Um Breve Estudo da Adição de um Quarto Nível de Cache em Processadores Multicore

Carlos Eduardo Benevides Bezerra[1], **Cláudio Fernando Resin Geyer**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

Abstract. *The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. The abstract may be up to 3 inches (7.62 cm) long. Leave two blank lines after the Abstract, then begin the main text.*

1. Introdução

Nos últimos anos, a freqüência de relógio dos processadores tem aumentado bastante [FIXME:ref]. Paralelamente a isso, e devido a uma certa estagnação da velocidade dos cores [FIXME:ref], os fabricantes têm optado pela produção de processadores com vários núcleos de processamento (processadores *multicore*, com diversos *cores*). Exemplos disso são as arquitetura da Intel (*Dunnington* [FIXME:ref], *Nehalem* [FIXME:ref] etc.), AMD (*Barcelona* [FIXME:ref], *Budapest* [FIXME:ref], *Deneb* [FIXME:ref] etc.) e Sun (*Niagara* [FIXME:ref], *Victoria Falls* [FIXME:ref] etc.).

[FIXME:fig.arch.Dunnington]

As memórias cache servem para reduzir o número de ciclos de latência que um processador deve esperar para acessar um determinado dado na memória. Reduzindo o número de ciclos, reduz-se o tempo necessário para acessar um dado. Por esta razão, o tempo de acesso a dados na memória cache é consideravelmente menor do que o tempo necessário para acessar a memória principal, por exemplo.

Contudo, ultimamente as aplicações têm cada vez um volume maior de dados a processar. Essa quantidade maior de dados, se não puderem ser acessados rapidamente pela unidade de processamento, podem implicar um pior desempenho do sistema [FIXME:ref]. Outra questão é que, nas arquiteturas multicore, vários núcleos podem estar utilizando a memória ao mesmo tempo. Por causa dessas razões, é recomendável aumentar o tamanho da memória cache (que tem acesso rápido) proporcionalmente ao aumento do volume de dados das aplicações e ao número de núcleos do processador.

A grande questão é que aumentar a memória cache, pura e simplesmente, pode não ser suficiente para reduzir o tempo de execução das tarefas realizadas pelo processador. Na verdade, um aumento sem critérios do espaço da memória cache pode implicar uma redução da velocidade do sistema. Isso acontece porque quanto maior a memória, maior é o seu tempo de acesso, devido a uma estrutura mais complexa de endereçamento [FIXME:ref]. Além disso, se vários núcleos estão tentando acessar a memória ao mesmo tempo, haverá uma contenção pelo barramento de acesso àquela memória, fazendo com que um ou mais processadores esperem até que o barramento seja liberado. Uma má

escolha de estrutura e tamanho de memórias cache pode fazer com que uma arquitetura com mais memória seja mais lenta do que uma arquitetura com menos memória, porém estruturada de maneira mais eficiente.

Tendo esses aspectos em vista, neste trabalho foi feito um breve estudo da influência da estrutura de cache em processadores multicore, comparando o desempenho de uma hierarquia com 3 níveis de memória, comparada com outras duas hierarquias, ambas com 4 níveis, introduzindo um nível L4 compartilhado entre todos os núcleos. Como contribuição, neste trabalho foi testado o uso de uma hierarquia de memória entrelaçada – o que será descrito nas próximas seções –, com o objetivo de reduzir a profundidade da busca realizada por um núcleo para buscar um dado compartilhado com outro núcleo.

O texto está organizado da seguinte forma: na seção 2, é apresentado um pouco do contexto dentro do qual este trabalho se insere; na seção 3, é apresentada em maiores detalhes a proposta deste trabalho, incluindo a idéia de entrelaçar a hierarquia da memória cache; na seção 4, são apresentados os detalhes da modelagem e parâmetros das simulações que foram realizadas; na seção 5, são mostrados os resultados e uma breve análise dos mesmos e, na seção 6, são apresentadas as conclusões a que se chegou com este trabalho.

2. Contexto e motivação

3. Proposta

Este trabalho teve dois objetivos, que são procurar evidências de que a adição de um quarto nível de memória cache é mais benéfico do que aumentar a memória cache L3 e avaliar o ganho de desempenho conseguido ao se entrelaçar a hierarquia de memórias cache. Para analisar a diferença de desempenho com uma cache L3 grande, contra a adição de um módulo de memória L4, foram comparadas as arquiteturas seguintes:

- uma arquitetura multicore, inspirada na arquitetura Dunnington, utilizada no Intel Xeon, que dispunha de 8 núcleos de processamento, com quatro módulos de memória cache L2, cada um compartilhado por um par de núcleos, e uma grande memória L3 (com 32 megabytes de espaço), compartilhada entre todos os oito núcleos (Figura 1);
- uma arquitetura multicore, semelhante à anterior, também com oito núcleos, porém os 32 megabytes que eram da L3 foram divididos em: duas memórias L3 de 8 megabytes, cada uma compartilhada por metade dos núcleos, e uma memória L4 de 16 megabytes, compartilhada por todos (Figura 2).

