Exploração do Projeto de Sistemas Baseados em GPU ciente de *Dark Silicon*

Rhayssa Sonohata, Casio Krebs, Daniela Catelan, Liana Duenha, Diego Segovia, Mateus Tostes, Ricardo Santos

[1]Faculdade de Computação (FACOM)
Universidade Federal de Mato Grosso do Sul (UFMS)
Campo Grande – MS – Brasil

{rhayssa.sonohata,casiokrebs,garcia.segovia171, mateustostesdossantos}@gmail.com

{daniela, lianaduenha, ricardo}@facom.ufms.br

Abstract. *This paper proposes an infrastructure for the design space exploration of platforms with both graphics processing units and general-purpose cores. The goal is to mitigate the dark silicon and increase system performance at design time. The GPGPUSim simulator has been extended to perform dark silicon estimates of GPU platforms and then integrated into the MultiExplorer framework. Additionally, we propose a strategy to estimate the performance of GPU platforms, and we also model a database that uses both GPU and general-purpose cores, thus enabling the design space exploration for heterogeneous GP-GPU architectures.*

Resumo. *Este artigo propõe uma infraestrutura para realizar a exploração do espaço de projetos de sistemas computacionais com unidades de processamento gráfico (GPUs) em conjunto com núcleos para processamento de propósito geral, com o objetivo de reduzir dark silicon e aumentar o desempenho do sistema em tempo de projeto. A ferramenta GPGPUSim de simulação e estimativa física de projeto foi estendida para realizar estimativas de dark silicon das plataformas de GPUs e, em seguida, foi integrada ao framework MultiExplorer. Adicionalmente, foi desenvolvida uma estratégia para estimativa de desempenho das plataformas de GPU e a modelagem de bases de dados que passaram a utilizar tanto núcleos de GPU quanto de plataformas multicore (núcleos de propósito geral), possibilitando, assim, a exploração do espaço de projeto buscando arquiteturas heterogêneas GP-GPUs.*

1. Introdução

O aumento da capacidade de processamento por meio da evolução do processo tecnológico na indústria de semicondutores tem sido, há décadas, guiada pela Lei de Moore [Schaller 1997] e a escala de Dennard [Dennard et al. 1974]. Entretanto, Dennard [Dennard et al. 2007] e outros já relataram que projetos de circuitos com transistores fabricados em processos tecnológicos abaixo de $90nm$, não seguem a escala original, devido ao efeito exponencial da corrente de fuga, o que acaba por aumentar a densidade de potência e potência dissipada total do *chip*. Dessa forma, originalmente, esses projetos contêm regiões que não poderão estar ativas

(em frequência máxima) concomitantemente com o restante do circuito [Nejatollahi and Salehi 2015, Shafique et al. 2014, Raghunathan et al. 2013], esse efeito é denominado *utilization wall*. A área do *chip* que deve ser mantida "desligada" ou funcionando em frequência aquém do esperado é denominada *dark silicon* [Hardavellas et al. 2011]. Uma forma para tratar esse problema é identificar áreas do *chip* em *dark silicon* e oferecer alternativas arquiteturais que possam ocupar essa área sem afetar a densidade e a potência total dissipada pelo *chip* [Santos et al. 2016].

Diante das limitações de projeto em sistemas com tecnologias abaixo de $90nm$, fabricantes buscaram alternativas para contornar o problema de *dark silicon* e continuar melhorando o desempenho dos sistemas. Os aceleradores como Unidades de Processamento Gráfico (*Graphic Processing Units* - GPUs) passaram a ser muito utilizados em sistemas heterogêneos que mesclam múltiplos núcleos de processamento (*multicore*) e GPUs (sistemas GP-GPU) de forma colaborativa para processamento de demandas de propósito geral [Sanders and Kandrot 2010].

Este trabalho propõe uma infraestrutura para realizar a exploração de projetos de sistemas baseados em GPU considerando restrições de área e densidade de potência visando mitigar o *dark silicon* e objetivando maximizar desempenho. Nesse sentido, apresenta-se aqui o desenvolvimento de um sistema de exploração de espaço de projetos que considera a utilização de plataformas de GPU e núcleos de sistemas *multicore*. Esse desenvolvimento é realizado sobre as ferramentas GPGPUSim [Fung et al. 2007] e MultiExplorer [Santos et al. 2018b]. Além do desenvolvimento, um amplo conjunto de experimentos foi realizado visando mostrar a evolução de *dark silicon* em sistemas compostos exclusivamente por GPUs e, após isso, exploramos o projeto utilizando tanto GPUs quanto sistemas GP-GPU.

Ressalta-se que em nossa revisão bibliográfica não encontramos ferramentas que façam a exploração de espaço de projetos tendo como objetivo sistemas heterogêneos GP-GPUs livres de *dark silicon*. Isto nos motivou a incluir funcionalidades ao MultiExplorer para modelagem, simulação, detecção de *dark silicon* em GPUs e exploração do espaço de projetos heterogêneos.

