Routing in the Internet

- The Global Internet consists of Autonomous Systems (AS) interconnected with each other:
  - **Stub AS**: small corporation
  - **Multihomed AS**: large corporation (no transit)
  - **Transit AS**: provider

- Two-level routing:
  - **Intra-AS**: administrator is responsible for choice
  - **Inter-AS**: unique standard

Internet AS Hierarchy

- Intra-AS border (exterior gateway) routers
- Inter-AS interior (gateway) routers

Intra-AS Routing

- Also known as **Interior Gateway Protocols (IGP)**
- Most common IGPs:
  - **RIP**: Routing Information Protocol
  - **OSPF**: Open Shortest Path First
  - **IGRP**: Interior Gateway Routing Protocol (Cisco propr.)

RIP (Routing Information Protocol)

- Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops)
  - *why?*
- Distance vectors: exchanged every 30 sec via Response Message (also called advertisement)
- Each advertisement: route to up to 25 destination nets
RIP: Link Failure and 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 routed (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 router
- Advertisements disseminated to entire AS (via flooding)

OSPF “advanced” features (not in RIP)

- Security: all OSPF messages authenticated (to prevent malicious intrusion); TCP connections used
- Multiple same-cost paths allowed (only one path in RIP)
- For each link, multiple cost metrics for different TOS (eg, satellite link cost set “low” for best effort; high for real time)
- 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 ASs.

**IGRP (Interior Gateway Routing Protocol)**

- **Cisco** proprietary; successor of RIP (mid 80s)
- Distance Vector, like RIP
- Several cost metrics (delay, bandwidth, reliability, load etc)
- Uses TCP to exchange routing updates
- Loop-free routing via Distributed Updating Alg. (DUAL) based on diffused computation

**Inter-AS routing**
Internet inter-AS routing: BGP

- **BGP** (Border Gateway Protocol): *the de facto* standard
- **Path Vector** protocol:
  - similar to Distance Vector protocol
  - each Border Gateway broadcast to neighbors (peers) *entire path* (i.e., sequence of ASs) to destination
  - E.g., Gateway X may send its path to dest. Z:
    
    \[
    \text{Path (X,Z)} = X,Y_1,Y_2,Y_3,...,Z
    \]

Suppose: gateway X send its path to peer gateway W

- W may or may not select path offered by X
  - cost, policy (don’t route via competitors AS), loop prevention reasons.
- If W selects path advertised by X, then:
  
  \[
  \text{Path (W,Z)} = w, \text{Path (X,Z)}
  \]

Note: X can control incoming traffic by controlling its route advertisements to peers:

- e.g., don’t want to route traffic to Z -> don’t advertise any routes to Z

BGP messages exchanged using TCP.

- **BGP** messages:
  - OPEN: opens TCP connection to peer and authenticates sender
  - UPDATE: advertises new path (or withdraws old)
  - KEEPALIVE keeps connection alive in absence of UPDATES; also ACKs OPEN request
  - NOTIFICATION: reports errors in previous msg; also used to close connection

Why different Intra- and Inter-AS routing?

- **Policy:**
  - Inter-AS: admin wants control over how its traffic routed, who routes through its net.
  - Intra-AS: single admin, so no policy decisions needed
- **Scale:**
  - hierarchical routing saves table size, reduced update traffic
- **Performance:**
  - Intra-AS: can focus on performance
  - Inter-AS: policy may dominate over performance
**IP datagram format**

- IP protocol version number
- Header length (bytes)
- "Type" of data
- Max number of remaining hops (decremented at each router)
- Upper layer protocol to deliver payload to

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td>4 bits, version number</td>
</tr>
<tr>
<td>Header length</td>
<td>4 bits, length of header</td>
</tr>
<tr>
<td>Type of service</td>
<td>8 bits, service type</td>
</tr>
<tr>
<td>Identifier</td>
<td>16 bits, identifier for fragmentation/reassembly</td>
</tr>
<tr>
<td>Flags</td>
<td>1 bit, fragment flag</td>
</tr>
<tr>
<td>Offset</td>
<td>13 bits, offset</td>
</tr>
<tr>
<td>Time to live</td>
<td>16 bits, time to live</td>
</tr>
<tr>
<td>Protocol</td>
<td>8 bits, upper layer protocol</td>
</tr>
<tr>
<td>Source IP addr</td>
<td>32 bits, source IP address</td>
</tr>
<tr>
<td>Destination IP addr</td>
<td>32 bits, destination IP address</td>
</tr>
<tr>
<td>Options</td>
<td>(if any)</td>
</tr>
<tr>
<td>Data</td>
<td>(variable length, typically a TCP or UDP segment)</td>
</tr>
</tbody>
</table>

**IP Fragmentation & Reassembly**

- Network links have MTU (max transfer size) - largest possible link-level frame.
- Different link types, different MTUs
- Large IP datagram divided ("fragmented") within net
  - One datagram becomes several datagrams
  - "reassembled" only at final destination
  - IP header bits used to identify, order related fragments

**IP Fragmentation and Reassembly**

- One large datagram becomes several smaller datagrams

**ICMP: Internet Control Message Protocol**

- Used by hosts, routers, gateways to communicate network-level information
- Error reporting:
  - Unreachable host, network, port, protocol
  - Echo request/reply (used by ping)
- Network-layer "above" IP:
  - ICMP msgs carried in IP datagrams
  - ICMP messages: type, code plus first 8 bytes of IP datagram causing error

<table>
<thead>
<tr>
<th>Type</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Echo reply (ping)</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>Dest. network unreachable</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Dest. host unreachable</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Dest. protocol unreachable</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Dest. port unreachable</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>Dest. network unknown</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>Dest. host unknown</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>Source quench (congestion control - not used)</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>Echo request (ping)</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>Route advertisement</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>Router discovery</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>TTL expired</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>Bad IP header</td>
</tr>
</tbody>
</table>