Jianming Tong

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

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**

Industry Experience

Fully Homomorphic Encryption (FHE) Acceleration Architecture Design DAMO Academy Alibaba Inc., Beijing
- Mentor: Jiansong Zhang; **Research Intern,** Jun 2021 – Aug 2021
- Designed and Implemented architecture for multiplication in FHE on FPGA.
- Categorize existing architectures for Number Theoretic Transform (NTT).

End-to-end Framework for Inference Pacific Northwest National Laboratory, WA
- Mentor: Roberto Gioiosa; **Research Intern,** Jun 2022 – Aug 2022
- Design and implement inference accelerator 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.

Peer-reviewed Publications

- **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**)*

- **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**)*

- **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

- **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].

- **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**.*

- **FastSwtich: 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.***

Project

MAERI 2.0: An End-to-end Inference Framework for enabling accelerator compiler research
Advisor: **Tushar Krishna** and **Alexey Tumanov**, **Georgia Tech** Jan 2020 – now

- Extend MAERI to fully scalable, parameterized and verified architecture template using Xilinx HLS.
- Enable deployment of convolution layer on MAERI by designing scalable on-chip memory hierarchy and compiler.
- https://maeri-project.github.io/ Give Tutorial at ICS 2022.
- The **Multi-tiling processing nature** of the MAERI 2.0 serves as infrastructure for compiler research.

High-performance Network-on-Chip (NoC) Design enabling arbitrary multicasting and unicasting
Advisor: **Tushar Krishna** and **Saibal Mukhopadhyay**, **Georgia Tech** Jan 2020 – now

- Designed scalable hierarchy NoC supporting arbitrary multicasting and unicasting.
- Synthesized, implemented and deploy the designed NoC to ASIC (already taped out).

TorchFHE: An PyTorch library for FHE based Machine Learning and Acceleration
Advisor: **Tushar Krishna** and **Callie Hao**, **Georgia Tech** Jan 2022 – now

- Enable secure Neural Network (NN) inference under Fully Homomorphic Encryption (FHE) on CPU.
- Supported optimization including Residue Number System (RNS), Number Theoretic Transform (NTT).
- Open high-level plug-in-and-play optimization interface and deployment on heterogeneous platforms.

Award

Huawei Scholarship (top 6/60)	May 2019
National Encouragement Scholarship (top 5%)	Sep 2017, Sep 2018, Sep 2019
Finalist in Qualcomm Fellowship	May 2022

Skills

Language	Chinese (Native), English (Fluent)
Software Languages	C/C++, JAVA, VBA, Matlab, Shell, Bash, and Python
Hardware languages	(System) Verilog, High-level-Synthesis (HLS) and Bluespec System Verilog
FPGA/ASIC tools	Xilinx Vivado, Altera Quartus and Xilinx Vitis (AI), Cadence Innovus, Synopse
Deep Learning	PyTorch, Keras (Tensorflow backended), Reinforcement Learning
Domain-specific Accelerator	MAERI, Eyeriss, Sigma, ExTensor, SCNN, Xilinx DPU, Xilinx AIE
Compiler	LLVM, Clang, Linux Kernel
Heterogeneous Programming	OpenCL, MCL
On-chip Network	GEM5, Garnet