第3章 以太网技术

CHAPTER

当前应用最广泛的局域网是以太网家族。以太网系列技术是目前局域网组网首选的网络技术。以太网家族包括成员如下。

(1) 10Mb/s 以太网(简记为 10ME)：它是最初被标准化的以太网，它的数据率是 10Mb/s。

(2) 快速以太网(Fast Ethernet，FE)：它的数据率是 100Mb/s。

(3) 千兆以太网(Gigabit Ethernet，GE)：它的数据率是 1Gb/s。

(4) 万兆以太网(10Gigabit Ethernet，10GE)：它的数据率是 10Gb/s。

以太网由 IEEE 802.3 委员会进行标准化，其标准如表 3-1 所示。

表 3-1 以太网标准摘要

标准 / 类型	IEEE 标准	物理层标准	批准时间(年)
10ME	802.3	10Base-5	1983
	802.3a	10Base-2	1988
	802.3i	10Base-T	1990
	802.3j	10Base-F	1992
FE	802.3u	100Base-FX	1995
	802.3u	100Base-TX	1995
	802.3u	100Base-T4	1995
	802.3x&y	100Base-T2	1997
GE	802.3z	1000Base-X	1998
	802.3ab	1000Base-T	1998
10GE	802.3ae	10GBase-LR/LW	2002
	802.3ae	10GBase-ER/EW	2002

注：表中有关概念将在本章后面的内容中介绍。

3.1 以太网通信方式

早期以太网采用总线型拓扑结构组网，网络中的主机共享数据通道。以太网解决多台主机共享信道的方法是载波侦听多路存取(Carrier Sense Multi Access，CSMA)。

3.1.1 CSMA 数据发送规则

CSMA 规定，当主机发送数据时，它首先侦听信道上是否有载波(尽管在采用基带传输的总线型局域网上不存在“载波”)。若没有载波，则表明信道上没有数据在传输(没有主机使用信道)，主机便将数据发送到信道上。若有载波，主机按以下工作方式操作。

(1) 主机仍然侦听信道直到信道空闲，然后把数据发送出去。这种方式称为 1-坚持 CSMA。

(2) 主机不再侦听信道，而是等待一个随机时间后，再侦听信道。这种方式称为非坚持 CSMA。

(3) 还有一种 p-坚持 CSMA，它应用于划分时隙的信道。在 CSMA 中，信道可以被附加上时间属性，使信道的访问在特定的时间间隔内进行。这种时间间隔就是时隙(有的文献称为时间槽、间隙)。主机在某些特定的时隙使用信道，而在其他时隙等待或侦听。在 p-坚持 CSMA 中，如果信道空闲，则侦听的主机以概率 p 发送，而以概率 $1-p$ 把这次发送推迟到下一个时隙。在信道空闲时，此过程一直重复，直到发送成功或者另外的主机开始发送。如果主机检测出信道忙，它就等到下一时隙再开始侦听。

可见，CSMA 能够做到在信道忙时，不会有其他主机打乱正在进行的发送。由于采用了信道侦听，因此在一定程度上减少了冲突发生的次数，从而提高了信道的利用率和整个网络的吞吐量。然而，CSMA 并不能完全避免冲突，原因很简单，即当主机 A 认为信道空闲时，主机 B 也可能认为信道空闲，因此两台主机差不多同时发送数据，导致冲突发生。

为了进一步提高效率，CSMA 增加了冲突检测功能，使发送方能够尽快地检测出冲突，并停止数据的发送。带冲突检测的 CSMA(CSMA with collision detection)记为 CSMA/CD。

在 CSMA/CD 中，发送者要边发送边监听(listen while talking)，只要检测到冲突，则发送者就立即停止发送数据，并发出一个特殊信号，通知其他主机发生了冲突。然后等待一段时间重新开始。

