A 16nm 0.042mm^2 0.66μJ/Ops Lightweight MLWE PQC KEM with Cryptanalysis-ASIC Co-optimization

Author's Version

Archisman Ghosh*[1], Ming-Che Li*[1], Lingke Ding[1], Suparna Kundu[2], Abdullah Sayeed[1], Angshuman Karmakar[3], Harshit Naman[1], Sarthak Antal[1], Ingrid Verbauwhede[2] and Shreyas Sen[1]

[1] Purdue University, West Lafayette, IN
{ghosh69,li4509,ding359,sayeeda,hnaman,santal,shreyas}@purdue.edu
[2] COSIC, KU Leuven, Kasteelpark Arenberg 10, Bus 2452, B-3001 Leuven-Heverlee, Belgium
{suparna.kundu,ingrid.verbauwhede}@esat.kuleuven.be
[3] Department of Computer Science and Engineering, IIT Kanpur, India
angshuman@cse.iitk.ac.in

Abstract. Resource-limited devices demand post-quantum cryptography (PQC) within stringent power, area, and memory limits. This work introduces the first Cryptanalysis-ASIC co-optimized PQC IC, achieving minimal area, power, and memory footprint. The approach delivers 8×–509× area reduction with respect to the state-of-the-art M-LWE accelerator and generic PQC processor, respectively. Fabricated in Intel16, this IC operates from 0.1 MHz to 1.03 GHz and achieves 8.5× lower energy-area product.

Keywords: Post-Quantum Cryptography · Key-Encapsulation Mechanism · Lightweight Cryptography · Lattice-Based Cryptography · ASIC · IC

1 Introduction

Resource-limited devices such as IoT edge devices, medical implants, wearables, etc., are increasingly intertwined with sensitive personal data, ranging from audio & video streams to physiological data and sensitive payment credentials. These devices form the backbone of emerging digital infrastructure, but their adoption remains hindered by the absence of advanced cryptographic primitives that can simultaneously ensure post-quantum security and meet stringent hardware resource budgets. Existing post-quantum cryptographic (PQC) algorithms [1, 2], designed for general-purpose platforms usually require large logic instances (>M gate instances), extensive memory (>100 KB), and significant energy, which is not suitable for resource-limited nodes (e.g. < hundreds of microwatts, <100k gate instances). This challenge is compounded by the siloed development of algorithm and hardware design. Most circuit designers primarily implement algorithms designed by cryptographers, the National Institute of Standard and Technology (NIST), or other standard bodies and try to optimize them from implementation perspective. However, algorithmic parameter choices disproportionally impact power, area, and memory footprint. Therefore, without Cryptanalysis-ASIC co-optimization, integrating PQC into the resource-limited device efficiently is highly impractical.

*Starred authors contributed equally to this research.

NIST has recently (2025) standardized ASCON [3] for lightweight devices as an alternate to AES (symmetric key encryption) and Keccak used as a hash function, extended output function (XOF) or pseudo-random number generator (PRNG). A similar evolution for PQC Key Encapsulation Mechanism (KEM) for lightweight devices, minimizing area, power and memory is anticipated (Fig. 1).

