Análise de desempenho e eficiência energética de aceleradores NVIDIA Kepler

Emilio Hoffmann, Bruno M. Muenchen, Taís T. Siqueira, Edson L. Padoin e Philippe O. A. Navaux

Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUI)
Universidade Federal do Rio Grande do Sul (UFRGS)

ERAD 2015 – Gramado, RS

Ciência da Computação UNIJUÍ

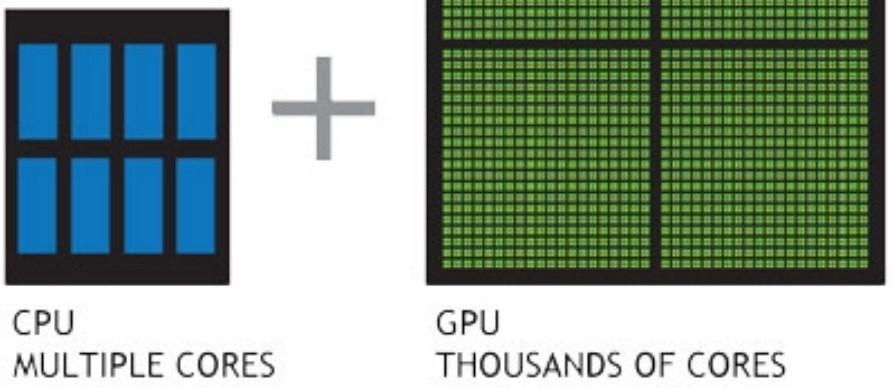
UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

UNIJUÍ UNIVERSIDADE REGIONAL

Sumário

- Introdução
- Arquitetura das GPUs
  - NVIDIA
- Ambiente de testes
- Resultados
- Conclusão
- Trabalhos futuros
- Bibliografia

CPU vs GPU GPGPUs

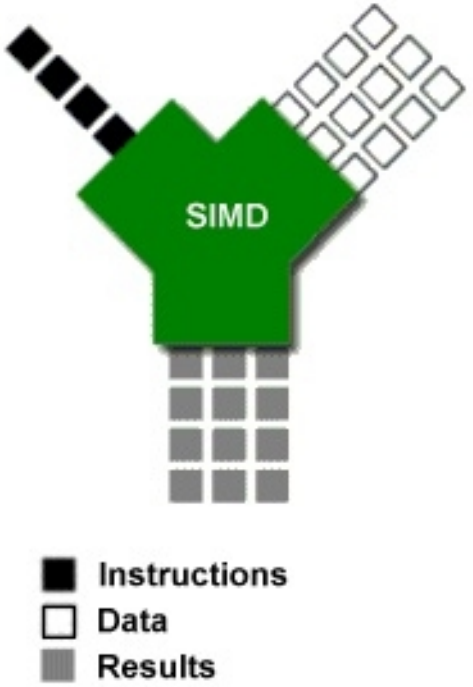


- Em 2000, NVIDIA transforma GPUs em GPGPUs
  - GPUs (Graphics Processing Units) com função fixa, processamento gráfico
  - GPGPUs (General Purpouse Graphics Processing Units) passam a ser programáveis

Fonte: NVIDIA

Arquitetura de uma GPU

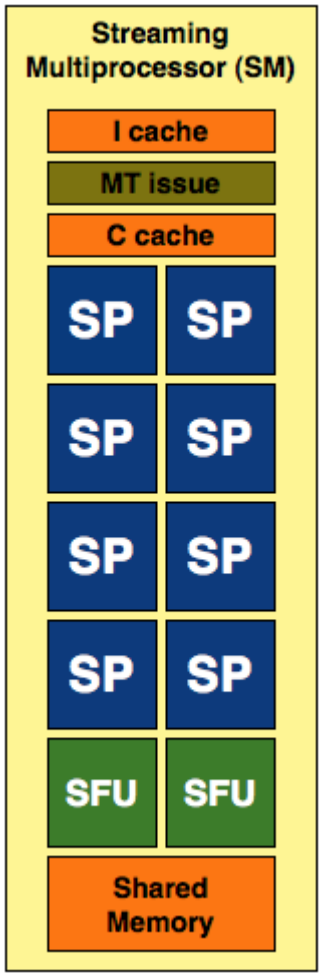
- Classificadas como SIMD (Single Instruction Multiple Data)
- SIMT (Single Instruction Multiple Threads)
  - Execução de instruções por grupo de Threads



Fonte: NVIDIA

Arquitetura das GPUs NVIDIA

- SP (Streaming Processor)
  - Similar ao núcleo de uma CPU
  - Possui Pipeline completo
  - Não tem memória cache
- SM (Streaming Multiprocessor)
  - Vetor de SPs
  - Memórias Cache
  - Escalonadores
  - SFU (Special Function Units)
- Quantidade de SMs e SPs



