Zhuohao Li

📱 (+1) 310-425-4122 • ✉ zhuohaol@ucla.edu
🌐 www.linkedin.com/in/zhuohaoli/

Education

University of California, Los Angeles **Los Angeles, CA, United States**
Ph.D., Electrical & Computer Engineering *Start from Sep 2023*
Research Interests: ML & LLM Systems, Distributed Training and LLM Inference, SW-HW Codeisgn
Advisor: Prof. Yuan Tian , Prof. Tony Nowatzki

Shanghai Jiao Tong University **Shanghai, China**
BE., Microelectronics (with Honor); BE., Computer Science, GPA: 3.99/4.3 *Sep 2019 – Jun 2023*
Advisor: Prof. Jingwen Leng, Prof. Haibo Chen

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

- **Samueli School of Engineering Fellowship** **UCLA**(*2023-2024*)
- **Department of Computer Science Engineering Fellowship** **HKUST**(*2023*)
- **Shanghai Outstanding Graduate** **Shanghai Jiao Tong University**(*2023*)
- **SenseTime AI Scholarship**(**30** nationally in China) **SenseTime Inc.**(*2022*)
- **Irving T. Ho PhD Scholarship**(**4** in ˜**15k** SJTU) **Irving T. Ho PhD Foundation**(*2022*)
- **EECS Overseas Research Scholarship** **Shanghai Jiao Tong University**(*2022*)
- **SJTU Academic Excellence Scholarship** **Shanghai Jiao Tong University**(*2020,2021,2022*)
- **Excellence Medal in Google Cup** **Google LLC.**(*2021*)
- **First Prize in National Mathematics Olympic Competition** **Chinese Mathematics Society**(*2019*)

Publications

[2]. *"FaaSwap: SLO-Aware, GPU-Efficient Serverless Inference via Model Swapping"*, 2024 European Conference on Computer Systems (**EuroSys 2024**). Minchen Yu, Ao Wang, Dong Chen, Haoxuan Yu, Xiaonan Luo, **Zhuohao Li**, Wei Wang, Ruichuan Chen, Dapeng Nie, Haoran Yang.

[1]. *"RealNet: Combining Optimized Object Detection with Information Fusion Depth Estimation on IoT Devices"*, arXiv preprint. 2022. **Zhuohao Li**, Fandi Gou, Qixin De, etc.

Research Experience

UCLA Samueli School of Engineering **UCLA/Duke/NUS/Microsoft**
Graduate Researcher, advised by Prof. Danyang Zhuo *from Sep 2023*

- Worked on **X-training**, a distributed training framework for MoE with efficient fault-tolerance paradigms.
  - Exploited an emerging model resharding scheme to manage models efficiently when job failure occurs.
  - Resharding rule guarantees each machine stays the same status as before while as less as state movement.
  - Implemented a standalone atop Gemini and DeepSpeed, exploiting AWS EC2 spot-instances to serve.

Big Data Institute **HKUST/Bell Labs/Alibaba Cloud**
Research Intern, advised by Prof. Wei Wang and Ruichuan Chen *Sep 2022 - Jun 2023*

- Worked on **Xpor**, an efficiently disaggregated GPU system for ML inference on serverless cloud.
  - Proposed *memory swapping* to swap model memory of GPU instances between host memory and GPUs.
  - Developed an algorithm to perform job scheduling, eviction, and cluster worker node management.
  - Developed a standalone from scratch to enable asynchronized CUDA API remoting.
  - Designed asynchronous, model pipelining features to optimize data transmission performance.

Speedway Group **The University of Texas at Austin**
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

Shanghai AI Laboratory **Shanghai, China**
Research Intern, Deap Learning Compiler. Mentor: Xiuhong Li, Yun Liang *Apr 2023 - Aug 2023*

- Invloved in Sensetime DeepLink (OpenComputeLab) research.

NVIDIA **Remote due to Covid**
Software Engineering Intern, CUDA SW-GPU(Tesla Architecture), Mentor: Leong He *Jun 2022 - Dec 2022*

- CUDA Toolkit(r11.8, r12.0) and Recommended Driver Development and Testing on DGX .
- Verified specific kernels for DGX/HGX datacenter platforms (Redstone, Delta).
- Developed docker containers for isolated test and shell scripts for automatic test.
- Code is contributed into NVIDIA Git repository .

Alibaba Group **Hangzhou, China**
Software Engineering Intern, Security Group *Jul 2021 - Nov 2021*

- Developed encryption module in voice security amd voiceprint classification, audio classification.

Services

External reviewers **(USENIX Security'24)***(2024)*

High-performance Computing(HPC) Team **Shanghai Jiao Tong University***(2021 – 2023)*

Technical Skill Set

Programming: CUDA, C/C++, Python, Verilog, x86/RISC-V/MIPS
Software: Docker, Pytorch, DP/PP/TP/MP distributed training, Deepspeed, Megatron-LM, vLLM
EDA/Simulation: Xilinx Vivado, Cadence Virtuoso, ChampSim, Zsim, GPGPU-sim
Language: English, Mandarin