Universidade Federal de Pelotas
Instituto de Física e Matemática
Departamento de Informática
Bacharelado em Ciência da Computação

Arquitetura e Organização de Computadores II

Aula 14

Memória Cache: uso da localidade espacial, projeto de um sistema de memória para suportar cache.

Prof. José Luís Güntzel

guntzel@ufpel.edu.br

www.ufpel.edu.br/~guntzel/AOC2/AOC2.html

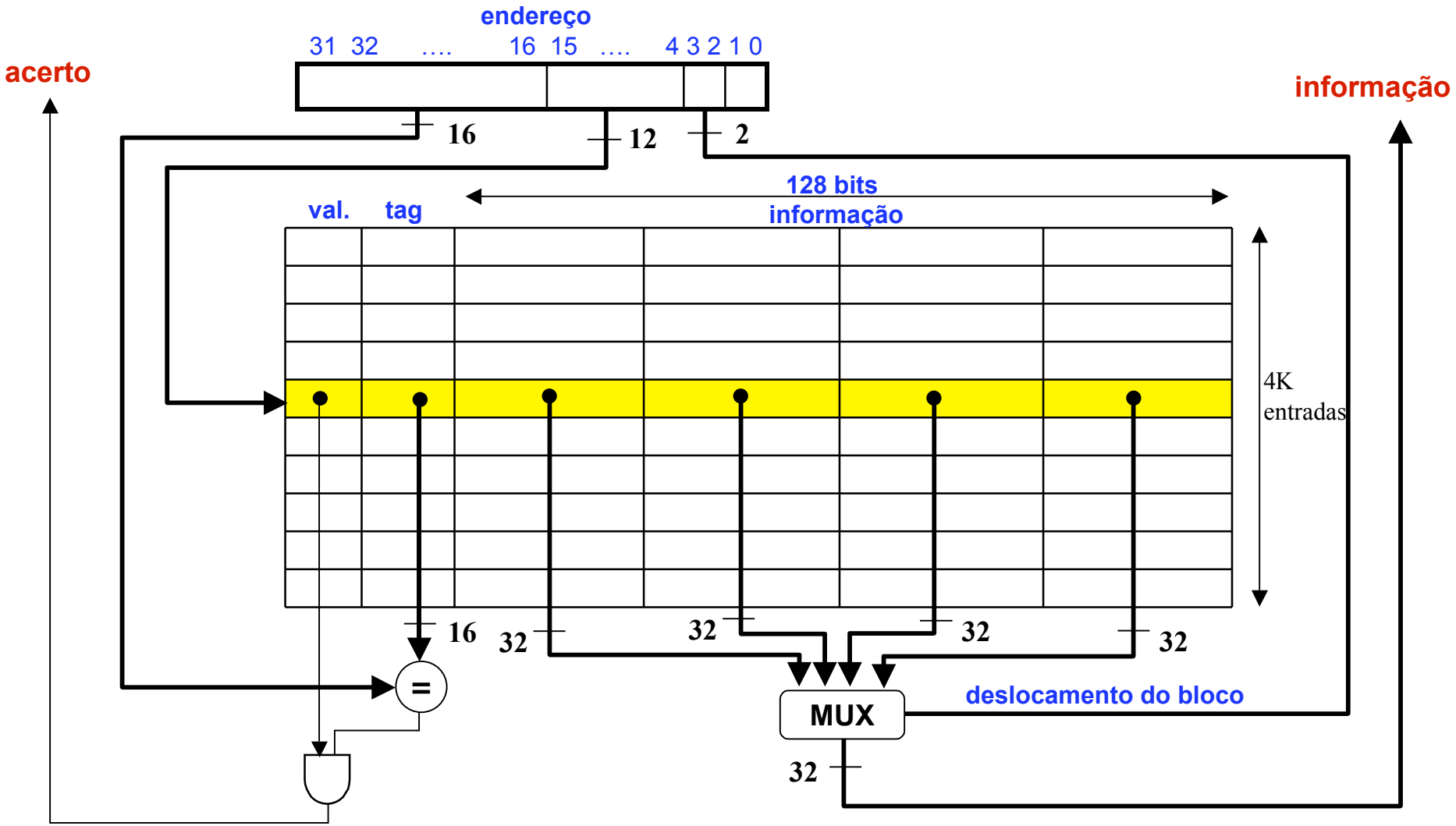
3. Hierarquia de Memória: memória cache

Uso da Localidade Espacial

- **A execução de um programa exibe localidade espacial**
- **Para poder tirar proveito da localidade espacial é preciso que a cache seja organizada em blocos com mais de uma palavra cada**
- **Ao ocorrer uma falta, buscam-se várias palavras adjacentes entre si, as quais têm grande probabilidade de serem necessárias em breve…**

3. Hierarquia de Memória: memória cache

Uma Cache de 64KB (blocos de 4 palavras cada)



3. Hierarquia de Memória: memória cache

Uso da Localidade Espacial

Esta cache:

- **Favorece a exploração da localidade espacial existente nos programas**
- **Apresenta um uso mais eficiente do espaço de armazenamento, pois:**
  - **O número total de bits de flag + bits de validade é menor do que na cache de capacidade equivalente, mas na qual cada bloco tem apenas uma palavra…**

3. Hierarquia de Memória: memória cache

Uso da Localidade Espacial

Como encontrar um bloco correspondente a um endereço em particular?

