13015 计算机系统原理

第五章考点解析（1）

13015 计算机系统原理【第五章】

考点1 存储器分类

按信息的可更改性

读/写存储器

只读存储器 → ROM

按断电后信息的可保存性

非易失性（不挥发）存储器 【包括：ROM、磁盘、光盘】

易失性（挥发）存储器 【包括：主存、cache】

主存，由**动态RAM芯片**组成

cache，由静态**RAM芯片**组成，位于主存和CPU之间，存储速度接近CPU的工作速度，用来存放当前CPU经常使用到的指令和数据。

13015 计算机系统原理【第五章】

考点2 存储器的层次化结构

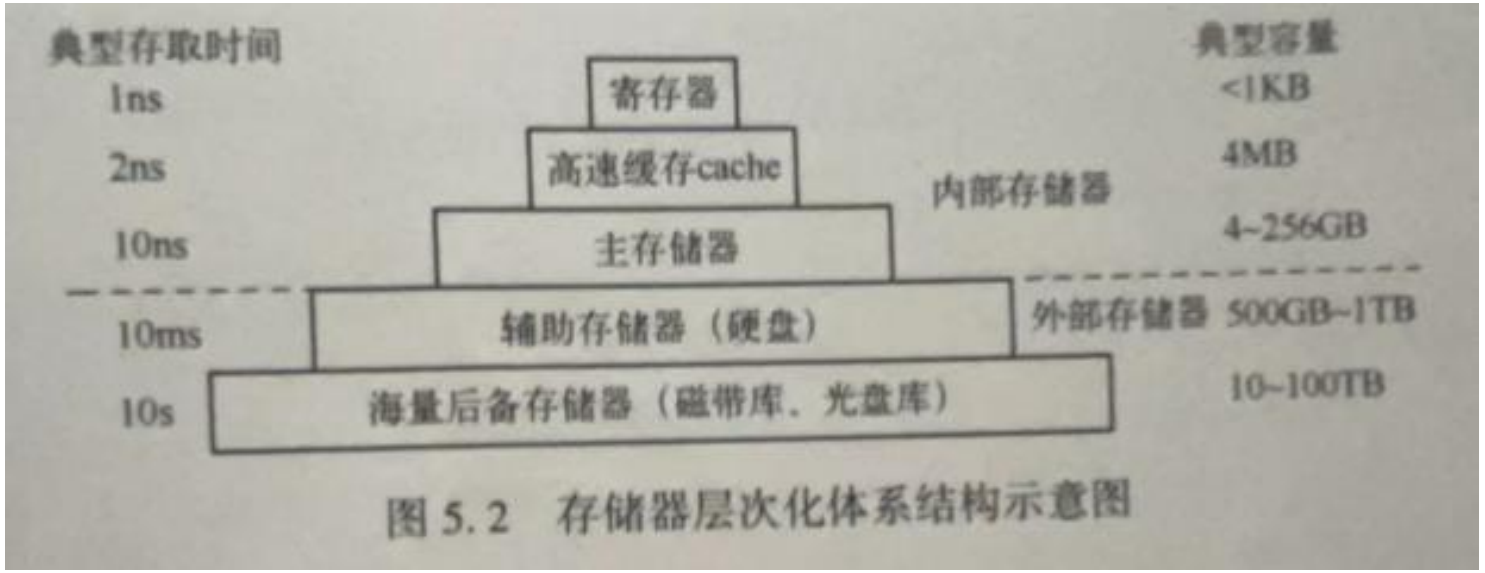


图 5.2 存储器层次化体系结构示意图

Cache和主存之间传送的**主存块（Block）**大小通常为**几十字节**

主存与硬盘之间传送的**页（Page）**大小通常为**几千字节**以上

在层次结构存储系统中，CPU需要访问存储器时，先访问cache，若不在cache，再访问主存，若不在主存，则访问硬盘，此时，从硬盘中读出信息送主存，然后再从主存送cache。

13015 计算机系统原理【第五章】

考点3 程序访问的局部性

程序产生的访存地址往往集中一个很小的范围，这种现象称为**程序访问的局部性**。

包括：

时间局部性：被访问的存储单元在较短时间内很可能被**重复访问**

空间局部性：被访问的存储单元的**临近单元**在较短时间内很可能被访问

13015 计算机系统原理【第五章】

例 5.1 假定数组元素按行优先存放，以下两段伪代码程序段 A 和 B 中：(1) 对于数组 a 的访问，哪一个空间局部性更好？哪一个时间局部性更好？(2) 变量 sum 的空间局部性和时间局部性各如何？(3) 对于指令访问来说，for 循环体的空间局部性和时间局部性如何？

程序段 A

```
1  int sum_array_rows(int a[M][N])
2  {
3      int i, j, sum=0;
4      for (i=0; i<M;i++)
5          for (j=0; j<N;j++)
6              sum+=a[i][j];
7      return sum;
8  }
```

程序段 B

```
1  int sum_array_cols(int a[M][N])
2  {
3      int i, j, sum=0;
4      for (j=0; j<N; j++)
5          for (i=0; i<M; i++)
6              sum+=a[i][j];
7      return sum;
8  }
```

