

Exploração do Espaço de Projetos de Sistemas GPGPUs ciente de Dark-Silicon

Laura B. Ludgero¹, Liana D. Duenha¹

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

{laura.ludgero, liana.duenha}@ufms.br

Abstract. *This work presents an extension of the MultiExplorer tool for GPGPU design space exploration. For this purpose, new simulators and estimators were integrated into the tool flow; the performance, area and consumption of six GPU models running CUDA and Rodinia applications were characterized; a bank of alternative cores required for architectural exploration was modeled, and the interface was extended to adapt to the new domain.*

Resumo. *Este trabalho apresenta uma extensão da ferramenta MultiExplorer para a exploração do espaço de projetos GPGPUs. Para tal, foi realizada a integração de novos simuladores e estimadores ao fluxo da ferramenta, a caracterização de desempenho, área e consumo de seis modelos de GPUs executando aplicações CUDA e Rodinia, a modelagem de um banco de núcleos alternativos necessários para a exploração arquitetural e a extensão da interface adaptada ao novo domínio.*

1. Introdução

A busca incessante por processadores cada vez mais rápidos, impulsionada pela Lei de Moore [Moore 1998], está atingindo seus limites. O aumento significativo do número de transistores em um único chip está gerando um consumo de energia cada vez maior, culminando no fenômeno conhecido como *Dark-Silicon*, caracterizado pela crescente porcentagem de transistores inativos em um chip, devido as restrições de sistemas computacionais, o que limita o desempenho e a eficiência energética desses dispositivos.

Neste contexto, surgiram as arquiteturas heterogêneas com múltiplos núcleos e os aceleradores, como as Unidades de Processamento Gráfico - GPUs, que vêm se tornando cada vez mais acessíveis para o processamento de propósito geral. No entanto, a complexidade dessas novas arquiteturas demandam ferramentas de análise e otimização sofisticadas, que, todavia, ainda são escassas. Outros desafios encontrados, são a dificuldade em se obter dados de diversos parâmetros arquiteturais na documentação dos fabricantes e a dependência que os simuladores têm de executarem em *hosts* com GPUs reais, entre outros.

Este trabalho apresenta a extensão da ferramenta Multiexplorer [Santos et al. 2018] para o domínio GP-GPU visando mitigar o efeito *dark-silicon* de placas gráficas por meio do projeto de GPUs heterogêneas. Inicialmente, foram caracterizados os parâmetros de desempenho, estimativas físicas e de consumo de seis placas de GPUs em suas versões de mercado, em distintas configurações e litografias.

Em seguida, foi elaborado um banco de núcleos de GPU, para servirem como núcleos alternativos para uma exploração arquitetural.

Este artigo está organizado como segue: A Seção 2 apresenta os trabalhos prévios na área de exploração do espaço de projetos de sistemas *multicore*; a Seção 3 descreve as etapas realizadas para estender o MultiExplorer para o domínio GPGPU; a Seção 4 mostra a caracterização das placas de GPU; a Seção 5 apresenta a aplicação da metodologia de estimativa de *dark-silicon* das placas de GPU; a seção 6 apresenta e detalha a metodologia e os resultados da exploração de projetos; por fim, a Seção 7 traz as considerações finais.

2. MultiExplorer

Embora haja trabalhos relacionados a avaliação de desempenho de aplicações em GPUs [Baghsorkhi et al. 2010], em nossa revisão bibliográfica, não foram encontradas ferramentas ou modelos para realizar a exploração do espaço de projetos de placas de GPUs, tampouco estimativas de *dark-silicon* para estas arquiteturas. O que nos motivou a dar continuidade na integração do fluxo de Design Space Exploration (DSE) com arquiteturas de GPUs no software MultiExplorer.

O MultiExplorer é uma ferramenta de exploração de espaço de projetos de sistemas multiprocessados em chip (MPSoCs) desenvolvida pelo grupo de pesquisa do Laboratório de Sistemas de Alto Desempenho (LSCAD) da Faculdade de Computação (FACOM/UFMS) [Santos et al. 2018]. Inicialmente foi projetado para a avaliação de desempenho dos componentes da plataforma, bem como fornecia estimativas de área, consumo e *dark-silicon* de arquiteturas *multicore*, com possibilidade de exploração de alternativas arquiteturais heterogêneas, livres de *dark-silicon*, com a menor redução de desempenho possível. A ferramenta se destaca por sua flexibilidade, permitindo a exploração de um amplo espaço de projetos de MPSoCs. As extensões do MultiExplorer para GPGPU e otimização de recursos em máquina virtuais [Arigoni et al. 2022], abrem novas possibilidades de aplicação da ferramenta.

Um estudo sobre a utilização do MultiExplorer para fins educacionais foi apresentado em [Amorim et al. 2023] e o código-fonte da ferramenta foi disponibilizado à comunidade para contribuições (<https://github.com/lscad-facom-ufms/multiexplorer>).

A Figura 1 exemplifica o fluxo de exploração do espaço de projetos do MultiExplorer. A ferramenta utiliza como entrada a descrição de uma plataforma-base com múltiplos núcleos de uma mesma arquitetura (plataforma homogênea), a sua tecnologia de fabricação e um conjunto de aplicações que são simuladas funcionalmente. Os dados de desempenho são então repassados para um estimador de parâmetros físicos, já estendido para incluir a estimativa de *dark-silicon* da plataforma. Se o projeto-base apresentou parte de sua área em *dark-silicon*, pode-se realizar a exploração do espaço de projetos para encontrar alternativas heterogêneas livres de *dark-silicon*, ordenadas do maior para o menor desempenho.

