Accelerating Fully Homomorphic Encryption via FPGA–CGRA Heterogeneous Architectures

Lingyu Gong[(1)]

[(1)] Bernoulli Institute and CogniGron, University of Groningen, The Netherlands

Fully Homomorphic Encryption (FHE) enables secure computation on encrypted data without decryption [1], offering strong potential for privacy-preserving applications, especially in machine learning and data analytics. Among various FHE schemes, CKKS is particularly suitable for approximate arithmetic and has been widely studied for privacy-preserving AI workloads [2]. However, the practical deployment of FHE is limited by its extremely high computational overhead, primarily due to intensive polynomial arithmetic and data movement.

While prior research has largely focused on algorithm-level optimization, key operations in FHE, including polynomial multiplication and Number Theoretic Transform, exhibit highly regular and repetitive computation patterns with significant data-level parallelism. These characteristics allow the decomposition of computations into uniform processing elements with predictable data access patterns, enabling efficient mapping onto parallel hardware architectures [3]. However, effectively exploiting these properties requires careful coordination between algorithm design and hardware architecture. To address this challenge, we adopt a hardware and algorithm co-design perspective, identifying key computational components in CKKS-based workloads that can benefit from architectural optimization in terms of parallelism, memory organization, and dataflow.

We explore the use of heterogeneous FPGA and CGRA architectures for accelerating FHE. FPGA platforms provide high flexibility and mature design ecosystems, while CGRA offers a promising balance between programmability and efficient execution of structured computations commonly found in AI-oriented workloads. By combining the strengths of both platforms, we aim to analyze and compare their performance, area time product, and energy efficiency, and to develop insights into novel hardware architectures for efficient FHE acceleration in privacy-preserving AI applications.

[1] C. Gentry et al., Proc. ACM STOC, 41, 169–178, 2009.

[2] J. Cheon et al., Proc. Asiacrypt, 10624, 409–437, 2017.

[3] Y. Gong et al., Cybersecurity, 7(1), 5, 2024.