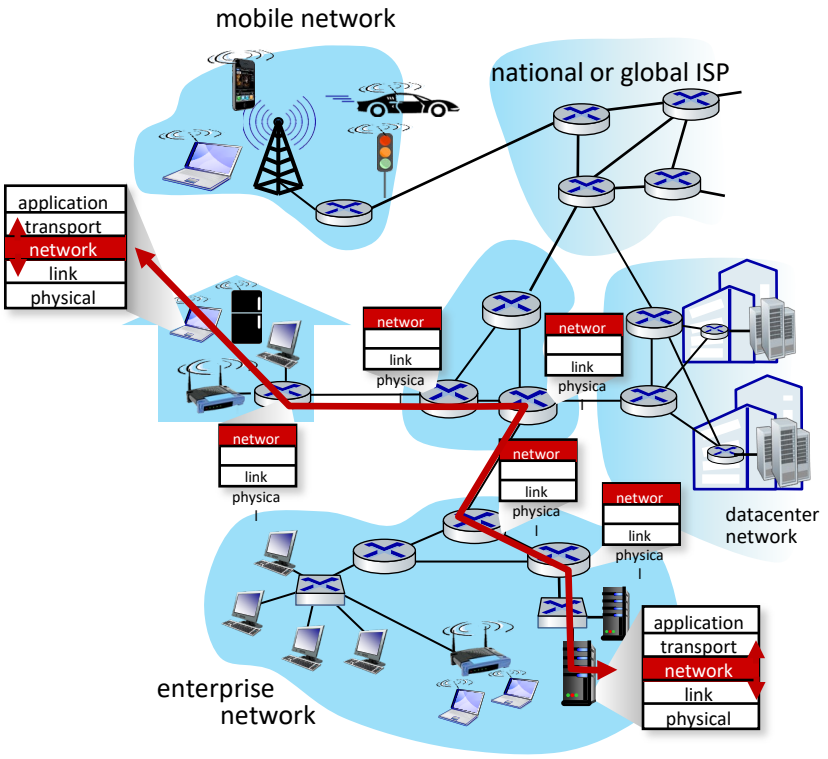
ECE 363
Communication Networks

Network Layer

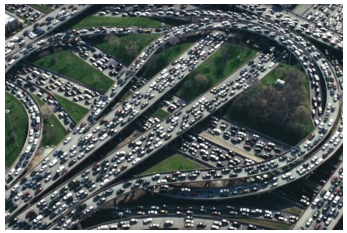
Network Layer Services and Protocols

- Transport segment 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

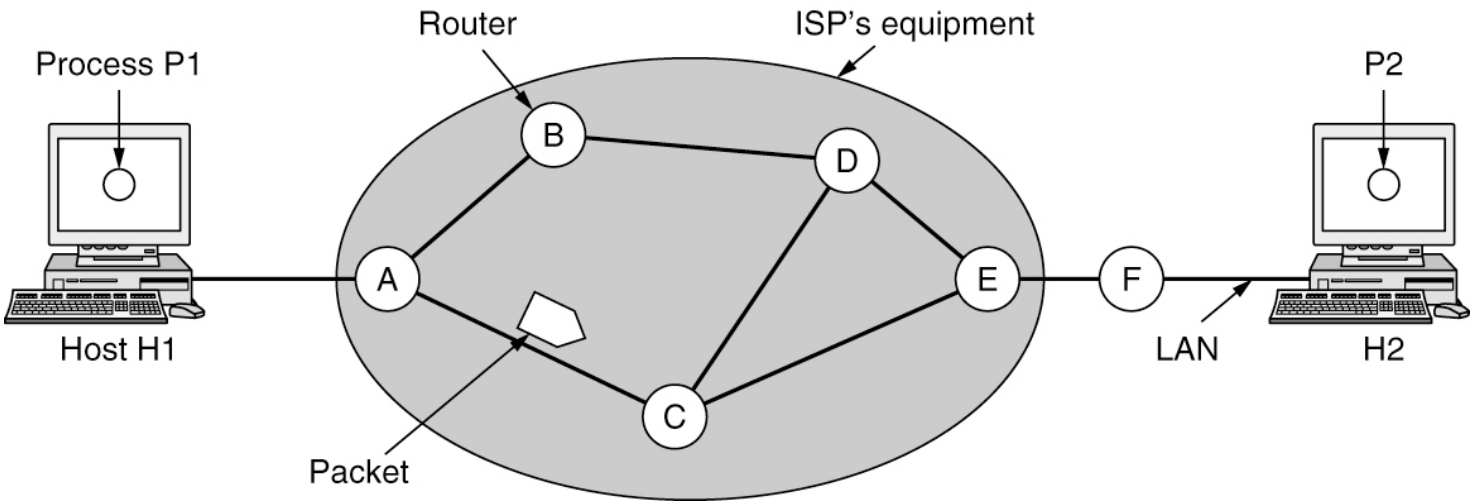
- Forwarding: Move packets from the router input to the appropriate router output
- Routing: Determine the route taken by packets from source to destination
- Analogy: Taking a trip
  - Forwarding: Process of getting through a single interchange
  - Routing: Process of planning the trip from source to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



forwarding

routing

Store-and-Forward Packet Switching



Packets are stored in routers before they are forwarded

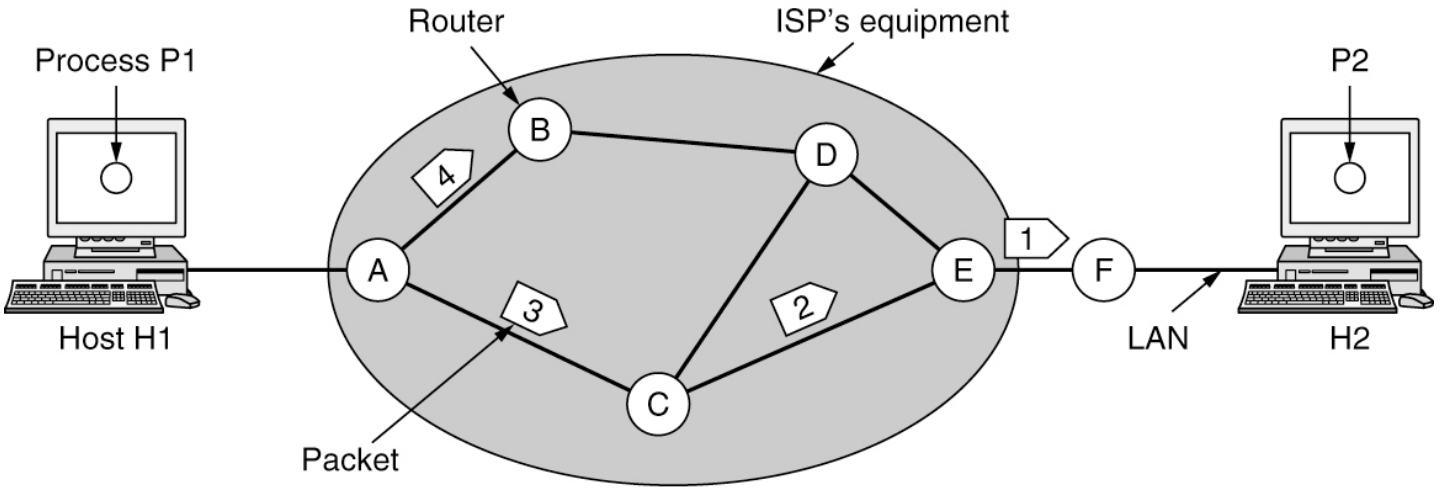
Services Provided to the Transport Layer