13015 计算机系统原理【第五章】

考点3 程序访问的局部性

从右图可知：指令和二维数组在主存的存放，是**按照行**存放的。

（1）

程序段A

空间局部性：访问顺序与存放顺序**一致**，故**空间局部性好**

时间局部性：**差**，因为每个数组元素都只被访问一次【无重复】

程序段B

空间局部性：访问顺序与存放顺序**不一致**，每次都要**跳过M个元素**，故**没有空间局部性**

时间局部性：**差**，因为每个数组元素都只被访问一次【无重复】

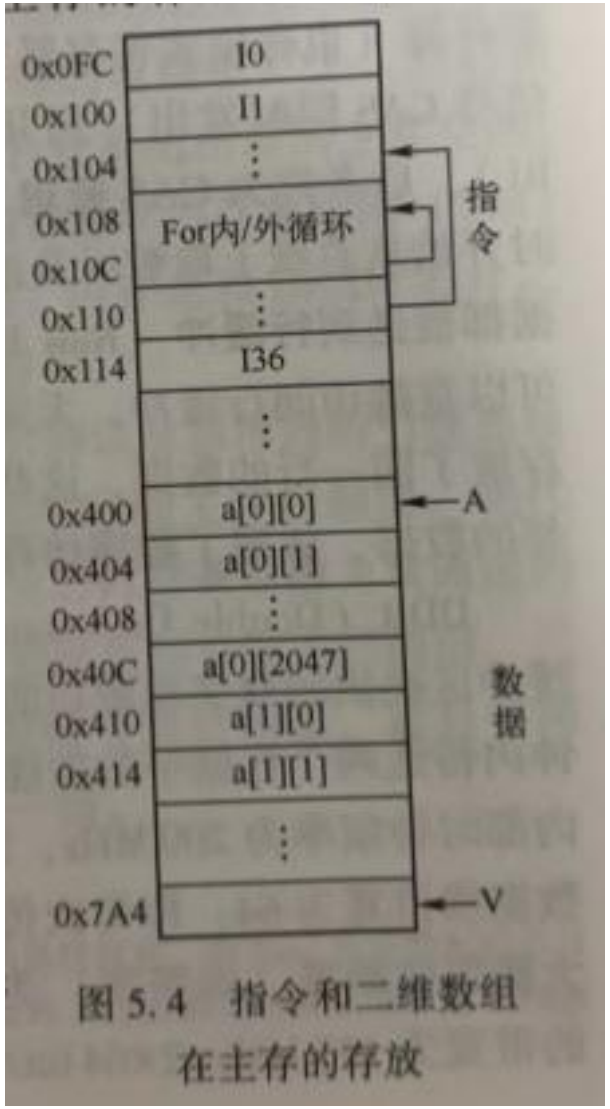


图 5.4 指令和二维数组在主存的存放

13015 计算机系统原理【第五章】

考点3 程序访问的局部性

(2)

变量sum(程序段A和B都一样)

空间局部性：单个变量没有意义

时间局部性：**好**，因为每次循环都要被访问，【重复访问】

(3)

for循环体(程序段A和B都一样)

