[面试·网络] TCP/IP（二）：IP协议

[面试·网络] TCP/IP（二）：IP协议

IP协议处于OSI参考模型的第三层——网络层，网络层的主要作用是实现终端节点间的通信。IP协议是网络层的一个重要协议，网络层中还有ARP(获取MAC地址)和ICMP协议(数据发送异常通知)

数据链路层的作用在于实现同一种数据链路下的包传递，而网络层则可以实现跨越不同数据链路的包传递。比如主机A通过Wi-Fi连接到路由器B，路由器B通过以太网连接到路由器C，而路由器C又通过Wi-Fi与主机D保持连接。这时主机A向D发送的数据包就依赖于网络层进行传输。

这篇文章主要介绍IP协议的基本知识和IP首部，IP协议可以分为三大作用模块：IP寻址、路由和IP分包。

IP地址

IP地址是一种在网络层用于识别通信对端信息的地址。它有别于数据链路层中的MAC地址，后者用于标识同一链路下不同的计算机。

举一个形象的例子，我要从镇江的家里去沈阳的东北大学，通信两端的地址分别是家和学校，他们相当于IP地址。然而没有交通工具可以让我从家直接去学校，所以我先要打车去火车站，然后坐高铁到沈阳站，再转公交去学校。这三次中转分别属于三种交通方式(数据链路)，每一次中转都有起点和终点，他们就相当于MAC地址。每次中转可以称为一跳(Hop)

IP地址由32位正整数表示，为了直观的表示，我们把它分成4个部分，每个部分由8位整数组成，对应十进制的范围就是0-255。

比如`172.20.1.1`可以表示为：`10101100 00010100 00000001 00000001`。转换规则很简单，就是分别把四个部分的十进制(0-255)与8位二进制数字进行转换。

从功能上看，IP地址由两部分组成：网络标识和主机标识。

网络标识用于区分不同的网段，相同段内的主机必须拥有相同的网络表示，不同段内的主机不能拥有相同的网络标识。

主机标识用于区分同一网段下不同的主机，它不能在同一网段内重复出现。

32位IP地址被分为两部分，到底前多少位是网络标识呢？一般有两种方法表示：IP地址分类、子网掩码。

IP分类

IP地址分为四个级别，分别为A类、B类、C类和D类。分类的依据是IP地址的前四位：

A类IP地址是第一位为“0”的地址。A类IP地址的前8位是网络标识，用十进制标识的话`0.0.0.0-127.0.0.0`是A类IP地址的理论范围。另外我们还可以得知，A类IP地址最多只有128个(实际上是126个，下文不赘述)，每个网段内主机上限为2的24次方，也就是16，777，214个。

B类IP地址是前两位为“10“的地址。B类IP地址的前16位是网络标识，用十进制标识的话`128.0.0.0-191.255.0.0`是B类IP地址的范围。B类IP地址的主机标记长度为16位，因此一个网段内可容纳主机地址上限为65534个。

C类IP地址是前三位为“110”的地址。C类IP地址的前24位是网络标识，用十进制标识的话`192.0.0.0-223.255.255.0`是C类IP地址的范围。C类地址的后8位是主机标识，共容纳254个主机地址。

