Uma Avaliação de Estruturas de Memória para Multiprocessadores

Raul Queiroz Feitosa

Resumo

Este trabalho compara diferentes organizações de memória para multiprocessadores. A contribuição de memórias locais e memórias cache privadas para aumento de desempenho é considerada. O estudo abrange sistemas com diferentes estruturas de interconexão: barramentos compartilhados, redes "crossbar" e redes de múltiplos estágios. O modelo porposto por Patel é extendido e aplicado a diferentes estruturas de memória. A análise indica que o miss ratio da cache é, entre os parâmetros do modelo, aquele que mais influencia o desempenho relativo das estruturas de memória consideradas. Os resultados indicam que a organização com cache e memória local é a que apresenta melhor desempenho, que é, no entanto, próximo ao apresentado pela organização que contém apenas a cache.

Abstract

This study compares different memory organizations for multiprocessor systems. The influence of local and/or private cache memories on processor activities is considered. Different interconnection structures are used: busses, mono- and multistage interconnection networks. Patel's model is extended to represent different memory organizations. The analysis indicates that the model parameter which most influences the results is the cache miss ratio. For realistic assumptions, the organization with cache and local memory is the best, but with little distance to the model with cache only.

Raul Queiroz Feitosa *formou-se em Engenharia Eletrônica em 1979 no Instituto Tecnológico de Aeronáutica, São José dos Campos; recebeu o título de Mestre em Engenharia em 1983 no mesmo Instituto. Em 1988 lhe foi outorgado o título de "Doktor der Ingenieurwissenschaften" pela Universidade de Erlangen-Nürnberg, República Federal da Alemanha. Suas áreas de interesse incluem arquitetura de computadores e sistemas paralelos, com ênfase em estruturas de memória para máquinas LISP, PROLOG e multiprocessadores. Trabalha atualmente como Professor Assitente na Pontifícia Universidade Católica do Rio de Janeiro - Departamento de Engenharia Elétrica - Rua Marquês de São Vicente, 225 - CEP: 22453 - Rio de Janeiro - RJ; Tel.: (021) 5299505; Correio Eletrônico - USERRQFT@LNCC.BITNET*

1. Introdução

Apesar de estruturas de memória complexas (hierarquia, interleaving, etc.), para determinadas aplicações os sistemas monoprocessados têm seu desempenho limitado muito mais pela velocidade da memória do que pela velocidade do processador. Particularmente importante é a escolha adequada de uma estrutura de memória para multiprocessadores. Nestes sistemas a memória deve assumir uma carga adicional que advém da comunicação entre os elementos que operam em paralelo.

Analisando-se possíveis organizações dos elementos de memória num multiprocessador, pode-se encontrar diversas alternativas que se distinguem, por exemplo, nos seguintes aspectos:

- composição dos elementos de memória: utilização ou não de memórias cache ou memórias locais, estrutura e estratégias de gerenciamento da memória cache [SMIT82]
- rede de interconexão: barramento compartilhado, redes de múltiplos estágios, redes "crossbar", conexão completa, ligação tipo "nearest neighbor", etc. [BOHÃ82]

As exigências que a estrutura de memória escolhida deverá satisfazer dependem da estratégie de paralelização empregada e de parâmetros específicos da aplicação.

O presente trabalho apresenta uma análise de diferentes estruturas de memória para um multiprocessador com uma memória global compartilhada, constituída de diversos módulos independentes. Sobre esta base são analisadas diversas formas de interconexão e diversas hipóteses quanto às características dos programas executados ("workload").

O modelo empregado é o modelo proposto por Patel [PATE82]. Através de uma atribuição adequada de valores a alguns dos parâmetros do modelo, representa-se estruturas de memória diferentes daquela a que o modelo original está orientado.

O artigo apresenta inicialmente as estruturas de memórias analisadas. Em seguida o modelo de Patel é descrito de forma sucinta. As modificações do modelo necessárias para representar as estruturas de memória estudadas são introduzidas a seguir. Finalmente são apresentados e discutidos os resultados obtidos pela aplicação do modelo.

2. As Estruturas de Memória Analisadas

O multiprocessador utilizado como base para este estudo, é representado esquematicamente na figura 1. N módulos de processamento idênticos (PM_i), se comunicam entre si por uma memória compartilhada composta de M módulos de memória idênticos (MM_j). Os acessos dos módulos de processamento aos módulos de memória são executadas através de uma rede de interconexão.

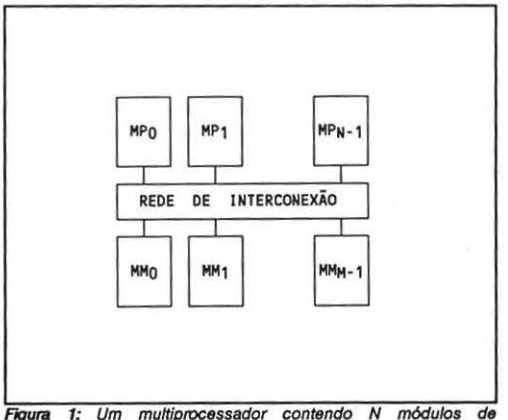

Figura 1: Um multiprocessador contendo N módulos de processamento (PM_i) que M módulos de memória (MM_j) através de uma rede de interconexão.

As cinco estruturas de memória com as quais este trabalho se ocupa são mostradas na figura 2. Cada estrutura é representada apenas por um módulo de processamento que, conforme a figura 1, está presente no sistema N vezes.

As estruturas de memória diferenciam-se umas das outras apenas na arquitetura da memória privada associada a cada módulo de processamento.

Na estrutura **a** da figura 2 o módulo de processamento não dispõe nem de uma memória local nem de um memória cache privada.

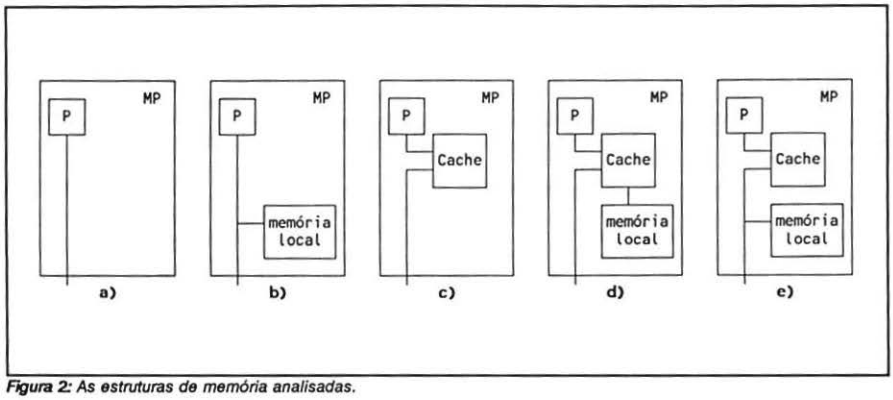

Figura 2: As estruturas de memória analisadas.

