Laboratoire de l'Informatique du Parallélisme
(UMR CNRS - ENS Lyon - UCB Lyon 1 - INRIA 5668)

équipe ROMA

master internship proposal

Analytical prediction of LLM inference throughput

Loris Marchal

Encadrant: Loris Marchal (senior researcher, CNRS, `Loris.Marchal@ens-lyon.fr`)

Lieu: Laboratory LIP, École Normale Supérieure de Lyon, team ROMA.

Team

The internship will take place within the ROMA team at the Laboratoire de l'Informatique du Parallélisme (UMR CNRS - ENS Lyon - UCB Lyon 1 - INRIA 5668), under the supervision of Loris Marchal. The ROMA team focuses on the design of parallel algorithms and scheduling for distributed computing platforms, particularly for scientific applications and LLM inference.

The internship will take place in the context of a collaboration with the Greenference company[1], which targets the conception of datacenters directly powered by solar energy to answer the growing demand for inference. Frequent discussions with Vincent Abad, co-founder of Greenference, will take place to feed the research with actual hardware and inference market knowledge.

Scope

Artificial Intelligence is becoming an essential tool in many fields, leading to a growing demand for Large Language Model (LLM) inference. As more users and applications rely on AI, the need for computing resources is increasing rapidly. Meeting this demand requires specific hardwares, hence many accelerators have recently been released to answer this demand, both from major manufacturers such as NVIDIA and AMD and from other hardware providers. However, predicting the inference performance of these platforms is a challenging task, even for LLMs based on standard Transformer architectures. The performance of a model, often measured in tokens per second, depends on many factors, and not only the sheer number of model parameters. This makes it difficult to estimate both the memory footprint and the inference speed of a given model on a given hardware platform.

Furthermore, LLM architectures are evolving rapidly, both to improve their accuracy and to increase their inference speed. New techniques and architectural changes, such as mixture of experts [3, 4], structured state space models [2] and multi-token prediction [1], aim to improve efficiency and reduce memory requirements. While these approaches can significantly improve inference performance, they also make performance prediction more complex. The memory footprint and computational requirements can vary greatly depending on the model architecture and the hardware platform.

The goal of this research internship is to develop a model that can predict the inference performance of LLMs across different hardware platforms. The main idea is one on hand to describe the computing platform and the inference engine using a small set of simple parameters that are easy to measure, and on the other size to carefully analyze the data flow of targeted LLM architectures. This approach should make it possible to predict the performance of a given model without requiring a detailed understanding of every hardware and software component for a given LLM.

A secondary objective is to determine whether specific inference engine optimizations can improve performance in a given situation. For example, techniques such as Tensor Parallelism and Model Parallelism can have different effects depending on the model, the hardware platform, and the workload. The prediction model should help identify when these optimizations are likely to be beneficial and provide a better understanding of the trade-offs involved.

[1] `greenferece.com`

The work will consist of the following steps:

1. Studying of LLM inference and existing performance models,
2. Characterization of hardware platforms and inference engines,
3. Development of a performance prediction model,
4. Experimental validation and performance model evaluation,
5. Evaluation of inference optimizations.

Skills

The intern must have strong skills in algorithms and programming. An understanding of machine learning, large language models, and proficiency in the classic tools used to implement them will be significant assets.

Opportunities and perspectives

The intern will gain important knowledge in a research domain that is of critical importance both for companies and academic research. She/he will interact both with academic researchers in top research groups, at ENS-Lyon and possibly through the collaboration with Inria Bordeaux and Mc Gill University through our associate team[2]), and with the industrial world through the collaboration with the Greenference company. Depending on how the internship goes, the work can be continued during a PhD thesis.

References

[1] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction, 2024.

[2] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024.

[3] Haiyang Huang, Newsha Ardalani, Anna Sun, Liu Ke, Hsien-Hsin S. Lee, Anjali Sridhar, Shruti Bhosale, Carole-Jean Wu, and Benjamin Lee. Towards MoE deployment: Mitigating inefficiencies in mixture-of-expert (MoE) inference, 2023.

[4] Zhongkai Yu, Yue Guan, Zihao Yu, Chenyang Zhou, Zhengding Hu, Shuyi Pei, Yangwook Kang, Yufei Ding, and Po-An Tsai. Patterns behind chaos: Forecasting data movement for efficient large-scale MoE LLM inference, 2026.

[2] https://perso.ens-lyon.fr/loris.marchal/ACACIA-EA.html