CSCI-1680
Network Layer:
Inter-domain Routing

Nick DeMarinis

Based partly on lecture notes by Rachit Agarwal, Rodrigo Fonseca, Jennifer Rexford,
Rob Sherwood, David Mazières, Phil Levis, John Jannotti
Administrivia

- IP: Due next Thursday (10/19)
- HW2: As soon as I can get there
Relationships between AS drive policy:

- **Customer->Provider:** Customer pays provider to advertise its routes, send it traffic
  - \( \Rightarrow \) Y pays C
  - \( \Rightarrow \) X pays B, C (multihomed)

- \( \Rightarrow \text{B is transit [provider] for X: Traffic destined for X goes through B} \)
- \( \Rightarrow \text{X is not transit for B, C: Traffic from B->C must not go through X!} \)

=> Why not? X gains nothing!

Example from Kurose and Ross, 5th Ed
How to turn this into a policy?

- **Selection Policy**: which path to use in your network

- **Export Policy**: which path to advertise
How to think about policies
Now to think about policies:

- **Control Plane:**
  - BGP updates from neighbors
  - Select policy
  - Find "best" route
  - Export policy
  - Forwarding table updates
    - Local route info
  - Data plane (per-packet)
  - Traffic sent out from this AS

- There are different BGP updates you send to your neighbors.
AS relationships

• Customer pays provider for connectivity
  – E.g. Brown contracts with OSHEAN
  – Customer is stub, provider is a transit

• Many customers are multi-homed
  – E.g., OSHEAN connects to Level3, Cogent

• Typical policies:
  – Provider tells all neighbors how to reach customer
  – Provider wants to send traffic to customers ($$$)
  – Customer does not provide transit service
Peer Relationships

• Peer ASs agree to exchange traffic for free
  – Penalties/Renegotiate if imbalance

• Tier 1 ISPs have no default route: all peer with each other

• You are Tier i + 1 if you have a default route to a Tier i

• Typical policies
  – AS only exports customer routes to peer
  – AS exports a peer’s routes only to its customers
  – Goal: avoid being transit when no gain
Typical route selection policy

In decreasing priority order:

1. Make or save money (send to customer > peer > provider)
2. Try to maximize performance (smallest AS path length)
3. Minimize use of my network bandwidth ("hot potato routing")
4. ...

\[ A - B - C \]
**How to Think About Export Policies**

**Ca.0 - Reford Principles**

**Given:** ISP A has:
- Customers: X, Y
- Peer with B, C
- Customer of Q

---

**If Prefix is advertised by:**
- **Customer (e.g., X,Y)**: Export Prefix to... Everyone! (X,Y, G, B, Q)
- **Peer (e.g., B)**: Customers only (Y, Y) (not, C, Q)
- **Provider (Q)**: Customer only (X, Y)

---

**Goal:** Don't become transit if no gain!
Typical Export Policy

<table>
<thead>
<tr>
<th>Destination prefix advertised by...</th>
<th>Export route to...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Customer</td>
<td>Everyone (providers, peers, other customers...)</td>
</tr>
<tr>
<td>Peer</td>
<td>Customers only</td>
</tr>
<tr>
<td>Provider</td>
<td>Customers only</td>
</tr>
</tbody>
</table>

Known as Gao-Rexford principles: define common practices for AS relationships
Gao-Rexford Model

• (simplified) Two types of relationships: peers and customer/provicer

• Export rules:
  – Customer route may be exported to all neighbors
  – Peer or provider route is only exported to customers

• Preference rules:
  – Prefer routes through customer ($$)

• If all ASes follow this, shown to lead to stable network
How to prevent X from forwarding transit between B and C?

X never tells B about C (or vice versa)

How to avoid transit between CBA?

B never tells A about C

Example from Kurose and Ross, 5th Ed
What can go wrong?
How to advertise your prefixes?

Try to aggregate (summarize) prefixes for networks you own, but not always possible

More specific prefix => More preferred
=> Can have policy, security implications…
How to advertise your prefixes?

Try to aggregate (summarize) prefixes for networks you own, but not always possible.
How to advertise your prefixes?

Try to aggregate (summarize) prefixes for networks you own, but not always possible.
IP PREFIXES/ROUTE AGGREGATION

138.16.0.0/16  138.16.x.x

IDEA: ALLOCATE SMALLER NETWORKS FROM ONE PREFIX.

