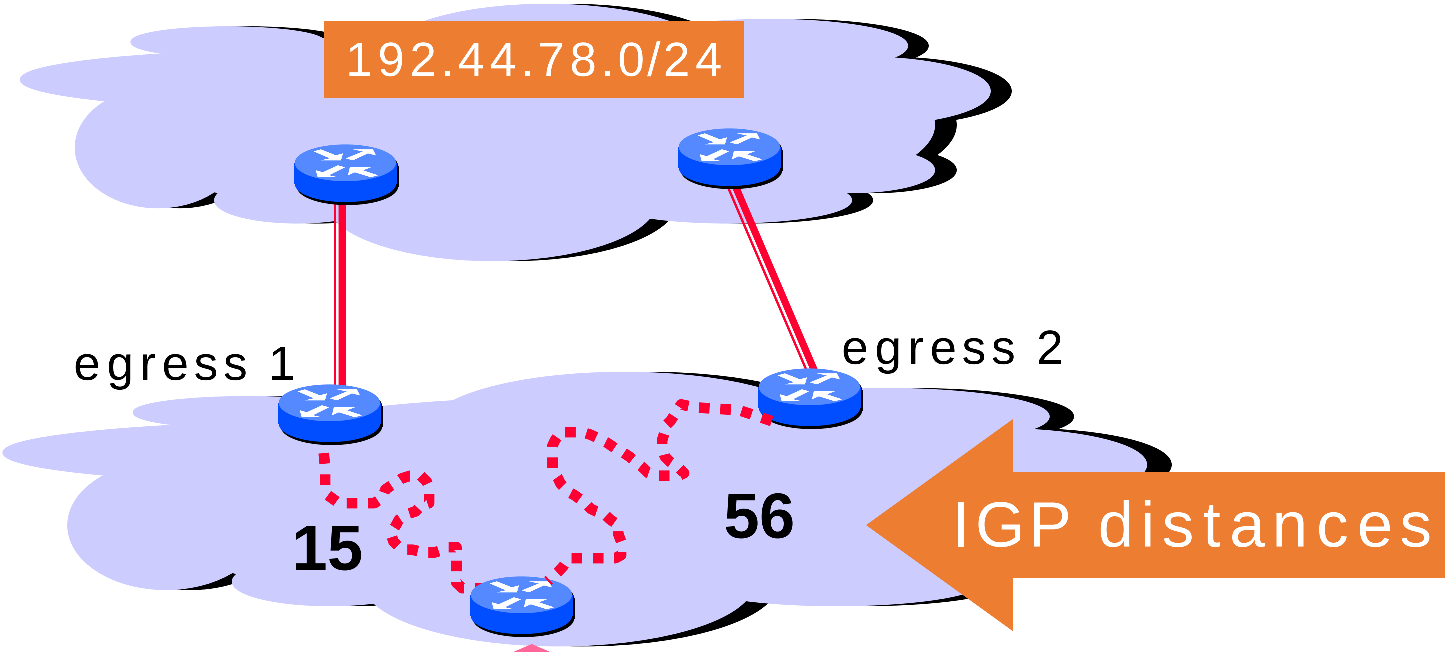
customer

AS 2

192.0.2.0/24

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

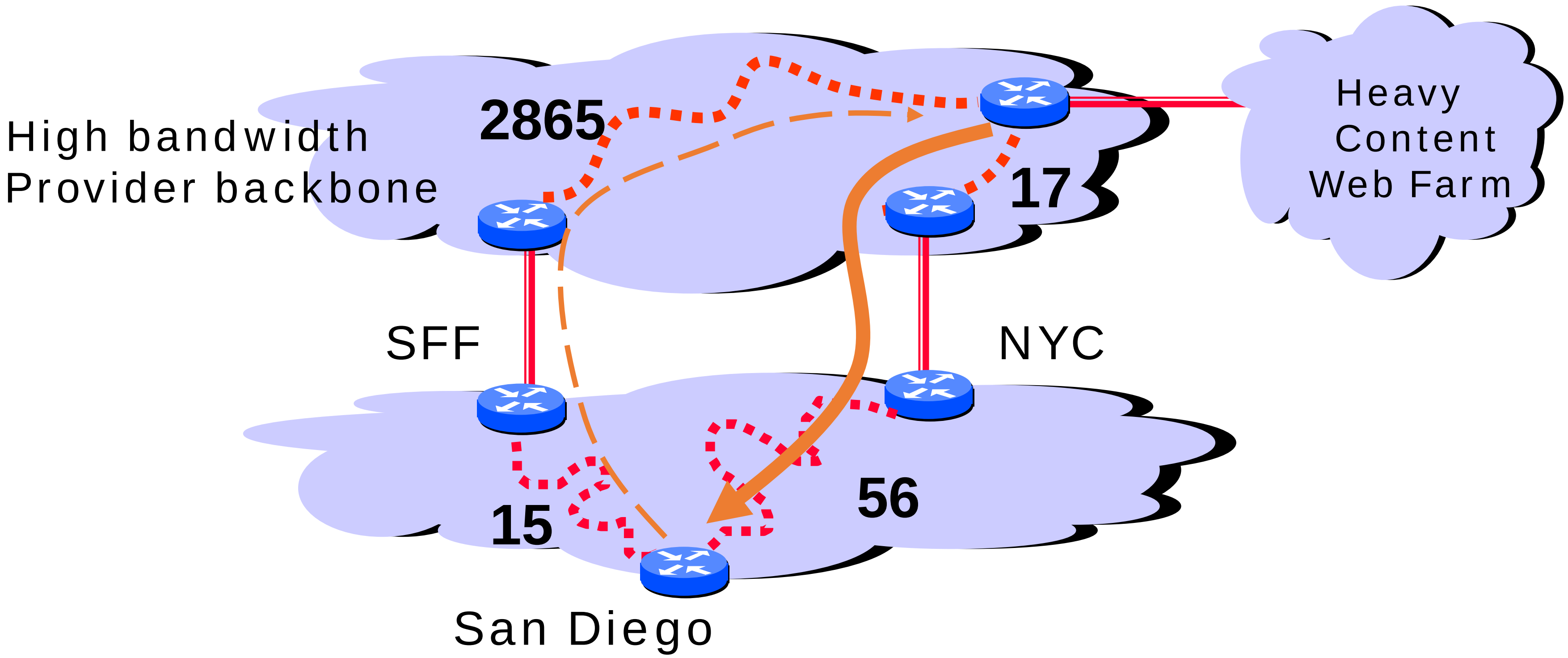


This Router has two BGP routes to 192.44.78.0/24.

