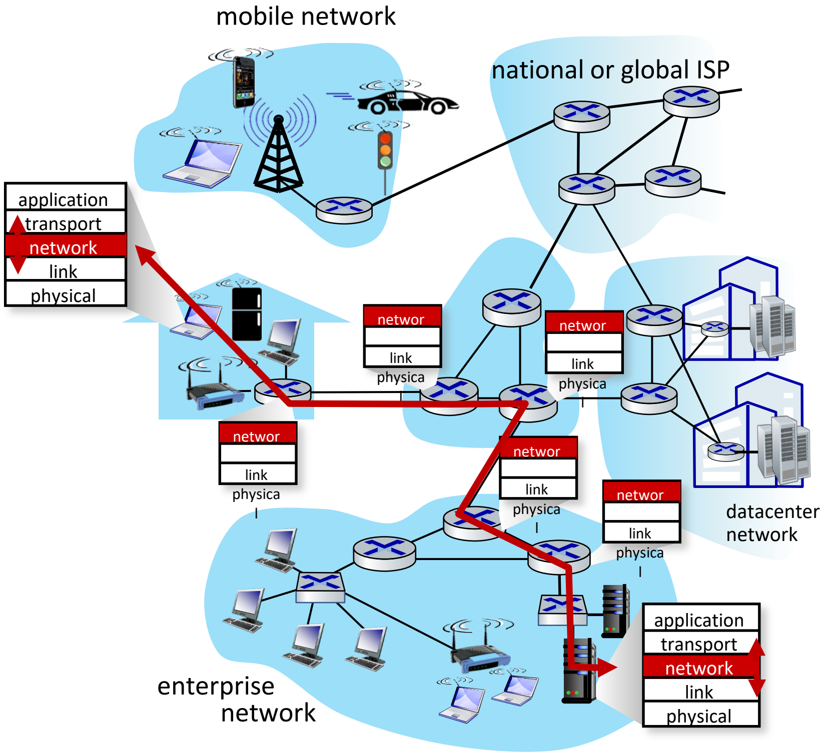
ECE 363
Communication Networks

Network Layer

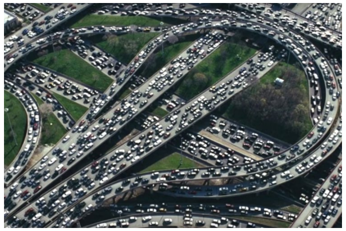
Network Layer Services and Protocols

- Transport segments from sending to receiving host
  - Sender: encapsulates segments into packets and passes them to the link layer
  - Receiver: delivers segments to the transport layer protocol
- Network layer protocols in every Internet host and router
- Routers
  - Examine the header fields in all IP packets passing through it
  - Move packets from input ports to output ports to transfer them along an end-to-end path



Key Network Layer Functions

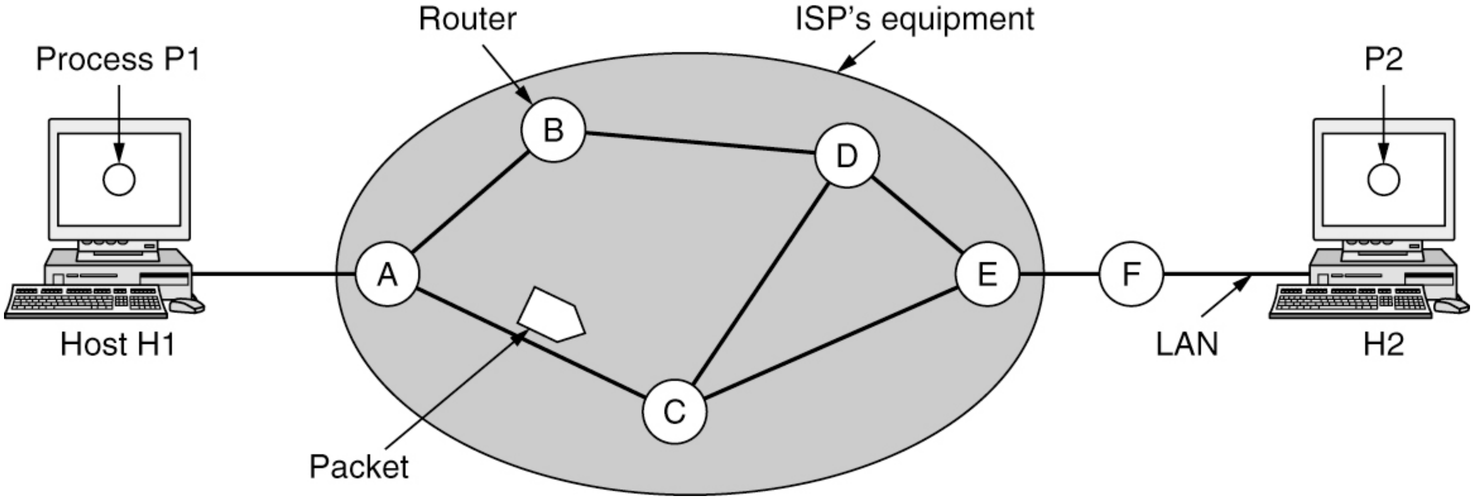
- Forwarding: Move packets from the router input to the appropriate router output
- Routing: Determine the route taken by packets from source to destination
- Analogy: Taking a trip
  - Forwarding: Process of getting through a single interchange
  - Routing: Process of planning the trip from source to destination



forwarding

routing

Store-and-Forward Packet Switching



Packets are stored in routers before they are forwarded

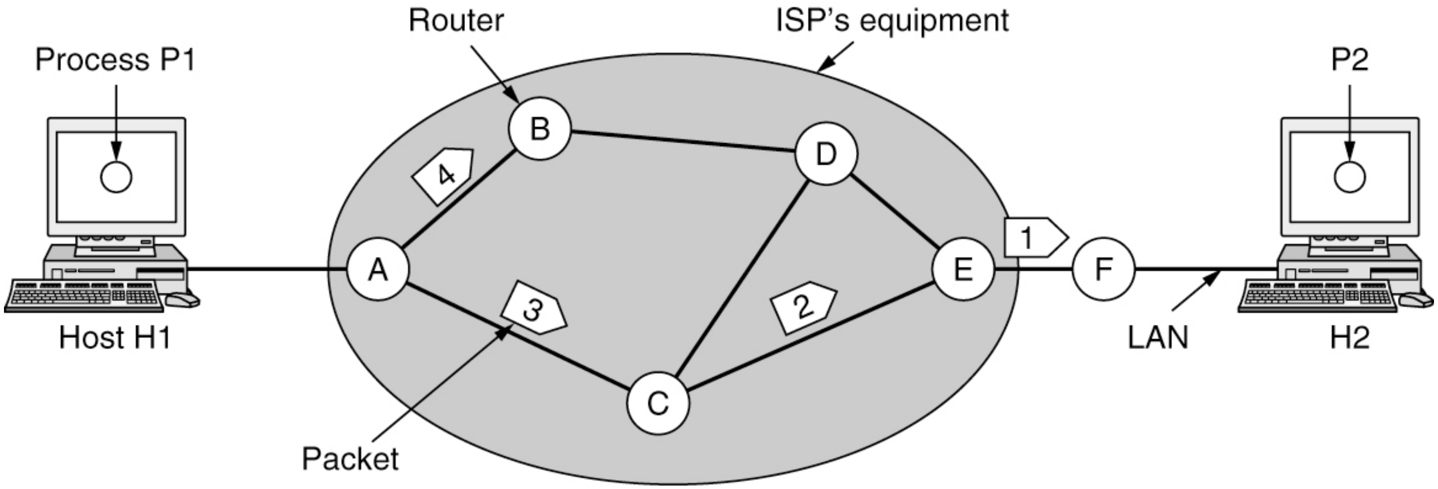
Services Provided to the Transport Layer

- Services independent of router technology
  - Packet delivery
    - Addressing and routing
  - Best effort
    - Packets lost, duplicated, out of order, corrupted
- Transport layer shielded from number, type, topology of routers
- Network addresses available to the transport layer use a uniform numbering plan

Services Provided by the Link Layer

- Frame delivery
  - Point-to-point links
- Medium Access Control (MAC)
  - Controlled access to the shared medium
- Error detection

Connectionless Service



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

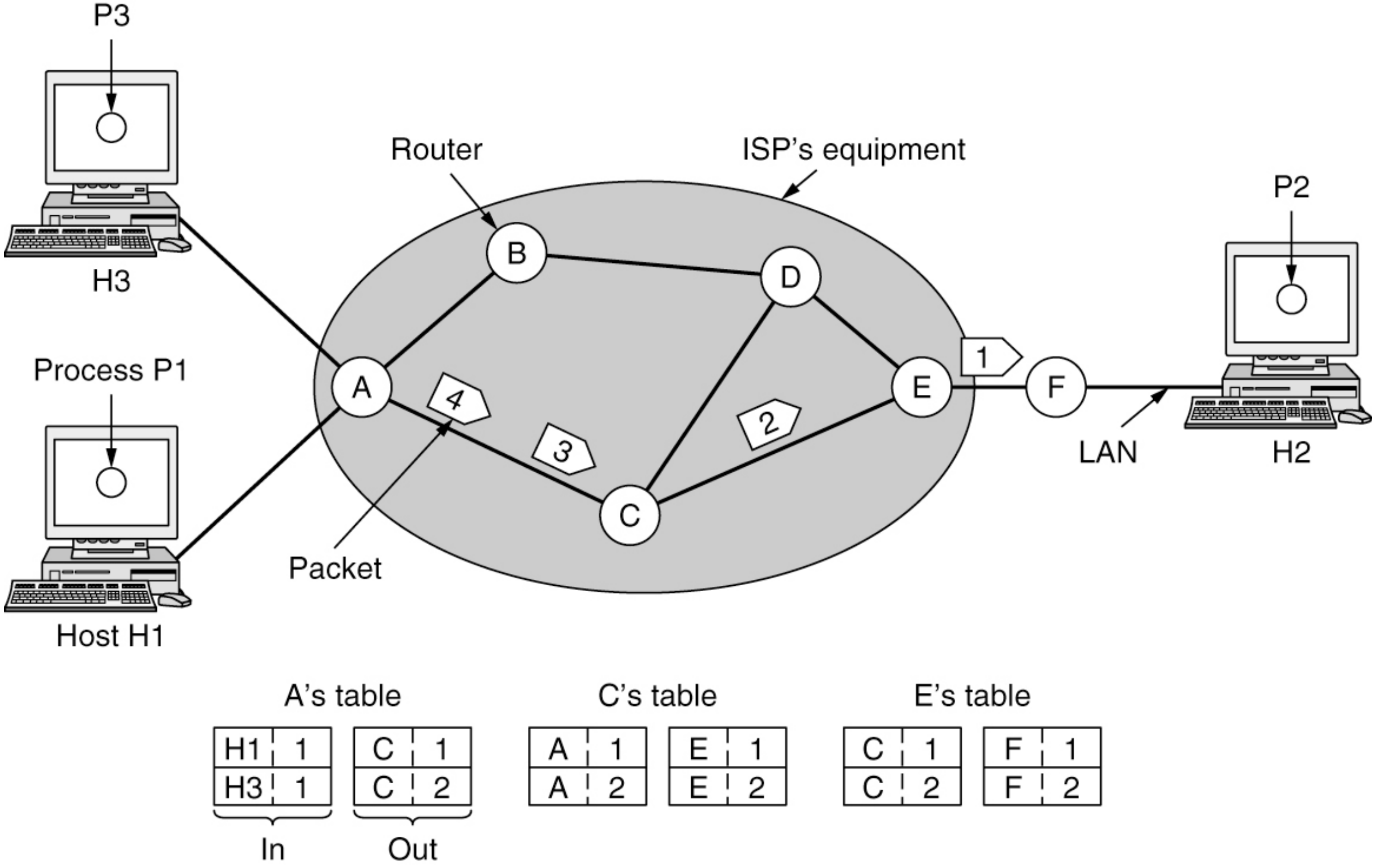
C's table

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

Connection-Oriented Service



A's table

In		Out	
H1	1	C	1
H3	1	C	2

C's table

A	1	E	1
A	2	E	2

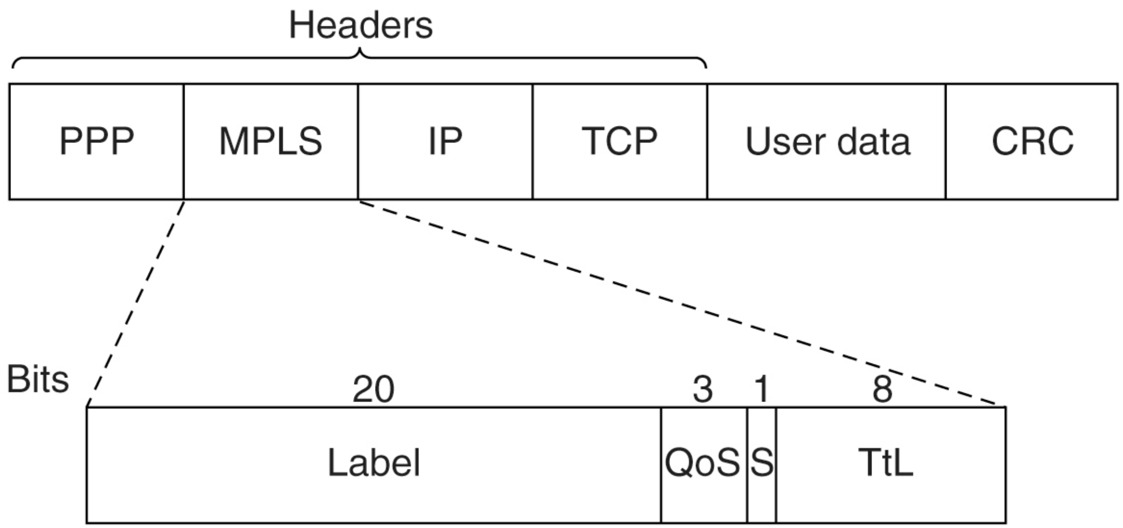
E's table

C	1	F	1
C	2	F	2

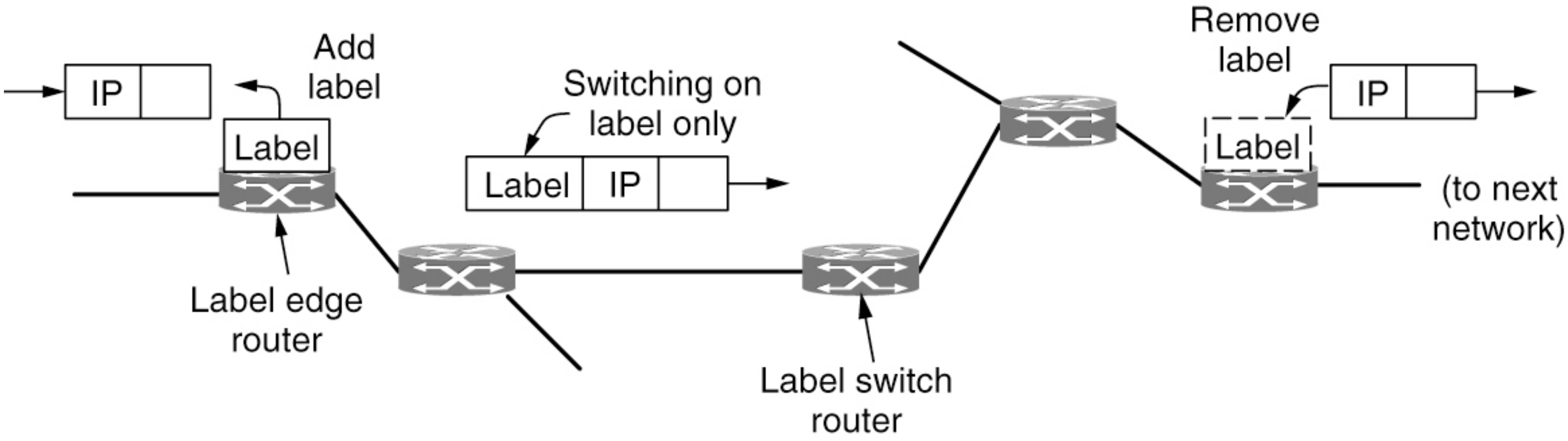
Routing within a virtual-circuit network

MultiProtocol Label Switching (MPLS)

- Virtual circuits used in ISPs
- Adds a label in front of each packet
- Forwarding based on the label (not the destination address)
- Allows forwarding to be done very quickly

