Jie Xin

[email] [github] [homepage]

Experience

NVIDIA Shanghai, China
Senior Compute Architect 2022 – Present

- Researching automated end-to-end model performance optimization on next-gen GPUs (Hopper, Blackwell, Rubin)
- Developed deep learning compilers (CuTile, Triton-to-CuTile) and MLPerf benchmark optimization
- Collaborating with PyG, Megatron-LM, and OpenAI Triton teams on compiler infrastructure

Education

Huazhong University of Science and Technology Wuhan, China
M.S. in Computer Science 2019 – 2022

Huazhong University of Science and Technology Wuhan, China
B.S. in Computer Science 2015 – 2019

Projects

Agent for Compiler 2026 – Present
Designing a lightweight harness system specifically for compiler optimization.

Triton-to-CuTile 2024 – 2025
OpenAI incubation project; compiler bridge connecting Triton frontend to CuTile backend. [code]
Enabling Triton users to target NVIDIA's optimized tile programming infrastructure.

CuTile 2023 – Present
CUDA Tile programming standard for NVIDIA GPUs. [code] [blog]
Designed and implemented optimization passes for deep learning workloads.

NvFuser 2022 – 2023
PyTorch JIT compiler achieving 1.2x–2.5x speedup on GNN models. [code] [blog]
Enabled gather/scatter/index_select graph operations for sparse workloads.

OpenFold2 Training 2023 – 2023
MLPerf HPC v3.1 #1 submission with 7.5x e2e speedup. [code] [paper] [blog]
Optimized protein structure prediction training on large-scale GPU clusters.

GPT-3 Training 2023 – 2023
MLPerf Training v4.0; contributed 1.02x training speedup on 175B model. [code]
Optimized distributed training pipeline for large language models.

PyTorch Geometric TorchScript 2022 – 2022
Added TorchScript support for PyG community, enabling JIT compilation for GNN models. [code]
Contributed core infrastructure for production deployment of graph neural networks.

SpMM on GPU – Champion, MIT/IEEE Graph Challenge 2021 – 2021
High-performance sparse matrix multiplication; up to 652x speedup over cuSPARSE. [code] [paper]
Won 1st place in the 2021 MIT/IEEE/Amazon Graph Challenge competition.

Current Focus

My research interest has fully shifted to AI agents – exploring how LLMs can autonomously optimize compilers, automate performance tuning, and transform traditional software development workflows.