ZIKUN LI

Mobile: (+1)4248320069 ◇ Email: zikunl@andrew.cmu.edu ◇ GitHub: zikun-li
Homepage: https://zikun-li.github.io

EDUCATION

School of Computer Science, Carnegie Mellon University (CMU) Pittsburgh, PA, US
Doctoral student in Computer Science Aug 2022 - Present

- Advised by Prof. Zhihao Jia
- Research interests: LLM Systems, Reinforcement Learning, Efficient LLM Inference

School of Electronic Engineering and Computer Science, Peking University (PKU) Beijing, China
Bachelor of Science in Computer Science Sep 2017 - Jul 2021

PUBLICATION

1. **Zikun Li***, Zhuofu Chen*, Remi Delacourt, Gabriele Oliaro, Zeyu Wang, Qinghan Chen, Shuhuai Lin, April Yang, Zhihao Zhang, Zhuoming Chen, Sean Lai, Xinhao Cheng, Xupeng Miao, and Zhihao Jia, *"AdaServe: Accelerating Multi-SLO LLM Serving with SLO-Customized Speculative Decoding"*, EuroSys '26 (* indicates equal contribution).
2. Lijie Yang, Zhihao Zhang, Zhuofu Chen, **Zikun Li**, and Zhihao Jia, *"Tidaldecode: Fast and accurate LLM decoding with position persistent sparse attention"*, ICLR '24.
3. **Zikun Li**, Jinjun Peng, Yixuan Mei, Sina Lin, Yi Wu, Oded Padon, and Zhihao Jia, *"Quarl: A Learning-Based Quantum Circuit Optimizer"*, OOPSLA '24.
4. Mingkuan Xu, **Zikun Li**, Oded Padon, Sina Lin, Jessica Pointing, Auguste Hirth, Henry Ma, Jens Palsberg, Alex Aiken, Umut A.Acar, and Zhihao Jia, *"Quartz: Superoptimization of Quantum Circuits"*, PLDI '22.
5. Zheng Zhong*, Shen Yan*, **Zikun Li***, Decheng Tan, Tong Yang, Bin Cui, 2021, *"BurstSketch: Finding Bursts in Data Streams"*, SIGMOD '21 (* indicates equal contribution).
6. Jizhou Li*, **Zikun Li***, Yifei Xu*, Shiqi Jiang, Tong Yang, Bin Cui, Yafei Dai, Gong Zhang, 2020, *"WavingSketch: An Unbiased and Generic Sketch for Finding Top-k Items in Data Streams"*, KDD '20 (* indicates equal contribution).

EXPERIENCES

Operator-Disaggregated LLM Serving over Heterogeneous GPUs Pittsburgh, PA (Remote with ByteDance Seed)
Student Researcher, ByteDance Seed-Training-Infra & CMU, Mentor: Dr. Ziheng Jiang Sep 2025 - Present

- Studied emerging **operator-disaggregated** LLM serving, quantifying its benefits on throughput and cost and identifying optimal disaggregation strategies.
- Developed a search algorithm that, given an LLM and heterogeneous GPUs, outputs the optimal end-to-end serving plan (operator partitioning, parallelization strategy, device assignment, and batch sizing).
- Built a distributed serving runtime on **vLLM** and **NCCL** to enable flexible operator-disaggregated execution across heterogeneous GPUs.
- Preliminary evaluation shows up to **30% cost reduction** under latency SLO requirements.

Robust Loop Scheduler in Machine Learning Compiler for Trainium Santa Clara, CA
Applied Scientist Intern, AWS Neuron Science Team, Mentor: Dr. Ziyang Xu May 2024 - Aug 2024

- Studied and verified the source of uncertainties in the latency of the scheduling units.
- Designed and implemented a robust loop scheduler that is resistant to uncertainties.
- Achieved up to **1.46× speedup** on attention kernels with the proposed algorithm.

TEACHING EXPERIENCE

Teaching assistant (Scheduled) 15-442/642 Machine Learning Systems Jan 2026 - May 2026

Teaching assistant 15-418/618 Parallel Computer Architecture and Programming Aug 2023 - Dec 2023

SERVICE

Program Committee Member, MLSys 2026

TECHNICAL SKILLS

Programming Languages	Python, C/C++, Triton
Tools	Git, GitHub, Weights & Biases, Docker, Nsight Systems, Nsight Compute, Hugging Face
Frameworks	PyTorch, CUDA, JAX