

WENQI JIANG

PhD Student

Department of Computer Science, ETH Zurich

Last Update: October 2023

STF G222

Stampfenbachstrasse 114

8092 Zurich, Switzerland

+41 076 585 8978

wenqi.jiang@inf.ethz.ch

<https://wenqijiang.github.io/>

RESEARCH INTERESTS

My research interests lie in the intersection of data management, computer architecture, and computer systems. Specifically, I am interested in designing efficient data systems in an era when Moore's Law no longer exists. This involves the development of cross-stack solutions that integrate algorithms, software systems, and underlying hardware to enhance overall system performance and efficiency. Recently, I have built Post-Moore data systems for large language models, vector retrieval via approximate nearest neighbor search, recommender systems, and spatial data processing.

EDUCATION

ETH Zurich, Switzerland

2021~2025 (*Expected*)

PhD in Computer Science

Affiliated with the Systems Group

Concentration in Data Processing on Heterogeneous Hardware

Advisor: Prof. Gustavo Alonso

Columbia University, USA

2018~2020

Master in Electrical Engineering

Concentration in Data-Driven Analysis and Computation

Advisor: Prof. Luca Carloni

Overall GPA: 4.0/4.0 (top 5%)

Huazhong University of Science and Technology, China

2014~2018

Bachelor in Automation

Concentration in Pattern Recognition

Overall GPA: 3.7/4.0

PROFESSIONAL APPOINTMENTS

Amazon Web Services

Oct. 2023 ~ Jan. 2024

Applied Scientist Intern

Santa Clara, USA

Alibaba Cloud

Sep. 2019 ~ Dec. 2019

Database Development Intern

Shenzhen, China

AWARDS AND HONORS

The Master's Award of Excellence (top 5%), Columbia University	2021
Outstanding Graduate, HUST	2018
Scholarship for Excellent Academic Performance, HUST	2015

UNDER SUBMISSION

- [1] **Wenqi Jiang**, Marco Zeller, Roger Waleffe, Torsten Hoefler, Gustavo Alonso, "Chameleon: a Heterogeneous and Disaggregated Accelerator System for Retrieval-Augmented Language Models." .
- [2] **Wenqi Jiang**, Martin Parvanov, Gustavo Alonso, "SwiftSpatial: Spatial Joins on Modern Hardware." .

CONFERENCE PAPERS

- [1] Shuai Zhang, **Wenqi Jiang**, "Data-Informed Geometric Space Selection." *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS'23)*.
- [2] **Wenqi Jiang**, Shigang Li, Yu Zhu, Johannes de Fine Licht, Zhenhao He, Runbin Shi, Cedric Renggli, Shuai Zhang, Theodoros Rekatsinas, Torsten Hoefler, and Gustavo Alonso, "Co-design Hardware and Algorithm for Vector Search." *The International Conference for High Performance Computing, Networking, Storage and Analysis (SC'23)*.
- [3] **Wenqi Jiang***, Zhenhao He*, Shuai Zhang, Kai Zeng, Liang Feng, Jiansong Zhang, Tongxuan Liu, Yong Li, Jingren Zhou, Ce Zhang, and Gustavo Alonso, "FleetRec: Large-Scale Recommendation Inference on Hybrid GPU-FPGA Clusters." *Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'21)*.
- [4] Yu Zhu, Zhenhao He, **Wenqi Jiang**, Kai Zeng, Jingren Zhou, and Gustavo Alonso, "Distributed Recommendation Inference on FPGA Clusters." *31th International Conference on Field-Programmable Logic and Applications (FPL'21)*.
- [5] **Wenqi Jiang**, Zhenhao He, Shuai Zhang, Thomas B. Preußer, Kai Zeng, Liang Feng, Jiansong Zhang, Tongxuan Liu, Yong Li, Jingren Zhou, Ce Zhang, and Gustavo Alonso, "MicroRec: Efficient Recommendation Inference by Hardware and Data Structure Solutions." *4th Conference on Machine Learning and Systems (MLSys'21)*.

JOURNAL PAPERS

- [1] Shaoxiong Ji, **Wenqi Jiang**, Anwar Walid, Xue Li, "Dynamic Sampling and Selective Masking for Communication-Efficient Federated Learning." *IEEE Intelligent Systems*, 2022 .

WORKSHOP PAPERS

- [1] **Wenqi Jiang**, Gustavo Alonso, "Chameleon: a Disaggregated CPU, GPU, and FPGA System for Retrieval-Augmented Language Models." *Ninth International Workshop on Heterogeneous High-performance Reconfigurable Computing (H2RC @ SC'23)*.

TUTORIALS

- [1] **Wenqi Jiang**, Dario Korolija, and Gustavo Alonso, “Data Processing with FPGAs on Modern Architectures.” *International Conference on Management of Data (SIGMOD’23)*.

TALKS

Modern Search Engines on Specialized Hardware

ETH Zurich Sept. 2022

Efficient Recommendation Inference on Heterogeneous Hardware

AMD March 2022

ETH Zurich March 2021

FleetRec: Large-Scale Recommendation Inference on Hybrid GPU-FPGA Clusters

SIGKDD Conference Aug. 2021

MicroRec: Efficient Recommendation Inference by Hardware and Data Structure Solutions

MLSys Conference April 2021

ETH Zurich June 2020

PROFESSIONAL SERVICE

Reviewer

IEEE Micro 2021

LANGUAGES

English: fluent

Mandarin (Chinese): native

SKILLS

Programming Languages: C/C++, Python, OCaml, System Verilog

Platforms & Frameworks: Vivado HLS, CUDA, OpenCL, OpenCV, TensorFlow, PyTorch, Keras, Spark Streaming, Apache Beam, MySQL, PostgreSQL

REFERENCES

Available upon request