Last Lecture: Network Layer

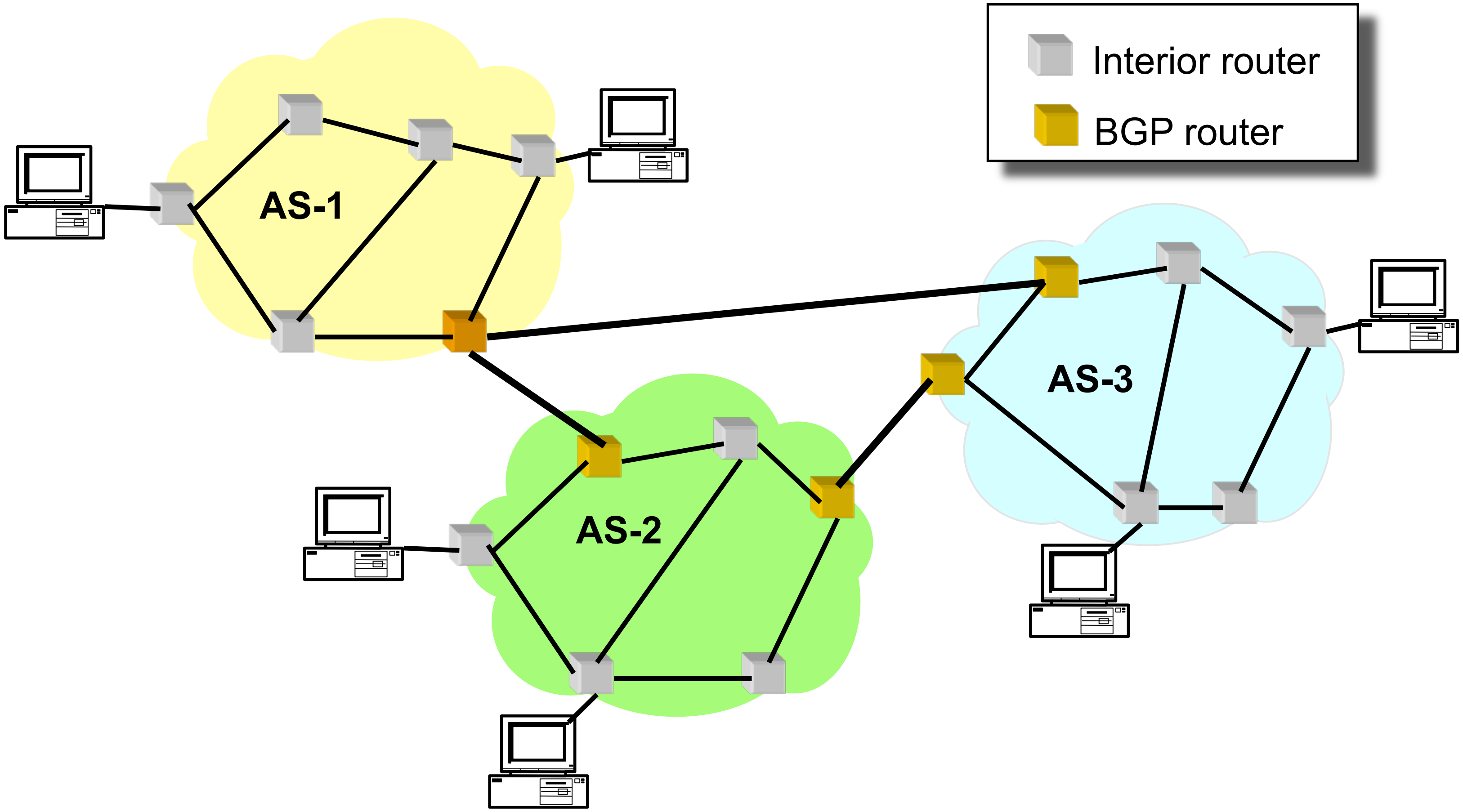
1. *Design goals and issues*
2. *Basic Routing Algorithms & Protocols*
3. *Addressing, Fragmentation and reassembly ✔*
   - *Hierarchical addressing ✔*
   - *Address allocation & CIDR ✔*
   - *IP fragmentation and reassembly*
4. *Internet Routing Protocols and Inter-networking*
5. *Router design*
6. *Congestion Control, Quality of Service*
7. *More on the Internet's Network Layer*

This Lecture: Network Layer

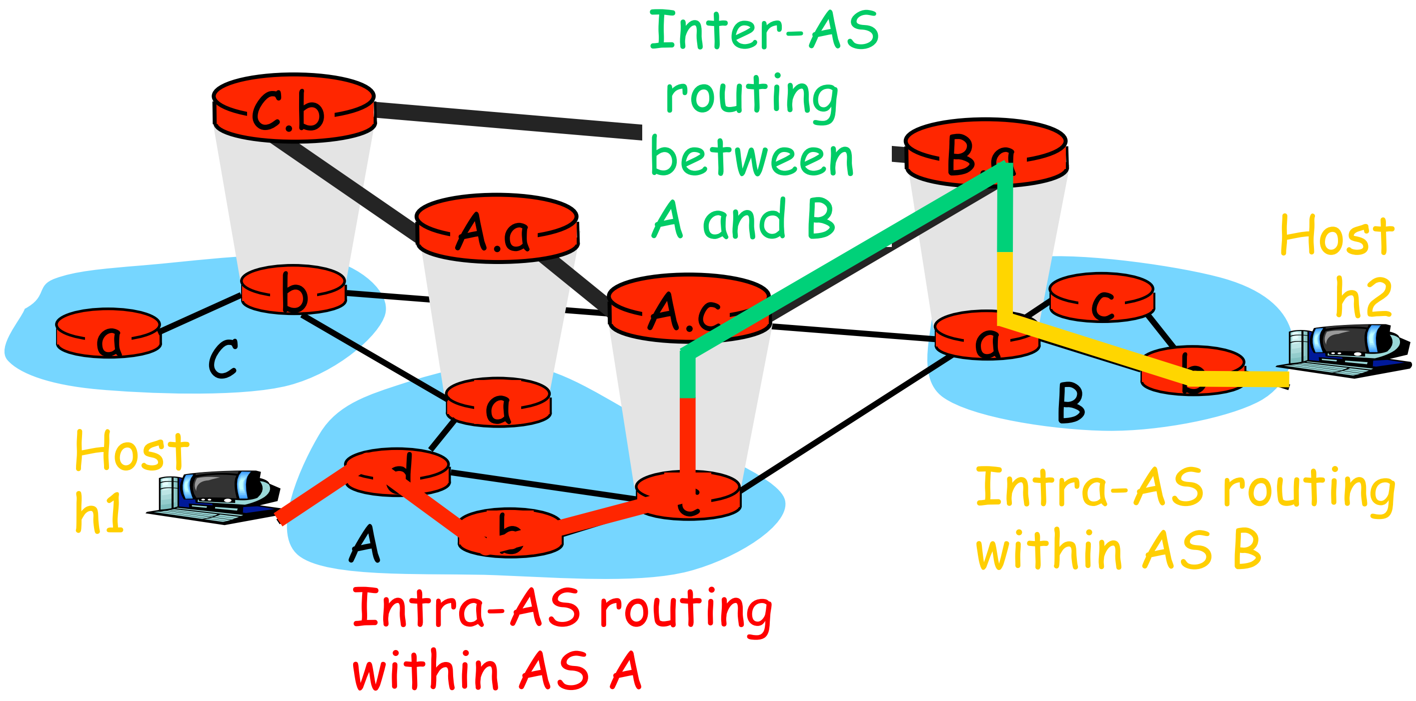
1. *Design goals and issues*
2. *Basic Routing Algorithms & Protocols*
3. *Addressing, Fragmentation and reassembly*
4. *Internet Routing Protocols and Inter-networking* ✔
   - ○ *Intra- and Inter-domain Routing Protocols* ✔
   - ○ *Introduction to BGP* ✔
   - ○ *Why is routing so hard to get right?*
   - ○ *Credits: slides taken from Jen. Rexford, Nick Feamster, Hari Balakrishnan, Tim Griffin ICNP'02 Tutorial*
5. *Router design*
6. *Congestion Control, Quality of Service*
7. *More on the Internet's Network Layer*

Flat View of Internet Hierarchy

AS = “Autonomous System”



Hierarchical Routing


Inter-AS routing between A and B
C.b
A.a
A.c
B.a
a
b
C
a
d
b
c
A
c
a
b
B
Host h1
Host h2
Intra-AS routing within AS B
Intra-AS routing within AS A

Commonly Used Protocols

- *Intra-AS* or *Interior Gateway Protocols* (IGPs)
  - **Static**: used in very small domains
  - [DV] **RIP**: used in some small domains (has limitations)
  - [LS] **OSPF**: widely used in enterprise networks
  - [LS] **IS-IS**: widely used in ISP networks
  - [DV] Cisco's **IGRP** and **EIGRP**
- *Inter-AS* or *Exterior Gateway Protocol* (EGPs)
  - **BGP** (v4) – de facto standard

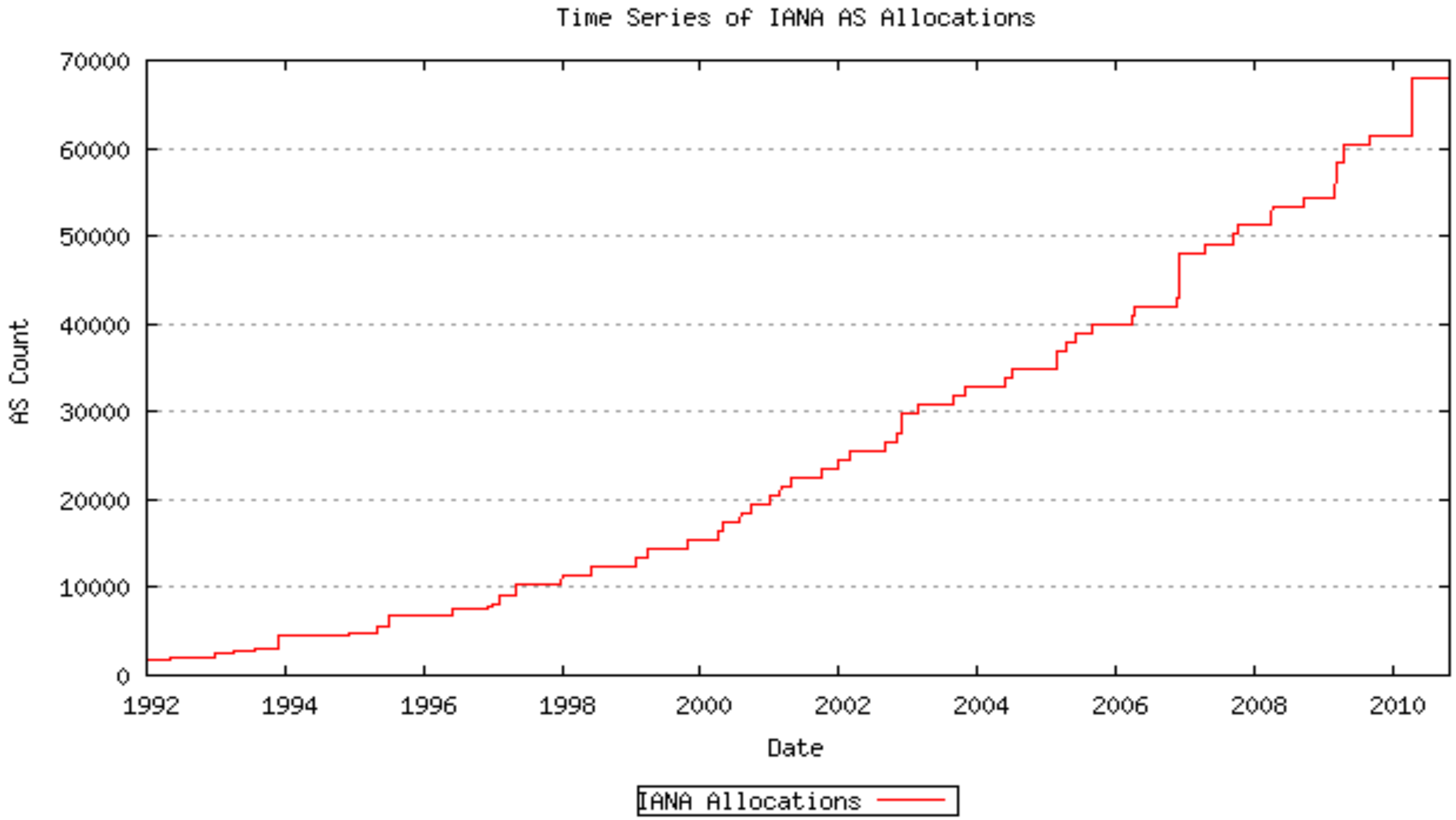
Why Hierarchical Routing?

Because single routing algorithm

- *Does not scale well*
  - 768 Mil destinations (Jul 2010) can’t be stored in memory
  - LS: overhead required to broadcast link status + reveals too much information
  - DV: likely never converge
- *Is politically infeasible*

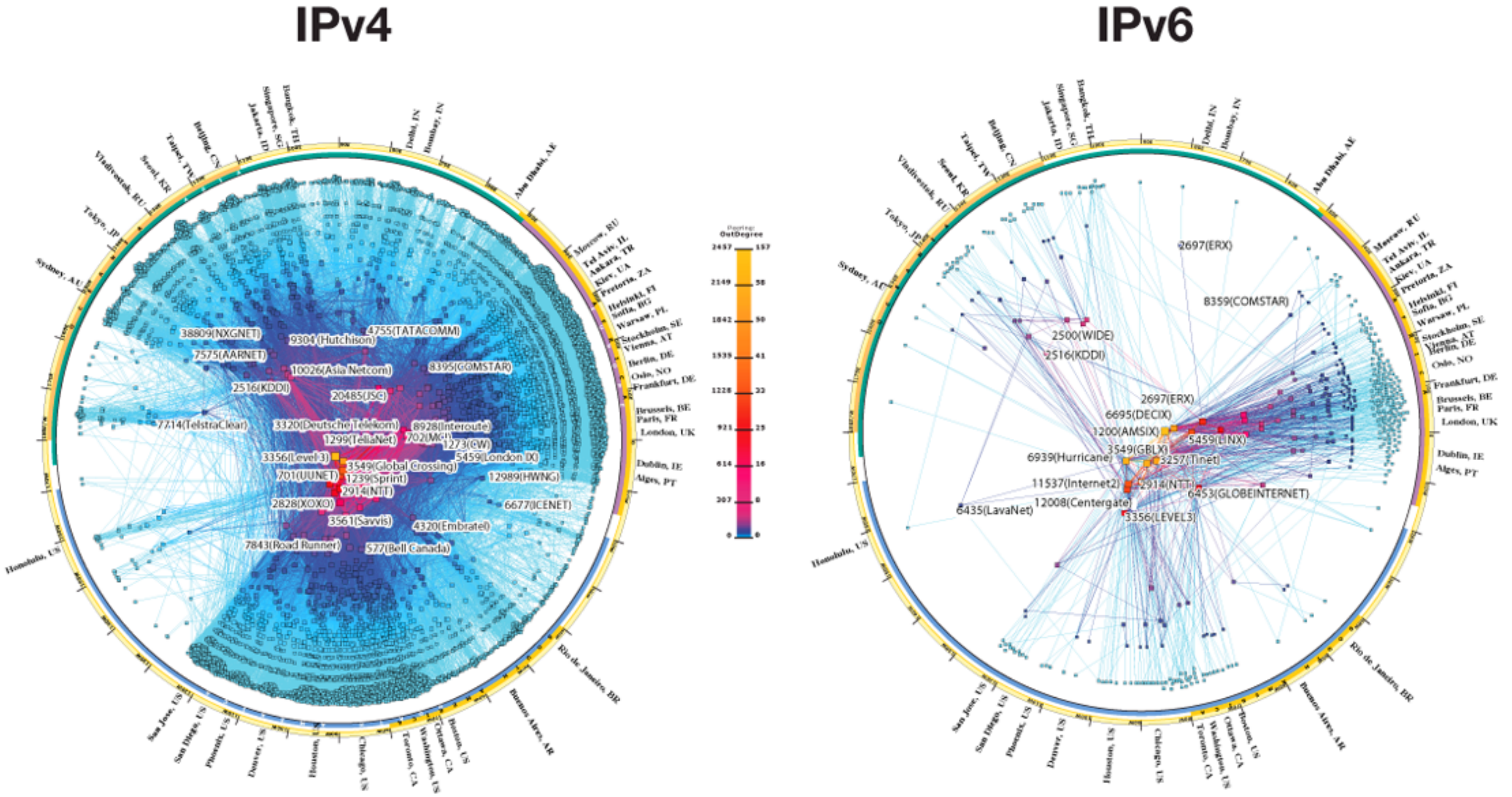
Even *with* hierarchical routing, scalability is a very hard problem to solve

