Aula 2

Professores:

Lúcia M. A. Drummond

Simone de Lima Martins

Conteúdo:

Subsistemas de memória

- Memória Cache
- Detalhes

Memória Cache

Diferença de velocidade UCP/MP

- MP transfere bits para a UCP em velocidades sempre inferiores às que a UCP pode receber e operar os dados - tempo de espera na UCP (wait state)
- Difícil de resolver. Enquanto o desempenho dos processadores dobra a cada 18 ou 24 meses, o mesmo não acontece com as memórias DRAM (aumento de 10% a cada ano)
- Tecnologia para aumentar a velocidade de MP é bem conhecida - problema: CUSTO!!

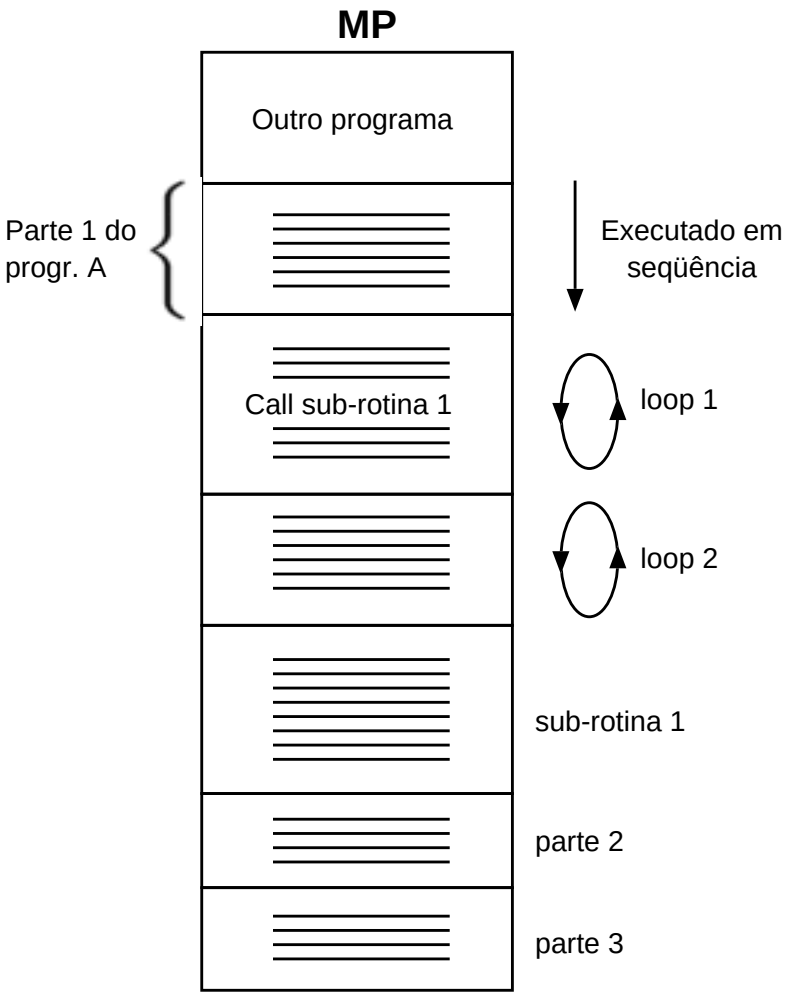
Memória Cache

Conceitos de Localidade

- A execução de programas se realiza, na média, em pequenos grupos de instruções.
- Duas facetas: espacial e temporal
- Localidade espacial: se o programa acessa uma palavra de memória, há uma boa probabilidade de que ele acesse uma palavra subseqüente ou de endereço adjacente
- Localidade temporal: Se um programa acessa uma palavra de memória, há uma boa probabilidade de que em breve ele acesse a mesma palavra novamente

Memória Cache

Conceitos de Localidade

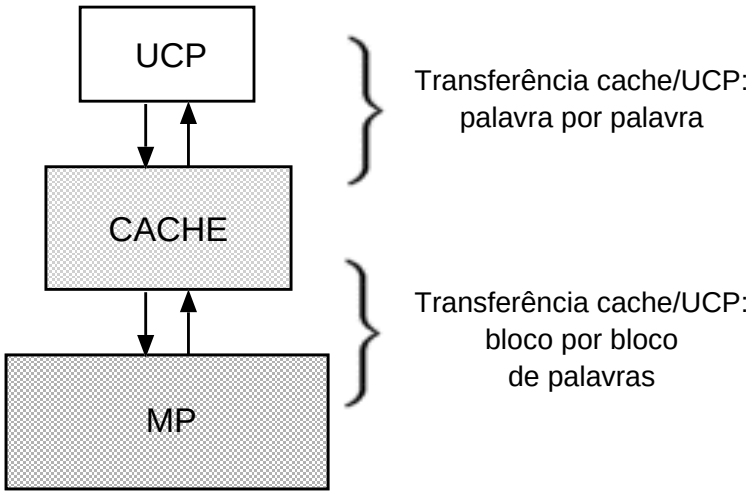


(Fig. 5.20 do livro texto)

Memória Cache

Utilização da Memória Cache

- Sempre que a UCP vai buscar uma nova instrução (ou dado), ela acessa a memória cache
- Se a instrução estiver na cache, chama-se de acerto (hit). Ela é transferida em alta velocidade
- Se a instrução não estiver na cache, chama-se de falta (miss).
  - O grupo de instruções a que a instrução pertence é transferida da MP para a cache, considerando o princípio da localidade

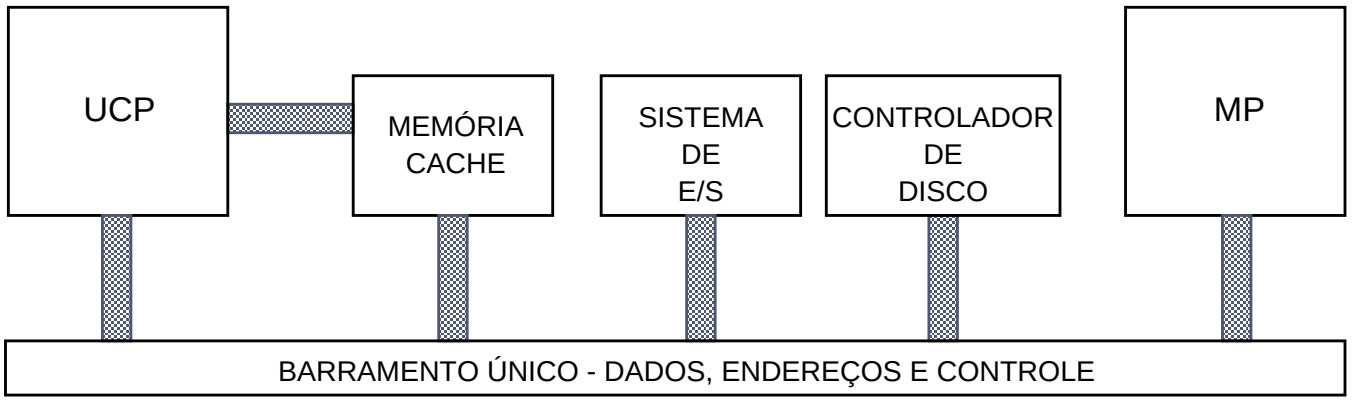


(Fig. 5.21 do livro texto)

Memória Cache

Utilização da Memória Cache

- Para melhorar o desempenho é necessário muito mais acertos do que faltas



(Fig. 5.22 do livro texto)

Memória Cache

Tipos de Memória Cache

Dois tipos básicos de emprego de cache:

- Na relação UCP/MP (cache de RAM)
- Na relação MP/disco (cache de disco)
  - Funciona segundo o mesmo princípio da cache de memória RAM porém em vez de utilizar a memória de alta velocidade SRAM para servir de cache, o sistema usa uma parte da memória principal, DRAM como se fosse um espaço em disco

