Junlin Chen

(+86)182-0080-2136 ◊ junlin.chen@buaa.edu.cn ◊ hiGiraffe.github.io

Education

Beihang University Beijing, China

Bachelor of Computer Science *Sept. 2021 - Jun. 2025(Expected)*

- ◇ GPA: 3.70/4
- ◇ Second Class Innovation and Entrepreneurship Scholarship, 2022-2023
- ◇ Second Class Outstanding Social Work Scholarship, 2022-2023
- ◇ Outstanding Student Cadre, 2022-2023

University of Macau Macau, China

Exchange Program *Aug. 2024 - May. 2025(Expected)*

Research Interests

Performance Optimization, Machine Learning System and **Distributed System**.

Research Publications

[1] Zizhao Mo, **Junlin Chen**, Huanle Xu, et al. "Serving Hybrid Loads for LLMs with SLO-awareness using CPU and GPU", to be submitted before December.

[2] **Junlin Chen***, Chaojing Liu*, Zhongzhi Luan, Ming Gong, Qingfeng Li, Depei Qian, "Large-Scale Parallelization and Optimization of Lattice QCD on Tianhe New Generation Supercomputer", The 25th IEEE High Performance Computing and Communications (HPCC 2023), Dec. 13-15, 2023, Melbourne, Australia.

Research Experience

Research Intern @ University of Macau Mar. 2024 – Present

Cloud Computing and Distributed Systems Laboratory *Macau, China*

- ◇ Conducted research on LLM inference system optimization and efficient scheduling under the guidance of Prof. Huanle Xu.
- ◇ Supplemented relevant cutting-edge knowledge, including LLM serving system and efficient scheduling for mlsys.
- ◇ Engaged in a research project of LLM serving system(in progress).
- ◇ Familiar with the vLLM related code and made further modifications to it.

Research Intern @ Beihang University Sept. 2022 – Feb. 2024

Sino-German Joint Software Institute *Beijing, China*

- ◇ Conducted research on performance optimization under the guidance of Prof. Zhongzhi Luan.
- ◇ Collaboratively completed a research project utilizing Tianhe new generation supercomputer to accelerate scientific computations, including parallel mode design and hardware optimization.
- ◇ Participated in the design of parallel computing patterns, experimentation, paper writing, scientific visualization and oral presentation.

Projects

A System for Serving Hybrid Workloads for LLMs with SLO Awareness | *Pytorch* Jun. 2024 – Present

- ◇ Achieve fine-grained performance isolation for mixed LLM requests and a lower service-level objective(SLO) violation rate.
- ◇ Effectively leveraged CPUs and GPUs to enhance system throughput in the context of significant performance disparities between CPU and GPU.
- ◇ This work involved incremental modifications to the vLLM code framework.

BattleByte: Online Programming Battle Platform | *Spring Boot* Feb. 2024 – Jun. 2024

- ◇ Analyzed product requirements, designed in-game mechanics, authored product documentation and coordinated team efforts as a Product Manager.
- ◇ Designed and developed the backend WebSocket real-time communication component.

Online Flea Market Platform | *Python, Flask* Sept. 2023 – Dec. 2023

- ◇ Utilized the Flask framework to complete the backend code for user center and flea market functionalities.
- ◇ Integrated the backend with databases using GaussDB for MYSQL and MYSQL.

SysY-to-LLVM Compiler Project | *C++* Sept. 2023 – Dec. 2023

- ◇ Developed a compiler that translates SysY language into LLVM language, encompassing lexical analysis, syntax analysis, semantic analysis, LLVM intermediate code generation, and error handling.

Accelerating Lattice QCD on Supercomputer | *C, OpenMP, MPI* Dec. 2022 – Dec. 2023

- ◇ Accelerated communication between two computational processes through Global Shared Memory and Array Memory.
- ◇ Accelerated vectorized calculations through the MT-3000 processor's Acceleration Array.
- ◇ Conducted performance analysis on global reductions, identifying bottlenecks and proposing adaptive strategies to optimize reduction frequency.

Multi-threaded Scheduling System | *Java* Feb. 2023 – Jun. 2023

- ◇ Developed a multi-threaded elevator scheduling system supporting elevator maintenance and elevator accessibility.
- ◇ Developed a local greedy approach to handle the addition of elevators and maintenance requests.
- ◇ Completed the development using the principles of object-oriented programming.

MIPS Pipeline Processor with Exception Handling Support | *Verilog* Sept. 2022 – Dec. 2022

- ◇ Implemented a MIPS five-stage pipeline CPU that supports branch prediction and hazard handling.
- ◇ Implemented external instruction memory and data memory.
- ◇ Introduced CP0, Bridge, and Timer to support interrupt and exception handling.

Other Information

Language Proficiency: English, Mandarin, Cantonese

Programming Languages: C++, Python, Java, Spring Boot, Verilog, LLVM IR, MIPS Assembly Language, LaTex

Frameworks and Tools: OpenMP, MPI, CUDA, PyTorch, Ray

Familiar Project Code: vLLM

Further personal skills are showcased on my personal notes website.