

Lightweight Hardware Accelerator for Post-Quantum Digital Signature CRYSTALS-Dilithium

Naina Gupta[✉], Arpan Jati[✉], Anupam Chattopadhyay[✉], *Senior Member, IEEE*, and Gautam Jha

Abstract—The looming threat of an adversary with quantum computing capability led to a worldwide research effort towards identifying and standardizing novel post-quantum cryptographic primitives. Post-standardization, all existing security protocols will need to support efficient implementation of these primitives. In this work, we contribute to these efforts by reporting the smallest implementation of CRYSTALS-Dilithium, one of the chosen post-quantum digital signature scheme for NIST standardization process. By invoking multiple optimizations to leverage parallelism, pre-computation and memory access sharing, we obtain an implementation that could be fit into one of the smallest Zynq FPGA. On Zynq Ultrascale+, our design achieves an improvement of about 36.7%/35.4%/42.3% in Area×Time (LUTs×s) trade-off for KeyGen/Sign/Verify respectively over state-of-the-art implementation. We also evaluate our design as a co-processor on three different hardware platforms and compare the results with software implementation, thus presenting a detailed evaluation of CRYSTALS-Dilithium targeted for embedded applications. Further, on ASIC using TSMC 65nm technology, our design requires 0.227mm² area and can operate at a frequency of 1.176 GHz. As a result, it only requires 53.7μs/96.9μs/57.7μs for KeyGen/Sign/Verify operation for the best-case scenario.

Index Terms—Post-quantum, cryptography, PQC, crystals-dilithium, FPGA, hardware, ASIC, hardware accelerator.

I. INTRODUCTION

THE threat of an adversary with quantum computing capability is getting increasingly realistic [1], [2], with rapid growth in the capacity of quantum computers, as well as, optimized implementations of the current cryptographic primitives using quantum circuit simulators [3], [4]. It is predicted that current public-key primitives could be broken within hours [5], thus necessitating the search for alternative cryptographic primitives in the era of quantum computing. This is systematically undertaken by NIST through its

Post-Quantum Cryptography (PQC) contest [6], which plans to roll out the winners of this contest as a standard. Naturally, there is a pressing need to study efficient hardware implementations of the PQC candidates, which not only plays an important role in the contest judgement process but also helps in the rapid adoption of the standard.

The third round finalists of the NIST PQC contest consisted of 3 candidates in digital signature scheme, one of which has recently been attacked [7]. There have been several previous works as well compromising the security of this scheme [8], [9], [10]. Further, NIST recently selected three algorithms CRYSTALS-Dilithium, Falcon and SPHINCS+ to be standardized as post quantum digital signature scheme [11]. Naturally, it is of high importance to study efficient implementations of these schemes. In this work, we targeted CRYSTALS-Dilithium [12]. Digital signature being an integral part of numerous security protocols, the use-cases and the platform constraints range from high-speed servers to highly resource-constrained IoT platforms. As a result, there have been several efforts towards optimizing Dilithium for different security parameters and different platforms such as FPGAs and ASICs targeting pure-hardware based implementation [13], [14], [15], [16], [17], HLS based implementations [18], [19], [20] or as a software-hardware co-design [21]. Further, few works focused on integration in TLS protocol [22], accelerated on a GPU [23] and for developing a quantum secure blockchain [24].

From our studies on the existing Dilithium implementations, we found that there is a significant room for improvement in terms of area-efficiency, which sets the motivation for this work. The main contributions and results of this work can be summarized as:

- 1) In this work, we designed a lightweight hardware accelerator for CRYSTALS-Dilithium. We used multiple optimization strategies such as resource and control logic sharing, fusion of modules, pre-computed LUTs etc.
- 2) By achieving a reduction of about 24% in Look-up Tables (LUTs) and Flip-Flops (FFs) than state-of-the-art implementation, this work presents the smallest hardware accelerator for Dilithium. As a result, it can now be fit into one of the smallest Zynq FPGA.
- 3) In our design, we have leveraged both pipelining as well as parallelism to achieve a good balance of performance and area. Thus, on Zynq Ultrascale+, our design achieves efficiency of more than 35% for Area×Time compared to existing implementation.

Manuscript received 4 July 2022; revised 15 November 2022 and 14 March 2023; accepted 24 April 2023. Date of publication 19 May 2023; date of current version 28 July 2023. This work was supported by NRF SOCure Project under Grant NRF2018NCR-NCR002-0001. This article was recommended by Associate Editor R. Azarderakhsh. (*Corresponding author: Naina Gupta.*)

Naina Gupta and Anupam Chattopadhyay are with the School of Computer Science and Engineering, NTU, Singapore 639798 (e-mail: naina003@e.ntu.edu.sg; anupam@ntu.edu.sg).

Arpan Jati is with the School of Physical and Mathematical Sciences, NTU, Singapore 639798 (e-mail: arpan.jati@ntu.edu.sg).

Gautam Jha is with the Department of Electronics and Electrical Communication Engineering, IIT Kharagpur, Kharagpur 721302, India (e-mail: gauti.jha37@gmail.com).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TCSI.2023.3274599>.

Digital Object Identifier 10.1109/TCSI.2023.3274599

- The paper is organized as follows. Section II starts with the notations and presents a brief background about Dilithium. Section III shows the overall system architecture along with the design decisions. It is then followed by the different experimental results and performance comparison with the state-of-the-art implementations in Section IV. The performance evaluation as a hardware accelerator is presented in Section V and Section VI finally concludes the work.

A. Notations

B. Protocol Description

The diagram illustrates the high-level architecture of the proposed scheme, divided into three main components: KeyGen, Sign, and Verify.

KeyGen (Key Generation): This process starts with a seed ρ and a function $\text{TRNG} + \text{HASH}(H)$. The seed is expanded into three parts: ρ , s_1 , and s_2 . s_1 is processed by NTT to produce s_1' , which is then combined with ρ via $\hat{t} = \hat{A} \circ s_1'$ and INTT to produce t . t is combined with s_2 via $t = t + s_2$ and Power2Round to produce t_1 . Finally, t_1 and t_0 are packed and hashed to produce the secret key sk .

Sign (Signing): This process takes a message M and the secret key sk as input. M is hashed and combined with sk to produce s_1 and s_2 . s_1 is processed by NTT to produce s_1' , which is then combined with ρ via $\hat{t} = \hat{A} \circ s_1'$ and INTT to produce t . t is combined with s_2 via $t = t + s_2$ and Power2Round to produce t_1 . Finally, t_1 and t_0 are packed and hashed to produce the signature sig .

Verify (Verification): This process takes a challenge ch and the signature sig as input. ch is hashed and combined with sig to produce s_1 and s_2 . s_1 is processed by NTT to produce s_1' , which is then combined with ρ via $\hat{t} = \hat{A} \circ s_1'$ and INTT to produce t . t is combined with s_2 via $t = t + s_2$ and Power2Round to produce t_1 . Finally, t_1 and t_0 are packed and hashed to produce the verification result $valid_sig$.

