Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Dissertação de Mestrado

Área: Inteligência Computacional

Título: Modelagem Preditiva do Comportamento Térmico de GPUs sob Cargas de Trabalho de Deep Learning

Orientador: Sérgio Murilo Maciel Fernandes (sergio.fernandes@upe.br)
Coorientador: Sidney Marlon Lopes de Lima (sidney.lima@ufpe.br)

Descrição – A ascensão de Large Language Models (LLMs) impôs uma pressão sem precedentes sobre a infraestrutura de hardware, transformando o consumo de energia e a dissipação térmica em gargalos críticos para a escalabilidade da Inteligência Artificial [2]. Diferente de cargas de trabalho computacionais tradicionais, o treinamento de modelos massivos gera flutuações de potência que levam as GPUs ao seu limite térmico (Thermal Design Power - TDP), disparando frequentemente o Thermal Throttling [2], [4]. Este fenômeno reduz a frequência de operação dos núcleos, resultando em uma degradação severa da vazão de processamento e aumentando significativamente o custo operacional (TCO) de grandes data centers, por exemplo [2], [4].

Historicamente, a modelagem térmica dependia de simulações de Dinâmica de Fluidos Computacional (CFD), que, embora precisas, possuem um custo computacional proibitivo para decisões em tempo real [5]. Trabalhos recentes de alto impacto propõem a transição para modelos substitutos (*surrogate models*) baseados em Deep Learning, como os Operadores Neurais de Fourier (FNO), que conseguem prever distribuições térmicas 3D com uma velocidade aos métodos tradicionais [5]. Além disso, a integração de leis da física com aprendizado de dados (técnicas como LEnPOD-GP) permitiu o mapeamento de *hotspots* dinâmicos em GPUs de muitos núcleos (como a arquitetura NVIDIA Volta/Hopper) [1].

Apesar desses avanços, a literatura recente aponta que o gerenciamento térmico reativo é insuficiente para lidar com a volatilidade de modelos de IA modernos [2], [4]. Pesquisas publicadas em fóruns de elite sugerem que o escalonamento de tarefas consciente da temperatura (*Thermal-aware Scheduling*) pode reduzir o pico térmico em até 12°C e os custos de resfriamento pela metade [3], [6]. No entanto, a fusão eficiente entre telemetria de hardware em tempo real (obtida via NVML) e o comportamento estocástico das fases de treinamento (como *feedforward* vs *backpropagation*) ainda carece de modelos preditivos robustos e generalizáveis para ambientes privados, nuvem e edge [3], [6].

Este projeto propõe investigar e desenvolver um *framework* de modelagem preditiva que utilize redes neurais recorrentes (LSTMs) para antecipar o comportamento térmico de GPUs sob cargas de DL. A pesquisa focará na correlação entre hiperparâmetros de software (como *batch size*) e o estresse térmico do hardware [2], [4]. Espera-se que a solução proposta permita uma orquestração proativa, otimizando a eficiência energética (Performance-per-Watt) e estendendo a vida útil dos hardwares de alto desempenho [3], [5].

Referências Bibliográficas

[1] L. Jiang, Y. Liu, and M.-C. Cheng, “Effective thermal modeling for large-scale many-core gpus using local physics-based data-learning approach,” Structural and Multidisciplinary Optimization, vol. 68, 2025.

[2] P. Patel, E. Choukse, C. Zhang, Í. Goiri, B. Warrier, N. Mahalingam, and R. Bianchini, “Characterizing power management opportunities for llms in the cloud,” Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3, 2024.

[3] S. Ilager, K. Ramamohanarao, and R. Buyya, “Thermal prediction for efficient energy management of clouds using machine learning,” IEEE Transactions on Parallel and Distributed Systems, vol. 32, pp. 1044–1056, 2020.

[4] J. Stojkovic, C. Zhang, Í. Goiri, E. Choukse, H. Qiu, R. Fonseca, J. Torrellas, and R. Bianchini, “Tapas: Thermal- and power-aware scheduling for llm inference in cloud platforms,” Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2, 2025.

[5] S. Sarkar, A. Guillen-Perez, Z. Carmichael, A. Naug, R. M. Çam, V. Gundecha, A. R. Babu, S. Ghorbanpour, and R. L. Gutiérrez, “Fast 3d surrogate modeling for data center thermal management,” 2025.
[6] T. Tan and G. Cao, “Thermal-aware scheduling for deep learning on mobile devices with npu,” IEEE Transactions on Mobile Computing, vol. 23, pp. 10706–10719, 2024.