Energy-efficient NTT Design with One-bank SRAM and 2-D PE Array

Jianan Mu[1,2,3], Huajie Tan[4], Jiawen Wu[4], Haotian Lu[4], Chip-Hong Chang[5],Shuai Chen[6], Shengwen Liang[1,2], Jing Ye[1,2,3], Huawei Li[1,2,3], Xiaowei Li[1,2,3]
[1]State Key Laboratory of Computer Architecture, Institute of Computing Technology Chinese Academy of Sciences, [2]University of Chinese Academy of Sciences, [3]CASTEST, [4]Tianjin University, [5]Nanyang Technological University, [6]Rock-Solid Security Lab, Fiberhome
mujianan19s@ict.ac.cn, {hj_tan, wujiawen, ht_lu092}@tju.edu.cn, ECHChang@ntu.edu.sg, chenshuai_ic@163.com, {liangshengwen, yejing, lihuawei, lxw}@ict.ac.cn

***Abstract*—In Number Theoretic Transform (NTT) operation, more than half of the active energy consumption stems from memory accesses. Here, we propose a generalized design method to improve the energy efficiency of NTT operation by considering the effect of processing element (PE) geometry and memory organization on the data flow between PEs and memory. To decrease the number of data bits that are required to be accessed from the memory, a two-dimensional (2-D) PE array architecture is used. A pair of ping-pong buffers are proposed to transposed swap the coefficients to enable a single bank of memory to be used with the 2-D PE array to reduce the average memory bit access energy without compromising the throughput. Our experimental results show that this design method can produce NTT accelerators with up to 69.8% saving in average energy consumption compared with the existing designs based on multi-bank SRAM and one-bank SRAM with one-dimensional PE array with the same number of PEs and total memory size.**

***Index Terms*—Number Theoretic Transform, Low-power design, SRAM, Memory access.**

I. Introduction

Energy and power efficient Number Theoretic Transform (NTT) designs have been approached in two main directions. The first strategy focuses on increasing the energy efficiency of the PE module of different PE structures by, for instance, clock gating and operand isolation methods [1], [2], [3]. The second approach focuses on reducing the energy consumption of the memory module as memory accesses take up 50% to 80% of the overall energy budget of a typical NTT accelerator [4], [5]. For example, Nejatollahi et al. [6] introduced ReRAM for in-memory computation to minimize energy consumption by reducing memory accesses in NTT computation. The drawbacks of emerging memory technologies like ReRAM are the high defects, lower manufacturing yield and higher cost [7]. Therefore, more reliable and cost-effective SRAM is still a prevailing and preferred choice for NTT accelerator designs. In the high-speed pipelined NTT calculation, SRAM is accessed in every cycle [8]. At the architectural level, the energy dissipated by memory access in NTT computation is affected by two key factors, bit_{acc} and e_{bit}. bit_{acc} is the total number of the data bits that are read from and written into the SRAM. These data bits arise from the input coefficients and outputs of the butterfly operations. e_{bit} is the average memory access energy per data bit.

It is observed that the structure of the PE array has an indirect impact on bit_{acc}. There are two main layouts of PE array, a one-dimensional (1-D) chain of PEs [9], [10] and a two-dimensional (2-D) array of PEs with different depth d_{PE} and width w_{PE} [8], [11]. Within the 1-D PE array, the results of each PE are written into the SRAM directly in each stage, whereas within the 2-D PE array the results of the PEs are not written to the memory directly, but consumed within the layers of depth d_{PE} in each stage.

e_{bit}, on the other hand, is directly affected by the SRAM structure. With w_{PE} PEs working in parallel, $2w_{PE}$ coefficients are consumed concurrently in each cycle. Two methods that allow conflict-free memory access to the desired coefficients for the PEs in each cycle are proposed in [12], [13]. The first method partitions the coefficients used by the PEs in either 1-D or 2-D array structure into different SRAM banks [12]. The second method packs the coefficients used by the PEs in a 1-D array into one SRAM address [13]. Compared with the first method, the second method needs only one bank of SRAM to store all coefficients required by the NTT operation. Due to the dominant switching activities of address decoders during the memory access, the saving of e_{bit} over the first method can be significant with reduced address decoders for a single bank SRAM. However, unlike the first method, the original method only works for the 1-D PE array.

