计算机组成原理与系统结构

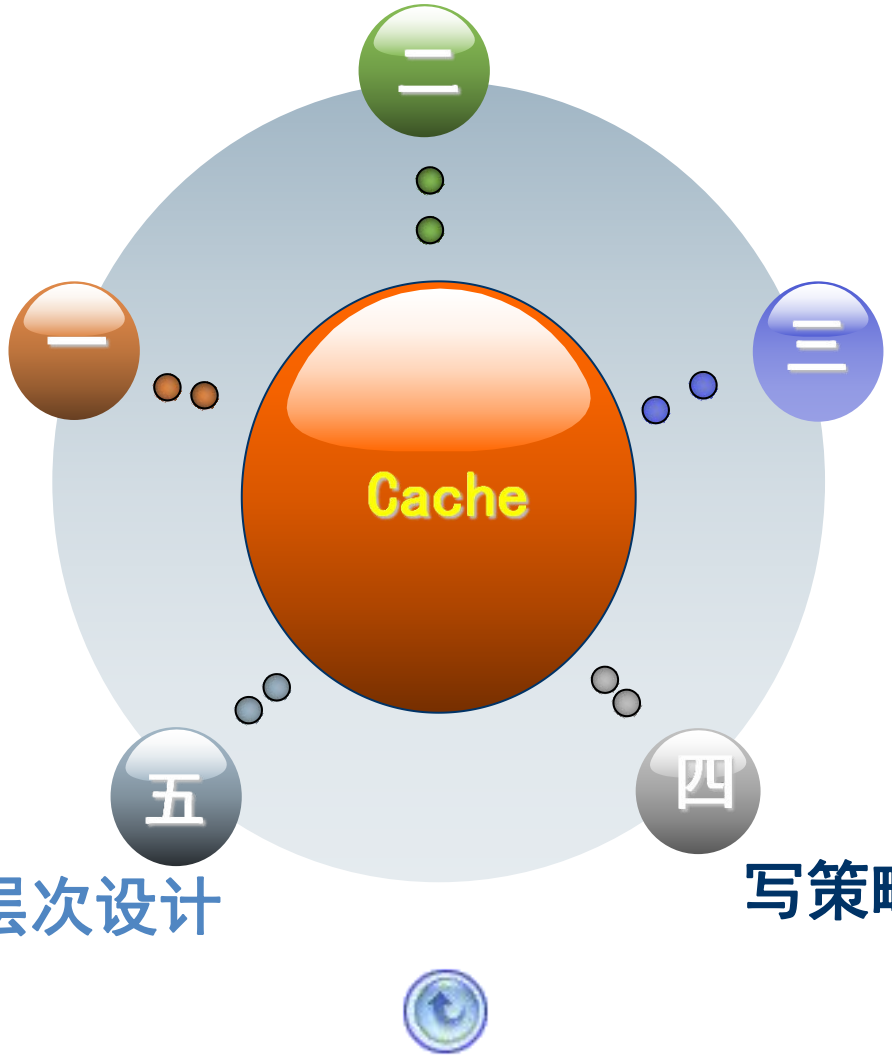
第五章　存储体系

http://www.icourses.cn/coursestatic/course_2859.html

杭州電子科技大学
HANGZHOU DIANZI UNIVERSITY

5.5 高速缓冲存储器Cache

主存与Cache的地址映射方式



Cache的基本原理

替换算法

Cache的多层次设计

写策略

一、Cache的基本原理

Cache的特点

Cache的工作原理

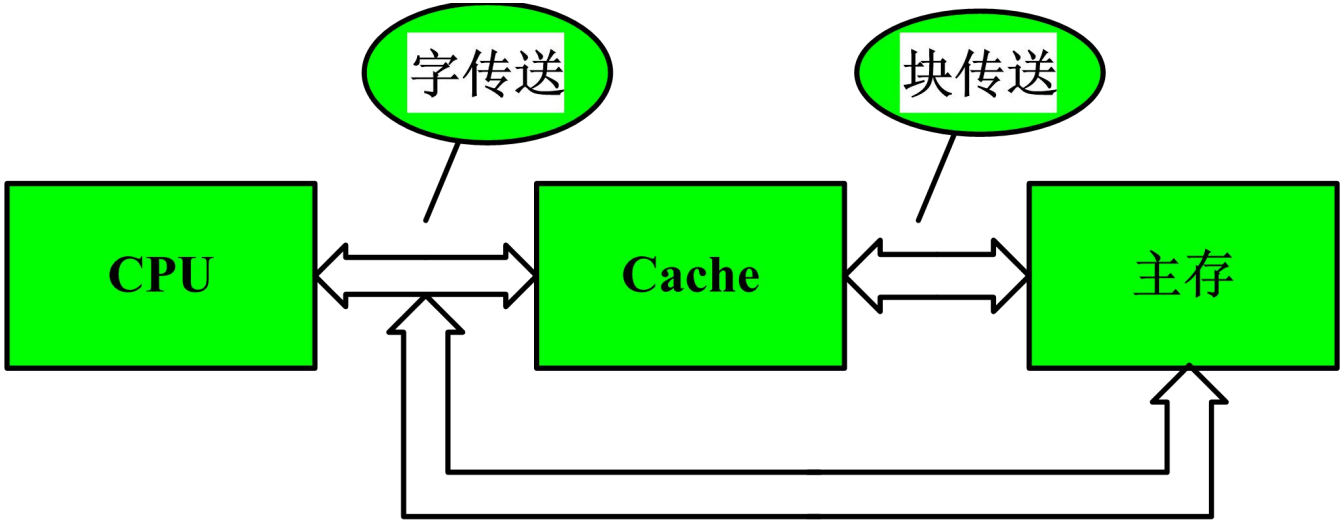
Cache的命中率

1、Cache的特点

- Cache是指位于CPU和主存之间的一个高速小容量的存储器，一般由SRAM构成。
- Cache功能：用于弥补CPU和主存之间的速度差异，提高CPU访问主存的平均速度。
- 设置Cache的理论基础，是程序访问的局部性原理。
- Cache的内容是主存部分内容的副本，Cache的功能均由硬件实现，对程序员是透明的。

