1 概述

1.1 转发和路由选择

网络层的两种功能：

1. 转发，将分组从一个输入链路接口转移到适当的输出链路接口的路由器本地动作。
2. 路由选择，这是一个网络范围的过程，用于决定分组从源到目的地所采取的端到端路径。

课本第 204 页有转发和路由的形象比喻。
网络层的三种设备：

- 分组交换机，是一台通用分组交换设备，它根据分组首部字段中的值，从输入链路借口到输出链路接口转移分组。
- 链路层交换机，根据链路层字段中的值做转发决定。
- 路由器，根据网络层字段中的值做转发决定。

网络层还有连接建立的功能：要求从源到目的地沿着所选择的路径彼此握手，以便在给定源到目的地连接中的网络层数据分组能够在开始流动之前建立起状态。

1.2 网络服务模型

有三种需要了解的网络服务模型：因特网、ATM CBR 和 ATM ABR 服务模型，在课本第 206 页表 4-1 很清楚地列出了这三种服务模型的特点。

2 虚电路和数据报网络

网络层连接和无连接服务是主机到主机之间的服务。
虚电路网络：在网路层提供连接服务的计算机网络。
数据报网络：在网络层提供无连接服务的计算机网络。
网络层连接服务不仅在端系统实现，还在路由器中实现。

2.1 虚电路网络

因特网是一个数据报网络，而 ATM 是一个虚电路网络。
虚电路的组成如下：

1. 源和目的主机之间的路径 (一系列的链路和路由器)。
2. VC 号，沿着该路径的每段链路的一个号码。属于虚电路的分组的首部中有一个 VC 号，每个中间路由器用下一段链路的 VC 号替代每个传输分组的 VC 号，这个 VC 号从转发表中获得。
3. 沿着该路径的每台路由器中的转发表表项。

课本第 208 页有一个 VC 号更换的例子。

一个分组沿着该路由不简单地保持相同 VC 号的原因:

1. 逐链路代替该号码减少了在分组首部中 VC 字段的长度。
2. 通过允许沿着该虚电路路径每条链路上有一个不同的 VC 号，大大简化了虚电路的建立。

虚电路网络中，路由器维持连接状态信息，该信息将 VC 号与输出接口号联系起来。当跨越一个路由器创建一个新连接时，就增加一个新的连接项，关闭一个连接时，删除该项。课本第 208 页有这个例子。

虚电路上的 3 个阶段:

1. 虚电路建立，步骤如下:
   - 发送运输层与网络层联系，指定接收方地址。
   - 网络层决定发送方与接收方之间的路径。
   - 网络层为沿着该路径的每条链路决定一个 VC 号。
   - 网络层在沿着路径的每台路由器的转发表中增加一个表项。
2. 数据传送，分组沿着虚电路传送。
3. 虚电路拆除，步骤如下:
   - 发送方通知网络层终止该虚电路。
   - 网络层通知网络另一侧的端系统结束呼叫。
   - 更新路径上每台路由器中的转发表。

信令报文：端系统向网络发送指示虚电路启动与终止的报文，以及路由器之间传递的用于建立虚电路的报文。

信令协议：用来交换这些报文的协议。

2.2 数据报网络

数据报网络分组传递方式：当一个端系统要发送分组，它就为该分组加上目的端系统的地址，然后将分组推进网络中。

数据报网络的工作方式：每台路由器都使用分组的目的地址来转发该分组，路由器使用该分组的目的地址在转发表中查找适当的输出链路接口，然后将分组向该输出链路接口转发。

数据报网络中的路由器在转发表中维持了转发状态信息, 转发表通过路由选择算法进行修改。

3 路由器工作原理

路由器由 4 个部分组成：

1. 输入端口
2. 交换结构
3. 输出端口
4. 路由选择处理器

3.1 输入端口

功能如下：

- 具有线路端接功能，用于执行将一条输入的物理链路与路由器相连接的物理层功能。
- 路由器在输入端口处，使用转发表来查找输出端口，使得到达的分组将能经过交换结构转发到该输出端口。
- 转发表的一份影子副本存放在每个输入端口。

