YU ZOU

yuzou93@outlook.com LinkedIn Profile Google Scholar Profile (+1)6892751440
(+86)18548929351

EDUCATION

2017 – 2021	Ph.D. Computer Engineering Dissertation: FPGA-Augmented Secure Crash-Consistent Non-Volatile Memory *Advisor: Mingjie Lin, Amro Awad*	University of Central Florida USA
2015 – 2017	M.S. Computer Engineering *Advisor: Herman Lam, Alan D. George*	University of Florida USA
2011 – 2015	B.S. Electrical Engineering *Advisor: Yu Wang (Tsinghua University)*	Beihang University CN

WORK EXPERIENCE

2024 – Present	System Security Research Engineer *Job Duty: Secure cloud infrastructure R&D*	ByteDance, CN
2021 – 2024	Research Scientist *Job Duty: Heterogeneous trusted execution environment (TEE) research*	Alibaba, CN

RESEARCH INTERESTS

Secure Computer Architecture – Explore secure and trusted computer architecture design to safeguard data confidentiality, integrity, availability, and recoverability. Propose efficient microarchitecture designs to minimize the performance overhead incurred by the data protection mechanisms. [C1][C2][C3][C4][C6][C9][J1][J2][J3]

Heterogeneous Computing - Explore the utilization of heterogeneous devices, e.g., FPGA and GPU, for efficient acceleration of computation. Through hardware-software co-design and optimizing data layout, data movement, and processing parallelism, fully exploit the potential of accelerators and achieve near-optimal performance.
[C2][C3][C4][C6][C7][C8][C9][C10][C11][C12][J1][J2][J3][J4][R1]

Secure Data Processing System – Explore the system-level design of secure and trusted data processing systems by leveraging emerging heterogeneous devices. Propose practical, high-performance, and data privacy policy compliant on-premise or cloud processing systems. [C1][C5]

RESEARCH PROJECT

System Security Research Engineer @ ByteDance (2024)
Confidential AI Accelerator Security Assurance

- Led threat modeling and comprehensive security analysis for ByteDance's confidential AI products. Conducted an in-depth examination of the entire AI stack, spanning from the chip power supply chain to reliable machine learning products like Doubao and Volcano Model-as-a-Service.
- Evaluated the design and coding security of 8+ in-house confidential AI accelerators at ByteDance, facilitating the early remediation of design vulnerabilities.
- Shepherded the product security certifications for CC and FIPS, ensuring compliance and integrity.

System Security Research Engineer @ ByteDance (2024)
Hardware Root-of-Trust (HRoT) R&D

- Led ByteDance hardware root-of-trust (HRoT) design to harden server platform firmware secure booting.
- Led system infrastructure design, including remote attestation and public key infrastructures.
- Deployed HRoT on ByteDance servers of 4 generations, 10,000+ counts in total.

Research Scientist @ Alibaba (2021-2024)
Research on Heterogeneous Trusted Execution Environment (TEE) & Confidential Computing

- Proposed and designed the first cost-efficient CPU-FPGA heterogenous TEE design. The product achieved a 13.4x speedup over Intel SGX and has been integrated into Alibaba FPGA-as-a-Service (FaaS) cloud product.
- Designed a customized isolated FPGA TEE card for heterogeneously processing sensitive user data stored in Alibaba encrypted database. The product achieved 10x SQL query performance improvement over Intel SGX baseline.
- Authored one top-tier conference paper. [C1]

Graduate Research Assistant @ UCF (2019-2022)

YU ZOU

yuzou93@outlook.com LinkedIn Profile Google Scholar Profile (+1)6892751440
(+86)18548929351

