Internet Protocol (IP)

Guevara Noubir

Textbook: Computer Networks: A Systems Approach,
L. Peterson, B. Davie, Morgan Kaufmann

Chapter 4.

Lecture Outline

Internet Protocol

Addressing

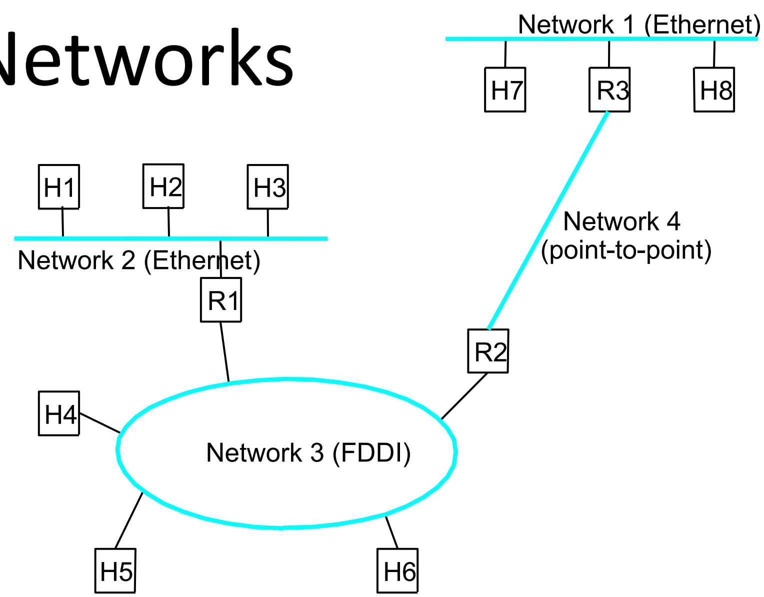
IP over LAN

Routing

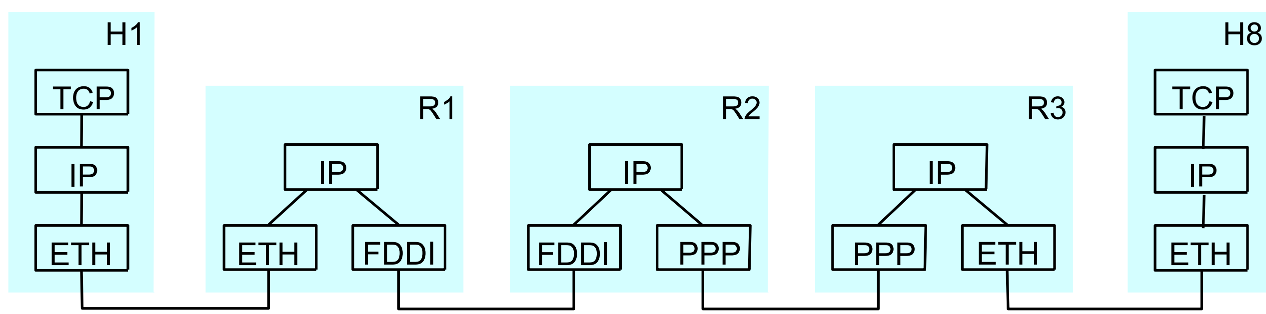
IPv6

IP Internet

- Concatenation of Networks

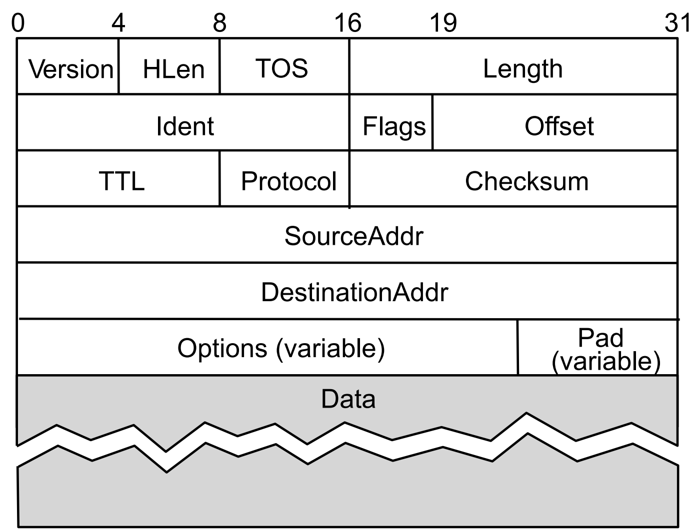


- Protocol Stack



Service Model

- Connectionless (datagram-based)
- Best-effort delivery (unreliable service)
  - packets are lost
  - packets are delivered out of order
  - duplicate copies of a packet are delivered
  - packets can be delayed for a long time
- Datagram format

