Weijian Zheng

Phone: (317) 370-4736 Email: wzheng@anl.gov / zwj3652006@gmail.com
Homepage: http://www.weijianzheng.net/

Research Interests:

Distributed Systems, AI4Science, Parallel Algorithms Design,
Parallel Machine Learning, High-Performance Computing, LLMs

Professional Experiences

Argonne National Laboratory Assistant Computer Scientist, Jan. 2026 - current — Lemont, IL, USA
Argonne National Laboratory Postdoc Appointee, Jan. 2023 - Dec. 2025 — Lemont, IL, USA
Oak Ridge National Laboratory Research Assistant, Aug. 2018 - Dec. 2022 — Oak Ridge, TN, USA
Oak Ridge National Laboratory Research Intern, May 2018 - Aug. 2018 — Oak Ridge, TN, USA
Indiana U Purdue U Indianapolis Graduate Research Assistant, Aug. 2014 - Dec. 2022 — Indianapolis, IN, USA
Ball State University Teaching Assistant, June 2013 - July 2013 — Muncie, IN, USA

Education

Ph.D. in Computer Science, Purdue University, Dec. 2022 — West Lafayette, IN, USA
B.S. in Computer Science (*Summa cum Laude*), Ball State University, July 2014 — Muncie, IN, USA

Publications

1. Ziyu Hu*, Jiamin Wang, Zhiqing Zhong, **Zheng, Weijian**, Hemant Sharma, Jun-Sang Park, Peter Kenesei, Antonino Miceli, Zhaorui Zhang, Rajkumar Kettimuthu, et al. Fastrei: Fast rare event identification on x-ray data with cross-stage optimizations. In *2025 IEEE International Conference on Big Data (BigData)*, pages 2169–2176. IEEE, 2025a. *Student advised

2. Austin Yunker*, **Weijian Zheng***, and Rajkumar Kettimuthu. Inferct: An efficient and generalizable framework to enable 3d machine learning for computed tomography. In *Proceedings of the SC'25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 71–77, 2025. *Equal contribution

3. Flavio Castro, **Zheng, Weijian**, Joaquin Chung, Ian Foster, and Raj Kettimuthu. To stream or not to stream: Towards a quantitative model for remote hpc processing decisions. In *Proceedings of the SC'25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis (best paper award)*, pages 834–840, 2025

4. Ziyu Hu*, Zhiqing Zhong, **Zheng, Weijian**, Zhijing Ye, Xuwei Tan, Xueru Zhang, Zheng Xie, Rajkumar Kettimuthu, and Xiaodong Yu. Dabench-llm: Standardized and in-depth benchmarking of post-moore dataflow ai accelerators for llms. In *2025 IEEE International Symposium on Workload Characterization (IISWC)*, pages 127–141. IEEE, 2025b. *Student advised

5. Dali Wang, Chen Wang, Qinglei Cao, Jayesh Krishna, Danqing Wu, **Zheng, Weijian**, Peter Schwartz, Fengming Yuan, Kathryn Mohror, and Peter Thornton. Scaling Ultrahigh-Resolution E3SM Land Model for Leadership-Class Supercomputers. In *2025 IEEE 25th International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW)*, pages 1–4. IEEE, 2025

6. Hanzhi Zhang, Sumera Anjum, Heng Fan, **Zheng, Weijian**, Yan Huang, and Yunhe Feng. Poly-FEVER: A Multilingual Fact Verification Benchmark for Hallucination Detection in Large Language Models. *arXiv preprint arXiv:2503.16541*, 2025

7. Chengming Zhang, Xinheng Ding, Baixi Sun, Xiaodong Yu, **Zheng, Weijian**, Zhen Xie, and Dingwen Tao. GFormer: Accelerating Large Language Models with Optimized Transformers on Gaudi Processors. *arXiv preprint arXiv:2412.19829*, 2024

8. **Zheng, Weijian**, Hemant Sharma, Ryan Chard, Peter Kenesei, Jun-Sang Park, Nicholas Schwarz, Antonino Miceli, Ian T Foster, and Rajkumar Kettimuthu. Model and Data Management for Machine Learning (M2ML): Integrating Instruments, Edge and HPC for Accelerated Machine Learning. In *2024 IEEE International Conference on Big Data (BigData)*, pages 4275–4282. IEEE, 2024c

9. **Zheng, Weijian**, Jack Kordas, Tyler J Skluzacek, Raj Kettimuthu, and Ian Foster. Globus service enhancements for exascale applications and facilities. *The International Journal of High Performance Computing Applications*, 38(6): 658–670, 2024a

10. **Zheng, Weijian**, J-S Park, Peter Kenesei, Ahsan Ali, Zhengchun Liu, Ian Foster, Nicholas Schwarz, Rajkumar Kettimuthu, Antonino Miceli, and Hemant Sharma. Rapid detection of rare events from in situ X-ray diffraction data using machine learning. *Journal of Applied Crystallography*, 57(4), 2024b

