Department of Computer Science and Engineering
School of Electrical and Computer Sciences
Indian Institute of Technology, Bhubaneswar

HIGH PERFORMANCE COMPUTER ARCHITECTURE

PROJECT PHASE 1

Analytical Modeling of LLM Computation and Communication on Real CPU/GPU Hardware

Team Members

Nali Bhavana - 24AI06013
Devesh Sharma - 24CS06002
Sapna Vishwakarma - 24CS06012
Puvvula Nikhileswari - 24AI06017

Analytical Modeling of LLM Computation and Communication on Real CPU/GPU Hardware

Abstract

Large Language Models (LLMs) such as GPT-3, LLaMA, and Falcon-40B have revolutionized NLP by enabling advanced text generation, summarization, and reasoning tasks. However, their deployment is constrained by high computational cost, memory footprint, and network communication overhead. The autoregressive nature of LLMs, memory-bound computations, and inter-GPU communication bottlenecks make inference and training inefficient on conventional hardware.

This project aims to develop an analytical performance model that predicts LLM execution time, memory consumption, and communication overhead based on hardware configurations, model parameters, and parallelization strategies. By profiling LLM execution on real CPU/GPU hardware and using Roofline model analysis, this study will identify performance bottlenecks and propose optimization techniques for energy-efficient AI computing.

1 Introduction

LLMs have become central to AI-driven applications, yet their performance is highly dependent on hardware efficiency. Training and inference of LLMs require massive computational resources, high-speed memory access, and optimized interconnect communication. The key challenges include:

- **Compute-Intensive Workloads**: LLMs rely on deep transformer layers, making them computationally expensive.
- **Memory Bottlenecks**: Storing billions of parameters leads to out-of-memory (OOM) errors and requires memory-efficient architectures.
- **Communication Overhead**: Multi-GPU setups experience delays due to synchronization inefficiencies in NCCL, NVLink, and InfiniBand.

Despite efforts like model quantization, speculative decoding, and MoE architectures, there is no unified analytical model that predicts how LLMs behave under different hardware setups. A systematic performance model will help optimize parallelization strategies, memory usage, and inference latency.

2 Problem Definition

Current hardware and software optimizations for LLMs lack a generalized performance prediction model. Training and deploying LLMs efficiently requires understanding the balance between:

- **Computation**: How different transformer operations impact execution time.
- **Memory Access**: The role of caching, prefetching, and attention optimization.
- **Communication Delays**: The impact of interconnect bandwidth on scaling performance.

This project seeks to answer:

1. How does LLM execution scale across single-GPU, multi-GPU, and multi-node systems?
2. What are the dominant bottlenecks—compute-bound, memory-bound, or communication-bound?
3. Can we develop a mathematical performance model to predict LLM inference/training time?

4. What are the best optimization strategies to reduce energy consumption and improve efficiency?

By addressing these questions, this project will bridge the gap between AI model efficiency and hardware optimization, contributing to scalable and sustainable AI computing.

3 Significance of Research

This study has significant implications in AI research and industry:

- **Performance Optimization**: AI companies can predict execution time and cost before training large models, reducing experimental overhead.
- **Scalability Analysis**: Helps determine the optimal number of GPUs and nodes required for LLM workloads.
- **Energy Efficiency & Sustainability**: AI workloads consume vast power—GPT-3 training required 1.3 GWh, equivalent to 120 US households' annual electricity usage. Optimizing execution supports Green AI initiatives.
- **Edge AI Deployment**: Efficient LLM execution benefits autonomous vehicles, drones, and smart edge devices, where power and compute resources are limited.

4 Expected Outcomes

This project will deliver:

- A predictive performance model for LLM execution on CPU/GPU hardware.
- Profiling results analyzing computation, memory, and communication overhead.
- Optimization strategies for reducing execution time and energy usage.

The findings will help optimize parallel computing frameworks (DeepSpeed, FSDP, ZeRO) and guide future AI accelerator designs.

5 Conclusion

The rapid growth of LLMs demands efficient execution strategies to improve scalability and sustainability. This research will bridge the gap between AI model performance and hardware-aware optimizations by developing an analytical framework for LLM computation and communication modeling. The insights from this study will guide efficient training, deployment, and sustainable AI computing, making AI more accessible, energy-efficient, and scalable.

References

[1] Zhihang Yuan, Yuzhang Shang, Yang Zhou, Zhen Dong, Zhe Zhou, Chenhao Xue, Bingzhe Wu, Zhikai Li, Qingyi Gu, Yong Jae Lee, Yan Yan, Beidi Chen, Guangyu Sun, and Kurt Keutzer. *LLM Inference Unveiled: Survey and Roofline Model Insights.* Infinigence-AI, Illinois Institute of Technology, Carnegie Mellon University, Peking University, Tencent AI Lab, Institute of Automation, CAS, University of Wisconsin, Madison, University of California, Berkeley, 2025.