Na estrutura **b** o módulo de processamento possui uma memória local, na qual são armazenados o código executado pelo respectivo processador e parte dos dados por ele acessados. Esta análise considera dois casos distintos para a estrutura **b**. No primeiro caso admite-se que a memória local e a memória compartilhada são construídas com base na mesma tecnologia e que possuem portanto o mesmo tempo de acesso, a menos dos atrasos introduzidos pela rede de interconexão e causados pela contenção. O segundo caso prevê uma memória local rápida (acessos sem ciclos de espera).

A estrutura **c** prevê uma memória cache privada, que pode conter cópias dos blocos da memória compartilhada.

Já na estrutura **d** o processador de cada módulo de processamento dispõe de uma cache e de uma memória local. Ao contrário da estrutura **c**, nesta estrutura a cache não contém cópias dos blocos da memória compartilhada, mas apenas dos blocos da memória local.

Na estrutura **e** a cache pode conter cópias tanto de blocos da memória compartilhada como da memória local.

Na seção 6 estas estruturas de memória são analisadas exclusivamente sob o ponto de vista do desempenho. Cada uma delas apresenta, no entanto, vantagens e desvantagens que não podem ser expressas exclusivamente através do desempenho associado. Todas as estruturas com uma memória local (**b**, **d** e **e**) apresentam a desvantagem de que o programador, de alguma forma, deve determinar a distribuição dos códigos e dados entre memória local e compartilhada. Isto se constitui numa fonte potencial de erros de software. Nas estruturas **c** e **e** as caches podem conter cópias de blocos da memória compartilhada. Em tais estruturas surge o problema de coerência ou consistência de memórias cache. Várias caches do multiprocessador podem conter uma cópia de um mesmo bloco da memória compartilhada. Pode ocorrer que um processador modifique uma destas cópias na sua cache privada sem que as outras cópias do sistemas sejam correspondentemente alteradas. Um outro processador pode fazer um acesso ao mesmo bloco e receber uma versão não atual daquela informação. Existem inúmeros mecanismos que garantem a coerência em tais sistemas. Cite-se aqui como exemplos [ARCH84], [ARCH85], [BITA86], [CENS78], [DUBO82], [KATZ85], [VERN86] e [FEIT90].

A estrutura **b** com uma memória local rápida pode ser vista como uma memória cache sem os seus mecanismos de gerenciamento (estratégia de substituição, de atualização da memória principal, mapeamento de endereços, "prefetch", etc...). Comparada com uma cache convencional, a construção de uma memória local rápida é mais simples. São, no entanto, estes mecanismos de gerenciamento da cache os responsáveis pelo elevado hit ratio (veja definição a seguir) que se obtém mesmo com memórias cache de pequena capacidade. Em antecipação à análise que se segue, saliente-se aqui, que, de uma forma geral, a estrutura **b** com uma memória local rápida pode apresentar um desempenho semelhante ao da estrutura **c**, se tiver uma capacidade maior e se for consequentemente mais cara do que a cache da estrutura **c**.

3. O Modelo de Patel

Antes de apresentar o modelo proposto por Patel, faz-se necessário definir alguns conceitos:

- *Miss* é o termo utilizado para caracterizar a ocorrência de um acesso de um processador a um bloco, do qual sua cache privada não contém uma cópia.
- *Miss ratio* é a relação entre o número de misses e o número total de acessos gerados por um processador ao executar um determinado programa.
- *Hit* é o termo utilizado para caracterizar a ocorrência de um acesso de um processador a um bloco, do qual sua cache privada possui uma cópia.
- *Hit ratio* é a relação entre o número de hits e o número total de acessos gerados por um processador ao executar um determinado programa. É claro que o hit ratio é igual a 1 menos o miss ratio.
- *Bloco* é todo conjunto de 2^k (k é um número inteiro) bytes consecutivos da memória principal, cujos endereços podem ser diferentes apenas nos k bits menos significativos. A troca de informações entre cache e memória principal se faz bloco a bloco.
- *Linha* é a unidade da cache onde são armazenadas cópias de blocos da memória principal. Uma linha pode conter a cada instante a cópia de um único bloco.
- *Tamanho de linha* ou *tamanho de bloco* caracteriza o número de bytes que uma linha ou um bloco contém. Tamanho de bloco e tamanho de linha são utilizados neste trabalho como sinônimos.

O Modelo

Vários modelos analíticos foram propostos que procuram prever a redução de desempenho provocada em multiprocessadores devido à contenção (p. ex. [RAVI72], [BHAN75], [YEN-82], [SMIL85]). Para esta análise escolheu-se o modelo proposto em [PATE82], por ser particularmente orientado a sistemas com memórias cache privadas, cujos efeitos sobre o desempenho do sistema é um dos aspectos que se deseja analisar neste trabalho. São introduzidas pequenas alterações no modelo original, de modo a representar outras estruturas de memória distintas daquela na qual o modelo original se baseia.

A figura 3 mostra o sistema representado pelo modelo de Patel. Ele corresponde ao sistema da figura 1, onde os módulos de processamento têm a estrutura **c** da figura 2. Tem-se N módulos de processamento (MP_i), constituídos cada um de um processador (P_i) e uma memória cache privada (C_i), e M módulos de memória (MM_i) compartilhados pelos módulos de processamento. Admite-se que os acessos dos processadores estão uniformemente distribuídos entre todos os módulos de memória.

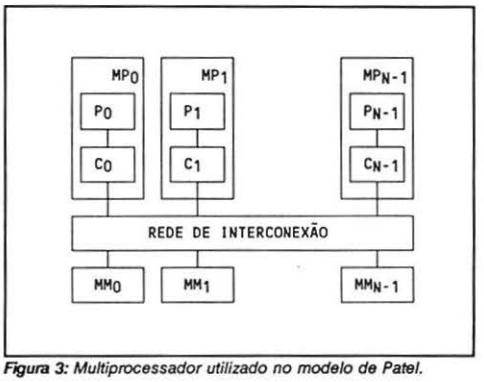

Figura 3: Multiprocessador utilizado no modelo de Patel.

Para simplificação da análise, admite-se que algum mecanismo específico resolve o problema de coerência no sistema, e que sua execução tem pouca influência sobre o desempenho global.

