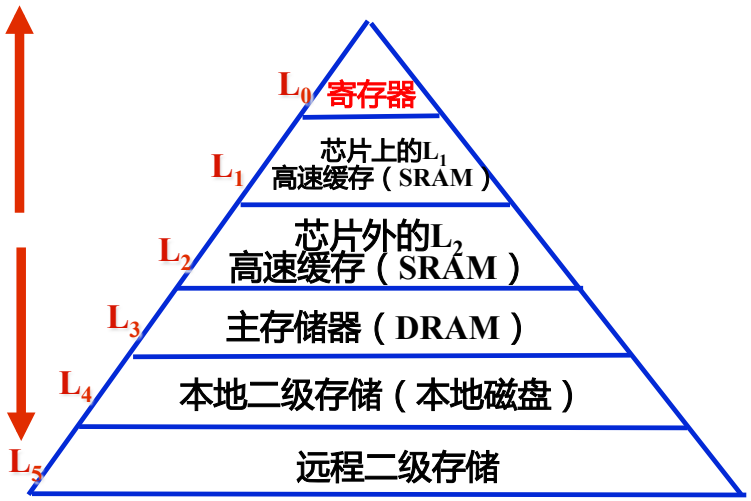
5.3.2 Cache(高速缓存)是什么样的？

刘 芳 副教授
国防科学技术大学计算机学院

加速访存措施一：引入Cache

在CPU和主存之间设置一个快速、小容量的存储器，总是存放最活跃(即，被频繁访问)的程序块和数据

早期的计算机，Cache用来代表存储层次结构中处理器和主存之间的特殊层次



存储器是一个塔式层次结构

- 数据只有在第i+1层存在，才能在第i层被访问到
- 处理器的访存时间主要由层次1决定，而整个存储器的容量却和层次n一样大

思考1：实现Cache机制需解决哪些问题？

- 如何分块？
- 主存块和Cache之间如何映射？ 映射策略
- Cache已满时，怎么办？ 淘汰/替换策略
- 写数据时，怎样保证Cache和MM一致性？ 一致性问题：回写/多处理器问题
- 给出的主存地址怎么样转换为Cache地址？

思考2：Cache对程序员(编译器)是否透明？

- 透明的。程序员(编译器)在编写/生成低级语言程序时，无需了解Cache是否存在或如何设置
- 了解Cache有助编写出高效程序！

Cache结构

- Cache是小容量、高速缓冲存储器，由SRAM组成
- Cache直接制作在CPU芯片内，速度几乎与CPU一样快
- 一般将Cache和主存的存储空间都划分为若干大小相同的块（主存中称为：块Block、Cache中称为：行line）

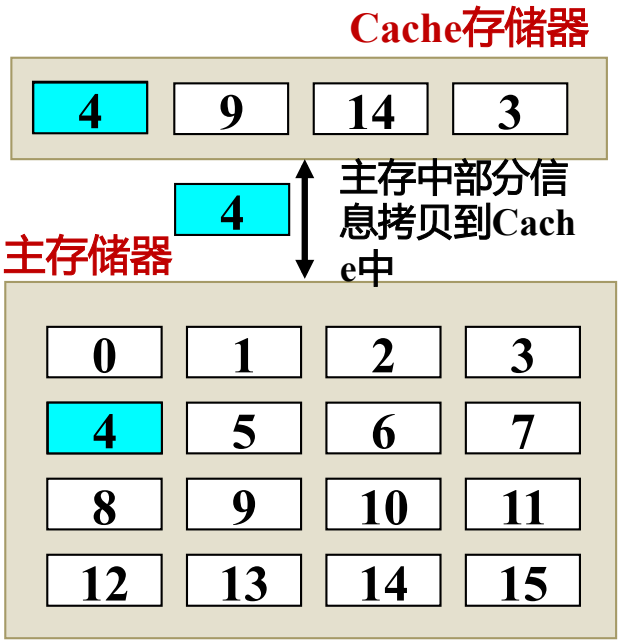
Cache原理

程序运行时，CPU使用的一部分数据/指令会预先成批拷贝到Cache中，Cache的内容是主存储器中部分内容的映象(副本)

