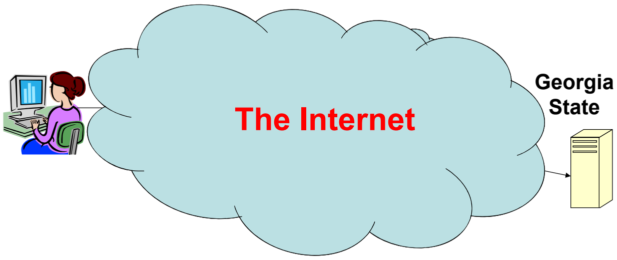
Interdomain Routing

Internet Routing



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

AS Numbers (ASNs)

ASNs are 16 bit values.
64512 through 65535 are "private"

Currently around 30,000 in use.

- Level 3: 1
- MIT: 3
- Harvard: 11
- Yale: 29
- Princeton: 88
- AT&T: 7018, 6341, 5074, ...
- UUNET: 701, 702, 284, 12199, ...
- Sprint: 1239, 1240, 6211, 6242, ...
- ...

ASNs represent units of routing policy

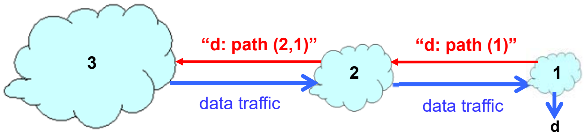
Challenges for Interdomain Routing

- Scale
  - Prefixes: 250,000, and growing
  - ASes: 30,000, and growing
  - Routers: at least in the millions…
- Privacy
  - ASes don't want to divulge internal topologies
  - … or their business relationships with neighbors
- Policy
  - No Internet-wide notion of a link cost metric
  - Need control over where you send traffic
  - … and who can send traffic through you

Policy-Based Path-Vector Routing

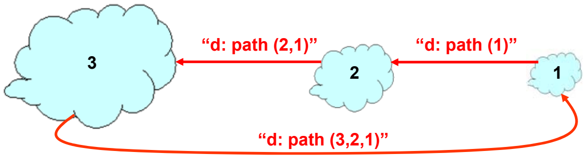
Path-Vector Routing

- Extension of distance-vector routing
  - Support flexible routing policies
  - Reduce convergence time (avoid count-to-infinity)
- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per dest d
  - Path vector: send the *entire path* for each dest d



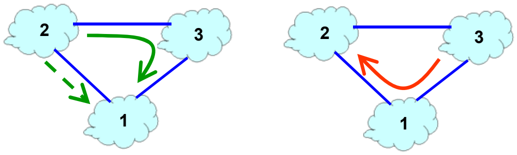
Faster Loop Detection

- 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



Flexible Policies

- Each node can apply local policies
  - Path selection: Which path to use?
  - Path export: Whether to advertise the path?
- 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"

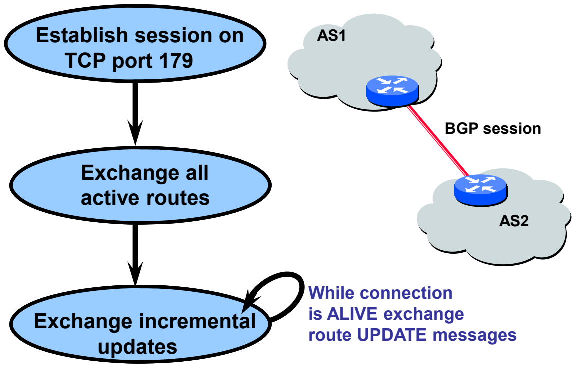


Border Gateway Protocol

- Prefix-based path-vector protocol
- Policy-based routing based on AS Paths
- Evolved during the past 18 years

- 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
- 2006 : BGP-4 [RFC 4271], update

BGP Operations

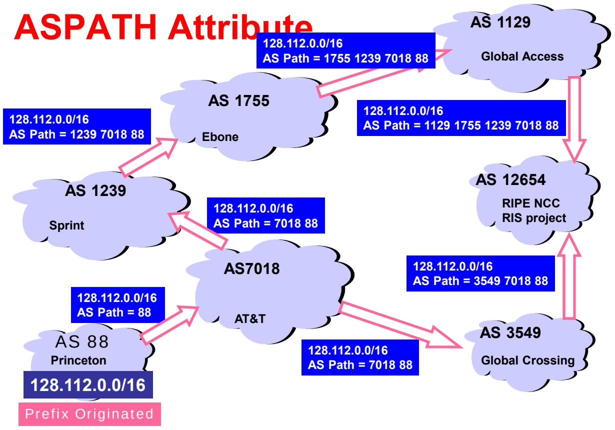

Establish session on TCP port 179
Exchange all active routes
Exchange incremental updates
AS1
BGP session
AS2

