CS 640: Introduction to Computer Networks
Aditya Akella

Lecture 11 -
Inter-Domain Routing -
BGP (Border Gateway Protocol)

Intra-domain routing

- The Story So Far...
  - Routing protocols generate the forwarding table
  - Two styles: distance vector, link state
  - Scalability issues:
    - Distance vector protocols suffer from count-to-infinity
    - Link state protocols must flood information through network

- Today's lecture
  - How to make routing protocols support large networks
  - How to make routing protocols support business policies

Inter-domain Routing: Hierarchy

- "Flat" routing not suited for the Internet
  - Doesn't scale with network size
    - Storage: Each node cannot be expected to store routes to every destination network
    - Convergence times increase
    - Communication: Total message count increases
  - Administrative autonomy
    - Each internet network may want to run its network independently
      - E.g. hide topology information from competitors

- Solution: Hierarchy via autonomous systems
Internet’s Hierarchy

- What is an Autonomous System (AS)?
  - A set of routers under a single technical administration
  - Use an interior gateway protocol (IGP) and common metrics to route packets within the AS
  - Connect to other ASes using gateway routers
  - Use an exterior gateway protocol (EGP) to route packets to other ASes
  - IGP: OSPF, RIP (last class)
  - Today’s EGP: BGP version 4

An example

- Intra-AS routing algorithm + Inter-AS routing algorithm → Forwarding table

The Problem

- Easy when only one link leading to outside AS
- Much harder when two or more links to outside ASes
  - Which destinations reachable via a neighbor?
  - Propagate this information to other internal routers
  - Select a “good route” from multiple choices
- Inter-AS routing protocol
  - Communication between distinct ASes
  - Must be the same protocol
History

• Mid-80s: EGP
  - Reachability protocol (no shortest path)
  - Did not accommodate cycles (tree topology)
  - Evolved when all networks connected to NSF backbone

• Result: BGP introduced as routing protocol
  - Latest version = BGP 4
  - BGP-4 supports CIDR
  - Primary objective: connectivity not performance

BGP Preliminaries

• Pairs of routers exchange routing info over TCP connections (port 179)
  - One TCP connection for every pair of neighboring gateway routers
  - Routers called “BGP peers”
  - BGP peers exchange routing info as messages
  - TCP connection + messages → BGP session

• Neighbor ASes exchange info on which CIDR prefixes are reachable via them

Choices for Routing

• How to propagate routing information?

• Link state or distance vector?
  - No universal metric - policy decisions
  - Problems with distance-vector:
    - Very slow convergence
  - Problems with link state:
    - Metric used by ISPs not the same → loops
    - LS database too large - entire Internet

• BGP: Path vector
AS Numbers (ASNs)

ASNs are 16 bit values. 64512 through 65535 are "private". Currently over 15,000 in use.

- Geniux: 1
- MIT: 3
- CMU: 9
- UC San Diego: 7377
- AT&T: 7018, 6341, 5074, ...
- UUNET: 701, 702, 284, 12199, ...
- Sprint: 1239, 1240, 6211, 6242, ...

ASNs represent units of routing policy.

Distance Vector with Path

- Each routing update carries the entire AS-level path so far.
  - "AS.Path attribute"

- Loops are detected as follows:
  - When AS gets route, check if AS already in path
    - If yes, reject route
    - If no, add self and (possibly) advertise route further.
      - Advertisement depends on metrics/cost/preference etc.

- Advantage:
  - Metrics are local - AS chooses path, protocol ensures no loops.

Hop-by-hop Model

- BGP advertises to neighbors only those routes that it uses.
  - Consistent with the hop-by-hop Internet paradigm.
  - Consequence: hear only one route from neighbor.
    - (although neighbor may have chosen this from a large set of choices)
    - Could impact view into availability of paths.
Policy with BGP

- BGP provides capability for enforcing various policies

- Policies are not part of BGP; they are provided to BGP as configuration information

  * **Enforces** policies by
    - Choosing appropriate paths from multiple alternatives
    - Controlling advertisement to other ASs

Examples of BGP Policies

- A multi-homed AS refuses to act as transit
  - Limit path advertisement

- A multi-homed AS can become transit for some AS's
  - Only advertise paths to some AS's

- An AS can favor or disfavor certain AS's for traffic transit from itself

BGP Messages

- **Open**
  - Announces AS ID
  - Determines hold timer - interval between keep alive or update messages, zero interval implies no keep alive

- **Keep alive**
  - Sent periodically (but before hold timer expires) to peers to ensure connectivity
  - Sent in place of an UPDATE message

