

Xi Xie

Homepage: xiexi51.github.io

Github: github.com/xiexi51

Email: xi.xie@uconn.edu

Mobile: (860)-771-9769

EXPERIENCE

- University of Connecticut Storrs, CT, USA
Ph.D. in Computer Science and Engineering; GPA: 4.0/4.0 *Aug. 2022 - Present*
Research Interests: GPU systems, CUDA kernel design, privacy-preserving machine learning.
- Beijing Yuxing Software Co., Ltd. Beijing, China
Software Engineer *Oct. 2020 - Jun. 2022*
Developed core systems for digital avatar technology.
- Institute of Geophysics, China Earthquake Administration Beijing, China
Master's Degree in Geophysics *Sep. 2015 - Oct. 2020*
Research Interests: Seismology machine learning, aftershock detection.
- China Earthquake Networks Center Beijing, China
Software Engineer *Aug. 2013 - Oct. 2017*
Conducted automated statistical analysis of earthquake precursors.
- Beijing University of Technology Beijing, China
Bachelor's Degree in Software Engineering *Sep. 2009 - Jul. 2013*

RESEARCH PROJECTS

- **Research Assistant, University of Connecticut** **Aug. 2022 - Present**
 - Designed CUDA kernels for neural network acceleration.
 - Improved SpMM kernel design for graph neural networks (GNNs), utilizing lightweight graph preprocessing, block-level partitioning, and a combined warp strategy, achieving an average of $1.17\times$ speedup over cuSPARSE.
 - Designed novel variants of SpMM kernels that accelerate row-wise balanced sparsity injected into the right-hand matrix during the SpMM operation. These kernels can achieve up to a $10.6\times$ speedup over the original SpMM operation, resulting in a $3.5\times$ speedup in GNN training without any accuracy degradation.
 - Developed a lightweight C++/CUDA SpMM testing framework.
 - Received over 100 stars on my two GitHub repositories for the above implementations:
 - github.com/xiexi51/ICCAD-Accel-GCN
 - github.com/xiexi51/MaxK-GNN
 - (Ongoing) Developing an ultra-fast row-wise top-k kernel design based on binary search, optimized for parallel top-k selection on large batches of limited vectors. This design achieves a $3.9\times$ speedup over PyTorch (SOTA) with a precision of $1e-4$.
 - Publications: 23' ICCAD^[2], 24' ASPLOS^[1], arXiv preprint^[3].
 - Fully/Partial Polynomial Model Design for Privacy-Preserving Computation Acceleration
 - (Ongoing) Propose PolyNorm, which provides strong numerical constraints on data flow and supports fully polynomial replacement for deep neural networks. It can achieve highly stable and accurate training on large-scale datasets such as ImageNet.
 - Designed pixel-wise partial polynomial replacement methods for deep neural networks. By utilizing a smoothing loss function for thresholding, the replacement pattern is automatically selected, achieving both high replacement ratios and high accuracy.
 - Publications: 24' ICCAD^[4], 23' ICCV^[5], 23' AAAI workshop^[6]

SKILLS

- **Programming language:** CUDA; C/C++; Python; ASM; Verilog.
- **Software:** PyTorch; TensorFlow; Vivado (HLS); MATLAB; Bash scripting; Compiling chain.
- **Hardware:** CUDA kernel design; FPGA kernel design; Compiling chain.

HONORS AND AWARDS

- **Eversource Fellowship** by Eversource Energy Center **08/2024**
- **Cigna Fellowship** by UConn School of Engineering **02/2023**
- **1st place in accuracy and 4th place overall**, as team member, ACM/IEEE TinyML Design Contest **11/2022**
- **Cigna Fellowship** by UConn School of Engineering **08/2022**
- **1st prize in finals**, as executive team leader, Zhixin Cup National AI Robot Competition, hosted by CAAI **12/2021**
- **Semi-finals**, Aftershock Detection Artificial-Intelligence Contest, hosted by IGPCEA & Alibaba Cloud **07/2017**
- **2nd prize in finals**, Blue Bridge Cup Programming Contest, hosted by MIITEC **05/2012**
- **1st prize in semi-finals**, Blue Bridge Cup Programming Contest, hosted by MIITEC **05/2011**

PUBLICATIONS

1. [24'ASPLOS] **X. Xie***, H. Peng*, K. Shivdikar, M. A. Hasan, J. Zhao, S. Huang, O. Khan, D. Kaeli, C. Ding. MaxK-GNN: Towards Theoretical Speed Limits for Accelerating Graph Neural Networks Training. 2024 ACM International Conference on Architectural Support for Programming Languages and Operating Systems.
2. [23'ICCAD] **X. Xie**, H. Peng, M. A. Hasan, S. Huang, J. Zhao, H. Fang, W. Zhang, T. Geng, O. Khan, C. Ding. Accel-GCN: High-Performance GPU Accelerator Design for Graph Convolution Networks. 2023 IEEE/ACM International Conference On Computer-Aided Design.
3. [24'arXiv] **X. Xie**, Y. Luo, H. Peng, C. Ding. RTop-K: Ultra-Fast Row-Wise Top-K Algorithm and GPU Implementation for Neural Networks. arXiv preprint arXiv:2409.00822, 2024. *Not publicly available for now, can be accessed through the following link.* (<https://drive.google.com/file/d/1djHgwro2sXkHfj5k8Fn82XiBnlU5JlcV/view?usp=sharing>)
4. [24'ICCAD] T. Zhou, J. Zhao, Y. Luo, **X. Xie**, W. Wen, C. Ding, X. Xu. AdaPI: Facilitating DNN Model Adaptivity for Efficient Private Inference in Edge Computing. 2024 IEEE/ACM International Conference on Computer-Aided Design.
5. [23'ICCV] H. Peng, S. Huang, T. Zhou, Y. Luo, C. Wang, Z. Wang, J. Zhao, **X. Xie**, A. Li, T. Geng, K. Mahmood, W. Wen, X. Xu, C. Ding. AutoReP: Automatic ReLU Replacement for Fast Private Network Inference. 2023 International Conference on Computer Vision.
6. [23'AAAI workshop] H. Peng, S. Zhou, Y. Luo, N. Xu, S. Duan, R. Ran, J. Zhao, S. Huang, **X. Xie**, C. Wang, T. Geng, W. Wen, X. Xu, C. Ding. RRNet: Towards ReLU-Reduced Neural Network for Two-party Computation Based Private Inference. 2023 AAAI Workshop on DL-Hardware Co-Design for AI Acceleration.
7. [23'arXiv] K. Thorat, J. Zhao, Y. Liu, H. Peng, **X. Xie**, B. Lei, J. Zhang, C. Ding. Advanced Language Model-Driven Verilog Development: Enhancing Power, Performance, and Area Optimization in Code Synthesis. arXiv preprint arXiv:2312.01022, 2023.
8. [Master's Thesis] Use TensorFlow to implement an automatic phase picking method based on the nearest neighbor method, 2020.

PROFESSIONAL ACTIVITIES

- Reviewer for Conferences/ Journals
 - Great Lakes Symposium on VLSI 2024 (Program Committee)
 - Alexandria Engineering Journal
 - Journal of Organizational and End User Computing
 - Jordanian Journal of Computers and Information Technology
 - Journal of Systems Architecture