Jianming Tong

+1 470-357-8082 | jianming.Tong@gatech.edu | jianmingTong.github.io

Research Interest

Privacy-preserving AR/VR; System-Hardware-Software Co-designed ML Acceleration; On-chip Network; FPGA/VLSI;

Education

Visiting Ph.D. Student, Aug 2023 – Dec 2023 **Massachusetts Institute of Technology (MIT), Boston**

- Advisor: Dr. **Tushar Krishna**

Ph.D. in Computer Science, Jan 2021 – Dec 2025 **Georgia Institute of Technology (GT), Atlanta**

- Advisor: Dr. **Tushar Krishna**
- GPA: 4.0/4.0

Visiting Ph.D. Student, July 2020 – Jan 2021 **Tsinghua University (THU), Beijing**

- Advisor: Dr. **Yu Wang**

B.S. in Electrical Engineering, Sep 2016 – June 2020 **Xi`an JiaoTong University (XJTU), Xi`an**

- Advisor: Dr. **Pengju Ren**
- GPA: 3.91/4.3, Rank: 6/361

Award

Best Poster Award @ IAP2023	Sep 2023
Qualcomm Innovation Fellowship	Jul 2023
Finalist in Qualcomm Innovation Fellowship	Oct 2022
2nd Georgia Tech SCS Poster Competition	Apr 2022
National Encouragement Scholarship **(top 5%)**	Sep 2017, Sep 2018, Sep 2019
Huawei Scholarship **(top 6/60)**	May 2019

Peer-reviewed Publications (* Equal Contribution)

- **Hardware-Software co-design for real-time latency-accuracy navigation in tinyML applications**
  Payman Behnam*, **Jianming Tong***, Alind Khare, Yangyu Chen, Yue Pan, Abhimanyu Bambhaniya, Pranav Gadikar, Tushar Krishna, and Alexey Tumanov
  IEEE micro**, **Sep 2023

- **On Continuing DNN Accelerator Architecture 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 Reconfigurable Accelerator with Data Reordering Support for Low-Cost On-Chip Dataflow Switching**
  Jianming Tong, Anirudh Itagi, Tushar Krishna
  *3rd On-Device Intelligence Workshop, @ In Proc of Sixth Conference on Machine Learning and Systems **(MLSys), Jun 2023***

- **ReLU-FHE: Low-cost Accurate ReLU Polynoimal Approximation in Fully Homomorphic Encryption Based ML Inference**
  Jingtian Dang*, **Jianming Tong***, Anupam Golder, Callie Hao, Tushar Krishna
  *3rd On-Device Intelligence Workshop, @ In Proc of Sixth Conference on Machine Learning and Systems **(MLSys), Jun 2023***

- **SUSHI: SUbgraph Stationary Hardware-software Inference Co-design**
  Payman Behnam***, Jianming Tong*,** Alind Khare, Yangyu Chen, Yue Pan, Pranav Gadikar, Abhimanyu Rajeshkumar Bambhaniya, Tushar Krishna, Alexey Tumanov
  *In Proc of Sixth Conference on Machine Learning and Systems **(MLSys), Jun 2023***
  ++ Qualcomm Innovation Fellowship 2023
  ++ Best Poster Award

- **FPGA-Based High-Performance 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 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***

- **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, 2022****.*

- **A Configurable Architecture 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****), 2022.*

- **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
  *2021 IEEE International Conference on Robotics and Automation (****ICRA 2021, oral****)* [code], [Demo Link].

- **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 2021, Full Paper****)* [code].

- **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****)*

- **COCOA: 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 2020****)*

Books

- **On Chip Networks, Second Edition [Translated Book]**
  Natalie Enright Jerger, Tushar Krishna, and Li-Shiuan Peh
  Synthesis Lectures on Computer Architecture, Morgan Claypool Publishers, June 2017
  Chinese Translator: Pengju Ren, Tian Xia, **Jianming Tong [Project Leader]**, Pengcheng Zong, Haoran Zhao.
  Publishing House of Electronics Industry, Jan 2021

PROJECTS

Privacy-preserving Pixel Codec Avatar (PiCA) for AR/VR
Advisor: **Tushar Krishna**, **Edward Suh**, **Peter Capak** (Meta)**, Krishnakumar Nair** (AMD). Jul 2023 – now
- Exploring quantization schemes to reduce compute demand of PiCA enabling multi-users VR meetings.
- Exploring new model partitioning to secure user-defined content through privacy-preserving processing.

