计算机网络复习整理（第四章）

###### 网络层

本章最重要的内容是：

1. 虚拟互连网络的概念。
2. IP地址与物理地址的关系。
3. 传统的分类的IP地址（包括子网掩码）和无分类域间路由选择CIDR。
4. 路由选择协议的工作原理

4.1网络层提供的两种服务

有些人认为应当模仿打电话所使用的面向连接的通信方式，当两台计算机进行通信时，也应当先建立连接（在分组交换中是建立一条**虚电路VC**），以预留双方通信所需的一切网络资源。然后双方就沿虚电路发送分组，分组首部不需要填写完整的目的主机地址，只需填写虚电路编号（一个不大的整数），因此减小了分组开销。这种通信方式再使用可靠的网络协议就可使发送的分组无差错按序到达终点也不丢失、不重复在通信结束后要释放连接。

互联网的设计者认为计算具有很强的差错处理能力，因此采用了这样的设计思路：**网络层向上只提供简单灵活的、无连接的、尽最大努力交付的数据报**（即IP数据报或分组）**服务**。

网络在发送分组时不需要先建立连接，每个分组独立发送。网络层不提供服务质量的承诺，因此所传送分组可能出错、丢失、重复和失序，也不保证到达时限。如果主机（端系统）中的进程之间的通信需要的是可靠的，那么就由网络的主机中的运输层负责（差错处理、流量控制等）。

OSI体系支持者曾推出虚电路服务的著名标准X.25建议书，但早已过时。

虚电路服务：

1. 可靠通信应由网络层保证。
2. 必须有连接。
3. 终点地址仅在建立连接阶段使用，每个分组使用短的虚电路号。
4. 属于同一条虚电路的分组均按照同一路由转发。
5. 所有出故障的结点的虚电路均不能工作。
6. 分组按序到达。
7. 差错控制和流量控制可以由网络负责，也可以由用户主机负责。

数据报服务：

1. 可靠通信应由用户主机保证。
2. 不要连接。
3. 每个分组都有终点的完整地址。
4. 每个分组独立选择路由转发。
5. 出故障的结点可能会丢失分组，一些路由可能会发生变化。
6. 不一定按序到达。
7. 差错控制和流量控制由主机负责。

4.2 网际协议IP

与IP协议配套的三个协议：

1. **地址解析协议ARP**
2. **网际控制报文协议ICMP**
3. **网际组管理协议IGMP**

逆地址解析协议RARP已被淘汰。

ICMP，IGMP经常要使用IP协议,IP协议经常要使用ARP协议。由于网际协议IP是用来使互连起来的许多计算机网络能够进行通信的，因此TCP/IP体系中的网络层也叫**网际层**或**IP层**。

1. 虚拟互连网络

由于用户需求是多样的，**没有一种单一的网络能适应所有用户的需求**。

将网络互相连接起来要使用一些**中间设备**，根据层次可划分为：

1. 物理层的中间设备叫做**转发器**
2. 数据链路层的中间设备叫做**网桥**或**桥接器**
3. 网络层的中间设备叫**路由器**
4. 网络层以上的中间设备叫**网关**，网关连接不兼容的系统要在高层进行协议转换

由于参加互连的计算机都使用相同的**网际协议IP**，因此可以把互连后的计算机网络看成一个**虚拟互连网络**，也就是逻辑互联网络，互连起来的各种物理网络的异构性本来是客观存在的，但利用IP协议可以使这些性能各异的网络**在网络层看起来好像是一个统一的网络**。

当源主机H1要把一个IP数据报发送给目的主机H2时，根据分组交换的存储转发概念，H1先查看自己的路由表，如果目的主机就在本网络则采用**直接交付**，如果不是就转发给路由器进行**间接交付**，直到某路由器知道自己是和H2连接在同一个网络上就把数据表**直接交付**给H2。路由器的协议栈只有下三层，**互联网可以由多种异构网络互连组成**。

1. 分类的IP地址

**IP地址及其表示方法**

整个的互联网就是一个**单一的、抽象的网络**。IP地址就是给互联网的每一台主机（或路由器）的每一个接口分配一个在全世界内范围唯一的32位标识符。IP地址由**互联网名字和数字分配机构ICANN**分配。

IP地址的方法经过三个阶段：

1. **分类的IP地址**。最基本的编址方法。
2. **子网的划分**。最基本编址方法的改进。
3. **构成超网**。比较新的无分类编址方法。

所谓分类的IP地址就是将IP地址划分为若干个固定类，其中的A类、B类和C类地址都由两个固定长度的字段组成，其中第一个字段是**网络号**，标志主机或路由器所连接到的网络。一个网络号必须是唯一的。第二个字段是**主机号**，它标志主机或路由器，一台主机号在它前面网络号所指明的网络范围内必须是唯一的。因此一个IP地址**在整个互联网范围内唯一**。

A类、B类和C类地址的网络号字段分别为1个、2个和3个字节长，而在网络号字段的最前面有1-3位的**类别位**，分别规定为0，10和110。

A类、B类和C类地址的主机号字段分别为3个、2个和1个字节长。

D类地址（前4位是1110）用于**多播**（一对多通信）。

E类地址（前4位是1111）保留位以后使用。

IP地址不仅指明主机，还指明了主机所连接的网络。当某个单位申请到一个IP地址时，实际上是获得了具有同样网络号的一块地址。其中各台主机号则由该单位自行分，只要保证在范围内无重复主机号即可。

将32位IP地址的每8位插入一个空格（**机器中并没有**），并用等效十进制数字表示，并在这些数字之间加上一个点，就叫做**点分十进制记法**。

**常用的三种IP地址**

A类地址的网络号字段占1个字节，只有7位可供使用，可指派网络号是126个。减2的原因是：第一，IP地址中全0表示“这个”，是保留地址，意思是本网络。第二，网络号01111111保留作为本地软件**环回测试**本主机的进程之间通信之用。

A类地址的主机号占3个字节，因此每个A类网络中最大主机数是224-2，减2的原因是：全0的主机号表示该IP地址是本主机所连接到的**单个网络地址**，而全1表示**所有的**，即该网络上的所有主机。

IP地址共有232个地址，A类地址共有231个地址，占整个IP地址空间的50%。

B类网络地址的网络号字段有2个字节，但前两位（1 0）已经固定。B类地址的网络数为214-1，从128.1.0.0开始，B类地址的最大主机数是216-2=65534个，减2是因为要扣除全0和全1的主机号，整个B类地址有230个，占整个IP地址的25%。

C类地址有3个字节的网络号段，最前面的3位是（1 1 0），还有21位可分配。C类地址可指派网络总数221-1个，最小从192.0.1.0开始。最大主机数是254个，整个C类地址空间共229个，占整个IP地址空间的12.5%。

