Eficiência energética x Desempenho entre um coprocessador Intel Xeon Phi e uma GPGPU NVIDIA *

Francisco Berti da Cruz[1], Emilio Hoffmann[2], Edson L. Padoin[1,3], Philippe O. A. Navaux[1], Jean-François Méhaut[3]

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

{francisco.cruz,emilio.hoffmann,padoin}@unijui.edu.br

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

navaux@inf.ufrgs.br

[3]Laboratoire d'Informatique de Grenoble (LIG), Université de Grenoble - France

Jean-Francois.Mehaut@imag.fr

Resumo. *A computação heterogênea está presente na grande maioria dos atuais sistemas Petaflops. Almejando a construção dos futuros sistemas Exaflops e atento às restrições de potência, aceleradores e coprocessadores tem sido utilizados como alternativas. Desse modo, este trabalho visa apresentar um comparativo de desempenho e eficiência energética entre um coprocessador Intel Xeon Phi e uma GPGPU da NVIDIA.*

1. Introdução

Nos últimos anos o esgotamento dos recursos naturais está cada vez mais evidente e vem sendo discutido em diversas áreas. Na comunidade de *High Performance Computing* (HPC), estudos vem sendo realizados para melhorar a eficiência energética, visando reduzir o tempo de execução das aplicações e a demanda de potência dos sistemas.

Considerando o histórico das últimas décadas, o aumento do desempenho foi obtido através do aumento na velocidade do *clock* dos processadores e com a criação da tecnologia *multi-core*. Entretanto, tais melhorias aumentaram a demanda de potência dos sistemas, chegando a limites inaceitáveis. Deste modo, a utilização de núcleos mais simples, agregados à uma coleção de núcleos mais sofisticados, seguindo o modelo de programação *Single Instruction, Multiple Data* (SIMD), passa a ser uma tendência para a construção destes sistemas [Borkar and Chien 2011].

Neste contexto, aceleradores gráficos e coprocessadores surgem como uma alternativa para aumento da eficiência energética dos sistemas. Assim sendo, a motivação deste trabalho é analisar o desempenho e a eficiência energética de um coprocessador Intel Xeon Phi e uma GPGPU da NVIDIA.

Na Seção 2 será abordado os trabalhos relacionados. Na Seção 3 serão apresentados as configurações dos equipamentos e o *benchmark* utilizado nos testes. Na Seção 4 apresenta-se os resultados alcançados e, por fim, na Seção 5, é exposto uma breve conclusão e possíveis trabalhos futuros.

*Trabalho parcialmente apoiado por CNPq, CAPES, FAPERGS e FINEP. Pesquisa realizada no contexto do Laboratório Internacional Associado LICIA e tem recebido recursos do programa EU H2020 e do MCTI/RNP-Brasil sob o projeto HPC4E de número 689772.

2. Trabalhos Relacionados

Muitos trabalhos tem sido realizados almejando analisar o desempenho entre aceleradores. Li, em [Li et al. 2014], apresenta resultados de desempenho e eficiência energética entre um processador convencional, um acelerador gráfico, e um Intel Xeon Phi.

Similar, Wang, em [Wang et al. 2013], compara um Xeon Phi e uma GPGPU NVIDIA utilizando diferentes *benchmarks*. Já Karl, em [Rupp 2016], apresenta uma análise considerando a largura de banda de memória de diferentes processadores e coprocessadores Intel de diferentes arquiteturas.

Diferente dos trabalhos mencionados, esta pesquisa visa analisar uma GPGPU da NVIDIA com a arquitetura Kepler, e um coprocessador Intel Xeon Phi da arquitetura KNC, por meio de diferentes implementações dos *benchmakrs* FFT e GEMM do SHOC *benchmark suite*.

3. Metodologia

Para realização dos testes foram utilizados 2 aceleradores, uma GPGPU NVIDIA da arquitetura Kepler, e um Intel Xeon Phi da arquitetura *Knights Corner* (KNC). O coprocessador Intel Xeon Phi é de modelo 3120A, que possui 57 *cores* de 1.10 GHz, instalado em um sistema com um processador Intel Xeon E5-2630 de 6 *cores*, de *clock* de 2.3 GHz. O sistema operacional instalado é o CentOS Linux versão 7.2.1511, com a versão do *kernel* 3.10.0-327.28.2.el7.x86_64. Para rodar os *benchmarks* do SHOC, estava configurado na máquina o compilador Intel ICC versão 16.0.3, que é compatível com o GCC 4.8.5. Utilizou-se também a biblioteca da Intel *Math Kernel Library* (MKL), que é necessária para rodar alguns *benchmarks* da suíte do SHOC OpenMP. Para rodar o *benchmark* em OpenCL foi utilizado o OpenCL versão intel-opencl-1.2-6.0.0.1049.

