

Network Layer

Chapter 5

- Design Issues
- Routing Algorithms
- Congestion Control
- Quality of Service
- Internetworking
- Network Layer of the Internet

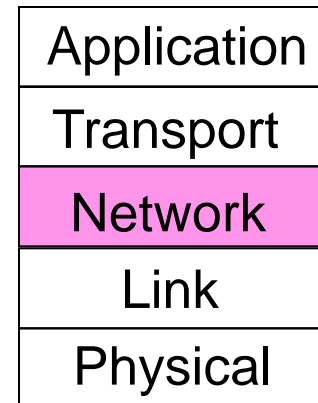
Revised: August 2011

The Network Layer

Responsible for delivering packets between endpoints over multiple links

Network Layer is the lowest layer in the OSI Reference Model that deals with end-to-end transmission.

It provides services to the Transport Layer.

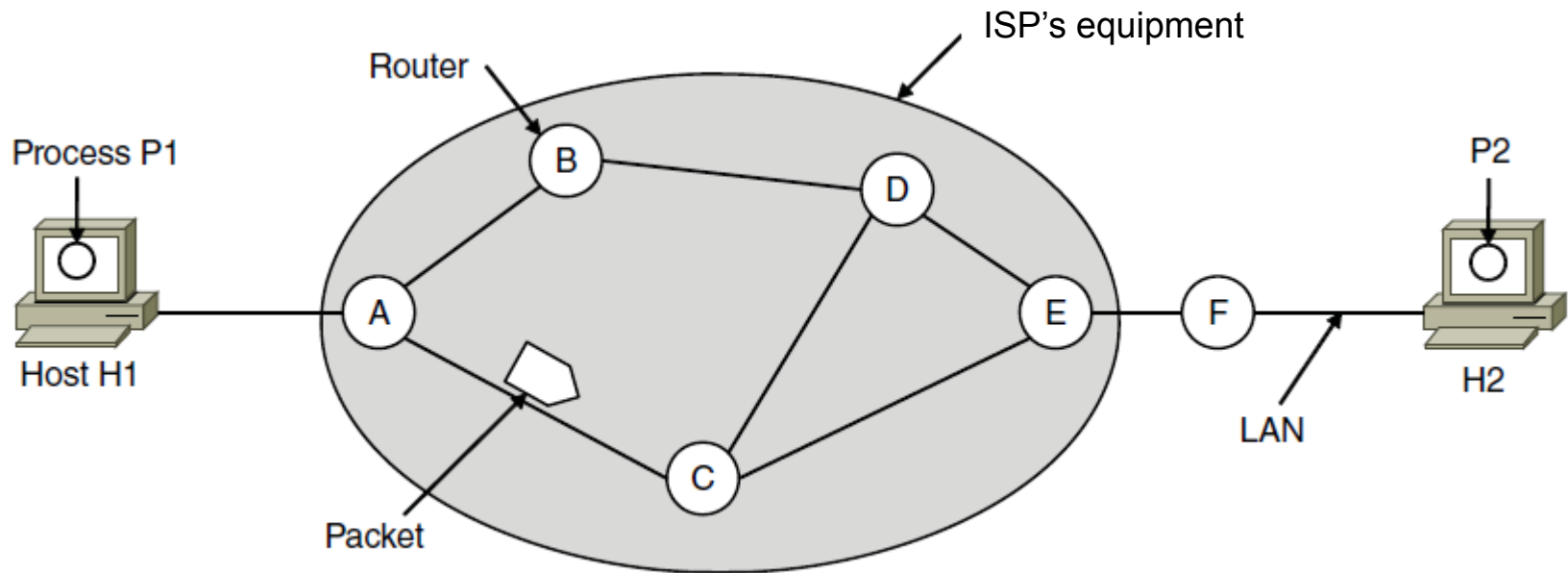


Design Issues

- Store-and-forward packet switching »
- Connectionless service – datagrams »
- Connection-oriented service – virtual circuits »
- Comparison of virtual-circuits and datagrams »

Store-and-Forward Packet Switching

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



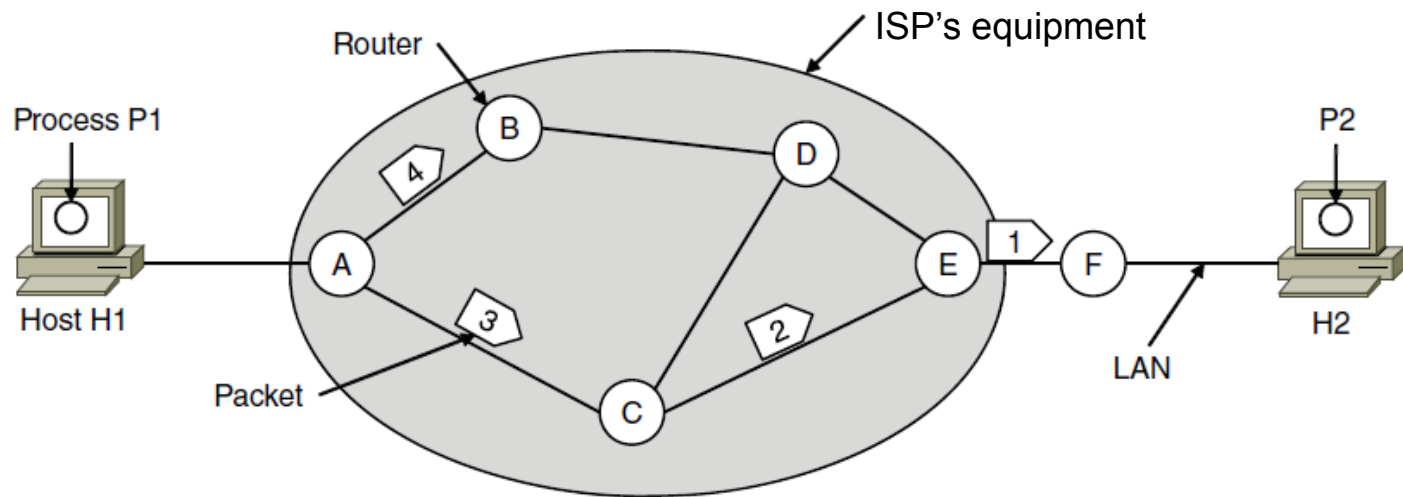
Questions: If P1 on Host H1 is sending a message to P2 on H2, for the packet at Host H1:

- What is the destination address for the packet's network layer?
- What is the destination address for the packet's data link layer?

Connectionless Service – Datagrams

Packet is forwarded using destination address inside it

- Different packets may take different paths



A's table (initially)

A	
B	B
C	C
D	B
E	C
F	C

Dest. Line

A's table (later)

A	
B	B
C	C
D	B
E	D
F	D

C's Table

A	A
B	A
C	
D	E
E	E
F	E

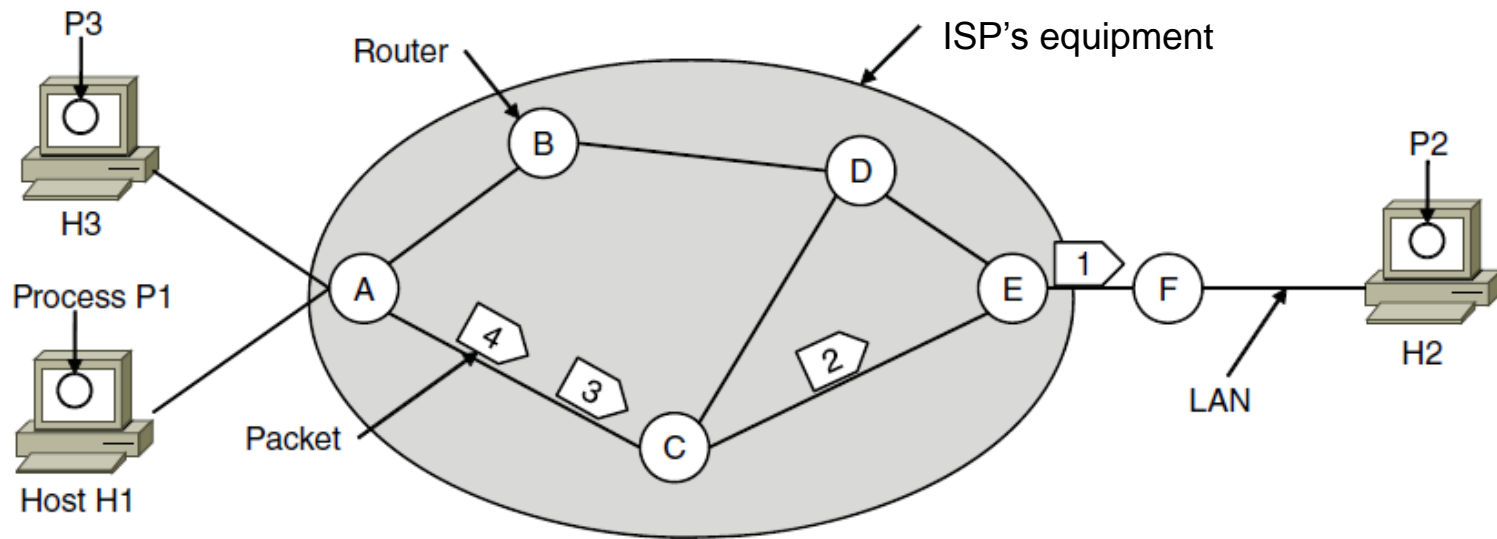
E's Table

A	C
B	D
C	C
D	D
E	
F	F

Connection-Oriented – Virtual Circuits

Packet is forwarded along a virtual circuit using tag inside it

- Virtual circuit (VC) is set up ahead of time



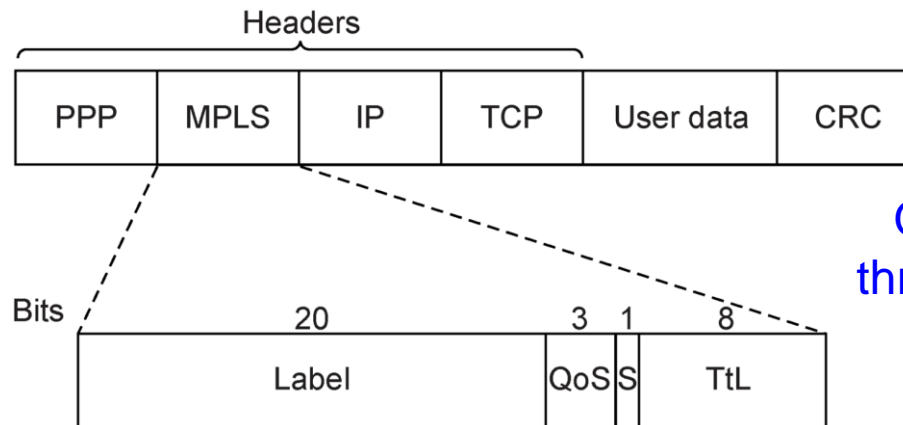
A's table				C's Table				E's Table			
H1	1	C	1	A	1	E	1	C	1	F	1
H3	1	C	2	A	2	E	2	C	2	F	2
In: Line		Tag	Line	In: Line		Tag	Line	In: Line		Tag	Line

Question: For the Internet Protocol Suite, is there ANY connection-oriented protocol at the Network Layer whatsoever?

CONS in the Internet

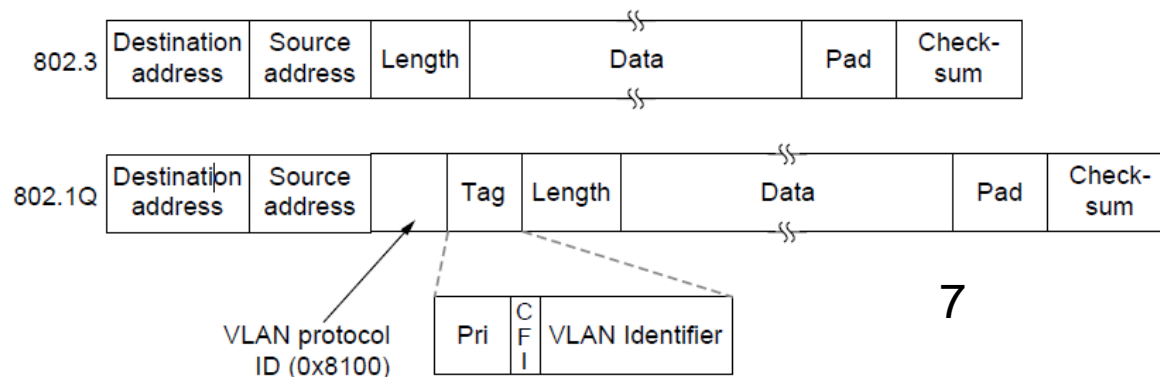
In addition to telephony, Tannenbaum argues that there are at least two other examples of connection-oriented protocols in the Internet:

1. MultiProtocol Label Switching (MPLS) – see pages 471-474



Question: Are any of these three Network Layer protocols within the Internet Protocol Suite?

2. Virtual LANS (VLANs) – see pages 342-349



Comparison of Virtual-Circuits & Datagrams

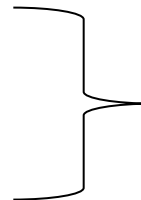
Issue	Datagram network	Virtual-circuit network
Circuit setup	Not needed	Required
Addressing	Each packet contains the full source and destination address	Each packet contains a short VC number
State information	Routers do not hold state information about connections	Each VC requires router table space 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

Routing Algorithms (1)

Routing logically comprises two processes:

1. **Forwarding**: processing arriving packets by looking up appropriate outgoing link to use from routing tables
2. Filling in and updating the routing tables. This is where **routing algorithms** occur.

- Optimality principle »
- Shortest path algorithm »
- Flooding »
- Distance vector routing »
- Link state routing »
- Hierarchical routing »
- Broadcast routing »
- Multicast routing »
- Anycast routing »
- Routing for mobile hosts »
- Routing in ad hoc networks »



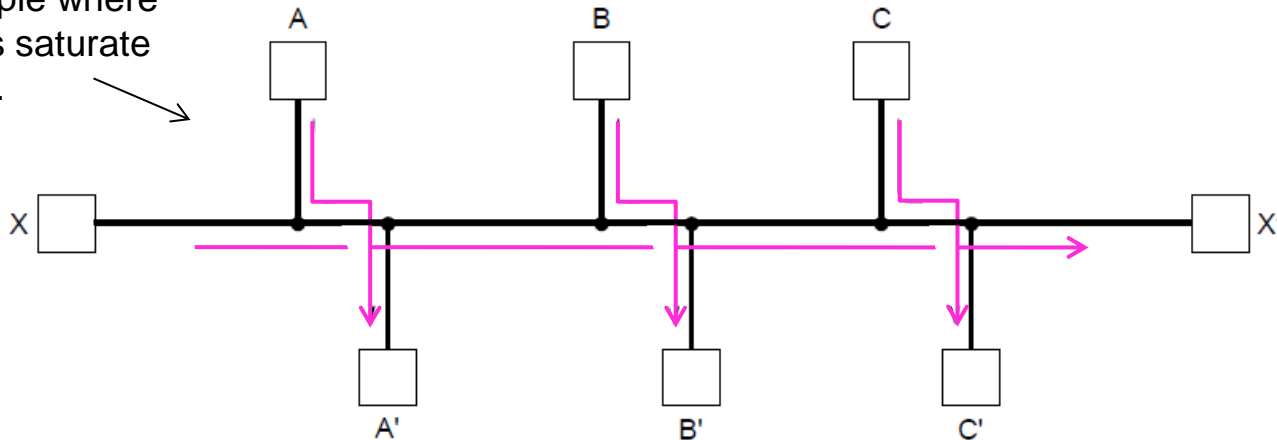
Routing Algorithms

Routing Algorithms (2)

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)
- Update routes for changes in topology (e.g., failures)

Fairness Example where
vertical Comms saturate
horizontal links.

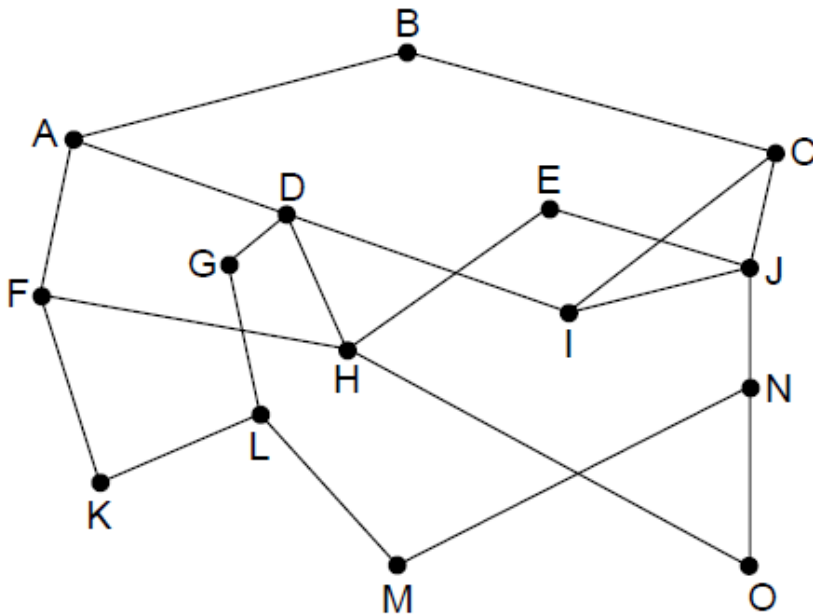


Forwarding is the sending of packets along a path

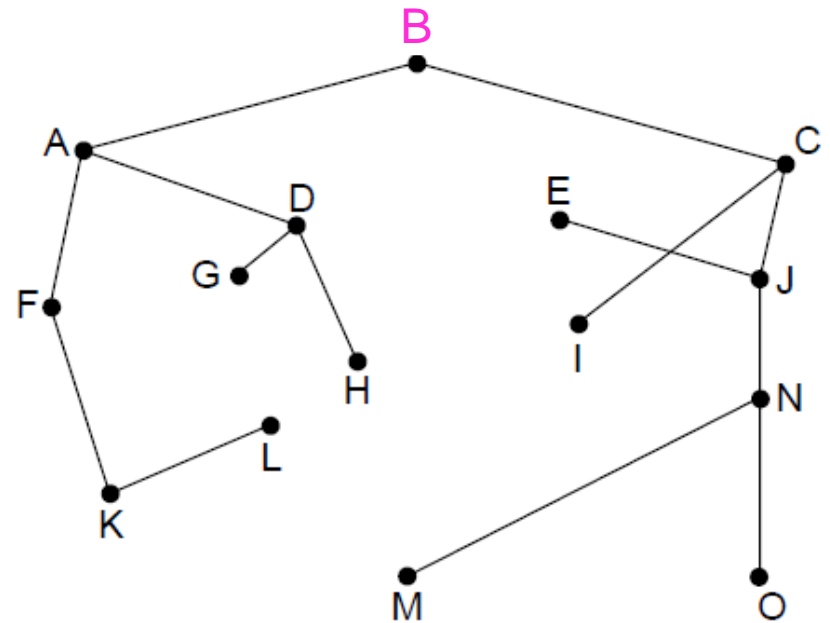
The Optimality Principle

Each portion of a best path is also a best path; the union of them to a router is a tree called the sink tree

- Best means fewest hops in the example



Network



Sink tree of best paths to router B

The goal of all routing algorithms is to discover and use either sink trees or Directed Acyclic Graphs (DAG) to eliminate routing loops for all routers. DAGs are like sink trees except they allow all non-looping possible paths to be chosen in graphs.

Shortest Path Algorithm (1)

Shortest path selects the most efficient path through a graph in terms of a specific metric used by that Autonomous System (AS, e.g., number hops, distance, latency, bandwidth, average delay, comm cost, measured delay).

