Weiming Hu

Email: huwm1@shanghaitech.edu.cn
Tel: 18800319117
Homepage: https://huweim.github.io/

Education

ShanghaiTech University Shanghai, China
Master in Computer Science 2020.9 - 2023.6 (Expected)

- Core Course: **Computer Architecture II** (A-), **Computer Architecture III** (A), Digtal VLSI Design Project (A), Artificial Intelligence Computer System (B+)
- GPA: 3.47/4.0. Advisor: Prof. Xin Lou, Dr. Yuanfeng Wang

Sichuan University Chengdu, Sichuan, China
B. Eng in Civil Engineering 2016.9 - 2020.6

- Rank: Top 30%.

My research interests lie in **Computer Architecture** with a focus on **GPU Micro-architecture**, Hardware/Software Co-Design and Neural Network Quantization.

Publication

Weiming Hu, Yi Zhou, Ying Quan, Yuanfeng Wang, Xin Lou, ”Cache-locality Based Adaptive Warp Scheduling for Neural Network Acceleration on GPGPUs”, *to appear in IEEE 35th International System-on-Chip Conference* (**SOCC 2022**)

Experience

Glenfly Tech Co., Ltd. (Shanghai Zhaoxin Semiconductor Co., Ltd., GPU Department) Shanghai, China
GPU Architecture R&D Intern, Core Pipeline Group 2021.8 - Now

- Assist the performance team to develop **performance analysis tools**, which used to visualize data and analyze the bottleneck by **hardware counter** of each GPC.
- Review the topic about GPU architecture and read related paper. Design a statistical mechanism to quantify **intra-warp locality** and **inter-warp locality** of application. The proposed mechanism select warp scheduling policy according to locality information.
- Implement it in GPGPU-Sim, evaluate the performance improvement under the proposed warp scheduling policy and write a manuscript.

Project

Low-bit Deep Neural Network Quantization 2022.4-2022.7

- Maintain a quantization framework built with PyTorch, and verify the accuracy of several CNN models and ViT under low-bit inference.
- Test the performance and power consumption of the quantization architecture by DRAM simulator CACTI and the GPU simulator GPGPU-Sim.
- Convert various network models into corresponding GEMMs to verify the benefits of low-bit quantization.

Convolution Kernel ASIC 2021.4-2021.6

- Design an ASIC for convolution according to paper **An Energy-Efficient Precision-Scalable ConvNet Processor in 40-nm CMOS**. Aims to reduce the number of accessing DRAM and improve the rate of **data reuse**.
- Kernel size 4 × 4, input feature map size 64 × 64, output feature map size 61 × 61. The number of channels and kernels is adjustable between 8-32.
- Implement RTL-level design with Verilog, compile it with **VCS** and synthesis with **Design Compiler**. And it passes the formal verification. [Github Link]

Awards

2021 Outstanding Administrative Assistant.
2019 First prize of Sichuan Province in National Mathematics Competition for College.
2017 Individual Second-class Scholarship.
2018 Individual First-class Scholarship.

Skill

Programming Languages: C/C++, Python.
Framework & Toolchain & Simulator: CUDA, PyTorch, Verilog, GPGPU-Sim, Accel-Sim.
Tool: Docker, Vim, GDB, LaTex.