While connection is ALIVE exchange route UPDATE messages

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 node 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

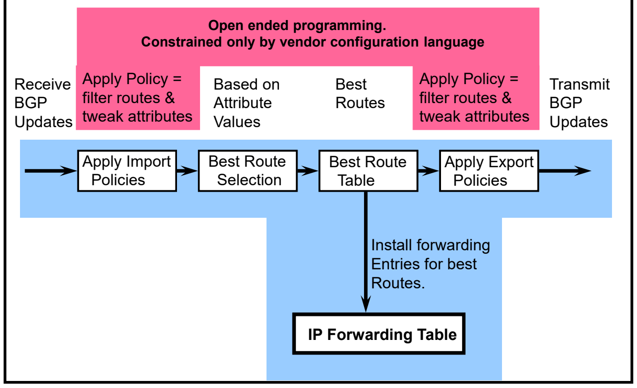
ASPATH Attribute



BGP Policy: Applying Policy to Routes

- Import policy
  - Filter unwanted routes from neighbor
    - E.g. prefix that your customer doesn't own
  - Manipulate attributes to influence path selection
    - E.g., assign local preference to favored routes
- Export policy
  - Filter routes you don't want to tell your neighbor
    - E.g., don't tell a peer a route learned from other peer
  - Manipulate attributes to control what they see
    - E.g., make a path look artificially longer than it is

BGP Policy: Influencing Decisions



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

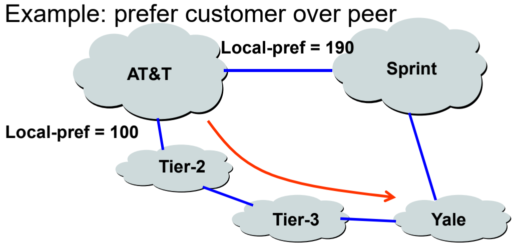
Apply Export Policies

Install forwarding Entries for best Routes.

IP Forwarding Table

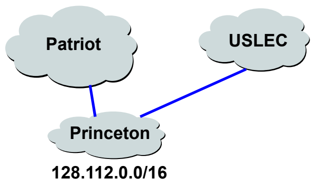
Import Policy: Local Preference

- Favor one path over another
  - Override the influence of AS path length
  - Apply local policies to prefer a path
- Example: prefer customer over peer



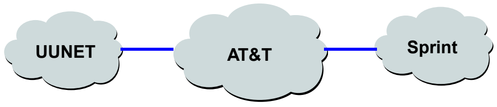
Import Policy: Filtering

- Discard some route announcements
  - Detect 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



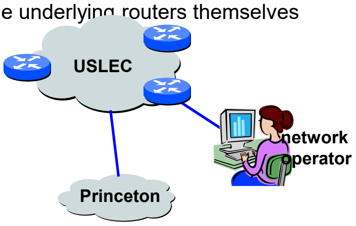
Export Policy: Filtering

- Discard some route announcements
  - Limit propagation of routing information
- Examples
  - Don’t announce routes from one peer to another



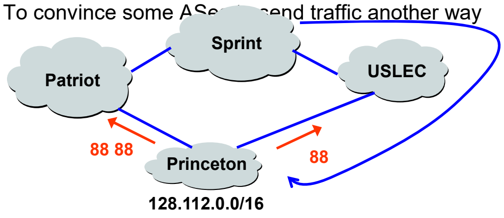
Export Policy: Filtering

- 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

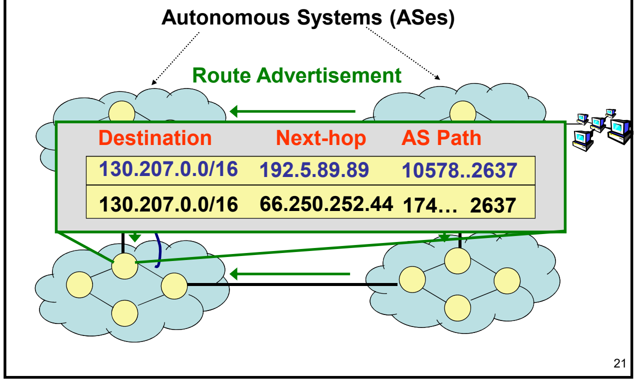
- Modify attributes of the active route
  - To influence the way other ASes behave
- Example: AS prepending
  - Artificially inflate the AS path length seen by others
  - To convince some ASes to send traffic another way



