Chapter 7
网络层协议

zibuyu

January 10, 2006

1 网络层概述

1.1 ISO定义

网络层为一个网络连接的两个传送实体间交换网络服务数据单元提供功能和规程的方法,它使传送实体独立于路由选择和交换的方式.

网络层是是信子网的最高层,处理端到端传输的最低层.

网络层要解决的关键问题是了解通信子网的拓扑结构.

1.2 网络层为传输层提供服务

1. 面向连接的服务:传统电信的观点,认为通信子网应当提供可靠的、面向连接的服务.(ATM)
2. 无连接的服务:观点,通信子网无论怎么设计都是不可靠的,因此网络层只提供无连接的服务.(IP)

1.3 虚电路子网和数据报子网

都属于分组交换,采用存储转发机制.

1. 路由器内空间与带宽的权衡
   - 虚电路方式:路由器需要维护虚电路的状态信息.
   - 数据报方式:每个数据报都写黛完整的目的/源地址，浪费带宽.
2. 连接建立时间与地址查找时间的权衡
   - 虚电路需要在建立连接时花费时间.
   - 数据报在每次路由时花费时间.
3. 服务质量与可靠性的权衡

- 虚电路方式很容易保证服务质量(QoS Quality of Service),但可靠性不高.
- 数据报不容易保证QoS,但对通信线路故障,适应性较强.

1.4 注意

通信子网提供的服务(面向连接或无连接)与通信子网结构(虚电路与数据报)没有必然联系.

2 路由算法的最优化原则

- 采用数据报分组交换方式,每个包都要进行路由选择.
- 采用虚电路分组交换方式,只需在建立连接时做一次路由选择.

2.1 路由算法分类

- 非自适应算法,静态路由算法.
- 自适应算法,动态路由算法.

2.2 最优化原则(Optimality Principle)

如果路由器J在从路由器I到K的最优路由上,则从J到K的最优路由会落在同一路由上.

2.3 汇集树(sink tree)

从所有的源结点到一个给定的目的结点的最优路由的集合形成了一个以目的结点为根的树，称为汇集树.

路由算法的目的是找出并使用汇集.

3 静态路由算法

3.1 最短路径路由算法(Shortest Path Routing)-Dijkstra算法

1. 每个结点用从源结点沿已知最佳路径到本结点的距离来标注，标注分为临时性标注和永久性标注;
2. 初始时，所有结点都为临时性标注，标注为无穷大;
3. 将源结点标注为0，且为永久性标注，并令其为工作结点;
4. 检查与工作结点相邻的临时性结点，若该结点到工作结点的距离与工作结点的标注之和小于该结点的标注，则用新计算得到的和重新标注该结点;

5. 在整个图中查找具有最小值的临时性标注结点，将其变为永久性结点，并成为下一轮检查的工作结点;

6. 重复第四、五步，直到目的结点成为工作结点;

注意 :程序与算法的区别是：从目的结点开始.

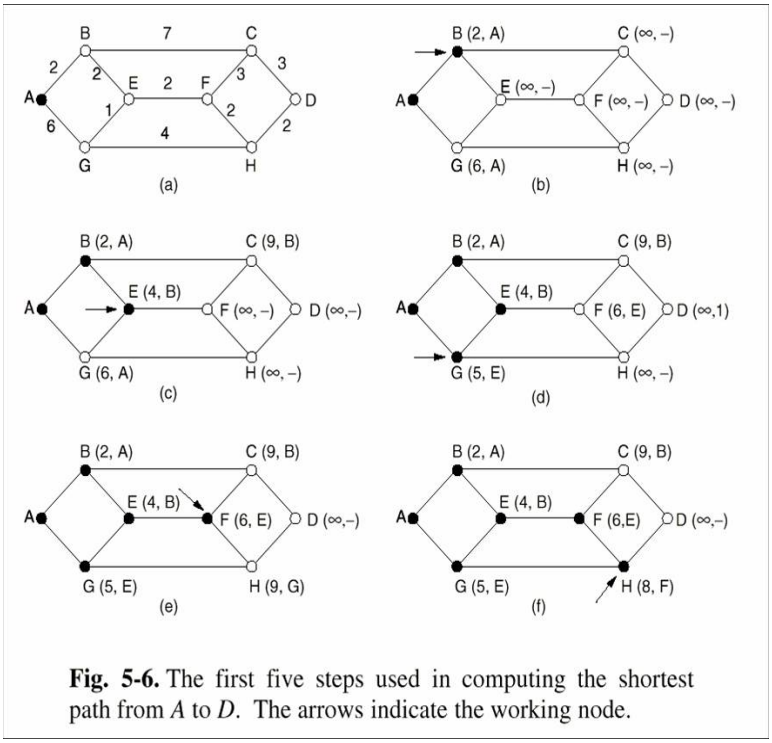


Fig. 5-6. The first five steps used in computing the shortest path from *A* to *D*. The arrows indicate the working node.