Scale: AS Numbers in Use as of Nov 01, 2010



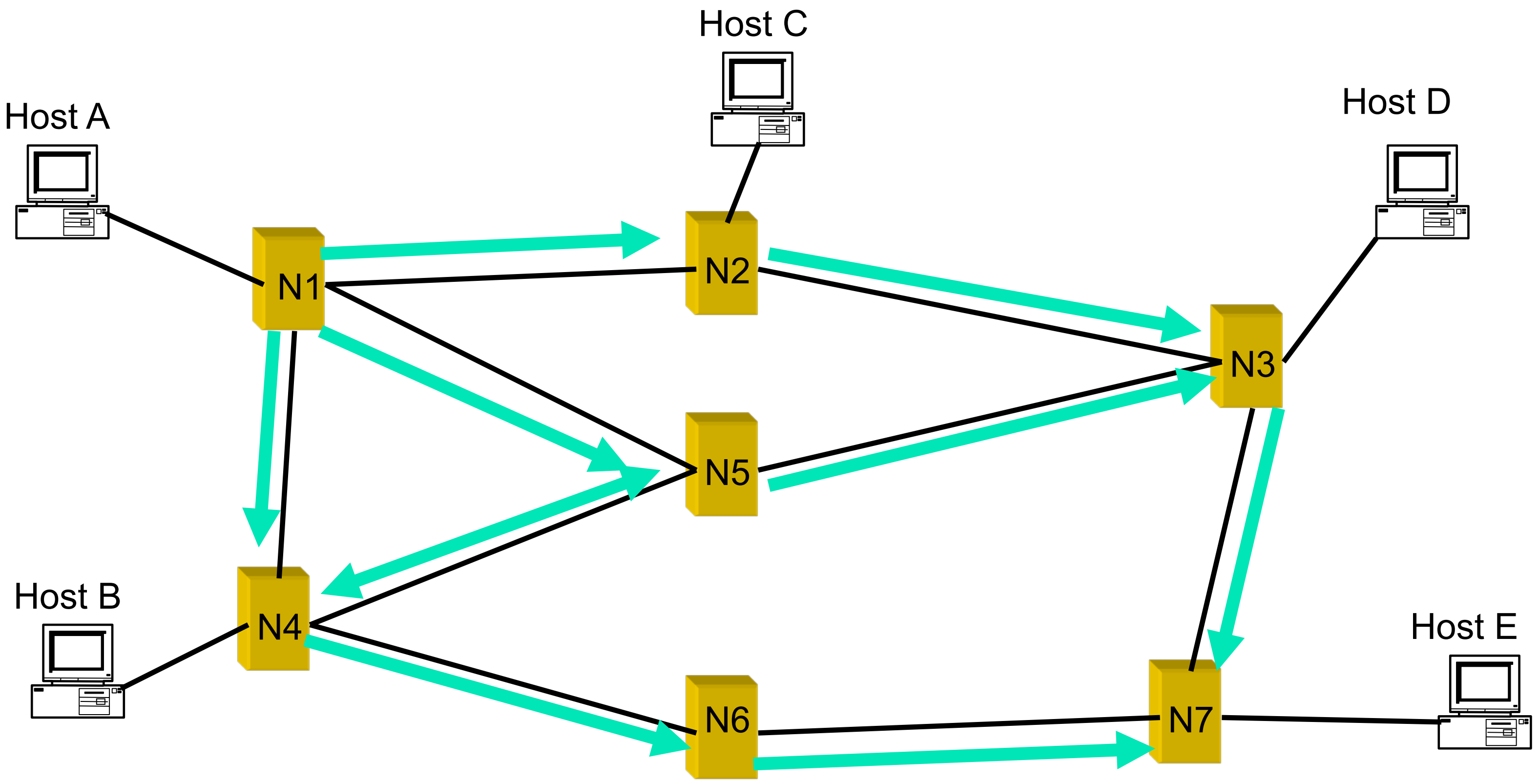
IPv4 & IPv6 INTERNET TOPOLOGY MAP JANUARY 2009

AS-level INTERNET GRAPH



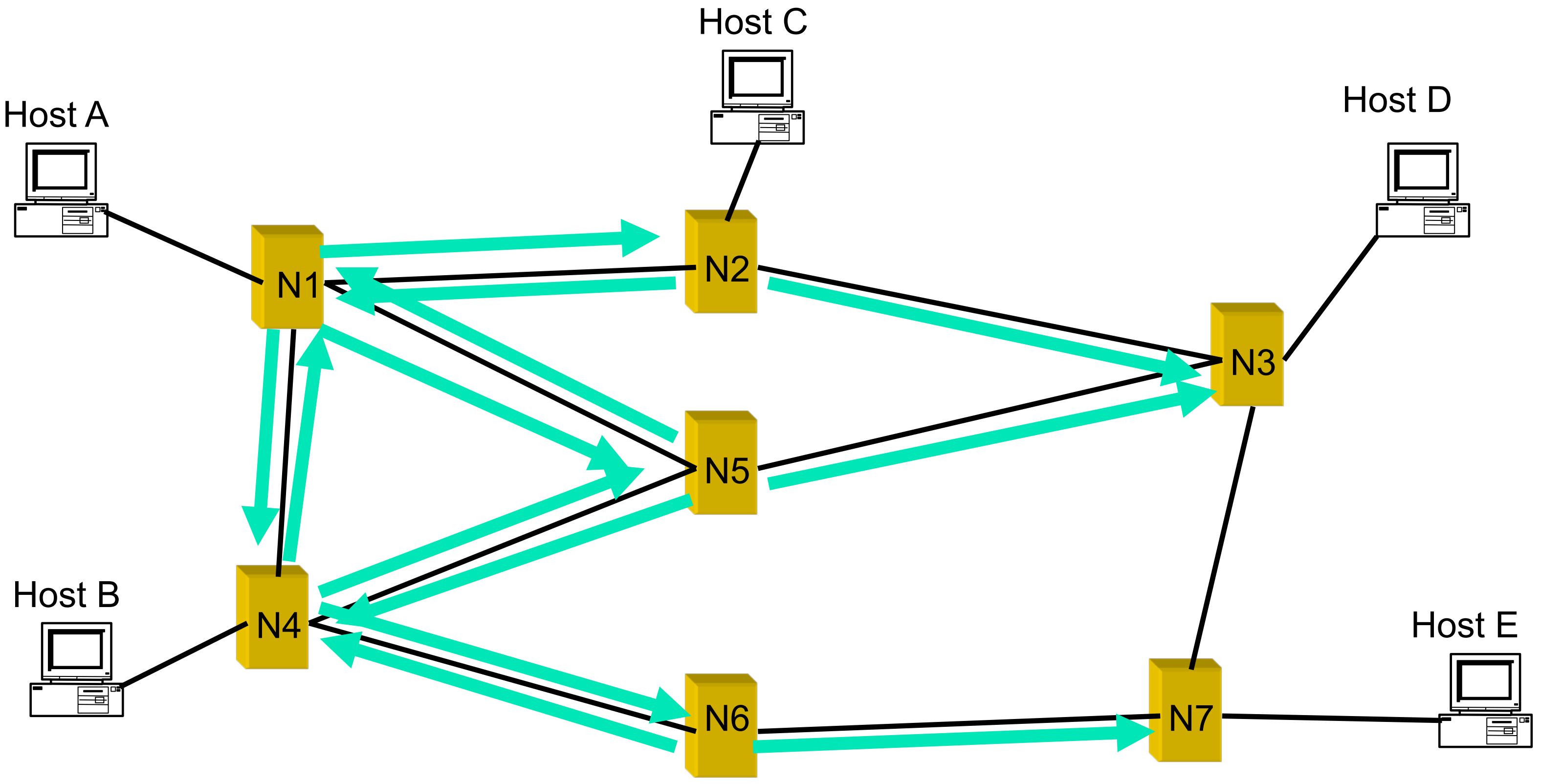
Reminder: Link State

- Each node floods its neighborhood information to all



Reminder: Distance Vector

- Each node sends its table to its neighbors
- Then updates its table based on information from neighbors



Intra-domain Routing

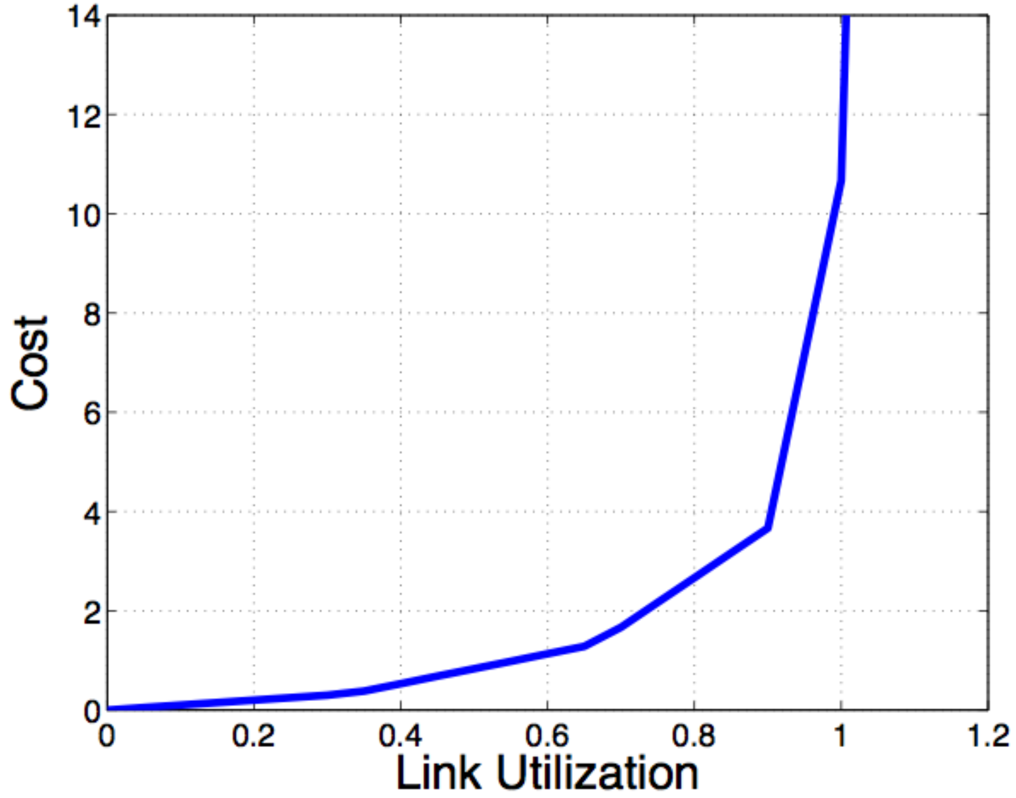
- Within an AS (operated by the same organization), *traffic engineering* is important (and feasible)
  - "Engineer" traffic to optimize some objective(s)

- This is an example of *many* research topics related to intra-AS routing

- The basic question is
  - *How do we set the link weights?*
  - What is the objective function anyway?

Objective of Traffic Engineering

- Generally, a convex function
  - Ex 1: minimize the maximum link utilization
  - Ex 2: minimize sum of (link) *congestion cost*
    - Model queueing delay, “proportional” to congestion



Selecting Link Weights in OSPF

- OSPF splits traffic evenly among shortest paths
- Finding the best link weights (to minimize congestion cost) is NP-Hard
  - Proved in [Fortz-Thorup 2000]
  - Heuristics proposed based on local search
- If we insist on splitting traffic evenly, then optimal traffic engineering can't be achieved
- However, smarter splitting can!
  - Xu-Chiang-Rexford, INFOCOM 2008
  - A gain of 15% in capacity utilization over OSPF was demonstrated

Inter-Domain Routing

1. Inter-AS routing: LS or DV?

2. What is the major engineering objective?
   - Trickier: *who* has the right to define the objective?

Link-State is Problematic as an EGP

- *Topology information is flooded*
  - High bandwidth and storage overhead
  - Forces nodes to divulge sensitive information
- *Entire path computed locally per node*
  - High processing overhead in a large network
- *Minimizes some notion of total distance*
  - Works only if policy is shared and uniform

- Thus, typically used only inside an AS
  - E.g., OSPF and IS-IS

Distance Vector is on the Right Track

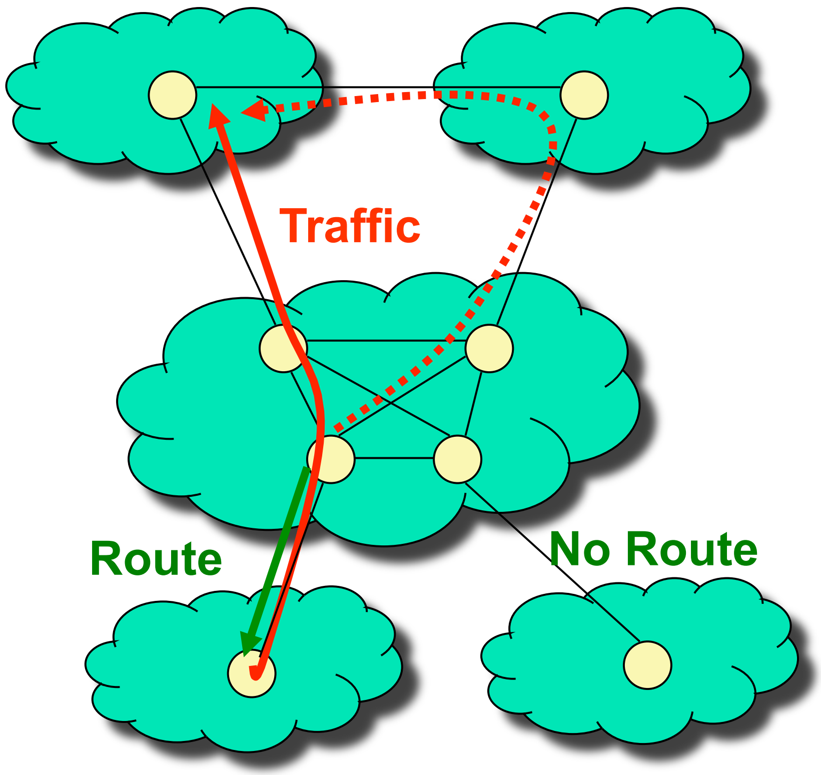
- *Advantages*
  - Hides details of the network topology
  - Nodes determine only “next hop” toward the dest
- *Disadvantages*
  - Minimizes *some* notion of total distance, which is difficult in an inter-domain setting
  - Slow convergence due to the counting-to-infinity problem (“bad news travels slowly”)
- *Idea*: extend the notion of a distance vector
  - Make it easier to detect loops
  - Thus avoid count to infinity

