ARQUITETURA DE COMPUTADORES E

SISTEMAS OPERACIONAIS | UNIDADE 3

Aula 2 | Memória Cache e Memória Principal

PROFESSOR(A): BEATRIZ C SANTANA

1. Tempo médio de acesso

2. Tecnologia de fabricação de memória

3. Memória cache - Localidade

4. Mapeamento da memória cache

Tempo Médio de Acesso

- ➔ A memória funciona numa velocidade menor que a do processador.
- ➔ Uma forma de acelerar o acesso às informações é organizar as memórias de forma hierárquica, colocando as mais rápidas mais próximas ao processador.
- ➔ A tentativa de acesso é feita sempre partindo do nível mais alto da hierarquia de memórias. Como essas memórias são mais rápidas, evidentemente o tempo de acesso será menor. Caso haja a possibilidade de a informação requerida não estar no nível de memória acessado, nesse caso, é feita uma busca no nível abaixo da hierarquia de memória.

Tempo Médio de Acesso

- Quando uma informação procurada em um nível de memória é encontrada, chamamos de acerto (hit), quando a informação não é encontrada, chamamos de falha (miss).

$$t_a = (t_h \times p_h) + (t_m \times p_m)$$

- ta é o tempo médio de acesso a uma informação em um nível de memória, th é o tempo de acesso no caso de um acerto, ph é a probabilidade de acerto de acesso, tm é o tempo de acesso no caso de uma falha, ou seja, o tempo de acesso aos níveis inferiores na hierarquia de memória; e pm é a probabilidade de falha de acesso.

Tempo Médio de Acesso

- ➔ Exemplo
- ➔ Se um nível de memória tem 75% de chance de conter uma informação, e o tempo de acesso a essa memória seja de 10ns, em caso de acerto, e 100ns, em caso de falha, teremos

$$ta = (10 \times 0{,}75) + (100 \times 0{,}25) = 7{,}5ns + 25ns = 32{,}5ns.$$

1. Tempo médio de acesso

2. Tecnologia de fabricação de memória

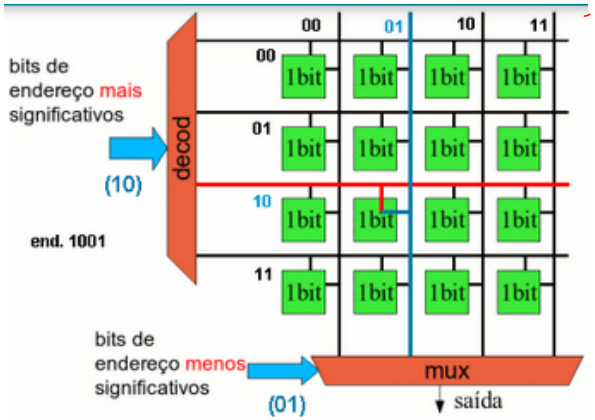
3. Memória cache - Localidade

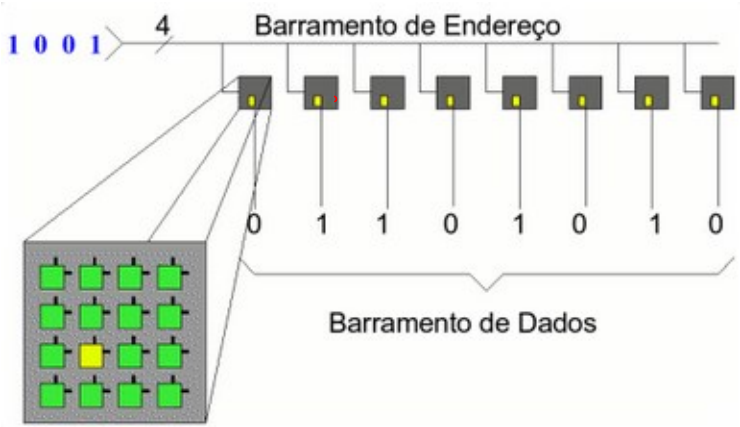
4. Mapeamento da memória cache

Tecnologias de Fabricação de Memória

- ➔ Em termos de tecnologia de fabricação de memória RAM, temos dois tipos principais:
- ➔ A memória RAM estática (SRAM) e a memória RAM dinâmica (DRAM).
- ➔ A primeira é mais antiga, porém, até hoje, a mais rápida e, consequentemente, mais cara, é usada principalmente para a construção de memórias cache, já a segunda, mais barata, é utilizada para a fabricação de memórias principais (MP).