Fonte: NVIDIA

Arquitetura NVIDIA Kepler

- SMX, Nova versão dos SMs
  - 192 CUDA Cores de precisão simples – SPs
  - GK110 + 64 Unidades de precisão dupla;
  - 32 SFUs
  - 4 Escalonadores – 8 Despachantes de instruções
    - 2 Instruções por Warp (Grupo de Threads)

Ambiente de Testes

- Hardware
  - NVIDIA Technology Center – PSG Cluster
    - 30 nós com GPUs
      - Dual sockets
        - Ivy Bridge (10 cores)
        - Sandy Bridge (6 ou 8 cores)

- Software
  - Sistema operacional CentOS 6.4 OS
  - gcc/4.6.4
  - cuda/5.0/toolkit

- Conexão
  - VPN
  - SSH

Aceleradores Utilizados

	K10m.g1	K20m	K20Xm	K40m
Memória GDDR5 (GB):	4	5	6	12
Interface memoria (Bits)	256	320	384	384
Clock da Memória (MHz):	2500	2600	2600	3004
CUDA Cores:	1536	2496	2688	2880
CLOCK SMs (MHz):	745	706	732	875
Consumo máximo (W):	117,5	225	235	235

Fonte: NVIDIA

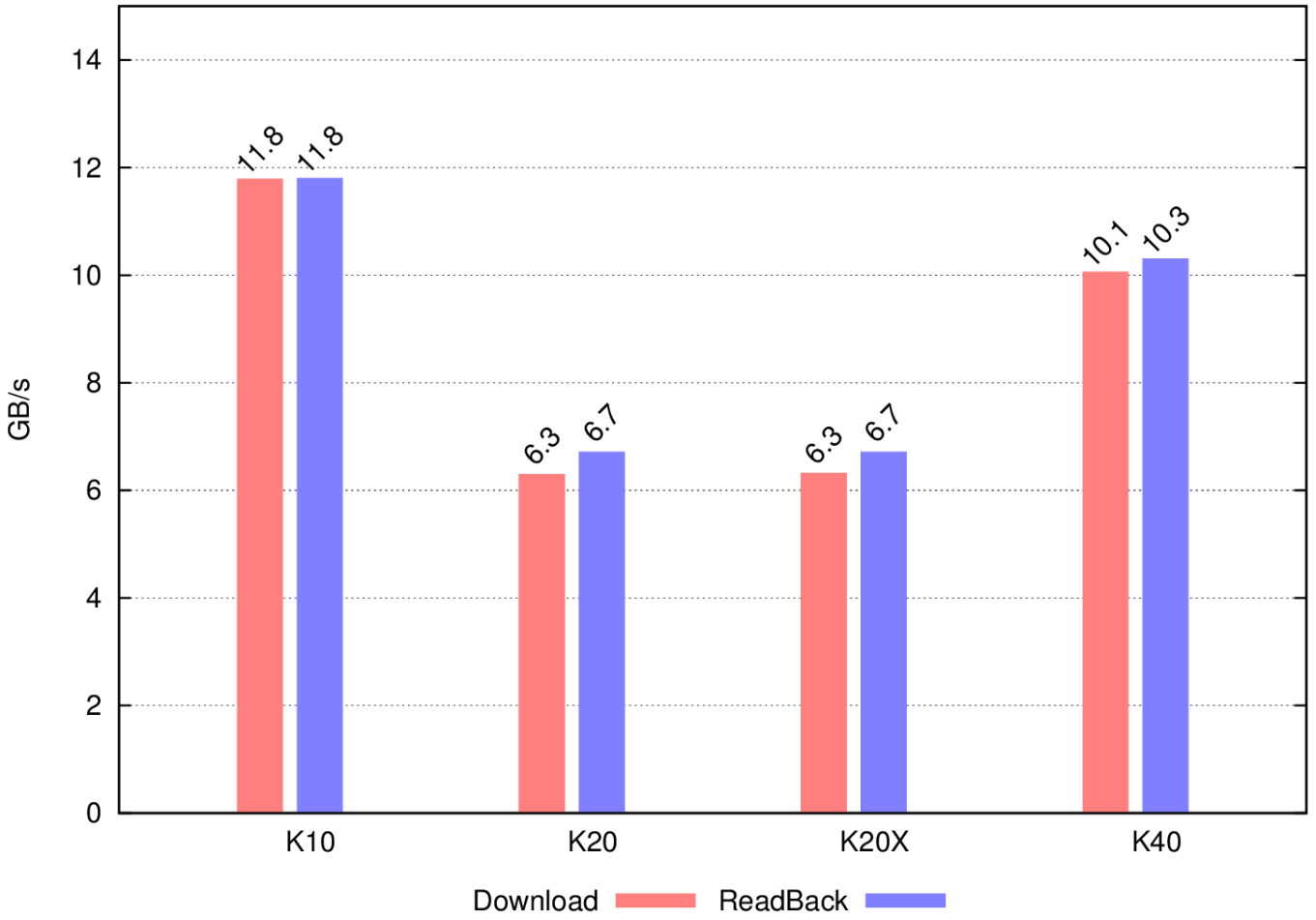
Benchmark SHOC

- **S**calable **H**eter**O**geneous **C**omputing Benchmark suite (SHOC)
  - Dividido por categorias
    - **Categoria 0** – Medições de nível baixo – velocidades, alimentação e memória
      - BusSpeed
      - DeviceMemory
    - **Categoria 1** – Medições de performance com nível mais alto
      - FFT
      - GEMM
    - **Categoria 2** – Medições com aplicações reais

