Quantitative and Statistical Performance Evaluation of Arbiter Physical Unclonable Functions on FPGAs

Yohei Hori*, Takahiro Yoshida†, Toshihiro Katashita* and Akashi Satoh*
Research Center for Information Security (RCIS)
National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan
Email: {hori.y, t-katashita, akashi.satoh}@aist.go.jp
†*Research and Development Initiative, Chuo University, Tokyo, Japan*
Email: t-yoshida@imailab.jp

***Abstract*—The quantitative performance indicators of Physical Unclonable Functions (PUFs)—Randomness, Steadiness, Correctness, Diffuseness and Uniqueness—are strictly defined and applied to the evaluation of 45 arbiter PUFs on Virtex-5 FPGAs. The indicators effectively reflect the characteristics of PUFs ranging from 0 to 1 with 1 being the highest performance. The indicators enable the easy measurement and intuitive understanding of PUF performances. The experimental results shows that the arbiter PUFs have excellent overall intra-device performances though a slight bit bias is indicated. The inter-device performance is moderate and will suffice for the practical use of PUFs for device authentication and so on. Additionally, the reliability of the obtained PUF performances is statistically discussed in terms of the Confidence Interval and the number of devices. This paper presents in detail the definitions of the performance indicators and the quantitative and statistical evaluation results of the arbiter PUFs.**

***Keywords*-Physical Unclonable Function (PUF); Arbiter PUF; Virtex-5; FPGA; performance evaluation;**

I. Introduction

A Physical Unclonable Function (PUF) [1] is a function that outputs a device-specific response by extracting its intrinsic physical characteristics such as particulate diffusion and microscopic variation of silicon devices. The physical characteristics of the devices are practically unclonable, and therefore a PUF is expected to output unique responses used for device authentication and secret key generation.

A Silicon PUF is a circuit constructed on semiconductor and outputs a unique ID utilizing device variation of gate length, threshold voltage and so on. The examples of Silicon PUFs are the Arbiter PUF [2], Ring-oscillator (RO) PUF [3], SRAM PUF [4], Butterfly PUF [5] and Tri-state PUF [6].

Silicon PUFs are significantly important for the security of reconfigurable devices such as FPGAs. The configuration data of FPGA are stored in an electric file and downloaded to the device when it is configured/reconfigured. As is the similar problem to the software piracy and viruses, in some cases the configuration data must be protected against illegal eavesdropping and tampering. Although some of the high-end FPGAs have an AES/TDES core for bitstream decryption and HMAC core for bitstream authentication, lots of low-cost FPGAs do not have cryptographic engine for the bitstream security. Additionally, even in these high-end devices, the cryptographic cores are deliberately disabled when the dynamic partial reconfiguration feature is used. Furthermore, in some devices the secret key is stored in volatile memory with a battery, which could be unsuitable for the long-term usage and also be space-consuming. Other common way to store the secret key is to use non-volatile memory which, however, could arouse apprehension for the information leakage by, for example, physical attack [7] and side-channel attack [8], [9].

Considering the above situation, PUFs can be the solution to the security issues in reconfigurable devices. As the past studies reported, a PUF is useful for device authentication and IP core protection [2], [4], [10]. Even when the configuration data are eavesdropped by the attacker on the network, the attacker cannot obtain the secret information from his/her own device since the PUF generates a device-specific key.

However, there seems little consensus about which PUF is more suitable for the particular purpose or particular device. To enable the objective comparison, the *performance* of a PUF must be first defined. So far, the information-theoretical evaluation method of Optical PUFs [11] and Coating PUFs [12], and the empirical evaluation with 125 RO PUFs [13] have been reported.

In this paper, we evaluate Arbiter PUFs on 45 Virtex-5 FPGAs [14] using strictly defined performance indicators: Randomness, Steadiness, Correctness, Diffuseness and Uniqueness. All the indicators range from 0 to 1 with 1 being the highest performance. With these indicators we can easily and intuitively perceive the performance of PUFs. The rigorous definitions of the indicators and their application to the evaluation of 45 Arbiter PUFs are explained in the following sections.

