Inter-Domain Routing

• Border Gateway Protocol (BGP)
• Assigned reading
  • [LAB00] Delayed Internet Routing Convergence
  • [Nor00] Internet Service Providers and Peering

Outline

• Routing hierarchy
• Internet structure
• External BGP (E-BGP)
• Internal BGP (I-BGP)

Routing Hierarchies

• Flat routing doesn’t scale
  • Each node cannot be expected to have routes to every destination (or destination network)
• Key observation
  • Need less information with increasing distance to destination
• Two radically different approaches for routing
  • The area hierarchy
  • The landmark hierarchy (discuss in routing alternatives)

Areas

• Divide network into areas
  • Areas can have nested sub-areas
  • Constraint: no path between two sub-areas of an area can exit that area
• Hierarchically address nodes in a network
  • Sequentially number top-level areas
  • Sub-areas of area are labeled relative to that area
  • Nodes are numbered relative to the smallest containing area

The Area Hierarchy
Routing

• Within area
  • Each node has routes to every other node
• Outside area
  • Each node has routes for other top-level areas only
  • Inter-area packets are routed to nearest appropriate border router
• Can result in sub-optimal paths

Path Sub-optimality

Outline

• Routing hierarchy
• Internet structure
• External BGP (E-BGP)
• Internal BGP (I-BGP)

Internet’s Area Hierarchy

• What is an Autonomous System (AS)?
  • A set of routers under a single technical administration, using an interior gateway protocol (IGP) and common metrics to route packets within the AS and using an exterior gateway protocol (EGP) to route packets to other AS’s
  • Sometimes AS’s use multiple IGPs and metrics, but appear as single AS’s to other AS’s
  • Each AS assigned unique ID
  • AS’s peer at network exchanges

Example

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
A Logical View of the Internet?

- After looking at RIP/OSPF descriptions
  - End-hosts connected to routers
  - Routers exchange messages to determine connectivity
  - NOT TRUE!

A Logical View of the Internet?

- RIP/OSPF not very scalable \(\rightarrow\) area hierarchies
  - But, ISP’s aren’t equal
    - Size
    - Connectivity
  - NOT TRUE EITHER!

A Logical View of the Internet

- Tier 1 ISP
  - “Default-free” with global reachability info
- Tier 2 ISP
  - Regional or country-wide
- Tier 3 ISP
  - Local

Transit vs. Peering

Choices

- Link state or distance vector?
  - No universal metric – policy decisions
- Problems with distance-vector:
  - Bellman-Ford algorithm may not converge
- Problems with link state:
  - Metric used by routers not the same – loops
  - LS database too large – entire Internet
  - May expose policies to other AS’s

Outline

- Routing hierarchy
- Internet structure
- External BGP (E-BGP)
- Internal BGP (I-BGP)
Solution: Distance Vector with Path

- Each routing update carries the entire path
- 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
- Advantage:
  - Metrics are local - AS chooses path, protocol ensures no loops

Interconnecting BGP Peers

- BGP uses TCP to connect peers
- Advantages:
  - Simplifies BGP
  - No need for periodic refresh - routes are valid until withdrawn, or the connection is lost
  - Incremental updates
- Disadvantages
  - Congestion control on a routing protocol?
  - Poor interaction during high load

Hop-by-hop Model

- BGP advertises to neighbors only those routes that it uses
  - Consistent with the hop-by-hop Internet paradigm
  - e.g., AS1 cannot tell AS2 to route to other AS’s in a manner different than what AS2 has chosen (need source routing for that)

Policy with BGP

- BGP provides capability for enforcing various policies
- Policies are not part of BGP: they are provided to BGP as configuration information
- BGP enforces policies by choosing paths from multiple alternatives and controlling advertisement to other AS’s

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 Common Header

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marker (security and message delineation)</td>
<td>16 bytes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length (2 bytes)</td>
<td>Type (1 byte)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE
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
  - Metrics
  - All prefixes advertised in message have same path attributes

Path Selection Criteria

- Information based on path attributes
- Attributes + external (policy) information
- Examples:
  - Hop count
  - Policy considerations
  - Preference for AS
  - Presence or absence of certain AS
  - Path origin
  - Link dynamics

LOCAL PREF

- Local (within an AS) mechanism to provide relative priority among BGP routers

LOCAL PREF – Common Uses

- Handle routes advertised to multi-homed transit customers
  - Should use direct connection
- Peering vs. transit
  - Prefer to use peering connection, why?
  - In general, customer > peer > provider
  - Use LOCAL PREF to ensure this

AS_PATH

- List of traversed AS’s
CIDR and BGP

What should T announce to Z?

Options
- Advertise all paths:
  - Path 1: through T can reach 197.8.0.0/23
  - Path 2: through T can reach 197.8.2.0/24
  - Path 3: through T can reach 197.8.3.0/24
- But this does not reduce routing tables! We would like to advertise:
  - Path 1: through T can reach 197.8.0.0/22

Sets and Sequences
- Problem: what do we list in the route?
  - List T: omitting information not acceptable, may lead to loops
  - List T, X, Y: misleading, appears as 3-hop path
- Solution: restructure AS Path attribute as:
  - Path: (Sequence (T), Set (X, Y))
  - If Z wants to advertise path:
    - Path: (Sequence (Z, T), Set (X, Y))
  - In practice used only if paths in set have same attributes

Multi-Exit Discriminator (MED)
- Hint to external neighbors about the preferred path into an AS
  - Non-transitive attribute (we will see later why)
  - Different AS choose different scales
- Used when two AS’s connect to each other in more than one place

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

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 ignores MED from ISP2
  - ISP2 obeys MED from ISP1
  - ISP2 ends up carrying traffic most of the way
Other Attributes
- **ORIGIN**
  - Source of route (IGP, EGP, other)
- **NEXT_HOP**
  - Address of next hop router to use
- Check out [http://www.cisco.com](http://www.cisco.com) for full explanation

Decision Process
- **Processing order of attributes:**
  - Select route with highest LOCAL-PREF
  - Select route with shortest AS-PATH
  - Apply MED (if routes learned from same neighbor)

Outline
- Routing hierarchy
- Internet structure
- External BGP (E-BGP)
- Internal BGP (I-BGP)

Internal vs. External BGP
- BGP can be used by R3 and R4 to learn routes
- How do R1 and R2 learn routes?
  - Option 1: Inject routes in IGP
    - Only works for small routing tables
  - Option 2: Use I-BGP

Internal BGP (I-BGP)
- Same messages as E-BGP
- Different rules about re-advertising prefixes:
  - Prefix learned from E-BGP can be advertised to I-BGP neighbor and vice-versa, but
  - 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.
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 Lecture: New Routing Ideas

- Border Gateway Protocol (BGP) cont.
- Overlay networks
- Active networks
- Assigned reading
  - [S+99] The End-to-End Effects of Internet Path Selection
  - [W99] Active network vision and reality: lessons from a capsule-based system