IP地址的特点：

1. 每个IP地址都由网络号和主机号两部分组成，从这个意思来说IP地址是一种**分等级的地址结构**。好处：IP管理机构只分配网络号，路由器仅根据目的主机所连接的网络号来转发分组，**减小了路由表所占的存储空间和查找路由表的时间**。
2. IP地址是标志一台主机或路由器和一条链路的**接口**。当一台主机同时连到两个网络上该主机就必须具有两个相应的IP地址，称为**多归属主机**。一个路由器至少应当连接到两个网络，因此一个路由器至少应有两个不同的IP地址。
3. 一个网络是指具有相同网络号的主机的集合，**因此用转发器或网桥连接起来的若干个局域网仍为一个网络**。
4. 所有分配到网络号的网络都是**平等**的。

同一个局域网上的主机或路由器的IP地址中的网络号必须是一样的。

当两个路由器直接相连时，可以分配也可以不分配IP地址，如分配了IP地址，这样的特殊网络就叫**无编号网络**或**无名网络**。

1. IP地址与硬件地址

**物理地址是数据链路层和物理层的地址**，**IP地址是网络层及其之上各层的地址**，是一种**逻辑地址**。

IP地址放在IP数据报的首部，而硬件地址放在MAC帧的首部。

**在IP层抽象的互联网上只能看到IP数据报**。

虽然在IP数据报首部有源站IP地址，但**路由器只根据目的站的IP地址的网络号进行路由选择**。

**在局域网的链路层，只能看到MAC帧**。MAC帧在不同网络传送时，其首部的源地址和目的地址也要变化。

尽管互连在一起的网络的硬件地址体系各不相同，**但IP层抽象的互联网却屏蔽了下层这些复杂细节**。

1. 地址解析协议ARP

地址解析协议ARP用来寻找已知IP地址主机的硬件地址。

ARP在主机ARP高速缓存中存放一个从IP地址到硬件地址的映射表，并且这个映射表还经常动态更新（新增或超时删除）。

每一台主机都设有一个**ARP高速缓存**，里面有**本局域网**上的各主机和路由器的IP地址到硬件地址的映射表，这些都是该主机目前知道的一些地址。

当主机A要向本地局域网的某台主机B发送IP数据报时，先在其ARP高速缓存查看有无主机B的IP地址，如有就在ARP高速缓存中查出其对应的硬件地址再把它写入MAC帧，通过局域网把MAC帧发往此硬件地址。

若差不到主机B的IP地址，主机A就自动运行ARP，然后按以下步骤查出主机B的IP地址：

1. ARP进程在本局域网上广播发送一个ARP请求分组，ARP分组的主要内容是：我的IP地址是...，我的硬件地址是...，我想知道IP地址为...的硬件地址。
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的IP地址与ARP请求分组中要查询的IP地址一致，收下这个ARP请求分组，并向主机A发送ARP响应分组。ARP响应分组的主要内容是：我的IP地址是...，我的硬件地址是...。ARP请求分组是广播，但响应是单播。
4. 主机A收到ARP响应后，在ARP高速缓存中写入主机B的IP地址到硬件地址的映射。

为减小网络通信量，当A发送其ARP请求时就把自己的IP地址到硬件地址的映射写入ARP请求分组，B收到后将其写入自己的ARP高速缓存。

ARP对保存在高速缓存中的每一个映射地址项目都设置**生存时间**，超过生存时间的项目将被删除。ARP是解决同一个局域网上的主机或路由器的IP地址和硬件地址的映射问题，若不在一个局域网上就无法解析。

从IP地址到硬件地址的解析是自动进行的，**主机的用户对这种过程是不知道的**。

由于全世界存在各种网络，它们使用不同的硬件地址。要使这些异构网络能相互通信就必须进行**非常复杂的硬件地址转换**工作，IP编址解决了此问题，因此不直接用硬件地址而额外使用抽象的IP地址。

1. IP数据报的格式

一个IP数据报由首部和数据两部分组成，首部的前一部分是**固定长度**，共20字节，是所有IP数据报必须具有的。在首部的固定部分的后面是一些**可选字段**，长度可变。

**IP数据报首部固定部分中的各字段**

1. **版本** 占4位，指IP协议的版本。通信双方所用版本必须一致。
2. **首部长度** 占4位
3. **区分服务** 占8位，在旧标准中叫做**服务类型**，一般情况不使用
4. **总长度** 指首部和数据长度之和，16位，因此数据报的最大长度为216-1为65535字节。在进行分片时，数据报首部中的总长度字段是指分片后的每一个分片的首部长度与该分片的数据长度之和。
5. **标识** 占16位，IP软件在存储器中维持一个计数器，每产生一个数据报，计数器加1，并将此值赋给标识字段。当数据报长度由于超过MTU而必须分片后，标识字段的值就被复制到所有的数据报片的标识字段，相同标识字段的值使分片后的各数据报片最后能正确地重装为原有的数据报。
6. **标志** 占3位，只有2位有意义。最低位**MF**，1表示还有分片，0表示无。中间一位记作**DF**，意思是**不能分片**，DF=0时才允许分片。
7. **片偏移** 占13位，较长的分组在分片后，某片在原分组中地相对位置。除最后一个数据报片外，每个分片地长度一定是8字节的整数倍。
8. **生存时间** 占8位 常用英文缩写TTL，表明数据报在网络中的寿命。每经过一个路由器就减去其消耗的时间，为0时丢弃。

后来功能改为“**跳数限制**”（但**名称不变**），现在TTL的单位不是秒而是**跳数**。数据报能在互联网中经过的路由器最大数值为255，若初值设1则只能在本局域网传送。

1. **协议** 占8位 指出此数据报携带的数据使用何种协议，以便使目的主机IP层知道应该将数据上交给哪个协议处理。
2. **首部检验和** 占16位，只**检验数据报的首部不包括数据部分**，主要使用反码。
3. **源地址** 占32位。
4. **目的地址** 占32位。

**IP数据报首部的可变部分**

用来支持排错、测量以及安全等措施，内容丰富。

1. IP层转发分组的流程

在路由表中，对每一条路由最主要的是以下两个信息：（**目的网络地址**，**下一跳地址**）。

IP数据报最终一定可以找到目的主机所在目的网络上的路由器（可能需要通过多次间接交付）。只有到达最后一个路由器时，才试图向目的主机进行直接交付。

虽然互联网上所有的分组转发都是基于**目的主机所在的网络**，但在大多数情况下都允许有这样的特例，即对特定的目的主机指明一个路由，这种路由叫做**特定主机路由**。

路由器还可采用**默认路由**以减小路由表所占空间和搜索路由表所用的时间。

当路由器收到一个待转发的数据报，在从路由表中得出下一跳路由器的IP地址后，不是把这个地址填入IP数据报，而是交送数据链路层的网络接口软件。网络接口软件把下一跳路由器的IP地址转换成硬件地址（必须使用ARP），并将此硬件地址放在链路层MAC帧的首部，然后根据这个硬件地址找到下一跳路由器。