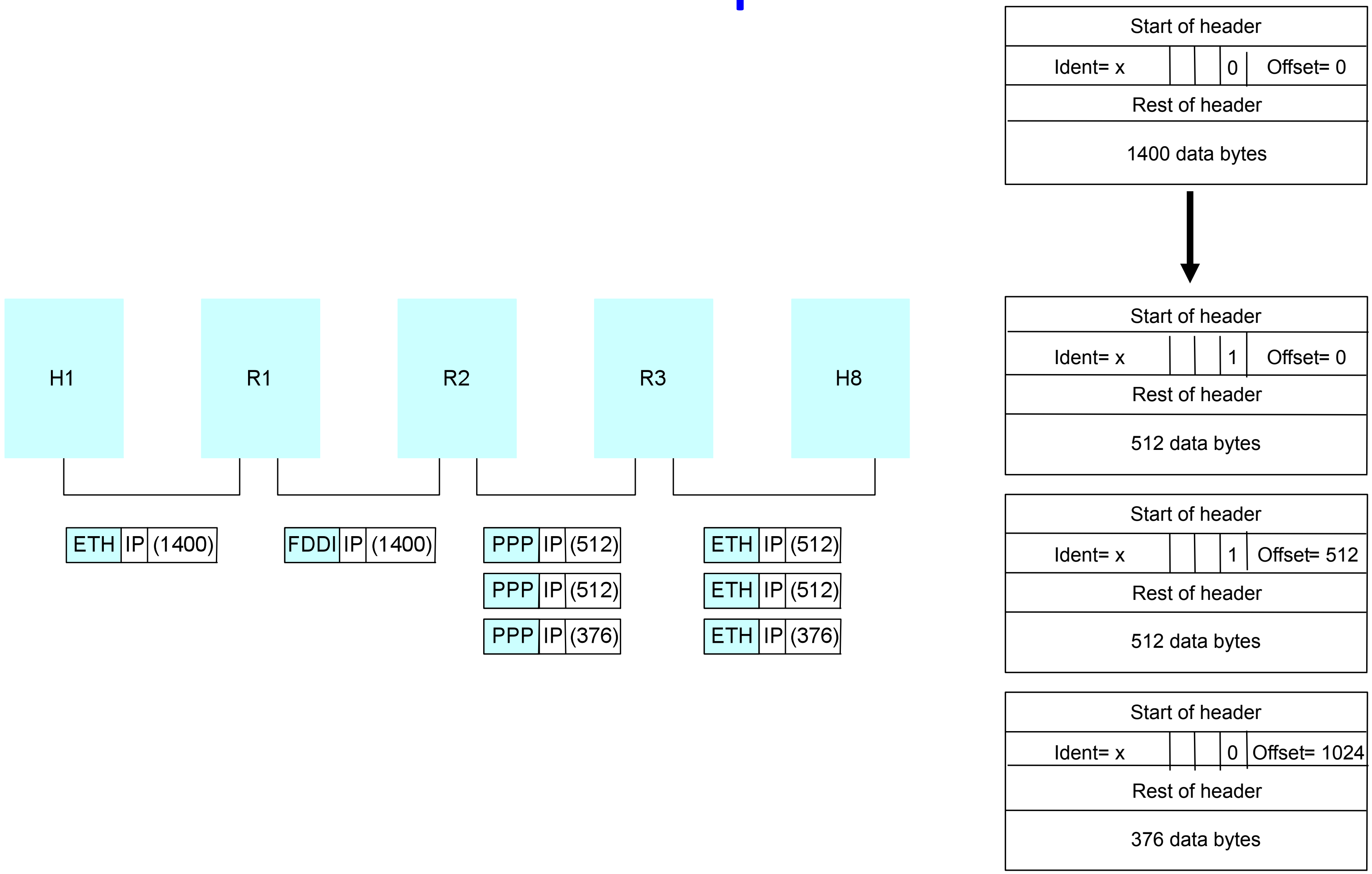


Fragmentation and Reassembly

- Each network has some MTU
- Strategy
  - fragment when necessary (MTU < Datagram)
  - re-fragmentation is possible
  - fragments are self-contained datagrams
  - use CS-PDU (not cells) for ATM
  - delay reassembly until destination host
  - do not recover from lost fragments

