Abstraction Model of Probing and DFA Attacks on Block Ciphers[⋆]

Yuiko Matsubara[1], Daiki Miyahara[1,2], Yohei Watanabe[1,2], Mitsugu Iwamoto[1], and Kazuo Sakiyama[1]

[1] The University of Electro-Communications, Chofu, Tokyo 182–8585, Japan
yuiko.matsubara@mail.uec.jp, sakiyama@uec.ac.jp
[2] National Institute of Advanced Industrial Science and Technology (AIST), Koto, Tokyo 135–0064, Japan

Abstract. A thread of physical attacks that try to obtain secret information from cryptographic modules has been of academic and practical interest. One of the concerns is determining its efficiency, e.g., the number of attack trials to recover the secret key. However, the accurate estimation of the attack efficiency is generally expensive because of the complexity of the physical attack on a cryptographic algorithm. Based on this background, in this study, we propose a new abstraction model for evaluating the attack efficiency of the probing and DFA attacks. The proposed model includes an abstracted attack target and attacker to determine the amount of leaked information obtained in a single attack trial. We can adapt the model flexibly to various attack scenarios and can get the attack efficiency quickly and precisely. In the probing attack on AES, the difference in the attack efficiency is only approximately 0.3% between the model and experimental values, whereas that of a previous model is approximately 16%. We also apply the probing attack on DES, and the results show that DES has a high resistance to the probing attack. Moreover, the proposed model works accurately also for the DFA attack on AES.

Keywords: Physical attack · Probing attack · Differential fault analysis · Advanced encryption standard · Information leakage.

1 Introduction

There are two main types of security evaluations for block ciphers. One is the theoretical analysis or cryptanalysis, in which the intermediate value in cryptographic processing is unknown to an attacker. We evaluate the amount of public information and the analysis time required to recover the secret key and compare them to the existing computational resources to verify its level of security. The other is the so-called physical attack. The attacker uses a physical side channel to obtain information that cannot be obtained through cryptanalysis. Because more information can be used for the analysis, the data and computation complexities required to identify the secret key are much lesser than in a theoretical attack.

[⋆] An earlier version of this paper was presented and appeared as conference papers [2, 3].

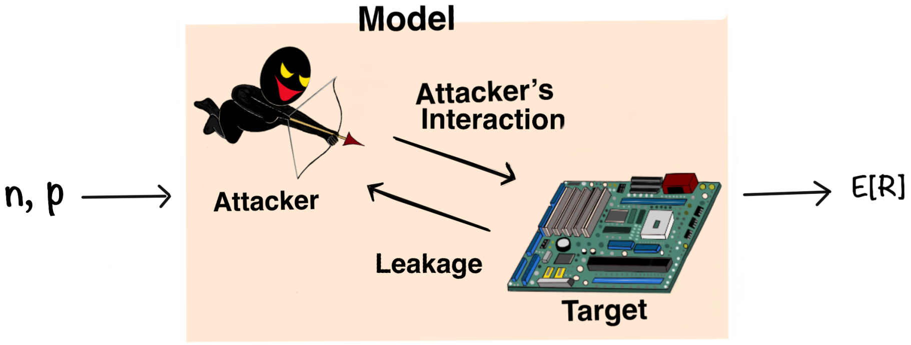


Fig. 1: Abstraction model for the physical attacks on block ciphers

The most significant difference between the theoretical and physical attacks is whether or not the attacker can obtain a cryptographic module that is the target of the physical attacks. In the era of the internet of things (IoT), IoT devices equipped with cryptographic functions are placed within reach of the attackers, and the attack surface of the physical attacks is steadily increasing. It is essential to accurately determine the physical attack resistance when evaluating the security of block ciphers. Furthermore, the discussion and modeling of ideal attack resistance will be used as a security indicator in designing future block ciphers.

In this study, we propose a new abstraction model for evaluating the physical attack resistance of block ciphers. We aim to build a model with a good balance in simulation time, cost, and accuracy. The model assumes an ideal abstracted attack target and an ideal attacker at a high level of abstraction, which enables us to perform millions of accurate simulations only in a few seconds. According to this model, the attacker interacts with an attack target and obtains information leakage from the target, as shown in Fig. 1. The attacker attempts to identify the secret key of the cryptographic algorithm implemented in the target device. The model parameters are the survival probability of each key candidate p and the bit size of the key to be recovered n. The probability p is determined by the physical attack scenario and the cryptographic algorithm. Consequently, the model provides the expected number of attack trials to derive the secret keys, $E[R]$, where R denotes a random variable for the number of trials. It is worth noting that the attack efficiency is also regarded as the attack resistance of the attack target, i.e., the block cipher algorithm.

1.1 Previous studies

Studies on block ciphers mainly focus on the advanced encryption standard (AES) [4], which is now widely used in applications such as wireless communication and integrated circuit (IC) cards [5, 6, 7, 8, 9, 10]. Although the key-recovery attacks, where an attacker attempts to recover a secret key, are feasible with both the cryptanalysis and the physical attack, the physical attacks are considered powerful due to their low computational cost. Therefore, the key-recovery attacks are intensively discussed over the hardware and software implementations of AES.

The physical information called side-channel information is leaked information from a cryptographic device. Well-known side-channel information are power consumption and electromagnetic radiation. The attacker analyzes the side-channel information to obtain the intermediate values of the cryptographic algorithm to narrow down the keyspace to identify the secret key. The direct probing of wires in electronic devices also enables attackers to read out the intermediate values [11, 12]. Stack-at-fault injection can be used for obtaining the intermediate values by checking for the occurrence of a fault [13]. Another powerful attack is known as the differential fault analysis (DFA) [14]. The DFA attack analyzes the difference in the value of an erroneous output during fault injection and reveals the secret key.

In the physical attacks, the number of attack trials to the target device is determined by the attacker's ability and the attack resistance of the cryptographic algorithm. Therefore, it could be a security indicator for attack efficiency. For example, in power analysis, one of the side-channel attacks using power consumption, we refer to attack efficiency in terms of the number of power traces acquired during cryptographic operations. In the probing attack, the attack efficiency is the number of probings, and in the fault attacks, it is the number of fault injections.

Most studies on the side-channel attacks have observed that environmental noise significantly affects attack efficiency. Therefore, they focus on recovering the most convenient intermediate value from the noisy side-channel information. Due to this fact, we often do not deeply discuss the security of cryptographic algorithms [15, 16, 17, 18, 19] against the side-channel attacks. For the above reason, the key-recovery attack using the probing and DFA attacks is considered suitable for evaluating the security of cryptographic algorithms. The fact that we can utilize a vast amount of knowledge in cryptanalysis also helps in validating the attack efficiency of the probing and DFA attacks. Note that even in the probing and DFA attacks, there are errors owing to the influence of noise during the attack; however, the noise is not as dominant as that in the side-channel information.

In [1], the authors estimated the upper limit of the amount of leaked information in the DFA attacks. Moreover, they evaluated the efficiency of the DFA attacks using the actual cryptographic algorithms of the previous study. More specifically, the minimum number of attack trials required to identify the secret key was derived, and the optimality of the attacks in previous experiments was verified. The results were consistent with the actual experiments when the amount of information leaked in a single trial is significant, i.e., when the attack efficiency is high and the secret key can be uniquely identified using a small number of attack trials. However, the model in [1] has a problem that the results do not match the experimental values when the leakage information in one fault injection is small.

