UniEVANGÉLICA
UNIVERSIDADE EVANGÉLICA DE GOIÁS

CURSO DE ENGENHARIA DE SOFTWARE

Disciplina: Arquitetura e Organização de Computadores

MEMÓRIA CACHE

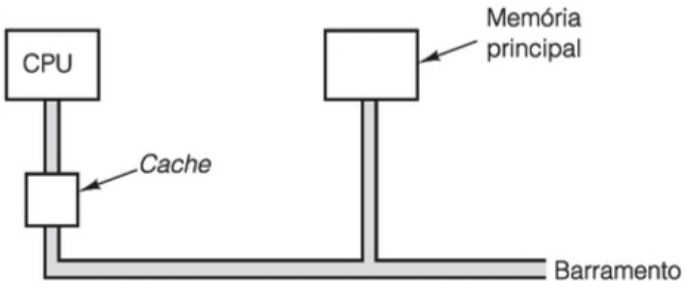
Prof. Alexandre Tannus

Objetivos

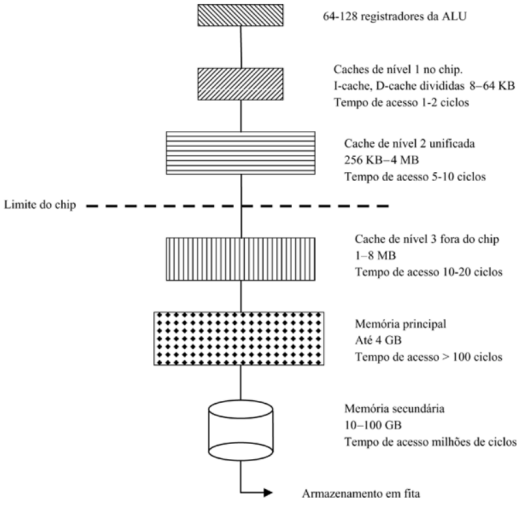
- Descrever a função das memórias cache
- Relatar a evolução da memória cache
- Explicar os princípios de localidade
- Calcular a eficiência de uma memória cache
- Distinguir entre as políticas de alocação, substituição e escrita

Memória Cache

- Memória mais próxima dos registradores
  - Acesso mais rápido
  - Capacidade pequena



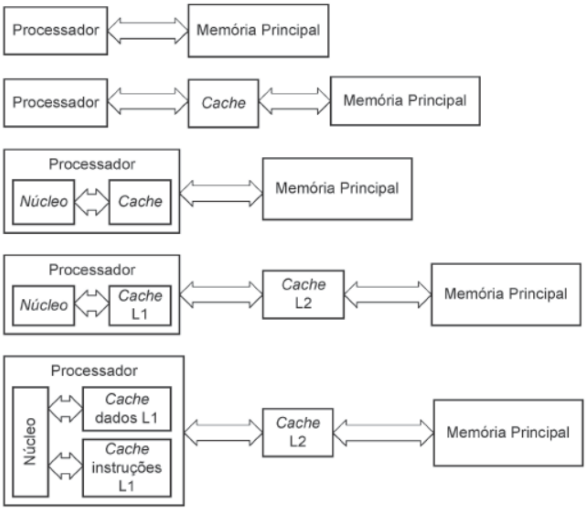
Memória Cache



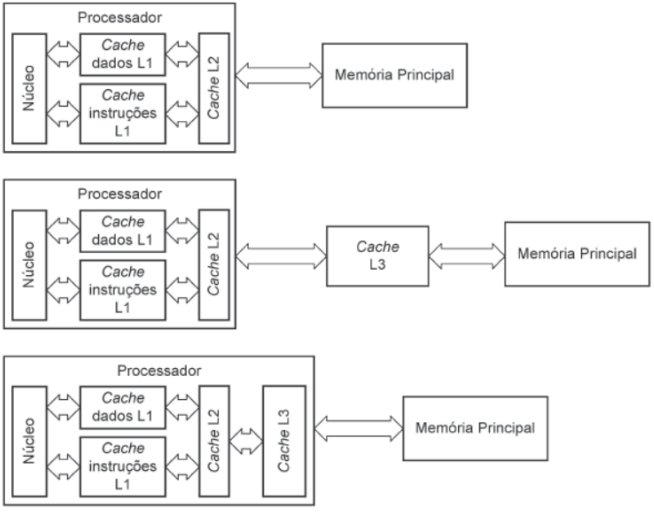
Memória Cache

- Primeiro local que o processador busca informações (dados ou instruções)
- Pode estar localizada dentro ou fora do processador
- Possibilidade de vários níveis
  - Itanium 64 bits: 4 níveis (L1, L2, L3, L4)