Este trabalho está organizado como segue: a Seção 2 ilustra trabalhos relacionados com a exploração de espaço de projetos e outras ferramentas de simulação com suporte a GPUs; a Seção 3 descreve o fluxo original do MultiExplorer e as alterações necessárias para inclusão do suporte à modelagem e simulação de GPUs com GPGPUSim; a Seção 4 mostra os experimentos e resultados obtidos para validação da proposta e a Seção 5 conclui este artigo.

2. Trabalhos Relacionados

Considerando o contexto de projetos com *dark silicon*, ressalta-se o trabalho proposto por Turakhia et al. [Turakhia et al. 2013], que apresenta a ferramenta Hades para otimização iterativa de desempenho em MPSoCs (*MultiProcessors System-on-Chip*) com *dark silicon* que determina a quantidade ideal de núcleos para formar um sistema multiprocessador heterogêneo. A ferramenta realiza exploração arquitetural considerando restrições de área e potência. Nota-se, entretanto, que o seu objetivo não é eliminar ou mesmo mitigar o *dark silicon*, e sim melhorar o desempenho do sistema. Além disso, Hades atua em sistemas *multicore* e seu escopo de

heterogeneidade não inclui o uso de GPUs.

No contexto de técnicas ou ferramentas para modelagem e simulação de sistemas com GPUs, citamos os seguintes trabalhos: o GPGPUSim [Fung et al. 2007] é um simulador com precisão de ciclo que, devido à extensão de suporte ao conjunto de instruções CUDA PTX (*Parallel Thread Execution*) [Bakhoda et al. 2009] passou a dar suporte à modelagem de arquiteturas das GPUs atuais. Além disso, Leng et al [Leng et al. 2013] adicionaram o GPUWattch ao GPGPUSim, que deu ainda mais robustez à ferramenta original, pois este passou a prover parâmetros físicos do projeto, além dos parâmetros e estatísticas de desempenho.

Jia et al. [Jia et al. 2012] propôs o Stargazer, um *framework* automatizado baseado em uma modelagem por regressão. Por meio de parâmetros obtidos pela simulação de modelos de GPU, a ferramenta fornece estimativa de desempenho do sistema e traça os parâmetros mais importantes da arquitetura com menos de 1,1% de erro.

Alguns trabalhos se baseiam não somente em GPUs, mas também em outras alternativas de sistemas heterogêneos, por exemplo, *chips* FPGAs (*Field Programmable Gate Array*), para melhoria de desempenho com foco em eficiência energética. O trabalho de Morris e Aubury [Morris and Aubury 2007] explora a utilização de FPGAs em comparação com GPUs em exploração de espaço de projeto por meio da simulação de Monte Carlo [Mooney 1997].

3. Suporte a Sistemas com GPUs no MultiExplorer

A Figura 1 ilustra o fluxo de modelagem, simulação e DS-DSE (*Dark Silicon aware - Design Space Exploration*) da ferramenta MultiExplorer. O objetivo desta ferramenta é dar suporte à exploração, em tempo de projeto, de uma plataforma computacional, desde a simulação de desempenho até a exploração do espaço de projetos ciente de *dark silicon*. O usuário ou projetista especifica parâmetros iniciais para a modelagem da plataforma, por exemplo, informações sobre a arquitetura do sistema, estágios de *pipeline*, quantidade de núcleos, arquitetura e políticas de *caches*, entre outros parâmetros. Em seguida, o usuário pode escolher entre três simuladores funcionais ou simuladores de desempenho (MPSoCBench [Duenha et al. 2014], Sniper [Carlson et al. 2011] e Multi2Sim [Ubal et al. 2012]) que fornecem como saída relatórios de desempenho da plataforma. Os resultados estatísticos da simulação funcional finalizam a primeira etapa do fluxo como exploração de desempenho.

Na etapa denominada de *exploração física*, os resultados da simulação de desempenho são fornecidos para a ferramenta de estimativa física McPAT [Li et al. 2013], a qual retorna informações de temporização, área e consumo de potência de cada componente da plataforma. No trabalho de [Santos et al. 2016] foi apresentada uma extensão junto ao MultiExplorer, especificamente no fluxo da ferramenta McPAT para realizar, em tempo de projeto, a estimativa de *dark silicon*. Essa estimativa consiste em identificar a densidade de potência sobre um *chip* e compará-lo com a densidade de potência de um projeto de referência (processador projetado com tecnologia de $90nm$). Se houver aumento da densidade de potência em projetos baseados nas mesmas características físicas e de desempenho, mas com diferentes

