CSE 222
Graduate Networking

Winter 2001

Lecture 8: Internet Routing

Geoffrey M. Voelker

Overview

- Last time we looked at the theory behind distributed routing protocols
- Today we'll look at how they apply to routing in the Internet
  - Routing among ISPs/ASes using BGP
  - Peering and transit relationships among ISPs
  - End-to-end routing behaviors
  - Note: First two will be repeats for those in Stefan's class…

Autonomous System (AS)

- An Autonomous System: Roughly, an entity with uniform administrative control over its network
  - Transit AS (ISPs)
    - » ASes that carry traffic across their network
  - Stub AS
    - » ASes that are sources/sinks for traffic
- Every AS is indentified using a unique number (ASN)
  - UUNet (AS 701), AT&T (7018), Sprint (1239), etc.

Hierarchical Routing

- Internet routing is hierarchical
  - One protocol used inside of a network (AS), another to route among networks (ASes)
  - Interior Gateway Protocol (IGP): link-state: OSPF, ISIS
  - Exterior Gateway Protocol (EGP): BGP
- IGPs potentially faster and more stable, but…
  - Can't do good static load balancing
    - » Shortest path gets all of the traffic, second shortest none
  - Doesn't do any dynamic load balancing
    - » Metrics are static, load insensitive
  - Can't express policy
    - » Only allow traffic from X to go to Y

Border Gateway Protocol

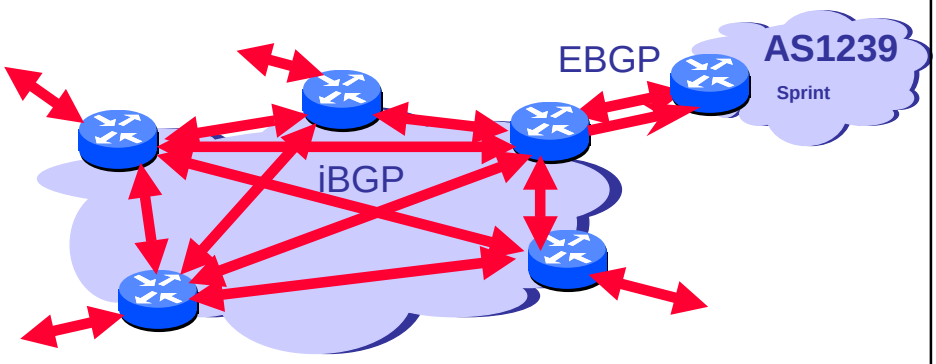
- BGP is used to route among ASes
- It is a distance vector protocol that exchanges reachability information
  - No metrics
  - Not about optimizing anything
  - All about policy (business and politics)

BGP Algorithm

- Border routers in each AS communicate with neighboring routers in other ASes
- BGP route announcements say:
  - I can reach this network, and this is the path of ASN's I heard this from
  - Plus some attributes I choose to tell you
- Can't accept route if your ASN is in it
  - Prevents loops

EBGP vs. iBGP

- Routers between ASes speak EBGP
- Routers within AS use iBGP to synchronize tables



BGP Decision Process

- Local policy controls everything
  - Routes you listen to
  - Routes you install in forwarding table
  - Routes you advertise
- Standard selection procedure
  - Highest local preference, the shortest AS path, most local origin, etc.
  - Ugly, hard to reason about
- Bottom line
  - Top-level Internet routing is ruled by policy ($$$ and politics)
  - Performance matters in as much as it affects $$$

Norton: ISPs and Peering

- BGP is the protocol by which ASes announce and exchange routes
- How is it used in the Internet today?
- Not a research paper, but it's hard to find descriptions of how this stuff works

History

- 1994-1995 NSFnet transitions to commercial Internet
  - NSF ceases to manage/control Internet backbone
  - Commercial carriers (MCI, Sprint, UUNet, etc.) start to sell IP backbone service
  - Interconnected with each other and regional networks at several public exchange points, network access points (NAP)
    - » MAE-West, MAE-East, etc.
  - At exchange points, everyone talks to everyone else
- Then five years go by….

Telephony

- In the telephone world
  - Local exchange carriers (LECs)
  - Inter-exchange carriers (IXCs)
- LECs must provide IXCs access to customers according to government regulation
- When a call goes from one phone company to another
  - Call billed to caller
  - Money is split up among the phone systems (known as "settlement"

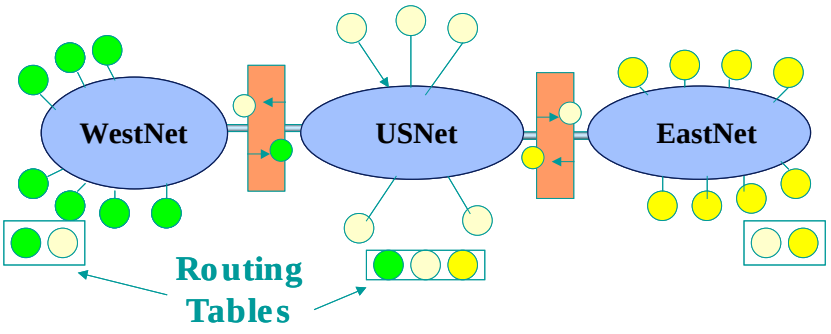
Internet

- No regulation
  - One ISP doesn't have to talk to another
- Founded on "shared goodwill"
  - Pay for connectivity, not per packet
  - Not clear who should pay anyway ("caller" or "callee"?)
- No standard settlement