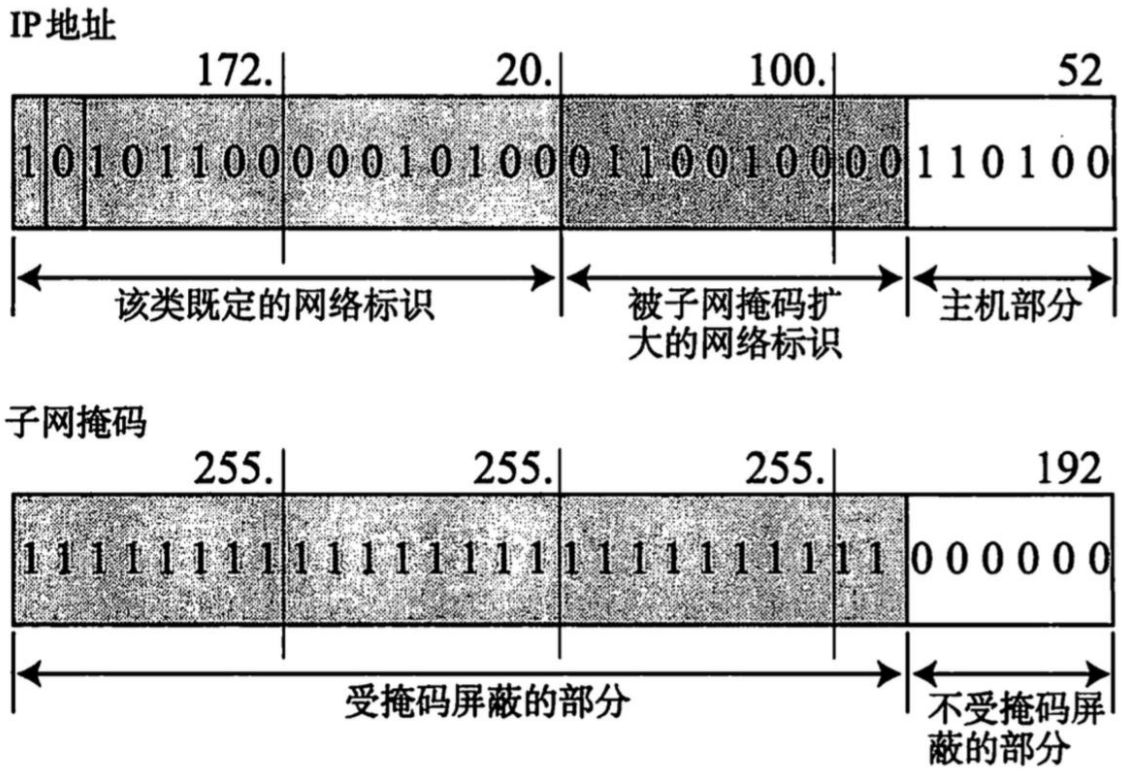
D类IP地址是前四位为“1110”的地址。D类IP地址的网络标识长32位，没有主机标识，因此常用于多播。

子网掩码

IP地址总长度32位，它能表示的主机数量有限，大约在43亿左右。而IP地址分类更是造成了极大的浪费，A、B类地址一共也就一万多个，而世界上包含主机数量超过254的网段显然不止这么点。

我们知道IP地址分类的本质是区分网络标识和主机标识，另一种更加灵活、细粒度的区分方法是使用子网掩码。

子网掩码长度也是32位，由一段连续的1和一段连续的0组成。1的长度就表示网络标识的长度。以IP地址`172.20.100.52`为例，它本来是一个B类IP地址(前16位是网络标识)，但通过子网掩码，它可以是前26为为网络标识的IP地址：