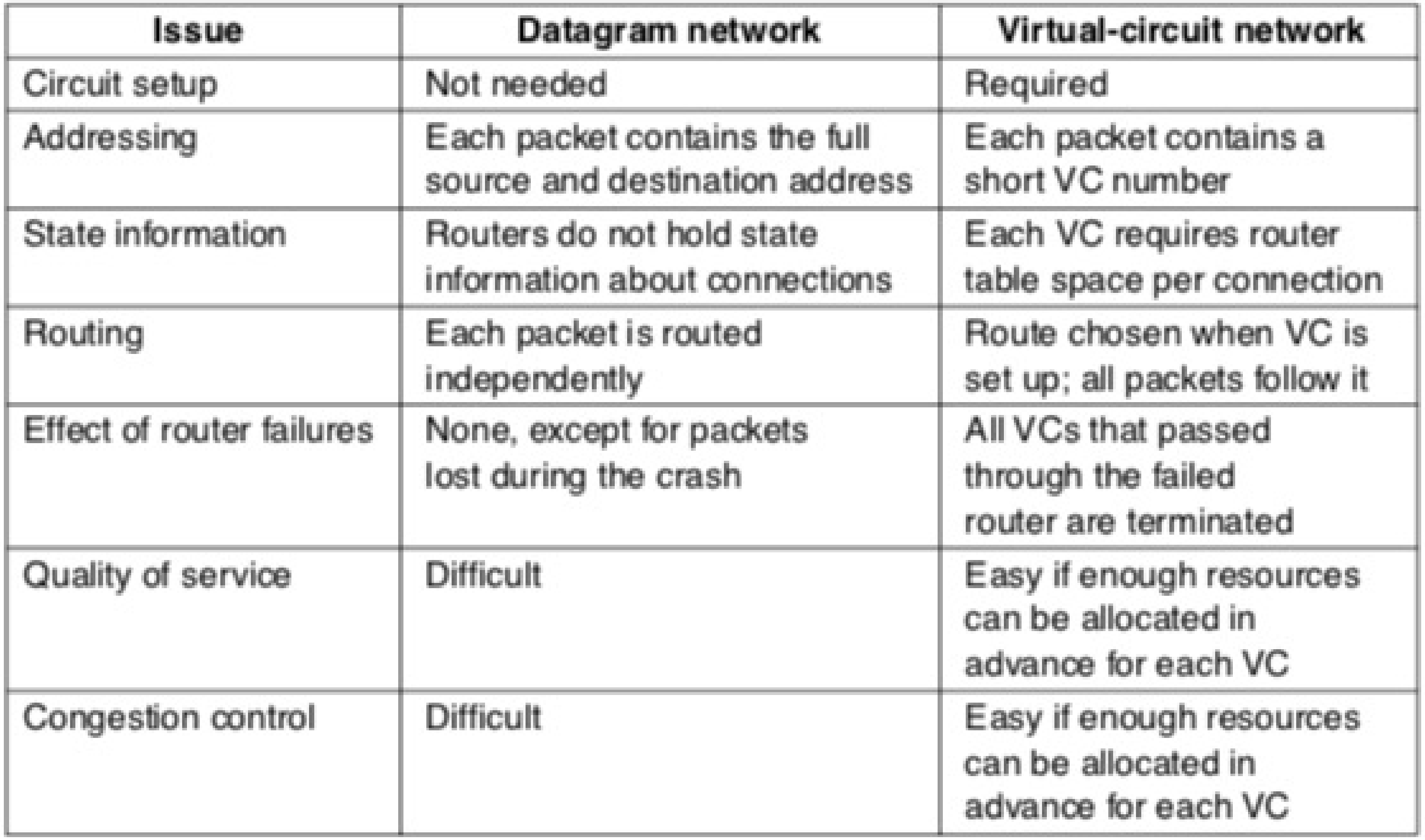


MultiProtocol Label Switching (MPLS)



Forwarding an IP packet through an MPLS network

Datagram and Virtual-Circuit Networks

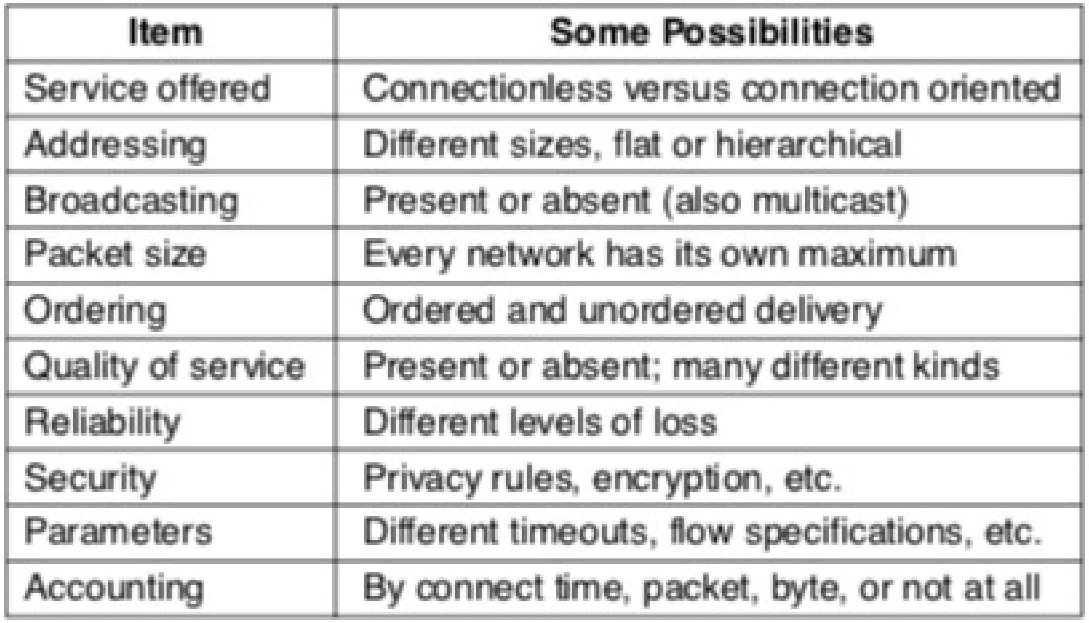


Issue	Datagram network	Virtual-circuit network
Circuit setup	Not needed	Required
Addressing	Each packet contains the full source and destination address	Each packet contains a short VC number
State information	Routers do not hold state information about connections	Each VC requires router table space per connection
Routing	Each packet is routed independently	Route chosen when VC is set up; all packets follow it
Effect of router failures	None, except for packets lost during the crash	All VCs that passed through the failed router are terminated
Quality of service	Difficult	Easy if enough resources can be allocated in advance for each VC
Congestion control	Difficult	Easy if enough resources can be allocated in advance for each VC

Internetworking

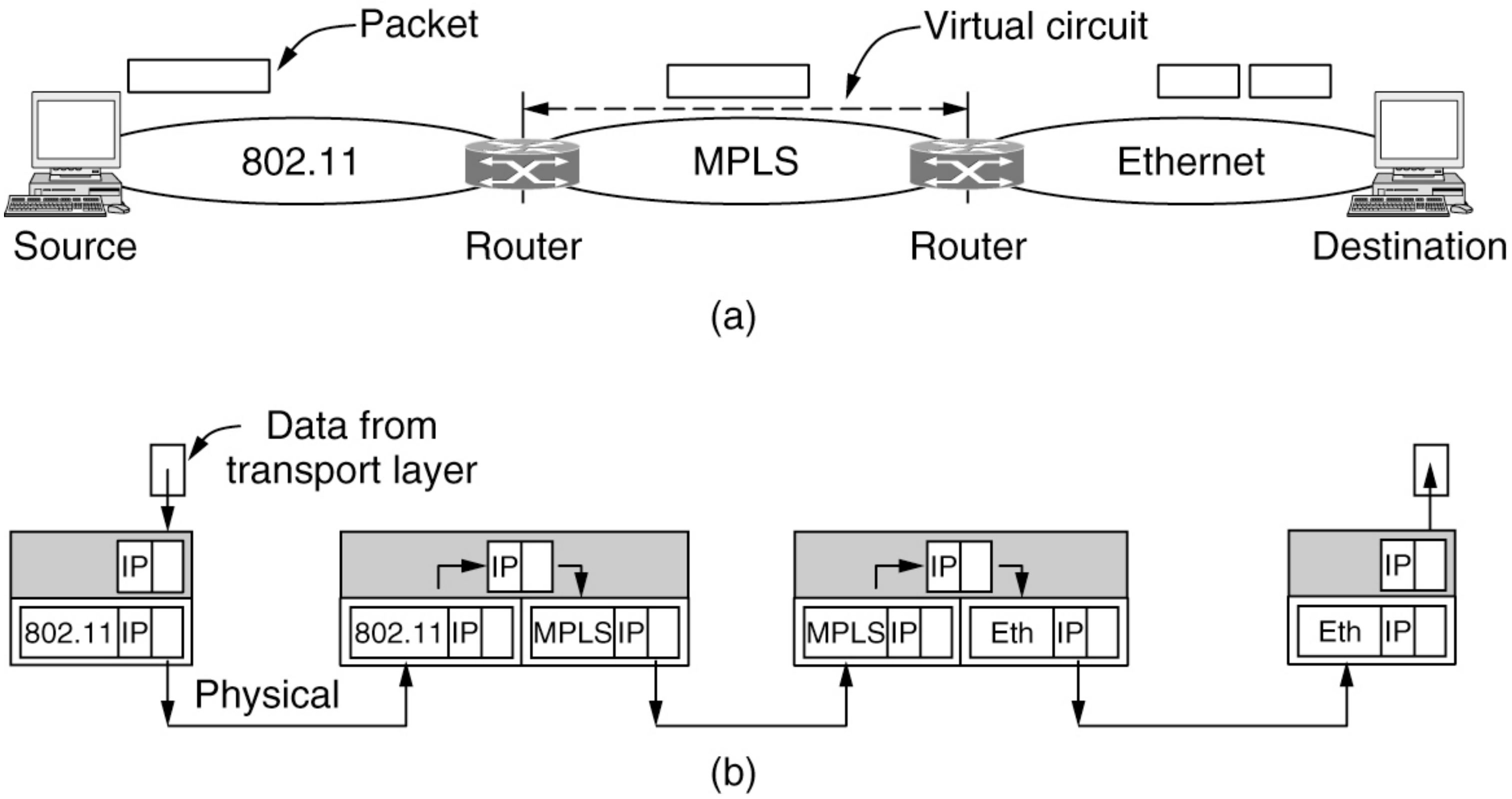
- Internetworking is the connecting of multiple distinct computer networks together to form a larger, unified network, so that devices on different networks can communicate with each other
- The goal is to enable communication across heterogeneous networks (different technologies, topologies, or protocols)
- The result is a network of networks allowing hosts in different physical or logical networks to exchange data transparently
- The Internet is the prime example of an internetwork

How Networks Differ



Item	Some Possibilities
Service offered	Connectionless versus connection oriented
Addressing	Different sizes, flat or hierarchical
Broadcasting	Present or absent (also multicast)
Packet size	Every network has its own maximum
Ordering	Ordered and unordered delivery
Quality of service	Present or absent; many different kinds
Reliability	Different levels of loss
Security	Privacy rules, encryption, etc.
Parameters	Different timeouts, flow specifications, etc.
Accounting	By connect time, packet, byte, or not at all

Internetworking

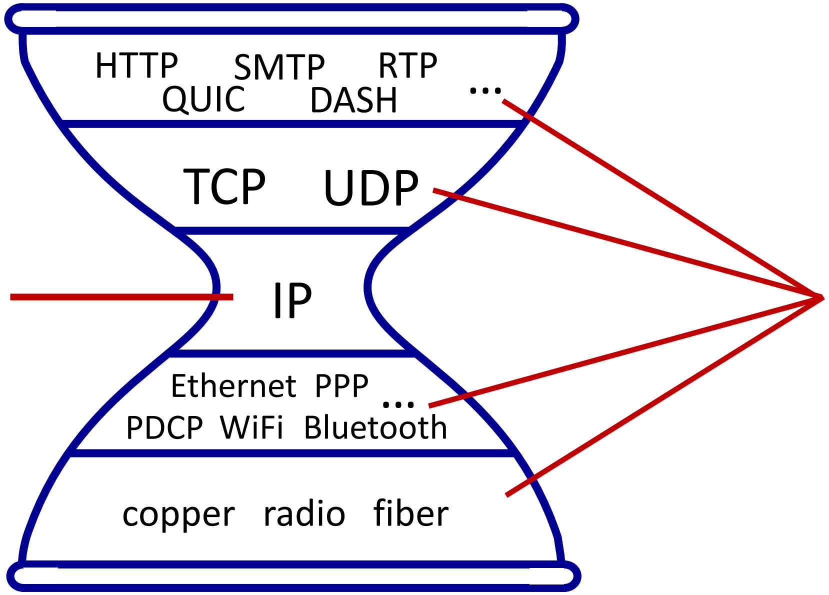


(a) A packet crossing different networks.

(b) Network and link layer protocol processing.

Internet Protocol (IP)

Implemented by every (billions) Internet-connected device

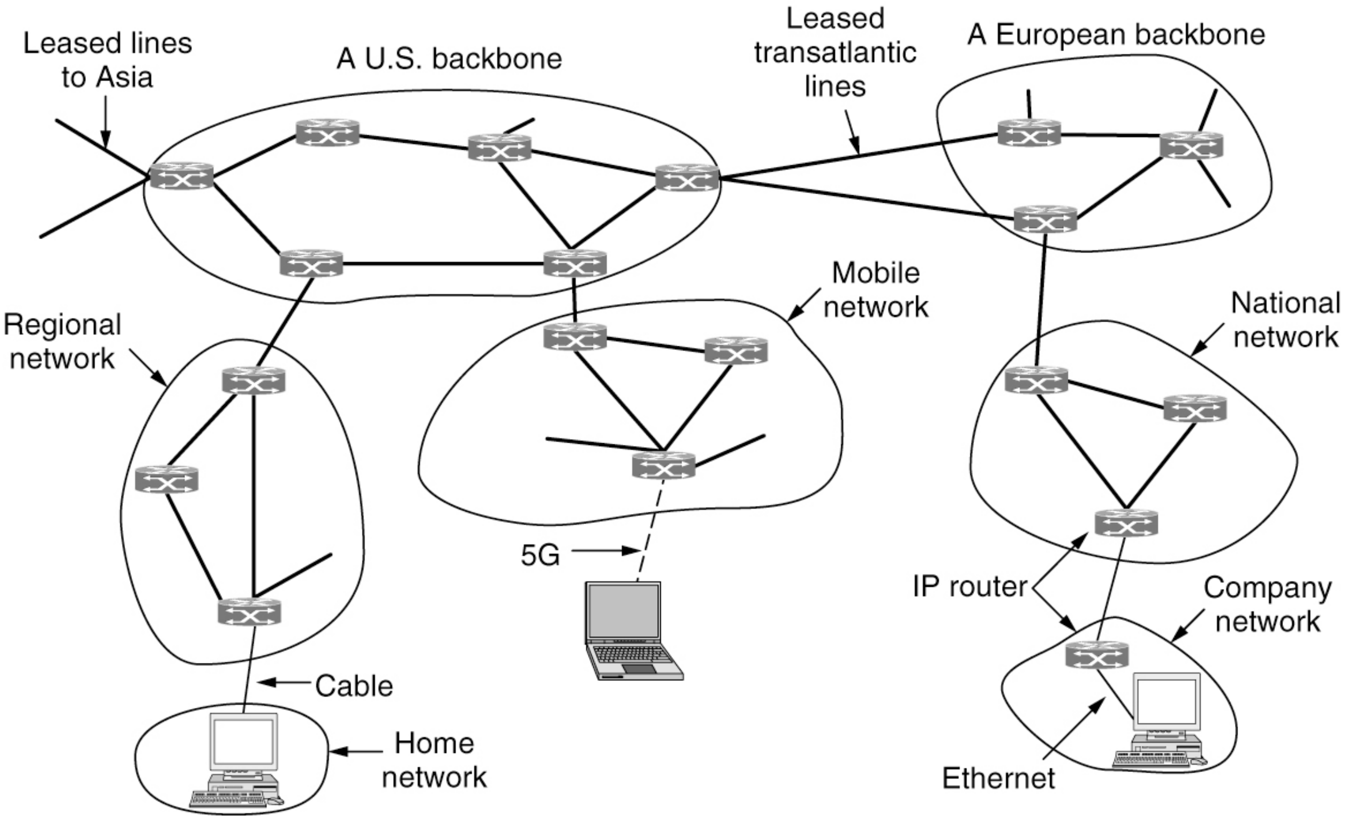


Many protocols in the physical, link, transport, and application layers

Internet Network Layer

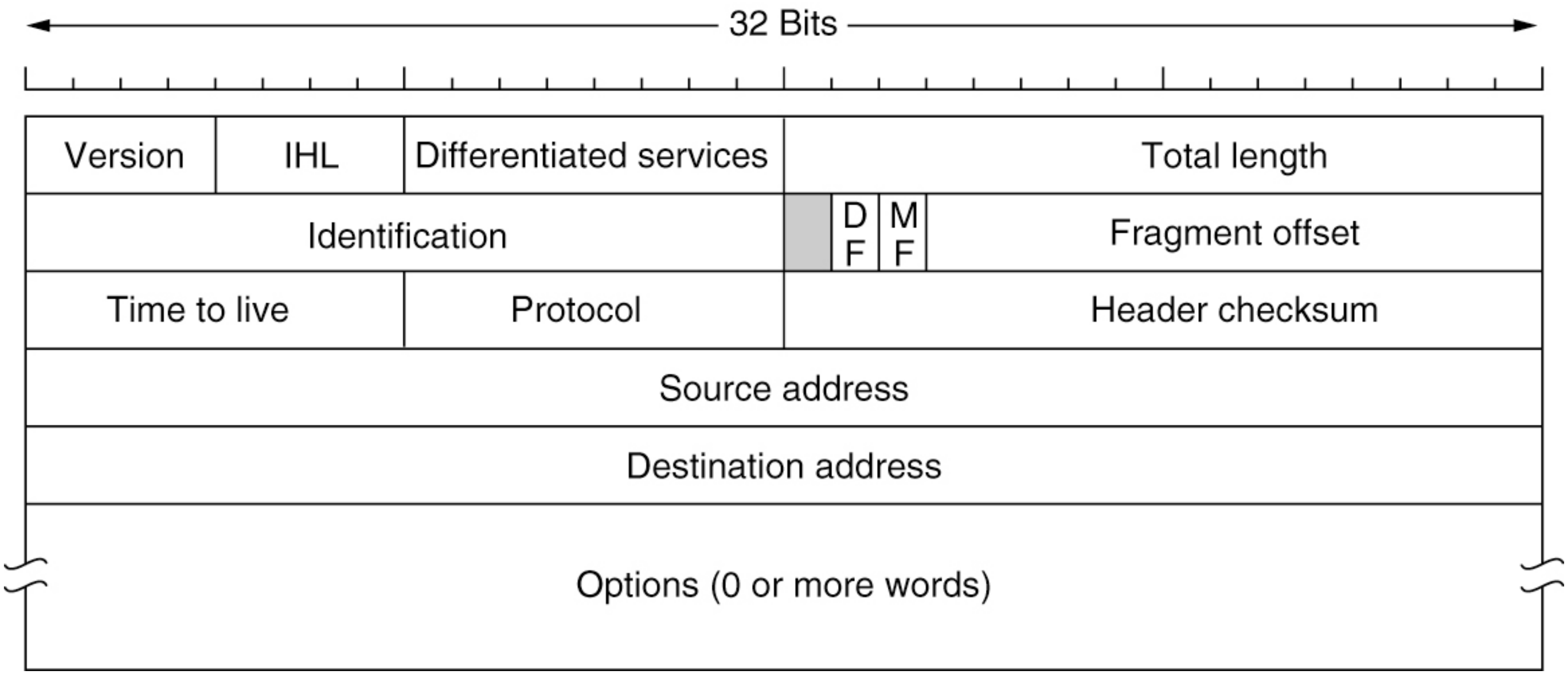
- IP Version 4 Protocol
- IP Addresses
- Internet control protocols
- OSPF—An interior gateway routing protocol
- BGP—The exterior gateway routing protocol

