Jiaming Tang

jmtang42@gmail.com / https://jiamingtang.me

Education

Shanghai Jiao Tong University — Shanghai, China
Bachelor of Computer Science — *Sept. 2020 - Present*

- Member of ACM Honors Class, which is an elite CS program for top 5% talented students
- **Avg. Score** (All): 89.78/100, Ranking: 10/36
- **Avg. Score** (Grade 2&3): 90.99/100, Ranking: 3/36
- Scores of some courses:
  - Computer Architecture: 96/100, Ranking: 1/36
  - Machine Learning: 98/100, Ranking: 1/36
  - Natural Language Processing: 98/100, Ranking: 1/36

Experience

Massachusetts Institute of Technology — Cambridge, MA, USA
Research Assistant, advised by Prof. Song Han — *Mar. 2023 - Present*
Research Topic: Efficient Algorithms and Systems for Large Language Models.

Shanghai Jiao Tong University — Shanghai, China
Undergraduate Researcher, advised by Prof. Jingwen Leng — *June. 2022 - Mar. 2023*
Research Topic: Efficient Software-Hardware Co-design for Large Language Models.

Publications

OliVe: Accelerating Large Language Models via Hardware-friendly Outlier-Victim Pair Quantization
C. Guo, **J. Tang***, W. Hu, J. Leng, C. Zhang, F. Yang, Y. Liu, M. Guo, Y. Zhu (*equal contributions)*

- Accepted by **ISCA 2023**.
- We propose a novel outlier-aware quantization method and an efficient architectural implementation, which surpasses the previous SOTA outlier-aware accelerator by **4.5× speedup and 4.0× energy reduction**.

AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration 

J. Lin, **J. Tang***, H. Tang, S. Yang, X. Dang, S. Han (*equal contributions)*

- Submitted to **MLSys 2024**.
- We propose a hardware-friendly approach for LLM low-bit weight-only quantization. This method achieve **more than 3× speedup** over the Huggingface FP16 implementation on both desktop and mobile GPUs.
- **AWQ has shown great impact in the industry and community!** It is integrated into vLLM, FastChat, TensorRT-LLM, Huggingface Transformers, and LMDeploy!

Projects

RISC-V CPU Implemented in Verilog RTL
SJTU ACM Class Computer Architecture 2021 Assignment (MS108 Course Project)
A Tomasulo RISC-V cpu with iCache and branch predictor with 2-bit saturating counter. My architecture design and implementation achieved the **top performance** in ACM Class 2020.

Compiler for Mx* Language
SJTU ACM Class Compiler Design and Implementation 2022 Assignment (MS208 Course Project)
A Compiler from Mx* language (which is a C++ & Java like language) to RV32I Assembly. I designed a new IR simplified from LLVM IR to implement my self-designed optimizations.

Superhuman Board Game AI with AlphaZero Algorithm
SJTU ACM Class Machine Learning 2022 Assignment (CS420 Course Project)
A board game AI achieved super-human performance in 13x13 Gomoku and 9x9 Go. Train only via self-play with limited resources of a single home PC. This project got the **highest score** in ACM Class 2020.

Honors & Awards

Programming Competition

- Gold Medal, The 2020 ACM-ICPC Asia Nanjing Regional Contest
- Gold Medal, The 2020 ACM-ICPC Asia Yinchuan Regional Contest
- Second Runner up, The 2020 CCPC Qinhuangdao Programming Contest

Scholarship

- 2020, 2021, 2022 Zhiyuan Honorary Scholarship (Top **2%** in Shanghai Jiao Tong University)

Other Experience

Shanghai Jiao Tong University ACM team
Co-coach *June. 2022 - June. 2023*

Principle and Practice of Computer Algorithms
Teaching Assistant *June. 2021 - Aug. 2021*

Computer Programming
Teaching Assistant *Sept. 2020 - Dec. 2020*

Technical Skills

Languages: Mandarin (native), English (TOEFL: 108).
Programming Languages: Proficient in C/C++, CUDA, Python, Java, and Verilog.