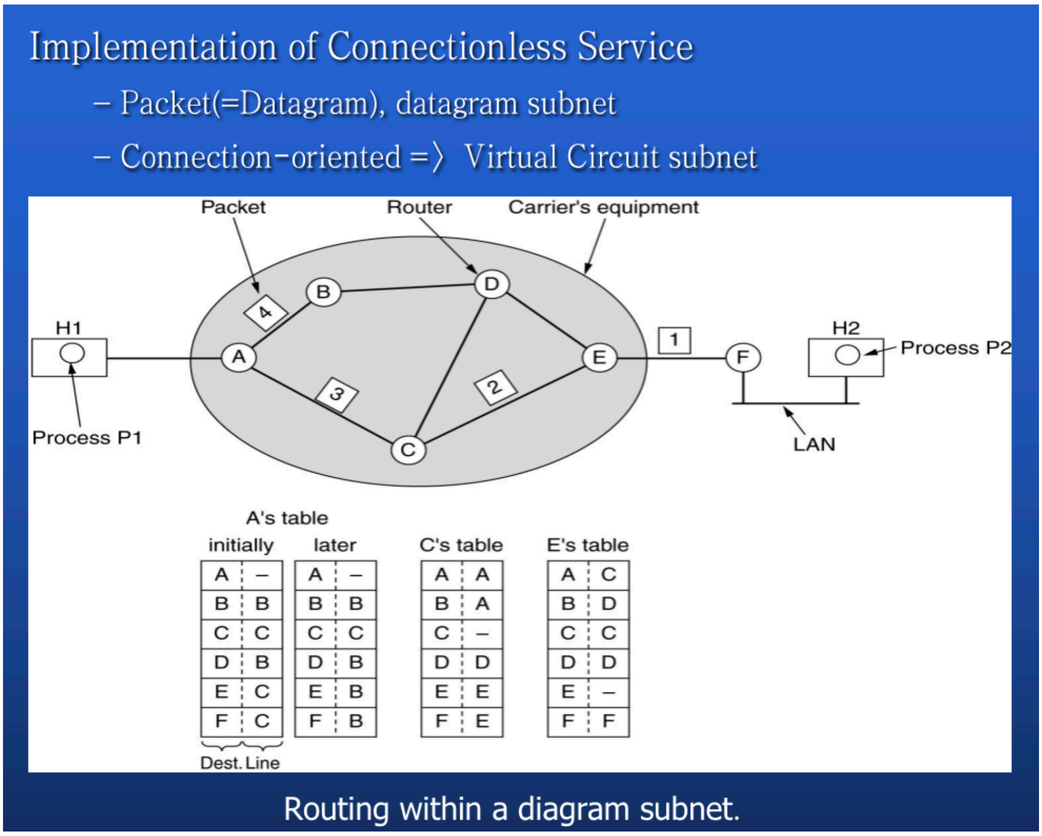
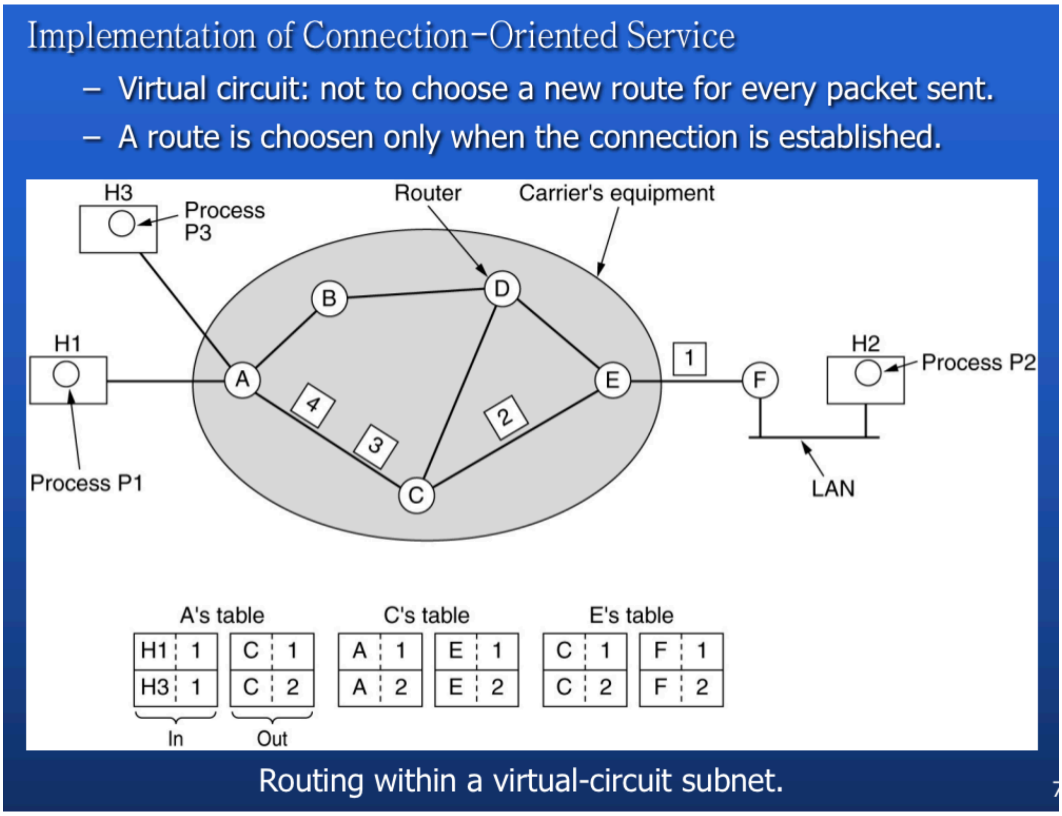
1 Network Layer design issues

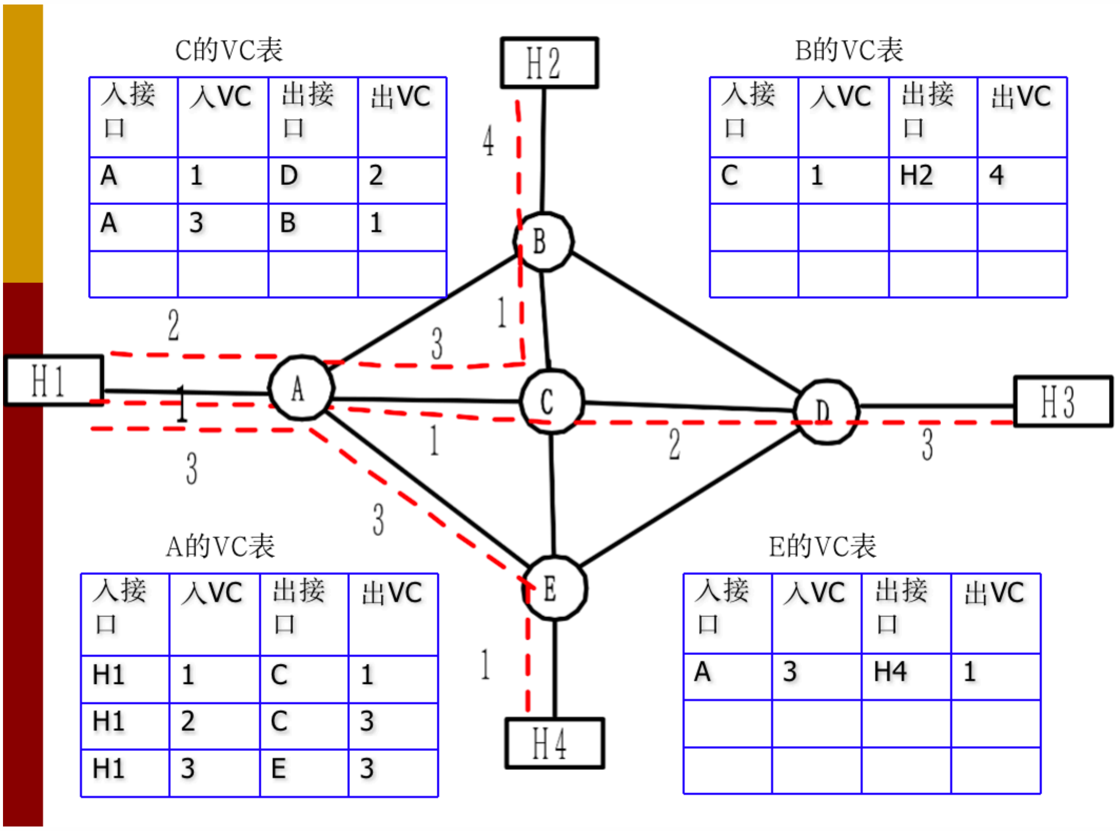
1.1 Internet Community: Connectionless service



