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; Large language models; Distributed systems.

Professional Skills

Languages: English: Toefl 107; Mandarin

Computer skills: Python, C++/C, CUDA, Verilog, SQL; Pytorch, Cutlass, vllm, GEM5, MATLAB

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**

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
- Paper in submission to ISCA 2025

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

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: **_*Exploit query-aware sparsity in long-context inference of LLM_** _Supervisor: Prof.Baris Kasikci_

- Profile sparsity in attention score for motivation
- Utilize inherent distribution of key vectors by building data structure

***Co-lead the project.**

Publication

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.