SHIYI CAO

shiyicao314@gmail.com ⋄ +86 18190733275

EDUCATION

UC Berkeley *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 computer systems and high-performance computing, with a focus on understanding and accelerating emerging applications on heterogeneous 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.

RESEARCH EXPERIENCE

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

- Leading the end-to-end implementation for enabling generalized parallelism training strategies.

Accelerating Data SerDes with In-Network Compute *Aug. 2021 - Aug. 2022*
SPCL Lab, ETH Zurich, Advisor: Salvatore Di Girolamo and Prof. Torsten Hoefler

- Offloaded data deserialization (ProtoBuf) to SmartNIC for efficient RPC framework.
- Designed the deserialization algorithm to enable parallel and streaming processing on the NIC.
- Published the paper on SC'22 ExaMPI Workshop as the *first author*.

Deep Reinforcement Learning in Distributed Metadata Management *Sept. 2018 - Jan. 2019*
Advanced Network Laboratory, SJTU, Advisor: Prof. Xiaofeng Gao

- Introduced for the first time deep reinforcement learning in distributed metadata management.
- Proposed an adaptive fine-grained metadata management scheme AdaM, leveraging deep reinforcement learning.
- Conducted experiments on real-world data traces and compared AdaM with strong baselines.
- Demonstrated that AdaM can address the trade-off between load balance and locality preservation cost-effectively and is highly adaptive to time-varying access pattern.

- Published the paper in ICPP'19 (International Conference on Parallel Processing) as the *first author*.

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++, SSE/AVX, Python, PyTorch, Tensorflow
Softwares	Latex, Matlab, Unity3D
GitHub	https://github.com/caoshiyi