4.3.2 Cache—主存地址映射

由主存地址映射到 Cache 地址称为地址映射。地址映射方式很多,有直接映射(固定的映射关系)、全相联映射(灵活性大的映射关系)、组相联映射(上述两种映射的折中)。

1. 直接映射

图 4.54 示出了直接映射方式主存与缓存中字块的对应关系。

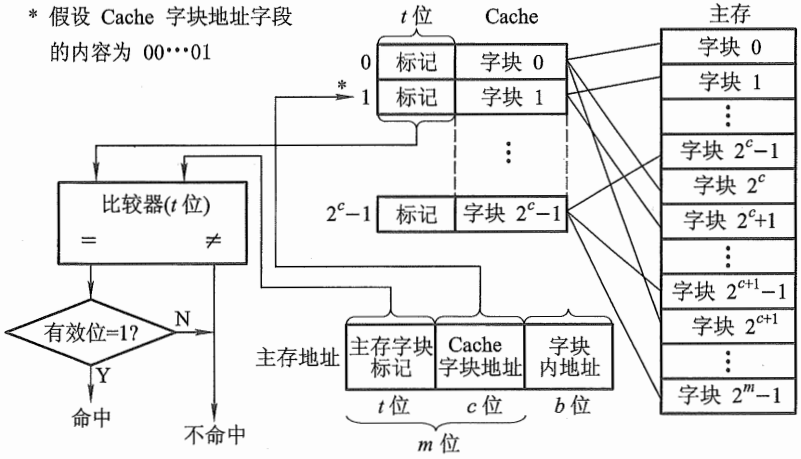


图 4.54 直接映射

图中每个主存块只与一个缓存块相对应,映射关系式为

$$i=j \bmod C \qquad 或 \qquad i=j \bmod 2^C$$

其中,i 为缓存块号,j 为主存块号,C 为缓存块数。映射结果表明每个缓存块对应若干个主存块,如表 4.4 所示。

表 4.4 直接映射方式主存块和缓存块的对应关系

缓存块	主 存 块
0	$0, C, \cdots, 2^m - C$
1	$1, C+1, \cdots, 2^m - C + 1$
…	…
$C-1$	$C-1, 2C-1, \cdots, 2^m - 1$

这种方式的优点是实现简单,只需利用主存地址的某些位直接判断,即可确定所需字块是否在缓存中。由图 4.54 可见,主存地址高 m 位被分成两部分:低 c 位是指 Cache 的字块地址,高 t 位($t=m-c$)是指主存字块标记,它被记录在建立了对应关系的缓存块的“标记”位中。当缓存接

到 CPU 送来的主存地址后，只需根据中间 c 位字段（假设为 00…01）找到 Cache 字块 1，然后根据字块 1 的“标记”是否与主存地址的高 t 位相符来判断，若符合且有效位为“1”（有效位用来识别 Cache 存储块中的数据是否有效，因为有时 Cache 中的数据是无效的，例如，在初始时刻 Cache 应该是“空”的，其中的内容是无意义的），表示该 Cache 块已和主存的某块建立了对应关系（即已命中），则可根据 b 位地址从 Cache 中取得信息；若不符合，或有效位为“0”（即不命中），则从主存读入新的字块来替代旧的字块，同时将信息送往 CPU，并修改 Cache“标记”。如果原来有效位为“0”，还得将有效位置成“1”。

直接映射方式的缺点是不够灵活，因每个主存块只能固定地对应某个缓存块，即使缓存内还空着许多位置也不能占用，使缓存的存储空间得不到充分的利用。此外，如果程序恰好要重复访问对应同一缓存位置的不同主存块，就要不停地进行替换，从而降低命中率。

2. 全相联映射

全相联映射允许主存中每一字块映射到 Cache 中的任何一块位置上，如图 4.55 所示。这种映射方式可以从已被占满的 Cache 中替换出任一旧字块。显然，这种方式灵活，命中率也更高，缩小了块冲突率。与直接映射相比，它的主存字块标记从 t 位增加到 $t+c$ 位，这就使 Cache“标记”的位数增多，而且访问 Cache 时主存字块标记需要和 Cache 的全部“标记”位进行比较，才能判断出所访问主存地址的内容是否已在 Cache 内。这种比较通常采用“按内容寻址”的相联存储器（见附录 4A）来完成。

