An eDRAM-Based In-Situ-Computing Processor for Homomorphic Encryption Evaluation on the Edge

Luchang Lei[1], Yongqing Zhu[2*], Xinhang Zou[1*], Yifan He[1], Zhou Zhang[2], Zhenyu Guan[2], Huazhong Yang[1], Yongpan Liu[1], Song Bian[2], and Hongyang Jia[1†]
*Equally Credited Authors. [1]Tsinghua University, Beijing, China, [2]Beihang University, Beijing, China (Contact Email: hjia@tsinghua.edu.cn)

2024 IEEE European Solid-State Electronics Research Conference (ESSERC) | 979-8-3503-8813-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ESSERC62670.2024.10719576

***Abstract*— This work presents an in-situ-computing accelerator for homomorphic encryption (HE) evaluation on the edge, with eight eDRAM-based in-situ HE processing cores integrated with RISC-V CPU and HE-specialized NoC. The key features are: (1) the architecture-level fusion with customized ISA chains HE operators, avoiding excessive intermediate ciphertext movements; (2) the circuit-level fusion of eDRAM bitcells and dynamic digital datapath enables energy-efficient in-memory HE evaluation with high storage density; (3) intra-/inter-core hybrid automorphism network greatly reduces communication overheads. The silicon prototype in 28nm CMOS achieves 2.33 Mb/mm^2 HE storage density, with peak energy efficiency of 329 nJ/NTT at 4096-points-19-bit and 15.1 classification per second with a linear support vector machine in CKKS-RNS scheme, exceeding previously reported ASIC works and CPUs.**

Keywords*— *Homomorphic Encryption, In-Situ Computing, eDRAM

I. Introduction

Homomorphic Encryption (HE) is a promising candidate for privacy-preserving computation, where untrusted parties can directly execute arbitrary programs over data encrypted under HE without decryption, making the data available but invisible. However, the number-theoretic nature of HE introduces magnitudes of storage and computation inflation over ciphertexts. These have motivated pioneer research on hardware acceleration of the encryption, decryption, and evaluation of HE.

Fig. 1 (a) shows a typical protocol of HE-based privacy-preserving outsourcing computation, in which input data and inference models should be kept private. The data owner encrypts input data and sends it to the model owner. The model owner executes the inference model and sends back the encrypted results, without access to the original data and the result due to the cryptographic nature of HE algorithms. However, even for the parallelism-friendly HE schemes packing multiple plaintexts in one polynomial using Ring Learning With Error (RLWE) [1], ciphertext has a ~40x expansion, while this number goes to 10^3~10^4 for other schemes [2]. Thus, data transmission between the parties becomes one of the major bottlenecks for efficient HE protocol execution [3], restricting HE applications in heavy communication- and energy-restricted IoT scenarios.

Fig. 1 (b) shows an alternative for IoT scenarios with a homomorphic secret sharing procedure. For lightweight inference models with fewer parameters such as SVM and PCA, instead of communicating massive encrypted input data to the server, the encrypted model is sent to the data owner once and stored in the edge client of the data owner. For each inference task, HE evaluation is performed locally at the edge without expensive out-of-domain communication. A homomorphic secret sharing procedure based on other crypto schemes like AES is attached at the end of the protocol to ensure data privacy without massive data expansion [4].

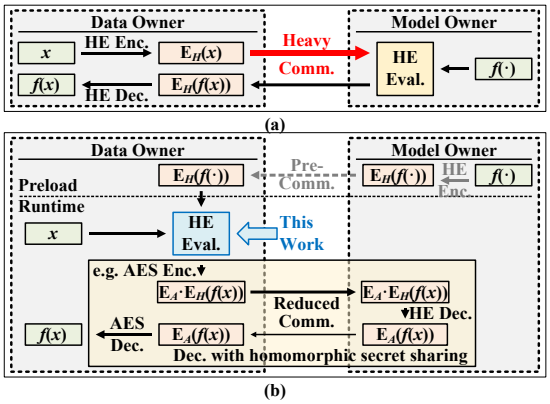


Fig. 1. Dataflow of (a) HE-based outsource privacy-preserving computation; (b) HE with homomorphic secret sharing procedure.

While homomorphic secret sharing is a mature technology with many practical solutions [3], HE evaluation on edge devices still faces the "memory wall" issue. In particular, expensive hierarchical memory access prevents HE evaluations from deployment at usable speed and a feasible power budget [5], which necessitates extreme hardware acceleration with higher local storage density near compute and better energy efficiency.