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

Note: Local_Pref, AS_PATH are the same -> IGP cost

Hot potato routing used at provider too

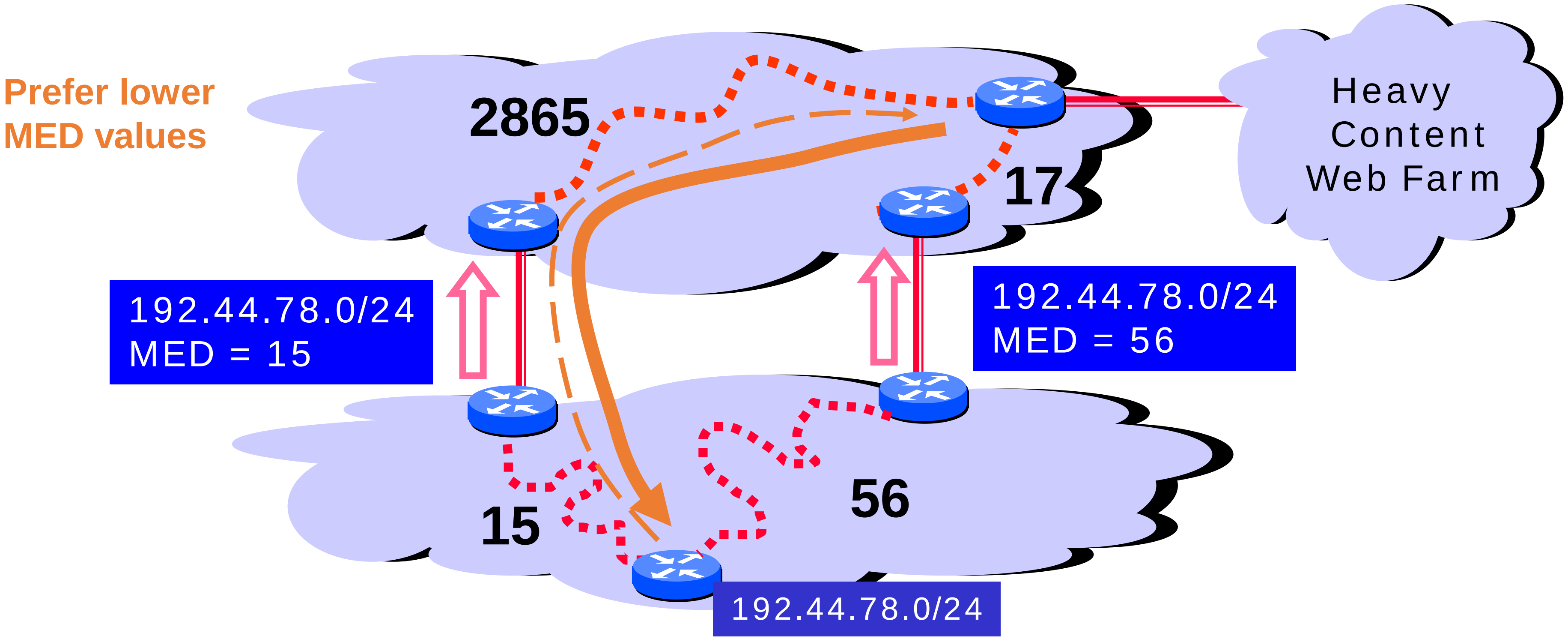


Many customers want their provider to carry the bits!

