CSCI 251
Systems and Networks

Network Layer

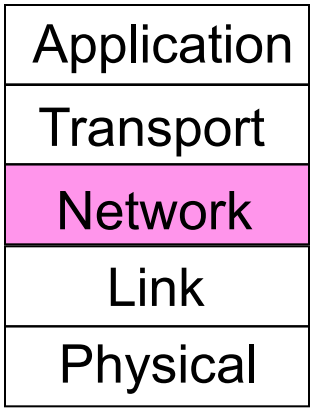
Humayun Kabir
Professor, CS, Vancouver Island University, BC, Canada

Outline

- Store and Forward Packet (Datagram) Switching
- Routers
- Routing Algorithms
  - Shortest Path Routing
  - Distance Vector Routing
  - Link State Routing
- Internet Protocol (IP)
  - IP Packet
  - IP Address, Subnet, and CIDR
  - Network Address Translation (NAT)
  - Internet Control Message Protocol (ICMP)
  - Address Resolution Protocol (ARP)
  - Dynamic Host Configuration Protocol (DHCP)

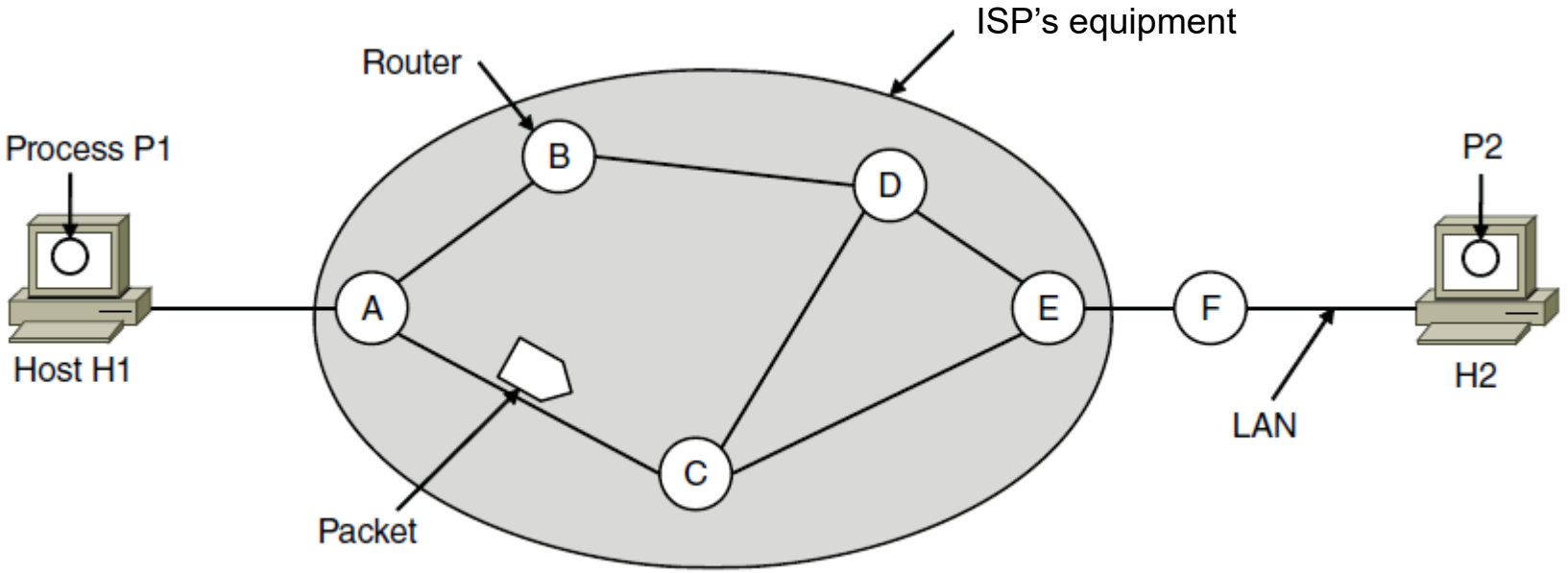
The Network Layer

Responsible for routing datagrams (packets) from source to destination network (eventually source to destination nodes) over multiple hops (networks).



Store-and-Forward Packet Switching

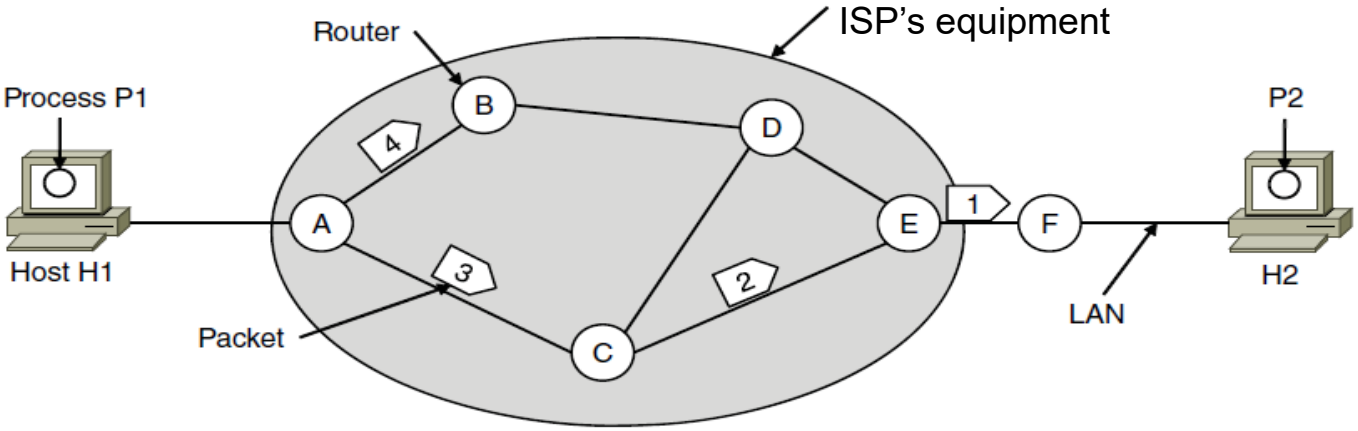
Hosts send packets into the network; packets are forwarded by routers



Connectionless Service – Datagrams

Packet is forwarded using destination address inside it

- Different packets may take different paths



A's table (initially)

Dest.	Line
A	
B	B
C	C
D	B
E	C
F	C

A's table (later)

A	
B	B
C	C
D	B
E	B
F	B

C's Table

A	A
B	A
C	
D	E
E	E
F	E

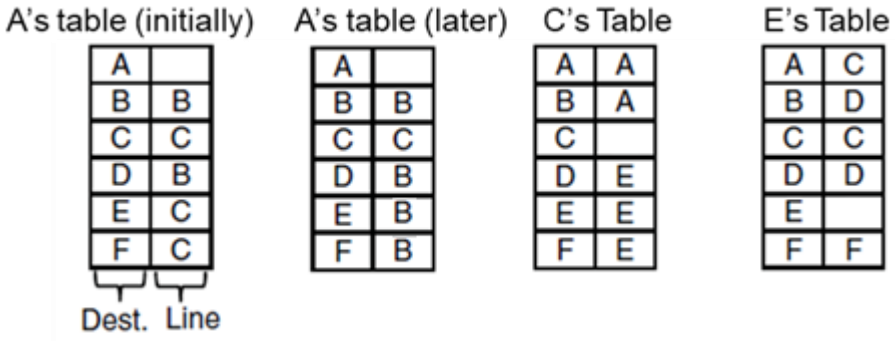
E's Table

A	C
B	D
C	C
D	D
E	
F	F

Routing and Forwarding

Routing is the process of discovering network paths

- Model the network as a graph of nodes and links
- Decide what to optimize (e.g., fairness vs efficiency)
- Build routing tables in each router
- Update routes for changes in topology (e.g., failures)



Dest.	Line
A	
B	B
C	C
D	B
E	C
F	C

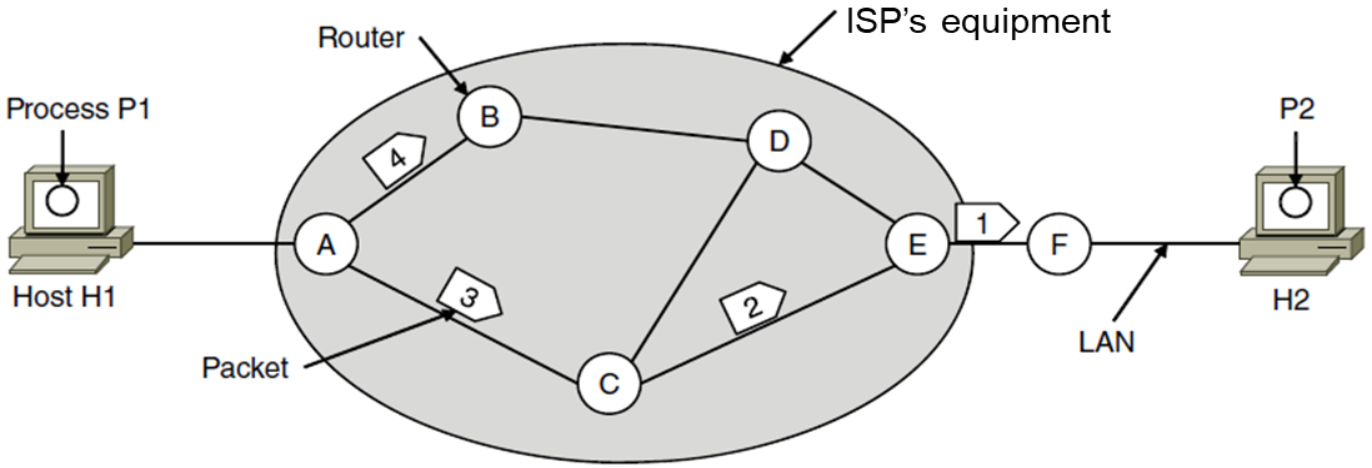
Dest.	Line
A	
B	B
C	C
D	B
E	B
F	B

Dest.	Line
A	A
B	A
C	
D	E
E	E
F	E

Dest.	Line
A	C
B	D
C	C
D	D
E	
F	F

Routing and Forwarding

Forwarding is the sending of packets along a path using the routing table



A's table (initially)

Dest.	Line
A	
B	B
C	C
D	B
E	C
F	C

A's table (later)

A	
B	B
C	C
D	B
E	B
F	B

C's Table

A	A
B	A
C	
D	E
E	E
F	E

E's Table

A	C
B	D
C	C
D	D
E	
F	F

Flooding

A simple method to send a packet to all network nodes

Each node floods a new packet received on an incoming link by sending it out all of the other links

