A Systematic Method to Evaluate and Compare the Performance of Physical Unclonable Functions

Abhranil Maiti, Vikash Gunreddy, and Patrick Schaumont

Virginia Tech
Blacksburg, Virginia, USA
{abhranil, gvikash7, schaum}@vt.edu

Abstract. In this work, we propose a systematic method to evaluate and compare the performance of Physical Unclonable Functions (PUFs). The need for such a method is justified by the fact that various types of PUFs have been proposed so far. However, there is no common method that can fairly compare them in terms of their performances. We first propose three generic dimensions of PUF measurements. We then define several parameters to quantify the performance of a PUF along these dimensions. We also analyze existing parameters proposed by other researchers. Based on our analysis, we propose a compact set of parameters that will be used as a tool to evaluate as well as compare the performance of different PUFs. To make the method independent of the underlying PUF technique, we focus on the statistical properties of the binary PUF responses. Finally, we show a detailed comparison analysis between two PUFs: ring-oscillator-based PUF (RO PUF) and Arbiter-based PUF (APUF) using measured PUF data.

Keywords: Physical Unclonable Function, Challenge, Response, Ring Oscillator, Arbiter, Parameter, Performance.

1 Introduction

An on-chip Physical Unclonable Function (PUF) is a chip-unique challenge-response mechanism exploiting manufacturing process variation inside integrated circuits (ICs). The relation between a challenge and the corresponding response is determined by complex, statistical variation in logic and interconnect in an IC. A PUF has several applications in the field of hardware-oriented security. For example, it can be used in device authentication and secret key generation [20]. Guajardo et al. discussed the use of PUFs for Intellectual Property (IP) protection, remote service activation and secret-key storage [5]. A PUF-based RFID tag has also been proposed to prevent product counterfeiting [3, 2].

Since the inception of the idea of PUFs, different types of PUFs have been proposed so far. For example, Lim et al. proposed Arbiter PUF that exploits the delay mismatch between a pair of identically laid-out delay paths [11]. A PUF which is based on random start-up values of SRAM cells was proposed

by Guajardo et al [6]. A PUF based on an array of identically laid-out ring oscillators has also been proposed [20]. There are several other PUFs which are either an enhanced version of a previously proposed PUF or introduce a new method of generating PUF challenge-response pairs (CRPs).

The availability of several different PUFs gives us choices to select a particular one suitable for an application. However, it also raises few practical questions: how do we know if a PUF is efficient or not? How do we compare one PUF with another? Currently, there is no readily available method to fairly compare one PUF with another. A concrete as well as easy-to-use evaluation-comparison method will be useful for a designer who may want to employ a PUF in her design. Armknecht et al. expressed the same view in their work on formalizing the security features of PUFs [1]. In this paper, we propose a systematic method to evaluate the performance of PUFs and to make a fair comparison among them. As part of this work, we have identified the following goals:

First, we need to clearly define parameters that will quantify the performance of a PUF in a concrete manner. In order to do that, we not only propose our own PUF parameters but also explore several other parameters defined by other researchers. No analysis has been carried out so far to compare these parameters to estimate how effective they are in evaluating performance of a PUF. It might be possible that many of these parameters are similar in nature or redundant. We aim to find any such cases to define a compact set of PUF parameters while removing redundancy.

Second, we aim to make the comparison method independent of the underlying PUF circuit. For example, it should be able to compare a delay-based PUF like an RO PUF [20] with a memory-based PUF such as SRAM PUF [6]. Therefore, we focus on the statistical properties of the binary PUF responses. This is possible because every PUF produces binary responses (or responses can be converted to binary form) when supplied with challenges irrespective of the underlying technique. With these goals in mind, we present the following contributions in this paper.

- We first propose three measurement dimensions of a PUF. They are : device, time and space. The PUF performance parameters will be defined along these dimensions. We explain the significance of these dimensions in detail and show how several PUF parameters can be defined based on them.

- We propose that a set of m parameters be used to evaluate and compare the performance of different PUFs while each PUF may have different number of challenge and response bits. This is possible as the parameters purely rely on the statistical properties of the binary PUF responses. A simple view of the idea is presented in Figure 1. As a preliminary effort, we propose seven parameters: uniqueness, reliability, randomness, steadiness, bit-aliasing, diffuseness and probability of misidentification as part of the method. We have defined some of these parameters while others have been selected from the works done by other researchers. We explain in detail why these parameters are useful in evaluating the performance of PUFs.

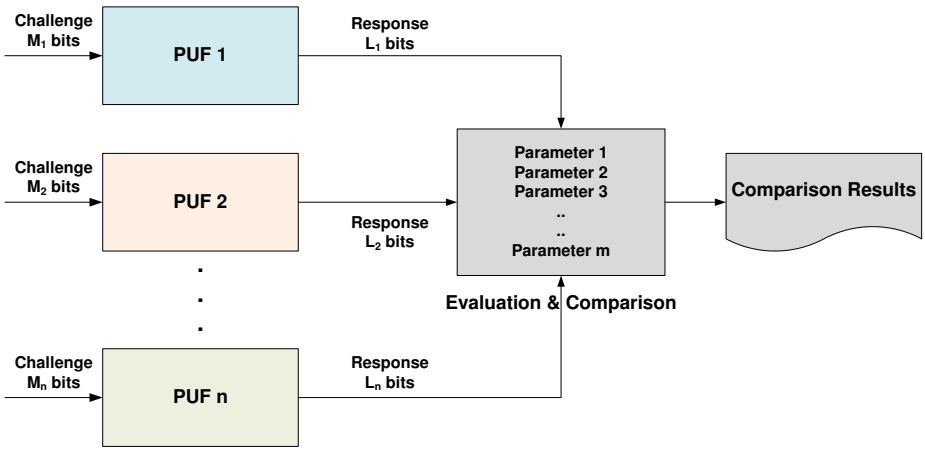

Fig. 1. Basic idea of a PUF Evaluation and Comparison Method

- Finally, we compare two different PUFs: the RO PUF and the APUF using the above mentioned parameters. We used measured PUF data for this comparison. A detailed comparison result is presented in this paper to validate the proposed method.

The rest of the paper is organized as follows. Section 2 gives an overview of different PUFs proposed so far in the research community. It also discusses several works related to the evaluation and comparison of performance of different PUFs. Section 3 introduces the dimensions of PUF measurements, defines four PUF parameters and analyzes few PUF parameters from the literature. Based on our analysis, we propose a final set of parameters as the main building block of the evaluation-comparison method. Section 4 presents an analysis to compare the RO PUF with the APUF. Finally, we conclude the paper in Section 5.

