Mr. Zhiwen Mo 莫志文
Born in Hangzhou, China | +86-13067900235 | zhiwen.mo25@ic.ac.uk

EDUCATION BACKGROUND

Shanghai Jiaotong University **Bachelor in Microelectronics** 2018-2022
Average score: 91.74/100 *Ranking:1/62*
GPA:3.96/4.3 *Ranking:1/62*

Shanghai Jiaotong University **PhD in EE (Zhiyuan Honor Program, top 1%)** 2022-2024(Quit)
Research Interest: Reconfigurable Computing Architecure

Imperial College London **PhD in Computing (Full Scholarship from UKRI)** 2025-2029(Est.)
Supervised by: Dr. Hongxiang Fan, Prof. Wayne Luk
Research Interest: Domain Specific Architecture, GPU Microarchitecure, Machine Learning Systemr

RESEARCH & INTERNSHIP & Project EXPERIENCE

Microsoft Research Asia (MSRA) 2023-2025
Research Intern, System Research Group, AI Compiler Team

- Leading Project: a tile-based performance modeling tool achieving 5% error rate than of real GPU running LLMs (under review in SOSP'25, first author)
- Leading Project: LUT-based tensor core accelerates mixed-precision GEMM by an order of magnitude (ISCA'25, first author)
- Building cost models for a batch of compiler projects: PipeThreader: Software-Defined Pipelining for Efficient DNN Execution(OSDI'25); Tilelang: A tile-based programming language for GPUs (to submit in NIPS'25)

Shanghai Artificial Laboratory 2022-2023
Intern, SoC Design

- Project : Pizza, a PFlops wafer-scale chip with ultra-high inter-bandwidth
  Participated in determining the spec of the wafer-scale chip
  Analyzed Cerebras & Tesla's solution for redundancy, scale-up, scheduling, and reliability

Mapping of neural networks such as LLAMA, Bert, RepVGG on CGRA 2023

- Searched dataflow for multi-tensor fusion, enabling less DDR I/O and better performance
- Participated in the implementation of a C#-based simulator and designed a set of xml-based mapping description templates, making it easier for mapping of algorithms and quantitative analysis of performance

A Design Of Sparse Convolution Accelerator Based On Kernel Fusion 2022
Bachelor's degree thesis

- Customized a special 3*3+1 operator to fully support channel fusion of RepVGG style networks

Arm processor core-based SoC design for smart medical ward detection (Arm Cup) 2021

- *Hls-based accelerator design integrated, for gesture recognition*
- First-class prize, East China District

Publications

Zhiwen Mo, et al, "TileSight: Facilitating GPU Kernel Optimization in the LLM Era Through Tile-Centric Modeling" (under review, SOSP'25)
Zhiwen Mo, et al, "LUT Tensor Core: Lookup Table Enables Efficient Low-Bit LLM Inference Acceleration" (ISCA'25)
Yu Cheng, ……, **Zhiwen Mo,** ...., et al. "PipeThreader: Software-Defined Pipelining for Efficient DNN Execution" 19th USENIX Symposium on Operating Systems Design and Implementation (OSDI'25)
Lei Xv, **Zhiwen Mo**, et al. "Enabling Multiple Tensor-wise Operator Fusion for Transformer Models on Spatial Accelerators", Design Automation Conference, 2024 (DAC'24).

HONORS&AWARDS

UK Research and Innovation (UKRI) Fully-Funded PhD Scholarship (£ 22780 per year) 2025-2029
Chinese National Scholarship 2020-2022
A-class Shanghai Jiaotong University Scholarship(top 1%) 2020,2021
First Prize, Arm Cup, National Student IC Innovation and Entrepreneurship Competition, East China Region 2021
The 5th SenseTime Scholarship (30 people per year, selected in the field of artificial intelligence nationwide) 2021
Excellent Graduates in Shanghai 2022

ADDITIONAL INFORMATION

Proficient in: Verilog, System Verilog, Python, Cuda toolkit (Especially NCU)
Familiar with: C/C++, Linux Shell, MATLAB
English: IELTS 7.0, GRE 323(V153, Q170)+3.5

HOBBIES

Piano (with perfect pitch)
Football(center back)
Singing (chorus)
Biking (hands free)