3. Extensão do MultiExplorer para GPGPUs

Esta seção aborda as atividades realizadas durante a extensão do MultiExplorer para o domínio de GPGPUs. Após essa extensão, o MultiExplorer deve ser capaz de receber como entrada a descrição de uma placa de GPU, sua tecnologia de fabricação e

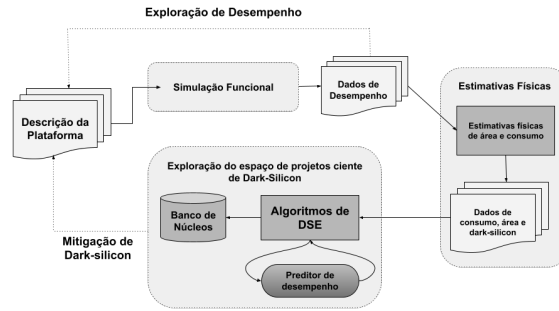


Figura 1. Fluxo do MultiExplorer com GPU

uma aplicação. Com essas entradas o MultiExplorer realiza a simulação da execução da aplicação na placa, as estimativas físicas de área e consumo dos componentes da placa. Em caso de aparecimento de *dark-silicon*, o usuário pode optar pela exploração do espaço de projetos de GPUs heterogêneas, mesclando núcleos de distintas arquiteturas, em busca de soluções livres de *dark-silicon* e com máximo desempenho. Neste trabalho, um *núcleo* ou *core* da GPU é definido como uma *streaming multiprocessor* (SM) da placa.

3.1. Aplicações

Foram validados nove aplicações dos pacotes *benchmark CUDA Library* [NVIDIA 2021] e Rodinia [Che et al. 2009], as quais permitem uma avaliação das placas em cenários diferentes. A Tabela 1 mostra a quantidade de instruções em cada uma e o peso de cada aplicação no *benchmark*, após simulação com uma SM por placa no GPGPU-Sim [Bienia et al. 2008].

Aplicações	Instruções	Peso	Benchmark
clock	1.950.272	0,15 %	CUDA Library
asyncAPI	218.103.808	16,99%	CUDA Library
vectorAdd	1.051.936	0,08%	CUDA Library
nn	1.201.052	0,09%	Rodinia
backprop	134.545.760	10,48%	Rodinia
needle	102.311.808	7,97%	Rodinia
dwt2d	185.349.936	14,43%	Rodinia
bfs	551.590.436	42,96%	Rodinia
hotspot	87.962.832	6,85%	Rodinia
Total	1.284.067.840	100,00%	-

Tabela 1. Quantidade de instruções de cada uma das aplicações e seu peso (ou representatividade) no *benchmark*.

3.2. Integração do Simulador de Desempenho

O simulador de desempenho escolhido para a integração foi o GPGPU-Sim [Bienia et al. 2008], ferramenta desenvolvida pela *University of British Columbia*, que permite a modelagem e simulação de diversas placas de GPUs utilizando-se de *benchmarks* CUDA SDK e Rodinia [Khairy et al. 2020].

Ao simular diferentes arquiteturas e configurações de GPU alterando as aplicações e os parâmetros do arquivo *json* como o número de SMs, litografia, frequência, registradores, memória, entre outros, é possível analisar o impacto dessas variações no desempenho das aplicações utilizadas, permitindo identificar gargalos e melhores configurações para otimização.

Para a extensão do MultiExplorer, foram modeladas seis placas de GPUs (QV100, TITANV, RTX2060, TITANX, GK100, GTX480) que cobrem cinco arquiteturas distintas (Volta, Turing, Pascal, Kepler, Fermi). A Tabela 2 apresenta as informações originais de cada uma dessas placas incluindo arquitetura, tecnologia de fabricação, frequência, quantidade de unidades de computação (SMs) e outros parâmetros relevantes.

Características do Modelo da Placa						
Placa	QV100	TITANV	RTX2060	TITANX	GK100	GTX480
Arquitetura	Volta	Volta	Turing	Pascal	Kepler	Fermi
Tecnologia (ηm)	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

Tabela 2. Parâmetros das placas originais

3.3. Integração do estimador de parâmetros de área, consumo e *dark-silicon*

A ferramenta McPAT [Li et al. 2013], originalmente utilizada para estimativas físicas de área e consumo de CPUs, foi adaptada e incluída ao GPGPU-Sim com o nome de GPUWattch [Leng et al. 2013]. Uma vez que o GPGPU-Sim foi inserido no fluxo do MultiExplorer, esta ferramenta passou a servir para para estimar desempenho, consumo e área de placas de GPU, utilizando o arquivo de saída *XML*. Para prover como saída a estimativa de área de *dark-silicon* nas placas de GPU, o código da ferramenta McPAT que há dentro do GPUWattch foi estendida com a metodologia validada em [Sonohata et al. 2023], a qual é baseada na densidade de potência da placa e de seus componentes internos. O fluxo de execução do MultiExplorer foi alterado para que o arquivo de saída do McPAT fosse gerado automaticamente e aparecesse na pasta de saída da simulação (*output*). Também foram adicionados novos parâmetros que mostram claramente dados como *dark silicon*, densidade de potência e informações originais das placas simuladas, que antes era preciso calcular manualmente.

4. Caracterização das Placas e do Banco de Núcleos

O MultiExplorer usa quatro arquivos como entrada. O arquivo *.json* no qual fornece os parâmetros arquiteturais de cada placa, tais como os números de SMs, tecnologia de fabricação, a largura de banda de memória, entre outros. O arquivo *.icnt* no qual fornece informações com topologia da rede, algoritmos de roteamento, controle de fluxo e detalhes específicos do tráfego e da arquitetura do roteador para o GPGPU-Sim. O arquivo *.config* é utilizado para configurar o GPGPU-Sim, permitindo que o simulador defina parâmetros que controlam o comportamento da simulação e, assim, replique o comportamento de uma GPU real. Por fim, o arquivo, *.xml* que é modificado com base nas informações do arquivo *.json*.

