

Jianming Tong

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

Research Interest

Computer Architecture; Hardware-software co-design; Deep Learning Compiler and Accelerator; On-chip Network; FPGA; VLSI Design

Education

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

Visiting Ph.D. Student, July 2020 – Jan 2021

Tsinghua University (THU), Beijing

- Advisor: Dr. [Yu Wang](#)

Ph.D. in Computer Science, Jan 2021 – Dec 2025

Georgia Institute of Technology (GT), Atlanta

- Advisor: Dr. [Tushar Krishna](#)
- GPA: 4.0/4.0

Award

Finalist in **Qualcomm Fellowship**

Oct 2022

National Encouragement Scholarship (**top 5%**)

Sep 2017, Sep 2018, Sep 2019

Huawei Scholarship (**top 6/60**)

May 2019

Project

MAERI 2.0: An end-to-end Inference Framework for enabling accelerator compiler research

Advisor: [Tushar Krishna](#) and [Alexey Tumanov](#), Georgia Tech

Jan 2021 – now

- Extended MAERI with memory-hierarchy, compiler and running time to verified end-to-end inference framework.
- Enabled “searched mapping from MAESTRO” deployment and verified on Xilinx VCK, Alevo and ZYNQ FPGA.
- Designed fully scalable, parameterized and verified architecture template using Xilinx HLS.
- <https://maeri-project.github.io/> (**Tutorial @ICS 2022**).

(IARPA) Large-scale Network-on-Chip (NoC) Modeling and Simulation in SST Simulator.

Advisor: [Tushar Krishna](#), Georgia Tech, collaborated with Intel.

Nov 2022 – now

- Modeling, simulating and testing large-scale multi-dimension NoC in [SST simulator toolkit](#).

(DARPA) High-performance Network-on-Chip (NoC) Design 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 24 nm.

TorchFHE: An PyTorch library for privacy-preserving neural network inference using FHE.

Advisor: [Tushar Krishna](#) and [Callie Hao](#), Georgia Tech

Mar 2022 – now

- Enabled secure neural network inference in fully homomorphic encryption domain using BFV/CKKS schemes.

Industry Experience

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.

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

- Mentor: [Jiansong Zhang](#); **Research Intern**, Jun 2021 – Aug 2021
- Designed efficient homomorphic encryption acceleration architecture and deployed on Xilinx Alevo U280 for PoC.
- Categorized serial and parallel accelerator architectures for Number Theoretic Transform (NTT).

Peer-reviewed Publications (* Equal Contribution)

- **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 [Translator Leader]**, Pengcheng Zong, Haoran Zhao.
Publishing House of Electronics Industry, Jan 2021

Skills

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