Wenyi Wang

(872)-806-9983 | 5730 S. Ellis Ave, Chicago | wenyiw@uchicago.edu | wenyiwang-us.github.io

Research Interests

My core research interests are at the intersection of computer systems and machine learning, centered on building fast, scalable, and efficient solutions by exploiting extremely fine-grained parallelism. My current projects focus on large-scale, efficient LLM inference on many-node HPC systems, alongside the advancement of fine-grained parallelism techniques in OpenMP tasking models.

Projects

MLSys

MPI-vLLM: *Large Scale LLM Serving*

- Enabling fast model-switching capabilities with MPI backend on limited GPU resources.
- Designing scalable workflows to deploy MPI-vLLM on hundred-nodes HPC systems.
- Conducting comprehensive experiments to study MPI-vLLM's scalability and performance

HPC

GNU-XTask: *Optimizing Fine-Grained Parallelism Through Dynamic Load Balancing on Multi-socket Many-core Systems*

- Enabled extremely fine-grained parallelism leveraging lock-less data structures XQueue on GNU-OpenMP.
- XQueue-enabled GNU-OpenMP shows up to 1522.8× performance improvement on an eight-socket machine with 192 threads.
- Developed GNU-XTask that is equipped with NUMA-aware, lock-less dynamic load balancing techniques that can bring 4× additional performance improvement over XQueue-enabled GNU-OpenMP

UDSHMEM: *High-performance SHMEM Library for UpDown Accelerators (Advised by Yanjing Li and Andrew A. Chien)*

- Designed and implemented UDSHMEM data movement library based on SHMEM's specification on top of the novel UpDown computer system, written in UpDown assembly language.
- UDSHMEM achieved optimal system-wide bandwidth, saturating system's capacity while respecting data layout in memory banks and data locality across physical nodes.

Interweaving Project @ Northwestern University: *Paths to OpenMP in the Kernel*

- Achieved an average performance gain of 22% (geometric mean) across scales and benchmarks for runtime in kernel implementation by inspecting runtime behavior.
- Customized LLVM/OpenMP runtime library libomp and implemented pthread-embedded library to make libomp function within Nautilus kernel.
- Discovered a Floating-Point logic error in Nautilus codebase by benchmarking Gaussian elimination.
- Ported different benchmarks including NAS Parallel Benchmarks

Other Projects

Project Us @ MIT Media Lab: *Connecting Us*

- Led full-stack development of real-time cloud-based AI emotion recognition system with React, Flask, MongoDB and TensorFlow for initial client evaluation, pushing the project to participate in the MIT delta-v program.

AITom @ CMU: *AI Tools for Tomography*

- Led the research on 3D saliency detection for Cryo-ET by applying attention mechanism and teacher-student model in an unsupervised environment.

Education

University of Chicago 2022 — Present
Ph.D. Computer Science, GPA: 4.0/4.0, Advisors: Kyle Chard, Ian Foster *Chicago, IL*

Northwestern University 2019 — 2021
M.S. Computer Science, GPA: 3.9/4.0, Advisor: Peter Dinda *Evanston, IL*

Northeastern University 2015 — 2019
B.E. Software Engineering, Major GPA: 3.9/4.0, Advisor: Tao Ren *Shenyang, China*

Publications

[**SC Posters'25**] **W. Wang**, M. Gonthier, H. Lai, P. Nookala, H. Pan, I. Foster, I. Raicu, K. Chard, **"Exploring Fine-Grained Parallelism in Data-flow Runtime Systems on Many-Core Systems"**, *The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC25), Research Poster*, 2025 .

[**SC Workshop'25**] A. Fell, Y. Wang, T. Su, M. Nourian, **W. Wang**, J. M. Monsalve-Diaz, A. S. Rajasukumar, J. Su, R. Xu, R. Khandelwal, T. Zhang, D. Gleich, Y. Li, H. Hoffmann, A. A. Chien, **"KVMSR+UDWeave: Extreme-Scaling with Fine-grained Parallelism on the UpDown Graph Supercomputer"**, *The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC25), Workshop*, 2025 .

[**IPDPS'25**] **W. Wang**, M. Gonthier, P. Nookala, H. Pan, I. Foster, I. Raicu, K. Chard, **"Optimizing Fine-Grained Parallelism Through Dynamic Load Balancing on Multi-Socket Many-Core Systems"**, *2025 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 2025 , Pages: 81-93.

[**arXiv'24**] A. S. Rajasukumar, J. Su, T. Su, M. Nourian, J. M. Monsalve-Diaz, T. Zhang, J. Ding, **W. Wang**, Z. Zhang, M. Jeje, H. Hoffmann, Y. Li, A. A. Chien, **"UpDown: Programmable fine-grained Events for Scalable Performance on Irregular Applications"**, *preprint*, 2024 .

[**LCPC Workshop'23**] Y. Wang, A. S. Rajasukumar, T. Su, M. Nourian, J. M. Monsalve-Diaz, A. Pervaiz, J. Ding, C. Colley, **W. Wang**, Y. Li, D. F. Gleich, H. Hoffmann, A. A. Chien, **"Efficiently Exploiting Irregular Parallelism Using Keys at Scale"**, *International Workshop on Languages and Compilers for Parallel Computing (LCPC Workshop)*, 2023 .

[**BDCAT'23**] N. C. Hudson, J. G. Pauloski, M. Baughman, A. Kamatar, M. Sakarvadia, L. Ward, R. Chard, A. Bauer, M. Levental, **W. Wang**, W. Engler, O. P. Skelly, B. Blaiszik, R. Stevens, K. Chard, I. Foster, **"Trillion parameter ai serving infrastructure for scientific discovery: A survey and vision"**, *Proceedings of the IEEE/ACM 10th International Conference on Big Data Computing, Applications and Technologies*, 2023 .

[**SC'21**] J. Ma, **W. Wang**, A. Nelson, M. Cuevas, B. Homerding, C. Liu, Z. Huang, S. Campanoni, K. Hale, P. Dinda, **"Paths to OpenMP in the Kernel"**, *International Conference for High Performance Computing, Networking, Storage and Analy (SC21)*, 2021 .

Honors

- Intel Student Ambassador, 2025
- Crerar Fellowship, 2022
- Exceptional Funding of the Nation (China), awarded to the top 5%, the 12th National Innovation Training Program for College Students, 2018
- Gold Award, China College Students' Entrepreneurship Competition in Liaoning Province, 2018
- Nationwide Second Prize, China, "Innovation has a future" University AI Innovation Grand Competition, 2018

Teaching

- MPCS 52040 (Winter 2025) Distributed Systems
- MPCS 52040 (Autumn 2025) Distributed Systems

Skills

- Frameworks/libs: vLLM, PyTorch, OpenMP, Taskflow, SHMEM, MPI, Tensorflow, Flask, React, MongoDB.
- Programming Language:C, C++, Python, Java, Rust.