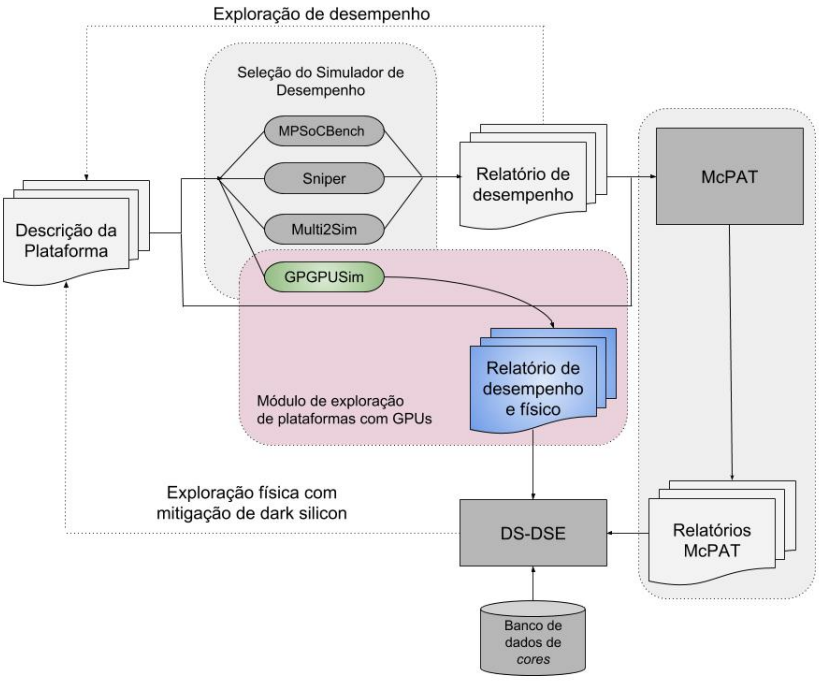

Figura 1. Fluxo de projeto da ferramenta MultiExplorer, com destaque para a extensão de modelagem e simulação de GPUs

processos tecnológicos, então, isso gerará aumento da potência total sobre o *chip* e, considerando que os projetos possuem a mesma área, um incremento da densidade de potência indicando, então, a presença de *dark silicon* no projeto. Os cálculos para estimativa de *dark silicon* encontram-se em [Santos et al. 2016] e [Santos et al. 2018b].

Neste trabalho, um novo módulo de simulação de GPUs foi adicionado ao fluxo MultiExplorer, que abrange as etapas de exploração de desempenho e exploração física a partir da integração da ferramenta GPGPUSim. A ferramenta simula GPUs em um nível de ciclos e disponibiliza informações de performance e físicas sobre o modelo de GPU em questão [Bakhoda et al. 2009].

Com esse novo módulo, o fluxo do MultiExplorer foi alterado, pois o GPGPUSim é um software que já tem embutido o simulador de características físicas (GPUWattch). Desta forma, caso o usuário descreva uma plataforma com GPUs, o simulador selecionado deve ser o GPGPUSim e os relatórios de saída são tratados e repassados diretamente ao módulo de exploração de espaço de projetos (DS-DSE), sem necessidade de passar pelo estimador de parâmetros físicos McPAT, como mostra a Figura 1.

3.1. Banco de Dados de Núcleos de Processadores e Unidades de Computação

Para dar suporte à exploração do espaço de projetos de sistemas heterogêneos GP-GPUs, faz-se necessário desenvolver um banco de dados com informações

de caracterização dos circuitos de referência disponíveis para que o módulo DS-DSE possa utilizá-los para explorar alternativas arquiteturais (Figura 1). Baseando em [Santos et al. 2018b], que utilizaram um banco de dados de núcleos de plataformas *multicore*, este trabalho considerou, como circuito de referência, unidades de computação (UCs) de GPUs e núcleos de plataformas *multicore*. A Tabela 1 apresenta os núcleos de processadores, os quais já se encontravam no banco originalmente, assim como as configurações das unidades de computação de GPUs incluídos neste trabalho. A última coluna utiliza a métrica IPC (instruções por ciclo) para expressar o desempenho das plataformas a partir da execução dos *benchmarks* apresentados na Tabela 2.

Tabela 1. Núcleos e unidades de computação presentes no banco de dados

Núcleos	**Tecnologia**	**Frequência** (Ghz)	**Potência** (W)	**Área do núcleo** (mm^2)	**Desempenho de núcleos/UCs**
1 - Smithfield	90nm	2,8	8,9	111,12	6428
2 - Quark x1000	32nm	0,4	1,06	11,32	502,53
3 - ARM A53	22nm	1,6	5,5	6,82	3125,68
4 - ARM A57	22nm	2,0	12,13	6,79	4006,64
5 - Atom Silvermont	22nm	0,5	2,51	6,15	648,47
6 - Quadro FX5600	65nm	650	16,00	32,41	35814,41
7 - 8800GTX	65nm	576	19,69	13,98	47619,45
8 - 9800GTX	65nm	675	18,65	16,31	55862,05
9 - GTX285	65nm	648	41,06	18,81	52144,24
10 - GTX480	65nm	700	28,52	77,29	256833,49
11 - GTX580	65nm	772	30,05	77,29	288959,60
12 - GTX680	65nm	1006	34,96	178,11	64618,49
13 - GTX780ti	65nm	876	66,78	182,97	54918,42