综上所述，CSMA/CD 的工作流程如图 3-1 所示。

读者可能会问，发送者如何检测出冲突呢？做法是这样的：发送者把从信道上收到的信号[①]的特性，如电压值、脉冲宽度等，与发送的信号的特性做比较，如果一致，则认为无冲突，否则认为发生了冲突。冲突次数少时，还可以继续发送帧，但在重新发送之前要发送特殊的信号通知其他机器已经发生了冲突(强化冲突)。然后等待一段随机时间再去

① 该信号一定是自己的，因为在同一时刻信道只能被一方使用。信号是被广播的，发送者也能收到。

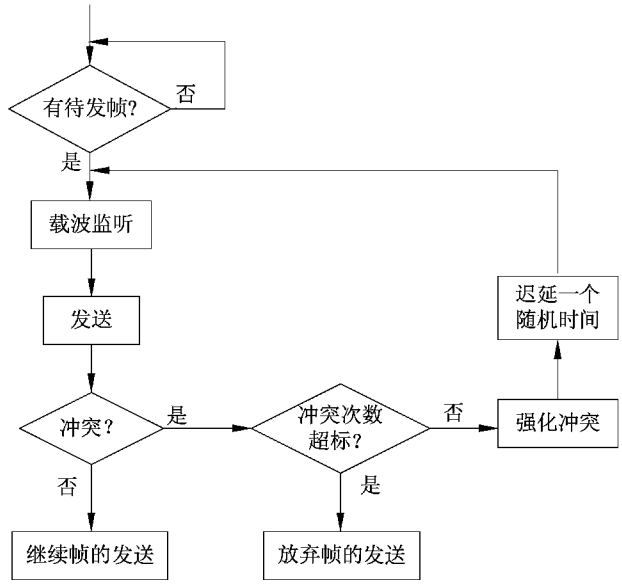


图 3-1　CSMA/CD 工作流程

侦听信道。但是当冲突次数超过一个门限值时，就要放弃发送。

CSMA/CD 可以划分时隙，也可以不划分时隙。非时隙的操作不必在时隙的起点进行。时隙操作要在时隙的起点处进行，所有主机要参考同一个时钟，因此网络中要进行时间同步。

3.1.2　影响 CSMA 效率的因素

从上面的叙述中可知，避免冲突对提高信道利用率是至关重要的，下面分析与冲突有关的因素。在分析这些因素之前，先讨论分时 CSMA/CD 的时间模型，以便后面内容的叙述。该模型如图 3-2 所示。

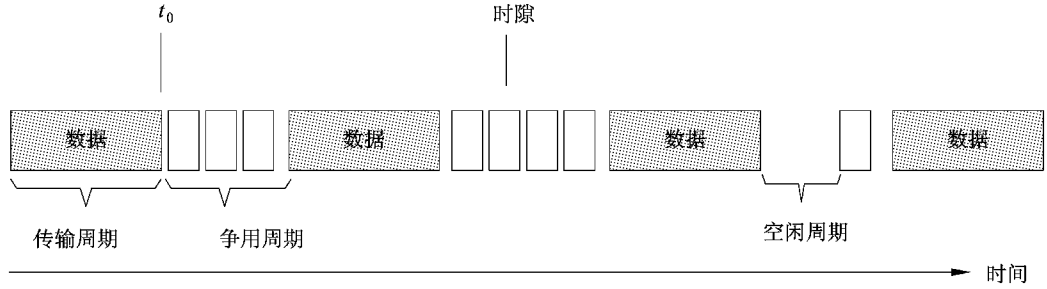


图 3-2　分时 CSMA/CD 时间模型

从图 3-2 中可以看出，对信道的使用在 3 种时间间隔（周期）内进行。这 3 种周期如下。

（1）传输周期。该时间内信道上有数据在传输。

（2）争用周期。有数据要发送的主机在这段时间内侦听信道。

（3）空闲周期。该时间内，信道上没有数据传输，也没有主机侦听信道。

每个周期由若干个时隙组成。传输周期结束后信道进入争用周期或空闲周期，有主机开始发送意味着争用周期结束，空闲周期结束后信道进入争用周期。

下面讨论这样一个问题：假如一台主机A在争用周期的第一个时隙开始（即图3-2的t_0时刻）发送数据，那么其他主机需要等待多长时间后发送数据才能避免冲突？显然，讨论这个问题对于提高信道利用率具有重要意义。

