

Organização de Computadores

Aula 17

Memória cache segunda parte

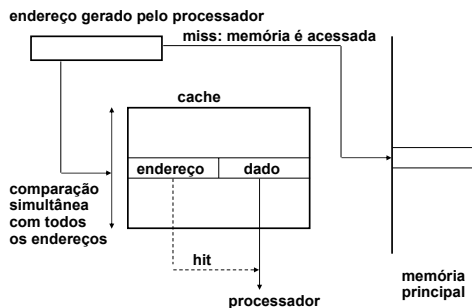
INF01113 - Organização de Computadores

Memória cache segunda parte

1. Mapeamento completamente associativo
2. Mapeamento direto
3. Mapeamento conjunto - associativo

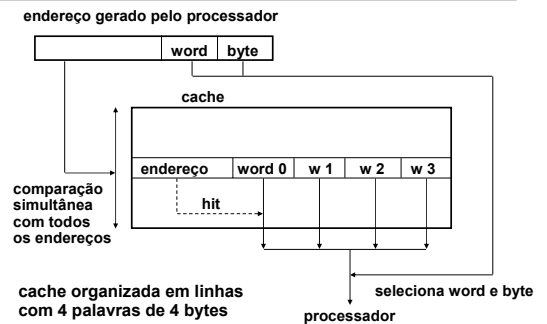
INF01113 - Organização de Computadores

1. Mapeamento completamente associativo



INF01113 - Organização de Computadores

Mapeamento completamente associativo



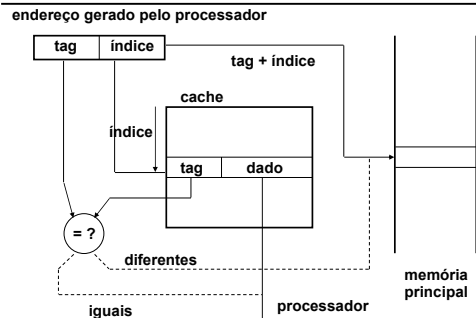
INF01113 - Organização de Computadores

Mapeamento completamente associativo

- vantagem: máxima flexibilidade no posicionamento de qualquer palavra (ou linha) da memória principal em qualquer palavra (ou linha) da cache
- desvantagens
 - custo em hardware da comparação simultânea de todos os endereços armazenados na cache
 - algoritmo de substituição (em hardware) para selecionar uma linha da cache como consequência de um miss
- utilizado apenas em memórias associativas de pequeno tamanho
 - tabelas

INF01113 - Organização de Computadores

2. Mapeamento direto



INF01113 - Organização de Computadores

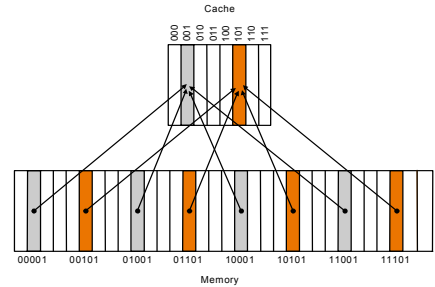
Mapeamento direto

- endereço é dividido em 2 partes
 - parte menos significativa: índice, usado como endereço na cache onde será armazenada a palavra
 - parte mais significativa: tag, armazenado na cache junto com o conteúdo da posição de memória
- quando acesso é feito, índice é usado para encontrar palavra na cache
 - se tag armazenado na palavra da cache é igual ao tag do endereço procurado, então houve hit
- endereços com mesmo índice são mapeados sempre para a mesma palavra da cache