Therefore, there are still some critical viewpoints in modeling the amount of information leakage.

1) Disadvantages brought by the model's simplicity [1]

If the model for the attack and the target, i.e. cryptographic algorithm, implementation, and so on, are too simple, it will not match the experimental value since the description is extremely simple. Therefore, even if the optimality for the number of attacks can be discussed to some extent, the attacker in a physical attack does not extend to understanding the relationship between the attacker's ability and the amount of leaked information, as well as the attack resistance of cryptographic algorithms. In addition, variations in attacks cannot be expressed, and all the cryptographic algorithms are identical.

2) Advantages brought by model simplicity [1]

The amount of leaked information can be calculated, and an intuitive understanding of the specific physical attack resistance can be obtained.

1.2 Research motivation

In this study, we propose an abstraction model to overcome the disadvantage of [1] over keeping its advantage. Namely, the proposed model needs to be capable of performing lightweight simulations with high accuracy. The proposed model parameterizes the attacker's ability, i.e., the interaction of intermediate data such as the number of simultaneous probing bits and the attack target, i.e., the target cryptographic algorithm. Consequently, the discussion will accelerate the various physical attacks on different block ciphers to evaluate the amount of leaked information and the physical security of cryptographic algorithms.

1.3 Organization

The notations in this study are listed in Table 1. Section 2 describes a model for the probing attack, and Sect. 3 discusses the difference between the model and experimental results. Section 4 deals with the case of the DFA attack, and Sect. 5 summarizes the results using the model and experiments. Section 6 discusses the power of the proposed model, and finally Sect. 7 concludes the study.

2 Abstraction Model for the Probing Attack

In our model, we assume that the attacker can neither change the probing position nor directly consider the key register[3]; however, they can make several queries to obtain plaintext or ciphertext. Therefore, in this study, we set the attacker more realistically than in [1].

2.1 Target of the abstraction model

Figure 2 describes the abstraction model for the probing attack. The four cases shown in Fig. 3 are conceptual diagrams of typical attack targets based on AES and DES encryptions[4]. The attacker obtains leakage information π and derives secret information

[3] Generally, it is difficult to drill the target register of an IC device and directly read heavily guarded memory where the key is stored.

[4] For simplicity, we focus on attacks during encryption, although the proposed model is applicable to the cases for decryption.

Table 1: Notations used in the study

Notation	Explanation
n	Number of bits for input and output values of a bijective function
n_i	Number of bits for an input value of a surjective function
n_o	Number of bits for an output value of a surjective function
$f : \{0,1\}^n \to \{0,1\}^n$	n-bit bijection function,
$x \in \{0,1\}^n$	Input value of function f
$y \in \{0,1\}^n$	Output value of function f
$k \in \{0,1\}^n$	Key candidate
k^*	Correct secret key
$\mathcal{K}$	Set of key candidates
k_i	Key for i-th round
r	Number of attack trials required to derive the correct key
m	Amount of information obtained from an attack trial
$\mathcal{B} \subseteq \{1,2,\ldots,n\}$	Set of bit positions for probing,
$\pi \in \{0,1\}^{\lvert\mathcal{B}\rvert}$	Information leaked by probing
$\overline{\mathcal{B}}$	Complement of $\mathcal{B}$
$x_{\mathcal{B}}$	A value of x at bit position $\mathcal{B}$
p	Probability that an incorrect key remains after an attack trial
q	Probability that a ciphertext is obtained
A_r	Number of false-key candidates after the rth attack trial
$\mathcal{K}(y,\pi)$	Set of key candidates after the first attack trial
$\mathcal{Y}(k,\pi)$	Set of ciphertexts given π and a key candidate
Δx	Input difference of a bijective function
Δy	Output difference of a bijective function
$\Delta\mathcal{X}$	Set of Δx

based on probings, where $|\mathcal{B}|$ is the number of probing bits, and q is the probability that the attacker can obtain the probing data after one attack trial. A measurement error that an attacker can detect happens with probability q. Note that as will be discussed in Sect. 2.4, an ineffective fault attack (IFA) [21] is regarded as a probing attack because the attacker knows the intermediate values only when the injected faults, e.g. stuck-at-fault, do not affect the computation results.

Depending on whether the function is bijective or surjective, there are two types of abstraction models for the target.

1) Bijective function f

Bijection means that for every element y, there exists one x such that $f(x) = y$. Regarding this case, input bit size is equal to the output bit size. Figure 3a illustrates the attack using the ciphertext. The ciphertext y and probe information π are the public information, and intermediate value x and the secret key k are the secret information. The n-bit input date x is calculated with function f, which is XORed with k and becomes the output data y. Figure 3b shows the attack using plaintext. The plaintext x and probe information π are the public information, and the intermediate value y and the secret key k are the secret information. The n-bit input date x is XORed with k, which is cal-

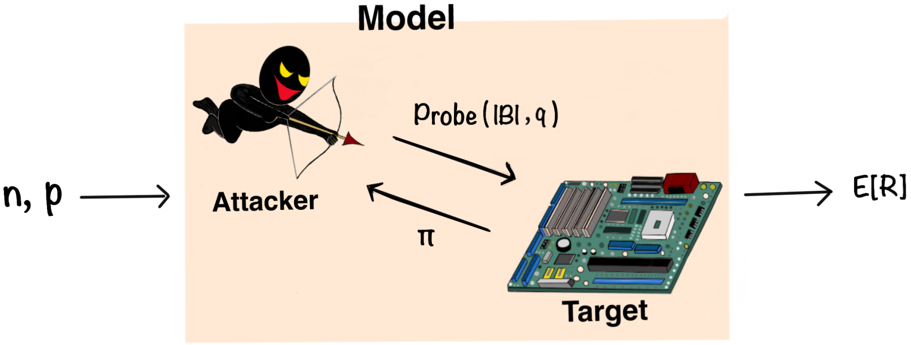


Fig. 2: Abstraction model for the probing attack, where attacker's interaction is a probing to the target devise and the leakage is bit values derived from the probing positions.

culated with function f and becomes the output data y.

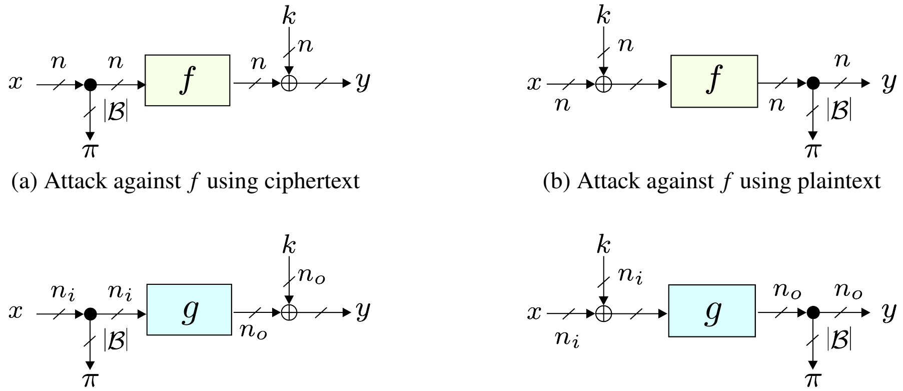
2) Surjective function g
Surjection means that for every y, there exists an x such that $f(x) = y$. In this case, we have $n_i > n_o$. Figure 3c shows the attack using ciphertext. Because n_i is larger than n_o and we cannot derive a unique key, we need to assume a reverse lookup table in this case. Figure 3d shows the attack using plaintext.