3.2 洪泛算法(Flooding)

基本思想 :把收到的每一个包,向除了该包到来的线路外的所有输出线路发送.

主要问题 :会产生大量重复包.

解决措施 :

- 每个包头包含站点计数器，每经过一站计数器减1，为0时丢弃包.(防止包无限制传递).
- 记录包经过的路径,防止重复发送给这些分组.

选择性洪泛算法(selective flooding) ：将进来的每个包仅发送到与正确方向接近的线路上.

应用 ：

- 对路由器和线路都过于浪费.
- 具有极好健壮性,可用于军事应用.

3.3 基于流量的路由算法(Flow-Based Routing)

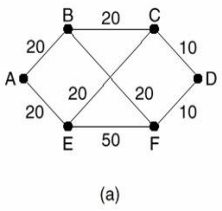
基本思想 ：

- 既考虑拓扑结构，又兼顾网络负载.
- 前提：每对结点间的平均数据流失相对稳定和可预测的.
- 根据网络带宽和平均流量，可得出平均包延迟.因此路由选择问题归结为寻找产生网络最小延迟的路由选择算法.

需要的信息 ：

- 网络拓扑结构.
- 通信量矩阵F_{ij}.
- 线路带宽矩阵C_{ij}.
- 路由算法.

个人理解 :根据某一路由算法得到一个路由,然后计算得到该路由的网络平均延迟,从而评价该路由的优劣.通过对不同路由的比较,找到平均延迟最小的路由.



(a)

Destination

Source	A	B	C	D	E	F
A		9 AB	4 ABC	1 ABFD	7 AE	4 AEF
B	9 BA		8 BC	3 BFD	2 BFE	4 BF
C	4 CBA	8 CB		3 CD	3 CE	2 CEF
D	1 DFBA	3 DFB	3 DC		3 DCE	4 DF
E	7 EA	2 EFB	3 EC	3 ECD		5 EF
F	4 FEA	4 FB	2 FEC	4 FD	5 FE	

(b)

Fig. 5-8. (a) A subnet with line capacities shown in kbps. (b) The traffic in packets/sec and the routing matrix.

i	Line	λ_i (pkts/sec)	C_i (kbps)	μC_i (pkts/sec)	T_i (msec)	Weight
1	AB	14	20	25	91	0.171
2	BC	12	20	25	77	0.146
3	CD	6	10	12.5	154	0.073
4	AE	11	20	25	71	0.134
5	EF	13	50	62.5	20	0.159
6	FD	8	10	12.5	222	0.098
7	BF	10	20	25	67	0.122
8	EC	8	20	25	59	0.098

Fig. 5-9. Analysis of the subnet of Fig. 5-8 using a mean packet size of 800 bits. The reverse traffic (*BA*, *CB*, etc.) is the same as the forward traffic.

4 动态路由算法

4.1 距离向量路由算法(Distance Vector Routing)

也称Bellman-Ford或Ford-Fulkerson算法,最初用于ARPANET,后被RIP协议采用 .

基本思想 :

- 每个路由器维护一张表，表中给出了到每个目的地的已知最佳距离和线路，并通过与相邻路由器交换距离信息来更新表;
- 以子网中其它路由器为表的索引，表项包括两部分：到达目的结点的最佳输出线路，和到达目的结点所需时间或距离;
- 每隔一段时间，路由器向所有邻居结点发送它到每个目的结点的距离表，同时它也接收每个邻居结点发来的距离表;
- 邻居结点X发来的表中，X到路由器i的距离为X_i，本路由器到X的距离为m，则路由器经过X到i的距离为$X_i + m$。根据不同邻居发来的信息，计算$X_i + m$，并取最小值，更新本路由器的路由表.

注意 :本路由器中的老路由表在计算中不被使用.

无限计算问题 **对**好消息反应迅速，对坏消息反应迟钝.