II. Arbiter Physical Unclonable Function

An Arbiter PUF extracts the device-specific variation as delay differential between two selector chains. The example structure of the Arbiter PUF is illustrated in Figure 1. The stimuli are simultaneously input to the upper and lower

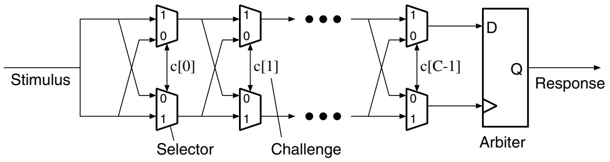


Figure 1. A structure of the arbiter PUF.

Table I
THE NOTATION USED IN THE ARTICLE.

Notation	Explanation
N	The number of devices.
K	The number of different IDs generated per device.
T	The number of tests performed per ID.
L	The length of an ID.
n	The index of a device. The n-th device is denoted by n for simplicity if not confusing. $1 \leq n \leq N$.
k	The index of an ID. The k-th ID is denoted by k for simplicity if not confusing. $1 \leq k \leq K$.
t	The index of a test. The t-th test is denoted by t for simplicity if not confusing. $1 \leq t \leq T$.
l	The bit position of an ID. The l-th bit is denoted by l for simplicity if not confusing. $1 \leq l \leq L$.
$\boldsymbol{ID}_{n,k}$	The correct ID k expected to be generated in device n.
$\boldsymbol{ID}_{n,k,t}$	The empirically generated ID k in test t in device n.
$b_{n,k,l}$	The correct response bit l of ID k expected to be generated in device n. $b_{n,k,l} \in \{0,1\}$.
$b_{n,k,t,l}$	The empirically generated response bit l of ID k in test t in device n. $b_{n,k,t,l} \in \{0,1\}$.

selector chains and the challenge is input as the selection signal of the selector. The stimulus will travel various patterns of paths according to the input challenge. The response is 1 if the stimulus to the D port is faster than that to the clock port, and conversely the response is 0 if the stimulus to the clock port is faster.

The delays of the two selected paths are significantly affected by the intrinsic device-specific variation, and therefore the Arbiter PUF is expected to output unique IDs to the device.

III. DEFINITIONS OF THE PUF PERFORMANCE

To enable the quantitative performance evaluation of PUFs, we propose five performance indicators: Randomness, Steadiness, Correctness, Diffuseness, and Uniqueness. The first four indicators are the *intra-device* performance and Uniqueness is the *inter-device* performance. The procedure for obtaining these indicators is illustrated in Figure 2. To analyze the performance of PUFs, we need to know the *correct IDs* which are expected to be steadily generated in the device. This section first describes the procedure for determining correct IDs, and then gives the definitions of the performance indicators in detail. The notation used in this paper is listed in Table I.

A. Correct IDs

Suppose a PUF in device n outputs K different IDs from K different challenge sets. Let $\boldsymbol{ID}_{n,k}$ be the correct ID k in device n. $\boldsymbol{ID}_{n,k}$ is empirically determined through T tests where the same ID k is tried to be generated T times.

Let $\boldsymbol{ID}_{n,k,t}$ be the ID k generated in test t in device n, and $b_{n,k,t,l} \in \{0,1\}$ be the l-th bit value of $\boldsymbol{ID}_{n,k,t}$. Then, $\boldsymbol{ID}_{n,k,t}$ is expressed as follows:

$$\boldsymbol{ID}_{n,k,t} = \{0,1\}^L = b_{n,k,t,1}||b_{n,k,t,2}||\ldots||b_{n,k,t,L} \qquad (1)$$

where the operator $||$ denotes concatenation of the operands.

Let $p_{n,k,l}$ be the relative frequency of the l-th bit of $\boldsymbol{ID}_{n,k,t}$ being 1. $p_{n,k,l}$ is defined as the average number of 1's in $\boldsymbol{ID}_{n,k,t}$ generated through T tests:

$$p_{n,k,l} = \frac{1}{T}\sum_{t=1}^{T} b_{n,k,t,l}. \qquad (2)$$

Let $b_{n,k,l}$ be the l-th bit of the correct ID $\boldsymbol{ID}_{n,k}$. Then $b_{n,k,l}$ is defined as follows:

$$b_{n,k,l} = \lfloor(p_{n,k,l} + 0.5)\rfloor = \begin{cases} 1 & \text{if } \ p_{n,k,l} \geq 0.5, \\ 0 & \text{otherwise.} \end{cases} \qquad (3)$$

With the definition of $b_{n,k,l}$, the correct ID $\boldsymbol{ID}_{n,k}$ is defined as follows:

$$\boldsymbol{ID}_{n,k} = b_{n,k,1}||b_{n,k,2}||\ldots||b_{n,k,L}. \qquad (4)$$

B. Randomness

A PUF is expected to output 0 and 1 ideally in the same probability. *Randomness* indicates the balance of 0 and 1 in the responses of the PUF.