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

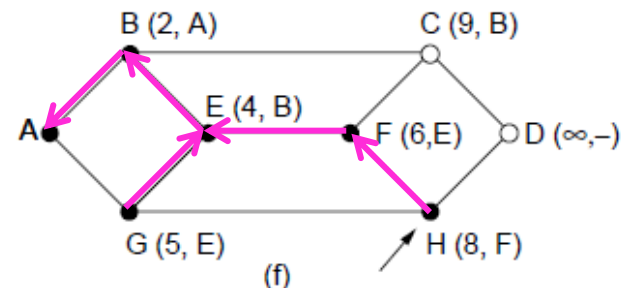
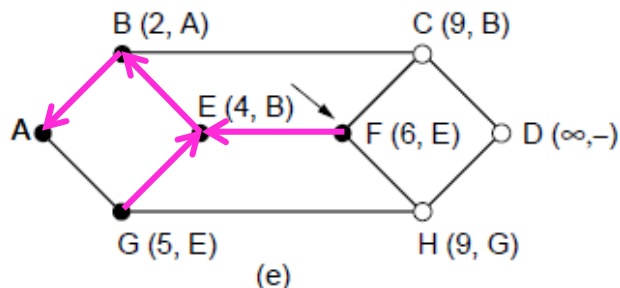
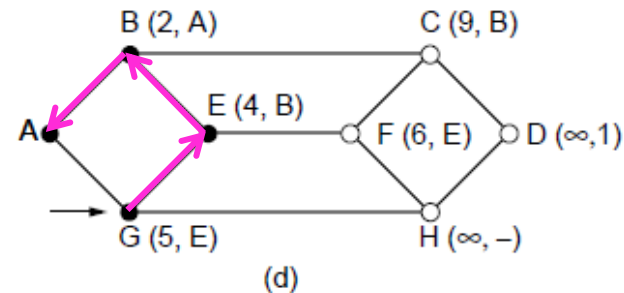
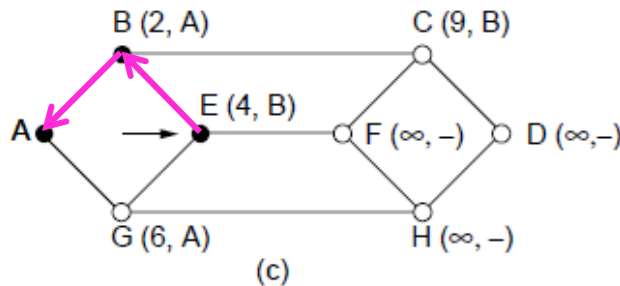
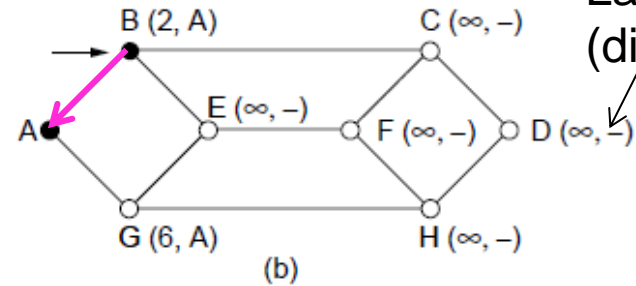
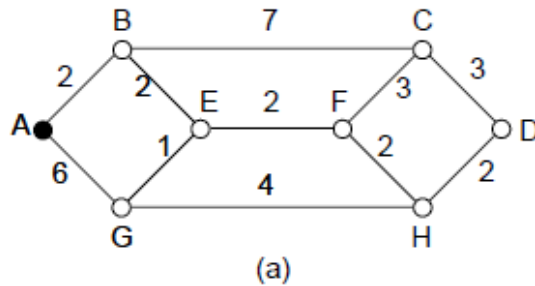
Algorithm:

- Start with sink, set distance at other nodes to infinity
- Relax (i.e., evaluate) distance to adjacent nodes
- Pick the lowest adjacent distance node, add it to sink tree
- Repeat until all nodes are in the sink tree

Shortest Path Algorithm (2)

Labels =
(distance, path)

Start
at Sink
and compute
backwards



A weighted, undirected graph of a network and the first five steps in computing the shortest paths from A to D. Pink arrows show the sink tree so far.

Note: Dijkstra's Algorithm == Shortest Path Algorithm

Shortest Path Algorithm (3)

...

```
for (p = &state[0]; p < &state[n]; p++) {  
    p->predecessor = -1;  
    p->length = INFINITY;  
    p->label = tentative;  
}
```

```
state[t].length = 0; state[t].label = permanent;
```

```
k = t;
```

```
do {
```

```
    for (i = 0; i < n; i++)
```

```
        if (dist[k][i] != 0 && state[i].label == tentative) {
```

```
            if (state[k].length + dist[k][i] < state[i].length) {
```

```
                state[i].predecessor = k;
```

```
                state[i].length = state[k].length + dist[k][i];
```

```
            }
```

```
        }
```

...

Start with the sink,
all other nodes are
unreachable

Relaxation step.
Lower distance to
nodes linked to
newest member of
the sink tree

Shortest Path Algorithm (4)

...

```
k = 0; min = INFINITY;  
for (i = 0; i < n; i++)  
    if (state[i].label == tentative && state[i].length < min) {  
        min = state[i].length;  
        k = i;  
    }  
    state[k].label = permanent;  
} while (k != s);
```

Find the lowest distance, add it to the sink tree, and repeat until done

Flooding

Flooding is SOLEY used by routing protocols at the IP Layer. For example, it is used by the Protocol Independent Multicast – Dense Mode (PIM-DM) routing protocol (i.e., flood and prune to create multicast paths). Flooding is NOT a service that is available to end users.

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 flood; even using a hop limit can blow up exponentially

Distance Vector Routing (1)

The Border Gateway Protocol (BGP) uses distance vector routing. BGP is the Inter-Domain Routing Protocol used by the Internet (i.e., the protocol used to route between Autonomous Systems (AS)).

Distance Vector Routing uses the Bellman-Ford routing algorithm.

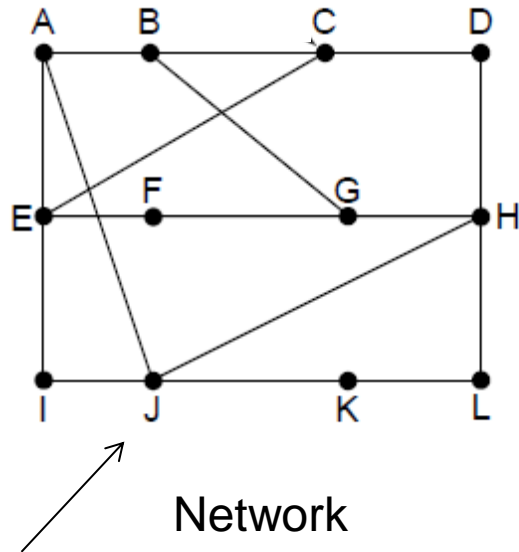
Distance vector is a distributed routing algorithm

- Shortest path computation is split across nodes (each router maintains its own routing table giving the best known distance (and link to use) to every router in the network).

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 (2)



					New estimated delay from J	
To	A	I	H	K	↓ 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

Distance Vector (DV) algorithm has a convergence issue in that it can converge to a correct routing map slowly because it reacts rapidly to good news but leisurely to bad news

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

A	B	C	D	E	
•	•	•	•	•	Initially
	•	•	•	•	After 1 exchange
	1	•	•	•	After 2 exchanges
	1	2	•	•	After 3 exchanges
	1	2	3	•	After 4 exchanges
	1	2	3	4	After 5 exchanges

Good news of a path to A spreads quickly

Router A is 4 routers away from Router E. The example is in terms of the implications to Router's routing entry for A and metric is routing hops.

A	B	C	D	E	
•	•	•	•	•	Initially
	1	2	3	4	After 1 exchange
	3	2	3	4	After 2 exchanges
	3	4	3	4	After 3 exchanges
	5	4	5	4	After 4 exchanges
	5	6	5	6	After 5 exchanges
	7	6	7	6	After 6 exchanges
	7	8	7	8	After 7 exchanges
	7	8	7	8	After 8 exchanges
	7	8	7	8	After 9 exchanges
	7	8	7	8	After 10 exchanges
	7	8	7	8	After 11 exchanges
	7	8	7	8	After 12 exchanges
	7	8	7	8	After 13 exchanges
	7	8	7	8	After 14 exchanges
	7	8	7	8	After 15 exchanges
	7	8	7	8	After 16 exchanges
	7	8	7	8	After 17 exchanges
	7	8	7	8	After 18 exchanges
	7	8	7	8	After 19 exchanges
	7	8	7	8	After 20 exchanges
	7	8	7	8	After 21 exchanges
	7	8	7	8	After 22 exchanges
	7	8	7	8	After 23 exchanges
	7	8	7	8	After 24 exchanges
	7	8	7	8	After 25 exchanges
	7	8	7	8	After 26 exchanges
	7	8	7	8	After 27 exchanges
	7	8	7	8	After 28 exchanges
	7	8	7	8	After 29 exchanges
	7	8	7	8	After 30 exchanges
	7	8	7	8	After 31 exchanges
	7	8	7	8	After 32 exchanges
	7	8	7	8	After 33 exchanges
	7	8	7	8	After 34 exchanges
	7	8	7	8	After 35 exchanges
	7	8	7	8	After 36 exchanges
	7	8	7	8	After 37 exchanges
	7	8	7	8	After 38 exchanges
	7	8	7	8	After 39 exchanges
	7	8	7	8	After 40 exchanges
	7	8	7	8	After 41 exchanges
	7	8	7	8	After 42 exchanges
	7	8	7	8	After 43 exchanges
	7	8	7	8	After 44 exchanges
	7	8	7	8	After 45 exchanges
	7	8	7	8	After 46 exchanges
	7	8	7	8	After 47 exchanges
	7	8	7	8	After 48 exchanges
	7	8	7	8	After 49 exchanges
	7	8	7	8	After 50 exchanges
	7	8	7	8	After 51 exchanges
	7	8	7	8	After 52 exchanges
	7	8	7	8	After 53 exchanges
	7	8	7	8	After 54 exchanges
	7	8	7	8	After 55 exchanges
	7	8	7	8	After 56 exchanges
	7	8	7	8	After 57 exchanges
	7	8	7	8	After 58 exchanges
	7	8	7	8	After 59 exchanges
	7	8	7	8	After 60 exchanges
	7	8	7	8	After 61 exchanges
	7	8	7	8	After 62 exchanges
	7	8	7	8	After 63 exchanges
	7	8	7	8	After 64 exchanges
	7	8	7	8	After 65 exchanges
	7	8	7	8	After 66 exchanges
	7	8	7	8	After 67 exchanges
	7	8	7	8	After 68 exchanges
	7	8	7	8	After 69 exchanges
	7	8	7	8	After 70 exchanges
	7	8	7	8	After 71 exchanges
	7	8	7	8	After 72 exchanges
	7	8	7	8	After 73 exchanges
	7	8	7	8	After 74 exchanges
	7	8	7	8	After 75 exchanges
	7	8	7	8	After 76 exchanges
	7	8	7	8	After 77 exchanges
	7	8	7	8	After 78 exchanges
	7	8	7	8	After 79 exchanges
	7	8	7	8	After 80 exchanges
	7	8	7	8	After 81 exchanges
	7	8	7	8	After 82 exchanges
	7	8	7	8	After 83 exchanges
	7	8	7	8	After 84 exchanges
	7	8	7	8	After 85 exchanges
	7	8	7	8	After 86 exchanges
	7	8	7	8	After 87 exchanges
	7	8	7	8	After 88 exchanges
	7	8	7	8	After 89 exchanges
	7	8	7	8	After 90 exchanges
	7	8	7	8	After 91 exchanges
	7	8	7	8	After 92 exchanges
	7	8	7	8	After 93 exchanges
	7	8	7	8	After 94 exchanges
	7	8	7	8	After 95 exchanges
	7	8	7	8	After 96 exchanges
	7	8	7	8	After 97 exchanges
	7	8	7	8	After 98 exchanges
	7	8	7	8	After 99 exchanges
	7	8	7	8	After 100 exchanges

Bad news of no path to A is learned slowly

System not know only path is thru B, B thinks there is a path thru C

B knows it has no link to A so it chooses one of its neighbors that is 3 hops away

Link State Routing (1)

Link state routing is often used for intra-domain routing protocols such as IS-IS and OSPF. These routing protocols are used for routing within an AS.

Link state is an alternative to distance vector

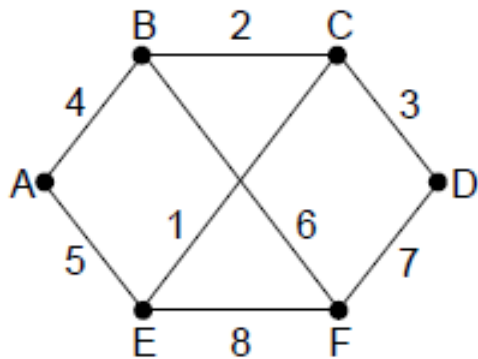
- More computation but simpler dynamics
- Widely used in the Internet (OSPF, ISIS)

Algorithm:

- Each node floods information about its neighbors in LSPs (Link State Packets); **all nodes learn the full network graph with identical view of network topology**
- Each node runs Dijkstra's algorithm to compute the path to take from itself to each destination

Link State Routing (2) – LSPs

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



Sender ID
Sequence #
Age

List neighbor
and cost

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

Network

LSP for each node

1. When a router is booted, it learns who its neighbors are by sending a Hello packet via each of its NICs. Adjacent router replies giving its names.
 - Routers on Broadcast LANs select a designated router to reply for the LAN – LANs are therefore treated as if it were a single node.
2. Each link has the same distance or cost metric. Delay can be determined by ECHO packets for systems that use delay as a metric.
3. Link State Packets (LSP – see above) are then constructed
4. Routers flood their LSP to all routers in the system. Age field decremented once per second and packet discarded once age hits zero

Link State Routing (3) – Reliable Flooding

Seq. number and age are used for reliable flooding

- New LSPs are acknowledged on the lines they are received and sent on all other lines
- Example shows the LSP database at router B

Source	Seq.	Age	Send flags			ACK flags			Data
			A	C	F	A	C	F	
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	← E info arrived twice: EAB and EFB
C	20	60	1	0	1	0	1	0	
D	21	59	1	0	0	0	1	1	

One row of the database is used for each recently arrived but not as yet fully processed LSP. 1 in Send flag indicates the link that info needs to be sent on and 1 in ACK indicates where receipt of info needs to be ack to.

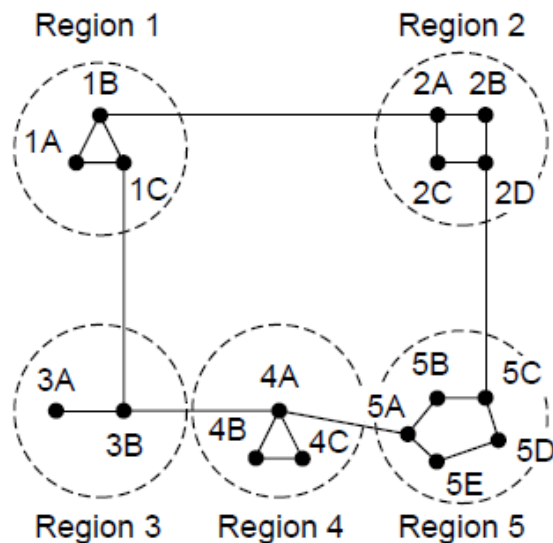
5. Next step is to have each node locally run the Dijkstra Algorithm on the received info. Therefore, possible that different directions of same path might have different costs.

Hierarchical Routing

Routing tables grow as networks grow which may cause issues. HR divides routers into regions for 2-level hierarchies; 3-level or more possible.

- Kamoun and Kleinrock – optimal number of levels for N router network is $\ln N$

Hierarchical routing reduces the work of route computation but may result in slightly longer paths than flat routing



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

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

Best choice to reach nodes in 5 except for 5C

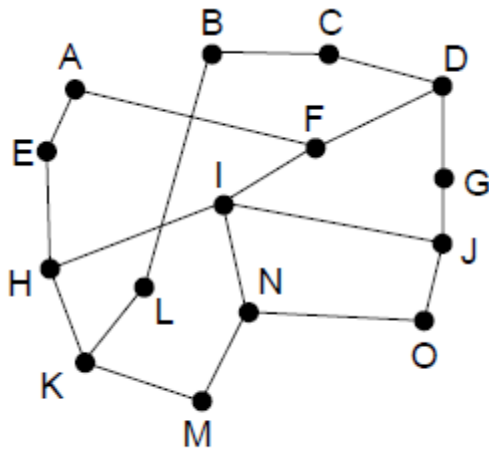
Tannenbaum's Use of "Broadcast" at the NW Layer

- Tannenbaum confusingly uses "broadcast" to describe how routers support Multicast (MC).
 - Broadcast is **NOT** a service available to the end user at the network layer within the Internet protocol suite.
- Routing Algorithms support Multicast via two alternative methods:
 1. Flood packets and then prune back to create a spanning tree
 2. Create a spanning tree from a common root location, known as Core Based Trees
- Routing forwarding for MC may use Reverse Path Forwarding (RPF)
- End users (including applications) have 3 service alternative choices at the Network Layer:
 1. Unicast
 2. Multicast
 3. Anycast.

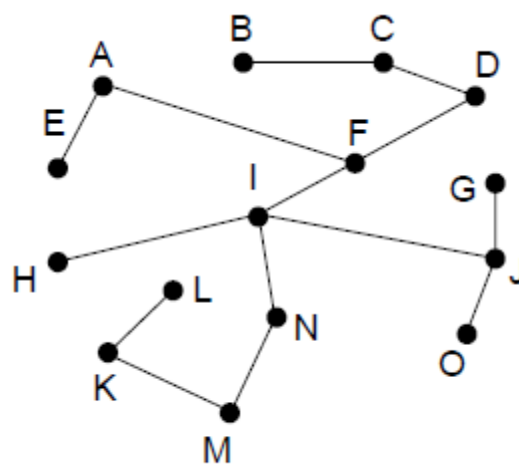
Broadcast Routing

Broadcast sends a packet to all nodes simultaneously

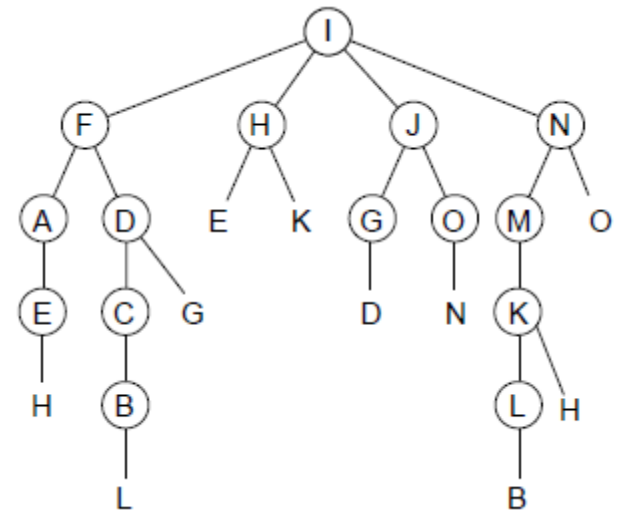
- **RPF** (Reverse Path Forwarding): send broadcast received on the link to the source out all remaining links
 - When a MC packet arrives at a router, the router checks the reverse path of packet to see if it is normally used to send MC packets. If router finds a matching routing entry for source IP addr, the RPF check passes and the packet is forwarded to all other interfaces of that MC group otherwise the packet is dropped. RPF can be used by distance vector routing systems
- Alternatively, can build and use sink trees (using link state) at all nodes



Network



Sink tree for I is
efficient broadcast

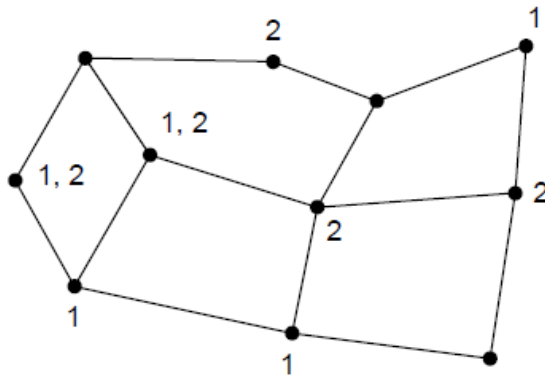


RPF from I is larger than
sink tree

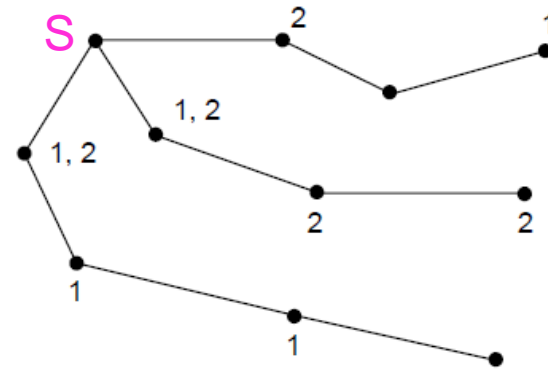
Multicast Routing (1)

Multicast sends to a subset of the nodes called a group

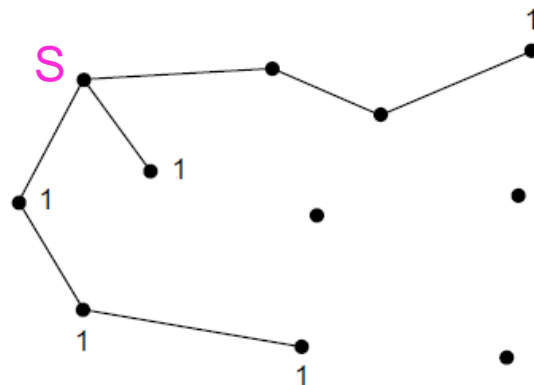
- Uses a different tree for each group and source



