Jinghan Yao

2nd Year Ph.D, Dept. of Computer Science and Engineering
The Ohio State University, Columbus, Ohio

LinkedIn | Google Scholar | Personal Web | yjhmitweb@gmail.com | GitHub

Research Interests:

High-performance Computing & Communication, Machine Learning Systems, Computer Vision

Selected Research Experience:

High-performance Computing &. Communication in Generative Model Inference Sept.22 - Present

- **ExFlow, IPDPS'24** - Conducted an in-depth analysis of expert routing preferences within state-of-the-art Mixture-of-Experts Large Language Models (LLMs), leading to the novel introduction of **inter-layer expert affinity** and **coherent KV cache** strategies. These innovations significantly mitigate Alltoall routing overhead during distributed LLM inference, offering applicability across various GPT-like models and achieving an enhancement of up to 220% in inference throughput.
- **Flover, HiPC'23** - Designed a full-stack LLM inference framework, based on NVIDIA's FasterTransformer, in which I introduced temporal fusion (also known as in-flight batching) to increase the serving throughput and reduce per-request latency significantly. I proposed an efficient memory shuffle algorithm to guarantee a compact and contiguous KV cache. Flover outperforms NVIDIA's latest TensorRT-LLM and already has more than 150 downloads. This work has been selected for the NVIDIA GTC 2024 Oral Presentation.

Efficient Transformer for Computer Vision and Language Models Dec.20 - Aug.22

- **SOFT, NeurIPS'21** - We proposed a new Gaussian kernel-based attention that outperforms the vanilla dot-product-based attention. This attention exhibits symmetric positive semi-definite property which allows us to perform efficient low-rank approximation, largely increasing the training and inference throughput for both computer vision and NLP tasks.

Selected Publications:

1. **Jinghan Yao**, Quentin Anthony, Aamir Shafi, Hari Subramoni, Dhabaleswar K. Panda. "Exploiting Inter-Layer Expert Affinity for Accelerating Mixture-of-Experts Model Inference" Advances in IEEE International Parallel & Distributed Processing Symposium 38 (IPDPS 2024)
2. **Jinghan Yao**, Nawras Alnaasan, Tian Chen, Aamir Shafi, Hari Subramoni, Dhabaleswar K. Panda. "Flover: A Temporal Fusion Framework for Efficient Autoregressive Model Parallel Inference" Advances in IEEE International conference on High Performance Computing, Data, & Analytics 30 (HiPC 2023)
3. Lu, Jiachen, **Jinghan Yao**, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. "Soft: Softmax-free transformer with linear complexity." Advances in Neural Information Processing Systems 34 (NeurIPS 2021)
4. **Jinghan Yao**, Yu, Jun, Jian Zhang, Zhou Yu, and Dacheng Tao. "SPRNet: single-pixel reconstruction for one-stage instance segmentation." IEEE Transactions on Cybernetics

Selected Awards:

- **Oral Presentation at NVIDIA GTC' 24** - California, U.S
  - Flover: A Temporal Fusion Framework for Efficient Autoregressive Model Parallel Inference
- **Best Poster Award in ISC' 23** - Hamburg, Germany
  - MPI4Dask: Efficient MPI-based Communication For Scalable Accelerated Dask Applications
- **Spotlight(Top 3%) Paper Award in NeurIPS' 21** - Virtual Conference
  - SOFT: Softmax-free Transformer with Linear Complexity

Education:

Ph.D.	2027 (Expected)	The Ohio State University	Computer Science and Engineering	Advisor: Dhabaleswar K. Panda
RA	2022	Fudan University	School of Big Data	Advisor: Li Zhang
BS	2019	Hangzhou Dianzi University	Computer Science and Engineering	Advisor: Jun Yu