**BGP - Autonomous Systems**

- BGP protocol
- iBGP configuration
- AS-graph
- Internet topology

---

**Autonomous System**

- Definition: internet is network of networks glued by IP
- Within a network (intra-domain) any routing policy can be chosen
- A common routing policy is needed when routing between networks or domains
- A Domain is a network that has unified administrative routing policy
- Autonomous System (domain) or AS: Has a number assigned to it and provides routing information to other ASes

---

**Internet structure**

- AS provide reachability information to other ASes
- Within AS, local routing protocols used (optimize path metric)
- Inter-AS concerned with reachability and policy implementation
  - Usually $$$ involved with relationships

---

**References**

1. BGP tutorial - BPG4 case studies by Sam Halabi
2. BGP routing policies in ISP networks by Mathew Caesar and Jennifer Rexford
**Autonomous system**

- The actual entity that participates in interdomain routing
- Has a unique 16 bit number assigned
- Examples:
  - RUTGERS: 46, STANFORD; 32, MIT: 3, CMU: 9
  - AT&T: 6431, ...
  - Quest: 209, ...
  - Sprint: 1239, ...

How do ASes interconnect to provide global connectivity?
How does routing information get exchanged?
How is policy specified and implemented?

---

**Internet routing**

- Intra domain
  - OSPF, RIP
  - Route on IP addresses
  - Path metrics
- Inter domain
  - BGP
  - Route on AS numbers
  - Policy and business relations based

---

**BGP: basic idea**

- AS1 needs to inform AS2 that it can route to 192.8/16 and AS2 needs to inform AS1 that it can route to 128.16/16
- After this, what else
  - Route updates/changes
  - Policy: what is AS1 does not want to route to anyone else but its own domain?
  - What paths should be preferred?
  - This is essentially BGP

---

**BGP protocol**

- BGP uses TCP as its transport protocol, on port 179. On connection start, BGP peers exchange complete copies of their routing tables, which can be quite large. However, only changes (deltas) are then exchanged, which makes long running BGP sessions more efficient than shorter ones.
- Four Basic messages:
  - Open:
    - Establishes BGP session (uses TCP port #179)
  - Notification:
    - Report unusual conditions
  - Update:
    - Inform neighbor of new routes that become active
    - Inform neighbor of old routes that become inactive
  - Keepalive:
    - Inform neighbor that connection is still viable
OPEN Message

- During session establishment, two BGP speakers exchange their
  - AS numbers
  - BGP identifiers (usually one of the router’s IP addresses)
  - Select hold timer : max time before declaring peer is down
- A BGP speaker has option to refuse a session
- authentication information (optional)

<table>
<thead>
<tr>
<th>Version</th>
<th>My AS number</th>
<th>Hold Time</th>
<th>BGP Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optional Parameters</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTIFICATION and KEEPALIVE Messages

- NOTIFICATION
  - Indicates an error
  - terminates the TCP session
  - gives receiver an indication of why BGP session terminated
  - Examples: header errors, hold timer expiry, bad peer AS, bad BGP identifier, malformed attribute list, missing required attribute, AS routing loop, etc.

- KEEPALIVE
  - protocol requires some data to be sent periodically.
  - If no UPDATE to send within the specified time period, then send KEEPALIVE message to assure partner that connection is still alive

UPDATE Message

- used to either advertise and/or withdraw previously announced prefixes
- path attributes: list of attributes that pertain to ALL the prefixes in the Reachability Info field

<table>
<thead>
<tr>
<th>FORMAT:</th>
<th>Withdrawn route (2 octets)</th>
<th>Withdrawn routes (variable length)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total path attributes length (2 octets)</td>
<td>Path Attributes (variable length)</td>
</tr>
<tr>
<td></td>
<td>(NLRI) Reachability Information (variable length)</td>
<td></td>
</tr>
</tbody>
</table>

BGP update message

- Withdrawn Routes: Length field 2 Bytes
- Withdrawn route list
- Path attributes: Length field 2 bytes
- Path attributes list
- NLRI list : a list of entries
  - Length field (1 byte), Prefix (variable length)
- Path attributes apply to all the prefixes in the NLRI list
BGP attributes

- BGP protocol announcements carries with it several attributes
- Attribute describes characteristics of a prefix
- BGP chooses a single path for a given prefix based on attributes (can choose to ignore)
- BGP always announces the best path to neighbors

