Topics for This Week

• Routing Protocols in the Internet
  – OSPF, BGP

• More on IP
  – Fragmentation and Reassembly
  – ICMP

• Readings
  – Sections 5.6.4-5.6.5
Hierarchical Routing

• aggregate routers into regions, “autonomous systems” (AS)
• routers in same AS run same routing protocol
  – “intra-AS” routing protocol
  – routers in different AS can run different intra-AS routing protocol

**gateway routers**

• special routers in AS
• run intra-AS routing protocol with all other routers in AS
• *also* responsible for routing to destinations outside AS
  – run *inter-AS routing* protocol with other gateway routers
Intra-AS and Inter-AS routing

Gateways:
- perform inter-AS routing amongst themselves
- perform intra-AS routing with other routers in their AS

Inter-AS, intra-AS routing in gateway A.c

Routing Table

- network layer
- link layer
- physical layer
Intra-AS and Inter-AS routing

Inter-AS routing between A and B

Intra-AS routing within AS A

Intra-AS and Inter-AS routing

Host h1

Host h2

A

B

A.a

A.c

B.a

C.b

C

Intra-AS routing within AS B
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
Intra-AS Routing (stop)

• 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 proprietary)
RIP (Routing Information Protocol)

- Distance vector algorithm
- Included in BSD-UNIX Distribution in 1982
- Distance metric: # of hops (max = 15 hops)
  - Can you guess 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
  - Gated
- 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)
Neighbor Discovery and Maintenance

- OSPF Hello protocol
- Sends Hello packets on all its interfaces
  - Every Hello Interval (default 10 sec)
- Helps detect the failure of neighbors
- Neighbor is designated as failed
  - If no Hello for Dead Interval (40 sec)
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 (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.
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:

\[
\text{Path } (X,Z) = X,Y_1,Y_2,Y_3,\ldots,Z
\]
Internet Inter-AS routing: BGP

**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 \(\rightarrow\) don’t advertise any routes to Z
Internet Inter-AS routing: BGP

- 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
The Internet Network layer

- Routing protocols
  - path selection
  - RIP, OSPF, BGP

- IP protocol
  - addressing conventions
  - datagram format
  - packet handling conventions

- ICMP protocol
  - error reporting
  - router "signaling"

- Transport layer: TCP, UDP

- Network layer

- Link layer

- physical layer
More on IP
# IP Packet Format

<table>
<thead>
<tr>
<th>Field</th>
<th>Bit Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version (VERS)</td>
<td>0</td>
</tr>
<tr>
<td>Header Length (H. LEN)</td>
<td>4</td>
</tr>
<tr>
<td>Service Type</td>
<td>8</td>
</tr>
<tr>
<td>Protocol Type</td>
<td>16</td>
</tr>
<tr>
<td>Total Length</td>
<td>19</td>
</tr>
<tr>
<td>Identification</td>
<td>24</td>
</tr>
<tr>
<td>Flags</td>
<td>31</td>
</tr>
<tr>
<td>Fragment Offset</td>
<td>0</td>
</tr>
<tr>
<td>Time to Live</td>
<td>4</td>
</tr>
<tr>
<td>Type</td>
<td>8</td>
</tr>
<tr>
<td>Header Checksum</td>
<td>16</td>
</tr>
<tr>
<td>Source IP Address</td>
<td>19</td>
</tr>
<tr>
<td>Destination IP Address</td>
<td>24</td>
</tr>
<tr>
<td>Option (if any)</td>
<td>31</td>
</tr>
<tr>
<td>Padding</td>
<td>0</td>
</tr>
<tr>
<td>Beginning of Data</td>
<td>4</td>
</tr>
<tr>
<td>Data</td>
<td>8</td>
</tr>
</tbody>
</table>
Fields in IP Packet

• IP protocol version
  – Current version is 4

• Header length
  – Number of 32-bit quantities in the header

• Type of Service
  – 3-bit Priority
  – Delay, Throughput, Reliability bits

• Total length
  – Including header (maximum 65535 bytes)
Fields in IP Packet

• Identification
  – All fragments of a packet have same identification
• Flags
  – Don’t Fragment, More Fragments
• Fragment offset
  – Where in the original packet (count in 8 byte units)
• Time to live
  – Life time of packet
• Protocol Type
  – TCP, UDP etc
IP Fragmentation & Reassembly

