Organização de Computadores
Unidade 2.2 – Memória Cache

Prof. Rafael Milbradt

Programa

- 2.2.1 – Introdução
- 2.2.2 – Conceito de Localidade
- 2.2.3 – Organização e Funcionamento da Cache
- 2.2.4 – Elementos de Projeto das Memórias Cache
- 2.2.5 – Algoritmos de Substituição de Dados na cache

Introdução

- Diferença de Velocidade entre Processador e MP;
  - Processador chega a conseguir executar 100 instruções enquanto é feito um acesso à MP.
  - Capacidade dos processadores dobra a cada 18 meses, já a velocidade das memórias aumenta 10% a cada ano;
  - DRAM: RAM dinâmica;
    - Lenta e barata;
    - 1 transistor + 1 capacitor;
    - Precisa ser refrescada de tempos em tempos;
  - SDRAM: DRAM Síncrona → está sempre pronta.
  - DDR2, DDR3, RAMBus → continuam tendo 1 transistor + 1 capacitor, por isto continuam mais lentas que a UCP;

Introdução

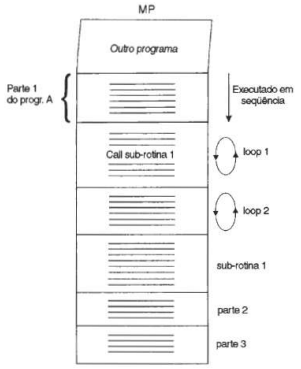
- Como cada instrução precisa fazer pelo menos um acesso à MP, o problema é muito grave;
- SRAM: RAM Estática
  - 7-10 transistores por bit;
  - Mais espaço, mais energia, maior aquecimento;
  - Maior custo;
  - Velocidade compatível com a do processador;
- SRAM pode ser utilizada de forma redundante com a MP:
  - Espaço de endereçamento de acesso rápido;
- Se a SRAM não pode ter o tamanho da MP, ela trará benefício à UCP?

Localidade

- Programas são compilados e colocados na MP sem muitas alterações significativas;
  - Instruções ocupam espaço contíguo na MP e vão sendo executadas uma após a outra;
    - CI vai sendo incrementado;
  - A menos que ocorra um desvio: condicional, loop, jump, call, o próximo acesso à MP é previsível;

Localidade

- Pesquisadores perceberam que diferentes aplicações como as científicas, comerciais, etc. apresentam as características de localidade de acesso.
- Este o **princípio da localidade**;

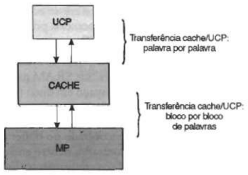


Localidade

- Localidade pode ser decomposta em duas modalidades:
  - Localidade espacial:
    - O programa acessa uma palavra e tem uma grande chance de acessar a palavra seguinte, ou uma palavra próxima;
    - O próprio hardware do processador é construído levando em conta este princípio – após a busca de uma instrução CI é incrementado;
  - Localidade temporal:
    - O programa acessa uma palavra e tem uma grande chance de acessar aquela mesma palavra em um tempo próximo;
    - Loops são exemplo da modalidade temporal;

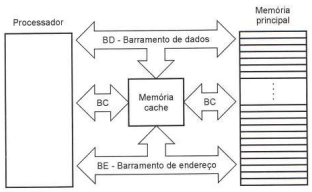
Organização da Cache

- Como aproveitar os dois princípios da localidade?
- O projetista cria um elemento entr a UCP e a MP, chamada memória cache.
  - Veloz e pequena, porém suficientemente grande para armazenar partes inteiras de um programa;
  - Obter o máximo rendimento dos princípios da localidade com baixo custo;
  - Para aproveitar a localidade espacial a MP é dividida em blocos de células;



Organização da Cache

- Funcionamento Genérico de Acesso:
  - Procedimento referente a leitura de 1 byte de dados pela UCP (1 célula da MP);