归纳**分组转发算法**如下：

1. 从数据报首部提取出目的主机的IP地址D，得出目的网络地址N。
2. 若N就是与此路由器直接相连的某网络，则直接交付，否则间接交付，执行（3）。
3. 若路由表中有目的地址为D的特定主机路由，则把数据报传送给路由表中所指明的下一跳路由器，否则执行（4）。
4. 若路由表中有到达网络N的路由，则把数据报传送给路由表中所指明的下一跳路由器，否则执行（5）。
5. 若路由表有默认路由，则把数据报传送给路由表中所指明的默认路由器，否则执行（6）。
6. 报告转发分组出错。

4.3划分子网和构造超网

划分子网

1. **从两级IP地址到三级IP地址**

早期IP地址设计不合理处：

1. **空间利用率有时很低**。
2. 给每一个物理网络分配一个网络号**会使路由表变得太大**而降低网络性能。
3. **两级IP地址不灵活**。

为解决上述问题，1985年起在IP地址中增加了一个“**子网号字段**”，这种做法叫做**划分子网**，或**子网寻址**或**子网路由选择**。划分子网已成为互联网正式标准协议。

划分子网的基本思想：

1. 一个拥有许多物理网络的单位，可将所属物理网络划分为若干**子网**。划分子网纯属一个单位内部的事情，本单位以外的网络**看不见**这个网络是由多少个子网组成，因为这个单位**对外仍表现为一个网络**。
2. 从网络的主机号借用若干位作为子网号，当然主机号就相应地减少同样的位数。于是两级IP地址在**本单位内部**就变为三级IP地址：网络号，子网号和主机号。
3. 凡是从其他网络发送给本单位某台主机的IP数据报，仍然是根据IP数据报的目的网络号找到连接在本单位网络上的路由器。但此路由器收到IP数据报后，再按目的网络号和子网号找到目的子网，把IP数据报交付目的主机。
4. **子网掩码**

从IP数据报的首部**无法看出**源主机或目的主机所连接的网络是否进行了子网划分，因此需要使用**子网掩码**。

使用子网掩码的好处：不管网络有没有划分子网，只要把子网掩码和IP地址进行逐位的**与**运算，就立即得出网络地址来。

如果一个网络不划分子网，那么该网络的子网掩码就使用默认的子网掩码。

A类地址的默认子网掩码是255.0.0.0。

B类地址的默认子网掩码是255.255.0.0。

C类地址的默认子网掩码是255.255.255.0。

子网掩码是一个网络或一个子网的重要属性。路由器在和相邻路由器交换路由信息时，必须把自己所在的网络或子网的子网掩码告诉相邻路由器。

在采用固定长度子网划分时，所划分的**所有**子网掩码的子网掩码都是相同的。子网号不能为全1或全0，但随着无分类域间路由CIDR的广泛使用，现在全1和全0的子网号也能使用了。划分子网增加了灵活性，但却减少了能够连接在网络上的主机总数。

同样的IP地址和不同的子网掩码可以得出相同的网络地址。但是不同的掩码效果是不同的。

使用子网时分组的转发

使用子网划分后，路由表必须包含以下三项内容：**目的网络地址**、**子网掩码**和**下一跳地址**。

划分子网的情况下，路由器转发分组的算法如下：

1. 从收到的数据报首部提取目的IP地址D。
2. 先判断是否直接交付。对路由器直接相连的网络逐个检查，用子网掩码与D进行与运算，看结果是否和相应的网络地址匹配。若匹配直接交付，否在就是间接交付，执行3。
3. 若路由表中有目的地址为D的特定主机路由，就把数据报传送给路由表中指明的下一跳路由器，否则执行4。
4. 对路由表每一行，用其中的子网掩码和D逐位与，结果为N。如果N与该行目的地址网络适配，则把数据报传送给改行指明的下一跳路由器，否则执行5。
5. 若路由表有默认路由，则传送给默认路由器，否则报告转发分组错误。

无分类编址CIDR

1. 网络前缀

1992年互联网面临的问题：

1. B类地址已分配近一半。（2）互联网主干上路由表项目数急剧增长。（3）整个IPv4的地址空间最终将全部耗尽。

IETF研究出**无分类编址**的方法解决前两个问题。它的正式名字时**无分类域间路由选择CIDR**。

CIDR最主要的特点：

1. **消除了传统的A类、B类、C类地址以及划分子网的概念**，因而能更有效的分配IPv4的地址空间，并在新的IPv6使用之前容许互联网规模继续增长。CIDR把32位IP地址划分为前后两部分，前面是**网络前缀**，用来指明网络，后面用来指明主机。因此CIDR使IP地址从三级编址又回到了两级编址（**无分类的两级编址**）。

使用**斜线记法**，或称**CIDR记法**，即在IP地址后加上斜线，写上网络前缀所占位数。

1. CIDR把**网络前缀相同**的连续IP地址组成一个CIDR地址块，只要知道地址块中任何一个地址就能知道其中的起始地址和最大地址，以及地址块数。

为方便路由选择，CIDR使用32位的**地址掩码**。地址掩码由一串1和一串0组成，1的个数是网络前缀的长度，也可继续称为子网掩码。斜线记法中，**斜线后面的数字就是地址掩码中的1的个数**。

由于一个CIDR地址块中有很多单位，所以路由表就利用CIDR地址块来查找目的网络，这种地址的聚合常称为**路由聚合**，也称为**构成超网**。

**网络前缀越短，其地址块所包含的地址数就越多**。而在三级结构的IP地址中，划分子网是使网络前缀变长。

1. 最长前缀匹配

使用CIDR时，路由表中的每个项目由**网络前缀**和**下一跳地址**组成。但在查找路由表时**可能会得到不止一个匹配结果**。**应当从匹配结果中选择具有最长网络前缀的路由**，这叫做**最长前缀匹配**，因为网络前缀越长，地址块就越小，因而路由就越具体。最长前缀匹配又称**最长匹配**或**最佳匹配**。

1. **使用二叉线索查找路由表**

为了进行更有效的查找，最常用的是二叉线索。为了简化二叉线索的构造，可以先找出对应于每一个IP地址的唯一前缀。

二叉线索只是提供了一种可以快速在路由表中找到匹配的叶节点的机制，但这是否和网络前缀匹配，还要和子网掩码进行一次逻辑与的运算。

为提高二叉线索的查找速度，广泛使用了各种压缩技术。

4.4网际控制报文协议ICMP

为了更有效地转发IP数据报和提高交付成功的机会，在网际层使用了网际控制报文协议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，不是高层协议，作为IP层数据报的数据，加上数据报的首部，组成IP数据报。

