Yongqin Wang

+1 (425)-445-5266 | yongqin@usc.edu | Pasadena CA. 91103 | Google Scholar

Education

University of Southern California Los Angeles, CA
Ph.D. Candidate in Computer Engineering

University of Southern California Los Angeles, CA
Master of Science in Computer Engineering

University of Washington Seattle, WA
Bachelor of Science in Electrical Engineering

Research Interests

- Secure Multi-party Computing
- Oblivious RAM
- Trusted Execution Environment

Work Experiences

Research Assistant Los Angeles, CA
University of Southern California *Sept, 2019 – present*
Conduct research on privacy-preserving technologies including but not limited to Trusted Execution Environments, Oblivious RAM, and Multi-party Computation.

Research Intern Pasadena, CA
Meta Platform *May, 2021 – December 2021*
Studies major Transformer-based model inference runtime overheads and potential optimizations when Multi-party computing (MPC) is implemented. Results are published in ISPASS 2022.

Publications

Fastrack: Fast IO for Secure ML using GPU TEEs. **[under review]**
Authors: ***Yongqin Wang****, Rachit Rajat*, Murali Annavaram
ML training/inference can suffer significant IO overheads when running inside GPU TEE systems. This paper introduces an efficient IO communication scheme between CPU and GPU TEE, such that the CPU/GPU IO overheads are significantly reduced, resulting in 85% up to end-to-end runtime reduction.

PIGEON: A Framework for Private Inference of Neural Networks. **[under review]**
Authors: Christopher Harth-Kitzerow, ***Yongqin Wang***, Rachit Rajat, Murali Annavaram
This paper introduces an efficient framework for privacy-preserving MPC ML. It utilizes a novel ABG programming model that switches between Arithmetic vectorization, Bitslicing, and GPU offloading to optimize performance for different neural network layers. PIGEON significantly improves ReLU throughput by two orders of magnitude, reduces GPU memory usage, and supports larger batch sizes compared to state-of-the-art methods. It also achieves up to 70% saturation for 25Gbps networks, offering a scalable, modular, and protocol-agnostic framework for efficient private inference.

CompactTag: Minimizing Computation Overheads in Actively-Secure MPC for DNN. **[under review]**
Authors: ***Yongqin Wang***, Pratik Sarkar, Nishat Koti, Arpita Patra, Murali Annavaram

MPC for active malicious adversaries uses MACs to check the final results. MAC-related computation increases runtime by around 30%. In this paper, I propose a compression technique that eliminates most of the MAC-related computations, providing up to 1.41x runtime speedups.

High-Throughput Secure MPC with an Honest Majority in Various Network Settings. **[PETS 2025]**
Authors: Christopher Harth-Kitzerow, Ajith Suresh, ***Yongqin Wang***, Hossein Yalme, Georg Carle, Murali Annavaram
In this work, we present novel protocols over rings for semi-honest secure three-party computation (3PC) and malicious four-party computation (4PC) with one corruption. Our protocols address these issues by tolerating multiple arbitrarily weak network links between parties without any substantial decrease in performance. Additionally, they significantly reduce computational complexity by requiring up to half the number of basic instructions per gate compared to related work. These improvements lead to up to twice the throughput of state-of-the-art protocols in homogeneous network settings and even larger performance improvements in heterogeneous settings.

MPC-Pipe: An Efficient Pipeline Scheme for Semi-honest MPC Machine Learning. **[ASPLOS 2024]**
Authors: ***Yongqin Wang***, Rachit Rajat, Murali Annavaram
MPC-Pipe is a novel and efficient MPC framework that utilizes communication and computation overlaps to reduce ML model inference runtime latency. There are three major pipeline schemes introduced in the paper: 1) inter-linear pipeline, 2) inner-layer pipeline, and 3) inter-batch pipeline. MPC-Pipe achieves 33% throughput improvement and 13% latency improvement for the state-of-art DNNs.

LAORAM: A Look Ahead ORAM Architecture for Training Large Embedding Tables. **[ISCA 2023]**
Authors: Yongqin Wang*, Rachit Rajat*, Murali Annavaram
LAORAM proposes a private training method for CPU-based large embedding tables using ORAM. LAORAM proposes an aggressive superblock formation mechanism that significantly reduces the number of access to the CPU-based large embedding tables and uses a fat-tree organization to mitigate contentions over the stash in the GPU client. Those schemes combined can reduce the ORAM access latency by up to 5.4x.

PageORAM: An Efficient DRAM Page Aware ORAM Strategy. **[MICRO 2022]**
Authors: Rachit Rajat, ***Yongqin Wang***, Murali Annavaram
In this work, we introduce a new method to read paths in PathORAM to reduce stash management costs by fetching additional data nodes in the same subtree in the ORAM tree organizations to have more opportunities for data block evictions, reducing the contention to the stash space.

