INTERNATIONAL STANDARD

ISO/IEC 20897-2

First edition
2022-05

Information security, cybersecurity and privacy protection — Physically unclonable functions —

Part 2: Test and evaluation methods

Sécurité de l'information, cybersécurité et protection de la vie privée — Fonctions non clonables physiquement —

Partie 2: Méthodes d'essai et d'évaluation

Reference number
ISO/IEC 20897-2:2022(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents). Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 27, *Information security, cybersecurity and privacy protection*.

A list of all parts in the ISO/IEC 20897 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

This document specifies the test methods for physically unclonable functions (PUFs) for generating non-stored cryptographic parameters.

Cryptographic modules generate the certain class of critical security parameters such as a secret key using a random bit generator within the modules. Such modules may store generated security parameters in embedded non-volatile memory elements. For a higher security, a combination of tamper response and zeroisation techniques may be used for protecting stored security parameters from active unauthorized attempts of accessing such parameters. As the reverse-engineering technology advances, however, the risk of theft of such stored security parameters has become higher than ever.

The rapidly pervading technology called a PUFs is promising to mitigate the above-mentioned risks by enabling security parameter management without storing such parameters. PUFs are hardware-based functions providing mathematical unclonability, steadiness and randomness of their outputs and physical unclonability of the functions themselves, taking advantage of intrinsic subtle variations in the device's physical properties, which are also considered objects' fingerprints. PUFs may be used for security parameter (e.g. key, initialization vector, nonce and seeds) generation, entity authentication or device identification in cryptographic modules. More detailed information about the characteristics and security requirements of the PUF are given in ISO/IEC 20897-1 and this document only describes test and evaluation methods.

Now, security requirements of PUFs should be considered at system level, meaning that they should consider many possible attack paths, as detailed further in this document. The purpose of this document is to specify how to test those security requirements for assuring an adequate level of quality of the provided PUFs in cryptographic modules. This document is supposed to be used for the following purposes:

a) In the procurement process of a PUF-equipped product, the procurement body specifies the security requirements of the PUF in accordance with ISO/IEC 20897-1. The product vendor evaluates the PUF in accordance with this document whether the PUF satisfies all the specified security requirements, and reports the evaluation results to the procurement body.

b) The vendors evaluate the security of their PUF in accordance with this document, publicize the evaluation results and clarify the security of their PUF.

It should be noted that all of the security requirements defined in ISO/IEC 20897-1 are not necessarily quantitatively evaluable.

Information security, cybersecurity and privacy protection — Physically unclonable functions —

Part 2: Test and evaluation methods

1 Scope

This document specifies the test and evaluation methods for physically unclonable functions (PUFs). The test and evaluation methods consist of inspection of the design rationale of the PUF and comparison between statistical analyses of the responses from a batch of PUFs or a unique PUF versus specified thresholds.

This document is related to ISO/IEC 19790 which specifies security requirements for cryptographic modules. In those modules, critical security parameters (key) and public security parameters (product serial number, identification code, etc.) are the assets to protect. PUF is one solution to avoid storing security parameters, thereby increasing the overall security of a cryptographic module.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 19790, *Information technology — Security techniques — Security requirements for cryptographic modules*

ISO/IEC 20897-1, *Information security, cybersecurity and privacy protection — Physically unclonable functions — Part 1: Security requirements*

3 Terms, definitions and abbreviated terms

For the purposes of this document, terms, definitions and abbreviated terms given in ISO/IEC 20897-1, ISO/IEC 19790 and following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1 Abbreviated terms

BER Bit error rate.

iid Independent and identically distributed.

IID

NRBG Non-deterministic random bit generator

4 Symbols

For the purposes of this document, the following symbols apply.

$\forall$ A math symbol representing "for all" or "for any."

$\in$ A math symbol representing set membership.

$\boldsymbol{D}^{\text{inter}}$ A vector representing Inter-HD between PUF responses.

$\boldsymbol{D}^{\text{intra}}$ A vector representing Intra-HD between PUF responses.

i, j The index for the PUF instances. $1 \leq i, j \leq N_{\text{PUF}}$.

k The index for the challenge. $1 \leq k \leq N_{\text{chal}}$.

N The sequence size (bit length) of PUF responses.

N_{c} The largest number of identical responses (correct responses).

N_{chal} The number of different challenges given to a PUF.

N_{meas} The number of measurements of responses repeatedly collected for a single challenge.

N_{PUF} The number of PUF instances.

N_{res} The length of PUF response obtained from a single challenge.

σ A standard deviation of a random value.

Σ A sum of all values in the specified range.

t The index for the response measurements. $1 \leq t \leq \text{N}_{\text{meas}}$.

μ A mean of a random value.

x A challenge.

y A response.

$y_i^{(t)}(x)$ The t-th response of the i-th PUF instance obtained by giving a challenge x.

5 Tests of PUFs

5.1 General

In this document, testing a PUF means verifying the security requirements defined in ISO/IEC 20897-1. As already mentioned in ISO/IEC 20897-1, for the purpose of the ISO/IEC 20897 series, the responses from multiple PUFs are arranged into a cube as shown in Figure 1. The repetitive calls to a PUF are illustrated in Figure 2. The single small cube describes a 1-bit response from a PUF. The three axes of the cube and the time are described hereafter, as directions:

— direction B: "#bits" shows the bit length of the response obtained from a single challenge. In a 1-bit response PUF, e.g., arbiter PUF, the dimension B collapses.

— direction C: "#challenges" shows the number of different challenges given to a PUF. In a no-challenge PUF (or, more rigorously, a one-challenge PUF), e.g., SRAM PUF[1], the dimension C collapses.

— direction D: "#PUF" shows the number of different PUF devices under test.

— direction T: "#query" shows the number of query iterations under the fixed PUF device and challenge.