Inter-Domain Routing

1. Inter-AS routing: LS or DV?
   - Answer: Path Vector (PV)
2. What is the major engineering objective?
   - Answer: engineering objective is secondary to political/economic objective/policies
   - PV can help with that

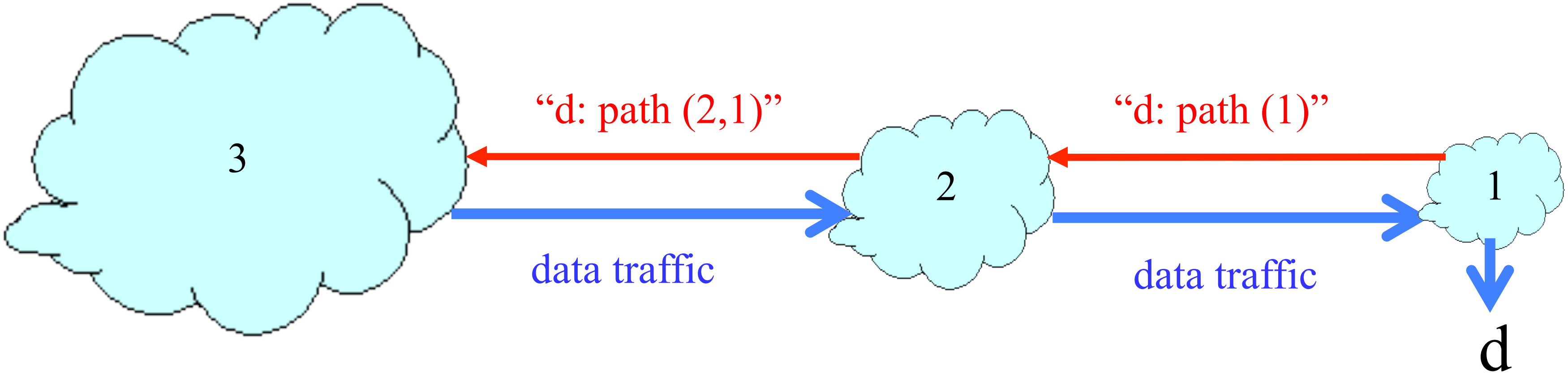
What Kind of Policy Are You Talking About?

- Which neighboring networks can send traffic
- Where traffic enters and leaves the network
- How routers within the network learn routes to external destinations
- And many others



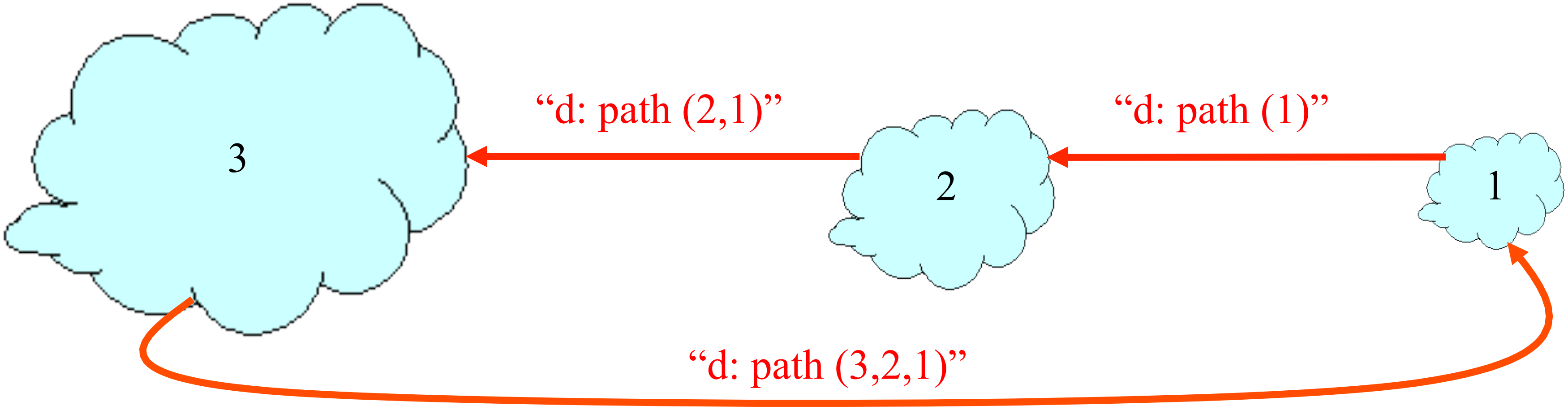
Path-Vector (PV) Routing

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



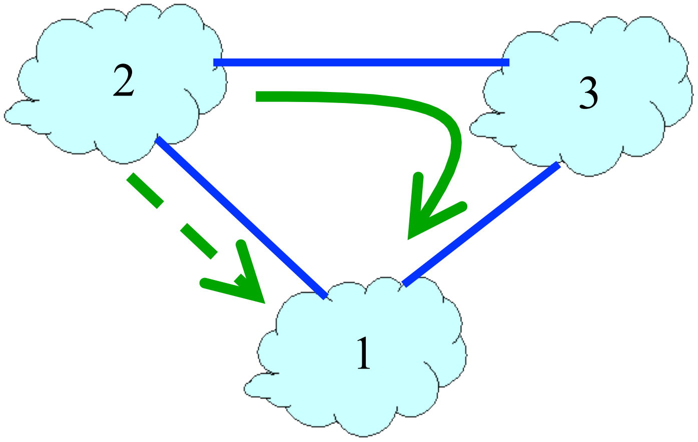
PV Pro: Faster Loop Detection

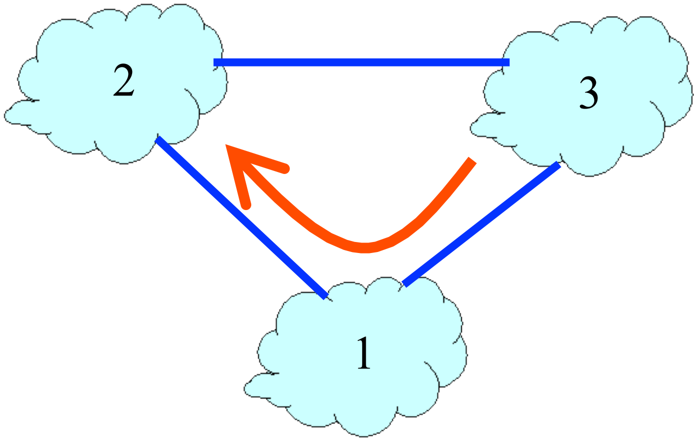
- Node can easily detect a loop
  - Look for its own node identifier in the path
  - E.g., node 1 sees itself in the path “3, 2, 1”
- Node can simply discard paths with loops
  - E.g., node 1 simply discards the advertisement



PV Pro (?): Flexible Policies

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



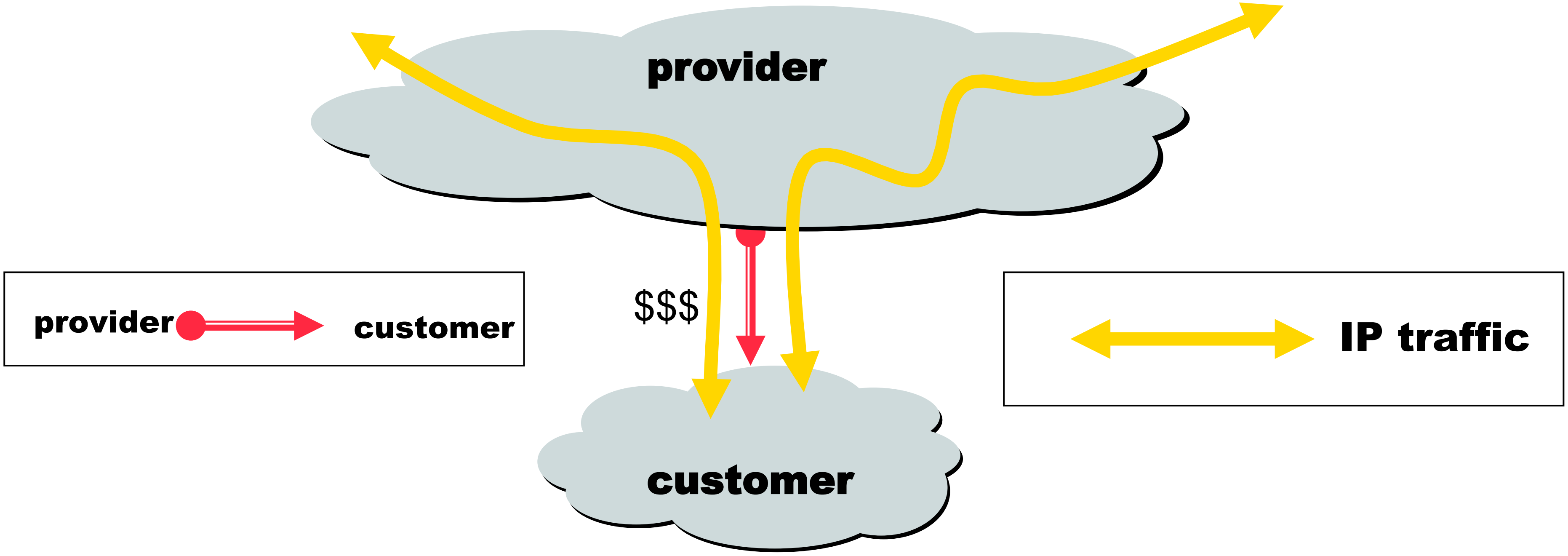


Why Are There All These Path “Preferences”?

Need to go back to see how autonomous systems are connected in the first place

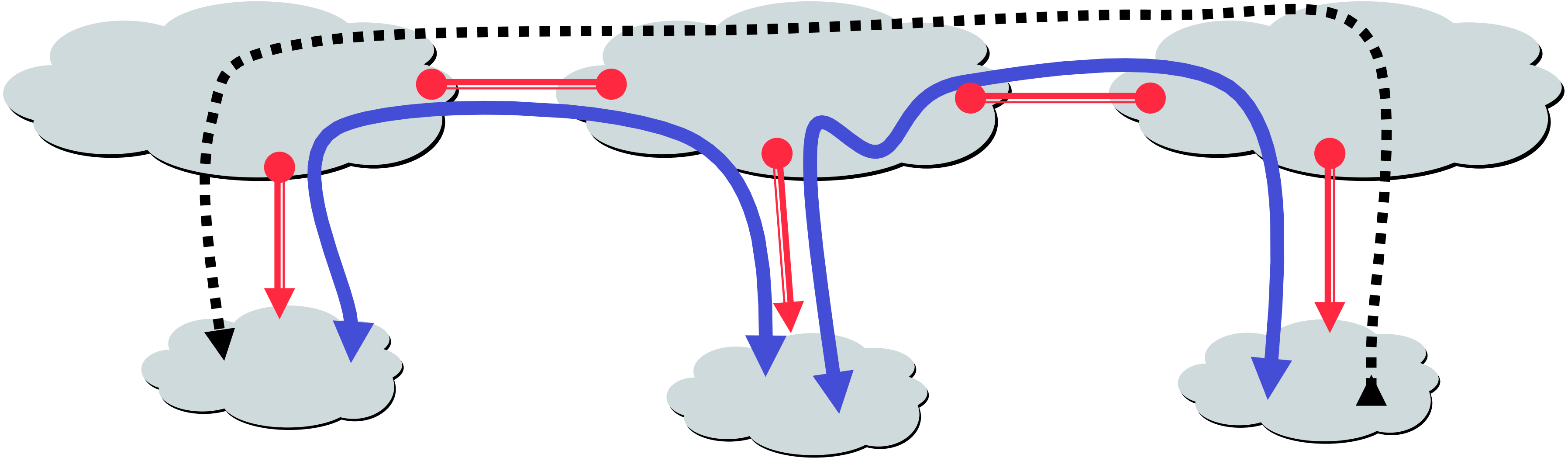
- Customer-Provider
- Peering

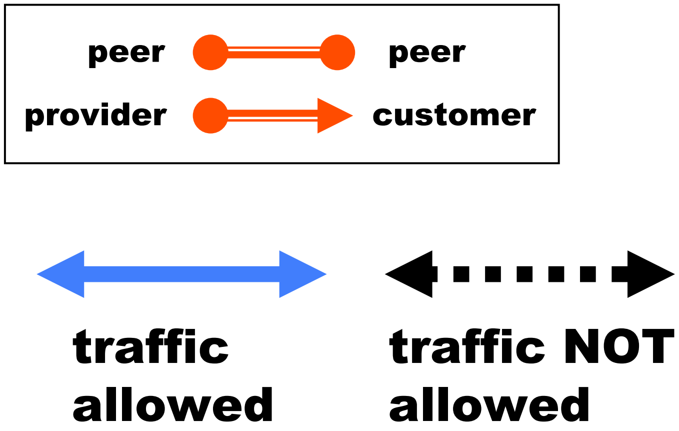
Customers and Providers (aka “Transit”)



Customer pays provider for access to the Internet

“Peering”



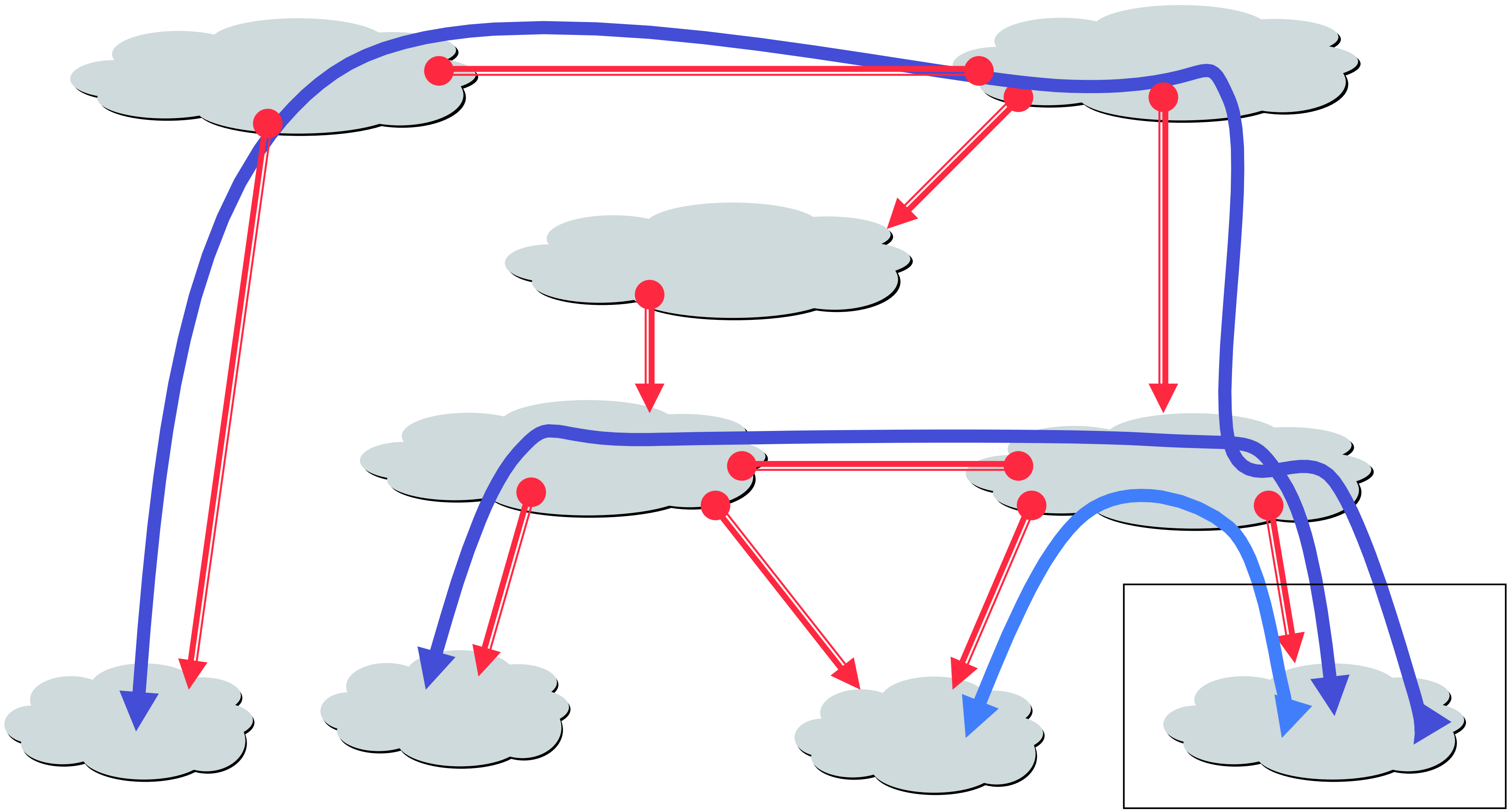


Peers provide transit between their respective customers

<u>Peers do not provide transit between peers</u>

Peers (often) do not exchange $$$, especially when traffic ratio is NOT highly asymmetric (< 4:1)

Peering Provides “Shortcuts”



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

peer ●═● peer

provider ●⇒ customer

To Peer or Not To Peer, That's the Problem

Peer

- Reduces upstream transit costs
- Can increase end-to-end performance
- May be the only way to connect your customers to some part of the Internet ("Tier 1")

Don't Peer

- You would rather have customers
- Peers are usually your competitors
- Peering relationships may require periodic renegotiation

Peering struggles are by far the most contentious issues in the ISP world!