- 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

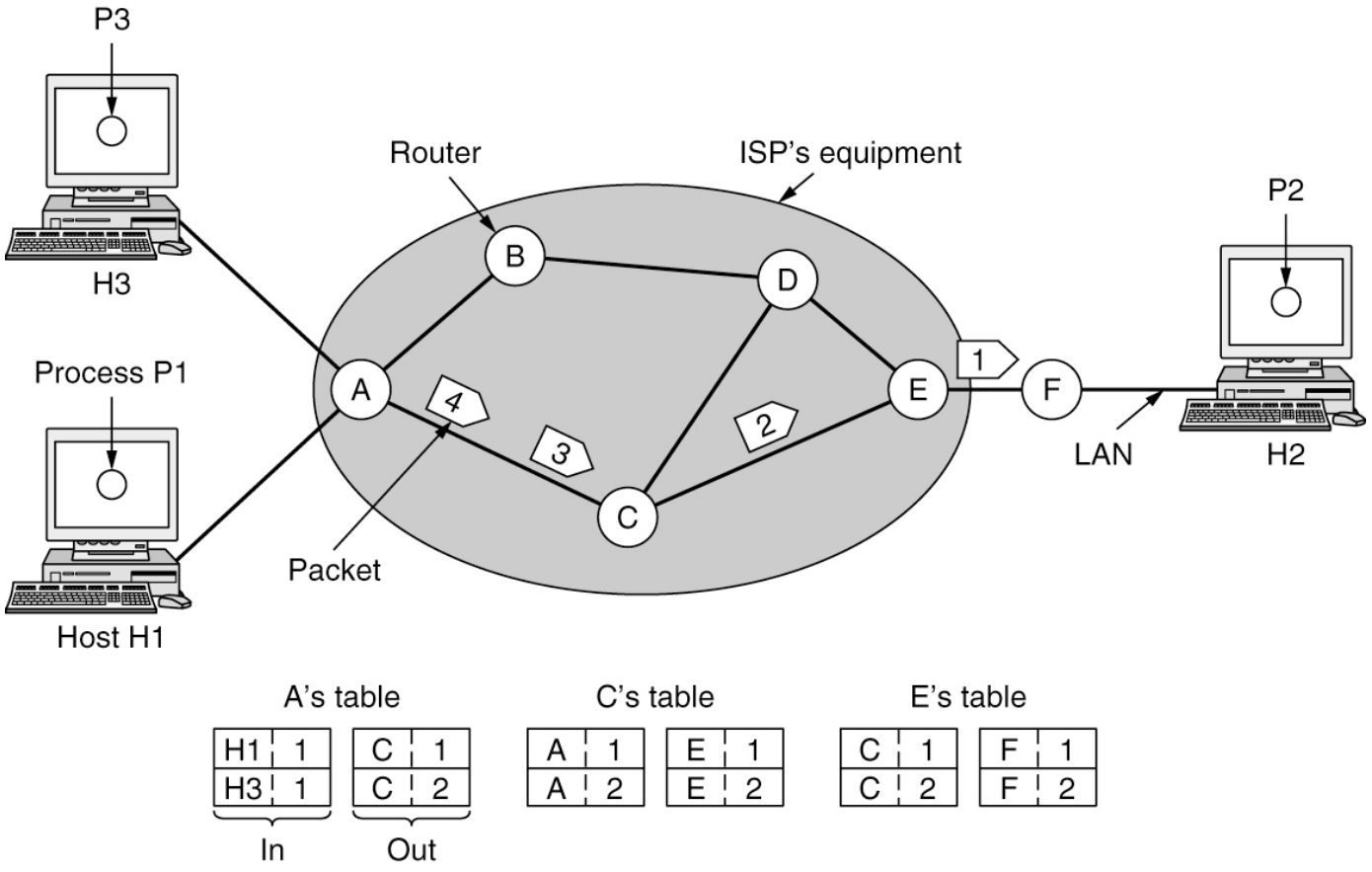
C's table

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

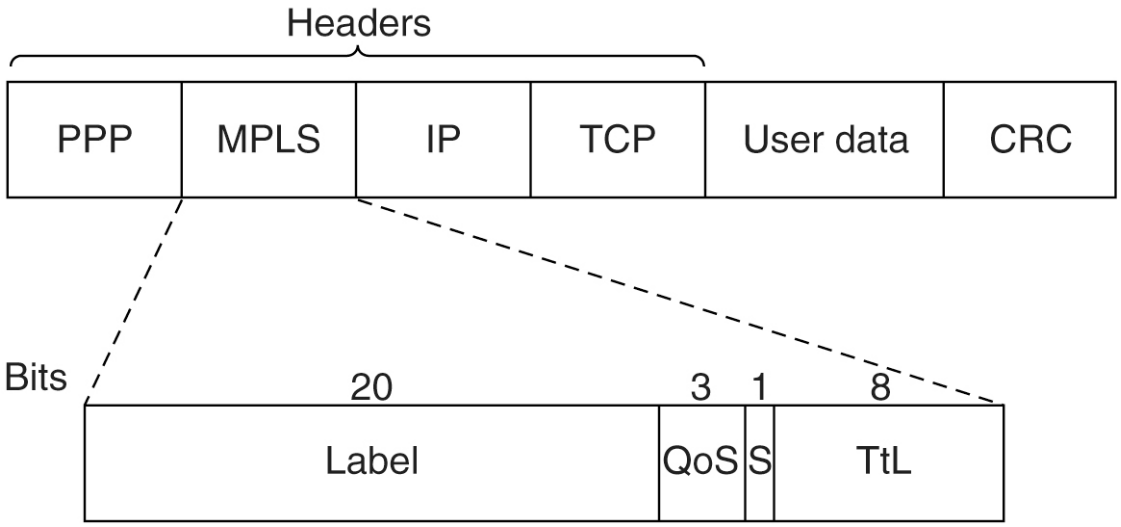
Connection-Oriented Service



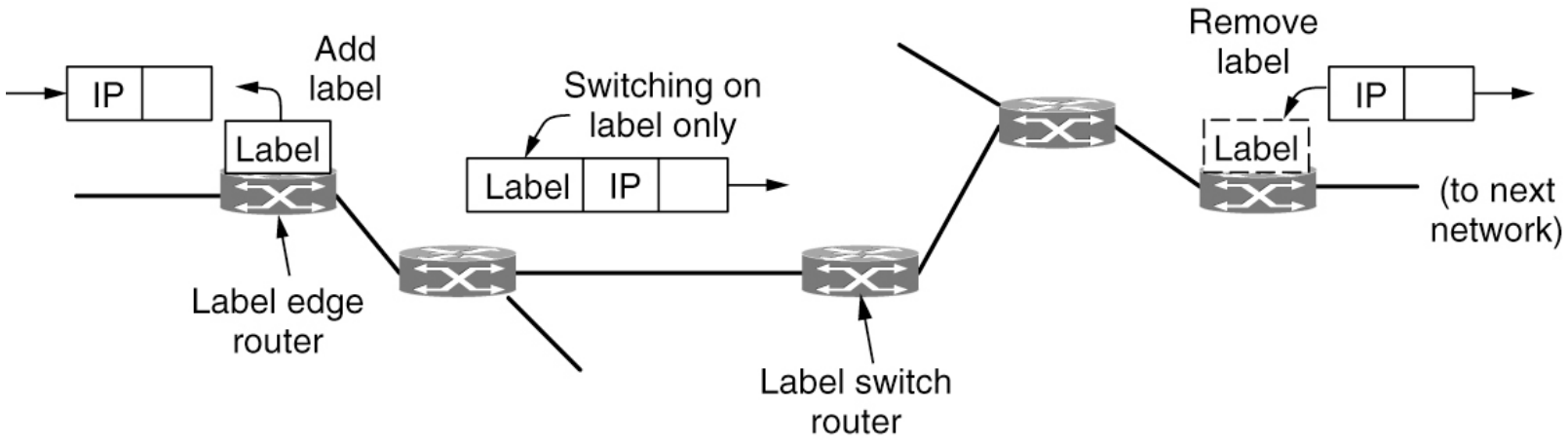
Routing within a virtual-circuit network

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

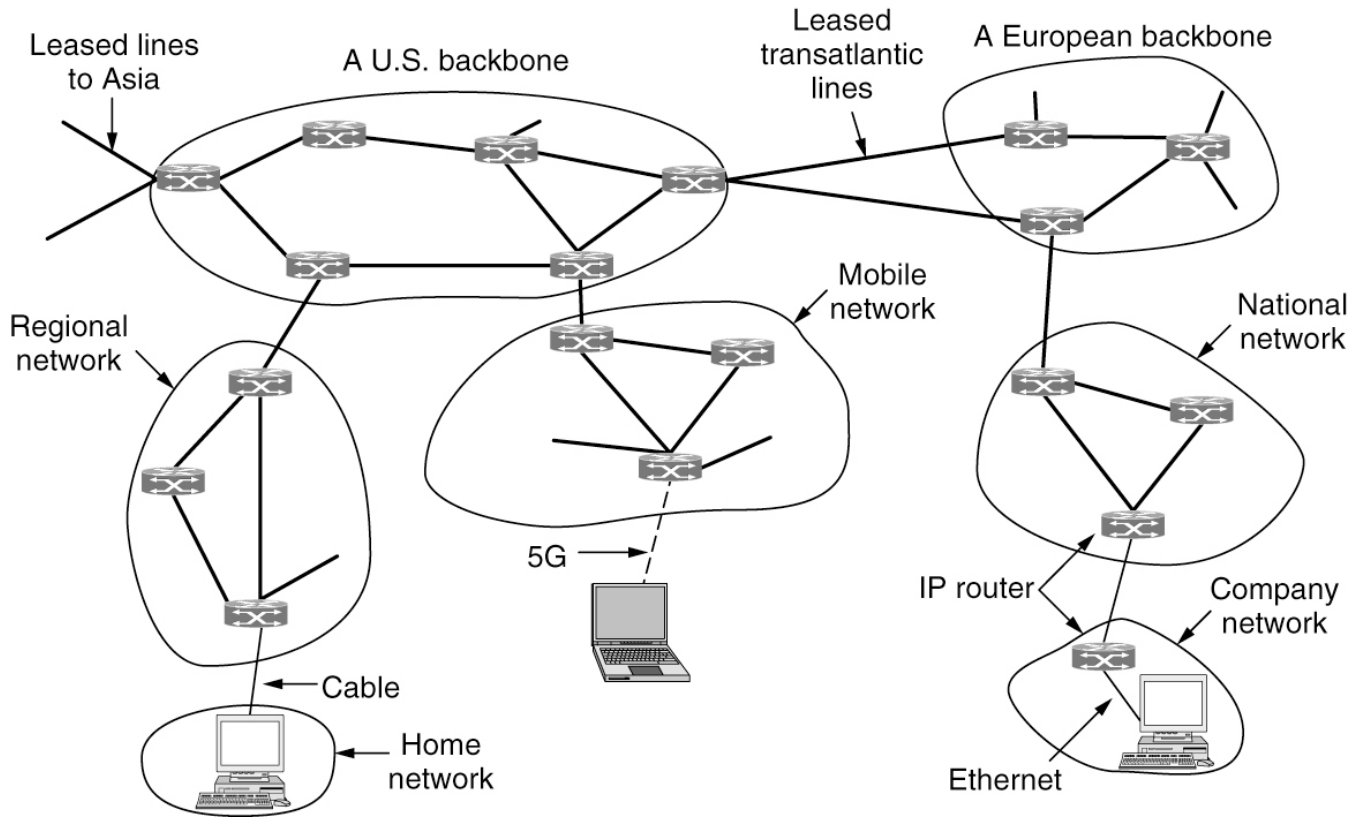
Datagram and Virtual-Circuit Networks

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

The Network Layer in the Internet

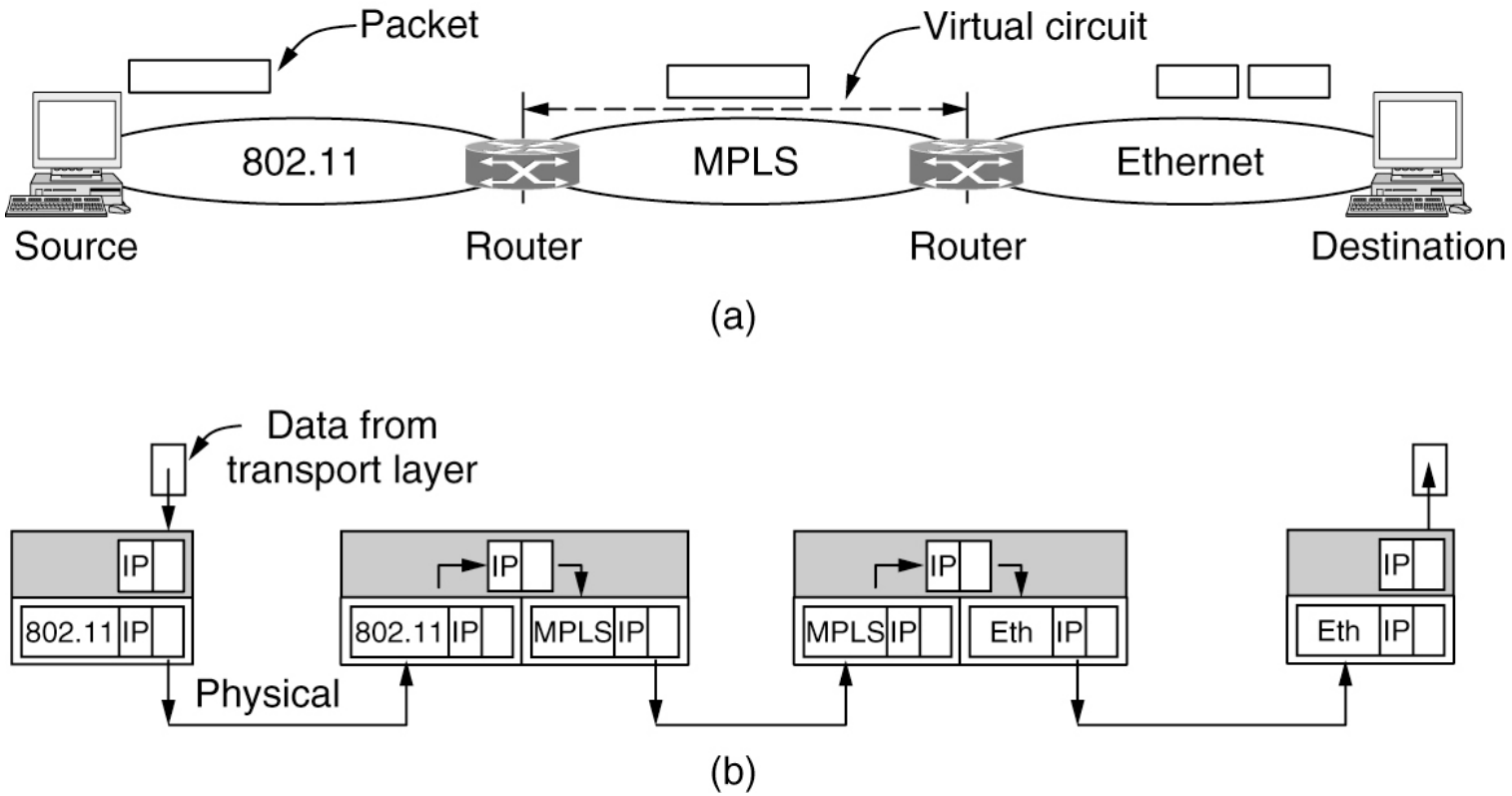


The Internet is an interconnected collection of networks

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

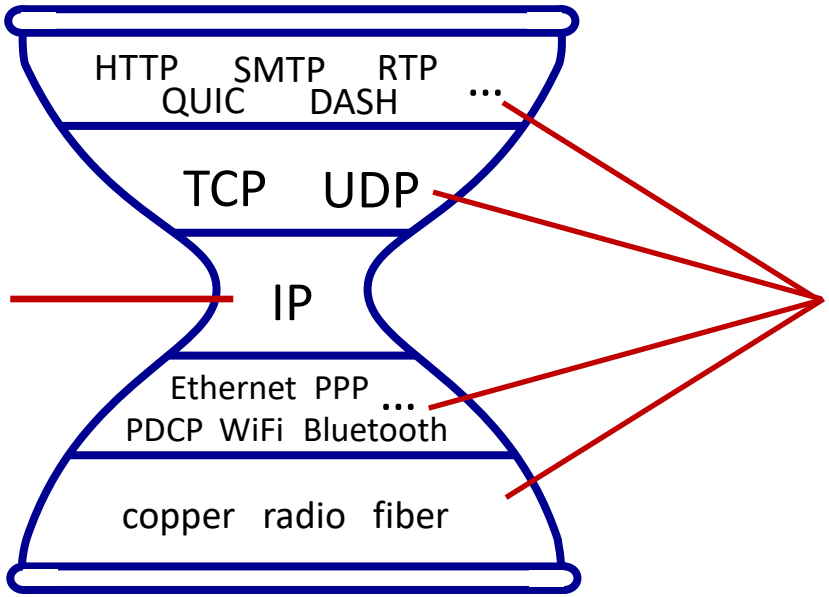


(a) A packet crossing different networks.

(b) Network and link layer protocol processing.

