第四章 网络层

#一、网络层的几个重要概念

1. 网络层提供的两种服务
   - **面向连接的虚电路服务：**
     通信之前先**建立虚电路**，以确保双方通信所需要的一切网络资源。
     如果再使用可靠传输的网络协议，可以使发生的分组无差错按序到达终点，不缺失，不重复。
     虚电路只是一条**逻辑上的连接**，分组都沿着这条逻辑连接，按照存储转发方式传送，并不是真正建立一条物理连接
   - **无连接尽最大努力交付的数据报服务：**
     网络在发送分组的时候**不需要先建立连接**；
     每一个分组(IP数据报)**独立发送**，与前后的分组无关；
     网络层不提供是服务质量的承诺，即所**传送的分组可能出错、丢失、重复和失序，也不保证分组传送的时限**；
     由运输层负责可靠通信
     - 虚电路服务与数据报服务的对比

对比的方面	虚电路服务	数据报服务
思路	可靠通信应当由网络来保证	可靠通信应当由用户主机来保证
连接的建立	必须有	不需要
终点地址	仅在连接建立阶段使用，每个分组使用短的虚电路号	每个分组都有终点的完整地址
分组的转发	属于同一条虚电路的分组均按照同一路由进行转发	每个分组独立选择路由进行转发
当结点出故障时	所有通过出故障的结点的虚电路均不能工作	出故障的结点可能会丢失分组，一些路由可能会发生变化
分组的顺序	总是按发送顺序到达终点	到达终点时不一定按发送顺序
端到端的差错处理和流量控制	可以由网络负责，也可以由用户主机负责	由用户主机负责

2. 网络层的两个层面
   不同网络中的两个主机进行通信需要经过若干路由转发分组完成
   - **数据层面：**
     路由器**根据本路由器生成的转发表，把收到的分组从查找到的对应接口转发出去**；
     独立工作；
     采用硬件进行转发，快速；
   - **控制层面：**
     根据**路由选择协议所用的路由算法计算路由**。创建出本路由器的路由表；
     许多路由器协同动作；
     采用软件计算，缓慢；

- 软件定义网络SDN
  在控制层面，使用远程控制器计算出最佳路由，在每个路由器上生成正确的转发表，
  在数据层面，路由器查找转发表转发对应分组

#二、网络协议IP

与网际协议IPv4配套的三个协议：**地址解析协议ARP，网际控制报文协议ICMP，网际组管理协议IGMP**

1. 虚拟互联网络
   要实现对异构网络的互联互通应该使用中间设备，互联网可以使用多种异构的网络组成

层	中间设备
运输层及以上	网关 (gateway)
网络层	路由器 (router)
数据链路层	网桥或桥接器 (bridge)，交换机 (switch)
物理层	转发器 (repeater)

   使用转发器、网桥或交换机仅仅是把一个网络扩大了，其实仍然是一个网络，这种不称为网络互连，**网络互连使用的是路由器**
2. IP地址
   IP地址32位进制码，每八个一组转化为十进制数，采用**点分十进制记法**
   每台主机或者路由器的每个接口分配一个**全世界唯一的IP地址**
   IP地址采取二级结构： $IP地址 = < 网络号 >, < 主机号 >$
   IP地址在整个互联网范围是唯一的，IP地址指明了连接在某一网络的一个主机
   主机号不能为全 0 和全 1

   - 主机位全为0表示的是**这个网络中主机所处的网络地址**，主机间能不能直接通信就是要看这两台主机是不是在同一个网络中！所以这个不能分配。
   - 地址位全为1的地址是这个网络中**主机的广播地址**，当网络中一台主机向这个ip地址发送信息是，所有与这台主机在同一个网络的主机都能收到它发送的信息，所以这个也不能具体分配给一台主机。
     分类IP地址的优点与缺点：
     - 优点：
       管理简单；
       使用方便；
       转发分组迅速；
       划分子网，灵活地使用。

- 缺点：
  设计上不合理：
  大地址块，浪费地址资源；
  即使采用划分子网的方法，也**无法解决 IP 地址枯竭的问题**

  CIDR无分类域间路由选择:
- **网络前缀：网络前缀位数n不固定，可以在0到32之间任意选择**
- 地址块：将网络前缀都相同的连续地址组成一个地址块，可以分配的IP地址数目取决于网路前缀的位数：**可以指派的地址数为**$(2^{32-n}-2)$**个**，其中n是网络前缀位数；IP地址128.14.35.7/20中，20指明了网络前缀是二十位，该地址是128.14.32.0/20 地址块中的一个地址。
- **子网掩码**(地址掩码)：位数为32位，目的是让机器快速从IP地址中计算出网路地址，由一串0和1组成，**1的个数就是网络前缀的长度。网络地址就是通过IP地址与子网掩码进行且运算得到**

三个特殊的 CIDR 地址块

网络前缀长度	点分十进制	说明
/32	255.255.255.255	就是一个 IP 地址。这个特殊地址用于主机路由
/31	255.255.255.254	只有两个 IP 地址，其主机号分别为 0 和 1。 这个地址块用于点对点链路
/0	0.0.0.0	同时 IP 地址也是全 0，即 0.0.0.0/0。用于默认路由。

路由聚合
对于若干前缀存在重复的地址进行合并，统一将其聚合为一个地址

IP地址的特点：
- 每个IP地址都是**由网络前缀和主机号两部分组成**
- IP地址是**标志一台主机(或路由器)和一条链路的接口**（一个路由器至少连接两个网络，因此路由器至少有两个不同的IP地址）
- **转发器或交换机连接起来的若干个局域网仍是一个网络**
- 在IP地址中，所有分配到网络的前缀是平等的
- 同一个局域网上的主机和路由器的IP地址的网络号必须一样

3. IP地址与MAC地址
   网络层的IP数据报传输到数据链路层加上首部MAC地址与尾部组成MAC帧
   - 尽管互连在一起的网络的 MAC 地址体系各不相同，但 IP 层抽象的互联网却屏蔽了下层这些很复杂的细节
   - 只要我们在网络层上讨论问题，就能够使用统一的、抽象的 IP 地址研究主机和主机或路由器之间的通信。
   - 主机或者路由器如何知道应当在MAC帧的首部填入什么样的MAC地址?
     这里就需要使用**地址解析协议ARP**
4. 地址解析协议ARP
   实现IP通信的时候使用了两个地址：

- IP地址(网络层地址)
- MAC地址(数据链路层地址)
  地址解析协议ARP的功能就是从IP地址中解析出来MAC地址
  ARP高速缓存：
- 存放IP地址到MAC地址的映射表。<IP地址；MAC地址；生存时间；类型；...>
- 映射表动态更新(新增或超时删除)

ARP的工作：**当主机A想要从本局域网上的某个主机B发送IP数据报的时候，首先需要从ARP高速缓存上查找主机B的IP地址，找到就去取出MAC地址，并且将MAC地址写入MAC帧的目的地址，并且发送MAC帧，如果没有找到那就自动运行ARP寻找到主机B的MAC地址，并且更新ARP的高速缓存，重新查找。**

