第三章　单片机高级特性

时至今日，单片机的技术已经发展到前所未有的地步，PC 流行大旗刚刚树起的九十年代，主频终于突破 100MHz，简称 586 的奔腾一代开始用软解压向人们结结巴巴的演示多媒体的未来，就是 INTEL 自己也为这一进步激动不已，从此电视广告中“Beng Beng Beng Beng”的旋律成为 INTEL 的象征。

让我们来看看当时让 INTEL 如此激动的奔腾电脑的摸样：

1996 年

100M 主频 Intel Pentuim CPU

16M 内存

1M 显存显卡

850M 硬盘

14 寸彩显

大概需要 8000~10000 人民币

再来看一看现在 iPhone 使用的三星 64xx 的 MCU（以某开发板为例）：

Samsung S3C6410，ARM1176JZF-S 内核，主频 533MHz/667MHz

128M Bytes DDR RAM

256M Bytes NAND Flash

2M Bytes NOR FLASH

100Mbps 以太网接口

USB HOST 接口

USB Device 接口

AC97 接口

双高速 SD 卡接口

双 LCD 接口

VGA 接口

TV OUT 接口

S-VIDEO 接口

双摄像头接口

2D/3D 硬件加速

带 800*480 的低成本液晶屏开发板成本大约为 300~400 人民币

只要简单对比就可以知道今天的高端单片机在性能方面已经远超当年的奔腾电脑，单片机要发展到这一步肯定不能拘泥在早期单片机技术的框架当中，需要不断引入一系列新技术，这些技术有可能是早期电脑才能采用“昂贵”技术，随着技术的不断进步才逐渐平民化为单片机所用，这一章让我们来一起了解单片机的这些高级技术。

本章的内容如果你看不明白并不要紧，你糊里糊涂的看就行了，知道有这么回事，等到有一天你面对这些技术时突然有恍然大悟的感觉时再回来与你的理想做对比。有告诉你一个秘密，这一章中的内容其实我自己也不大明白，就是许多专家也不完全明白。

3.1. Cache

首先得清楚什么 Cache，Cache 是英文中对高速缓存系统的称谓，Cache 的概念在硬件和软件中都存在。这里我借鉴别人的一个例子来解释 Cache 的作用：软件高速缓存的作用产生于人们使用数据不平均时，我们虽然常常拥有大量数据，但最经常使用的往往只有其中一小部分。如国标汉字不到 7000 个，但经常使用的只有 2000~3000 个，其中几百个又占了 50%以上的使用频率，如果将这几百个放到存取最快的地方，就可以用很小的代价大大提高工作速度。我们知道内存的存取速度比硬盘快得多，程序一起启动我们就将常用几百个字模装入内存指定区域，当使用这部分字的时候直接从内存取字，其余的才会去读硬盘。我们知道内存的读取速度为硬盘的数万倍，假设我们有一本书需要显示，预装几百个字模到内存指定区域的方法差不多将平均读取速度提高一倍，如果将预装的字模数增加到常用 2000~3000 个，读取速度甚至可以提高十倍。

这里我们要说的 Cache 是指一种用来加速存储器读写操作的硬件存储器，象买电脑时常说的一级高速缓存/二级高速缓存就是 CPU 内部的这种硬件存储器，和软件高速缓存比虽然是两种不同的方式，但其作用是一样的，都是为了提高读写速度。

可能有人会有这样的疑问，明明 RAM 已经是一种存取速度非常快的硬件，为什么还需要 Cache 呢，是的现在的 RAM 可以提供超过 100M 的读写速率，但这个速度同 CPU 的处理速度相比并不存在优势，甚至远小于 CPU 的处理速度，象 S3C6410 工作在 667MHz 主频下，就是一条需要 4 个周期的指令执行也只需要 6ns，而 RAM 的读写时间需要 10ns，显然 RAM 的速度无法满足 CPU 的高速处理要求。

如何解决 CPU 与 RAM 之间的这种速度差异问题？通常有下列方法：

一、在基本总线周期中插入等待，当 CPU 需要读写 RAM 数据的时候，先向 RAM 发送读写命令，再等待 RAM 处理好总线数据完成一些读写操作，这样做显然会浪费 CPU 的能力，就象我们设计了速度都可以达到 120 公里/小时的汽车和高速公路，可高速公路每隔十公里就设置一个收费站，这样再好的汽车也无法跑出快的速度来。

二、采用存取时间较快的 SRAM 或其它新型存储器材作存储器，这样虽然解决了 CPU 与存储器间速度不匹配的问题，但却大幅提升了系统成本。另外还有一个问题，大容量 RAM 作为外部器件，需要通过外部连线将其与 CPU 连接起来，这些外部连线因为分布电容等问题使得 RAM 与 CPU 之间的

最高传输速率有限制，如果对 PCB 布板要求过高不利于生产推广，这个问题同样可以用高速公路的例子来理解，汽车的速度可以继续提高，收费站也可以撤掉，但实际生活中高速公路不可能设计成笔直宽阔的大道，所以还是不能满足汽车速度的需求。

三、在慢速的 RAM 和快速 CPU 之间插入一速度较快、容量较小的 SRAM，起到缓冲作用，使 CPU 既可以以较快速度存取 RAM 中的数据，又不使系统成本上升过高，这就是 Cache 法。目前，一般采用这种方法，它是在不大增加成本的前提下，使 CPU 性能提升的一个非常有效的技术。

当然 Cache 的实现并不是简单的插入一块小容量高速存储器那么简单，是基于程序统计规律通过一系列复杂控制技术才得以实现，而且它并不是万能的，同样存在缺陷，后面我们会详细讲述这些细节。

先看一下 ARM 关于存储器的结构图。

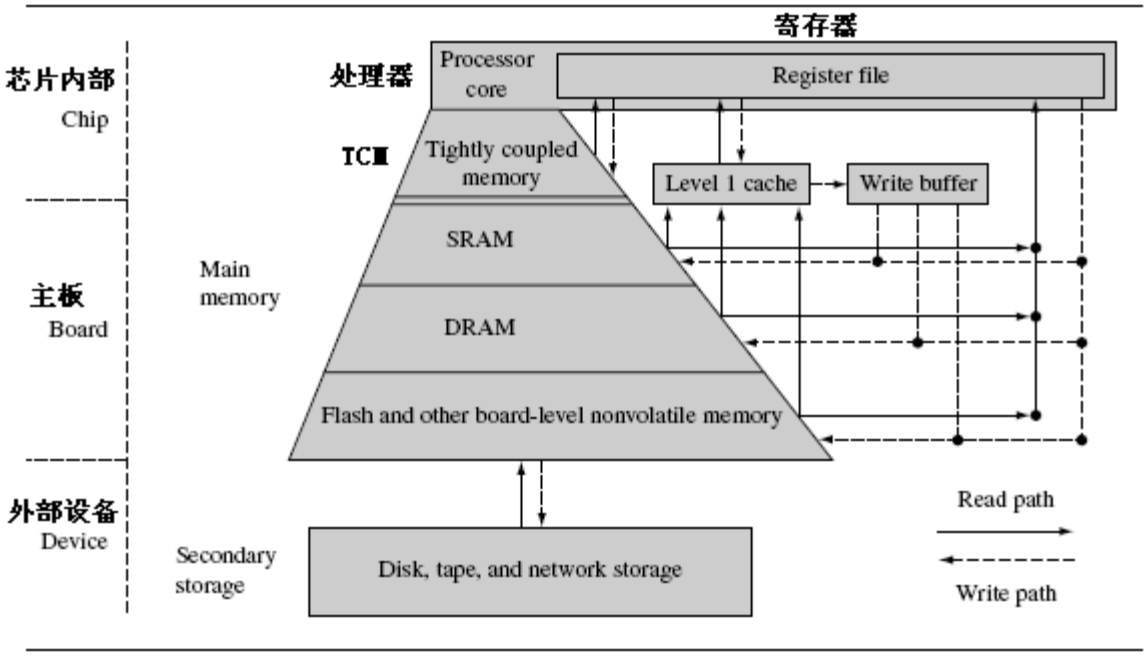

图 3.1.-1　ARM 存储器示意图

TCM（摘自 ARM 论坛）：

TCM 是一个固定大小的 RAM，紧密地耦合至处理器内核，提供与 cache 相当的性能，相比于 cache 的优点是，程序代码可以精确地控制什么函数或代码放在那儿(RAM 里)。当然 TCM 永远不会被踢出主存储器，因此，他会有一个被用户预设的性能，而不是象 cache 那样是统计特性的性能提高。

TCM 对于以下几种情况的代码是非常有用、也是需要的：可预见的实时处理（中断处理）、时间可预见（加密算法）、避免 cache 分析（加密算法）、或者只是要求高性能的代码（编解码功能）。随着 cache 大小的增加以及总线性能的规模，TCM 将会变得越来越不重要，但是他提供了一个让你权衡的机会。

那么，哪一个更好呢？他取决于你的应用。Cache 是一个通用目的的加速器，他会加速你的所

有代码,而不依赖于存储方式。TCM 只会加速你有意放入 TCM 的代码,其余的其他代码只能通过 cache 加速。Cache 是一个通用目的解决方案，TCM 在某些特殊情况下是非常有用的。假如你不认为需要 TCM 的话，那么你可能就不需要了，转而加大你的 cache，从而加速运行于内核上的所有软件代码。

可以看出 Cache 位于芯片内部，通过内部总线与 CPU 相连，图中自上而下的存储器离处理器越远读写速度就越慢，Cache 本质也是 SRAM，只是对其增加了一些特殊的读写控制方法。同样是 SRAM，片外的 SRAM 速度比片内要慢，这就是外部总线的影响。Cache 是不能独立当作存储器使用的，对于程序员来说，它并没有特定的地址可以进行访问，只是处理器提供了一些控制指令可以让程序员对 Cache 进行控制方法的设定，所以为了针对某些特殊应用芯片厂商会在芯片内部另外会放置一小段 SRAM，这段 SRAM 对于程序员来说就有特定的地址与之对应，程序可以当作普通 RAM 进行读写。

Cache 的工作原理

通常程序代码都是连续的，代码执行时都是一条接一条的连续执行，程序中跳转操作所占的比例并不高，即便是跳转指令，大多数时候跳转的距离都不会太远，加上指令地址的分布本来就是连续的，另外象程序中的循环体要重复执行多次，这样在一个较短的时间间隔内，由程序产生的地址往往集中在存储器地址空间的很小范围之内，因此对这些地址的访问就自然地具有时间上集中分布的倾向，对大量典型程序运行情况的统计分析结果也验证了这一点。

数据分布的这种集中倾向没有程序代码明显，程序中的数据读写操作虽然大多数时候也是处在相邻区域，但间距大过程序代码几率要高，不过数组的存储和访问还是会让存储器地址相对集中。

```
UINT32 i;
UINT32 data_buf1[1024];
UINT32 data_buf2[1024];

for(i=0;i<1024;i++)
{
  data_buf1[i]=i;
}
for(i=0;i<1024;i++)
{
  data_buf2[i]=data_buf1[i];
}
```

假定样例代码中的变量 i 和数组 data_buf1[]与 data_buf2[]分布是连续的，两段循环代码可以肯定是连续分布。

第一个循环：

```
for(i=0;i<1024;i++)              //需要读写 i
{
  data_buf1[i]=i;                //顺序写数组 data_buf1[]的每一个成员
}
```

很明显循环体的代码量非常小，执行这段代码完全满足代码地址在一小段区域之内的要求，但对数据的读写则有点不同，当 i 在 0 附近时，data_buf1[i]=i 的操作 RAM 地址间隔并不大，但当 i 逐渐增大情况就发生了变化，比如 i 为 1000 时，data_buf1[i]=i 的操作对 RAM 的操作跳转就会变得比较大，按假定条件 i 和写 data_buf1[1000]在 RAM 中的位置间隔有 4000 字节，虽然这 4000 个字节并不是特别大，但如果数组大小从 1024 变为 1024*1024 呢？间隔就会非常之大。

第二个循环：

```
for(i=0;i<1024;i++)              //需要读写 i
{
  data_buf2[i]=data_buf1[i];  //顺序读写数组 data_buf1[]和 data_buf2[]的每一个成员
}
```

和循环一相比情况数据在 RAM 的读写跳跃间隔会更大，每一次循环都要在两个数组之间切换，这样跳跃的间隔为数组的大小 4096 字节，变量 i 与数组 data_buf2[]之间的间隔全部超过 4096 字节。

Cache 的工作原理正是是基于程序访问的局部性来实现的，如果把较短时间间隔内的代码从外部 RAM 放到做为内部 Cache 的 SRAM 中执行，这段代码显然会因为不需要等待外部 RAM 的存取操作而获得更高的执行效率，但 Cache 的容量有限，只能放置少量的代码，还需要通过某种方法让整个程序都是在 Cache 中执行才有实际意义。Cache 对于程序代码的效果总体上看要优于数据。

Cache 的实现思路并不复杂，由处理器硬件不断地将与当前代码相关联的一小段后续代码从 RAM 中读到 Cache，然后再与 CPU 高速传送，从而达到速度匹配。CPU 对存储器进行数据请求（包含代码执行和数据读写）时，通常先访问 Cache。通过前面的分析我们知道由于局部性原理并不能保证所请求的数据百分之百地在 Cache 中，这里便存在一个命中率，即 CPU 在任一时刻从 Cache 中可靠获取数据的几率。

命中率越高，正确获取数据的可能性就越大，命中率和 Cache 的容量是成正比的。一般来说，Cache 的容量比 RAM 的容量小得多。从成本考虑，是 Cache 的容量越小越好，最好是不用，但 Cache，太小会使命中率太低；站在最大发挥 CPU 功效的角度，最好是能达到 100%的命中率，这样就希望 Cache 的容量尽可能的大，但过大不但会增加成本，因为命中率和容量之间的关系不是线性比例关系，当容量超过一定值后，命中率随容量的增加将会变得不明显。

只要 Cache 的空间与 RAM 空间在一定范围内保持适当比例的映射关系，是可以保证 Cache 的命中率达到一个比较高的比例。统计分析的结论告诉我们，当 Cache 与 RAM 的空间比例关系在 1：256 时，即 4kBytes Cache 映射 1Mbbytes RAM 既能得到较高的命中率，又不至于因为 Cache 导致成本上升太多。在这种情况下，命中率可以达到 90%以上，至于没有命中的数据，CPU 只好直接从内存