This paper introduces an eDRAM-based in-situ accelerator customized for HE evaluation in homomorphic secret sharing applications on the edge, featuring architecture-level operator fusion which helps to chain HE operators in place. The proposed accelerator also supports HE encryption/decryption, enabling more general tasks with traditional outsourced HE computation.

II. Operator Fusion in HE Evaluation

Fig. 2 (a) illustrates a typical operator sequence in HE evaluation, in which the output of the former HE operator goes straight to the next. Following the conventional organization of the von Neumann architecture, many prior ASIC works employ a hierarchical buffering system with fixed function units dedicated to each HE operator [6]. As shown in Fig. 2 (b), different function units are attached to a higher-level buffer. Massive inter-operator data transmission storing and loading the local buffer dominates energy consumption, limiting the feasibility of those designs on energy-limited edge devices.

The highly parallel dataflow with less dimension reduction in the ciphertext domain motivates the chaining of HE operators in an in-situ computing scheme. As illustrated in Fig. 2 (c), with the processing unit able to execute various HE operators, cascaded dataflow can be performed without energy-consuming inter-unit data

This work was partly supported by the National Key R&D Program of China 2023YFB3106200 and NSFC Grant 62374101, 62202028.

transmission. This work proposed an In-Situ HE Processing Core (ISHEPC), in which the complex HE operators are decomposed into basic operators enabling in-place processing of large-dimensional operation series with minimal external data movement. It also applies an eDRAM-based near-computing-storage design to achieve energy-efficient and high-density in-situ processing (ISP).

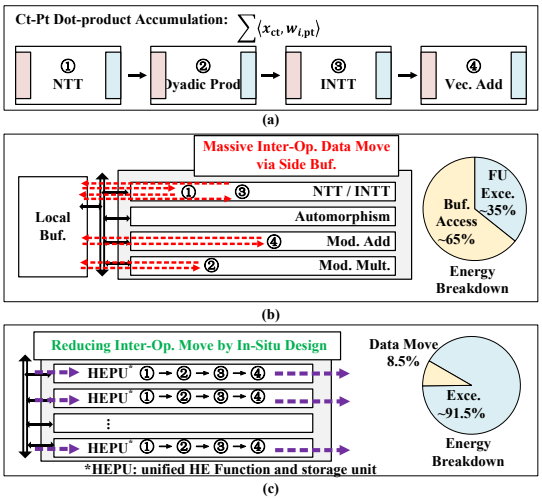


Fig. 2. (a) A typical cascade HE dataflow, computing a ciphertext-plaintext dot-product accumulation in SVM inference; (b) traditional architecture with separated function units and local buffer; (c) proposed in-situ architecture design.

There are three major challenges for designing the proposed in-situ architecture: (1) scheduling of massive decomposed operators increases controlling cost; (2) highly parallel operations require huge spatial logic, degrading the effective local storage density; (3) the complex structured data moving in NTT operations raise communication bottlenecks.

To mitigate the aforementioned challenges, the proposed ISHEPC is characterized by: (1) fine-grained operator decomposition with HE-specific ISA design for efficient computing scheduling control; (2) the fusion of eDRAM local storage with dynamic-logic-based datapath forms a dynamic in-memory processing engine (DIMPE) with significantly improved local storage density and parallelism. (3) the flexible intra-/inter-core Hybrid Automorphism Network (HAN) performs high bandwidth in-memory swapping and scalable exchange, alleviating communication overheads for large-scale automorphism.

III. Architectural Overview

The demonstrated eDRAM-based ISP HE evaluation accelerator incorporates architecture- and circuit-level fusion enabling energy-efficient high-density in-situ HE processing to overcome the aforementioned challenges. Fig. 3 illustrates the overall architecture of the proposed design. The eight ISHEPCs are the key components of the demonstrated HE accelerator, with a total of 1.2Mb local storage for HE operands, supporting a maximum polynomial dimension of 4096. A custom instruction decoder with a 4kB instruction memory holds instructions with an HE-specialized RISC-style ISA. The prototype SoC also comprises a 48kB HE data buffer, and a RISC-V CPU [7] with peripherals as testing interface and on-chip controller for light-weight HE tasks.

