Análise de desempenho e eficiência energética de aceleradores NVIDIA Kepler*

Emilio Hoffmann de O.[1], **Bruno M. Muenchen**[1], **Taís T. Siqueira**[1], **Edson L. Padoin**[1,2], **Philippe O. A. Navaux**[2]

[1]Universidade Reg. do Noroeste do Estado do Rio G. do Sul (UNIJUI) - Ijuí - RS - Brasil

emiliohoffmann@hotmail.com, {bruno.mokan,tais.thomas,padoin}@unijui.edu.br

[2]Universidade Federal do Rio Grande do Sul (UFRGS) - Porto Alegre - RS - Brasil

navaux@inf.ufrgs.br

Resumo. *Este artigo apresenta uma análise de desempenho e eficiência energética de aceleradores NVIDIA com o benchmark SHOC. O objetivo do estudo é analisar a utilização de aceleradores como alternativa ao aumento de desempenho e eficiência energética. Foram utilizadas 4 GPGPUs da arquitetura Kepler para a análise e realização dos testes, sendo observado nos testes um desempenho de até* 3706 *GFlops e uma eficiência energética de* 17 *GFlops/w.*

1. Introdução

No cenário tecnológico, fabricantes competiam entre si em busca de processadores cada vez mais velozes. Nesta competição, a frequência de relógio foi aumentando e chegando ao limite físico dos materiais empregados com altíssimo consumo de energia. Frente a isso, novas soluções começam a ser utilizadas para aliar poder computacional e economia de energia, como o surgimento do mecanismo de pipeline e a exploração do paralelismo, a qual demandou a necessidade de mudança dos códigos sequenciais existentes.

Na busca por um maior desempenho e um menor consumo de energia, a NVIDIA, já conhecida pela fabricação de placas de vídeo, desenvolveu as GPGPUs (*General Purpose Graphics Processing Units*), em meados de 2000. Inicialmente, as GPUs eram unidades com uma função fixa, construídas sobre um pipeline gráfico que se sobressaiam apenas em processamento gráfico. Desde sua criação, as empresas tem melhorado seu hardware com o foco também para o aspecto programável da GPU. O resultado desta criação é uma unidade de processamento com grande capacidade aritmética, não mais restrito à operações gráficas [Zanotto et al. 2012].

Assim, o objetivo deste trabalho é analisar o desempenho e a eficiência energética dos aceleradores GPGPU por meio do *benchmark* SHOC e analisar a sua possível utilização como uma alternativa para a computação de alto desempenho.

2. Trabalhos Relacionados

Existem vários trabalhos que abordam análise de desempenho de GPUs. Huang em [Huang et al. 2009] compara o desempenho entre CPUs e GPUs e conclui que sistemas heterogêneos obtêm um desempenho superior aos sistemas com CPUs. Muenchen em

*Esta pesquisa tem sido parcialmente financiada pela bolsa da CAPES sob processo número 3471-13-6.

[Muenchen et al. 2014] realiza comparação entre aceleradores abordando desempenho e consumo energético. Danalis em [Danalis 2010] apresenta um estudo do *benchmark* SHOC para a análise de desempenho de vários aceleradores e CPUs. Seus resultados mostram que o SHOC é uma ferramenta com grande variedade de testes para sistemas heterogêneos. Este trabalho, diferente dos demais aqui citados, apresenta um comparativo de desempenho entre os modelos mais recentes de aceleradores NVIDIA da família Kepler.

3. Metodologia

Os aceleradores usados são descritos na Seção 3.1. Para a avaliação de desempenho dos aceleradores selecionamos o *benchmark* SHOC descrito na Seção 3.2.

3.1. Ambiente de Testes

Os experimentos foram realizados no *NVIDIA Technology Center* também conhecido como PSG Cluster. Os aceleradores testados foram as GPUs NVIDIA Tesla K10m, K20m, K20Xm e K40m. Na Tabela 1 pode-se observar as especificações de cada acelerador, a TDP apresentada é a TDP configurada pelo servidor, limitando o consumo das GPUs. As informações mais importantes da arquitetura podem ser consultadas em [NVIDIA 2012].