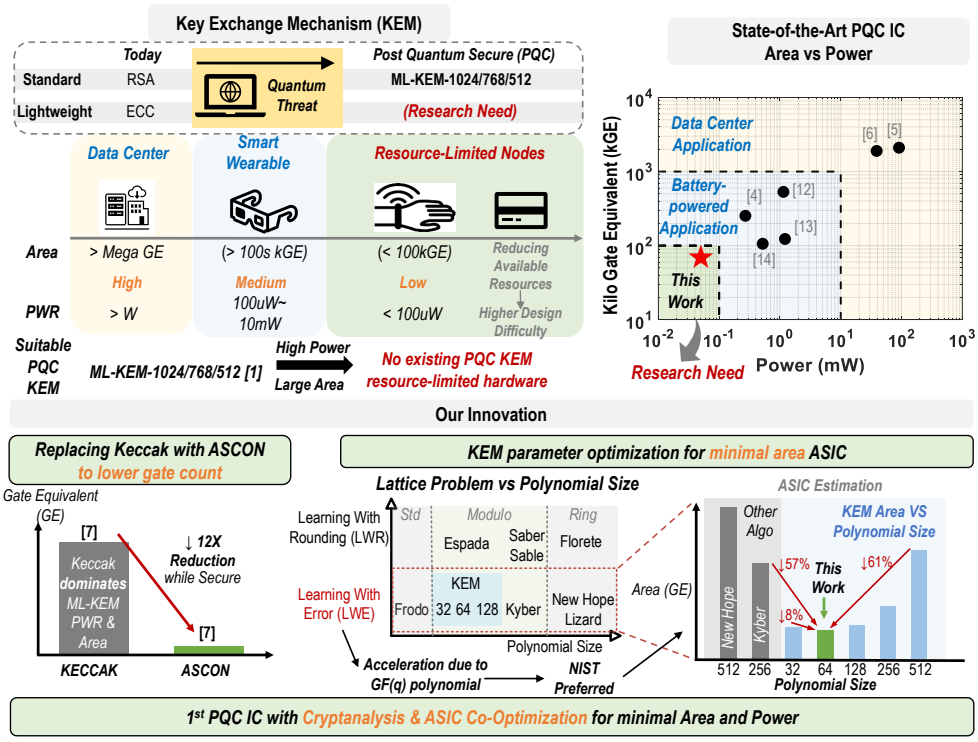


Figure 1: Challenges of PQC design for resource-constrained devices. This work designs a lightweight crypto core by using the lightweight standard ASCON-based PRNG/XOF/Hash function, KEM parameter optimization for minimum area, and is the first PQC IC with cryptanalysis and ASIC co-optimization for low area and power.

The smallest area in the State-of-the-Art (SOTA) PQC core [4] Kyber was achieved through digital circuit optimization. However, it still relied on the resource-intensive Keccak for PRNG, XOF, and hashing due to strict adherence to NIST standard parameters, leading to significant area & power overhead, with 0.34 mm^2 area [4]. While massively parallel datapaths lead to agile PQC processors [5, 6], the associated logic instances exceed 2 MGE and energy above 3 μK per operation, making them prohibitively unsuitable for resource-limited devices. In contrast, this IC, fabricated in Intel16, introduces an algorithm-IC co-optimization flow for minimum area implementation that replaces Keccak with ASCON (Fig. 1), explores parameters of Module Learning with Error (MLWE) beyond the standard sets while being quantum secure, employs a single-butterfly datapath with efficient modulo arithmetic, and minimizes memory via tiled data processing. These innovations collectively yield active area of 0.042 mm^2, 8x better than the current SOTA Kyber accelerator [4], and 509x better than generic SOTA PQC processor [5] while achieving 0.028 $\mu K/op \cdot mm^2$ Energy-Area-Product (EAP), 8.5x improvement compared to [4], with a wide operation frequency range of 100 kHz-1.03 GHz. To date, this work is the smallest PQC IC reported and the first to explicitly target resource-limited nodes via Cryptanalysis-ASIC co-optimization strategy.

2 Architecture

Three tightly integrated innovations (Fig. 2) drive the compactness of this architecture. First, an ASCON-based PRNG and hash replaces the area-intensive Keccak. Theoretical study (NIST) [7] shows a potential 12x improvement in area through replacing Keccak with ASCON. This substitution not only saves area but also reduces power by 9.5x compared to Keccak-based implementations (estimated from [5]) in iso-frequency, iso-VDD. Second, a minimum computation datapath centered on a single butterfly engine for number theoretic transform (NTT) and inverse-NTT (INTT) is designed. Custom shift-and-add-based modulo reduction achieves up to 3x smaller area than the standard Montgomery reduction. With shift-and-add modulo division (MD), 9x area reduction is realized with respect to multiplier-based design [8]. Third, we optimize memory with concurrent data generation and computation with tiled data processing by two independent datapaths and in-place memory reuse with access scheduling for NTT/INTT, reducing total memory down to only 6.7 KB. Together, these innovations enable the first PQC KEM IC that achieves both low area and low energy, paving the way for mass adoption in ultra-low-power IoT nodes and wearables.