Durante a etapa de exploração de desempenho é realizada a seleção da placa de GPU apropriada para o experimento, juntamente com a aplicação específica que se deseja avaliar. Após a execução da simulação, a ferramenta fornece os dados mais significativos sobre o desempenho, como a quantidade de instruções executadas, o número de ciclos por *kernel*, métricas de desempenho como o IPC e *runtime*. No entanto, para se obter uma visão mais completa e detalhada dos resultados, o arquivo de saída gerado pelo GPGPU-Sim é redirecionado para a *rundir*, onde pode-se encontrar os dados de saída completos, incluindo métricas adicionais que não são apresentadas no resumo inicial, permitindo uma análise mais aprofundada do desempenho da GPU sob as condições simuladas.

A Tabela 3 apresenta os resultados da simulação de desempenho e estimativas físicas das seis placas. Os dados de desempenho foram obtidos a partir da execução das nove aplicações nas seis placas, com apenas 1 SM por placa. Foram utilizadas duas métricas de desempenho: IPC (instruções por ciclo), e MIPS (milhões de Instruções por segundo). As estimativas físicas de área e potência da Tabela 3 foram reportadas para a partir da quantidade de SMs da placa original (descritas na Tabela 2).

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
<i>Runtime (ms)</i>	99	48	106	107	118	305
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	693,44	404,51	826,27
Potência (W)	212,68	240,14	112,24	131,16	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	28,89	31,07	55,08
Potência (W)	2,66	3,00	3,74	5,46	5,71	5,32
Dens. pot. (W/ mm^2)	0,17	0,19	0,21	0,19	0,17	0,10

Tabela 3. Caracterização de desempenho e estimativas físicas de seis placas de GPUs com 1 SM.

O banco de núcleos do MultiExplorer foi estendido com os núcleos das placas de GPUs caracterizadas na Tabela 3. O banco contém os valores da tripla (D,C,A), caracterizando o Desempenho em MIPS (D), o consumo (C) em Watts e a área (A) em mm^2 , para cada placa na sua tecnologia original e em tecnologias mais avançadas, para cada uma das aplicações disponíveis. Cada placa está caracterizada pelo seu nome, seguida da tecnologia (45, 32 ou 22 nanômetros).

As Equações 1 e 2 mostram como o desempenho de cada placa é calculado com base na métrica MIPS do *benchmark* completo. Primeiramente, na Equação 1 calcula-se o desempenho de cada aplicação (D_{app_i}) multiplicando-se o IPC obtido como saída da simulação de cada aplicação e a frequência de operação da placa ($freq$) em MHz ou, equivalentemente, em milhões de ciclos por segundo. Em seguida, calcula-se o desempenho de cada placa (D_{placa}) para o *benchmark* completo como um média ponderada dos desempenhos das aplicações, utilizando, como fator de ponderação, o peso que cada aplicação tem no *benchmark* ($peso_{app_i}$), que é relação entre os ciclos de execução da aplicação na

placa e o total de ciclos do *benchmark* completo [Sonohata et al. 2023]

$$D_{app_i} = IPC_{app_i} \times freq \quad (1)$$

$$D_{placa} = \sum_i (D_{app_i} \times peso_{app_i}) \quad (2)$$

O consumo em W foi descrito utilizando-se a soma de três parâmetros de potência (*Peak Dynamic*, *Subthreshold Leakage* e *Gate Leakage*) que, divididos pelo total de SMs da placa, resulta no consumo por SM. Por último, utiliza-se a área de uma SM da placa, para cada litografia disponível. Essas informações estão disponíveis no arquivo de saída *txt*, que é gerado quando executamos o McPAT manualmente, utilizando o arquivo *xml* na pasta de saída gerada pelo GPGPU-sim.

Aplicação		QV100-22	TITANV-22	RTX2060-22	TITANX-32	TITANX-22
	C	2,66	3,01	3,74	5,66	3,01
	A	15,43	15,24	17,90	30,45	15,24
clock	D	80806,80	185240,33	101387,10	63667,80	52326,44
assyncAPI	D	43892,99	59455,96	46714,08	32750,94	33234,10
vectorAdd	D	32390,52	59100,29	46307,16	34556,55	27323,01
nn	D	32859,02	71619,08	37502,82	30893,64	21974,15
backprop	D	43393,52	90050,92	60463,97	38791,68	30548,30
needle	D	2512,46	4845,64	1515,37	2266,00	1787,66
dwt2d	D	21553,00	41155,59	281348,21	17768,51	34387,66
bfs	D	14019,17	34245,05	19834,75	13776,12	12804,21
hotspot	D	50293,03	132330,65	19834,75	65245,81	58660,46
Aplicação		GK100-32	GK100-22	GTX480-45	GTX480-32	GTX480-22
	C	5,71	2,81	5,32	4,79	2,33
	A	31,07	15,52	55,09	28,90	14,50
clock	D	52326,44	52326,44	31535,20	31535,20	31535,20
assyncAPI	D	33234,10	33234,10	19045,07	19045,07	19045,07
vectorAdd	D	27323,01	27323,01	24633,03	24633,03	24633,03
nn	D	21974,15	21974,15	28497,61	28497,61	28497,61
backprop	D	30548,30	30548,30	16272,48	16272,48	16272,48
needle	D	1787,66	1787,66	1033,32	1033,32	1033,32
dwt2d	D	34387,66	34387,66	6609,19	6609,19	6609,19
bfs	D	12804,21	12804,21	3561,50	3561,50	3561,50
hotspot	D	58660,46	58660,46	23908,12	23908,12	23908,12

Tabela 4. Banco de Núcleos de GPUs do MultiExplorer.