BGP Policy Configuration

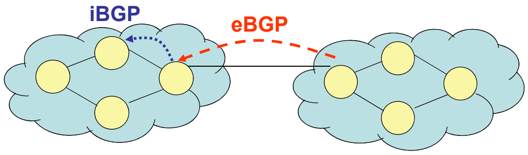
- Routing policy languages are vendor-specific
  - Not part of the BGP protocol specification
  - Different languages for Cisco, Juniper, etc.
- Still, all languages have some key features
  - Policy as a list of clauses
  - Each clause matches on route attributes
  - … and either discards or modifies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

Internet Routing Protocol: BGP



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

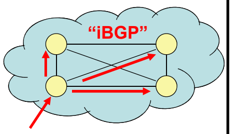


- **External BGP (eBGP):** exchanging routes *between* ASes
- **Internal BGP (iBGP):** disseminating routes to external destinations among the routers *within an AS*

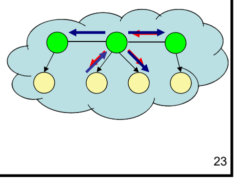
Question: **What's the difference between IGP and iBGP?**

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

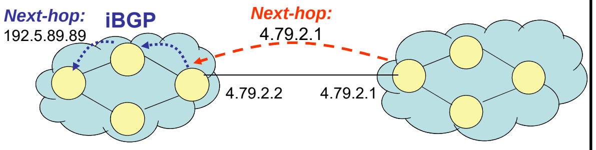
Next-hop

Routing Attributes and Route Selection

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

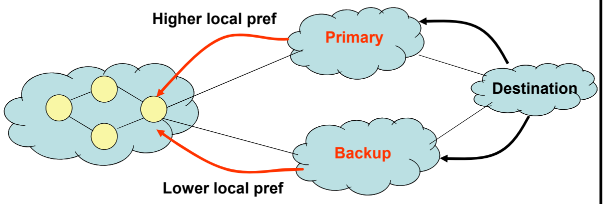
- **Local preference:** numerical value assigned by routing policy. Higher values are more preferred.
- **AS path length:** number of AS-level hops in the path
- **Multiple exit discriminator ("MED"):** allows one AS to specify that one exit point is more preferred than another. Lower values are more 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



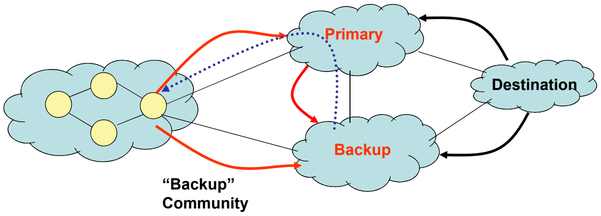
- **Next-hop:** IP address to send packets en route to destination.
- **Community value:** Semantically meaningless. Used for passing around "signals" and labeling routes. More in a bit.

Local Preference



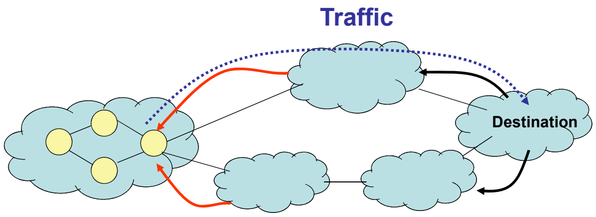
- **Control over *outbound* traffic**
- *Not* transitive across ASes
- Coarse hammer to implement route preference
- Useful for preferring routes from one AS over another (*e.g.*, primary-backup semantics)