Memória Cache

Níveis de Cache da Memória RAM

Para não aumentar muito o custo da cache, conforme o aumento da sua capacidade: sistema hierárquico de caches

- Nível 1 ou L1 sempre localizada no interior do processador
- Nível 2 ou L2 localizada em geral na placa mãe, externa ao processador
- Nível 3 ou L3 existente em poucos processadores, localizada externamente ao processador

Cache também pode ser dividida: dados e instrução

Memória Cache

Níveis de Cache da Memória RAM

Processadores	Fabricante	Tamanho
486	Intel	8KB, unificado
C6	Cyrix	64KB, dividido
K5	AMD	24KB, dividido
K7	AMD	128KB, dividido
Pentium	Intel	16KB, dividido
Pentium MMX	Intel	32KB, dividido
Pentium PRO	Intel	16KB, dividido
Power PC601	Motorola/IBM	32KB
Pentium III	Intel	32KB, dividido

Memória Cache

Elementos de Projeto de uma Memória Cache

- Definição do tamanho das memórias cache, L1 e L2
- Função de mapeamento de dados MP/cache
- Algoritmos de substituição de dados na cache
- Política de escrita pela cache

Memória Cache

Elementos de Projeto de uma Memória Cache

Tamanho da Memória Cache (fatores):

- Tamanho da memória principal
- Relação acertos/faltas
- Tempo de acesso da MP
- Custo médio por bit, da MP, e da memória cache L1 ou L2
- Tempo de acesso da cache L1 ou L2
- Natureza do programa em execução (princípio da localidade)

Memória Cache

Elementos de Projeto de uma Memória Cache

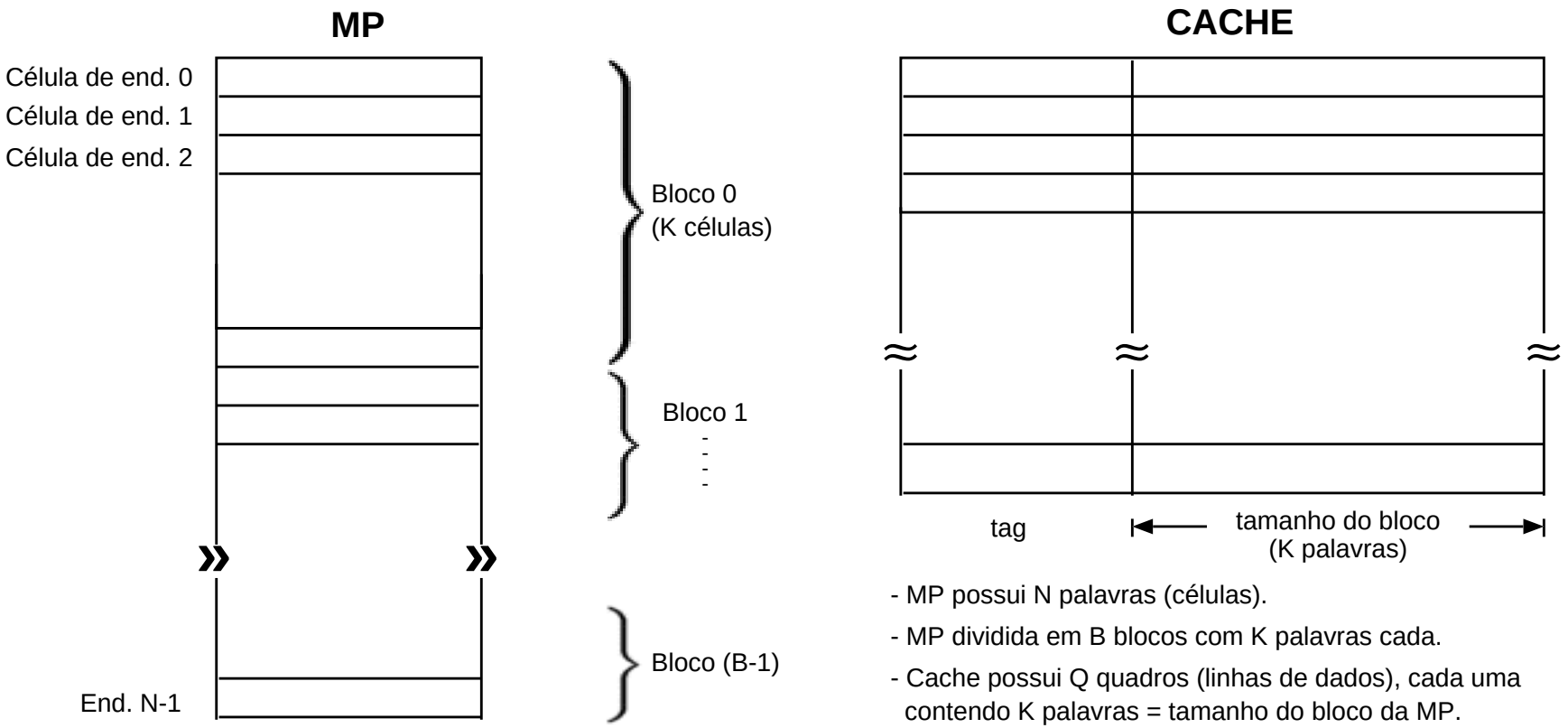
Mapeamento de dados MP/Cache:

- A memória RAM está dividida em conjuntos de B blocos, cada um com K células e a cache com Q linhas, cada uma com K células.
- Q é muito menor do que B
- Para garantir acerto de 90% a 95% - conceito da localidade

Memória Cache

Elementos de Projeto de uma Memória Cache

Mapeamento de dados MP/Cache:



- MP possui N palavras (células).
- MP dividida em B blocos com K palavras cada.
- Cache possui Q quadros (linhas de dados), cada uma contendo K palavras = tamanho do bloco da MP.
- Cada quadro possui um campo(campo TAG) que contém o número de identificação do bloco que está nele armazenado.

(Fig. 5.23 do livro texto)

