Universidade: presente!

UFRGS PROPESQ

XXXI SIC

Salão UFRGS 2019 – Conhecimento o Formação o Inovação o

21 . 25 . OUTUBRO . CAMPUS DO VALE

Caracterização de acessos não-temporais à memória em arquiteturas de processadores multicores

Lucas Lauck dos Passos

RESUMO

As aplicações com baixa localidade de dados são muitas vezes penalizadas pela extensa hierarquia de memória cache, tanto pela latência de acesso quanto pelo alto consumo de energia que é dispensável. Estas aplicações são popularmente chamadas de *streaming*, comuns na área de transmissão contínua de vídeo e áudio. Este trabalho tem como objetivo caracterizá-las em termos do grau de reuso de dados em memórias cache.

INVESTIGAÇÃO

Foram retirados os seguintes dados dos programas:

- Cold Reference (primeira referência a um endereço de memória)
- Distância de reuso na memória cache.
- Número de acessos e misses.
- Stride (distância de leituras subsequentes na memória).

RESULTADOS PRELIMINARES

Nas aplicações streaming percebe-se um número mais significativo de Cold Reference (CR) decorrente da falta de reuso (Fig. 1). Isolando as 4 aplicações que mais se diferenciam entre si em distância de reuso na cache (Fig. 2) percebe-se um número muito semelhante de CR e miss rate no primeiro nível de cache de dados (L1d) nas streaming. Além disso, a taxa de miss rate no último nível de cache de dados (LLd) se mostra altíssima e todos os misses são de um novo endereço nunca acessado. Com esses dados foi calculado um grau de streaming para as aplicações (Quadro). Válida para aplicações com miss rate LLd muito próximo de 100%:

- Quanto mais perto o valor de grau de streaming for do miss rate L1d menor a localidade temporal da aplicação.
- Quanto mais próximo de 100% o grau de streaming, menor a localidade espacial da aplicação.

Uma segunda característica que se observa em todos os programas é o grau de reuso em uma mesma linha de cache e numa distância um (Fig. 1). Esse pico é um dos mostradores da localidade espacial nos testes. Por esse motivo apesar de streaming ter pouco reuso temos muitos hits na L1d. Outro forte indicativo de localidade espacial é o alto índice de leituras contínuas na memória menor que 8 B (Fig. 3). Mostrando até mesmo uma proporção entre eles.

Na tabela é possível ver a comparação entre todos os testes. Note que as aplicação não streamings se beneficiam demasiadamente da hierarquia de memória.

Programas	Miss rate L1d	Miss rate LLd	G. streaming	CR	Dist [0,2[	Stride < 8 B
Busca em vetor	6,25%	99,99%	6,25%	6,25%	93,72%	99,95%
Daxpy	6,25%	99,30%	6,25%	6,25%	93,72%	100%
Add	6,25%	99,96%	6,25%	6,25%	93,69%	99,94%
L.E(ORD)	6,19%	94,66%	5,64%	6,07%	81,70%	86,37%
L.E(RAND)	21,98%	93,92%	19,91%	21,11%	43,21%	37,25%
Stencil-2d	0,30%	0,23%	0%	0%	89,01%	99,98%
Floyd–Warshall	0,30%	0,10%	0%	0%	83,55%	99,97%

INTRODUÇÃO

A localidade pode aparecer de duas formas:

- Localidade temporal: quando os dados se repetem ao longo do tempo.
- Localidade espacial: quando os dados são lidos de forma sequencial na memória.

IMPLEMENTAÇÃO

Os códigos foram compilados no GCC 9.2 com flag de otimização -O3, sem qualquer tipo de vetorização, flag -fno-tree-vectorize. Foram usados códigos dos benchmarks PolyBench e Stream Bench.

