Introduction to Routing

Evi Nemeth, Univ. of Colorado

CAIDA's IEC Workshop
San Diego, August 1999

Addressing

IP

Routing

Routing Protocols

Topics - Addressing

- Addresses
- CIDR
- Netmasks

Addressing

- **For people: hostname**
  - □ **32 characters, A-Z, 0-9, -, _**
  - □ **Case insensitive**
- **For software: IP address**
  - □ **32 bits, dotted decimal representation**
- **For hardware: Link level address**
  - □ **48 bits, hexidecimal representation**
    - → **Ethernet's first 24 bits identify manufacturer of interface card**
    - → **Last 24 bits a unique serial number**
    - → **FDDI addresses are 48 bits too**

IP Address Details

- 32 bits, dotted decimal form
  - 128.138.243.151
- Identifies a "computer's" network interface
- "computer" can be
  - normal computer
  - network device
  - printer
  - toaster (soon)

IP Address, cont.

- Network interface can be:
  - □ Ethernet interface, cost $50-100
  - □ FDDI interface, cost $2000-4000
  - □ Serial port for SLIP/PPP, cost $8
  - □ ATM interface, cost$500-1000
  - □ WAN connection
    - → T1 (1.5mb/sec), cost $1500-3000
    - → T3 (45mb/sec), cost $20,000

Address Classes

- Address originally artificially divided into classes

- Made separating the network part and host part easier

- Inefficient use of address space

- Designed when "The network (ARPANet at the time) might grow to 200 hosts maximum"

Class A Network: N.H.H.H

- N byte assigned to you by the NIC (Network Information Center); H bytes yours to assign locally

- First byte 1-127, first bits 01

- Only 127 networks, but 2**24 hosts

- Huge network, too big

- Subnetted to N.S.S.H usually

Class B - N.N.H.H

- First byte 128-191, first bits 10
- 63*255 networks, 2**16 hosts
- Size about right for medium to large organization
- Usually subnetted to N.N.S.H

Class C - N.N.N.H

- First byte 192-223, first bits 110
- 32*255*255 networks
- But only 255 hosts each
- Size too little for most organizations
- Used to be no way to subnet

Other Classes

- Class D
  - □ First byte 224-239, first bits 1110
  - □ Used for multicast
- Class E
  - □ First byte 240-254, first bits 11110
  - □ Used for experiments

Address Depletion

- In 1993 a look at IMR (Internet Monthly Report) growth trends showed:
  - Running out of addresses
  - Out of class B's by mid '95
  - Routing tables too large
    - → Each class B is 1 routing table entry (OK)
    - → Each class C is 1 routing table entry (Not OK)
  - Network growing exponentially

Address Depletion

- Short term fix: CIDR - Classless Inter-domain Routing

- Long term fix: Bigger IP addresses, new IP protocol
  - Called IPv6
  - 128 bit addresses

- Last protocol change:
  - Occurred in 1978-81 NCP (Network Control Protocol) to TCP/IP
  - Took 3-4 years to complete the change
  - The network was infinitely smaller then

CIDR

• Don't use classes to determine network part

• Use netmask variable and notation /xx for xx bits of network address

• Traditionally:

- □ Class A network is /8
- □ Class B is /16
- □ Class Cis /24

• Under CIDR /xx is OK for almost any xx

Implications of CIDR

- Routing used initial bits of address to determine network

- Had to change to use netmask

- Routing protocols had to include netmask

- People had to learn binary/hex better

- More efficient allocation of addresses fixes depletion, what about routing table explosion

CIDR Blocks

- Suppose a site requests 100 class C addresses

- CIDR says no:
  - You can have 64
  - You can have 128
  - But not 100, say you take 64 = 2**6

- Creates block of addresses with

```
<-------->|<---------->    <--->
  constant     6 bits      host
            that vary
           from 000000
            to 111111
```

CIDR Block, cont.