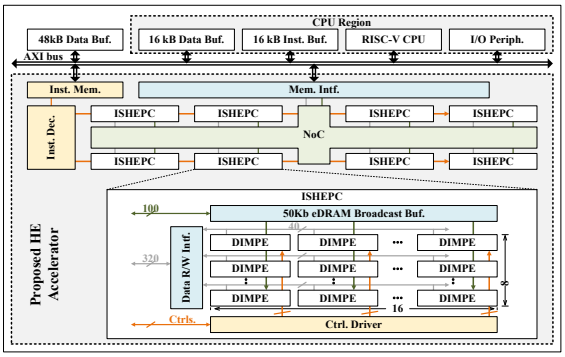


Fig. 3. Architectural overview.

Each ISHEPC comprises 8×16 DIMPEs cascaded by hierarchical read and write bit-lines and word-lines, forming a unified eDRAM memory array with interface on one side. All ISHEPCs execute simultaneously, making the total computation parallelism 1024, significantly higher than the prior near-memory computation work [8]. A 50kb eDRAM-based broadcast buffer is attached to support vertical operand reuse, such as twiddle factors in (I)NTT operators, as marked in the green lines in Fig. 3.

IV. Microarchitecture Design

A. HE Operator Decomposition and ISA Design

The design of ISHEPC firstly decomposes HE operators into base compute and data movement instructions. Those base instructions are further split into atomic logic for higher hardware utilization, as illustrated in Fig. 4 (a) and (b). The operator decomposition enables programmable performing of highly parallel complex HE functions using a unified datapath, and thus, leads to in-situ HE evaluation at the place the polynomial generated.

Instructions		HE Operator						Shared Atomic Logic				
		Mod. Add.	Mod. Mult.	NTT	Boot.	…	eDRAM Refr.	XOR	AND	OR	SFT	BUF
Compute	Add		√	√	√			√	√	√		
	Invert	√	√	√	√			√				
	Multiply		√	√	√				√			
	Truncate		√	√	√				√			
	Shift		√	√	√						√	
	Reset	√	√	√	√							√
Local Data Move		√	√	√	√							√
eDRAM Refresh							√					√

(a)

Add
Cycle 1:
$Cell_0 \leftarrow Cell_0$ XOR $Cell_1$
$Cell_1 \leftarrow Cell_0$ AND $Cell_1$
Cycle 2:
$Cell_0 \leftarrow Cell_0$ XOR C_{in}
$C_{out} \leftarrow (Cell_0$ AND $C_{in})$ OR $Cell_1$

(b)

Instruction bit index	15	14	13-8	7	6	5-0
Long Instruction	1	0	long inst.			
Short Instruction	0	0	short inst. 1	0	x	short inst. 0

Long Instructions:
- Local Data Movement
- Branch and Control

Short Instructions:
- In-Situ Computation

(c)

Fig. 4. (a) Operator decomposition and shared atomic logic; (b) example of logic sharing, executing $Cell_0 \leftarrow Cell_0 + Cell_1$; (c) 16b ISA design with compressed instruction format.

An HE-specific RISC ISA is proposed to alleviate control costs while maintaining programmability. The 16b ISA is designed in a compressed instruction format as shown in Fig. 4 (c), supporting both in-situ computation, data movement, and flow control. Enabled by the ISA's programmability, the proposed HE accelerator executes most polynomial operations with a native configuration of 4096-point-19-bit, enough for lightweight CKKS-RNS HE evaluation, encryption, and decryption. Furthermore, the programmability from the fine-grained instructions helps extend the configuration for bit-width greater than 19, by spitting wide numbers and composing native operations into wider ones, e.g. 38b modular multiplication. The polynomial dimension can also be extended with higher bit support.

B. Dynamic In-Memory Processing Engine

Fig. 5 (a) represents the circuit-level design of DIMPE, which is the fused storage and computing engine of ISHEPCs. Each DIMPE includes 40×21 bit-cells and programmable dynamic logic along the vertical side, forming a basic unit for in-situ in-memory HE computation. During execution, stored bits are loaded on RBLs and straightly sent to dynamic logic. While computing results are driven on WBLs. Thanks to the structural similarity between eDRAM cells and dynamic logic, the storage and computation are fused in an area-efficient layout, maximizing storage and compute density. DIMPE natively supports 38b add/sub and 19b bit-serial multiply, with the higher bit-width operation performed by stitching over multiple eDRAM entries.