For example, could divide into two networks:

1. 138.16.9.0/17
2. 138.16.128.0/17

IDEA: AS 3 combines, or aggregates prefixes for its customers.

⇒ USE HIERARCHY OF ADDRESSES!

IN PRACTICE... NOT SO EASY
How to advertise your prefixes?

Try to aggregate (summarize) prefixes for networks you own, but not always possible.

Problem: smaller allocations => more prefixes in table
=> Forwarding table size limited by fast memory (TCAM) inside routers.
What can lead to table growth?

- More addresses being allocated
- Fragmentation
  - Multihoming
  - Change of ISPs
  - Address re-selling
Map of the Internet, 2021 (via BGP)
OPTE project
BGP Table Growth

Source: bgp.potaroo.net
BGP Table Growth for v6

Source: bgp.potaroo.net
How big can the table get?

- August 12, 2014: the full IPv4 BGP table reached 512k prefixes
- March 5, 2019: 768k prefixes
How big can the table get?

- August 12, 2014: the full IPv4 BGP table reached 512k prefixes
- March 5, 2019: 768k prefixes

Older routers run out of space
=> Outages
Peering Drama

• Cogent vs. Level3 were peers
• In 2003, Level3 decided to start charging Cogent
• Cogent said no
• Internet partition: Cogent’s customers couldn’t get to Level3’s customers and vice-versa
  – Other ISPs were affected as well
• Took 3 weeks to reach an undisclosed agreement
BGP can be fragile!

- Individual router configurations and policy can affect whole network

- Consequences sometimes disastrous…
BGP Problems and Security Issues
Who owns a prefix?

- Allocated by Internet authorities
  - Regional Internet Registries (ARIN, RIPE, APNIC)
  - Internet Service Providers

- Ideally, AS who owns prefix (or its providers) should advertise it

- However: BGP does not verify this

\[ \text{\underline{NO BUILT-IN WAY TO VERIFY OWNERSHIP.}}\]
The Five RIRs

- ARIN
- RIPE NCC
- AFRINIC
- LACNIC
- APNIC
Prefix Hijacking

- Problem: WHO "owns" a prefix?

- WHO is ALLOWED to ORIGINATE a prefix?

- BGP by default does NOT verify announce match THE network that ORIGINATES.
  - AS's have their own security policies but not enforced.

WHAT CAN HAPPEN?

An actor who does not own a prefix can advertise it to INTERCEPT TRAFFIC.
WHAT CAN YOU DO?
- INTERCEPT OR REDIRECT PACKETS FOR A
- SNOOPING
- MODIFY/ SLOW DOWN TRAFFIC.

HARD TO DEBUG BECAUSE MIGHT ONLY BE VISIBLE FROM CERTAIN PARTS OF NETWORK!

IF B ADVERTISES MORE SPECIFIC PREFIX, IT WINS! ALL TRAFFIC GOES THERE, BECAUSE MORE SPECIFIC PREFixes ARE USUALLY PREFERABLE.
What can go wrong?
Some Notable incidents

June 24, 2019: Misconfigured small customer router accepted lots of transit traffic

Jérôme Fleury

[URGENT] Route-leak from your customer

To: CaryNMC-IP@one.verizon.com, peering@verizon.com, help4u@verizon.com,

At this level, solving problems involves a lot of human expertise!
Pakistan Youtube incident

- Youtube’s has prefix 208.65.152.0/22
- Pakistan’s government order Youtube blocked
- Pakistan Telecom (AS 17557) announces 208.65.153.0/24 in the wrong direction (outwards!)
- Longest prefix match caused worldwide outage
- http://www.youtube.com/watch?v=IzLPKuAOe50
• ISP outage in Russian-occupied city of Kherson, Ukraine
• Comes back several days later... with traffic routed through a Russian ISP

Many other incidents

- China incident, April 8\(^\text{th}\) 2010
  - China Telecom’s AS23724 generally announces 40 prefixes
  - On April 8\(^\text{th}\), announced ~37,000 prefixes
  - About 10% leaked outside of China
  - Suddenly, going to www.dell.com might have you routing through AS23724!

