Zhaodong Chen

805-722-6987 chenzd15thu@gmail.com google scholar github.com/apuaaChen

Education

UC Santa Barbara, Electrical & Computer Engineering **Sep. 2019 – Mar. 2024**
Doctor of Philosophy *GPA: 3.97/4.00*

UC Santa Barbara, Electrical & Computer Engineering **Sep. 2019 – Dec. 2021**
Master of Science *GPA: 3.97/4.00*

Tsinghua Uiversity, School of Economics and Management **Sep. 2016 – July 2019**
Bachelor of Management (Second Bachelor's Degree) *GPA: 3.53/4.00*

Tsinghua Uiversity, Department of Precision Instrument **Sep. 2015 – July 2019**
Bachelor of Engineering *GPA: 3.67/4.00, Rank: 7/45*

Technical Skills

Languages: CUDA C++/PTX, C++, Python, Java, AMDGCN Assembly, Latex, Verilog
Technologies/Frameworks: PyTorch, Accel-SIM, MLIR, TVM, Nsight Compute/System

Experience

NVIDIA Corporation **May 2024 – Present**
Deep Learning Library Performance Software Engineer *Austin, TX*

- Develop high performance kernels for NVIDIA CUTLASS library.

UC Santa Barbara **Sep. 2019 – Mar. 2024**
Graduate Student Researcher *Santa Barbara, CA*

- Conduct research and publish paper in the area of GPU and Deep Learning, design and develop compilers and high-performance CUDA kernels that accelerate Deep Learning on GPU.

NVIDIA Corporation **June 2023 – Sep. 2023**
Deep Learning Library Performance Software Engineer Intern *Santa Clara, CA*

- Developed the Epilogue Visitor Tree on NVIDIA Ampere GPU under the CUTLASS 3.0 style. Achieved better performance than original CUTLASS 2.x epilogues written by experts. https://github.com/NVIDIA/cutlass/tree/main/include/cutlass/epilogue/threadblock/fusion
- Developed a Just-In-Time compiler (https://github.com/NVIDIA/cutlass/tree/main/python/cutlass/backend/evt) to automatically construct fused kernels for NVIDIA Ampere & Hopper GPUs.
- Adopted by multiple internal and external users of CUTLASS.

NVIDIA Corporation **June 2022 – Sep. 2022**
Deep Learning Library Performance Software Engineer Intern *Santa Clara, CA*

- Developed the CUTLASS Python Interface (https://github.com/NVIDIA/cutlass/tree/main/python) from scratch. Enabled specifying, emitting, compiling, launching, and profiling CUTLASS kernels from Python with client-provided tensor types.
- Proposed the CUTLASS Epilogue Visitor Tree for more flexible CUTLASS epilogue fusion.

Cloud9 Technology **June 2021 – Sep. 2021**
Operation Intern *Belmont, CA*

- Developed state-of-the-art Ethash mining kernel on AMD Vega GPU from scratch with AMDGCN Assembly. Reduced the power consumption by 1.48% under the same hashrate.

UC Santa Barbara **March 2020 – June 2020**
Teaching Assistant of ECE 152A (Digital Design Principle) *Santa Barbara, CA*

- Led lab sessions, held office hour, designed and graded lab, homework, exams.

First-Author Publications

[**ASPLOS'24**] Zhaodong Chen, Andrew Kerr, Richard Cai, Jack Kosaian, Haicheng Wu, Yufei Ding, and Yuan Xie. Evt: Accelerating deep learning training with epilogue visitor tree. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2024 (code)

[**PPoPP'23**] Zhaodong Chen, Zheng Qu, Yuying Quan, Liu Liu, Yufei Ding, and Yuan Xie. Dynamic n: M fine-grained structured sparse attention mechanism. In *Proceedings of the 28th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, pages 369–379, 2023 (link, code)

[**SC'21**] Zhaodong Chen, Zheng Qu, Liu Liu, Yufei Ding, and Yuan Xie. Efficient tensor core-based gpu kernels for structured sparsity under reduced precision. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–14, 2021 (link, code)

[**ICCAD'20**] Zhaodong Chen, Mingyu Yan, Maohua Zhu, Lei Deng, Guoqi Li, Shuangchen Li, and Yuan Xie. fusegnn: Accelerating graph convolutional neural network training on gpgpu. In *Proceedings of the 39th International Conference on Computer-Aided Design*, pages 1–9, 2020 (link, code)

[**IEEE TPAMI**] Zhaodong Chen, Lei Deng, Bangyan Wang, Guoqi Li, and Yuan Xie. A comprehensive and modularized statistical framework for gradient norm equality in deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):13–31, 2020 (link, code)

[**IEEE TNNLS**] Zhaodong Chen, Lei Deng, Guoqi Li, Jiawei Sun, Xing Hu, Ling Liang, Yufei Ding, and Yuan Xie. Effective and efficient batch normalization using a few uncorrelated data for statistics estimation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):348–362, 2020 (link)

Co-Author Publications

[**IEEE TC**] Liu Liu, Zheng Qu, Zhaodong Chen, Fengbin Tu, Yufei Ding, and Yuan Xie. Dynamic sparse attention for scalable transformer acceleration. *IEEE Transactions on Computers*, 71(12):3165–3178, 2022 (link)

[**DATE'22**] Ling Liang, Zhaodong Chen, Lei Deng, Fengbin Tu, Guoqi Li, and Yuan Xie. Accelerating spatiotemporal supervised training of large-scale spiking neural networks on gpu. In *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 658–663. IEEE, 2022 (link)

[**ASPLOS'22**] Zheng Qu, Liu Liu, Fengbin Tu, Zhaodong Chen, Yufei Ding, and Yuan Xie. Dota: detect and omit weak attentions for scalable transformer acceleration. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 14–26, 2022 (link)

[**ATC'22**] Boyuan Feng, Tianqi Tang, Yuke Wang, Zhaodong Chen, Zheng Wang, Shu Yang, Yuan Xie, and Yufei Ding. Faith: An efficient framework for transformer verification on {GPUs}. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)*, pages 167–182, 2022 (link)

[**IEEE TCAD**] Ling Liang, Zheng Qu, Zhaodong Chen, Fengbin Tu, Yujie Wu, Lei Deng, Guoqi Li, Peng Li, and Yuan Xie. H2learn: High-efficiency learning accelerator for high-accuracy spiking neural networks. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(11):4782–4796, 2021 (link)

[**ICML'20**] Liu Liu, Lei Deng, Zhaodong Chen, Yuke Wang, Shuangchen Li, Jingwei Zhang, Yihua Yang, Zhenyu Gu, Yufei Ding, and Yuan Xie. Boosting deep neural network efficiency with dual-module inference. In *International Conference on Machine Learning*, pages 6205–6215. PMLR, 2020 (link)

[**IEEE CAL**] Mingyu Yan, Zhaodong Chen, Lei Deng, Xiaochun Ye, Zhimin Zhang, Dongrui Fan, and Yuan Xie. Characterizing and understanding gcns on gpu. *IEEE Computer Architecture Letters*, 19(1):22–25, 2020 (link)

[**CVPR'19 Oral**] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 72–81, 2019 (link, code, video)

Talks

[**GTC'22**] Zhaodong Chen and Zheng Qu. Accelerating structured sparse attention with tensor core. In *GPU Technology Conference*, 2022 (link)