The dilithium signature scheme consists of three algorithms key generation (KeyGen), signature generation (Sign) and verification (Verify) shown in Fig. 1.

- 1) **KeyGen()**: Key generation algorithm generates a keypair consisting of a public verification key (pk) and a private signing key (sk). It utilizes two random seeds public seed (ρ) and noise seed (ρ') and expands them using a variant of SHAKE-128 to generate the matrix \mathbf{A} and two vectors \mathbf{s}_1 and \mathbf{s}_2 . It then computes $\mathbf{t} = \mathbf{A}\mathbf{s}_1 + \mathbf{s}_2$ to generate the final keypair.
- 2) **Sign(sk, M)**: The purpose of this algorithm is to take the message M and the signing key sk and generate the signature sig . For this, first a masking vector \mathbf{y} is generated using SHAKE-256 and then it is used to compute $\mathbf{w} = \mathbf{A}\mathbf{y}$. The high-order bits of \mathbf{w} (denoted as \mathbf{w}_1) are then hashed with the message M to generate the challenge c . This challenge is used to generate the signature as $\mathbf{z} = \mathbf{c}\mathbf{s}_1 + \mathbf{y}$. Apart from generating \mathbf{z} , the algorithm also generates some hints h for the verifier. If the signature passes all the correctness and security checks, then sig consisting of $(\mathbf{c}, \mathbf{z}, \mathbf{h})$ is sent as the final signature to the verifier. Otherwise, the signature is re-computed again as shown in the rejection loop.
- 3) **Verify(pk, sig, M)**: The verifier computes $\mathbf{A}\mathbf{z} - \mathbf{c}\mathbf{t}_1$ and set the high-order bits in \mathbf{w}_1 . It is then hashed with the message M to generate the challenge \mathbf{c} . This challenge is compared with the one received in the signature. If the challenges match and also the norm of \mathbf{z} is valid, then the signature is accepted and the algorithm returns *valid sig* as true, otherwise signature is rejected.

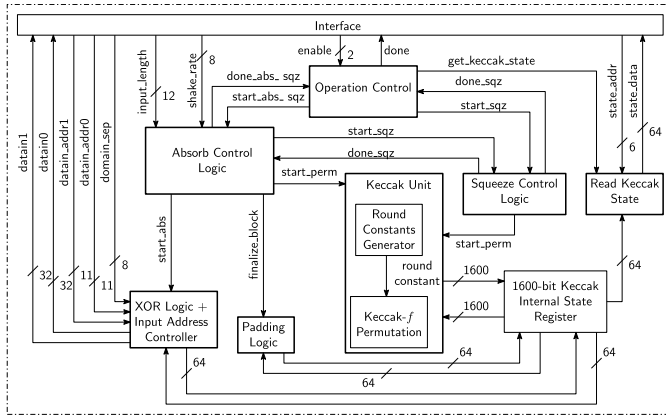


Fig. 4. Unified SHAKE wrapper.

D. SHAKE

Here, we describe the implemented architecture for the SHAKE module. Dilithium uses SHAKE-128 and SHAKE-256 extendable-output functions (XOFs) of the SHA-3 [26] family which is based around the Keccak permutation [27] for different functionalities with minor variations. For instance, SHAKE-256 and its variations are used to generate random seeds from an initial seed, for the hashing (H) and also to generate the challenge (c). Whereas, to generate Matrix A , vectors s_1 , s_2 and masking vector y_1 , variants of SHAKE-128 are used.

As the XOFs are one of the most expensive and time consuming operations, it is important to design them carefully. Consequently, there can be multiple possible design choices. One can have multiple instances of the Keccak as in [16], cascade two rounds of Keccak [14] to increase performance or have a single shared instance to achieve all the functionalities to save resources at the expense of some performance. In our design, we chose the latter approach and created and implemented a unified wrapper around Keccak from the ground-up to support all the required variations. For the Keccak- f permutation shown in Fig. 4, we used the implementation provided by the designers [28]. As Dilithium is a digital signature scheme, the architecture should support variable input message length. Also, the requested output length from the XOFs is different depending on the operation being performed. Such variations with different input/output length, different padding logic, nonce support, etc. makes it quite challenging to have one module which is efficient in hardware as well as serve all the required functionalities. As a result, the overall logic for this module is quite complex. In order to provide support for variable input and output message length, the design has three configurable modes of operation - perform absorb followed by one squeeze (mode = 1), perform only squeeze (mode = 2) and read Keccak state (mode = 3). The modes are extremely useful when the number of executions for squeeze is unknown (for instance, in case of rejection sampling). Also, mode = 3 allows the Keccak state to be accessed from outer modules, as a result we do not need extra resources to store additional copies of the 1600-bit register elsewhere compared to the work in [14].

The internal state remains valid after each squeeze operation unless reset from outside. Fig. 4 shows the architecture for the implemented SHAKE wrapper module. The interface is used to set the corresponding mode using enable signal, provide necessary information such as number of bytes to absorb (input_length), shake rate (shake_rate), etc. and to communicate with the different outer modules.

The control logic for all the modes are for starting the corresponding operation. For example, for a SHAKE-128 operation, the outer module first sets the enable signal to mode = 1 along with the input_length and shake_rate. The operation control decodes the mode and transfers the control to absorb control logic. This starts the absorb phase by reading in the domain separator and input. The XORing of the input with the Keccak state and the generation of corresponding input addresses are handled by XOR Logic + Input Address Controller. In our design, we are absorbing data in chunks of 64-bits per clock cycle. The absorb control logic is responsible for monitoring the absorbed input length and starts the permutation if required (when the input message length is greater than the shake rate). Further, when the input is completely absorbed, it enables the padding logic to finalize the absorbed block. The absorb control logic then starts a squeeze operation by setting start_sqz signal. The squeeze block starts the Keccak permutation and waits for it to finish execution using a dedicated counter (24 clock cycles as Keccak has 24 rounds). It then sends the done_sqz signal to the absorb control logic which then triggers the done_abs_sqz signal to operation control. Once the outer module receives the done signal, it starts reading the Keccak state by setting mode = 3, and providing the corresponding state address. In a single clock cycle, 64-bit data can be fetched from the wrapper. The outer module processes this data first and if more data is required, it simply increments the address, otherwise it sends a read done signal to reset the internal Keccak state. If the required output length is more than the shake rate, then after reading enough data (= shake rate), the outer module sets mode = 2 to start another squeeze operation and then read the data using mode = 3. Such an approach allows us to call squeeze back and forth only when required and when we have exhausted all the available Keccak output. Thus, saving us clock cycles as well as area as we do not need to pre-compute and store extra output bytes for example, in case of rejection sampling.

E. Sampling Modules

Dilithium requires four different types of sampling logic to generate matrix A , the vectors s_1 , s_2 , challenge c and the masking vector y_1 . Even though all of them have different sampling logic, but the coefficients are sampled based on the output of a variant of XOF function. Hence, the sampling modules in our design only contain the sampling and control logic to start and fetch coefficients from the Keccak wrapper as discussed in III-D.

