Mingkai Zheng

Email: zhengmingkai616@gmail.com | Phone: +1 (646) 784-2461 | LinkedIn

EDUCATION BACKGROUND

Rutgers University, New Brunswick, NJ 01/2024 – 12/2027
(Anticipated)
Ph.D. in Electrical and Computer Engineering - High Performance Computing (HPC)

Cornell University, Ithaca, NY 08/2021 – 12/2022
Master of Engineering in Electrical and Computer Engineering

University of Liverpool (UoL), Liverpool, UK 09/2016 – 07/2020
Bachelor of Engineering in Electrical Engineering

SKILLS

Programming: Python, Java, C/C++, HTML, CSS, JavaScript, TypeScript, shell, MATLAB, SystemVerilog HDL
Tools and Frameworks: SLURM, PyTorch, **pthread**, **OpenMP**, **MPI**, **CUDA**, **Triton**, Git, Docker, Apptainer

WORK EXPERIENCE

Research Scientist Intern, Infra Team, GenBio AI, Palo Alto, CA 05/2025 – 01/2026
Research Engineer, Texas Advanced Computing Center (TACC), Austin, TX 01/2023 – 01/2024

PROJECT EXPERIENCES

SCAPE: Accurate and Efficient LLM Training with Extremely Sparse Communication [Under Review]
Project Lead, Rutgers University 11/2025 – 05/2026

- Designed SCAPE, a sparse-communication distributed optimizer for LLM pre-training that sparsifies AdamS first-momentum instead of raw gradients, supporting **90%** and **99%** sparsity without quality loss
- Built partitioned momentum-derived top-k mask refresh, delayed mask reuse, and asynchronous mask synchronization to overlap sparse communication with computation in **Megatron-LM**
- Optimized sharded training with single-buffer second-moment reconstruction, CPU-offloaded error-feedback buffers, bit-packed masks, and Triton kernels for memory-efficient sparse synchronization
- Supports both Data Parallelism (DP) and DP with sharded model and optimizer in Megatron-LM
- Reduced Llama-500M end-to-end training time by 43.3% (**1.76×** throughput) and improved Llama-1.8B per-step training throughput by up to **3.26×** on 64 NVIDIA GH200 GPUs connected by InfiniBand with improved train loss and minimal degradation on evaluation scores of downstream tasks

Radius: Accelerate LLM Pre-training by Leveraging Gradients' Temporal Locality [Accepted by MLSys25]
Individual project, Rutgers University 01/2024 – 11/2024

- Designed a new top*k*-based gradient sparsification algorithm, Radius, to reduce the communication overhead caused by using data parallel to pre-train Large Language Models (LLMs)
- Implemented Radius based on **Megatron-LM** framework and optimized Radius's memory overhead
- Empirically verified the convergence and tested the convergence speed of Radius by pre-training GPT-355M and GPT-2.0B with OpenWebText dataset using different sparsity rates on Perlmutter supercomputer
- When pre-training GPT-2.0B on 64 NVIDIA A100-80GB GPUs, Radius can provide **1.9×** per-step speedup and **1.2×** overall speedup on wall-clock time without loss of performance on downstream tasks

TACC-GPT (Now hpcGPT) [Accepted by PEARC26, Best Student Paper Award]
Team lead, Texas Advanced Computing Center (TACC) 04/2023 – Present

- Trained and deployed TACC-GPT, a ChatGPT-like model, to answer tickets submitted by TACC users
- Used and cleaned the history tickets and replies maintained in TACC's database. More specifically, tickets with no useful info are removed, and the greeting line (e.g., Hi/Hello), closing line (e.g., Best/Best regards) and signature in each email are cleaned to avoid user and TACC staff confidential information leakage
- Fine-tuned llama-3.1 8B using **ZeRO stage 3** and **DeepSpeed-Ulysses** with the cleaned tickets on VISTA
- The fine-tuned models have been deployed on Cyclone to collect rankings for training reward model (link)

Short/Long-Term Prediction for Buildings Power Load on Cornell Campus

Master of Engineering Degree Project, Cornell University 03/2022 – 05/2022

- Used real data collected and maintained by Cornell University and performed **data-cleaning** before the training
- Utilized **LSTM model** in **PyTorch** with **Flower** federated learning framework to simulate the training process
- Performed **hyper-parameters tuning** to improve the performance while reducing the training time
- Realized both 1h and 24h ahead prediction and achieved **accuracies over 90%** on min-max normalized data

PUBLICATIONS

1. J. Chen, D. Dong, H. Di, H. Jia, J. Wu, H. Xie, **M. Zheng**, Y. Li, L. Chen, H. Wang, E. P. Xing, and H. Wang, "Broken Symmetry in BF16 Attention: Why FlashAttention Gradients Blow Up Late in Training," 2026, arXiv:2609.34272. [Online]. Available: https://arxiv.org/abs/2609.34272
2. **M. Zheng**, H. Xie, J. Chen, and Z. Zhang, "SCAPE: Accurate and Efficient LLM Training with Extreme Sparse Communication", [Online]. Available: https://arxiv.org/abs/2607.01678
3. H. Xie, J. Chen, **M. Zheng**, and Z. Zhang, "PHOENIX: Resilient LLM Training with Hot-Swapping via Zero-Overhead Checkpoint", [Online]. Available: https://arxiv.org/html/2607.01646v2
4. **M. Zheng**, F. Linghu, S. Li, M. Kandes, N. Gaffney, I. Foster, and Z. Zhang, "A Comparative Study on LLM-enabled HPC User Support", [Online]. Available: https://dl.acm.org/doi/10.1145/3785462.3815796
5. M. Mogilevsky, H. Zaky, Y. Sun, **M. Zheng**, L. Yang, and Z. Zhang, "Quantifying Performance Variability in GPU Clusters," *IEEE Transactions on Parallel & Distributed Systems*, vol. 37, no. 06, pp. 1400–1411, Jun. 2026. [Online]. Available: https: //doi.ieeecomputersociety.org/10.1109/TPDS.2026.3684387
6. **M. Zheng** and Z. Zhang, "Radius: Range-based Gradient Sparsity for Large Foundation Model Pre-training," in *Proceedings of Machine Learning and Systems*, M. Zaharia, G. Joshi, and Y. Lin, Eds., vol. 7. MLSys, 2025. [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2025/file/54dd9e0cff6d9214e20d97eb2a3bae49-Paper-Conference.pdf
7. B. Zhang, **M. Zheng**, D. Ganguly, X. Zhang, V. Singh, V. Chaudhary, and Z. Zhang, "Efficient Fine-Grained GPU Performance Modeling for Distributed Deep Learning of LLM," in *2025 IEEE 32nd International Conference on High Performance Computing, Data, and Analytics (HiPC)*, 2025, pp. 22–32. [Online]. Available: https://doi.org/10.1109/HiPC66333.2025.00012
8. B. Sun, W. Liu, J. G. Pauloski, J. Tian, J. Jia, D. Wang, B. Zhang, **M. Zheng**, S. Di, S. Jin, Z. Zhang, X. Yu, K. A. Iskra, P. Beckman, G. Tan, and D. Tao, "COMPSO: Optimizing Gradient Compression for Distributed Training with Second-Order Optimizers," in *Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, ser. PPoPP '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 212–224. [Online]. Available: https://doi.org/10.1145/3710848.3710852