Nodes need to keep track of flooded packets to stop the explosion; even using a hop limit can blow up exponentially

Distance Vector Routing

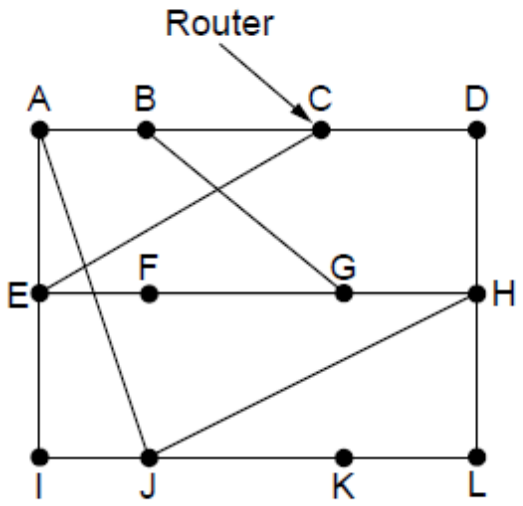
Distance vector is a distributed routing algorithm

- Shortest path computation is split across nodes
- Often used in Internet (RIP)

Algorithm:

- Each node knows distance of links to its neighbors
- Each node advertises vector of lowest known distances to all neighbors
- Each node uses received vectors to update its own
- Repeat periodically

Distance Vector Routing



Network

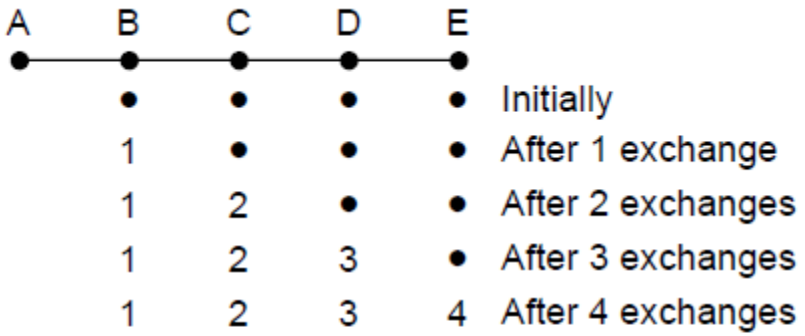
To	A	I	H	K	New estimated delay from J	Line
A	0	24	20	21	8	A
B	12	36	31	28	20	A
C	25	18	19	36	28	I
D	40	27	8	24	20	H
E	14	7	30	22	17	I
F	23	20	19	40	30	I
G	18	31	6	31	18	H
H	17	20	0	19	12	H
I	21	0	14	22	10	I
J	9	11	7	10	0	–
K	24	22	22	0	6	K
L	29	33	9	9	15	K
	JA delay is 8	JI delay is 10	JH delay is 12	JK delay is 6		

New vector for J

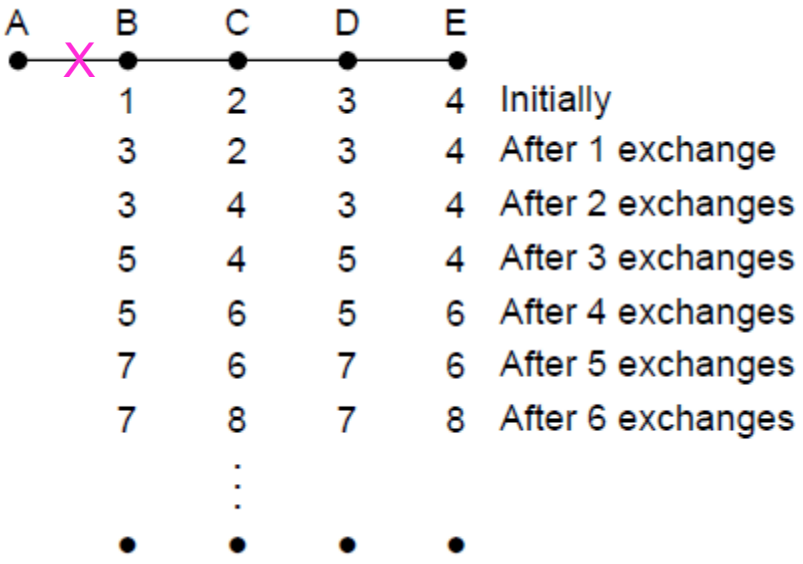
Vectors received at J from Neighbors A, I, H and K

The Count-to-Infinity Problem

Failures can cause DV to “count to infinity” while seeking a path to an unreachable node



Good news of a path to *A* spreads quickly



Bad news of no path to *A* is learned slowly

Shortest Path Algorithm

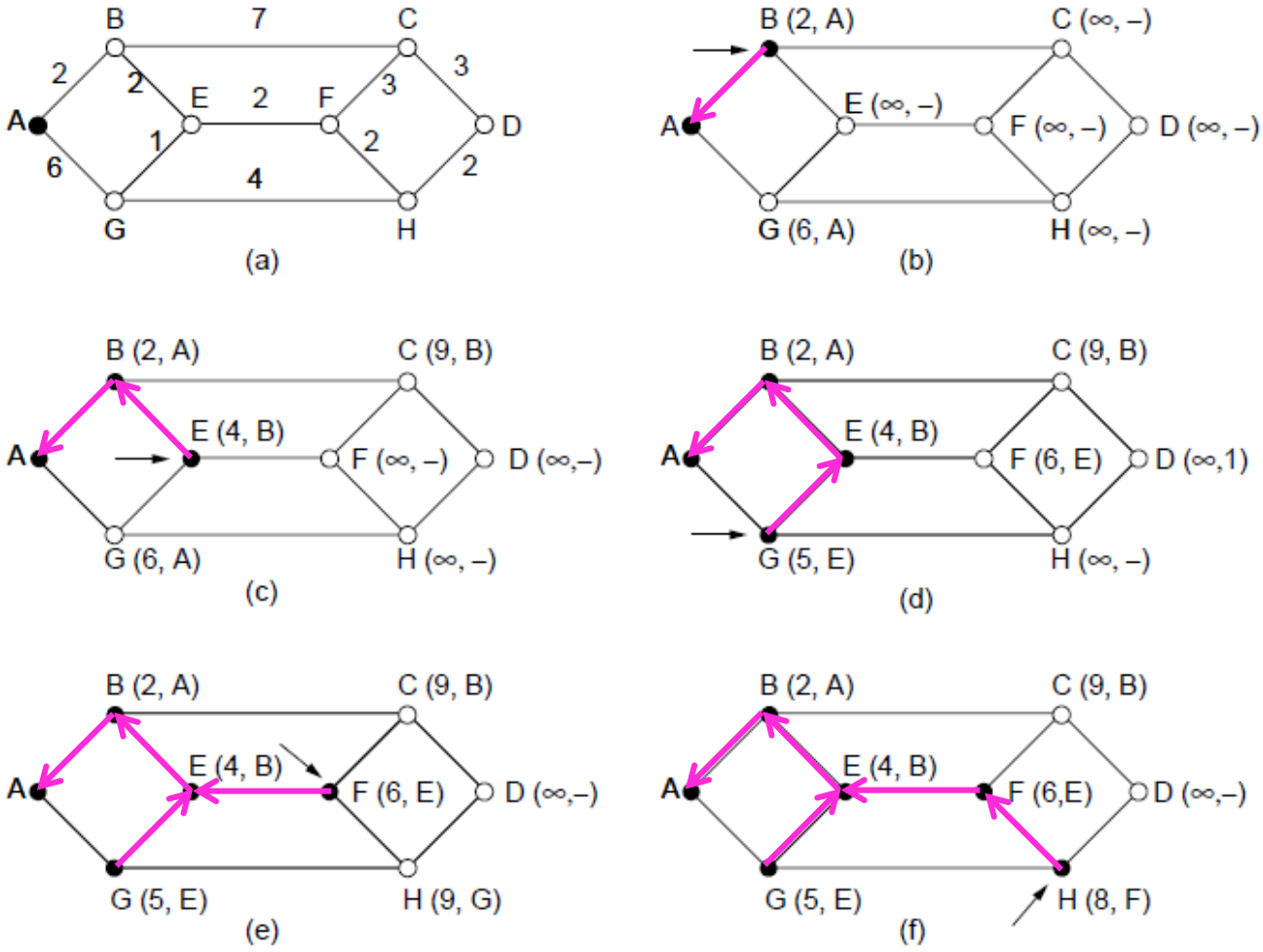
Dijkstra's algorithm computes a sink tree on the graph:

- Each link is assigned a non-negative weight/distance
- Shortest path is the one with lowest total weight
- Using weights of 1 gives paths with fewest hops

Algorithm:

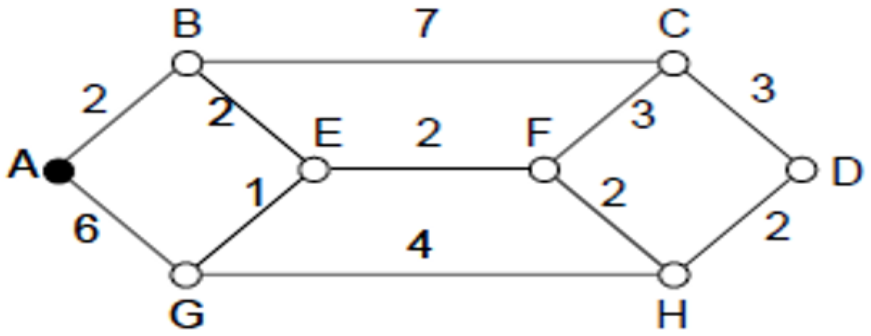
- Start with sink, set distance at other nodes to infinity
- Relax distance to other nodes
- Pick the lowest distance node, add it to sink tree
- Repeat until all nodes are in the sink tree

Shortest Path Algorithm



A network and first five steps in computing the shortest paths from A to D. Pink arrows show the sink tree so far.

