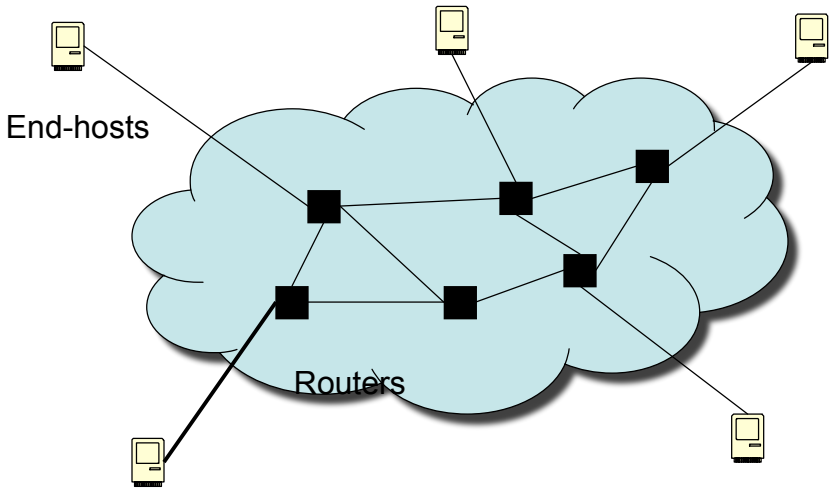
Interdomain Routing

David Andersen
15-744 Spring 2007
Carnegie Mellon University

Outline

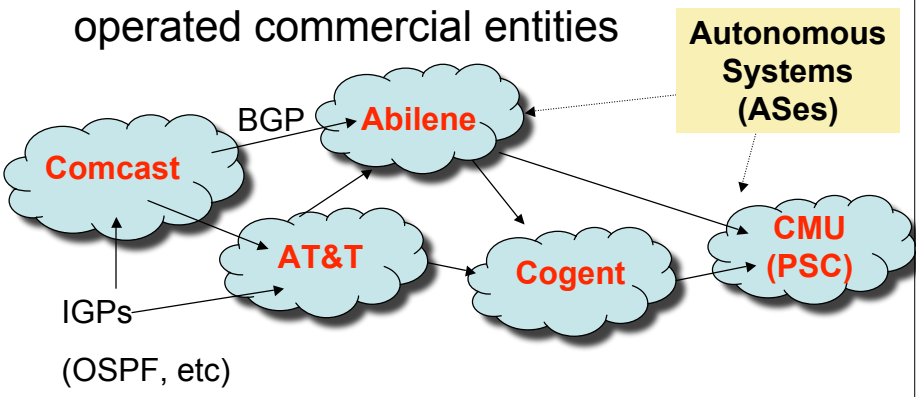
- What does the Internet look like?
- Relationships between providers
  - Enforced by: Export filters and import ranking
- BGP: The Border Gateway Protocol
  - Design goals
  - Protocol basics
    - Updates, withdrawals, path vector concept
  - eBGP and iBGP
  - Scaling with confederations and route reflectors
  - BGP decision process, MEDs, localpref, and path length; load balancing
  - Failover and scalability
    - Multi-homing and address allocation
    - Convergence problems
    - Preview of stability

The Internet: Attempt 1



The Internet: Zooming In

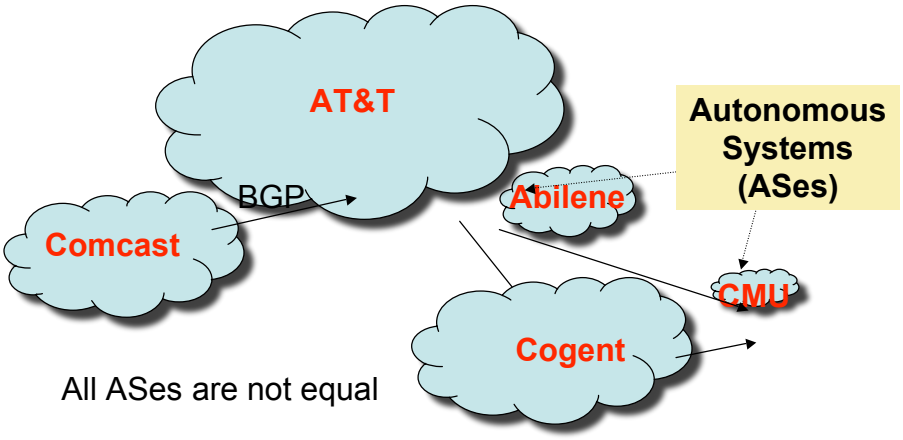
- ASes: Independently owned & operated commercial entities



ASes

- Economically independent
- All must cooperate to ensure reachability
- Routing between: BGP
- Routing inside: Up to the AS
  - OSPF, E-IGRP, ISIS (You may have heard of RIP; almost nobody uses it)
- Inside an AS: Independent policies about nearly everything.