Memória Cache

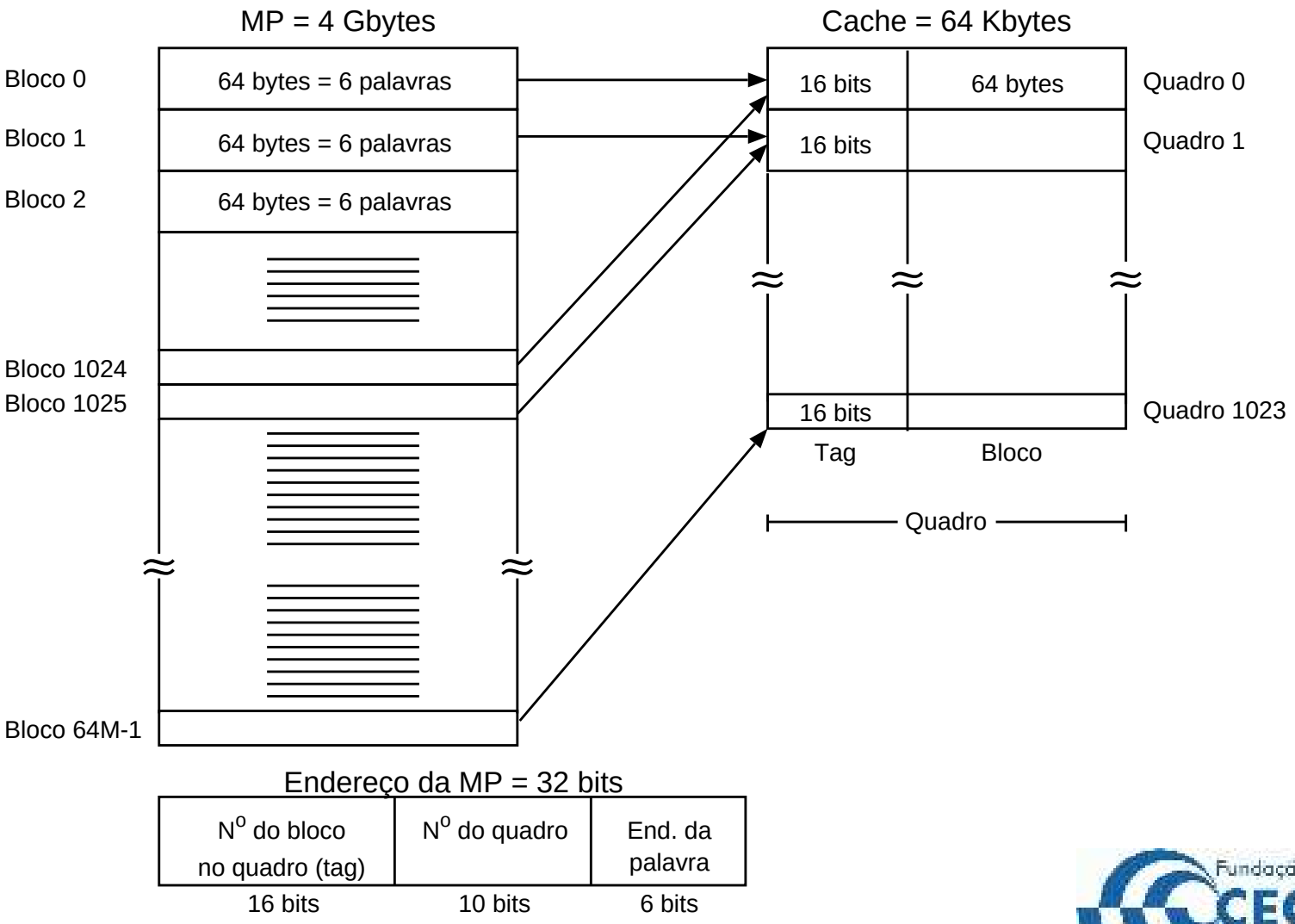
Elementos de Projeto de uma Memória Cache

Para efetuar a transferência de um bloco da MP para uma específica linha da memória cache, escolhe-se um das 3 alternativas:

- Mapeamento Direto
- Mapeamento Associativo
- Mapeamento Associativo por conjuntos

Memória Cache

Mapeamento Direto



(Fig. 5.24 do livro texto)

Memória Cache

Mapeamento Direto

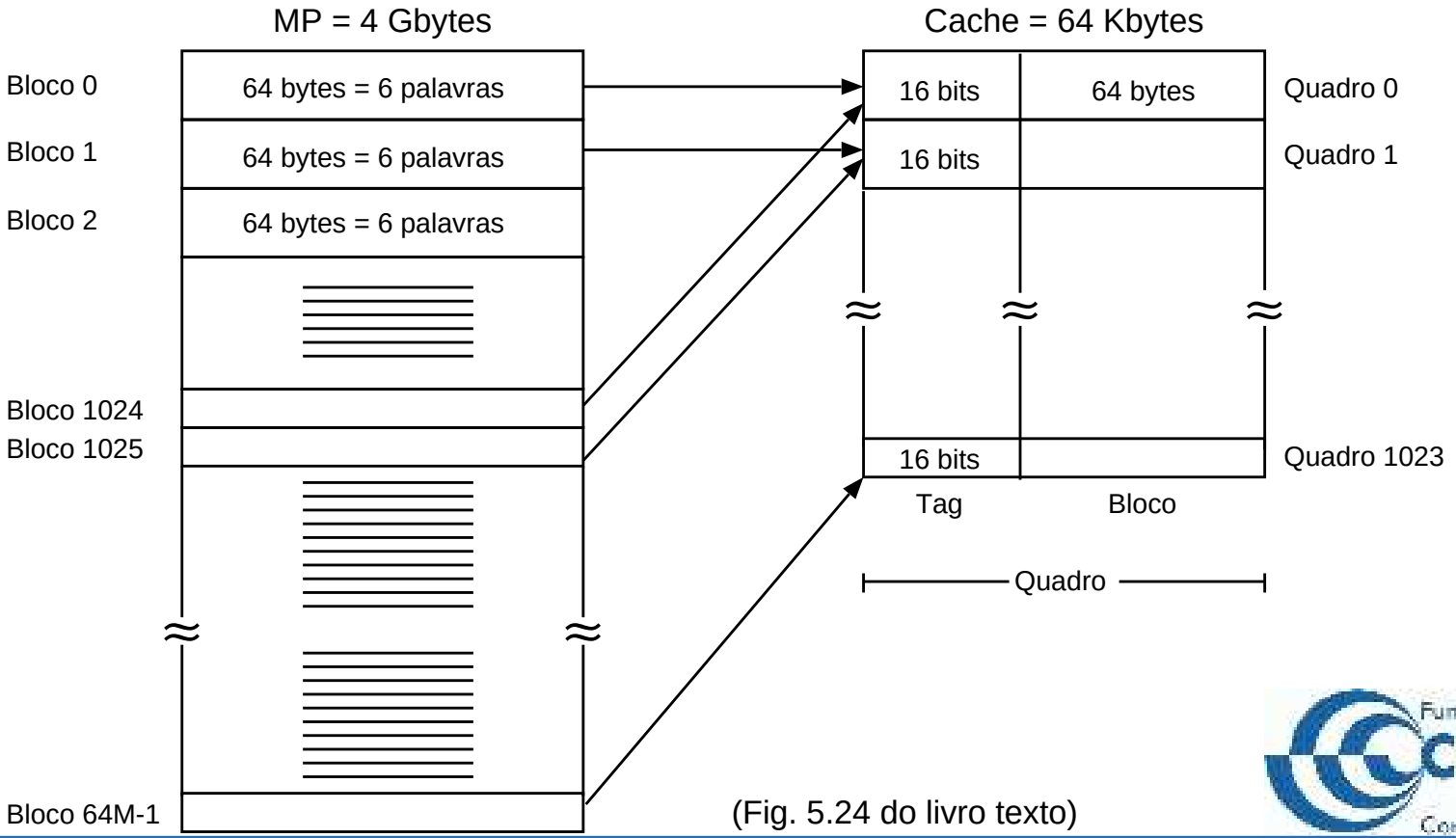
- Cada bloco da MP tem uma linha de cache
- Como há mais blocos do que linhas de cache, muitos blocos vão ser destinados a uma mesma linha

Memória Cache

Mapeamento Direto

Exemplo:

1. MP com 4G palavras (N) e endereços de 32 bits
2. Cache com 64Kbytes, 1024 linhas (Q) com 64 bytes de dados cada uma (K)
3. Número de blocos da MP = B = N/K = 4G/64= 64 M blocos
4. Para localizar um endereço de MP (end) em Cache (EQ): EQ= end módulo 1024