11. Kareem Shaik, Dali Wang, **Zheng, Weijian**, Qinglei Cao, Heng Fan, Peter Schwartz, and Yunhe Feng. S3LLM: Large Scale Scientific Software Understanding with LLMs Using Source, Metadata, and Document. In *International Conference on Computational Science*, pages 222–230. Springer, 2024

12. Shihui Song, Yafan Huang, Peng Jiang, Xiaodong Yu, **Zheng, Weijian**, Sheng Di, Qinglei Cao, Yunhe Feng, Zhen Xie, and Franck Cappello. Ceresz: Enabling and scaling error-bounded lossy compression on cerebras cs-2. In *Proceedings of the 33rd International Symposium on High-Performance Parallel and Distributed Computing*, pages 309–321, 2024

13. Jim Pruyne, Valerie Hayot-Sasson, **Zheng, Weijian**, Ryan Chard, Justin M Wozniak, Tekin Bicer, Kyle Chard, and Ian T Foster. Steering a fleet: Adaptation for large-scale, workflow-based experiments. *arXiv preprint arXiv:2403.06077*, 2024

14. Chengming Zhang, Baixi Sun, Xiaodong Yu, Zhen Xie, **Zheng, Weijian**, Kamil A Iskra, Pete Beckman, and Dingwen Tao. Benchmarking and in-depth performance study of large language models on habana gaudi processors. In *Proceedings of the SC'23 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis*, pages 1759–1766, 2023

15. Aniket Tekawade, Viktor Nikitin, Yashas Satapathy, Zhengchun Liu, Xuan Zhang, Peter Kenesei, **Zheng, Weijian**, Francesco De Carlo, Ian Foster, and Rajkumar Kettimuthu. Tomo2Mesh: Fast porosity mapping and visualization for synchrotron tomography. In *2023 IEEE 19th International Conference on e-Science (e-Science)*, pages 1–10. IEEE, 2023

16. Yunhe Feng, Sreecharan Vanam, Manasa Cherukupally, **Zheng, Weijian**, Meikang Qiu, and Haihua Chen. Investigating Code Generation Performance of Chat-GPT with Crowdsourcing Social Data. In *Proceedings of the 47th IEEE Computer Software and Applications Conference*, pages 1–10, 2023

17. **Zheng, Weijian**, Dali Wang, and Fengguang Song. A Distributed-GPU Deep Reinforcement Learning System for Solving Large Graph Optimization Problems. *ACM Transactions on Parallel Computing*, 2023

18. Dali Wang, Peter Schwartz, Fengming Yuan, Peter Thornton, and **Zheng, Weijian**. Towards Ultra-high-resolution E3SM Land Modeling on Exascale Computers. *Computing in Science & Engineering*, 2022

19. Jian Zhou, **Zheng, Weijian**, Dali Wang, and David W Coit. A resilient network recovery framework against cascading failures with deep graph learning. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, page 1748006X221128869, 2022

20. **Zheng, Weijian**, Dali Wang, and Fengguang Song. Designing a parallel Feel-the-Way clustering algorithm on HPC systems. *The International Journal of High Performance Computing Applications (IJHPCA)*, 35(2):154–169, 2021.

21. **Zheng, Weijian**, Dali Wang, and Fengguang Song. OpenGraphGym: A parallel reinforcement learning framework for graph optimization problems. In *International Conference on Computational Science (ICCS)*, pages 439–452. Springer, 2020.

22. **Zheng, Weijian**, Dali Wang, and Fengguang Song. FQL: An extensible feature query language and toolkit on searching software characteristics for HPC applications. In *Tools and Techniques for High Performance Computing (SE-HER)*, pages 129–142. Springer, 2019b.

23. **Zheng, Weijian**, Dali Wang, and Fengguang Song. XScan: An Integrated Tool for Understanding Open Source Community-Based Scientific Code. In *International Conference on Computational Science (ICCS)*, pages 226–237. Springer, 2019a.

24. **Zheng, Weijian**, Fengguang Song, Lan Lin, and Zizhong Chen. Scaling Up Parallel Computation of Tiled QR Factorizations by a Distributed Scheduling Runtime System and Analytical Modeling. *Parallel Processing Letters (PPL)*, 28(01): 1850004, 2018.

25. Dali Wang, **Zheng, Weijian**, and Fengguang Song. Application Software Analytics Toolkit for Facilitating the Understanding, Componentization, and Refactoring of Large-Scale Scientific Models. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 2018.

26. **Zheng, Weijian**, Fengguang Song, and Lan Lin. Designing a Synchronization-reducing Clustering Method on Manycores: Some Issues and Improvements. In *Proceedings of the Machine Learning on HPC Environments (MLHPC)*, page 9. ACM, 2017.

27. **Zheng, Weijian**, Fengguang Song, Lan Lin, and Zizhong Chen. suCAQR: A Simplified Communication-Avoiding QR Factorization Solver Using the TBLAS Framework. In *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)*, pages 1092–1099. IEEE, 2016.