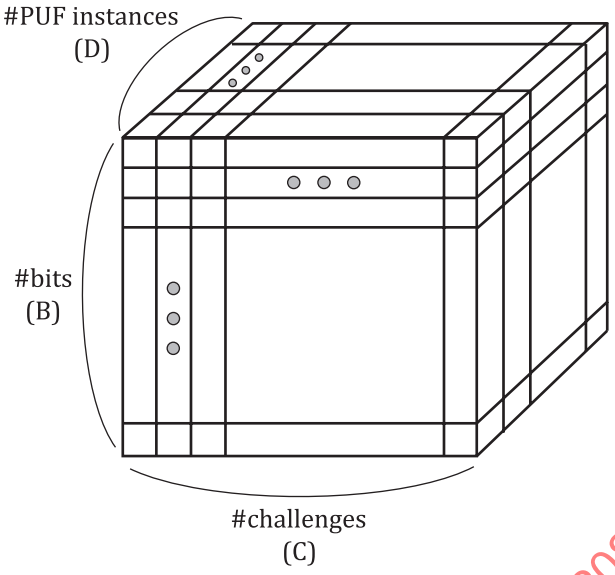

Figure 1 — Cube representation of the response sequences from multiple PUFs

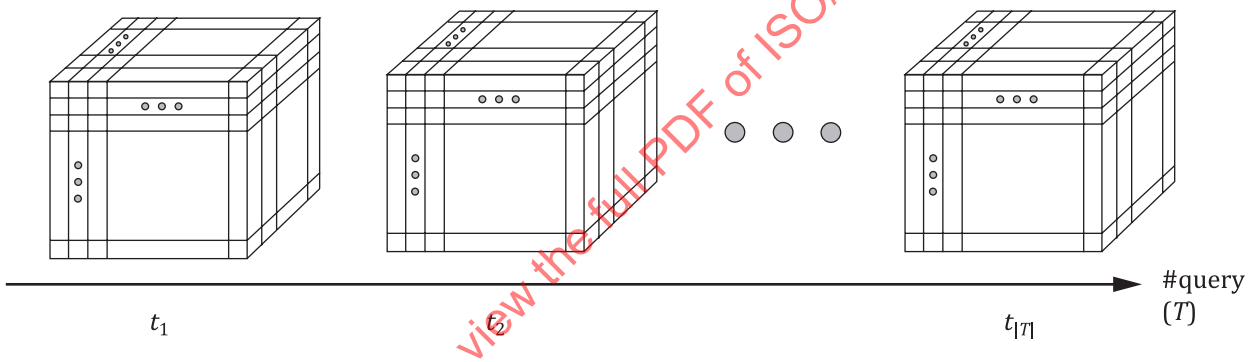

Figure 2 — Responses obtained by repetitive calls to the PUFs.

Among the defined security requirements, the steadiness, randomness and uniqueness may be tested by measuring responses from actual device(s) (Figures 1 and 2). If the empirical approaches are insufficient for evaluation, a stochastic model can be applied. If a stochastic model is used, the vendor shall provide the detailed document which explains the correctness of the model and the validity of the use of the model.

The tests based on stochastic models are defined in BSI AIS 31 standard and Reference [3]. They refer to an abstraction of the device where probabilistic aspects are clearly described. The model allows to derive ideal results, in terms of expectations over distributions (not only estimations based on limited sampling).

Notice that stochastic models shall be used in these two conditions:

— when a metric is not otherwise testable, owing to the prohibitive number of measurements which would be required,

— when a predictive (asymptotically) value of a metric is required.

For the stochastic model to be relevant, it shall rely on analogue random properties (such as delays, voltages, etc.). These measurements would typically be quantified to get the response bits. Some PUF structures may feature the capability to quantify the analogue properties, for example, when the response bits are obtained by a logical process. For instance, in the loop-PUF, the number of loops is counted. The response bit is subsequently computed based on a comparison between two (or more)

loop numbers. Thus, the information of the number of loops is available for analysis in a stochastic model.

Some PUF structures could add the capability of measurement quantification in a so-called "test mode." The quantification capability is thus intentionally added in an artificial manner for the sake of computing the metric. The addition of the quantification capability can be justified in situations where the PUF responses might be used only if the PUF itself is trusted.

5.2 Test conditions

The metrics shall be estimated according to several environmental conditions, including temperature, voltage, and humidity. Extreme values defined from the device standard operating conditions, e.g., the temperature and supply voltage specified in 7.7.4.3 of ISO/IEC 19790:2012, shall be tested. Also, some cycles between extreme values shall be performed, so as to check that the device is still functioning as intended. The accelerated aging techniques in Reference [5] may be used for this purpose.

5.3 Security tests

5.3.1 General

5.3 provides the concrete test and evaluation methods of a PUF. The security tests described in 5.3.2 through 5.3.7 corresponds to the security requirements defined in ISO/IEC 20897-1.

To claim that the PUF satisfies one or more of the security requirements, the vendor shall document the reason for that based on the conducted tests, evaluation results, and/or design rationale. The document may include the logic diagram of the PUF building blocks, e.g., entropy source, entropy extractor, pre-processing block, post-processing block, and so forth. For the example of security requirements, evaluation criteria, and test conditions, see Annex D.

5.3.2 Test of steadiness

Steadiness is the repeatability over T measurements in Figure 2, that is, the estimation of a probability of failure based on repeated measurements. For devices which require a very high steadiness (e.g., for mission critical applications), the experimental approach can be insufficient in terms of error bars (estimation is too crude). In such a case, a stochastic model may be applied.

A PUF shall meet the steadiness requirement if the application of the PUF requires it (cf. 5.6, ISO/IEC 20897-1:2020). If a vendor claims that the PUF satisfies the steadiness requirement, the vendor shall document the reason for that. The steadiness shall be quantified by at least one metric: the intra-HD, bit error rate (BER), stochastic model, and so forth. The chosen value of challenge bits shall be precisely described, but is arbitrary. A vendor shall be responsible for determining the criterion of the steadiness considering the applications and operating environment of the PUF. It is the vendor's responsibility, if necessary, to check with the designer, manufacturer, and so forth for detailed information of the PUF in order to establish the evaluation criteria for steadiness. The concept of the test of steadiness is illustrated in Figure 3. For the example test procedure for steadiness, see Annex A.

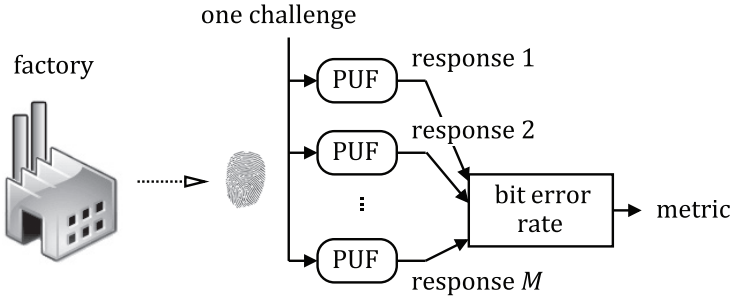

Figure 3 — Principle of the test for steadiness

5.3.3 Test of randomness

Randomness is the variability across B-C plane in Figure 1. A PUF shall meet the randomness requirement if the application of the PUF requires it (cf. 5.6, ISO/IEC 20897-1:2020). If a vendor claims that the PUF satisfies the randomness requirement, the vendor shall document the reason for that. A vendor shall be responsible for determining the criterion of the randomness considering the applications and operating environment of the PUF. It is the vendor's responsibility, if necessary, to check with the designer, manufacturer, and so forth for detailed information of the PUF in order to establish the evaluation criteria for randomness.

The randomness shall be tested by applying a statistical randomness test or entropy estimation to the B-C plain of the responses (see Figure 1). When the dimension C collapses (e.g., in confined PUFs), the test is applied to the bit sequence in dimension B. Similarly, when the dimension B collapses (e.g., in an arbiter PUF), the test is applied to the bit sequence in dimension C. Note that the randomness tests cannot apply to the cases without enough number of PUF responses. The tests may be the NIST FIPS 140-1[2], BSI AIS 31[3], SP800-22[6], SP800-90B[7] and so forth. The chosen value of challenge bits shall be precisely described, but is arbitrary. For the example test procedure for randomness, see Annex B.

NOTE FIPS 140-1 has been withdrawn, but that its entropy tests are still reliable methods which can be repurposed to evaluate a PUF.