Com relação ao problema da atualização da memória principal admite-se que as caches empregam a estratégia "buffered write back"-BWB, ou seja, nos acessos de escrita o dado referenciado é alterado somente na cache, caso ela contenha dele uma cópia, e a memória principal só é atualizada quando este bloco é substituído na cache. Se a cache não contém uma cópia do bloco a ser alterado, este é carregado da memória principal e, em seguida é executada a escrita. Na ocorrência de um miss, o bloco a ser substituído na cache é escrito num "buffer" intermediário. O bloco solicitado é então trazido

da memória principal e o bloco substituído é em seguida transferido do "buffer" para a memória principal. Neste modelo admite-se que o "buffer" pode gerenciar vários misses e que possui um número infinito de posições.

A memória cache é modelada da seguinte forma. Num miss a cache transfere o bloco substituído para o "buffer" em t_b unidades de tempo. Em seguida é gerada uma solicitação de leitura que, devido à contenção, somente é atendida depois de w unidades de tempo. A leitura do novo bloco é executada em t unidades de tempo. O bloco no "buffer" é escrito então na memória principal. A cache gera uma solicitação de escrita que é atendida depois de w unidades de tempo. A escrita requer, da mesma forma que a leitura, t unidades de tempo. Durante este tempo o processador não fica bloqueado, a menos que ocorra um novo miss antes que o processo de escrita tenha sido completado. Neste caso, a cache espera até que a escrita se complete e envia uma nova solicitação de leitura. O valor médio do atraso assim gerado é igual a $\dot{w}$ unidades de tempo.

O comportamento do processador é modelado como segue. Distingue-se entre ciclo ativo e ciclo passivo do processador. Passivos são os ciclos de espera introduzidos em decorrência da baixa velocidade da memória ou de contenção. Ativos são todos os demais ciclos que não caem na classe anterior. A duração de um ciclo, seja ele ativo ou passivo, é igual a uma unidade de tempo. Quando são introduzidos ciclos passivos, depois de uma sequência de ciclos ativos, o processador suspende o processamento e não executa qualquer ação produtiva. No próximo ciclo ativo o processador prossegue com o processamento a partir do ponto onde este foi suspenso. Admite-se que o número de ciclos ativos necessários para a execução de uma tarefa independe do número de ciclos passivos introduzidos ao longo do processamento.

Define-se *Utilização do Processador* (U) como a relação entre o número de ciclos ativos e o número total de ciclos do processador durante a execução de uma determinada tarefa. Se a cada k ciclos ativos do processador são introduzidos em média $\bar{k}$ ciclos passivos, a Utilização do Processador é igual a:

$$U = \frac{k}{k + \bar{k}} \qquad (1)$$

Admite-se no modelo que a cache é capaz de atender os acessos do processador sem ciclos de espera. Isto implica em que a Utilização do Processador no sistema esquematizado na figura 3 será igual a 1, se o miss ratio for igual a zero.

Seja m a probabilidade de que o processador gere um miss num ciclo ativo. m é portanto igual à probabilidade de que o processador gere um acesso num ciclo ativo e que a informação referenciada não esteja disponível na cache.

Seja Q a probabilidade de que o processador gere um acesso num ciclo ativo, e seja f o miss ratio das caches. Admite-se que o valor de f é uma boa aproximação para a probabilidade de ocorrência de um miss. A seguinte equação é portanto válida:

$$m = Q \cdot f \qquad (2)$$

Como a cada ciclo ativo é gerado um miss com probabilidade m, tem-se em média mk misses a cada k ciclos de processamento produtivo. Os mk misses adicionam m k ($t_b + \dot{w} + w + t$) ciclos passivos aos k ciclos ativos do processador. Tendo em vista a equação (1), conclui-se que a Utilização do Processador é igual a:

$$U = \frac{k}{k + m\,k\,(t_b + \dot{w} + w + t)} = \frac{1}{1 + m\,(t_b + \dot{w} + w + t)} \qquad (3)$$

Os k ciclos ativos geram mk acessos de escrita (atualização do bloco substituído na cache) e mk acessos de leitura (carga do bloco solicitado na cache). Durante as $k + m\,k\,(t_b + \dot{w} + w + t)$ unidades de tempo, os módulos de memória do multiprocessador com N módulos de processamento atendem portanto um total de 2 N m k acessos. Cada acesso ocupa um módulo de memória durante t unidades de tempo. O número médio de módulos de memória ocupados é igual a:

$$B = \frac{2 N m t}{1 + m (t_b + \dot{w} + w + t)} = 2 N m t U. \qquad (4)$$

O atraso w pode ser calculado da seguinte forma. Admite-se que os acessos de leitura têm prioridade mais alta do que os acessos de escrita. A escrita pode ser bloqueada durante os w ciclos de espera, mas, assim que a ligação através da rede de interconexão se estabeler, a escrita não poderá mais ser interrompida até se completar. Supondo-se que um acesso de escrita está em andamento e ocorre um miss logo após a primeira unidade de tempo, o acesso de leitura deverá aguardar (t - 1) unidades de tempo. Isto ocorre com uma probabilidade igual a m. Quando a solicitação de leitura ocorre, não após a primeira, mas após a segunda unidade de tempo, a leitura esperará (t - 2) unidades de tempo. A probabilidade deste evento é igual a (1- m) m. Continuando com este raciocínio conclui-se que os acessos de escrita contribuem com

$$m (t - 1) + m (1 - m)(t - 2) + \ldots + m (1 - m)^{i-1}(t - i) + \ldots + m (1 - m)^{t-2}$$

unidades de tempo para o atraso total. Obtém-se assim, o seguinte valor para o atraso $\dot{w}$:

$$\dot{w} = m \sum_{i=1}^{t} (1 - m)^{i-1}(t - i) \qquad (5)$$

Patel constatou através de simulação que a Utilização do Processador (U) pode ser aproximada por uma função do produto (m t). A Utilização depende na realidade não apenas de (m t) mas também de m e t individualmente. O efeito de m e t, é, no entanto, pequeno quando comparado com o efeito do produto (mt). Em outras palavras, a Utilização do Processador (U) depende preponderantemente da intensidade do tráfego e pouco da natureza deste tráfego.

Patel substituiu o modelo descrito até aqui por um outro modelo, onde a intensidade de tráfego é a mesma, mas que permite um tratamento analítico mais simples.

No modelo original o caminho entre os módulos de processamento e de memória é mantido por t unidades de tempo enquanto o acesso é executado. Esta ocorrência pode ser vista como t solicitações consecutivas ao mesmo módulo de memória, sendo que cada solicitação consome uma unidade de tempo de serviço.

