CV of Tian Tang

tiantang673@gmail.com tang-t21

Education

Tsinghua University **Sep 2021 – present**

Undegraduate in Yao Class (Honored Computer Science class), IIIS in Tsinghua *GPA: 3.87/4*

Research Interest

Machine learning systems; Distributed systems; Large language models.

Professional Skills

Languages: English: Toefl 107 (R29, L27, S23, W28); Mandarin
Computer skills: Python, C++/C, CUDA, Verilog, SQL; Pytorch, Cutlass, vllm, llama.cpp, GEM5, MATLAB

Research Experience

IDEAL Lab, Tsinghua University **July 2023 – Feb 2024**

Project: ***Scalable and Flexible Accelerator for Modern Cryptographic Primitives*** *Supersivor: Prof.Mingyu Gao*

- Identified the fixed pattern of running FHE algorithms on hardware
- Constructed operator graph and applied pipeline and co-locate techniques to find optimal schedule
- Implemented ResNet and Logistic Regression in encrypted version and evaluated it on our method

Efes Lab, University of Washington **Feb 2024 – Jun 2024**

Project: ***Towards Optimal Large Language Model Serving Throughput*** *Supervisor: Prof.Baris Kasikci*

- Constructed kernel wrapper and linked them into pipeline mode
- Evaluation data collection and visualization
- Paper in submission to ODSI 2025

Efes Lab, University of Washington **Feb 2024 – Aug 2024**

Project: ****Heterogeneous Architecture for Inference of Mixture-of-Experts Models*** *Supervisor: Prof.Baris Kasikci*

- Designed an inference system that finds the optimal execution strategy using both the GPU and CPU
- Added beam search feature to mixtral model and evaluate it on the system
- Optimized the computation of expert on CPU using AVX512 instruction set
- Paper in submission to ICLR 2025

Efes Lab, University of Washington **Aug 2024 – present**

Project: ****Dynamic Thresholding for Sparse Attention in Long-Context Models*** *Supervisor: Prof.Baris Kasikci*

- Profile sparsity in attention score for motivation
- Utilize inherent distribution of key vectors by building data structure
- Since it's an on-going project, if you want more details, drop me an email!

***Co-lead the project.**

Publications

Orchestrating Heterogeneous Architecture for Fast Inference of Mixture-of-Experts Models

Keisuke Kamahori,* ***Tian Tang****, Yile Gu, Kan Zhu, Baris Kasikci. (*equal contribution), in submission to ICLR 2025*

- We designed an inference system for MoE models for heterogeneous architecture, that finds the optimal execution strategy using both the GPU and CPU.

NanoFlow: Towards Optimal Large Language Model Serving Throughput

K.Zhu, Y. Zhao, L.Zhao, G.Zuo, Y.Gu, D.Xie, Y.Gao, Q.Xu, ***T.Tang*** *,..., A.Krishnamurthy, B.Kasikci*

- A detailed analysis and validation of the workload characteristics and the theoretically optimal throughput of LLM serving systems.
- Intra-device parallelism, a novel parallelism paradigm that exploit nano-batching to maximize hardware utilization.

Honors and Awards

Friends of Tsinghua - Lingjun Pilot Scholarship **2021-2022**

Tsinghua Academy Talent Training Program **2021-2025**

Second Class Scholarship for Freshmen **2021-2025**

Gold medal in the National Physics Olympiad for high school students **2020**

Projects

Database Management System | *C++, SQL Github Repo* **March-June 2023**

- The project is led by a course taught by Prof. Huanchen Zhang, which is similar as CMU-15445. The DBMS (Database management system) incorporates B+ tree Indexing, volcano model of execution, optimizer and concurrency control.

BlockChain System | *Go, Github Repo* **Nov 2023-Jan 2024**

- This project aims at building a basic version of BlockChain system, for details check the report. This is a course project of Distributed System, taught by Prof. Wei Xu.