Peering agreements are almost always confidential.

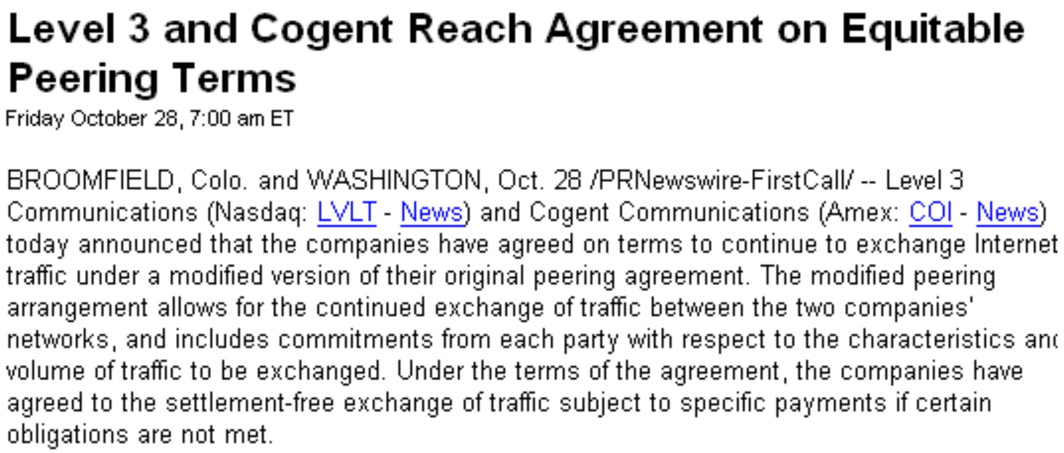
The Business Game & Depeering

- **31 Jul 2005:** Level 3 Notifies Cogent of intent to disconnect
- **16 Aug 2005:** Cogent begins massive sales effort and mentions a 15 Sept. expected depeering date.
- **31 Aug 2005:** Level 3 Notifies Cogent again of intent to disconnect (according to Level 3)
- **5 Oct 2005 9:50 UTC:** Level 3 disconnects Cogent. Mass hysteria ensues up to, and including policymakers in Washington, D.C.
- **7 Oct 2005:** Level 3 reconnects Cogent

During the "outage", Level 3 and Cogent's singly homed customers could not reach each other. (~ 4% of the Internet's prefixes were isolated from each other)

Depeering Continue …

Resolution…



Level 3 and Cogent Reach Agreement on Equitable Peering Terms

Friday October 28, 7:00 am ET

BROOMFIELD, Colo. and WASHINGTON, Oct. 28 /PRNewswire-FirstCall/ -- Level 3 Communications (Nasdaq: LVLT - News) and Cogent Communications (Amex: COI - News) today announced that the companies have agreed on terms to continue to exchange Internet traffic under a modified version of their original peering agreement. The modified peering arrangement allows for the continued exchange of traffic between the two companies' networks, and includes commitments from each party with respect to the characteristics anc volume of traffic to be exchanged. Under the terms of the agreement, the companies have agreed to the settlement-free exchange of traffic subject to specific payments if certain obligations are not met.

…but not before an attempt to steal customers!

As of 5:30 am EDT, October 5th, Level(3) terminated peering with Cogent without cause (as permitted under its peering agreement with Cogent) even though both Cogent and Level(3) remained in full compliance with the previously existing interconnection agreement. Cogent has left the peering circuits open in the hope that Level(3) will change its mind and allow traffic to be exchanged between our networks. **We are extending a special offering to single homed Level 3 customers.** →

Cogent will offer any Level 3 customer, who is single homed to the Level 3 network on the date of this notice, one year of full Internet transit free of charge at the same bandwidth currently being supplied by Level 3. Cogent will provide this connectivity in over 1,000 locations throughout North America and Europe.

The Gang of Four

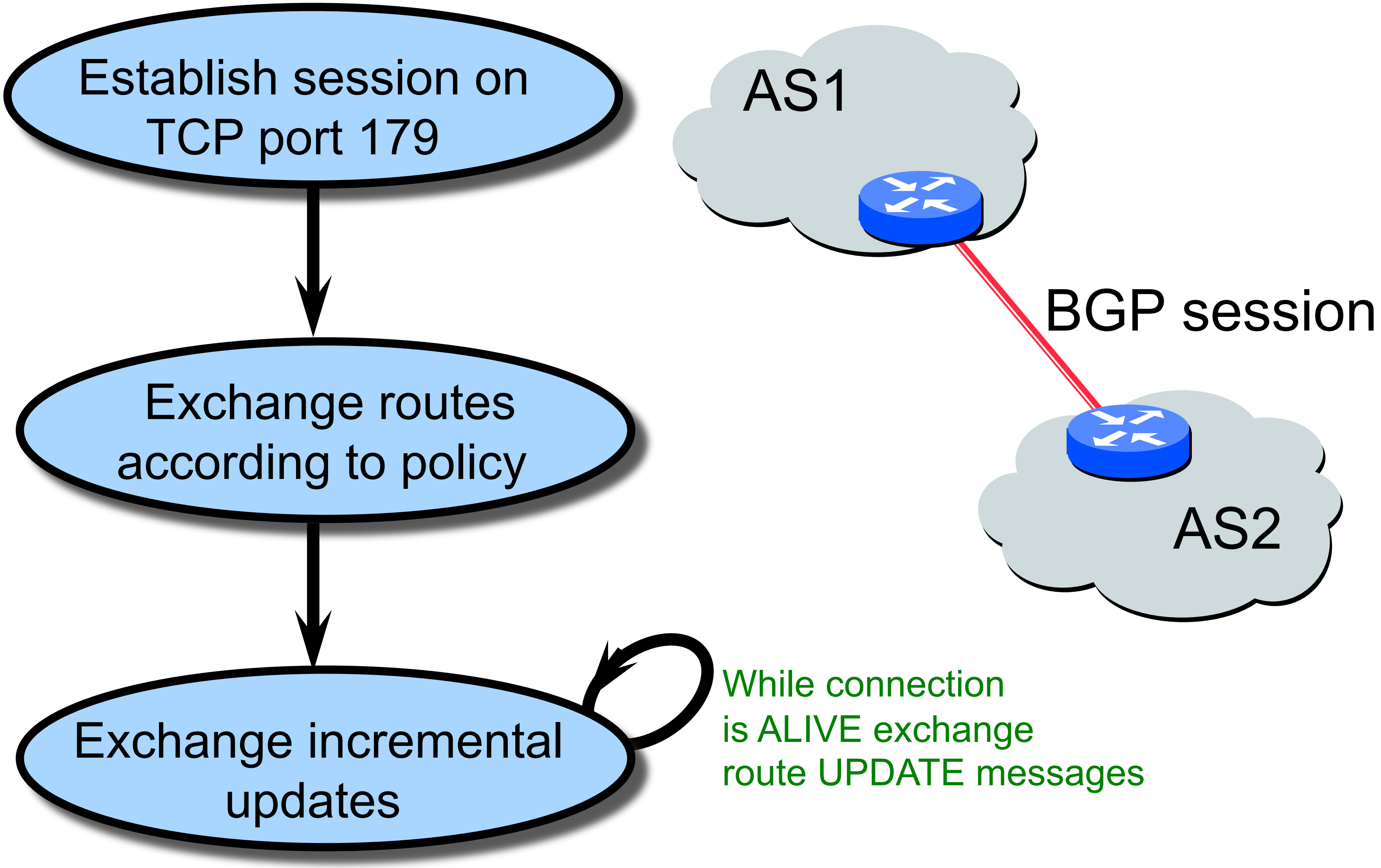
	Link State	Vectoring
IGP	OSPF IS-IS	RIP EIGRP
EGP		BGP

Border Gateway Protocol (BGP v4)

- *The* inter-domain routing protocol
  - Prefix-based path-vector protocol
  - Policy-based routing based on AS Paths
  - Evolved during the past 20 years
  - Take *years* to master

- **1989 : BGP-1 [RFC 1105], replacement for EGP**
- **1990 : BGP-2 [RFC 1163]**
- **1991 : BGP-3 [RFC 1267]**
- **1995 : BGP-4 [RFC 1771], support for CIDR**
- **2009 : BGP-4 [RFC 4271], update**

BGP Basic Operations



Establish session on TCP port 179

Exchange routes according to policy

Exchange incremental updates

While connection is ALIVE exchange route UPDATE messages

AS1

BGP session

AS2

Incremental Protocol

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

BGP Message Types

- **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.

