Area and Power Efficient Implementation of Configurable Ring Oscillator PUF

Enas Abulibdeh\textsuperscript{1}, Hani Saleh\textsuperscript{1}, Baker Mohammad\textsuperscript{1}, Mahmoud Al-Qutayri\textsuperscript{1}, and Asif Veeran\textsuperscript{1}

\textsuperscript{1}Department of Computer and Communication Engineering, SoC Lab, Khalifa University

April 02, 2024
Area and Power Efficient Implementation of Configurable Ring Oscillator PUF

Enas Abulibdeh, Hani Saleh, Baker Mohammad, Mahmoud Al-Qutayri, Asif Veeran

SoC Lab, Department of Computer and Communication Engineering, Khalifa University, Abu Dhabi, UAE

{ 100059804, hani.saleh, baker.mohammad, mahmoud.alqutayri, asif.veeran } @ku.ac.ae

Abstract—Physically Unclonable Function (PUF) is an emerging hardware security primitive that provides a promising solution for lightweight security. PUFs can be used to generate a secret key that depends on the random manufacturing process variation of the device for lightweight authentication and device identification. This work proposes an optimized version of the Configurable Ring Oscillator (CRO) PUF that aims to reduce power consumption and area overhead. The proposed design eliminates the duplication of ROs, reduces the switching activity, and introduces the inter-stage delay as an additional source of randomness. The proposed PUF has been implemented in 22nm FDSOI technology using the Synopsys tools. A comprehensive security analysis has been acquired utilizing Challenge-Response Pairs collected from 8 chips. Results show an average of 49.42%, 38.25%, 9.95%, and 45.5% for uniformity, diffuseness, reliability, and uniqueness, respectively. Compared with the state-of-the-art, the proposed design achieves an area and power reduction of 75% and 65.1%, respectively. With the proposed PUF delivering $10^{12}$ CRPs, it is classified as a strong PUF. Additionally, the proposed design passes NIST tests and achieves an average prediction accuracy of 67.1% of machine learning modeling.

Index Terms—PUF, CROPUF, ASIC, area, power, uniformity, uniqueness, diffuseness, reliability, machine learning, modeling attack, correlation, NIST.

