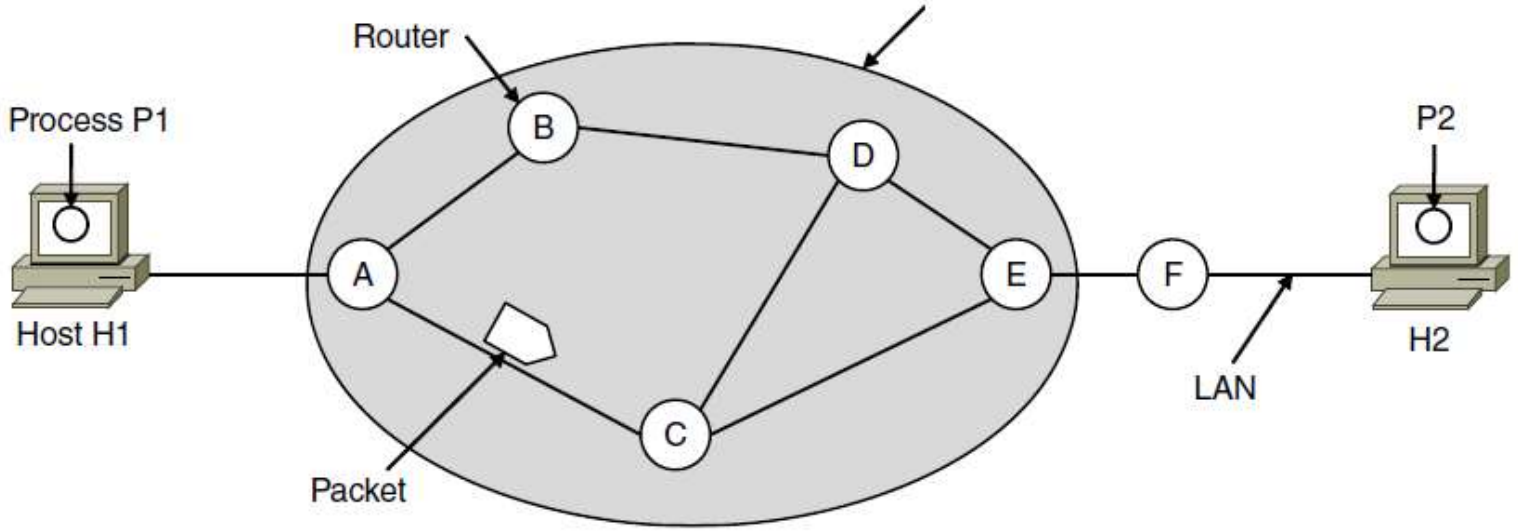
CHAPTER 5
THE NETWORK LAYER

- Network Layer Design Issues（网络层设计概述）
- Routing Algorithms　(路由算法)
- Congestion Control Algorithms （拥塞控制算法）
- Quality of Service (服务质量)
- Internetworking (网络互连)
- The Network Layer in the Internet (互联网中的网络层）

5.1 NETWORK LAYER DESIGN ISUES

1. **Store-and-Forward Packet Switching**
   (存储转发分组交换)
2. **Services Provided to the Transport Layer**
   （为传输层提供的服务）
3. **Implementation of Connectionless Service**
   （无连接服务的实现）
4. **Implementation of Connection-Oriented Service**
   （面向连接服务的实现）
5. **Comparison of Virtual-Circuit and Datagram Subnets**
   （虚电路或数据报子网的比较）

5.1.1 Stored-and-forward packet switching



A host with a packet to send transmits it to the nearest router.

The packet is received, verified, and stored.

Then it is forwarded to the next router.

This step can be repeated many times.

Finally the packet reaches the destination host. **(存储转发分组交换)**

5.1.2 Services provided to the transport layer

- The network layer services should been designed with the following goals in mind.
  1. The services should be independent of the router technology.
  2. The transport layer should be shielded from the number, type, and topology of the routers present.
  3. The network addresses made available to the transport layer should use a uniform numbering plan, even cross LANs and WANs.
- Two types of network layer services:
  - **Connection-oriented service vs connected-less service**

5.1.2 Services provided to the transport layer

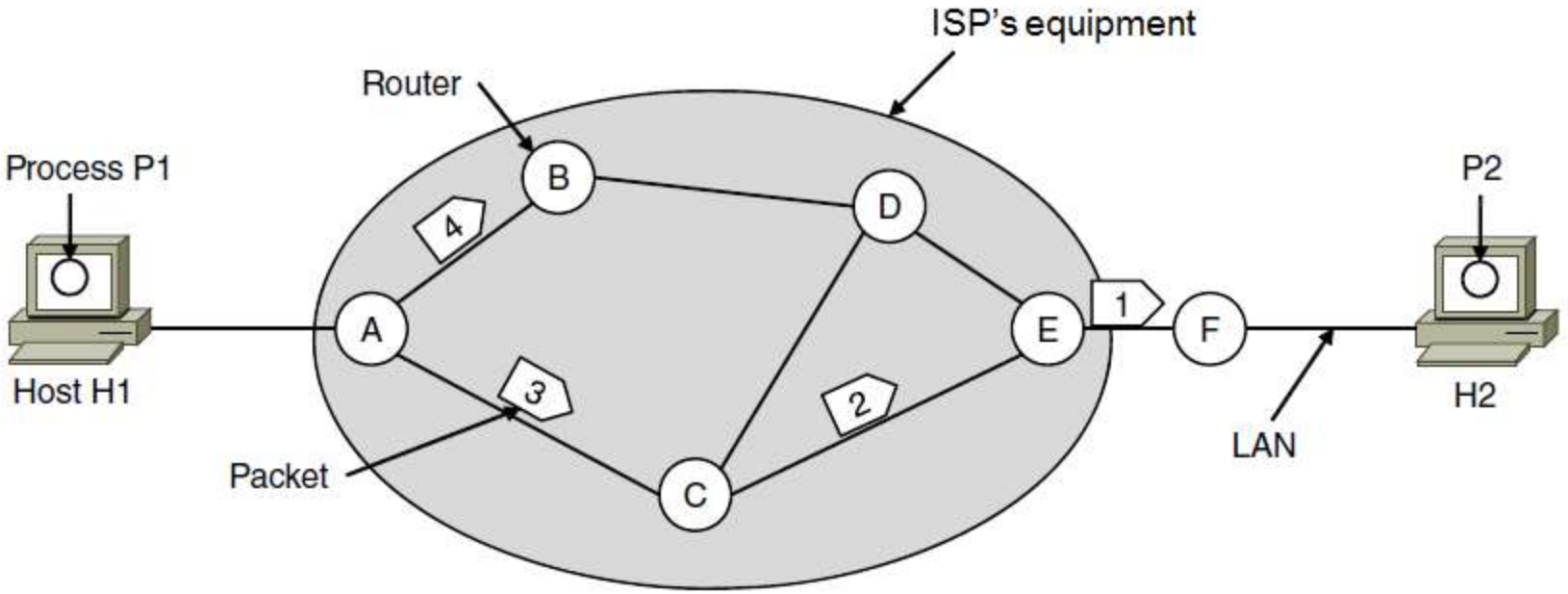
- Two-type of services:
  - **Connection-less services (无连接服务➔Datagram 数据报)**:
    - 40+ years of experience with the computer network, unreliable internet, hosts doing error control and flow control.
  - **Connection-oriented services（面向连接服务 ➔Virtual Circuit虚电路 ）:**
    - 100+ years of experience with the worldwide telephone system, quality of service.

→ **Connection-less + connection-oriented services**.

5.1.2 Services provided to the transport layer

- **Implementation of connectionless services**
  - No advance setup is needed.
  - Packets are injected into the subnet individually and routed independently of each other.
  - The packets are frequently called datagrams (in analogy with telegrams) and the subnet is called a datagram subnet.
- **Implementation of connection-oriented services**
  - A path from the source router to the destination router must be established before any data packets can be sent.
  - This connection is called a VC (virtual circuit), similar to physical circuits set up by the telephone system,
  - The subnet is called a virtual-circuit subnet.

5.1.3 Implementation of Connectionless Service



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

Routing within a datagram network

5.1.3 Implementation of Connectionless Service

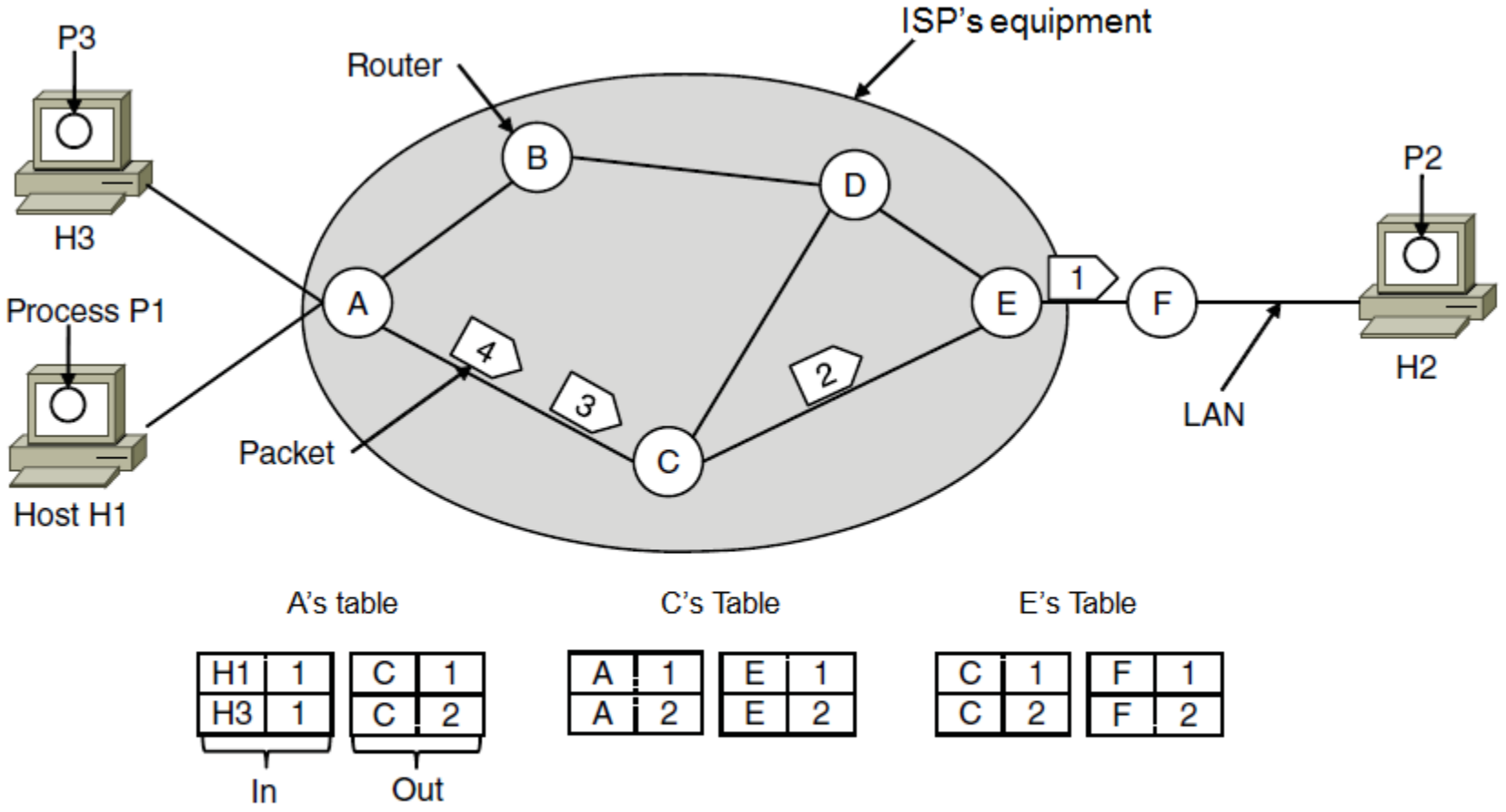
P1 on H1 → P2 on H2

- P1:application layer → H1: transport layer
- H1:transport layer → H1: network layer (disassemble)

- H1:network layer → H2: network layer
  - The routes for packets 1,2,3
    H1 → A → C → E → F → H2
  - The routes for packets 4
    H1 → A → B → D → E → F → H2

- H2: network layer → H2: transport layer (assemble)
- H2: transport layer → P2: application layer

5.1.4 Implementation of Connection-Oriented Service



Routing within a virtual-circuit network

5.1.5 Comparison of Virtual-Circuit and Datagram Subnets

Issue	Datagram subnet	Virtual-circuit subnet
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

5.2 ROUTING ALGORITHMS
（路由选择算法）

1. The Optimality Principle (最优化原则)
2. Shortest Path Routing (最短路径路由)
3. Flooding (扩散路由)
4. Distance Vector Routing (距离向量路由)
5. Link State Routing (链路状态路由)
6. Hierarchical Routing (分层路由)
7. Broadcast Routing (广播路由)
8. Multicast Routing (多点传送路由)
9. Anycast Routing (任意路由)
10. Routing for Mobile Hosts (移动主机的路由)
11. Routing in Ad Hoc Networks (特定网络的路由)

Routing Algorithms: Introduction

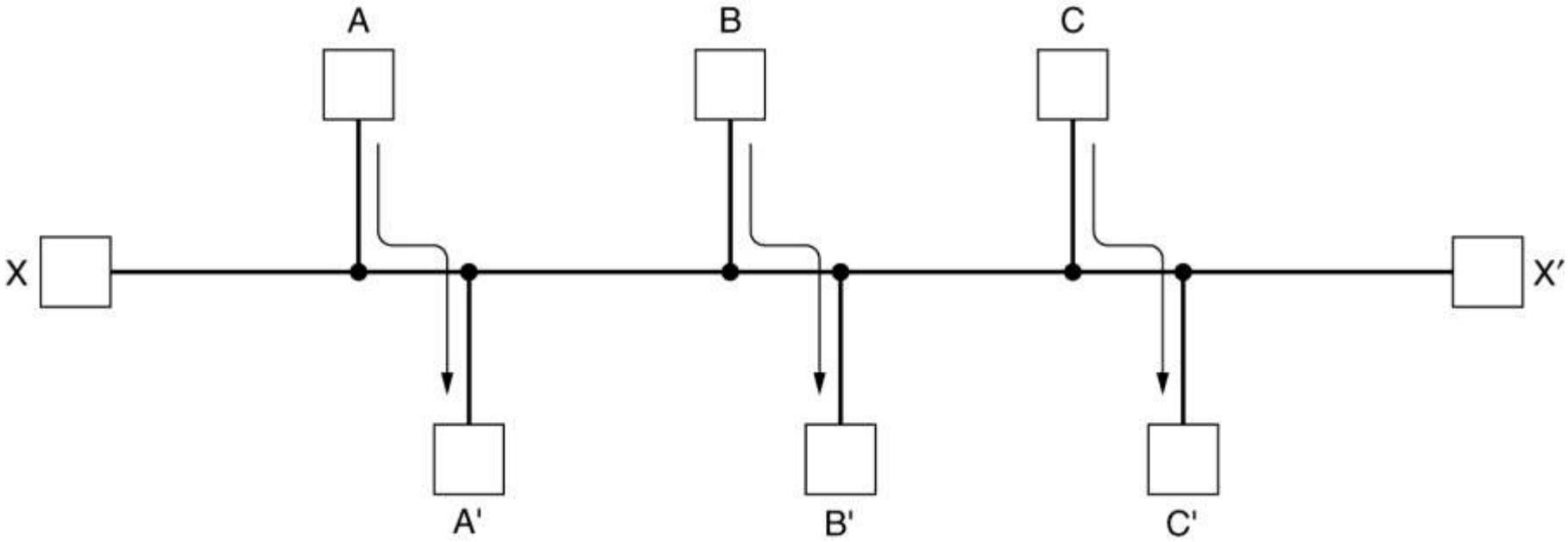
- The main function of the network layer is to route packets from the source machine to the destination machine.
- The **routing algorithm** is used to decide which output line an incoming packet should be transmitted on
  - **For datagram networks**, this decision must be made anew for every arriving data packet since the best route may have changed since last time.
  - **For virtual-circuit networks**, this decision is made only when a new virtual circuit is being set up. Thereafter, data packets just follow the previously established route.

Routing Algorithms: Introduction

- A router performs two tasks:
  - To forward the incoming packet according to the routing table (**Forwarding**)
  - To fill in and update the routing table (**Routing**)
- Desirable properties in a routing algorithm:
  - **Correctness (正确性), simplicity (简单性):** no comment
  - **Robustness (健壮性):** The routing algorithm should cope with changes in the topology and traffic without requiring all processes in all hosts to be aborted and the network to be rebooted every time some router crashes.
  - **Stability(稳定性):** A stable algorithm reaches equilibrium and stays there.
  - **Fairness (公平性), Efficiency (高效率):** Conflict between fairness and efficiency.

Routing Algorithms: Introduction

Conflict between fairness and efficiency.

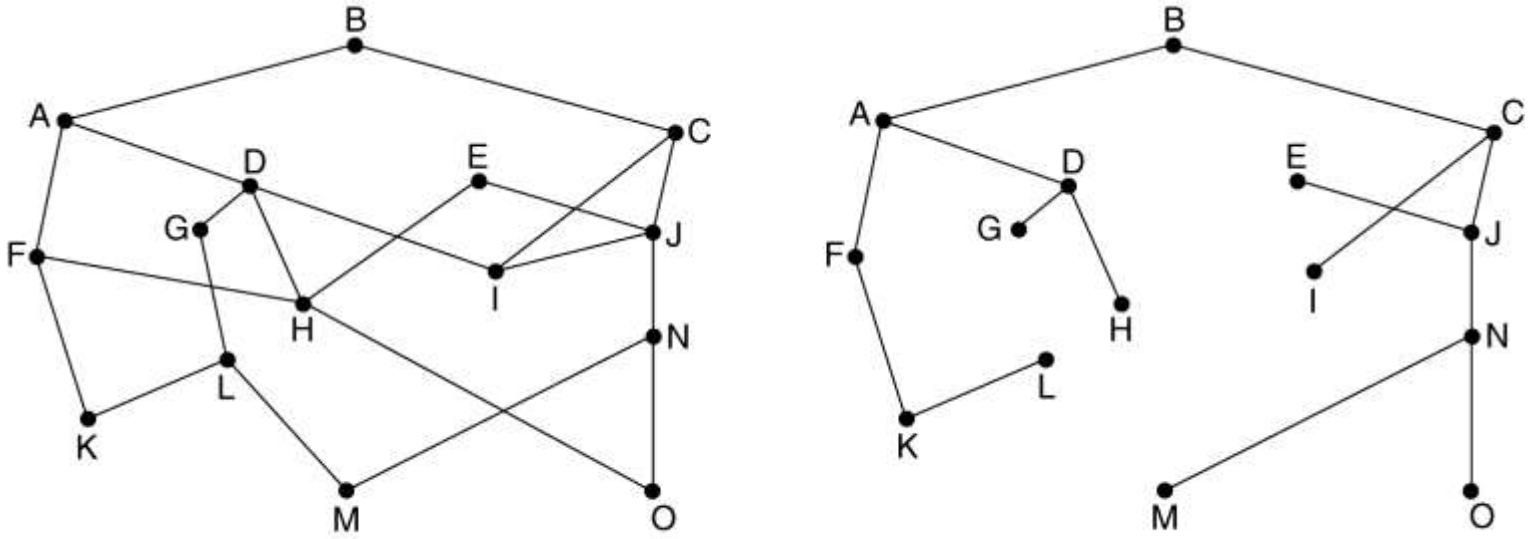


Routing Algorithms: Introduction

- Classes of routing algorithms
  - **Nonadaptive algorithms (非自适应算法)** do not base their routing decisions on measurements or estimates of the current traffic and topology. Instead, the choice of the route to use to get from I to J (for all I . and J) is computed in advance, off-line, and downloaded to the routers when the network is booted.
  - **Adaptive algorithms (自适应算法),** in contrast, change their routing decisions to reflect changes in the topology, and usually the traffic as well. They differ in
    - where they get their information,
    - when they change the routes, and
    - what metric is used for optimization.

Routing Algorithms: The optimality principle

- **Optimality principle(最优化原则):** If router J is on the optimal path from router I to router K (I→J→K), then the optimal path from J to K also falls along the same route.
- **Sink tree (汇集树):** The set of optimal routes from all sources to a given destination form a tree rooted at the destination. (a) A subnet. (b) A sink tree for router B.



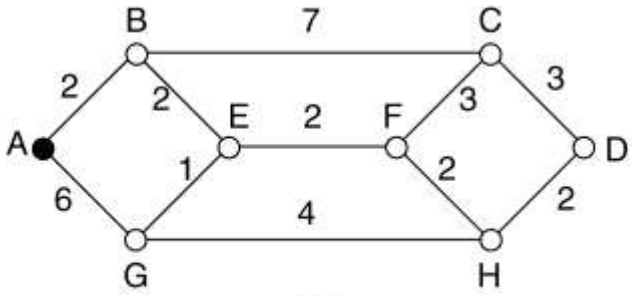
5.2.1Routing Algorithms: Shortest path routing

- **To build a graph of the subnet**,
  - with each graph node representing a router and
  - each graph edge representing a communication line.
- **To choose a route between a given pair of routers**, the algorithm just finds the shortest path between them.
- **How to measure path length**:
  - Hops (结点数量), Physical distance (物理距离), Bandwidth (带宽), traffic (流量), cost (费用), measured delay (测量延迟), mean queue length (平均队列长度) and
  - Other factors or combinations of these factors.

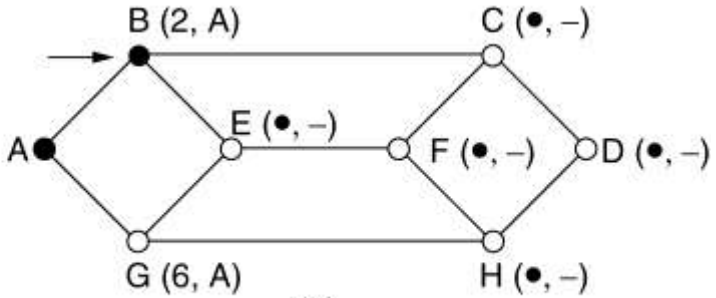
5.2.2Routing Algorithms: Shortest path routing

- **Use Dijkstra's algorithm to compute the shortest path.**
  - Each node is labeled with its distance from the source node along the best known path.
  - Initially, no paths are known, so all nodes are temporarily labeled with infinity.
  - As the algorithm proceeds and paths are found, the labels may change, reflecting better paths. A label may be either temporary or permanent.
  - When it is discovered that a label represents the shortest possible path from the source to that node, it is made permanent and never changed thereafter.

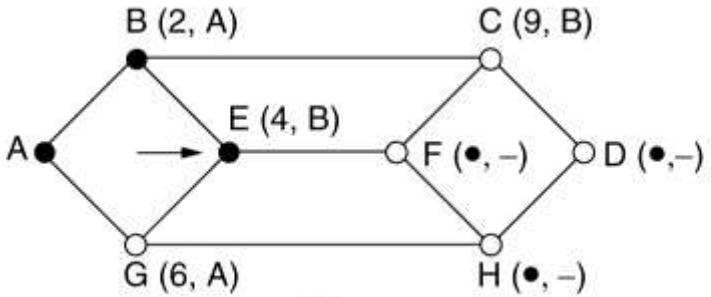
Routing Algorithms: Shortest path routing



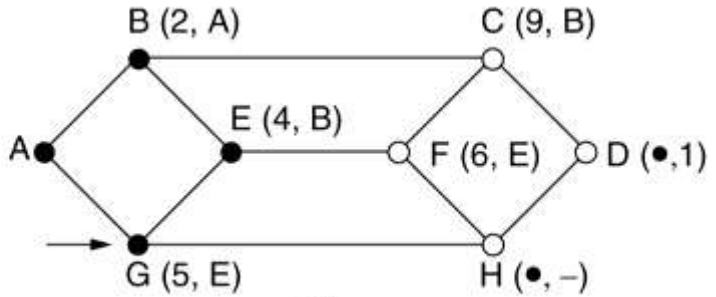
(a)



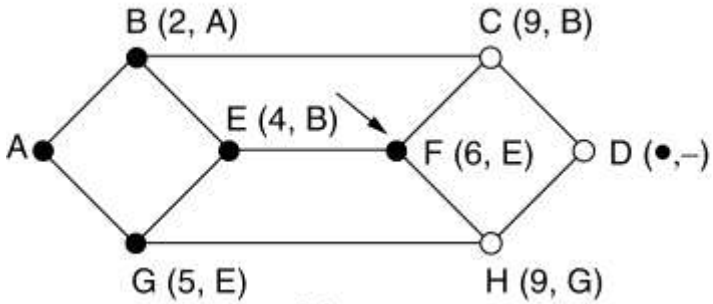
(b)



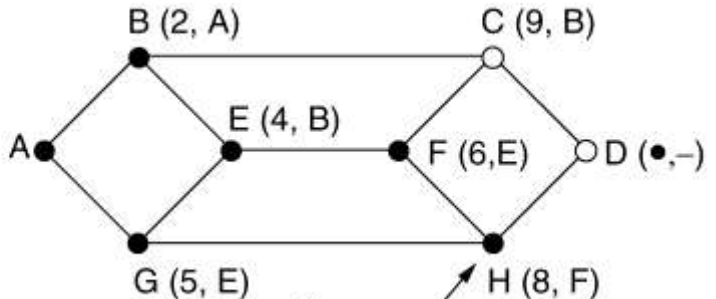
(c)



(d)



(e)



(f)

Routing Algorithms: Shortest path routing

```
#define MAX_NODES 1024                          /* maximum number of nodes */
#define INFINITY 1000000000                     /* a number larger than every maximum path */
int n, dist[MAX_NODES][MAX_NODES];              /* dist[i][j] is the distance from i to j */

void shortest_path(int s, int t, int path[])
{ struct state {                                /* the path being worked on */
      int predecessor;                          /* previous node */
      int length;                               /* length from source to this node */
      enum {permanent, tentative} label;        /* label state */
  } state[MAX_NODES];

  int i, k, min;
  struct state *
               p;
  for (p = &state[0]; p < &state[n]; p++) {     /* initialize state */
      p->predecessor = -1;
      p->length = INFINITY;
      p->label = tentative;
  }
  state[t].length = 0;  state[t].label = permanent;
  k = t;                                        /* k is the initial working node */
  do {                                          /* Is there a better path from k? */
      for (i = 0; i < n; i++)                   /* this graph has n nodes */
          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];
              }
          }

      /* Find the tentatively labeled node with the smallest label. */
      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);

  /* Copy the path into the output array. */
  i = 0;  k = s;
} do {path[i++] = k; k = state[k].predecessor; } while (k >= 0);
```

5.2.3 Routing Algorithms: Flooding

- **Flooding（扩散法）：** Every incoming packet is sent out on every outgoing line except the one it arrived on.
- How to damp the flooding process:
  - One is to have a hop counter contained in the header of each packet, which is decremented at each hop.
  - The other is to keep track of which packets have been flooded, to avoid sending them out a second time.
- A variation of flooding that is slightly more practical is selective flooding.
- Flooding is not practical in most applications, but it does have some uses such as military applications.

5.2.4Routing Algorithms: Distance vector routing

- **Distance vector routing（距离矢量路由）(也叫Bellman-Ford routing)**:
  - Each router maintains a vector or table giving
    - the best known distance to each destination and
    - which line to use to get there.
  - These tables are updated by exchanging information with the neighbors. To update these tables
    - Measure its distance to its neighbors
    - Receive the vectors from its neighbors
    - Compute its own new vector.

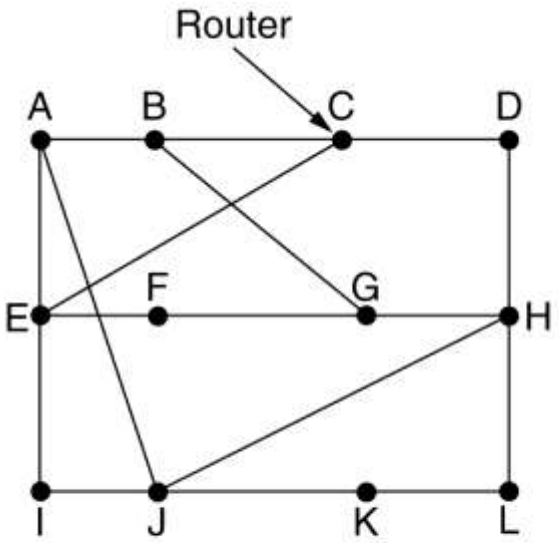
Routing Algorithms: Distance vector routing

- Assume
  - that delay is used as a metric (度量) and that the router knows the delay to each of its neighbors.
  - Once every T msec, each router sends to each neighbor a list of its estimated delays to each destination. It also receives a similar list from each neighbor.
- Imagine
  - one of these tables has just come in from neighbor *X*, with X_i being's estimate of how long it takes to get to router *i*.
  - If the router knows that the delay to *X* is *m* msec,
  - Then it can reach router *i* via *X* in $X_i + m$ msec.
- By performing this calculation for each neighbor, a router can find out which estimate seems the best and use that estimate and the corresponding link in its new routing table.
- Note that the old routing table is not used in the calculation.

Routing Algorithms: Distance vector routing

(a) A subnet

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



(a)

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 routing table for J

Vectors received from J's four neighbors

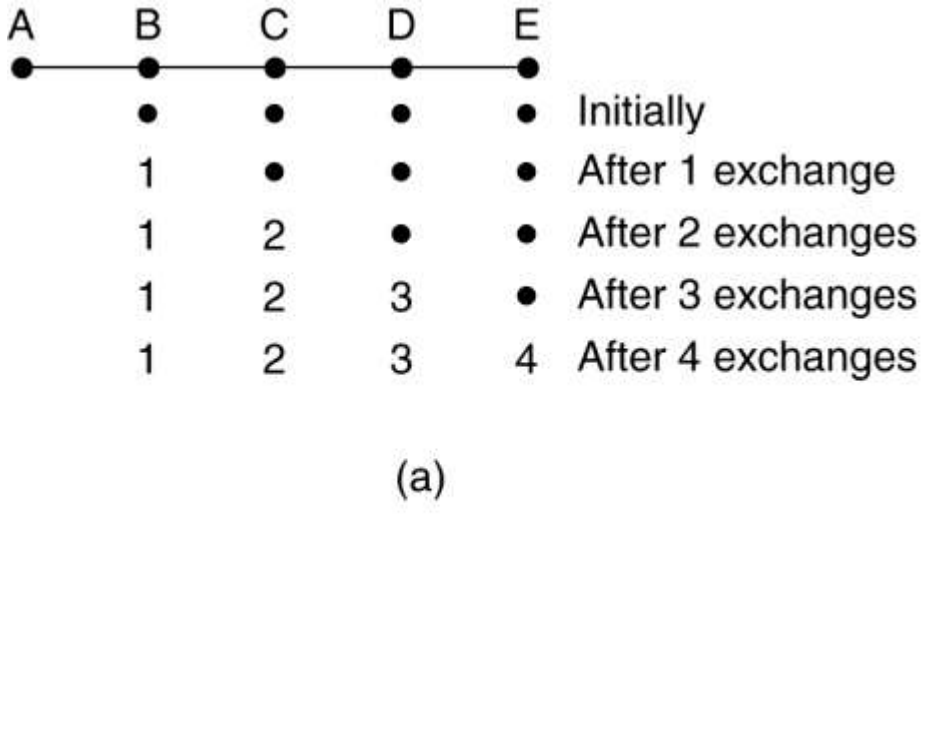
(b)

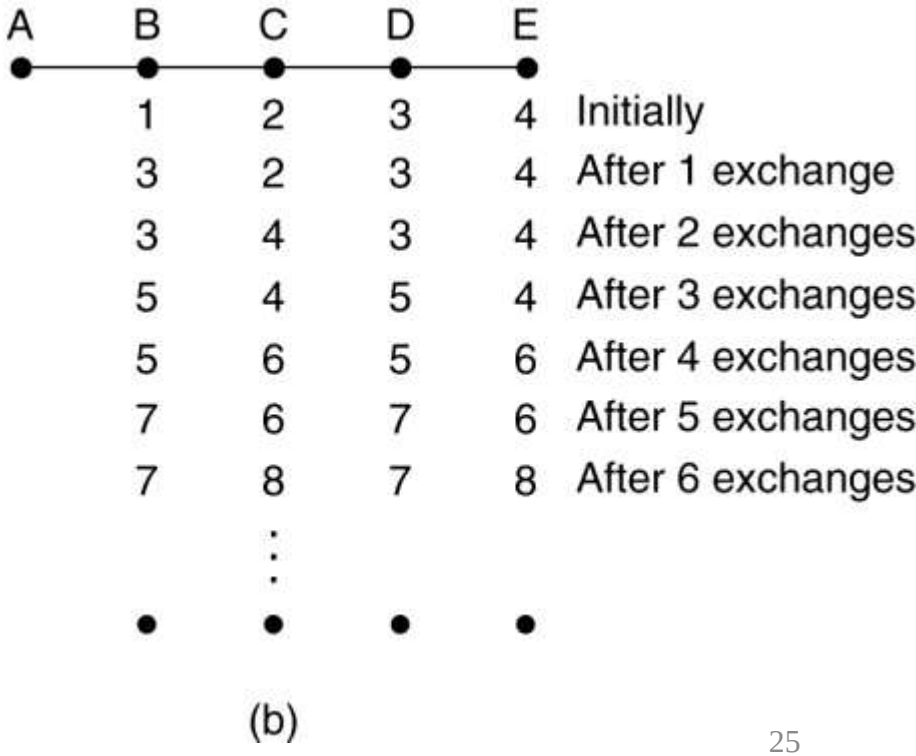
Routing Algorithms: Distance vector routing

The count-to-infinity problem (无穷计数问题):

It reacts rapidly to good news,

but leisurely to bad news.





Routing Algorithms: Distance vector routing

Two main applications:

- **RIP protocol**
- **BGP protocol**

5.2.5 Routing Algorithms: Link state routing

- Problems with distance vector routing
  - The delay metric was queue length, thus it did not take line bandwidth into account when choosing routes.
  - The algorithm often took too long to converge.
- → Link state routing: Each router must do the following:
  1. Discover its neighbors, learn their network address.
  2. Set the distance or cost metric to each of its neighbors.
  3. Construct a packet containing all it has just learned.
  4. Send this packet to and receive packets from all other routers.
  5. Compute the shortest path to every other router.

Routing Algorithms: Link state routing

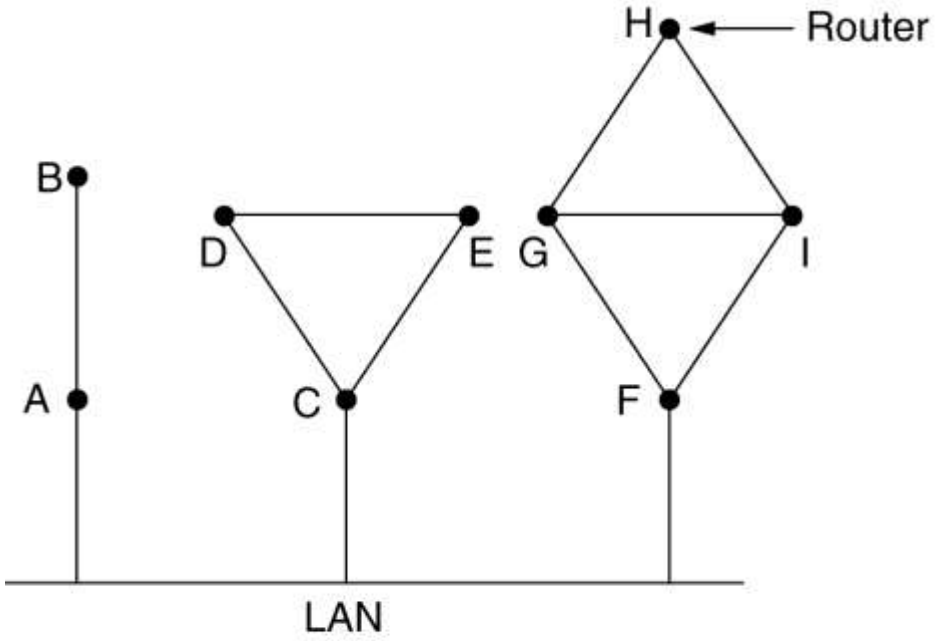
- **Step 1: Learning about the neighbors**
  - One router sends a special HELLO packet on each point-to-point line.
  - The router on the other end is expected to send back a reply telling who it is.
  - These names must be globally unique.
  - → The info about the neighbors can be found out.

Routing Algorithms: Link state routing

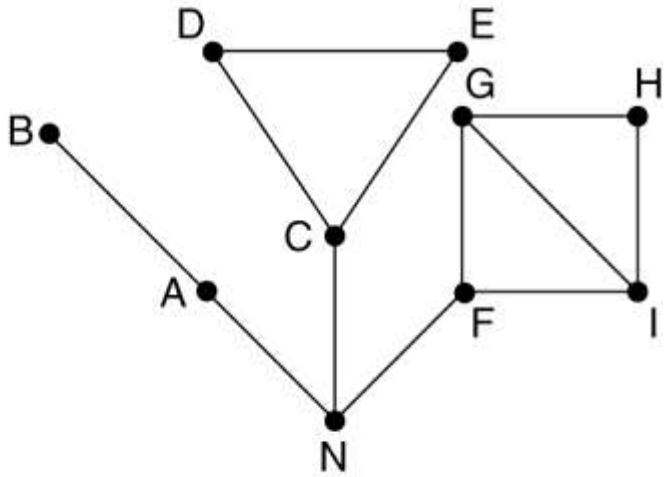
Designated Router (指定路由器) and
Backup Designated Router (后备指定路由器)

(a) Nine routers and a LAN.

