Jiaao (Mason) MA

Ph.D. Candidate, Department of Computer Science, Duke University

jiaao.ma@duke.edu | 949-668-5668 | https://jiaaom.github.io/ | Durham, NC

Research Interests

My research interests include **Computer Architecture**, **Domain-specific Accelerators**, **Applied Cryptography**, **Deep Neural Network Compilers**, and **Privacy-preserving Machine Learning (PPML)**.

Currently, I focus on efficient software/hardware co-design for privacy-preserving computing, with the aim of improving computational efficiency and public accessibility.

I specialize in advancing **secure natural language processing (NLP)** and **data analysis** through **fully homomorphic encryption (FHE)**, focusing on optimization across **arithmetic circuits, compiler design, and hardware acceleration**.

Education

Duke University, Ph.D. Candidate in Computer Science, Application-driven Programmable Efficient Accelerated Systems (APEX) Lab - apexlab-duke.github.io — Sept 2021 - Present

- **Advisor**: Prof. Lisa Wu Wills
- **TA Experiences**: Undergraduate and Graduate-level Computer Architecture courses
- **Relevant Coursework**: Computational Complexity, Cryptography, NLP, Distributed System and Networking

University of California, Irvine, Bachelor of Science in Computer Engineering — Sept 2017 - Jun 2021

- GPA: 3.825/4.0, Latin Honor of Cum Laude

Research Experience

Ph.D. Research Assistant at Duke University – Durham, NC — Nov 2021 - Present
Advised by Prof. Lisa Wills

ML Compiler Framework for High-Performance Systolic Array Simulation

- Designed **the first compiler framework for heterogeneous systolic array** accelerators, enabling cycle-accurate evaluation of architectural and microarchitectural design choices.
- Optimizes specifically for state-of-the-art **language models** (BERT, GPT-2, GPT-Neo, etc.) and Transformer-based **U-Net models**. The modular pipeline takes PyTorch models and produces low-level hardware instructions that optionally integrate with the Beethoven framework.
- Features flexible **operations fusion** support that optimizes the distribution of workload between heterogeneous functional units based on affinity while minimizing data movement costs.
- Provides seamless integration with PyTorch's ecosystem through `TorchDynamo` and `FX`, enabling efficient model compilation and optimization for specialized accelerator architecture.

Hardware Accelerator for Multi-bit Fully Homomorphic Encryption

- Developed a specialized hardware accelerator for multi-bit Torus-FHE, achieving $\mathbf{2600}\times$ and $\mathbf{1200}\times$ **speedup** compared to CPU and GPU platforms, respectively, on real-world workloads including LLM inference, CNN inference, and regression models.
- Features the first heterogeneous FFT cluster design in FHE accelerators for fast large polynomial multiplication.
- The proposed architecture effectively addresses scaled ciphertext dimensions and excessive memory bandwidth requirements. Compared to the previous state of the art, it achieves $\mathbf{2.8}\times$ **better throughput per unit area**.

ML Compiler for Multi-bit FHE Hardware Accelerator

- Developed the **first compiler for FHE hardware accelerators** based on *Multi-Level Intermediate Representation* (MLIR) and `FHELinAlg` dialect, with integration to the `Concrete` toolchain.
- Introduces multi-level operation deduplication and data reuse that reduces up to 47.12% key-switching operations and memory requirement by 11.28`GB/s`.
- For the first time, a quantized large language model (`GPT-2`) is compiled targeting hardware accelerators, enabling **real-time privacy-preserving LLM inference**.

High-performance CUDA and Distributed CPU Execution Engine for Boolean FHE

- **The first execution engine** for Boolean FHE programs that works on both distributed CPU and GPU platforms.
- The CUDA backend shows up to **120× speedup** compared to the prior GPU executor by leveraging CUDA Graphs and improved scheduling and dependency management.
- The distributed backend shows up to **60× speedup** compared to a single-node executor, enabling high-performance distributed FHE execution for the first time.

PyTorch Neural Network to Verilog Generator

- Developed the *ChiselTorch* generator to facilitate privacy-preserving neural network model implementation.
- Ensured correctness by providing pre-built Chisel modules for common neural network layers (e.g., convolution, pooling, activation, normalization).
- Maximized performance by enforcing a data-oblivious computational model, translating computation DAGs into fused combinatorial forms to minimize operation count and enhance efficiency.

On-going Projects

- Hardware-software Co-design for Privacy-preserving *deep packet inspection* (DPI) for enterprise networking protection.
- Compiler that enables efficient large language model inference by leveraging multi-bit Torus-FHE arithmetic.

Industry Experience

ML Systems Intern at Ambarella Inc. – Santa Clara, CA May 2025 - Aug 2025

- Built the first multi-SoC framework for **LLM inference** across Ambarella **CV-** and **N-series** SoCs, combining **tensor** and **pipeline** parallelism.
- Fully compatible with Ambarella's existing **compiler toolchains**; integrates with current platforms and next-generation products.
- Transport-agnostic design operates over existing inter-SoC links and benefits from higher-bandwidth fabrics.

Publications

Mansi Choudhary, Chris Kjellqvist, *Jiaao Ma*, Lisa Wills, **A Cycle-Accurate Simulator for Heterogeneous Systolic Array Architectures** Mar 2025
2025 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Ghent, Belgium

Jiaao Ma, Ceyu Xu, Lisa Wills, **A Scalable Architecture for Efficient Multi-bit Fully Homomorphic Encryption** Oct 2024
Under submission for ASPLOS 2025

Jiaao Ma, Ceyu Xu, Lisa Wills, **PyTFHE: An End-to-End Compilation and Execution Framework for Fully Homomorphic Encryption Applications** Apr 2023
Best paper award in *2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Raleigh, NC, USA* - 10.1109/ISPASS57527.2023.00012

Honors and Rewards

- ISPASS-2023 Best Paper Award
- Undergraduate Research Opportunities Program (UROP) Scholarship at UCI, 2020
- Summer Undergraduate Research Program (SURP) Scholarship at UCI, 2020
- Dean's Honor List recipient at UCI, 2018-2020

Knowledge Area

Programming: C/C++, Rust, Python, CUDA, Java, Scala, Chisel HDL, Verilog

Compilers and EDA Tools: LLVM, MLIR, Yosys, FIRRTL, Synopsys and Cadence toolchains

ML / FHE Toolchain: PyTorch, Candle ML, DeepSpeed, Concrete-ML, OpenFHE, TFHE-rs