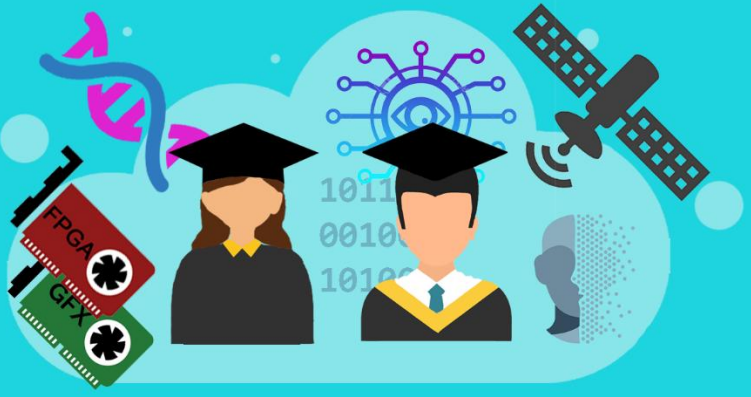
Diploma Thesis

Microproccessors and Digital Systems Laboratory

&



HUAWEI

Efficient Use of Hardware Accelerators for LLMs

Huawei Paris Research Center - Internship Proposition

September 30, 2025

Context

Huawei is a leading global information and communications technology solutions provider. Through our dedication to customer-centric innovation and strong partnerships, we have estab- lished end-to-end advantages in telecom networks, devices and cloud computing. We are committed to creating maximum value for telecom operators, enterprises and consumers by providing compet- itive solutions and services. Our products and solutions have been deployed in over 170 countries, serving more than one third of the world's population. The **Huawei Paris Research Center** is responsible for advanced research in the fields of algorithm and software design, 5G, networks, aesthetics, to create and design the innovative technologies and software platforms for our brand.

1 Introduction

The ever-increasing requirements of LLM (Large Language Model) deployment establish the need for high-performance solutions. While aggressively optimized libraries and frameworks are at the heart of state-of-the-art solutions for high-performance LLMs, recent advances also rely on minute alterations of the original designs. In consequence, novel optimizations for LLMs are twofold: new high-level ideas can be explored on the condition that low-level efficient implementations can follow suit. By essence, such research directions require specialized operations that can not be (efficiently) covered by existing software. Moreover, an additional challenge lies in the need to target multiple hardware architectures. To that effect, recent research has been focusing on various methods for approximating computation LLMs. In particular, one such class of approximation is to introduce sparsity at various levels, for instance during Attention computation [JLZ+24]. The purpose of this internship is to join and assist local efforts towards high-performance LLMs, especially on Ascend NPUs. This means participating in the design or implementation of novel approximation ideas with due consideration for the target hardware's

constraints and particularities.

2 Scope of Work

The student may:

- study state-of-the-art research on optimizations for LLMs
- suggest new ideas to improve LLM performance
- adapt existing or new algorithms to specific accelerator architectures
- optimize existing accelerator kernels or implement new ones for specific accelerator architectures
- run end-to-end experiments for validation

3 Skills Required

- Excellent knowledge of parallel programming paradigms and frameworks (appreciated)
- Experience on hardware accelerators programming (GPU, NPU, FPGA, ASIC, etc.) (appreci- ated)
- Excellent skills in C++17 or later and Python (essential)
- Experience with modern AI frameworks (PyTorch, MindSpore, etc.)
- Solid knowledge of build systems (e.g. cmake), version control tools (e.g. git) and software development conventions
- Proficiency in English (essential) and French (appreciated)

4 Duration

6 months

References

[JLZ+24] Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H Abdi, Dongsheng Li, Chin-Yew Lin, et al. Minference 1.0: Accel- erating pre-filling for long-context llms via dynamic sparse attention. *Advances in Neural Information Processing Systems*, 37:52481–52515, 2024.

Contact Information:

Dimosthenis Masouros Ph.D.: (dmasouros@microlab.ntua.gr)

Prof. Dimitrios Soudris: (dsoudris@microlab.ntua.gr)