Ziren Wang

wang-zr22@mails.tsinghua.edu.cn Wazrrr
Google Scholar Personal Web

Biography

I am Ziren Wang, a senior undergraduate in **Yao Class** at Tsinghua University with GPA 3.86/4.00, majoring in Computer Science.

I am currently a research intern at **Efeslab** / **SyFI** at University of Washington, advised by Prof. **Baris Kasikci**, and also advised by Prof. **Mingyu Gao** at Tsinghua University.

Currently, I am working on the development of a flexible and high-performance Python framework for LLM inference, featuring fine-grained intra-GPU resource management. My research interests center around distributed systems and machine learning systems, with a focus on building efficient and scalable infrastructures for modern AI workloads.

Research Publications

1. K. Zhu et al., "Nanoflow: Towards optimal large language model serving throughput," in *19th USENIX Symposium on Operating Systems Design and Implementation (OSDI 25)*, 2025, pp. 749–765.

Projects

LLM Inference

Leading development of a flexible, high-performance Python framework for LLM inference with fine-grained intra-GPU resource management.[code]
We model SMs, memory bandwidth, and PCIe transfers as separable resources on independent streams, enabling kernel co-scheduling and overlap to maximize GPU utilization while minimizing cross-kernel interference.

Network on Chip

Reproduced GOAL Algorithm on 3D Torus Network [code] [report][slides]
I reproduced the current SOTA algorithm for torus networks GOAL, implemented different VCs control policies and evaluated the experiment results, which shows global load balance by randomly choosing the direction to route in each dimension and therefore achieves local load balance by routing adaptively.

Backend Development

Enrollment Website[report & demo]
Our team developed an enrollment system, where we can publish announcements, and it also allows users to take exams. Additionally, there are some design tricks, such as masking, security design, and so on.

Numerical Analysis

A New Randomized Cholesky Parallel Decomposition Algorithm[report]
We presented a new parallel decomposition algorithm that utilizes the sampling algorithm of RChol in conjunction with Multifrontal, dynamically managing the dependencies between threads and nodes. Experiments show that this algorithm can effectively improve the matrix decomposition rate when the matrix has high parallelism; however, it does not accelerate matrices that are inherently difficult to compute in parallel.

Awards and Achievements

2022	**Golden Medal in Asian Physics Olympiad (APHO)**, 2nd place in Global Ranking
2024	**Honorable Mention in Interdisciplinary Contest in Modeling**

Experiences

2025.2 – Present	**Research Internship**, University of Washington, following Baris Kasicki.

Education

2022 – Present	**B.S., Tsinghua University**, Yao Class.

Skills

English	**TOEFL Score:** 96
Programming Language	**Python, C++, CUDA**