Shortest Path Algorithm



From A

A	B	C	D	E	F	G	H
0, -	2,A	∞, -	∞, -	∞, -	∞, -	6, A	∞, -

From B

A	B	C	D	E	F	G	H
0, -	2,A	9, B	∞, -	4, B	∞, -	6, A	∞, -

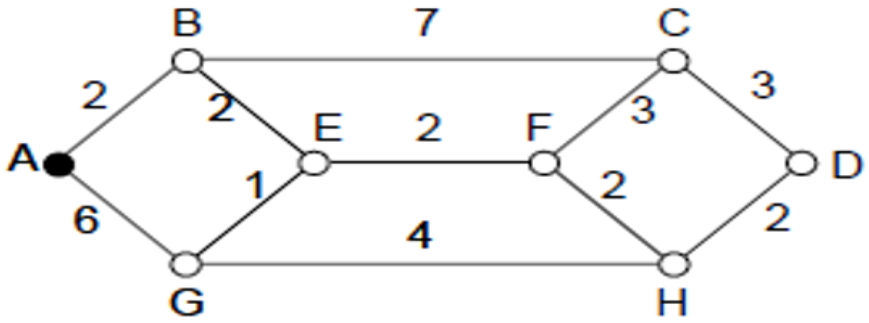
From E

A	B	C	D	E	F	G	H
0, -	2,A	9, B	∞, -	4, B	6, E	5, E	∞, -

From G

A	B	C	D	E	F	G	H
0, -	2,A	9, B	∞, -	4, B	6, E	5, E	9, G

Shortest Path Algorithm



From F

A	B	C	D	E	F	G	H
0, -	2,A	9, B	∞, -	4, B	6, E	5, E	8, F

From H

A	B	C	D	E	F	G	H
0, -	2,A	9, B	10,H	4, B	6, E	5, E	8, F

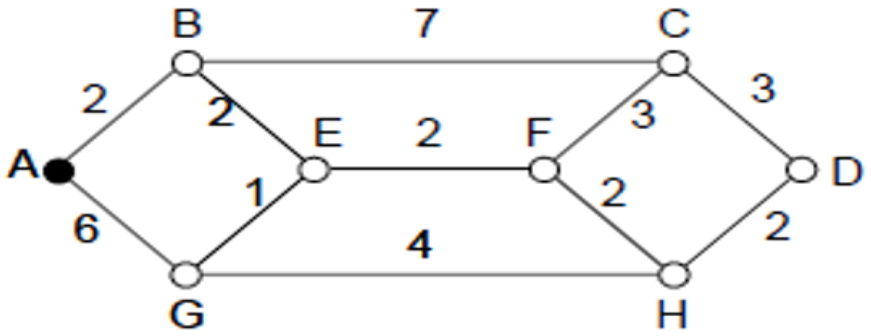
From C

A	B	C	D	E	F	G	H
0, -	2,A	9, B	10,H	4, B	6, E	5, E	8, F

From D

A	B	C	D	E	F	G	H
0, -	2,A	9, B	10,H	4, B	6, E	5, E	8, F

Routing Table from Shortest Path Algorithm



A's Routing Table

Destination	Cost	Next Hop
A	0	-
B	2	B
C	9	B
D	10	B
E	4	B
F	6	B
G	5	B
H	8	B

Link State Routing

Link state is an alternative to distance vector

- More computation but simpler dynamics
- Widely used in the Internet (OSPF, ISIS)

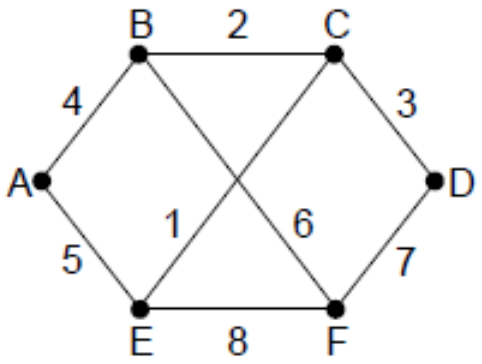
Algorithm:

- Each node floods information about its neighbors in **LSP**s (Link State Packets); all nodes learn the full network graph
- Each node runs Dijkstra's algorithm to compute the path to take for each destination

Link State Routing – LSPs

LSP (Link State Packet) for a node lists neighbors and weights of links to reach them

- LSPs are flooded across the network.
- New LSPs are acknowledged on the lines they are received and sent on all other lines



Network

Link State Packets

A	
Seq.	
Age	
B	4
E	5

B	
Seq.	
Age	
A	4
C	2
F	6

C	
Seq.	
Age	
B	2
D	3
E	1

D	
Seq.	
Age	
C	3
F	7

E	
Seq.	
Age	
A	5
C	1
F	8

F	
Seq.	
Age	
B	6
D	7
E	8

LSP for each node

Link State Routing – Reliable Flooding

Sequence number and age are used for reliable flooding

- At the source router **sequence number** field is incremented for each new LSP and **age** field is set to a max value, i.e., the higher the sequence number the newer the LSP.
- Each router maintains a LSP database (LSPDB), example shows the LSP database at router B

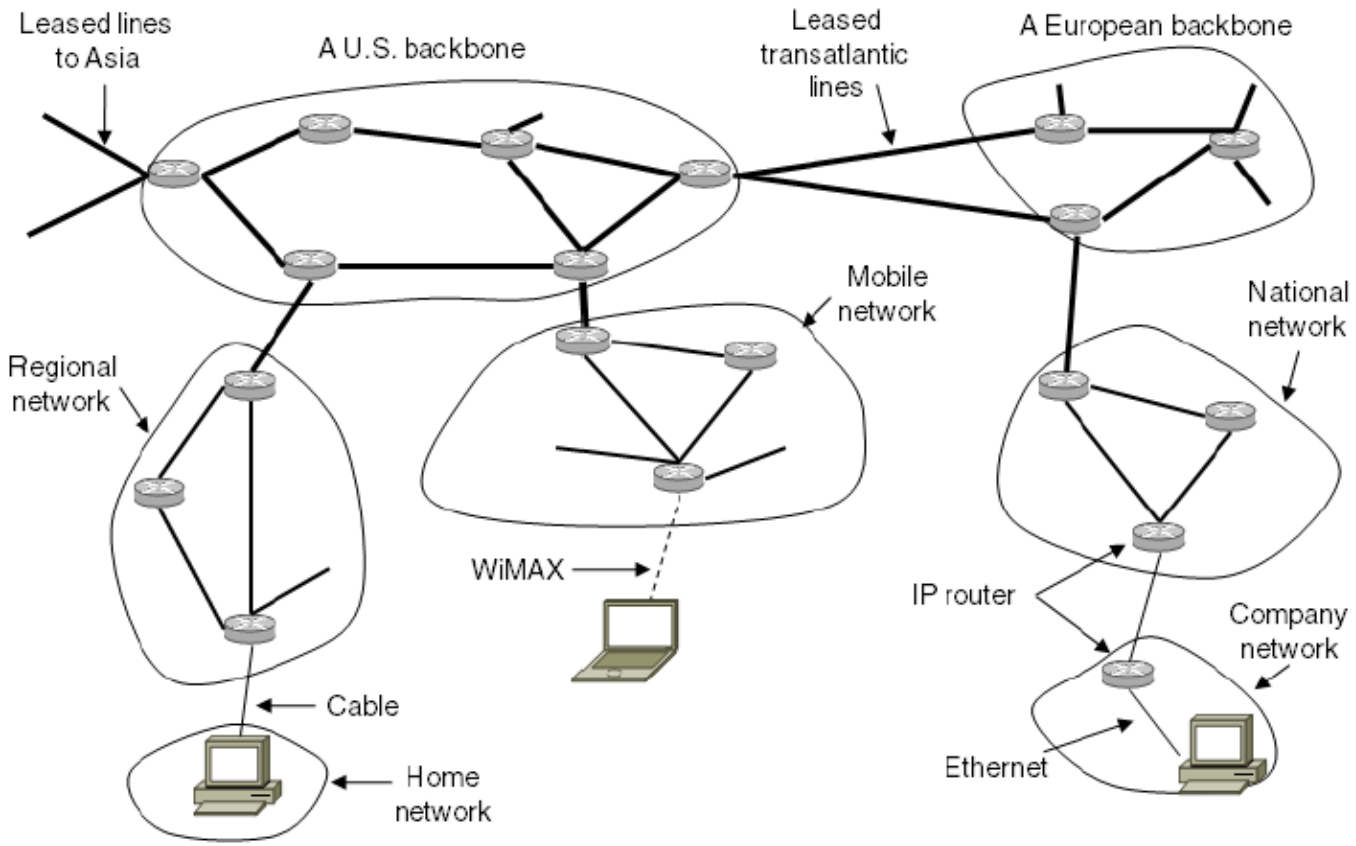
			Send flags			ACK flags			
Source	Seq.	Age	A	C	F	A	C	F	Data
A	21	60	0	1	1	1	0	0	
F	21	60	1	1	0	0	0	1	
E	21	59	0	1	0	1	0	1	
C	20	60	1	0	1	0	1	0	
D	21	59	1	0	0	0	1	1	

Link State Routing – Reliable Flooding

- LSPDB is updated upon the reception of a new LSP.
- The reception of an old LSP is ignored at each router.
- The age field of an LSP on the flight and in LSPDB is decremented by the routers every second and if the age of an LSP reaches to zero that LSP is immediately discarded.

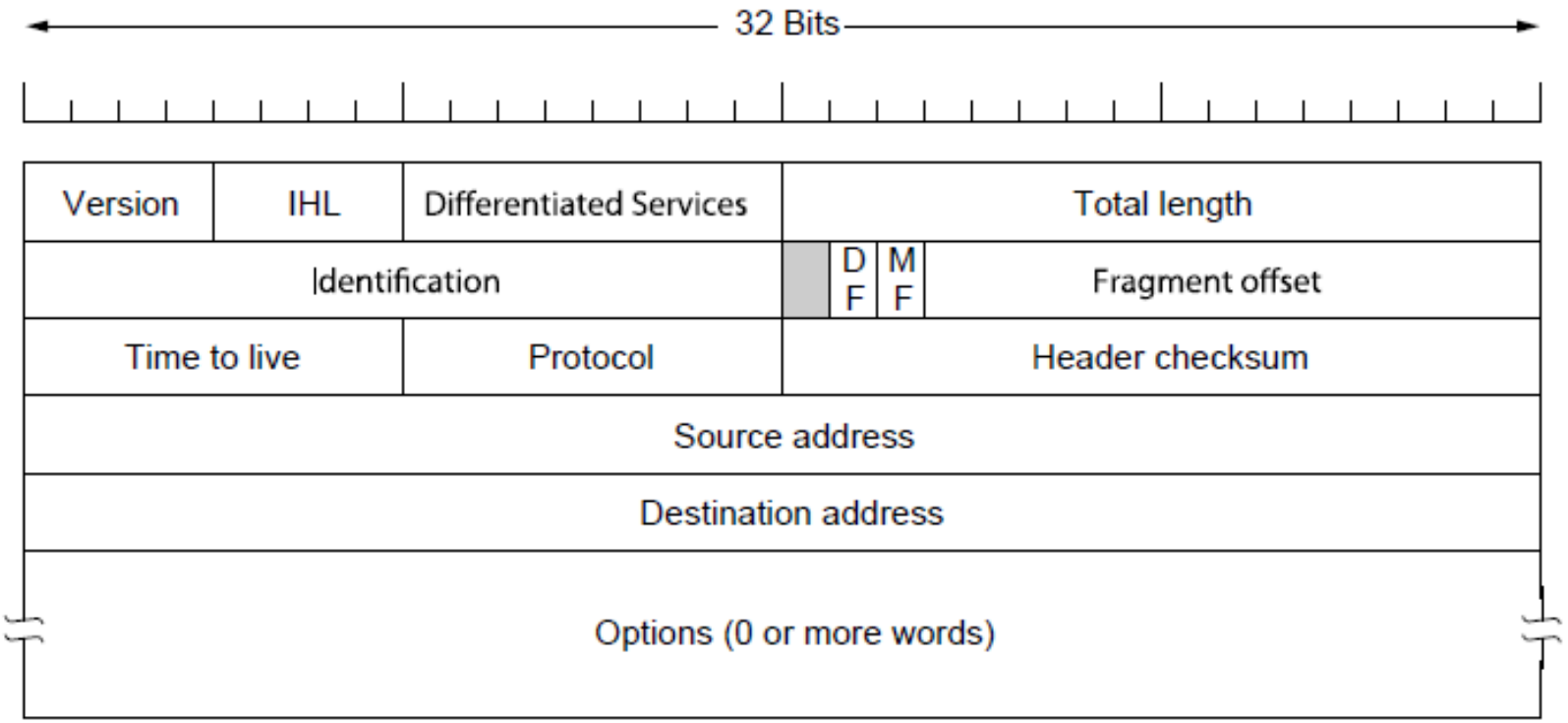
Internet Protocol (IP)

