ΗΥ 335
Φροντιστήριο 8ο

Χειμερινό Εξάμηνο 2009-2010

Παπακωνσταντίνου Άρτεμις
artpap@csd.uoc.gr

4/12/2009

Roadmap

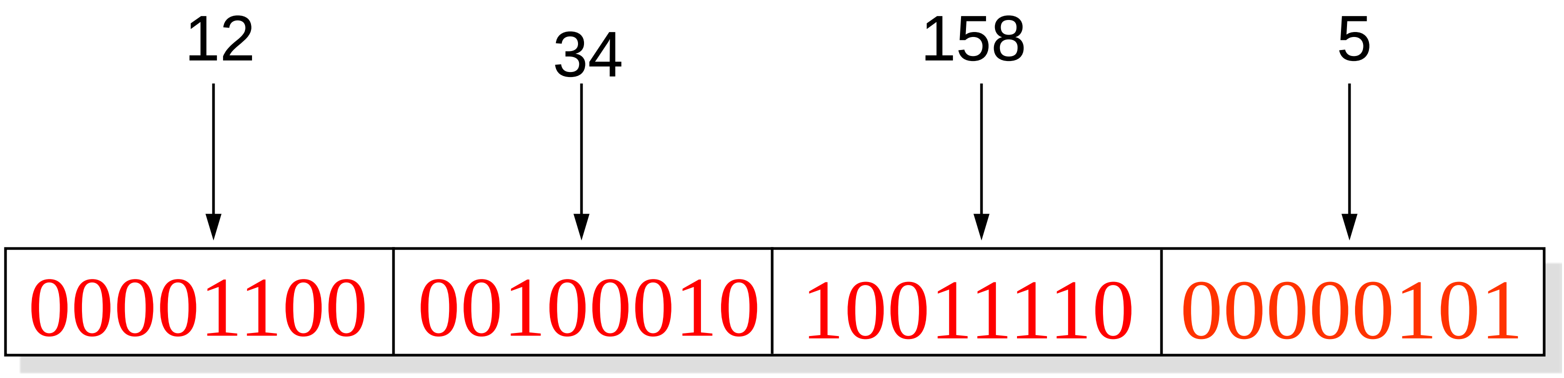
- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

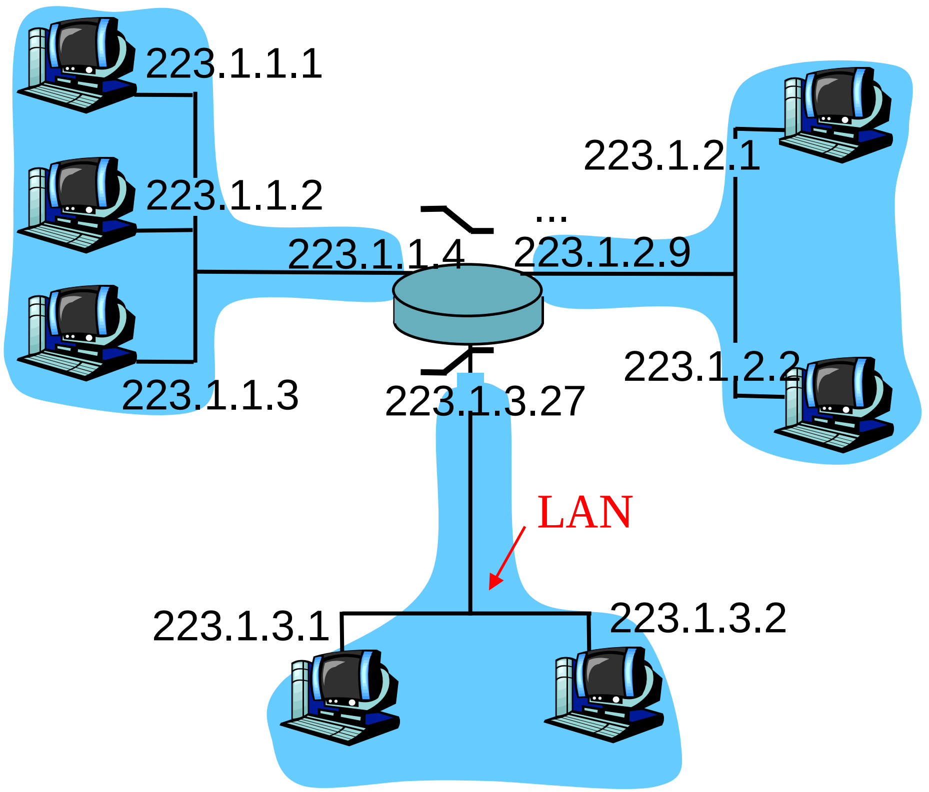
IP Address (IPv4)

- A unique 32-bit number
- Identifies an interface (on a host, on a router, …)
- *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
- Represented in dotted-decimal notation

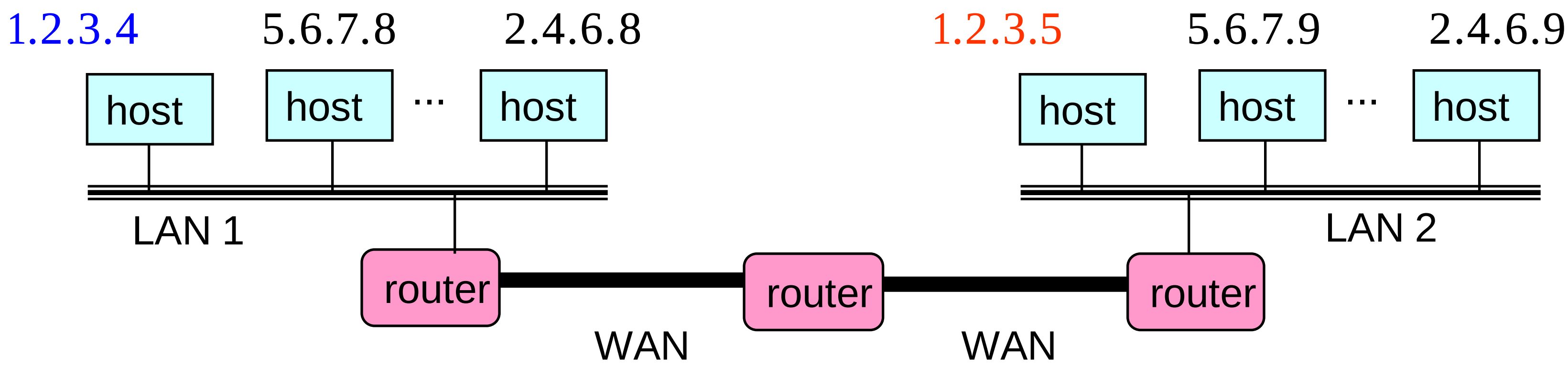


Grouping Related Hosts

- The Internet is an "inter-network"
  - Used to connect *networks* together, not *hosts*
  - Needs a way to address a network (i.e., group of hosts)



Scalability Challenge



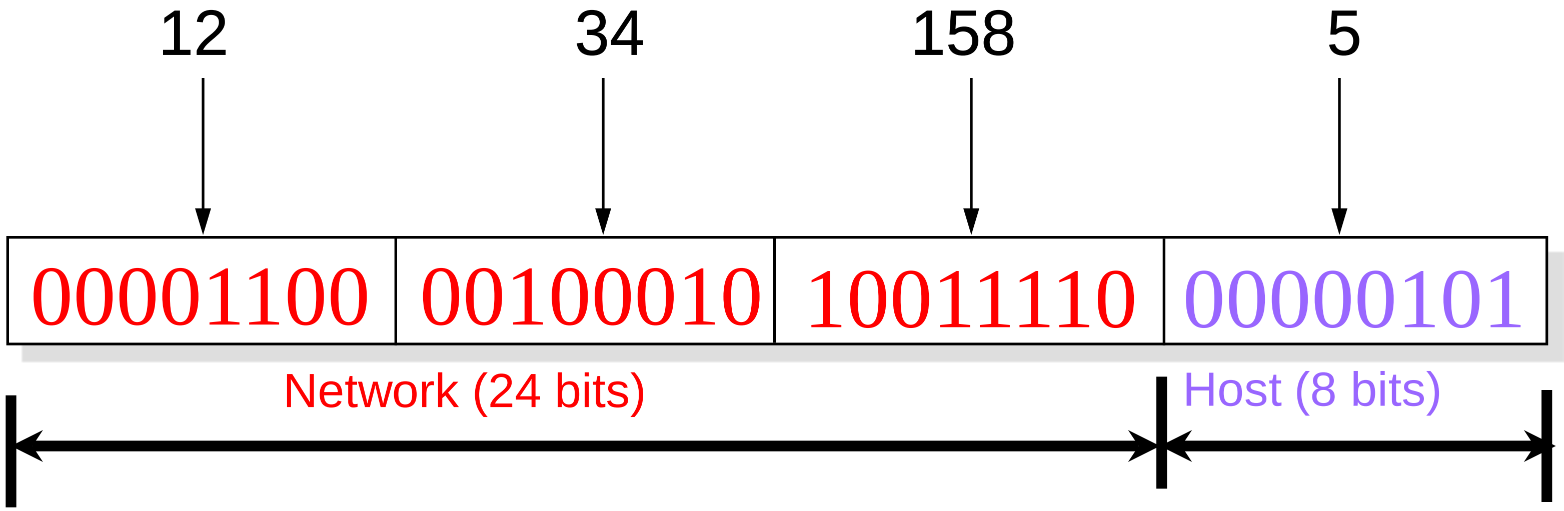
- Suppose hosts had arbitrary addresses
  - Then every router would need a lot of information
  - ...to know how to direct packets toward the host