Let p_n be the relative frequency of 1 appearing in all the response bits generated in device n. Then p_n is given by

$$p_n = \frac{1}{K \cdot T \cdot L}\sum_{k=1}^{K}\sum_{t=1}^{T}\sum_{l=1}^{L} b_{n,k,t,l}. \qquad (5)$$

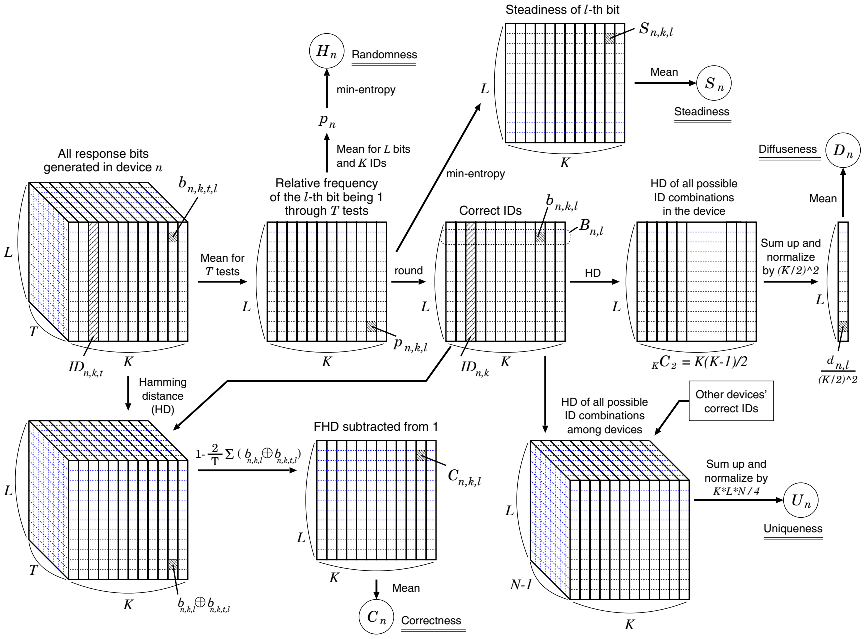
Then, the device Randomness H_n is defined as follows:

$$H_n = -\log_2 \max(p_n, 1 - p_n) \qquad (6)$$

with $\log_2(0) := 0$. Since outputs of a PUF are expected to be uniformly distributed, H_n is defined using min-entropy which is suitable for evaluating the randomness of a bit sequence. H_n takes the highest value 1 when $p_n = 0.5$, and the lowest value 0 when $p_n = 0$ or $p_n = 1$.

Note that H_n is the performance of a single device n. The apprehension is that the PUF could show good performance only in the particular device and not in other devices. Ideally, H_n should be calculated for all the devices where the PUF of the same structure will be implemented, nevertheless it is practically impossible. All the possible PUFs are considered as the *population*, and the devices under test are considered as the *sample* taken from the population. We estimate the population mean of Randomness from the sample of size N.

As dealing with the entropy is intractable, we first analyze p_n instead. The output of a device is independent of other devices and here PUFs are assumed to be implemented on



H_n
Randomness
min-entropy
p_n
Mean for L bits and K IDs
All response bits generated in device n
$b_{n,k,t,l}$
Relative frequency of the l-th bit being 1 through T tests
Mean for T tests
round
$p_{n,k,l}$
L
T
K
$ID_{n,k,t}$
Hamming distance (HD)
Steadiness of l-th bit
$S_{n,k,l}$
Mean
S_n
Steadiness
min-entropy
Correct IDs
$b_{n,k,l}$
$B_{n,l}$
$ID_{n,k}$
HD
HD of all possible ID combinations in the device
${}_KC_2 = K(K\text{-}1)/2$
Sum up and normalize by $(K/2)$^2
Diffuseness
D_n
Mean
$\frac{d_{n,l}}{(K/2)\text{^}2}$
FHD subtracted from 1
$1\text{-}\frac{2}{T}\Sigma\,(b_{n,k,l} \oplus b_{n,k,t,l})$
$C_{n,k,l}$
$b_{n,k,l} \oplus b_{n,k,t,l}$
Mean
C_n
Correctness
HD of all possible ID combinations among devices
Other devices' correct IDs
Sum up and normalize by $K*L*N/4$
U_n
Uniqueness
N-1