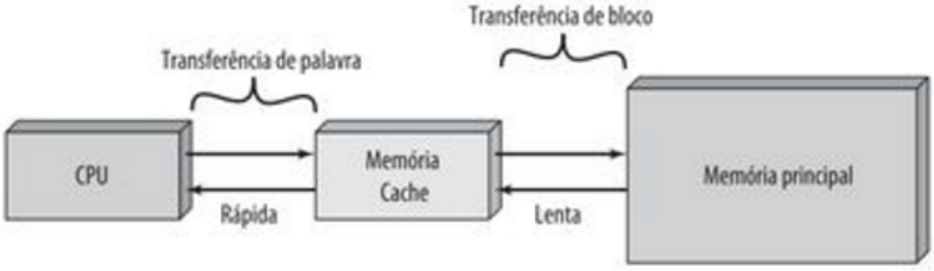
Evolução do cache



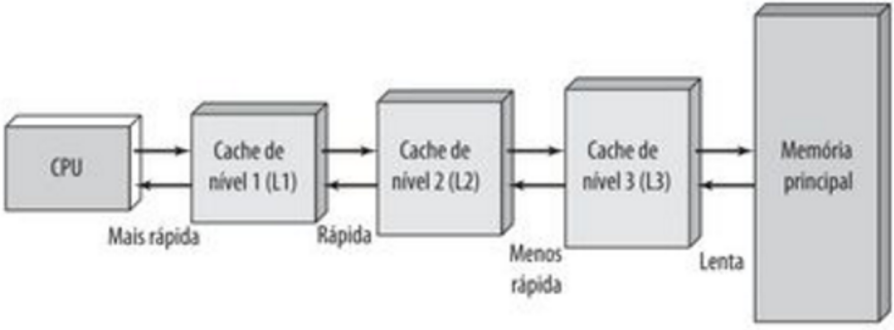
Evolução do cache



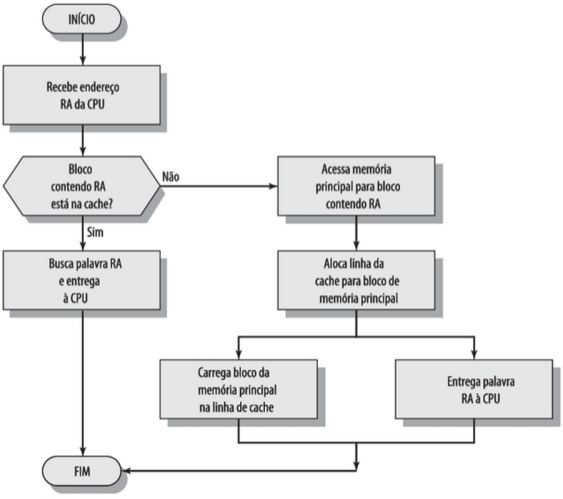
Memórias cache de 1 nível



Memórias cache de vários níveis



Funcionamento



Princípio da localidade

- Localidade espacial
  - Quando um determinado item é referenciado, itens com endereços de memória próximo a ele tendem a ser logo referenciados

- Localidade temporal
  - Quando um determinado item é referenciado, a tendência é que ele seja novamente referenciado dentro de um curto período de tempo

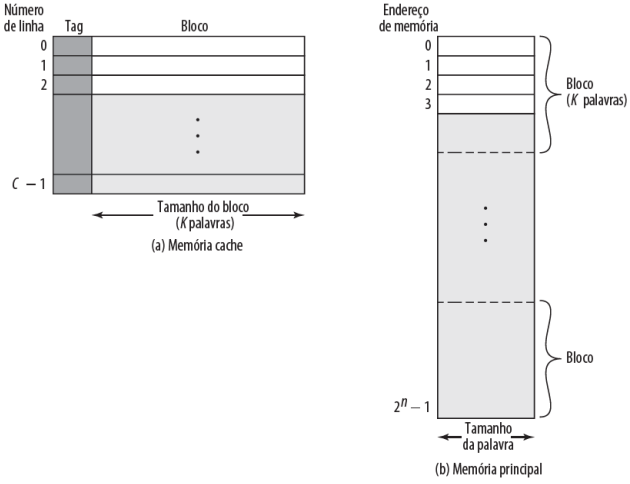
Parâmetros importantes

- Tamanho da cache
- Tamanho dos blocos
- Taxa de sucesso (*hit ratio*) / insucesso (*miss ratio*)
- Política de alocação
- Política de substituição
- Política de escrita

