Shengyu Liu

Email address: interestingLSY@gmail.com | **Website:** https://interestinglsy.github.io/

ABOUT ME

I am a junior student in Turing Class at the School of Electronic Engineering and Computer Science (EECS), Peking University.
My advisor is Prof. Xin Jin, and my research interests include **Machine Learning Systems (MLSys)** and **Distributed Systems**.

EDUCATION AND TRAINING

01/09/2021 – CURRENT Beijing, China

BACHELOR Peking University (PKU)

Address No.5 Yiheyuan Road, Haidian District, 100871, Beijing, China | **Website** https://www.pku.edu.cn/

ADDITIONAL INFORMATION

PUBLICATIONS

My advisor is Prof. Xin Jin

Link https://xinjin.github.io/

DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving
– 2023
Submitted to OSDI'24.
I am the second author of the paper.

We propose DistServe, a novel large language model serving system that disaggregates prefill and decoding to optimize goodput under certain latency constraints (SLOs).

I am responsible for implementing our ideas and conducting experiments. In total, I wrote an additional 2.5k lines of Python code on top of our previous serving system, which was also written by me in 8k CUDA/C++ and 3.5k Python.

Link https://arxiv.org/abs/2401.09670

Iteration-Level Preemptive Scheduling for Large Language Model Inference – 2023
Submitted to NSDI'24 Spring and got a "revise-and-resubmit" status.
I am the second author of the paper.

We propose FastGen, a distributed inference serving system for LLMs that exploits the autoregressive pattern of LLM inference to enable preemption at token-level granularity.

I am responsible for implementing the entire serving system. In total, 11.5k lines of code were written by me, including 8k C++/CUDA and 3.5k Python.

HONOURS AND AWARDS

01/10/2023
National Scholarship – China Highest Honor for undergraduates in China. Top 1% in Peking University.

Merit Student of Beijing

01/10/2023
Pacemaker Award for Merit Student – Peking University Top 1% in Peking University

01/09/2022
John Hopcroft Scholarship of Peking University – Peking University

20/09/2022
Academic Excellence Award – Peking University

15/05/2023
First place at the 10th ASC Student Supercomputer Challenge – The ASC Committee I am the team leader of Peking University Supercomputing Team and lead us to the champion.

Link https://www.hpcwire.com/2023/06/14/asc23-who-won-why/

17/11/2023
Second place at the SC23 Student Cluster Competition – The SC Committee I am the team leader of Peking University Supercomputing Team.

17/11/2023
The Highest Linpack Benchmark Award at the SC23 Student Cluster Competition – The SC Committee The High Performance Linpack (HPL) benchmark is the standard way to measure the performance of a supercomputer, and we got the highest Linpack benchmark result at the SC23 Student Cluster Competition.

Link https://studentclustercompetition.us/2023/index.html

17/11/2023
Community Impact Award at the SC23 Student Cluster Competition – The SC Committee In recognition for significant open-source contributions towards the MLPerf Inference project

19/07/2019
Silver Medal in National Olympiad in Informatics (NOI) – China Computer Federation (CCF) I attended the National Olympiad in Informatics (NOI) in China during high school, ranked 144 out of ~300 and was awarded a silver medal.

PROJECTS

DistServe A novel large language model serving system that disaggregates prefill and decoding to optimize goodput under certain latency constraints (SLOs).
Built on SwiftTransformer with an additional 5000+ lines of Python.

SwiftTransformer SwiftTransformer is a tiny yet powerful and flexible implementation of the transformer neural network. It aims at providing a framework for researchers to try on their novel ideas. During my junior year, my senior colleague and I implemented two of our ideas on FastServe and submitted papers based on those ideas to OSDI and ATC.
Contains 10000+ lines of C++/CUDA, with the majority (>90%) of the code being authored by myself.

IntOJ - A Online Judge System Written in Python & Flask During the first year in my high school, I wrote a online judge system (like codeforces) out of interest.

Link https://github.com/intoj

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. It is a mining pool built on nodejs.

Link https://github.com/interestingLSY/IntPool

A GOOD GRADE

Doing Well in My Class

I do well in my class, with a GPA of 3.866. In the last years, my GPA ranking was 3 out of 135 students. I do really well in courses which have a strong connection with my major, like Operating Systems(Honor Track) (94 pts), Introduction to Parallel and Distributed Computing (98.5 pts), Algorithm Design and Analysis(Honor Track) (98 pts), Advanced Algebra(I) (93 pts), and Practice of Programming in C&C++ (99 pts).

ORGANISATIONAL SKILLS

Monitor of My Class I serve as the monitor of the Turing Class and have successfully organized numerous activities for the class, including board game salons and on-campus orienteering.

HOBBIES AND INTERESTS

High Performance Computing Besides computer systems, I am also passionate about High Performance Computing (HPC). As the team leader of the Peking University Supercomputing Team, I led us to the champion at the ASC Supercomputing Challenge and the second place at the SC23 Student Cluster Competition (both are world-famous supercomputing competitions).

Maintaining a Server in My Dormitory My forte lies in both hands-on and practical skills. During my freshman year of college, I collected computer hardware components from the second-hand market, assembled a desktop server, and deployed many engaging and beneficial applications onto it. This immersive experience significantly enhanced my understanding and familiarity with various aspects of technology, including the workings of computer systems and computer networking.

Having a Passion for Technologies Related to Computer Science I have a passion for technologies related to computer science all the time. As a youth, I eagerly engaged in programming single-chip microcomputers and tinkering with embedded hardware systems. Upon entering high school, I proactively self-taught Python and JavaScript, leveraging these skills to develop two projects: an online judging system for competitive programming, and a cryptocurrency mining pool.

COMMUNICATION AND INTERPERSONAL SKILLS

Teaching CS Courses for Tutoring Agencies I love teaching. During my summer and winter vacations, I often teach courses related to OI (Olympiad in Informatics) for a tutoring agency called Qingbei Xuetang. Students love my course, and their feedbacks are positive, like "Your courses are really easy to comprehend!" and "I'm really looking forward to the next time you tutor us!". And they give me high marks in the official feedback session.