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:** 3.91/4.00.
- **Selected Awards:** UChicago Summer Quad Research Scholar (2023), UChicago Quad Research Scholar (2022-2023), UChicago Dean's List (2021-2022, 2022-2023), UChicago Jeff Metcalf Award (May 2022).
- **Selected Courses:** Advanced Operating Systems, Computer Architecture, Introduction to Computer Systems, Machine Learning for Computer Systems, Introduction to Computer Science I-II, Reading and Research in Computer Science, 3D Geometry Processing & Computer Vision, Introduction to Human-Computer Interaction, Introduction to Formal Languages, Theory of Algorithms, Discrete Mathematics, Numerical Linear Algebra, Statistical Theory and Methods.

PUBLICATIONS

1. **Cyrus Zhou**, Zack Hassman, Ruize Xu, Dhirpal Shah, and Yanjing Li, "YFlows: Systematic Dataflow Exploration and Code Generation for Efficient Neural Network Inference using SIMD Architectures on CPUs", March 2024 (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

- Developed Tempura (Testing Engine for Monitoring and Proving Uncovered Retry Areas), a tool that leverage unit tests to validate and track the retry behavior within large-scale distributed systems (currently supporting Java and C#) through code injection.
- 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).

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

Architectural Support for Mixed Precision Neural Networks June 2023 - Present

- Led and coordinated a team of 5 students.
- Added Heterogenous SIMD support (with precision information stored in the metadata register) to the GEM5 simulator.
- Replicated existing mixed-precision schemes on SIMD architecture.

Data Scheduling for Neural Networks on CPUs using SIMD Architecture October 2022 - July 2023

- We addressed 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 demonstrated 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 2.7x speedup for 8-bit neural networks, and up to 4.8x speed up for binary neural networks, respectively, over the state-of-the-art SIMD implementations of neural networks today.

- Proposed the research idea, extended discussion on neural network dataflows, developed a code generation framework for CNNs and Transformers, 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, reinforcement learning, and transfer learning for improved recommendation personalization.
- Developed 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.

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

- Co-advised by Professor Shan Lu and Professor Blase Ur, see above for information on this project.

SELECTED PROJECTS

University of Chicago Chicago, US

Advanced Operating Systems, Professor Shan Lu Spring 2023

DEFT: SLO-Driven Preemptive Scheduling for Containerized DNN Serving

- Applied dynamic scheduling to containerize DNN serving on shared GPU servers, enabling kernel-level preemption and minimizing SLO violations.
- Developed a Job Remaining Time (JRT) 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.
- Compared to existing solutions like Clockwork, DEFT achieved fewer SLO violations in preliminary experiments, showing its effectiveness in preempting DNN inference jobs.

Machine Learning for Computer Systems, Professor Nick Feamster Autumn 2022

Detecting Webcams' Spying Behavior via Machine Learning

- Applied machine learning techniques to detect if any spying application is running on the host computer.
- Empirically collected experiment data performed both multiclass classification (control, only spy application, WeChat Video Call, WeChat + Spy, Zoom Video Conferencing, Zoom + Spy and binary classification (spy vs non-spy).
- Presented our project as a report using Sphinx using nbSphinx to embed Jupyter Notebook.

Computer Architecture, Professor Fred Chong Autumn 2022

ARM Instruction Set Simulator

- 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.

Human-Computer Interaction, Professor Ken Nakagaki Autumn 2022

VR Diver Interface

- Used Unity to design and program an augmented-reality interface for diving into a Virtual Reality Setting.
- The development process includes: prototyping, 3D modelling, programming, debugging, presenting and demoing.

Human-Computer Interaction, Professor Ken Nakagaki Autumn 2022

AudioPong

- Developed an eyes-free and hands-free pong game with audio as the only form of both input and output.

Theory of Algorithms, Professor Lorenzo Orecchia Spring 2022

Best Route in Theme Parks

- Designed and implemented three algorithmic solutions (greedy, dynamic programming, and backtracking) for an NP-hard problem: Orienteering Problems with Time Windows.
- Wrote a program to find the optimal solution among these three, and optimized results with local search
- Ranked Top 6 among about 41 teams in total.

Software Development, Professor Borja Sotomajor Spring 2022

Chiventure - A Game Engine

- Led the graphics team, designed, implemented, and integrated graphical display programs for scenes, non-player characters, game stats, maps, input boxes, and quests.
- Collaborated with corresponding teams to meet the requirements.
- Proposed and developed GDL (Graphics Description Language), a JSON-format data structure that enabled game developers to specify and customize graphical details which were not directly retrievable from the game state.

Introduction to Computer Systems, Professor Shan Lu Winter 2022

Computer Systems Projects

- Took charge of bit-level manipulation & arithmetic and simple shell (argument parsing, fork programs, output redirection).
- Debugged binary bombs by "deciphering" assembly lines.
- Cached optimization with matrix transposition.