ICMP报文的种类有两种，即**ICMP差错报告报文**和**ICMP询问报文**。

ICMP的前四个字节是统一的格式，共有三个字段：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，最后面是数据字段，长度取决于ICMP的类型。

ICMP差错控制报文共有四种，即：

1. **终点不可达** 当路由器或主机不能交付数据时就向源点发送此报文
2. **时间超过** 当路由器收到生存时间为0的数据报时，除丢弃该数据报外还要向源点发送该报文。当终点在预先规定时间内不能收到一个数据报的全部数据报片时，就把已收到的数据报片都丢弃，并向源点发送此报文。
3. **参数问题** 当路由器或目的主机收到的数据报的首部中有的字段不正确时，就丢弃该数据报，并向源点发送此报文。
4. **改变路由**（**重定向**） 路由器把改变路由报文发送给主机，让主机知道下次应将数据报发送给另外的路由器。

所有的ICMP差错报文报告中的数据字段都具有相同格式，把收到的需要进行差错报告的IP数据报的首部和数据字段的前8个字节提取出来，作为ICMP报文的数据字段，再加上ICMP差错报告报文的前8个字节，就构成了ICMP差错报告报文。

下面是不应发送ICMP差错报告报文的情况：

1. 对ICMP差错报告报文，不再发送
2. 对第一个分片的数据报片的所有后续数据报片，都不发送
3. 对具有多播地址的数据报，都不发送
4. 对具有特殊地址的数据报，不发送

常用的ICMP询问报文有两种：

1. **回送请求和回答** ICMP回送请求报文是由主机或路由器向一个特定的目的主机发出的询问。收到此报文的目的主机必须给源主机或路由器发送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其有关状态。
2. **时间戳请求和回答** ICMP时间戳请求报文是请某台主机或路由器回答当前的日期和时间。用于时钟同步和时间测量。

ICMP的一个重要应用就是分组网间探测**PING**（Packet Internet Groper），用来测试两台主机之间的连通性。PING使用了ICMP回送请求和回送回答报文。PING是应用层直接使用网络层ICMP的一个例子，没有通过运输层的TCP或UDP。

Windows操作系统的用户在命令提示行输入ping+主机名或ip地址就可测试连通性。

例PING一个服务器时，PC一连发出4个ICMP回送请求报文，如果服务器正常工作而且响应这个ICMP回送请求报文（有的主机为防止恶意攻击而不理睬外界发送的这种报文），那么它就发回ICMP回送回答报文。由于往返的ICMP报文上都是时间戳，因此很容易得出往返时间。最后显示出的是统计结果：发送到哪个机器（IP地址），发送的、收到的和丢失的分组数（但不给出分组丢失原因）以及往返时间的最小值、最大值和平均值。

另一个非常有用的应用是traceroute（这是UNIX系统中的名字），它用来跟踪一个分组从源点到终点的路径，在Windows中这个命令是tracert。

原理：Traceroute从源主机向目的主机发送一连串的IP数据报，数据报中封装的是无法交付的UDP用户数举报（因为使用了非法端口号）。第一个数据报的TTL设为1，当其到达路径上第一个路由器时，路由器先收下它，接着把TTL减1，由于TTL=0，路由器丢弃数据报并向源主机发送一个**ICMP时间超过**差错报告报文。

接着源主机发送第二个数据报，并把TTL设为2...以此类推直到最后一个数据报刚到达目的主机时，数据报的TTL=1，主机不转发数据报也不把TTL减1，但因为IP数据报中封装的时无法交付的运输层的UDP数据报，因此目的主机要向源主机发送**ICMP终点不可达**差错报告报文。这样源主机就知道了路由信息——到达目的主机所经过的路由器的IP地址以及到达其中每个路由器的往返时间。每一行有3个时间出现是因为对应于每个TTL，源主机要发送三次同样的IP数据报。

4.5 互联网的路由选择协议

理想的路由算法特点：

1. **算法必须是正确的和完整的**。正确指沿着各路由表所指引的路由，分组一定能最终到达目的网络和目的主机。
2. **算法在计算上应简单**，不应使网络通信量增加太多额外开销。
3. **算法能适应通信量和网络拓扑的变化**，要有**自适应性**。有时称这种自适应性为稳健性。
4. **算法应具有稳定性**，网络通信量和拓扑稳定时不应使路由不停变化。
5. **算法应公平**。
6. **算法应是最佳的**，找到最好的路由使分组平均时延最小而网络吞吐量最大。

若从路由算法能否随网络的通信量或拓扑自适应地进行调整变化来划分，则只有两大类，即**静态路由选择策略**（非自适应路由选择）和**动态路由选择策略**（自适应路由选择）。

分层次的路由选择协议

互联网采用的路由选择协议主要是自适应的（动态的）、分布式路由选择协议，主要因为：

1. 互联网规模非常大，如果让所有路由表知道所有网络应怎样到达，开销太大。
2. 许多单位不愿意外界了解自己网络的布局细节和采用的路由选择协议但又希望连接互联网。

为此可把互联网划分为许多较小的**自治系统**AS。AS是在单一技术管理下的一组路由器，而这些路由器使用一种自治系统内部的路由选择协议和共同的度量。一个AS对其他AS表现出的是一个**单一的和一致的路由选择策略**。

目前的互联网中，一个大的ISP就是一个自治系统。这样，互联网就把路由选择协议划分为两大类：

1. **内部网关协议IGP** 包括RIP和OSPF协议
2. **外部网关协议EGP** 目前使用最多的外部网关协议是BGP

自治系统之间的路由选择叫**域间路由选择**，而在自治系统内部的路由选择叫**域内路由选择**。

内部网关协议RIP

RIP是内部网关协议IGP中最先得到广泛使用的协议，中文名为路由信息协议，但很少使用。RIP是一种分布式的**基于距离向量**的路由选择协议，最大优点是简单。

RIP协议要求网络中每一个路由器都要维护从它自己到其他每个目的网络的距离记录（这**一组距离**即**距离向量**）。距离定义如下：

从路由器到直接相连的网络距离为1，从路由器到非直接相连的网络距离为经过的路由器数加1。RIP的距离也成为跳数，RIP允许一条路径最多只能包含15个路由器，因此距离等于16时相当于不可达，因此**RIP只适用于小型互联网**。

RIP协议的特点：

1. **仅和相邻路由器交换信息**。
2. 路由器交换的信息是**当前本路由器知道的全部信息**，**即自己现在的路由表**。
3. **按固定时间间隔交换路由信息**。

路由表中最主要的信息就是：到某个网络的距离（最短距离），以及应经过的下一跳地址。路由表更新的原则是找出到每个目的网络的**最短距离**，这种更新算法又称**距离向量算法**。