Figure 2. The procedure for obtaining PUF performances.

the same kind of devices, and thus the output $b_{n,k,t,l}$ is assumed to be independently and identically distributed (*iid*) random variables. Hence, by the central limit theorem, the probability distribution function (*pdf*) of p_n is assumed to be the normal distribution.

Since the population variance σ_p^2 is unknown, we guess the population mean μ_p by interval estimate with the sample mean $\overline{p}$ and unbiased sample variance $\hat{s}_p^2$. p and $\hat{s}_p^2$ are defined as follows:

$$\overline{p} = \frac{1}{N}\sum_{n=1}^{N} p_n \tag{7}$$

$$\hat{s}_p^2 = \frac{1}{N-1}\sum_{n=1}^{N}(p_n - \overline{p})^2. \tag{8}$$

The the normalized value t given by

$$t = \frac{p_N - \overline{p}}{\hat{s}_p/\sqrt{N}} \tag{9}$$

follows the t-distribution with the degree of freedom $N-1$. Then, the two-sided 95% confidence interval (*CI*) for μ_p is

$$[\overline{p} - t_{0.025;(N-1)} \times \frac{\hat{s}_p}{\sqrt{N}},\ \overline{p} + t_{0.025;(N-1)} \times \frac{\hat{s}_p}{\sqrt{N}}], \tag{10}$$

where $t_{0.025;(N-1)}$ denotes the 2.5 percentile in the t-distribution with the degree of freedom $N-1$.

After the CI for μ_p is determined, the CI for μ_H is defined using Equation (6). Since the higher indicator shows the better performance, the lower bound of the CI guarantees the worst-case performance with the probability 95%. Therefore, the narrower CI under the sufficiently high confidence coefficient is desirable.

C. Steadiness

When generating the same ID T times in the same device, the l-th bits of the output IDs are expected to be identical. *Steadiness* indicates how stably a PUF outputs the same responses to the same challenge sets, in other words, how strongly $p_{n,k,l}$ is biased toward 0 or 1.

Let $S_{n,k,l}$ be the Steadiness of the l-th bit of $\boldsymbol{ID}_{n,k}$. Then, $S_{n,k,l}$ is defined using min-entropy as follows:

$$S_{n,k,l} = 1 + \log_2 \max(p_{n,k,l}, 1 - p_{n,k,l}) \tag{11}$$

with $\log_2(0) := 0$. Then, the device Steadiness S_n is defined by taking the mean of $S_{n,k,l}$ as follows:

$$\begin{aligned} S_n &= \frac{1}{K \cdot L}\sum_{k=1}^{K}\sum_{l=1}^{L} S_{n,k,l} \\ &= 1 + \frac{1}{K \cdot L}\sum_{k=1}^{K}\sum_{l=1}^{L}\log_2 \max(p_{n,k,l}, 1 - p_{n,k,l}). \end{aligned} \tag{12}$$

S_n takes the highest value 1 when $p_{n,k,l} = 0$ or $p_{n,k,l} = 1$, and the lowest value 0 when $p_{n,k,l} = 0.5$. By the similar discussion to Randomness, the pdf of S_n is assumed to be normal distribution and the population mean μ_S is estimated from the sample mean $\overline{S}$ and unbiased sample variance $\hat{s}_S^2$ with the 95% CI.

D. Correctness

Correctness is the degree of accuracy of the PUF outputs. Suppose that a part of the device is broken for some reason after the correct ID is determined. Then, a PUF on the device could always output a wrong bit value for a particular challenge. In this case, the bit is *stable* but *incorrect*. Correctness would be useful to find such device defects or degradation by aging.