- 路由表：从某个站点发送数据包到另一个站点要走哪个站点过
- 如果ping出问题，那要查一下路由表

1.2 Telephone companies: Connection-oriented service





- 直接存了走的路径，对输入的接口和VC查找目标转发接口和VC

1.3 Comparison

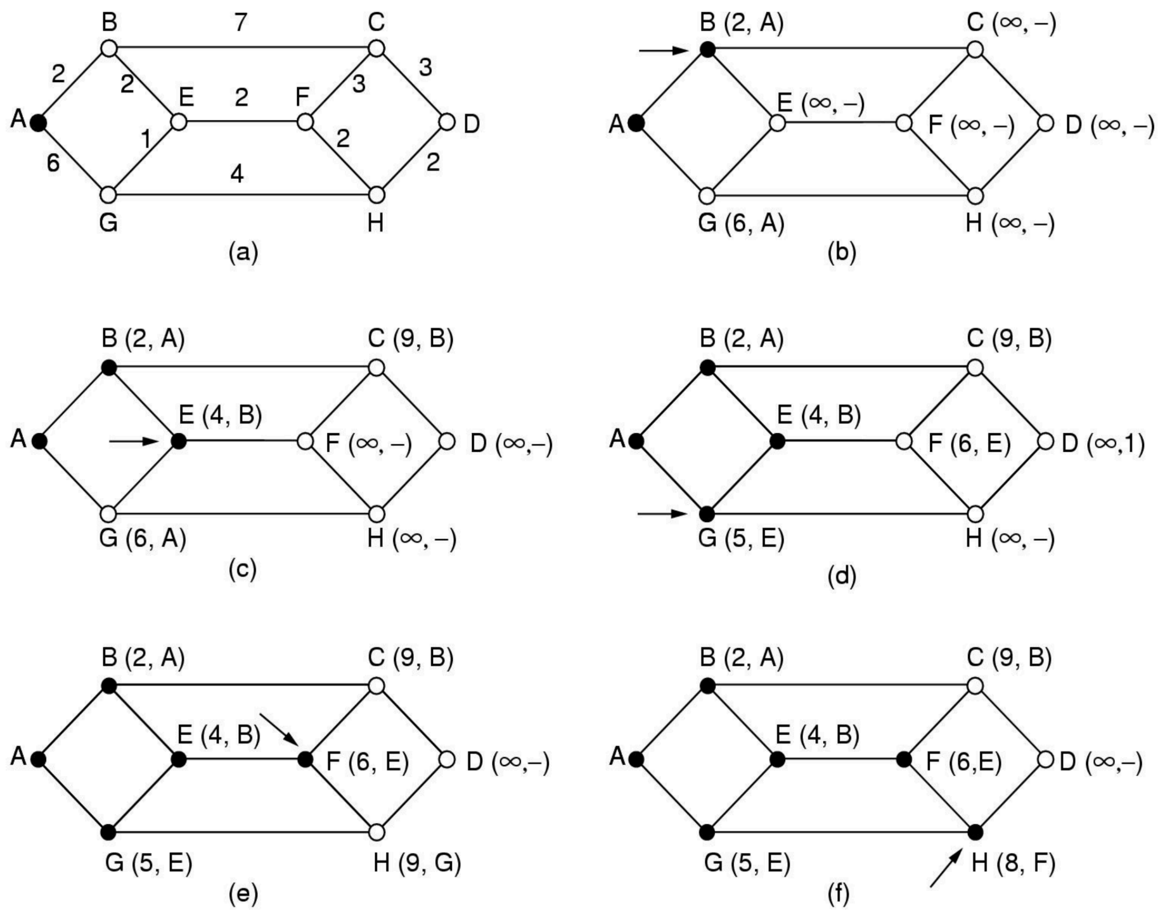
Issue	Datagram subnet	Virtual-circuit subnet
Circuit setup	Not needed	Required
Addressing	Each packet contains the full source and destination address	Each packet contains a short VC number
State information	Routers do not hold state information about connections	Each VC requires router table space per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

Comparison of Virtual-Circuit and Datagram Subnets

2 Routing Algorithms 路由算法

- Forwarding: to handle each incoming packet, look up the route table,then forward to a output line.（查找转发）
- Filling and updating the route table.

2.1 Dijkstra’s algorithm



The shortest path from A to D - From source to destination：

- Step 1(initialization): Marking node A as permanent,indicated by a filled-in circle.
- Step 2: relabeling each node that is adjacent to the newly marked node. Two values are rlabeled: the distance to the source node along the best known path, and the next node to reach to the source node.
- Step 3: selecting one with the smallest label in all tentative nodes (empty circles) as a permanent node, and mark it by a filled-in circle.
- Go to step 2 until destination node is marked as permanent.
- the best known path is the better path between the old path and the new path via the newly marked node

2.2 Flooding 洪泛法

- 在修改之前不好，因此要对其改造：
  - Solution 1: A hop counter contained in the header of each packet，packet is discarded when the counter reaches zero.
  - 方法1: 设定最高的转发次数

- Solution 2: Every node keeps track of which packets have been flooded, to avoid sending them out a second time.
- 方法2: 每个站点都会登记转发过的数据包的标识（谁发的，发给谁），如果该包已经登记过了，那就扔掉，否则就登记转发，避免了兜圈子的问题。这个包会到达每个路由器一次，不会多的。

2.3 Distance Vector Routing

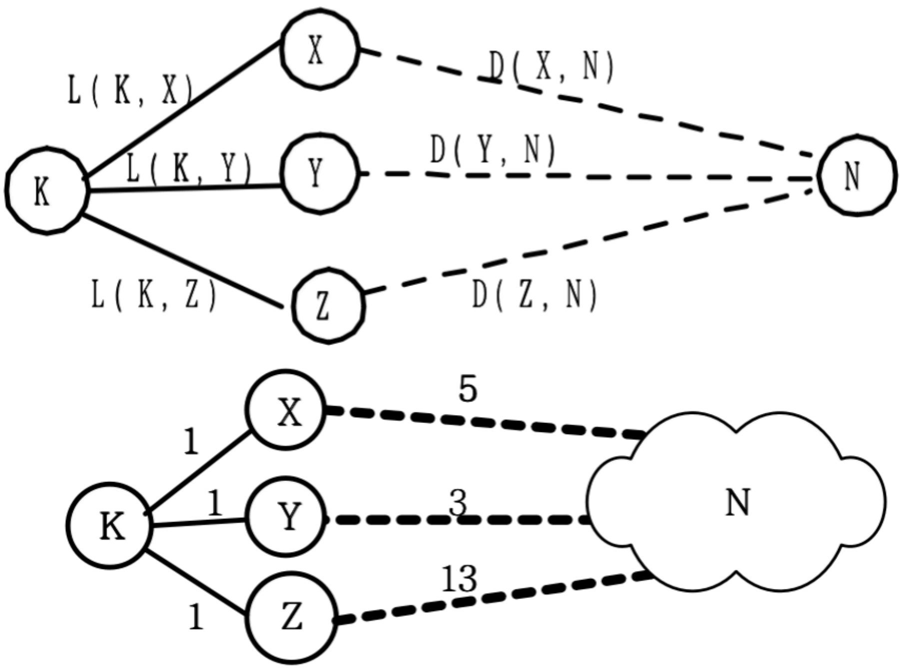
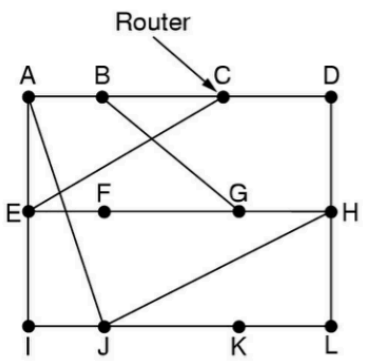


fig k7-4 RIP的工作原理

- 算法描述： -[see fig 5-9]
- 假定以延迟作为度量标准。
- 每隔T ms，任一路由器把它估计的到每个目的路由器的延迟清单送给每个邻居路由器，它也从每个邻居处收到一份类似的清单。
- 某一路由器 K，收到所有相邻路由器（设为 X,Y,Z)的路由表后，更新到某一路由器（设为 N ）的路由：
  - 记D（I，J）为路由器 I 至目的路由器J 的距离，L(K,X)、 L(K,Y)、 L(K,Z)分别为K 至邻居X、Y、Z的延迟.
  - K 经X 至路由器 N 的距离为 L(K,X)+D(X,N）
  - K 经Y 至路由器 N 的距离为 L(K,Y)+D(Y,N）
  - K 经Z 至路由器 N 的距离为 L(K,Z)+D(Z,N）

- L(K,X)+D(X,N)、 L(K,Y)+D(Y,N)、L(K,Z)+D(Z,N) 三者中的最小值就是K 至N 的最短距离。

- 即使路由器K 中老的 D(K,N) 比新的D(K,N)更小，D(K,N)也要更新。
- K对所有其他路由器进行如同路由器N 的更新运算。



(a)

JAC, JIC, JHC, JKC

8+25, 10+18, 12+19, 6+36

33, 28, 31, 42

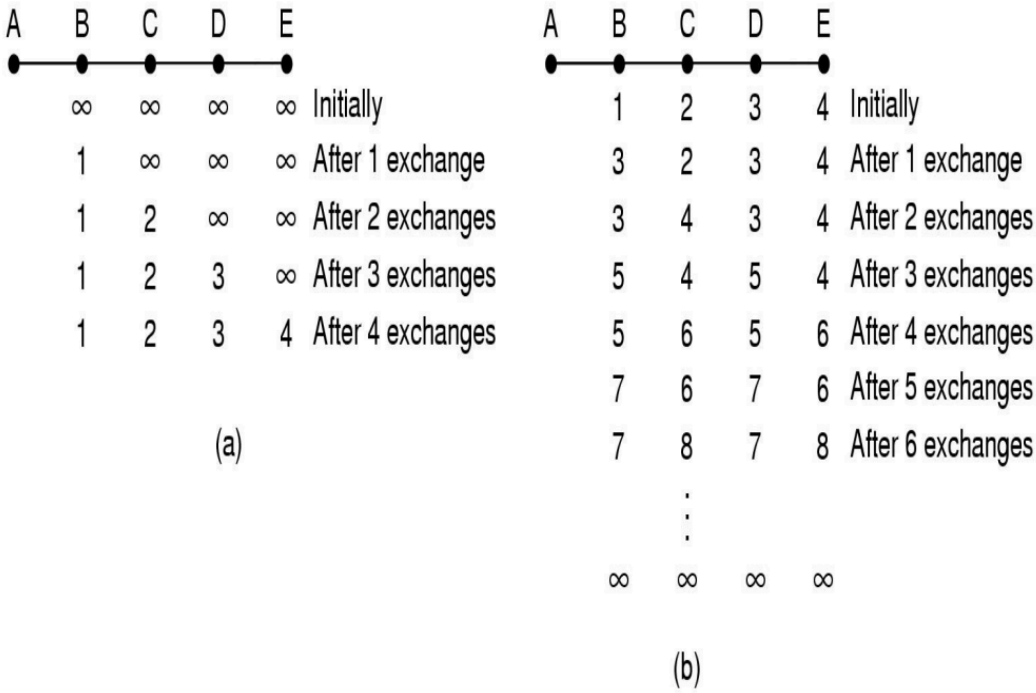
To	A	I	H	K	New estimated delay from J	Line
A	0	24	20	21	8	A
B	12	36	31	28	20	A
C	25	18	19	36	28	I
D	40	27	8	24	20	H
E	14	7	30	22	17	I
F	23	20	19	40	30	I
G	18	31	6	31	18	H
H	17	20	0	19	12	H
I	21	0	14	22	10	I
J	9	11	7	10	0	–
K	24	22	22	0	6	K
L	29	33	9	9	15	K
	JA delay is 8	JI delay is 10	JH delay is 12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

Vectors received from J's four neighbors

(b)

