

18-345 – Fall 08

Lecture 19

Network Layer: IP and Routing in the Internet

Peter Steenkiste

Readings: Sections 8.1, 8.2, 8.3

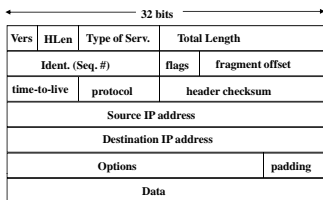


Overview Network Layer

- Network layer
- Packet network topology
- Distance vector routing
- Link state routing
- Internet addressing
- Internet protocol
- Routing in the Internet
- Putting things together



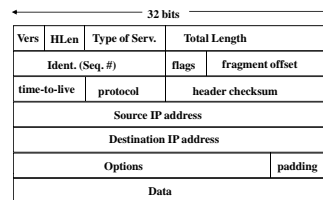
IP Datagram Format



- Protocol version allows coexistence of different protocol generations
- HLen give header length in multiples of 32-bit words
- Type of Service: reliability, speed, throughput; not used
- Total length of packet, including header
- All fragments of a datagram have same id #
- Fragment Offset specifies where in datagram a fragment belongs
- DF (don't fragment) flag bit; Min fragment is 576 bytes



Datagram Format

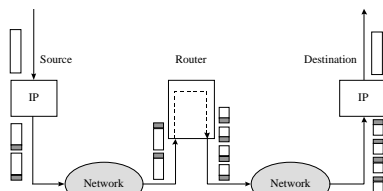


- Time-to-live counts down number of gateway hops traversed
- Protocol indicates which transport process datagram is to be delivered to, i.e. TCP, UDP, ...
- Header checksum (recomputed at each hop)
- Options allow additional information (very rare):
 - Security; Strict Source Routing; Loose Source Routing
 - Record Route; Time Stamp



Fragmentation & Reassembly

- **Maximum Transmission Unit (MTU)** maximum packet size that can be handled by a given network
- Sending IP entity must fragment packets into IP packets smaller than MTU
- Destination IP entity responsible for reassembly of fragments into original packet



Internet Control Message Protocol

- **Internet Control Message Protocol** allows hosts to interact with gateways, and hosts and gateways to interact with internet monitoring and control centers
- **Destination Unreachable**: router or subnet cannot locate destination; DF bit set and packet cannot pass network
- **Time Exceeded**
- **Parameter Problem**: illegal value in header
- **Redirect**: router notices incorrect packet routing
- **Echo request**; **Echo reply**
- **Timestamp request**; **Timestamp reply**



IPv6

New version of IP developed to deal with impending problems:

- Exhaustion of address space
- Size of routing tables
- Simplify packet processing, increase speed
- Improve security
- Quality of service support
- Aid multicasting
- Allow mobility
- Allow for evolution and smooth transition

Electrical & Computer ENGINEERING

IPv6 Header Format

← 32 bits →			
Vers	Priority	Flow Label	
Payload Length	Next header	Hop limit	
Source Address (16 bytes)			
Destination Address (16 bytes)			

- Header simplified relative to v4 (7 fields vs. 13 fields before)
- Version: 6 for v6; 4 for v4 (4 bits)
- Priority: 0-7 can be flow controlled; 8-15 real-time flows (4 bits)
- Routers examine flow label to determine type of treatment for packet; enables pseudo connections and QoS support in connectionless net
- Next header specifies type of additional extension headers which provide features previously available in v4 header; When header of last packet for a datagram, specifies transport protocol handler to which datagram is to be passed (8 bits)

Electrical & Computer ENGINEERING

← 32 bits →			
Vers	Priority	Flow Label	
Payload Length	Next header	Hop limit	
Source Address (16 bytes)			
Destination Address (16 bytes)			

- Hop limit decremented at each hop; to remove circulating packets
- Fixed-length 16 byte addresses:
 $2^{128} = 3 \times 10^{38} = 7 \times 10^{23}$ /square meter of earth's surface
- IPv4 addresses prefixed by 80 zeros + 16 ones
- Provider-based addresses
- Geographic-based addresses
- Multicast addresses (112 bit group identifier)
- Anycast address: routing to first host in a group of addresses
- New Address Notation: 8 groups of 4-hexadecimal digits
 4BF5:AA12: ... (8 such groups total)

Electrical & Computer ENGINEERING

Tunneling: IP-to-IP Encapsulation

- Can be used to force packet to travel through a specific node
- Intermediate routers may not understand inner IP header
 - E.g. internal IP addresses, new features, etc.

Electrical & Computer ENGINEERING

Tunneling Example

Electrical & Computer ENGINEERING

Mobile IP

- Proliferation of mobile devices: PDAs, laptops, cellphones, ...
- As user moves, point-of-attachment to network necessarily changes
- Problem: IP address specifies point-of-attachment to Internet
 - Changing IP address involves terminating all connections & sessions
- *Mobile IP (RFC 2002)*: device can change point-of-attachment while retaining IP address and maintaining communications

Electrical & Computer ENGINEERING

Routing in Mobile IP

- Home Agent (HA) keeps track of location of each Mobile Host (MH) in its network; HA periodically announces its presence
- If an MH is in home network, e.g. MH#1, HA forwards packets directly to MH
- When an MH moves to a Foreign network, e.g. MH#2, MH obtains a care-of-address from foreign agent (FA) and registers this new address with its HA

Routing in Mobile IP

- Correspondent Host (CH) sends packets as usual (1)
- Packets are intercepted by HA which then forwards to Foreign Agent (FA) (2)
- FA forwards packets to the MH
- MH sends packet to CH as usual (3)
- How does HA send packets to MH in foreign network?