A Tabela 4 mostra os dados resultantes de área e consumo dos projetos de GPUs modeladas até o presente momento, bem como os resultados de desempenho conforme métrica descrita pelas Equações 1 e 2. Estes resultados são apenas uma base de dados para trabalhos futuros sobre a ferramenta, para que sirvam de referência para otimizações futuras.

5. Evolução tecnológica das placas

Foram realizados experimentos nos seis modelos de GPUs da Tabela 3, modelados na tecnologia de fabricação original. O projeto original foi evoluído para tecnologias mais recentes, até o limite máximo de $22nm$, escalando o número de SMs, buscando manter a área o mais próximo possível da área do chip original, a mesma frequência de operação e mesma voltagem. A Tabela 5 mostra a evolução tecnológica das placas *TITANX*. A limitação do processo tecnológico em $22nm$ é da ferramenta McPAT.

Verifica-se que há um significativo aumento na densidade de potência nos distintos processos tecnológicos, as quais se justificam pelo aumento na potência total (dinâmica e estática). A Tabela 5 mostra as estimativas de *dark-silicon* das duas placas de GPU, considerando como circuito de referência os dados de SM. Como exemplo, a plataforma TITANX (22nm) tem estimativa de *dark-silicon* de 6,11%. Considerando como plataforma atual a TITANX em 22nm e a plataforma-base com 32nm, a estimativa da área em *dark-silicon* foi calculada da seguinte forma [Sonohata et al. 2023]:

$$P_{exc} = DP_{atual} - DP_{base} \times A_{chip} = (0.201 - 0.189) \times 693,44 \approx 8,32$$

$$A_{DS} = \frac{P_{exc}}{P_{circ}} \times A_{circ} = \frac{8,32}{2,68} \times 13,33 \approx 41,38$$

$$DS = \frac{A_{DS}}{A_{chip}} = \frac{41,38}{693,44} \approx 0.061 = 6,11\%$$

TitanX (Pascal)		
Número de SMs	24	52
Tecnologia (nm)	32	22
Frequência (MHz)	1000	1000
Área do chip (mm ²)	693,44	693,44
Potência da placa/chip (W)	131,16	139,69
Densidade de Potência (W/mm ²)	0,189	0,201
Área do circuito de referência (mm ²)	28,89	13,33
Potência do circuito de referência (W)	5,46	2,68
Porcentagem de dark-silicon (%)	-	6,11%

Tabela 5. Evolução tecnológica da TITANX.

Se a área correspondente a esta porcentagem for inutilizada, o chip de GPU teria a mesma potência do projeto original (projetada em 32nm). É importante ressaltar que as estimativas de *dark-silicon* são relativas à plataforma-base, a qual consideramos, para aplicação da metodologia, livre de *dark-silicon*.

6. DS-DSE

A Exploração do Espaço de Projetos (DSE) é uma análise sistemática e orientada por objetivos e restrições, realizadas antes da implementação de um sistema computacional, com o intuito de otimizar o projeto. O principal objetivo da DSE é maximizar indicadores como desempenho e largura de banda [Zaccaria et al. 2020], ao mesmo tempo em que minimiza o consumo energético e a área, respeitando limitações físicas. Essa atividade é considerada crucial para o projeto de MPSoCs, dada a complexidade crescente desses chips. Para lidar com o vasto espaço de projeto, o problema é frequentemente modelado como uma otimização multiobjetivo, utilizando componentes como um explorador para gerar novas configurações e um avaliador para simular e prever o desempenho. A ferramenta MultiExplorer foi desenvolvida justamente para apoiar a pesquisa nessa área, permitindo a exploração de projetos para mitigar o efeito do dark silicon, que é a parte de um chip que está desativada devido a limitações energéticas.

A usabilidade da ferramenta MultiExplorer, usada para a exploração de espaço de projetos, foi aprimorada com a implementação de uma Interface Gráfica de Usuário (GUI) [Amorim et al. 2023]. Anteriormente, a interação com a ferramenta era feita exclusivamente via linha de comando, um fator que se mostrava como um dos principais

obstáculos para a adoção da ferramenta por alunos de graduação. Para superar essa dificuldade, a concepção e prototipação da nova interface foram guiadas pela metodologia Design Sprint, que possibilitou o desenvolvimento de uma fachada intuitiva, focada na experiência do usuário. A implementação, realizada em Python com o módulo Tkinter, foi estruturada de forma modular, com uma arquitetura de software parcialmente baseada no padrão MVC, permitindo que a interface seja responsiva e facilmente adaptável a novos fluxos de exploração (VM e Iot) [Amorim et al. 2023].

Para demonstrar o fluxo de utilização da ferramenta MultiExplorer, será apresentada uma atividade prática de simulação e exploração do espaço de projetos. Presume-se que a ferramenta e seus pré-requisitos já foram devidamente instalados e configurados [Rodrigues and Santos 2024]. O processo de instalação pode ser feito diretamente em um ambiente Linux, como o Ubuntu 18.04, sem a necessidade de usar Docker. No entanto, a utilização do Docker é recomendada, pois garante um ambiente homogêneo com todas as bibliotecas e ferramentas específicas já configuradas, facilitando a experiência do usuário. O processo de uso se inicia com a execução da interface gráfica, que é acessada por meio do diretório da ferramenta com o comando.

```
$ python ME.py
```

A partir da execução do comando, a interação com o MultiExplorer se dá por meio da interface gráfica, que se abre em uma janela inicial como demonstrado na Figura 2. Nela, o usuário tem a opção de obter mais informações sobre a ferramenta clicando no botão "About" ou, como ponto de partida para a exploração, selecionar um dos três fluxos disponíveis: GPGPU - Design Space Exploration, CPU - Design Space Exploration, e VM - Design Space Exploration. Para a atividade proposta, o fluxo de GPGPU será selecionado.