INF01113 - Organização de Computadores

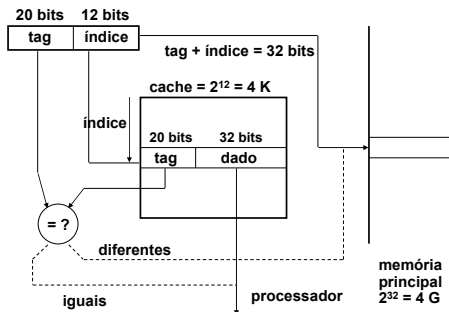
Mapeamento direto

- mapeamento: endereço é o módulo do número de blocos na cache



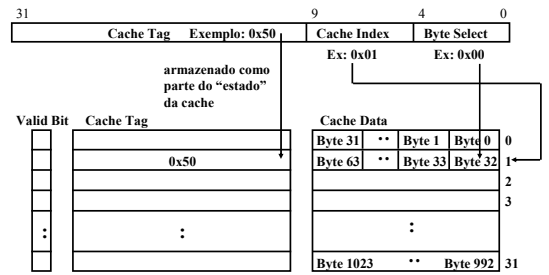
INF01113 - Organização de Computadores

Mapeamento direto – exemplo



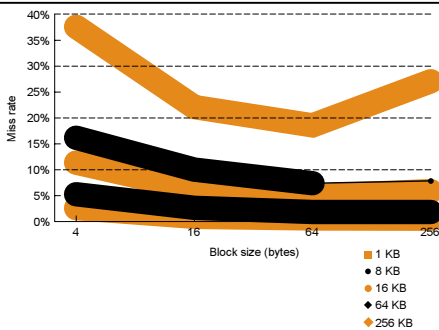
INF01113 - Organização de Computadores

Mapeamento direto – uso de linhas



INF01113 - Organização de Computadores

Tamanho da linha x miss ratio

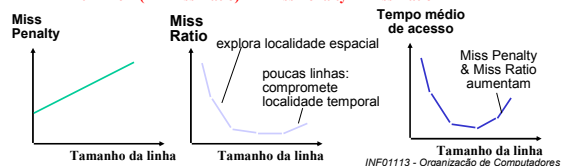


INF01113 - Organização de Computadores

Tamanho da linha

- em geral, uma linha maior aproveita melhor a localidade espacial **MAS**
 - linha maior significa maior miss penalty
 - demora mais tempo para preencher a linha
 - se tamanho da linha é grande demais em relação ao tamanho da cache, miss ratio vai aumentar
 - muito poucas linhas
- em geral, tempo médio de acesso =

$$\text{Hit Time} \times (1 - \text{Miss Ratio}) + \text{Miss Penalty} \times \text{Miss Ratio}$$



INF01113 - Organização de Computadores

Quantos bits tem a cache no total?

- supondo cache com mapeamento direto, com 64 KB de dados, linha com uma palavra, endereços de 32 bits
- 64 KB \rightarrow 16 Kpalavras, 2^{14} palavras, neste caso 2^{14} linhas
- cada linha tem 32 bits de dados mais um tag (32-14-2 bits) mais um bit de validade:
 $2^{14} \times (32 + 32 - 14 - 2 + 1) = 2^{14} \times 49 = 784 \times 2^{10} = 784 \text{ Kbits}$
- 98 KB para 64 KB de dados, ou 50% a mais

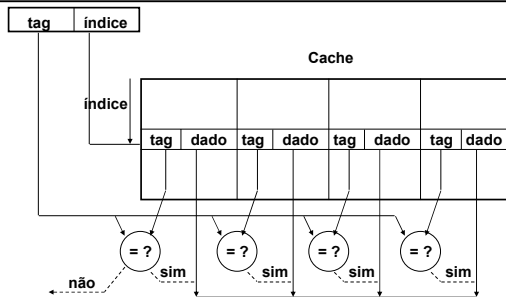
INF01113 - Organização de Computadores

Mapeamento direto

- vantagens
 - não há necessidade de algoritmo de substituição
 - hardware simples e de baixo custo
 - alta velocidade de operação
 - especulação é possível
- desvantagens
 - desempenho cai se acessos consecutivos são feitos a palavras com mesmo índice
 - *hit ratio* inferior ao de caches com mapeamento associativo
- demonstra-se no entanto que *hit ratio* aumenta com o aumento da cache, aproximando-se de caches com mapeamento associativo
 - tendência atual é de uso de caches grandes

INF01113 - Organização de Computadores

3. Mapeamento conjunto – associativo



INF01113 - Organização de Computadores

Mapeamento conjunto – associativo

- mapeamento direto: todas as palavras armazenadas na cache devem ter índices diferentes
- mapeamento associativo: linhas podem ser colocadas em qualquer posição da cache
- compromisso: um n° limitado de linhas, de mesmo índice mas diferentes tags, podem estar na cache ao mesmo tempo (num mesmo conjunto)
- n° de linhas no conjunto = associatividade