The Network Layer in the Internet



The Internet is an interconnected collection of networks

IP Version 4 Protocol



32 Bits

Version | IHL | Differentiated services | Total length

Identification | D F | M F | Fragment offset

Time to live | Protocol | Header checksum

Source address

Destination address

Options (0 or more words)

The Internet Protocol version 4 (IPv4) header

Header Checksum

- The checksum field is 16 bits
- Add all the 16 bit halfwords using one's complement addition
- Example:

  4500 0073 0000 4000 4011 B861 C0A8 0001 C0A8 00C7

  - The sum excluding the checksum is

  2479C → 479C + 2 = 497E

- The checksum is the one's complement of the result

  0100 0111 1001 1100

  1011 1000 0110 0011 → B861

- To verify the header, add all the header halfwords

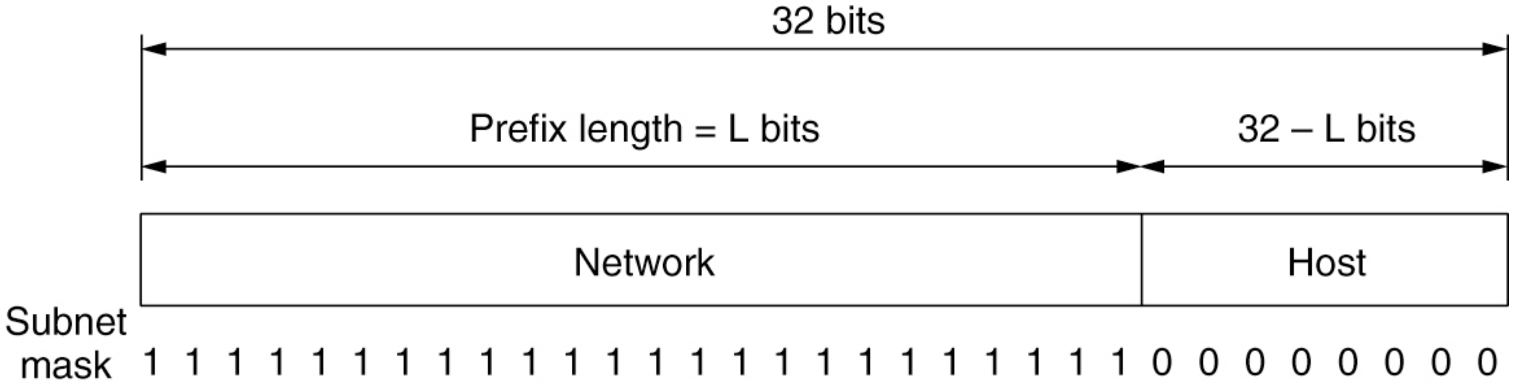
  2FFFD → FFFD + 2 = FFFF

- Taking the one's complement gives 0000 so the header is verified

IP Addresses

- Prefixes
  - A contiguous block of IP address space
- Subnets
- CIDR—Classless InterDomain Routing
- Classful and special addressing
- NAT—Network Address Translation

Prefixes



32 bits

Prefix length = L bits

32 – L bits

Network

Host

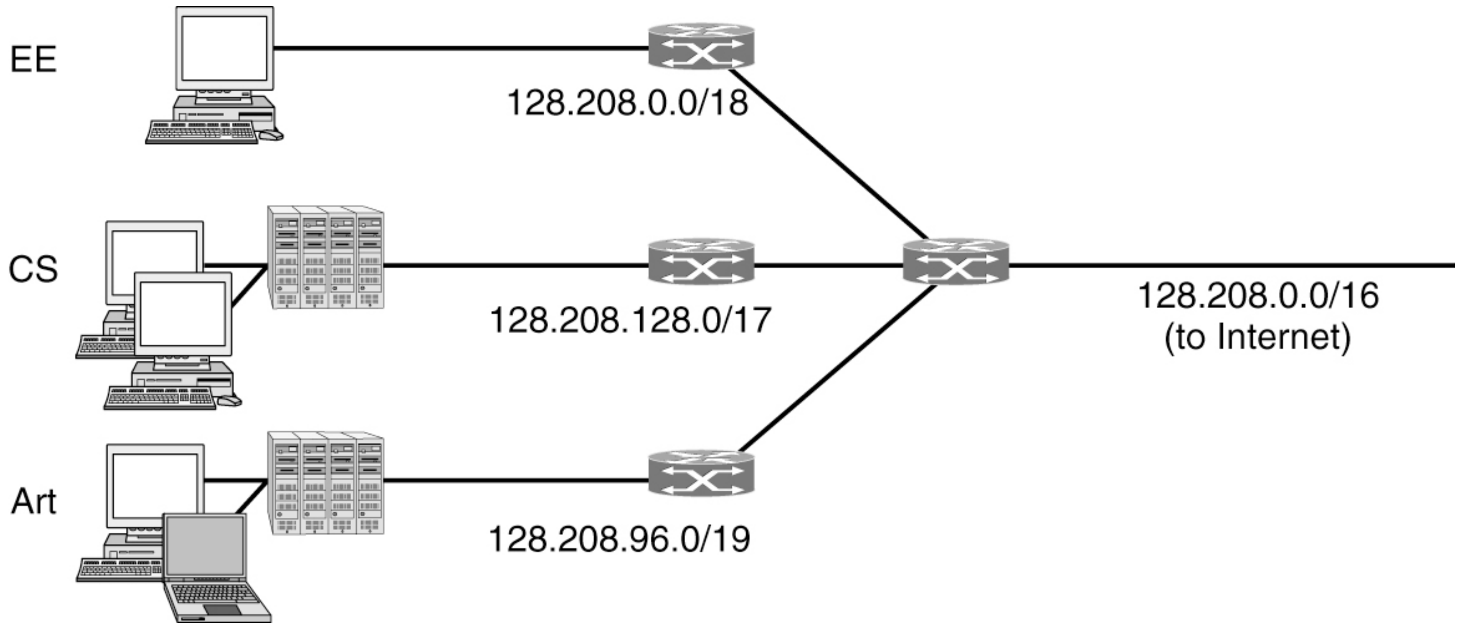
Subnet mask 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0

A prefix and a subnet mask

IP Address Allocation

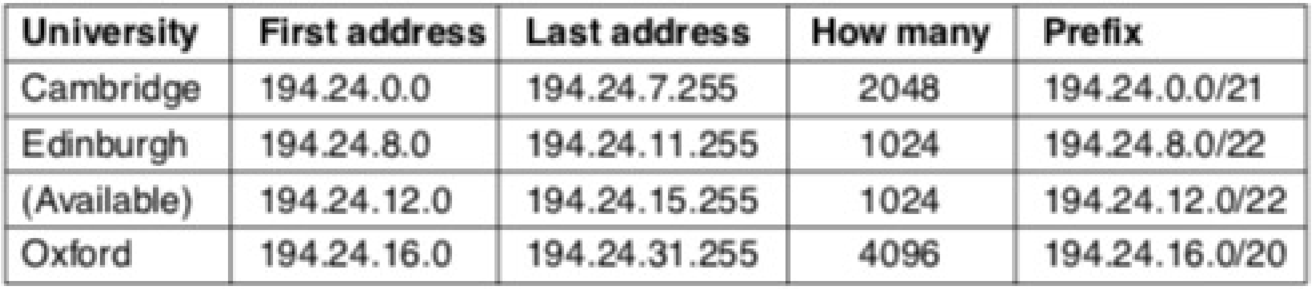
- The Internet Corporation for Assigned Names and Numbers (ICANN) is a nonprofit organization responsible for coordinating key parts of the global Internet naming and numbering infrastructure
- Coordinates global IP address allocation through Regional Internet Registries (RIRs)
    - ARIN (North America)
    - RIPE NCC (Europe)
    - APNIC (Asia-Pacific)
    - LACNIC (Latin America)
    - AFRINIC (Africa)
- Global IPv4 exhaustion occurred in 2011
- All RIRs have since exhausted their normal IPv4 reserves so they are now
    - Recycled
    - Strictly rationed
- They are also transferred on secondary markets (bought, sold, or transferred between organizations)
- Growth today depends on IPv6

Subnets



Splitting an IP prefix into separate networks with subnetting

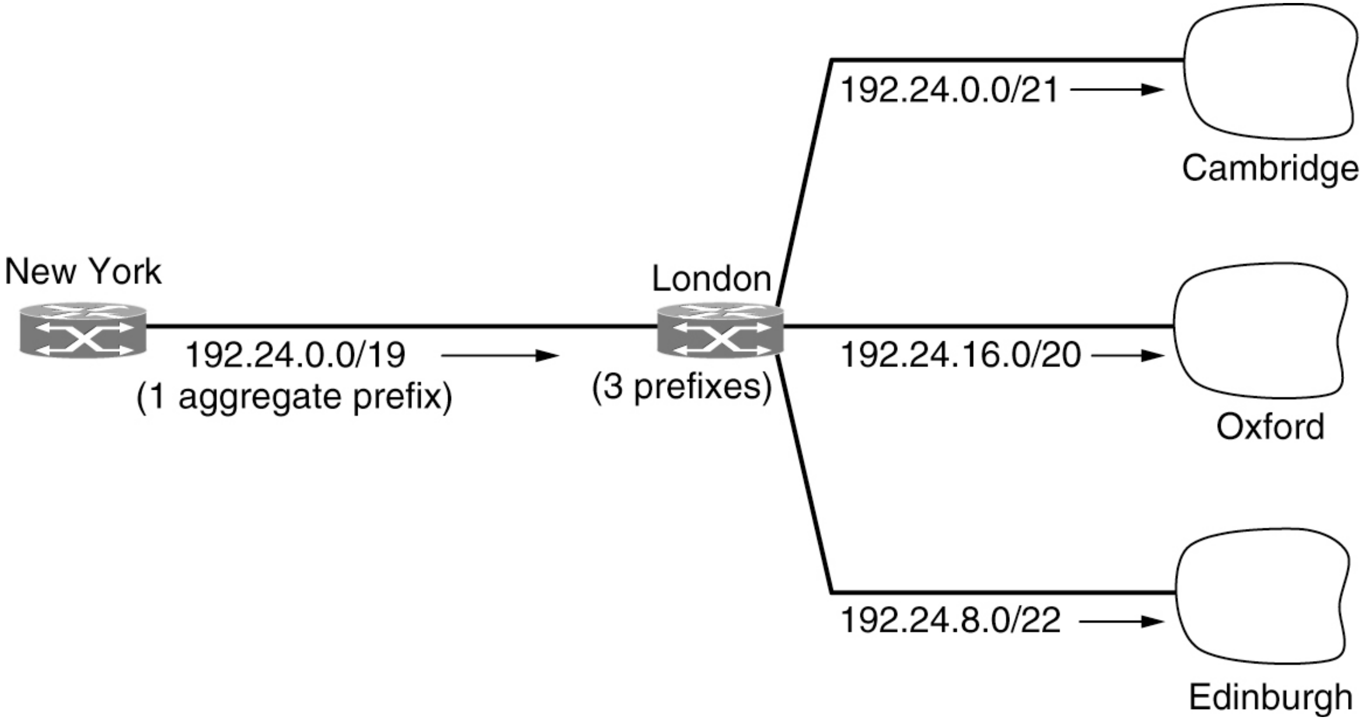
CIDR—Classless InterDomain Routing



University	First address	Last address	How many	Prefix
Cambridge	194.24.0.0	194.24.7.255	2048	194.24.0.0/21
Edinburgh	194.24.8.0	194.24.11.255	1024	194.24.8.0/22
(Available)	194.24.12.0	194.24.15.255	1024	194.24.12.0/22
Oxford	194.24.16.0	194.24.31.255	4096	194.24.16.0/20

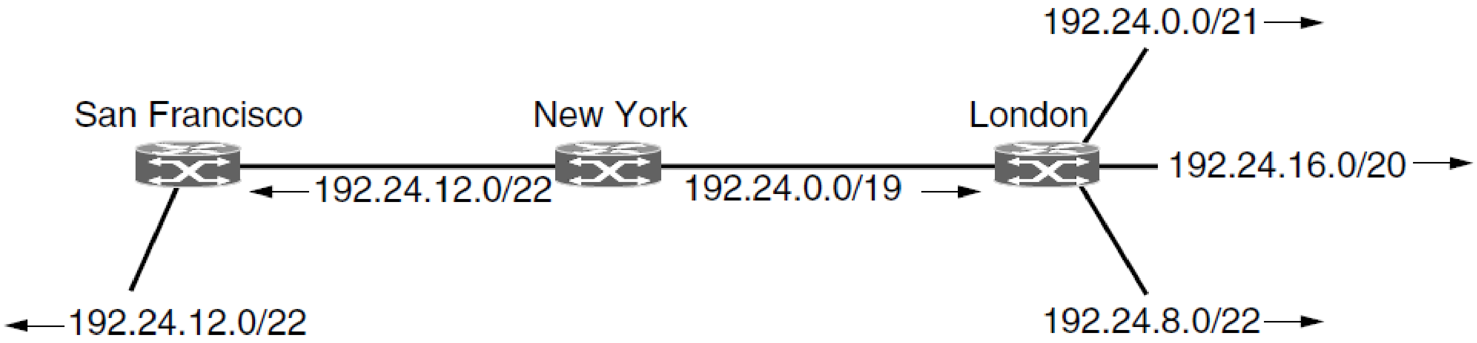
A set of IP address assignments

CIDR—Classless InterDomain Routing



Aggregation of IP prefixes

CIDR—Classless InterDomain Routing



Longest matching prefix routing at the New York router

Longest Matching Prefix

When looking for the forwarding table entry for a destination address, use the longest address prefix that matches the destination address

Destination Address Range	Link interface
11001000 00010111 00010*** ********	0
11001000 00010111 00011000 ********	1
11001000 00010111 00011*** ********	2
otherwise	3

examples:

11001000 00010111 00010110 10100001 which interface?

11001000 00010111 00011000 10101010 which interface?

Longest Matching Prefix

When looking for the forwarding table entry for a destination address, use the longest address prefix that matches the destination address

Destination Address Range	Link interface
11001000 00010111 00010*** ********	0
11001000 00010111 00011000 ********	1
11001000 00010111 00011*** ********	2
otherwise	3

match!

examples:

11001000 00010111 00010110 10100001 which interface?

11001000 00010111 00011000 10101010 which interface?

Longest Matching Prefix

When looking for the forwarding table entry for a destination address, use the longest address prefix that matches the destination address

Destination Address Range	Link interface
11001000 00010111 00010*** ********	0
11001000 00010111 00011000 ********	1
11001000 00010111 00011*** ********	2
otherwise	3

match!

examples:

11001000 00010111 00010110 10100001 which interface?

11001000 00010111 00011000 10101010 which interface?

Longest Matching Prefix

When looking for the forwarding table entry for a destination address, use the longest address prefix that matches the destination address

Destination Address Range	Link interface
11001000 00010111 00010*** ********	0
11001000 00010111 00011000 ********	1
11001000 00010111 00011*** ********	2
otherwise	3

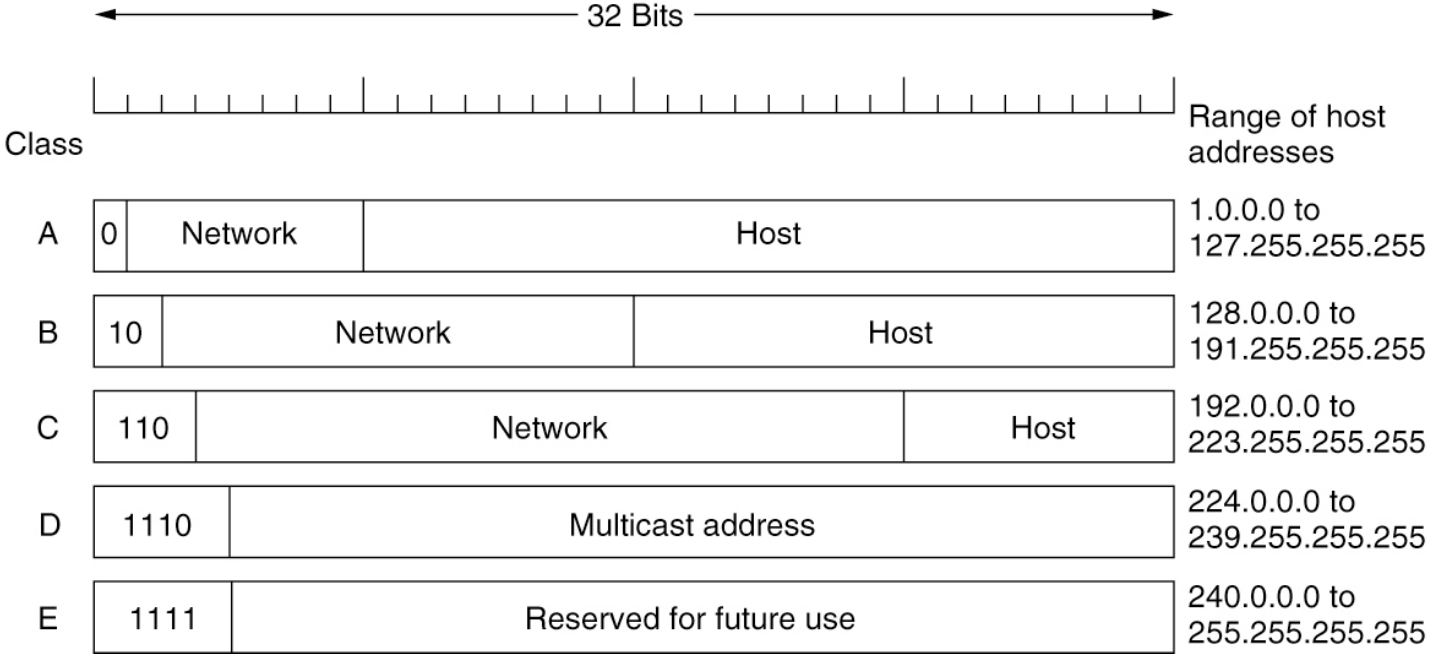
match!

examples:

11001000 00010111 00010110 10100001 which interface?

