Interdomain Routing

Nick Feamster
CS 7260
January 22, 2007

Administrivia

- PS 1 will go out tonight (3 problems).

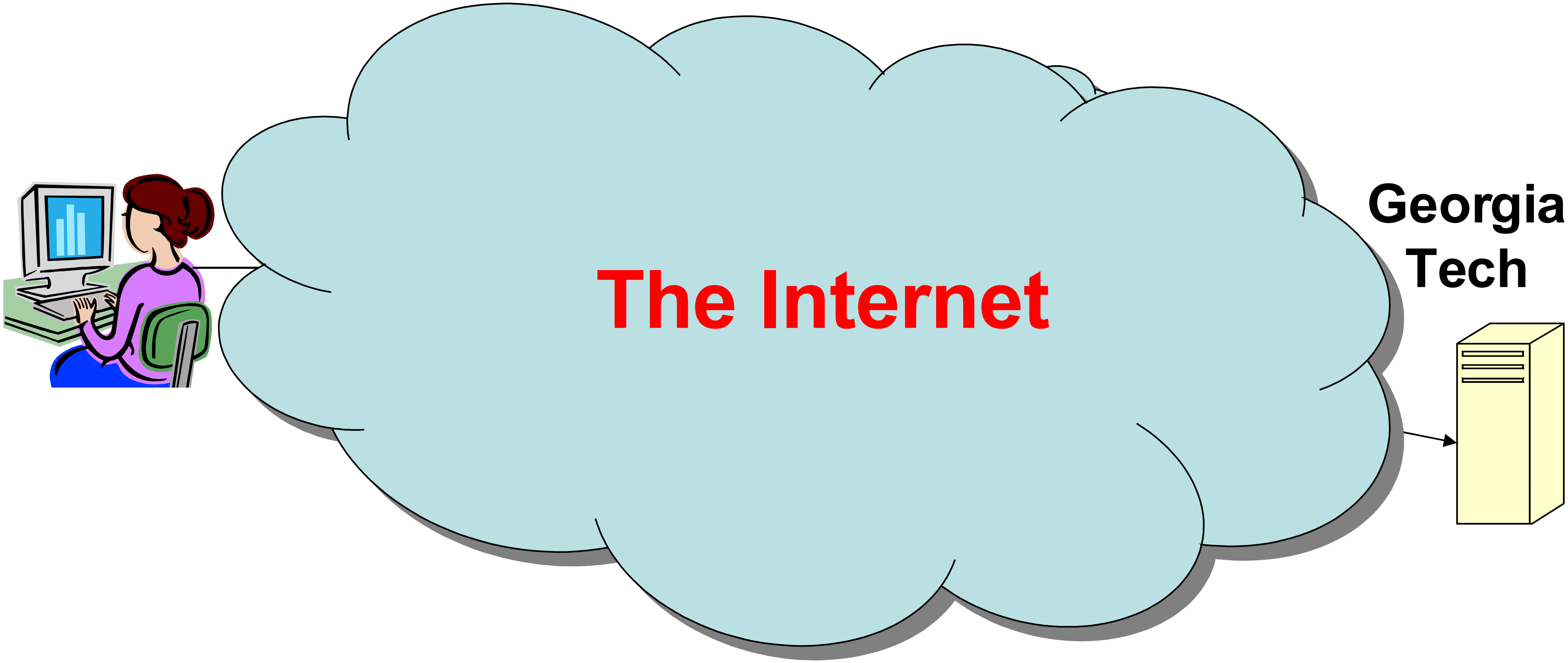
- Send project groups by Wednesday.

Today’s Lecture: Interdomain Routing

- Today’s interdomain routing protocol: BGP
  - BGP route attributes
    - Usage
    - Problems
  - Business relationships
- Today’s Paper: *Stable Internet Routing without Global Coordination*
  - Main ideas
  - Extensions

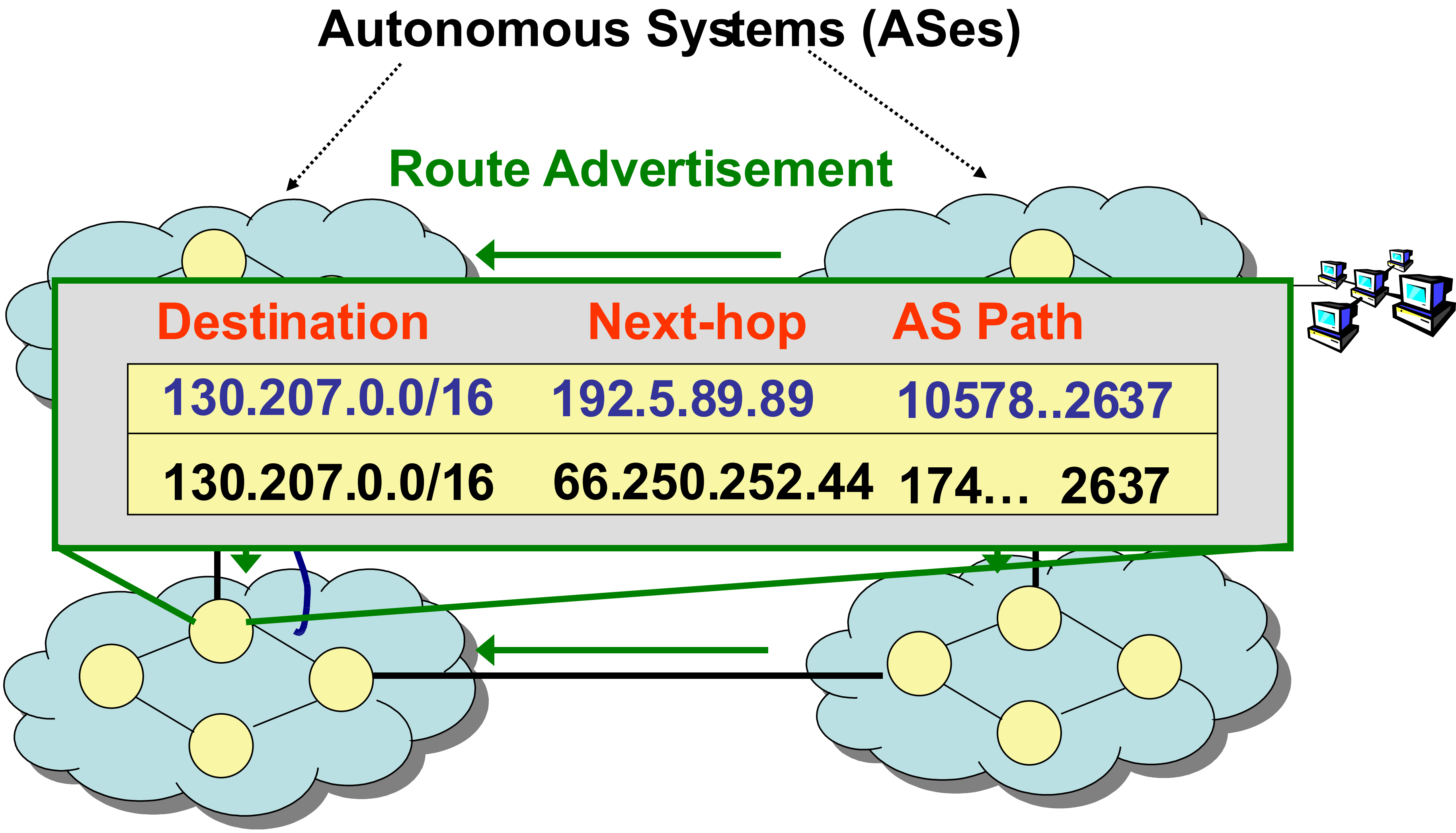
See http://nms.lcs.mit.edu/~feamster/papers/dissertation.pdf (Chapter 2.1-2.3) for good coverage of today’s topics.

Internet Routing



- **Large-scale:** Thousands of autonomous networks
- **Self-interest:** Independent economic and performance objectives
- But, must cooperate for global connectivity

Internet Routing Protocol: BGP

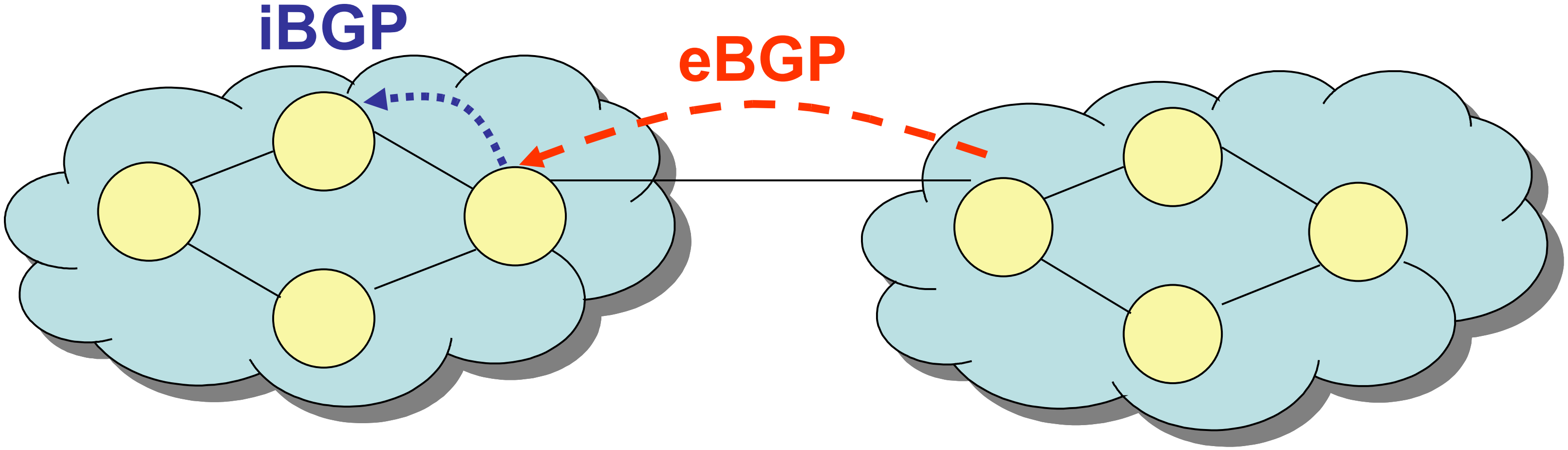


Autonomous Systems (ASes)

Route Advertisement

Destination	Next-hop	AS Path
130.207.0.0/16	192.5.89.89	10578..2637
130.207.0.0/16	66.250.252.44	174… 2637