Correctness is defined using the *fractional Hamming Distance* (*FHD*). FHD is the average Hamming distance (*HD*) among IDs, that is, the sum of Hamming distance (*SHD*) normalized by T, K, L and so on. Let $c_{n,k,l}$ be the SHD between the correct bit $b_{n,k,l}$ and the generated bit $b_{n,k,t,l}$ through T tests:

$$c_{n,k,l} = \sum_{t=1}^{T} b_{n,k,l} \oplus b_{n,k,t,l}. \tag{13}$$

Since $b_{n,k,l}$ is determined by majority voting as shown in Equation (3), the following inequations hold for any n, k, and l in the defined domain:

$$0 \leq \sum_{t=1}^{T} (b_{n,k,l} \oplus b_{n,k,t,l}) \leq \frac{T}{2}. \tag{14}$$

Therefore, $c_{n,k,l}$ should be normalized by the possible maximum value $T/2$. Let $C_{n,k,l}$ be Correctness of the l-th bit of ID k in device n. Correctness $C_{n,k,l}$ is defined by the FHD subtracted from 1:

$$C_{n,k,l} = 1 - \frac{\sum_{t=1}^{T} (b_{n,k,l} \oplus b_{n,k,t,l})}{T/2}. \tag{15}$$

Then, the device Correctness C_n is given by taking the mean of $C_{n,k,l}$ for L and K:

$$\begin{aligned} C_n &= \frac{1}{K \cdot L} \sum_{k=1}^{K} \sum_{l=1}^{L} C_{n,k,l} \\ &= 1 - \frac{2}{K \cdot T \cdot L} \sum_{k=1}^{K} \sum_{t=1}^{T} \sum_{l=1}^{L} (b_{n,k,l} \oplus b_{n,k,t,l}). \end{aligned} \tag{16}$$

C_n takes the highest value 1 if all the generated IDs are identical to the correct IDs, and the lowest value 0 if only a half of the generated bits are correct. By the similar discussion to Randomness, the population mean μ_C is estimated from the sample mean $\overline{C}$ and unbiased sample variance $\hat{s}_C^2$ with the 95% CI.

E. Diffuseness

A PUF is expected to output different responses from different challenge sets. *Diffuseness* is the degree of difference among the IDs generated from the different challenge sets in the same device. Diffuseness of a device is determined by calculating FHD for all the possible ID-combinations generated in the device.

Note that the upper bound of the SHD of all the possible ID-combinations is $L \cdot (K/2)^2$, not $L \cdot ({}_K\mathrm{C}_2)$. Before we prove it, we first prove the following lemma.

Lemma 1: Suppose K ($K \geq 2$) correct IDs are determined in device n and only the set of their l-th bits, $\boldsymbol{B}_{n,l}$, is currently of consideration. $\boldsymbol{B}_{n,l}$ is expressed as follows:

$$\boldsymbol{B}_{n,l} = b_{n,1,l} || b_{n,2,l} || \cdots || b_{n,K,l}. \tag{17}$$

Let $d_{n,l}$ be the SHD of the possible bit-combinations in $\boldsymbol{B}_{n,l}$:

$$d_{n,l} = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} (b_{n,i,l} \oplus b_{n,j,l}). \tag{18}$$

Then, for $d_{n,l}$ the following inequations hold:

$$0 \leq d_{n,l} \leq (\frac{K}{2})^2. \tag{19}$$

Proof: Let x be the number of 1's in $\boldsymbol{B}_{n,l}$ and $K - x$ be the number of 0's ($0 \leq x \leq K$). Then, the number of the bit-combinations that give HD=1 is $x(K - x)$. Therefore,

$$d_{n,l} = x(K - x). \tag{20}$$

Since $x \geq 0$ and $K - x \geq 0$, the first inequation of (19) clearly holds with the equality if and only if $x = 0, K$.

Next, as for the second inequation of (19),

$$\begin{aligned} (\frac{K}{2})^2 - d_{n,l} &= (\frac{K}{2})^2 - x(K - x) \\ &= \frac{1}{4}(K - 2x)^2 \geq 0. \end{aligned} \tag{21}$$

Therefore, the second inequation of (19) holds with the equality if and only if $x = K/2$. ■

Since each bit of L-bit IDs is independent of other bits, the upper bound of the SHD of the possible ID-combinations is $L \cdot (K/2)^2$. Considering the upper bound of the SHD, the device Diffuseness D_n is defined as follows:

$$\begin{aligned} D_n &= \frac{1}{L} \sum_{l=1}^{L} \frac{d_{n,l}}{(K/2)^2} \\ &= \frac{4}{L \cdot K^2} \sum_{l=1}^{L} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} (b_{n,i,l} \oplus b_{n,j,l}). \end{aligned} \tag{22}$$

D_n takes the highest value 1 if the SHD of the possible ID-combinations reaches the upper bound, and the lowest value 0 if all the generated IDs are identical. By the similar discussion to Randomness, the population mean μ_D is estimated with the sample mean $\overline{D}$ and unbiased sample variance $\hat{s}_D^2$ with the 95% CI.