- J要找通路，从他邻居入手，找通过邻居的点转发的最短路径，每30s更新一次，更新一次路由表

2.3.1 RIP Drawback



- it converges（收敛） slowly：（good news travels rapidly, bad news travels slowly）好的消息走得快，坏的消息走得很慢
  - （a）中的是建立连接，很快就好了
  - （b）中的A-B的网络连接断掉了，因此信息不断更新，一直到无穷大才会判断出来，会花大量的时间

2.4 Link state routing 链路状态路由算法

2.4.1 Link state routing Principle（for each router）

1. Discover its neighbors and learn their network address 发现邻居，学习其网络地址
2. Measure the delay or cost to each of its neighbors 确定时间延迟等
3. Construct a packet telling all it has just learned 建立一个路由器刚掌握知识的包
4. Flood this packet to all other routers 把这个新包洪泛发送给所有路由器
5. Compute the shortest path to every other router 重计算最短距离

2.4.2 实现

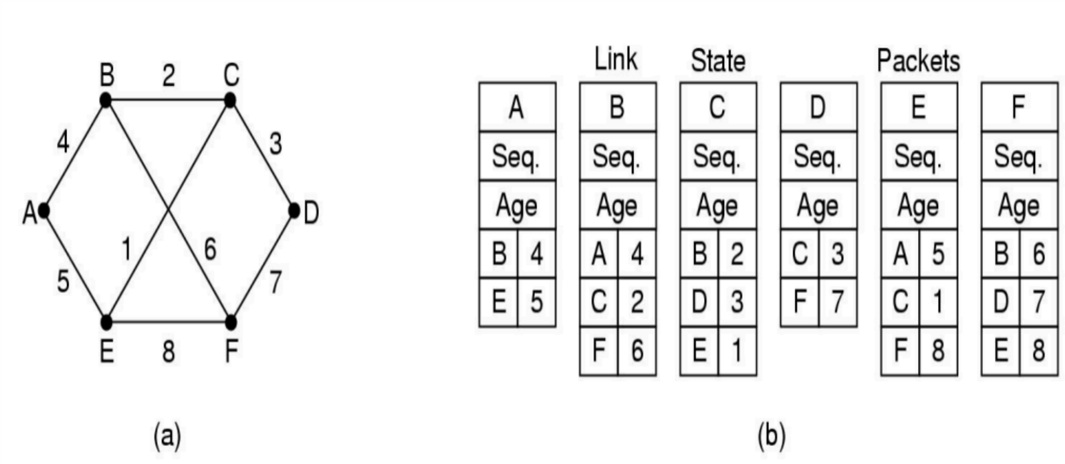
2.4.2.1 Learning about the neighbors

- Accomplishes this goal by sending a special **<u>Hello packet</u>**, and neighboring router **<u>sends back a reply</u>**.

2.4.2.2 Measuring delay or line cost.

- Send ECHO packet，**<u>round-trip time/2</u>**
- Whether or not to take the load into account? better performance, but route table may oscillate 振荡。最短通路会因负载变化而改变，出现振荡，很可能会出现兜圈子的问题。
  - solution: spread load over both lines.（？？？）

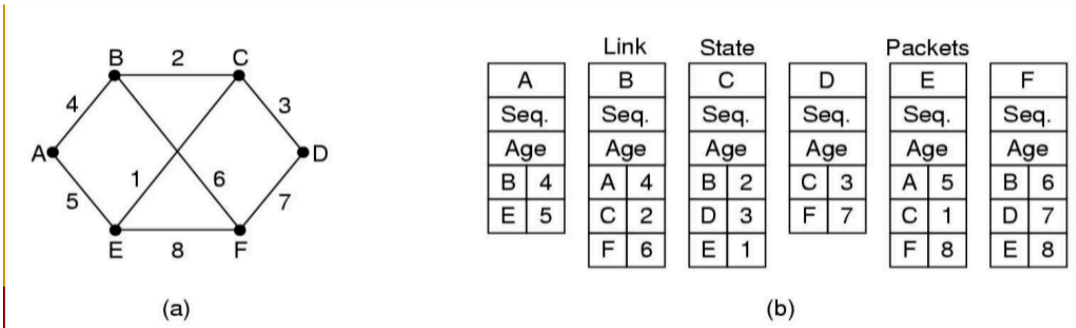
2.4.2.3 Building the link state packets 链路状态包



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

- When to build?
  - periodically（周期性地） or
  - when some significant event (increase or decrease of lines or routers) occurs（网络的拓扑结构改变，路由器上电or断线了）. Used in OSPF

2.4.2.4 Distributing link state packet

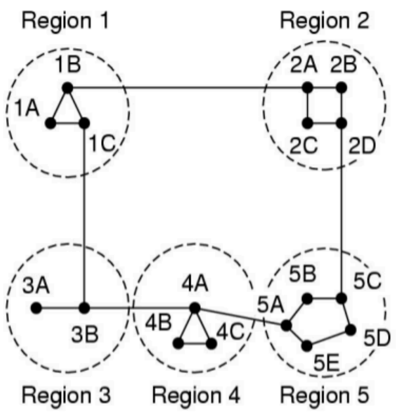


			Send flags			ACK flags			
Source	Seq.	Age	A	C	F	A	C	F	Data
A	21	60	0	1	1	1	0	0	
F	21	60	1	1	0	0	0	1	
E	21	59	0	1	0	1	0	1	
C	20	60	1	0	1	0	1	0	
D	21	59	1	0	0	0	1	1	

- 注意，上图是B的packet buffer！！！是以A为源的发送而来的包，seq是收到的包的序号，age是最大可停留时间
- Fundamental ideas：use flooding to distribute the link state packet,**<u>each packet contains a sequence number that is incremented for each new packet sent</u>**. 用洪泛法发送链路状态包，每次发送的时候，该链路状态包都会增加一次sequence number。
- If an packet arrives with seq number <= early incoming seq, it is discarded. 如果接收到的包的序号比自己之前收到的小，那么就丢掉，更新的包的序号会一直递增。
- Problems
  - Sequence numbers wrap around. 如果数字太小，那么序号就会到一定值开始归0，导致误扔。
    - Solution: 32bit sequence number. One packet/second, then 137 year. 设置足够大的序号。
  - If Router crash, it will lose track of its sequence number 如果断电，归0了。
    - age是会不断减小的，如果60s内的包都没有更新，那么这一行的记录就会被删掉，以防掉电等情况，0号包无法重新进来。
  - If Sequence number's transmission is wrong
    - Solution
      - Include the age field in each packet, the age is recorded in a table. -[see fig 5-14]
      - larger sequence number used before router crashes, or transferred incorrectly, will be obsolete.

2.5 Hierarchical routing

- Networks grow, the router tables grow proportionally
- Routers divided into regions-[see fig 5-15]



(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)

Hierarchical routing *

- 对不同的地区分成区域，这样路由表就会被大大压缩，只用看dest（destination）的区域在哪

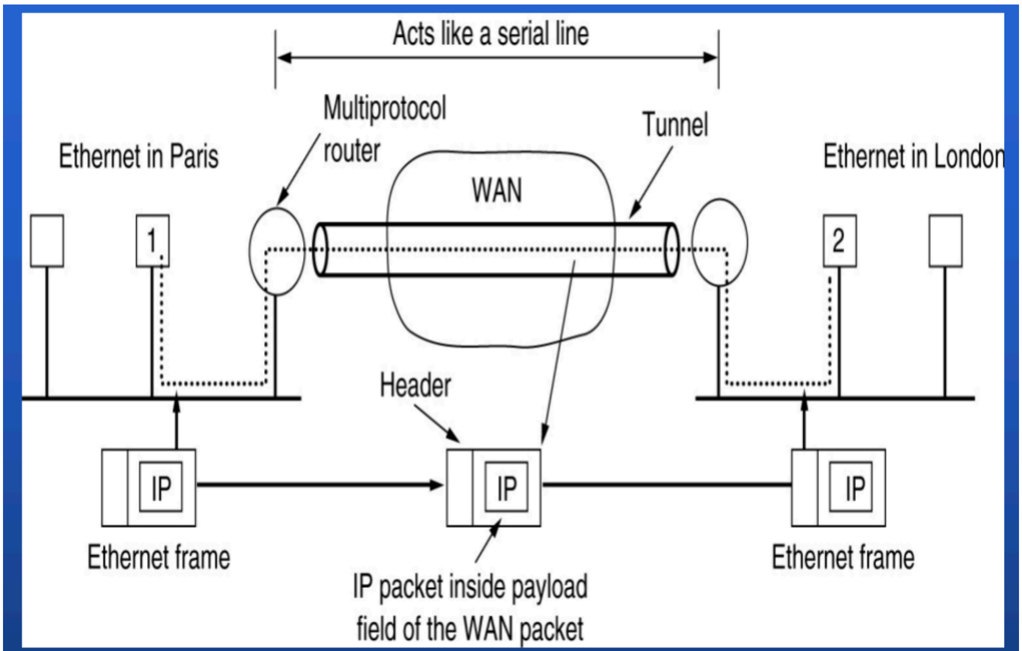
3 QUALITY OF SERVICE

- 比较少考，可以翻一下ppt
- 5.3.2 不考

4 Internetworking

4.1 Tunneling 隧道

- **Handling the general case of making two different networks interwork** is exceedingly difficult.
  - An example tunneling a packet from Paris to London -[See Fig 5-47]
- The solution to this problem is a technique called **tunneling**.
- The WAN can be seen as a big tunnel extending from one multiprotocol（多协议的） router to the other.
- An analogy may make tunneling clearer. -[See Fig 5-48]



Tunneling a packet from Paris to London.

