Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Computação
027359 - Arquitetura e Organização de Computadores 1
Prof. Luciano Neris

Exercícios 08 - Respostas

1) Por que um sistema de hierarquia de memória, onde um dispositivo pequeno como o cache fica perto do processador, é eficiente?
SOLUÇÃO:
Devido ao princípio da localidade de referência, onde o processador não faz referência (endereçamento) às palavras de forma aleatória, mas sim de forma localizada, devido a existência de estruturas de dados, armazenamento de programas em endereços contíguos de memória, a execução sequencial de instruções, existência de laços em programas, etc.

2) Quando um programa sequencial é executado, que tipo de localidade de referência (espacial ou temporal) pode ser explorado pelo cache?
SOLUÇÃO:
Como num programa sequencial as instruções estão em endereços subsequentes da memória, a localidade de referência que pode ser explorada pelo cache é a localidade espacial.

3) Quando num programa, um laço é executado repetidas vezes, que tipo de localidade de referência pode ser explorada pelo cache, no que se refere ao acesso às instruções repetidamente?
SOLUÇÃO:
Se as instruções são referenciadas repetidamente, pode ser explorada a localidade temporal.

4) Seja um cache de mapeamento direto com o tamanho de 4 Kbytes, ou 1024 palavras, e tamanho de slot de 4 palavras, inicialmente vazio. Se uma área de memória de 8 Kbytes com endereço inicial igual a 4096 é referenciada pelo processador MIPS em sequência, apenas para leitura, quantos erros e acertos devem ocorrer?
SOLUÇÃO:
Para cada bloco acessado, temos 4 palavras carregadas no cache e portanto, 1 erro e 3 acertos. No total de 8 Kbytes ou 2048 palavras, temos 512 erros e 1536 acertos.

5) Se no exercício anterior, o processador referenciar duas vezes a área de memória de 8 Kbytes, ou seja, faz a leitura de 8 Kbytes completa e repete a operação a) qual seria o número de erros e acertos? b) e se o tamanho do cache fosse de 2048 palavras?
SOLUÇÃO:
a) na repetição da operação de leitura completa de 8 Kbytes, os blocos anteriormente carregados no cache já foram substituídos, pois o tamanho do cache é de 1024 palavras, portanto, o número de erros e acertos é o dobro da questão 4, ou seja, 1024 erros e 3072 acertos.
b) se o tamanho do cache fosse de 2048 palavras, os blocos anteriormente carregados no cache ainda estão no cache, portanto o número de erros é zero, para a repetição da operação de leitura, e o número de acertos seria de 2048. O número total de erros e acertos deve ser a soma dos erros e acertos das duas referências completas, ou seja, 512 erros e 3584 acertos.

6) Para uma organização do cache fosse associativo por conjunto de 2 vias, com tamanho de bloco de 4 palavras, sendo o tamanho do cache de 8Kbytes, se uma área de memória de 8 Kbytes com endereço inicial igual a 4096 é referenciada pelo processador MIPS em sequência, apenas para leitura, quantos erros e acertos devem ocorrer?
SOLUÇÃO:
Se o tamanho do cache é de 8Kbytes e portanto de 2048 palavras, como o cache associativo por conjunto de 2 vias tem 2 slots por conjunto, as primeiras 1024 palavras seriam carregadas em 256 slots do cache preenchendo totalmente um dos slots de cada conjunto. Para as 1024 palavras restantes, o outro slot de cada conjunto seria preenchido, preenchendo totalmente o cache. O número de erros é de 1 para cada carregamento do bloco. As três referências seguintes para as palavras do bloco são acertos. O número total de erros é portanto de 512 contra 1536 acertos.

7) Se no exercício anterior, o processador referenciar duas vezes a área de memória de 8 Kbytes, ou seja, faz a leitura de 8 Kbytes completa e repete a operação a) qual seria o número de erros e acertos?
SOLUÇÃO:
Caso o processador referenciar duas vezes a área de memória de 8 Kbytes, o número de erros seria igual a zero, e o número de acertos seria de 2048, para a segunda referência completa. O número total de erros e acertos deve ser a soma dos erros e acertos das duas referências completas, ou seja, 512 erros e 3584 acertos.

8) Por que num sistema de memória virtual, é usado o TLB (Translation Look-aside Buffer)?

SOLUÇÃO: no sistema de memória virtual, o TLB é usado para a tradução rápida de endereço de memória virtual para a memória física. O TLB é um cache especialmente projetado para essa finalidade e deve conter uma parte da tabela de página mais usada.