Internet Protocol (IP)

Implemented by every (billions) Internet-connected device

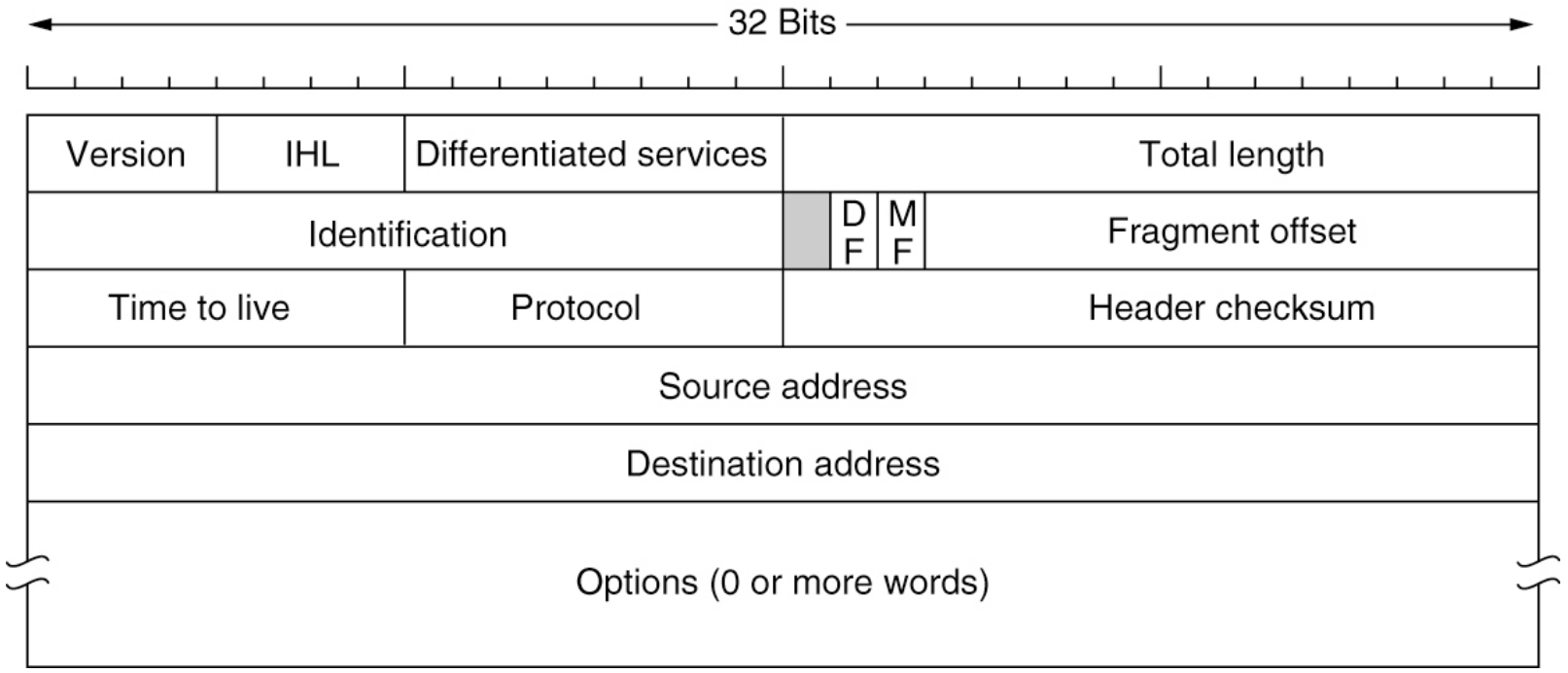


Many protocols in the physical, link, transport, and application layers

Internet Network Layer

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

IP Version 4 Protocol



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

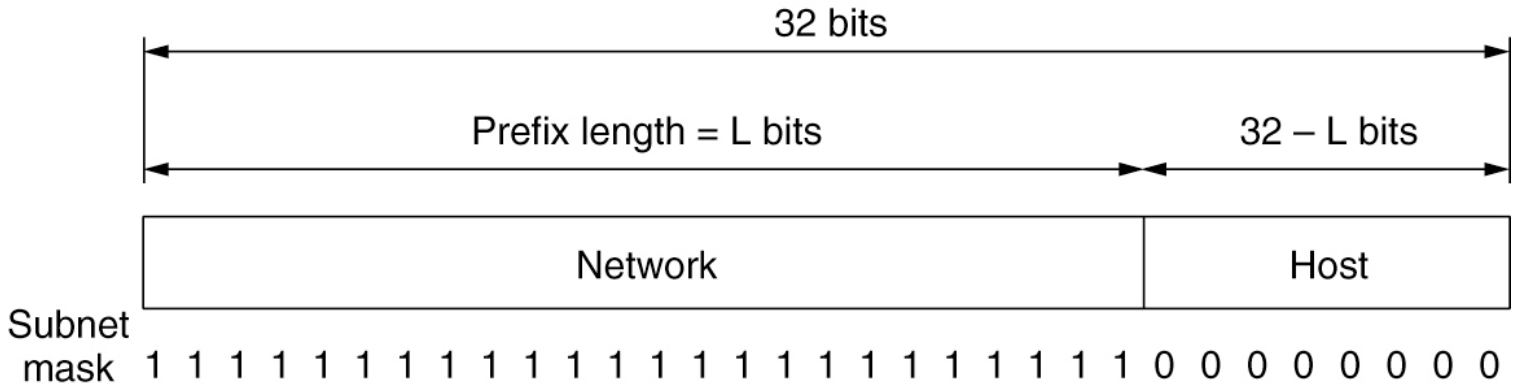
  2FFFD → FFFD + 2 = FFFF

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

IP Addresses

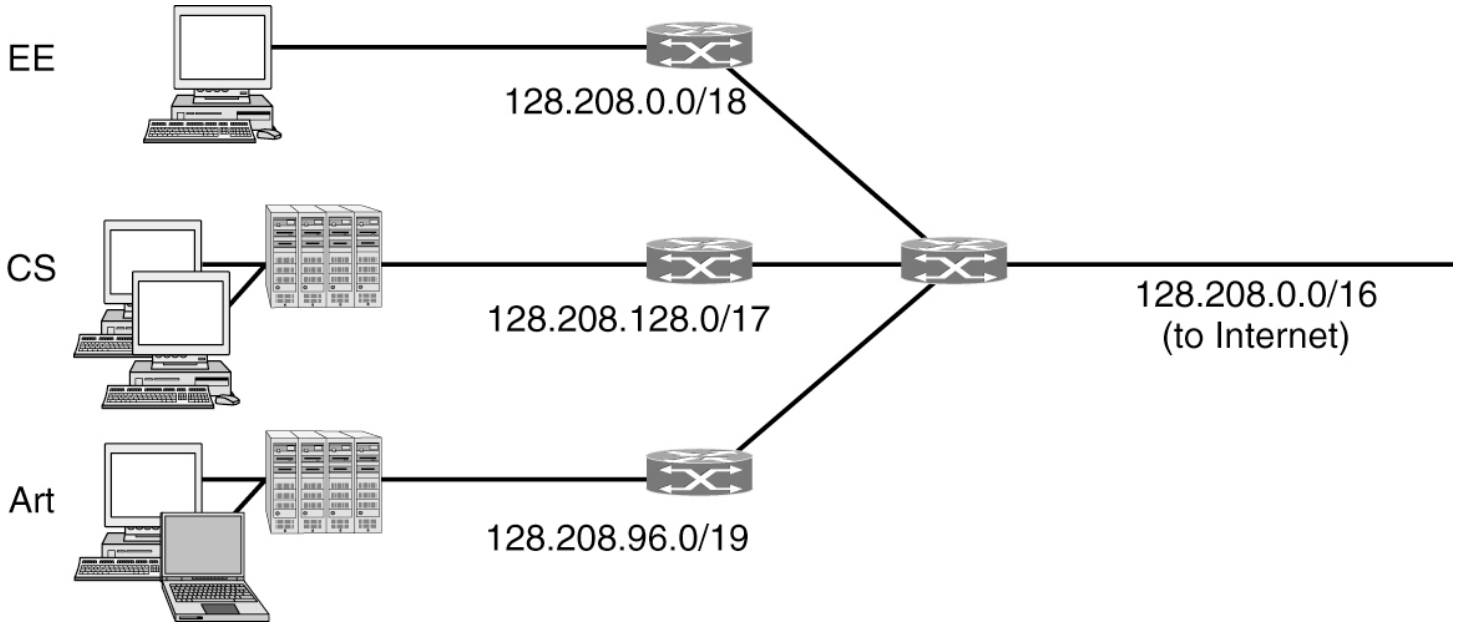
- Prefixes
  - A contiguous block of IP address space
- Subnets
- CIDR—Classless InterDomain Routing
- Classful and special addressing
- NAT—Network Address Translation

Prefixes



A prefix and a subnet mask

Subnets



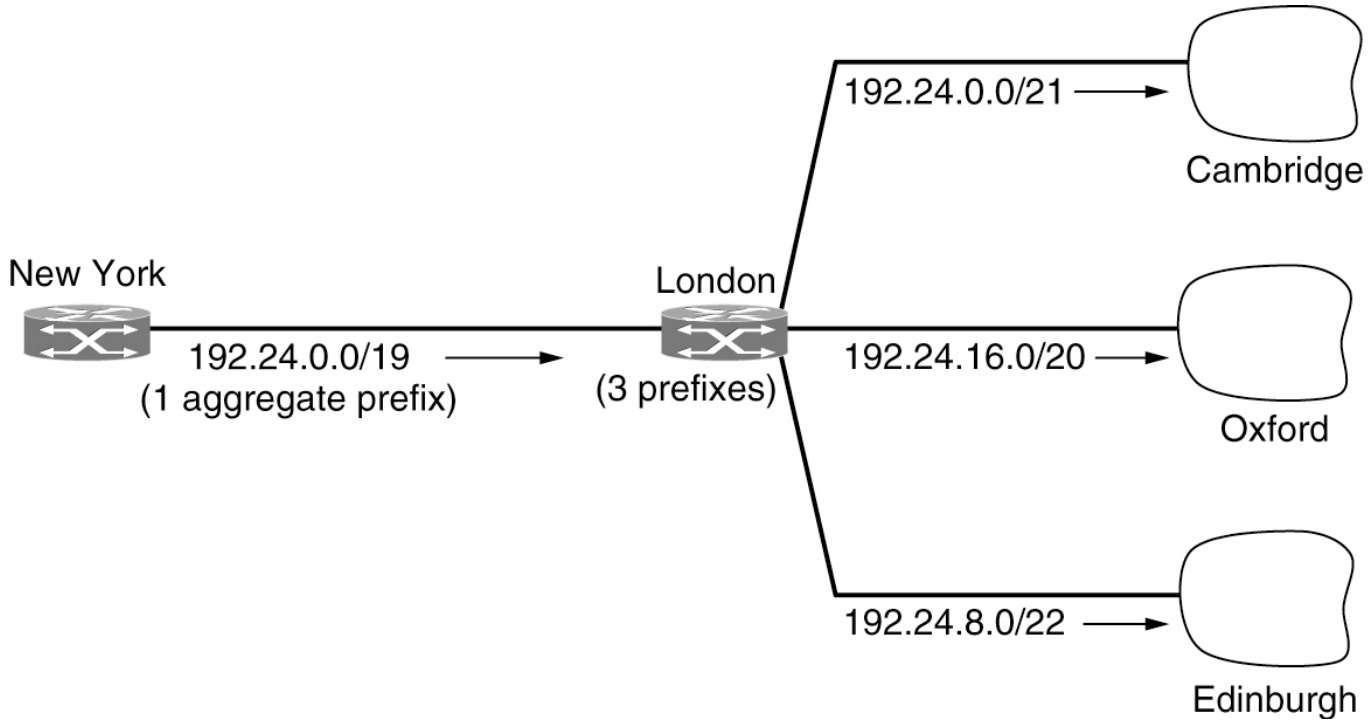
Splitting an IP prefix into separate networks with subnetting

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

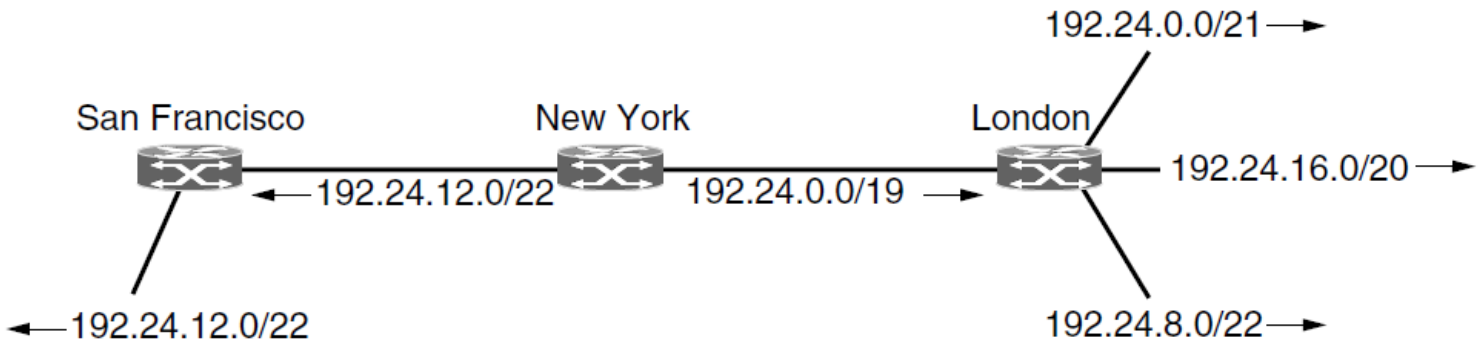
A set of IP address assignments

CIDR—Classless InterDomain Routing



Aggregation of IP prefixes

CIDR—Classless InterDomain Routing



Longest matching prefix routing at the New York router

Longest Matching Prefix

longest prefix match

when looking for forwarding table entry for given destination address, use longest address prefix that matches 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

longest prefix match

when looking for forwarding table entry for given destination address, use longest address prefix that matches 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

longest prefix match

when looking for forwarding table entry for given destination address, use longest address prefix that matches destination address.