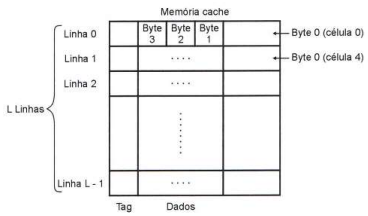


Organização da Cache

1. Processador inicia a operação de leitura colocando o endereço no BE;
2. Sistema de Controle da Cache intercepta e interpreta o conteúdo do BE;
3. Da interpretação a cache verifica se o dado do endereço está nela. Se estiver, imediatamente o coloca no BD, sinaliza no BC e contabiliza um acerto (hit);
4. Se não estiver contabiliza uma falta (miss).
   1. Controle da MP é acionado para localizar a bloco que contém a célula desejada;
   2. Bloco da MP é transferido para uma linha da cache;
   3. Célula é transferida para a UCP → acesso demorado (cache miss);
5. Considerando o princípio da localidade onde próximos acessos serão a endereços próximos o transporte de dados da cache/MP é realizado em blocos/linhas.

Organização da Cache

- O que se deseja é um máximo de *hits* e um mínimos de *misses.*
- *Eficiência da Cache = Acertos / total de acessos * 100*



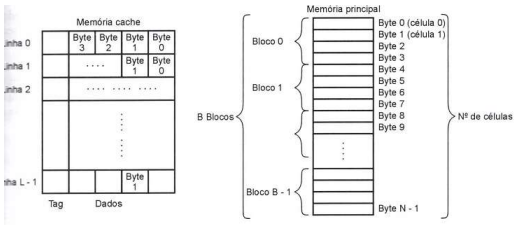
Organização da Cache

- Organização genérica das memórias cache:
  - Uma linha da MP consiste em um bloco que é transferido de uma só vez da MP para a cache quando ocorre um *miss.*
  - Bloco deve ser da mesma largura em células da linha;
  - MP é organizada em N células de um byte cada;
  - Para funcionar integrada à cache, a MP é organizada em B blocos de X bytes cada um, cada um deles com um endereço B_i

Organização da Cache

- A memória cache é organizada em um conjunto de L linhas, cada uma constituída por um conjunto de X bytes (tamanho do bloco da MP).
  - As linhas tem um endereço de 0 até L -1;
  - Cada linha possui um campo *tag* ou rótulo, com o endereço do bloco que está armazenado nela naquele instante;
- Como existem X bytes dentro do bloco, cada célula é endereçada pelo endereço do bloco mais o deslocamento dentro do bloco;
- Mesmo existindo poucas linhas nas caches em relação à quantidade de blocos da MP (4 GB vs. 2MB) devido ao princípio da localidade as caches atuais de eficiência da ordem de 98%.
  - Programas não estruturados podem não respeitar o princípio da localidade;

Organização da Cache



Tipos de Uso da Cache

- O termo cache se aplica a outros tipos de memória, que não apenas a utilizada na relação entre a UCP/MP;
- Acessos à disco também podem usar a MP como cache, já que a mesma é milhares de vezes mais rápida que o disco;
- A cache de disco, mesmo usando DRAM ao invés de SRAM pode trazer um aumento excepcional no desempenho do sistema.

Elementos de Projeto da Cache

- Alguns parâmetros da memória cache são decisões de projeto que podem ter grande influência na eficiência destas memórias como:
  - Função de mapeamento de dados MP/Cache;
  - Algoritmos de substituição de dados na cache;
  - Política de escrita pela cache;
  - Níveis de cache;
  - Definição do tamanho das caches L1, L2 e L3;
  - Escolha da largura de linha da cache;

Exercício

1. Defina e diferencie os princípios da localidade.
2. No contexto das memórias cache, defina hit e miss.
3. Defina e diferencie linhas e blocos.
4. Quais são as principais características de projeto de uma memória cache? Explique.
5. O que é a TAG? Pra quê serve?
6. A construção de um programa não estruturado poderá afetar o desempenho de uma memória cache? Explique.