Deve-se notar que os valores dos parâmetros físicos dizem respeito a estimativas realizadas pela ferramenta GPGPUSim configurada para estimar os parâmetros de uma unidade de computação de cada arquitetura de GPU e de um núcleo de cada plataforma *multicore*. A caracterização das plataformas no banco de dados inclui também o desempenho de todos os modelos, obtidos por simulação utilizando os *benchmarks* e ferramentas apresentados na Tabela 2.

Tabela 2. Aplicações utilizadas para a caracterização de desempenho das plataformas

Benchmark	**Aplicações**	**Simulador**
SPLASH-2	Barnes, FFT, Radix	Sniper
PARSEC	Fluidanimate, Swaptions	Sniper
CUDA Math	MonteCarlo, simpleMultiGPU, scalarProd, clock	GPGPUSim

A métrica de desempenho adotada para cada núcleo/UC considera o IPC (Instruções por ciclo). A simulação de cada modelo gera o valor de IPC relativo a cada aplicação. O IPC relativo a todo o *benchmark* é calculado por meio da média ponderada dos IPCs obtidos para cada aplicação separadamente, considerando como fator de ponderação a carga de trabalho de cada aplicação do conjunto. Assim, aplicações mais robustas serão mais relevantes no cômputo do IPC final do que as aplicações mais leves.

3.2. Estimativa de *Dark Silicon* em GPUs

A estimativa de *dark silicon* em GPUs baseia-se na metodologia original apresentada em Santos et al. [Santos et al. 2016] e posteriormente aplicada na exploração do espaço de projetos de sistemas *multicore* [Santos et al. 2018b].

Utiliza-se uma GPU modelada com tecnologia de $90nm$ e considera-se, com base na escala de Dennard [Dennard et al. 2007], que o projeto está livre de *dark silicon*. Então, evolui-se esse projeto utilizando processos tecnológicos de $65nm$, $45nm$, $32nm$ e $22nm$ adicionando mais unidades de computação para atingir uma área próxima à área do projeto original. A presença de *dark silicon* no projeto será estimado a partir da densidade de potência do novo projeto, quando comparado com o projeto original livre de *dark silicon*. A densidade de potência da GPU é calculada dividindo-se à potência pela área do *chip*, antes e depois da evolução tecnológica. Se a diferença entre a densidade de potência do projeto depois da evolução tecnológica e do projeto original for maior do que zero, considera-se que esse projeto possui área em *dark silicon*. Nesse caso, calcula-se também o total de potência excedida.

Na solução proposta, cada plataforma computacional possui componentes com diferentes densidades de potência. Tais componentes são os candidatos para uma estratégia de exploração de espaço de projeto, uma vez que mudanças na configuração desses componentes alterará a densidade de potência total da plataforma e, como consequência, a área em *dark silicon* no *chip*. Com base nos dados de densidade de potência e área do circuito de referência e do total de potência excedida do projeto, estima-se a área de *dark silicon* do projeto. Em plataformas compostas por GPUs, as unidades de computação (*compute units*) são os componentes com maior densidade de potência.

Partindo, então, do princípio de que abaixo de $90nm$ há aumento energético considerável no *chip* [Dennard et al. 2007], experimentos foram realizados para estimar *dark silicon*. Para cada projeto de GPU, considerou-se fixar o processo tecnológico em $90nm$, e usar os resultados desse processo como base para a evolução em $65nm$, $45nm$, $32nm$ e $22nm$ (limite máximo suportado pela ferramenta GPGPUSim). A Tabela 3 apresenta os resultados das estimativas físicas.

A Tabela 3 apresenta as configurações de quatro plataformas de GPU apresentadas como parte do banco de dados (Tabela 1). Entretanto, visando estimar a presença de *dark silicon* mediante a evolução do processo tecnológico dessas plataformas, a Tabela 3 apresenta a quantidade de unidades de computação utilizadas em cada tecnologia, a frequência de *clock*, a área do *chip* e a densidade de potência observada. O planejamento desse experimento utilizou os parâmetros quantidade de unidades de computação e área do *chip* o mais próximo possível das versões comerciais dessas plataformas. Por outro lado, optou-se por manter a frequência de *clock* constante entre os processos tecnológicos abaixo de $90nm$ uma vez que a evolução desse parâmetro, seguindo a escala de Dennard (S, S = fator de escala), gera uma estimativa de *dark silicon* pessimista [Santos et al. 2018b].