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

match!

examples:

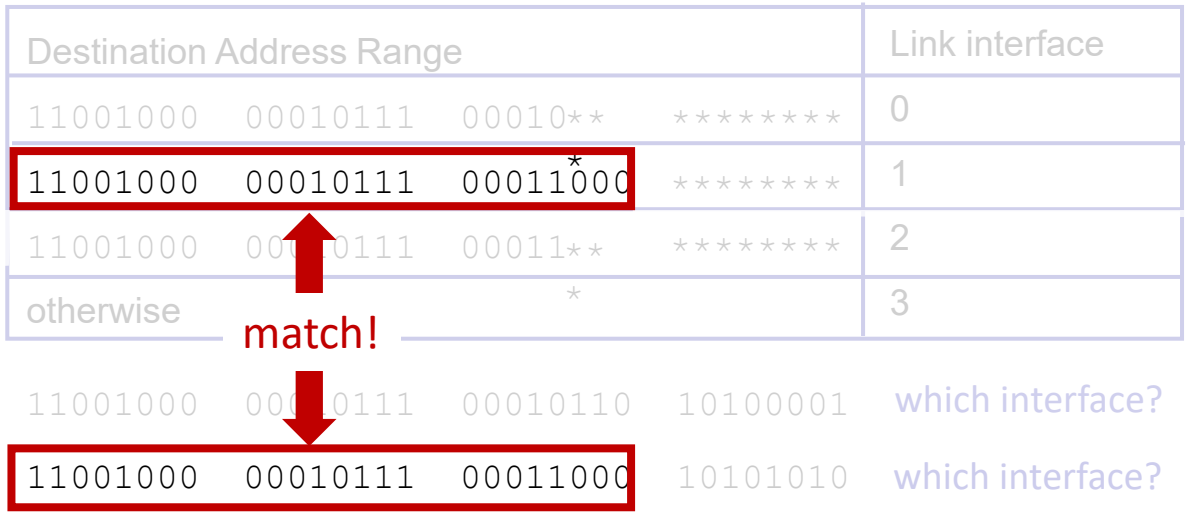
11001000 00010110 10100001 which interface?

11001000 00010111 00011000 10101010 which interface?

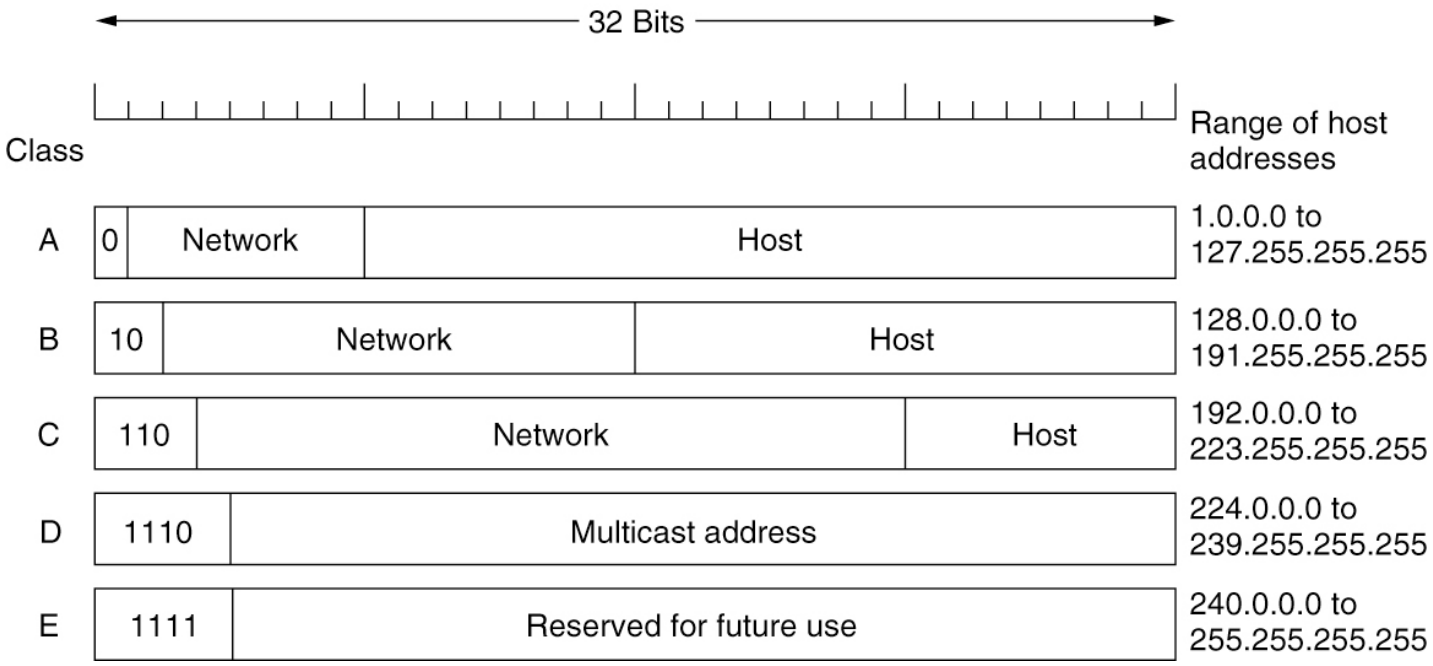
Longest Matching Prefix

longest prefix match

when looking for forwarding table entry for given destination address, use longest address prefix that matches destination address.

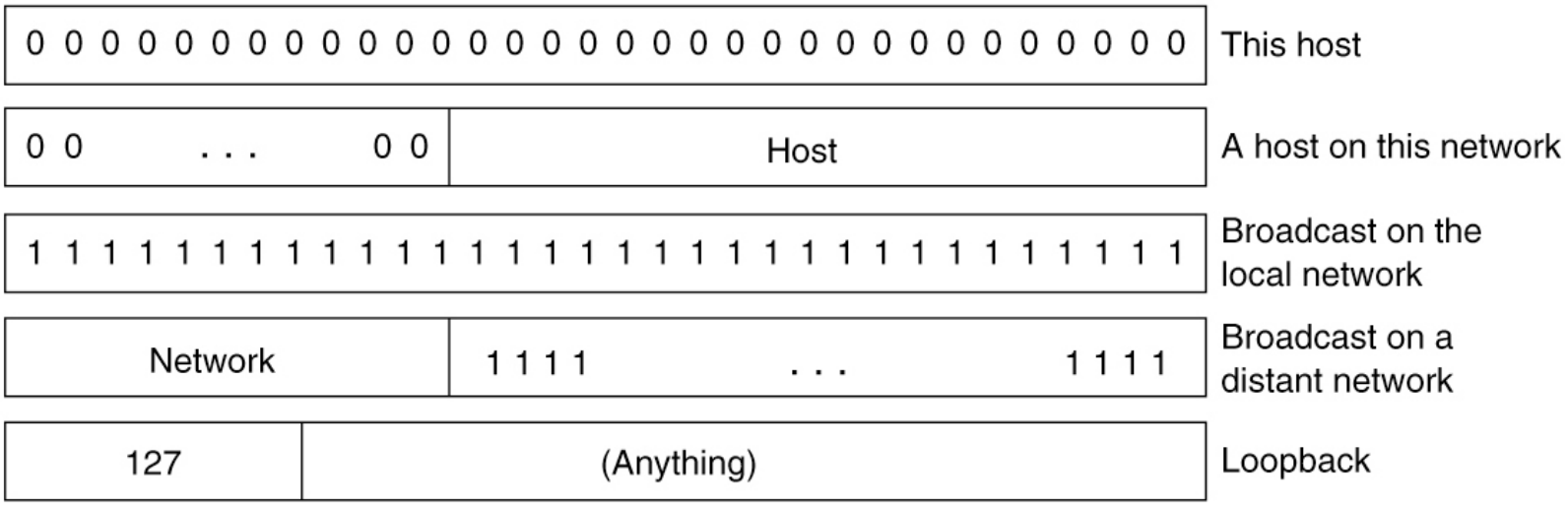


Classful and Special Addressing



IP address formats

Classful and Special Addressing



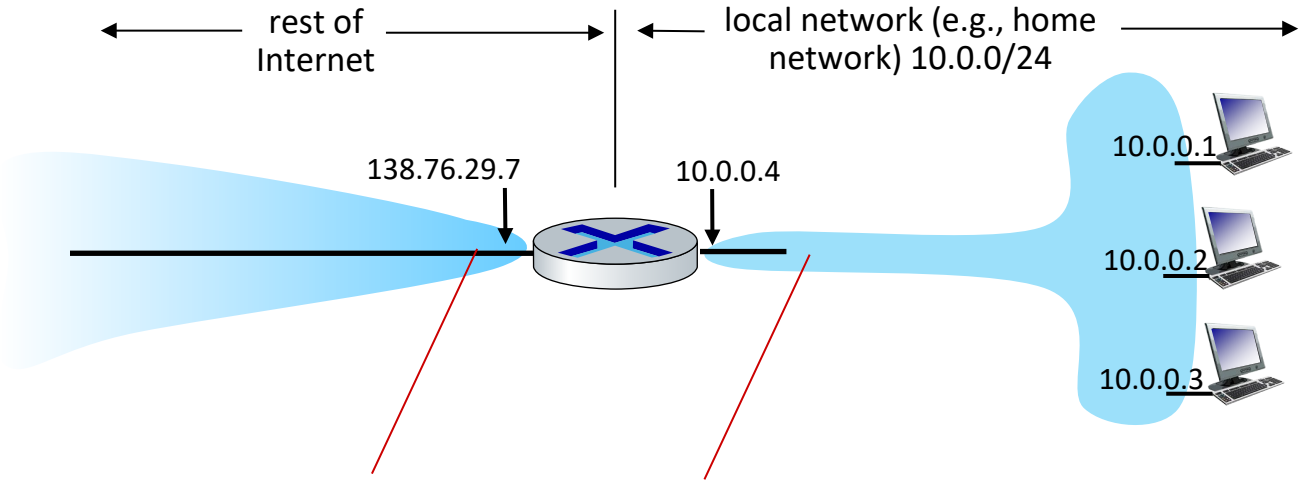
Special IP addresses

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

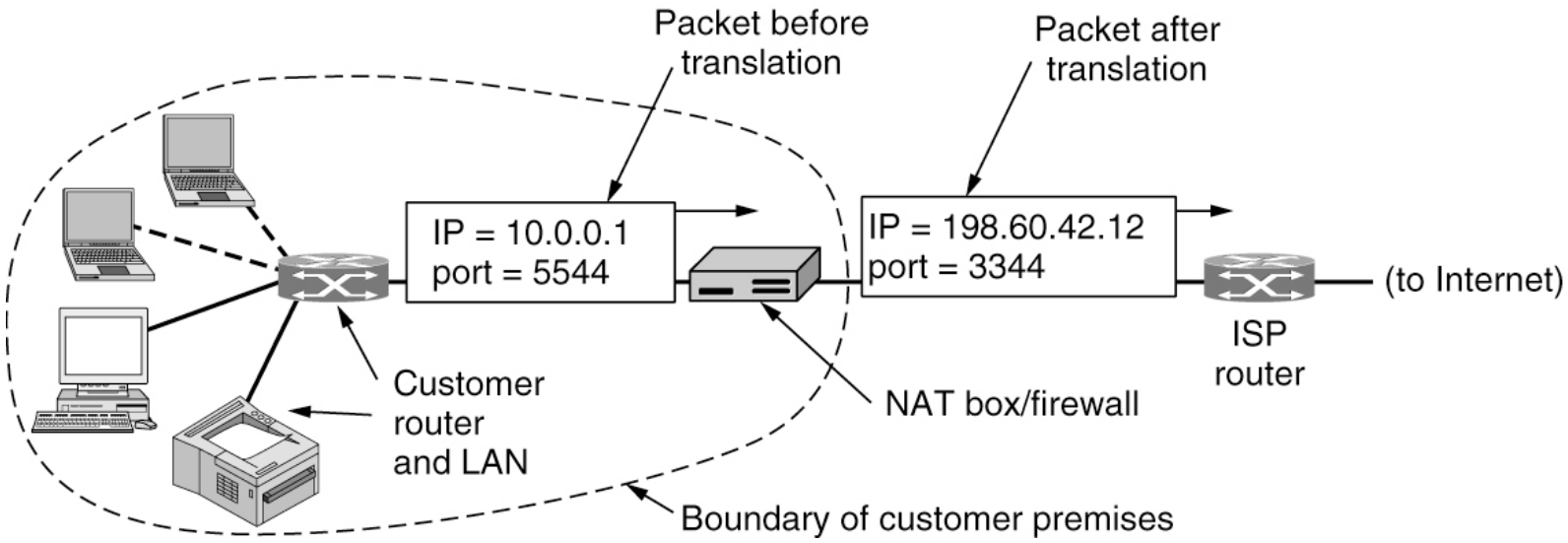
NAT: all devices in the local network share just one IPv4 address as far as outside world is concerned