(b) A graph model of (a)



(a)



(b)

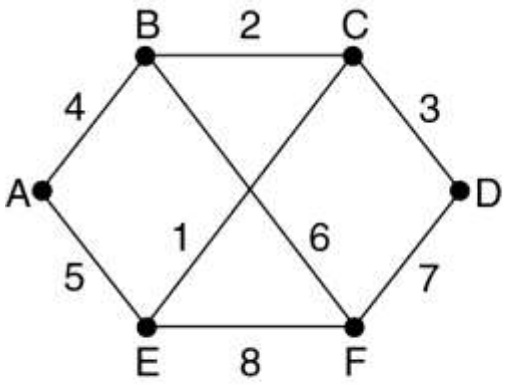
Routing Algorithms: Link state routing

- **Step 2: Measuring line cost**
  - To determine the delay is to send over the line a special ECHO packet that the other side is required to send back immediately.
  - By measuring the round-trip time and dividing it by 2, the sending router can get a reasonable estimate of the delay.
  - Average delay value can be better.

Routing Algorithms: Link state routing

- **Steps 3: Building link state packets**
  - Each router builds a packet containing all the data.
  - The packet starts with the identity of the sender, followed by a sequence number and age and a list of neighbors.
  - To build them is easy, but when to build them is difficult to determine:
    - To build them periodically (at regular intervals)
    - To build them when some significant event occurs, such as a line or neighbor going down or coming back up again or changing its properties appreciably.

Routing Algorithms: Link state routing



(a)

Link State Packets

A	
Seq.	
Age	
B	4
E	5

B	
Seq.	
Age	
A	4
C	2
F	6

C	
Seq.	
Age	
B	2
D	3
E	1

D	
Seq.	
Age	
C	3
F	7

E	
Seq.	
Age	
A	5
C	1
F	8

F	
Seq.	
Age	
B	6
D	7
E	8

(b)

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

Routing Algorithms: Link state routing

- **Steps 4: Distributing the link state packets**
  - To use flooding to distribute the link state packets
  - To keep track of all the (source router, sequence) pairs they see.
  - To include the age of each packet after the sequence number and decrement it once per second.

Routing Algorithms: Link state routing

The packet buffer for router B

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	

Routing Algorithms: Link state routing

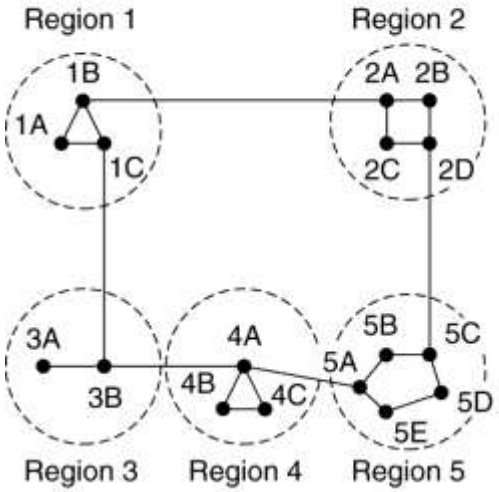
- **Steps 5: Computing the new routes**
  - Once a router has accumulated a full set of link state packets, it can construct the entire subnet graph because every link is represented.
  - Dijkstra's algorithm can be run locally to construct the shortest path to all possible destinations. The results of this algorithm can be installed in the routing tables.
  - Some applications
    - IS-IS (Intermediate System – Intermediate System)
    - OSPF (Open Shortest Path First)

5.2.6 Routing Algorithms: Hierarchical routing

- As networks grow in size, the router routing tables grow proportional, and so do router memory and computing power.
- **Two level routing:** Every router knows
  - all the details about how to route packets to destinations within its own region
  - but knows nothing about the internal structure of **other regions.**
- **Multiple-level routing:**
  - Regions → clusters → zones → groups → …

Routing Algorithms: Hierarchical routing

Reduction of routing tables



(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)

Routing Algorithms: Hierarchical routing

- **How many levels should the hierarchy have?**
  - Consider a subnet with 720 routers
    - No hierarchy: every router needs 720 routing table entries.
    - 30 routers/region x 24 regions : every router needs 30 for local entries + 23 for other regions = 53 table entries.
    - 10 routers/region x 9 regions/cluster x 8 clusters : every router needs 10 + 8 + 7 = 25 table entries.
  - ➔Kamount and Kleinrock (1979): The optimal number of levels for an *N* router subnet is *lnN*, requiring a total of e *ln N* entries per router.

5.2.7 Routing Algorithms: Broadcast routing

- **Broadcasting: to send a packet to all destinations simultaneously.**
  - The source simply sends a distinct packet to every destination.
  - Flooding.
  - Multidestination routing (each packet contains either a list of destinations or a bit map indicating the desired destinations.) (One router pays full fare and the rest ride free.)
  - To make use of the sink tree for the router initiating the broadcast.
  - Reverse path forwarding.

Routing Algorithms: Broadcast routing

– **Reverse path forwarding:**

- When a broadcast packet arrives at a router, the router checks to see if the packet arrived on the line that is normally used for sending packets to the source of the packets.
- If so, there is an excellent chance that the broadcast packet itself followed the best route from the router and is therefore the first copy to arrive at the router. Then the router forwards copies of it onto all lines except the one it arrived on.
- If no, it is discarded as a duplicate.

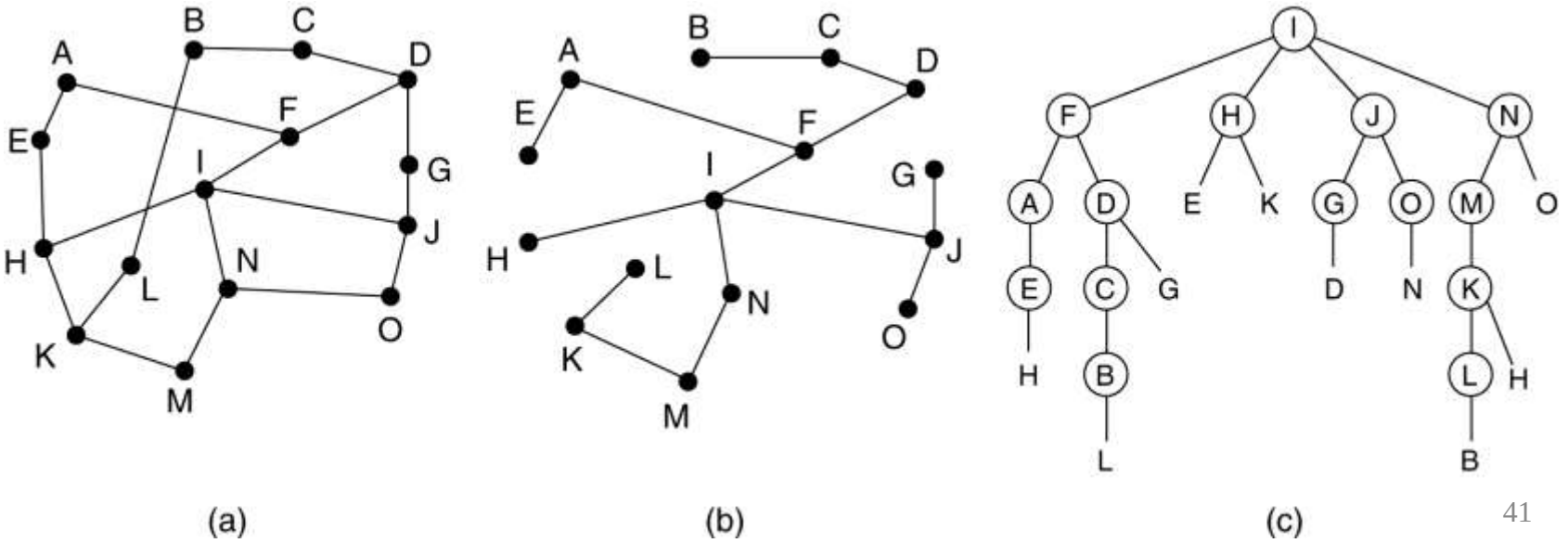
Routing Algorithms: Broadcast routing

Reverse path forwarding.

(a) A subnet.

(b) a Sink tree.

(c) The tree built by reverse path forwarding.



5.2.8 Routing Algorithms: Multicast routing

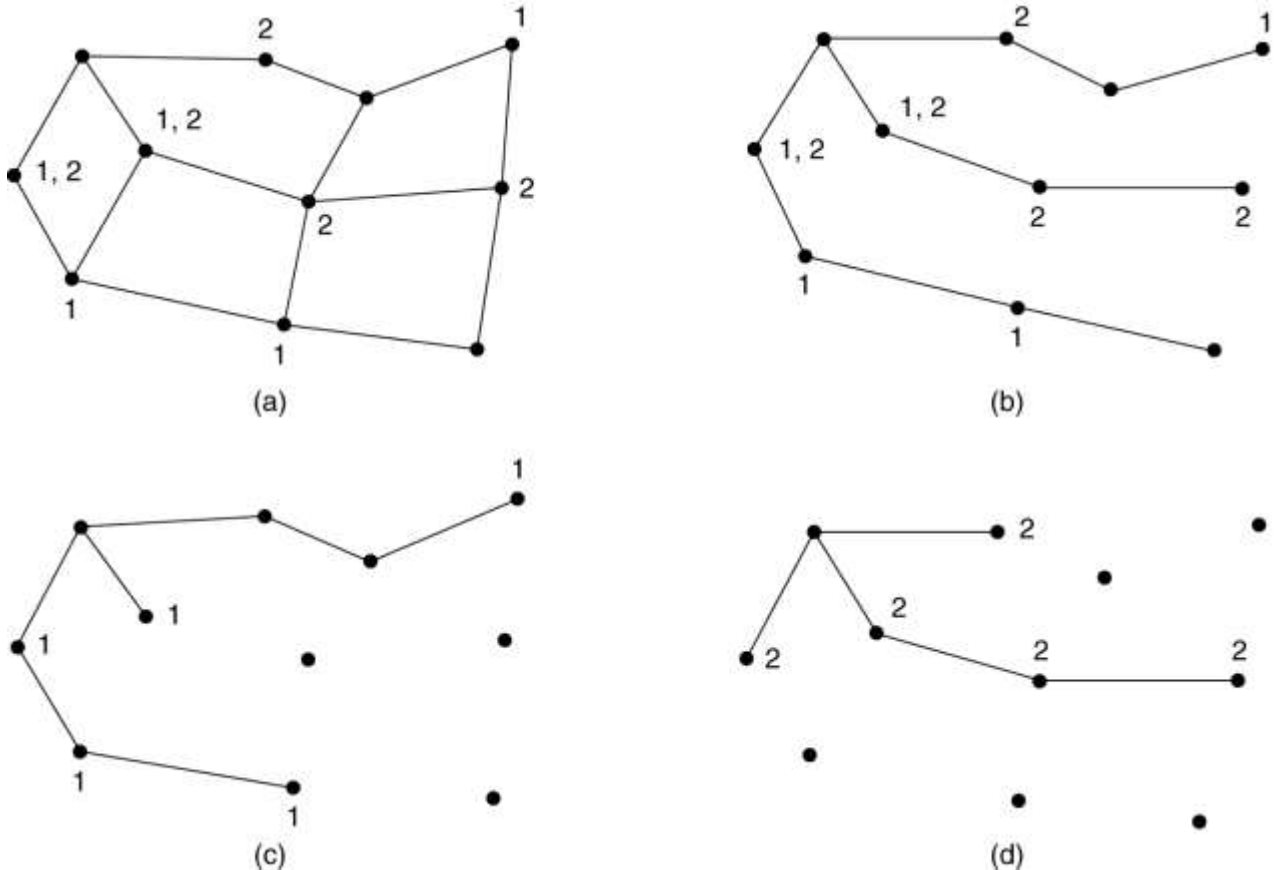
- **Multicasting (组播):** to send messages to well-defined groups that are numerically large in size but small compared to the network as whole.
  - Group management: Some way is needed to create and destroy groups, and to allow processes to join and leave groups.
  - Computing a spanning tree covering all other routers.
  - Multicast routing is to prune the spanning tree.
    - When a process sends a multicast packet to a group,
    - The first router examines its spanning tree
    - and prunes it, removing all the lines that do not lead to hosts that are members of the group.

Routing Algorithms: Multicast routing

(a) A network. (b) A spanning tree for the leftmost router.

(c) A multicast tree for group 1.

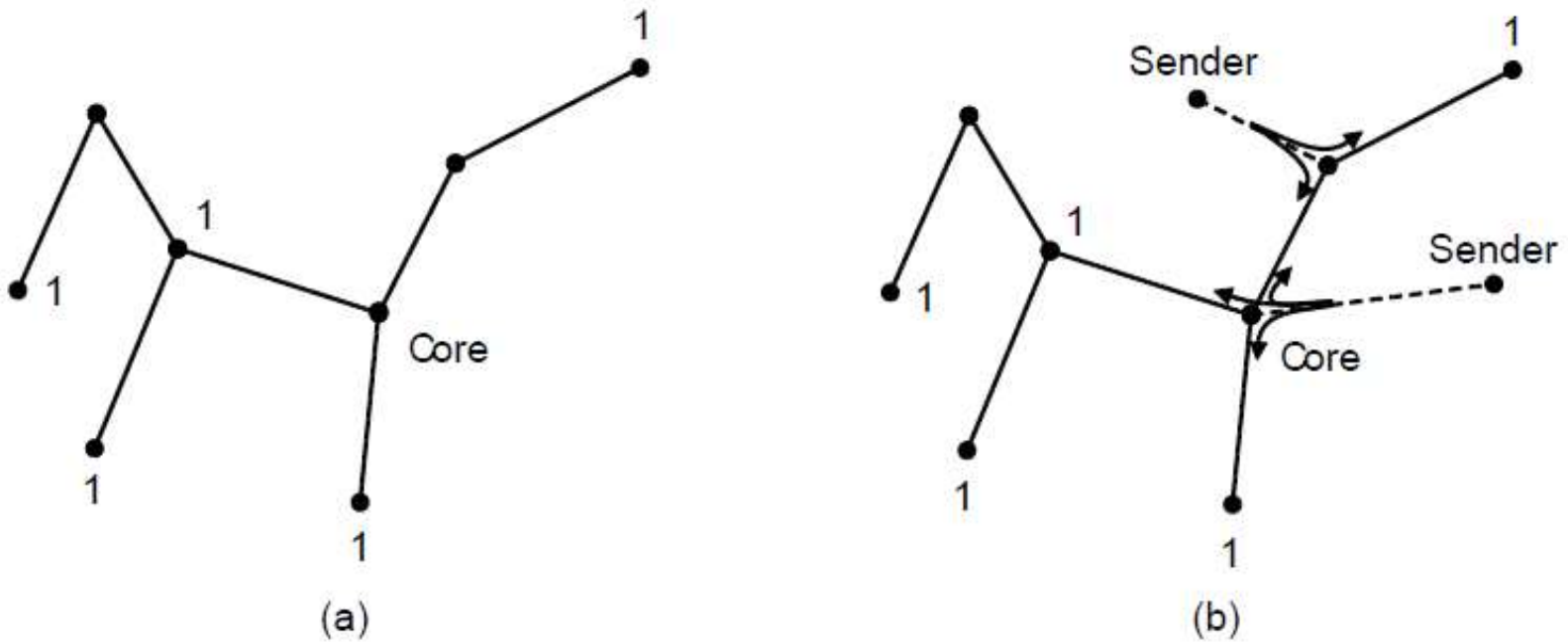
(d) A multicast tree for group 2.



Routing Algorithms: Multicast routing

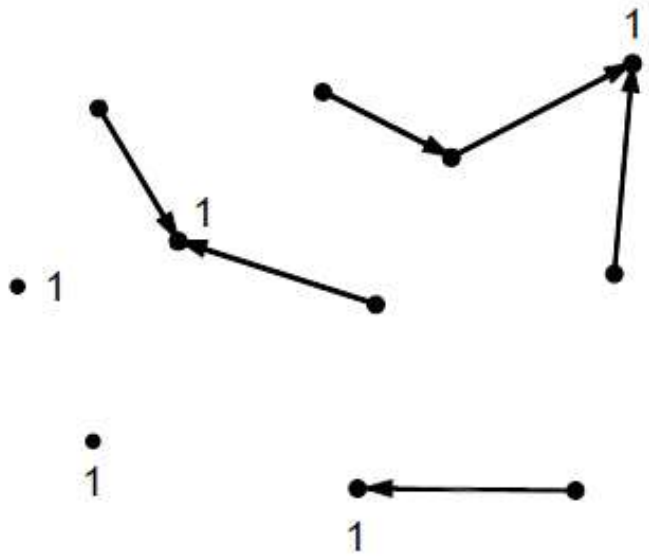
(a)Core-based tree for group 1.

(b)Sending to group 1.

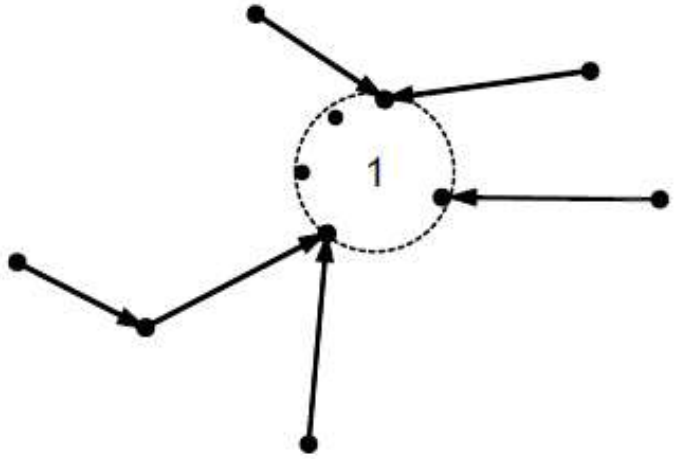


5.2.9 Routing Algorithms: Anycast routing

- In anycast, a packet is delivered to the nearest member of a group. Schemes that find these paths are called anycast routing.

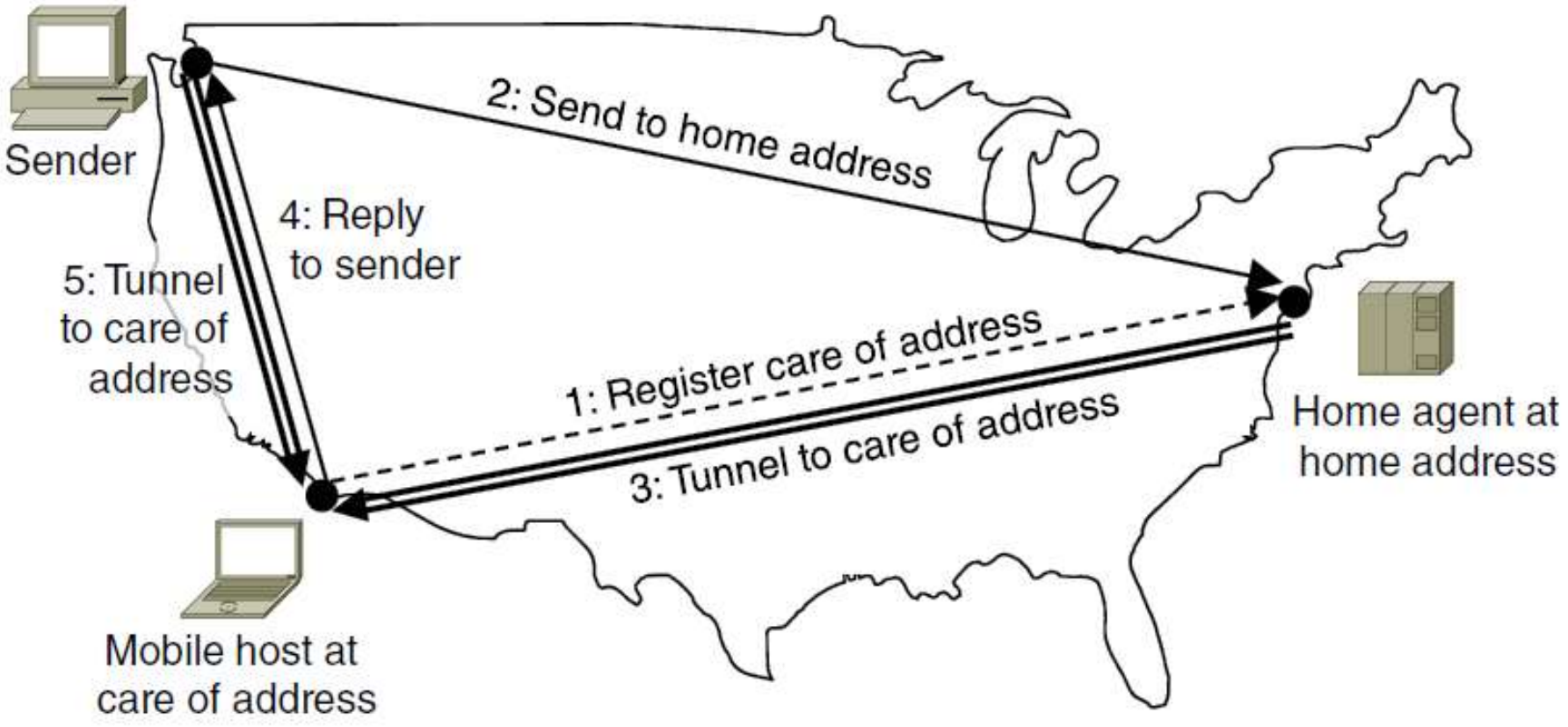


(a)



(b)

Routing Algorithms: Routing the mobile hosts



5.2.11 Routing Algorithms:

Routing in Ad Hoc Networks: Route discovery

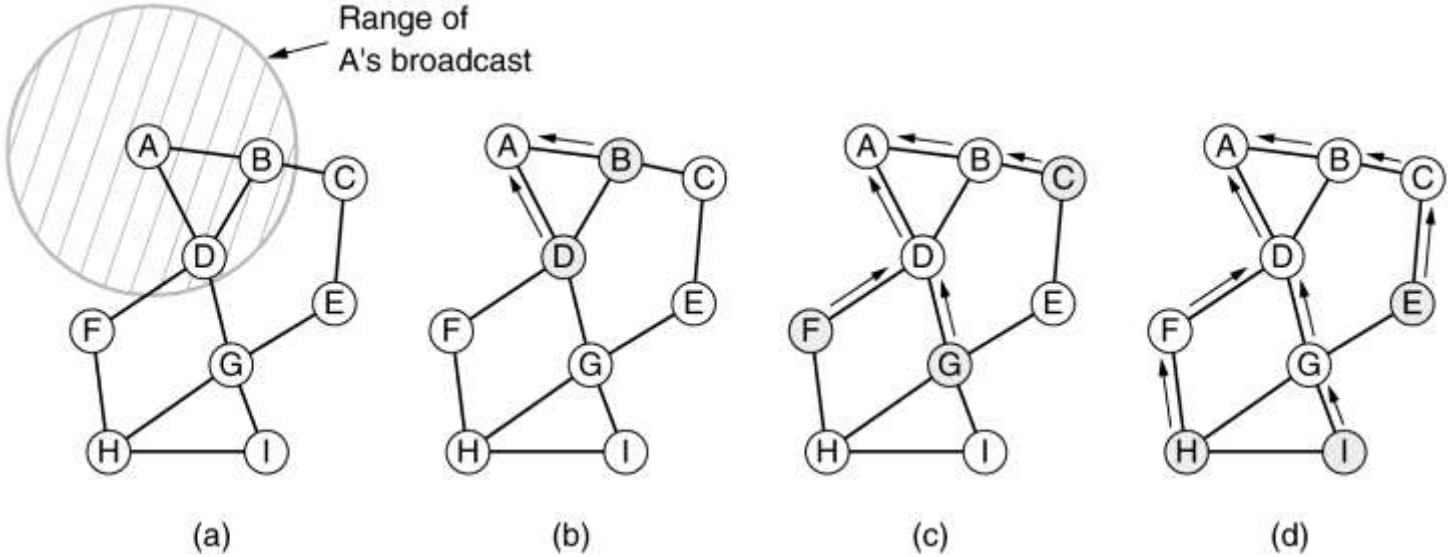
(a) Range of A's broadcast.

(b) After B and D have received A's broadcast.

(c) After C, F, and G have received A's broadcast.

(d) After E, H, and I have received A's broadcast.

Shaded nodes are new recipients. Arrows show possible reverse routes.

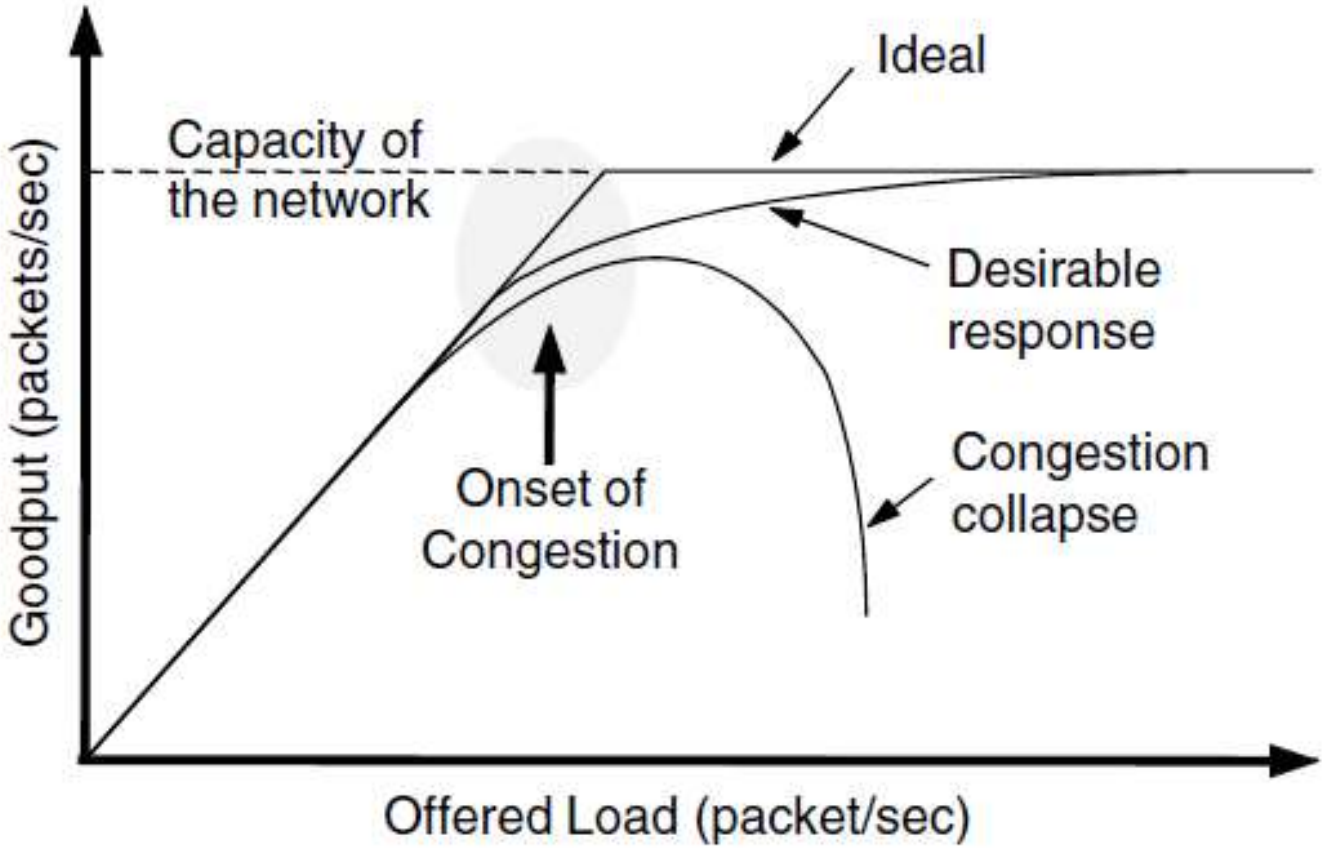


5.3 CONGESTION CONTROL ALGORITHMS (拥塞控制算法) (*)

- Approaches to Congestion Control
- Traffic-Aware Routing
- Admission Control
- Traffic Throttling
- Load Shedding

Congestion Control Algorithms: Introduction

Congestion: When too much traffic is offered, congestion sets in and performance degrades sharply.



Congestion Control Algorithms: Introduction

Congestion causes (拥塞原因):

- Burst packets on one output line.
  - If all of a sudden, streams of packets begin arriving on three or four input lines and all need the same output line, a queue will build up. If there is insufficient memory to hold all of them, packets will be lost. Adding more memory will not help a little but not much.
- Mismatch between parts of the system
  - Insufficient memory.
  - Slow CPU.
  - Low bandwidth.

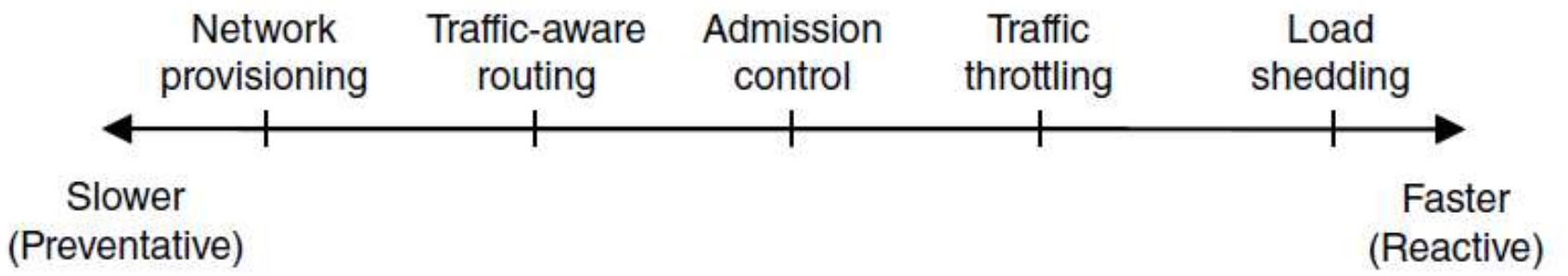
Congestion Control Algorithms: Introduction

Congestion control and flow control

- Differences:
  - Congestion control: Subnet, global.
    - A store-and-forward network with 1-Mbps lines and 1000 large computers, half of which are trying to transfer files at 100kbps to the other half.
  - Flow control: end-to-end traffic, local.
    - An example: 1000Gbps (supercomputer) → 1Gbps (personal computer).
- Similarity: Both congestion control and flow control can tell the sender to slow down
  - because the **receiver** cannot handle the load or
  - because the **network** cannot handle it.

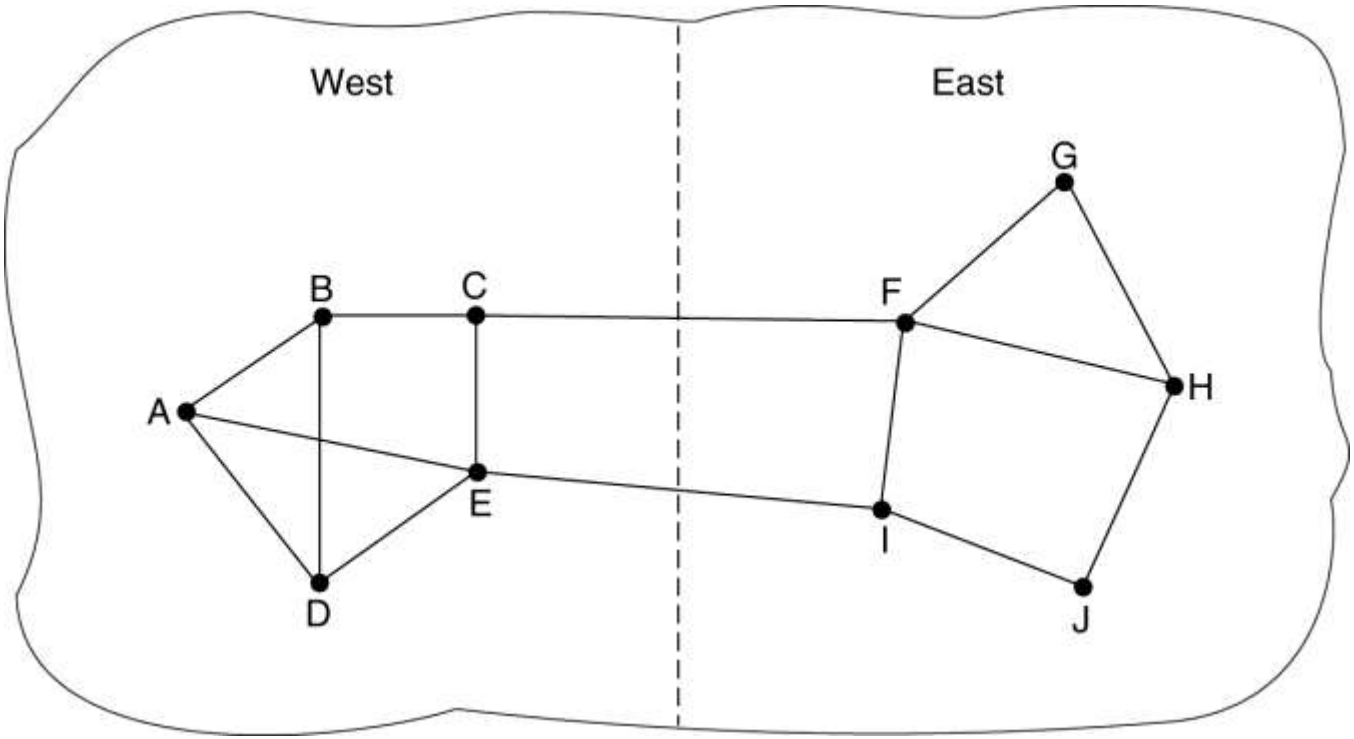
5.3.1 Congestion Control Algorithms: Approaches to Congestion Control

- Time scales of approaches to congestion control
  - Network provisioning (网络供应)： months
  - Traffic-aware routing: hours
  - Admission control: minutes
  - Traffic throttling: seconds
  - Load shedding: seconds



5.3.2 Congestion Control Algorithms: Traffic-Aware Routing

A subnet in which the East and West parts are connected by two lines.

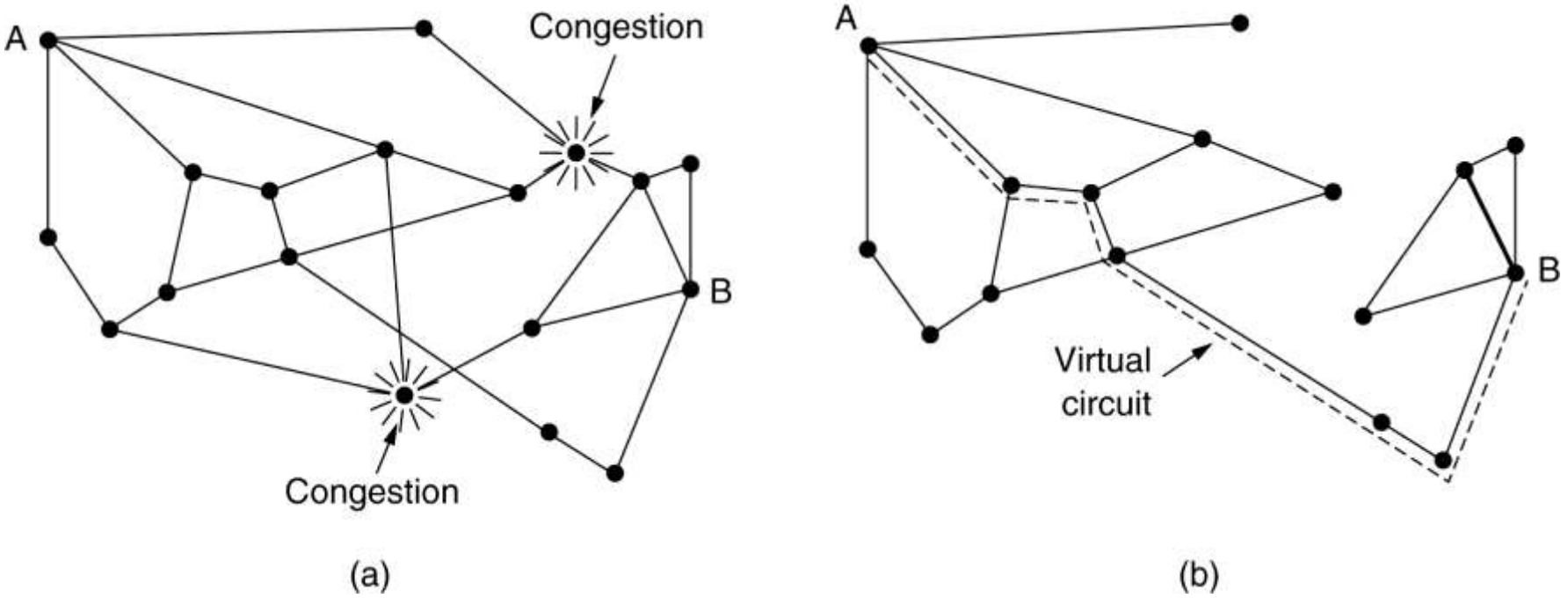


5.3.3 Congestion Control Algorithms:
Admission control

- Traffic is often described in terms of its rate and shape.
- A commonly used descriptor that captures this effect is the **leaky bucket** or **token bucket**.
- Armed with traffic descriptions, the network can decide whether to admit the new virtual circuit.

