Kaiqiang Xu

andyxukq@gmail.com • https://kqxu.com/

BIOGRAPHY

My research focuses on improving the efficiency and accessibility of machine learning systems and AI computing infrastructure. My work designs new abstractions, parallelization strategies, and scheduling algorithms that exploit the unique characteristics of machine learning workloads.

I believe a systemic understanding of computing infrastructure is key to effective optimization. I have publications in top conferences on computer networks, architecture, operating systems and databases.

EDUCATION

Ph.D in Computer Science The Hong Kong University of Science and Technology (2020 - Aug 2025)

Thesis Title: *Towards Efficient and Accessible Systems for Distributed Machine Learning*

B.S. in Computer Science Renmin University of China (2012-2017)

Google EXPERIENCE

(Go link to my Google CV: go/xkq)

Student Researcher **Seattle, WA, USA**

Google Cloud - ML, Systems and Cloud AI (MSCA) March 2025 to May 2025

- Worked on TPU compiler and JAX/Pallas kernel optimization, submitted 25+ CLs during 12-week term
- TPU SparseCore Compiler: XLA assembly bundle analysis and visualization → go/xla-sc-visualization
- JAX and Pallas on TPU: transformer model kernel performance profiling and tuning → go/pallas-perf
- Recognized with a peer bonus (certificate) from the XLA compiler team

ACADEMIC PUBLICATIONS

ACM ASPLOS 2025: Design and Operation of Shared Machine Learning Clusters on Campus

First Author. Accepted to and Presented at ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '25)

The rise of large AI models drives investment in GPU clusters, but universities lack the expertise to fully utilize these resources, leading to barriers to democratized ML access. This work simplifies shared GPU cluster management with a 4-layer architecture, covering job scheduling, user isolation, and resource allocation, designed for low-maintenance operation while improving resource fairness and utilization.

USENIX NSDI 2025: GREEN: Carbon-efficient Resource Scheduling for Machine Learning Clusters

First author. Accepted to and presented at USENIX Symposium on Networked Systems Design and Implementation (NSDI '25)

Machine learning workloads currently account for 8% of global data center demand. Traditional cluster resource schedulers optimize job completion time but do not consider environmental impacts. This work proposes a cluster resource scheduling algorithm that, through energy-aware task scaling and temporal workload shifting, reduces the carbon footprint by up to 41.2% while maintaining overall time efficiency.

ACM SIGMOD 2025: Sequoia: An Accessible and Extensible Framework for Privacy-Preserving Machine Learning over Distributed Data

First author. Accepted to and presented at ACM International Conference on Management of Data (SIGMOD '25)

Privacy-preserving machine learning (PPML) enables secure collaborative model training across distributed datasets, addressing privacy concerns. This work decouples secure computation from ML models, and enhances PPML accessibility and efficiency through a compiler-executor architecture and decentralized scheduling, reducing code complexity by up to 92% and speed up execution by up to 252%.

ACM SIGMOD 2023: Scalable and Efficient Full-Graph GNN Training for Large Graphs

Co-first author. Accepted to and presented at ACM International Conference on Management of Data (SIGMOD '23)

Graph Neural Networks (GNNs) struggle to scale efficiently on billion-edge graphs due to computation dependencies across partitions. This work enables scalable full-graph GNN training using hybrid parallelism and pipeline scheduling, achieving up to 2.24× speedup and 6% higher accuracy.

WORK EXPERIENCE

Vincross Robotics **Beijing, China**

Chief Operating Officer May 2016 to Dec 2019

- Led a global team to build HEXA—the first programmable all-terrain robot designed for individuals to create their own robotic applications. HEXA, which empowers users to explore, build, and shape the future of robotics, was featured in Wall Street Journal, WIRED, The Verge, and IEEE.
- Led the design and development of MIND OS, a Linux-based robotics OS with an SDK for 3rd party application development, managing processes, networking, and peripherals. It also enables robot communication with mobile and IoT devices via its cloud service.

Zhihu.com (NYSE: ZH) **Beijing, China**

Algorithm Engineer Intern Oct 2013 to Sept 2014

- Built real-time full-text search engine and recommendation system, utilizing NLP knowledges including user query analysis, query segmentation, and query expansion.
- Designed and developed an efficient parallel algorithm for large-scale content analysis.

AWARDS

ACM-ICPC, International Collegiate Programming Contest (Asia Regional), 2012 Silver Medal

National Olympiad in Informatics, China, 2011 Bronze Medal

Forbes 30 Under 30 List 2020 (China), 2022 (Asia)