存储器·高速缓冲存储器

xyfJASON

---

1 概述

1.1 工作原理

设主存有 2^n 个可编址字，被分为 $M=2^m$ 块，每块长度为 $B=2^b$ 个字，即有：$n=m+b$。由于主存地址有 n 位，我们取低 b 位作为块内地址，高 m 位作为块地址（主存块号）。同样的，cache 被分为 $C=2^c$ 块且 $C \ll M$，每块长 $B=2^b$ 个字，其地址的低 b 位为块内地址，高 c 位为块号。此外，cache 中每个缓存块都有一个标记，标记这个缓存块对应主存中的块号。

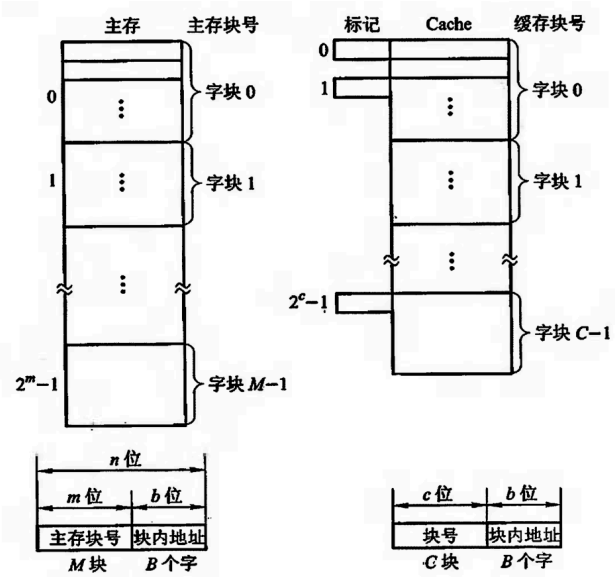

图 4.49　Cache－主存存储空间的基本结构

cache 命中：CPU 需要访问的块已经被调入 cache 中；

cache 未命中：CPU 需要访问的块未被调入 cache 中。

命中率：$h=\dfrac{N_c}{N_c+N_m}$，其中 N_c 表示访问 cache 命中的次数，N_m 表示访问主存的次数（即 cache 未命中）。

平均访问时间：$t_a=ht_c+(1-h)t_m$，其中 t_c 为命中时访问 cache 的时间，t_m 为未命中时访问主存的时间。

访问效率：$e=\dfrac{t_c}{t_a}\times 100\%=\dfrac{t_c}{ht_c+(1-h)t_m}\times 100\%$

1.2 基本结构

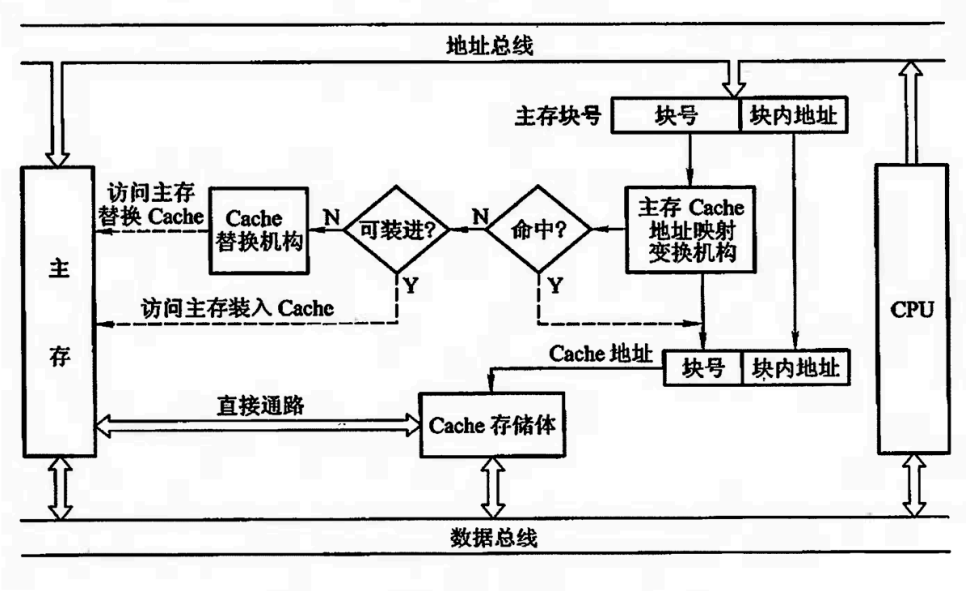

图 4.50　Cache 的基本结构原理框图

读操作：见上图流程

写操作：注意写入 cache 的信息必须与主存中对应位置的信息完全一致，有两种方法：

- 写直达法：写数据时既写入 cache 又写入主存，能保证 cache 与主存的数据随时一致，但增加了访存次数；
- 写回法：写数据时只写入 cache，不写入主存，当 cache 中的数据被替换出去时再写入主存。为此，每个缓存块需要增加一个标志位 (dirty)，标志这个块的内容是否被修改过而与主存不一致。

2 Cache-主存地址映射