4.2 Internetwork Routing

- Routing algorithms:
  - A two-level routing algorithm
    - Interior gateway protocol (这里“网络”实际上是指自治系统）内部网关协议，只连内部的自

治系统，比如A4

- Exterior gateway protocol，外部网关协议，就是路由器连另一个自治系统，比如A2
- AS (autonomous system)

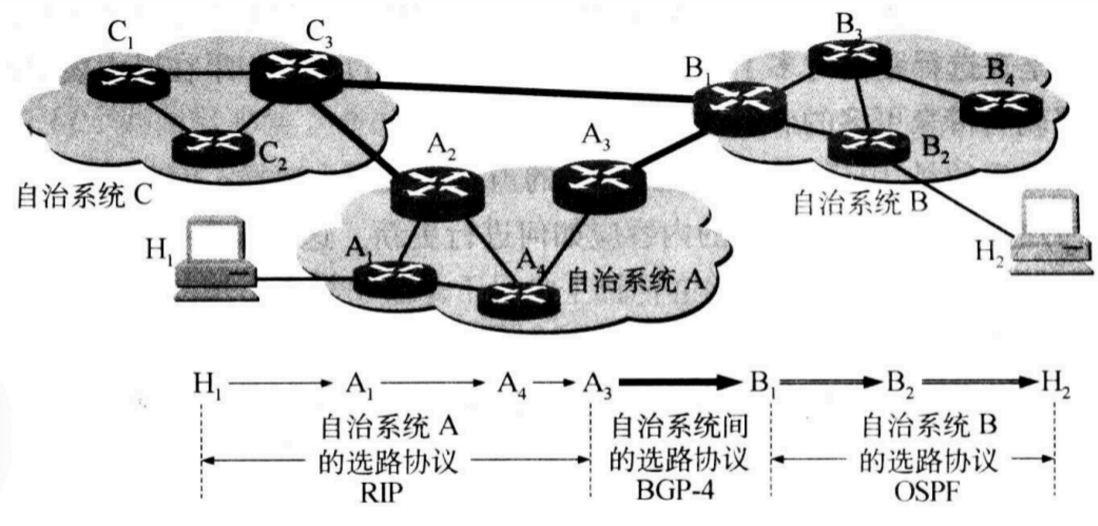


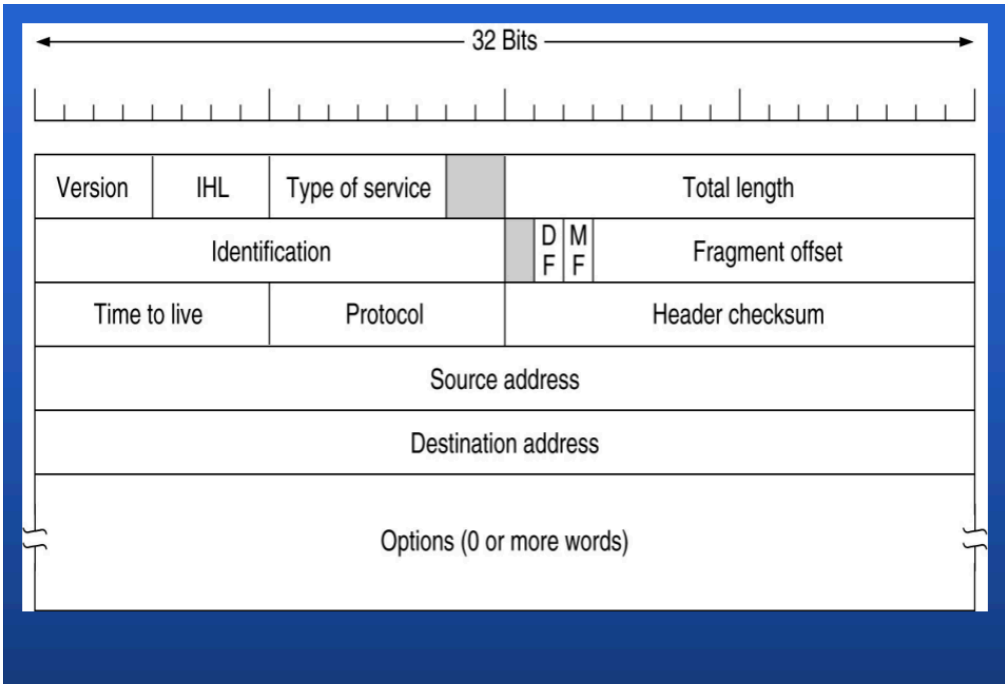
图 7-14　自治系统和内部网关协议、外部网关协议

- RIP 就是类似于Dijkstra算法；OSPF

5.4.3-5.4.5简单考

5 The Network Layer in the Internet

5.1 IP Version 4 Protocol（IPv4）20+x



The IPv4 (Internet Protocol) header.

- version：0100，就是IPv4；0110，就是IPv6
- **首部长：20字节+option的字节**
- 第2行是分片有关的。比如一个数据包2000字节，但是以太网的数据包最大1500字节，所以要分片。DF为0，是分片的意思。MF（more fragment）：前面的分片都是1，最后一个片就是0，意思是说后面没有分片了。offset是-1、-2、-3...，用来区分分片的顺序
- protocol：是指数据字段是什么。**protocol: 1--ICMP, 6--TCP, 17--UDP, 89--OSPF, ...**
- 中间转发的路由器不会改变source和destination的内容，只会抓出目的地址然后查路由表

- （补充）例如：原始数据报首部长度28字节，数据长度410字节。将此数据报分成3个报片，其数据长度分别为160,160,90字节，这样一来就能通过最大允许的数据报总长为28+160=188字节的网关，
- 按照格式(IHL，Total Length，Identification,MF,DF, Fragment Offset),
  - 原始数据报的上述值为：（7，438，51037，0，0，0)
  - 3个报片的上述值为：

    1#报片：（7，188，51037，1，0，0)

    2#报片：（7，188，51037，1，0，20)

    3#报片：（7，118，51037，0，0，40)

- **Fragment Offset 是以8个字节为单位，因此第二个报片偏移量是160字节，offset是20。**
- IHL、Identification都会继承下来
- MF；**DF = 0 代表有分片**

- header中的校验和、TTL（time to live，其实存的是条数）要改变（-1）
- 这种题有可能会考

5.2 IP Addresses

5.2.1 分类 ABCDE

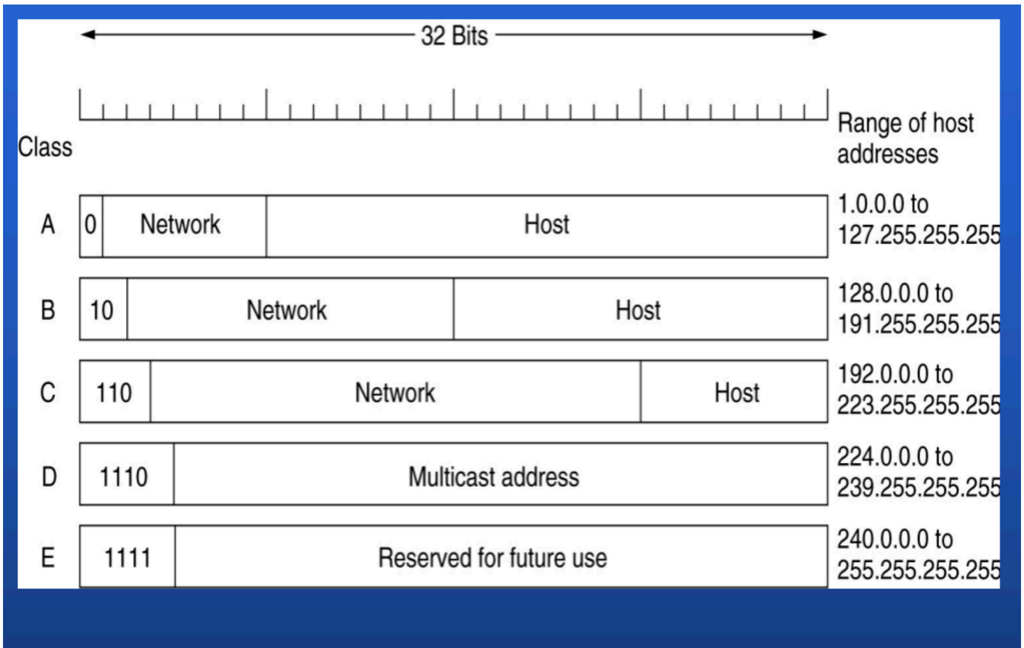
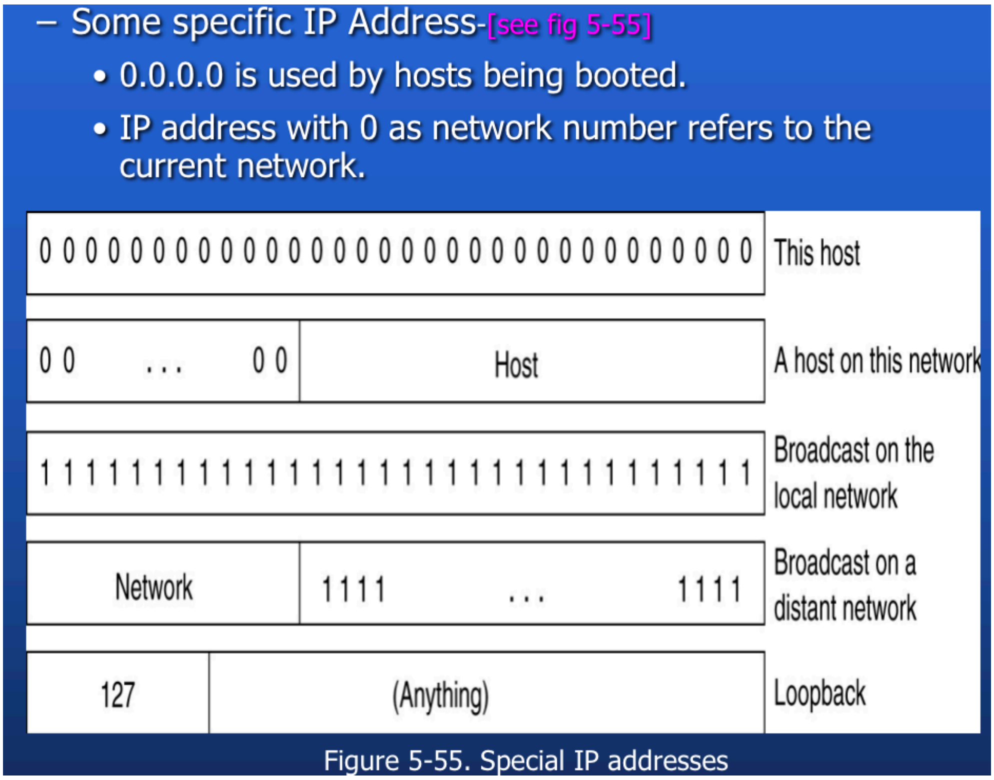