为了更有代表性地说明问题，假设一种最坏的情况，即主机A位于信道的一端，而等待发送的另一台主机B位于信道的另一端。之所以说这是最坏的情况，是因为A与B之间的传输延时τ最大，称τ为端端延时。其他位于A和B之间的主机到A的延时不大于B到A的延时。因此，如果它们也等待B要等待的时间的话，也不会造成冲突。

假设在t_0时刻，A的第一个比特离开A开始向B传输，经过τ后到达B。如图3-3所示，如果B在t_0到$t_0+\tau$之间发送数据，就会导致冲突。

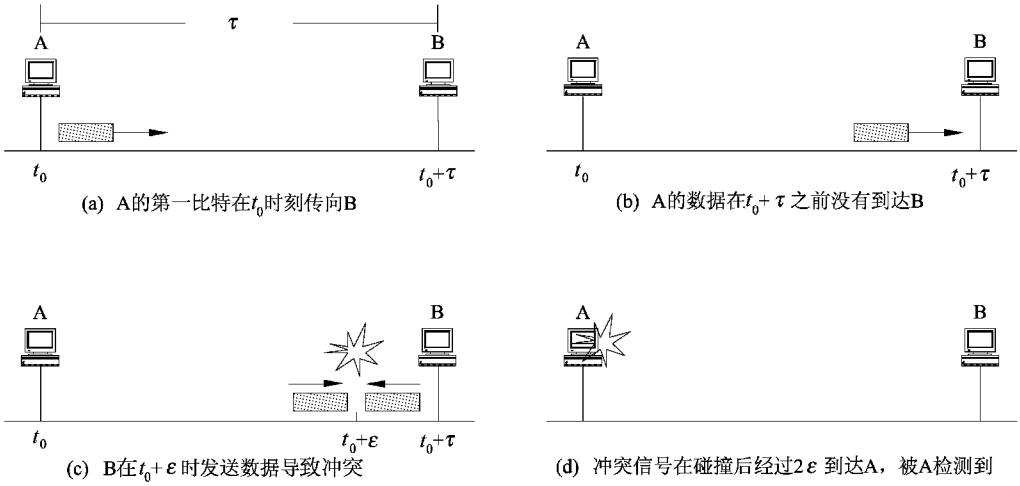


图3-3 冲突在t_0到$t_0+\tau$之间发生

冲突信号需要经过2ε到达A，此时A才能检测出冲突。可见，从t_0时刻起，要想避免与A发生冲突，其他主机至少要在$t_0+\tau$之后发送数据。而A最晚可能在$t_0+2\tau$时刻检测到冲突。

显然，τ与A到B之间距离d和信号传输速度r有关，即

$$\tau=d/r$$

而传输速度r又与信道有关。对于UTP电缆信道，电磁波在电缆中的传播速度r最大只有在自由空间的65%左右，因此，1km电缆的$\tau\approx5.12\times10^{-6}\text{s}=5.12\mu\text{s}$。

读者可能会问这样的问题：数据发送需不需要时间？实际上主机发送数据也需要时间。从数据的第一个比特离开主机到最后一个比特离开主机之间的时间称为发送时间。单位时间内发送的比特数，称为比特率。因此，当A的第一个比特到达B时，可能A发送数据的最后一个比特还在A中。同样，主机接收数据也需要时间。