1.2.3.4	←
1.2.3.5	→
⋮	

forwarding table

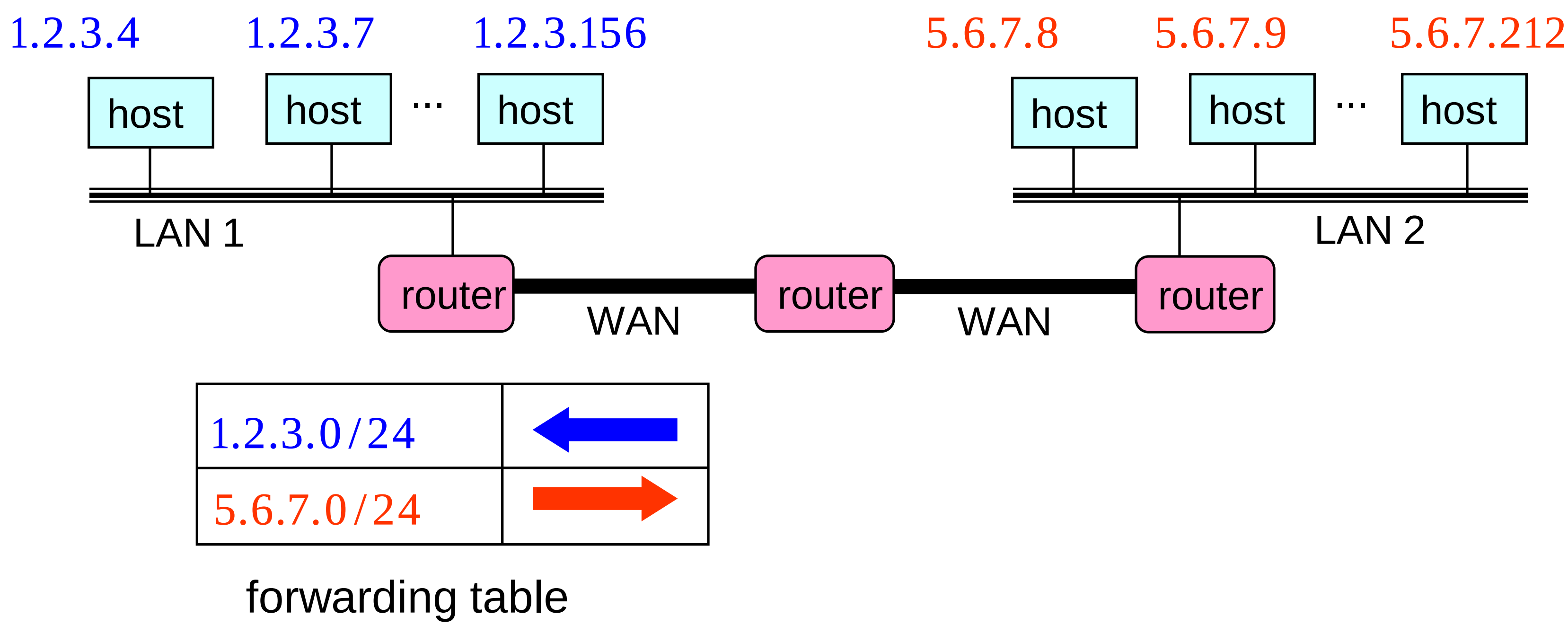
Hierarchical Addressing: IP Prefixes

- Divided into network & host portions (left and right)
- Forming *subnets:*
  - device interfaces with same network part of IP address
  - can physically reach each other without intervening router
- 12.34.158.0/24 is a 24-bit prefix with 2^8 addresses



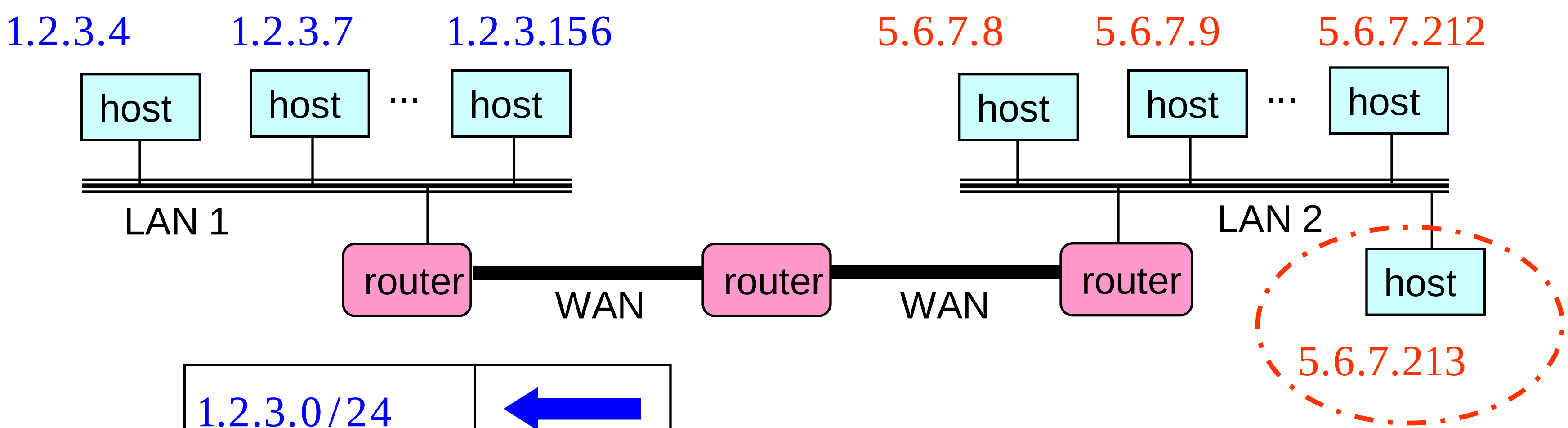
Scalability Improved

- Group related hosts from a common subnet
  - 1.2.3.0/24 on the left LAN
  - 5.6.7.0/24 on the right LAN



Easy to Add New Hosts

- No need to update the routers
  - E.g., adding a new host 5.6.7.213 on the right doesn't require adding a new forwarding-table entry



1.2.3.0/24	←
5.6.7.0/24	→

forwarding table

Class-full Addressing

- In the older days, only fixed allocation sizes
  - Class A: 0*
    - Very large /8 blocks (e.g., MIT has 18.0.0.0/8)
  - Class B: 10*
    - Large /16 blocks (e.g,. Princeton has 128.112.0.0/16)
  - Class C: 110*
    - Small /24 blocks (e.g., AT&T Labs has 192.20.225.0/24)
  - Class D: 1110*
    - Multicast groups

IP addressing: CIDR

- Class-full 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

network part | host part

11001000 00010111 00010000 00000000

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
- DHCP: Dynamic Host Configuration Protocol: dynamically get address from server
  - "plug-and-play"

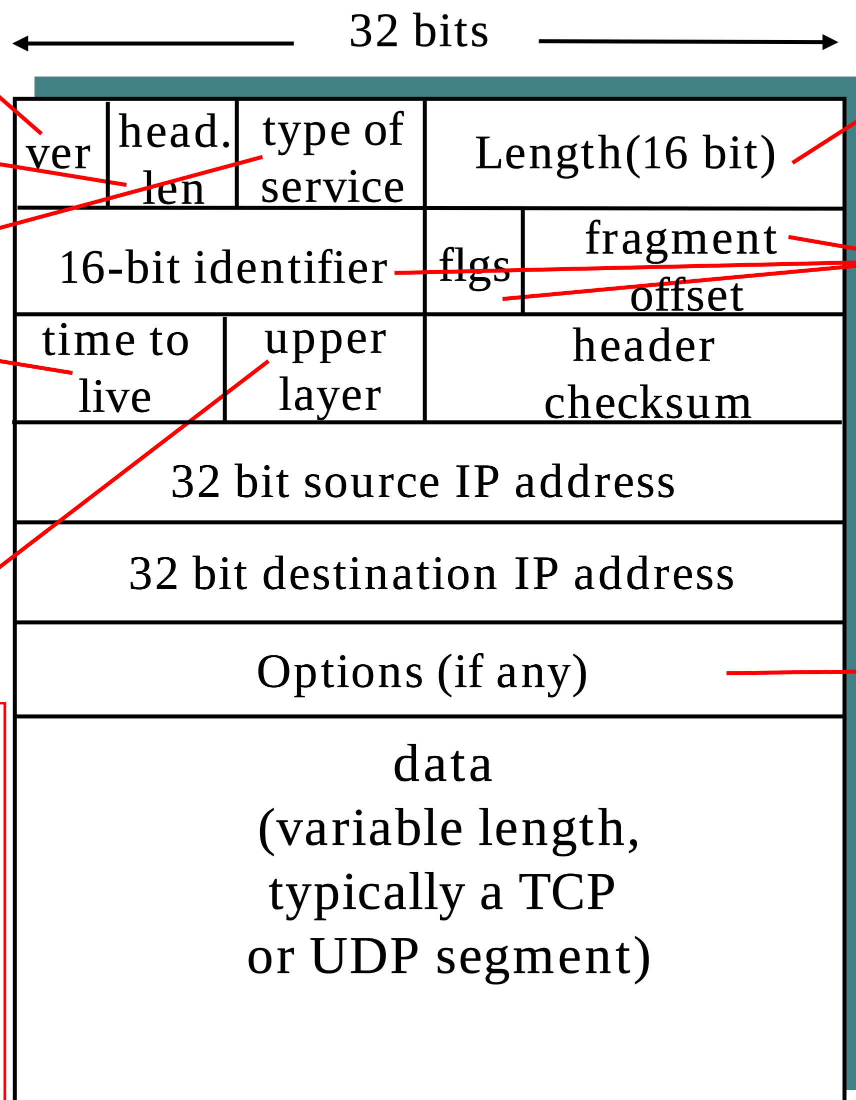
Obtaining a Block of Addresses

- Separation of control
  - Prefix: assigned *to* an institution
  - Addresses: assigned *by* the institution to their nodes
- Who assigns prefixes?
  - Internet Corporation for Assigned Names and Numbers
    - Allocates large address blocks to Regional Internet Registries
  - Regional Internet Registries (RIRs)
    - E.g., ARIN (American Registry for Internet Numbers)
    - Allocates address blocks within their regions
    - Allocated to Internet Service Providers and large institutions
  - Internet Service Providers (ISPs)
    - Allocate address blocks to their customers
    - Who may, in turn, allocate to their customers…

Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

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

total datagram length (bytes)

for fragmentation/ reassembly

E.g. timestamp, record route taken, specify list of routers to visit.

<u>how much overhead with TCP?</u>

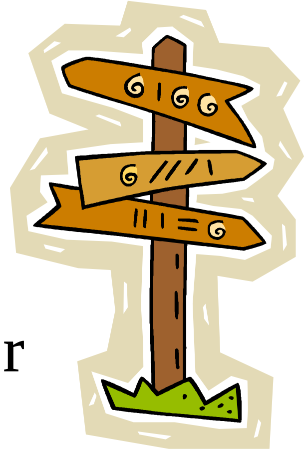
- 20 bytes of TCP
- 20 bytes of IP
- = 40 bytes + app layer overhead

Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

Hop-by-Hop Packet Forwarding

- Each router has a forwarding table
  - Maps destination addresses…
  - … to outgoing interfaces
- Upon receiving a packet
  - Inspect the destination IP address in the header
  - Index into the table
  - Determine the outgoing interface
  - Forward the packet out that interface
- Then, the next router in the path repeats
  - And the packet travels along the path to the destination



Getting a datagram from source to dest.