- - → tiny http request

⟶ huge http reply

Use MED to change provider

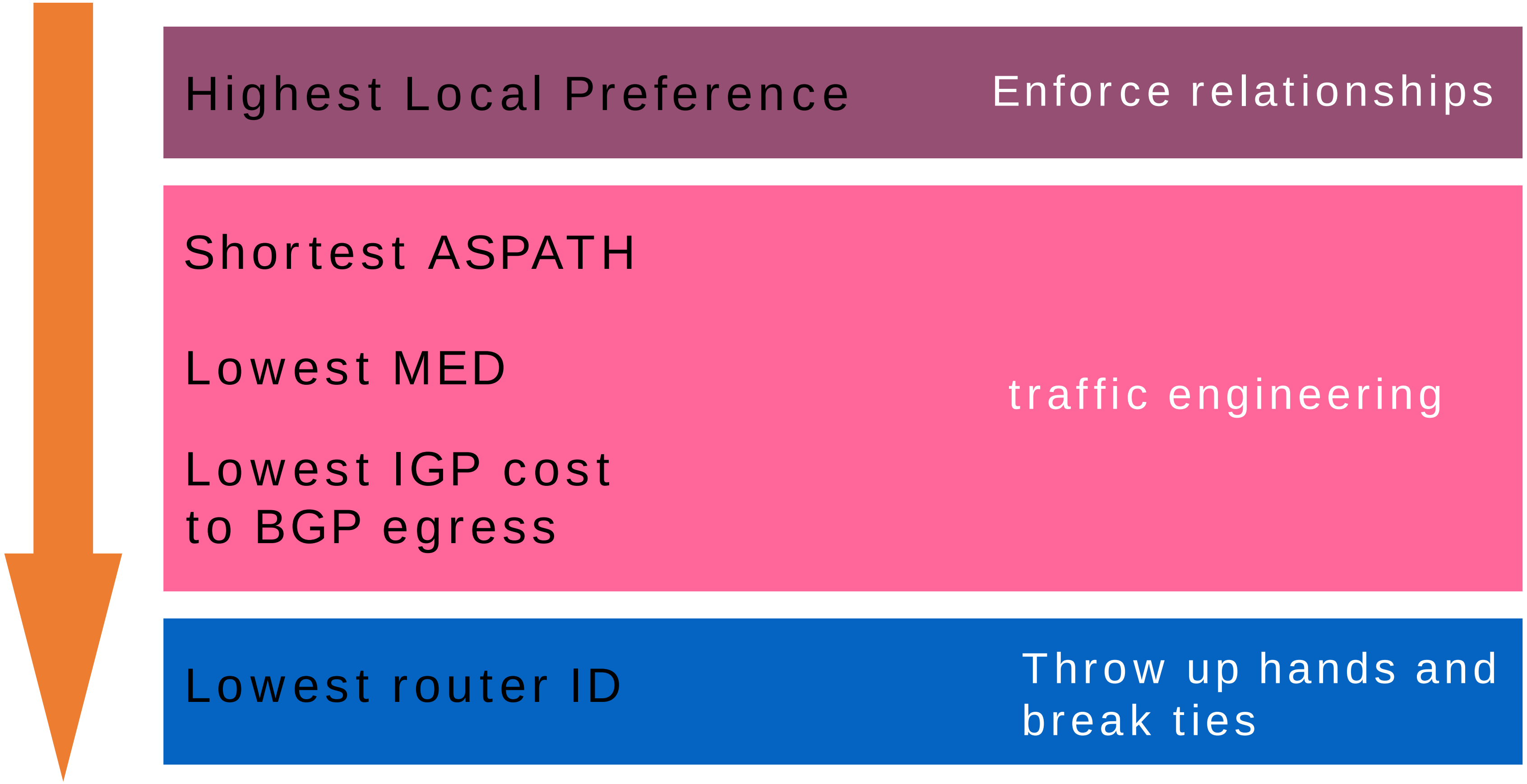


MED = Multi-Exit Discriminator

This means that MEDs must be considered BEFORE IGP distance!