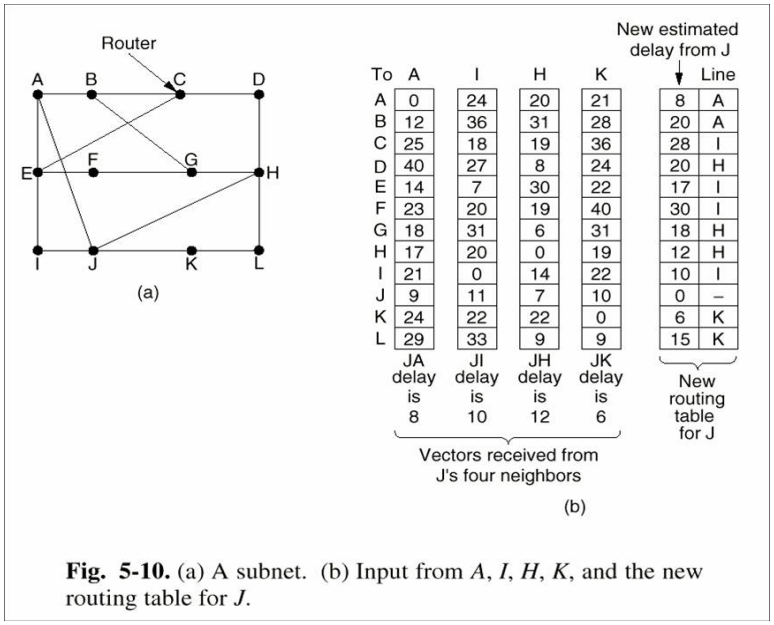


Fig. 5-10. (a) A subnet. (b) Input from *A*, *I*, *H*, *K*, and the new routing table for *J*.

水平分裂算法 ：工作过程与距离向量算法相同,区别在于,到X的距离不向真正通向X的邻居节点报告,使得坏消息传播的也快.但仍然会有失败.

问题关键 :在于当X告诉Y它有一条路径的时候,Y无法知道自己是否在这条路径上.

4.2 链状路由算法(Link State Routing)

距离向量算法主要问题 ：

- 选择路由时,没有考虑线路带宽
- 路由收敛速度慢

算法 ：

1. 发现邻居结点，并学习它们的网络地址;
   - 路由器启动后，通过发送HELLO包发现邻居结点
   - 两个或多个路由器连在一个LAN时,引入人工结点
2. 测量到每个邻居结点的延迟或开销
   - 一种直接的方法是：发送一个要对方立即响应的ECHO包，来回时间除以2即为延迟(默认双向延迟相同)
3. 将所有学习到的内容封装成一个包
   - 包以发送方的标识符开头，后面是序号、年龄和一个邻居结点列表

- 列表中对应每个邻居结点，都有发送方到它们的延迟或开销
- 链路状态包定期创建或发生重大事件时创建

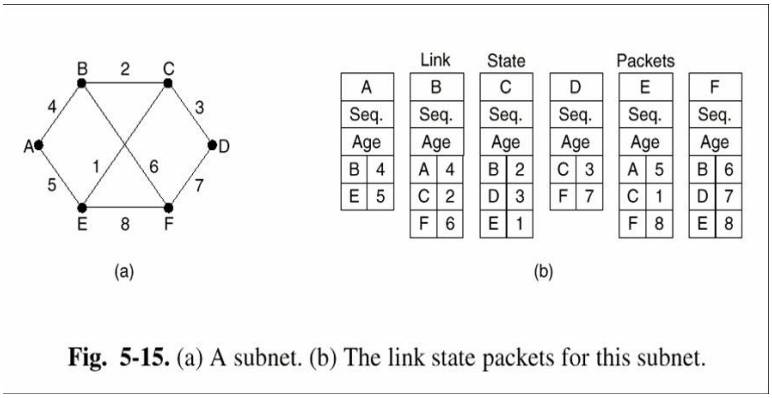


Fig. 5-15. (a) A subnet. (b) The link state packets for this subnet.

4. 将这个包发送给所有其它路由器

   - 基本思想：洪泛链路状态包，为控制洪泛，每个包包含一个序号，每次发送新包时加1。路由器记录信息对(源路由器，序号)，当一个链路状态包到达时，若是新的，则分发；若是重复的，则丢弃；若序号比路由器记录中的最大序号小，则认为过时而丢弃
   - 改进

     Question 序号循环使用会混淆

     Solution 使用32位序号

     Question 路由器崩溃后，序号重置,序号会出错

     Solution 增加年龄（age）域，每秒钟年龄减1，为零则丢弃

     其他改进1 链路状态包到达后，延迟一段时间，并与其它已到达的来自同一路由器的链路状态包比较序号，丢弃重复包，保留新包

     其他改进2 链路状态包需要确认

5. 计算到每个其它路由器的最短路径

   - 在各路由器中根据全部链路状态分组,可以构造完整的通信子网图,采用Dijkstra算法计算最短路径

4.3 LS和DV比较

1. 路由信息复杂性

   - LS 路由信息向全网发送
   - DV 仅在相邻节点发送

2. 收敛(Convergence)速度
    - LS 使用最短路径优先算法;可能出现路由振荡
    - DV 收敛时间不确定;可能出现路由循环以及count-to-infinity(无限计算)问题
3. 健壮性(如果路由器不能正常工作会怎样)
    - LS 结点会广播错误链接开销,每个节点只计算自己的路由表
    - DV 结点会广播错误路径开销,每个结点的路由表被别的结点使用,错误传播全网

5 分层路由(Hierarchical Routing)

思想 :分而治之,将路由器分为区域、聚类、区和组.