Two Flavors of BGP

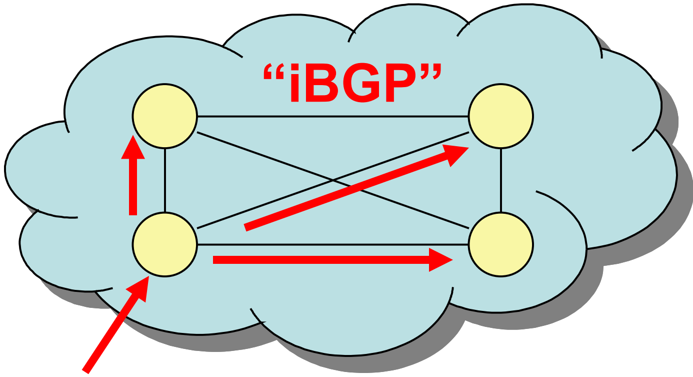


- **External BGP (eBGP):** exchanging routes *between* ASes
- **Internal BGP (iBGP):** disseminating routes to external destinations among the routers *within an AS*

Question: **What’s the difference between IGP and iBGP?**

Internal BGP (iBGP)

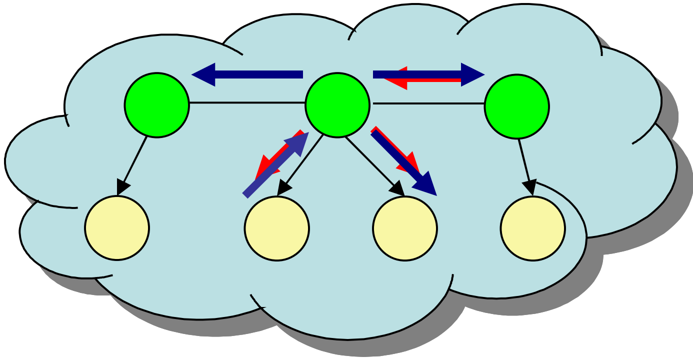
Default: “Full mesh” iBGP.
Doesn’t scale.



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.



Example BGP Routing Table

The full routing table

```
> show ip bgp

 Network             Next Hop          Metric LocPrf Weight Path
*>i3.0.0.0           4.79.2.1               0    110      0 3356 701 703 80 i
*>i4.0.0.0           4.79.2.1               0    110      0 3356 i
*>i4.21.254.0/23     208.30.223.5          49    110      0 1239 1299 10355 10355 i
* i4.23.84.0/22      208.30.223.5         112    110      0 1239 6461 20171 i
```

Specific entry. Can do longest prefix lookup:

```
> show ip bgp 130.207.7.237
BGP routing table entry for 130.207.0.0/16
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Not advertised to any peer
  10578 11537 10490 2637
    192.5.89.89 from 18.168.0.27 (66.250.252.45)
      Origin IGP, metric 0, localpref 150, valid, internal, best
      Community: 10578:700 11537:950
      Last update: Sat Jan 14 04:45:09 2006
```

Prefix

AS path

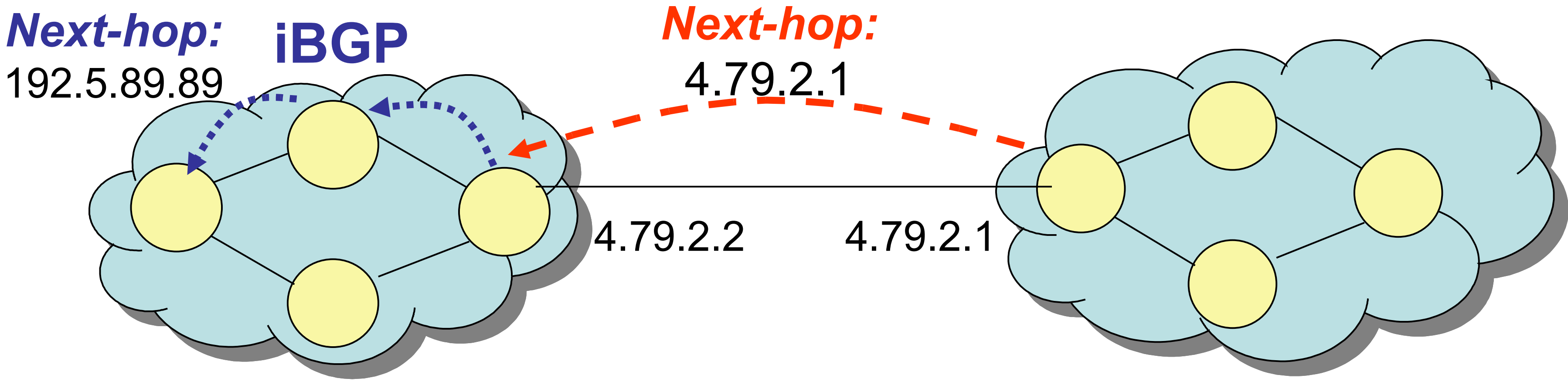
Next-hop

Routing Attributes and Route Selection

BGP routes have the following attributes, on which the route selection process is based:

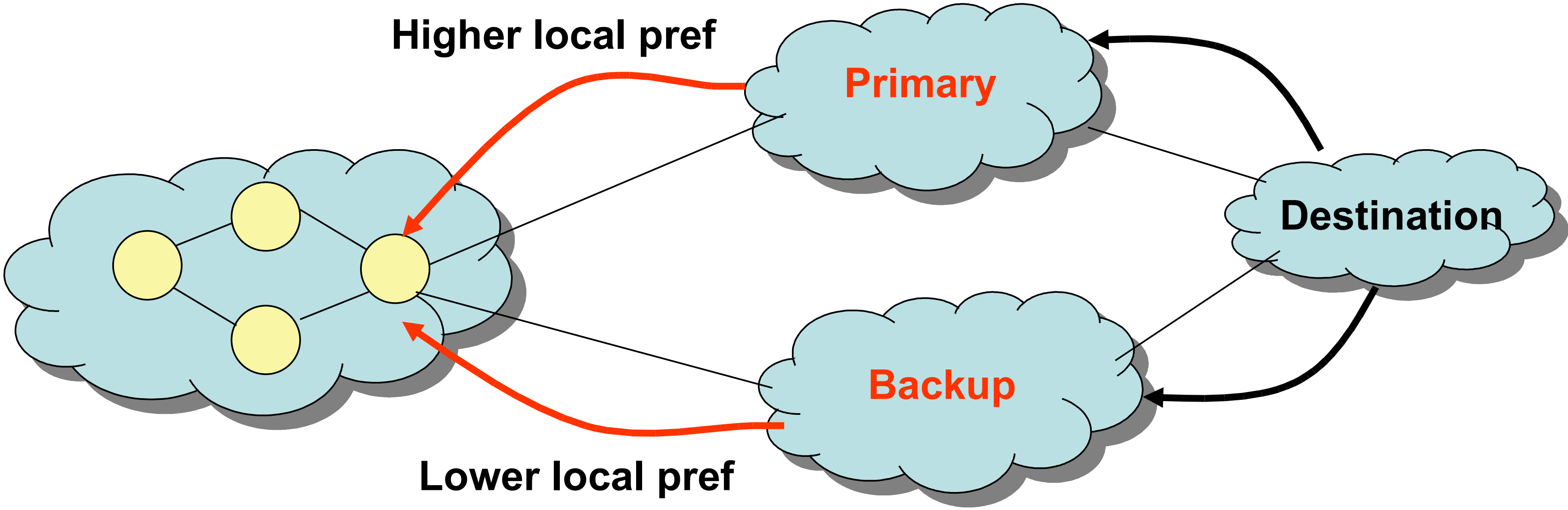
- **Local preference:** numerical value assigned by routing policy. Higher values are more preferred.
- **AS path length:** number of AS-level hops in the path
- **Multiple exit discriminator ("MED"):** allows one AS to specify that one exit point is more preferred than another. Lower values are more preferred.
- **Shortest IGP path cost to next hop:** implements "hot potato" routing
- **Router ID tiebreak:** arbitrary tiebreak, since only a single "best" route can be selected

Other BGP Attributes



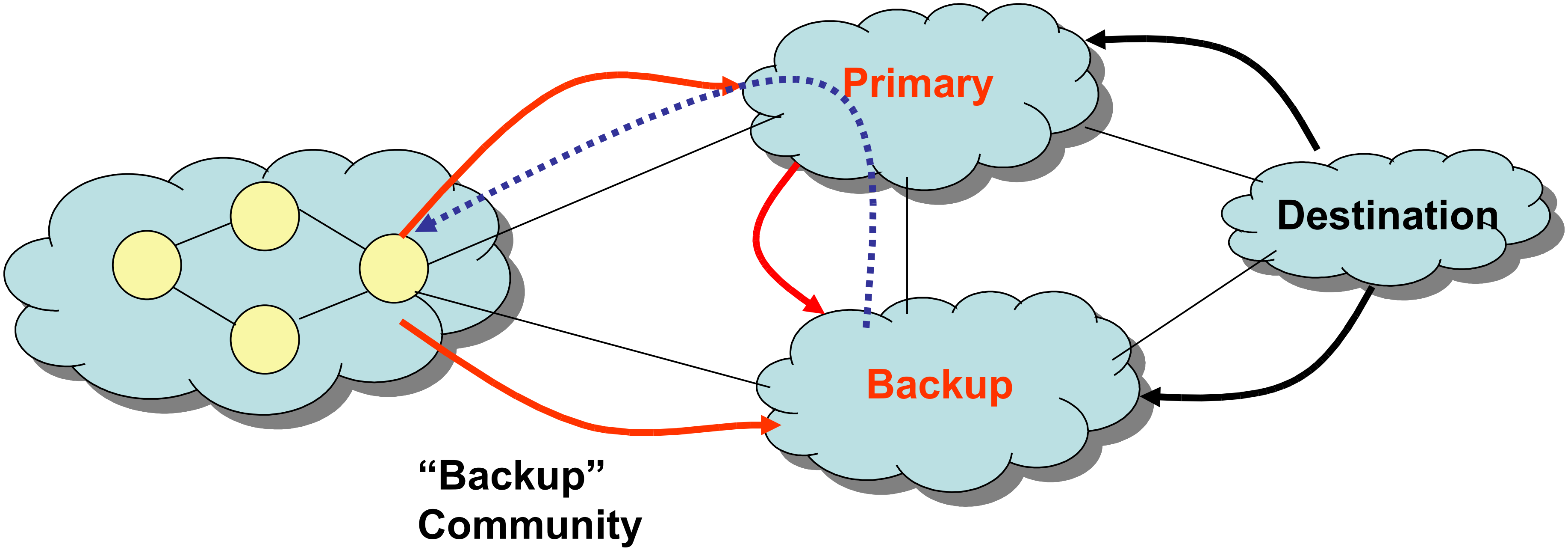
- **Next-hop:** IP address to send packets en route to destination. (*Question:* How to ensure that the next-hop IP address is reachable?)
- **Community value:** Semantically meaningless. Used for passing around “signals” and labelling routes. More in a bit.