Do ponto de vista do módulo de processamento, o conjunto formado pela rede de interconexão e os módulos de memória requer, (w + t) unidades de tempo para cada acesso, independentemente de se tratar de um acesso de leitura ou de escrita.

A cada miss a rede de interconexão recebe (w + t) solicitações consecutivas de uma unidade de tempo de serviço para a atualização na memória principal do bloco substituído na cache, e mais (w + t) solicitações para a carga do bloco referenciado na cache. A taxa com que solicitações de unidades de tempo de serviço são enviadas à rede de interconexão é portanto igual a:

$$\dot{m} = \frac{2 m (w + t)}{1 + m (t_b + \dot{w} + w + t)} = 2 m (w + t) U. \qquad (6)$$

A aproximação introduzida por Patel é basicamente que as (w + t) solicitações consecutivas a um mesmo módulo de memória são substituídas por (w + t) solicitações separadas, independentes, e aleatória e uniformemente distribuídas entre todos os módulos de memória.

No passo seguinte considera-se a estrutura de interconexão utilizada. São considerados a seguir dois casos: uma rede de interconexão tipo "crossbar" e uma rede de múltiplos estágios.

Rede "crossbar"

Numa rede "crossbar" cada módulo de memória é referenciado por um processador num determinado instante de tempo com uma probabilidade igual a $\dot{m}/M$. A probabilidade de que nenhuma cache do sistema solicite serviço de um módulo de memória particular é $(1 - \dot{m}/M)^N$. Conclui-se que a probabilidade de que, pelo menos uma cache esteja enviando uma solicitação a um determinado módulo

de memória, é igual a $1 - (1 - m'/M)^N$, valendo portanto a equação:

$$B = M\,[1 - (1 - m'/M)^N] \qquad (7)$$

Substituindo as equações (3) e (4) em (7) obtém-se:

$$2\,N\,m\,t\,U - M\,[1 - (1 - m'/M)^N] = 0 \qquad (8)$$

Este sistema de equações pode ser resolvido através de métodos iterativos convencionais. Um bom valor inicial pode ser obtido da equação (3) fazendo w igual a 0, ou seja,

$$U = \frac{1}{1 + m\,(t_b + w' + t)} \qquad (9)$$

que corresponde à máxima Utilização do Processador.

Rede de Interconexão Delta

Patel incluiu em seu trabalho também um modelo para a rede de interconexão tipo Delta. Trata-se de uma rede de múltiplos estágios, construída a partir de redes "crossbar" elementares que possuem a entradas e b saídas. Uma descrição detalhada deste tipo de rede de interconexão é apresentada em [PATE81] e [SIEG85]. Ao contrário da rede "crossbar", pode ocorrer contenção na rede de interconexão Delta, mesmo que os acessos sejam dirigidos a módulos de memória distintos.

A modificação a ser introduzida para incluir este tipo de rede na análise consiste em calcular a equação (7) recursivamente para cada estágio da rede Delta. Obtém-se assim:

$$B = M\,m_n, \qquad (10)$$

onde

$$m_{i+1} = 1 - (1 - m_i/b)^a \qquad 0 \le i < n \qquad (11)$$

e

$$m_o = m'. \qquad (12)$$

Na análise de redes de interconexão de múltiplos estágios admite-se que a = b e N = M.

Os resultados obtidos a partir deste modelo foram comparados com resultados de simulações onde os parâmetros m, t, M e N foram variados dentro de uma ampla faixa de valores. Para a Utilização do Processador a diferença absoluta entre os resultados obtidos analiticamente e por simulação permanecem abaixo de 0,02.

Algumas Observações quanto ao Modelo de Patel

O modelo apresentado despreza os efeitos do mecanismo de coerência sobre o desempenho do sistema. Todos os mecanismos de coerência conhecidos afetam negativamente o desempenho. As causas são basicamente três:

1. elevação do miss ratio [DUBO87],
2. tráfego adicional sobre a rede de interconexão, causado por eventuais trocas de comandos entre caches e memória principal, e

3. eventuais interrupções da operação normal da cache.

Cada mecanismo de coerência é afetado diferentemente por cada um destes fatores. Há mecanismos em que nem todos estes fatores estão presentes. Na seção 6 o fator 1 é considerado, ao variar-se o miss ratio. Representar sistemas onde os efeitos 2 e 3 são significativos, exigiria modificações não triviais no modelo original que seriam, em muitos casos, específicas para um determinado mecanismo de coerência ou para um conjunto de mecanismos. Tal análise excede o escopo deste trabalho.

Outra observação quanto ao modelo de Patel diz respeito ao "buffer" infinito. No modelo atribui-se sempre maior prioridade aos acessos de leitura do que aos de escrita. Nos casos reais, quando o "buffer" finito se enche, a escrita passa a ter maior prioridade, de modo a criar espaço para novos blocos.

O modelo de Patel permite portanto que, em situações de carga elevada, o número de blocos armazenados no "buffer" cresça indefinidamente. Nestes casos, os $mk[1 + m(t_b + w' + w + t)]$ ciclos são suficientes para que a cache execute os mk acessos de leitura à memória principal, mas a cache não consegue executar todos os mk acessos de escrita, provocando o enchimento do "buffer". Esta situação ocorre quando:

$$2m(w + t) > 1 + m(t_b + w' + w + t), \text{ ou } m' > 1$$

Uma situação tão distante da prática como esta é excluída das considerações seguintes. Ao se aplicar o modelo, verifica-se, se $m' < 1$. Em caso afirmativo o resultado produzido é aceito, caso contrário desprezado.

Patel admite ainda em seu modelo que num miss a cache só iniciará o acesso, que carregará o bloco solicitado, após ter escrito o bloco substituído no "buffer". Uma alternativa a este procedimento é iniciar a escrita no "buffer" e a carga do bloco solicitado pelo processador simultaneamente de modo que ambos os acessos ocorram paralelamente. É razoável supor que o tempo de acesso ao "buffer" é próximo do tempo de acesso à cache e portanto muito menor do que o tempo de acesso à memória principal. Assim, quando o bloco da memória principal estivesse disponível, o bloco substituído já teria sido transferido para o "buffer". Desta forma, o tempo de acesso ao "buffer" não produziria qualquer atraso adicional, o que, para efeito do modelo, corresponde a igualar o parâmetro t_b a zero. Esta alternativa, além de ser mais atraente do ponto de vista de desempenho, não apresenta dificuldades significativas de implementação. Por esta razão admite-se daqui em diante que o parâmetro t_b é igual a zero.