- **64 consecutive network #'s**

- **But, don't need 64 routing table entries**

- **Just list routing info as one entry with netmask 6 bits less.**

```
<-------->   <------->
constant      6 bits
bits          in CIDR block

<-------------------->
customer network length

<----------->
routing network length
```

CIDR Example

- Allocate the 64 Class C networks:

    192.135.0.H
    192.135.1.H
    192.135.2.H
    . . .
    192.135.63.H

as 192.135.0.0./18 and use the netmask 255.255.192.0

- Arithmetic gets messier if not on byte boundary and not starting with 0

Special IP addresses

• Certain addresses are reserved or have special meanings

- □ 127.0.0.1 local host, loopback address
- □ 127.x.x.x. still local host
- □ host bits all =0 network address
- □ host bits all =1 broadcast address
- □ 0.0.0.0 this host on this network

Special Addresses Example

- **For our CIDR block**
  - □ **192.135.0.0/18 is net address externally**
  - □ **192.135.1.0/24 is an internal class C**
  - □ **192.135.2.0/26 is an internal class C divided into 4 pieces (first piece)**
  - □ **192.135.2.64/26 is the second piece**
  - □ **192.135.2.127is the broadcast address of the second piece**

Today's IP Addresses

- Class A- rarely allocated

- Class B- very tight, trying to reclaim under-utilized ones

- Class C
  - Old ones allocated called the 192-swamp
  - New ones allocated in CIDR blocks

- Everyone getting good at netmasks

- Everyone getting stingy at giving them out in large blocks

- Allocation agencies have changed

IP Allocation

- Was centrally done by internic to end users
- Now in US
  - NIC replaced by ARIN who sells to ISPs
  - ISPs sell/give to customers
  - Sometimes many levels, not just 2 shown
- In Europe - RIPE, in Asia - APNIC
- Who owns an address
  - ISP
  - End customer
- How can customer change ISPs -- renumber?

IP Allocation, cont.

- Cost of renumbering huge
  - Took SGI 2 years, several staff
- ISPs mostly say they won't "give" addresses
  - Address stays with ISP
- IETF came up with "fix"
  - Identified range of numbers for internal use, called private networks
  - Not routable
  - Not unique
  - Proxy machine converts to real #'s
  - Called NAT - Network Address Translation

IP - Topics

- Network Layer
- Fragmentation
- IP Packet Format
- IP Protocol

Network Layer

- Gets the packet to the right machine
  - □ Header checksum guarantees IP header OK
  - □ Broken packets discarded, rather than delivered to the wrong place
- Not reliable, best effort delivery
  - □ No checksum over the data, just over the header
- Biggest chore of this layer is routing

IP Datagram

- Basic unit of transfer called a datagram

```
-------------------------------------------------------------
|   Frame Header  |   IP header     |          Data        |
-------------------------------------------------------------
```

- Datagram contains:
  - □ Frame header, from link layer
  - □ IP header, from network layer
  - □ Data, from higher layers

- What size should datagram be
  - □ MTU determines maximum packet size for different media
  - □ Don’t know which media packet will traverse
    - → Intermediate routers must be able to fragment packet
    - → Destination host re-assembles packet

IP Packet Format

```
1      4      8             16      19                          32
|------|------|--------------|-----------------------------------|
| ver  | hdr  | type of serv |       total length                |
|      | len  |              |                                   |
|---------------------------|-------|----------------------------|
|     identification        | flags |    fragment offset         |
|                           |       |                            |
|-------------|--------------|------------------------------------|
|time to live |   protocol   |       header checksum              |
|-----------------------------------------------------------------|
|              IP source address, 32 bits                         |
|-----------------------------------------------------------------|
|              IP destination address                             |
|-----------------------------------------------------------------|
|              IP options if any (0-320 bits)                     |
|-----------------------------------------------------------------|
|                                                                 |
|              user data                                          |
|                                                                 |
|-----------------------------------------------------------------|
1      4      8             16      19                          32
```

Fields of the IP Header

- Version
  - 4 bits, current value is 4
  - Version 6 is defined
- Header Length
  - 4 bits, units are 32 bit words
  - If there are no options, value is 5 or 20 bytes
  - Maximum header size is 60 bytes

IP Header, cont.

- Type of Service
  - □ Not widely used, yet
  - □ 3 bits for precedence; 4 bits for type of service, 1 bit unused
  - □ TOS bits, from left to right
    - → Low delay (minimize), for interactive traffic
    - → High throughput (maximize), for file transfer
    - → High reliability
    - → Low cost

IP Header, cont.