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

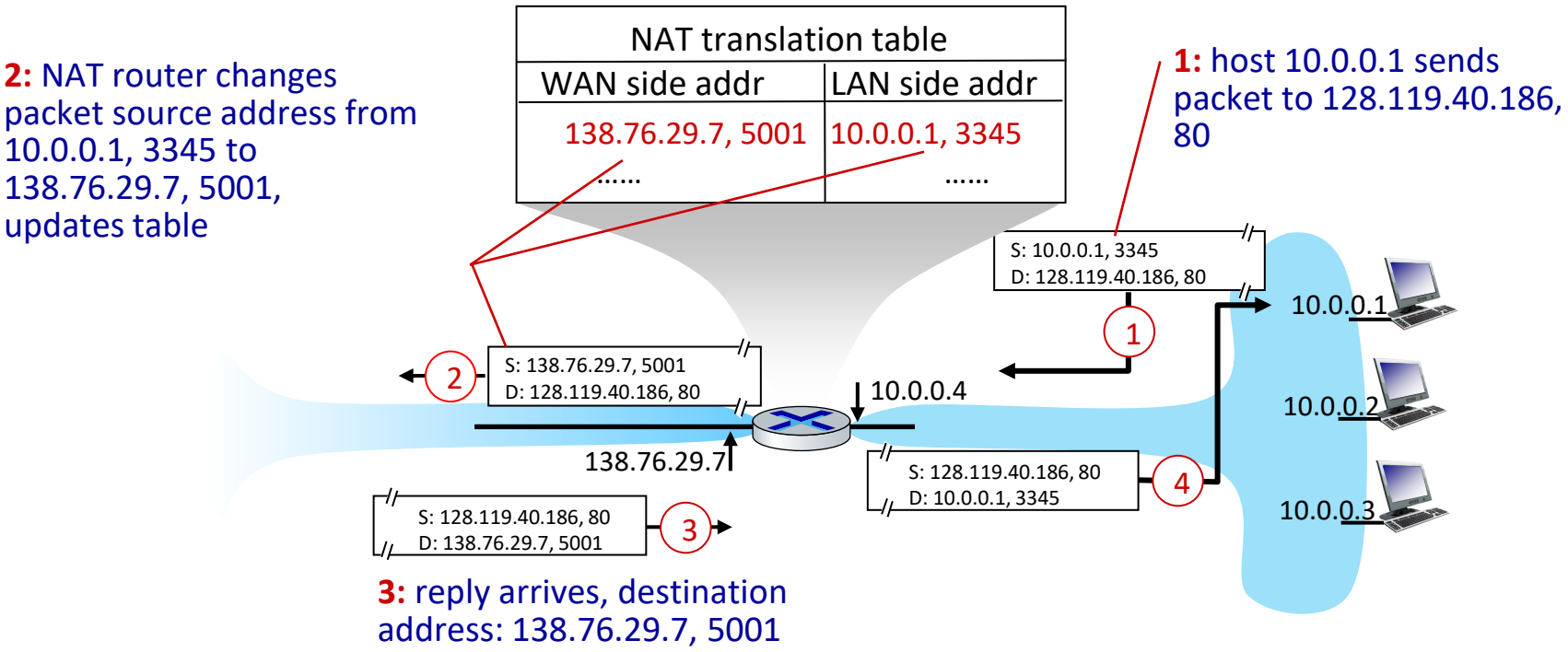


Placement and operation of a NAT box

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 corresponding (source IP address, port #) stored in the NAT table

NAT—Network Address Translation



NAT—Network Address Translation

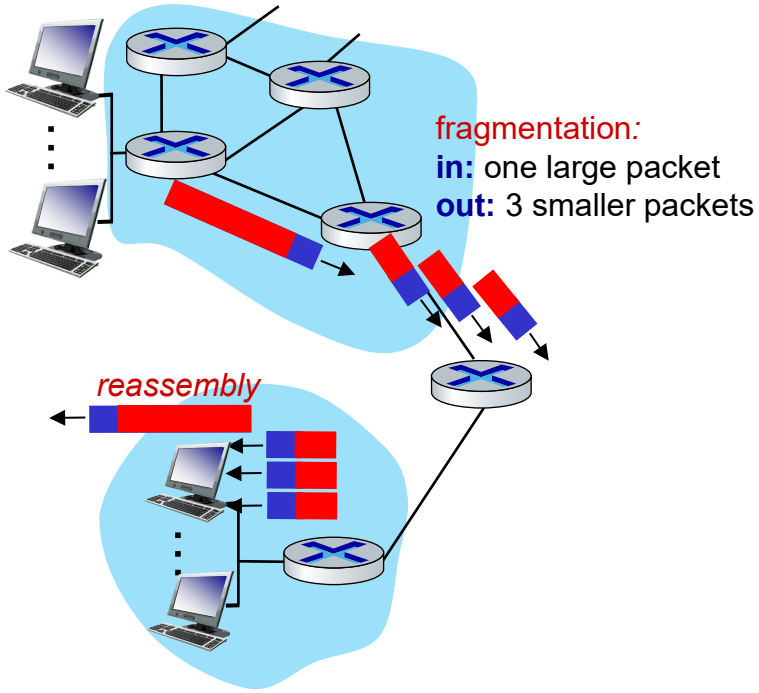
- NAT has been controversial
  - Routers should only process up to layer 3
  - Violates layer independence
  - Address shortage should be solved by IPv6
  - Violates end-to-end argument (port # manipulation by network-layer device)
  - Processes may use a protocol other than TCP or UDP
- But NAT is here 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 error induced retransmissions
  - Desire to prevent a packet from occupying the channel too long

Packet Fragmentation

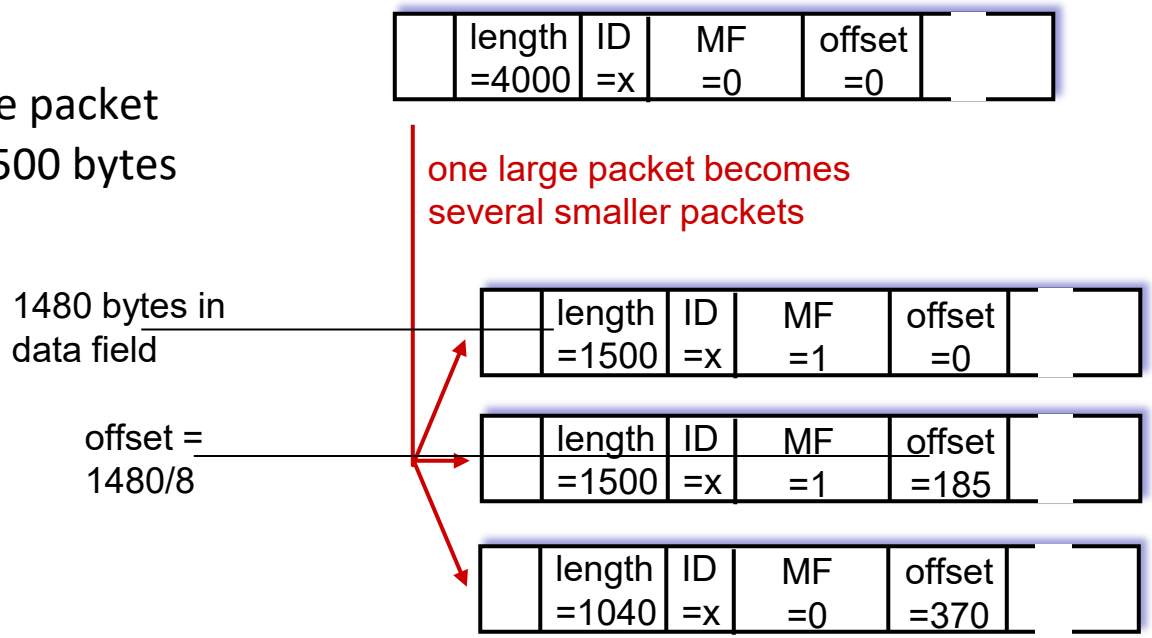
- Network links have a Maximum Transmission Unit (MTU)
  - Largest possible link-level frame
  - Different link types, different MTUs
- A large IP packet is divided (fragmented) so it becomes several packets
  - Reassembled only at the destination



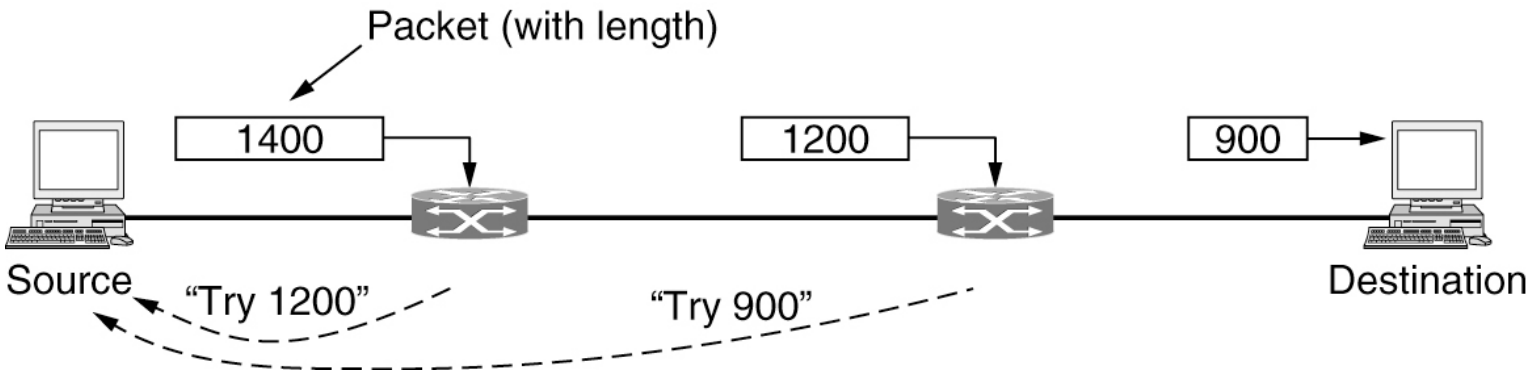
IP Packet Fragmentation

Example

- 4000 byte packet
- MTU = 1500 bytes



Packet Fragmentation



Path MTU discovery

Internet Control Protocols

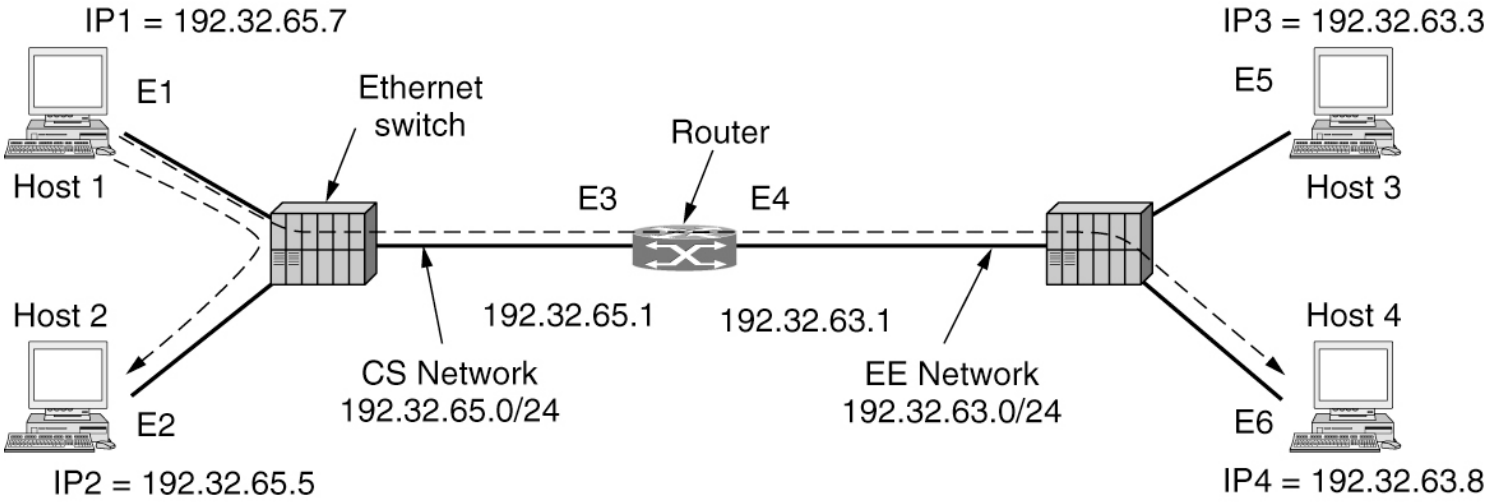
- ICMP—The Internet Control Message Protocol
- ARP—The Address Resolution Protocol
- DHCP—The Dynamic Host Configuration Protocol

ICMP—Internet Control Message Protocol

Message type	Description
Destination unreachable	Packet could not be delivered
Time exceeded	Time to live field hit 0
Parameter problem	Invalid header field
Source quench	Choke packet
Redirect	Teach a router about geography
Echo and echo reply	Check if a machine is alive
Timestamp request/reply	Same as Echo, but with timestamp
Router advertisement/solicitation	Find a nearby router