Ainda associado ao "buffer" existe um outro fator a ser considerado. Ao obter a equação (6), Patel admite que para cada acesso de leitura e para cada acesso de escrita a cache produzirá em média $(w + t)$ solicitações ao conjunto composto pela rede de interconexão e módulos de memórias. Ao tratar os acessos de leitura e escrita igualmente, Patel está na realidade introduzindo uma simplificação no modelo. Se ao tentar escrever na memória principal ocorrer um miss, antes da cache ganhar o acesso ao módulo de memória, a escrita cede lugar à leitura. Completada a leitura, a cache reinicia suas tentativas de estabelecer o acesso ao módulo de memória correspondente para executar a escrita pendente. Isto pode se repetir várias vezes até que a escrita se realize efetivamente. Conclui-se que, a rigor, o tempo médio que a cache fica tentando ganhar o acesso ao módulo de memória para a escrita é maior do que para a leitura. Em [FEIT88] são apresentadas as modificações a serem introduzidas no modelo de modo a considerar este efeito. Naquele trabalho constatou-se, contudo, que, na faixa realística de valores dos parâmetros do modelo (vide discução na seção 5), a influência destas alterações sobre os resultados é muito pequena, e que a aproximação de Patel é portanto bastante boa. Por estas razões e em benefício da simplicidade, optou-se aqui por manter a simplificação de Patel quanto aos acessos de escrita.

4. Modelamento de outras Estruturas de Memória

O modelo de Patel refere-se apenas à estrutura **c** da figura 2. As outras estruturas de memória mostradas na figura 2 podem ser representadas através de uma definição adequada de alguns dos parâmetros do modelo original, conforme será mostrado a seguir.

Devido à memória local, presente em algumas das estruturas mostradas na figura 2, introduz-se aqui a definição de alguns novos parâmetros.

Seja t_l e t_c respectivamente os tempos de acesso à memória local e à memória compartilhada, representada pelos módulos de memória. Seja z_l e z_c respectivamente a probabilidade de que um acesso gerado pelo processador referencie a memória local ou a memória compartilhada. Evidentemente $z_l = 1 - z_c$.

A estratégia "Buffered Flagged Write Back - BFWB" apresenta algumas vantagens sobre a estratégia BWB, considerada no modelo original de Patel. Na estratégia BFWB nem todos os misses causam a atualização da memória principal. Somente se o bloco substituído foi alterado desde a última vez em que foi trazido para a cache, ocorre a atualização da memória principal. Para considerar esta estratégia no modelo, deve-se introduzir as seguintes modificações nas equações (3), (4) e (5). Seja σ a probabilidade de que o bloco substituído na cache foi modificado depois da última vez em que foi carregado na cache.

Dois aspectos devem ser considerados aqui. Primeiramente o atraso $(w' + t_b)$ não afeta mais todos os misses, mas apenas aqueles nos quais ocorre a atualização da memória principal. Em segundo lugar, o número de acessos à memória principal não é mais o dobro mas apenas $(1 + \sigma)$ vezes o número de misses. A modificações necessárias nas equações (3) a (5) resultam nas equações (13) a (16), que modelam a estrutura **c** da figura 2.

Estrutura c:

As seguintes equações modelam a estrutura **c**:

$$U_c = \frac{1}{1 + Q\ f\ [\sigma\ w_c' + w + t_c]} \tag{13}$$

$$m_c' = (1 + \sigma)\ Q\ f\ (w + t_c)\ Uc \tag{14}$$

$$B_c = N\ Q\ f\ (1 + \sigma)\ t_c\ Uc \tag{15}$$

O atraso médio no "buffer" da cache pode ser calculado pela equação (5), onde m é subsituído por Q f e t por t_c:

$$w_c' = Q\ f \sum_{i=1}^{t_c} (1 - Q\ f)^{i-1} (t_c - i) \tag{16}$$

Estrutura a:

A estrutura **a** pode ser vista como um caso especial da estrutura **c**, no qual o miss ratio é igual a 1. Conforme a equação (2), o parâmetro m assume o valor de Q. Além disso, σ é igual a zero, já que nunca ocorre uma substituição. Deve-se ainda atentar para o fato de que a cache no modelo de Patel é capaz de atender os acessos do processador sem ciclos de espera na ocorrência de um hit. Isto significa que o processo de procura dentro da cache pelo bloco referenciado é executado num único ciclo. Este processo se realiza tanto quando ocorre um hit, como quando ocorre um miss. Na estrutura **a**, em que não há cache, não se tem a fase da procura. Numa hipótese pessimista em relação à cache, admite-se aqui que, na estrutura **a** se economiza um ciclo nos acessos á memória local. Assim, tem-se:

$$U_a = \frac{1}{1 + Q\ (w + t_c - 1)} \tag{17}$$

$$m_a' = Q\ (w + t_c)\ U_a \tag{18}$$

$$B_a = N\ Q\ t_c\ U_a \tag{19}$$

Estrutura b:

Na estrutura **b** também não existe uma memória cache (m = Q), e o atraso w não afeta os acessos à memória local. Obtém-se portanto:

$$U_b = \frac{1}{1 + Q\,[z_l t_l + z_c(w + t_c) - 1]} \tag{20}$$

$$m_b' = Q\, z_c\, (w + t_c)\, U_b \tag{21}$$

$$B_b = N\, Q\, z_c\, t_c\, U_b \tag{22}$$

Estrutura d:

As equações para a estrutura **d** podem ser obtidas a partir das seguintes considerações. A cada k ciclos ativos são gerados k Q z_c acessos à memória compartilhada e k Q f z_l misses na cache. Multiplicando-se o número de acessos à memória local e à memória compartilhada pelos correspondentes atrasos, e dividindo por k, obtém-se Q z_l f (σ w' + t_l) + Q z_c(w + t_c - 1) ciclos passivos para cada ciclo ativo. Disso resultam as seguintes equações para a Utilização do Processador, Taxa de Solicitação à Rede de Interconexão e Número Médio de Módulos de Memória Ocupados:

$$U_d = \frac{1}{1 + Q\, z_l\, f\, (\sigma\, w_l' + t_l) + Q\, z_c (w + t_c - 1)} \tag{23}$$

$$m_d' = Q\, z_c\, (w + t_c)\, U_d \tag{24}$$

$$B_d = N\, Q\, z_c\, t_c\, U_d \tag{25}$$

O atraso médio no "buffer" da cache pode ser calculado pela equação (5), onde a probabilidade de que um miss ocorra exatamente depois da primeira unidade de tempo de um acesso de escrita (na equação (5), o próprio m) é igual a Q f z_l.