The Internet: Zooming In 2x

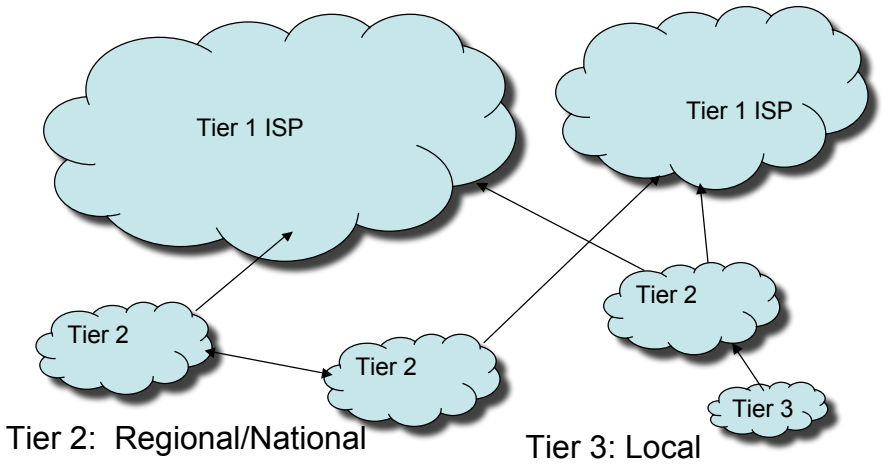


All ASes are not equal

AS relationships

- Very complex economic landscape.
- Simplifying a bit:
  - Transit: "I pay you to carry my packets to everywhere" (provider-customer)
  - Peering: "For free, I carry your packets to my customers only." (peer-peer)
- Technical defn of tier-1 ISP: In the "default-free" zone. No transit.
  - Note that other "tiers" are marketing, but convenient. "Tier 3" may connect to tier-1.

Zooming in 4



Economics of Packets

- Transit: Customer pays the provider
- Who is who? Usually, the one who can "live without" the other. AT&T does not need CMU, but CMU needs *some* ISP.
- What if both need each other? Peering.
  - Instead of sending packets over $$ transit, set up a direct connection and exchange traffic for free! (traceroute www.pitt.edu)

- Tier 1s must peer by definition
- Peering *can* give:
  - Better performance
  - Lower cost
  - More "efficient" routing (keeps packets local)
- But negotiating can be very hairy!

Business and peering

- 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)

Formalizing Relationships

- Provider:
  - **Sends:** all routes to customer and customer's routes to everyone
  - **Prefers:** Route to customers over peers/providers
- Peering:
  - **Sends:** to customers but not to other peers or providers.
  - **Prefers:** Route to peer over providers
- Customer:
  - **Sends:** to customers but not to peers or other providers
  - **Prefers:** Anything else.

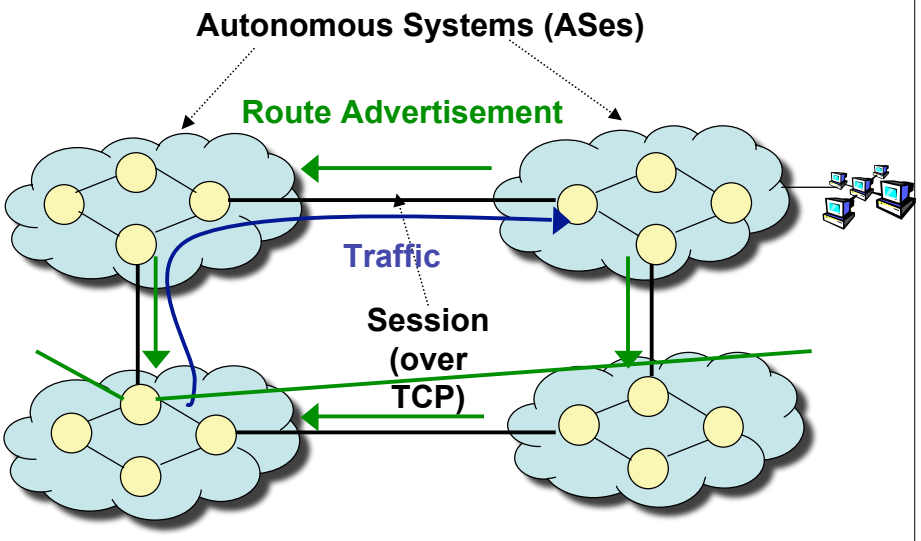
Enforcing relationships

- Two mechanisms:
- Export filters
  - Control what you send over BGP
- Import *ranking*
  - Controls which route you prefer of those you hear.
  - "LOCALPREF" – Local Preference. More later.
- Terminology nit: Both people at the BGP session level are called "BGP peers" regardless of business relationship. So you have a BGP peering session with your provider…