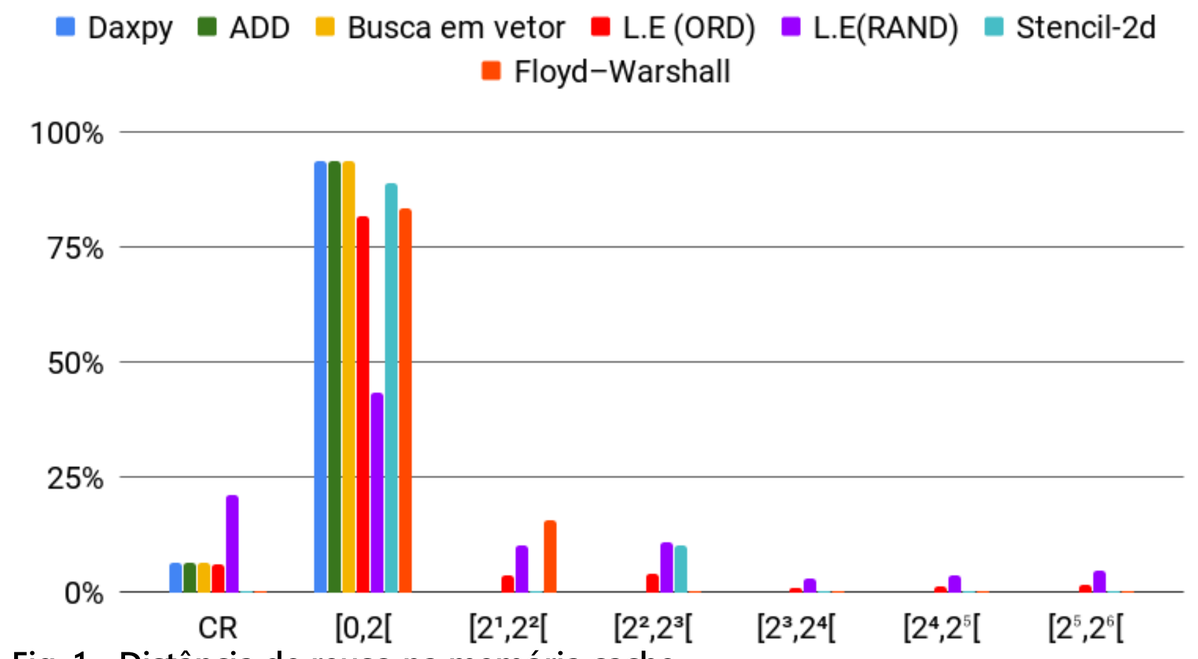

Fig. 1 - Distância de reuso na memória cache.

$$CRmisses = \frac{CR}{misses\ LLd}, \quad CRtotal = \frac{CR}{Total\ acessos}, \quad Gs = CRmisses * CRtotal * 100$$

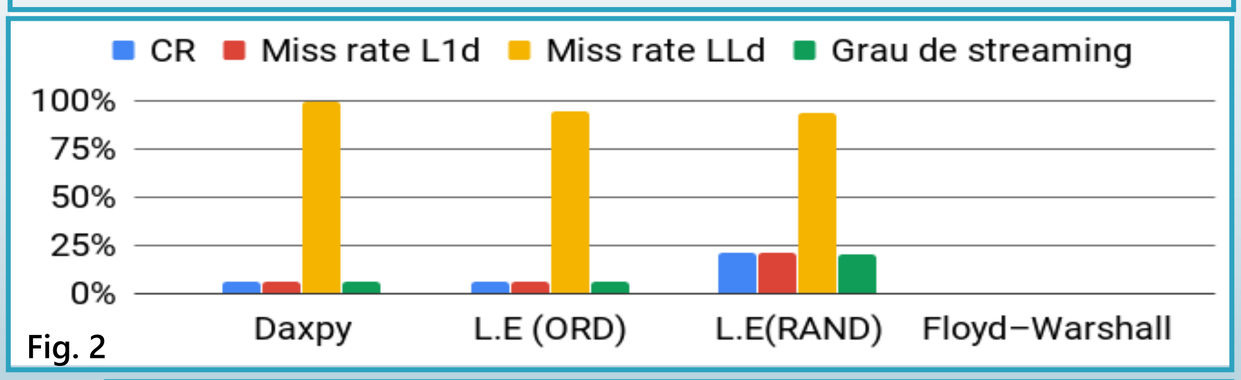

Fig. 2

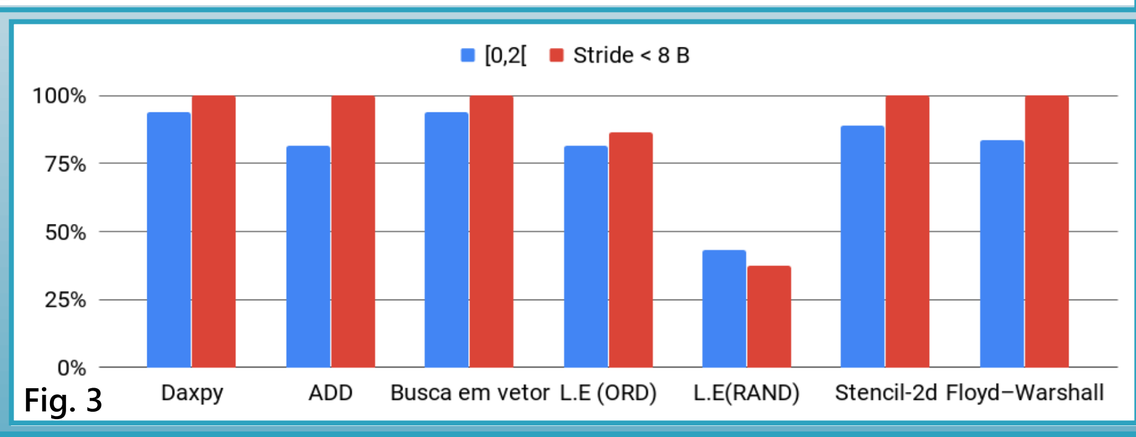

Fig. 3

CONCLUSÃO

Identificar o comportamento das aplicações é essencial para podermos usar artifícios de otimização como o bypass, onde o processador ignora a presença da cache com o objetivo de economizar energia ou, até mesmo, aumentar o desempenho. Apesar de a cache de primeiro nível ser usada intensamente devido a particularidades das aplicações analisadas, o último nível de cache não foi utilizado em qualquer momento. Essas métricas foram obtidas apenas em tempo de execução, o que motiva trabalhos futuros para incluir análises com base em contadores de hardware em bibliotecas para otimização de acessos não-temporais à memória.