当CPU需从主存读(写)数据或指令时，先查看Cache

- 若有，则直接从Cache中取，不用访问主存
- 若没有，则直接访问主存

数据访问过程：



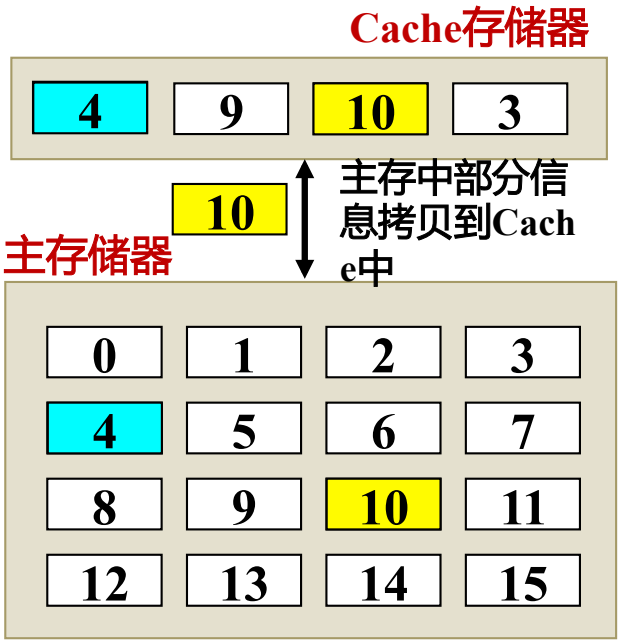
Cache原理

程序运行时，CPU使用的一部分数据/指令会预先成批拷贝到Cache中，Cache的内容是主存储器中部分内容的映象(副本)

