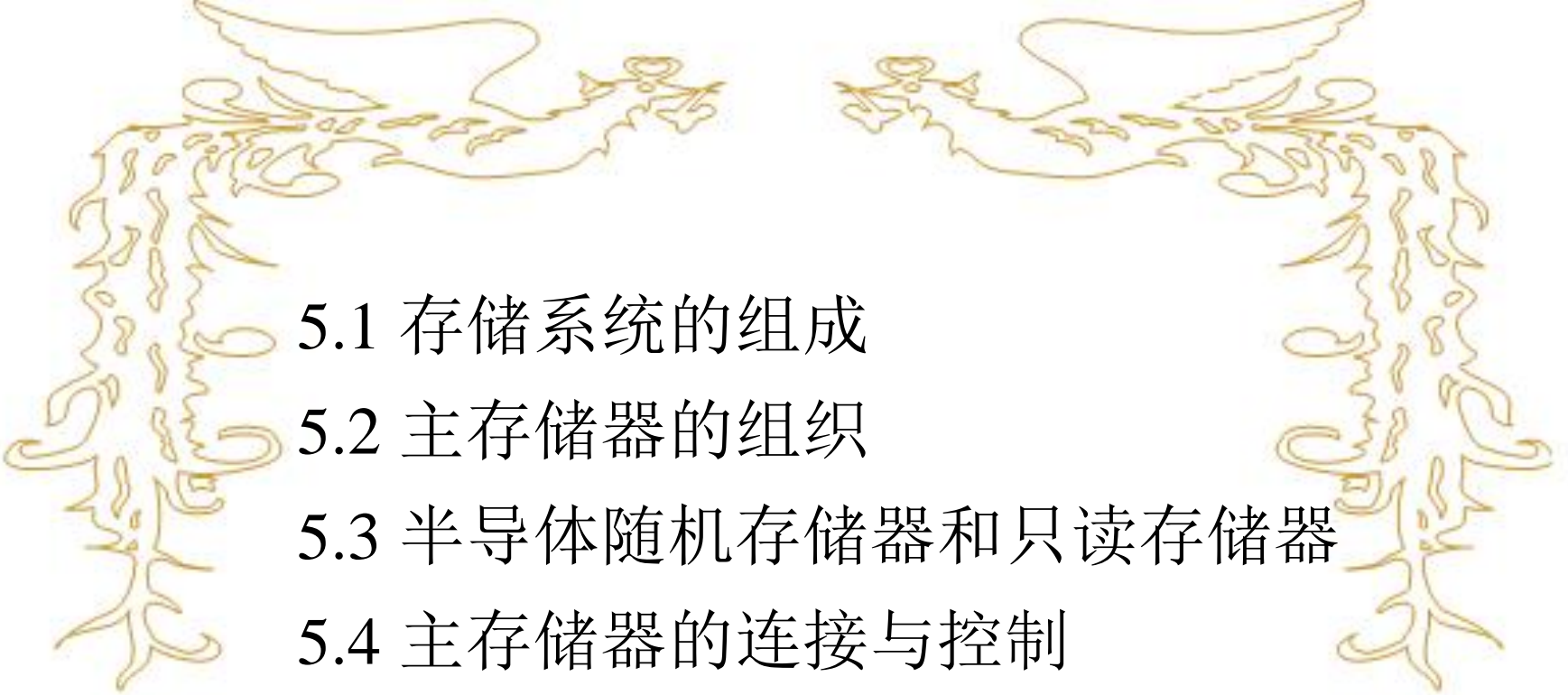



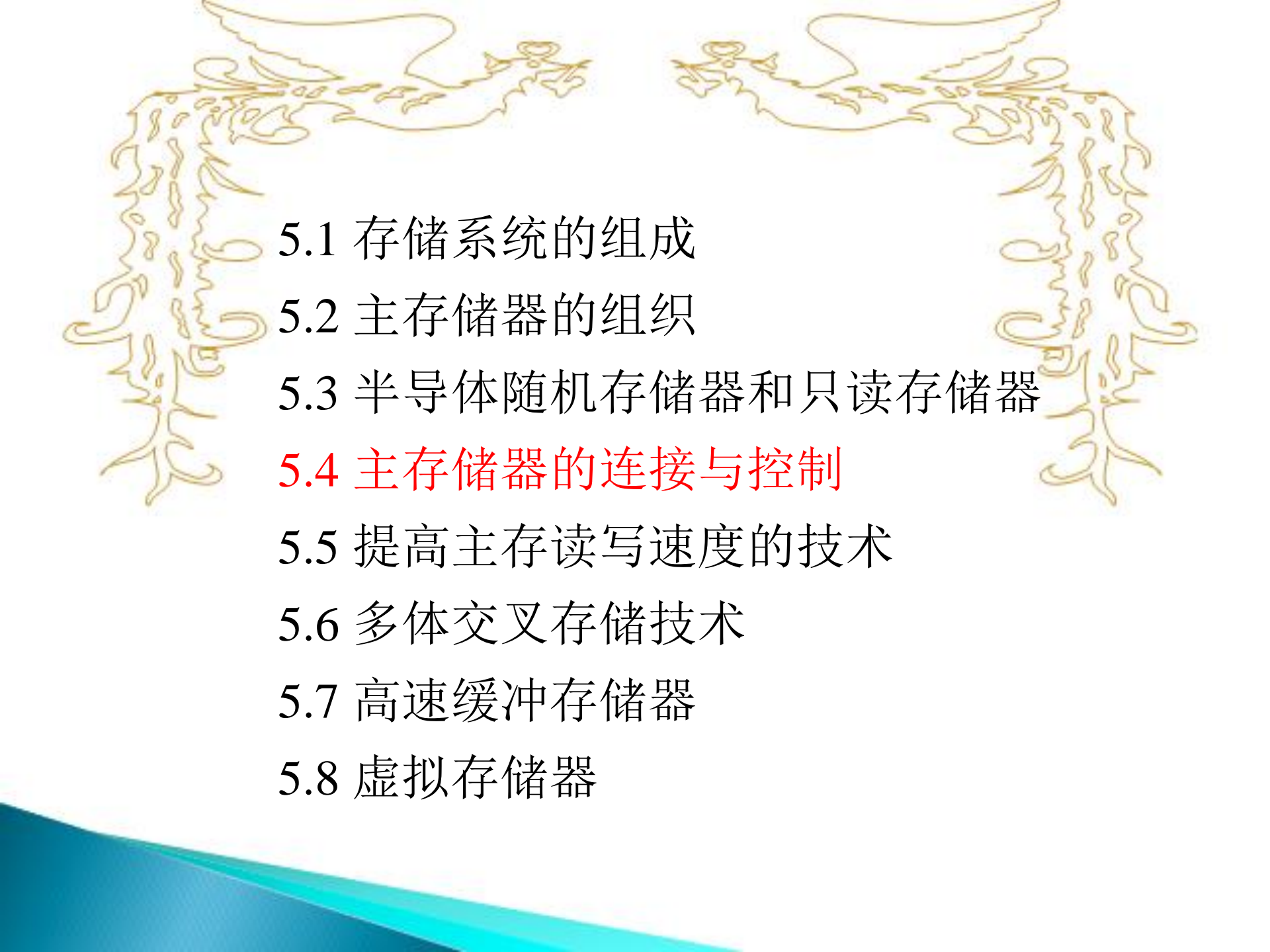


河北师范大学软件学院
Software College of Hebei Normal University

计算机组成原理

第五章 存储系统和结构

- 
- 5.1 存储系统的组成
 - 5.2 主存储器的组织
 - 5.3 半导体随机存储器和只读存储器
 - 5.4 主存储器的连接与控制
 - 5.5 提高主存读写速度的技术
 - 5.6 多体交叉存储技术
 - 5.7 高速缓冲存储器
 - 5.8 虚拟存储器
- 