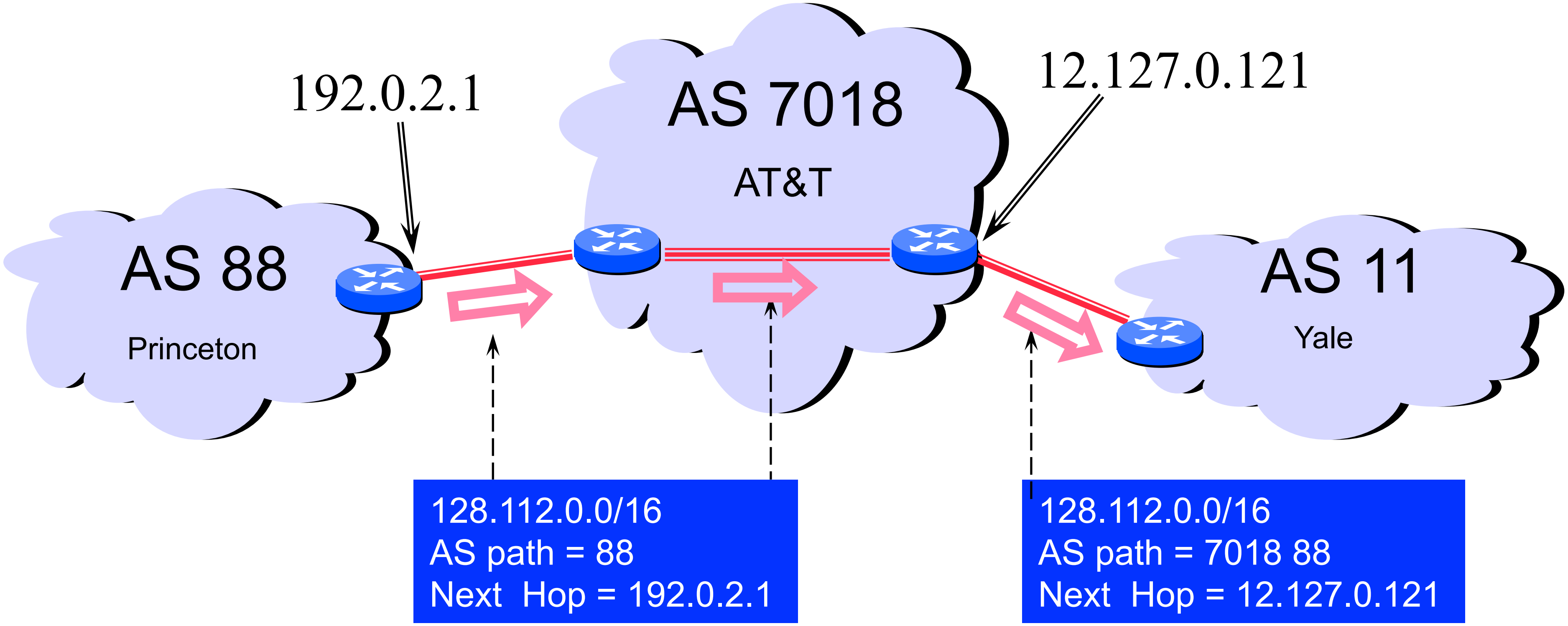
announcement

=

prefix + <u>attributes values</u>

BGP Route Advertisement

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



128.112.0.0/16
AS path = 88
Next Hop = 192.0.2.1

128.112.0.0/16
AS path = 7018 88
Next Hop = 12.127.0.121

BGP Route Attributes

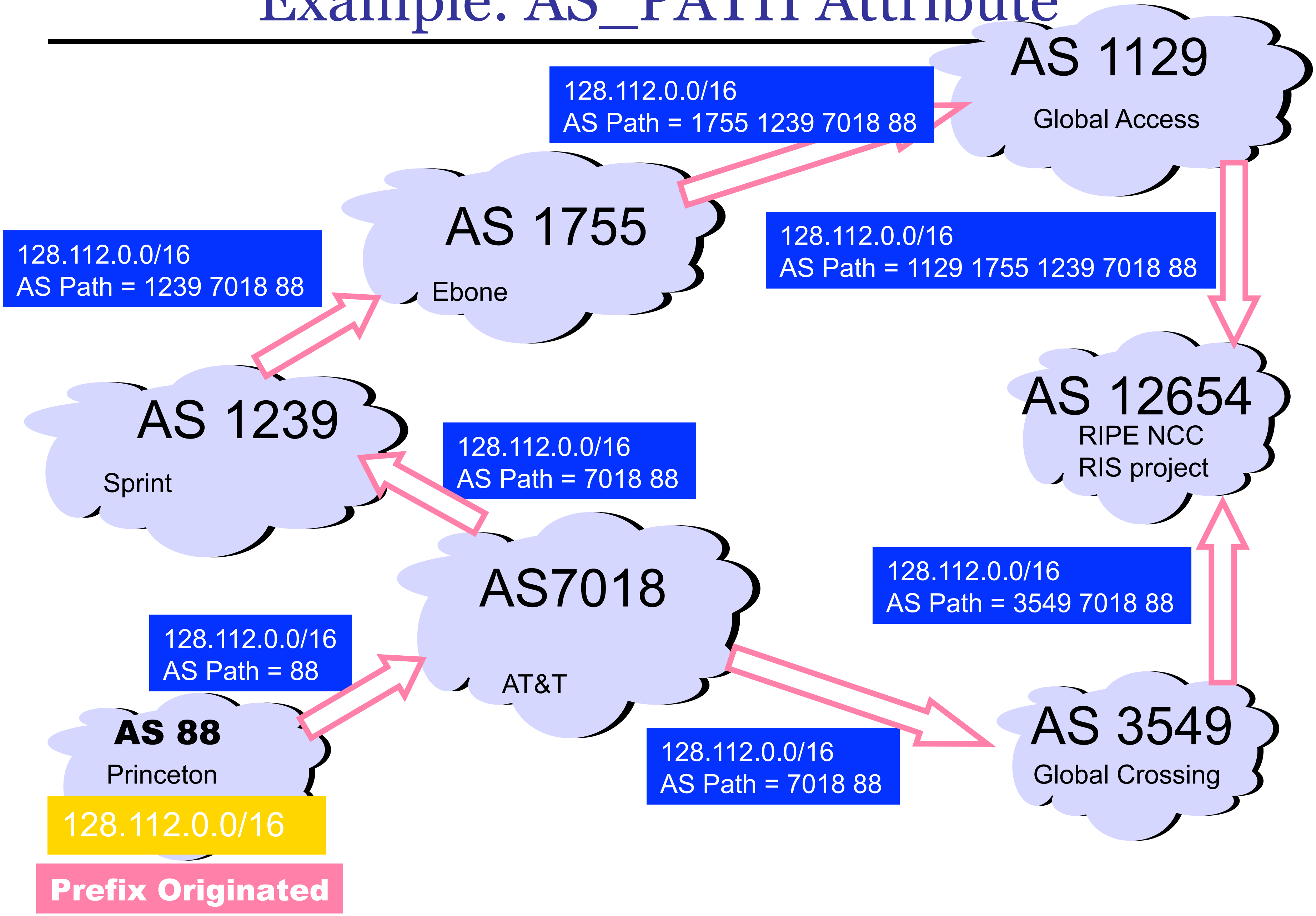
```
Value      Code                                  Reference
-----      ---------------------------------     ---------
   1       ORIGIN                                [RFC1771]
   2       AS_PATH                               [RFC1771]
   3       NEXT_HOP                              [RFC1771]
 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

Example: AS_PATH Attribute



AS 1129
Global Access

128.112.0.0/16
AS Path = 1755 1239 7018 88

AS 1755
Ebone

128.112.0.0/16
AS Path = 1239 7018 88

128.112.0.0/16
AS Path = 1129 1755 1239 7018 88

AS 12654
RIPE NCC
RIS project

AS 1239
Sprint

128.112.0.0/16
AS Path = 7018 88

AS7018
AT&T

128.112.0.0/16
AS Path = 3549 7018 88

128.112.0.0/16
AS Path = 88

AS 88
Princeton

128.112.0.0/16

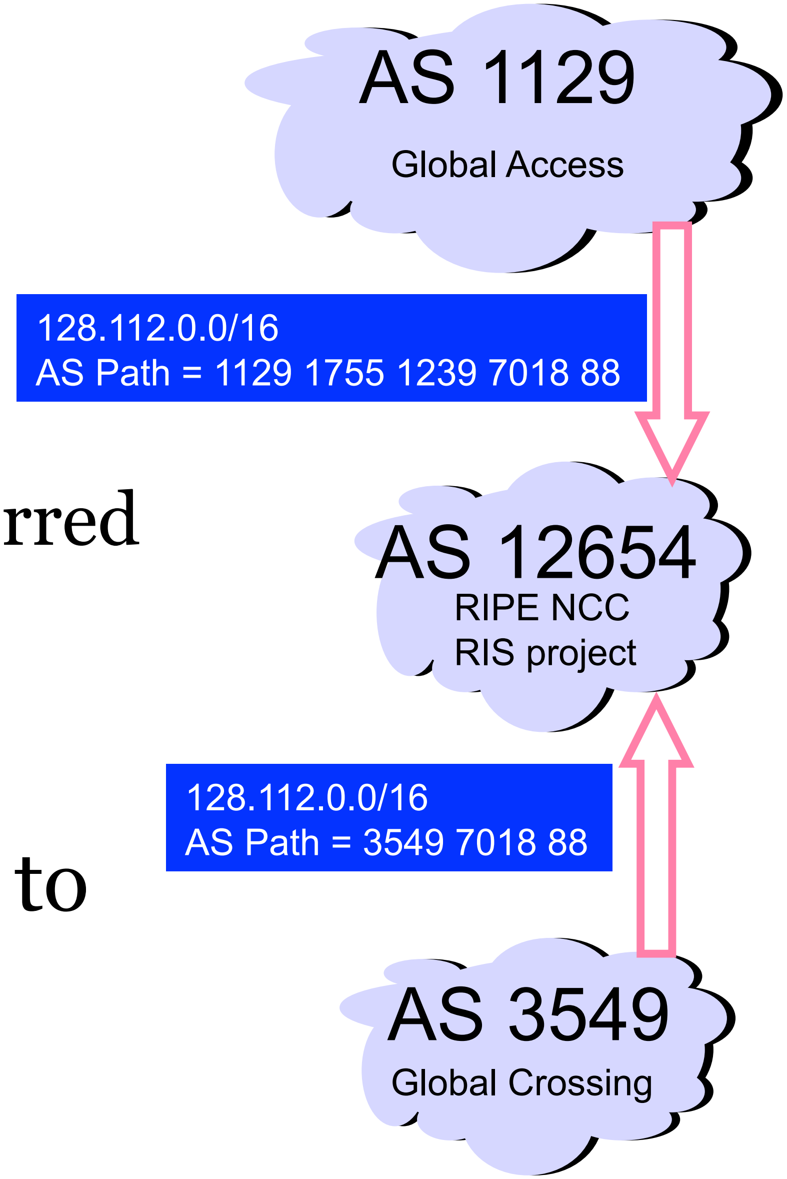
128.112.0.0/16
AS Path = 7018 88

AS 3549
Global Crossing

Prefix Originated

BGP Path Selection

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



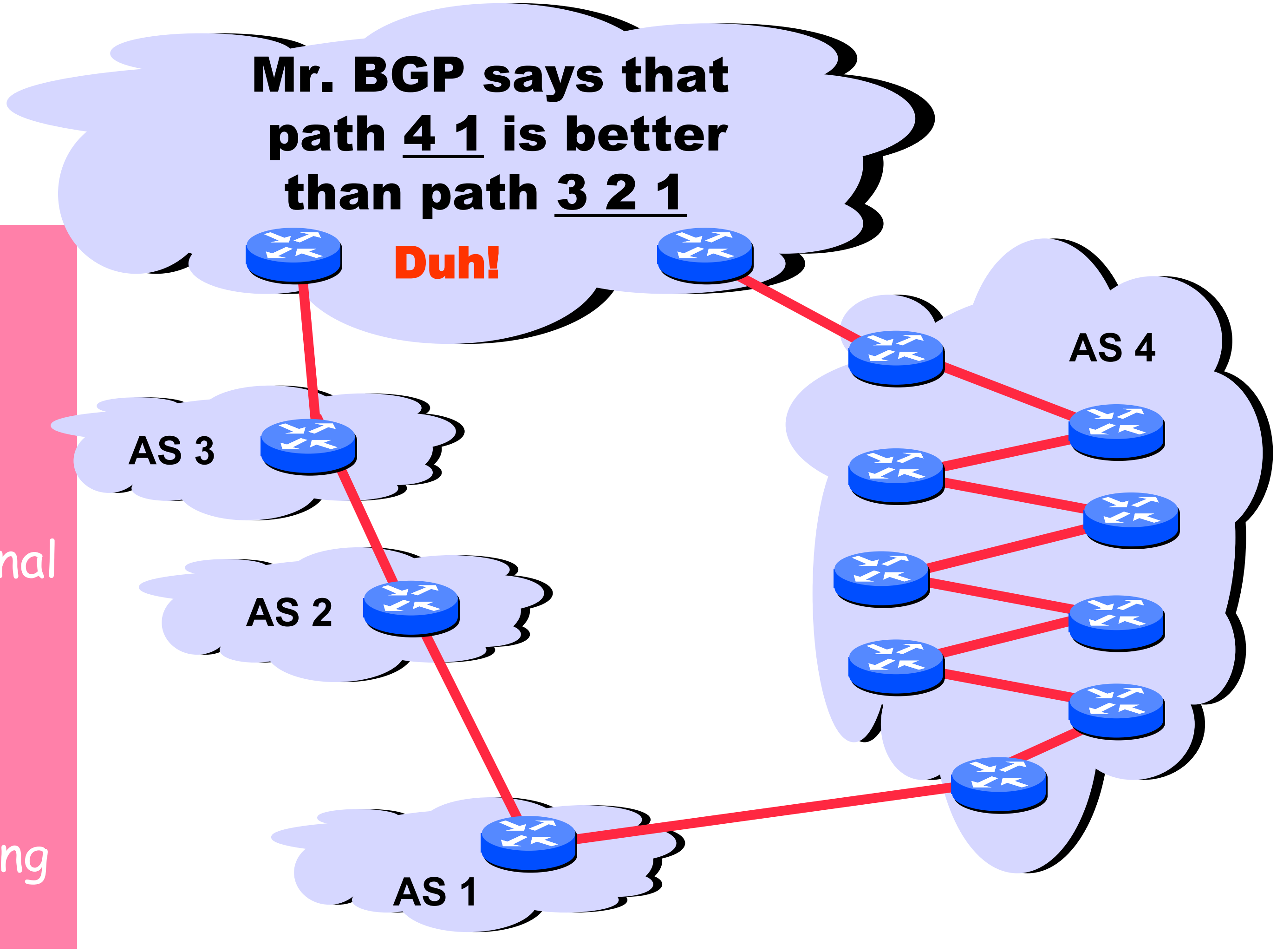
128.112.0.0/16
AS Path = 1129 1755 1239 7018 88

128.112.0.0/16