Blocos de memória

- A memória cache é organizada em blocos (linhas de cache)
- Armazena cópias de parte da memória principal
- Contém um identificador (*tag*) da posição da memória principal onde se encontram os dados

Blocos de Memória



Taxa de sucesso/insucesso

- Taxa de sucesso (*hit ratio*)
  - Percentual de vezes que a palavra buscada está presente na cache
  - Ideal: acima de 95
- Taxa de insucesso (*miss ratio*)
  - Percentual de vezes que a palavra buscada está ausente na cache

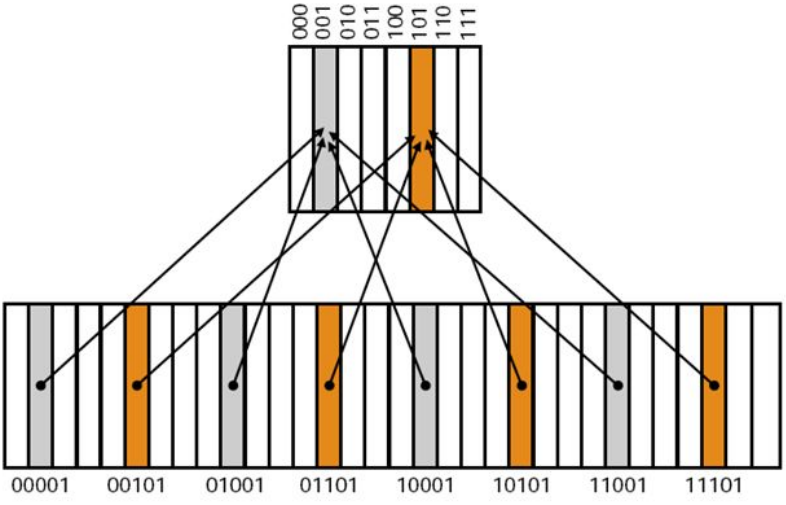
Política de Alocação

- Determina o local de armazenamento de uma linha de dados
- Políticas mais flexíveis geram maior custo de hardware

Mapeamento Direto

- Forma mais simples de localização do bloco
- Utiliza os P bits menos significativos do número do bloco como índice
- Cada bloco da memória principal mapeia uma linha exclusiva de cache
- Vantagem
  - Simplicidade de implementação
- Desvantagem
  - *Thrashing* - Baixa taxa de sucesso e contínua troca de blocos na cache

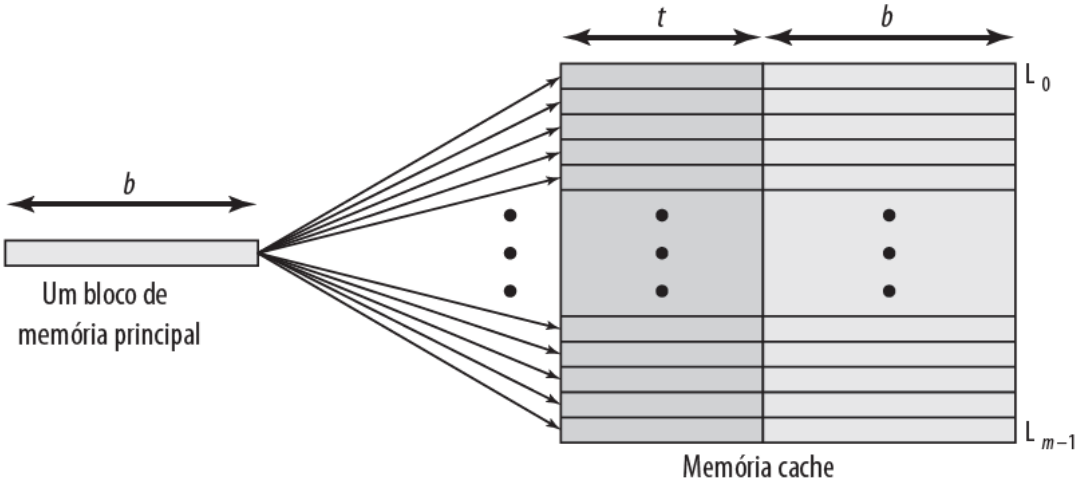
Mapeamento Direto



Mapeamento Associativo

- Permite o carregamento de um bloco em qualquer linha de cache
- Linhas são substituídas somente em caso de memória cheia
- Vantagem
  - Melhor utilização da cache
  - Maior taxa de acertos
- Desvantagem
  - Circuitos de controle mais complexos
  - Diminuição da capacidade de armazenamento

