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)*.
2. Haochen Pan, **Cyrus Zhou**, Shan Lu, Suman Nath, Madan Musuvathi, "Understanding and Detecting Idempotency Violations in Serverless Applications", *In Preparation.*
3. 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)*.

Research Experience

University of Chicago Chicago, US
Research Assistant to Professor Shan Lu's Group

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

- Co-developed wasabi, a tool that leverages unit tests to validate and track the retry behavior within large-scale distributed systems (currently supporting Java and C#) 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

- This project aimed to clarify the retry mechanisms of major serverless providers and present the first analysis of idempotency violations in serverless applications. Specifically, this study exposed the pervasiveness and severity of the problem as well as illustrated the root causes, symptoms, and fixes. It generalized idempotency violation patterns from statically checkable instances in bug study and used these patterns to find more cases.
- Collected 100+ 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 tested framework for symptom diagnosis.

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

- This project aimed to research the first empirical study of users' end-to-end TAP debugging process, focusing on obstacles users face in debugging TAPs and how well users ultimately fix incorrect automation.
- 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 Yanjing Li's Group

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

- Led and coordinated a team of 5 students.
- Designed and compared multiple encoding methods (i.e., ways to store precision information) for executing mixed-precision neural networks on CPUs.
- Designed CPU SIMD architecture for executing mixed-precision operations, modified the architecture and instructions of the GEM5 simulator to model the behavior of the modified CPU.
- Designed a code generator for mixed-precision neural networks to exploit the modified CPU.
- Replicated and adapted existing mixed-precision schemes on SIMD architecture.

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

- We address the challenges associated with deploying neural networks on CPUs, with a particular focus on minimizing inference time while maintaining accuracy. Our novel approach is to use the dataflow (i.e., computation order) of a neural network to explore data reuse opportunities using heuristic-guided analysis and a

code generation framework, which enables exploration of various Single Instruction, Multiple Data (SIMD) implementations to achieve optimized neural network execution. Our results demonstrate that the dataflow that keeps outputs in SIMD registers while also maximizing both input and weight reuse consistently yields the best performance for a wide variety of inference workloads, achieving up to 3x speedup for 8-bit neural networks, and up to 4.8x speedup for binary neural networks, respectively, over the optimized implementations of neural networks today.

- Proposed the research idea, extended discussion on neural network dataflows, developed a code generation framework for neural networks, set up ARM server and, on top of it, the Gem5 simulator, ran experiments, analyzed results, 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

- 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 preexisting codebases for efficient classification of file sensitivity and making group deletion recommendations.

Selected Projects

DEFT: SLO-Driven Preemptive Scheduling for Containerized DNN Serving Spring 2023

- Applied dynamic scheduling to containerize DNN serving on shared GPU servers, enabling kernel-level preemption and minimizing SLO violations.
- Developed a Job Remaining Time estimator that dynamically estimates remaining time for DNN jobs, making preemption decisions based on the remaining time instead of static weights or individual kernel call duration.
- Implemented a scheduling system compatible with Kubernetes, without requiring changes to user applications.

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.

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, designed, implemented, and integrated graphical interfaces.

Teaching experience

University of Chicago Chicago, US

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

Evaluations can be found here: https://registrar.uchicago.edu/registration/college-process/course-feedback/

Inspire! Education Shanghai, China

Essay Tutor February 2021 - June 2021

Activities, Services and Leadership

UChicago Student Summer CS Research Fellowship Program, Student Lead June - September 2023

The Greater Chicago Area Systems Research Workshop (GCASR), Student Volunteer April 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 June - August 2022

Vision Knight Capital, Investment Intern June - August 2021

Kinzon Capital, Investment Intern (Technology Sector) March - June 2021