距离向量算法，基础是Bellman-Ford算法或Ford-Fulkerson算法。

步骤：

1. 对地址X的相邻路由器发来的RIP报文，先修改此报文中的所有项目，把下一跳地址改为X，并把所有距离加1。每个项目都有三个关键数据，目的网络N，距离d，下一跳地址X。
2. 对修改后的项目，若路由表没有目的网络N则添加，若有且下一跳地址是X则替换，若有且下一跳地址不是X，如果d小于原来的距离则更新，否则什么也不做。
3. 3分组内没收到相邻路由表的更新路由表，则记为不可达，即设置距离16。
4. 返回。

RIP协议使得从**每一个路由到每一个目的网络的路由都是最短的**。

RIP2支持变长子网掩码和无分类域间路由选择CIDR，还提供简单鉴别支持多播。RIP协议使用UDP进行传送（UDP端口520）。

RIP报文由首部和路由组成。

RIP2报文中的路由部分由若干路由信息组成，每个路由信息需要用20字节。**地址族标识符**（又称**地址类别**）字段用来标志所使用的地址协议。路由标记填入**自治系统号ASN**，这是考虑使RIP有可能收到本自治系统外的路由选择信息。再后面指出某个**网络地址**、该网络的**子网掩码**、**下一跳路由器地址**以及**到此网络距离**。一个RIP报文最多包括25个路由，因而RIP报文最大长度504字节。

RIP问题：当网络故障时，要经过比较长的时间才能将信息传送到所有路由器。这一特点叫：**好消息传播得快，坏消息传播得慢**。总之RIP协议最大的优点就是**实现简单**，**开销较小**，缺点也较多，限制了网络规模，对于规模较大的网络应使用OSPF协议。

内部网关协议OSPF

这个协议的名字是**开放最短路径优先OSPF**，为克服RIP缺点开发。**开放**表明OSPF协议不是受一家厂商控制，而是公开发表。**最短路径优先**是因为使用了**Dijkstra提出的最短路径算法SPF**。

主要特征：使用分布式的链路状态协议，而不是RIP那样的距离向量协议。

1. 向本自治系统的**所有路由器**发送信息，用的是**洪泛法**。
2. 发送的信息就是与本路由器**相邻的所有路由器的链路状态**，这只是路由器知道的部分信息。链路状态就是说明本路由器都和哪些路由器相邻，以及该链路的**度量**（费用、距离、时延、带宽等）。
3. 只有当链路状态**发生变化**时，路由器才向所有路由器用**洪泛法**发送此信息。

由于各路由器之间频繁交换链路信息，因此所有路由器最终都能建立一个**链路状态数据库**，实际上就是**全网的拓扑结构图**。RIP协议的每个路由器虽然知道到所有网络的距离以及下一跳路由器，但不知道全网的拓扑结构。OSPF的链路状态数据库能较快更新，使各路由器及时更新路由表，OSPF的更新过程收敛得快是其重优点。

为使OSPF能够用于大型网络，OSPF将一个自治系统划分为若干**区域**。OSPF使用**层次结构的区域划分**，在上层的区域叫**主干区域**，作用使连通其他下层区域，从其他区域来的信息都由**区域边界路由器**进行概括。在主干区域内的路由器叫**主干路由器**，一个主干路由器可以同时是多个区域边界路由器。在主干区域还有一个路由器专门和本自治系统之外的其他自治系统交换路由信息，叫做**自治系统边界路由器**。

OSPF不用UDP而是直接用IP数据报传送（IP数据报首部协议字段值89）。OSPF分组使用24字节的固定首部长度，分组的数据部分可以是五种类型分组中的一种。

OSPF首部各字段的意义：

1. **版本** 当前版本号为2
2. **类型** 五种类型分组中的一种
3. **分组长度** 包括OSPF首部在内的分组长度，以字节为单位
4. **路由器标识符** 标志发送该分组的路由器的接口的IP地址
5. **区域标识符** 分组属于的区域
6. **检验和** 用来检验分组中的差错
7. **鉴别类型** 0（不用）和1（口令）
8. **鉴别** 鉴别类型为0时填入0，为1时填入8个自负的口令

OSPF的其他特点：

1. 对于不同类型的业务可计算不同的路由，链路代价为1-65535
2. 如果到同一个目的网络有多条相同路径，那么可以将通信量分配给这几条路径，这叫做**多路径间的负载平衡**。
3. 所有在OSPF路由器之间交换的分组都具有鉴别的功能。
4. OSPF支持可变长度的子网划分和CIDR。
5. OSPF让每个链路状态都带上一个32位的序号，序号越大状态越新。

OSPF的五种分组类型:

1. 类型1，**问候分组**。
2. 类型2，**数据库描述分组**。
3. 类型3，**链路状态请求分组**。
4. 类型4，**链路状态更新分组**。
5. 类型5，**链路状态确认分组**。

外部网关协议BGP

1989年公布了新的外部网关协议——边界网关协议BGP。

内部网关协议主要设法使数据报在一个AS中尽可能有效从源站传送到目的站，在一根AS内也不需要考虑其他方面策略。然而BGP使用环境不同，主要因为：

1. 互联网规模太大，使得自治系统AS之间路由选择非常困难。
2. 自治系统AS之间的路由选择必须考虑有关策略。

由于上述情况，边界网关协议BGP只能力求寻找一条能够到达目的网络且**比较好**的路由，而**并非寻找一条最佳路由**。BGP采用了**路径向量路由选择协议**。

配置BGP时，每个AS管理员要选择至少一个路由器作为该AS的BGP发言人，一般来说，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。

一个BGP发言人与其他AS的BGP发言人要交换路由信息就要先建立TCP连接（端口号179），然后在此连接上交换BGP报文以建立**BGP会话**，利用BGP会话交换路由信息。使用TCP连接能提供可靠服务，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**临站**或**对等站**。

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自治系统，当BGP发言人相互交换了网络可达性信息后，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自治系统的较好录由。

BGP协议交换路由信息的结点数量级时**自治系统个数**的量级，要比自治系统的**网络数**少很多。

BGP支持无分类域间路由选择CIDR，因此BGP的路由表也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自治系统序列。

BGP的四种报文：

1. OPEN打开报文，用来与相邻的另一个BGP发言人建立关系，通信初始化。
2. UPDATE 更新报文，用来通告某一路由的信息以及列出要撤销的多条路由。
3. KEEPALIVE保活报文，用来周期性证实临站的连通性。
4. NOTIFICATION通知报文，用来发送检测到的差错。

四种类型的BGP报文具有同样的通用首部，长度19字节。通用首部分为三各字段。标记字段16字节，用来鉴别收到的BGP报文，当不使用鉴别时，标记字段置为全1。长度字段指出包括通用首部在内的整个BGP报文以字节为单位的长度，最小值19，最大值4096。类型字段的值1-4。

