Yuxuan Zhang

Last update on October 9, 2025

zyuxuan@seas.upenn.edu • 734-846-1069 • zyuxuan0115.github.io • 3333 Chestnut St., Philadelphia, PA

Education

University of Pennsylvania, Philadelphia, PA
PhD in Computer Science, 2019 - 2026
Advisor: Sebastian Angel
GPA: 4.0/4.0

University of Michigan, Ann Arbor, MI
MS in Electrical Engineering, 2015 - 2017

Harbin Institute of Technology, Harbin, China
BS in Electrical Engineering, 2008 - 2012

Employment History

Google LLC, Sunnyvale, CA
Software Engineer Intern, May 2025 - August 2025

VMWare Corp., Boston, MA
Software Engineer Intern, May 2022 - Aug 2022

Microsoft Research Asia, Beijing, China
Research Intern, Jan 2018 - July 2019
Mentor: Yongqiang Xiong

NVIDIA Corporation, Beijing, China
Software Engineer Intern, May 2017 - Sep 2017

Publications

Quilt: Resource-aware Merging of Serverless Workflows.
Y. Zhang, S. Angel.
Proc. ACM Symposium on Operating Systems Principles (***SOSP***), Oct. 2025

RPG2: Robust Profile-Guided Runtime Prefetch Generation.
Y. Zhang, N. Sobotka, S. Park, S. Jamilan, T. A. Khan, B. Kasikci, G. Pokam, H. Litz, J. Devietti.
International Conference on Architectural Support for Programming Languages and Operating Systems (***ASPLOS***), May. 2024.

Online COde Layout OptimizationS via Ocolos.
Y. Zhang, T. A. Khan, G. Pokam, B. Kasikci, H. Litz, J. Devietti.
IEEE Micro Volume 43, Issue 4, "Top Picks From the 2022 Computer Architecture Conferences", July. 2023.

OCOLOS: Online COde Layout OptimizationS.
Y. Zhang, T. A. Khan, G. Pokam, B. Kasikci, H. Litz, J. Devietti.
Proc. International Symposium on Microarchitecture (***MICRO***), Oct. 2022.

Honor & Awards

Paper selected for IEEE Micro Top Picks in Computer Architecture from 2023

Outstanding Graduates Awards, *Harbin Institute of Technology, 2012*

Fuji Xerox Scholarship, *Harbin Institute of Technology, 2011*
Undergraduate GPA ranking top 1 for one academic year

Suzhou Industry Park Scholarship, *Harbin Institute of Technology, 2010*
Undergraduate GPA ranking top 2 for one academic year

Talks

RPG^2: Robust Profile-Guided Runtime Prefetch Generation. May. 2024
International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), San Diego, CA.

OCOLOS: Online COde Layouy OptimizationS. Oct. 2022
International Symposium on Microarchitecture (MICRO), Chicago, IL.

Professional Service

Journal Reviewer

IEEE Transactions on Computers (TOC), 2023

Teaching

Teaching Assistant for CIS505 Distributed Systems, University of Pennsylvania, Fall 2020, Fall 2021

Selected Research Projects

Quilt: Resource-aware Merging of Serverless Workflows

Distributed System Lab, University of Pennsylvania

Built a feedback-driven serverless runtime optimizer that determines how functions within a workflow are deployed and executed. It dynamically merges serverless functions into a single process, guided by call graph analysis and resource consumption data obtained through distributed tracing and runtime profiling. The goal is to minimize function invocation overhead and improve both throughput and resource utilization.

RPG^2: Robust Profile-Guided Runtime Prefetch Generation

Architecture+Compilers group, University of Pennsylvania

Built an online data cache prefetching system that can profile and analyze the behavior of data memory accesses and then make the decision of whether and where to insert the prefetch instructions into the running process. After prefetches inserted, RPG^2 can monitor and tune the prefetches to maximize performance.

OCOLOS: Online COde Layout Optimization System

Architecture+Compilers group, University of Pennsylvania

Built a code layout optimization tool that optimizes the code layout of datacenter applications at runtime. The optimization process involves profiling and analyzing the application to generate an optimized code layout, which is then injected into the program's text section to enhance performance.

Glane on GPU

Networking Research Group, Microsoft Research Asia

Built a Linux module that can expose an NVIDIA GPU's physical memory for direct data transfer, and a hardware stack for GPUs in a device-centric cluster to buffer and transfer data. Prototyped CUDA code to perform GPU computation and data transfer in parallel without host CPU involvement.

Skills

Programming: C/C++, Python, JavaScript, Rust, System Verilog

Language: Chinese (native), English (professional working proficiency), Japanese (intermediate level)