Listing 1 shows a small code snippet from the software reference implementation for rejection sampling. It is used to generate vectors s_1 and s_2 from the output of SHAKE-128.


```

t0 = buf[pos] & 0x0F;
if (t0 < 15) {
    t0 = t0 - (205 * t0 >> 10) * 5;
    a[ctr++] = 2 - t0;
}

```

Listing 1. Rejection eta.

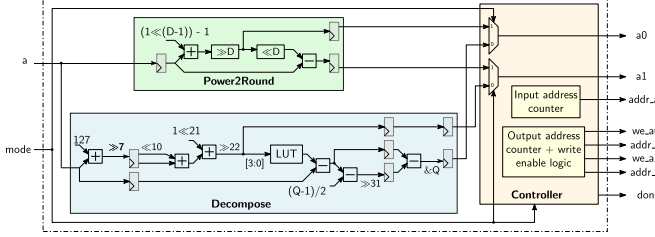


Fig. 5. Power2Round and decompose module.

Initial value of t_0 (and t_1) comes from a 4-bit uniform random and can as well assume the value 15, in which case it's rejected. Thus, the possible output values for polynomial coefficient can be pre-computed and stored in a LUT. This resulted in saving a lot of resources. The same logic is used for another value t_1 . Similar optimizations are used for other sampling modules wherever possible along with pipelining for good performance.

F. Combined Power2Round and Decompose

Fig. 5 shows the combined module for Power2Round and Decompose modules. As can be seen from the figure, the controller unit is shared between both the modules, thus saving about 180 LUTs and 120 FFs. The controller unit is used to enable the operation depending on the mode and generate address and write enable signals after receiving done from the respective module. For simplicity we have not shown the individual enable and done signals in the diagram. The software reference implementation of decompose module for Dilithium-V requires two multiplication operations to realize high-order and low-order bits of the polynomial coefficient. One is $(a1 * 1025)$ and other one is $(a1 * 2 * \text{GAMMA2})$ where $\text{GAMMA2} = 261888$, resulting in an expensive implementation for Decompose. In order to prevent DSP usage for the two cases, we used two different strategies. We realized the first multiplication using an addition operation as $(a1 * 1025)$ is equivalent to $(a1 * 1024 + a1)$ and $(a1 * 1024)$ is very efficient in hardware and can be realized by shift logic. For the second multiplication, we used a small 4-bit lookup-table (LUT) as the input $a1$ can only have 16 possible values. We pre-computed the output of $(a1 * 2 * \text{GAMMA2})$ for all 16 possible values and used the LUT whenever required. The rest of the operations are mostly either shift operations or addition/subtraction with a constant value. Because of the optimizations employed, the resource utilization of decompose is thus significantly reduced to only about 120 LUTs and 203 FFs. Power2Round is a very simple module requiring only one addition and subtraction operation. We have performed such control logic unification with multiple modules and benefited in terms of area.

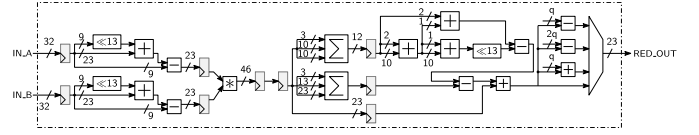


Fig. 6. Efficient modular reduction module.

G. Modular Reductions

The work in [29] and [30] proposed an efficient method for modular reduction in hardware for NewHope-NIST and CRYSTALS-Kyber. In this work, we followed a similar approach for efficient modular reduction in Dilithium. In order to compute $a \bmod q$, we used Dilithium modulus property $2^{23} = 2^{13} - 1 \pmod{8380417}$ recursively. The corresponding equations are as shown below:

$$\begin{aligned}
 a &= 2^{23}a[45 : 23] + a[22 : 0] \\
 &= 2^{13}a[45 : 23] - a[45 : 23] + a[22 : 0] \\
 &= 2^{23}a[45 : 33] + 2^{13}a[32 : 23] - a[45 : 23] + a[22 : 0] \\
 &= 2^{13}a[45 : 33] - a[45 : 33] + 2^{13}a[32 : 23] \\
 &\quad - a[45 : 23] + a[22 : 0] \\
 &= 2^{23}a[45 : 43] + 2^{13}a[42 : 33] + 2^{13}a[32 : 23] \\
 &\quad - (a[45 : 33] + a[45 : 23]) + a[22 : 0] \\
 &= 2^{13}(a[45 : 43] + a[42 : 33] + a[32 : 23]) \\
 &\quad - (a[45 : 43] + a[45 : 33] + a[45 : 23]) + a[22 : 0] \\
 &= 2^{13}c - e + a[22 : 0] \pmod{q}
 \end{aligned}$$

where $c = (a[45 : 43] + a[42 : 33] + a[32 : 23])$ and $e = a[45 : 43] + a[45 : 33] + a[45 : 23]$. Using the same modulus property again, c can be further reduced as:

$$\begin{aligned}
 2^{13}c &= 2^{23}c[11 : 10] + 2^{13}c[9 : 0] \\
 &= 2^{13}(c[11 : 10] + c[9 : 0]) - c[11 : 10] \\
 &= 2^{13}f - c[11 : 10] \pmod{q}
 \end{aligned}$$

Further reduction is possible for f as shown below:

$$\begin{aligned}
 2^{13}c &= 2^{23}f[10 : 10] + 2^{13}f[9 : 0] - c[11 : 10] \\
 &= 2^{13}(f[10] + f[9 : 0]) - (f[10] + c[11 : 10])
 \end{aligned}$$

Fig. 6 shows the implemented reduction module. The extra multiplication operation required before modular reduction operation is also integrated in this module. In order to reduce the input to 23-bit before multiplication, we first performed an initial reduction using the same recursive property. This helped in the reduction of 2 DSP resources per such module. The back-to-back flip flops correspond to the pipeline delays required for optimal DSP implementation. Two such modules are instantiated in hardware to speed up operations.

H. NTT/INTT

The NTT/INTT is one of the most computationally expensive operations in Dilithium. Fig. 7 shows the basic architecture utilized in this implementation. As the NTT operations in Dilithium allows implementations with two simultaneous butterfly operations we have utilized two

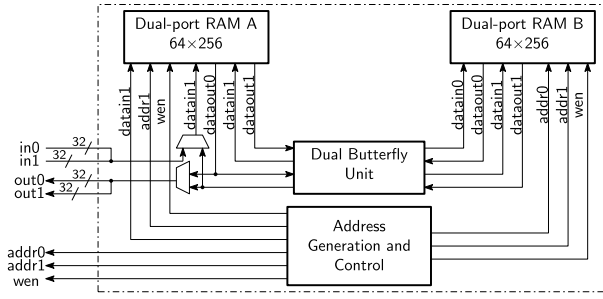


Fig. 7. NTT/INTT architecture.

64×256 dual-port memories attached with a *dual butterfly* unit. This allows for two butterfly-operations to be performed per clock cycle. The *dual butterfly* unit internally utilizes two multiplication-reduction units as described in section III-G.