OPEN报文有6个字段，即版本（1字节，当前值4）、本自治系统（2字节）、保持时间（2字节）、BGP标识符（2字节）、可选参数长度和可选参数。

UPDATE报文共有5个字段，即不可行路由长度（2字节）、撤销的路由、路径属性总长度（2字节）、路径属性和网络层可达性NLRI。

KEEPALIVE报文只有BGP的19字节长的通用首部。

NOTIFICATION报文有3个，即差错代码（1字节）、差错子代码（1字节）和差错数据。

路由器的构成

路由器是一种具有多个输入端口和多个输出端口的专用计算机，其任务是转发分组。路由器的转发分组是网络层的主要工作。路由器的结构可划分为两部分：**路由选择**部分和**转发分组**部分。

路由选择部分也叫**控制部分**，其核心构件是路由选择处理机。路由选择处理机的任务是根据所选定的路由选择协议构造出路由表，同时经常或定期和相邻路由器交换路由信息而不断地更新和维护路由表。

分组转发部分由三部分组成：**交换结构**、一组**输入端口**和一组**输出端口**。这里的端口指硬件接口。

交换结构又称为**交换组织**，它的作用是根据转发表对分组进行处理，将某个输入端口进入的分组从一个合适的输出端口转发出去。可将交换结构看成在路由器中的网络。

**转发**就是路由器根据转发表把收到的IP数据报从路由器的合适端口转发出去，仅仅涉及一个路由器，但**路由选择**涉及到很多路由器，路由表则是多个路由器协同工作的结果。路由表一般仅包含从目的网络到下一跳的映射，而转发表是从路由表得出的。路由表总是用软件实现的，但转发表则甚至可用特殊的硬件实现。**当讨论路由选择原理时，往往不去区分转发表和路由表的分别**。

若分组处理的速率赶不上分组进入的速率，则队列的存储空间最终必定减少到零，这就使后面再进入队列的分组由于没有存储空间而只能被丢弃。以前所提的分组丢失就是发生在路由器的输入或输出队列产生溢出的时候。设备或线路故障也可能使分组丢失。

交换结构是路由器的关键构件，正是这个交换结构把分组从一个输入端口转移到某个合适的输出端口。实现这样的交换有多种方法。

1. 最早使用的路由器就是利用普通计算机，用计算机的CPU作为路由器的路由选择处理机。许多现代的路由器也通过存储器进行交换，与早期路由器的区别是，目的地址的查找和分组在存储器中的缓存都是在输入端口中进行的。
2. 通过总线进行交换，采用这种方式时，数据报从输入端口通过共享总线直接传送到合适的输出端口，而不需要路由选择处理机的干预。
3. 通过纵横交换结构进行交互，这种交换结构常称为**互连网络**，它有2N条总线，可以使N个输入端口和N个输出端口相连接，这取决于相应的交叉结点是使水平总线和垂直总线接通还是断开。输入端口将收到的分组发送到与输入端口相连的水平总线，若垂直总线空闲就将其与水平总线连通并转发到输出端口，否则在输入端口排队。

4.6 IPV6

解决IP地址耗尽的根本措施就是采用具有更大地址空间的新版本的IP，即IPV6。

IPv6仍支持无连接的传送，但将协议数据单元PDU称为分组，而不是IPv4的数据报。

IPv6的主要变化：

1. **更大的地址空间**，从IPv4的32位增大到128位。
2. **扩展的地址层次结构**。
3. **灵活的首部格式**。IPv6数据报的首部和IPv4不兼容。
4. **改进的选项**。IPv6允许数据报包含有选项的控制信息，但IPv6的首部长度是固定的。
5. **允许协议继续扩充**。
6. 支持即插即用（自动配置）。不需要使用DHCP。
7. **支持资源的预分配**。
8. IPv6的首部改为**8字节对齐**（即首部长度必须是8字节的整数倍）。

IPv6数据报由两大部分组成，即**基本首部**和后面的**有效载荷**（净负荷），有效载荷允许有零个或多个**扩展首部**，再后面是数据部分。

与IPv4相比，IPv6对首部的某些字段进行了如下更改：

1. 取消了首部长度字段，因为它的首部长度固定为40字节
2. 取消了服务类型字段
3. 取消了总长度字段
4. 取消了标志、标识和片偏移字段
5. 把TTL字段改为限制条数字段，功能不变
6. 取消了协议字段
7. 取消了检验和字段，加快了路由器处理数据报的速度
8. 取消了选项字段

IPv6首部中各字段的作用：

1. **版本** 4位，指明协议版本，值为6
2. **通信量类** 8位，为了区分不同IPv6数据报的类别或优先级
3. **流标号** 20位，IPv6提出**流**的抽象概念，所谓**流就是互联网上从特定源点到特定终点（单播或多播）的一系列数据报（如实时音频或视频传输），而在这个流所经过的路径上的路由器都保证指明的服务质量**。属于同一个流的数据报都具有同样的流标号。
4. **有效载荷长度** 16位，指明IPv6数据报除基本首部以外的字节数，最大值64KB。
5. **下一个首部** 8位，相当于IPv4的协议字段或可选字段。当没有扩展首部时，下一个首部字段的作用和IPv4的协议字段作用一样，6代表交付运输层TCP，17代表交付UDP。

出现扩展首部时，下一个首部字段的值标识后面第一个扩展首部的类型。

1. **跳数限制** 8位，防止数据报在网络无限期存在，最大跳255。
2. **源地址** 128位，数据报发送端的IP地址。
3. **目的地址** 128位，数据报接收端的IP地址。

IPv6把原来IPv4首部中选项的功能都放在扩展首部，并把扩展首部留给路径两端的源点和终点的主机处理，而数据报途径中经过的**路由器都不处理这些扩展首部**（除了逐跳选项扩展首部），大大**提高了路由器的处理效率**。

RFC2460定义了以下6种扩展首部：（1）逐跳选项（2）路由选择（3）分片（4）鉴别（5）封装安全有效载荷（6）目的站选项

IPv6的地址是以下三种之一：

1. **单播** 传统的点对点通信
2. **多播** 一点对多点的通信，数据报发送到一组计算机中的而每一个，IPv6没有采用广播的术语，而是将广播看作多播的一个特例
3. **任播** 终点是一组计算机，但数据报只交付其中一个，通常是距离最近的一个。

IPv6把实现IPv6的主机或路由器均称为结点。

向IPv6过渡只能采取逐步演进的办法，同时还必须使新安装的IPv6系统能够向后兼容。这就是说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。

向IPv6过渡的策略：

1. **双协议栈**

是指在完全过渡到IPv6之前使一部分主机或路由器装有双协议栈：一个IPv4和一个IPv6。它使用域名系统DNS来查询目的主机使用的IP地址类型。

1. **隧道技术**

这种方法的要点就是在IPv6数据报要进入IPv4网络时，把IPv6数据报封装称IPv4数据报，要使双协议栈的主机知道IPv4数据报里封装的是一个IPv6数据报就必须把IPv4首部的协议字段值设为41。