Local Preference



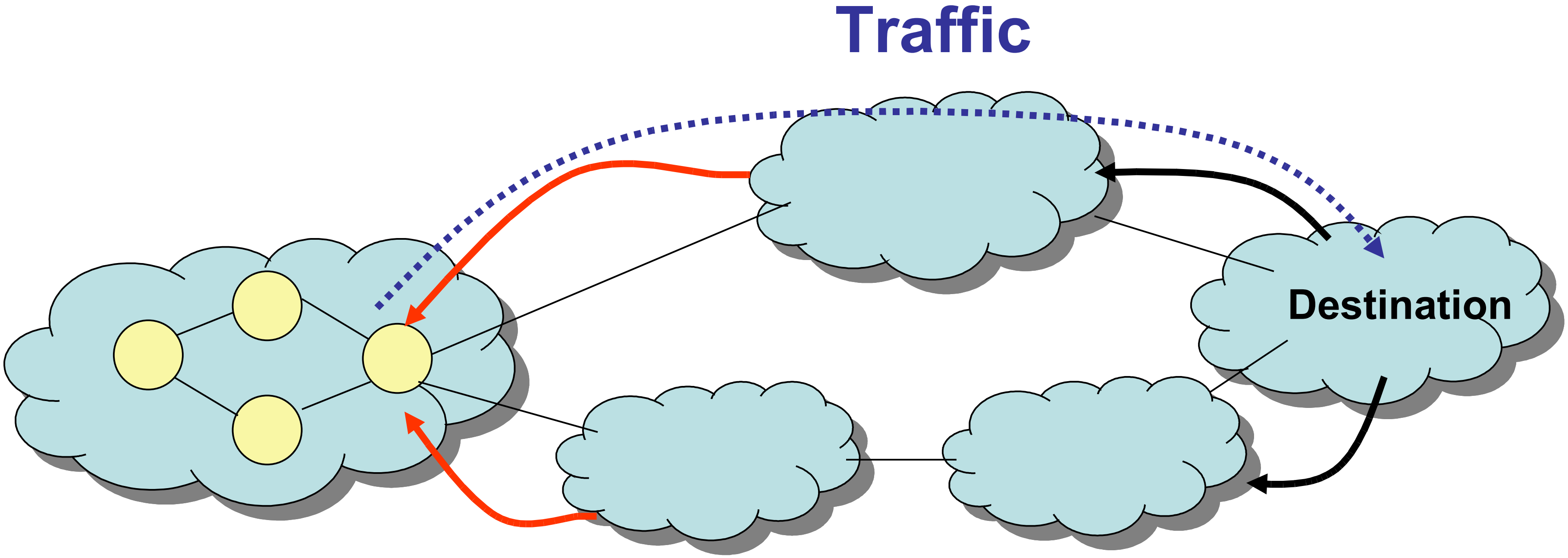
- **Control over *outbound* traffic**
- *Not* transitive across ASes
- Coarse hammer to implement route preference
- Useful for preferring routes from one AS over another (*e.g.*, primary-backup semantics)

Communities and Local Preference



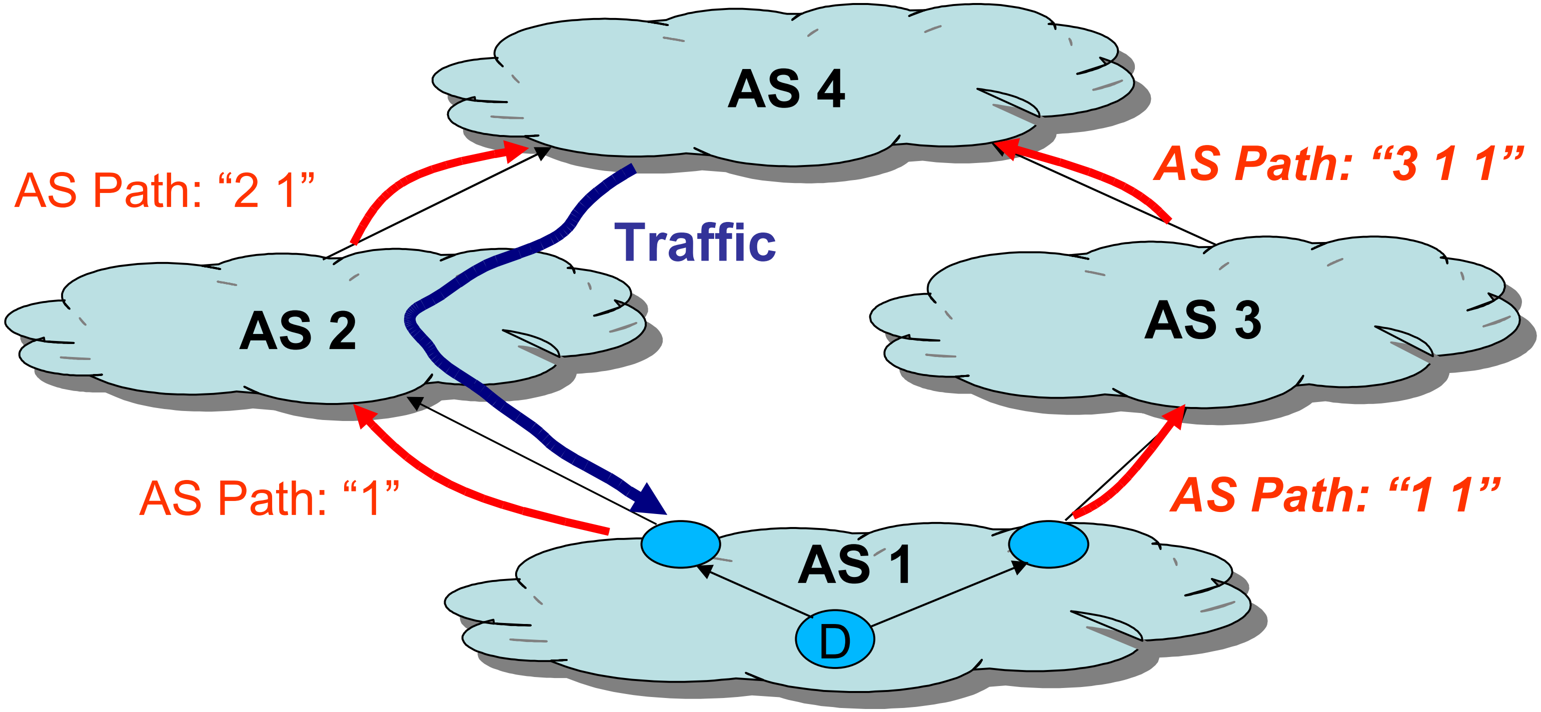
- Customer expresses provider that a link is a backup
- Affords *some* control over inbound traffic
- More on multihoming, traffic engineering in Lecture 7

AS Path Length



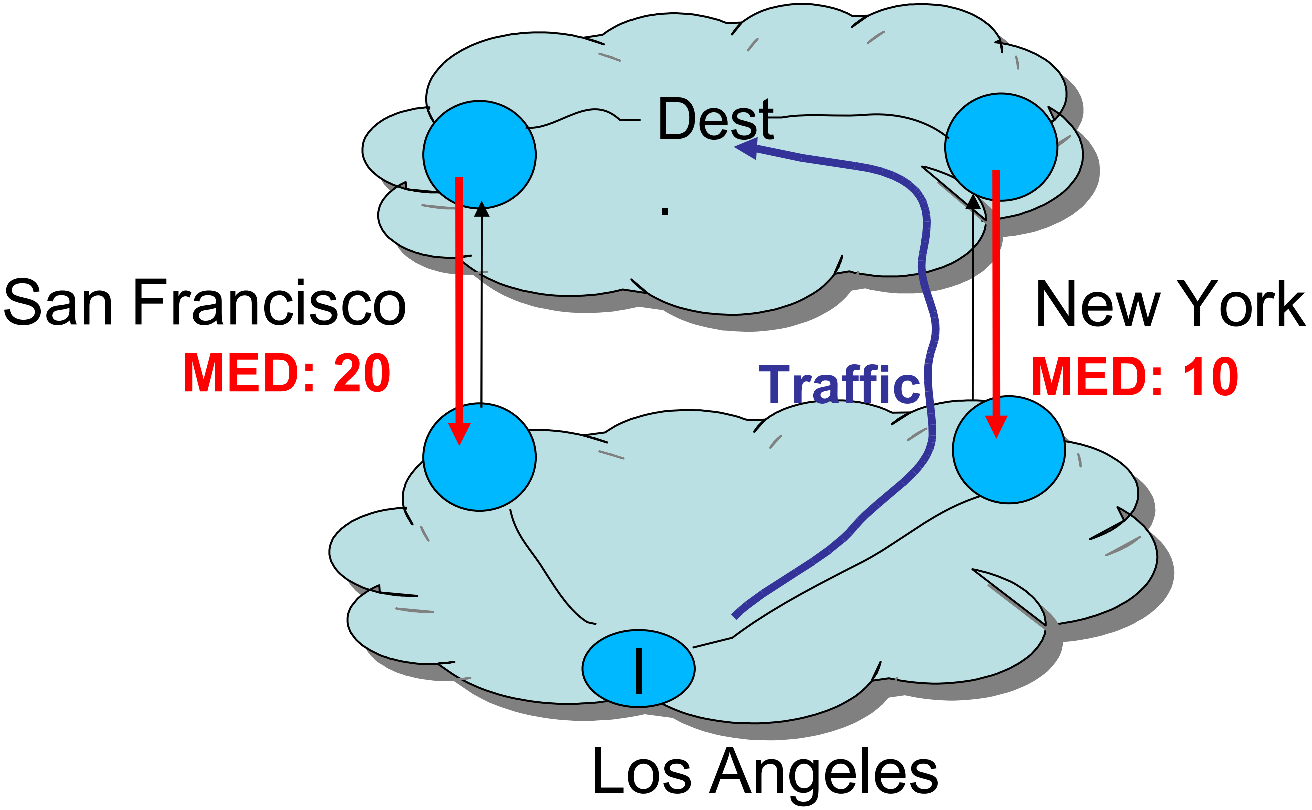
- Among routes with highest local preference, select route with shortest AS path length
- Shortest AS path != shortest path, for *any* interpretation of “shortest path”

AS Path Length Hack: Prepending



- Attempt to control inbound traffic
- Make AS path length look artificially longer
- How well does this work in practice vs. *e.g.*, hacks on longest-prefix match?