Fonte: DANALIS

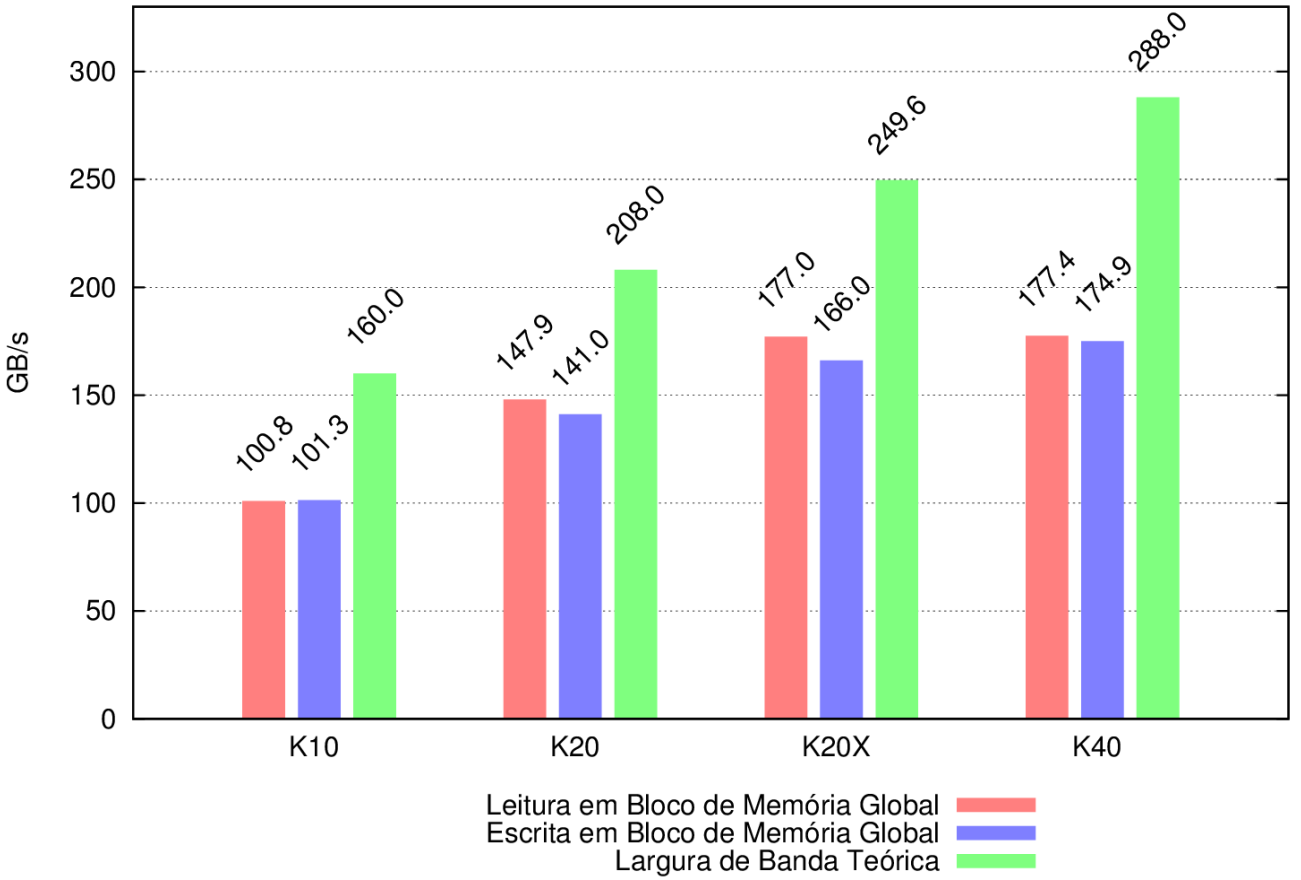
Benchmark: BusSpeed
Velocidade do Barramento

