Jianming Tong

+1 470-357-8082 | jianming.tong@gatech.edu | jianming@csail.mit.edu | jianmingtong.github.io

Research Interest

Computer Architecture: Privacy-preserving ML Acceleration; System-Hardware-Software Co-design

Education

2021 – 2025 **Georgia Institute of Technology**, Ph.D. in Computer Science
2023 – present **Massachusetts Institute of Technology**, Visiting Ph.D. in EECS

- *Advisor:* ***Tushar Krishna****; Field: Domain-specific Accelerator and System.*

2016 – 2020 **Xi'an JiaoTong University**, B.E. in Electrical Engineering

- *Advisor:* ***Pengju Ren***; *Thesis:* FPGA Accelerated High-Accuracy SLAM (FPT'21)

Professional Experience

May'18 – Aug' 23 **Rivos Inc,** Mountain View, CA, USA
Engineer Intern, Mentor: ***Gautham Chinya***, (Outcome: Performance Modeling – NDA)

Jul' 22 – Aug' 22 **Pacific Northwest National Laboratory,** Remote, GA, USA
Research Intern, Mentor: ***Roberto Gioiosa***, (Outcome: first-author tutorial @ ICS'22)

Jun'21 – Aug'21 **DAMO Academy Alibaba Inc.**, Beijing, China
Research Intern, Mentor: ***Jiansong Zhang***, (Outcome: open-sourced code DAC'23)

2021 – present **Georgia Institute of Technology**, Atlanta, GA, USA
Graduate Research Assistant

Aug'20 – Jan'21 **Tsinghua University**, Beijing, China
Research Intern, Mentor: ***Yu Wang***, (Outcome: co-first author paper @ ICRA'21)

Sep'18 – Jul'20 **Xi'an JiaoTong University**, Xi'an, ShannXi, China
Undergraduate Research Assistant, (Outcome: project lead – paper @ FPT'21)

Selected Awards and Honors [Full list]

Sep'23 **Best Poster Award** – SUSHI @ Industry-Academia Partner Workshop (IAP'23)
The top voted poster among over 20+ candidates by industry partners.

Jul'23 **Qualcomm Innovation Fellowship** – SUSHI
18 winners out of 182 submissions in north America

Selected Publications (* Equal Contribution) [Full list]

[1] **A Reconfigurable Accel. with Data Reordering Support for Low-Cost On-Chip Dataflow Switching**
Jianming Tong, Anirudh Itagi, Prasanth Chatarasi, Tushar Krishna. (In Submission)
[2] **Leveraging ASIC AI Chips for Homomorphic Encryption**
Jianming Tong, Leo De Castro, Tianhao Huang, Anirudh Itagi, Jingtian Dang, Anupam Golder, Arvind, Edward Suh, Tushar Krishna. (In Submission)
[3] **SUSHI: SUbgraph Stationary Hardware-software Inference Co-design** (MLSys'23)
Payman Behnam*, *Jianming Tong**, Alind, Yangyu, Yue, Pranav, Abhimanyu, Tushar, Alexey Tumanov
[4] **Hardware-Software Co-design for Real-time Latency-Accuracy Navigation in TinyML** (IEEE Micro'23)
Payman Behnam*, *Jianming Tong**, Alind, Yangyu, Yue, Pranav, Abhimanyu, Tushar, Alexey Tumanov

Selected Projects

End-to-end High-performance Machine Learning Inference Acceleration Framework Jan'21 – now

- Spotted layout switching as a performance-critical but often ignored issue in reconfigurable accelerators. A lack of layout consideration could result in up-to **120×** theoretical-practice **performance gap**.
- Proposed HW architecture for hiding layout switching behind data reduction to enable real-time dataflow-layout co-switching at layer granularity, enabling **2.89× speedup over** fix dataflow-layout **Xilinx DPU**.
- Implemented **end-to-end FPGA framework** enabling dataflow-layout co-search, compile & deployment.

Enabling AI Accelerator for Fully Homomorphic Encryption (FHE) Acceleration Jan'23 – now

- Charactered workloads and spotted Number Theory Transform and Basis Conversion as performance bottleneck in FHE (**86.3% latency**) for requiring high-precision large-degree polynomial multiplication.
- Proposed pure scheduling methods to deploy high-precision polynomial multiplication on **MITA** with low-precision matrix multipliers and vector processors (**no HW changes, 4.5× faster than GPU**)

Privacy-preserving Pixel Codec Avatar (PiCA) for AR/VR Jul'23 – now

- Designed algorithmic and system solutions to preserve user's privacy in multi-user virtual reality meeting.
- (Algorithmic) Proposed dynamic privacy filtering to enable **"privacy – costs" tradeoff space navigation**.
- (System) Developed model partitioning strategies for **privacy-preserving model** inference **outsourcing**.

SELECTED TALKS

Enable Efficient AI/FHE Inference on Real-time Practical System

- **HAN Lab @ MIT** – Host: Hanrui Wang, Song Han Oct'23
- **EIC Lab @ GaTech** – Host: Celine Lin Jul'23

Enable Best ML Inference and Training: A systematic Approach @ EIC Lab – GaTech Mar'23

Full-Stack ML Dataflow, Mapping and SW/HW Co-Design and Search @ NICSEFC – Tsinghua Nov'22

SKILLS

Programming	(System) Verilog, Xilinx HLS, C/C++, Python, OpenCL, LLVM, MLIR, Clang
Tools	Xilinx Vivado, Vitis HLS (AI), Cadence, Synapse

SERVICES

Artifact Evaluation Committee (AEC)	ASPLOS'24
Steering Team Member	Computer Architecture Student Association (CSSA)

MENTORSHIP

Name	First employment
Yangyu Chen (Master@GT 2023)	Apple ASIC Verification Designer
Yue Pan (undergrade@GT 2022)	UCSD Ph.D.
Yuqi He (undergrade@GT 2022)	Apple ASIC Designer
Jingtian Dang (undergrade@GT 2022)	CMU ECE Master
Yingkun Liu (undergrade@XJTU 2021)	SJTU Ph.D.
Kedai Zuo (undergrade@XJTU 2021)	UCSD Master
Cheng Wang (undergrade@XJTU 2021)	Tsinghua Ph.D.

Last Updated: Dec 29, 2023