

Jianming Tong

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

RESEARCH INTEREST

Computer Architecture; Model-System-Hardware Co-design for Privacy-preserving Machine Learning.

EDUCATION

- 2021 – present **Georgia Institute of Technology**, Ph.D. in Computer Science
- Advisor: [Tushar Krishna](#)
- 2016 – 2020 **Xi'an JiaoTong University**, B.E. in Electrical Engineering
- Advisor: [Pengju Ren](#); Thesis: FPGA Accelerated High-Accuracy SLAM ([FPT'21](#))

PROFESSIONAL EXPERIENCE

- Sep'23 – present **Visiting Researcher** in **Massachusetts Institute of Technology**, Cambridge, MA
Focus on Algorithmic, System and hardware acceleration for FHE, host: [Arvind](#).
- Jan'24 – May'1 **Teaching Assistant** in **Massachusetts Institute of Technology**, Cambridge, MA
Constructive Computer Architecture (6.192) – Instructor: [Arvind](#), [Tushar Krishna](#)
- Jan'21 – present **Graduate Research Assistant** in **Georgia Institute of Technology**, Atlanta, GA
Software-System-Hardware Codesign for Edge ML ([MLSys'23](#), [IEEE MICRO'23](#))
- May'23 – Aug' 23 **Engineer Intern** in **Rivos Inc.** Mountain View, CA, Mentor: [Gautham Chinya](#)
Inter-chiplet performance modelling (NDA)
- Jan'23 – May'23 **Teaching Assistant** in **Georgia Institute of Technology**, Atlanta, GA
Processor Design (CS 3220) – Homework / Project Design
- Jul' 22 – Aug' 22 **Research Intern** in **Pacific Northwest National Laboratory**, Remote
A single-author end-to-end FPGA framework for AI inference ([Tutorial@ICS'22](#))
- Jun'21 – Aug'21 **Research Intern** in **DAMO Academy Alibaba Inc.** China, Mentor: [Jiansong Zhang](#)
Designed and implemented FPGA accelerator for Homomorphic Encryption scalar/vector multiplication and addition (open-sourced with paper [DAC'23](#))
- Aug'20 – Jan'21 **Visiting Ph.D.** in **Tsinghua University**, Beijing, China, Mentor: [Yu Wang](#)
Designed multi-robot collaborative exploration alg. (co-first author paper [ICRA'21](#))
- Sep'18 – Jul'20 **Undergraduate Research Assistant** in **Xi'an JiaoTong University**, Xi'an, China
Lead NoC book translation and robotic HW acceleration (open-sourced paper [FPT'21](#))

SELECTED AWARDS AND HONORS [[Full list](#)]

- Jun'24 **ML and System Rising Star** @ [MLCommon](#)
- Apr'24 **DAC Young Fellow** @ Design Automation Conference ([DAC'24](#))
- 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** – Latency/Accuracy Navigation in Edge ML
18 winners out of 182 submissions (three rounds, nationwide)
- Oct'22 **Qualcomm Innovation Fellowship Finalist** – ML Accel. Side-channel Attack and Mitigation

LEADING PUBLICATIONS (* EQUAL CONTRIBUTION)

As Principal Contributor and Leading Author

----- Computer Architecture ASIC/FPGA (Hardware) -----

- **A Reconfigurable Accel. with Data Reordering Support for Low-Cost On-Chip Dataflow Switching**
[Jianming Tong](#), Anirudh Itagi, Prasanth Chatarasi, Tushar Krishna.
International Symposium on Computer Architecture ([ISCA](#)), Jun 2024

System

- **SUSHI: SUBgraph Stationary Hardware-software Inference Co-design**
Payman Behnam*, [Jianming Tong*](#), Alind, Yangyu, Yue, Pranav, Abhimanyu, Tushar, Alexey Tumanov
In Proc of Sixth Conference on Machine Learning and Systems (**MLSys**), Jun 2023
++ **Qualcomm Innovation Fellowship**
++ **Best Poster Award**
- **Hardware-Software Co-design for Real-time Latency-Accuracy Navigation in TinyML**
Payman Behnam*, [Jianming Tong*](#), Alind, Yangyu, Yue, Pranav, Abhimanyu, Tushar, Alexey Tumanov
(**IEEE Micro**), Sep 2023
- **FastSwitch: Enabling Real-time DNN Switching via Weight-Sharing**
[Jianming Tong](#), Yangyu Chen, Yue Pan, Abhimanyu Bambhaniya, Alind Khare, Taekyung Heo, Alexey Tumanov, and Tushar Krishna.
The 2nd Architecture, Compiler, and System Support for Multi-model DNN Workloads Workshop@**ISCA**.

