

TINGFENG LAN

🏠 antlera.github.io 🔄 Antlera ✉ erc8gx@virginia.edu

Research Interests

Keywords: LLM infra, storage systems for AI, serverless computing

- The overarching goal of my research is to enable fundamentally sustainable, scalable, and performant AI applications and platforms by exploring the intersection of AI and systems.
- Current research focuses: (1) rethinking the LLM storage backend through new system abstractions that better align with modern model and workload characteristics, and (2) cost-effective training and inference systems on modern and heterogeneous hardware platforms.

Education

University of Virginia

Ph.D. in Computer Science, Advisor: Prof. Yue Cheng

Sep 2024 – Present

VA, USA

Sichuan University

B.Eng. in Computer Engineering, Advisor: Prof. Mingjie Tang

Sep 2020 – Jun 2024

Sichuan, China

Industry Experience

AntGroup AI Infra

Research Intern, Manager: Jian Sha

Sep 2023 – Jul 2024

- 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

Tingfeng Lan, Zirui Wang, Zhaoyuan Su, Yunjia Zheng, Juncheng Yang, Yue Cheng. “**TStore: Towards Practical Compression-Aware Model Storage with Tensor-Centric Fingerprinting and Clustering.**”

In submission to SOSP'26.

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.**”

In submission to IJCAI'26.

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.**”

Adopted into DeepSpeed and featured on PyTorch blog. In submission to SIGMOD'26.

ACL

Demo'26

Yinghao Tang, Yupeng Xie, Yingchaojie Feng, Tingfeng Lan, Wei Chen. “**IGenBench: Benchmarking the Reliability of Text-to-Infographic Generation.**”

- ACL'26** Yinghao Tang, Xueding Liu, Boyuan Zhang, [Tingfeng Lan](#), Yupeng Xie, Jiale Lao, Yiyao Wang, Haoxuan Li, Tingting Gao, Bo Pan, Luoxuan Weng, Xiuqi Huang, Minfeng Zhu, Yingchao-jie Feng, Yuyu Luo, Wei Chen. “**IGenBench: Benchmarking the Reliability of Text-to-Infographic Generation.**”
- MLSys'26** Zhaoyuan Su, [Tingfeng Lan](#), Zirui Wang, Juncheng Yang, Yue Cheng. “**Efficient and Workload-Aware LLM Serving via Runtime Layer Swapping and KV Cache Resizing.**”
- MLSys'26** 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.**”
In submission to ICDE'26.
- ICDE'26** Ziling Huang, Zhengmao Ye, Qingsong Cai, Zelong Huang, Bo Sang, Haitao Zhang, Jian Sha, [Tingfeng Lan](#), Hui Lu, Yuanchun Zhou, Mingjie Tang. “**DLRover-LM: LLM Pre-Training Framework with Thousands of Accelerators in AntGroup.**”
In Proceedings of the 42nd IEEE International Conference on Data Engineering.
- 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.
- 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 $5\times$ end-to-end speedup, $2\times$ 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 autotuner system with fault-tolerance awareness

Jun 2023 - March 2024

Received 1.6k+ ★ 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**; Total: \$2,000 GPU credits; Led the proposal writing and submission;

Service & Activities

2025-2026 **Shadow PC for EuroSys'26**

2025-2026 **Artifact Evaluation Committee for EuroSys'26, NSDI'26, OSDI'26**

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

Talks

2025 **ZenFlow: Enabling Stall-Free Offloading Training via Asynchronous Updates;**

Invited talk: DeepSpeed Team

2025 **ZenFlow: Enabling Stall-Free Offloading Training via Asynchronous Updates;**

Poster presentation: UVA 2025 AI/ML Resource Fair

Awards

2026 **EuroSys Distinguished AEC**