# Yichuan Wang

yichuanjiaoda@sjtu.edu.cn or yichuanmistygrass@gmail.com — https://yichuan520030910320.github.io/

## Education

**Shanghai Jiao Tong University** Shanghai, China

*Bachelor of Computer Science* *Sept. 2020 - Present*

- Member of ACM Honors Class, which is an elite CS program for top 5% talented students
- GPA (All): 88.13/100, GPA (year 2&3): 90.03/100, rank: 11/36
- Scores of some courses:
  - Natural Language Processing: 98/100, Ranking: 1/34
  - Machine Learning: 97/100, Ranking: 2/34
  - Mathematical Logic: 97/100, Ranking: 1/34
  - Computational Complexity: 98/100, Ranking: 1/34

## Experience

**Shanghai Jiao Tong University** Shanghai, China

*Undergraduate Researcher in Machine Learning System, advised by Prof. Quan Chen* *June. 2022 - present*

1. Optimize the Inference of Dynamic Neural Networks. I was primarily responsible for the single-machine optimization in the paper. I proposed some ideas inspired by traditional compilers and implemented the corresponding code

2. Exploring another project on how to deploy dynamic neural networks in distributed scenarios and topics related to serving Large Language Models (LLMs), from the perspective of Mixture of Experts (MoE).

**New York University** New York, NY, USA.

*Research Assistant in Machine Learning System, advised by Prof. Jinyang Li* *Jan. 2023 - Present*

Currently working on scaling up Graph Neural Networks (GNNs) on disk with algorithm and system co-design, optimizing the data transfer and making optimizations based on the training pattern of GNN mini-batches. Specifically addresses the read amplification in SSD reading, and uses some algorithms and system designs to overlap reading data from SSDs. I finalized the proof of concept with substantial end-to-end improvements, achieving a 13x increase in data transfer speed and a 4.1x enhancement in overall performance, all without any loss in accuracy.

**AWS Shanghai AI lab** Shanghai, China

*Research Intern with Dr. Minjie Wang* *Oct. 2023 - Present*

Currently serving as an unofficial collaborator for SSD GNN training.

## Publications

**Optimizing Dynamic Neural Networks with Brainstorm**

*W.Cui, Z.Han, L.Ouyang,* ***Y.Wang****, N.Zheng, L.Ma, Y.Yang, F.Yang, J.Xue, L.Qiu, L.Zhou, Q.Chen, H.Tan, M.Guo.*

- Accepted by **OSDI 2023** [pdf]
- We mainly propose a better abstraction for dynamic neural networks to achieve more optimization space. We optimize the execution time and memory utilization for dynamic neural networks like MoE.
- As the main developer, **I primarily accomplished the design and implementation of the ideas** in Dynamic Horizontal Fusion, Speculative Routing, and Speculative Weight Preloading in DFG and CUDA level.

## Teaching Experience

**Principle and Practice of Computer Algorithms**

*Teaching Assistant guided by Prof.Yong Yu* *June. 2022 - Aug. 2022*

Advised students to implement a RISC-V simulator.Helped students with ray tracing projects, designed some additional tracks, and handed on every student with the outline of the project and details.

**Compiler design**

*Teaching Assistant guided by Prof.Yong Yu* *Sept. 2022 - Feb. 2023*

Help students design their compiler and teach them my own experience and knowledge in compiling.

## Projects

**Personal Project: RayTracer**

*About 8K LoC in RUST [repo]* *Jun. 2021 - Jul. 2021*

A system that can build its scene and track the refraction and reflection of light in it. Several compilation-related optimizations were undertaken to enhance the overall performance.

**Course Project: Compiler for Mx* Language**

*About 8K LoC in JAVA [repo]* *Sept. 2021 - Feb. 2022*

Using the technology of ANTLR, semantic checking, code generation, and optimization to develop a compiler that compiles C-and-Java-like language (Mx∗) to RV32I Assembly.

**Course Project: RISC-V CPU Implemented in Verilog RTL**

*About 3K LoC in Verilog [repo]* *Sept. 2021 - Dec. 2021*

Designed a RISC-V CPU with Write Buffer, ICache, DCache, and Branch Prediction. Supports RV32I instruction set (2.1-2.6 in RISC-V User Manual). Used Vivado to generate bitstream and program the Basys3 FPGA board.

**Course Project: Ticket System**

*About 4K LoC in C++[repo]* *Apr. 2021 - Jun. 2021*

Designed a train ticket system with multi-user support and privilege management. I implemented the backend and built a Bplustree storage. All used C++ STL data structures are from scratch (including map, and queue).

## Honors & Awards

**Scholarship**

- 2020, 2021, 2022 Zhiyuan Honorary Scholarship (Top **2%** in Shanghai Jiao Tong University, in all 2K USD)
- 2021, 2022 Excellence Scholarship for Undergraduates
- 2022 Commercial Sponsorship Scholarship. (Top 8 in ACM class. 1.5K USD)
- 2023 Ruiyuan - Hongshan Talent Development Fund. (Top 70 persons in SJTU one year one year. 3K USD)

## Technical Skills

**Programming Languages**: C/C++, Python, Java, Rust, CUDA, RISC-V Assembly, Verilog, MatLab
**Framework**: Torch, DGL, Pyg, MongoDB, SQL, Vivado, Latex, Markdown

## Hobby

Playing basketball, traveling, watching sports events (including but not limited to NBA, English Premier League, UEFA).