  - hosts are encouraged to perform “path MTU discovery”

Example



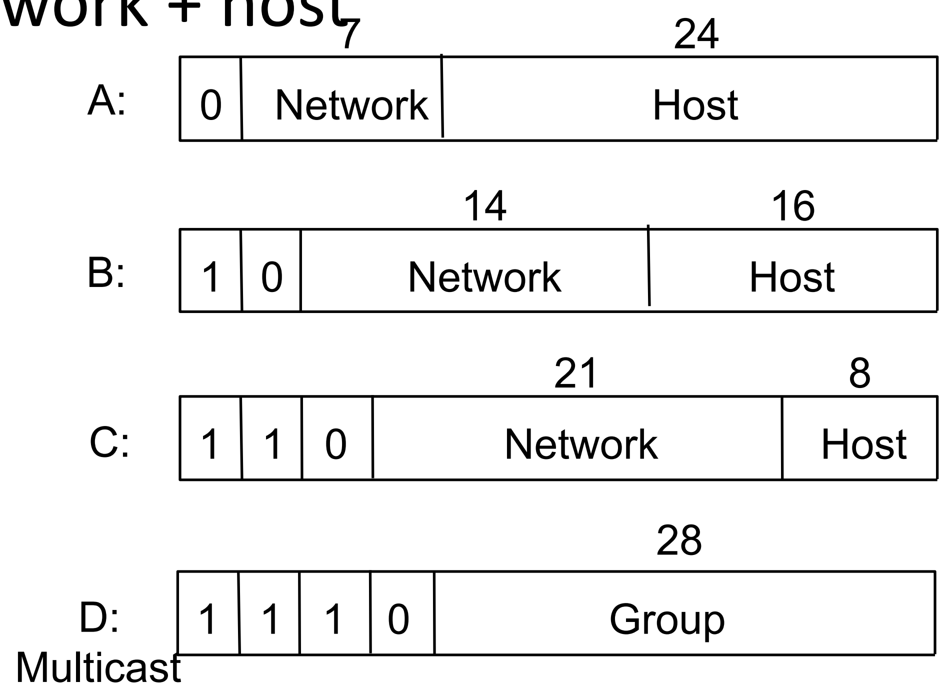
Start of header
Ident= x
0
Offset= 0
Rest of header
1400 data bytes
H1
R1
R2
R3
H8
ETH IP (1400)
FDDI IP (1400)
PPP IP (512)
PPP IP (512)
PPP IP (376)
ETH IP (512)
ETH IP (512)
ETH IP (376)
Start of header
Ident= x
1
Offset= 0
Rest of header
512 data bytes
Start of header
Ident= x
1
Offset= 512
Rest of header
512 data bytes
Start of header
Ident= x
0
Offset= 1024
Rest of header
376 data bytes

Internet Control Message Protocol (ICMP) RFC 792

- Integral part of IP but runs as ProtocolType = 1 using an IP packet
- Codes/Types:
  - Echo (ping)
  - Redirect (from router to source host)
  - Destination unreachable (protocol, port, host, cannot fragment)
  - TTL exceeded (so datagrams don't cycle forever)
  - Cannot fragment
  - Checksum failed
  - Reassembly failed

Global Addresses

- Properties
  - globally unique
  - hierarchical: network + host
- Dot Notation
  - 10.3.2.4
  - 128.96.33.81
  - 192.12.69.77



Datagram Forwarding

- Strategy
  - every datagram contains destination's address
  - if directly connected to destination network, then forward to host
  - if not directly connected to destination network, then forward to some router
  - forwarding table maps network number into next hop
  - each host has a default router
  - each router maintains a forwarding table
- Example (R2)

Network Number	Next Hop
1	R3
2	R1
3	interface 1
4	interface 0

Address Translation

- Map IP addresses into physical addresses
  - destination host
  - next hop router
- Techniques
  - encode physical address in host part of IP address
  - table-based
- ARP
  - table of IP to physical address bindings
  - broadcast request if IP address not in table
  - target machine responds with its physical address
  - table entries are discarded if not refreshed

ARP Details

- Request Format
  - HardwareType: type of physical network (e.g., Ethernet)
  - ProtocolType: type of higher layer protocol (e.g., IP)
  - HLEN & PLEN: length of physical and protocol addresses
  - Operation: request or response
  - Source/Target-Physical/Protocol addresses
- Notes
  - table entries timeout in about 15 minutes
  - update table with source when you are the target
  - update table if already have an entry
  - do not refresh table entries upon reference