显然，数据的发送时间与数据的长度 L 和比特率 R 有关（R 又与数据编码方式有关）。如果要发送的数据长度太小，数据很快就发送完毕，此时 A 处没有数据，A 发出的数据在广播信道上。因此，A 认为没有冲突发生，但实际上存在潜在的冲突。为此要对数据的长度有个最小限制，保证在发送完毕之前，能够检测到最晚到来的冲突信号。前面已经说过，最晚到来的冲突信号从最远端过来，即经过 τ（秒）后到达。这样数据从离开到冲突到来最大用时为 2τ。因此，数据帧的最小长度 $L_{\min}$ 应该满足关系：

$$2\tau = L_{\min}/R \tag{3-1}$$

即

$$L_{\min} = 2\tau R$$

又由于

$$\tau = d/r$$

所以

$$L_{\min} = 2dR/r$$

其中 d 为网络共享链路的长度（常被称为网络直径，单位是 m），R 是数据率（单位是 b/s），r 是信号传输速率（单位是 m/s）。

注意：通过上面的分析，可以得出以下结论。

（1）共享信道上的主机要在 A 发送数据后的 τ 时间后发送数据才可能避免冲突。

（2）A 发送数据的长度要大于 $2tR$，才能避免潜在的冲突。

（3）从式（3-1）可以看出，为了避免冲突、提高信道的效率，应该在 d，R，r 之间做出合理的匹配。换句话说，当网络规模（d）不同时，使用的信道（r）和采用的编码方式（R）也要做相应的调整。

3.1.3　IEEE 802.3 标准的 CSMA/CD

IEEE 802.3 以太网使用 CSMA/CD 发送数据。IEEE 802.3 使用的时间模型如图 3-4 所示。

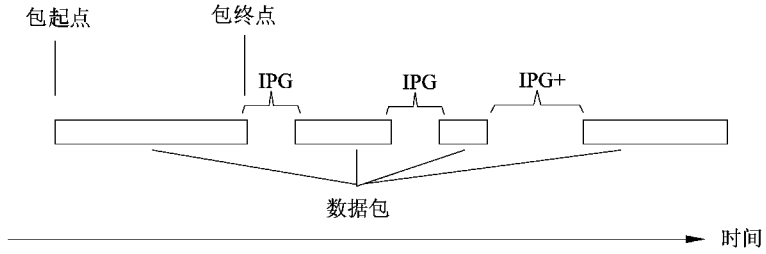


图 3-4　IEEE 802.3 CSMA/CD 时间模型

图 3-4 中的 IPG 称为包间隔（Inter Packet Gap）。实际上，此处包是指封装了 MAC 帧的物理层数据单元，前导/SFD 就是包头部[①]。在 IEEE 802.3 10Mb/s 以太网标准中，

① 参见 3.2.1 节。

IPG＝9.6μs；在快速以太网（FE）中，IPG＝0.96μs；在千兆以太网（GE）中，IPG＝0.096μs。IPG相当于争用周期。IEEE 802.3 CSMA/CD必须在一个IPG的起点开始发送，而不是在其中间某时刻。帧和帧之间至少要有一个IPG。

在IEEE 802.3 CSMA/CD中，当检测到冲突时，发送方要发送一个32b的JAM信号，其他主机收到此信号后即认为发生了冲突。然后要等待一段随机时间，再去侦听信道。这个过程称为后退。等待的随机时间由二进制指数后退算法计算。计算公式是：

$$\text{后退时间}=2^{\min(N,10)}R\tau$$

其中，N是发送尝试次数，R是一个随机整数，R的作用是保证后退时间的随机性，避免冲突方因后退了相同的时间而再同时发送引起的冲突。IEEE 802.3规定为了成功地发送一个帧可以尝试15次（一共发送16次）。若再不成功，则放弃发送。τ是时间槽（端端延时）。可见，后退时间是时间槽的整数倍。

综上所述，IEEE 802.3 CSMA/CD详细工作流程如图3-5所示。

图中判断是否耗时IPG和延迟剩余的IPG是保证在传输周期的起始时刻发送数据。

下面，观察式(3-1)中的参数在IEEE 802.3 CSMA/CD实现的情况。由于在链路上编码信号的衰减，链路的最大长度有一定的限制。在最早的IEEE 802.3a粗同轴电缆以太网标准中，链路最大长度为500m，帧最小发送时间为51.2μs。对于10Mb/s数据率的以太网，最小帧长度＝最小发送时间/比特宽度＝51.2μs/(1/10Mb/s)＝51.2μs×10b/μs＝512b，即64B。10Mb/s以太网的比特宽度为1/10μs。