- **Notification**
  - Used for error notification
  - TCP connection is closed immediately after notification
BGP UPDATE Message

- List of withdrawn routes
- Network layer reachability information
  - List of reachable prefixes
- Path attributes
  - Origin
  - Path
  - LocalPref
  - MED
  - Metrics
- All prefixes advertised in message have same path attributes

Path Selection Criteria

- Attributes + external (policy) information
- Examples:
  - Policy considerations
    - Preference for AS
    - Presence or absence of certain AS
  - Hop count
  - Path origin

LOCAL PREF

- Local (within an AS) mechanism to provide relative priority among BGP exit points
- Prefer routers announced by one AS over another or general preference over routes
**AS_PATH**

- List of traversed AS's

```
AS 200
170.10.0.0/16
AS 300
AS 500
180.10.0.0/16
```

**Multi-Exit Discriminator (MED)**

- Hint to external neighbors about the preferred path into an AS
  - Different AS choose different scales
- Used when two AS's connect to each other in more than one place
  - More useful in a customer provider setting
  - Not honored in other settings
  - Will see later why

**MED**

- Hint to R1 to use R3 over R4 link
- Cannot compare AS40's values to AS30's

```
MED = 120
MED = 200
```
MED

- MED is typically used in provider/subscriber scenarios
- It can lead to unfairness if used between ISP because it may force one ISP to carry more traffic:

ISP1  ISP2
- ISP1 ignores MED from ISP2
- ISP2 obeys MED from ISP1
- ISP2 ends up carrying traffic most of the way

Decision Process (First cut)

- Rough processing order of attributes:
  - Select route with highest LOCAL-PREF
  - Select route with shortest AS-PATH
  - Apply MED (to routes learned from same neighbor)

- How to set the attributes?
  - Especially local_pref?
  - Policies in action

A Logical View of the Internet

- Tier 1 ISP
  - "default-free" with global reachability info
- Tier 2 ISP
  - Regional or country-wide
  - Typically route through Tier-1
  - Customer
- Tier 3/4 ISPs
  - Local
  - Route through higher tiers
- Stub AS
  - End network such as IBM or UW-Madison
Inter-ISP Relationships: Transit vs. Peering

These relationships have the greatest impact on BGP policies.

Illustrating BGP Policies

Which route should Frank pick to 13.13.0.0/16?

Policy I: Prefer Customer routing

Set appropriate "local pref" in policy preferences:
Higher local preference values are preferred.
Policy II: Import Routes

- From provider route
- From peer route
- From customer route
- From ISP route

Policy II: Export Routes

- To provider route
- To peer route
- To customer route
- From ISP route

Policy II: Valley-Free Routes

- "Valley-free" routing
  - Number links as (+1, 0, -1) for provider, peer and customer
  - In any valid path should see sequence of +1, followed by at most one 0, followed by sequence of -1
  - Why?
    - Consider the economics of the situation
- How to make these choices?
  - Prefer-customer routing: LOCAL_PREF
  - Valley-free routes: control route advertisements (see previous slide)
BGP Route Selection Summary

<table>
<thead>
<tr>
<th>Highest Local Preference</th>
<th>Enforce relationships E.g. prefer customer routes over peer routes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shortest AS Path</td>
<td>traffic engineering</td>
</tr>
<tr>
<td>Lowest MED</td>
<td></td>
</tr>
<tr>
<td>i-BGP &lt; e-BGP</td>
<td>traffic engineering</td>
</tr>
<tr>
<td>Lowest BGP cost to BGP ingress</td>
<td></td>
</tr>
<tr>
<td>Lowest router ID</td>
<td>Throw up hands and break ties</td>
</tr>
</tbody>
</table>

Internal vs. External BGP

- BGP can be used by R3 and R4 to learn routes
- How do R1 and R2 learn best routes?

- Use I-BGP
- Create a full mesh
- TCP connections
- Use this to exchanged BGP route information

Link Failures

- Two types of link failures:
  - Failure on an E-BGP link
  - Failure on an I-BGP Link
- These failures are treated completely different in BGP
- Why?
Failure on an E-BGP Link

- If the link R1-R2 goes down
- The TCP connection breaks
- BGP routes are removed
- This is the desired behavior

Failure on an I-BGP Link

- If link R1-R2 goes down, R1 and R2 should still be able to exchange traffic
- The indirect path through R3 must be used
- Thus, E-BGP and I-BGP must use different conventions with respect to TCP endpoints

Next Class

- Multicast
  - Service model
  - IGMP
  - IP Multicast routing protocols
  - Overlay-based multicast