Shaojie Xiang

xiang.elec@gmail.com | github.com/hecmay | linkedin.com/in/shaojie-xiang

Education

Cornell University Aug 2018 - May 2024
Ph.D. in Electrical and Computer Engineering
Research Area: AI compiler, FPGA/GPU accelerators
Committee: Zhiru Zhang, Christina Delimitrou, Adrian Sampson

Huazhong University of Science and Technology Sep 2014 - May 2018
B.Eng. in Electrical Engineering
GPA: 4.98/5.0 ranking 1/423

Work Experiences

AWS AI Research, Applied Scientist Intern May-Aug 2023
Optimizing Large Language Models (LLM) on GPUs and Trainium

- Implemented block-sparse flash-attention kernels for GPUs and AWS Trainium using Triton/Neuron SDK, improved attention kernel inference time by 3-5x

Nvidia Machine Learning Compiler Group, SDE Intern Sep-Dec 2020
Automatic Scheduling and Optimizing Neural Networks on GPUs

- Developed a heuristic optimizer based on TVM/Ansor that automatically schedules the tensor programs to achieve better performance on GPUs

Intel Parallel Computing Lab, Research Intern May-Aug 2020
AI Compiler for Spatial FPGA Accelerators

- Designed and implemented a domain-specific programming language and compiler to accelerate ML workloads on FPGA with systolic array

Selected Publications

Shaojie Xiang, Yi-Hsiang Lai, Yuan Zhou, Hongzheng Chen, Niansong Zhang, Debjit Pal, Zhiru Zhang. *HeteroFlow: An Accelerator Programming Model with Decoupled Data Placement for Software-Defined FPGAs*. ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA) 2022

Yi-Hsiang Lai, Ecenur Ustun, Shaojie Xiang, Zhenman Fang, Hongbo Rong, and Zhiru Zhang. *Programming and Synthesis for Software-Defined FPGA Acceleration: Status and Future Prospects*. ACM Transactions on Reconfigurable Technology and Systems (TRETS) 2021

Nikita Lazarev, Shaojie Xiang, Neil Adit, Zhiru Zhang, Christina Delimitrou. *Dagger: Efficient and Fast RPCs for Cloud Microservices with Near-Memory Reconfigurable NICs*. 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS) 2021

Ecenur Ustun, Shaojie Xiang, Jinny Gui, Cunxi Yu, Zhiru Zhang. *LAMDA: Learning-Assisted Multi-stage Autotuning for FPGA Design Closure*. 27th International Symposium on Field-Programmable Custom Computing Machines (FCCM) 2019

Awards

Christen Fellowship, Cornell University 2018
National Scholarship, Ministry of Education, China 2015/2016/2017

Skills

Python, C/C++, Rust, CUDA, OpenCL, Verilog
PyTorch, TensorFlow, DeepSpeed, Triton, TVM, LLVM/MLIR, Vitis HLS