2021 18th International SoC Design Conference (ISOCC) | 978-1-6654-0174-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISOCC53507.2021.9614034

Configurable Butterfly Unit Architecture for NTT/INTT in Homomorphic Encryption

Phap Duong-Ngoc, Tuy Nguyen Tan, and Hanho Lee
Dept. of Information and Communication Engineering
Inha University, Incheon, Korea
hhlee@inha.ac.kr

***Abstract*—This paper proposes a configurable architecture of butterfly unit (BU) supporting number theoretic transform (NTT) and inverse NTT (INTT) accelerators in the ring learning with error based homomorphic encryption. The proposed architecture is fully pipelined and carefully optimized the critical path delay. To compare with related works, several BU designs of different bit-size specific primes are synthesized and successfully placed-and-routed on the Xilinx Zynq UltraScale+ ZCU102 FPGA platform. Implementation results show that the proposed BU designs achieve 3× acceleration with more efficient resource utilization compared with previous works. Thus, the proposed BU architecture is worthwhile to develop NTT/INTT accelerators in advanced homomorphic encryption systems.**

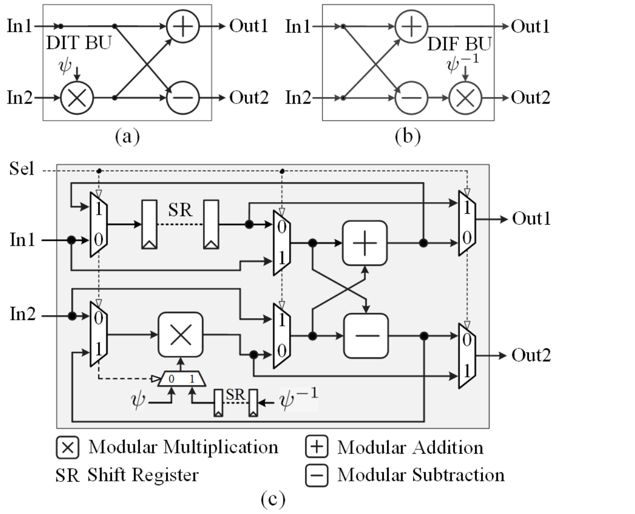
***Index Terms*—Number theoretic transform (NTT), homomorphic encryption, ring learning with error, butterfly unit**

I. Introduction

Ring learning with error (RLWE) based homomorphic encryption (HE) has emerged as an ideal solution to enable computation on encrypted data. These HE schemes can support fully homomorphic computations, in which polynomial multiplication is the most computational intensive operation. Number theoretic transform (NTT) and inverse NTT (INTT) are often utilized to accelerate the polynomial multiplication, and designing an efficient butterfly unit (BU) architecture is significant to develop NTT and INTT accelerators.

Modular multiplier is the most expensive operation in BU architectures. Kim *et al.* presented the FPGA implementation of various modular multipliers based on Barrett and Shoup algorithms [1]. Riazi *et al.* proposed the architectures of NTT and INTT cores in HEAX [2], a high performance accelerator for CKKS-based HE scheme on Intel FPGA devices. These cores' operations are based on the Shoup modular multiplication (MM) algorithm and require three expensive integer multiplications. Xin *et al.* have recently proposed a multi-functional BU structure of 40-bit prime for NTT, INTT, and MM [3]. Due to requiring extra control logic, their approach might not be pipelined efficiently and reduce the performance.

In this work, we propose a configurable BU architecture suitable for unified NTT/INTT accelerators in RLWE-based HE systems. The proposed BU architecture supports NTT and INTT computations using Cooley-Turkey (CT) decimal-in-time (DIT) and Gentleman-Sandy (GS) decimal-in-frequency (DIF) methods, respectively [4]. We utilize the Barrett MM algorithm for lightweight primes [1] to reduce the number of digital signal processing (DSP) slices. The unified design is realized with full pipeline and critical path delay optimization to improve the throughput. Compared with existing studies, the proposed BU design is valuable for configurable NTT/INTT accelerators in practical HE-based applications.



(a)

(b)

(c)

Fig. 1. Operations of (a) CT BU and (b) GS BU methods, and (c) proposed configurable BU architecture. Some of the pipeline registers are omitted for the sake of simplicity.

The rest of this paper is organized as follows. Section II describes the proposed configurable BU architecture. The implementation results and comparison are presented in Section III. Finally, Section IV gives the conclusion.

II. Proposed Configurable BU Architecture

Fig. 1 (c) shows the configurable BU architecture supporting CT BU and GS BU methods. The BU structure includes three major modular components: multiplier, adder, and subtractor. Additional multiplexers are added to select the execution order of these components. Control signal "Sel" is set to "0" for the CT BU operation and "1" for the GS BU operation.

In this work, we realize two configurable BU architectures of specific 40-bit and 60-bit primes. The integer multiplication

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1011232), and in part by the KEIT grant funded by the MOTIE, Korea (No. 20010589).