Mapeamento associativo



Política de substituição

- Determina qual bloco será sobrescrito caso a cache esteja cheia
  - *First In First Out* - FIFO
    - O bloco mais antigo na cache será substituído pelo novo
  - *Least Recently Used* - LRU
    - O bloco menos utilizado recentemente é substituído
  - *Least Frequently Used* - LFU
    - O bloco menos frequentemente utilizado é substituído
  - Escolha Aleatória

Políticas de Escrita

- Escrita imediata (*write-through*)
- Escrita atrasada (*write-back*)

Escrita imediata (write-through)

- Memória principal e cache são atualizadas simultaneamente.
- Vantagem
  - Consistência da cache em relação à memória principal
- Desvantagem
  - Pode causar alto tráfego de memória

Escrita atrasada (write-back)

- Escrita apenas na memória cache
- Memória principal é atualizada quando a linha é substituída
- Vantagem
  - Consistência da cache em relação à memória principal
- Desvantagem
  - Pode causar alto tráfego de memória

Coerência de cache

- Problema: Memórias compartilhadas podem apresentar inconsistência nos dados

- Soluções para o problema (coerência)
  - Observação do barramento com *write-through*
  - Transparência de hardware
  - Memória não cacheável

Questão - TJ-RJ 2012

As designações L1 e L2 são utilizadas em referência à memória de computadores. A seu respeito é correto afirmar que

a memória L1 tem menor latência que memória L2.

b memória L1 tem maior latência que memória L2.

c todo computador tem ambos os tipos de memória.

d nenhum computador pode ter ambos os tipos de memória.

e L1 e L2 designam níveis de memória virtual.

Se a referência à memória é para um endereço determinado, é possível que a próxima referência à memória seja feita nas adjacências desse endereço. Trata-se de uma afirmação relevante ao princípio que forma a base de todos os sistemas cache, denominado princípio da

a referência.
b localidade.
c temporalidade.
d latência.
e velocidade.

Qual característica NÃO se refere à memória cache de processadores?

a Tem o objetivo de reduzir o tempo de acesso à memória principal.
b Os dados nela armazenados são cópias de parte da memória principal.
c É implementada pelo sistema operacional com suporte do hardware.
d Pode ser inserida diretamente no chip do processador.
e É comumente encontrada em processadores RISC.

I O número de blocos da memória principal é igual ao número de linhas da memória cache.

II No mapeamento direto, é possível que dois acessos recentes façam referência a blocos alocados para mesma linha da memória cache, o que provoca a retirada de um bloco que acabou de ser trazido da memória principal.

III Na estratégia de mapeamento associativo, o bloco trazido da memória principal pode ser alocado em qualquer linha da memória cache, de acordo com uma política de substituição de linhas definida.

IV Denomina-se hit quando um dado solicitado não está armazenado na memória cache e, neste caso, o bloco da memória principal que contém o byte desejado é transferido para a memória cache.

V A eficiência da memória cache de um sistema de computação em que ocorrem 94 hits a cada 100 acessos é de 6

Questão - EBSERH 2016

Constantemente em material técnico de hardware encontra-se o termo técnico cache. Assinale, das alternativas abaixo, a única que identifica corretamente uma breve descrição de cache:

a uma área de armazenamento temporária onde os dados frequentemente utilizados são armazenados para acesso rápido.

b para garantir a qualidade dos dados que circulam entre memórias e processador, o cache armazena temporariamente os dados, processa algoritmos de segurança e somente repassa dados seguros para o próximo dispositivo.

c uma memória de grande capacidade de armazenamento, de alta velocidade e custo bastante baixo.

d uma área de armazenamento semelhante a ROM onde os dados são frequentemente acedidos por meio de estatísticas dos dados mais acessados.

e uma memória de pequena capacidade de armazenamento, de baixa velocidade e custo bastante alto.

Questão - Petrobras 2012

A técnica de atualização da memória cache, na qual as escritas são feitas apenas nessa memória, e a memória principal só é atualizada se o bit de atualização do bloco substituído tiver o valor 1, é denominada

a write-through
b write-back
c write-on-update
d write-if-updated
e write-when-updated

Bibliografia

Irv Englander.
A arquitetura de hardware computacional, software de sistema e comunicação em rede: uma abordagem da tecnologia da informação.
LTC, Rio de Janeiro, 2011.

Renato Rodrigues Paixão.
Arquitetura de computadores.
Érica, São Paulo, 2014.

William Stallings.
Arquitetura e Organização de Computadores.
Pearson, São Paulo, 8 edition, 2010.


UniEVANGÉLICA
UNIVERSIDADE EVANGÉLICA DE GOIÁS