Hao Wu

☎+1 (571)653-9582 hwu27@gmu.edu

Education

01/2024–Present **George Mason University** Fairfax, VA, USA
Degree: Ph.D. in Computer Science
Advisor: Dr. Keren Zhou
Area of Research: Profiler and Debugger for ML Programs

09/2020–06/2023 **University of Science and Technology of China** Hefei, Anhui, China
Degree: M.Sc in Computer Science
Thesis: Distributed ML Systems

09/2016–06/2020 **University of Science and Technology of China** Hefei, Anhui, China
Degree: B.Sc in Computer Science

Professional Experiences

05/2025–08/2025 Software Engineer Intern at **Meta** Bellevue, WA, USA
06/2023–01/2024 Software Engineer II at **Alibaba Cloud Intelligence** Beijing, China
08/2021–11/2021 Software Engineer Intern at **Baidu** Beijing, China
12/2019–09/2020 Software Engineer Intern at **ByteDance** Shanghai, China

Publications

Conferences

[**C1**] Shenggan Cheng, Shengjie Lin, Lansong Diao, **Hao Wu**, Siyu Wang, Chang Si, Ziming Liu, Xuanlei Zhao, Jiangsu Du, Wei Lin, Yang You. *Concerto: Automatic Communication Optimization and Scheduling for Large-Scale Deep Learning*. In: Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2025

[**C2**] Qidong Zhao, **Hao Wu***, Yuming Hao, Zilingfeng Ye, Jiajia Li, Xu Liu, Keren Zhou. (Qidong and Hao are co-first authors.) *DeepContext: A Context-aware, Cross-platform, and Cross-framework Tool for Performance Profiling and Analysis of Deep Learning Workloads*. In: Proceedings of the 31th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2026

Posters

[**P1**] Qidong Zhao, **Hao Wu**, and Keren Zhou, Torch-Monitor: A Comprehensive Call Path Profiling Tool for PyTorch. In PyTorch conference (PyTorch), 2024

Journals

[**J1**] **Hao Wu**, Shiyi Wang, Youhui Bai, Cheng Li, Quan Zhou, Jun Yi, Feng Yan, Ruichuan Chen, and Yinlong Xu. *A Generic, High-Performance, Compression-Aware Framework for Data Parallel DNN Training*. In: IEEE Transactions on Parallel and Distributed Systems (TPDS), 2023

Projects

01/2024–Present **George Mason University** Fairfax, VA, United States

DeepContext—A Context-Aware, Cross-Platform ML Profiler

- Implemented **callback-based** instrumentation on JAX for performance profiles and call path information.
- Evaluated DeepContext's overhead on 10 different ML workloads across PyTorch and JAX, deployed on NVIDIA and AMD GPUs.
- Demonstrated typical optimizations by 8 case studies. Achieved speedup ranging **from 1.06× to 1.66×**.
- Submitted to ASPLOS 2025 Fall, under review.

06/2023–01/2024 **Platform of A.I., Alibaba Cloud Intelligence** Beijing, China

EasyDist—An Automatic Distributed Parallel Training Framework

- Built a customized CUDA memory allocator that **traces** all tensor allocations and **maps** them with corresponding tensors. Integrated into our DL compiler framework *Concerto*, published in ASPLOS 2025.
- Implemented rule-based strategies and dominant tree algorithm to reduce the complexity of EasyDist's sharded computation graph. Achieved **up to 7.2**× graph size reduction on a typical Resnet model.

05/2022–02/2023 **University of Science and Technology of China** Hefei, Anhui, China

HiPress—A Gradient Compression Framework for Data Parallel Training

- Implemented PowerSGD, a low-rank gradient compression algorithm.
- Built a task coordinator to overlap the compression-based cross-node communication with DNN computation, maximizing training performance.
- Compared to PowerSGD baselines provided by TorchDDP, our framework achieved **21.8%-23.0**% throughput improvement on a cluster of 128 Tesla V100s.

08/2021–11/2021 **Deep Learning Technology Platform, Baidu** Beijing, China

GPU Memory Optimization for LLM training

- Implemented **ZeRO stage-2** atop PaddlePaddle, Baidu's distributed DL training framework; Validated loss curve on GPT2-xl model in a single machine with 4 V100 GPUs.
- ZeRO enabled PaddlePaddle to support large language models with up to **10 billion** parameters.

12/2019–09/2020 **Interactive Entertainment Services, Bytedance** Shanghai, China

Resso—TikTok's music streaming App

- Implemented animation effects for Resso's campus promotion and playlist sharing, with an **approved patent (CN111970571A)**.
- Implemented WebView preloading module, reducing the average loading time **from 4.1s to 1.8s**.
- Refactored Setting pages, playing pages and subscription pages.

Awards and Patents

2020—2022	Scholarships for Master's Degree Students
2020	Patent: A new method of video making (CN111970571A)
2018	No.3 USTC Hackergame, 3rd Prize (Top 5%)
2018	Intel Parallel Application Challenge, Best Application Bronze Prize (Top 5%)
2018	Member of the USTC Swangeese Undergraduate Supercomputing Contest Team
2017	Freshman Seminar Best Paper Award
2016	No. 33 National High School Physics Olympiads (Shanghai), 1st Prize (Top 1%)