Shengyu Liu

Email: shengyu.liu@stu.pku.edu.cn | Portfolio: interestinglsy.github.io

Research Interests

I am broadly interested in Computer Systems and Networking. My research currently focuses on Machine Learning Systems (MLSys), Machine Learning Compilers, and Distributed Systems.

Education

Peking University, BS in Computer Science — Sep 2021 – Current

- GPA: 3.872/4.0, ranking 8 out of 134 students.
- Serve as the monitor of the Turing Class.

Experience

Computer Systems Research Group, Peking University — Sep 2021 – Present

- Advisor: Associate Professor Xin Jin.
- Research area: large language model systems (LLMSys).
- Contributed to the LoongServe (in SOSP, 2nd author), DistServe (in OSDI, 2nd author), and FastServe (4th author) project, aiming at speeding up LLM inference. Responsibilities included generating innovative ideas, implementing the complete inference system, and conducting comprehensive experiments.

Catalyst Group, Carnegie Mellon University, Visiting Scholar — Jun 2024 – Sep 2024

- Advisor: Assistant Professor Zhihao Jia.
- Research area: machine learning compiler.
- Contributed to the Mirage project (the first multi-level superoptimizer for tensor programs).
- Independently conceived, designed, and implemented Mirage's CUDA transpiler.

Team Leader, Peking University Supercomputing Team — Mar 2022 – Dec 2023

- Directed the team to the the First Place at the 10th ASC and Second Place at SC23, two of the world's leading supercomputing competitions.
- Earned the "Community Impact Award" from the Student Cluster Competition (SCC) committee in SC23 for significant contributions to renowned open-source projects such as MLPerf.

Publications

LoongServe: Efficiently Serving Long-context Large Language Models with Elastic Sequence Parallelism — Apr 2024

Bingyang Wu, ***Shengyu Liu***, Yinmin Zhong, Peng Sun, Xuanzhe Liu, Xin Jin
In SOSP'24, arxiv.org/abs/2404.09526

DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving — Dec 2023

Yinmin Zhong, ***Shengyu Liu***, Junda Chen, Yibo Zhu, Xuanzhe Liu, Xin Jin, Hao Zhang
In OSDI'24, www.usenix.org/conference/osdi24/presentation/zhong-yinmin

RLHFuse: Efficient RLHF Training for Large Language Models with Inter- and Intra-Stage Fusion — Sep 2024

Yinmin Zhong*, Zili Zhang*, Bingyang Wu*, ***Shengyu Liu***, Yukun Chen, Changyi Wan, Hanpeng Hu, Lei Xia, Ranchen Ming, Yibo Zhu, Xin Jin
Submitted to NSDI, arxiv.org/abs/2409.13221

Fast Distributed Inference Serving for Large Language Models — Sep 2023

Bingyang Wu*, Yinmin Zhong*, Zili Zhang*, ***Shengyu Liu***, Fangyue Liu, Yuanhang Sun, Xuanzhe Liu, Xin Jin

Submitted to NSDI, arxiv.org/abs/2305.05920

Awards

Second Place and Nov 2023
Highest Linpack (HPL) Award, SC23 Student Cluster Competition
As the team leader. SC (SuperComputing) is a world-famous conference for high-performance computing.

Community Impact Award, SC23 Student Cluster Competition Nov 2023
For contribution to the open-source community, including contribution to the MLPerf project and the Zaychik Server project.

Champion at the 10th ASC Student Supercomputer Challenge Apr 2023
As the team leader. ASC is the largest student supercomputer competition in the world.

National Scholarship (2024) Oct 2024
The highest honor for undergraduates in China. Top 1% in Peking University.

National Scholarship (2023) Oct 2023
The highest honor for undergraduates in China. Top 1% in Peking University.

CCF Elite Collegiate Award Sep 2024
Only 2 award winners each year at Peking University.

Merit Student of Beijing Dec 2023

Pacemaker Award for Merit Student Oct 2023
Top 1% in Peking University

John Hopcroft Scholarship of Peking University Oct 2022

Academic Excellence Award Oct 2022

Projects

LoongServe github.com/LoongServe/LoongServe
Efficiently serving long-context large language models with elastic sequence parallelism.
Paper in SOSP.

SwiftLLM github.com/interestingLSY/swiftLLM
A tiny yet powerful LLM inference system tailored for researching purposes.
vLLM-equivalent performance with only 2k lines of code (2% of vLLM).

Tiny SYSY Compiler github.com/interestingLSY/sysy-compiler
A compiler for the SYSY language (a subset of C).
Got the first place in the performance benchmark among all students.

NeuroFrame github.com/interestingLSY/NeuroFrame
A DNN training framework written in C++/CUDA.
Train Resnet 150 with 95% of PyTorch's performance.

DistServe github.com/LLMServe/DistServe
Disaggregates prefill and decoding to optimize goodput under certain latency constraints (SLOs) for LLM serving.
Paper in OSDI.

SwiftTransformer github.com/LLMServe/SwiftTransformer
A tiny yet powerful and flexible implementation of the transformer neural network.
10000+ lines of C++/CUDA.

IntPool github.com/interestingLSY/IntPool
A cryptocurrency mining pool written in Nodejs.