获取。获取的同时，也把它拷进 Cache，以备下次访问。

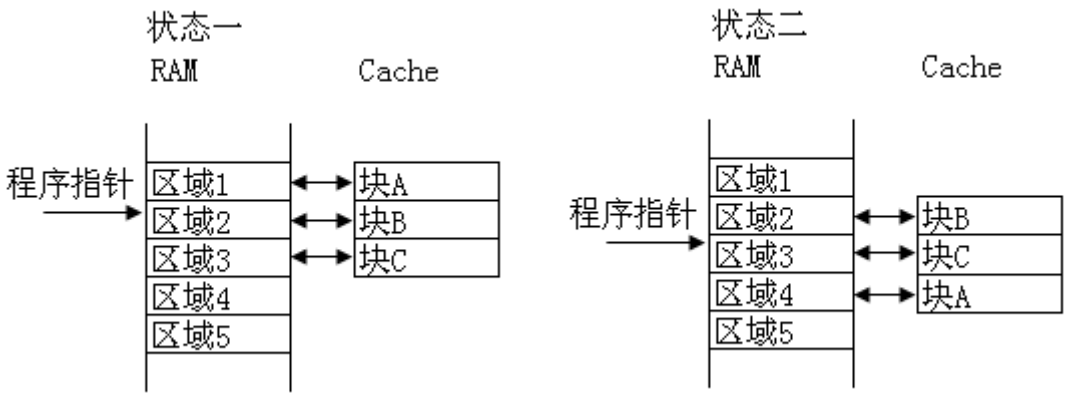

图 3.1.-2　Cache 映射示意图

图示为一种简单的 Cache 控制方式，CPU 从 RAM 存取数据（包括程序代码和程序数据）时均通过 Cache 进行，Cache 分成三块，块 B 对应当前程序指针所在的区域，块 A 和块 C 则分别对应其前后的相邻区域。通常程序不会产生大距离的跳转，所以程序大部分时候下一条指令所在位置都会位于这三个区域之中，当程序运行到程序指针会换区时，硬件会自动更新 Cache 的映射状态，这样就能继续维持程序指针始终指向 Cache 中间块。图示状态一到状态二程序指针从 RAM 的区域 2 换到区域 3，此时硬件自动将块 A 原本映射区域 1 的内容释放给区域 1，改为装载区域 4 的内容并建立映射关系，虽然这个更新过程同外部 RAM 同步数据需要一定时间，当前程序指针所指向的块自己有一定空间可以耗费程序一段时间以实现更新，当 CPU 需要区域 4 中的内容时，大部分时候 Cache 的更新已经完成。

这种简单的 Cache 控制方法在效果方面不是特别理想，对于 CPU 存取数据的地址跳转情况没有做出一个比较好的应对方法，对于一些特殊的程序模式命中率会比较低。

```
if(flag==1)
{
  func();
}
flag=0;
```

代码中的条件执行效果就会不怎么好，程序会依据 flag 的值选择是继续执行还是调用函数 func()，一般来说函数入口地址与当前所执行代码的位置间距都不会小，也就是说对于程序选择执行函数 func()的情况命中率很大机会不够理想，差的时候甚至是 0。

对于这种效果不理想的情况并非没有应对良策，正常编写的程序一定满足局部集中性的规律，

除非写出全是跳转的变态程序，如果将 Cache 的分块分得更为精细，每一块的空间尽可能的变小，所映射的区域不要求必须连续，而是通过某种方法预测知道后面执行代码的可能位置，提前将这些位置所在的区域建立映射，同样还是可以有效的保证高命中率。

象例中代码执行完 if(flag==1)语句后就有两种可能，一是调用函数 func()，二是执行 flag=0，什么方法可以让我们知道这两种可能的存在呢，现在我们看代码一下就可以看出来，如果能让 CPU 做到这一点自然也就可以实现，一种叫流水线的技术实现了我们的期望。这种技术一是可以避免 CPU 需要取指、译码、执行步骤串行进行产生的时间等待；二就是一定程度上可以解决程序跳转 Cache 命中率低的问题，在进行 if(flag==1)比较之后肯定会有一条条件转移指令，在执行比较之前流水线技术就可以知道当前比较指令之后一定有跳转指令，将两种可能地址的内容都用 Cache 的精细小块建立起映射关系，这样无论程序比较结果是哪一种都能保证后续的代码已经映射到 Cache 之中。

当然要真正做到这一步并不是我所说的那么简单，需要一系列的复杂技术才可以实现，这里是为了便于大家理解而做出的最简解释，如果想完全理解透有关 Cache 的实现技术，还需要大家自己查阅相关资料，接下来为大家介绍一些 Cache 的技术要点。

Cache 的基本结构

Cache 通常由相联存储器实现，相联存储器的每一个存储块都具有额外的存储信息，称为标签(Tag)。当访问相联存储器时，将地址和每一个标签同时进行比较，从而对标签相同的存储块进行访问。如果地址没有找到与之匹配的标签，则需要将原有的标签按一定规则丢弃一个，然后将其改映射到新地址。

Cache 的三种基本结构如下：

全相联 Cache

在全相联 Cache 中，存储的块与块之间，以及存储顺序或保存的存储器地址之间没有直接的关系。程序可以访问很多的子程序、堆栈和段，而它们是位于主存储器的不同部位上。

因此，Cache 保存着很多互不相关的数据块，Cache 必须对每个块和块自身的地址加以存储。当请求数据时，Cache 控制器要把请求地址同所有地址加以比较，进行确认。

这种 Cache 结构的主要优点是，它能够在给定的时间内去存储主存器中的不同的块，命中率高；缺点是每一次请求数据同 Cache 中的地址进行比较需要相当的时间，速度较慢。

直接映像 Cache

直接映像 Cache 不同于全相联 Cache，地址仅需比较一次。

在直接映像 Cache 中，由于每个主存储器的块在 Cache 中仅存在一个位置，因而把地址的比较次数减少为一次。其做法是，为 Cache 中的每个块位置分配一个索引字段，用 Tag 字段区分存放在

Cache 位置上的不同的块。

单路直接映像把主存储器分成若干页，主存储器的每一页与 Cache 存储器的大小相同，匹配的主存储器的偏移量可以直接映像为 Cache 偏移量。Cache 的 Tag 存储器(偏移量)保存着主存储器的页地址(页号)。

以上可以看出，直接映像 Cache 优于全相联 Cache，能进行快速查找，其缺点是当主存储器的组之间做频繁调用时，Cache 控制器必须做多次转换。

组相联 Cache

组相联 Cache 是介于全相联 Cache 和直接映像 Cache 之间的一种结构。这种类型的 Cache 使用了几组直接映像的块，对于某一个给定的索引号，可以允许有几个块位置，因而可以增加命中率和系统效率。

单纯从字面可能不容易理解这三种结构到底有何不同，用一个 8kBytes Cache 和 32Mbytes RAM 的情况来解释一下这三种结构（和实际情况可能有所不同)。

全相联将 8kBytes 的 Cache 分成 16Bytes 大小 512 小段，每段映射 32Mbytes RAM 中的一个地址，CPU 存取数据时从这 512 个 Cache 小段中查询是否命中。

直接映像将 32Mbytes RAM 按 8kBytes 的大小分成 4096 页，这样每页的大小与 Cache 一致，当 CPU 需要从 RAM 存取数据只要判断存取数据的地址是否在页面映射的区域之内就知道数据是否命中。

组相联将 8kBytes 的 Cache 分成 1kBytes 大小的 8 等份，现在 32Mbytes RAM 应该分成大小为 1kBytes 的 32768 页，CPU 存取数据时从这 8 个 Cache 分区中查询是否命中。

Cache 与 RAM 的数据一致性

在 CPU 与 RAM 之间增加了 Cache 之后，便存在数据在 Cache 和 RAM 中一致性的问题。

对 Cache 的读写有两种 2 种方式：

直写法(write through)

直写法是当 CPU 在写 Cache 的同时，写入 Cache 中的新内容也会随之更新 RAM 中对应的内容，这样 RAM 和 Cache 中的内容始终是一致的，因为需要写 RAM，所以速度会慢一点，但比没有 Cache 的情况要快，如果没有 Cache 由 CPU 直接写 RAM 需要 CPU 等待 RAM 写成功，会受到 RAM 读写速度的限制，但有了 Cache 不需要这个等待时间，写 RAM 的操作会在程序执行后续代码的同时由 Cache 控制器自动完成。

回写法(write back)

回写法和直写法不同在于当 CPU 在写 Cache 的时候并不同步更新 RAM 中对应的内容，只是在特

定位置出一个刚才所写位置已经被新内容改写的标志，这个标志位叫做脏位，直到 Cache 需要抛弃当前 RAM 位置改映射新 RAM 位置时才由 Cache 将新内容更新到 RAM 中去。

当一段程序频繁使用某些临时局部变量的时候，由于这些变量是临时的，所以根本不需要写进 RAM 中去，这样回写法效率就会非常高，不用写 RAM 就可以完成整段程序功能，对于全局变量同样也会有效果，如果一个频繁被改写的全局变量对于程序来说也不需要每次都写进新内容，只需要在抛弃 Cache 映射关系时更新最后的内容就可以。

这里将局部变量和全局变量分开说是因为全局变量因为 Cache 的使用引入了一个新问题，RAM 和与之对应的 Cache 存在内容不一致的可能，当程序没有抛弃当前的 Cache 映射关系时，程序所修改的变量实际上只修改 Cache 中的内容，RAM 里面的内容保持不变，两者不相一致。这个不一致会不会对应用产生不良影响呢？答案是肯定的。

一种情况就会导致错误，当 MCU 的硬件可以不通过 CPU 与 RAM 交换数据时，错误就会产生，象我们通过 DMA 将 RAM 指定位置的内容传送到其它位置或者外围接口时，DMA 取的数据是 RAM 中的内容，而程序更改的是 Cache 里面的内容，这样 DMA 传送的数据并不是程序最新的结果。用例子解释会更清楚一些，程序将某个全局变量从 0 开始往上累加，DMA 则将这个变量传递给 UART 输出到电脑显示，这里程序累加的是 Cache 里面的值，RAM 中始终会保持 0，导致结果是电脑收到的始终是 0，直到抛弃当前的 Cache 映射关系才改为输出新的值。

这是 MCU 中的某种设备由总线绕过 CPU 操作直接读 RAM 而产生的数据不一致错误，另外一种情况刚好相反，是设备使用总线绕过 CPU 操作直接写 RAM 而导致程序处理的数据不是最新数据，将上面的情况反过来从 PC 向 MCU 发送数据，UART 收到的数据通过 DMA 直接存入 RAM，MCU 显示自己接收到的数据不能保证为最新状态。

注：后面关于问题调试与分析的章节中有与此相关的实例。

Cache 的分级

目前处理器发展趋势是 CPU 主频越做越快，系统架构越做越先进，但主存 RAM 的结构和存取时间改进相对偏慢。因此如何将 CPU 的高速特性展现出来，Cache 技术就成为不二选择，但芯片面积和成本等因素的限制不能满足 Cache 做得足够大的愿望，所以 Cache 的设计提出了分级的概念。

微处理器性能由如下几种因素估算：

性能=k(f×1/CPI－(1－H)×N)

式中 k 为比例常数，f 为工作频率，CPI 为执行每条指令需要的周期数，H 为 Cache 的命中率，N 为存储周期数。

要想提高处理器的性能，就应该提高工作频率，减少执行每条指令需要的周期数，提高 Cache 的命中率。减少 CPI 值可通过同时分发多条指令和采用乱序控制的方法实现，采用转移预测和增加 Cache 容量则可以提高 H 值，减少存储周期数 N 通常是采用高速总线接口和不分块 Cache 技术。

以前为了提高处理器的性能，主要采用提高工作频率和指令并行度这类直接方法，开始效果是非常明显，但随着改进的深入瓶颈也随之出现，也就是说靠提高工作频率和指令并行度对效率的提升效果不再明显，于是改进方向转向了提高 Cache 的命中率，正是这样的背景下设计出无阻塞 Cache 分级结构。

Cache 分级结构的主要优势在于，对于一个典型的一级缓存系统的 80％的内存申请都发生在 CPU 内部，只有 20％的内存申请是与外部内存打交道。而这 20％的外部内存申请中的 80％又与二级缓存打交道。因此，只有 4％的内存申请定向到 RAM 中。Cache 分级结构的不足在于高速缓存组数目受限，需要占用线路板空间和一些支持逻辑电路，会使成本增加所以目前采用 Cache 分级结构的 MCU 还比较少。

I-Cache 和 D-Cache

从数据集中性的分析中我们知道虽然 CPU 的指令和数据都满足集中性规则，但指令比数据会更符合这一规则，所以在集中性方面指令和数据虽然符合统一基本规则，但各自又都具有相对独立的特征。

基于这个规律有些芯片公司将 Cache 设计成 I-Cache（指令 Cache）与 D-Cache（数据 Cache）两种。这种双路高速缓存结构减少了争用高速缓存所造成的冲突，改进了处理器效能，可以让数据访问和指令调用在同一时钟周期内进行。另外对于程序通常数据和指令在内存中的位置是以数据或者指令的方式归类成块分步，采用 I-Cache 和 D-Cache 可以减少因为数据和指令的位置不同而导致的 Cache 更新操作。

```
if(flag==1)
{
  func();
}
flag=0;
```

这里读写 flag 是在数据段中进行，函数 func()和 flag=0 的指令是在代码段中，采用 I-Cache 和 D-Cache 分离方式就不会出现数据段和代码段间隔大而产生的 Cache 更新操作，对提高 Cache 效率有着明显的作用。

PC 的 Cache 技术

PC 中 Cache 的发展是以 80386 为界。

现在计算机系统中都采用高速 DRAM（动态 RAM）芯片作为主存储器。早期的 CPU 速度比较慢，CPU 与内存间的数据交换过程中，CPU 不需要进行额外的等待，以早期的 8MHz 的 286 为例，其时钟周期为 125ns，而 DRAM 的存取时间一般为 60~100ns，因此 CPU 与主存交换数据无须等待。这种情

况称为零等待状态，所以CPU与内存直接打交道是完全不影响速度的。

近年来CPU的时钟频率的发展速度远远超过DRAM，几年内CPU的时钟周期从100ns加速到几个ns，而DRAM经历了FPM、EDO、SDRAM几个发展阶段，速度只不过从几十ns提高到10ns左右，DRAM和CPU之间的速度差，使得CPU在存储器读写总线周期中必须插入等待周期；由于CPU与内存的频繁交换数据，这极大地影响了整个系统的性能，这使得存储器的存取速度已成为整个系统的瓶颈。

当然采用高速的静态RAM（SRAM）可以作为主存储器与CPU速度匹配，问题是SRAM结构复杂，不仅体积大而且价格昂贵。因此除了大力加快DRAM的存取速度之外，当前解决这个问题的最佳方案是采用Cache技术。Cache即高速缓冲存储器，它是位于CPU和DRAM主存之间的规模小的速度快的存储器，通常由SRAM组成。

Cache的工作原理是保存CPU最常用数据，当Cache中保存着CPU要读写的数据时，CPU直接访问Cache。由于Cache的速度与CPU相当，CPU就能在零等待状态下迅速地实现数据存取，只有在Cache中不含有CPU所需的数据时CPU才去访问主存。Cache在CPU的读取期间依照优化命中原则淘汰和更新数据，可以把Cache看成是主存与CPU之间的缓冲适配器，借助于Cache，可以高效地完成DRAM内存和CPU之间的速度匹配。

386以前的芯片一般都没有Cache，对后来的486以及奔腾级甚至更高级芯片，已把Cache集成到芯片内部，称为片内Cache。片内Cache的容量相对较小，可以存储CPU最常用的指令和数据。别看容量小,片内Cache灵活方便,对系统效率有相当的提高,如果在BIOS中关掉CPU的内部Cache，会让系统性能下降一半甚至更多。