2 Background

In this section, we briefly discuss the history of the PUF technology since it was introduced. We also discuss several research works that are related to PUF performance evaluation.

2.1 Chronology of PUFs

One of the seminal works in the area of PUF is that of Lofstrom et al. in 2000 exploiting mismatch in silicon devices for identification of ICs [12]. Though the authors did not call it a PUF, the objective of their work was very similar to that of a PUF. Around the same time (in 2001), Pappu et al. presented the concept of physical one-way function which led to the idea of PUF [18]. After that, many

Table 1. Different Types of PUFs

2000	IC identification using device mismatch [12]
2001	Physical One-way Function [18]
2002	Physical Random Function [4]
2004	Arbiter PUF [11]
2006	Coating PUF [22]
2007	Ring Oscillator PUF[20], SRAM PUF [6]
2008	Butterfly PUF [10]
2009	PUF using power distribution system of an IC [7]
2010	Glitch PUF [21]
2011	Mecca PUF [9]

different types of PUFs have been proposed. Almost every year, there has been at least one new PUF circuit that was proposed. Table 1 shows a year-wise list of different PUFs starting from the year 2000. Though this list is not exhaustive, it shows us a picture about how the researchers tried to build different PUFs. For more information, one may refer to the work by Maes et al. which presents a comprehensive discussion on different PUFs proposed so far [14].

Variety of PUFs does support the need to build a systematic method to evaluate and compare their performances. Despite the existence of multiple PUF techniques, not many of them have been actually integrated in a system so far. An evaluation-comparison method may make it easier for a system designer to select a PUF that suits best for a particular application leading to more utilization of the PUF technology.

2.2 Related Work

We discuss few works that primarily focus on the performance of PUFs. Majzoobi et al. proposed several parameters to test the security of PUFs [16]. They mainly tested three security properties of a PUF: predictability, sensitivity to component accuracy and susceptibility to reverse engineering. Two variants of the Arbiter PUF, linear and feed-forward, were tested. However, no comparison of different PUFs were done.

In another work, Armknecht et al. formalized three properties of a PUF: robustness, unclonability and unpredictability [1]. The analysis result presented in this work is based on SRAM-based PUF proposed by Gujardo et al. [6]. No comparison between multiple types of PUFs were presented in this work.

Van der Leest et al. tested the performance of D flip-flop-based PUF implemented on ASIC using several parameters [23]. This PUF was originally proposed by Maes et al. using flip-flops on FPGAs [13]. This work applied Hamming weight test, inter-chip uniqueness test, NIST randomness test on PUF responses. This paper presented a comparison between SRAM-based PUF, Butterfly PUF [10] and DFF-based PUF. However, this comparison was not the primary focus

of this work, and did not use PUF parameters based on statistical properties of PUF responses. Instead, the comparison was made based on the number of gates used to implement the PUF, platform used (ASIC/FPGA), spread on die and entropy reported in the respective works.

A comprehensive performance evaluation of APUF has been done by Hori et al [8]. In this work, several PUF parameters were defined systematically and were validated based on a large set of PUF data. The PUF parameters proposed by this work are uniqueness, randomness, correctness, steadiness and diffuseness. We have studied this work as a part of the PUF comparison effort in this paper. We will discuss this work in detail in the subsequent sections. However, this work also did not present any comparison analysis on different types of PUFs.

The work by Maes et al. presented a detailed comparative analysis between several PUFs [14]. One of the comparisons presented in this work was made using several qualitative properties of PUFs such as evaluability, uniqueness, reproducibility, physical unclonability, mathematical unclonability, unpredictability, one-way-ness and tamper evidence. Another comparison analysis, presented in this work, includes randomness, challenge-response mechanism, CRP space, implementation platform, average inter-chip and average intra-chip variation in PUF responses, entropy, tamper-evidence and model building as the comparison metrics. The results presented for the CRP space, the entropy, the inter- and intra-chip variation were quantitative while the others were qualitative.

Our contribution in this paper is different in many ways from the related work discussed. First, we concretely define a basis for the PUF performance measurement/evaluation. We propose three measurement dimensions for this purpose. The PUF measurement dimensions have not been formally defined before. We explain how these dimensions capture useful information in order to evaluate the performance of PUFs. We, then, define parameters using the proposed dimensions to evaluate the statistical property of the PUF responses. Hori et al. defined their PUF parameters using a similar approach but they did not explicitly define the dimensions [8]. Second, we analyze several PUF evaluation parameters to point out any redundancy that may exist among them. Based on our analysis, we propose a compact set of parameters for PUF evaluation as well as comparison. We did not find any work in the literature that has made a similar effort. Finally, we compare two different PUFs: the RO PUF and the Arbiter PUF. Unlike previous efforts, our comparison is done entirely quantitatively using the parameters we proposed.

3 PUF Evaluation and Comparison Method

In this section, we first introduce the PUF measurement dimensions. Then we define four PUF parameters to quantify several important quality factors of a PUF. Next, we analyze few parameters defined by other researchers. We compare them with the parameters we defined. Finally, we propose a set of parameters as the components of the evaluation-comparison method.

3.1 PUF Measurement Dimensions

Figure 2 shows three different dimensions of PUF measurement along three axes: device, space and time. The inter-chip variation in PUF responses is captured using the device axis. The two other axes are used to capture the intra-chip variation.

The device axis represents the population dimension of PUF measurements. A PUF not only needs to generate random responses for a chip, the generated responses also need to uniquely identify the chip among several other chips of the same type. To estimate this property, one needs to measure a population of PUF instantiations on several chips/devices. Hence, we included the device axis. A group of k devices have been shown in Figure 2 to represent a population.

The space axis stands for the location of a single-bit response, r in an m-bit response string, R. The rationale behind naming it the space axis is that most of the PUF response bits are generated at different physical locations on a chip. For example, in an SRAM PUF, the SRAM cell that produces a response i has an on-chip location that is different from that of another SRAM cell that produces a response j. For Arbiter PUF, the location of the Arbiter that produces the response is fixed. However, the locations of the stimulated delay paths change depending on the challenge (whether straight connections or criss-cross connections through a switch). To estimate the randomness of a PUF, we examine multiple response bits from a PUF. Hence, the space dimension becomes useful.

Finally, the time-dependent properties of a PUF are captured along the time axis. One critical attribute of a PUF is the reliability of the responses. It estimates how consistently the responses can be generated against varying operating conditions such as variable ambient temperature and fluctuating supply voltage. To estimate the reliability, we take multiple samples of the responses at different instances of time. Samples of PUF responses are also useful in estimating the circuit aging effect on PUFs.

