The Internet Network layer

Host, router network layer functions:

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

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

- **ICMP protocol**
  - error reporting
  - router "signaling"

<table>
<thead>
<tr>
<th>Transport layer: TCP, UDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routing protocols</td>
</tr>
<tr>
<td><em>path selection</em></td>
</tr>
<tr>
<td>RIP, OSPF, BGP</td>
</tr>
<tr>
<td>IP protocol</td>
</tr>
<tr>
<td><em>addressing conventions</em></td>
</tr>
<tr>
<td><em>datagram format</em></td>
</tr>
<tr>
<td><em>packet handling conventions</em></td>
</tr>
<tr>
<td>ICMP protocol</td>
</tr>
<tr>
<td><em>error reporting</em></td>
</tr>
<tr>
<td><em>router &quot;signaling&quot;</em></td>
</tr>
</tbody>
</table>

**IP Addressing: introduction**

- **IP address**: 32-bit identifier for host, router interface
- **interface**: connection between host/router and physical link

- router's typically have multiple interfaces
- host may have multiple interfaces
- IP addresses associated with each interface

| 223.1.1.1 = 11011111 00000001 00000001 00000001 |
| 223 1 1 1 1 |

Network Layer 4-2
IP Addressing

- IP address:
  - network part (high order bits)
  - host part (low order bits)
- What's a network?
  (from IP address perspective)
  - device interfaces with same network part of IP address
  - can physically reach each other without intervening router

network consisting of 3 IP networks
(for IP addresses starting with 223, first 24 bits are network address)

How to find the networks?

- Detach each interface from router, host
- create "islands of isolated networks"

Interconnected system consisting of six networks
## IP Addresses

given notion of "network", let's re-examine IP addresses:

"class-full" addressing:

<table>
<thead>
<tr>
<th>Class</th>
<th>Network</th>
<th>Host</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>host</td>
<td>1.0.0.0 to 127.255.255.255</td>
</tr>
<tr>
<td>B</td>
<td>10</td>
<td>network</td>
<td>128.0.0.0 to 191.255.255.255</td>
</tr>
<tr>
<td>C</td>
<td>110</td>
<td>network</td>
<td>192.0.0.0 to 223.255.255.255</td>
</tr>
<tr>
<td>D</td>
<td>1110</td>
<td>multicast address</td>
<td>224.0.0.0 to 239.255.255.255</td>
</tr>
</tbody>
</table>

32 bits

## IP addressing: CIDR

### Classful addressing:
- inefficient use of address space, address space exhaustion
- e.g., class B net allocated enough addresses for 65K hosts, even if only 2K hosts in that network

### CIDR: Classless InterDomain Routing
- network portion of address of arbitrary length
- address format: `a.b.c.d/x`, where `x` is # bits in network portion of address

```
  11001000  00010111  00010000  00000000
   network part                     host part
```

200.23.16.0/23
**IP addresses: how to get one?**

**Q:** How does host get IP address?

- hard-coded by system admin in a file
- Wintel: control-panel->network->configuration->tcp/ip->properties
- UNIX: /etc/rc.config
- DHCP: Dynamic Host Configuration Protocol: dynamically get address from as server
  - "plug-and-play"
  (more shortly)

**IP addresses: how to get one?**

**Q:** How does network get network part of IP addr?

**A:** gets allocated portion of its provider ISP's address space

<table>
<thead>
<tr>
<th>ISP's block</th>
<th>11001000_00010111_00010000_00000000</th>
<th>200.23.16.0/20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organization 0</td>
<td>11001000_00010111_00010000_00000000</td>
<td>200.23.16.0/23</td>
</tr>
<tr>
<td>Organization 1</td>
<td>11001000_00010111_00010010_00000000</td>
<td>200.23.18.0/23</td>
</tr>
<tr>
<td>Organization 2</td>
<td>11001000_00010111_00010100_00000000</td>
<td>200.23.20.0/23</td>
</tr>
<tr>
<td>Organization 7</td>
<td>11001000_00010111_00011110_00000000</td>
<td>200.23.30.0/23</td>
</tr>
</tbody>
</table>
Hierarchical addressing: route aggregation

Hierarchical addressing allows efficient advertisement of routing information:

- **Organization 0**: 200.23.16.0/23
- **Organization 1**: 200.23.18.0/23
- **Organization 2**: 200.23.20.0/23
- **Organization 7**: 200.23.30.0/23

"Send me anything with addresses beginning 200.23.16.0/20"

Fly-By-Night-ISP

"Send me anything with addresses beginning 199.31.0.0/16"

ISPs-R-Us

Internet

Hierarchical addressing: more specific routes

ISPs-R-Us has a more specific route to Organization 1

- **Organization 0**: 200.23.16.0/23
- **Organization 2**: 200.23.20.0/23
- **Organization 7**: 200.23.30.0/23
- **Organization 1**: 200.23.18.0/23

"Send me anything with addresses beginning 200.23.16.0/20"

Fly-By-Night-ISP

"Send me anything with addresses beginning 199.31.0.0/16 or 200.23.18.0/23"

ISPs-R-Us

Internet
IP addressing: the last word...