但是片内Cache容量有限，在CPU内集成大量的SRAM会极大的降低CPU的成品率，增加CPU的成本。在这种情况下，采取的措施是在CPU芯片片内Cache与DRAM间再加Cache，称为片外二级Cache(Secondary Cache)。片外二级Cache实际上是CPU与主存之间的真正缓冲。由于主板DRAM的响应时间远低于CPU的速度，如果没有片外二级Cache，就不可能达到CPU的理想速度。片外二级Cache的容量通常比片内Cache大一个数量级以上。

主板上的片外Cache工作在CPU的外频下，与CPU主频速度通常相差几倍。为了进一步提高系统性能，在CPU片内Cache和主板Cache之间加入真正的二级Cache，这就是片内二级Cache。它通常以CPU主频的半速或全速工作，容量一般为128K～512K，新的至强处理器则达到2M以上。

全速的二级Cache可以极大地加速大型密集性程序的运行速度,带有同速的Cache的Pentium II至强、 Pentium Pro系列处理器是大型服务器的首选CPU。但集成高密度的二级Cache同样会加大CPU的成本，所以这一类的处理器都是价格昂贵的产品。去掉二级Cache的处理器性能虽然有不少下降，但价格可以降得很多，市场上的赛扬处理器就是一个很好的例子。

3.2. 总线

总线的专业解释是一种描述电子信号传输线路的结构形式，是一类信号线的集合，是子系统间

传输信息的公共通道。通过总线能使整个系统内各部件之间的信息进行传输、交换、共享和逻辑控制等功能。在单片机系统中，它是 CPU、内存、输入、输出设备传递信息的公用通道，MCU 的各个模块通过内部总线相连接，外部设备则通过相应的接口电路再与总线相连。

总线英文叫作“BUS”，对应中文意思为“公交车”，这是一个形象的比喻。为了更形象你可以将总线理解成一座独木桥，和常见的独木桥不同的是这个独木桥中间有许多分支，每条分支都连接一座房子，房子里面可能只住一个人，也可能住着许多人。如果一个人想从一座房子到另外一座房子里面去，就要出来经过独木桥才可以到达，但独木桥同一时刻只能走一个人，一个人出来之前就应该先看看桥上有没有人在走，没有人才可以出发。

总线分类的方式有很多，下面是几种最常见的分类方法。

按连接方式分：

按连接方式可分为内部总线和外部总线。内部总线和外部总线在功能上可以完全相同，没有本质的区别，只是在速度、抗干扰能力等指标性能方面会存在不同。内部总线在芯片内部连接各功能模块，因为总线不出芯片，所以可以工作到相对更高的速度，外部总线从通过引脚从芯片内部引出来，加上接口驱动电路后就可以连写接口功能一致的外部设备，因为外部引线会受到物理电气特性的限制，最高速度相对会慢一些。

内部总线的最高速度不是一定会高过外部总线，有一些芯片内部的工作主频比较低，这时外围的接口驱动电路最高限制速度就有可能高过总线可设置到的最高速度，此时外部总线的最高速度可以和内部总线相同。

单片机因为 CPU 只是芯片的一部分，所以除了在内部有内部总线连接各个功能模块，也会有不少单片机支持外部总线方式，这样可以让用户自行决定是否使用某些可选功能模块，外部总线加上接口驱动电路就可以和同样功能其它设备进行通讯。PC 的总线主要以外部总线方式在主板上体现，这是由 PC 的构架决定的，CPU 主要是完成运算处理功能，需要通过总线和外部交换数据。

按功能分：

最常见的是从功能上分为地址总线（address bus）、数据总线（data bus）和控制总线（control bus）。在有的系统中，数据总线和地址总线可以在地址锁存器控制下被共享，采用的是地址和数据复用方式，这种方式在一些简单的 MCU 中较为常见。

地址总线顾名思义是用来传送地址的，实际应用中最常见的是 CPU 地址总线来选用存储器的存储地址。地址总线的位数往往决定了存储器存储空间的大小，所支持的最大空间大小为 2 的总线位数次幂，象 8 位/16 位/32 位的地址总线对应的其最大可存储空间为 256/64k/4G Bytes。地址总线越宽芯片设计需要的体积越大，为了降低实现成本，一些简单的 MCU 会采用 8 位地址总线，但 8 位地址总线所支持的寻址空间只有 256Bytes，这些 MCU 会采用 PAGE/BANK 之类的技术来增大芯片的寻址空间（可参阅章节“单片机 PAGE/BANK 概念”）。

数据总线用于传送数据信息，它又有单向传输和双向传输数据总线之分，双向传输数据总线通常采用双向三态形式的总线。数据总线的位数通常与 MCU 的字长一致，8 位的 MCU 字长 8 位，其数据总线宽度也是 8 位。在实际工作中，数据总线上传送的并不一定是完全意义上的数据，象 CPU 取指令操作就是先将地址总线设为指令所在地址，然后通过数据总线将具体指令取回，此时数据总线传送的是指令而不是数据。

控制总线用于传送控制信号和时序信号。象 MCU 对外部存储器 SDRAM 进行读写操作就要先通过控制总线发出读/写信号、片选信号和读入中断响应信号等。控制总线一般是双向的，其传送方向由具体控制信号而定，其位数也要根据系统的实际控制需要而定。

另外也有分成系统总线和非系统总线的分法，不过实际应用中很少使用这种分法。

按传输方式分：

按照数据传输的方式划分，总线可以被分为串行总线和并行总线。理论上并行传输方式要优于串行传输方式，传输速率会远远高过串行方式，但其成本上会有所增加。这一点在以外部接口方式出现时更为明显，因为外部接口为了保证连接的可靠，就需要采用性能良好的接头，在接头插槽接触点位置，需要镀金等技术保证接触良好。另外并行方式接口过多的连接点会导致连接的可靠性下降，因为接头的每一个接触点出故障的几率是等同的，随着触点增多可靠性自然就会低下来。

常见的串行总线有 SPI、I2C、USB、UART、CAN、SIO 等；并行总线则有 PCI、LPT、CSI、TFT、IDE 等。

按时钟方式分：

按照时钟信号是否独立，可以分为同步总线和异步总线。同步总线的时钟信号独立于数据，也就是说要用一根单独的线来作为时钟信号线；而异步总线没有独立的时钟信号，通常是在信号中约定一个同步触发信号，这个同步触发信号被当做时间基点，然后总线上的各个模块用自己内部的时钟信号得出总线控制时序的时间轴。

因为同步总线有专门的时钟信号线，所以同步总线进行通讯时每一步都由该时钟信号线进行同步，所以同步总线的通讯时序上任意时刻各模块间的同步性是一致的。异步总线由于没有时钟信号线，各个模块的内部时钟不可能做到绝对一致，相互之间会存在误差，这个误差会在控制时序的时间轴上进行累加，累加到一定程度就有可能出错，所以异步方式每隔一段时间都需要重新同步。

总线技术指标：

评价总线的主要技术指标是总线的带宽（即传输速率）、数据位的宽度（位宽）、工作频率和传输数据的可靠性、稳定性等。

总线的带宽指的是单位时间内总线上传送的数据量，即每秒可以传送最大数据传输率。总线的位宽指的是总线能同时传送的二进制数据的位数，或数据总线的位数，即 16 位、32 位等总线宽度的概念；总线的位宽越宽，数据传输速率越大，总线的带宽就越宽。总线的工作时钟频率以 MHz 为

单位，它与传输的介质、信号的幅度大小和传输距离有关。在同样硬件条件下，我们采用差分信号传输时的频率常常会比单边信号高得多，这是因为差分信号的的幅度只有单边信号的一半而已。

MCU 有最高工作频率限制，所以对于任何 MCU 其内部总线的带宽也是有限度的，有的 MCU 内部可能包含许多功能模块，大多数时候只会用到一部分功能模块，为了降低芯片成本芯片厂商在设计总线的时候会尽量降低最高工作频率，这种情况如果将所有模块都设置到极限工作状态，就会出现总线速率跟不上的情况，从外部看是 MCU 的整体性能急剧下降。

总线信号复用方式：

依据前面对总线的定义可知总线的基本作用就是用来传输信号，为了各模块的信息能及时有效的被传送，就需要避免各模块彼此间的信号相互干扰和物理空间上过于拥挤，解决此问题最好的办法是采用多路复用技术。所谓多路复用就是指多个用户共享公用信道的一种机制，目前最常见的主要有时分复用、频分复用和码分复用等。

时分复用（TDMA）是将信道按时间加以分割成多个时间段，不同来源的信号会要求在不同的时间段内得到响应，彼此信号的传输时间在时间坐标轴上是不会重叠。

频分复用（FDMA）就是把信道的可用频带划分成若干互不交叠的频段，每路信号经过频率调制后的频谱占用其中的一个频段，以此来实现多路不同频率的信号在同一信道中传输。而当接收端接收到信号后将采用适当的带通滤波器和频率解调器等来恢复原来的信号。

码分复用（CDMA）是所被传输的信号都会有各自特定的标识码或地址码，接收端将会根据不同的标识码或地址码来区分公共信道上的传输信息，只有标识码或地址码完全一致的情况下传输信息才会被接收。

总线并不是一项陌生的技术，单片机诞生的那一天它就一直存在于单片机之中，只不过在很长的一段时间内单片机都只采用总线技术最简单的部分，这一阶段对工程师来说只要知道总线的存在就行，完全不需要了解其中的技术细节。直到 32bits 的 MCU 成为市场的一大支柱，总线技术从幕后跃居台前，实现技术也从简单变得复杂，编写底层驱动程序的时候也逐渐需要工程师去了解总线的工作方式，甚至是设定总线的控制模式。

接下看一看一个 16 位 MCU 和另外一个 32 位 MCU 的总线连接示意图，16 位 MCU 单片机总线实现可以说最为简单，而 32 位 MCU 则相对要复杂一些。

16 位 MCU 总线示例：

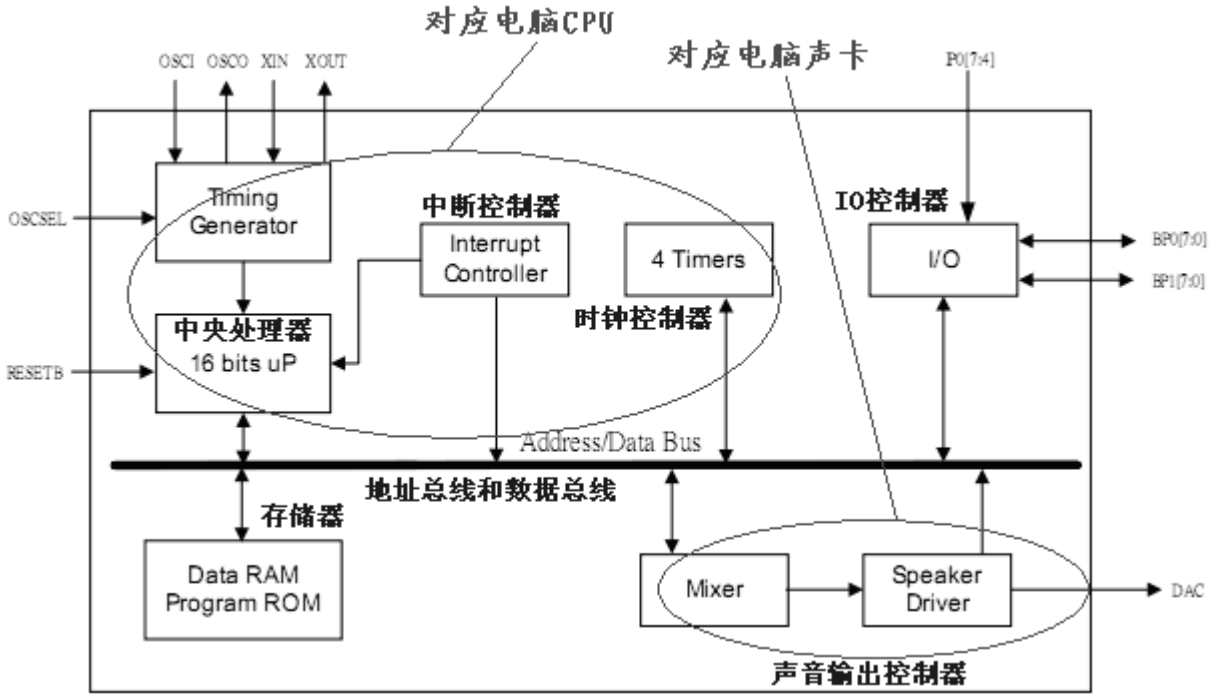

图 3.2.-1　16 位 MCU 样例总线示意图

该 16 位 MCU 所用的总线比较简单，中央处理器通（CPU）过总线与存储器（RAM 和 ROM）、中断控制器、时钟控制器、IO 控制器和声音输出控制器（SPU）相连，其中中央处理器和声音输出控制器可以独立访问存储器 RAM 和 ROM，中断控制器、时钟控制器、IO 控制器只能由中央处理器进行控制访问。

不同模块间的访问通过总线实现，存储器、中断控制器、时钟控制器、IO 控制器和声音输出控制器各自都有自己的空间地址段，不同模块间的地址段相互独立且不重复，图中的 MCU 提供 24 位地址总线，所以其支持范围为 0x000000~0xFFFFFF 的 16MBytes 寻址空间，显然该 MCU 的应用程序不会有 16Mbytes 这么大，之所以支持这么大的范围是为了给声音输出控制器提供足够的声音数据。

来看一下 MCU 的地址空间映射关系：

0x000000　0x007FFF　内部 32kBytes 的 RAM 空间

0x008000　0x008FFF　系统配制寄存器空间

0x009000　0x009FFF　中断控制寄存器空间

0x00A000　0x00AFFF　时钟控制寄存器空间

0x00B000　0x00BFFF　IO 控制寄存器空间

…

0x010000　0x1FFFFF　内部 ROM 空间（接近 2MBytes）

0x200000　0xFFFFFF　外部可扩展存储器空间（ROM 或 RAM）

当中央处理器执行程序时，先通过地址总线设定相应地址从存储器中得到指令和数据，然后执行相应指令，如果指令的操作对象是其它控制寄存器时，基本流程和存储器中读写数据过程相同，

只是需要设定不同的地址。

如果指令和数据共用同一套地址总线，取指令和操作数需要对地址总线设定相应地址，如果操作对象为 RAM 变量或寄存器则还需要对地址总线设定另外的地址以读取或存储相应数据，这样一条指令就有可能需要对地址总线进行多次设定，中央处理器每做一步操作至少需要一个系统时钟周期，这样就难以得到高的代码执行效率。针对这种问题有的 MCU 地址总线有两套，一套专门用于取指令操作，另外一套则用于数据读写，这样做理论上可以让代码执行效率提高将近一倍，但内部结构和成本会略有上升。

声音输出控制器也能独立从存储器空间读取数据，具体方法是在声音输出控制器提供一系列的寄存器供中央处理器进行设置，这些寄存器包含有声音输出开始和结束地址等信息，当程序设置好这些寄存器后声音输出控制器就开始自主工作，按规定的间隔从存储器读取数据并通过 DAC 输出，这样就不需要中央处理器再进行干预就能将指定的声音输出。