Peering vs Transit

- Peering
  - Two ISPs provide connectivity to each others customers (typically for free, sometimes for $$$ when load imbalanced)
  - Reciprocal, but non-transitive relationship
- Transit
  - One ISP provides connectivity to every place it knows about (usually for $$$)
  - It sells destinations to a customer, announces customer to all its peers and transits

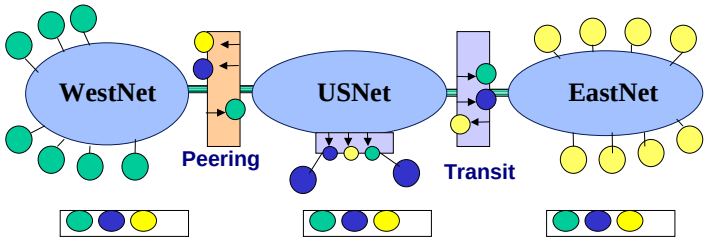
Example: Peering



- WestNet sees USNet customers and vice-versa
- EastNet sees USNet customers and vice-versa
- But WestNet does not see EastNet customers

Example: Transit

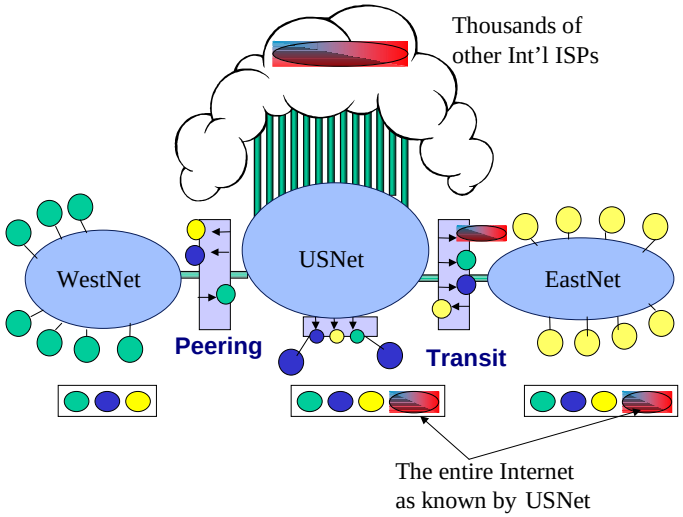
- EastNet purchases transit from USNet
  - USNet announces all of its routes to EastNet
  - USNet announces routes to EastNet to its peers and transits
  - Note that WestNet sees EastNet customers now…and without paying for transit!



The Value of Transit

- Not just paying for the fiber, also paying for the routes
  - Routes to destinations, and routes back to you
  - There is no single “backbone” to hook up to
  - If you’re an ISP, how do your customers get to Yahoo!?
  - Make a business arrangement with another ISP that can
- Implications
  - Large ISPs have more value to offer small ISPs
  - How do small ISPs get off ground?
    - » Start by buying transit
    - » Grow, grow, grow
    - » Leverage public exchange points and peering relationships to reduce dependence on transit

The Value of Transit (2)



Aside

- Peering and transit are two popular points on a continuum
- Some places sell "partial transit"
- Other places sell "usage-based" peering
- Bottom line:
  - Which routes do you give away and which do you sell? To whom? Under what conditions?

Public vs. Private Peering

- Public peering
  - Connection via shared switch or network at a “public” exchange point
    - » Place anyone can be if they pay money
  - Still negotiated bilaterally
- Private peering
  - Private point-to-point link between peers

Why Peer?

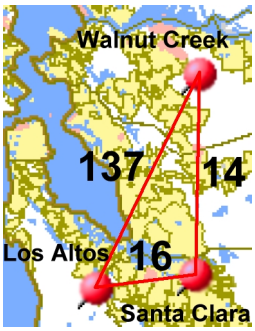
- Transit is expensive
  - $150,000/month for an OC3 (155Mbps) transit link
    - » Remember, it’s not just the bandwidth, it’s also the routes
- Peering with other ISPs can reduce use of transit link
  - » Potentially lower latency (if traffic headed in right direction)
- Communication patterns aren’t uniform
  - More of your traffic is exchanged with some networks than others
    - » Web hosting vs. Web browsing
  - Try to peer with other ISPs whose customers exchange traffic frequently/equally with your customers

Why Not Peer?

- Traffic asymmetry
  - More traffic goes one way than the other
  - Peer who carries more traffic feels cheated
- Hassle (operationally, administratively)
- Big ISPs have no interest in helping small ISPs
  - The “Big Boys” all peer with each other at no/little cost
- Harder to deal with problems without strong financial incentives

Discussion

- AS connectivity is all about policy, not performance
- What are the implications on…
  - Performance? Plenty of anomalies…
  - Reliability?
  - Routing behavior?



