Jinghan Yao

4th year Ph.D., Department of Computer Science and Engineering
The Ohio State University

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

Education & *Work* Experiences:

Ph.D.	2027 (Expected)	The Ohio State University	Computer Science	Advisor: Dhabaleswar K. Panda
Research Intern	2026 (Ongoing)	*Anyscale*	*Ray Inference Team*	*Mentor: Kourosh Hakhamaneshi*
Research Intern	2025	*Microsoft*	*AI Frameworks Team*	*Mentor: Masahiro Tanaka*
Research Intern	2024	*Microsoft*	*DeepSpeed Team*	*Mentor: Sam Ade Jacobs*
Research Assistant	*2022*	*Fudan University*	*Zhang Vision Lab*	*Advisor: Li Zhang*
BS	2019	Hangzhou Dianzi University	Computer Science	Advisor: Jun Yu

Research Interests:

High-performance Computing & Communication, Machine Learning Systems

Selected Research Experience:

High-performance Computing & Communication in Large Language Model Training and Inference Sept.22 - Present

- ***Kernel-based GPU RDMA Design in MVAPICH*** - Actively involved in MVAPICH, one of the most widely used CUDA-Aware MPI libraries across US national laboratories and research facilities, where I provide support for GPU kernel-based designs in accelerating large-scale GPU communication through common collectives such as All-to-All, ReduceScatter, Allgather, etc, through RDMA and GPU Direct Async (GDA).
- ***MAC-Attention*, MLSys'26** – Co-designed a fidelity- and access-preserving long-context decoding method for LLMs in collaboration with **Microsoft AIFX**. **MAC-Attention** accelerates decode by reusing prior attention computations for semantically similar recent queries through a match-amend-complete pipeline, reducing KV-cache accesses while maintaining full-attention quality. Compared to the latest FlashInfer library, **MAC-Attention** reduces KV accesses by up to **99%**, cuts token generation latency by over **60%** at 128K context, and delivers over **14.3**× attention-phase speedups with up to **2.6**× end-to-end gains, while preserving full-attention accuracy. This work was accepted to MLSys 2026.
- ***NIMBLE*, IPDPS'26** – Designed a runtime communication orchestration system for skewed traffic patterns on modern GPU clusters with heterogeneous intra-node and inter-node interconnects. **NIMBLE** adaptively redistributes traffic across intermediate GPUs and rail-matched NICs through a capacity-normalized minimum-congestion optimization and CUDA-aware GPU kernel-based RDMA pipelining, while preserving ordering, determinism, and low overhead. **NIMBLE** outperforms NCCL and MPI by up to **5.2**× on skewed All-to-Allv workloads and **1.35**× on end-to-end LLM MoE blocks.
- ***DeepSpeed-FPDT*, MLSys'25, NVIDIA GTC'25** - Designed a resource-efficient LM training framework tailored for ultra-long context. The proposed design supports over **2 Million** context training with the latest 8B LLaMA model on **4 GPUs**, while reaching over **55% MFU**. This work has been integrated into Microsoft DeepSpeed framework (The dominant training framework for Large Language Models). This work has also been accepted to MLSys 2025. This work has also been selected by NVIDIA GTC 2025 conference.
- ***ExFlow*, IPDPS'24, NVIDIA GTC'25** - 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. This work has been accepted to IPDPS 2024, one of 48 accepted papers around the world. This work has also been selected by NVIDIA GTC 2025 conference.
- ***Flover*, HiPC'23, NVIDIA GTC'24** - 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. This work has also been selected by NVIDIA GTC 2024 conference as oral session.

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

- ***Softmax-free Transformer with Linear Complexity*, NeurIPS'21** - Proposed a new Gaussian kernel-based attention that outperforms the vanilla dot-product-based attention. This attention exhibits a 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, showing state-of-the-art performance on ImageNet, MS COCO, and Long Range Arena.

Selected Publications:

1. **Jinghan Yao**, Sam Ade Jacobs, Walid Krichene, Masahiro Tanaka, Dhabaleswar K Panda. "MAC-Attention: a Match-Amend-Complete Scheme for Fast and Accurate Attention Computation" In proceeding of The Ninth Annual Conference on Machine Learning and Systems (MLSys 2026)
2. **Jinghan Yao**, Kaushik Kandadi, Bharath Ramesh, Hari Subramoni, Dhabaleswar K Panda. "From Skew to Symmetry: Node-Interconnect Multi-Path Balancing with Execution-time Planning for Modern GPU Clusters" In proceeding of IEEE International Parallel & Distributed Processing Symposium 40 (IPDPS 2026)
3. **Jinghan Yao**, Sam Ade Jacobs, Masahiro Tanaka, Olatunji Ruwase, Aamir Shafi, Hari Subramoni, Dhabaleswar K Panda. "Training Ultra Long Context Language Model with Fully Pipelined Distributed Transformer" In proceeding of The Eighth Annual Conference on Machine Learning and Systems (MLSys 2025)
4. **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)
5. **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)
6. Jiachen Lu, **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)

Selected Awards:

- **Best Poster Award in ISC' 26** - Hamburg, Germany
  - NIMBLE: Node-Interconnect Multi-Path Balancing with On-the-fly Orchestration for High Bandwidth GPU Clusters
- **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