Russian hackers intercept Amazon DNS, steal $160K in cryptocurrency
Egypt Incident

Number of Egyptian networks

<table>
<thead>
<tr>
<th>Date</th>
<th>00:00</th>
<th>02:00</th>
<th>16:00</th>
<th>20:00</th>
<th>00:00</th>
<th>18:00</th>
<th>22:00</th>
<th>10:00</th>
<th>12:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>11-01-27</td>
<td>2903</td>
<td>327</td>
<td>239</td>
<td>241</td>
<td>242</td>
<td>243</td>
<td>134</td>
<td>2539</td>
<td>2825</td>
</tr>
</tbody>
</table>

Source: BGPMon (http://bgpmon.net/blog/?p=480)
What can be done?

Originally: Internet Routing Registries (IRRs): public database listing IP allocations

route: 10.0.0.0/8
descr: University of Blogging
descr: Anytown, USA
origin: AS65099
mnt-by: MNT-UNIVERSITY
notify: person@example.com
changed: person@example.com 20180101
source: RADB

But, database not verified and often incomplete/wrong
What can be done?

$whois -h whois.radb.net AS14325
aut-num:       AS14325
as-name:       ASN-OSHEAN
descr:          OSHEAN, Inc.
import:         from AS14325:AS-MBRS accept PeerAS
mp-import:      from AS14325:AS-MBRS accept PeerAS
export:         to AS-ANY announce AS14325:AS-MBRS
mp-export:      to AS-ANY announce AS14325:AS-MBRS
admin-c:        Tim Rue
tech-c:         Ventsislav Gotov
notify:         vgotov@oshean.org
mnt-by:         MAINT-AS14325
changed:        vgotov@oshean.org 20210512
source:         RADB
Proposed Solution: RPKI

• Based on a public key infrastructure
• Address attestations
  – Claims the right to originate a prefix
  – Signed and distributed out of band, checked on BGP updates
  – Checked through delegation chain from ICANN
• Can avoid
  – Prefix hijacking
  – Addition, removal, or reordering of intermediate ASes
Proposed Solution: RPKI

• Every AS adds **signature** of its route info in database
  – Max prefix size, etc.

• Other ASes using routes can **cryptographically verify** advertised routes against signature

• Can avoid
  – Prefix hijacking
  – Addition, removal, or reordering of intermediate ASes
RPKI deployment

RPKI-ROV Analysis of Unique Prefix-Origin Pairs (IPv4)

Valid: 35.12%
Invalid: 0.74%

Unique P-O
TOTAL: 996,018
Not-Found: 638,780
64.13%

Valid:349,820
Not-Found:638,780
Invalid:7,418

Not-Found: 64.13%
RPKI at Brown?

FAILURE

Your ISP (Verizon, AS701) does not implement BGP safely. It should be using RPKI to protect the Internet from BGP hijacks. Tweet this →

Details

- fetch https://valid.rpki.cloudflare.com
  ✔ correctly accepted valid prefixes

- fetch https://invalid.rpki.cloudflare.com
  ✗ incorrectly accepted invalid prefixes
EXTRA CONTENT
WE DID NOT
COVER
What can be done?

Brenda's ISP

Can contain some info on this AS's policy.

In theory, should reflect how BGP announcements are sent.
Proposed Solution: RPKI

- Based on a public key infrastructure
- Address attestations
  - Claims the right to originate a prefix
  - Signed and distributed out of band, checked on BGP updates
  - Checked through delegation chain from ICANN
- Can avoid
  - Prefix hijacking
  - Addition, removal, or reordering of intermediate ASes

\[\text{Can work, if everyone cooperates.}\]
BGP Protocol Details

• **BGP speakers**: nodes that communicates with other ASes over BGP

• Speakers connect over TCP on port 179

• Exact protocol details are out of scope for this class; most important messages have type UPDATE
Prefixes

- Nodes in local network share prefix
  - Key to decide whether to send message locally
- Prefixes can also aggregate multiple networks
  - E.g., 100.20.33.128/25, 100.20.33.0/25 -> 100.20.33.0/24
- If networks connected hierarchically, can have significant aggregation
- But allocations aren’t so hierarchical… what does this mean?
Anatomy of an UPDATE

- Withdrawn routes: list of withdrawn IP prefixes
- Network Layer Reachability Information (NLRI)
  - List of prefixes to which path attributes apply
- Path attributes
  - ORIGIN, AS_PATH, NEXT_HOP, MULTI-EXIT-DISC, LOCAL_PREF, ATOMIC_AGGREGATE, AGGREGATOR, ...
  - Extensible: can add new types of attributes
Example

- NLRI: 128.148.0.0/16
- AS-Path: ASN 44444 3356 14325 11078
- Next Hop IP
- Various knobs for traffic engineering:
  - Metric, weight, LocalPath, MED, Communities
  - Lots of voodoo
Demo: AS11078
BGP Security Goals

