Exploração do espaço de projeto em arquiteturas heterogêneas cientes de *dark silicon* e utilizando computação aproximada

Daniela Catelan, Ricardo Santos

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

{daniela, ricardo}@facom.ufms.br

Resumo. *O problema de dark silicon surgiu com o incremento da corrente de fuga (leakage current) em consequência da miniaturização dos transistores. Pesquisas a fim de encontrar soluções para mitigar o dark silicon têm sido estudadas, muitas delas propondo a heterogeneidade de dispositivos de processamento. Contudo, o aumento da diversidade de dispositivos e dos objetivos na definição de sistemas heterogêneos e de alto desempenho, tornam o projeto de tais sistemas mais complexo, exigindo mecanismos automatizados de exploração de espaço de projeto cientes de dark silicon. Uma solução promissora é a utilização da computação aproximada, na qual, componentes de hardware e software utilizam a aproximação ao invés da precisão das operações, aceitando perda de qualidade de saída para melhorar a eficiência energética e obter ganhos de desempenho. Este trabalho objetiva obter soluções eficientes para o problema de exploração de projetos de processadores cientes de dark silicon, utilizando módulos de computação aproximada como elementos factíveis de um sistema computacional heterogêneo.*

1. Introdução

O problema de limitações físicas ocasionadas pelo incremento da corrente de fuga, em consequência da miniaturização dos transistores, fizeram com que surgisse o *dark silicon*, que é uma porção de área do *chip* que deve ser desligada ou funcionar em frequência aquém do estipulado. Para manter a viabilidade do projeto, pesquisas a fim de encontrar soluções para mitigar o *dark silicon* têm sido estudadas, muitas delas propondo a heterogeneidade de dispositivos de processamento.

O uso dos recursos heterogêneos tem tornado o projeto de um sistema computacional mais complexo e exigente de ferramentas de projeto que consigam lidar com características heterogêneas para os processadores com diversas variáveis de restrições. Neste ponto, há uma oportunidade para implementação de novos algoritmos de exploração de espaço de projeto que possibilitem a determinação de recursos arquiteturais para utilizar a área de *dark silicon* mantendo as restrições físicas e de desempenho do projeto.

A computação aproximada (CA) tem sido uma opção para a investigação em diferentes níveis de abstração, desde as propriedades físicas dos transistores, as funções lógicas de operadores aritméticos até as alterações nas arquiteturas dos sistemas computacionais. Deste modo, torna-se mais flexível à relação entre a implementação e a especificação em um sistema de computação. Ao trocar a exatidão dos resultados numéricos pela redução no consumo energético na área utilizada ou no atraso, visto que esta é uma das técnicas para a exploração de computação aproximada, é possível realizar o

ajuste de tensão de alimentação além dos limites considerados seguros para a manutenção de um resultado preciso [Chippa et al. 2014].

Este trabalho pretende realizar uma investigação científica sobre soluções para o problema de projetos de processadores cientes de *dark silicon*, com soluções eficientes de *Dark Silicon aware - Design Space Exploration* (DS-DSE) utilizando computação aproximada. Vislumbra-se utilizar, como plataforma para desenvolvimento do projeto, a ferramenta MultiExplorer com ênfase na adequação da mesma para dar suporte ao trabalho, ressaltando que a ferramenta MultiExplorer foi proposta em 2014 pelo grupo LSCAD e encontra-se ainda em desenvolvimento. Com isto, este trabalho além de atuar sobre uma ferramenta ainda em desenvolvimento irá contribuir para a sua expansão.

2. Exploração do Espaço de Projeto

Com o aumento da complexidade dos multiprocessadores e a variedade dos parâmetros arquiteturais a serem explorados no momento do projeto para encontrar a melhor solução entre os vários objetivos concorrentes (como energia, atraso, largura de banda, etc.), o espaço de projeto é enorme. Deve-se levar em consideração todas as combinações possíveis de cada parâmetro (número de processadores, largura de emissão do processador, tamanhos de cache, etc.). Faz-se necessário explorar o espaço de projeto visando obter a melhor configuração da arquitetura de hardware para uma determinada aplicação.

Usualmente o resultado da exploração não será uma única solução, mas um conjunto de soluções que compõem o conjunto de Pareto [1]. Em projetos de processadores, a etapa de exploração consiste em um problema de otimização multiobjetivo e frequentemente utiliza métodos heurísticos para sua resolução. Logo, o processo de otimização será composto por vários parâmetros de sistemas e microarquiteturas que envolvem a minimização ou maximização de múltiplos objetivos, tornando a otimização não exclusiva.

Relacionado a este trabalho de doutorado foi desenvolvido um artigo [Santos et al. 2019] que 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, integrada ao *framework* MultiExplorer. Foi desenvolvida concomitantemente, uma estratégia para a 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*, possibilitando, assim, a exploração do espaço de projeto buscando arquiteturas heterogêneas GP-GPUs.

3. Computação Aproximada

As técnicas de circuito aproximado, no campo de pesquisa têm muitas e diferentes propostas ([Gupta et al. 2013], [Yang et al. 2013], [Almurib et al. 2016], [Jiang et al. 2015], [Gorantla and Deepa 2019]), focando principalmente na aritmética aproximada, como somadores e subtratores. Alguns estudos apresentam circuitos com apenas um bit,

[1] O trabalho original de Pareto é "Cours d'economie politique, F. Rouge, Lausanne, 1896"

mostrando seus detalhes como posicionamento de erro, uso da área, potência e atraso. A maioria das pesquisas apresenta apenas um único circuito onde o objetivo é controlar a precisão. Como exemplo, [Muthulakshmi et al. 2018] apresenta quatro tipos de subtratores (APSC4-APSC7), onde a diferença entre eles é apenas a posição onde o erro foi inserido na tabela verdade.

