Curriculum Vitae - Shiyang Chen

Ph.D. Student
Rutgers, The State University of New Jersey,
New Brunswick, NJ 08854

978-856-9583
shiyang.chen@rutgers.edu
GitHub homepage

EDUCATION

2023 - Present
Ph.D. Student in **High-Performance Computing**
Department of Electrical & Computer Engineering
Rutgers, The State University of New Jersey
Advisor: Hang Liu

2020 - 2022
Master Student in **High-Performance Computing**
Department of Electrical & Computer Engineering
Stevens Institute of Technology
Advisor: Hang Liu

2015 - 2019
B.E. in **Electric Science and Technology**
School of Optical and Electronic Information
Huazhong University of Science & Technology

EXPERIENCES

2023 - Present
Graduate Research Assistant
Rutgers, The State University of New Jersey
Adviser: Hang Liu

Summer 2020
Research Intern
Lawrence Livermore National Laboratory
Mentor: Chunhua Liao
Project: **Monte Carlo sampling with emerging hardware.**

- Developed a GPU-accelerated Monte Carlo sampling framework, leveraging ray tracing cores and associated acceleration structures on NVIDIA GPUs.
- Reformulated sampling as a rendering problem, utilizing OptiX framework, Tensor Cores, and CUDA kernels to build and preprocess data structures.
- Achieved significant reduced turn-around time in quantum mechanism simulations.

2022.6-12
Applied Scientist Intern
Amazon Web Service
Mentor: Da Zheng
Project: **Online inference for temporal Graph Neural Network.**

- Built the system for low-latency, staleness-bounded temporal GNN inference on distributed, large-scale graphs.
- Designed a staleness-bounded request scheduling algorithm, enabling efficient real-time GNN inference with dynamic graph updates.

- Delivered the system, making it the first deployed solution supporting online temporal GNN inference in production.

Summer 2023 **Applied Scientist Intern**
Amazon Web Service
Mentor: Xiang Song
Project: **Large-scale Graph Neural Network inference.**

- Developed an end-to-end large-scale GNN inference system, optimizing both graph construction and node embedding computation.
- Designed GPU kernels for efficient graph processing and GNN computations across distributed machines, scaling to up to 48 EC2 instances.
- The system was deployed in production and integrated into the open-sourced GraphStorm project. It reduces end-to-end processing time from days to under an hour.

2023 - 2024 **External Collabrator**
Microsoft Deepspeed
Mentor: Shuaiwen Leon Song
Project: **Deepspeed AI for Science. Large Language Model inference.**

- Extended the DeepSpeed framework for AI-for-science workloads, optimizing models like OpenFold and AI2BMD with customized GPU kernels.
- Developed GPU kernels and system optimizations for DeepSpeed-Inference, especially for model quantization.

Summer 2024 **Software Engineer Intern**
ByteDance
Mentor: Wei Xu
Project: **Disaggregated LLM inference framework.**

- Designed a PoC system based on vLLM to disaggregate prefill and decode stages in LLM inference.
- Developed a high-throughput, low-overhead KV cache transfer network library, enabling efficient distributed inference.
- Achieved near-linear latency reduction as the system scaled across multiple nodes.

RESEARCH - PUBLICATIONS

2021 **Shiyang Chen**, Shaoyi Huang, Santosh Pandey, Bingbing Li, Guang Gao, Long Zheng, Caiwen Ding and Hang Liu. "E.T.: Rethinking Transformer Models on GPUs". In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (**SC**)*. ACM, 2021.

2021 Hongwu Peng, **Shiyang Chen**, Zhepeng Wang, Junhuan Yang, Scott Weitze, Tong Geng, Ang Li, Jinbo Bi, Minghu Song, Weiwen Jiang, Hang Liu and Caiwen Ding. "Optimizing FPGA-based Accelerator Design for Large-Scale Molecular Similarity Search". In *Proceedings of the International Conference On Computer Aided Design (**ICCAD**)*. IEEE/ACM, 2021.

2021

Shaoyi Huang, **Shiyang Chen**, Hongwu Peng, Daniel Manu, Zhenglun Kong, Geng Yuan, Lei Yang, Shusen Wang, Hang Liu and Caiwen Ding. "HMC-TRAN: A Tensor-core Inspired Hierarchical Model Compression for Transformer-based DNNs on GPU". In *Proceedings of Proceedings of the 2021 on Great Lakes Symposium on VLSI (**GLVLSI**)*. ACM, 2021.

