Shuoer Wang

Shuoer Wang | shuoer0wang@gmail.com | +1 437-341-5717

Work Experience

Huawei Technologies Canada Co., Ltd.

Compiler Engineer @ Ascend AI Processor Team Feb 2025 - Present

- Commercial Release: Ascend 910B series chips for server, mobile, and automotive.
- Working on target specific optimization for **DeepSeek** kernels on **SIMD** architecture(910D).
- Proficiency in **loop** optimizations, **memory** alias analysis, and auto **synchronization**.
- Proactively identify issues of compiler **pipeline design** and deliver solutions.
- Experience with graph-based parallelism strategies with **vLLM** and **GSPMD**.
- Experience with LLM model deployment using **TorchDynamo**, **OpenXLA** and **IREE**.

Compiler Engineer @ Maleoon GPU team July 2023 - Jan 2025

- Commercial Release: Maleoon 910 GPU supporting **OpenGL(GLSL)** and **Vulkan(SPIR-V)**
- Experience with **LLVM** at almost all levels including: Clang, MLIR, IR, ISel, TableGen, MIR.
- Familiar with ISA and hardware architecture of in-house **GPU** and **RISC-V** co-processor.
- Experience with graphics and compute API including: Vulkan, OpenGL, CUDA.
- CodeGen support for ISA evolution across **SIMT** and **RISC-V** backends.
- Optimized the thread uniform variable handling pipeline, delivered a **7% performance boost**.
- Implemented FastISel for -O0, achieved over **10% compilation time improvement**.
- Developed solutions for cloud collaborative **global illumination** and Vulkan **Task/Mesh** pipeline.
- Trained multiple **Neural Super Sampling** models using **PyTorch** with data extracted from **Unreal Engine** and tested on device using **IREE** and **ONNX**.
- Implemented high performance pipeline of **3D Gaussian Splatting** and optimize it at compiler backend.
- Proposed a set of extension functions guiding CodeGen that can boost performance for more than **5%**.
- Designed and prototyped a **new programming model** targeting both graphics and AI, including a runtime library and a compiler frontend. Leveraging a unified host-device context and modern c++ feature to optimize scheduling efficiency and streamline synchronization. This project evolved from a personal innovation initiative into a team-wide strategic objective, securing funding for additional headcount.

Compiler Engineer(Co-op Intern) @ Maleoon GPU team May 2021 - Aug 2022

- Refactored compiler pipeline for uniform analysis and lowered new instructions to RISC-V target.
- Proactively resolved issues and implemented optimizations to enhance performance and eliminate bugs.
- Developed testing tools for compiler output comparison and deployed on Jenkins system.

Personal Projects

NerF A Pytorch implementation of Instant Neural Graphics Primitives. Link to Code

LLM Fine-tuned a **T5** model converting Chinese alphabets(Pinyin) to character. Link to Code

Path Tracer A simple path tracer with only spheres with **CUDA**. Link to Code

Other Experience

- Co-designed a serious game with physicians for physiotherapy training. Supervised by Steve Engels.
- Optimized matrix-vector mutiplication as a research assistant. Supervied by Maryam Mehri Dehnavi.
- Worked as teaching assistant in the course csc207 software design at University of Toronto.

Education

2018 - 2023 Computer Science Bachelor's Degree at **University of Toronto**. (GPA: 3.92/4.0)