Outra contribuição do trabalho foi que também se propôs o entrelaçamento da hierarquia das memórias cache. Para buscar algum resultado que indique se há algum benefício ao ser utilizado esse tipo de abordagem, foi comparadas as seguintes arquiteturas:

- uma arquitetura com quatro níveis de cache (L1, L2, L3 e L4), onde cada módulo de cache só tem um módulo de nível inferior na hierarquia de memória, tal qual na Figura 2;
- com outra arquitetura, na qual cada cache L2 tem duas L3 no nível inferior da hierarquia; além disso, a outra mudança com relação à arquitetura

  o primeiro era . A razão para isso era

  Please read the following carefully.

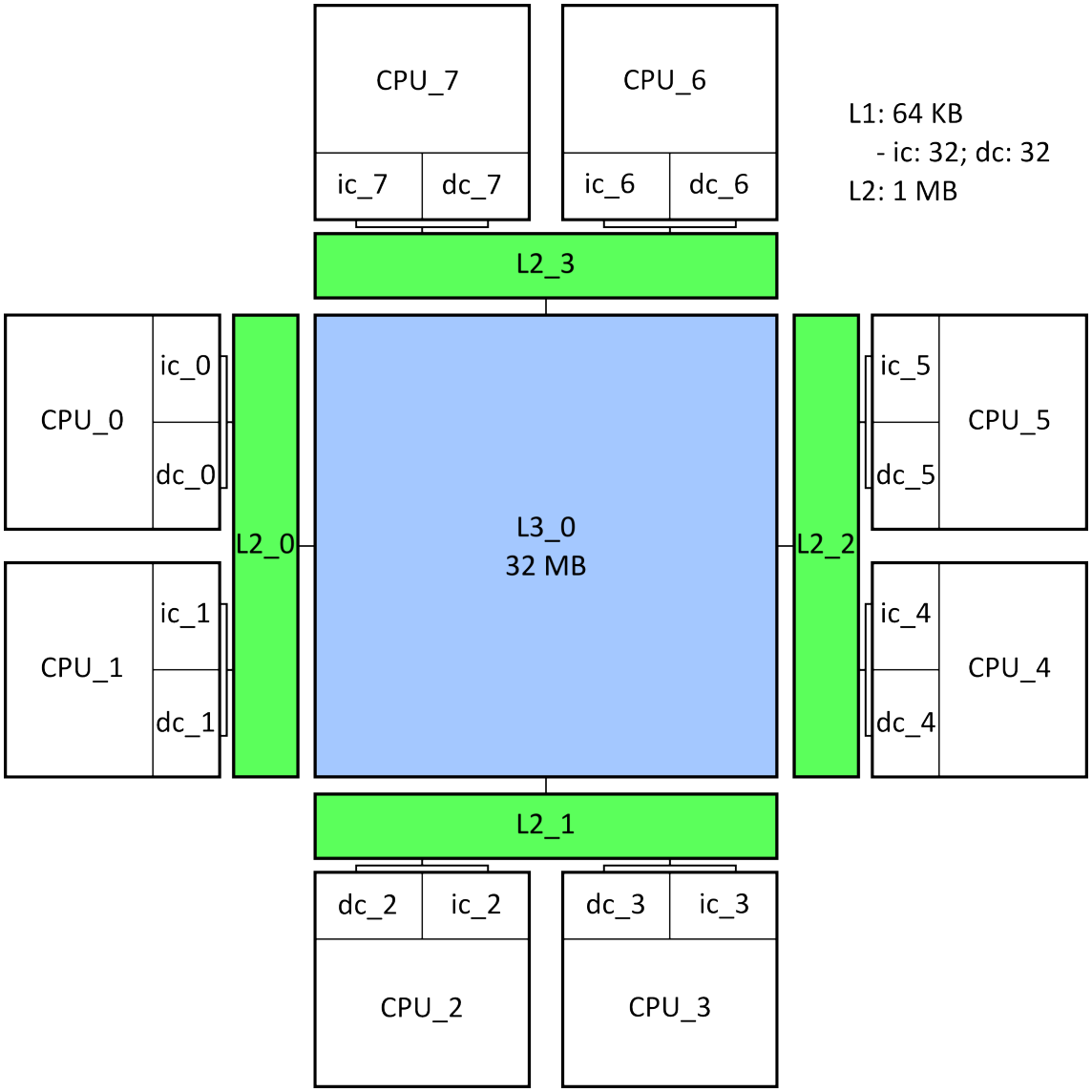

Figura 1. Arquitetura com 3 níveis de memória cache

4. Modelagem e simulações

All manuscripts must be in English.