可是中央处理器会占用总线，那声音输出控制器如何通过总线得到声音数据呢？答案很简单，总线并不是时刻都被中央处理器占用，如果不是 RISC 结构的 MCU，至少在指令被执行的那一个系统时钟周期内中央处理器是不会占用总线的，这样在任意一条指令执行过程中都至少存在一个系统时钟周期中央处理器是不会占用总线，声音输出控制器就可以在这个时间间隙内使用总线完成数据读取。

那如果 MCU 是 RISC 结构怎么办？这个问题就涉及到了总线管理一些更复杂的技术，这里我们先不进行讨论，留到后面 32 位 MCU 例子中再做解释。这个例子的总线控制方法相对比较简单，能够主动申请总线操作的只有中央处理器和声音输出控制器，所以只是类似时分复用的简单方法就实现了总线的管理。

32 位 MCU 总线示例：

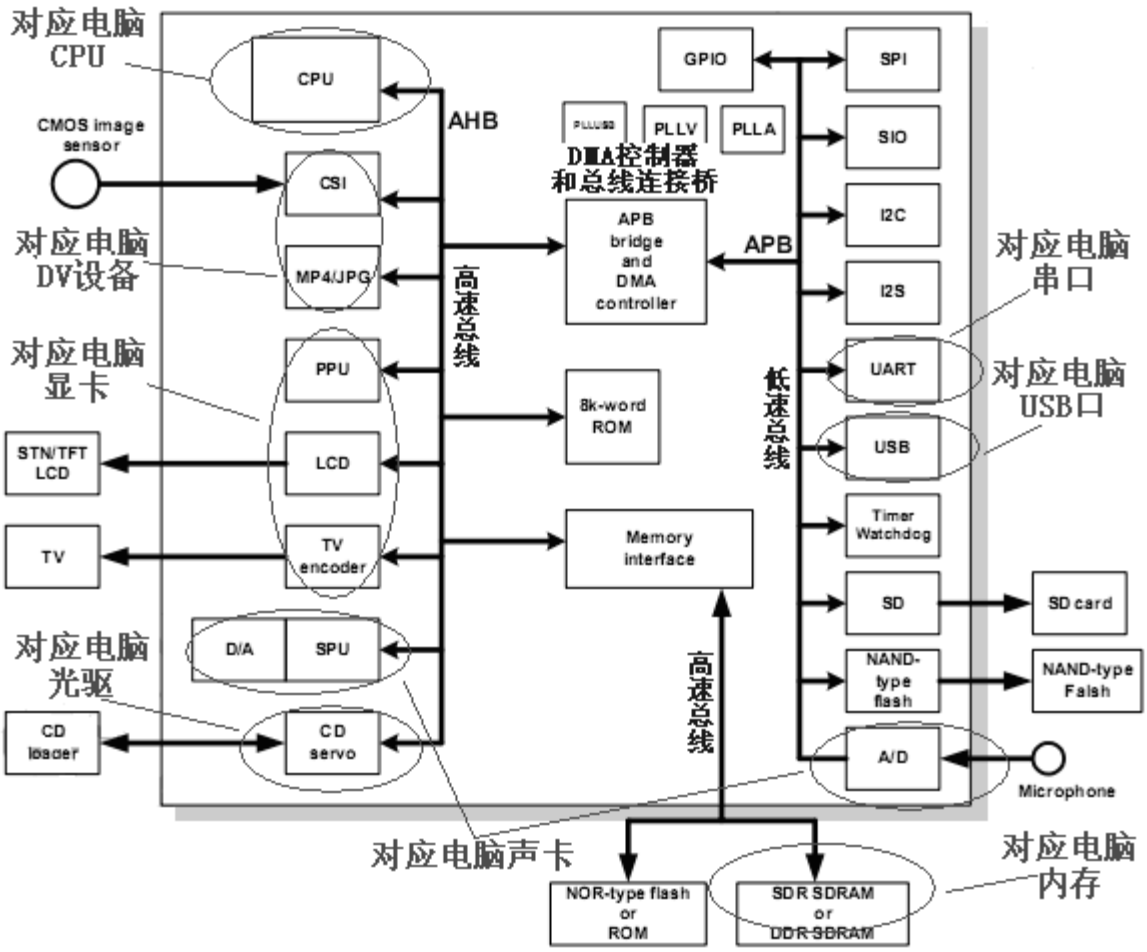

图 3.2.-2　32 位 MCU 样例总线示意图

与 16 位的 MCU 相比该 32 位 MCU 内部的总线连接明显要复杂，首先是总线有了高速和低速之分，其次多了总线控制器（总线连接桥），另外还多了和总线控制器融合在一起的 DMA 控制器。DMA 控制器我们在后面一节中会详细讲述，这里只要知道 DMA 可以独立于 CPU 之外自主完成数据传输功能。

前面 16 位 MCU 只有中央处理器和声音输出控制器会主动去申请总线传送数据，这里 32 位的 MCU 会申请总线操作的内部模块远多于 16 位 MCU 的两种。下列操作会主动申请总线，为简化说明这里将存储器全当做 RAM，实际上所有读 RAM 操作对于读 ROM 一样可行。

●CPU 总线双向读写 RAM 操作，完成 CPU 取指令、读写数据等操作。

●CSI 总线单向写 RAM 操作，将摄像头的数据自动填入相应的 RAM buffer。

●PPU 总线双向读写 RAM 操作，将各个虚拟屏的数据自动读出合成到输出 RAM buffer。

●LCD 总线单向读 RAM 操作，自动读取 PPU 输出的数据并通过 LCD 接口输出液晶屏。

●TVE 总线单向读 RAM 操作，自动读取 PPU 输出的数据并通过 TV 接口输出到电视

●SPU 总线单向读 RAM 操作，自动读取 RAM 中的声音数据并通过 Audio 接口输出。

●DMA 申请总线 RAM 与 RAM 间传送操作，自动完成 RAM 到 RAM 间的数据块传送。

●DMA 申请总线 RAM 与接口模块间传送操作，可以自动完成 SPI、SIO、I2C、I2S、UART、USB、SD、NAND 这些接口模块的输入输出操作，另外可以自动完成 CD、ADC 接口模块的输入操作。

既然可以主动申请总线操作的模块源远大于两种，继续采用类似 16 位 MCU 的总线管理方法肯定行不通，所以在这个 MCU 中出现了用于总线管理的 DMA 控制器和总线连接桥。需要留意的是 DMA 控制器并不只是单单管理前面所列出的 DMA 申请类别，对于 CSI、PPU、LCD、TVE、SPU 这几个模块它们内部隐藏有自己专用的 DMA 通道，不可以被其它模块使用，这些私有 DMA 通道同样也需要经过这个 DMA 控制器进行管理。后面两种 DMA 申请是几个公用 DMA 通道供所列出的模块共享，当某个模块需要使用 DMA 时，需要从这公用 DMA 通道中申请一个空闲通道，如果没有空闲通道则申请失败，然后等待别的模块释放通道或者停止一个正在使用的通道供其使用。

象 CSI、PPU、LCD、TVE、SPU 这几个模块需要传送的数据量都相当大，而 CPU 也是随时需要使用总线，这样相互间同时申请总线的几率自然就会高，于是总线申请冲突产生，从原理上将这种冲突是无法避免的，总线控制器就是用于总线调度管理，以消除这些冲突。通常总线控制器不允许某个模块独自长时间占用总线，会轮询产生了总线申请的各个模块，然后让这些模块分时使用总线，如果模块需要连续传送数据，总线控制器会将这一过程分割成许多小段，每段传送一小块数据，这样就可以让所有产生总线申请的模块都通过总线完成数据传送。

即便是总线控制器采用轮巡方法让各个模块分时公享总线，也还存在问题，如果是高速运行的程序在时间上偶然产生一个细小的延迟，用户可能很难察觉到，但对于 PPU、SPU 输出的图像和声音可能就不一样，也许中间数据传送的一个小小延迟会让图像和声音出现噪点和杂音。这样就要求对于这些模块同时对总线产生的申请应该由总线控制器制定出优先次序，同时产生的申请先响应优先级高的模块，将冲突导致的时间延迟尽量加到用户不容易察觉的申请类型中。

通常 MCU 是依据各模块对数据实时性依赖的程序在内部建立各模块对总线申请的优先级表，少数功能强大的 MCU 会允许程序员对优先级进行配置，对于这类 MCU 需要程序员对模块功能和总线管理非常熟悉，一般建议采用 MCU 的默认模式。

很显然，分时轮巡加优先级表的方法解决了 16 位 MCU 例中简单的总线控制方法对于 RISC 结构 MCU 不适用的矛盾，当然实际中的总线管理并不是我所说的那么简单，还需要许多复杂的技术才能实现，比如当延迟产生后如何让申请总线的模块等待延迟结束、如何让总线传输和模块的操作同步等等，这里就不一一细述。

再来看一下总线带宽的极限情况，假定该 32 位 MCU 最高工作频率 100MHz，总线位宽为 32 位，其 PPU 支持 4 层虚拟屏，如果我们将这 4 层虚拟屏全部打开、屏幕为 VGA（640*480）、16bits 颜色、同时输出到 LCD 和 TV、打开摄像头（VGA 模式），看看每秒输出 30 帧的时候显示功能会占用总线多少资源。

每层虚拟屏需要 640*480*2=614400Bytes。

4 层虚拟屏、LCD/TV 输出和摄像头输入共需要 614400*7=4300800Bytes。

每秒 30 帧则需要 4300800*30=129024000Bytes（约为 129MBytes）。

MCU 最高工作频率 100MHz，总线位宽为 32 位，带宽=最高工作频率*总线位宽/8=400MBytes。但是实际应用中虽然 MCU 内核为 32 位，为降低成本往往外部只接一片 SDRAM，此时总线访问外部 SDRAM 实际上是 16 位模式，所以带宽还需要除以 2 为 400MBytes/2=200MBytes。另外总线控制器在不同模块间进行管理切换需要时间，需要通过总线向 SDRAM 发送一些控制指令等，所以实际上有效带宽大概只有 200MBytes*70%=140MBytes。

这个数值和 129MBytes 已经相差不大，也就是说此时显示功能几乎耗尽了 MCU 的总线带宽，已经难以为程序提供正常运行所需带宽。实际测试结果也验证了这一分析结果：当 MCU 工作在前面状态下摄像头的图像在屏幕上回显会出现许多小彗星一样的飞点，如果减少一个虚拟屏或者将 LCD/TV 一个关闭输出，飞点消失。飞点消失后将主频减半，飞点重新出现，而且更加严重。

总线的通信协议

总线除了从电气层面定义接口方式外，通信协议也是非常重要的，任何一种总线都会有通信协议与之对应，相关文档白皮书则会从最基本的模型开始介绍总线的实现方式。就单片机应用来说，了解总线协议并不需要过多探究七层协议之类的理论，重要的是了解总线传送数据的时序，知道总线是如何传递数据就完全满足应用需求。

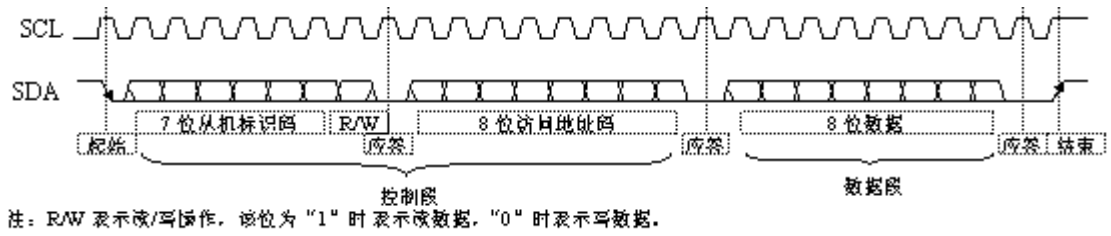

图 3.2.-3　I2C 总线时序图

图示是常见的 I2C 总线上传输的一字节数据的数据帧，其总线形式是由数据线 SDA 和时钟 SCL 构成的双线制串行总线，并联在总线上的各个设备既然可作为发送器（主机）也可作为接收器（从机）。帧数据中除了控制码（包括从机标识码和访问地址码）与数据码外还包括起始信号、结束信号和应答信号。

起始信号：SCL 为高电平时 SDA 由高电平向低电平跳变。

控制码：也叫地址码，用来选择操作目标，只有内部地址码与之相同的从设备才会继续响应后面数据。

数据码：是主机向从机发送的具体的有用的数据。

应答信号：接收方收到 8bits 数据后，向发送方发出低电平做为确认信号。

结束信号：SCL 为高电平时，SDA 由低电平向高电平跳变表示数据帧传输结束。

按协议规定在 SDA 和 SCL 上应该接 4.7k 上拉电阻，这一要求和应答信号和结束信号的定义是一致的。应答信号为低是接收方收到数据后主动将 SDA 拉低，如果没有响应上拉电阻会将 SDA 拉成高。结束信号表示总线操作结束，于是总线会被释放掉，所以定为从底变为高不会与总线释放变为高产生冲突。

3.3. DMA

DMA 是 Direct Memory Access（存储器直接访问）的缩写，它是一种高速的数据传输操作，允许在外部设备和存储器之间直接读写数据，既不通过 CPU，也不需要 CPU 干预。数据传输过程的操作管理通过 DMA 控制器实现，当其进行数据传输时，只需要 CPU 在数据传输开始和结束时对 DMA 控制器进行相关设定，然后 DMA 控制器自行完成指定的数据传输工作，在传输过程中 CPU 可以解放出来进行其它工作。

这样设计实质就是将某些数据搬运工作由硬件完成，不需要消耗 CPU 的软件资源，能这样设计是因为实际应用程序 CPU 大部分时间并不占用总线，在这部分时间段内 MCU 实际上是可以通过总线来间隔传输数据，所以可以将 DMA 控制器与 CPU 对总线的操作设计为并行状态，两者相互交替使用总线，这样就可以将 CPU 没有占用总线的那段空余时间利用起来，使整个系统的效率大为提高。

就这样单纯的说效率可大大提高并没有说服力，相信不少人还是会疑惑效率到底提高在什么地方呢？现在手机大都带有照相功能，也可以摄录一些视频短片，只要手机工作到照相机模式，就会将摄像头的实时画面显示在屏幕上，加入现在你是开发手机的工程师，对于这项功能你会怎样实现？

如果没有 DMA 功能，只能是编写程序从摄像头（CMOS Sensor）将实时画面的图像数据取回来，然后将这些数据通过 LCD 显示出来，图像数据从 CMOS Sensor 搬运到 LCD 的工作需要由程序来完成，假定我们每次搬运一个点的颜色数据，就算是完成 QVGA/30 帧这样的效果也需要一秒传输 320*240*30=2304000 个点。

那完成一个点的数据搬运需要 CPU 做多少事情呢？最少需要下面步骤：

a.依据当前点位置判断是否向 CMOS Sensor 给出行场同步脉冲信号

b.向 CMOS Sensor 给出时钟信号

c.读当前点的颜色数据

d.依据当前点位置判断向 CMOS Sensor 给出行场同步脉冲信号

e.向 CMOS Sensor 给出时钟信号

f.写当前点颜色数据到 LCD

g.更新下一点的位置继续循环

就算每一步平均需要两条指令一个点会耗费 14 条指令，完成实时图像数据的搬运每秒需要执行 2304000*14=32256000 条指令，实际情况比这个数会更多，无疑占用了太多的 CPU 资源。

