Wang Ziren | 王梓人

wang-zr22@mails.tsinghua.edu.cn • https://wazrrr.github.io/ • https://github.com/Wazrrr

Last updated on 2025.11.10

EDUCATION

Tsinghua University, Beijing, China

Senior undergraduate in Yao Class, Institute for Interdisciplinary Information Sciences **2022 - Present**

- **GPA:** 3.86/4.0 (Top 30% in Yao Class)
- **Relevant Courses:** Computer Architecture (A-), Operating System and Distributed System (A), Machine learning (A-)

HONOURS AND AWARDS

Golden Medal in Asian Physics Olympiad (APHO), 2nd place in Global Ranking **2022**
Comprehensive Merit Scholarship of Tsinghua **2023**
Honorable Mention in Interdisciplinary Contest in Modeling **2024**

SKILLS AND INTERESTS

Programming Languages: C & C++, Python, CUDA.
Current Research Interests: Distributed Systems and Machine Learning Systems.
Other Interested Areas: Robotics, Machine Learning Algorithms.
Research Goal: Building efficient and scalable infrastructures for modern AI workloads.

RESEARCH EXPERIENCES (HIGHLIGHT)

Nanoflow: Towards Optimal Large Language Model Serving Throughput
https://www.usenix.org/system/files/osdi25-zhu-kan.pdf
Supervised by Baris Kasikci, University of Washington

- We propose NanoFlow, a novel serving framework that exploits intra-device parallelism, which overlaps the usage of heterogeneous resources within a single device.
- Split inputs into smaller nano-batches and duplicate operations to operate on each portion independently, enabling overlapping.
- Automatically identifies the number, size, ordering, and GPU resource allocation of nano-batches to minimize the execution time, while considering the interference of concurrent operations.
- Published in **OSDI 2025**.

SchedFlow: Transparent and Flexible Intra-Device Parallelism via Programmable Operator Scheduling
Supervised by Baris Kasikci, University of Washington

- Introduce a framework that enables the transparent and flexible integration of intra-device parallelism by decoupling the logical model definition from the physical execution schedule.
- Introduce a flexible frontend with annotations for graph partitioning and a programmable interface for defining custom intra-device parallelism strategies.

- Its efficient backend manages complex control/data-flow asynchronously, uses custom memory management to eliminate copy overheads, and preserves compatibility with optimizations like CUDA Graphs and TorchInductor.
- Submitted to **MLsys 2026**.

Improving the efficiency of LLM inference with PIM architectures

Supervised by Mingyu Gao, Tsinghua University

- Address this limitation by modeling the KV cache imbalance that arises from requests of varying lengths across GPUs.
- Implement a simulator to simulate the time consumption of various operations, including both ML inference and PIM processing steps.
- My contribution: Conducting literature reviews on recent PIM-based acceleration research. Participating in weekly discussions with senior researchers to refine our methodology and analysis.

Exploring sim-to-real transfer for robotic manipulation

Supervised by Huazhe Xu, Tsinghua University

- Building simulation pipelines, surveying, and comparing state-of-the-art vision-based methods.
- My contribution: Help with running ablation and making plots.

TODO: GOOD OPEN SOURCE PROJECTS

COURSE PROJECTS

Reproduced GOAL Algorithm on 3D Torus Network

https://github.com/Wazrrr/NoC_Project/tree/working

"AI+X Computing Acceleration: From Algorithms Development, Analysis, to Deployment" Course Project

- Reproduce the current SOTA algorithm for torus networks GOAL.
- Implement different VCs control policies and evaluated the experiment results, which show global load balance by randomly choosing the direction to route in each dimension, and therefore achieve local load balance by routing adaptively.

A Variant of Randomized Cholesky Parallel Decomposition Algorithm

link to report

"Numerical Analysis" Course Project

- Present 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.

Enrollment Website

link to Google Drive

"Type-safe Modern System Practice" Course Project

- Develop 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.

EXTRACURRICULAR EXPERIENCE

CPHOS (A Non-Profit Organization for Physics Olympiads). **2022.1 - 2024.7**

Tech Leader

- Lead the Theory Group.
- Guide the design, direction, and academic rigor of the theoretical examination papers.

Tsinghua Drama Troupe **2022.9 - Present**

- Majors in Acting.
- Training focuses on Stanislavski's system.