

Kan Zhu

✉ kanzhu@cs.washington.com ☎ 734-596-2015

About Me

I am Kan Zhu, a second year PhD student at University of Washington's Paul G. Allen School of Computer Science and Engineering, co-advised by [Baris Kasikci](#) and [Arvind Krishnamurthy](#).

Research Interests

I develop systems and methodologies for optimizing Large Language Model (LLM) inference. The widespread adoption of LLMs presents unique challenges for on-device inference and cost-effective large-scale serving due to their substantial computational demands. To address these issues, I am interested in designing innovative hardware, algorithms, and frameworks tailored for both edge devices and data center environments.

Education

PhD	University of Washington , Computer Science and Engineering • Advisor: Prof. Baris Kasikci and Prof. Arvind Krishnamurthy	Sept 2023 – Now
BS	University of Michigan , Computer Engineering • Award: University Honors, Dean's List, James B. Angell Scholar	Sept 2021 - Sept 2023
BS	Shanghai Jiao Tong University , Electrical and Computer Engineering • Award: Outstanding Graduate, Undergraduate Excellent Scholarship	Sept 2019 - Sept 2021

Awards

Allen School Computer Science & Engineering Research Fellowship	Mar 2023
ACM Student Research Competition 1st Place Award (MICRO'22) • Presented a poster and gave 10 min talk on micro-architectural implications of Google applications	Oct 2022
OSDI Travel Grant	Jul 2024

Publications

NanoFlow: Towards Optimal Large Language Model Serving Throughput. <i>Kan Zhu</i> , 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	arXiv'24
QUEST: Query-Aware Sparsity for Efficient Long-Context LLM Inference. Jiaming Tang, Yilong Zhao, <i>Kan Zhu</i> , Guangxuan Xiao, Baris Kasikci, Song Han	ICML'24
Atom: Low-Bit Quantization for Efficient and Accurate LLM Serving. Yilong Zhao, Chien-Yu Lin, <i>Kan Zhu</i> , Zihao Ye, Lequn Chen, Size Zheng, Luis Ceze, Arvind Krishnamurthy, Tianqi Chen, Baris Kasikci	MLSys'24
Fiddler: CPU-GPU Orchestration for Fast Inference of Mixture-of-Experts Models. Keisuke Kamahori, Yile Gu, <i>Kan Zhu</i> , Baris Kasikci	CoRR abs'24
Can Storage Devices be Power Adaptive? Dedong Xie, Theano Stavrinou, <i>Kan Zhu</i> , Simon Peter, Baris Kasikci, Thomas E. Anderson	HotStorage'24

Open Source Projects

NanoFlow, a throughput-oriented LLM serving framework

[efeslab/Nanoflow](https://github.com/efeslab/Nanoflow) 

- Constructed a high-performance serving pipeline using CUDA.
- Achieved up to 72% of optimal serving throughput.

Talks

NanoFlow: Towards Optimal Large Language Model Serving Throughput.

- Carnegie Mellon University
- ByteDance

Oct 2024
Sept 2024

Research Mentoring

Yilong Zhao (SJTU BS -> UCB PhD)

2022-2023

- Activation and weight quantization for LLM

Yuqi Mai (Umich BS -> Cornell PhD)

2022-2023

- Cache prefetcher throttling for Google Traces

Yuewen Hou (Umich BS -> Umich PhD)

2022-2023

- Optimal cache replacement policy for generic cache

Teaching

Shanghai Jiao Tong University, VG 101

Summer 2021

- Teaching Assistant for undergraduate course, Introduction to Programming.