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 — *August, 2019 – current*

University of Washington — Seattle, WA
Bachelor of Science in Electrical Engineering — *September, 2015 – June, 2019*

Research Interests

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

Internship Experience

Research Intern — Pasadena, CA
Meta AI — *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

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.

MPC-Pipe: An Efficient Pipeline Scheme for Semi-honest MPC Machine learning. **[under review]**
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***, 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.

Byzantine-Robust and Privacy-Preserving Framework for FedML. **[ICLR Workshop 2021]**
Authors:, Hanieh Hashemi, ***Yongqin Wang***, Murali Annavaram
This is a federated machine learning framework that provides security against a subset of malicious clients that may send inaccurate gradient data to undermine model accuracy and convergence, and information-theoretic data privacy clients cannot access other clients' gradients. Untrusted servers only can access encoded gradients. We provide a rigorous analysis to guarantee bounds of information leakage are infinitesimally small.

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. **[IEEE 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.

Academic Service

- Student volunteer for HPCA 2021
- Student volunteer for ISCA 2023
- AE committee member for ISCA 2024