Attributes
- 1 ORIGIN
- 2 AS_PATH
- 3 NEXT_HOP
- 4 MED
- 5 LOCAL_PREF
- 6 WEIGHT
- 7 COMMUNITY
- 8 AGGREGATOR

PATH ATTRIBUTES

- ORIGIN(TYPE CODE = 1):
  - Who originated the announcement? Where was a prefix injected into BGP?
  - Manually configured, directly connected, or other intra routing protocols
  - IGP, EGP, default – improbable (learned from some other means)

- AS-PATH (TYPE CODE = 2):
  - A list of ASs through which the announcement for a prefix has passed
  - Each AS prepends its AS to the AS-PATH attribute when forwarding an announcement
  - Used to detect and prevent loops
  - AS length can be used to select among routes unless a LOCAL_PREF attribute overrides

Use of local pref

- Used to indicate preference among multiple paths for the same prefix anywhere in the internet.
- The higher the value the more it is preferred
- Default value is 100
- Local to the AS (non-transitive)
- Often used to select a specific exit point for outbound traffic
- Override influence of AS path length

Attribute: Local Preference (type code = 5)

- Used to indicate preference among multiple paths for the same prefix anywhere in the internet.
- The higher the value the more it is preferred
- Default value is 100
- Local to the AS (non-transitive)
- Often used to select a specific exit point for outbound traffic
- Override influence of AS path length
Attribute: NEXT HOP (code=3)

When AS boundaries are crossed, the next hop field is changed and replaced with the address of the border router.

I.BGP address of external neighbour
I.BGP address of external neighbour

I.BGP address of external neighbour

E? 128.64.8.1

Use of next hop

Next hop=138.64.8.2 for 198.6.8/24

AS1

MED=100

MED=500

AS2

MED values from the same AS are compared
A lower MED value is preferred
MED values exchanged between ASs - non-transitive

Ref: BGP Tutorial – Sam Halabi- Cisco

Attribute: Multi-Exit Discriminator (MED) (code=4)

- when AS’s interconnected via 2 or more links
- AS path length are same
- AS announcing prefix, sets MED value
- enables AS2 to indicate its preference (lower MED is better)
- AS receiving prefix uses MED to select link
- a way to specify how close a prefix is to the link it is announced on

Link B

Link A

AS1

AS2

Med values from the same AS are compared
A lower MED value is preferred
MED values exchanged between ASs - non-transitive
Routing Process Overview

Input policy engine → Decision process → Routes used by router → Output policy engine

Routes received from peers → Routing Information Base

Routing Information Base
- Store all BGP routes for each destination prefix
- Withdrawal message: remove the route entry
- Announcement message: update the route entry

Selecting the best route
- Consider all BGP routes for the prefix
- Apply rules for comparing the routes
- Select the one best route
  - Use this route in the forwarding table
  - Send this route to neighbors

BGP Decision Process: Path Selection on a Router

1. Choose route with highest LOCAL-PREF
2. If have more than 1 route, select route with shortest AS-PATH
3. If have more than 1 route, select according to lowest ORIGIN type where IGP < E-BGP < default
4. If have more than 1 route, select route with lowest MED value
5. Select e-BGP learned over i-BGP learned path
6. Select min cost path to NEXT HOP using IGP metrics (lowest IGP cost to BGP egress)
7. If have multiple internal paths, use lowest BGP Router ID to break the tie.

- See: BGP routing policies in ISP networks by Caesar and Rexford

Internal vs. External BGP

- Internal-BGP or i-BGP used to distribute routes within AS
- Egress routers use E-BGP or BGP
- R4 and R3 learn routes from R2 using i-BGP
- R1 and R2 talk e-BGP (different AS)
- R2, R4 and R2, R3 and R3, R4 talk i-BGP (same AS)
Internal BGP (I-BGP)

- Same messages as E-BGP
- Different rules about re-advertising prefixes:
  - Rule #1: Prefix learned from E-BGP can be advertised to I-BGP neighbor and vice-versa, but
  - Rule #2: Prefix learned from one I-BGP neighbor cannot be advertised to another I-BGP neighbor
  - Reason: no AS PATH within the same AS and thus danger of looping.
  - Means each I-BGP speaker must be connected directly with every other I-BGP within the same AS
  - Full MESH!!!