WORK EXPERIENCE

Moody's New York, US

Data Management & Back-end Engineer Intern, Moody's Shared Service June - August 2022

- Utilized the Spring Boot API of Collibra to implement its integration with local files, AWS, and databases in general.
- Proposed, drafted, and prototyped a new format that resorted to JSON information files for more lubricated and automated data actions, alleviating the burden on data managers.
- Wrote general templates and programming guides for succeeding developers, connected and visualized metadata from data inventories in PowerBI, and composed several delta files.

TEACHING EXPERIENCE

University of Chicago Chicago, US

Teaching Assistant to Professor Borja Sotomajor Winter 2023

CMSC 14200 - Introduction to Computer Science

- Taught introductory data structures and software development to undergraduate students.
- Led lab discussions, held office hours, answered Ed Discussion questions, and helped set up infrastructures.

Teaching Assistant to Professor Haryadi Gunawi Fall 2022

CMSC 15400 - Introduction to Computer Systems

- Held office hours for four weeks about Operating Systems, answered questions on Ed Discussion and helped students with computer setup and course logistics.

Essay Tutor Shanghai, China

Inspire! Education March - July 2021

- Taught essay writing fundamentals to high school students, covering topics in Economics ("what will happen if there is no government intervention in the education sector?"), Psychology (heuristics vs systematic thinking modes), and Philosophy ("does jury duty count as modern slavery?").
- Conducted weekly discussions and advised students on personal writing projects.

PARTICIPATED SEMINARS

2023 UChicago Undergraduate Research Symposium Chicago, US

Quad Research Scholar May 2023

- Presented our work on scheduling the computations of neural networks on SIMD CPUs for fast inference in 2023 UChicago Undergraduate Research Symposium, the largest campus-wide interdisciplinary undergraduate research symposium. Poster available in the title link.

10th Greater Chicago Area Systems Research Workshop (GCASR 2023) Chicago, US

Participant April 2023

- Presented our work on scheduling the computations of neural networks on SIMD CPUs for fast inference.
- Helped coordinate the workshop.

Quantum & Advanced Technology Career Launcher Chicago, US
Participant March - April 2024

SERVICES AND LEADERSHIP

Student Summer CS Research Fellowship Program Chicago, US
Student Mentor June - September 2021

- Helped organize a summer research program for international students, led by Professor Shan Lu.

CreditEase GenZ Forum Beijing, China
Co-Organizer April - July 2021

- Collaboratively planned and organized the first CreditEase Gaia Planet GenZ Forum, attracting 500+ GenZ applicants, admitted about 300 of them, coordinated 20+ guests, and received an 8.7/10 rating.
- Designed and developed a WeChat Mini Program for the forum, composed the case for business competition, and co-formulated and operated the forum currency.

Anthropology Study on Ghanaian Returnees Accra/Kumasi, Ghana
Individual Researcher June - September 2019

- Inspired by Brit(ish) by Afua Hirsch, conducted an anthropology study on the socio-economic status and identified the crisis of Ghanaian returnees.
- Conducted field studies in Accra and Kumasi and wrote reports on findings and thoughts.

Niwo Friends Sex Education X LSE CSSA Xi'an, China
Volunteer December 2020 to April 2021

- Participated in the recording of a series of sex education courses for elementary school students.

OTHER AWARDS AND HONORS

- **Winner, John Locke Essay Competition** International, 2019
  First Chinese Awardee in history.
  Essay about Neuroscience.
- **Gold Medal, Canadian Chemistry Olympiad** International, 2019
- **Gold Prize, United Kingdom Mathematics Trust Math Challenge** International, 2019
- **Honorable Mention, New York Times Summer Reading Contest** International, 2019
- **Third Prize, China Brain Bee** National, 2019

ADDITIONAL INFORMATION

- **Computer and Language Skills:**
  - ✓ **Proramming Languages:** C/C++, ARM/x86/RISCV Assembly/Intrinsics, Python, Java, SQL, C#, JavaScript, R, Swift, Racket, HTML, CSS, Unix scripting.
  - ✓ **Tools:** GEM5, GCC, Clang, TVM, Arm C++ Compiler, Intel C++ Compiler, Docker, GIT, SVN, AWS, Azure, Google Cloud, CodeQL, Jacoco, Venv, Unity, Fruitloops, Vocaloid, VSCode, Spring Boot, Spring Web, Thymeleaf, Pandas, Numpy, Django, Angular, React, Matlab, Postgres, Power BI, SSH, Sphinx, Jupyter Notebook.
  - ✓ **Libraries/Packages:** OpenCV, Tensorflow, PyTorch, Keras, CUDA, LLVM, GEM5, SimGrid, MPI (Message Passing Interface), Hadoop, Spark, Pandas, Numpy, Scipy, Matplotlib, Seaborn, Bazel, Caffe, ONNX, Kubernetes, Docker, Git.
  - ✓ **Frameworks/Programming Paradigms:** AWS Lambda, Azure Serverless, Trigger-Action Programming/IFTTT, Django, Spring Frameworks, React, Angular, Swift UI.
  - ✓ **Operating Systems:** Linux, MacOS, Windows.

- **Language:** Chinese (native), English (proficient), German (Basic).

- **Interests:** Weightlifting, Traveling, Basketball (High School Varsity), Jazz, R&B, Cooking