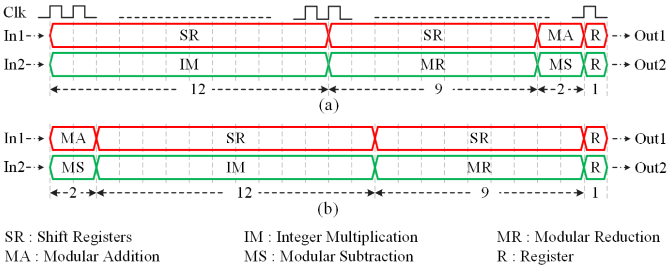


Fig. 2. Pipeline processing of (a) CT BU and (b) GS BU operations of 60-bit prime. Red and green flows illustrate the pipeline operations of positive and negative butterfly circuits, respectively. The numbers indicate the number of CCs of corresponding functions.

of specific 40-bit prime requires six DSP slices whereas that of specific 60-bit prime needs twelve ones. The modular reductions only use bit-shift and addition with specific primes. In this implementation, we take a sample of lightweight primes with four Hamming weight from the widely-used HE library namely Microsoft SEAL [5].

Fig. 2 (a) and (b) analyze pipeline processing of the proposed BU architecture for specific 60-bit prime. When the pipeline is fulfilled, the CT BU and GS BU operations have the same number of execution clock cycles (CCs). Some shift registers (SRs) are required to bypass the positive butterfly circuit. Output coefficients are generated every CC after a delay. The CC delay is calculated by the total cycles of integer multiplication, modular reduction, modular subtraction, and output register (e.g., 12 + 9 + 2 + 1 = 24 CCs).

III. Implementation Results and Comparison

We modeled the proposed BU architectures using Verilog HDL on Xilinx Vivado© tool (2020.1). The implementation results were placed and routed on Xilinx Zynq UltraScale+ ZCU102 (xczu9eg-ffvb1156-2-e) FPGA platform. The efficacy of the proposed BU designs were compared with previous works in terms of utilized resources and execution time. We reported the maximum clock frequency (short for $f_{\max}$) and calculated the latency by the number of CCs per $f_{\max}$.

Table I shows the resource consumption of the proposed configurable BU architecture of specific 40-bit prime compared with [3]. The proposed approach eliminates extra control logic for constant multiplication and multiply-accumulate operations, that reduces 23% of used look-up tables (LUTs) in [3]. Additionally, the modular multiplier requires only six DSP slices for the integer multiplication while more flip flops (FFs) are used for the modular reduction. Additional registers are required for bit-shift and pipeline operation, which are not described in detail in [3]. With careful optimization of the critical path, the proposed design achieves 3.2× higher clock frequency than that of [3].

Table II relatively compares the implementation results of the proposed BU architecture of specific 60-bit prime with the NTT and INTT cores of 54-bit prime in HEAX [2]. The second and third columns show that the proposed approach can save a lot of LUTs and FFs. The proposed approach requires a little more number of DSP slices as shown in the fourth column although the Shoup MM algorithm in [2] requires three integer multiplications. This difference comes from the Intel DSP slices supporting the 27 × 27-bit integer multiplication while the Xilinx DSP slices support 27 × 18-bit integer multiplication. The last two columns compare the maximum clock frequency and execution CCs of corresponding approaches. With higher clock frequency and fewer number of CCs, the proposed BU design achieves approximate 3× acceleration compared with NTT and INTT cores.

TABLE I
FPGA Resource Consumption of the Proposed 40-bit Configurable BU Design and [3] on Xilinx ZCU102 Platform.

Design	**LUT**	**FF**	**DSP**	$f_{\max}$**(MHz)**	**CCs**
Xin [3]	1037	336	15	150	-
This work	804	1257	6	479	17

TABLE II
Comparisons of the Proposed 60-bit Configurable BU Design with 54-bit NTT and INTT cores in [2].

NTT and INTT cores on Intel Arria 10 GX 1150 [2]					
Core	***ALM***[(*)]	***REG***	***DSP***	$f_{\max}$***(MHz)***	***CCs***
NTT	2066	6297	10	300	50
INTT	2119	5449	10	300	49

Configurable BU design on Xilinx Zynq UltraScale+ ZCU102					
Core	***LUT***	***FF***	***DSP***	$f_{\max}$***(MHz)***	***CCs***
This work	1242	2356	12	445	24

[(*)] Adaptive Logic Module (ALM) contains two combinational adaptive LUTs, a two-bit full adder, and four 1-bit registers.

IV. Conclusion

This work presented a configurable BU architecture for unified NTT/INTT accelerators in RLWE-based HE schemes. The proposed BU design effectively utilizes the hardware resources and achieves significant acceleration. The comparison results confirm that the proposed approach is worthwhile to further develop NTT/INTT accelerators for practical HE-based applications.

References

[1] S. Kim, K. Lee, W. Cho, J. H. Cheon and R. A. Rutenbar, "FPGA-based Accelerators of Fully Pipelined Modular Multipliers for Homomorphic Encryption", 2019 International Conference on ReConFigurable Computing and FPGAs (ReConFig), pp. 1-8, Dec. 2019.

[2] M. S. Riazi, K. Laine, B. Pelton, and W. Dai, "HEAX: High-Performance Architecture for Computation on Homomorphically Encrypted Data in the Cloud", The Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), New York, USA, pp. 1295–1309, Mar. 2020.

[3] G. Xin, Y. Zhao and J. Han, "A Multi-Layer Parallel Hardware Architecture for Homomorphic Computation in Machine Learning," IEEE International Symposium on Circuits and Systems (ISCAS), May. 2021, doi: 10.1109/ISCAS51556.2021.9401623.

[4] P. Duong-Ngoc, T. N. Tan, and H. Lee, "Efficient NewHope Cryptography Based Facial Security System on a GPU", IEEE Access, vol. 8, pp. 108158-108168, Jun. 2020.

[5] Microsoft SEAL (release 3.6), https://github.com/Microsoft/SEAL, Microsoft Research, Redmond, WA, Nov. 2020.