ARP查找IP地址对应的MAC地址的过程：

- 本局域网**广播发送ARP请求**(路由器不转发ARP请求)
- **ARP请求分组**：包含发送方硬件地址/发送方IP地址/目标硬件地址(未知时填0)/目标方IP地址
- **单播ARP响应分组**：包含发送方硬件地址/发送方IP地址/目的方硬件地址/目的方IP地址
- ARP分组封装在以太网帧中传播

ARP高速缓存的作用：

- 存放最近获得IP地址到MAC地址的绑定
- 减少ARP的广播的通信量
- 为进一步减少ARP的通信来那个，主机A在发送ARP请求分组时，就将自己的IP地址到MAC地址的映射写入ARP请求分组
- 当主机B收到ARP的请求分组的时候，就将主机A的IP地址以及其对应的MAC地址映射写入主机自己的ARP高速缓存中，不必再发送ARP请求

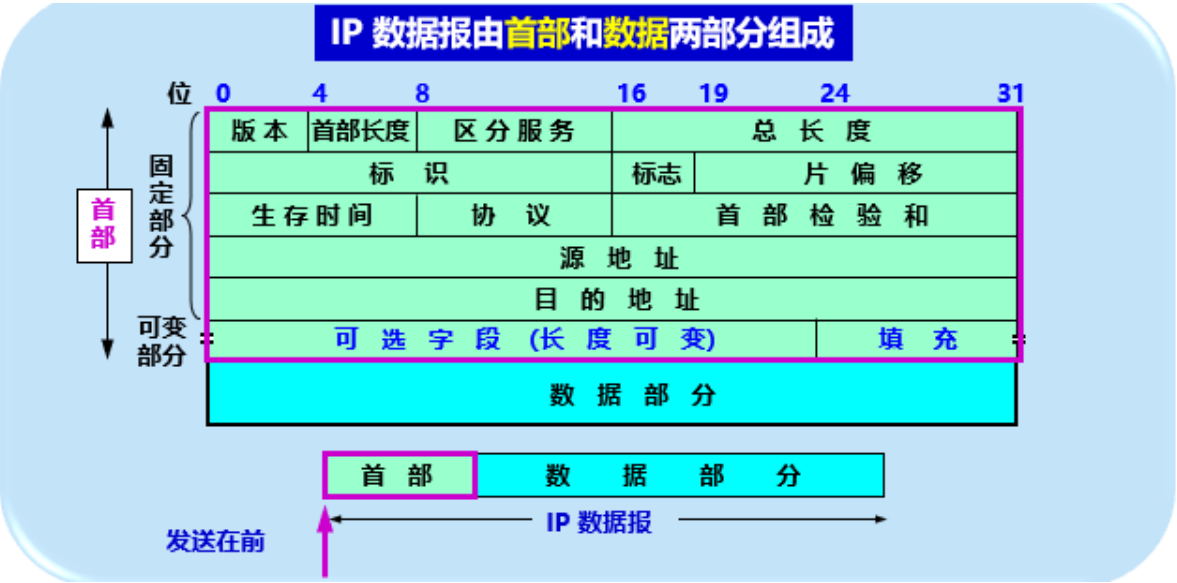
使用ARP的四种典型情况：

- 发送方是主机，要把IP数据报发送到本网络上的另一个主机。这时用ARP找到目的主机的硬件地址
- 发送方是主机，要把IP数据报发送到另一个网络上的一个主机，这时候用ARP找到本网络的一个路由器的硬件地址，剩下的工作由路由器完成
- 发送方是路由器，要把IP数据报转发给本网络的一个主机，这时候用ARP找到目的主机的硬件地址
- 发送方是路由器，要把IP数据报转发到另一个网路的一个主机，这时候用ARP找到本网络另一个路由器的硬件地址，剩下的工作由这个路由器完成

为什么使用两种地址：IP地址和MAC地址
- 不同的MAC地址之间的转化非常复杂
- 对于以太网的MAC地址的寻址也是及其困难的

- 连接到互联网的主机只需各自拥有一个唯一的 IP 地址，它们之间的通信就像连接在同一个网络上那样简单方便，即使必须多次调用 ARP 来找到 MAC 地址，但这个过程都是由计算机软件自动进行的，对用户来说是看不见的。

5. IP数据报格式



版本指的是IP协议的版本，协议版本号为4那就是IPv4，首部长度最大值为60，区分服务，只有在使用区分服务的时候该字段才会起作用，总长度是指首部与数据之和的长度，最大长度不超过最大传送单元MTU，标识是一个计数器是用来产生IP数据报的表示，片偏移：首部在原分组的相对位置，片偏移以八个字节为偏移单位，生存时间记为TTL指示数据报在网络中可以通过路由器的最大值，协议指该数据报携带的数据使用的是何种协议，首部检验和，这里只检验数据报的首部，不检验数据部分

#三、IP层转发分组的过程

1. 基于终点的转发
   分组在互联网中是**逐跳转发的**。基于终点转发是指基于分组首部中的**目的地址**的传送和转发
   为了压缩转发表的大小，转发表的主要路由是(**目的网络地址，下一跳地址**)，**查找转发表的过程就是逐行寻找前缀匹配**
   将目的地址与该网络的网络掩码进行AND运算，如果前缀匹配则路由器直接给下一跳进行交付
2. **最长前缀匹配**
   使用CIDR的时候查找转发表可能会得到不止一个匹配结果，最长匹配原则：**选择前缀最长的一个作为匹配的前缀**
   网络的前缀越长，其地址块就越小，因而路由就越具体
   把前缀最长的排在转发表的第一行
   转发表的2种特殊的路由
   - 主机路由：又叫做特定主机路由，是对特定目的主机的IP地址专门指明的一个路由，网络的前缀就是**a.b.c.d/32** ，放在转发表的最前面

- **默认路由**：不管分组的最终目的网络在哪里，都由指定的路由器R来处理，用特殊前缀0.0.0.0/0表示。**如果计算出目的网络不是主机路由，那么就交付给默认路由器**

3. 使用二叉线索查找转发表
   二叉线索：一种特殊结构的树，可以快速在转发表中找到匹配的叶节点
   从二叉线索的根节点自顶向下的深度最多有32层，每一层对应于IP地址的一位
   为了简化二叉线索的结构，可以使用唯一前缀来构造二叉线索
   为了提高二叉线索的查找速度，广泛使用各种压缩技术
   二叉线索的构造规则：**先检查IP地址的左边第一位，如果为0则第一层的节点位于根节点的左下方，否则在右下方。然后检查地址的第二位，构造出第二层的节点以此类推**
   为检查网络的前缀是否匹配，必须使二叉树线索中的每一个叶节点包含所对应的网络前缀和子网掩码
   在二叉线索中寻找IP地址
   1.根据IP地址的前缀寻找叶节点，2.将目的IP地址和该叶节点的子网掩码进行按位AND运算，看结果是否与叶节点的网络前缀相匹配。3.如果匹配就按照下一跳的接口转发该分组，否则就丢弃该分组
   根据IP地址的前缀寻找IP地址的过程中，如果在二叉线索中找不到匹配的，则说明这个地址不在二叉线索中，检查是否存在默认路由，如果有则将分组传送给默认路由，否则丢弃该分组