Routing in the Internet

- Autonomous Systems (AS)
- Routing Information Protocol (RIP)
- Open Shortest Path First (OSPF)
- Border Gateway Protocol (BGP)

- Reading: Section 8.6
 - But only material covered in class
 - No message formats, etc.

Autonomous Systems

- Global Internet viewed as collection of autonomous systems.
- **Autonomous system (AS)** is a set of routers or networks administered by a single organization
- Same routing protocol need not be run within the AS
- But, to the outside world, an AS should present a *consistent picture of what ASs are reachable* through it
- **Stub AS**: has only a single connection to the outside world.
- **Multihomed AS**: has multiple connections to the outside world, but refuses to carry transit traffic
- **Transit AS**: has multiple connections to the outside world, and can carry transit and local traffic.

AS Number

- For exterior routing, an AS needs a globally unique AS 16-bit integer number
- Currently, there are about 11,000 registered ASs in Internet (and growing)
- *Stub AS*, which is the most common type, does not need an AS number since the prefixes are placed at the provider's routing table
- *Transit AS* needs an AS number
- Request an AS number from the ARIN, RIPE and APNIC

Inter and Intra Domain Routing

Interior Gateway Protocol (IGP): routing within AS

- RIP, OSPF

Exterior Gateway Protocol (EGP): routing between AS's

- BGPv4

Border Gateways perform IGP & EGP routing

Routing Information Protocol (RIP)

- RFC 1058
- Distributed in BSD UNIX
- Uses the **distance-vector algorithm**
- Runs on top of UDP, port number 520
- Metric: number of hops
- Max limited to 15
 - suitable for small networks (local area environments)
 - value of 16 is reserved to represent infinity
 - small number limits the *count-to-infinity* problem

Electrical & Computer ENGINEERING

RIP Operation

- Router sends update message to neighbors every 30 sec
- A router expects to receive an update message from each of its neighbors within 180 seconds in the worst case
- If router does not receive update message from neighbor X within this limit, it assumes the link to X has failed and sets the corresponding minimum cost to 16 (infinity)
- Uses **split horizon with poisoned reverse**
- Convergence improved by triggered updates
 - neighbors notified immediately of changes in distance vector table

Electrical & Computer ENGINEERING

Open Shortest Path First

- RFC 2328 (v2)
- Fixes some of the deficiencies in RIP
- Each router monitors the *link state* to each neighbor and floods the link-state information to other routers
 - Limits flooding using combination of TTL, sequence number, and list of visited routers
- Allows router to build shortest path tree with router as root
- OSPF typically converges faster than RIP when there is a failure in the network

Electrical & Computer ENGINEERING

OSPF Features

- *Multiple routes* to a given destination, one per type of service
- Support for *variable-length subnetting* by including the subnet mask in the routing message
- More *flexible link cost* - can range from 1 to 65,535
- Distribution of traffic over *multiple paths* of equal cost
- *Authentication* to ensure routers exchange information with trusted neighbors
- Uses *notion of area* to partition sites into subsets

Electrical & Computer ENGINEERING

OSPF Network

- To improve scalability, AS may be partitioned into areas
 - Area is identified by 32-bit Area ID
 - Router in area only knows complete topology inside area & limits the flooding of link-state information to area
 - *Area border routers* summarize info from other areas
- Each area must be connected to *backbone area* (0.0.0.0)
 - Distributes routing info between areas
- *Internal router* has all links to nets within the same area
- *Area border router* has links to more than one area
- *backbone router* has links connected to the backbone
- *Autonomous system boundary (ASB) router* has links to another autonomous system.

Electrical & Computer ENGINEERING

OSPF Areas

ASB: 4
 ABR: 3, 6, and 8
 IR: 1,2,7
 BBR: 3,4,5,6,8

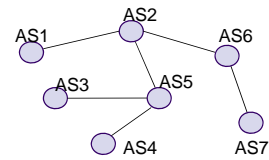
R = router
N = network

Electrical & Computer ENGINEERING

Routing in the Internet

- Autonomous Systems (AS)
- Routing Information Protocol (RIP)
- Open Shortest Path First (OSPF)
- Border Gateway Protocol (BGP)

Border Gateway Protocol v4



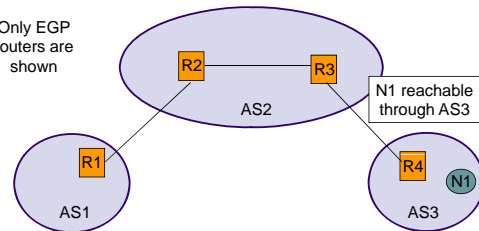
- BGP (RFC 1771) an EGP routing protocol to exchange network reachability information among BGP routers
- Network reachability info contains sequence of ASs that packets traverse to reach a destination network
- Info exchanged between BGP routers allows a router to construct a graph of AS connectivity
 - Routing loops can be pruned
 - Routing policy at AS level can be applied