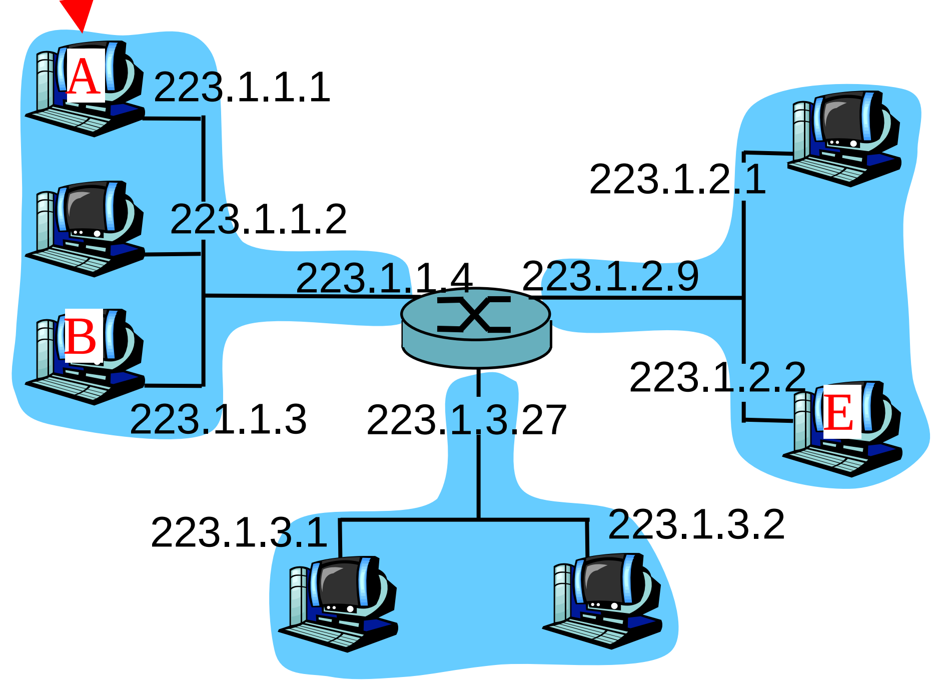
misc fields	223.1.1.1	223.1.1.3	data

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

forwarding table in A

Dest. Net.	next router	Nhops
223.1.1		1
223.1.2	223.1.1.4	2
223.1.3	223.1.1.4	2



Getting a datagram from source to dest.

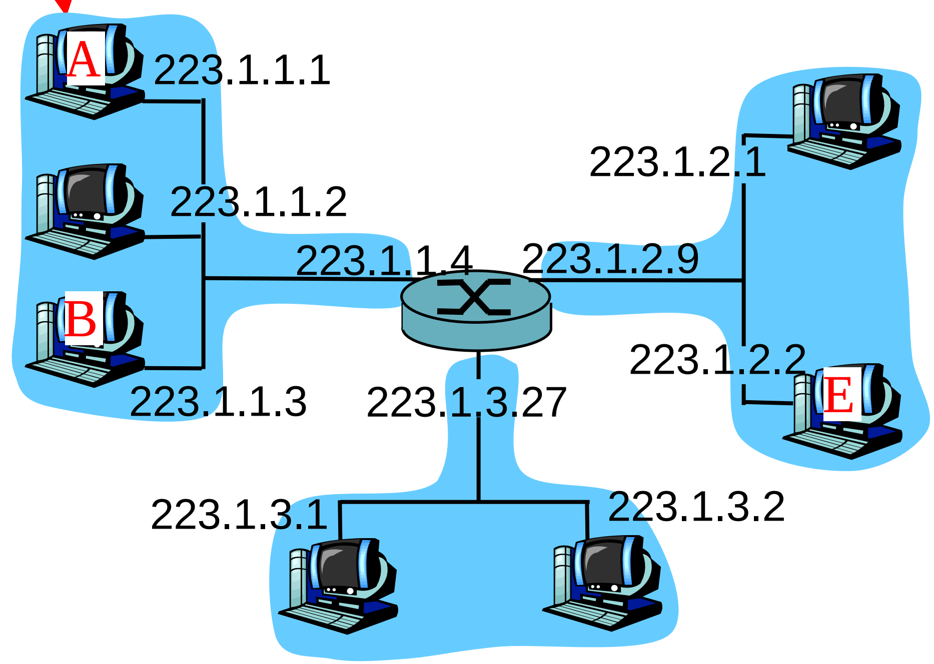
misc fields	223.1.1.1	223.1.2.2	data

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.....

forwarding table in A

Dest. Net.	next router	Nhops
223.1.1		1
223.1.2	223.1.1.4	2
223.1.3	223.1.1.4	2

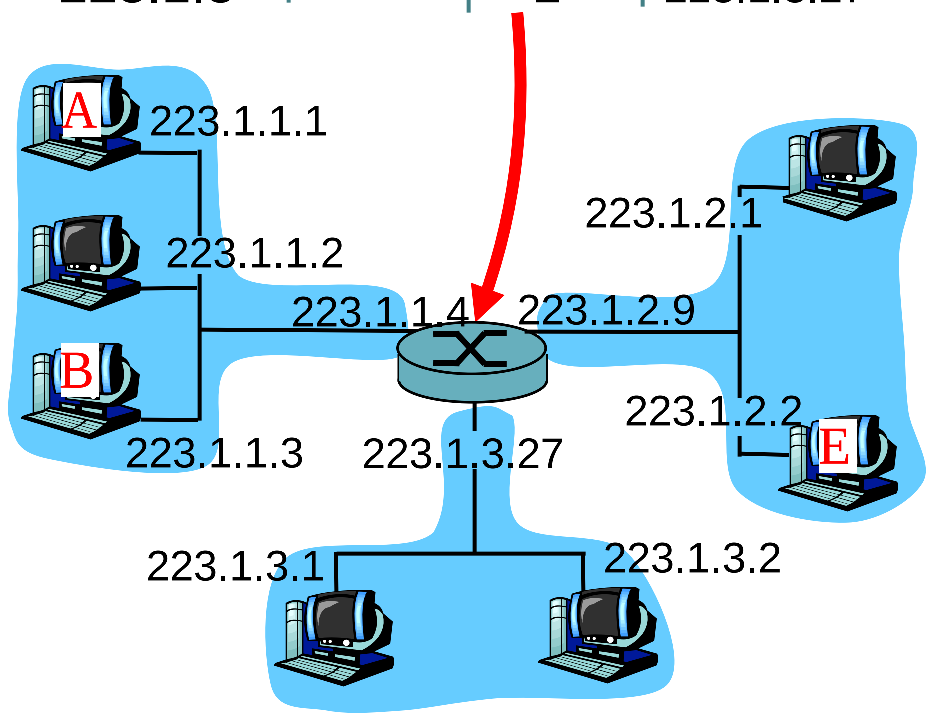


Getting a datagram from source to dest.

misc fields	223.1.1.1	223.1.2.2	data

forwarding table in router

Dest. Net	router	Nhops	interface
223.1.1	-	1	223.1.1.4
223.1.2	-	1	223.1.2.9
223.1.3	-	1	223.1.3.27



Arriving at 223.1.4, destined for 223.1.2.2

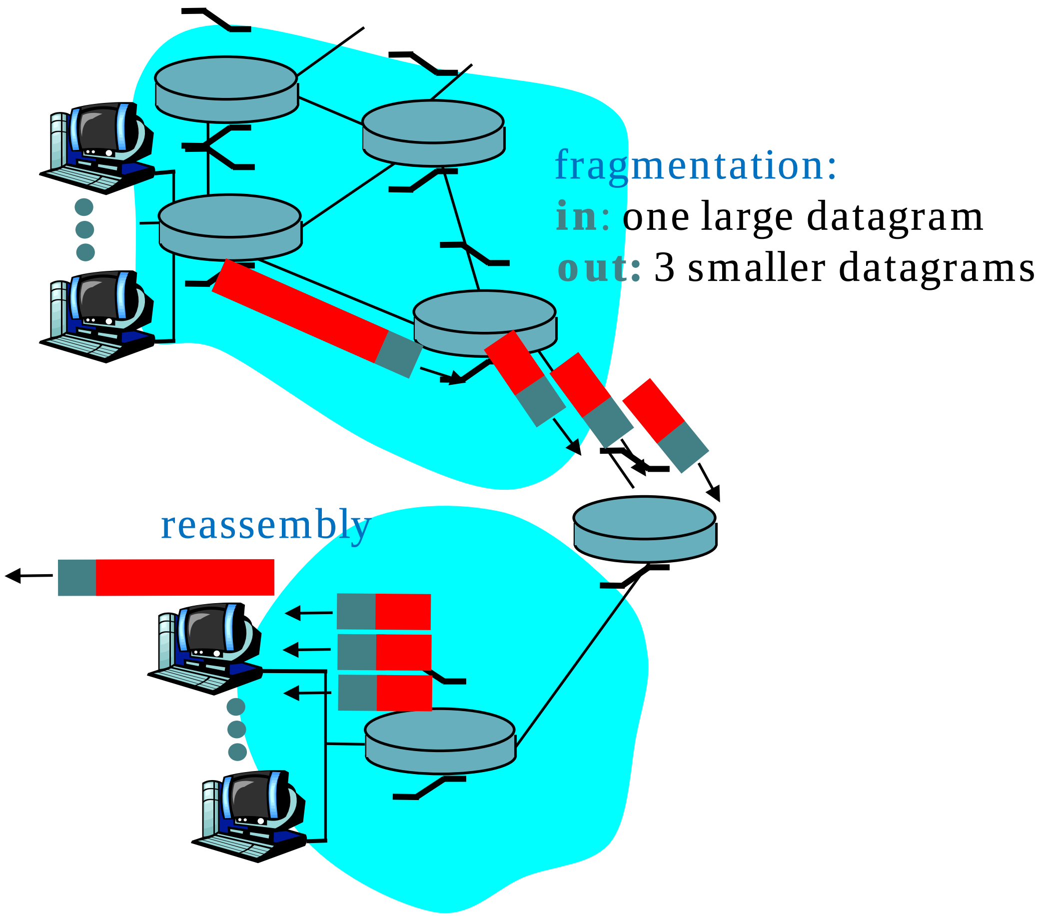
- look up network address of E in router's forwarding table
- E on *same* network as router's interface 223.1.2.9
  - router, E directly attached
- link layer sends datagram to 223.1.2.2 inside link-layer frame via interface 223.1.2.9
- datagram arrives at 223.1.2.2!!!

Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

IP Fragmentation & Reassembly

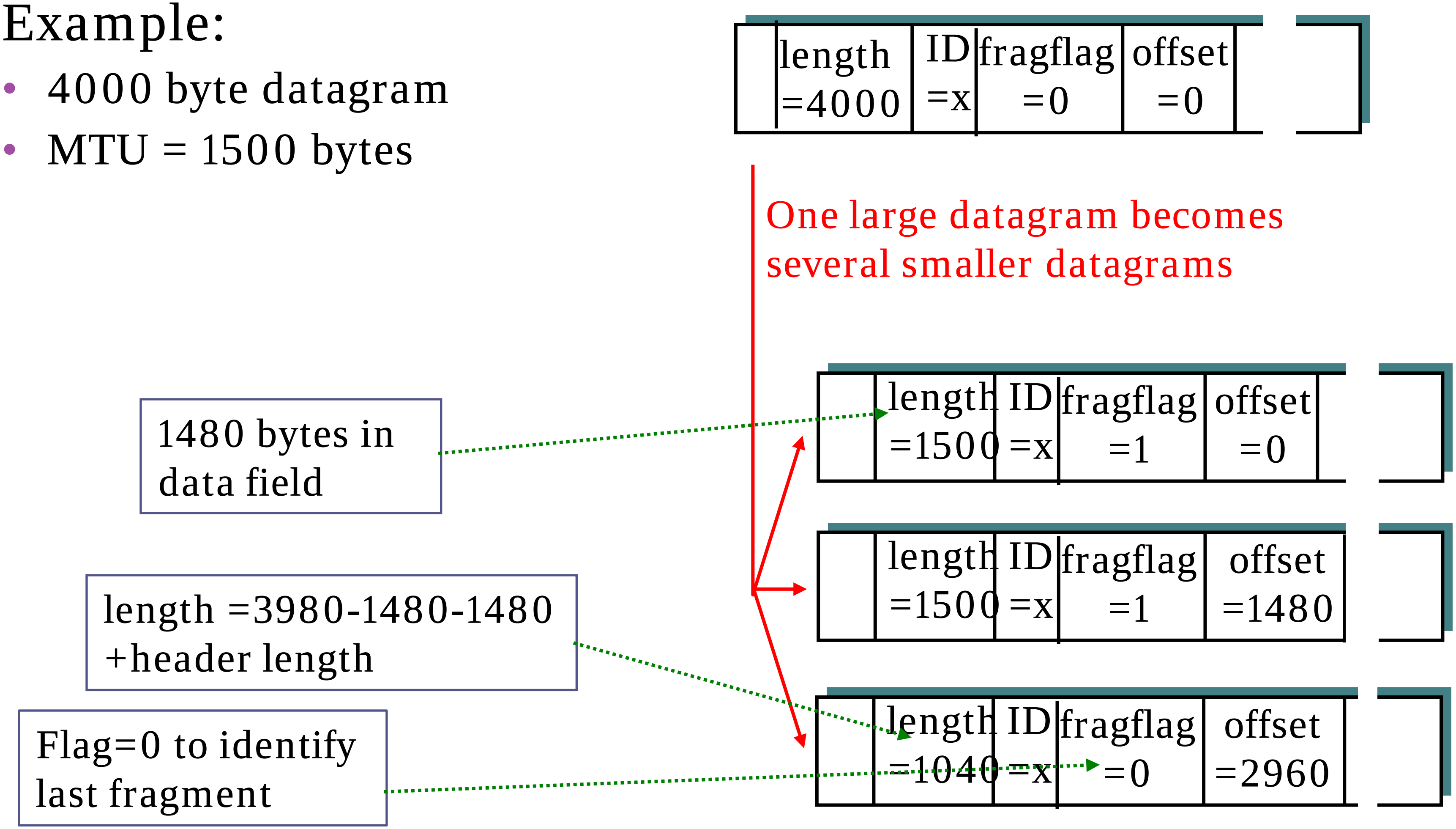
- network links have MTU (max.transfer unit) - largest possible link-level frame.
  - different link types, different MTUs
- large IP datagram divided ("fragmented") within net
  - one datagram becomes several datagrams
  - "reassembled" only at final destination
  - IP header bits used to identify, order related fragments



IP Fragmentation & Reassembly

Example:

- 4000 byte datagram
- MTU = 1500 bytes



Fragment Loss

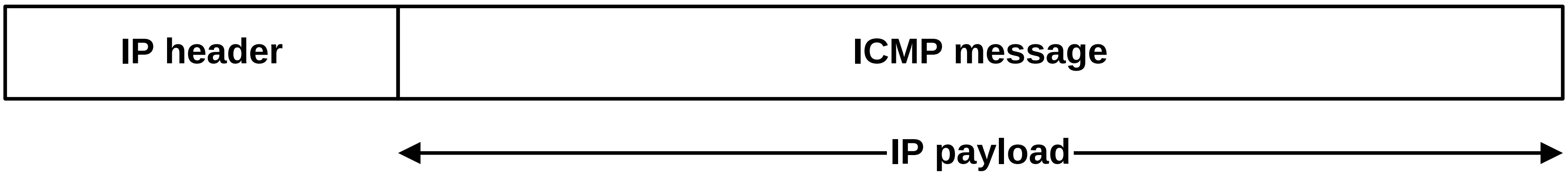
- IP does not guarantee datagram delivery
- Some fragments may be delayed or lost
- Datagrams with lost fragments cannot be reassembled
- If TCP is used in the transport layer the original datagram can be retransmitted
- Fragments may be saved temporarily.
- IP specifies a maximum time to hold fragments.
- After a timer expires, saved fragments are discarded.

Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

ICMP Overview

- The **Internet Control Message Protocol (ICMP)** is a helper protocol that supports IP with:
  - Error reporting (unreachable host, network, port, protocol)
  - Simple queries (echo request/reply, used by ping)
- ICMP message: type, code plus first 8 bytes of IP datagram causing error
- ICMP messages are encapsulated as IP datagrams:

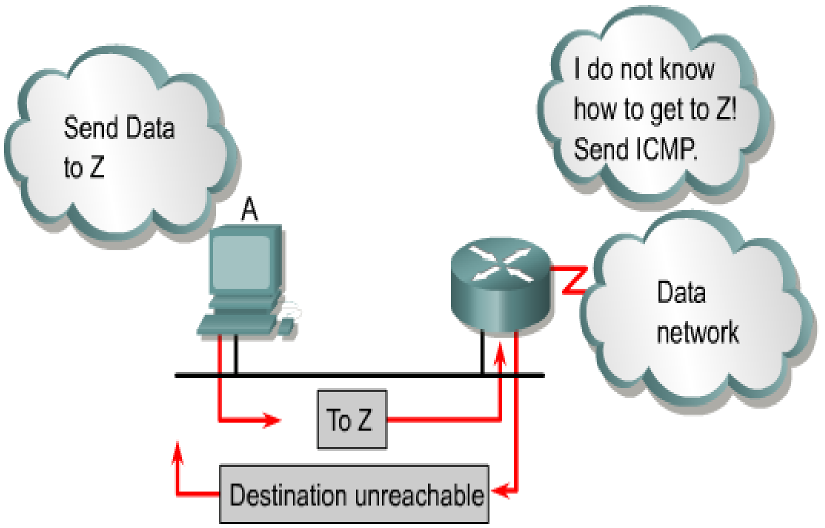


ICMP Message Types

Type	Code	description
0	0	echo reply (ping)
3	0	dest. network unreachable
3	1	dest host unreachable
3	2	dest protocol unreachable
3	3	dest port unreachable
3	6	dest network unknown
3	7	dest host unknown
4	0	source quench (congestion control - not used)
8	0	echo request (ping)
9	0	route advertisement
10	0	router discovery
11	0	TTL expired
12	0	bad IP header

Examples of errors/problems

- Unreachable Network
  - Sender sends datagram to a non-existent IP address
  - Destination device is disconnected from its network.
  - Router's connecting interface is down
  - Router does not have the information necessary to find the destination network.
- Port Unreachable
  - No process is waiting in destination port of destination host



An ICMP destination unreachable meassage is sent if:

- Host or port unreachable
- Network unreachable

ICMP use in Traceroute

- Command to determine the active route to a destination address
- How?
  - Send a UDP message to an unused port on the target host with ttl=1
  - When ttl becomes 0, router has to return an ICMP time exceed massage
    - It includes IP address & name of router
  - Traceroute set ttl = 2 and retransmits, this time go one more hop
  - ttl++ until UDP reach the destination
  - The target returns an ICMP service unreachable because there is no UDP port service

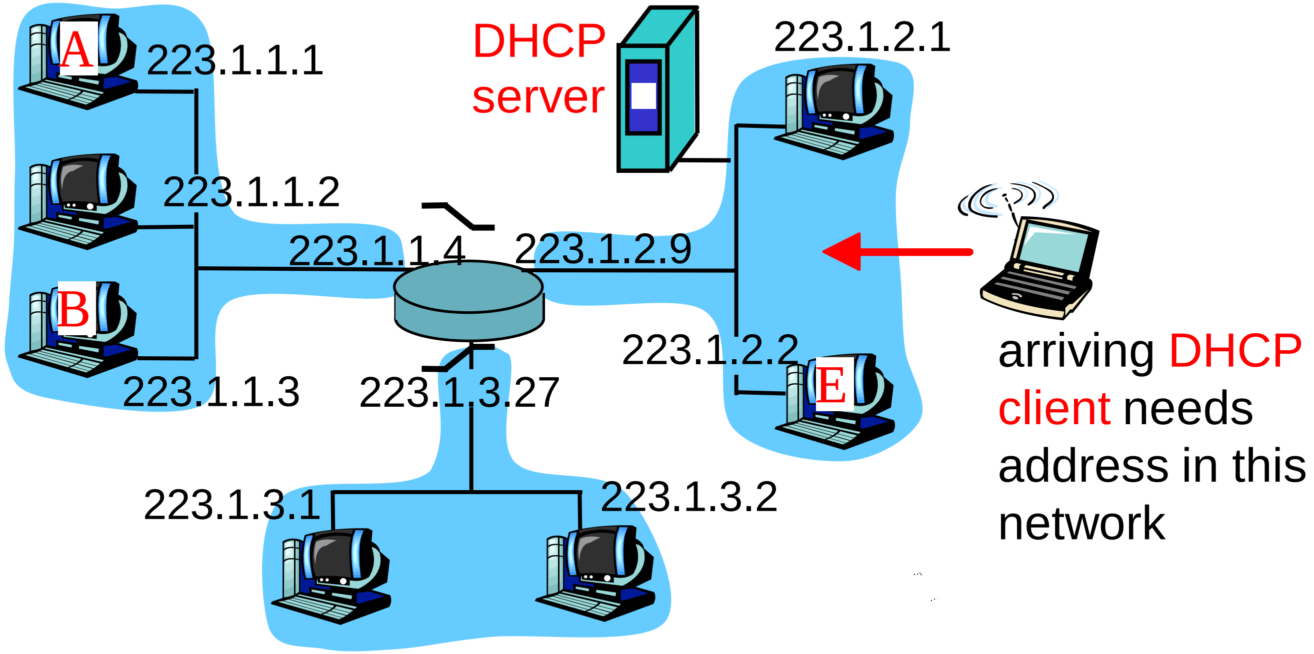
Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