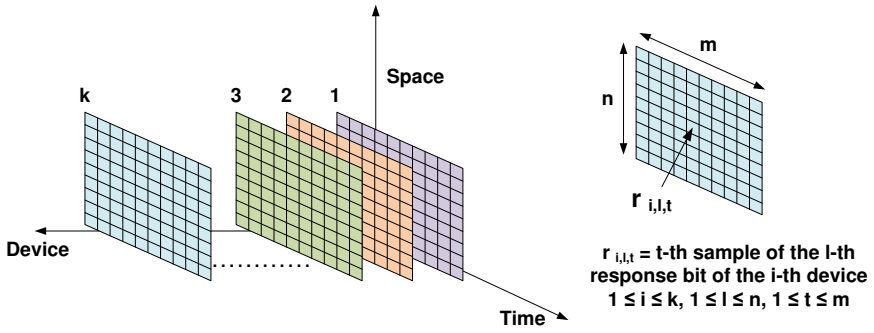

Fig. 2. Dimensions of PUF Measurement

3.2 Defining PUF Parameters

Now, we will define four PUF parameters based on the measurement dimensions introduced. These parameters are: uniqueness, reliability, uniformity and bit-aliasing. We formalized these parameters in one of our previous works on PUF characterization [15]. Each of these parameters quantifies an essential quality factor of a PUF as we will explain them in detail.

1) Uniqueness: Uniqueness represents the ability of a PUF to uniquely distinguish a particular chip among a group of chips of the same type. We use Hamming distance (HD) between a pair of PUF identifiers to evaluate uniqueness. If two chips, i and j ($i \neq j$), have m-bit responses, R_i and R_j respectively for the challenge C, the average inter-chip HD among k chips is defined as

$$\text{Uniqueness } = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{HD(R_i, R_j)}{n} \times 100\% \tag{1}$$

It is an estimate of the inter-chip variation in terms of the PUF responses, and not the actual probability of the inter-chip process variation. In Figure 3, it is shown that the inter-chip HD is estimated along the device axis. One comparison is shown in dark gray between the device 3 and the device k. Another comparison is shown in light gray between the device 1 and the device 3.

2) Reliability: PUF reliability captures how efficient a PUF is in reproducing the response bits. We employ intra-chip HD among several samples of PUF response bits to evaluate it. To estimate the intra-chip HD, we extract an n-bit reference response (R_i) from the chip i at normal operating condition (at room temperature using the normal supply voltage). The same n-bit response is extracted at a different operating condition (different ambient temperature or different supply voltage) with a value R'_i. m samples of R'_i are collected. For the chip i, the average intra-chip HD is estimated as follows.

$$\text{HD}_{\text{INTRA}} = \frac{1}{m} \sum_{t=1}^{m} \frac{HD(R_i, R'_{i,t})}{n} \times 100\% \tag{2}$$

where $R'_{i,t}$ is the t-th sample of R'_i. HD_{INTRA} indicates the average number of unreliable/noisy PUF response bits. In other words, the reliability of a PUF can be defined as:

$$\text{Reliability} = 100\% - \text{HD}_{\text{INTRA}} \tag{3}$$

Figure 4 shows how the reliability of a PUF is evaluated using the time dimension of PUF measurement. The two intra-chip Hamming distance measurements are shown along the time axis for the device 3 and the device k with light gray and dark gray respectively.

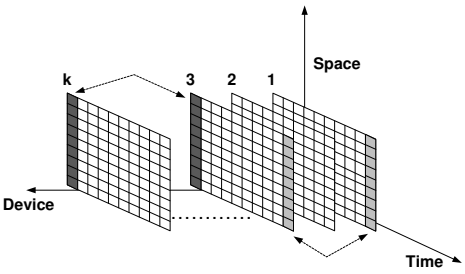

Fig. 3. PUF Uniqueness Evaluation

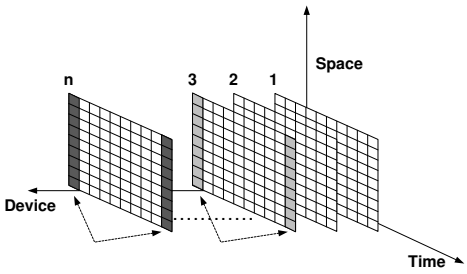

Fig. 4. PUF Reliability Evaluation

3) Uniformity Uniformity of a PUF estimates how uniform the proportion of '0's and '1's is in the response bits of a PUF. For truly random PUF responses, this proportion must be 50%. We define uniformity of an n-bit PUF identifier as the percentage Hamming Weight(HW) of the n-bit identifier:

$$(\text{Uniformity})_i = \frac{1}{n}\sum_{l=1}^{n} r_{i,l} \times 100\%$$

where $r_{i,l}$ is the $l-$th binary bit of an $n-$bit response from a chip i. (4)

In Figure 5, the uniformity of the device k and the device 3 are evaluated along the space axis (marked in dark gray and light gray respectively).

4) Bit-aliasing If bit-aliasing happens, different chips may produce nearly identical PUF responses which is an undesirable effect. We estimate bit-aliasing of the l-th bit in the PUF identifier as the percentage Hamming Weight(HW) of the l-th bit of the identifier across k devices:

$$(\text{Bit} - \text{aliasing})_l = \frac{1}{k}\sum_{i=1}^{k} r_{i,l} \times 100\%$$

where $r_{i,l}$ is the $l-$th binary bit of an $n-$bit response from a chip i. (5)

In Figure 6, the bit-aliasing is evaluated along the device axis for two different bit locations (marked in dark gray and light gray).