5.3.4 Test of uniqueness

Uniqueness is the variability across B-D plane in Figure 1. A PUF shall meet the uniqueness requirement if the application of the PUF requires it (cf. 5.6, ISO/IEC 20897-1:2020). If a vendor claims that the PUF satisfies the uniqueness requirement, the vendor shall document the reason for that. A vendor shall be responsible for determining the criterion of the uniqueness considering the applications and operating environment of the PUF. It is the vendor's responsibility, if necessary, to check with the designer, manufacturer, and so forth for detailed information of the PUF in order to establish the evaluation criteria for uniqueness.

The uniqueness shall be assessed by evaluating the inter-Hamming distance of the responses among the PUF devices, or by the statistical ways similar to the randomness. In practice, it is not always possible to prepare a sufficient number of devices for the test of uniqueness. In such a case, a stochastic model may leverage the test (see C.4). If the uniqueness is evaluated by the statistical tests, the applicable test includes the NRBG health test in ISO/IEC 18031:2011[8] which is based on FIPS 140-1[2] and AIS-31[3], and NIST SP800-90B[7]. The concept of the test of uniqueness is illustrated in Figure 4. The chosen value of challenge bits shall be precisely described, but is arbitrary. For the example test procedure for uniqueness, see Annexes B and C.

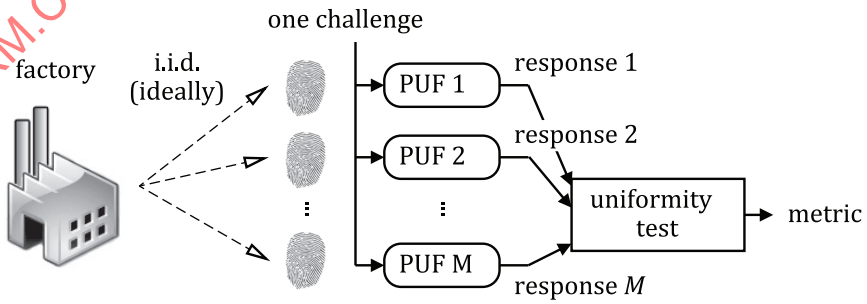

Figure 4 — Test of uniqueness

5.3.5 Test of Tamper-resistance

The test of the tamper-resistance verifies whether the PUF does not leak secret data nor lose requisite properties by invasive, semi-invasive and non-invasive physical attacks. The examples of the physical attacks include side-channel attacks, reverse engineering, using fault analysis tools such as LVP and FIB.

A PUF shall meet the tamper-resistance requirement if the application of the PUF requires it (cf. 5.6, ISO/IEC 20897-1:2020). If a vendor claims that the PUF satisfies the tamper-resistance requirement, the vendor shall document the reason for that. A PUF vendor shall be responsible for adducing the rationale for the tamper-resistance based on the conducted tests, evaluation results, design and implementation of the PUF. It is the vendor's responsibility, if necessary, to check with the designer, manufacturer, and so forth for detailed information of the PUF in order to establish the evaluation criteria for tamper-resistance.

5.3.6 Test of Mathematical unclonability

The test of the mathematical unclonability verifies whether the PUF's challenge-response behaviour is not simulated or emulated by other devices. A mathematically unclonable PUF generates responses which are difficult to be corelated to the challenge, design and implementation of the PUF. A PUF is mathematically unclonable if it is impossible to disclose the mapping table or function of the CRPs by for example dictionary attacks, machine learning attacks, and so on.

A PUF shall meet the mathematical unclonability if the application of the PUF requires it (cf. 5.6, ISO/IEC 20897-1:2020). If a vendor claims that the PUF satisfies the mathematical unclonability, the vendor shall document the reason for that. A PUF vendor shall be responsible for adducing the rationale for the mathematical unclonability based on the conducted tests, evaluation results, design and implementation of the PUF. It is the vendor's responsibility, if necessary, to check with the designer, manufacturer, and so forth for detailed information of the PUF in order to establish the evaluation criteria for mathematical unclonability.

5.3.7 Test of Physical unclonability

The test of the physical unclonability verifies whether the physical clone of the PUF is impractical to be manufactured. The physical unclonability ensures that there are no two PUFs that have the same input-output behaviour. The physical unclonability is assessed by examining whether the PUF surely utilizes the entropy source derived from the uncontrollable device variation.

A PUF shall meet the physical unclonability if the application of the PUF requires it (cf. 5.6, ISO/IEC 20897-1:2020). If a vendor claims that the PUF satisfies the physical unclonability, the vendor shall document the reason for that. A PUF vendor shall be responsible for adducing the rationale for the physical unclonability based on the conducted tests, evaluation results, design or implementation of the PUF. It is the vendor's responsibility, if necessary, to check with the designer, manufacturer, and so forth for detailed information of the PUF in order to establish the evaluation criteria for physical unclonability.

Annex A
(informative)

Tests of the steadiness

A.1 General

The steadiness should be evaluated by at least one of the following metrics: Intra-HD, the bit error rate (BER), the stochastic model, and so forth. This Annex provide the detail procedure to evaluate the steadiness based on the BER, intra-HD, and stochastic model.

A.2 Bit error rate

Steadiness may be quantified by the average bit error rate of the response sequence arranged in T direction (see Figure 2). The bit error rate is the number of differences obtained during multiple queries divided by the number of queries. It is optimally expressed in logarithmic scale. A vendor should be responsible for determining the criterion of the steadiness considering the applications and operating environment of the PUF.

NOTE For extensive PUFs, some challenges are devoted to challenge-response authentication. The steadiness test is preferable to be performed on the challenges which are not used for the challenge-response authentication.

The test consists in taking N ($=N_{\mathrm{res}} \times N_{\mathrm{chal}}$) measurements. N_c denotes the largest number of identical responses among the N measurements. All other responses (subtraction N-N_{c} in total) are considered incorrect and are thus counted as errors. Therefore, the steadiness rate is quantified by the ratio N_c/N.

Let ϵ be BER of a PUF in which a wrong response is obtained (owing to noise) at maximum once every $1/\epsilon$ queries. The steadiness test is thus to check that the expected value of $\hat{p} = N_{\mathrm{c}} / N$ over n queries (assuming the responses are iid) is such that

$$1 - \epsilon \leq \hat{p} \leq 1$$

with high probability.

Let p be the population rate of $\hat{p}$, then the following inequation is obtained with the confidence level of z:

$$\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Thus, the constraint is on the error interval length

$$2z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \epsilon.$$

Using the relationship

$$\hat{p}(1-\hat{p}) \leq \epsilon \leq \left(\frac{1}{2}\right)^2,$$

the error interval can be bounded to

$$2z\sqrt{\frac{1}{4n}} \leq \epsilon.$$

Therefore, the required number of CRPs under the bit error rate ϵ is given by the following inequation:

$$n \geq \frac{z^2}{\epsilon^2}.$$

A.3 Intra-Hamming distance

Steadiness is often quantified by the intra-HD, especially in academic papers[9][10].