Na primeira aba de entradas do usuário, é preciso configurar a arquitetura que será simulada no fluxo de exploração de GPUs, definindo a aplicação, o modelo de GPU e o número de cores, conforme ilustrado na Figura 3

Para iniciar, é fundamental selecionar uma GPU que, em sua evolução tecnológica, tenha demonstrado o fenômeno do dark silicon. Utilizaremos a placa TITANX como exemplo, um modelo caracterizado em nossos experimentos com tecnologia de 32nm e 24 SMs. Com uma área de 693,44 mm² e uma densidade de potência de 0,189 W/mm² na configuração original, a placa já é uma base para a exploração de projetos. Para a evolução tecnológica, o objetivo é manter a área de chip constante, enquanto se reduz a litografia para 22nm, um processo que resulta na necessidade de mais SMs para preencher a mesma área. Essa evolução gera o *dark-silicon* como demonstrado na Tabela 5.

Na aba de configurações de DSE, o usuário pode modificar a calibragem do algoritmo evolutivo de busca (NSGA-II). É possível ajustar o número de núcleos que a solução alternativa pode ter, bem como parâmetros de NSGA-II como a taxa de cruzamento, a taxa de mutação, o número de gerações e o tamanho da população, permitindo um controle fino sobre o processo de otimização. Além de determinar as restrições máximas de área e densidade de potência do chip simulado que devem ser retirados da evolução tecnológica da Tabela 5, como demonstrado na Figura 4.

A interface de execução permite que o usuário acompanhe o progresso das simulações e da busca. Após a conclusão de todas as etapas, o usuário pode prosseguir