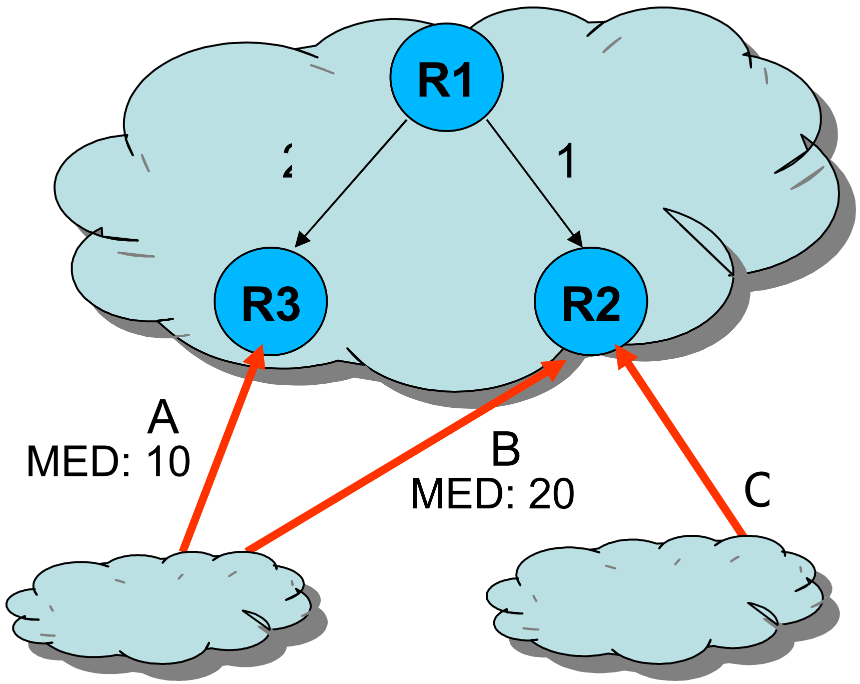
Multiple Exit Discriminator (MED)



- Mechanism for AS to control how traffic enters, given multiple possible entry points.

Problems with MED

- **Safety:** No persistent oscillations
  - Routing system should “settle down”, assuming the system’s inputs are not changing

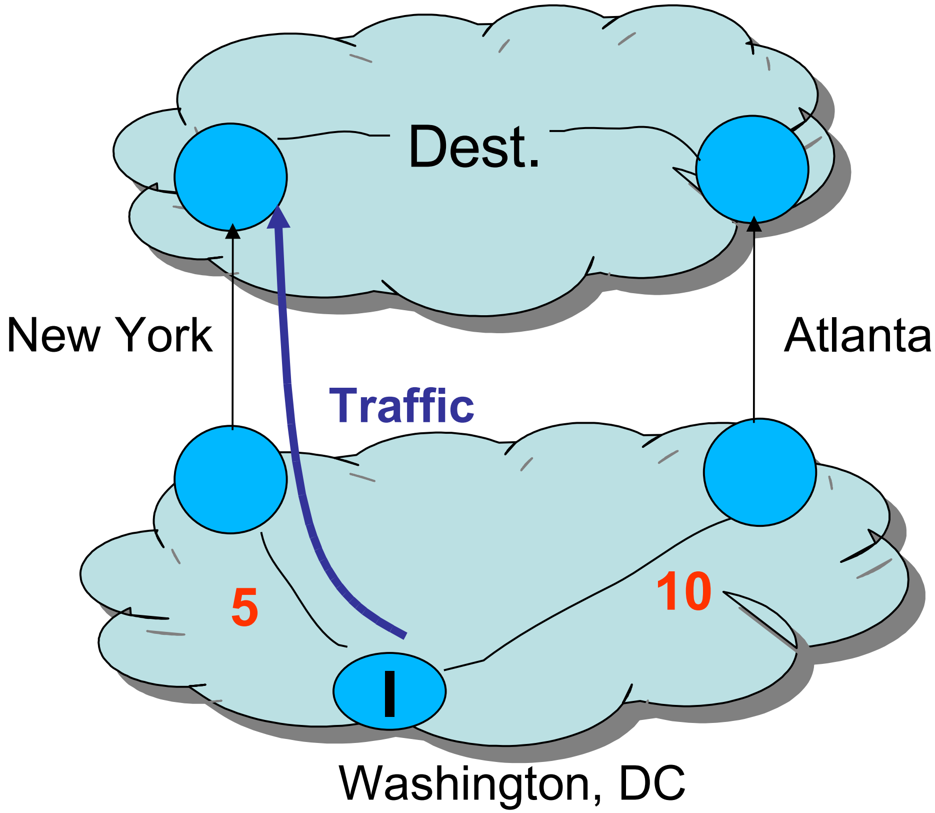


- R3 selects *A*
- R1 advertises *A* to R2
- R2 selects *C*
- R1 selects *C*
  - (R1 withdraws *A* from R2)
- R2 selects *B*
  - (R2 withdraws *C* from R1)
- R1 selects *A*, advertises to R2

Preference between *B* and *C* at R2 depends on presence or absence of A.

Hot-Potato Routing

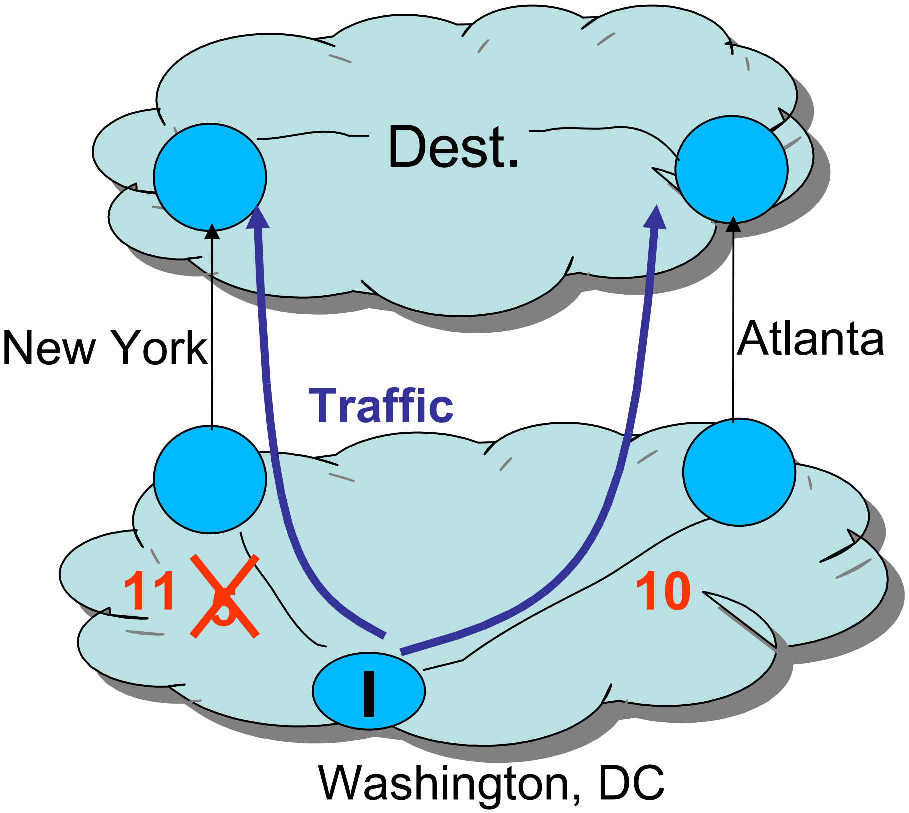
- Prefer route with shorter IGP path cost to next-hop
- *Idea:* traffic leaves AS as quickly as possible



Common practice: Set IGP weights in accordance with propagation delay (*e.g.*, miles, etc.)

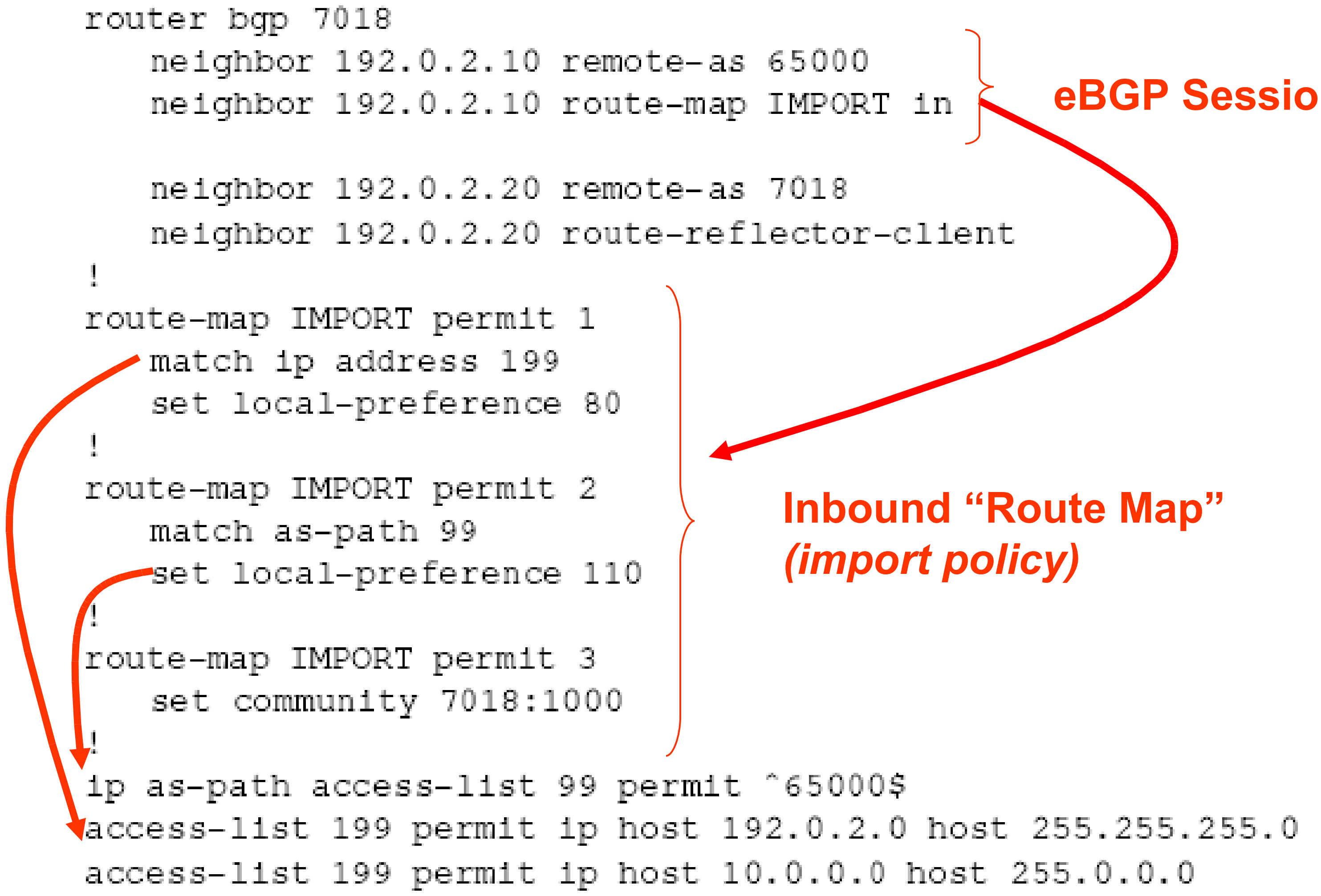
Problems with Hot-Potato Routing

- Small changes in IGP weights can cause large traffic shifts



Question: Cost of sub-optimal exit vs. cost of large traffic shifts

What policy looks like in Cisco IOS



```
router bgp 7018
   neighbor 192.0.2.10 remote-as 65000
   neighbor 192.0.2.10 route-map IMPORT in

   neighbor 192.0.2.20 remote-as 7018
   neighbor 192.0.2.20 route-reflector-client
!
route-map IMPORT permit 1
   match ip address 199
   set local-preference 80
!
route-map IMPORT permit 2
   match as-path 99
   set local-preference 110
!
route-map IMPORT permit 3
   set community 7018:1000
!
ip as-path access-list 99 permit ^65000$
access-list 199 permit ip host 192.0.2.0 host 255.255.255.0
access-list 199 permit ip host 10.0.0.0 host 255.0.0.0
```