Internet is an interconnected collection of many networks that is held together by the IP protocol



IP Version 4 Protocol Header

IPv4 (Internet Protocol) header is carried on all packets and has fields for the key parts of the protocol:



IP Version 4 Protocol Header

- **Version** (4 bits): IP version (IPv4)
- **IHL** (4 bits): IP header length
  - IP header can carry zero or more optional headers.
  - IHL indicates the total header (fixed and optional) length in 32-bit words or 4-byte words. IHL minimum and maximum are 5 and 15 (20 and 60 bytes) respectively.

IP Version 4 Protocol Header

- **Differentiated services** (8 bits): First 6 bits for representing service classes and the last 2 bits for explicit congestion notification (ECN).
  - Service classes examples
    - Default forwarding (DF) – Best effort traffic
    - Expediated forwarding (EF) – Low-loss and low-latency traffic
    - Assured forwarding (AF) – Traffic that needs delivery assurance within its subscription rate limit.

IP Version 4 Protocol Header

- **Total length** (16 bits): The total length of everything in the IP packet, both header and data, in bytes. i.e., maximum 65,535 bytes.
- **Identification** (16 bits): Indicates the identification of a non-fragmented IP packet.
  - A sender sets an incremented identification for each new IP packet that it sends.
  - When an IP packet is fragmented by a router, all fragments of the same IP packet carries the same identification value, which helps the destination to reassemble the IP packet from its fragments.

IP Version 4 Protocol Header

- **DF** (1 bit): Don't fragment – If set by a sender, it instructs the router not to fragment the IP packet.
- **MF** (1 bit): More fragment – If set by a router, it instructs the destination that more fragment of an IP packet is on the way.
  - Router sets MF bit of each fragment except the last one of an IP packet.
  - Helps the destination to know when to start the reassembly process.

IP Version 4 Protocol Header

- **Fragment offset** (13 bits): Tells where in the IP packet this fragment belongs..
  - Has no meaningful use in non-fragmented IP packet.
  - Expressed as the multiple of 8-bytes, i.e. a value 100 means that the fragment starts at 800 bytes position in the original IP packet.
  - Helps the destination to reassemble the original IP packet from its fragments.
- **Time to live** (8 bits):
  - Set to 255 by the sender and decremented by each router, i.e., hop counting.
  - An IP packet is thrown away when TTL reaches to zero.

IP Version 4 Protocol Header

- **Protocol** (8 bits): Indicates the upper layer protocol; TCP (6) or UDP (17).
- **Header checksum** (16 bits):
  - Modulo 2^{16} sum of the header.
  - Both sender and receiver add up all the 16-bit halfwords of the header using one's complement arithmetic and take one's complement of the result.
  - Sender uses zeros for the checksum field while computing the checksum.
  - Receiver uses received checksum value while computing the checksum.
  - The header of an IP packet is believed to be error free when the receiver computed checksum is zero.

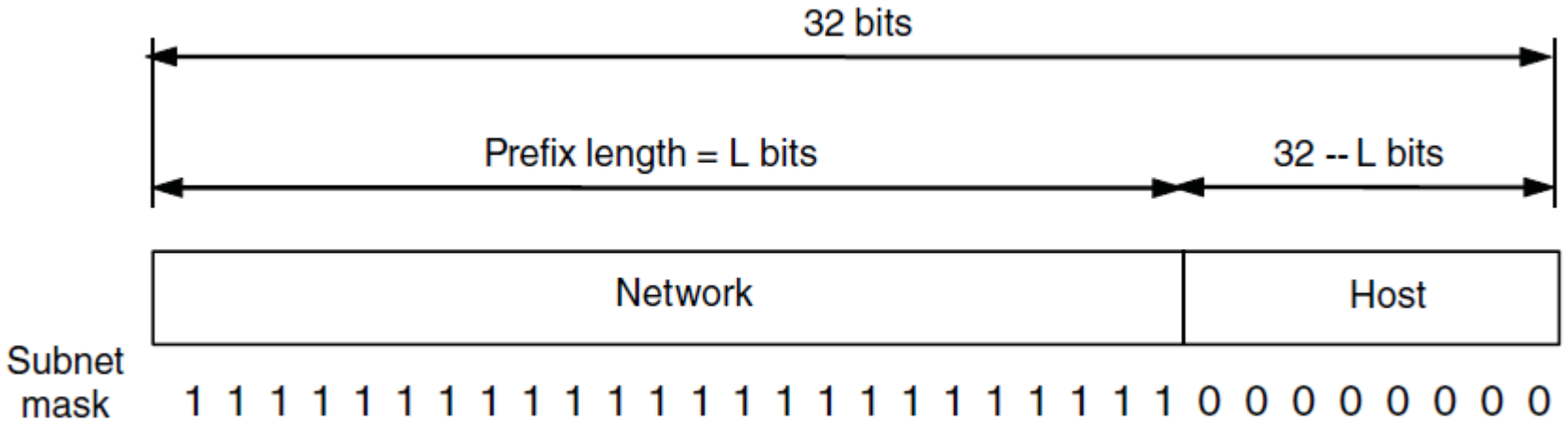
IP Version 4 Protocol Header

- **Source address** (32 bits): The IP address of the source network interface.
- **Destination address** (32 bits): The IP address of the destination network interface.
- **Options:** Optional Headers
  - Each optional header must be multiple of 32-bit word
  - Max10 words or 40 bytes for all optional headers of an IP packet.
  - Example optional headers: Security, Strict source routing, Loose source routing, Record route, and Timestamp.

IP Addresses – Prefixes

Addresses are allocated in blocks called prefixes

- Prefix is determined by the network portion
- Has 2^L addresses aligned on 2^L boundary
- Written address/length, e.g., 18.0.31.0/24

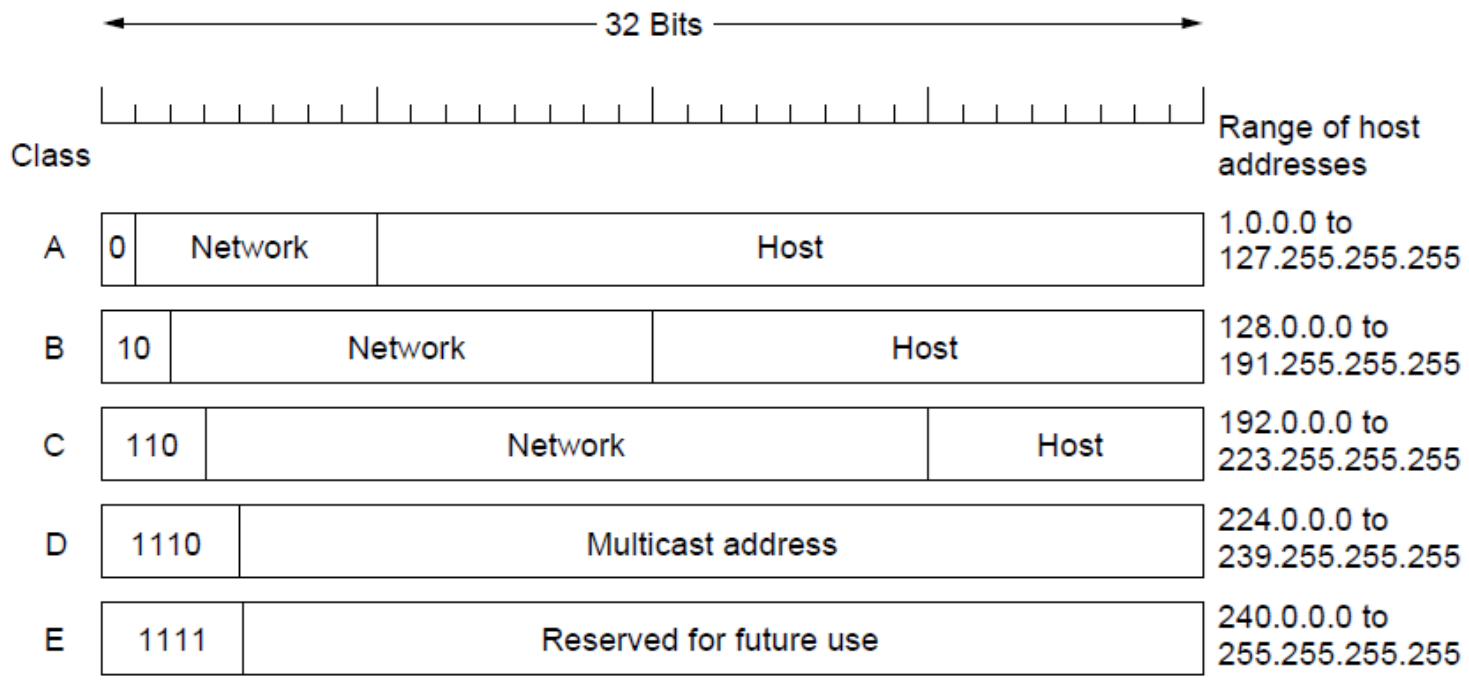


- Prefix length is sometime represented by subnet mask, e.g., 18.0.31.0, **255.255.255.0**

IP Addresses – Classful Addressing

Old addresses came in blocks of fixed size (A, B, C)

- Carries size as part of address, but lacks flexibility
- Called class full (vs. classless) addressing