F. Uniqueness

When the same challenge sets are given to different PUFs, the output IDs are expected to be different. *Uniqueness* indicates how different the generated IDs are among the devices. Since Uniqueness is the inter-device performance, all the possible device-combinations should be considered. By the similar discussion to lemma 1, the SHD of the possible ID-combinations among all devices does not exceed $L \cdot (N/2)^2$. Considering the upper bound of SHD, the device Uniqueness U_n is defined as follows:

$$U_n = \frac{4}{K \cdot L \cdot N} \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{j=1, j \neq n}^{N} (b_{n,k,l} \oplus b_{j,k,l}). \tag{23}$$

Table II
THE EXPERIMENTAL RESULTS OF THE EVALUATION OF 45 FPGAS.

Performance		Mean	$\hat{s}$	95% CI	Width
Randomness	H	.8469	.02670	[.8388, .8546]	.01586
(Probability	p)	(.5561)	(.01017)	([.5530, .5591])	(.006111)
Steadiness	S	.9848	.07401	[.9626, 1.000]	.04134
Correctness	C	.9829	.08293	[.9579, 1.000]	.04206
Diffuseness	D	.9839	.01021	[.9810, .9870]	.006134
Uniqueness	U	.3675	.5150	[.2127, .5222]	.3095

Then, the sample mean $\overline{U}$ is defined as follows:

$$\begin{aligned}\overline{U} &= \frac{1}{N}\sum_{n=1}^{N} U_n \\ &= \frac{4}{K \cdot L \cdot N^2}\sum_{k=1}^{K}\sum_{l=1}^{L}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(b_{i,k,l} \oplus b_{j,k,l}). \qquad (24)\end{aligned}$$

$\overline{U}$ takes the highest value 1 when the SHD of the possible ID-combinations reaches the upper bound, and the lowest value 0 when the IDs in device n are completely identical to the others. With the above definitions, the similar discussion to the intra-device indicators is also applied to Uniqueness. The pdf of U_n is assumed to be the normal distribution, and the population mean μ_U is estimated from the sample mean $\overline{U}$ and unbiased sample variance $\hat{s}_U^2$ with the 95% CI.

IV. PUF PERFORMANCE EVALUATION

A. Implementation

We implemented the Arbiter PUFs on 45 FPGAs using SASEBO-GII [15] evaluation boards. SASEBO-GII is equipped with a Xilinx Virtex-5 xc5vlx30-ffg324 and a Spartan-3A [16] xc3s400a-ftg256, and the core voltage of the Virtex-5 can be directly adjusted by the variable resistor on the board. The PUFs are implemented on Virtex-5 and their core voltage are set to 1.050 V. The number of selector stages is set to 64 and therefore the challenge length is 64 bits. The challenge sets are generated using the C# `Random` class. The development tools used are Xilinx PlanAhead v10.1.8 for manual placement and ISE 10.1.03 for compilation. One of the two selector chains is placed from SLICE_X24Y1 to SLICE_X24Y64 and connected to the D input of the arbiter D-FF; the other chain is placed from SLICE_X25Y1 to SLICE_X25Y64 and connected to the clock input of the arbiter D-FF.

B. Experimentation

In the experimentation, 1024 kinds of 128-bit IDs are generated 1024 times per ID. Therefore, the test parameters are $N = 45$, $K = 1024$, $T = 1024$ and $L = 128$. The experimental results of the mean, standard deviation, and 95% CI of the PUF performances are shown in Table II. The CI is the estimated population mean with the confidence coefficient 95% with $t_{0.025;44} = 2.0154$. The results shown here are discussed in detail in the following subsections.