2.2 Procedure of the probing attack

The attack target shown in Fig. 3a is discussed here. Algorithm 1 describes the process of selecting a set of key candidates $\mathcal{K}$ for each probing. The attacker obtains leaked information $\pi \in \{0,1\}^{|\mathcal{B}|}$ from the bit position $\mathcal{B} \subseteq \{1, 2, \ldots, n\}$ and the corresponding output value y, computing intermediate value $x_{\mathcal{B}} = f^{-1}(y \oplus k)_{\mathcal{B}}$ from y and the guessed key $k \in \mathcal{K}$. If $x_{\mathcal{B}} \neq \pi$, then k is eliminated from $\mathcal{K}$. The same process is repeated for all n bits. $\mathcal{K}$ can be gained by one probing, and the attacker can perform this procedure several times.

Algorithm 1 Phase 1 of the probing attack

```
Input: leaked information $\pi$, output public value $y$, and an $n$-bit bijective function $f$, the position of probing $\mathcal{B}$
Output: a set of key candidates $\mathcal{K}$
$\mathcal{K} \leftarrow \{0, \ldots, 2^n - 1\}$
for $k = 0$ to $2^n - 1$ do
    $x_{\mathcal{B}} \leftarrow f^{-1}(y \oplus k)_{\mathcal{B}}$
    if $x_{\mathcal{B}} \neq \pi$ then
        $\mathcal{K} \leftarrow \mathcal{K} \setminus \{k\}$
    end if
end for
return $\mathcal{K}$
```



(a) Attack against f using ciphertext

(b) Attack against f using plaintext

(c) Attack against g using ciphertext ($n_i > n_o$)

(d) Attack against g using plaintext ($n_i > n_o$)

Fig. 3: Conceptual diagram for the probing attack during encryption

Algorithm 2 Phase 2 of the probing attack

```
Input: a set of key candidate $\mathcal{K}^{(1)}$ obtained with Algorithm 1 with a probing attack
Output: correct key $k^*$ and the number of probing trials $r$
$\mathcal{K} \leftarrow \mathcal{K}^{(1)}$, $r \leftarrow 1$
while $|\mathcal{K}| > 1$ do
    $r \leftarrow r + 1$
    $\mathcal{K}^{(r)} \leftarrow$ Algorithm 1 with a fresh attack setting, i.e., different probing point or plaintext
    $\mathcal{K} \leftarrow \mathcal{K} \cap \mathcal{K}^{(r)}$
end while
return $k^* \in \mathcal{K}$ and $r$
```

Algorithm 2 shows the procedure for deriving the correct key. While $|\mathcal{K}|$ is more than one, the key candidates are narrowed down by intersection with $\mathcal{K}$ and $\mathcal{K}^{(r)}$, where $\mathcal{K}^{(r)}$ is gained by r-th probing in Alg. 1. This is repeated until a unique key candidate remains, and the correct key k and number of probes r are returned. In this study, we assume that one key always remains.

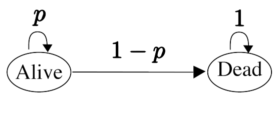


Fig. 4: Markov process for each false key

2.3 Ideal cipher

We assume that a set of candidates of $\mathcal{K}$ is chosen randomly, and a probability that a candidate is included in $\mathcal{K}$ after Step 4 in Alg. 2 is constant for any candidate. We refer to this probability p as the survival probability of each false key and assume that the ideal cipher has a property such that p is constant. Such a cipher is resistant to physical attacks because it does not have strong or weak keys against probing attacks. The false key is alive with a probability p and dead with a probability $1 - p$, as shown in Fig. 4.

Hereafter, we explain the model's features in detail. The attacker obtains y and π in a trial. Regarding Alg. 1, the size of $\mathcal{K}$ is always restricted to $2^{n-|\mathcal{B}|}$. Let $\mathcal{Y}(k, \pi)$ be a set of possible ciphertexts after one trial, written as

$$\mathcal{Y}(k, \pi) := \{y \mid f^{-1}(y \oplus k)_{\mathcal{B}} = \pi\}. \tag{1}$$

An ideal cipher satisfies that $|\mathcal{Y}(k, \pi)|$ is constant for $k \neq k^*$, i.e., the probability p that a false key candidate survives after one trial is constant.

2.4 Deriving the number of attack trials

An ineffective fault attack [21] is an attack that uses fault-free information when a fault is injected. A ciphertext is available only when a specific intermediate value is zero at the moment of fault injection [22]. This is because the calculated results do not change, even if faults are injected. We model the attack as a Markov process, such that the ciphertext is obtained with probability q after an attack trial, and a false key remains with probability p.

Figure 5 shows the entire process of key derivation using a Markov process. If the elements of the key candidate $\mathcal{K}$ are chosen randomly in each trial in Alg. 1, the process of narrowing the key in Alg. 2 can be regarded as a Markov process. The circled state ϵ indicates that the ϵ false keys remain. Here, $\beta = 2^n - 1$ and $\epsilon = 2^{n-|\mathcal{B}|} - 1$. The state 0 indicates that there are either zero false keys or the correct key has been derived. The arrow pointing from state ϵ to state $\epsilon - 1$ shows the probability of moving from state ϵ to state $\epsilon - 1$ by one probe. Firstly, starting from the initial state β, the process moves to the subsequent state after obtaining the ciphertext with probability q and deriving a key candidate in Alg. 1. Except for the initial state, the same state is continued when any keys do not remain, or any ciphertext is not gained after one proving trial. The number of key candidates always becomes ϵ after the first proving trial because the first trial is a bijection. Regarding the second trial, the bijection collapses; therefore, we assume that false keys remain with probability p. The trials are continued until reaching the state 0. Let A_r denote the number of false-key candidates after the r-th attack trial. The probability that $A_r = i$ is

$$\Pr[A_{r+1} = i] = (1 - q) \Pr[A_r = i] + q \sum_{j \geq i} \Pr[A_r = j] \binom{i}{j} p^i (1 - p)^{j-i}. \tag{2}$$

Then, the probability that A_r becomes zero at the r-th trial is $\Pr[A_{r+1} = 0] - \Pr[A_r = 0]$. Therefore, the expected number of trials $E[R]$ is

$$E[R] = \sum_{r=0}^{\infty} r(\Pr[A_{r+1} = 0] - \Pr[A_r = 0]).$$

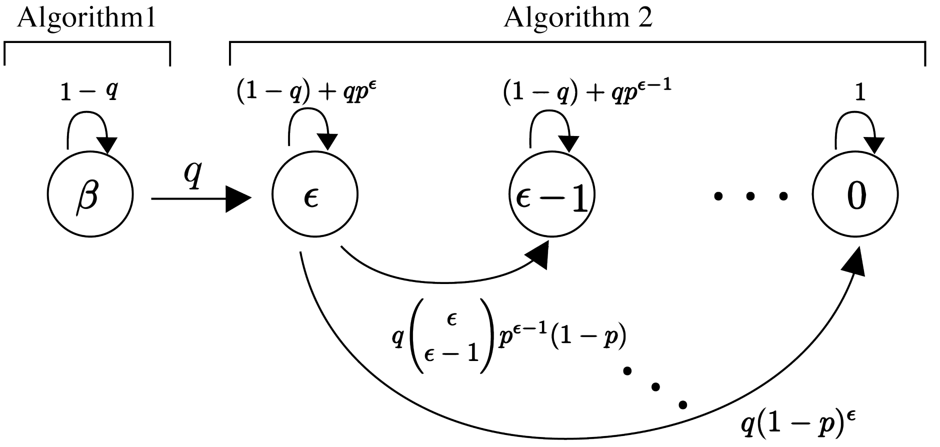


Fig. 5: Markov process of reducing keyspace

2.5 Derivation of probability *p*

We review the derivation of the probability p that a false key survives after a trial. We define $p = \Pr[Y \in \mathcal{Y}(k,\pi)]$. Let Y be random variables with possible values of $y \in \{0,1\}^n$.