- TOS (Type of Service) not currently used
  - Proposed use for differentiated service (diffserv)
  - Busy router doesn't have many choices
    - → Throw packet away
    - → Forward with normal priority
    - → Forward with higher priority
- RSVP protocol being developed to provide end-to-end quality of service guarantees
  - Bogged down in standardization process
  - May be replaced by differentiated services work (Van Jacobson)

IP Header, cont

- Total Length
  - 16 bits
  - Includes data and IP header, but not frame header
- Identification
  - 16 bits
  - Used to be sure that the re-assembly process glues the right fragments back together

IP Header, cont

- **Flags**
  - □ **3 bits, 2 low order ones control fragmentation**
    - → **Don't fragment bit: 1=don't, 0=OK to**
    - → **More fragments bit: 1=yes more, 0=no last one**

- **Fragment Offset**
  - □ **13 bits, units are 8 bytes (the other 3 bits needed)**
  - □ **Offset in the original unfragmented datagram where this fragment starts**
    - → **Always relative to the beginning of the packet**
    - → **Can fragment fragments**

IP Header, cont.

- TTL - time to live
  - □ 8 bits
  - □ To avoid infinite loops in the network
  - □ Decremented by each machine that forwards the packet
  - □ Usually set to 32 or 64, sometimes 255
- Protocol
  - □ 8 bits
  - □ Identifies the higher level protocol (TCP, UDP) sending the packet
    - → Protocol numbers given out by the IANA

IP Header, cont.

- **Header Checksum**
  - **16 bits**
  - **Just over the header, not the user's data**
  - **If it doesn't match at the receivers end, datagram is thrown away**
    - → **No error message is sent**
    - → **Higher level protocol must deal with lost packet**
- **Source IP Address**
  - **32 bits, IP address of the sender**
- **Destination IP Address**
  - **32 bits, IP address of the recipient**

IP Header, cont.

- Options
  - ☐ 0-320 bits (0 - 40 bytes) of optional stuff
  - ☐ Security handling restrictions
  - ☐ Record route
  - ☐ Time stamp
  - ☐ Loose source routing
    - → Designed for testing
    - → If misused, creates security problems
  - ☐ Strict source routing
- Options are rarely used, they slow a Cisco router down by a factor of 100-1000 depending on the model.

IP Limitations

- Total length field 16 bits long implies the maximum packet size is 64k, supercomputers want more

- TTL field is 8 bits long, implies maximum is 256 hops, is that too big?

- Checksum is over header only, not data
  - Insures that packet is delivered to the right place
  - No protection against errors in the data

IP Limitations, cont.

- **Header length field is only 4 bits, max value 15 words, so 60 bytes**
  - **Not enough for timestamp or record route options in todays Internet**

- **Options make the header variable length**
  - **If none, header is 20 bytes**
  - **Max is 60 bytes**
  - **Variable length things are slower to process than fixed length**

Routing - Topics

- Connecting Networks
- Routing Models
- Routing Protocols

Background Information -- Routing

- **Remember the ISO/OSI 7 layer model**
  - **Layer 7 - application**
  - **Layer 6 - presentation**
  - **Layer 5 - session**
  - **Layer 4 - transport, TCP/UDP protocols - conversations**
  - **Layer 3 - network, IP protocol, routing done here - packets**
  - **Layer 2 - link, media access - frames**
  - **Layer 1 - physical, wire, fiber, microwave, etc. - bits**

Connecting Networks

- **At layer 1**
  - □ **Connectors splicing cables together**
  - □ **Repeaters, or hubs**
  - □ **Unkind to the network, packets are re-produced on each segment**
- **At layer 2**
  - □ **Bridges or switches**
  - □ **Network protocol independent**
  - □ **Can learn which hosts are connected to which interfaces**
  - □ **Forwards broadcasts**
- **At layer 3**
  - □ **Routers**
  - □ **Protocol specific, eg. IP router**
  - □ **Kindest to the network, does not for-ward broadcasts**

Protocols

- Some protocols can be routed or bridge; others must be bridged
- IP - can be routed
- IPX - can be routed
- Appletalk - can route, usually encapsulated
- Netbeui (Microsoft) - must bridge
- Netbios (IBM) - must bridge
- LAT (Dec) - must bridge