Organização da Memória DRAM





1. Tempo médio de acesso

2. Tecnologia de fabricação de memória

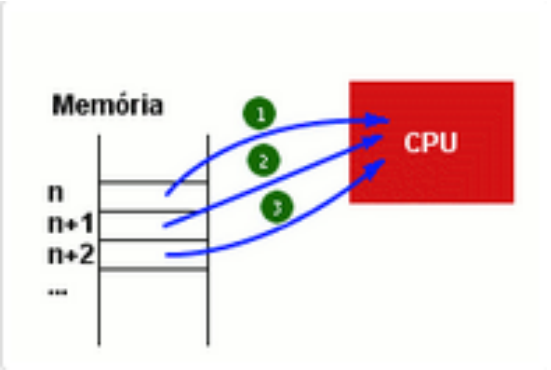
3. Memória cache - Localidade

4. Mapeamento da memória cache

Memória Cache - Localidade

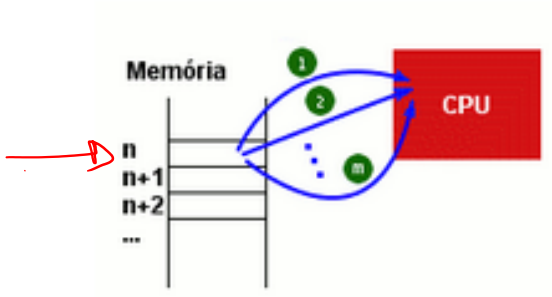
- **A localidade espacial** quer dizer que, "se um programa acessa uma informação na memória, é provável que venha a acessar outra informação próxima a essa em um curto prazo".

- Isso quer dizer que, se o programa acessa um endereço n, é bem provável que venha a acessar um endereço n+1, n+2, ..., logo em seguida.



Memória Cache - Localidade

- **A localidade temporal** quer dizer que, "se um programa acessa uma informação na memória, é provável que venha a acessar essa mesma informação em um curto prazo".
- Veja o exemplo: suponha que um programa deve ler um valor da memória e somar esse valor dez vezes, serão feitos dez acessos ao mesmo valor na memória.



Uso das Localidades Espacial e Temporal

- ➔ O conceito de localidade é aplicado da seguinte forma:
- ➔ Como a memória cache é uma memória bem menor que a memória principal, não poderá conter todas as informações que são solicitadas, mas conterá as últimas informações acessadas.
- ➔ Para tal, o processador procurará o dado primeiro na memória cache, se o dado estiver lá (hit cache), o processador não precisará acessar a memória principal, caso contrário (miss cache), o processador faz um acesso à memória principal em busca do dado solicitado.

Uso das Localidades Espacial e Temporal

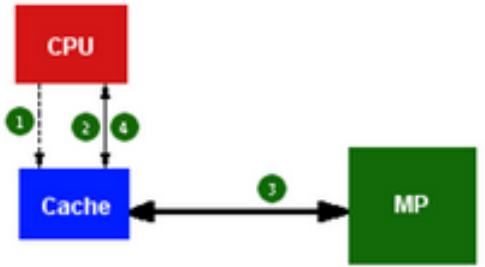
➔ Possibilidades de acesso a cache:

1- O processador verifica se a palavra solicitada está na cache.

2- Se ocorrer hit cachê, a palavra é acessada.

3- Senão, o bloco da MP que contém a palavra é carregado na cache.

4- A palavra é acessada.



1. Tempo médio de acesso

2. Tecnologia de fabricação de memória

3. Memória cache - Localidade

4. Mapeamento da memória cache

Mapeamento de Memória Cache

- Para o correto funcionamento do esquema cache/memória principal, é importante considerarmos algoritmos ou políticas de substituição de blocos e o mapeamento de cache.

- FIFO (First In First Out) – este algoritmo funciona basicamente como uma fila, ou seja, o bloco que está há mais tempo na cache (o primeiro que chegou) é escolhido para ser substituído.

- LFU (Least Frequently Used – menos frequentemente usado) – de acordo com esse algoritmo, o bloco a ser substituído é aquele que tiver a menor quantidade de acessos.