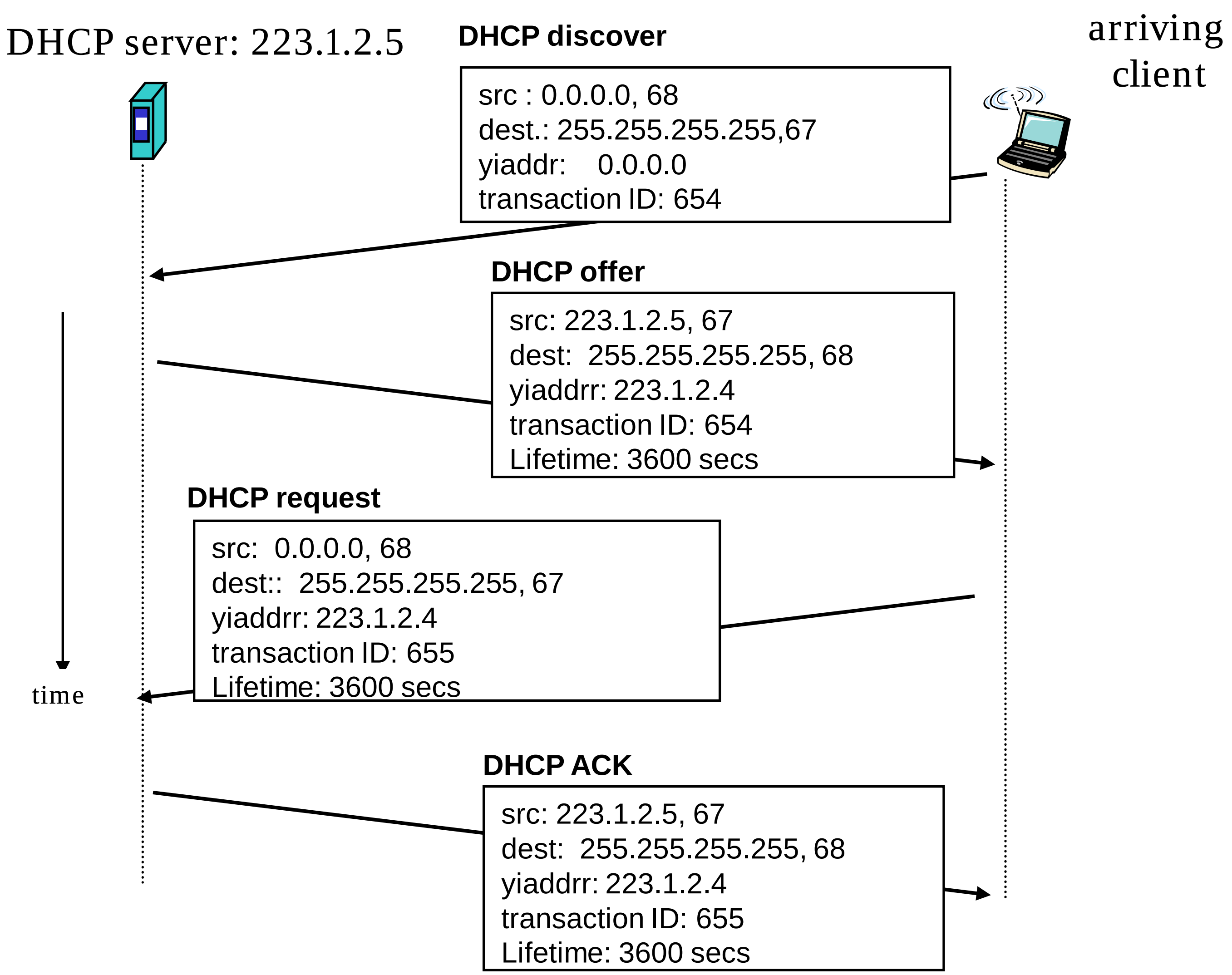
DHCP: Dynamic Host Configuration Protocol

- Allows 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)
  - Support for mobile users who want to join networks
- DHCP Overview
  - host broadcasts “DHCP discover” msg
  - DHCP server responds with “DHCP offer” msg
    - Several servers may respond
  - host requests IP address: “DHCP request” msg
  - DHCP server sends address: “DHCP ack” msg

DHCP client-server scenario



DHCP client-server scenario



Roadmap

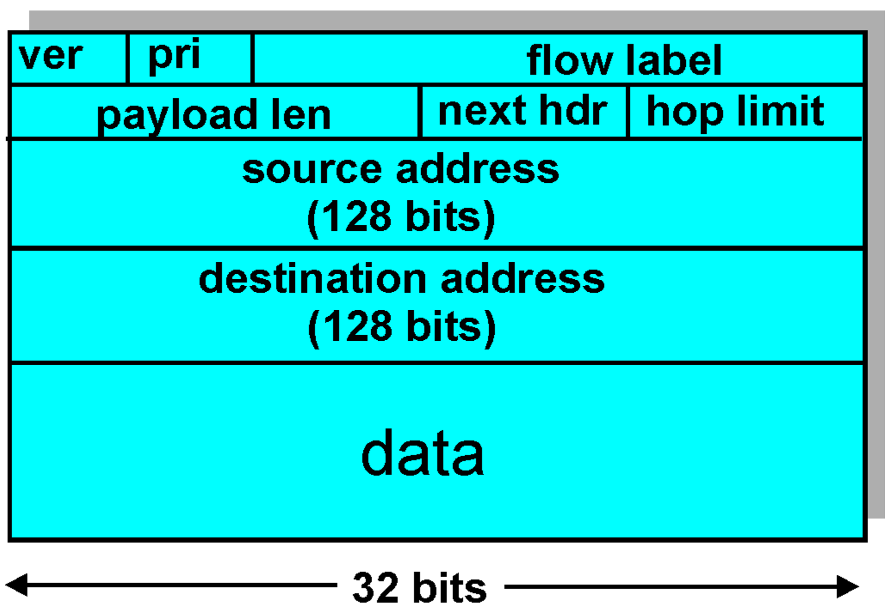
- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

IPv6

- Initial motivation
  - 32-bit address space soon to be completely allocated
  - 2^{32} = 4,294,967,296 (just over four billion)
  - Plus, some are reserved for special purposes
  - Great need for IPs(Computers, PDAs, routers, mobiles..)
- Additional motivation:
  - header format helps speed processing/forwarding
  - header changes to facilitate QoS
- IPv6 has 128-bit addresses ($2^{128} = 3.403 \times 10^{38}$)
  - every grain of sand on the planet can be IP-addressable!
- Short-term solutions: limping along with IPv4
  - Network address translation (NAT)
  - Dynamically-assigned addresses (DHCP)
- IPv6 datagram format:
  - fixed-length 40 byte header
  - no fragmentation allowed

IPv6 Header

- *Priority:* identify priority among datagrams in flow or give priority to datagrams from certain apps (ICMP)
- *Flow Label:* identify datagrams in same “flow.”
  - Special handling for some flows (e.g. real time app.)
  - Flows of high priority users (paying for better service)
- *Next header:* identify upper layer protocol for data (TCP/UDP)