在100Mb/s快速以太网中，比特宽度为1/100μs，而由于最小帧长度没有增大，还是64B，那么式(3-1)中哪些指标做了修改呢？事实上是d做了修改，两个节点间的最大距离是100m。另外，编码做了修改，用4B/5B编码代替了10Mb/s以太网使用的曼彻斯特编码。

对于传输速率为1Gb/s的千兆以太网，比特宽度为1/1000μs。为了保持与早期10Mb/s，100Mb/s以太网的兼容和降低升级的成本，GE没有对d＝100m和链路介质做出修改，而是修改了帧长度，增加了扩展字段，增加了发送时间，从而抵消了比特宽度缩短带来的负面影响。

为了保证与以前所有的以太网兼容，万兆以太网（10GE）采用了另一种修改帧格式的策略。对于10Gb/s以太网，比特宽度更小，仅为GE的1/10。但10GE仅增加了2B的HEC字段，这样做的原因在于10Gb/s以太网采用光纤链路技术（如SDH），传输延时和d的限制放宽了。

CSMA/CD由局域网的MAC子层实现。IEEE 802.3 MAC子层实现CSMA/CD的细节包括以下内容。

(1) 时间槽τ：10ME是512比特时间（＝51.2μs），FE是512比特时间（＝51.2μs），GE是512×8比特时间（＝4.096ms）。

(2) MAC帧长度：10ME和FE，GE，10GE是64～1518B，GE通过载波扩展后帧的长度是512～1966B。

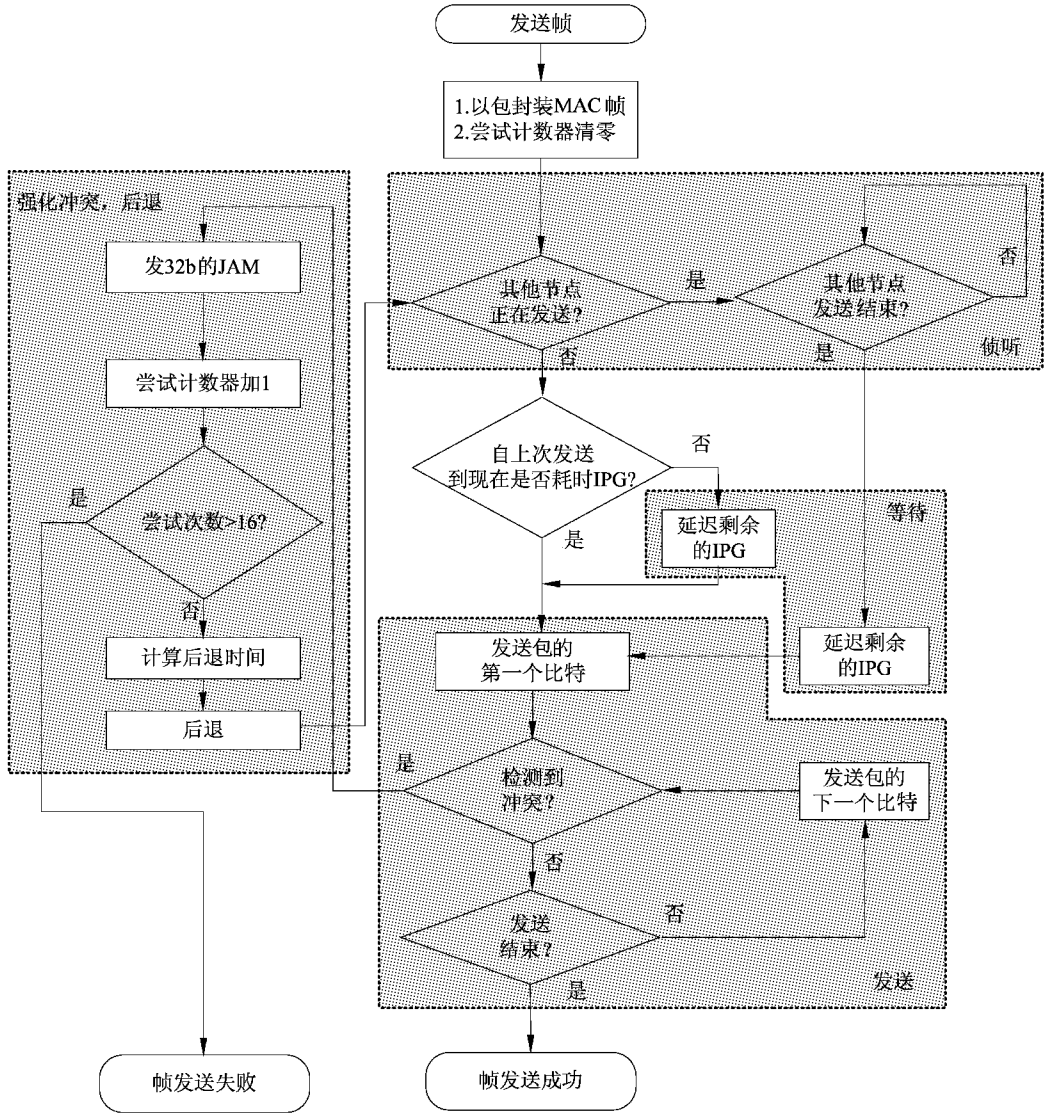


图 3-5　IEEE 802.3 CSMA/CD 工作流程