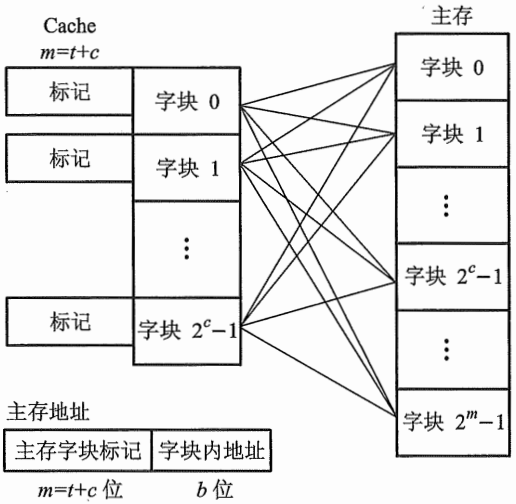


图 4.55　全相联映射

总之，这种方式所需的逻辑电路甚多，成本较高，实际的 Cache 还要采用各种措施来减少地址的比较次数。

3. 组相联映射

组相联映射是对直接映射和全相联映射的一种折中。它把 Cache 分为 Q 组，每组有 R 块，并有以下关系：

$$i=j \bmod Q$$

其中，i 为缓存的组号，j 为主存的块号。某一主存块按模 Q 将其映射到缓存的第 i 组内，如图4.56所示。

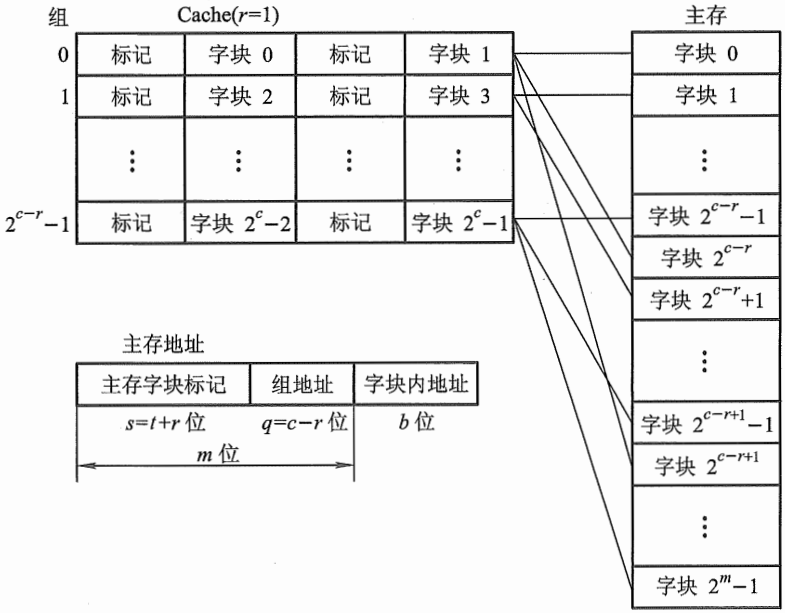


图 4.56 组相联映射

组相联映射的主存地址各段与直接映射(参见图 4.54)相比，还是有区别的。图 4.54 中 Cache 字块地址字段由 c 位变为组地址字段 q 位，且 $q=c-r$，其中 2^c 表示 Cache 的总块数，2^q 表示 Cache 的分组个数，2^r 表示组内包含的块数。主存字块标记字段由 t 位变为 $s=t+r$ 位。为了便于理解，假设 $c=5$，$q=4$，则 $r=c-q=1$。其实际含义为：Cache 共有 $2^c=32$ 个字块，共分为 $2^q=16$ 组，每组内包含 $2^r=2$ 块。组内 2 块的组相联映射又称为二路组相联。

根据上述假设条件，组相联映射的含义是：主存的某一字块可以按模 16 映射到 Cache 某组的任一字块中。即主存的第 0，16，32…字块可以映射到 Cache 第 0 组 2 个字块中的任一字块；主存的第 15，31，47…字块可以映射到 Cache 第 15 组中的任一字块。显然，主存的第 j 块会映射到 Cache 的第 i 组内，两者之间一一对应，属直接映射关系；另一方面，主存的第 j 块可以映射到 Cache 的第 i 组内中的任一块，这又体现出全相联映射关系。可见，组相联映射的性能及其复杂性介于直接映射和全相联映射两者之间，当 $r=0$ 时是直接映射方式，当 $r=c$ 时是全相联映射方式。

