INNOVATION PARTNERSHIPS
UNIVERSITY OF MICHIGAN

DRAM-Based Process-In-Memory System for GPT Inference Acceleration

TECHNOLOGY NUMBER: 2023-569

OVERVIEW

High-Performance DRAM-Based System for GPT Inference Acceleration

- Enables significant speedup and energy savings for GPT model inference
- Real-time text generation, conversational AI, language processing models

BACKGROUND

Traditional Transformer models have advanced the field of natural language processing by executing tasks like text generation and translation. Despite their benefits, these models demand significant computing resources, prompting the development of hardware accelerators. Current accelerators often incur high costs and are limited by their focus on specific subcomponents, overlooking comprehensive workflow optimization. Moreover, previous designs often target encoder-only models, leaving decoder-only models with less efficient solutions. Addressing this gap, there is a necessity for an integrated hardware-software solution capable of handling the enormous data processing demands of these models without compromising speed or energy efficiency.

INNOVATION

Researchers at the University of Michigan have developed PIM-GPT system that integrates computing capabilities within DRAM, minimizing data movement, and reducing performance bottlenecks. This architecture leverages ASICs for intensive computations, such as softmax operations, while mapping software functions to maximize data locality, enhancing parallelism. The PIM-GPT system significantly accelerates inference for extensive models like GPT-2 and GPT-3-XL, achieving substantial improvements in speed and energy use compared to traditional hardware. With potential implementations in artificial intelligence applications, this invention provides a scalable, efficient solution to deploying resource-intensive models across various sectors, including machine translation and text analysis.

ADDITIONAL INFORMATION

REFERENCES

Technology ID

2023-569

Category

Hardware
Engineering & Physical Sciences
Semiconductor, MEMS, and Electronics

Inventor

Wei Lu
Yuting Wu
Ziyu Wang

Further information

Joohee Kim
jooheek@umich.edu

View online page

Wu, Y., Wang, Z. & Lu, W.D. PIM GPT a hybrid process in memory accelerator for autoregressive transformers. npj Unconv. Comput. 1, 4 (2024). https://doi.org/10.1038/s44335-024-00004-2

INTELLECTUAL PROPERTY

US20250028563 "Accelerator Architecture For A Transformer Machine Learning Model"