WENQI JIANG

Assistant Professor
National University of Singapore
Last Update: July 2026

AS6 Room 05-12
11 Computing Dr
Singapore 117416

jiangwenqi233@gmail.com
https://wenqijiang.github.io/

EDUCATION

ETH Zurich 2021~2025
PhD in Computer Science
Advisors: Gustavo Alonso and Torsten Hoefler
Committee: Gustavo Alonso, Torsten Hoefler, Ana Klimovic, and Christos Kozyrakis
🏆 ACM SIGMOD Jim Gray Doctoral Dissertation Award

Columbia University 2018~2020
Master in Electrical Engineering
The Master's Award of Excellence (top 5%)

Huazhong University of Science and Technology 2014~2018
Bachelor in Automation
Outstanding Graduate

PROFESSIONAL APPOINTMENTS

National University of Singapore July 2026 ~ Now
Assistant Professor Singapore

NVIDIA Research Jan. 2026 ~ July 2026
Postdoctoral Researcher Santa Clara, USA
Mentor: Christos Kozyrakis

Google Cloud July 2024 ~ Dec. 2024
Student Researcher Sunnyvale, USA
Mentors: Vidushi Dadu, Suvinay Subramanian, and Amir Yazdanbakhsh

Amazon Web Services Oct. 2023 ~ Jan. 2024
Applied Scientist Intern Santa Clara, USA
Mentors: Shuai Zhang and Boran Han

Alibaba Cloud Sep. 2019 ~ Dec. 2019
Software Engineering Intern Shenzhen, China

AWARDS AND HONORS

ACM SIGMOD Jim Gray Doctoral Doctoral Dissertation Award 2026

VLDB'25 Best Paper Award (Scalable Data Science Track) 2025

ML and Systems Rising Stars Award 2024

AMD HACC Outstanding Researcher Award	2023
The Master's Award of Excellence (top 5%), Columbia University	2021
Outstanding Graduate, HUST	2018
Scholarship for Excellent Academic Performance, HUST	2015

Publications

[1] **Wenqi Jiang**, Jason Clemons, Rowland O'Flaherty, Hugo Hadfield, Alperen Degirmenci, Shuran Song, Yashraj Narang, and Christos Kozyrakis, "ROSA: A Robotics Foundation Model Serving System for Robot Factories." *arXiv preprint (arXiv'26)*

[2] **Wenqi Jiang**, Jason Clemons, Karu Sankaralingam, and Christos Kozyrakis, "How Fast Can I Run My VLA? Demystifying VLA Inference Performance with VLA-Perf." *arXiv preprint (arXiv'26)*

[3] Qijing Huang, **Wenqi Jiang**, Christos Kozyrakis, and Jason Clemons, "Enabling the Robotic Revolution: Bridging Performance Gap between Present and Future." *IEEE/JSAP Symposium on VLSI Technology and Circuits (VLSI'26)*

[4] You Peng, Youhe Jiang, **Wenqi Jiang**, Chen Wang, and Binhang Yuan, "HEXGEN-FLOW: Optimizing LLM Inference Request Scheduling for Agentic Text-to-SQL." *IEEE International Conference on Data Engineering (ICDE'26)*

[5] **Wenqi Jiang**, Suvinay Subramanian, Cat Graves, Gustavo Alonso, Amir Yazdanbakhsh, and Vidushi Dadu, "RAGO: Systematic Performance Optimization for Retrieval-Augmented Generation Serving." *Proceedings of 52nd Annual International Symposium on Computer Architecture (**ISCA'25**)*

[6] **Wenqi Jiang**, Marco Zeller, Roger Waleffe, Torsten Hoefler, and Gustavo Alonso, "Chameleon: a Heterogeneous and Disaggregated Accelerator System for Retrieval-Augmented Language Models." *Proceedings of the VLDB Endowment (**VLDB'25**)*
🏆 **Best Scalable Data Science Paper Award**

[7] **Wenqi Jiang**, Hang Hu, Torsten Hoefler, and Gustavo Alonso, "Fast Graph Vector Search via Hardware Acceleration and Delayed-Synchronization Traversal." *Proceedings of the VLDB Endowment (**VLDB'25**)*

[8] **Wenqi Jiang**, Oleh-Yevhen Khavrona, Martin Parvanov, and Gustavo Alonso, "SwiftSpatial: Spatial Joins on Modern Hardware." *International Conference on Management of Data (**SIGMOD'25**)*

[9] **Wenqi Jiang**, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska, "PipeRAG: Fast Retrieval-Augmented Generation via Adaptive Pipeline Parallelism." *Proceedings of the 31th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD'25**)*

[10] Qi Chen, Xiubo Geng, Corby Rosset, Carolyn Buractaon, Jingwen Lu, Tao Shen, Kun Zhou, Chenyan Xiong, Yeyun Gong, Paul Bennett, Nick Craswell, Xing Xie, Fan Yang, Bryan Tower, Nikhil Rao, Anlei Dong, **Wenqi Jiang**, Zheng Liu, Mingqin Li, Chuanjie Liu, Zengzhong Li, Rangan Majumder, Jennifer Neville, Andy Oakley, Knut Magne Risvik, Harsha Vardhan Simhadri, Manik Varma, Yujing Wang, Linjun Yang, Mao Yang, and Ce Zhang, "MS MARCO Web Search: A Large-scale Information-rich Web Dataset with Millions of Real Click Labels." *International World Wide Web Conference (**WWW'24**)*

[11] Shuai Zhang and **Wenqi Jiang**, "Data-Informed Geometric Space Selection." *Thirty-seventh Conference on Neural Information Processing Systems (**NeurIPS'23**)*

[12] **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**)*

[13] **Wenqi Jiang**, Dario Korolija, and Gustavo Alonso, "Data Processing with FPGAs on Modern Architectures." *International Conference on Management of Data (SIGMOD'23 Tutorial)*

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

[15] **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**)*

[16] 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)*

[17] **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**)*

Teaching

Guest Lecturer:

Cloud Computing Architecture	Spring 2025
Data Modelling and Databases	Spring 2024
Hardware Acceleration for Data Processing	Fall 2021, 2023

Teaching Assistant:

Big Data	Fall 2023
Big Data for Engineers	Spring 2022, 2023
Systems Programming and Computer Architecture	Fall 2021, 2022

Professional Service

Program Committee: VLDB'27, EuroSys'27, ATC'26, HCDS@ASPLOS'26

Journal Reviewer: CACM'26, VLDBJ'26, TOCS '25