The principal ICMP message types

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

DHCP—Dynamic Host Configuration Protocol

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

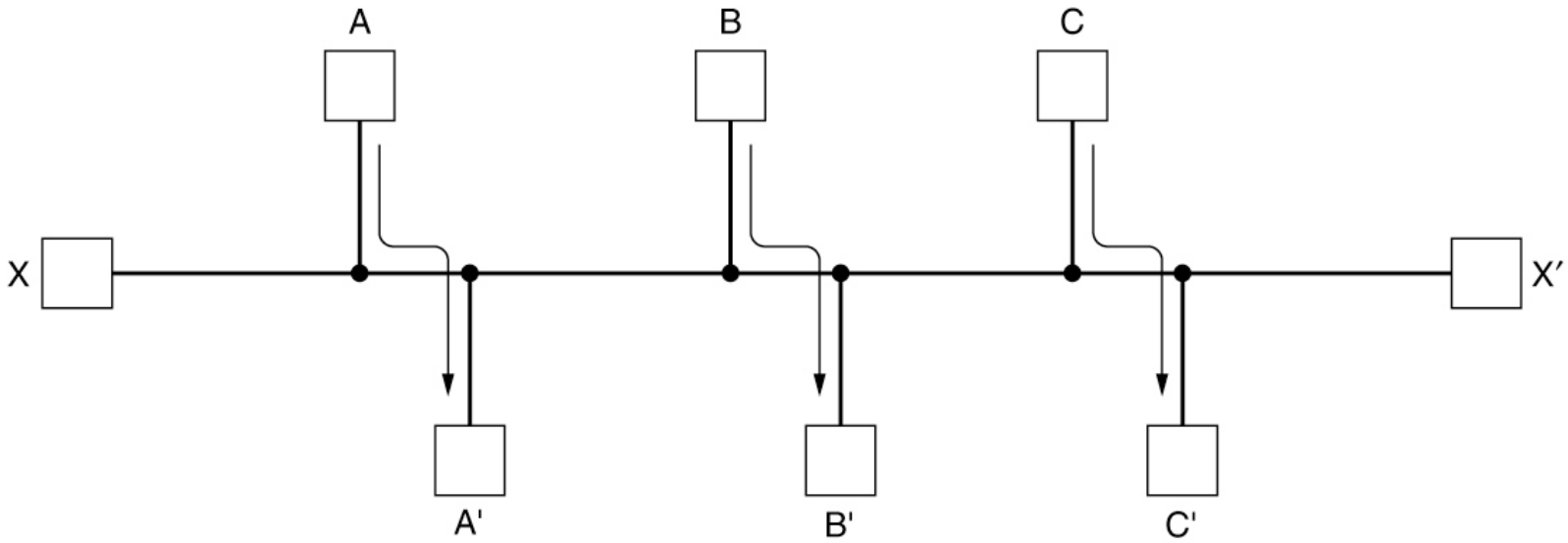
- Can renew its lease on address in use
- Allows reuse of addresses (only hold address while connected/on)
- Support for mobile users who join/leave network

- 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
  - Host requests IP address: DHCP request message
  - DHCP server sends address: DHCP ack message

Routing Algorithms

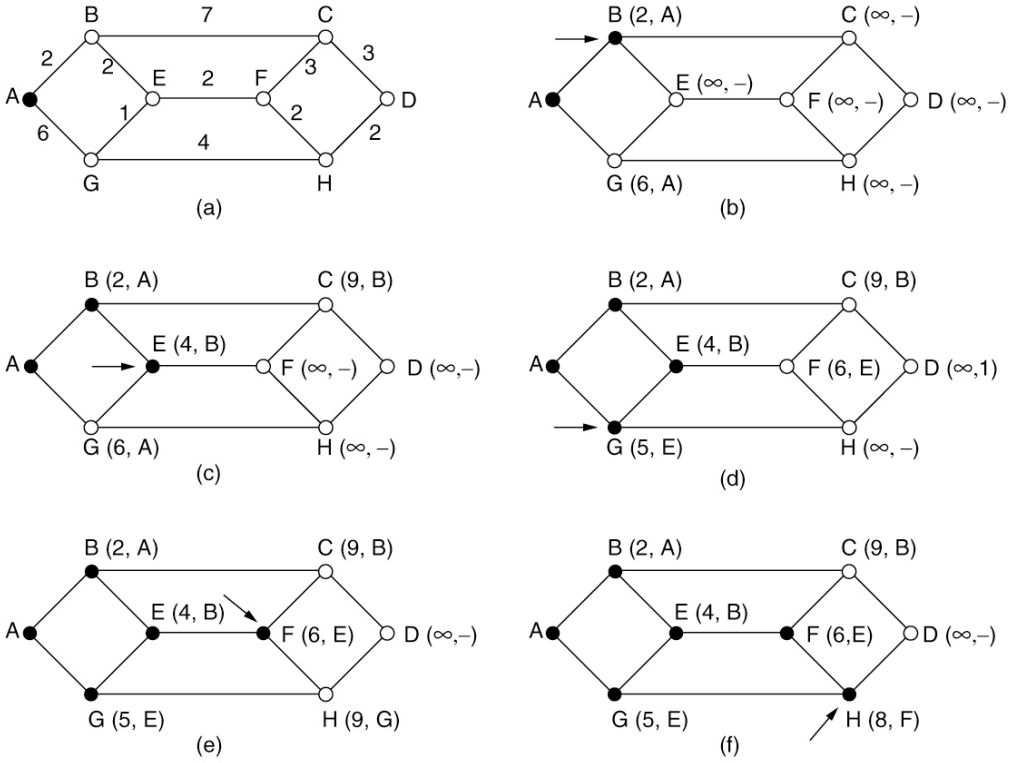
- Shortest path (Dijkstra's) algorithm
- Flooding
- Distance vector routing
- Link state routing

Routing Algorithms



Network with a conflict between fairness and efficiency

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 cost (measure) 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 cost 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

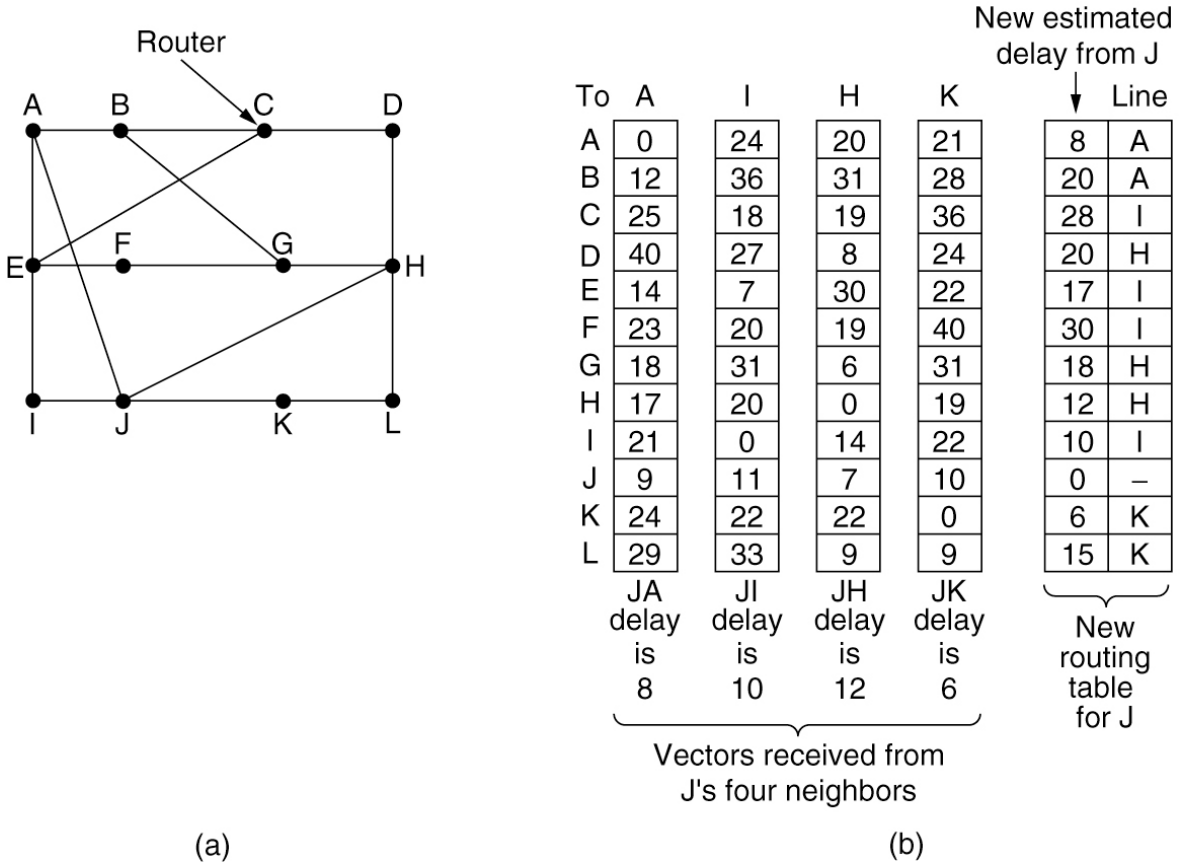
Distance Vector Routing

- Route exchange
  - A: I can reach X at cost (A,X)
  - B: I can reach X at cost (B,X)
  - A: I am cost (A,B) away from B
- Shortest path calculation
  - A: min{cost (A,X), cost (A,B) + cost (B,X)}

Distance Vector Routing

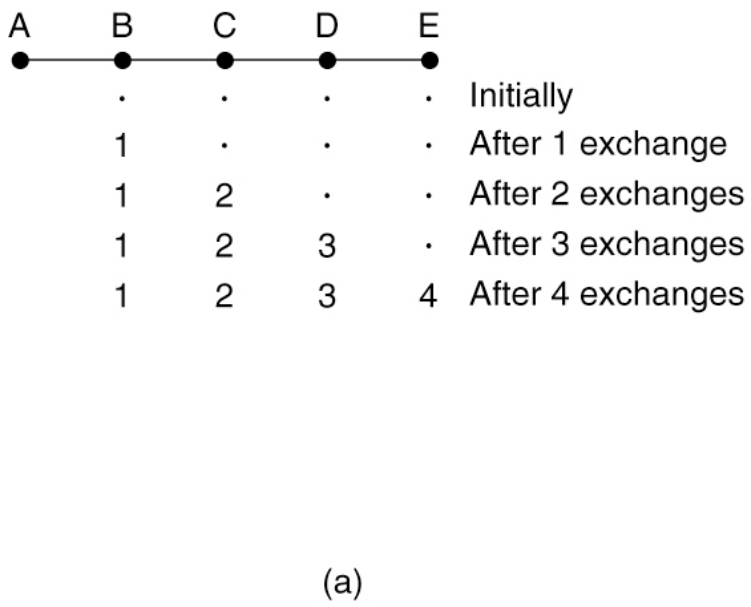
- This process continues until the network converges.
- This algorithm is simple to implement, and has low computational complexity and memory requirements
- Suitable for small or stable networks
- It can have slow convergence, especially after failures
- Count-to-infinity problem
  - Mitigation: Route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

Distance Vector Routing

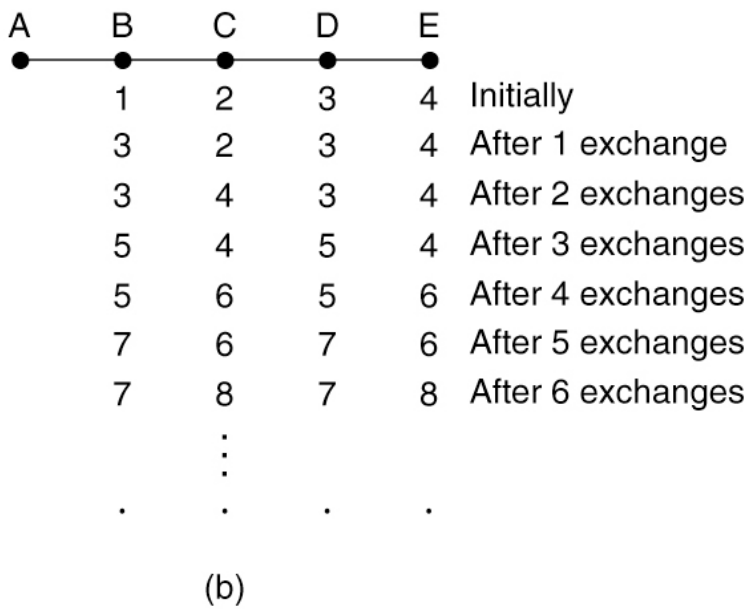


(a) A network. (b) Input from *A*, *I*, *H*, *K*, and the new routing table for *J*.