为了提高查找的速度，采用了如下技术：

- 使用硬件执行查找
- 对大型转发表使用超出简单线性搜索的技术
- 减少内存访问事件，使用 DRAM、SRAM 或 TCAM 存储器。

3.2 交换结构

有多种实现方式的交换结构：

- 经内存交换：
  1. 分组到达一个输入端口时，该端口会通过中断方式向路由选择处理器发出信号。
  2. 该分组复制到处理器内存。
  3. 路由选择处理器从分组首部提取目的地址，找出适当的输出端口。
  4. 将分组复制到输出端口的缓存中。
- 经总线交换：
  1. 输入端口为分组计划一个交换机内部标签。
  2. 分组在总线上传送，该分组能有所有输出端口接收到。
  3. 只有域标签匹配的端口才能保存该分组，然后标签在输出端口被去除。
- 经互联网络交换：
  1. 当分组到达端口 A，需要转发到端口 Y。
  2. 交换机控制器闭合总线 A 和 Y 交叉部分的交叉点。
  3. 分组在总线 A 上发送，仅由总线 Y 接收。

3.3 何处出现排队

对缓存长度的需求：缓存数量应该等于平均往返时延乘以链路的容量。

主动队列管理算法：在缓存填满前便丢弃一个分组，以便向发送方提供一个拥塞信号。

随机早期检测 (RED 算法) 是主动队列管理算法中的一种：

1. 为输出队列长度维护着一个加权平均值。
2. 当平均队列长度小于最小阈值，则接纳这个分组。如果平均队列长度大于最大阈值，则丢弃这个分组。
3. 如果平均队列在这两者之间，则该分组以某种概率被丢弃。

如果使用 FCFS 策略，那么可能出现线路前部阻塞。一旦输入链路上的分组到达速率达到缓存容量的 58%，那么输入队列长度将无限制地增大。

4 网际协议：因特网中的转发和编址

4.1 数据报格式

数据报格式在课本第 221 页～ 223 页有详细而清楚的解释。

4.1.1 IP 数据报分片

最大传送单元 MTU：一个链路层帧能承载的最大数据量。

当 MTU 小于 IP 数据报时的解决方案：将 IP 数据报中的数据分片成两个或更多个较小的 IP 数据报，用单独的链路层帧封装这些较小的 IP 数据报，然后向输出链路上发送这些帧。

重新组装分片的机制：

1. IP 数据报首部中放有标识、标志和片偏移字段，当生成一个数据报时，发送主机为该数据报贴上标识号。
2. 当路由器需要对一个数据报分片，则每个数据报有初始数据报的标识号。最后一个片的标志比特设为 0，其他片的标志被设为 1。使用偏移字段指定该片应放在初始 IP 数据报的哪个位置。

Dos 攻击的两种行为：

- 攻击者向目标主机发送了小片的流，这些片中没有一个片的偏移量为 0。
- 发送交迭的 IP 片，这些片的偏移量值被设置得不能够适当地排列起来。

4.2 IPv4 编址

子网的定义：主机群与路由器端口分开后如果能形成一个隔离的网络岛，那么这个主机群就叫做一个子网。路由器与路由器之间连接的路径也称为子网。课本第 227 页的图 4-17 有一个形象的例子。

子网掩码：IP 编址如果为一个子网分配一个地址为 223.1.1.0/24，那么其中的/24 记法就称为子网掩码。子网掩码用于将某个 IP 地址划分成网络地址和主机地址两部分，比如 24 记法的子网掩码为 255.255.255.0。

子网寻址：32 比特的 IP 地址被划分为两部分，并且具有点分十进制数形式 a.b.c.d/x，其中 x 指示了地址的第一部分中的比特数。IP 地址的第一部分称为网络前缀。

分类编址：将 IP 地址的网络部分限制为 8、16 或 24 比特，将具有 8、16 和 24 比特子网地址的子网称为 A、B 和 C 类网络。

主机获得 IP 地址的方式：1. 直接配置一个地址。2. 动态主机配置协议。3，网络地址转换。4.UPnP。

4.2.1 获取一块地址

网络管理员从 ISP 获取一块 IP 地址，而 ISP 从 ICANN 组织 (因特网名字和编号分配组织) 获取 IP 地址。

4.2.2 动态主机配置协议

DHCP 协议又称为即插即用协议。DHCP 协议是一个 4 个步骤的过程：

- DHCP 服务器发现。DHCP 客户生成包含 DHCP 发现报文的 IP 数据报，其中目的地址为 255.255.255.255，源地址为 0.0.0.0。客户在 UDP 分组中向 67 端口发送该报文。
- DHCP 服务器提供。DHCP 服务器使用 DHCP 提供报文作出响应，目的地址为 255.255.255.255，报文中包含发现报文的事务 ID、向客户推荐的 IP 地址、网络掩码以及 IP 地址租用期。
- DHCP 服务器请求。客户向 DHCP 服务器发送一个 DHCP 请求报文，回显配置参数。
- DHCP ACK。服务器用 DHCP ACK 报文对客户进行响应。

