Quarta Lista de Exercícios – Memória

Problemas 5.4, 5.5 da Segunda Edição em inglês

5.4) Você comprou um computador ACME com as seguintes características: (i) 95% de todas as referências são encontradas na cache, (ii) blocos da cache tem duas palavras, e o bloco completo é lido numa falta, (iii) a CPU envia referências à cache na taxa de 10^9 palavras/segundo, RDs ou WRs, (iv) 25% de todas as refs são WRs, (v) o sistema de memória pode suportar 10^9 pals/segundo, RDs ou WRs, (vi) o barramento transfere uma palavra a cada vez (o sist de memória não lê ou escreve duas palavras de uma vez), (vii) suponha que a qualquer tempo, 30% dos blocos na cache tenham sido modificados, (viii) a cache usa *write-allocate* numa falta na escrita. Você deseja adicionar um periférico ao sistema, e deseja saber quanto da largura de banda do sistema de memória já está sendo utilizado. Calcule a fração da banda do sist de memória que está sendo utilizada se: (a) a cache é com escrita forçada; (b) a cache é com escrita preguiçosa. Explicite suas assunções.

5.5) Uma diferença entre caches com escrita forçada e escrita preguiçosa é o tempo necessário para escrever. Durante o primeiro ciclo, detecta-se se um acerto vai ocorrer, e durante o segundo ciclo (supondo acerto) o dado é gravado na cache. Suponha que 50% dos blocos são sujos numa cache com escrita preguiçosa, e que a fila de escrita nunca bloqueia o processador (a fila não fica cheia). Um acerto na leitura custa 1 ciclo, a penalidade por falta é 50 ciclos, e a escrita de um bloco da cache na memória custa 50 ciclos. Finalmente, suponha que a taxa de faltas na cache de instruções é 0.5%, e a taxa de faltas na cache de dados é 1%.
(a) Usando as percentagens de LDs e STs listadas abaixo, estime o desempenho de uma cache com escrita forçada com escrita em 2 ciclos *versus* uma cache com escrita preguiçosa também de 2 ciclos para cada um dos programas listados.
(b) Repita a comparação, mas agora suponha que a cache com escrita forçada tem um *pipeline* no circuito de escrita, de forma a que uma escrita leva somente um ciclo.
P1: 5% WRs, 20% RDs, P2: 10% WRs, 15% RDs, P3: 15% WRs, 20% RDs.

1) Quais parâmetros de projeto de memórias cache são associados à localidade temporal? Quais parâmetros são associados à localidade espacial? Justifique sua resposta nos dois casos.

2) Faça um diagrama detalhado de uma memória cache com 1 Mbytes, associatividade quaternária (*4-way set-associative*), 8 palavras por bloco, escrita preguiçosa. O processador emite endereços de 32 bits. Indique como um endereço é interpretado pelo controlador da cache (i.e especifique a função de *hashing* da cache).

3) Faça um projeto detalhado de uma fila de escrita com capacidade para quatro referências pendentes. Considere duas possibilidades para a largura da escrita: (a) uma palavra (i.e. par $\langle$endereço, valor$\rangle$, e (b) um bloco da cache (i.e. tupla $\langle$endereço, valor[TAM_BLOCO]$\rangle$. Discuta a relação custo-desempenho das duas implementações. Qual organização é a melhor?

4) Considere um processador que executa uma instrução por ciclo (CPI=1.0) se ligado a um sistema de memória ideal e perfeito. O relógio do processador é 1 GHz. A hierarquia de memória possui uma cache, com tempo de acerto de 1 ciclo (1ns). Cada bloco da cache tem capacidade para 8 palavras; a escrita é forçada e o barramento entre cache e memória têm duas palavras de largura. A busca de instruções nunca causa faltas. Calcule o CPI deste processador quando ligado ao sistema de memória da questão acima se a taxa de acertos na cache de dados é de 95% quando o processador executa código com um padrão de acessos com 25% de escritas e 75% de leituras e 40% das instruções são referências à dados.