带来问题 :路由表中的路由不一定是最优路由.

6 移动主机路由

基本概念 :

- 移动用户(mobile users)
- 家乡位置(home location)
- 外部代理(foreign agent):记录正在访问该区域的移动用户
- 家乡代理(home agent):记录家乡在该区域,但目前正在访问其他区域的用户

算法

1. 移动用户进入一个新区域时，必须首先向外部代理注册
2. 移动用户的路由转发过程:发送方=¿家乡局域网=¿家乡代理(并告知发送方以后直接通过外部代理发送给用户)=¿隧道技术=¿外部代理=¿移动用户

7 拥塞(congestion)控制

拥塞控制和流量控制

- 拥塞控制:确保通信子网能够承载用户提供的通信量,是个全局性的问题,涉及主机、路由器等很多因素.
- 流量控制(Flow Control):与点到点通信量有关，主要解决快速发送方与慢速接收方的问题.局部问题，一般基于反馈进行控制.

基本原理 :基于控制论控制方法分为两类

- 开环控制:通过好的设计解决问题,避免拥塞发生;拥塞控制时不考虑当前网络状态.
- 闭环控制:基于反馈机制.

衡量拥塞的参数

- 缺乏缓冲区造成的丢包率
- 平均队列长度
- 超时重传的包数目
- 平均包延迟
- 报延迟变化

7.1 拥塞控制算法

7.1.1 拥塞预防策略

属于开环控制.
采取影响拥塞的网络设计策略来预防拥塞.

7.1.2 流量整形(Traffic Shaping)

基本思想 :

- 造成拥塞的主要原因是网络流量通常是突发性的
- 强迫包以一种可预测的速率发送
- 在ATM网中广泛使用

7.1.3 漏桶算法(The Leaky Bucket Algorithm)

基本思想 :将用户发出的不平滑的数据包流转变成网络中平滑的数据包流.

适用: 可用于不固定包长协议(ATM),也可用于变包长协议(IP).

缺点: 无论负载突发性如何，漏桶算法强迫输出按平均速率进行，不灵活.

7.1.4 令牌桶算法(The Token Bucket Algorithm)

基本思想 :漏桶存放令牌，每ΔT秒产生一个令牌，令牌累积到超过漏桶上界时就不再增加。包传输之前必须获得一个令牌，传输之后删除该令牌.

与漏桶算法的区别 ：

- 流量整形策略不同：漏桶算法不允许空闲主机积累发送权，以便以后发送大的突发数据；令牌桶算法允许，最大为桶的大小;
- 漏桶中存放的是数据包，桶满了丢弃数据包；令牌桶中存放的是令牌，桶满了丢弃令牌(传输容量或传输许可)，不丢弃数据包.