ARP Packet Format

0 8 16 31

Hardware type = 1		ProtocolType = 0x0800
HLen = 48	PLen = 32	Operation
SourceHardwareAddr (bytes 0 – 3)		
SourceHardwareAddr (bytes 4 – 5)		SourceProtocolAddr (bytes 0 – 1)
SourceProtocolAddr (bytes 2 – 3)		TargetHardwareAddr (bytes 0 – 1)
TargetHardwareAddr (bytes 2 – 5)		
TargetProtocolAddr (bytes 0 – 3)		

ATMARP

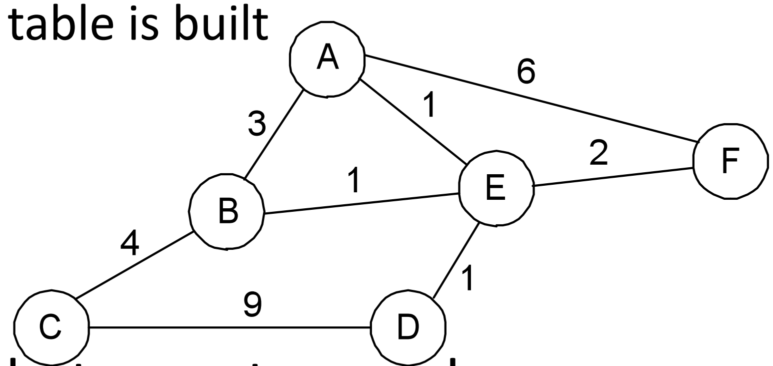
- ATM is not a broadcast network. There is a need for a specific address resolution mechanism.
- Use an ARP server:
  - Each node in the Logical IP Subnet (LIS) is configured with the ATM address of the ARP server
  - Each establishes a VC to the ARP server and register its <IP-ADDR, ATM-ADDR >
  - All address resolution requests are sent to the ARP server

Dynamic Host Configuration Protocol (DHCP)

- IP addresses of interfaces cannot be configured when manufactured (like for Ethernet)
- Configuration is an error-prone process
- Solution: centralize the configuration information in a DHCP server:
  - DHCP server discovery: broadcast a DHCPDISCOVER request
  - Request are relayed (unicast) to the server by DHCP relays
  - DHCP server broadcast replies with <HWADDR, IPADDR, lease-info>

Routing Overview

- Forwarding vs Routing
  - forwarding: to select an output port based on destination address and routing table
  - routing: process by which routing table is built
- Network as a Graph



- Problem: Find lowest cost path between two nodes
- Factors
  - static: topology
  - dynamic: load

Distance Vector

- Each node maintains a set of triples
  - **`(Destination, Cost, NextHop)`**
- Exchange updates directly connected neighbors
  - periodically (on the order of several seconds)
  - whenever table changes (called *triggered* update)
- Each update is a list of pairs:
  - **`(Destination, Cost)`**
- Update local table if receive a "better" route
  - smaller cost
  - came from next-hop
- Refresh existing routes; delete if they time out