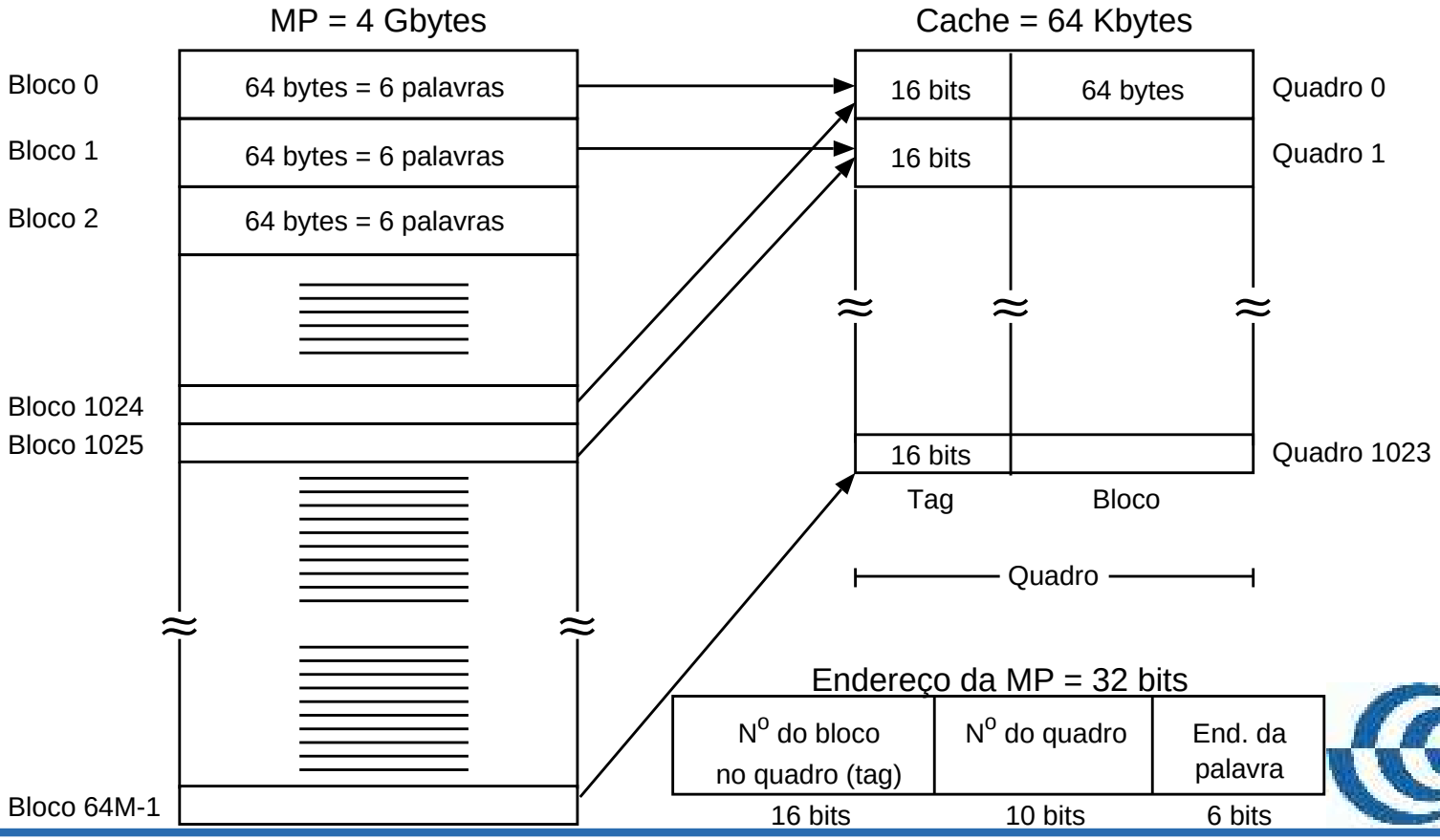
(Fig. 5.24 do livro texto)

Memória Cache

Mapeamento Direto

Cada endereço da memória pode ser dividido da seguinte forma:

- 6 bits menos significativos: indicam a palavra (2^6= 64 palavras no bloco B e na linha Q)
- 10 bits do meio: indicam o endereço da linha da cache (2^{10} = 1024 linhas)
- 16 bits mais significativos: qual o bloco dentre os 64 K blocos que podem ser alocados na linha

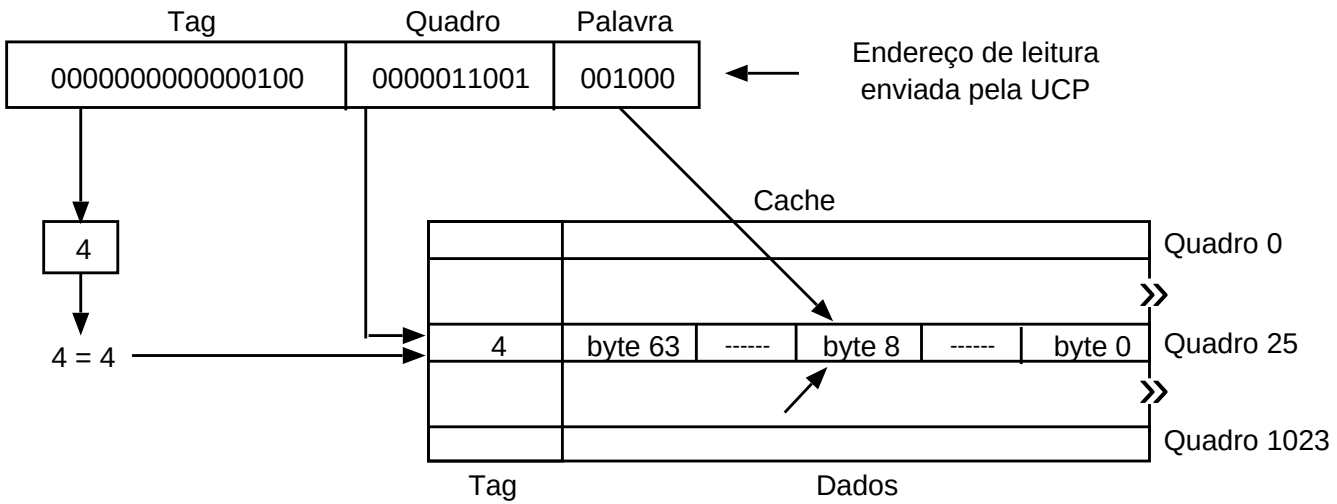


Memória Cache

Mapeamento Direto

Exemplo:

- UCP apresenta endereço de 32 bits ao circuito da cache: 00000000000001000000011001001000
- Parte 1: 0000000000000100 (comparado com o tag do quadro 25 da cache)
- Parte 2: 0000011001 (quadro 25)
- Parte 3: 001000 (palavra 8 é acessada)

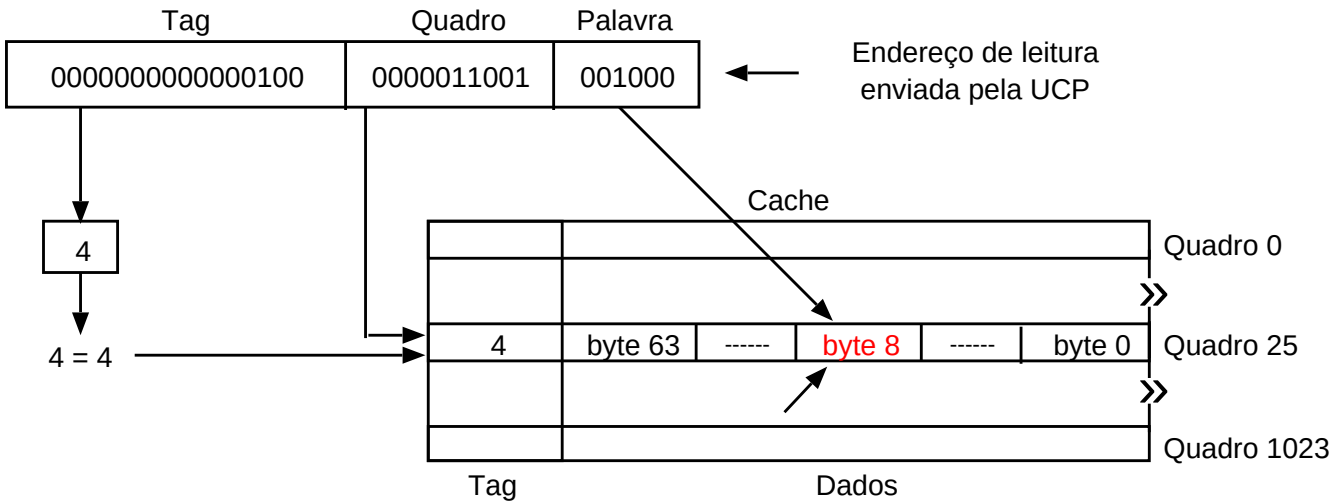


(Fig. 5.25 do livro texto)