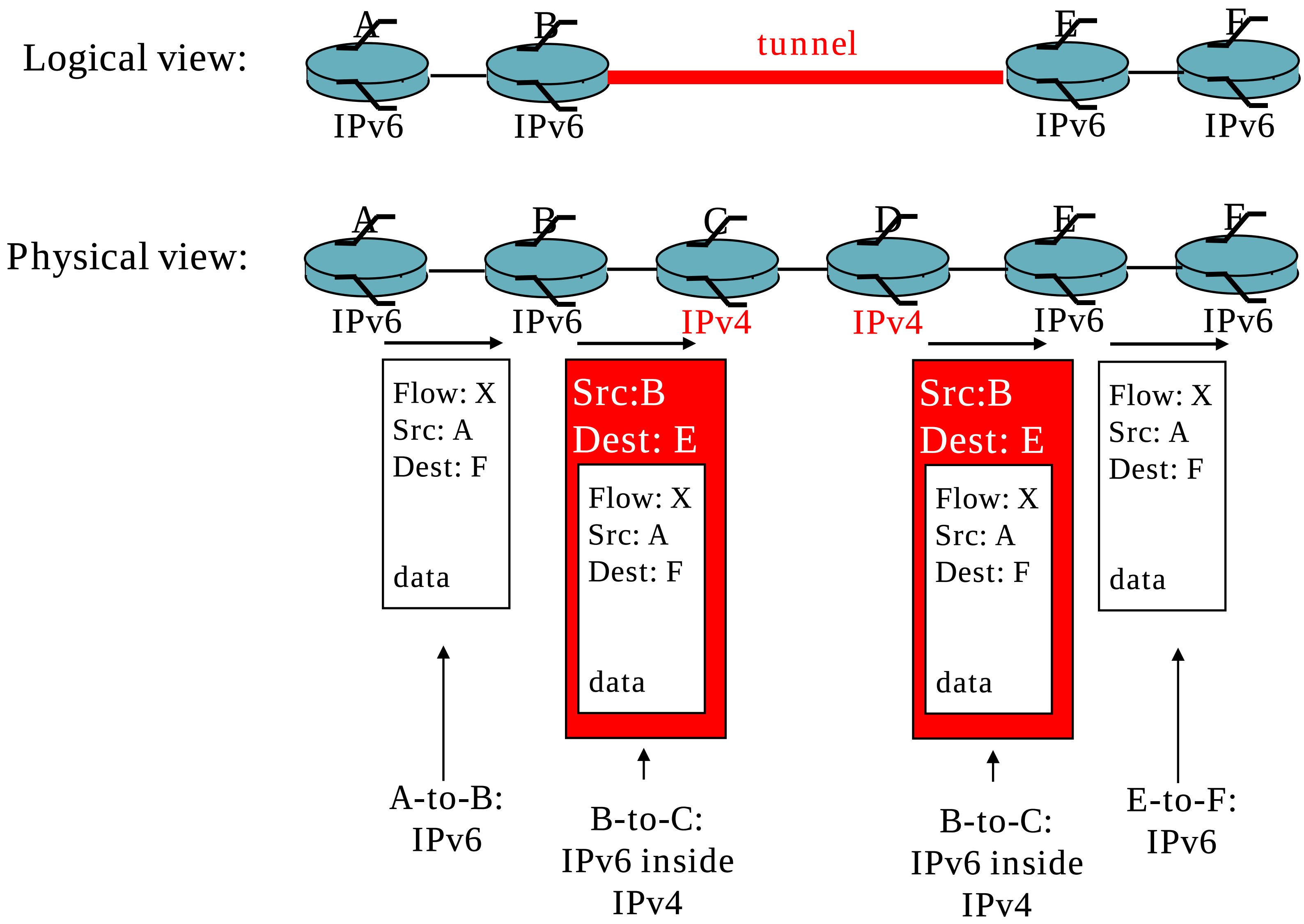
Other Changes from IPv4

- *Checksum:* removed entirely to reduce processing time at each hop (after change of TTL)
- *Options:* allowed, but outside of header, pointed to by "Next Header" field
- *ICMPv6:* new version of ICMP
  - additional message types, e.g. "Packet Too Big"
  - multicast group management functions

Transition From IPv4 To IPv6

- Not all routers can be upgraded simultaneous
  - no “flag days” & huge size of Internet
  - How will the network operate with mixed IPv4 and IPv6 routers?
- Dual Stack approach
  - IPv6 nodes also have a complete IPv4 implementation
  - Nodes must have both IPv6 & IPv4 addresses
  - Must be able to determine if other nodes are IPv6 capable
- *Tunneling: entire* IPv6 packet carried as payload in IPv4 datagram among IPv4 routers

Tunneling



Logical view:
A
B
E
F
tunnel
IPv6
IPv6
IPv6
IPv6
Physical view:
A
B
C
D
E
F
IPv6
IPv6
IPv4
IPv4
IPv6
IPv6
Flow: X
Src: A
Dest: F
data
Src:B
Dest: E
Flow: X
Src: A
Dest: F
data
Src:B
Dest: E
Flow: X
Src: A
Dest: F
data
Flow: X
Src: A
Dest: F
data
A-to-B:
IPv6
B-to-C:
IPv6 inside
IPv4
B-to-C:
IPv6 inside
IPv4
E-to-F:
IPv6

Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

Routing in the Internet

- Internet is organized as a set of independent Autonomous Systems (AS)
  - AS: collection of networks under single administration
- The AS appears to the outside world to have coherent routing plan and presents unique view what destination are reachable through it
- Routers in same AS run same routing protocol
  - "intra-AS" routing or Interior Gateway Protocols (IGP)
  - Different AS can have different intra-AS routing protocols
  - RIP, OSPF
- A separate protocol is used to transfer information between AS
  - "inter-AS" routing or Exterior Routing Protocol (EGP)
  - BGP

Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

RIP - Routing Information Protocol

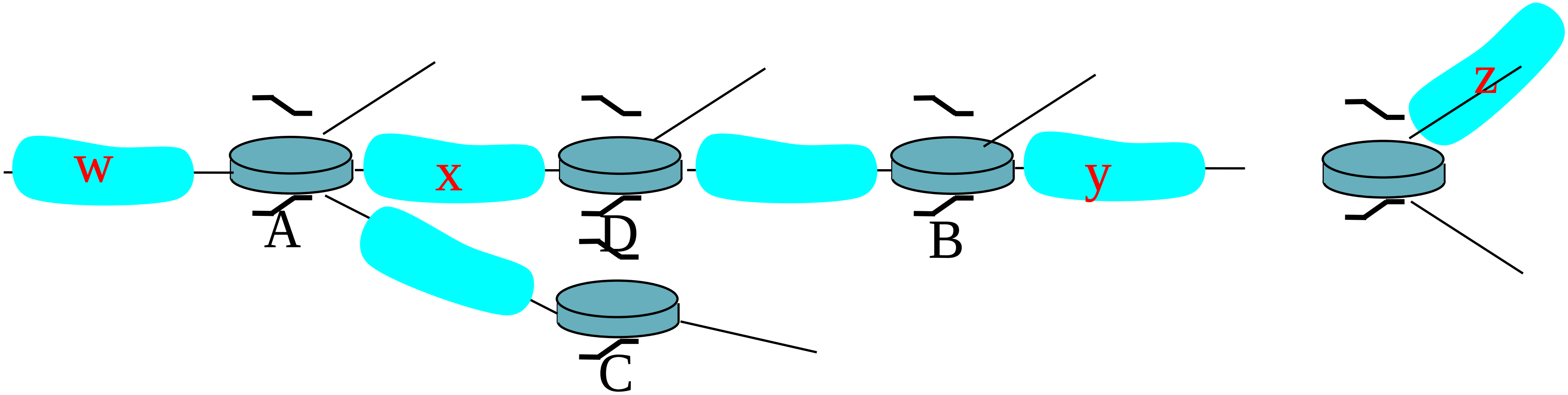
- A simple intra-AS routing protocol
- Uses Distance Vector Algorithm
  - The information is exchanged only between adjacent routers
  - Fixed (hop) metrics
  - "count to infinity" problem
- Each router advertises its distance vector every 30 seconds (or whenever its routing table changes) to all of its neighbors
  - Each advertisement: list of up to 25 destination subnets within AS
- RIP always uses 1 as link metric
- Maximum cost of path is 15, with "16" equal to "∞"
- Routes are declared dead (set to 16) after 3 minutes if no advertisement heard from neighbor

Routing with RIP

- **Initialization:** Send a request packet on all interfaces:
    - RIPv1 uses broadcast if possible,
    - RIPv2 uses multicast address 224.0.0.9, if possible

  requesting routing tables from neighboring routers
- **Request received**: Routers that receive above request send their entire routing table
- **Response received**: Update the routing table
- **Regular routing updates**: Every 30 seconds, send all or part of the routing tables to every neighbor in an response message
- **Triggered Updates:** Whenever the metric for a route change, send entire routing table.

RIP Example



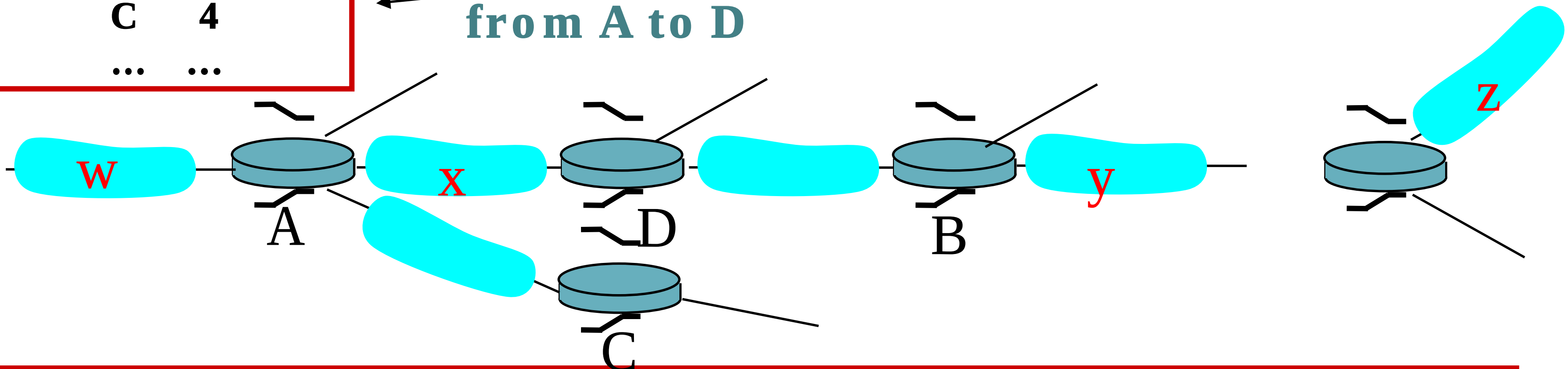
Destination Network	Next Router	Num. of hops to dest.
w	A	2
y	B	2
z	B	7
x	--	1
….	….	….

Routing/Forwarding table in D

RIP Example (Cont)

Dest	Next	hops
w	-	1
x	-	1
z	C	4
….	…	…

Advertisement from A to D

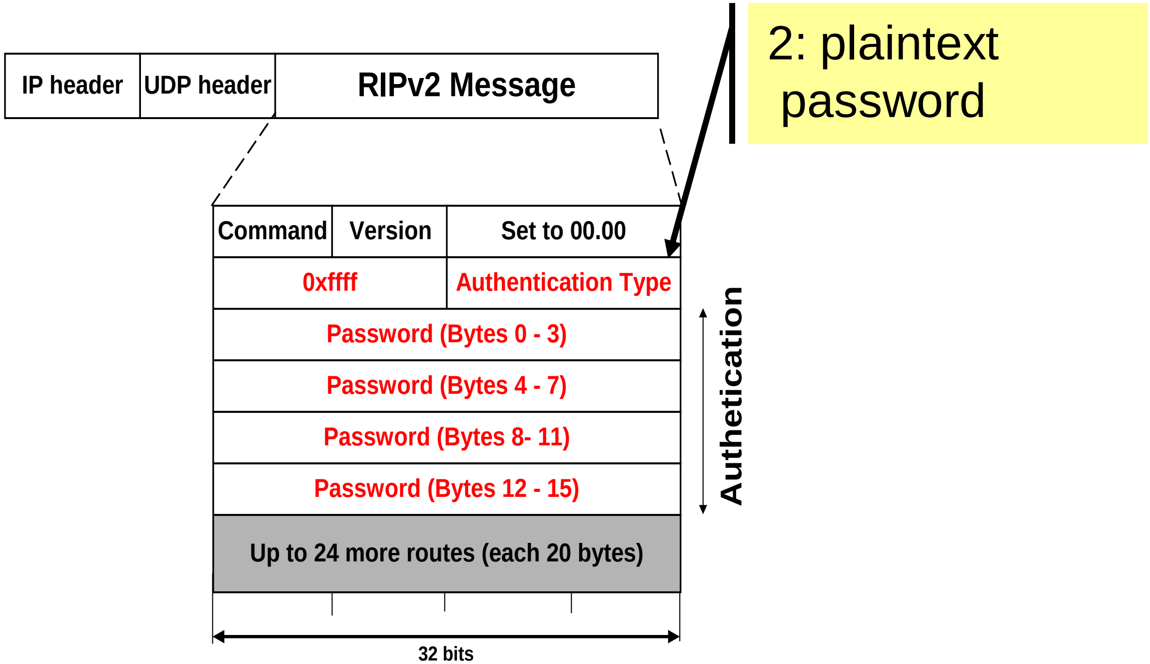


Destination Network	Next Router	Num. of hops to dest.
w	A	2
y	B	2
z	~~B~~ A	~~7~~ 5
x	--	1
….	….	….

