Internet Technology

09. Routing on the Internet

Paul Krzyzanowski
Rutgers University
Spring 2016
Summary

• Routing
  – Enable a host to determine the next hop on a least-cost route to a destination
  – Graph traversal problem
    • Graph $G = (N \text{ nodes, } E \text{ edges}) \Rightarrow \text{Network of } N \text{ hosts and } E \text{ links}$

• Global knowledge
  – Link State (LS) = Dijkstra’s algorithm
    • Each iteration, replace distances with more accurate values

• Local (neighbor) knowledge
  – Distance-Vector algorithm
  – Construct a distance vector to all nodes
  – Exchange information with neighbors until no changes to vector
A problem of scale

- There are over a billion hosts* on the Internet
  - That’s a LOT of routing information to store
  - Sending Link State updates would consume a lot of bandwidth
  - Distance Vector algorithm may never converge
    - Time to converge vs. time between any route changes

- Organizations may not want arbitrary routing through their infrastructure

What do we do?

*https://www.isc.org/network/survey/
Autonomous Systems (ASes)

Autonomous System

- Collection of routers and hosts that are under common administrative control
  - Typically one network service provider or large company
- Collection of subnets (routing prefixes ⇒ route aggregation)
- Present a common routing policy to the Internet
- Identified by an AS Number:
  - Internet Assigned Numbers Authority (IANA) → Regional Internet Registry (RIR)
Autonomous System

- Collection of routers and hosts that are under common administrative control
  - Typically one network service provider or large company
- Collection of subnets (routing prefixes ⇒ route aggregation)
- Present a common routing policy to the Internet
- Identified by an AS Number:
  - Internet Assigned Numbers Authority (IANA) → Regional Internet Registry (RIR)
<table>
<thead>
<tr>
<th><strong>Autonomous System Number</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
</tr>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td><strong>Handle</strong></td>
</tr>
<tr>
<td><strong>Organization</strong></td>
</tr>
<tr>
<td><strong>Registration Date</strong></td>
</tr>
<tr>
<td><strong>Last Updated</strong></td>
</tr>
<tr>
<td><strong>Comments</strong></td>
</tr>
<tr>
<td><strong>RESTful Link</strong></td>
</tr>
<tr>
<td><strong>See Also</strong></td>
</tr>
<tr>
<td><strong>See Also</strong></td>
</tr>
</tbody>
</table>
Autonomous Systems (ASes)

- Routing algorithm within AS
  - Routers in an AS all run the same routing algorithm
  - Routers within an AS know about all the routers inside the AS
  - Intra-AS routing protocol (either LS or DV)
• **Gateway routers**: forward packets outside the AS

• If there is just one gateway router with one link, the forwarding decision is easy
  … it becomes the other AS’s problem

• If multiple gateway routers
  – AS needs to know which destinations are reachable via which AS
  – Configure internal routing tables to route to the appropriate gateway
  – An *Inter-AS routing protocol* figures this out
What if a subnet is accessible via AS1 & AS3?

- AS2 can route to either one
- Send the packet to the gateway router that has the lowest routing cost
- Hot potato routing: pass traffic onto another AS as quickly as possible
Autonomous system types

- **Stub AS**
  - Carries only traffic for which it is a source or a destination
  - Does not route traffic between ASes

- **Multihomed stub AS**
  - Like a stub AS but connected to multiple other ASes
  - Provides fault tolerant connectivity for systems in the AS but does not offer routing from other ASes

- **Transit AS**
  - Provides connections through itself to other networks
Intra-AS Routing: RIP
Routing Information Protocol (RIP)

- Intra-AS protocol = Interior Gateway Protocol (IGP)
- **RIP**: distance-vector routing protocol – used as an IGP

  - Hop count is used as a cost metric (cost of each link = 1)
    - Cost = # hops from the source router to a destination subnet (including the subnet)
    - Minimum cost = 1
    - Maximum cost = 15 (to avoid routing loops)
How RIP works