When y and π are obtained, a set of key candidates after one trial is described as

$$\mathcal{K}(y,\pi) \coloneqq \{k \mid f^{-1}(y \oplus k)_{\mathcal{B}} = \pi\}, \tag{3}$$

where π is $|\mathcal{B}|$ bits and y is n bits. The size of $\mathcal{K}(y,\pi)$ is

$$\forall y, \forall \pi, |\mathcal{K}(y,\pi)| = \frac{2^n}{2^{|\mathcal{B}|}} = 2^{|\overline{\mathcal{B}}|}. \tag{4}$$

Here, $\overline{\mathcal{B}}$ is the complement set of $\mathcal{B}$. The total number of $|\mathcal{K}(y,\pi)|$ for all y is

$$\forall \pi, \sum_{y \in \{0,1\}^n} |\mathcal{K}(y,\pi)| = 2^n 2^{|\overline{\mathcal{B}}|}. \tag{5}$$

By the definitions of $\mathcal{Y}(k,\pi)$ and $\mathcal{K}(y,\pi)$, we obtain

$$\sum_{k \in \{0,1\}^n} |\mathcal{Y}(k,\pi)| = \sum_{y \in \{0,1\}^n} |\mathcal{K}(y,\pi)|. \tag{6}$$

Because $f^{-1}(y \oplus k)_{\mathcal{B}} = \pi$ always holds for $k = k^*$, we have

$$\forall \pi, |\mathcal{Y}(k^*,\pi)| = 2^n. \tag{7}$$

Considering Eqs. (5) and (6),

$$\begin{aligned}\sum_{k \neq k^*} |\mathcal{Y}(k, \pi)| &= 2^n 2^{|\overline{\mathcal{B}}|} - 2^n \\ &= 2^n (2^{|\overline{\mathcal{B}}|} - 1). \end{aligned} \tag{8}$$

As previously mentioned in Sect. 2.3, we assume an ideal function such that $|\mathcal{Y}(k, \pi)|$ is constant for $k \neq k^*$. Thus, we obtain

$$\begin{aligned} p &= \Pr[Y \in \mathcal{Y}(k, \pi)] \\ &= \frac{1}{2^n} |\mathcal{Y}(k, \pi)| \\ &= \frac{2^{|\overline{\mathcal{B}}|} - 1}{2^n - 1}. \end{aligned} \tag{9}$$

For instance, if $|\mathcal{B}| = 1$ and $n = 8$, we have

$$p = \Pr[Y \in \mathcal{Y}(k, \pi)] = \frac{127}{255}. \tag{10}$$

In the case of the block ciphers with surjective functions, as shown in Fig. 3d, we assume that the bit-size difference in input and output is complemented equally, i.e., the output probed value $|\mathcal{B}|$ is equivalent to $(n_i/n_o)|\mathcal{B}|$ at the input of the function g. In this way, based on Eq. (9), the survival probability p is derived as

$$p = \frac{2^{n_i - \frac{n_i}{n_o}|\mathcal{B}|} - 1}{2^{n_i} - 1}. \tag{11}$$

Similarly, p for Fig. 3c is given by

$$p = \frac{2^{n_o - \frac{n_o}{n_i}|\mathcal{B}|} - 1}{2^{n_o} - 1}. \tag{12}$$

Notice that Eq. (11) is the same as Eq. (9) when $n_i = n_o$. Furthermore, Fig. 3d corresponds to the probing attack against DES encryption by setting $n_i = 6$ and $n_o = 4$.

3 Case Study of the Probing Attack

This section applies the proposed model to various block ciphers and compares the model results with the experimental value, which is a result obtained from simulations using a detailed description of a cryptographic algorithm. The results are summarized in Table 2. The experimental value is the average of the values derived from one million simulations.

We performed a one-bit probing simulation with the proposed model by changing the key length from $n = 4, 5, 6, 7, 8, 9$. Corresponding to the key length, the survival probability p is calculated with Eq. (9) and sent to the model. We set the model's parameters as $q = 1, |\mathcal{B}| = 1$. Also, we conducted one-million simulations for the one-bit

Table 2: $E[R]$ to derive the secret key to seven-block ciphers by one-bit probing attack

Block ciphers	Attack target	Key length	$\lvert\mathcal{B}\rvert$	p	Previous model [1]	Proposed model	Experiments
PRINCE [24]	Fig. 3b	4	1	$4.67{\times}10^{-1}$	4.00	5.26	5.43
PRESENT [25]	Fig. 3b	4	1	$4.67{\times}10^{-1}$	4.00	5.26	5.43
FIDES [26]	Fig. 3b	5	1	$4.83{\times}10^{-1}$	5.00	6.28	5.98
FIDES [26]	Fig. 3b	6	1	4.92×10^{-1}	6.00	7.28	7.11
MISTY [27]	Fig. 3b	7	1	$4.96{\times}10^{-1}$	7.00	8.28	8.28
CAMELLIA [28]	Fig. 3b	8	1	$4.98{\times}10^{-1}$	8.00	9.28	9.31
MISTY [27]	Fig. 3b	9	1	$4.99{\times}10^{-1}$	9.00	10.28	10.50
AES	Fig. 3b	8	1	$4.98{\times}10^{-1}$	8.00	9.28	9.31
AES	Fig. 3b	8	2	2.47×10^{-1}	4.00	4.88	4.90
AES	Fig. 3b	8	3	$1.21{\times}10^{-1}$	3.00	3.41	3.42
AES	Fig. 3b	8	4	$5.88{\times}10^{-2}$	2.00	2.65	2.67
AES	Fig. 3b	8	5	$2.74{\times}10^{-2}$	2.00	2.18	2.19
AES	Fig. 3b	8	6	$1.12{\times}10^{-2}$	2.00	2.03	2.04
AES	Fig. 3b	8	7	$1.17{\times}10^{-2}$	2.00	2.00	2.00
AES	Fig. 3b	8	8	0	1.00	1.00	1.00
AES (IFA attack)	Fig. 3b	8	1	$4.98{\times}10^{-1}$	16.00	18.56	18.81
DES	Fig. 3c	6	1	$3.43{\times}10^{-1}$	6.00	5.27	8.19
DES	Fig. 3c	6	2	$1.11{\times}10^{-1}$	3.00	3.02	4.04
DES	Fig. 3c	6	3	$2.90{\times}10^{-2}$	2.00	2.19	2.83
DES	Fig. 3c	6	4	0	2.00	2.00	2.31

probing attacks against five block ciphers, PRINCE [24], PRESENT [25], FIDES with 5- and 6-bit S-box [26], MISTY with 7- and 9-bit S-box [27], and CAMELLIA [28] in addition to AES. From Table 2, assuming that the proposed model provides ideal results, the obtained experimental results indicate that all the five block ciphers have ideal physical attack resistance against the probing attacks.