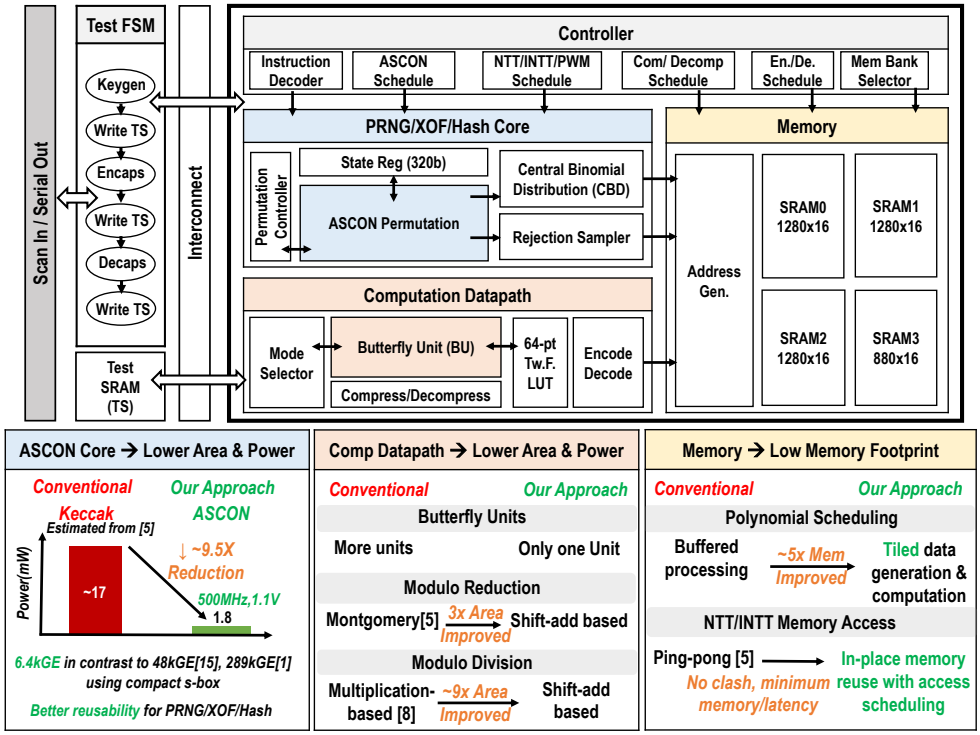


Figure 2: Full system architecture; Benefit of ASCON in reducing area and power; Improvements in computation data path in terms of butterfly unit, modulo reduction technique and modulo division technique; Memory reduction technique through inplace memory reuse with access scheduling.

The overall architecture (Fig. 2) integrates these innovations together into a compact system. The computation datapath and ASCON permutation datapath are coordinated by a shared scheduler, enabling parallel operation without idle cycles. A test FSM provides high-level control and is used for validation. 4 SRAM banks totaling 6.7 KB support computation, with two banks active during NTT/INTT, one used for secret or matrix generation, and one for output and test storage. The datapath is designed to

sustain concurrent PRNG and arithmetic computation without memory access clashing, as the SRAM follows a 1-write/1-read (1W1R) scheme for minimum area and energy, and simultaneous multi-access would otherwise introduce additional stall cycles.

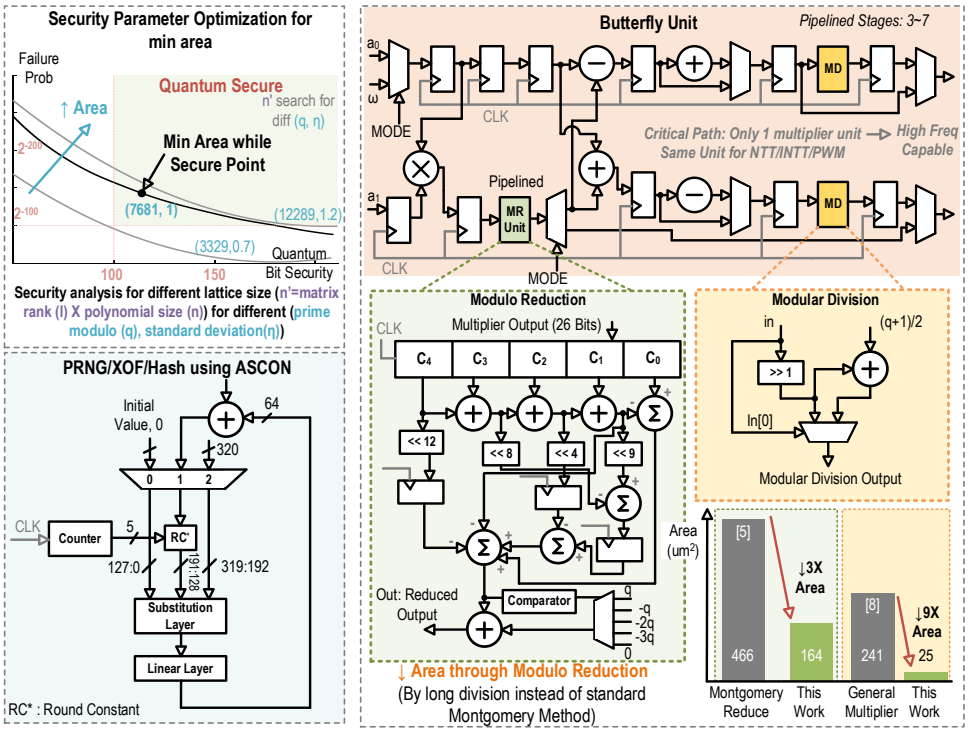


Figure 3: Security parameter optimization for minimum lattice area; ASCON implementation; Butterfly unit implementation for NTT/INTT/PWM and critical path reduction for high frequency capability with modulo reduction and modular division techniques highlighted.