The Count-to-Infinity Problem



(a)



(b)

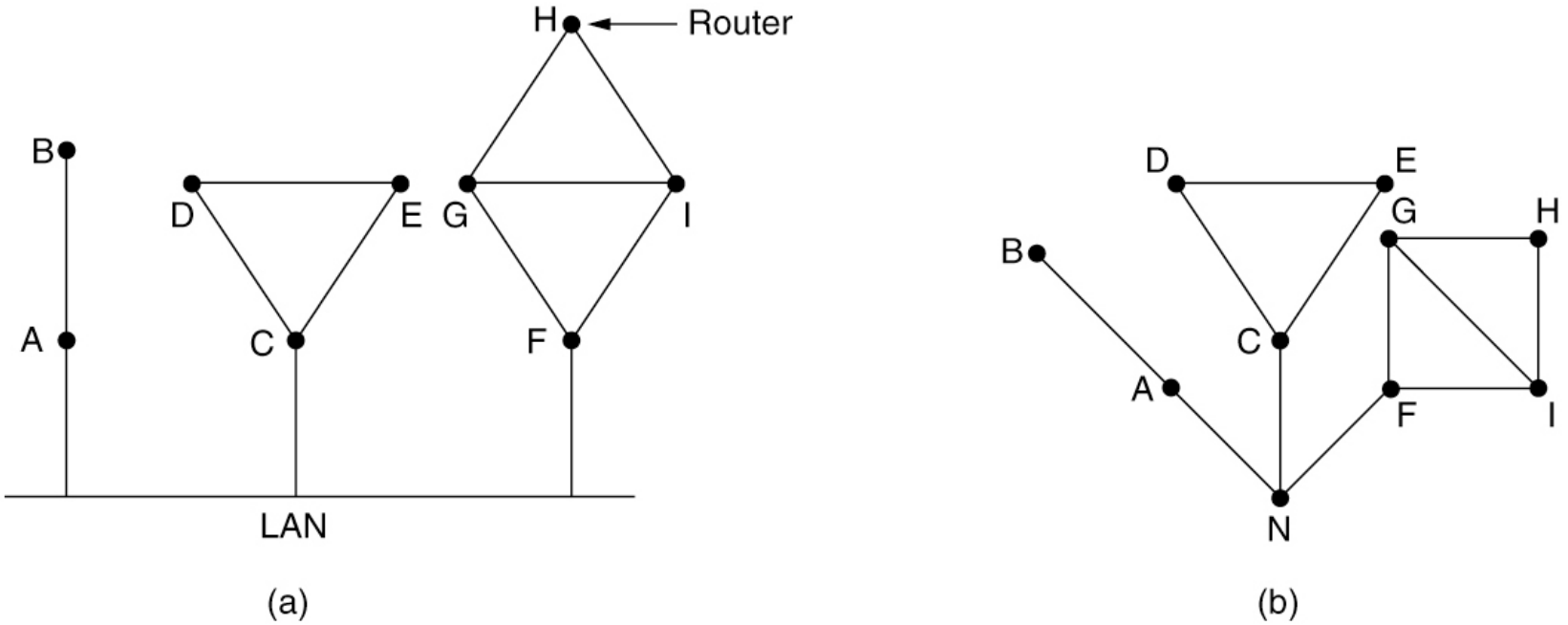
Route Poisoning

- Choose a suitable value for infinity
- Poisoned 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

Link State Routing

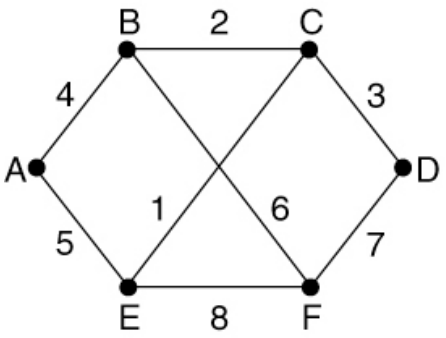
- Discover neighbors, learn network addresses
- Set distance/cost metric to each neighbor
- Construct packet telling all it has learned
- Send packet to, receive packets from other routers
- Compute shortest path to every other router

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

Source	Seq.	Age	Send flags A	Send flags C	Send flags F	ACK flags A	ACK flags C	ACK flags 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 varies

Speed of convergence

LS: $O(n^2)$ algorithm, $O(n^2)$ messages
- may have oscillations

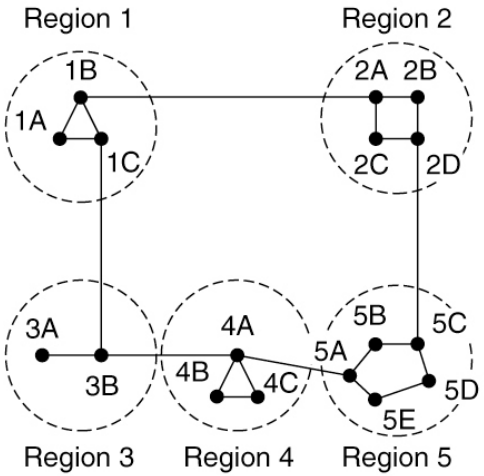
DV: convergence time varies
- may have routing loops
- count-to-infinity problem

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-holing*
- each router's DV is used by others: error propagate thru network

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 its own network

Internet Approach to Scalable Routing

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

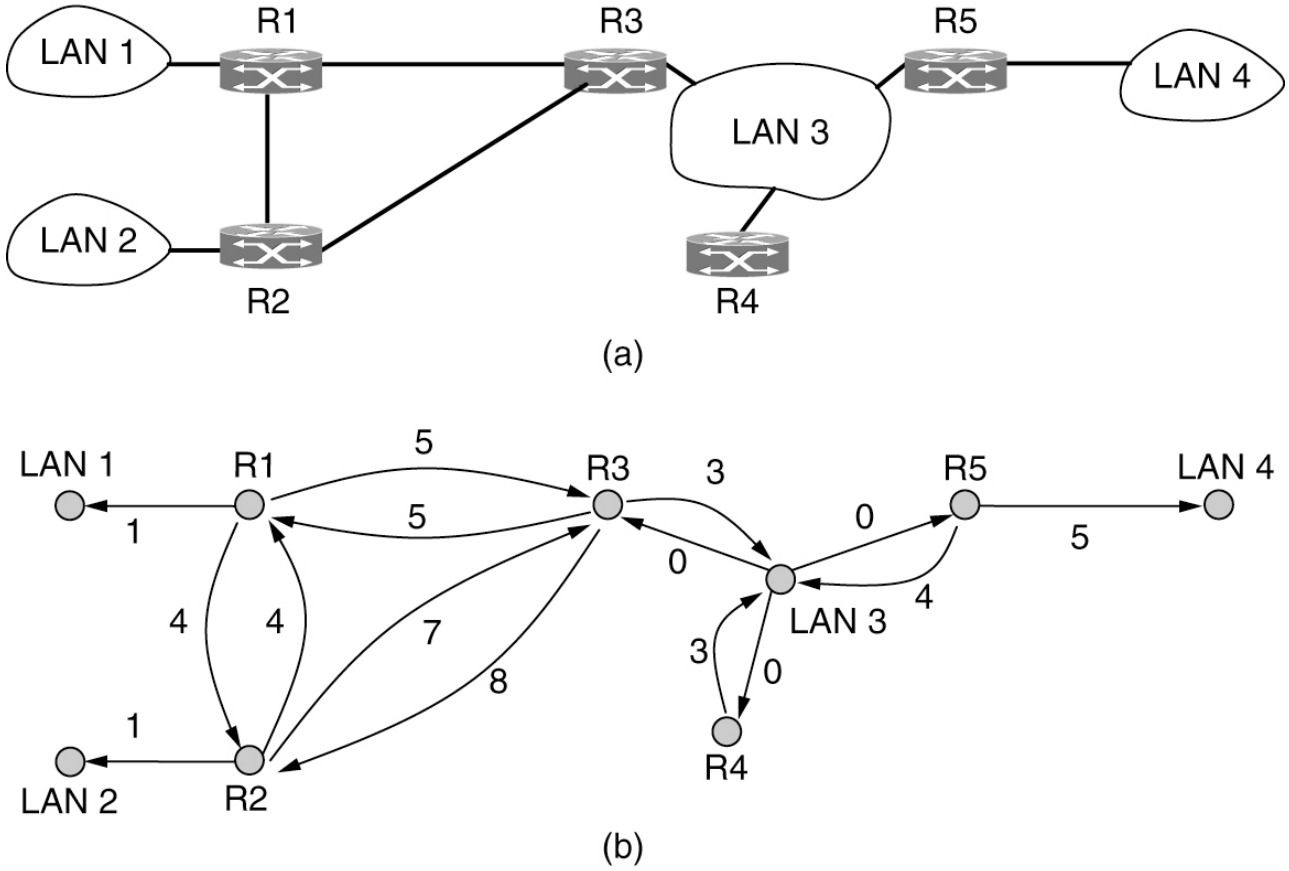
Routing Information Protocol (RIP)

- Uses distance vector routing
- Included in BSD Unix in 1982; max hops: 15
- Distance vector
  - Exchanged between neighbors every 30 s
  - Up to 25 destinations within an RIP packet (UDP 520)
- 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

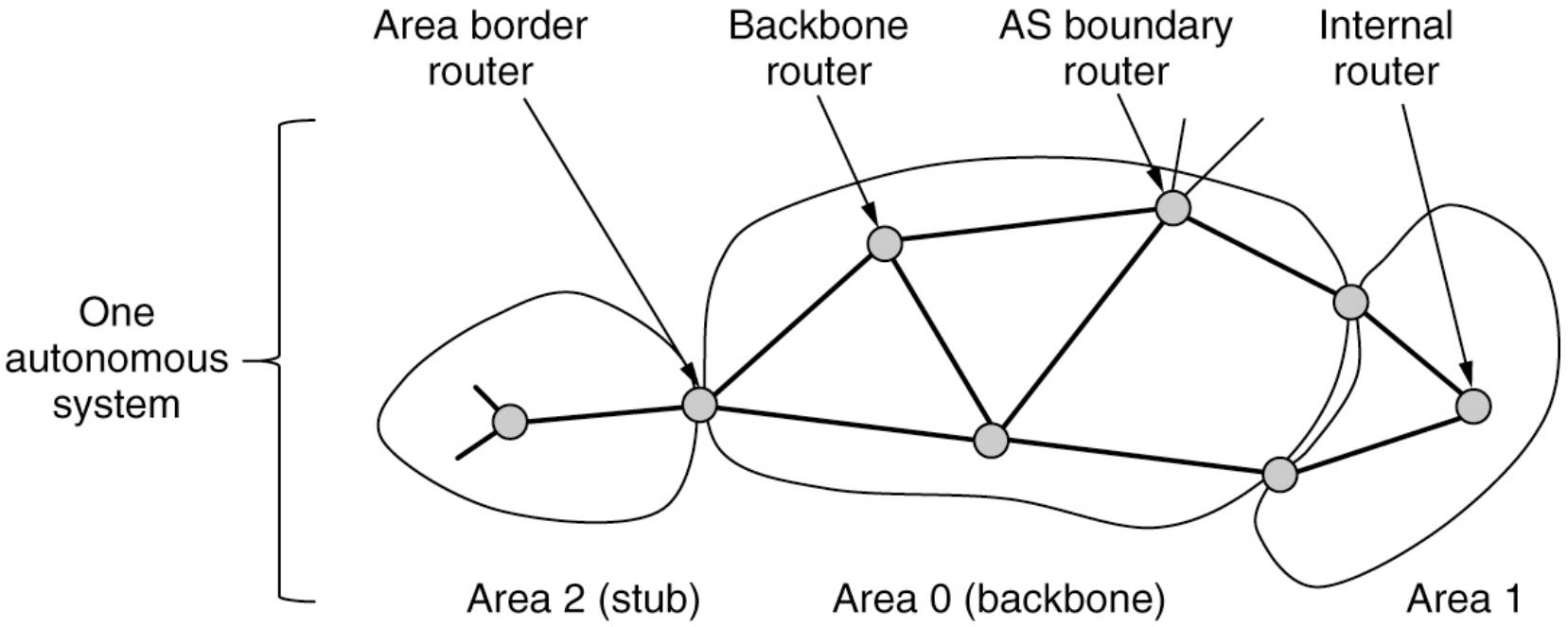
- Open: publicly available
- Link-state based
  - Each router floods OSPF link-state advertisements directly over IP to all other routers in entire AS
  - Supports a variety of distance metrics, e.g. bandwidth, delay
  - Each router has full topology, uses Dijkstra's algorithm to compute forwarding table
- 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) An autonomous system. (b) A graph representation of (a).

