NEWS LETTER

14

October | 2024

CLOUD STARS

Pol Garcia, at IBM-NY research offices in Manhattan, NY

Accelerating Small Language Models in the Cloud

by Pol García, Jordi Torres, Alberto Gutiérrez, Josep Ll. Berral, Ramon Nou
(Barcelona Supercomputing Center – Universitat Politècnica de Catalunya)

*The collaboration between the Barcelona Supercomputing Center and IBM T.J. Watson research center in NY leads to advances in GPU management and optimization towards Small Language Models and other potential AI applications in the Cloud. The BSC student, Pol Garcia, published recently the study "**Towards Pareto Optimal Throughput in Small Language Model Serving**" in the 4th Workshop on Machine Learning and Systems (EuroMLSys), focusing on the use and benchmarking of GPUs towards Cloud AI-oriented services.*

In the fall of 2023, the Barcelona Supercomputing Center (BSC) initiated its first round of research secondments at the IBM Thomas J. Watson Research Center in New York. Pol Garcia, PhD student at BSC has been collaborating since then with the "Cloud Native AI Platform" research team, optimizing the inference performance of language models in AI deployments. This joint effort has resulted in a scientific publication and presentation at EuroMLSys '24 workshop, a specialized venue in conjunction with EuroSys 2024 in Athens.

The conducted research is aimed to improve the efficiency of serving language models in real-world applications, towards balancing throughput, latency, and resource usage in AI-oriented Cloud services. Large models like GPT-3 demand immense computational resources, and their internal architecture – generating one word at a time – creates performance bottlenecks. The performed study explores how different batch sizes and serving techniques like dynamic batching affected performance.

An important key finding focusing on Small Language Models (SLM) is

that such models could achieve optimal throughput on a single GPU, eliminating the need for costly multi-GPU setups. And additionally, replicating models on the same hardware further enhanced efficiency. The joint IBM-BSC research team also measured energy consumption, showing that smaller models often used significantly less energy while maintaining performance, critical for large-scale AI deployments.

Such collaboration has continued along 2024, improving results and aimed to continue through 2025 in the frame of CLOUDSTARS. Furthermore, building on the success of this first collaboration on the topic of Cloud-oriented high-performance AI applications, during 2024 other two PhD students at BSC, Ferran Agulló and Joan Oliveras, have continued with the ongoing research, with the expectation to provide new publishable results by the end of 2024 and 2025.

The published work can be found in the “Proceedings of the 4th Workshop on Machine Learning and Systems”, pages 144 – 152, 22nd April 2024, with on-line access link HERE.

cloudstars.eu | twitter.com/Cloudstars_2023 | github.com/cloudstars-eu

CLOUDSTARS project has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101086248