$$w_l' = Q\, f\, z_l \sum_{i=1}^{t_l} (1 - Q\, f\, z_l)^{i-1} (t_l - i) \tag{26}$$

Estrutura e:

Admitindo-se que a probabilidade de que um bloco substituído na cache pertence à memória compartilhada é igual a z_c[1], obtém-se as seguintes equações para a estrutura **e**:

$$Ue = \frac{1}{1 + Q\, f\, [\sigma\, (z_c\, w_c' + z_l\, w_l') + z_l t_l + z_c (w + t_c)]} \tag{27}$$

$$m_e' = Q\, f\, z_c\, (1 + \sigma)\, (w + t_c)\, U_e \tag{28}$$

$$B_e = N\, Q\, f\, z_c\, (1 + \sigma)\, t_c\, U_e \tag{29}$$

w_c' e w_l', que representam o atraso devido ao "buffer" respectivamente para acessos à memória compartilhada e local, são calculadas pelas seguintes expressões:

---

[1] *Esta hipótese corresponde à situação em que a distribuição de códigos e dados entre memória local e compartilhada se faz de modo uniforme. O que ocorre nos casos reais, no entanto, é que os programas são armazenados na memória local, de tal forma que a maioria das atualizações ocorrem na memória compartilhada.*

$$w_l' = Q\ f \sum_{i=1}^{t_l} (1 - Q\ f)^{i-1}(t_l - i) \qquad (30)$$

$$w_c' = Q\ f \sum_{i=1}^{t_c} (1 - Q\ f)^{i-1}(t_c - i) \qquad (31)$$

As equações (7) para "crossbar" e as equações (10) a (12) para rede de múltiplos estágios são válidas para todas as estruturas.

5. Os Parâmetros do Modelo

O próximo passo é determinar valores plausíveis para cada um dos parâmetros do modelo. Nas linhas seguintes são escolhidas faixas de valores para cada parâmetro com base em resultados apresentados em vários trabalhos publicados.

5.1 Parâmetros Associados ao Sistema

A exceção de sistemas com um barramento compartilhado[2], considera-se aqui multiprocessadores com 32 módulos de processamento e 32 módulos de memória. É de conhecimento geral que em sistemas com um barramento compartilhado ocorre a saturação do barramento já com muito menos do que 32 módulos de processamento. Por isso os valores apresentados na seção 6 para Utilização do Processador em sistemas com barramentos compartilhados supõem apenas 4 módulos de processamento.

Admite-se que a tecnologia (e consequentemente o tempo de acesso) da memória local e dos módulos de memória é a mesma, a menos que seja mencionado explicitamente o contrário. Esta hipótese não implica, no entanto, em valores idênticos para t_l e t_c. O tempo de propagação dos sinais através da rede de interconexão deve ser considerado em t_c, de tal sorte que t_c é maior do que t_l. Se este tempo de propagação é representado pelo parâmetro μ, tem-se:

$$t_c = t_l + \mu \qquad (32)$$

O valor de μ depende preponderantemente da tecnologia e do número de estágios da rede de interconexão. Na seção 6 admite-se que o atraso provocado por cada estágio da rede de interconexão é igual a uma unidade de tempo, de tal forma que o atraso provocado pela rede de interconexão é numericamente igual ao número de estágios que a compõem. Admite-se que as chaves elementares que constituem a rede de múltiplos estágios possuem 2 entradas e 2 saídas (a = b = 2).

Para simplificar a comparação entre as estruturas **a** e **b** e entre as estruturas **c** e **d**, admite-se que a via de dados tem largura igual ao número de bits de um bloco. Isto significa que toda uma linha pode ser carregada na cache, ou escrita na memória principal de uma vez.

Tendo em vista a velocidade das memórias dinâmicas atuais, é razoável admitir valores para $t_l = t_c - \mu$ entre 3 e 8. Esta faixa é coerente com os valores apresentados em [WOOD86] e [PFIS85a].

5.2 Parâmetros Associados ao "Workload"

A medida do miss ratio para diversos tipos e organizações de memórias cache foi objeto de inúmeros trabalhos publicados (p. ex. [KAPL83], [SMIT82]). Estes trabalhos indicam que valores de miss

---

2 *O multiprocessador com um barramento compartilhado é modelado fazendo M = 1.*

ratio em torno de 10% podem ser alcançados mesmo por caches de pequena capacidade.

De [MGRE84] pode-se inferir, que o valor do parâmetro Q para o processador Motorola MC68020 fica entre 65% e 85% em sistemas com um único processador. Admite-se que o valor deste parâmetro para multiprocessadores é aproximadamente o mesmo.

A probabilidade σ, de que um bloco substituído na cache deva ser atualizado na memória principal é igual a 1 na estratégia BWB. Smith [SMIT85] constatou que para a estratégia BFWB cerca de 50% dos blocos que contém dados apresentam conteúdo distinto da memória principal, no momento em que são substituídos. Os resultados apresentados na seção 6 correspondem a σ = 25%, já que nem todos os blocos substituídos na cache contém dados.

A probabilidade z_c, de que o processador nas estruturas **b**, **d** e **e** solicite uma informação da memória compartilhada, depende em grande medida da aplicação. É em geral fácil armazenar programas nas memórias locais, de tal forma que a memória compartilhada só seja referenciada nos acessos a operandos. Esta medida por si só já reduz o valor de z_c a menos do que 50% na maioria das aplicações [SMIT85]. Nesta análise atribui-se o valor de 30% a z_c.

6. Aplicação do Modelo

O modelo apresentado foi aplicado para os parâmetros variando numa ampla faixa de valores, definida em conformidade com a discussão da seção anterior.

O desempenho do multiprocessador é proporcional à Utilização dos Processadores. Por isso, decidiu-se comparar as estruturas de memória com base nas respectivas Utilizações do Processador.

Constatou-se que o parâmetro que mais influencia o desempenho relativo das estruturas consideradas é o miss ratio. Por esta razão, o desempenho relativo dos multiprocessadores para as diferentes estruturas de memória é apresentado nas figuras 4 a 6 em função do miss ratio.