Tabela 1. Especificações Técnicas dos Aceleradores

	K10m	K20m	K20Xm	K40m
CUDA Cores	3072	2496	2688	2880
Frequência de Clock (MHz)	745	706	732	745
Memória (GB)	8	5	6	12
Clock de Memória (MHz)	2500	2600	2600	3000
Interface PCI-E	3.0 x16	2.0 x16	2.0 x16	3.0 x16
TDP Configurada (W)	117,5	225	235	235

3.2. Benchmarks

O *benchmark* utilizado para mensuração do desempenho dos aceleradores foi o *Scalable Heterogeneous Computing Benchmark Suite* (SHOC). Este *benchmark* é composto por uma coleção de *benchmarks* que avaliam o desempenho e a estabilidade de sistemas que utilizam dispositivos não convencionais para computação. Seu foco é em sistemas que utilizam GPUs e processadores multi-core [Danalis 2010]. De acordo com o autor, o SHOC é dividido em três níveis:

- **Nível 0**: Os testes neste Nível medem as características de baixo nível do hardware. Todos os testes usam *kernels* artificiais, e os resultados representam um limite superior empírico das características testadas.
  - **MaxFlops** - Mede o desempenho máximo atingido com operações de ponto flutuante, tanto para precisão simples quanto dupla.
  - **BusSpeed** - Mede a banda de transferência de dados entre processador e acelerador com repetidas operações alterando o tamanho dos dados, de 1 KB a 64 MB. Inclui o tempo de transferência da interface de conexão.
  - **DeviceMemory** - Mede a banda de memória no acesso de diferentes níveis da hierarquia de memória e diferentes padrões de acesso. As medições não incluem o tempo de transferência da interface de conexão.

- **Nível 1**: Os testes neste Nível utilizam algoritmos básicos, como *Fast Fourrier Transform* (FFT), que mede o desempenho de FFT unidimensional. O *benchmark* calcula FFTs com 512 elementos complexos.
- **Nível 2**: Os testes neste Nível medem *kernels* de aplicações reais, como o S3D, que computa a taxa de várias reações químicas através de uma grade 3D.

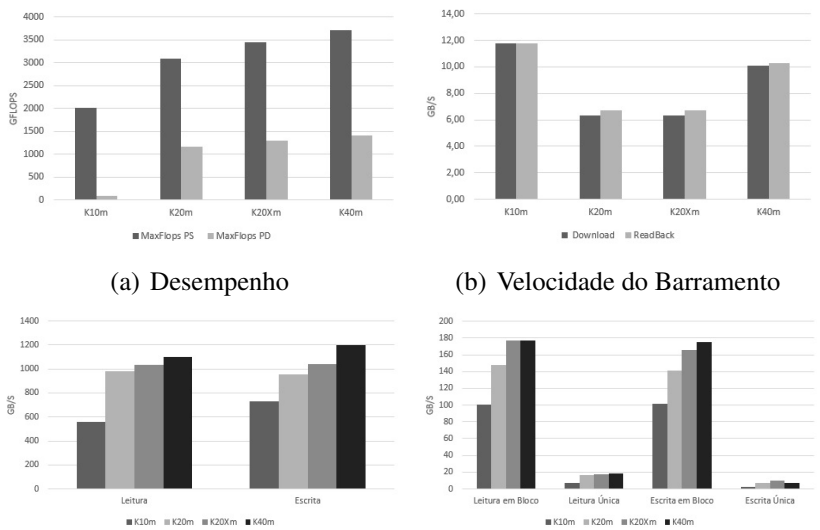
4. Resultados

Os resultados representam uma média de cinco execuções em cada acelerador, sendo que o *benchmark* executa cada aplicativo dez vezes e exibe o maior resultado de cada. Os dados serão apresentados da seguinte forma: Na Seção 4.1 são analisados os resultados do Nível 0. Na Seção 4.2, são analisados os resultados do Nível 1 e sua eficiência energética.

4.1. Nível 0

Os resultados apresentados na Figura 1(a) demonstram que, mesmo a GPU K10m tendo um maior número de núcleos que a K40m, com a mesma frequência, a K40m ainda foi superior em desempenho. Isso é justificado por sua construção ser a união de duas GPUs em uma só, criando um gargalo na comunicação.