FPGA-Secured Persistent Memory
- Led a student research group on a DARPA/SPAWAR funded project to explore efficient implementations of novel secure architectures through hardware prototyping.
- Explored to use FPGA as a transparent middleware to secure new-generation non-volatile main memory, protecting confidentiality, integrity, availability, and recoverability of data residing in persistent memory.
- Proposed efficient and hardware-friendly crash-consistent/inconsistent data structures to guarantee data.
- Mentored two undergraduate students to develop Linux drivers for efficient cacheable PCIe memory mapping
- Authored and co-authored 5 top-tier conference papers and 3 top-tier journal papers. [C2][C3][C4][C6][C9][J1][J2][J3]

Graduate Research Assistant @ UCF (2020)
FPGA-Accelerated NVMe SSD
- Designed and open-sourced the first RTL NVMe controller IP to directly connect NVMe SSD with FPGA.
- Proposed a new FPGA-accelerated framework to optimize OS access latency to NVMe SSD. NVMe access performance was improved by 1.5x over state-of-the-art Intel SPDK.
- Proposed a near-storage framework by incorporating an in-line FPGA accelerator on SSD datapath. By offloading streaming tasks into in-line FPGA, performance was improved by 1.9x to 782.5x.
- Authored 1 top-tier conference paper and 1 top-tier journal paper. [C7][J4]

Graduate Research Assistant @ UCF (2019)
Massively Simulating Adiabatic Bifurcations with FPGA to Solve Combinatorial Optimization
- Proposed an edge-centric graph-based simulated bifurcation algorithm to solve sparse Ising models.
- Designed a dedicated hardware architecture and proposed algorithmic optimizations to accelerate graph processing.
- Designed an FPGA-based combinatorial optimization problem solver and achieved 10.91x speedup over the state-of-the-art GPU baseline implementation.
- Authored 1 top-tier conference paper. [C8]

Graduate Research Assistant @ UCF (2018)
Exploiting Hidden Parallelism of Non-Stencil Computation in High-Level Synthesis
- Pioneered a new optimization direction to accelerate non-stencil kernel computing.
- Proposed a systematic and automatic way to construct graphs from a high-level non-stencil kernel code.
- Designed a workflow to regroup and reschedule data and computation by operating on constructed graphs with graph theory.
- Prototyped an end-to-end transformation tool which directly synthesizes a high-level code into a hardware architecture and deploys on the device. The proposed approach obtained 2-13x performance speedup over baseline.
- Authored 1 top-tier conference paper. [C10]

Graduate Research Assistant @ UF (2016-2017)
FPGA-Based Custom Memory Cube (CMC) Emulation Platform
- Explored the first hardware-in-the-loop emulation of custom memory cube.
- Developed a 3D-stacked DRAM custom logic emulator, accelerating development and validation of CMC logic, leveraging Micron HMC (Hybrid Memory Cube) techniques.
- Designed performance measurement and a mathematical model for the platform.
- Implemented DRE (data rearrangement/reordering engine), bloom filter, and sorting algorithms on the platform using Micron's hybrid-threading (HT) language/toolset and SystemC.

TEACHING

EEL4930/EEL5934 (Reconfigurable Computing)	Guest Lecturer, UF (Spring 2017) Gave one invited lecture introducing in-memory processing to undergraduate students.

PROFESSIONAL SERVICE

Reviewer	● Design Automatic Conference ● IEEE Transactions on Computers

PUBLICATION

Conference

[C1] **Zou, Y**. Salus: A Practical Trusted Execution Environment for CPU-FPGA Heterogeneous Cloud Platforms. *To appear in ASPLOS*

YU ZOU

yuzou93@outlook.com LinkedIn Profile Google Scholar Profile (+1)6892751440
(+86)18548929351

2025.

[C2] Shadab, R. M., **Zou, Y.**, & Lin, M. (2024, May). CTR+: A High-Performance Metadata Access Scheme for Secure Embedded Memory in Heterogeneous Computing Systems. In *2024 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)* (pp. 304-308). IEEE.

[C3] Shadab, R. M., **Zou, Y.**, Gandham, S., & Lin, M. (2023, May). OMT: A run-time adaptive architectural framework for bonsai merkle tree-based secure authentication with embedded heterogeneous memory. In *2023 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)* (pp. 191-202). IEEE.