#四、网际控制报文协议ICMP

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是互联网的标准协议。**ICMP不是高层协议，而是IP层的协议

1. ICMP报文的种类
   两种分别为：**差错报告报文、询问报文**

几种常用的 ICMP 报文类型

ICMP 报文种类	类型的值	ICMP报文的类型
差错报告报文	3	终点不可达
	11	时间超过
	12	参数问题
	5	改变路由 (Redirect)
询问报文	8 或 0	回送 (Echo) 请求或回答
	13 或 14	时间戳 (Timestamp) 请求或回答

   差错报告报文
   收到的IP数据报的首部和后面八个字节加上ICMP的前八个字节构成ICMP差错报告报文
   不应发送ICMP差错报告报文的几种情况
   - 对ICMP差错报告不再发送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**

- **对具有特殊地址的(如127.0.0.0或者0.0.0.0)的数据报不发送ICMP差错报告报文**
  询问报文
  （1）**回送请求和回答**
- **由主机或路由器向一个特定的目的主机发送的询问**
- 收到此报文的主机必须给源主机或者路由器发送ICMP回送回答报文
- 这种询问报文用来**测试目的站是否可达，以及了解有关的状态**
  （2）**时间戳请求和回答**
- **请某台主机或者路由器回答当前的日期或时间**
- 时间戳回答报文中有一个32位的字段，其中写入整数代表从1900年1月1日到现在一共多少秒
- 时间戳请求与回答**可以用于时间同步和时间测量**

2. ICMP的应用案例
   PING用来测试两个主机之间的连通性
   使用了ICMP回送请求与回送回答报文
   是应用层直接使用网络层ICMP的例子，没有通过TCP或者UDP
   tracert用来跟踪一个分组从源点到终点的路径，它利用IP数据报中的TTL字段，ICMP时间超过差错报告报文和ICMP终点不可达差错报告报文实现从源点到终点的路径的跟踪

#五、IPv6

IP是互联网的核心协议
IPv4地址耗尽，为了解决问题，采用了具有**更大的地址空间的IPv6**

1. IPv6的基本首部
   IPv6仍然支持无连接的传送
   将协议数据单元PDU称为分组

   主要的变化：

   - **更大的地址空间，将地址从32位增大到128位**
   - **扩展的地址层次结构**，可划分为**更多的层次**
   - **灵活的首部格式**，定义了许多可选的**扩展首部**
   - **改进的选项**，允许数据报中包含选项的控制信息，其选项放在有效载荷中
   - **允许协议继续扩充**，更好的适应新的应用
   - **支持即插即用**，自动配置，不需要使用DHCP
   - **支持的资源预分配**，支持实时视像等要求，保证一定的带宽和时延的应用
   - **IPv6首部改为8字节对齐**，首部长度必须是8字节的整数倍

   IPv6数据报有两大部分组成：
   - 基本首部：固定40个字节，八个首部字段。有版本(4)、通信量类(8)、流标号(20)、有效载荷长度(16)、下一个首部(8)、跳数限制(8)、源地址(128)、目的地址(128)

IPv6 对首部的主要更改

• 取消了首部长度字段； • 取消了服务类型字段； • 取消了总长度字段，改用有效载荷长度字段；	• 把 TTL 字段改称为跳数限制字段； • 取消了协议字段，改用下一个首部字段； • 取消了检验和字段； • 取消了选项字段，而用扩展首部来实现选项功能。

- 有效载荷：有效载荷也称为净负荷，有效载荷允许有多个或者0个扩展首部，再后面是数据部分。六种扩展首部：逐跳选项、路由选择、分片、鉴别、封装安全有效载荷、目的站选项

2. IPv6的地址

三种基本类型：

- 单播：传统的点对点通信
- 多播：一点对多点的通信
- 任播：IPv6增加的一种类型。任播的终点是一组计算机，数据报在交付的时候只能交付一个，一般是按照路由算法最近的一个

IPv6将实现了IPv6的主机和路由器均称为节点，一个节点可能有多个与链路连接的接口，**IPv6地址是分配给接口的，一个具有多个接口的节点可以有多个单播地址**，其中任何一个地址都可以当做到达该节点的目的地址

冒号十六进制记法：

冒号十六进制记法中16位的值用**十六进制**值表示，**各个值之间用冒号进行分隔**，并且使用**零压缩，若该分段里面都是0那么只需要写一个：：来代替，在任一地址中只能使用一次零压缩**

IPv6 地址分类

地址类型	二进制前缀	IPv6记法
未指明地址	00...0（128位），仅此一个	::/128
环回地址	00...1（128位），仅此一个	::1/128
多播地址	11111111（8位），功能和 IPv4 的一样	FF00::/8
本地链路单播地址	1111111010（10位），未连接到互联网，不能和互联网上的其他主机通信	FE80::/10
全球单播地址	除上述四种外，所有其他的二进制前缀	

IPv6单播地址划分方法：全球路由选择前缀(n)，子网标识符(m)，接口标识符(128-n-m)

3. 从IPv4到IPv6过渡

方法：**逐步演进，向后兼容**

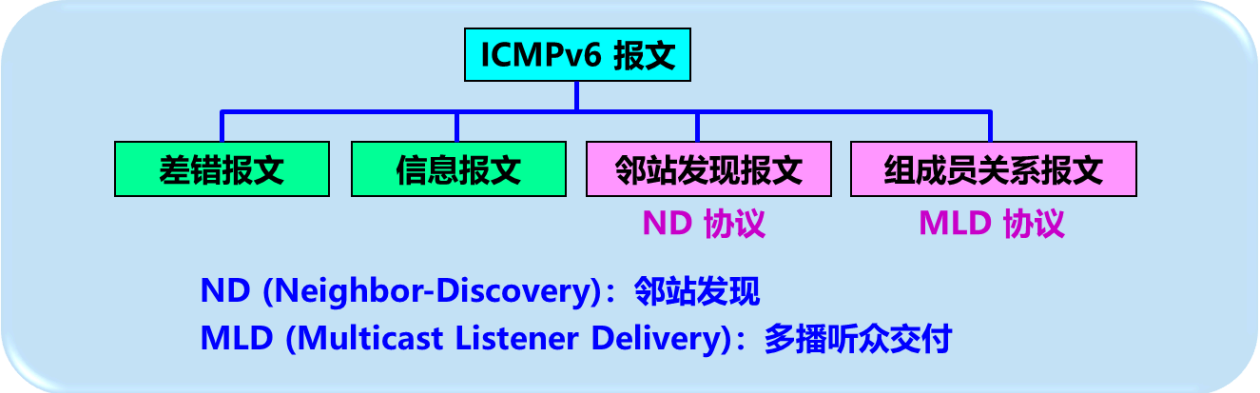
向后兼容指的是：**IPv6系统必须能够接收和转发IPv4分组，并且能够为IPv4分组选择路由**