Example

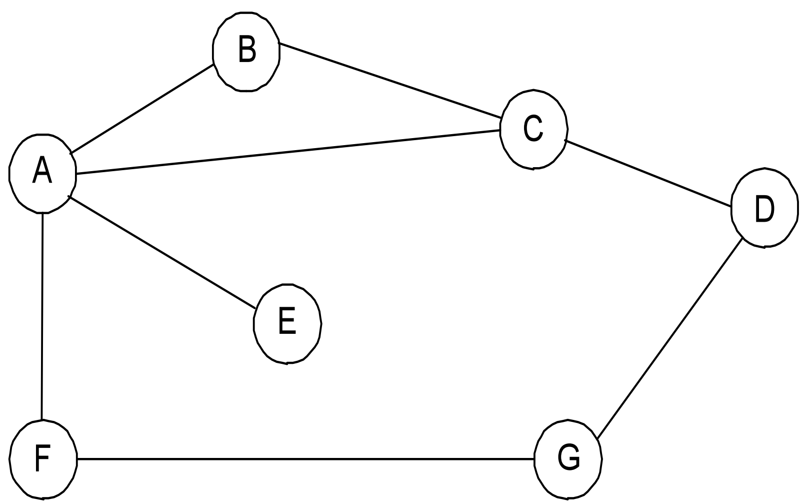


Table for node B

Destination	Cost	NextHop
A	1	A
C	1	C
D	2	C
E	2	A
F	2	A
G	3	A

Routing Loops

- Example 1
  - F detects that link to G has failed
  - F sets distance to G to infinity and sends update to A
  - A sets distance to G to infinity since it uses F to reach G
  - A receives periodic update from C with 2-hop path to G
  - A sets distance to G to 3 and sends update to F
  - F decides it can reach G in 4 hops via A
- Example 2
  - link from A to E fails
  - A advertises distance of infinity to E
  - B and C advertise a distance of 2 to E
  - B decides it can reach E in 3 hops (through C); advertises this to A
  - A decides it can read E in 4 hops (through B); advertises this to C
  - C decides that it can reach E in 5 hops…

Loop-Breaking Heuristics

- Set infinity to 16
- Split horizon
- Split horizon with poison reverse
- Waiting upon hearing failure
- Sequence number

Routing Information Protocol (RIP)

- Uses Bellman-Ford’s algorithm
- Protocol over UDP, port 520
- Distance-vector protocol
- Protocol overview:
  - Init: send a request packet over all interfaces
  - On response reception: update the routing table
  - On request reception:
    - if request for complete table (*address family*=0) send the complete table
    - else send reply for the specified address (infinity=16)
  - Regular routing updates:
    - every 30 seconds part/entire routing table is sent (broadcast) to neighboring routers
  - Triggered updates: on metric change for a route
  - Simple authentication scheme

Link State

- Strategy
  - send to all nodes (not just neighbors) information about directly connected links (not entire routing table)
- Link State Packet (LSP)
  - id of the node that created the LSP
  - cost of link to each directly connected neighbor
  - sequence number (SEQNO)
  - time-to-live (TTL) for this packet

Link State (cont)

- Reliable flooding
  - store most recent LSP from each node
  - forward LSP to all nodes but one that sent it
  - generate new LSP periodically
    - increment SEQNO
  - start SEQNO at 0 when reboot
  - decrement TTL of each stored LSP
    - discard when TTL=0

Route Calculation

- Dijkstra's shortest path algorithm
- Let
  - *N* denotes set of nodes in the graph
  - *l* (*i*, *j*) denotes non-negative cost (weight) for edge (*i*, *j*)
  - *s* denotes this node
  - *M* denotes the set of nodes incorporated so far
  - *C*(*n*) denotes cost of the path from *s* to node *n*

```
M = {s}
for each n in N - {s}
    C(n) = l(s, n)
while (N != M)
    M = M union {w} such that C(w) is the minimum for
        all w in (N - M)
    for each n in (N - M)
        C(n) = MIN(C(n), C (w) + l(w, n ))
```

Open Shortest Path First

- IP protocol (not over UDP), reliable (sequence numbers, acks)
- Protocol overview: link state protocol
  - The link status (cost) is sent/forwarded to all routers (LSP)
  - Each router knows the exact topology of the network
  - Each router can compute a route to any address
  - Simple authentication scheme
- Advantages over RIP
  - Faster to converge
  - The router can compute multiple routes (e.g., depending on the type of services, load balancing)
  - Use of multicasting instead of broadcasting (concentrate on OSPF routers)

Metrics

- Original ARPANET metric
  - measures number of packets enqueued on each link
  - took neither latency nor bandwidth into consideration
- New ARPANET metric
  - stamp each incoming packet with its arrival time (**`AT`**)
  - record departure time (**`DT`**)
  - when link-level ACK arrives, compute