有了 DMA 情况会截然不同，这每秒 32256000 条用来搬运数据的指令可以全部省掉，这类手机为了支持 CMOS Sensor 和 LCD，芯片会提供相应专用接口，接口可以自动完成同步、时钟信号的处理，同时将输入数据写进指定位置或者从指定位置读出并输出。可能在 MCU 的内部结构图或数据手册上并没有明确这些数据的传送是 DMA 完成，从实质上讲这类操作都可以归纳为 DMA 操作，只是表现形式不同。

现在只需要程序通过 CPU 设定好 CMOS Sensor 和 LCD 的工作参数，好让摄像头和屏幕工作起来，这些参数一定包含有设定数据空间的操作，象 CMOS Sensor 和 LCD 需要设定数据缓冲区的起始地址、图像的宽和高以及图像的颜色深度等信息。有了这些设定，当 CMOS Sensor 开始工作就会由硬件自动将数据填入所设定的数据缓冲区地址，LCD 对应数据缓冲区的数据则会由硬件自动读出并输出给液晶屏，只要两者参数相互适应且数据缓冲区地址相同，CMOS Sensor 的实时画面就可以不受 CPU 干预自动在屏幕显示出来。

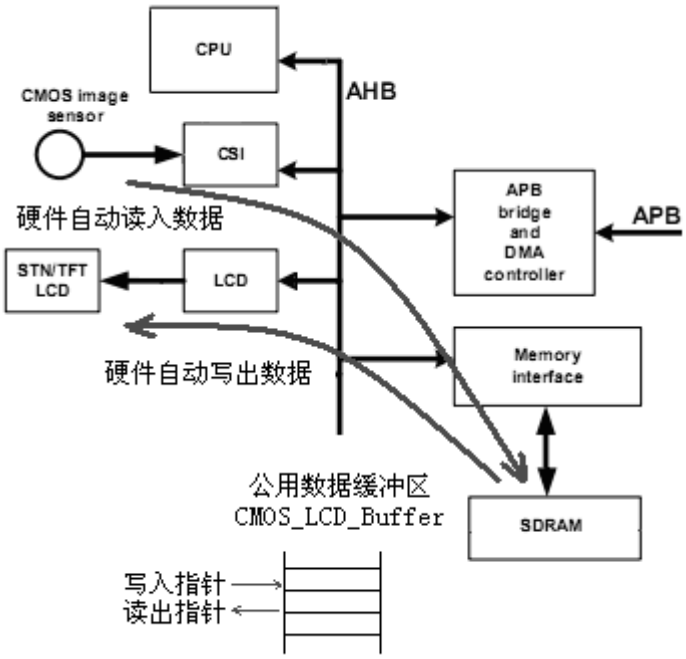

图 3.3.-1　DMA 操作示意图

在讲 Cache 时曾提 DMA 传输数据不经由 CPU，从图示可以看出从 CMOS Sensor 进来的数据直接由 DMA 通过 AHB 高速总线传进 SDRAM，确实不需要经过 CPU 进行传递，LCD 显示也是同样情况，DMA 控制器通过 AHB 高速总线将 SDRAM 中的数据直接传递给 LCD 显示。

那 DMA 是不是可以完成任意方式的数据搬运操作呢？答案是否定的，DMA 控制器不可能设计得非常复杂，基本上都是设定好起始地址和所需搬运数据的长度和方式就可以自动开始进行传输，每完成一次传输硬件会自动将地址递增或递减，这样 DMA 的传输过程实际上就只适用地址连续的数据块传输，间隔传输则无法实现，更不用说想做到随机地址传输。

对于例子中的 CMOS Sensor 和 LCD 的 DMA，只能由这两个模块专用，因此在设定上相当简单，只需要设定传输地址和传输数据大小，不用管传输数据宽度等其它设定。但通用的 DMA 不能设计得过于死板，会允许用户自己设定传输数据位宽、传输模式等，这样同一个通用 DMA 通道可以依据实际情况进行配置，以期得到最佳效果。

比如现在想让 RAM 之间的 DMA 数据传输速度最快，就可以将 DMA 的传输数据位宽和传输数据块长度设最大。如果设置成传输数据位宽 32bits、传输模式为每次传 32 次块传输，那总线每让 DMA 占用一次就可以传送 4*32=128 字节；要是设置成传输数据位宽 8bits、传输模式为每次传 1Bytes 单点传输，总线每让 DMA 占用一次则只能传送 1*1=1 字节。显然前一种设置速度会更快，既然前一种速度要快，而 DMA 的设计目的也是更快传输数据，为什么还可以出现后一种设置呢？

前一种设置速度虽然有优势，但每一次传输实际上传送了 128 字节，这样 DMA 传输的长度必须是 128 的整数倍，否则设定会出错，而后一种可以设置任意传输长度。另外 DMA 的块传输意思是 DMA 占用总线后要完成规定次数传输才释放总线，前一种设置每次 DMA 占用总线时间都是后一种的 32 倍，这样如果有其它模块需要使用总线传输数据前一种设置就需要等更长的时间才有机会申请到总线，要是这个模块对数据的实时性要求很高就会有不良影响。

通常 DMA 会包括这些组成功能：设定传输数据地址、设定传输数据数量、设定 DMA 的控制 / 状态逻辑、DMA 总线请求触发、DMA 数据缓冲、管理 DMA 中断。对 DMA 的设计并没有固定模式，只是要求通过简单设定就能完成地址连续的数据块传输，所以不同的 MCU 设计 DMA 往往会依据自己芯片的应用方向而各具特色。

前面例子中图像输入输出的数据格式没有统一标准，象 LCD 和 CMOS Sensor 有的支持 YUV 格式，有的支持 RGB 格式。YUV 格式显示效果要好， RGB 格式但对于程序员直观，可以知道图像中任意点的颜色，如果程序员希望处理这些数据，YUV 格式需要编写程序转换成 RGB 格式才容易理解。对于 LCD 和 CMOS Sensor，厂家为了降低成本，可能只支持 YUV 和 RGB 格式中的一种，虽然 MCU 大都同时支持 YUV 和 RGB 格式，但如果产品选用的 CMOS Sensor 只支持 YUV 格式而 LCD 只支持 RGB 格式同样存在问题，也需要编写程序转换数据格式才能正确显示。可是软件转换需要逐点转换，转换公式是一个 3*3 的矩阵计算，需要相当多的 CPU 指令才可以完成，所以软件转换的效果不怎么好。

台湾一些芯片设计公司发现了视频图像数据传输处理方面这一特殊需求，于是他们在设计一些芯片 DMA 功能的时候特意加上了相应处理，象针对电视游戏机市场的一些 MCU 他们让 DMA 完成数据传输的同时支持格式转换，因为 YUV 和 RGB 格式转换的公式是恒定的，用程序转换耗费 CPU 资源多是需要将所有的点都按转换公式计算一遍，如果在芯片内部设计有专门的计算电路就可以省去软件

计算矩阵所耗费的时间，这样这些 MCU 在进行 DMA 数据传输的时候可以由程序员选择是否同时由硬件进行数据转换，解决了软件转换效率低下的问题。

武侠影视剧中常看到大侠在空中飞来飞去，对于这种场景喜欢问为什么的人会疑惑，这不是违背了牛顿三大定律么？是否违背牛三定律我不管，影视剧嘛能带给观众良好的视听享受就行，没必要和科学较真，但我们可以了解这些场景到底是怎么拍出来的。飞来飞去大家都清楚，用细钢丝绳将演员吊起来拉来拉去，钢丝绳细，距离远一点摄像机就不会拍出来。但一些在悬崖之类的危险地方飞来飞去难道也是吊几条钢丝绳在悬崖上？肯定不是，要是那样做多危险，大都是让演员先吊在摄影棚里拍，然后拍后面的背景，再把两个场景合成在一起。

如何把两个场景合成在一起好象是一件挺奇妙的事情，娱乐新闻里面常会看到这样的场景，演员吊在一面蓝色（或其它颜色）的背景墙之前做各种动作，但最后出来的影视作品演员就变成了在悬崖、沙漠中，着实让人有点不可思议。其实刚才我所说的那些台湾芯片公司在设计 DMA 附加功能时也能实现场景合成这么奇妙的工作，我们知道，颜色是由三基色构成，任何颜色都可以由 RGB 三原色组合而成，这些芯片可以由程序员定义一种颜色，在该颜色三基色附近的颜色当成透明色，当 DMA 传输的时候就会把这些颜色数据过滤掉。比如 24bits 色为了滤掉蓝色背景，我们就可以定义红色和绿色分量小于 10、蓝色分量大于 245 的颜色为透明色，只要背景的蓝够蓝，用该 DMA 传输出来的数据就能将背景墙过滤掉。

要想用好 DMA，就要熟悉总线的运作方式，DMA 和总线两项技术是相互相成，总线提供可以进行数据快速传递的前提条件，DMA 则是总线实现数据快速传递的具体方式。

3.4. 存储器管理

单片机上电后 PC 指针会自动指向一个固定地址，通常这个地址为 0，然后从该地址开始执行具体程序。简单的单片机的地址分配都很简单，MCU 自身内部所带的 ROM/RAM 会占用固定的地址空间，如果外部可以扩展也一般只支持单片存储器扩展，所扩展的地址空间也都是从特定位置开始，大小不超过外部扩展存储器容量的空间。

当然也可以通过一些特殊方法实现外部多片存储器的扩展，比如用不同 IO 口去选择外部不同存储器的 CS 脚，以实现对扩展接口的地址和数据总线的共享。但这么做需要软件做出相应处理，这些存储器就 MCU 来说是相同的地址空间，类似于一些小单片机为了增大存储空间所采用的 PAGE、BANK 方法。

简单的单片机程序都是在 ROM 中直接运行，程序所需的 ROM 和 RAM 空间一般不会太大，所以这些简单的单片机大都提供出容量一定的存储空间给用户使用，如果空间不够就只能选择另外的单片机进行开发。随着 32 位单片机和嵌入式系统的兴起，不同产品程序所需的 ROM 和 RAM 空间可能相去甚远，小的程序只需要几十 k 的空间就够用，而大的可能需要几十 M 甚至上百 M 才够用，这样 32 位的单片机为了有更好的适用性，改成了自身所带存储器空间非常小，主要由外部存储器扩展的方

式。这样做的好处是有利于降低成本，可以由用户自行决定所需存储器的大小，避免出现成本过高或者空间浪费的情况。

为了支持大容量的存储空间，这些单片机往往可以支持多片外部存储器扩展，和简单单片机不同之处在于这些单片机外部有多条用于存储器扩展的 CS 脚，每一条 CS 脚都映射有一段空间，这样单片机硬件自己就可以通过这些 CS 脚的选择自动访问外部不同的存储器，不需要简单单片机中用程序改变外部存储器的选择。

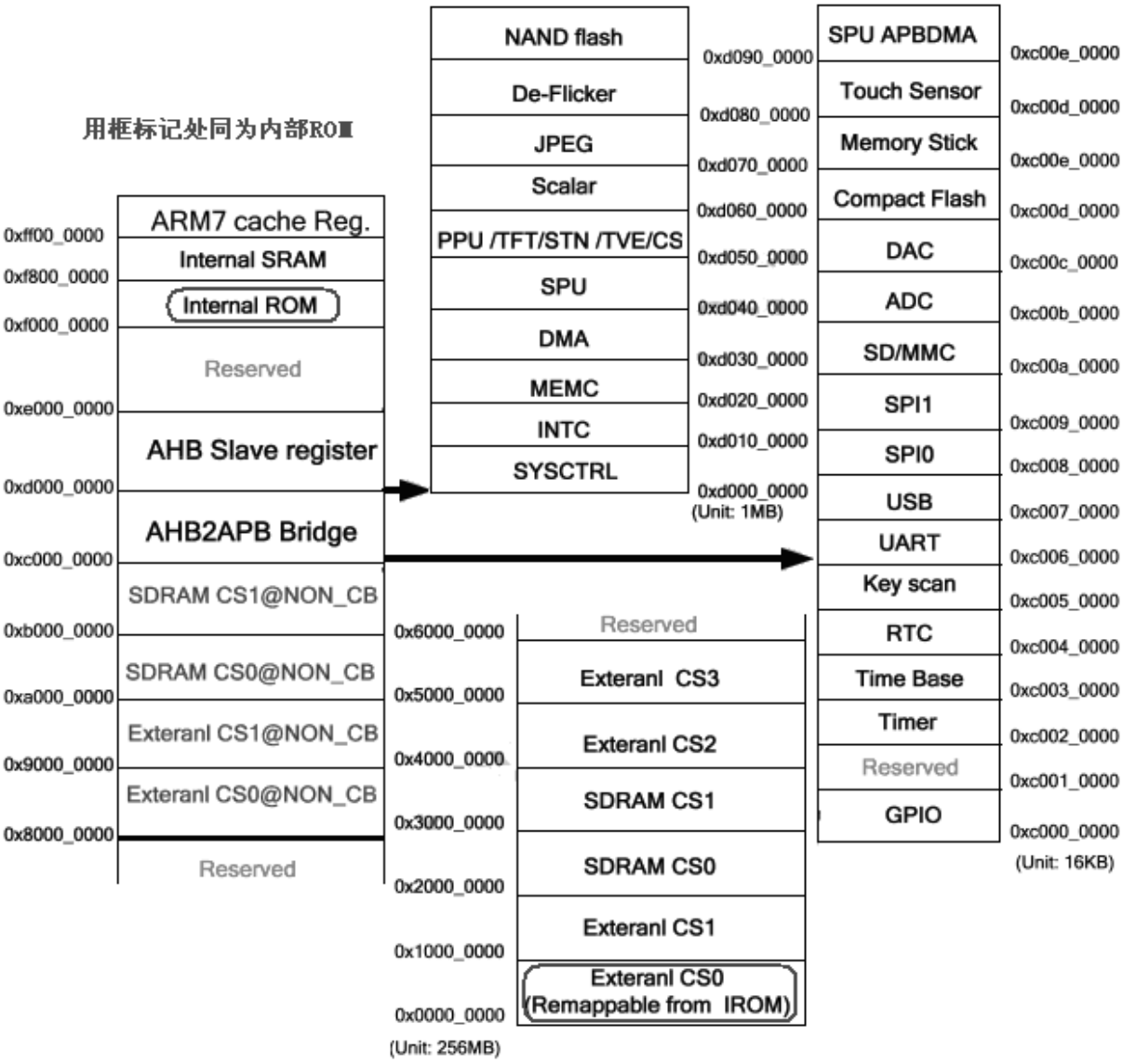

图 3.4.-1 MCU 存储器空间地址表

图示中的 MCU 最多可以扩展 6 片外部存储器，从图看该 MCU 支持的地址总线为 32 位，实际上其地址总线的最高位用来做其它用途，实际上有效地址总线为 31 位，也就是说只有 bit31 不同的两个地址实际上是同一个地址，所以在图中可以看到每条 CS 对应高低两段空间，象 SDRAM CS0 就

对应 0x20000000 和 0xA0000000 两段空间，这两段空间实际上是同一片外部 SDRAM。当程序访问 0x20000000 时，硬件就会自动访问由 SDRAM CS0 选择的外部 SDRAM，不需要程序再做其它任何控制。