To achieve compactness while ensuring security, an exhaustive Pareto parameter search (Fig. 3) of LWE (NIST-preferred hardware-friendly quantum hard problem) across the algorithm-ASIC design space is conducted. Although a similar approach has been reported on FPGA [9], a thorough IC-level exploration is required to capture the constraints of resource-limited devices. The chosen Pareto frontier point of matrix rank (l) = 9, polynomial size (n) = 64, underlying lattice dimension (n') = 576, modulus (q) = 7681, and standard deviation (η)=1 provides 100bit core SVP security (NIST level 1) with a failure probability of 2-128. Security is rigorously evaluated against both the classical and quantum adversaries using SOTA cryptanalysis techniques [10], including the Block Korkine-Zolotarev (BKZ) lattice reduction algorithm [11] and the Core Shortest Vector Problem (Core-SVP) estimator [10]. This exploration revealed that standard parameter sets used by NIST candidates could be optimized further for resource-limited hardware. By systematically navigating this design space, we identified a Pareto frontier point where area, memory, and energy are minimized while security is preserved. Any deviation from this operating point either increases the area or reduces the effective bit security below the required NIST level 1.

Fig. 3 also demonstrates key architectural components. ASCON permutation block is reused across PRNG/XOF/Hash operations, yielding a unified lightweight primitive aligned with the standardized lightweight security requirements. The second innovation is a minimum computation datapath designed around a single butterfly unit (BU) reused

for NTT, INTT and point-wise multiplication (PWM). Conventional PQC accelerators often instantiate multiple BUs in highly parallel architectures, which significantly inflate area and power. This design pipelines NTT into five stages and INTT into seven stages, the latter requiring additional modulo division due to the Gentleman–Sande algorithm. To keep the datapath compact, we introduce a shift-and-add modulo reduction (MR) unit which reduces area by 3x compared to the standard Montgomery reduction. A long-division-inspired modular division only needs one shift, one mux and one adder. Unlike prior massively parallel designs [5], the datapath achieves the required throughput in a fraction of the area, delivering efficiency tailored to resource-limited devices.

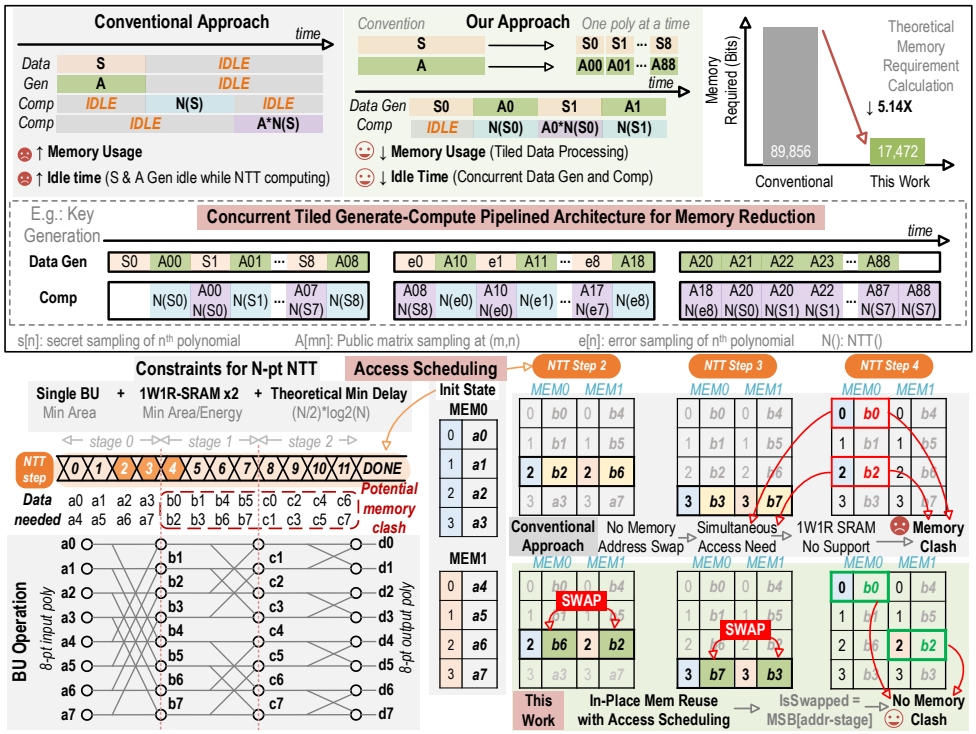


Figure 4: Concurrent tiled generate-compute pipelined architecture in secret, public matrix and error generation for minimum memory usage; N-pt NTT implementation with in-place memory reuse with access scheduling to fit resource-limited design.