Let $\boldsymbol{D}^{\text{intra}}$ be the 3D vector that consists of all measured responses for N_{puf}, N_{chal} and N_{meas}.

$$D^{\text{intra}} = \left[\text{HD}\left(y_i^{(j_1)}(x_k), y_i^{(j_2)}(x_k)\right)\right]$$

$$\forall 1 \leq i \leq N_{\text{puf}}, \forall 1 \leq k \leq N_{\text{chal}}, \forall 1 \leq j_1 \neq j_2 \leq N_{\text{meas}}$$

The mean of the $\mathbf{D}^{\text{intra}}$ is calculated as follows:

$$\mu^{\text{intra}} = \text{E}\left[D^{\text{intra}}\right] = \frac{2}{N_{\text{puf}} \cdot N_{\text{chal}} \cdot N_{\text{meas}} \cdot (N_{\text{meas}} - 1) \cdot N_{\text{res}}} \Sigma\Sigma\Sigma\Sigma D^{\text{intra}}$$

The standard deviation of the $\mathbf{D}^{\text{intra}}$ is calculated as follows:

$$\sigma^{\text{intra}} = \sqrt{\frac{2}{N_{\text{puf}} \cdot N_{\text{chat}} \cdot N_{\text{meas}} \cdot (N_{\text{meas}} - 1) \cdot N_{\text{res}} - 2} \Sigma\Sigma\Sigma\Sigma\left(D^{\text{intra}} - \mu^{\text{intra}}\right)^2}$$

Generally, when both μ^{intra} and σ^{intra} are around zero, the reproducibility of the response of the PUF is high, and therefore, the steadiness is high.

A.4 Stochastic model

The stochastic model for steadiness can consist in the following steps:

— Estimation of the noise standard deviation A by evaluating the variability in the value of *x*, over few measurements

— Estimation of the intrinsic variability £ , by estimating the distribution of *x* over several devices D

The probability of error (measurement of steadiness) can therefore be extrapolated as the computation of an error function (customarily denoted as "erf").

An example of such computation is provided in Reference [9]. Also, some entropy estimation techniques exist based on models, such as Bayesian approach[12][13]. A vendor should provide detailed documentation or a reference to detailed documentation of the stochastic model if it is used. It is the vendor's responsibility, if necessary, to check with the designer, manufacturer, and so forth for detailed information on the PUF to explain the used stochastic model.

Annex B
(informative)

Tests of the randomness

B.1 General

Randomness can be attested by statistical tests. However, as of today, the tests aim at verifying the randomness offline, for long sequences, but not online, during the nominal operation of the device. Indeed, in online tests, the sequences of random number are short, and the standard tests cannot apply. Therefore, this Annex aims at making standard tests more flexible in terms of random sequence size.

The generalization made here will still match with the standard for the prescribed sequence length, but in addition will apply in the same spirit (i.e., with similar confidence internal) for different lengths.

B.2 Current standards

B.2.1 DieHarder

DieHarder[14] is today considered the ancestor of NIST FIPS SP 800-22 (see B.2.3). The DieHarder test suite requires a lot of data, typically 80 million of bits.

B.2.2 NIST FIPS 140-1

The random tests in FIPS 140-1[2] originally consisted in four tests, namely monobit, poker, runs, long runs. The sequence is short (only 20 000 bits).

These tests are today deprecated. See B.2.3 through B.2.6 for up-to-date tests. Still, current version of FIPS 140-1 keeps a simple health test: the first 16-bit (or greater) block is compared to the previous first 16-bit block of the next sequence. An alarm is generated if they are the same, which would be a possible testimony for a stuck-at issue.

B.2.3 NIST SP 800-22

The NIST FIPS SP 800-22[6] lists 15 demanding tests, most of them inherited from DieHarder, some requiring 1 Mbit of data, and others up to 1 Gbit.

B.2.4 NIST SP 800-90B

The standard NIST FIPS SP 800-90B[7] is methodological. In particular, it aims at understanding whether some implicit hypotheses assumed by other analyses are true in practice. For instance, the iid test is prescribed. It helps clarify the reason for tests failure, if any.

B.2.5 BSI AIS 31

The German BSI innovates in AIS 31[3] mostly by requiring the tests to be performed on the randomness source in addition to the TRNG output. It also introduces the notion of stochastic model in section 2.4.1. Regarding the statistical tests, there are 9 of them. They require only 20 000 bits (like NIST FIPS 140-1, recall 2.2) to yield interesting results.

— Test T0 (disjointness test): birthday paradox for substrings

— Test T1 (Monobit Tests): same as frequency (Monobit) test

— Test T2 (same as Poker test)

— Test T3 (same as runs test)

— Test T4 (long run test)

— Test T5 (autocorrelation test)

— Test T6 (uniform distribution test)

— Test T7 (comparative test for multinomial distributions, also known as "test for homogeneity")

— Test T8 (entropy estimation)

B.2.6 ISO/IEC 20543

The abovementioned tests are either ad hoc (e.g., DieHarder) or national standards (e.g., USA for NIST documents, Germany for AIS 31, etc.). Notice that other countries also emit recommendations, such as France and its general reference on security (RGS: *Référentiel Général de Sécurité*).

For this reason, an international standardization project has been launched. ISO/IEC 20543[15] is an ISO project aiming (in particular) at unifying NIST SP 800-90C[16] and BSI AIS 31. It was developed by ISO/IEC JTC 1/SC 27/WG 3. It insists on the method, but stresses that tests are required. In addition, it does require rationale evidence through the documentation of *stochastic models.*

B.3 Comparison of tests and mathematical background

B.3.1 Comparison of tests

The comparison of the tests is shown in Table B.1. It can be seen that standards differ in the number of bits they require to compute a metric, and also differ in the number of tests. It can clearly be seen that the different methods do not require the same amount of bits N to compute the tests. "Methodological" means that the security is not analyzed through numerical applications, but through a reasoned analysis which justifies qualitatively the security level.

Table B.1 — Comparison of tests

Tests	Publication year	Number of tests	Sequence size N bits
DieHarder[14]	2006	31	≥ 80 000 000
NIST FPIS 140-2[2]	2001	4	20 000
NIST SP 800-22[6]	2010	15	≈ 1 000 000 000
NIST SP 800-90B[7]	2018	Methodological	N/A
BSI AIS 31[3]	2011	9	20 000
ISO/IEC 20543[15]	2018	Methodological	N/A

B.3.2 Mathematical background on statistical test

Different approaches can be taken to develop a battery of statistical tests of randomness in a binary sequence. The most used ones involve computing a test statistic for the generated sequence and the decision rule states that "the sequence fails a test if the test statistic falls outside of a range". The use of this approach implies that significance levels and acceptable ranges are pre-computed. If significance levels are modified in the future, the range values should be recomputed. To accomplish a statistical test, some notions have to be defined:

— The Null hypothesis H_0: The assertion that the "provided sequence is random".

— The test statistic S: A numerical summary that reduces the observed data to one value.

— The p-value: The probability that, when H_0 is true, the statistical test would be the same as or of greater magnitude than the actual observed results.

— The significance level α : The probability of rejecting H_0, given that it were true.

Given these definitions, a statistical test is conducted as follows:

— If p-value $\geq \alpha$, then H_0 fails to be rejected;

— If p-value < α , then H_0 is rejected.

The p-value can be computed given the statistical model of the test statistic *S*. It is to note that according to this approach, different standardized tests are no different from each other given the length of the observed sequence as a parameter.