Route reflectors

Mesh does not Scale
O(N^2) sessions

Route reflectors

- Problem: requiring a full mesh of I-BGP sessions between all pairs of routers is hard to manage for large AS’s.
- Solution:
  - group routers into clusters.
  - Assign a leader to each cluster, called a route reflector (RR).
  - Members of a cluster are called clients of the RR
  - I-BGP Peering
    - clients peer only with their RR
    - RR’s must be fully meshed

Route Reflectors: Rule on

- If received from RR, reflect to clients
- If received from a client, reflect to RRs and clients
- If received from E-BGP, reflect to all - RRs and clients
- RR’s reflect only the best route to a given prefix, not all announcements they receive.
  - helps size of routing table
  - sometimes clients don’t need to carry full table
Announcement loop

CISCO manual on BGP configuration

Caution: Incorrectly setting BGP attributes for a route reflector can cause inconsistent routing, routing loops, or a loss of connectivity. Setting BGP attributes for a route reflector should be attempted only by an experienced network operator.

<table>
<thead>
<tr>
<th>Command</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>Router(config-router)# no bgp client-to-client reflection</code></td>
<td>Disables client-to-client route reflection.</td>
</tr>
</tbody>
</table>

RFC 4456 - BGP Route Reflectors

When a route is reflected, it is possible through misconfiguration to form route redistribution loops.

Avoiding Loops with Route Reflectors

- Loops cannot be detected by traditional approach using AS-PATH because AS-PATH not modified within an AS.
- Announcements could leave a cluster and re-enter it.
- Two new attributes introduced:
  - ORIGINATOR_ID: router id of route’s originator in AS
  - CLUSTER_LIST: a sequence of cluster id’s, set by RRs.

Announcement loops prevention

Inter-AS relationships

Peering vs transit

Determining AS relationships

- Drop if Originator ID= router id
- Drop if Cluster List contains ClusterID
Internet AS-structure

- ISPs or AS provide reachability information to other ASs
- Different sizes (Tier-1, Tier-2, Tier-3)
  - Global/transit, Regional, local
- Different types (transit, hub, multi-homed)
- Within AS, local routing protocols used (optimize path metric)
- Inter-AS concerned with reachability and policy implementation
  - Usually $$$ involved with relationships

AS graphs

- Model links to peers as edges
- Send path probes to various routers
- Find IP -> Name mapping from DNS
- Send probes from multiple vantage points
- Get an idea of out-degree for various AS
- Create internet map based on Route views and other tools
- Two Papers:
  - Heuristics for internet map discovery, Govindan et al, Infocom 2000
  - Measuring ISP topologies with rocket fuel, ACM SIGCOMM 2002

ISP Size Distribution

Total number of ISPs: 2700

http://www.caida.org/research/topology/as_core_network/
AS Commercial Agreements

- Provider-customer:
  - customer pays its provider for transit services
- Peer-peer:
  - exchange traffic between customers
  - Free of charge
- Sibling-sibling:
  - have mutual transit agreement
  - Reachability for the entire internet (load balancing)
  - Backup connectivity in the event of a failure of one of the providers
**AS Relationship Graph**

- **Provider-Customer**
  - Provider needs to provide all known routes to the customer
  - Provider needs to announce customer routes to everyone

- **Peer-Peer**
  - Peers exchange their route and customer routes
  - Peering is not transitive
  - E.g., Origin Prefix for P1 is x, customer C2 prefix is y
  - Origin prefix for P3 is a and customer C3 prefix is b
  - P3 will know x and y; P1 will know a and b
  - P2 will know ....

- **Transit vs Peering**
  - Peering ASs provide mutual reachability information about customers
    - Non-transitive relationships
    - Good for mutual agreements/direct connection
  - Transit ASs provide global reachability information
    - Involves money for carrying traffic
    - Can reach any destination
**Import Routes**

- Provider route
- Customer route
- Peer route
- ISP route

**Export Policy**

- An AS exports only best paths to its neighbors:
  - Guarantees that once the route is announced the AS is willing to transit traffic on that route.
- To Customers:
  - Announce all routes learned from peers, providers and customers, and self-origin routes.
- To Providers:
  - Announce routes learned from customers and self-origin routes; does not export its provider or peer routes.
- To Peers:
  - Announce routes learned from customers and self-origin routes, but does NOT export its provider routes or peer routes.
- To Sibling:
  - Announce its routes and routes of customers, and as well as its provider or peer routes.

**Export Routes**

- Provider route
- Customer route
- Peer route
- ISP route

**BGP export rules**

<table>
<thead>
<tr>
<th>Own Routes</th>
<th>Customer’s Routes</th>
<th>Sibling’s Route</th>
<th>Provider’s Route</th>
<th>Peer’s Route</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exporting to a Provider</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Exporting to a Customer</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Exporting to a Peer</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
</tr>
<tr>
<td>Exporting to a Sibling</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
**Route Propagation Policy**

- Constrained by contractual commercial agreements between administrative domains

![Diagram of ISPs and C]

- e.g., An AS does not provide transit services between its providers

**Determining AS Relationships**

- Relationships are confidential
- Bound by contractual agreements
- ISPs hide information about what they do
- Very difficult to get a global view of AS relationship graph
- No Facebook account for ISPs!!!