- K10 e K40 Possuem PCI-E 3.0
- Resultando em quase o dobro de velocidade



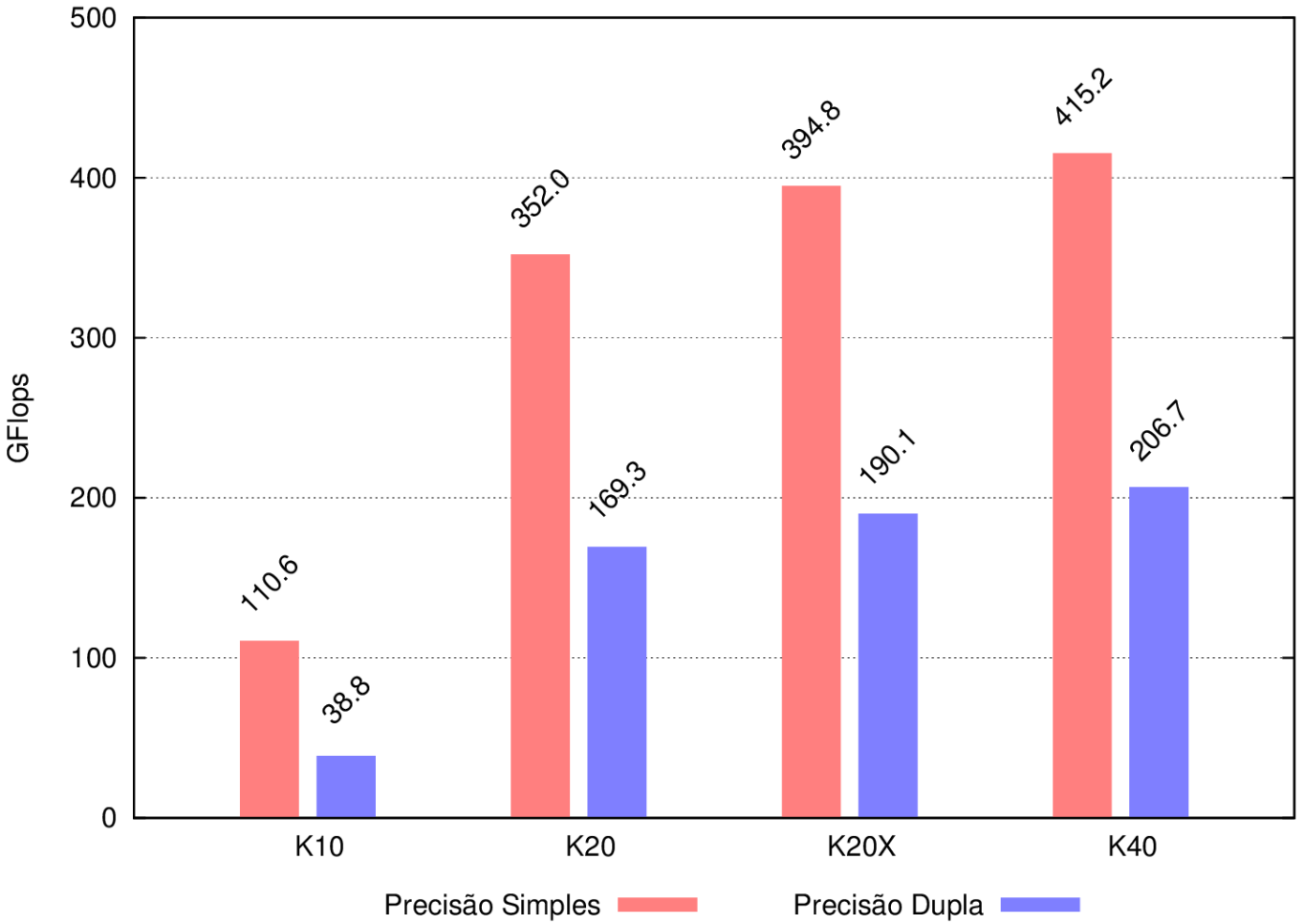
Benchmark: DeviceMemory
Velocidade de Memória

- Teórica = Interface * Frequência
- Com o SHOC obtêm-se entre **60%** e **70%** da largura teórica