Congestion Control Algorithms:
Admission control

(a) A congested subnet. (b) A redrawn subnet, eliminates congestion and a virtual circuit from A to B.



5.3.4 Congestion Control Algorithms: Traffic Throttling(流量节流）

- In the Internet and many other computer networks, senders adjust their transmissions to send as much traffic as the network can readily deliver.
  1. Routers must determine when congestion is approaching, ideally before it has arrived.
  2. Routers must deliver timely feedback to the senders that are causing the congestion.

Congestion Control Algorithms:
Traffic Throttling

- Each router monitors the utilization of its output lines and other resources.

$$u_{new} = au_{old} + (1-a)f$$

- Whenever *u* moves above the threshold, the output lines enters a “warning stat”.
- Possible actions:
  - Warning bits (警告位)
  - Choke packets (抑制包)
  - Hop-by-hop choke packets (单跳抑制包)

Congestion Control Algorithms:
Traffic Throttling

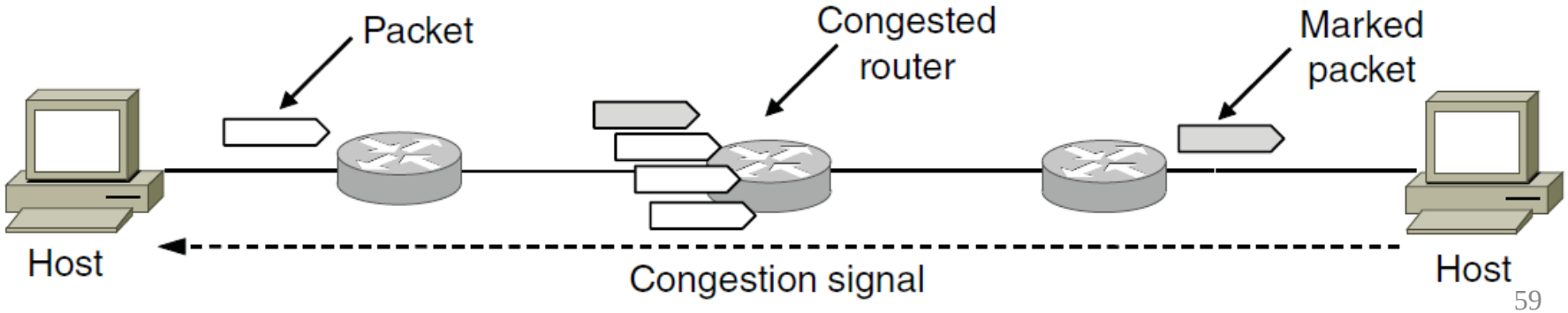
Choke packets （抑制包）

- If congestion occurs, the router sends a choke packet back to the source host,
- When the source gets the choke packet, it reduces the traffic sent to the specified destination by X percent. And the source will ignore other choke packets referred to that destination for a fixed time interval. ……
- Types of choke packets:
  - a mild（温和） warning,
  - a stern（严格）warning,
  - an ultimatum（强制）

Congestion Control Algorithms:
Congestion Control in Datagram Subnets

Explicit Congestion Notification

- Two bits in the IP packet header are used to record whether the packet has experienced congestion.
- If any of the routers they pass through is congested, that router will then mark the packet as having experienced congestion as it is forwarded.
- The destination will then echo any marks back to the sender as an explicit congestion signal



Congestion Control Algorithms:
Congestion Control in Datagram Subnets

Hop-by-hop choke packets

- At high speeds or over long distances, sending a choke packet to the source hosts does not work well because the reaction is so slow.
  - For example, a host in San Francisco (router A) sends packets to a host in New York at 155Mbps.
  - When the New York host begins to run out of buffers, it will take about 40msec for a choke packet to get back to San Francisco to tell it to slow down.
  - In those 40 msec, another 6.2 megabits will have been sent.
- To have the choke packet take effect at every hop it passes through.

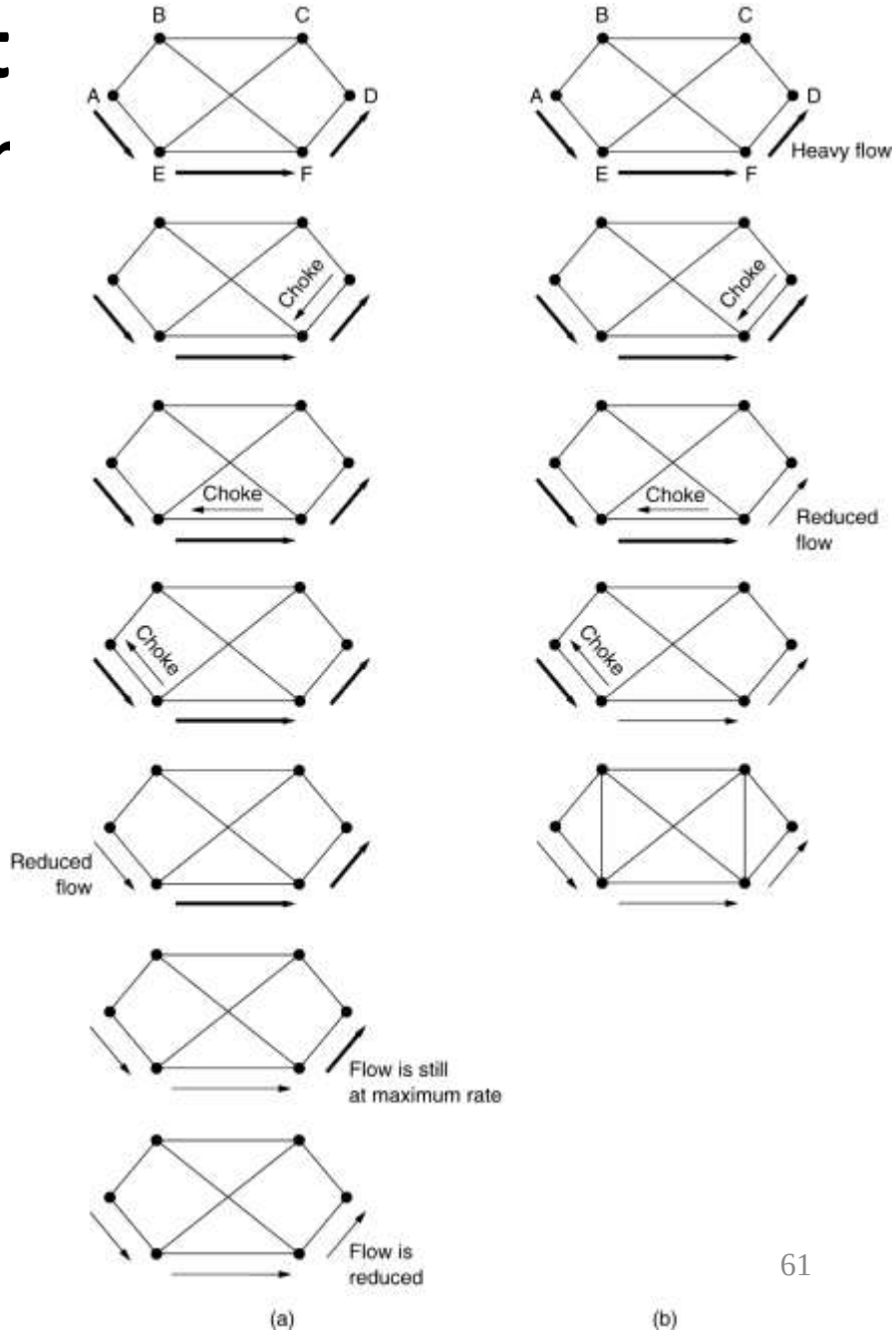
Congestion Cont

Congestion Control ir

- **Hop-by-Hop Choke Packets**

(a) A choke packet that affects only the source.

(b) A choke packet that affects each hop it passes through.



5.3.5 Congestion Control Algorithms: Load shedding

- **Load shedding** is a fancy way of saying when routers are being drowned淹没 by packets that they cannot handle, they just throw them away. (Electricity)
- Which packets to drop?
  - At random
  - Application: To drink wine (old is better than new) or milk (new is better than old)?
  - Compression: full frame or modification?
  - Priority: High or low? (Unless there is some significant incentive to mark packets as anything other than VERY IMPORTANT – NEVER NEVER DISCARD, nobody will do it.

Congestion Control Algorithms: Load shedding

- RED (Random Early Detection)
  - Why early?
    - Dealing with congestion after it is first detected is more effective than letting it gum up the works and then trying to deal with it.
    - Try to discard packets before all the buffer space is really exhausted.
  - Why random?
    - Since the router probably cannot tell which source is causing most of the trouble, picking a packet at random from the queue that triggered the action is probably as good as it can do.

5.4 QUALITY OF SERVICE (服务质量)

- Requirements of QoS
- Techniques for QoS
- Integrated Services
- Differentiated Services
- Label Switching and MPLS

Quality Of Service: Requirements

- A **flow** is a stream of packets from a source to a destination.
  - In a connection-oriented network, all the packets belonging to a flow follow the same route;
  - In a connection-less network, they may follow different routes.
- The **needs (or requirements) of each flow** can be characterized by four primary parameters:
  - **Reliability(可靠性),**
  - **Delay (延迟),**
  - **Jitter (抖动),**
  - **Bandwidth (带宽).**

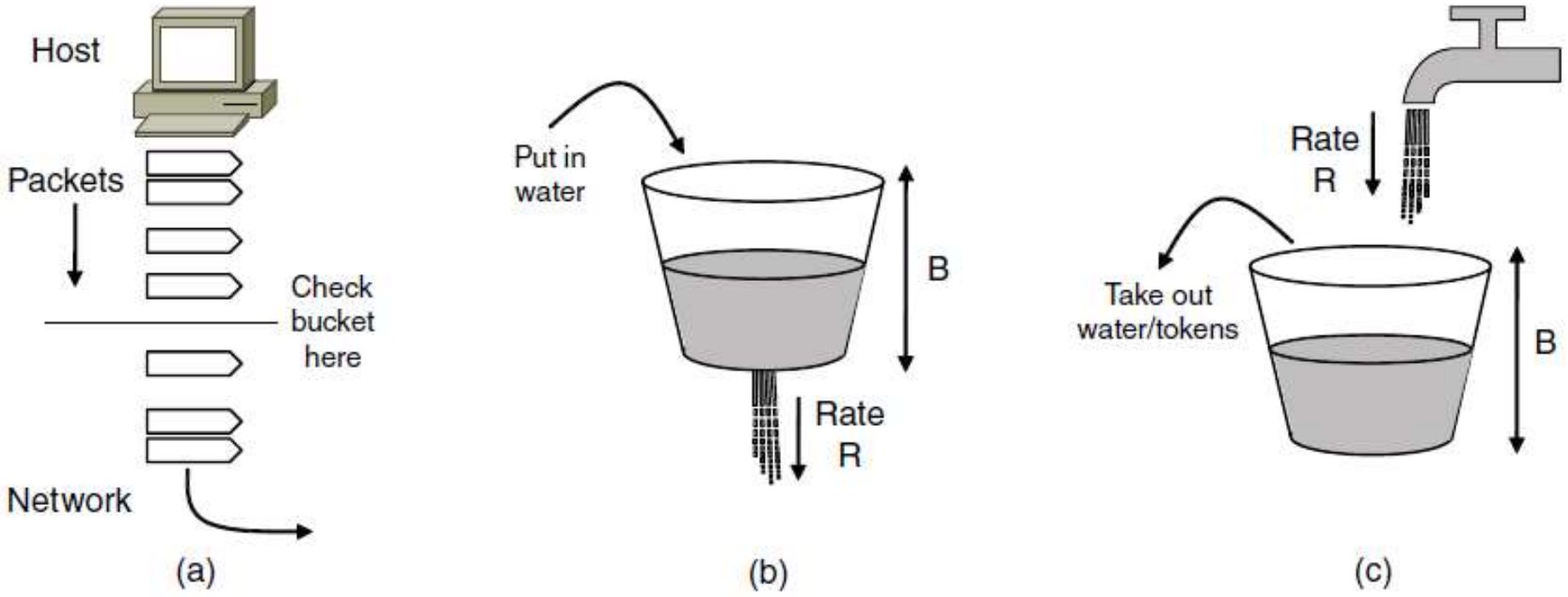
5.4.1 Quality Of Service: Requirements

How stringent the quality-of-service requirements are.

Application	Reliability	Delay	Jitter	Bandwidth
E-mail	High	Low	Low	Low
File transfer	High	Low	Low	Medium
Web access	High	Medium	Low	Medium
Remote login	High	Medium	Medium	Low
Audio on demand	Low	Low	High	Medium
Video on demand	Low	Low	High	High
Telephony	Low	High	High	Low
Videoconferencing	Low	High	High	High

5.4.2 Quality Of Service: Traffic shaping

(a) Shaping packets. (b) A leaky bucket. (c) A token bucket



- Leaky bucket algorithm
- Token bucket algorithm

Quality Of Service: Traffic shaping

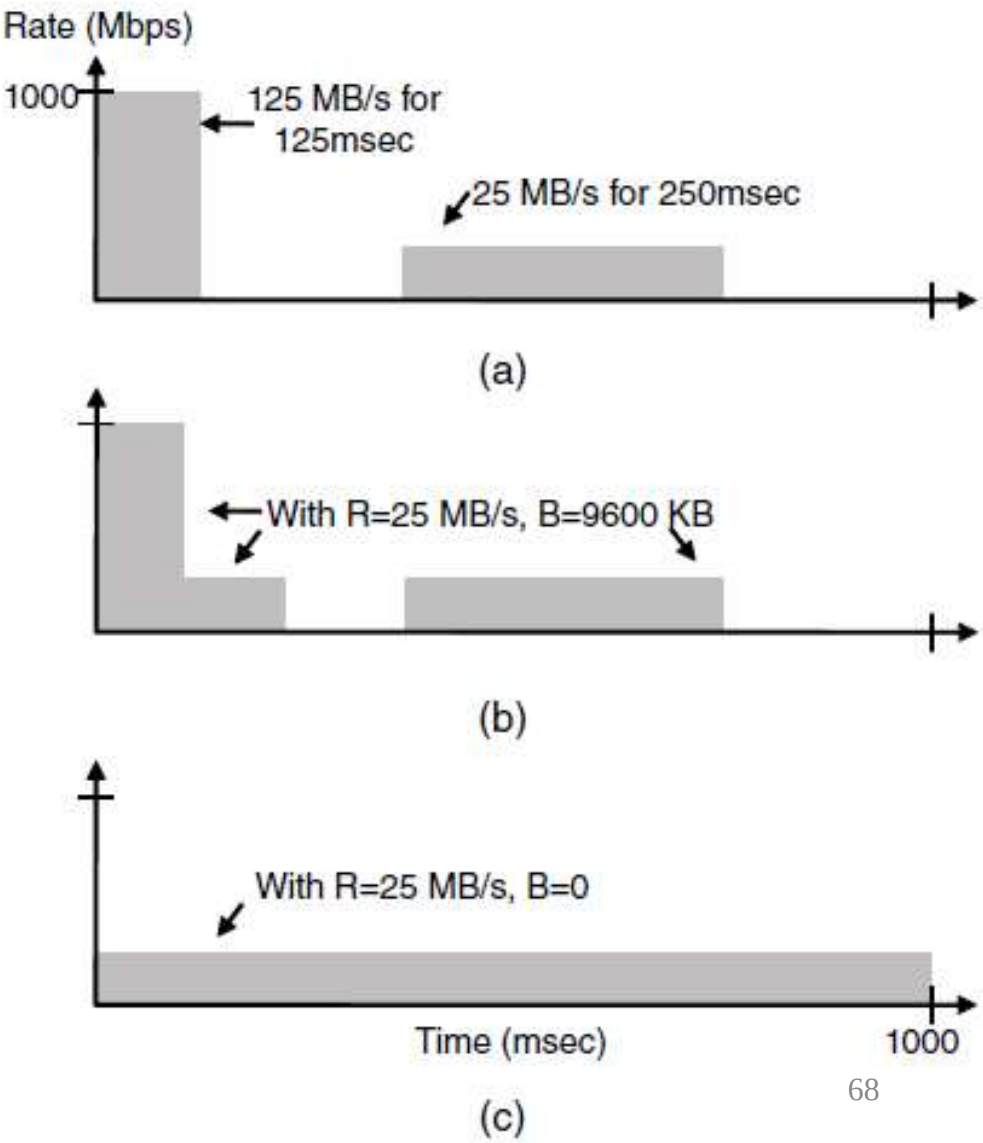
(a) Traffic from a host.

Output shaped by a token bucket of rate 200 Mbps and capacity

(b) 9600 KB,

(c) 0 KB.

Assume： The tap is running at rate R and the bucket has a capacity of B



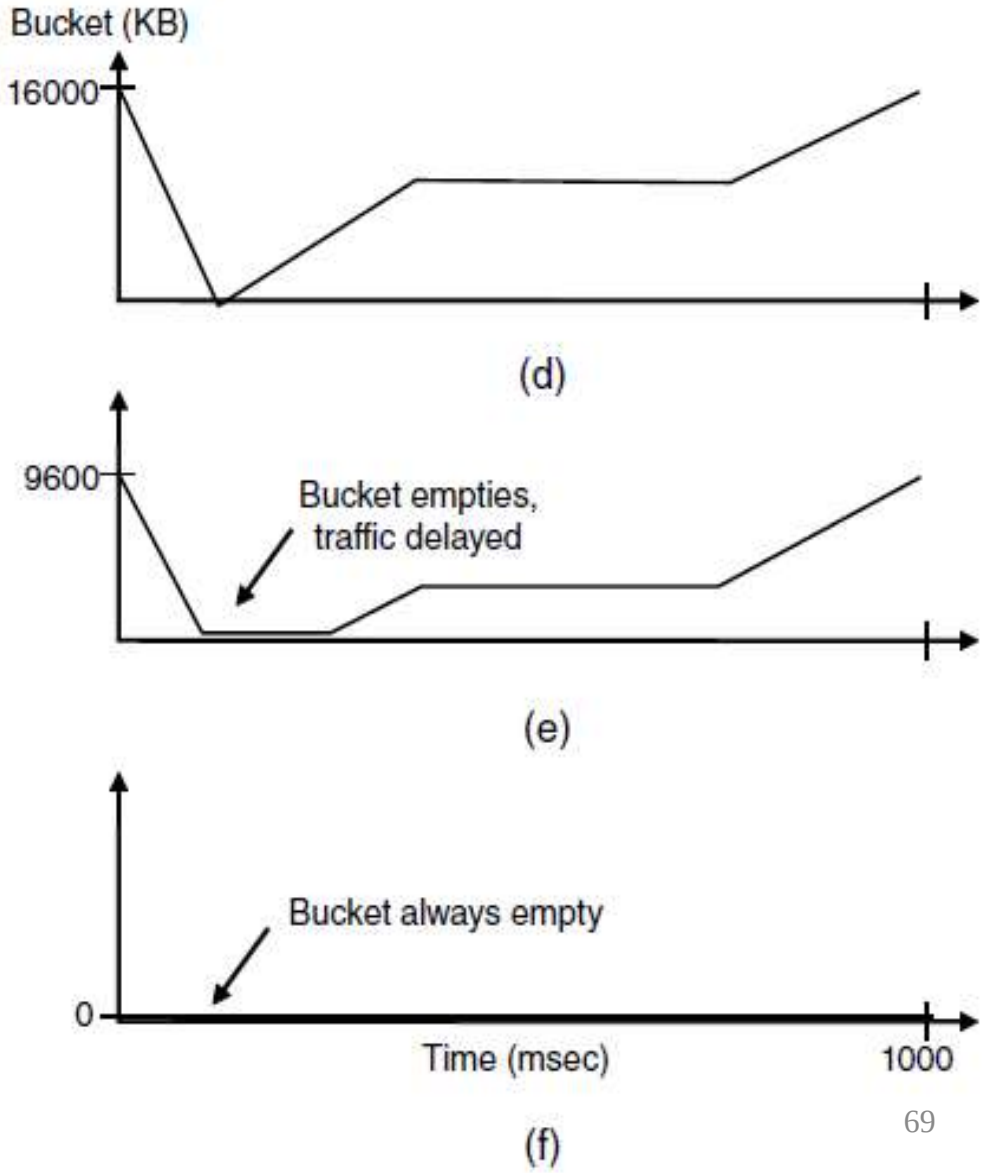
Quality Of Service: Traffic shaping

Token bucket level for shaping with rate 200 Mbps and capacity

(d) 16000 KB,

(e) 9600 KB, and

(f) 0KB..

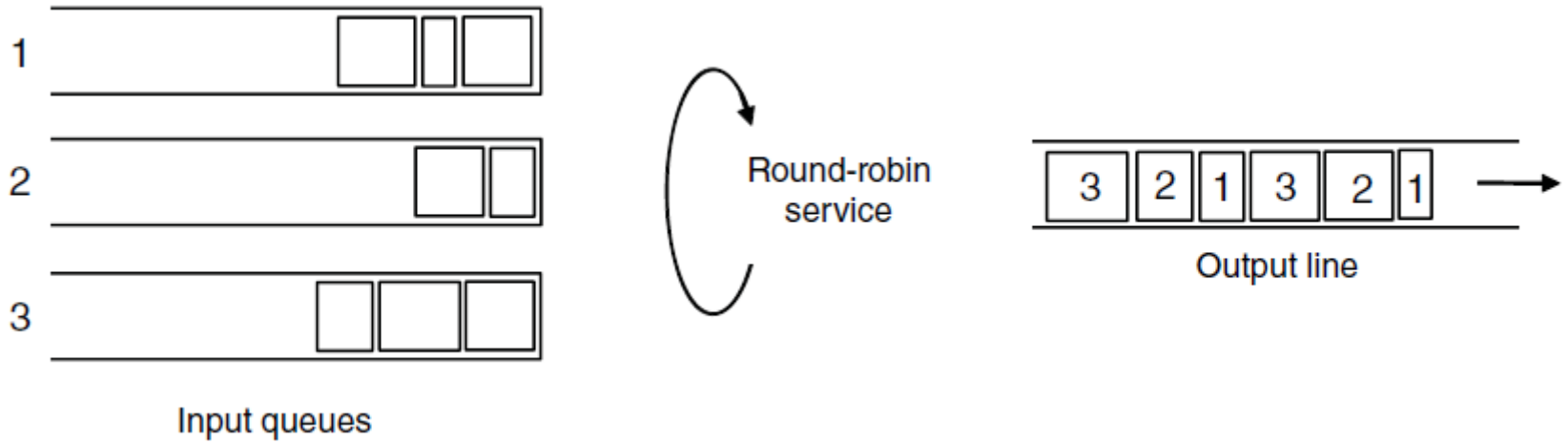


5.4.3 Quality Of Service: Packet Scheduling

- **Resource reservation**
  - Suppose there is a specific route for a flow, it becomes possible to reserve resources along the route to make sure the needed capacity is available.
- What resources to reserve?
  - Bandwidth
  - Buffer space
  - CPU cycles

Quality Of Service: Packet Scheduling

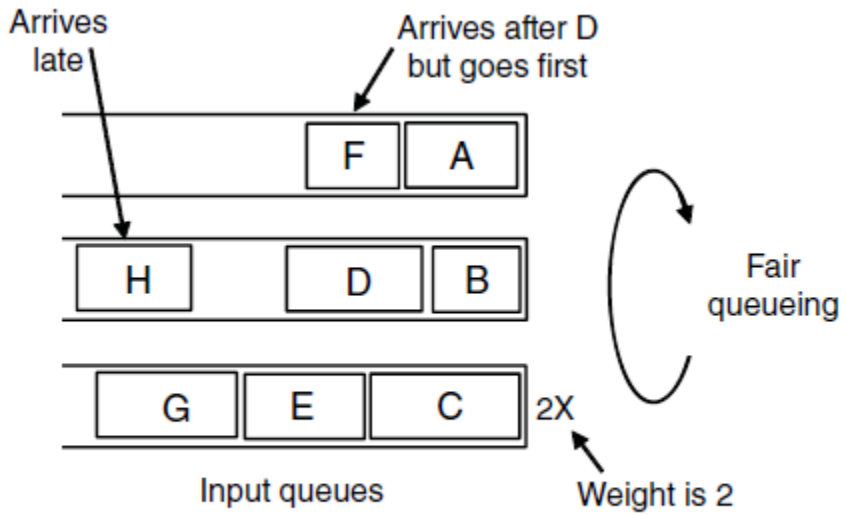
Round-robin Fair Queuing



Quality Of Service: Packet Scheduling

(a)Weighted Fair Queueing.

(b)Finishing times for the packets.



(a)

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

(b)

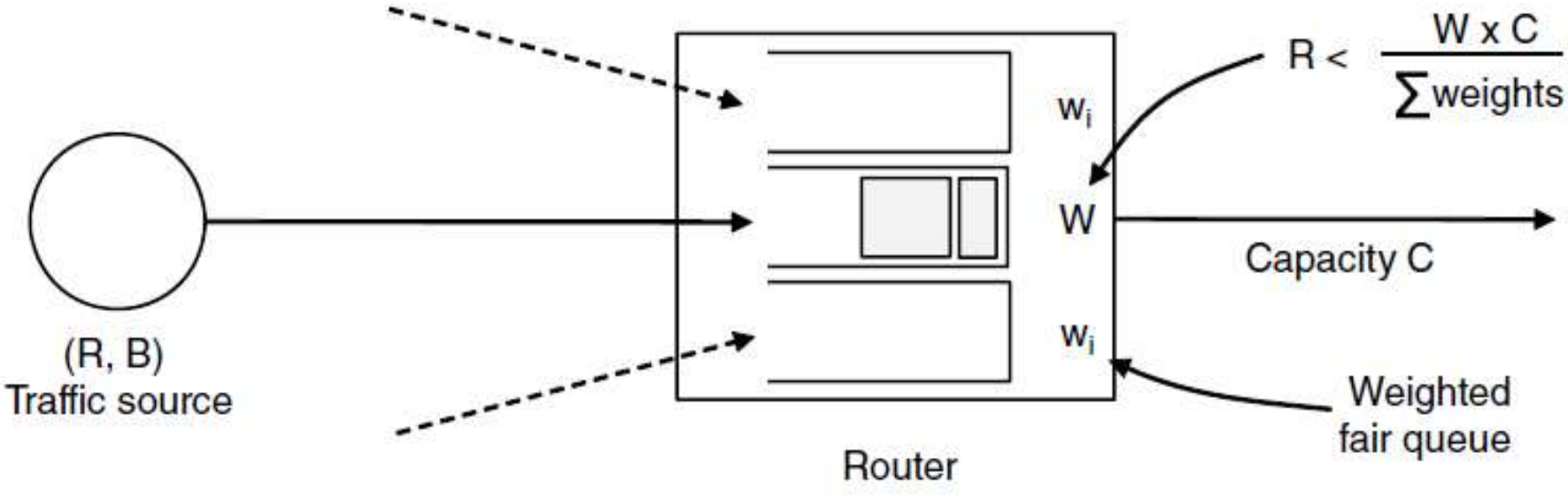
5.4.4 Quality Of Service: Admission Control

An example flow specification

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

Quality Of Service: Admission Control

Bandwidth and delay guarantees with token buckets and WFQ(Weighted Fair Queueing).

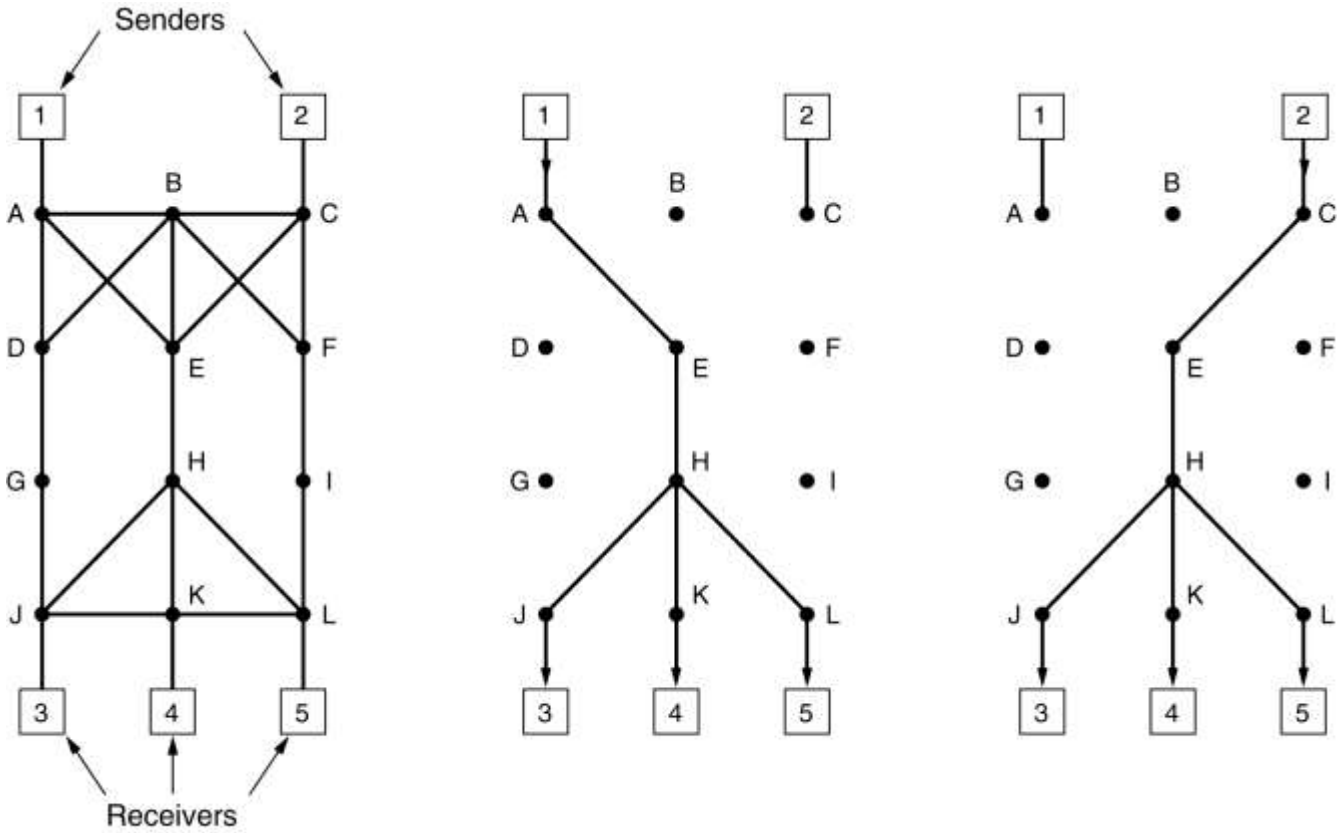


Quality Of Service: Integrated services

- How to stream multimedia?
  - Integrated services (Flow-based algorithms)
  - Differential services (Class-based algorithms)
- How about having the senders reserve bandwidth in advance? Too many destinations
- RSVP (Resource reSerVation Protocol资源预留协议): to allow multiple senders to transmit to multiple groups of receivers, permits individual receivers to switch channels freely, and optimizes bandwidth use while at the same time eliminating congestion.

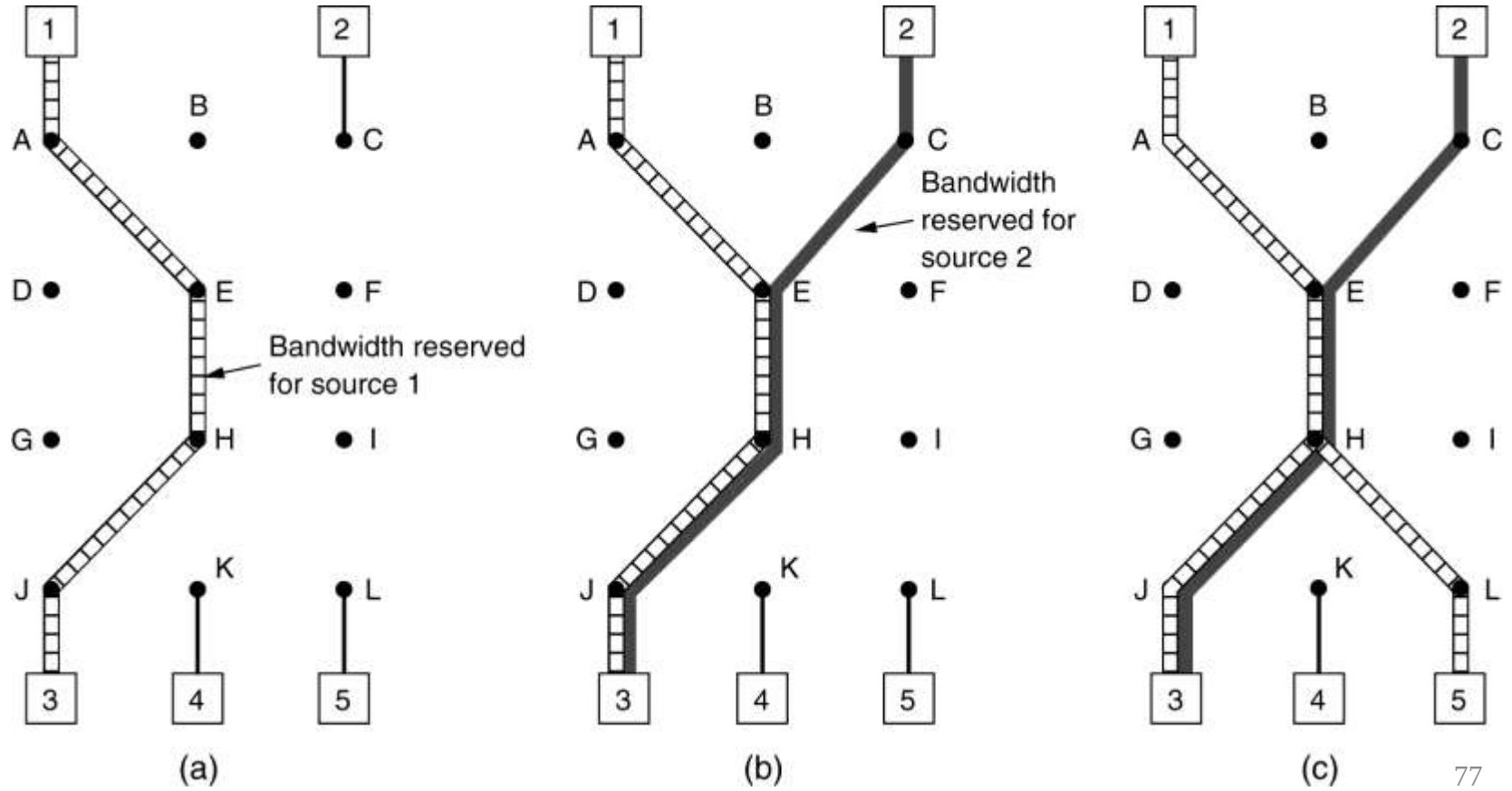
Quality Of Service: Integrated services

(a) A network, (b) The multicast spanning tree for host 1.
(c) The multicast spanning tree for host 2.



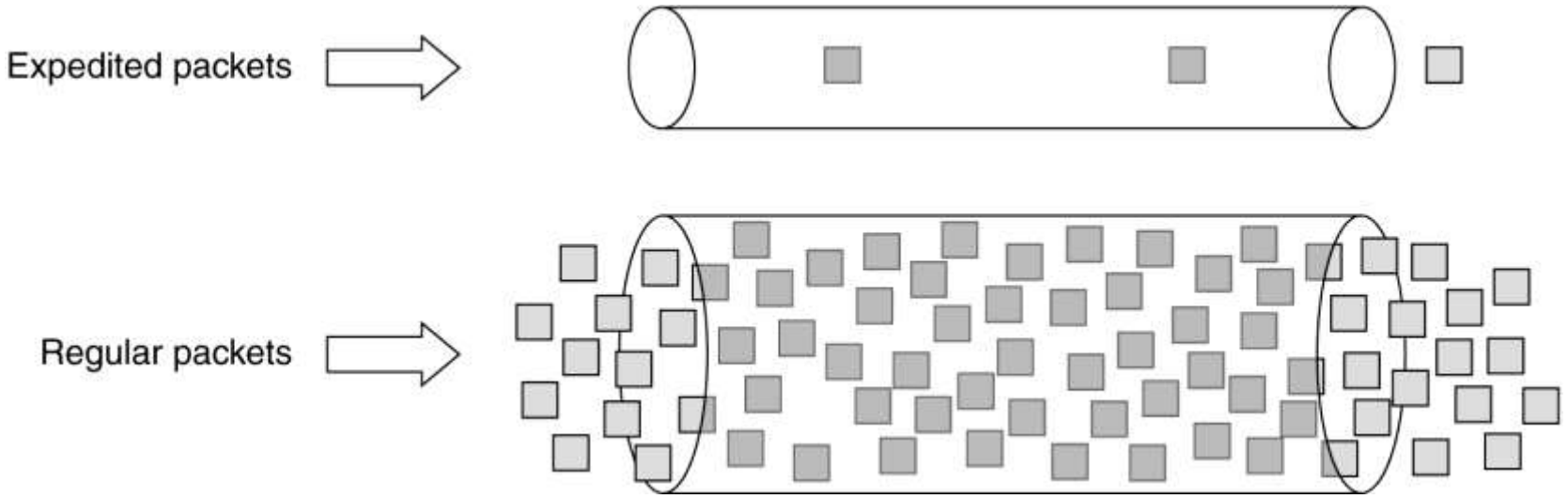
5.4.5 Quality Of Service: Integrated services

(a) Host 3 requests a channel to host 1. (b) Host 3 then requests a second channel, to host 2. (c) Host 5 requests a channel to host 1.