[C4] Shadab, R. M., **Zou, Y.**, Gandham, S., & Lin, M. (2023, February). OMT: A Demand-Adaptive, Hardware-Targeted Bonsai Merkle Tree Framework for Embedded Heterogeneous Memory Platform. In *Proceedings of the 2023 ACM/SIGDA International Symposium on Field Programmable Gate Arrays* (pp. 47-47).

[C5] Wang, S., Li, Y., Li, H., Li, F., Tian, C., Su, L., ... & **Zou, Y**. (2022). Operon: An encrypted database for ownership-preserving data management. *Proceedings of the VLDB Endowment*, *15*(12), 3332-3345.

[C6] **Zou, Y.**, Awad, A., & Lin, M. (2021, December). Hermes: Hardware-efficient speculative dataflow architecture for bonsai merkle tree-based memory authentication. In *2021 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)* (pp. 203-213). IEEE.

[C7] **Zou, Y.**, & Lin, M. (2021, May). FERMAT: fpga-accelerated heterogeneous computing platform near nvme storage. In *2021 IEEE 29th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)* (pp. 262-262). IEEE.

[C8] **Zou, Y.**, & Lin, M. (2020, February). Massively simulating adiabatic bifurcations with FPGA to solve combinatorial optimization. In *Proceedings of the 2020 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays* (pp. 65-75).

[C9] **Zou, Y.**, & Lin, M. (2019, July). Fast: A frequency-aware skewed merkle tree for fpga-secured embedded systems. In *2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (pp. 326-331). IEEE.

[C10] **Zou, Y.**, & Lin, M. (2019, June). Graph-Morphing: exploiting hidden parallelism of non-stencil computation in high-level synthesis. In *Proceedings of the 56th Annual Design Automation Conference 2019* (pp. 1-6).

[C11] **Zou, Y.**, & Lin, M. (2018, December). GridGAS: an I/O-efficient heterogeneous FPGA+ CPU computing platform for very large-scale graph analytics. In *2018 International Conference on Field-Programmable Technology (FPT)* (pp. 246-249). IEEE.

[C12] **Zou, Y.**, & Lin, M. (2018, July). Very large-scale and node-heavy graph analytics with heterogeneous fpga+ cpu computing platform. In *2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)* (pp. 638-643). IEEE.

Journal

[J1] Shadab, R. M., **Zou, Y.**, Gandham, S., Awad, A., & Lin, M. (2023). A secure computing system with hardware-efficient lazy bonsai merkle tree for fpga-attached embedded memory. *IEEE Transactions on Dependable and Secure Computing*.

[J2] Shadab, R. M., **Zou, Y.**, Gandham, S., Awad, A., & Lin, M. (2023). Hmt: A hardware-centric hybrid bonsai merkle tree algorithm for high-performance authentication. *ACM Transactions on Embedded Computing Systems*, *22*(4), 1-28.

[J3] **Zou, Y.**, Zubair, K. A., Alwadi, M., Shadab, R. M., Gandham, S., Awad, A., & Lin, M. (2022). ARES: Persistently secure non-volatile memory with processor-transparent and hardware-friendly integrity verification and metadata recovery. *ACM Transactions on Embedded Computing Systems (TECS)*, *21*(1), 1-32.

[J4] **Zou, Y.**, Awad, A., & Lin, M. (2022). DirectNVM: Hardware-accelerated NVMe SSDs for high-performance embedded computing. *ACM Transactions on Embedded Computing Systems (TECS)*, *21*(1), 1-24.

Technical Report

[R1] Wang, G., Lam, H., **Zou, Y.,** Xavier, R., Gundecha, S., & George, A.D. (2016, Aug). A research platform for custom memory cube. *Workshop on Modeling & Simulation of Systems and Applications (ModSim).* University of Seattle, Seattle, WA.