

## Jianming Tong

+1 470-357-8082 | [jianming.tong@gatech.edu](mailto:jianming.tong@gatech.edu) | [jianmingtong.github.io](https://jianmingtong.github.io)

### RESEARCH INTEREST

**Computer Architecture** with major interest on full-stack optimizations—spanning software, systems, and hardware—aimed at enhancing both the efficiency and privacy of AI systems.

### EDUCATION

- Jan'21 – May'26 **Georgia Institute of Technology, PhD** in Computer Science  
Advisor: [Tushar Krishna](#), *Thesis*: Privacy-preserving ML Acceleration
- Jan'21 – May'24 **Georgia Institute of Technology, MS** in Computer Science  
Advisor: [Tushar Krishna](#), *Thesis*: Reconfigurable Dataflow Accelerator [[ISCA'24](#)]
- Sep'16 – May'20 **Xi'an JiaoTong University, BE** in Electrical Engineering  
Advisor: [Pengju Ren](#), *Thesis*: FPGA Accelerated High-Accuracy SLAM [[FPT'21](#)]

### PROFESSIONAL EXPERIENCE

- Feb'25 – present **Visiting Researcher** in **Massachusetts Institute of Technology**, Cambridge, MA  
Focus on Authenticated ML Acceleration, host: [Mengjia Yan](#)
- Aug'24 – Apr'25 **Student Researcher** in **Google**, Cambridge, MA  
Leveraging Google TPU for Homomorphic Encryption, host: [Asra Ali](#), [Jevin Jiang](#)
- Sep'23 – Feb'25 **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 Chinaya](#)  
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
- Jun' 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](#))
- Jul'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

### PUBLICATIONS (\* EQUAL CONTRIBUTION) -- CONFERENCE

- In Submission** Enabling Multi-scalar Multiplication over AI ASICs for ZKP  
**+Deployed in Google Cloud (TPU series)** **Jianming Tong**, Jingtian Dang, Simon Langowski, Tianhao Huang, Asra Ali, Jeremy Kun, Jevin Jiang, Sridhar Devadas, Tushar Krishna  
[\[Code\]](#)

<b>In Submission</b> +Deployed in Google Cloud (TPU series) +SotA Performance	Leveraging ASIC AI Chips for Homomorphic Encryption <b>Jianming Tong</b> , Tianhao Huang, Leo De Castro, Anirudh Itagi, Jingtian Dang, Anupam Golder, Asra Ali, Jevin Jiang, Arvind, G. Edward Suh, Tushar Krishna Preprinted, Jan 2025 <a href="#">[Code]</a>
<b>In Submission</b>	Privatar: Enabling Privacy-preserving Real-time Multi-user VR through Secure Outsourcing <b>Jianming Tong</b> , Hanshen Xiao, Hao Kang, Krishnakumar Nair, Ashish Sirasao, G. Edward Suh, Tushar Krishna <a href="#">[Code]</a>
<b>ISCA 2024</b> +LayoutLoop Integrated into NVLabs/timeloop	FEATHER: A Reconfigurable Accel. with Data Reordering Support for Low-Cost On-Chip Dataflow Switching <b>Jianming Tong</b> , Anirudh Itagi, Prasanth Chatarasi, Tushar Krishna. International Symposium on Computer Architecture, Jun 2024 <a href="#">[Code]</a> <a href="#">[Talk]</a>
<b>MLSys 2024</b>	SmartPAF: Accurate Low-degree Polynomial Approximation of non-Polynomial Operators for Fast Private Inference in Homomorphic Encryption <b>Jianming Tong*</b> , Jingtian Dang*, Anupam Golder, Tushar Krishna. In Proc of Seventh Conference on Machine Learning and Systems, Jun 2024 <a href="#">[Code]</a>
<b>MLSys 2023</b> +Qualcomm Innovation Fellowship +Best Poster Award	SUSHI: SUBgraph Stationary Hardware-software Inference Co-design Payman Behnam*, <b>Jianming Tong*</b> , Alind, Yangyu, Yue, Pranav, Abhimanyu, Tushar, Alexey Tumanov. In Proc of Sixth Conference on Machine Learning and Systems, Jun 2023
<b>IEEE Micro 2023</b>	Hardware-Software Co-design for Real-time Latency-Accuracy Navigation in TinyML Payman Behnam*, <b>Jianming Tong*</b> , Alind, Yangyu, Yue, Pranav, Abhimanyu, Tushar, Alexey Tumanov. IEEE Micro, Sep 2023
<b>TRadar 2024</b>	Real-time Digital RF Emulation – II: A Near Memory Custom Accelerator Xiangyu Mao, Mandovi Mukherjee, Nael Mizanur Rahman, Coleman B DeLude, Joseph W. Driscoll, Sudarshan Sharma, Payman Behnam, Uday Kamal, Jongseok Woo, Daehyun Kim, Sharjeel M. Khan, <b>Jianming Tong</b> , Jamin Seo, Prachi Sinha, Madhavan Swaminathan, Tushar Krishna, Santosh Pande, Justin Romberg, and Saibal Mukhopadhyay. IEEE Transactions on Radar Systems, Sep 2024.
<b>TVLSI 2023</b>	On Continuing DNN Accelerator Arch. Scaling Using Tightly-coupled Compute-on-Memory 3D ICs Gauthaman Murali, Aditya Iyer, Lingjun Zhu, <b>Jianming Tong</b> , Francisco Munoz Martinez, Srivatsa Rangachar Srinivasa, Tanay Karnik, Tushar Krishna, Sung Kyu Lim IEEE Transactions on Very Large Scale Integration Systems, Jul 2023
<b>RadarConf 2023</b>	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, <b>Jianming Tong</b> , Prachi Sinha, Santosh Pande, Tushar Krishna, Justin Romberg, Madhavan Swaminathan, and Saibal Mukhopadhyay. In Proc of IEEE Radar Conference, May 2023
<b>SENSORS 2023</b>	SNATCH: Stealing Neural Network Architecture from ML Accelerator in Intelligent Sensors Sudarshan Sharma, Uday Kamal, <b>Jianming Tong</b> , Tushar Krishna, and Saibal Mukhopadhyay. IEEE SENSORS conference, Aug 2023.