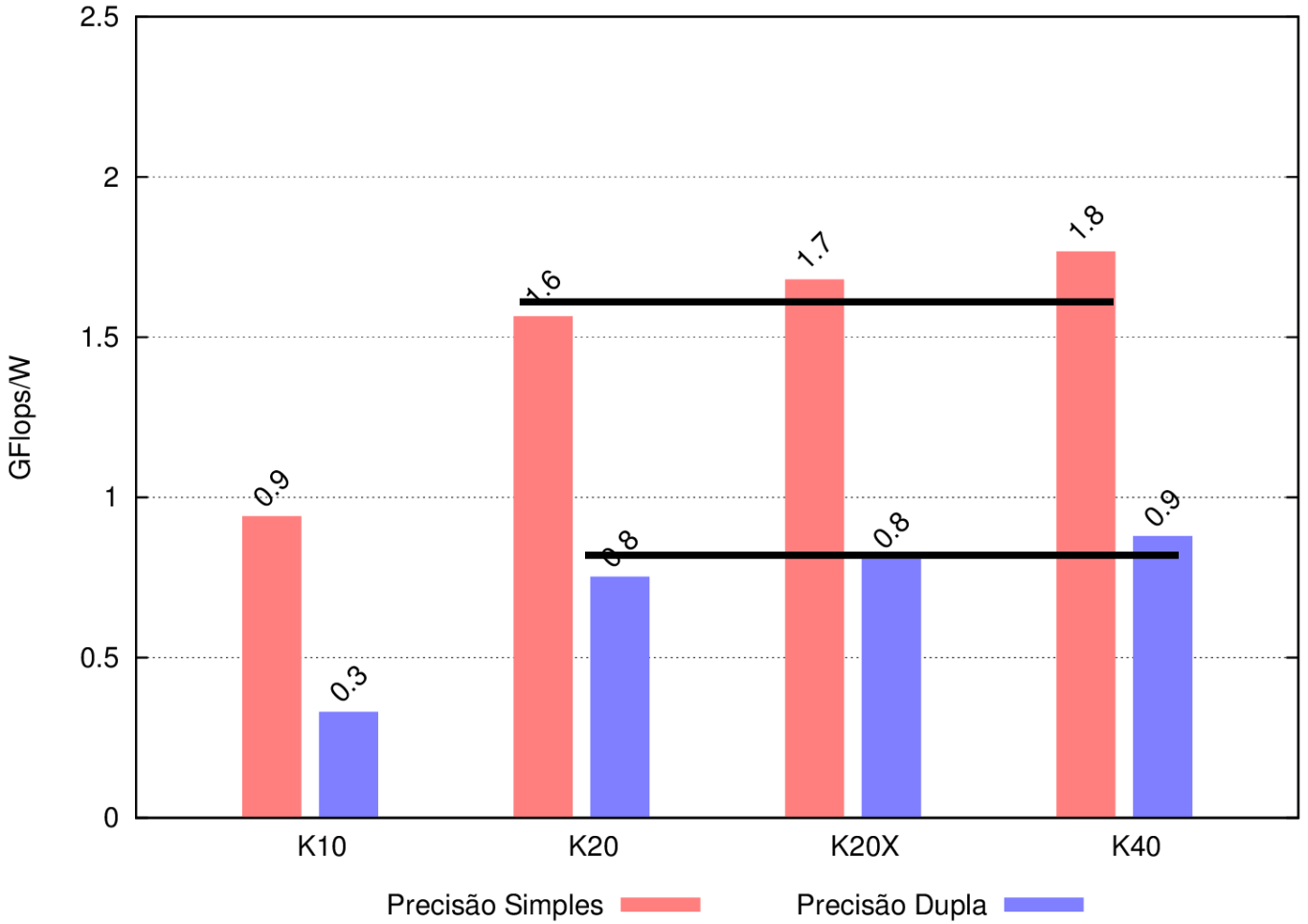
Benchmark: FFT Performance

- K10 >70 MFlops/CUDA Core
- K20, K20x e K40 > 140 MFlops/CUDA Core
- Interface de memória