Figure 5-54. IP address formats

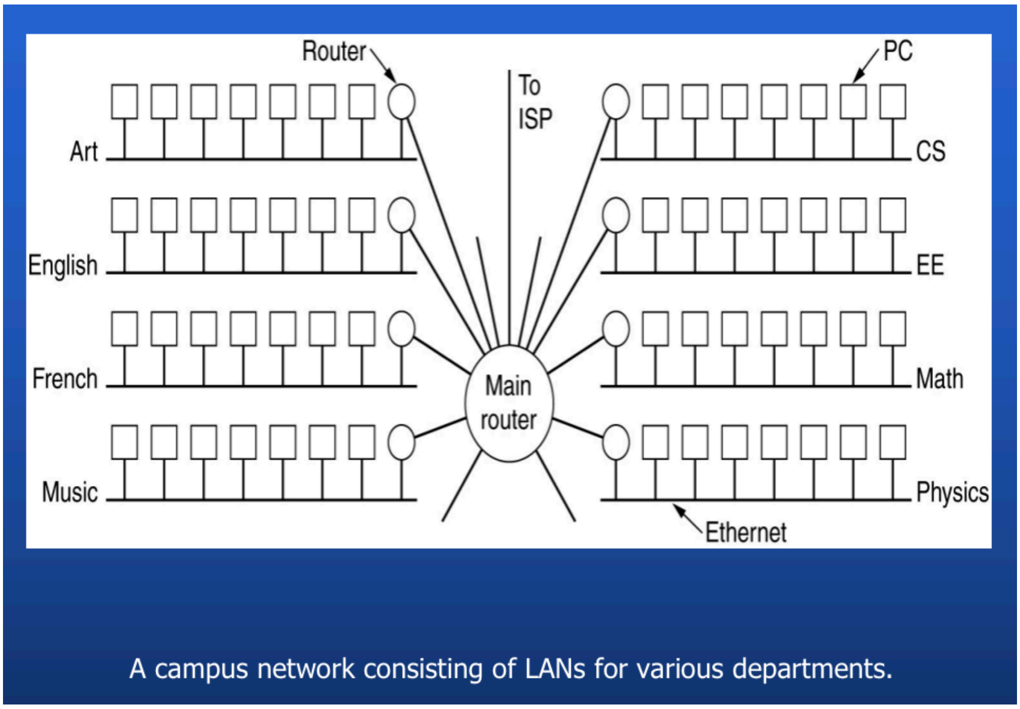
- 根据class的序号不同，分为5类（1-127；128-191；192-223；224-239；240-255）；
- A类的网络云最大，主机最多。
- ps：host / network 全0或者全1都是不能用的，都是有特殊分配的。
- ps：一朵网络云里是没有路由器的