B.4 Extrapolation of FIPS 140-1 standard tests

B.4.1 General

Without loss of generality to the statistical approach above-mentioned, FIPS 140-1 tests[2] (which also correspond to T1, T2, T3, and T4 tests from BSI AIS 31[3]) are chosen to be dealt with B.4. They all have a significance level "around 10^{-6}" (which will be explained in more detail in B.4.2 to B.4.5). One interesting feature about such four tests is that they can be evaluated "online" using some accumulators: those are refreshed for any incoming bit, and the evaluation of the tests can be finalized at each selected moment in time *N*. Hence low memory requirements, as required for embedded online tests of TRNGs.

In B.4, the goal is to provide a mathematically rigorous extrapolation given a sequence of bits $(b_i)_{1\leq i\leq N}$ of unfixed length *N*. The response bits collection should be carried out in various environmental conditions, as claimed by the PUF vendor. The modus operandi of the PUF responses can depend on the product as well:

— they can be the result of application of challenge bits provided "in order",

— they can result from the alternation of a given number of response bits and PUF reset (to mimic the test of NIST SP 800-90B).

Subsequently, the tests described in B.4.2 to B.4.5 are applied on the collect sequence of bits.

B.4.2 T1: the Monobit test

The Monobit test (Frequency test) determines the proportion of the number of ones and zeros in a bit sequence. For a random sequence, one expects that the average number of ones (and consequently zeros) is 1/2. The Monobit test examines a sequence of *N* = 20 000 bits by calculating the sum T_1 as follows:

$$T_1 = \sum_{i=1}^{20\,000} b_i$$

The test fails if T_1 falls outside the range [9 654, 10 346]. This test is generalized by the following evaluation approach. The statistical test is defined to be $S_N = \sum_{i=1}^{N} 2b_i - 1 = \sum_{i=1}^{N} (-1)^{b_i}$ and the significance level $\alpha = 9{,}6\times 10^{-7}$ (AIS 31). Let $S_{\text{obs}} = |S_N|/\sqrt{2N}$ be the observation to be tested. The reference distribution of this observation is half normal (zero-mean) for large *N*. If the sequence is random, then the plus and minus ones will tend to cancel each other out so that the test statistic is about 0. The p-value is then $\text{efrc}(S_{\text{obs}})$ where $\text{efrc}(z) = \frac{2}{\sqrt{\pi}} \int_z^{+\infty} e^{-u^2} du$ the complementary Gauss error function. Notice that this continuous function is strictly decreasing (hence bijective) from $\mathbb{R}$ to

[0, 1]. According to the evaluation approach, H_0 is accepted if p-value $\geq \alpha$, which leads to the following general boundaries:

$$\frac{N}{2} - \sqrt{\frac{N}{2}} \times \text{erfc}^{-1}(\alpha) \leq T_1 \leq \frac{N}{2} + \sqrt{\frac{N}{2}} \times \text{erfc}^{-1}(\alpha).$$

B.4.3 T2: the Poker test

The test starts by dividing a sequence of N = 20 000 bits into 5 000 contiguous 4-bit segments, then counts and stores the number of occurrences f(i) of each of the possible 4-bit values, for $i \in \{0, 1, \cdots, 15\}$. The sequence passes the test if 1,03 < T_2 < 57,37 where,

$$T_2 = \frac{16}{5\ 000}\left(\sum_{i=0}^{15} \text{f}(i)^2\right) - 5\ 000.$$

The statistical test is

$$S_{N,M} = \frac{2^M}{N/M}\left(\sum_{i=0}^{2^M-1} \text{f}(i)^2\right) - N/M,$$

where N the length of the sequence and M is the length of the contiguous blocks (for simplicity, $M < N$ is assumed here). The significance level of this test is chosen to be $\alpha = 1,014 \times 10^{-6}$ (AIS 31). The purpose of this test is to evaluate the proportion of ones within each M bit block. The frequency of ones in every M bit block should be $M/2$ as would be expected for a random sequence. The number of occurrences f(i) are normally distributed, so $S_{N,M}$ follows a "chi square" (denoted χ^2) distribution of $2^M - 1$ degrees of freedom. To determine whether there is a significant difference between the expected frequencies and the observed frequencies, the observed χ^2_{obs} may be limited within two critical values of $\chi^2_{2^M-1}$ of $2^M - 1$ degrees of freedom, given a significance level α. In AIS31 standard, the significance levels are chosen to be $0,3 \times \alpha$ and $0,7 \times \alpha$. This choice means that the lower bound is $\chi^2_{2^M-1}(0,3 \times \alpha)$ and that the upper bound is $\chi^2_{2^M-1}(1 - 0,7 \times \alpha)$. This analysis leads to the general formula of T_2, given N and M:

$$\chi^2_{2^M-1}(0,3 \times \alpha) \leq T_2(N, M) \leq \chi^2_{2^M-1}(1 - 0,7 \times \alpha).$$

The chi square distribution has only one parameter which determines its degree of freedom. This means that the boundaries are the same whatever N is, unless M (= 4) changes.

B.4.4 T3: The Runs test

A run is defined as a maximal sequence of consecutive bits of either all ones or all zeros, which is part of the N = 20 000 bit sample stream. The occurrences of runs (for both consecutive zeros and consecutive ones) of all lengths ($k \geq 1$) in the sample stream should be counted and stored. The test is passed if the number of runs that occur (of lengths k from 1 to 6) is each within the corresponding interval specified below. This should hold for both the zeros and the ones; that is, all 12 counts should lie within the specified interval. For the purpose of this test, runs of greater than 6 are considered to be of length 6. Let us denote $T_3(k, p)$ the run of length k of the bit $p \in \{0,1\}$.

Intuitively, the purpose of the runs test is to observe whether ones and zeros are not changing too fast (e.g., 0101010101) or too slow (e.g., 0000011111). Following the evaluation approach, a statistical test is defined to be the number of runs of length k denoted $T_3(N, k)$ where N is the sequence length and the significance level is $\alpha = j \times 10^{-6}$, where $j \in \{3,25; 1,33; 0,85; 0,40; 0,10; 0,10\}$ (AIS31). Assuming that the bits in the sequence are independently identically distributed, then similar to the analysis in the case of the Monobit test, the test statistic $T_3(N; k)$ is considered to be normally distributed. The mean of this distribution is $\mu_{N,k} = N/2^{k+2}$, because the probability of each run of 0's or 1's is the same, so in $N/2$

there are k-runs of 0's or 1's and each bit has a probability of 1/2 to occur. This is said, the mean value expected is $(N/2)/2^{k+1} = N/2^{k+2}$. The sum of all the k-runs if the experiment is done an infinite number of times should be $N/2$:

$$\sum_{k=0}^{+\infty} \frac{N}{2^{k+2}} = \frac{N}{4} \sum_{k=0}^{+\infty} \left(\frac{1}{2}\right)^k = \frac{N}{4} \times \frac{1}{(1-1/2)} = \frac{N}{2}$$