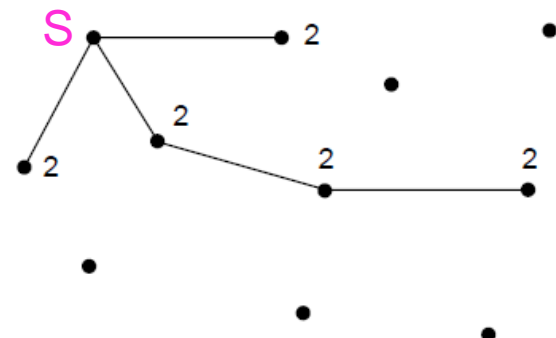
Network with groups 1 & 2



Spanning tree from source S



Multicast tree from S to group 1

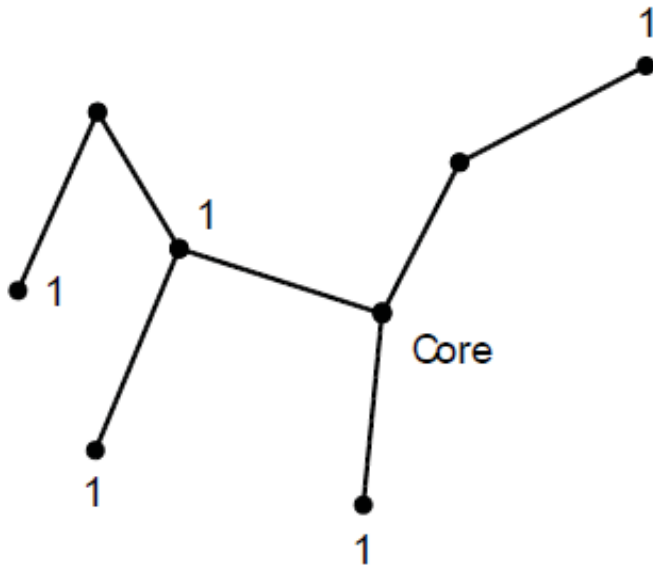


Multicast tree from S to group 2

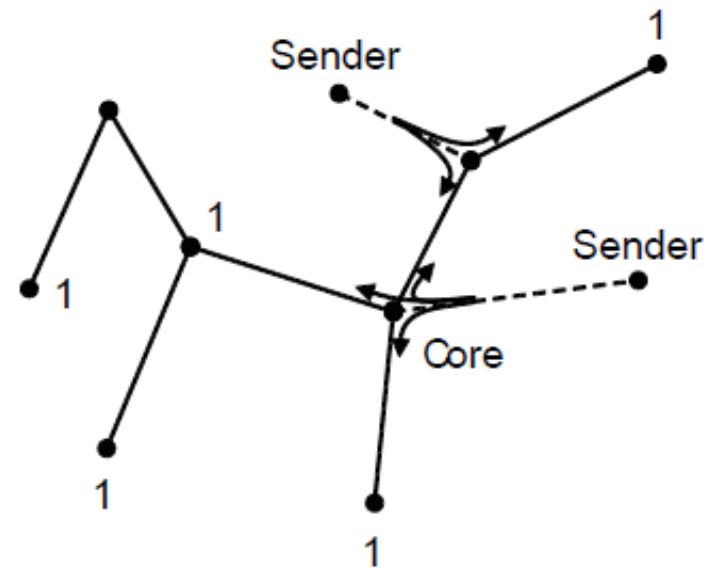
Multicast Routing (2) – Sparse Case

CBT (Core-Based Tree) uses a single tree to multicast

- Tree is the sink tree from core node to group members
- Multicast heads to the core until it reaches the CBT



Sink tree from core to group 1



Multicast is send to the core then down when it reaches the sink tree

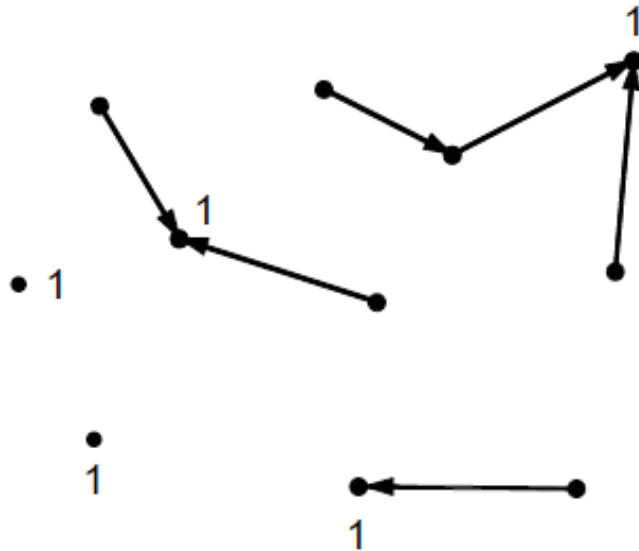
Used by PIM-SM

Anycast Routing

Anycast can be used by services – packet sent to the nearest member of a group (the group all use the same well-known IP address). E.g., DNS

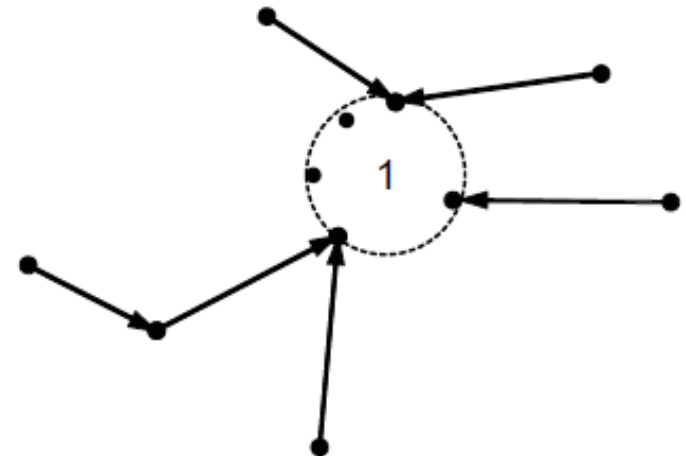
Anycast sends a packet to one (nearest) group member

- Falls out of regular routing with a node in many places
 - Distance vector will send packet to shortest path of that addr
 - Link state distinguishes between routers and host. It also can resolve anycast addr as long as Anycast nodes are in different parts of the network from each other (e.g., in different network areas, ASes).



Anycast routes to group 1

Example pretends that 1 is a valid IP address



Apparent topology of sink tree to "node" 1

Mobility

Routers, data links, and humans may have a different concept of what “mobility” is.

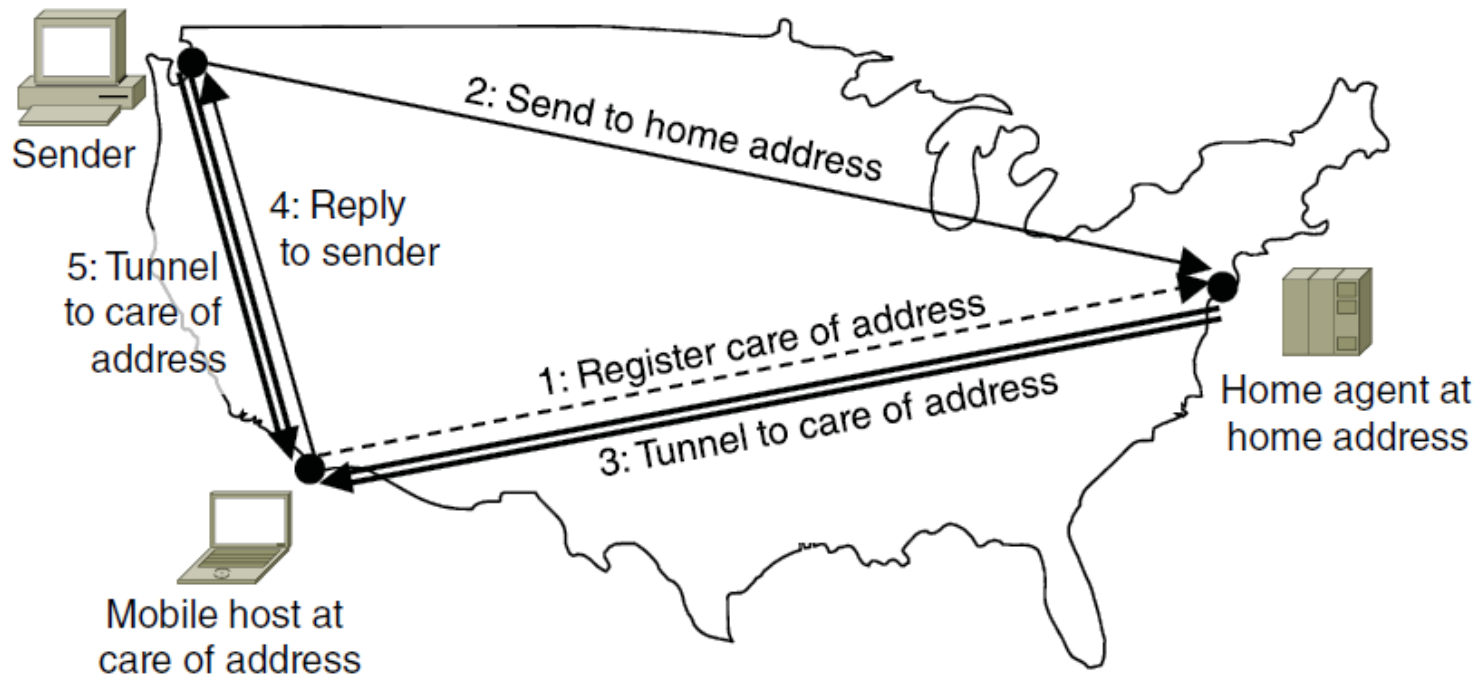
- Humans think “mobility” means changing locations.
- Networks only think “mobility” exists when the same IP address is used outside of its normal topological location. (Recall “Keys to Kingdom” lecture that IP addresses are locators, not identifiers.)
- Consider:
 - User moves within a satellite’s “beam” is not considered mobile from satellite’s perspective even if it is a move over a substantial geographical distance
 - User moves within the cell phone system is handled by cell protocols – not considered mobility from IP’s perspective
 - User moves between wi-fi (IEEE 802.11) hotspots. From IP’s perspective:
 - Not mobility if user gets a new IP address at that new hotspot
 - Is mobility if user doesn’t get a new IP address at that new hotspot

Routing for Mobile Hosts

Mobile IP – for many apps (VoIP, VPN) sudden changes of IP addr cause problems. The Mobile IP protocol is often used when users carry mobile devices across multiple LAN subnets (e.g., IP over DVB, WLAN, WIMAX, BWA)

Mobile hosts can be reached via a home agent

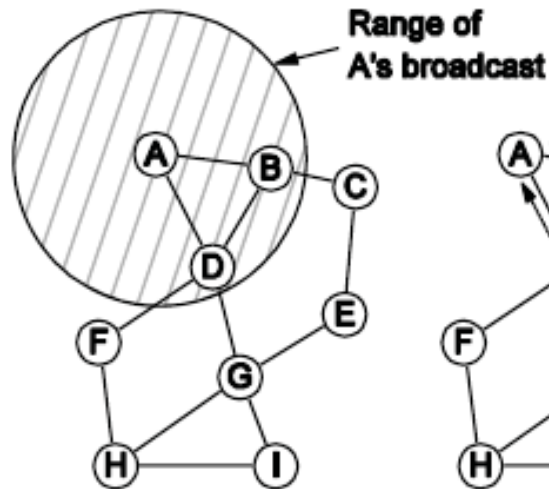
- Fixed home agent tunnels packets to reach the mobile host; reply can optimize path for subsequent packets
- No changes to routers or fixed hosts



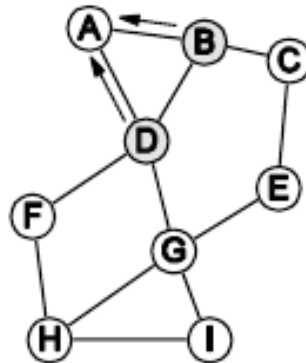
Routing in Ad Hoc Networks

The network topology changes as wireless nodes move

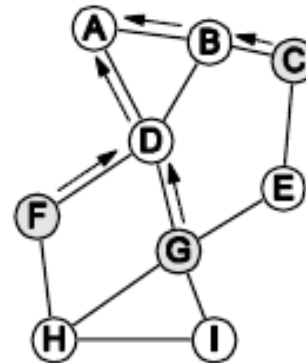
- Routes are often made on demand, e.g., AODV (below)



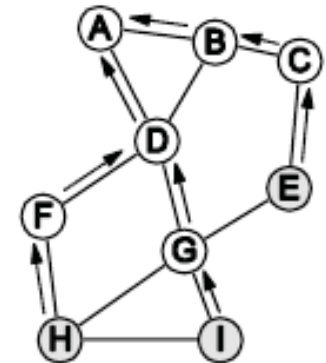
A's starts to
find route to I



A's broadcast
reaches B & D



B's and D's
broadcast
reach C, F & G



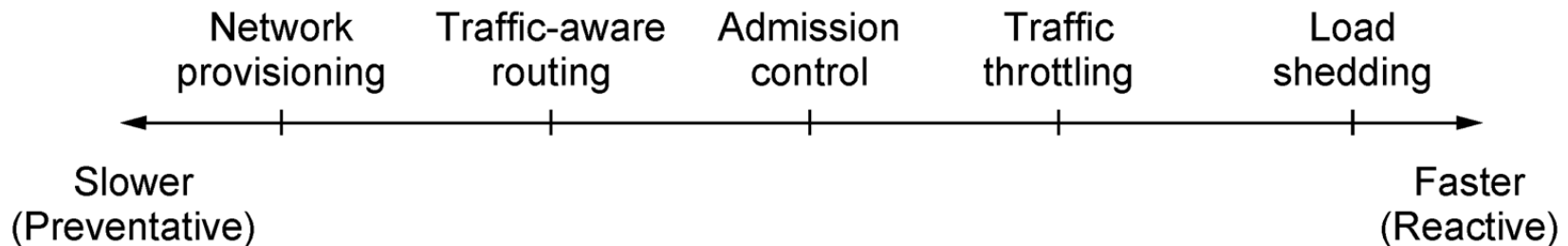
C's, F's and G's
broadcast
reach H & I

Congestion Control (1)

Congestion causes packet delay and loss that degrades performance.

Handling congestion is the responsibility of the Network and Transport layers working together

- We look at the Network portion here
- Traffic-aware routing » Section 5.3.2 in textbook
- Admission control » Section 5.3.3 in textbook
- Traffic throttling » Section 5.3.4 in textbook
- Load shedding » Section 5.3.5 in textbook

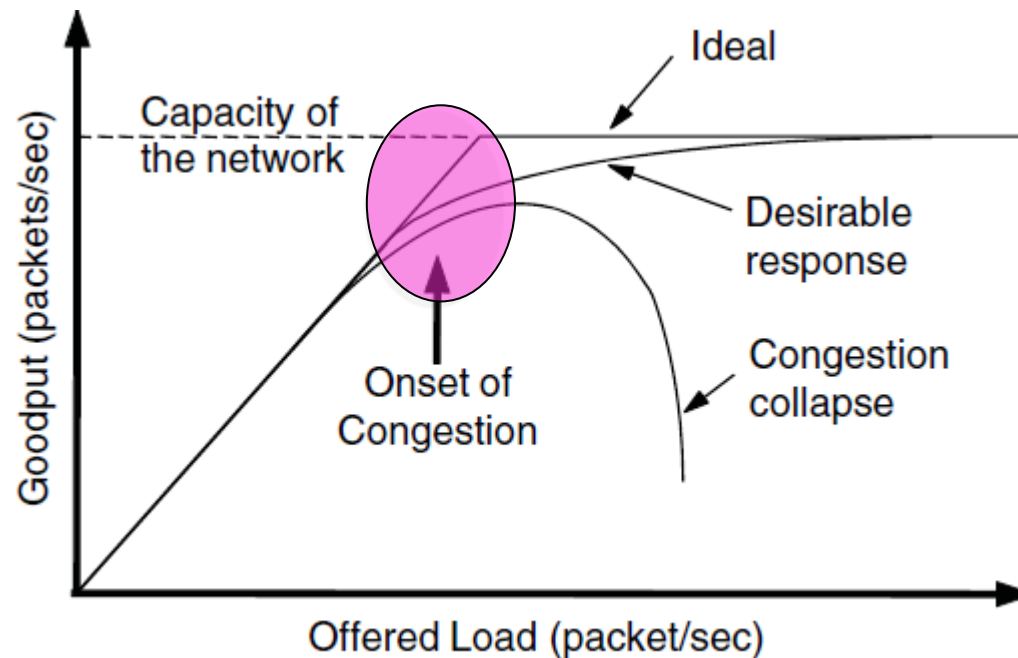


Timescales of approaches to congestion control.

Congestion Control (2)

Congestion results when too much traffic is offered; performance degrades due to loss/retransmissions

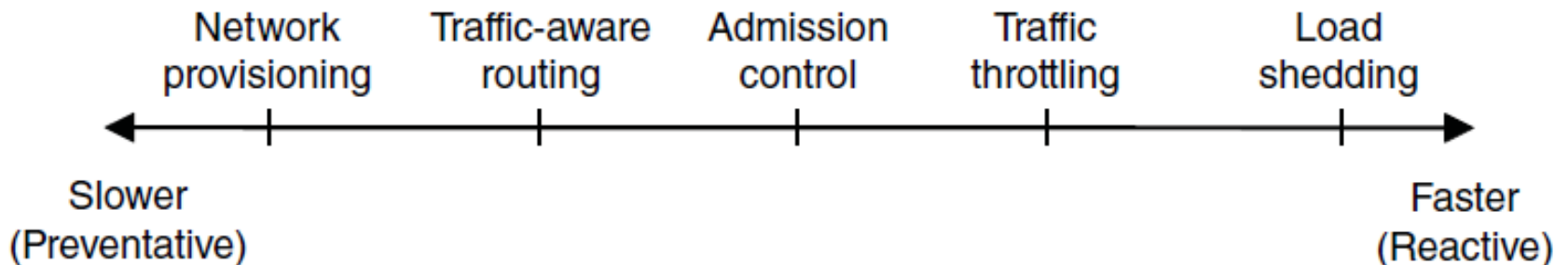
- Goodput (=useful packets) trails offered load



Congestion Control (3) – Approaches

Network must do its best with the offered load

- Different approaches at different timescales
- Nodes should also reduce offered load (Transport)



Provisioning – network deployment

Traffic Aware – e.g., splitting traffic across multiple paths

Admission Control – decrease network load (i.e., traffic entering the network)

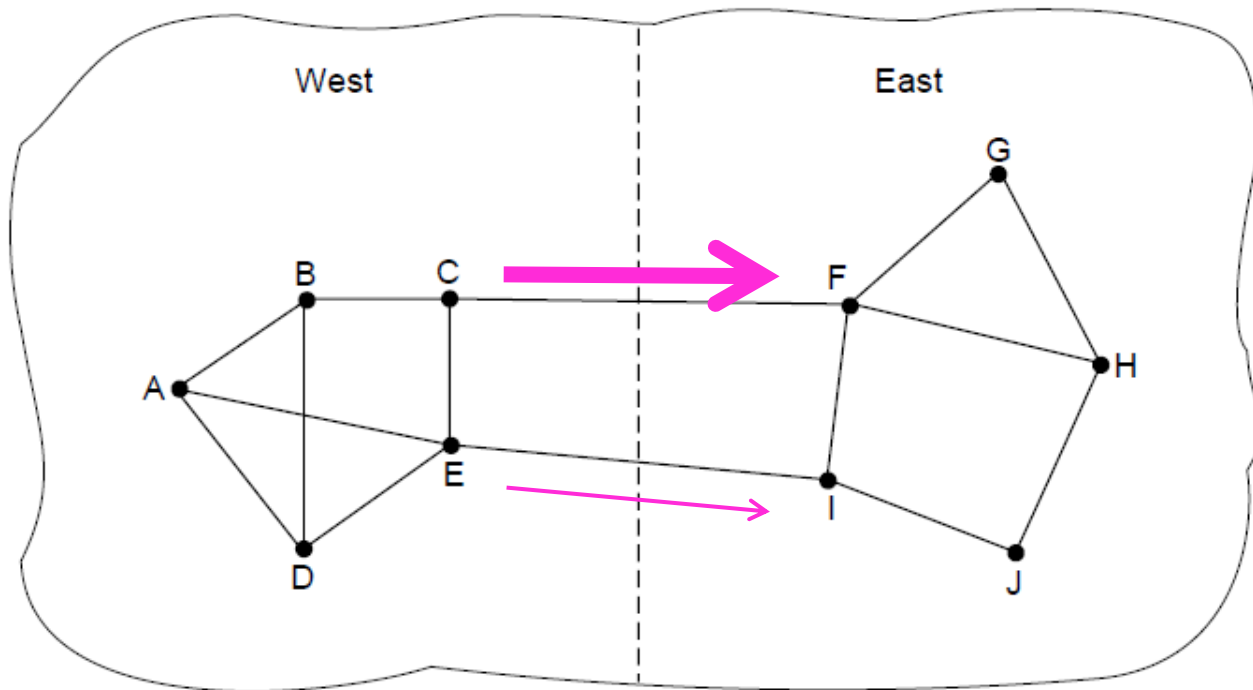
Traffic Throttling – e.g., explicit congestion notification (ECN)