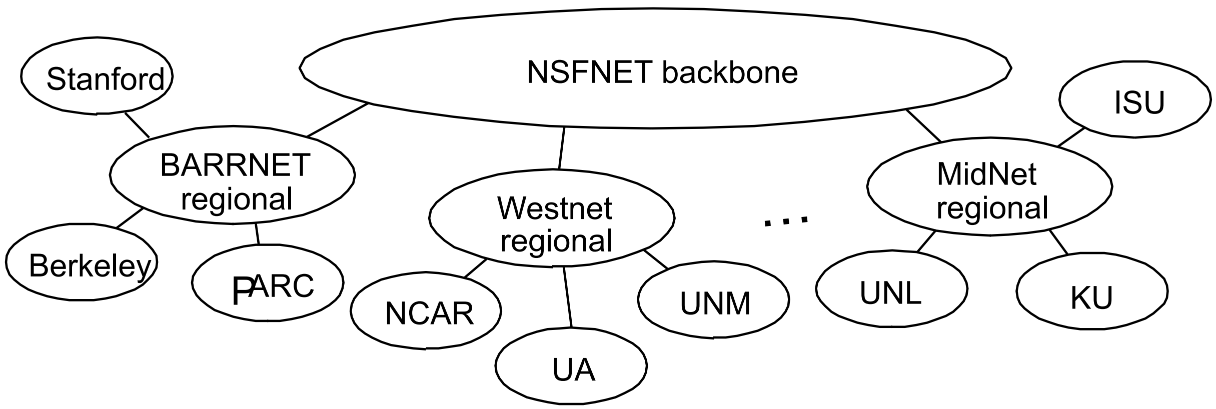
    `Delay = (DT - AT) + Transmit + Latency`
  - if timeout, reset **`DT`** to departure time for retransmission
  - link cost = average delay over some time period
- Fine Tuning
  - compressed dynamic range
  - replaced **`Delay`** with link utilization

Popular Interior Gateway Protocols

- RIP: Route Information Protocol
  - distributed with Unix
  - distance-vector algorithm
  - based on hop-count
- OSPF: Open Shortest Path First
  - recent Internet standard
  - uses link-state algorithm
  - supports load balancing
  - supports authentication

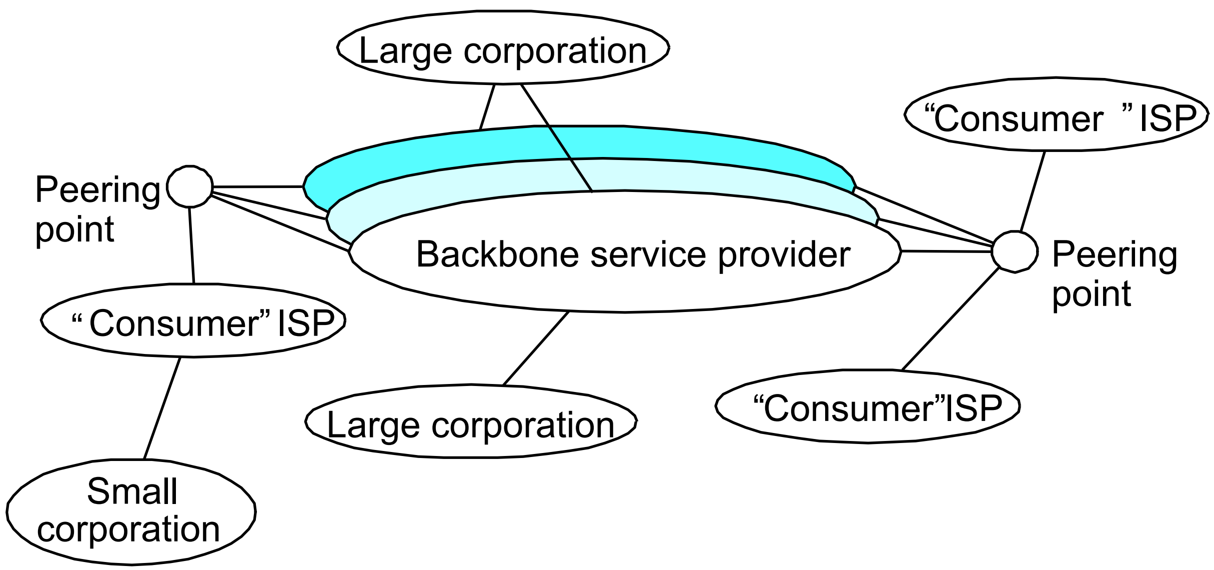
Internet Structure

Recent Past



Internet Structure

Today



How to Make Routing Scale

- Flat versus Hierarchical Addresses
- Inefficient use of Hierarchical Address Space
  - class C with 2 hosts (2/255 = 0.78% efficient)
  - class B with 256 hosts (256/65535 = 0.39% efficient)
- Still Too Many Networks
  - routing tables do not scale
  - route propagation protocols do not scale

Subnetting

- Add another level to address/routing hierarchy: *subnet*
- *Subnet masks* define variable partition of host part
- Subnets visible only within site