Let the observation be $S_{\text{obs}} = \left|\left(T_3(N,k) - \mu_{N,k}\right) / \sqrt{\mu_{N,k}}\right|$, so the p-value is $\text{erfc}\left(\left|\left(T_3(N,k) - \mu_{N,k}\right) / \sqrt{\mu_{N,k}}\right|\right)$ and it is compared to the significance level α. The test passes if p-value $\geq \alpha$. This leads to

$$\mu_{N,k} - \sqrt{2\mu_{N,M}} \times \text{erfc}^{-1}(\alpha) \leq T_3(N,k) \leq \mu_{N,k} + \sqrt{2\mu_{N,k}} \times \text{erfc}^{-1}(\alpha).$$

B.4.5 T4: the Longest Run test

A long run is defined to be a run of length 34 or more of either zeros or ones. On a sample of 20 000 bits, the test is passed if there are NO long runs. The acceptance region ($\leq$ 34) is based on the analysis of a N = 20 000 bit stream. The evaluation approach abovementioned (see 4.3) can be used to calculate the probability for this test using a Runs test within N-bit blocks, but a more efficient method to calculate the long runs probability is found. Let $P_N(k)$ be the probability of a run (0's or 1's) of length longer than or equal to k to appear in a sequence of N bits. This can be expressed by the addition of two mutually exclusive event probabilities. One of them is the probability that the run longer than k appears in (N - 1) bits and the other is the probability that no run longer than k appears in (N - 1) bits but appears in the right most k bits of N (see Figure B.1).

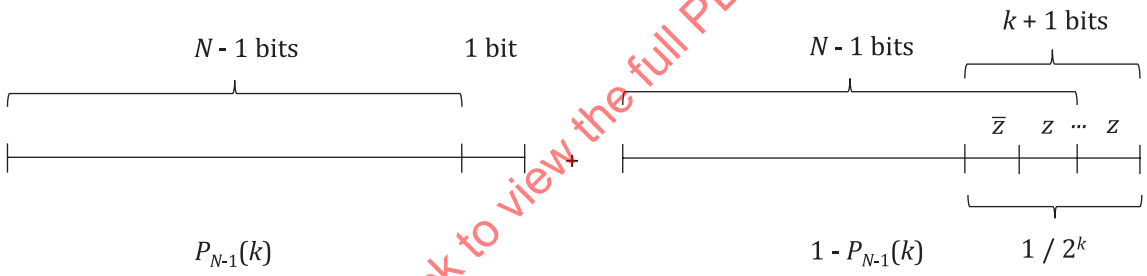

Figure B.1 — The probability of a k-length run

This means that:

$$P_N(k) = P_{N-1}(k) + \left(1 - P_{N-1}(k)\right) \times \frac{1}{2^k},$$

$$P_N(k)=\left(1-\frac{1}{2^k}\right)\times P_{N-1}(k)+\frac{1}{2^k}\ ,$$

which leads to the general expression of $P_N(k)$ given N

$$\begin{cases} 0 & \text{if } N \le k-1 \\ \dfrac{1}{2^{k-1}} & \text{if } k = N \\ \left(1-\dfrac{1}{2^k}\right)\times P_{N-1}(k)+\dfrac{1}{2^k} & \text{if } N \ge k+1 \end{cases}$$

This is an arithmetic-geometric progression $U_N = a\times U_{N-1}+b$ and it is easily found that the general term $U_N = a^N\times(U_0-r)+r$ where $r=\frac{b}{1-a}$, thus:

$$P_N(k)=\left(1-\frac{1}{2^k}\right)^N\times(P_0-r)+r$$

here $U_0 = 0$ and $r = 1$. So $P_N(k)$ is obtained as follows:

$$P_N(k)=1-\left(1-\frac{1}{2^k}\right)^N$$

Therefore, the longest run k is found as follows (assuming $P_N(k) = \alpha$):

$$k=-\log_2\left(1-(1-\alpha)^{1/N}\right).$$

B.5 Entropy estimation Using NIST SP800-90B

The randomness of a PUF may be evaluated by the entropy estimation provided in NIST SP 800-90B[7]. The data set for the entropy estimation in the randomness evaluation may be taken from the B-C plain of the aligned responses as shown in Figure B.2. For the detailed data construction, see the restart data in 3.1.4.1 in Reference [7].

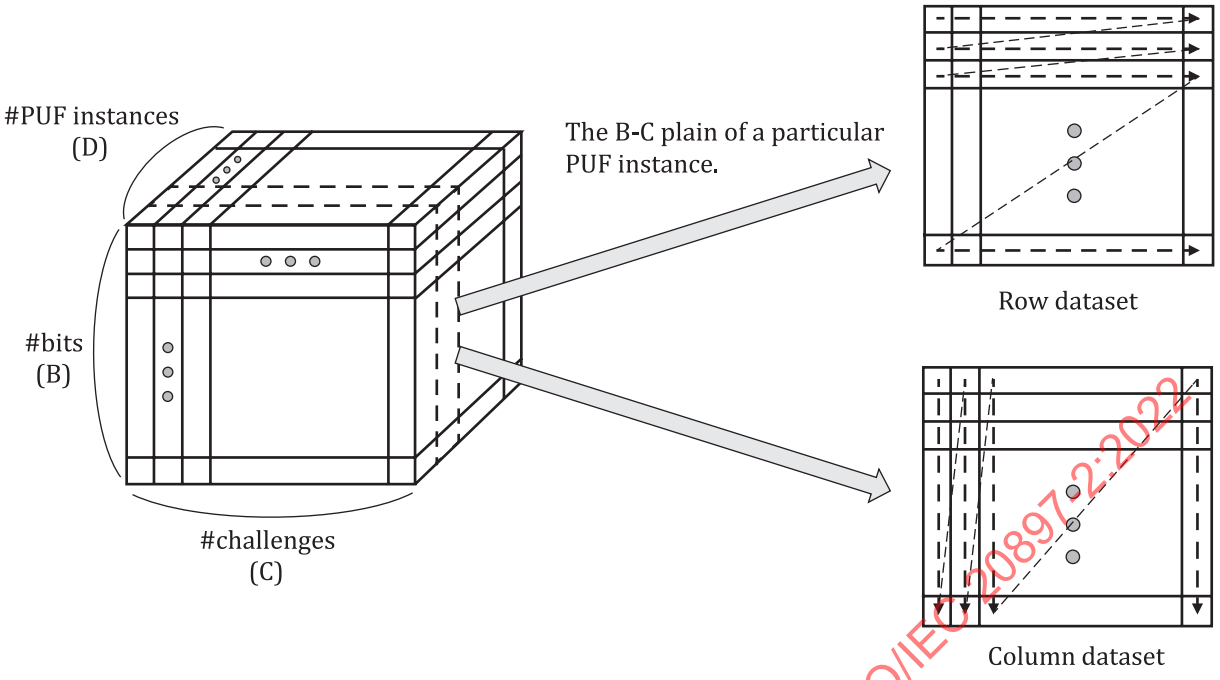

Figure B.2 — Test data construction example of the entropy estimation in the randomness evaluation.

Annex C
(informative)

Tests of the uniqueness

C.1 General

Uniqueness may be quantified by the inter-HD, statistical tests, or stochastic model. In C.2 through C.4, methods to evaluate the uniqueness are exemplified in detail.