Os valores retornados pela ferramenta com respeito à área do *chip* são valores estimados, baseando-se na configuração da plataforma e processo tecnológico. O GPGPUSim utiliza, como motor de estimativas físicas, a ferramenta McPAT [Li et al. 2009], que possui imprecisões de estimativas físicas, com ênfase na área (erros máximos de 20%), conforme já declarado em [Li et al. 2013]. A estimativa de densidade de potência considera a razão entre a potência total dissipada pelo *chip* (soma de potência dinâmica e potência estática) pela área do *chip*. Nota-se, pela Tabela 3,

Tabela 3. Estimativas físicas e frequência de clock constante dos modelos

GTX285					
Quantidade de UCs	30	59	121	226	435
Tecnologia (nm)	90	65	45	32	22
Frequência (MHz)	648	648	648	648	648
Área do *chip* (mm^2)	1199,08	1207,40	1197,91	1197,33	1200,40
Densidade de potência $(\frac{W}{mm^2})$	0,15	0,17	0,22	0,39	0,38
9800GTX					
Quantidade de UCs	16	29	66	123	231
Tecnologia (nm)	90	65	45	32	22
Frequência (MHz)	675	675	675	675	675
Área do *chip* (mm^2)	539,33	533,53	536,51	540,81	540,10
Densidade de potência $(\frac{W}{mm^2})$	0,14	0,17	0,23	0,40	0,38
8800GTX					
Quantidade de UCs	16	31	65	120	226
Tecnologia (nm)	90	65	45	32	22
Frequência (MHz)	576	576	576	576	576
Área do *chip* (mm^2)	493,04	487,87	495,28	494,45	493,23
Densidade de potência $(\frac{W}{mm^2})$	0,14	0,17	0,22	0,38	0,37
QuadroFX5600					
Quantidade de UCs	8	16	33	64	129
Tecnologia (nm)	90	65	45	32	22
Frequência (MHz)	650	650	650	650	650
Área do *chip* (mm^2)	564,13	570,53	560,66	561,01	563,53
Densidade de potência $(\frac{W}{mm^2})$	0,10	0,11	0,15	0,56	0,57

um aumento significativo desse resultado ao longo dos processos tecnológicos. Como a área não é aumentada entre os processos tecnológicos, o aumento da densidade de potência é em função do aumento da potência total (dinâmica e estática). Analisando os dados dos relatórios internos da ferramenta, conclui-se que a potência estática e, especialmente, a potência devido à corrente de fuga, são as principais responsáveis pelo aumento da potência total. Entretanto, tal aumento corrobora com as estimativas de Dennard sobre o aumento da potência em tecnologias abaixo de $90nm$ e, como consequência, da existência de *dark silicon* no *chip*.

Realizamos um novo experimento no fluxo da ferramenta MultiExplorer com o objetivo de estimar a área do *chip* em *dark silicon*. A Tabela 4 ilustra as estimativas de *dark silicon* para as mesmas plataformas de GPU da Tabela 3, usando uma unidade de computação como circuito de referência. Deve-se observar que os valores de área do circuito e potência do circuito correspondem à unidade de computação (circuito de referência) e não ao *chip* da plataforma de GPU.

Importante notar que há, para algumas plataformas, um pequeno aumento ou mesmo redução de *dark silicon* com $22nm$ que é justificado pelos erros da ferramenta McPAT [Li et al. 2013] nas estimativas de área e potência em processos tecnológicos menores. Deve-se observar também que o valor estimado de *dark silicon* considera a metodologia apresentada em [Santos et al. 2018b] em que a partir da diferença de densidade de potência entre plataformas com diferentes processos tecnológicos, obtém-se a área em *dark silicon* e, tendo a área do circuito de referência, estima-se a porcentagem de *dark silicon* sobre a plataforma. Como exemplo, considere a plataforma GTX8800 em seu processo tecnológico de $90nm$ com uma porcentagem de *dark silicon* de $16,71\%$. Considere os seguintes parâmetros:

- ΔDP = diferença entre densidade de potência do *chip* atual e *chip* de refe-

Tabela 4. Estimativas de dark silicon usando uma unidade de computação de GPU como circuito de referência

GTX285				
Tecnologia (nm)	65	45	32	22
Área do *circuito* (mm^2)	18,81	9,42	5,13	2,70
Potência do *circuito* (W)	2,95	2,07	2,02	1,03
Porcentagem de *Dark Silicon* (%)	14,73	34,86	61,39	60,17
9800GTX				
Tecnologia (nm)	65	45	32	22
Área do *circuito* (mm^2)	16,31	7,60	4,22	2,28
Potência do *circuito* (W)	2,58	1,75	1,71	0,87
Porcentagem de *Dark Silicon* (%)	13,97	37,38	62,45	60,80
8800GTX				
Tecnologia (nm)	65	45	32	22
Área do *circuito* (mm^2)	13,97	7,08	3,92	2,11
Potência do *circuito* (W)	2,17	1,54	1,53	0,78
Porcentagem de *Dark Silicon* (%)	16,71	35,73	61,64	59,71
QuadroFX5600				
Tecnologia (nm)	65	45	32	22
Área do *circuito* (mm^2)	32,40	15,99	8,42	4,25
Potência do *circuito* (W)	3,27	2,32	4,78	2,46
Porcentagem de *Dark Silicon* (%)	9,90	32,37	80,27	80,77

rência (plataforma com $90nm$).