9) Seja o exemplo do cache em mapeamento direto da Fig. 1:
e) qual o tamanho de um bloco em bytes?
f) onde as palavras de endereços 0, 5, 17 e 66 são mapeados?
g) onde o bloco de número 8192 é mapeado? qual é o tag nesse caso?
h) qual o endereço da primeira palavra do bloco de número 128?

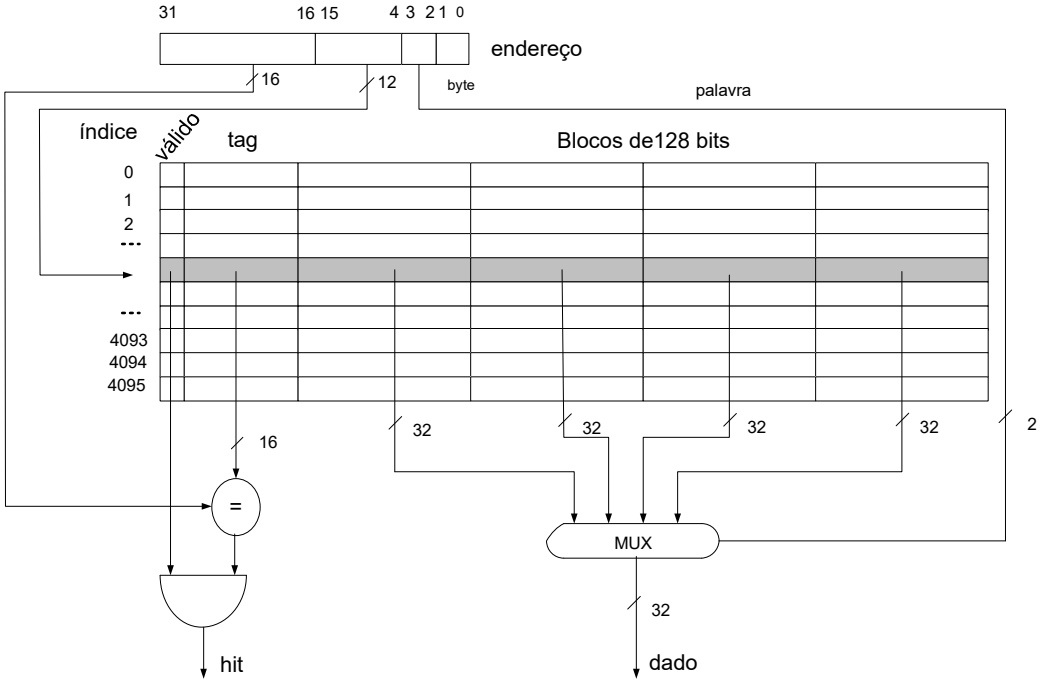

Figura 1. Cache em mapeamento direto.

SOLUÇÃO:
a) qual o tamanho de um bloco em bytes?
Resp: O cache contém 4 palavras por bloco, portanto 16 bytes
b) em que slots as palavras de endereços 0, 5, 17 e 66 são mapeados?
Resp: cada palavra contém 4 bytes portanto, as palavras de endereços 0, 4, 8 e 12 são subsequentes e estão no mesmo bloco. Usando esse raciocínio, as palavras de endereços 0, 5, 17 e 66 estão nos slots de índices: 0, 0, 1, e 4, respectivamente.
c) onde o bloco de número 8192 é mapeado? qual é o tag nesse caso?
Resp: O bloco de número 8192 corresponde ao 8192-ésimo bloco de memória. Como o cache contém slots numerados de 0 a 4095, o 8192-ésimo bloco é mapeado no primeiro slot do cache, ou seja, no slot de índice 0, com tag igual a 2.
d) qual o endereço da primeira palavra do bloco de número 128?
Resp: Lembrando que um bloco contém 4 palavras e cada palavra contém 4 bytes, o número de bloco deve ser multiplicado por 16 para se obter o endereço da primeira palavra do bloco de número 128, portanto, 128 x 16 = 2048.

10) Por que as memórias estáticas são usadas se o custo por Mbyte varia em torno de 100 reais, enquanto que as memórias dinâmicas tem um custo 100 vezes menor?
SOLUÇÃO: Porque apesar do custo das memórias estáticas ser alto, a velocidade é de 5 a 10 vezes maior, o que compatibiliza o seu uso como memória cache.