2022 Shaoyi Huang, Dongkuan Xu, Ian Yen, Yijue Wang, Sung-En Chang, Bingbing Li, **Shiyang Chen**, Mimi Xie, Sanguthevar Rajasekaran, Hang Liu and Caiwen Ding. "Sparse Progressive Distillation: Resolving Overfitting under Pretrain-and-Finetune Paradigm". In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (**ACL**)*. ACL, 2022.

2022 Hongwu Peng, Shaoyi Huang, **Shiyang Chen**, Bingbing Li, Tong Geng, Ang Li, Weiwen Jiang, Wujie Wen, Jinbo Bi, Hang Liu and Caiwen Ding. "A Length Adaptive Algorithm-Hardware Co-design of Transformer on FPGA Through Sparse Attention and Dynamic Pipelining". In *2022 59th ACM/IEEE Design Automation Conference (**DAC**)*. IEEE, 2022.

2023 Yifei Wang, **Shiyang Chen**, Guobin Chen, Ethan Shurberg, Hang Liu, Pengyu Hong. "Motif-Based Graph Representation Learning with Application to Chemical Molecules". In *Informatics Vol. 10. No. 1*. MDPI, 2023.

2023 **Shiyang Chen**, Da Zheng, Caiwen Ding, Chengying Huan, Yuede Ji and Hang Liu. "Tango: rethinking quantization for graph neural network training on GPUs". In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (**SC**)*. ACM, 2023.

2023 Wang Feng, **Shiyang Chen**, Hang Liu and Yuede Ji. "PeeK: A Prune-Centric Approach for K Shortest Path Computation". In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (**SC**)*. ACM, 2023.

2023 Shuaiwen Song, Bonnie Kruft, Minjia Zhang, Conglong Li, **Shiyang Chen**, Chengming Zhang, Masahiro Tanaka, Xiaoxia Wu, Mohammed AlQuraishi, Gustaf Ahdritz, Christina Floristean, Rick Stevens, Venkatram Vishwanath, Arvind Ramanathan, Sam Foreman, Kyle Hippe, Prasanna Balaprakash, Yuxiong He. "DeepSpeed4Science Initiative: Enabling Large-Scale Scientific Discovery through Sophisticated AI System Technologies". In ***NeurIPS** 2023 AI for Science Workshop.*

2023 Xiaoxia Wu, Haojun Xia, Stephen Youn, Zhen Zheng, **Shiyang Chen**, Arash Bakhtiari, Michael Wyatt, Reza Yazdani Aminabadi, Yuxiong He, Olatunji Ruwase, Leon Song, Zhewei Yao. "ZeroQuant(4+2): Redefining LLMs Quantization with a New FP6-Centric Strategy for Diverse Generative Tasks". *Arxiv Preprint.*

2024 Chengying Huan, Yongchao Liu, Heng Zhang, Shuaiwen Song, Santosh Pandey, **Shiyang Chen**, Xiangfei Fang, Yue Jin, Baptiste Lepers, Yanjun Wu, Hang Liu. "TEA+: A Novel Temporal Graph Random Walk Engine With Hybrid Storage Architecture". In *ACM Transactions on Architecture and Code Optimization (**TACO**)*.

2024 Haojun Xia, Zhen Zheng, Xiaoxia Wu, **Shiyang Chen**, Zhewei Yao, Stephen Youn, Arash Bakhtiari, Michael Wyatt, Donglin Zhuang, Zhongzhu Zhou , Olatunji Ruwase, Yuxiong He, Shuaiwen Leon Song. "FP6-LLM: Efficiently Serving Large Language Models Through FP6-Centric Algorithm-System Co-Design". In *2024 USENIX Annual Technical Conference (**ATC**)*, USENIX, 2024.

2024 Ahdritz, Gustaf, et al. "OpenFold: Retraining AlphaFold2 yields new insights into its

learning mechanisms and capacity for generalization." *Nature Methods*: 1-11.

2024 Hongwei Chen, **Shiyang Chen**, Joshua J. Turner, Adrian Feiguin. "Kernel fusion in atomistic spin dynamics simulations on Nvidia GPUs using tensor core." *Journal of Computational Science* (2024): 102357.