Quanto a GPGPU, os testes foram realizados em um servidor disponibilizado pela NVIDIA no NVIDIA Technology Center, também conhecido como PSG Cluster. O sistema operacional do servidor é o CentOS 6.4 OS. Para os testes foram utilizados os módulos do GCC 4.6.4, CUDA Toolkit 5.0 e a versão do driver utilizado foi a 331.62. Maiores detalhes dos 2 equipamentos são apresentados na Tabela 1.

Tabela 1. Especificações GPGPU NVIDIA Tesla k40m e Xeon Phi 3120A

Modelo	NVIDIA Tesla K40m	Xeon Phi 3120A
Quantidade de SMX	15	-
Número de *cores*	2880	57
Frequência dos *cores* (MHz)	745	1126
Número de *threads* por *core*	-	228
Interface de conexão	PCI-E 3.0 x16	PCI-E 2.0 x16
Thermal Design Power (TDP) (W)	235	300
Tamanho máximo da memória (GB)	12	6
Largura de banda máxima da memória (GB/s)	288	240

Para o estudo foi selecionando o *benchmark Scalable Heterogeneous Computing* (SHOC) devido a sua disponibilidade de diferentes implementações. Assim, na GPGPU foi usado a versão *CUDA* e no Xeon Phi as versões implementadas em *OpenMP* e *OpenCL* [SHOC 2012, Intel 2013].

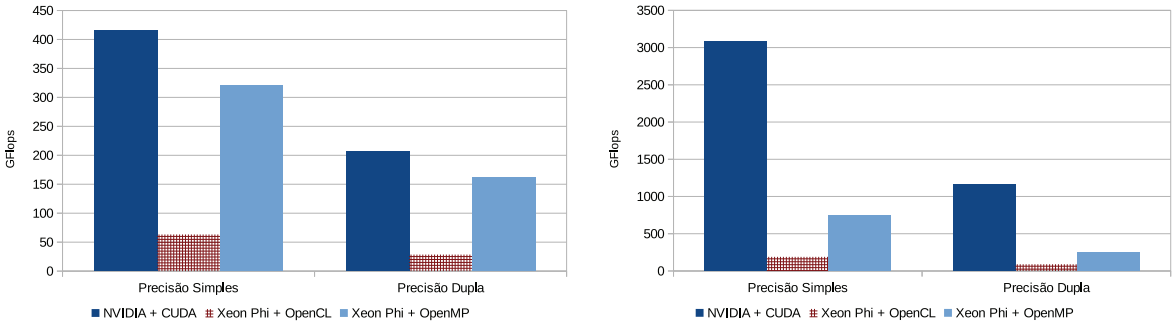
Os testes de desempenho deste *benchmark* são subdivididos considerando sua complexidade. Assim, selecionou-se do nível 1 os testes *Fast Fourier Transform* (FFT), que mede o desempenho da FFT para valores de precisão simples e dupla, e *General Matrix Multiply* (GEMM), que realiza a multiplicação em matrizes quadradas. Os resultados apresentados representam uma média de 5 execuções em cada acelerador, sendo que o *benchmark* executa cada teste 10 vezes.

4. Resultados

Observando os resultados do teste FFT (Figura 1(a)), percebe-se que o Xeon Phi + OpenCL alcançou $63,23$ GFlops para valores de precisão simples e $28,50$ GFlops para valores de precisão dupla. Por outro lado, o Xeon Phi + OpenMP alcançou $320,53$ GFlops para valores de precisão simples e $161,34$ GFlops para valores de precisão dupla. Ou seja, com a mudança de OpenCL para OpenMP sobre Xeon Phi tem-se uma diferença de $5,07$ vezes para valores com precisão simples e $5,66$ vezes para precisão dupla. Esta diferença é justificada devido a diferente forma de implementação dos algoritmos dos *benchmarks*.

Analisando os resultados do FFT alcançados na GPGPU da NVIDIA e no Xeon Phi + OpenMP, percebe-se que a primeira alcançou $415,21$ GFlops para valores de precisão simples, e $206,68$ GFlops para valores de precisão dupla. Ambos os resultados são mais altos se comparados com o Xeon Phi + OpenMP. Isto ocorre devido ao fato deste teste ser considerado mais complexo, onde o resultado depende de várias especificações dos aceleradores, como número de *cores*, a sua frequência, o barramento de memória.