• Each subnet has its own MTU size
  – Maximum Transmission Unit

• An IP packet is chopped into smaller pieces if
  – Packet size is greater than network MTU
  – Don’t fragment option is not set

• Each datagram has unique identification
  – All fragments carry original datagram id

• All fragments except the last have more flag set
IP Fragmentation & Reassembly

- Datagram assembly done only at destination
  - Why not at a router?
- Use datagram id to put pieces together
  - The last piece indicated with more bit 0
  - Offset plus the length tell whether any
    - Holes missing in the middle
- Setup a reassembly timer after first fragment
  - If all pieces in time, pass the pkt to upper layer
  - If some do not arrive in time, discard the fragments
- No recovery from lost fragments (why?)
IP Fragmentation & Reassembly

• large IP datagram 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

| length =4000 | ID =x | moreflag =0 | offset =0 |
| length =1500 | ID =x | moreflag =1 | offset =0 |
| length =1500 | ID =x | moreflag =1 | offset =1480 |
| length =1040 | ID =x | moreflag =0 | offset =2960 |
Fragmentation Example
Internet Control Message Protocol

• An error reporting mechanism
  – Time exceeded
    • Packet discarded because TTL was 0
  – Destination unreachable
    • Router cannot locate destination
  – Source quench
    • Buffer overflow, request source to reduce rate
  – Redirect
    • Suggest a better router
ICMP Message Transport

- ICMP messages carried in IP datagrams
- Treated like any other datagram
  - But no error message sent if
    - ICMP message causes error
- Message sent to the source
  - 8 bytes of the original header included
ICMP Usage

• Testing reachability
  – ICMP echo request/reply
  – ping

• Tracing route to a destination
  – Time-to-live field
  – Traceroute

• Path MTU discovery
  – Don’t fragment bit
Protocol Configuration

• Items to be configured
  – IP address
  – Default router address
  – Subnet mask
  – DNS server address

• Reverse Address Resolution Protocol (RARP)
• ICMP address mask request and router discovery
• Bootstrap protocol (BOOTP)
IP addresses: how to get one?

- Hard-coded by system admin in a file
- **Dynamic Host Configuration Protocol**
  - dynamically get address: “plug-and-play”
  - host broadcasts “**DHCP discover**” msg
  - DHCP server responds with “**DHCP offer**” msg
  - host requests IP address: “**DHCP request**” msg
  - DHCP server sends address: “**DHCP ack**” msg
Network Layer Summary

• Network service
  – datagram vs virtual circuit

• Routing protocols
  – Link state and distance vector
  – RIP, OSPF, PGP

• Case studies
  – ATM, IPv4
Hierarchical OSPF
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.
Traceroute yahoo.com

1  rsfc-v15.eswitch.umn.edu (128.101.35.253)
2  192.168.99.30 (192.168.99.30)
3  tc3x.router.umn.edu (160.94.26.70)
4  tc2x.router.umn.edu (160.94.26.98)
5  otr-tc2.northernlights.gigapop.net (192.42.152.134)
6  otr-onvoy.northernlights.gigapop.net (192.42.152.14)
7  core1-ge1-1-0.msc.mr.net (137.192.3.254)
Traceroute yahoo.com

8  p4-0.chcgil1-cr1.bbnplanet.net (4.24.149.97)
9  p5-0.chcgil1-br1.bbnplanet.net (4.24.5.241)
10 so-3-0-0.chcgil2-br1.bbnplanet.net (4.24.9.69)
11 p1-0.chcgil2-cr1.bbnplanet.net (4.24.7.134)
12 pos1-2.core1.Chicago1.Level3.net
   (209.0.225.33)
13 so-4-0-0.mp2.Chicago1.Level3.net
   (209.247.10.169)
Traceroute yahoo.com

14  so-3-0-0.mp1.SanJose1.Level3.net
   (64.159.1.129)
15  gigabitethernet9-2.ipcolo4.SanJose1.Level3.net
   (64.159.2.138)
16  ***
17  ge-3-3-0.msr1.pao.yahoo.com (216.115.101.42)
18  vlan29.bas2-m.snv.yahoo.com
   (216.115.100.126)
19  w9.snv.yahoo.com (216.115.102.81)