例 4.8　假设主存容量为 512 KB，Cache 容量为 4 KB，每个字块为 16 个字，每个字 32 位。

（1）Cache 地址有多少位？可容纳多少块？

（2）主存地址有多少位？可容纳多少块？

（3）在直接映射方式下，主存的第几块映射到 Cache 中的第 5 块（设起始字块为第 1 块）？

（4）画出直接映射方式下主存地址字段中各段的位数。

解：（1）根据 Cache 容量为 4 KB（2^{12}=4 K），Cache 地址为 12 位。由于每字 32 位，则 Cache 共有 4 KB/4 B=1 K 字。因每个字块 16 个字，故 Cache 中有 1 K/16=64 块。

（2）根据主存容量为 512 KB（2^{19}=512 K），主存地址为 19 位。由于每字 32 位，则主存共有 512 KB/4 B=128 K 字。因每个字块 16 个字，故主存中共 128 K/16=8 192 块。

（3）在直接映射方式下，由于 Cache 共有 64 块，主存共有 8 192 块，因此主存的 5，64+5，2×64+5，…，2^{13}-64+5 块能映射到 Cache 的第 5 块中。

（4）在直接映射方式下，主存地址字段的各段位数分配如图 4.57 所示。其中字块内地址为 6 位（4 位表示 16 个字，2 位表示每字 32 位），缓存共 64 块，故缓存字块地址为 6 位，主存字块标记为主存地址长度与 Cache 地址长度之差，即 19-12=7 位。

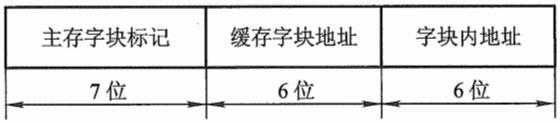


图 4.57　例 4.8 主存地址各字段的分配

例 4.9　假设主存容量为 512 K×16 位，Cache 容量为 4 096×16 位，块长为 4 个 16 位的字，访存地址为字地址。

（1）在直接映射方式下，设计主存的地址格式。

（2）在全相联映射方式下，设计主存的地址格式。

（3）在二路组相联映射方式下，设计主存的地址格式。

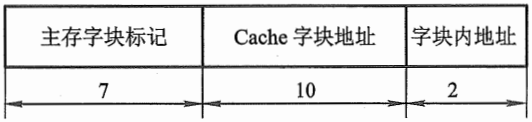
（4）若主存容量为 512 K×32 位，块长不变，在四路组相联映射方式下，设计主存的地址格式。

解：（1）根据 Cache 容量为 4 096=2^{12}字，得 Cache 字地址为 12 位。根据块长为 4，且访存地址为字地址，得字块内地址为 2 位，即 b=2，且 Cache 共有 4 096/4=1 024=2^{10}块，即 c=10。根据主存容量为 512 K=2^{19}字，得主存字地址为 19 位。在直接映射方式下，主存字块标记为 19-12=7。主存的地址格式如图 4.58（a）所示。

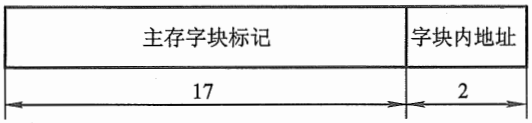
（2）在全相联映射方式下，主存字块标记为 19-b=19-2=17 位，其地址格式如图4.58（b）所示。

（3）根据二路组相联的条件，一组内有 2 块，得 Cache 共分 1 024/2=512=2^q 组，即 q=9，主存字块标记为 19-q-b=19-9-2=8 位，其地址格式如图 4.58（c）所示。

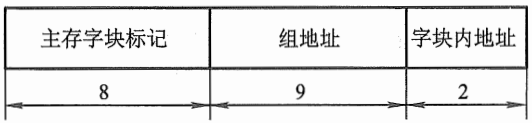
（4）若主存容量改为 512 K×32 位，即双字宽存储器，块长仍为 4 个 16 位的字，访存地址仍为字地址，则主存容量可写为 1 024 K×16 位，得主存地址为 20 位。由四路组相联，得 Cache 共分 1 024/4 = 256 = 2^q 组，即 $q=8$。对应该条件下，主存字块标记为 20−8−2 = 10 位，其地址格式如图 4.58(d)所示。