BGP version 4

- Design goals:
  - Scalability as more networks connect
  - Policy: ASes hould be able to enforce business/routing policies
    - Result: Flexible attribute structure, filtering
  - Cooperation under competition:
    - ASes should have great autonomy for routing and internal architecture
    - But BGP should provide global reachability

BGP

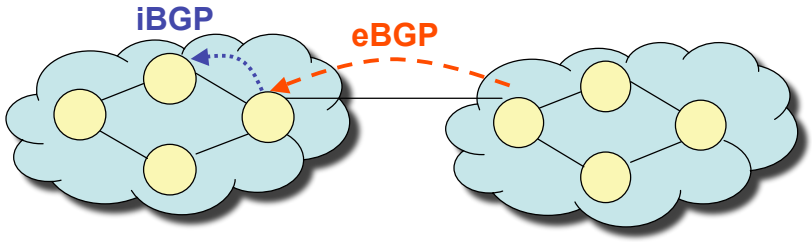


- BGP messages
  - OPEN
  - UPDATE
    - Announcements
      - Dest Next-hop AS Path … other attributes …
      - 128.2.0.0/16 196.7.106.245 2905 701 1239 5050 9
    - Withdrawals
  - KEEPALIVE
    - Keepalive timer / hold timer
- Key thing: The Next Hop attribute

Path Vector

- ASPATH Attribute
  - Records what ASes a route went through
  - Loop avoidance: Immediately discard
  - Short path heuristics
- Like distance vector, but fixes the count-to-infinity problem

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*

Two flavors?

- Most ASes have more than one “border” router that talks to other peers
- Must disseminate information inside the AS and through the AS.
  - Must be loop-free.
  - Must have complete visibility.
    - AS is a monolithic entity, so routers must be consistent
    - For every external destination, each router picks the same route that it would have picked had it seen the best routes from each eBGP router in the AS.

iBGP and the gooey insides

- iBGP is not an IGP
  - Does *not* set up forwarding state internally!
  - Requires that you have an IGP so that all routers can talk to all other routers
- Original: Full mesh iBGP
  - Simple! All routers see all routes
  - But, causes scaling problems.
    - Route Reflectors and Confederations

Scaling iBGP w/Route Reflectors

- A BGP router with clients
  - RR selects a single best-path for each prefix and sends it to all clients
    - If RR learns a route via eBGP or iBGP from a client, it re-advertises to everyone
    - If RR learns via non-client iBGP, it only advertises to clients
    - (Figure. ☺)

Problems with RRs

- Only 1? Single point of failure, etc.
  - But that's pretty easy to fix.
- More serious: All clients take same path
  - May mess up *traffic engineering*
- Common solution: Hierarchicy of RRs
- Note: Properly configuring iBGP, RRs, eBGP and your IGP to be loop-free is actually quite difficult!
  - Not all RR configurations provide visibility
  - Inconsistency can lead to loops

Picking Paths

- By agreed-upon convention:
  - Weight (don't worry about it)
  - Local Preference
    - Used to express peer/provider/cust, etc., and outbound load balancing
  - Local route
  - AS Path length
  - Origin
  - MED – Multi-exit discriminator
    - Set by the route sender as an *inbound* load balancing / locality suggestion

Load balancing

- Outbound is "easy" (you have control)
  - Set localpref according to goals
- Inbound is tough (nobody has to listen)
  - AS path prepending
  - MEDs
    - Hot and Cold Potato Routing (picture)
    - Often ignored unless contracts involved
    - Practical use: tier-1 peering with a content provider

Failover

- BGP is designed for scaling more than fast failover
  - Many mechanisms favor this balance
  - Route flap damping, for example.
    - If excess routing changes ("flapping"), ignore for some time.
    - Has unexpected effects on convergence times.
  - Route advertisement/withdrawal timers in the 30 second range
  - Effect: tens of seconds to many minutes to recover from "simple" failures.
  - 15-30 minute outages not uncommon.

Multi-homing

- (More later in the semester)
- Connect to multiple providers
  - Goal: Higher availability, more capacity
- Problems:
  - Provider-based addressing breaks
  - Everyone needs their own address space

BGP problems

- Convergence: BGP may explore many routes before finding the right new one.
  - Labovitz et al., SIGCOMM 2000
- Correctness: routes may not be valid, visible, or loop-free.
- Security: There is none!
  - Some providers filter what announcements their customers can make. Not all do.
  - See paper discussion site for pointers

Summary

- Internet is a set of federated independent networks
- They must play nicely for everybody to be connected to everybody else
- Routing is done using BGP
  - Simple protocol
  - Extremely complex configuration flexibility
  - Many open research problems in policy, scalability, failover, configuration, correctness, security.

Acknowledgements

- Much of the outline of this lecture is derived from a similar course by Hari Balakrishnan.
- Several slides are from Nick Feamster