两种过渡策略：

- 使用**双协议栈：将IPv6数据报映射到IPv4数据报上**

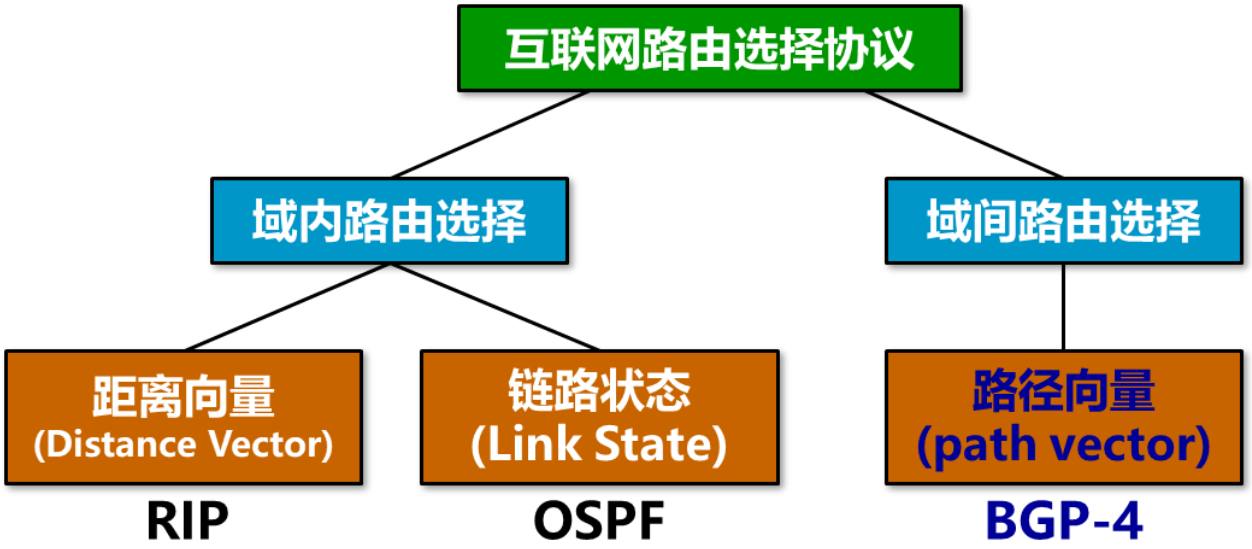
- 使用**隧道技术：使用IPv6隧道**

4. ICMPv6



#六、互连网的路由选择协议

1. 路由选择协议的基本概念
   路由选择协议属于**网络层控制层面**的内容
   路由选择算法：
   静态路由选择策略，非自适应路由选择，简单，开销小
   动态路由选择策略，自适应路由选择，复杂，开销大
   两大路由选择协议：
   内部网关协议IGP，在**一个自治系统内部使用的路由选择协议，常用：RIP，OSPF**
   外部网关协议EGP，在**不同自治系统之间进行路由选择时候使用的协议**，使用最多的是BGP-4
   自治系统之间的路由选择叫做**域间路由选择**
   自治系统内部的路由选择叫做**域内路由选择**
   自治系统AS：是在单一技术管理下的许多网络，IP地址以及路由器1，而这些路由器使用的一种自治系统内部的路由选择协议和共同的度量，每一个AS对其他AS表现出的是一个单一和一致的路由选择策略



2. **内部网关协议RIP**
   路由信息协议RIP是一种分布式的、基于距离向量的路由选择协议，是互联网的标准协议，最大的优点是：简单，**要求网路中的每个路由器都要维护从它自己到其他每一个目的网络的距离记录**
   RIP距离的规定：**路由器到直接连接的网络的距离=1**，路由器到非直接连接的网络的距离=所经过的路由器数+1。RIP协议中距离也称为跳数，每经过一个路由器，跳数加一
   RIP所选择的最佳路由是距离最短的路由，**一条路径最多含有15个路由器，距离的最大值16时相当于不可达**，RIP不能在两个网络之间同时使用多条路由，只能选择距离最短的路由
   RIP协议的三个特点：
   - **仅和相邻的路由器交换信息**
   - **交换的信息是当前本路由器所知道的全部信息，即自己的路由表**
   - **按固定时间间隔交换路由信息，例如每隔30秒。当网络拓扑结构发生变化也能及时向相邻的路由器通告拓扑变化的路由消息**
     路由表建立过程：
   - 路由器刚开始工作的时候路由表是空的
   - 然后得到直接相连的网络距离
   - 然后根据数目有限的相邻路由器交换并且更新路由信息
   - 经过若干次更新，所有的路由都最终知道到达本自治系统中任何一个网络的最短距离和吓一跳的路由器的地址
   - RIP的收敛速度较快，这种收敛就是在自治系统中所有的结点都能够得到正确路由选择信息的过程
     对于每一个相邻的路由器(假设地址为X)发送过来的RIP报文，路由器：
   - 修改RIP报文中所有的项目，将下一跳字段都修改为X，并且把所有的距离字段的值加一
   - 对修改后的RIP报文中的每一个项目重复一下步骤
     - 若路由表中**没有目的网络N，则把该项目添加到路由表中**
     - 若**原路由表中网络N的下一跳路由器为X，则用收到的项目替换原路由表中的项目**
     - 若**收到的项目的距离小于路由表中的距离，则用收到的项目更新原路由表中的项目**，否则什么也不做
   - **三分钟还没收到相邻路由器的更新路由表，则将此相邻路由器记为不可达路由器，即将距离置为16**
   - 返回
     RIP2报文
     组成：首部和路由两个部分
     路由部分：由于若干个路由信息组成，每个路由信息共20个字节，地址族标识符，用来标志所使用的地址协议
     路由标记填入自治系统的号码
     后面为具体路由，指的某个网络地址，该网络的子网掩码，下一跳的路由器地址以及到该网络的距离
     一个RIP报文最多可以包含25个路由，所以一个RIP报文的最大长度为504个字节

RIP协议的特点：好消息传播速度快，**坏消息传播速度慢**
RIP协议的优缺点：

- 优点：
  实现简单，开销较小
- 缺点：
  网络规模有限，最大的距离为15
  交换的路由信息为完整路由表，开销较大
  坏消息传播慢，收敛时间长

3. 内部网关协议OSPF
   开放的最短路径优先OSPF是为了克服RIP缺点在1989年开发出来的
   原理很简单，根据Dijkstra提出了最短路算法SPF，采用分布式链路状态协议
   三个主要特点：
   - **采用洪泛法，向本自治系统的所有路由器发送信息**
   - **发送的信息是与本路由器相邻的所有路由器的链路状态**，但这只是路由器所知道的部分信息
   链路状态：说明本路由器都和哪些路由器相邻，以及该链路的度量
   - 当链路状态发生变化或者每隔一段时间。路由器才使用洪泛法向所有的路由器发送此信息
   链路状态数据库：每个路由器最终都能建立全网的拓扑结构图，在全网的范围内是一致的，**每个路由器使用链路状态数据库的数据来构造自己的路由表**。链路状态数据库能够较快的进行更新，使得各个路由器可以及时更新路由器，重要的特点是：**OSPF更新过程收敛速度快**
   OSPF将自治系统划分为两种不同的区域，多个路由器相互连接的区域为主干区域(分为区域边界路由器ABR，主干路由器BR，自治系统边界路由器ASBR)
   划分区域的优点和缺点：
   - 优点：
   减少了整个网络上的通信量，每个区域内部交换路由信息的通信量大大减小
   减少了需要维护的状态数量
   - 缺点：
   交换信息的种类增多了
   使OSPF协议更加复杂了
   OSPF的五种分组类型：
   - 问候分组
   - 数据库描述分组
   - 链路状态请求分组
   - 链路状态更新分组
   - 链路状态确认分组
   OSPF分组使用IP数据报进行传送，将IP数据报首部加到OSPF分组前面，就构成IP数据报。其中IP数据报的协议字段的值为89