C. Intra-device Evaluation

As Table II shows, Randomness, Steadiness, Correctness, and Diffuseness of the PUFs are of good properties with the mean values close to 1 and small deviations. Additionally from the statistical analysis with 45 FPGAs, the CI for these indicators are quite narrow. The CI guarantees with probability 95% that the performance of the same structure of PUFs on the same kind of FPGAs will be at least the lower confidence bound. All of the lower confidence bounds for the intra-device indicators are satisfactorily large, and eventually, the PUFs are considered to have high intra-device performances.

Randomness of the PUFs are 0.8469, slightly lower than other intra-indicators are. This Randomness corresponds to the probability 55.61%. Therefore, the number of 1's in the responses is slightly larger than 0. This would be because the delays of the two selector chains are not very close, nevertheless equalizing the two delays is almost impracticable since the architecture of the FPGA is fixed. Despite the slight bit bias toward 1, Diffuseness is definitely high (0.9839), and therefore we can safely judge that the PUF will have high intra-device performances.

The bit bias could be improved by adjusting the location of the selectors in the FPGA. To study a method to construct a well-balanced arbiter PUF would be future work.

D. Inter-device Evaluation

The mean of the device Uniqueness is 0.3675 with relatively large standard deviation 0.5150. This Uniqueness corresponds to the average HD$\simeq$12 per 128-bit ID. This value seems to be small, nonetheless the possible number of IDs whose HD is 12 from a particular ID is ${}_{128}\mathrm{C}_{12} = 2.3726 \times 10^{16} > 2^{54}$. Therefore, the size of the possible response space with Uniqueness 0.3675 is considered extremely large. This size of response space would suffice for the practical use of PUFs for the device authentication where several IDs are queried. Even with the lower confidence bound 0.2127, the corresponding HD is about 7 and the response space is still quite large: ${}_{128}\mathrm{C}_{7} = 9.4526 \times 10^{10} > 2^{36}$.

Actually, the total number of the unique correct IDs determined through our experiment is 45,198, while its upper bound is 46,080 (= 128 devices×1024 IDs). This means that more than 98% of the correct IDs are unique among the devices. When a PUF is used for generating 128-bit IDs, the two IDs are equally judged *different* even if their HD is 1 or 128. Thus, the number of unique IDs is not necessarily reflects the *degree* of difference. There could be a more suitable indicator of the device uniqueness, and exploring such indicators is left as future work.

E. Discussion

Maiti et al. emphasized the importance of the large sample size by the empirical window method with the window size 16 [13]. The sample means of RO frequency of 16 devices

Table III
THE REQUIRED SAMPLE SIZE AND CORRESPONDING CI WIDTH.

Target CI width	$\hat{s}$	$0.8\,\hat{s}$	$0.6\,\hat{s}$	$0.4\,\hat{s}$	$0.2\,\hat{s}$
Required sample size	17	26	45	101	402
CI width of μ_H	.0107	.008136	.006102	.004068	.002034
μ_S	.07401	.05921	.04440	.02960	.01480
μ_C	.08293	.06634	.04976	.03317	.01659
μ_D	.01021	.008166	.006125	.004083	.002042
μ_U	.5150	.4120	.3090	.2060	.1030

are noticeably deviated from that of 125 devices. In our experimentation, it is statistically explained by Equation (10). The narrower the CI width is, the more likely the lower bound of the performance will increase.

Suppose the unbiased sample variance $\hat{s}^2$ is given by the preliminary experiment with the sample of size N. Let δ be the target CI width. Then, the required sample size N^* is the least integer N satisfying the following inequation [17]:

$$2 \times t_{0.025,(N-1)} \times \frac{c^*\hat{s}}{\sqrt{N^*}} \leq \delta, \qquad (25)$$

$$\text{where} \quad c^* = \frac{\sqrt{2}\,\varGamma(N/2)}{\sqrt{N-1}\,\varGamma((N-1)/2)}. \qquad (26)$$

Table III shows the required sample size N^* determined by the preliminary experiment with N=45 and the target CI width of each performances. The sample size N^*=125 and N^*=16 in [13] correspond to the CI width about δ = $0.35\,\hat{s}$ and $0.99\,\hat{s}$, respectively. The sample size N^* = 45 in our experimentation corresponds to δ = $0.60\,\hat{s}$. As Table III shows, the CI width of the performances are narrowed as the sample size increases, and therefore, a large sample size is important for the reliable performance evaluation. However, while the CI width of Uniqueness is noticeably narrowed, those of the other performances are not drastically shrunk. Thus, a relatively small sample size will suffice for the evaluation of intra-device performances.

