TIAN-YU GUO 郭天宇

guoty9@mail2.sysu.edu.cn · "gty111.github.io/info" · "github.com/gty111"

Education

Sun Yat-sen University, Guangzhou, China 2022 – 2025 (expected)

Master student in Computer Science (CS)

Xidian University, Shaanxi, China 2018 – 2022

B.S. in Computer Science (CS)

Publications

- **Tianyu Guo**, Xuanteng Huang, Kan Wu, Xianwei Zhang and Nong Xiao, "SMILE: LLC-based Shared Memory Expansion to Improve GPU Thread Level Parallelism", The 61st ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, United States, June 2024.

Experience and Projects

Fine-grained cache Management in GPU 2023

To achieve accurate/fast objection recognition and detection, various machine learning models arise. After characterizing those models, we achieve fine-grained cache management to speed up the inference procedure through modifying cache control bits in the binary level before the launch of each kernel.

Shared Memory Expansion to Improve GPU TLP 2023

GPU thread level parallelism tends to be bounded by shared memory or registers. To expose the potential performance restricted by hardware resources, We purpose to harness the abundant last level cache(LLC) to be shared memory extension. Experiments in Accel-Sim shows that it can greatly improve TLP and reduce the execution time.

Optimize GEMM step by step 2023

"GEMM MMA" first implementates a naive kernel of GEMM by CUDA mma.sync and then optimize it step by step(using vectorization, asynchronous copy, conflict-free shared memory access, consolidated memory access and so on), which achieves above 70% of peak performance relative to CUTLASS in the final version.

Teaching Assistant of "SYSU-DCS3013 : Computer Architecture" 2022

Release "SYSU-ARCH LAB" which focuses on simulators(gem5, GPGPU-Sim and Accel-Sim).

Design PTX-EMU 2022

"PTX-EMU" is an simulator for NVIDIA's virtual instruction set PTX. You can use it to generate image by simulating rendering program.

Design CNN framework on CPU and GPU 2022

"CovNN" is a CNN framework(train and inference) support on CPU and GPU(built on CUDNN). To validate its availability, CNNs are built to solve MNIST or CIFAR-10 training on GPU and achieve 98% or 70% accuracy respectively.

Skills

- Programming Languages: C, C++, CUDA, Python, Java
- English: CET6(517/750)

♡ Honors and Awards

The Second Prize in ACTIC of A3 track operator implementation and performance optimization	2023
Top 4.2% nationwide | Top 2.9% worldwide in "leetcode contest"	2023
The Second Prize Scholarship in SYSU	2022-2023
The Second Prize Scholarship in XDU	2018-2019, 2019-2020, 2020-2021
Top 7.58% in CCF CSP	2021
Provincial first prize of CUMCM	2020