Ao avaliar pequenos circuitos, esses trabalhos perdem a oportunidade de analisar o comportamento dessas soluções de computação aproximada na presença de longas entradas e saídas e até mesmo em uma plataforma de projeto do mundo real.

Uma Unidade Lógica Aritmética (ULA) desempenha um papel importante no desempenho do processador, pois é responsável por executar a maioria das instruções de um programa. Além disso, os circuitos aritméticos usam grandes porções da área e potência do projeto do hardware [Sassi 2013]. Aplicando técnicas de circuitos aproximados em um projeto de ULA, os programas serão capazes de usar técnicas aproximadas, economizando assim o consumo de energia com algum custo de precisão de instrução.

Também no âmbito desta proposta de doutorado, realizou-se um estudo detalhado sobre circuitos aritméticos aproximados projetados em uma plataforma *field-programmable gate array* (FPGA) com o objetivo de explorar como circuitos aproximados poderiam ser ajustados a plataformas de prototipagem amplamente utilizadas e flexíveis, como FPGAs, mas com menos flexibilidade no gerenciamento de energia e área. Os resultados mostraram que alguns circuitos estão bem adaptados às plataformas FPGA e podem tirar vantagem de sua organização melhor do que outros.

Ao avaliar esses circuitos considerando precisão, área e dissipação de potência, tem-se o objetivo de caracterizar circuitos aritméticos aproximados mostrando que, por um lado, existe uma compensação entre precisão e parâmetros físicos e, por outro lado, circuitos aproximados são dependentes de hardware projetados para fornecer melhor área e economia de energia.

4. Conclusão

Com a necessidade de mitigar o *dark silicon*, este trabalho pretende realizar uma investigação científica sobre soluções para o problema de projetos de processadores cientes de *dark silicon*, utilizando soluções eficientes de DS-DSE utilizando computação aproximada

Para embasar este trabalho, já foram realizados experimentos iniciais com circuitos aproximados somadores, subtratores e multiplicadores, caracterizados com base na precisão, área e métricas de dissipação de potência, em tamanhos que variam de 8 a 64 bits, comparados com os circuitos exatos, servem de base para a utilização da técnica da CA.

Como exemplo, o circuito AMA1 misto (circuitos somadores exatos junto com circuitos aproximados) tem um erro relativo de $54,7\%$ com 8 bits e $99,5\%$ de erro relativo com 64 bits. Considerando o uso da área, o circuito somador aproximado AMA4 tem um uso de apenas $29,1\%$.

Nossos resultados preliminares mostram que os circuitos de aproximação projetados sob medida trazem níveis úteis de precisão, uso de área e dissipação de energia em circuitos pequenos, mas os benefícios são reduzidos para circuitos com muitas entradas.

Os resultados da caracterização física mostram que os melhores resultados no uso da área e dissipação de energia dependem da plataforma de hardware. Uma plataforma de prototipagem rápida como FPGA pode não trazer a eficiência de área e energia que seriam necessárias para obter os benefícios de um grande circuito aproximado.

Nossos trabalhos futuros objetivam aprofundar o estudo e experimentação com hardwares aproximados, com enfoque em circuitos aritméticos mais complexos como multiplicadores, divisores e aritmética de ponto-flutuante. Adicionalmente, ênfase será dada no estudo teórico da fronteira de Pareto aplicado à exploração de projetos, a modelagem matemática e técnicas eficientes de resolução desse problema de exploração.

References

Almurib, H. A. F., Kumar, T. N., and Lombardi, F. (2016). Inexact designs for approximate low power addition by cell replacement. In *Proceedings of the 2016 Conference on Design, Automation & Test in Europe*, DATE '16, page 660–665, San Jose, CA, USA. EDA Consortium.

Chippa, V., Mohapatra, D., Roy, K., Chakradhar, S., and Raghunathan, A. (2014). Scalable effort hardware design. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 22:2004–2016.

Gorantla, A. and Deepa, P. (2019). Design of approximate subtractors and dividers for error tolerant image processing applications. *Journal of Electronic Testing*, pages 1–7.

Gupta, V., Mohapatra, D., Raghunathan, A., and Roy, K. (2013). Low-power digital signal processing using approximate adders. *IEEE Trans. on CAD of Integrated Circuits and Systems*, 32(1):124–137.

Jiang, H., Han, J., and Lombardi, F. (2015). A comparative review and evaluation of approximate adders. In *Proceedings of the 25th Edition on Great Lakes Symposium on VLSI*, GLSVLSI '15, page 343–348, New York, NY, USA. Association for Computing Machinery.

Muthulakshmi, S., Dash, C., and Prabaharan, S. (2018). Memristor augmented approximate adders and subtractors for image processing applications: An approach. *AEU - International Journal of Electronics and Communications*, 91.

Santos, R., Sonohata, R., Krebs, C., Catelan, D., Duenha, L., Segovia, D., and Santos, M. (2019). Exploração do projeto de sistemas baseados em gpu ciente de dark silicon. In *Anais Principais do XX Simpósio em Sistemas Computacionais de Alto Desempenho*, pages 358–369, Porto Alegre, RS, Brasil. SBC.

Sassi, A. B. (2013). *Projeto de uma ULA de inteiros e de baixo consumo em tecnnologia CMOS*. PhD thesis, Escola de Engenharia de São Carlos da Universidade de São Paulo.

Yang, Z., Jain, A., Liang, J., Han, J., and Lombardi, F. (2013). Approximate xor/xnor-based adders for inexact computing. *Proceedings of the IEEE Conference on Nanotechnology*, pages 690–693.