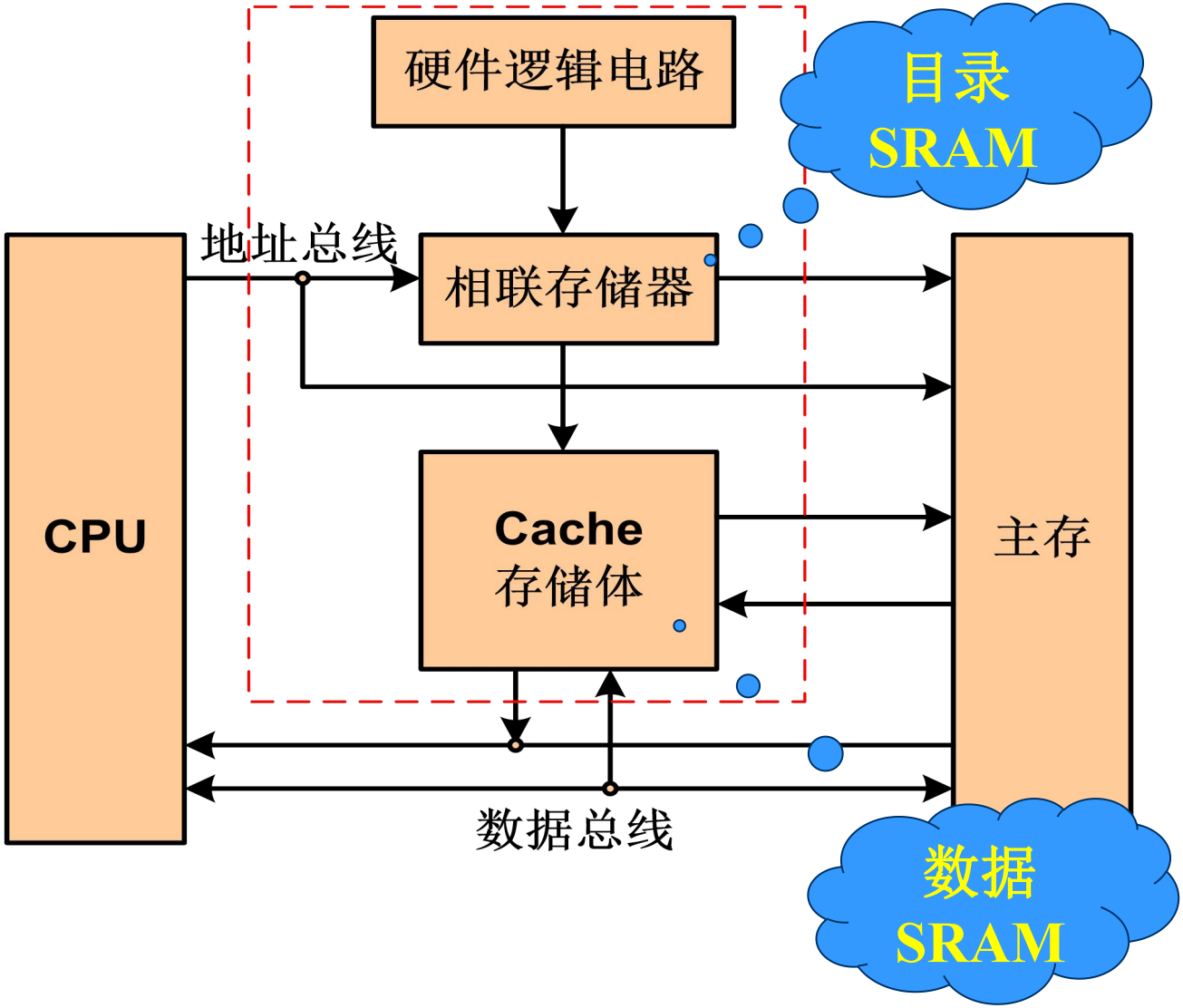
2、Cache的工作原理

- Cache的速度比主存快5－10倍。



Cache、主存与CPU的关系

Cache的原理图



Cache的读写操作

1. CPU在读写存储器时，Cache控制逻辑首先要依据地址来判断这个字是否在Cache中，若在Cache中，则称为“命中”；若不在，则称为“不命中”。