3.1 Multi-bit probing attack against AES

Next, we explored the multi-bit probing attack on AES. We provided $n = 8$ and the corresponding survival probability p to the model. The model's parameters, $q = 1, |\mathcal{B}| = 2, 3, 4, 5, 6, 7, 8$, are set in this case. As can be seen from Table 2, the results show that AES is resistant to multi-bit probing attacks with ideal properties.

3.2 IFA attack against AES

We also evaluated the one-bit IFA attack resistance of AES, where $n = 1$ and $p = 127/255$ or 4.98×10^{-1}. Notice that we set $q = 1/2$ under an assumption that ineffective faults are injected with a probability of $1/2$. It is observed that AES also has an ideal resistance to the IFA attack.

3.3 Multi-bit probing attack against DES

We compared the number of attacks in the multi-bit probing attack on data encryption standard (DES) [23]. DES has an S-box with six-bit input and four-bit output, as

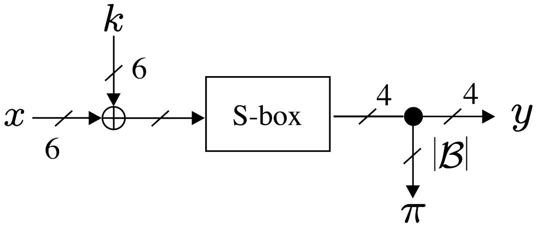


Fig. 6: Conceptual diagram for the probing attack against DES

shown in Fig. 6. Due to the DES structure, we adopted Eq. (11) derived from Fig. 3(d)[5]. The results show that DES has stronger probing attack resistance, for $|\mathcal{B}| = 1, 2, 3, 4$, compared to the ideal physical attack resistance by the proposed model.

4 Abstraction Model of the DFA Attack

The DFA attack has a characteristic similar to the probing attack in which the key is extracted with the information about the intermediate values and public information. Therefore, we can straightforwardly apply the proposed abstract model simply by providing n and by changing the survival probability p. The abstraction model for the DFA attack is shown in Fig. 7. The attacker injects intended faults during block cipher's encryption in which the differential property is set ideal. The attack target leaks ΔC, which is the difference between faulty and correct ciphertexts, as the leakage information.

[5] Note that the attack target illustrated in Fig. 3(c) is not corresponding to the probing attack attacking DES.

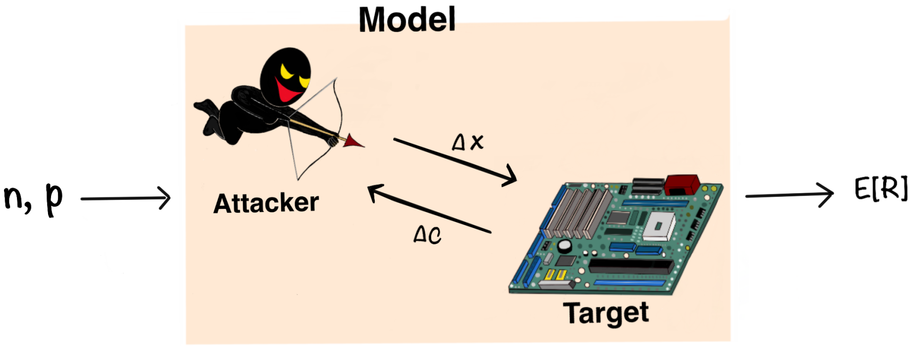


Fig. 7: Abstraction model for the DFA attack

4.1 Deriving the Number of Attack Trials

The DFA attack attempts to derive the secret key by inducing an error to an intermediate value x and obtains $\Delta y = x \oplus \Delta x$ from the ciphertext, where Δx denotes the input difference and Δy denotes the output difference. Depending on the attacker's ability to control faults, the range of Δx is limited, and the size of a set of possible Δx, $\Delta\mathcal{X}$, is determined. More precisely, a set of key candidates $\mathcal{K}$ for the DFA case is given by

$$\mathcal{K} = \{k \mid f^{-1}(y \oplus k) \oplus f^{-1}(y \oplus \Delta y \oplus k) \in \Delta\mathcal{X}\}, \tag{13}$$

which can be expressed like Eq. (3). The size of $\mathcal{K}$ depends on the difference distribution table (DDT) in an actual block cipher.

Similar to the case for the probing attacks, the proposed model eliminates false keys with a certain probability[6]. Here, we assume a cipher with the optimal attack resistance for the DFA attack, and employ an ideal DDT in which the number of key candidates α is derived by averaging the total number of key candidates from the function f, as shown in Fig. 8. The proposed ideal DDT is shown in Table 4. The averaged values α are uniformly distributed, excluding $\Delta x = 0$ and $\Delta y = 0$ because there are no faults at $\Delta x = 0$ and $\Delta y = 0$. By averaging the values in the difference distribution table, the differential probability of the function f is equal for any Δx and Δy. This is regarded as an optimal property for counteracting the DFA attacks because such an S-box satisfies the assumption that false key candidates are removed with equal probabilities. We also show the DDT of AES in Table 3 for S-box in Fig. 9.

4.2 Deriving Probability *p*

Subsequently, we explain how to derive p. For the DFA attack against the target with the ideal DDT, the total number of k that satisfies Eq. (13) for any Δx and Δy is given by

$$\sum_{\Delta x} \sum_{\Delta y} |\mathcal{K}| = (2^n)^2. \tag{14}$$

If we exclude the cases for $\Delta x = 0$ and $\Delta y = 0$, i.e., when the attacker can always inject a fault, as summarized in Table 4, we have

$$\alpha = \frac{(2^n)^2 - 2^n}{(2^n - 1)^2} = \frac{2^n}{2^n - 1} \ (\approx 1). \tag{15}$$

The attacker who obtains $\Delta\mathcal{X}$ and Δy can reduce the keyspace from $\alpha(2^n - 1)$ to $\alpha|\Delta\mathcal{X}|$, which is obtained from Table 4. Thus, we can use the following as the survival probability[7].

$$p = \frac{|\Delta\mathcal{X}|}{2^n - 1}. \tag{16}$$

[6] Algorithm 1 is skipped; we set $\epsilon = 2^n - 1$ and start with ϵ. We execute Alg. 2 but select a set of $\mathcal{K}^{(r)}$ by Eq. (13) at step 4.

[7] From the derived probability, we see that α is unnecessary in the model.