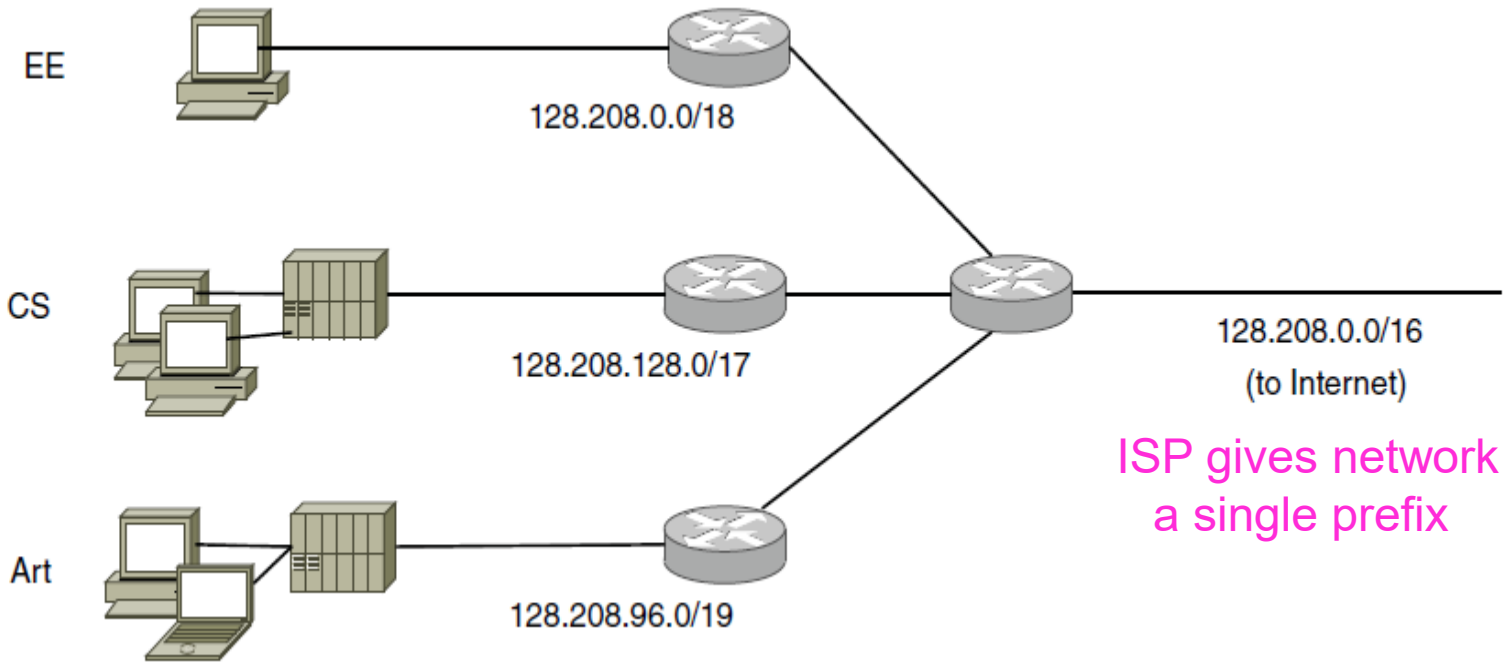
IP Addresses – CIDR

- In Classless Inter-Domain Routing (CIDR), IP address prefixes are not of fixed sizes but varying sizes.
- Routers have the corresponding prefix information of an IP address to make routing decision, e.g., 128.120.0.0/16 or 128.120.0.0, 255.255.0.0

CIDR IP Addresses – Subnets

Subnetting splits up IP prefix to help with management

- Looks like a single prefix outside the network



Network divides it into subnets internally

IP Addresses – CIDR

- 128.208.0.0/16
- 16 bit network prefix 128.208
- 16 bits left for host numbering, 2^{16} **hosts** if placed in a single network.
- 1 or more left most host bits are used to divide the network into subnets.
- Using **1 left most host** bit **two subnets** are created as follows and one subnet is **assigned** to **CS** department.

128.208.**0**0000000.000000000	128.208.**0**.0/17	**Other**
128.208.**1**0000000.000000000	128.208.**1**28.0/17	**CS**

IP Addresses – CIDR

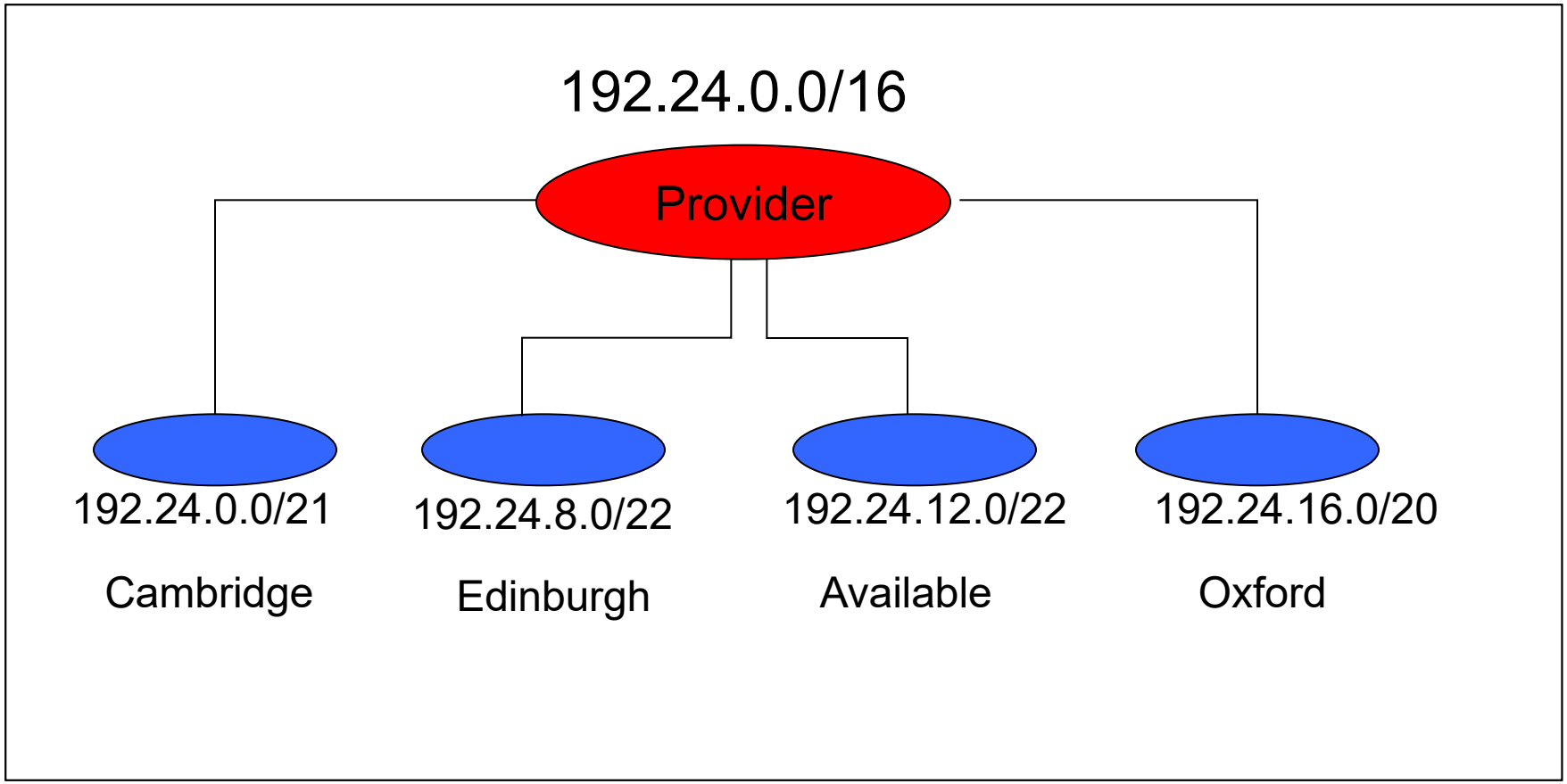
- Using **1 left most host** bit of the **Other** subnet **two more subnets** are created as follows and one subnet is assigned to **EE department.**

128.208.**00**000000.000000000	128.208.**0**.0/18	**EE**
128.208.**01**000000.000000000	128.208.**64**.0/18	**New Other**

- Using **1 left most host** bit of the **New Other** subnet **two more subnets** are created as follows and one subnet is assigned to **EE department.**

128.208.**010**00000.000000000	128.208.**64**.0/19	**Available**
128.208.**011**00000.000000000	128.208.**96**.0/19	**Arts**

CIDR IP Address: ISP Assignments



CIDR IP Address: ISP Assignments

- 192.24.0.0/16
- 16 bit network prefix 192.24
- 16 bits left for host numbering, 2^{16} **hosts** if placed in a single network.
- Using **4 left most host** bit **2^4** or **16 subnets** are created as follows and the second subnet is **assigned** to **Oxford**.

192.24.**0000**0000.000000000	192.24.**0**.0/20	First subnet
192.24.**0001**0000.000000000	192.24.**16**.0/20	**Oxford**
192.24.**0010**0000.000000000	192.24.**32**.0/20	Third subnet
192.24.**1111**0000.000000000	192.24.**240**.0/20	Sixteenth subnet

CIDR IP Address: ISP Assignments

- Using **1 left most host** bit of the **first** subnet **two more subnets** are created as follows and the first subnet is assigned to **Cambridge.**