对于外部存储器的扩展并没有连接次序或数目的限制，SDRAM 可以从 SDRAM CS0 和 SDRAM CS1 中自由选择组合，只扩展一片可以选 SDRAM CS0，同样也可以选 SDRAM CS1，只需工程师自己知道所对应的地址空间。至于外部扩展存储器的空间大小和总片数也是只要不超过图中的限制就行，象 SDRAM CS0 扩展一片 16MBytes 的 SDRAM，而 SDRAM CS1 却扩展一片 32MBytes 的 SDRAM 都是可以的，只是从 0x20000000 起 16MBytes 空间才有效，从 0x30000000 起是 32MBytes 有效。

既然这类 MCU 可支持的存储器空间急剧增加，空间大小分布又不是必须满足等大小和连续的规则，如果还要求程序员了解系统的存储器空间分配和具体硬件实现后再去写程序显然不太方便，所以最好 MCU 自身能提供出一整套完备的存储器管理的方法，设定好以后由硬件自动完成相关管理，程序员只需知道他可以使用的存储器空间大小就可以编写应用程序。

当然 MCU 不提供硬件对存储器空间管理功能也还是可以编写程序的，来看看这样对程序员编程会造成多大的麻烦。我以图示中 MCU 设计一种极端情况，六个 CS 分别扩展 1/2/4/8/16/32MBytes 的存储器，这种情况得到的实际有效存储器空间如下：

按 1/2/4/8/16/32MBytes 顺序扩展	按 32/16/8/4/2/1MBytes 顺序扩展
0x50000000~0x51FFFFFF	0x50000000~0x500FFFFF
0x40000000~0x40FFFFFF	0x40000000~0x401FFFFF
0x30000000~0x307FFFFF	0x30000000~0x303FFFFF
0x20000000~0x203FFFFF	0x20000000~0x207FFFFF
0x10000000~0x101FFFFF	0x10000000~0x10FFFFFF
0x00000000~0x000FFFFF	0x00000000~0x01FFFFFF

图 3.4.-2　外部存储器扩展表

对于这种有效地址的分布情况，程序员需要时刻防止自己的程序进入到无效空间，一旦外部扩展方式改变，需要重新编写整个程序。MCU 提供的硬件存储器空间管理功能可以让这些烦恼一扫而空，即便是外部扩展方式改变，也只需在系统底层做少量修改，完全可以做到不更改应用程序。

MMU（Memory Management Unit）

MMP 字面意思为存储器管理单元，其主要功能就是解决前面所提出的应用程序和实际地址空间如何协调的问题。先提出两个名词：物理地址和虚拟地址（也可将虚拟地址称为逻辑地址）。物理地址就是在硬件层面看存储器所处的空间位置，物理地址等于访问存储器的地址总线上的地址内容，前面表中的地址均为物理地址。虚拟地址（逻辑地址）则是站在应用程序层面看的存储器的空间位置，它可以等同物理地址，也可以与之不同。

MPU 的做法是将其所能支持的物理地址空间分成许多小块，然后以这些小块为单位将其实际物理地址映射成另外一个地址。比如物理起始地址为 0x00000000 的 64kBytes 空间，MPU 就可以将其映射到起始地址为 0x10000000 或 0x20080000 的位置。CPU 程序需要访问存储器时候，先不将需要访问的地址写入地址总线，而是交给 MPU 的一个转换器，这个转换器先进行虚拟地址到物理地址的逆转换，然后依照逆转换出来的实际物理地址访问存储器。比如 MPU 将物理地址为 0x00000000 的 64kBytes 空间映射成逻辑地址 0x20080000，当 CPU 访问地址 0x20080000 开始的 16kBytes 空间时，实际访问的物理地址并不是 0x2008XXXX，而是 0x0000XXXX。

需要留意的这个地址转换是以某个大小为单位的块为基本单位，MPU 不可能做到将所有的地址单个一一对应，芯片设计人员希望 MPU 的内部转换器越小越好，这样可以节省芯片空间。所以通常块大小都大于 4kBytes，而且允许用户可以自行选择设定块的大小，最终可以将所有的物理地址空间都映射到一个连续的虚拟地址空间。

按 1/2/4/8/16/32MBytes 顺序扩展	虚拟地址（逻辑地址）
0x50000000~0x51FFFFFF	0x01F00000~0x01FFFFFF
0x40000000~0x40FFFFFF	0x00F00000~0x00FFFFFF
0x30000000~0x307FFFFF	0x00700000~0x00EFFFFF
0x20000000~0x203FFFFF	0x00300000~0x006FFFFF
0x10000000~0x101FFFFF	0x00100000~0x002FFFFF
0x00000000~0x000FFFFF	0x00000000~0x000FFFFF

图 3.4.-3　外部存储器虚拟地址映射表

在系统启动的时候先运行固化在系统板上的启动配置程序，完成对转换表的相关设定，启动起来后再装载应用程序就可以按虚拟地址访问存储器，如果系统存储器硬件有改变，也只需更改这部分启动配置代码，完全可以做到应用程序不做任何修改。

Remap

嵌入式系统和通用单片机的程序运行方式不一样,通用单片机的程序大都是在 ROM 中直接执行，虽然有少数应用会将程序放在 RAM 中执行，但都是特殊实现方式，通用单片机的设计思想就是让程序在 ROM 中直接执行。而嵌入式不同，几乎所有的应用程序都是放在 RAM 中执行的，这就存在一个问题，单片机断电后 RAM 里面的内容不会被保存，所以嵌入式系统需要解决断电后可重新运行程序这一问题。

要想保存 RAM 中的内容不丢失，只有在断电后继续保持 RAM 供电，方法无非就是用备用电池，这样做对于第一次上电和取走电池的情况无效，所以用电池的方法行不通。于是 MCU 采用了另外一种方法，在芯片内部自带一小片内部 FALSH，另外还有一小片内部 SRAM，上电时这片内部 ROM 的地址会映射到 0x00000000 位置，看前面的图用方框标示的 IROM 就是它，在地址映射图中有对应到地

址 0x00000000 位置,它自己的真实物理地址是在 0xF0000000。芯片上电后会从固定地址 0x00000000 执行程序，这样就会执行 IROM 中的程序。

不过内部 ROM 中的程序编写存在一些特殊限制，通常这段代码（常被称为 bootloader）是由汇编代码写成，为什么要用汇编是因为如果用 C 无法保证代码中不使用 RAM 变量，而在刚上电的时候物理地址和虚拟地址之间的映射关系还没有建立，只能是通过绝对物理地址来访问存储器。而且外部扩展的存储器不是接上去就可以使用的，还需要对接口进行一些设定才能可靠访问。用汇编则可以保证不使用 RAM 变量，如果非用 RAM 变量不可也可以直接在内部的 SRAM 中按绝对物理地址直接定义。

这段汇编代码需要完成系统时钟、中断等资源的初始化，另外还需要配置好外部扩展存储器的接口参数，到这一步系统可以使用外部扩展的存储器，不过还需要按绝对物理地址进行访问。接下来的工作是将真正的程序装载进来，程序的位置和大小等信息事先需要约定好，汇编代码按照约定将真正的程序装载到 SDRAM 的指定物理地址。为了加快程序的装载速度这里会要求打开 Cache，如果不开速度可能相差数倍。

装载完程序启动代码的工作基本完成，接下来应该开始执行所装载的程序，只需要把程序指针指向所装载程序的入口地址即可。但现在所装载的程序物理地址可能并不等于程序层面所需的虚拟地址，所以跳转前还需要做一个重要操作，就是 Remap，这步操作其实很简单，将某个寄存器的一个控制位置上就可以，当然之前已经设定好 MMU 的映射表，此后 MMU 就开始工作。这里对程序还有一个特殊要求，因为在 Remap 之前程序是以物理地址基准，Remap 之后变成虚拟地址，实际上程序已经转到另外的物理空间去，所以需要保证新的逻辑地址中后一条指令依然相同。这个特殊要求可以这样小技巧实现，启动程序和所装载的程序都按一定格式进行编写，比如按照后面的固定格式定义从地址 0x00000000 开始的代码内容与结构，这样启动程序和所装载的程序开始一段代码的格式是一样的，每个位置对应的都是相同信息，所以 Remap 之后就能得到正确的跳转指令。

```
.org 0x0
ROMbase:
  b  Reset_Handler
  ldr  pc,=mmUndefinedInstructionEntry
  ldr  pc,=mmSWIEntry
  ldr  pc,=mmInstructionAbortEntry
  ldr  pc,=mmDataAbortEntry
  nop
  ldr  pc,=mmInterruptEntry
  ldr  pc,=mmFastInterruptEntry
.org 0x40
__mmUndefinedInstructionEntry:
```

```
  .long  mmUndefinedInstructionEntry
__mmSWIEntry:
  .long  mmSWIEntry
__mmInstructionAbortEntry:
  .long  mmInstructionAbortEntry
__mmDataAbortEntry:
  .long  mmDataAbortEntry
__mmInterruptEntry:
  .long  mmInterruptEntry
__mmFastInterruptEntry:
  .long mmFastInterruptEntry
.globl  Reset_Handler
.type  Reset_Handler, function
Reset_Handler:
  bl  _kmc_asm_init     /*初始化系统*/
  bl  _init_cache_mmu   /*开Cache*/
  bl  _copy_text_data   /*装载程序*/
  bl  _fill_bss         /*初始化变量区*/
  ldr  pc,=start       /*跳转到程序入口*/
```

不是所有MCU 都有内部FLASH，有的MCU 会省掉内部 FLASH，直接用外部扩展的FLASH 来启动，这种设计存在某些FLASH 不可以使用的可能，因为芯片上电后扩展接口工作在默认状态，如果选用的FLASH 不支持该工作设定就可能不能使用，不过基本上所有FLASH 的工作接口方式都相同，所以一般不存在问题。前面图示例子也可以不用内部FLASH 启动，该MCU 有一条管脚选择内部ROM 还是外部ROM，所以图中可以看到IROM 在0x00000000 位置和一组扩展CS 位置重叠，实际应用中为二选一，并不矛盾。

MPU（Memory Protection Unit）

MPU 字面意思为存储器保护单元，嵌入式系统的程序需要一个基础框架来支撑，就好比电脑的Windows 程序需要在Windows 支撑一样，这个基础框架就是操作系统，应用程序是不允许对系统所在的位置进行读写操作的，否则可能导致系统崩溃，这就需要对系统所在空间进行保护，MPU 就是提供这类功能。

保护可以通过软件来实现，但这样做需要对软件的编写提出额外的保护要求，而且所写的软件如果出错保护就无法实现，所以软件保护不是一个方便可靠的方法。MPU 所提供的是硬件保护，系

统由专门的营建来检测和限制系统资源的访问，当任何程序去访问任意地址之前，MPU 会依照所制定的访问权限控制规则检查当前程序是否有权限进行访问。

如果没有相应权限而程序试图进行访问操作，MPU 会产生一个异常信号，该信号会触发处理器执行异常中断处理程序，这就是嵌入式系统常常遇到发生异常中断的一个主要原因，程序试图访问一个它没有权限的地址。当然 MPU 的保护对于此类错误并不能避免或者从错误中恢复，只能告诉调试的程序员当前发生了超越权限的访问，要完全解决错误还需要程序员自己查找出超越权限代码的位置。

来看一个带 MPU 的 MCU 的存储器映射例子，该芯片内部自带 256kBytes 的 RAM，地址范围为 0x00000~0x3FFFF，外部可扩展的的存储器为 0x10000000~0x12000000 的 32MBytes 空间。

该例子具备这些特点：

●系统小于 64kBytes，向量表、异常处理程序位于此空间内，系统软件空间用户模式下的程序不可访问，以免应用程序破坏系统。

●有一个不超过 64kBytes 的共享系统空间，用于存放系统提供的通用库以及用户任务间的数据传递。

●最多支持 3 个独立功能的用户任务，每个任务占用的空间最多 32kBytes，任务间完全独立，不可以相互访问。

功能	访问级别	起始地址	大小	区域
外部扩展空间	系统	0x10000000	32MBytes	4
受保护的系统	系统	0x00000000	4GByte	1
共享的系统	用户	0x00010000	64kBytes	2
用户任务 1	用户	0x00020000	32kBytes	3
用户任务 2	用户	0x00028000	32kBytes	3
用户任务 3	用户	0x00030000	32kBytes	3

图 3.4.-4 MCU 存储器空间权限分配表

区域表示空间处于访问控制权限表中的位置，不同 MPU 映射关系会不同，例子中的 MPU 总共支持 16 级权限控制区域供程序员设定。

管理员	用户	区域编号
不可访问	不可访问	0
读/写	不可访问	1

读/写	只读	2
读/写	读/写	3
不可预知	不可预知	4
只读	不可访问	5
只读	只读	6
不可预知	不可预知	7
不可预知	不可预知	8~15

图 3.4.-5 MCU 存储器访问权限表

系统软件空间只允许管理员进行读写，用户编写的应用程序是不可进行任何读写操作；系统共享空间因为要给应用程序提供库函数，所以应用程序具有读权限，但是不可以写，如果任务间需要传递数据必须调用相应系统函数才能完成；三个用户任务各自拥有一段空间，这段空间对于任务自己和管理员是完全敞开的，可以进行读写操作。外部扩展空间定义成不可预知是现在不知道外部所接存储器的类型，如果是 SDRAM，读写操作都是允许的，如果是 NOR FLASH 显然不可以进行写操作，所以只能定义成不可预知。

如何设定存储器的保护不同 MCU 在数据手册的 MPU 部分会有详细描述，主要是通过设定一系列特殊寄存器来实现。通过这些寄存器可以依照特定的地址、空间大小转换规则建立一张映射表，然后 MPU 依照映射表进行保护。如果用户不需要 MPU 功能，可以选择关闭 MPU 功能。

3.5. 嵌入式与操作系统

单片机发展到今天，嵌入式和操作系统已经成为单片机产品开发的一大主流趋势，掌握嵌入式系统的相关知识已经成为一名电子工程师的必备素质，这一节我们一起来谈谈嵌入式操作系统，嵌入式样操作系统可以说是目前单片机技术在软件方面所发展到的最高层次，不是三言两语就能说清楚的，所以这一节中我只是简单的讲一点我个人对嵌入式系统的理解，如果你想深入了解嵌入式核心，那还得要你自己去多找一些关于嵌入式的资料进行钻研。

什么是嵌入式

什么是嵌入式？看似简单的一个个问题，实际却没几个人能回答得清晰透彻，即便是一些在嵌入式领域有着丰富经验的人也只是知道好像就是那么一种单片机方面的开发应用，到底具体指什么也不能很肯定。我自己实际上也认为嵌入式是一个含糊的概念，它渗透在电子产品开发之中，与传

统常规单片机开发并没有严格明晰的界限。

IEEE（国际电气和电子工程师协会）对嵌入式系统的定义是：Devices Used to Control，Monitor or Assist the Operation of Equipment，Machinery or Plants，意思为用于控制、监视或者辅助操作机器和设备的装置。这个定义相对比较抽象，从字面意思理解所有电子控制设备都属于嵌入式，哪怕就是一个振荡器控制LED闪烁的电路都是，这种理解显然过于广泛。