Quanto a velocidade de barramento, que testa operações do processador para o acelerador (*Download*) e do acelerador para o processador (*ReadBack*), observada na Figura 1(b), ambas K10m e K40m tem melhor desempenho por possuírem interface de conexão PCI-E 3.0, ao invés de 2.0, como as demais.



(a) Desempenho (b) Velocidade do Barramento

(c) Memória Local (d) Memória Global

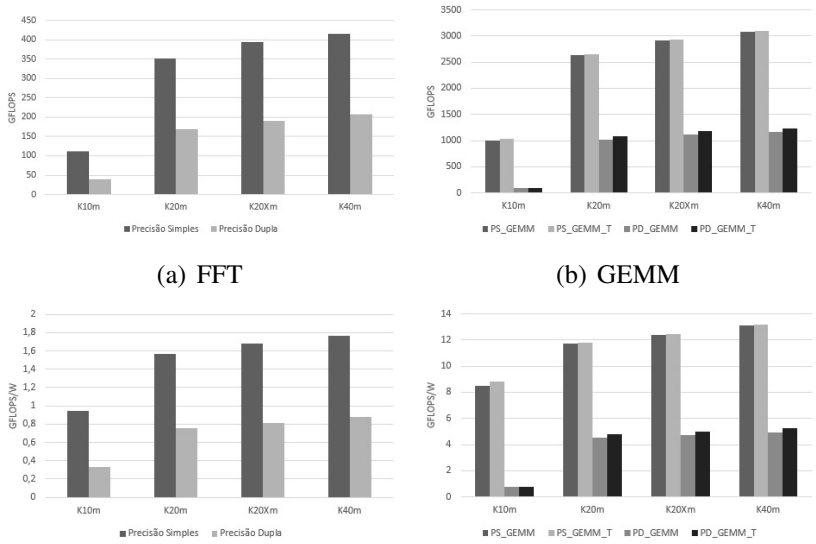
Figura 1. Resultados de Desempenho e Barramento

Nas Figuras 1(c) e 1(d) pode-se relacionar o aumento da taxa de transferência pelas diferenças do *clock* e interface de memória dos dispositivos, já que o teste não leva em consideração a velocidade da interface de conexão.

4.2. Nível 1

Nas Figuras 2(a) e 2(b), observa-se uma grande diferença de desempenho da K10m para as demais, o que pode ser resultado de uma gargalo na comunicação entre as duas GPUs

internas. Nas demais GPUs, observa-se um padrão de crescimento que pode ser relacionado com a quantidade de CUDA Cores, frequência destes e a velocidade da memória.



(a) FFT (b) GEMM

(c) Eficiência Energética - FFT (d) Eficiência Energética - GEMM

Figura 2. Resultados dos Testes com o Benchmark SHOC

Nas Figuras 2(c) e 2(d) pode ser visto o mesmo crescimento padrão apresentado nos resultados anteriores, podendo relacionar ao número de CUDA Cores, sua frequência e o desempenho da memória.

5. Conclusão e Trabalhos Futuros

Esse artigo apresentou uma comparação de desempenho e a eficiência energética entre GPUs NVIDIA de mesma arquitetura, voltadas para computação de alto desempenho. Nos algoritmos testados a NVIDIA k40m apresentou um melhor desempenho e eficiência energética, demonstrando um futuro promissor para a computação de alto desempenho. Como trabalhos futuros, pretende-se comparar os resultados aqui obtidos com outros aceleradores para *High Performance Computing* como, por exemplo, o Intel Xeon PHI.

Referências

Danalis, A. e. a. (2010). The scalable heterogeneous computing (SHOC) benchmark suite. In *Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units*, pages 63–74. ACM.

Huang, S., Xiao, S., and Feng, W.-c. (2009). On the energy efficiency of graphics processing units for scientific computing. In *Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*, pages 1–8. IEEE.

Muenchen, B. M., Padoin, E. L., and Navaux, P. O. A. (2014). Análise de desempenho de Aceleradores da AMD, NVIDIA e INTEL com a biblioteca OpenCL. *Erad 2014*.

NVIDIA (2012). Nvidia's Next Generation CUDA Compute Architecture: Kepler GK110.

Zanotto, L., Ferreira, A., and Matsumoto, M. (2012). Arquitetura e programação de gpu nvidia.