The design choice of computing two coefficients per clock cycle is motivated by two reasons. Firstly, a 2×2 NTT based design like [16] can yield twice the NTT performance but at significantly increased area both in terms of LUT as well as DSPs. Secondly, we are able to share the DSP based multiplications to perform polynomial multiplications as well. Duplicating arithmetic circuits would also be required to fully utilize the additional instantiated DSPs. This will result in significant increase in area.

The NTT operation is performed in three stages. First, the data from external memory is copied over to RAM-A. Second, the data in RAM-A is processed (128 butterfly operations) and written over to RAM-B and reverse, until all the layers have been processed. Finally the result is written back to the external RAM. The entire operation requires multiple addresses, write enables and zetas to be generated and provided to various modules. This is performed by the address generation and control state machine. At the end of INTT, the required scaling by $1/n$ is performed by the butterfly unit as well. So, no additional resources or clock cycles are utilized.

IV. RESULTS AND PERFORMANCE COMPARISON

In this section, we present resource utilization and performance results for both FPGA and ASIC implementations. We also compare our results with the state-of-the-art implementations for Dilithium-V. The implemented design is realized in Verilog and the results are presented after synthesis and place and route using Vivado 2020.1 targeting two platforms Xilinx Artix-7 (XC7A200T-2) and Zynq UltraScale+ (XCZU9EG-2). The correctness of the implementation is verified using the KAT provided in the NIST package. From here on, Artix-7 is referred as A7 and Zynq UltraScale+ is referred as ZUS+. For ASIC, we synthesized the design for TSMC 65nm process technology. We used Synopsys Design Compiler version R-2020.09-SP5 for synthesis and Cadence Innovus version V19.10-P002 for place-and-route of the design. We used CCS-based standard cell library for accurate results. Synopsys DesignWare library components were used wherever applicable.

TABLE I
RESOURCE UTILIZATION FOR FPGA AND ASIC. THE ASIC RESOURCES ARE REPORTED IN GATE EQUIVALENTS (GES)

| Sub-module | FPGA | | | | ASIC | Reference |
|------------|-----------------|-----------------|-----|------|---------|-----------|
| | LUT | F/F | DSP | BRAM | GEs | |
| MakeHint | 2389 | 740 | 0 | 0 | - | [16] |
| | 67 | 85 | - | - | 1423.7 | This Work |
| UseHint | 6453 | 2808 | 0 | 0 | - | [16] |
| | 186 | 279 | 0 | 0 | 4740 | This Work |
| Encode + | 1626 | 461 | 0 | 0 | - | [16] |
| Pack | 650 | 603 | 0 | 0 | 8884 | This Work |
| Decode + | 2189 | 239 | 0 | 0 | - | [16] |
| Unpack | 694 | 568 | 0 | 0 | 9458.3 | This Work |
| Decompose | 1437 | 680 | 0 | 0 | - | [16] |
| | 120 | 203 | 0 | 0 | 3028.7 | This Work |
| NTT + | 4509×2 | 3146×2 | 16 | 0 | - | [16] |
| PolyArith | 5676 | 1218 | 41 | 1 | - | [17] |
| | 2759 | 2037 | 4 | 7 | 40182 | This Work |
| SampleA + | 3548 | 1015 | 0 | 0 | - | [16] |
| SampleS | 1479 | 189 | - | - | - | [17] |
| | 360 | 355 | 0 | 0 | 4710.3 | This Work |
| SampleY | 2220 | 630 | 0 | 0 | - | [16] |
| | 469 | 48 | - | - | - | [17] |
| | 99 | 199 | 0 | 0 | 2353 | This Work |
| SampleC | 1856 | 868 | 0 | 0 | - | [16] |
| | 384 | 662 | - | - | - | [17] |
| | 289 | 244 | 0 | 0 | 3782.7 | This Work |
| Keccak | 5483×3 | 4451×3 | 0 | 0 | - | [16] |
| | 3708 | 1623 | - | - | - | [17] |
| | 4202 | 1800 | 0 | 0 | 42155.3 | This Work |

A. Resource Utilization

Table I shows resource utilization for the major components required in Dilithium. Through extensive resource sharing between the modules, we obtained a significant reduction in overall resource utilization. One can note that the resource utilization for Sampling is quite low compared to the other existing implementations. This is mainly because the different Sampling modules contain only the minimal arithmetic and control logic with the XOFs shared between them. The slice registers usage is somewhat higher in our case because of the deep pipelines needed for good performance on both FPGA and ASIC. The various modules were individually optimized for hardware using multiple strategies wherever applicable.

B. FPGA Implementation

Table II presents detailed results for our implementation as well as existing implementations for Dilithium-V for the best-case scenario. In our case, the combined architecture (for KeyGen, Sign and Verify) requires about 13.9k LUTs, 6.8k Slice registers, 35 BRAMs and 4 DSPs. Due to a deeply pipelined architecture, our design can run at a maximum frequency of 163 MHz on Artix-7 and 391 MHz on Zynq UltraScale+. As a result, we require $387\mu s$, $699\mu s$ and $416\mu s$ for KeyGen, Sign and Verify operations on Artix-7. For

Zynq Ultrascale+, the respective operations can be finished in $161\mu s$, $291\mu s$ and $173\mu s$. We also report number of operations that can be performed per second (OP/s) for all the implementations. Further, we report the results for our implementation excluding packing/unpacking modules.

Currently, to the best of our knowledge, there are four known hardware implementations which report results for Dilithium Security Level V. Hence, we compare our results with only these implementations.

1) *Comparison With [17]*: The authors in [17] target to reduce area in terms of LUTs and FFs by utilizing more DSP resources in their architecture. Further, the modules related to packing and unpacking operations were not implemented as part of the hardware architecture. Their design requires $3.8\times$, $2.44\times$ and $11.25\times$ the LUTs, FFs and DSP resources and 4 less BRAMs compared to our design. Even though, our implementation has slightly more latency for the three operations, the overall improvement in terms of efficiency (Area \times Time) is quite high. Compared to ours, their design has a lower Area \times Time trade-off efficiency by 280%, 192% and 287% for KeyGen, Sign and Verify operations respectively.

2) *Comparison With [14]*: Zhao et. al. proposed a compact and high-performance hardware design. They did several optimizations such as segmented pipeline processing to achieve a high-performance architecture. The authors also did not implement the packing and unpacking modules in their architecture. Further, as the clock cycles reported in Section 5.1, Table 4 [14] either do not account for pre-calculation cycles or are provided for the average case. As a result, in order to accurately calculate clock cycles for the best-case sign operation, we used $(\text{Sign}_{\text{avg}}^b - \text{Sign}_{\text{avg}}^a + \text{Sign}_{\text{min}}^a)$. Similarly, for verify operation, we combined $\text{Verify}_{\text{avg}}$ with the time execution for computing $\text{NTT}(t_1)$ using $(256\times k + 296 \text{ clock cycles})$. The final numbers are shown in Table II. Their design can perform KeyGen/Sign/Verify in $90\mu s/228\mu s/117\mu s$ which is $4.05\times/2.87\times/3.15\times$ faster than our design. Even though their design offers high-performance, it comes at an expense of consuming large amounts of resources. Compared to our design, their implementation requires 155% and 82.5% more LUTs and FFs. As a result, the Area \times Time trade-off is comparable in case of Sign and Verify operation whereas it is better in [14] for KeyGen.

