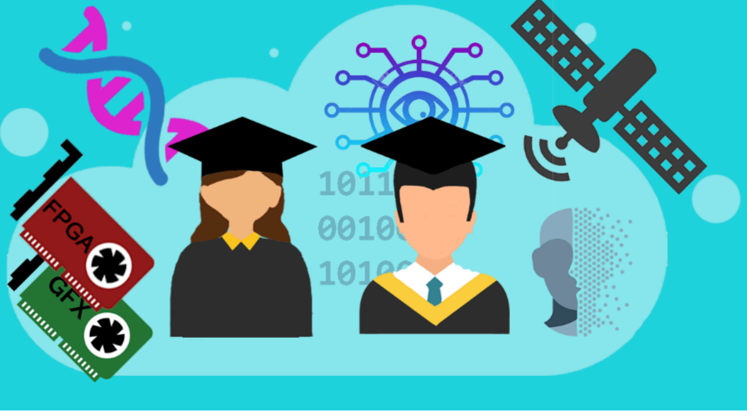
Diploma Thesis

Microproccessors and Digital Systems Laboratory



Performance and Power Modeling of Embedded AI Computing Platforms

The rapid advancements in AI and the proliferation of IoT devices have led to the emergence of Edge AI, where AI models are processed locally on devices at the edge of the network, rather than in centralized cloud servers. This decentralization minimizes latency, reduces bandwidth usage, and enhances data privacy. Edge AI is commonly deployed on embedded devices such as microcontrollers, sensors, and AI-optimized hardware like NVIDIA's embedded GPUs and specialized AI accelerators. This thesis focuses on the latter. These devices are typically resource-constrained but capable of performing complex AI tasks like image recognition, natural language processing, and predictive analytics. This makes Edge AI an attractive solution for applications in areas such as autonomous vehicles, smart homes, healthcare, and industrial automation, where speed, privacy, and power efficiency are critical.

In order to enable reliable architecture Design Space Exploration in the field of Embedded AI Computing Platforms, several parameters must be taken into account, most notably **performance and power consumption**. The above platforms usually consist of a multi-core CPU integrated with a specialized GPU. While performance and power models exist for both in isolation they: a) Are not tuned for the above architectures b) Do not function synergistically. To this end, the purpose of this thesis is **to tune the respective models on the above platforms' specialized components and integrate the resulting CPU and GPU models** in order to provide these metrics for the execution of whole Machine Learning workloads.

Accel-Sim and AccelWattch are state-of-the-art cycle-accurate tools for GPU performance and power modeling respectively. Accel-Sim is a configurable framework for detailed, validated GPU performance simulation across GPU generations. AccelWattch, integrated with Accel-Sim, models dynamic, static, and constant power, capturing complex power gating. **Gem5** is a modular, open-source cycle-accurate simulator for processors, including ARM, x86, RISC-V, and MIPS. It models CPUs, memory systems, and full-system environments, supporting performance, power efficiency, and design trade-off analysis.

PREREQUISITES:

Familiarity with: C++, Python.

Desirable qualifications: CUDA, Bash scripting.

RELATED MATERIAL:

https://info.nvidia.com/rs/156-OFN-742/images/Jetson_AGX_Xavier_New_Era_Autonomous_Machines.pdf, https://www.gem5.org/, https://ieeexplore.ieee.org/document/9138922, https://paragon.cs.northwestern.edu/papers/2021-MICRO-AccelWattch-Kandiah.pdf,

CONTACT INFORMATION:

Asst. Prof. Sotirios Xydis (sxydis@microlab.ntua.gr)

Konstantinos Iliakis (kiliakis@microlab.ntua.gr)

Panagiotis Eleftherakis (pelef@microlab.ntua.gr)