Load Shedding – packet drop approaches and algorithms

Traffic-Aware Routing

Choose routes depending on traffic, not just topology

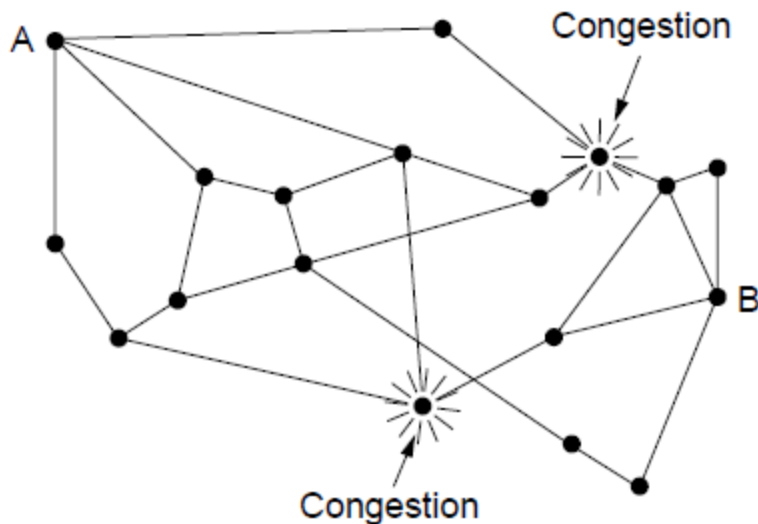
- E.g., use *EI* for West-to-East traffic if *CF* is loaded
- But take care to avoid oscillations



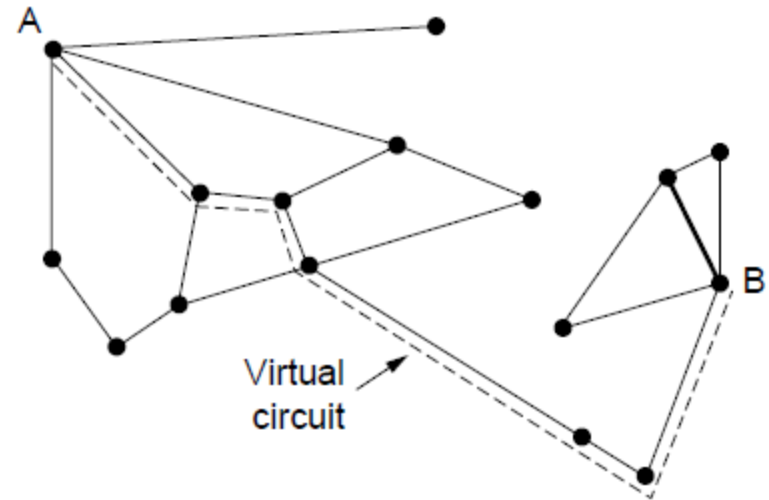
Admission Control

Admission control allows a new traffic load only if the network has sufficient capacity, e.g., with virtual circuits

- Can combine with looking for an uncongested route



Network with some congested nodes

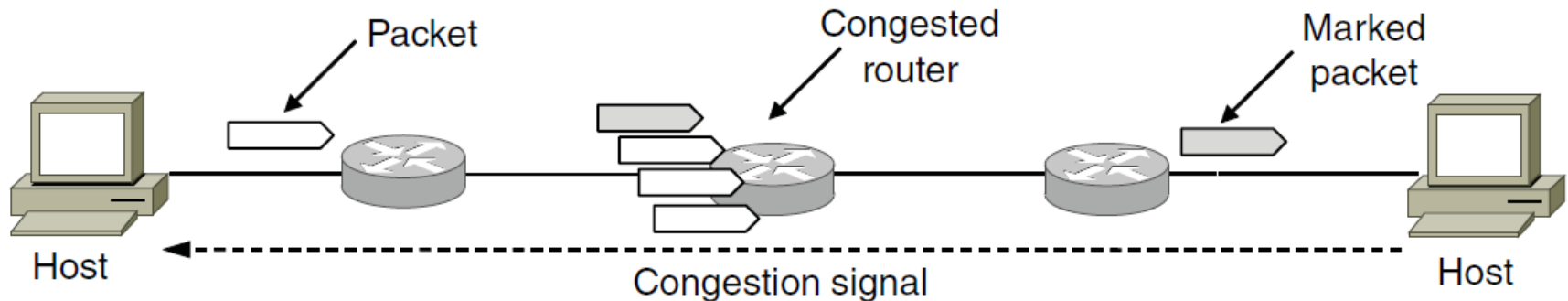


Uncongested portion and route AB around congestion

Traffic Throttling

Congested routers signal hosts to slow down traffic

- ECN (Explicit Congestion Notification) marks packets and receiver returns signal to sender

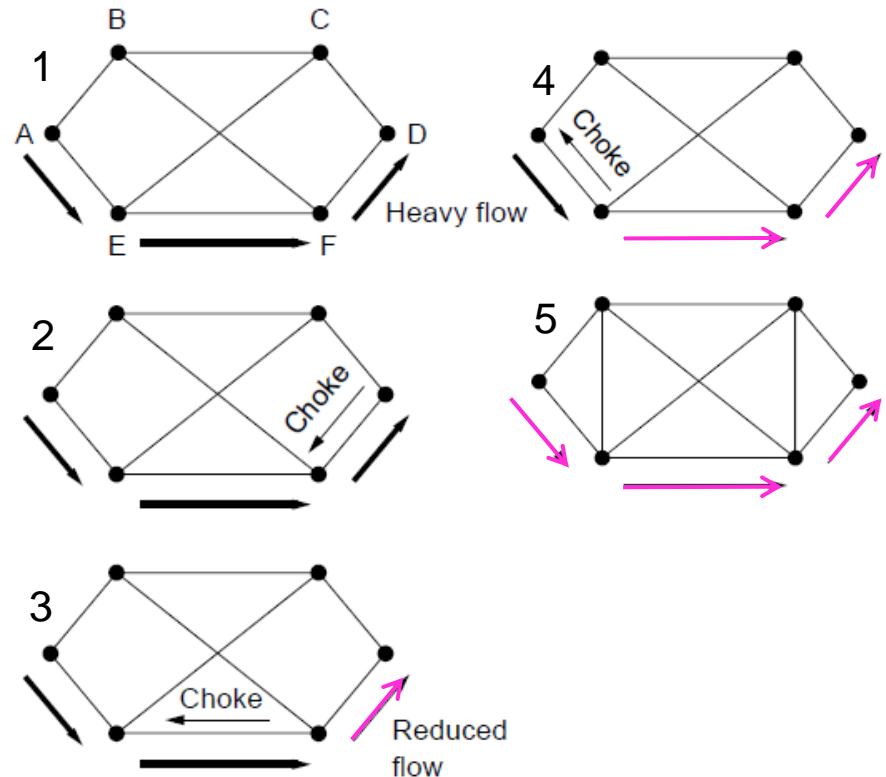


Load Shedding (1)

When all else fails, network will drop packets (shed load)

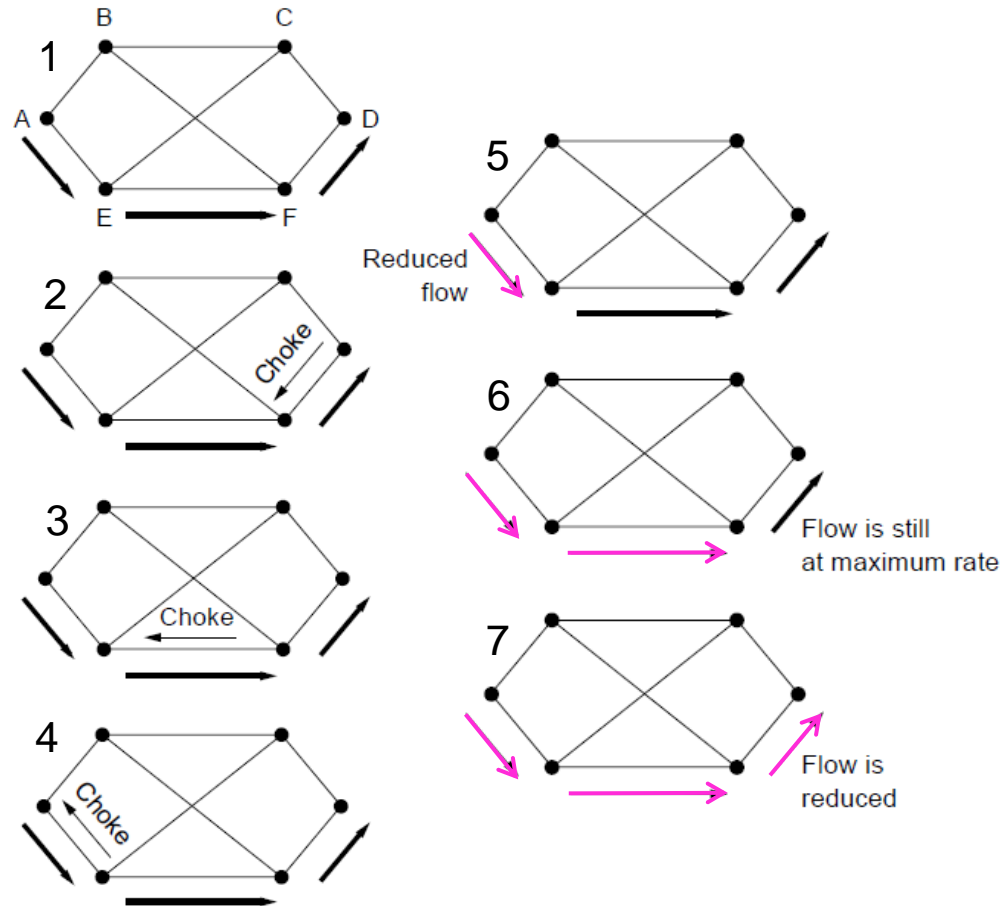
Can be done end-to-end or link-by-link

Link-by-link (right) produces rapid relief



Load Shedding (2)

End-to-end (right) takes longer to have an effect, but can better target the cause of congestion



Quality of Service

- Application requirements »
- Traffic shaping »
- Packet scheduling »
- Admission control »
- Integrated services »
- Differentiated services »

Application Requirements (1)

Different applications care about different properties

- We want all applications to get what they need

Application	Bandwidth	Delay	Jitter	Loss
Email	Low	Low	Low	Medium
File sharing	High	Low	Low	Medium
Web access	Medium	Medium	Low	Medium
Remote login	Low	Medium	Medium	Medium
Audio on demand	Low	Low	High	Low
Video on demand	High	Low	High	Low
Telephony	Low	High	High	Low
Videoconferencing	High	High	High	Low

“High” means a demanding requirement, e.g., low delay

Application Requirements (2)

Network provides service with different kinds of QoS (Quality of Service) to meet application requirements

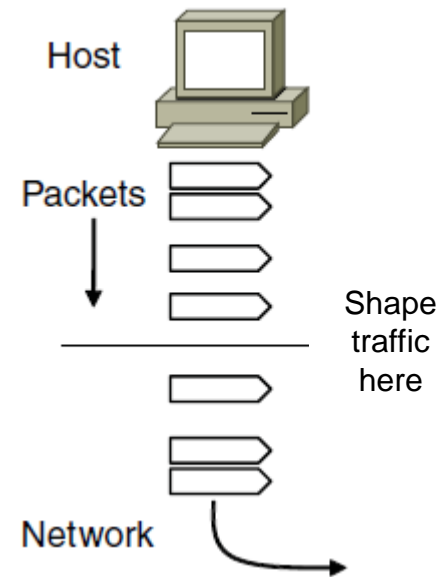
Network Service	Application
Constant bit rate	Telephony
Real-time variable bit rate	Videoconferencing
Non-real-time variable bit rate	Streaming a movie
Available bit rate	File transfer

Example of QoS categories from ATM networks

Traffic Shaping (1)

Traffic shaping regulates the average rate and burstiness of data entering the network

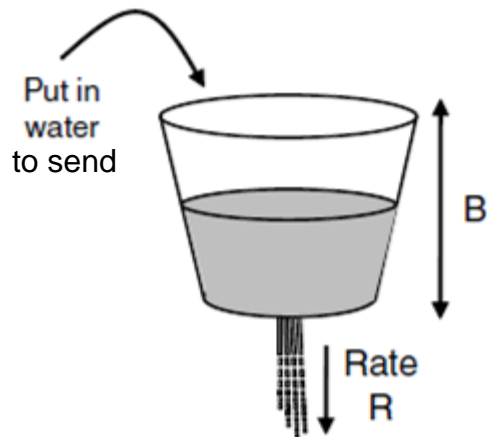
- Lets us make guarantees



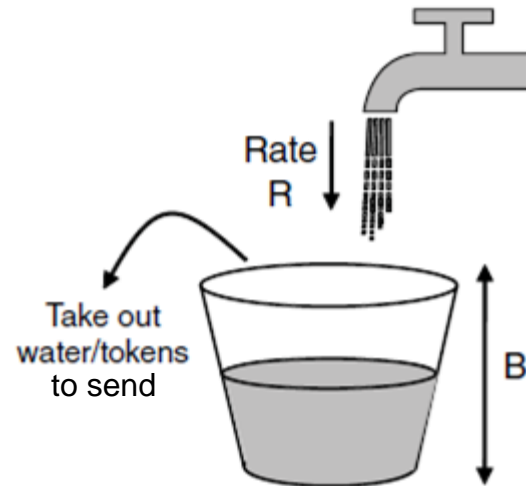
Traffic Shaping (2)

Token/Leaky bucket limits both the average rate (R) and short-term burst (B) of traffic

- For token, bucket size is B , water enters at rate R and is removed to send; opposite for leaky.



Leaky bucket
(need not full to send)



Token bucket
(need some water to send)

Traffic Shaping (3)

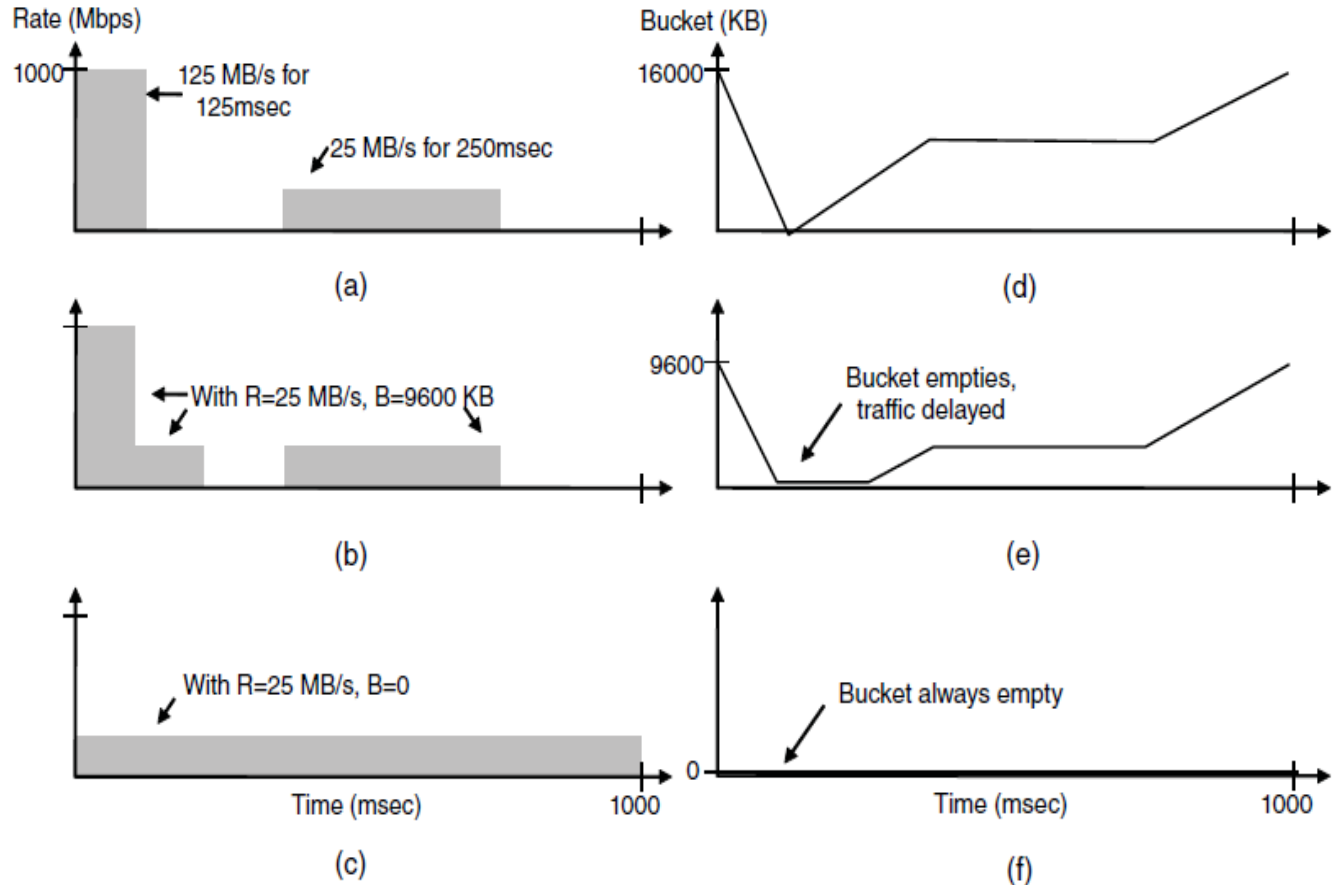
Host traffic
 $R=200$ Mbps
 $B=16000$ KB



Shaped by
 $R=200$ Mbps
 $B=9600$ KB



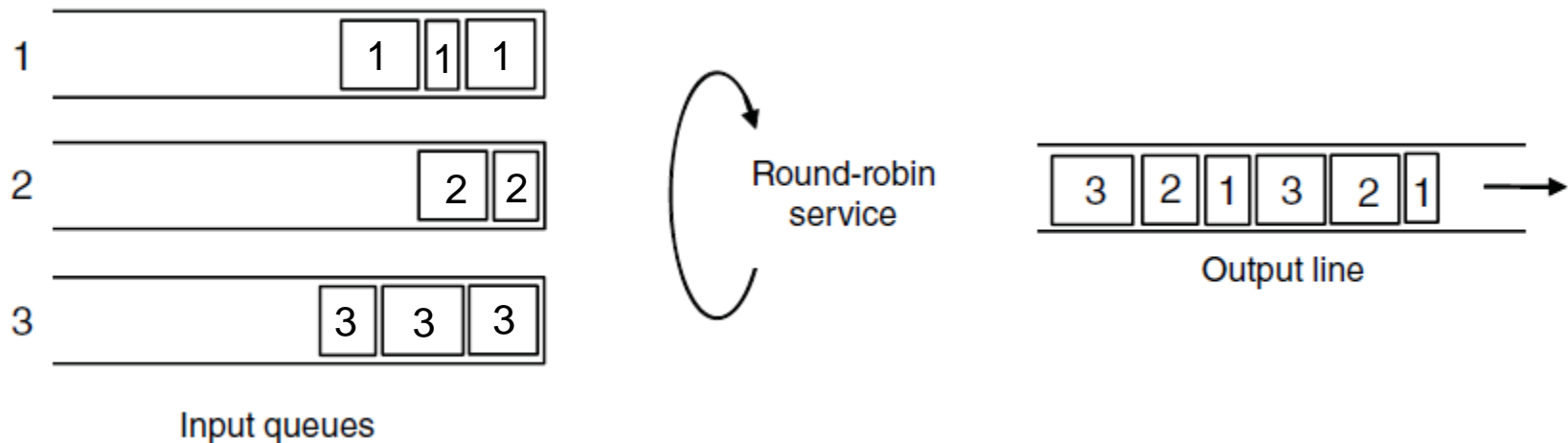
Shaped by
 $R=200$ Mbps
 $B=0$ KB



Smaller bucket size delays traffic and reduces burstiness

Packet Scheduling (1)

Packet scheduling divides router/link resources among traffic flows with alternatives to FIFO (First In First Out)

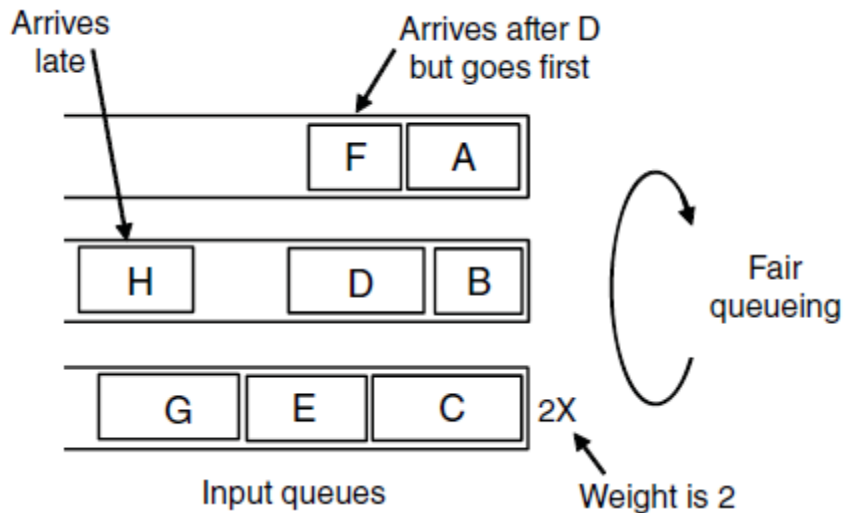


Example of round-robin queuing

Packet Scheduling (2)

Fair Queueing approximates bit-level fairness with different packet sizes; weights change target levels

- Result is WFQ (Weighted Fair Queueing)