3.3 Analysis of PUF Parameters Defined by Other Researchers

In this section, we explore different PUF parameters defined by other researchers in the community. We also try to find out if there is any redundancy among these parameters. Table 2 shows few examples of different PUF parameters proposed by several research groups. In this work, we will be presenting an analysis of the parameters proposed by Hori et al. in [8] with a comparison of the parameters introduced in Section 3.1. The reason we chose this work for analysis is that it

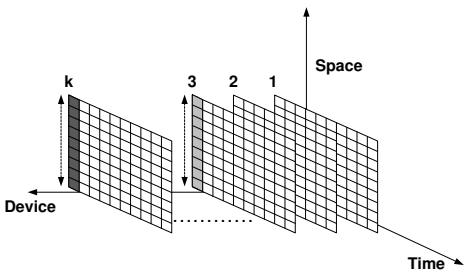

Fig. 5. PUF Uniformity Evaluation

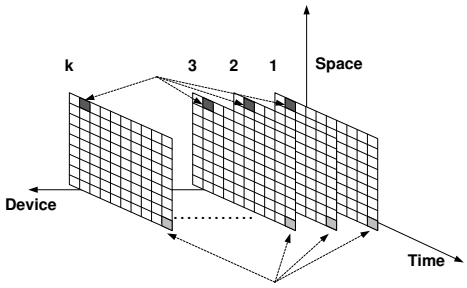

Fig. 6. PUF Bit Aliasing Evaluation

has a similar effort to define PUF parameters like ours. Moreover, the authors of this work made a large PUF dataset, based on APUF, available for analysis. This motivated us to use the APUF dataset in carrying out a detailed comparison analysis with a similarly large dataset based on RO PUF that we generated. We also analyze the parameter "probability of misidentification" proposed by Su et al [19]. This parameter estimates the rate of false positives in chip identification for a given number of noisy bits in the PUF responses.

There are several other parameters existing in the literature. For example, Majzoobi et al. defined parameters such as single-bit probability and conditional probability [16]. A parameter called variety has been proposed by Yamamoto et al [25] [1]. As part of the future effort, we plan to analyze many of these parameters to enhance the evaluation-comparison method.

Table 2. Different PUF Parameters

Hori et al. [8]	Randomness Steadiness Correctness Diffuseness Uniqueness
Maiti et al. [15]	Uniformity Bit-aliasing Uniqueness Reliability
Su et al. [19]	Probability of Misidentification
Majzoobi et al. [16]	Single-bit Probability Conditional Probability
Yamamoto et al. [25]	Variety

[1] This term was introduced in the slides for the presentation of the paper [25] by the authors in CHES,2011. [24]

At first, we introduce few notations that will be used to describe the parameters. The notations are:

$$
\begin{aligned}
N &= \text{ total number of chips} \\
n &= \text{ index of a chip } (1 \leq n \leq N) \\
K &= \text{ total number of identifiers(IDs) generated per chip} \\
k &= \text{ index of an ID in a chip } (1 \leq k \leq K) \\
T &= \text{ total number of samples measured per ID} \\
t &= \text{ index of a sample } (1 \leq t \leq T) \\
L &= \text{ total number of response bit in an ID} \\
l &= \text{ index of a response bit } (1 \leq l \leq L) \\
M &= \text{ total number of ring oscillators} \\
m &= \text{ index of an oscillator } (1 \leq m \leq M)
\end{aligned}
$$

The above notations, apart from m and M for the ring oscillators, have been proposed by Hori et al. [8]. We decide to keep these notations to define any PUF parameters except the fact that we use r in place of b (used in [8]) to denote a single response bit from a PUF. We notice that the parameters in Section 3.1 used k as the size of the device population, n as the total number of response bits from a PUF and m as the number of samples of the response bits. While we compare the parameters defined by the two groups, we will express all the PUF parameters using the notations above.

We have proposed four parameters as described in Section 3.1. They are uniqueness, reliability, bit-aliasing and uniformity. The five parameters proposed by Hori et al. are uniqueness, randomness, correctness, steadiness and diffuseness. Upon analyzing these parameters, we have found that there are similarities among these parameters. Figure 7 shows the relation between these parameters. There are three parameters from each group that are similar in definition while others are different. We explain all these parameters in detail.

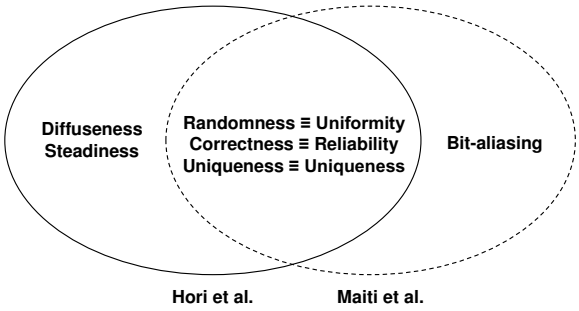

Fig. 7. Relation between parameters defined by Hori et al. and this work

Randomness vs Uniformity: The randomness by Hori et al. is defined below.

$$- log_2 max(p_n, 1 - p_n) \tag{6}$$

$$\text{where } p_n = \frac{1}{K.T.L} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{l=1}^{L} r_{n,k,t,l} \tag{7}$$

On the other hand, the uniformity parameter by us has been defined as follows.

$$\frac{1}{K.L} \sum_{k=1}^{K} \sum_{l=1}^{L} r_{n,k,l} \tag{8}$$

It can be noticed that the randomness includes T samples of the response bits while the uniformity does not include the samples and is based on the reference/correct bits only. Also, the randomness is expressed in the min-entropy form. Otherwise, the two definitions are very similar in nature. Both of them estimate the ratio of '1' vs '0' in all the response bits generated by a PUF.

Correctness vs Reliability: The correctness parameter is defined as follows.

$$1 - \frac{2}{K.T.L} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{l=1}^{L} (r_{n,k,l} \oplus r_{n,k,t,l}) \tag{9}$$

On the other hand, the reliability is defined as below.

$$1 - \frac{1}{K.T.L} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{l=1}^{L} (r_{n,k,l} \oplus r_{n,k,t,l}) \tag{10}$$

The reliability parameter is different from the correctness only in terms of the factor by which it is normalized. The reliability is calculated based on the average value of the intra-chip HD whereas the correctness is normalized by the maximum value of the fractional Hamming distance between the correct ID (the ID which is considered as the reference) and the sample IDs. There are few points that are not clearly mentioned in the definition of the correctness parameter.

It is defined using the sum of Hamming Distances (SHD) normalized by T, K and L. A parameter $c_{n,k,l}$ has been defined as the SHD between the correct bit $r_{n,k,l}$ and the generated bit $r_{n,k,t,l}$ through T tests.

$$c_{n,k,l} = \sum_{t=1}^{T} r_{n,k,l} \oplus r_{n,k,t,l} \tag{11}$$

However, it is not clear from the definition what the time instances of the measurements are. Since it is used to capture the effect of device defect or aging, it can be assumed that the correct bit $r_{n,k,l}$ is measured before the aging or defect

whereas $r_{n,k,t,l}$ is measured after the aging effect. If that is the case then the following inequality (mentioned in [8])does not necessarily hold good.

$$0 \leq c_{n,k,l} = \sum_{t=1}^{T} r_{n,k,l} \oplus r_{n,k,t,l} \leq \frac{T}{2} \tag{12}$$

This is because there might be a case when the aging or device defect might flip a correct bit in such a way that the subsequent T samples produce a complementary value for more than $T/2$ samples. This inequality holds good *always* only if both $r_{n,k,l}$ and $r_{n,k,t,l}$ are measured during the same sampling instance which contradicts the definition of correctness.