本质上令牌桶所做的事情是 :允许突发流量，但是不能超过一个预定的最大长度。

公式 :设突发事件为S秒，令牌桶的容量是C字节，令牌的到达速率为ρ字节/秒，令牌最大传出速率为M字节/秒，则有：

$$C + \rho S = MS$$

得公式如下：

$$S = C/(M - \rho)$$

如图所示：

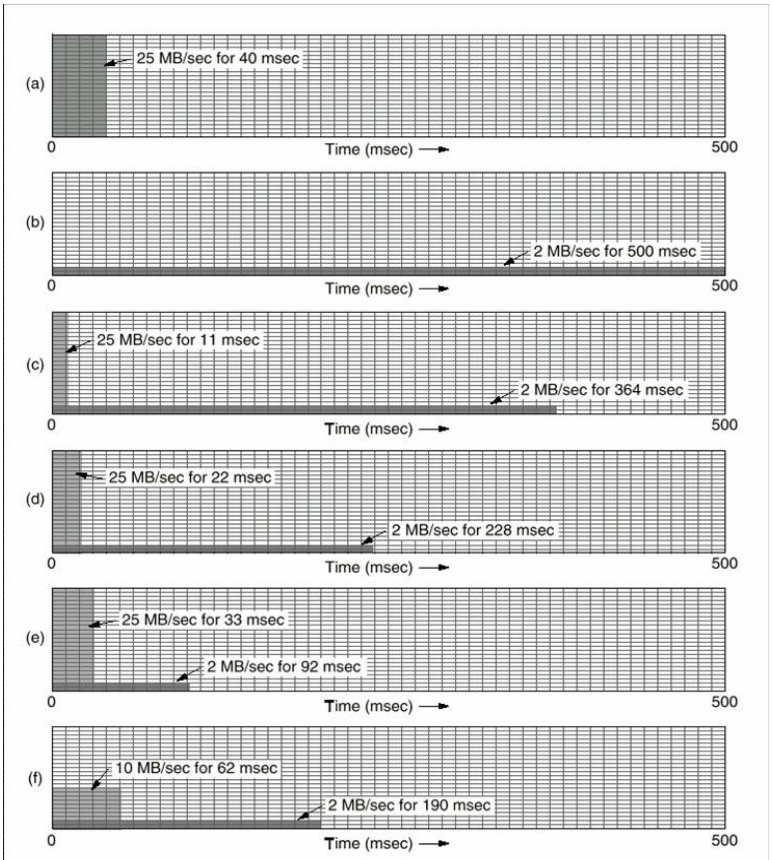


Fig. 5-25. (a) Input to a leaky bucket. (b) Output from a leaky bucket. (c) - (e) Output from a token bucket with capacities of 250KB, 500KB, and 750KB. (f) Output from a 500KB token bucket feeding a 10 MB/sec leaky bucket.

7.1.5 加权公平队列(Weighted Fair Queueing)

公平队列(Fair Queueing)算法 :

- 路由器的每个输出线路有多个队列;
- 路由器循环扫描各个队列，发送队头的包;
- 所有队列具有相同优先级;
- 一些ATM交换机、路由器使用这种算法;

加权公平队列算法 :给不同主机以不同的优先级；优先级高的主机在一个轮讯周期内获得更多的时间片。

7.1.6 负载丢弃(Load Shedding)

其他算法都不能消除拥塞时，路由器只得将包丢弃 .

针对不同服务，可采取不同丢弃策略 :

文件传输 优先丢弃新包，wine策略;

多媒体服务 优先丢弃旧包，milk策略;

7.2 流说明(Flow Specification)

:一个数据流的发送方、接收方和通信子网三方认可的、描述发送数据流的模式和希望得到的服务质量的数据结构.

对于发送方的流说明 :子网和接收方可以做出三种答复：同意、拒绝、其它建议.

7.3 虚电路子网的拥塞控制

方法 :

许可控制(Admission Control): 一旦发生拥塞，在问题解决之前，不允许建立新的虚电路;

另一种方法: 可以建立新的虚电路，但要绕开发生拥塞的地区;

资源预留: 建立虚电路时，主机与子网达成协议，子网根据协议在虚电路上为此连接预留资源.

7.4 数据报子网的拥塞控制

7.4.1 抑制包(Choke Packets)

思想 :

- 路由器监控输出线路及其它资源的利用情况，超过某个阈值，则此资源进入警戒状态;
- 每个新包到来，检查它的输出线路是否处于警戒状态;
- 若是，则向源主机发送抑制包，包中指出发生拥塞的目的地址。同时将原包打上标记（为了以后不再产生抑制包），正常转发;
- 源主机收到抑制包后，按一定比例减少发向特定目的地的流量，并在固定时间间隔内忽略指示同一目的地的抑制包。然后开始监听，若此线路仍然拥塞，则主机在固定时间内减轻负载、忽略抑制包；若在监听周期内没有收到抑制包，则增加负载;

注意 :通常采用的流量增减策略是：减少时，按一定比例减少，保证快速解除拥塞；增加时，以常量增加，防止很快导致拥塞.

7.4.2 逐跳抑制包(Hop-by-Hop Choke Packets)

在高速、长距离的网络中，由于源主机响应太慢，抑制包算法对拥塞控制的效果并不好 .

思想 :

- 抑制包对它经过的每个路由器都起作用;
- 能够迅速缓解发生拥塞处的拥塞;

注意 :上游路由器(临近发送端的路由器)要求有更多的缓冲区.

8 互联网络(internet)设备

- 中继器(repeater)
  - 物理层设备，在电缆段之间拷贝比特
  - 对弱信号进行放大或再生，以便延长传输距离
- 网桥(bridge)
  - 数据链路层设备，在局域网之间存储转发帧
  - 网桥可以改变帧格式
- 多协议路由器(multi protocol router)
  - 网络层设备，在网络之间存储转发包
  - 必要时，做网络层协议转换
- 传输网关(transport gateway)
  - 传输层设备，在传输层转发字节流
- 应用网关(application gateway)
  - 应用层设备，在应用层实现互连

8.1 级联虚电路(Concatenated Virtual Circuits)

多个虚电路子网互联.

8.2 无连接网络互连(Connectionless Internetworking)

多个数据报子网互联.

9 隧道(Tunneling)技术及防火墙

源和目的主机所在网络类型相同，连接它们的是一个不同类型的网络，这种情况下可以采用隧道技术 .

9.1 互联网络路由(Internetworking Routing)

互联子网络路由类似于单独子网的路由过程.

两极路由算法 :

- 内部网关协议IGP(Interior Gateway Protocol)
- 外部网关协议EGP(Exterior Gateway Protocol)
- 自治系统AS(Autonomous System):只每个独立于其他网络的子网.

