Yile (Michael) Gu

Address : 5555 14th Ave NW, Seattle, WA   Mobile : +1 734-881-5477
Email : yilegu@cs.washington.edu / guyile1998@gmail.com   Personal Website : https://ikace.github.io/

Education

University of Washington, Seattle, WA
Ph.D. in Computer Science and Engineering Sep 2023 - Present
Advisor: Prof. Baris Kasikci
Research Interests: Systems Reliability, Machine Learning Systems

University of Michigan, Ann Arbor, MI
M.S.E. in Computer Science and Engineering, Cumulative GPA: 4.00/4.00 Aug 2021 - May 2023
B.S.E. in Computer Science, Cumulative GPA: 3.96/4.00 Aug 2019 - May 2021
Award: James B. Angell Scholar, EECS Scholar, Dean's List
Coursework: Compiler Construction, Advanced Operating System, Distributed System, Advanced Computer Vision

Shanghai Jiao Tong University, Shanghai, China
B.E. in Electrical and Computer Engineering, Cumulative GPA: 3.82/4.00, Rank: 11/253 Sep 2017 - Aug 2021
Award: Outstanding Graduate, Merit Student, Undergraduate Excellent Scholarship (Top 10%)
Coursework: Programming & Data Structures, Intro to Signals & Systems, Intro to Logic Design, Electronic Circuits

Research Experience

Efes Lab, University of Washington May 2022 - Present
Project: Application-level Crash-consistency Bug Detection *Supervisor: Prof. Baris Kasikci*

- Spearheaded the development of an application level crash-consistency bug detection tool to address current issues with sub-optimal testing space.
- Leveraged redundancy in programs' update behaviors to build dependency graphs for pruning testing space.
- Developed a Pin tool to trace syscalls & mmap-IOs, and designed an algorithm to test systems with hybrid protocols.
- Assessed the efficacy of the tool on POSIX and persistent-memory applications, leading to 54 new bug discoveries.
- Built an optimized exhaustive testing baseline to demonstrate that our tool can achieve a 32x crash-state reduction.

Symbiotic Lab, University of Michigan Aug 2022 - May 2023
Project: Energy Scheduling in Large Model Training *Supervisor: Prof. Mosharaf Chowdhury*

- Discovered the existence of energy bloat in large model training caused by fundamental computation imbalance, where GPUs waste energy by running unnecessarily faster than the critical path of the computation.
- Represented training schedule as a directed acyclic graph (DAG), and designed a graph cut-based algorithm that exclusively and efficiently enumerates all energy schedules on the "iteration time-energy" Pareto frontier.
- Evaluated on large models including GPT3 that our system reduces energy consumption by up to 28.5% without slowdown in training time, with negligible 6.5-minutes average time for the algorithm. Open-sourced at Perseus.

Project: Privacy-enhancing Federated Learning (FL) Platform

- Identified issues with existing FL platforms which lack fundamental support for privacy accounting under different workloads and various types of heterogeneity.
- Containerized core components of FedScale using Docker for flexible deployment to different operating systems.
- Built a Kubernetes-based coordinator to enable load-balancing support and handle simulated and real FL workloads.
- Designed a privacy-accounting client selector prototype maximizing FL job utility while respecting privacy budget.

Work Experience

Microsoft Research, Redmond, WA, USA June 2024 – Sep 2024
Research Intern *Mentors: Jonathan Mace, Yifan Xiong*

- Proposed an accurate, efficient and explainable agent-based time-series anomaly detection system for AI infra.
- Evaluated the system on public anomaly detection dataset, showing a 26% increase in avg F1 score for 27 KPI metrics.

ByteDance Ltd, Shanghai, China May 2020 – Aug 2020
Software Engineering Intern *Mentors: Jilong Liu, Dong Li*

- Contributed to a cross-platform mobile application framework with native UI features using C++ and Objective-C.
- Detected and resolved performance bugs in the framework, including a serious memory leak due to circular reference.
- Developed customized components with improved efficiency in rendering logic for mobile application developers.