Exterior Gateway Protocols

- Within each AS, there is a consistent set of routes connecting the constituent networks
- The Internet is woven into a coherent whole by *Exterior Gateway Protocols (EGPs)* that operate between AS's
- EGP enables two AS's to exchange routing information about:
 - The networks that are contained within each AS
 - The AS's that can be reached through each AS
- EGP path selection guided by policy rather than path optimality
 - Trust, peering arrangements, etc

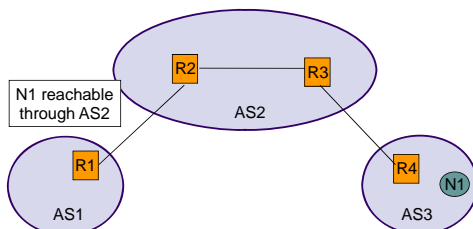
EGP Example

Only EGP routers are shown



- R4 advertises that network N1 can be reached through AS3
- R3 examines announcement & applies *policy* to decide whether it will forward packets to N1 through R4
- If yes, routing table updated in R3 to indicate R4 as next hop to N1
- IGP propagates N1 reachability information through AS2

EGP Example



- EGP routers within an AS, e.g. R3 and R2, are kept consistent
- Suppose AS2 willing to handle *transit* packets from AS1 to N1
- R2 advertises to AS1 the reachability of N1 through AS2
- R1 applies its policy to decide whether to send to N1 via AS2

EGP Requirements

- Scalability to global Internet
 - Provide connectivity at global scale
 - Link-state does not scale
 - Should promote address aggregation
 - Fully distributed
- EGP path selection guided by policy rather than path optimality
 - Trust, peering arrangements, etc
 - Key: only carry packets if it matches a business arrangement, e.g. you get paid
 - EGP should allow flexibility in choice of paths

BGP Features

- BGP is *path vector protocol*: advertises sequence of AS numbers to the destination network
- Path vector info used to prevent routing loops
- BGP enforces policy through selection of different paths to a destination and by control of redistribution of routing information
- Uses CIDR to support aggregation & reduction of routing information

Electrical & Computer ENGINEERING

Peering and Inter-AS connectivity

The diagram illustrates the hierarchy of inter-AS connectivity. At the top is a Peering Center. Below it are two Tier 1 ISPs (Transit AS). These Tier 1 ISPs connect to various Tier 2 (transit AS) and non-transit ASes. A Content or Application Service Provider (Non-transit) is shown connected to a Tier 2 (transit AS). Client ASes obtain service from Tier 2 ISPs.

- Non-transit AS's (stub & multihomed) do not carry transit traffic
- Tier 1 ISPs peer with each other, privately & in peering centers
- Tier 2 ISPs peer with each other & obtain transit services from Tier 1s
- Tier 1's carry transit traffic between their Tier 2 customers
- Client AS's obtain service from Tier 2 ISPs

Electrical & Computer ENGINEERING

Hop-by-hop Model

- BGP advertises to neighbors only those routes that it uses
 - Consistent with the hop-by-hop Internet paradigm
 - e.g., AS1 cannot tell AS2 to route to other AS's in a manner different than what AS2 has chosen (need source routing for that)
- BGP enforces policies by:
 - choosing paths from multiple alternatives
 - controlling advertisement to other AS's

Electrical & Computer ENGINEERING

Examples of BGP Policies

- A multi-homed AS refuses to act as transit
 - Limit path advertisement
- A multi-homed AS can be transit for some ASes
 - Only advertise paths to some ASes
- An AS can favor or disfavor certain AS's for traffic transit from itself
- Results in "valley-free" routing: path goes up and then down in the AS hierarchy
 - Means that you only carry traffic for which you get paid
 - Private peering can add "horizontal" links

Electrical & Computer ENGINEERING

Transit vs. Peering

The diagram shows four ISPs: Z, X, P, and Y. ISP Z is connected to ISP X via peering. ISP X is connected to ISP P via transit (\$\$\$). ISP P is connected to ISP Y via transit (\$\$\$). ISP Z is connected to ISP P via transit (\$\$\$). ISP X is connected to ISP Y via transit (\$). ISP Z is connected to ISP X via transit (\$\$). ISP X is connected to ISP P via transit (\$\$). ISP Y is connected to ISP P via transit (\$\$ 1/2).

Electrical & Computer ENGINEERING

BGP Policy

- Examples of policy:
 - Never use AS X
 - Never use AS X to get to a destination in AS Y
 - Never use AS X and AS Y in the same path
- *Import policies* to accept, deny, or set preferences on route advertisements from neighbors
- *Export policies* to determine which routes should be advertised to which neighbors
 - A route is advertised only if AS is willing to carry traffic on that route

Electrical & Computer ENGINEERING