Characterization of MPC-based Private Inferences for Transformer-based Models. **[ISPASS 2022]**
Authors: ***Yongqin Wang***, Edward Suh, Wenjie Xiong, Benjamin Lefaudeux, Brian Knott, Murali Annavaram, Hsien-Hsin Lee
In this work, we provide an in-depth character study of the performance overhead for running the now popular Transformer models with secure multi-party computing (MPC). Three unique challenges are identified: 1) significant Softmax runtime, 2) significant embedding table lookups, and 3) fixed point numerical stability issue.

DarKnight: A Data Privacy Scheme for Training and Inference of Deep Neural Networks. **[MICRO 2021]**
Authors: Hanieh Hashemi, ***Yongqin Wang***, Murali Annavaram
This work provides a cloud machine learning training & inference computation scheme that achieves input image privacy and performance improvements. We use coded computing and trusted execution environments to achieve input image privacy and GPU to achieve performance gains.

Origami inference: Private inference using hardware enclaves. **[CLOUD 2021]**
Authors: Krishna Giri Narra, Zhifeng Lin, ***Yongqin Wang***, Keshav Balasubramanian, Murali Annavaram
This work provides a cloud machine learning inference computation mode, which distributes layers in machine learning model inference to different computation units to achieve performance and input image privacy. Layers whose inputs are highly correlated with original images are computed on trusted hardware to achieve privacy and

layers whose inputs have low correlation with original images are computed on GPU to improve performance. Our model can achieve considerate performance improvements on our baseline.

* Equal contributions.

Workshops/Posters

MPC-Pipe: An Efficient Pipeline Scheme for Semi-honest MPC Machine Learning. **[PETS 2024 Poster]**
Authors: ***Yongqin Wang***, Rachit Rajat, Murali Annavaram

CompactTag: Minimizing Computation Overheads in Actively-Secure MPC for DNN. **[PETS 2024 Poster]**
Authors: ***Yongqin Wang***, Pratik Sarkar, Nishat Koti, Arpita Patra, Murali Annavaram

Characterizing MPC-based Inference for Transformer-based Models. **[NeurIPS 2021 Workshop]**
Authors: ***Yongqin Wang***, Edward Suh, Wenjie Xiong, Benjamin Lefaudeux, Brian Knott, Murali Annavaram, Hsien-Hsin Lee

Look Ahead ORAM: Obfuscating Addresses in Recommendation Model Training. **[ISCA 2021 Workshop]**
Authors: Rachit Rajat*, ***Yongqin Wang***, Murali Annavaram

Byzantine-Robust and Privacy-Preserving Framework for FedML. **[ICLR 2021 Workshop]**
Authors: Hanieh Hashemi, ***Yongqin Wang***, Murali Annavaram

Privacy and Integrity Preserving Training Using Trusted Hardware. **[ICLR 2021 Workshop]**
Authors: Hanieh Hashemi, ***Yongqin Wang***, Murali Annavaram

* Equal contributions

Talks

Optimizing ML MPC from System & Theoretical Perspective.
NIST Workshop on Privacy-Enhancing Cryptography 2024

CompactTag: Minimizing Computation Overheads in Actively-Secure MPC for DNN.
Intel PrivateAI Workshop 2023 Fall

MPC-Pipe: An Efficient Pipeline Scheme for Semi-honest MPC Machine Learning.
Intel PrivateAI Workshop 2023 Spring

Teaching

Co-instructor University of Southern California
Advanced Topics on Micro-architecture *Fall 2024*

- A graduate-level class.
- 5 lectures on Trusted Execution Environments.
- 3 lectures on Oblivious RAM.

Teaching Assistant University of Southern California
Computer Systems Architecture *Spring 2024*

Co-instructor University of Southern California
Hardware Foundation for Machine Learning *Fall 2023*

- A new graduate-level course designed in ML system.

- 4 lectures on ML primitives.
- 2 lectures on Trusted Execution Environments.
- 2 lectures on Multi-party Computation.

Teaching Assistant University of Southern California
Advanced Topics on Micro-architecture *Spring 2022*

Teaching Assistant University of Southern California
Introduction to Computer Networks *Fall 2021*

Mentoring

WiSE Mentorship Program *Fall 2024*
Mentee: one first-year PhD student
This program is to provide guidance to the first-year women PhD student. I provide advice and guidance to adjust and thrive in their PhD career.

Summer High School Intensive in Next-Generation Engineering *Summer 2024*
Mentee: Two high school students
I advised two high school students during summer 2024, assigning them preliminary research projects on CPU simulators and recommendation models. By the end of the program, they successfully presented posters showcasing their research findings.

Directed Research *Fall 2022*
Mentee: Nancy Arora
I guided a student in characterizing Intel SGX 3.0, focusing on performance profiling and bottleneck identification. Through this project, she gained in-depth knowledge of Trusted Execution Environments and valuable hands-on research experience. This work played a crucial role in helping her secure a position at Qualcomm.

Academic Services

- Reviewer for HPCA 2025
- Artifact evaluation committee member for ASPLOS 2025
- Artifact evaluation committee member for ISCA 2024
- Student volunteer for ISCA 2023
- Student volunteer for HPCA 2021