Cyrus (Zikai) Zhou

+1 708-631-1373 | Email: zhouzk@uchicago.edu

Education

University of Chicago Chicago, US
BS in Computer Science September 2020 - Present

- **GPA:** 3.9/4.0, **Major:** 3.93/4.00.
- **Selected Awards:** UChicago Summer Quad Research Scholar (S2023), UChicago Quad Research Scholar (2022-2023), UChicago Dean's List (2021-2022, 2022-2023), UChicago Jeff Metcalf Award (May 2022).
- **Selected Courses:** Grad Deep Learning Systems, Grad Advanced Operating Systems, Grad Machine Learning for Computer Systems, Grad 3D Computer Vision, Computer Architecture, Systems Programming, Algorithms.

Publications & Preprints

1. **Cyrus Zhou**, Zack Hassman, Ruize Xu, Dhirpal Shah, Vaughn Richard, and Yanjing Li, "SIMD Dataflow Co-optimization for Efficient Neural Networks Inferences on CPUs", March 2024 (accepted pending revision), *IEEE/ACM International Symposium on Code Generation and Optimization (CGO). [arxiv]*
2. Lefan Zhang, **Cyrus Zhou**, Michael L. Littman, Blase Ur, and Shan Lu, "Helping Users Debug Trigger-Action Programs", Jan 2023, *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT/UbiComp). [pdf]*

Pre-Submission Papers

3. *Hardware-software Co-design for Serving Mixed Precision Neural Networks on CPUs with SIMD.*
4. *If At First You Don't Succeed, Try, Try, Again...? Automatically Detecting Retry Bugs in Distributed Applications.*
5. *Understanding and Detecting Idempotency Violations in Serverless Applications.*

Research Experience

University of Chicago Chicago, US
Research Assistant to Professor Yanjing Li's Group

Hardware-software Co-design for Fast, Efficient, and Accurate Neural Networks June 2023 - Present

- ***Description:*** *Current CPUs and frameworks lack support for both ultra-low precision mixed-precision neural networks. We provide the first hardware-software co-design for these networks on CPUs' SIMD architecture.*
- Led a team of 4 students, proposed the research idea.
- Designed and compared multiple encoding methods for efficiently storing precision information.
- Designed modified SIMD architecture to support mixed-precision patterns.
- Designed new instructions and memory transformation algorithms for correctly running neural networks.
- Implemented a code generator and a simulator for executing neural networks based on all modifications.
- Composed the paper.

SIMD Dataflow Co-optimization for Efficient Neural Networks Inferences October 2022 - September 2023

- ***Description:*** *Compiler support for SIMD programs remains lacking. Existing compilers usually fail to optimize SIMD code or parallelize code with SIMD instructions. We develop a framework for generating efficient SIMD code for serving neural networks, substantially lowering the costs and barriers of deploying neural networks on edge devices and commercial servers, democratizing the use of AI.*
- Led a team of 5 students.
- Proposed the research idea, extended the notion of dataflows to CPUs, analyzed further reuse opportunities.
- Developed a SIMD code generation framework for generating efficient programs for neural networks.
- Designed and implemented end-to-end optimization algorithms.
- Ran experiments, collected data, performed data analysis, compared different dataflows, composed the paper.
- Substantially outperformed state-of-the-art frameworks.

Research Assistant to Professor Shan Lu's Group

Analyzing and Exposing Retry Bugs in Large-Scale Distributed Systems June 2023 - Present

- ***Description:*** *Many developers are not aware of bugs related to retries in distributed systems. We find bugs in these systems using a combination of static checking, large language models, and exception injection.*
- Refined and conducted experiments using Wasabi, a tool that leverages unit tests to validate and track the retry behavior within large-scale distributed systems through exception injection, analyzed bug reports.
- Developed Coconut (Code Coverage for iNdividual Unit Tests), a tool that leverages JaCoCo to find the line coverage of each individual unit test, which is intended to find out useful unit tests to simulate and eventually expose retry bugs in large-scale distributed systems (e.g., Hadoop).
- Helped compose the paper.