Fig. 4 illustrates the memory footprint reduction techniques including the concurrent tiled generate-compute pipelined architecture and in-place memory reuse with access scheduling. Conventional PQC implementations treat secret and matrix generation as stages separate from NTT/INTT computation, which leads to idle datapaths and significantly larger memory requirements. In this design, secret and public matrix generation are decomposed into tiles, enabling memory reuse and minimizing buffer size. A concurrent generate-compute pipelined architecture further reduces idle time. This optimization is a key factor contributing to a reduction of 5.14x in memory footprint.

While concurrency reduces idle time, it is important to implement a minimum-area N(=64)-point NTT. Thus, three key constraints are imposed. First, only two 1W1R SRAMs are used to minimize area and energy. Second, only a single BU is employed. Third, the operation must be completed within the theoretical minimum latency ($(N/2) * \log N + \Delta$ cycles to fill the pipeline). Conventional techniques either require additional memory/buffer or additional BU. To demonstrate, in a simple case of three-stage 8-pt NTT (real implementation is 64-pt) with input data stored in MEM0 (a0-a3) and MEM1

(a4-a7), stage 0 output data is computed and stored back in-place (b0-b7). However, this leads to memory access clashing as stage 1 computation requires data stored in the same 1W1R SRAM (Fig. 4), necessitating additional memory units or stalling cycles. This work adopts a unique memory swapping technique to detect the clash preemptively and swaps the required memory when writing back at each stage (Fig. 4 bottom). Thus, the computation data for stages 1 and 2 are stored separately in MEM0 and MEM1 (rather than sharing a single memory), thereby preventing memory access clashes. The same technique is also applied to the INTT operation.

3 Results

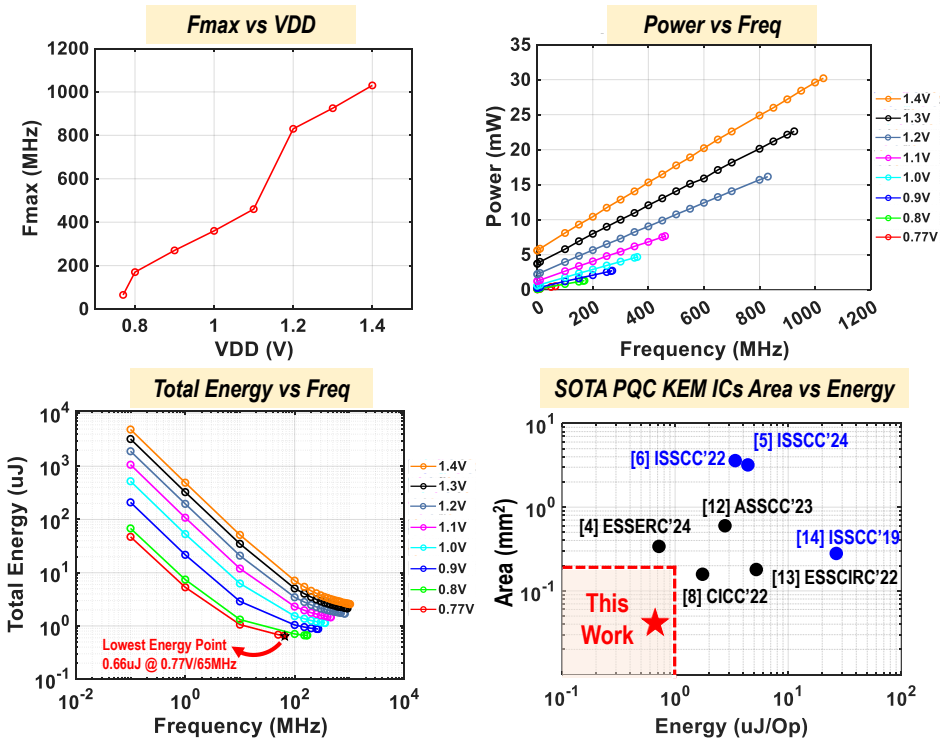


Figure 5: Measurement results. This work extends EAP for state-of-the-art PQC KEM IC by 8.5x and achieves 509x better EAP than state-of-the art generic PQC processor.