- IMS 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, Jun 2023
- IMS 2021** 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, Sep 2022
- TOC 2021** 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, Jul 2021
- FPT 2021** ac2SLAM: FPGA Accelerated High-Accuracy SLAM with Heapsort and Parallel Keypoint Extractor  
Cheng Wang, Yinkun Liu, Kedai Zuo, **Jianming Tong**, Yan Ding, and Pengju Ren.  
International Conference on Field-Programmable Technology, Jul 2021 [\[Code\]](#)
- ICRA 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.  
IEEE International Conference on Robotics and Automation, Jan 2022 [\[Code\]](#)[\[Demo\]](#)
- GLSVLSI 2020** 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, Mar 2020
- ACS-DNN 2022** 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

---

## **PUBLICATIONS (\* EQUAL CONTRIBUTION) -- BOOK**

### **On Chip Networks, Second Edition [Translated Book in Mandarin]**

Natalie Enright Jerger, Tushar Krishna, and Li-Shiuan Peh

Translator: Pengju Ren, Tian Xia, **Jianming Tong**, Pengcheng Zong, Haoran Zhao.

Publishing House of Electronics Industry, Jan 2021 [\[Link\]](#)[\[OriginalVersion\]](#)

---

## **SELECTED TALKS**

### **Leveraging ASIC AI chips for Homomorphic Encryption**

Google Host: [Jeremy Kun](#) (May'24)

IBM Host: [Manoj Kumar](#), [Pradip Bose](#) (Aug'24)

NYU Host: [Brandon Reagon](#) (Nov'24)

### **Enabling Reconfigurable (Dataflow, Layout) CoSwitching in AI Accelerator**

MIT Host: [Vivienne Sze](#) (Jun'24)

NVIDIA Host: [Angshuman Parashar](#) (Jul'24)

GaTech Host: [Alexandros Daglis](#) (Aug'24)

### **Enabling Real-time Accuracy Latency Navigation in Multi-Query AI Inference**

HAN Lab @ MIT Host: [Hanrui Wang](#), [Song Han](#) (Oct'23)

EIC Lab@GaTech Host: [Celine Lin](#) (Jul'23)

## **A Sparse and Irregular GEMM Accelerator with Flexible Interconnects**

Tsinghua Host: [Shulin Zeng](#) (Nov'22)

### **SELECTED PROJECTS**

#### **PROVE -- Leveraging ASIC AI chips for Zero-Know Proof (ZKP)**

Dec'24 – now

- Proposed Lazy Reduction, achieving up to an 18.5x speedup over high-precision Barrett Reduction.
- Proposed systematical memory intrinsics to explicitly explore dataflow factoring memory overhead.
- Deployed in Google cloud, achieving 111x end-to-end Multi-scalar Multiplication speedup over CPUs, enabling immediate Zero-knowledge Proof acceleration on commodity AI accelerators like TPUs [\[link\]](#).

#### **CROSS -- Leveraging ASIC AI chips for Homomorphic Encryption (HE)**

Jan'23 – now

- Proposed Basis Aligned Transformation to convert high-precision multiplication as low-precision MatMul.
- Proposed Layout Invariant 3-step Number Theory Transformation to convert NTT as Matrix Multiplication.
- CROSS achieves immediate privacy-preserving ML via TPUs (**126x, 5x faster than CPU, GPU-A100**).
- Integrated into Google privacy-preserving library and deployed in Google jaxite [\[link\]](#).

#### **Enabling Privacy-preserving Real-time Multi-User VR Through Secure Outsourcing**

Jul'23 – Jan'25

- Proposed **Horizontal Partitioning (HP)** to split multi-user VR flow into local-cloud for offloading less private data to untrusted cloud with noisy perturbation privacy protection for supporting more users.
- Developed, **Privatar**, the first framework leverages both local and untrusted cloud to concurrently achieve privacy-preserving multi-user VR, with **1.5x~2.27x** higher accuracy and **3.75x** more users support than the SotA completely model outsourcing. Such benefits only come at a negligible **9%** energy consumption.
- Applied PAC Privacy to multi-user VR, reducing noise intensity by up to **158x** compared to state-of-the-art differential privacy, achieving stronger privacy with minimal accuracy loss.

