

Yue Jin

Senior Software Engineer, Ant Group, Hangzhou, China
jinyue.derek@gmail.com — +86 15088682238 — <https://jinderek.github.io/>

RESEARCH INTERESTS

Machine Learning Systems, Parallel Computing, GPU Computing, Graph Learning, and Compiler Optimization.

INDUSTRY EXPERIENCE

Ant Group

Senior Software Engineer

Hangzhou, China
Oct. 2018 — Present

- Developed G-Sparse, a compiler-based GNN accelerator for GPU.
- Developed GraphGen, a high-performance distributed graph sampling engine.
- Developed Woodpecker, a high-performance compiler-based deep learning framework.

Alibaba Group

Software Engineer

Hangzhou, China
Apr. 2015 — Oct. 2018

- Developed JSNI, the first standardized native Interface for JavaScript and Native C/C++ code interactions, widely adopted by Alibaba Group and other industry companies.
- Enhanced the Multithreaded V8 JavaScript Virtual Machine Project by optimizing its garbage collection module.

EDUCATION

Zhejiang University, Hangzhou, China

Master of Engineering in Electrical Engineering

Sep. 2012 — Mar. 2015

Zhejiang University, Hangzhou, China

Bachelor of Engineering in Electronic & Information Engineering

Sep. 2008 — Jun. 2012

PUBLICATIONS

G-Sparse: compiler-driven acceleration for generalized sparse computation for graph neural networks on modern GPUs.

Y. Jin, C. Huan, H. Zhang, Y. Liu, S. L. Song, R. Zhao, Y. Zhang, C. He, W. Chen.
PACT 2023.

TEA+: a novel temporal graph random walk engine with hybrid storage architecture.

C. Huan, Y. Liu, H. Zhang, S. Song, S. Pandey, S. Chen, X. Fang, Y. Jin, B. Lepers, H. Liu, Y. Wu.
ACM TACO 2024.

GraphRPM: risk pattern mining on industrial large attributed graphs.

S. Tian, X. Zeng, Y. Hu, B. Wang, Y. Liu, Y. Jin, C. Meng, C. Hong, T. Zhang, W. Wang.
ECML PKDD 2024.

GraphGen: a distributed graph sample generation framework on industry-scale graphs.

Y. Jin, S. Tian, Y. Liu, C. Hong.
EuroSys 2024 (poster track).

GPC: compiler-based optimization for sparse computations in graph neural networks.

Y. Jin, Y. Liu.
EuroSys 2023 (poster track).

Woodpecker-DL: Accelerating Deep Neural Networks via Hardware-Aware Multifaceted Optimizations.

Y. Liu*, Y. Jin*, Y. Chen, T. Teng, H. Ou, R. Zhao, Y. Zhang.
arXiv preprint, arXiv:2008.04567, 2020.

SELECTED PROJECTS

Large Scale Graph Chain of Thought with LLMs - GraphCoT

Mar. 2024 — Present

Hangzhou, China

- Responsible for the development of the large-scale GraphCoT engine at Ant Group, which is integrated with Large Language Models (LLM).

Compiler-based GNN Accelerator on GPU – G-Sparse

Hangzhou, China

Mar. 2022 — Present

- Led the development of G-Sparse, a GPU-accelerated compiler framework for generalized sparse computations in GNNs, achieving a $2.4\times$ speedup on training and inference and a $1.3\times$ to $4.8\times$ speedup on key operators (g-SpMM and g-SDDMM) over DGL and NVIDIA cuSparse.
- Empowered real-time graph analytics in production systems, bridging cutting-edge compiler techniques with practical deployment.
- Published papers in PACT 2023 and contributed to open-source libraries and frameworks (Halide, TuGraph).
- Integrated into a Python package, enabling users to accelerate DGL and PYG training with a single line of code.

High-performance Distributed Graph Sampling Engine - GraphGen

Hangzhou, China

Mar. 2022 — Present

- Project lead of GraphGen, a high-performance distributed graph sampling engine, achieving 10 million nodes per second performance— $20\times$ faster than SQL-based solutions—and significantly improving sample generation for industry-scale graphs.

High-performance compiler-based Deep Learning Framework - Woodpecker

Hangzhou, China

Oct. 2018 — Mar. 2022

- Tech lead of the deep learning high-performance operator library.
- Developed a domain-specific language (DSL) compiler (based on Halide) and ML-based cost model, reducing auto-tuning time from minutes to seconds.
- Achieved a $1.5\times$ to $10\times$ speedup compared to NVIDIA standard libraries (cuDNN, cuBLAS, cuSPARSE) on core operations (Conv, Matmul, LayerNorm, ArgMax, ArgMin, SpMM, SDDMM, etc.).
- Achieved $1.2\times$ to $1.7\times$ speedup on DNN and GNN models such as ResNet-50, DeepFM, Transformer, GAT, GCN and GraphSage, based on compiler optimization and auto-fusion techniques.
- Presented work at GTC 2020 and the Stanford DAWN Retreat 2019, receiving recognition for contributions to heterogeneous computing optimizations.

TALKS

G-Sparse: Compiler-driven acceleration for generalized sparse computation for graph neural networks on modern GPUs.
PACT Conference 2023.

Model-based cost estimation and its application in deep learning operation optimizations.
GPU Technology Conference 2020 (GTC 2020), China.

Woodpecker-DL: An Efficient Compiler for Accelerating Deep Learning on Heterogeneous Computing Architectures.
GPU Technology Conference 2019 (GTC 2019), China.

Woodpecker Project Presentation.
Stanford Fall DAWN Retreats 2019, Quadrus, Menlo Park, CA, USA.

AWARDS

Most Innovative Spirit Award

Hangzhou, China

Excellent Engineer: Most Innovative Spirit Award, Ant Group.

2021

Recognized for contributions to AI infrastructure and high-performance computing.

ENGLISH

TOEFL (iBT): 102 (overall score)

Test date: Oct. 2024

SKILLS

- Programming Languages: C/C++, CUDA, Python, JavaScript
- Frameworks/Tools: Halide, PyTorch, TensorFlow, DGL, Triton, TVM, MLIR
- Systems: Node.js/V8, Linux/GNU, ARM/x86

REFERENCES

Dr. Yongchao Liu

Staff Engineer at Ant Group, Hangzhou, China, **E-mail:** yongchao.ly@antgroup.com

Dr. Yao Zhang

Principal Architect at Fireworks AI, California, United States, **E-mail:** yaozhang@fireworks.ai

Prof. Heng Zhang

Associate Professor at Institute of Software, Chinese Academy of Sciences, **E-mail:** zhangheng17@iscas.ac.cn

Prof. Dekuang Su

Professor at the School of Mathematical Sciences, Zhejiang University, Hangzhou, China **E-mail:** sdk001@zju.edu.cn