Model/Algorithm (Software)

- **SmartPAF: Accurate Low-degree Polynomial Approximation of non-Polynomial Operators for Fast Private Inference in Homomorphic Encryption**
[Jianming Tong*](#), [Jingtian Dang*](#), Anupam Golder, Tushar Krishna. (**MLSys**), Jun 2024
- **SMMR-Explore: SubMap-based Multi-Robot Exploration System with Multi-robot Multi-target Potential Field Exploration Method**
Jincheng Yu*, [Jianming Tong*](#), Yuanfan Xu, Zhilin Xu, Haolin Dong, Tianxiang Yang and Yu Wang
IEEE International Conference on Robotics and Automation (**ICRA**, oral), Jan 2022 [[Code](#)][[Demo](#)]

Book – On-chip Networks

- **On Chip Networks, Second Edition [Translated Book]**
Natalie Enright Jerger, Tushar Krishna, and Li-Shiuan Peh
Translator: Pengju Ren, Tian Xia, [Jianming Tong](#) [Translator Leader], Pengcheng Zong, Haoran Zhao.
Publishing House of Electronics Industry, Jan 2021

COLLABORATIVE PUBLICATIONS (* EQUAL CONTRIBUTION)

• As Collaborator or Mentor

- **On Continuing DNN Accelerator Arch. Scaling Using Tightly-coupled Compute-on-Memory 3D ICs**
Gauthaman Murali, Aditya Iyer, Lingjun Zhu, [Jianming Tong](#), Francisco Munoz Martinez, Srivatsa Rangachar Srinivasa, Tanay Karnik, Tushar Krishna Sung Kyu Lim
IEEE Transactions on Very Large Scale Integration Systems (**TVLSI**), Jul 2023
- **A High-Performance Computing Architecture for Real-Time Digital Emulation of RF Interactions**
Mandovi Mukherjee*, Nael Mizanur Rahman*, Coleman B. DeLude*, Joseph W. Driscoll*, Uday Kamal, Jongseok Woo, Jamin Seo, Sudarshan Sharma, Xiangyu Mao, Payman Behnam, Sharjeel M. Khan, Daehyun Kim, [Jianming Tong](#), Prachi Sinha, Santosh Pande, Tushar Krishna, Justin Romberg, Madhavan Swaminathan, and Saibal Mukhopadhyay
In Proc of IEEE Radar Conference (**RadarConf**), May 2023
- **FPGA-based High-Perf. Real-Time Emulation of Radar System using Direct Path Compute Model**
Xiangyu Mao*, Mandovi Mukherjee*, Nael M. Rahman*, Uday Kamal, Sudarshan Sharma, Payman Behnam, [Jianming Tong](#), Jongseok Woo, Coleman B DeLude, Joseph W. Driscoll, Jamin Seo, Santosh Pande, Tushar Krishna, Justin Romberg, Madhavan Swaminathan, and Saibal Mukhopadhyay
In Proc of IEEE MTT-S International Microwave Symposium (**IMS**), Jun 2023
- **A Configurable Arch. for Efficient Sparse FIR Computation in Real-time Radio Frequency Systems**
Jamin Seo, Nael Mizanur Rahman, Mandovi Mukherjee, Coleman DeLude, [Jianming Tong](#), Justin Romberg, Tushar Krishna, and Saibal Mukhopadhyay.
IEEE Microwave and Wireless Technology Letters (**IMS**), Sep 2022
- **PIT: Processing-In-Transmission with Fine-Grained Data Manipulation Networks**
Tian Xia, Pengchen Zong, Haoran Zhao, [Jianming Tong](#), Wenzhe Zhao, Nanning Zheng and Pengju Ren
IEEE Transactions on Computers (**TOC**), Jul 2021
- **ac2SLAM: FPGA Accelerated High-Accuracy SLAM with Heapsort and Parallel Keypoint Extractor**
Cheng Wang, Yinkun Liu, Kedai Zuo, [Jianming Tong](#) [Project Leader], Yan Ding, and Pengju Ren.
International Conference on Field-Programmable Technology (**FPT**, Full Paper), Jul 2021 [[Code](#)]

- **Content-Oriented Configurable Architecture based on Highly-Adaptive Data Transmission Networks**
Tian Xia, Pengchen Zong, Haoran Zhao, [Jianming Tong](#), Wenzhe Zhao, Nanning Zheng and Pengju Ren
The 30th edition of the ACM Great Lakes Symposium on VLSI ([GLSVLSI](#)), Mar 2020

SELECTED PROJECTS

End-to-end Reconfigurable Flexible Machine Learning Accelerator (FEATHER, ISCA'24) 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 **120x** 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 **2.89x 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 compilation techniques to map high-precision polynomial multiplication on **TPU** with low-precision matrix multipliers and vector processors (**no HW changes, 4.5x 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 outsourcing** using FHE

SELECTED TALKS

Hardware Architecture Innovation to Enable (Dataflow, Layout) CoSwitching

- **Nvidia @ Westford** – Host: [Angshuman Parashar](#) Jul'24
- **MIT** – Host: [Vivienne Sze](#) Jun'24

Enable Multi-Query AI Inference on Real-time Practical System

- **HAN Lab @ MIT** – Host: [Hanrui Wang](#), [Song Han](#) Oct'23
- **EIC Lab @ GaTech** – Host: [Celine Lin](#) Jul'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'23, ASPLOS'24, ISCA'24
Steering Team Member Computer Architecture Student Association (CSSA)

MENTORSHIP

Name

Anirudh Itagi (Master@GT 2024)
Yangyu Chen (Master@GT 2023)
Yue Pan (undergrade@GT 2022)
Yuqi He (undergrade@GT 2022)
Jingtian Dang (undergrade@GT 2022)
Yingkun Liu (undergrade@XJTU 2021)
Kedai Zuo (undergrade@XJTU 2021)
Cheng Wang (undergrade@XJTU 2021)

First employment

Microsoft Azure AI Infrastructure
Apple ASIC Verification Designer
UCSD Ph.D.
Apple ASIC Designer
CMU ECE Master -> Now Ph.D. at GaTech
SJTU Ph.D.
UCSD Master
Tsinghua-XJTU Ph.D.