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 - Apr 2023
B.S.E. in Computer Science, Cumulative GPA: 3.96/4.00 Aug 2019 - Apr 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: Benchmarking Generative AI Applications on End-User Devices *Supervisor: Prof. Baris Kasikci*

- Proposed a comprehensive benchmarking framework for evaluating GenAI applications under realistic multi-application workloads on resource-constrained consumer devices. Open-sourced at ConsumerBench.
- Built a DAG–based workflow engine to orchestrate heterogeneous GenAI applications (e.g., chatbots, image generation, live captioning), capturing both application-level SLOs and system-level metrics (GPU/CPU utilization, memory bw).
- Revealed key inefficiencies in resource sharing, including starvation under greedy GPU allocation and underutilization under static partitioning, motivating the need for dynamic, SLO-aware scheduling and backend-aware kernel designs.

Project: Efficent and Accurate Application-level Crash-consistency Bug Detection

- Spearheaded the development of an application level crash-consistency bug detection tool to address current issues with sub-optimal testing space. Open-sourced at Pathfinder.
- 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 - Apr 2023
Project: Reducing Energy Bloat 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 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.

Work Experience

Google LLC, Seattle, WA, USA Jan 2026 – Aug 2026
Student Researcher *Mentor: Stanko Novakovic*

- Designing and implementing an automated system for detecting and diagnosing performance bugs in ML infrastructure.

University of Washington, Seattle, WA, USA Sep 2023 – Present
Research Assistant

- Conduct research in systems reliability and machine learning systems, focusing on improving the reliability, efficiency, and performance of modern computing infrastructure.

- Design and build systems for benchmarking generative AI applications on resource-constrained devices, including workflow engines that orchestrate applications and collect application-level and system-level performance metrics.
- Conduct research on crash-consistency bug detection by developing program analysis and tracing techniques to reduce testing space and identify reliability bugs in storage applications.

Amazon, Santa Clara, CA, USA June 2025 – Dec 2025
Applied Scientist Intern *Mentors: Zhen Zhang, Mason Fu*

- Designed and implemented a differential debugging framework (Ekka) that automates the diagnosis of silent accuracy bugs in production LLM serving engines by systematically aligning components with reference implementations.
- Achieved a 24% improvement in diagnosis accuracy over state-of-the-art systems on a curated benchmark of real-world accuracy bugs, and successfully diagnosed 4 new bugs in production frameworks that were confirmed by developers.

Microsoft Research, Redmond, WA, USA June 2024 – Sep 2024
Research Intern *Mentors: Jonathan Mace, Yifan Xiong*
Project: Agentic Time-Series Anomaly Detection with Autonomous Rule Generation

- Designed and implemented an agentic time-series anomaly detection system that autonomously trains explainable and reproducible rules from time-series data using large language models. Open sourced at Argos.
- Proposed a multi-agent training loop with rule mutation and top-k selection for accurate and efficient rule generation.
- Evaluated Argos on public cloud metric datasets (KPI, Yahoo) and a large internal AI infra dataset, achieving up to 29.3% F_1 score improvement over state-of-the-art systems and 3.0×–34.2× inference time speedups.
- Demonstrated Argos's practicality by detecting real-world incidents (e.g., GPU NCCL hangs) in production cloud monitoring, preventing resource waste and service degradation in time.

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 2025/2026/2027, ICML 2026, NeurIPS 2026, AAAI 2027, TMLR
- **Artifact Evaluation Committee:** OSDI 2023, ATC 2023
- **Student Volunteer:** NSF NeTS PI Meeting 2023

Peer-reviewed Publications

[1] The Streaming Batch Model for Efficient and Fault-Tolerant Heterogeneous Execution. Frank Sifei Luan*, Ron Yifeng Wang*, **Yile Gu**, Ziming Mao, Charlotte Lin, Amog Kamsetty, Hao Chen, Cheng Su, Balaji Veeramani, Scott Lee, SangBin Cho, Clark Zinzow, Eric Liang, Ion Stoica, Stephanie Wang. (To Appear) NSDI 2027, Providence, RI, May 2027. https://arxiv.org/abs/2501.12407.

[2] ConsumerBench: Benchmarking Generative AI Applications on End-User Devices. **Yile Gu***, Rohan Kadekodi*, Sahi Chitrapu, Aradhya Agrawal, Camille Sawa, Hoang Nguyen, Keisuke Kamahori, Yiyu Liu, Baris Kasikci. COLM 2026, San Francisco, CA, Oct 2026. https://arxiv.org/abs/2506.17538.

[3] Ekka: Automated Diagnosis of Silent Errors in LLM Inference. **Yile Gu**, Zhen Zhang, Shaowei Zhu, Xinwei Fu, Jun Wu, Yida Wang, Baris Kasikci. ICML 2026, Seoul, South Korea, July 2026. https://arxiv.org/abs/2606.04594.

[4] DynaFlow: Transparent and Flexible Intra-Device Parallelism via Programmable Operator Scheduling. Yi Pan, **Yile Gu**, Jinbin Luo, Yibo Wu, Ziren Wang, Hongtao Zhang, Ziyi Xu, Shengkai Lin, Baris Kasikci, Stephanie Wang. MLSys 2026, Bellevue, WA, USA, May 2026.

[5] 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, Rohan Kadekodi, Stephanie Wang, Arvind Krishnamurthy, Rohan Kadekodi, Luis Ceze, Baris Kasikci. MLSys 2026, Bellevue, WA, USA, May 2026. https://arxiv.org/abs/2502.20969.

[6] 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. ICLR 2026, Rio de Janeiro, Brazil, April 2026. https://arxiv.org/abs/2502.12216.

[7] Mitigating Application Resource Overload with Targeted Task Cancellation. Yigong Hu, Zeyin Zhang, Yicheng Liu, **Yile Gu**, Shuangyu Lei, Baris Kasikci, Peng Huang. SOSP 2025, Seoul, Republic of Korea, November 2025. https://dl.acm.org/doi/10.1145/3731569.3764835.

[8] 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. OOPSLA 2025, Singapore, October 2025. https://arxiv.org/abs/2503.01390.

[9] NanoFlow: Towards Optimal Large Language Model Serving Throughput. Kan Zhu, Yufei Gao, Yilong Zhao, Liangyu Zhao, Gefei Zuo, **Yile Gu**, Dedong Xie, Zihao Ye, Keisuke Kamahori, Chien-Yu Lin, Ziren Wang, Stephanie Wang, Arvind Krishnamurthy, Baris Kasikci. OSDI 2025, Boston, MA, USA, July 2025. https://arxiv.org/abs/2408.12757.

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

[11] 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.

Preprints

[1] 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.

[2] Semantic Scheduling for LLM Inference. Wenyue Hua*, Dujian Ding*, **Yile Gu**, Yujie Ren, Kai Mei, Minghua Ma, William Yang Wang. https://arxiv.org/abs/2506.12204.

[3] AgentFlux: Decoupled Fine-Tuning & Inference for On-Device Agentic Systems. Rohan Kadekodi*, Zhan Jin*, Keisuke Kamahori, **Yile Gu**, Sean Khatiri, Noah H Bayindirli, Sergey Gorbunov, Baris Kasikci. https://arxiv.org/abs/2510.00229.

Teaching

TA of Systems for All, University of Washington	Jan 2025 - Mar 2025
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

Awards

1st, 2nd, and 3rd place across three tracks in FlashInfer AI Kernel Generation Contest, MLSys 2026	May 2026
Azure GenAI for Science Hub (15K USD)	Jan 2025