Communities and Local Preference



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

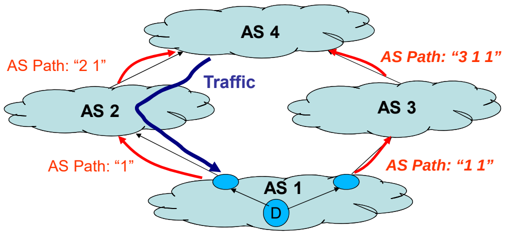
AS Path Length



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

29

AS Path Length Hack: Prepending



- Attempt to control inbound traffic
- Make AS path length look artificially longer

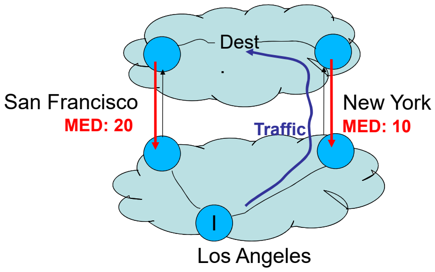
30

Multi-Exit Discriminator (MED)

- Hint to external neighbors about the preferred path into an AS
  - Non-transitive attribute
  - Different AS choose different scales
- Used when two AS’s connect to each other in more than one place

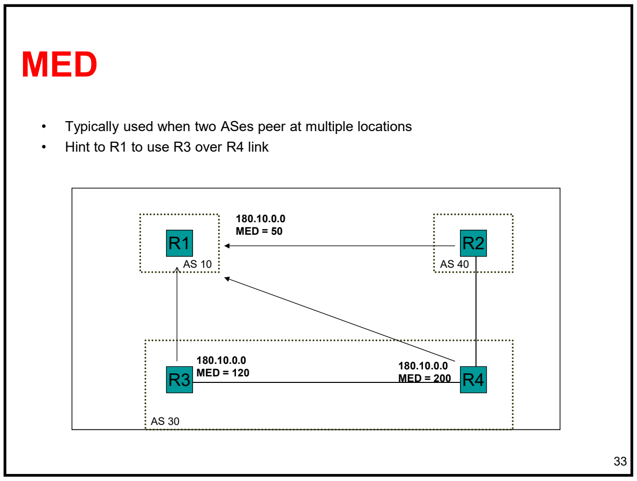
31

Multiple Exit Discriminator (MED)



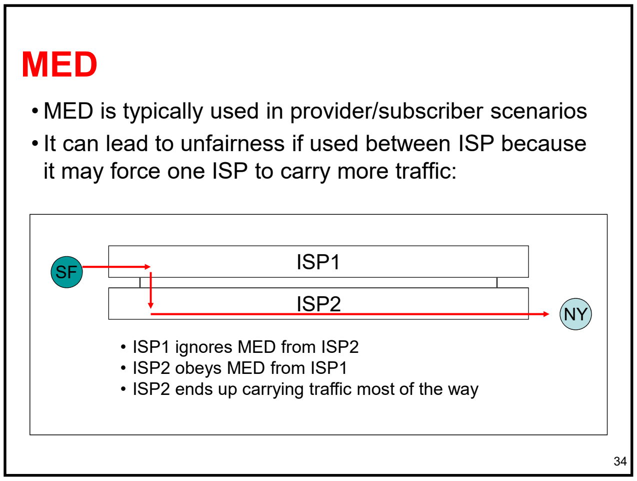
- Mechanism for AS to control how traffic enters, given multiple possible entry points.

32



MED

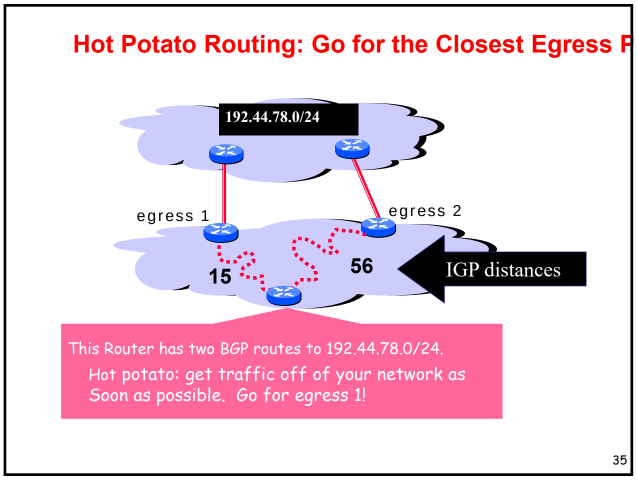
- Typically used when two ASes peer at multiple locations
- Hint to R1 to use R3 over R4 link