OSPF工作过程：
- **确定邻站可达**
- 相邻的路由器每隔10秒钟交换一次问候分组

- 若40秒钟没有收到某个相邻的路由器发来的问候分组，则可认为该相邻路由器是不可达的
- **同步链路状态数据库**
- 同步指的是不同路由器的链路状态数据库的内容是一样的
- 两个同步的路由器叫做完全邻接的路由器
- **更新链路状态**
- **只要链路状态发生变化，路由器就使用链路状态更新分组，采用可靠的洪泛法向全网更新链路状态**
- 为了确保链路状态数据库与网络的状态保持一致，OSPF还规定，**每间隔一段时间，如30分钟，就要刷新一次数据库中的链路状态**

OSPF链路状态只涉及相邻路由器，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候，OSPF协议要比距离向量协议RIP好得多**，OSPF没有"坏消息传得慢"的问题，收敛速度快

指定路由器DR
- 多点接入的局域网采用了指定的路由器DR的方法，使广播的信息量大大减少
- 指定的路由器代表局域网那个上所有的链路向连接该网络的各路由器的发送状态信息

4. **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协议的主要特点：
- 用于自治系统AS之间的路由选择
- 只能是**力求选择出一条能够到达目的网络且比较好的路由**，而并非计算出一条最佳路由
- 互联网的规模太大，使自治系统AS之间路由选择十分困难
- 自治系统AS之间的路由选择必须考虑有关策略
- 采用了路径向量路由选择协议

BGP发言者(边界路由器)在AS之间交换信息

eBGP连接和iBGP连接

在AS之间，BGP发言者在半永久性TCP连接上建立BGP会话，这种连接又叫做eBGP连接

在AS内部，任何相互通信的路由器之间必须要有一个逻辑连接(TCP)连接，AS内部所有路由器之间的通信是全连通的，这种连接通常称为iBGP连接

eBGP连接：**运行eBGP协议，在不同的AS之间交换路由信息**

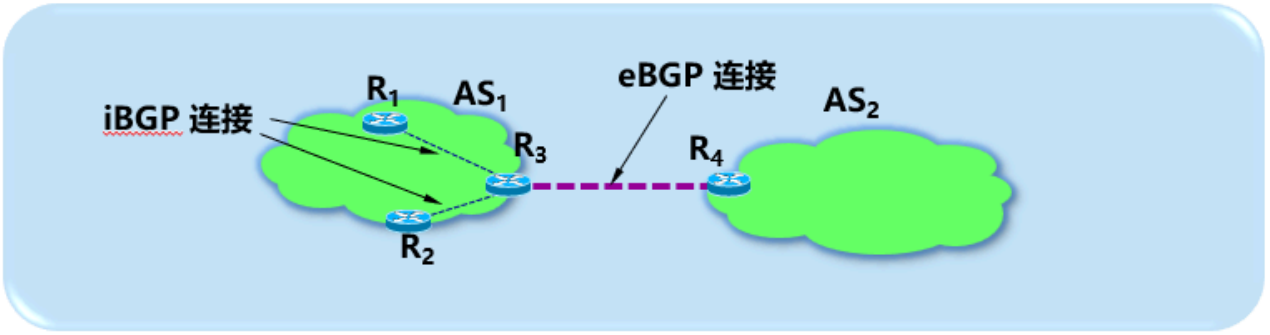
iBGP连接：**运行iBGP协议，在AS内部路由器之间交换BGP路由信息**

在AS内部运行：内部网关协议IGP(OSPF或者RIP)，协议iBGP

在AS外部运行：协议eBGP

同一个协议BGP使用的报文类型、使用的属性、使用的状态机等都完全一样

但它们在通报前缀时采用的规则不同：
- 在**eBGP连接的对等端得知的前缀信息，可以通报给一个iBGP连接的对等端，反过来也是可以的**
- **但是iBGP连接的对等端得知的前缀信息，则不能够通报给另一个iBGP对等端**



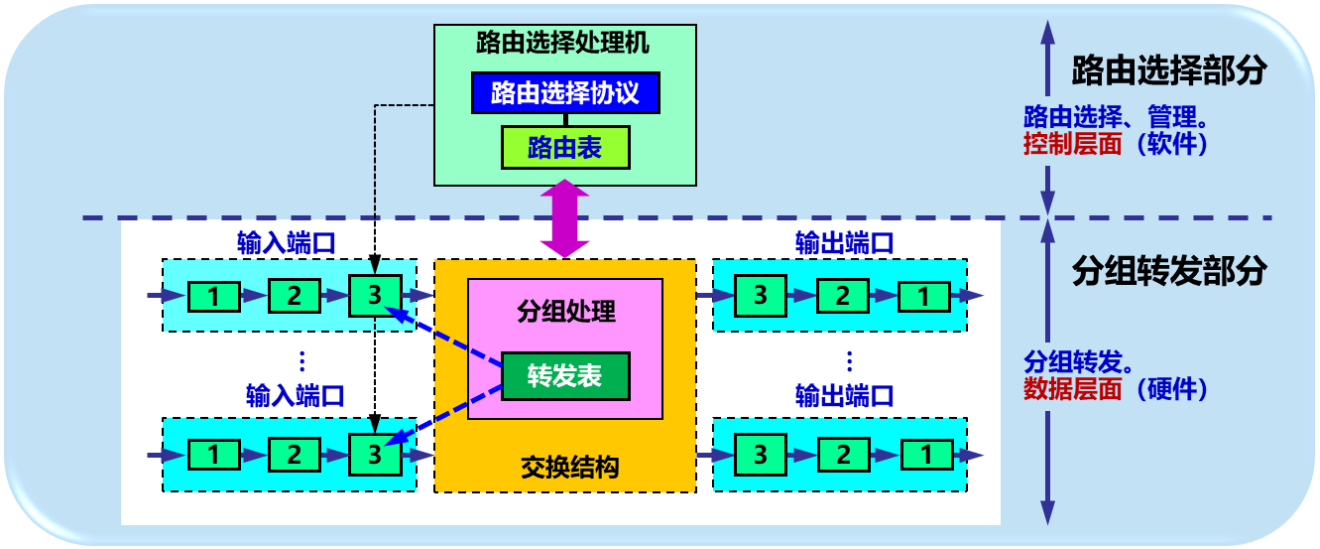
- **R_3 从 eBGP 连接的对等端 R_4 得到的前缀信息可以通报给 iBGP 连接的对等端 R_1 或 R_2。**
- **R_3 从 iBGP 连接的对等端 R_1 和 R_2 得到的前缀信息可以通报给 eBGP 连接的对等端 R_4。**
- **但 R_3 从 iBGP 连接的对等端 R_1 得到的前缀信息不允许再通报给另一个 iBGP 连接的对等端 R_2。**