2. 针对命中/不命中、读/写操作，Cache的处理是不同的：

- 读命中：立即从Cache读出送给CPU；
- 读不命中：通常有两种解决方法：

⑩ 将主存中该字所在的数据块复制到Cache中，然后再把这个字传送给CPU；

Cache的读写操作

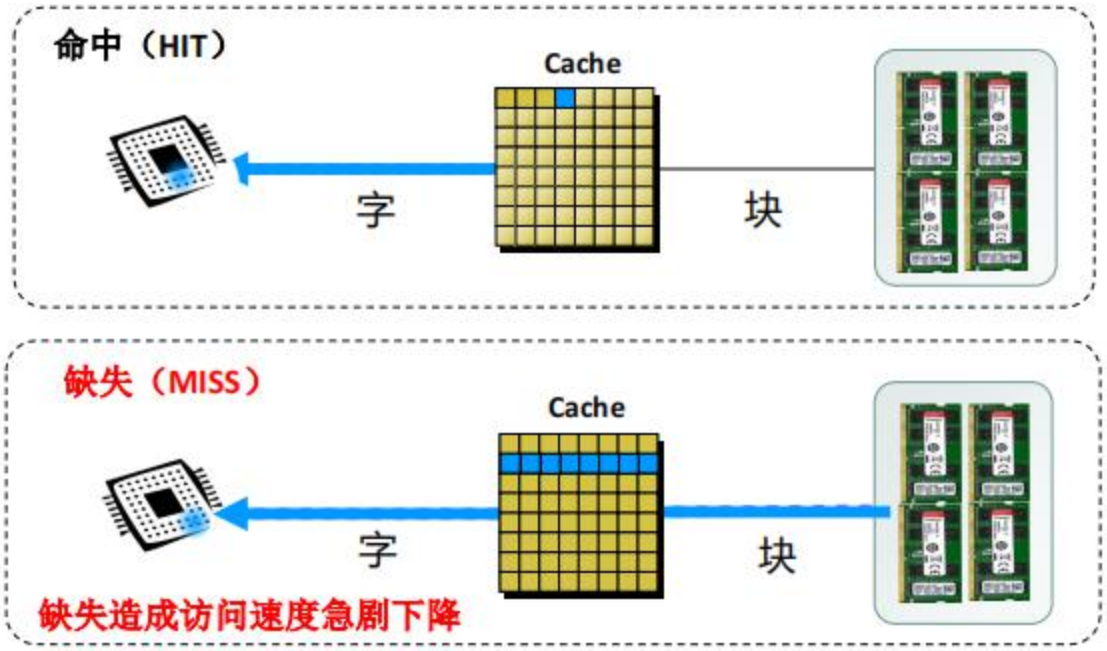
②把此字从主存读出送到CPU，同时，把包含这个字的数据块从主存中读出送到Cache中。