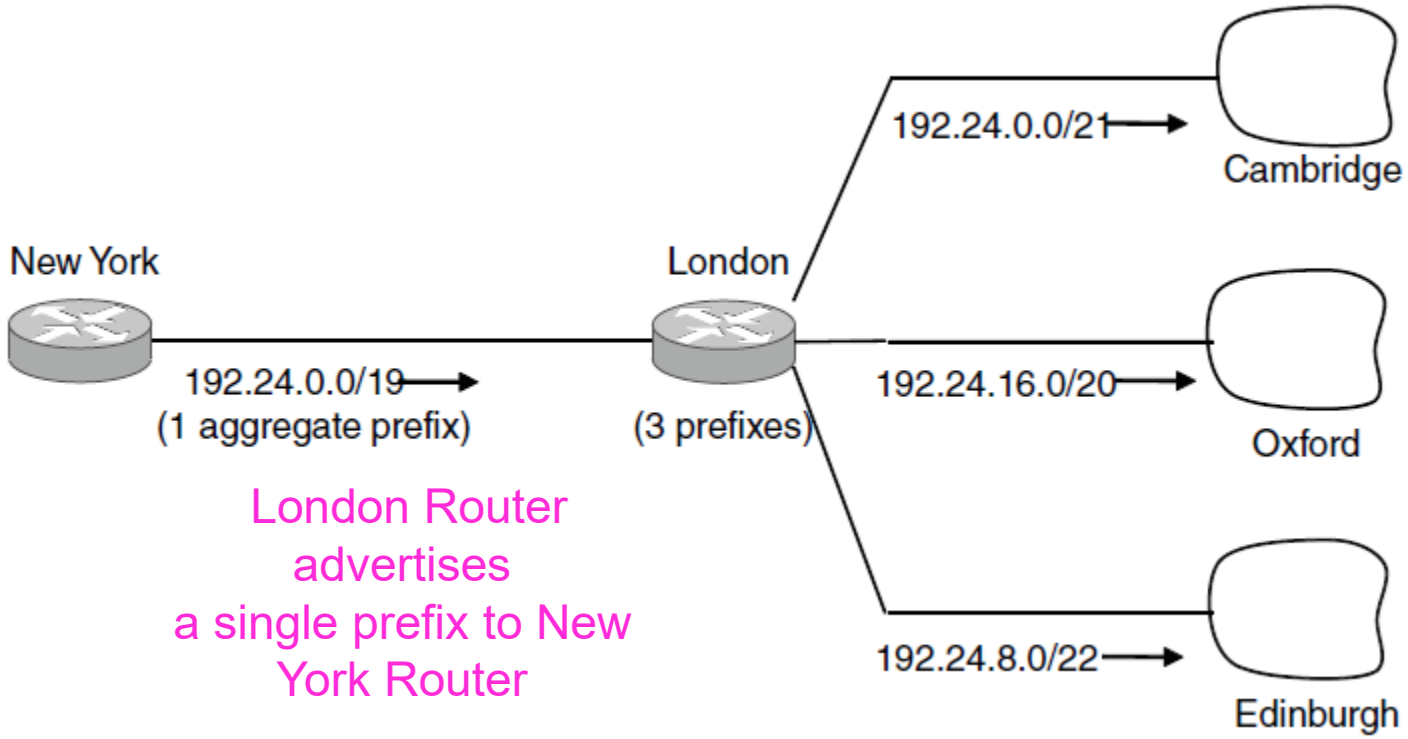
192.24.**00000**000.000000000	192.24.**0**.0/21	**Cambridge**
192.24.**00001**000.000000000	192.24.**8**.0/21	Second sub-subnet

- Using **1 left most host** bit of the **second sub-subnet** two more **subnets** are created as follows and one sub-sub-subnet is assigned to **Edinburgh** and the other sub-sub-subnet is left available**.**

192.24.**00001**000.000000000	192.24.**8**.0/22	**Edinburgh**
192.24.**00001**100.000000000	192.24.**12**.0/22	Available

CIDR IP Addresses – Aggregation

CIDR enables route aggregation to reduce the number of routing table entries. Aggregation joins multiple IP prefixes into a single larger prefix to reduce routing table size.



London Router advertises a single prefix to New York Router

ISP customers have different prefixes

CIDR IP Address: Aggregation

192.24.**00000**000.000000000	192.24.**0**.0/21	**Cambridge**
192.24.**0001**0000.000000000	192.24.**16**.0/20	**Oxford**
192.24.**00001**000.000000000	192.24.**8**.0/22	**Edinburgh**

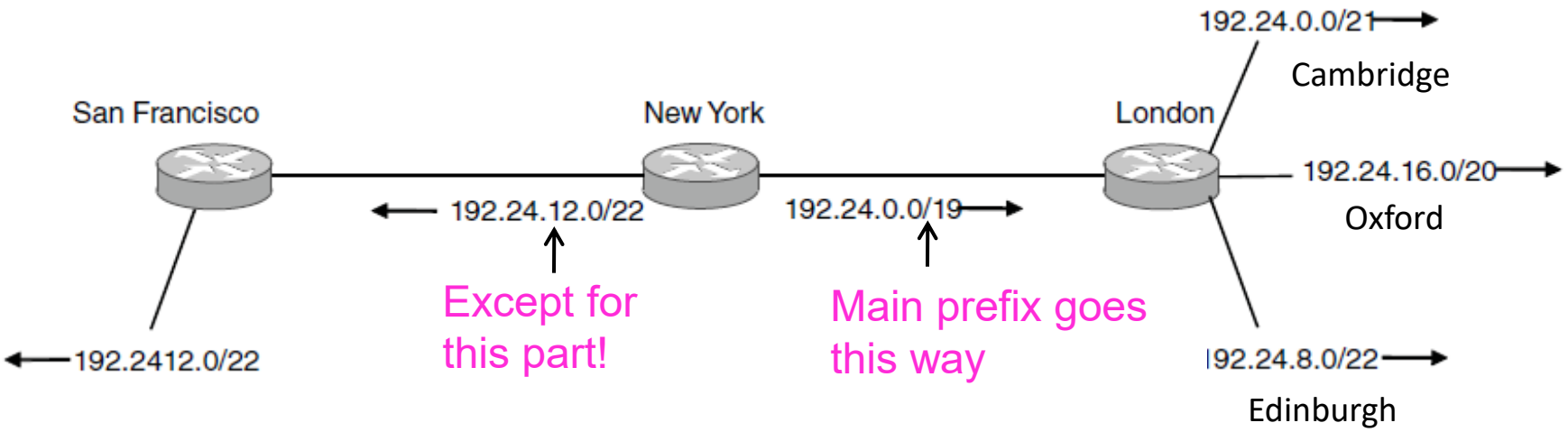
192.24.**000**00000.000000000	192.24.0.0/19	Aggregated Prefix by **London router**

CIDR IP Addresses – Longest Matching Prefix

A router may have multiple entries with common prefix but with different prefix lengths.

Packets are forwarded to the entry with the longest matching prefix

Complicates forwarding but adds flexibility



CIDR IP Address: Longest Matching Prefix

- **Oxford**, **Cambridge,** and **Edinburgh** subnets are connected by **London** router.
- **London** router aggregates 3 subnets into 192.24.0.0/19 aggregated prefix and advertises it to **New York** router.
- One of the subnet 192.24.12.0/22 available before is assigned to **San Francisco.** San Francisco router advertises it to **New York** router.
- **New York** router has two entries having with common prefix but varying prefix lengths (19 and 22).

192.24.**000**00000.000000000	192.24.0.0/19	**London**
192.24.**00001100**.000000000	192.24.**12**.0/22	**San Francisco**

CIDR IP Address: Longest Matching Prefix

192.24.**000**00000.000000000	192.24.0.0/19	**London**
192.24.**00001**100.000000000	192.24.**12**.0/22	**San Francisco**

- When a packet comes to **New York** router its **destination address** is compared with the **longest prefix entry** 192.24.12.0/22 first and forward the packet to **San Francisco** if matches.
- **New York** router compares packet **destination address** with the next **longest prefix entry** 192.24.0.0/19 next and forward the packet to **London** if matches.

CIDR IP Address: Longest Matching Prefix

- If a packet comes with the **destination address** 192.24.15.252
- It is compared with the **longest prefix entry** 192.24.12.0/22 first and matches.

192.24.**000011**00.000000000	192.24.**12**.0/22	**San Francisco**
192.24.000011**11**.11111100	192.24.**15**.252	

- The packet is forwarded to **San Francisco.**

CIDR IP Address: Longest Matching Prefix

- If another packet comes with the **destination address** 192.24.20.248
- It is compared with the **longest prefix entry** 192.24.12.0/22 first and **does not match**.

192.24.**00001100**.000000000	192.24.**12**.0/22	**San Francisco**
192.24.00010100.11111000	192.24.**20**.248	

- It is compared with the **next longest prefix entry** 192.24.0.0/19 next and **matches**.

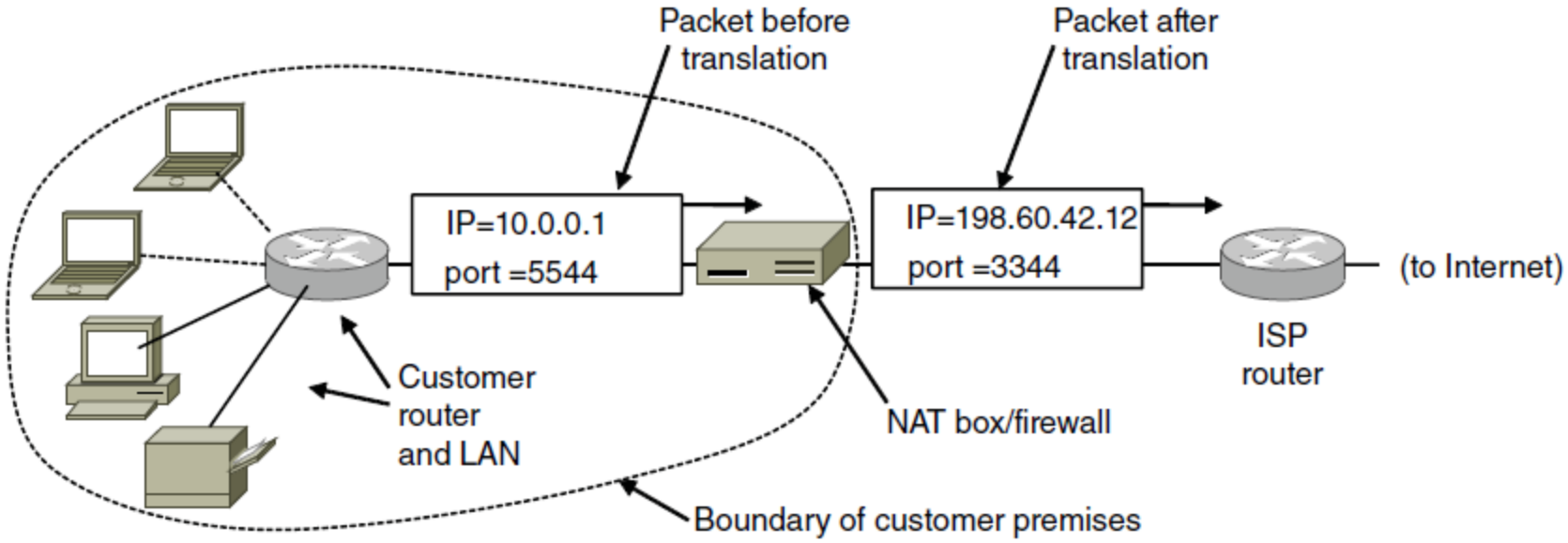
192.24.**000000**00.000000000	192.24.**0**.0/19	**London**
192.24.00010100.11111000	192.24.**20**.248	

- The packet is forwarded to **London.**

IP Addresses – NAT

NAT (Network Address Translation) box maps one external IP address to many internal IP addresses

- Uses TCP/UDP port to tell connections apart
- Violates layering; very common in homes, etc.