BGP路由=【前缀，BGP属性】=【前缀，AS-PATH,NEXT-HOP】
前缀：指明到哪一个子网
BGP属性：自治系统路径AS-PATH，下一跳NEXT-HOP
三种不同的自治系统AS：
末梢AS：不会将来自其他AS的分组再转发给另一个AS，必须向所连接的AS付费
多归属AS：同时连接到两个或者两个以上的AS，增加连接的可靠性
穿越AS：为其他AS有偿转发分组
对等AS：经过事先协商的两个AS彼此之间发送或者接收分组都不收费
BGP路由**当检查到说到的BGP路由的AS-PATH中已经有了自己，就立即删除这条路由，从而避免圈子路由的出现**

BGP的路由选择：
- 本地偏好值(local preference)最高的路由
- **AS跳数最小的路由**
- 使用热土豆路由选择算法
- 路由器BGP的ID数值最小的路由，具有多个接口的路由器有多个IP地址，BGP ID就使用该路由器的IP地址数值中最大的一个

BGP-4的四种报文：
- OPEN(打开)：用来与相邻的另一个BGP发言者建立关系，使通信初始化
- UPDATE(更新)：用来通告某一路由信息，以及列出需要撤销的多条路由
- KEEPALIVE(保活)：用来周期性地证实临站的连通性
- NOTIFICATION(通知)：用来发送检测到的差错

5. 路由器的构成
路由器工作在网络层，用于互连网络
是互联网中的关键设备
路由器的主要工作：**转发分组**，把从某个输入端口收到的分组，按照分组要去的目的网络，把分组从路由器的某个合适的输出端口转发给下一跳路由器



转发与路由选择的区别

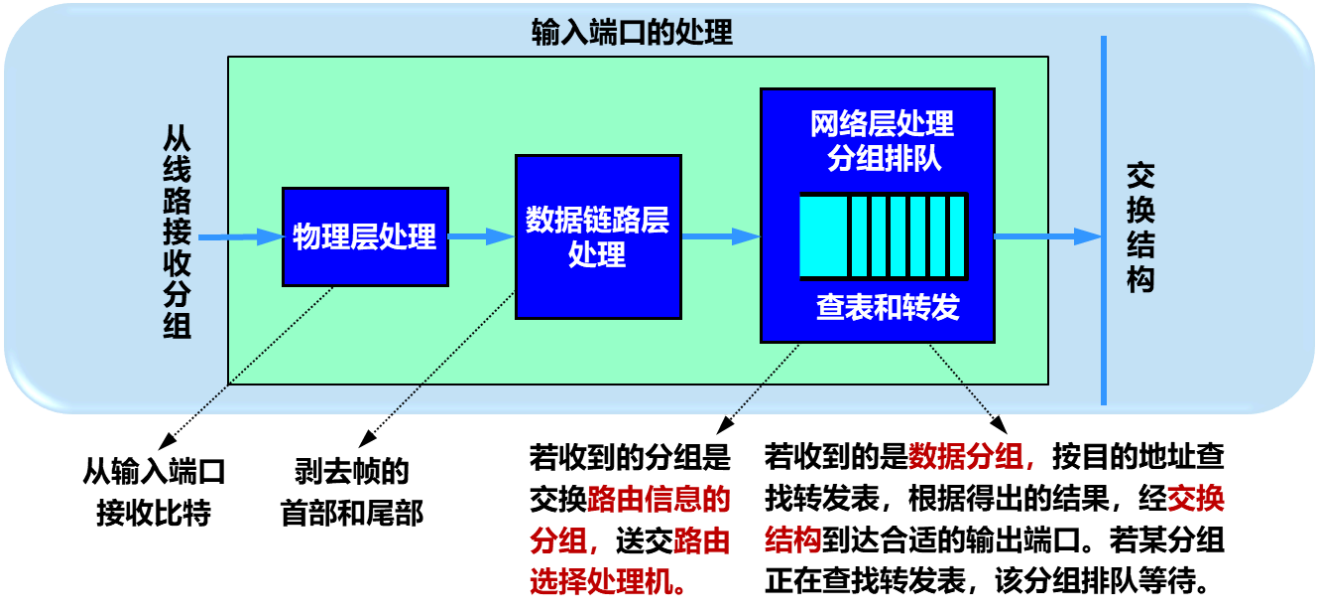
转发：

- **根据转发表将用户的IP数据报从合适的端口转发出去**
- **仅涉及到一个路由器**
- 转发表是**从路由表中得出的**
- 转发表必须包含完成转发功能所必须的信息，每一行必须包含从要到达目的网络到输出端口和某些MAC地址信息(如下一跳的以太网地址)的映射

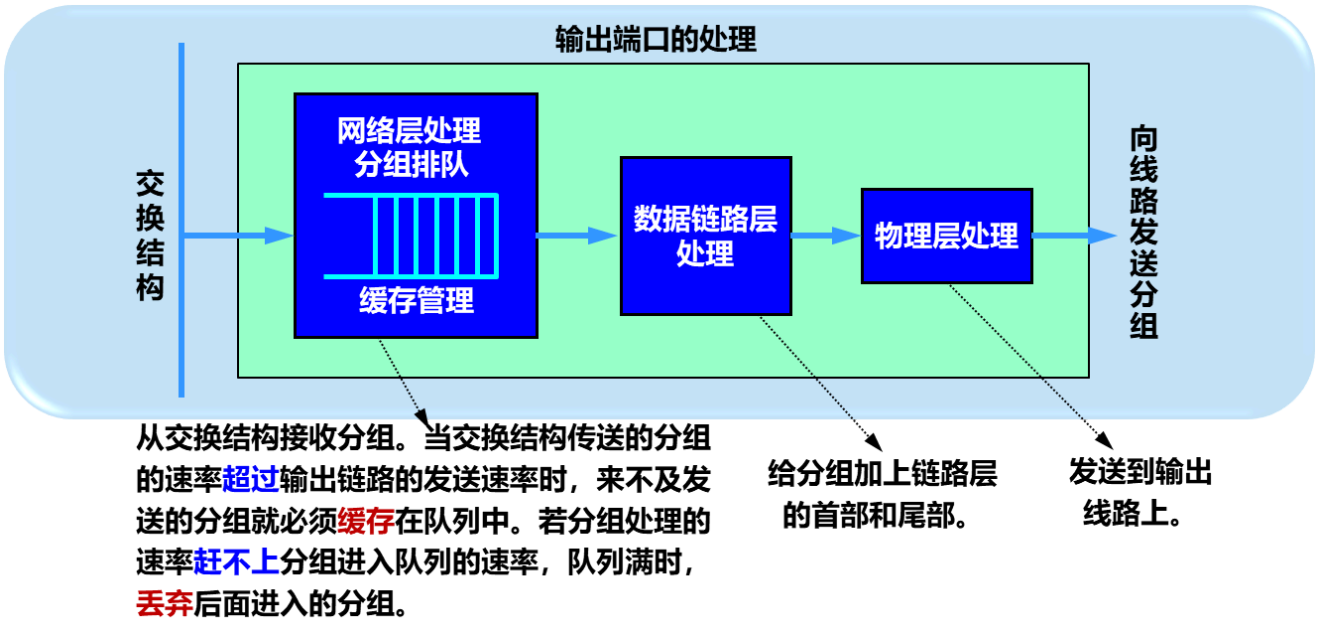
路由器选择：

- **按照路由选择算法，根据网络拓扑结构的变化情况，动态地改变所选择的路由，并由此构造出整个路由表**
- 涉及到很多路由器
- **路由表一般进包含从目的网络到下一跳(用IP地址表示)的映射**