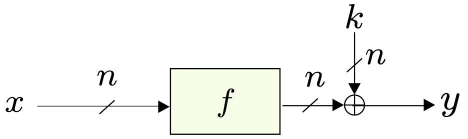


Fig. 8: Conceptual diagram for the DFA attack using ciphertext

Table 3: AES S-box differential distribution table

		Δy										
		0	1	2	3	4	5	6	7	8	$\cdots$	255
	0	256	0	0	0	0	0	0	0	0	$\cdots$	2
	1	0	2	0	0	2	0	2	0	2	$\cdots$	2
	2	0	0	0	2	2	2	2	2	0	$\cdots$	2
Δx	3	0	0	2	0	2	2	0	0	2	$\cdots$	2
	4	0	0	0	0	0	0	0	0	0	$\cdots$	2
	5	0	0	0	0	2	0	0	0	4	$\cdots$	2
	6	0	2	0	0	2	2	0	0	0	$\cdots$	2
	7	0	2	0	0	0	0	2	0	2	$\cdots$	2
	8	0	0	2	2	0	0	0	0	0	$\cdots$	2
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\cdots$	2
	255	0	2	2	2	0	0	0	2	0	$\cdots$	2

5 Case Study of the DFA Attack

We picked up the tenth, ninth, and eighth round DFA attacks on AES, for example. As in the case of the probing attacks, we provided the key length *n* and the survival probability *p* derived from α in the ideal DDT. Table 5 shows the result using the proposed model and the experimental results of simulations with detailed description of AES algorithms.

5.1 Case for the Tenth Round DFA Attack

We evaluated the DFA attack resistance for 1- to 8-bit random fault models at the tenth round of AES. We applied the encryption process at the tenth round to function f, as shown in Fig. 9. For example, $|\Delta X| = \binom{8}{2} = 26$ for the 2-bit fault model because we have 8-bit positions to inject 2-bit faults. Except for the 8-bit random fault case, the results

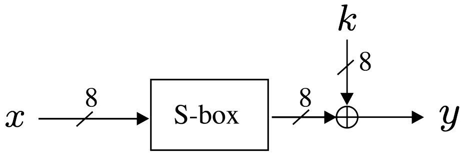


Fig. 9: Conceptual diagram for the DFA attack at the tenth round of AES

Table 4: Ideal differential distribution table

		Δy										
		0	1	2	3	4	5	6	7	8	$\cdots$	$2^n - 1$
	0	2^n	0	0	0	0	0	0	0	0	$\cdots$	0
	1	0	α	α	α	α	α	α	α	α	$\cdots$	α
	2	0	α	α	α	α	α	α	α	α	$\cdots$	α
Δx	3	0	α	α	α	α	α	α	α	α	$\cdots$	α
	4	0	α	α	α	α	α	α	α	α	$\cdots$	α
	5	0	α	α	α	α	α	α	α	α	$\cdots$	α
	6	0	α	α	α	α	α	α	α	α	$\cdots$	α
	7	0	α	α	α	α	α	α	α	α	$\cdots$	α
	8	0	α	α	α	α	α	α	α	α	$\cdots$	α
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\cdots$	α
	$2^n - 1$	0	α	α	α	α	α	α	α	α	$\cdots$	α

show that AES has ideal resistance judging from the proposed model. The larger $|\Delta\mathcal{X}|$ is, the smaller the difference between the number of key candidates in the actual DDT and the proposed DDT model. This characteristics is true for 1 to 7 bits, and hence the model and experimental results were well matched.

The possible reasons for the mismatch for the 8-bit fault attack are as follows. The secret key cannot be identified from the experiments using the actual AES S-box because the values in DDT are two or four as far as an appropriate fault is injected. Therefore, the false key will be left after a trial. On the other hand, DDT in the proposed model is filled with averaged values of $\alpha \approx 1$, and most of the attacks can identify the key in a single trial.

5.2 Case for the Ninth Round DFA Attack with 1-byte Fixed-Position Fault Model

We performed the DFA attack with the 1-byte fault model in the ninth round of AES. The attack target f is shown in Fig. 10. The input and output of function f are 32 bits. The 1-byte fixed position fault is injected before MixColumns, so we have $|\Delta\mathcal{X}| = 2^8 - 1 = 255$. Table 5 shows that the experimental result is slightly higher than the model value; however, we see that the model provides enough precision.

5.3 Case for the Ninth Round DFA Attack with 1- to 4-byte Fixed-Position Fault Model

We applied the fault models in [29], which inject 1- to 4-byte faults at a specific position in the ninth round. The attack target f is shown in Fig. 10, where the input and output are 32 and we have $|\Delta\mathcal{X}| = (2^8 - 1)^4 = 255^4$ bits, respectively. The results based on the proposed model is 1477, while the experimental result reported in [29] with the actual AES S-box is 1495.

We also calculated the number of attacks using the proposed model excluding the case for 1- to 3-byte faults, i.e., using a 4-byte fault only. In this case, $n = 32, p =$

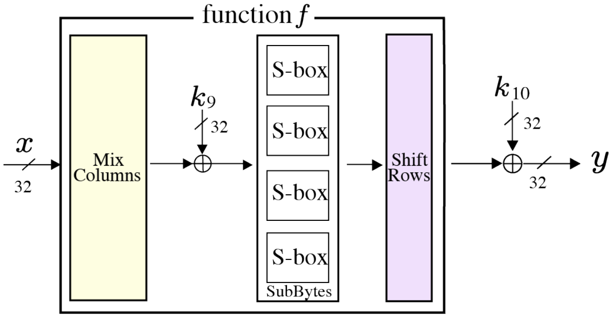


Fig. 10: Function f for 1- to 4-byte fault model in the ninth round of AES

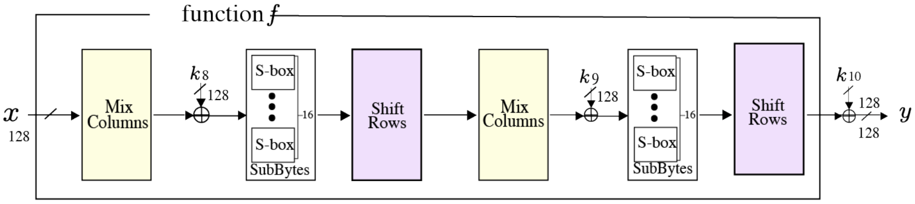


Fig. 11: Function f for the fault model in the eight round of AES

0.9847, and the result of the model value becomes 1477, which is close to the experimental value in [29]. Assuming that the proposed model is accurate enough for modeling the attacks on AES, it might be possible that only the ciphertext with a 4-byte fault is used in [29].

5.4 Case for the Eighth Round DFA Attack with 1-byte Random Fault Model

Finally, we evaluated the one-byte fault model for the eighth round of AES proposed by [30]. The attack target f is shown in Fig. 11. The input and output bytes of function f are 128 bits, and $|\Delta X| = 255^4 \times 4$ because a 1-byte falut is injected into four positions. We find that the AES has an ideal attack resistance from Table 5.

6 Discussion

Figure 12 depicts the model values with changing survival probability p and key length n and the experimental values of probing attack against seven block ciphers and the DFA attack against AES. Figure 13 focuses on the case of $n = 32$ [29] from Fig. 12.