空间局部性：访循环体内指令按序连续存放，所以**空间局部性好**

时间局部性：因都是**M×N**次，所以**时间局部性好**

13015 计算机系统原理【第五章】

考点4 Cache高速缓冲存储器

整个访存过程如下：

判断信息是否在cache

若**是**，则直接从cache取信息

若**否**，则从主存取一个主存块到cache，如果对应的cache行已满，则需要替换cache中的信息。

注意：cache中的内容是主存中部分内容的**副本。**

每个cache行有一个**有效位**

清0淘汰某cache行中的主存块

装入一个新主存块时**置1**

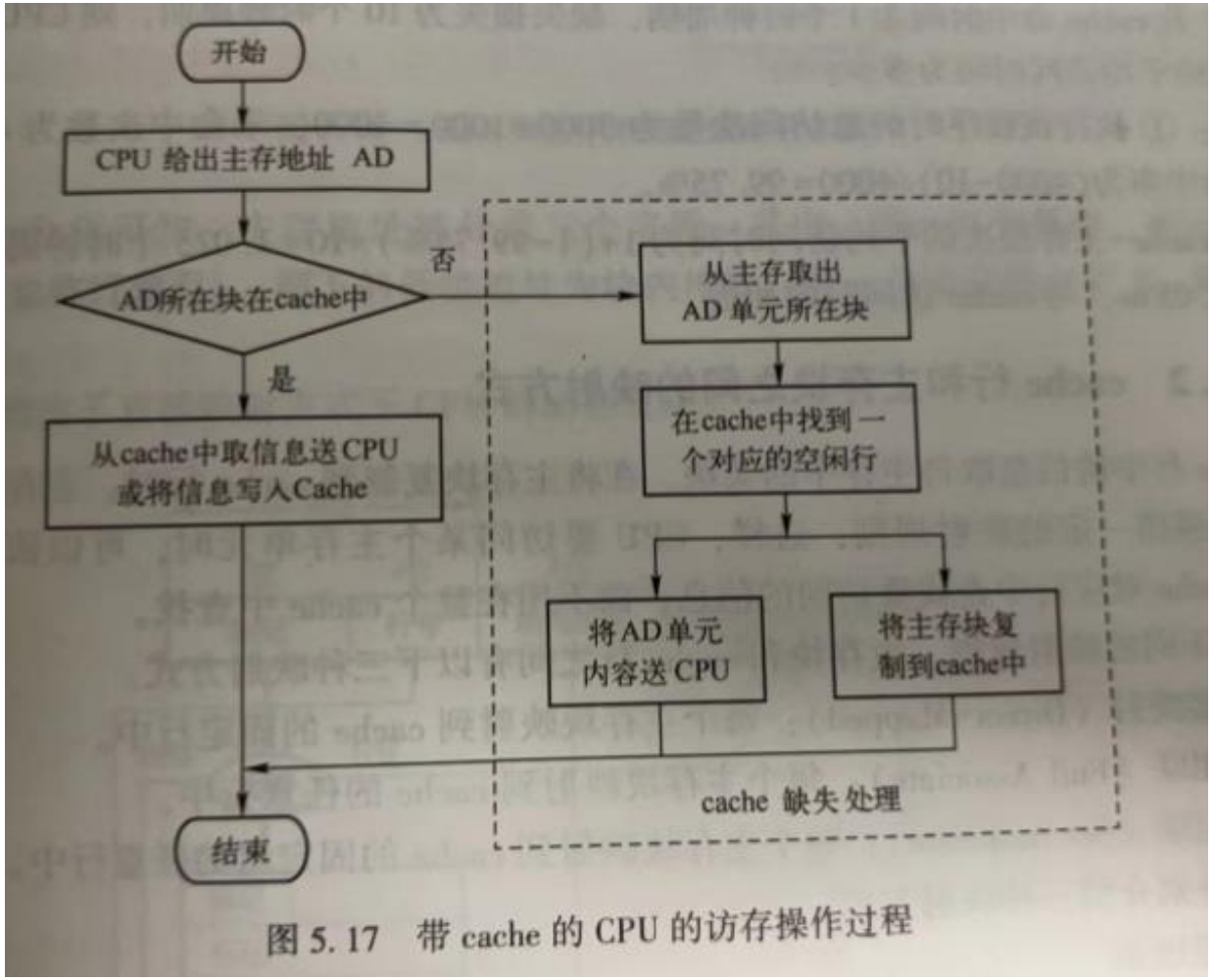


图 5.17 带 cache 的 CPU 的访存操作过程

13015 计算机系统原理【第五章】

考点4 Cache高速缓冲存储器

若CPU访问单元所在的块

若在cache中，则称**cache命中**，命中概率称为命中率p，命中时间Tc

它等于**命中次数**与**访问总次数**之比，即：**命中次数/访问总次数**

命中时，CPU在cache中直接存取信息，所用时间即为cache访问时间**Tc**，称**命中时间**

若不在cache中，则称**不命中**或**缺失**，其概率称为缺失率，访问时间Tm

它等于**不命中次数**与**访问总次数**之比，即：**不命中次数/访问总次数**

缺失时，需要从主存读取一个主存块送cache，并同时将所需信息送CPU，因此，所用时间为主存访问时间**Tm**和cache访问时间**Tc**之**和**。

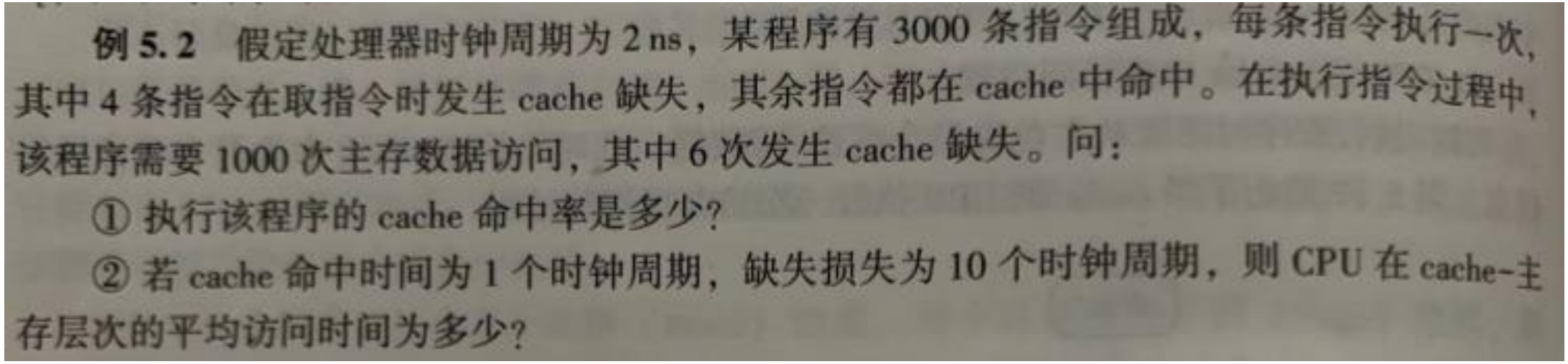
注意：通常把**Tm**称为**缺失损失**

CPU在**cache-主存层次**的**平均访问时间**为:

$Ta = p \times Tc + (1 - p) \times (Tm + Tc) = \mathbf{Tc + (1 - p) \times Tm}$

13015 计算机系统原理【第五章】

考点4 Cache高速缓冲存储器



例 5.2 假定处理器时钟周期为 2 ns，某程序有 3000 条指令组成，每条指令执行一次，其中 4 条指令在取指令时发生 cache 缺失，其余指令都在 cache 中命中。在执行指令过程中，该程序需要 1000 次主存数据访问，其中 6 次发生 cache 缺失。问：

① 执行该程序的 cache 命中率是多少？

② 若 cache 命中时间为 1 个时钟周期，缺失损失为 10 个时钟周期，则 CPU 在 cache-主存层次的平均访问时间为多少？

①取指3000条指令，每条执行一次，共3000次；执行1000次，总共3000 + 1000 = 4000次，未命中 4 + 6 = 10次，所以：cache命中率是：(4000 - 10) / 4000 = 99.75%。

②由题可知：Tc = 1，Tm = 10

CPU在**cache-主存层次**的**平均访问时间**为:

Ta = **Tc + (1 - p) × Tm** = 1 + (1 - 99.75%) × 10 = 1.025个时钟周期，即1.025 × 2 = 2.05ns

13015 计算机系统原理【第五章】

考点5 直接映射

每个**主存块**映射到cache的**固定行**中。也叫**模映射**，其映射关系如下：

cache行号 = 主存块号 mod cache行数

例如：若cache有16行(数)，根据100 mod 16 = 4可知，主存第**100块**映射到cache第4行(号)

例 5.3 假定 cache 采用直接映射方式，主存块大小为 64 B，按字节编址。cache 数据区大小为 1 KB，主存空间大小为 256 KB。问：主存地址如何划分？要求用图表示主存块和 cache 行之间的映射关系，假定 cache 当前为空，说明 CPU 对主存单元 0240CH 的访问过程。

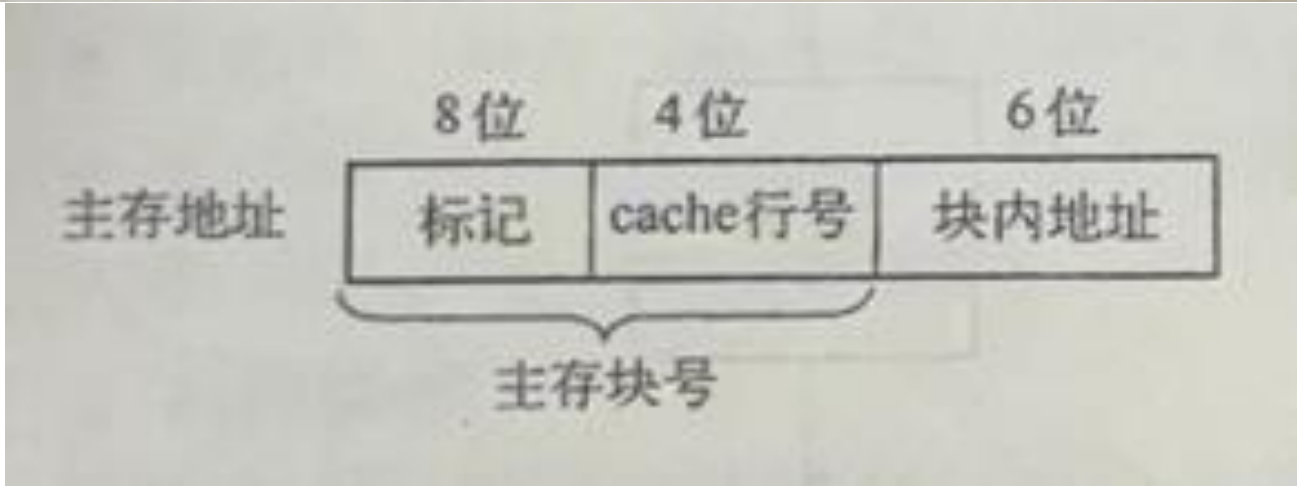
解题思路：[明确以下四个概念]