Memória Cache

Mapeamento Direto

- Se os valores dos campos tag, do endereço e da linha cache não forem iguais - bloco deve ser transferido da MP para o quadro 25, substituindo o atual bloco.
- Em seguida, o byte 8 é transferido para a UCP
- 26 bits mais significativos são utilizados como o endereço do bloco desejado (2^{26} = 64M)



Mapeamento Direto

Voltar

Memória Cache

Mapeamento Direto

Considerações:

- simples, de baixo custo, não acarreta sensíveis atrasos de processamento de endereços
- Problema: fixação da localização para os blocos (65.536 blocos destinados a uma linha)

Se durante a execução houver repetidas referências a palavras situadas em blocos alocados na mesma linha: muitas substituições de blocos

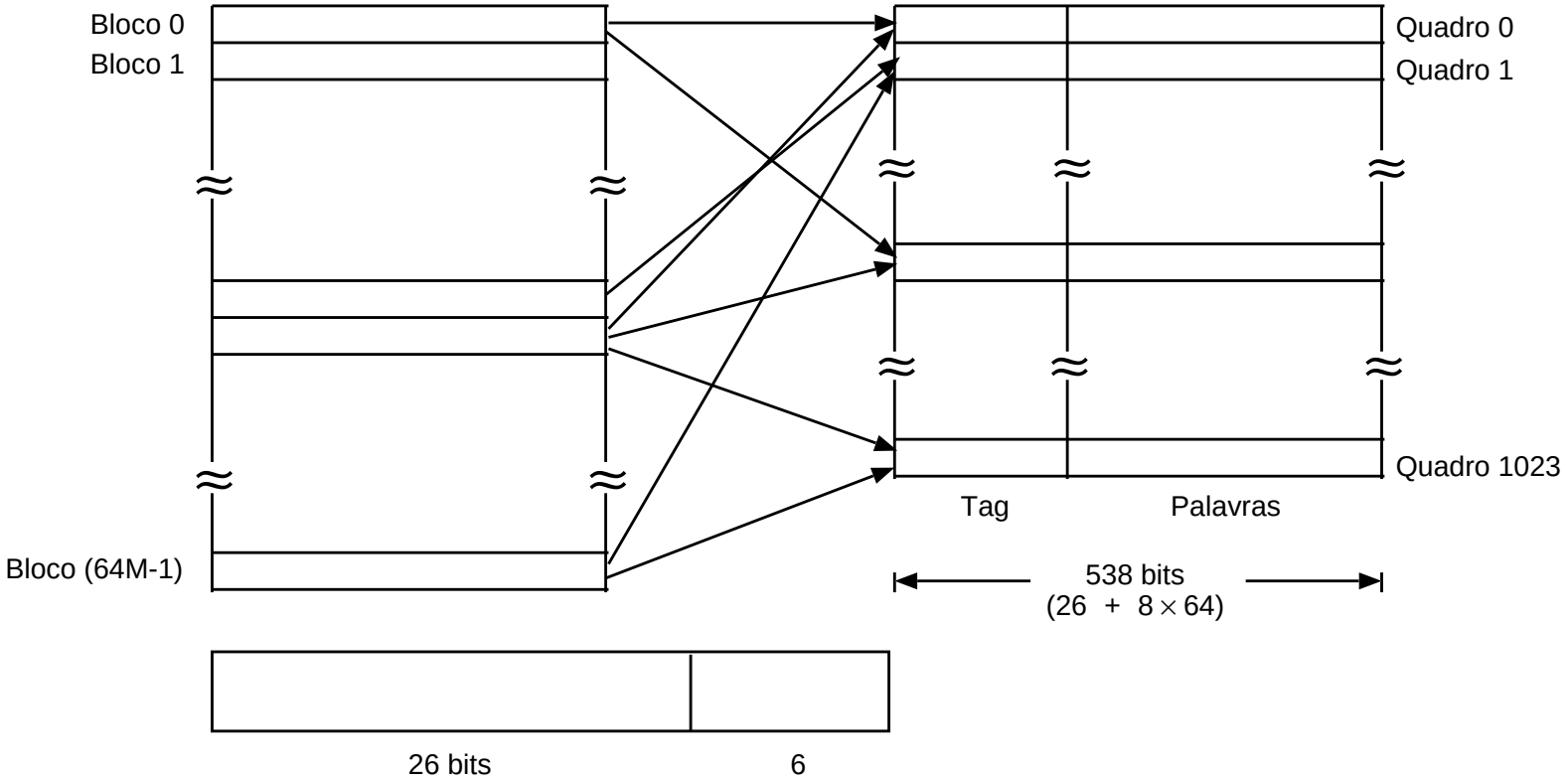
Memória Cache

Mapeamento Associativo

- Os blocos não têm uma linha fixada previamente para seu armazenamento
- Se for verificado que o bloco não está armazenado na cache, este será transferido, substituindo um bloco já armazenado
- Endereço da MP é dividido em duas partes:
- 6 bits menos significativos: palavra desejada
- 26 bits restantes: endereço do bloco desejado

Memória Cache

Mapeamento Associativo



(Fig. 5.26 do livro texto)

Sempre que a UCP realizar um acesso, o controlador da cache deve examinar e comparar os 26 bits de endereço do bloco com o valor dos 26 bits do campo de tag de todas as 1024 linhas.

Memória Cache

Mapeamento Associativo

Considerações:

- Evita a fixação de blocos às linhas
- Necessidade de uma lógica complexa para examinar cada campo de tag de todas as linhas de cache

Memória Cache

Mapeamento Associativo por Conjuntos

- Compromisso entre as duas técnicas anteriores:
  tentar resolver o problema de conflito de blocos e
  da busca exaustiva e comparação do campo tag
- Organiza as linhas da cache em grupos, denominados conjuntos
- Nos conjuntos, as linhas são associativas

Memória Cache

Mapeamento Associativo por Conjuntos

A cache é dividida em C conjuntos de D linhas:

- Quantidade de linhas Q = C x D
- Endereço da linha no conjunto K = E módulo C

Memória Cache

Mapeamento Associativo por Conjuntos

O algoritmo estabelece que:

- O endereço da MP é dividido da seguinte forma:

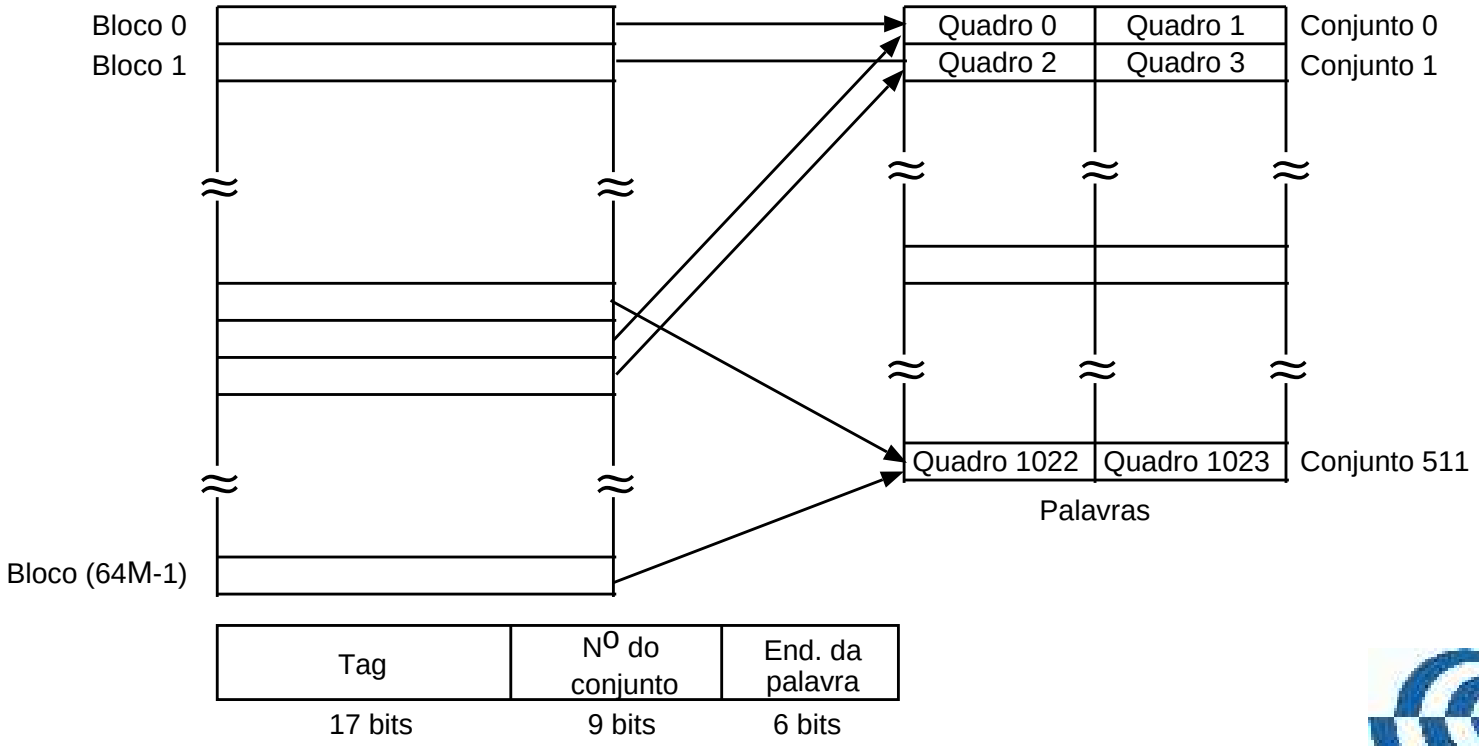
Tag	Número do conjunto	Endereço da palavra
17 bits	9 bits	6 bits

Memória Cache

Mapeamento Associativo por Conjuntos

O algoritmo estabelece que:

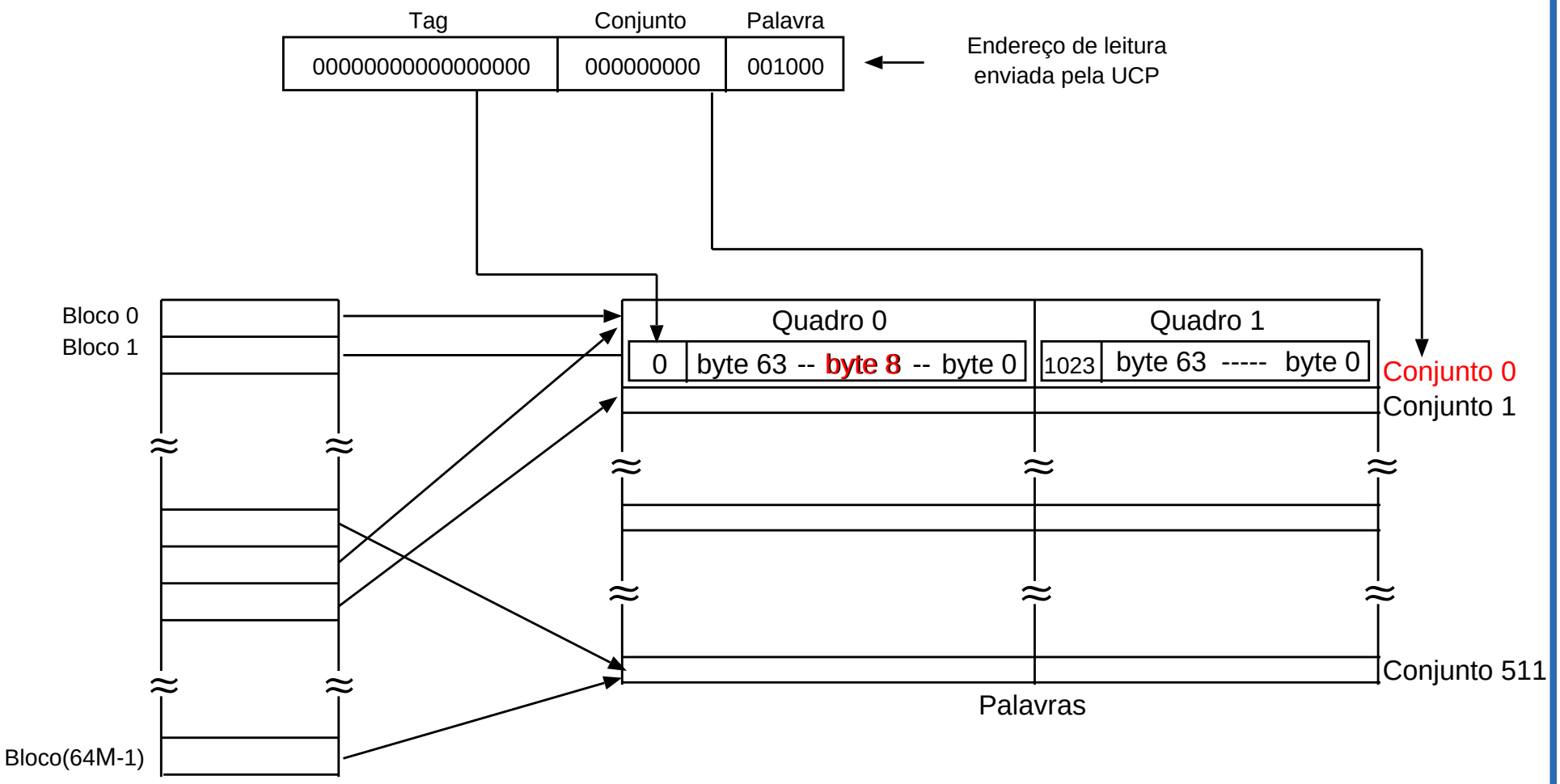
- Ao se iniciar uma operação de leitura, o controlador da cache interpreta os bits do campo de conjuntos para identificar qual o conjunto desejado.
- Em seguida, o sistema compara, no conjunto encontrado, o valor do campo tag do endereço com o valor do campo tag de cada linha do conjunto encontrado.



(Fig. 5.27 do livro texto)

Memória Cache

Mapeamento Associativo por Conjuntos



Exemplo

Voltar

Memória Cache

Algoritmos de Substituição de Dados na Cache

Definir qual dos blocos atualmente armazenados na cache deve ser retirado para dar lugar a um novo bloco que está sendo transferido (já que Q<<B)

Memória Cache

Algoritmos de Substituição de Dados na Cache

Dependendo de qual técnica de mapeamento se esteja usando, pode-se ter algumas opções de algoritmos:

- Se o método de mapeamento for o direto, somente há uma única linha possível para um dado bloco
- Para os outros dois métodos - associativo e associativo por conjunto - existem várias opções

Memória Cache

Algoritmos de Substituição de Dados na Cache

- LRU: o sistema escolhe para ser substituído o bloco que está há mais tempo sem ser utilizado
- FILA: o primeiro a chegar é o primeiro a ser atendido. O sistema escolhe o bloco que está armazenado há mais tempo na cache.
- LFU: o sistema escolhe o bloco que tem tido menos acessos por parte da CPU
- Escolha aleatória: trata-se de escolher aleatoriamente um bloco para ser substituído

Memória Cache

Política de Escrita pela Memória Cache

- Toda vez que a UCP realiza uma operação de escrita, esta ocorre imediatamente na cache.
- Quando atualizar a MP?