11) Por que as memórias dinâmicas são lentas e precisam ser reavivadas a um intervalo de alguns milisegundos?
SOLUÇÃO: devido ao armazenamento dos bits em forma de capacitores, cujo efeito de carga e descarga é lento em relação aos transistores usados nos inversores das memórias estáticas. O reavivamento é necessário exatamente porque os capacitores perdem as suas cargas com o tempo, portanto os mesmos devem ser recarregados.

12) O que significa princípio de localidade de referência? Quais tipos de localidade de referência existem?
SOLUÇÃO:
O princípio da localidade de referência refere-se à forma com que os processadores fazem referência à memória durante a execução de programas. A memória é fisicamente um vetor de palavras, porém, o processador não faz referência (endereçamento) a essas palavras de forma aleatória, mas sim de forma localizada, devido a existência de estruturas de dados, armazenamento de programas em endereços contíguos de memória, a execução sequencial de instruções, existência de laços em programas, etc.
Existem dois tipos de localidade de referência: 1) o primeiro tipo é a localidade espacial. A localidade espacial refere-se aos endereçamentos de palavras próximas fisicamente entre uma referência e outra. 2) o segundo tipo é a localidade temporal. A localidade temporal refere-se aos endereçamentos das mesmas palavras referenciadas anteriormente.

13) Como funciona o cache de mapeamento direto?
SOLUÇÃO:
A memória principal é dividida em blocos de tamanho igual ao slot do cache, onde slot é o espaço no cache onde será carregado um bloco. O cache tem um certo número de slots, ordenados sequencialmente. Como a memória principal é maior que o cache, em cada slot do cache deve ser carregado blocos diferentes da memória. O mapeamento direto é o tipo de mapeamento em que a cada slot consecutivo do cache são destinados os blocos consecutivos da memória para carregamento. Quando é atingido o tamanho máximo do cache, os próximos blocos consecutivos da memória são destinados a partir do slot inicial do cache, novamente. Assim, os blocos de endereços múltiplos do tamanho do cache são mapeados no slot inicial do cache, e os blocos subsequentes a esses endereços múltiplos são mapeados nos slots subsequentes do cache.

14) Como funciona o cache de mapeamento associativo por conjunto?
SOLUÇÃO:
O cache de mapeamento associativo por conjunto é diferente em relação ao mapeamento direto, devido a existência de mais do que um slot no cache associado a um bloco de memória. Por exemplo, um cache associativo por conjunto de duas vias, tem dois slots que um determinado bloco de memória pode ser mapeado. Um cache associativo por conjunto de quatro vias, tem quatro slots. O circuito do sistema de mapeamento associativo fica um pouco mais complexo em relação ao mapeamento direto, porque quando é preciso descobrir se um determinado bloco está no cache, é preciso verificar todos os slots correspondentes a aquele bloco, pois o bloco pode estar em qualquer um dos slots.

15) O que significa um mapeamento totalmente associativo?
SOLUÇÃO:
Significa um mapeamento onde todos os slots do cache estão disponíveis para todos os blocos da memória. Assim, quando é preciso verificar se um determinado bloco está no cache, é preciso consultar todos os slots do cache, pois o bloco pode estar em qualquer um dos slots. Portanto é o sistema de mapeamento mais complexo de custo elevado.

16) Para que serve o campo tag do endereço de memória? Como o campo tag é usado para verificar se um dado bloco está no cache?
SOLUÇÃO:
O campo tag do endereço de memória fornecido pelo processador contém os bits que conseguem identificar se o bloco contido num slot do cache é realmente o bloco procurado. Para a verificação se um determinado bloco

de memória está no cache, os bits do campo tag do endereço devem ser comparados com os bits do tag que estão contidos no diretório do cache. O diretório do cache é a parte do circuito do cache onde ficam as informações sobre o conteúdo. Toda vez que o cache recebe um bloco de memória o diretório é atualizado quanto ao slot que recebeu o bloco, com a informação de tag, e com a indicação de slot com bloco válido.

17) O que acontece quando o processador faz uma referência de leitura a uma palavra cujo bloco de memória não é encontrado no cache? O que acontece se no slot do cache onde o bloco deve ser carregado existir um bloco válido?
SOLUÇÃO:
O sistema de memória deve buscar na memória principal o bloco, carregar esse bloco no cache e disponibilizar a palavra que consta dentro do bloco para o processador.
Caso no slot do cache onde o bloco deve ser carregado existir um bloco válido, esse bloco deve ser substituído. Caso esse bloco a ser substituído tivesse sido escrito, e a memória principal não tivesse sido atualizada, a memória principal deve ser atualizada (write-back).