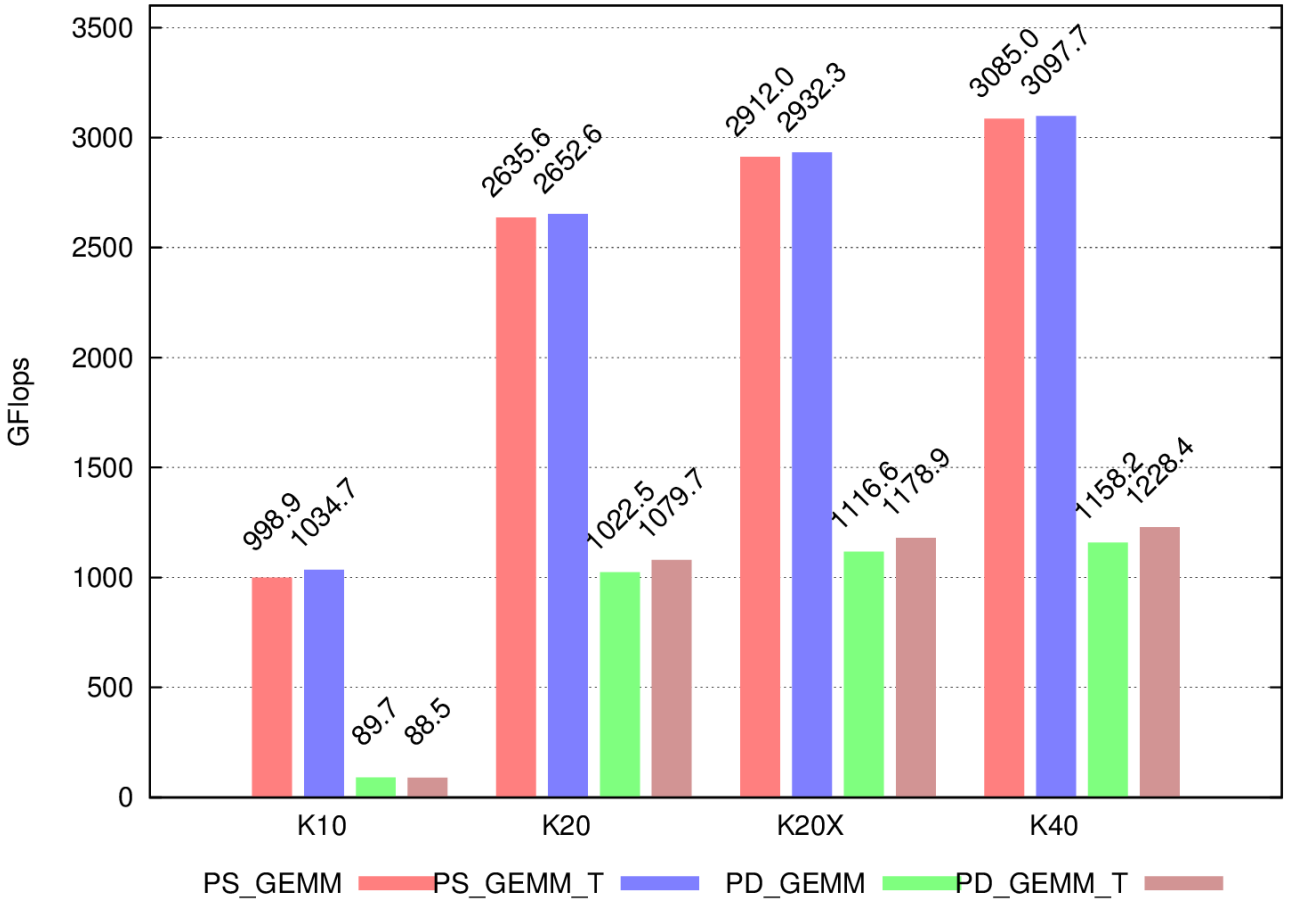
Benchmark: FFT Eficiência energética

- K10 - baixa performance = baixa eficiência
- Outras tem performance e potencia parecidas