Fig. 5 shows the performance evaluation of the implemented IC. It is fabricated in Intel16 FinFET and characterized across 0.77-1.4 V. The core consumes a power of 50 uW at 0.77 V while operating at 100 kHz and 30.2 mW operating at 1.4 V, 1.03 GHz. A highly pipelined and optimized critical path helps to achieve this maximum frequency, which is the highest among the existing SOTA PQC KEM IC. This IC occupies only 0.042 mm^2 active area including the hash functions and memory, as well as computation circuitry making it the smallest PQC IC reported till date. We introduce EAP as figure-of-merit for compactness of PQC IC. We achieved 8.5x better EAP with respect to the smallest published Kyber core [4]. Fig. 6 presents a comparative analysis with SOTA implementations, alongside a Shmoo plot illustrating the relationship between supply voltage and operating frequency, and a detailed breakdown of area distribution across architectural components. The die micrograph with specification and the test setup is shown in Fig. 7. With these techniques, this work establishes the smallest and most

	This work	ESSERC'24 [4]	ISSCC'24 [5]	ASSCC'23 [12]	ESSCIRC'22 [13]	ISSCC'22 [6]	CICC'22 [8]	ISSCC'19 [14]
Technology	Intel16	40nm	28nm	40nm	28nm	28nm	65nm	40nm
Voltage (V)	0.77-1.4	0.62-1.2	0.7-1.1	0.8-1.2	0.65-1.35	0.9	0.7-1.1	0.68-1.1
Freq. (MHz)	0.1-1030	10-270	275-750	80-115	35-190	500	40-160	12-72
Active Area (mm^2)	0.042	0.34	3.2	0.6	0.18	3.6	0.158[f]	0.28[g]
Logic Gates (kGE) / Instance (k)	155/15	253[d]	2,100	532	123[e]	1,900	NA	106[g]
Memory (KB)	6.7	7	228.5	14	31	448	10.19	40.25
Power (mW)	0.05-30.2	0.27-35.0	91-420	1.14-27.57	1.21	39-368	0.33	0.52
Cycles	87,598	21,143	7,197	12,245	83,148	10,433	57,014	347,826
Energy (uJ/Op)	0.66[a]	0.72	4.4	2.76	5.22	3.4	1.75	26.6
Throughput (Ops/s)	11,758[b] / 9,475[c]	12,821	69,473	9,434	1,924	47,925	2,841	207
EAP (uJ/ops*mm^2)	0.028	0.24	14.08	1.7	0.94	12.24	0.28	7.45

a: Operating condition 0.77V/65MHz b: Operating condition 1.4V/1.04GHz c: Operating condition 1.2V/830MHz d: All circuit included SHA-3, on-chip SRAM, ROM, comm interface, and test circuit e: Cache and poly bank memory are not included f: Only computing circuit, excluding PRNG (Keccak) and SRAM g: Only computing circuit, w/o PRNG (Keccak) and SRAM

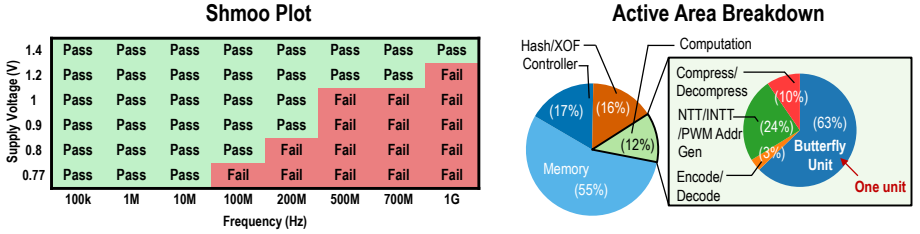
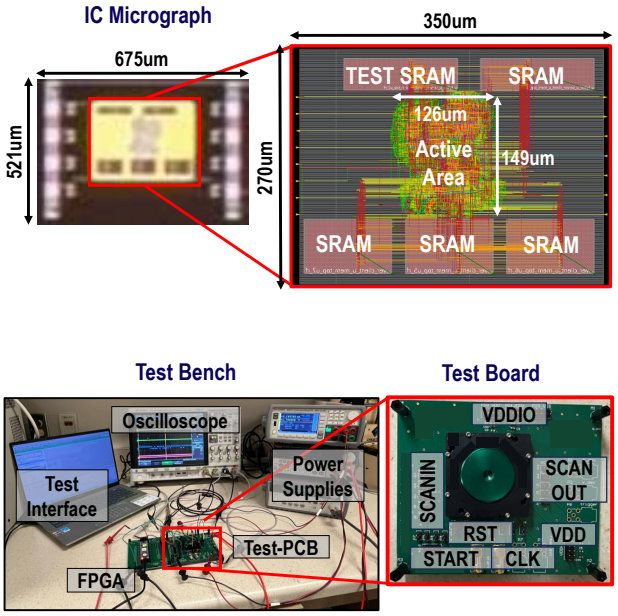


