E6998-02: Internet Routing

Lecture 16
Border Gateway Protocol, Part V

John Ioannidis
AT&T Labs – Research

ji+ir@cs.columbia.edu
Announcements

Lectures 1-16 are available.

Homework 4 is available, due 11/14.
Multihoming

• Connecting to multiple providers.
• Backup links (we’ve already examined this).
  – The backup link is idle unless the primary goes down.
  – Slow is better than dead!
  – We’ve already covered this.
• Load sharing / load balancing / redundancy.
  – To the same provider.
  – To different providers.
Redundancy Issues

• Not just two ISPs!
• Redundant telco lines.
• Redundant power.
• Redundant exit points from the building!
• Redundant routers.
  – Make sure any additional hardware does not become a single point of failure!
• Redundant …
Multihoming Issues

- Addressing.
  - Pick addresses from upstream (main) provider.
  - Use addresses from both providers.
  - Get addresses allocated from ARIN/RIPE/APNIC.
- Routing.
  - Where/how to advertise prefixes.
    - Affects incoming traffic.
  - Where/how to set up own IGP.
    - Affects outgoing traffic.
- DNS
- Higher-layer protocols.
Dual Links

- Simplest cast: two distinct telco lines between the same pair of routers.
- Protects against link failure.
Dual Routers

- Different Configurations protects against router or link failure.
- A1/A2, B1/B2, C1/C2 are “near” each other.
  - IGP handles everything.
  - No BGP tricks involved.
Dual \{Links, Routers\} cont’d

- These configurations add redundancy.
- Also enable load sharing/load balancing between the links.
- Traffic is (usually) split on a \textbf{per-flow} basis.
  \begin{itemize}
  \item \textit{Flow}: (protocol, src, dst, src-port, dst-port).
  \item Performance reasons (can be done on the linecard).
  \item Per-packet split possible at much higher CPU burden.
    \begin{itemize}
    \item Or by using MUXes or multipoint PPP (below the network layer).
    \item Packet ordering maintained.
    \end{itemize}
  \item At least across the redundant hop.
  \end{itemize}
- OSPF can use equal-cost paths.
Multihoming to a Single Provider

- ... when access links are “far” from each other.
- ISP advertises defaults to customer.
  - Customer’s IGP ensures packets take the closest egress router (B or D).
- Customer advertises more-specifics with MED to force cold-potato routing.
Cold-Potato with MEDs

- MED takes precedence over IGP distance.
Multihoming to Multiple Providers

Customer
AS 666
??????/??

ISP A
AS 7018
12.0.0.0/8

ISP B
AS 1
4.0.0.0/8

ISP S
AS 1239
Own Address Space

• Great if you can get it!
  – And if you’re big enough.
• If the prefix is too long (> /24), it may not get through filters.
  – Lose connectivity from parts of the Internet.
• It does get redundancy.
• Does it get us good load-sharing?
  – Depends on the relative sizes of ISP A and ISP B.
• If equally “important”
  – roughly half the traffic will be coming from each
  – roughly half the announcements will be “better” from one of the two
    • resulting in outbound load sharing.
• Otherwise, may use AS_PATH padding to shed some traffic.
Address Space from Both ISPs

• With the service agreement comes address space.
  – 12.96.16.0/20 from ISP A.
  – 4.99.32.0/21 from ISP B.
• Announce the 12… space to A, and the 4… space to B.
  – (or not announce at all).
• Load sharing depends on source/destination of bulk of traffic.
• No redundancy.
  – If one link goes down, half of Customer’s address space is unreachable.
  – And unusable (no return routes).
• Use DNS round-robin to respond with addresses from both spaces.
  – Incoming connections will chose an address at random.
  – Not optimal in half the cases.
• How to pick address for outgoing connection?
  – Allocate address by region.
  – Random.
• Problems if ISPs do ingress filtering.
• Use of NAT has been suggested (arrrgggggghhhh!)
Address Space from one ISP

- Outgoing traffic **from** Customer is not affected.
What does AS3 Advertise?

- Customer AS 3
  - 5.2.0.0/16

- ISP A
  - AS 1
  - 5.0.0.0/8

- ISP B
  - AS 2
  - 9.0.0.0/8

- ISP S
  - AS 4
• Customer advertises its prefix to both its ISPs.
• ISP A (and its customers) now knows how to reach 5.2.0.0/16.
• ISB B (and its customers) also knows how to reach 5.2.0.0/16.
  – Although it gets 5.0.0.0/8 from ISP A.
  • Longest-prefix match.

[ ISP B could in some situations filter 5.2.0.0 ]
• What does ISP S (and the rest of the net know)?

