

# Shengyu Liu

shengyu.liu@stu.pku.edu.cn | interestinglysy.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](https://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](https://www.usenix.org/conference/osdi24/presentation/zhong-yinmin)

**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](https://arxiv.org/abs/2305.05920)

## Awards

---

**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 The Highest Linpack (HPL) Award** and Nov 2023

**Community Impact Award** at the SC23 Student Cluster Competition  
As the team leader. SC (SuperComputing) is a world-famous conference for high-performance computing.

<b>National Scholarship</b> The highest honor for undergraduates in China. <u>Top 1%</u> in Peking University.	Oct 2023
<b>Pacemaker Award for Merit Student</b> <u>Top 1%</u> in Peking University	Oct 2023
<b>Champion at the 10th ASC Student Supercomputer Challenge</b> As the <u>team leader</u> . ASC is the largest student supercomputer competition in the world.	Apr 2023
<b>John Hopcroft Scholarship of Peking University</b>	Oct 2022
<b>Academic Excellence Award</b>	Oct 2022

## Projects

---

<b>SwiftLLM</b> A tiny yet powerful LLM inference system tailored for researching purpose. vLLM-equivalent performance with only 2k lines of code (2% of vLLM).	github.com/interestingLSY/swiftLLM
<b>Tiny SYSY Compiler</b> 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.	github.com/interestingLSY/sysy-compiler
<b>NeuroFrame</b> 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".	github.com/interestingLSY/NeuroFrame
<b>DistServe</b> 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.	github.com/LLMServe/DistServe
<b>SwiftTransformer</b> SwiftTransformer is a tiny yet powerful and flexible implementation of the transformer neural network. 10000+ lines of C++/CUDA.	github.com/LLMServe/SwiftTransformer
<b>IntPool</b> A mining pool written in Nodejs. During the third year in my high school, I wrote a mining pool as a matter of interest.	github.com/interestingLSY/IntPool