课本第 232 页有一个 DHCP 客户-服务器交互的例子。

4.2.3 网络地址转换

课本第 234 页有 NAT 的工作例子。
NAT 会妨碍 P2P 应用程序，解决方法称为 NAT 穿越，步骤如下：

1. A 不在 NAT 后面，B 在 NAT 后面，C 也不在 NAT 后面。
2. C 和 B 已经建立了 TCP 连接，则 A 可以通过 C 请求对等方 B。

4.2.4 UPnP

UPnP 称为即插即用，用于允许主机发现并配置邻近 ANT，工作原理如下：

1. UPnP 使得应用程序能够为某些请求的公共端口号号请求一个 NAT 映射，该映射位于（专用 IP 地址，专用端口号）和（公共 IP 地址，公共端口号）之间。
2. 有了这个映射，来自外部的结点能够发起到（公共 IP 地址，公共端口号）的 TCP 连接。
3. UPnP 还能让应用程序知道（公共 IP 地址，公共端口号），使得该应用程序可以向外部世界通告它。

4.3 因特网控制报文协议

因特网控制报文 (ICMP 报文) 有一个类型字段和一个编码字段，还包含引起该 ICMP 报文首次生成的 IP 数据报的首部和前 8 字节内容。

ICMP 报文可以用于差错报告和其他应用。课本第 236 页图 4-23 描述了 ICMP 报文类型及其应用。

ICMP 报文的两个应用:

1. ping 程序，发送类型 8 的 ICMP 报文，请求回显。然后目的主机发回类型 0 的 ICMP 报文，用于回显回答。
2. traceroute 程序，向目的主机发送一系列普通的 IP 数据报，每个数据报 TTL 字段逐一增加。TTL 过期时，路由器丢弃该数据报并发送一个 ICMP 警告报文 (类型 11 的 ICMP 报文) 给源主机。该警告报文包含了路由器的名字与它的 IP 地址。当警告报文到达源主机时，源主机从定时器获得往返时延，从 ICMP 报文中得到第 n 台路由器的名字与 IP 地址。

4.4 IPv6

课本第 238 页～第 239 页有 IPv6 的数据报格式，很详细而且清楚。

课本第 239 页在 IPv4 数据报中出现的、而在 IPv6 数据报中没有出现的字段，很详细而且给出了删除这些字段的原因。

注意，IPv6 的首部是定长的 40 字节。

4.4.1 从 IPv4 到 IPv6 的迁移

迁移到 IPv6 的方法叫做双栈方式，有两种形式

1. 使用 IPv6/IPv4 结点，有发送 IPv4 和 IPv6 两种数据报的能力。
2. 建立隧道，将 IPv6 整个数据报放到一个 IPv4 数据报的数据字段，将该 IPv4 数据报的地址指向隧道接收端的 IPv6 结点。

4.4.2 涉足 IP 安全性

IPsec 是一种提供安全性服务的网络协议，仅需要在通信的两台主机中可用。

工作方式: 在发送端，运输层向 IPsec 传递一个报文段，IPsec 加密这个报文段，在报文段上添加附加的安全性字段，并且在一个普通的 IP 数据报中封装得到的有效载荷。在接收端，IPsec 解密报文段并将脱密的报文段传送给运输层。

IPsec 提供的服务:

- 密码技术约定，让两台通信主机约定加密算法和密钥。
- IP 数据报有效载荷的加密，这个有效载荷仅能被在接收主机中的 IPsec 解密。
- 数据完整性，可以保证数据报在传输过程中没有被修改过。
- 初始鉴别，用于鉴别数据报中的源 IP 地址是否是该数据报的实际源。

5 路由选择算法

默认路由器：和主机直接相连的路由器，又称为该主机的第一跳路由器。

源路由器：源主机的默认路由器。目的路由器：目的主机的默认路由器。

路由选择算法的目的：给定一组路由器以及连接路由器的链路，路由选择算法要找到一条从源路由器到目的路由器的最优路径。

5.1 路由选择算法的分类

根据该算法是全局式的还是分散式的来区分：

- 全局式路由选择算法，用完整的、全局性的网络知识计算出从源到目的地之间的最低费用路径。该算法常称为链路状态算法。
- 分散式路由选择算法，以迭代、分布式的方法计算出最低费用路径。