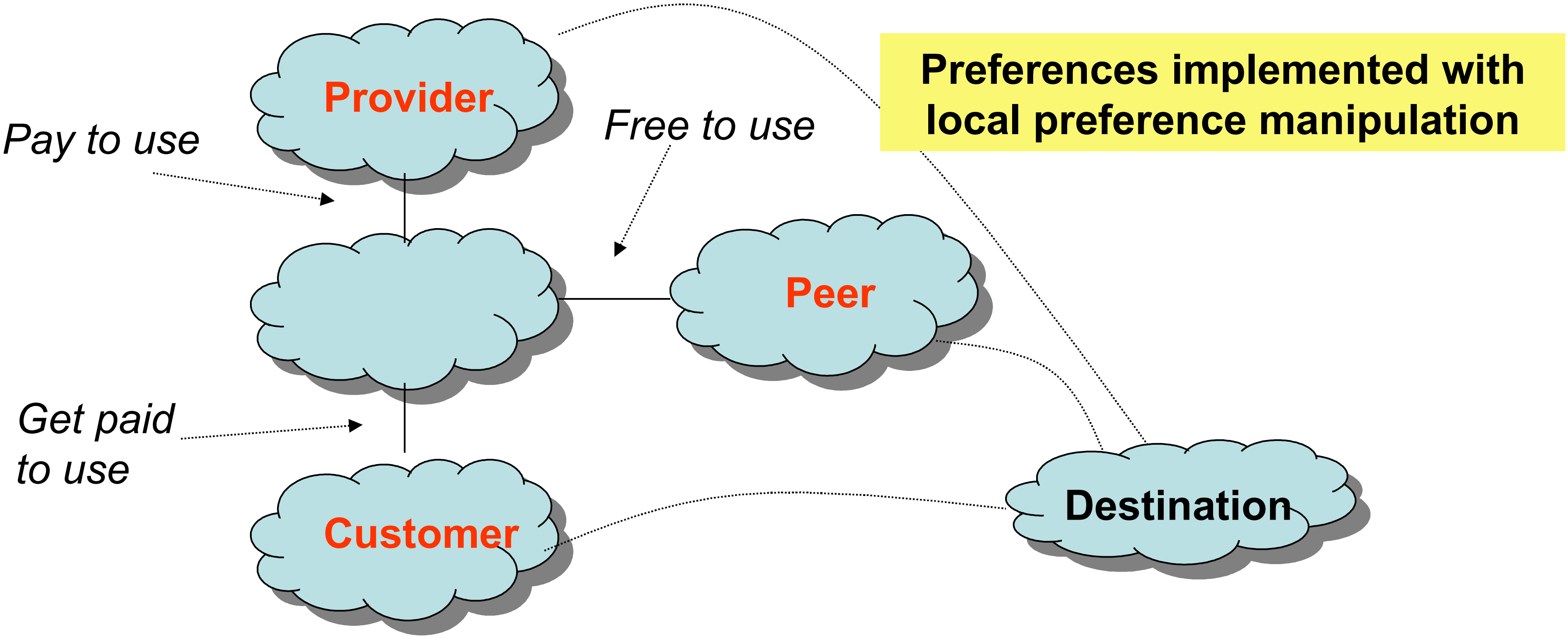
eBGP Session

Inbound “Route Map”
(import policy)

General Problems with BGP

- **Convergence**
- **Security**
  - Too easy to "steal" IP address space
    - http://www.renesys.com/blog/2006/01/coned_steals_the_net.shtml
    - Regular examples of suspicious activity (see Internet Alert Registry)
  - Hard to check veracity of information (*e.g.*, AS path)
  - Can't tell where data traffic is actually going to go
- **Broken business models**
  - "Depeering" and degraded connectivity: universal connectivity depends on cooperation. *No guarantees!*
- **Policy interactions**
  - Oscillations (*e.g.*, today's paper)

Internet Business Model (Simplified)



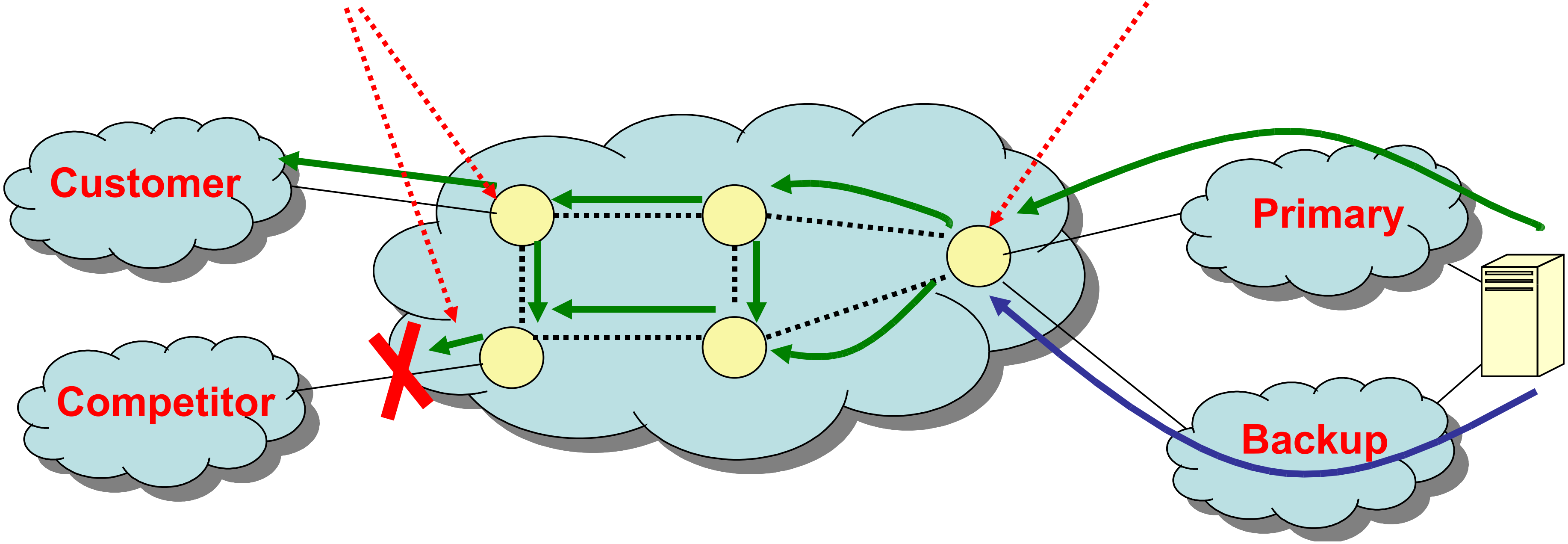
Preferences implemented with local preference manipulation

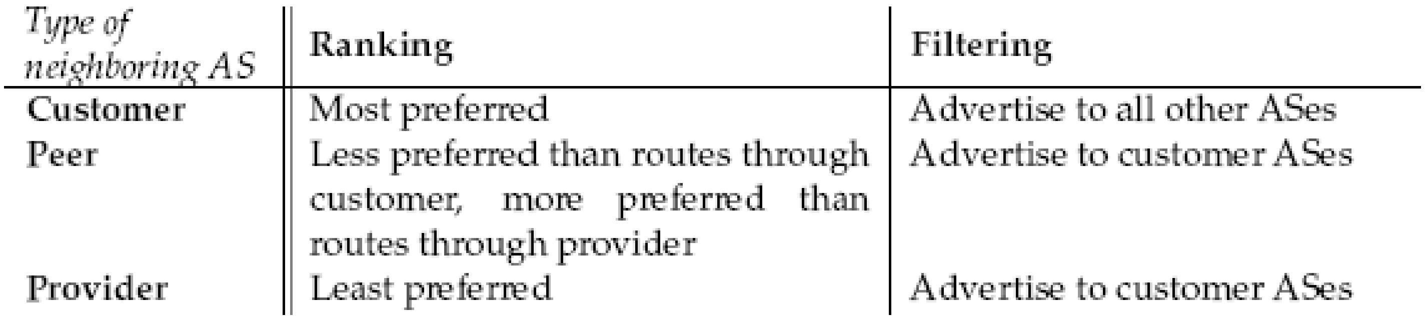
- **Customer/Provider:** One AS pays another for reachability to some set of destinations
- **"Settlement-free" Peering:** Bartering. Two ASes exchange routes with one another.

Filtering and Rankings

Filtering: route advertisement

Ranking: route selection





Type of neighboring AS	Ranking	Filtering
Customer	Most preferred	Advertise to all other ASes
Peer	Less preferred than routes through customer, more preferred than routes through provider	Advertise to customer ASes
Provider	Least preferred	Advertise to customer ASes

The Business Game and Depeering

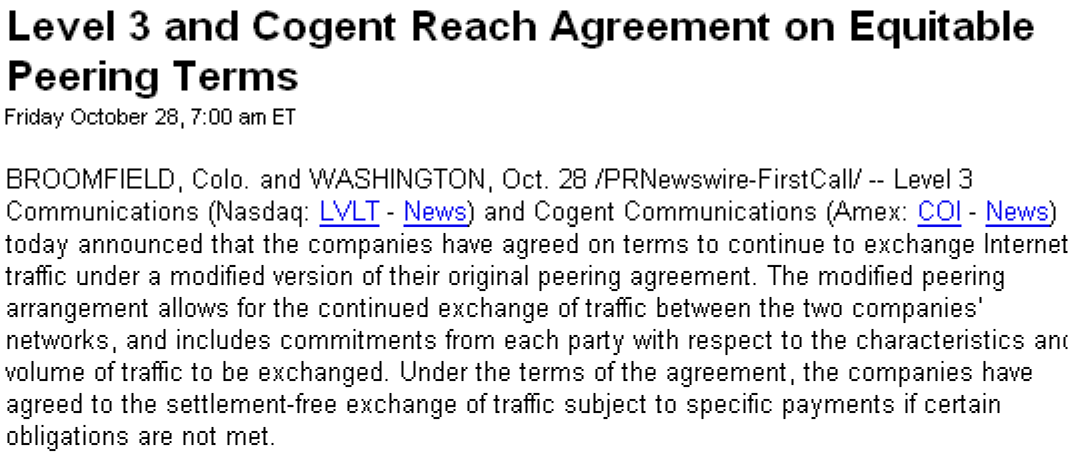
- Cooperative competition (brinksmanship)
- Much more desirable to have your peer's customers
  - Much nicer to get paid for transit
- Peering "tiffs" are relatively common

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 Continued

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 and 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.