11001000 00010111 00011000 10101010 which interface?

Classful Addressing

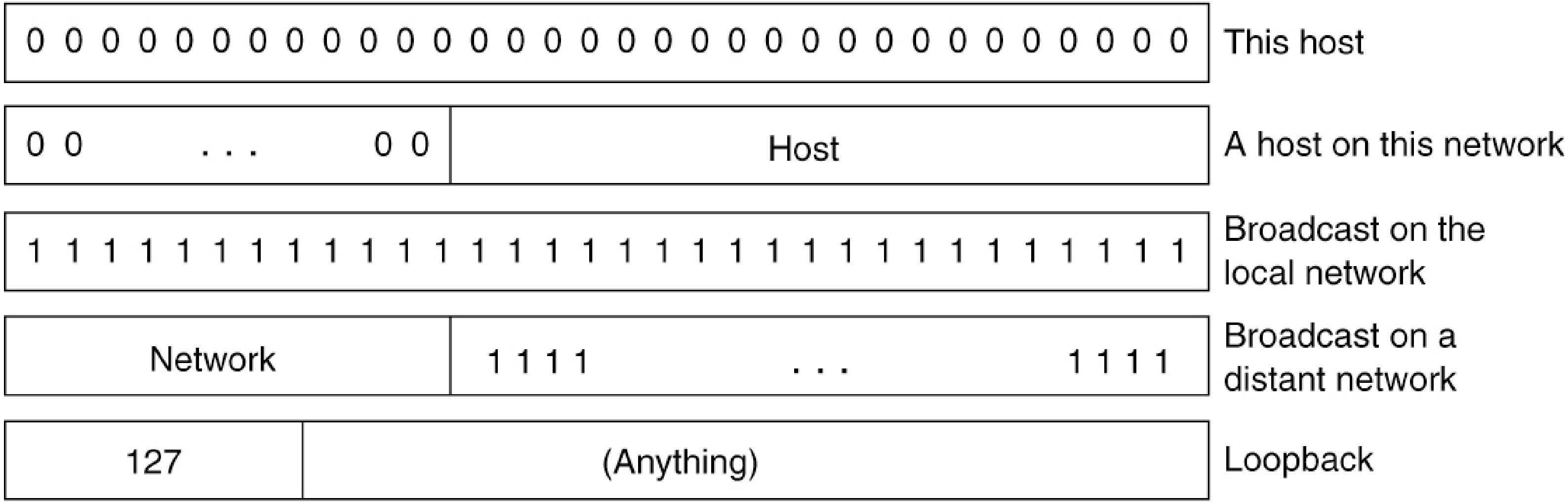


32 Bits

Class				Range of host addresses
A	0	Network	Host	1.0.0.0 to 127.255.255.255
B	10	Network	Host	128.0.0.0 to 191.255.255.255
C	110	Network	Host	192.0.0.0 to 223.255.255.255
D	1110	Multicast address		224.0.0.0 to 239.255.255.255
E	1111	Reserved for future use		240.0.0.0 to 255.255.255.255

IP address formats

Special Addressing



0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 — This host

0 0 . . . 0 0 | Host — A host on this network

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 — Broadcast on the local network

Network | 1 1 1 1 . . . 1 1 1 1 — Broadcast on a distant network

127 | (Anything) — Loopback

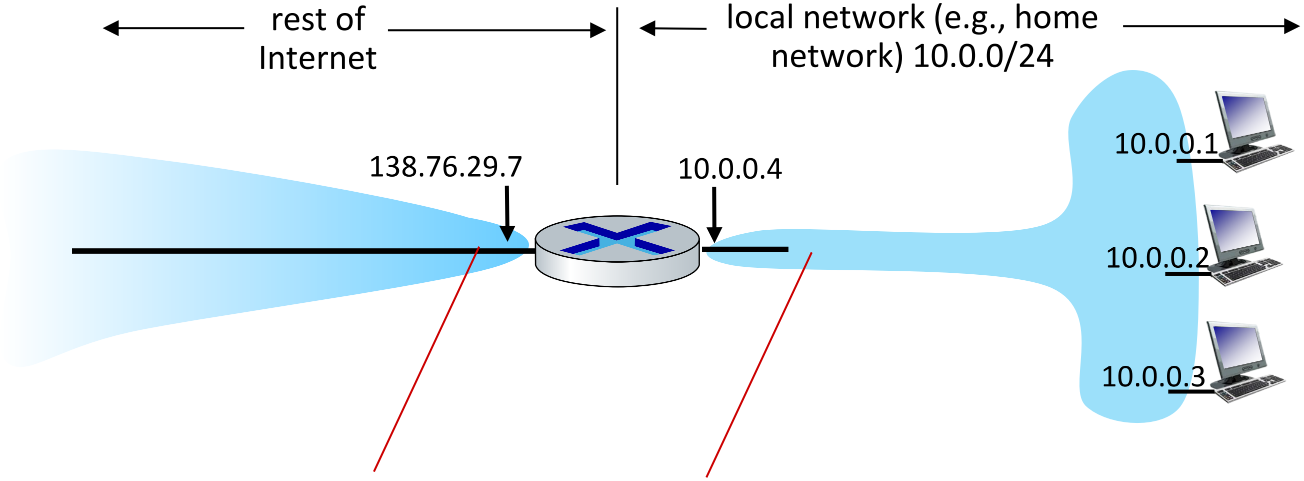
Special IP addresses

NAT—Network Address Translation

- All devices in the local network have 32-bit addresses in a private IP address space (10/8, 172.16/12, 192.168/16 prefixes) that can only be used in the local network
- Advantages
  - Just one IP address needed from provider ISP for all devices
  - Can change the addresses of hosts in local network without notifying outside world
  - Can change ISP without changing addresses of devices in local network
  - Security: devices inside the local network are not directly addressable or visible by the outside world

NAT—Network Address Translation

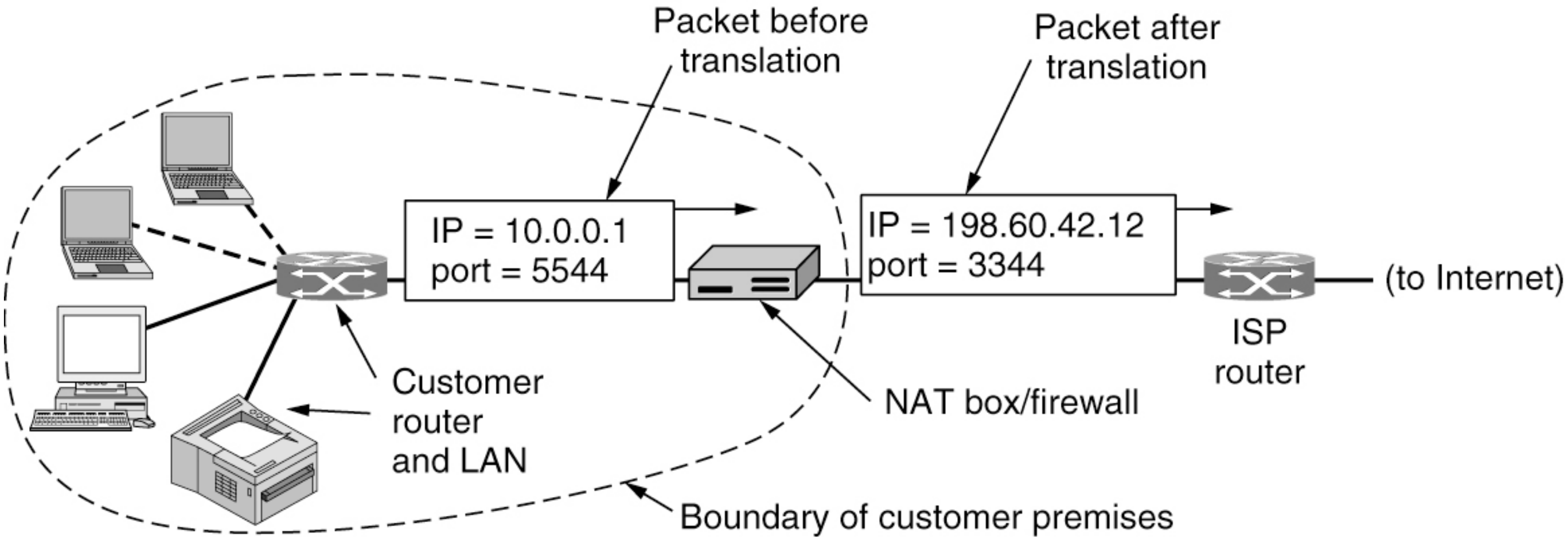
NAT: all devices in the local network share just one IPv4 address as far as the outside world is concerned



All packets leaving the local network have the same source NAT IP address: 138.76.29.7, but different source port numbers

Packets with source or destination in this network have a 10.0.0/24 address for source, destination (typical)

NAT—Network Address Translation



Placement and operation of a NAT box

NAT—Network Address Translation

Implementation: NAT router must (transparently)

- Outgoing packets: replace (source IP address, port #) of every outgoing packet to (NAT IP address, new port #)
  - Remote clients/servers will respond using (NAT IP address, new port #) as the destination address
- Place in NAT translation table every (source IP address, port #) to (NAT IP address, new port #) translation pair
- Incoming packets: replace (NAT IP address, new port #) in destination fields of every incoming packet with the corresponding (source IP address, port #) stored in the NAT table

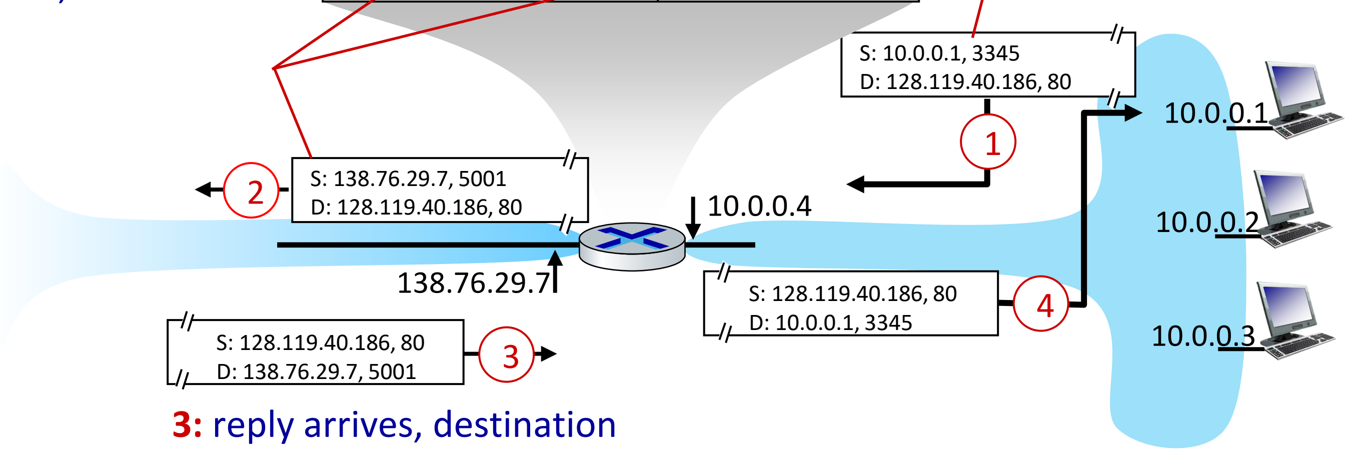
NAT—Network Address Translation

2: NAT router changes packet source address from 10.0.0.1, 3345 to 138.76.29.7, 5001, updates table

NAT translation table

WAN side addr	LAN side addr
138.76.29.7, 5001	10.0.0.1, 3345
……	……

1: host 10.0.0.1 sends packet to 128.119.40.186, 80



3: reply arrives, destination address: 138.76.29.7, 5001

NAT—Network Address Translation

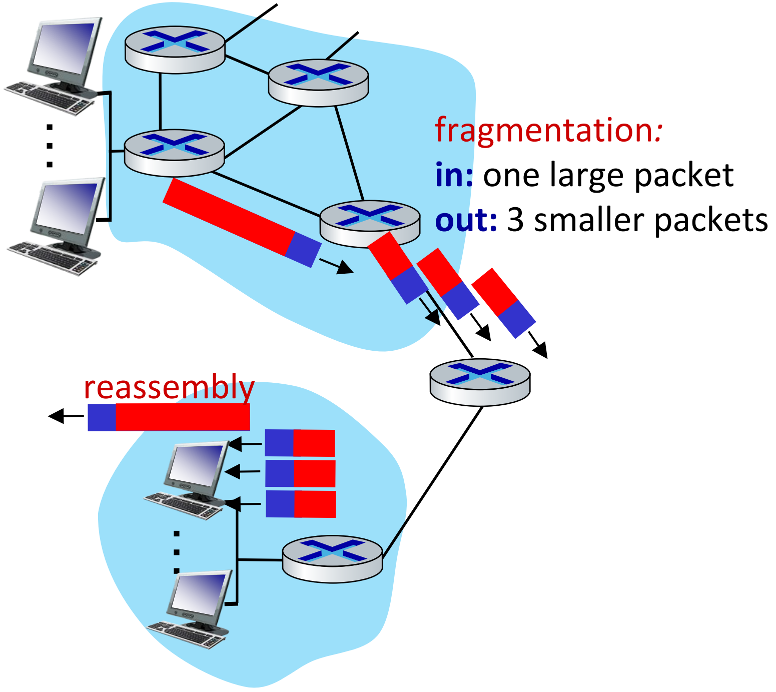
- NAT has been controversial
  - Routers should only process at the network layer
  - Violates layer independence
  - Address shortage should be solved by IPv6
  - Violates end-to-end argument (port # manipulation by network-layer device)
  - Processes may use a protocol other than TCP or UDP
- But NAT is here to stay
  - Extensively used in home and institutional networks as well as 4G/5G cellular networks

Packet Fragmentation

- Each network or link imposes some maximum size on its packets
- These limitations have various causes such as
    - Hardware
        - Size of an Ethernet frame
    - Operating system
        - All buffers are size 512 bytes
    - Protocol
        - The number of bits in the packet length field
    - National or international standards
    - Desire to reduce retransmissions due to errors
    - Desire to prevent a packet from occupying the channel too long

Packet Fragmentation

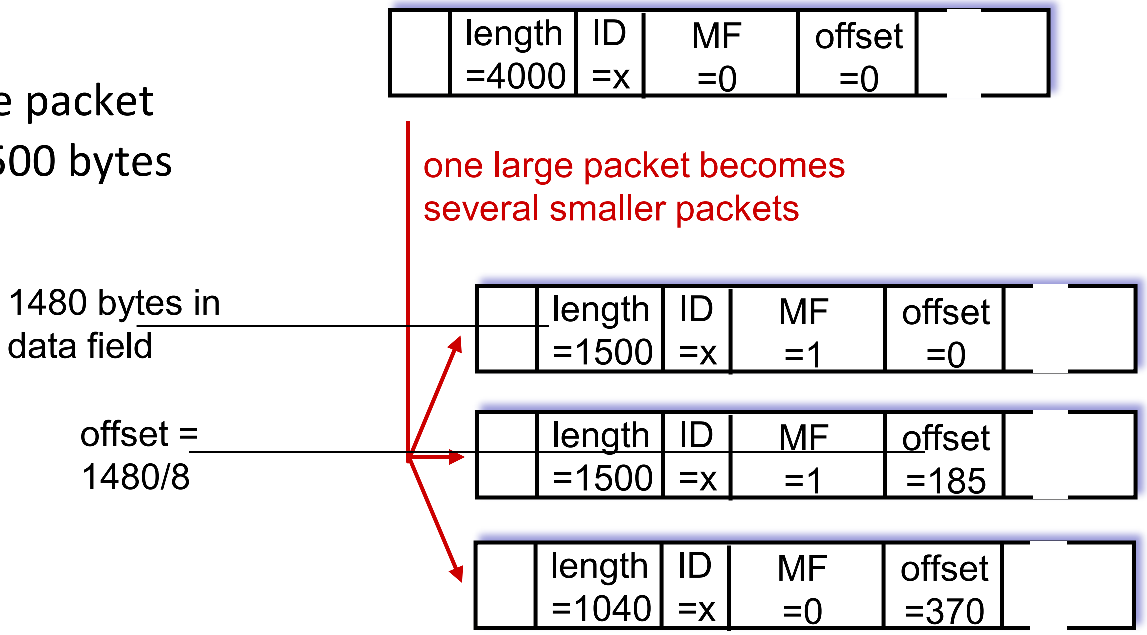
- Network links have a Maximum Transmission Unit (MTU)
  - Largest possible link-level frame
  - Different link types, different MTUs
- A large IP packet is divided (fragmented) so it becomes several packets
  - Reassembled only at the destination