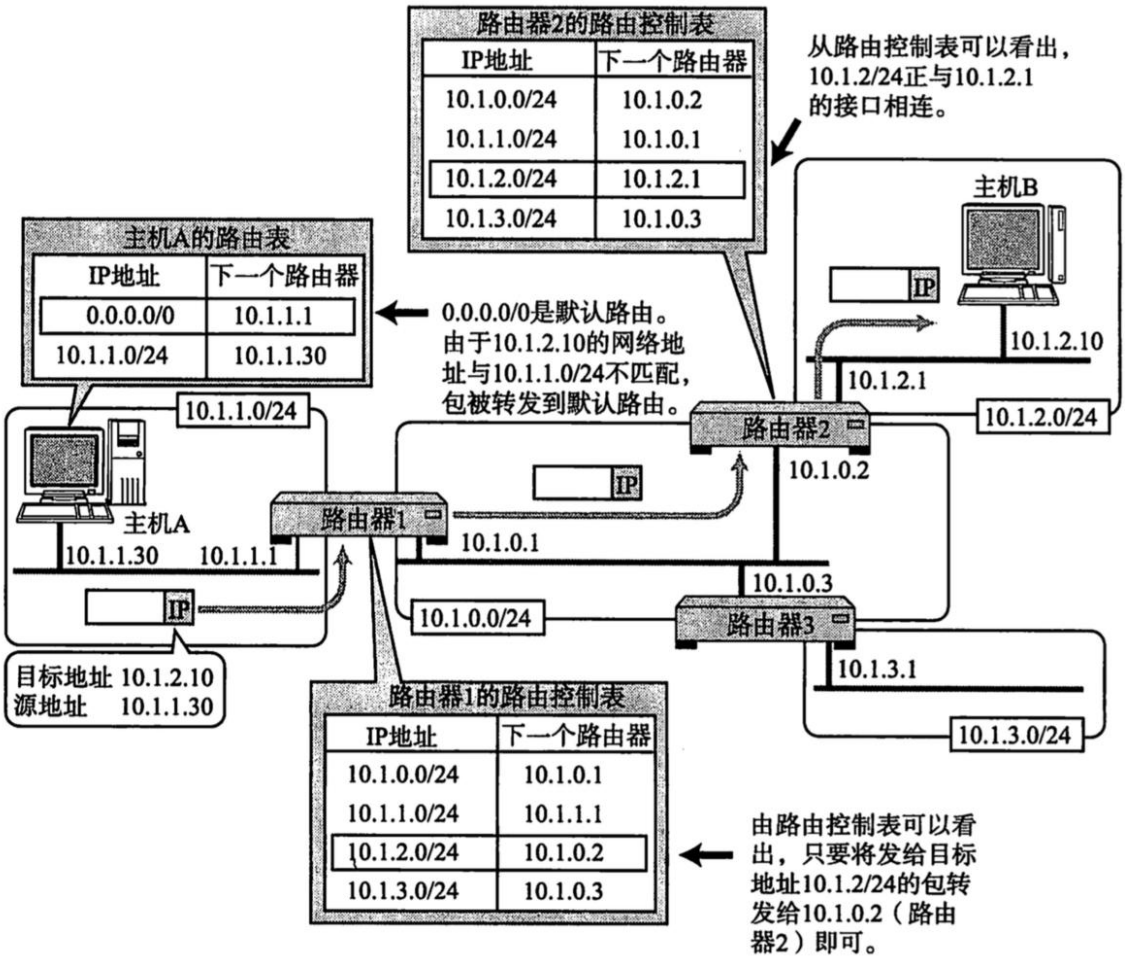


子网掩码

路由控制

路由控制(Routing)是指将分组数据发送到目标地址的功能，这个功能一般由路由器完成。(不要与家里用的小型无线路由器混为一谈)

路由器中保存着路由控制表，它在路由控制表中查找目标IP地址对应的下一个路由器地址。下图描述了这一过程：



路由控制

主机A的地址是`10.1.1.30`,要把数据发往地址为`10.1.2.10`的主机。在主机A的路由表中，保存了两个字段，由于目标地址`10.1.2.10`与`10.1.1.0/24`段不匹配，所以它被发往默认路由`10.1.1.1`也就是图中路由器1的左侧网卡的IP地址。

路由器1继续在它自己的路由控制表中查找目标地址`10.1.2.10`，它发现目标地址属于`10.1.2.0/24`这一段，因此将数据转发至下一个路由器`10.1.0.2`，也就是路由器2的左侧网卡的地址。

路由器2在自己的路由控制表中查找目标地址`10.1.2.10`，根据表中记录将数据发往`10.1.2.1`接口，也就是自己的右侧网卡的IP地址。主机B检查目标IP地址和自己相同，于是接收数据。

路由控制表

路由控制的关键在于路由控制表，路由控制表可以由管理员手动设置，称为静态路由控制，但是估计大部分人没这么干过。这是因为路由器可以喝其他路由器互换信息比即使自动刷新路由表，这个信息交换的协议并没有在IP协议中定义，而是由一个叫做“路由协议”的协议管理。

环路

上图中，假设主机A向一个不存在的IP地址发送数据，并且路由器1、2、3设置的默认路由形成了一个循环，那么数据将在网络中不断转发最终导致网络拥堵。这个问题将在下文分析IP首部时得到解决。

IP报文分割重组

在数据链路层中，我们已经提到过不同的数据链路有不同的最大传输单元(MTU)。因此IP协议的一个任务是对数据进行分片和重组。分片由发送端主机和路由器负责，重组由接收端主机负责。