- P_{exc} = potência excedente no *chip*.
- A_{DS} = área em *dark silicon* no *chip* considerando a potência e área do circuito de referência (P_{circ} e A_{circ}).
- DS= estimativa em porcentagem da área do *chip* em *dark silicon*.

Assim, a porcentagem de *dark silicon* do *chip* é calculada da seguinte forma:

$$\Delta DP = DP_{atual} - DP_{base} = 0,173 - 0,147 = 0,026 \tag{1}$$

$$P_{exc} = \Delta DP \times A_{chip} = 0,026 \times 487,87 \approx 12,67 \tag{2}$$

$$A_{DS} = \frac{P_{exc}}{P_{circ}} \times A_{circ} = \frac{12,67}{2,17} \times 13,97 \approx 81,56 \tag{3}$$

$$DS = \frac{A_{DS}}{A_{chip}} = \frac{81,56}{487,87} \approx 0,1671 = 16,71\% \tag{4}$$

Por fim, reforça-se que a informação de *dark silicon* apresentada na Tabela 4 aponta que uma área correspondente a essa porcentagem deve ser "desligada" para que a plataforma, na tecnologia especificada, possa ter a mesma potência de sua plataforma-base ($90nm$). Essa informação de *dark silicon* consiste na devida motivação para se realizar um processo de exploração do projeto dessas plataformas, uma vez que, a partir de $32nm$, a maioria das plataformas tem mais da metade da área do *chip* comprometida com *dark silicon*.

4. Resultados com a Exploração do Espaço de Projeto

Esta seção apresenta os experimentos e resultados que demonstram a integração da ferramenta GPGPUSim ao fluxo do *framework* MultiExplorer para exploração do espaço de projetos ciente de *dark silicon* em plataformas contendo GPUs.

A exploração de espaço de projeto da ferramenta MultiExplorer foi realizada a partir dos núcleos e unidades de computação apresentados na Tabela 1. Por questões de limitação de espaço do texto, são apresentados apenas os resultados de 4 das 8 plataformas de GPU em processos tecnológicos de $65nm$ e $45nm$. Ressalta-se que para tal experimento, novas bases de dados foram geradas considerando todos os núcleos e UCs da Tabela 1 com $65nm$ e $45nm$.

As Tabelas 5-8 apresentam os resultados de soluções encontradas pelo algoritmo de força bruta (o qual gera resultados a partir da busca de todas as soluções possíveis) obtidos com as plataformas 8800GTX ($65nm$), 9800GTX ($45nm$), GTX285 ($45nm$) e QuadroFX5600 ($65nm$). As tabelas apresentam o número de unidades de computação originais (N_O) da plataforma após a proposta alternativa, a quantidade de novas unidades de computação sugeridas (N_{IP}), o tipo dessas unidades de computação de acordo com a Tabela 1 (T_{IP}), a área ($AT(mm^2)$), densidade de potência ($DP(\frac{W}{mm^2})$) e o desempenho ($Performance$) da nova plataforma.

Tabela 5. 8800GTX: Soluções com o algoritmo de Força Bruta - $65nm$**. Restrições:** $AT \leq 493,04$ **e** $DP \leq 0,14$

N_O	N_{IP}	T_{IP}	$AT(mm^2)$	$DP(\frac{W}{mm^2})$	$Performance$
2	2	11	461,14	0,09	673158
2	2	10	461,14	0,09	608905
2	1	10	299,13	0,13	352072
1	6	7	479,92	0,04	333336
3	4	1	446,36	0,13	168570

Os resultados apresentados nessas tabelas indicam as soluções arquiteturais alternativas para as plataformas originais (ver Tabela 3) considerando as restrições de área e densidade de potência. Por exemplo, na Tabela 5, pode-se observar que as soluções possuem área menor ou igual a 493,04 e densidade de potência menor ou igual a 0,14. Esses valores de referência são obtidos da Tabela 3. Ainda considerando como exemplo os resultados da plataforma 8800GTX, a primeira linha da Tabela 5 exibe que a primeira plataforma sugerida pelo método de força bruta, em comparação com a plataforma original 8800GTX foi gerada a partir de 2 UCs 8800GTX ($N_O = 2$) e 2 UCs GTX580 ($N_{IP} = 2$ e $T_{IP} = 11$ (ver Tabela 1).