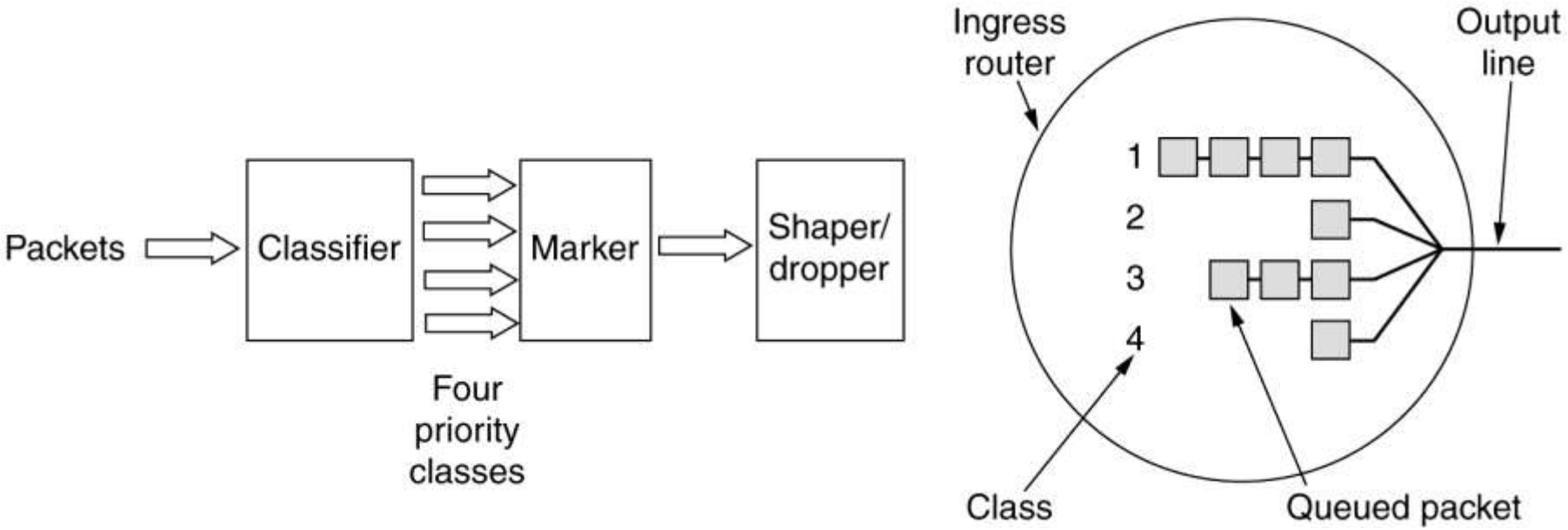
5.4.6 Quality Of Service: Differential services

Expedited (畅通的, 迅速的) packets experience a traffic-free network. (RFC 3246)



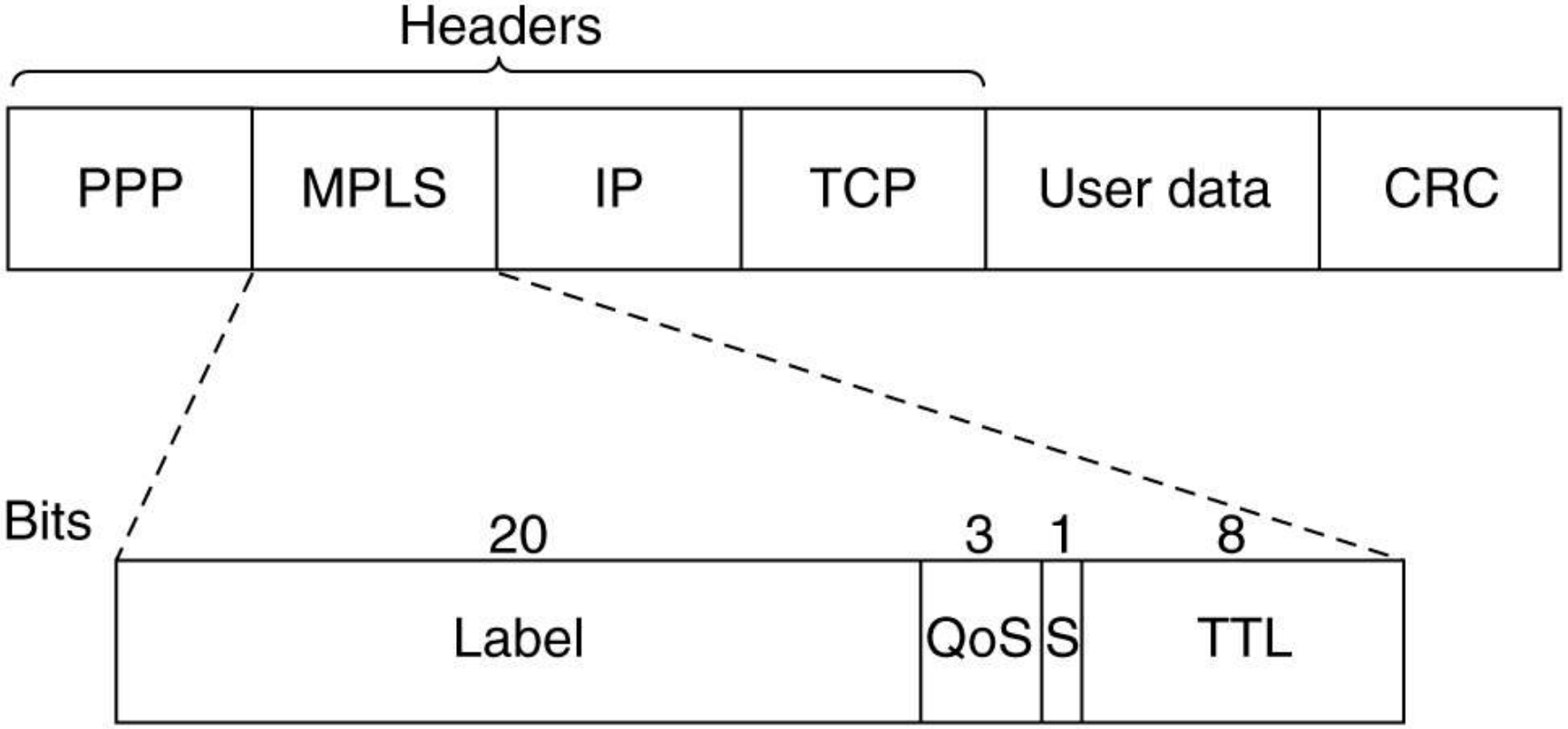
Quality Of Service: Differential services

- Assured Forwarding (确定推进) (RFC2597)

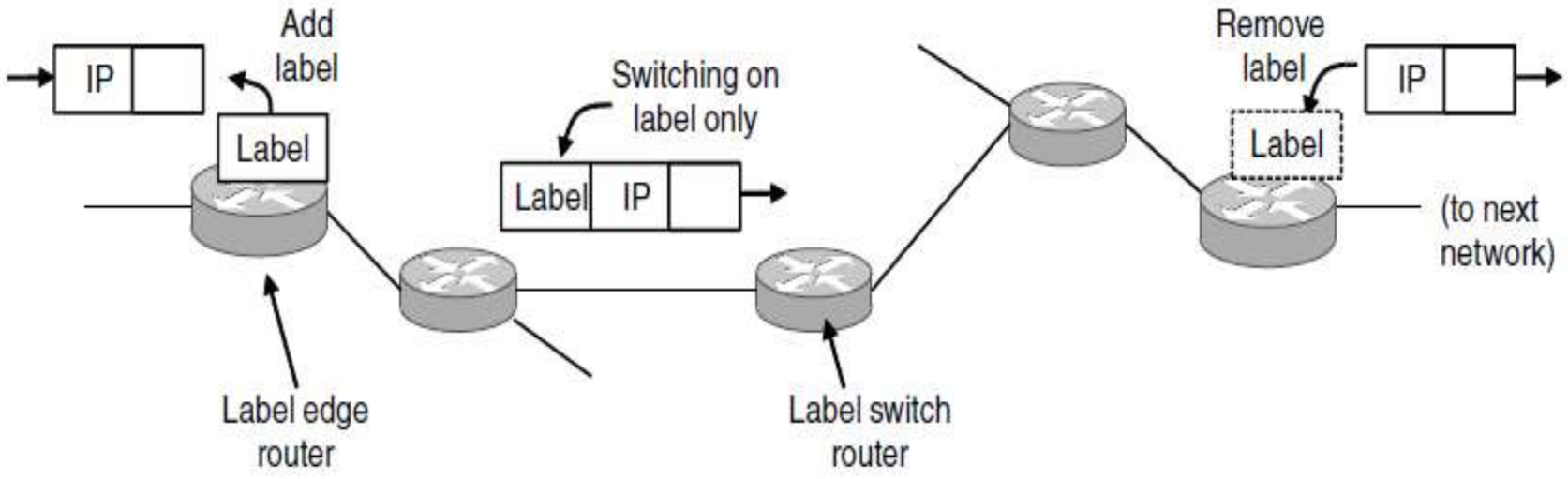


5.4.7 Quality Of Service: Label switching and MPLS

Transmitting a TCP segment using IP, MPLS, and PPP.



Quality Of Service:
Label switching and MPLS



Forwarding an IP packet through an MPLS network

5.5 INTERNETWORKING (网络互连)

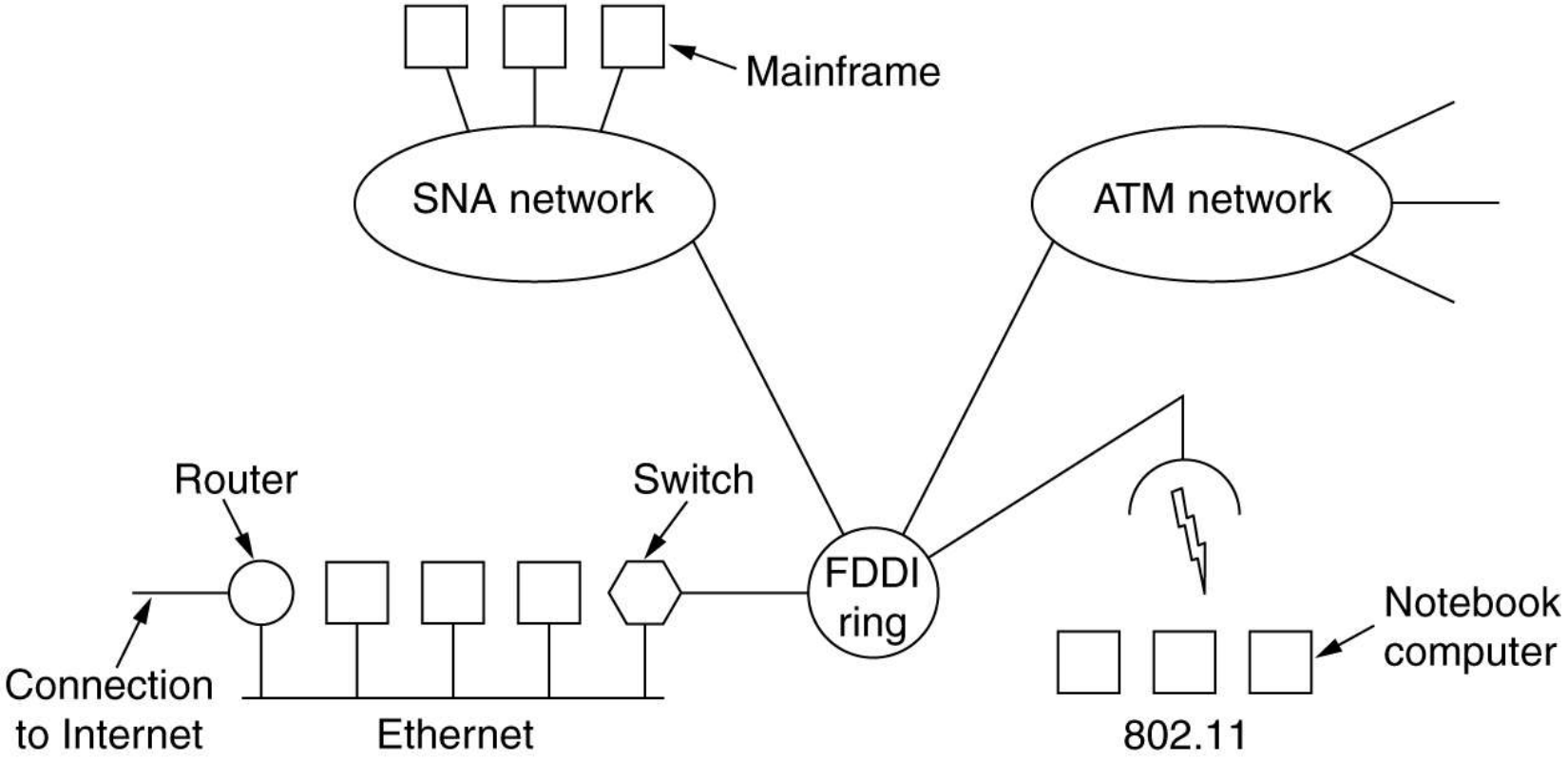
- How Networks Differ
- How Networks Can Be Connected
- Tunneling
- Internetwork Routing
- Packet Fragmentation

Internetworking: Introduction

- A variety of different networks (and thus protocols) will be around
  - The installed base of different networks is large and growing.
  - As computers and networks get cheaper, the place where decisions get made moves downward.
  - Different networks have radically different technology, so it should not be surprising that as new hardware developments occur, new software will be created to fit the new hardware.

Internetworking: Introduction

A collection of interconnected networks.



Internetworking: Introduction

- The interconnection of different networks
  - LAN-LAN: A computer scientist downloading a file to engineering.
  - LAN-WAN: A computer scientist sending mail to a distant physicist.
  - WAN-WAN: Two poets exchanging sonnets.
  - LAN-WAN-LAN: Engineers at different universities communicating.

5.5.1 Internetworking: How Networks Differ

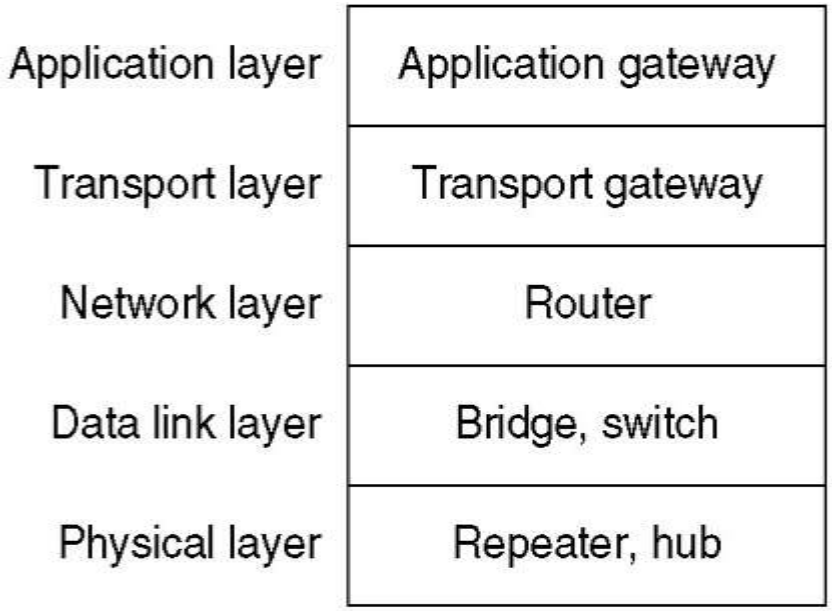
Some of the many ways networks can differ

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

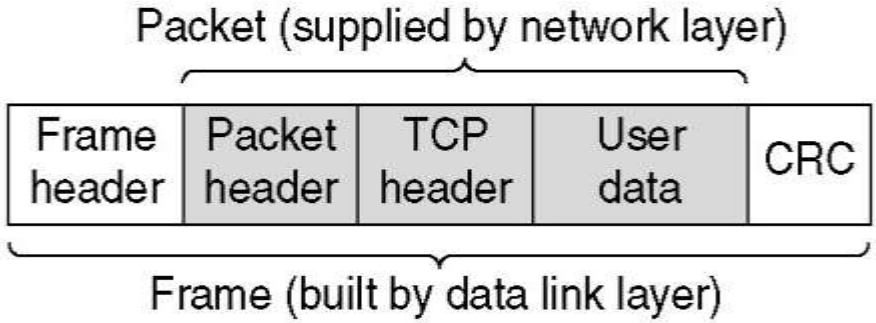
5.5.2 Internetworking: How Networks Can Be Connected

(a) Which device is in which layer.

(b) Frames, packets, and headers.



(a)

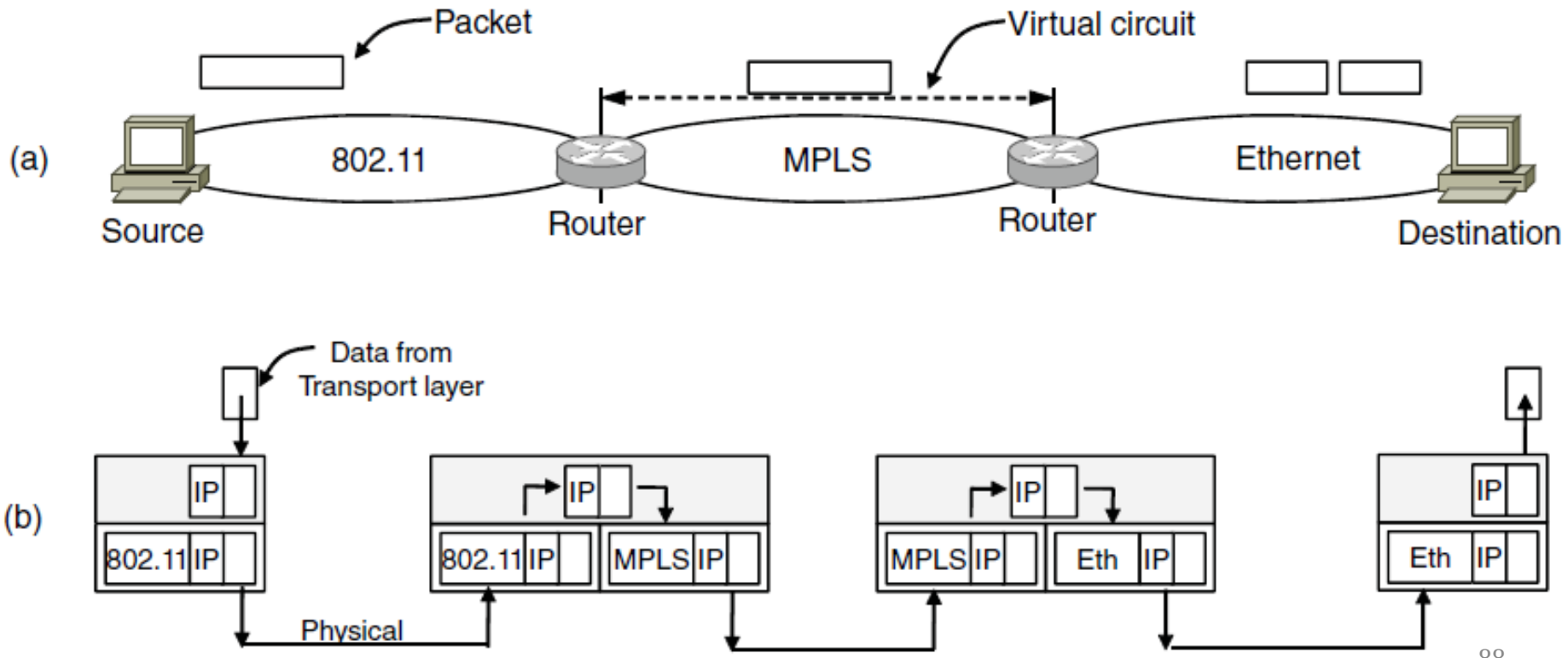


(b)

Internetworking:
How Networks Can Be Connected

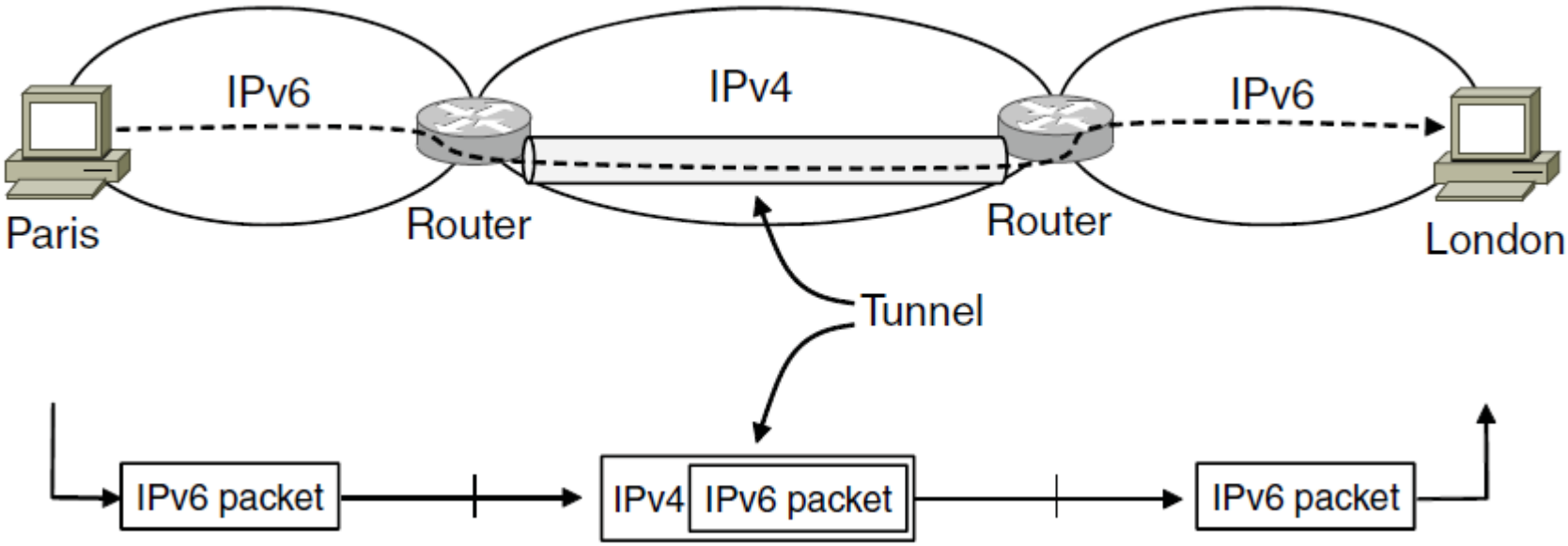
(a)A packet crossing different networks.

(b)Network and link layer protocol processing.



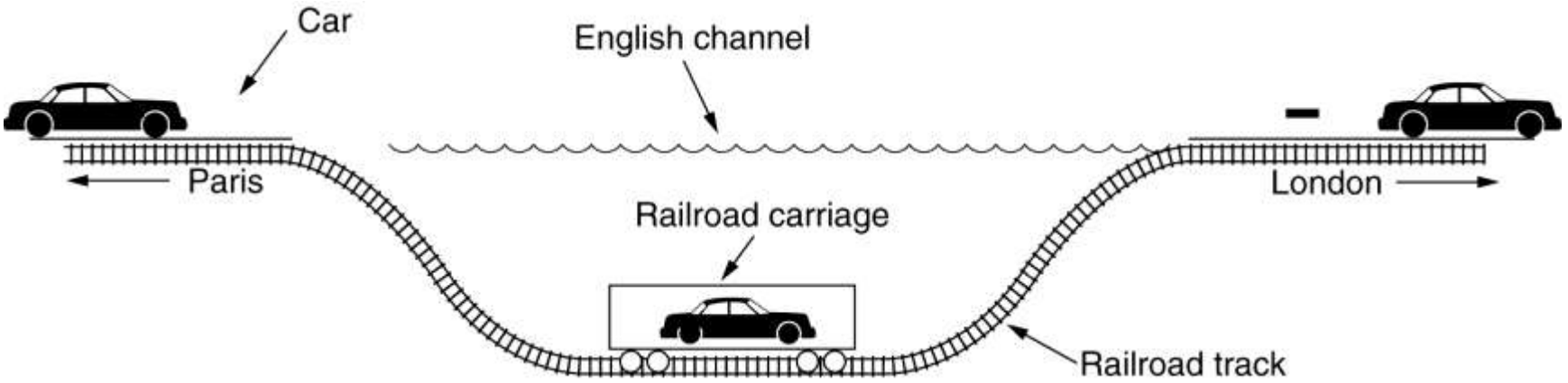
5.5.3 Internetworking: Tunneling (隧道)

Tunneling a packet from Paris to London.



Internetworking: Tunneling

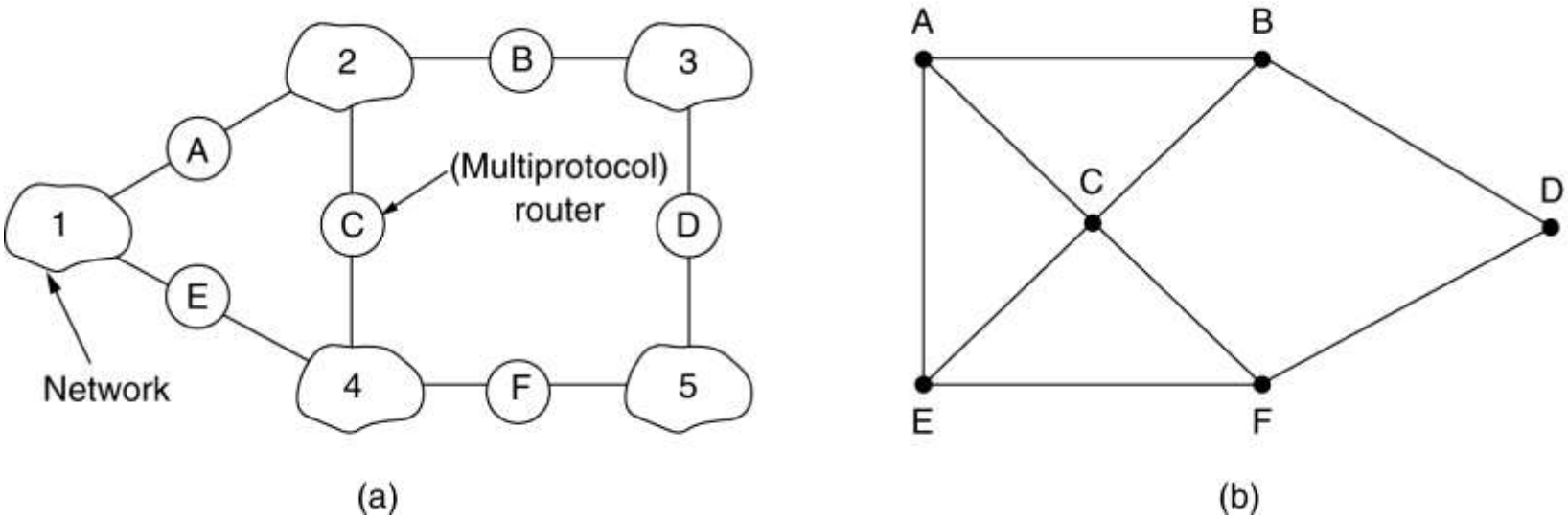
Tunneling a car from France to England.



5.5.4 Internetworking: Internetwork Routing

(a) An internetwork.

(b) A graph of the internetwork.



Internetworking: Internetwork Routing

- Two-level routing
  - Internet routing: Exterior gateway protocol
  - Intranet routing: Interior gateway protocol
- To route a packet
  - A typical internet packet starts out on its LAN addressed to the local multiprotocol router.
  - After it gets there, the network layer code decides which multiprotocol router to forward the packet to, using its own routing tables.
    - Direct forwarding using native network protocol
    - Tunneling using the intervening network protocol
  - This process repeats until the packet reaches the desination network.
- AS and BGP

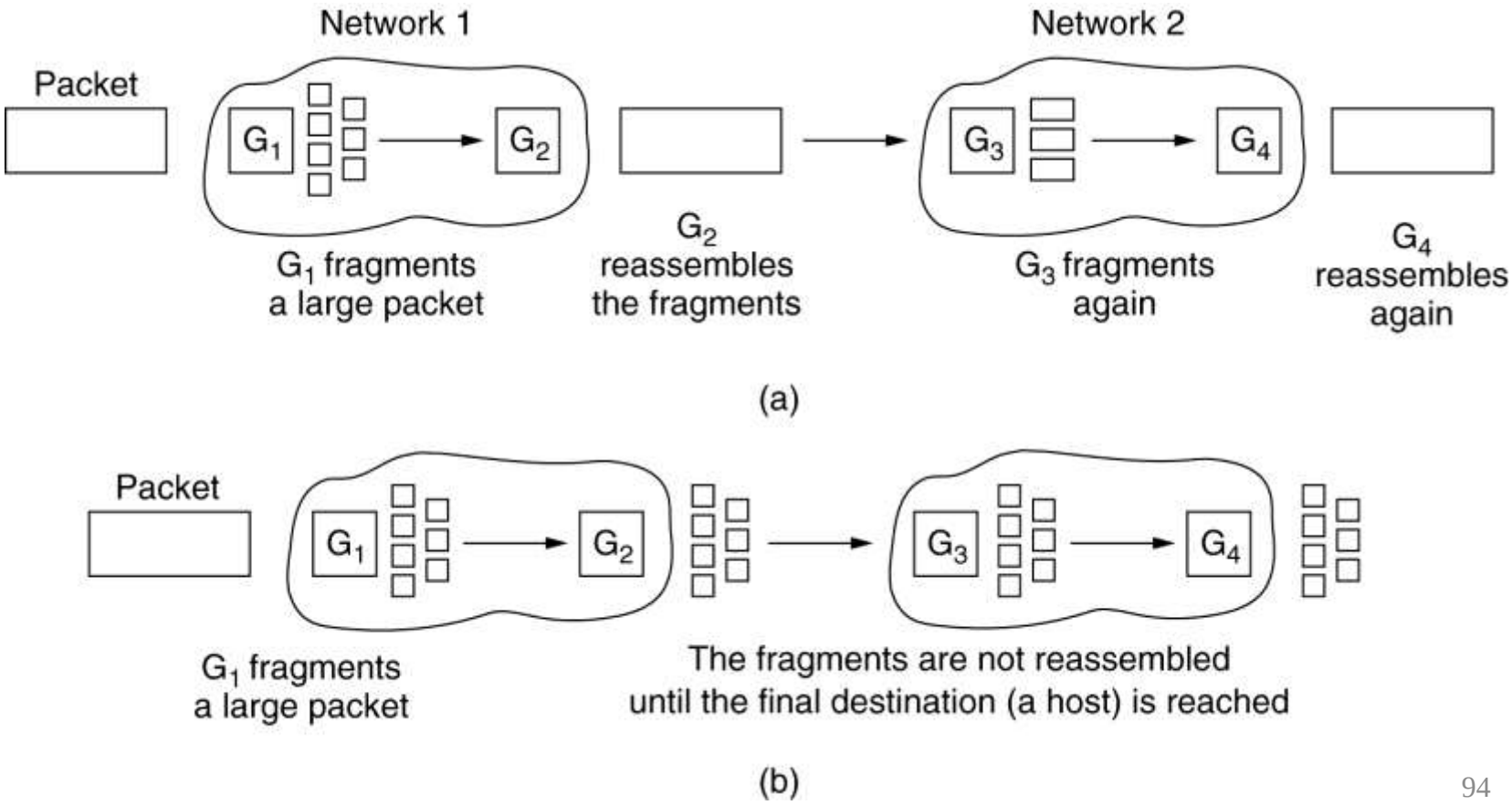
5.5.5 Internetworking: Fragmentation

- Each network imposes some maximum size on its packets. These limits have various causes, among them:
  - Hardware (e.g., the width of a TDM transmission slot).
  - Operating system (e.g., all buffers are 512 bytes).
  - Protocols (e.g., the number of bits in the packet length field).
  - Compliance with some (inter)national standard.
  - Desire to reduce error induced retransmissions to some level.
  - Desire to prevent one packet from occupying the channel too long.

Internetworking: Fragmentation

(a) Transparent fragmentation. (ATM)

(b) Nontransparent fragmentation. (IP)

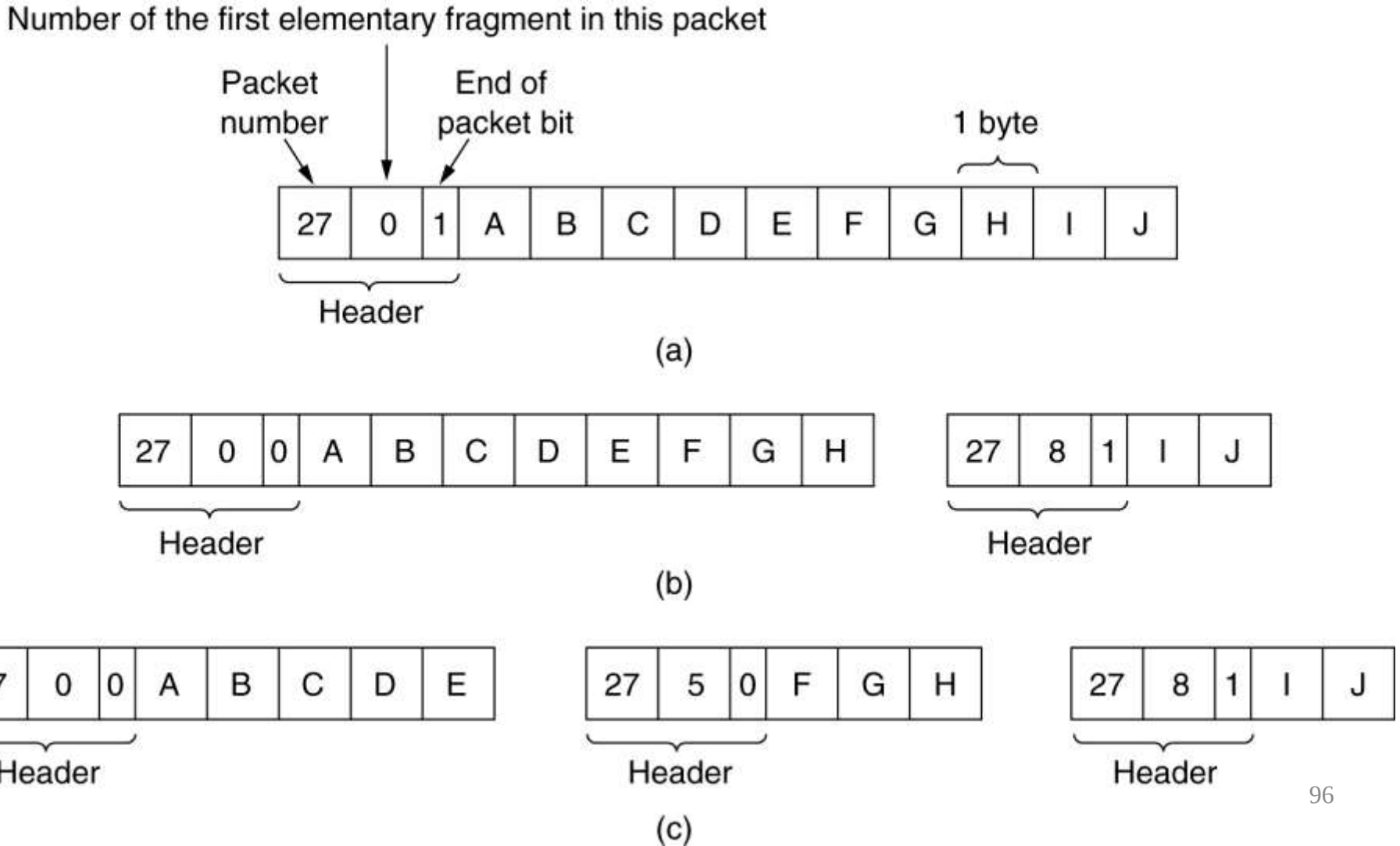


Internetworking: Fragmentation

- When a packet is fragmented, the fragments must be numbered in such a way that the original data stream can be reconstructed.
  - To define an elementary fragment size small enough that the elementary fragment can pass through every network.
  - When a packet is fragmented, all the pieces are multiple of the elementary fragment size.
  - The internet header provide
    - an original packet number and
    - the number of the (first) elementary fragment contained in the packet and
    - a bit indicating the last piece of the original packet.