Note: some providers will not listen to MEDs

The order of considering different attributes



Highest Local Preference

Enforce relationships

Shortest ASPATH

Lowest MED

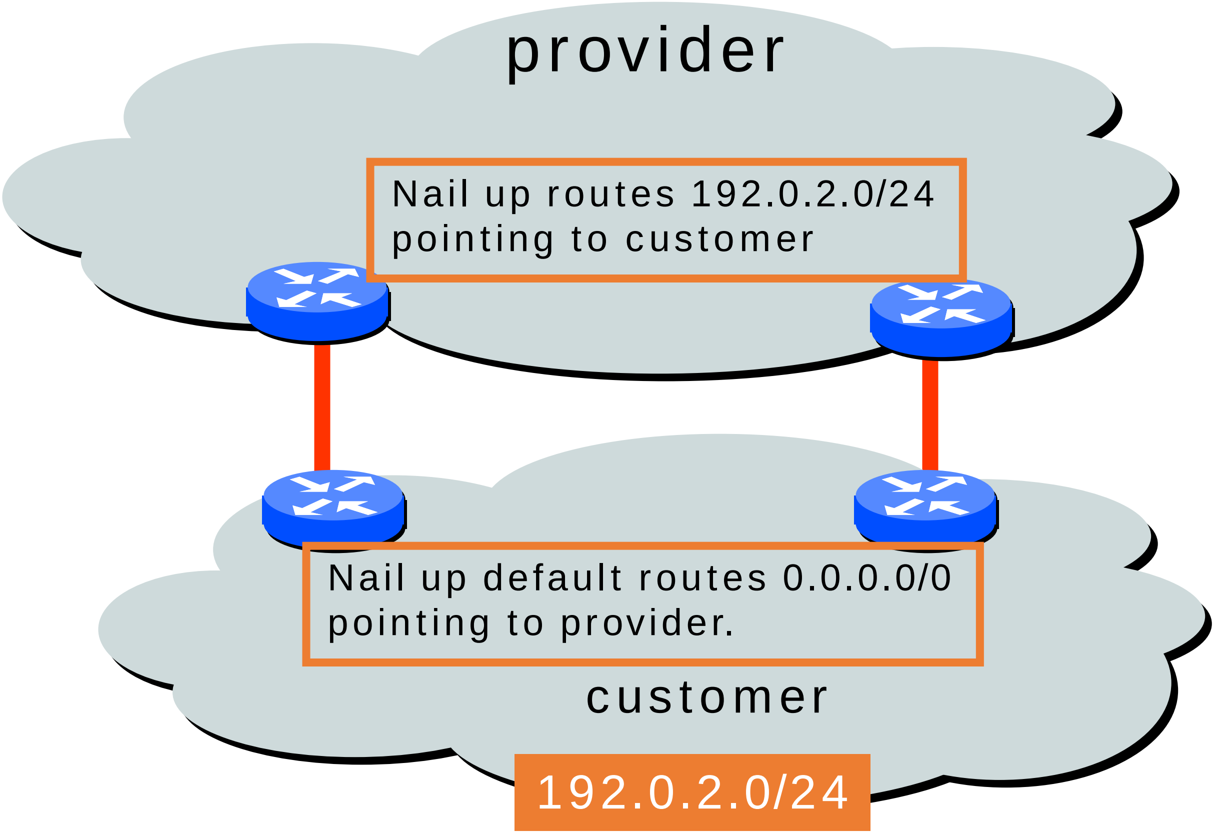
Lowest IGP cost
to BGP egress

traffic engineering

Lowest router ID

Throw up hands and
break ties

Simple customers can use default route instead of BGP


provider
Nail up routes 192.0.2.0/24 pointing to customer
Nail up default routes 0.0.0.0/0 pointing to provider.
customer
192.0.2.0/24

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 ...

Summary of inter-domain routing

- Inter-domain routing needs to implement peering agreements (who provide transit for who)
- BGP uses various attributes to allow route selection, whether for shortest path, or load balancing; also considered traffic engineering