Q: How does an ISP get block of addresses?
A: ICANN: Internet Corporation for Assigned Names and Numbers
  - allocates addresses
  - manages DNS
  - assigns domain names, resolves disputes

Getting a datagram from source to dest.

IP datagram:

<table>
<thead>
<tr>
<th>misc fields</th>
<th>source IP addr</th>
<th>dest IP addr</th>
<th>data</th>
</tr>
</thead>
</table>

- datagram remains unchanged, as it travels source to destination
- addr fields of interest here

forwarding table in A

<table>
<thead>
<tr>
<th>Dest. Net.</th>
<th>next router</th>
<th>Nhops</th>
</tr>
</thead>
<tbody>
<tr>
<td>223.1.1</td>
<td>223.1.4</td>
<td>1</td>
</tr>
<tr>
<td>223.1.2</td>
<td>223.1.3</td>
<td>2</td>
</tr>
<tr>
<td>223.1.3</td>
<td>223.1.4</td>
<td>2</td>
</tr>
</tbody>
</table>
Getting a datagram from source to dest.

Starting at A, send IP datagram addressed to B:
- look up net. address of B in forwarding table
- find B is on same net. as A
- link layer will send datagram directly to B inside link-layer frame
  - B and A are directly connected

Getting a datagram from source to dest.

Starting at A, dest. E:
- look up network address of E in forwarding table
- E on different network
  - A, E not directly attached
- routing table: next hop router to E is 223.1.1.4
- link layer sends datagram to router 223.1.1.4 inside link-layer frame
- datagram arrives at 223.1.1.4
- continued.....
**Getting a datagram from source to dest.**

### Forwarding Table in Router

<table>
<thead>
<tr>
<th>Dest. Net</th>
<th>router</th>
<th>Nhops</th>
<th>interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>223.1.1</td>
<td>-</td>
<td>1</td>
<td>223.1.4</td>
</tr>
<tr>
<td>223.1.2</td>
<td>-</td>
<td>1</td>
<td>223.1.2.9</td>
</tr>
<tr>
<td>223.1.3</td>
<td>-</td>
<td>1</td>
<td>223.1.3.27</td>
</tr>
</tbody>
</table>

#### Arriving at 223.1.4, destined for 223.1.2.2

- Look up network address of E in router's forwarding table.
- E on same network as router's interface 223.1.2.9.
- E router, E directly attached.
- Link layer sends datagram to 223.1.2.2 inside link-layer frame via interface 223.1.2.9.
- Datagram arrives at 223.1.2.2!!! (hooray!)

### IP Datagram Format

- **IP Protocol Version Number**
- **Header Length** (bytes)
- "**Type**" of data
- **Max Number Remaining Hops** (decremented at each router)
- **Upper Layer Protocol to Deliver Payload To**

**Header Components:**

- **32-bit Source IP Address**
- **32-bit Destination IP Address**
- **16-bit Identifier**
- **Flags**
- **Fragment Offset**
- **Time to Live**
- **Internet Checksum**
- **Total Datagram Length (bytes)**
- **Options (if any)**

**Data:**

- (variable length, typically a TCP or UDP segment)

**Example Overhead:**

- 20 bytes of TCP
- 20 bytes of IP
- 40 bytes + app layer overhead

---

*Network Layer 4-15*
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

Example
- 4000 byte datagram
- MTU = 1500 bytes

One large datagram becomes several smaller datagrams
ICMP: Internet Control Message Protocol

- used by hosts, routers, gateways to communication 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>

DHCP: Dynamic Host Configuration Protocol

**Goal:** allow host to dynamically obtain its IP address from network server when it joins network

- Can renew its lease on address in use
- Allows reuse of addresses (only hold address while connected an "on"
- Support for mobile users who want to join network (more shortly)

DHCP overview:

- 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
DHCP client-server scenario
**NAT: Network Address Translation**

- **rest of Internet**
- **local network (e.g., home network)**
  - 10.0.0/24
  - 10.0.0.1
  - 10.0.0.2
  - 10.0.0.3
  - 10.0.0.4
  - 138.76.29.7