AS Path = 3549 7018 88

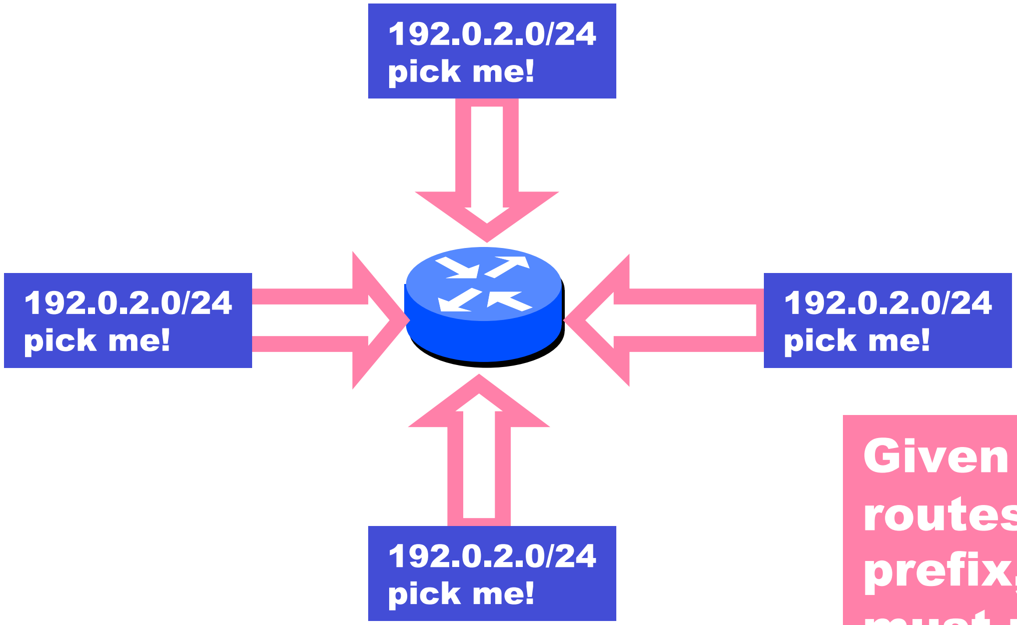
Problem Even in the Simplistic Case

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

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



Reality: Path Selection is Much More Complex



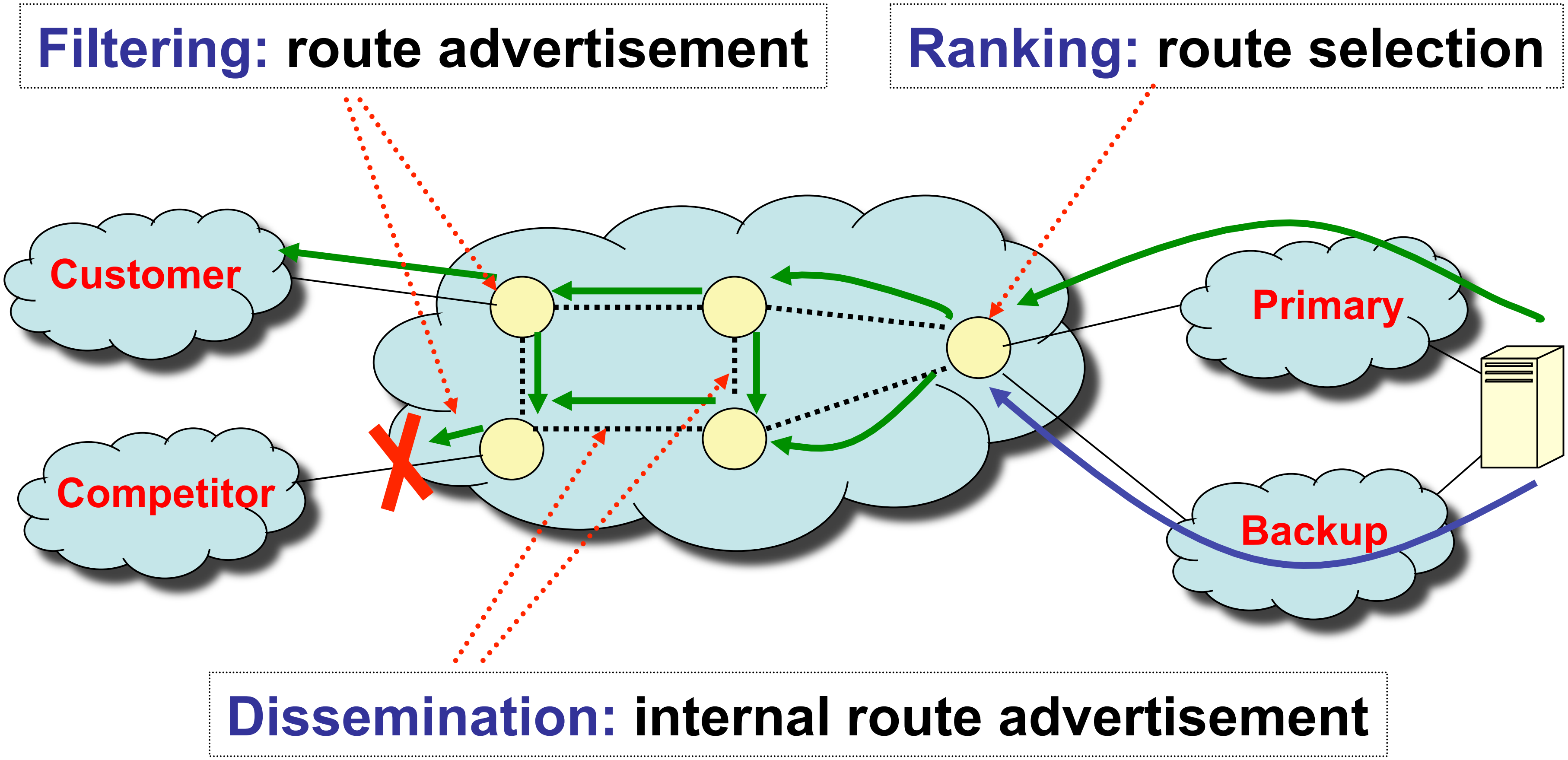
Given multiple routes to the same prefix, a BGP speaker must pick at most <u>one</u> best route

(Note: it could reject them all!)

Policy-Based Path Selection

- Complex business relationships
  - Your customer needs to be reachable by everyone
  - Your provider can't route traffic through you
  - You may not want your traffic through a competitor
  - You may want to dump all your traffic through a competitor
  - You export only customer routes to peers
  - You export peer routes only to your customer
- Hard part:
  - How does BGP realize the *routing policies*?
  - Many mechanisms, including *route import/export policies*

Configuration Semantics

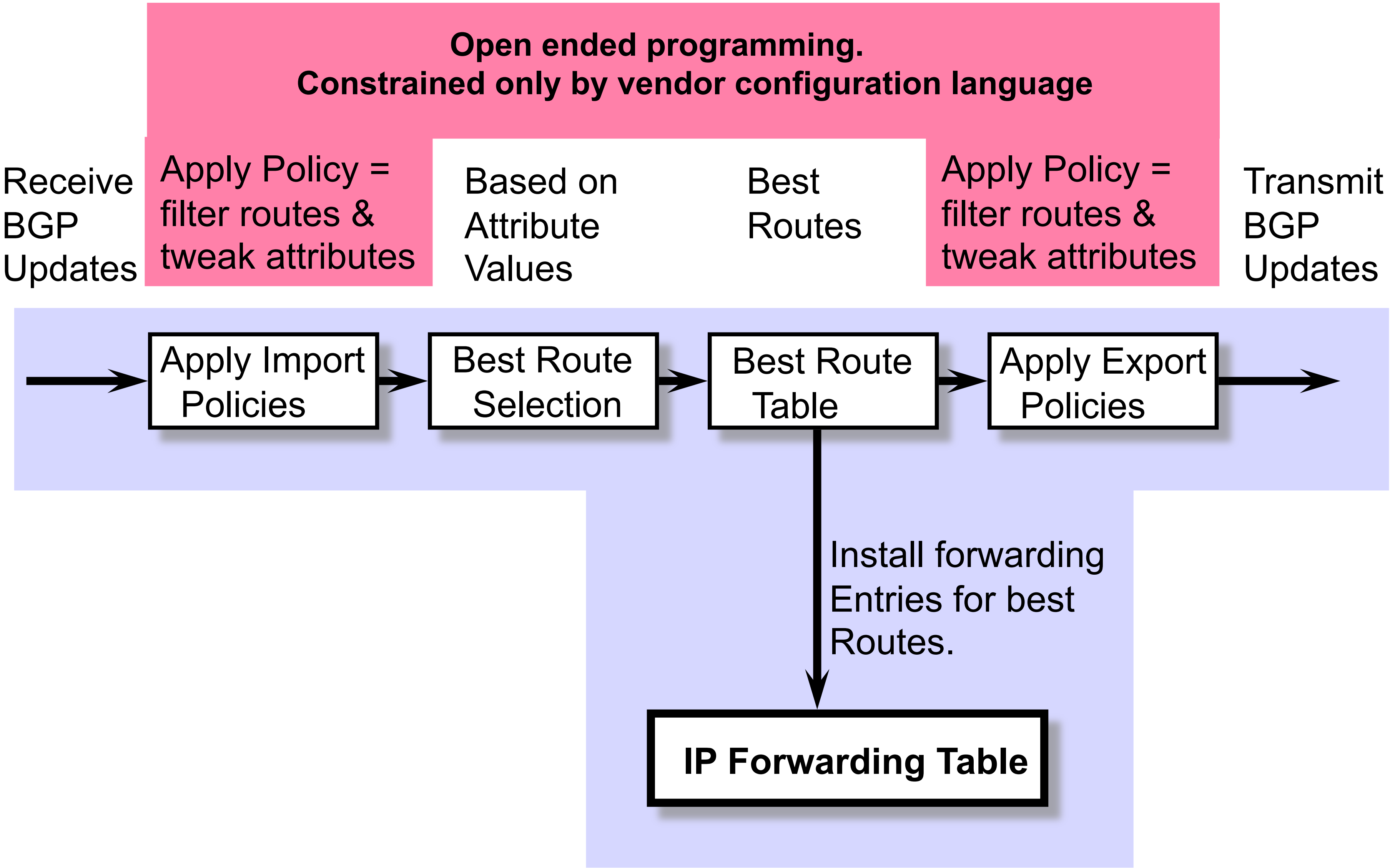


Filtering: route advertisement

Ranking: route selection

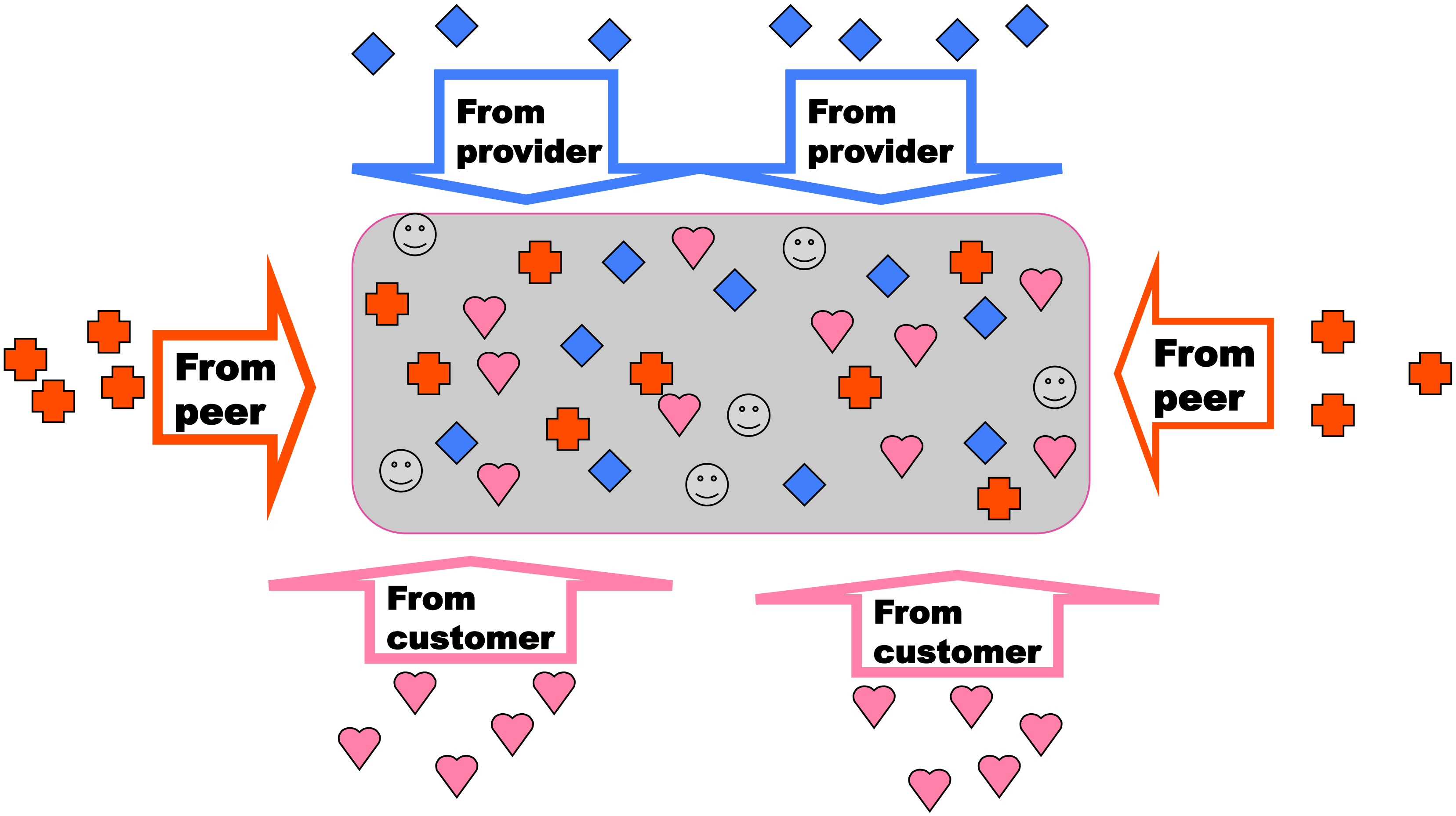
Dissemination: internal route advertisement

BGP Route Processing: Summary


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

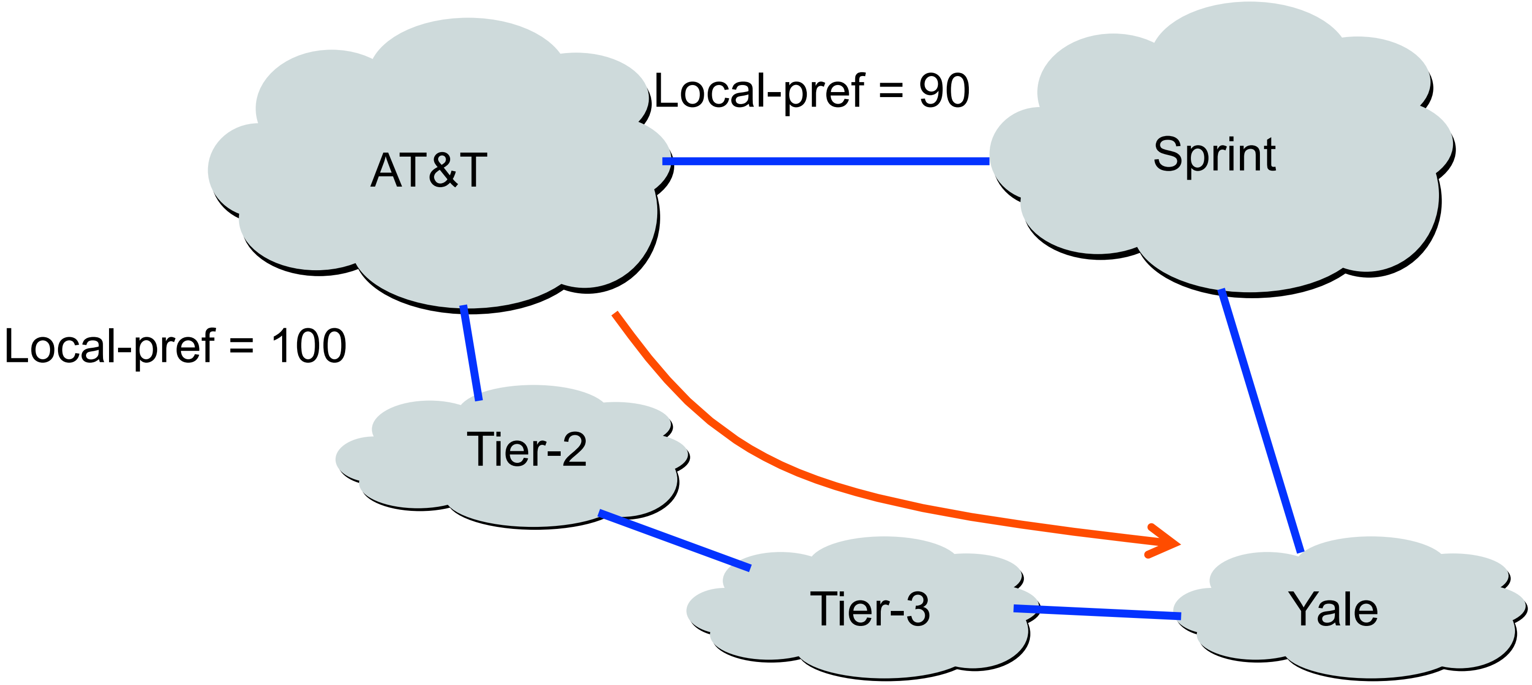
Import Routes

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



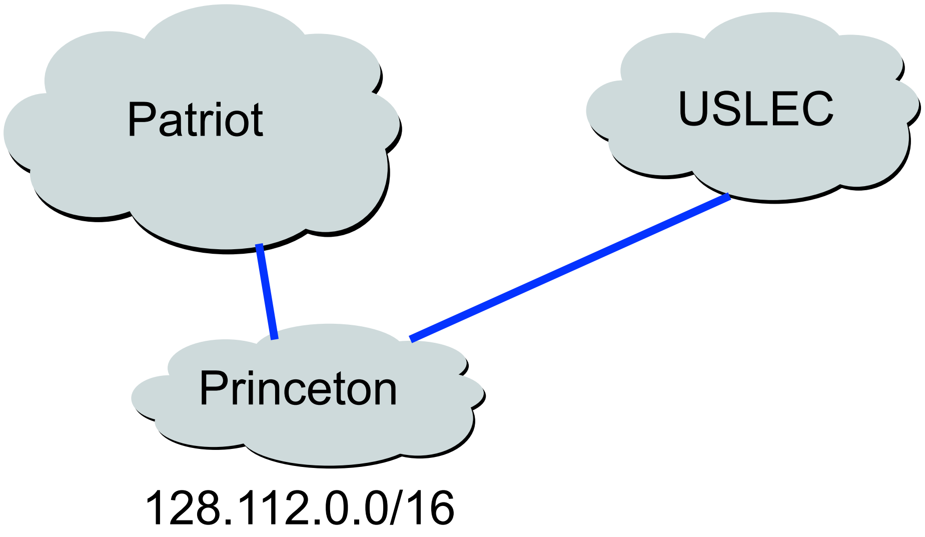
Import Policy: Local Preference