- 写不命中：直接将该字写入主存中，且不再调入Cache；
- 写命中：通常也有两种方法进行处理：

⑩写贯穿方法：同时对Cache和主存进行写操作；

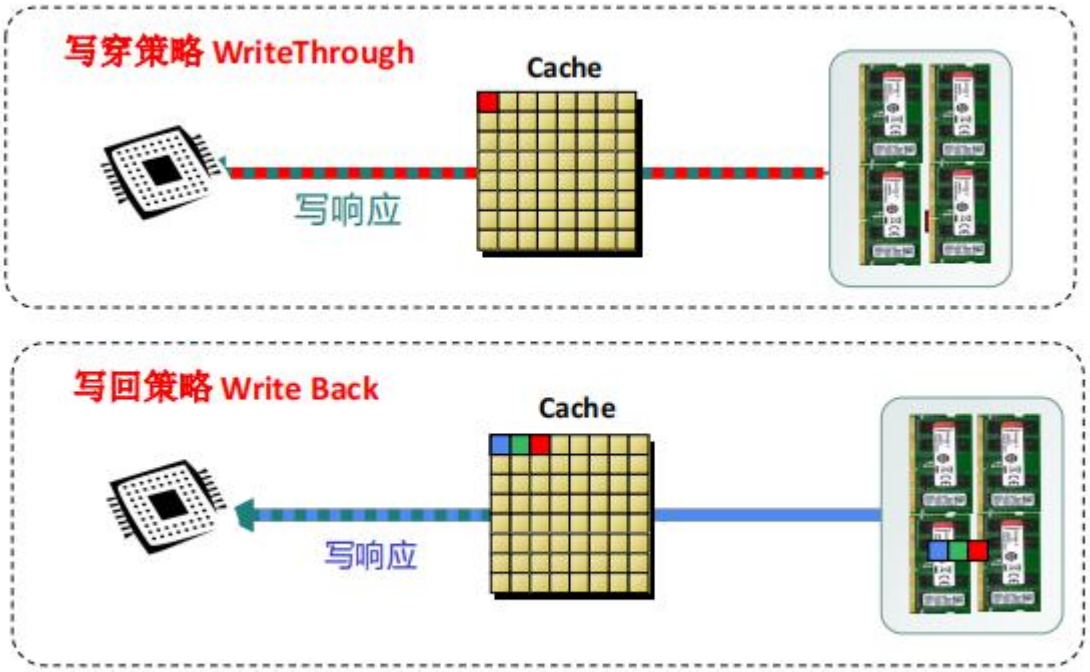
⑩写回：只写Cache，仅当此Cache块被替换时，才将该块写入主存

Cache 的工作工作过程



Cache的读操作

Cache 的工作工作过程



Cache的写操作

四、写策略

常用的写策略通常有写贯穿和写回两种

1. 写贯穿策略

当CPU写Cache命中时，所有写操作既对Cache也对主存进行；当CPU写Cache不命中时，直接写主存，有两种做法：

- 其一，不将该数据所在的块拷贝到Cache行，称为WTNWA法；
- 其二，将该数据所在块拷贝到Cache的某行，称为WTWA法。

四、写策略

2. 写回策略（Write Back）

- 当CPU写Cache命中时，写操作只是对Cache进行，而不修改主存的相应内容，仅当此Cache行被换出时，相应的主存内容才被修改；当CPU写Cache不命中时，先将该数据所在块拷贝到Cache的某行，余下操作与Cache写命中时相同。
- 为了区别Cache行是否被改写过，应为每个Cache行设置一个修改位，CPU修改Cache行时，标记其修改位，当此Cache行被换出时，判别此Cache行的修改位，从而决定是否将Cache行数据写回主存相应单元。

四、写策略

3. 两种写策略比较

- 写贯穿策略保证了主存数据总是有效，写回策略可能导致Cache和主存数据不一致；
- 写回策略的效率高于写贯穿策略；
- 写回策略的控制比写贯穿策略的控制复杂。

3、Cache的命中率

1. 命中率指CPU访问主存数据时，命中Cache的次数，占全部访问次数的比率；失效率就指不命中Cache的次数，占全部访问次数的比率。命中率h取决于程序的行为、Cache的容量、组织方式、块大小。