Tabela 6. 9800GTX: Soluções com o algoritmo de Força Bruta - $45nm$**. Restrições:** $AT \leq 539,33$ **e** $DP \leq 0,14$

N_O	N_{IP}	T_{IP}	$AT(mm^2)$	$DP(\frac{W}{mm^2})$	$Performance$
2	5	11	537,44	0,05	1556522
2	5	10	537,44	0,05	1395891
4	5	10	531,64	0,09	1250782
1	5	11	404,53	0,03	1211700
6	13	5	523,45	0,13	343602

Os resultados apresentados nas Tabelas 5-8 demonstram a factibilidade de aplicação da estratégia DS-DSE para obter plataformas alternativas livres de *dark silicon*, algumas delas, inclusive, arquiteturas GP-GPU. Observou-se que o desempenho obtido com essas plataformas alternativas pode ser também superior ao de-

Tabela 7. GTX285: Soluções com o algoritmo de Força Bruta - $45nm$. Restrições: $AT \leq 1199,08$ e $DP \leq 0,15$

N_O	N_{IP}	T_{IP}	$AT(mm^2)$	$DP(\frac{W}{mm^2})$	$Performance$
1	12	11	1153,59	0,02	3519659
1	11	11	1063,05	0,03	3230699
1	12	10	1153,59	0,02	3134146
4	10	11	1173,84	0,09	3098172
1	8	1	273,67	0,11	103568

Tabela 8. QuadroFX5600: Soluções com o algoritmo de Força Bruta - $65nm$. Restrições: $AT \leq 564,13$ e $DP \leq 0,10$

N_O	N_{IP}	T_{IP}	$AT(mm^2)$	$DP(\frac{W}{mm^2})$	$Performance$
2	2	11	492,9	0,07	649548
2	2	10	492,9	0,07	585295
1	6	8	545,48	0,03	370986
3	4	8	560,68	0,09	330891
2	1	10	330,89	0,10	328462

sempenho da plataforma com *dark silicon*. Por exemplo, a simulação da plataforma GTX8800 com $65nm$ possui desempenho de 359172 enquanto que a solução alternativa de melhor desempenho (primeira linha da Tabela 5) obteve 673158, um *speedup* de 1,87. Baseando-se nessa mesma comparação, o *speedup* obtido com as soluções alternativas para as GPUs GTX9800, GTX280 e QuadroFX5600 foram respectivamente 4,17, 5,32 e 3,41. Ressalta-se, entretanto, que a estimativa de desempenho gerada pelo módulo DS-DSE da ferramenta MultiExplorer consiste numa abordagem otimista que considera o aumento linear do desempenho nas configurações de GPU à medida em que aumentamos a quantidade de núcleos.

5. Conclusões e Trabalhos Futuros

Este trabalho apresentou o desenvolvimento de uma extensão da ferramenta de exploração de espaço de projetos (MultiExplorer), originalmente validada para sistemas *multicore* e *manycore*, para exploração do projeto de sistemas com GPUs. O desenvolvimento consistiu na extensão da ferramenta GPGPUSim para estimativa de *dark silicon* e na integração com MultiExplorer visando possibilitar que plataformas de GPU suportadas por GPGPUSim sejam utilizadas no fluxo MultiExplorer. Adicionalmente, foi desenvolvida uma estratégia para estimativa de desempenho, baseada no IPC alcançado pelas plataformas de GPU e a modelagem de bases de dados que utilizavam tanto núcleos de GPU quanto de plataformas *multicore* (núcleos *General Purpose*-GP) possibilitando assim a exploração do espaço de projeto buscando arquiteturas heterogêneas GP-GPU.

Os experimentos e resultados obtidos permitiram avaliar o comportamento evolutivo do projeto de GPUs quanto à presença de *dark silicon*. Também possibilitaram verificar que arranjos arquiteturais entre GPUs ou mesmo de núcleos com GPUs (sistemas GP-GPU) são alternativas viáveis para mitigar a presença de *dark silicon* (Tabelas 5-8). Interessante observar que as soluções propostas pela estratégia DS-DSE desenvolvida na ferramenta MultiExplorer não apenas gerou alternativas

viáveis quanto ao *dark silicon* mas também com desempenho superior (*speedups* de até $5,32$) comparado às plataformas com *dark silicon*.

Como trabalhos futuros vislumbra-se a utilização de uma estratégia mais precisa e acurada para estimativa de desempenho das soluções geradas pelo DS-DSE. Uma possibilidade é aplicar um preditor de desempenho robusto, baseado em regressores de vetor de suporte, como o apresentado em [Santos et al. 2018a]. Adicionalmente, explorar soluções baseadas em algoritmos genéticos como em [Santos et al. 2018b] visando minimizar o tempo na execução do DS-DSE.

Referências

Bakhoda, A., Yuan, G. L., Fung, W. W., Wong, H., and Aamodt, T. M. (2009). Analyzing cuda workloads using a detailed gpu simulator. In *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on*, pages 163–174. IEEE.