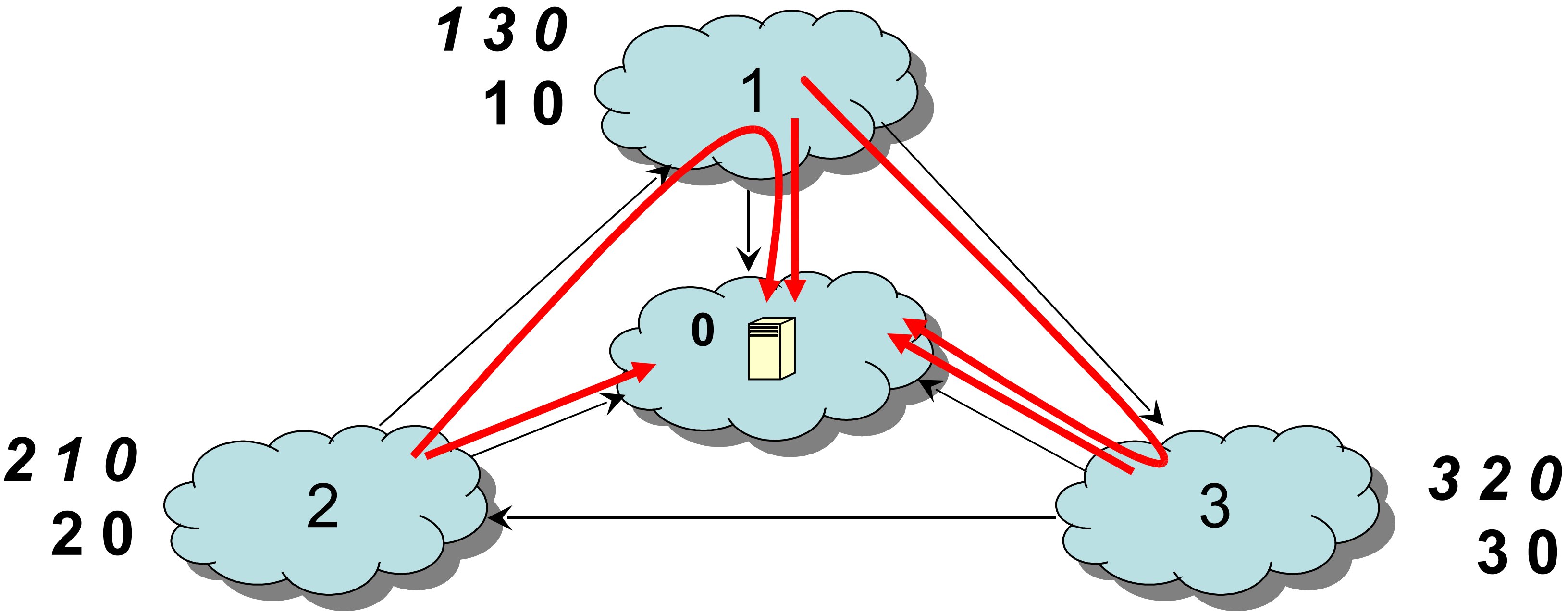
General Problems with BGP

- **Security** (more in Lecture 8, Feb. 6)
  - Too easy to “steal” IP address space
    - Happened again just yesterday
    - http://www.renesys.com/blog/2006/01/coned_steals_the_net.shtml
  - Hard to check veracity of information (*e.g.,* AS path)
  - Can’t tell where data traffic is actually going to go

- **Broken business models**
  - “Depeering” and degraded connectivity: universal connectivity depends on cooperation. *No guarantees!*

- **Policy interactions**
  - Oscillations (*e.g.,* today’s paper)

Policy Interactions



Varadhan, Govindan, & Estrin, "Persistent Route Oscillations in Interdomain Routing", 1996

Strawman: Global Policy Check

- Require each AS to publish its policies
- Detect and resolve conflicts

Problems:

- ASes typically unwilling to reveal policies
- Checking for convergence is NP-complete
- Failures may still cause oscillations

Think Globally, Act Locally

- Key features of a good solution
  - Safety: guaranteed convergence
  - Expressiveness: allow diverse policies for each AS
  - Autonomy: do not require revelation/coordination
  - Backwards-compatibility: no changes to BGP

- *Local* restrictions on configuration semantics
  - Ranking
  - Filtering

Main Idea of Today's Paper

- Permit only two business arrangements
  - Customer-provider
  - Peering

- Constrain both filtering and ranking based on these arrangements to guarantee safety

- Surprising result: these arrangements correspond to today's (common) behavior

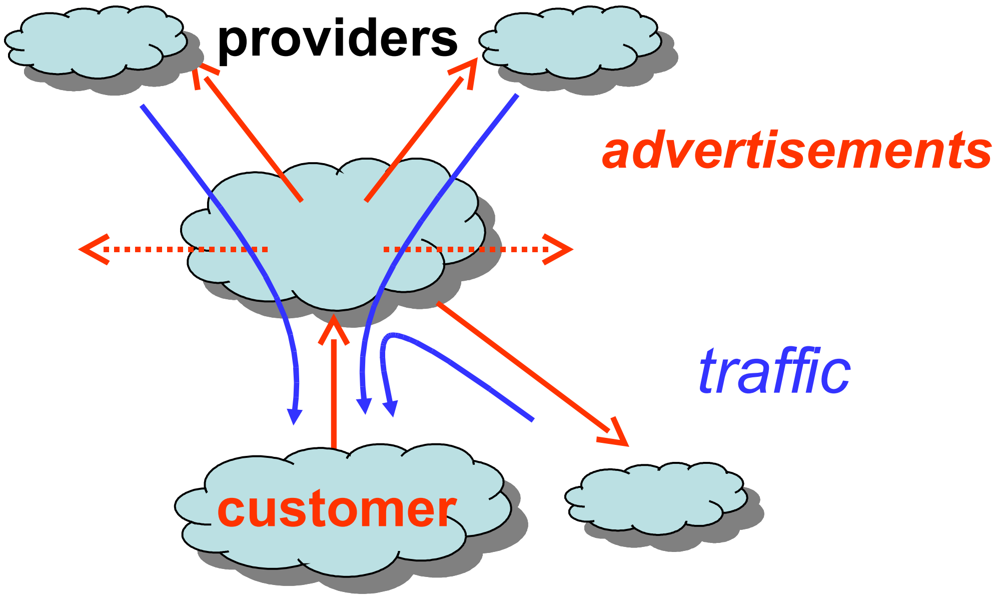
Gao & Rexford, "Stable Internet Routing without Global Coordination", *IEEE/ACM ToN*, 2001

Relationship #1: Customer-Provider

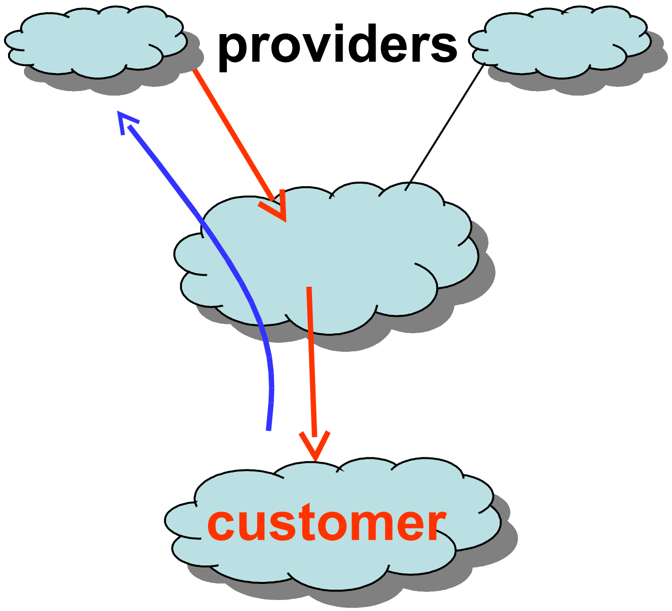
Filtering

- Routes from customer: to *everyone*
- Routes from provider: only to *customers*