18) O que acontece quando um processador faz uma referência de escrita a uma palavra?
SOLUÇÃO:
Quando o processador faz uma referência de escrita a uma palavra, essa palavra deve ser escrita no cache, no bloco correspondente. Além disso, a palavra deve ser atualizada na memória principal, e nesse caso tem dois critérios: 1) write-through – em que a palavra na memória principal é atualizada a cada vez que é escrita no cache; 2) write-back – em que a palavra na memória principal só é atualizada quando o bloco correspondente no cache deve ser substituído (ver questão 9).

19) O que se entende por intercalação de bancos de memória, ou interleaving?
SOLUÇÃO:
Uma implementação onde existem vários bancos de memória, e as palavras consecutivas estão contidas em bancos diferentes, é denominada de intercalação de bancos. Isso porque para se obter os dados de um determinado bloco referenciado pelo processador todos os bancos de memória devem ser lidos ao mesmo tempo e os resultados de leitura devem ser intercalados para formar o bloco.

20) Por que no sistema de memória virtual, a atualização do disco usa o critério de write-back?
SOLUÇÃO: no sistema de memória virtual, os blocos são maiores que no cache, e tem tamanhos da ordem de alguns Kbytes e são denominados páginas. Se a cada escrita na memória física tivesse que atualizar o disco (write-through) o custo seria muito alto, portanto usa-se o write-back, em que o disco só é atualizado quando uma página da memória física deve ser substituída, e essa página tivesse sido escrita.

21) Para que serve a tabela de páginas num sistema de memória virtual?
SOLUÇÃO:
Num sistema de memória virtual a memória é dividida em páginas, de tamanho da ordem de alguns Kbytes. Como existe um número muito grande de páginas, as mesmas ficam num sistema de armazenamento como o disco magnético. O processador quando faz uma referência a uma palavra, a página que contém essa palavra é carregada do disco para a memória principal, denominada de memória física. Isso porque um acesso à memória principal é mais rápido que um acesso ao disco. A página fica na memória física até que ela precise ser substituída, usando algum critério de substituição. Assim, as páginas mais referenciadas tem uma cópia na memória física. Para que o sistema saiba se uma determinada palavra referenciada tem uma cópia na memória física é usada a tabela de página. Nela contém informações da existência ou não daquela página virtual na memória física e do número do slot da memória física onde fica a cópia, se for o caso.

22) Seja o exemplo do cache em mapeamento direto da Fig. 3:
i) qual o tamanho de um bloco em bytes?
j) onde as palavras de endereços 0, 4, 16 e 64 são mapeados?
k) onde o bloco de número 4096 é mapeado? qual é o tag nesse caso?
l) qual o endereço da primeira palavra do bloco de número 4096?