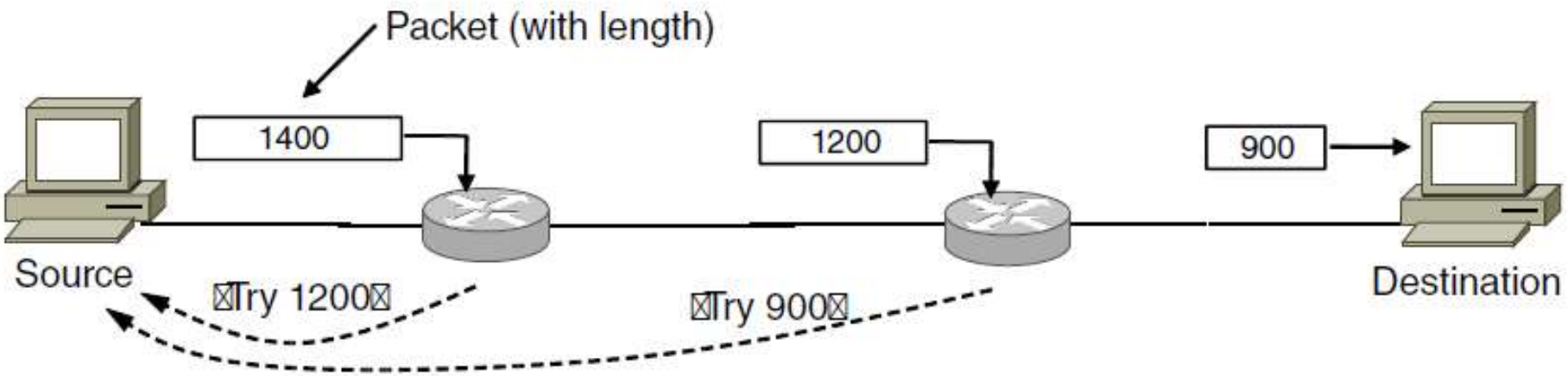
Internetworking: Fragmentation

Fragmentation when the elementary data size is 1 byte.



Internetworking: Fragmentation

Path MTU Discovery



5.6 THE NETWORK LAYER IN THE INTERNET

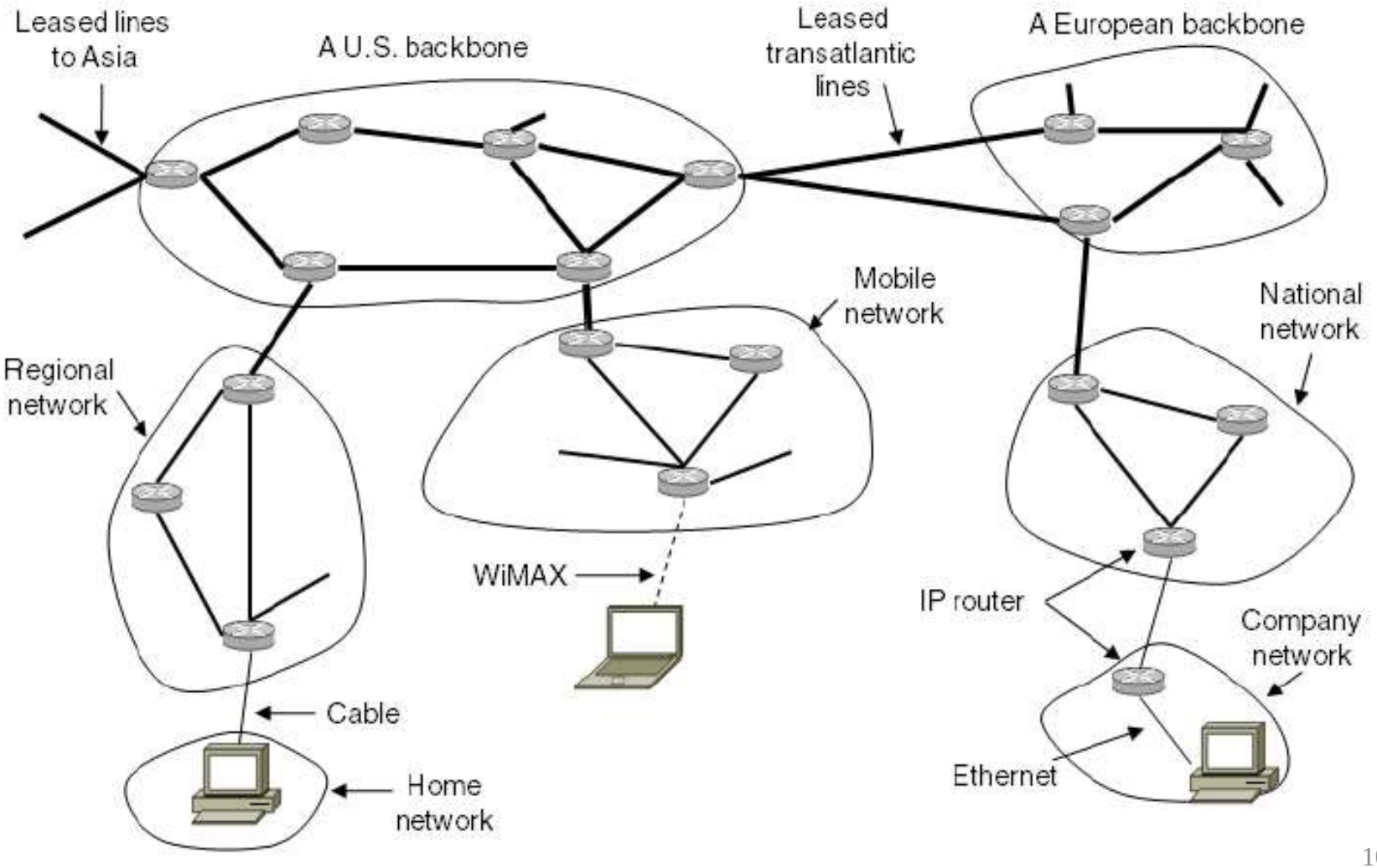
- The IPv4 Protocol
- IP Addresses
- The IPv6 Protocol
- Internet Control Protocols
- OSPF – The Interior Gateway Routing Protocol
- BGP – The Exterior Gateway Routing Protocol
- Internet Multicasting
- Mobile IP

The Network Layer in the Internet:
Top 10 principles for the Internet

1. Make sure it works.
2. Keep it simple.
3. Make clear choices.
4. Exploit modularity.
5. Expect heterogeneity.
6. Avoid static options and parameters.
7. Look for a good design; it need not be perfect.
8. Be strict when sending and tolerant when receiving.
9. Think about scalability.
10. Consider performance and cost.

The Network Layer in the Internet:

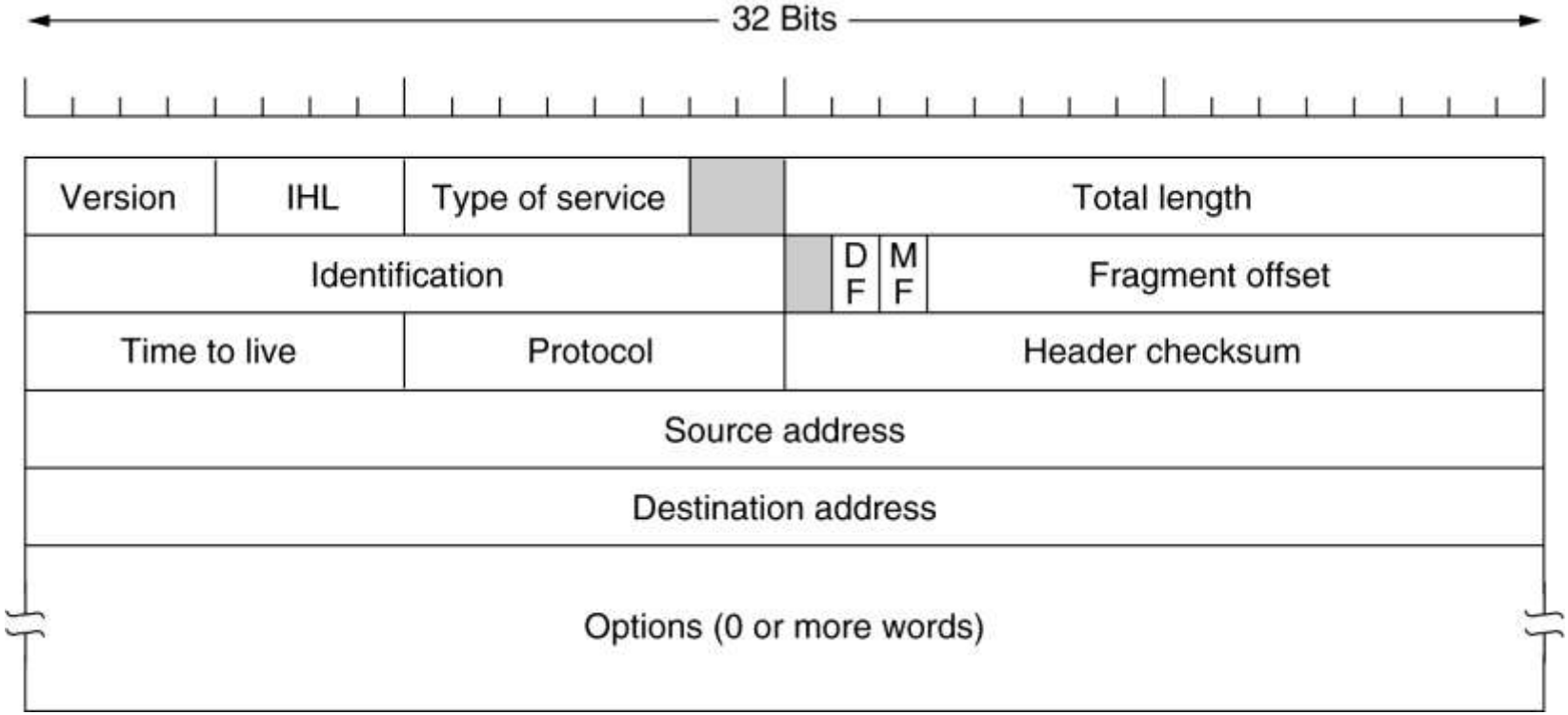
The Internet: Collections of Subnetworks or ASes



5.6.1 The Network Layer in the Internet: The IPv4 Protocol

An IP datagram consists of a header part and a text part.

The IPv4 (Internet Protocol) header.



The Network Layer in the Internet:
The IPv4 Protocol: Header fields

- ***Version***: to keep track of which version of the protocol the datagram belongs to.
- ***IHL:*** to tell how long the header is, in 32-bit words. 5 <= IHL <= 15.
- ***Differentiated service****:*
  - ***Type of service服务类型:*** 3 for priority, 3 for D, T, and R, and 2 unused.
  - ***Differentiated services区分服务***: Defined by IETF in 1998， 6 for service class, 2 for congestion.
- ***Total length***: the length of header and data. The maximum length is 65,535 bytes.
- ***Identification:*** the ID of the datagram.
- ***DF***: Don't Fragment 只有当DF=0时才允许分段
- (**最小MTU=576**，即假设 **512（传输层）+60（最大IP头）+4（冗余）**)
- ***MF:*** More Fragment (=1 for all fragments except last one
  =0 for last fragment)

The Network Layer in the Internet:
The IPv4 Protocol: Header fields

- ***Fragment offset***: to tell where in the current datagram this fragment belongs. All fragments except the last one in a datagram must be a multiple of 8 bytes, the elementary fragment unit.
- ***Time to Live***: a counter used to limit packet lifetimes.
- ***Protocol:*** which transport process to give this datagram to. (http://www.iana.org/assignments/protocol-numbers)
- ***Header checksum***: to verify the header only
- ***Source and destination address***: to indicate the network number and host number.
- ***Options***

The Network Layer in the Internet:
IP Header Options

Some of the IP options.

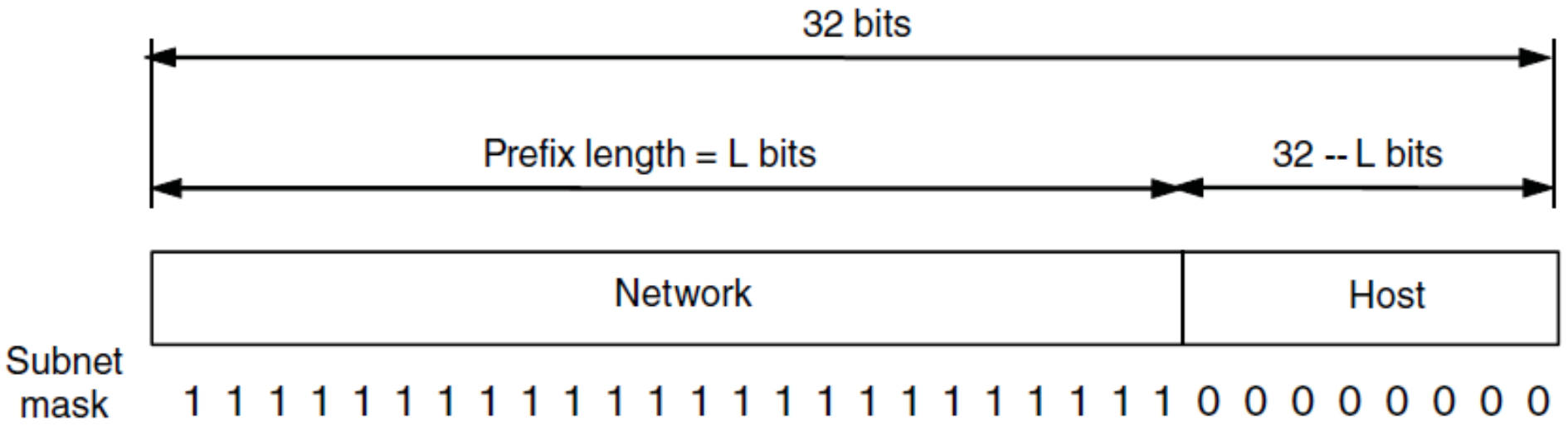
Option	Description
Security	Specifies how secret the datagram is
Strict source routing	Gives the complete path to be followed
Loose source routing	Gives a list of routers not to be missed
Record route	Makes each router append its IP address
Timestamp	Makes each router append its address and timestamp

5.6.2 The Network Layer in the Internet: IP Addresses

- Every Internet interface has an IP address, which encodes its network number and host number. The combination is unique: no two interfaces have the same IP address.
- All IP addresses are 32 bits long and are used in the source address and Destination address fields of IP packets
- IP addressing
  - Prefixes
  - Subnets (division),
  - CIDR(mergement),
  - Classful and Special Addressing
  - NAT

The Network Layer in the Internet:
IP Addresses: Prefixes

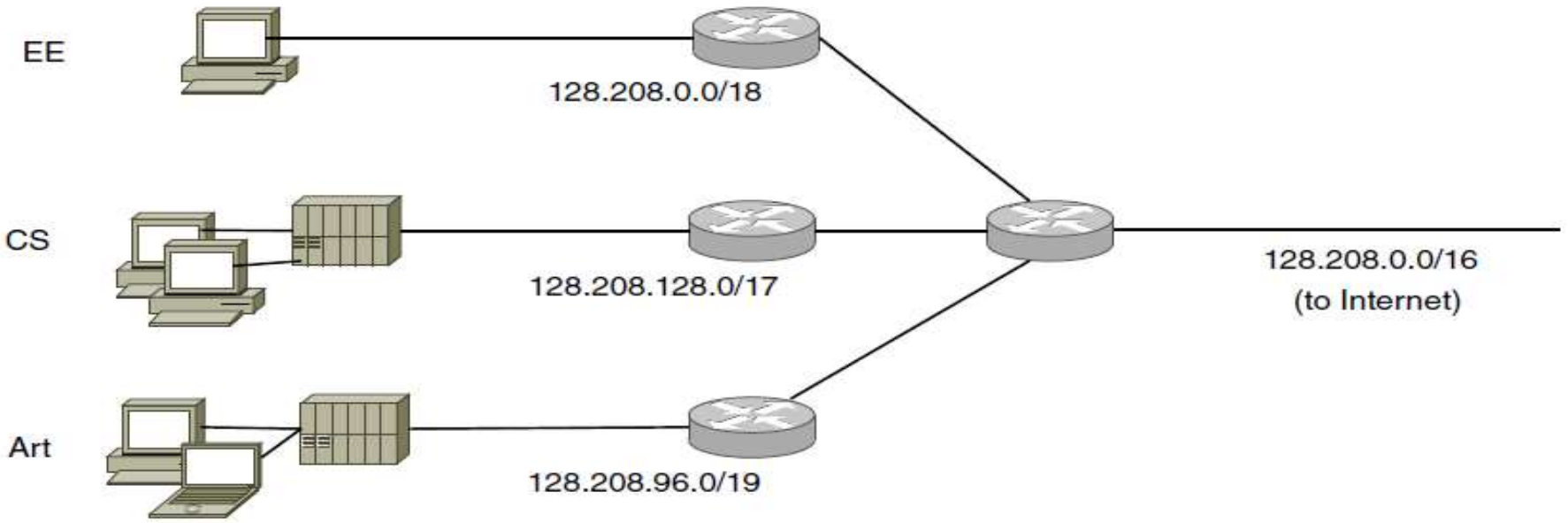
An IP prefix.



The Network Layer in the Internet:
IP Addresses: Prefixes

Splitting an IP prefix into separate networks with subnetting.

Computer Science:	10000000	11010000	1|xxxxxxx	xxxxxxxx
Electrical Eng.:	10000000	11010000	00|xxxxxx	xxxxxxxx
Art:	10000000	11010000	011|xxxxx	xxxxxxxx



The Network Layer in the Internet:
IP Addresses: CIDR

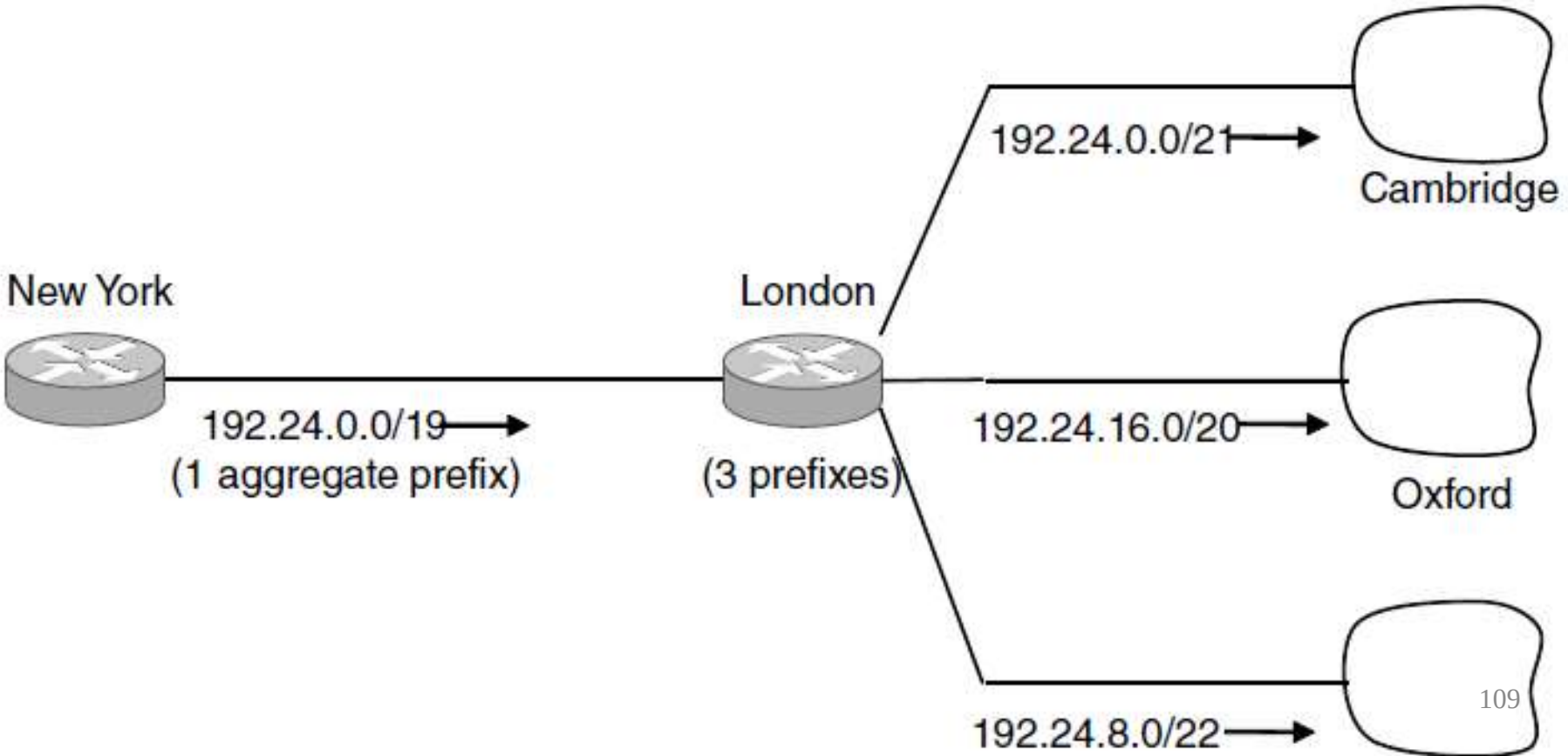
A set of IP address assignments

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	194.24.12/22
Oxford	194.24.16.0	194.24.31.255	4096	194.24.16.0/20

The Network Layer in the Internet:
IP Addresses: CIDR

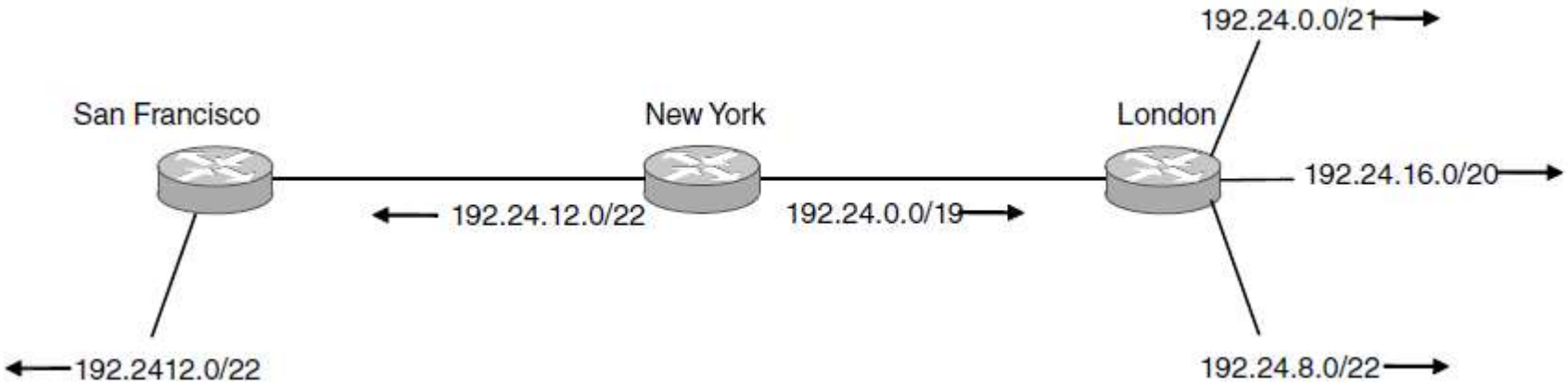
Aggregation of IP prefixes

supernet（超网）vs subnet



The Network Layer in the Internet:
IP Addresses: CIDR

Longest matching prefix routing at the New York router.



The Network Layer in the Internet: IP Addresses: CIDR

试题： A router has the following (CIDR) entries in its routing table:

Address/mask	Interface	Next hop
135.46.56.0/22	Interface 0	
135.46.60.0/22	Interface 1	
192.53.40.0/23	router 1	
default	router 2	

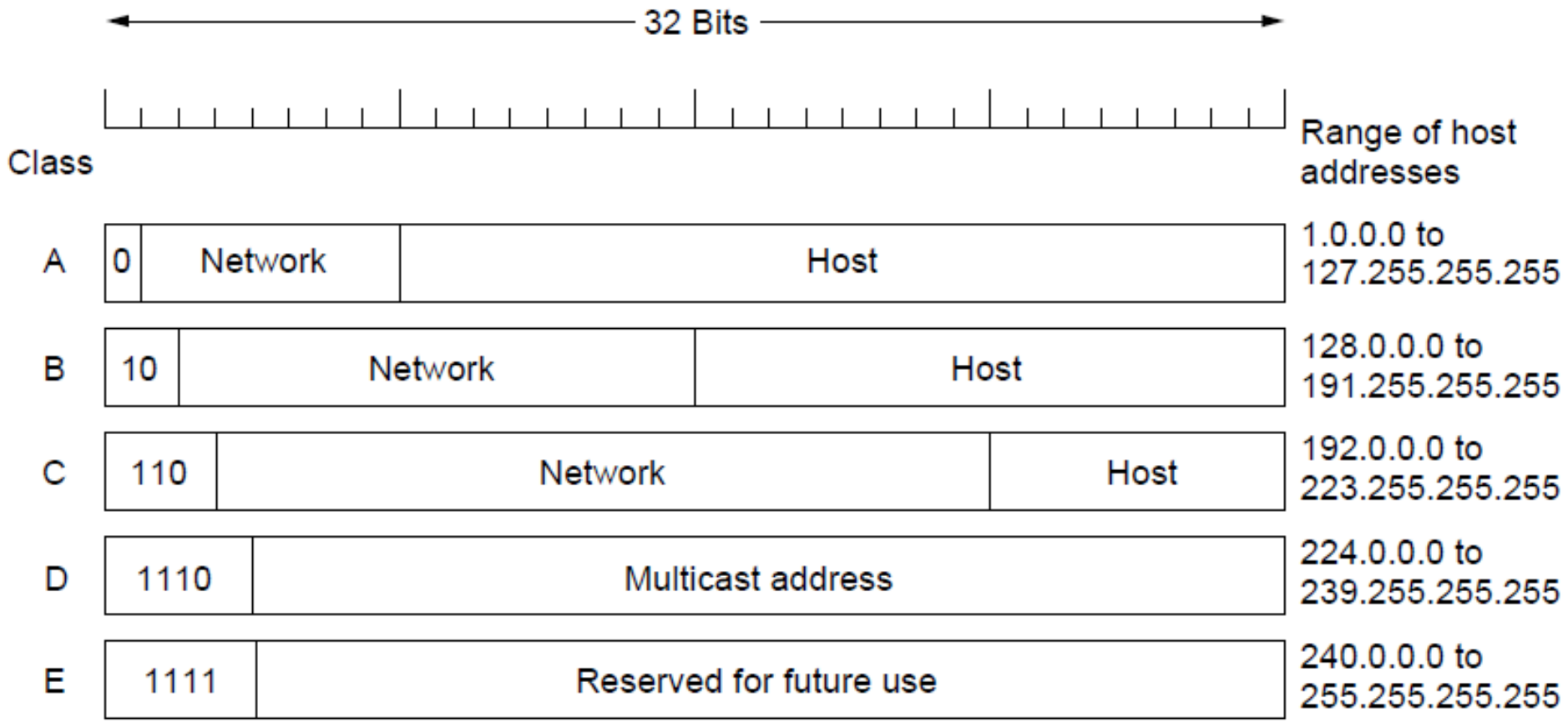
for each of the following IP addresses, which one does the router select for the next hop if a packet with that address arrives? *(5 bonus)*

(1) 135.46.63.10 (2) 135.46.57.14 (3) 135.46.52.2

(4) 192.53.40.7 (5) 192.53.56.7

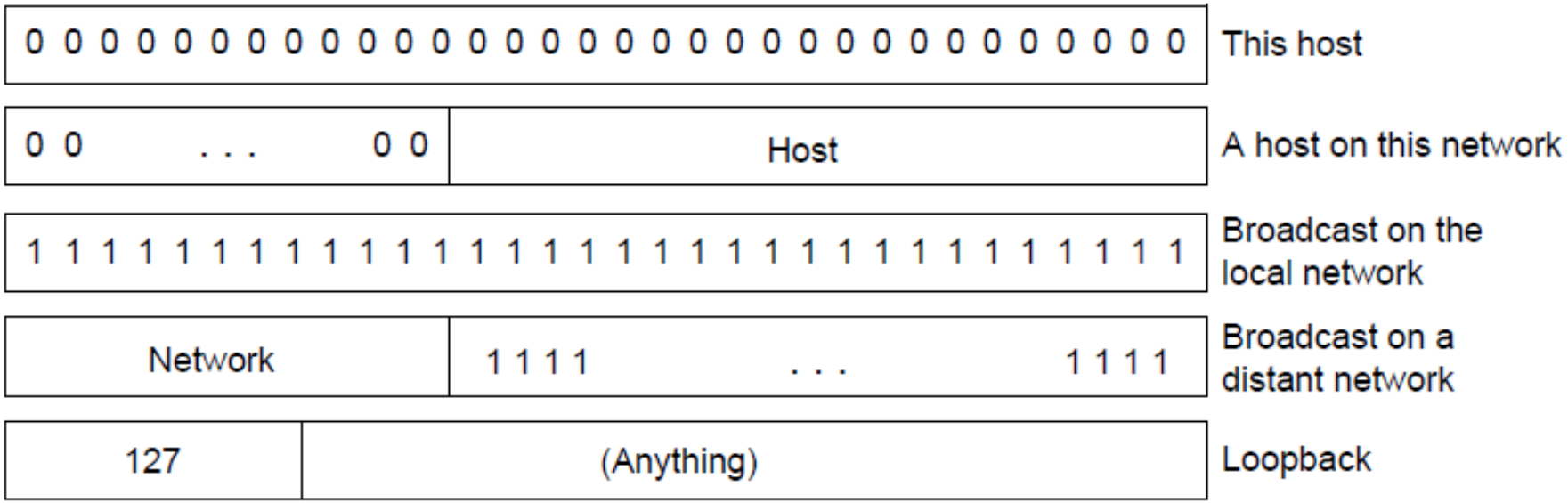
The Network Layer in the Internet:
IP Addresses: Classful and Special Addressing

IP address formats



The Network Layer in the Internet:
IP Addresses: Classful and Special Addressing

Special IP addresses



RFC 1918 指明的专用地址(private address)

- **10.0.0.0 到 10.255.255.255**
- **172.16.0.0 到 172.31.255.255**
- **192.168.0.0 到 192.168.255.255**
- 这些地址只能用于一个机构的内部通信，而不能用于和因特网上的主机通信。
- 专用地址只能用作本地地址而不能用作全球地址。在因特网中的所有路由器对目的地址是专用地址的数据报一律不进行转发。

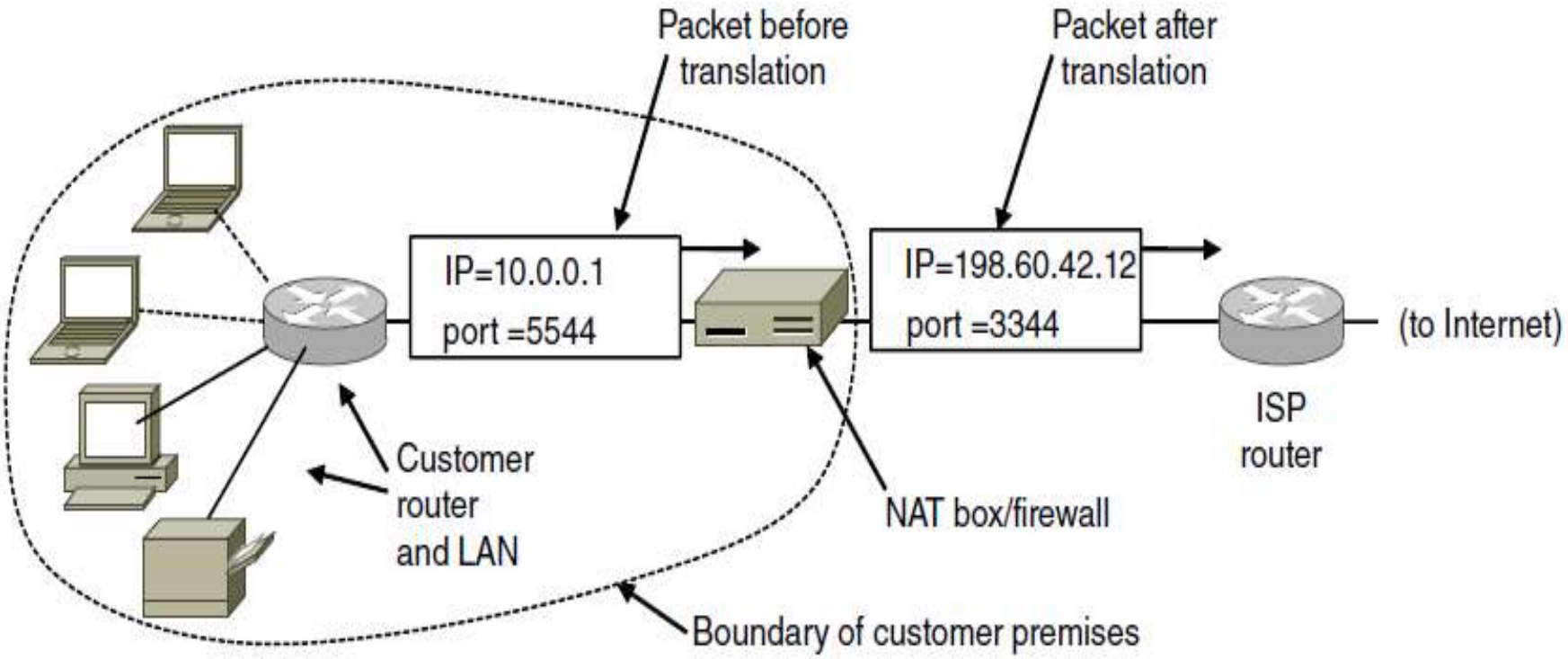
The Network Layer in the Internet:

IP Addresses: NAT(Network Address Translation)

- Addressing
  - Permanent addresses: too few
  - Temporary addresses (DHCP): large users
- Direct addressing (1-level) → indirect addressing (two-level)
- NAT (Network Address Translation)
  - To assign each company a single IP address (or at most, a small number of them) for Internet traffic
  - Within the company, every computer gets a unique IP address (10.x.x.x, 172.16.x.x, 192.168.x.x)

The Network Layer in the Internet:
IP Addresses: NAT

Placement and operation of a NAT box.



The Network Layer in the Internet:
IP Addresses: NAT

- How NAT works internally
  - (IP add_src, port_src) ←→ (IP add_dst, port_dst)
  - For outgoing packets
    - (10.0.0.1, 5544) → (198.60.42.12, 3344)
    - (198.60.42.12, 3344) → (210.32.32.32, 4433)

    - (198.60.42.12, 3344) ← (210.32.32.32, 4433)
    - (10.0.0.1, 5544) ← (198.60.42.12, 3344)

5.6.3 The Network Layer in the Internet: IPv6

Major goals for IPV6

- Support billions of hosts.
- Reduce the size of the routing tables.
- Simplify the protocol, to allow routers to process packets faster.
- Provide better security (authentication and privacy).
- Pay more attention to type of service.
- Aid multicasting by allowing scopes to be specified.
- Make it possible for a host to roam without changing its address.
- Allow the protocol to evolve in the future.
- Permit the old and new protocols to coexist for years

The Network Layer in the Internet: IPv6

- In 1990, IETF started work on a new version of IP.
- IETF issued a call for proposals and discussion in RFC 1550(IP: Next Generation (IPng) White Paper Solicitation).
- 21 in 1990 → 7 in 1992 → 3 in 1993 → 1
- IPv6 meets the goals fairly well.
  - IPv6 has longer addresses than IPv4. They are 16 bytes long.
  - The simplification of the header.
  - Better support for options.
  - Security
  - More attention has been paid to type of service than in the past.