Figure 6: Comparison with state-of-the-art. Measured Shmoo plot at different voltage and frequency. Area division of different architectural components.

energy-efficient PQC IC reported to date, paving the way for adoption in resource-limited ultra-low-power IoT nodes.



IC Specification	
Technology	Intel16 FinFET
Package	FCBGA
Scheme	Lightweight ML-KEM
Security Level	AES-128 (NIST Level 1)
Total[a] Area (um^2)	94,500
Active area with test circuit (um^2)	46,733
Active[b] Area (um^2)	42,059
Logic Gates (kGE)[c]	155 kGE / 69 kGE[d]
Instance Count	15k
On-Chip SRAM	6.7KB
Voltage	0.77-1.4 V
Frequency	0.1MHz @ 0.77V, 1.03GHz @ 1.4V
Power	0.05mW @0.77V 30.2mW@1.4V

a: Include filler cells b: Include SRAMs + core KEM logic c: NAND2 equivalent d: SRAM excluded

Figure 7: IC micrograph; IC Specification and test setup.

Acknowledgments

The work of Suparna Kundu has been partially supported by the Horizon 2020 ERC Advanced Grant (101020005 Belfort), the Cybersecurity Research Program Flanders (CRPF) with reference number VOEWICS02, and Intel Corporation. The work of Angshuman Karmakar is supported by the Research-I foundation from Infosys, the Initiation grant from IIT Kanpur, Prime Minister Early Career Research Grant from Anusandhan National Research Foundation, and the Google India research fellowship.

References

[1] NIST Kyber Draft reference NIST: FIPS 203 Module-Lattice-Based Key-Encapsulation Mechanism Standard. Online. Accessed 2nd September, 2025 (2024), https://doi.org/10.6028/NIST.FIPS.203. http://doi.org/10.6028/NIST.FIPS.203

[2] NIST Dilithium Draft reference NIST: FIPS 204 Module-Lattice-Based Digital Signature Standard. Online. Accessed 2nd September, 2025 (2024), https://doi.org/10.6028/NIST.FIPS.204. http://doi.org/10.6028/NIST.FIPS.204

[3] NIST ASCON reference NIST. NIST Special Publication 800. NIST SP 800-232. Ascon-Based Lightweight Cryptography Standards for Constrained Devices Authenticated Encryption, Hash, and Extendable Output Functions. Online. Accessed 2nd September, 2025 (2025), https://doi.org/10.6028/NIST.SP.800-232. http://doi.org/10.6028/NIST.SP.800-232

[4] A. Li et al., "A 273μW 0.34mm^2 Efficient CRYSTALS-KYBER Processor for PQC Towards Edge Computing," 2024 IEEE European Solid-State Electronics Research Conference (ESSERC), Bruges, Belgium, 2024, pp. 472-475, doi: 10.1109/ESSERC62670.2024.10719541. keywords: Random access memory;Standardization;Energy efficiency;Polynomials;System-on-chip;Cryptography;Reliability;Read only memory;Edge computing;Arithmetic;Post-quantum cryptography;CRYSTALS-KYBER;low power;energy-efficient,. http://doi.org/10.1109/ESSERC62670.2024.10719541

[5] Y. Zhu et al., "16.2 A 28nm 69.4kOPS 4.4μJ/Op Versatile Post-Quantum Crypto-Processor Across Multiple Mathematical Problems," 2024 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2024, pp. 298-300, doi: 10.1109/ISSCC49657.2024.10454332. keywords:Standardization;Parallel processing;NIST;Throughput;Energy efficiency;Solid state circuits;Cryptography,. http://doi.org/10.1109/ISSCC49657.2024.10454332

