Zhuohao Li

zhuohaol@ucla.edu • zhuohaoli.com

Education

University of California, Los Angeles — **Los Angeles, CA, United States**
Ph.D., Computer Engineering — *Sep 2023 – (Expected) Aug 2028*
Research Interests: Computer Architecture, Systems Security, Machine Learning Systems
Advisor: Prof. Yuan Tian

Shanghai Jiao Tong University — **Shanghai, China**
BSc., Computer Engieering, GPA: 3.99/4.3 — *Sep 2019 – Jun 2023*
Advisor: Prof. Jingwen Leng

The Hong Kong University of Science and Technology — **Hong Kong S.A.R.**
Exchange, Computer Science — *Sep 2022 - Jun 2023*
Advisor: Prof. Wei Wang

Awards & Honors

- **HKUST Fellowship** ($45,000) — **CSE, HKUST** (*2023*)
- **SenseTime Scholarship** (**30** nationally in China) — **SenseTime Inc.** (*2022*)
- **Irving T. Ho PhD Scholarship** (**4** among ~**15,000** SJTU) — **Irving T. Ho PhD Foundation** (*2022*)
- **Overseas Research Scholarship** — **EECS, Shanghai Jiao Tong University** (*2022*)
- **SJTU Academic Excellence Scholarship** — **Shanghai Jiao Tong University** (*2020,2021,2022*)
- **Golden Medel in Google Cup** — **Google** (*2021*)
- **First Prize in National Mathematics Olympic Competition** — **Chinese Mathematics Society** (*2019*)

Research Experience

Big Data Institute — HKUST

Visiting Research Intern, advised by Prof. Wei Wang and Ruichuan Chen (Bell Lab) — *Sep 2022 - Jun 2023*

- Worked on **Serverless GPU**, an efficiently disaggregate GPU system on cloud.
  - Proposed *memory swapping* to swap GPU memory of an instance between main memory or other GPUs.
  - Designed a scheduler to perform swap performance and manage cluster nodes.
  - Developed a standalone to enable CUDA API remoting.
  - Designed asynchronous, model pipelining features to optimize data transmission in scheduling.

Speedway Group — The University of Texas at Austin

Remote Research Intern, advised by Prof. Calvin Lin — *Apr 2022 - Sep 2022*

- Worked on **mBelady**, an optimal multi-level cache replacement policy to sense hierarchical cache system.
  - Proposed an augmented design paradigm to online policies including Hawkeye, Harmony, Mockingjay to be aware of cache hierarchy.
  - Modeled cache system formally and performed a math proof of its optimality.
  - Simulated the policies in ChampSim and developed APIs including *promotion*, *demotion*, *selective insertion* and *bypassing*. Evaluated the system on a subset of SPEC benchmark suite.

Emerging Parallel Computing Center — Shanghai Jiao Tong University

Undergraduate Researcher, advised by Prof. Jingwen Leng — *Nov 2021 - Apr 2022*

- Worked on **Sparsifier**, a MLsys combining algorithmic and hardware architecture co-optimization to exploit layer-wise N:M structured sparsity in the activation during post-training inference, post-fine-tuning inference, and training process.
  - Designed TopK(satisfy N:M structured sparsity pattern) and Embedded Index Engine of Sparsifier.
  - Designed a heuristic greedy algorithm to determine the sparsity ratio of each layer.
  - Less lower memory footprint and flops reduction for the pre-trained AI models without any fine-tuning.

Professional Experience

SenseTime Research **Shanghai, China**
Research Intern, Deap Learning Compiler *Mar 2023 - Jul 2023*

NVIDIA **Shanghai, China**
Software Engineering Intern, SW-GPU(Tesla) Architecture *Jun 2022 - Dec 2022*
Mentor: Leong He, Eko Liu
- CUDA Toolkit(r11.8, r12.0) and Recommended Driver Development and Testing.
- Verified specific APIs for DGX/HGX datacenter platforms (Redstone, Delta).
- Developed docker containers and scripts for stress test. Merged into NVIDIA Git main stream.

Alibaba Group **Hangzhou, China**
Software Engineering Intern, Cloud Security *Jul 2021 - Nov 2021*
- Developed encryption module in voice security amd voiceprint classification, audio classification.
- Developed an OCR for target monitor with **Alibaba Damo Academy**.

Services

Computer Architecture Student Association (CASA) **ACM SIGARCH**
Reviewer, Senior Member Students@System, IEEE Student Member *2022 – now*

High-performance Computing(HPC) Team **Shanghai Jiao Tong University**
Senior Member, ACM Parallel Computing Competion'22, ASC *2021 – 2023*

Technical Skill Set

Programming: C++, CUDA, Verilog, Python, Rust, SPICE, x86/RISC-V/MIPS
EDA/Simulation: Xilinx Vivado, Cadence Virtuoso, ChampSim, Zsim, GPGPU-sim
Software: Docker, Pytorch
Language: English, Mandarin, French