Tian Tang ✉ tiantang673@gmail.com tang-t21

Education

Tsinghua University **Beijing, China**
Bachelor of Engineering (B.E.) in Computer Science and Technology *September 2021 – present*

- GPA: **3.87** / 4.0
- **Yao Class** (Honors Computer Science Program), Interdisciplinary Institute for Information Science (IIIS).
- Selected Coursework: Linear Algebra (A), Calculus (A), Introduction to Computer Systems (A), Operating Systems and Distributed Systems (A), Database Systems (A), Introduction to Programming in C/C++ (A+).

Research Interest

Systems for Machine Learning, Distributed Systems, Networks, Large language models.

Research Experience

IDEAL Lab, Tsinghua University **July 2023 – February 2024**
Project: ***Scalable and Flexible Accelerator for Modern Cryptographic Primitives*** *Supersivor: Prof.Mingyu Gao*

- Identified fixed patterns in execution of FHE algorithms on hardware
- Constructed operator graph and applied pipeline and co-locate techniques to determine optimal schedule.
- Implemented encrypted versions of ResNet and Logistic Regression, evaluating their performance with our method.

Efes Lab, University of Washington **February 2024 – June 2024**
Project: ***Towards Optimal Large Language Model Serving Throughput*** *Supervisor: Prof.Baris Kasikci*

- Constructed kernel wrapper, integrating it into pipeline mode.
- Collected, visualized data. Submitting to OSDI 2025

Efes Lab, University of Washington **February 2024 – August 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.
- Integrated beam search to Mixtral model and evaluated performance.
- Optimized the computation of expert on CPU using AVX512 instruction set. Paper in submission to ICLR 2025

Efes Lab, University of Washington **August 2024 – present**
Project: ****Dynamic Thresholding for Sparse Attention in Long-Context Models*** *Supervisor: Prof.Baris Kasikci*

- Dynamic budget to reach given attention weight ratio.
- Leverage inherent distribution of key vectors by building data structure.

***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, Tsinghua University	**2021-2022**
Tsinghua Academy Talent Training Program, Tsinghua University	**2021-2025**
Second Class Scholarship for Freshmen, Tsinghua University	**2021-2025**
Gold medal, National High School Physics Olympiad	**2020**

Projects

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

- This is a course project led by Prof. Huanchen Zhang, which mirrors CMU-15445. The project integrates B+ tree Indexing, volcano model of execution, optimizer, transactions and concurrency control.

BlockChain System | *Go, Github Repo* **November 2023-January 2024**

- Developed a basic version of blockchain system as part of Prof. Wei Xu's Operating Systems and Distributed Systems course. Project includes core blockchain functionalities such as transaction validation, block creation, chain linking, attack and defense strategy. Check report for details.

Additional Skills

Languages: English: TOEFL 107 (Reaing: 29, Listening: 27, Speaking: 23, Writing: 28); Mandarin
Computer skills: Python, CUDA, C++/C, Go, Verilog, SQL; Pytorch, Cutlass, vllm, llama.cpp, GEM5, MATLAB