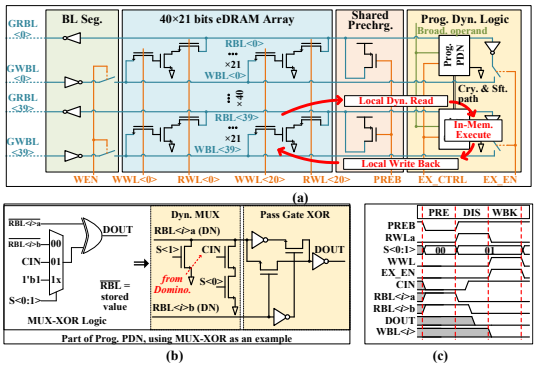


Fig. 5. (a) Circuit design of DIMPE; (b) example of the fused dynamic-based datapath; (c) waveform example, compute *1* at RBL<*i*>a (load from eDRAM by RWLa) XOR *1* at RBL<*i*>b (from adjacent PDN by Domino logic) and write result *0* at WBL<*i*> (to eDRAM chosen by WWL).

The atomic logic operations are implemented with dynamic, domino, and pass-gate logic to ensure a minimum digital logic area, as shown in Fig. 5 (b). It fuses the pull-down networks (PDNs) with the read circuitry of 3T eDRAM cells sharing dynamic nodes to enhance the storage density of DIMPE further. The fused dynamic-based digital datapath shows 3× density gain over standard cell implementation. Furthermore, the eDRAM dynamic load sequence and dynamic logic executing are fused into 3 phases as demonstrated in Fig 5 (c).

C. Intra-/Inter-Core HAN for Fast Automorphism

As shown in Fig. 6 (a) and (b), the proposed intra-/inter-core HAN enables in-memory swap and high-bandwidth exchange. This helps ISHEPC to handle local and long-range butterflies in staged (I)NTT computation and matrix-transpose operations in automorphism.

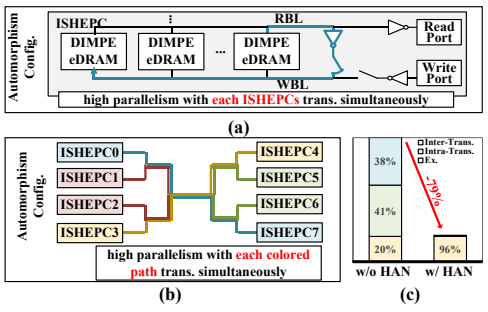


Fig. 6. (a) Intra-core transmission via in-memory swapping; (b) flexible inter-core transmission via NoC design. (c) latency breakdown of a typical NTT task and the latency saving of HAN.

For native 4096-point-19-bit configuration, staged (I)NTT operator is organized into 32-16-8 divisions, enabled by the proposed HAN design. Each DIMPE can execute 32-point (I)NTT solely. The next 16-point data swapping is realized with an intra-core transmission path via local BL write back, making 512-point (I)NTT executable in a single ISHEPC. 8 ISHEPCs compute 4096-point (I)NTT by inter-core swapping via the proposed NoC design. As shown in Fig. 6 (c), the high parallel HAN reduced the latency of NTT by 79% compared to swapping all data via the CPU memory interface, making the (I)NTT procedure computing bound.

V. Measurements and Demonstrations

The silicon prototype chip is fabricated in a 28nm HKMG CMOS technology, as shown in Fig. 7. The 8 ISHEPCs occupy 0.518mm^2 with a total capacity of 1240kb, enough to store 8 polynomials of 4096-dimension-19-bit configuration, without the help of on-chip HE data buffer. The architecture- and circuit-level fusion achieves a 3.3× gain on the FoM considering both storage density and energy efficiency, compared to a 12nm simulation-based ASIC design [6], and a 65× gain over a 65nm simulation-based NMC design [8].

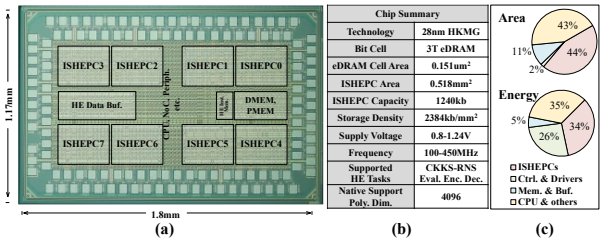


Fig. 7. (a) Die photo; (b) chip summary; (c) area and energy breakdown, measured under consecutive NTT task workload with 0.9V voltage.

Fig. 8 shows the voltage scaling of the chip and retention attribute of the eDRAM cell. The chip works at 100-450MHz with 0.6-1.2V for digital and 0.8-1.24V for ISHEPC. The energy efficiency varies from 0.5 to 3.04 MOPs/W in different voltage-frequency situations. The retention time of the eDRAM cell introduces negligible parallel refresh overhead at room temperature.