注意 :这里的 " 网关 " 是 " 路由器 " 的老式叫法.

9.2 分段(Fragmentation)

每个网络对最大包长有限制 .

注意 :当大包通过小包网络时，网关要将大包分成若干段(fragment),每段作为独立的包传输.

9.3 段重组策略

分段重组过程对其他网络透明 :

- 网关将大包分段后，每段都要经过同一出口网关，并在那里重组;
- 带来问题
  - 出口网关需要知道何时所有分组都到齐;
  - 所有分组必须从同一出口网关发送出去;
  - 大包经过一系列小包网络时，需要反复地分段重组，开销大.

分段重组过程对其他网络不透明 :

- 中间网关不做重组，而由目的主机做
- 带来问题
  - 对主机要求高，能够重组;
  - 每个段都要有一个包头，网络开销增大;

标记段 ：

1. 树型标记法
   - 例:包0分成三段，分别标记为0.0, 0.1, 0.2，段0.0构成的包被分成三段，分别标记为0.0.0, 0.0.1, 0.0.2
   - 存在问题:(1)段标记域足够长(2)如果重传,两次分段长度前后要一致.
2. 偏移量法
   - 定义一个基本段长度,使得基本段能够通过所有网络
   - 包分段时，除最后一个段小于等于基本段长度外，所有段长度都等于基本段长度
   - 一个包可以包括几个段，包头中包括:原始包序号，包中第一个基本段的偏移量，最后段指示位(指示该包中包含的最后一个基本分段是否为原始包的最后一个分段)

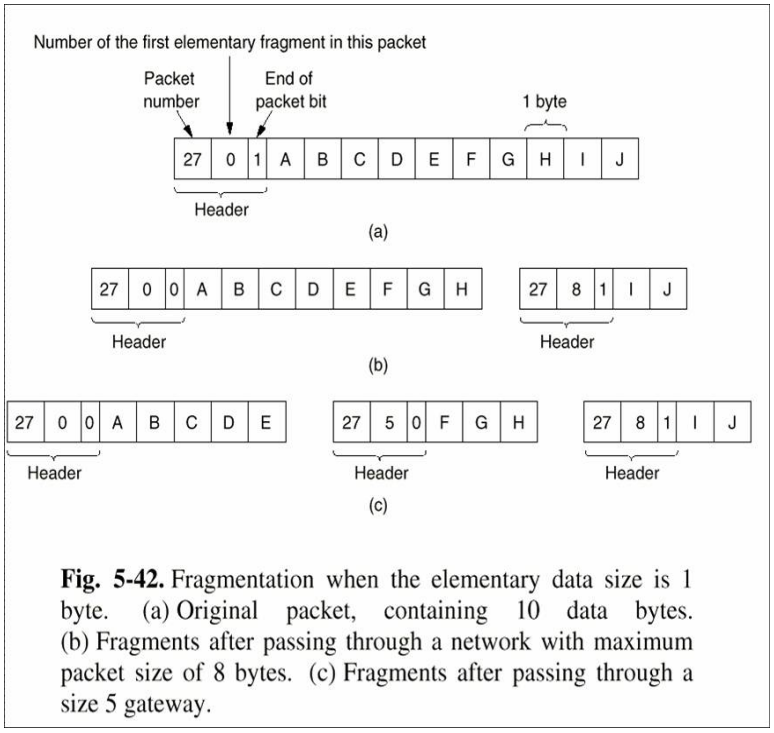


Fig. 5-42. Fragmentation when the elementary data size is 1 byte. (a) Original packet, containing 10 data bytes. (b) Fragments after passing through a network with maximum packet size of 8 bytes. (c) Fragments after passing through a size 5 gateway.

9.4 防火墙(Firewalls)

防火墙的一种常用配置

- 两个路由器，根据某种规则表，进行包过滤；
- 一个应用网关，审查应用层信息。

10 IP协议等

10.1 IP协议

10.1.1 IP头

包括20个字节的固定部分和变长(最多40字节)部分.

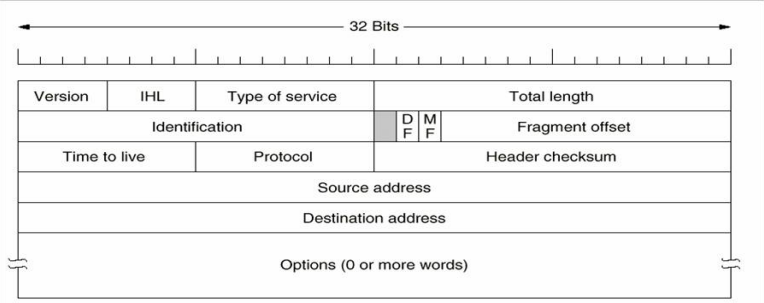


Fig. 5-45. The IP (Internet Protocol) header.