From other destinations
To the customer



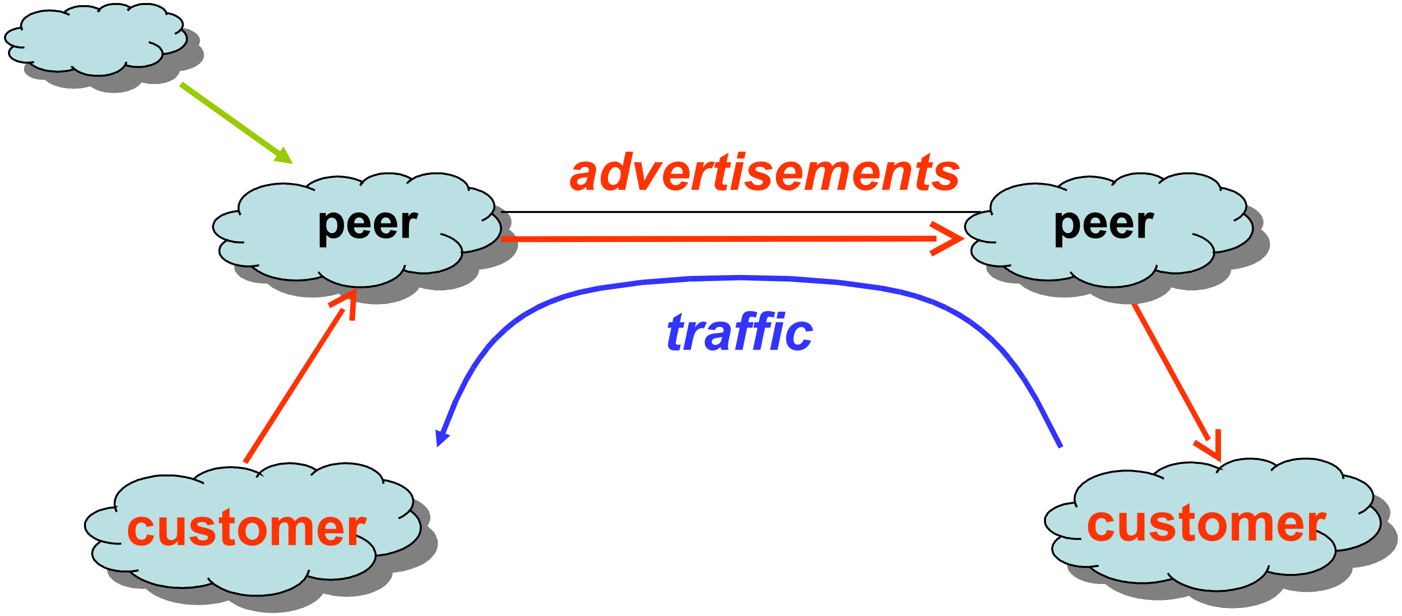
From the customer
To other destinations



Relationship #2: Peering

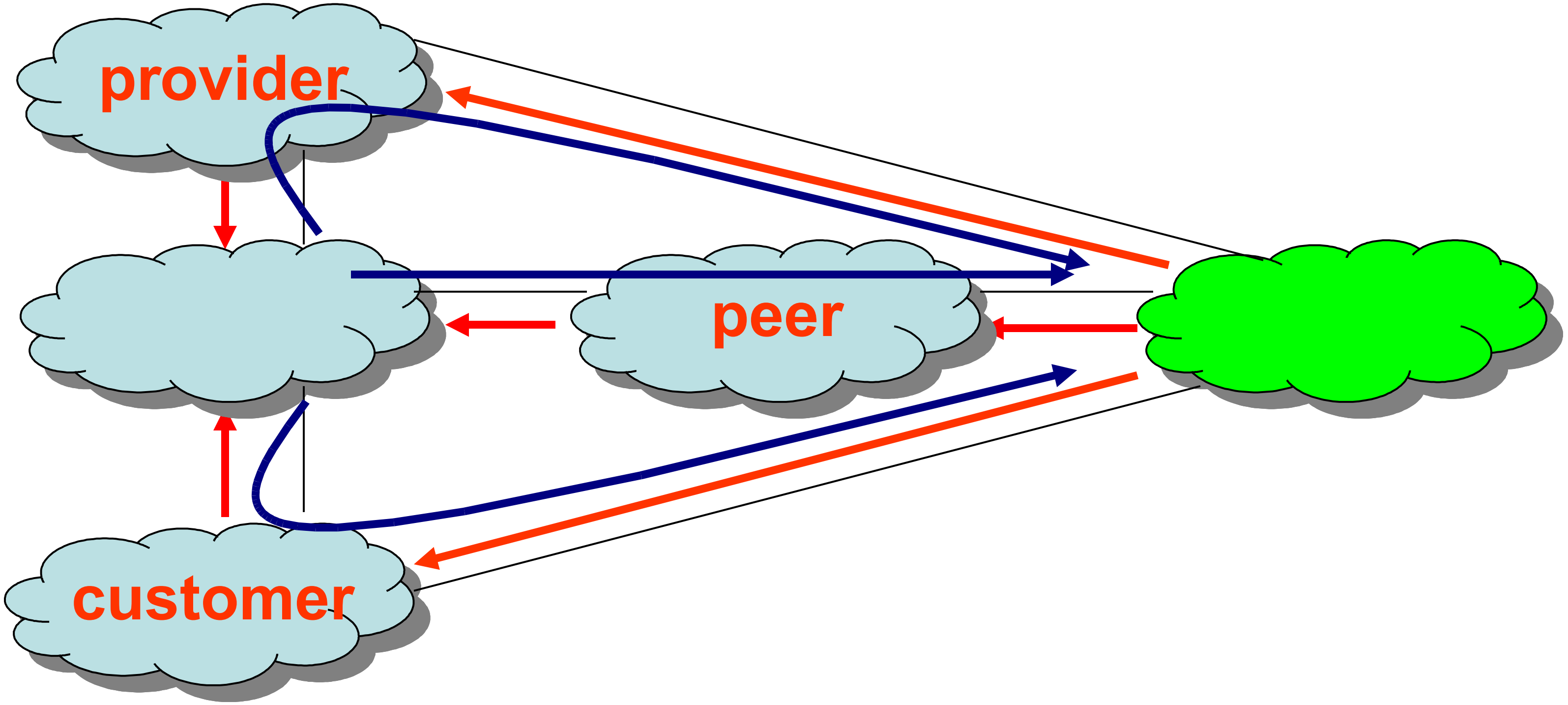
Filtering

- Routes from peer: only to customers
- No routes from other peers or providers

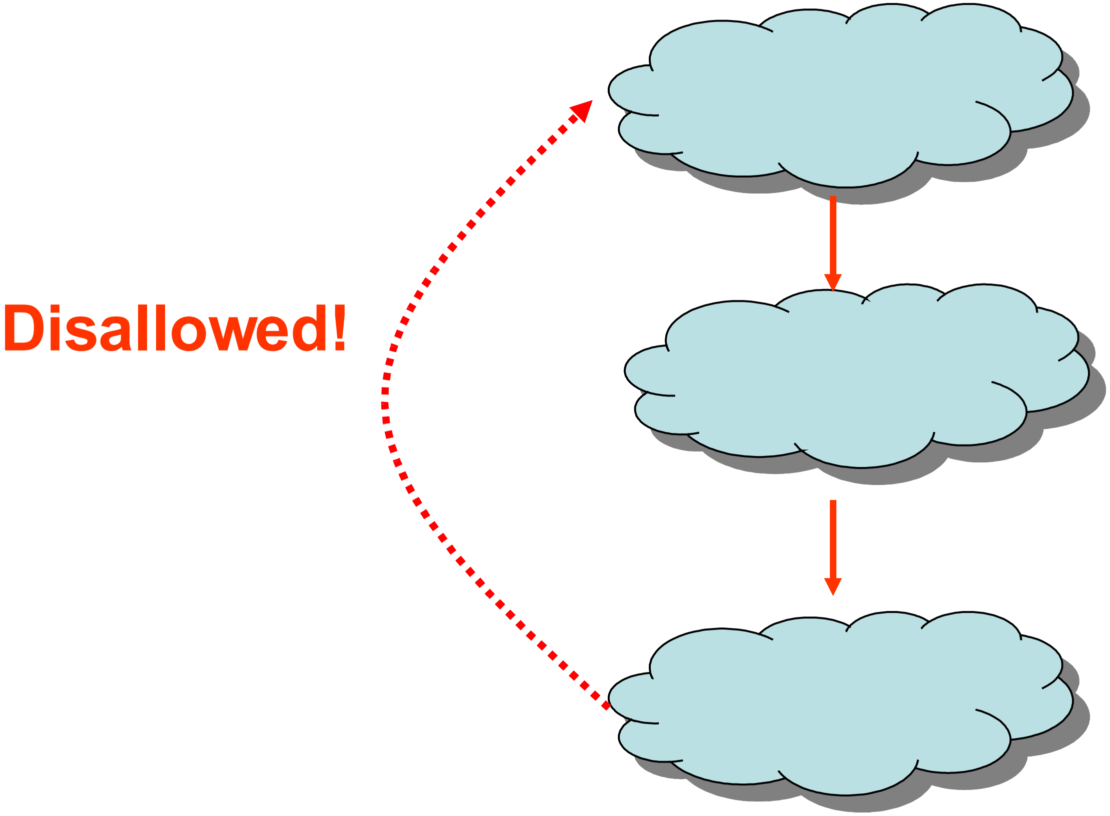


Rankings

- Routes from customers over routes from peers
- Routes from peers over routes from providers



Additional Assumption: Hierarchy

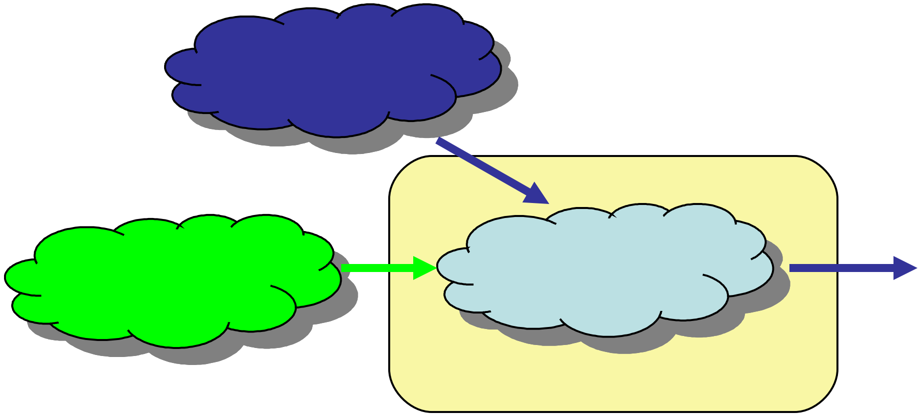


Safety: Proof Sketch

- **System state:** the current route at each AS
- **Activation sequence:** revisit some router's selection based on those of neighboring ASes

Activation Sequence: Intuition

- **Activation:** emulates a message ordering
  - Activated router has received and processed all messages corresponding to the system state



