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 - Aug 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 (1.4K stars on GitHub), deploying the SpecInfer algorithm.
- (FlexLLM project): Use ML compilation and novel batching techniques, together with optimized CUDA kernels, to efficiently multiplex LLM inference and fine-tuning. Reduce GPU memory activations overhead by up to 8×.
- (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.

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.

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 (29.7K stars): ray-project/ray

SELECT PUBLICATIONS

(ArXiv '24b) FlexLLM: A System for Co-Serving Large Language Model Inference and Parameter-Efficient Finetuning. Xupeng Miao*, Gabriele Oliaro*, Xinhao Cheng, Mengdi Wu, Colin Unger, Zhihao Jia. [pdf]

(ArXiv '24a) 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]

(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]

(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

(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.