10.1.2 IP地址

IP地址= **网**络号+主机号.

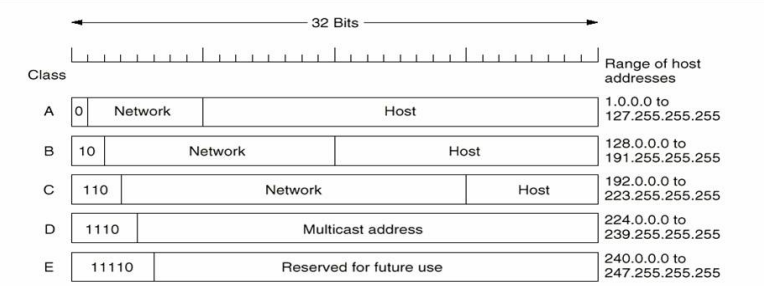


Fig. 5-47. IP address formats.

全0和全1的特殊含义 :

- 全0:表示本网络或本主机
- 全1:表示广播地址

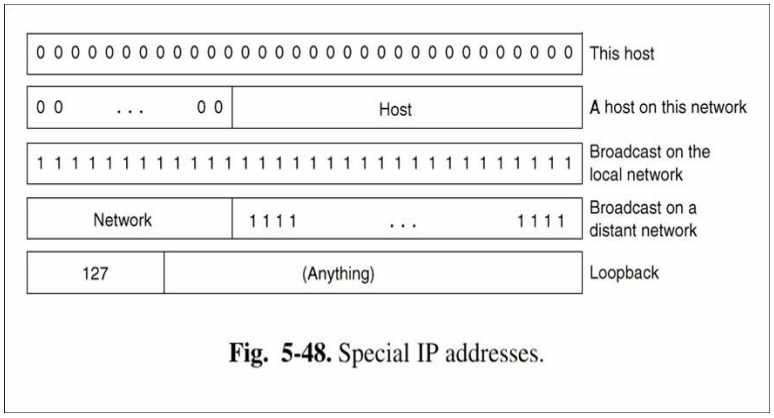


Fig. 5-48. Special IP addresses.

10.1.3 子网(subnets)

分而治之的思想：为了便于管理和使用，可以将网络分成若干供内部使用的部分，称为子网。对外界，该网络还是一个单独的网络。

子网掩码 :代表“网络+子网号”与主机号的分割。见课本372页。

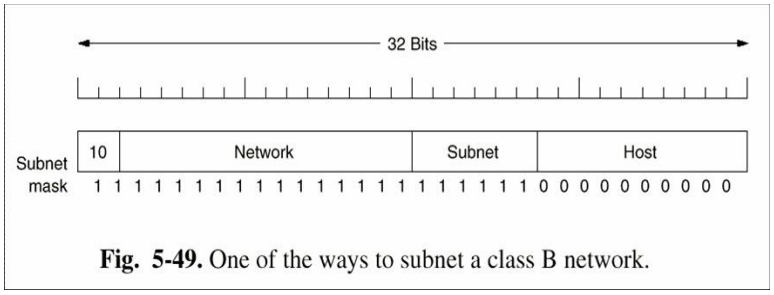


Fig. 5-49. One of the ways to subnet a class B network.

10.2 Internet控制报文协议ICMP

Internet Control Message Protocol

用途 :

- 主要用来报告出错和测试
- 报文类型
- ICMP报文封装在IP包中

10.3 地址解析协议ARP(Address Resolution Protocol)

用途 :

- 解决网络层地址(IP地址)与数据链路层地址(MAC地址)的映射问题;

- 工作过程
  - 主机建立一个ARP表,表中存放(IP地址，MAC地址)对;
  - 若目的主机在同一子网内，用目的IP地址在ARP表中查找，否则用缺省网关的IP地址在ARP表中查找。若未找到，则发送广播包，目的主机收到后给出应答，ARP表增加一项;
  - 每个主机启动时，广播它的(IP地址，MAC地址)映射;
  - ARP表中的表项有生存期，超时则删除.
- 在不同以太网上的主机互相发送信息,需要
  - 用代理ARP(proxy ARP),在源主机ARP表中添加(目标IP地址+该代理路由器MAC地址) 项.
  - 让主机马上看到目标主机在另外一个远程网上,并将所有这样的流量发送到默认以太网址,由其负责所有远程流量

10.4 反向地址解析协议RARP(Reverse Address Resolution Protocol)

用途 :

- 解决数据链路层地址(MAC地址)与网络层地址(IP地址)的映射问题;
- 主要用于无盘工作站启动;
- 缺点由于路由器不转发广播帧，RARP服务器必须与无盘工作站在同一子网内。一种替代协议BOOTP（使用UDP）。

11 RIP协议等

11.1 RIP(Routing Information Protocol)

内部网关协议IGP: 自治系统(AS)内使用的路由算法为:RIP,OSPF.
外部网关协议EGP: 自治系统之间使用的路由算法:BGP.