Routing Chores

- Routing involves 2 chores
  - □ Looking up an address in a table and forwarding packets based on that lookup
  - □ Maintaining the tables
    - → Statically by human configuration
    - → Dynamically with a routing daemon speaking a routing protocol to other routers
- Bridging involves learning algorithm
  - □ Happens automatically, just plug it in
  - □ Doesn't scale, large bridged networks are not robust
- Switching is bridging with many networks instead of just two

Routing Mechanism

- Packet contains:
  - Frame (hardware) addresses
  - IP addresses
- Don't know the hardware address of a destination across the network
  - Can't find out
- Hardware address used only on the local network
  - Destination IP address fixed by application
  - Destination hardware address filled in with the hardware address of the next hop router by the IP layer

Routing Algorithm

- Search the routing table for a complete match with IP destination
  - If matched send packet out indicated interface
- If not matched, search for network part of the address
  - If matched, send packet out
- If not matched, search for default entry
  - If found, ship the packet out
  - Else, report an error
- On the backbone routers
  - There are no default routes
  - Must find longest match in table of about 65,000 entries several hundred thousand times a second

Routing Table

- What data does it need?
  - □ Destination address, host address or network address
  - □ IP address of next hop-router
  - □ Which interface to send packet on to get to next-hop router
  - □ Cost metric
  - □ Netmask to help interpret destination address

Routing Basics

- Two basic choices of routing models
  - Static
  - Dynamic

- Static works, but doesn't scale well and is hard to maintain

- Done in UNIX with the "route" command

```
route add/delete network/default gateway-machine metric

route add 128.138.243.0 128.138.245.254  1

route add default 128.138.245.254  1

route delete 128.138.243.0 128.138.245.254
```

  - Gateway-machine must be on local network
  - No netmasks expressed or understood by the route command

Routing Basics, cont.

• Routing table can be displayed with "netstat" command

```
netstat -nr  (from ieng9.ucsd.edu, static routing)


Routing Table:
  Destination      Gateway         Flags  Ref   Use   Interface
-------------------------------- ----- ----- ------ ---------
127.0.0.1       127.0.0.1        UH    0   1668089  lo0
132.239.56.0    132.239.50.58    UG    0    449064
132.239.60.0    132.239.50.249   UG    0    215443
132.239.50.0    132.239.50.9     U     3     54981  nf0
224.0.0.0       132.239.50.9     U     3         3  nf0
default         132.239.50.56    UG    0   2578822


netstat -nr   (from nag.cs.colorado.edu, dynamic routing)


Routing tables
Destination     Gateway         Flags  Refs   Use  If     Pmtu PmtuTime
127.0.0.1       127.0.0.1       UH       0 123652  lo0    4608
128.138.130.100 128.138.202.1   UGHD     0   3329  lan0   1500
128.138.130.101 128.138.202.1   UGHD     0   2210  lan0   1500
128.138.130.102 128.138.202.1   UGHD     0  23162  lan0   1500
128.138.130.105 128.138.202.1   UGHD     0    503  lan0   1500
128.138.130.106 128.138.202.1   UGHD     0   4158  lan0   1500
128.138.130.111 128.138.202.1   UGHD     0   1753  lan0   1500
128.138.130.112 128.138.202.1   UGHD     0   2048  lan0   1500
128.138.130.114 128.138.202.1   UGHD     0    707  lan0   1500
128.138.130.117 128.138.202.1   UGHD     0   2091  lan0   1500
128.138.130.118 128.138.202.1   UGHD     0   5537  lan0   1500
128.138.130.121 128.138.202.1   UGHD     1   6538  lan0   1500
   ... another 50 or so entries
```


Routing vs. Routed Protocols

- Routing Protocols are protocols that implement routing algorithms
  - RIP
  - OSPF
  - BGP
  - IGRP, EIGRP
  - And lots of others
- Routed protocols are transport protocols that can be routed over an internetwork, for example
  - IP
  - DECnet
  - IPX

Routing Protocol Families

• There are two families of distributed routing algorithms

- □ Distance Vector
- □ Link State

• All popular routing network protocols are based on one of these

• Distance Vector is also sometimes called "Bellman-Ford" routing, after the algorithm it uses