IP Addresses – NAT

Three range of IP addresses are declared private.

10.0.0.0	– 10.255.255.255/8	(16,777,216 hosts)
172.16.0.0	– 172.31.255.255/12	(1,048,576 hosts)
192.168.0.0	– 192.168.255.255/16	(65,536 hosts)

Internet routers do not forward any IP packet with these private address as the destination.

IP Addresses – NAT

NAT table is used for translating private-to-public and public-to-private IP addresses.

Index	Source Port	Source IP
0	……..	……
1	……..	……
2	……..	……
….	……..	……
3344	5544	10.0.0.1

IP Addresses – NAT

Outgoing IP Packet Before NAT

Source IP	Destination IP	Source Port	Destination Port	Payload
10.0.0.1	142.104.2.1	5544	80	Payload

Outgoing IP Packet After NAT

Source IP	Destination IP	Source Port	Destination Port	Payload
196.60.42.12	142.104.2.1	3344	80	Payload

Incoming IP Packet Before NAT

Source IP	Destination IP	Source Port	Destination Port	Payload
142.104.2.1	196.60.42.12	80	3344	Payload

Incoming IP Packet After NAT

Source IP	Destination IP	Source Port	Destination Port	Payload
142.104.2.1	10.0.01	80	5544	Payload

Internet Control Protocols

IP works with the help of several control protocols:

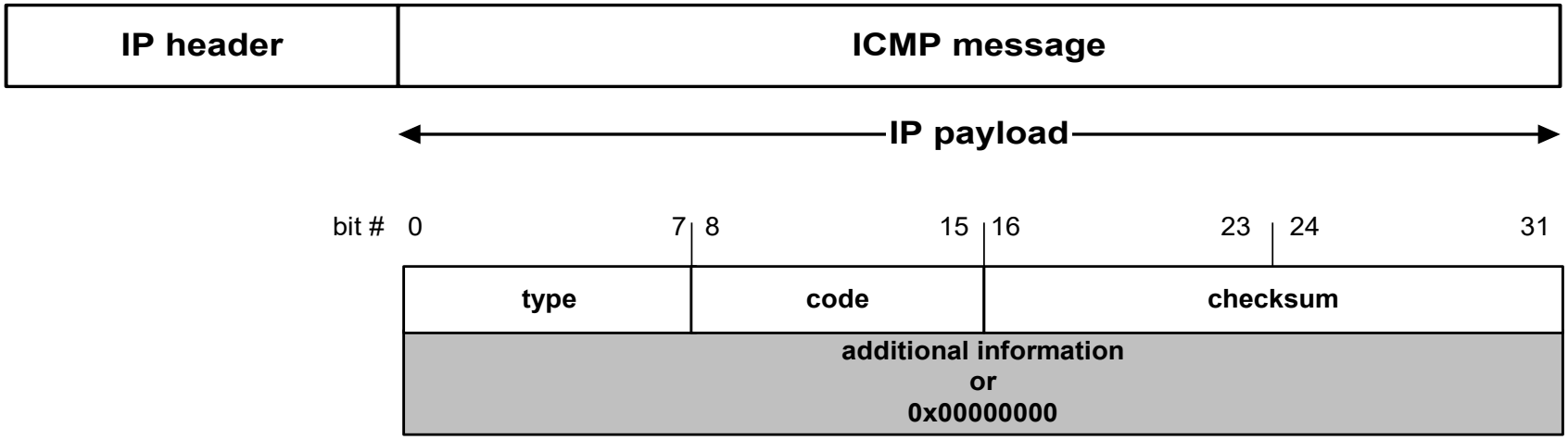
- ICMP is a companion to IP that returns error info
  - Required, and used in many ways, e.g., for traceroute
- ARP finds Ethernet address of a local IP address
  - Glue that is needed to send any IP packets
  - Host queries an address and the owner replies
- DHCP assigns a local IP address to a host
  - Gets host started by automatically configuring it
  - Host sends request to server, which grants a lease

ICMP

The **Internet Control Message Protocol (ICMP)** is a helper protocol that supports IP with facility for

- Error reporting
- Simple queries

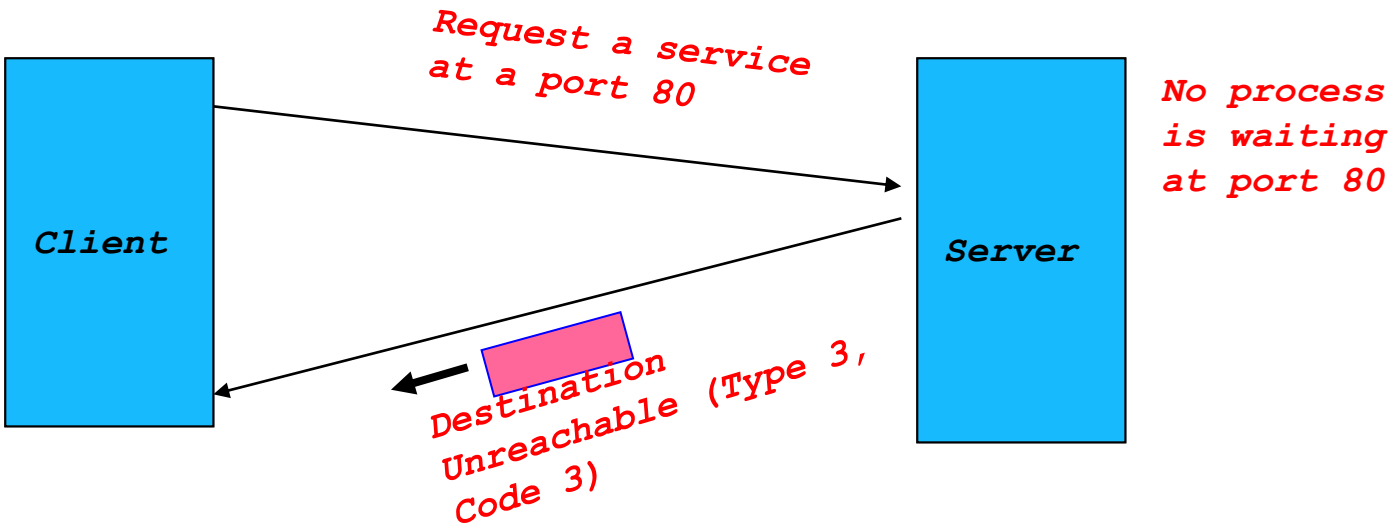
ICMP messages are encapsulated as IP datagrams:



While **type** specifies the message type, **code** specifies the further details within the specified type.

ICMP Error Reporting

If, in the destination host, the IP module cannot deliver the datagram because the indicated protocol module or process port is not active, the destination host may send a destination unreachable message (type 3) to the source host.



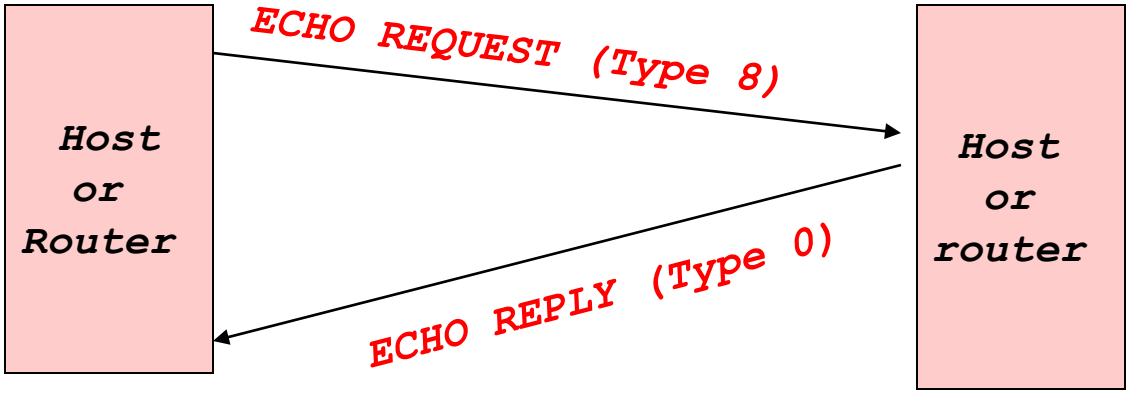
Code 0 specifies network unreachable.
Code 1 specifies host unreachable.
Code 2 specifies protocol unreachable.

ICMP Request and Reply

Ping's are handled directly by the kernel

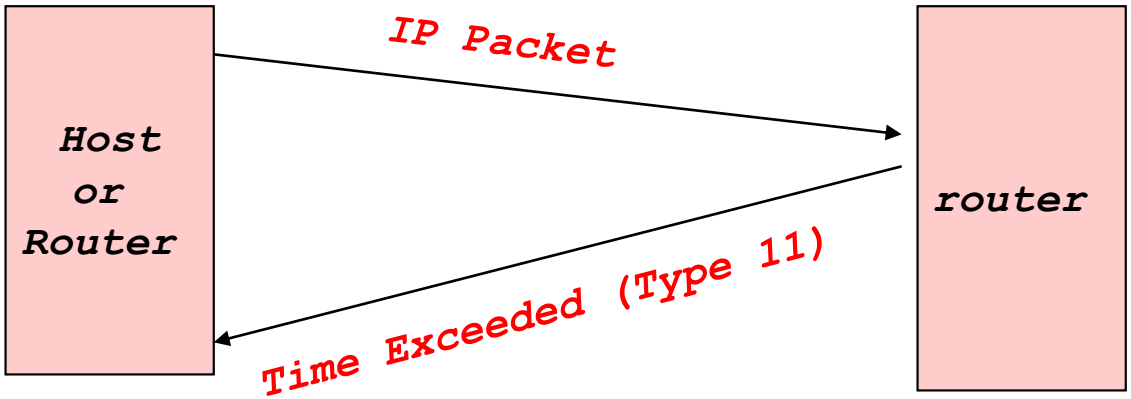
Each Ping is translated into an Echo Request

The Ping'ed host responds with an Echo Reply



ICMP Time Exceeded

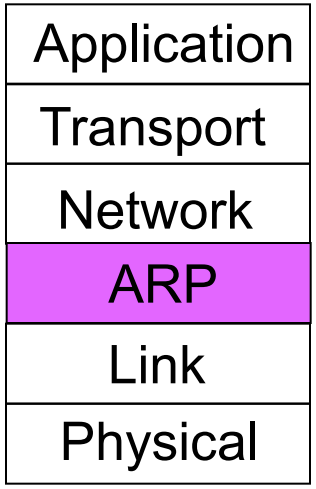
If IP packet's TTL reaches to zero



ARP

ARP (Address Resolution Protocol) operates below the network layer as a part of the interface between the network (IPv4) and the data link layer (Ethernet).