输入端口对收到的分组的处理



输出端口对收到的分组的处理



交换结构：常用的交换方法有三种：**通过存储器，通过总线，通过纵横交换机构**

- 通过存储器：
- 当路由器的某个输入端口接收到一个分组的时候，就用中断的方式通知路由器选择处理机，然后分组就从输入端口**复制到存储器中**
- 路由器处理机从分组首部提取目的地址，查找路由表，再将分组复制到合适的输出端口的缓存中
- 若存储器的带宽(读或者写)为每秒M个分组，那么路由器的交换速率(即分组从输入端口传送到输出端口的速率)一定小于M/2
- 通过总线：
- 数据报从输入端口通过**共享的总线**直接传送到合适的输出端口，而**不需要**路由选择机的干预
- 当分组到达输入端口时若发现**总线忙**，则被阻塞而不能通过交换结构，并在**输入端口排队等待**
- 因为每一个要转发的分组都需要通过这一条总线，因此**路由器的转发带宽就受到总线速率的限制**
- 通过纵横交换结构
- 常被称为互连网络
- 它有**2N条总线，控制交叉节点可以使N个输入端口和N个输出端口连接**
- 当输入端口收到一个分组的时候，就将它发送到水平总线上
- 若通向输出端口的垂直总线空闲，则将垂直总线于水平总线接通，把该分组转发到这个输出端口。若输出端口已被占用，分组在输出端口排队排队
- 特点：**是一种无阻塞的交换结构，分组可以转发到任何一个输出端口**，只要这个输出端口没有被别的分组占用

#七、IP多播

1. IP多播的基本概念
   多播目的是为了更好地支持一对多通信，可以**大大节约网络资源**
   在互联网上进行多播就叫做IP多播，互联网范围的多播要靠路由器来实现，**能够运行多播**

协议的路由器叫做多播路由器
多播IP地址：多播的标识符就是IP地址中的多播地址，地址范围是：
224.0.0.0~239.255.255.255
每一个D类地址标志一个多播组，**多播地址只能作为目的地址**
多播数据报与IP数据报的区别：
目的地址：使用D类IP地址
协议字段=2，表明使用网际组管理协议IGMP
尽最大努力交付，不保证一定能够交付多播组内的所有成员
对多播数据报**不产生ICMP差错报文，在PING命令后键入多播地址永远不会收到响应**

2. 在局域网进行硬件多播
   IANA拥有以太网地址块的24位为00-00-5E
   TCP/IP协议使用的以太网地址块的范围是从00-00-5E-00-00-00到00-00-5E-FF-FF-FF
   IANA只拿出01-00-5E-00-00-00到01-00-5E-7E-FF-FF(2^{23}个地址)作为以太网的多播地址，或者说，在48位的多播地址中，前25位不变，只有后23位可以用作多播
   收到多播数据报的主机，还要在IP层对IP地址进行过滤，把不是本主机要接收的数据报抛弃
3. **网际组管理协议IGMP**和**多播路由选择协议**
   网际组管理协议IGMP使多播路由器知道多播成员的信息
   多播路由协议，使多播路由器协同工作，把多播数据报用最小的代价传送给多播组的所有成员
   IGMP使多播路由器知道多播组成员信息：**IGMP协议是让连接在本地局域网上的多播路由器知道本局域网上是否有主机参加或者退出了某个多播组**，IGMP不知道IP多播组包含的成员数，也不知道这些成员都分布在哪些网络上

多播转发的特点：

- 多播转发必须**动态地适应多播成员的变化**，因为每一台主机可以随时加入或则离开一个多播组
- 多播路由器在转发多播数据报的时候，不仅仅根据多播数据报的目的地址，还要考虑这个多播数据报从哪里来，到哪里去
- 多播数据报可以由没有加入多播的主机发出，也可以通过没有组成员的接入网络

IGMP的工作可以分为两个阶段：
第一阶段：**加入多播组**

- **当某一个主机加入多播组的时候，该主机向多播地址发送IGMP报文，声明自己要成为该组的成员**
- **本地的多播路由器收到IGMP报文之后，将组成员关系转发给互联网的其他多播路由器**
  第二阶段：探寻组成员变化情况
- **本地多播路由器周期性地探寻本地局域网上的主机**，以便知道这些主机是否还继续是该组的成员
- 只要对某个组由一个主机响应，则多播路由器救人位该组是活跃的
- 但是一个组经过**几次探询后仍然没有一个主机响应**，则不再将该组的成员关系转发给其他的多播路由器

IGMP采用了一些具体措施，以**避免增加大量开销**

- 对所有通信都使用IP多播，**只要有可能，都使用硬件多播来传送**
- 对所有的组**只发送一个请求信息的询问报文**，默认询问速率是每125秒发送一次
- 当同一个网络上连接有多个多播路由器的时候，能迅速和有效地选择其中一个来探询主机的成员关系
- 分散响应：在IGMP的询问报文中有一个数值N，它指明一个最长响应时间(默认为10秒)。收到询问时，主机在0到N之间随机选择发送响应所需要经过的延迟。若一台主机同时参加了几个多播组，则主机对每一个多播组选择不同的随机数，**对应最小时间延迟的响应应该最先发送**
- 采用抑制机制。同一个组内的每个主机都要监听响应，**只要有本组的其他主机先发送了响应，自己就不再发送响应了**

多播路由选择：
实际上就是要找出以源主机为根节点的多播转发树
不同的多播组对应不同的多播转发树
同一个多播组，对不同的源点也会有不同的多播转发树
M个源，N个多播组，需要M * N棵以源为根的多播转发树
转发多播数据报时使用三种方法：

- 洪泛与剪除
  适合于较小的多播组，所有组成员接入的局域网也是相邻接的。开始时，路由器转发多播数据报的时候使用的是洪泛的方法，为了避免兜圈子，采用反向路径广播RPB的策略
  RPB要点：
  检查转发的数据报是否是经过最短路径传送来的，如果不是就丢弃。如果**存在多条同样长度的最短路径，选择IP地址最小的**，最后就得到了以源为根节点的，用来转发多播数据报的多播转发树
  剪枝：如果在多播转发树上的某个路由器发现它的下游树枝(即叶节点方向)已**没有多播组的成员，就把它和下游的树枝一起剪除**
  嫁接：当某**个树枝又新增加的组成员的时候，可以再接入多播转发树上**
- 隧道技术
  隧道技术用于多播组的位置在**地理上很分散**的情况
- 基于核心的发现技术
  对于多播组的大小在较大范围内变化都适合
  对于每一个多播组G**指定一个核心路由器**，并给出它的IP单播地址。**核心路由器创建出对应多播组的转发树**，为一个多播组构建一棵转发树，而不是为每个（源，组）组合构建一棵转发树
  如果一个路由器向核心路由器发送数据，那么它在途中经过的每一个路由器都要检查其内容
  当数据报到达参加多播组的路由器的时候，就由该路由器处理该数据，如果发送的是一个多播数据报，接收的路由器就会向多播组的其他成员发送该数据报。如果发送的是请求加入的数据报，就会将这个信息加入它的路由中，并且使用隧道技术向源路由器转发每一个多播数据报的副本

