Lixun **Zhang**

☐ (972) 804-9382 | ✉ lixun.zhang@amd.com | ⌨ https://github.com/zhanglx13

Education

The University of Texas at Austin   *Austin, Texas*

PH.D. IN COMPUTER SCIENCE   *Jan. 2013 - May. 2022*

- Dissertation: A real-time throughput model based particle filter program generator on GPU: a real-time analysis

Tsinghua University   *Beijing, China*

B.S. IN ENGINEERING, AUTOMATION   *Sep. 2008 - Jul. 2012*

Work Experience

Advanced Micro Devices, Inc.   *Austin, Texas*

COMPILER ENGINEER   *May. 2022 - current*

- Supporting OpenAI Triton on AMD GPUs.
- Studied the factors affecting verification errors for convolution end-to-end tests with fp16 data type.
- Improved CI infrastructure for multi-arch testing.
- Developed an automation framework to generate convolution end-to-end tests.

Mathworks   *Natick, MA*

COMPILER ENGINEER INTERN   *Jun. 2016 - Aug. 2016*

- Developed a C++ pass to analyze for loops in the IR to find loop invariant variables.
- Generated intrinsic functions which expand to OpenMP pragmas in the code generation phase.
- Performed timing experiments of the generated code on multi-core systems.

Research Projects

Implementation of Monte Carlo Localization on Heterogeneous Systems   *May. 2021 - Aug. 2021*

- Implemented the Monte Carlo Localization algorithm in CUDA.
- Partitioned the particles among multiple threads: one launches the GPU kernel and others run on CPU in a multi-threaded manner.
- Mutex and condition variables were used to synchronize between threads.
- Simulated the localization of the f1tenth autonomous car system using ROS.

Implementation of Particle Filter on Heterogeneous Systems   *Jan. 2021 - Apr. 2021*

- Partitioned the particles among two processes: one runs on CPU and the other launches the GPU kernel.
- POSIX shared memory and System V semaphore were used to communicate and synchronize between the two processes.
- Conducted experiments on NVIDIA GTX TITAN and Jetson TX2 to obtain the optimal partition among CPU and GPU.

An Analytical Performance Model for GPGPU kernels   *Jan. 2019 - Dec. 2020*

- Developed a mathematical model to estimate the execution time of simple kernels based on static analysis of CUDA assembly code and knowledge of GPU architecture.

Laser Power Control for Selective Laser Sintering   *Sep. 2016 - Aug. 2017*

- Achieved motion detection of galvanometer by comparing two consecutive images taken by infrared cameras in LabVIEW.
- Built an automatic laser power control system at both vector-level and layer-level control granularity to eliminate thermal gradients in the post-sintering temperature on the Laser Additive Manufacturing Pilot System at the department of mechanical engineering.

A MATLAB to CUDA translator for Particle Filter Applications   *Sep. 2014 - Aug. 2015*

- Developed a front-end in C to translate MATLAB code of particle filter estimator into CUDA code.
- Developed domain-specific optimization passes in C for the generated CUDA code.

CUDA Implementation of Particle Filter for Real World Applications   *Jun. 2013 - Aug. 2014*

- Applications include Vacuum Arc Remelting and Early Kick Detection.
- Implemented sampling, importance, and resampling modules in CUDA.
- Analyzed and improved GPU kernel performance through profiling.

Presentation

University of Nebraska Collaboration Initiative   *University of Nebraska Omaha*

PRESENTER FOR <REAL-TIME THROUGHPUT MODEL FOR PARTICLE FILTER PROGRAM ON GPU>   *Nov. 2019*

Skills

Programming   C/C++, bash, CUDA, Hip, LaTeX, MATLAB, ROS, MPI, tikz, gnuplot