

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

Microsoft Research Asia.

May. 2024 – Nov. 2024

Research Intern, System Group

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.

Research Experience

WallFacer:

Guiding Transformer Model Training Out of the Long-Context Dark Forest with N-body Problem

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. (Python)

- This paper is currently under review.
- 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 WallFacer 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.
- Utilizing the action list, Hanayo's runtime system can support nearly all pipeline parallel algorithms while optimizing performance through features such as asynchronous communications.
- 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.

EnergionAI: An Inference System for 10-100 Billion Parameter Transformer Models

Advisor: Presidential Young Prof. You Yang

Apr. 2022 – Dec. 2022

Objective: With the large Transformer models trending, we develop a new inference system that support multiple parallelism(Tensor, Data, Pipeline and so on) and use various techniques to speed up the process. (Python)

- Design and implement checkpoint saving and loading system that supports various parallel schemes.
- Design and implement dynamic batch warping algorithm to speed up the process of multiple inference requests.
- Improve the implement of models like GPT and Bert to fit in with our parallel schemes.

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.

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

Junfeng Lin*, **Ziming Liu***, Yang You, Jun Wang, Weihao Zhang, Rong Zhao

To appear on 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

To appear on ASPLOS 2025, ACM International Conference on Architectural Support for Programming Languages and Operating Systems

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

WallFacer:

Guiding Transformer Model Training Out of the Long-Context Dark Forest with N-body Problem

Ziming Liu, Shaoyu Wang, Shenggan Cheng, Zhongkai Zhao, Yang Bai, Xuanlei Zhao, James Demmel, Yang You

Under Review, Arxiv: 2407.00611, 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

DSP: Dynamic Sequence Parallelism for Multi-Dimensional Transformers

Xuanlei Zhao, Shenggan Cheng, Zangwei Zheng, Zheming Yang, **Ziming Liu**, and Yang You
2024

Under Review, Arxiv: 2403.10266, 2024

Skills

Languages: Python, C, C++, Latex

Frameworks: Pytorch, Huggingface, Megatron, Deepspeed, Colossal AI