3) *Comparison With [16]*: The work by Beckwith et. al. present results for multiple platforms as well as for all the security levels. For a fair comparison, we compare our results with the similar FPGA platform. The authors in [16] targeted a high-performance implementation. As a result, they can perform Keygen, Sign and Verify in $121\mu s$, $210\mu s$ and $126\mu s$ achieving a reduction of about $3.2\times$, $3.33\times$ and $3.30\times$ compared to our implementation. Even though the performance achieved is better, their design requires $3.81\times$, $4.14\times$ and $4\times$ the LUTs, FFs and DSPs compared to our design. This is because their architecture utilizes multiple cores to perform expensive operations such as NTT, Keccak whereas we are using only one core for almost all the modules. Thus, the high latency in our design is compensated by low area utilization as well as higher achievable frequency. As a result, their design has a higher Area \times Time trade-off metric

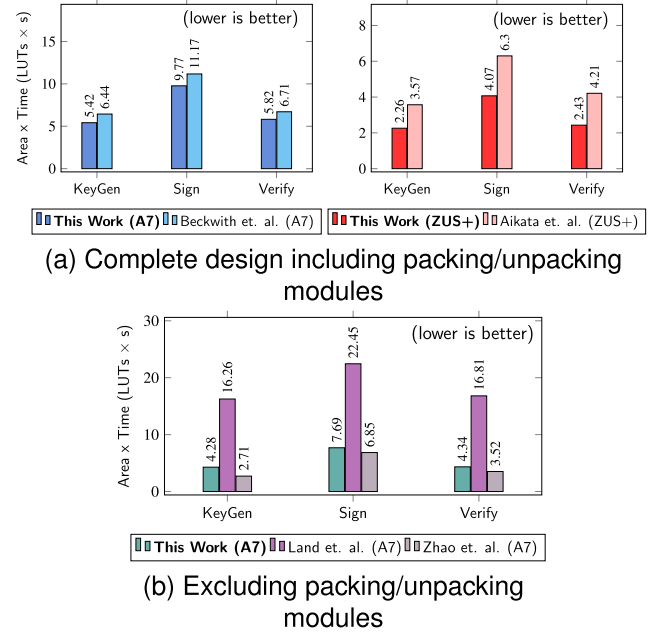


Fig. 8. **Area \times Time (LUTs \times s) trade-off**: When comparing the results from (a), our design shows better Area \times Time trade-off for both Artix-7 and Ultrascale+. When excluding packing / unpacking modules (b), the implementation by Zhao et. al. [14] has better trade-off for KeyGen, whereas our design is comparable for Sign and Verify operations. One interesting thing to note is that our result on Zynq Ultrascale+ is best across all platforms.

of 18.8%, 14.3% and 15.3% for KeyGen, Sign and Verify operations respectively.

4) *Comparison With [15]*: The design by Mert et. al. presents a unified architecture for Dilithium and Saber. The authors in [15] target a low area implementation. Hence, the Keccak core and the polynomial multiplier are shared between the two algorithms. But, there are many specialized modules required just for Dilithium, similarly for Saber. As a result, their implementation requires 31.7% and 36.2% more LUTs and FFs compared to our design but the BRAM utilization is comparatively low in their implementation. One interesting thing to note is our design can run at about twice the speed of their design. As a result, we are able to achieve an improvement of 17%, 14.9% and 24.5% in the execution time of KeyGen, Sign and Verify operations. Further, the Area \times Time trade-off efficiency is lower by 57.9%, 54.8% and 73.3% compared to ours for the respective operations.

Efficiency metrics: In order to collate all the results, we compare our work with the existing implementations using two metrics - Area \times Time (LUTs \times s) trade-off (shown in Fig. 8) and Number of operations (KeyGen/Sign/Verify) that can be performed per second per LUT (shown in Fig. 9).

Both metrics are used to quantify the efficiency of an implementation. In the former case, a lower value denotes a better design, whereas for the latter, a higher value is considered to be good.

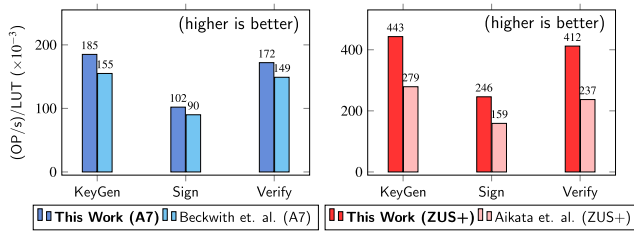
C. Post-Layout ASIC Implementation

Table II shows the results for the ASIC implementation. Our design requires an area of about 0.227 mm^2 ($\approx 157\text{k}$ GEs) after place-and-route excluding on-chip

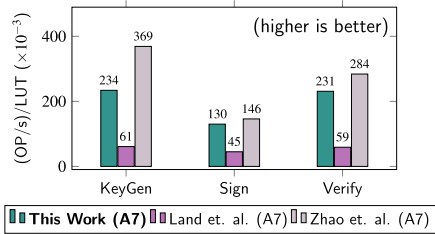
TABLE II

PERFORMANCE AND COMPARISON FOR DILITHIUM-V FOR BEST-CASE SCENARIO (SIGNATURE IS VALID AFTER THE FIRST ITERATION) ON FPGA AND ASIC. THE AREA FOR ASIC IS EXCLUDING ON-CHIP MEMORY AND IS REPORTED IN mm^2 AND GATE EQUIVALENTS (GES)