C.2 Inter-Hamming distance

Uniqueness is often quantified by the inter-HD, especially in academic papers[9][10]. Let $\boldsymbol{D}^{\text{inter}}$ be the vector that consists of all the measured N_{res}-bit responses for N_{puf}, N_{chal} and N_{meas}.

$$\boldsymbol{D}^{\text{inter}} = \left[HD\left(y_{i1}^{(j)}(x_k); y_{i2}^{(j)}(x_k) \right) \right]$$

$$\forall 1 \le i_1 \ne i_2 \le N_{\text{puf}},\ \forall 1 \le k \le N_{\text{chall}},\ \forall 1 \le j \le N_{\text{meas}}.$$

The mean of the $\mathbf{D}^{\text{inter}}$ is calculated as follows:

$$\mu^{\text{inter}} = \mathrm{E}\left[\boldsymbol{D}^{\text{inter}} \right] = \frac{2}{N_{\text{puf}} \cdot \left(N_{\text{puf}} - 1 \right) \cdot N_{\text{chal}} \cdot N_{\text{meas}} \cdot N_{\text{res}}} \sum\sum\sum\sum \boldsymbol{D}^{\text{inter}}$$

The standard deviation of the $\mathbf{D}^{\text{inter}}$ is calculated as follows:

$$\sigma^{\text{inter}} = \sqrt{\frac{2}{N_{\text{puf}} \cdot \left(N_{\text{puf}} - 1 \right) \cdot N_{\text{chal}} \cdot N_{\text{meas}} \cdot N_{\text{res}} - 2} \sum\sum\sum\sum \left(\boldsymbol{D}^{\text{inter}} - \mu^{\text{inter}} \right)^2}.$$

The evaluation of the uniqueness using the above formulae assumes that the response is iid. As pointed out in,[17] when the responses are non-iid, the standard deviation of inter-HD varies greatly depending on the length of the response block. According to Reference [17] if the responses are non-iid, the standard deviation of Inter-HD becomes larger as the block size increases. If the inter-HD is used for the evaluation of the uniqueness, the block size should appropriately be set based on the expected PUF application.

C.3 Entropy estimation

The uniqueness of a PUF may be evaluated by the entropy estimation provided in NIST SP 800-90B[7]. The data set for the entropy estimation in the uniqueness evaluation may be taken from the B-D plain of the aligned responses as shown in Figure C.1. For the detailed data construction, see the restart data in 3.1.4.1 in Reference [7].

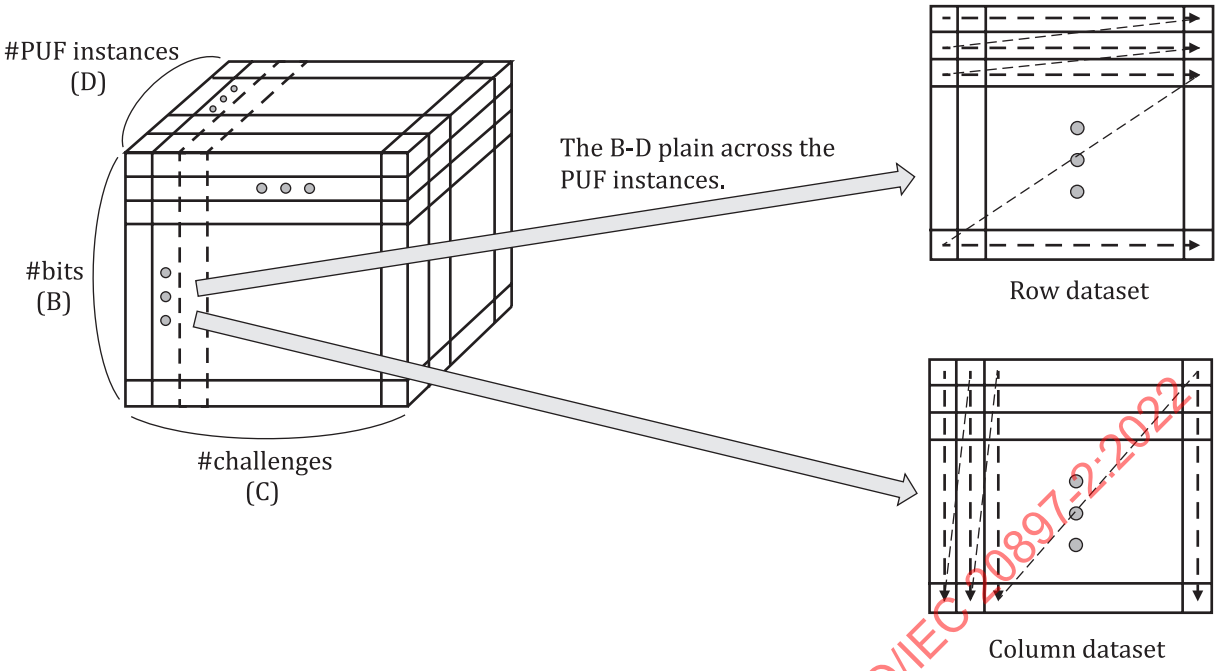

Figure C.1 — Test data construction example of the entropy estimation in the uniqueness evaluation.

C.4 Stochastic model

A study to evaluate the uniqueness of a PUF using a stochastic model has been reported in [18]. In this paper, a loop PUF, which is one of the delay-based extensive PUFs, is modelled as a chain of M delay elements that follows normal distributions. When the number of the loop PUFs is L, the $M \cdot L$ delay elements follow the global distribution D; the i-th delay elements in L PUFs follow the distribution D_i^L $(i \in L)$. Then, the common area between the distribution D and D_i^L quantifies the uniqueness.

Annex D
(informative)

Example of the test of the PUF security requirements

D.1 General

This Annex provides an example of tests of the PUF security requirements. The devices under tests (DUTs) are electronic control units (ECUs) in the automotive industry market. Note that these ECUs are not real products; they are fictional ones only used in this document to exemplify the tests of PUF security requirements. The evaluation criteria described later in D.2 and D.3 are also imaginary, and thus they are not necessarily applicable for real products. In actual tests, the test procedure, test conditions, and evaluation criteria should be clearly explained and justified in a document by a PUF vendor based on the applications and operation environment of the PUFs.

D.2 Example of the specifications, evaluation criteria, and test conditions of DUT

D.2.1 Specifications of the ECU

The DUTs are ECUs equipped with PUFs in the automotive industry market. The PUF can be used in the context of, for example, secure boot and key-encryption-key, whereby the PUF value allows to derive a master key. This master key allows to check for the integrity of the firmware installed in the ECU. The specifications of the ECU with a PUF are shown in Table D.1.

Table D.1 — Example of the specifications of the ECU under tests

Product model number	XXXX-ECU-YYYY (ZZZZ company)
Environment of use	In-vehicle
PUF Type	128-bit challenge Arbiter PUF
PUF usage	256-bit key generation
Technology	180 nm CMOS
Operating clock frequency	100 MHz
Operating voltage range	1,0 V ± 5 %
Operating temperature range	-40 ~ 125 °C

D.2.2 PUF interfaces

The PUF in the DUT has the interfaces listed in Table D.2 for data transfer and control.

