Tingfeng Lan

antlera.github.io Antlera erc8gx@virginia.edu

Research Interests

- I am broadly interested in co-designing systems and algorithms for **efficient large-scale machine learning**, with a focus on foundation models (e.g., GPT, LLaMA).

- Current research: 1) rethinks the design of large-scale systems for LLM applications in the interaction between computing and storage systems, and 2) optimizes/offloads/accelerates critical operations of LLM apps to the most appropriate hardware to harmonize heterogeneity, efficiency, and performance.

Education

University of Virginia Sep 2024 – Present
Ph.D. in Computer Science, Advisor: Prof. Yue Cheng *VA, USA*

Sichuan University Sep 2020 – Jun 2024
B.Eng. in Computer Engineering, Advisor: Prof. Mingjie Tang *Sichuan, China*

Industry Experience

AntGroup AI Infra Sep 2023 – Jul 2024
Research Intern, Manager: Jian Sha

*- Designed and implemented **DLRover-RM** (VLDB'24), a resource-aware optimization system for large-scale recommendation-model training that improves resource utilization and reduces training cost in cloud environments.*

*- Designed and implemented **m-LoRA** (VLDB'25), a multi-tenant LoRA training framework that enables parallel multi-adapter fine-tuning via pipeline parallelism, reducing memory redundancy and improving training throughput.*

Publications

Preprint Yinghao Tang, Tingfeng Lan, Xiuqi Huang, Hui Lu, Wei Chen. "**SCORPIO: Serving the Right Requests at the Right Time for Heterogeneous SLOs in LLM Inference**."

Preprint Tingfeng Lan, Yusen Wu, Bin Ma, Zhaoyuan Su, Rui Yang, Tekin Bicer, Masahiro Tanaka, Olatunji Ruwase, Dong Li, Yue Cheng. "**ZenFlow: Enabling Stall-Free Offloading Training via Asynchronous Updates**."
ZenFlow had been adopted into DeepSpeed.

Preprint Minchen Yu, Rui Yang, Chaobo Jia, Zhaoyuan Su, Sheng Yao, Tingfeng Lan, Yuchen Yang, Yue Cheng, Wei Wang, Ao Wang, Ruichuan Chen. "**λScale: Enabling Fast Scaling for Serverless Large Language Model Inference**."

Preprint Jiale Lao, Yinghao Tang, Tingfeng Lan, Mingjie Tang, Yuanchuan Zhou, Jianguo Wang. "**PathBee: Accelerating Shortest Path Querying via Graph Neural Networks**."

NSDI'26 Zirui Wang, Tingfeng Lan, Zhaoyuan Su, Juncheng Yang, Yue Cheng. "**ZipLLM: Efficient LLM Storage via Model-Aware Synergistic Data Deduplication and Compression**." *In Proceedings of the 23rd USENIX Symposium on Networked Systems Design and Implementation (to appear).*

VLDB'25 Zhengmao Ye*, Dengchun Li*, Zetao Hu, Tingfeng Lan, Jian Sha, Sicong Zhang, Lei Duan, Jie Zuo, Hui Lu, Yuanchun Zhou, Mingjie Tang. "**mLoRA: Fine-Tuning LoRA Adapters via Highly-Efficient Pipeline Parallelism in Multiple GPUs**." *In Proceedings of 51th International Conference on Very Large Data Bases*

VLDB'24 Qinglong Wang*, Tingfeng Lan*, Yinghao Tang, Bo Sang, Haitao Zhang, Jian Sha, Hui Lu, Ke Zhang, Mingjie Tang. "**DLRover-RM: Resource Optimization for Deep Recommendation Models Training in the Cloud**." *In Proceedings of 50th International Conference on Very Large Data Bases*

* denotes equal contribution

Open Source Projects

DeepSpeed-ZenFlow: A stall-free offloading framework for LLM fine-tuning Oct 2024 - Present

Available on DeepSpeed, Received 40k+ ⭐ on GitHub

*- Designed and implemented **ZenFlow**, an importance-aware asynchronous offloading system that decouples GPU and CPU updates to eliminate GPU stalls. Achieved up to 5× end-to-end speedup, 2× reduction in PCIe traffic, and over 85% stall elimination while preserving accuracy.*

mLoRA: A efficient multi-tenant LoRA training system Sep 2023 - May 2024

Received 300+ ⭐ on GitHub

- Designed and implemented a training mechanism "BatchLoRA" which allows multiple LoRA adapters to share the pre-trained base model concurrently with reduced kernel launch overhead.

DLRover: An efficient autodl system with fault-tolerance awareness Jun 2023 - March 2024

Received 1.5k+ ⭐ on GitHub, Joined LF AI & Data Foundation ⚡

- Designed and implemented a hyper-parameter autotuner to optimize performance-relevant configurations, like micro-batch size, for maximum hardware utilization. Achieved over 95% memory utilization within a 30s estimation and re-configuration time; An elastic trainer, allowing for real-time hyper-parameter configuration during training sessions, thereby eliminating the restart overheads typically necessary in conventional training frameworks.

Funding and Grants

2025 **Modal Research Grant**

Service & Activities

EXTERNAL SERVICE

2025-2026 **Artifact Evaluation Committee for EuroSys'26**

2023–2024 **Journal Reviewer for IEEE TBD'24**