• Each router maintains a routing table
  – Contains the router’s distance vector & the forwarding table
    • Each subnet identifies the next router & # hops to the destination

• RIP advertisements
  – Each router sends a RIP advertisement to its neighbors approximately every 30 seconds
  – UDP port 520
  – The advertisement contains the router’s routing table
  – If a router does not hear from a neighbor in 180 seconds
    • It assumes the neighbor is dead or disconnected
    • Removes the neighbor from its routing table & propagates info to neighbors

• Upon receiving an advertisement
  – Merge the received table with your own table
    • Choose the smallest # of hops to each destination
    • Add any new destination subnets
RIP Example

Routing table at router D

<table>
<thead>
<tr>
<th>Destination subnet</th>
<th>Next router</th>
<th>Hops to destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>z</td>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>x</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

from p. 385-387 of the text with small mods
RIP Example

Advertisement from A

<table>
<thead>
<tr>
<th>Destination subnet</th>
<th>Next router</th>
<th>Hops to destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>w</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>x</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Routing table at router D

<table>
<thead>
<tr>
<th>Destination subnet</th>
<th>Next router</th>
<th>Hops to destination</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>y</td>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>z</td>
<td>A</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

What do we merge?
- Destination z via A is 5 hops vs. 7
- We know of a destination to w (2 hops via A)
Running RIP

• On UNIX/BSD/Linux
  – RIP runs as a background process called *routed* ("route daemon")
  – Application layer process that can modify routing tables

• On routers
  – RIP runs in the control plane

• Downsides of RIP
  – Converges slowly
  – Does not scale to very large networks
  – Insecure (plain text authentication)

• But it’s still widely used
Intra-AS Routing: OSPF
Open Shortest Path First (OSPF)

• Another interior gateway protocol (intra-AS routing)
  – Designed as a successor to RIP
  – Typically used in large enterprise networks

• RIP is based on the Distance-Vector algorithm

• OSPF is based on Dijkstra’s shortest-path (Link State) algorithm
  – Each router constructs a complete graph of the entire AS
  – Each router runs Dijkstra’s algorithm to determine the shortest path to all subnets with itself as the root node
    • Costs of links are configured by the admin (simplest case: each link = 1)
  – If the link state of a router changes (connectivity or cost)
    • It broadcasts the change to all routers in the AS, not just the neighbors

• OSPF implemented as a special upper-layer protocol
  – Protocol 89 in the IP protocol field
    (TCP=6, UDP=17, ICMP=1)
Features of OSPF

• Security
  – Shared secret key among routers
  – Send MD5 hash(OSPF packet content, shared_key)
  – Receiver validates the hash to ensure that the contents have not been modified
  – Each message includes a sequence number to prevent replay attacks
• Allow multiple paths to be used if they have the same cost
• Support multicast routing
• Allow an AS to be configured into a hierarchy: OSPF Areas
OSPF Areas: “ASes within an AS”

- **OSPF Areas**
  - Subdivision of an OSPF autonomous system
  - Each area
    - Runs its own OSPF link state routing algorithm
    - Has one or more area border routers (ABR) to route outside the area

- **Backbone area:**
  - Contains all area border routers in the AS (and possibly others)
  - Inter-area routing
    - route to an ABR, through the backbone, and to the ABR in the destination area
Inter-AS Routing: BGP
Border Gateway Protocol: BGP

• RIP & OSPF: interior gateway protocols (IGP)
  – intra-AS protocols

• Border Gateway Protocol: exterior gateway protocol (EGP)
  – inter-AS protocol: routes between autonomous systems (AS)
  – BGP version 4 is the standard inter-AS protocol in the Internet
BGP Sessions

• Pairs of routers exchange information via semi-permanent TCP connections
  – One connection for each link between gateway routers
  – Two routers with a BGP connection are BGP peers
BGP Sessions

- Pairs of routers exchange information via semi-permanent TCP connections
  - One connection for each link between gateway routers
    - External BGP (eBGP) session
  - Two routers with a BGP connection are BGP peers
  - Also BGP TCP connections between routers inside an AS
    - Typically between each pair of routers
    - Internal BGP (iBGP) session
Learning destinations

- BGP destinations are CIDR prefixes
  - Range of IP addresses representing one or more subnets

136.16.64.0/24
136.16.65.0/24
136.16.66.0/24
136.16.67.0/24

Gateway router for AS3

Advertise 136.16.64.0/22 to AS2

Route aggregation

Subnets in AS3

Sent via iBGP

Gateway router for AS2

Sent via eBGP
Learning reachable destinations

• What if 136.16.67.0/24 was in AS1?