和IPv4一样，IPv6也不保证数据报的可靠交付，它也需要使用ICMP来反馈一些差错信息，新的版本称为ICMPv6，它比ICMPv4要复杂许多，地址解析协议ARP和网际组管理协议IGMP的功能都已被合并到ICMPv6中。

4.7 IP多播

与单播相比，在一对多的通信中，多播可大大节约网络资源。

在互联网范围的多播要靠路由器实现，这些路由器必须增加一些能够识别多播数据报的软件，能够运行多播协议的路由器称为**多播路由器**。

在互联网上进行多播就叫IP多播。IP多播所传送的分组需要使用多播IP地址。

在多播数据报的目的地址写入的是多播组的标识符，然后设法让加入这个多播组的主机IP地址与多播组的标识符关联起来。其实多播组的标识符就是IP地址中的D类地址。D类地址范围是224.0.0.0到239.255.255.255，用每一个D类地址标志一个多播组。多播数据报和一般IP数据报的区别是它使用D类IP地址作为目的地址，并且首部中的协议字段值为2，表明使用网际管理组协议IGMP。

**多播地址只能用于目的地址，而不能用于源地址**。对多播报文不产生ICMP差错报文，因此若在PING命令后键入多播地址将永远不会收到响应。

IP多播分为两种：（1）在本局域网上进行硬件多播（2）在互联网范围进行多播。前一种虽然简单但很重要，因为现在大部分主机都是通过局域网接入互联网。在互联网上进行多播的最后阶段还是要把多播数据报在局域网上用硬件多播交付多播组的所有成员。

IP多播需要两种协议

1. 网际组管理协议IGMP

工作可分为两阶段：

第一阶段：当某台主机加入新的多播组时，该主机向多播组的多播地址发送一个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报文后，还要利用多播路由选择协议把这种组成员关系转发给互联网上的其他多播路由器。

第二阶段：组成员关系是动态的，本地多播路由器要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组的成员。只要有一台主机对某个组响应，那么多播路由器就认为这个组是活跃的。但一个组在多次探询后仍然没有一台主机响应，多播路由器就认为本网络上的主机都离开了这个组，也就不再转发组关系至其它多播路由器。

1. 多播路由选择协议

在多播过程中一个多播组中的成员是动态变化的。多播路由实际上就是要找出以源主机为根节点的**多播转发树**。

转发多播数据报时用到了三种方法：（1）**洪泛与剪除**。适合于小的多播组，而所有组成员接入的局域网也是相邻的。为了避免兜圈子，采用了**反向路径广播RPB。**（2）**隧道技术**，适用于多播组的位置在地理上很分散的情况。（3）基于核心的发现技术，这种方法对于多播组的大小在较大范围内变化时都适合。

4.8 虚拟专用网VPN和网络地址转换NAT

假定一个机构内部的计算机通信也采用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自行分配其IP地址，这就是说让这些计算机使用仅在本机构内有效的IP地址（本地地址）而不需要向互联网的管理机构申请全球唯一的IP地址（全球地址），这样可大大节约IP地址资源。

**在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**。

2013年4月，RFC6890给出了所有特殊用途的IPv4地址：

（1）10.0.0.0到10.255.255.255 即10.0.0.0/8

（2）172.16.0.0到172.31.255.255 即172.16.0.0/12

（3）192.168.0.0到192.168.255.255 即192.169.0.0/16

采用这样的专用IP地址的互连网洛称为**专用互联网**或**本地互联网**，或**专用网**。专用IP地址也叫做**可重用地址**。

机构利用公用的互连网作为本机构各专用网之间的通信载体，这样的专用网叫做**虚拟专用网VPN**。之所以叫专用网是因为这种网络时为本机构的主机用于机构内部的通信，而不适用于和网络外费本机构的主机通信。如果专用网不同网点间通信必须经过公用互联网，但又有保密要求，那么**所有通过互联网传送的数据都必须加密**。虚拟表示好像时，但实际上不是，因为现在并没有真正使用通信专线，而VPN只是在**效果上**和真正专用网一样。

VPN类型：内联网、外联网（内联网和外联网都基于TCP/IP）、远程接入

网络地址转换NAT

这种方法需要在专用网连接到互联网上的路由器安装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。这样所有使用本地地址的主机在和外界通信时，都要在NAT路由器上将其本地地址转换成全球IP地址，才能和互联网通信。

通过NAT路由器的通信必须由专用网内的主机发起，专用网内的主机不能充当服务器，因为互联网上的客户无法请求专用网内的服务器提供服务。为更加有效利用NAT路由器上的IP地址，现在常用的NAT转换表把运输层的端口号也加上，这样就可使多个拥有本地地址的主机，公用一个NAT路由器上的IP地址，因而可以同时和互联网上的不同主机通信。

采用了端口号的NAT也叫做**网络地址与端口号转换NAPT**，而不使用端口号的NAT就叫做NAT。普通路由器转发IP数据报时，源地址或目的地址都不改变，但NAT路由器在转发IP数据报时一定要更换其IP地址，且还要查看和转换运输层端口号，不仅仅工作在网络层。

4.9 多协议标记交换MPLS

MPLS利用面向连接技术，使每个分组携带一个叫做**标记**的小整数，当分组到达交换机时，交换机读取分组的标记，并用标记值来检索分组转发表，这样就比查找路由表来转发分组要快得多。

特点：（1）支持面向连接的服务质量（2）支持流量过程，平衡网络负载（3）有效支持虚拟专用网VPN

MPLS的一个重要特点就是在MPLS域的入口处，给每一个IP数据报打上固定长度的标记，**然后对打上标记的IP数据报用硬件转发**，大大加快了IP数据报的转发过程。采用硬件技术对打上标记的IP数据报进行转发就成为标记交换。交换也表示在转发时不再上升到第三层查找转发表，而是**根据标记在第二层（链路层）用硬件转发**。MPLS可使用多种链路层协议：PPP、以太网、ATM以及帧中继等。

MPLS域指该域中有许多彼此相邻的路由器，并且所有路由器都是支持MPLS技术的**标记交换路由器LSR**。LSR同时具有标记交换和路由选择两种功能。

4.10 本章的重要概念

TCP/IP体系中的网络层向上只提供简单灵活、无连接的、尽最大努力交付的数据报服务。网络层不提供服务质量的承诺，不保证分组交换的时限，所传送的分组可能出错、丢失、重复或失序。进程之间通信的可靠性由运输层负责。

IP网是虚拟的，因为从网络层上看，IP网就是一个统一的、抽象的网络（实际上是异构的）。IP层抽象的互联网屏蔽了下层网络复杂的细节使我们能够使用统一的、抽象的IP地址处理主机之间的通信问题。