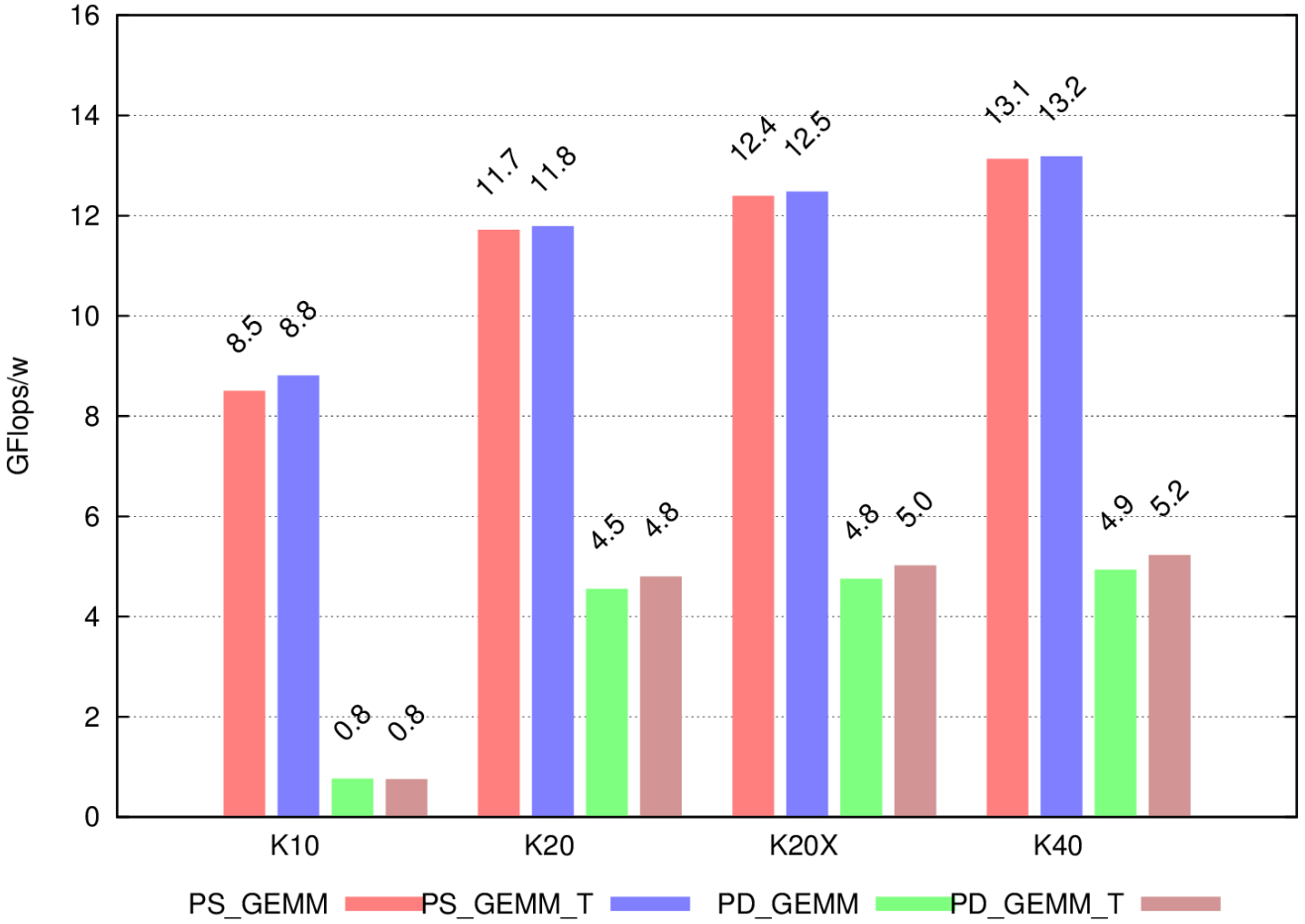
Benchmark: GEMM Performance

- GK 104 na k10 – baixa performance precisão dupla
- Resultado semelhante ao FFT



Benchmark: GEMM Eficiência Energética

- GK 104 na k10 – baixa performance precisão dupla
- Resultado semelhante ao FFT

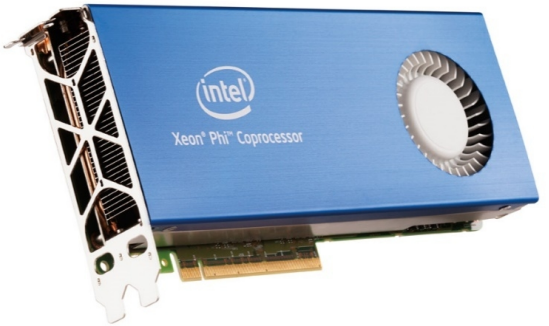


Conclusão

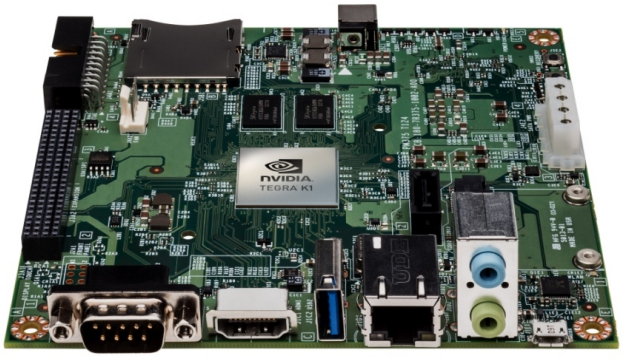
- Utilizado o benchmark SHOC para avaliar quatro aceleradores NVIDIA
  - K10, K20, K20x e K40
- Performance:
  - 3 TFlops com precisão simples
  - 1,2 TFflops com precisão dupla.
- Eficiência energética:
  - 13,2 GFlops/w com precisão simples
  - 5,2 GFlops/w com precisão dupla
- Aplicações científicas utilizam dados com precisão dupla

Trabalhos Futuros

- Ampliar aceleradores
  - Intel Xeon Phi
- Submissão de um artigo para WSCAD 2015
  - Jetson TK1
- Benchmarks da categoria 2

