Cunyang Wei

Email: cunyang@umd.edu

EDUCATION

University of Maryland, College Park **09/2024 – Present**
Ph.D. in Computer Science

University of Chinese Academy of Sciences **09/2020 – 06/2023**
M. S. in Computer Technology

Zhengzhou University **09/2016 – 06/2020**
B. S. in Mathematics and Applied Mathematics

RESEARCH INTERESTS

- High-performance computing, Systems for Machine Learning, Distributed and Parallel Training

RESEARCH EXPERIENCE

Unmasking Network-Induced Performance Variability in GPU-Accelerated Supercomputers **09/2024 – Present**

- Modern HPC systems are increasingly challenged by performance variability that significantly impacts both scientific simulations and AI training, with even minor delays on a single node causing widespread job slowdowns. This issue is exacerbated by heterogeneous hardware, software jitter, and especially network contention, leading to inefficient resource usage and higher operational costs.
- Our study is the first to systematically investigate network-induced performance variability on modern GPU clusters, revealing that network delays are the dominant factor affecting overall system performance.
- We conducted a longitudinal study on production systems such as Perlmutter and Frontier, collecting extensive real-world data across both traditional MPI applications and distributed deep learning workloads. These novel insights provide actionable strategies for mitigating network bottlenecks, underscoring the originality and importance of our work in advancing HPC and AI system efficiency.

Taming Billion-edge Graphs with 3D Parallel Full-graph GNN Training **09/2024 – Present**

- Proposed a novel 3D parallel algorithm to address memory, communication, and load-balancing challenges in large-scale GNN training, enabling efficient distribution of graph data and computation across thousands of GPUs.
- Designed a performance model to automatically select optimal 3D virtual GPU grid configurations and designed a double permutation scheme to achieve near-perfect load balancing for sparse graph data.
- Achieved unprecedented scalability up to 2048 GPUs on the Frontier and Perlmutter supercomputers, delivering up to a 54.2x speedup over state-of-the-art frameworks.

Optimization of LLM Inference Framework on Mobile GPU **07/2023 – 01/2024**

- Accelerated LLaMA-7B inference on mobile GPUs (Qualcomm Adreno 740) by co-designing computation scheduling and memory optimization strategies.
- Optimized tall-and-skinny matrix multiplication kernels for the prefill phase computational bottleneck, achieving 4.0× performance improvement over CLBlast baseline through sophisticated tiling algorithms and strategic on-chip memory utilization.
- Enhanced GEMV operation efficiency in the decode phase, delivering >90% peak memory bandwidth utilization through targeted algorithmic improvements and hardware-aware optimization techniques.

IrGEMM: An Input-Aware Tuning Framework for Irregular GEMM on ARM and X86 CPUs **10/2021 – 04/2023**

- Generated hundreds of highly optimized assembly kernels for diverse irregular GEMM types based on computing templates, the instruction mapping rules between templates and assembly codes, and pipeline optimization strategies.
- Abstracted tiling problems in GEMM into bin packing problems and applied a dynamic programming approach to

minimize memory access and maximize computational efficiency.
- Built a load-balanced multithreaded scheduling framework for processing batch matrix multiplication to achieve the ultimate multi-threaded speedup.
- Implemented a high-performance irregular matrix multiplication library for ARMv8 and Intel cascade Lake architectures.
- Achieved 2.3x, 2.7x, and 2.5x speedups for three representative shapes compared to Intel MKL, ARMPL, LIBXSMM, and BLIS, respectively.

High-performance Image Processing Algorithms Optimization Based on ARMv8 CPUs **10/2020 – 10/2021**
- Sorted image processing algorithms into three types (data irrelevant algorithm, data sharing algorithm and irregular memory access algorithm).
- Built a high-performance image processing algorithms library by writing the underlying code with Arm Neon Intrinsic and optimizing multi-threaded performance with OpenMP.
- Presented optimized image processing algorithm library based on ARMv8 architecture and substantially improved the image processing performance by optimizing the algorithms, memory access, SIMD, and assembly instruction.
- Increased the speed-up ratio of cvtColor, Resize and Filter modules to 1.2x, 2x, and 2x in comparison to the OpenCV algorithms library.

PUBLICATIONS

- Aditya K. Ranjan, Siddharth Singh, **Cunyang Wei**, Abhinav Bhatele. *Plexus: Taming Billion-edge Graphs with 3D Parallel GNN Training*. Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis (SC) 2025.
- **Cunyang Wei**, Haipeng Jia, Yunquan Zhang, Jianyu Yao, et al. *IrGEMM: An Input-Aware Tuning Framework for Irregular GEMM on ARM and X86 CPUs.* IEEE Transactions on Parallel and Distributed Systems (TPDS) 2024.
- Luhan Wang, Haipeng Jia, Lei Xu, **Cunyang Wei**, Kun Li, et al. *VNEC: A Vectorized Non-Empty Column Format for SpMV on CPUs.* IEEE International Parallel and Distributed Processing Symposium (IPDPS) 2024.
- Rongyuan Guo, Haipeng Jia, Yuanquan Zhang, Mingsen Deng, **Cunyang Wei**, et al. *SA_TRSM: A Shape-Aware Auto-Tuning Framework for Small-Scale Irregular-Shaped TRSM.* IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS) 2023.
- **Cunyang Wei**, Haipeng Jia, Yunquan Zhang, Liusha Xu, and Ji Qi. *IATF: An Input-Aware Tuning Framework for Compact BLAS Based on ARMv8 CPUs*. In 51st International Conference on Parallel Processing (ICPP) 2022.
- **Cunyang Wei**, Haipeng Jia, Yunquan Zhang, et al. *LBBGEMM: A Load-Balanced Batch GEMM Framework on ARM CPUs*. IEEE International Conference on High Performance Computing & Communications (HPCC) 2022.
- Luhan Wang, Haipeng Jia, Yunquan Zhang, Kun Li, **Cunyang Wei**. *EgpuIP: An Embedded GPU Accelerated Library for Image Processing*. IEEE International Conference on High Performance Computing & Communications (HPCC) 2022.

• HONORS AND AWARDS

- MVAPICH User Group Conference Travel Grant **2025**
- Dean’s Fellowship, University of Maryland, College Park **2024**
- Outstanding Graduate of Beijing, Beijing Municipal Education Commission **2023**
- Outstanding Graduate, University of Chinese Academy of Sciences **2023**
- National Scholarship (top scholarship in China), Ministry of Education of the People's Republic of China **2022**

PROFESSIONAL SERVICE

Session Chair for IEEE HPCC’22 **12/2022**
TPC reviewer for IEEE ISPA’25 **07/2025**

TEACHING EXPERIENCE

Teaching Assistant for DATA200 - Knowledge in Society: Science, Data and Ethics **Spring 2025**
Teaching Assistant for CMSC216 - Introduction to Computer Systems **Fall 2024**