① **确定主存地址位数；**

② **确定块内地址位数；**

③ **确定cache行号位数；**

④ **确定标记位数；** **①-②-③**



13015 计算机系统原理【第五章】

考点5 直接映射

解题思路：[明确以下四个概念]

① 确定**主存地址**位数；例如：主存空间大小为256KB，所以$\log_2(256\times1024)=$ **18位**

② 确定**块内地址**位数；例如：主存块大小为64B，所以$\log_2 64 =$ **6位**

③ 确定**cache行号**位数；例如：cache数据区大小为1KB，主存块大小为64B，所以1KB/64B=16行，所以$\log_2 16 =$ **4位**

④ 确定**标记**位数；①-②-③ = **8位**

注意：标记，即：块群

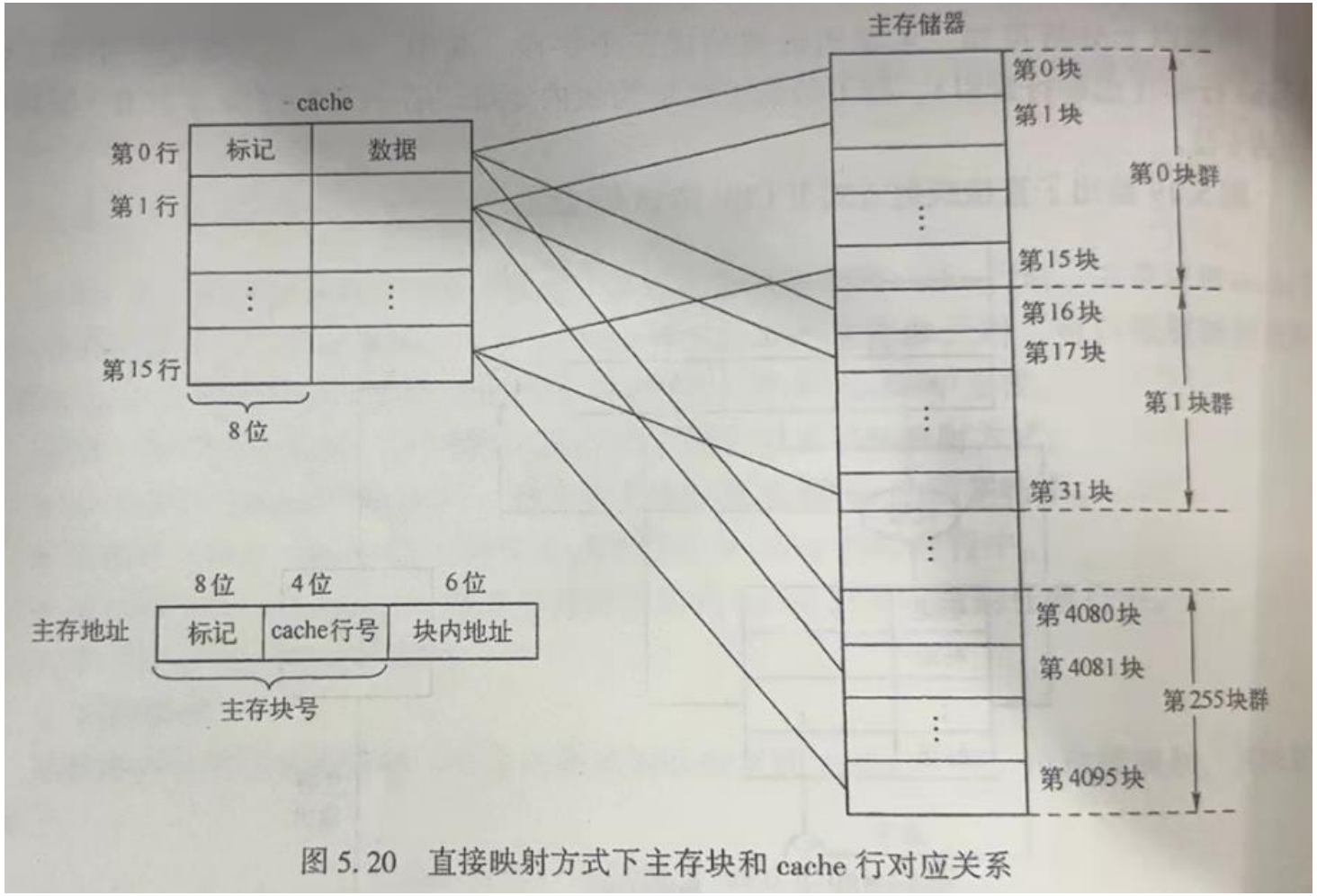


图 5.20 直接映射方式下主存块和 cache 行对应关系

13015 计算机系统原理【第五章】

考点5 直接映射

CPU对主存单元0240CH的访问过程如下：

0000 0010 0100 0000 1100

保留18位，最高位的00舍掉，为：00 0010 0100 0000 1100

按照 8位 + 4位 + 6位 的方式截取：

00 0010 01	00 00	00 1100

所以：主存块号是前12位，00 0010 0100 00，二进制转十进制是：144（第144块）

是**第9块群**中的**第0块**，映射到的**cache行号为0000（第0行）**。

由题知假定cache当前为空，访问0240CH单元的**过程如下**：

首先根据地址中间4位cache行号0000，找到cache第0行，因为**cache当前为空**，所以，每个cache行的有效位都为0，因此，不管第0行的标志是否等于00001001，**都不命中**。此时，将0240CH单元所在的主**存第144块复制到cache第0行**，并置**有效位为1**，置**标记**为00001001（表示信息取自主存第9块群）助记：**cache空不命中**，读数，装cache，置1置**标记。**

