NEWS LETTER

18

September | 2025

IBM Dublin and VU Amsterdam Collaboration on LLM Systems

by Matthijs Jansen and Zebin Ren (Vrije Universiteit Amsterdam)

Large Language Models (LLMs) are rapidly transforming our digital society, impacting education, healthcare, and industry. However, their significant

computational, memory, and storage demands pose serious challenges to existing cloud infrastructure. As part of the CLOUDSTARS project, we aim to enhance the scalability and efficiency of European cloud infrastructure in executing LLM-driven AI workloads. This newsletter highlights the ongoing research collaboration between IBM Dublin and Vrije Universiteit Amsterdam, focused on two key areas: scheduling LLM workloads on heterogeneous GPU clusters and enabling LLM inference using disaggregated CXL memory.

Matthijs Jansen visited IBM Dublin for four months in 2024 to research how cloud providers can efficiently schedule LLM applications on GPU clusters as part of machine learning services. Modern GPU clusters are often heterogeneous due to short release cycles, providing many options for executing LLM applications. When users submit LLM-based applications with varying models and datasets to cloud services, providers must decide how them across available hardware. Yet detailed knowledge of the performance and resource consumption of LLM-based applications is often missing, despite being critical for scheduling decision-making. Using empirical data on application performance and resource usage collected within IBM, Matthijs developed a framework to predict runtime and memory requirements. Ongoing research shows that integrating these predictions into existing schedulers can improve GPU placement efficiency and user latency.

Zebin Ren visited IBM Dublin for three months in 2024 to investigate how to offload GPU memory usage efficiently during LLM inference. Modern LLMs often contain tens to hundreds of billions of parameters and process millions of tokens to achieve complex tasks such as repository-level code analysis. However, this scale of inference requires memory capacities beyond that of a single GPU, especially when maintaining large model weights and key-value (KV) caches. While multi-GPU inference provides higher memory capacities, it comes at high hardware costs. During the exchange, Zebin explored the performance implications of offloading LLM models and KV caches during inference to main memory, CXL memory, and NVMe storage devices. His work provides an in-depth view on how GPU memory offloading affects the performance of LLM inference and poses suggestions on how to offload GPU memory during LLM inference efficiently.

Following their stay in Dublin, both researchers continued developing their work. Zebin published his ongoing research titled “An I/O Characterizing Study of Offloading LLM Models and KV Caches to NVMe SSD” in the Workshop on

Challenges and Opportunities of Efficient and Performant Storage Systems (https://doi.org/10.1145/3719330.3721230). Matthijs presented his work "Efficient Cluster Scheduling for Fine-tuning LLMs Using Historical Configuration Data" at the Cloud Control Workshop and the NWO ICT.OPEN conference. Two additional researchers from the VU are scheduled to visit IBM Dublin in 2025 and beyond to continue this promising line of research.

CLOUD STARS

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