- Confidential message exchange between neighbors
- Validity of routing information
  - Origin, Path, Policy
- Correspondence to the data path
Origin: IP Address Ownership and Hijacking

• IP address block assignment
  – Regional Internet Registries (ARIN, RIPE, APNIC)
  – Internet Service Providers

• Proper origination of a prefix into BGP
  – By the AS who owns the prefix
  – ... or, by its upstream provider(s) in its behalf

• However, what’s to stop someone else?
  – Prefix hijacking: another AS originates the prefix
  – BGP does not verify that the AS is authorized
  – Registries of prefix ownership are inaccurate
Prefix Hijacking

- Consequences for the affected ASes
  - Blackhole: data traffic is discarded
  - Snooping: data traffic is inspected, and then redirected
  - Impersonation: data traffic is sent to bogus destinations
How to Hijack a Prefix

• The hijacking AS has
  – Router with eBGP session(s)
  – Configured to originate the prefix

• Getting access to the router
  – Network operator makes configuration mistake
  – Disgruntled operator launches an attack
  – Outsider breaks into the router and reconfigures

• Getting other ASes to believe bogus route
  – Neighbor ASes not filtering the routes
  – … e.g., by allowing only expected prefixes
  – But, specifying filters on peering links is hard
Many other incidents

- Spammers steal unused IP space to hide
  - Announce very short prefixes (e.g., /8). Why?
  - For a short amount of time

- China incident, April 8th 2010
  - China Telecom’s AS23724 generally announces 40 prefixes
  - On April 8th, announced ~37,000 prefixes
  - About 10% leaked outside of China
  - Suddenly, going to www.dell.com might have you routing through AS23724!
Attacks on BGP Paths

• Remove an AS from the path
  – E.g., 701 3715 88 -> 701 88

• Why?
  – Attract sources that would normally avoid AS 3715
  – Make path through you look more attractive
  – Make AS 88 look like it is closer to the core
  – Can fool loop detection!

• May be hard to tell whether this is a lie
  – 88 could indeed connect directly to 701!
Attacks on BGP Paths

• Adding ASes to the path
  – E.g., 701 88 -> 701 3715 88

• Why?
  – Trigger loop detection in AS 3715
    • This would block unwanted traffic from AS 3715!
  – Make your AS look more connected

• Who can tell this is a lie?
  – AS 3715 could, if it could see the route
  – AS 88 could, but would it really care?
Proposed Solution: S-BGP

• Based on a public key infrastructure
• Address attestations
  – Claims the right to originate a prefix
  – Signed and distributed out of band
  – Checked through delegation chain from ICANN
• Route attestations
  – Attribute in BGP update message
  – Signed by each AS as route along path
• S-BGP can avoid
  – Prefix hijacking
  – Addition, removal, or reordering of intermediate ASes
S-BGP Deployment

• Very challenging
  – PKI (RPKI)
  – Accurate address registries
  – Need to perform cryptographic operations on all path operations
  – Flag day almost impossible
  – Incremental deployment offers little incentive

• But there is hope! [Goldberg et al, 2011]
  – Road to incremental deployment
  – Change rules to break ties for secure paths
  – If a few top Tier-1 ISPs
    • Plus their respective stub clients deploy simplified version (just sign, not validate)
    • Gains in traffic => $ => adoption!
FAILURE

Your ISP (Verizon, AS701) does not implement BGP safely. It should be using RPKI to protect the Internet from BGP hijacks. Tweet this →

Details

Fetch https://valid.rpki.cloudflare.com

✔ correctly accepted valid prefixes

Fetch https://invalid.rpki.cloudflare.com

✗ incorrectly accepted invalid prefixes
Data Plane Attacks

- Routers/ASes can advertise one route, but not necessarily follow it!
- May drop packets
  - Or a fraction of packets
  - What if you just slow down some traffic?
- Can send packets in a different direction
  - Impersonation attack
  - Snooping attack
- How to detect?
  - Congestion or an attack?
  - Can let ping/traceroute packets go through
  - End-to-end checks?
- Harder to pull off, as you need control of a router
BGP Recap

• Key protocol that holds Internet routing together
• Path Vector Protocol among Autonomous Systems
• Policy, feasibility first; non-optimal routes
• Important security problems
Next Class

- Network layer wrap up