Título do Artigo: Comprehensive Multivariate Extrapolation Modeling of Multiprocessor Cache Miss Rates
Citação Bibliográfica: ILYAGLUHOVSKY,DAVIDVENGEROV,and BRIANOíKRAFKA; "Comprehensive Multivariate Extrapolation Modeling of Multiprocessor Cache Miss Rates"; ACM Transactions on Computer Systems,Vol.25, No.1, February 2007
Autor do Resumo: Alexandre K. I. De Mendonça (075537)

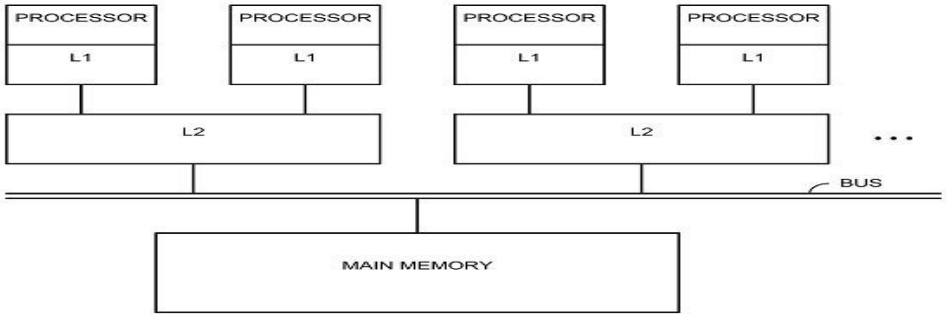

Fig. 1. High level diagram of a memory system.

Nos primeiros estágios do processo de desenvolvimento de multiprocessadores, é uma prática comum usar modelos em alto-nível para explorar uma larga quantidade de opções de projeto. Estes modelos em alto-nível necessitam de estimativas de taxas de erros de chache para os workloads alvo. Uma hierarquia típica de memória que é demonstrada na fig. 1, tem múltiplos níveis de cache com diferentes quantias de compartilhamento de cache. Cada nível tem opções como tamanho, tamanho da linha, associatividade, compartilhamento, latência e bandwith.

Em resumo, o problema de estimativa da taxa de miss é o seguinte: havendo a capacidade de simular apenas um subconjunto das configurações de cache, necessitamos de um modelo multivariável para todos os miss rates requeridos por todas as configurações de cache baseadas, em um limitado número de dados de simulação.

Os autores observaram que os miss rates de caches normalmente seguem uma certa combinação de requisitos: Elas são naturalmente não negativas, os rates de misses em clean lines são esperadas a diminuir quando o tamanho da cache e associatividade aumentam desta maneira minimizando os retornos tipicamente observados. Isto sugere encaixar um modelo convexo não negativo e não crescente nos parâmetros acima mencionados.

Por outro lado, taxas de transferência cache-cache tipicamente aumentam junto com o tamanho e a associatividade pois em maiores tamanhos, menos são as linhas compartilhadas de escrita-leitura são substituídas antes de serem acessadas por outra cache. Este efeito tipicamentediminui conforme o tamanho e a associatividade aumenta, tendo assim um modelo côncavo e não decrescente em ambos os parâmetros

Então os autores modelam e provam uma função matemática baseada em Splines para as estrapolações, estas splines tem a vantagem de ter uma representação compacta pois formam um espaço linear, e são as melhores alternativas para modelos de regressão. Como as splines são definidas pedaço por pedaço, são ótimas para modelagem local, como a concatenação das peças é suave, esta classe é apropriada para a modelagem de funções suaves como a de taxa de erros de memórias cache.

Para a simulação, os autores utilizaram os seguintes parâmetros: Tamanho da cache, Tamanho da Linha, Associatividade, Compartilhamento(Processadores/Cache), Número de Processadores. Bem como 2 workloads diferentes o *sap* o *trans,* no primeiro teste com o workload trans foi testado o miss rate de loads clean, os dados resultantes mostram que quanto mais threads compartilham a mesma cache, os miss rates mostram um aumento quando as diferentes threads interferem nas outras. No segundo teste foi comparado os miss rates x compartilhamento para diferentes tamanhos, ambos os workloads mostram aumentos das miss rates conforme o compartilhamento aumenta, e o efeito é menor em caches com tamanhos maiores. No próximo teste é feito uma extrapolação de como seriam os resultados com ate 512 threads. Os dados para ambos os workloads indicam que as clean miss rates são insensitivas ao número de threads, mostram um pequeno aumento para um pequeno número de threads, e são praticamente planas para os maiores números.

Após são realizados os mesmos testes porém para escritas cache-cache e os autores concluem que a taxa de transferência cache-cache melhoram com o compartilhamento, enquanto as clean rates diminuem.