Routing/Forwarding table in D

RIP Security

- Issue: Sending bogus routing updates to a router
- RIPv1: No protection
- RIPv2: Simple authentication scheme
  - Simple password
  - MD5



Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

OSPF-Open Shortest Path First

- Intra-AS routing protocol
- Uses Link State Algorithm
  - LS packet dissemination
  - topology map at each node
  - route computation using Dijkstra's algorithm
- Every OSPF router sends periodically 'hello' packets
  - Hello packets used to determine if neighbor is up
  - Hello packets are small easy to process

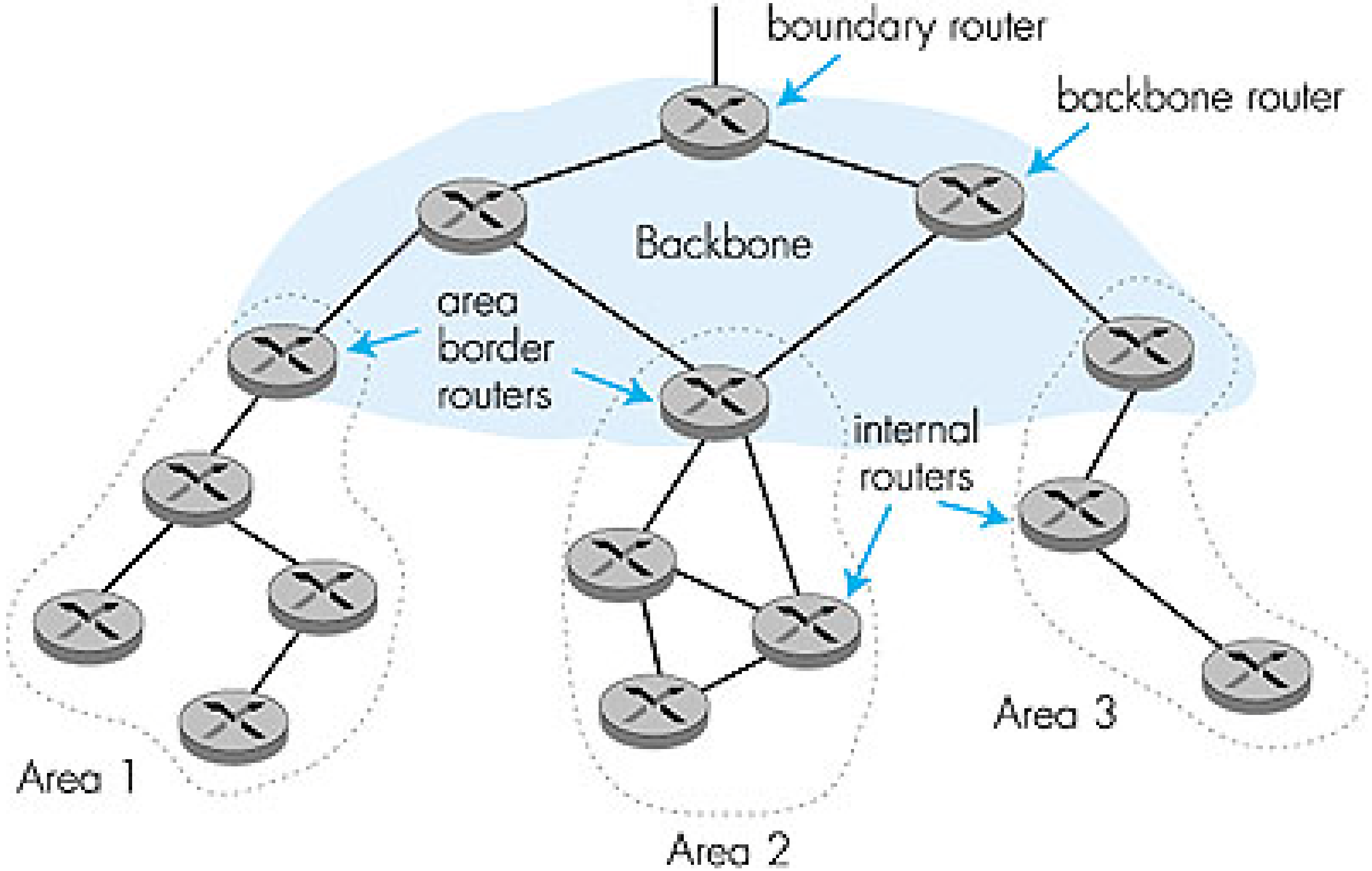
OSPF Operation

- Once an adjacency is established, trade information with your neighbor
- Topology information is packaged in a "link state announcement“
  - OSPF advertisement carries one entry per neighbor router
- LSA-Updates are distributed to all other routers via Reliable Flooding
  - If a received LSA does not contain new information, the router will not flood the packet
  - Exception: Infrequently (every 30 minutes), a router will flood LSAs even if there are not new changes.

OSPF “advanced” features (not in RIP)

- Provides authentication of routing messages
- Enables load balancing by allowing traffic to be split evenly across routes with equal cost
- Type-of-Service routing allows to setup different routes dependent on the TOS field
- Integrated uni- and multicast support:
  - Multicast OSPF (MOSPF) uses same topology data base as OSPF
- Allows hierarchical routing

Hierarchical OSPF



Hierarchical OSPF

- two-level hierarchy: local area, backbone.
  - Link-state advertisements only in area
  - Each nodes has detailed area topology; only know direction (shortest path) to nets in other areas.
- Backbone Area:
  - Role: route traffic between the other areas in the AS
  - Contains all area border routers in the AS and may contain non border routers as well.
- *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.

Roadmap

- IP: The Internet Protocol
  - IPv4 Addressing
  - Datagram Format
  - Transporting a datagram from source to destination
  - IP Fragmentation & Reassembly
  - ICMP
  - DHCP
  - IPv6
- Routing in the Internet
  - RIP
  - OSPF
  - BGP

BGP-Border Gateway Protocol

- Inter-AS routing protocol
- Uses TCP to send routing messages
- Is neither a link state, nor a distance vector protocol. Routing messages in BGP contain complete routes.
- Network administrators can specify routing policies

BGP message types

- Open
  - Sent after the TCP connection is established
  - Includes
    - hold time - the maximum time between consecutive keep alive messages
    - router ID
      - Router is identified and authenticated
- Keep alive
  - Sent periodically (I am alive but have nothing new to send!)
- Update
  - Contains information about one path
- Notification
  - Sent in case of error condition

BGP Speakers

- Router running BGP is called BGP speaker
- BGP speakers establish TCP connection to exchange routing information in a BGP session
  - If the two BGP speakers belong to different AS they are running external BGP (eBGP)
    - They have to be directly connected
  - If the two speakers belong to the same AS they are running internal BGP (iBGP)
    - They do not have to be directly connected
    - IGP protocol must be in place to assure connectivity between BGP internal neighbours
- At startup BGP speakers exchange full routing tables, then only changes are advertised
- When AS2 advertises a prefix to AS1:
  - AS2 *promises* it will forward datagrams towards that prefix.
  - AS2 can aggregate prefixes in its advertisement

Path attributes & BGP routes

- Advertised prefix includes BGP attributes.
- Two important attributes:
  - AS-PATH: contains ASs through which prefix advertisement has passed: e.g, AS 67, AS 17
  - NEXT-HOP: indicates specific internal-AS router to next-hop AS. (may be multiple links from current AS to next-hop-AS)
- When gateway router receives route advertisement, uses import policy to accept/decline.

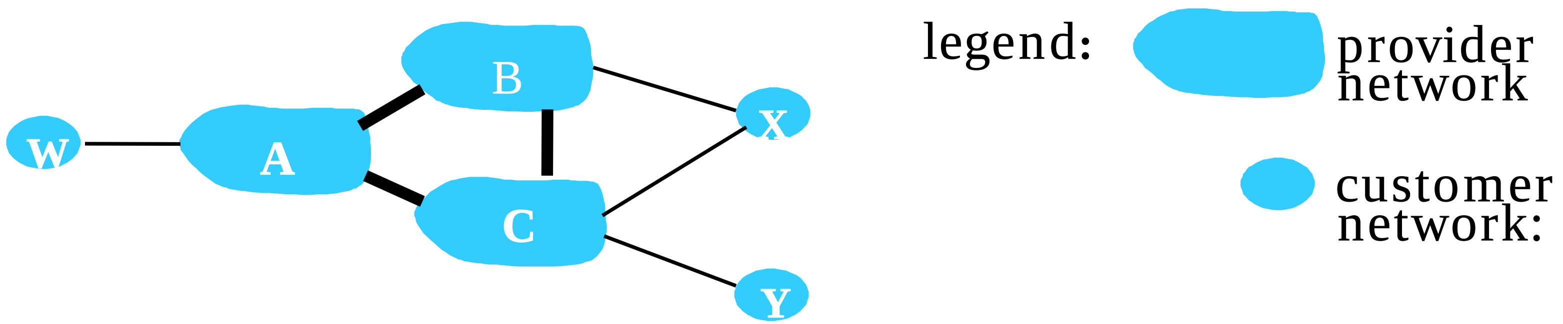
BGP route selection

- Router may learn about more than 1 route to some prefix. Router must select route.
- Elimination rules:
  1. local preference value attribute: policy decision
  2. shortest AS-PATH
  3. closest NEXT-HOP router: hot potato routing
  4. additional criteria