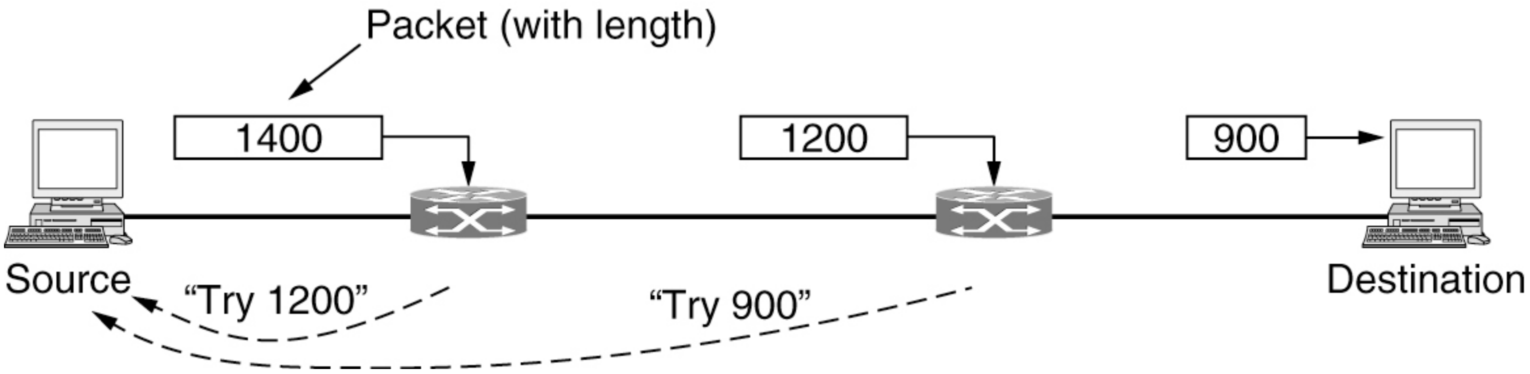
IP Packet Fragmentation

Example

- 4000 byte packet
- MTU = 1500 bytes



Packet Fragmentation



Path MTU discovery

Differentiated Services

- DS field (8 bits) in the IP header
- DSCP (6 bits) provides traffic prioritization using traffic classes
  - Edge routers classify & mark packets (set DSCP)
  - Core routers apply Per-Hop Behavior (PHB) policies for
    - Prioritization
    - Queueing
    - Drop precedence
  - Used in enterprise WANs, ISP backbones, MPLS networks, and data centers
  - DSCP enables class-based QoS by marking packets and applying priority handling at each hop
- ECN (2 bits) is used to signal network congestion without dropping packets
  - ECN reduces packet loss, improves latency, and avoids unnecessary retransmissions
  - Particularly useful in data centers and low-latency networks

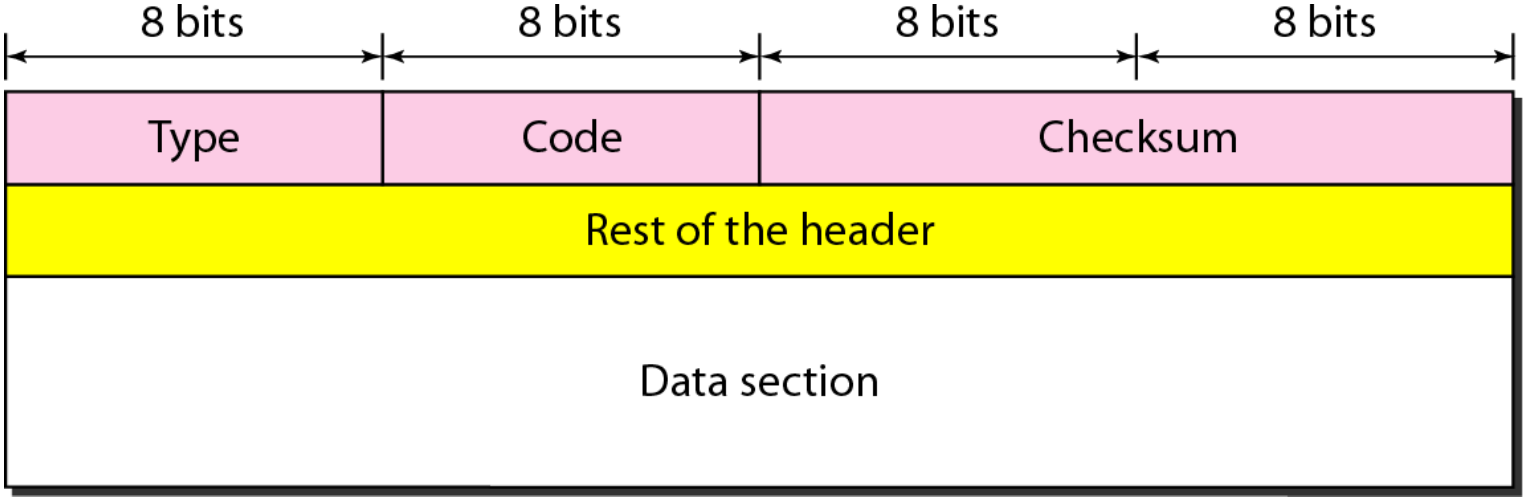
Internet Control Protocols

- ICMP—The Internet Control Message Protocol
- ARP—Address Resolution Protocol
- DHCP—The Dynamic Host Configuration Protocol

ICMP—Internet Control Message Protocol

- IP has no error-reporting or error-correcting mechanisms
- The IP protocol also lacks a mechanism for host and management queries
- ICMP has been designed to compensate for these deficiencies
- It is a companion to the IP protocol
- Protocol = 1

ICMP Format



8 bits | 8 bits | 8 bits | 8 bits

Type | Code | Checksum

Rest of the header

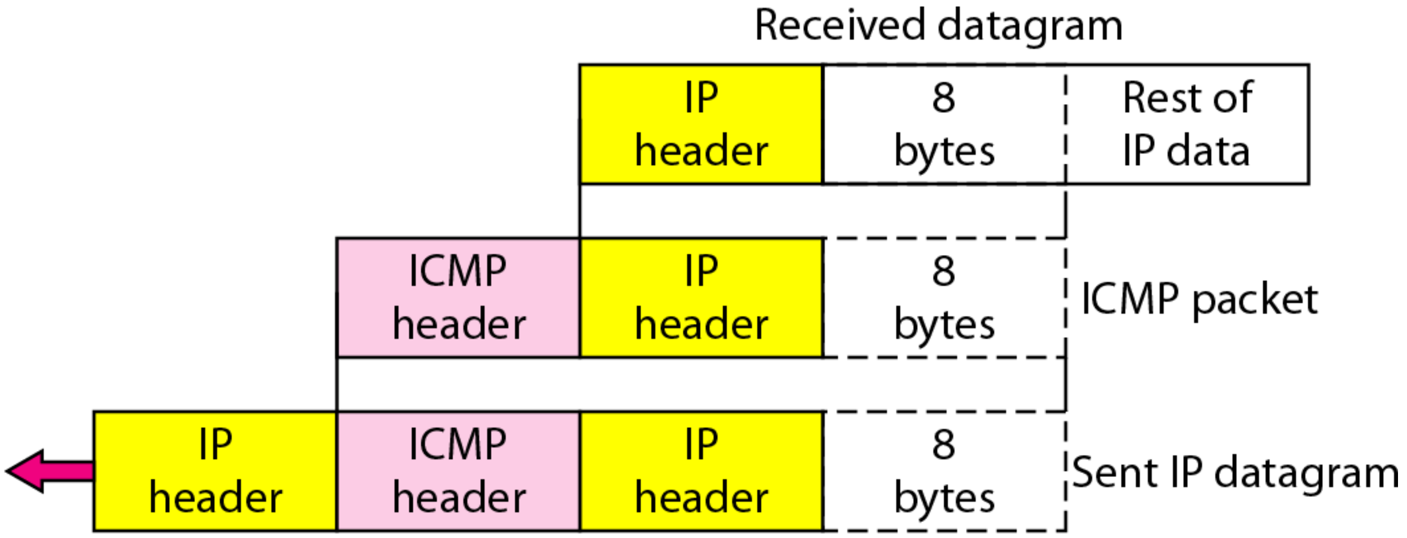
Data section

- If there is no additional data, the 4 bytes in the rest of the header are set to zero
- Each ICMP messages is at least 8 bytes long

ICMP Messages

- ICMP messages are divided into two broad categories
    - Error reporting messages
    - Query messages
- Type field specifies the kind of message
- In error reporting messages, the data section carries information for finding the original packet that had the error
    - ICMP sends error messages to the original source
- In query messages, the data section carries extra information based on the type of query

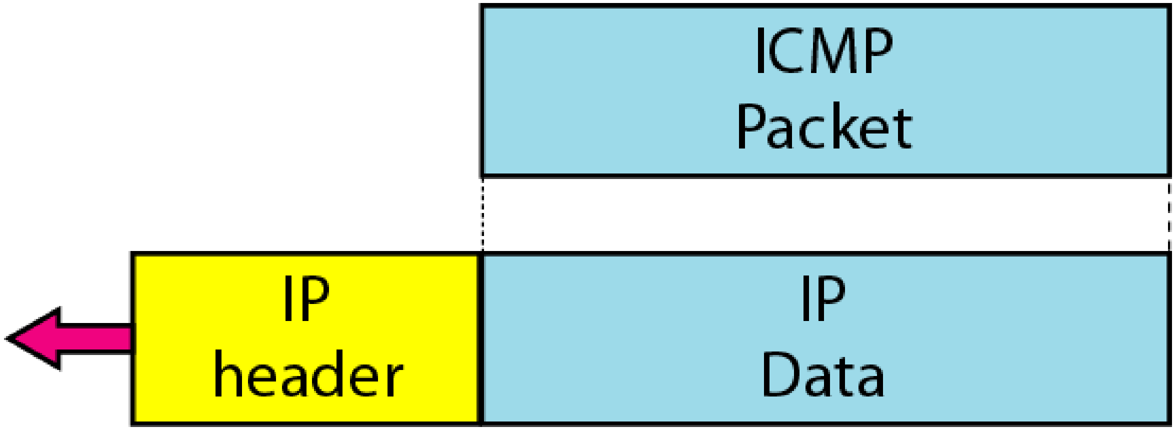
ICMP Error Messages



ICMP Query Messages

- Echo and Echo Reply
    - The echo and echo reply messages are used for diagnostic purposes.
- Timestamp Request/Reply
    - Two machines (hosts or routers) can use the timestamp request and timestamp reply messages to determine the round-trip time needed for an IP packet to travel between them. It can also be used to synchronize the clocks in two machines.
- Router Advertisement/Solicitation
    - A host that wants to send data to a host on another network needs to know the address of routers connected to its own network.

ICMP Query Messages



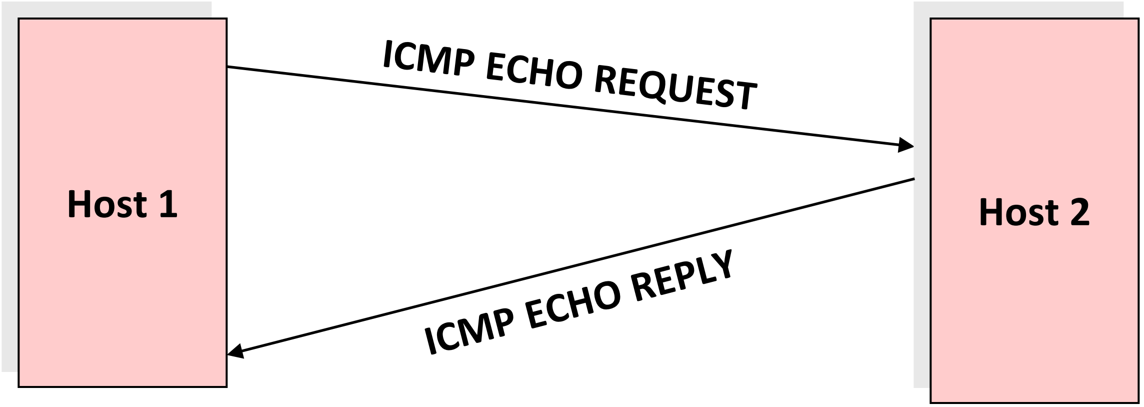
ICMP—Internet Control Message Protocol

Message	**Type**	**Description**
Destination unreachable	3	Packet could not be delivered
Time Exceeded	11	Time to live field hit 0
Parameter problem	12	Invalid header field
Redirect	5	Teach a router about geography
Echo and echo reply	8,0	Check if a machine is alive
Timestamp request/reply	13,14	Same as Echo but with timestamp
Router advertisement/solicitation	9,10	Find a nearby router

The principal ICMP message types

Ping

- Each ping is translated into an ICMP Echo Request
- The Pinged host responds with an ICMP Echo Reply



Ping Example

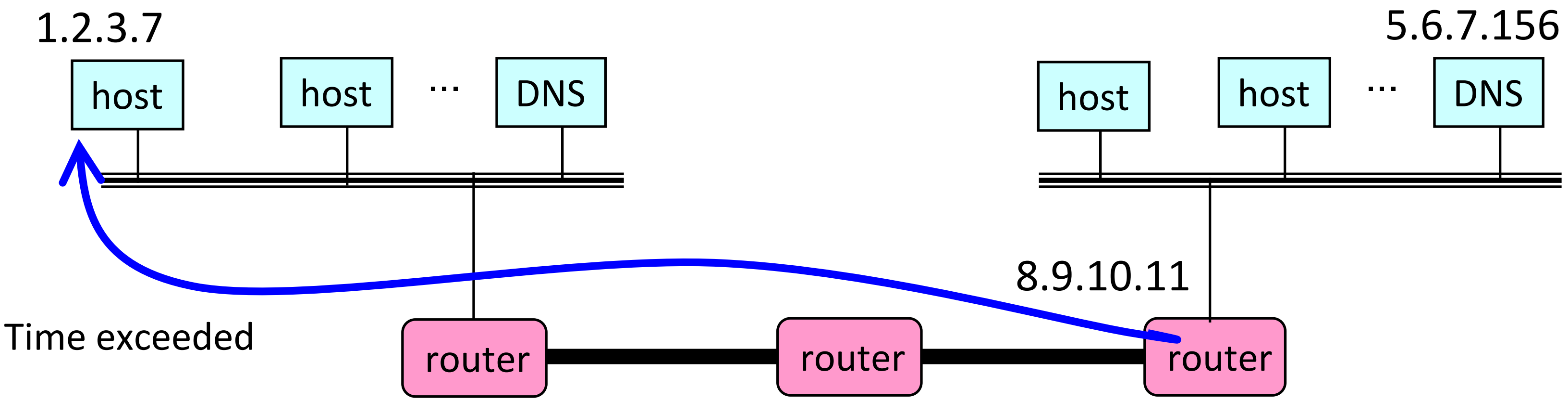
```
C:\Users\agullive>ping google.com

Pinging google.com [142.251.46.78] with 32 bytes of data:
Reply from 142.251.46.78: bytes=32 time=12ms TTL=119
Reply from 142.251.46.78: bytes=32 time=20ms TTL=119
Reply from 142.251.46.78: bytes=32 time=19ms TTL=119
Reply from 142.251.46.78: bytes=32 time=19ms TTL=119

Ping statistics for 142.251.46.78:
Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
Minimum = 12ms, Maximum = 20ms, Average = 17ms
```

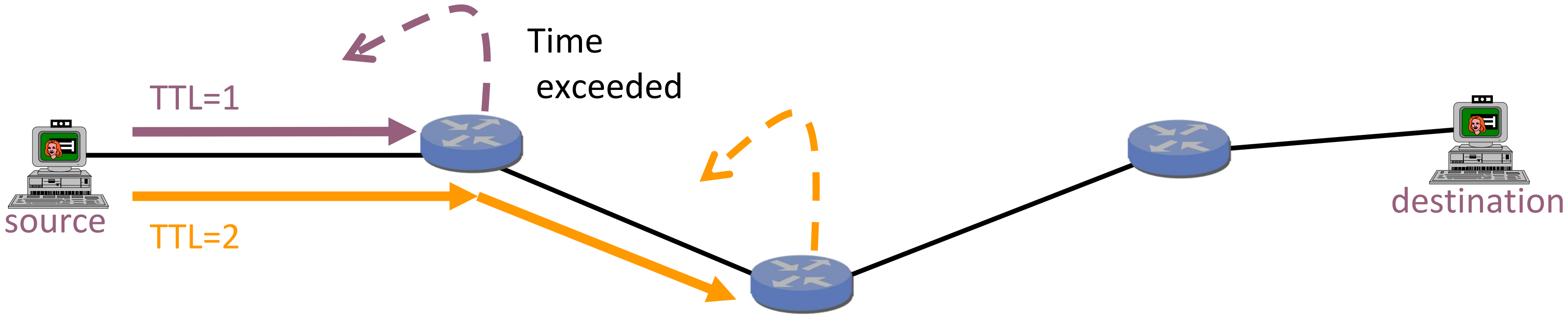
Time Exceeded

- Host sends an IP packet
  - Each router decrements the Time To Live (TTL)
- If TTL reaches 0
  - Router sends an ICMP time exceeded message back to the source



Traceroute: Exploiting Time Exceeded

- TTL field in IP packet header
  - Source sends a packet with a TTL of *n*
  - Each router along the path decrements the TTL
  - Time exceeded message sent when TTL reaches 0
- Traceroute exploits this TTL behavior
  - Send packets with TTL=1, 2, … and record source of Time exceeded message



Traceroute Example

```
C:\Users\agullive>tracert www.google.com

Tracing route to www.google.com [142.251.33.196] over a maximum of 30 hops:

 1     7 ms     1 ms     1 ms  192.168.0.1
 2   475 ms   157 ms   111 ms  24.69.160.1
 3    16 ms    15 ms    11 ms  rd1lp-be107-1.gv.shawcable.net [64.59.162.229]
 4    30 ms    32 ms    34 ms  24.244.58.77
 5    38 ms    12 ms    13 ms  24.244.61.97
 6    14 ms    12 ms    10 ms  rd1lp-be3.gv.shawcable.net [24.244.61.106]
 7    12 ms    13 ms    15 ms  rc1wt-be40.wa.shawcable.net [66.163.68.18]
 8    25 ms    17 ms    21 ms  72.14.221.102 (Google)
 9    27 ms    15 ms    22 ms  142.251.229.135
10    30 ms    30 ms    16 ms  142.251.50.243
11    19 ms    12 ms    21 ms  iad23s96-in-f4.1e100.net [142.251.33.196]

Trace complete.
```