Motivated by the above analysis, this paper explores the general 2-D PE array structure and the "one-bank" memory access scheme to reduce both bit_{acc} and e_{bit} simultaneously to increase the energy efficiency of NTT computation.

II. Energy-efficient 2-D PE Array and One Bank Memory Access Scheme

A 2-D PE array with d_{PE} layers and $w_{PE} = 2^{d_{PE}-1}$ PEs on each layer is proposed in [8]. The transpose memory mapping can be generalized to higher d_{PE} stages using only one single memory bank. An efficient architecture to accomplish this is shown in Fig. 1. It consists of six main parts: a control module, a PE array, an SRAM module, two ping-pong buffers

978-3-9819263-7-8/DATE23/© 2023 EDAA

configured as 2-D register arrays MTX_0 and MTX_1, a twiddle factor memory, and interface logic. The control module provides the overall NTT operation sequencing. The SRAM stores the polynomial coefficients used for the NTT butterfly operations. The PE array has d_{PE} layers and $w_{PE} = 2^{d_{PE}-1}$ PEs on each layer. Each of the ping-pong buffers (MTX_0 and MTX_1) has $2w_{PE} = 2^{d_{PE}}$ rows and $2w_{PE} = 2^{d_{PE}}$ columns. The twiddle factors are pre-computed and stored in TF MEM. As only a pair of coefficients is read into a PE on layer 0 in each round, with w_{PE} PEs on layer 0, the width of the SRAM is set to $2w_{PE} \times \lceil log2(q) \rceil$, and the size of the SRAM is $N \times \lceil log2(q) \rceil$.

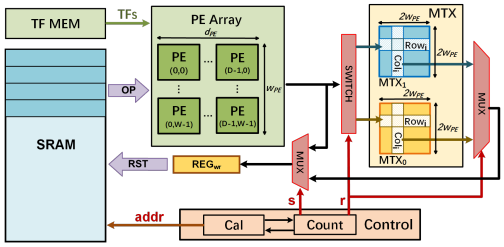

Fig. 1: Overall architecture.

III. Experimental Results

To evaluate and compare the energy consumption, NTT modules of different parameter settings of N, q, w_{PE} and d_{PE} are implemented in Verilog HDL. The energy consumption of SRAM module is evaluated by CACTI 7.0 [14] in 65nm CMOS process and the remaining modules by Synopsys Design Compiler in TSMC 65nm CMOS process.

Four main categories of NTT design methods are listed in Table I based on the memory access scheme and the PE array structure. Examples of existing works that fall in each of these categories are also listed. It is infeasible to draw useful conclusion and unbiased comparison on energy efficiency due to data transfer between PE array and memory for different categories of PE and memory configurations from experimental data reported in literature of existing NTT accelerators. Our comparison focuses on the energy efficiency of NTT computation due to the difference in memory and PE array configurations, i.e., 1-D versus 2-D PE array and single bank versus multiple banks of memory. To this end, we have implemented four NTT hardware modules, one for each category of Table I. For the 2-D PE array, $d_{PE} = 2$.

TABLE I: The categories of the NTT designs

Category	Work	Memory Scheme	PE Array
MB1D	[10], [15]	Multi-bank	1-D
MB2D	[8], [11]	Multi-bank	2-D
OB1D	[1], [2], [4]	One-bank	1-D
OB2D	This Work	One-bank	2-D

1) OB2D design versus OB1D design: Our proposed OB2D design saves on average 31.9% energy consumption compared with the OB1D design. The main saving comes from the SRAM dynamic power dissipation, which accounts for 56.6% of power reduction over that of the OB1D. This is because bit_{acc} of NTT implemented with a PE array of $d_{PE} = 2$ is half that of NTT implemented with a 1-D PE array.