Figura 2. Tela Inicia Do MultiExplorer

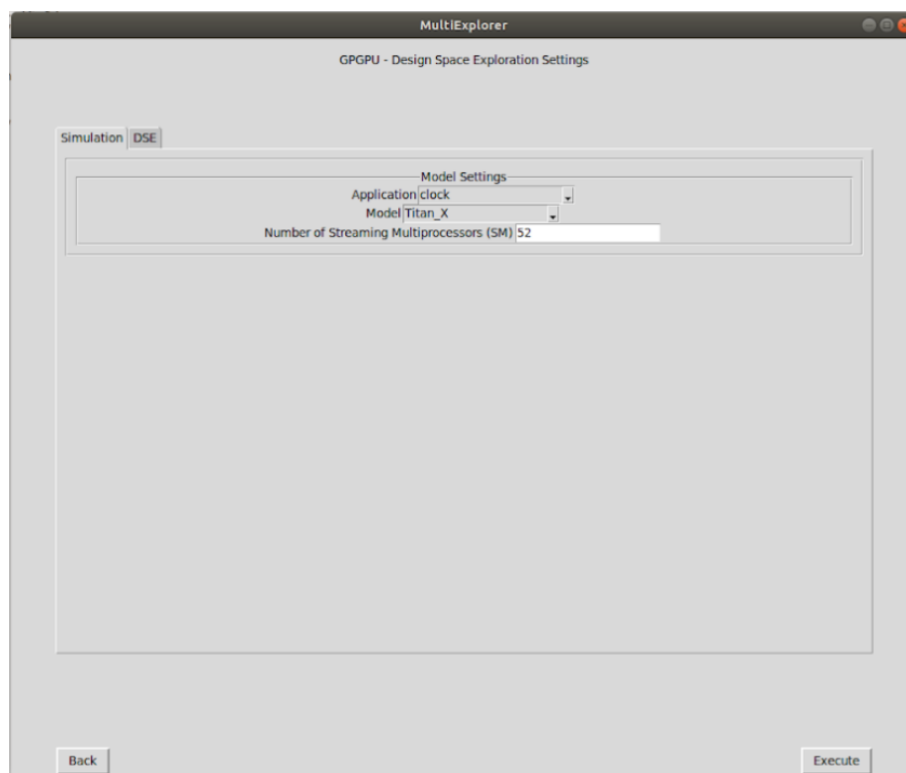


Figura 3. Entradas de usuário do fluxo de exploração de GPUs Heterogêneas

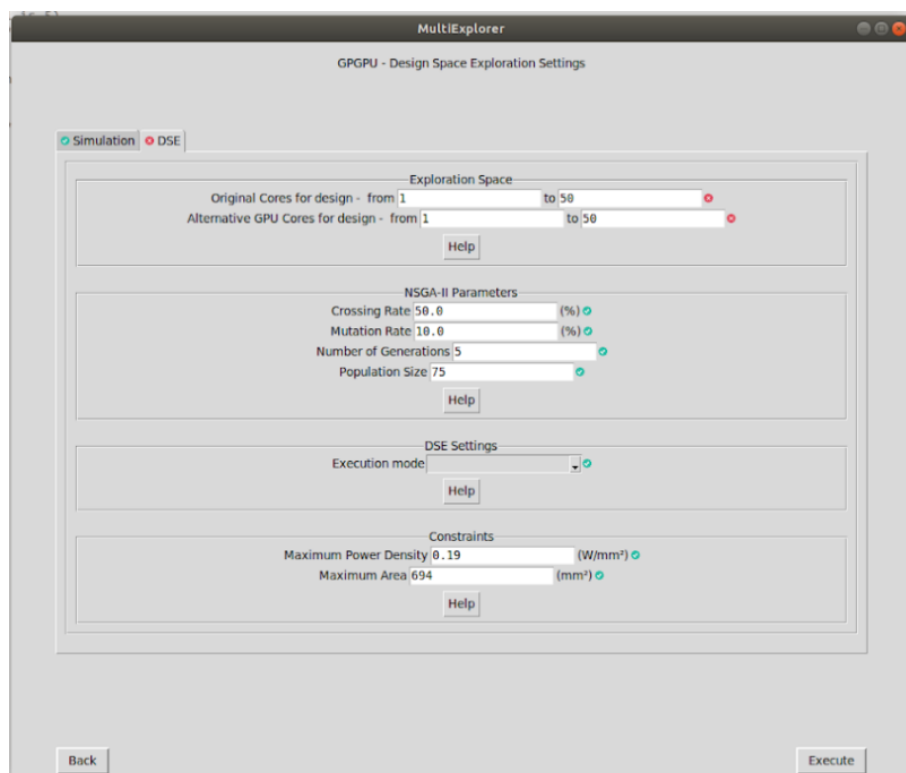


Figura 4. DSE

para a tela de apresentação de resultados. O fim de cada etapa permite a visualização de resultados parciais ("Preview Results"). Após a conclusão de todas as etapas da execução o botão "See Results" é habilitado.

Os resultados da ferramenta incluem um perfil de desempenho e parâmetros físicos das arquiteturas encontradas pelo NSGA-II e pelo algoritmo de força bruta, além de um gráfico que ilustra a comparação entre essas soluções como apresentado nas Figuras 5 a 8. As Figuras 5 e 6 ilustram as soluções encontradas pelo algoritmo genético NSGA-II (aproximado), contendo, para cada solução, o seu desempenho, área e densidade de potência. A Figura 7 mostra soluções encontradas pelo algoritmo de força-bruta (exato); a Figura 8 mostra um gráfico gerado automaticamente pela ferramenta, apresentando os mesmos dados em um formato visual.

É importante ressaltar que todas as soluções oferecidas pelos algoritmos de DSE atendem restrições de densidade de potência e área de uma placa original sem dark-silicon. As soluções são ordenadas com relação à performance alcançada.

Os resultados da exploração de projetos demonstram a eficácia dos algoritmos de busca. A ferramenta disponibiliza tanto as soluções encontradas pelo algoritmo evolutivo NSGA-II quanto as soluções ótimas identificadas pelo método exaustivo de força bruta, que é aplicável em espaços de busca razoavelmente pequenos. Ao comparar os conjuntos de resultados, observa-se que o NSGA-II se aproxima consistentemente do conjunto de soluções encontradas pela força bruta. Em um dos experimentos, por exemplo, o NSGA-II retornou uma solução idêntica à do força bruta: uma arquitetura de 32x TITANX 22nm e 12x TITAN V 22nm, mostrando que, mesmo sendo um método heurístico, ele é capaz

MultiExplorer			
NSGA-II Generated Architectures			
Architecture	Performance	Area	Power Density
32x TITANX 22nm & 12x TITANV 22nm	1017810.13 s ⁻¹	637.96 mm ²	0.19 W/mm ²
32x TITANX 22nm & 11x TITANV 22nm	1010763.22 s ⁻¹	623.47 mm ²	0.19 W/mm ²
32x TITANX 22nm & 10x TITANV 22nm	1000000.37 s ⁻¹	608.97 mm ²	0.19 W/mm ²
32x TITANX 22nm & 9x TITANV 22nm	984894.15 s ⁻¹	594.47 mm ²	0.19 W/mm ²
36x TITANX 22nm & 9x TITANV 22nm	973082.21 s ⁻¹	652.47 mm ²	0.19 W/mm ²
32x TITANX 22nm & 6x TITANV 22nm	913045.07 s ⁻¹	550.98 mm ²	0.19 W/mm ²
36x TITANX 22nm & 5x TITANV 22nm	859585.22 s ⁻¹	594.49 mm ²	0.19 W/mm ²
23x TITANX 22nm & 13x KEPLER 22nm	694284.31 s ⁻¹	535.27 mm ²	0.18 W/mm ²
15x TITANX 22nm & 13x KEPLER 22nm	692128.08 s ⁻¹	419.27 mm ²	0.18 W/mm ²
10x TITANX 22nm & 13x KEPLER 22nm	678649.26 s ⁻¹	346.77 mm ²	0.18 W/mm ²
12x TITANX 22nm & 21x KEPLER 22nm	661929.57 s ⁻¹	499.94 mm ²	0.18 W/mm ²
10x TITANX 22nm & 21x KEPLER 22nm	654183.9 s ⁻¹	470.94 mm ²	0.18 W/mm ²
6x TITANX 22nm & 13x KEPLER 22nm	646629.97 s ⁻¹	288.77 mm ²	0.18 W/mm ²
10x TITANX 22nm & 24x KEPLER 22nm	639402.76 s ⁻¹	517.5 mm ²	0.18 W/mm ²
6x TITANX 22nm & 21x KEPLER 22nm	622164.61 s ⁻¹	412.94 mm ²	0.18 W/mm ²
5x TITANX 22nm & 24x KEPLER 22nm	594281.54 s ⁻¹	445.0 mm ²	0.18 W/mm ²
4x TITANX 22nm & 23x KEPLER 22nm	582710.35 s ⁻¹	414.98 mm ²	0.18 W/mm ²
4x TITANX 22nm & 24x KEPLER 22nm	578135.68 s ⁻¹	430.5 mm ²	0.18 W/mm ²
Back			