- **All datagrams leaving local network have same single source NAT IP address: 138.76.29.7, different source port numbers**
- **Datagrams with source or destination in this network have 10.0.0/24 address for source, destination (as usual)**

**Motivation:**
- local network uses just one IP address as far as outside world is concerned:
  - no need to be allocated range of addresses from ISP: just one IP address is used for all devices
  - can change addresses of devices in local network without notifying outside world
  - can change ISP without changing addresses of devices in local network
  - devices inside local net not explicitly addressable, visible by outside world (a security plus).
**NAT: Network Address Translation**

**Implementation:** NAT router must:

- **outgoing datagrams:** replace (source IP address, port #) of every outgoing datagram to (NAT IP address, new port #)
  
  . . . remote clients/servers will respond using (NAT IP address, new port #) as destination addr.

- **remember (in NAT translation table)** every (source IP address, port #) to (NAT IP address, new port #) translation pair

- **incoming datagrams:** replace (NAT IP address, new port #) in dest fields of every incoming datagram with corresponding (source IP address, port #) stored in NAT table
NAT: Network Address Translation

- 16-bit port-number field:
  - 60,000 simultaneous connections with a single LAN-side address!
- NAT is controversial:
  - routers should only process up to layer 3
  - violates end-to-end argument
    - NAT possibility must be taken into account by app designers, eg, P2P applications
  - address shortage should instead be solved by IPv6

Routing in the Internet

- The Global Internet consists of Autonomous Systems (AS) interconnected with each other:
  - **Stub AS**: small corporation: one connection to other AS’s
  - **Multihomed AS**: large corporation (no transit): multiple connections to other AS’s
  - **Transit AS**: provider, hooking many AS’s together

- Two-level routing:
  - **Intra-AS**: administrator responsible for choice of routing algorithm within network
  - **Inter-AS**: unique standard for inter-AS routing: BGP
# 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 Intra-AS routing protocols:
  - 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 among neighbors every 30 sec via Response Message (also called advertisement)
- Each advertisement: list of up to 25 destination nets within AS

**RIP: Example**

<table>
<thead>
<tr>
<th>Destination Network</th>
<th>Next Router</th>
<th>Num. of 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>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Routing table in D
### RIP: Example

<table>
<thead>
<tr>
<th>Dest</th>
<th>Next hops</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td>A</td>
</tr>
<tr>
<td>x</td>
<td>-</td>
</tr>
<tr>
<td>z</td>
<td>C</td>
</tr>
<tr>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Destination Network | Next Router | Num. of hops to dest.
--- | --- | ---
| w | A | 2 |
| y | B | 2 |
| z | A | 5 |
| x | - | 1 |
| ... | ... |

Routing table in D

### 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 route-d (daemon)
- advertisements sent in UDP packets, periodically repeated

![Diagram](Network Layer 4-35)

RIP Table example (continued)

Router: giroflee.eurocom.fr

<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.</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.</td>
<td>193.55.114.6</td>
<td>U</td>
<td>3</td>
<td>58503</td>
<td>lo0</td>
</tr>
<tr>
<td>192.168.3.</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>lo0</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>

- 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)
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)
  - Carried in OSPF messages directly over IP (rather than TCP or UDP)

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; high for real time)
- Integrated uni- and multicast support:
  - Multicast OSPF (MOSPF) uses same topology database 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.
**Inter-AS routing in the Internet: BGP**

![Diagram showing inter-AS routing with AS1, AS2, AS3, and AS4.]

**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 AS's) to destination
  - BGP routes to networks (ASs), not individual hosts
  - E.g., Gateway X may send its path to dest. Z:

\[
\text{Path (X,Z) = X,Y1,Y2,Y3,...,Z}
\]
**Internet inter-AS routing: BGP**

**Suppose:** gateway X sends 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:
  Path \((W,Z) = w\), Path \((X,Z)\)
- Note: X can control incoming traffic by controlling it route advertisements to peers:
  - e.g., don’t want to route traffic to Z -> don’t advertise any routes to Z

---

**BGP: controlling who routes to you**

- A, B, C are provider networks
- X, W, Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - X does not want to route from B via X to C
  - ... so X will not advertise to B a route to C
**BGP: controlling who routes to you**

- A advertises to B the path AW
- B advertises to X the path BAW
- Should B advertise to C the path BAW?
  - No way! B gets no "revenue" for routing CBAW since neither W nor C are B's customers
  - B wants to force C to route to W via A
  - B wants to route only to/from its customers!

---

**BGP operation**

_Q: What does a BGP router do?_

- Receiving and filtering route advertisements from directly attached neighbor(s).
- Route selection.
  - To route to destination X, which path of several advertised) will be taken?
- Sending route advertisements to neighbors.
**BGP messages**

- 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