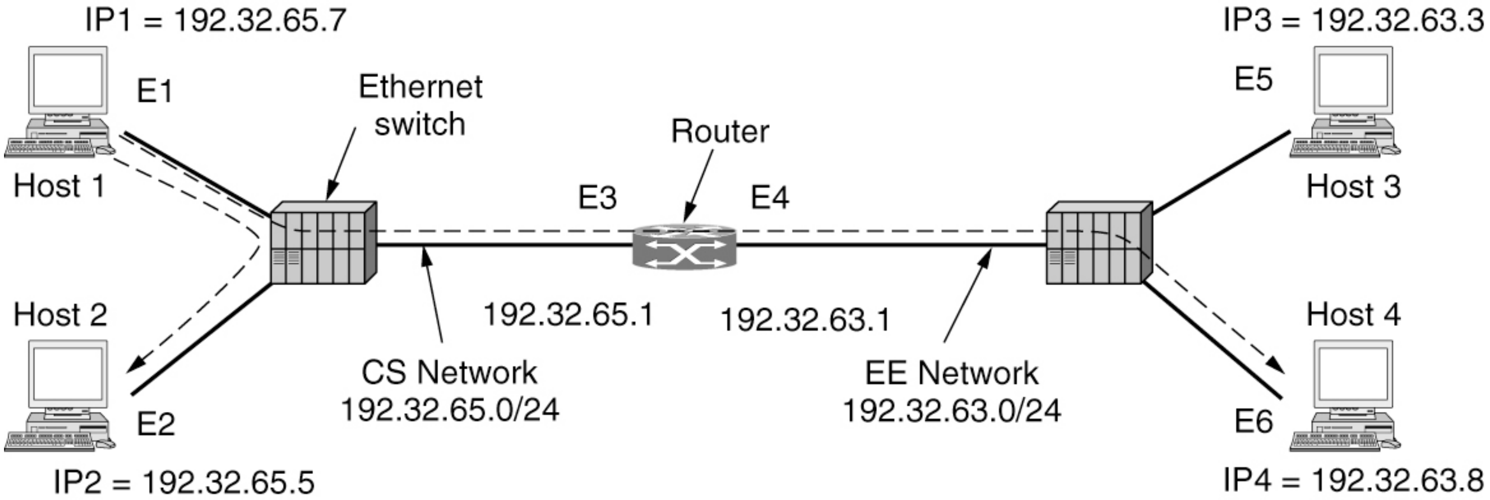
Path MTU Discovery

- The goal is to determine the largest packet size (MTU) that can travel from sender to receiver without fragmentation
    - IP fragmentation is inefficient and undesirable
- Sender transmits packets with the DF bit set
- If a router encounters a link with a smaller MTU
    - Router drops the packet
    - Router sends an ICMP Fragmentation Needed message
        - Type 3: Destination Unreachable
        - Code 4: Fragmentation Needed and DF Set
- The sender reduces the packet size and retransmits
- Process repeats until packets pass successfully
- Sender then transmits packets ≤ Path MTU for that connection

MAC Addresses

- MAC address allocation administered by IEEE
- Manufacturers buy a portion of the MAC address space (to assure uniqueness)
- Analogy
  - MAC address: Social Insurance Number
  - IP address: postal address
- MAC flat address: portability
  - Can move an interface from one LAN to another
- IP addresses are not portable: they depend on the IP subnet to which the node is attached

ARP—Address Resolution Protocol



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

Two switched Ethernet LANs joined by a router

IP Addresses: How To Get One?

- How does a network get an IP address for itself (network part of the address)?
  - Allocated a portion of the ISP provider address space
- How does a host get an IP address within its network (host part of the address)?
  - Hard-coded by a system administrator in the config file
  - Dynamically get an address from a DHCP server

DHCP—Dynamic Host Configuration Protocol

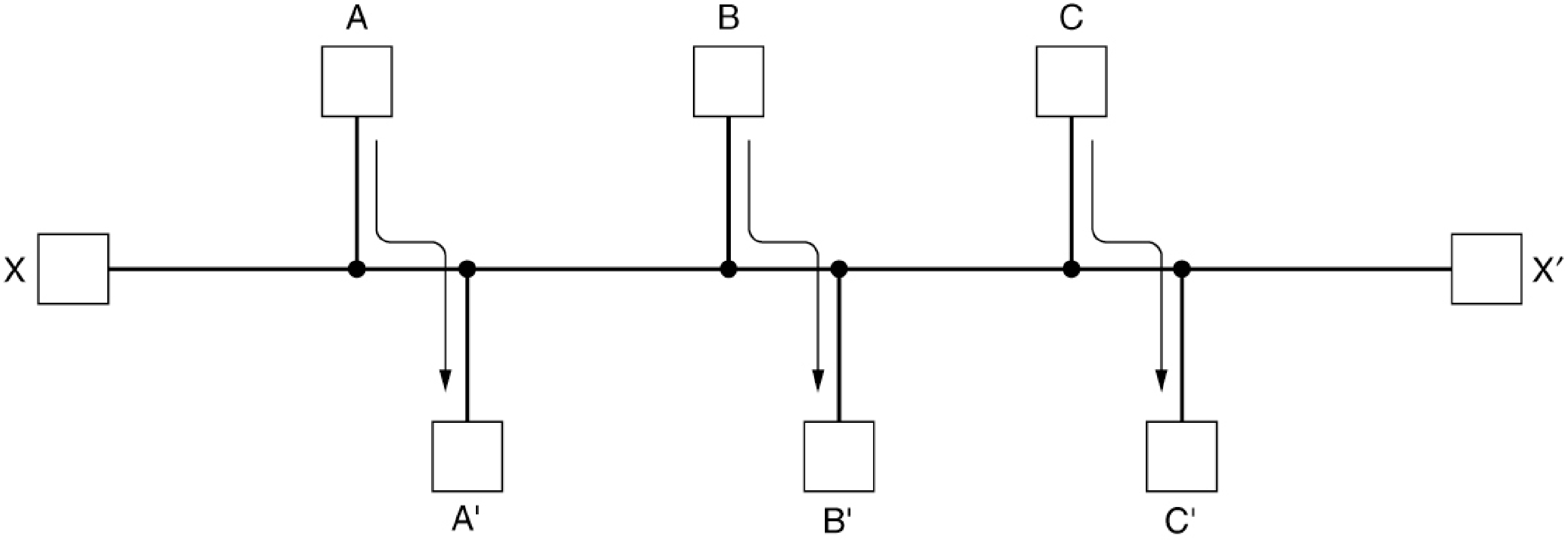
A host dynamically obtains an IP address from the network DHCP server when it joins the network

- Host broadcasts DHCP discover message
- DHCP server responds with DHCP offer message
- If a host remembers and wishes to reuse a previously allocated network address (renew lease)
  - Host requests IP address: DHCP request message
  - DHCP server sends address: DHCP ack message
- Allows reuse of addresses (only hold address while connected/on)
- Support for mobile users who join/leave the network

Routing Algorithms

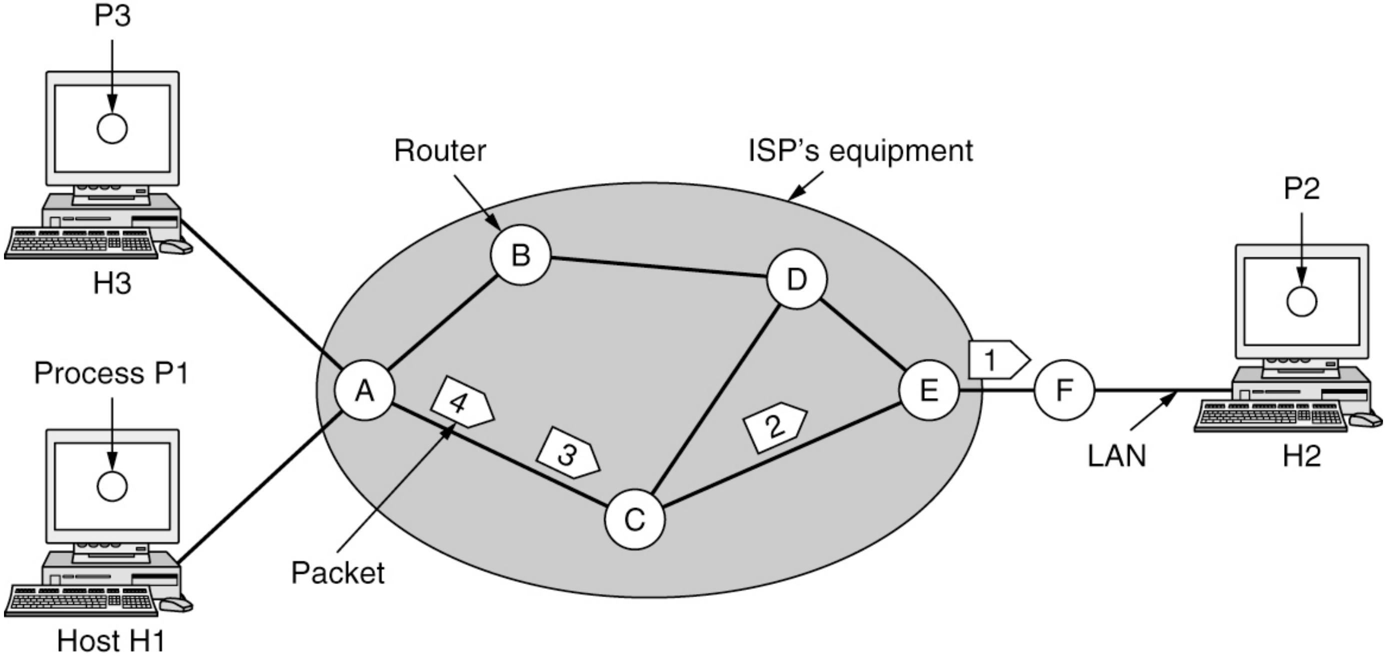
- Shortest path (Dijkstra’s) algorithm
- Flooding
- Distance vector routing
- Link state routing

Routing Algorithms

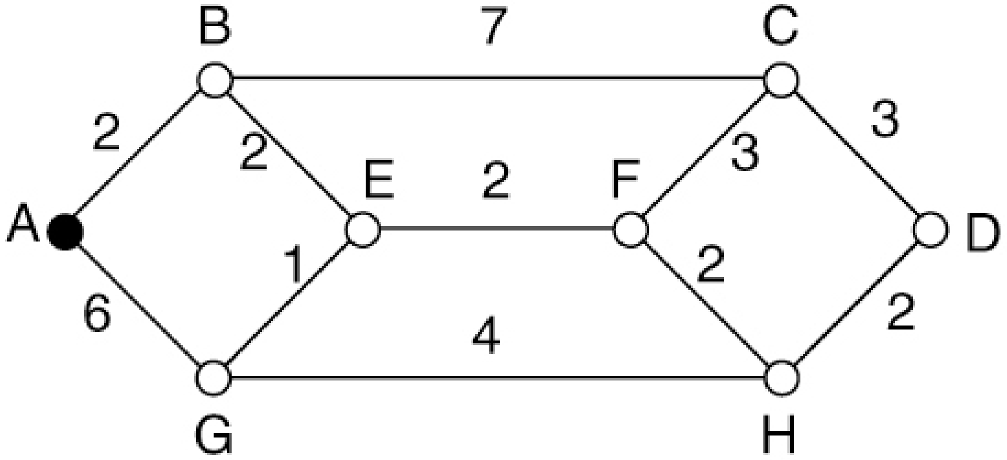


Network with a conflict between fairness and efficiency

Routing Algorithms

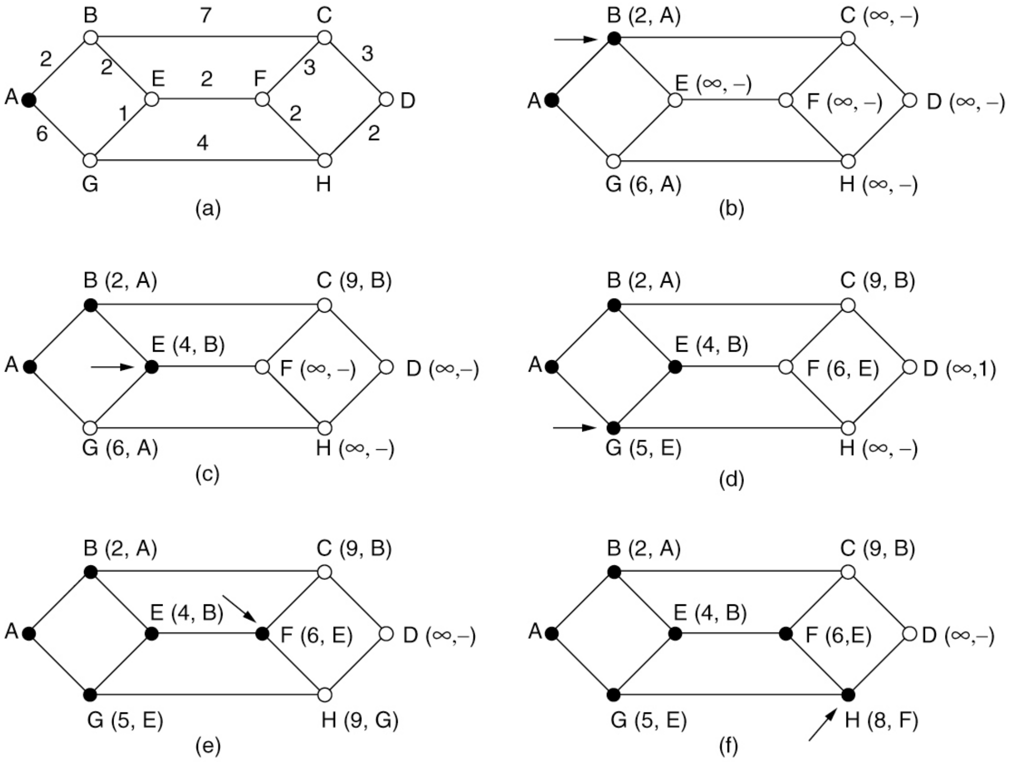


Shortest Path



- If the number of hops in the path is the metric, paths ABC and ABE are equally long
- If geographic distance is the metric, path ABC is longer than ABE if the weights represent distance

Dijkstra’s Algorithm



The first six steps used in computing the shortest path from *A* to *D*. The arrows indicate the working node.

Dijkstra’s Algorithm

1. **Initialization**: Set the distance to the source node to 0 and to all other nodes to infinity (∞). Make the source node permanent and the other nodes tentative.
2. **Select Node**: Select the tentative node with the smallest distance to the source node and make it permanent. This is the working node.
3. **Update Neighbors (Relaxation):** For the working node, calculate the distance from the source to the neighboring tentative nodes. If the new distance for a node is smaller than the previously recorded distance, update the node distance.
4. **Loop**: Repeat Steps 2–3 until all nodes are visited or the smallest distance among unvisited nodes is infinity (indicating unreachable nodes).

Flooding

- Every incoming packet is sent out on every outgoing line (except the one it arrived on)
- A simple, robust, and fast way to disseminate information throughout a network
  1. Guaranteed delivery: if a path exists, flooding will find it
  2. Fast network-wide dissemination: information reaches all nodes quickly
  3. Simplicity: no routing tables needed and easy to implement.
  4. Used to build routing information: Link-State Advertisements (LSAs) in link-state routing
  5. Used for path discovery: e.g. shortest path

Distance Vector Routing

- Each router maintains a distance vector
  - a table listing the best known distance (cost) to reach every known destination in the network and the next-hop router to use
- Common measures include hop count, delay, bandwidth, and reliability
- Each router knows only the distance to its neighbors
- Routers periodically exchange their distance vectors with neighbors
- The distance vectors received from the neighbors are used to recalculate the best paths and update the routing table
- This process continues until the network converges

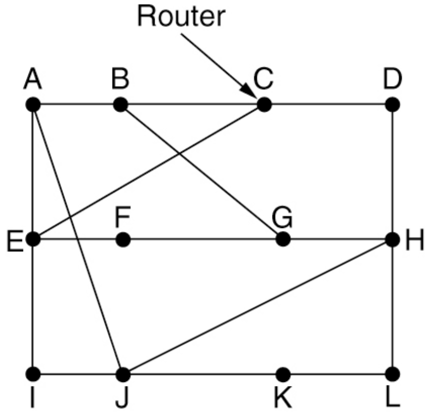
Distance Vector Routing

- Route exchange
  - B: I can reach X at cost(B,X)
  - C: I can reach X at cost(C,X)
  - A: I am cost(A,B) away from B
  - A: I am cost(A,C) away from C
- Shortest path calculation from A to X

  A: min{cost(A,X), cost(A,B) + cost(B,X), cost(A,C) + cost(C,X)}

  where cost (A,X) is the cost when there is a direct link between A and X

Distance Vector Routing



(a)

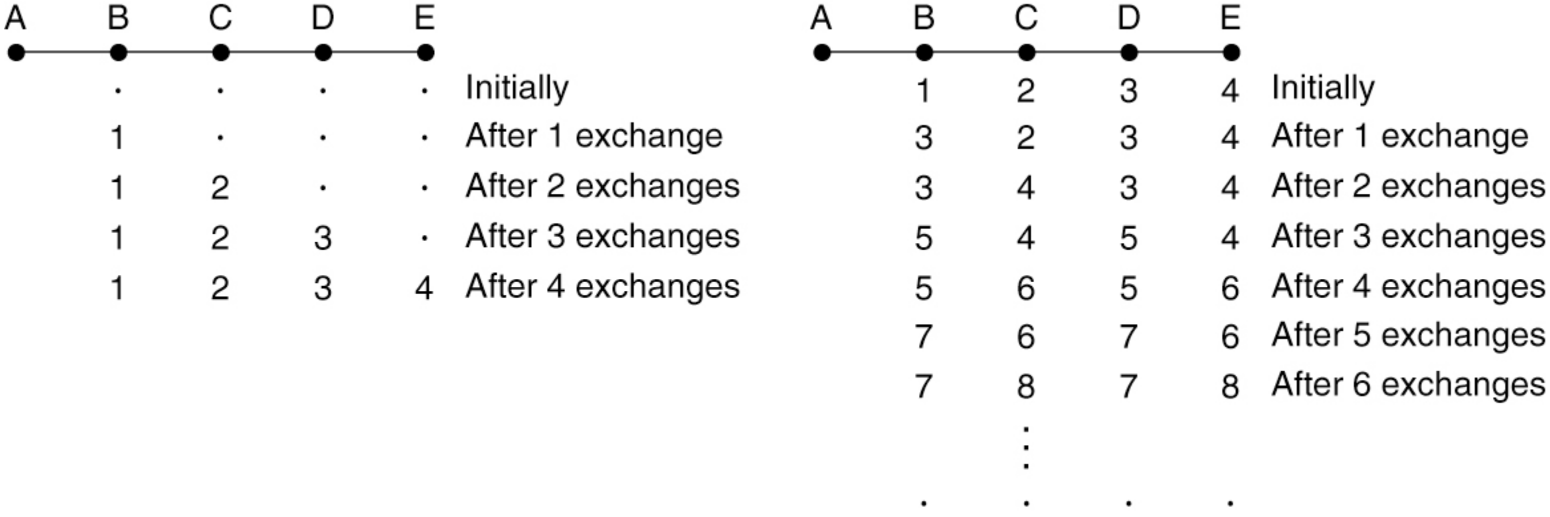
To	A	I	H	K	New estimated delay from J	Line
A	0	24	20	21	8	A
B	12	36	31	28	20	A
C	25	18	19	36	28	I
D	40	27	8	24	20	H
E	14	7	30	22	17	I
F	23	20	19	40	30	I
G	18	31	6	31	18	H
H	17	20	0	19	12	H
I	21	0	14	22	10	I
J	9	11	7	10	0	–
K	24	22	22	0	6	K
L	29	33	9	9	15	K
	JA delay is 8	JI delay is 10	JH delay is 12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

Vectors received from J's four neighbors

(b)

(a) A network. (b) Input from *A*, *I*, *H*, *K*, and the new routing table for *J*.