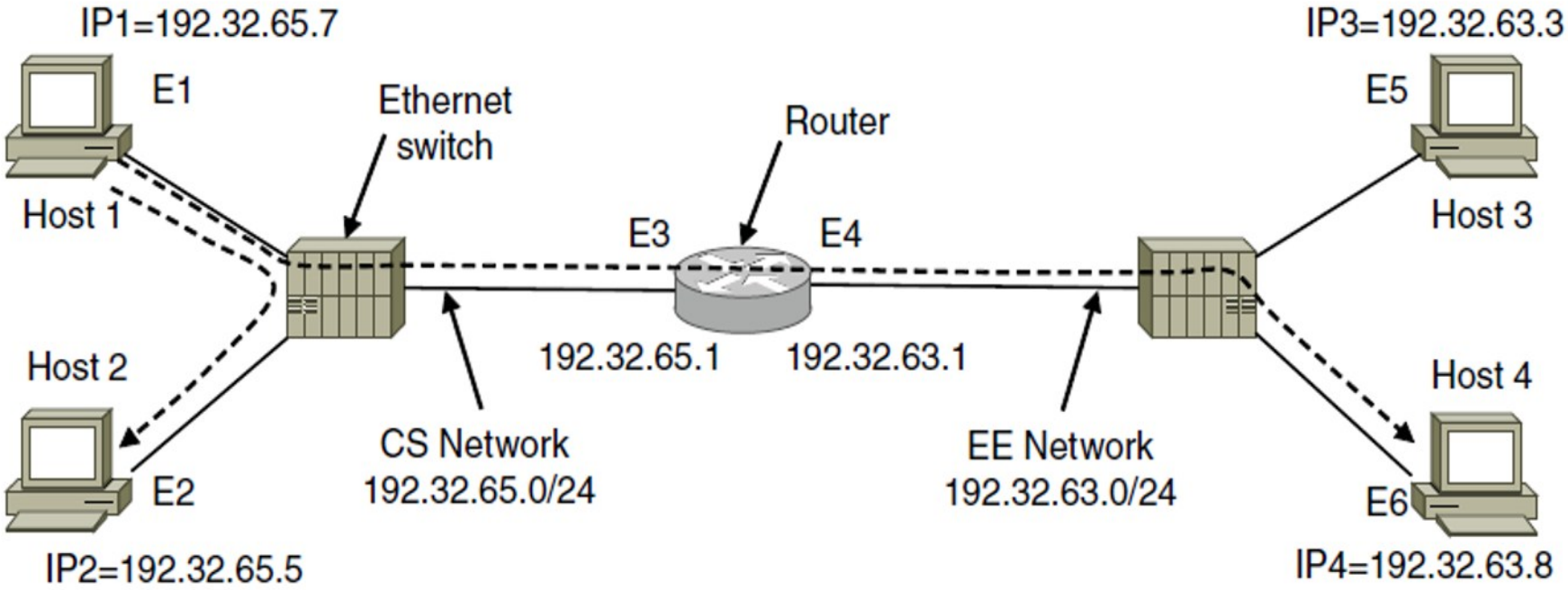
ARP lets nodes find target Ethernet addresses from their IP addresses.



ARP

- ARP **Request** and **Reply** messages (broadcast).
- ARP **Cache**
  - Reduces ARP broadcast.
- **Gratitude** ARP
  - ARP request against self IP address
- **Proxy** ARP
  - Gateway (router) machine replies against out of network IP address.

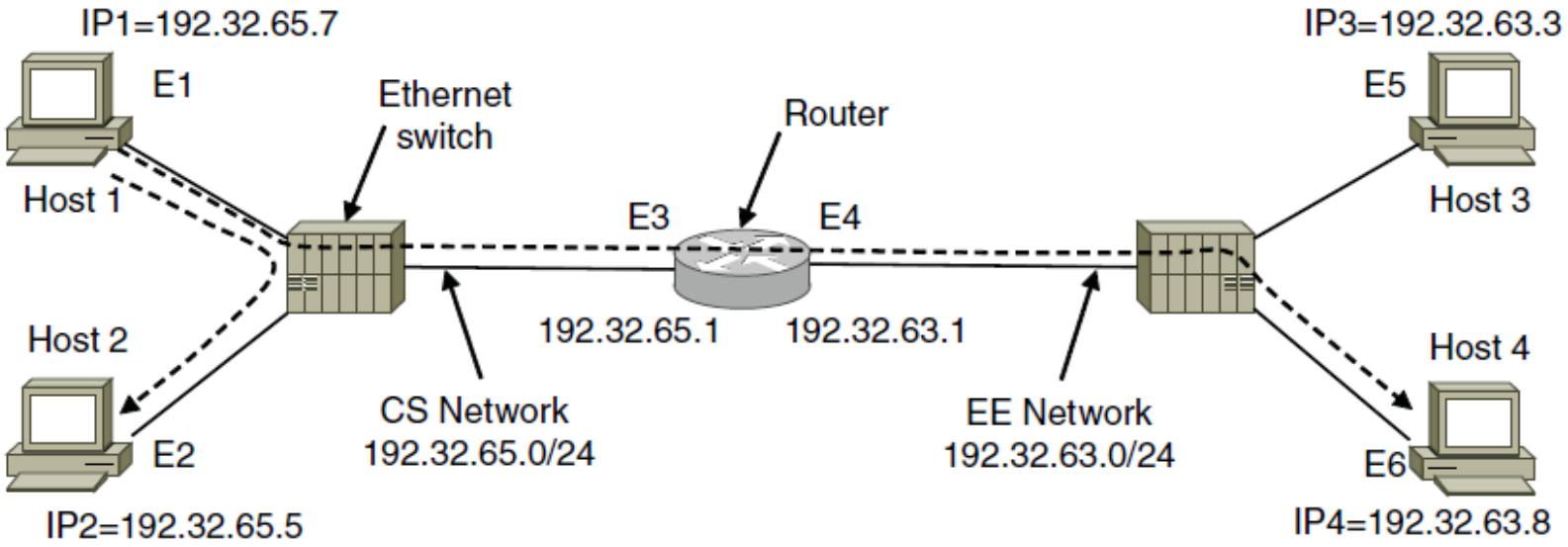
ARP



Destination IP	Destination Ethernet
192.32.65.5	E2
192.32.65.1	E3 (router/gateway)
192.32.63.3	E3 (proxy)
192.32.63.8	E3 (proxy)

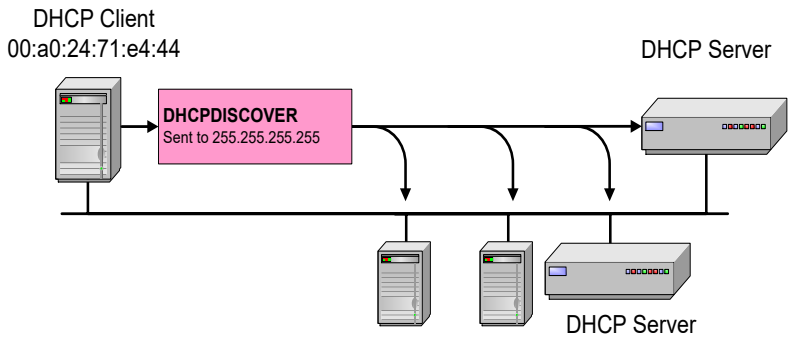
ARP Cache at Host 1 in CS Network

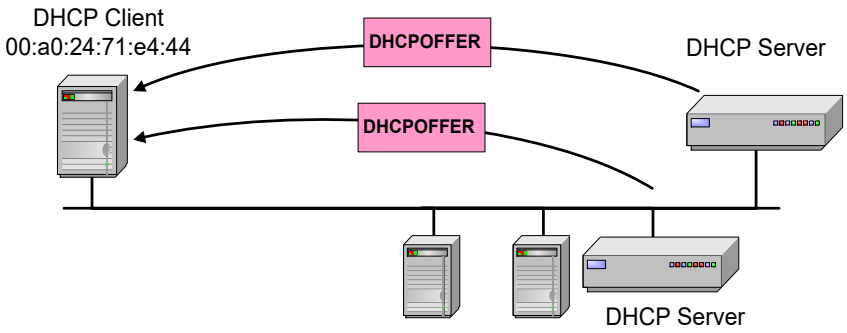
ARP

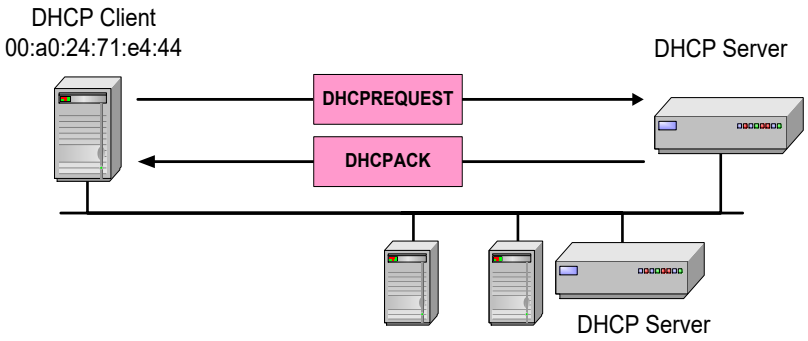


Frame	Source IP	Source Eth.	Destination IP	Destination Eth.
Host 1 to 2, on CS net	IP1	E1	IP2	E2
Host 1 to 4, on CS net	IP1	E1	IP4	E3
Host 1 to 4, on EE net	IP1	E4	IP4	E6

DHCP


DHCP Client
00:a0:24:71:e4:44
DHCP Server
DHCPDISCOVER
Sent to 255.255.255.255
DHCP Server


DHCP Client
00:a0:24:71:e4:44
DHCPOFFER
DHCP Server
DHCPOFFER
DHCP Server


DHCP Client
00:a0:24:71:e4:44
DHCP Server
DHCPREQUEST
DHCPACK
DHCP Server

Summary

- Store and Forward Packet Switching
- Datagrams
- Routers
- Routing Algorithms
  - Shortest Path Routing
  - Distance Vector Routing
  - Link State Routing
- Internet Protocol (IP)
  - IP Packet
  - IP Address, Subnet, and CIDR
  - Network Address Translation (NAT)
- Internet Control Message Protocol (ICMP)
- Address Resolution Protocol (ARP)
- Dynamic Host Configuration Protocol (DHCP)

Next

Transport Layer

- User Datagram Protocol (UDP)
- Transport Control Protocol (TCP)
  - TCP Segment Header
  - TCP Connection
  - TCP Flow Control
  - TCP Congestion Control
- TCP Retransmission Timer