

刘炼

✉ liulian211@mailsucas.ac.cn · ☎ (+86) 159-8257-6242 · 🏠 <https://leliyiliu.github.io/> ·

🎓 教育背景

计算技术研究所, 中科院, 中国科学院大学, 北京 2021.09 – 至今

- 在读博士研究生 计算机系统结构, 预计 2026 年 6 月毕业
- 导师: 王颖研究员, 李华伟研究员

南开大学, 天津 2017.09 – 2021.06

- 学士 物联网工程
- 导师: 李涛教授, 卢冶副教授
- 绩点: 3.88/4, 排名: 1/45

🔍 研究兴趣

面向 LLM 的架构设计, 基于存算的系统优化, 基于算法和系统的协同优化框架

📄 论文

■ 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).

■ 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
Accepted by ICML'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).

■ 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).

在投论文

■ 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, 小修).

■ 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 MICRO'25, 2025. (CCF-A, In submission).

* Co-first author

🧑‍🔬 科研项目

- 📄 基于 CXL-PIM 的异构池化 LLM Serving 系统, 一作 2024.11 – 至今
- 基于池化的思路, 构建了特定的 **NPU 池** 和 **PIM 池** 来实现解耦合的优化思路
 - 基于 CXL controller 的硬件扩展和层次化优化, 实现 **PIM 池** 的灵活调度
- 📄 基于统一虚拟内存管理扩展的本地 LLM 部署系统, 一作 2024.10 – 至今
- 针对于 Attention Sparsity 中的优化孤立问题, 提出了一套**统一虚拟内存管理和扩展策略**
 - 通过基于 **UVM driver** 的底层优化, 实现了对于层次化内存系统的**用户不可知协同管理**
- 📄 基于层次化存算扩展的 LLM Serving 系统, 共同一作 2024.09 – 2025.03
- 低成本解决 LLM Serving 中 **KV cache** 计算, 访存和存储挑战的系统设计
- 提出了一套新颖的跨存储层异构计算策略, 以有效利用上下文局部性和**层次化存算 (PIM)** 设计
 - 全栈解决方案设计, 包括计算方法, 调度优化和系统扩展
 - 相比于 DGX-H100 vLLM 系统, 实现 **7.20× 性能提升 & 7.17× 计算效能优化 (tokens/\$)**
- 📄 基于近存 DIMM 设计的 LLM 本地部署系统, 共同一作 2024.03 – 2024.10
- 基于近存设计的高效且低成本的大模型推理系统 (RTX 4090 + NDP-DIMM), 以实现**高效的本地部署**
- 基于异构计算系统的两阶段 (**离线 & 在线**) **负载均衡调度**
 - 实现了 **13.75 tokens/s for LLaMA2-70B**, **75.25× 加速** 相比于现有系统
- 📄 基于聚类思路的 LLM KV cache 层次化压缩策略研究, AMD 实习期间 2024.07 – 2024.12
- 提出了一种基于 **KV cache** 融合的压缩策略, 支持灵活的压缩粒度调整
 - 相比于现有推理系统设计, 能够实现 **2.21× ~ 3.94× 效率提升**
- 📄 轻量化大模型权重剪枝方法研究, AMD 实习期间 2024.03 – 2024.09
- 系统分析现有剪枝策略在 LLM 剪枝中导致的不均匀分布问题
 - 基于轻量化搜索的大模型剪枝标准的**自动优化框架设计**, **减少 2.49 的混淆度**
- 📄 LLM 的细粒度混合精度量化和推理优化系统, 一作 2023.06 – 2024.06
- 提出了一个全新的细粒度混合精度量化算法来**减少计算和访存开销**
 - 基于 CUTLASS 实现了第一个在 GPU 上支持细粒度混合精度计算的 **GEMM 算子**, COMET-W4Ax
 - 相比于现有推理系统, 在 A100 上实现了 **2.03× 加速比**
- 📄 基于动态混合精度量化的 NN 加速器设计, 一作 2022.09 – 2023.09
- 面向通用神经网络模型的**动态精度量化算法设计**
 - 支持灵活动态精度矩阵计算的**可配置 NPU 架构设计**
 - 相比于现有 NPU 设计, 实现 **2.85× 加速 & 3.12× 功耗节省**
- 📄 通用动态神经网络加速器设计, 一作 2022.03 – 2023.11
- 基于**贪心调优策略**的在线调度探索, 以实现对动态张量计算的高效处理
 - 基于**关联性的预取方案设计**, 以减少动态神经网络执行过程中冗余数据搬运
 - 相比于面向静态神经网络的 NPU 设计, 实现 **3.48× 加速 & 3.03× 功耗节省**

♡ 获奖情况

计算所 所长优秀奖	2023 年 12 月
南开大学 优秀毕业生, 优秀毕业论文	2021 年 6 月
南开大学 国家奖学金	2019 年 10 月
三等奖, “龙芯杯” 系统能力培养大赛	2019 年 8 月