路径MTU发现

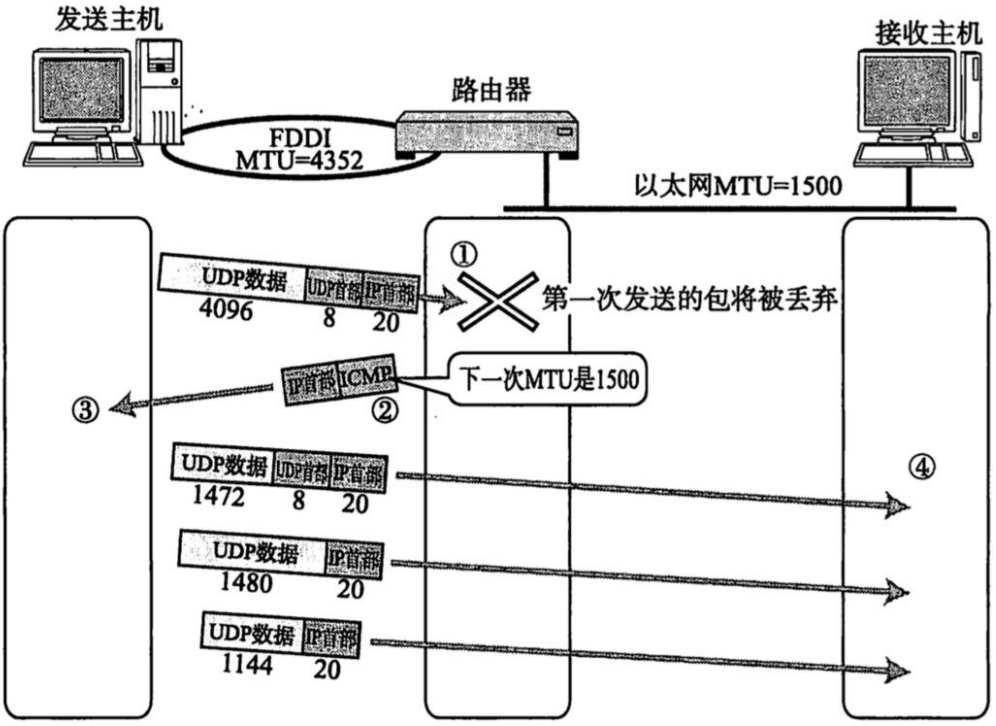
分片会加重路由器的负担，因此只要条件允许，我们都不希望路由器对IP数据包进行分片处理。另外，如果一个分片丢失，整个IP数据报都会作废。

解决以上问题的技术是“路径MTU发现”。主机会首先获取整个路径中所有数据链路的最小MTU，并按照整个大小将数据分片。因此传输过程中的任何一个路由器都不用进行分片工作。

为了找到路径MTU，主机首先发送整个数据包，并将IP首部的禁止分片标志设为1.这样路由器在遇到需要分片才能处理的包时不会分片，而是直接丢弃数据并通过ICMP协议将整个不可达的消息发回给主机。

主机将ICMP通知中的MTU设置为当前MTU，根据整个MTU对数据进行分片处理。如此反复下去，直到不再收到ICMP通知，此时的MTU就是路径MTU。

以UDP协议发送数据为例：



① 发送时IP首部的分片标志位设置为不分片。路由器丢包。
② 由ICMP通知下一次MTU的大小。
③ UDP中没有重发处理。应用在发送下一个消息时会被分片。具体来说，就是指UDP层传过来的“UDP首部+UDP数据”在IP层被分片。对于IP，它并不区分UDP首部和应用的数据。
④ 所有的分片到达目标主机后被重组，再传给UDP层。
（数字表示数据长度，单位为字节）