Professional Service

- **Reviewer:** ICLR 2025
- **Artifact Evaluation Committee:** OSDI 2023, ATC 2023
- **External Reviewer:** SOSP 2023, OSDI 2024, MICRO 2024, SOSP 2024
- **Student Volunteer:** NSF NeTS PI Meeting 2023

Peer-reviewed Publications

[1] Scalable and Accurate Application-level Crash-Consistency Testing via Representative Testing. **Yile Gu***, Ian Neal*, Jiexiao Xu, Shaun Christopher Lee, Ayman Said, Musa Haydar, Jacob Van Geffen, Rohan Kadekodi, Andrew Quinn, Baris Kasikci. (To Appear) OOPSLA 2025, Singapore, October 2025. https://arxiv.org/abs/2503.01390.

[2] Fiddler: CPU-GPU Orchestration for Fast Inference of Mixture-of-Experts Models. Keisuke Kamahori*, Tian Tang*, **Yile Gu**, Kan Zhu, Baris Kasikci. (To Appear) ICLR 2025, Singapore, May 2025. https://arxiv.org/abs/2402.07033.

[3] Perseus: Removing Energy Bloat from Large Model Training. Jae-Won Chung, **Yile Gu**, Insu Jang, Luoxi Meng, Nikhil Bansal, Mosharaf Chowdhury. SOSP 2024, Austin, TX, USA, November 2024. https://doi.org/10.1145/3694715.3695970.

[4] Argos: Agentic Time-Series Anomaly Detection with Autonomous Rule Generation via Large Language Models. **Yile Gu**, Yifan Xiong, Jonathan Mace, Yuting Jiang, Yigong Hu, Baris Kasikci, Peng Cheng. https://arxiv.org/abs/2501.14170.

[5] Tactic: Adaptive Sparse Attention with Clustering and Distribution Fitting for Long-Context LLMs. Kan Zhu*, Tian Tang*, Qinyu Xu*, **Yile Gu**, Zhichen Zeng, Rohan Kadekodi, Liangyu Zhao, Ang Li, Arvind Krishnamurthy, Baris Kasikci. https://arxiv.org/abs/2502.12216.

[6] TeleRAG: Efficient Retrieval-Augmented Generation Inference with Lookahead Retrieval. Chien-Yu Lin*, Keisuke Kamahori*, Yiyu Liu, Xiaoxiang Shi, Madhav Kashyap, **Yile Gu**, Rulin Shao, Zihao Ye, Kan Zhu, Stephanie Wang, Arvind Krishnamurthy, Rohan Kadekodi, Luis Ceze, Baris Kasikci. https://arxiv.org/abs/2502.20969.

[7] NanoFlow: Towards Optimal Large Language Model Serving Throughput. Kan Zhu, Yilong Zhao, Liangyu Zhao, Gefei Zuo, **Yile Gu**, Dedong Xie, Yufei Gao, Qinyu Xu, Tian Tang, Zihao Ye, Keisuke Kamahori, Chien-Yu Lin, Stephanie Wang, Arvind Krishnamurthy, Baris Kasikci. https://arxiv.org/abs/2408.12757.

Selected Projects

Enabling Loop Fusion in LLVM by Moving Intervening Code, University of Michigan

- Observed that current LLVM implementation of loop fusion, a powerful compiler optimization that enables better loop distribution and software pipelining, requires unnecessarily strict matching criteria.
- Designed an algorithm attempting to move intervening code before loop fusion if two loop candidates are not adjacent by analyzing data dependencies and determining the correct location for the intervening code.
- Evaluated on microbenchmarks that our algorithm achieves a 40% reduction in running time and a 12% decrease in dynamic instructions executed on average.

Understanding Data Privacy and Byzantine Resilience in Distributed ML, University of Michigan

- Observed theoretical upper bound for combining data privacy and Byzantine resilience with batch size as a bottleneck.
- Determined that a large batch size is required for the convergence of CNN models under Gaussian noise injection.
- Applied gradient sparsification for privacy amplification to account for the fundamental privacy-utility trade-off.
- Discovered that batch size directly correlates with attackers' ability to reconstruct individual images from gradients.

Teaching

GSI of Foundation of Computer Science, University of Michigan	Jan 2022 – May 2022 & Aug 2022 - Dec 2022
IA of Academic Writing II and Fantasy Literature, UM-SJTU Joint Institute	Feb 2019 - Aug 2019

Skills

- **Programming:** C++, Python, JavaScript, SQL **Markup Languages:** HTML, LaTeX, Markdown