Uniqueness by Hori et al. vs Uniqueness by Maiti et al.: The uniqueness is defined by Hori et al. as:

$$\frac{1}{K.L}\frac{4}{N^2}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(r_{i,k,l} \oplus r_{j,k,l}) \tag{13}$$

The uniqueness is defined by us as:

$$\frac{1}{K.L}\frac{2}{N(N-1)}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(r_{i,k,l} \oplus r_{j,k,l}) \tag{14}$$

In the case of the uniqueness, two definitions differ from each other with respect to the normalization factor. Hori et al. used the sum of Hamming distance (SHD) of all the possible combinations of the PUF identifiers as the normalization factor. For a population of n chips, the value of that factor is $K.L.N^2/4$. On the other hand, we used the total number of all possible pairwise combinations of response bits as the normalizing factor whose value is $K.L.N.(N-1)/2$. For a large value of N, the normalization factor used by us. is approximately two times bigger than that used by Hori et al.

Parameters uniquely defined by both the groups: The term bit-aliasing is uniquely defined by us. On the other hand, the steadiness and the diffuseness are uniquely defined by Hori et al. The diffuseness is same as the uniqueness except the fact that the diffuseness is defined inside a single chip among several different IDs while the uniqueness is measured across several chips.

Steadiness: The steadiness measures the degree of bias of a response bit towards '0' or '1' over T samples. It is defined as:

$$S_n = 1 + \frac{1}{K.L}\sum_{k=1}^{K}\sum_{l=1}^{L} log_2 max(p_{n,k,l}, 1 - p_{n,k,l}) \tag{15}$$

$$where\ p_{n,k,l} = \frac{1}{T}\sum_{t=1}^{T} r_{n,k,t,l} \tag{16}$$

This parameter is somewhat similar to the correctness parameter. A lower value of steadiness will produce a lower correctness. In this case also, the time stamps of the sample measurements are important. This is because the steadiness of a PUF may change when operating conditions change. However, Hori et al. did not discuss the effect of time on the steadiness parameter [8].

Diffuseness: The diffuseness is defined as:

$$\frac{1}{L}\frac{4}{K^2}\sum_{l=1}^{L}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}(r_{n,i,l} \oplus r_{n,j,l}) \tag{17}$$

One question arises regarding the diffuseness parameter. Since the diffuseness is estimated among K signatures (L bits each) generated in a chip, its value might change depending on how we create a group of bits to produce an ID. For example, there are a total of $K \times L$ binary bits in K L-bit signatures. These $K \times L$ bits can be divided in many possible K groups with L bits each. Therefore, the same set of $K \times L$ bits can lead to different values of diffuseness based on the combination we select. Since, the PUF challenges are selected by a software program [8], the diffuseness can be controlled deterministically.

Bit-aliasing: The bit-aliasing parameter is defined as:

$$(\text{Bit} - \text{aliasing})_{k,l} = \frac{1}{N}\sum_{n=1}^{N} r_{n,k,l} \tag{18}$$

(This parameter has been defined once in Section 3.2. We express it again here in terms of the common notations that have been introduced in the beginning of this section.)

Probability of misidentification (PMSID): Here, we introduce another parameter, probability of misidentification, defined by Su et al [19]. It is a useful parameter to estimate the probability of a chip being falsely identified as another chip due to noise in the response bits.

Suppose a chip X has a reference PUF identifier R_X with L bits. At some other point in time, it produces an identifier R'_X. Therefore, the number of unreliable bits in that PUF is $\text{HD}(R_X, R'_X)$. Now, if there exists another chip Y such that $HD(R_Y, R'_X) \leq HD(R_X, R'_X)$, there will be a misidentification. For a PUF identifier with L response bits, if p is the fraction of the unreliable bits (fractional intra-chip HD) and h is the value of HD between the L-bit identifier of the chip and that of another chip ($h \leq L$), the probability of misidentification is defined as [19]:

$$\sum_{h=0}^{L}\left[\binom{L}{h}0.5^h(1-0.5)^{L-h}.\sum_{h/2}^{h}p^{h/2}(1-p)^{h-h/2}\right] \tag{19}$$

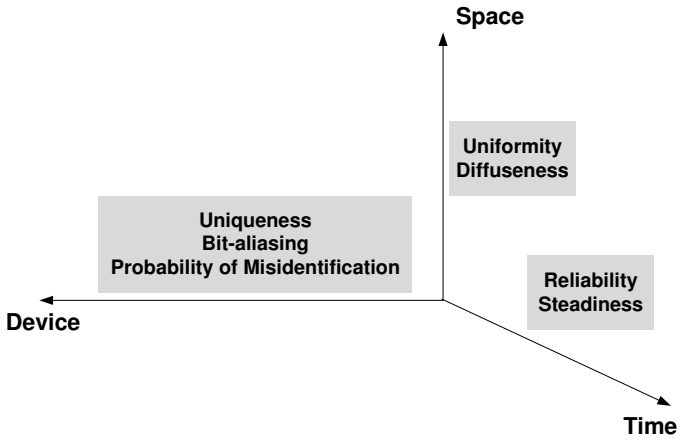

Fig. 8. Final Parameters mapped on the PUF measurement dimension

3.4 Final Set of PUF Parameters

We have discussed several parameters that quantify the quality/performance of a PUF. As a conclusion, we suggest a set of parameters with some modifications as the components of the PUF evaluation-comparison method. It includes seven parameters: uniformity, reliability, steadiness, uniqueness, diffuseness, bit-aliasing and probability of misidentification. Basically, we excluded the redundant parameters while analyzing the parameters proposed by Hori et al., Su et al. and us. This set of parameters will serve as the starting point of the evaluation-comparison method. However, we believe that this set is subject to further modification and enhancement. Figure 8 shows these seven parameters mapped along the proposed PUF measurement dimensions. We briefly discuss why we inlcuded these parameters as the components of the evaluation-comparison method.

1) Uniformity As the analysis shows, uniformity does not include samples of response bits unlike randomness does. If samples are included, it becomes dependent on time. Since we want to evaluate the entropy of the PUF on an average i.e. the entropy of the reference response bits, uniformity is a good fit. Moreover, to estimate the time dependency of a PUF response, we have other parameters such as reliability and steadiness.