Packets may be sent out of arrival order

Packet	Arrival time	Length	Finish time	Output order
A	0	8	8	1
B	5	6	11	3
C	5	10	10	2
D	8	9	20	7
E	8	8	14	4
F	10	6	16	5
G	11	10	19	6
H	20	8	28	8

$$F_i = \max(A_i, F_{i-1}) + L_i/W$$

Finish virtual times determine transmission order

Admission Control (1)

Admission control takes a traffic flow specification and decides whether the network can carry it

- Sets up packet scheduling to meet QoS

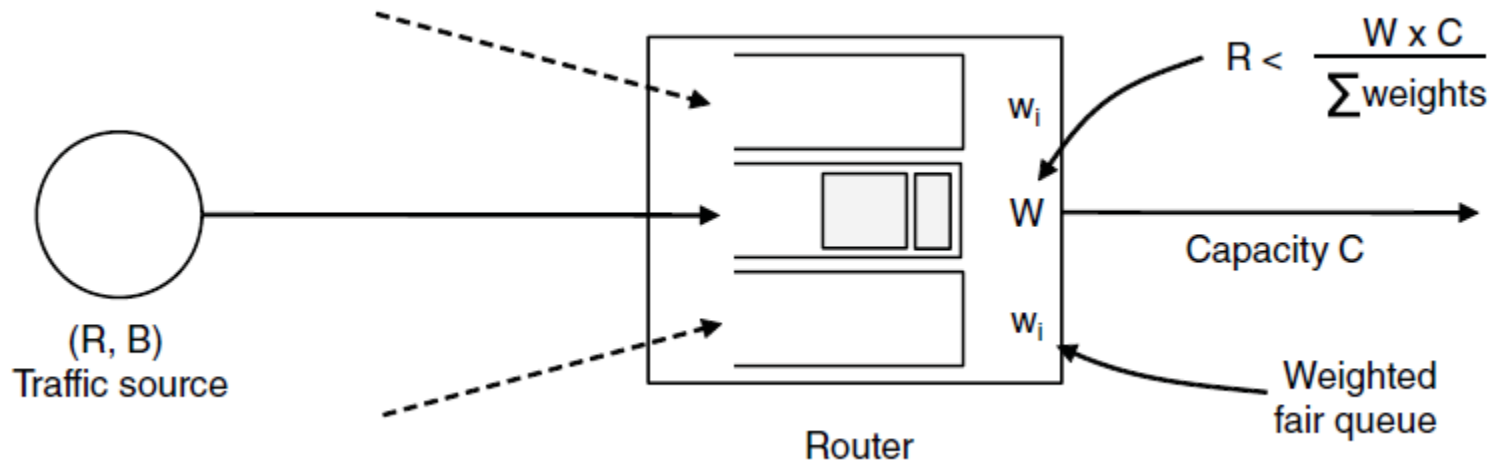
Parameter	Unit
Token bucket rate	Bytes/sec
Token bucket size	Bytes
Peak data rate	Bytes/sec
Minimum packet size	Bytes
Maximum packet size	Bytes

Example flow specification

Admission Control (2)

Construction to guarantee bandwidth B and delay D:

- Shape traffic source to a (R, B) token bucket
- Run WFQ with weight W / all weights $> R/\text{capacity}$
- Holds for all traffic patterns, all topologies



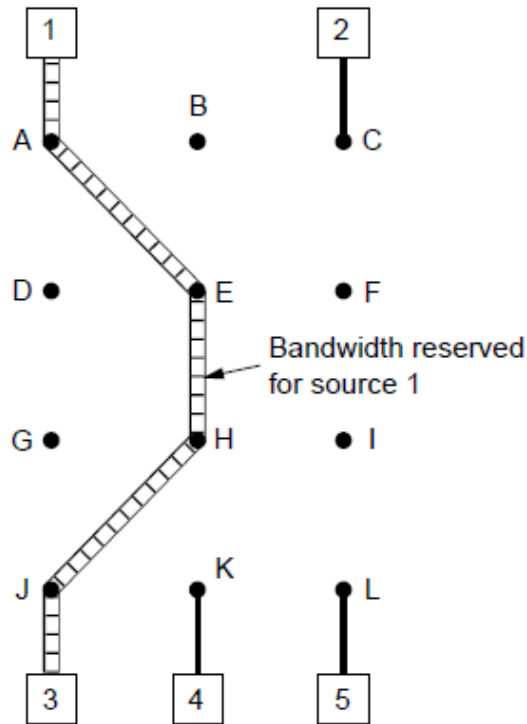
Integrated Services (1)

Design with QoS for each flow; handles multicast traffic.

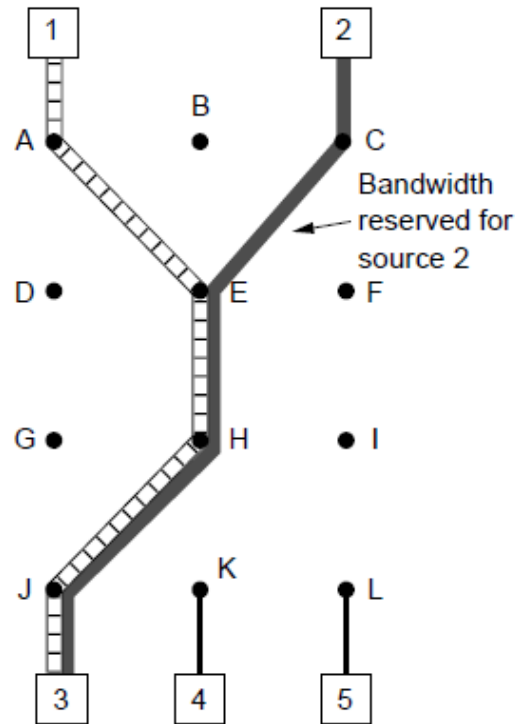
Admission with RSVP (Resource reSerVation Protocol):

- Receiver sends a request back to the sender
- Each router along the way reserves resources
- Routers merge multiple requests for same flow
- Entire path is set up, or reservation not made

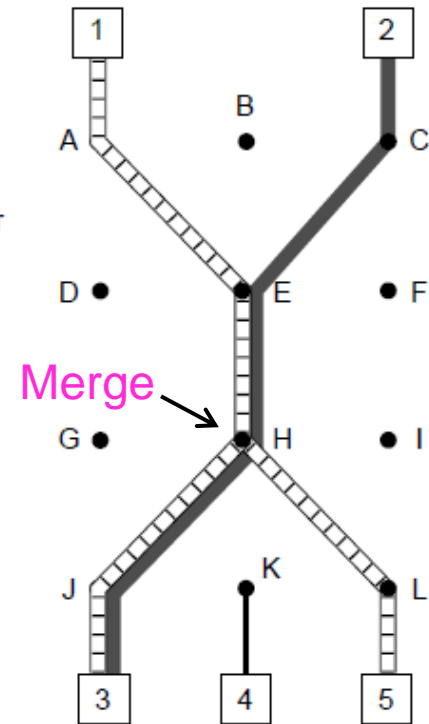
Integrated Services (2)



R3 reserves flow from S1



R3 reserves flow from S2

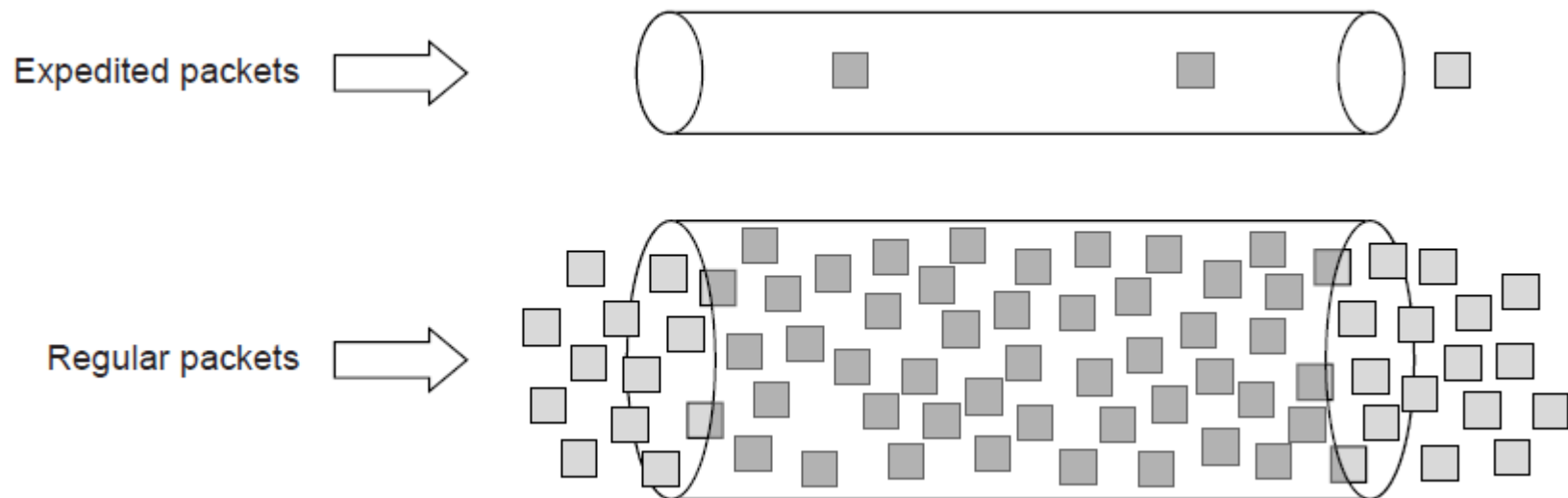


R5 reserves flow from S1; merged with R3 at H

Differentiated Services (1)

Design with classes of QoS; customers buy what they want

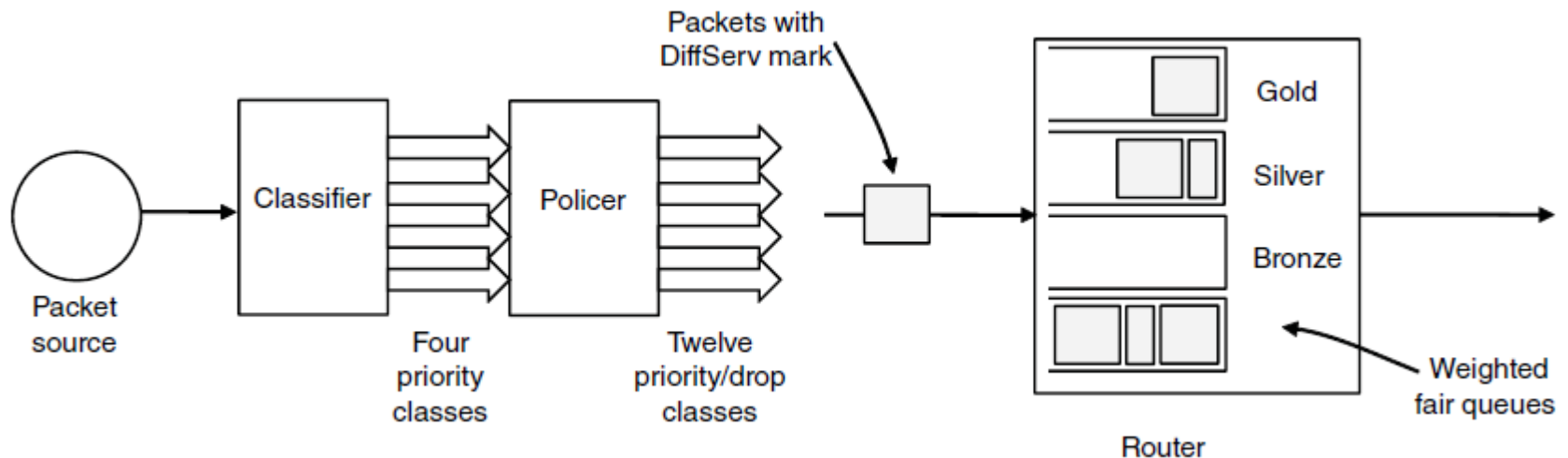
- Expedited class is sent in preference to regular class
- Less expedited traffic but better quality for applications



Differentiated Services (2)

Implementation of DiffServ:

- Customers mark desired class on packet
- ISP shapes traffic to ensure markings are paid for
- Routers use WFQ to give different service levels



Internetworking

Internetworking joins multiple, different networks into a single larger network

- How networks differ »
- How networks can be connected »
- Tunneling »
- Internetwork routing »
- Packet fragmentation »

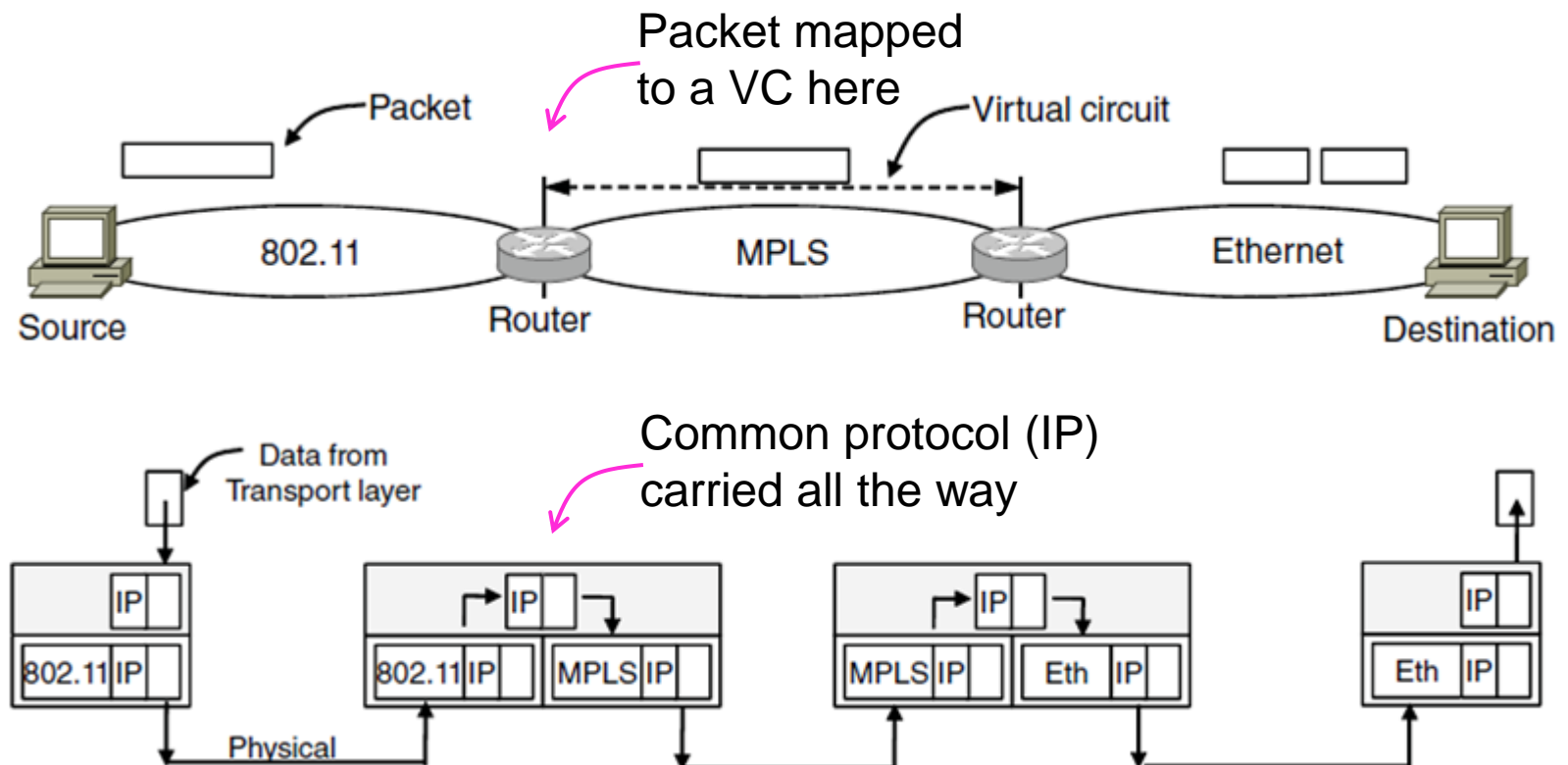
How Networks Differ

Differences can be large; complicates internetworking

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

How Networks Can Be Connected

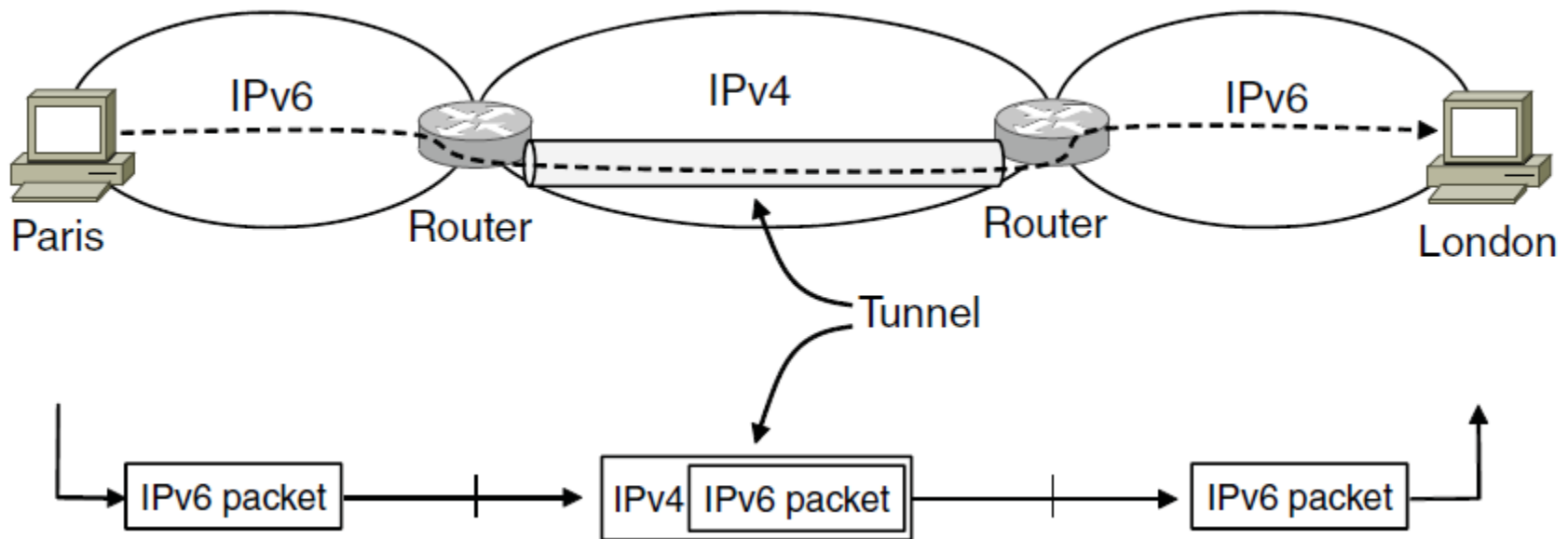
Internetworking based on a common network layer – IP



Tunneling (1)

Connects two networks through a middle one

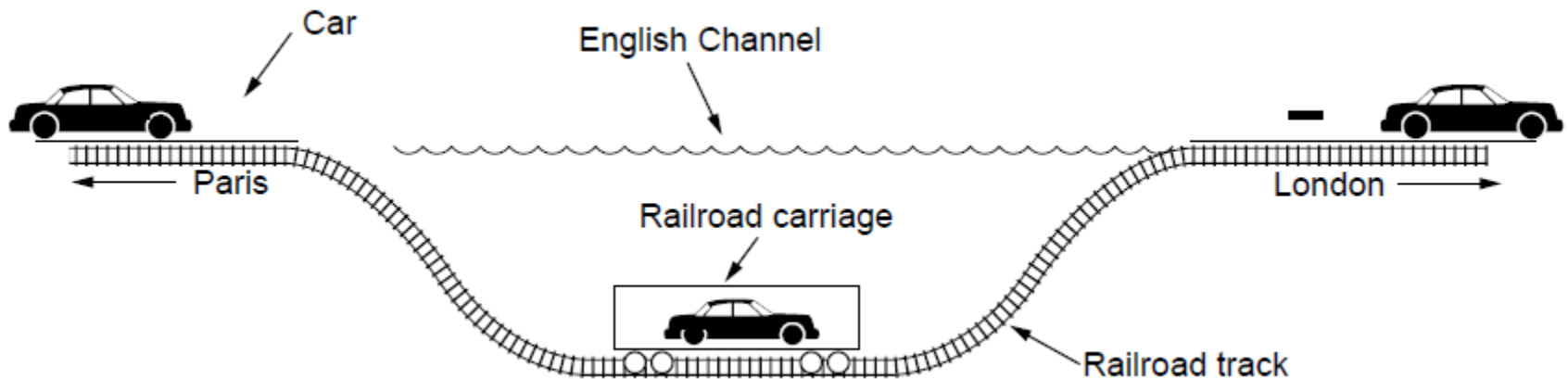
- Packets are encapsulated over the middle



Tunneling (2)

Tunneling analogy:

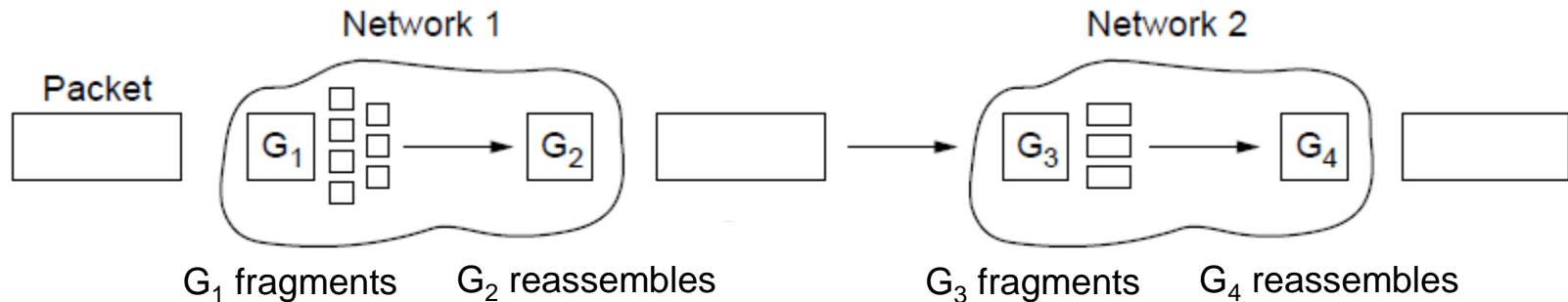
- tunnel is a link; packet can only enter/exit at ends



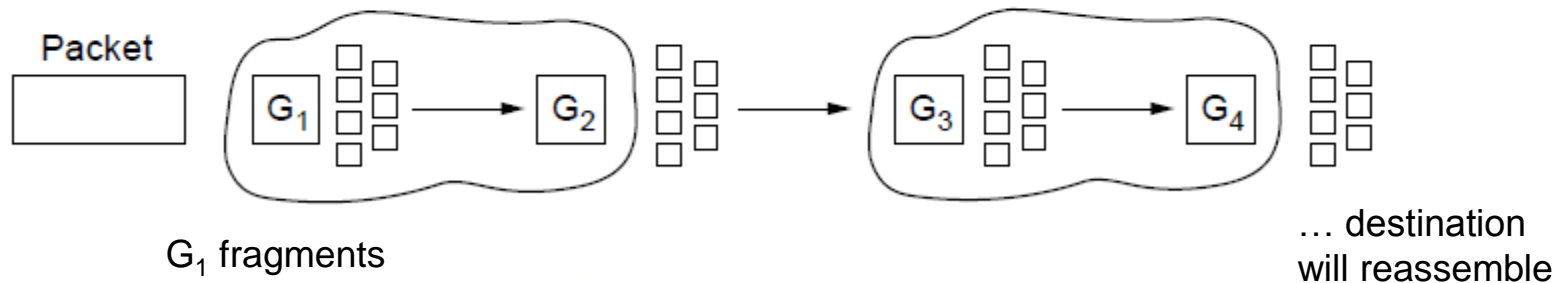
Packet Fragmentation (1)

Networks have different packet size limits for many reasons

- Large packets sent with fragmentation & reassembly



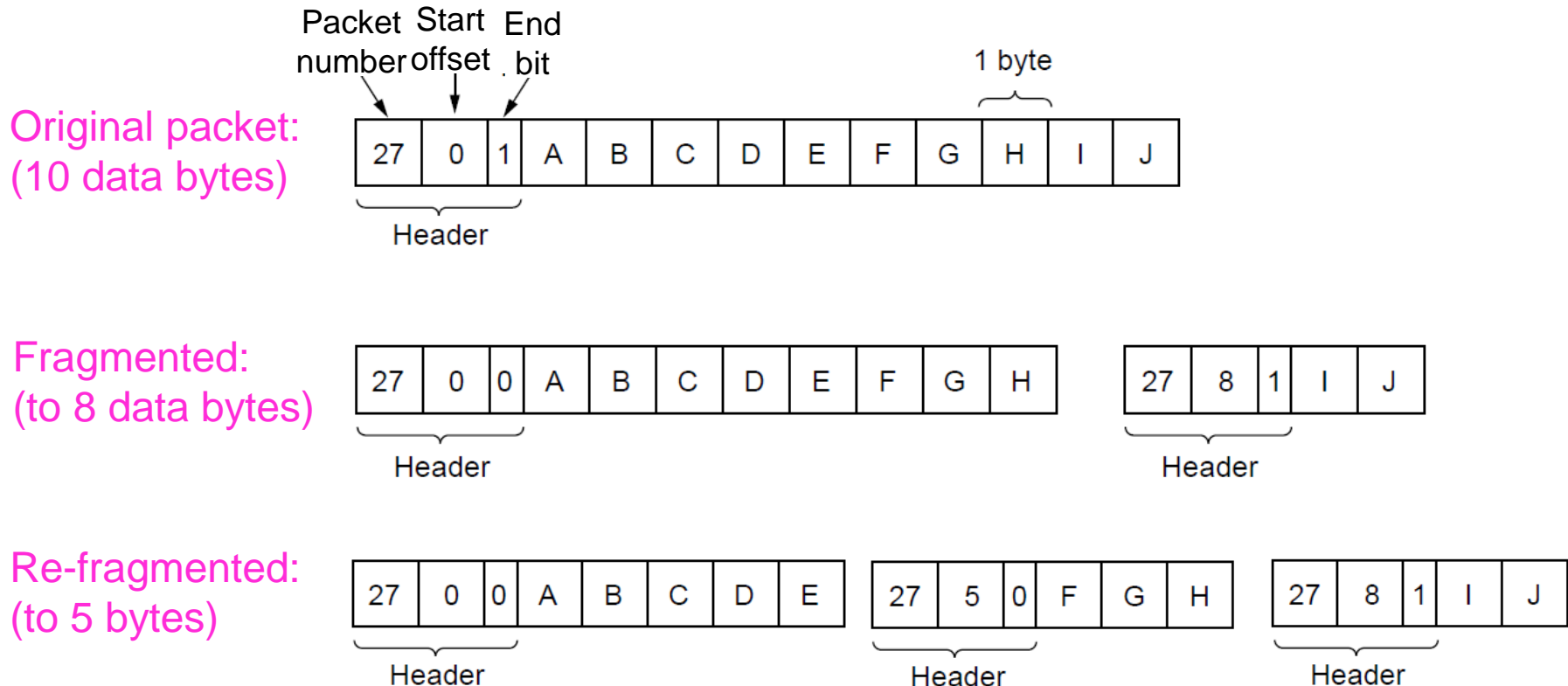
Transparent – packets fragmented / reassembled in each network



Non-transparent – fragments are reassembled at destination

Packet Fragmentation (2)

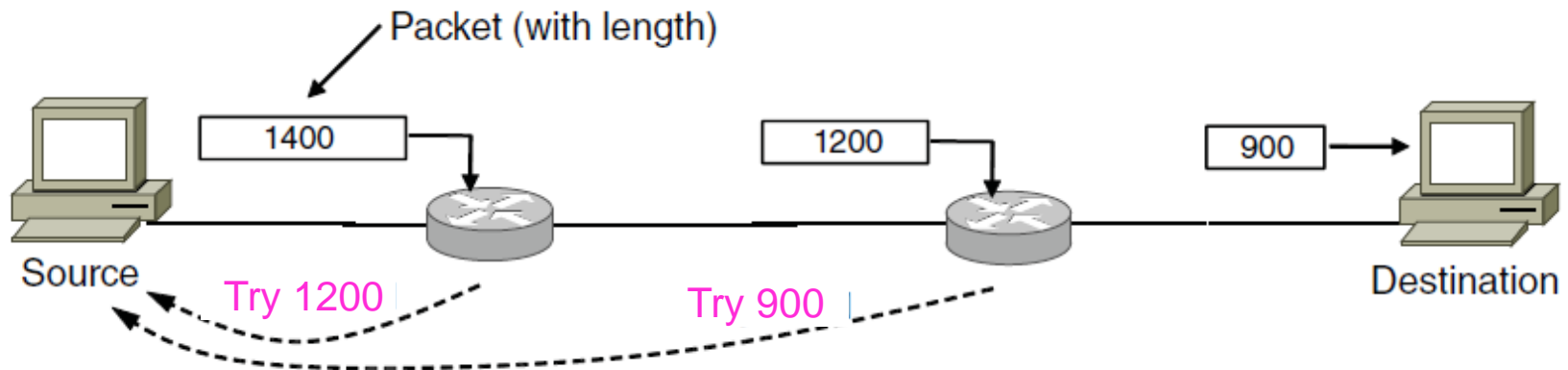
Example of IP-style fragmentation:



Packet Fragmentation (3)

Path MTU Discovery avoids network fragmentation

- Routers return MTU (Max. Transmission Unit) to source and discard large packets



Network Layer in the Internet (1)

- IP Version 4 »
- IP Addresses »
- IP Version 6 »
- Internet Control Protocols »
- Label Switching and MPLS »
- OSPF—An Interior Gateway Routing Protocol »
- BGP—The Exterior Gateway Routing Protocol »
- Internet Multicasting »
- Mobile IP »

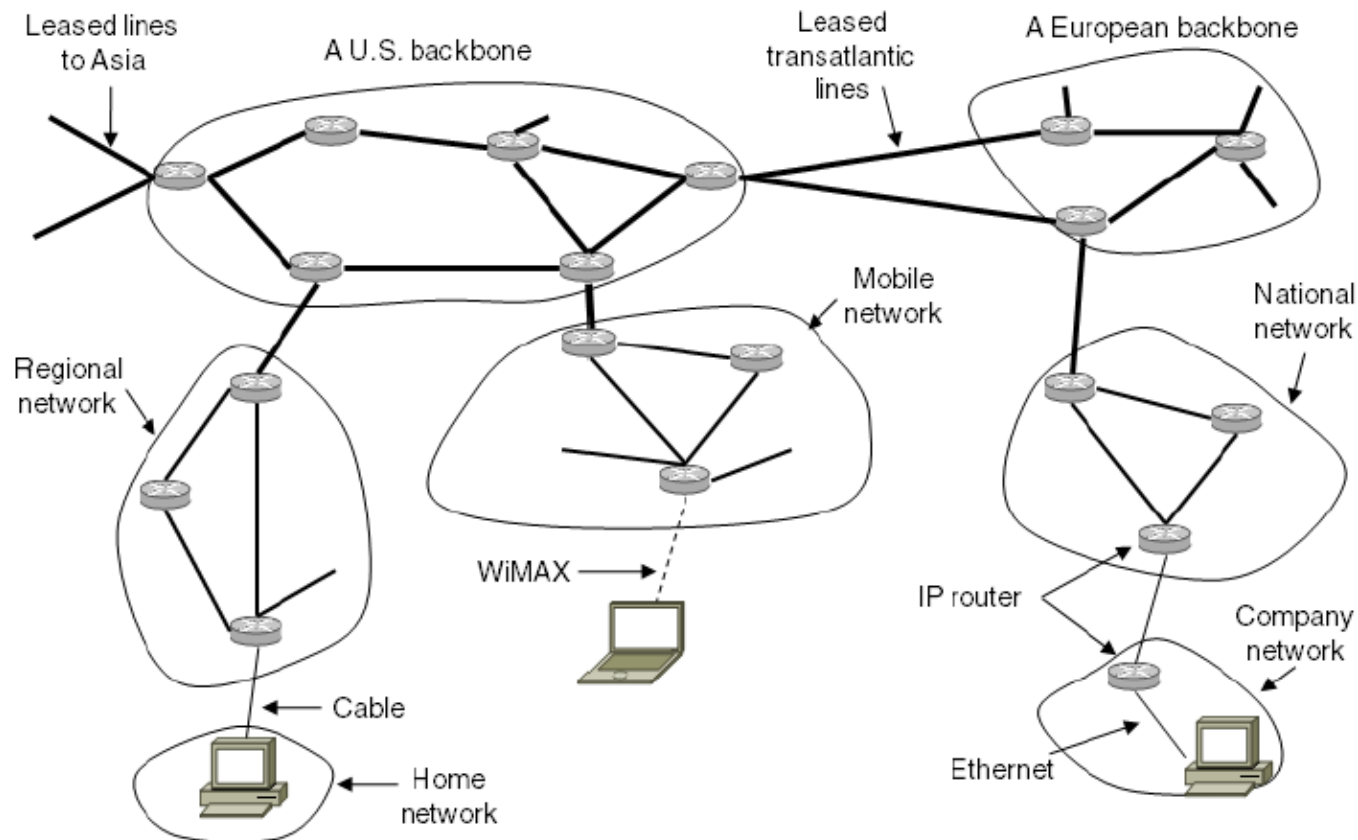
Network Layer in the Internet (2)

IP has been shaped by guiding principles:

- Make sure it works
- Keep it simple
- Make clear choices
- Exploit modularity
- Expect heterogeneity
- Avoid static options and parameters
- Look for good design (not perfect)
- Strict sending, tolerant receiving
- Think about scalability
- Consider performance and cost

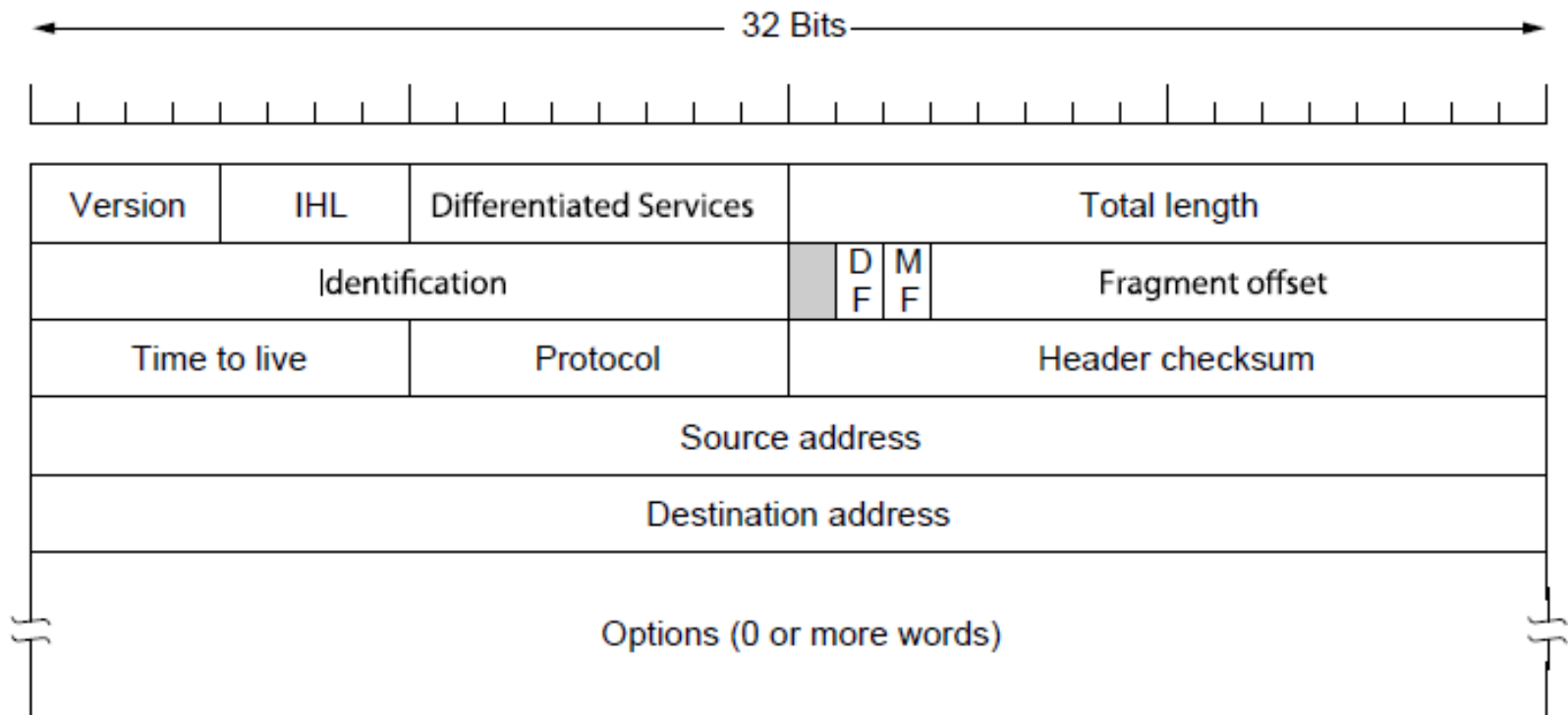
Network Layer in the Internet (3)

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



IP Version 4 Protocol (1)

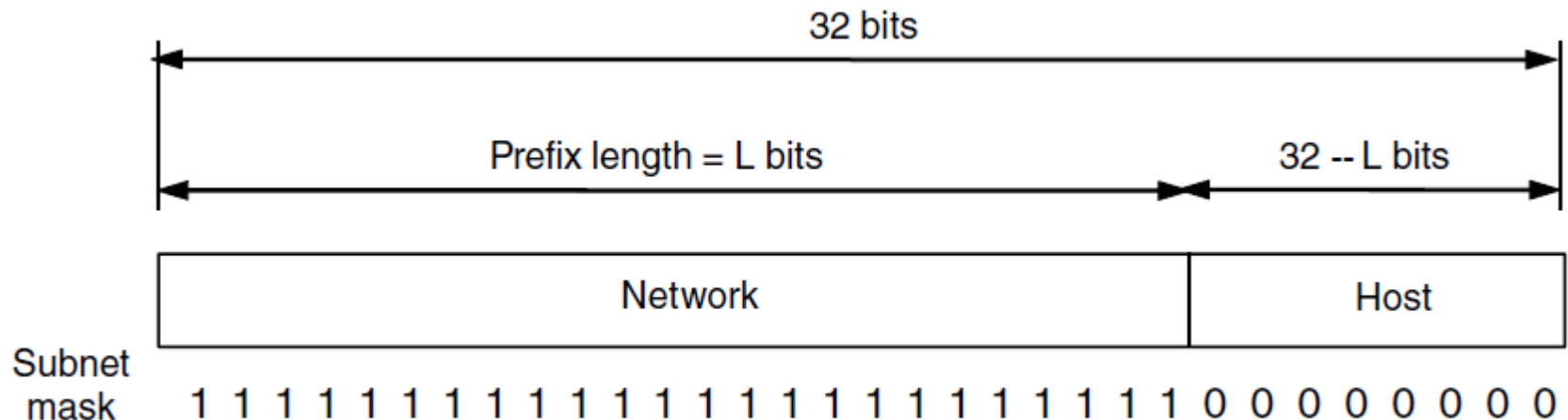
IPv4 (Internet Protocol) header is carried on all packets and has fields for the key parts of the protocol:



IP Addresses (1) – Prefixes

Addresses are allocated in blocks called prefixes

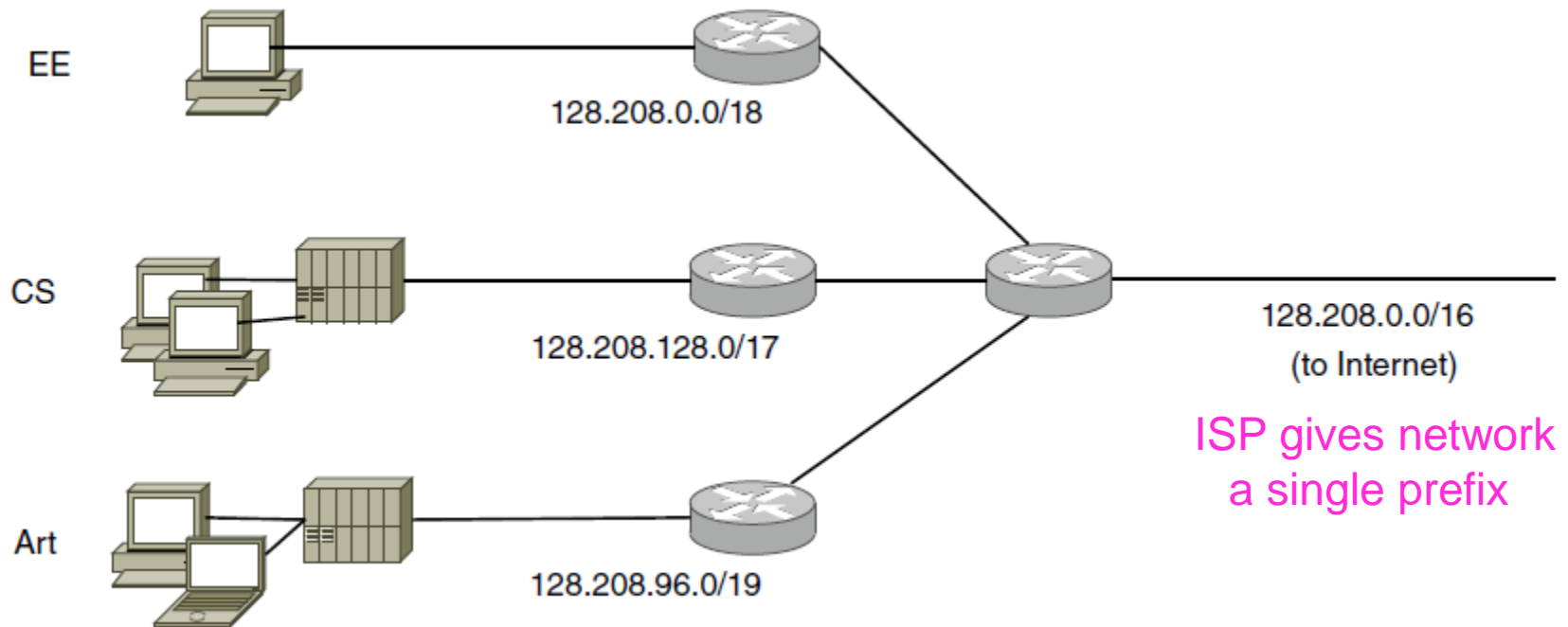
- Prefix is determined by the network portion
- Has 2^L addresses aligned on 2^L boundary
- Written address/length, e.g., 18.0.31.0/24