I. INTRODUCTION

The security and privacy of IoTs are crucial due to their diversity, distribution, and ease of access. The inherent trade-off between hardware capabilities and security in limited-resource systems makes them vulnerable to different types of attacks, including physical breaches. Physically Unclonable Function (PUF) is a hardware primitive that presents a promising solution for low-cost security [1]. PUF devices convert the random physical variations of the underlying hardware into unique and unpredictable digital responses (outputs of PUF). Lightweight authentication, protection against IC counterfeiting, and secure key generation are examples of security applications that utilize PUF keys [2]. A variety of PUF architectures have been proposed in the literature. Their source of variation and target applications rely on their characteristics. Configurable Ring Oscillator (CRO) PUF is a digital-based PUF that evaluates the delay variation between a selected pair of ROs and extracts a response according to the frequency difference [3]. Generally, CRO PUF has a relatively simple design, is easy to implement, and delivers a large space of CRPs, allowing the classification of the block as a strong PUF. However, the duplication of ROs in the CRO structure consumes a relatively large area, which is a challenge to justify especially for systems with limited resources [4]. For resource reduction, prior works focused on introducing different delay units and amplifying their variations. A PUF cell of tri-gate CMOS is used in [5] and [6] as a lightweight delay component, where a 4-bit challenge is used to configure the cell. However, the staked physical layers complicate the post-processes required to resolve biased and unstable bits inherited by the manufacturing process. Hybrid PUF [7] introduces four delay structures based on logic gates as a rich source of entropy, despite the negative impact on space utilization. Following the same approach, [8] replaces the inverters in the RO PUF with a tristate matrix, which has relatively less complexity than [7]. Still, the area overhead due to matrix replication is significant. S. Taneja et al. [9] define the accumulated jitter in ROs as a dynamic source of entropy and use a Time-to-Digital Converter (TDC) to convert pulse-width into random digital output. Temperature-aware PUF [10] uses a set of buffer-based ROs and a phase/frequency detector (PFD) followed by a charge pump (CP) employed to evaluate the difference of the selected RO pair. PFD proves its ability to extract stable responses at different environmental temperatures. A similar approach has been adopted by COMS PUF [11], which evaluates differences in the frequency and duty cycle of ROs. Nevertheless, the counter of conventional design is simpler as a digital circuit compared to the complexity of PFD and CP. M. Korona et al. [12] reduce the number of ROs and employ a Linear-Feedback-Shift-Register (LFSR) as a scrambler to generate response bits sequentially. Yet, the design achieves low quality in terms of uniqueness. H. Kareem et al. [13] propose an XOR-based delay unit, configured by a 3-bit challenge. However, the proposed approach relies on reducing the challenge size for area reduction, posing scalability concerns. This work proposes an optimized design that focuses on minimizing both area and power consumption by replacing the ROs with a single pair of ROs. To compensate for the source of variation that arises from RO duplication, the complex networks of logic are used as delay units inside the RO structure, delivering 16 delay configurations per two connected gates. The evaluation time of the response is the average delay of the delay path. The delay stage is a delay unit of the proposed RO and is configured by a 4-bit challenge (input of PUF). Hence, the RO is constructed from a set of delay stages and configured by a $4^n$-bit challenge. The variation in the output of the delay units relies on the unit delay and the signal propagates through the RO. Consequently, the delay configurations and phase variations at the Mux inputs are characterized as diverse sources of variation compared to conventional CRO PUF.
designs. The proposed PUF delivers Challenge-Response Pairs (CRPs) of $10^{32}$, meeting the requirements for lightweight authentication [14].

A. Contributions

The contributions of this work include the following:

• Compact design: Replaces the set of ROs with a single pair, achieving a 75% reduction in the area.
• Power-efficient design: Achieves power reduction along with area reduction. Additionally, the switching activity across RO is reduced by 50%.
• Fast generation: Requires minimal oscillations to emphasize stage variations, with responses generated at 1 GHz.
• A Comprehensive study is conducted on the design, which was implemented, and fabricated in silicon using 22 nm FDSOI technology.

• Surpassing state-of-the-art designs in both area and power efficiency, while maintaining superior performance. In comparison with the state-of-the-art, the proposed design significantly reduces area and power consumption by 75% and 65.1%, respectively. Furthermore, the experimental results demonstrate the proposed block is in good shape for security metrics, comparable to existing works, which can be effectively used for security applications and key generation.

The remainder of the paper is organized as follows: Section II explains the proposed approach for area and power reduction and verifies associated concepts. Section III evaluates the design in different aspects. Finally, Section IV summarizes the findings of this work.

II. PROPOSED DESIGN

This work proposes a lightweight design for CRO PUF that leverages various sources of variation inherent in existing CRO architecture [7]. The logic-based architecture uses a 2-input logic gate, where one of the inputs is the configurable bit; the other bit fluctuates/bypasses the oscillation signal. Physically, at least four transistors are connected in n-well and p-well networks according to their functionality. Therefore, the accumulated delay variation of their connection grows accordingly. Moreover, the connection of the inputs at the transistor level and the transition of output to the equivalent states vary the switching delay and the phase shift, respectively. Hence, this study exploits variations in frequency and phase shift at the level of the delay units instead of RO. To reduce both area and power consumption, the set of ROs is replaced with a pair of ROs, and the delay unit is constructed from hybrid logic gates and defined as a delay stage. The delay stage S in Fig. 1A is configured by a 4-bit challenge and the overall structure of the proposed design is shown in Fig. 1B. n stages are connected sequentially, and the output of each is applied as one of the Mux inputs. The counters are triggered by Mux output and enabled for a certain period defined by the operating time of PUF, set empirically during the post-manufacturing process.