2) OB2D design versus MB2D design: Our proposed design saves 35.4% of average energy consumption over the MB2D design. Specifically, the SRAM dynamic power has been reduced by 62.2% over that of the MB2D design. The reduction mainly comes from the reduced e_{bit}, due to our more energy efficient memory access scheme.

3) OB2D design versus MB1D design: Our proposed design saves 69.8% of average energy consumption over the MB1D design. The SRAM dynamic power dissipation has been reduced significantly by 86.6%. This is because 1) the use of 2-D PE array reduces bit_{acc} by half from the 1-D PE array and 2) e_{bit} is also reduced in one-bank address scheme due to its lower energy consumption cost in address decoding.

The above analysis corroborates that memory and PE array configurations play an instrumental role in the energy efficiency of NTT computation.

IV. Acknowledge

This paper is supported in part by the National Natural Science Foundation of China (NSFC) under grant No.(62090024, U20A20202). The corresponding authors are Jing Ye and Xiaowei Li.

References

[1] T. Fritzmann and J. Sepúlveda, "Efficient and flexible low-power NTT for lattice-based cryptography," in *2019 IEEE HOST*, 2019, pp. 141–150.
[2] U. Banerjee, T. S. Ukyab, and A. P. Chandrakasan, "Sapphire: A configurable crypto-processor for post-quantum lattice-based protocols," *IACR TCHES*, vol. 2019, no. 4, p. 17–61, Aug. 2019.
[3] W. Guo and S. Li, "Area-efficient modular reduction structure and memory access scheme for NTT," in *2021 IEEE ISCAS*. IEEE, 2021, pp. 1–5.
[4] N. Zhang et al., "NTTU: An area-efficient low-power NTT-uncoupled architecture for ntt-based multiplication," *IEEE TC*, vol. 69, no. 4, pp. 520–533, 2019.
[5] D. Li, A. Pakala, and K. Yang, "MeNTT: A compact and efficient processing-in-memory Number Theoretic Transform (NTT) accelerator," *IEEE TVLSI*, vol. 30, no. 5, pp. 579–588, 2022.
[6] N. Hamid et al., "CryptoPIM: in-memory acceleration for lattice-based cryptographic hardware," in *2020 ACM/IEEE DAC*. IEEE, 2020, pp. 1–6.
[7] Y. Chen, "ReRAM: History, status, and future," *IEEE TED*, vol. 67, no. 4, pp. 1420–1433, 2020.
[8] J. Mu et al., "Scalable and Conflict-free NTT Hardware Accelerator Design: Methodology, Proof and Implementation," *IEEE TCAD*, 2022.
[9] S. S. Roy et al., "FPGA-based high-performance parallel architecture for homomorphic computing on encrypted data," in *2019 IEEE HPCA*, 2019, pp. 387–398.
[10] N. Zhang et al., "Highly efficient architecture of NewHope-NIST on FPGA using low-complexity NTT/INTT," *IACR TCHES*, vol. 2020, no. 2, p. 49–72, Mar. 2020.
[11] X. Chen et al., "CFNTT: Scalable radix-2/4 NTT multiplication architecture with an efficient conflict-free memory mapping scheme," *IACR TCHES*, vol. 2022, no. 1, p. 94–126, Nov. 2021.
[12] L. Johnson, "Conflict free memory addressing for dedicated FFT hardware," *IEEE TCAS-II*, vol. 39, no. 5, pp. 312–316, 1992.
[13] S. S. Roy et al., "Compact Ring-LWE cryptoprocessor," in *2014 IACR CHES*, L. Batina and M. Robshaw, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 371–391.
[14] B. Rajeev et al., "CACTI 7: New tools for interconnect exploration in innovative off-chip memories," *ACM TACO*, vol. 14, no. 2, pp. 1–25, 2017.
[15] H.-F. Lo, M.-D. Shieh, and C.-M. Wu, "Design of an efficient FFT processor for DAB system," in *2001 IEEE ISCAS*, vol. 4, 2001, pp. 654–657 vol. 4.