#### **Approximating Non-linear Layers in ML Models for Homomorphic Encryption**

Dec'22 – Jan'24

- Aim at reducing polynomial approximation degree of non-linear ML layers while preserving accuracy.
- Proposed, **SmartPAF**, the **first** training framework to replace non-linear operators with low-degree Polynomial Approximation Function and recover accuracy via ML fine-tuning, achieving **7.81x** speedup.
- Published [MLSys'24](#) with open-sourced [code](#), WIP to be integrated in Google [HEIR](#) compiler.

#### **End-to-end Reconfigurable Flexible Machine Learning Accelerator (FEATHER, ISCA'24)**

Jan'21 – Jan'24

- Spotted layout switching as a performance-critical but often ignored issue in reconfigurable accelerators. A discordant layout slows down the theoretical performance of dataflow by up-to **120x**.
- Proposed **FEATHER**, the first architecture enabling (dataflow, layout) coswitching via novel NoC, **BIRRD**.
- Proposed **functional arbitrary reordering** to enable arbitrary layout switching and **implementational reordering in reduction** to hide layout reordering latency behind critical path.
- **Deployed on real FPGA**, achieving **2.65~4.56x** end-to-end throughput/PE improvement over SotAs.
- Published [ISCA'24](#) with open-sourced [code](#), layout modeling deployed in NVIDIA [Timeloop](#) library.

#### **Enable Real-time Latency/Accuracy Navigation in Edge Applications**

Mar'21 – Oct'22

- Worked on scheduling and hardware of multi-query inference system to improve performance.
- Proposed **SubGraph Stationary** to reuse shared weights of weight-shared networks across queries.
- Designed **SUSHI**, a multi-query inference serving system enabling SubGraph Stationary with novel hardware (SushiAccel) and software (SushiSched), improving **latency / accuracy** by **25% / 0.98%**.
- Published [MLSys'23](#), [IEEE Micro'23](#), wins [Qualcomm Innovation Fellowship](#), [Best Poster Award](#)@IAP'23.

#### **Scalable Arbitrary Unicasting and Multicasting On-chip Network**

Jan'21 – May'22

- Designed a scalable multi-stage on-chip network for **arbitrary multicasting** across hundred or **thousand nodes**, achieving  **$O(N \log N)$  scalability** with the number of nodes  $N$ , better than  $O(N^2)$  of crossbar.
- **Taped out** a realistic test chip for with 16 nodes under TSMC 28nm, verified on a real FPGA prototype.
- Published [TRadar'24](#), [IMS'23](#), [IMS'21](#), [RadarConf'23](#).

## **SKILLS**

---

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

## **SERVICES**

---

**Reviewers** ICRA'24, IROS'24, MLSys'25, CAL'25  
**AEC** ASPLOS'23, ASPLOS'24, ISCA'24  
**Steering Team** Computer Architecture Student Association ([CASA](#))

## **MEDIA COVERAGE**

---

**Semiconductor Engineering** My paper “**Leveraging ASIC AI Chips for Homomorphic Encryption**” is added into Semiconductor Engineering’s library for its potential impact to boarder industry.  
**ACE News** Jianming Tong: Spotlight from DARPA SRC JUMP2.0 Program ACE center (Aug’24)  
**GaTech News** Jianming Tong: Ph.D. Students Named Rising Stars in Machine Learning (Jul’24)  
**GaTech News** Jianming Tong: Ph.D. Students Won Qualcomm Innovation Fellowship (Jul’23)  
**GaTech News** Jianming Tong won 2<sup>nd</sup>-place in SCS Poster Competition (Apr’23)

## **MENTORSHIP**

---

	<b>Name</b>	<b>First employment</b>
<b>MS MIT 2025</b>	Nathan Xiong	MIT Master now
<b>MS MIT 2025</b>	Jan Strzeszynski	MIT Master now
<b>MS GT 2025</b>	Yujie Li	GaTech ECE Master now.
<b>MS GT 2024</b>	Anirudh Itagi	Microsoft Azure AI Infrastructure
<b>MS GT 2023</b>	Yangyu Chen	Apple ASIC Verification Designer
<b>UG GT 2022</b>	Yue Pan	UCSD Ph.D.
<b>UG GT 2022</b>	Yuqi He	Apple ASIC Designer
<b>UG GT 2022</b>	Jingtian Dang	CMU ECE Master -> Now Ph.D. at GaTech
<b>UG XJTU 2021</b>	Yingkun Liu	SJTU Ph.D.
<b>UG XJTU 2021</b>	Kedai Zuo	UCSD Master
<b>UG XJTU 2021</b>	Cheng Wang	Tsinghua-XJTU Ph.D.