(3) IPG：96 比特时间。比特时间随数据率变化。

(4) 最大重发次数：15 次。

(5) 随机等待时间(后退时间)：等于 $2^{\min(N,10)}R\tau$。N 是重发次数。

(6) 最大后退次数：等于 $\min(N,10)=10$。

(7) 帧突发计时时间：最大为 65 536 比特时间(帧突发将在 3.4.4 节介绍)。

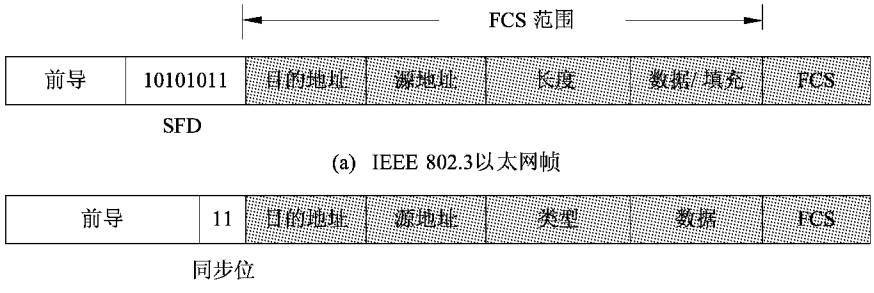
(8) 冲突检测机制：电压超出冲突阈值或发送器和接收器同时变为活动状态。

(9) JAM 信号长度：32b。

3.2 10Mb/s以太网

3.2.1 帧结构

10Mb/s以太网帧结构如图3-6所示。



(a) IEEE 802.3以太网帧

(b) DIX以太网帧

图3-6 10ME帧(阴影部分为MAC帧)

阴影部分的帧字段含义如下。

(1) 源地址/目的地址。它占2～6B,在基带信号[①]标准中只使用6B长度的地址。地址字段的内容是MAC地址,该地址由IEEE分配,唯一地标识MAC帧的发送者和接收者。

(2) 长度。它占2B,由于IEEE 802.3 MAC帧数据字段是变长的,因此要指明其实际的长度,该长度值放在长度字段中。接收MAC帧时,接收方要进行长度检查,MAC帧被放入缓冲区的同时对其位数进行计数。当时间槽结束后,检查计数值是否在512b(最小长度)至12 144b(最大长度)之间,如果是,则认为帧长度合法,否则认为帧是冲突碎片或超长帧,并将之从缓冲中消除。