Once the model parameters are determined, the curves can be used as a security indicator of physical attack resistance. The experimental values of FIDES, CAMELLIA, MISTY, and AES (both probing and DFA attacks) are on the model value curves. These ciphers have ideal properties for the assumed physical attack. The observed experimental values of PRINCE and PRESENT are slightly higher than the model value curve,

Table 5: $E[R]$ for the DFA attack on AES

Block cipher	Fault model	Key length	p	Proposed model	Experiment
AES	10th-round 1-bit random	8	$8/255$	2.230	2.240
AES	10th-round 2-bit random	8	$28/255$	3.282	3.286
AES	10th-round 3-bit random	8	$56/255$	4.540	4.542
AES	10th-round 4-bit random	8	$70/255$	5.233	5.235
AES	10th-round 5-bit random	8	$56/255$	4.540	4.542
AES	10th-round 6-bit random	8	$28/255$	3.282	3.285
AES	10th-round 7-bit random	8	$8/255$	2.230	2.230
AES	10th-round 8-bit	8	$1/255$	1.637	2.009
AES	9th-round 1-byte fixed	32	$255/(256^4 - 1)$	2.000	2.017
AES	9th-round 4-byte fixed (using 1- to 4-byte fault)	32	$255^4/(256^4 - 1)$	1454	-
AES	9th-round 4-byte fixed (using only 4-byte fault) [29]	32	9.847×10^{-1}	1477	1495
AES	8th-round 1-byte random	128	$(255^4 \times 4)/(256^{16} - 1)$	2.000	2.020[†]

[†] From [30], 98% of the experimental values are the results of two attacks to identify the secret key, and 2% are the results of three attacks to identify the secret key.

and hence, they have slightly strong physical attack resistance. The experimental values of the DES are far from the model curve, and hence we assume extremely strong against physical attacks. The experimental values excluding DES are on the model curves. This indicates that the proposed model performs well in evaluating the attack efficiency.

Assuming that there exists a trade-off between the physical attack and cryptanalysis, our results infer that AES is well-balanced. Nonetheless, DES seems to have a strong resistance to the probing attack and a weak resistance to cryptanalysis. The difficulty of the physical attack stems from the fact that the data are not mixed well by S-box.

7 Conclusion

In this study, we proposed an evaluation model to determine the number of attacks on block ciphers using the secret key length n and the survival probability p of the key candidates. Assuming the elements of the key candidates are chosen uniformly at random in the probing attack procedure; we modeled the number of attacks in an ideal cipher, where a false key survives with p. The difference in the attack efficiency between the model and experimental values of the one-bit probing attack on AES was approximately 0.3 %. Moreover, we obtained a more precise value compared with the previous model. We evaluated six block ciphers with a bijective function for the probe attacks and AES for the DFA attacks. The model values of these ciphers were very close to the experimental values, and these ciphers have the ideal properties that we assumed. Furthermore, we extended the proposed model to ciphers that have a surjective function and evaluated a multi-probing attack on a DES with a value significantly higher than the model value. Assuming that there is a trade-off in resistance between the physical attack and cryptanalysis, our results imply that AES has a balanced trade-off. Although DES has a strong resistance to physical attacks, it does not have enough resistance to cryptanalysis. Our future work includes the cases for more noisy data in the probing and DFA attacks. Also, we will apply the proposed model to other ciphers including stream ciphers.

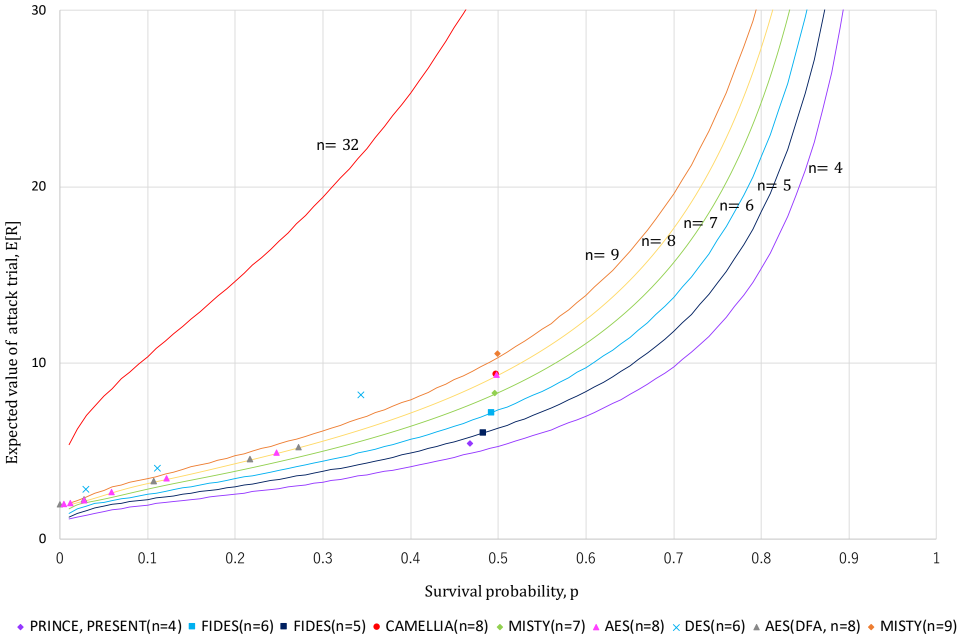


Fig. 12: Attack efficiency for survival probability p

Acknowledgments This study was supported by the Japan Society for the Promotion of Science (JSPS) and Grants-in-Aid for Scientific Research (KAKENHI), Grant Number JP18H05289.

References

1. K. Sakiyama, Y. Li, M. Iwamoto, and K. Ohta, "Information Theoretic Approach to Optimal Differential Fault Analysis," IEEE Transactions on Information Forensics and Security. 7 (1), pp.109–120, 2012.
2. N. Shoji, T. Sugawara, M. Iwamoto, and K. Sakiyama, "An Abstraction Model for 1-bit Probing Attack on Block Ciphers," IEEE 4th International Conference on Computer and Communication Systems (ICCCS), pp. 502–506, 2019.
3. H. Sugimoto, R. Hatano, N. Shoji, and K. Sakiyama, "Validating the DFA Attack Resistance of AES (Short Paper)," In: Benzekri A., Barbeau M., Gong G., Laborde R., Garcia-Alfaro J. (eds) Foundations and Practice of Security. FPS 2019. Lecture Notes in Computer Science, vol. 12056, pp. 371–378, 2019.
4. National Institute of Standards and Technology, "FIPS 197: Announcing the Advanced Encryption Standard (AES)," `http://nvlpubs.nist.gov/`.
5. A. Biryukov, D. Khovratovich and I. Nikolić, " Distinguisher and Related-Key Attack on the Full AES-256," In: Halevi, S. (ed.) CRYPTO 2009. LNCS, vol. 5677, pp. 231–249. Springer, Heidelberg (2009)

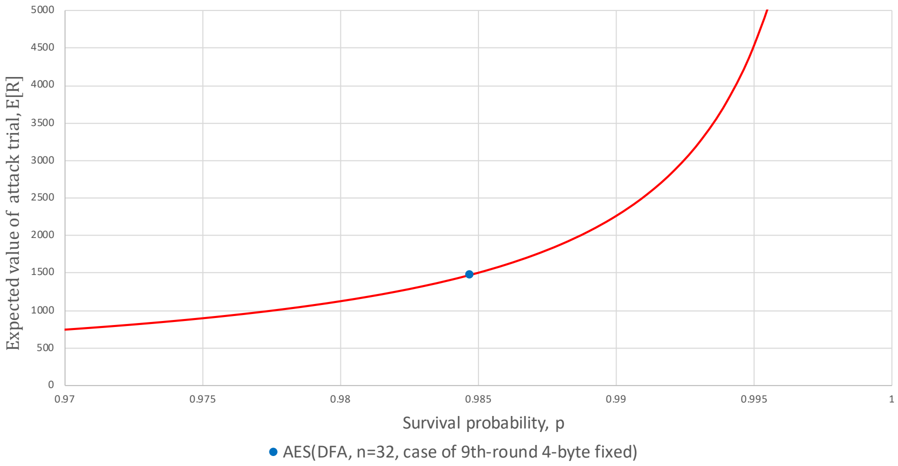


Fig. 13: Attack efficiency for survival probability p focusing on key length $n = 32$ (enlarged view)