Detecting and Exposing Idempotency Violations in Serverless Computing October 2022 - June 2023

- ***Description:*** *Serverless computing has become popular for event-driven applications. However, idempotency bugs are often overlooked by developers, leading to undesired effects. We strive to expose and fix these bugs.*
- Collected Problematic Codebases from Open Source Github Repositories, categorized them by application type, API type, root cause, and symptoms, and found general anti-patterns leading to idempotency violations.
- Deployed 20+ repositories of serverless software, tested static checkers for anti-patterns, and refined and ran experiments with the testing framework for symptom diagnosis.

Helping Users Debug Trigger-Action Programs (TAP) March - August 2022

- ***Description:*** *This project is the first empirical study of users' end-to-end TAP debugging process, focusing on obstacles users face in debugging TAPs and providing them with three debugging tools.*
- Designed user study methodology, proposed, and refined the user obstacles mental model.
- Tested and refined the software system.
- Conducted user study, analyzed the result, and composed the paper.

Research Assistant to Professor Blase Ur's Group

Personalized Classifier for Finding and Deleting Sensitive Files on Google Drive March 2023 - Present

- ***Description:*** *Users often mistakenly upload sensitive files to Google Drive. We develop a recommendation system for deleting these files to greatly enhance user security and privacy.*
- Led a team of 10 students from UChicago and UIC.
- Enhanced the existing sensitive file classifier by integrating multiclass classification and reinforcement learning for improved deletion recommendation personalization.
- Deployed Llama-2 on local GPU, studied prompt engineering methods to best extract features using it.
- Designed a new rule-based classifier using ILP and formal reasoning techniques.
- Devised and conducted user studies for training classifiers more effectively.
- Deployed codebases from preceding works, performed feasibility analysis of backend extension.

Selected Course Projects

Detecting Webcams' Spying Behavior via Machine Learning Autumn 2022

- Applied machine learning techniques to detect if any spying application is running on the host computer.
- Empirically collected experiment data, trained and cross-validated machine learning models, performed both multiclass classification and binary classification, presented the report using Sphinx using nbSphinx.

ARM Instruction Set Simulator Autumn 2022

- Implemented a CPU-based ARM ISA Simulator in the following cumulative stages: (1) Single-stage, (2) Pipelined, (3) Branch Predictor, and (4) Memory Hierarchy. Topped the class in accuracy.

VR Diver Interface Autumn 2022

- Led a team of 4 students, implemented an augmented reality (AR) diver interface on Unity using C#.

Orienteering Problem with Time Windows Spring 2022

- Designed and implemented three algorithmic solutions (greedy, dynamic programming, and backtracking) for an NP-hard problem, wrote a program to find the best solutions, further optimized results with local search.

Chiventure Game Engine Spring 2022

- Led the graphics team (5 members), designed, implemented, and integrated graphical interfaces.

Simple Shell & Cache Winter 2022

- Implemented a Shell in C that performs argument parsing, process forking, and output redirection (all correct)
- Implemented a simulator of cache (behavior all correct) and optimized its performance (topped the class).

Teaching experience

Teaching Assistant to Professor Yanjing Li, *CMSC 14400 - Systems Programming II* Autumn 2023
Teaching Assistant to Professor Borja Sotomayor, *CMSC 14200 - Introduction to Computer Science* Winter 2023
Teaching Assistant to Professor Haryadi Gunawi, *CMSC 15400 - Introduction to Computer Systems* Autumn 2022

Activities, Services and Leadership

UChicago Student Summer CS Research Fellowship Program, Student Lead June - September 2023
UChicago Undergraduate Research Symposium, Quad Scholar Presenter April 2023
CreditEase GenZ Forum, Co-organizer March - September 2021
NiwoFriends Sex Education for Underprivileged Kids, Volunteer January - June 2021

Work Experience

Moody's, Database Infrastructure & Middleware Intern, under Ryan Galloway, SVP June - August 2022
Vision Knight Capital, Investment Intern June - August 2021
Kinzon Capital, Investment Intern (Technology Sector) March - June 2021