3、Cache的命中率

2. 在一个程序执行期间，设Nc表示Cache完成存取的总次数，Nm表示主存完成存取的总次数，则命中率：

$$h = \frac{N_c}{N_c + N_m} \times 100\%$$

若tc表示Cache的访问时间，tm表示主存的访问时间，则Cache/主存系统的平均访问时间ta为：

$$t_a = ht_c + (1-h)\ t_m$$

Cache/主存系统的访问效率e：

$$e = \frac{t_c}{t_a}$$

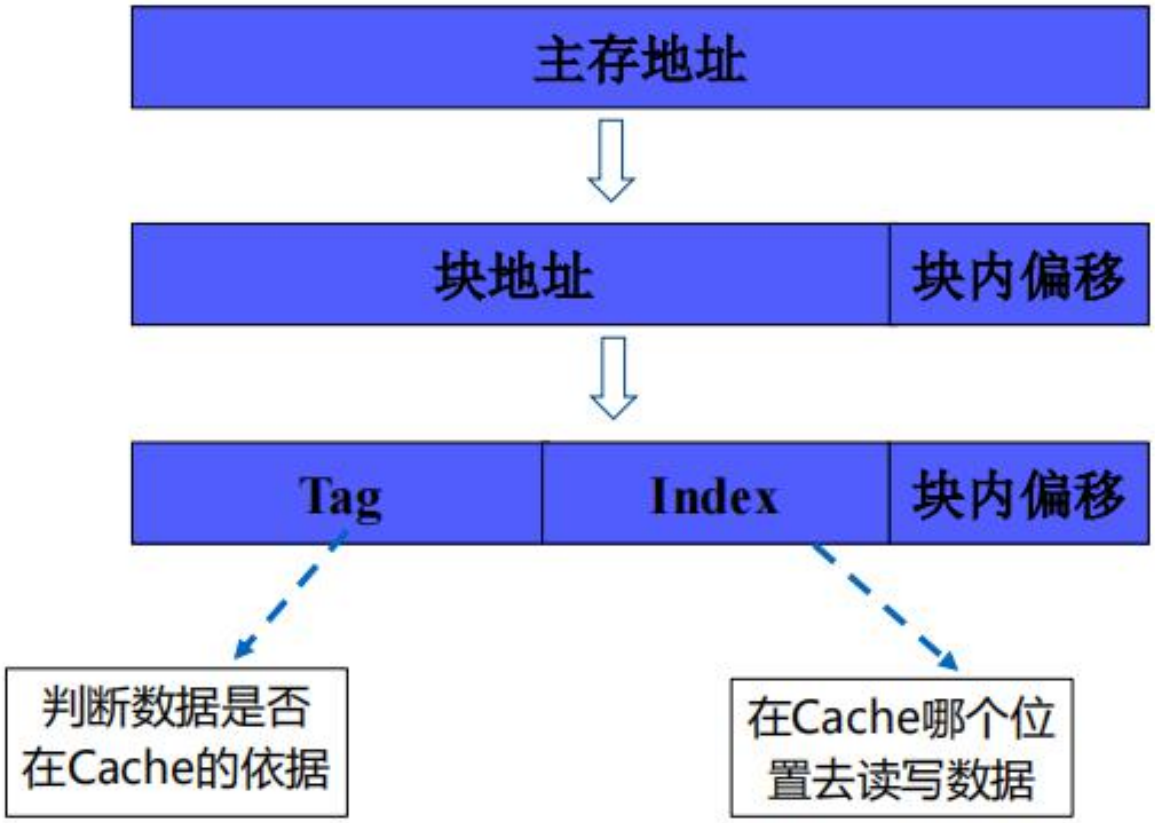
二、主存与Cache的地址映射方式

- **讨论的问题：**如何根据主存地址，判断Cache有无命中并变换为Cache的地址，以便执行读写。有三种地址映射方式：

1. 直接映射
2. 全相联映射
3. 组相联映射

- **讨论前提：**Cache的数据块称为行，主存的数据块称为块，行与块是等长的；主存容量为2^m块，Cache容量为2^c行，每个字块中含2^b字。

Cache地址映射机制



Cache的结构

- Cache被分成若干行，每行的大小与主存块相同
- Cache每行包含四部分，是Cache要保存的信息。Tag从CPU访问主存的地址中剥离得到、Data是与主存交换的数据块、Valid表示Cache中的数据是否有效、 Dirty表示主存中的数据是最新。