几种多播路由器选择协议：

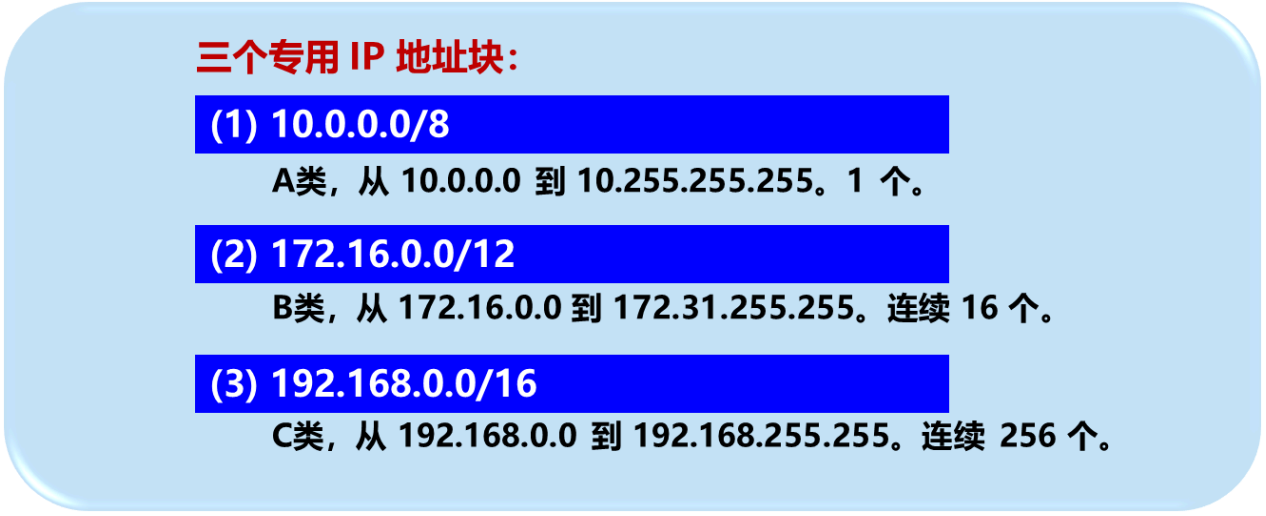
- 距离向量多播路由选择协议DVMRP，互联网上使用的第一个多播路由器选择协议
- 基于核心的转发树CBT
- 开放最短通路优先的多播扩展MOSPF
- **协议无关多播-稀疏方式PIM-SM，唯一成为互联网标准的协议**
- 协议无关多播-密集方式PIM-DM

#八、虚拟专用VPN和网络地址转换NAT

1. 虚拟专用网VPN
   由于IP地址紧缺，一个机构能够申请到的IP地址数量远小于本机构拥有的主机数，考虑到互联网并不安全，机构也不会将所有的主机都接入到外部的互联网，如果一个机构内部的计算机通信也是采用TCP/IP协议，那么这些**仅在机构内部使用的计算机就可以由本机构自行分配IP地址**

   - 本地地址：仅在机构内部使用的IP地址，可以由本机自行分配，而且不需要向互联网的管理机构申请
   - 全球地址：全球唯一的IP地址，必须向互联网的管理机构申请
     专用地址只能用作本地地址，而不能用作全球地址
     互联网中的所有路由器**对目的地址是专用地址的数据报一律不进行转发**

   专用IP地址



   采用专用IP地址的称为专用互联网或者本地互联网，或者更简单点，就叫做专用网
   专用IP地址也叫做可重用地址
   利用公用互联网作为本机构各专用网之间的通信载体，这样的专用网又叫做虚拟专用网
   专用网：指这种网络是为本机构的主机用于机构内部的通信，而不是用于和网络外非本机构的主机通信
   虚拟：表示没有使用通信专线，只是在效果上和真正的专用网一样
   如果专用网的不同网点之间的通信必须经过公用的互联网，但是又有保密的要求，那么所有通过互联网传送的数据都必须加密

必须为每一个场所购买专门的硬件和软件，并且进行配置，使每一个场所的VPN系统都知道其他场所的地址

VPN类型：
- 内联网：同一个机构的内部网络所构成的VPN
- 外联网：一个机构和某些外部机构共同建立的
- 远程接入VPN，允许外部流动员工通过直接接入VPN建立VPN隧道访问公司的内部网络，好像就是公司内部本地网络直接访问一样

2. 网络地址转换NAT
需要在专用网连接到互联网的路由器上安装NAT软件
装有NAT软件的路由器叫做NAT路由器，它至少有一个有效的外部全球IP地址
所有使用本地地址的主机在和外界通信的时候，都要在**NAT路由器上将本地地址转换为全球IP地址，才能和互联网连接**
在内部主机与外部主机通信的时候，在NAT路由器上发生了两次地址转换
离开专用网的时候：替换源地址，将内部地址替换为全球地址
进入专用网的时候，替换目的地址，将全球地址替换为内部地址

当NAT路由器具有n个全球IP地址的时候，专用网内最多可以同时拥有n台主机接入到互联网
可以使用专用网内较多的数量的主机轮流使用NAT路由器有限数量的IP地址
通过NAT路由器的路由器的通信必须是由专用网的主机发起的，因此专用网内部的主机不能充当服务器使用
NAT并不能节省IP地址

NAPT可以使多台拥有本地地址的主机，共用一个全球IP地址，同时和互联网上的不同主机进行通信
使用运输层端口号的NAT叫做网络地址与端口转换NART，而不是使用端口的NAT就叫做传统的NAT
NART把专用网内不同的源IP地址转换为相同的全球IP地址，将TCP源端口号转换为新的TCP端口号
收到从互联网发来的应答的时候，从IP数据报的数据部分找出运输层的端口号，从NAPT转换表中找到真正的目的主机

#九、课后作业

1. 路由器转发分组的根据是报文的IP地址
2. IP协议和ARP协议是属于网络层的协议
3. 在IP字段中，与首部与分片和重组有关的字段是标识，标志，片偏移
4. CIDR地址块192.168.10.0/20所包含的IP地址范围是：20个网络位，则有12个地址位，12-8=4所以在第三个字节从右往左数四位里面最多有四个1,0到15
11000000.10101000.0000__.__
=192 .168 . 0~15 .0~255

5. 当主机需要发送信息，但是ARP表中并没有目的IP地址与MAC地址的映射关系的时候需要启用ARP请求
6. 主机A发送IP数据报到主机B，途中经过五个路由器，在此过程中使用了六次ARP。每一跳之前都要使用一次ARP
7. ARP的功能是根据IP地址查询MAC地址
8. 用于域间路由协议是BGP
9. 一个网络连接的主机数为$2^{\text{主机位数}}-2$