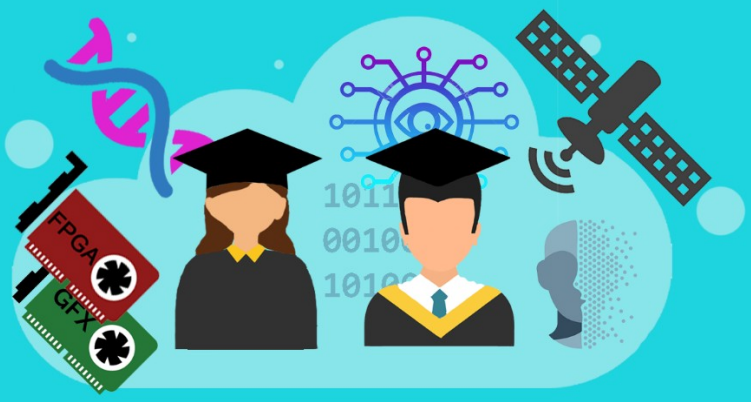
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Multi-Core In-Memory Computing / Near-Threshold Computing Modeling

Specialized hardware architectures have evolved alongside the rapid growth of Deep Neural Networks (DNNs) to effectively accelerate their inference. By utilizing dedicated computing clusters and optimizing data sharing and reuse, they have managed to significantly outperform conventional cores [1]. Recent designs are transitioning from single specialized cores to multi-core architectures that offer increased compute parallelism and improved dataflow flexibility [2]. This enables designers to map Machine Learning (ML) workloads across various accelerators, enhancing the overall system's efficiency.

Yet, memory interfacing remains a performance-limiting aspect of modern neural accelerators. Compute-In-Memory (CiM) presents a promising solution for accelerating DNNs by minimizing energy-intensive weight movements and utilizing memory arrays to enable low-energy, high-density computations [3]. This significantly minimizes access overheads and unlocks extensive parallelization opportunities, potentially leading to orders-of-magnitude improvements in energy efficiency and throughput [4].

Considering DNNs' inherent tolerance to errors, investigating strategies that leverage this property to enhance performance or reduce energy consumption offers significant potential. Low-voltage computing, and especially near-threshold voltage computing (NTC), offers substantial energy efficiency gains by operating transistors near their threshold voltage. However, the implementation of NTC also introduces challenges, including increased parametric variation, a higher failure rate, and potential performance degradation [5, 6].

The diploma thesis focuses on the following areas (but is not limited to):

- Modeling the performance of IMC accelerators within heterogeneous multi-core systems.
- Assessing the effects of near-threshold voltages on IMC accelerators across various technology scales (e.g., 45nm, 22nm).
- Determining the optimal operating voltage for each accelerator (especially IMC accelerators) in a multi-core system to enhance energy efficiency, while ensuring that DNNs' accuracy remains at high levels.
- Identifying fault-tolerant layers within DNN workloads to enable their effective mapping onto NTC IMC cores.

PREREQUISITES:

- Experience with Python
- Familiarity with Machine Learning / Deep Neural Networks
- Basic knowledge of Digital Design

RELATED MATERIAL:

[1] Y. Chen et al., “A survey of accelerator architectures for deep neural networks,” Engineering, vol. 6, no. 3, pp. 264–274, 2020.

[2] A. Symons, L. Mei, S. Colleman, P. Houshmand, S. Karl, and M. Verhelst, "Stream: A modeling framework for fine-grained layer fusion on multi-core DNN accelerators," in *2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2023, pp. 355–357.

[3] K. Asifuzzaman, N. R. Miniskar, A. R. Young, F. Liu, and J. S. Vetter, "A survey on processing-in-memory techniques: Advances and challenges," *Memories-Materials, Devices, Circuits and Systems*, vol. 4, pp. 100022, 2023.

[4] P. Houshmand, J. Sun, and M. Verhelst, "Benchmarking and modeling of analog and digital SRAM in-memory computing architectures," arXiv preprint arXiv:2305.18335, 2023. [Online]. Available: https://arxiv.org/abs/2305.18335.

[5] Sparsh Mittal. 2016. A Survey of Architectural Techniques for Near-Threshold Computing. J. Emerg. Technol. Comput. Syst. 12, 4, Article 46 (July 2016), 26 pages. https://doi.org/10.1145/2821510

[6] S. Kalra and K. Singh, "Performance evaluation of near-threshold ultradeep submicron digital CMOS circuits using approximate mathematical drain current model," *Journal of Integrated Circuits and Systems*, vol. 19, no. 2, pp. 1–14, 2024.

CONTACT INFORMATION:

- Alexis Maras Ph.D Student, NTUA: (amaras@microlab.ntua.gr)
- Panagiotis Chaidos, Junior Researcher , NTUA: (pchaidos@microlab.ntua.gr)
- Professor Dimitrios Soudris, NTUA: (dsoudris@microlab.ntua.gr)
- Assistant Professor Sotirios Xydis, NTUA: (sxydis@microlab.ntua.gr)