Estimativas de desempenho, consumo e *dark-silicon* em projetos GPGPU

Gislayne G. Damasceno[1], Laura B. Ludgero[1], Samuel S. Rodrigues[1], Ricardo R. dos Santos[1], Liana D. Duenha[1]

[1]Faculdade de Computação – Universidade Federal do Mato Grosso do Sul (UFMS)

{gislayne_garabini, laura_ludgero, samuel_rodrigues ricardo.santos, liana.duenha}@ufms.br

Abstract. *This paper presents an extension of the MultiExplorer tool for the analysis and exploration of the GPGPU designs. New simulators were integrated to characterize the performance, area, and power consumption of six GPU models running CUDA and Rodinia applications. Additionally, a core database was modeled to explore different scenarios.*

Resumo. *Este artigo apresenta a ampliação da ferramenta MultiExplorer para análise e exploração de GPUs. Foram integrados novos simuladores para a caracterização de desempenho, área e consumo de seis GPUs executando aplicações CUDA e Rodinia. Adicionalmente, um banco de núcleos foi modelado para explorar diferentes cenários.*

1. Introdução

A busca incessante por processadores cada vez mais rápidos, impulsionada pela Lei de Moore [MOORE 1998], está atingindo seus limites, resultando em um consumo de energia cada vez maior, conhecido como *dark-silicon*, onde parte dos transistores de um chip permanece inativo devido as limitações computacionais. Para contornar essas retrições, surgiram as arquiteturas heterogêneas e aceleradores, como as Unidades de Processamento Gráfico (GPUs).

A extensão da ferramenta Multiexplorer [SANTOS et al. 2018] para o domínio GPGPU tem como o objetivo mitigar o *dark-silicon* em placas gráficas por meio de GPUs heterogêneas. Neste estudo, foram caracterizados parâmetros de desempenho, estimativas físicas e de consumo de seis GPUs comerciais da NVIDIA, em diferentes configurações e litografias, também desenvolve um banco de núcleos de GPU para possibilitar a exploração arquitetural.

2. Multiexplorer

O MultiExplorer [Devigo et al. 2015] é uma ferramenta de exploração de projetos de sistemas MPSoCs. Elaborado para avaliar o desempenho de componentes da plataforma, estimando área, consumo e o *dark-silicon* de arquiteturas *multicore*, permitindo a exploração de alternativas heterogêneas, maximizando o desempenho com redução mínima. Os dados de desempenho são repassados para um estimador físico, que inclui a estimativa de *dark-silicon*, possibilitando a busca por opções livres dessa condição.

Com essa extensão, o MultiExplorer pode simular a execução de uma aplicação em uma GPU, estimando área e consumo. Em caso de *dark-silicon*, o usuário pode explorar arquiteturas heterogêneas mesclando núcleos diferentes em busca de melhor desempenho. No contexto deste artigo, um *core* de GPU é definido como um *streaming multiprocessor* (SM).

3. Caracterização das Placas, o Banco de Núcleos e a Evolução tecnológica

Nove aplicações dos pacotes CUDA *Library* e Rodinia foram validados, bem como seis placas NVIDIA, que cobrem cinco diferentes arquiteturas. O GPGPU-Sim foi integrado ao MultiExplorer para a simulação das GPUs, permitindo a análise do impacto de diferentes parâmetros arquiteturais no desempenho e otimização das configurações. A ferramenta McPAT, adaptada como GPUWattch, estima desempenho, consumo e *dark-silicon*.

A Tabela 1 apresenta a caracterização das placas, bem como os resultados da simulação de desempenho, estimativas físicas e as características das placas. Os dados de desempenho foram obtidos a partir da execução das nove aplicações nas seis placas, com apenas 1 SM por placa. Utilizando duas métricas de desempenho: IPC (instruções por ciclo), e MIPS (milhões de instruções por segundo).

