Modalidade do trabalho: Relatório técnico-científico
Evento: XXII Seminário de Iniciação Científica

COMPUTAÇÃO PARALELA COM ACELERADORES GPGPU[1]

Emilio Hoffmann De Oliveira[2], Edson Luiz Padoin[3].

[1] Trabalho de Conclusão de Curso
[2] Aluno do Curso de Ciência da Computação - emiliohoffmann@hotmail.com
[3] Professor-Orientador, Mestre em Ciência da Computação – padoin@unijui.edu.br

1. Introdução

Nos últimos anos o debate a respeito do esgotamento dos recursos naturais está cada vez mais presente em todas as áreas, em busca de diminuir o consumo dos recursos energéticos, os programadores e cientistas da computação almejam melhorar a performance de softwares para reduzir o tempo de processamento e o consumo de energia. Uma das formas de tornar isso possível é utilizando a paralelização do processamento, recentemente um novo modo de se paralelizar vem se destacando, que são as unidades de processamento gráfico de propósito geral, ou GPGPUs. Antigamente as GPUs eram utilizadas apenas para processamento gráfico, já as GPUs mais recentes estão totalmente programáveis e estão ganhando cada vez mais espaço do ranking dos computadores mais potentes do mundo.

O nosso objetivo com esse trabalho é avaliar a eficiência energética e performance dos aceleradores GPGPU por meio de benchmarks. Nas seções seguintes serão apresentados os aceleradores, e os benchmarks utilizados.

2. Metodologia

Nesta seção será apresentada a metodologia utilizada nos testes, os equipamentos testados e os benchmarks utilizados. Primeiramente vamos ver alguns detalhes do hardware dos aceleradores GPGPUs.

2.1. GPGPUs

Uma grande diferença entre as CPUs e GPUs é que as CPUs se dedicam a quantidades de circuitos de controle e a GPU é mais focada as unidades aritméticas, se tornando assim mais eficientes no paralelismo. As GPU são construídas com mais ênfase ao throughput do que na latência, por essas razões ambos tem caminhado para direções diferentes (FERREIRA, 2012).

2.1.1. Nvidia

As GPUs Nvidia estão organizadas em vários SMs (Streaming Multiprocessors) e cada um executa um grupo de threads chamados de warps. Cada SM possui vários núcleos, chamados de CUDA

cores. Todo CUDA core possui pipelines completos de operações aritméticas e de ponto flutuantes. Cada SM possui memória cache L1 comum privada somente aos núcleos de um SM e todos os cores tem acesso a memória cache L2 (NVIDIA, 2014).
As primeiras placas programáveis em 2006 tinham 128 CUDA cores distribuídos entre 8 SMs. Atualmente as mais recentes placas, lançadas em 2012 com a arquitetura Kepler, os SMs que antes tinham 16 CUDA cores cada, passaram a se chamar SMX e ter 192 CUDA cores cada, com 8 SMX totalizam 1536 CUDA cores. As suas placas podem chegar a 2880 CUDA cores (NVIDIA, 2014).

2.1.2. Intel
Recentemente a Intel entrou no mercado dos aceleradores com os Coprocessadores Intel Xeon Phi. Aceleradores esses que podem conter até 61 núcleos de processamento, podendo utilizar 244 threads concorrentes (INTEL, 2012).
A arquitetura Many Integrated Core (MIC) combina vários núcleos de processadores Intel em um único chip. Podem utilizar linguagens de alto nível como C, C++ e FORTRAN, o mesmo código compilado em um acelerador Intel Xeon Phi pode ser executado da mesma forma em um processador Intel Xeon.

2.1.3. AMD
Os mais novos aceleradores da AMD estão utilizando a arquitetura Graphics Core Next (GCN) que no seu modelo mais completo tem um total de 32 Compute Units, que totalizam 2048 Stream Processors em uma única GPU. (WOLOGROSKI, WALLOSSEK, 2014).

2.2. Benchmarks
Benchmarks são ferramentas utilizadas para avaliação de recursos computacionais, trata-se de um conjunto de testes realizados no computador que levam em conta a capacidade do hardware ou de um software específico. Levam o hardware ao seu limite em busca de descobrir a sua capacidade de processamento.
No meio cientifico a performance do hardware geralmente é medida em flop/s, que seria o número de cálculos de ponto flutuante que consegue executar por segundo. No nosso trabalho serão abordados os benchmarks SHOC e Linpack. Nas seções abaixo serão caracterizados estes benchmarks.

2.2.1. Scalable HeterOgeneous Computing Benchmark Suite
O Scalable HeterOgeneous Computing Benchmark suite (SHOC) é uma coleção de benchmarks que avaliam a performance e a estabilidade de sistemas que utilizam dispositivos não convencionais para computação de propósito geral. Seu foco principal é em sistemas que utilizam unidades de processamento gráfico (GPUs) e processadores multi-core. Pode ser utilizado em cluster e hosts individuais (DANALIS et al, 2014).

Modalidade do trabalho: Relatório técnico-científico
Evento: XXII Seminário de Iniciação Científica

A linguagem padrão utilizada é o OpenCL porém, recentemente passou a utilizar o CUDA para comparações de performance. É compatível com as plataformas Linux e Mac OS X, além dos sistemas individuais o SHOC pode ser utilizado em clusters utilizando OpenMPI ou mpich2 (DANALIS et al, 2014).