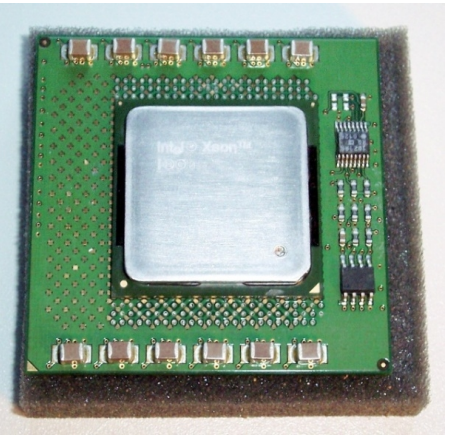


Fonte: Intel



Fonte: NVIDIA

Fonte: AMD



Fonte: Intel

Bibliografia

- AHMED, M. F.; HARIDY, O. A comparative benchmarking of the FFT on Fermi and Evergreen GPUs. In: IEEE. Performance Analysis of Systems and Software (ISPASS), 2011 IEEE International Symposium. [S.l.], 2011. p. 127–128.
- BELL, N.; GARLAND, M. Efficient Sparse Matrix-Vector Multiplication on CUDA . Dezembro 2008.
- BELL, N.; GARLAND, M. Análise de Desempenho da Arquitetura CUDA Utilizando os NAS Parallel Benchmarks. Dezembro 2009.
- DANALIS, A. et al. The scalable heterogeneous computing (SHOC) benchmark suite. In: Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units. [S.l.: s.n.], 2010. p. 63–74.
- DETOMINI, R. C. Exploração de Paralelismo em Criptografia Utilizando GPUs. Tese (Doutorado) — Universidade Estadual Paulista, 2010.
- FUJIMOTO, N. Dense matrix-vector multiplication on the CUDA architecture. Parallel Processing Letters, World Scientific, v. 18, n. 04, p. 511–530, 2008.
- GARLAND M. LE GRAND, S. . N. J. . A. J. . H. J. . M. S. . P. E. . Y. Z. . V. V. Parallel Computing Experiences with CUDA. 2008.
- GREEN 500. Acesso em: 03.nov.2014. Disponível em: <http://www.green500.org>.
- HEMSOTH, N. Mission Possible — Greening the HPC Data Center. 2009. Acesso em: 22.out.2014. Disponível em: <https://computing.llnl.gov/tutorials/pthreads>.
- KIRK, D. B.; WEN-MEI, W. H. Programming massively parallel processors: a hands-on approach. [S.l.]: Newnes, 2012.
- LINDHOLM JOHN NICKOLLS, S. F. O. J. M. E. Tesla: A Unified Graphics and Computing Architecture. IEEE Micro, p. 39–55, Junho 2008.

Bibliografia

- MUENCHEN, B. M. GPGPU: comparação de aceleradores AMD, NVIDIA e INTEL utilizando a biblioteca OPENCL. 2013.
- NVIDIA. NVIDIA Tesla C870 GPU Computing Processor Board. Abril 2008.
- NVIDIA. Tesla C1060 Computing Processor Board . Setembro 2008.
- NVIDIA. NVIDIA's Next Generation CUDA Compute Architecture: FERMI. 2009.
- NVIDIA. Tesla M2050 and Tesla M2070/M2070Q Dual-Slot Computing Processor Modules. Agosto 2010.
- NVIDIA. TESLA C2050 AND TESLA C2070 COMPUTING PROCESSOR BOARD. 2011.
- NVIDIA. Tesla C2075 Computing Processor Board. Setembro 2011.
- NVIDIA. Tesla M2090 Dual-Slot Computing Processor Module. Junho 2011.
- NVIDIA. NVIDIA's Next Generation CUDA Compute Architecture: KEPLER. 2012.
- NVIDIA. Tesla K10 GPU Accelerator. Junho 2012.
- NVIDIA. Tesla K20X GPU Accelerator. Novembro 2012.
- NVIDIA. CUDA C Programming Guide v5.5. Julho 2013.
- NVIDIA. O que é computação com GPU? 2013. Acesso em: 01.nov.2013. Disponível em: <http://www.nvidia.com.br/object/what-is-gpu-computing-br.html>.
- NVIDIA. Tesla K20 GPU Accelerator. Julho 2013.

Bibliografia

- NVIDIA. Tesla K40 GPU Accelerator. Novembro 2013.
- NVIDIA. CUDA C Programming Guide v6.5. Agosto 2014.
- NVIDIA. CUDA GPUs. 2014. Acesso em: 18.out.2014. Disponível em: <https://developer.nvidia.com/cuda-gpus>.
- OHSHIMA, S. et al. Parallel processing of matrix multiplication in a CPU and GPU heterogeneous environment. In: High Performance Computing for Computational Science-VECPAR 2006. [S.l.]: Springer, 2007. p. 305–318.
- ROSE, C. A. F. D. Caderno dos Cursos Permanentes. 2006.
- SHOC. 2012. Acesso em: 21.out.2014. Disponível em: <https://github.com/vetter/shoc>.
- SPAFFORD, K. et al. Accelerating S3D: a GPGPU case study. In: SPRINGER. Euro-Par 2009–Parallel Processing Workshops. [S.l.], 2010. p. 122–131.
- STRATTON, J.; STONE, S.; HWU, W. mei. Mcuda: An efficient implementation of CUDA kernels on multi-cores. [S.l.], 2008.
- TEAM, D. SHOC: The Scalable HeterOgeneous Computing Benchmark Suite. 2011.
- YANO, L. Avaliação e comparação de desempenho utilizando CUDA. Junho 2010.
- ZANOTTO, L.; FERREIRA, A.; MATSUMOTO, M. Arquitetura e Programação de GPU Nvidia. 2012.
- ZHANG, Y. Performance and Power Comparisons Between Fermi and Cypress GPUS. Dezembro 2013.

Análise de desempenho e eficiência energética de aceleradores NVIDIA Kepler

Obrigado!

Emilio Hoffmann, Bruno M. Muenchen, Taís T. Siqueira, Edson L. Padoin e Philippe O. A. Navaux

Universidade Regional do Noroeste do Estado do Rio Grande do Sul (UNIJUI)
Universidade Federal do Rio Grande do Sul (UFRGS)

ERAD 2015 – Gramado, RS

Ciência da Computação UNIJUÍ

UFRGS UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

UNIJUÍ UNIVERSIDADE REGIONAL