Weiming Hu

Email: weiminghu@sjtu.edu.cn
Tel: 86-18800319117
Homepage: https://huweim.github.io/

Education

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

- First prize of Sichuan Province in National Mathematics Competition for College Students.
- 2017 Individual Second-class Scholarship. 2018 Individual First-class Scholarship.

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

- Core Course: **Computer Architecture II** (A-), **Computer Architecture III** (A), Digtal VLSI Design Project (A)

Shanghai Jiao Tong University Shanghai, China
Ph.D. in Computer Science 2023.9 - 2027.6 (Expected)

Publication

- **Weiming Hu**, Haoyan Zhang, Cong Guo, Yu Feng, Renyang Guan, Zhendong Hua, Zihan Liu, Yue Guan, Minyi Guo, Jingwen Leng, 'MANT: Efficient Low-bit Group Quantization for LLMs via Mathematically Adaptive Numerical Type', to appear in IEEE International Symposium on High-Performance Computer Architecture (**HPCA 2025**)
- Cong Guo*, Jiaming Tang*, **Weiming Hu**, Jingwen Leng, Chen Zhang, Fan Yang, Yunxin Liu, Minyi Guo, Yuhao Zhu, 'OliVe: Accelerating Large Language Models via Hardware-friendly Outlier-Victim Pair Quantization', Proceedings of the 50th Annual International Symposium on Computer Architecture (**ISCA 2023**)
- **Weiming Hu**, Yi Zhou, Ying Quan, Yuanfeng Wang, Xin Lou, 'Cache-locality Based Adaptive Warp Scheduling for Neural Network Acceleration on GPGPUs', IEEE 35th International System-on-Chip Conference (**SOCC 2022**)

Internship

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 maintain and develop **performance analysis tools**, which used to visualize data and analyze the bottleneck by **hardware counter** of each GPC.
- Research advanced GPGPU architecture, focus on warp scheduling, data management, L1 cache, etc.

Research Experience

GPGPU/GPU Architecture Design Shanghai, China

- 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.

Skill

Programming Languages: C/C++; Verilog, CUDA (novice); python, shell, Triton.
Simulator & Framework: GPGPU-Sim, Accel-sim, Sniper, GPUWattch, PyTorch.
Tool: Docker, Vim, GDB, Git, tmux.