- **Usar a fórmula**

  (Endereço do bloco) módulo (Número de blocos da cache)

- **O endereço do bloco é simplesmente o endereço da palavra dividido pelo número de palavras no bloco (ou equivalente-mente, o endereço do byte dividido pelo número de bytes no bloco)**

3. Hierarquia de Memória: memória cache

Uso da Localidade Espacial

Exemplo

Considere uma cache com 64 blocos, cada um com 16 bytes. Em qual dos blocos desta cache o endereço 1.200 está mapeado?

Solução:

O bloco é dado por

(Endereço do bloco) módulo (Número de blocos da cache)

onde o endereço do bloco é dado por

(Endereço a byte)/(bytes por bloco)

Considerando a existência de 16 bytes por bloco e uma memória que endereça byte, o endereço 1.200 corresponde ao bloco 1200/16 = 75 , o qual é mapeado na cache no bloco (75 módulo 64) = 11

3. Hierarquia de Memória: memória cache

Uso da Localidade Espacial

- **As faltas geradas por leitura são processadas da mesma maneira em caches com bloco de uma palavra ou com blocos com mais de uma palavra (porém, uma falta causa a transferência de um bloco inteiro para a cache…)**
- **No caso de escrita, acertos e faltas precisam ser tratados de maneira diferente do tratamento feito na DECStation 3100**
- **Considerando o esquema *write-through*, comparar os rótulos do endereço e da entrada da cache:**
  - Se não forem iguais, ocorre uma falta de escrita. Um bloco deverá ser trazido da memória principal e só então a palavra que causou a falta poderá ser escrita na cache.

3. Hierarquia de Memória: memória cache

Uso da Localidade Espacial

O aumento do tamanho do bloco se justifica pela exploração da localidade espacial

- **Em geral, a taxa de faltas cai com o aumento do tamanho do bloco**
- **Supondo memória endereçada a bytes, e cache com blocos de 4 palavras, a falta do endereço 16 vai trazer para a cache o bloco com os endereços 16, 20, 24 e 28. Portanto, será gerada uma única falta para as 4 referências.**

Uso da Localidade Espacial

Taxas de Faltas Geradas pela Execução dos Programas gcc e spice

Programa	Tamanho do bloco, em palavras	Taxa de faltas no acesso a instruções	Taxa de faltas no acesso a dados	Taxa de faltas combinada
gcc	1	6,1%	2,1 %	5,4 %
gcc	4	2,0%	1,7 %	1,9 %
spice	1	1,2 %	1,3 %	1,2 %
spice	4	0,3 %	0,6 %	0,4 %

- A taxa de faltas da cache de instruções cai a uma razão aproximadamente igual ao acréscimo do tamanho do bloco
- O maior decréscimo na taxa de faltas da cache de instruções (em relação à cache de dados) deve-se à melhor localidade espacial apresentada pelas referências a instruções

3. Hierarquia de Memória: memória cache

Uso da Localidade Espacial

- **Mas a taxa de faltas pode aumentar caso o bloco representar uma fração considerável do tamanho total da cache pois**
  - **O número de blocos que podem ser mantidos na cache será pequeno**
  - **Um bloco será retirado da cache antes que muitas de suas palavras tenham sido acessadas**
- **Ver figura 7.12 (página 329 do HW-SW Interface, 2ª Edição)**

3. Hierarquia de Memória: memória cache

Penalidade por Falta

- **À medida que o tamanho do bloco aumenta, aumenta o custo de uma falta**
- **A penalidade por falta é determinada pelo tempo necessário à busca de um bloco no nível imediatamente inferior na hierarquia, carregando-o na cache**
- **O tempo de busca é dividido em duas partes:**
  - **Latência para a busca da primeira palavra**
  - **Tempo de transferência do resto do bloco**
- **O tempo de transferência aumenta com o tamanho do bloco**
- **Solução: projetar o sistema de memória para que este consiga transferir blocos grandes de maneira mais eficiente…**

3. Hierarquia de Memória: memória cache

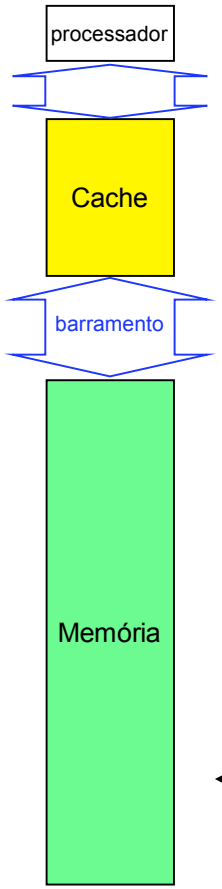
Projeto de um Sistema de Memória
para Suportar Caches