| Reference | Family | Freq. | Area | | | | KeyGen | | | Sign | | | Verify | | |
|--|--------------------|-------|--------------|--------|---------|-----|-----------------------------|--------------------|-------|-----------------------------|--------------------|-------|-----------------------------|--------------------|-------|
| | | (MHz) | LUT | FF | RAM | DSP | Cycles ($\times 10^3$) | Time (μ S) | OP/s | Cycles ($\times 10^3$) | Time (μ S) | OP/s | Cycles ($\times 10^3$) | Time (μ S) | OP/s |
| FPGA Results (Including packing/unpacking modules) | | | | | | | | | | | | | | | |
| This Work | Artix-7 | 163 | 13,975 | 6,845 | 35 | 4 | 63.2 | 387 | 2580 | 113.9 | 699 | 1430 | 67.9 | 416 | 2401 |
| Beckwith et. al. [16] | Artix-7 | 116 | 53,187 | 28,318 | 29 | 16 | 14.0 | 121 | 8263 | 24.4 | 210 | 4762 | 14.6 | 126 | 7922 |
| Beckwith et. al. [16] | Kintex-7 | 173 | 54,468 | 28,639 | 29 | 16 | 14.0 | 81 | 12324 | 24.4 | 141 | 7102 | 14.6 | 85 | 11815 |
| This Work | Zynq Ultrascale+ | 391 | 13,975 | 6,845 | 35 | 4 | 63.2 | 161 | 6189 | 113.9 | 291 | 3431 | 67.9 | 173 | 5759 |
| Aikata et. al. [15] | Zynq Ultrascale+ | 200 | 18,406 | 9,323 | 24 | 4 | 38.8 | 194 | 5149 | 68.5 | 342 | 2920 | 45.8 | 229 | 4368 |
| Beckwith et. al. [16] | Virtex Ultrascale+ | 256 | 53,907 | 28,435 | 29 | 16 | 14.0 | 55 | 18238 | 24.4 | 95 | 10509 | 14.7 | 57 | 17483 |
| FPGA Results (Excluding packing/unpacking modules) | | | | | | | | | | | | | | | |
| This Work | Artix-7 | 168 | 11,744 | 5,661 | 35 | 4 | 61.3 | 365 | 2742 | 109.9 | 655 | 1527 | 62.1 | 369 | 2707 |
| Land et. al. [17] | Artix-7 | 140 | 44,653 | 13,814 | 31 | 45 | 51.0 | 364 | 2746 | 70.4 | 503 | 1989 | 52.7 | 377 | 2656 |
| Zhao et. al. [14] | Artix-7 | 96.9 | 29,998 | 10,336 | 11 | 10 | 8.8 | 90 | 11055 | 22.1 | 228 | 4382 | 11.4 | 117 | 8513 |
| ASIC Results | | | | | | | | | | | | | | | |
| This Work | TSMC 65nm | 1176 | 0.227 mm^2 | | 157 kGE | | 63.2 | 53.7 | 18614 | 113.9 | 96.9 | 10320 | 67.9 | 57.7 | 17322 |
| Aikata et. al. [15] | UMC 65nm | 400 | 0.317 mm^2 | | 220 kGE | | 38.8 | 97.1 | 10298 | 68.5 | 171.3 | 5839 | 45.8 | 114.5 | 8735 |



(a) Complete design including packing/unpacking modules



(b) Excluding packing/unpacking modules

Fig. 9. **Operations/second/LUT:** In case of Zynq Ultrascale+, we are able to achieve an improvement of about 58.78%, 54.72% and 73.84% for KeyGen, Sign and Verify operations over the best known state-of-the-art implementation [15]. Whereas, comparing our design on Artix-7 when including pack/unpack modules, we achieved an improvement of 19.4%, 13.33% and 15.43% respectively. While excluding packing/unpacking modules, our design has competitive performance for sign and verify. The improvement in ZUS+ is quite high for all the operations because our design is better in terms of both area as well as performance than the existing implementation.

memory. Because of the highly pipelined architecture, the design can run at 1.176 GHz. As a result, the KeyGen/Sign/Verify takes only 53.7 μs /96.9 μs /57.7 μs . Compared to state-of-the-art implementation, we achieve a reduction of 1.4 \times in area and improve the runtime of KeyGen/Sign/Verify by 1.81 \times /1.77 \times /1.98 \times .

The physical layout of the implemented design after place-and-route is shown in Fig. 10. We have highlighted the regions for individual modules in Dilithium. One can see that majority of the area in the design is consumed by Keccak, Sampling, Reduction and NTT modules with Keccak being the largest. In our design, we are using 2 instantiations of

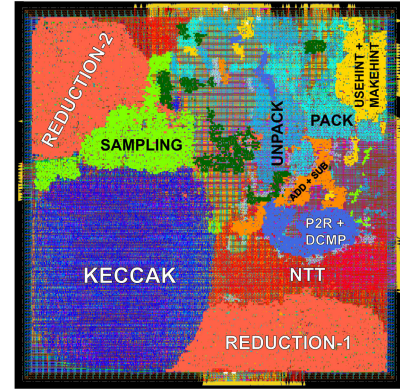


Fig. 10. Physical layout of the Dilithium-V Core after place-and-route.

reduction modules (marked as Reduction-1 and Reduction-2). The remainder of the space is utilized by the multiplexers connecting the various modules and the FSM based control logic.

V. HARDWARE EVALUATION AS AN ACCELERATOR

To demonstrate the effectiveness of the developed hardware accelerator, we evaluated our design on three test platforms: Artix-7 (XC7A75T-2) with Microblaze@100MHz processor, Zynq-7000 (XC7Z010-1) with ARM Cortex-A9@667MHz and Zynq Ultrascale+ (XCZU3EG-1) with ARM Cortex-A53@1200MHz. For software implementation, we used the reference implementation available in the NIST Round 3 package of Dilithium [31] and compiled it using Xilinx Vitis 2020.1 with default settings. It is important to quantify the real practical benefits of a hardware accelerator in a complete system as in many situations the full potential of an accelerator might not be achieved. Fig. 11 shows setup for the Zynq Ultrascale+ platform using the Ultra96 board.

Fig. 12 shows the block diagram of the accelerator connected as an AXI Peripheral applicable for both the Zynq Platforms. A very similar block design is used for the Microblaze setup as well. We created an AXI4-lite memory mapped register based peripheral and implemented support for control and data logic to connect with the designed

TABLE III

PERFORMANCE AS AN ACCELERATOR FOR **BEST-CASE SCENARIO**: SIGNATURE IS VALID IN THE FIRST ITERATION AND IN THE **WORST-CASE SCENARIO**: REJECTION LOOP IS EXECUTED 17 TIMES BEFORE A VALID SIGNATURE IS GENERATED. THE NUMBERS ARE CPU CLOCK CYCLES ($\times 10^3$) ELAPSED WHEN THE OPERATION IS EXECUTED ON HARDWARE VERSUS ON SOFTWARE

| FPGA | Processor | KeyGen | | | Sign: Best-case | | | Sign: Worst-case | | | Verify | | |
|-----------|----------------|--------|---------|---------|-----------------|---------|---------|------------------|---------|---------|--------|---------|---------|
| | | HW | SW | Speedup | HW | SW | Speedup | HW | SW | Speedup | HW | SW | Speedup |
| XC7A75T-2 | Microblaze | 48.2 | 12613.3 | 261.7 | 85.9 | 16210.8 | 188.7 | 796.7 | 84363.9 | 105.9 | 51.4 | 12857.7 | 250.1 |
| XC7Z010-1 | ARM Cortex-A9 | 337.5 | 5079.6 | 15.05 | 607.9 | 6907.6 | 11.36 | 5661.3 | 39156.1 | 6.92 | 362.2 | 5248.9 | 14.49 |
| XCZU3EG-1 | ARM Cortex-A53 | 254.1 | 2134.5 | 8.4 | 456.3 | 3747.8 | 8.21 | 4245.9 | 27240.7 | 6.42 | 271.7 | 2347.7 | 8.64 |

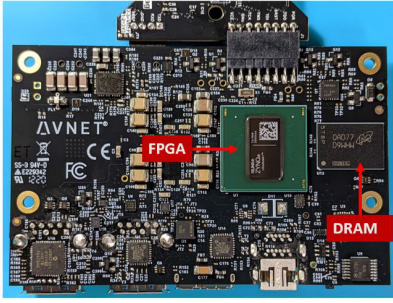


Fig. 11. Setup with Ultra96-v2 (XCZU3EG-1) as one of the hardware evaluation board.

