Ziming Liu (刘子铭)

E-mail: liuziming@comp.nus.edu.sg
Homepage: https://maruyamaaya.github.io/
Twitter/X: @lzm_mlsys
National University of Singapore / Peking University

Education

National University of Singapore, School of Computing

- Ph.D. in Computer Science **Jan. 2023 – Present**

National University of Singapore, School of Computing

- Master's degree in computer science (Artificial Intelligence) **Aug. 2021 – Jan. 2023**

Peking University, School of Electronics Engineering and Computer Science

- B.S. in Computer Science and Technology **Sep. 2016 – Jul. 2020**

Industry Experience

Qiji Zhifeng (Startup) **Jan. 2025 – Now**

Research Intern *Currently working on large-scale MoE Serving.*

Microsoft Research Asia. **May. 2024 – Nov. 2024**

Research Intern, System Group *Rewarded "Star of Tomorrow" certificate (**Top 10% intern**)*

HPC-AI Tech. **May. 2022 – Dec. 2022**

Research Intern

ByteDance Inc. **Aug. 2020 – Jul. 2021**

Machine Learning Engineer, Lark

Research Interests

Machine Learning System and High Performance Computing.

Including distributed model training (parallelism schemes) / inference and serving systems. Also working on efficient training/inference with sparsity.

Highlight Research Experience

StarTrail:
Concentric Ring Parallelism for Efficient Near-Infinite-Context Transformer Model Training

Advisor: Presidential Young Prof. You Yang, Prof. James Demmel ***Dec. 2023 – June.2024***

Objective: We develop a multi-dimensional sequence parallel system to reduce the communication volume and improve overall efficiency for long-sequence Transformer model training.

- This paper is accepted by NeurIPS2025.
- We conceptualize Attention computation as a novel instance of the traditional n-body problem, providing fresh insights into optimizing and parallelizing Attention computation.
- We introduce a near-infinite-context training system for Transformer models, featuring a groundbreaking multi-ring sequence parallelism scheme.
- Preliminary results indicate that our **StarTrail** system outperforms Ring Attention by up to 77.12%, showcasing its

efficacy and scalability.

Hanayo:Harnessing Wave-like Pipeline Parallelism for Enhanced Large Model Training Efficiency

Advisor: Presidential Young Prof. You Yang ***Dec. 2022 – Apr.2023***

Objective:We develop a new pipeline parallel technique to solve the problem the bubbles in existing pipeline model training techniques and achieve SOTA results in multiple tasks.

- This paper has been accepted by SC '23(The International Conference for High Performance Computing, Networking, Storage, and Analysis).
- We introduce a wave-like pipeline scheme that achieves a low bubble ratio and high performance in large model training.
- Experimental results demonstrate that Hanayo achieves up to a 30.4% performance improvement over the current state-of-the-art pipeline parallelism implementation.

WeiPipe: Weight Pipeline Parallelism for Communication-Effective Long-Context Large Model Training

Advisor: Presidential Young Prof. You Yang and Prof. Rong Zhao ***Apr. 2024 – Nov.2024***

Objective: We introduce weight-pipeline parallelism (WeiPipe) that transitions from an activation-passing pipeline to a weight-passing pipeline in long-context scenarios to reduce the communication volume and enhance efficiency.

- This paper has been accepted by PPoPP '25(ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming).
- We propose WeiPipe, WeiPipe-Interleave, and WeiPipe-Zero-Bubble that reduces the bubble ratio and relieve the communication requirements
- Experimental results demonstrate that WeiPipe can improve training efficiency by about 30%-80% compared to state-of-the-art PP and maintain weak and strong scalability under long-context scenarios.

Expert-as-a-Service: Towards Efficient, Scalable, and Robust Large-scale MoE Serving

May. 2024 – Dec.2024

Objective: Efficient, scalable, and robust MoE Serving System.

- Our system disaggregates MoE modules into independent, stateless services. This design enables fine-grained resource scaling and provides inherent fault tolerance by decoupling compute units.
- The architecture is powered by a high-performance, CPU-free peer-to-peer communication library that ensures minimal overhead and high throughput.
- EaaS incurs less than a 2% throughput reduction under simulated hardware failures that would otherwise halt monolithic architectures. It further saves up to 37.5% of computing resources through dynamic fine-grained adaptation to serving traffic, demonstrating strong resilience for large-scale MoE deployment in production.