在互联网上的交互有两种：在本网络的直接交付（不经过路由器）和到其他网络的间接交付（至少经过一个路由器但最后一次一定是直接交付）。

一个IP地址在整个互联网范围内是唯一的，分类的IP地址包括A类、B类、C类地址（单播地址）、以及D类地址（多播地址），E类地址未使用。

分类的IP地址由网络号字段（指明网络）和主机号字段（指明主机）组成。网络号字段最前面的类别位指明IP地址的类别。

IP地址是一种分等级的地址结构。IP地址管理机构在分配IP地址时只分配网络号，主机号则由得到该网络号的单位自行分配。路由器仅根据目的主机所连接的网络号来转发分组。

IP地址标志一台主机（或路由器）和一条链路的接口。多归属主机同时连接到两个或更多的网络上。这样的主机同时具有两个或更多的IP地址，其网络号必须是不同的。由于一个路由器至少应连接到两个网络，因此一个路由器至少应当有两个不同的IP地址。

按照互联网的观点，用转发器或网桥连接起来的若干局域网仍为一个网络，所有分配到网络号的网络（无论范围大小）都是平等的。

物理地址（硬件地址）是数据链路层和物理层使用的地址，而IP地址是网络层和以上各层使用的地址，是一种逻辑地址（用软件实现），在数据链路层看不见数据报的IP地址。

IP数据报分为首部和数据两部分，首部的前一部分是固定长度20字节，是所有IP数据报必须具有的（源地址、目的地址、总长度等重要字段都在固定首部中）。一些长度可变的可选字段放在固定首部的后面。

IP首部中的生存时间字段给出了IP数据报在互联网中所能经过的最大路由数，可防止IP数据报在互联网中无限制兜圈。

地址解析协议ARP把IP地址解析为硬件地址，它解决同一个局域网上的主机或路由器的IP地址和硬件地址的映射问题。ARP的高速缓存可以大大减少网络上的通信量。

互联网中我们无法根据硬件地址寻找在某个网络上的某台主机，因此从IP地址到硬件地址的解析是非常必要的。

无分类域间路由选择CIDR是解决目前IP地址紧缺的一个好办法。CIDR记法把IP地址后面加上斜线“/”，然后写上前缀所占的位数。前缀用来指明网络，前缀后面的部分是后缀，用来指明主机。CIDR把前缀都相同的连续IP地址组成一个CIDR地址块，IP地址的分配都以CIDR地址块为单位。

CIDR的32为地址掩码由一串1和一串0组成，1的个数是前缀的长度。只要把IP地址和地址掩码逐位进行与运算，就很容易得出网络地址。A类地址的默认地址掩码是255.0.0，B类地址的默认地址掩码是255.255.0.0，C类地址的默认地址掩码是255.255.255.0。

路由聚合（把许多前缀相同的地址用一个来代替）有利于减少路由表中的项目，减少路由器之间的路由选择信息的交换，从而提高整个网络的性能。

转发和路由选择有区别，转发是单个路由器的动作，路由选择是许多路由器共同协同的过程，这些路由器相互交换信息，目的是生成路由表，再从路由表导出转发表。若采用自适应的路由选择算法，则网络拓扑变化时，路由表和转发表都能够自动更新。许多情况下，可不考虑路由表和转发表的区别而都使用路由表这一名词。

自治系统AS就是在单一的技术管理下的一组路由器，一个自治系统对其他自治系统表现出的是一个单一的和一致的路由选择策略。

路由选择协议有两大类：内部网关协议（或自治系统内部的路由选择协议），如RIP和OSPF；外部网关协议（自治系统之间的路由选择协议），如BGP-4。

RIP是分布式的基于距离向量的路由选择协议，只适用于小型互联网。RIP按固定的时间间隔与相邻路由器交换信息。交换的信息是自己当前的路由表，即到达本自治系统中所有网络的最短距离，以及到每个网络应经过的下一跳路由器。

OSPF是分布式的链路状态协议，适用于大型互联网。OSPF只在链路状态发生变化时，才向本自治系统中的所有路由器用洪泛法发送与本路由器相邻的所有路由器的链路状态信息。链路状态指明本路由器都和哪些路由器相邻，以及该链路的度量。度量表示费用、时延、带宽等，可通称为代价。所有的路由器最终都能建立一个全网的拓扑结构图。

BGP-4是不同AS的路由器之间交换路由信息的协议，是一种路径向量路由选择协议。BGP力求寻找一条能够到达目的网络（可达）且比较好的路由（不兜圈子），而并非寻找一条最佳路由。

网际控制报文协议ICMP是IP层的协议。ICMP报文作为IP数据报的数据，加上首部后组成IP数据报发送出去。使用ICMP并非为了实现可靠传输，ICMP允许主机或路由器报告差错情况和提供有关异常的报告。ICMP报文的种类有两种，即ICMP差错报告报文和ICMP询问报文。

ICMP的一个重要应用就是分组网间探测PING，用来测试两台主机之间的连通性。PING使用了ICMP回送请求和回送回答报文。

要解决IP地址耗尽的问题，最根本的办法就是采用具有更大地址空间的新版本IP协议即IPv6。

IPv6的主要变化是：（1）更大的地址空间（2）灵活的首部格式（3）改进的选项（4）支持即插即用（5）支持资源的预分配（6）IPv6的首部改为8字节对齐。

IPv6数据报在基本首部的后面允许有零个或多个扩展首部，再后面是数据。所有的扩展首部和数据合起来叫做数据报的有效载荷或净负荷。

IPv6数据报的目的地址可以是以下三种基本类型之一：单播、多播和任播。

IPv6的地址使用冒号16进制记法。

向IPv6过渡只能采用逐步演进的办法，必须使新安装的IPv6系统能够向后兼容。向IPv6过渡可以使用双协议栈或隧道技术。

与单播相比，在一对多的通信中，IP多播可大大节约网络资源。IP多播使用D类IP地址，IP多播需要使用网际组管理协议IGMP和多播路由选择协议。

虚拟专用网VPN利用公用的互联网作为本机构各专用网之间的通信载体。VPN内部使用互联网的专用地址，一个VPN至少要有一个路由器具有合法的全球IP地址，这样才能和本系统另一个VPN通过互联网进行通信。所有通过互联网传说的数据都必须加密。

使用网络地址转换NAT技术，可以在专用网络内部使用专用IP地址，而仅在连接到互联网的路由器使用全球IP地址，这样就大大节约了IP地址资源。

MPLS的特点：（1）支持面向连接的服务质量（2）支持流量工程，平衡网络负载（3）有效地支持虚拟专用网VPN

MPLS在入口结点给每一个IP数据报打上固定长度的标记，然后根据标记在第二层（链路层）用硬件进行转发（在标记交换路由器中进行标记对换），因而转发速率大大加快。