The Count-to-Infinity Problem



A	B	C	D	E	
	·	·	·	·	Initially
	1	·	·	·	After 1 exchange
	1	2	·	·	After 2 exchanges
	1	2	3	·	After 3 exchanges
	1	2	3	4	After 4 exchanges

(a)

A	B	C	D	E	
	1	2	3	4	Initially
	3	2	3	4	After 1 exchange
	3	4	3	4	After 2 exchanges
	5	4	5	4	After 3 exchanges
	5	6	5	6	After 4 exchanges
	7	6	7	6	After 5 exchanges
	7	8	7	8	After 6 exchanges
		⋮			
	·	·	·	·	

(b)

Route Poisoning

- Choose a suitable value for infinity
- Poison reverse
    - A: I can reach X through B for cost T
    - A tells B
    - I can reach X for infinity cost, since I reach X through you!
- Can fail for loops of length ≥ 3

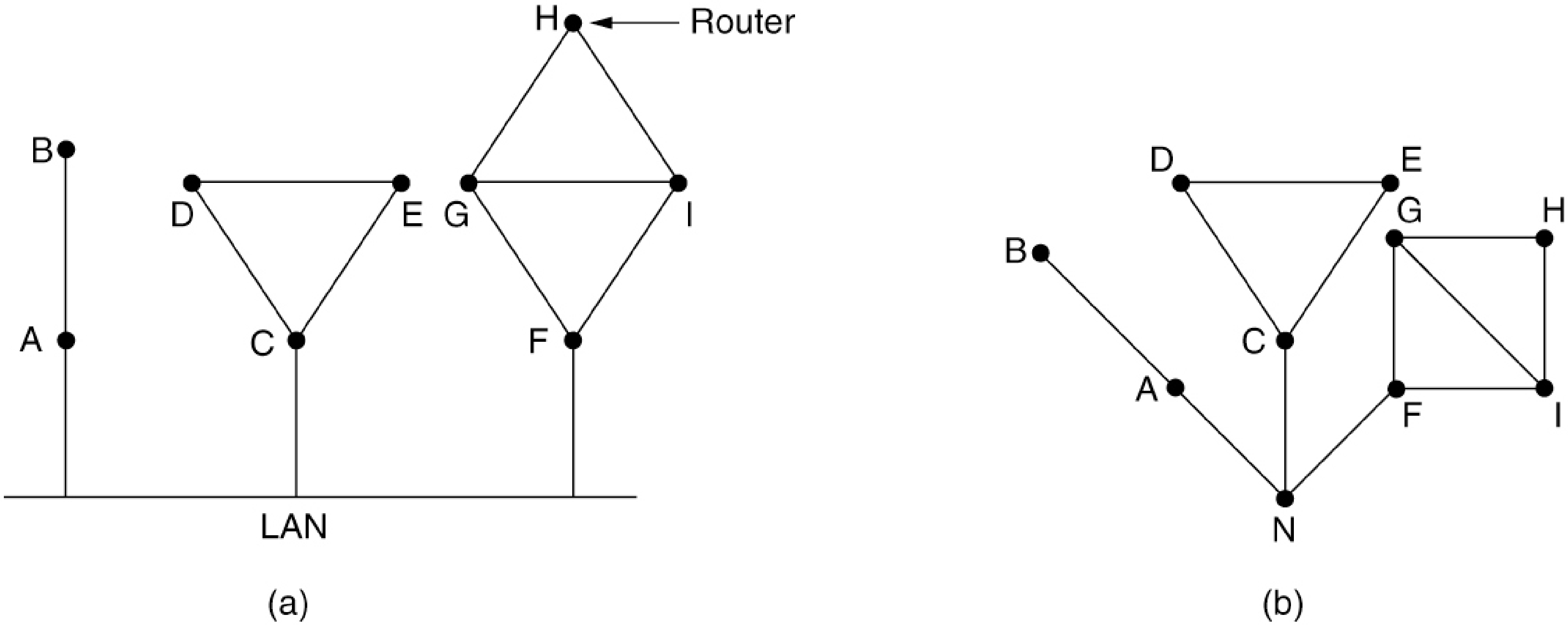
Distance Vector Routing

- This algorithm is simple to implement, and has low computational complexity and memory requirements
- Suitable for small or stable networks
- It can have slow convergence, especially after failures
- Count-to-infinity problem
  - Mitigation: Route poisoning
- Used in the Internet Routing Information Protocol (RIP)

Link State Routing

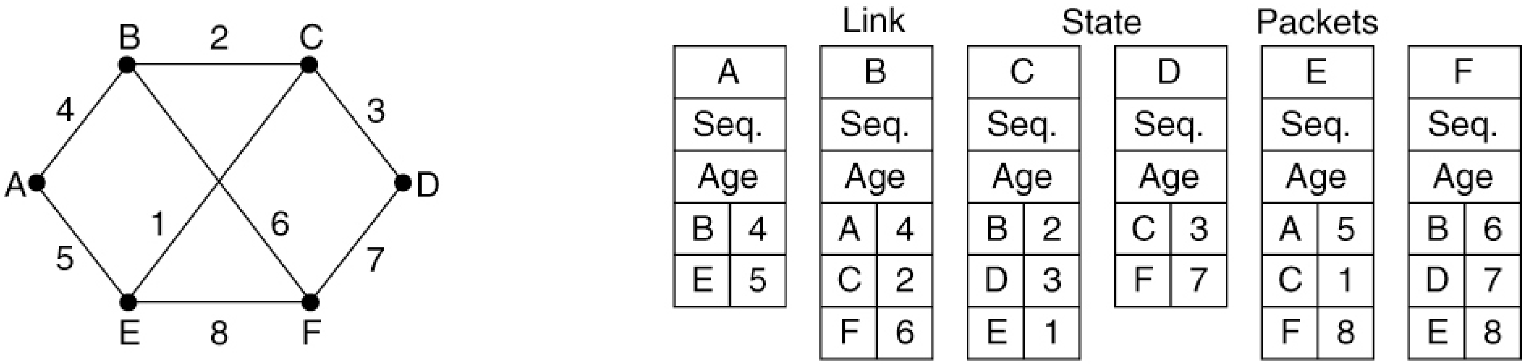
- Discover neighbors, learn network addresses
- Set distance (cost) to each neighbor
- Construct a packet telling what it has learned
- Send this packet to and receive packets from all other routers
- Compute the shortest path to every other router

Learning about Neighbors



(a) Nine routers and a broadcast LAN. (b) A graph model of (a).

Building Link State Packets



(a)

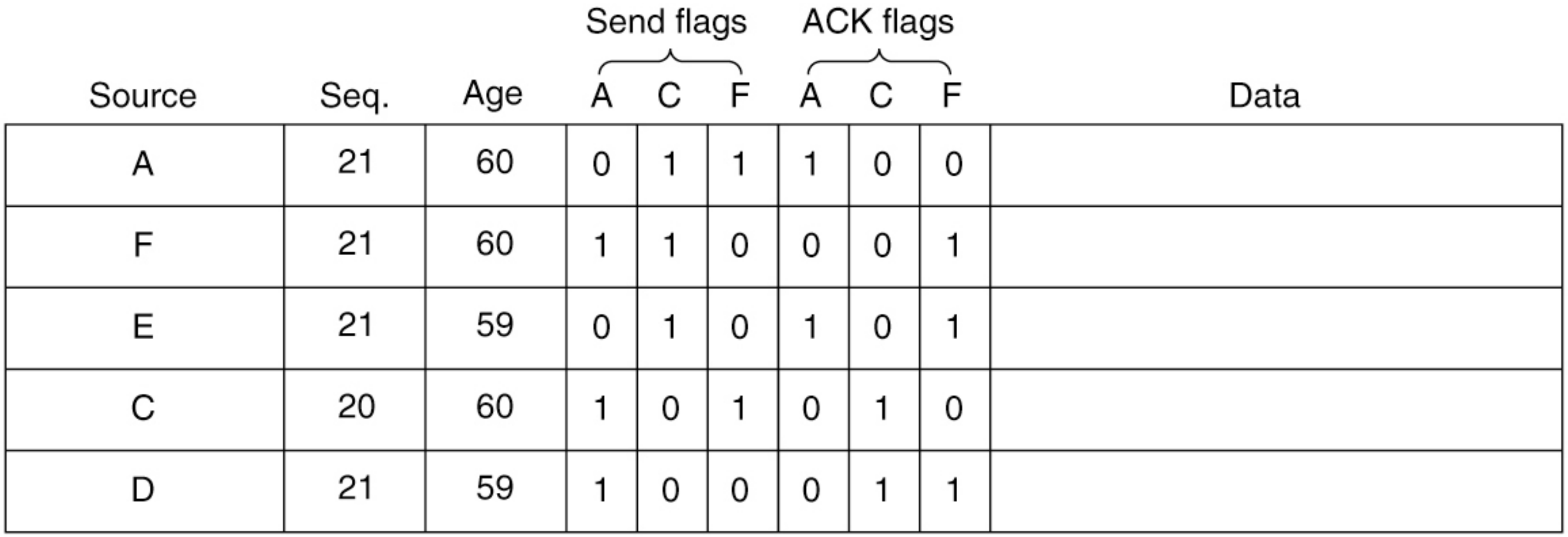
Link State Packets

A		B		C		D		E		F	
Seq.		Seq.		Seq.		Seq.		Seq.		Seq.	
Age		Age		Age		Age		Age		Age	
B	4	A	4	B	2	C	3	A	5	B	6
E	5	C	2	D	3	F	7	C	1	D	7
		F	6	E	1			F	8	E	8

(b)

(a) A network. (b) The link state packets for this network.

Distributing the Link State Packets



			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	

The packet buffer for router *B*

Comparison of LS and DV Algorithms

Message complexity

LS: *n* routers, O(n^2) messages sent

DV: Exchange between neighbors; convergence time and thus the number of exchanges varies

Speed of convergence

LS: O(n^2) algorithm, O(n^2) messages

- May have oscillations

DV: convergence time varies

- May have routing loops
- Count-to-infinity problem

Robustness: what happens if a router malfunctions or is compromised?

LS:

- Router can advertise incorrect link cost
- Each router computes only its own table

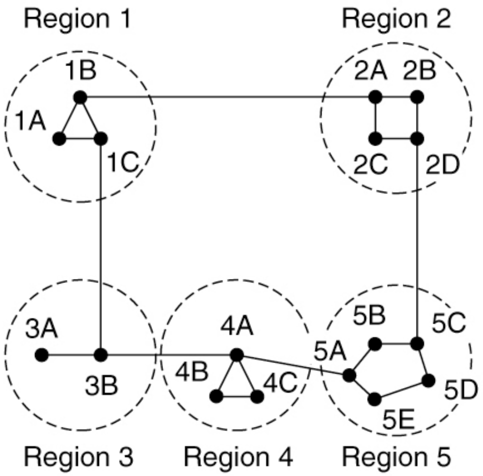
DV:

- DV router can advertise incorrect path cost (I have a really low-cost path to everywhere): black holes
- Each router’s DV is used by others: errors propagate through the network

Black Holes

- Distance Vector (DV) protocols (such as RIP) rely on routers periodically exchanging routing tables with neighbors
- Because updates propagate slowly, routers may temporarily have inconsistent routing information
- When a link fails, one router detects the failure, but other routers still believe the route exists
- Traffic continues to be forwarded on the failed path
- Packets reach a router that has no valid route, and that router drops the packets
- This creates a black hole

Hierarchical Routing



(a)

Full table for 1A

Dest.	Line	Hops
1A	–	–
1B	1B	1
1C	1C	1
2A	1B	2
2B	1B	3
2C	1B	3
2D	1B	4
3A	1C	3
3B	1C	2
4A	1C	3
4B	1C	4
4C	1C	4
5A	1C	4
5B	1C	5
5C	1B	5
5D	1C	6
5E	1C	5

(b)

Hierarchical table for 1A

Dest.	Line	Hops
1A	–	–
1B	1B	1
1C	1C	1
2	1B	2
3	1C	2
4	1C	3
5	1C	4

(c)

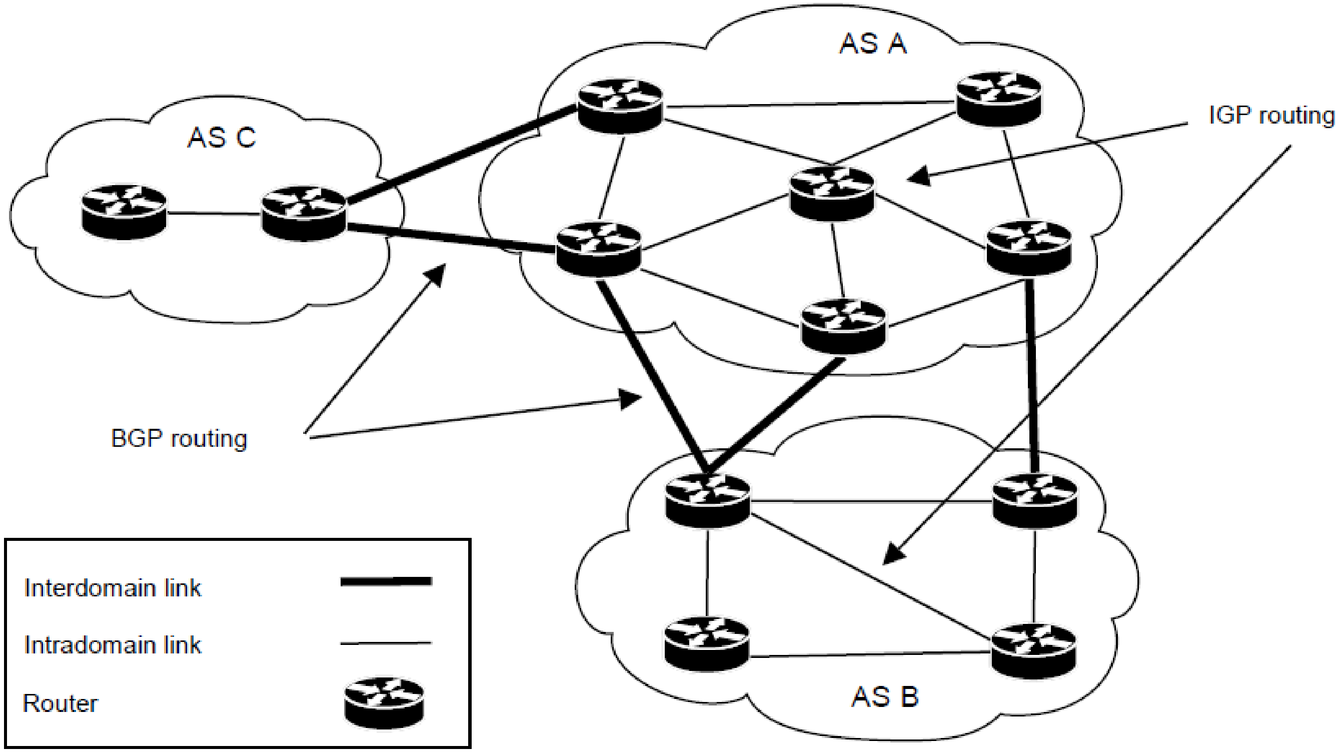
Making Routing Scalable

- Routing discussion thus far has been idealized
  - All routers are identical
  - Network is flat
- Not true in practice
- Scale: Billions of destinations
  - Can’t store all destinations in routing tables
  - Routing table exchange would overload links
- Administrative autonomy
  - Internet: A network of networks
  - Each network administrator may want to control routing in their own network

Internet Approach to Scalable Routing

- Aggregate routers into regions known as Autonomous Systems (ASs) or domains
- Intra-AS (intradomain)
  - Routing among routers within the same AS
  - All routers run the same intradomain protocol
  - Routers in different ASes can run different intradomain routing protocols
  - Gateway router: at AS edge, has link(s) to router(s) in other Ases
- Inter-AS (interdomain) routing among ASes
  - Gateways perform interdomain routing (as well as participate in intradomain routing)

Interdomain and Intradomain Routing



Intradomain Routing

- Intradomain routing
    - IGP (Interior Gateway Protocol)
- RIP (Routing Information Protocol)
    - Works well in small systems
- OSPF (Open Shortest Path First)
    - Widely used in company networks
- IS-IS (Intermediate-System to Intermediate-System)
    - Widely used in ISP networks

