GABRIELE OLIARO

goliaro@cs.cmu.edu

EDUCATION

Carnegie Mellon University — Aug 2022 - Present
Ph.D. in Computer Science. Advisor: Zhihao Jia

Tsinghua University — Aug 2021 - June 2023
M.S in Advanced Computing. Advisor: Jidong Zhai — *Awarded full-ride merit scholarship*

Harvard University — Aug 2017 - May 2021
B.S. in Electrical Engineering. Advisor: Minlan Yu — *Cum Laude with High Honors (GPA 3.83)*

EXPERIENCE

CMU Catalyst Group — Aug 2022 - present
PhD Researcher with Prof. Zhihao Jia — *Carnegie Mellon University*

- (SpecInfer project): Co-design ML and systems optimizations to reduce LLM inference latency by 1.5× - 3.5×. Build multi-node, multi-GPU distributed runtime with efficient CUDA kernels. Lead development of open-source FlexFlow Serve (1.8K stars on GitHub), deploying the SpecInfer algorithm. 🏆 First author ASPLOS '24 paper, cited 250+ times.
- (FlexLLM project): Designed FlexLLM, the first system to co-serve LLM inference and parameter-efficient finetuning, leveraging token-level finetuning, dependent parallelization, and graph pruning to reduce GPU memory overhead by up to 8× and preserve over 80% finetuning throughput under heavy inference loads.
- (AdaServe project): Co-developed AdaServe, a novel LLM serving system that enables fine-grained SLO-customized inference through speculative decoding, achieving up to 73% higher SLO attainment and 74% higher goodput compared to state-of-the-art systems.
- (QST project): Mentor the design of a quantization-based LLM fine-tuning technique requiring up to 2.3× less GPU memory and up to 3× lower fine-tuning time compared to PEFT. 🏆 Outstanding Paper Award at ACL '24

Snowflake AI Research — May 2024 - Present
Research Intern — *San Mateo, CA*

- Designed and implemented SuffixDecoding, a model-free speculative decoding method that accelerates LLM inference using suffix trees to predict token sequences, achieving up to 2.9× higher throughput and 3× lower latency compared to state-of-the-art methods.

Whist Technologies, Inc — Sept 2021 - Aug 2022
Software Engineer — *New York, NY*

- Implemented client-server profiling suite for a streaming application. Built scalable back-end components in Go to efficiently and securely handle user data in the cloud. Profiled and improved performance of low-latency network streaming protocol in C/C++.

Systems + Theory Lab @ Harvard University — Sept 2020 - Feb 2022
Undergraduate Researcher with Prof. Minlan Yu — *Cambridge, MA*

- Designed scalable RDMA protocol for real-time in-band network telemetry in large datacenters. Implemented protocol in P4, deployed on Tofino, and ran simulations in Mininet and ns-3. Co-authored papers at SIGCOMM '23 and HotNets '21.

RISE Lab @ UC Berkeley — Jun 2020 - Aug 2020
Undergraduate Researcher with Prof. Ion Stoica — *Berkeley, CA*

- Designed a work-stealing mechanism, together with task pipelining technique to improve the throughput of distributed Ray framework. Committed to the GitHub repo (35.2K stars): ray-project/ray

RECENT PUBLICATIONS

First-author papers

- **(ArXiv '24) SuffixDecoding: A Model-Free Approach to Speeding Up Large Language Model Inference** Gabriele Oliaro, Zhihao Jia, Daniel Campos, Aurick Qiao. [pdf]
- **(ArXiv '24) FlexLLM: A System for Co-Serving Large Language Model Inference and Parameter-Efficient Fine-tuning**. Xupeng Miao*, Gabriele Oliaro*, Xinhao Cheng, Mengdi Wu, Colin Unger, Zhihao Jia. [pdf]
- **(ASPLOS '24) SpecInfer: Accelerating Generative Large Language Model Serving with Speculative Inference and Token Tree Verification**. Xupeng Miao*, Gabriele Oliaro*, Zhihao Zhang*, Xinhao Cheng*, Zeyu Wang, Rae Ying Yee Wong, Alan Zhu, Lijie Yang, Xiaoxiang Shi, Chunan Shi, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, Zhihao Jia. [pdf] 🏆 Cited 250+ times

Collaborations

- **(ArXiv '25) SpecReason: Fast and Accurate Inference-Time Compute via Speculative Reasoning**
  Rui Pan, Yinwei Dai, Zhihao Zhang, Gabriele Oliaro, Zhihao Jia, Ravi Netravali. [pdf]
- **(ArXiv '25) AdaServe: SLO-Customized LLM Serving with Fine-Grained Speculative Decoding**
  Zikun Li*, Zhuofu Chen*, Remi Delacourt, Gabriele Oliaro, Zeyu Wang, Qinghan Chen, Shuhuai Lin, April Yang, Zhihao Zhang, Zhuoming Chen, Sean Lai, Xupeng Miao, Zhihao Jia. [pdf]
- **(ACL '24 Oral) Quantized Side Tuning: Fast and Memory-Efficient Tuning of Quantized Large Language Models**. Zhengxin Zhang, Dan Zhao, Xupeng Miao, Gabriele Oliaro, Qing Li, Yong Jiang, Zhihao Jia. [pdf] 🏆 Outstanding Paper Award
- **(ASPLOS '24) Optimal Kernel Orchestration for Tensor Programs with Korch**. Muyan Hu, Ashwin Venkatram, Shreyashri Biswas, Balamurugan Marimuthu, Bohan Hou, Gabriele Oliaro, Haojie Wang, Liyan Zheng, Xupeng Miao, Jidong Zhai, Zhihao Jia
- **(ACM Comput. Surv.) Towards Efficient Generative Large Language Model Serving: A Survey from Algorithms to Systems**. Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Hongyi Jin, Tianqi Chen, Zhihao Jia. [pdf]
- **(SIGCOMM '23) Direct Telemetry Access**. Jonatan Langlet, Ran Ben Basat, Gabriele Oliaro, Michael Mitzenmacher, Minlan Yu, Gianni Antichi. [pdf]
- **(HotNets '21) Zero-CPU Collection with Direct Telemetry Access**. Jonatan Langlet, Ran Ben Basat, Sivaram Ramanathan, Gabriele Oliaro, Michael Mitzenmacher, Minlan Yu, Gianni Antichi. [pdf]

* Equal contribution

SELECT PROJECTS

Chickadee OS: Designed a multi-core OS kernel in C/C++. Implemented virtual memory, buddy allocator, processes, threads, wait queues, file system, disk support, buffer cache, signals and system calls.

LSM-Tree based Key-Value Store: Designed and implemented in C++ a NoSQL key-value store using a Log-Structured Merge Tree and an in-memory skip list.

TMALL Repeat Buyers Prediction: Built an ensemble model based on gradient boosting to predict repeat buyers. Achieved Top 0.6% ranking on Tianchi challenge portal on aliyun.com

XRDict: Implemented a basic reverse dictionary leveraging XLM-R, a cross-lingual understanding language model.