Network number	Host number

Class B address

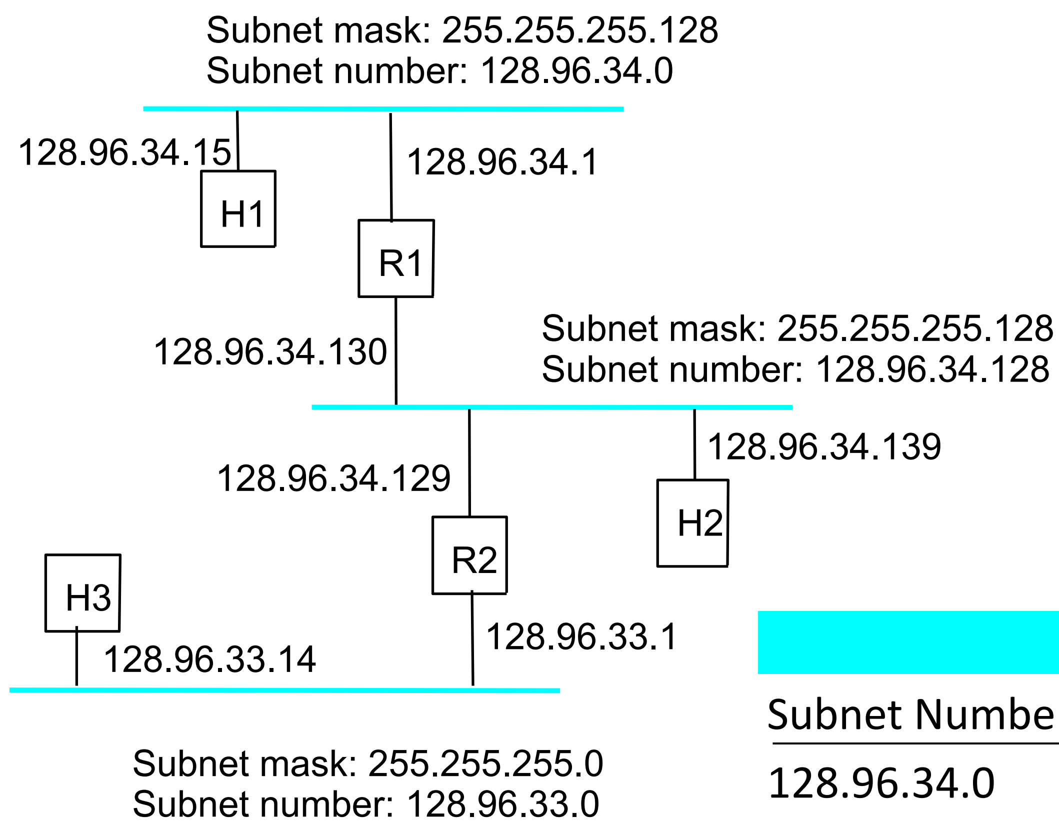
111111111111111111111111	00000000

Subnet mask (255.255.255.0)

Network number	Subnet ID	Host ID

Subnetted address

Subnet Example



Subnet Number	Subnet Mask	Next Hop
128.96.34.0	255.255.255.128	interface 0
128.96.34.128	255.255.255.128	interface 1
128.96.33.0	255.255.255.0	R2

Forwarding Algorithm

```
D = destination IP address
for each entry (SubnetNum, SubnetMask, NextHop)
   D1 = SubnetMask & D
   if D1 = SubnetNum
      if NextHop is an interface
         deliver datagram directly to D
      else
         deliver datagram to NextHop
```

- Use a default router if nothing matches
- Not necessary for all 1s in subnet mask to be contiguous
- Can put multiple subnets on one physical network
- Subnets not visible from the rest of the Internet

Supernetting

- Assign block of contiguous network numbers to nearby networks
- Called CIDR: Classless Inter-Domain Routing
- Represent blocks with a single pair

  `(first_network_address, count)`
- Restrict block sizes to powers of 2
  - E.g., 192.4.16 – 192.4.31
- Use a bit mask (CIDR mask) to identify block size
- All routers must understand CIDR addressing

Route Propagation

- Know a smarter router
  - hosts know local router
  - local routers know site routers
  - site routers know core router
  - core routers know everything
- Autonomous System (AS)
  - corresponds to an administrative domain
  - examples: University, company, backbone network
  - assign each AS a 16-bit number
- Two-level route propagation hierarchy
  - intradomain routing protocol (each AS selects its own)
  - interdomain routing protocol (Internet-wide standard)