13015 计算机系统原理【第五章】

考点6 全相联映射

每个**主存块**映射到cache的**任意行**中。

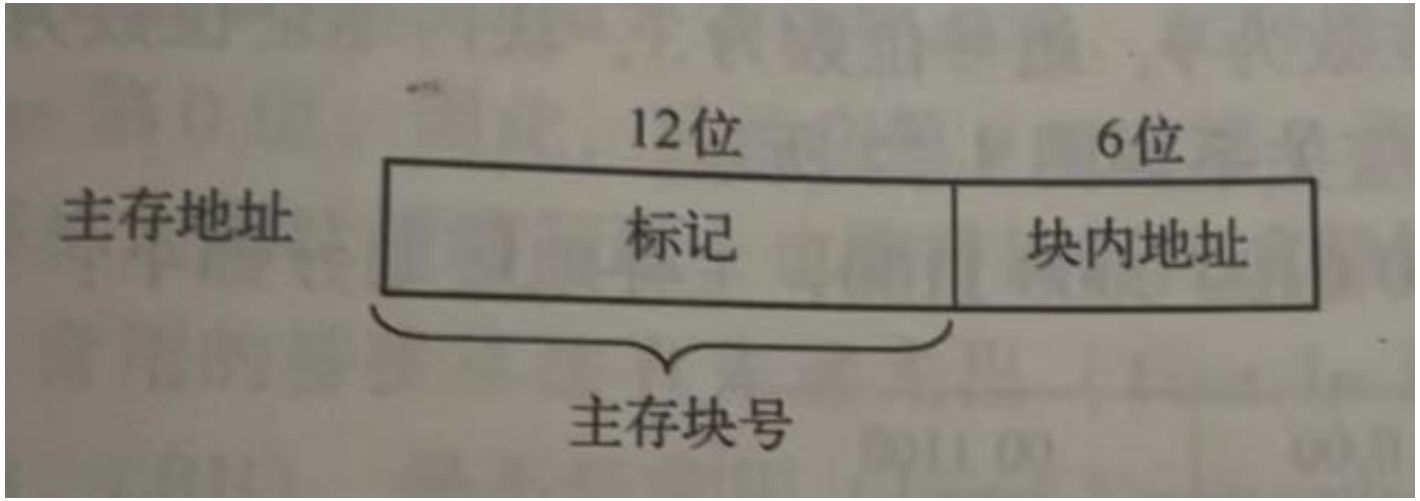
例 5.4 假定 cache 采用全相联方式，主存块大小为 64 B，按字节编址。cache 数据区大小为 1 KB，主存空间大小为 256 KB。问：主存地址如何划分？要求用图表示主存块和 cache 行之间的映射关系，并说明 CPU 对主存单元 0240CH 的访问过程。

解题思路：[明确以下三个概念]

① **确定主存地址位数；**

② **确定块内地址位数；**

③ **确定标记位数；** **①-②**



13015 计算机系统原理【第五章】

考点6 全相联映射

解题思路：[明确以下三个概念]

① 确定**主存地址**位数；例如：主存空间大小为**256KB**，所以$\log_2(256\times1024)=$ **18位**

② 确定**块内地址**位数；例如：主存块大小为**64B**，所以$\log_2 64 =$ **6位**

③ 确定**标记**位数；①-② = **12位**

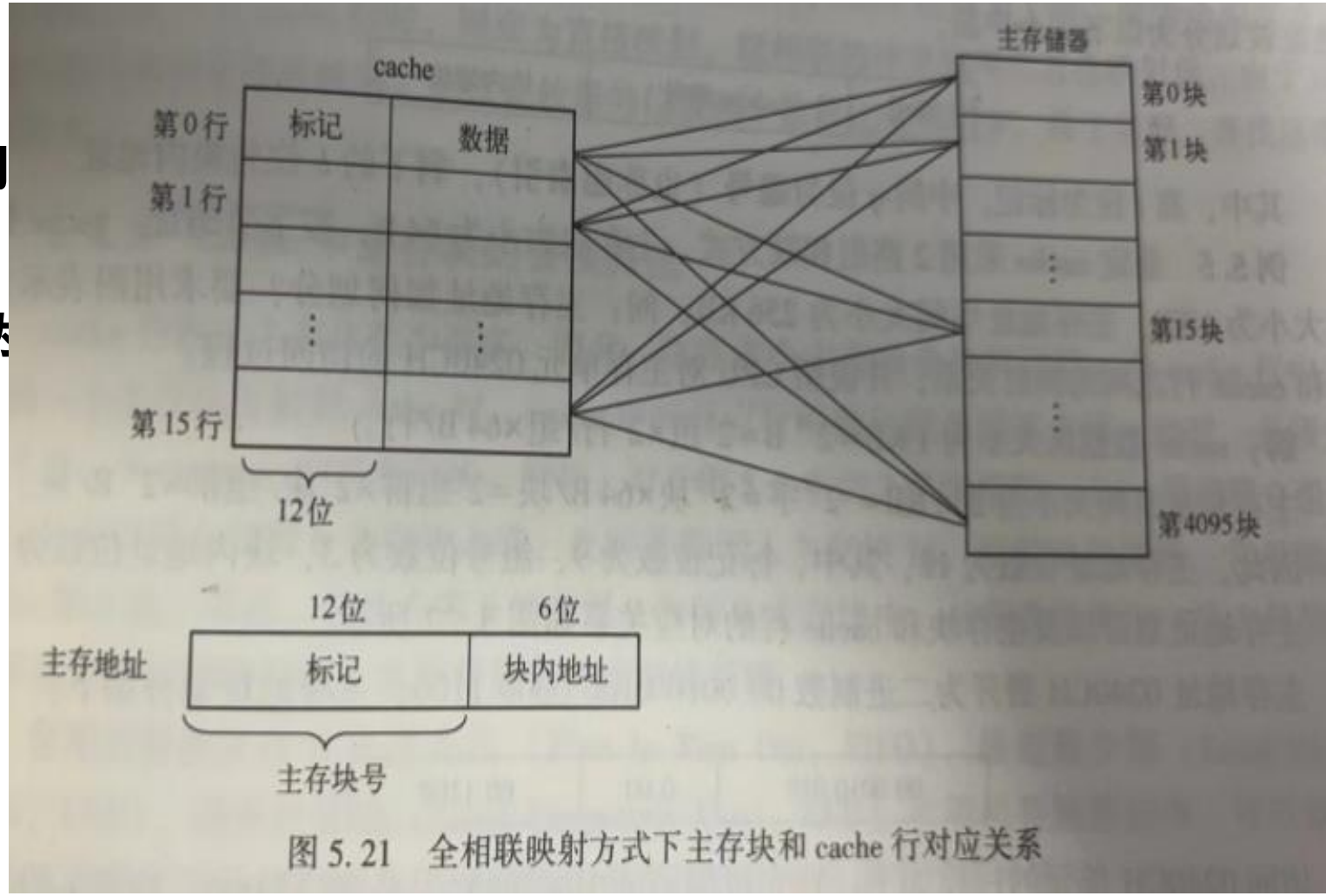


图 5.21 全相联映射方式下主存块和 cache 行对应关系

13015 计算机系统原理【第五章】

考点6 全相联映射

CPU对主存单元0240CH的访问过程如下：

00 0010 0100 0000 1100

按照 12位 + 6位 的方式截取：

00 0010 0100 00	00 1100

所以：标记是前12位，00 0010 0100 00，二进制转十进制是：144（第144块）

由题知假定cache当前为空，访问0240CH单元的**过程如下**：

首先将高12位标记与每个cache行标记进行比较，若有一个相等且有效位为1，则命中，此时，CPU根据块内地址**00 1100**从该行中取出信息；**若都不相等，则不命中**，此时，将0240CH单元所在的主**存第144块读出，并装入任意一个空闲cache行中**，置**有效位为1**，置**标记**为000010010000（表示信息取自主存第144块）

