Jinwei **Yao**

Computer Science · Machine Learning System

École Polytechnique Fédérale de Lausanne (EPFL), Route Cantonale, 1015 Lausanne, SWITZERLAND

jinwei.yao1114@gmail.com | monstertail.github.io | Monstertail | jinwei-yao-541164190

"Write the code, change the world."

Education

École Polytechnique Fédérale de Lausanne (EPFL)

Lausanne, Switzerland

Ph.D. student in Computer Science — *2022 - 2023*

- Got the **Ph.D. fellowship** from CS Dept of EPFL.
- Got **5.75/6** in System for Data Management and Data Science, an advanced course for PhD students
- **Programming skills**. Experienced: Python, Scala, Java; Familiar: Golang, C, MATLAB, VHDL

Zhejiang University

Hangzhou, China

B.Eng. in Electronic Science and Technology — *2018 - 2022*

- **Overall GPA:** 3.89/4.00 (87/100).
- **Honors Minor(40/6800+):**: Advanced Class of Engineering Education(ACEE), **Chu Kochen Honors College**.
- **Graduated with Outstanding Thesis Award**.

Research Interest

System Design for Machine Learning(especially System for LLMs), Distributed Systems

Experience

[System for LLM]

Accelerating LLM inference for reasoning tasks.

Westlake University, China

Research Assistant at LINS Lab, Westlake University — *Oct. 2023 - (ongoing)*

- **Mentor: Prof.Tao Lin(Westlake Univ.), Prof.Binhang Yuan(HKUST), Prof.Zeke Wang(Zhejiang Univ.), Prof.Jiaxuan You(UIUC)**
- **Motivation:** SOTA frameworks for reasoning like Tree-of-Thoughts or Graph-of-Thoughts are not efficient from a system perspective due to accumulative network delays.
- **Insight:** We can build the tree during the decoding to improve the efficiency of a single reasoning request in a new paradigm– Tree-based decoding with sequence granularity.
- **Goal:** improve memory-efficiency and calculation-efficiency for Tree-based decoding with sequence granularity.
- **Skills: Python/Pytorch, Triton, Shell**

[Programming Language & Distributed System]

Performance Optimization for Distributed Agent-based Simulation Engine

EPFL, Switzerland

Research Intern at DATA lab. — *Apr. 2023 - Jun.2023*

- **Mentor: Prof.Christoph Koch**
- **Motivation:** Large-scale agent-based simulations are popular for data science, while they have seen very little foundational research from core computer science.
- **Challenges:** Scaling agent-based simulations up and out is challenging for its complex data transformation, aggregation, and time series analysis.
- **Goal:** Optimize the performance of agent-based simulations based on previous work-CloudCity.
- **Related Knowledge :** distributed computing, programming language, databases, and computer architecture.
- **Skills: Scala,Java**

[Distributed System]

A Starvation-free Consensus for Decentralized Cross-shard Transaction Commit

EPFL, Switzerland

Research Intern at Decentralized and Distributed Systems Lab(DEDIS). — *Oct. 2022 - Feb. 2023*

- **Mentor: Prof.Bryan Ford**
- **Motivation:** Most commit protocols for cross-shard transactions in decentralized systems have starvation problems and are also expensive in communication.
- **Insight:** Starvation is mainly caused by the contentions for the exclusive locks. Transaction atomic commit and consensus have many similarities.
- **Challenge:** 1. How to resolve contention without introducing a global or even an intra-shard leader for ordering. 2. How to reduce communication costs.
- **Solution:** Fair contention resolution to enable starvation-free. Unifying consensus and atomic commit to reducing communication costs.
- **Skills: GoLang**

[System for ML]

Accelerating Graph Neural Network Sampling by Customizing FPGA Subsystem — *Zhejiang University, China*

Undergraduate Research Intern at RC4ML Lab, Zhejiang University — *Nov. 2021 - May.2022*