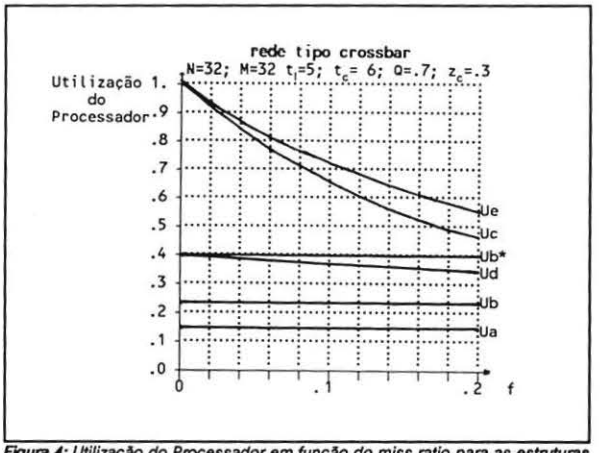

Figura 4: Utilização do Processador em função do miss ratio para as estruturas de memória analisadas, considerando uma rede de interconexão "crossbar".

A estrutura indicada por **b*** corresponde à estrutura **b** com memória local rápida. Neste caso atribui-se ao parâmetro t_i o valor 1, o que implica em zero ciclo de espera para acessos à memória local.

Os resultados mostrados na figura 4 dizem respeito a um sistema com uma rede de interconexão tipo "crossbar". Observando-se os valores indicados na figura, verifica-se a seguinte ordem segundo a Utilização do Processador:

$$U_e > U_c > U_b^* > U_d > U_b > U_a$$

Deve-se salientar que esta ordem não se altera mesmo variando-se os demais parâmetros do modelo numa ampla faixa de valores.

A introdução de uma memória local, que transforma a estrutura **a** na estrutura **b** apresenta-se como responsável por um aumento de cerca de 50% no desempenho do sistema. Este aumento se deve preponderantemente à redução da contenção na memória compartilhada. Observe-se que para sistemas com barramento compartilhado (vide fig. 6), onde o problema da contenção é mais crítico, o aumento

foi de cerca de 200%.

A redução do tempo de acesso à memória local (estrutura **b*** com t_l = 1) produz, para rede "crossbar", um aumento de cerca de 60% sobre a estrutura **b** com memória local lenta. Comparando com os resultados da figura 6, a redução do tempo de acesso à memória local resulta em pequeno aumento de desempenho. Em tais sistemas a contenção é portanto o fator determinante do desempenho.

Observe-se ainda que a estrutura **d** apresenta um desempenho muito próximo ao da estrutura **b** com memória local rápida, para todas as estruturas de interconexão consideradas. Uma escolha entre elas será determinada pelo custo, o que favorecerá na maioria dos casos a estrutura **d**.

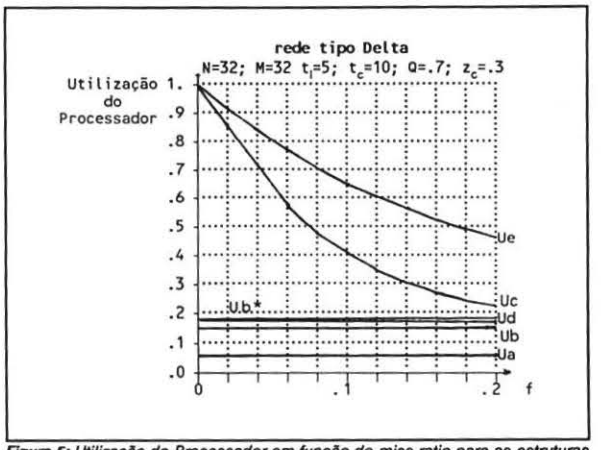

Figura 5: Utilização do Processador em função do miss ratio para as estruturas de memória analisadas, considerando uma rede de interconexão Delta de 5 estágios.

Para f ≈ .1, um valor pouco otimista para o miss ratio, a cache da estrutura **c**, que pode conter dados dos módulos de memória, já apresenta desempenho cerca de 80% a 100% superior ao da estrutura **d** com uma cache dedicada à memória local. Isto evidencia a conclusão já apresentada em [PATE82] de que a memória cache (cf. fig. 2c) promove um expressivo aumento de desempenho no sistema.

Na figura 4, nota-se que a introdução de uma memória local na estrutura **c**, para assim construir a estrutura **e**, produz, para pequenos valores do miss ratio, pouca vantagem. Em termos de desempenho, para f ≈ .1, a estrura **e** supera a estrura **c** em cerca de 10%. Observe-se ainda que a superioridade da estrutura **e** sobre a estrutura **c** se torna cada vez menor com a redução do miss ratio. Uma memória cache de capacidade elevada com um mecanismo de coerência eficiente pode portanto simplesmente dispensar o uso de uma memória local.

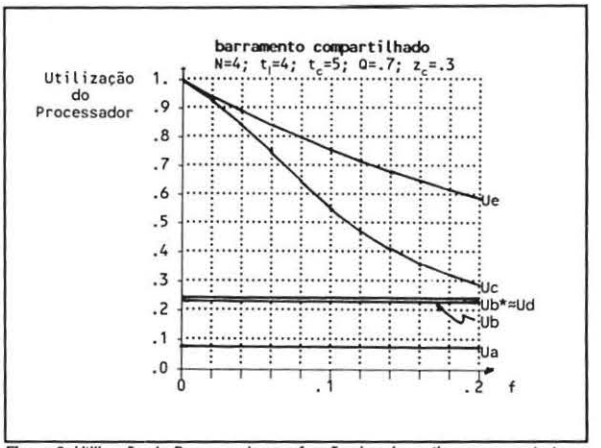

Figura 6: Utilização do Processador em função do miss ratio para as estruturas de memória analisadas, considerando um sistema com um barramento compartilhado.

A ordem das estruturas apresentadas nas linhas anteriores não se altera dentro da faixa de valores para o miss ratio mostrada nas figuras 4 a 6. Nota-se, contudo, que o desempenho da estrutra **c** se aproxima do das estruturas **a**, **b** e **d** à medida que f cresce. Valores de miss ratio superiores a 20% podem ser considerados pessimistas, somente observados em caches muito pequenas ou quando o mecanismo de coerência utilizado é muito ineficiente. Se o miss ratio, por um ou outro motivo, for significativamente superior a 20%, a estrutura **d** passa a ser a mais atrativa. Ela só é superada pela estrutura **b** com memória local rápida e pela estrutura **e**. A primeira é uma opção normalmente excluída por razões de custo, e a segunda exige, ao contrário da estrutura **d**, um mecanismo de coerência.

As figuras 5 e 6 apresentam respectivamente os resultados para uma rede de interconexão de múltiplos estágios e para um sistema com um barramento compartilhado. As conclusões apresentadas até aqui se verificam de um modo geral também para estes sistemas.

No caso de um sistema com rede de interconexão de múltiplos estágios só se pode abdicar da memória local sem incorrer em degradação significativa de desempenho se, devido às caracteristicas da cache e da aplicação, for razoável admitir que o miss ratio será baixo (f < .05).