• Link State routing uses Dijkstra's shortest path algorithm

Distance Vector Routing

- A distance vector routing protocol requires that each router maintain:
  - A list of each possible destination
  - The distance to each of those destinations
- Each router advertises to its neighbors its list of destinations
- Each router updates its tables based on the advertisements most recently received from its neighbors
  - Adjusts neighbors metric values to represent itself
  - Replaces routes in its tables if new metric values are better
- Gossipy protocol

Link State Routing

- A link state routing protocol requires that each router maintain:
  - A list of routers in its routing club
  - A list of their links
- Each router meets its club neighbors
  - Learns their names
  - Exchanges link state information with them
- Each router, having a full graph of the network, computes the shortest path to each destination

DV - Advantages/Disadvantages

- Simple to calculate
- Easy to propagate bad information
- Updates often broadcast based, can cause traffic spikes
- Slow to respond to changes in the network
- Example, RIP and the "count-to-infinity" problem

Count to Infinity Problem

- Chronic problem of distance vector protocols

```
          A
        /   \
       /     \
      /       \
    B -------- C
   /
  /
 D
```

- A advertises: A=0,B=1,C=1
- B advertises: B=0,A=1,C=1,D=1
- A updates tables to: A=0,B=1,C=1,D=2 and advertises it
- Backhoe cuts B-D link
- B advertises: B=0,A=1,C=1,D=infinity
- B hears A's advertisement of D=2, and advertises B=0,A=1,C=1,D=3
- A updates his tables and advertises A=0,B=1,C=1,D=4, ...
- After many 30 second update cycles D=infinity again and the count to infinity stops

Distance Vector Fixes

- To fix count-to-infinity
  - □ Split horizon: don't advertise routes out the link you learned them from
  - □ Poison reverse: advertise all routes, but with a metric of infinity out the interface you learned it on

- To fix broadcast storms
  - □ Triggered updates
    - → Don't mark routing entry down if it isnt updated regularly (< 3 min)
    - → Broadcast when there is a change, not every 30 sec.
  - □ Acknowledged updates
    - → If we use triggered updates, need acknowledgements
    - → Protocol still UDP, no guarantees

Distance Vector Fixes, cont.

- Black holes
  - Cisco router blasts out routing update every 30 seconds
    - → Back-to-back packets
    - → UDP so no reliable delivery
    - → Not even another Cisco can keep up to back-to-back packets
    - → Routes in later packets never make it out
    - → After 6 tries they time out, creating a black hole
    - → Routes on the fringes of making it out "flap" (are there, then gone, then there)

- Path hold down (IGRP)
  - Don't let routes flap
  - Ignore updates if the same route changes too often

LS - Advantages/Disadvantages

- Can be complex to calculate
- Link state routing makes network topology easier to discover
- Troubleshooting a network is easier with link state routing
- Updates are usually asynchronous and infrequent
- Link state routing converges more quickly when the network changes

Autonomous Systems

- An Autonomous System (AS) is a group of routers controlled by the same administrative entity

- Each AS has an AS number

- Some routing protocols use AS numbers
  - BGP
  - OSPF
  - EIGRP

- Some protocols propagate the AS number as part of their route state information
  - Must be globally unique AS number assigned by IANA

- Some protocols do not propagate AS information
  - AS number can be locally assigned
  - Must be unique within your site

Routing Protocols

- RIP - Routing Information Protocol

- IGRP/EIGRP - [Extended] Interior Gateway Routing Protocol

- OSPF - Open Shortest Path First

- BGP - Border Gateway Protocol, details this afternoon

RIP

- **Routing Information Protocol**
  - **The original routing protocol, vintage 1982 or so**
- **Distance-vector, interior gateway protocol**
- **Grew out of Xerox XNS (late 70s)**
- **Used by *routed* on many workstations**
- **Maximum metric of 15**
- **Broadcast routing protocol**
- **Routing updates every 30 seconds**
  - **"I’m a router to network xxx.xxx.xxx.xxx in yy hops"**

RIP - cont.

