LIU ZIHAN (Altair)

ilovehanhan1120@hotmail.com, +86 159 0215 7531

Personal Site: http://subjectnoi.github.io/about/ *

Education

2015.09-2019.06 | Bachelor, Dept. of Computer Science, East China Normal University

- GPA: 3.81/4.00, Rank: 5/116, scholarship in 2018/2019. Awards in MCM, CCCC programming contest, etc.

2019.09-2022.03 | Master, Dept. of Computer Science and Engineering, Shanghai Jiao Tong University

- ReArch Lab, I research on compiler optimization, AI hardware and system architecture. My tutor is Prof. Jingwen Leng.
- Courses taken: Advanced Computer Architecture (A+), Matrix Theory (A), Stochastic Process (B+), Parallel Computing (A), Programming Language Theory (A), Compiler Principle (A), Computational Complexity (A-), Modern Encryption Algorithm (A-), Computer Organization (TA, 2020).

2022.03-2026.03 | PhD, Dept. of Computer Science and Engineering, Shanghai Jiao Tong University

- I will continue researching on AI system and compiler optimization under the guidance of Prof. Jingwen Leng.

Job

2022.07-2022.Now Intern (GPU IP Front-end Design Verification), **GFX MI HW, AMD**

- I participated in the design and verification of the cache sub-system in AMD GPU IP. I'm responsible for coverage report, failed-case debugging, etc. Later I will take part in the micro-architecture design.

2021.07-2022.07 Researcher, **Shanghai QiZhi Institution**

- I mainly research on the optimization of multi-tenant deep learning service from compiler and scheduler aspects for higher service capability, my work is accepted by ASPLOS 2022[3].

2020.06-2021.06 Intern (LLVM CodeGen), **IAGS, Intel**

- I participated in the development of compiler back-end optimization for next generation Intel CPU (`LLVM`), and I mainly worked on the intrinsics related to new matrix instructions.
- I participated in the research of a more intellectual profiling guided optimization (PGO) project combined with some machine learning techniques, with which we can hit a higher branch accuracy at compiling stage.

2019.01-2019.06 Intern (GPU SM Arch), **NVIDIA**

- I participated in the development of a cycle level simulator for functionality verification and performance modeling for next generation GPU circuit based on `C++, PTX, SASS`, and I mainly worked on a series of new matrix instructions.

2018.08-2019.01 Intern (Java Development), **DBS(IBSO), SAP**

- I participated in the development and deployment of `S/4 HANA` cloud applications based on `Java, SAPUI5` with corresponding tools `Spring, OData, MongoDB`, I'm also responsible for the unit and integrated test based on `karma, QUnit, Opa5`.

*Github: https://github.com/SubjectNoi

Project Experience

Compiling and optimizing tool chain for military intelligent chip.

2018.12-2020.07, ICT-CAS, Logistic Department of Central Military Commission of P.R.C., National Key Research Projects

- I implemented a full-stack compiler tool chain for Cambricon MLU-100 in `C++, Python` with `ONNX` and `TVM` [1].
- I researched and developed a series of hardware-aware graph level optimization procedures[2].

Heterogeneous cloud AI chip compiler stack

2020.09-2021.03, Montage Tech, R&D Projects

- I researched and implemented multiple front-end and back-end design schemes with communication enabled between sub-components on a heterogeneous cloud AI ASIC.
- I construct an evaluation and simulation platform based on `onnxruntime` and `TVM`.

Bachelor/Master Projects

2015.09-2022.03, ECNU/SJTU, Course Projects

- Profiling of new generation GPU with Tensor Core and corresponding AI framework optimizations.
- Implementing a C-like language compiler front-end based on `lex` and `yacc` with generated IR executed on an interpreter.
- A 3D third-person action game and graphics optimization based on Unreal Engine 4.
- ML based future spread strategies research (`LSTM, Gaussian Process`) and design with the support of Optiver™.

Skills

C/C++/Assembly, CUDA/OpenCL, TVM/LLVM, Verilog HDL/SystemVerilog/UVM, Hardware and Computer System Arch., LaTeX, Python, Java, SQL/MongoDB, Unreal Engine 4.

Interests

Computer system architecture, compiler, chip designing, game.

Publications

1. Cong Guo, Chen Zhang, Jingwen Leng, **Zihan Liu**, Fan Yang, Yunxin Liu, Minyi Guo, Yuhao Zhu. 2022. ANT: Exploiting Adaptive Numerical Data Type for Low-bit Deep Neural Network Quantization. In: 55th IEEE/ACM International Symposium on Microarchitecture (MICRO). IEEE/ACM.
2. **Zihan Liu**, Jingwen Leng, Zhihui Zhang, Quan Chen, Chao Li and Minyi Guo. 2022. VELTAIR: Towards High-Performance Multi-Tenant Deep Learning Services via Adaptive Compilation and Scheduling. In: 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), ACM, 388-401.
3. **Zihan Liu**, Jingwen Leng, Quan Chen, Chao Li, Wenli Zheng, Li Li, and Minyi Guo. 2020. DLFusion: An Auto-Tuning Compiler for Layer Fusion on Deep Neural Network Accelerator. In: 18th IEEE International Symposium on Parallel & Distributed Processing with Applications (ISPA). IEEE, 118–127.
4. **Zihan Liu**, Jingwen Leng, Guandong Lu, Chenhui Wang, Quan Chen, and Minyi Guo. 2020. Survey and design of paleozoic: a high-performance compiler tool chain for deep learning inference accelerator. CCF Trans. High Perform. Comput. 2, 4 (2020), 332–347.