助记：比较，相等1**命中（行）**，否则不命中，读数，装cache，置1置**标记。**

13015 计算机系统原理【第五章】

考点7 组相联映射

每个**主存块**映射到cache的**固定组的任意行**中。其映射关系如下：

cache组号 = 主存块号 mod cache组数

例如：若1KB的cache划分为2^3组×2行/组×64B/行，则主存第**100块**映射到cache第4组的任意一行中，因为100 mod 2^3 = 4。

例 5.5 假定 cache 采用 2 路组相联方式，主存块大小为 64 B，按字节编址。cache 数据区大小为 1 KB，主存地址空间大小为 256 KB。问：主存地址如何划分？要求用图表示主存块和 cache 行之间的映射关系，并说明 CPU 对主存单元 0240CH 的访问过程。

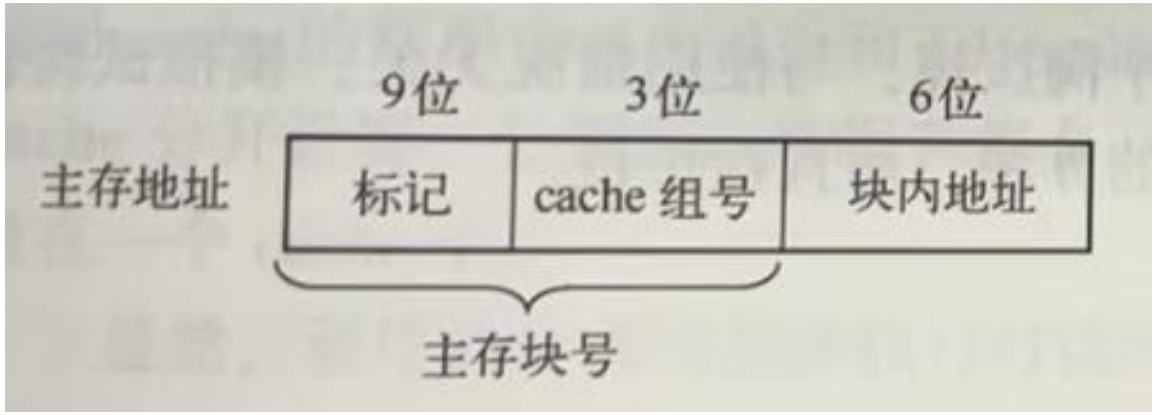
解题思路：[明确以下四个概念]

① 确定**主存地址**位数；

② 确定**块内地址**位数；

③ 确定**cache组号**位数；

④ 确定**标记**位数；①-②-③



13015 计算机系统原理【第五章】

考点7 组相联映射

解题思路：[明确以下四个概念]