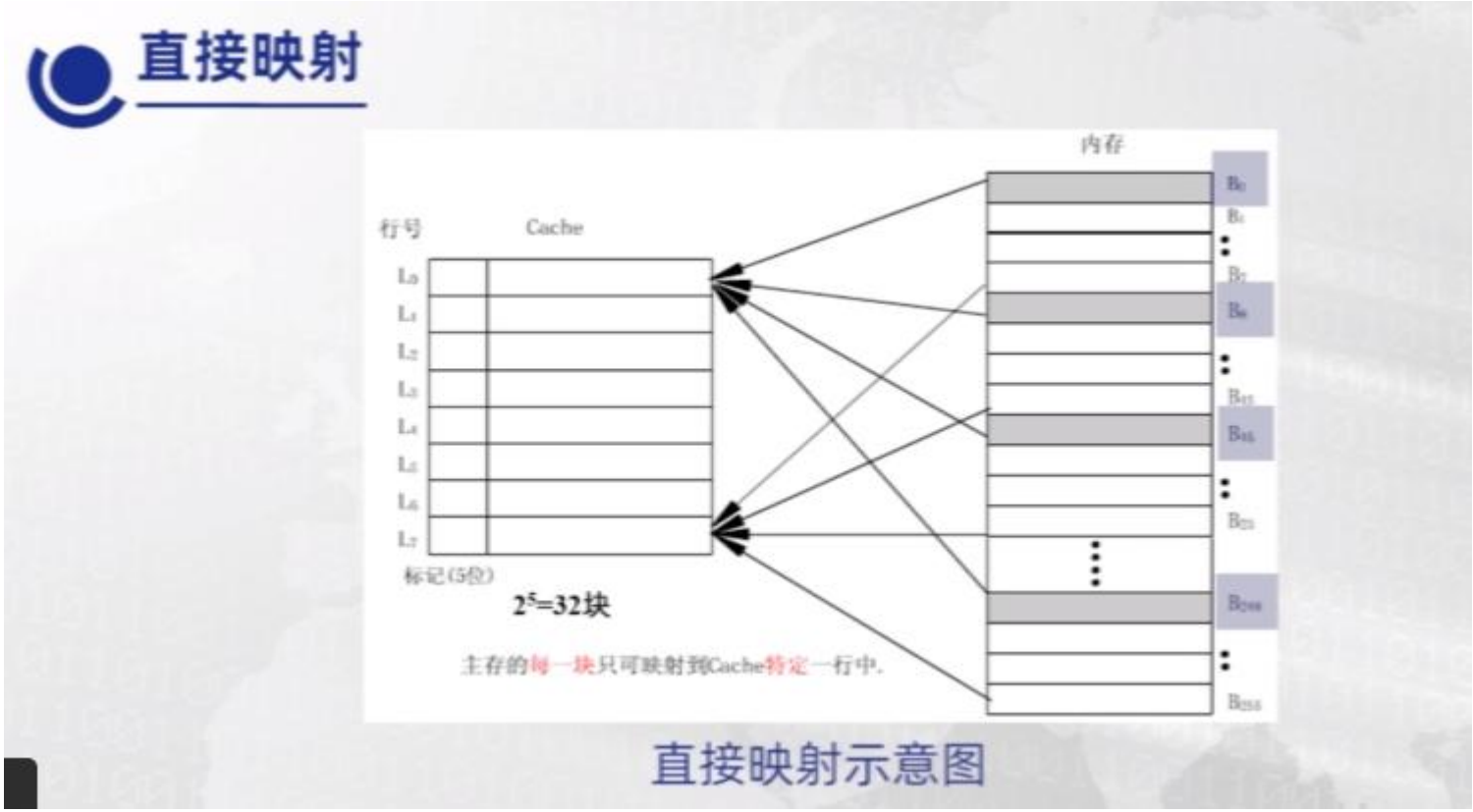
1、直接映射

❖ **特点：** 是一种多对一的映射关系

①优点：映射方式简单，易实现。

②缺点：机制不灵活，Cache命中率低。

1、直接映射


直接映射
内存
行号
Cache
L_0
L_1
L_2
L_3
L_4
L_5
L_6
L_7
标记(5位)
2^5=32块
主存的每一块只可映射到Cache特定一行中.
B_0
B_1
B_7
B_8
B_{15}
B_{16}
B_{23}
B_{248}
B_{255}
直接映射示意图