Lian Liu

liulian211@mails.ucas.ac.cn · (+86) 159-8257-6242 · https://leliyliu.github.io/ ·

Education

Institute of Computing Technology (ICT), CAS, Beijing, China 2021 – Present

- *PhD student* in Computer Science (CS), expected June 2026
- Advisor: Prof. Ying Wang, Prof. Huawei Li

Nankai University (NKU), Tianjin, China 2017 – 2021

- *B.S.* in the Internet of Things (IoT)
- Advisor: Prof. Tao Li
- GPA: 3.88/4, Rank: 1/45

Research Interests

- LLM-centric Architecture Design
- Processing-in-Memory System
- Algorithm-System Co-design

Publications

COMET: Towards Practical W4A4KV4 LLMs Serving
Lian Liu, Long Cheng, Haimeng Ren, Mengdi Wang, Xiaowei Li, Yinhe Han, and Ying Wang
Accepted by **ASPLOS'25**, 2025. (CCF-A).

Make LLM Inference Affordable to Everyone: Augmenting GPU Memory with NDP-DIMM
Lian Liu*, Shixin Zhao*, Bing Li, Mengdi Wang, Xiaowei Li, Yinhe Han, and Ying Wang
Accepted by **HPCA'25**, 2025. (CCF-A).

DAQU: An Automatic Neural Network Architecture-and-Quantization Joint Optimization Framework for Efficient Model Inference
Lian Liu, Ying Wang, Weiwei Chen, Xiandong Zhao, Huawei Li*, Xiaowei Li, and Yinhe Han
Accepted by **TCAD'24**, 2024. (CCF-A).

Drift: Leveraging Distribution-based Dynamic Precision Quantization for Efficient Deep Neural Network Acceleration
Lian Liu, Zhaohui Xu, Yintao He, Ying Wang, Huawei Li, Xiaowei Li, and Yinhe Han
Accepted by **DAC'24**, 2024. (CCF-A).

PAM: Processing Across Memory Hierarchy for Efficient KV-centric LLM Serving System
Shixin Zhao*, **Lian Liu***, Yutian Zhou, Yintao He, Mengdi Wang, Yinhe Han, and Ying Wang
Submitted to **ISCA'25**, 2025. (CCF-A, In submission).

BaWA: Automatic Optimizing Pruning Metric for LLMs with Balanced Weight and Activation
Lian Liu, Xiandong Zhao, Guanchen Li, Dong Li, Mengdi Wang, Yinhe Han, and Ying Wang
Submitted to **ICML'25**, 2025. (CCF-A, In submission).

DNA: A General Dynamic Neural Network Accelerator
Lian Liu, Jinxin Yu, Mengdi Wang, Xiaowei Li, Yinhe Han, and Ying Wang
Submitted to **TC'25**, 2025. (CCF-A, In submission).

Enhanced One-Shot Pruned Pre-trained Language Models through Sparse-Dense-Sparse Mechanism
Guanchen Li, Xiandong Zhao, **Lian Liu**, Zeping Li, Dong Li, Lu Tian, Jie He, Ashish Sirasao, Emad Barsoum
Accepted by **COLING'25**, 2025. (CCF-B).

On-Line Fault Protection for ReRAM-Based Neural Networks
Wen Li, Ying Wang, Cheng Liu, Yintao He, **Lian Liu**, and Huawei Li
Accepted by **TC'23**, 2023. (CCF-A).

* Co-first author

Research Experience

KV-centric LLM serving with tiered PIM Design, *Co-first Author* 2024.09 – 2025.03

Addressing the memory-bound KV-related operations with PIM extension and system optimization.

- A **Novel Processing Manner** with tiered PIM design and Context Locality in LLMs.
- **Full-stack solution** including algorithmic innovations, scheduling optimizations, and PIM-extension.
- 7.20× Speedup & 7.17× Cost Efficiency (tokens/$) than vLLM

Accelerating LLM Inference with NDP-DIMMs extended System, *Co-first Author* 2024.03 – 2024.10

An efficient and affordable LLM inference system with NDP design.

- A Heterogeneous Processing Architecture with NDP Design
- Load Balanced Scheduling on GPU and NDP-DIMMs
- 13.75 tokens/s for LLaMA2-70B & 75.24 × Speedup

Automatic and Lightweight LLM Pruner for N:M Sparsity, *First Author* 2024.03 – 2024.09

A novel pruning metric that balances weight & activation distribution.

- **Comprehensive Analysis** of Bias on Existing Pruning Metrics.
- **Automatic Optimization Strategy** for Searching the Optimal Pruning Metric.
- Reduce Comprehension Perplexity by 2.49

Accelerating LLM with Fine-Grained Mixed-Precision Quantization, *First Author* 2023.06 – 2024.06

An automatic framework to support mixed-precision quantization for LLM on **Commercial GPUs**.

- Fine-Grained Mixed-Precision Quantization for LLM to **Reduce Memory and Computation Overheads.**
- Automatic Framework with **Fine-Grained Scheduling & Data Conversion** for Efficient Kernel
- 2.03× Speedup on Commercial GPUs

Dynamic Precision Quantization for Neural Network Acceleration, *First Author* 2022.09 – 2023 09

An **algorithm-hardware co-design work** to accelerate widely used NN models, including LLMs.

- Dynamic Precision Quantization Method for All NN Models.
- Novel Architecture to Support Dynamic Precision Computation
- 2.85× Speedup & 3.12× Energy Saving

General Dynamic Neural Network Accelerator Design, *First Author* 2022.03 – 2023.11

The first work to design an accelerator for **general dynamic neural network** models.

- **Transverter-based Online Scheduling** for Efficient Compilation for Dynamic Tensor
- **Predictor-based Prefetching** to Reduce the Data Movement for Dynamic NNs.
- 3.48× Speedup & 3.03× Energy Saving

Neural Architecture and Quantization Joint-Optimization, *First Author* 2021.09 – 2023.03

A novel **automatic framework** to find the optimal combination of NN architecture and quantization strategy.

- **Differentiable Optimization Framework**, Precision Transfer Strategy for Unbiased Search
- 3.5× Speedup without Accuracy Loss

Honors and Awards

National Scholarship Sep. 2019
3rd Prize, Award on National System Design Competition Aug. 2019
Outstanding Graduates Jun. 2021
Excellent Student of ICT Sep. 2023