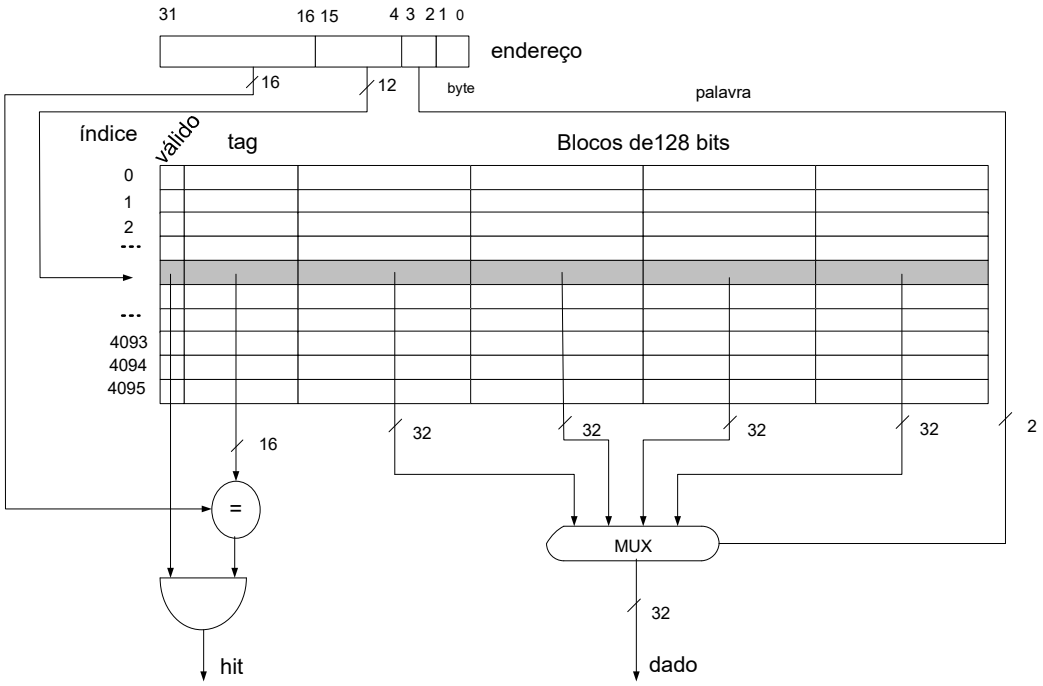

Figura 3. Cache em mapeamento direto.

SOLUÇÃO:

e) qual o tamanho de um bloco em bytes?

Resp: O cache contém 4 palavras por bloco, portanto 16 bytes

f) em que slots as palavras de endereços 0, 4, 16 e 64 são mapeados?

Resp: cada palavra contém 4 bytes portanto, as palavras de endereços 0, 4, 8 e 12 são subsequentes e estão no mesmo bloco. Usando esse raciocínio, as palavras de endereços 0, 4, 16 e 64 estão nos slots de índices: 0, 0, 1, e 4, respectivamente.

g) onde o bloco de número 4096 é mapeado? qual é o tag nesse caso?
Resp: O bloco de número 4096 corresponde ao 4096-ésimo bloco de memória. Como o cache contém slots numerados de 0 a 4095, o 4096-ésimo bloco é mapeado no primeiro slot do cache, ou seja, no slot de índice 0, com tag igual a 1. O tag seria 2 se o bloco fosse 8192-ésimo.

h) qual o endereço da primeira palavra do bloco de número 4096?

Resp: Lembrando que um bloco contém 4 palavras e cada palavra contém 4 bytes, o número de bloco deve ser multiplicado por 16 para se obter o endereço da primeira palavra do bloco de número 4096, portanto, 4096 x 16 = 65536.

23) Considerando-se que é executado um programa num computador MIPS, que lê quatro vezes em seguida, um vetor na memória, de comprimento 20, que se inicia no endereço 16, e que o cache em mapeamento direto da Figura 3 é usado na hierarquia de memória, calcular a quantidade de acertos (hit) e de erros (miss), apenas referente ao acesso às palavras do vetor. Observação: considerar inicialmente, o cache totalmente vazio.

SOLUÇÃO: O vetor contem 20 palavras, em endereços subsequentes a partir de 16. Cada slot do cache contem um bloco de 4 palavras, portanto, são usados 20/4 = 5 slots do cache. a) no primeiro acesso ao vetor, a cada palavra lida, é carregado um bloco, e ocorre um erro. As 3 palavras subsequentes já estão no cache, portanto, 3 acertos. Subtotal = 5 erros e 15 acertos.

b) nos acessos subsequentes, ocorrem 20 acertos. Subtotal = 60 acertos.

c) Total geral = 5 erros e 75 acertos.

24) Ainda, usando o mesmo cache em mapeamento direto da Figura 3, executando um programa no computador MIPS, que lê duas vezes, alternadamente, dois vetores na memória, A e B, de mesmo comprimento, 20, sendo que o vetor A se inicia no endereço 16, e o vetor B, no endereço 16400, calcular a quantidade de acertos e erros, no que se refere ao acesso às palavras dos vetores A e B (Observação: 1) considerar inicialmente o cache totalmente vazio; 2) sequência de leitura dos vetores -> A, B, A, B); 3) o mapeamento do vetor B coincide com o do vetor A, pois o endereço 16400 é igual a 4x4096 +16.

SOLUÇÃO:
a) primeiro acesso ao vetor A: 5 erros e 15 acertos; b) primeiro acesso ao vetor B: 5 erros e 15 acertos; c) segundo acesso ao vetor A: 5 erros e 15 acertos, pois o slot do cache do vetor B é o mesmo para o vetor A; d) segundo acesso ao vetor B: 5 erros e 15 acertos. e) Total geral= 20 erros e 60 acertos.

25) Dado o cache associativo por conjunto da Fig.4:
a) onde se carregam os blocos de números 0, 256, 512 e 1024? quais são os respectivos tags?
b) como poderia melhorar esse cache para explorar a localidade espacial?
c) Explicar como seria a seleção do dado quando um bloco fosse constituído de 2 palavras.