[6] Y. Zhu et al., "A 28nm 48KOPS 3.4μJ/Op Agile Crypto-Processor for Post-Quantum Cryptography on Multi-Mathematical Problems," 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2022, pp. 514-516, doi: 10.1109/ISSCC42614.2022.9731783. keywords:Encapsulation;Program processors;Memory management;NIST;Throughput;Energy efficiency;Cryptography,.http://doi.org/10.1109/ISSCC42614.2022.9731783

[7] Tolga Yalcin and Samaneh Ghandali, “Need for Low-latency Ciphers: A Comparative Study of NIST LWC Finalists” https://csrc.nist.gov/csrc/media/Presentations/2022/need-for-low-latency-ciphers-a-comparative-study-o/images-media/session-1-yalcin-need-for-low-latency-ciphers.pdf. https://csrc.nist.gov/csrc/media/Presentations/2022/need-for-low-latency-ciphers-a-comparative-study-o/images-media/session-1-yalcin-need-for-low-latency-ciphers.pdf

[8] A. Ghosh et al., "A 334uW 0.158mm^2 Saber Learning with Rounding based Post-Quantum Crypto Accelerator," 2022 IEEE Custom Integrated Circuits Conference (CICC), Newport Beach, CA, USA, 2022, pp. 1-2, doi: 10.1109/CICC53496.2022.9772859. keywords: Computers;Application specific integrated circuits;Quantum computing;Conferences;NIST;Public key cryptography;Silicon,. http://doi.org/10.1109/CICC53496.2022.9772859

[9] Suparna Kundu, Archisman Ghosh, Angshuman Karmakar, Shreyas Sen, and Ingrid Verbauwhede. 2025. Rudraksh: A compact and lightweight post-quantum key-encapsulation mechanism. IACR Transactions on Cryptographic Hardware and Embedded Systems 2025, 2 (Mar. 2025), 647–680. doi:10.46586/tches.v2025.i2.647-680. http://doi.org/10.46586/tches.v2025.i2.647-680

[10] Core SVP Dana Dachman-Soled, Léo Ducas, Huijing Gong, and Mélissa Rossi. LWE with side information: Attacks and concrete security estimation. In Daniele Micciancio and Thomas Ristenpart, editors, Advances in Cryptology - CRYPTO 2020 - 40th Annual International Cryptology Conference, CRYPTO 2020, Santa Barbara, CA, USA, August 17-21, 2020, Proceedings, Part II, volume 12171 of Lecture Notes in Computer Science, pages 329–358. Springer, 2020. http://doi.org/10.1007/978-3-031-38554-4_24

[11] Dennis Hofheinz, Kathrin Hövelmanns, Eike Kiltz: Dennis Hofheinz, Kathrin Hövelmanns, and Eike Kiltz. A Modular Analysis of the Fujisaki-Okamoto Transformation. In Yael Kalai and Leonid Reyzin, editors, Theory of Cryptography - 15th International Conference, TCC 2017, Baltimore, MD, USA, November 12-15, 2017, Proceedings, Part I, volume 10677 of Lecture Notes in Computer Science, pages 341–371. Springer, 2017. http://doi.org/10.1007/978-3-319-70500-2_12

[12] A. Li, J. Lu, D. Liu, X. Li et al., "A 40nm 2.76μJ/Op Energy-Efficient Secure Post-Quantum Crypto-Processor for Crystals-Kyber on Module LWE," A-SSCC, pp. 1-3, Nov. 2023. http://doi.org/10.1109/A-SSCC58667.2023.10347915 [13] B. Kim, J. Park, S. Moon, K. Kang et al., "Configurable energy-efficient lattice-based post-quantum cryptography processor for iot devices," ESSCIRC, pp. 525-528, Sep. 2022. http://doi.org/10.1109/ESSCIRC55480.2022.9911531

[13] U. Banerjee, A. Pathak and A. P. Chandrakasan, "2.3 An Energy-Efficient Configurable Lattice Cryptography Processor for the Quantum-Secure Internet of Things," 2019 IEEE International Solid-State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2019, pp. 46-48, doi: 10.1109/ISSCC.2019.8662528. keywords: Cryptography;Program processors;Random access memory;Protocols;Lattices;Hardware;Computer architecture,. http://doi.org/10.1109/ISSCC.2019.8662528

[14] Guido Bertoni, Joan Daemen, Michaël Peeters, and Gilles Van Assche. Keccak. In Thomas Johansson and Phong Q. Nguyen, editors, Advances in Cryptology - EUROCRYPT 2013, 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, May 26-30, 2013. Proceedings, volume 7881 of Lecture Notes in Computer Science, pages 313–314. Springer, 2013. http://doi.org/10.1007/978-3-642-38348-9_19