根据算法是静态的还是动态的分类：

- 静态路由选择算法，路由的变化很缓慢，需要人工干预进行调整。
- 动态路由选择算法，当网络流量负载或拓扑发生变化时改变路由选择算法。

根据算法是负载敏感的还是负载迟钝的分类：

- 负载敏感算法，链路费用会动态地变化以反映出底层链路的当前拥塞水平。
- 负载迟钝算法，链路费用不会明显地反映出当前的拥塞水平。

5.2 链路状态路由选择算法

在链路状态路由选择算法 (LS 算法) 中，网络拓扑和所有的链路费用都是已知的。

以上的信息通过链路状态广播算法获得，让每个结点向网络中所有其他节点广播链路状态分组来完成，其中链路状态分组包含它所连接的链路的特征和费用。

典型的算法是 Dijkstra 算法，在课本的第 246 ～ 247 页有讨论，下面是它的伪代码：

```
Initialization:
    N' = {u}
    for all nodes V
        if V is a neighbor of u
            D(V) = c(u, v)
        else
            D(V) = INF

Loop
    find a minimum value w not in N'
    add w to N'
    for every neighbor V of w and V not in N'
        D(V) = min{D(V), D(w) + c(w, V)}
until N' == N
```

LS 算法可能产生振荡问题：课本第 247 ～ 248 页举了一个例子。

解决振荡问题的方案：

1. 强制链路费用不依赖于所承载的流量。这个方法不可接受，因为路由选择本身就是要避开高度拥塞的链路。
2. 确保并非所有的路由器都同时运行 LS 算法。

自同步现象：即使路由器初始时以同一周期但在不同时刻执行算法，算法执行时机在路由器上会变成同步并保持。

避免自同步的方法：让每台路由器发送链路通告的时间随机化。

5.3 距离向量路由选择算法

令 $d_x(y)$ 是从结点 x 到结点 y 的最低费用路径的费用，$d_x(y) = min_v\{c(x,y) + d_v(y)\}$。

距离向量：结点 x 到 N 中所有目的地 y 的费用的估计值，$\boldsymbol{D_x} = [\boldsymbol{D_x(y)} : \boldsymbol{y \in N}]$。

DV 算法中，每个结点 x 维持着如下的信息：

- 对于每个邻居，从 x 到直接相连邻居 v 的费用 $\boldsymbol{c(x,v)}$。
- 它每个邻居对目的地 y 的费用的估计值，$\boldsymbol{D_v} = [\boldsymbol{D_v(y)} : \boldsymbol{y \in N}]$。
- x 到 N 中所有目的地 y 的费用的估计值，$\boldsymbol{D_x} = [\boldsymbol{D_x(y)} : \boldsymbol{y \in N}]$。

更新信息的步骤：

1. 每个结点不时地向它的邻居发送距离向量副本。
2. 结点 x 收到距离向量副本后更新自己的距离向量：$\boldsymbol{D_x(y)} = \boldsymbol{min_v\{c(x,y) + D_v(y)\}}$。

3. 如果结点 x 的距离向量得到更新，则它向每个邻居发送其更新后的距离向量。

以上更新信息的步骤随着不断地迭代，每个距离向量都会收敛到 $d_x(y)$。

课本第 250 到第 252 页有 DV 算法工作的例子。

5.3.1 距离向量算法：链路费用改变与链路故障

当链路费用增加时，将可能产生无穷计数问题，课本的第 252 ～ 253 页有举相应的例子。

5.3.2 距离向量算法：增加毒性逆转

为了解决 3 个结点的无穷计数问题，我们可以使用毒性逆转技术。

毒性逆转技术的思想：如果 z 通过 y 路由达到目的地 x，那么 z 将告诉 y，它 (z) 到 x 的距离是无穷大的。

5.4 层次路由选择

需要将路由器组织进自治系统 (AS)，以免将网络只看作一个互联路由器的集合。

自治系统内部路由选择协议：在一个自治系统内运行的路由选择算法。

网关路由器：在一个 AS 内的一台或多台路由器将有另外的任务，也就是负责向在本 AS 之外的目的地转发分组。

自治系统间路由选择协议：从相邻 AS 获取可达性信息和向该 AS 中所有路由器传播可达性信息。