For validation purposes, a RO of four stages has been connected as Fig. 1B, and the output of the stages has been analyzed for the same 4-bit challenge utilizing the Cadence simulation tool and using standard cells of 22nm technology. The duty cycle of stages 1, 2, 3, and 4 are 44.33%, 49.83%, 55.33%, and 61.03%, respectively, as shown in Fig. 2. To amplify inter-stage variations, the operating frequency is set based on the average stage delays such that $T_{stage_{min}} < T_{operating} < T_{stage_{max}}$, which on average evaluates 50% of responses as logic-1 and the other half as logic-0. Practically, the operating frequency is calculated by evaluating the quality of the outputs. Fig. 2 demonstrates the response extraction approach. The counters are synchronously enabled for $T_{operating}$, which defines the evaluation window. Fig. 2 shows that stage2 and stage3 are included in the evaluation window, while the phase variation of stage1 and stage4 plays a crucial role in determining whether stage1 and stage4 are part of the evaluation process or not. Consequently, Mux selection lines select the final components that trigger or not the related counters. During the generation process, the counter is irregularly enabled/disabled multiple times to amplify the variation between ROs. The response generation time is set to minimal to prevent a high number of oscillations in ROs so that the variation in frequencies does not overwhelm the variation in stage phases. Because the randomness comes from changes in both phase and frequency, the suggested design lowers power usage. About 50% of the delay units bypass the input signal without negation. Additionally, the minimal number of oscillations helps decrease switching activity, as indicated by the formula $P_{switching} = \alpha CV_{dd}^2 f$, where $\alpha$ is the activity factor, $C$ the effective capacitance, $V_{dd}$ is the supply voltage, $f$ the switching frequency.

III. EXPERIMENTAL RESULTS

The proposed design was implemented in 22 nm FDSOI technology using the Synopsys tools. The proposed PUF has been taped out as part of RISC-V based SOC reported in [15] and shown in Fig. 3. A Serial-to-Parallel Interface (SPI) has been developed to communicate with the block through the Genesys-2 FPGA board to collect a large set of CRPs.
A. Functionality

PUF block is characterized by a set of metrics that define the efficiency of the produced response from a single PUF device, which in sequences has a significant impact on post-processes to guarantee a secure and reliable key. PUF quality is defined by uniformity, diffuseness, reliability, and uniqueness. The uniformity represents the Hamming Weight (HW) of the generated response. The diffuseness evaluates the difference between responses generated by the same device utilizing different challenges. Intra Hamming Distance (IntraHD) or reliability reflects the stability of the response under different environmental conditions, including temperature and voltage variations. Inter Hamming Distance (InterHD) or uniqueness evaluates the differences between responses generated from different devices utilizing the same challenge. Given that, \( R_{ijt} \) is the \( j^{th} \) response of the \( i^{th} \) device, generated by applying the challenge under certain environmental conditions defined by \( t \). \( J \) is the total number of applied challenges (generated responses) and \( I \) is the total number of PUF devices (excluding the device under test). Then, Uniformity\(_j\) = HW(\( R_{ij} \)), Diffuseness\(_j\) = HD(\( R_{ij} \), \( R_{im} \), \( m \neq j, m = \{1, 2, ..., J\} \)), IntraHD\(_j\) = HD(\( R_{ij} \), \( R_{jt} \)) and InterHD\(_i\) = HD(\( R_{ij} \), \( R_{ij} \), \( i \in I \)). On average across eight chips and utilizing 1M CRPs from each, the proposed design achieves 49.42%, 38.25%, 9.95%, and 45.5% for uniformity, diffuseness, reliability, and uniqueness, respectively. Considering the ideal values for uniformity, diffuseness, and uniqueness are 50% and 0% for reliability. The performance of each chip is reported in Table I.