Considerando os resultados obtidos no algoritmo GEMM, (Figura 1(b)), o acelerador da NVIDIA alcançou $3.097,74$ GFlops para valores de precisão simples, e $1.228,41$ GFlops para valores de precisão dupla. O Xeon Phi + OpenMP obteve $740,21$ GFlops para valores de precisão simples, e $248,80$ GFlops para valores de precisão dupla. Já o Xeon Phi + OpenCL, alcançou resultados de $150,25$ GFlops para valores de precisão simples, e $70,86$ GFlops para valores de precisão dupla. Percebe-se que o acelerador da NVIDIA obteve os resultados mais altos, acredita-se que isto ocorreu devido ao mesmo motivo dos testes realizados com o algoritmo FFT, visto na Figura 1(a). A implementação OpenCL no Xeon Phi novamente apresentou os resultados mais baixos.



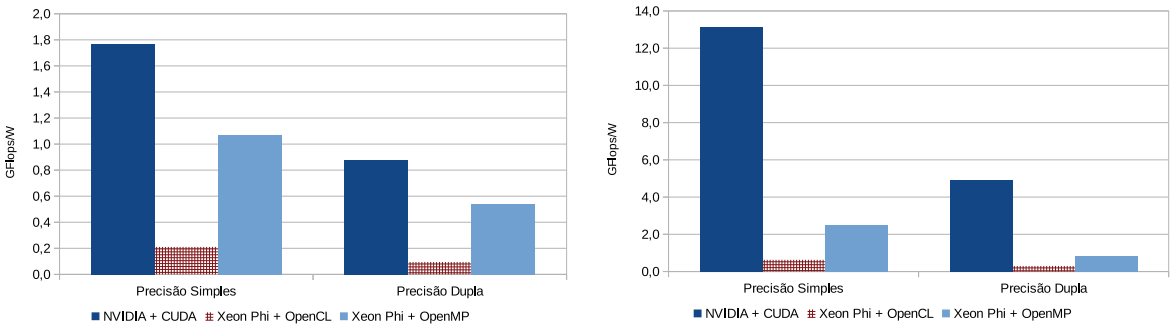
(a) FFT (b) GEMM

Figura 1. Resultados de Desempenho

Quanto a eficiência energética, foi considerado o TDP que consta nas especificações do fabricante de cada componente. No teste FFT, vide Figura 2(a), a GPGPU chegou a $1,77$ GFlops/W para valores de precisão simples, e $0,88$ GFlops/W

para valores de dupla. Enquanto o Xeon Phi + OpenCL alcançou $0,21$ GFlops/W para valores de precisão simples, e $0,10$ GFlops/W para valores de precisão dupla, e o Xeon Phi + OpenMP chegou a $1,07$ GFlops/W para valores de precisão simples, e $0,54$ GFlops/W para valores de precisão dupla. Justifica-se os resultados mais altos na GPGPU o fato dela possuir uma TDP mais baixa, 235 W contra 300 W do coprocessador.

Considerando os testes no Xeon Phi + OpenCL, eles mostraram-se menos eficientes que as outras duas opções no FFT e no GEMM, isto ocorreu devido ao TDP mais alto se comparado com o GPGPU, e também por ter obtido os resultados mais baixos nos testes de desempenho, que podem ser vistos na Figura 1(a) e na Figura 1(b).



(a) FFT (b) GEMM

Figura 2. Resultados de Eficiência Energética

5. Conclusão e Trabalhos Futuros

Almejando o uso de aceleradores gráficos e coprocessadores na computação de alto desempenho, este artigo apresenta uma analise de desempenho e eficiência energética de um coprocessador Intel Xeon Phi e uma GPGPU da NVIDIA. Para trabalhos futuros, sugere-se a realização de testes considerando a implementação do SHOC utilizando apenas sua versão OpenCL.

Referências

Borkar, S. and Chien, A. A. (2011). The future of microprocessors. *Commun. ACM*, 54(5):67–77.

Intel (2013). *The Scalable Heterogeneous Computing Benchmark Suite (SHOC) for Intel Xeon Phi*.

Li, B., Chang, H.-C., Song, S., Su, C.-Y., Meyer, T., Mooring, J., and Cameron, K. W. (2014). The power-performance tradeoffs of the intel xeon phi on hpc applications. In *Parallel & Distributed Processing Symposium Workshops (IPDPSW), 2014 IEEE International*, pages 1448–1456. IEEE.

Rupp, K. (2016). Knights landing vs. knights corner, haswell, ivy bridge, and sandy bridge: Stream benchmark results.

SHOC (2012). *The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite*.

Wang, Y., Qin, Q., SEE, S. C. W., and Lin, J. (2013). Performance portability evaluation for openacc on intel knights corner and nvidia kepler. *HPC China*.