Para sistemas com barramento compartilhado o desempenho apresentado pelas estruturas **c** e **e** se tornam próximos apenas para valores muito pequenos do miss ratio. Em tais sistemas a estrutura **e** se impõe como a melhor.

7. Conclusão

Um modelo analítico para avaliação de diferentes estruturas de memória para multiprocessadores foi apresentado. Ele se baseia num modelo já proposto anteriormente. Por este modelo foi calculado a Utilização dos Processadores para diferentes estruturas de memória. O modelo foi aplicado considerando 3 tipos diferentes de estruturas de interconexão: rede "crossbar", rede de múltiplos estágios e barramento compartilhado.

Os resultados indicaram:

- Uma memória cache que pode conter cópias de blocos da memória compartilhada (estruturas **c** e **e**), produz um significativo aumento de desempenho.
- Entre os parâmetros do modelo, aquele que influencia de modo mais expressivo os valores de desempenho relativo é o miss ratio da cache. Constatou-se que a redução do miss ratio provoca um significativo aumento de desempenho do sistema.
- Para sistemas com uma rede de interconexão com poucos estágios, a introdução de uma memória local num sistema que já dispõe de memórias cache privadas (passagem da estrutura **c** para a estrutura **e**) produz um pequeno aumento de desempenho, que pode não justificar a utilização da memória local.
    - se a rede de interconexão possui muitos estágios ou se constitui de um único barramento compartilhado, a utilização de uma memória local ao lado da cache traz benefícios apreciáveis, mesmo quando a cache opera com miss ratio reduzido.

8. Bibliografia

[ARCH84] Archibald, J.; Baer, J.L.: An Economical Solution on the Cache Coherence Problem. 11th Ann. Int. Symp. Comp. Arch., June 1984, pp. 335-362.

[ARCH85] Archibald, J.; Baer, J.L.: An Evaluation of Cache Coherence Solutions in Shared-Bus Multiprocessors. Tech. Rep. 85-10-05, Department of Computer Science, University of Washington, Seattle, WA 98185.

[BHAN75] Bhandarkar, D.P.: Analysis of Memory Interference in Multiprocessors. IEEE Trans. Comp., Vol. C-24, Sept. 1975, pp. 897-908.

[BITA86] Bitar, P.; Despain, A.M.: Multiprocessor Cache Synchronization, Issues, Innovations, Evolution. 13th Ann. Int. Symp. Comp. Arch., June 1986, pp. 424-433.

[BOHA83] Bode, A.; Handler, W.: Rechnerarchitektur II, Strukturen. Springer Verlag, 1983.

[CENS78] Censier,L.M.; Feautrier, P.: A New Solution to Coherence Problems in Multicache Systems. IEEE Trans. Comp., Vol. C-27, No. 12, Dec., 1978, pp. 1112-1118.

[DUBO82] Dubois, M.; Briggs, F.A.: Effects of Cache Coherency in Multiprocessors. IEEE Trans. Comp., Vol. C-31, No. 11, Nov. 1982, pp. 312-328.

[DUBO87] Dubois, M.: Effect of Invalidations on the Hit Ratio of Cache-based Multiprocessors, Proc. Int. Conf. Parallel Processing, 1987, pp. 255-257.

[FEIT88] Feitosa, R. Q.: Speicherstrukturen von Speichergekoppelten Multiprozessoren, Arbeitsberichte des IMMD, Universitaet Erlangen Nuernberg, Tese de Doutorado.

[FEIT90] Feitosa, R. Q.: O Problema de Coerência de Memórias Cache Privadas em Grandes Multiprocessadores para Aplicações Numéricas: Uma Nova Solução, Anais do XVII SEMISH, 1990.

[KAPL73] Kaplan, K.R.; Winder, R.O.: Cache based Computer Systems. Computer, Mar. 1973, pp. 30-36.

[KATZ85] Katz, R.H.; Eggers, S.J.; Wood, D.A.; Perkins, D.L.; Sheldon, R.G.: Implementing Cache Consistency Protocol, 12th Int. Ann. Symp. Comp. Arch., June 1985, pp. 276- 283.

[MGRE84] MacGregor, D.; Mothersole, D.; Moyer, B.: The Motorola 68020. IEEE Micro, Vol. 4, No. 4, Aug. 1984, pp. 101- 118.

[PATE81] Patel, J.H.: Performance of Processor-Memory Interconnection for Multiprocessors. IEEE Trans. Comp., Vol. C-30, Oct. 1981, pp. 771-780.

[PATE82] Patel, J.H.: Analysis of Multiprocessors with Private Cache Memories. IEEE Trans. Comp., Vol. C-31, April 1982, pp. 296-304.

[PFIS85] Pfister, G.F.; Brantley, W.C.; George, D.A.; Harvey, S.L.; Kleinfelder, W.J.; McAuliffe, K.P.; Melton, E.A.; Norton, V.A.; Weiss, J.: The IBM Research Parallel Processor Prototype (RP3): Introduction and Architecture. 12th Ann. Int. Symp. Comp. Arch., 1985, pp. 764-771.

[RAVI72] Ravi, C.V.: On the Bandwidth and Interference in Interleaved Memory Systems. IEEE Trans. Comp., Vol. C- 21, Aug. 1972, pp. 899-901.

[SIEG85] Siegel, H.J.: Interconnection Network for Large-Scale Parallel Processing, Lexington Books, 1985.

[SMIL85] Smilauer, B.: General Model for Memory Interference in Multiprocessors and Mean Value Analysis. IEEE Trans. Comp., Vol. C-34, No. 8, Aug. 1985, pp. 744-751.

[SMIT82] Smith, A.J.: Cache Memories. Computing Surveys 14, Sept. 1982, pp. 473-530.

[SMIT85] Smith, A.J.: Cache Evaluation and the Impact of Workload Choice. 12th Ann. Int. Symp. Comp. Arch., 1985, pp. 64- 73.

[VERN86] Vernon, M.K.; Holliday, M.A.: Performance Analyses of Multiprocessor Cache Consistency Protocols Using Generalized Timed Petri Nets. Performance Evaluation Review, Vol. 14, No. 1, May 1986, pp. 9-17.

[WOOD86] Wood, D.A., et al.: A In-Cache Address Translation Mechanism. 13th Ann. Int. Symp. Comp. Arch., June 1986, pp. 365-385.

[YEN-82] Yen, D.W.L.; Patel, J.H.; Davidson, E.S.: Memory Interference in Synchronous Multiprocessor Systems. IEEE Trans. Comp., Vol. C-31, No. 11, Nov. 1982, pp. 1116-1121.