SHIYI CAO

shicao@berkeley.edu ◇ (650) - 304 7475 ◇ https://shiyicao.com/

EDUCATION

UC Berkeley *Aug. 2023 - current*
Department of Computer Science
Ph.D. in Computer Science

ETH Zurich *Sept. 2020 - Jun. 2023*
Department of Computer Science
M.S. in Computer Science

Shanghai Jiao Tong University *Sept. 2016 - 2020*
School of Electronic Information and Electrical Engineering
B.S. in Computer Science and Technology

RESEARCH INTEREST

My general research interests lie in the fields of distributed systems and high-performance computing, with a focus on understanding and accelerating emerging applications on heterogeneous systems. I am currently working on building efficient and scalable LLM inference/training systems.

SELECTED PUBLICATIONS

1. **Shiyi Cao**, Salvatore Di Girolamo, and Torsten Hoefler. Accelerating Data Serialization/Deserialization Protocols with In-Network Compute. In *Workshop on Exascale MPI, ExaMPI@SC*, 2022.
2. **Shiyi Cao**, Yuanning Gao, Xiaofeng Gao, and Guihai Chen. Adam: An adaptive fine-grained scheme for distributed metadata management. In *International Conference on Parallel Processing (ICPP)*, 2019.

SELECTED WORK UNDER SUBMISSION

Fairness in Serving Large Language Models *Oct. 2023 - Dec. 2023*
Sky, Berkeley, Advisor: Ion Stoica and Joseph E. Gonzalez

- Devised a novel scheduling algorithm that achieves fairness guarantee for LLM serving.

S-LoRA: Serving Thousands of Concurrent LoRA Adapters *Aug. 2023 - Nov. 2023*
Sky, Berkeley, Advisor: Ion Stoica and Joseph E. Gonzalez

- Developed a scalable and efficient system for serving thousands of LoRA adapters concurrently, optimizing the batched LoRA computation and memory management.

High-performance Quantum Circuits Simulation *Jan. 2023 - Apr. 2023*
Catalyst, CMU, Advisor: Zhihao Jia

- Developed a scalable and efficient system for quantum circuits simulation on GPUs, exploiting data locality and optimizing communication cost.

Graph Pipeline Parallelism for DL Model Training *July. 2022 - Dec. 2022*
Catalyst, CMU, Advisor: Zhihao Jia

- Led the end-to-end implementation for enabling graph pipeline parallelism training strategies on FlexFlow.

SELECTED PROJECTS

Barrelfish OS Development *Mar. 2022 - Jun. 2022*
Advanced Operating System Course by David Cock and Prof. Timothy Roscoe

- Implemented our own memory management, paging, message passing, inter-core communication etc. on Barrelfish research operating system.
- Implemented and benchmarked the Network stack.

Distributed DL Training on Bagua *Oct. 2021 - Jan. 2022*
DS3Lab, ETH, Advisor: Jiawei Jiang and Prof. Ce Zhang

- Port, improve and benchmark existing distributed deep compression training algorithms to Bagua, a deep learning trainging acceleration framework for PyTorch.

High-performance Image Compression Implementation *Mar. 2021 - June. 2021*
Advanced System Lab Course Project

- Designed highly optimized implementations of the whole SPIHT image compression pipeline, leveraging techniques such as SIMD vectorization, memory rearrangement, and blocking.
- Our best optimized version achieves a runtime speedup of 100x and 200x for encoding and decoding respectively compared with the baseline implementation.

TALKS & PRESENTATIONS

Participant, Workshop on Exascale MPI @ SC *Nov. 2022*

- Made oral presentation for the accepted paper Accelerating Data Serialization/Deserialization Protocols with In-Network Compute.

Participant, International Conference on Parallel Processing *Aug. 2019*

- Made oral presentation for the accepted paper Adam: An adaptive fine-grained scheme for distributed metadata management.

AWARDS

- Academic Excellence Scholarship (Second Class), 2016-2017
- Academic Excellence Scholarship (Third Class), 2018-2019
- Meng Minwei International Exchange Fund (12000RMB), 2019

SKILLS

English Proficiency	GRE: 329 + 4.0 (V:160 Q:169 AW:4.0), TOEFL: 109
Programming	C, C++, Python, PyTorch, CUDA/Triton, SSE/AVX, Tensorflow
Softwares	Latex, Matlab, Unity3D
GitHub	https://github.com/caoshiyi