Características do Modelo da Placa						
Placa	QV100	TITANV	RTX2060	TITANX	GK110	GTX480
Arquitetura	Volta	Volta	Turing	Pascal	Kepler	Fermi
Tecnologia (nm)	22	22	22	32	32	45
Frequência (Mhz)	1200	1200	1365	1000	800	700
Qte de SMs da placa	80	80	30	24	14	15
Blocos por SM	32	32	32	32	32	8
Threads por SM	2048	2048	2048	1536	2048	1920
Regs. por core	65.536	32.768	65.536	65.536	65.536	16.384
Estimativas de Desempenho executando as nove aplicações com 1 SM por placa						
Placa	QV100	TITANV	RTX2060	TITANX	GK100	GTX480
Instruções	1.284M	1.284M	1.284M	1.284M	1.284M	1.284M
Ciclos	118.294 M	57.116 M	144.428 M	107.197 M	94.638 M	213.835 M
Instr. por ciclo (IPC)	10,86	22,48	8,89	11,99	13,57	6,00
Runtime (ms)	99	48	106	107	118	305
D_*placa* (MIPS)	13.026	26.978	12.136	11.978	10.855	4.203
Estimativas Físicas das Placas Originais						
Área (mm^2)	1234,34	1219,42	536,94	809,02	404,51	826,27
Potência (W)	212,68	240,14	112,24	153,02	70,15	79,20
Dens. pot.(W/mm^2)	0,17	0,19	0,21	0,19	0,17	0,10
Estimativas Físicas de 1 SM por placa						
Área (mm^2)	15,43	15,24	17,90	30,45	31,07	55,08
Potência (W)	2,66	3,00	3,74	5,66	5,71	5,32
Dens. pot.(W/mm^2)	0,17	0,20	0,21	0,19	0,18	0,10

Tabela 1. Caracterização de desempenho e estimativas físicas com 1 SM.

As Equações 1 e 2 descrevem o cálculo de desempenho das placas baseado na métrica MIPS [Sonohata et al. 2023].

$$D\_app_i = IPC_{app_i} \times freq \tag{1}$$

$$D\_placa = \sum_i (D_{app_i} \times peso_{app_i}) \tag{2}$$

O consumo em W é a soma de três parâmetros de potência (*Peak Dynamic*, *Subthreshold Leakage* e *Gate Leakage*), dividido pelo total de SMs, resultando no consumo por SM. A área de uma SM é utilizada para cada litografia disponível, com as informações provenientes do arquivo de saída gerado pelo McPAT no GPGPU-sim.

Realizou-se experimentos em seis modelos de GPUs descritas na Tabela 1, evoluindo o projeto original para tecnologias mais recentes até $22nm$ (limite máximo suportado pela ferramenta McPAT) onde se escala os números de SMs, mantendo a área do chip original, a mesma frequência de operação e tensão constantes.

Com isso, realizou-se a evolução tecnológica da placa *GTX480* a partir dos cálculos descritos em [Sonohata et al. 2023] e assim obtendo a % de *dark-silicon*.

GTX480 (Fermi)							
Número de SMs	15	28	60	Densidade de Pot. *(W/mm2)*	0,097	0,162	0,169
Tecnologia *(nm)*	45	32	22	Área do circuito de ref. *(mm^2)*	55,08	29,51	13,77
Frequência *(MHz)*	700	700	700	Pot. do circuito de ref. *(W)*	5,32	4,78	2,33
Área/chip *(mm^2)*	826,27	826,27	826,27	*Dark-silicon (%)*	-	40,12%	42,55%
Potência *(W)*	79,80	133,94	139,53				

Tabela 2. Evolução tecnológica da GTX480.

4. Conclusão

Este trabalho estendeu a ferramenta MultiExplorer para explorar sistemas GPGPU, integrando o simulador GPGPU-Sim, ajustando a interface gráfica e adicionando suporte para estimar dark-silicon em GPUs. Seis GPUs de cinco arquiteturas diferentes foram modeladas e nove aplicações CUDA e Rodinia foram integradas à ferramenta. Simulações e estimativas físicas foram realizadas e a estimativa de *dark-silicon* foi implementada. No entanto, a validação do fluxo de exploração do espaço de projetos para GPUs ainda está em desenvolvimento, bem como a integração com IoTs e computação aproximada.

Referências

Devigo, R., Duenha, L., Azevedo, R., and Santos, R. (2015). Multiexplorer: A tool set for multicore system-on-chip design exploration. In *2015 IEEE 26th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, pages 160–161.

MOORE, G. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85.

SANTOS, R., DUENHA, L., SILVA, A., SOUSA, M., TEDESCO, L., MELGAREJO, J., SANTOS, T., AZEVEDO, R., and MORENO, E. (2018). Dark-silicon aware design space exploration. *Journal of Parallel and Distributed Computing*, 120:295–306.

Sonohata, R., Arigoni, D. C. A., Fernandes, E. R., Ribeiro dos Santos, R., and Dessandre Duenha, L. (2023). Performance predictors for graphics processing units applied to dark-silicon-aware design space exploration. *Concurrency and Computation: Practice and Experience*, 35(17):e6877.