Yipeng(Leo) Shen

github | mysite | yipengshenn@gmail.com | +86 18310211513

Education

Zhejiang University *BS in Computer Science* Sept 2022 - June 2026

- **GPA:** 3.98/4.3 or 88.5/100
- **Relevant Course:** High Performance Computing, Computer Organization, Computer Architecture, Operating System, Computer Networks, Digital Logic Design, Machine Learning and Data Analysis, Introduction to Computing Systems, Natural Language Processing, Compiler Principles.
- **Research Interest**: **Machine Learning System**, **Serving Engine**, **Agent System**.

Publications

[1] Anonymous(Second Author), "Scaling multi-agent simulation," in *Submitted to The 9th Annual Conference on Machine Learning and Systems*, under review, 2025.

[2] Anonymous(Second Author), "Unlocking long-term dependencies in spiking neural networks with a recurrent LIF memory module," in *Submitted to The Fourteenth International Conference on Learning Representations*, under review, 2025.

[3] Pan Zaifeng, PATEL AJJKUMAR, Shen Yipeng, Hu Zhengding, Guan Yue, Li Wan-Lu, Qin Lianhui, Wang Yida, and Ding Yufei, "KVFlow: Efficient prefix caching for accelerating LLM-based multi-agent workflows," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Research Experience

UCSD Picasso Lab *Advised by Prof. Yufei Ding* Mar 2025 - present

- Optimize the KV cache and LoRA usage for mulit-agent workflow under SGLang Framework. Mainly take response of modifying the sglang backend engine, write new cache structure, cache writing queue, interrupt module and lora scheduler.

Zhejiang University CCNT Lab *Advised by Prof. Peng Lin* June 2024 - Nov 2025

- Designed and build a high-accuracy, low-power neural network that integrate both LIF's spike feature and LSTM's gate circuit base on the specific features of lab's neuromorphic hardware.

Zhejiang University RC4ML Lab *Advised by Prof. Zeke Wang* Nov 2024 - Mar 2025

- Accelerated convolution using large number multiplication algorithms on DSP48.
- Evaluate the performance of AMD VERSAL Network on Chips on Vivado.

Projects

NUS Cloud Computing Summer Workshop Github Page

- Led the 'Cloud on Cloud' project, building a real-time weather forecasting platform and information exchanging community. Our team won the project's second prize.
- Utilized Kafka for stream data processing, Spark for weather prediction, Kubernetes for managing cloud resources on AWS. Responsible for cloud system design and module integration.

MiniOS Github Page

– Implemented a fully functional Mini-OS kernel run on RISC-V 64. Key responsibilities involved setting up exception, handling trap, developing thread scheduler, building SV39 virtual memory scheme, loading ELF program, kernel-to-user mode switching, and file system integration (VFS and FAT32).

SysY Compiler ZJU Git

– Implemented a full-stack compiler for the SysY language, covering all phases from lexical and syntax analysis (generating an AST) to semantic analysis and code generation. Successfully translated SysY source code into optimized RISC-V 32 assembly.

Skills

Tools: Git, Docker, Kubernetes, Vivado, Matlab
Language: Chinese(Native); English(TOEFL=R30 L30 S21 W24)
Programming: C/C++, Python, Pytorch, Verilog, Risc-V, Cuda, OpenMP, MPI

Honors

Scholarship 2023, 2024, 2025

– Zhejiang University Third Prize Scholarship.

Social Work 2024, 2025

– President of Zhejiang University Student Guitar Association.

Volunteer Work 2023

– Head volunteer of MPC of the 19th Hangzhou Asian Games and Asian Para Games.

– Over 300 volunteer hours during undergraduate years.