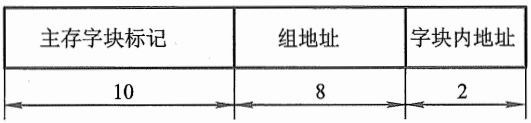
（a）直接映射方式主存地址格式



（b）全相联映射方式主存地址格式



（c）二路组相联映射方式主存地址格式



（d）四路组相联映射方式双字宽主存地址格式

图 4.58　例 4.9 主存地址格式

例 4.10　假设 Cache 的工作速度是主存的 5 倍，且 Cache 被访问命中的概率为 95%，则采用 Cache 后，存储器性能提高多少？

解：设 Cache 的存取周期为 t，主存的存取周期为 $5t$，则系统的平均访问时间为

$$t_a = 0.95\times t + 0.05\times 5t = 1.2t$$

性能为原来的 $5t/1.2t = 4.17$ 倍，即提高了 3.17 倍。

例 4.11　设某机主存容量为 16 MB，Cache 的容量为 8 KB。每字块有 8 个字，每字 32 位。设计一个四路组相联映射的 Cache 组织。

（1）画出主存地址字段中各段的位数。

（2）设 Cache 初态为空，CPU 依次从主存第 0，1，2，…，99 号单元读出 100 个字（主存一次读出一个字），并重复此次序读 10 次，问命中率是多少？

(3) 若 Cache 的速度是主存速度的 5 倍,试问有 Cache 和无 Cache 相比,速度提高多少倍?

(4) 系统的效率为多少?

解:(1) 根据每个字块有 8 个字,每个字 32 位,得出主存地址字段中字块内地址字段为 5 位,其中 3 位为字地址,2 位为字节地址。

根据 Cache 容量为 8 KB = 2^{13} B,字块大小为 2^5 B,得 Cache 共有 2^8 块,故 $c = 8$。根据四路组相联映射 $2^r = 4$,得 $r = 2$,则 $q = c - r = 8 - 2 = 6$ 位。

根据主存容量为 16 MB = 2^{24} B,得出主存地址字段中主存字块标记为 24-6-5 = 13 位。

主存地址字段各段格式如图 4.59 所示。

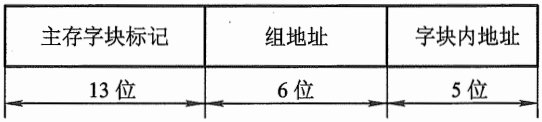


图 4.59　例 4.10 主存地址字段

(2) 由于每个字块中有 8 个字,而且初态 Cache 为空,因此 CPU 读第 0 号单元时,未命中,必须访问主存,同时将该字所在的主存块调入 Cache 第 0 组中的任一块内,接着 CPU 读 1~7 号单元时均命中。同理,CPU 读第 8,16,…,96 号单元时均未命中。可见 CPU 在连续读 100 个字中共有 13 次未命中,而后 9 次循环读 100 个字全部命中,命中率为

$$\frac{100\times10-13}{100\times10}=0.987$$

(3) 根据题意,设主存存取周期为 $5t$,Cache 的存取周期为 t,没有 Cache 的访问时间为 $5t\times1\,000$,有 Cache 的访问时间为 $t\,(1\,000-13)+5t\times13$,则有 Cache 和没有 Cache 相比,速度提高的倍数为

$$\frac{5t\times1\,000}{t(1\,000-13)+5t\times13}-1\approx3.75$$

(4) 根据(2)求得的命中率 0.987,主存的存取周期为 $5t$,Cache 的存取周期为 t,得系统的效率为

$$\frac{t}{0.987\times t+(1-0.987)\times5t}\times100\%=95\%$$

4.3.3　替换策略

当新的主存块需要调入 Cache 并且它的可用空间位置又被占满时,需要替换掉 Cache 的数据,这就产生了替换策略(算法)问题。在直接映射的 Cache 中,由于某个主存块只与一个 Cache 字块有映射关系,因此替换策略很简单。而在组相联和全相联映射的 Cache 中,主存块可以写入