当CPU需从主存读(写)数据或指令时，先查看Cache

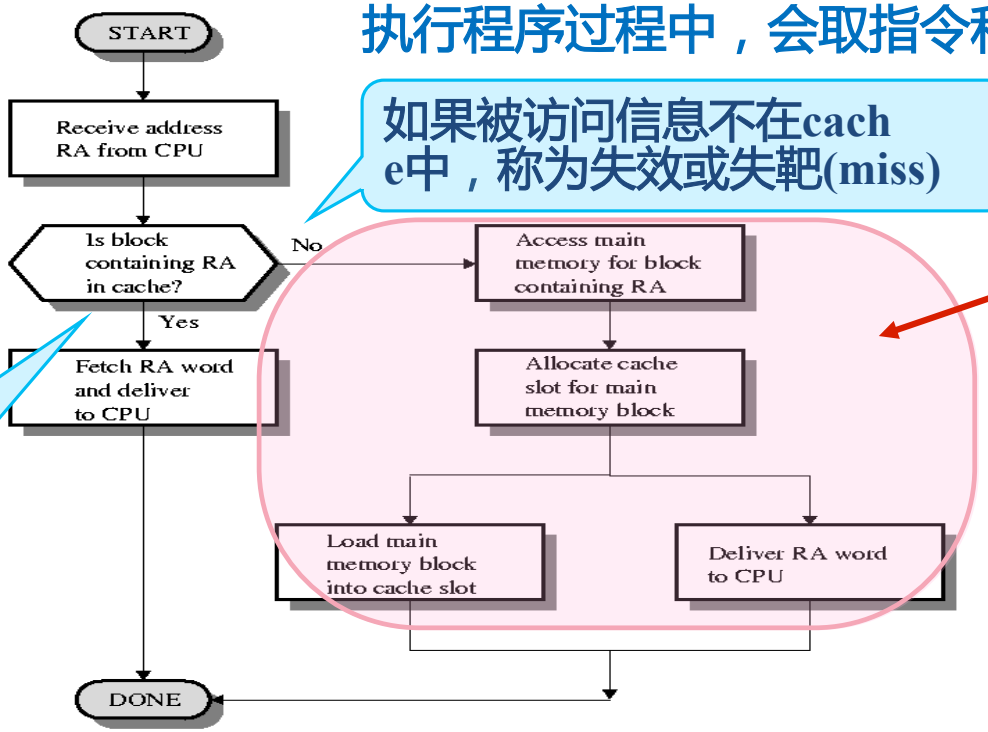
- 若有，则直接从Cache中取，不用访问主存
- 若没有，则直接访问主存

数据访问过程：



Cache操作过程

执行程序过程中，会取指令和读写数据！



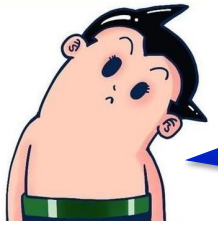
如果被访问信息不在cache中，称为失效或失靶(miss)

失效处理

在cache中为获得的数据分配一个空间

如果被访问信息在cache中，称为命中(hit)

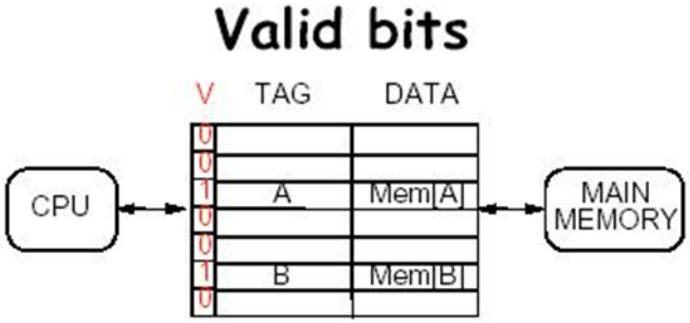
5.3.2 Cache(高速缓存)是什么样的？



系统加电启动后，Cache内无有效信息，如何标识？
信息从主存复制到Cache时，Cache中的有效信息，如何标识？

解决方法

- 每个TAG域增加1位 — 有效位V(Valid Bit)
- 开机或复位时， All V=0
- 命中的Cache行，V=1
- Flush Cache行，V=0
- 新装入Cache行，V=1



Cache - 主存层次的平均访问时间

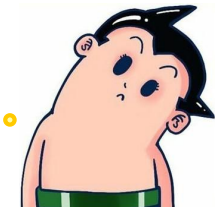
命中（Hit）：要访问的信息在Cache中

- Hit Rate(命中率 *p*)： 在Cache中的概率
- Hit Time (命中时间 Tc) ：访问Cache所需时间，包括：判断时间 + Cache访问

失效（Miss）：要访问的信息不在Cache中

- Miss Rate (失靶率/失效率 1 - p) = 1 - (Hit Rate)
- Miss Penalty (失效损失 Tm)：从主存将一块信息替换到Cache所需时间，包括访问主存块，向上逐层传输块直至将数据块放入发生缺失的那一层所需时间。

命中时间<<失效损失



平均访问时间

提高平均访问速度，必须提高命中率！

命中率对平均访问时间的影响

平均访问时间 (average memory access time, AMAT)

$$T = p \times T_C + (1 - p)(T_C + T_M) = T_C + (1 - p)T_M$$

例1. 若命中率p=0.85， T_C=1 ns，T_M =20ns，则平均访问时间T为多少？

答： T = 4ns

例2. 若命中率p提高到0.95，结果如何？

答： T = 2ns

例3. 若命中率为0.99？

答: T = 1.2ns

访问速度与命中率的关系非常大！

How high of a hit ratio?

Suppose we can easily build an on-chip static memory with a 4 nS access time, but the fastest dynamic memories that we can buy for main memory have an average access time of 40 nS. How high of a hit rate do we need to sustain an average access time of 5 nS?

$$\alpha = 1 - \frac{t_{ave} - t_c}{t_m} = 1 - \frac{5 - 4}{40} = 97.5\%$$