图 4.50 中，主存 Cache 地址映射变换机构将主存的块号映射为 cache 的块号并确定是否命中，这样的映射方法不唯一，主要有以下三种方式：直接映射，全相联映射，组相联映射。

- 直接映射：$i = j \bmod C$，其中 i 为缓存块号，j 为主存块号，C 为缓存块数。

  这种方法下，cache 中的标记显然就是主存地址的高 $m - c$ 位。

  优点是实现简单，缺点是不够灵活、命中率不高。

- 全相联映射：主存的任何一块可以映射到 cache 的任何一块。

  这种方法下，cache 中的标记需要增加到 m 位。

  优点是灵活、命中率高。

  查找是否命中需要遍历整个 cache，常采用“按内容寻址”的相联存储器完成。

- 组相联映射：前两种的结合。将 cache 分为 $Q = 2^q$ 组，每组 $R = 2^r$ 块 $(c = q + r)$，组间直接映射（主存第 j 块映射到 cache 第 $j \bmod Q$ 组），组内全相联。

  这种方式下，主存地址的高 m 位又可以分为低 q 位的组地址和高 $m - q$ 位主存字块标记，cache 中的标记需要 $m - q$ 位。

  直接映射是一个块作为一组的特例，即 $q = c, r = 0$;

  全相联映射是所有块作为一组的特例，即 $q = 0, r = c$.

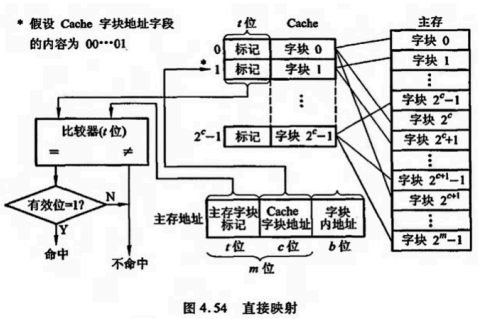

图 4.54 直接映射

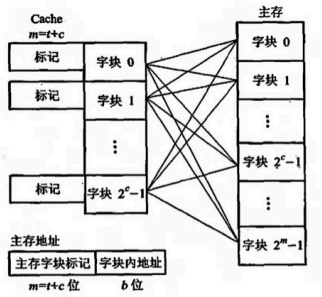

图 4.55 全相联映射

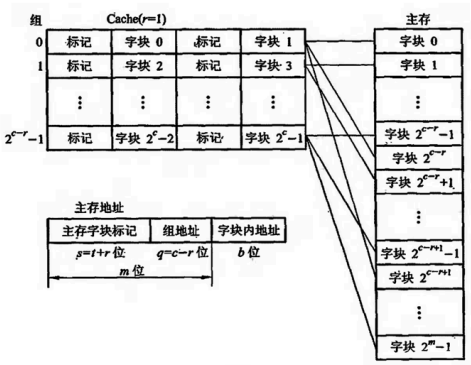

图 4.56 组相联映射

3 替换策略

图 4.50 中，替换机构完成主存和 cache 中数据块的替换操作。替换策略一般有以下三种：

- 先进先出 (FIFO)：容易实现、开销小，但没有利用局部性原理，不能提高命中率；
- 最近最少使用 (LRU)：平均命中率比 FIFO 高；
- 随机法：比较简单，但没有利用局部性原理，不能提高命中率。