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

équipe ROMA

internship proposal

Expert allocation for the inference of large language models

Loris Marchal & Olivier Beaumont

Encadrants: Loris Marchal (senior researcher, CNRS, Loris.Marchal@ens-lyon.fr) Olivier Beaumont (senior researcher, Inria Bordeaux, Olivier.Beaumont@inria.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. The internship will also be supervised by Olivier Beaumont, from the Inria Topal team in Bordeaux. Topal develops efficient algorithms and tools for linear and tensor algebra and for machine learning.

Scope

Text generation with large language models (LLMs) requires significant computing power and memory. The Mixture of Experts (MoE) paradigm has been proposed to reduce the amount of data required for inference [3]: each layer of the model is composed of several experts, and only a small number of them are used to produce a token. The DeepSeek model thus has 64 experts in each layer, only 8 of which are active in producing a token [2].

When inferring on a distributed platform consisting of several GPUs, the experts can be distributed among the GPUs and possibly replicated. We generally seek to process several text generation requests (or *prompts*) in parallel to improve computational efficiency, but the number of requests is limited by the memory required to store their context.

The objective of this internship is to optimize the distribution of experts across GPUs, their possible replication, and the allocation of experts needed for different queries on GPUs, in order to improve generation throughput. In an initial study [1], we showed that using the frequency of simultaneous use of subsets of experts makes it possible to increase the level of parallelism for a single request, and therefore the inference throughput. The objective of this internship is to extend this study when multiple queries are processed simultaneously, which is both a more realistic use case and allows for increased inference performance. This will involve identifying trade-offs between inference throughput and latency for each of these queries.

The work will consist of:

- Precisely modeling the optimization problem related to the distribution of experts and the allocation of queries, as well as conducting bibliographic research on existing strategies;
- Proposing effective methods, possibly with performance guarantees, starting with simple cases (a single request and/or no replication) and then generalizing them;
- Testing the proposed solutions in simulation, on traces of expert activation obtained by running the corresponding models on existing datasets.
- Possibly implement some of these strategies in an existing inference system.

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. Preliminary knowledge of operations research and/or high-performance computing will also be appreciated.

Collaborations and perspectives

The internship will take place in Lyon, at LIP, but will be co-supervised by Olivier Beaumont (Inria Bordeaux). Discussions with Olivier Beaumont will initially take place via video conference, then visits may be organized to work together. Depending on how the internship goes, the work can be continued during a PhD thesis.

We are working on these topics in collaboration with Oana Balmau and Mark Coates from McGill University in Montreal (Canada). If the intern continues with a thesis, it may be carried out under the joint supervision of McGill University and ENS Lyon.

References

[1] Olivier Beaumont, Raphaël Bourgouin, Maxime Darrin, Loris Marchal, and Pablo Piantanida. Leveraging expert usage to speed up LLM inference with expert parallelism. In *31st European Conference on Parallel and Distributed Processing (Euro-Par)*, volume 15900 of *Lecture Notes in Computer Science*, pages 145–158. Springer, 2025.

[2] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

[3] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *ICML 2022*, 2022.