Table D.2 — Interfaces of the PUF in the DUT

Name	Direction	Width [bit]	Description
CLK	IN	1	100 MHz clock signal.
RSTn	IN	1	Active low reset signal.
EN	IN	1	Active high enable signal.
DIN	IN	128	Challenge and helper data input to the PUF.
CHA_RDY	IN	1	Active high signal to assert when a challenge data is ready.

Table D.2 *(continued)*

Name	Direction	Width [bit]	Description
HLP_RDY	IN	1	Active high signal to assert when a helper data is ready.
START	IN	1	Active high signal to start the PUF.
DOUT	OUT	128	Error-corrected response data output from the PUF.
BUSY	OUT	1	Active high signal asserted during at least one of the arbiter PUF core and the fuzzy extractor is running.
END	OUT	1	Active high signal asserted for one clock cycle when the 128-bit error-corrected PUF output is ready.

D.2.3 PUF building blocks

The building blocks of the PUF in the DUT are shown in Figure D.1. The PUF consists of the arbiter PUF core (main block) and fuzzy extractor (post-processing block). The helper data preparation is implemented on the vendor side, and therefore, the pre-processing block is not implemented on the chip. Here, the target of evaluation is the Arbiter PUF core (main block). In some products, the target of evaluation may be the entire PUF block including the Fuzzy Extractor (post-processing block).

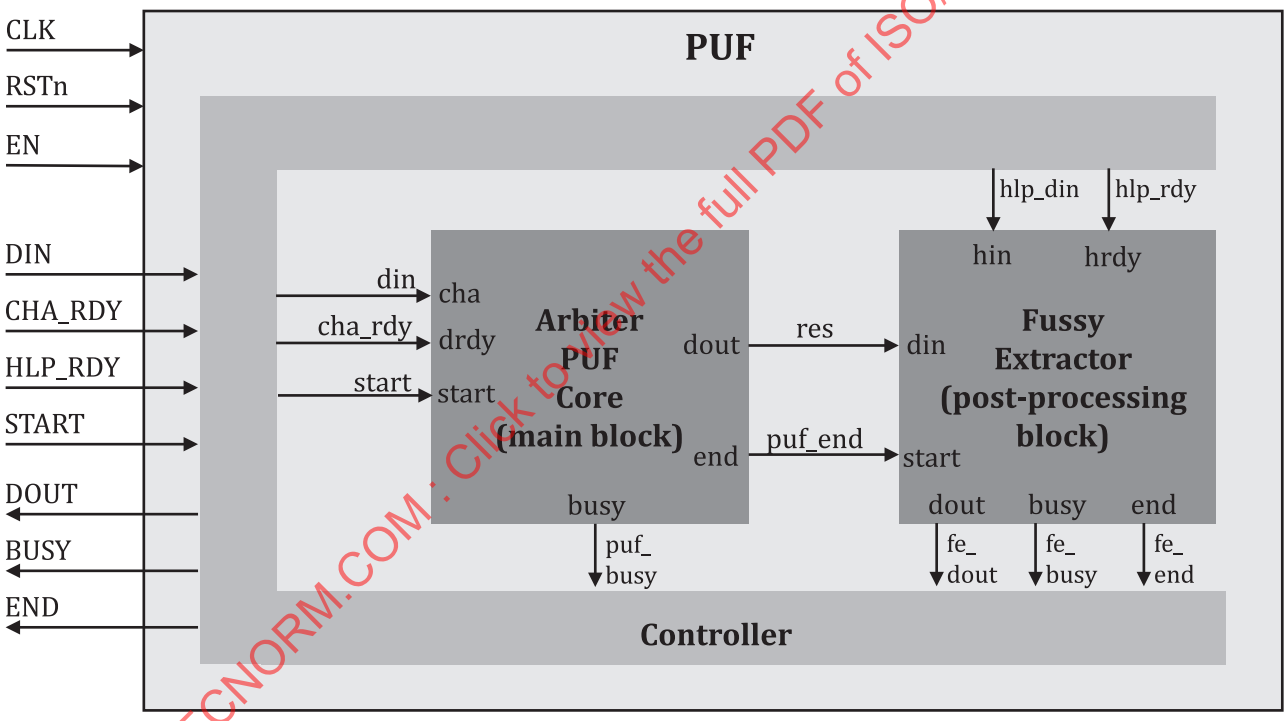

Figure D.1

Figure D.1 — Building blocks of the PUF in the DUT.

D.2.4 Security requirements and evaluation criteria

The security requirements of the PUFs embedded in the ECUs and their evaluation criteria are shown in Table D.3. Indeed, steadiness is the most relevant parameter, as it allows make sure the master key (in the case of secure boot) works correctly, without needing to reboot several times until success.

Second, the randomness is important, as it allows for instance to make sure each firmware has an unpredictable key. This allows to make sure the system is untamperable, in terms of signature forging.

The uniqueness allows to make sure that one attacked device, for which the key is disclosed, could not be used by another device (the so-called "attack once, break all" issue).

Table D.3 — Example of the security requirements and evaluation criteria of the ECUs

Security requirements	Evaluation criteria
Steadiness	ex. 1: The expectation of the BER of the DUT is smaller than 10^{-9}. ex. 2: The mean of the normalized intra-HD of the DUT is less than 2×10^{-2} and the standard deviation is less than 2×10^{-2}.
Randomness	ex. 1: The DUT passes the T1~T4 tests under the parameter N_{chal} given in D.2.2 and the significance level $\alpha = 10^{-7}$ (NIST default value). ex. 2: The min-entropy per bit of the responses is larger than 0,9 under the parameter N_{chal} given in D.2.2.
Uniqueness	ex.: The mean of the normalized inter-HD of the responses is in the range of [0,45, 0,55] and the standard deviation is less than 0,5.
Tamper-resistance	ex.: The secret information is not obtained by invasive and semi-invasive attacks (e.g., milling and delayer of the chip).
Mathematical unclonability	ex. 1: If the CRPs of this ECU do not go outside the chip, mathematical unclonability need not be considered for the chip. ex. 2: The prediction accuracy of the machine learning attack for unknown challenges is less than 60 %. The number of training data for machine learning is at least 5 000 000.
Physical Unclonability	ex.: The responses are generated using uncontrollable and unclonable physical variation.

D.2.5 Example of the test conditions

The operation voltage of the ECU is 1,0 V and the operation temperature range is -40 °C ~ 125 °C as provided in Table D.1. The test conditions of the ECUs are set as follows:

— Nominal test: Steadiness, randomness and uniqueness are tested at 1,0 V and 25 °C.

— Voltage fluctuation test: Steadiness is tested at 25 °C and 0,95 V ~ 1,05 V with 0,05 V steps.

— Temperature fluctuation test: Steadiness is tested at 1,0 V and the following temperatures [°C]: {-40, -20, 0, 25, 50, 75, 100, 125}.

— VT Corner test: Steadiness is tested at (0,95 V, -40 °C), (0,95 V, 125 °C), (1,05 V, -40 °C), (1,05 V, 125 °C).

Other test parameters are as follows:

— N_{puf}: 1 000

— N_{chal}: 1 000 000

— N_{meas}: 1 000

— N_{res}: 1