目前国内普遍认同的嵌入式系统定义为：以应用为中心，以计算机技术为基础，软硬件可裁剪，适应应用系统对功能、可靠性、成本、体积、功耗等严格要求的专用计算机系统。

嵌入式系统（Embedded System），一般由嵌入式微处理器、外围硬件设备、嵌入式操作系统以及用户的应用程序等四个部分组成，用于实现对其他设备的控制、监视或管理等功能。和传统单片机电子产品相比，嵌入式样更能体现功能细分的时代潮流特性。

传统单片机产品是工程师在特定硬件平台下需要完成所有的代码工作，即便是芯片厂商提供有代码样例，也只是适用于与之相适应的硬件，如果更换硬件则需要重新编写代码。不同的厂商芯片代码编写没有统一规程，每家都是按自己的意愿设计样例代码，各种接口的驱动代码也是同样境况。如果想把一个基于东家MCU的产品程序移植到西家MCU上，基本上可以移植的只有程序流程和框架，原来的代码只有参考意义。

传统单片机厂商所为了便于工程师迅速开始编写程序所提供的无功能样例，样例包含有中断、主循环等关键组成单元，工程师只要在上面填写自己的代码就可以正式开始程序编写，可以省掉学习如何搭建程序架构的过程。

传统单片机程序构架样例：

```
#include <HT46R01C.H>
#pragma vector isr_4 @ 0x4
#pragma vector isr_8 @ 0x8
#pragma vector isr_c @ 0xc
//中断服务函数定义，如果需要中断在对应函数内添加相应代码
void isr_4(){} //external ISR
void isr_8(){} //timer/event0
void isr_c(){} //ADC convert
//----------------------------------------
//安全初始化，可以去掉此函数自己初始化系统
//----------------------------------------
void safeguard_init()
{
  _wdts = 0x00;
  _intc0 = 0x00;
```

```
  _tmr0c = 0x00;
  _tmr1c = 0x00;
  _ctrl0 = 0x00;
  _adcr = 0x00;
}
//------------------------------------------
//主程序
//------------------------------------------
void main()
{
  safeguard_init();
  while(1)
  {
    //添加用户代码
  }
}
```

有了这样的例子工程师就可以在此基础上编写自己的代码，即便改变芯片型号只要是同公司同一系列的芯片，也都很容易用照葫芦画瓢的方式建立自己的新项目工程。

嵌入式系统实际上就是起类似样例框架代码作用，是一些专业公司针对某些硬件平台设计出基本程序框架，框架本身不包含实际的具体的功能实现，但要求尽可能的支持硬件平台的所有功能，也就是说其它工程师在此框架基础之上通过接口驱动函数可以使用硬件平台的所有功能。既然嵌入式系统由专业公司提供，所提供的功能自然相当强大，除了支持硬件所具备的各项特性外，还在软件方面做出了许多功能扩展，比如多任务控制管理等。

嵌入式系统要做到支持全部硬件功能，势必需要一定数量的代码方可实现，对于任意一个实际产品可能只是用到硬件平台的部分功能，这样会造成程序空间的浪费，所以嵌入式系统除了功能支持外还需要支持功能的可选择，对于不需要的功能，可以通过规定的方法从系统中将相应代码移除掉，这就是嵌入式系统的可裁剪性。象 UC/OS 裁剪出最小内核只需要 2kBytes 的空间，而如果加上对硬件平台的功能支持、TCP/IP 等通讯协议的实现则需要几百 kBytes，如果是自己设计一个最简支持多任务控制的操作系统模型，一两百 kBytes 都可以实现。

提示：可以将嵌入式系统理解为带操作系统的单片机产品

嵌入式操作系统大都支持多任务，所以将嵌入式的操作系统称为多任务操作系统也是可以的。

如果用电脑来打比方的话，传统的单片机程序就是早期的 MSDOS，单任务，所有的系统资源都可归当前任务所有，嵌入式系统则是现在的 Windows，多任务，系统资源需要经由 Windows 来统一进行调度。

嵌入式系统因为设计思想大体与 Windows 相同，为了便于程序员理解系统以及进行程序开发，所以在驱动程序设计方面也是尽量与电脑方式一致，当程序员针对嵌入式系统进行开发或者移植工作时，会感觉到许多地方程序的规则、结构和风格和电脑非常相似。WinCE 是微软针对便携式电子产品推出的嵌入式操作系统，如果将 WinCE 放在一个有键盘和屏幕显示的电子产品上，会让使用者觉得这完全就是一个装了 Windows 的小电脑，甚至电脑上的程序只要做少量修改用 WinCE 的开发工具编译后就能直接在上面运行，有兴趣的朋友可以用采用 WinCE 的多普达等智能手机看看。

嵌入式系统为了便于扩展和平台移植，通常都是使用分层的方法设计系统，图示就是 eCos 的大体构架图，该图在 eCos 官方示意图的基础根据实际情况做出了一点修改，官方示意图左边的“设备驱动程序”并不与“目标硬件平台”相邻，这里改成了同时与“目标硬件平台”和“硬件抽象层”相邻。（相邻表示相互之间可以进行通讯）

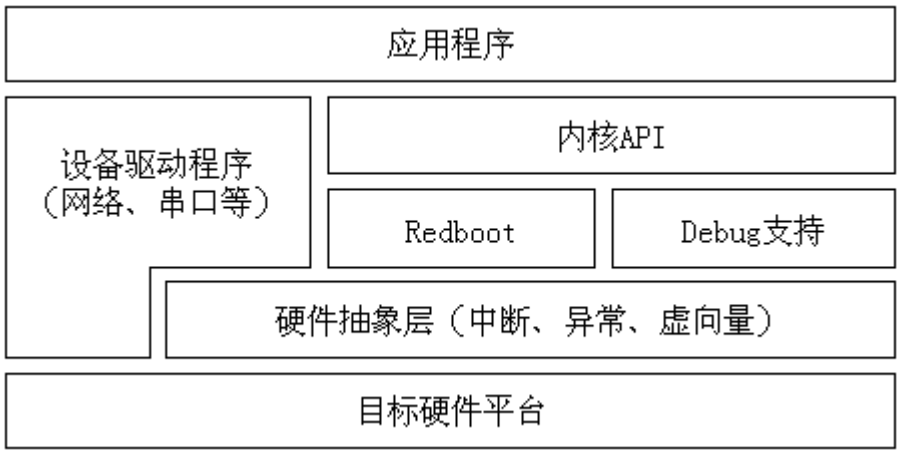

图 3.5.-1　eCos 操作系统构架示意图

如果严格遵循嵌入式操作系统的要求，应用程序应尽量避免直接操作硬件，而是要求按照统一规则编写驱动函数。来看一下 eCos 所提供的标准驱动函数是什么样子。

cyg_io_lookup(const char *name, cyg_io_handle_t *handle)

lookup 函数用来在设备表中查找 name 参数指定的设备，并在参数 handle 中返回句柄指针，如果没有找到指定设备，函数返回设备未找到的错误信息。

name 通常为“/dev/serial0”这种形式。

cyg_io_write(cyg_io_handle_t handle,void *buf,cyg_uint32 *len)

cyg_io_read(cyg_io_handle_t handle,void *buf,cyg_uint32 *len)

write 和 read 函数是对句柄所对应的设备进行写或读操作，buf 和 len 分别是数据缓冲区和数据长度，如果操作失败函数返回相应错误信息。

cyg_io_set_config(cyg_io_handle_t handle,cyg_uint32 key,void *buf,cyg_uint32 *len)

cyg_io_get_config(cyg_io_handle_t handle,cyg_uint32 key,void *buf,cyg_uint32 *len)

set 和 get config 函数对句柄对应的设备进行配置，其中 key 表示配置的具体类型，另外两个指针参数存放配置的具体参数。

如果对 VC 熟悉就会发现这几个函数和 VC 里面的 CreateFile()、ReadFile()和 WriteFile()去控制串口等外设的方式非常相似，都是通过设备名得到一个句柄，然后对句柄进行相应读写操作。

假定现在我们用 eCos 的系统函数控制硬件串口，在底层系统串口名被定义为“/dev/serial0”，该名称在底层系统是独立唯一的，与之对应有一系列的底层函数，这些底层函数可以直接完成对串口硬件的任意操作，但对于应用程序来说这些底层函数是不公开的，所以应用程序不能直接调用底层函数，只能通过前面的 eCos 标准接口函数来间接调用。

eCos 在内部建立有几张表，表应该包含下列内容（和实际有区别）：

设备名	/dev/serial0	/dev/serial1	/dev/spi	/dev/i2c	……
实际设备	硬件串口一	硬件串口二	硬件 SPI 口	硬件 I2C 口	……

图 3.5.-2 eCos 设备名关系示意表

key 值	函数名	功能
0	uart_set_bandrate()	设置波特率
1	uart_set_length()	设置数据位长度
2	uart_set_check_bit()	设置校验位方式
3	uart_set_stop_bit()	设置停止位方式
…	…	…

图 3.5.-3 eCos 串口底层函数示意图

当应用程序调用 eCos 所带的标准接口函数时，会先从设备名表中找出设备名“/dev/serial0”对应的设备为硬件串口一，然后依据 key 值知道所需要进行的具体操作，参数则通过接口函数中的指针传递进来，只需要应用程序和底层函数都按统一格式传递参数就可保证参数正确。

将这些底层函数融合到操作系统的过程叫做驱动程序编写，任何一个操作系统推出时本身都只会支持某几种型号的 MCU，如果想要操作系统可以在新的 MCU 上运行就需要编写针对新 MCU 的驱动程序，常见做法是以一个操作系统支持的 MCU 为蓝本，在其基础之上实现对新 MCU 的支持，这个工作被称为操作系统平台移植。

操作系统平台移植不是一项简单的工作，这一过程需要严格遵循操作系统制定的各项规则，参与移植工作的人需要不但需要对操作系统了解透彻，还要对新的 MCU 的各种硬件特性了如指掌，所以平台移植是一件费时费力的事情，少则三五个月，稍微慢点就可能需要一年半载，对于大多数公司来说这无法接受。

正是基于这个原因前面 eCos 系统的构架图我做了修改，目的是满足公司产品开发时间紧迫的要求，如果应用程序不需要调用其它软件公司针对操作系统提供的第三方软件，底层驱动就不一定必须完全按照操作系统的规则进行编写。

第三方软件是软件公司针对某一类功能提供的专业软件包，也就是适用于应用层的上层驱动。为了通用，第三方软件所调用的接口函数必须完全满足操作系统所制定的规则。比如现在有一家第三方公司针对 eCos 编写了上层图形显示的 API，用户通过其提供的 API 可以非常方便的显示各种图形，第三方公司并不知道图形显示的具体方式，只是将处理好的数据通过 cyg_io_write()交给用 cyg_io_lookup()得到的显示设备句柄，如果第三方公司的 API 需要知道显示设备所能支持的数据格式、显示宽高等信息可以通过 cyg_io_get_config()得到。要支持这家公司的 API 显示驱动程序就必须按照 eCos 的要求编写，显示设备可以是“/dev/lcd”，也可以是“/dev/tv”，但都必须封装成 eCos 的标准形式。

不是所有的应用程序都需要使用此类第三方软件，这时可以采用一种简化方法。因为操作系统对任务调度、中断管理这些操作主要是以软件为主，硬件只提供少量最基本的支持，所以这部分移植工作量相对要简单一些，我们只是将这部分代码按操作系统要求进行移植，对于设备驱动程序不进行移植。设备驱动程序还是按照传统的方法编写代码，提供下面的驱动函数。

lcd_init()

lcd_set_config()

lcd_get_con()

lcd_write_data()

应用程序需要显示的时候直接调用这些驱动函数，不过编写这些驱动函数时要留意防止不同任务同时调用同一个驱动函数，在函数内部需要有保护代码。这种方式可以将开发时间大为缩短，不足是不满足操作系统的通用性原则，象第三方公司提供的 API 就无法使用。

总体说掌握嵌入式操作系统还是一件较有难度的事情，单凭一两篇文章就想弄清楚操作系统是不现实的，理解操作系统细节唯一的方法是阅读理解其提供的源代码，有一定基础知识之后自己再尝试移植一个简单的系统，并写出一两个符合操作系统要求的驱动程序，当完这些事情都你就会发现嵌入式操作系统已经嵌入到你的头脑之中。

嵌入式误区之不死机

嵌入式系统一词在国内流行开的时间并不长，刚开始并不叫嵌入式，我记得九十年代末还习惯被称为RTOS（实时多任务操作系统）。自我第一次接触到RTOS，给我印象最深刻的是其宣称的高可靠性不死机的超强特性，培训中无论是 RTOS 软件公司的市场和技术人员还是本公司的前辈都一再强调这一点，使得刚参加工作的我虽然有些不理解但不敢质疑。

记得当时讲解如何实现不死机这一特性主要是依靠RTOS 公司的专业性和 RTOS 的多任务机制得以实现。

首先 RTOS 是由专业软件公司开发完成的，这样的公司技术人员具有丰富的经验，所以写出来的程序可靠性高，出错的几率比较小，而我们这样的技术人员尤其是刚参加工作的，经验欠缺，所写的程序自然容易出错，这一点想想确实如此，没有可质疑的地方。

其次一个 RTOS 在推出之前经过了严格的测试，进一步降低了出错的几率，一般公司写的程序都没有经过如此严格的测试，这一点好象也是那么回事。

最后一点是不用 RTOS 写的程序为单任务模式，通常程序需要完成键盘扫描、显示和功能控制等操作，这些功能模块是串联模式，一旦有一个因异常导致死循环，整个循环就会一同死掉，如果是多任务可以将这些功能分别放在不同任务当中，即使有死循环产生其它任务还能继续工作，不至于死掉，好象说得也有道理。

解释完负责培训的人员还会列举使用 RTOS 的例子：美国的航天飞机控制程序必须采用RTOS，美国军方的一些控制设备也要求使用RTOS，意思就是世界上使用技术最先进、对产品质量要求最高的地方都是要用RTOS。到这个时候就是再多的半信半疑也只能是咽到肚子里去，总不能跑到美国去求证真伪。

我的记忆可能不准确，就算没有记错也有可能是当时培训的人员为了推荐产品自己有一些夸张，或者是他自己理解不够准确，所以我前面的叙述并不能肯定 RTOS 不死机的说法真的存在。这里我从网络中摘录了一些关于 RTOS 的概念陈述，发现大都同样有提到不死机的特性，看来这一说话确实存在。

1.实时多任务操作系统(RTOS)

(1)更加面向硬件系统，而不是操作者

嵌入式系统处理器一般都是独立工作的，没有人的直接参与；即使参与，也没有大量的文字信息输出，这是和桌面计算机有所不同的。因此 RTOS 着重面向的是硬件，而不是具有完整的人机界面。

*(2)**实时性***

单片机系统的监测、控制、通信等工作都要求实时性，一旦出现有关情况，CPU 能够及时响应，刻不容缓。为此，一个实用的RTOS 都应具有完善的中断响应机制，保证中断响应潜伏时间足够短。