IPv6能提供多少个IP地址

数量名称	科学记数法	零的数量
1000	10^3	1,000
100 万	10^6	1,000,000
10 亿	**10^9**	**1,000,000,000**
1 万亿	10^12	1,000,000,000,000
100 万的四次方	10^15	1,000,000,000,000,000
100 万的五次方	10^18	1,000,000,000,000,000,000
100 万的六次方	10^21	1,000,000,000,000,000,000,000
100 万的七次方	10^24	1,000,000,000,000,000,000,000,000
100 万的八次方	10^27	1,000,000,000,000,000,000,000,000,000
100 万的九次方	10^30	1,000,000,000,000,000,000,000,000,000,000
100 万的十次方	10^33	1,000,000,000,000,000,000,000,000,000,000,000
100 万的十一次方	**10^36**	**1,000,000,000,000,000,000,000,000,000,000,000,000**

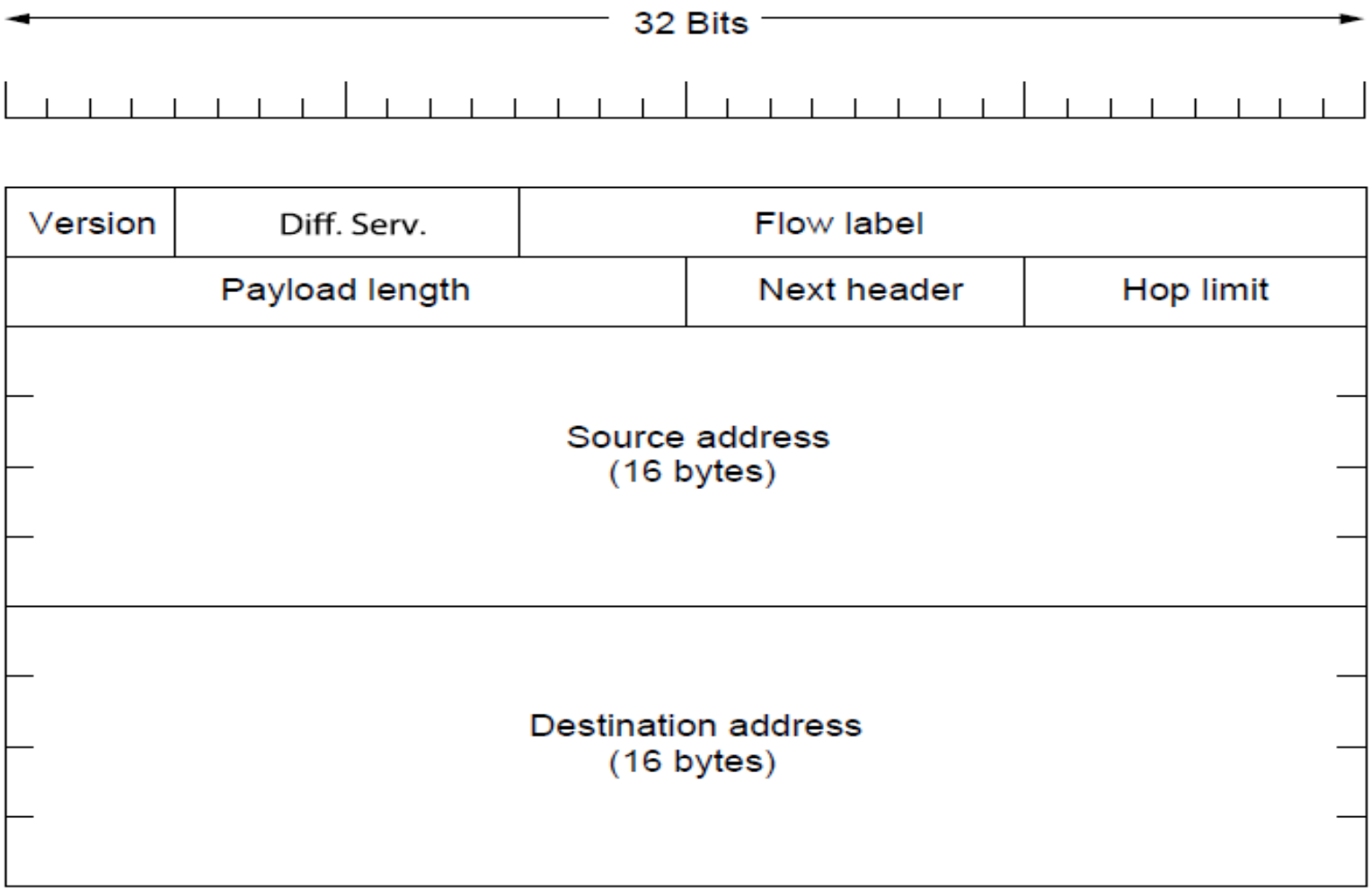
图例

有 40 亿个 IPv4 地址

有 340 涧个 IPv6 地址

The Network Layer in the Internet: IPv6

The IPv6 fixed header (required).



The Network Layer in the Internet: IPv6

IPv6 extension headers.

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

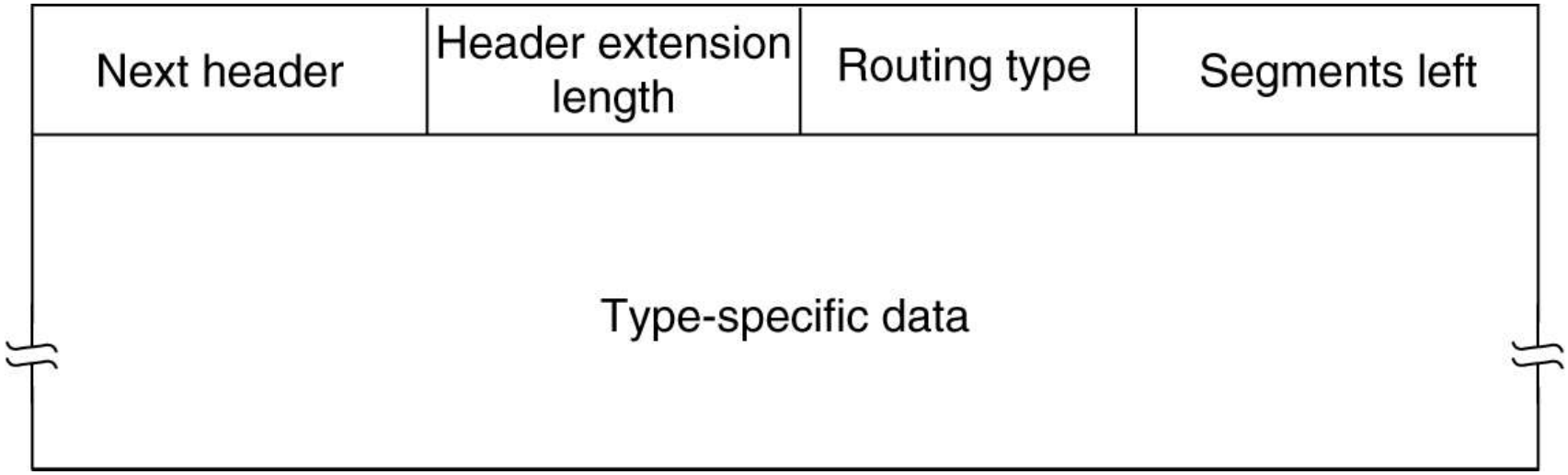
The Network Layer in the Internet: IPv6

The hop-by-hop extension header for large datagrams (jumbograms).

Next header	0	194	4
Jumbo payload length			

The Network Layer in the Internet: IPv6

IPv6 extension headers.



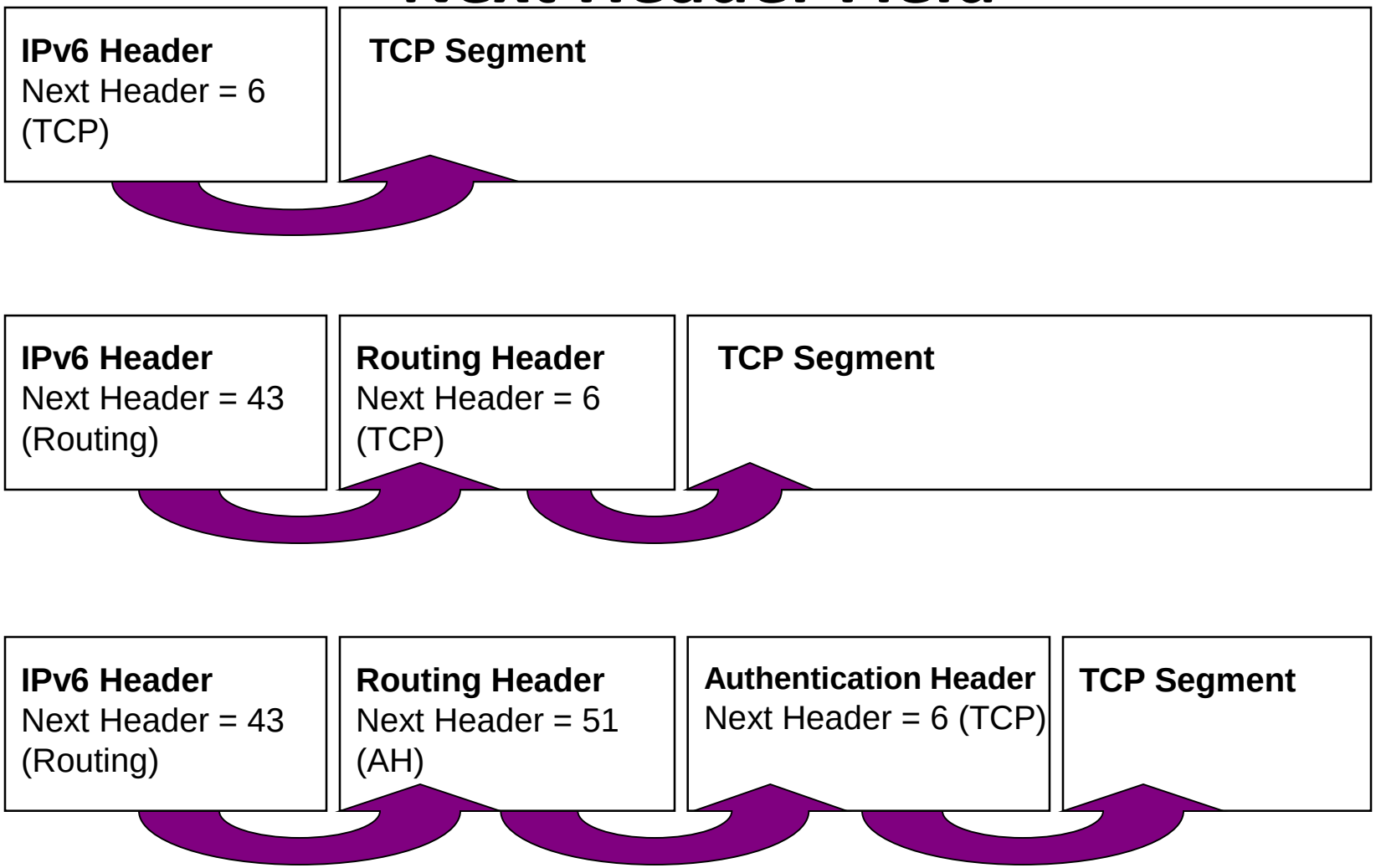
The Network Layer in the Internet: IPv6

Controversies

- Address length: 8 byte, 16 byte, 20 bytes → 16bytes
- Hop limit: 8 bits or more → 8 bits
- Maximum packet size: 64 KB or larger → normal 64kB and permit jumbograms.
- Checksum: needed or not → not needed any more.
- Mobile support: yes or no →no but.
- Security: yes or no → no but.

- Huitema, C. 1998. “IPv6: The New Internet Protocol” Prentice-Hall.

The Chain of Pointers Formed by the Next Header Field



冒号十六进制记法
(colon hexadecimal notation)

- 每个 16 位的值用十六进制值表示，各值之间用冒号分隔
  68E6:8C64:FFFF:FFFF:0:1180:960A:FFFF
- 零压缩(zero compression)，即一连串连续的零可以为一对冒号所取代。
- FF05:0:0:0:0:0:0:B3　可以写成：
- FF05::B3

点分十进制记法的后缀

- 0:0:0:0:0:0:128.10.2.1
  再使用零压缩即可得出： ::128.10.2.1
- CIDR 的斜线表示法仍然可用。
- 60 位的前缀 12AB00000000CD3 可记为：
  12AB:0000:0000:CD30:0000:0000:0000:0000/60
  或12AB::CD30:0:0:0:0/60
  或12AB:0:0:CD30::/60

5.6.4 The Network Layer in the Internet:

Internet Control Protocols: ICMP

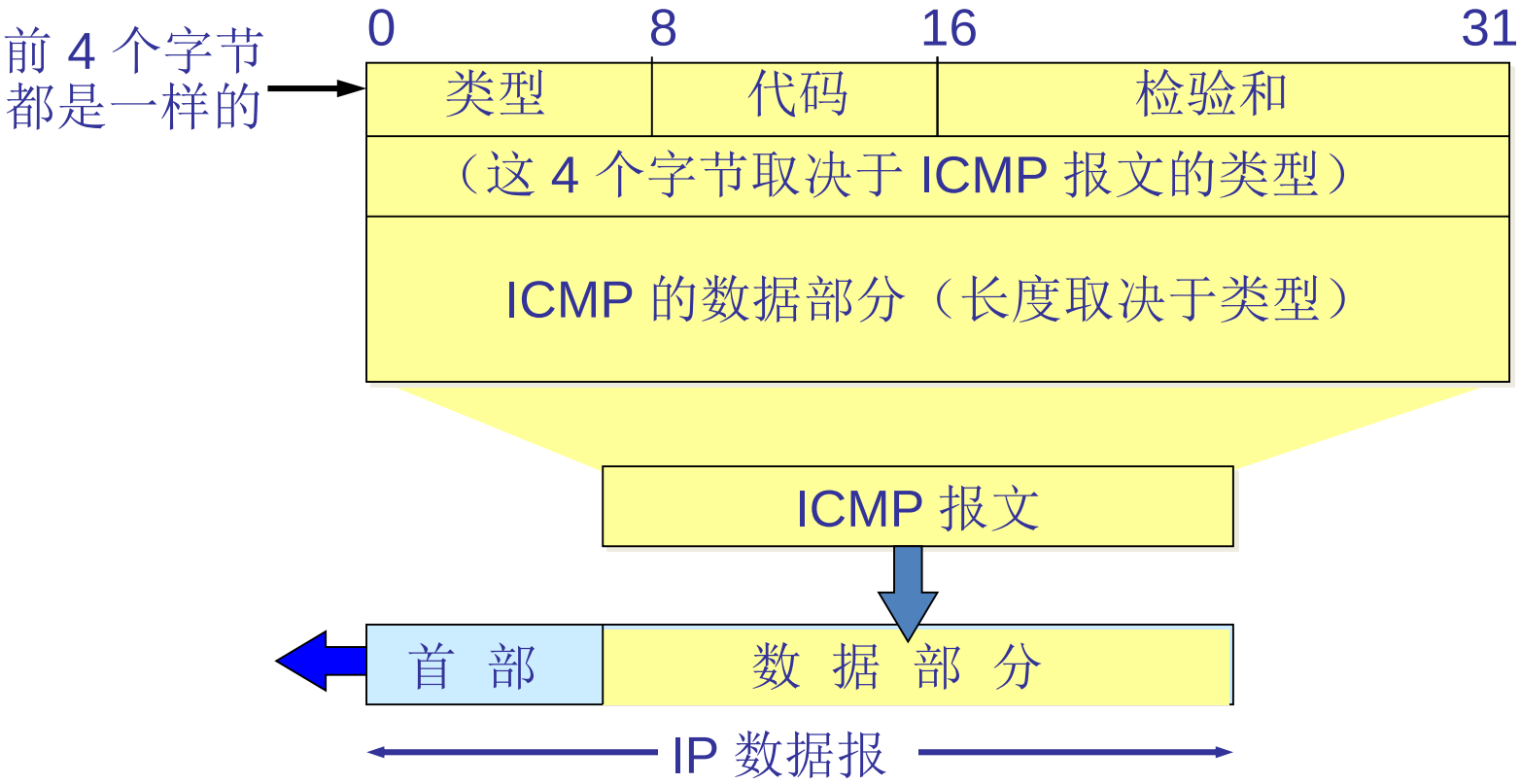
The principal ICMP message types.

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

网际控制报文协议 ICMP

- 为了提高 IP 数据报交付成功的机会，在网际层使用了网际控制报文协议 ICMP (Internet Control Message Protocol)。
- ICMP 允许主机或路由器报告差错情况和提供有关异常情况的报告。
- ICMP 不是高层协议，而是 IP 层的协议。
- ICMP 报文作为 IP 层数据报的数据，加上数据报的首部，组成 IP 数据报发送出去。

ICMP 报文的格式


前 4 个字节
都是一样的
0
8
16
31
类型
代码
检验和
（这 4 个字节取决于 ICMP 报文的类型）
ICMP 的数据部分（长度取决于类型）
ICMP 报文
首 部
数 据 部 分
IP 数据报

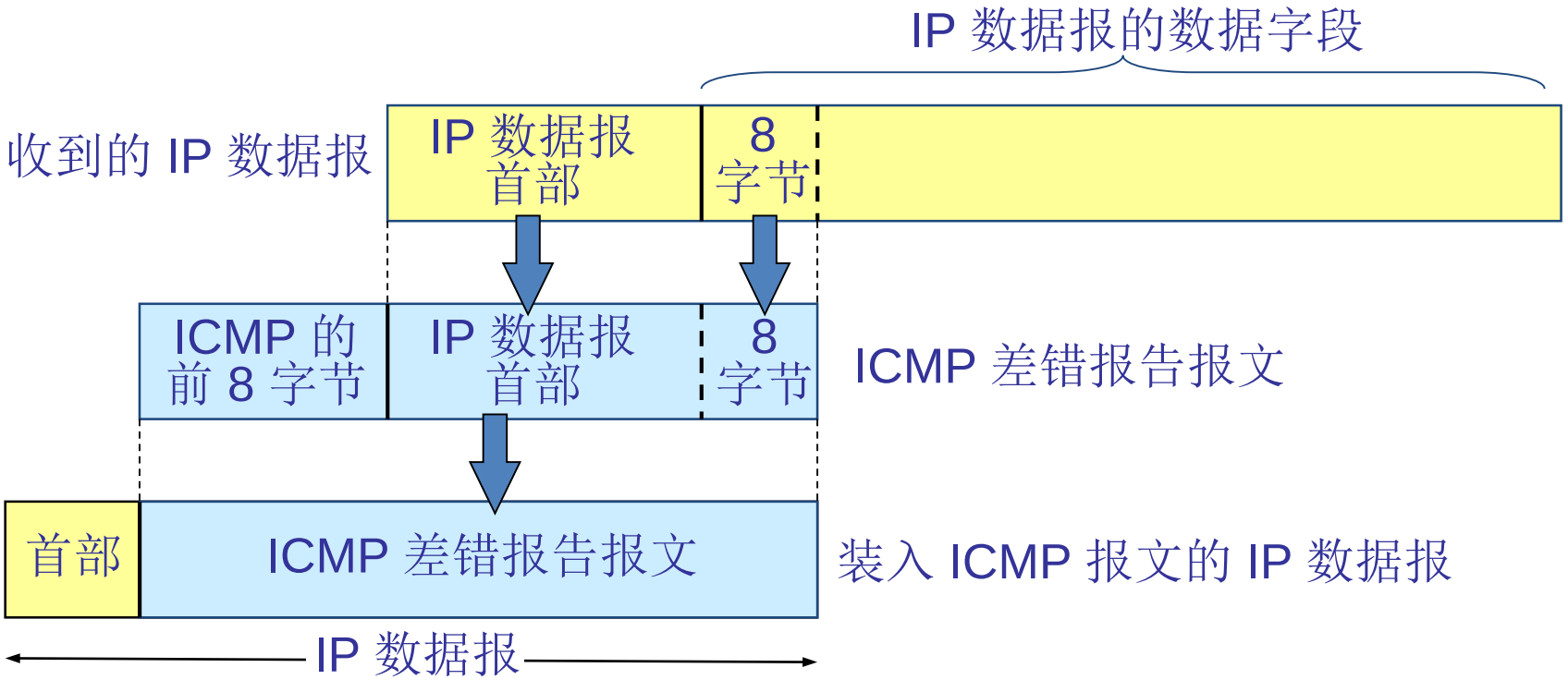
ICMP 报文的种类

- ICMP 报文的种类有两种，即 ICMP 差错报告报文和 ICMP 询问报文。
- ICMP 报文的前 4 个字节是统一的格式，共有三个字段：即类型、代码和检验和。接着的 4 个字节的内容与 ICMP 的类型有关。

ICMP 差错报告报文共有 5 种

- 终点不可达
- 源点抑制(Source quench)
- 时间超过
- 参数问题
- 改变路由（重定向）(Redirect)

ICMP 差错报告报文的数据字段的内容



不应发送 ICMP 差错报告报文的几种情况

- 对 ICMP 差错报告报文不再发送 ICMP 差错报告报文。
- 对第一个分片的数据报片的所有后续数据报片都不发送 ICMP 差错报告报文。
- 对具有多播地址的数据报都不发送 ICMP 差错报告报文。
- 对具有特殊地址（如127.0.0.0 或 0.0.0.0）的数据报不发送 ICMP 差错报告报文。

ICMP 询问报文有两种

- 回送请求和回答报文
- 时间戳请求和回答报文

下面的几种 ICMP 报文不再使用

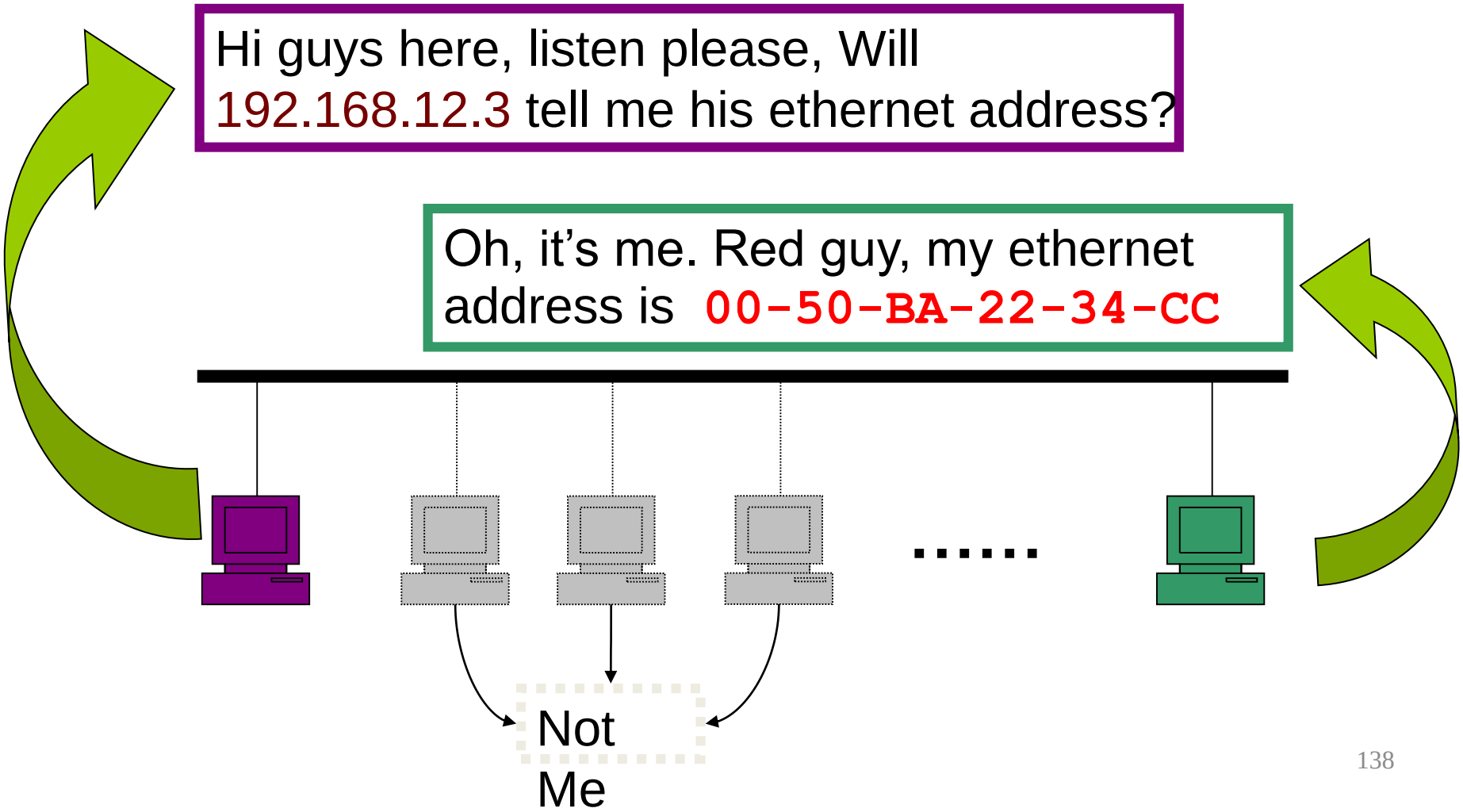
- 信息请求与回答报文
- 掩码地址请求和回答报文
- 路由器询问和通告报文

ICMP的应用举例
PING (Packet InterNet Groper)

- PING 用来测试两个主机之间的连通性。
- PING 使用了 ICMP 回送请求与回送回答报文。
- PING 是应用层直接使用网络层 ICMP 的例子，它没有通过运输层的 TCP 或UDP。

The Network Layer in the Internet:

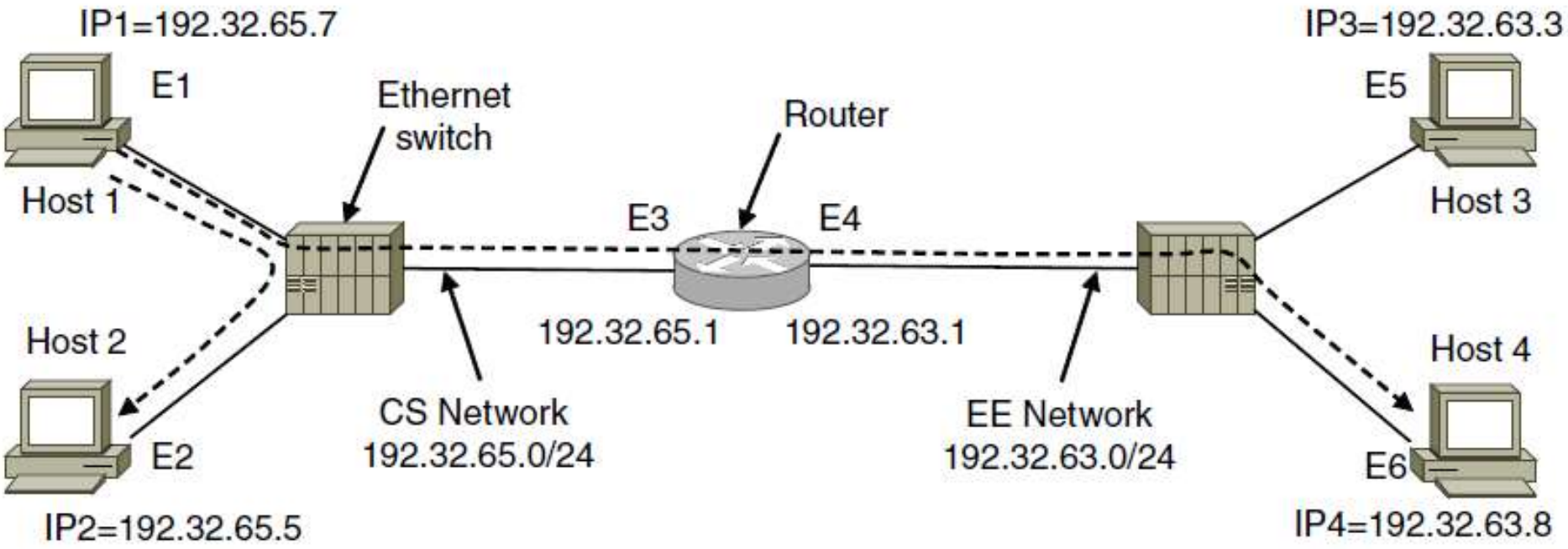
Internet Control Protocols: **ARP**



The Network Layer in the Internet:

Internet Control Protocols: ARP

Two switched Ethernet LANs joined by a router



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

The Network Layer in the Internet:

Internet Control Protocols: ARP

How does a user on host 1 send a packet to a user on host 2?

- Find the IP address for host 2 (e.g. DNS)
- Build a packet with 192.31.65.5 in the Destination address field
- Find the destination's Ethernet address: Conf File or ARP
- Build an Ethernet frame addressed to E2 and dump it into the Ethernet.
- The Ethernet board of host 2 detects this frame, recognizes it as a frame for itself, and causes an interrupt.
- The Ethernet driver extracts the IP packet from the payload and passes it to the IP software.

The Network Layer in the Internet:

Internet Control Protocols: ARP

How host 1 sends a packet to host 4 (192.31.63.8).

- Host 1 packets the IP packet and sends the frame to E3.
- When the CS router gets the Ethernet frame, it finds the physical address E3 by ARP. It then inserts the packet into the payload field of a frame addressed to E6 and puts it on the net.
- When the Ethernet frame arrives at host 4, the packet is extracted from the frame and passed to the IP software for processing.

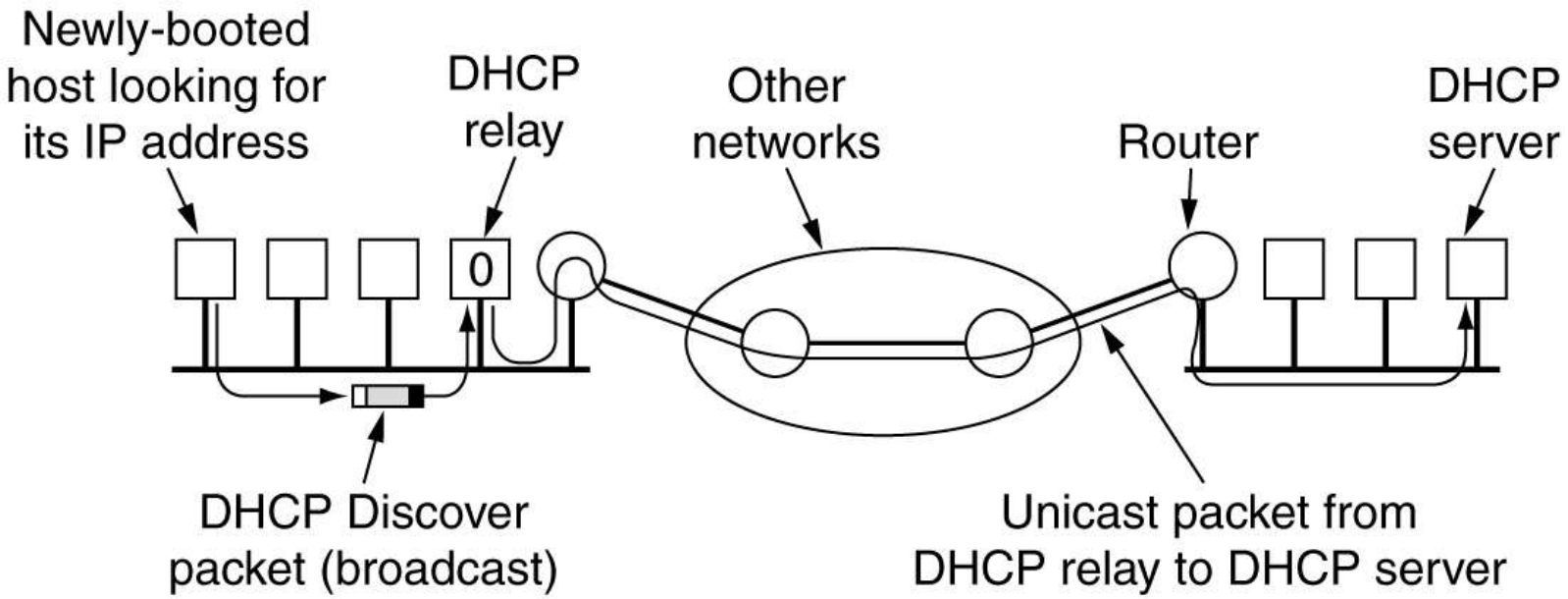
The Network Layer in the Internet:
Internet Control Protocols: DHCP

- **ARP** (Address Resolution Protocol, 地址辨析协议)
- **RARP** (Reverse Address Resolution Protocol，反向地址辨析协议)：Given a Link address, what is the corresponding IP address?
- **BOOTP** (Bootstrap Protocol) is better than RARP, can be forwarded by a router.
- **DHCP** (Dynamic Host Configuration Protocol, 动态主机配置协议) allows both manual IP address assignment and automatic assignment.

The Network Layer in the Internet:

Internet Control Protocols: DHCP

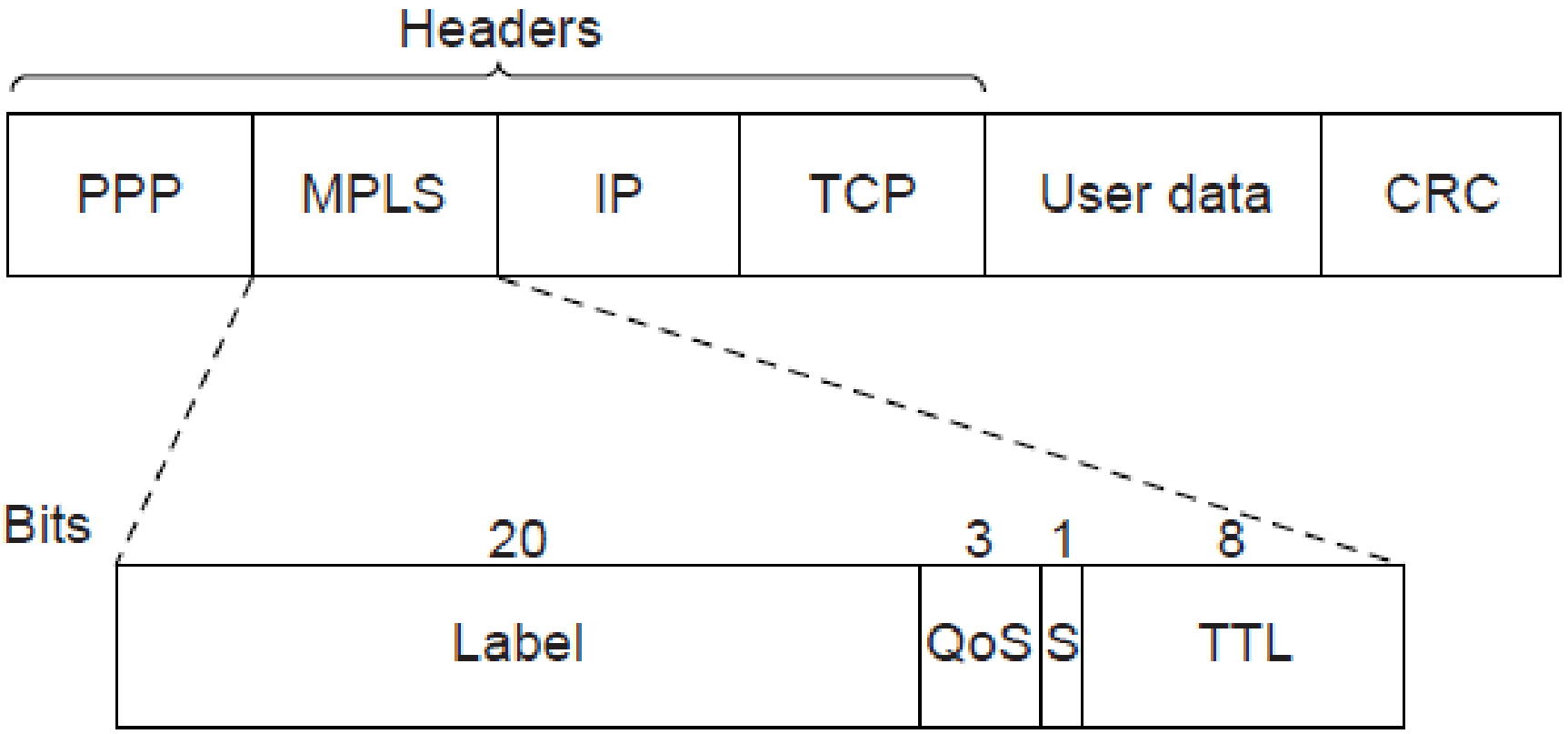
Operation of DHCP.