BGP Policy Routing

- BGP's goal is to find any loop free path (not an optimal one). Since the internals of the AS are never revealed, finding an optimal path is not feasible.
- For each AS, BGP distinguishes:
  - **local traffic** = traffic with source or destination in AS
  - **transit traffic** = traffic that passes through the AS
  - **Stub AS** = has connection to only one AS, only carry local traffic
  - **Multihomed AS** = has connection to >1 AS, but does not carry transit traffic
  - **Transit AS** = has connection to >1 AS and carries transit traffic

BGP Routing Policy Example



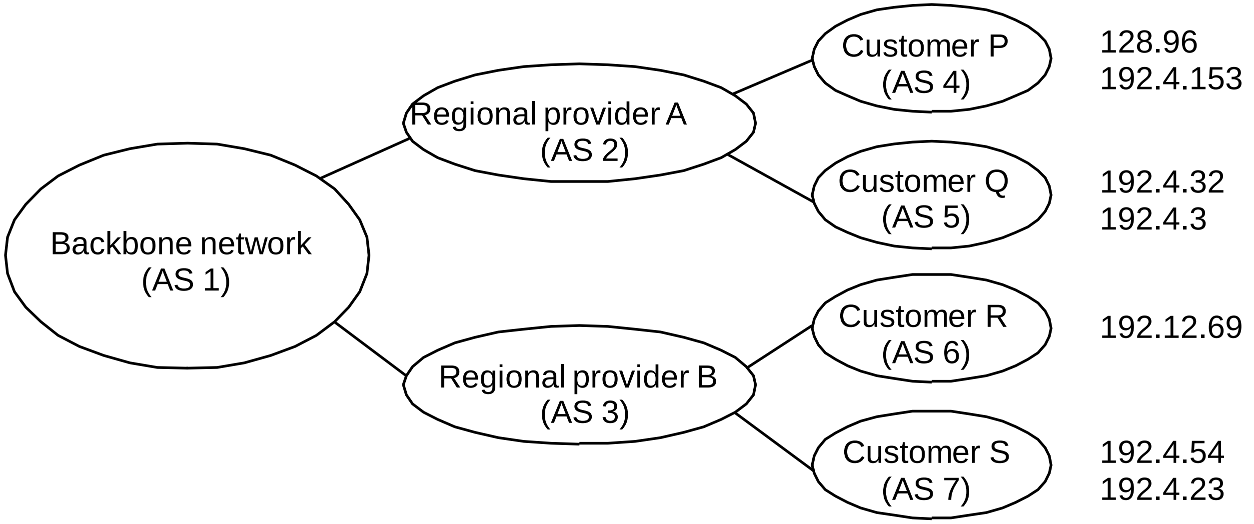
- A advertises path AW to B
- B advertises path BAW to X
- Should B advertise path BAW to C?
- No! B gets no "revenue" for routing CBAW since neither W nor C are B's customers
  - B wants to force C to route to w via A
  - B wants to route *only* to/from its customers!

BGP – IGP Interaction

- AS has to be consistent about the routes it advertises
  - If eBGP advertises a route before all routers in AS have learned about it, AS might receive traffic that some routers cannot route
- BGP waits until IGP has propagated routing information across AS (Synchronization)

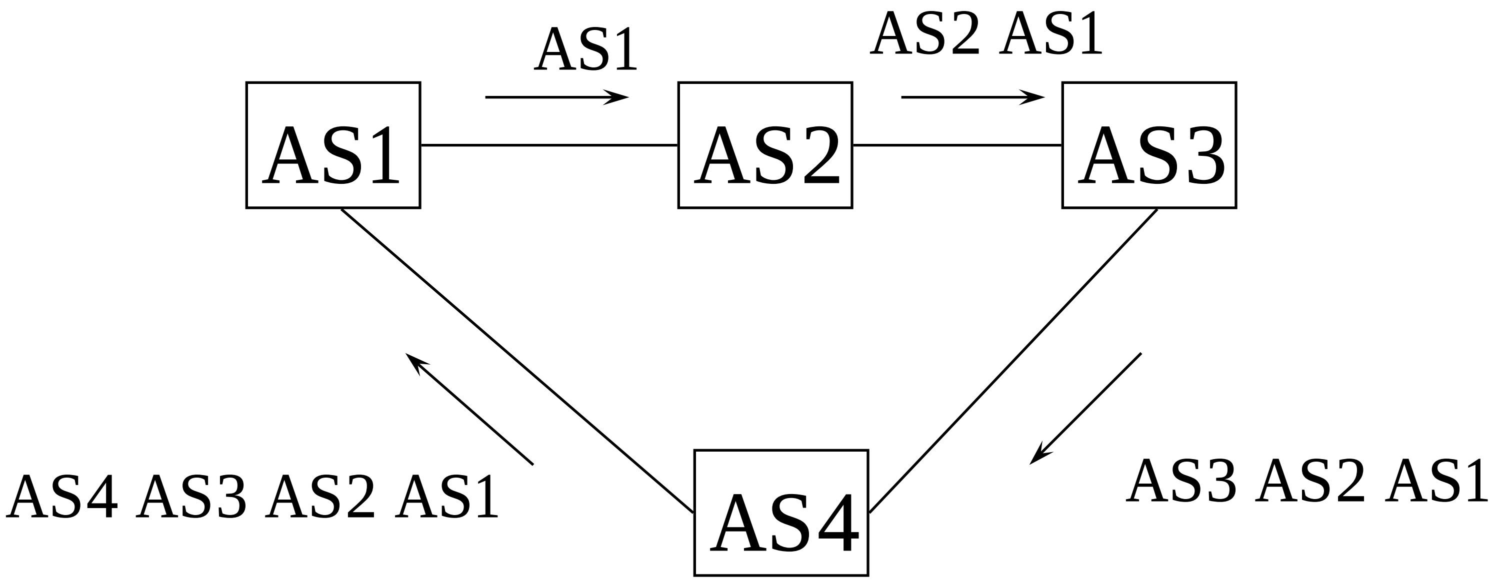
BGP Example

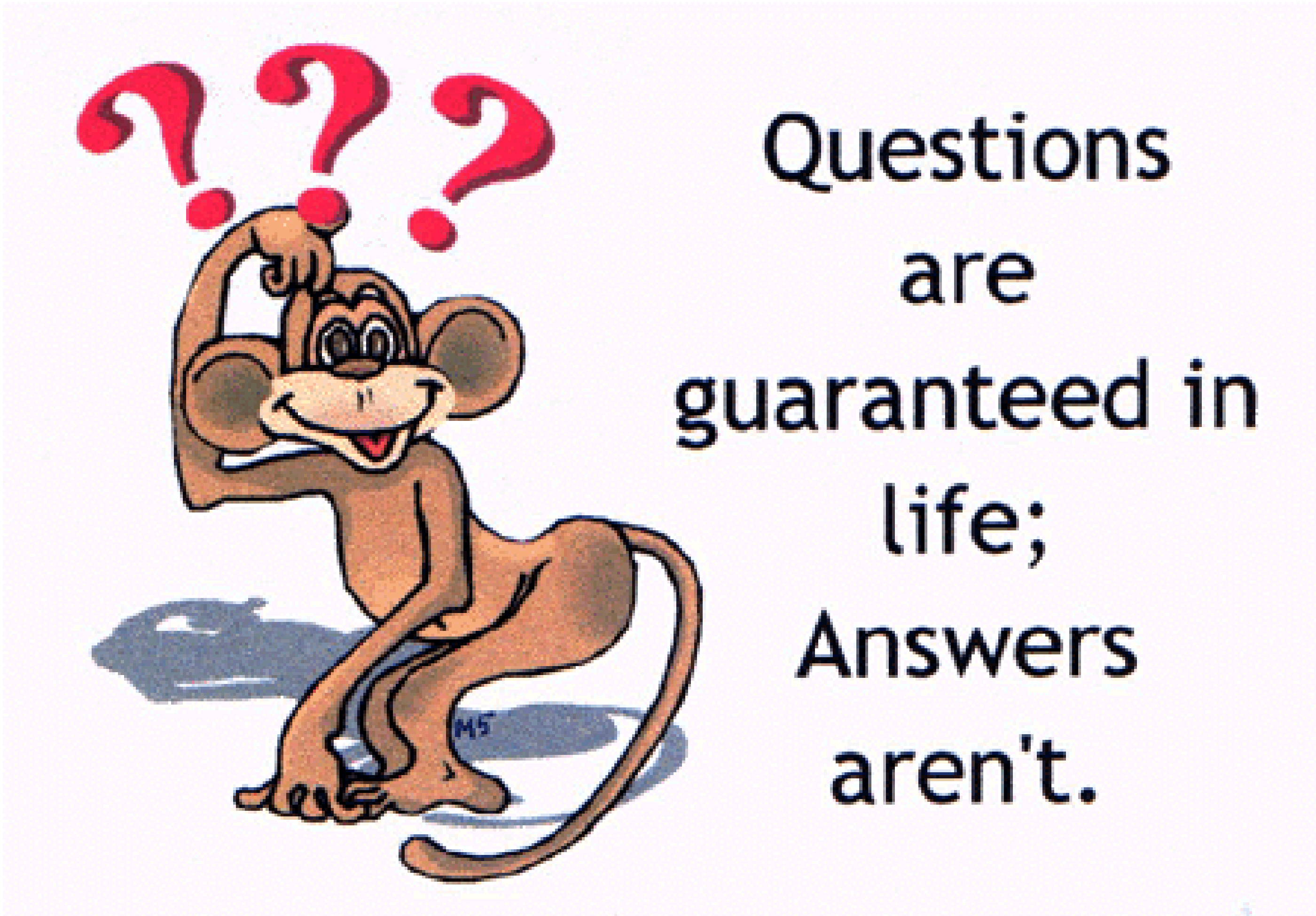
- Speaker for AS2 advertises reachability to P and Q
  - network 128.96, 192.4.153, 192.4.32, and 192.4.3, can be reached directly from AS2
- Speaker for backbone advertises
  - networks 128.96, 192.4.153, 192.4.32, and 192.4.3 can be reached along the path (AS1, AS2).



Loop Avoidance

- Routing information sent from AS1 to AS2, to AS3, to AS4 and back to AS1 will be ignored by AS1




Questions
are
guaranteed in
life;
Answers
aren't.