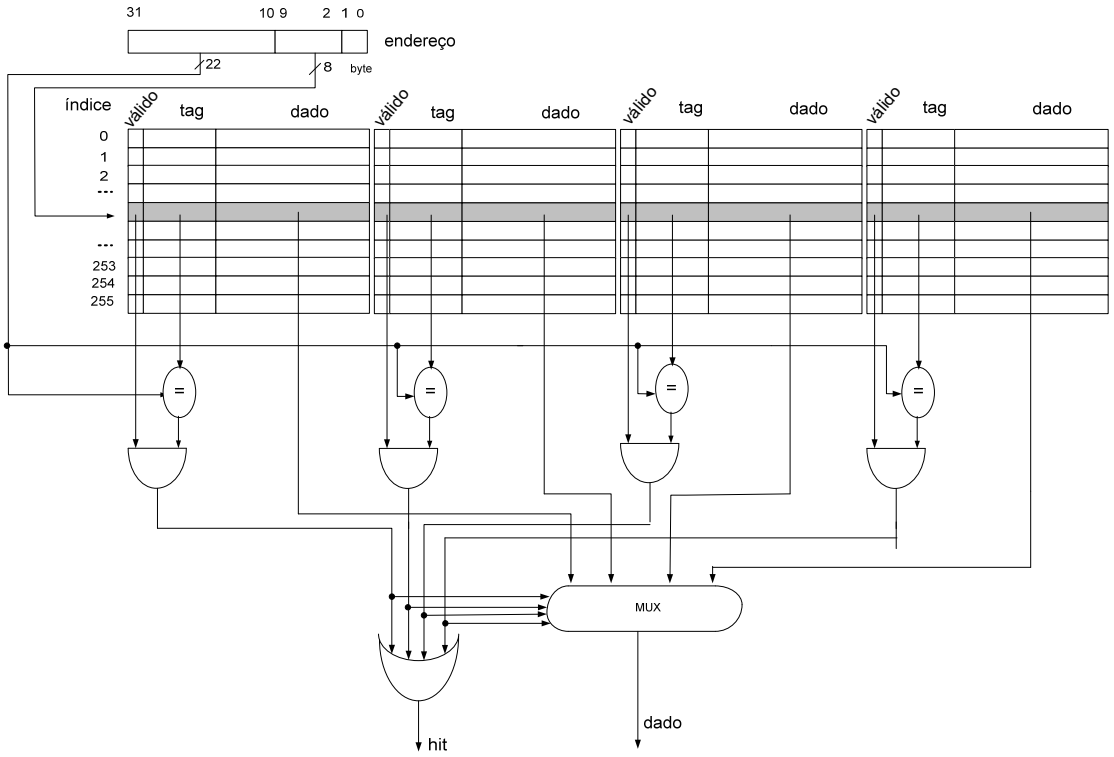

Figura 4. Cache em mapeamento associativo por conjunto.
SOLUÇÃO:
a) onde se carregam os blocos de números 0, 256, 512 e 1024? quais são os respectivos tags?
Resp: esses blocos se carregam nos conjuntos de índice 0. Os respectivos tags são: 0, 1, 2 e 4.
b) como poderia melhorar esse cache para explorar a localidade espacial?
Resp: fazendo com que cada bloco contenha mais do que uma palavra.
c) Explicar como seria a seleção do dado quando um bloco fosse constituído de 2 palavras.
Resp: inserindo 4 multiplexadores 2x1 logo abaixo de cada um dos 4 slots do conjunto, para selecionar a palavra dentro do bloco. A saída desses 4 multiplexadores entram no mux 4 x 1 da Figura 4.

26) Usando o cache em mapeamento associativo por conjunto da Figura 4, executando um programa no computador MIPS, que lê duas vezes, alternadamente, dois vetores na memória, A e B, de mesmo comprimento, 20, sendo que o vetor A se inicia no endereço 16, e o vetor B, no endereço 16400, calcular a quantidade de acertos e erros, no que se refere ao acesso às palavras dos vetores A e B.
SOLUÇÃO:
a) No primeiro acesso a A, ocorrem 20 erros e nenhum acerto.
b) no primeiro acesso a B, ocorrem 20 erros e nenhum acerto.
c) no segundo acesso a A ocorrem 20 acertos e nenhum erro, pois os slots ocupados por B, pertencem aos mesmos conjuntos de A, mas são diferentes.
d) no segundo acesso a B, ocorrem 20 acertos e nenhum erro.
e) Total: 40 erros e 40 acertos.