Memória Cache

Política de Escrita pela Memória Cache

Considerações:

- MP pode ser acessada tanto pela cache quanto por elementos de E/S. É possível que uma palavra da MP tenha sido alterada só na cache, ou um elemento de E/S pode ter alterado a palavra da MP e a cache esteja desatualizada
- MP pode ser acessada por várias UCP´s. Uma palavra da MP é atualizada para atender à alteração de uma cache específica e as demais caches estarão desatualizadas.

Memória Cache

Política de Escrita pela Memória Cache

Técnicas:

- Escrita em ambas (write through): cada escrita em uma palavra de cache acarreta escrita igual na palavra correspondente da MP
- Escrita somente no retorno (write back):atualiza a MP apenas quando o bloco for substituído e se tiver ocorrido alguma alteração na cache. Uso do bit ATUALIZA
- Escrita uma vez (write once): é uma técnica apropriada para sistemas multi UCP/cache, que compartilhem o mesmo barramento. Primeira atualização: write through + alerta os demais componentes que compartilham o barramento único.

Memória Cache

Política de Escrita pela Memória Cache

Comparações:

- Com write through pode haver uma grande quantidade de escritas desnecessárias na MP
- Com write back, a MP fica desatualizada para dispositivos de E/S, por exemplo, o que os obriga a acessar o dado através da cache (problema!)
- write once é conveniente para sistemas com múltiplas UCP´s

Estudos mostram que a percentagem de escritas na MP é baixa (15%), o que aponta para uma simples política write through

Detalhes

Localização da Célula desejada

- O processo de localização de uma determinada célula para efeito de uma operação de leitura ou escrita é o mesmo, qualquer que seja a tecnologia de fabricação de MP

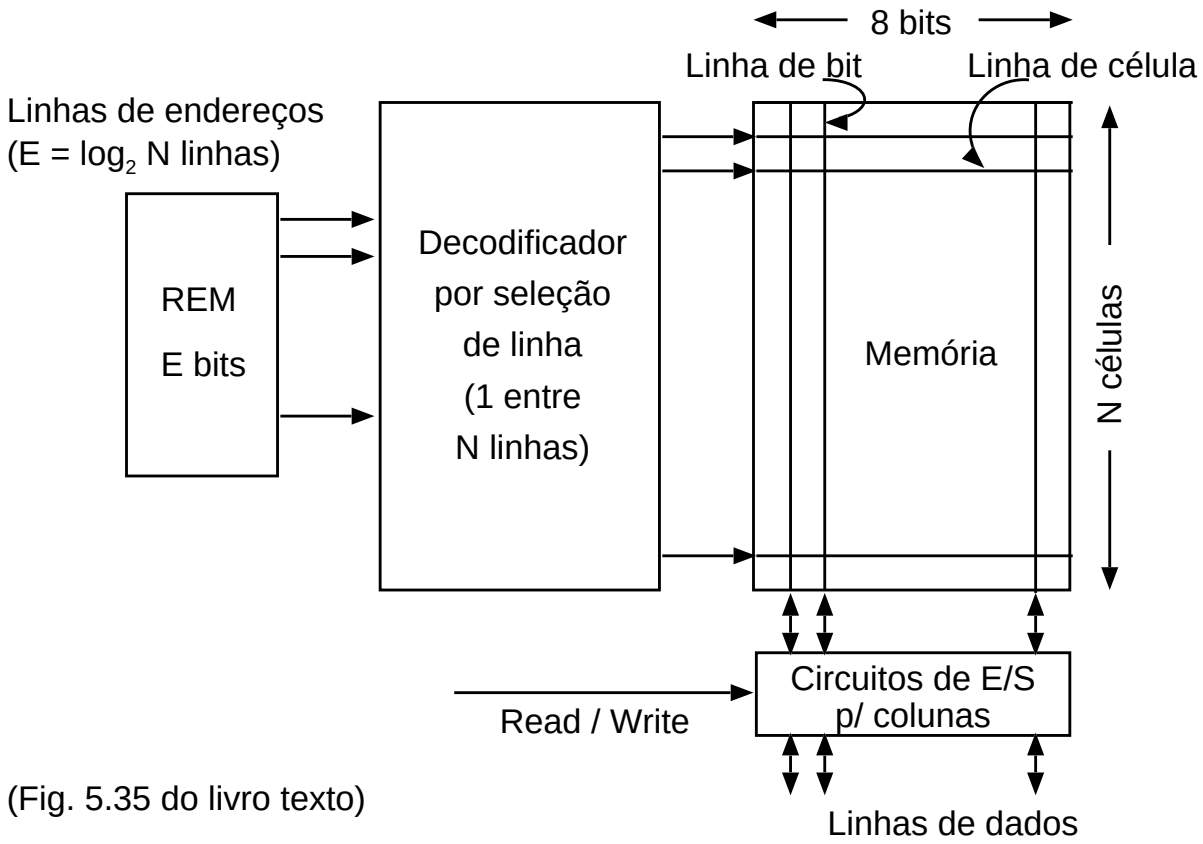
Detalhes

Organização de memória
Tipo Seleção Linear - 1 Dimensão

- Com esta técnica todos os bits de uma dada palavra estão na mesma pastilha
- Arranjo físico é igual ao arranjo lógico: o conjunto é organizado em N palavras de M bits cada
- Ex: uma pastilha de 16 K bits pode conter 1024 palavras de 16 bits cada
- Elementos de cada conjunto são conectados por linhas horizontais e verticais