5.1 存储系统的组成

5.2 主存储器的组织

5.3 半导体随机存储器和只读存储器

5.4 主存储器的连接与控制

5.5 提高主存读写速度的技术

5.6 多体交叉存储技术

5.7 高速缓冲存储器

5.8 虚拟存储器



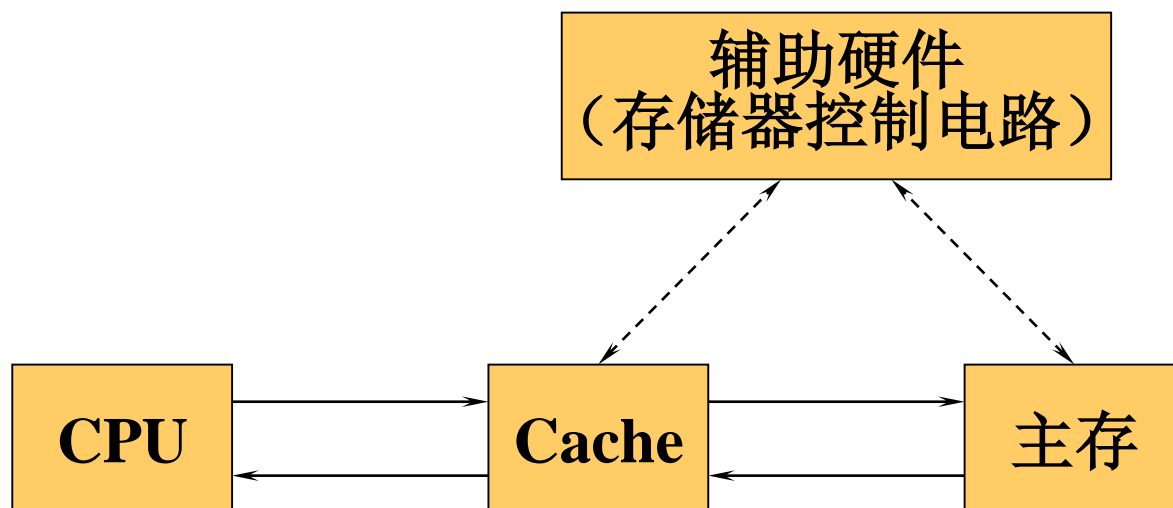
5.7.1 存储系统的层次结构

为了解决容量、速度和价格之间的矛盾，把各种不同存储容量，不同存取速度，不同价格的存储器，按一定的体系结构组织起来，使所存放的程序和数据按层次分布在各存储器中，形成一**多层次**的存储系统。

5.7.2 高速缓冲存储器

1.Cache—主存层次

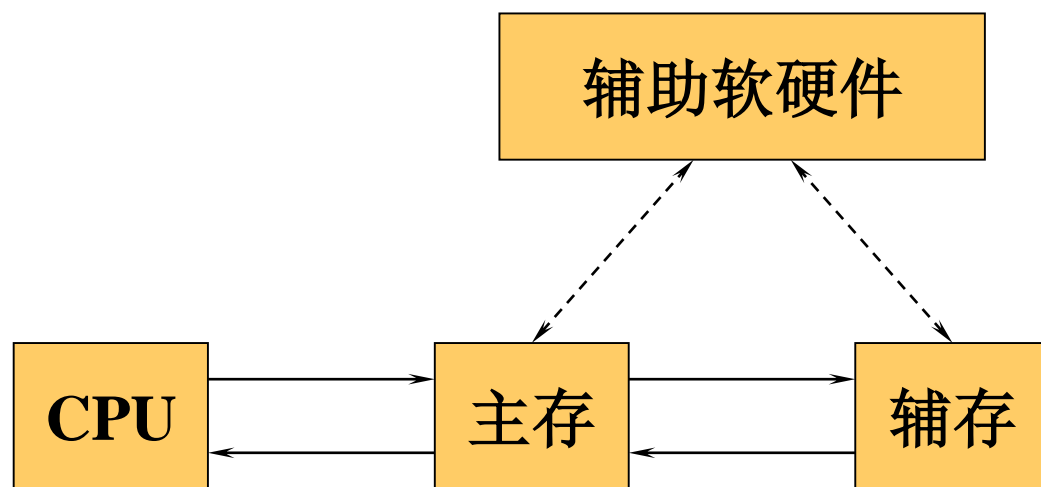
在CPU和主存之间设置了一级容量不大，但速度很高的高速缓冲存储器（Cache），简称高速缓存。



- ▶ 从整体看，Cache—主存层次的存取速度接近于Cache的存取速度，但容量是主存的，每位价格接近于主存的每位平均价格。
- ▶ 解决了高速度和低成本之间的矛盾。
- ▶ 这个层次完全由硬件实现，对用户是透明的。

2.主—辅存层次

主—辅存层次通过附加的硬件及存储管理软件来控制。从整体看，主—辅存层次的存取速度接近于主存的存取速度，容量是辅存的容量，而每位平均价格接近于廉价的辅存平均价格，从而解决了大容量和低成本间的矛盾。



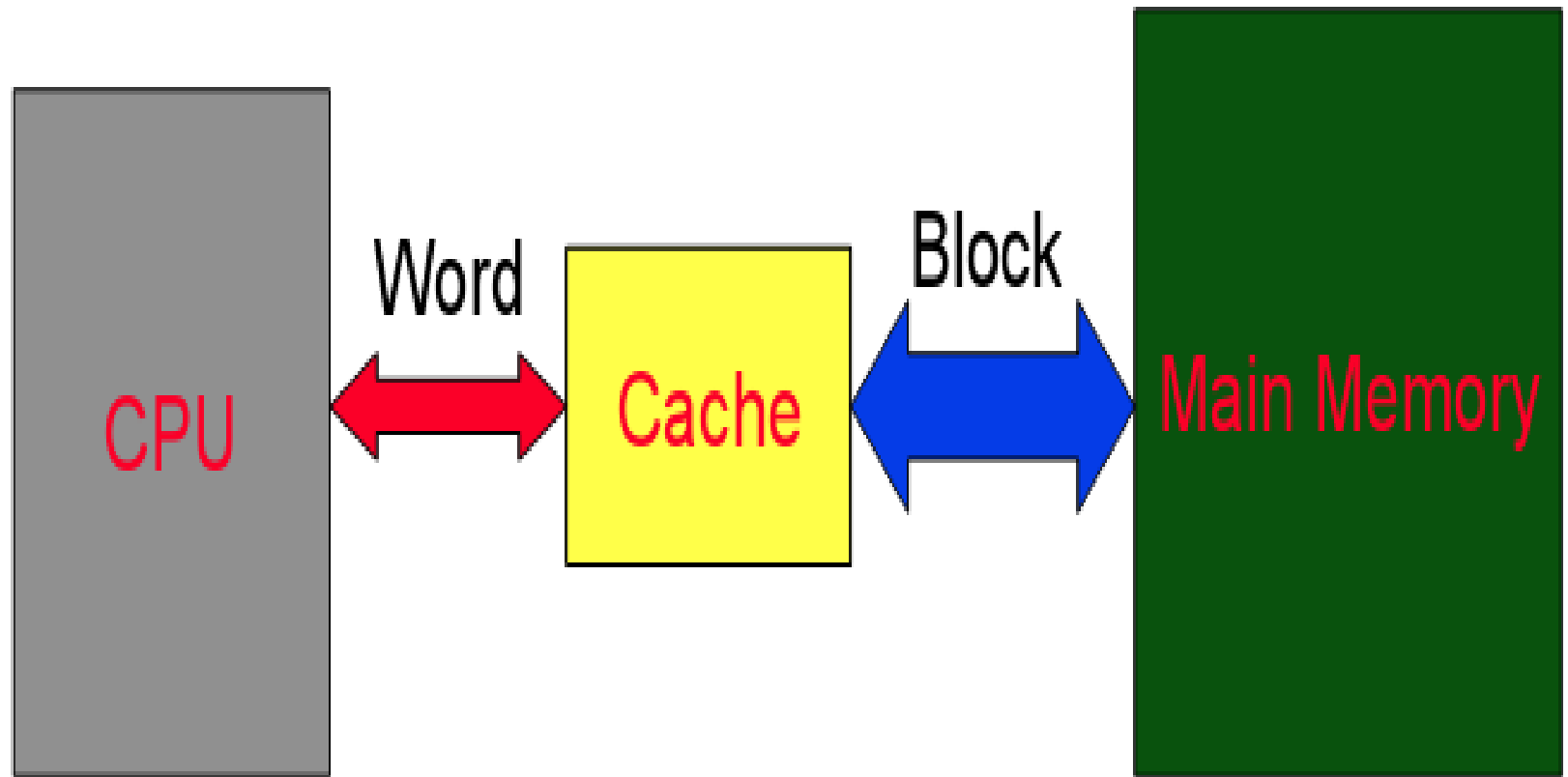
高速缓冲存储器的工作原理

1.程序访问的局部性:

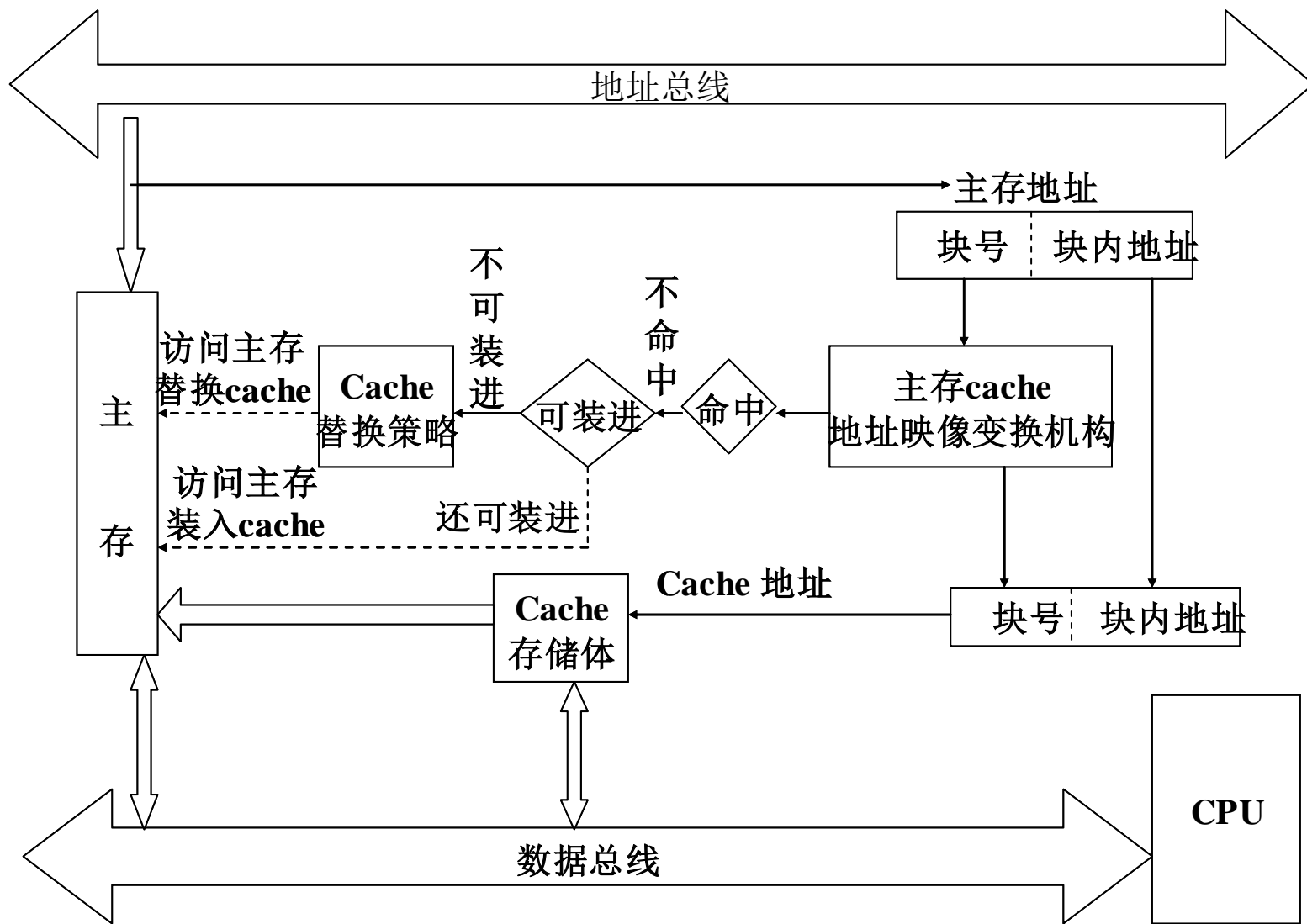
- ◆时间上的局部性指的是当前正在使用的信息很可能是后面立刻还要使用的信息，原因是程序循环（重复执行一段代码）。
- ◆空间上的局部性是指在一个较短的时间间隔内，如果一个单元被访问，则该单元邻近的单元也可能很快被访问。原因是程序的连续存放以及顺序执行，程序中的数组操作。

2.高速缓存的工作原理:

- ▶ **Cache采用SRAM器件，构成小容量高速存储器。**
- ▶ **把程序中经常使用的部分存放在Cache中。**
- ▶ **使CPU的访存操作大多数在Cache中命中，从而使程序的执行速度大大提高。**



- ▶ **Cache**和主存都被分成若干个大小相等的块，每块由若干字节组成。
- ▶ 当**CPU**发出访存请求时，以主存地址同时访问**Cache**和主存。
- ▶ 如果**Cache命中**，就直接对**Cache**进行读写操作；
- ▶ 如果**Cache不命中**，则从主存将该信息所在的块信息一次从主存调入**Cache**内。
- ▶ 根据程序访问的局部性原理，该块信息在一段时间内被访问的概率很高，保证保存程序活跃部分。
- ▶ 若此时**Cache**已满，则需根据某种替换算法，用这个块替换掉**Cache**中原来的某块信息。



- ▶ 命中率 H 是指对该级存储器来说，要访问的信息正好在这一级中的概率，即命中的访问次数与总访问次数之比。
- ▶ 如何保证较高的命中率？地址映射规则和替换算法

高速缓冲存储器的地址映像方法:

地址映像方法决定主存信息放到Cache中的位置。

应用某种函数把主存地址映像为Cache地址，称为地址映像。

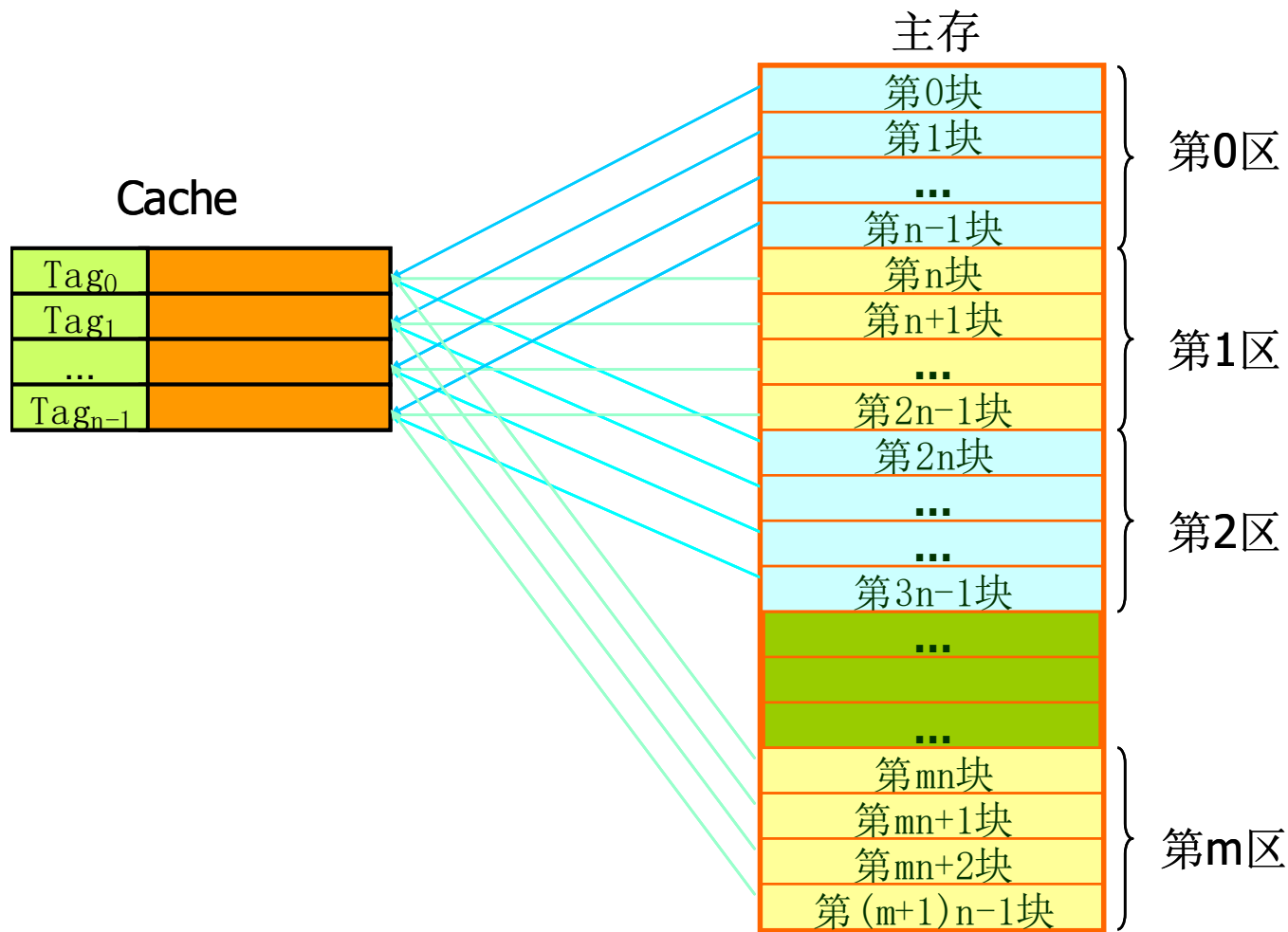
根据主存地址，按照映像算法将主存地址变换成cache地址叫做地址变换

(1)直接映像

(2)全相联映像

(3)组相联映像

1.直接映射



映象规则:

主存分割成若干个与cache大小相同的区，Cache块号b与主存块号B的对应关系如下：

$$b = B \bmod C_b$$

M_b 应是 C_b 的整数倍。

C_b 是Cache中的块数， M_b 主存块数是

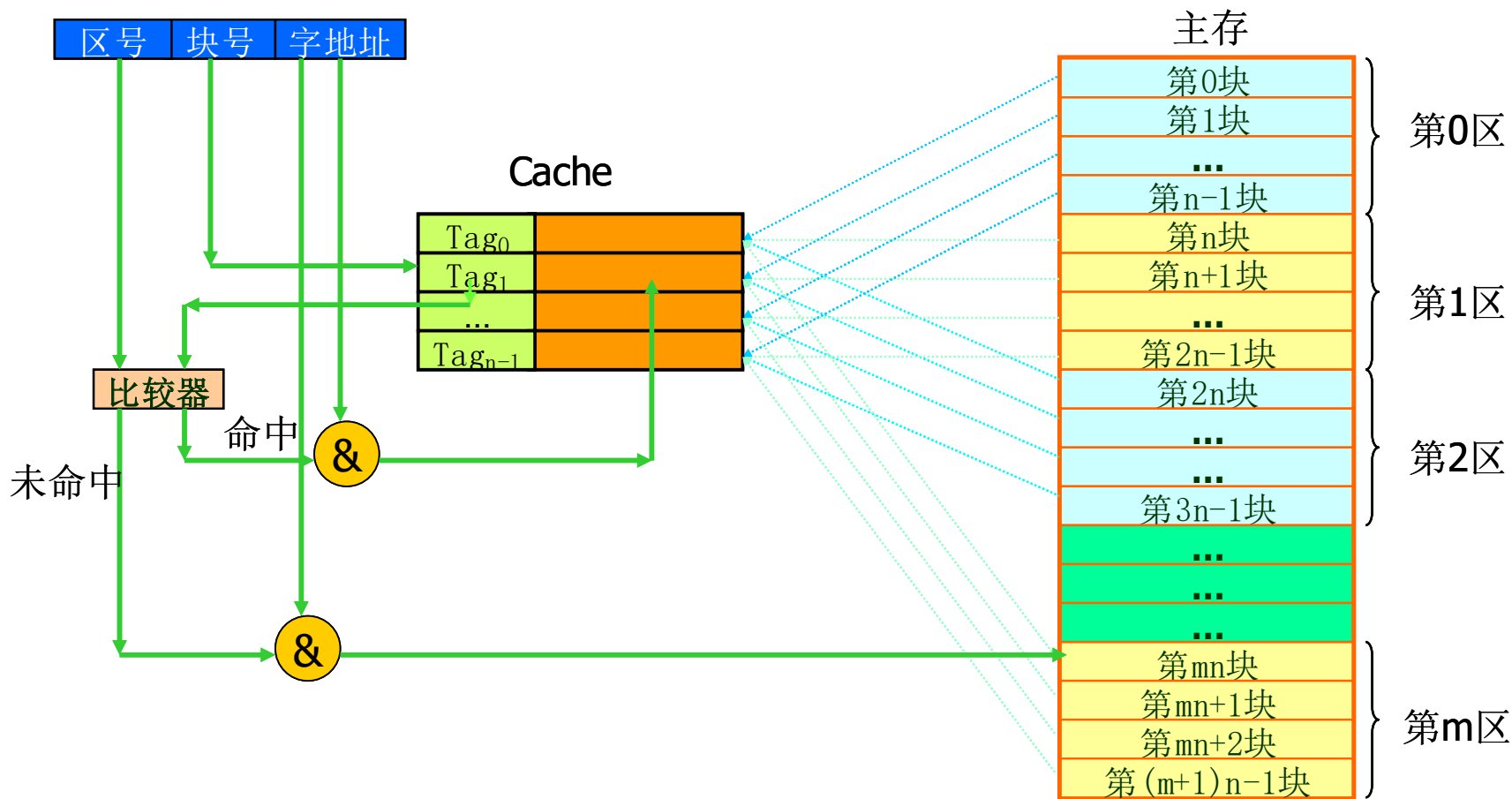
▶ 主存地址:

区号E	块号B	块内地址W
-----	-----	-------

▶ Cache地址:

块号b	块内地址w
-----	-------

- ▶ Tag内容是映射到该位置的主存块的主存地址中的区号，因为比较时只需确定映射的是哪一个区中的块。



Cache直接映射地址变换过程

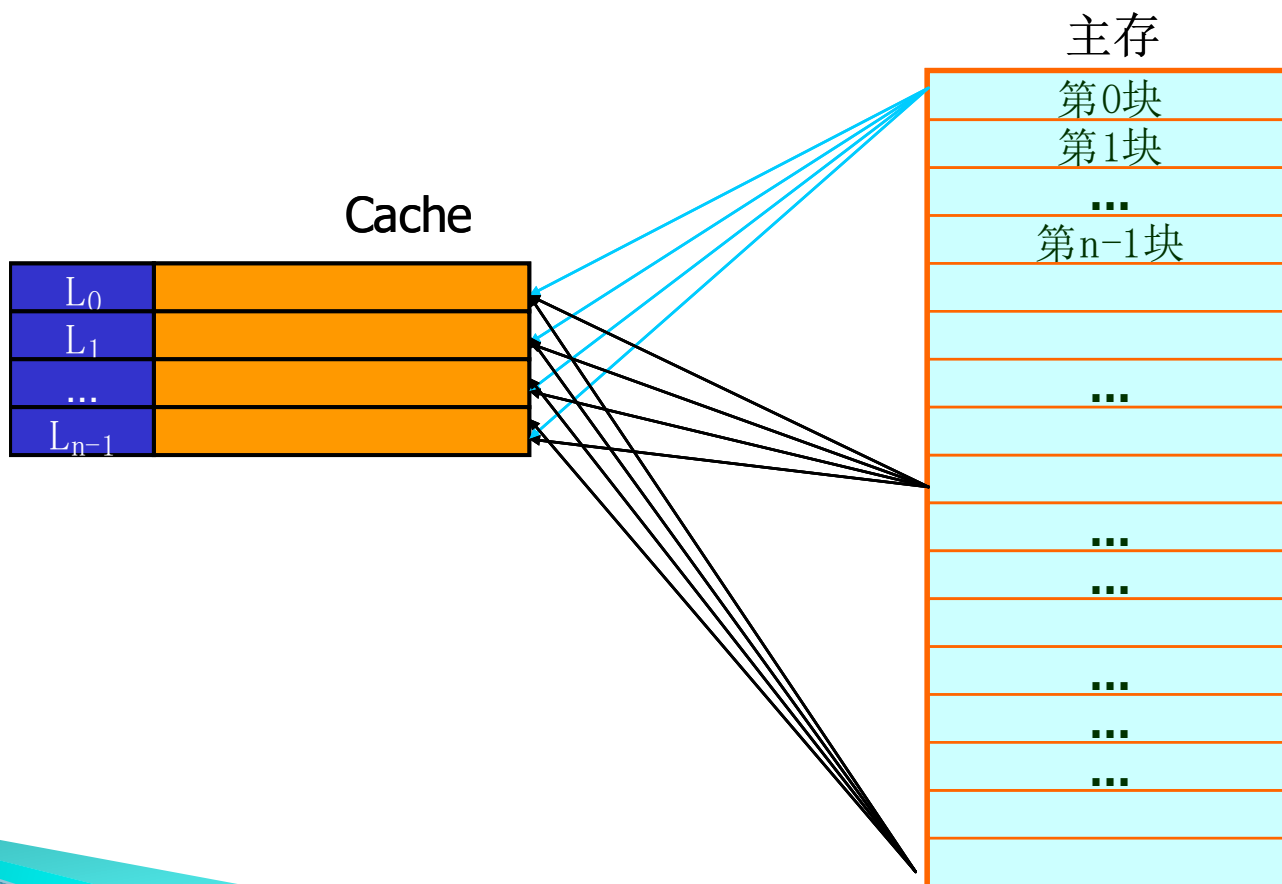
地址转换过程：

用程序中给出的主存地址中的块号找到Cache中对应的块，读出块的Tag标记与主存地址给出的区号进行比较，按照以下几种情况进行判断：

- ▶ 如果与主存地址给出的区号相等，且有效位为1，命中。
- ▶ 如果区号相等，有效位为0，失效（作废）。
- ▶ 区号不相等，有效位为0，cache块是空的，可以直接装入
- ▶ 区号不相等，有效位为1，该快的内容有用，写回后，替换

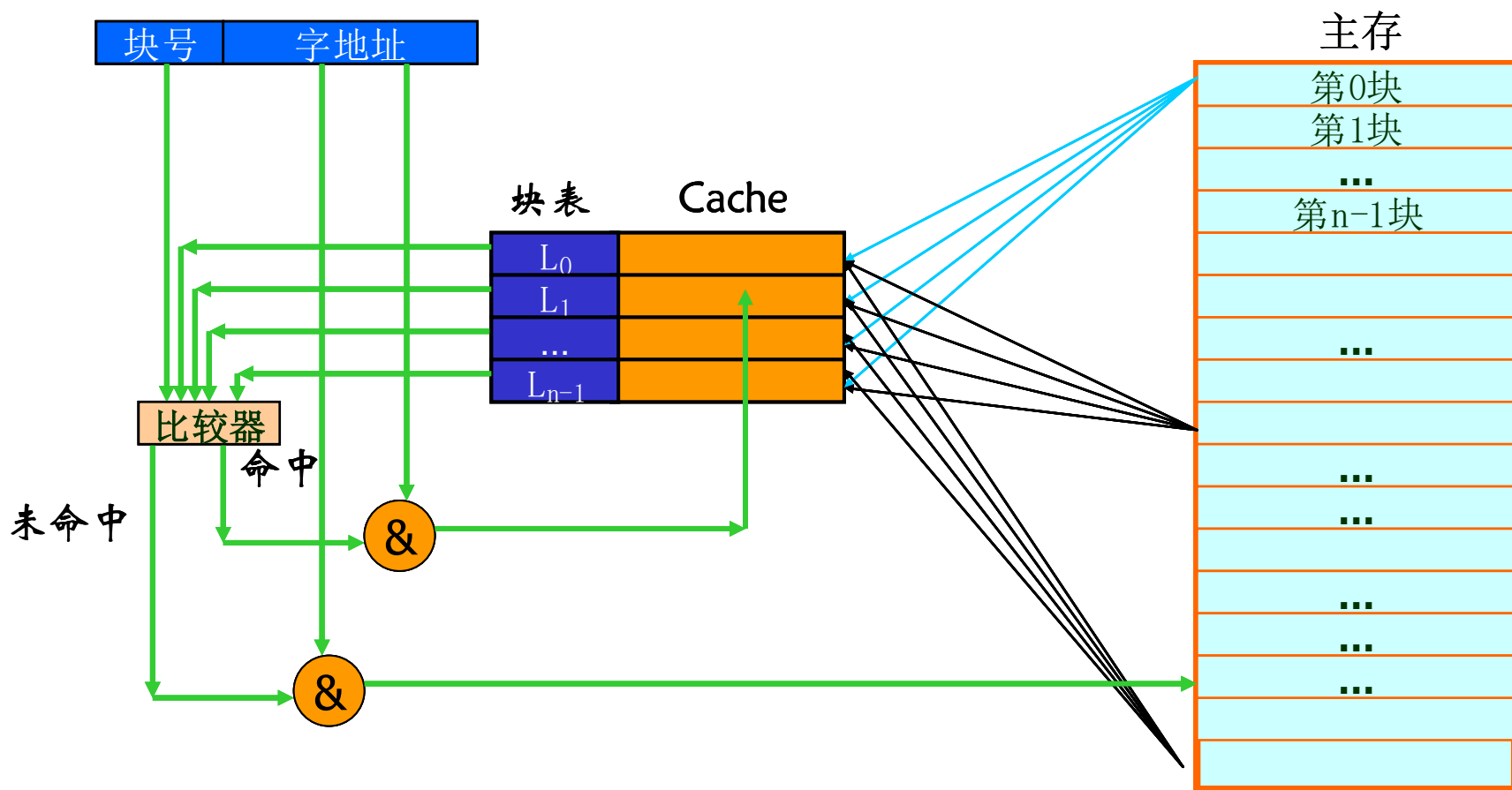
2. 全相联映射

主存分为若干Block，Cache按同样大小分成若干Block，Cache中的Block数目显然比主存的Block数少得多。主存中任何一块均可定位于Cache中的任意一块，可提高命中率，但是硬件开销增加



Cache的Tag内容:

- ▶ 主存中与该Cache数据块对应的数据块的块地址。
- ▶ 标记位数等于主存块号位数。



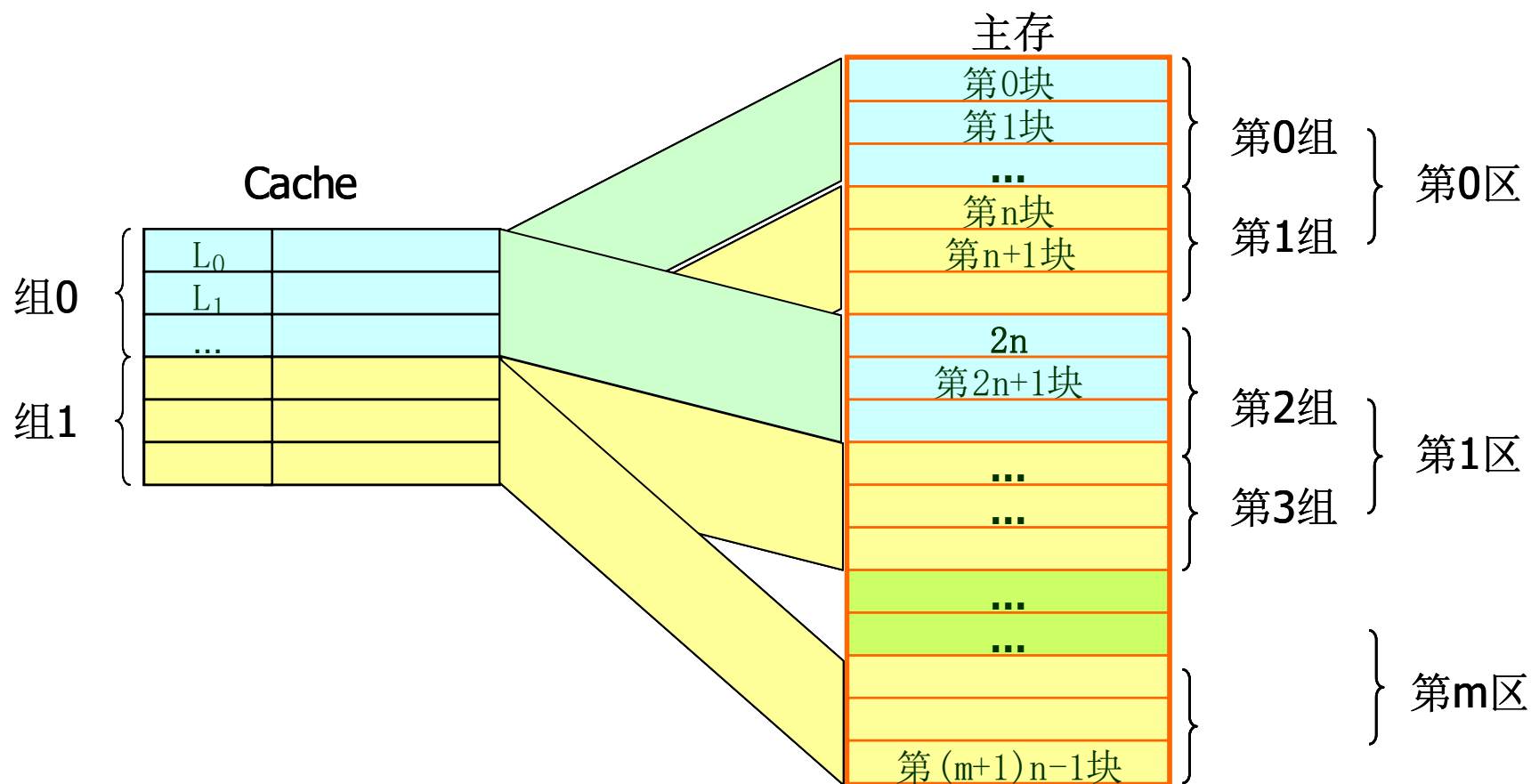
Cache全相联映射地址变换过程

地址转换过程：

因为全相联映射的主存块可能映射在任何Cache块内，所以根据主存块号，与所有cache块的标记进行比较，有相等的，说明命中，读出cache块号访问cache。

需要一个目录表来存放映射关系，目录表容量为cache的块数，字长等于CACHE块的标记、cache块号、有效位之和。

3. 组相联映射



映像规则:

主存分割成若干个与cache大小相同的区,Cache和主存各区内再分割成若干组, 每组若干块。组到组是直接映像, 组内是全相联映像。

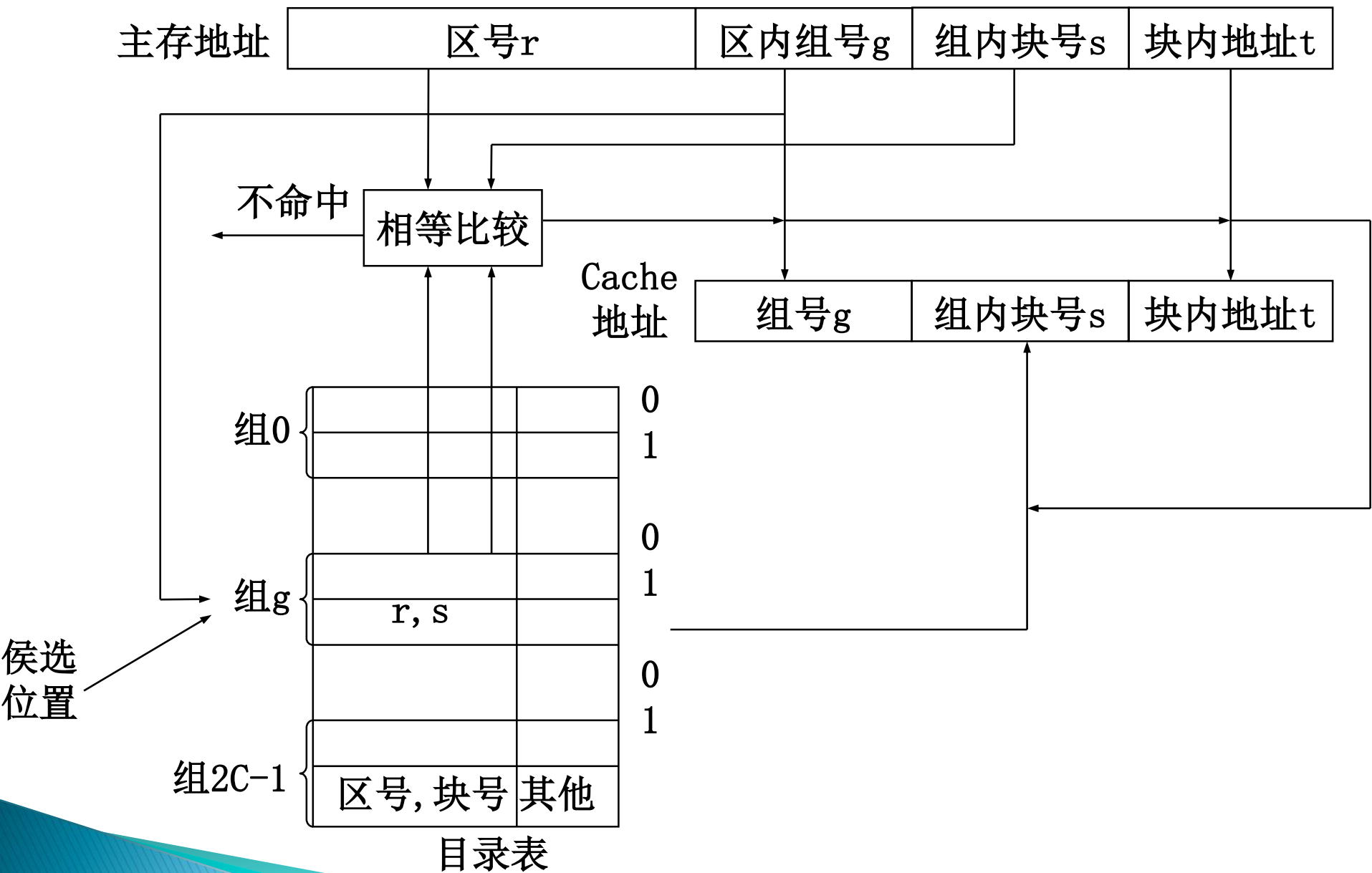
主存地址:

区号E	组号G	块号B	块内地址W
-----	-----	-----	-------

Cache地址:

组号g	块号b	块内地址w
-----	-----	-------

Tag内容是主存区号E和组内块号B，因为某一个Cache块中的内容可能来自任何一个区，确定了区号后，区内组号是一定的，（因为组到组是直接映像）还需要确定是组内哪一块（因为组内是全相联映像）



地址变换过程

- ▶ 需要一个块表，容量为cache的块数，字长等于主存地址的区号、组号、主存块号、cache块号等。
- ▶ 根据主存地址中的组号，按地址访问块表，读出组号相同的多个字。字的个数等于组内块数。把这些字中的区号E与块号B与主存地址的区号与块号进行比较，有相等的，说明命中。读出存储字中的cache块号b访问cache。
- ▶ 如果几路的标记同时比较，所以比较器的个数等于路数。
- ▶ 比较器的位数等标记的位数。

例7. 某计算机的Cache-主存层次采用组相联映射方式，块大小为128B，Cache容量为64块，按4块分组，主存容量为4096块，问求一个主存地址有多少位？一个cache地址有多少位？计算机主存地址格式中，区号，组号，块号和块内地址字段的位数。

(3) 说明层次结构的存储系统中CACHE和虚拟存储器的作用有何不同。

解：

(1) 主存容量 = $4096 \times 128\text{B} = 2^{19}$

故主存地址共有19位

块的大小为128B, 所以块内地址 = 7位

Cache容量为64块, 按4块分组, 组数为16, 所以
组地址 = 4位

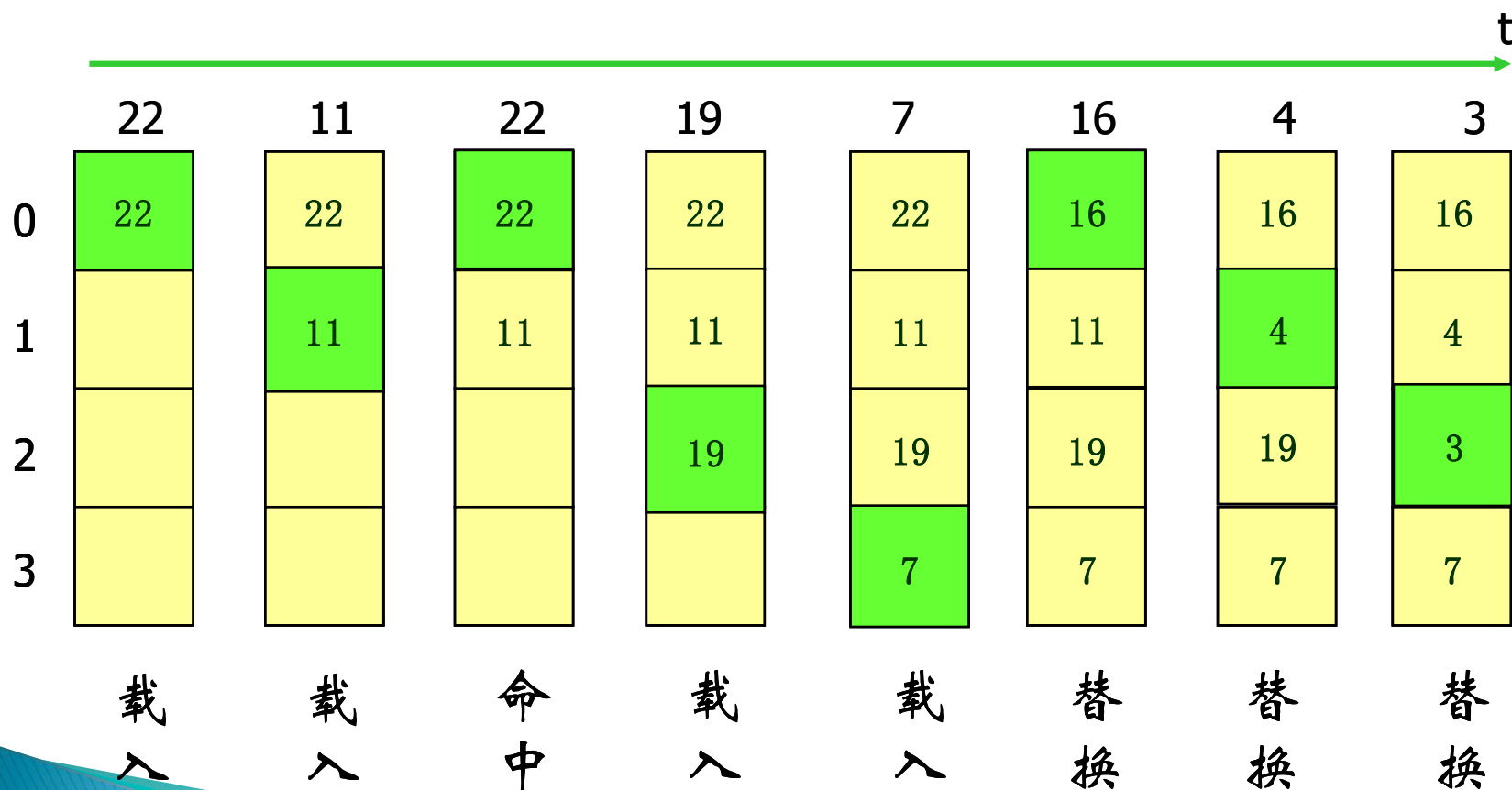
区地址 = $19 - 7 - 4 - 2 = 6$

(2) 由以上分析可知: 主存地址一共包括6位区号,
4位组号, 2位块号, 7位块内地址, 共19位,
cache地址一共包括4位组号, 2位块号, 7位块内
地址, 共13位。