INF01113 - Organização de Computadores

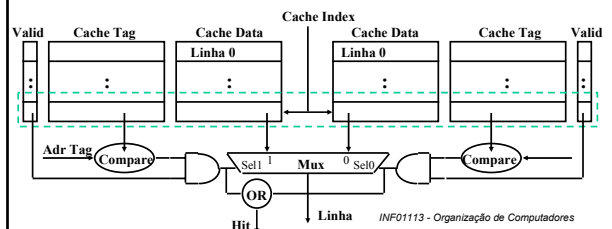
Mapeamento conjunto – associativo

- vantagem em relação ao mapeamento completamente associativo: comparadores são compartilhados por todos os conjuntos
- algoritmo de substituição só precisa considerar linhas dentro de um conjunto
- muito utilizado em microprocessadores
 - Motorola 68040: 4-way set associative
 - Intel 486: 4-way set associative
 - Pentium: 2-way set associative

INF01113 - Organização de Computadores

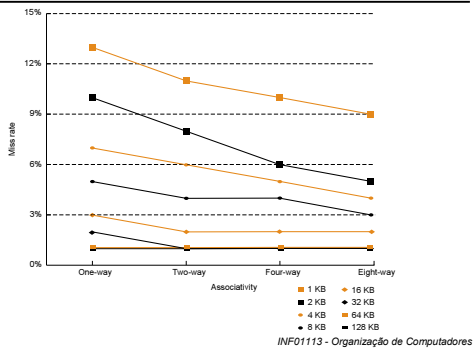
Desvantagem da cache conjunto-associativo

- conjunto-associativa N-way X mapeamento direto
 - dado tem atraso extra do multiplexador
 - dado vem **DEPOIS** da decisão *Hit/Miss* e da seleção do conjunto
- numa cache com mapeamento direto, linha da cache está disponível **ANTES** da decisão *Hit/Miss*
 - possível assumir um *hit* e continuar. Recuperar depois se for *miss*.



INF01113 - Organização de Computadores

Impacto da associatividade da cache

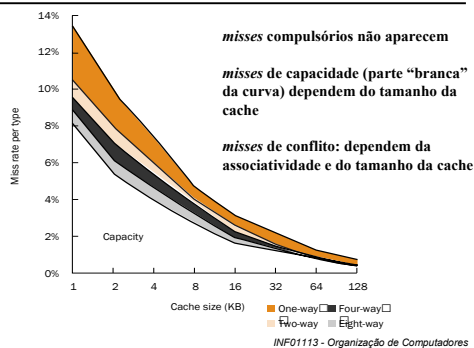


Fontes de misses

- **compulsórios** (*cold start* ou chaveamento de processos, primeira referência): primeiro acesso a uma linha
 - é um “fato da vida”: não se pode fazer muito a respeito
 - se o programa vai executar “bilhões” de instruções, *misses* compulsórios são insignificantes
- **de conflito** (ou colisão)
 - múltiplas linhas de memória acessando o mesmo conjunto da cache
 - conjunto-associativa ou mesma linha da cache com mapeamento direto
 - solução 1: aumentar tamanho da cache
 - solução 2: aumentar associatividade
- **de capacidade**
 - cache não pode conter todas as linhas acessadas pelo programa
 - solução: aumentar tamanho da cache
- **invalidação**: outro processo (p.ex. I/O) atualiza memória

INF01113 - Organização de Computadores

Fontes de misses



Quantidade de misses segundo a fonte

	Mapeam. direto	Conj.-associat. N-way	Compleat. associativa
Tamanho da cache	Grande	Médio	Pequeno
Misses compulsórios	Mesmo	Mesmo	Mesmo
Misses de conflito	Alto	Médio	Zero
Misses de capacidade	Baixo	Médio	Alto
Misses de invalidação	Mesmo	Mesmo	Mesmo

INF01113 - Organização de Computadores