EGP: Exterior Gateway Protocol

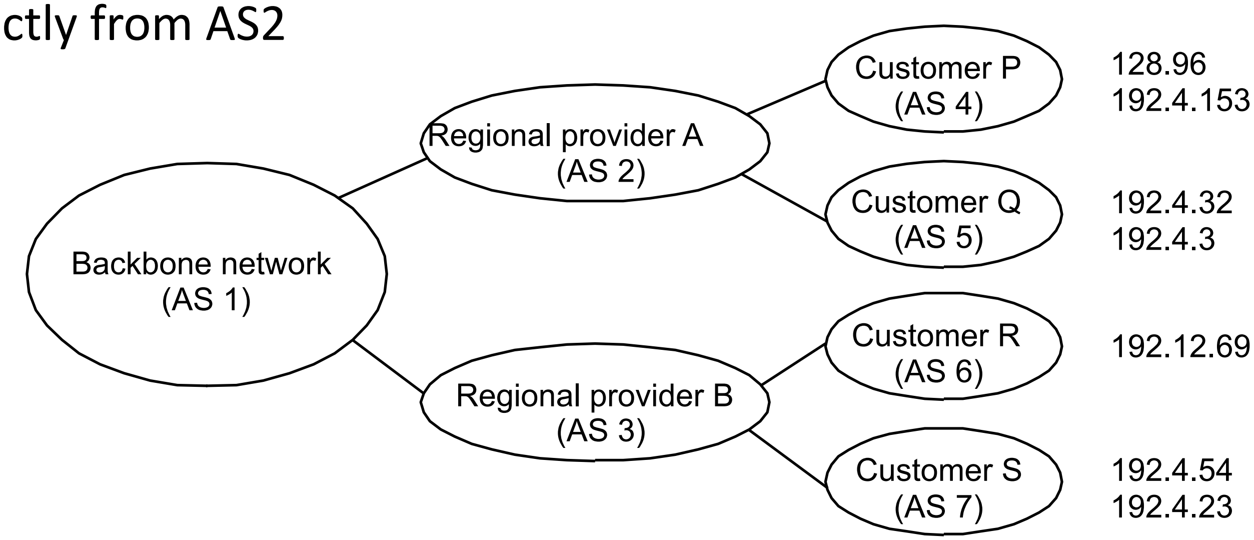
- Overview
  - designed for tree-structured Internet
  - concerned with *reachability*, and *policies* not optimal routes
- Protocol messages
  - neighbor acquisition: one router requests that another be its peer; peers exchange reachability information
  - neighbor reachability: one router periodically tests if the another is still reachable; exchange HELLO/ACK messages; uses a k-out-of-n rule
  - routing updates: peers periodically exchange their routing tables (distance-vector)

BGP-4: Border Gateway Protocol

- AS Types
  - stub AS: has a single connection to one other AS
    - carries local traffic only
  - multihomed AS: has connections to more than one AS
    - refuses to carry transit traffic
  - transit AS: has connections to more than one AS
    - carries both transit and local traffic
- Each AS has:
  - one or more border routers
  - one BGP *speaker* that advertises:
    - local networks
    - other reachable networks (transit AS only)
    - gives *path* information
- BGP-4 runs on top of TCP

BGP Example

- Speaker for AS2 advertises reachability to P and Q
  - network 128.96, 192.4.153, 192.4.32, and 192.4.3, can be reached directly from AS2



- Speaker for backbone advertises
  - networks 128.96, 192.4.153, 192.4.32, and 192.4.3 can be reached along the path (AS1, AS2).
- Speaker can cancel previously advertised paths

IP Version 6

- Features
  - 128-bit addresses (classless)
  - multicast
  - real-time service
  - authentication and security
  - autoconfiguration
  - mobility
  - end-to-end fragmentation
  - protocol extensions
- Addresses
  - notation: x:x:x:x:x:x:x:x where x is a hex representation of 16 bits
  - supports IPv4 addresses, multicast, link and site local addresses, anycast
- Header
  - 40-byte "base" header
  - version, priority, flow label, payload length, next header, hop limit, src, dst
  - no checksum
  - extension headers (fixed order, mostly fixed length)
    - e.g., NextHeader, Offset, M, Ident
    - fragmentation
    - source routing
    - authentication and security
    - other options
- Auto-configuration – concatenate
  - interface ID (e.g., MAC address)
  - prefix (e.g., for a printer use link local prefix 1111 1110 10)

Misc.

- Network Address Translation
- Virtual Private Networks
- Multi Protocol Label Switching (MPLS)