Grants and Awards

Awards

2025 Argonne Postdoctoral Performance Awards – 12 recipients in 2025

Grant Awards

"Multimodal AI Foundation Modeling of Turbulent, Multiphase, and Reacting Flows for Propulsion and Power Applications."
– Argonne Laboratory Directed Research & Development (LDRD) Prime Funding, 2026,
– $400,000, Role: Co-PI
"Use of Generative AI and LLMs for Accelerator Design and Optimization."
– Argonne Laboratory Directed Research & Development (LDRD) AI for Science Project Funding, 2025,
– $100,000, Role: Investigator
"Towards Fast Quantitative Regression Analysis with Vision Transformers and Bayesian Neural Networks."
– Argonne Laboratory Directed Research & Development (LDRD) Innovate (Seed) Funding, 2024,
– $25,000, Role: Sole PI

Computing Resource Awards

"Accelerating 3D Tomography Analysis with High-Performance Computing."
– ALCF Director's Discretionary (DD) Allocation, Argonne National Laboratory, 2025,
– 3,000 node-hours on Aurora Supercomputer, Role: PI
"Co-optimize AI Algorithm and Computer System for Accelerating High-resolution Image Generation.":
– ALCF Director's Discretionary (DD) Allocation, Argonne National Laboratory, 2024,
– 3,000 node-hours on Polaris and Sophia Supercomputer, Role: PI
"Advanced Applications and Performance Benchmarking of the Cerebras CS-2 in High-Volume Image Processing and Computational Analysis."
– NSF-funded ACCESS Allocation at Pittsburgh Supercomputing Center (PSC) 2024,
– 400,000 ACCESS credits on Neocortex, Role: PI
"Evaluation of NAS for X-ray applications on next generation AI hardware."
– ALCF Director's Discretionary (DD) Allocation, Argonne National Laboratory, 2024,
– AI-testbed (Cerebras, Graphcore, and SambaNova), Role: PI

Professional Activities

Conferences Program Committee Members and Reviewers

TPC for IEEE Cloud Summit, 2025, 26
TPC for IEEE Intl. Conf. on Distributed Computing Systems (ICDCS), 2025, 26
PC Member for Intl. Conf. on Dependable Systems and Networks (DSN), 2025
Reviewer for Intl. Conf. for High Performance Computing, Networking, Storage, and Analysis (SC), 2025
Reviewer for ACM Intl. Conf. on Knowledge Discovery and Data Mining (KDD), 2025, 26
Reviewer for IEEE Intl. Conf. on Cloud Computing (CLOUD), 2024, 2025
Reviewer for ACM Intl. Conf. on Supercomputing (ICS), 2025
Reviewer for IEEE Intl. Conf. on Advanced Networks and Telecommunications Systems (ANT), 2024
Reviewer for Intl. Conf. on Parallel Processing (ICPP), 2024
Reviewer for European Conf. on Parallel Processing (Euro-Par), 2024, 25
Reviewer for IEEE Intl. Symposium on Cluster, Cloud, and Internet Computing (CCGRID), 2023

Journal Reviewers

Reviewer for Information Sciences
Reviewer for IEEE Transactions on Systems, Man, and Cybernetics: Systems
Reviewer for IEEE Transactions on Big Data
Reviewer for IEEE Network Magazine
Reviewer for IEEE Internet of Things Journal
Reviewer for Journal of Signal Processing Systems
Reviewer for Digital Communications and Networks

Recent Talks and Presentations

"Scientific Computation and AI at Advanced Photon source",
– invited talk at the 6th annual GRP Workshop (6GRP)/IEEE International Conference on e-Science 2025, Sept. 25, Chicago, IL, USA

"AI-Driven Approaches for X-ray Diffraction: Rapid Plastic Deformation Detection and Scalable ML Infrastructure",
– invited talk at the APS Scientific Computation Seminar Series at Argonne, Mar. 25, Online

"Model and Data Management for Machine Learning (M2ML): Integrating Instruments, Edge and HPC for Accelerated Machine Learning",
– presentation at the 5th International Workshop on Big Data & AI Tools, Models, and Use Cases for Innovative Scientific Discovery (BTSD) at IEEE BigData 24, Dec. 24, Washington DC, USA

"Model and Data Management for Machine Learning (M2ML): Integrating Instruments, Edge and HPC for Accelerated Machine Learning",
– presentation at the XLOOP workshop at SC 24, Nov. 24, Atlanta, GA, USA

"Rapid Detection of Rare Events From In Situ X-ray Diffraction Data Using Machine Learning",
– invited talk at the Diffraction-based Microstructure Imaging Workshop at the APS / CNM Users Meeting 2024, May. 24, Lemont, IL, USA

"Anomaly Detection for High Energy Diffraction Methods (HEDM) and the Coordination between Experiment and HPC Facilities",
– invited talk at the Research Experiences for Undergraduates (REU) site named BigDataX: From theory to practice in Big Data computing at eXtreme scales, July. 24, Lemont, IL, USA

"HPC Accelerates", guest lecture at the University of North Texas (UNT), Mar. 23, Denton, TX, USA

"Scalable Parallel Machine Learning on High-Performance Computing Systems - Clustering and Reinforcement Learning",
– invited talk at the CS Seminar Series at the Argonne National Laboratory, Mar. 23, Lemont, IL, USA

Last updated: March 27, 2026