Shengyu Liu

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

About Me

I am a senior student in Turing Class at the School of EECS, Peking University (PKU) (2021.9-Present).

My advisor is Prof. Xin Jin, and my research interests include Machine Learning Systems (MLSys), Machine Learning Compilers, and Distributed Systems.

I was also the team leader of the Peking University Supercomputing Team, and we won the First Place at the 10th ASC and the Second Place at SC23 (both are world top-3 Supercomputing competitions).

Experience

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.
- 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. Being responsible for designing part of the ideas, implementing the whole inference system, and conduct experiments.

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.

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
In submission, arxiv.org/abs/2409.13221

Iteration-Level Preemptive Scheduling for Large Language Model Inference Sep 2023

Bingyang Wu*, Yinmin Zhong*, Zili Zhang*, ***Shengyu Liu***, Fangyue Liu, Yuanhang Sun, Xuanzhe Liu, Xin Jin
In submission, arxiv.org/abs/2305.05920

Awards

National Scholarship Oct 2024
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.

SenseTime Scholarship Jul 2024
20 students per year across China. SenseTime is a famous AI software provider.

Merit Student of Beijing Dec 2023

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

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

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

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

John Hopcroft Scholarship of Peking University Oct 2022

Academic Excellence Award Oct 2022

Projects

SwiftLLM github.com/interestingLSY/swiftLLM
A tiny yet powerful LLM inference system tailored for researching purpose.
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).
My homework for the course "compiler principles".
Got the first place in the performance benchmark.

NeuroFrame github.com/interestingLSY/NeuroFrame
A DNN training framework written in C++/CUDA.
Can train Resnet 150 with 95% of PyTorch's performance.
My homework for the course "programming in AI".

DistServe github.com/LLMServe/DistServe
A novel large language model serving system that disaggregates prefill and decoding to optimize goodput under certain latency constraints (SLOs). Built on SwiftTransformer.
Paper accepted by OSDI.
5000+ lines of Python.

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

IntPool github.com/interestingLSY/IntPool
A mining pool written in Nodejs.
During the third year in my high school, I wrote a mining pool as a matter of interest.