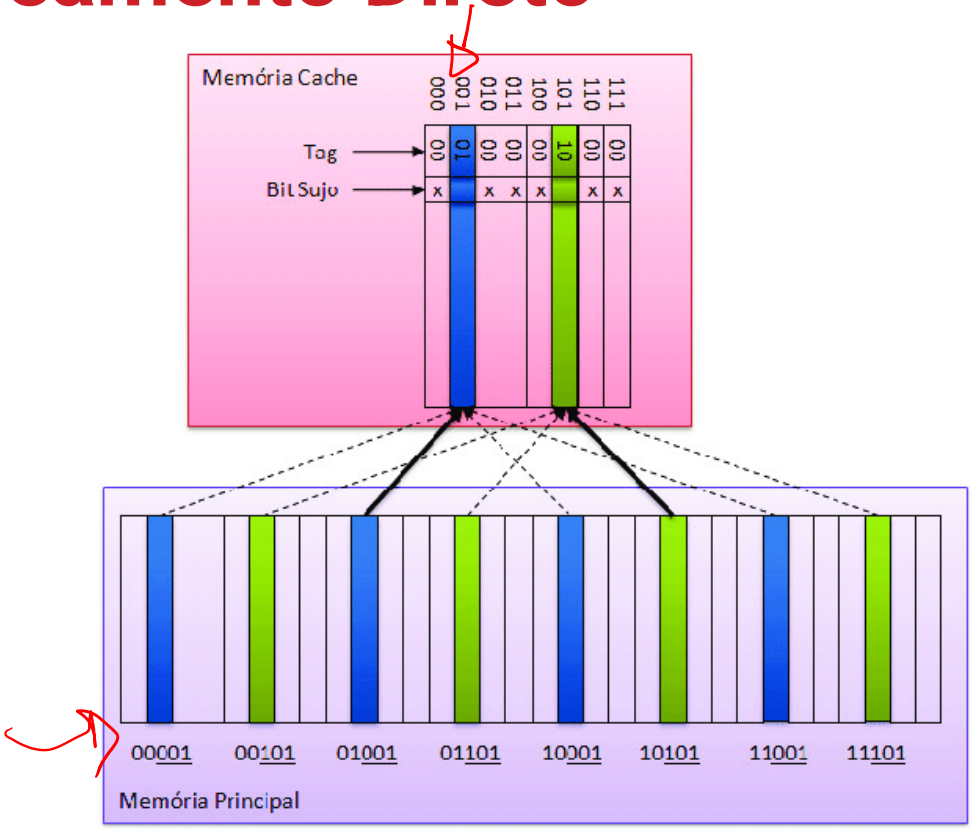
- LRU (Least Recently Used – menos recentemente usado) – nesse caso, o bloco a ser utilizado é aquele que tiver sido usado pela última vez há mais tempo, ou seja, aquele que está há mais tempo ocioso.

Mapeamento Direto

- Neste mapeamento, cada bloco da memória principal é mapeado para um quadro da cache. O quadro a ser usado é obtido pelo resto da divisão do endereço do bloco da memória principal pela quantidade de quadros da cache. Cada quadro da cache tem três campos: o índice, o tag e o endereço de memória. O tag é usado para validar se a linha procurada é a mesma que está na cache.

- Na Figura podemos observar que, para cada quadro da cache, teremos a possibilidade de quatro blocos da memória principal. Assim, precisamos de dois bits no tag para identificar qual dos blocos da memória principal está carregado ali.

Mapeamento Direto

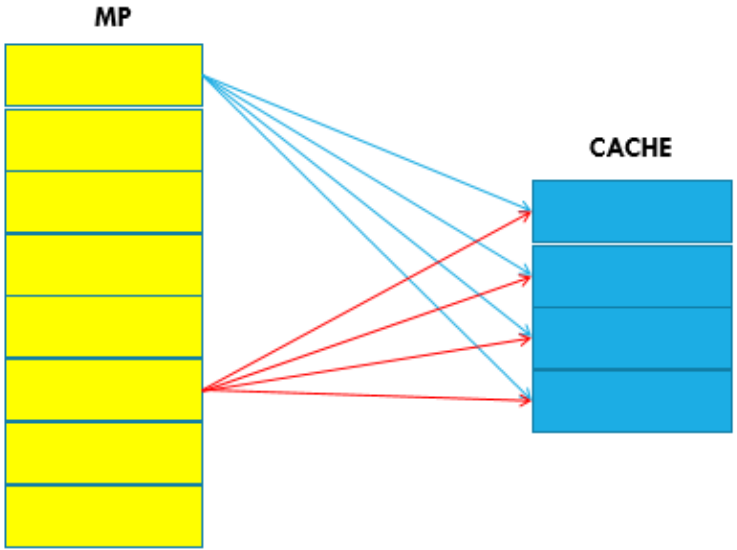


Fonte: https://www.researchgate.net/figure/Figura-39-Diagrama-de-uma-memoria-cache-com-mapeamento-direto-apresentando-a-tag-e-o_fig5_256293104

Mapeamento Totalmente Associativo

➔ Neste mapeamento, um bloco de memória pode estar em qualquer quadro da cache. Neste caso, sempre que precisar encontrar algo, o processador deverá varrer toda a cache até encontrar (se encontrar) o que procura.

Mapeamento Totalmente Associativo



Mapeamento Associativo por Conjuntos

- Este mapeamento é um híbrido entre os mapeamentos direto e totalmente associativo. Nele, os quadros da cache são divididos em conjuntos. O bloco de memória usará então o quadro de acordo com o resto da divisão entre o endereço da memória principal e a quantidade de conjuntos da cache (como no mapeamento direto). Dentro do conjunto, o bloco da memória principal poderá estar em qualquer um dos quadros (como no mapeamento totalmente associativo).

Mapeamento Associativo por Conjuntos