(3) 引入Cache结构的目的是为了解决主存和CPU
之间速度匹配问题, 而采用虚拟存储结构目的是解
决主存容量不足的问题

高速缓冲存储器的替换算法

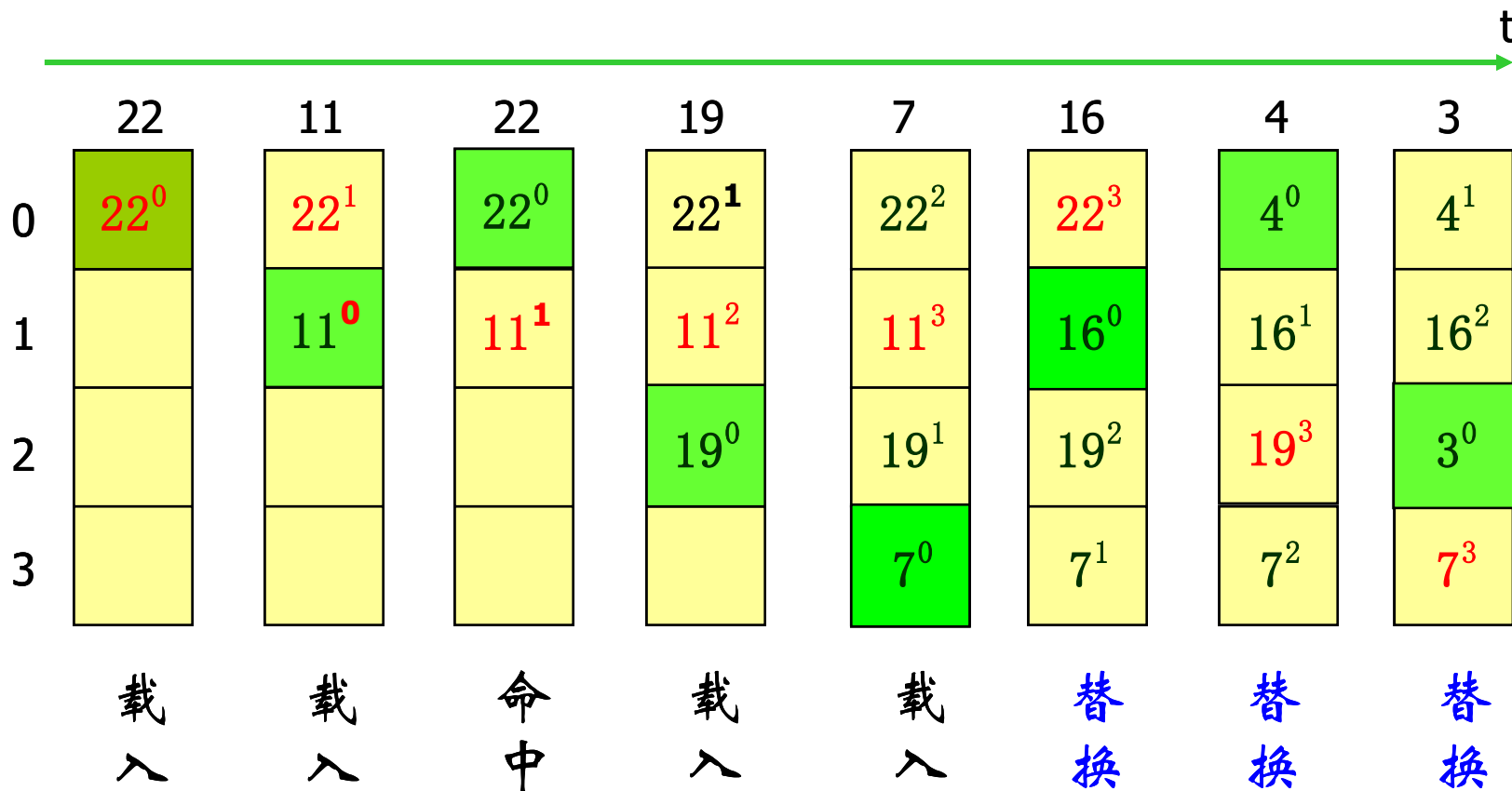
1. 先进先出 (FIFO) 算法



2.近期最少使用（LRU）算法：

每个cache单元需要一个计数器跟踪最近访问的数据，每次访问的数据相应的计数器置为0；其余的计数器依次加1。替换最少使用的数据。

LRU算法：



Cache一致性

1.写Cache命中时，如何保持Cache与主存中的内容一致？

(1) 写直达法：每次信息从CPU写入cache单元中时，也要写回相应的物理内存单元中。

(2) 写回法：即数据暂写入Cache，并用标志将该块注明，等需要将该块替换回到主存时，才写回主存，也称标志交换方式。

层次间应满足的原则

1. **一致性原则**：处在不同层次的同一个信息应保持相同的值。
2. **包含性原则**：处在内层的信息一定被包含在其外层的存储器中，反之则不成立，即内层存储器的全部信息，是其相邻外层信息的一部分的复制品。

2.写Cache不命中时:

(1) 按写分配法

把信息写入主存，同时将该块信息装入cache。

(2) 不按写分配法

直接更新物理内存中的值，而不把值装入cache

cache 性能评价

1、命中(Cache hit)率:

在一个程序执行期间，设 N_c 表示Cache完成存取的总次数， N_m 表示主存完成存取的总次数， h 定义为命中率。则有：

$$h = \frac{N_c}{N_c + N_m}$$

命中率跟程序、cache容量、组织方式、块的大小有关。

2、存储系统平均存取时间

cache存取时间为 t_c ，命中率为 h ，主存的存取时间为 t_m 则系统平均存取时间为：

$$t_a = h * t_c + (1 - h) * t_m$$

3、存储系统的访问效率

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

例：某计算机系统的内存存储系统是由cache和主存构成，cache的存取周期是45ns，主存的存取周期是200ns，已知在一段给定的时间内，CPU共访问存储系统2000次，其中访问主存100次，问：

- (1) cache的命中率是多少？**
- (2) CPU访问该内存存储系统的平均时间？**
- (3) cache--主存的效率是多少？**

$$(1) \quad h = N_c / (N_c + N_m) = (2000 - 100) / 2000 = 0.95$$

$$(2) \quad t_a = h * t_c + (1 - h) * t_m \\ = 0.95 * 45 + 0.05 * 200 = 60\text{ns}$$

$$(3) \quad e = t_c / t_a = 45 / 60 = 83.3\%$$

例： cache的存取周期是40ns，主存的存取周期是200ns， cache /主存系统平均访问时间为50ns，求 cache的命中率？

$$50 = h*40 + (1-h) * 200$$

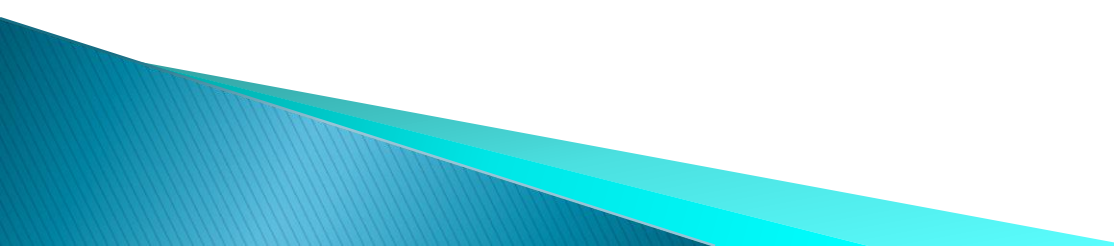
$$h = 93.8\%$$

5.7.3 虚拟存储器

虚拟存储器的基本概念

虚拟存储器是一种将大的逻辑空间映射到比它小得多的物理空间的机制。由CPU中的存储管理模块和操作系统中的相应模块共同支持

例题及习题

- 1、某计算机字长32位，其存储容量为64MB，若按字编址，它的存储系统的地址线至少需要（ ）条。
 - 2、对存储器的要求是容量大、速度快、成本低，为了解决这三方面的矛盾，计算机采用多级存储体系结构，即（ ）、（ ）、（ ）。
 - 3、主存储器的技术指标有（ ），（ ），（ ），（ ）。
 - 4、cache和主存构成了（ ），全由（ ）来实现。
 - 5、虚拟存储器分为页式、（ ）式、（ ）式三种。
- 

1、EEPROM是指（ ）。

- A 读写存储器 B 只读存储器
C 闪速存储器 D 电擦除可编程只读存储器

2、某SRAM芯片，其容量为 $1\text{M} \times 8$ 位，除电源和接地端外，控制端有E和R/W#，该芯片的管脚引出线数目是（ ）。

- A 20 B 28 C 30 D 32
- 

3、虚拟存储技术主要解决存储器的（ ）问题。

A 速度 B 扩大存储容量 C 成本 D 前三者兼顾

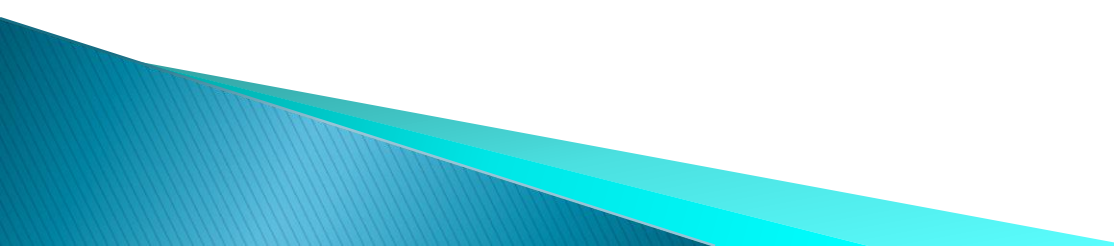
4、主存贮器和CPU之间增加cache的目的是（ ）。

A 解决CPU和主存之间的速度匹配问题

B 扩大主存贮器容量

C 扩大CPU中通用寄存器的数量

D 既扩大主存贮器容量，又扩大CPU中通用寄存器的数量



1、某计算机系统的存储器由cache和主存构成。已知在一段给定的时间内，CPU共访问内存5000次，其中400次访问主存。问cache的命中率是多少？