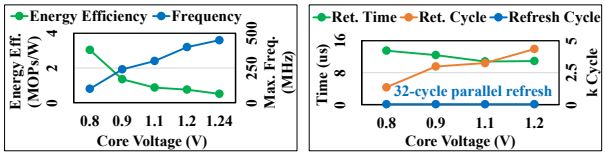
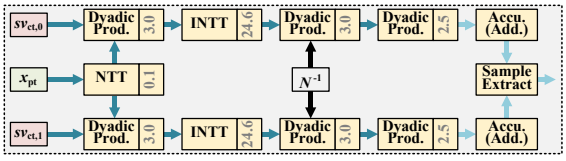


Fig. 8. (a) Voltage scaling, *Op* refers to an NTT task with 4096-point-19-bit configuration; (b) retention time, measured at room temperature.

Fig. 9 shows the demonstration task dataflow. For the proposed IoT HE evaluation on the edge, we evaluated the SVM inference with linear kernels, 512 support vectors, and a feature size of 512. The CKKS-RNS configuration is set as 4096-point-111-bit, minimizing computation complexity without losing accuracy. Experiments show that SVM with our parameters reaches 85.5% accuracy in human face classification tasks on the LFW dataset [9], which is the same as the plaintext evaluation result. The proposed accelerator can also execute CKKS-RNS encryption/decryption for traditional outsource scenarios,

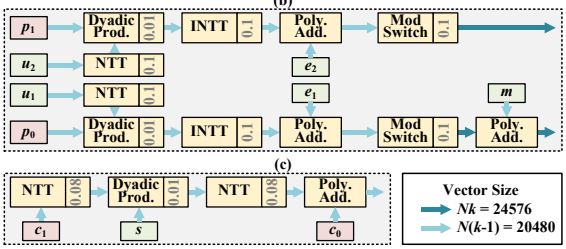
making the accelerator useful in more complex tasks. The comparison of HE encryption/decryption to a state-of-the-art ASIC work is also reported.



(a)

SVM kernel	This work		CPU	
	Linear	Quadratic	Linear	Quadratic
Peak Throughput (Class./s)	15.1 (155×)	0.583 (227×)	0.0971	0.00257
Peak Area Efficiency (Class./(s·mm²))	29.1 (7.5k×)	1.12 (10.9k×)	0.00388	0.000103
Peak Energy Efficiency (mJ/Class.)	1.36	35.2	-	-

(b)



(c)

(d)

Task	This work		ESSCIRC'23 [10]	
	CKKS		BFV	
	Enc.	Dec.	Enc.	Dec.
Peak Throughput (Op/s)	2.34k	5.89k	3.19k	5.10k
Peak Area Efficiency (Op/(s·mm²))	4.52k	11.4k	1.89k	3.02k
Peak Energy Efficiency (uJ/Op)	8.77	3.49	10.33	6.45

(e)

Fig. 9. Proposed ISP accelerator demonstration with execution time in milliseconds for CKKS-RNS with 4096-dimension-111-bit configuration and 6 RNS primes of [19, 19, 19, 18, 18, 18] bits. Negligible execution time like *Poly. Add.* is not shown. (a) linear-kernel SVM classification demonstration with 512 support vectors; (b) comparison with an i9-10920X single-core CPU with 64GB RAM; (c) encryption demonstration; (d) decryption demonstration; (e) comparison with [10].

VI. Conclusions

This work presents an eDRAM-based in-situ computing accelerator for HE evaluation on the edge, featuring the execution of multiple HE operators in place without excessive intermediate movements. It employs HE operator decomposition with specialized ISA, and circuit-level fusion of eDRAM and dynamic logic, achieving 2.33Mb/mm^2 local storage density and 329nJ-1.96uJ/NTT energy efficiency. The programmable accelerator supports HE encryption/decryption and more importantly, evaluation of CKKS-RNS on the edge. A facial classification demonstration shows 15.1 classifications per second and 1.36 mJ per classification.

Acknowledgment

The authors thank Prof. N. Sun (THU) for the support with testing.

References

[1] J. H. Cheon *et al*., "Homomorphic encryption for arithmetic of approximate numbers." Advances in Cryptology–ASIACRYPT 2017: 23rd International Conference on the Theory and Applications of Cryptology and Information Security, Hong Kong, China, December 3-7, 2017, Proceedings, Part I 23. Springer International Publishing, 2017.