① 确定主存地址位数；例如：主存空间大小为256KB，所以$\log_2(256\times1024)=$ 18位

② 确定块内地址位数；例如：主存块大小为64B，所以$\log_2 64 =$ 6位

③ 确定cache组号位数；例如：cache数据区大小为1KB，主存块大小为64B，所以1KB/64B=16行，2路组相联，所以有16/2=8组

$\log_2 8 =$ 3位

④ 确定标记位数；①-②-③ = 9位

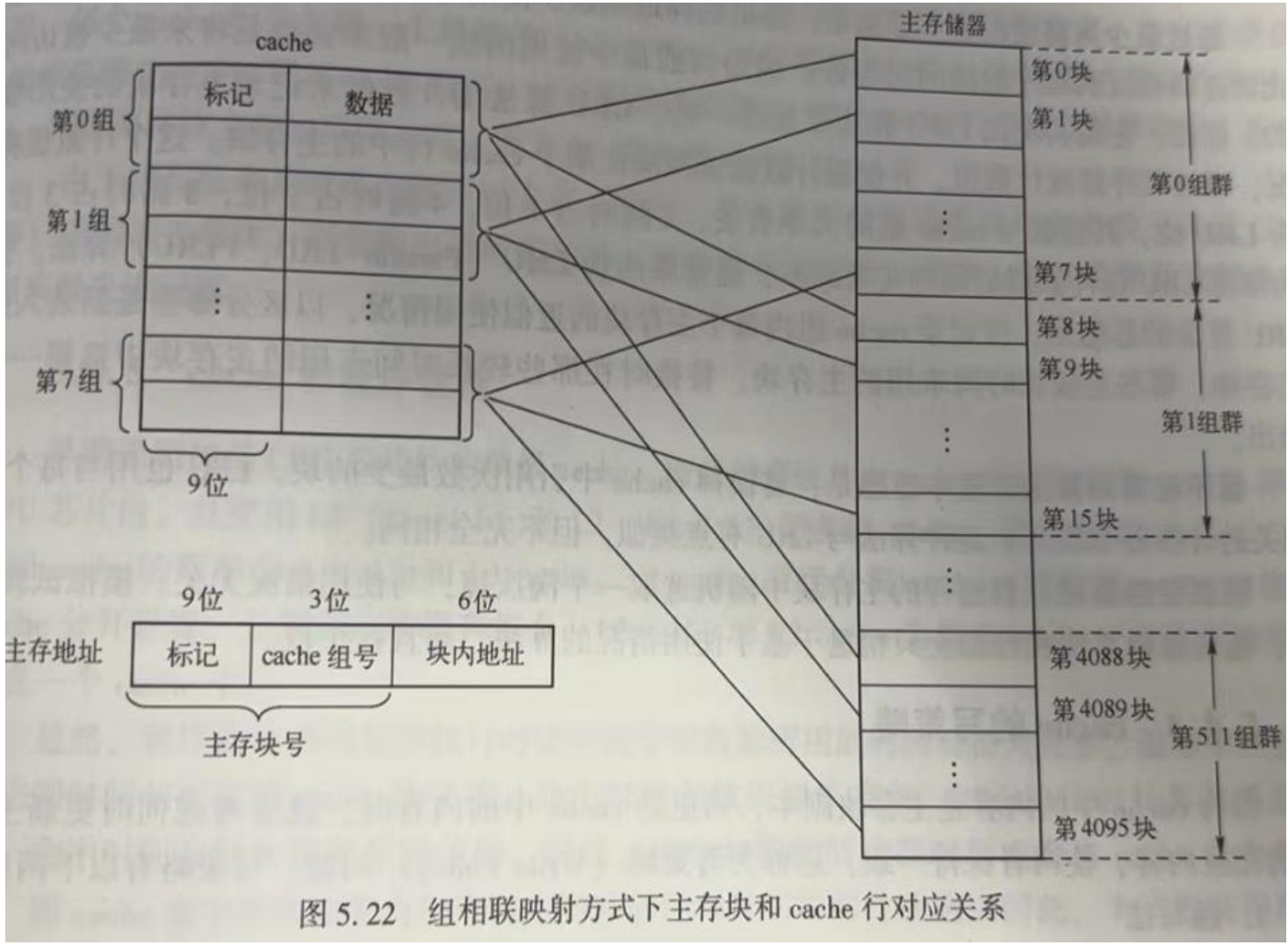


图 5.22 组相联映射方式下主存块和 cache 行对应关系

13015 计算机系统原理【第五章】

考点7 组相联映射

CPU对主存单元0240CH的访问过程如下：

00 0010 0100 0000 1100

按照 9位 + 3位 + 6位 的方式截取：

00 0010 010	0 00	00 1100

所以：主存块号是前12位，00 0010 0100 00，二进制转十进制是：144（第144块）

由题知假定cache当前为空，访问0240CH单元的**过程如下**：

首先根据cache组号**000**，找到cache第0组，将标记和第0组的两个cache行的标记进行比较，若有一个相等且有效位为1，则**命中**。此时，根据低6位块内地址从对应行中取出单元内容送CPU;都都不相等但有效位为0，则不命中。此时，将主**存第144块复制到cache第0组的任意一个空行中**，并置**有效位为1**，置**标记**为000010010（表示信息取自主存第18组群）

助记：比较，相等1**命中（组）**，否则不命中，读数，装cache，置1置**标记。**

13015 计算机系统原理【第五章】

考点8 直接映射、全相联映射和组相联映射优缺点

直接映射

优点：容易实现。

缺点：命中时间短，但由于多个主存块会映射同一个cache行，当访问集中在这些主存块时，就会引起频繁的调进调出，即使其他cache行都空闲，也无法充分利用。

例如：例5.3，第0块和第16块都只能映射到cache第0行，即使其他cache行都空闲，也无法充分利用。

全相联映射

优点：只要有空闲cache行，就不会发生冲突，因而块冲突概率低。

缺点：时间开销和所用元件开销都较大，因此全相联方式不适合容量较大的cache。

组相联映射

结合了**直接映射**和**全相联**的**优点**。

13015 计算机系统原理【第五章】

考点9 cache中主存块的替换算法

cache行数比主存块数少得多，因此，往往多个主存块会映射到同一个cache行中。当新的主存块复制到cache时，cache中的对应行可能已经全部被占满，此时，必须选择淘汰掉一个cache行中的主存块。具体如何选择称为**替换算法**。常用的替换算法有如下几种：

先进先出算法

基本思想：选择最早装入cache的主存块被替换掉。【不能正确反映程序的访问局部性】

最近最少使用算法

基本思想：选择近期最少使用的主存块被替换掉。【能正确反映程序的访问局部性】

最不经常使用算法

基本思想：选择替换掉cache中引用次数最少的块。

随机替换算法

基本思想：随机选择一个被替换掉。

13015 计算机系统原理【第五章】

考点10 cache的写策略

因为cache中的内容是主存块的副本，当更新cache中的内容时，就熬考虑何时更新主存中的相应内容，使两者保持一致，这称为 **写策略**问题。写策略有以下两种： **（2404考期-填空题）**

通写法

基本思想：当CPU写入数据命中时，**同时**将数据**写入主存和缓存**。这种策略可以保证缓存和主存中的数据一致性。不命中时，则**先写入主存**，然后以下两种情况

写分配法：分配一个cache行并装入更新后的主存块，**充分利用空间局部性**

非写分配法：不将主存块装入cache，**没有充分利用空间局部性**

回写法

基本思想：当CPU写入数据命中时，**只**将数据**写入缓存**，**不**立即**写入主存**。当缓存行被替换时，才将数据写回主存。写缺失时，分配一个cache行并装入主存块，然后更新该行的内容。**注意：回写法通常与写分配法组合使用。**

13015 计算机系统原理【第五章】

考点11 cache和程序性能

例 5.6 某计算机的主存空间大小为 256 MB，按字节编址。指令 cache 和数据 cache 分离，两种 cache 均有 8 个 cache 行，主存与 cache 交换的块大小为 64 B，数据 cache 采用 2 路组相联、通写法和 LRU 替换算法。现有两个功能相同的程序 A 和 B，其伪代码如图 5.23 所示。

程序 A：

```
int a[256][256];
……
int sum_array1 ( ){
        int i, j, sum = 0;
        for ( i = 0; i < 256; i++)
            for (j = 0; j < 256; j++)
                sum += a[i][j];
        return sum;
}
```

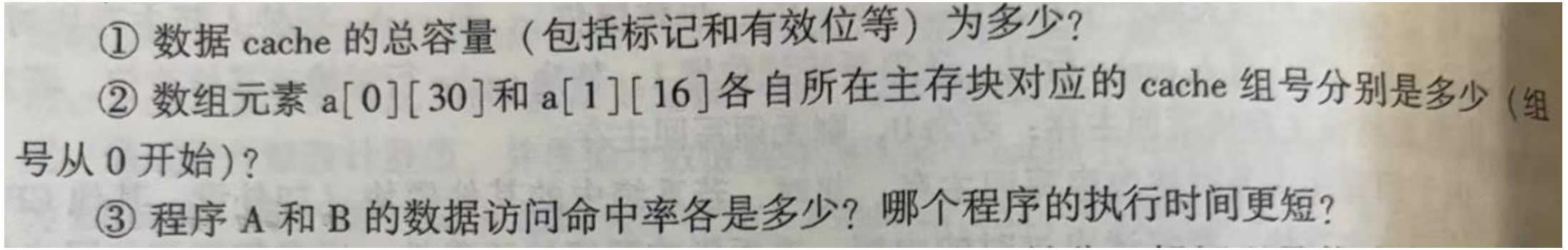
程序 B：

```
int a[256][256];
……
int sum_array2 ( ){
        int i, j, sum = 0;
        for ( j = 0; j < 256; j++)
            for ( i = 0; i < 256; i++)
                sum += a[i][j];
        return sum;
}
```