Figura 5. Resultado NSGAI

MultiExplorer			
Architecture	Performance	Area	Power Density
10x TITANX 22nm & 24x KEPLER 22nm	639402.76 s ⁻¹	517.5 mm ²	0.18 W/mm ²
6x TITANX 22nm & 21x KEPLER 22nm	622164.61 s ⁻¹	412.94 mm ²	0.18 W/mm ²
5x TITANX 22nm & 24x KEPLER 22nm	594281.54 s ⁻¹	445.0 mm ²	0.18 W/mm ²
4x TITANX 22nm & 23x KEPLER 22nm	582710.35 s ⁻¹	414.98 mm ²	0.18 W/mm ²
4x TITANX 22nm & 24x KEPLER 22nm	578135.68 s ⁻¹	430.5 mm ²	0.18 W/mm ²
2x TITANX 22nm & 13x KEPLER 22nm	571315.48 s ⁻¹	230.77 mm ²	0.18 W/mm ²
2x TITANX 22nm & 21x KEPLER 22nm	546850.11 s ⁻¹	354.94 mm ²	0.18 W/mm ²
24x TITANX 22nm & 23x GTX480 22nm	528898.12 s ⁻¹	881.43 mm ²	0.17 W/mm ²
15x TITANX 22nm & 24x GTX480 22nm	527336.9 s ⁻¹	565.43 mm ²	0.17 W/mm ²
15x TITANX 22nm & 27x GTX480 22nm	522450.83 s ⁻¹	608.92 mm ²	0.17 W/mm ²
13x TITANX 22nm & 28x GTX480 22nm	516292.91 s ⁻¹	594.42 mm ²	0.17 W/mm ²
12x TITANX 22nm & 28x GTX480 22nm	513865.1 s ⁻¹	579.92 mm ²	0.17 W/mm ²
10x TITANX 22nm & 24x GTX480 22nm	513858.08 s ⁻¹	492.93 mm ²	0.17 W/mm ²
5x TITANX 22nm & 21x GTX480 22nm	489693.03 s ⁻¹	376.94 mm ²	0.17 W/mm ²
4x TITANX 22nm & 21x GTX480 22nm	453547.17 s ⁻¹	362.44 mm ²	0.16 W/mm ²
10x TITANX 22nm & 9x GTX480 45nm	452052.95 s ⁻¹	640.76 mm ²	0.12 W/mm ²
5x TITANX 22nm & 9x GTX480 45nm	406931.73 s ⁻¹	568.26 mm ²	0.11 W/mm ²
4x TITANX 22nm & 9x GTX480 45nm	390785.87 s ⁻¹	553.76 mm ²	0.11 W/mm ²
Back			