• ISP B advertises the longer prefix to S.
• S now sends all traffic for 5.2.0.0/16 via B!
• This can lead to massive asymmetry!
  • Depends on relative amts of traffic from A vs. B+S
What is being advertised?

Customer
AS 3
5.2.0.0/16

ISP A
AS 1
5.0.0.0/8

ISP B
AS 2
9.0.0.0/8

ISP S
AS 4

5.2.0.0/16 3

5/8 1

5.2.0.0/16 3 1

5.2.0.0/16 3

5/8 1

5.2.0.0/16 3 2

9/8 2

5.2.0.0/16 3 2

9/8 2

5.2.0.0/16 3
• ISP A had to “punch a hole” in its aggregation policy.
• What is carried in ISP A’s I-BGP?
  – ISP-A knows that Customer is a proper subset.
  – If the access router does not readvertise inside I-BGP the more-specific, traffic for Customer would go out via ISP B!
    • Access router has to be configured accordingly.
• Customer and ISP A must run BGP.
  – I.e., A’s access router can’t just inject a static route.

• ISP S has the more-specific for Customer from both ISP A and ISP B.
  – Will route traffic for Customer properly.
Aggregation

• Address aggregation: announcing one less-specific prefix in lieu of many more-specific prefixes.

• Example:
  – Provider has a /12.
  – Customers are allocated /16s through /24s from that space.
  – Provider filters the more-specifics and only announces the /12 to its peers.

• More-specifics may still need to be carried inside I-BGP.
  – Finer-level aggregation on access routers.
  – (e.g.) Sixteen /24 customers are on an access router.
  – Access router advertises a /20 into the I-BGP mesh.

• More-specifics may still be announced (e.g., with NOEXPORT) to some peers.
Aggregation and Filtering

• External aggregation: provider only announces aggregates to its peers, not individual customer more-specifics.
• Internal aggregation: longer prefixes allocated to access routers, so that fewer routes are carried in I-BGP.
• Many times providers have to de-aggregate.
  – For multi-homed customers.
• Some providers do not allow in (filter) prefixes longer than /19 or /20 from aggregatable address space (post-CIDR allocations).
  – Contentious issue.
• Deaggregation leading cause of BGP table size.
  – “Grazing the commons”
Routing Table Size

- Source: http://bgp.potaroo.net/
- Active (used for the FIB) table.
BGP Scaling Issues

- Previous graph shows active routes (in the “Loc-RIB”).
- Many more routes floating around.
- Can’t just “add more memory”.
  - FIB memory is expensive, on linecards.
  - CPU/link capacity still an issue.
- Both the number of routes and the rate of UPDATEs (and their first derivatives) are scaling issues.
- Moore’s law only means we have to keep buying new routers!
- For a good time, go to telnet://route-views.oregon-ix.net/
- Chief problem: (at least) one route per advertised prefix.
  - De-aggregation due to multihoming a main source of the problem.
  - Switching to IPv6 doesn’t fix this!
  - Need a better routing architecture?
AS Numbers

• About 14K already.
• Increasing faster than linearly.
  – Current derivative: 2K/year.
• Source of new AS numbers:
  – New ISPs.
  – New multihomed customers.
• At this rate, we run out around 2007-2010.
  – IPv6 doesn’t fix this either!
• Suggestions:
  – 4-byte AS numbers (draft-ietf-idr-as4bytes-05.txt).
  – ASE (AS Number Substitution on Egress (AitFotL)).
    • Another cause of MOAS conflicts.
Route Flapping

- Routing instability.
- Route disappears, appears again, disappears again…
  - Withdrawal, announcement, withdrawal, announcement…
- Visible to the entire Internet.
  - Wastes resources, triggers more instability.
- Some causes of Route Flapping:
  - Flaky inter-AS links.
  - Flaky or insufficient hardware.
  - Link congestion.
  - IGP instability.
  - Operator error.
Link Instability

• The first three are examples of link instability.
  – Link itself fails.
  – Router/router interface fails.
  – Messages can’t get through.
• When a link goes down, routers withdraw routes associated with this link.
  – Customer-ISP.
  – ISP-ISP.
• Announcements travel throughout the default-free zone.
• Aggregation may mask downstream flapping.
  – Does not work for multihoming
IGP Instability

- IGP route-preference rule exports instability.
IGP Instability

- MEDs can export internal instability.
Route Flap Dampening

- Router detects route flapping.
- **Penalty:**
  - Increased each time a route flaps.
  - Decreased over time.
- If penalty threshold exceeded (*suppress limit*), route is suppressed.
- Until penalty drops below a certain level (*reuse limit*).