Determining AS relationships, Lixin Gao, ACM transactions on networking, 2001

---

**Why Infer AS Relationships?**

- Crucial part of Internet structure
  - Connectivity does not imply reachability
  - Connectivity alone cannot fully characterize structural properties of Internet
- No registry of AS relationships
  - Many ISPs are not willing to reveal their relationships to others in IRR
  - Relationships are evolving; hard to be up-to-date

**Applications of AS Relationships**

- Construct distance map
- Place proxy or mirror site servers
- Potentially avoid route divergence

- Help ISPs or domain administrators to achieve load balancing and congestion avoidance
- Help ISPs or companies to plan for future contractual agreements
- Help ISPs to reduce effect of misconfiguration and to debug router configuration files
**AS Relationship Graph**

![AS Relationship Graph]

**Routing Table Entry**

<table>
<thead>
<tr>
<th>Network</th>
<th>Next hop</th>
<th>AS Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.24.0/21</td>
<td>134.24.127.3</td>
<td>1740 1 i</td>
</tr>
<tr>
<td></td>
<td>194.68.130.254</td>
<td>5459 5413 1 i</td>
</tr>
<tr>
<td></td>
<td>158.43.133.48</td>
<td>1849 704 702 701 1 i</td>
</tr>
<tr>
<td></td>
<td>193.0.0.242</td>
<td>3333 286 1 i</td>
</tr>
<tr>
<td></td>
<td>144.228.240.93</td>
<td>1239 1 i</td>
</tr>
</tbody>
</table>

**Route Propagation/Traffic**

- Traffic flows opposite to the direction of route announcements
- Collect AS paths for various prefixes
- AS u and AS v have peering relationships:
  - u and v do not appear in paths for prefixes other than own
- AS u is a provider for AS v
  - u, v appears in prefixes
- AS u and v have sibling relationships:
  - u and v and v and u appear in paths for prefixes

**Traversal properties**

- If it is provider to customer edge, then it can be followed by another provider to customer or sibling-sibling edge
  - Downhill path
- If it customer to provider edge, then it can be followed by another customer to provider edge or sibling to sibling edge
  - Up hill path
- An uphill path can be followed by a downhill path
- Not allowed: a downhill path followed by another uphill path ➔ Valley free property
Valley-Free

Not Valley-Free

Heuristic Algorithms

- Infer provider-customer and sibling-sibling
  - basic
  - refined
- Infer peer-peer
  - final

Basic Algorithms

- Heuristics
  - Top provider has highest degree
  - A routing path with the top provider (u1, u2, utop, u3, u4) has customer to provider (u1, u2, utop) or provider to customer (utop, u3, u4)
- Infer:
  - u2 provides transit services for u1
  - u3 provides transit services for u4
- If in another path it is found that u1 provides transit services for u2 then conclude u1, u2 are siblings
**Refined Algorithm**

- **Heuristics**
  - What if BGP route announcements don’t adhere to valley-free property
  - BGP misconfigurations
  - Assume majority or a significant majority > L are correct
  - Majority rules
  - If majority concludes that
  - u2 provides transit services for u1, then Ignore paths that say u1 provides transit services u2
  - Conclude u2 provides transit services for u1
  - u2 to u1 is a provider to customer edge

---

**Peering classification**

- **Heuristics**
  - Peers don’t transit each other’s traffic
  - Degree of Peers are about the same
  - If every path is in the analysis
  - Eliminate AS pairs having transit relationships or not likely to peer
  - Atmost one consecutive pair can have peering relationship

---

**Why peer?**

- Reduce communication costs
  - Peering allows for sharing telco fees/pain!
- Lower latency
  - Direct connection is better
- Better Quality service

---

**Why not to peer**

- Traffic asymmetry/investment asymmetry
- Shared pain does not imply shared burden
- If link is down, who do you call?
- Transit may generate revenue as opposed to peer
- No Service level agreements/contractual obligations
Cost of peering vs transit

- Fixed cost of peering (colocation)
- Transit cost coming down
- Peering: better handle on performance

Remote Peering model

No CAPEX, No OPEX, only access cost ~1000$/month
Anatomy of a large European IXP – Anja Feldman SIGCOMM 2013

Network map 2012+

- IXP
- Regional Tier 2 Providers
- AS 1
- AS 2

- IXP central component
- Lots of local peering – rich fabric
- Even flatter AS topology than assumed