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

IP datagram format

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP protocol version number</td>
<td>4 bits, identifies the IP protocol (IPv4)</td>
</tr>
<tr>
<td>Header length (bytes)</td>
<td>1 byte, indicates the length of the header in 32-bit units</td>
</tr>
<tr>
<td>&quot;type&quot; of data</td>
<td>16 bits, identifies the upper layer protocol (e.g., TCP or UDP)</td>
</tr>
<tr>
<td>Max number of remaining hops</td>
<td>16 bits, decremented at each router, indicates the remaining number of hops</td>
</tr>
<tr>
<td>Source IP address</td>
<td>32 bits, identifies the source host IP address</td>
</tr>
<tr>
<td>Destination IP address</td>
<td>32 bits, identifies the destination host IP address</td>
</tr>
<tr>
<td>Options (if any)</td>
<td>1 or more options, e.g., time stamp, record route</td>
</tr>
<tr>
<td>Data</td>
<td>(variable length, typically a TCP or UDP segment)</td>
</tr>
<tr>
<td>Total datagram length (bytes)</td>
<td>Indicates the total length of the datagram (in bytes)</td>
</tr>
<tr>
<td>Fragment offset</td>
<td>Indicates the offset of the fragment if fragmentation is used</td>
</tr>
<tr>
<td>Fragment length</td>
<td>Indicates the length of the fragment if fragmentation is used</td>
</tr>
<tr>
<td>Internet checksum</td>
<td>16 bits, a checksum of the datagram to detect errors</td>
</tr>
<tr>
<td>Time to live</td>
<td>16 bits, decremented at each router, indicates the remaining time to live</td>
</tr>
<tr>
<td>Upper layer</td>
<td>1 byte, identifies the upper layer protocol (e.g., TCP or UDP)</td>
</tr>
<tr>
<td>Length</td>
<td>4 bits, indicates the length of the datagram in 32-bit units</td>
</tr>
<tr>
<td>Type of service</td>
<td>1 byte, identifies the type of service (e.g., time stamp)</td>
</tr>
<tr>
<td>Flags</td>
<td>1 or more flags, e.g., fragment, offset</td>
</tr>
<tr>
<td>Offset</td>
<td>Indicates the offset of the fragment if fragmentation is used</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

- \[ \text{length} = 4000 \]
- \[ \text{ID} = x \]
- \[ \text{fragflag} = 0 \]
- \[ \text{offset} = 0 \]

- \[ \text{length} = 1500 \]
- \[ \text{ID} = x \]
- \[ \text{fragflag} = 1 \]
- \[ \text{offset} = 0 \]

- \[ \text{length} = 1500 \]
- \[ \text{ID} = x \]
- \[ \text{fragflag} = 1 \]
- \[ \text{offset} = 1480 \]

- \[ \text{length} = 1040 \]
- \[ \text{ID} = x \]
- \[ \text{fragflag} = 0 \]
- \[ \text{offset} = 2960 \]
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 message: 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>

Internet AS Hierarchy
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 type scheme
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops)
  - Can you guess why?

- Distance vector: exchanged every 30 sec via a Response Message (also called Advertisement)
- Each Advertisement contains up to 25 destination nets
RIP (Routing Information Protocol)

<table>
<thead>
<tr>
<th>Destination Network</th>
<th>Next Router</th>
<th>Num. of hops to dest.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>B</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>--</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

RIP: Link Failure and Recovery

- If no advertisement heard after 180 sec, neighbor/link dead
- Routes via the neighbor are invalidated; new advertisements sent to neighbors
- Neighbors in turn send out new advertisements if their 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 an application process called route-d (daemon)
- Advertisements encapsulated in UDP packets (no reliable delivery required; advertisements are periodically repeated)
RIP Table example (continued)

RIP Table example
(at router giroflee.eurocom.fr):

- Three attached class C networks (LANs)
- Router only knows routes to attached LANs
- Default router used to “go up”
- Route multicast address: 224.0.0.0
- Loopback interface (for debugging)