(3) 数据。它占0～1500B。该字段存放MAC帧封装的上层协议的数据,即LLC帧。

(4) 填充。它占0～46B。当数据字段的长度小于46B时,需要填充数据使MAC帧的长度不小于64B,以区别其他无用帧(因冲突产生的碎片)。满足一定条件的最小帧长度可以减少冲突的发生。IEEE 802.3 MAC帧的最小长度是64B。

(5) 帧校验序列(Frame Correction Sequence,FCS)。它占4B,其内容是循环冗余校验算法(Cyclic Redundancy Check,CRC)计算出的监督比特序列。CRC算法的输入值是地址、长度、数据和填充字段等有效数据构成的二进制比特流,输出值就是包含FCS在内的整个MAC帧。

上述5个字段构成了IEEE 802.3 MAC帧。IEEE 802.3 MAC标准只处理地址、长度、数据/填充和FCS字段,而前导(preamble)和帧开始定界符(Start of Frame

① 没有经过调制的二进制信号。

Delimiter，SFD)字段由物理层处理，前导和 SFD 不是 MAC 帧的字段，它们和 MAC 帧构成以太网帧。物理层使用它们把 MAC 帧封装到物理层的数据结构中。IEEE 802.3 标准以太网帧的前导占 7B，每个字节的内容是 10101010。该字段二进制数据的曼彻斯特编码会产生 10MHz 持续 5.6μs 的方波[①]，从而使接收方与发送方的时钟同步。帧开始定界符 SFD 占 1B，其内容为 10101011，表明 MAC 帧即将开始。

图 3-6 中 DIX 以太网是由 DEC，Intel 和 Xerox 三家公司联合开发的。DIX 以太网帧的前导有 62b，它们由 1 和 0 交替排列组成。同步位是 11。因此，DIX 以太网帧和 IEEE 802.3 以太网帧左面 64b 是一样的。

下面，再把上面涉及的 MAC 地址和 CRC 校验码详细介绍一下。

1. MAC 地址

IEEE MAC 地址的结构如图 3-7 所示。

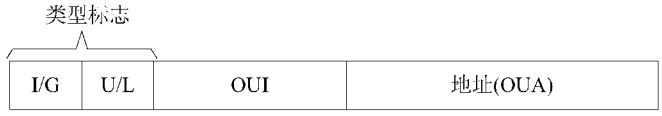


图 3-7　MAC 地址结构

MAC 地址字段含义如下。

(1) I/G。它占 1b，称为单个或组地址标志(individual or group address flag)。它指明该 MAC 地址是单个的地址(0)，还是组地址(1)。当 I/G＝1 时，MAC 地址的其余 47b 用于编码组地址，此时 U/L，OUI 和 OUA 字段不存在。

(2) U/L。它占 1b，称为通用或局部管理标志(universal or local administration flag)。它指明该 MAC 地址是全局唯一的(U/L＝0)，还是局部唯一的(U/L＝1)。由非网络接口设备厂商生产的设备，其 U/L＝1，此类设备一般专用于局部网络。

(3) OUI。它占 22b，称为组织唯一标识(organizationally unique identifier)。它指明具有该 MAC 地址的设备是由哪个供应商生产的。OUI 由 IEEE 分配给厂商，一个厂商可以得到 1 个或几个 OUI。

(4) OUA。它占 24b，称为组织唯一地址(organizationally unique address)。它是厂商给自己生产的接口设备分配的编号。同一厂商设备的 OUA 两两不同。因此任意两个 MAC 地址都不相同。

MAC 地址分为以下 3 类。

(1) 单个通用 MAC 地址。该 MAC 地址的 I/G＝0 且 U/L＝0，它全局唯一地寻址一个接口设备。当局域网数据帧只能发给一个接口设备时(此方式称为单播)，该帧的目的地址必须使用这种 MAC 地址。这种地址又称单播地址。

(2) 广播 MAC 地址。48b 全为 1 的 MAC 地址称为广播 MAC 地址。当局域网帧要发给网上所有接口设备时，该帧的目的地址必须使用这种地址。它可被看成是一种特殊的组地址。