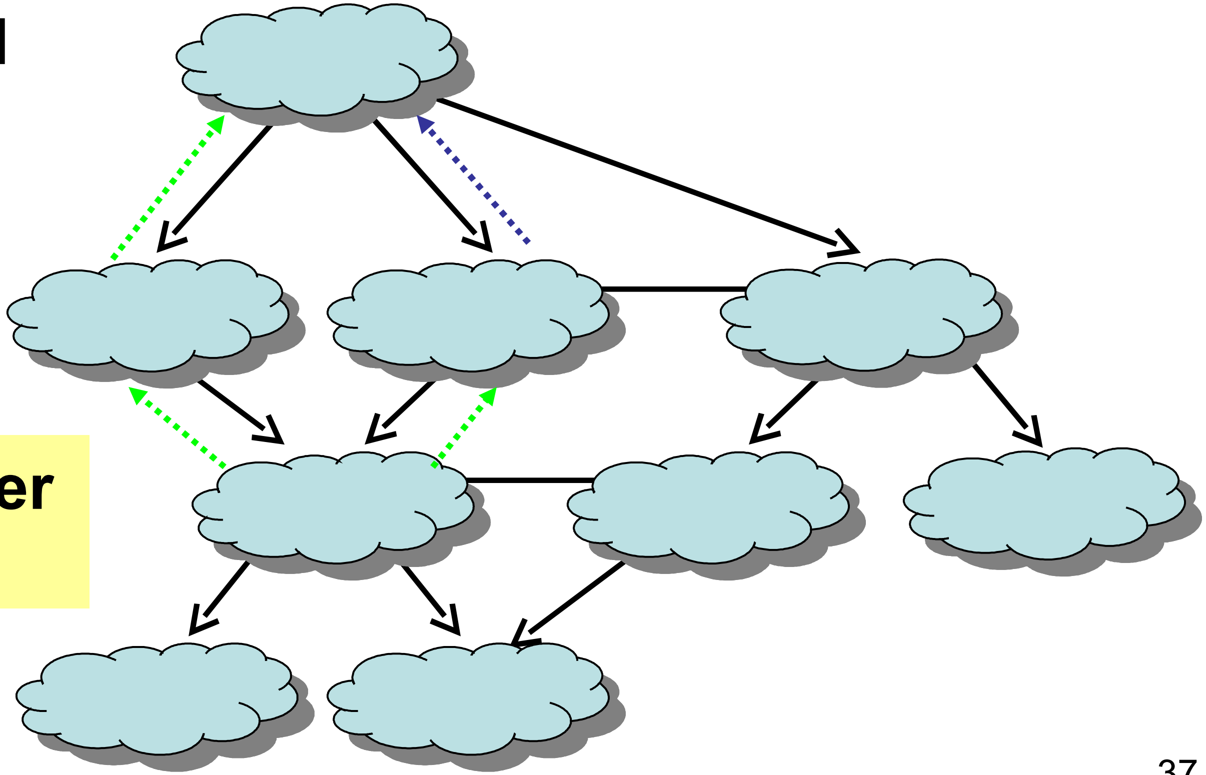
- **“Fair” activation:** all routers receive and process outstanding messages

Safety: Proof Sketch

- **State:** the current route at each AS
- **Activation sequence:** revisit some router’s selection based on those of neighboring ASes
- **Goal:** find an activation sequence that leads to a stable state
- **Safety:** satisfied if that activation sequence is contained within any “fair” activation sequence

Proof, Step 1: Customer Routes

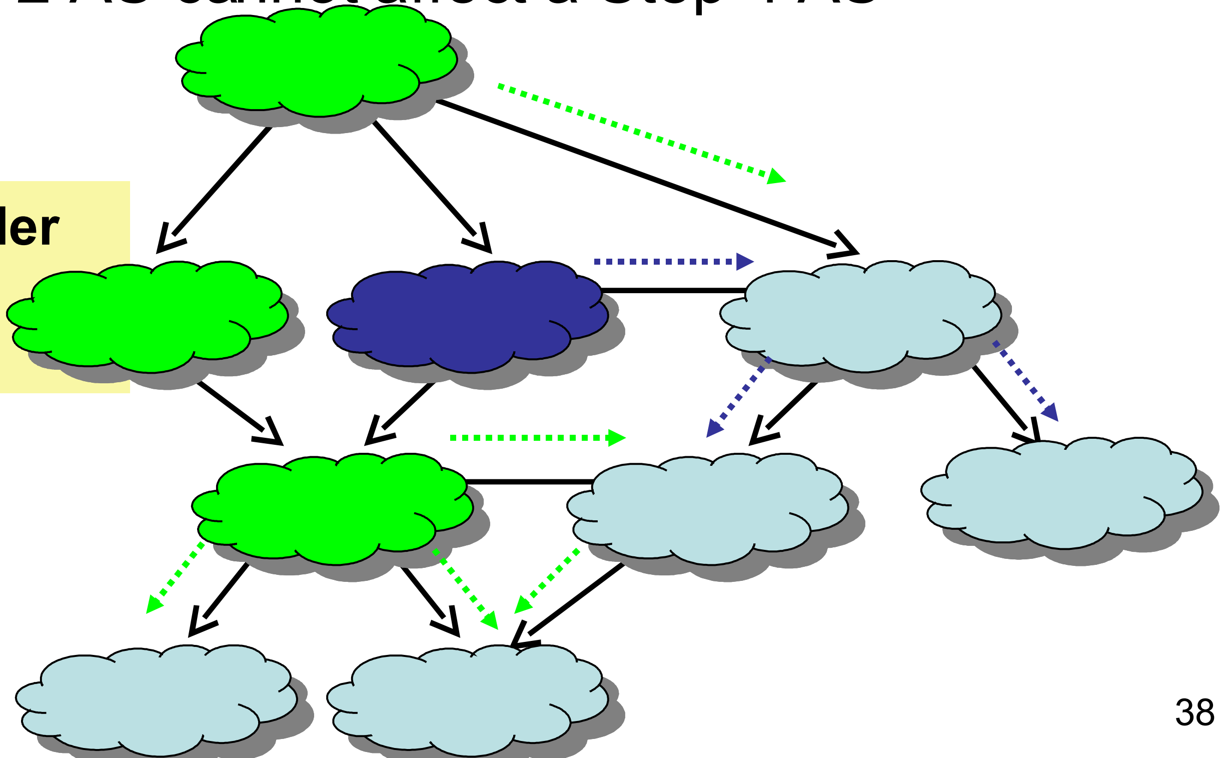
- Activate ASes from customer to provider
  - AS picks a customer route if one exists
  - Decision of one AS cannot cause an earlier AS to change its mind



An AS picks a customer route when one exists

Proof, Step 2: Peer & Provider Routes

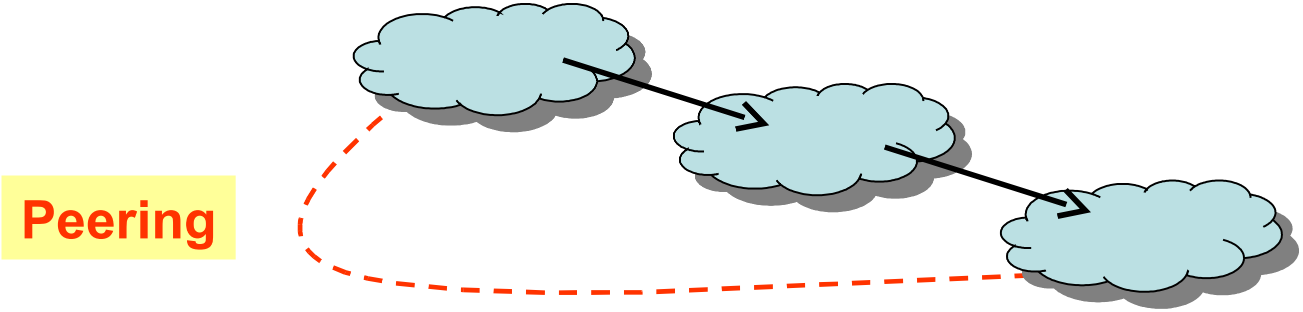
- Activate remaining ASes from provider to customer
  - Decision of one Step-2 AS cannot cause an earlier Step-2 AS to change its mind
  - Decision of Step-2 AS cannot affect a Step-1 AS



AS picks a peer or provider route when no customer route is available

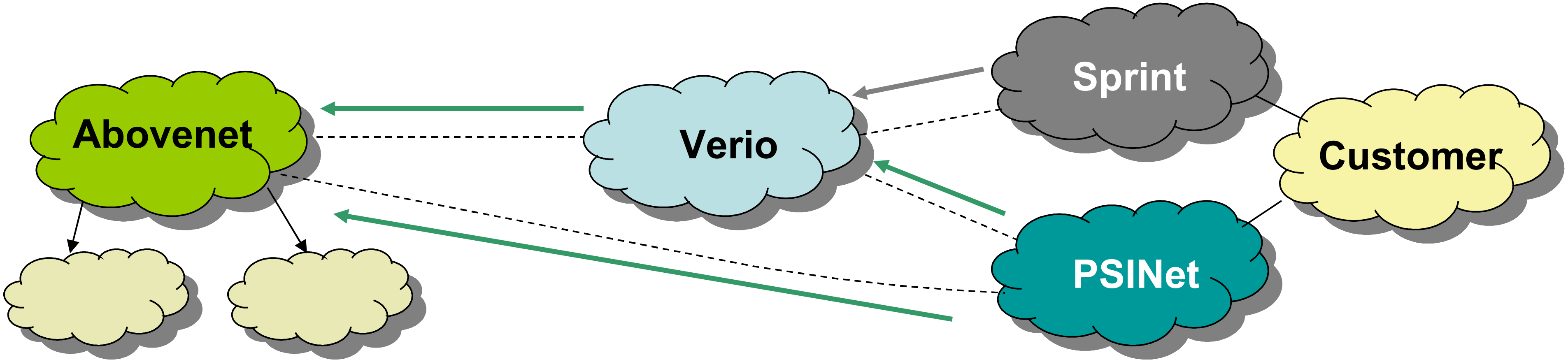
Ranking and Filtering Interactions

- Allowing *more flexibility* in ranking
  - Allow *same* preference for peer and customer routes
  - Never choose a peer route over a *shorter* customer route
- … at the *expense* of stricter AS graph assumptions
  - Hierarchical provider-customer relationship (as before)
  - No private peering with (direct or indirect) providers



Some problems

- Requires acyclic hierarchy *(global condition)*
- Cannot express many business relationships

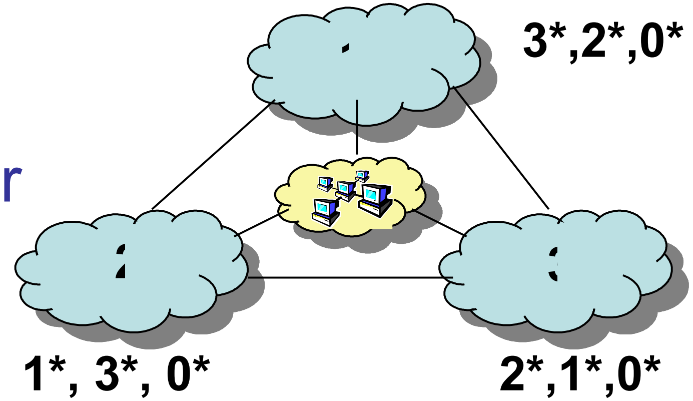


Question: Can we relax the constraints on filtering? What happens to rankings?

Other Possible Local Rankings

Accept only *next-hop rankings*

- Captures most routing policies
- Generalizes customer/peer/provider
- Problem: system not safe



Accept only *shortest hop count rankings*

- Guarantees safety under filtering
- Problem: not *expressive*

Feamster, Johari, & Balakrishnan, “Implications of Autonomy for the Expressiveness of Policy Routing”, *SIGCOMM 2005*

What Rankings Violate Safety?

Theorem.
Permitting paths of length *n+2* over paths of length *n* will violate safety under filtering.

Theorem.
Permitting paths of length *n+1* over paths of length *n* will result in a dispute wheel.

Feamster, Johari, & Balakrishnan, “Implications of Autonomy for the Expressiveness of Policy Routing”, *SIGCOMM 2005*