- What will things look like after another 5 years pass?

Paxson: E2E Routing Behavior

- Explore the large scale routing behavior in Internet
  - As seen by end hosts
- Trace data taken in 1994 and 1995
  - As Internet was transitioning from NSFNET to commercial
- Analyzed
  - Routing pathologies
  - Routing stability
  - Routing symmetry

Methodology

- Traceroute from A to B
  - Send out packets in rounds
  - Start with TTL 0, increment by 1 each round until reach dest
  - As packets go through network, routers decrement TTL
  - When TTL reaches 0, router replies with ICMP error packet
- Traced all routes between 37 hosts around world
  - Periodic, exponentially distributed, means of hours and days
  - Q: Is this representative? (most frequent eval question)
    - » Tiny fraction of all hosts, 8% of ASes, 50% of key ASes
- Thorough analysis of trace data

Traceroute Example

```
Tracing route to www.nytimes.com [208.48.26.200]
over a maximum of 30 hops:

  1   <10 ms   <10 ms   <10 ms  cse-fast-gateway.ucsd.edu [132.239.15.1]
  2   <10 ms   <10 ms   <10 ms  bigmama.ucsd.edu [132.239.254.5]
  3   <10 ms    10 ms   <10 ms  sdsc-gw.san-bb1.cerf.net [192.12.207.6]
  4   <10 ms    10 ms   <10 ms  atm2-0-6.san-bb1.cerf.net [134.24.12.25]
  5   <10 ms    10 ms    10 ms  pos0-0-155M.san-bb6.cerf.net [134.24.29.130]
  6   <10 ms    10 ms    10 ms  ge6-3-0.san-bb4.cerf.net [134.24.29.149]
  7    10 ms    10 ms    10 ms  pos3-0-622M.lax-bb5.cerf.net [134.24.29.14]
  8    10 ms    10 ms    10 ms  pos5-0-622M.lax-bb4.cerf.net [134.24.33.169]
  9    30 ms    20 ms    20 ms  pos6-0-622M.sfo-bb3.cerf.net [134.24.29.233]
 10    30 ms    20 ms    30 ms  pos5-0-0-155M.sjc-bb2.cerf.net [134.24.29.22]
 11    10 ms    20 ms    20 ms  208.50.172.61
 12    20 ms    10 ms    20 ms  so0-1-0-622M.cr1.pao2.gblx.net [208.48.118.121]
 13    80 ms    70 ms    80 ms  pos0-0-622M.cr1.NYC2.gblx.net [206.132.249.154]
 14    80 ms    80 ms    80 ms  pos4-0-622M.cr2.LGA2.gblx.net [208.48.234.106]
 15    71 ms    90 ms    80 ms  pos1-0-622M.hr3.LGA2.gblx.net [208.48.234.121]
 16    80 ms    80 ms    81 ms  208.48.26.200
```


Routing Pathologies

- Persistent loops – loop unresolved by end of tracert
  → 0.13-0.16%
- Temporary loops – loop resolved by end of tracert
  → 0.055-0.078%
- Erroneous routing – wrong path
  → 0.004% (1 instance)
- Mid-stream change – routing failure as it happened
  → 0.16 // 0.44%
- Infrastructure failure – route terminates inside network
  → 0.21 // 0.48%
- Outage >= 30 sec – repeatedly dropped probe packets
  → 0.96% // 2.2%
- Total: Significant and getting worse
  → 1.5% // 3.4% -- 1 in 30 your connection runs afoul of a bad route

Routing Stability

- How often do routes between two end-points change?
- Prevalence – How likely am I to observe the same route in the future?
  - Host median is 82% (strongly dominated by a single route)
- Persistence – How long before a route will change?
  - Routing changes occur over a range of time scales
  - 10s of minutes – 9%
  - 6+ hours – 23%
  - Days – 68%
  - Changes occur within a network (AS) though

Routing Asymmetry

- How often do the routes A→B and B→A differ?
  - Routes differing in cities traversed: 49% (!)
- Reflection of policy routing
  - If shortest path, asymmetry should be rare
- Aside: Asymmetry is a problem
  - Easier to measure in one direction but not the other
  - Would be nice to assume its representative of both – too bad
  - Q: Even with stable routes, why might delay HTTP request delay be shorter than response delay (UCSD to Yahoo!)?

Discussion

- If we redid experiment today, what do you think we would find?
  - Final project: Use CAIDA's skitter measurements?
- Why?

  1. Also, the paper just states the behavior, but it does not give the reasons why Internet behave like this? Is it because of the protocols or else?

  2. The study can identify the "symptoms" but not necessarily the "disease".

  → Excellent question, a topic of recent research (Labovitz)
- Really a crisis?
  - Would the authors still make the same claim if the rate had doubled from 0.5-1.0 %? Using the world doubling seems to make it worse than it is. I don't think that they would have had the same effectiveness if they said the chance increase by 1.9%?

For Next Time...

- Dmitrii is putting together a Web page describing how to use ns for the midterm project
- He is going to give an overview on Tuesday
- Read Web page before his overview