5) Faça um diagrama detalhado de uma cache de mapeamento de endereços (*translation lookaside buffer* ou TLB) com 1024 blocos e associatividade quaternária. Cada bloco contém um mapeamento. O processador emite endereços de 32 bits. O endereço físico possui 40 bits, e páginas virtuais tem 8 Kbytes. (a) Indique como um endereço é interpretado pelo controlador da TLB. (b) Qual o tamanho da Tabela de Páginas?

6) Considere o programa de multiplicação de matrizes abaixo. Suponha que as matrizes contém 1024x1024 elementos, cada elemento um double (8 bytes). O programa é executado num único processador com páginas de 4 Kbytes, e cache secundária de 2 MBytes.

(a) Descreva o comportamento do sistema de memória virtual durante a execução deste programa; (b) sugira uma ou mais maneiras de melhorar o desempenho da multiplicação de matrizes, envolvendo somente paginação; (c) sugira uma ou mais maneiras de melhorar o desempenho da multiplicação de matrizes, envolvendo somente caches.

```
/* a,b,c:  double×double */
1  for (i=0; i < 1024; i++) {
2    for (j=0; j < 1024; j++) {
3      for (sum=0.0, k=0; k < 1024; k++)
4        sum += a[i][k] * b[k][j];
5      c[i][j] = sum;
6    }
7  }
```

7) Usando a notação de Teoria dos Conjuntos, dê expressões que descrevem as relações entre os conteúdos dos registradores ($0 a $31), duas caches primárias (dados e instruções), cache secundária unificada, RAM, e área de swapping (suporte a memória virtual). Não esqueça das conseqüências de escritas em todos os níveis. Para as caches, escolha (e indique) o mecanismo de escrita que mais simplifica as expressões. Considere apenas o processo que está executando no processador e ignore os demais. As expressões devem descrever as circunstâncias nas quais as várias combinações de faltas e acertos podem ocorrer.

8) Uma CPU superescalar emite 4 instruções por ciclo, e tem relógio de 2 GHz (0.5ns/ciclo). Uma falta na L1 custa, no mínimo, 10 ciclos. Quais os números mínimo e máximo de instruções que poderiam ser executadas durante uma falta na L1? Uma falta na L2 custa, no mínimo 110 ciclos; quais os números máximo e mínimo? Justifique nos dois casos.

9) Considere as políticas de escrita preguiçosa e forçada, e alocação, não-alocação de um bloco numa falta na escrita. (a) Especifique cuidadosamente as quatro combinações possíveis (EP+aloc, EP+nãoAloc, EF+aloc, EF+nãoAloc). (b) Qual destas é a melhor para memória virtual? (c) Justifique.

10) Como são as localidades (temporal e espacial) nos acessos gerados da cache L1 para a L2? E da L2 para a memória? Posto de outra forma, descreva os padrões de acesso entre L1 e L2, e entre L2 e memória. Como isso influencia no projeto dos circuitos de memória nos três níveis? Como isso influencia na escolha dos CIs de DRAM?

11) Escreva, em pseudocódigo, uma função com o protótipo abaixo que percorre uma tabela de páginas de três níveis e retorna **1** se a página está em memória, ou **0** numa falta. O endereço físico é atribuído à `*enderfis` num acerto. Explicite quaisquer suposições que forem necessárias.

```
int buscatp( void *basetp, void* endervirt, void** enderfis );
```

12) Projete o algoritmo do controlador de uma cache primária com escrita preguiçosa (*write-back*). Os blocos são de 8 palavras e a cache entrega ao processador a palavra crítica primeiro. O algoritmo deve controlar as duas interfaces da cache: L1-CPU e L1-L2. Especifique as entradas, saídas e estruturas de dados.

13) *Sinônimos* ocorrem em sistemas de memória virtual quando dois endereços virtuais distintos mapeiam no mesmo endereço físico.

EVirtual	# pág virt	desloc
ender na cache	|–índice–|	
EFísico	# pág fís	desloc

Isso pode ocorrer quando dois processos compartilham uma área de memória. Uma maneira de evitar sinônimos, ou ao menos facilitar sua detecção, é garantir que a indexação da cache ocorra com bits que coincidem no EV e no EF, como mostrado no diagrama. Justifique esta solução.