To sum up, when the performance indicators are obtained from the PUFs on some devices, the CI of the performances are estimated by statistical analysis of the indicators. We can use the lower bound of CI as the threshold to judge whether the PUF has adequate properties. A sample of large size is preferable since the worst-case estimation will be improved. If the lower confidence bound of the performance does not reach the threshold, we judge that the PUF on the device does not have the required properties. To develop the rigorous evaluation criteria for PUF performances is left as future work.

V. CONCLUSIONS

We defined the five indicators—Randomness, Steadiness, Correctness, Diffuseness and Uniqueness—to quantitatively evaluate the performances of a PUF. These indicators are applied to the performance evaluation of 45 arbiter PUFs on Virtex-5 FPGAs. The experimental results show that the proposed indicators effectively reflect the characteristics of PUFs. The future work would be to develop more effective performance indicators especially for Uniqueness, and to develop the rigorous evaluation criteria to judge if the PUF is feasible for the security purposes such as device authentication and IP core protection.

ACKNOWLEDGMENT

This work is doen as a part of the project of Core Research for Evaluational Science and Technology (CREST) funded by Japan Science and Technology Agency (JST).

REFERENCES

[1] S.R. Pappu, Physical One-Way Functions, Ph.D thesis, MIT, 2001.

[2] D. Lim, J.W. Lee, B. Gassend, G.E. Suh, M. van Dijk, and S. Devadas, “Extracting secret keys from integrated circuits,” IEEE Trans. VLSI Syst., vol.13, no.10, pp.1200–1205, 2005.

[3] G.E. Suh, and S. Devadas, “Physical physical unclonable functions for device authentication and secret key generation,” DAC’07, pp.9–14, 2007.

[4] J. Guajardo, S.S. Kumar, G.J. Schrijen, and P. Tuyls, “FPGA intrinsic PUFs and their use for IP protection,” CHES’07, pp.63–80, 2007.

[5] S.S. Kumar, J. Guajardo, R. Maesyz, G.J. Schrijen, and P. Tuyls, “The butterfly PUF,” HOST’08, pp.67–70, 2008.

[6] E. Ozturk, G. Hammouri, and B. Sunar, “Physical unclonable function with tristate buffers,” ISCAS’08, pp.3194–3197, 2008.

[7] R. Anderson, M. Bond, J. Clulow, and S. Skorobogatov, “Cryptographic processors—a survey,” Proceedings of the IEEE, vol.94, no.2, pp.357–369, 2006.

[8] P.C. Kocher, “Timing attacks on implementations of diffie-hellman, RSA, DSS, and other systems,” Adances in Cryptography — CRYPTO’96, pp.104–113, 1996.

[9] P. Kocher, J. Jaffe, and B. Jun, “Differential power analysis,” CRYPTO’99, pp.388-397, 1999.

[10] C. Bosch, J. Guajardo, A.R. Sadegh, J. Shokrollahi, and P. Tuyls, “Efficient helper data key extractor on FPGAs,” CHES’08, pp.181–197, 2008.

[11] P. Tuyls, B. Škorić, S. Stallinga, A. Akkermans, and W. Ophey, “Information-theoretic security analysis of physical uncloneable functions,” FC’05, pp.141–155, 2005.

[12] B. Škorić, S. Maubach, T. Kevenaar, and P. Tuyls, “Information-theoretic analysis of capacitive physical unclonable functions,” J. Appl. Phys., vol.200, no.024902, 2006.

[13] A. Maiti, J. Casarona, L. McHale, and P. Schaumont, “A large scale characterization of RO-PUF,” HOST’10, pp.66–71, 2010.

[14] Xilinx, Inc., Virtex-5 Family Overview, 2009.

[15] “Side-channel attack standard evaluation board (sasebo),” http://www.rcis.aist.go.jp/special/SASEBO/Research Center for Information Security, National Institute of Advanced Industrial Science and Technology.

[16] Xilinx, Inc., Extended Spartan-3A FPGA Familiy Overview, 2010.

[17] Y. Nagata, Sample Size Determination, Asakura Shoten, 2003, in Japanese.