<table>
<thead>
<tr>
<th>Destination</th>
<th>Gateway</th>
<th>Flags</th>
<th>Ref</th>
<th>Use</th>
<th>Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>127.0.0.1</td>
<td>127.0.0.1</td>
<td>UH</td>
<td>0</td>
<td>26492</td>
<td>lo0</td>
</tr>
<tr>
<td>192.168.2.6</td>
<td>192.168.2.5</td>
<td>U</td>
<td>2</td>
<td>13</td>
<td>fa0</td>
</tr>
<tr>
<td>193.55.114.6</td>
<td>193.55.114.6</td>
<td>U</td>
<td>3</td>
<td>58503</td>
<td>le0</td>
</tr>
<tr>
<td>192.168.3.5</td>
<td>192.168.3.5</td>
<td>U</td>
<td>2</td>
<td>25</td>
<td>qaa0</td>
</tr>
<tr>
<td>224.0.0.0</td>
<td>193.55.114.6</td>
<td>U</td>
<td>3</td>
<td>0</td>
<td>le0</td>
</tr>
<tr>
<td>default</td>
<td>193.55.114.129</td>
<td>UG</td>
<td>0</td>
<td>143454</td>
<td></td>
</tr>
</tbody>
</table>
OSPF (Open Shortest Path First)

- “open”: publicly available
- Uses the Link State algorithm
  - LS packet dissemination
  - Topology map at each node
  - Route computation using Dijkstra’s alg
- OSPF advertisement carries one entry per neighbor router
- Advertisements disseminated to entire Autonomous System (via flooding)

OSPF “advanced” features (not in RIP)

- Security: all OSPF messages are 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 and backbone.
- Link-state advertisements do not leave respective areas.
- Nodes in each area have detailed area topology; they only know direction (shortest path) to networks in other areas.
- Area Border routers "summarize" distances to networks in the area and advertise them to other Area Border routers.
- Backbone routers run an OSPF routing alg limited to the 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
- routing tables exchanged only when costs change
- Loop-free routing achieved by using a Distributed Updating Alg. (DUAL) based on diffused computation
- In DUAL, after a distance increase, the routing table is frozen until all affected nodes have learned of the change.

Inter-AS routing
Inter-AS routing (cont)

- **BGP (Border Gateway Protocol):** the de facto standard
- **Path Vector** protocol: and extension of Distance Vector
- Each Border Gateway broadcast to neighbors (peers) the entire path (ie, sequence of ASs) to destination
- For example, Gateway X may store the following path to destination Z:

  \[ \text{Path} (X,Z) = X,Y_1,Y_2,Y_3,\ldots,Z \]

Inter-AS routing (cont)

- Now, suppose Gwy X send its path to peer Gwy W
- Gwy W may or may not select the path offered by Gwy X, because of cost, policy ($$$$) or loop prevention reasons.
- If Gwy W selects the path advertised by Gwy X, then:

  \[ \text{Path} (W,Z) = w, \text{Path} (X,Z) \]

  Note: path selection based not so much on cost (eg, # of AS hops), but mostly on administrative and policy issues (e.g., do not route packets through competitor's AS)
Inter-AS routing (cont)

- Peers exchange BGP messages using TCP.
- OPEN msg opens TCP connection to peer and authenticates sender.
- UPDATE msg advertises new path (or withdraws old).
- KEEPALIVE msg keeps connection alive in absence of UPDATES; it also serves as ACK to an OPEN request.
- NOTIFICATION msg reports errors in previous msg; also used to close a connection.

Why different Intra- and Inter-AS routing?

- **Policy**: Inter is concerned with policies (which provider we must select/avoid, etc). Intra is contained in a single organization, so, no policy decisions necessary.
- **Scale**: Inter provides an extra level of routing table size and routing update traffic reduction above the Intra layer.
- **Performance**: Intra is focused on performance metrics; needs to keep costs low. In Inter it is difficult to propagate performance metrics efficiently (latency, privacy etc). Besides, policy related information is more meaningful.

We need BOTH!