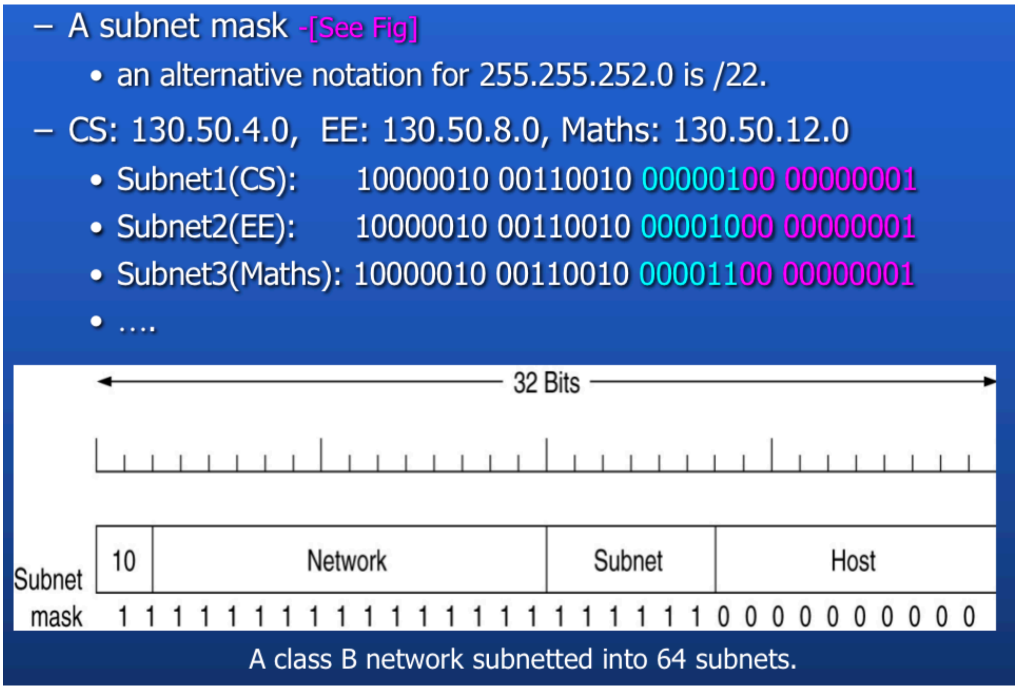
5.2.2 Some Specific IP Address



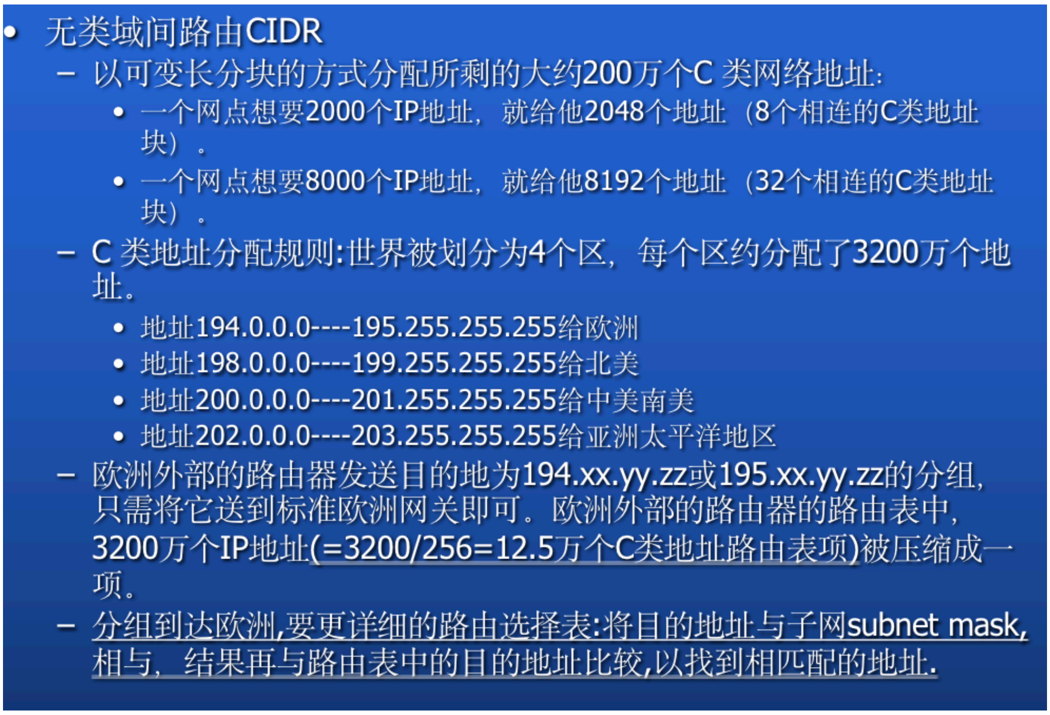
- 一些特定地址
- 127开头的是本地回环接口 Loopback

5.2.3 子网划分



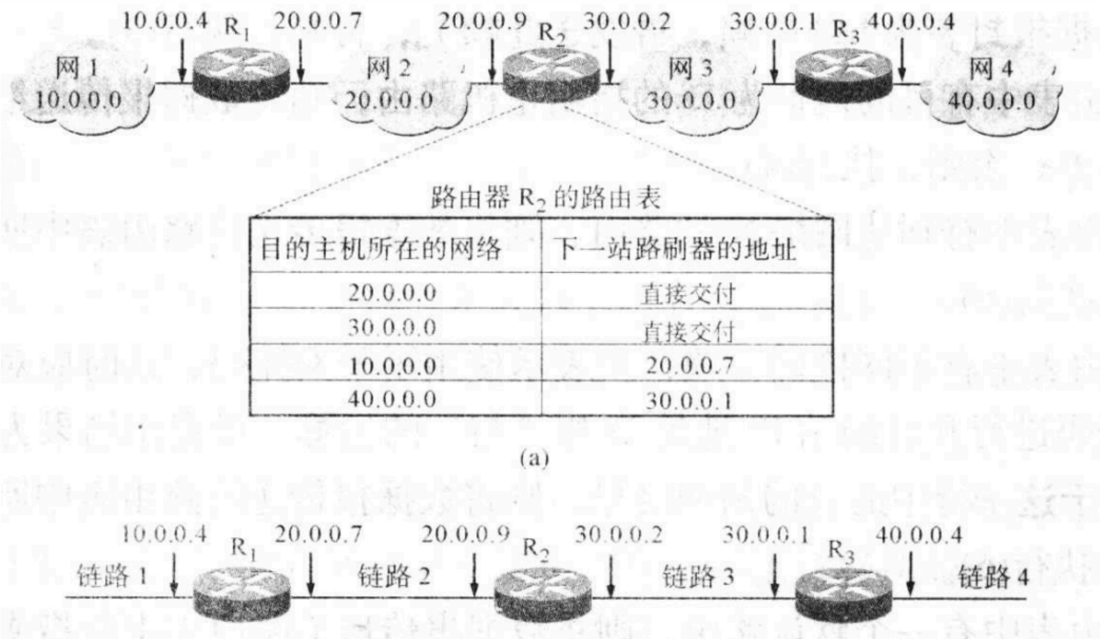


- 就是把host里前面一部分划分成网络号，以解决网络号划出来的主机太多的问题，这样就可以划出来更多的网络。



- 但是问题：<u>C类地址很多，网络号太多，对路由器来说很费力。所以把C类地址按区域分给不同地区。</u>

5.2.4 IP层处理数据报的流程



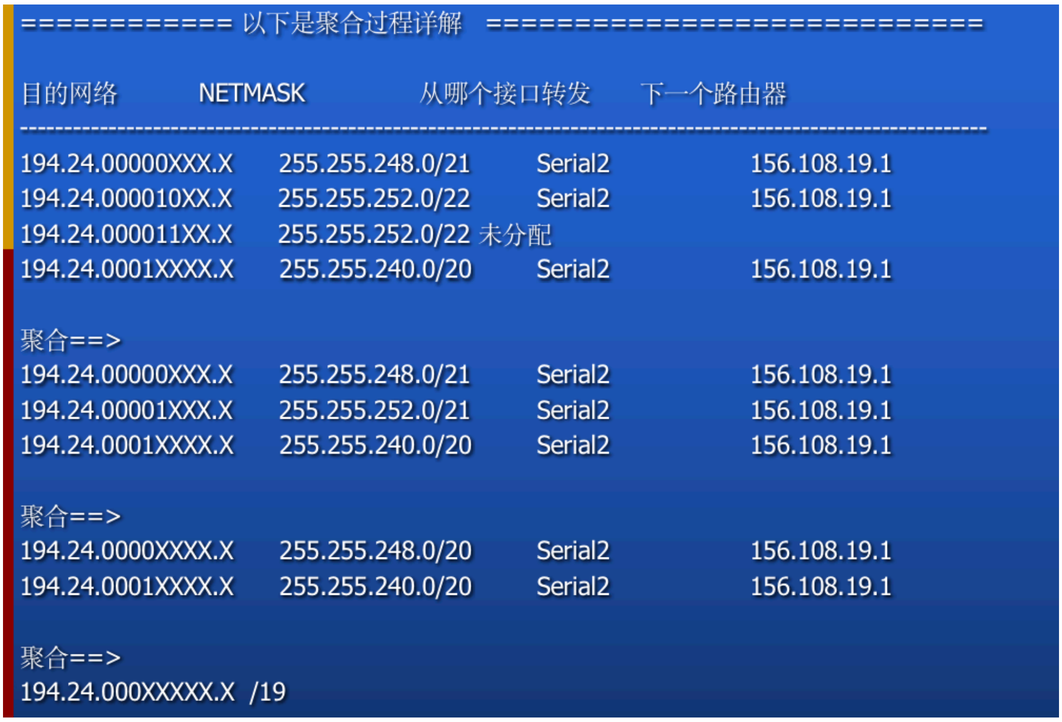
路由器R2的路由表举例：
(a)路由器R2的路由表；(b)将网络简化为一条链路 *

- 实例：
- 主机A，IP地址是20.2.2.2，位于网2内，其TCP/IP配置的缺省网关是20.0.0.9(即路由器R2）。就是要验证**destination**的**IP**是不是和本机处于同一个物理网络；如果不是，那么就把数据包发给缺省网关（路由器上的），下面由路由器查路由表发消息。
- 主机B，IP地址是40.4.4.4，位于网4内。
- **A**向**B**发送一个**IP**数据报**X**，数据报**X**的目的地址是**40.4.4.4**，不在**A**所在的网**2**内，于是发给缺省网关**R2**。数据报**X**的目的地址**40.4.4.4**不变。
- 路由器R2查路由表，然后转发给路由器R3。数据报X的目的地址40.4.4.4不变。
- 路由器R3查路由表，直接发送给主机B。
- 注意转发到目的地址的消息。下一站的路由器的接口要和本路由器的传出接口在同一个网络比如**30.0.0.2** 发到 **30.0.0.1**，而不能发到**40.0.0.4**

5.2.5 最长掩码的路由匹配优先规则

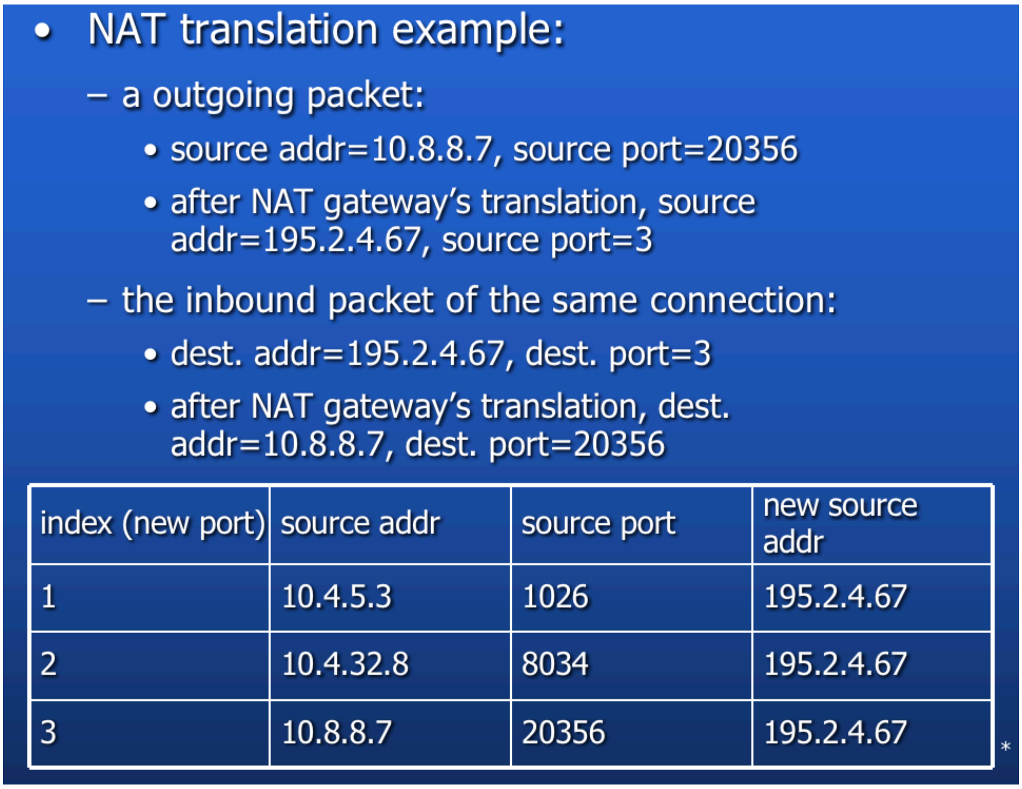
- 如果有多个匹配，子网掩码越长的那个会被选中，也就是越匹配的越会被选中。

5.2.6 route aggregation 路由聚合



- 路由器中的路由表太大了，就会做合并的过程，X类似卡诺图任意项，这样就可以将不同的目的网络地址都合并成一条路由表信息

5.2.7 NAT - Network Address Translation （? ? ? ）



- IP地址非常稀缺。只要一个IP地址，可以分给校内的2万台主机用。内网里有不同的分配的地址，但是出了内网，那么自己的这个源地址就只有一个，也就是之前分配的。由源端口号来使之unique
- 这种实际上就是一种翻译改造。
- 出去和进来的包的目的源地址&端口号是会反一下的

分配的IP地址 校内的IP地址	端口号	真正的IP地址 only one	端口号
10.8.8.7	20356	195.2.4.67	3
10.4.32.8	8034	195.2.4.67	2

5.3 Internet Control Protocols

5.3.1 ICMP

- The internet control message Protocol (ICMP)

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 request	Ask a machine if it is alive
Echo reply	Yes, I am alive
Timestamp request	Same as Echo request, but with timestamp
Timestamp reply	Same as Echo reply, but with timestamp

Figure 5-61 The principal ICMP message types

5.3.2 ARP -- 求MAC地址

IP地址--->硬件地址（MAC地址）的转换是由地址解析协议ARP（Address Resolution Protocol)来完成。

- 每一个主机都有一个ARP高速缓存(ARP cache)，当主机A欲向本局域网上的主机B发送一个IP数据报时，先在其ARP高速缓存中查看有无主机B的IP地址。如有，就将对应硬件地址写入MAC帧。
- 也有可能ARP cache中查不到主机B的IP地址。这可能是主机B才入网，也可能是主机A刚刚加电，其高速缓存还是空的。在这种情况下，主机A就自动运行ARP:
- ①ARP进程在本局域网上广播发送一个ARP请求分组，上面有主机B的IP地址。
- ②在本局域网上的所有主机上运行的ARP进程都收到此ARP请求分组。
- ③主机B在ARP请求分组中见到自己的IP地址，就向主机A发送一个ARP响应分组，上面写入自己的硬件地址。
- ④主机A收到主机B的ARP响应分组后，就在ARP高速缓存中写入主机B的IP地址到硬件地址的映射。

- ARP是广播包
- MAC地址:

> MAC地址也叫物理地址、硬件地址，由网络设备制造商生产时烧录在网卡(Network lnterface Card)的EPROM(一种闪存芯片，通常可以通过程序擦写)。IP地址与MAC地址在计算机里都是以二进制表示的，IP地址是32位的，而MAC地址则是48位的 。
>
> MAC地址的长度为48位(6个字节)，通常表示为12个16进制数，如：00-16-EA-AE-3C-40就是一个MAC地址，其中前3个字节，16进制数00-16-EA代表网络硬件制造商的编号，它由IEEE(电气与电子工程师协会)分配，而后3个字节，16进制数AE-3C-40代表该制造商所制造的某个网络产品(如网卡)的系列号。只要不更改自己的MAC地址，MAC地址在世界是唯一的。形象地说，MAC地址就如同身份证上的身份证号码，具有唯一性 。
>
> 寻址协议层不同。IP地址应用于OSI第三层，即网络层，而MAC地址应用在OSI第二层，即数据链路层。 数据链路层协议可以使数据从一个节点传递到相同链路的另一个节点上（通过MAC地址），而网络层协议使数据可以从一个网络传递到另一个网络上（ARP根据目的IP地址，找到中间节点的MAC地址，通过中间节点传送，从而最终到达目的网络）