- Convenient solution for LANs. Spoken by:
  - □ Printers
  - □ Hubs
  - □ PCs
  - □ UNIX boxes

- If misused, can cause extreme network congestion and chaos

- UNIX boxes should always run *routed -q* to avoid advertising false information

- If you must use RIP, use it only to advertise a handful of networks and perhaps a default route

RIP Packet Format

```
1             8             16                            32
|-------------|-------------|-----------------------------|
   command    | version (=1)|         must be zero
              |             |
|---------------------------|-----------------------------|
     address family (=2)    |         must be zero
                            |
|---------------------------------------------------------|
             IP address, 32 bits

|---------------------------------------------------------|
                must be zero

|---------------------------------------------------------|
                must be zero

|---------------------------------------------------------|
                metric (1-16)

|---------------------------------------------------------|

   up to 24 more routes, with same format, ie:
   address family, 0, ip address, 0, 0, metric
   ...

|---------------------------------------------------------|
```

- **Advertises (to all routers on the local networks) that he can reach that IP address in metric hops.**

- **Lots of wasted space in the packet**

- **No netmasks so doesn't work for CIDR addresses**

RIP2 Packet Format

- Fills in the "must be zero's" with something useful

```
1              8              16                              32
|-------------|--------------|------------------------------|
|   command   | version (=2) |        routing domain        |
|             |              |                              |
|----------------------------|------------------------------|
|     address family (=2)    |        route tag             |
|                            |                              |
|-----------------------------------------------------------|
|             IP address, 32 bits                           |
|                                                           |
|-----------------------------------------------------------|
|             subnet mask, 32 bits                          |
|                                                           |
|-----------------------------------------------------------|
|             next-hop IP address                           |
|                                                           |
|-----------------------------------------------------------|
|                metric (1-16)                              |
|                                                           |
|-----------------------------------------------------------|
|                                                           |
|   up to 24 more routes, with same format, ie:             |
|   family, tag, ip addr, subnet, nexthop, metric           |
|   ...                                                     |
|                                                           |
|-----------------------------------------------------------|
```

- Inter-operates with RIP version 1

- Routing domain identifies the routing daemon process

- Route tag can be AS number for BGP

Debugging RIP

- **Use a packet sniffer, or...**
- **Turn on *routed -t* on a UNIX box**

```
% in.routed –t
RESPONSE from 199.45.198.253.520:
        dst 0.0.0.0 metric 1
        dst 199.45.197.0 metric 1
RESPONSE from 199.45.198.10.520:
        dst 0.0.0.0 metric 1
        dst 199.45.197.0 metric 1
RESPONSE from 199.45.198.34.520:
        dst 0.0.0.0 metric 1
        dst 199.45.197.0 metric 1
RESPONSE from 199.45.198.254.520:
        dst 0.0.0.0 metric 1
        dst 199.45.197.0 metric 1
```

- **What a mess**
  - **Four different gateways advertising the same routes all with the same metric**
  - **But not all have the same capability (Sun 2 vs. Sparc Ultra)**
- **This site needs *gated* instead of *routed* to artificially weight metrics**

RIP on a Cisco Router

- To configure RIP on a Cisco router:

```
router rip
network 192.108.21.0
```

  - Says you will speak RIP and advertise the network 192.108.21.0

- You can use "show ip route" to view the routing table

IGRP/EIGRP

- (Enhanced) Interior Gateway Routing Protocol
- Developed by Cisco in mid-1980's to replace RIP
- Distance-vector, interior gateway protocol
- Uses ASs to identify routing groups
- Handy for all-cisco networks, no so handy for heterogeneous networks
- Works well in a WAN environment, can be easily distributed into RIP

IGRP/EIGRP cont.

- Routing updates every 90 seconds for IGRP, asynchronous for EIGRP

- Complex metric includes factors for delay, load, bandwidth, and reliability; varies from 1-255
  - Network manager gets to tune parameters in formula
  - Better metric than RIP's, makes IGRP more stable

- Uses split horizon and triggered updates

IGRP/EIGRP cont.

- Uses path holddown
  - No updates accepted after a failure
  - Waits a while (several update periods) before accepting updates again
- Route Poisoning
  - Treats every update that appears to be counting to infinity as unusable
- Keeps all paths, not just the best
  - Makes recovery faster when route goes down

Weaknesses of IGRP

- Hard to balance load across multiple paths

- No netmasks in routing info

EIGRP - Enhanced IGRP

- Has netmasks

- Can do CIDR aggregation

- New diffusion algorithm
  - Controls transient loops during update propagation

E/IGRP, cont.