### Table I: PUF Quality Across Eight Chips

<table>
<thead>
<tr>
<th>Chip</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniformity</td>
<td>48.4</td>
<td>46.6</td>
<td>49.2</td>
<td>48.7</td>
<td>47.5</td>
<td>52.5</td>
<td>44.7</td>
<td>61.2</td>
</tr>
<tr>
<td>Diffuseness</td>
<td>38.3</td>
<td>40.9</td>
<td>34.8</td>
<td>36.7</td>
<td>40.4</td>
<td>37.2</td>
<td>40.3</td>
<td>33</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>47.1</td>
<td>45.5</td>
<td>43.2</td>
<td>43.8</td>
<td>46.6</td>
<td>46.4</td>
<td>43.9</td>
<td>52.9</td>
</tr>
</tbody>
</table>

![Fig. 2](image1.png)  
Analyzing the output of RO’s stages at particular periods. The waveforms show the output of the stages with the active evaluation time defined by enabling counters for \( T_{ceating} \), which is the average delay of the involved stages.

![Fig. 3](image2.png)  
The layout of the proposed PUF (left), zoomed out from the 22nm die (right).

![Fig. 4](image3.png)  
The probability distribution of PUF quality metrics calculated from 1M CRPs.

![Fig. 5](image4.png)  
Raw BER values with repeated challenge up to 2500 times.

Data analysis and evaluation of the collected CRPs have been carried out in MATLAB and Python. Eight chips (PUF devices) have been subjected to CRPs collection and analysis, where 1K CRPs have been extracted from each chip utilizing the same set of challenges generated randomly. For intensive security analysis, 1M CRPs are extracted from one of eight chips that show good results in the primitive analysis.
Fig. 6: Probability distribution of IntraHD due to temperature (right) and voltage (left) variations to the reference ($25^oC$, $800mV$).

B. Randomness

The interdependence of multiple bitstreams has been investigated by correlation factor computation and the application of NIST 800-22 tests for random sequence. The spatial correlation is evaluated by the Pearson correlation, where $1K$ bitstreams have been collected from 32 adjacent PUF units and the correlation factor is computed for each possible pair as shown in Fig. [7]-left. Generally, most of the measurements exhibit a low correlation between adjacent cells, and the matrix does not display identifiable patterns. Moreover, the auto-correlation factor (ACF) is computed over consecutively collected $1K$ CRPs, which is equal to $\pm 0.006701$ at 95% confidence level as shown in Fig. [7]-right, where the light green illustrates the confidence bound. To assess the randomness of the generated sequence, 12 tests are applied considering the minimum input length required for every test and carried out by NIST 800-22 tool. The 128-bit sequence has been selected from each chip out of 8 chips. The sequence concatenates four 32-bit responses by applying four random challenges in the same order on all chips [16] [17]. Hence, the stream length is $128 \times 8 = 1024$ and the block size for related tests is 3. In Table II the listed tests are all passed with the P-values greater than 0.01, which means a 128-bit key is considered random across PUF devices with 99% confidence.

C. Security Attack

The resiliency of the proposed design against machine learning (ML) based modeling attacks has been investigated utilizing two machine models: Support Vector Machine (SVM) and Artificial Neural Network (ANN). $1M$ CRPs have been utilized, where 80% are used for training and 20% for testing. For both models, the challenge is the model's input that is used to predict the response bit individually. A linear kernel has been applied in SVM. ANN has an input layer of size 32 (challenge size) and 3 fully connected hidden layers of sizes: 256, 128, and 32 with the ReLU activation function. The output layer is evaluated via a sigmoid function. The batch size is 50, and the model is trained over 100 epochs. The prediction accuracy of both models is shown in Fig. [8].

Finally, the proposed architecture is compared with various structures in the literature in terms of design characteristics, performance, resource utilization, and ML resiliency, which is concluded in Table. III. The design achieved at least an area reduction in terms of the number of gates by 75% compared to [7], [21], [22], and in terms of space by 97% compared to [7], [19], [20]. Also, the power consumption is reduced by 65.1% compared to [7], [18], [19], [21].