---

① 曼彻斯特编码是 IEEE 802.3 以太网物理层采用的把 MAC 帧编码为基带信号的编码方式。

（3）组播 MAC 地址。I/G=1，但其余 47 位不全为 1 的组地址。当局域网帧发给网上多个接口设备时，就要在该帧的目的地址字段中填入代表这些设备的组地址。

局域网硬件在识别 MAC 地址时，实际上只关心它是单播地址还是组地址，而不去处理 U/L，OUI 或 OUA 字段。

MAC 地址标准被广泛地应用于 IEEE 局域网，而且也被其他标准的局域网采用。例如美国国家标准委员会 ANSI 制定的环型局域网光纤分布数据接口（Fiber Distributed Data Interface，FDDI）也使用了 MAC 地址。

由于 MAC 地址是协议参考模型第二层协议（MAC 协议）使用的地址，它用于寻址网络接口，而且被固化在网络接口的物理硬件中，不能被重新设置，因此它又经常被称为第二层地址、物理地址或硬件地址。

2. CRC 编码算法

由于数据在通信线路上传送时可能因为电磁干扰等原因使数据发生错误，因此，接收方在接收数据时要检查数据中是否包含错误，这称为差错控制。差错控制的基本思想是在发送有效数据之前应先把它编码，然后再发送。编码以后，数据中包含了若干位的监督比特。接收方收到编码数据后，按照一定的规则判断数据在传送过程中是否发生了错误。如果没发生差错，就进行解码，把有效数据从编码中还原出来。

CRC 编码算法的基本步骤如下。

（1）第一步，把 k 比特有效数据表示成 $M(x)=M_{k-1}x^{k-1}+M_{k-2}x^{k-2}+\cdots+M_1x+M_0$，其中 $M_i(i=0,\cdots,k-1)$就是 k 比特的值（0 或 1）。例如，如果有效数据是 10011，则对应的 $M(x)=x^4+x+1$。$M(x)$称为报文多项式。

（2）第二步，定义 $T(x)=X^{n-k}M(x)-C(x)$，其中 $C(x)$是 $X^{n-k}M(x)$除以 $G(x)$的余式，显然 $T(x)$能被 $G(x)$除尽。$G(x)$是生成多项式，IEEE 802.3 标准采用的 $G(x)$是：

$$G(x)=x^{32}+x^{26}+x^{23}+x^{22}+x^{16}+x^{12}+x^{11}+x^{10}+x^8+x^7+x^5+x^4+x^2+x+1$$

它能生成 32 位的监督比特。$T(x)$多项式的系数序列 $T=\{T_{n-1},T_{n-2},\cdots,T_0\}$就是发送方发送的编码数据。接收方收到 T 后，让 $T(x)$除以 $G(x)$，如果能除尽，则认为接收的比特没有错，否则认为有错误。

那么，接收方判定没有差错之后，如何从 $T=\{T_{n-1},T_{n-2},\cdots,T_0\}$中分离出有效数据呢？换句话说，在 $T=\{T_{n-1},T_{n-2},\cdots,T_0\}$中哪些比特是有效数据呢？实际上，从 $T(x)$的定义中不难得知，存在以下对应关系：$T_{n-1}=M_{k-1}$，$T_{n-2}=M_{k-2}$，$\cdots$，$T_{n-k}=M_0$。因此，接收方提取 $T_{n-1},T_{n-2},\cdots,T_0$的左边 k 比特作为有效数据即可。对于 IEEE 802.3 以太网 MAC 帧而言，这 k 比特就是源地址、目的地址、长度和数据/填充字段的数据，而右边的 $n-k(=32)$比特恰是 CRC 算法产生的 FCS。

3.2.2 网络接口结构

10Mb/s 以太网卡的结构原理如图 3-8 所示。

图 3-8 中主要结构的功能如下。

（1）接口控制器。接口控制器包含多个寄存器和锁存器，用于控制网卡的接口电路。接口电路用于接收来自主机的控制命令并按序执行。网卡接口电路包含两部分，一部分