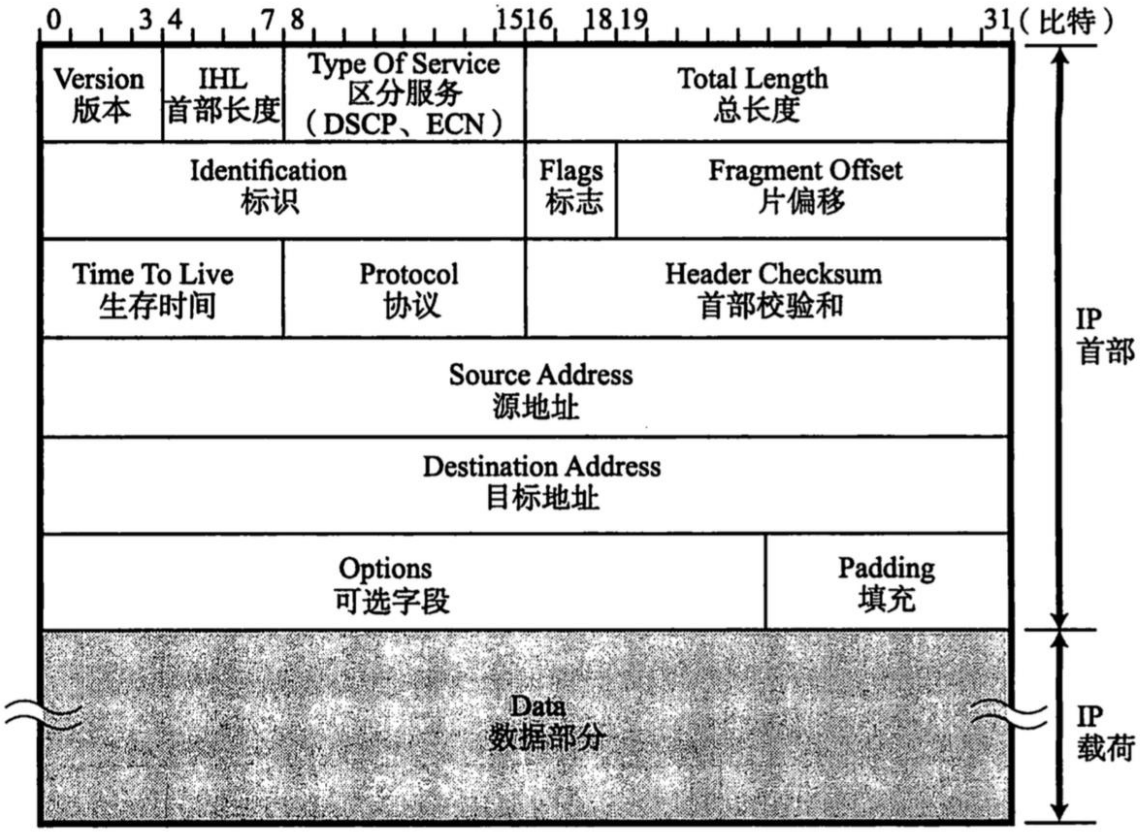
路径MTU发现

重组

接收端根据IP首部中的标志(Flag)和片偏移(Fragment Offset)进行数据重组。具体内容将在分析IP首部时详细解释。

IP首部(IPv4)

IP首部是一个有些复杂的结构，我们不用记忆它的结构，只需了解每个部分的作用即可，这样可以加深对IP协议的理解。



IP首部

其中几个重要的部分介绍如下：

- 总长度(Total Length)：表示IP首部与数据部分总的字节数，该段长16比特，所以IP包的最大长度为65535字节(2^16)。虽然不同数据链路的MTU不同，但是IP协议屏蔽了这些区别，通过自己实现的数据分片功能，从上层的角度来看，IP协议总是能够以65535为最大包长进行传输。

- 标识（ID：Identification）：用于分片重组。属于同一个分片的帧的ID相同。但即使ID相同，如果目标地址、源地址、上层协议中有任何一个不同，都被认为不属于同一个分片。

- 标志（Flags）：由于分片重组，由三个比特构成。

  第一个比特未使用，目前必须是0。

第二个比特表示是否进行分片，0表示可以分片，1表示不能分片。在路径MTU发现技术中就用到了这个位。

第三个比特表示在分片时，是否表示最后一个包。1表示不是最后一个包，0表示分配中最后一个包。

- 片偏移（FO: Fragment Offset）： 由13比特组成，表示被分片的段相对于原始数据的位置。它可以表示8192(2^13)个位置，单位为8字节，所以最大可以表示8 x 8192 = 65536字节的偏移量。

- 生存时间（TTL: Time To Live）： 表示包可以经过多少个路由器的中转。每经过一个路由器，TTL减1。这样可以避免前文提到的无限传递包的问题。

- 协议： 表示IP首部的下一个首部属于哪个协议。比如TCP协议的编号为6，UDP编号为17.

- 首部校验和：用于检查IP首部是否损坏

- 可选项：仅在试验或诊断时用，可以没有。如果有，需要配合填充（Padding）占满32比特。