Routing in the Internet
Hierarchical routing

- Our view of the Internet as a collection of indistinguishable routers is a bit simplistic.

- Scalability.
  - As the number of routers increases, the amount of over computing, storing, and communicating becomes prohibitive.

- Administrative autonomy.
  - Companies would like to run their routers as they please.

Autonomous systems

- These problems are addressed by organizing routers into autonomous systems (AS), each system under the same administrative control.

- Routers within the same AS run the same routing algorithm, called an intra-autonomous system routing protocol, and have information about each other.

- One or more gateway routers are responsible for connecting to the outside world.
Routing the Internet

There are two classes of Internet routing protocol
- **Intra-autonomous system routing** using **RIP** and **OSPF**
- **Inter-autonomous system routing** using **BGP**

RIP (Routing Information Protocol)

- Included in BSD-UNIX distribution in 1982
- Distance vector algorithm
  - distance metric: # hops (max = 15 hops), each link has cost 1
  - DVs exchanged with neighbors every 30 sec in response message (aka advertisement)
  - each advertisement: list of up to 25 destination subnets (in IP addressing sense)

<table>
<thead>
<tr>
<th>subnet</th>
<th>hops</th>
</tr>
</thead>
<tbody>
<tr>
<td>u</td>
<td>1</td>
</tr>
<tr>
<td>v</td>
<td>2</td>
</tr>
<tr>
<td>w</td>
<td>2</td>
</tr>
<tr>
<td>x</td>
<td>3</td>
</tr>
<tr>
<td>y</td>
<td>3</td>
</tr>
<tr>
<td>z</td>
<td>2</td>
</tr>
</tbody>
</table>

from router A to destination subnets:
RIP: example

Routing table in router D

<table>
<thead>
<tr>
<th>destination subnet</th>
<th>next router</th>
<th># hops to dest</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>B</td>
<td>7</td>
</tr>
<tr>
<td>x</td>
<td>--</td>
<td>1</td>
</tr>
</tbody>
</table>

....    ....    ....

RIP: example

A-to-D advertisement

<table>
<thead>
<tr>
<th>dest</th>
<th>next</th>
<th>hops to dest</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>x</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>z</td>
<td>C</td>
<td>4</td>
</tr>
</tbody>
</table>

....

Routing table in router D

<table>
<thead>
<tr>
<th>destination subnet</th>
<th>next router</th>
<th># hops to dest</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>B</td>
<td>5</td>
</tr>
<tr>
<td>x</td>
<td>--</td>
<td>1</td>
</tr>
</tbody>
</table>

....    ....    ....
RIP: link failure, recovery

• If no advertisement heard after 180 sec --> neighbor/link declared dead
  – routes via neighbor invalidated
  – new advertisements sent to neighbors
  – neighbors in turn send out new advertisements (if tables changed)
  – link failure info quickly (?) propagates to entire net
  – poison reverse used to prevent ping-pong loops (infinite distance = 16 hops)

RIP table processing

• RIP routing tables managed by application-level process called route-d (daemon)
• advertisements sent in UDP packets, periodically repeated
### OSPF (Open Shortest Path First)

- “open”: publicly available
- Uses link state algorithm
  - LS packet dissemination
  - topology map at each node
  - route computation using Dijkstra’s algorithm
- OSPF advertisement carries one entry per neighbor
- Advertisements flooded to entire AS
  - carried in OSPF messages directly over IP (rather than TCP or UDP)
- IS-IS routing protocol: nearly identical to OSPF

### OSPF “advanced” features (not in RIP)

- Security
  - all OSPF messages authenticated (to prevent malicious intrusion)
- Multiple same-cost paths allowed
  - only one path in RIP
- For each link, multiple cost metrics for different TOS (e.g., satellite link cost set “low” for best effort ToS; high for real time ToS)
- Integrated uni- and multicast support:
  - **Multicast OSPF (MOSPF)** uses same topology data base as OSPF
- Hierarchical OSPF in large domains.
Hierarchical OSPF

- Two-level hierarchy: local area, backbone.
  - link-state advertisements only in area
  - each node has detailed area topology; only know direction (shortest path) to nets in other areas.

- Area border routers
  - “summarize” distances to nets in own area, advertise to other Area Border routers.

- Backbone routers
  - run OSPF routing limited to backbone.

- Boundary routers
  - connect to other AS’s.
BGP routing

Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): the de facto inter-domain routing protocol
  - “glue that holds the Internet together”

- BGP provides each AS a means to:
  - eBGP: obtain subnet reachability information from neighboring ASs.
  - iBGP: propagate reachability information to all AS-internal routers.
  - determine “good” routes to other networks based on reachability information and policy.

