LIU ZIHAN 刘子汉, +86-159-0215-7531

ilovehanhan1120@hotmail.com, https://subjectnoi.github.io/about

Education

2015.09~2019.07	Bachelor	East China Normal University	
2019.09~2022.03	Master	Shanghai Jiao Tong University	Advisor: Prof. Jingwen Leng
2022.04~2025.06	Ph.D.	Shanghai Jiao Tong University	Lab: EPCC

Publications (https://orcid.org/0000-0002-0874-0682)

HPCA 2025	VQ-LLM: High-performance Code Generation for Vector Quantization Augmented LLM Inference (1st author)	LLM, Quantization, Code Generation
ASPLOS 2024	JUNO: Optimizing High-Dimensional Approximate Nearest Neighbour Search with Sparsity-Aware Algorithm and Ray-Tracing Core Mapping (1st author)	ANNS, Ray Tracing
ASPLOS 2022	VELTAIR: Towards High-Performance Multi-Tenant Deep Learning Services via Adaptive Compilation and Scheduling (1st author)	DNN Compiler, Multi-Tenant
ISPA 2020	DLFusion: An Auto-Tuning Compiler for Layer Fusion on Deep Neural Network Accelerator (1st author)	DNN Compiler, Auto Tuning, NPU
THPC 2020	Survey and Design of Paleozoic: a High-Performance Compiler Tool Chain for Deep Learning Inference Accelerator (1st author)	DNN Compiler, ONNX, IR
HPCA 2025	MANT: Efficient Low-bit Group Quantization for LLMs via Mathematically Adaptive Numerical Type	LLM, Quantization, Accelerator Design
TACO 2024	Potamoi: Accelerating Neural Rendering via a Unified Streaming Architecture	Neural Rendering, Accelerator Design
ISCA 2024	Cicero: Addressing Algorithmic and Architectural Bottlenecks in Neural Rendering by Radiance Warping and Memory Optimizations	Neural Rendering, Accelerator Design
ASPLOS 2024	GMLake: Efficient and Transparent GPU Memory Defragmentation for Large-scale DNN Training with Virtual Memory Stitching	LLM, Virtual Memory
CF 2023	AdaptGear: Accelerating GNN Training via Adaptive Subgraph-Level Kernels on GPUs	GNN, Code Generation
MICRO 2022	ANT: Exploiting Adaptive Numerical Data Type for Low-bit Deep Neural Network Quantization	Quantization, Accelerator Design

Jobs

2022.06~2022.12	AMD	DV Intern (GFX HW MI)	I'm responsible for part of the coverage report and verification/debugging of shader core components.
2020.06~2021.06	Intel	Compiler Dev. Intern (IAGS)	I'm responsible for test cases of AMX instructions, then I take part in the research of PGO with LLVM.
2019.02~2019.06	NVIDIA	GPU SM Arch Intern (Compute Arch)	I'm responsible for the implementation of a warp-group level matrix multiplication instruction of GA10b and Hopper in the C++ based simulator.
2018.08~2019.01	SAP	Java Intern (IBSO)	I'm responsible for part of the development of several cloud foundry applications in S/4 HANA.

Projects

2023 Now	NSFC Research Grant	Research on architecture and compiler design of dataflow architecture. I'm responsible for the dataflow compiler and programming model related research, I also participated part of the dataflow architecture DSE project.
2024 Now	R&D project From Industry	Research on high-performance code generation on new Hopper architecture GPUs for efficient LLM inference. This research deeply dives into the DSM/TMA, Triton, torch.compile(), CuTE, etc., and figures out optimal way to conduct aggressive compute kernel fusion.
2023 2024	R&D project From Industry	Research on LLM quantization, and I'm responsible for the KV-Cache compression via vector quantization techniques. For PPL we get a <0.05 worsen under equivalent 4-bit compression ratio. For inference performance, the proposed solution is accepted in HPCA'2025.
2021	R&D project From Industry	Research on compiler stack design of a heterogeneous high-end AI chip, integrated with R5CPU, vector core and AI (matrix) core. I provided a prototype via extending TVM, I added the support of communication related operator in both frontend translation and backend codegen.
2019 2020	NSFC Research Grant	Research on complier stack design and auto-tuning optimizations of Cambricon DNN accelerator (MLU-100). I implemented a frontend wrapper with vendor provided kernel library, with which I conduct a series of research on auto-fusion on this chip.
Before 2019	Course project(s)	RTL of RISC-V pipelined CPU. C-like language with lex and yacc. Profiling and optimizations of Tensor Core on Turing GPUs (B.S. Thesis).

Skills

C, C++, CUDA, PTX, Triton, CuTE, Computer Architecture, AI, LLM

Verilog/Verilator

Torch, TVM, LLVM, ONNXRuntime

Java, Spring, Python, Unreal Engine

Archery, Saxophone, Astrophotography, Badminton, Games (ACT, FPS, Flight Simulation)