- Favor one path over another
  - Ex: to override the influence of AS path length
- Favor one exit point over another
  - Ex: prefer customer over peer



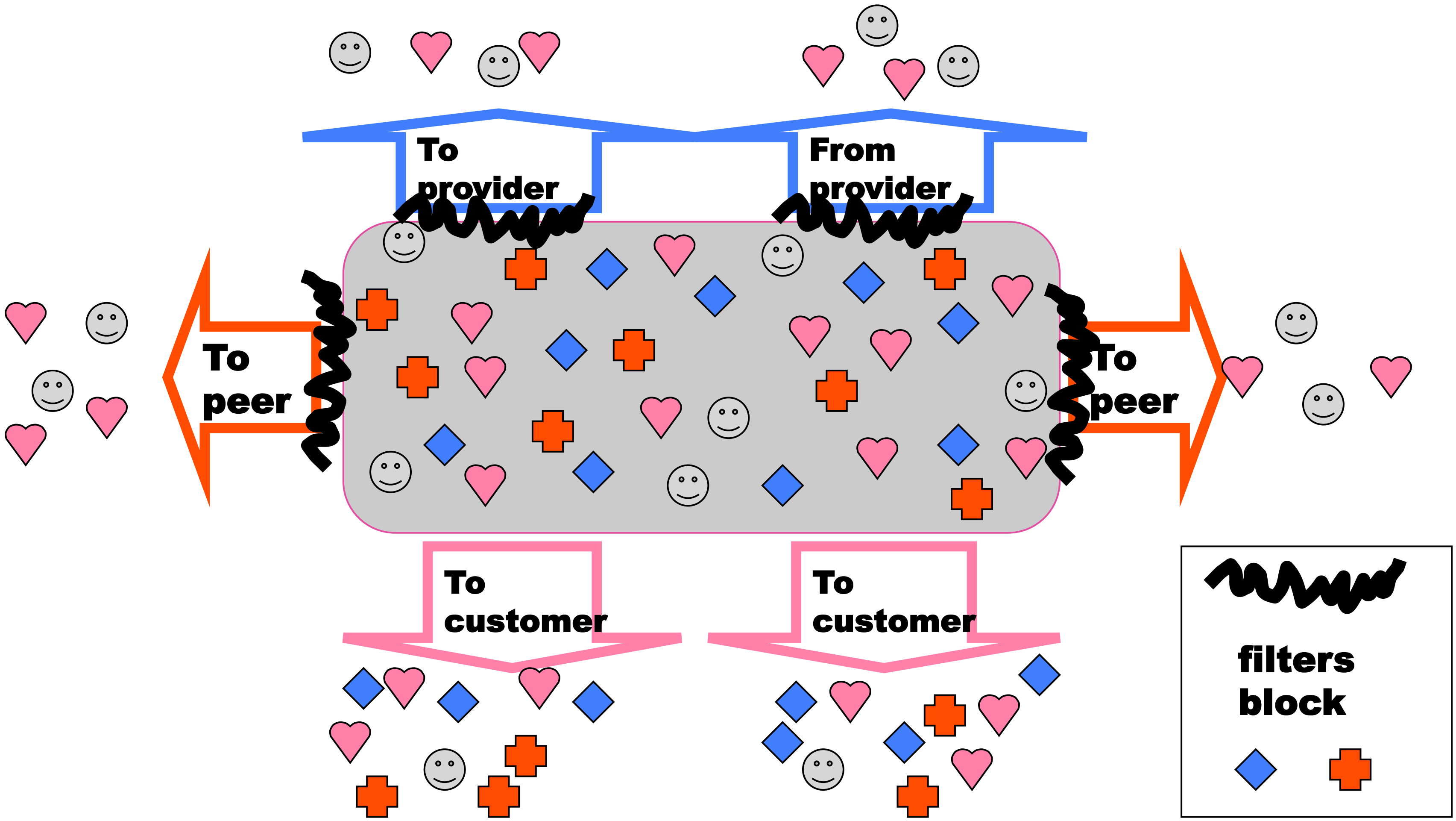
Import Policy: Filtering

- Discard some route announcements
  - E.g., after detecting configuration mistakes and attacks
- Examples on session to a customer
  - Discard route if prefix not owned by the customer
  - Discard route that contains other large ISP in AS path



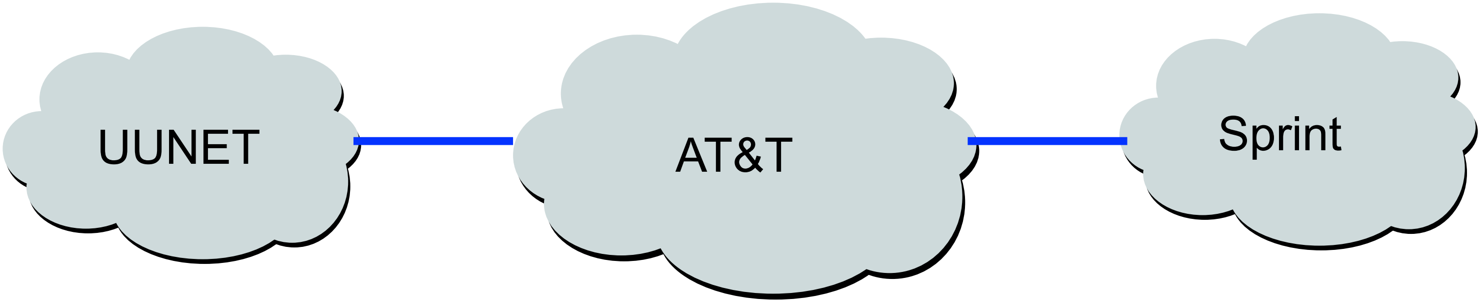
Export Routes

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



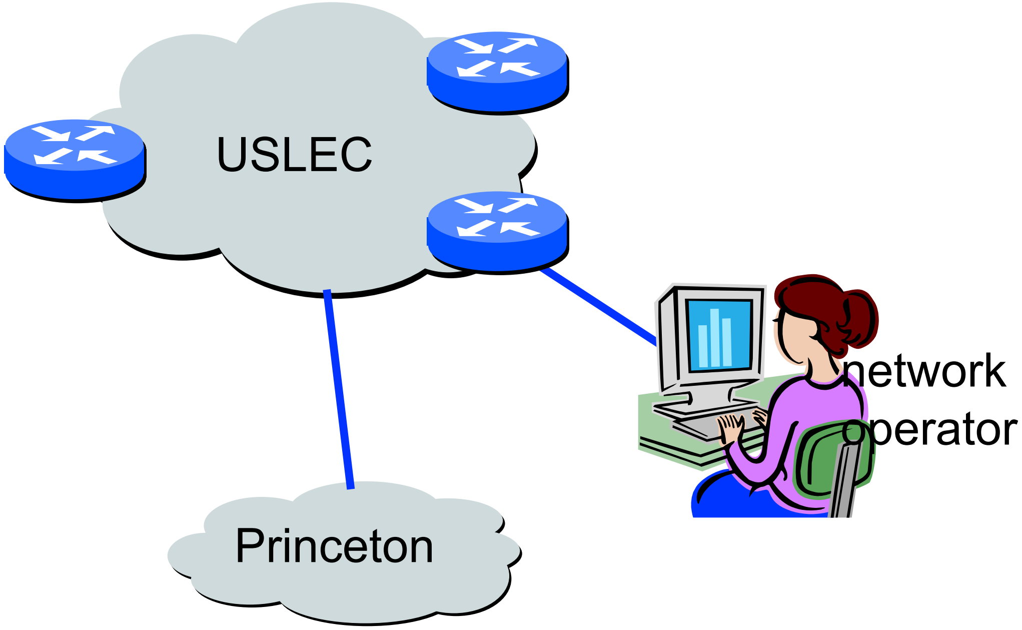
Export Policy: Filtering

- Major criterion: *do not transit packets for free!*
- Examples:
  - Prefer advertisements from customers over all else
  - Don’t announce routes from one peer to another
  - Don’t announce routes from provider to peer



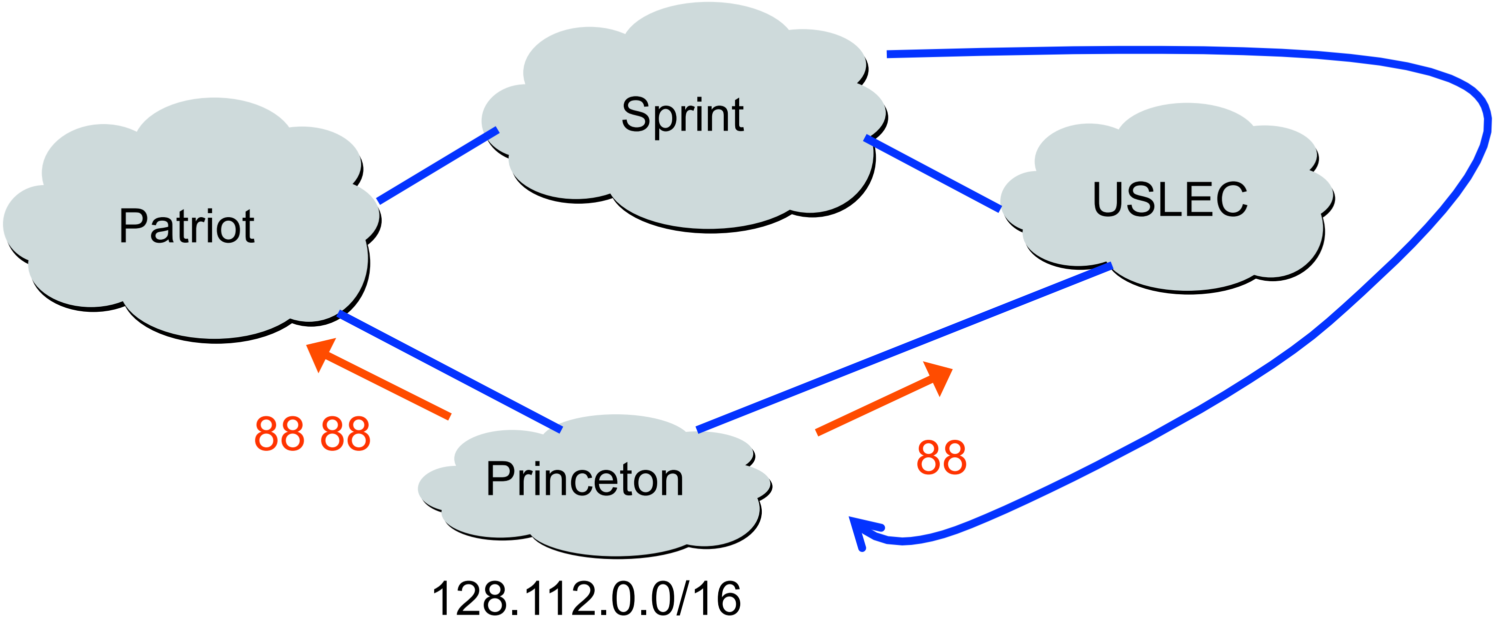
Export Policy: Filtering

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

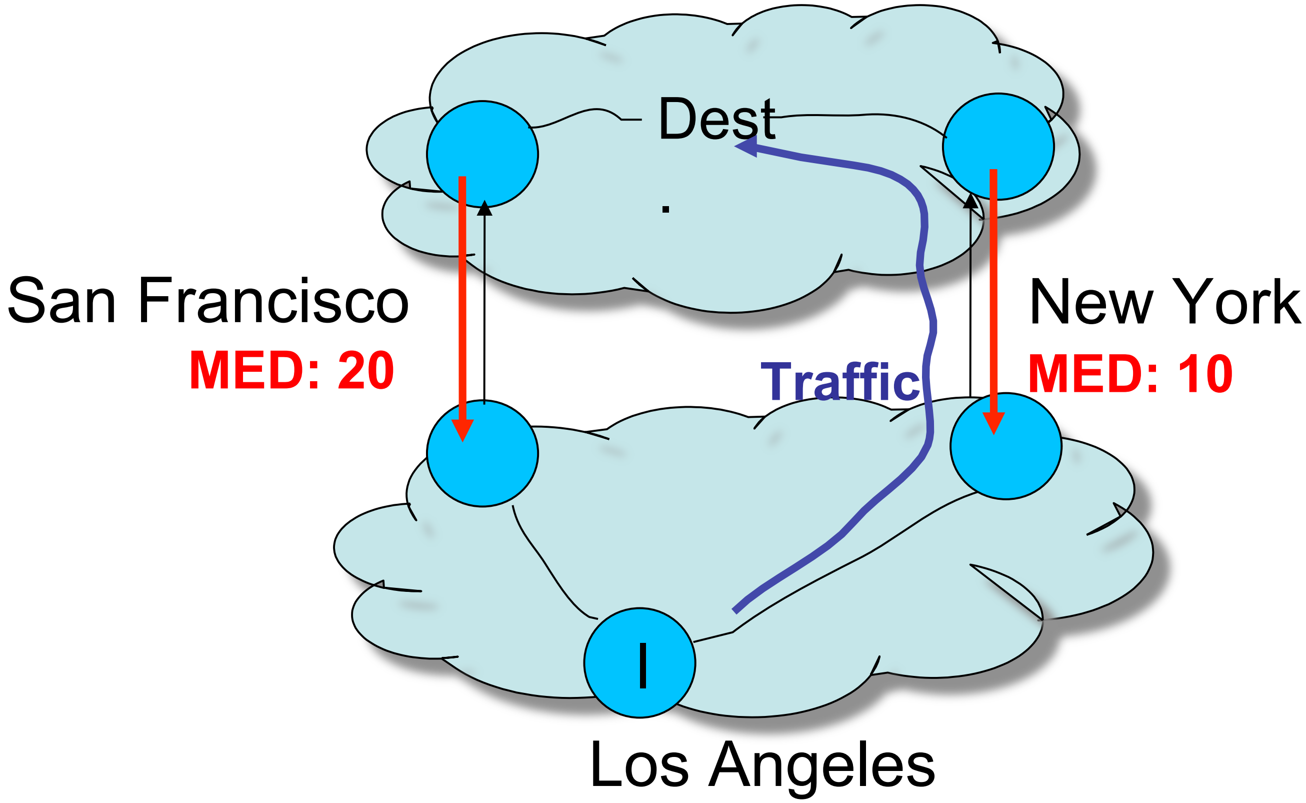


Export Policy: Attribute Manipulation

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

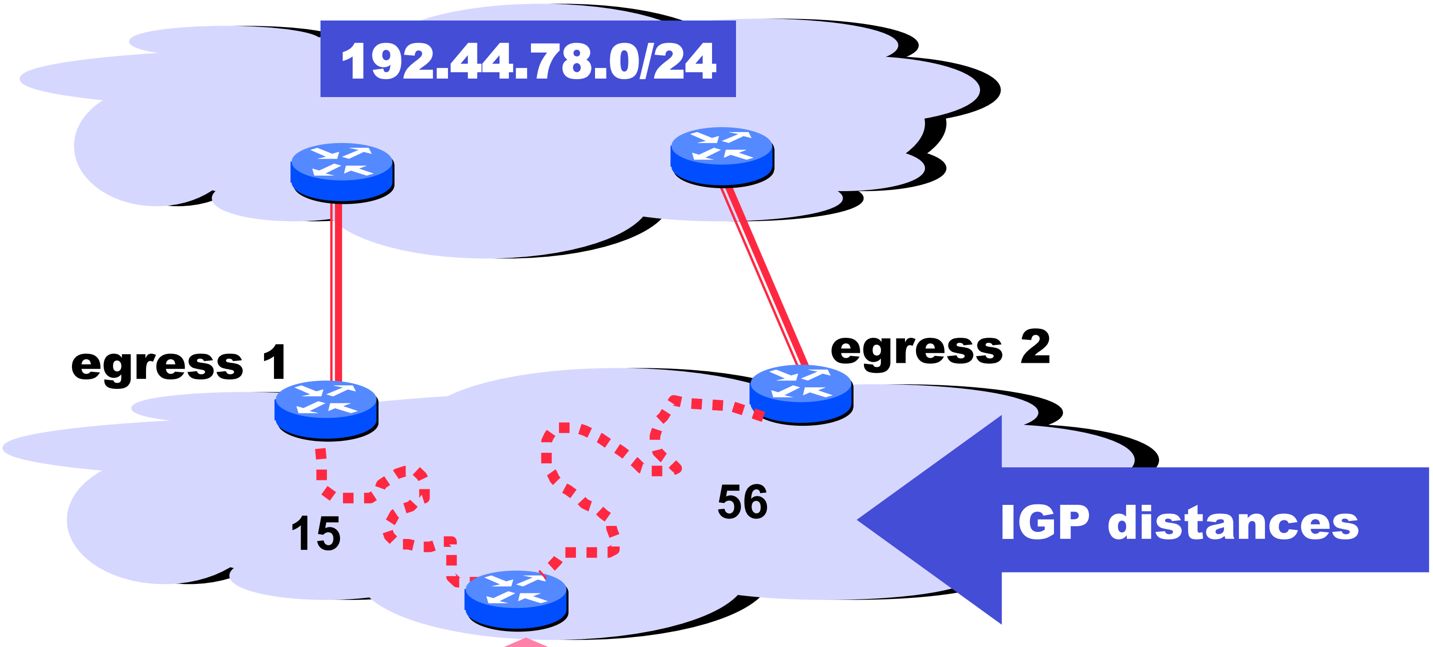


Export Policy: Multi-Exit Discriminator (MED)



- Mechanism for AS to control how traffic enters, given multiple possible entry points.
- Usually ignored when no finance is involved

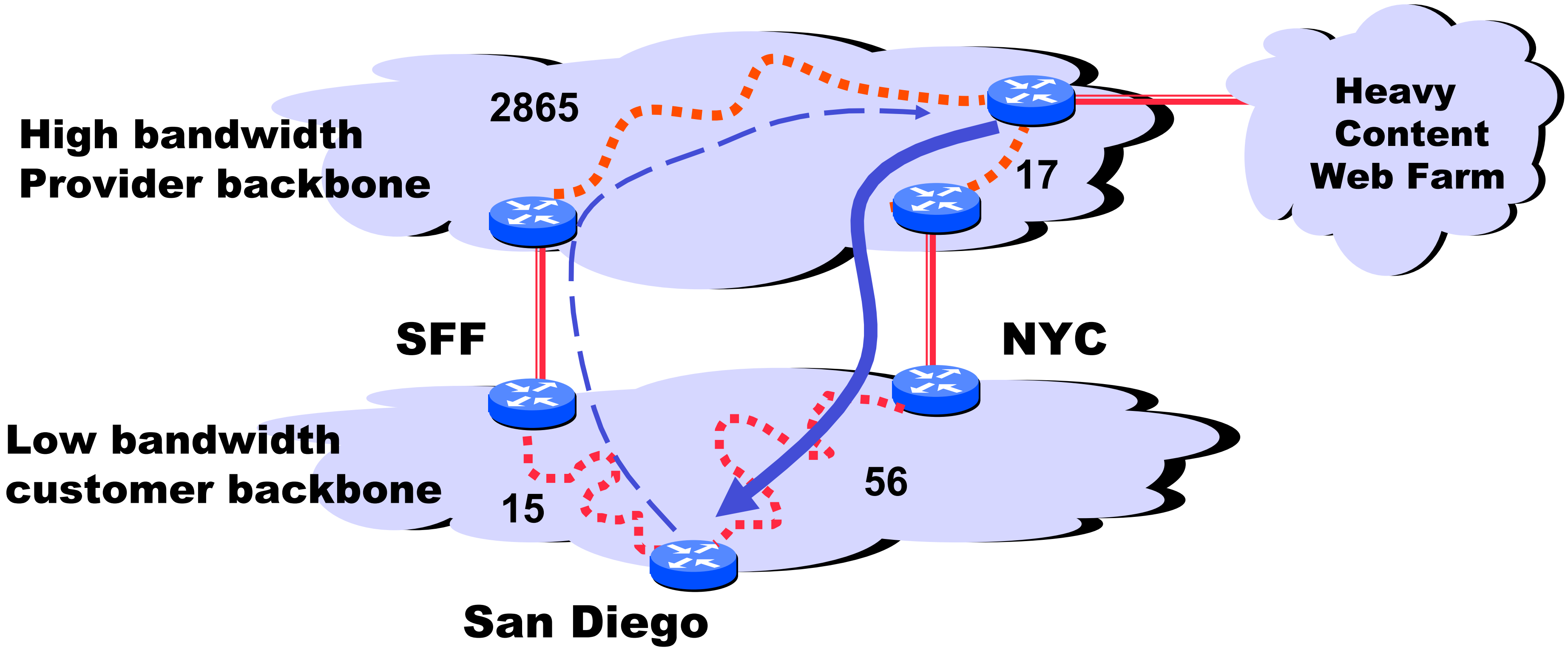