5. Resultados

Print your properly formatted text on high-quality, 8.5×11-inch white printer paper. A4 paper is also acceptable, but please leave the extra 0.5 inch (1.27 cm) at the BOTTOM of the page.

6. Conclusões

All printed material, including text, illustrations, and charts, must be kept within a print area 6-7/8 inches (17.5 cm) wide by 8-7/8 inches (22.54 cm) high. Do not write or print anything outside the print area. Number your pages lightly, in pencil, on the upper right-hand corners of the BACKS of the pages (for example, 1/10, 2/10, or 1 of 10, 2 of 10, and

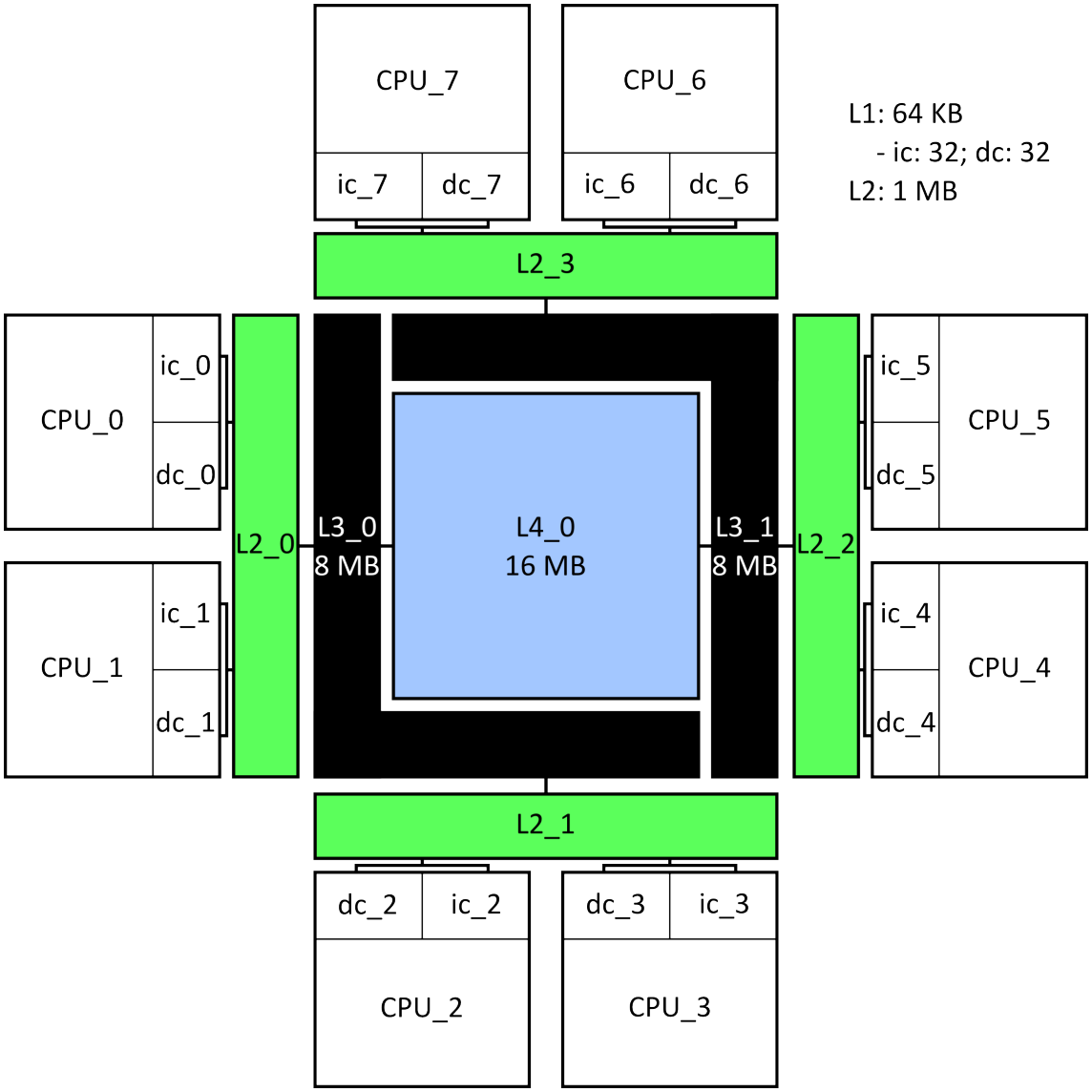

Figura 2. Arquitetura com 4 níveis de memória cache

so forth). Please do not write on the fronts of the pages, nor on the lower halves of the backs of the pages.

Agradecimentos