2) Reliability Our analysis showed that both the correctness and the reliability are defined in a similar way. However, the time reference is not defined well in case of the correctness. Moreover, the inequality (Eqn(12)) that determines the normalization factor for the correctness has been shown to be not holding good always. Hence, we propose to include the reliability parameter defined in Equation (10).

3) Steadiness Even though the steadiness parameters seems to be closely re-

lated to the reliability/correctness parameter, it represents the bias of individual response bits on an average. Therefore, we suggest to include this parameter as well. However, this parameter needs to be defined based on time stamps i.e. a chip may have different steadiness values depending on the time when it is measured.

4) Uniqueness For the uniqueness parameter, both Hori et al. and us employ average inter-chip HD of the PUF identifier except the normalization factor. The normalization factor used by Hori et al. represents the upper bound of the SHDs among all possible IDs which is an useful information about a PUF. Hence, we suggest to include the uniqueness defined by Hori et al.

5) Diffuseness The diffuseness, as discussed earlier, is very similar to the uniqueness. This parameter becomes useful when a PUF has a large CRP space like APUF and several identifiers can be produced from a single chip. Therefore, we suggest to use the diffuseness parameter when the PUF, being evaluated, has a large CRP space.

6) Bit-aliasing The bit-aliasing parameter is very useful in estimating the bias of a particular response bit across several chips. It may also give us information about any systematic, spatial effect across devices.

7) Probability of misidentification Finally, we also propose to include the probability of misidentification to estimate the rate of error in identification by a PUF. This parameter shows how chip identification may be affected by noise in the response bits.

4 Comparison of RO PUF with Arbiter PUF: A test case

In this section, we compare RO PUF with APUF using the parameters defined by Hori et al. and us as well as the probability of misidentification. We used two datasets for this purpose: a) one dataset consists of RO PUF responses from 193 FPGAs measured by us. b) the other dataset consists of APUF responses from 45 FPGAs measured by Hori et al. First, we briefly describe the two PUFs: the RO PUF and the APUF.

An RO PUF has m identically laid-out ROs and was proposed by Suh et al [20]. A pair of frequencies, f_a and f_b $(a \neq b)$ out of m RO outputs, are selected as challenge to create a response. Due to random process variation, f_a and f_b tend to differ from each other randomly. A response bit r_{ab} is produced by a simple comparison method:

$$r_{ab} = \begin{cases} 1 & \text{if } \ f_a > f_b \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

Since the variability in frequency is random, the response bits produce random binary values. Figure 9 shows a ring-oscillator-based PUF. On the other hand,

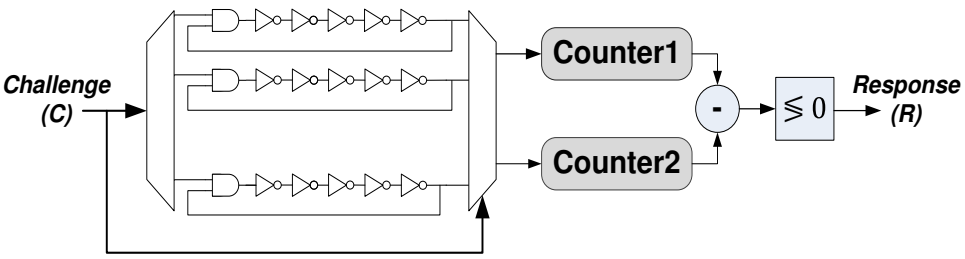

Fig. 9. Ring Oscillator-based PUF

an APUF, proposed by Lim et al., exploits the delay mismatch between a pair of identically routed paths to generate a response bit [11]. Depending on which of the delay path is faster, the arbiter flip-flop produces a '0' or '1' as PUF response. Due to random variation in delay paths, this response bit is random. Several pairs of delay paths can be configured by setting the inputs of the switch components (shown in Figure 10) used as challenge inputs.

Table 3 describes both the datasets that have been used for the analysis. It also includes the device technology used for the respective FPGA implementations. We measured the RO PUF at normal operating conditions. Since Hori et al. did not mention any variation in the operating condition during their measurement, we assume the APUF dataset is also measured under normal operating condition.

We first evaluate the parameters defined by Hori et al. using the RO PUF dataset and compare them with the results from APUF reported in [8]. Since the RO PUF we implemented produces only *one* 511-bit identifier (K=1), the diffuseness is not calculated for the RO PUF dataset. After that, we evaluate the parameters defined by us using the APUF dataset and compare them with the results based on the RO PUF dataset. We also compare both the datasets using the probability of misidentification. Finally, we summarize the comparison based on the set of seven parameters we selected.

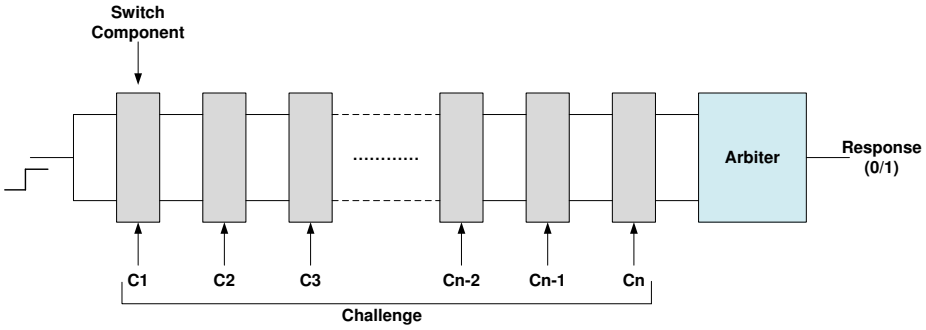

Fig. 10. Arbiter PUF

Table 3. Detail of the datasets used

	RO PUF	**APUF**
N	193	45
T	100	1024
K	1	1024
L	511	128
M	512	-
Device technology	90nm (Spartan 3E)	65nm (Virtex 5)

4.1 Comparison using Parameters defined by Hori et al.:

Table 4 shows the parameters defined by Hori et al. evaluated on both the datasets. The parameter values based on the APUF dataset have been taken from [8]. It can be noticed that the RO PUF shows better randomness compared to the APUF. This is supported by the fact that the average bit probability of RO PUF is close to 0.5 i.e. the RO PUF responses are more equally likely between '0' and '1'. Normally it is expected that a Virtex 5 FPGA on a 65-nm technology will have more variability (hence more randomness in PUF responses) than a Spartan 3E FPGA on 90-nm technology. However, this result shows that the RO PUF has better randomness than the APUF. This indicates that individual PUF technique may have significant influence on extracting the variability information. Another significant difference can be observed in the value of the uniqueness. The uniqueness of the RO PUF is distinctly much higher than that of the APUF. Lower value of the randomness in APUF is one of the reasons why the uniqueness is lower in it. In the next section, we will evaluate the bit-aliasing parameter that may explain this contrast in the uniqueness more. Apart from that, both the PUFs show similar values of the steadiness and the correctness. It implies that the two PUFs are in general highly tolerant to noise at normal operating condition.