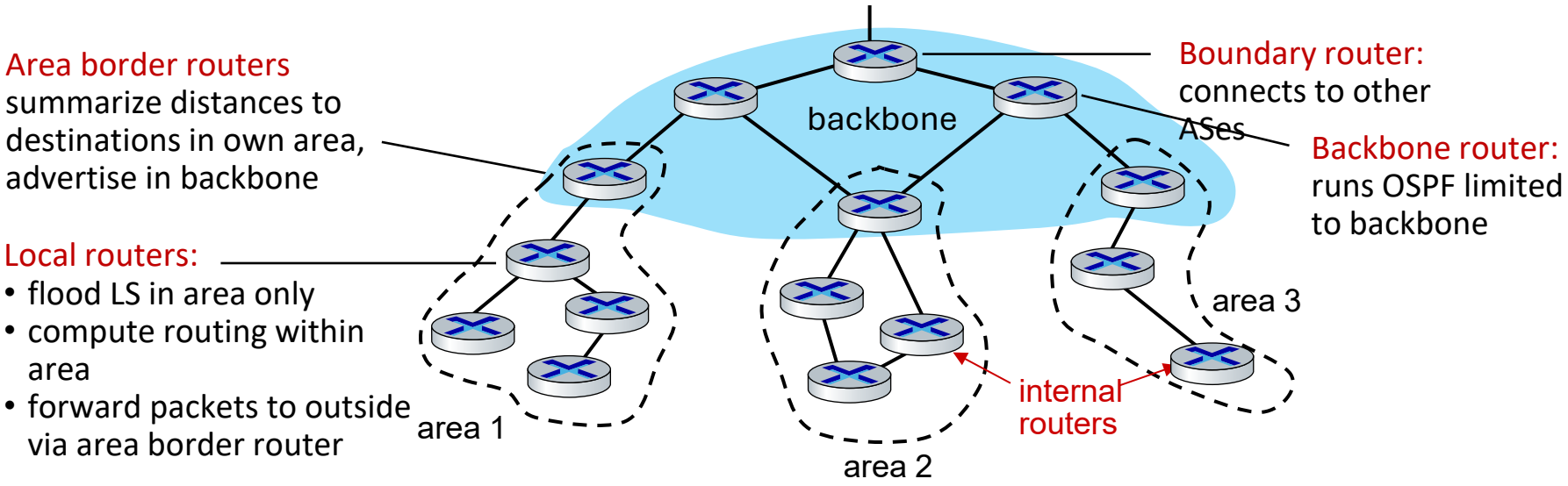
OSPF—Interior Gateway Routing Protocol



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



BGP—Exterior Gateway Routing Protocol

- The de facto inter-domain routing protocol
  - glue that holds the Internet together
- Allows subnets to advertise their existence, and the destinations it can reach, to 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 info 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 info

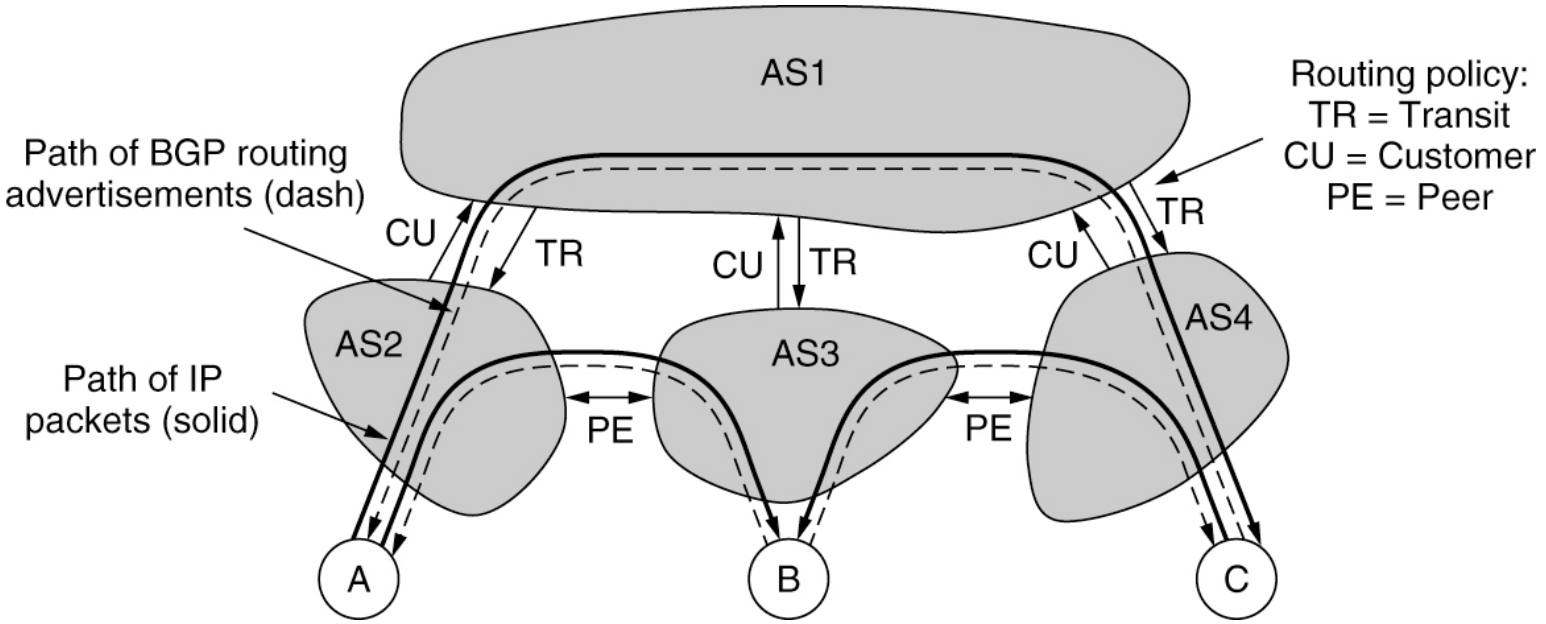
BGP—Exterior Gateway Routing Protocol

- 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

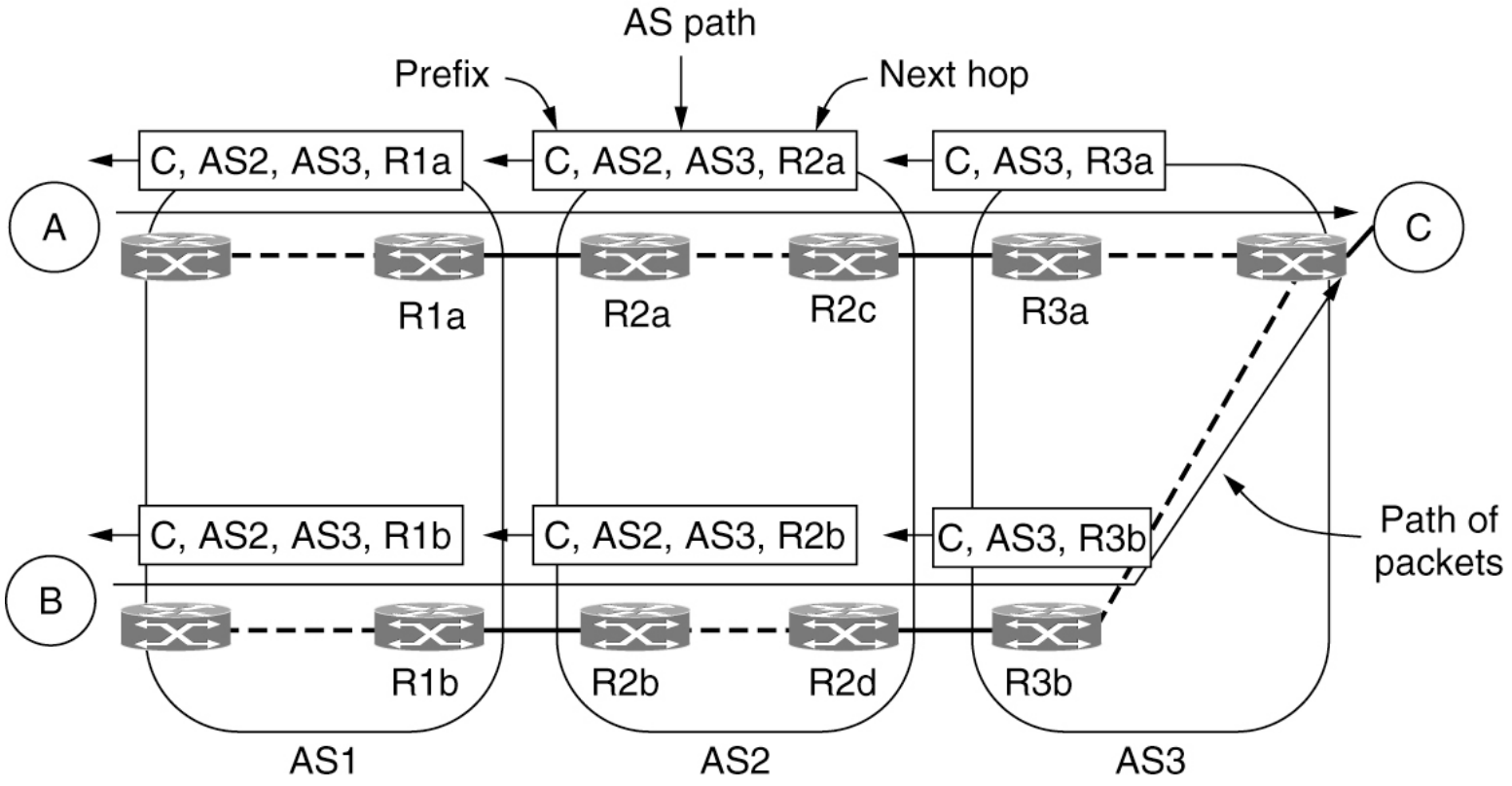
- BGP advertised route: prefix + attributes
  - prefix: destination being advertised
  - two important attributes
  - AS-PATH: list of ASes through which prefix advertisement has passed
  - NEXT-HOP: indicates specific internal-AS router to next-hop AS
- policy-based routing:
  - gateway receiving route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

BGP—Exterior Gateway Routing Protocol



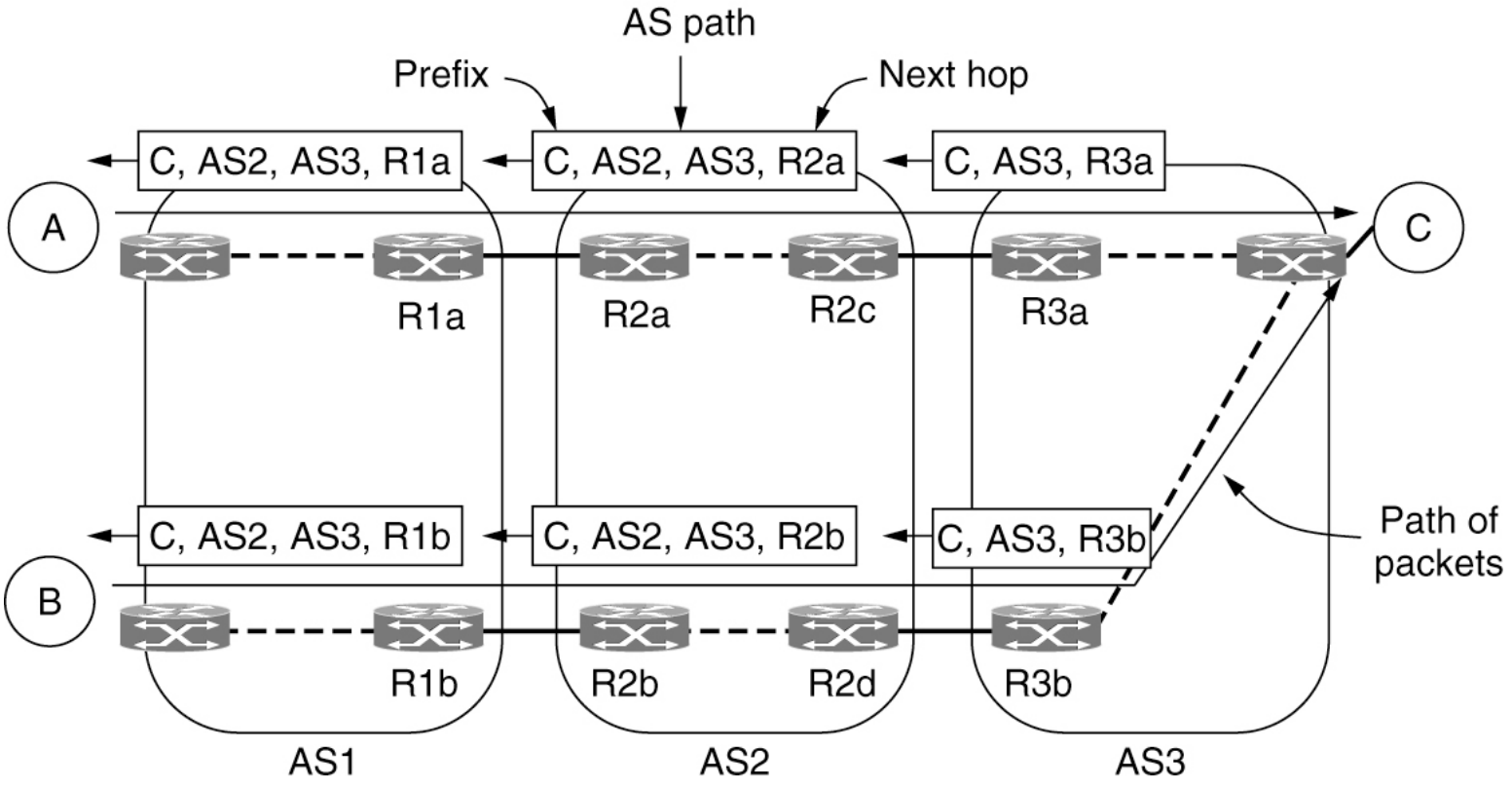
Routing policies between four autonomous systems

BGP—Exterior Gateway Routing Protocol



Propagation of BGP route advertisements

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