5.6.5 The Network Layer in the Internet:

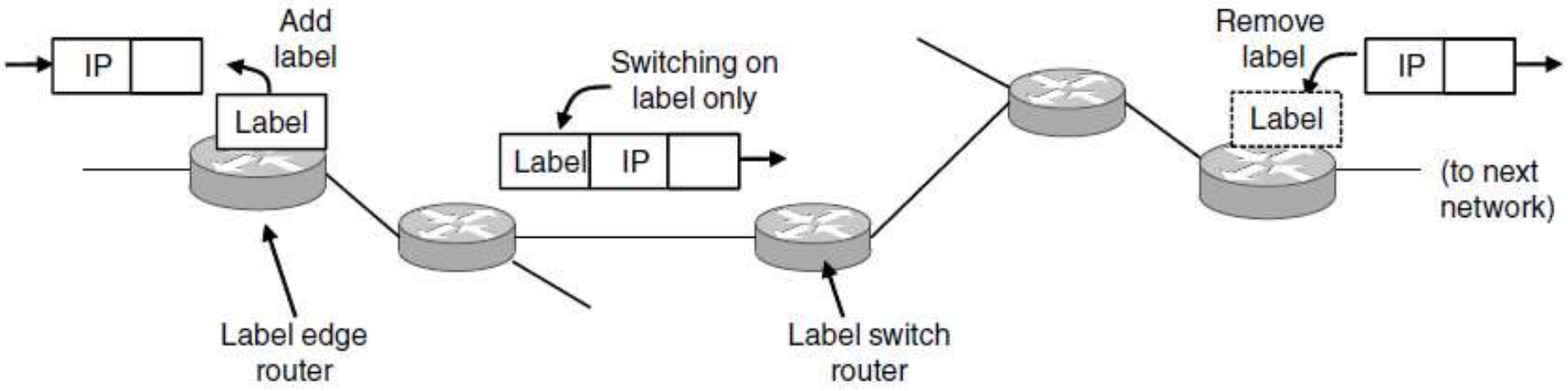
Label Switching and MPLS

Transmitting a TCP segment using IP, MPLS, and PPP



The Network Layer in the Internet:
Label Switching and MPLS

Forwarding an IP packet
through an MPLS network



5.6.6 The Network Layer in the Internet: OSPF – An Interior Gateway Routing Protocol

- The internet is made up of a large number of autonomous systems.
- A routing algorithm within an AS is called an **interior gateway protocol （IGP）**
  - Distance vector protocol (RIP)，采用UDP报文
  - Link state protocol (1979)
  - OSPF (Open Shortest Path First in 1990)，直接用 IP 数据报传送
- A routing algorithm between ASes is called an exterior gateway protocol
  - BGP (Border Gateway Protocol)

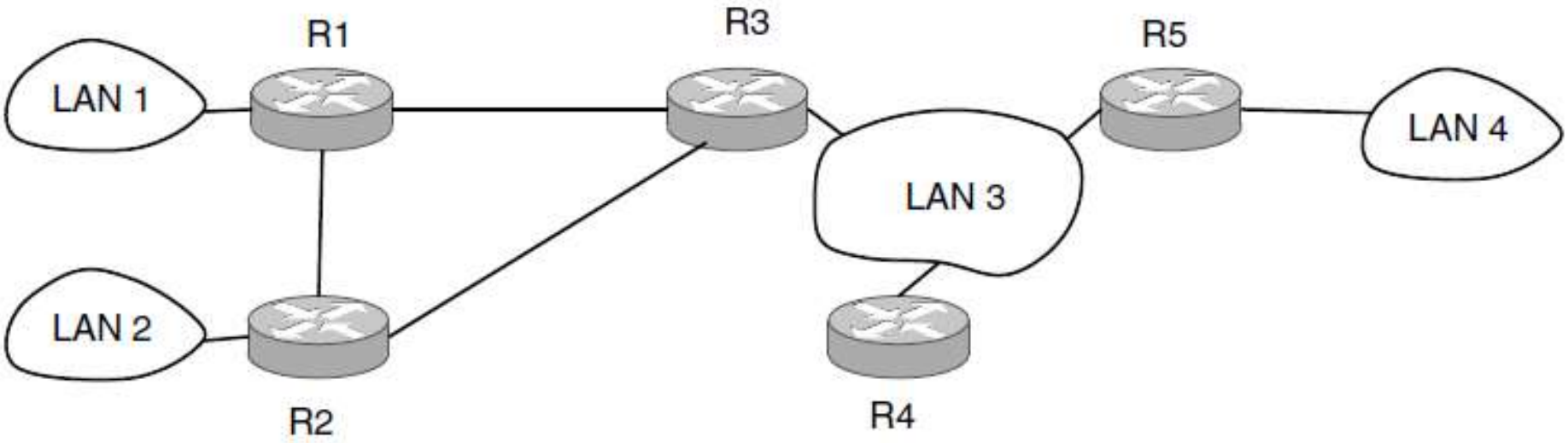
The Network Layer in the Internet: OSPF

Requirements for the new routing algorithms:

1. To be open, hence the "O" in OSPF.
2. To support a variety of distance metrics.
3. To be a dynamic algorithm.
4. To support routing based on type of service.
5. To do load balancing
6. To support hierarchical systems.
7. To support security
8. To support connection to the Internet via a tunnel

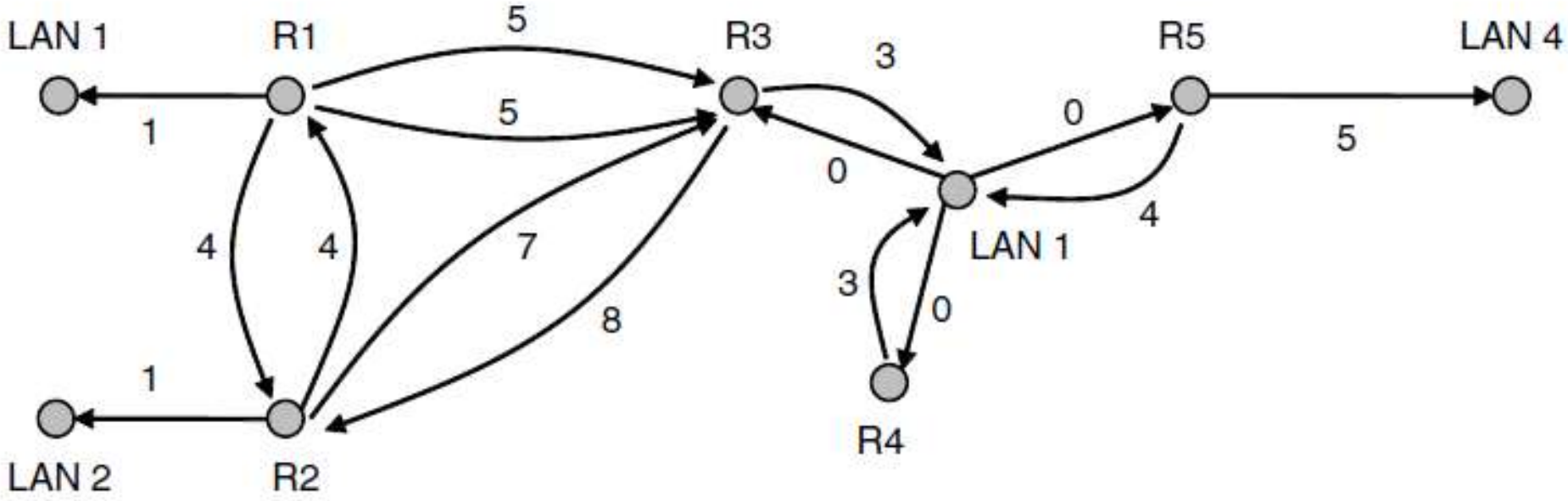
The Network Layer in the Internet: OSPF

An autonomous system



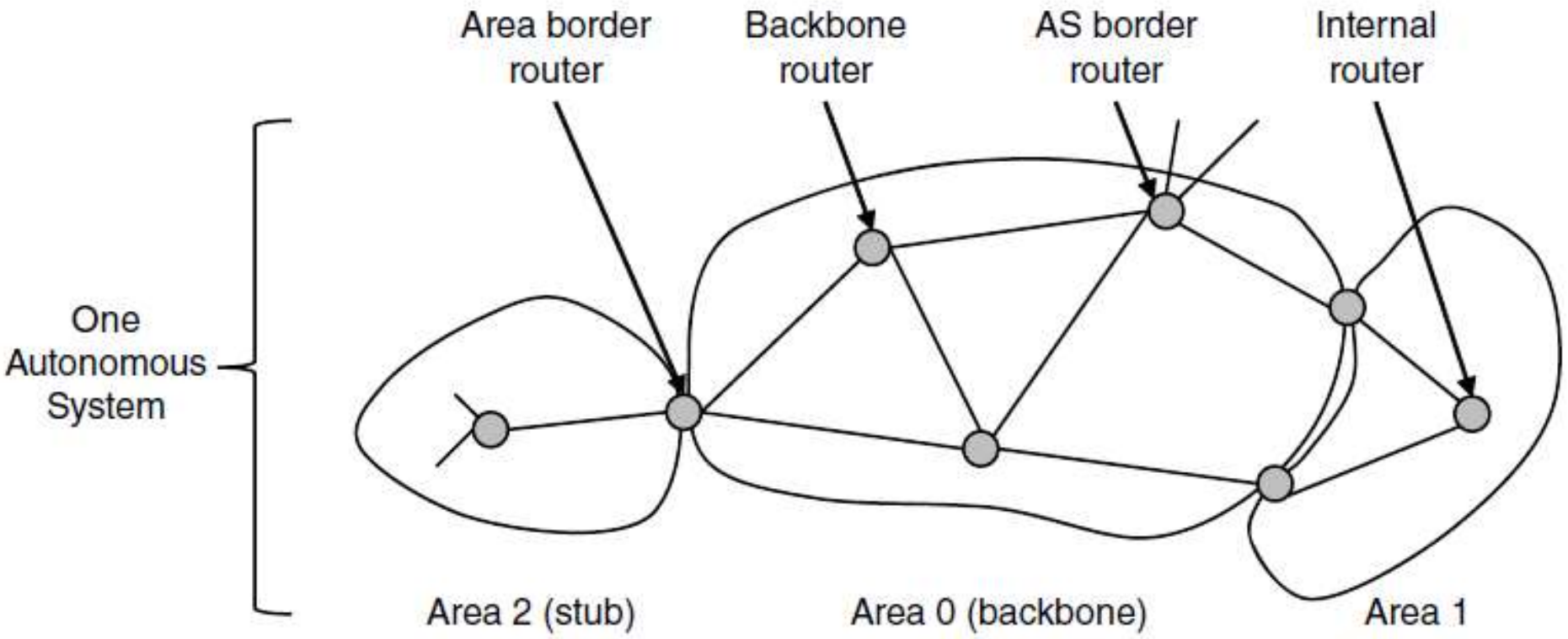
The Network Layer in the Internet:
OSPF

A graph representation of the previous slide.



The Network Layer in the Internet: OSPF

The relation between ASes, backbones, and areas in OSPF.



The Network Layer in the Internet:
OSPF

How a OSPF router works?

- When a router boots, it sends HELLO messages. From the response, each router learns who its neighbors are.
- Adjacent routers exchange information.
  - Each router periodically floods LINK STATE UPDATE messages to each of its adjacent routers. These must be acknowledged (LINK STATE ACK).
  - Either partner can request link state information from the other one using LINK STATE REQUEST messages.
  - Each router constructs the graph for its area(s) and compute the shortest path.

The Network Layer in the Internet: OSPF

The five types of OSPF messages.

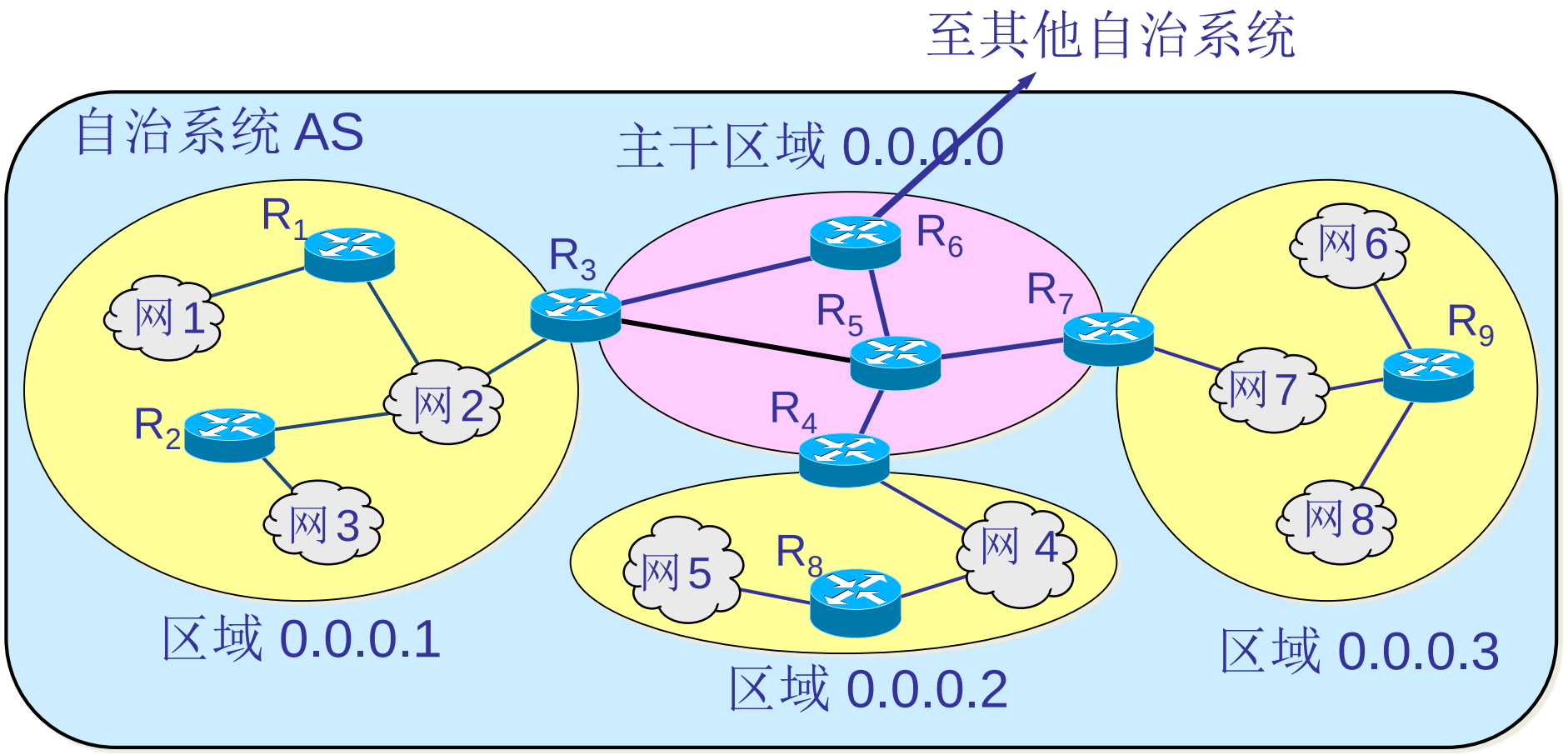
Message type	Description
Hello	Used to discover who the neighbors are
Link state update	Provides the sender's costs to its neighbors
Link state ack	Acknowledges link state update
Database description	Announces which updates the sender has
Link state request	Requests information from the partner

- Encapsulation protocol for OSPF protocol： IP
- Encapsulation protocol for RIP protocol： UDP

OSPF 的区域(area)

- 为了使 OSPF 能够用于规模很大的网络，OSPF 将一个自治系统再划分为若干个更小的范围，叫作区域。
- 每一个区域都有一个 32 位的区域标识符（用点分十进制表示）。
- 区域也不能太大，在一个区域内的路由器最好不超过 200 个。

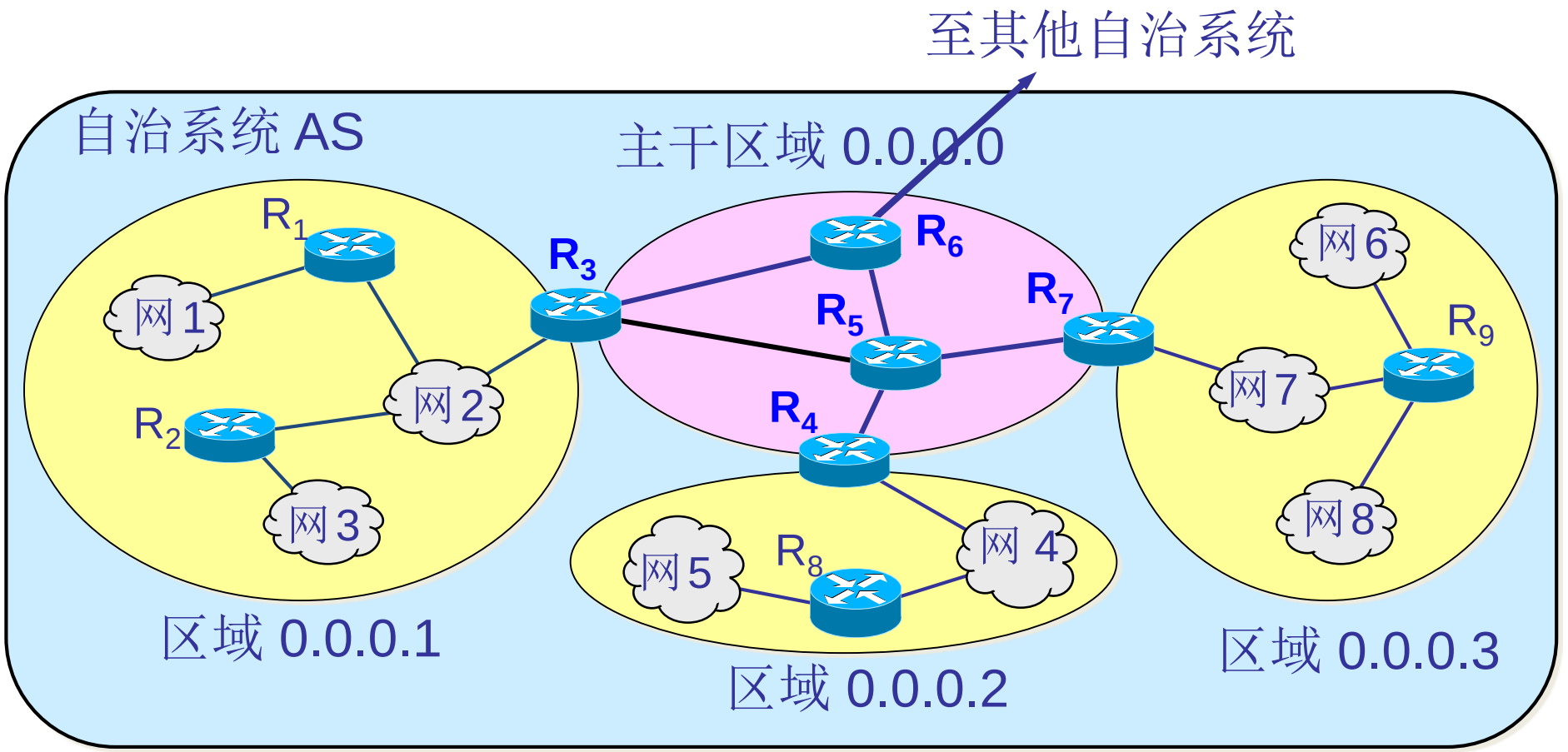
OSPF 划分为两种不同的区域



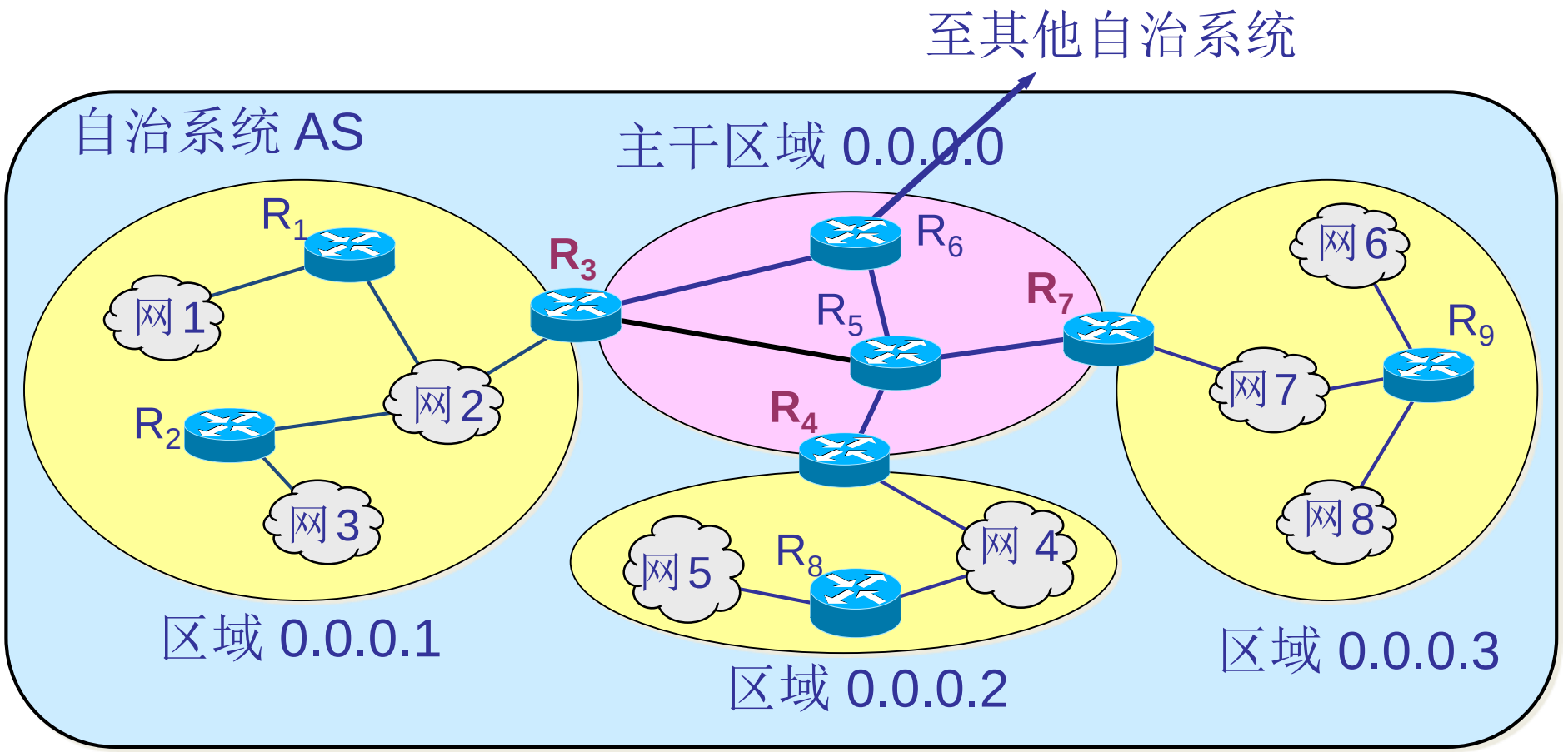
划分区域

- 划分区域的好处就是将利用洪泛法交换链路状态信息的范围局限于每一个区域而不是整个的自治系统，这就减少了整个网络上的通信量。
- 在一个区域内部的路由器只知道本区域的完整网络拓扑，而不知道其他区域的网络拓扑的情况。
- OSPF 使用层次结构的区域划分。在上层的区域叫作主干区域(backbone area)。主干区域的标识符规定为0.0.0.0。主干区域的作用是用来连通其他在下层的区域。

主干路由器


至其他自治系统
自治系统 AS
主干区域 0.0.0.0
R1
网 1
R2
网 2
网 3
区域 0.0.0.1
R3
R6
R5
R7
R4
网 5
R8
网 4
区域 0.0.0.2
网 6
R9
网 7
网 8
区域 0.0.0.3

区域边界路由器


至其他自治系统
自治系统 AS
主干区域 0.0.0.0
R1
网 1
网 2
R2
网 3
区域 0.0.0.1
R3
R6
R5
R7
R4
网 5
R8
网 4
区域 0.0.0.2
网 6
R9
网 7
网 8
区域 0.0.0.3

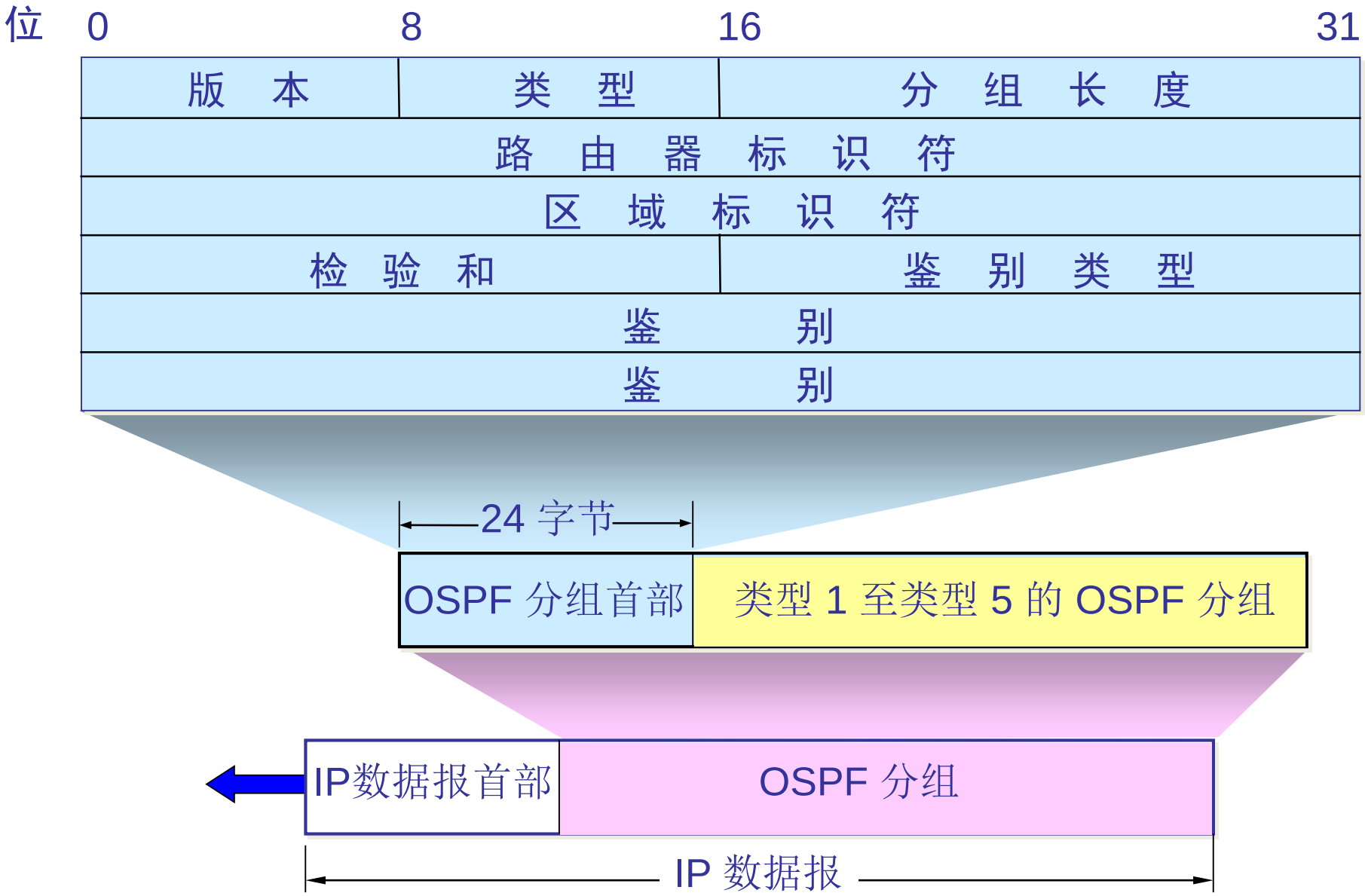
OSPF 直接用 IP 数据报传送

- OSPF 不用 UDP 而是直接用 IP 数据报传送。
- OSPF 构成的数据报很短。这样做可减少路由信息的通信量。
- 数据报很短的另一好处是可以不必将长的数据报分片传送。分片传送的数据报只要丢失一个，就无法组装成原来的数据报，而整个数据报就必须重传。

OSPF 的其他特点

- OSPF 对不同的链路可根据 IP 分组的不同服务类型 TOS 而设置成不同的代价。因此，OSPF 对于不同类型的业务可计算出不同的路由。
- 如果到同一个目的网络有多条相同代价的路径，那么可以将通信量分配给这几条路径。这叫作多路径间的负载平衡。
- 所有在 OSPF 路由器之间交换的分组都具有鉴别的功能。
- 支持可变长度的子网划分和无分类编址 CIDR。
- 每一个链路状态都带上一个 32 位的序号，序号越大状态就越新。

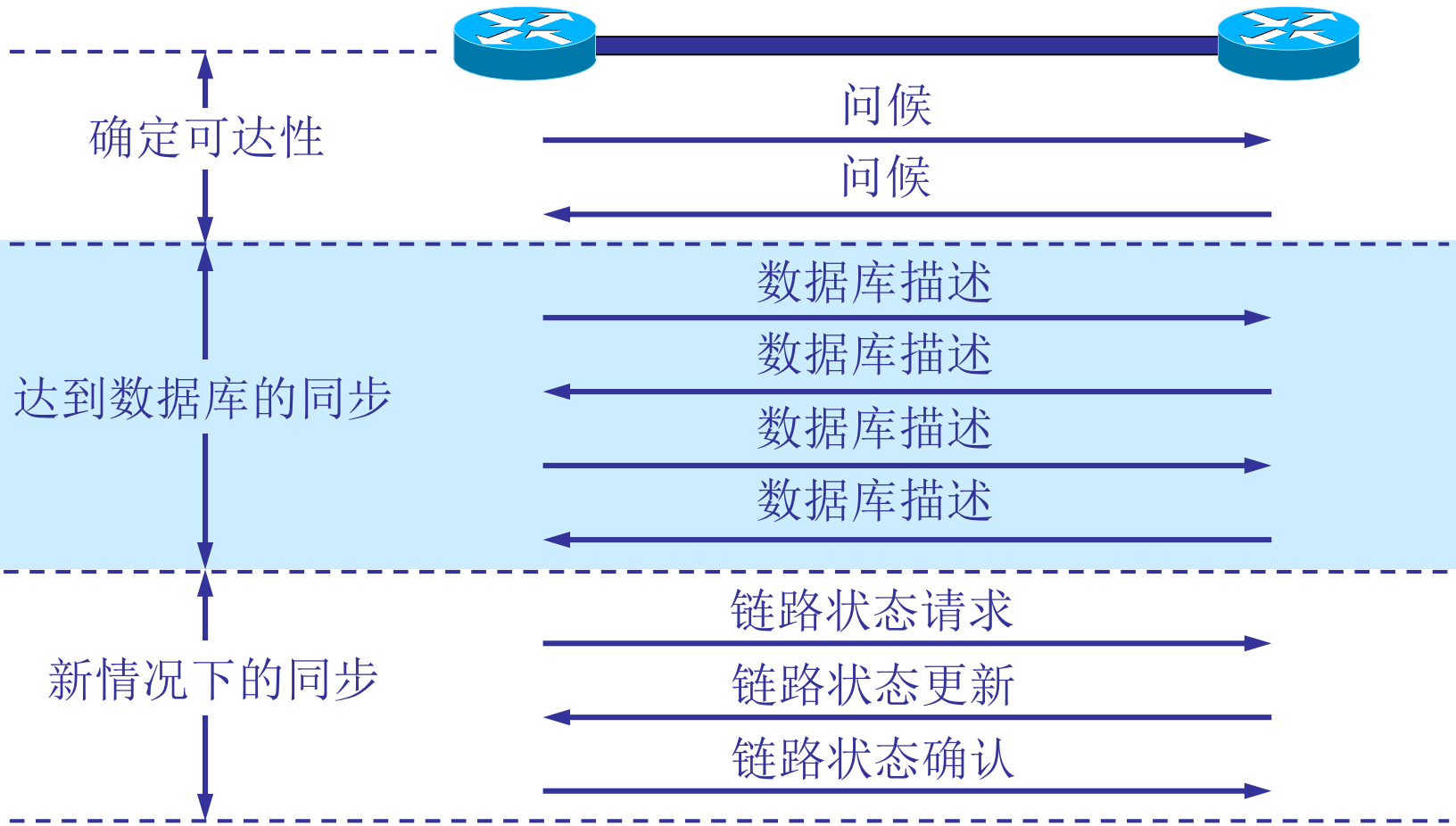
OSPF 分组


位
0
8
16
31
版 本
类 型
分 组 长 度
路 由 器 标 识 符
区 域 标 识 符
检 验 和
鉴 别 类 型
鉴 别
鉴 别
24 字节
OSPF 分组首部
类型 1 至类型 5 的 OSPF 分组
IP数据报首部
OSPF 分组
IP 数据报

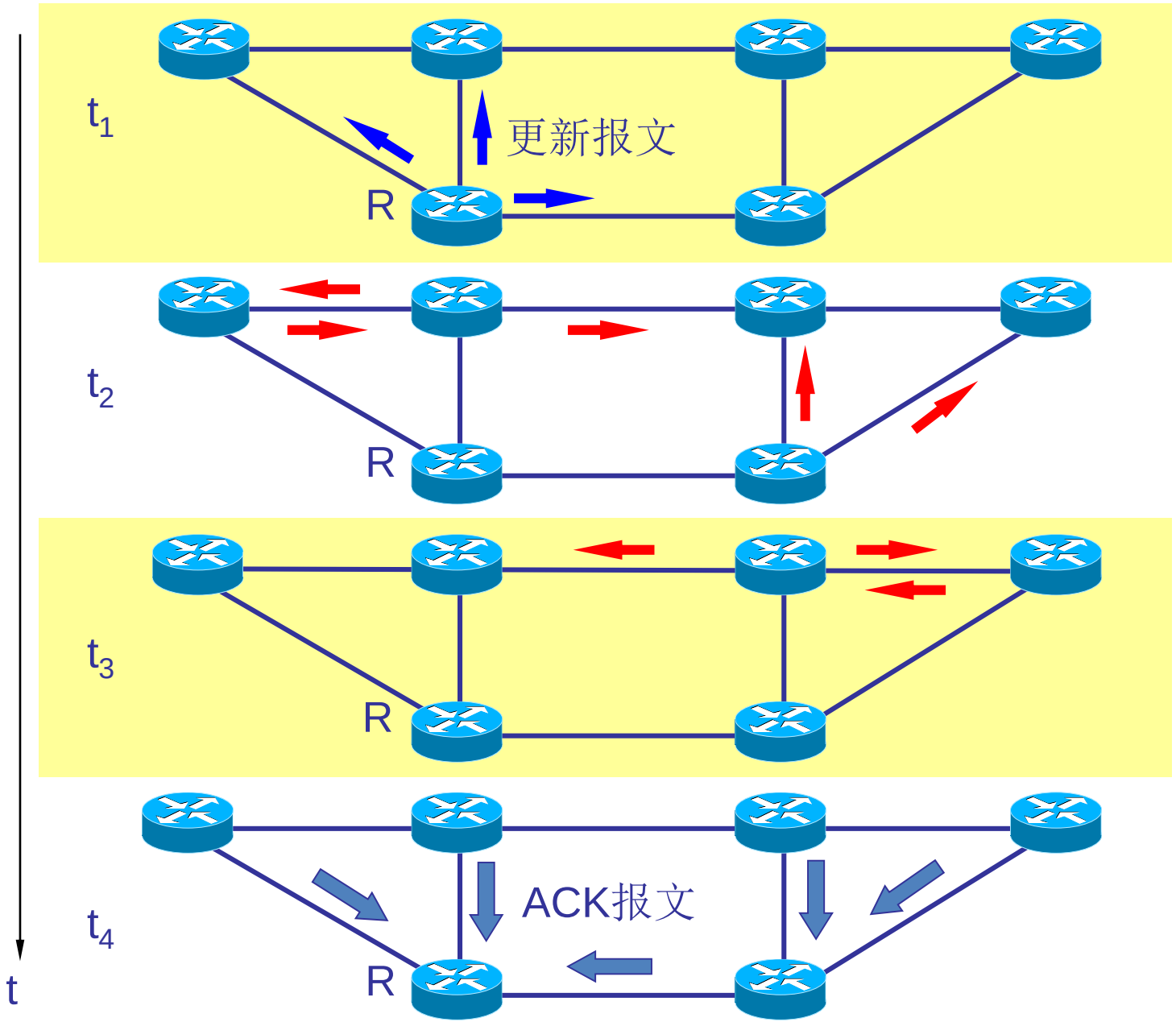
2. OSPF 的五种分组类型

- 类型1，问候(Hello)分组。
- 类型2，数据库描述(Database Description)分组。
- 类型3，链路状态请求(Link State Request)分组。
- 类型4，链路状态更新(Link State Update)分组，用洪泛法对全网更新链路状态。
- 类型5，链路状态确认(Link State Acknowledgment)分组。

OSPF的基本操作


确定可达性
问候
问候
达到数据库的同步
数据库描述
数据库描述
数据库描述
数据库描述
新情况下的同步
链路状态请求
链路状态更新
链路状态确认

OSPF 使用的是可靠的洪泛法



OSPF 的其他特点

- OSPF 还规定每隔一段时间，如 30 分钟，要刷新一次数据库中的链路状态。
- 由于一个路由器的链路状态只涉及到与相邻路由器的连通状态，因而与整个互联网的规模并无直接关系。因此当互联网规模很大时，OSPF 协议要比距离向量协议 RIP 好得多。
- OSPF 没有“坏消息传播得慢”的问题，据统计，其响应网络变化的时间小于 100 ms。

指定的路由器
(designated router)

- 多点接入的局域网采用了指定的路由器的方法，使广播的信息量大大减少。
- 指定的路由器代表该局域网上所有的链路向连接到该网络上的各路由器发送状态信息。

5.6.7 The Network Layer in the Internet:
BGP – The Exterior Gateway Routing Protocol

- Exterior gateway protocol routers have to worry about politics a great deal.
- Typical policies involve political, security, or economic considerations. A few examples of routing constraints are