6. A. Barenghi, L. Breveglieri, I. Koren, G. Pelosi, and F. Regazzoni, “Countermeasures against fault attacks on software implemented AES: effectiveness and cost,” In Proceedings of the 5th Workshop on Embedded Systems Security (2010), ACM, p. 7–16.
7. A. Bogdanov, D. Khovratovich, C. Rechberger, “Biclique Cryptanalysis of the Full AES,” In: Lee D.H., Wang X. (eds) Advances in Cryptology - ASIACRYPT 2011. ASIACRYPT 2011. Lecture Notes in Computer Science, vol 7073. Springer, Berlin, Heidelberg.
8. B. Johannes and S. -J. Pierre, “Fault Based Cryptanalysis of the Advanced Encryption Standard (AES),” IACR Cryptology ePrint Archive. 2002. 75.
9. P. Dusart, G. Letourneux and O. Vivolo, “Differential Fault Analysis on A.E.S,” 2016 IEEE International Symposium on Circuits and Systems (ISCAS), 2016, pp. 554–557.
10. C. Giraud, “DFA on AES,” in Proc. International Conference on the Advanced Encryption Standard, vol. 3373, pp. 27–41, 2005.
11. H. Handschuh, P. Paillier, and J. Stern, “Probing Attacks on Tamper-Resistant Devices,” CHES 1999, LNCS 1717, Springer-Verlag, pp. 303–315, 1999.
12. A. Moradi, M. T. M. Shalmani, M. Salmasizadeh , “A Generalized Method of Differential Fault Attack Against AES Cryptosystem,” In: Goubin L., Matsui M. (eds) Cryptographic Hardware and Embedded Systems - CHES 2006. CHES 2006. Lecture Notes in Computer Science, vol 4249. Springer, Berlin, Heidelberg.
13. S. M. Yen and M. Joye, “Checking before output may not be enough against fault-based cryptanalysis," in IEEE Transactions on Computers, vol. 49, no. 9, pp. 967–970, Sept. 2000.
14. E. Biham, A. Shamir (1997), “Differential fault analysis of secret key cryptosystems,” in Proc., Advances in Cryptology - CRYPTO ’97, LNCS 1294, pp. 513–525, 1997.
15. P. Kocher, J. Jaffe, and B. Jun, “Differential Power Analysis,” In Proc. Advances in Cryptology - CRYPTO ’99, 19th Annual International Cryptology Conference, pp.388–397, 1999.
16. P. Kocher, R. Lee, G. McGraw, A. Raghunathan and S. Ravi, "Security as a new dimension in embedded system design," Proceedings. 41st Design Automation Conference, 2004., 2004, pp. 753–760.
17. P. C. Kocher, “Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems.” In Proc. Int. Cryptology Conf. (CRYPTO’96), pages 104–113, 1996.

18. E. Brier, C. Clavier, F. Olivier, “Correlation Power Analysis with a Leakage Model,” In: Joye M., Quisquater JJ. (eds) Cryptographic Hardware and Embedded Systems - CHES 2004. CHES 2004. Lecture Notes in Computer Science, vol 3156. Springer, Berlin, Heidelberg.
19. S. Chari, J. R. Rao, P. Rohatgi, “Template Attacks,” In: Kaliski B.S., Koç .K., Paar C. (eds) Cryptographic Hardware and Embedded Systems - CHES 2002. CHES 2002. Lecture Notes in Computer Science, vol 2523. Springer, Berlin, Heidelberg.
20. E. Biham and A. Shamir, “Differential fault analysis of secret key cryptosystems,” In: Kaliski B.S. (eds) Advances in Cryptology — CRYPTO ’97. CRYPTO 1997. Lecture Notes in Computer Science, vol 1294.
21. C. Clavier and A. Wurcker, “Reverse Engineering of a Secret AES-like Cipher by Ineffective Fault Analysis,” In Proc. Workshop on Fault Diagnosis and Tolerance in Cryptography (FDTC 2013), pp. 119–128, 2013.
22. T. Sugawara, N. Shoji, K. Sakiyama, K. Matsuda, N. Miura, and M. Nagata, “Exploiting Bitflip Detector for Non-Invasive Probing and its Application to Ineffective Fault Analysis,” In Proc. Workshop on Fault Diagnosis and Tolerance in Cryptography (FDTC 2017), pp. 49–56, 2017.
23. National Institute of Standards and Technology, “Data Encryption Standard (DES),” https://csrc.nist.gov/csrc/media/publications/fips/46/3/archive/1999-10-25/documents/fips46-3.pdf, 1999.
24. J. Borghoff, A. Canteaut, T. Guneysu, E. B. Kavun, M. Knezevic, L. R. Knudsen, G. Leander, V. Nikov, C. Paar, C. Rechberger, P. Rombouts, S. S. Thomsen, and T. Yalcin, “PRINCE - A Low Latency Block Cipher for Pervasive Computing Applicationxs,” In Proc. Int. International Conference on the Theory and Application of Cryptology and Information Security (ASIACRYPT 2012), LNCS 7658, pp. 208–225, 2012.
25. A. Bogdanov, L. R. Knudsen, G. Leander, C. Paar, A. Poschmann, M. J. B. Robshaw, Y. Seurin, and C. Vikkelsoe, “PRESENT: An ultralightweight block cipher,” In Proc. Int. Workshop on Cryptographic Hardware and Embedded Systems (CHES 2007), LNCS 4727, pp. 450–466, 2007.
26. B. Bilgin, A. Bogdanov, M. Knezevic, F. Mendel, and Q. Wang, “FIDES: Lightweight Authenticated Cipher with Side-Channel Resistance for Constrained Hardware,” In Proc. Int. Workshop on Cryptographic Hardware and Embedded Systems (CHES 2013), LNCS 8086, pp. 142–158, 2013.
27. M. Matsui, “New block encryption algorithm MISTY,” In Proc. Fast Software Encryption (FSE 1997), LNCS 1267, pp. 54–68, 1997.
28. K. Aoki, T. Ichikawa, M. Kanda, M. Matsui, S. Moriai, J. Nakajima, and T. Tokita, “Camellia: a 128-bit block cipher Suitable for Multiple Platforms - Design and Analysis,” In Proc. Selected Areas in Cryptography (SAC 2000), LNCS 2012, pp. 39–56, 2001.
29. A. Moradi, M. T. M. Shalmani, and M. Salmasizadeh, “A generalized method of differential fault attack against AES cryptosystem,” In Proc. CHES’06, pp. 91–100, 2006.
30. G. Piret and J. J. Quisquater, “A Differential Fault Attack Technique against SPN Structures, with Application to the AES and Khazad,” Workshop on Cryptographic Hardware and Embedded Systems (CHES 2003), Springer-Verlag, 2003, pp. 77–88.