Table 5 shows the confidence interval(CI) proposed by Hori et al for a confidence level of 95%. for both the datasets. It can be noticed that the RO PUF dataset show significantly narrower CI compared to the APUF dataset. This

Table 4. Comparison of RO PUF and APUF using parameters defined by Hori et al.

	APUF	**RO PUF**	**Ideal Value**
Average Randomness	84.69%	96.81%	100%
Average Probability	55.61%	49.82%	50%
Average Steadiness	98.48%	98.51%	100%
Average Correctness	98.28%	98.29%	100%
Average Uniqueness	36.75%	94.07%	100%

Table 5. Confidence Interval Comparison Results with 95% Confidence Level

	RO PUF		**APUF**	
	Confidence Interval	Width	Confidence Interval	Width
Randomness	[0.9892, 0.9990]	0.00986	[0.8388, 0.8546]	0.01586
Bit Probability	[0.4962, 0.5003]	0.00407	[0.5530, 0.5591]	0.00611
Steadiness	[0.9846, 0.9857]	0.00110	[0.9626, 1.0000]	0.04134
Correctness	[0.9822, 0.9834]	0.00121	[0.9579, 1.0000]	0.04206
Uniqueness	[0.9334, 0.9481]	0.02940	[0.2127, 0.5222]	0.30950

indicates that the population size of the PUF dataset (RO-PUF having a larger dataset compared to APUF) has substantial impact on determining the confidence interval of the parameters.

4.2 Comparison using Parameters defined by Maiti et al.:

Table 6 shows the average values of the parameters defined by us for both the datasets. We considered 511 response bits for the RO PUF whereas 1024×128 = 131072 response bits were considered for the APUF.

The uniformity result resembles with the average probability reported in Table 4. For the bit-aliasing, the average in the RO PUF is close to the ideal value of 50% while the average in the APUF deviates significantly from 50%. Moreover, we found that the minimum value of bit-aliasing is 0% in case of APUF. This shows that there are bit positions that produce a value of 0 for all the 45 chips in the population. In fact, we found 21314 out of 131072 bit positions (nearly 16%) produced a value of 0%. These bits do not contain any useful information. This is consistent with a very low value of the uniqueness in APUF reported in Table 4. One reason for this may be the difficulty in ensuring routing symmetry in an APUF on an FPGA [17]. The sharp difference in the value of the uniqueness is visible in this case also. The reliability values are comparable for both the datasets indicating both RO PUF and APUF are equally reliable.

Table 6. Comparison of RO PUF and APUF using parameters defined by Maiti et al.

	APUF	**RO PUF**	**Ideal Value**
Uniformity	55.69%	50.56%	50%
Bit-aliasing	19.57%	50.56%	50%
Uniqueness	7.20%	47.24%	50%
Reliability	99.76%	99.14%	100%

Table 7. Comparison of Probability of Misidentification

	Minimum	Maximum	Average
RO PUF	2.81×10^{-71}	4.42×10^{-39}	1.18×10^{-40}
APUF	3.03×10^{-13}	3.91×10^{-12}	1.50×10^{-12}

4.3 Comparison using the Probability of Misidentification:

Table 7 shows the value of probability of misidentification for the two datasets. The RO PUF dataset shows a much lower value compared to the APUF dataset. However, this is due to the fact that the length of the identifier for the RO PUF is 511 whereas it is 128 for the APUF. In any case, even the Arbiter PUF shows a very low probability of misidentification. This is consistent with the fact that both the PUF showed a very high value of reliability indicating that the proportion of the noisy bits in both the PUFs are low.

4.4 Summary of Comparison between the RO PUF and the APUF

Table 8 shows the summary of the comparison between the APUF and the RO PUF in terms of the seven parameters we selected as part of the proposed evaluation-comparison method. The two PUFs exhibit comparable performance in terms of the uniformity, the reliability and the steadiness. However, the RO PUF shows much better performance compared to the APUF in terms of the uniqueness, the bit-aliasing and the PMSID. The diffuseness parameter could not be evaluated for the RO PUF.

5 Conclusions

In this work, we aimed at defining a method to evaluate as well as compare the performance of several PUFs irrespective of the underlying PUF techniques. Our

Table 8. Summary of Comparison between the RO PUF and the APUF

	APUF	**RO PUF**	**Ideal Value**
Uniformity	55.69%	50.56%	50%
Reliability	99.76%	99.14%	100%
Steadiness	98.48%	98.51%	100%
Uniqueness	36.75%	94.07%	100%
Diffuseness	98.39%	-	100%
Bit-aliasing	19.57%	50.56%	50%
PMSID	1.50×10^{-12}	1.18×10^{-40}	0%

approach relies on the statistical properties of the binary response bits of a PUF. We first proposed three dimensions of PUF measurement. Based on our analysis on parameters defined by us as well as by others, we proposed a set of seven PUF parameters as the primary building block of the evaluation-comparison method. Subsequently, we compared two different PUFs, namely the RO PUF and the APUF using these parameters based on measured PUF data. The RO PUF shows better performance than the APUF in terms of the uniqueness, the bit-aliasing and the probability of misidentification while other parameters yielded comparable results from both the PUFs.

As part of the future work, we plan to improve this method by studying other parameters that have not been covered in this work. We also plan to include implementation-related aspects of PUFs such as area, performance and power as part of the evaluation criteria. For easy usage of the proposed method, we plan to build a web application available to evaluate the performance of PUFs online (For more information, visit http://rijndael.ece.vt.edu/variability).