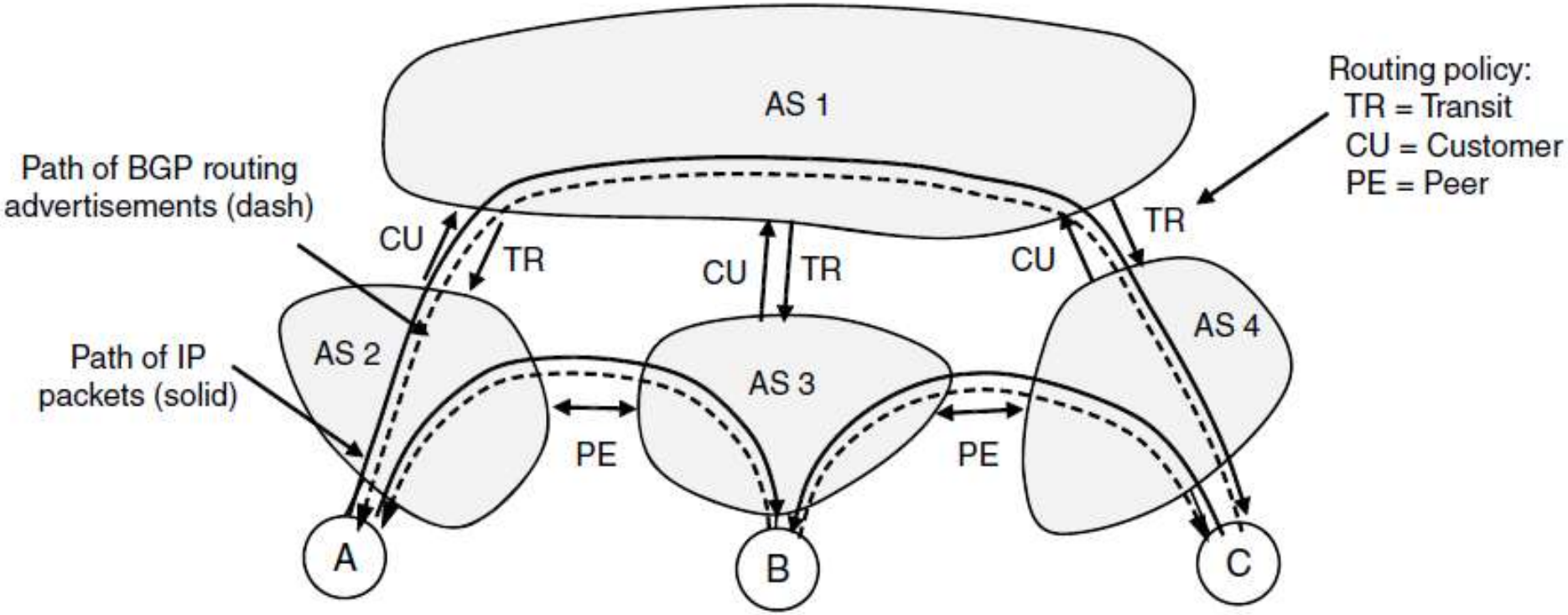
1. No commercial traffic for educational network
2. Never put Iraq on route starting at Pentagon
3. Choose cheaper network
4. Choose better performing network
5. Don't go from Apple to Google to Apple

The Network Layer in the Internet: BGP

- BGP routers communicate with each other by establishing TCP connections.
- BGP is fundamentally a distance vector protocol, but quite different from most others such as RIP.
  - Each BGP router keeps track of the path used (instead of maintaining just the cost to each destination)
  - Each BGP router tells its neighbors the exact path it is using (instead of periodically given each neighbor its estimated cost to each possible destination)

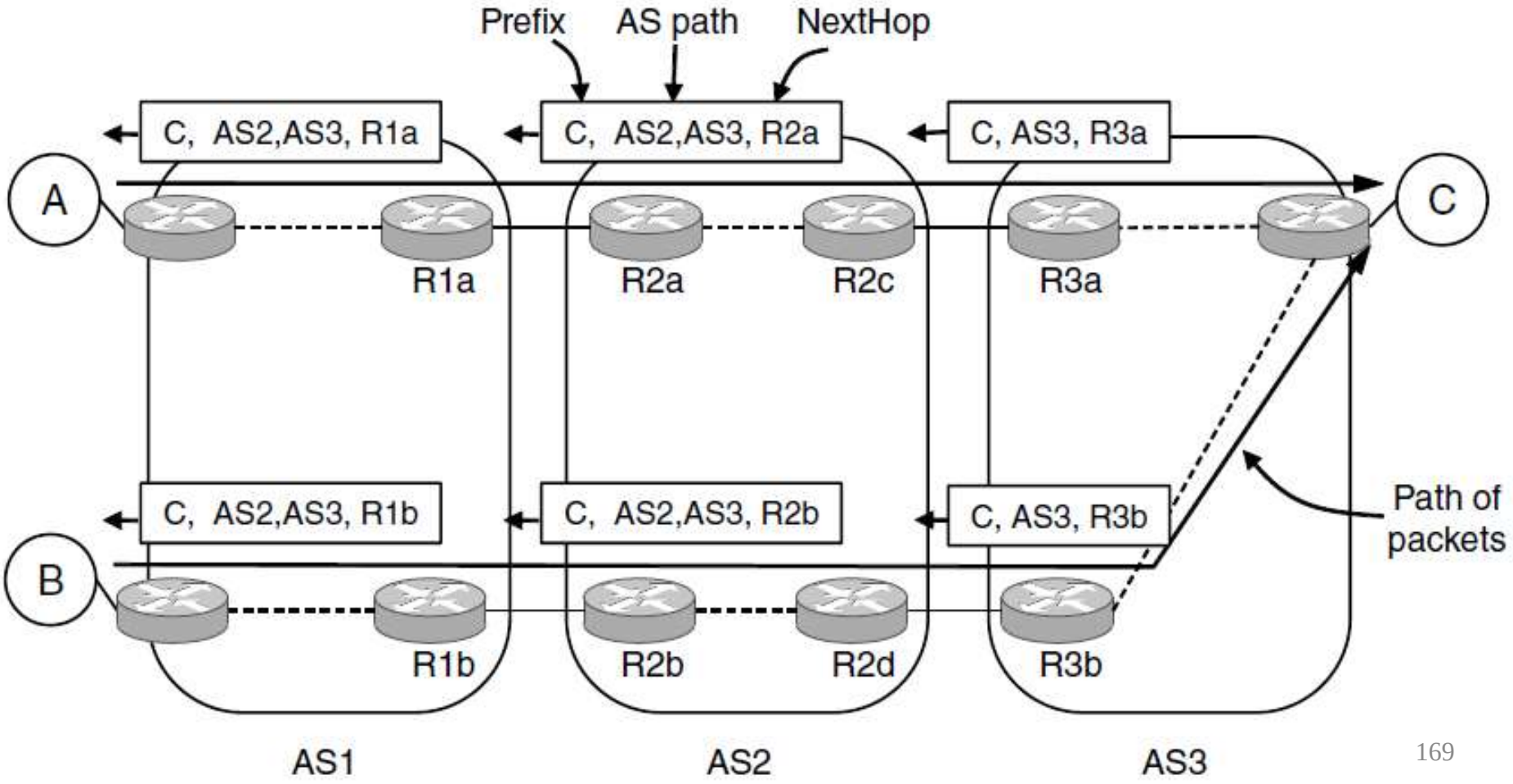
The Network Layer in the Internet: BGP

Routing policies between four Autonomous Systems



The Network Layer in the Internet:
BGP

Propagation of BGP route advertisements



外部网关协议 BGP

- BGP 是不同自治系统的路由器之间交换路由信息的协议。
- BGP 较新版本是 2006 年 1 月发表的 BGP-4（BGP 第 4 个版本），即 RFC 4271 ~ 4278。
- 可以将 BGP-4 简写为 BGP。

BGP 使用的环境却不同

- 因特网的规模太大，使得自治系统之间路由选择非常困难。对于自治系统之间的路由选择，要寻找最佳路由是很不现实的。
  - 当一条路径通过几个不同 AS 时，要想对这样的路径计算出有意义的代价是不太可能的。
  - 比较合理的做法是在 AS 之间交换“可达性”信息。
- 自治系统之间的路由选择必须考虑有关策略。
- 因此，边界网关协议 BGP 只能是力求寻找一条能够到达目的网络且比较好的路由（不能兜圈子），而并非要寻找一条最佳路由。

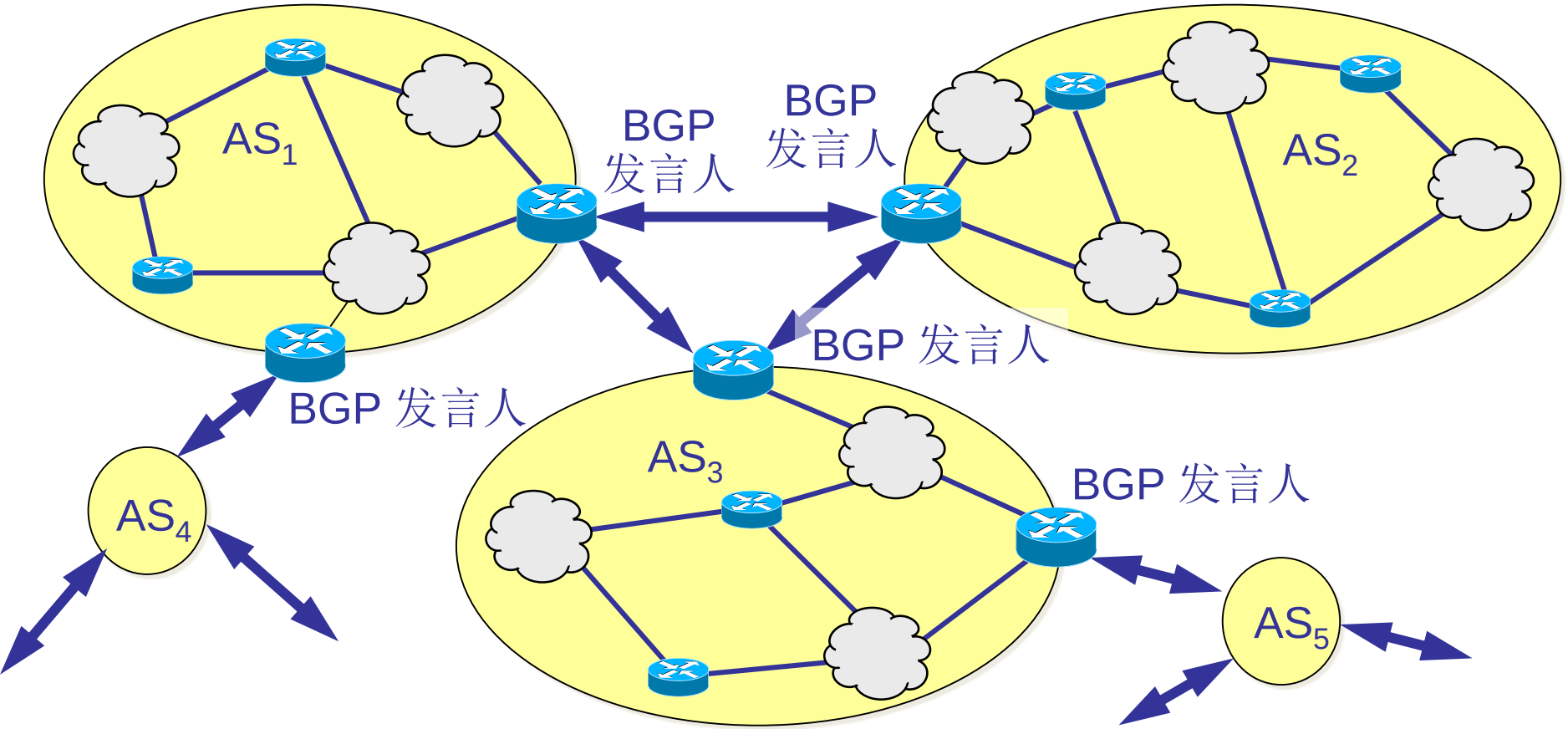
BGP 发言人
(BGP speaker)

- 每一个自治系统的管理员要选择至少一个路由器作为该自治系统的“ BGP 发言人”。
- 一般说来，两个 BGP 发言人都是通过一个共享网络连接在一起的，而 BGP 发言人往往就是 BGP 边界路由器，但也可以不是 BGP 边界路由器。

BGP 交换路由信息

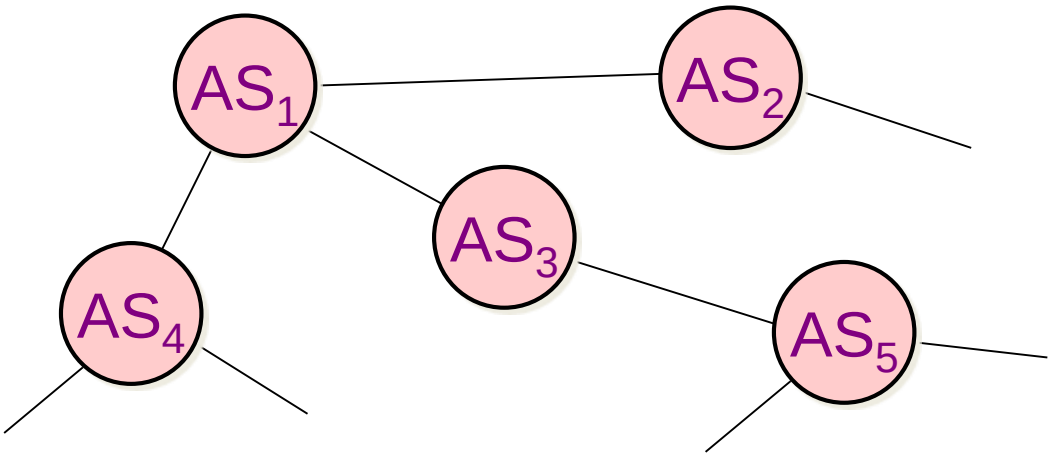
- 一个 BGP 发言人与其他自治系统中的 BGP 发言人要交换路由信息，就要先建立 TCP 连接，然后在此连接上交换 BGP 报文以建立 BGP 会话 (session)，利用 BGP 会话交换路由信息。
- 使用 TCP 连接能提供可靠的服务，也简化了路由选择协议。
- 使用 TCP 连接交换路由信息的两个 BGP 发言人，彼此成为对方的邻站或对等站。

BGP 发言人和
自治系统 AS 的关系



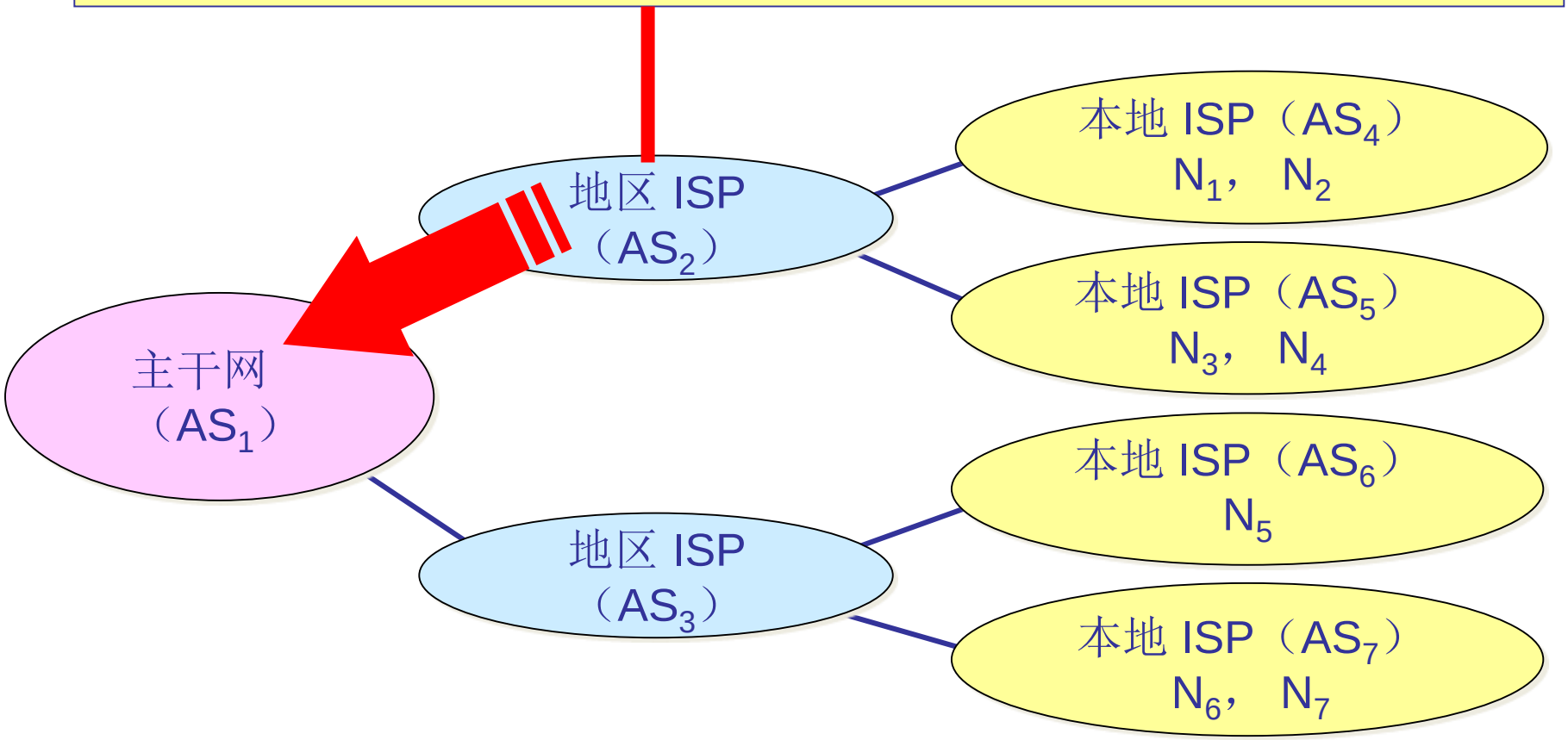
AS 的连通图举例

- BGP 所交换的网络可达性的信息就是要到达某个网络所要经过的一系列 AS。
- 当 BGP 发言人互相交换了网络可达性的信息后，各 BGP 发言人就根据所采用的策略从收到的路由信息中找出到达各 AS 的较好路由。



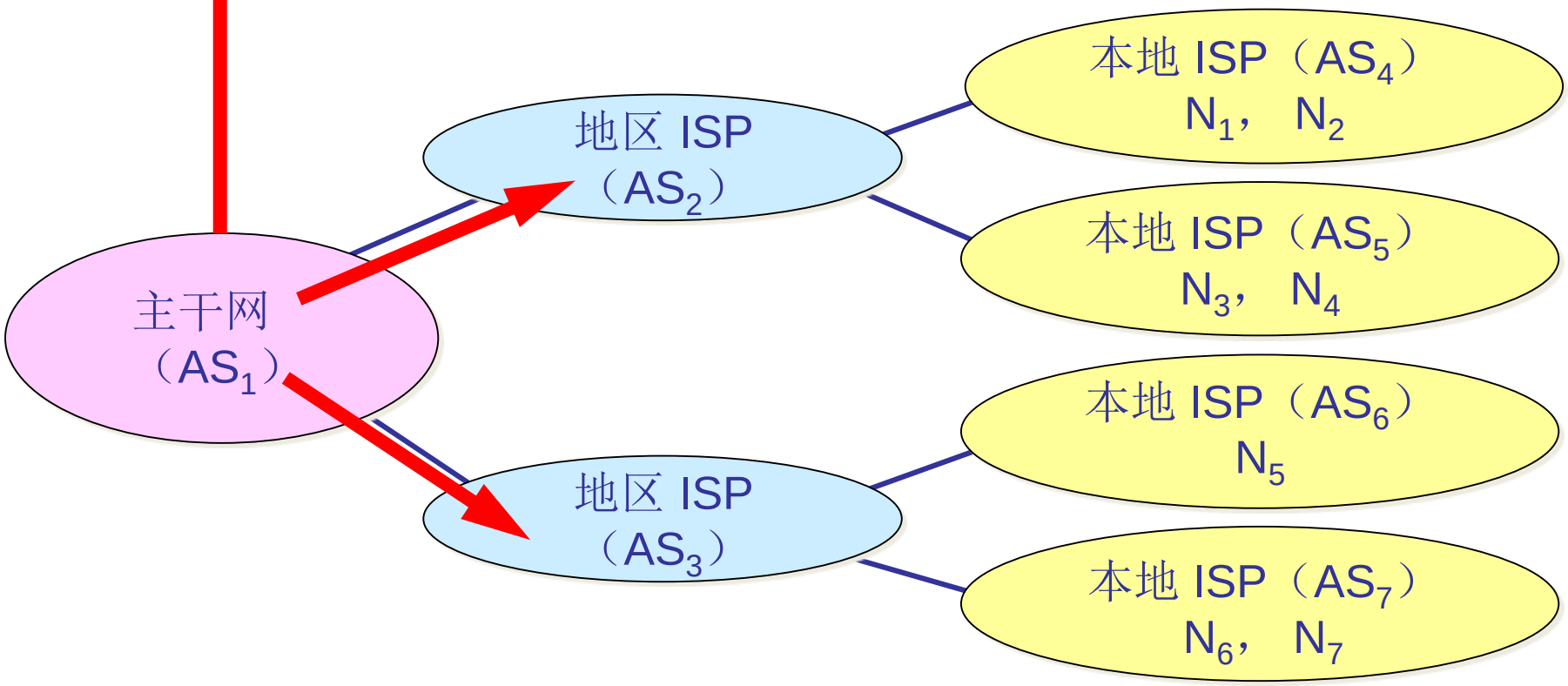
BGP 发言人交换路径向量

自治系统 AS_2 的 BGP 发言人通知主干网的 BGP 发言人：
“要到达网络 N_1, N_2, N_3 和 N_4 可经过 AS_2。”



BGP 发言人交换路径向量

主干网还可发出通知：“要到达网络 N_5, N_6 和 N_7 可沿路径（AS_1, AS_3）。”



BGP 协议的特点

- BGP 协议交换路由信息的结点数量级是自治系统数的量级，这要比这些自治系统中的网络数少很多。
- 每一个自治系统中 BGP 发言人（或边界路由器）的数目是很少的。这样就使得自治系统之间的路由选择不致过分复杂。

BGP 协议的特点

- BGP 支持 CIDR，因此 BGP 的路由表也就应当包括目的网络前缀、下一跳路由器，以及到达该目的网络所要经过的各个自治系统序列。
- 在BGP 刚刚运行时，BGP 的邻站是交换整个的 BGP 路由表。但以后只需要在发生变化时更新有变化的部分。这样做对节省网络带宽和减少路由器的处理开销方面都有好处。

BGP-4 共使用四种报文

(1) 打开(OPEN)报文，用来与相邻的另一个BGP发言人建立关系。

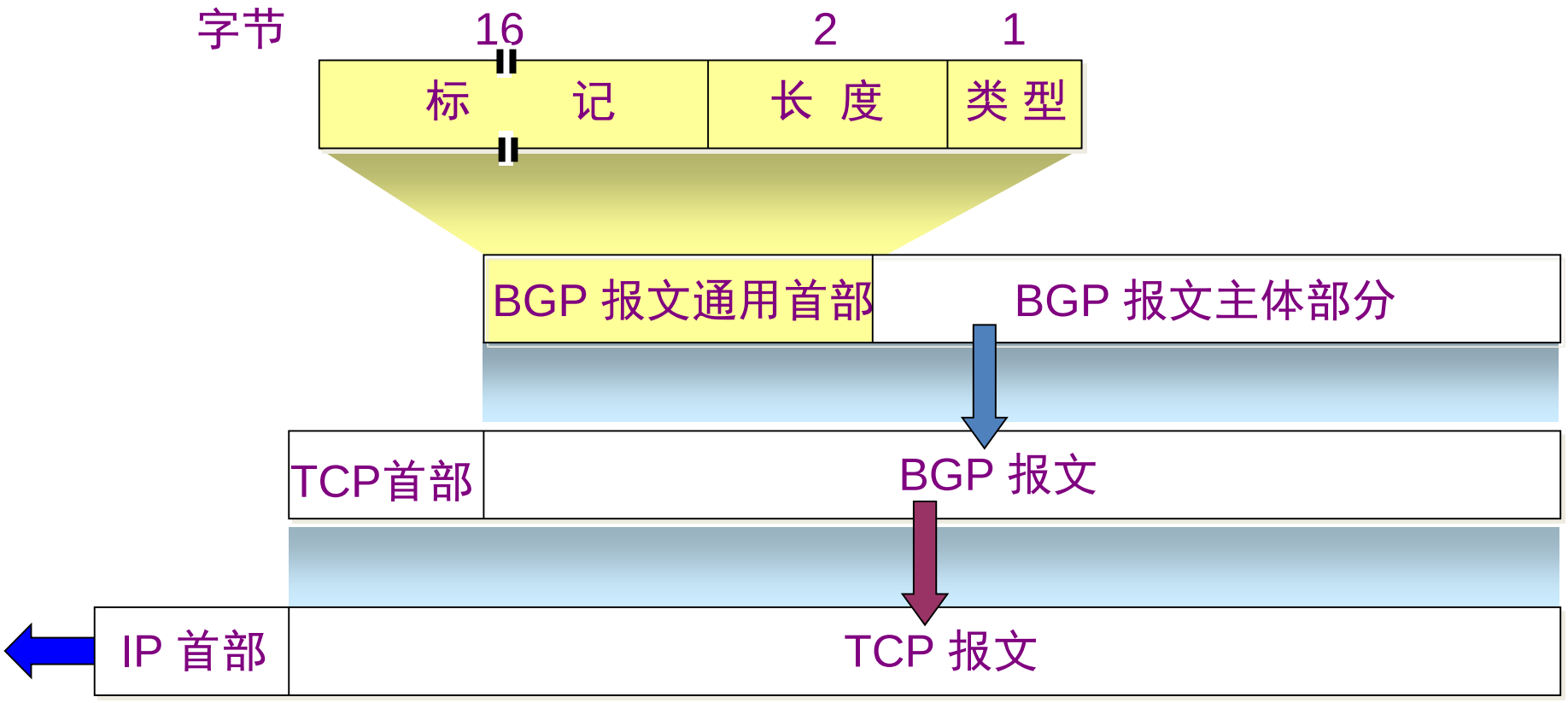
(2) 更新(UPDATE)报文，用来发送某一路由的信息，以及列出要撤消的多条路由。

(3) 保活(KEEPALIVE)报文，用来确认打开报文和周期性地证实邻站关系。

(4) 通知(NOTIFICATION)报文，用来发送检测到的差错。

- 在 RFC 2918 中增加了 ROUTE-REFRESH 报文，用来请求对等端重新通告。

BGP 报文具有通用的首部



5.6.8 The Network Layer in the Internet:

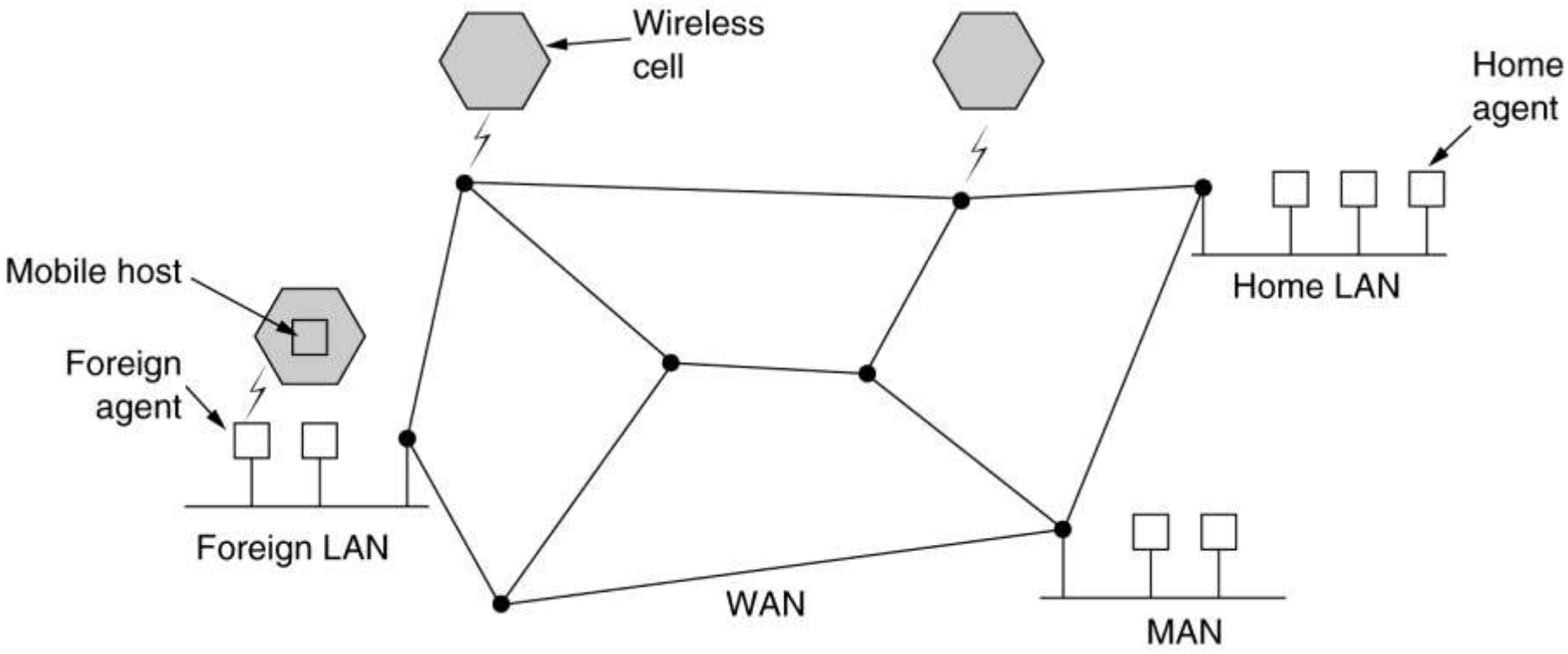
Internet Multicasting

- Databases, transmitting stock quotes to multiple brokers, and handling digital conference telephone calls
- IP supports multicasting, using class D address. 28 bits are available for identifying groups （最高4位是1110，共32位）
- Two kinds of group addresses are supported: permanent addresses and temporary ones.
- The range of IP addresses 224.0.0.0/24 is reserved for multicast on the local network. Some permanent group examples:
  - 224.0.0.1: All systems on a LAN
  - 224.0.0.2: All routers on a LAN
  - 224.0.0.5: All OSPF routers on a LAN
  - 224.0.0.6: All designated OSPF routers on a LAN
  - 224.0.0.251: All DNS servers on a LAN

The Network Layer in the Internet:
Internet Multicasting

- Multicasting is implemented by special multicast routers, which may or may not be collocated with the standard routers.
  - About once a minute, each multicast router sends a hardware multicast to the hosts on its LAN asking them to report back on the groups their processes currently belong to.
  - Each host sends back responses for all the class D addresses it is interested in.
  - This query and response packets use a protocol called IGMP (Internet Group Management Protocol).
- Multicast routing is done using spanning trees.

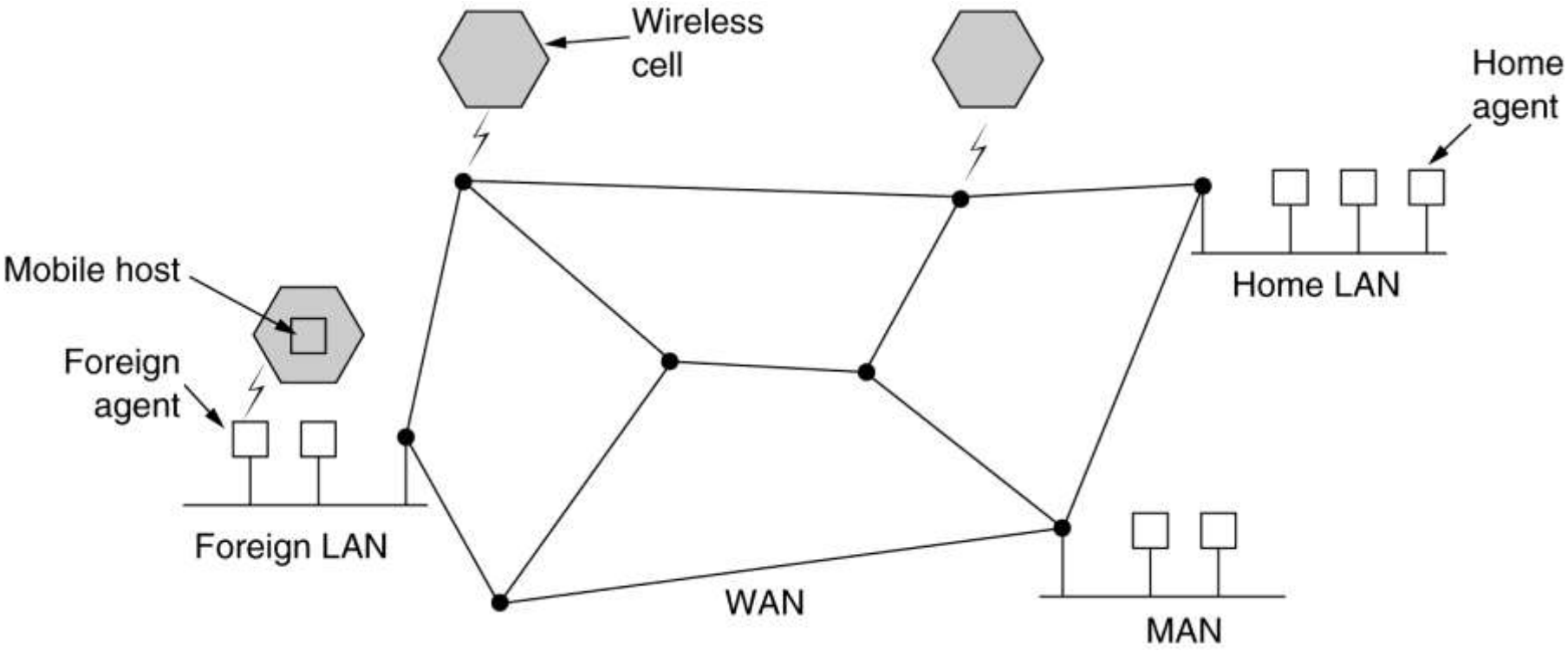
5.6.9 The Network Layer in the Internet:
Mobile IP



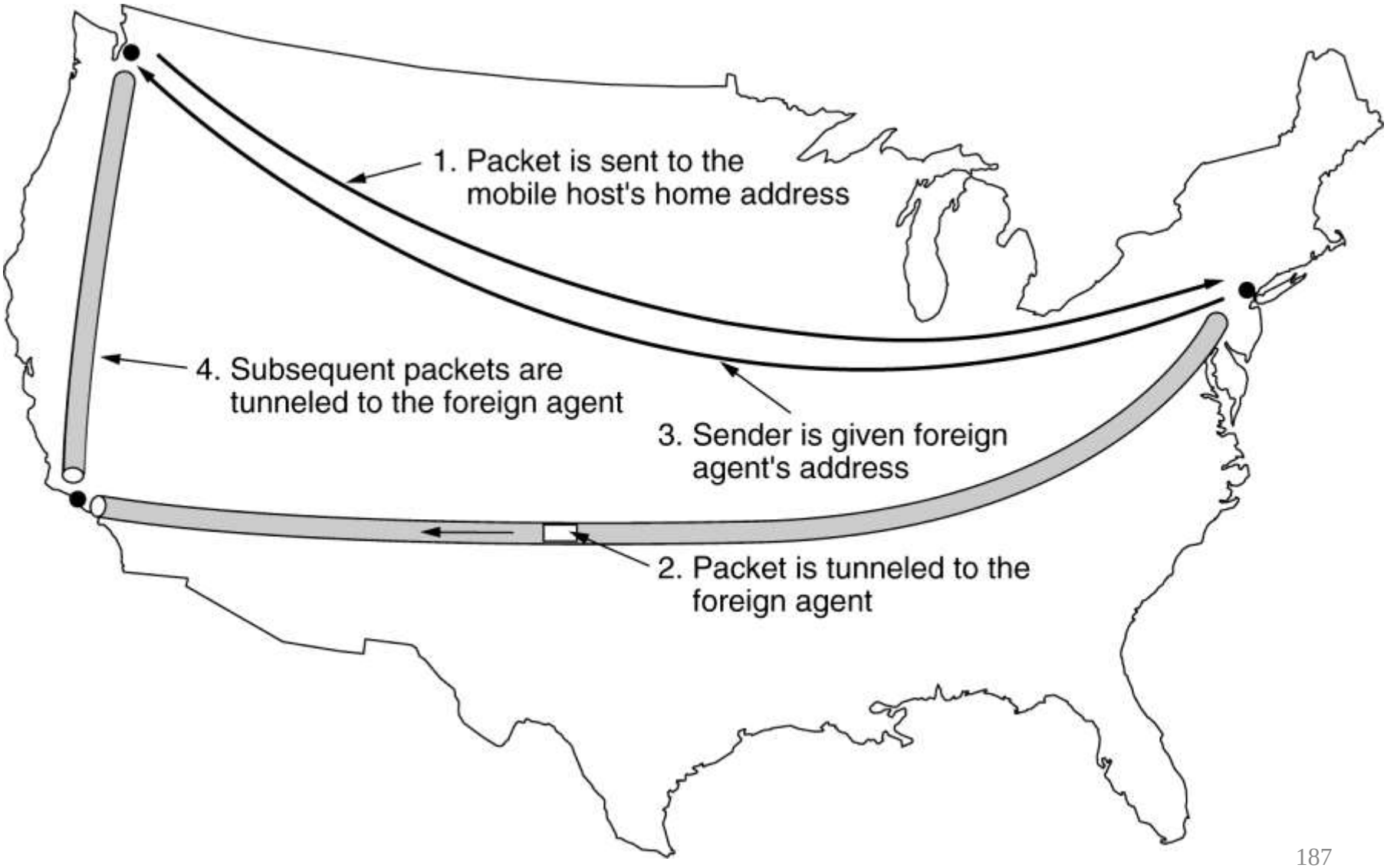
The Network Layer in the Internet: Mobile IP

- Every site that wants to allow its users to roam has to create a **home agent**.
- Every site that wants to allow visitors has to create a **foreign agent**.
- When a mobile host shows up at a foreign site,
  - it contacts the foreign agent there and register.
  - The foreign agent then contacts the user's home agent and gives it a care-of address, normally the foreign agent's own IP address.

The Network Layer in the Internet: Mobile IP



The Network Layer in the Internet: Mobile IP



Recommended Exercises

In 4th Edition：

- 1、6、9、17、22、23
- 27-30、34、36、38-44

In 5th Edition：

- **6, 23-33, 40-42**