### IV. Conclusion

The lightweight implementation of PUFs aligns with their use in resource-constrained IoTs. This work proposes an efficient approach to reduce area overhead and power consumption by shrinking the ROs set, minimizing switching activity, and emphasizing inter-stage delays. The accumulated delays configure the relative frequency, determining each stage's contribution to the generation process during the evaluation time. The proposed designs, developed and fabricated using 22nm FDSOI technology, exhibit strong performance, achieving, on average, 49.42%, 38.25%, 9.95%, and 45.5% for uniformity, diffuseness, reliability, and uniqueness, respectively. The proposed architecture passes the NIST test, shows low correlations, and exhibits no discernible patterns in adjacent or sequence outputs. Furthermore, the design’s resistivity against ML-based modeling attacks was tested using SVM and ANN.
TABLE III: PERFORMANCE COMPARISON

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology (nm)</td>
<td>28</td>
<td>65</td>
<td>40</td>
<td>-</td>
<td>45</td>
<td>28</td>
</tr>
<tr>
<td>Target Hardware</td>
<td>ASIC</td>
<td>ASIC</td>
<td>ASIC</td>
<td>FPGA</td>
<td>FPGA</td>
<td>FPGA</td>
</tr>
<tr>
<td>Entropy Source</td>
<td>NAND</td>
<td>Impedance Booster</td>
<td>XOR</td>
<td>Inverters</td>
<td>XOR</td>
<td>Logic Gates</td>
</tr>
<tr>
<td>Challenge Size (bit)</td>
<td>n</td>
<td>32</td>
<td>96</td>
<td></td>
<td>64</td>
<td>16</td>
</tr>
<tr>
<td>Area (µm²)</td>
<td>193.5</td>
<td>673.92</td>
<td>844.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Area/Bit (µm²/bit)</td>
<td>1.24</td>
<td>70</td>
<td>70</td>
<td>1.3</td>
<td>0.99</td>
<td>1.24</td>
</tr>
<tr>
<td>Power (mW)</td>
<td>378.9</td>
<td>61.4</td>
<td>0.03</td>
<td>48.1</td>
<td>-</td>
<td>119.4</td>
</tr>
<tr>
<td>Max No. CRPs</td>
<td>-</td>
<td>-</td>
<td>2.28</td>
<td>2.16</td>
<td>8.3</td>
<td>2.2</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>-40~150</td>
<td>0~80</td>
<td>-25~125</td>
<td>-</td>
<td>0~60</td>
<td>5~70</td>
</tr>
<tr>
<td>Voltage (V)</td>
<td>0.8~0.99</td>
<td>0.8~1.3</td>
<td>0.7~1.2</td>
<td>0.8~1.2</td>
<td>1.16~1.24</td>
<td>0.72~0.88</td>
</tr>
<tr>
<td>Reliability (%)</td>
<td>10.5</td>
<td>0.6</td>
<td>1.01</td>
<td>7.13</td>
<td>1.71</td>
<td>1.78</td>
</tr>
<tr>
<td>Uniformity (%)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Uniqueness (%)</td>
<td>49.78</td>
<td>49.8</td>
<td>50.1</td>
<td>44.64</td>
<td>44.38</td>
<td>46.76</td>
</tr>
<tr>
<td>Throughput (Mbps)</td>
<td>128</td>
<td>0.32</td>
<td>1.6</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Size of Trained CRPs</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ML Prediction Accuracy (%)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

G: gate, I: inverter, *: estimated or calculated

models, achieving a prediction accuracy of 67.1% utilizing 1M CRPs.

ACKNOWLEDGMENTS

This work is funded and supported by Technology Innovation Institute (TII) [grant: EX2021-005]. Additional support from Khalifa University, System on Chip Lab (SoC) [grant: RC2-2018-020]

REFERENCES