5.3.3 代理ARP

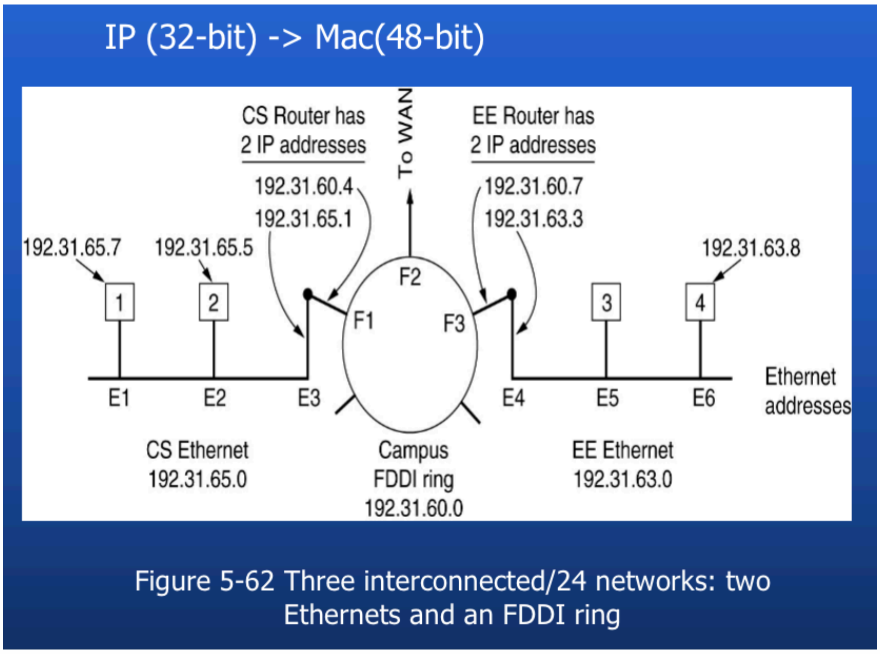
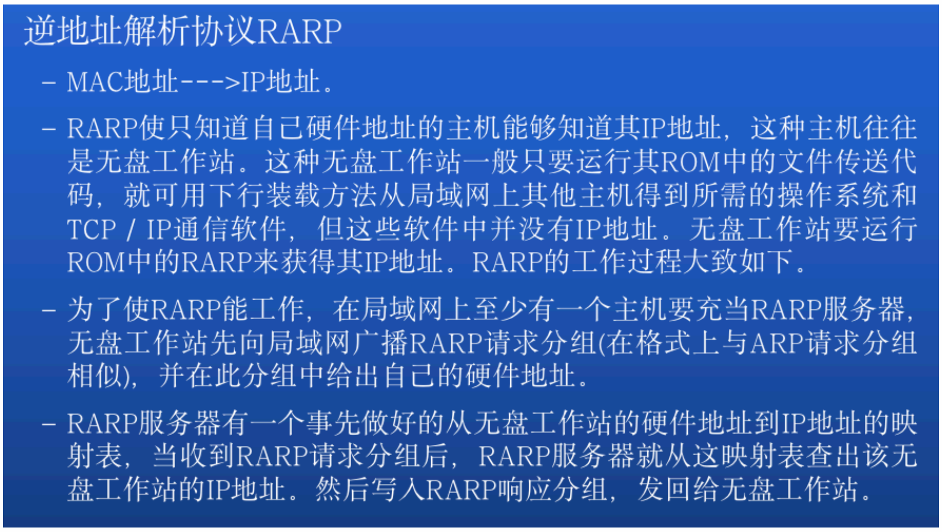


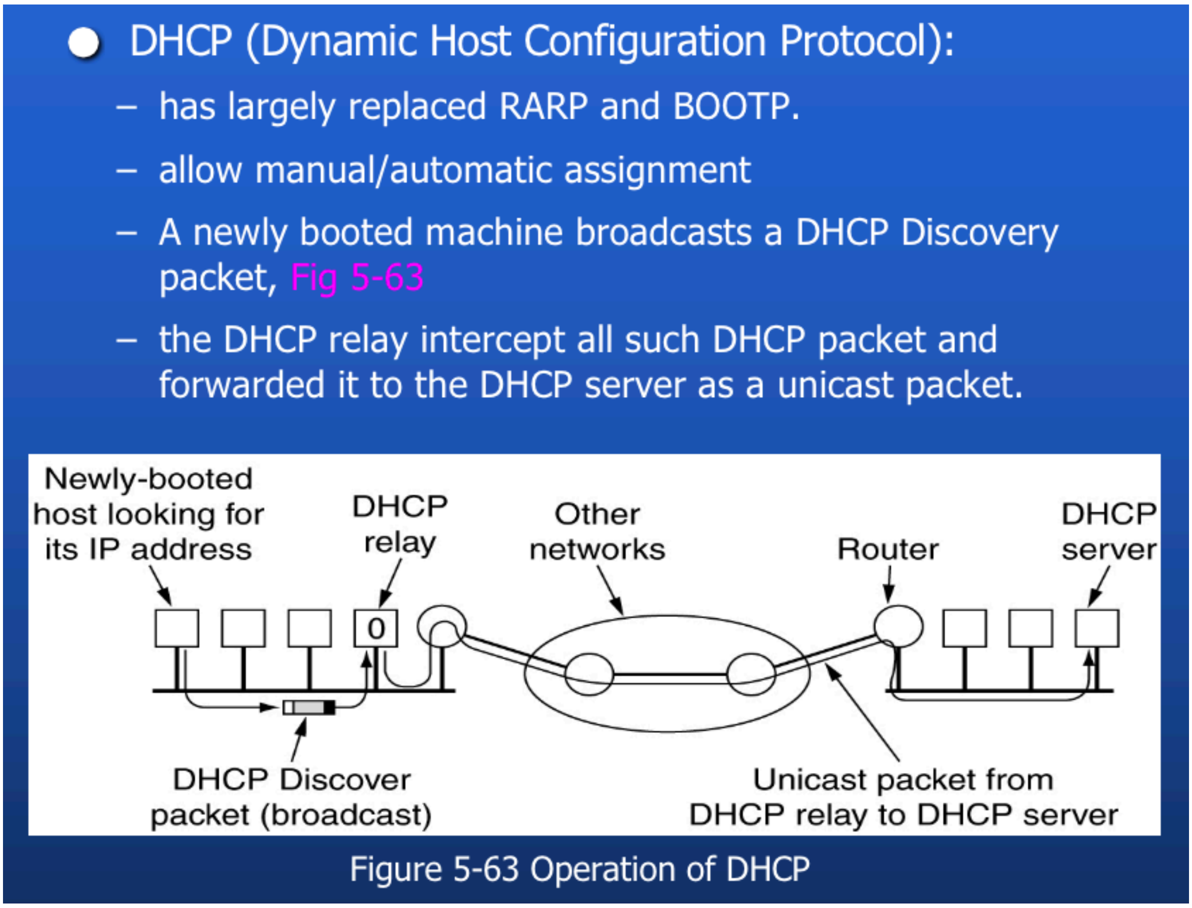
Figure 5-62 Three interconnected/24 networks: two Ethernets and an FDDI ring

- 比如在校园网内，如果目标地址不是本地局域网中的，那么路由器会出来工作，告诉源主机目标地址的MAC地址，就是路由器出来说的，而不是目标主机来说。

5.3.4 RARP



5.3.5 DHCP - Dynamic Host Configuration Protocol



- 就是动态分配IP地址。在上电后，会广播请求DHCP服务器分配一个IP地址给自己。

5.4 OSPF – The Interior Gateway Routing Protocol (? ? ?)

- Three kinds of routes： intra-area、interarea and interAS
- Distinguishes four classes of routers, they can be overlapped.–[see fig 5-53]
  - Internal routers are wholly within one area.
  - Area border routers connect two or more areas.
  - Backbone routers are on the backbone.
  - AS boundary routers talk to routers in other Ases.
- OSPF Works by exchanging information between adjacent routers to update routes. Adjacent routers are not the same as neighboring routers.
- Designated router: It is inefficient to have every router on a LAN talk to every other router on the LAN, one router is elected as the designated router.It is said to be adjacent to any router on the LAN and exchange information with them.neighboring routers that are not adjacent don't exchange information.
- 5 types of OSPF messages–[see fig 5-54]

117

- OSPF就是会和相邻路由器的进行信息交换，相邻路由器并不等同于邻居路由器? ? ?

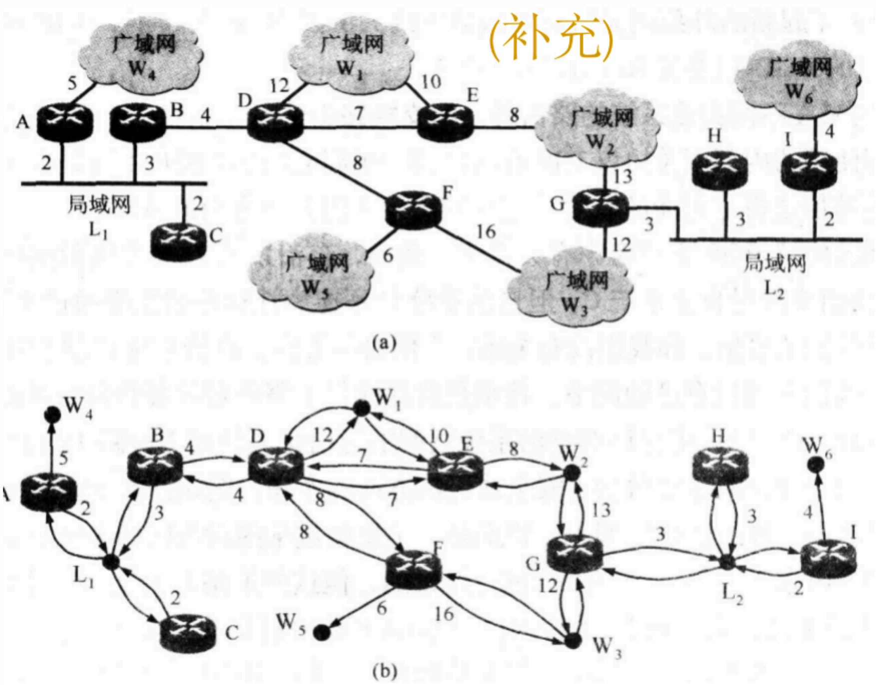


图7-19 OSPF支持的网络连接种类：(a)网络拓扑；(b)用有向图表示链路状态数据库

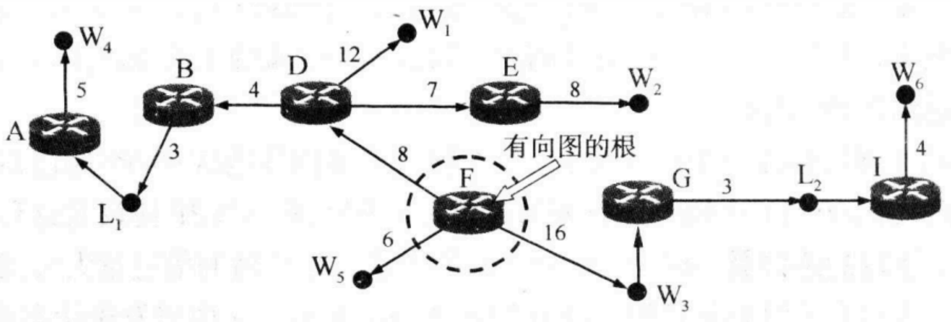


图 7-20 以路由器 F 为根的最短通路树（图 7-19 的例子）

- 将物理网络转换成一个有向图：找出最短生成树

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

Fig. 5-54. The five types of OSPF messages.