Figura 6. Resultado NSGAI

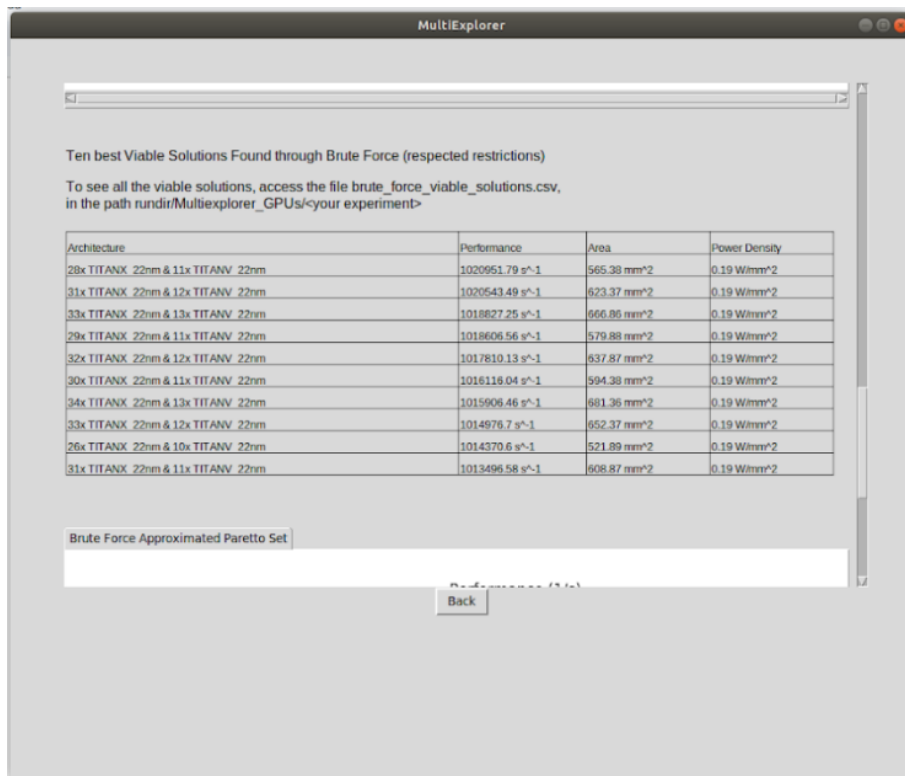


Figura 7. Resultado Força Bruta

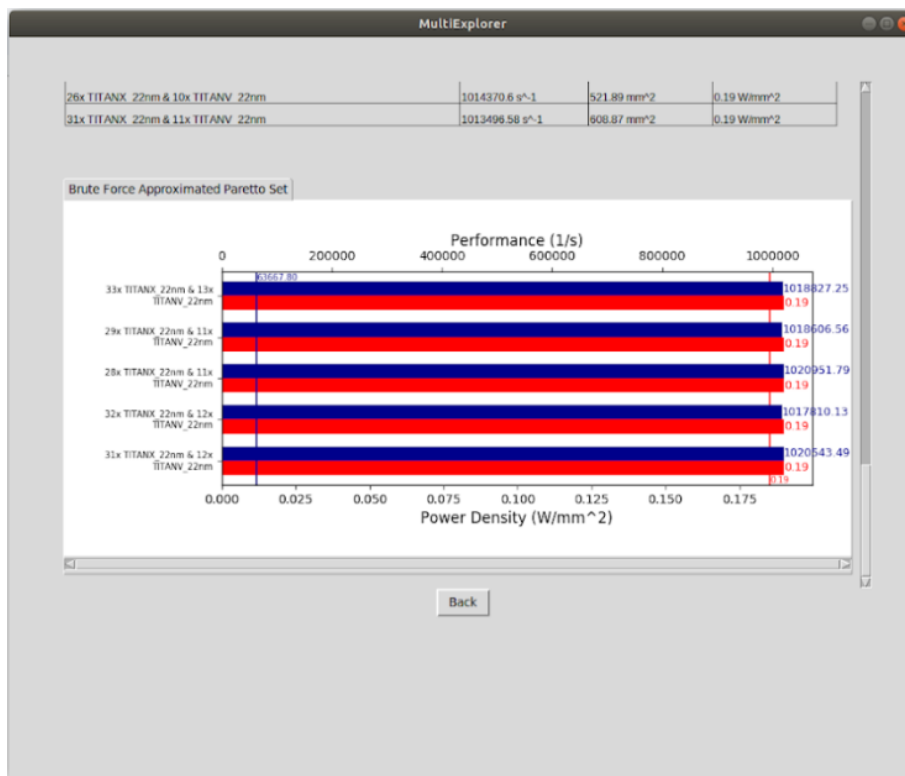


Figura 8. Resultado Força Bruta

de encontrar soluções de alta qualidade com um custo computacional significativamente menor.

7. Conclusão

Este Trabalho de Conclusão de Curso mostrou a extensão da ferramenta MultiExplorer para a exploração de projetos de sistemas GPGPU. Para alcançar esse objetivo, foi realizada a integração do simulador GPGPU-Sim ao fluxo do MultiExplorer. A interface gráfica da ferramenta também foi adaptada para o domínio das GPUs. O banco de núcleos do MultiExplorer, originalmente projetado para CPUs, foi estendido para incorporar núcleos de GPUs, caracterizados pelo seu desempenho, consumo e área.

Seis placas de GPU de cinco arquiteturas diferentes foram modelados e nove aplicações CUDA e Rodinia foram devidamente integradas à ferramenta. Os resultados das simulações e estimativas físicas de todas as placas foram apresentados. Por fim, a metodologia para estimativa de *dark-silicon* foi demonstrada em um processo de evolução tecnológica da placa TITANX.

Os resultados preliminares desse trabalho foram publicados em [Ludgero et al. 2024]. Posteriormente, o fluxo de DSE foi validado e uma versão mais completa será submetida na edição de 2025 do mesmo evento.

Como trabalhos futuros, cita-se a integração do fluxo para exploração para o domínio de IoTs e a atualização do McPAT para processos de fabricação de transistores abaixo de 22nm.

1

Referências

- Amorim, E. R., Santos, R., and Duenha, L. (2023). A tool for teaching and learning of design space exploration on processor architectures. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE.
- Arigoni, D., Santos, R., and Garanhani, L. (2022). Design space exploration of heterogeneous systems applied to the cloud resource allocation problem. In *Anais do XXIII Simpósio em Sistemas Computacionais de Alto Desempenho*, pages 169–180, Porto Alegre, RS, Brasil. SBC.
- Baghsorkhi, S. S., Delahaye, M., Patel, S. J., Gropp, W. D., and Hwu, W. (2010). An adaptive performance modeling tool for gpu architectures. In *Proceedings of the 15th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*, pages 105–114. ACM.
- Bienia, C., Kumar, S., Singh, J., and Li, K. (2008). The parsec benchmark suite: Characterization and architectural implications. In *Proceedings of the 17th International Conference on Parallel Architectures and Compilation Techniques*, pages 72–81. ACM.
- Che, S., Boyer, M., Meng, J., Tarjan, D., Sheaffer, J. W., Lee, S.-H., and Skadron, K. (2009). Rodinia: A benchmark suite for heterogeneous computing. In *2009 IEEE International Symposium on Workload Characterization (IISWC)*, pages 44–54. IEEE.

¹ Agradecemos ao CNPq e à CAPES pelo financiamento desta pesquisa.

- Khairy, M., Shen, Z., Aamodt, T. M., and Rogers, T. G. (2020). Accel-sim: An extensible simulation framework for validated gpu modeling. In *Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA '20)*, pages 473–486. IEEE Press.
- 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. *ACM SIGARCH Computer Architecture News*, 41(3):487–498.
- 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.
- Ludgero, L. B., Damasceno, G. G., Rodrigues, S. S., dos Santos, R. R., and Duenha, L. D. (2024). Exploração do espaço de projetos de sistemas gpgpus ciente de dark-silicon. In *Simpósio em Sistemas Computacionais de Alto Desempenho (WIC-SSCAD)*. SBC.
- Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85.
- NVIDIA (2021). Cuda toolkit documentation. <https://docs.nvidia.com/cuda/cuda-samples/index.html#cudalibraries>.
- Rodrigues, S. and Santos, R. (2024). Extensão da ferramenta multiexplorer para exploração de projetos de gpus e máquinas virtuais. In *Anais do 25º Workshop de Iniciação Científica do Simpósio em Sistemas Computacionais de Alto Desempenho (WIC-SSCAD)*. SBC.
- Santos, R., Duenha, L., Silva, A. C., Sousa, M., Tedesco, L. A., Melgarejo, J. C., 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., dos Santos, R. R., and Duenha, L. D. (2023). Performance predictors for graphics processing units applied to dark-silicon-aware design space exploration. *Concurrency and Computation: Practice and Experience*, 35(17):e6877.
- Zaccaria, V., Palermo, G., Castro, F., Silvano, C., and Mariani, G. (2020). Multicube explorer: An open source framework for design space exploration of chip multi-processors. In *23rd International Conference on Architecture of Computing Systems (ARCS)*, pages 1–7. VDE.