References

1. F. Armknecht, R. Maes, A.-R. Sadeghi, F.-X. Standaert, and C. Wachsmann. A Formal Foundation for the Security Features of Physical Functions. *IEEE Security and Privacy 2011*, 2011(1):16, 2011.
2. L. Bolotnyy and G. Robins. Physically unclonable function-based security and privacy in rfid systems. In *Pervasive Computing and Communications, 2007. PerCom '07. Fifth Annual IEEE International Conference on*, pages 211 –220, march 2007.
3. S. Devadas, E. Suh, S. Paral, R. Sowell, T. Ziola, and V. Khandelwal. Design and implementation of puf-based "unclonable" rfid ics for anti-counterfeiting and security applications. In *RFID, 2008 IEEE International Conference on*, pages 58–64, april 2008.
4. B. Gassend, D. Clarke, M. van Dijk, and S. Devadas. Silicon physical random functions. In *Proceedings of the 9th ACM conference on Computer and communications security*, CCS '02, pages 148–160, New York, NY, USA, 2002. ACM.
5. J. Guajardo, S. Kumar, G.-J. Schrijen, and P. Tuyls. Brand and ip protection with physical unclonable functions. In *Circuits and Systems, 2008. ISCAS 2008. IEEE International Symposium on*, pages 3186 –3189, may 2008.
6. J. Guajardo, S. S. Kumar, G.-J. Schrijen, and P. Tuyls. Fpga intrinsic pufs and their use for ip protection. In *Proceedings of the 9th international workshop on Cryptographic Hardware and Embedded Systems*, CHES '07, pages 63–80, Berlin, Heidelberg, 2007. Springer-Verlag.
7. R. Helinski, D. Acharyya, and J. Plusquellic. A physical unclonable function defined using power distribution system equivalent resistance variations. In *Proceedings of the 46th Annual Design Automation Conference*, DAC, pages 676–681, New York, NY, USA, 2009. ACM.
8. Y. Hori, T. Yoshida, T. Katashita, and A. Satoh. Quantitative and statistical performance evaluation of arbiter physical unclonable functions on fpgas. In *Reconfigurable Computing and FPGAs (ReConFig), 2010 International Conference on*, pages 298 –303, dec. 2010.

9. A. R. Krishna, S. Narasimhan, X. Wang, and X. Wang. Mecca: a robust low-overhead puf using embedded memory array. In *Proceedings of the 13th international conference on Cryptographic hardware and embedded systems*, CHES'11, pages 407–420, Berlin, Heidelberg, 2011. Springer-Verlag.
10. S. Kumar, J. Guajardo, R. Maes, G.-J. Schrijen, and P. Tuyls. Extended abstract: The butterfly puf protecting ip on every fpga. In *Hardware-Oriented Security and Trust, 2008. HOST 2008. IEEE International Workshop on*, pages 67 –70, 2008.
11. D. Lim, J. Lee, B. Gassend, G. Suh, M. van Dijk, and S. Devadas. Extracting secret keys from integrated circuits. *Very Large Scale Integration Systems, IEEE Transactions on*, 13(10):1200 – 1205, 2005.
12. K. Lofstrom, W. Daasch, and D. Taylor. Ic identification circuit using device mismatch. In *Solid-State Circuits Conference, 2000. Digest of Technical Papers. ISSCC. 2000 IEEE International*, pages 372 –373, 2000.
13. R. Maes, P. Tuyls, and I. Verbauwhede. Intrinsic pufs from flip-flops on reconfigurable devices. In *3rd Benelux Workshop on Information and System Security (WISSec 2008)*, page 17, Eindhoven,NL, 2008.
14. R. Maes and I. Verbauwhede. Physically unclonable functions: A study on the state of the art andfuture research directions. In *Towards Hardware-Intrinsic Security*. Springer, 2010.
15. A. Maiti, J. Casarona, L. McHale, and P. Schaumont. A large scale characterization of ro-puf. In *Hardware-Oriented Security and Trust (HOST), 2010 IEEE International Symposium on*, pages 94 –99, 2010.
16. M. Majzoobi, F. Koushanfar, and M. Potkonjak. Testing techniques for hardware security. In *Test Conference, 2008. ITC 2008. IEEE International*, pages 1 –10, 2008.
17. S. Morozov, A. Maiti, and P. Schaumont. An analysis of delay based puf implementations on fpga. In P. Sirisuk, F. Morgan, T. El-Ghazawi, and H. Amano, editors, *Reconfigurable Computing: Architectures, Tools and Applications*, volume 5992 of *Lecture Notes in Computer Science*, pages 382–387. Springer Berlin / Heidelberg, 2010.
18. R. S. Pappu, B. Recht, J. Taylor, and N. Gershenfeld. Physical one-way functions. *Science*, 297:2026–2030, 2002.
19. Y. Su, J. Holleman, and B. Otis. A digital 1.6 pj/bit chip identification circuit using process variations. *Solid-State Circuits, IEEE Journal of*, 43(1):69 –77, 2008.
20. G. E. Suh and S. Devadas. Physical unclonable functions for device authentication and secret key generation. In *Proceedings of the 44th annual Design Automation Conference*, DAC '07, pages 9–14, New York, NY, USA, 2007. ACM.
21. D. Suzuki and K. Shimizu. The glitch puf: a new delay-puf architecture exploiting glitch shapes. In *Proceedings of the 12th international conference on Cryptographic hardware and embedded systems*, CHES'10, pages 366–382, Berlin, Heidelberg, 2010. Springer-Verlag.
22. P. Tuyls, G.-J. Schrijen, B. Škorić, J. van Geloven, N. Verhaegh, and R. Wolters. Read-proof hardware from protective coatings. In *Cryptographic Hardware and Embedded Systems Workshop*, volume 4249 of *LNCS*, pages 369–383. Springer, October 2006.
23. V. van der Leest, G.-J. Schrijen, H. Handschuh, and P. Tuyls. Hardware intrinsic security from d flip-flops. In *Proceedings of the fifth ACM workshop on Scalable trusted computing*, STC '10, pages 53–62, New York, NY, USA, 2010. ACM.
24. D. Yamamoto, K. Sakiyama, M. Iwamoto, K. Ohta, T. Ochiai, M. Takenaka, and K. Itoh. Variety enhancement of puf responses based on the locations of

random outputting rs latches. http://www.iacr.org/workshops/ches/ches2011/presentations/Session8/CHES2011_Session8_3.pdf.
25. D. Yamamoto, K. Sakiyama, M. Iwamoto, K. Ohta, T. Ochiai, M. Takenaka, and K. Itoh. Uniqueness enhancement of puf responses based on the locations of random outputting rs latches. In *Proceedings of the 13th international conference on Cryptographic hardware and embedded systems*, CHES'11, pages 390–406, Berlin, Heidelberg, 2011. Springer-Verlag.

Acknowledgment

This work was supported by the National Science Foundation by grant no. 0964680 and grant no. 0855095.