- Allows subnet to advertise its existence to rest of Internet: “I am here”
BGP inter-domain routing

• Pairs of routers, called BGP peers, exchange information over semi-permanent TCP connections using port 179.

• External BGP (eBGP) sessions span two ASs; internal BGP (iBGP) sessions between peers in a single AS.

Route advertisements

• BGP destinations are not hosts, but CIDRized prefixes, with each prefix representing a subnet or collection of subnets.

*AS2 could aggregate these prefixes and use BGP to advertise the single prefix to 138.16.64/22 to AS1.
Longest-prefix matching

Fourth subnet moves:
138.16.67/24;

Attached subnets:
138.16.64/24;
138.16.65/24;
138.16.66/24;

*AS2 could still aggregate the three prefixes and use BGP to advertise the single prefix to 138.16.64/22. Why?

Distributing prefix reachability information

Using eBGP session between 3a and 1c, AS3 sends AS1 the list of its reachable prefixes. AS1 does likewise.

*We think of each prefix advertisement as a promise to forward any datagram destined for the prefix along a path toward the prefix.
Distributing prefix reachability information

Each gateway router uses its iBGP sessions to distribute the prefixes to the other routers in the AS.

Thus, gateway router 1b learns about AS3 from 1c via an iBGP connection and can share this with AS2 through its eBGP connection.

*Whenever a router (gateway or otherwise) learns about a new prefix, it creates an entry in its forwarding table.*

Autonomous system numbers

Stub AS carry only traffic for which it is a source or destination.

AS are assigned globally unique autonomous system numbers (ASNs) by ICANN.
AS2 advertises to AS1

Router advertisements across BGP sessions include a destination prefix together with a set of BGP attributes associated with a path to that destination.

When 1b receives an advert, it uses its import policy to decide whether to accept or filter the route and whether to set certain attributes such as the router preference metrics.

AS1 may choose to advertise to AS3 (or not)

AS1 advertises to AS3:
Destination: 138.16.64/24
AS-PATH: AS1; AS2
NEXT-HOP: IP address of 1c’s interface to AS3.

AS2 advertises to AS1:
Destination: 138.16.64/24
AS-PATH: AS2
NEXT-HOP: IP address of 2a’s interface to AS1.

*Why list all AS’s in path?
NEXT-HOP in action

AS3 advertises to AS1:
Destination: 192.24.32/24
AS-PATH: AS3
NEXT-HOP: IP address of 3a’s interface to AS1.

1. Router 1d learns about route to 192.24.32/24 through gateway 1c from iBGP.

2. Intra-AS routing algorithm has determined the least-cost path to all subnets attached to the routers in AS1.

3. Supposing 1b is first-hop of shortest path from 1d to subnet 3a-1c in AS1, then 1d enters (192.24.32/24, 1b) in its forwarding table.

Another use for NEXT-HOP

• NEXT-HOP attributes are also used to determine which peering link to use.

• For example, suppose a router in AS1 learns about two different routes to the same prefix x with the same AS-PATH.

• Using NEXT-HOP values and intra-AS routing algorithm, a router can determine the cost to each peering link, then apply hot-potato routing to find the interface.
If two or more routes get you there ...

- Use highest local priority number.*
- From remaining routes, use shortest AS-PATH.
- From remaining routes, use closest NEXT-HOP router (hot-potato).
- From remaining routes, use BGP identifiers.

*Assigned by the router, or learned from another router or set by the network administrator.

Routing policy

- A, B, and C are backbone provider networks, while w, x, and y are stub networks. All traffic entering a stub network must be destined for that network, and all traffic leaving a stub network originated there.
- X is a multi-homed stub network, but how is this enforced?

Simple, X tells its neighbors B and C that it has no paths to any destination except itself.
Economic and other consideration

- Suppose B has learned from A that A has a path AW to W. B would love to advertise the path BAW to its X customers.
- However, should B advertise the path BAW to C? If A, B, and C are all backbone providers, then B might feel aggrieved.

There are not rules, but

- The rule of thumb followed by commercial ISPs is that any traffic flowing across an ISP’s backbone must have either a source or a destination (or both) in a network that is a customer of that ISP.
- Otherwise the traffic would be getting a “free ride”.
### Why different Intra-, Inter-AS routing?

<table>
<thead>
<tr>
<th></th>
<th>Intra-AS</th>
<th>Inter-AS</th>
</tr>
</thead>
<tbody>
<tr>
<td>policy</td>
<td>single admin, so no policy decisions needed</td>
<td>admin wants control over how its traffic routed, who routes through its net</td>
</tr>
<tr>
<td>scale</td>
<td>hierarchical routing saves table size, reduced update traffic</td>
<td></td>
</tr>
<tr>
<td>performance</td>
<td>can focus on performance</td>
<td>policy may dominate over performance</td>
</tr>
</tbody>
</table>