Carlson, T. E., Heirman, W., and Eeckhout, L. (2011). Sniper: exploring the level of abstraction for scalable and accurate parallel multi-core simulation. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, page 52. ACM.

Dennard, R., Gaensslen, F., Yu, H.-N., Rideout, L., Bassous, E., and Leblanc, A. (1974). Design of ion-implanted mosfets with very small physical dimensions. *IEEE Journal of Solid-Circuits*, pages 256–267.

Dennard, R. H., Cai, J., and Kumar, A. (2007). A perspective on today's scaling challenges and possible future directions. *Solid-State Electronics*, 51(4):518–525.

Duenha, L., Guedes, M., Almeida, H., Boy, M., and Azevedo, R. (2014). MPSoC-Bench: A toolset for MPSoC system level evaluation. In *International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS XIV)*, pages 164–171. IEEE.

Fung, W. W., Sham, I., Yuan, G., and Aamodt, T. M. (2007). Dynamic warp formation and scheduling for efficient gpu control flow. In *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture*, pages 407–420. IEEE Computer Society.

Hardavellas, N., Ferdman, M., Falsafi, B., and Ailamaki, A. (2011). Toward dark silicon in servers. *IEEE Micro*, 31(4):6–15.

Jia, W., Shaw, K. A., and Martonosi, M. (2012). Stargazer: Automated regression-based gpu design space exploration. In *Performance Analysis of Systems and Software (ISPASS), 2012 IEEE International Symposium on*, pages 2–13. IEEE.

Leng, J., Hetherington, T., ElTantawy, A., Gilani, S., Kim, N. S., Aamodt, T. M., and Reddi, V. J. (2013). Gpuwattch: enabling energy optimizations in gpgpus. In *ACM SIGARCH Computer Architecture News*, volume 41, pages 487–498. ACM.

Li, S., Ahn, J., Strong, R., Brockman, J., Tullsen, D., and Jouppi, N. (2009). McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures. In *42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pages 469–480. IEEE.

Li, S., Ahn, J., Strong, R., Brockman, J., Tullsen, D., and Jouppi, N. (2013). The McPAT framework for multicore and manycore architectures: Simultaneously modeling power, area, and timing. *ACM Transactions on Architecture and Code Optimization (TACO)*, 10(1):5.

Mooney, C. Z. (1997). *Monte carlo simulation*, volume 116. Sage Publications.

Morris, G. W. and Aubury, M. (2007). Design space exploration of the european option benchmark using hyperstreams. In *Field Programmable Logic and Applications, 2007. FPL 2007. International Conference on*, pages 5–10. IEEE.

Nejatollahi, H. and Salehi, M. E. (2015). Voltage scaling and dark silicon in symmetric multicore processors. *The Journal of Supercomputing*, 71(10):3958–3973.

Raghunathan, B., Turakhia, Y., Garg, S., and Marculescu, D. (2013). Cherry-picking: exploiting process variations in dark-silicon homogeneous chip multi-processors. In *Proceedings of the DATE*, pages 39–44. EDA Consortium.

Sanders, J. and Kandrot, E. (2010). *CUDA by example: an introduction to general-purpose GPU programming*. Addison-Wesley Professional.

Santos, M. T., Oliveira, T., Sonohata, R., Krebs, C., Duenha, L., and Santos, R. (2018a). Modelo de predição de desempenho integrado à exploração do espaço de projetos. In *Anais do Workshop de Computação Heterogênea (WCH)*, pages 630–641.

Santos, R., Duenha, L., Silva, A. C., Sousa, M., Tedesco, L. A., Melgarejo, J. C., Santos, T., Azevedo, R., and Moreno, E. (2018b). Dark-silicon aware design space exploration. *Journal of Parallel and Distributed Computing*, 120:295–306.

Santos, T., Silva, A., Duenha, L., Santos, R., Moreno, E., and Azevedo, R. (2016). On the dark silicon automatic evaluation on multicore processors. In *Proceedings of the SBAC-PAD*, pages 166–173. IEEE.

Schaller, R. (1997). Moore's law: Past, present and future. *IEEE Spectrum*, 34(6):52–59.

Shafique, M., Garg, S., Henkel, J., and Marculescu, D. (2014). The eda challenges in the dark silicon era: Temperature, reliability, and variability perspectives. In *Proceedings of the 51st Annual DAC*, pages 1–6. ACM.

Turakhia, Y., Raghunathan, B., Garg, S., and Marculescu, D. (2013). Hades: Architectural synthesis for heterogeneous dark silicon chip multi-processors. In *Proceedings of the 50th Annual DAC*, page 173. ACM.

Ubal, R., Jang, B., Mistry, P., Schaa, D., and Kaeli, D. (2012). Multi2Sim: a simulation framework for CPU-GPU computing. In *Proceedings of the 21st international conference on Parallel architectures and compilation techniques*, pages 335–344. ACM.