图 5.23 例 5.6 中的伪代码程序

13015 计算机系统原理【第五章】

考点11 cache和程序性能


① 数据 cache 的总容量（包括标记和有效位等）为多少？

② 数组元素 a[0][30]和 a[1][16]各自所在主存块对应的 cache 组号分别是多少（组号从 0 开始）？

③ 程序 A 和 B 的数据访问命中率各是多少？哪个程序的执行时间更短？

①每个cache行除用于存放主存块外，还有有效位、标记以及修改位和使用位（如LRU位）等控制位。

由题可知用**通写法**，无须修改位，那就剩下有效位、标记和使用位（如LRU位）

解题思路：[明确以下四个概念]

① 确定主存地址位数；例如：主存空间大小为256MB，所以$\log_2(256\times1024\times1024)$= 28位

② 确定块内地址位数；例如：主存块大小为64B，所以$\log_2 64$ = 6位

③ 确定cache组号位数；例如：cache数据共8行，2路组相联，所以有8/2=4组，所以$\log_2 4$ = 2位

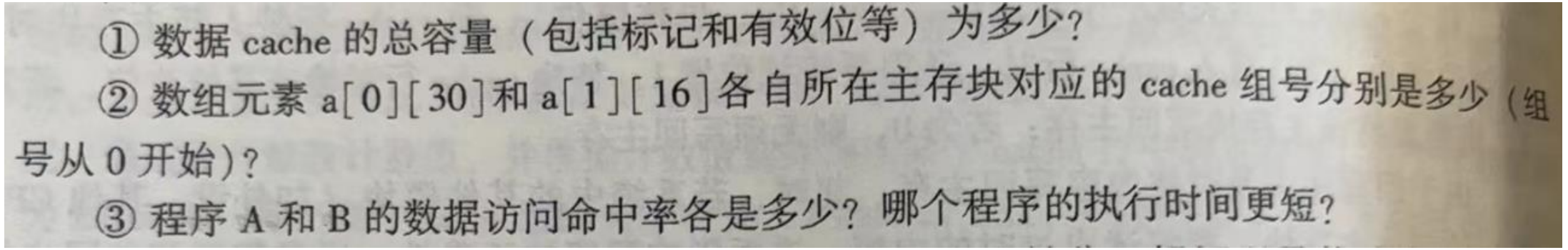
④ 确定标记位数；①-②-③ = 20位

所以cache的总容量为（64×8 + 20 + 1 + 1）× 8 = 4272位 = 534B

说明：（主存块 + 标记 + 有效位 + 使用位）

13015 计算机系统原理【第五章】

考点11 cache和程序性能


① 数据 cache 的总容量（包括标记和有效位等）为多少？
② 数组元素 a[0][30]和 a[1][16]各自所在主存块对应的 cache 组号分别是多少（组号从 0 开始）？
③ 程序 A 和 B 的数据访问命中率各是多少？哪个程序的执行时间更短？

②数组元素的**地址**计算方法：

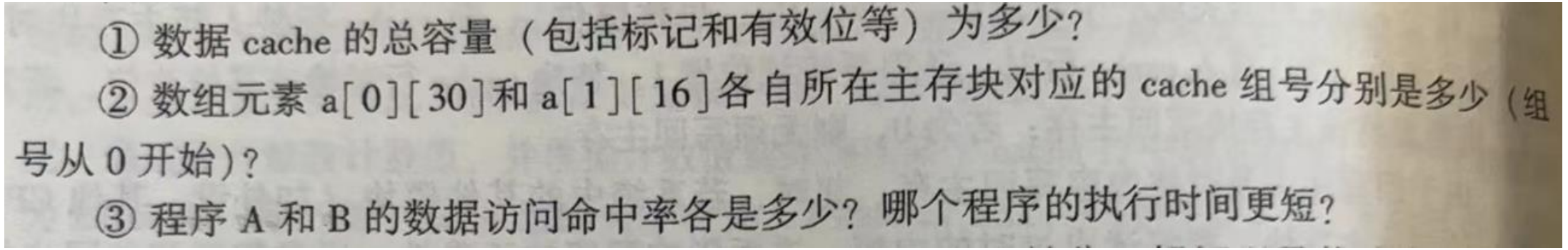
（行 × 总列 + 列）× 字节数 + 起始地址，这是int数组，所以字节数 = 4

a[0][30]：地址（0 × 256 + 30）× 4 + 320 = 440，对应的主存块号：440 / 64 = 6（取整），故cache组号为 6 mod 4 = 2

a[1][16]：地址（1 × 256 + 16）× 4 + 320 = 1408，对应的主存块号：1408 / 64 = 22（取整），故cache组号为 22 mod 4 = 2

13015 计算机系统原理【第五章】

考点11 cache和程序性能


① 数据 cache 的总容量（包括标记和有效位等）为多少？
② 数组元素 a[0][30]和 a[1][16]各自所在主存块对应的 cache 组号分别是多少（组号从 0 开始）？
③ 程序 A 和 B 的数据访问命中率各是多少？哪个程序的执行时间更短？

③编译时j，j，sum均分配在寄存器中，故数据访问命中率仅需要考虑数组a的访问情况。

程序A **（数组访问顺序与存放顺序相同，故依次访问的数据元素位于相邻单元）**

内存块数：256×256×4B / 64B = 4096块

每个内存块存放的元素：64B / 4B = 16个元素，顺序存放 a[0][0] 、 a[0][1]、 a[0][2] ... a[0][15]

共访存16次，**其中第一次不命中，以后15次都命中**，故15 / 16 = **93.75%**（每个主存块的命中情况都一样）程序B **（数组访问顺序与存放顺序不同）**

a[i][j]和a[i+1][j]，比如：a[0][0]和a[1][0]之间的存储地址相差256×4=1024B，1024B/64B=16块，因为16 mod 4 = 0，映射到同一个cache组，每个cache组有2行，后面的主存块替换前面的，故命中率为0

谢 谢 大 家