- **Mentor:** Prof.Zeke Wang
- **Motivation:** Graph sampling, as the first step for GNN training, is a time-consuming irregular computation on modern CPU.
- **Goal:** Customize an FPGA-based memory subsystem with High Bandwidth Memory (HBM) for graph sampling acceleration.
- **Hardware level:** High-bandwidth memory(HBM) can be applied to accelerate the sampling.
- **Software level:** 1)Pipelines and loop unrolling can increase parallelism. 2) The optimization of read and write strategies for graph information can squeeze the performance of HBM.
- **Win Outstanding Thesis Award for undergraduates(rank 1/200+).**
- **Skills: C++, Vitis, VHDL**

[FPGA Security]

Power-based Side-channel Disassembly Attacks on Shared FPGA — *EPFL, Switzerland*

Summer@EPFL Program Research Intern at Parallel Systems Architecture Laboratory, EPFL — *Jul. 2021 - Nov. 2021*

- **Mentor:** Dr. Mirjana Stojilovic
- **Supported by Summer@EPFL fellowship** (acceptance rate in 2021 was **1.5**%)
- **Motivation:**The voltage fluctuation generated by the softcore CPU deployed on the FPGA may leak side channel information when the instruction is running.
- **Goal:** Based on the voltage information, disassemble CPU instructions.
- **Insight:** The voltage fluctuations can be regarded as time-series signals.
- **Challenges:** 1) How to generate a robust dataset (noise between adjacent instructions needs to be considered)? 2)The critical paths of the same type of instructions (such as Add and Sub) are almost the same. What's worse, the sensor sampling rate on the FPGA is very low, and it is difficult for traditional ML methods to obtain enough information for classification.
- **Solution:** 1) Generating random but balanced training set by designing noise-tolerated templates for different instructions; 2) Using lightweight LSTM+CNN model to do the classification. LSTM is suitable for processing sequence data, and CNN can help LSTM learn more meaningful feature representations.
- **A paper has been accepted by Springer Journal in Hardware and Systems Security(HaSS'23).**
- **Skills: Python, Vivado**

Publications

1.[ICLR workshop AGI (Oral) 2024] DEFT: Flash Tree-Attention with IO-Awareness for Efficient Tree-search-based LLM Inference.
Yao, Jinwei*, Chen, Kaiqi*, Zhang, Kexun*, You, Jiaxuan, Yuan, Binhang, Wang, Zeke†, and Lin, Tao† preprint (arXiv:2404.00242), ICLR workshop AGI (Oral) 2024.
2.[HaSS'23] Instruction-Level Power Leakage Evaluation of Soft-Core CPUs on Shared FPGAs.
O. Glamocanin, S. Shrivastava, **J. Yao**, N. Ardo, M. Payer, and M. Stojilovic. Springer Journal in Hardware and Systems Security, special issue on Multitenant Computing Security Challenges and Solutions, 2023.

Honors & Awards

International

2022	**Ph.D. fellowship**, EPFL	*Switzerland*
2021	**Summer@EPFL 21 research fellowship(1.5% applications were awarded)**, EPFL	*Switzerland*

Domestic

2022	**Outstanding Thesis Award(rank 1/200+**, Graduation project for bachelor, Zhejiang University	*China*
2022	**Dean's Outstanding Honor**, Outstanding Graduate of Zhejiang University	*China*
2021	**Winner's Prize** , National College Student Information Security Contest	*China*
2021,2019	**Academic Fellowship and Academic Excellence Award**, Zhejiang University	*China*
2019	**Outstanding Student** , Zhejiang University	*China*

Extracurricular Skills

Language

English: TOEFL

- **103 – Listening 25/Reading 29/ Writing 24/ Speaking 25**
- **110(BEST) – Listening 30/Reading 29/ Writing 26/ Speaking 25**

Programming

Experienced: Python, Scala, Java; Familiar: Golang, C, MATLAB, VHDL

- **Framework: Tensorflow, PyTorch, Keras, Vivado, Vitis**