[2] G. Shi *et al*., "A 28nm 68MOPS 0.18μJ/Op Paillier Homomorphic Encryption Processor with Bit-Serial Sparse Ciphertext Computing." 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 242-244.

[3] J.-L. Watson, W. Sameer, and R. A. Popa, "Piranha: A GPU platform for secure computation." 31st USENIX Security Symposium (USENIX Security 22). 2022.

[4] C. Juveka, V. Vaikuntanathan, and A. Chandrakasan, "GAZELLE: A low latency framework for secure neural network inference." 27th USENIX security symposium (USENIX security 18). 2018.

[5] R. Agrawal and A. Joshi, "On architecting fully homomorphic encryption-based computing systems." Springer International Publishing, 2023.

[6] N. Samardzic *et al*., "F1: A fast and programmable accelerator for fully homomorphic encryption." MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture. 2021.

[7] P. D. Schiavone *et al*., "Slow and steady wins the race? A comparison of ultra-low-power RISC-V cores for Internet-of-Things applications." 2017 27th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS), Thessaloniki, Greece, 2017, pp. 1-8.

[8] D. Li, A. Pakala and K. Yang, "MeNTT: A Compact and Efficient Processing-in-Memory Number Theoretic Transform (NTT) Accelerator." in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 30, no. 5, pp. 579-588, May 2022.

[9] G. B. Huang *et al*., "Labeled faces in the wild: A database for studying face recognition in unconstrained environments." Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition. 2008.

[10] S. Dase *et al*., "A 10.33 μJ/encryption Homomorphic Encryption Engine in 28nm CMOS with 4096-degree 109-bit Polynomials for Resource-Constrained IoT Clients." ESSCIRC 2023- IEEE 49th European Solid-State Circuits Conference (ESSCIRC), Lisbon, Portugal, 2023, pp. 193-196.

[11] Y. Yang *et al*., "Poseidon: Practical Homomorphic Encryption Accelerator." 2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA), Montreal, QC, Canada, 2023, pp. 870-881.

[12] H. Lee, H. Kwon and Y. Lee, "16.1 A 2.7-to-13.3μJ/boot/slot Flexible RNS-CKKS Processor in 28nm CMOS Technology for FHE-Based Privacy-Preserving Computing." 2024 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2024, pp. 296-298.

TABLE I. Comparison Table

		[6] MICRO'21	[8] TVLSI'22	[11] HPCA'23	[2] ISSCC'23	[12] ISSCC'24	[10] ESSCIRC'23	This Work
Technology		14/12nm Sim.[4]	65nm Sim.[4]	16nm FPGA	28nm Tapeout	28nm Tapeout	28nm Tapeout	28nm Tapeout
Track		Digital ASIC	SRAM NMC	-	Digital ASIC	Digital ASIC	Digital ASIC	eDRAM In-Situ
Die Area (mm²)		151.4	0.36	-	42.96	11.28	2.56	1.17
Core Area (mm²)		63.52	0.35	-	34.37	10.54	1.69	0.52
Storage in HE Function Unit (kb)		65536	162.4	70448	2304	1984	3520	1240
Voltage (V)[1]		0.8	1.2	0.8	0.8-1.0	1.0	0.64-1.1	0.8-1.24[3]
Frequency (MHz)		1000	151	450	500	333	157	100-450
Supported HE Tasks		Eval.	NTT	Eval.	Enc., Dec., Eval.	Eval.	Enc., Dec.	Eval., Enc., Dec.
Local HE Storage Density (Mb/mm²)		1.01	0.45	-	0.065	0.18	0.92	2.33
NTT[2]	Peak Throughput (Op/s)	448M	9.09k	5.73k	-	-	24.2k	13.9k-62.5k
	Peak Area Eff. (Op/(s·mm²))	7.05M	26.0k	-	-	-	14.3k	26.8k-121k
	Peak Energy Eff. (J/Op)	440n	4.17u	-	-	85.6n	1.36u	329n-1.96u
	FoM[5] (Mb/mm²·Op/J)	2.30M	108k	-	-	2.10M	676k	1.19M-7.08M

1. Voltage supply for FPGA work is not reported, thus estimated from its nominal voltage
2. NTT normalized to 4096 points 19 bits
3. Digital voltage minimum at 0.6V
4. Logic synthesized result for Digital and SRAM, schematic level simulation for analog and memory
5. FoM = Local HE Storage Density / Peak Energy Eff.