Real-time Privacy-preserving Cloud Serving through Fully Homomorphic Encryption (FHE) Acceleration
Advisor: **Tushar Krishna,** and **Edward Suh** (Cornell, Meta AI) Jan 2023 – now
- Enabled secure neural network inference in fully homomorphic encryption domain using BFV/CKKS schemes.
- Proposed new algorithm-hardware co-design for efficient polynomial multiplication and addition.

Software-System-Hardware Co-design for runtime Latency/Accuracy Navigation on the Edge Device
Advisor: **Tushar Krishna** and **Alexey Tumanov**, Georgia Tech Jan 2021 – now

- Proposed new paradigm of SW/HW co-design for pareto-frontier models instead of a single model.
- Designed fully scalable, parameterized and verified architecture template using mix of Xilinx HLS + Verilog.
- Awarded Qualcomm Innovation Fellowship; Best Poster Award; (Published MLSys 2023, IEEE micro 2023)

(DARPA) High-performance Network-on-Chip (NoC) Enabling Arbitrary Multicasting and Unicasting
Advisor: **Tushar Krishna** and **Saibal Mukhopadhyay**, Georgia Tech Jan 2021 – Jan 2022

- Designed scalable NoC supporting arbitrary multicasting and unicasting from 1024 sources to 204 destinations.
- Synthesized, PNR and tapped out the designed NoC in scalable chip under TSMC 28 nm.
- Published IMS 2022, IMS 2023, RadarConf 2023.

Industry Experience

Performance Modeling **Rivos Inc, CA**

- Mentor: **Gautham Chinya**; **Research Intern,** May 2023 – Aug 2023
- Dedicated on-chip network performance modeling (Stealth-mode company, no details disclosed)

End-to-end Framework for Inference **Pacific Northwest National Laboratory, WA**

- Mentor: **Roberto Gioiosa**; **Research Intern,** Jun 2022 – Aug 2022
- Designed end-to-end inference framework on heterogeneous clusters with Xilinx VCK 5000, AMD CPU & GPU.
- Designed full-stack end-to-end inference including quantization, compiler, run-time and FPGA accelerator.
- Released end-to-end FPGA inference framework, **https://maeri-project.github.io/** (Tutorial @ICS 2022)

Fully Homomorphic Encryption (FHE) Acceleration Architecture Design **DAMO Academy Alibaba Inc., Beijing**

- Mentor: **Jiansong Zhang**; **Research Intern,** Jun 2021 – Aug 2021
- Designed efficient linear-operator architectures and deployed on Xilinx Alevo U280 for PoC.
- Categorized serial and parallel accelerator architectures for Number Theoretic Transform (NTT).
- Published DAC 2023 (code open-sourced [**link**])

Selected Talks

Enable Efficient AI/FHE Inference on Real-time Practical System @ HAN Lab – MIT Oct 2023

Enable Efficient AI/FHE Inference on Real-time Practical System @ CAG Lab – XJTU Jul 2023

Enable Best ML Inference and Training: A systematic Approach @ EIC Lab – GaTech Mar 2023

Full-Stack ML Dataflow, Mapping and SW/HW Co-Design and Search @ NICSEFC Lab – Tsinghua Nov 2022

Skills

Programming	C/C++, Python, OpenCL, MCL, LLVM, Clang, MLIR, (System) Verilog, Xilinx HLS
Tools	GEM5, SST, Scale-Sim, MAESTRO, Timeloop, Xilinx Vivado, Vitis HLS (AI), Cadence, Synapse

Services

Artifact Evaluation Committee (AEC)	ASPLOS’24

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: Oct 10, 2023