(3)多任务

半导体技术的发展和应用复杂性的增长促使CPU的处理能力越来越高，当今的一片16位或32位单片机，在运算速度、寻址能力等方面可以相当于8位单片机的几十片之和。在这样强大的处理器上运行应用程序，必然不是整块，而是根据所要实现的若干方面功能，划分为数个任务，这样有利于软件的开发和维护。

因此单片机系统中采用的RTOS必然是支持多任务的，并能够根据各个任务的轻重缓急，合理地在它们之间分配CPU和各种资源的占用时间。

(4)不同的典型外设驱动支持

单片机的典型片内外设为定时器、A/D、PWM、D/A、串行口、LCD/LED接口,CAN-bus、IC-bus等。根据处理器类型的不同，RTOS在出厂时一般附带若干上面硬件接口的驱动程度，而网卡等片外设备的驱动程序，以及其它一些高级驱动函数，如兼容DOS的文件系统、TCP/IP协议等，则需要另行选购。以RTOS为基础和接口标准，可以设计出大量的库函数驱动模块，并根据实际需要选择或裁剪。

*(5)**高可靠性***

*一般计算机的操作系统出现问题，例如**死机**，除数据丢失等外，不会有太大的问题；而单片机系统一般都是和工业控制、交通工具、医用器械等机电系统密切相关，不适当的输出甚至不及时的输出都可能会带来财产损失和安全问题。因此嵌入式系统中的RTOS要求高可靠性，发行之前必须经过严格的测试。这是一个耗费时间和精力的过程，也是RTOS价格普遍高于一般操作系统的原因之一。*

2.RTOS是一个内核

典型的单片机程序在程序指针复位后，首先进行堆栈、中断、中断向量、定时器、串行口等接口设置、初始化数据存储区和显示内容，然后就来到了一个监测、等待或空循环，在这个循环中，CPU可以监视外设、响应中断或用户输入。

这段主程序可以看作是一个内核，内核负责系统的初始化和开放、调度其它任务，相当于C语言中的主函数。

RTOS就是这样的一个标准内核，包括了各种片上外设初始化和数据结构的格式化，不必、也不推荐用户再对硬件设备和资源进行直接操作，所有的硬件设置和资源访问都要通过RTOS核心。硬件这样屏蔽起来以后，用户不必清楚硬件系统的每一个细节就可以进行开发，这样就减少了开发前的学习量。

一般来说，对硬件的直接访问越少，系统的可靠性越高。RTOS是一个经过测试的内核，与一般用户自行编写的主程序内核相比，更规范，效率和可靠性更高。对于一个精通单片机硬件系统和编程的“老手”而言，通过RTOS对系统进行管理可能不如直接访问更直观、自由度大，但是通过RTOS管理能够排除人为疏忽因素，提高软件可靠性。

*另外，**高效率**地进行多任务支持是RTOS设计从始至终的一条主线，采用RTOS管理系统可以统*

一协调各个任务，优化 CPU 时间和系统资源的分配，使之不空闲、不拥塞。针对某种具体应用，精细推敲的应用程序不采用 RTOS 可能比采用 RTOS 能达到更高的效率；但是对于大多数一般用户和新手而言，采用 RTOS 是可以提高资源利用率的，尤其是在片上资源不断增长、产品可靠性和进入市场时间更重要的今天。

3.RTOS 是一个平台

RTOS 建立在单片机硬件系统之上，用户的一切开发工作都进行于其上，因此它可以称作是一个平台。采用 RTOS 的用户不必花大量时间学习硬件，和直接开发相比起点更高。

RTOS 还是一个标准化的平台，它定义了每个应用任务和内核的接口，也促进了应用程序的标准化。应用程序标准化后便于软件的存档、交流、修改和扩展，为嵌入式软件开发的工程化创造了条件、减少开发管理工作量。嵌入式软件标准化推广到社会后，可以促进软件开发的分工，减少重复劳动，近来出现的建立于 RTOS 上的文件和通信协议库函数产品等就是实例。

RTOS 对于开发单位和开发者个人来说也是一种提高。引入 RTOS 的开发单位，相当于引入了一套行业中广泛采用的嵌入式系统应用程序开发标准，使开发管理更简易、有效。基于 RTOS 和 C 语言的开发，具有良好的可继承性，在应用程序、处理器升级以及更换处理器类型时，现存的软件大部分可以不经修改地移植过来。

对于开发人员来说，则相当于在程序设计中采用一种标准化的思维方式，提高知识创造的效率；同时因为具有类似的思路，可以更快地理解同行其它人员的创造成果。

4.RTOS 产生并得到迅速发展的原因

单片机处理器能力的提高和应用程序功能的复杂化、精确化，迫使应用程序划分为多个重要性不同的任务，在各任务间优化地分配 CPU 时间和系统资源，同时还要保证实时性。靠用户自己编写一个实现上述功能的内核一般是不现实的，而这种需求又是普遍的。在这种形势之下，由专业人员编写的、满足大多数用户需要的高性能 RTOS 内核就是一种必然结果了。

对程序实时性和可靠性要求的提高也是 RTOS 发展的一个原因。此外，单片机系统软件开发日趋工程化，产品进入市场时间不断缩短，也迫使管理人员寻找一种有利于程序继承性、标准化、多人并行开发的管理方式。从长远的意义上来讲，RTOS 的推广能够带来嵌入式软件工业更有效、更专业化的分工，减少社会重复劳动、提高劳动生产率。

5.RTOS 的基本特征

(1)任务

任务(Task)是 RTOS 中最重要的操作对象，每个任务在 RTOS 的调用下由 CPU 分时执行。激活的或当前任务是 CPU 正在执行的任务，休眠的任务是在存储器中保留其执行的上下文背景、一旦切换为当前任务即可从上次执行的末尾继续执行的任务。任务的调度目前主要有时间分片式(TimeSlicing)、轮流查询式(Round-Robin)和优先抢占式(Preemptive)三种，不同的 RTOS 可能支持

其中的一种或几种，其中优先抢占式对实时性的支持最好。

(2)任务切换

RTOS 管理下的系统 CPU 和系统资源的时间是同时分配给不同任务的，这样看起来就象许多任务在同时执行，但实际上每个时刻只有一个任务在执行，也就是当前任务。任务的切换有两种原因。当一个任务正常地结束操作时，它就把 CPU 控制权交给 RTOS，RTOS 则检查任务队列中的所有任务，判断下面那个任务的优先级最高，需要先执行。另一种情况是在一个任务执行时，一个优先级更高的任务发生了中断，这时 RTOS 就将当前任务的上下文保存起来，切换到中断任务。RTOS 经常性地整理任务队列，删除结束的任务，增加新的要执行任务，并将其按照优先级从大到小的顺序排列起来，这样可以合理地在各个任务之间分配系统资源。

(3)消息和邮箱

消息(Message)和邮箱(Mailbox)是 RTOS 中任务之间数据传递的载体和渠道，一个任务可以有多个邮箱。通过邮箱，各个任务之间可以异步地传递信息，没有占用 CPU 时间的查询和等待。当 RTOS 包含片上总线接口驱动功能时，各个单片机之间的通信也通过邮箱的方式来进行，用户并不需要了解更深的关于硬件的内容。

(4)旗语

旗语(Semaphore)相当于一种标志(Flag)，通过预置,一个事件的发生可以改变旗语。一个任务可以通过监测旗语的变化来决定其行动，在监测旗语变化的时候不消耗 CPU 时间，旗语对任务的触发是由 RTOS 来完成的。通过使用旗语，一个任务在等待事件变化的时候就可以不必不断查询，而把 CPU 时间出让给其它任务。

(5)存储区分配

RTOS 对系统存储区进行统一分配，分配的方式可以是动态的或静态的，每个任务在需要存储区时都要向 RTOS 内核申请。RTOS 通过使用存储分配类核心对象管理数据存储器，在动态分配时能够防止存储区的零碎化。

(6)中断和资源管理

*RTOS 提供了一种通用的设计用于中断管理，有效率而灵活，这样可以实现最小的中断潜伏时间和最大的中断响应度。RTOS 内核中的资源对象类则实现了对系统实体资源或虚拟资源的独占式访问，一个任务可以取得对资源的唯一访问权，其它任务在资源释放以前无法访问，这样可以避免资源冲突。设计完善的 RTOS 具有检查可能导致**系统死锁**的资源调用设计。*

阅读完这段陈述除了知道 RTOS 强调自己的高可靠性不死机外，还表明它实际上就是嵌入式操作系统，只是表述方法不同而已。

那 RTOS 到底有没有具备不死机的超强特性呢？这种说法在我看来是错误的，至少可以说不够严谨。可以说无论是硬件还是软件，目前还不存在完全解决了死机问题的方法，所有产品都只是尽可能的降低死机的几率，不可能将死机的几率降到零。

死机的原因五花八门，可能是硬件的，也可能是软件的，如果是硬件原因导致，单纯依靠软件

是无法解决的，而 RTOS 只是软件方面的产品，凭这一点就可以说 RTOS 不死机言过其实。在软件层面，RTOS 只是提供了一个程序框架，并不包含有实际功能的应用程序，要用到产品中就需要工程师在此框架基础上编写相应应用程序，就算 RTOS 自我非常完善，可应用程序的内容还是由应用工程师决定，如果应用程序出错 RTOS 同样无能为力。

虽然 RTOS 可以采用某些方式对系统自身进行保护，但程序运行起来后始终会出现 CPU 完全由应用程序控制的状态，这个时候 RTOS 如果没有硬件特殊功能（MPU）的支持，在有问题的应用程序面前同样如同一只任人宰割的羔羊。

```
UINT32 *p,i;
i=0x00000000;                  //i 进行这样一段复杂的运算是为了避免编译器优化
i=i+0x00001234;
i=i+0x56780000;
i=i&0x0000C1C1;                //到这里 i 的实际结果等于 0
p=(UINT32 *)i;                 //所以 p 这里也指向地址 0
for(i=0;i<0x00100000;i++)      //将从地址 0 开始的 4MBytes 空间清 0
{
  *p=0x00000000;
  p++;
}
```

通常 OS 都位于存储器从地址 0 开始的一段区域，如果执行这段代码会将 OS 所在区域清 0，也就是 OS 自身被应用程序破坏掉，试想 RTOS 面对这样的代码何以保证高可靠性？当然这样的代码是不允许存在的，但可以说明一个问题，应用程序在使用指针时如果指针出错，就存在将整个 RTOS 摧毁的可能。

对于编写应用程序的工程师，如果是在 RTOS 上进行编程，就应用程序本身来说，出错的几率和不使用 RTOS 是相同的，程序的质量主要靠工程师的素质来把握，和 RTOS 关系不大。因为 RTOS 多少对程序编写存在一些限制，所以对于工程师实际上会增加一些额外的负担，需要了解 RTOS 的相关知识，而且程序调试会因为 RTOS 的存在要麻烦一些。

所以 RTOS 高可靠性、不死机的说法是不正确的。实际上 RTOS 是适合逻辑流程相对复杂、而且需要同时处理多项事情的程序，如果是单任务方式，程序就需要非常多的判断、跳转操作，即便是经验丰富的工程师也可能会被过多的逻辑流程控制把自己绕晕，有了 RTOS 可以将不同的事情处理分离到独立的任务当中，任务之间通过 RTOS 传递交换数据，逻辑流程自然就明晰起来。

我们不用 RTOS 也可以实现 RTOS 类似的工作，比如我们可以在将每一个事项的处理分成许多小段程序，然后利用定时中断来依次执行每个事项的分段程序。这样做需要中断程序具备管理功能，以保证每次中断正确调用相应程序分段，这里的中断程序相当于一个功能最简的 RTOS。虽然中断程

序只是实现简单的管理功能，实际要做好并不容易。现有的 RTOS 正是解决了这个问题，将任务的管理工作很稳定的实现，高可靠性指的是这一点。

RTOS 不但在任务管理方面可靠性高，所提供的功能也是相当强大，几乎考虑了所有的用户需求，另外模块式的程序结构可以让用户自由进行功能裁剪，让用户制定出适合自己且经济高效的系统，应用层标准化接口更加有利技术的分工合作，使得软件公司开发基于 RTOS 的标准功能库成为现实。

所以不要因为我反对了 RTOS 的某一个宣传说法有所夸大就全面否定 RTOS，它的优点是符合技术发展的潮流趋势，现在不少产品都不能全靠自己的力量完成，适当引用其它公司的现有技术可以让产品开发周期和质量都得到提升。RTOS 确实是一个好东西，就好比 WINDOWS，和 MSDOS 相比虽然复杂许多，而且需要功能更强的硬件支持，但能给用户带来完全不一样的感受。

使用 RTOS，就如同站在巨人的肩膀上看世界，只是爬上巨人的肩膀需要多花一些力气。

嵌入式效率

嵌入式系统一贯宣称自己的实时、高效，在我看来这一点也是不正确的，最高效率是由硬件决定的，一旦硬件平台确定，任何软件都无法突破其最高效率。

也许有人会说软件高手编写的程序效率会比普通软件人员写的效率会高，嵌入式操作系统都是软件高手来完成的，说效率高很正常啊。这种说法采用的是概念转移的伪证法，将一个硬件平台所能达到的软件最高效率问题转换成不同程序员编写的代码效率高低。我们很容易将这种说法辩驳倒，让同一个写操作系统的程序员用两种方法来完成产品代码的编写，一种是程序员在操作系统之上编写，另外一种是程序员直接针对硬件编写，显然后一种方法得到的程序效率要高。

嵌入式操作系统是软件，软件的运行就要消耗 CPU 资源，同一个 CPU 其能提供的资源是恒定的，既然嵌入式系统运行耗费了部分资源，对于应用程序来说可用的资源就要减少，所以能达到的最大效率也自然要相应降低。另外操作系统为了实现自己的管理功能，需要打开 TIMER 中断为整个操作系统提供同步时钟信号，这些中断程序可能对产品实际功能并无多大作用，但要占用 CPU 运行时间，如果不用操作系统可以关闭掉。所以同样功能的软件，带操作系统的效率肯定要低过带操作系统的版本。

同样道理，硬件平台对于事件的最快响应时间也是恒定的，嵌入式样操作系统在快速实时性方面实际上并不理想，因为操作系统的构架方式对事件的响应会明显慢过硬件所支持的速度。因为中断是操作系统直接管理的资源，操作系统在中断向量表中放置的并不是中断程序服务程序的地质，而是操作系统的中断管理函数入口地址，当硬件中断信号产生后，CPU 不是直接执行相应中断服务程序，而是先执行操作系统的总中断管理函数，在该函数中再由软件决定何时执行中断服务程序。这种模式使得操作系统对中断的响应效率大打折扣。操作系统为了通用，所提供的接口都是 C 语言形式，C 语言一般来说代码效率没有汇编高，所以操作系统编程语言的选择又使得系统对中断的响应速度更慢了一些。

所以单纯从代码执行效率来看，嵌入式系统并没有高效的特点，这个方面反而是嵌入式系统的不足。但如果从另外一个角度来理解，也还是可以承认这种说法。现在硬件的速度越来越快，硬件速度的提升弥补了嵌入式系统运行效率的不足。嵌入式系统对多任务的支持，可以让原本复杂的逻辑流程变简单，加上基操作系统的大量第三方标准软件的出现，许多工作都可以直接使用别人的现有成果，在整个产品的开发角度来说无疑是更加高效的方法。