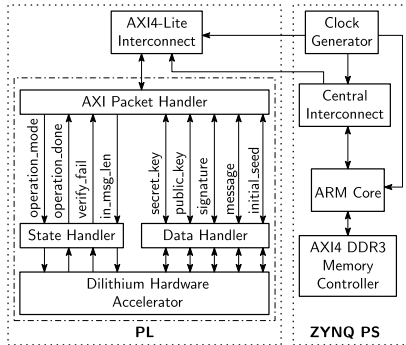


Fig. 12. Dilithium as an accelerator on Xilinx Zynq. The same AXI peripheral IP was used for the Microblaze processor as well. Due to connectivity using AXI4-lite bus, the accelerator can easily be connected with other processors as well such as RISC-V.

Dilithium core. The AXI packet handler is responsible for receiving and sending AXI commands from CPU core and decoding them. Based on the received command, it sends the corresponding operation request to the state handler. It also communicates with the data handler for address and data read/write operations. Both the state and data handler act as a bridge between the AXI packet handler and the Dilithium accelerator. Since, the accelerator is designed to be configurable from outside, one can also use it to only accelerate individual operations like KeyGen, Sign or Verify.

In Table III, we report results for the hardware accelerator. For comparison, we chose one best-case scenario where a valid signature is generated after the first iteration itself and a worst-case scenario where the valid signature is generated after the rejection loop is executed for 17 times. It is clear from the figure that the speedup obtained is inversely proportional to the performance of the associated CPU. In case of a modern high performance CPUs like ARM cortex A53 and A9 we obtain speedups of about 6.42 to 15.05 \times . Whereas, for Microblaze the speedup is significantly higher at 105-261 \times .

VI. CONCLUSION AND FUTURE WORK

By utilizing multiple optimization strategies such as resource and control logic sharing, pre-computed LUTs, fusion of modules etc. we achieved a reduction of 24% in LUTs and FFs in area compared to the best known implementation. As a result, our design requires only 13.9k LUTs, 6.8k FFs, 4 DSPs and 35 BRAMs. To the best of our knowledge, this work presents the smallest hardware accelerator for CRYSTALS-Dilithium which can now be fit into the smallest Zynq FPGA. We also present detailed comparison with existing implementations and show that our design achieves more than 35% efficiency for Area \times Time product on Zynq UltraScale+ for all the operations.

We also implemented our design on ASIC and report numbers for major components of Dilithium as well as the overall design. Compared to the state-of-the-art implementation, our design requires about 157 kGE with 0.227 mm^2 area, achieving a reduction of about 1.4 \times . In terms of performance, the implemented design can run at 1.176 GHz achieving an improvement of about 1.7 \times .

Further, this work presents the first hardware evaluation for complete Dilithium-V as an accelerator on three different platforms and demonstrate that achieved speedup is significantly high, about 105-261 \times for Microblaze and about 6.42-15.05 \times for ARM Cortex compared to the software implementations on these platforms.

In the future, we plan to integrate low-cost side-channel countermeasures in the design.

REFERENCES

- [1] J. Chow, O. Dial, and J. Gambetta. (2021). *IBM Quantum Breaks the 100-Qubit Processor Barrier*. [Online]. Available: <https://research.ibm.com/blog/127-qubit-quantum-process-or-eagle>
- [2] C. Wang et al., "Towards practical quantum computers: Transmon qubit with a lifetime approaching 0.5 milliseconds," *Npj Quantum Inf.*, vol. 8, no. 1, pp. 1–6, Jan. 2022.
- [3] Z. Wang, S. Wei, and G. Long, "A quantum circuit design of AES," 2021, *arXiv:2109.12354*.
- [4] J. Zou, Z. Wei, S. Sun, Y. Luo, Q. Liu, and W. Wu, "Some efficient quantum circuit implementations of camellia," *Quantum Inf. Process.*, vol. 21, no. 4, pp. 1–27, Apr. 2022.
- [5] C. Gidney and M. Ekerå, "How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits," *Quantum*, vol. 5, p. 433, Apr. 2021.
- [6] NIST. *Submission Requirements and Evaluation Criteria for the Post-Quantum Cryptography Standardization Process*. Accessed: Apr. 12, 2021. [Online]. Available: <http://csrc.nist.gov/groups/ST/post-quantum-crypto/documents/call-for-proposals-final-dec-2016.pdf>
- [7] W. Beullens, "Breaking rainbow takes a weekend on a laptop," in *Proc. 42nd Annu. Int. Cryptol. Conf. (CRYPTO)*. Santa Barbara, CA, USA: Springer, Aug. 2022, pp. 464–479.