Advertise 136.16.64.0/22 to AS2

Advertise 136.16.67.0/24 to AS2

Longest-prefix matching for forwarding makes the exception work!

136.16.64.0/24
136.16.65.0/24
136.16.66.0/24

gateway router for AS3

route aggregation

136.16.67.0/24

subnet in AS1

subnet in AS3
BGP reachability propagation via eBGP

- AS1 sends prefix reachability info to AS2
- AS2 sends prefix reachability info to AS1
- AS3 sends prefix reachability info to AS2
- AS2 sends prefix reachability info to AS3
BGP reachability propagation via iBGP

- When a gateway gets prefix reachability info via eBGP
  - It propagates the information to routers inside the AS via iBGP
Readvertising learned routes

• If a gateway router learns of new prefixes
  – It can re-advertise to its peers via eBGP
AS identification and BGP routes

• Each AS has a globally unique AS number (ASN)
  – Assigned by ICANN Regional Internet Registries

• BGP routers send route announcements
  – Destination address block (CIDR network)
  – Path of AS numbers the packet will take
  – BGP attributes

• Key attributes
  – AS-PATH
    • List of ASes through which the advertisement passed
    • If a router sees that its AS is contained in the list ⇒ loop ⇒ reject advertisement
  – NEXT-HOP
    • Identifies the router address outside the AS that sent the advertisement to our AS
      – Intra-AS routing algorithms know routes to internal nodes and attached subnets
      – NEXT-HOP identifies the address on the attached subnet
Use of the NEXT-HOP attribute

1. Router G advertises subnet x on the G-A eBGP session
2. Gateway Router A propagates this route to the intra-AS routers via iBGP
3. Router C needs to add this route to its table
4. NEXT-HOP attribute has the address of G’s IP address for the G-A connection (g’)
5. C creates a forwarding table entry for subnet x to the G-A link
6. It uses the intra-AS routing algorithm to find the next hop on the least-cost path from C to interface g’
Use of NEXT-HOP to resolve links

• Two peering links between AS1 and AS2; AS2 advertises prefix x
• A router in AS1 can get two route advertisements to a prefix x
  – The routes will have the same AS-PATH to x
  – NEXT-HOP will differ based on the eBGP gateway router on AS2
• Intra-AS routing algorithm can determine the cost of a path to each peering link
  – Choose route to h' or route to g’
BGP route selection

• BGP advertises routes through eBGP and iBGP
  – A gateway router may reject a route based on an import policy
  – A router may learn of multiple routes to a prefix

• Elimination criteria (in sequence order)
  – Pick route with the highest local preference value attribute
    • Local preference is a policy defined by an admin
  if multiple routes remaining,
  – Select the route with the shortest AS-PATH
    • BGP would use the distance-vector algorithm if this was the only criteria
  if multiple routes remaining,
  – Choose the route with the closest NEXT-HOP router
Policies are a core part of routing

- A, B, C: transit ASes – IPSs
- W, X, Y: stub ASes – customers

- X is a **multihomed** stub
  - Does not want to route traffic between C & B
  - Even if X knows of a path (e.g., XCY), it will only advertise paths to X

- B knows a path to W: B→A→W
  - Should it tell C?
  - C can route to C→B→A→W: extra burden on B
  - Typically, traffic through an ISP must either originate or terminate at an ISP’s address (customer of the ISP)
  - Peering agreements between ISPs can explicitly allow the route
The end