2.2.2 Linpack

O benchmark Linpack, mundialmente conhecido por ser utilizado pela lista TOP500 para definir os computadores com maior performance no mundo. Ele é determinado para resolver um denso sistema de equações lineares. Foi originalmente desenvolvido para dar aos usuários um pacote que fosse possível medir quanto tempo demoraria o seu computador para resolver um certo problema com matriz (DONGARRAY, LUSZCZEKY e PETITETZ, 2014).

Possui duas rotinas de precisão dupla, sendo elas: DGEFA e DGESL, e duas rotinas de precisão simples: SGEFA e SGESL. DGEFA realiza a decomposição e DGESL usa essa decomposição para resolver a equação linear. A maior parte do tempo de execução é gasto no DGEFA. Seus resultados refletem apenas sobre um problema, que é a solução de densos sistemas de equações (DONGARRAY, LUSZCZEKY e PETITETZ, 2014).

3. Resultados e Discussão

O benchmark SHOC foi utilizado por apresentar um conjunto de testes bem abrangente e sendo compatível com todos aceleradores, Linpack foi utilizado para fins comparativos da performance dos aceleradores e um processador. Os aceleradores testados foram apenas da Nvidia, na tabela 1 pode ser visto detalhadamente as especificações dos aceleradores testados.

GPU	M2090	K10	K20x	K40
CUDA Cores	512	2x 1536	2688	2880
Memória (GB)	6	8	6	12
TDP (W)	225	225	235	235

Tabela 1 - Especificações GPUs. Fonte: Nvidia.

Na tabela 2 pode ser visto os resultados dos testes de MaxFlops do SHOC e o cálculo da eficiência energética levando em conta o TDP (Thermal Design Power) do acelerador contido na tabela 1. Para comparação o benchmark Linpack foi executado em um processador Intel Core I5-3210M de dois núcleos e TDP de 35w, utilizando números de precisão dupla apenas.

Modalidade do trabalho: Relatório técnico-científico
Evento: XXII Seminário de Iniciação Científica

GPU	M2090	K10	K20x	K40	I5-3210M
Precisão Simples (GFlops)	1301	3840	3457	3698	34
Precisão Dupla (GFlops)	660	180	1300	1406	
PS - GFlops/W	5,782222	17,06667	14,71064	15,73617	0,971429
PD - GFlops/W	2,933333	0,8	5,531915	5,982979	

Tabela 2 – Resultados MaxFlops SHOC. Fonte: Autoria própria.

4. Conclusões

Nos testes, os aceleradores demonstram ter uma potência superior ao processador, são aceleradores de grande desempenho comparando a um processador comum. Mas se comparado a eficiência energética os aceleradores se mostram ser uma ótima opção para obter-se grandes performances.
Este trabalho apresentou uma comparação na performance de alguns aceleradores gráficos da Nvidia, juntamente com sua eficiência energética. Em trabalhos futuros, pretende-se ampliar a quantidade de aceleradores testados, incluindo outros fabricantes, e novos benchmarks para comparação.

5. Palavras-chave

SHOC; Benchmark; Nvidia; Tesla

6. Agradecimentos

Agradecemos a Nvidia por disponibilizar acesso à seu cluster de testes, onde encontram-se esses aceleradores que estão sendo estudados e testados.

7. Referências Bibliográficas

DANALIS, Anthony; MARIN, Gabriel; MCCURDY, Collin; MEREDITH, Jeremy; ROTH Philip; SPAFFORD, Kyle; TIPPARAJU, Vinod; VETTER, Jeffrey. “The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite”, Knoxville, Tennessee, 2014.
DONGARRA, Jack “Frequently Asked Questions on the Linpack Benchmark and Top500”, Disponível em: <http://www.netlib.org/utk/people/JackDongarra/faq-linpack.html>. Acesso em 05 jun. 2014.
DONGARRAY, Jack; LUSZCZEKY, Piotr; and PETITETZ, Antoine. “The LINPACK Benchmark: Past, Present, and Future”, Dez, 2001.
FERREIRA, Anselmo; LEANDRO, Zanotto; MARCELO, Matsumoto. “Arquitetura e Programação de GPU Nvidia”, Campinas, São Paulo, 2012.
INTEL, “Intel® Xeon Phi™ Core Micro-architecture”. Hilssboro, Oregon, 2012.

NVIDIA "Tesla GPU Accelerators for Servers", Disponível em <http://www.nvidia.com/object/tesla-servers.html>. Acesso em: 06 jun. 2014.

WOLOGROSKI, Don; WALLOSSEK, Igor. "AMD Radeon HD 7970", Disponível em: <http://www.tomshardware.com/reviews/radeon-hd-7970-benchmark-tahiti-gcn,3104-2.html>. Acesso em: 06 jun. 2014.

YOUNGE, A.; von LASZEWSKI, G.; WANG, L.; LOPEZ-ALARCON, S.; and CARITHERS, W. "Efficient resource management for cloud computing environments.", In International Conference on Green Computing, pages 357–364. IEEE. Rochester, Nova Iorque. 2010.