IP Addresses (2) – Subnets

Subnetting splits up IP prefix to help with management

- Looks like a single prefix outside the network

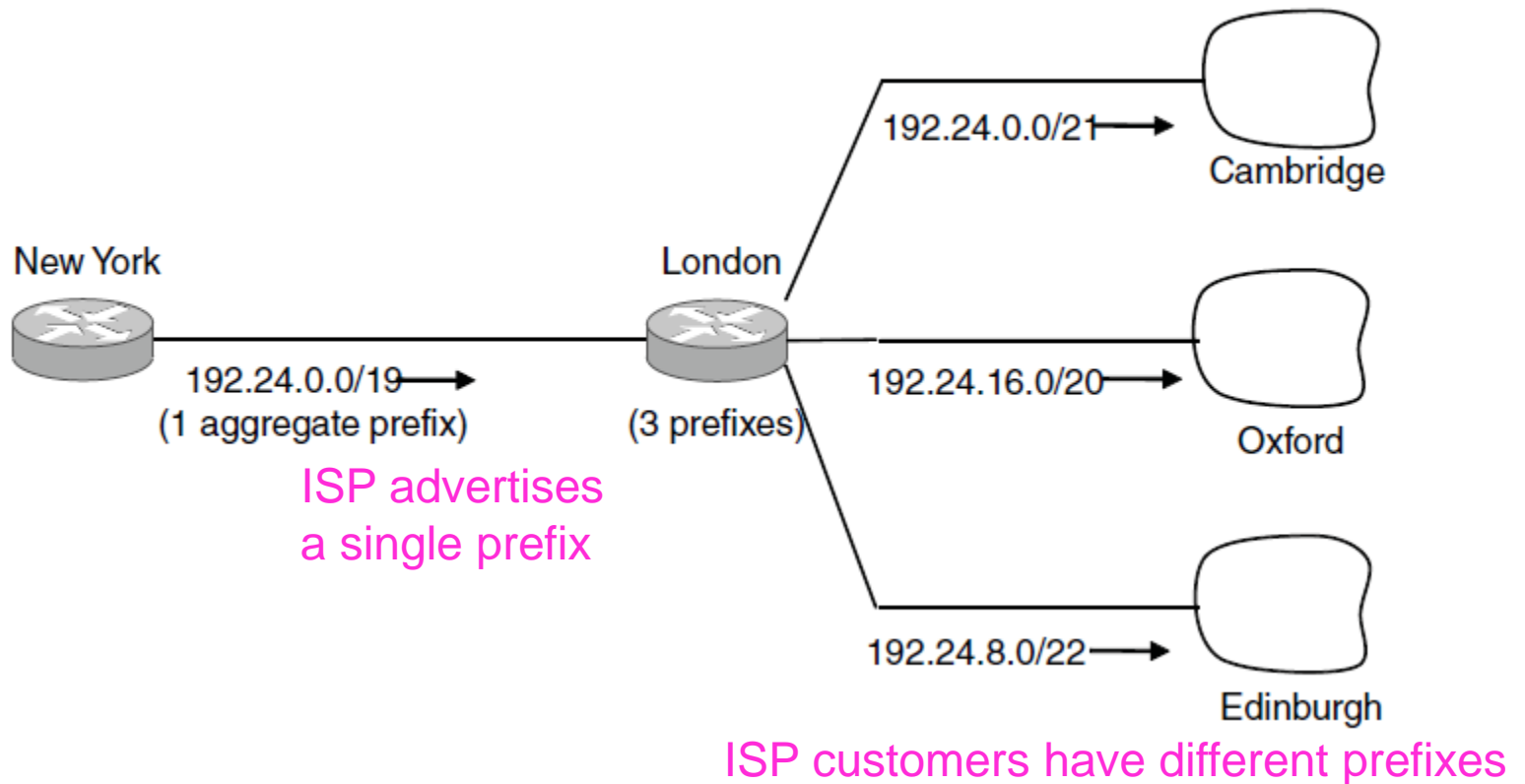


ISP gives network
a single prefix

Network divides it into subnets internally

IP Addresses (3) – Aggregation

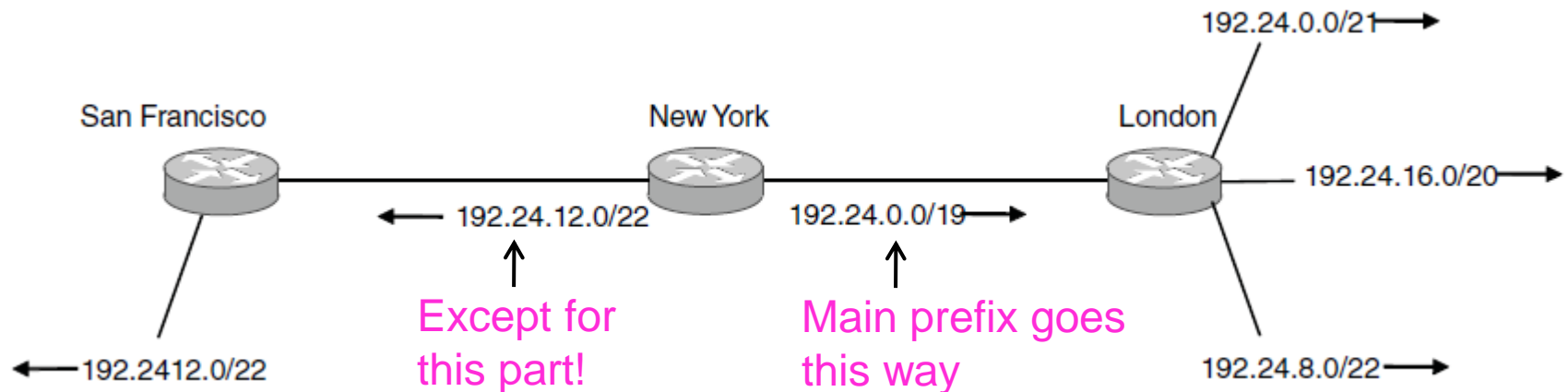
Aggregation joins multiple IP prefixes into a single larger prefix to reduce routing table size



IP Addresses (4) – Longest Matching Prefix

Packets are forwarded to the entry with the longest matching prefix or smallest address block

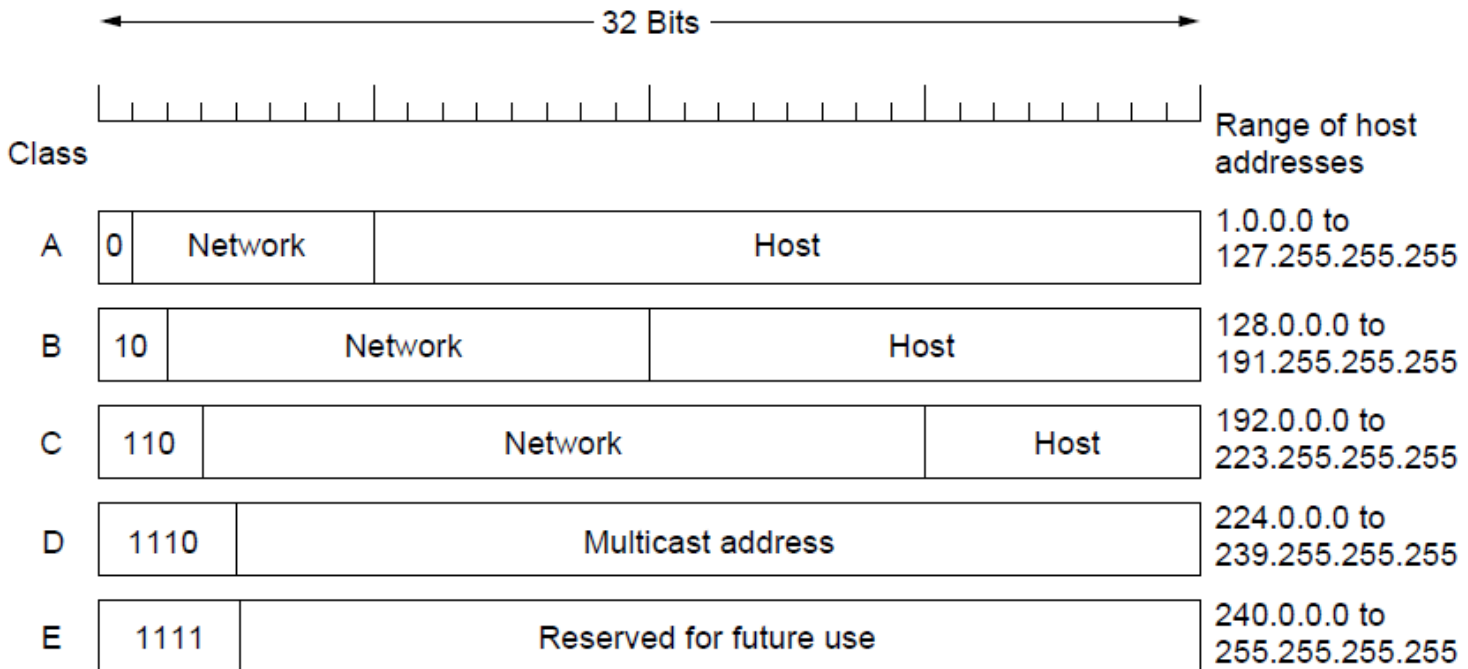
- Complicates forwarding but adds flexibility



IP Addresses (5) – Classful Addressing

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

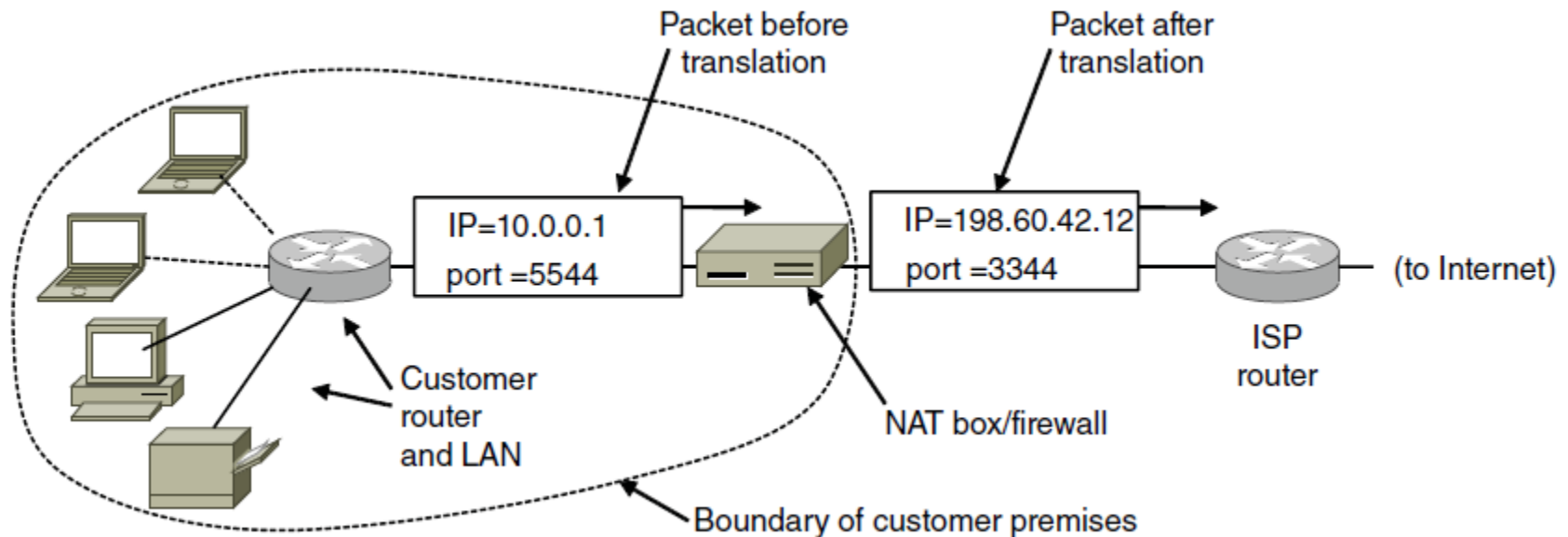
- Carries size as part of address, but lacks flexibility
- Called classful (vs. classless) addressing



IP Addresses (6) – 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 Version 6 (1)

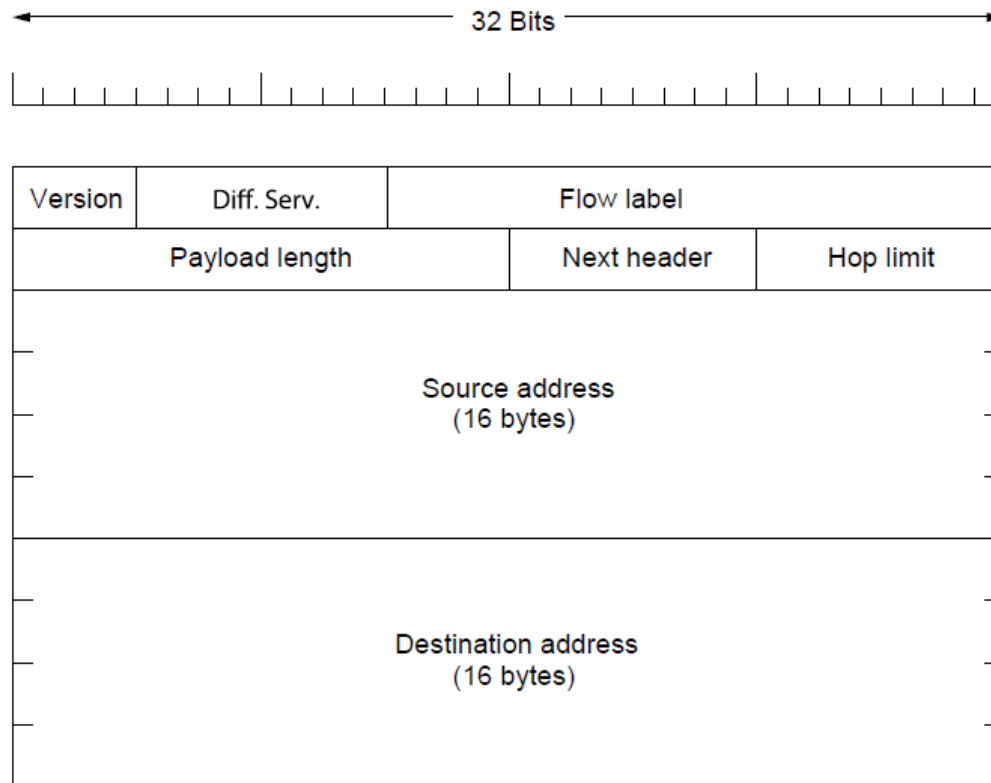
Major upgrade in the 1990s due to impending address exhaustion, with various other goals:

- Support billions of hosts
- Reduce routing table size
- Simplify protocol
- Better security
- Attention to type of service
- Aid multicasting
- Roaming host without changing address
- Allow future protocol evolution
- Permit coexistence of old, new protocols, ...

Deployment has been slow & painful, but may pick up pace now that addresses are all but exhausted

IP Version 6 (2)

IPv6 protocol header has much longer addresses (128 vs. 32 bits) and is simpler (by using extension headers)



IP Version 6 (3)

IPv6 extension headers handles other functionality

Extension header	Description
Hop-by-hop options	Miscellaneous information for routers
Destination options	Additional information for the destination
Routing	Loose list of routers to visit
Fragmentation	Management of datagram fragments
Authentication	Verification of the sender's identity
Encrypted security payload	Information about the encrypted contents

Internet Control Protocols (1)

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

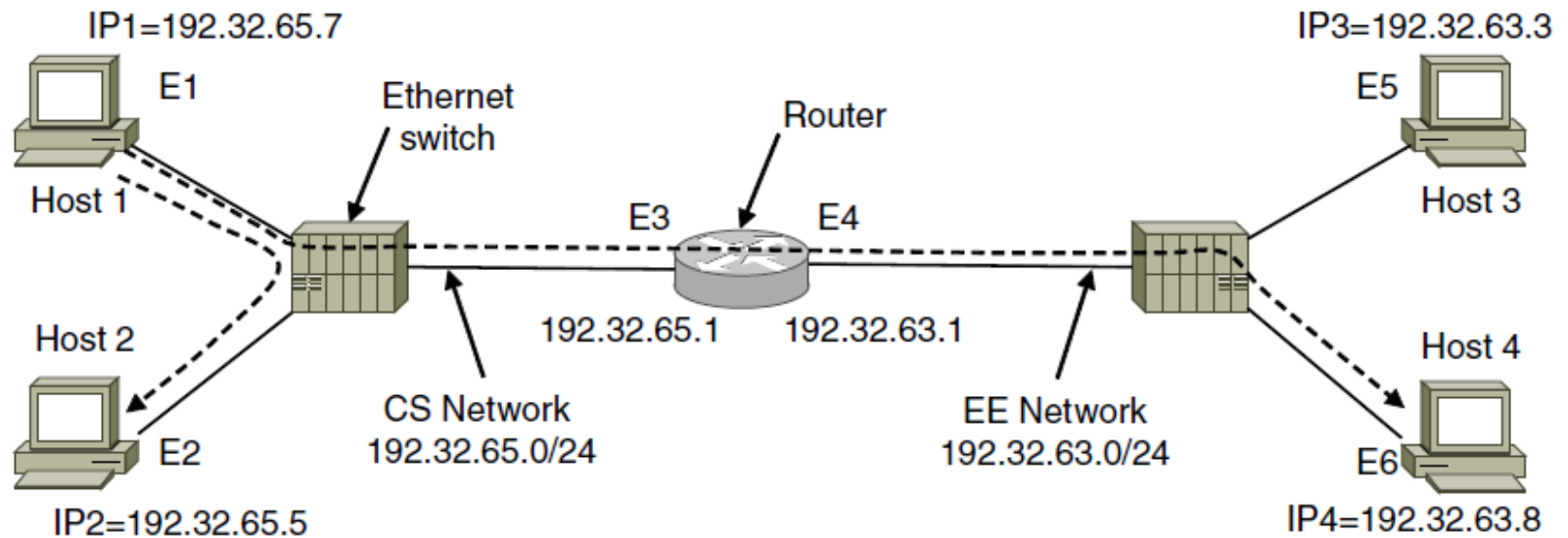
Internet Control Protocols (2)

Main ICMP (Internet Control Message Protocol) types:

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

Internet Control Protocols (3)

ARP (Address Resolution Protocol) lets nodes find target Ethernet addresses [pink] from their IP addresses

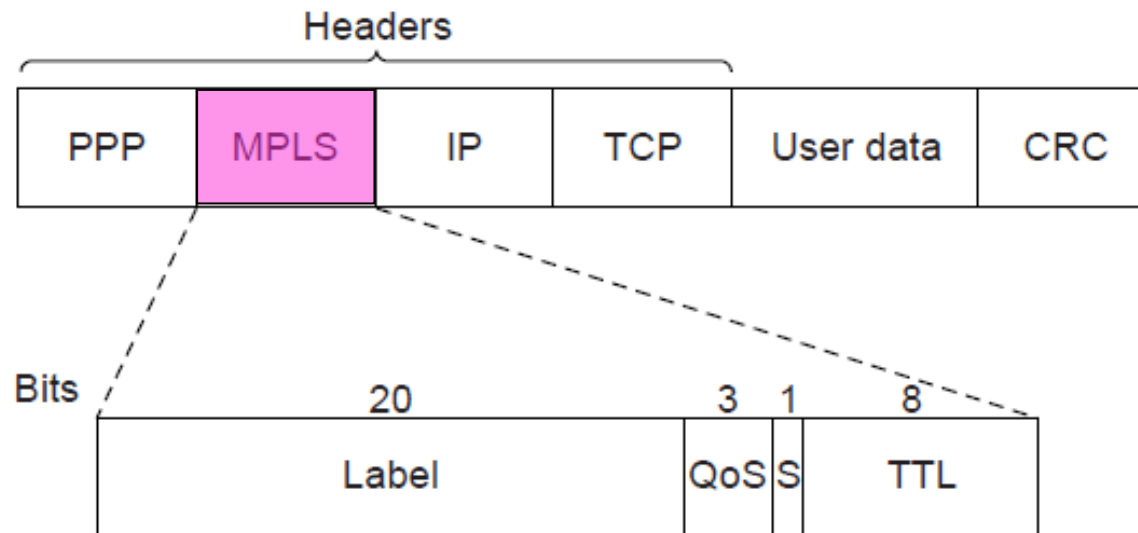


Frame	Source IP	Source Eth.	Destination IP	Destination Eth.
Host 1 to 2, on CS net	IP1	E1	IP2	E2
Host 1 to 4, on CS net	IP1	E1	IP4	E3
Host 1 to 4, on EE net	IP1	E4	IP4	E6