Region-Adaptive Sampling for Diffusion Transformers

Advisor: Dr.Yuqing Yang ***May. 2024 – Dec.2024***

Objective: Efficient Diffusion Transformer Inference.

- We introduce RAS, a novel, training-free sampling strategy that dynamically assigns different sampling ratios to regions within an image based on the focus of the DiT model.
- We evaluate RAS on Stable Diffusion 3 and Lumina-NextT2I, achieving speedups up to 2.36x and 2.51x, respectively, with minimal degradation in generation quality.

Publication

Hanayo: Harnessing Wave-like Pipeline Parallelism for Enhanced Large Model Training Efficiency
Ziming Liu*, Shenggan Cheng*, Haotian Zhou, and Yang You
SC '23, *In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2023
*: Equal Contribution.

StarTrail:
Concentric Ring Parallelism for Efficient Near-Infinite-Context Transformer Model Training
Ziming Liu, Shaoyu Wang, Shenggan Cheng, Zhongkai Zhao, Yang Bai, Xuanlei Zhao, James Demmel, Yang You
NeurIPS2025, *In Proceedings of The Thirty-Ninth Annual Conference on Neural Information Processing Systems.*

WeiPipe: Weight Pipeline Parallelism for Communication-Effective Long-Context Large Model Training
Junfeng Lin***, Ziming Liu*,** Yang You, Jun Wang, Weihao Zhang, Rong Zhao
PPoPP '25, *ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*
*: Equal Contribution.

Concerto: Automatic Communication Optimization and Scheduling for Large-Scale Deep Learning
Shenggan Cheng, Shengjie Lin, Lansong Diao, Hao Wu, Siyu Wang, Chang Si, **Ziming Liu**, Xuanlei Zhao, Jiangsu Du, Wei Lin, Yang You
ASPLOS 2025, *ACM International Conference on Architectural Support for Programming Languages and Operating Systems*

DSP: Dynamic Sequence Parallelism for Multi-Dimensional Transformers
Xuanlei Zhao, Shenggan Cheng, Zangwei Zheng, Zheming Yang, **Ziming Liu**, and Yang You
ICML2025, *In Proceedings of International Conference on Machine Learning 2025*

HeteGen: Efficient Heterogeneous Parallel Inference for Large Language Models on Resource-Constrained Devices
Xuanlei Zhao, Bin Jia, Haotian Zhou, **Ziming Liu**, Shenggan Cheng, and Yang You
MLSys 2024, *In Proceedings of Machine Learning and Systems 2024*

Preprints

Expert-as-a-Service: Towards Efficient, Scalable, and Robust Large-scale MoE Serving
Ziming Liu, Boyu Tian, Guoteng Wang, Zhen Jiang, Peng Sun, Zhenhua Han, Tian Tang, Xiaohe Hu, Yanmin Jia, Yan Zhang, He Liu, Mingjun Zhang, Yiqi Zhang, Qiaoling Chen, Shenggan Cheng, Mingyu Gao, Yang You, Siyuan Feng
Under Review, Arxiv:2509.17863, 2025

Region-Adaptive Sampling for Diffusion Transformers
Ziming Liu, Yifan Yang, Chengruidong Zhang, Yiqi Zhang, Lili Qiu, Yang You, Yuqing Yang

Under Review, Arxiv:2502.10389, 2024

EnergonAI: An Inference System for 10-100 Billion Parameter Transformer Models
Jiangsu Du, **Ziming Liu**, Jiarui Fang, Shenggui Li, and Yongbin Li, Yutong Lu, Yang You
Arxiv: 2301.08658 , 2022

ATP: Adaptive Tensor Parallelism for Foundation Models
Shenggan Cheng, **Ziming Liu**, Jiangsu Du, and Yang You
Arxiv: 2209.02341, 2023

Skills

Languages: Python, C, C++, Latex
Frameworks: Pytorch, Huggingface, Megatron, Deepspeed, SGLang.