Ruizhe Fu

Email: furuizheno1@gmail.com
Linkedin: https://www.linkedin.com/in/ruizhe-fu/
Website: https://frzno1.github.io/
Github: https://github.com/FrzNo1
Research Interest: Distributed System, Build System, Parallel Computing (GPU),
High-Performance Systems for Machine Learning

Education

Columbia University, New York City, NY

Bachelor of Science, **Computer Engineering** ***Sep 2023 - Current***

- GPA: 4.137/4.33
- Dean Lists

Grinnell College, Grinnell, IA

Bachelor of Art, **Computer Science** ***Sep 2020 – May 2023***

- GPA: 3.99/4.00
- Dean Lists
- Best Student Research Poster Award

Publication

Fast Distributed Selection with Graphics Processing Units

- Jeffrey D. Blanchard, **Ruizhe Fu**, Tristan Knoth
- Under review, IEEE Open Journal of the Computer Society

Research

Enhancing the Underlying System of Large Language Models

Columbia University, New York City, NY ***Sep 2024 – Current***

- Developed benchmarks to measure and analyze latency and memory usage during prefill and each decoding step for the vLLM model;
- Currently implementing intra-OP parallelism for the vLLM model to dynamically adjust GPU cores based on real-time memory and compute requirements;
- Aim to develop a new scheduling policy for the DistServe model that optimizes the balance between the prefill and decode states, moving beyond the currently used FCFS approach;
- Plan to implement a memory constraint on the total number of accepted requests to address the shortcoming of the current design for the Sarathi-Serve model.

Riker: Always-Correct and Fast Incremental Builds

Grinnell College, Grinnell, IA ***May 2023 – Aug 2023***

- Contributed to Riker, a forward build system that always guarantees fast and correct builds without manually listing any dependencies, using C/C++;
- Modeled the POSIX filesystem, directories and pipes to discover fast increment rebuild opportunities and guarantee every dependency is checked on each build;
- Added fresh flag and implemented Socket artifact to distribute Riker for tracing files across machines;
- Tested and built 14 open source packages including LLVM and Memcached, achieving average 94% of Make's speed on incremental builds with no risk of errors.
- Github repo: https://github.com/curtsinger-lab/riker

Fast Distributed Selection with Graphics Processing Units

Grinnell College, Grinnell, IA ***May 2022 – Aug 2024***

- Designed parallel algorithm with GPU for selecting thousands of order statistics from huge data sets, using C/C++ and CUDA with Thrust and Cub library;
- Distributed the parallel algorithms with Open MPI to select order statistics across machines without data set transformations, supporting Cloud Computing and improving speed and security measures;
- Achieved exponential increase in speed with larger vector size, ultimately reaching a 10k times speed-up for float vectors of length 228 compared to copy and select method;
- Released free software "DistributedSMOS" consisting of thousands of tests on over 20 distributions;
- Paper currently under review: IEEE Open Journal of the Computer Society.
- Github repo: https://github.com/FrzNo1/GGMS-Distributed

Experience

Software Engineer at Siemens AG(DISW division)

Siemens AG, Costa Mesa, CA ***May 2024 – Current***

- Develop and enhance features for NX, a leading CAD software, utilizing Object-Oriented programming approach with C/C++ and JSON;
- Resolve customer-reported issues and submitted change package to the 2412 release baseline through development testing processes and code review;
- Lead a project and collaborated with team to implement coating layer thickness visualization, adhering to the company's software development lifecycle with modification to 30+ files;
- Boost code testing coverage rate to 95% by designing, creating and executing unit tests, UI tests, and automated tests using Python, Java, and XML.

Software Engineer at State Grid Corporation of China

State Grid Corporation, Jiangsu, China ***May 2021 – Aug 2021***

- Collected and evaluated data from the tests of dry-charge of voltage transformer.
- Filtered raw data and ensured its consistent patterns to facilitate further data manipulations in MATLAB environment, using R and NoSQL.

Service

Teaching Assistant for COMS W4118 Operating System

Columbia University, New York City, NY ***Sep 2024 – Current***

- Hold office hours, grade assignments, update homework & solutions, revise midterm & final exams;
- Class website: https://www.cs.columbia.edu/~nieh/teaching/w4118/

Teaching Assistant for CSC 207 Data Structure and Algorithms

Grinnell College, Grinnell, IA ***Jan 2022 – May 2023***

- Hold office hours, grade assignments, update homework & solutions, revise midterm & final exams;
- Class website: https://jimenezp.cs.grinnell.edu/Courses/CSC207/2022Fa/syllabus/

Project

Linux Kernel Development

- Developed and integrated a Linux Round-Robin scheduler with SMP support and made it the default scheduler for all normal processes and threads in the kernel, using C with VMware;
- Modified Read-Write Lock in Kernel to support blocking, improving concurrency and system performance;
- Designed and implemented a file system with support for standard file operations, including mounting, directory listing, file/directory reading, modification, creation, and deletion.