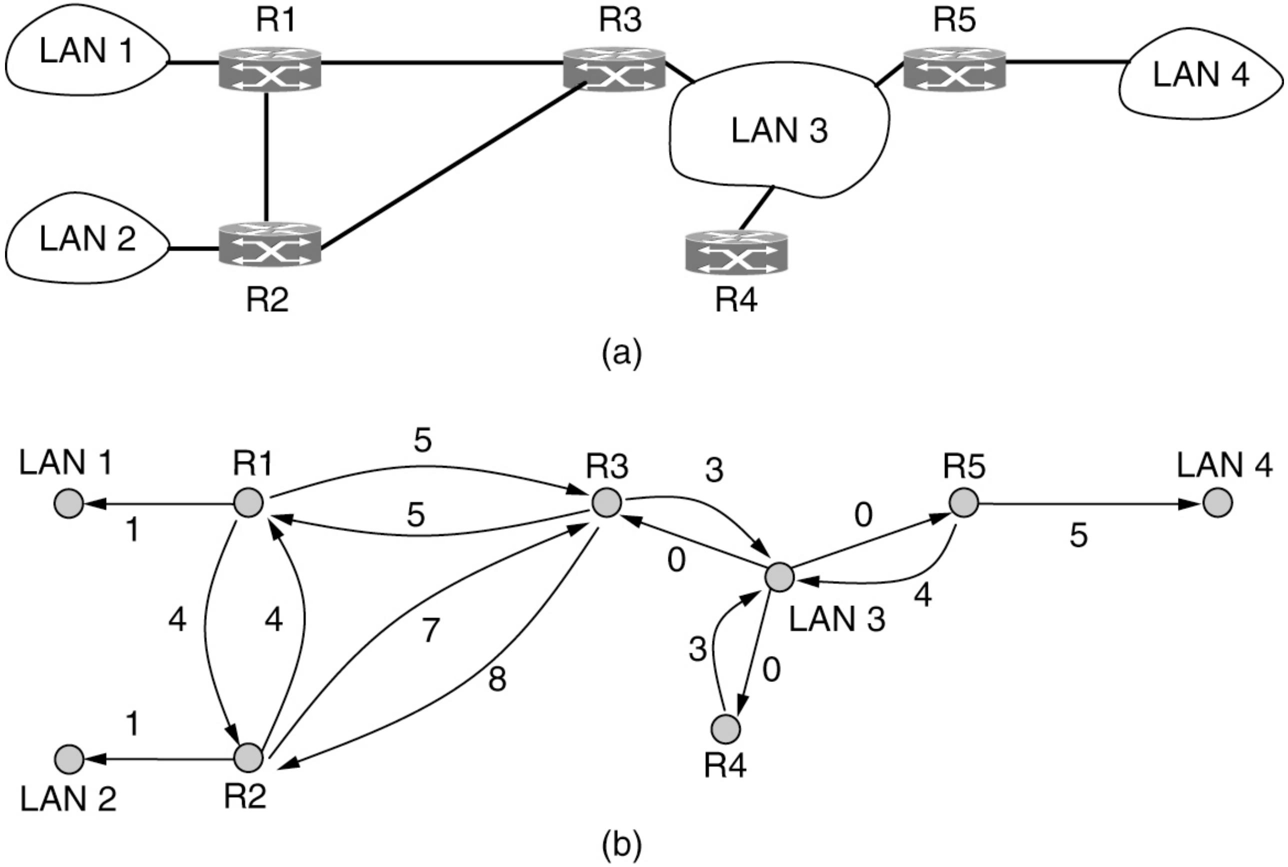
Routing Information Protocol (RIP)

- Uses distance vector routing
- Included in BSD Unix in 1982
  - Maximum number of hops: 15
- Distance vectors
  - Exchanged between neighbors every 30 s
  - Up to 25 destinations within an RIP packet
- if no advertisement for 180 s: neighbor is dead
  - Invalidate routes going through the neighbor
  - poisoned reverse to speed up bad news
  - infinity: 16 hops

OSPF—Interior Gateway Routing Protocol

- Open: publicly available
- Link-state based
  - Each router floods OSPF link-state advertisements directly over IP to all routers in the AS
  - Supports a variety of distance metrics, e.g. bandwidth, delay
  - Each router has the full topology and uses Dijkstra's algorithm to compute the forwarding table
- Performs load balancing, splitting the load over multiple lines (only one path allowed in RIP)
- Supports hierarchical systems
- Security: all OSPF messages are authenticated to prevent malicious intrusion

OSPF—Interior Gateway Routing Protocol

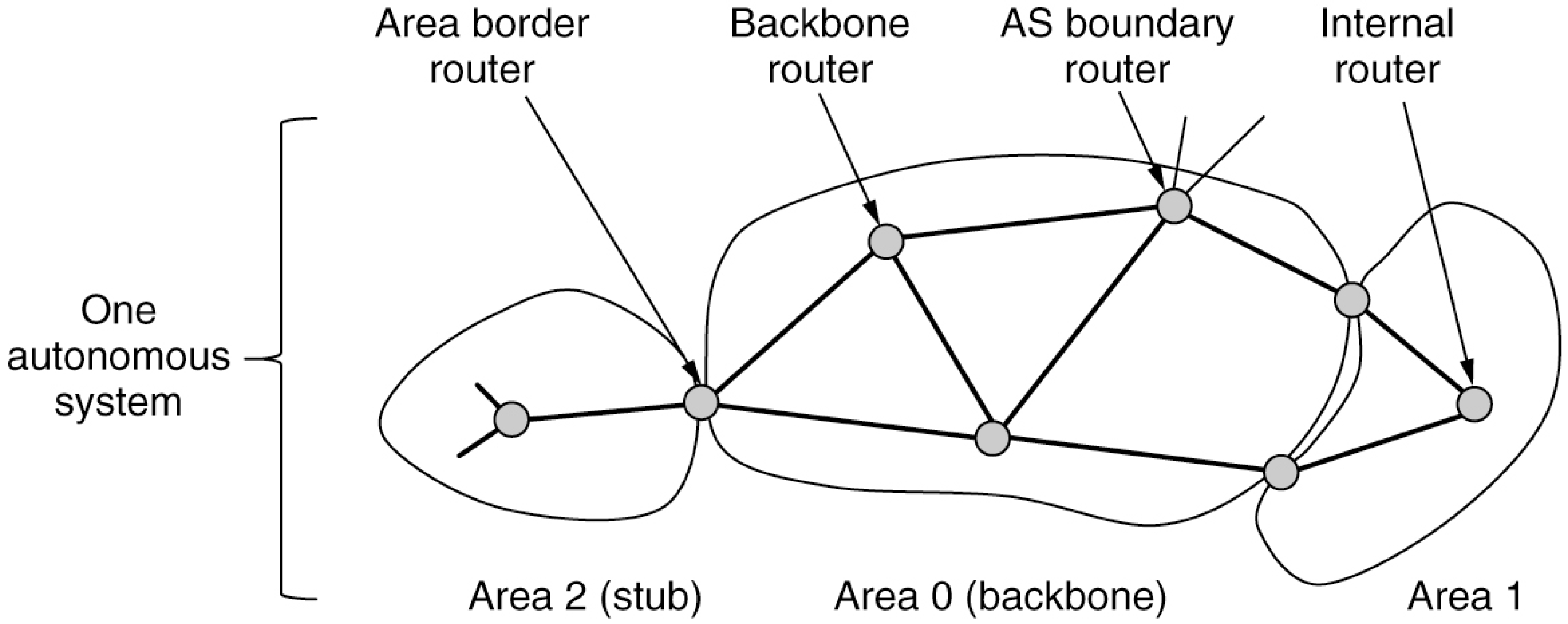


(a)

(b)

(a) An autonomous system. (b) A graph representation of (a).

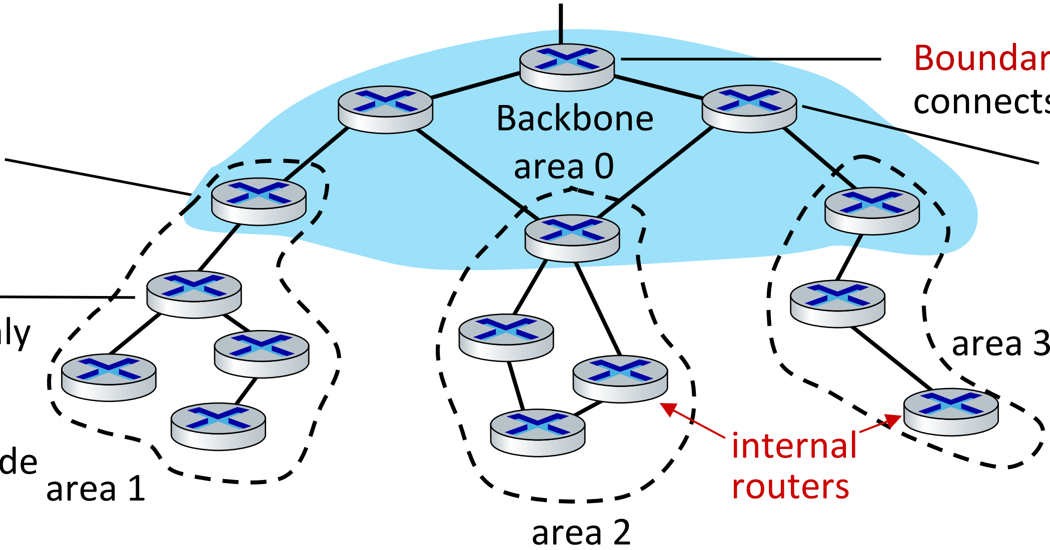
OSPF—Interior Gateway Routing Protocol



The relation between ASes, backbones, and areas in OSPF

Hierarchical OSPF

- Two-level hierarchy: local area and backbone
- Link-state advertisements flooded only in area or backbone
- Each node has detailed area topology; only knows direction to reach other destinations



Area border routers: summarize distances to destinations in own area, advertise in backbone

Internal routers:

- flood LSAs in the area only
- compute routing within the area
- forward packets to outside via area border router

Boundary router: connects to other ASes

Backbone router: runs OSPF limited to backbone

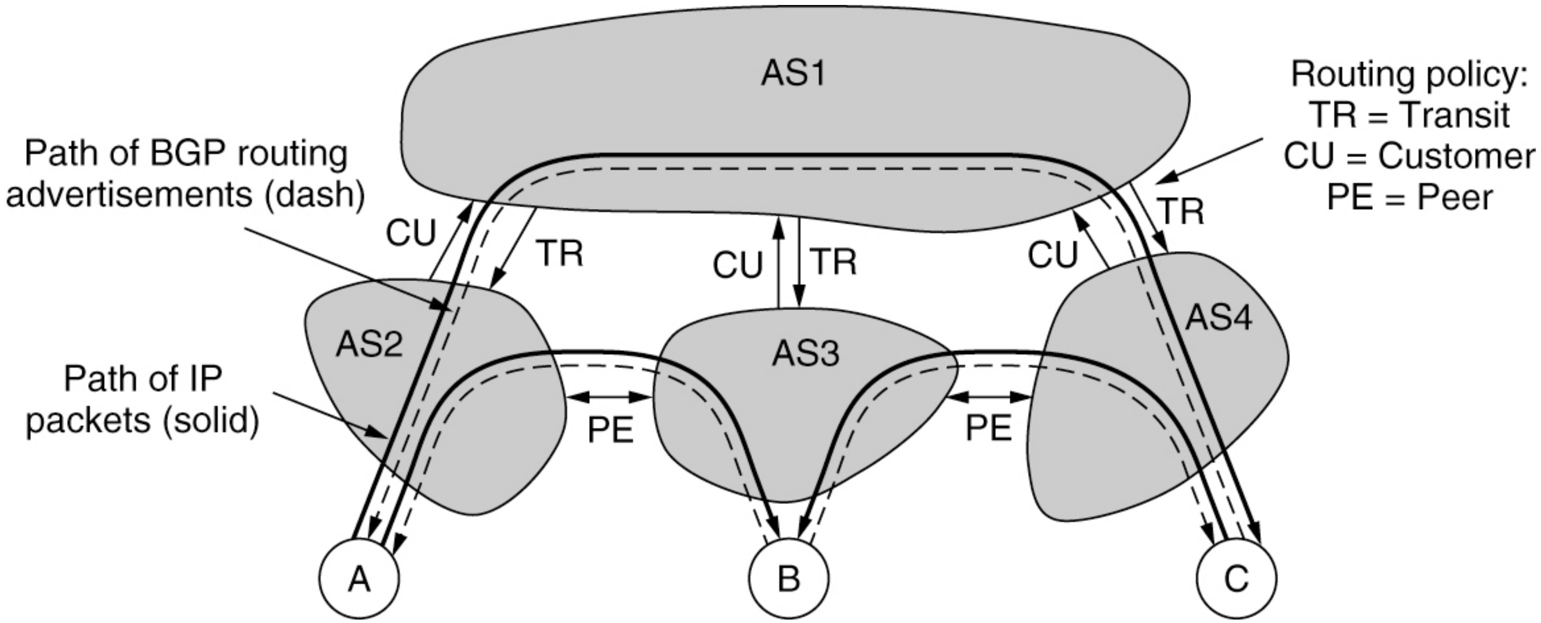
BGP—Exterior Gateway Routing Protocol

- The de facto interdomain routing protocol
  - Glue that holds the Internet together
- Allows subnets to advertise their existence, and the destinations it can reach, to the rest of Internet
- I am here, here is who I can reach, and how
- BGP provides each AS a means to
  - Obtain destination network reachability information from neighboring ASes
  - Determine routes to other networks based on reachability information and policy
  - Propagate reachability information to all AS-internal routers
  - Advertise (to neighboring networks) destination reachability information

BGP—Exterior Gateway Routing Protocol

- Possible routing constraints
  - Do not carry commercial traffic on the educational network
  - Never send traffic from the Pentagon on a route through Iraq
  - Use TeliaSonera instead of Verizon because it is cheaper
  - Don’t use AT&T in Australia because performance is poor
  - Traffic starting or ending at Apple should not transit Google

BGP—Exterior Gateway Routing Protocol

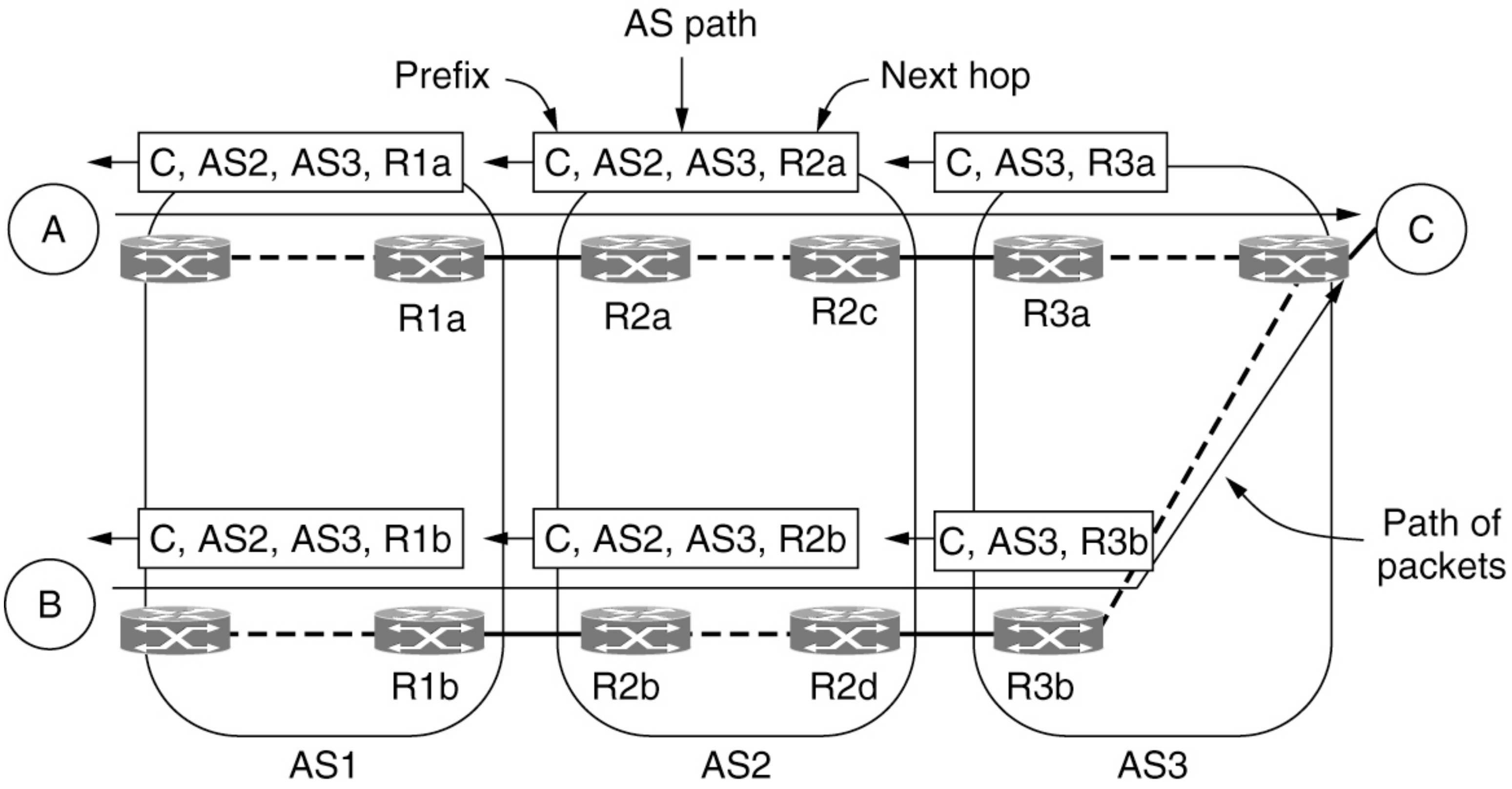


Routing policies between four autonomous systems

BGP—Exterior Gateway Routing Protocol

- BGP advertised route: prefix + attributes
  - Prefix: destination being advertised
  - Two important attributes
    - AS-PATH: list of ASes through which the prefix advertisement has passed
    - NEXT-HOP: indicates a specific internal-AS router to the next-hop AS
- Policy-based routing
  - gateway receiving route advertisement uses import policy to accept/decline path (e.g., never route through AS Y).
  - AS policy also determines whether to advertise the path to other neighboring ASes

BGP—Exterior Gateway Routing Protocol



Propagation of BGP route advertisements

BGP Route Selection

- A router may learn about more than one route to a destination AS
- The route can be selected based on
  - Peered networks
  - Local preference: policy decision
  - Lowest internal cost
  - Shortest AS PATH
  - Closest NEXT HOP router: hot potato routing