- [8] W. Beullens, "Improved cryptanalysis of UOV and rainbow," in *Proc. Annu. Int. Conf. Theory Appl. Cryptograph. Techn.* Cham, Switzerland: Springer, 2021, pp. 348–373.
- [9] D. Pokorny, P. Socha, and M. Novotny, "Side-channel attack on rainbow post-quantum signature," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Feb. 2021, pp. 565–568.
- [10] C. Tao, A. Petzoldt, and J. Ding, "Improved key recovery of the HFEv-signature scheme," *Cryptol. ePrint Arch., Tech. Paper 2020/1424*, 2020. [Online]. Available: <https://eprint.iacr.org/2020/1424>
- [11] M. V. Yesina, Y. V. Ostrianska, and I. D. Gorbenko, "Status report on the third round of the NIST post-quantum cryptography standardization process," *Radiotekhnika*, no. 210, pp. 75–86, Sep. 2022.
- [12] L. Ducas et al., "CRYSTALS-Dilithium: A lattice-based digital signature scheme," *IACR Trans. Cryptograph. Hardw. Embedded Syst.*, vol. 2018, pp. 238–268, Feb. 2018.
- [13] S. Ricci et al., "Implementing CRYSTALS-Dilithium signature scheme on FPGAs," in *Proc. 16th Int. Conf. Availability, Rel. Secur.*, Aug. 2021, pp. 1–11.
- [14] C. Zhao et al., "A compact and high-performance hardware architecture for CRYSTALS-Dilithium," *IACR Trans. Cryptograph. Hardw. Embedded Syst.*, vol. 2022, pp. 270–295, Nov. 2022.
- [15] A. C. Mert, D. Jacquemin, A. Das, D. Matthews, S. Ghosh, and S. S. Roy, "A unified cryptoprocessor for lattice-based signature and key-exchange," *IEEE Trans. Comput.*, early access, Oct. 17, 2022, doi: [10.1109/TC.2022.3215064](https://doi.org/10.1109/TC.2022.3215064).
- [16] L. Beckwith, D. T. Nguyen, and K. Gaj, "High-performance hardware implementation of lattice-based digital signatures," *Cryptol. ePrint Arch., Tech. Paper 2022/217*, 2022. [Online]. Available: <https://eprint.iacr.org/2022/217>
- [17] G. Land, P. Sasdrich, and T. Guneyso, "A hard crystal-implementing Dilithium on reconfigurable hardware," *Cryptol. ePrint Arch., Tech. Paper 2021/355*, 2021. [Online]. Available: <https://eprint.iacr.org/2021/355>
- [18] K. Basu, D. Soni, M. Nabeel, and R. Karri, "NIST post-quantum cryptography—A hardware evaluation study," *Cryptol. ePrint Arch., Tech. Paper 2019/047*, 2019. [Online]. Available: <https://eprint.iacr.org/2019/047>
- [19] D. Soni, K. Basu, M. Nabeel, N. Aaraj, M. Manzano, and R. Karri, *Hardware Architectures for Post-Quantum Digital Signature Schemes*. Berlin, Germany: Springer, 2021.
- [20] D. Soni, K. Basu, M. Nabeel, and R. Karri, "A hardware evaluation study of NIST post-quantum cryptographic signature schemes," in *Proc. 2nd PQC Standardization Conf.*, 2019, pp. 1–4.
- [21] Z. Zhou, D. He, Z. Liu, M. Luo, and K.-K.-R. Choo, "A software/hardware co-design of CRYSTALS-Dilithium signature scheme," *ACM Trans. Reconfigurable Technol. Syst.*, vol. 14, no. 2, pp. 1–21, Jun. 2021.
- [22] A. F. De Abiega-L'Eglise, K. A. Delgado-Vargas, F. Q. Valencia-Rodriguez, V. G. Gonzalez-Quiroga, G. Gallegos-Garcia, and M. Nakano-Miyatake, "Performance of new hope and CRYSTALS-Dilithium postquantum schemes in the transport layer security protocol," *IEEE Access*, vol. 8, pp. 213968–213980, 2020.
- [23] J. Wright, M. Gowanlock, C. Philabaum, and B. Cambou, "A CRYSTALS-Dilithium response-based cryptography engine using GPGPU," in *Proc. Future Technol. Conf.* Cham, Switzerland: Springer, 2021, pp. 32–45.
- [24] L. Sharma and A. Mishra, "Analysis of CRYSTALS-Dilithium for Blockchain security," in *Proc. 2nd Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, May 2021, pp. 160–165.
- [25] S. Bai. (2021). *CRYSTALS-Dilithium: Digital Signatures From Module Lattices*. [Online]. Available: <https://pq-crystals.org/dilithium/data/dilithium-specification-round3-20210208.pdf>
- [26] M. J. Dworkin, "SHA-3 standard: Permutation-based hash and extendable-output functions," *Federal Inf. Process. Stds. (NIST FIPS)*, Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Aug. 2015, DOI: [10.6028/NIST.FIPS.202](https://doi.org/10.6028/NIST.FIPS.202).
- [27] G. Bertoni, J. Daemen, M. Peeters, and G. Van Assche, "KECCAK sponge function family main document," *Submission NIST*, vol. 3, no. 30, pp. 320–337, 2009.
- [28] B. Guido, D. Joan, P. Michael, and V. A. Gilles. *KECCAK Hardware Implementation*. Accessed: Apr. 12, 2021. [Online]. Available: <https://keccak.team/hardware.html>
- [29] N. Zhang, B. Yang, C. Chen, S. Yin, S. Wei, and L. Liu, "Highly efficient architecture of NewHope-NIST on FPGA using low-complexity NTT/INTT," *IACR Trans. Cryptograph. Hardw. Embedded Syst.*, vol. 2020, pp. 49–72, Mar. 2020.
- [30] F. Yaman, A. C. Mert, E. Ozturk, and E. Savas, "A hardware accelerator for polynomial multiplication operation of CRYSTALS-KYBER PQC scheme," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Feb. 2021, pp. 1020–1025.
- [31] S. Bai. (2021). *CRYSTALS-Dilithium Round3 NIST Package*. [Online]. Available: <https://csrc.nist.gov/CSRC/media/Projects/post-quantum-cryptography/documents/round-3/submissions/Dilithium-Round3.zip>



Naina Gupta received the bachelor's degree from Guru Gobind Singh Indraprastha University (GGSIU), India, in 2013, the master's degree from IIIT-Delhi, India, in 2015, and the Ph.D. degree from the School of Computer Science and Engineering (SCSE), NTU, Singapore, in 2023. Since 2022, she has been a Research Associate with Temasek Laboratories, NTU. Prior to that she was a Researcher with IIIT-Delhi from 2015 to 2017 and as an Intern with NTU from 2017 to 2018. Her research interests include cryptography, side-channel attacks, hardware security and post-quantum cryptography, and efficient implementations of cryptographic algorithms.



Arpan Jati received the bachelor's degree from the West Bengal University of Technology (WBUT), India, in 2010, and the master's degree from IIIT-Delhi, India, in 2013, where he is currently pursuing the Ph.D. degree. Since 2018, he has been a Research Associate with the Physical Analysis and Cryptographic Engineering (PACE) Laboratory, School of Physical and Mathematical Sciences (SPMS), Nanyang Technological University (NTU), Singapore. His research interests include cryptography, side-channel attacks, hardware security, efficient implementations of cryptographic algorithms, blockchain protocols, and circuit design.



Anupam Chattopadhyay (Senior Member, IEEE) received the B.E. degree from Jadavpur University, India, in 2000, the M.Sc. degree from ALaRI, Switzerland, in 2002, and the Ph.D. degree from RWTH Aachen University, Germany, in 2008.

From 2008 to 2009, he was a member of Consulting Staff with CoWare Research and Development, Noida, India. From 2010 to 2014, he led the MPSOC Architectures Research Group, RWTH Aachen University, as a Junior Professor. Since September 2014, he has been an Assistant Professor with the School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU), Singapore, where he got promoted to an Associate Professor (with tenure) in August 2019. In the past, he was a Visiting Professor with EPFL, Switzerland, and with Indian Statistical Institute, Kolkata. His research advances has been reported in more than 150 conference/journal papers (ACM/IEEE/Springer), multiple research monographs and edited books (CRC, Springer) and open-access forums. Together with his doctoral students, he proposed novel research directions like, domain-specific high-level synthesis for cryptography, high-level reliability estimation flows for embedded processors, generalization of classic linear algebra kernels, and multi-layered coarse-grained reconfigurable architecture. His research in the area of emerging technologies has been covered by major news outlets across the world, including Asian Scientist, Straits Times, and The Economist. He received the Borchers' Plaque from RWTH Aachen University, for outstanding doctoral dissertation in 2008 and the nomination for the Best IP Award at DATE 2016.



Gautam Jha received the B.Tech. and M.Tech. degrees in electronics and electrical communication engineering under microelectronics and VLSI design specialization from IIT Kharagpur, Kharagpur, India, in 2022. His research interests include digital circuits, embedded systems, and computer architecture.