- 路由器刚上电会发送hello包
- 对于刚上电的路由器要建立路由表，database description和link state request，就是帮助本路由器从邻居路由器拿路由表和交换信息，更新其路由表信息。

5.5 BGP – The Exterior Gateway Routing Protocol（简单考）

- BGP(Border gateway protocol): used between ASes
- Have to worry about politics a great deal
  - Examples of routing constraints
- Policies are manually configured
- From the point of view of a BGP router, networks are grouped into one of three categories:
  - Stub network, which has only one BGP router to other AS.
  - Multiconnected networks: could be used for transit traffic, except that they refuse.
  - Transit networks: are usually backbones willing to handle third-party traffic.
- BGP is a distance vector protocol
- Keeps track of the exact path used-[see fig 5-67]
  - routing information from I and E are discarded
    - because they pass through F itself.
    - Thus count-to-infinity problem is solved.

- 可以选择性地避开一些国家地路由器（政治因素）

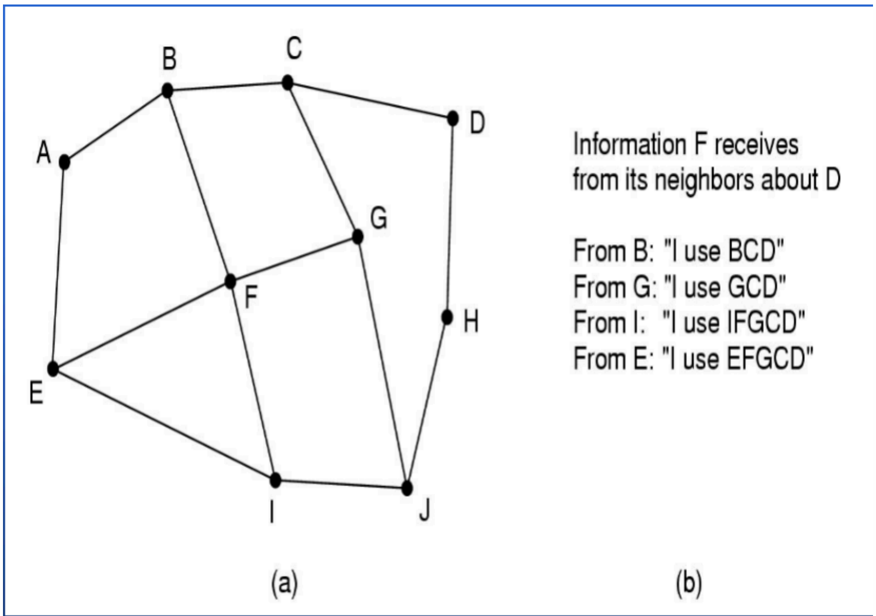


Figure 5-67. (a) A set of BGP routers. (b) Information sent to F

- 要发送一个信息给D，路由信息中对于需要经过自己的邻居给的information都扔掉，就是不采用这种路由路径。解决了坏消息传播慢，收敛慢的问题；解决了网络规模不能太大的问题。

5.6 Mobile IP

5.7 IPv6

- SIPP (Simple Internet Protocol Plus) namely IPv6:characteristic
  - Longer address
  - The simplification of the header
  - Better support for options
  - Security
  - More attention has been paid to type of service
- The main IPv6 header, fixed header -[see fig 5-68]
  - Traffic classes: identify different real-time delivery requirements.
  - Flow label: Each flow is designated by source IP addr, dest IP addr, and flow label.
  - Payload length: bytes count of payload, excluding fixed header
  - Next header: This field tell which of the 6 extended headers, if any, follows this one. If no extended headers follows, this field tells which transport protocol handler (TCP,UDP,...) to pass the packet to.

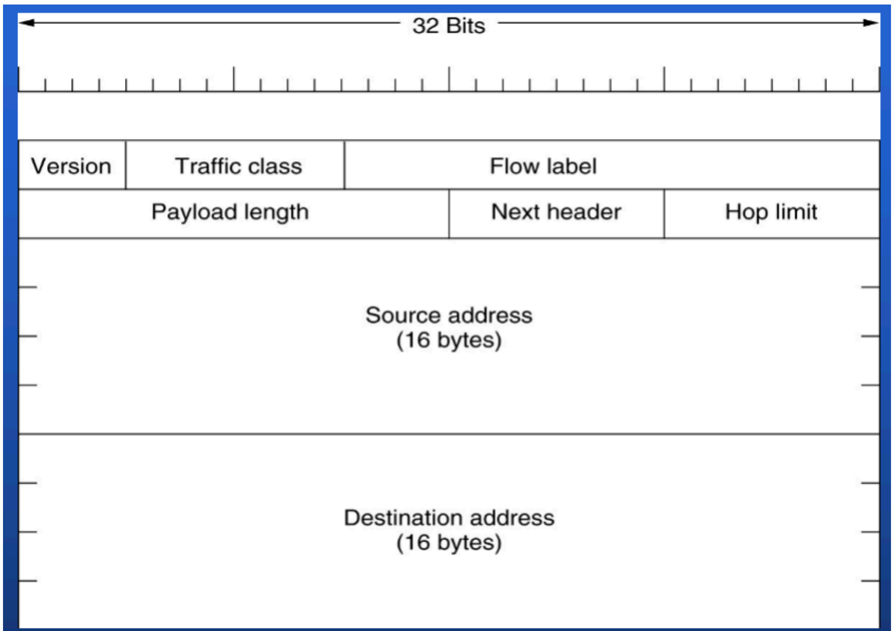


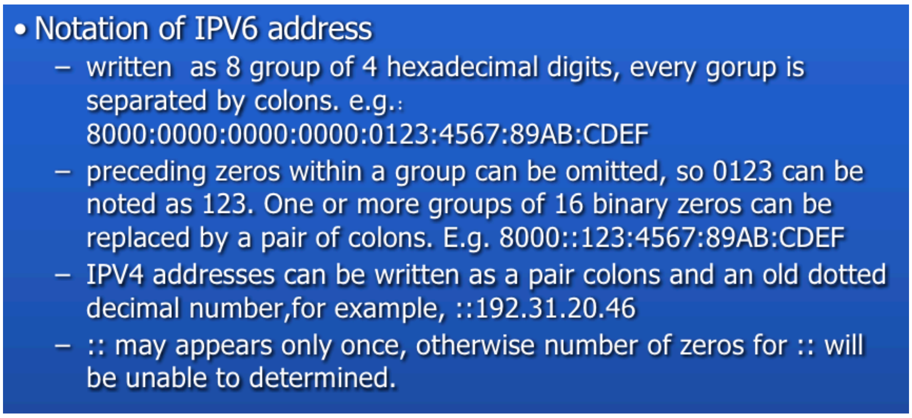
Figure 5-68 The IPv6 fixed header(required)

5.7.1 Extension header 扩展头

Packet 1	Packet 2	Packet 3	Packet 4
IPv6 header	TCP header + data		
Next Header = TCP			
IPv6 header	Routing header	TCP header +data	
Next Header = Routing	Next Header = TCP		
IPv6 header	Routing header	Fragment header	segment of TCP header + data
Next Header = Routing	Next Header = Fragment	Next Header = TCP	

- 很类似链表，一个头指出下一个头的信息
- 6 kinds of Extension Headers:
  - The hop-by-hop extension header-[see fig 5-70]
  - The extension header for routing-[see fig 5-71]
  - Fragmentation
  - Authentication
  - Encrypted security payload
  - Destination options

5.7.2 IPV6 address



5.7.3 flow

流形式：传输时有两个端点，然后有流标号。一个主机和多个主机/路由器通信，通过不同的流标号来区分信息发送渠道。

6 Tips

6.1 RIP & OSPF

6.1.1 RIP

RIP协议采用距离向量算法，在实际使用中已经较少适用。在默认情况下，RIP使用一种非常简单的度量制度：距离就是通往目的站点所需经过的链路数，取值为0~16，数值16表示路径无限长。RIP进程使用UDP的520端口来发送和接收RIP分组。RIP分组每隔30s以广播的形式发送一次，为了防止出现"广播风暴"，其后续的分组将做随机延时后发送。在RIP中，如果一个路由在180s内未被刷新，则相应的距离就被设定成无穷大，并从路由表中删除该表项。RIP分组分为两种：请求分组和响应分组。

6.1.2 OSPF

OSPF路由协议是一种典型的链路状态（Link-state）的路由协议，一般用于同一个路由域内。在这里，路由域是指一个自治系统（Autonomous System），即AS，它是指一组通过统一的路由政策或路由协议互相交换路由信息的网络。在这个AS中，所有的OSPF路由器都维护一个相同的描述这个AS结构的数据库，该数据库中存放的是路由域中相应链路的状态信息，OSPF路由器正是通过这个数据库计算出其OSPF路由表的。 [2]

作为一种链路状态的路由协议，OSPF将链路状态组播数据LSA（Link State Advertisement）传送给在某一区域内的所有路由器，这一点与距离矢量路由协议不同。运行距离矢量路由协议的路由器是将部分或全部的路由表传递给与其相邻的路由器。[2]

6.1.3 Comparison

RIP 就是把自己所有的路由表都发给邻居（中间解决了广播风暴（用序号标定）和断电无穷大问题（长时间未刷新移出路由表））。它本质上只能更新路由表来传递信息。

OSPF就是在一个AS内，很多路由器共同参与维护一个数据库，然后每个路由器建立一个最短路径树(Dijkstra)。当网络拓扑结构发生变化时，此协议会发送对应的信息包，收敛速度快很多