MED

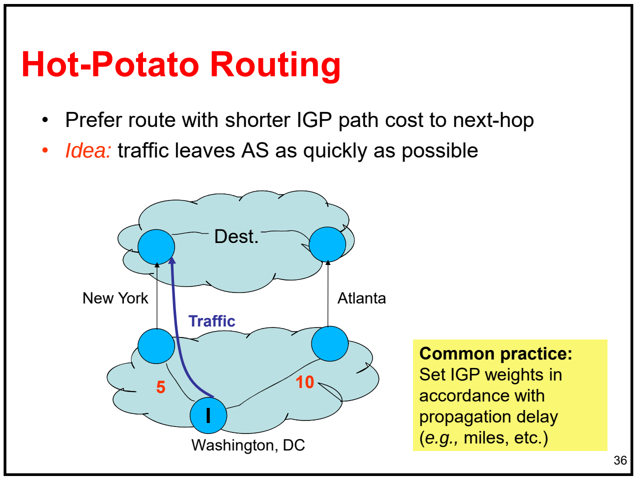
- MED is typically used in provider/subscriber scenarios
- It can lead to unfairness if used between ISP because it may force one ISP to carry more traffic:

- ISP1 ignores MED from ISP2
- ISP2 obeys MED from ISP1
- ISP2 ends up carrying traffic most of the way



Hot Potato Routing: Go for the Closest Egress P

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!



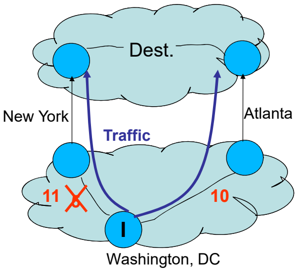
Hot-Potato Routing

- 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
```

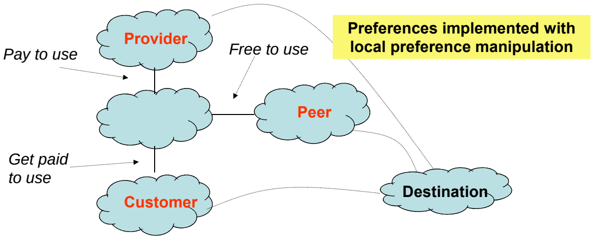
eBGP Session

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

Internet Business Model (Simplified)



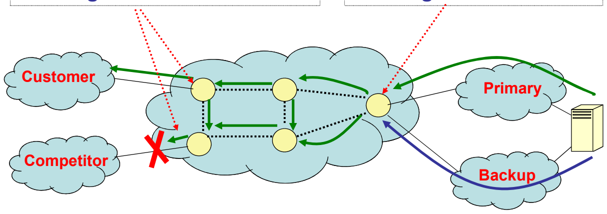
Preferences implemented with local preference manipulation

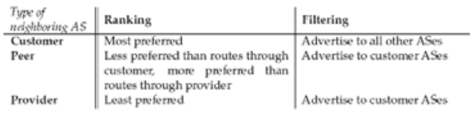
- **Customer/Provider:** One AS pays another for reachability to some set of destinations
- **"Settlement-free" Peering:** Bartering. Two ASes exchange routes with one another.

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

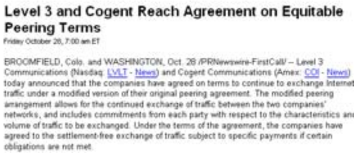
- Cooperative competition (brinksmanship)
- Much more desirable to have your peer’s customers
  - Much nicer to get paid for transit
- Peering “tiffs” are relatively common

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!

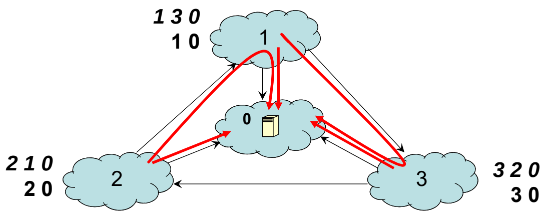
As of 5:30 am EDT, October 5th, Level(3) terminated peering with Cogent without cause (as permitted under its peering agreement with Cogent) even though both Cogent and Level(3) remained in full compliance with the previously existing interconnection agreement. Cogent has left the peering circuits open 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**
  - 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

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