特点 :

- 采用距离向量算法
- 基于UDP.
- 距离的衡量采用跳数,(max=15hops).
- 距离向量每30s交换一次.
- 采用poison reverse毒性逆转)避免ping-pong loops[大概是无限计算问题](infinite distance = 16hops).

11.2 开放最短路径优先OSPF(Open Shortest Path First)

特点 :

- 链路状态路由算法.
- 支持多种衡量尺度(如物理距离,延迟等).
- 动态算法.
- 负载平衡.
- 支持分层系统
- 支持隧道技术

构造有向拓扑图 :

- 根据实际的网络、路由器和线路构造有向图;
- 每个弧有一个开销值;
- 两个路由器之间的线路用一对弧来表示，弧权可以不同;
- 多路访问（multiaccess）网络，网络用一个结点表示，每个路由器用一个结点表示，网络结点、路由器结点的弧权为0

分层路由 :

- 自治系统AS可以划分区域(areas);
- 每个AS有一个主干（backbone）区域，称为区域0，所有区域与主干区域相连;
- 一般情况下，有三种路由:区域内,区域间,自治系统间;
- 四类路由器，允许重叠:(根据所在位置划分).

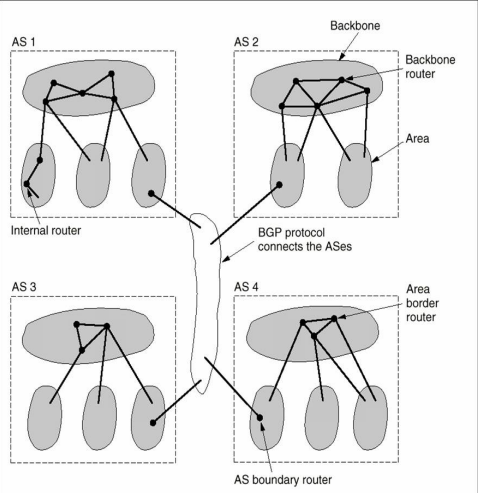


Fig. 5-53. The relation between ASes, backbones, and areas in OSPF.

11.3 边界网关协议BGP(Border Gateway Protocol)

特点 :

- 通过TCP连接传送路由信息;
- 采用路径向量（path vector）算法，路由信息中记录路径的轨迹;

12 无类域间路由CIDR(Classless InterDomain Routing)

CIDR提出 :Internet指数增长，IP地址即将用完;基于分类的IP地址空间的组织浪费了大量的地址.

基本思想 :

- 将剩余的C类地址分成大小可变的地址空间;
- 例如：需要2000个地址，则分配一个2048个地址（8个C类地址）的连续地址块，而不是一个B类地址;
- RFC1519 改变了过去C类地址的分配规则，将世界分成4个区，每个区分配一块连续的C类地址空间;
- 路由表中增加一个32位的掩码（mask）域;
- 最长匹配原则：路由查找时，若多个路由表项匹配成功，选择掩码长（1比特数多）的路由表项;
- CIDR思想可用于所有IP地址，没有A、B、C类之分.

注意掌握： **容**易结合IP地址分类，子网掩码等知识点考试.

13 IPv6

IPv6头结构 ：

- 地址变长，128位.
- IP头简化，由13个域减少为7个域，提高路由器处理速度.

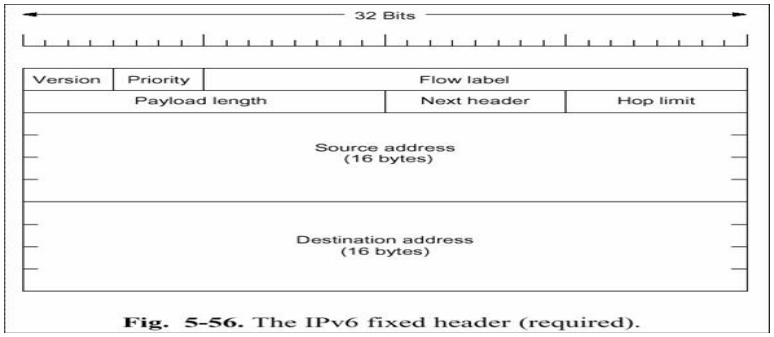


Fig. 5-56. The IPv6 fixed header (required).

地址表示 ：

- 16字节地址表示成用冒号（:）隔开的8组，每组4个16进制位;
- 含有多个0,可以优化表示.

与IPv4互联的三种方案 ：

- 双栈;
- 翻译
- 隧道：IPv6的报文作为IPv4报文的净荷在IPv4网络中传输.