- IGRP is commonly used when sharing information with other Cisco routers on your WAN, such as your ISP
  - You must enable IGRP routing, specify an autonomous system number and the networks which will use IGRP

```
router igrp 101
network 129.19.0.0
network 192.108.21.0
```

- IGRP then shares routing information with other routers also in the same autonomous system
  - Advertised/accepted routes can be controlled with access lists

OSPF

- Open Shortest Path First
- Interior gateway, link-state protocol
- Also designed as a replacement for RIP
- Can balance load across multiple links
- Unlike RIP, can operate within a hierarchy

OSPF, cont.

- Largest entity within a hierarchy is an AS, uses internal AS numbers, unique within the organization

- Each AS is divided into "areas"

- Link State information is exchanged by all routers within an area; summarized at borders of areas

- OSPF "backbone" (area 0) is responsible for exchanging information between areas

SPF Algorithm

- The Shortest Path First (SPF) [sometimes called Dijkstra] algorithm is the basis for operation of OSPF

- Once interfaces are initialized, a router sends OSPF HELLO to its neighbors

- Neighbor routers exchange link state information; once they're synchronized, they are said to be "adjacent"

- Link state information is re-transmitted periodically
  - Every 90 minutes
  - Whenever a router's state changes

OSPF’s Algorithm, cont.

- **Each router calculates a shortest path tree**
  - **Uses itself as the root of the tree**
  - **Builds the routing table with this SPF information**

- **Designated router**
  - **Network segments with more than one router elect a "designated router"**
  - **DR is responsible for maintaining link state information for that segment**
  - **New router joining the club gets one answer from the DR, not multiples**

- **One big feature of OSPF is support of variable-length subnet masks (RIP2 and EIGRP do too)**

- **Currently, *gated* is the best way to support OSPF on a UNIX box**

OSPF on a Cisco

- On a cisco router, your OSPF clause might look like:

```
router ospf 235
network 199.45.100.0  0.0.0.255  area 4
network 199.45.101.0  0.0.0.255  area 0
```

  - Local AS number chosen is 235
  - You are willing to speak OSPF on networks 199.45.100.0 and 199.45.101.0

- Output of "show ip ospf"

```
show ip ospf

Routing Process "ospf 235" with ID 199.45.141.67
Supports only single TOS(TOS0) routes
Number of areas in this router is 1
   Area 1.1.1.1
     Number of interfaces in this area is 1
     Area has no authentication
     SPF algorithm executed 29 times
     Area ranges are
     Link State Update Interval is 0:30:00 and due in 0:23:15
     Link State Age Interval is 0:20:00 and due in 0:03:15
```

OSPF Details

- **Shortest path computation centralized, not distributed like distance vector protocol**

```
         1                  3
A ---------------- B --------------- C
|                  |              ___/
| 3              4 |       5 ___/
|                  |     ___/
D ---------------- E ---/
         6
```

A says:

```
from A, to B, link 1, distance 1
from A, to D, link 3, distance 1
```

B says:

```
from B, to A, link 1, distance 1
from B, to C, link 2, distance 1
from B, to E, link 4, distance 1
```

OSPF Details

- Suppose backhoe cuts link 1 between A and B
  - A says to D

    `from A, to B, link 1, distance infinite`

  - B says to C and E

    `from B, to A, link 1, distance infinite`

  - Floods to the rest of the network, ie D says:

    `from A, to B, link 1, distance infinite`

  - E hears it twice, messages are time stamped and have a message number

OSPF Details, cont.

- Everybody updates their database and recomputes routes
  - Updates take message number into account so old messages don't pollute the database
  - Packets from A to B get re-routed thru D and E
- Security of the database
  - Flooding includes hop-by-hop acknowledgments
  - Can authentic/encrypt
  - Have a checksum in the packets
  - Timeouts remove entries that are not refreshed periodically

Strengths of Link State Protocol

- Fast, loopless convergence
- Precise metrics, can do multiple metrics
- Supports multiple paths to destination, can do load balancing
    - Done by having the SPF algorithm compute second best path too

BGP

- Border Gateway Protocol
  - □ Distance vector protocol, used on the backbones
  - □ Covered in detail in this afternoons sessions
- Routing algorithms used to want to route packets the fastest way
  - □ Metrics were related to speed
- Routing algorithms on the backbone related to policy
  - □ Metrics related to cost, hot potato routing
    - → Get rid of packet as soon as you can
    - → Provide as little transit as possible