And, There’s The Hot Potato Too



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!

Which Could Burn You

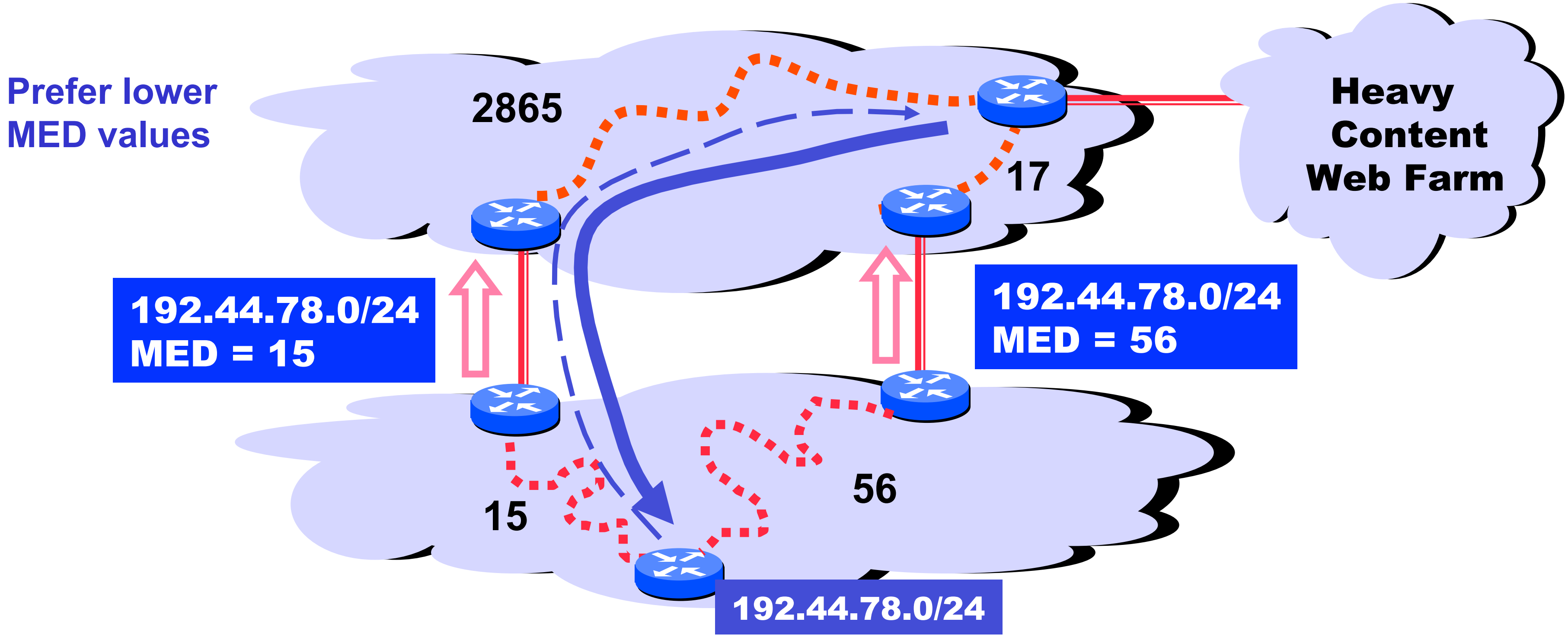


Many customers want their provider to carry the bits!

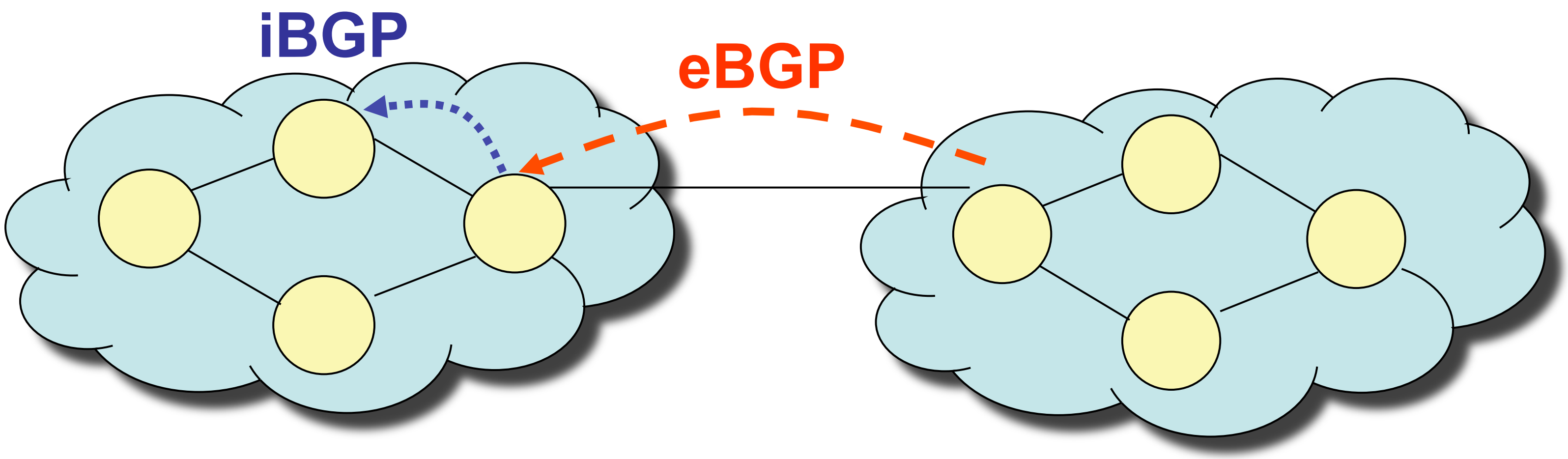
- - - → tiny http request

➡ huge http reply

Cold Potato Routing with MEDs



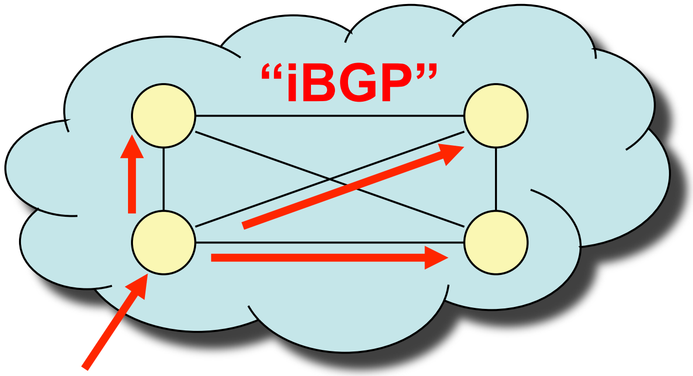
Two "Flavors" of BGP



- **External BGP (eBGP):** exchanging routes *between* ASes
- **Internal BGP (iBGP):** disseminating routes to external destinations among the routers *within an AS*

Internal BGP (iBGP)

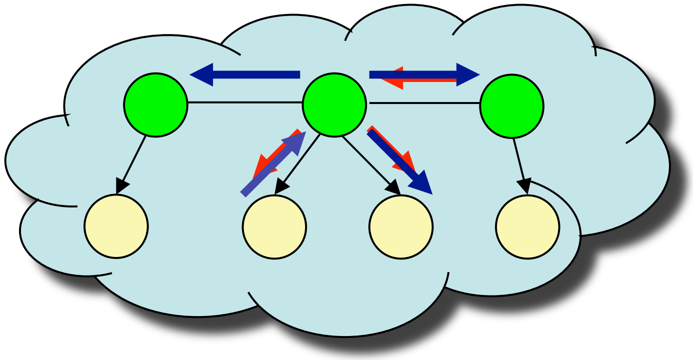
Default: “Full mesh” iBGP.
Doesn’t scale.



Large ASes use **“Route reflection”**

Route reflector:
non-client routes over client sessions;
client routes over all sessions

Client: don’t re-advertise iBGP routes.



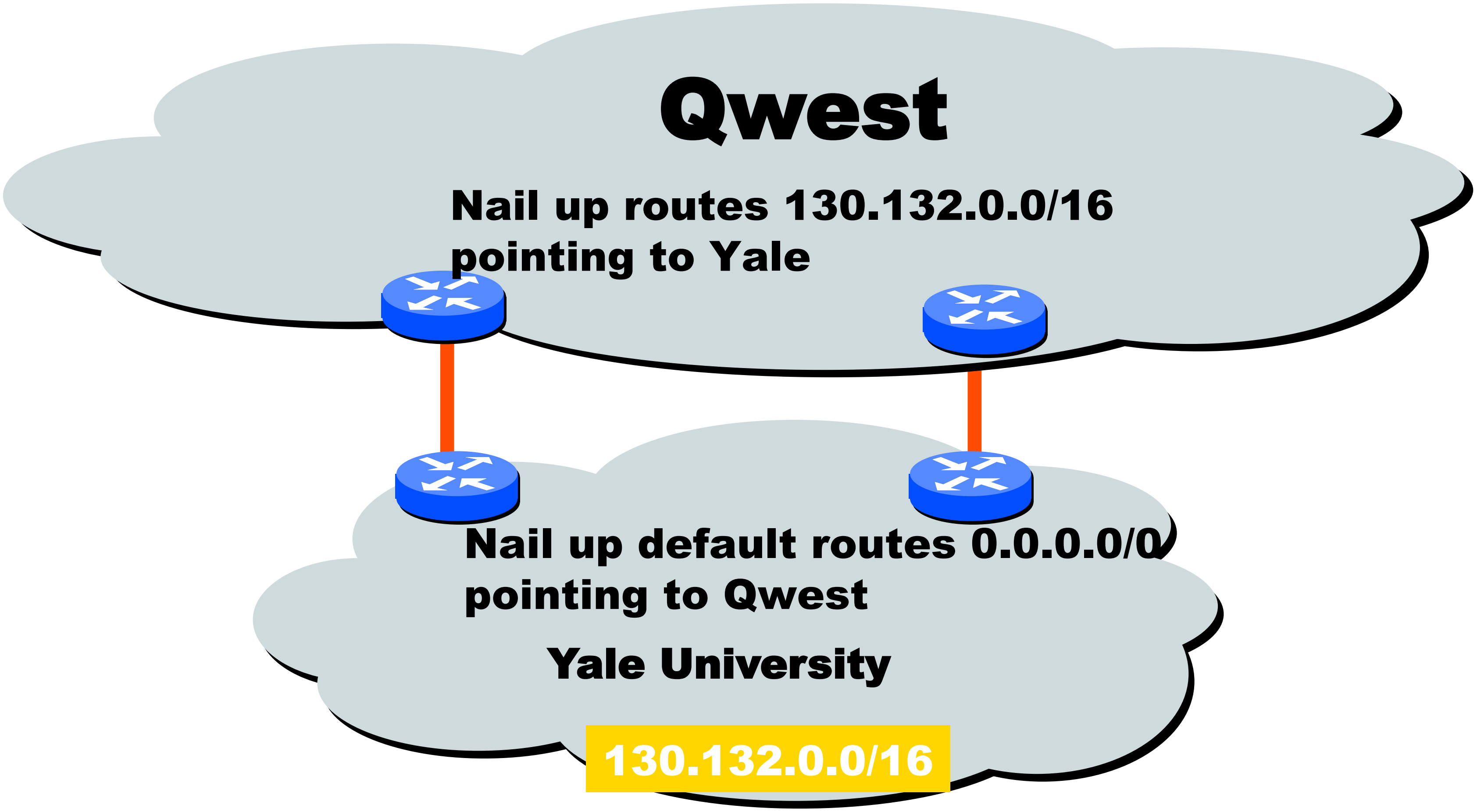
(A Simplified) Route Selection Rule

Priority	Rule	Remarks
1	LOCAL_PREF	Highest preferred
2	AS_PATH	Shortest preferred
3	MED	Lowest preferred
4	eBGP > iBGP	Did AS learn route via eBGP or iBGP
5	IGP path	Lower cost preferred
6	Router ID	Smaller preferred or random

BGP Policy Configuration

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

Don’t Always Need BGP!!!



Nail up routes 130.132.0.0/16 pointing to Yale

Nail up default routes 0.0.0.0/0 pointing to Qwest

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 …