Detalhes

Organização de memória
Tipo Seleção Linear - 1 Dimensão

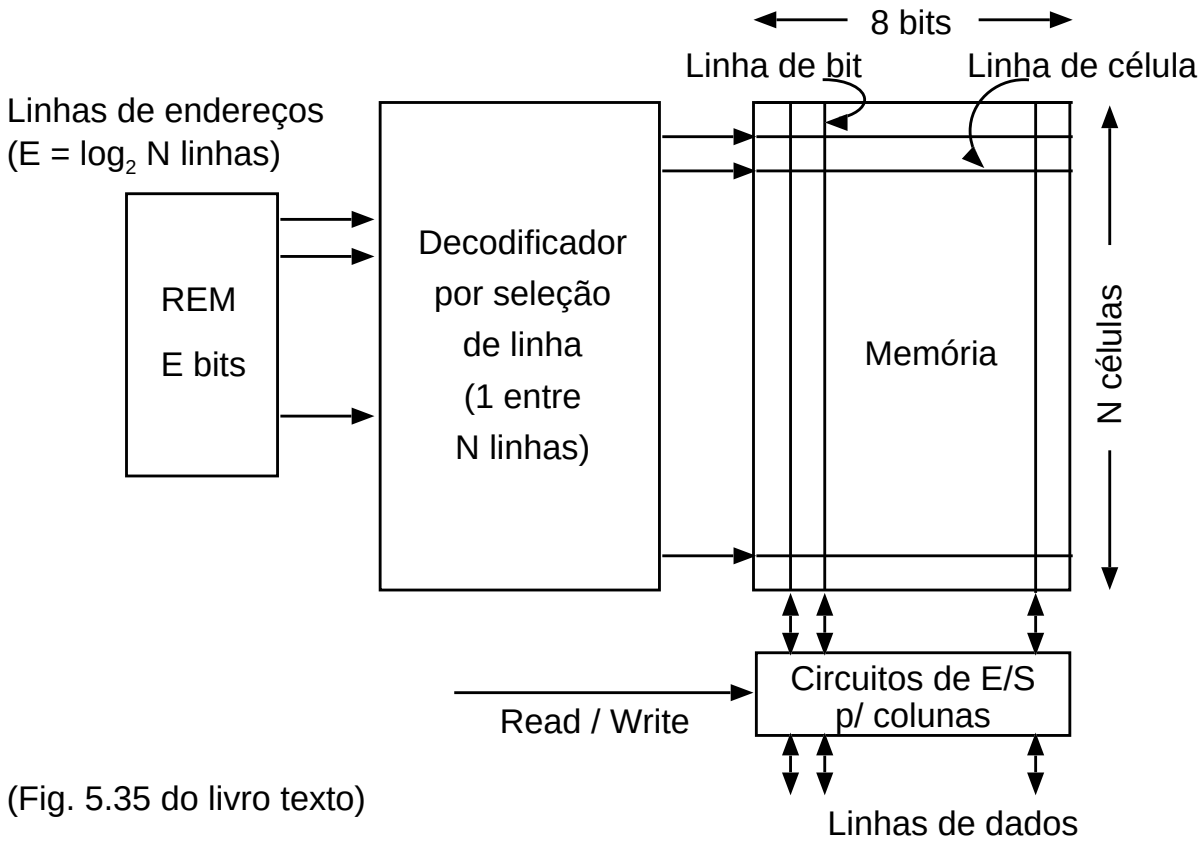


(Fig. 5.35 do livro texto)

- Cada linha horizontal da memória é uma saída do decodificador de linha e cada linha vertical da memória se conecta ao sensor de dados para receber ou enviar 1 bit de palavra
- O decodificador tem 2^E saídas para E entradas, isto é, se cada endereço armazenado no REM tem E bits, a ligação REM-decodificador tem E linhas

Detalhes

Organização de memória
Tipo Seleção Linear - 1 Dimensão



(Fig. 5.35 do livro texto)

- Ex: a pastilha de 16K bits teria um REM de 10 bits, 10 linhas de saída do REM para o decodificar e deste sairiam 1024 linhas, uma para cada célula da memória
- Tomemos, o endereço 12 decimal, armazenado no REM. Isto acarretaria uma saída 1 na 13ª linha do decodificador e as demais linhas seriam iguais zero.

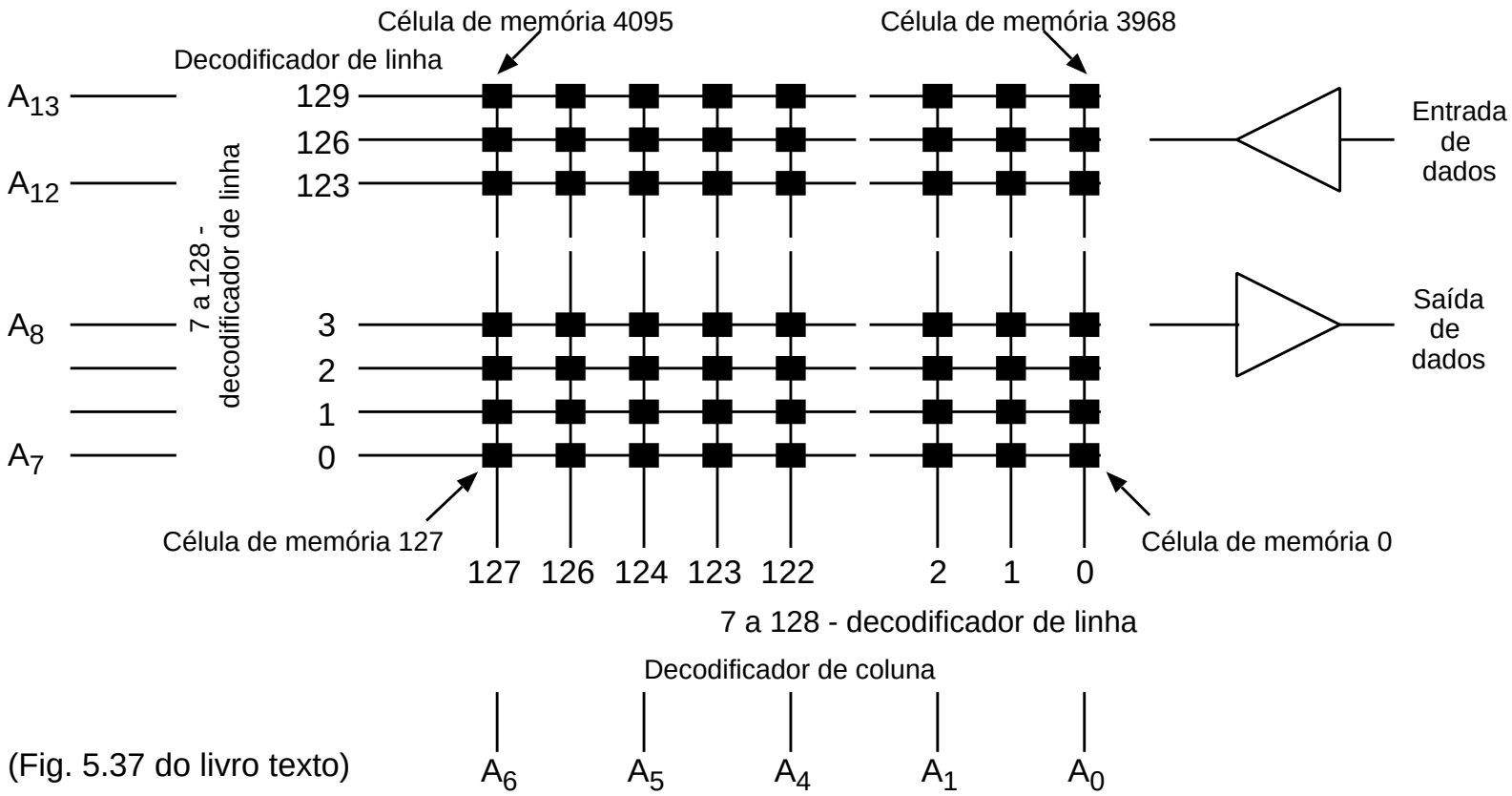
Detalhes

Organização de memória
Tipo Matriz Linha/Coluna

- Uma memória com 16K células, onde cada endereço é um número com 14 bits, visto que 2^{14}=16K
- Cada pastilha possui 14 linhas de endereço, identificadas pela nomenclatura A_0 a A_{13}
- Divididas em duas: A0 a A6 conectadas ao decodificador de colunas, A7 a A13 conectadas ao decodificador de linhas

Detalhes

Organização de memória
Tipo Matriz Linha/Coluna



(Fig. 5.37 do livro texto)

- O cruzamento de linha e coluna ativadas pelos respectivos decodificadores corresponde à célula desejada endereçada pelos 14 bits
- Ocorrem 128x128 cruzamentos = 16K células. Uma memória com 16K células, onde cada endereço é um número com 14 bits, visto que 2^{14} =16K

Detalhes

Organização de memória

Comparação

- Seleção linear: número de linhas de saída muito maior, mais circuitos em uma única pastilha
- Na técnica de matriz por linhas e colunas que seleciona 1 bit de palavra por pastilha são usadas várias pastilhas mais simples para totalizar a quantidade de bits por palavra.

Memória Cache

Exercícios:

Capítulo 5 do livro texto.

Questões:

1 até 15, 17, 20, 21, 23, 24 e 28.