Label Switching and MPLS (1)

MPLS (Multi-Protocol Label Switching) sends packets along established paths; ISPs can use for QoS

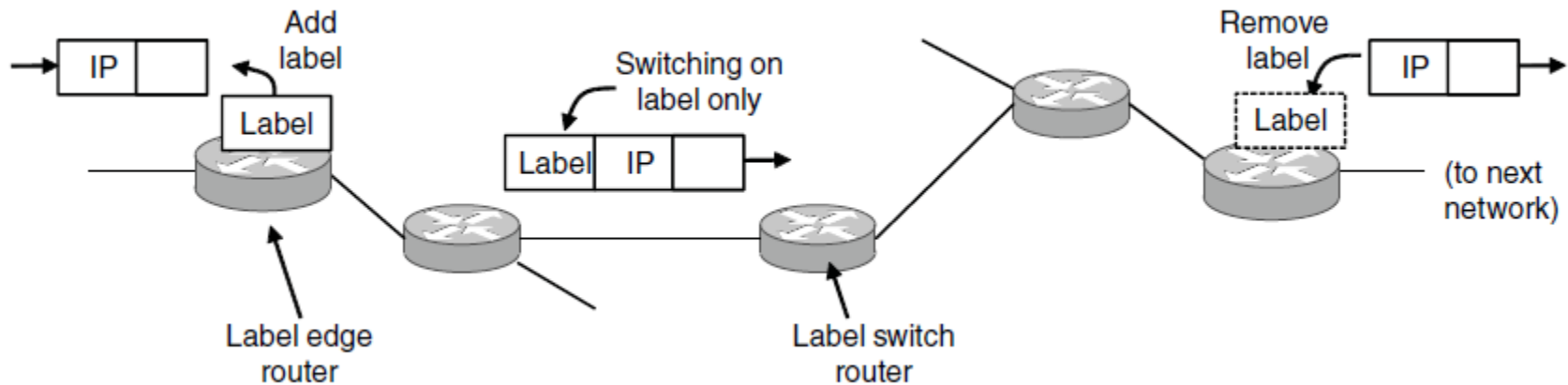
- Path indicated with label below the IP layer



Label Switching and MPLS (2)

Label added based on IP address on entering an MPLS network (e.g., ISP) and removed when leaving it

- Forwarding only uses label inside MPLS network

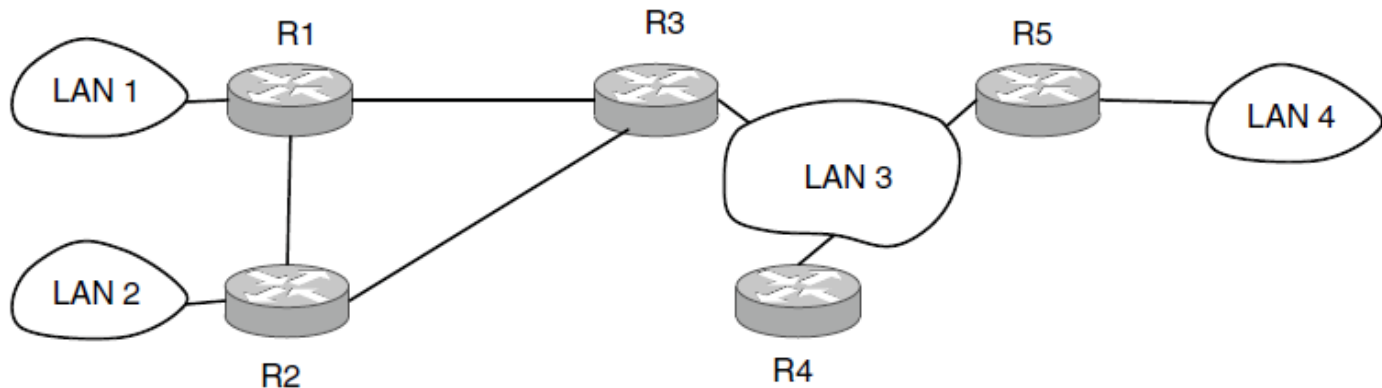


OSPF— Interior Routing Protocol (1)

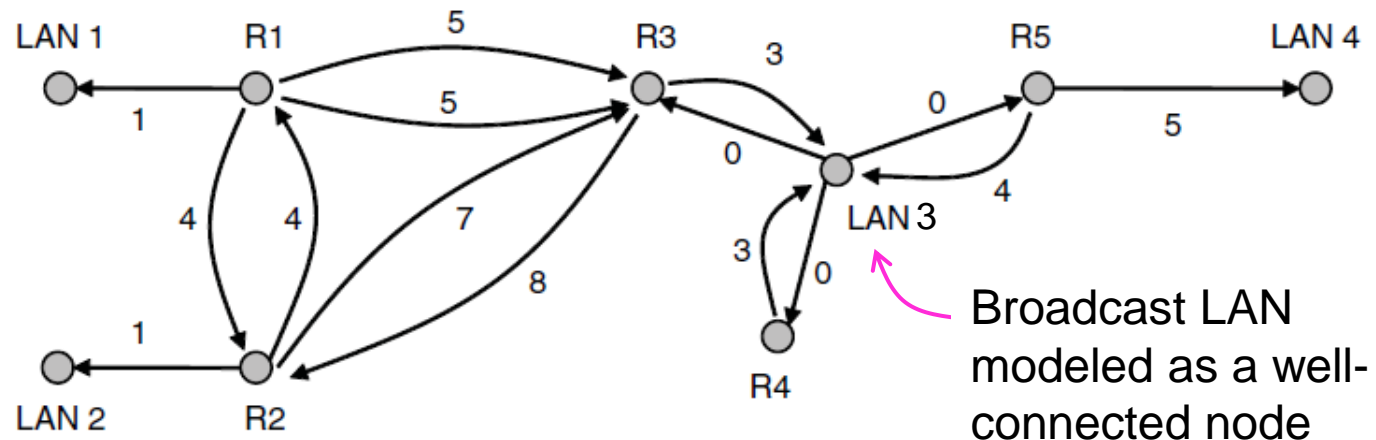
OSPF computes routes for a single network (e.g., ISP)

- Models network as a graph of weighted edges

Network:



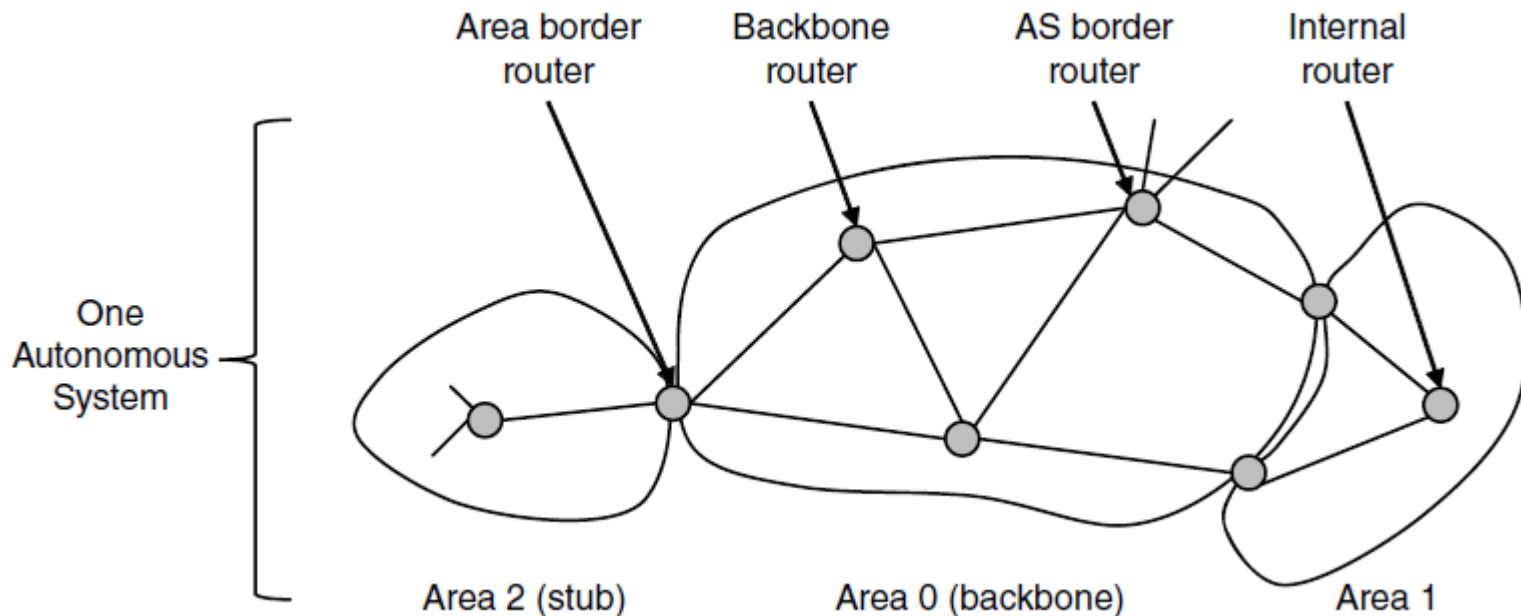
Graph:



OSPF— Interior Routing Protocol (2)

OSPF divides one large network (Autonomous System) into areas connected to a backbone area

- Helps to scale; summaries go over area borders



OSPF— Interior Routing Protocol (3)

OSPF (Open Shortest Path First) is link-state routing:

- Uses messages below to reliably flood topology
- Then runs Dijkstra to compute routes

Message type	Description
Hello	Used to discover who the neighbors are
Link state update	Provides the sender's costs to its neighbors
Link state ack	Acknowledges link state update
Database description	Announces which updates the sender has
Link state request	Requests information from the partner

BGP— Exterior Routing Protocol (1)

BGP (Border Gateway Protocol) computes routes across interconnected, autonomous networks

- Key role is to respect networks' policy constraints

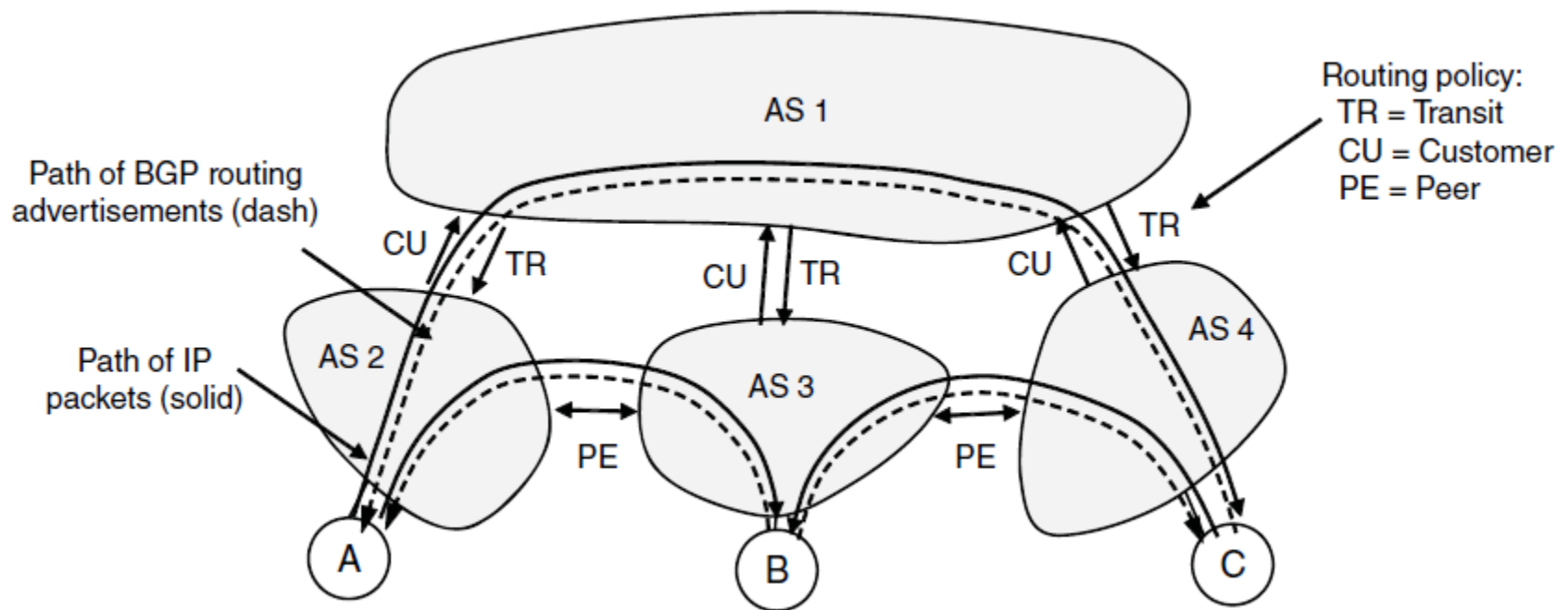
Example policy constraints:

- No commercial traffic for educational network
- Never put Iraq on route starting at Pentagon
- Choose cheaper network
- Choose better performing network
- Don't go from Apple to Google to Apple

BGP— Exterior Routing Protocol (2)

Common policy distinction is transit vs. peering:

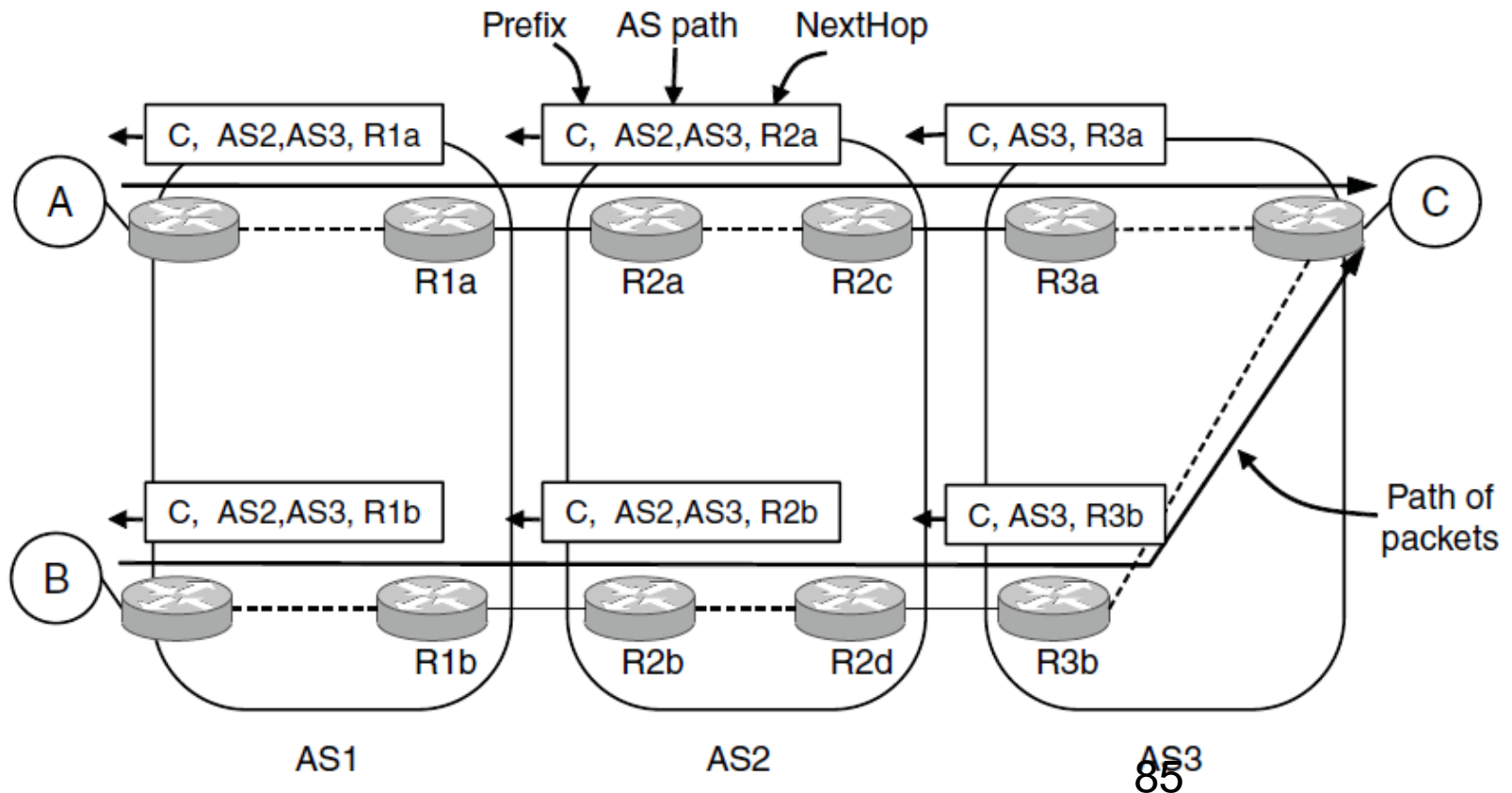
- Transit carries traffic for pay; peers for mutual benefit
- AS1 carries $AS2 \leftrightarrow AS4$ (Transit) but not AS3 (Peer)



BGP— Exterior Routing Protocol (3)

BGP propagates messages along policy-compliant routes

- Message has prefix, AS path (to detect loops) and next-hop IP (to send over the local network)



Internet Multicasting

Groups have a reserved IP address range (class D)

- Membership in a group handled by IGMP (Internet Group Management Protocol) that runs at routers

Routes computed by protocols such as PIM:

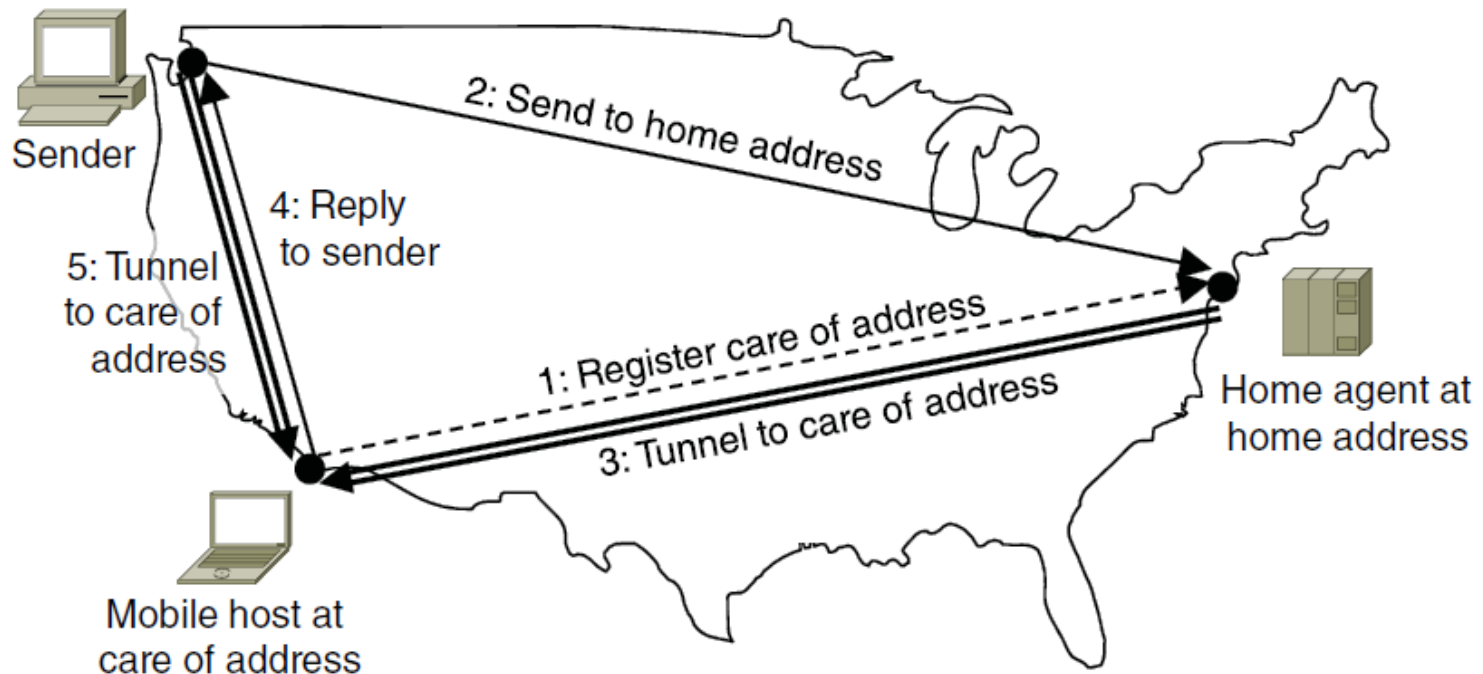
- Dense mode uses RPF with pruning
- Sparse mode uses core-based trees

IP multicasting is not widely used except within a single network, e.g., datacenter, cable TV network.

Mobile IP

Mobile hosts can be reached at fixed IP via a home agent

- Home agent tunnels packets to reach the mobile host; reply can optimize path for subsequent packets
- No changes to routers or fixed hosts



End

Chapter 5