- As faltas no acesso às caches são resolvidas pela memória principal, a qual é construída a partir de DRAMs
- O tempo de acesso nas DRAMs é maior do que nas SRAMs
- Logo, é difícil reduzir a latência da busca da primeira palavra da memória principal
- Mas podemos reduzir a penalidade por falta se **aumentarmos a banda passante da memória principal para a cache**
- Esta redução no custo da penalidade permite o uso de blocos maiores, a um custo próximo daquele obtido com blocos pequenos

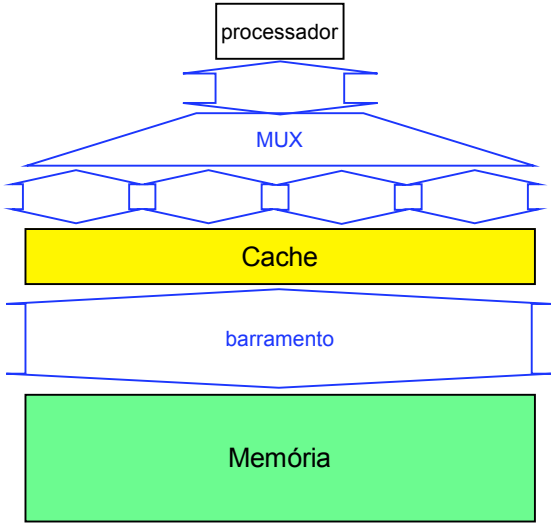
3. Hierarquia de Memória: memória cache

Projeto de um Sistema de Memória para Suportar Caches

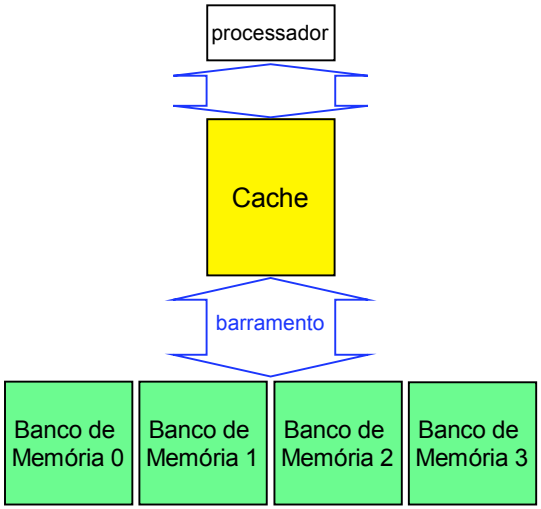
Três Opções para um Sistema de Memória



2. Organização com memória mais ampla e barramento com largura de 4 palavras



3. Organização com memória intercalada e barramento com largura de 1 palavra



← **1.** Organização com memória de uma palavra e barramento com largura de 1 palavra

3. Hierarquia de Memória: memória cache

Projeto de um Sistema de Memória para Suportar Caches

Custo da Penalidade por Falta versus Banda Passante da Memória

Suponha os seguintes tempos de acesso à memória:

- 1 ciclo de relógio para enviar um endereço
- 15 ciclos de relógio para cada acesso à DRAM (tempo de latência do acesso)
- 1 ciclo de relógio para transferência de uma palavra de dados

Suponha também que a cache possui bloco com 4 palavras

Caso	penalidade por falta	Nº de bytes transferidos/ciclos de relógio
1	1 + 4 x 15 + 4 x 1 = 65 ciclos de relógio	(4 x 4)/65 = 0,25 byte/ciclo de relógio†
2	1 + 1 x 15 + 1 x 1 = 17 ciclos de relógio*	(4 x 4)/17 = 0,94 byte/ciclo de relógio
3	1 + 1 x 15 + 4 x 1 = 20 ciclos de relógio	(4 x 4)/20 = 0,80 byte/ciclo de relógio

* supondo memória com 4 palavras e barrramento com largura de 4 palavras

† considerando palavras de 4 bytes, como no caso do MIPS

3. Hierarquia de Memória: memória cache

Projeto de um Sistema de Memória para Suportar Caches

O Tamanho da DRAM Cresce Quatro Vezes a Cada Três Anos

Ano de introdução	capacidade	$ por MB	Tempo de acesso a uma linha/coluna	Tempo de acesso à coluna para uma linha existente
1980	64 Kbit	1500	250 ns	150 ns
1983	256 Kbit	500	185 ns	100 ns
1985	1 Mbit	200	135 ns	40 ns
1989	4 Mbit	50	110 ns	40 ns
1992	16 Mbit	15	90 ns	30 ns
1996	64 Mbit	10	60 ns	20 ns