A unified formalism for side-channel and fault attacks on cryptographic circuits
Bruno Robisson, Hélène Le Bouder, Jean-Max Dutertre, Assia Tria

To cite this version:

HAL Id: emse-00742510
https://hal-emse.ccsd.cnrs.fr/emse-00742510
Submitted on 24 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A unified formalism for side-channel and fault attacks on cryptographic circuits

Bruno Robisson*, Hélène Le Bouder†, Jean-Max Dutertre‡ and Assia Tria*
*CEA-LETI †École Nationale Supérieure des Mines de Saint-Étienne
‡Département Systèmes et Architectures Sécurisés, Gardanne, France
{bruno.robianon, assia.tria}@cea.fr {lebouder, dutertre}@emse.fr

Abstract—Security is a key component for information technologies and communication. Security is a very large research area involved in the whole information technology, related to both hardware and software. This paper focuses on hardware security, and more specifically on hardware cryptanalysis whose aim is to extract confidential information (such as encryption keys) from cryptographic circuits. Many physical cryptanalysis techniques have been proposed in the last ten years but they always belong to one of those very distinct categories: fault and side channel attacks. In this article, a formal link between these two categories is proposed. To the best of our knowledge, this is the first time that a wide class of attacks is described in such a generic manner.


INTRODUCTION

Security is a key component for information technologies and communication. Among the security threats, a very important one is certainly due to vulnerabilities of the integrated circuits that implement cryptographic algorithms to ensure confidentiality, authentication or data integrity (such as smartcards). With the access to one of these circuits, the attacker tries either to reconstruct the functionality of the circuit (reverse engineering) or to recover cryptographic materials when the cryptographic algorithm is known (physical or hardware cryptanalysis). Both threats share a set of techniques. The first one, called side channel attacks, consists in observing some physical characteristics which are modified during the circuit’s computation. The second technique, called fault attacks, consists in disrupting the circuit’s behavior. The third one consists in getting information about the chip design by direct inspection of its structure. This inspection may be performed by using any kind of imaging techniques or by using destructive means such as abrasion, chemical etching or focused ion beam. Many protections have been proposed to improve security against those types of attacks. But every protection against physical cryptanalysis usually lacks of genericity and is usually very specific to one precise type of attack.

Even if physical cryptanalysis methods seem very different at first sight, several works has been proposed to described them with a commun formalism [MR03], [MOS09]. But these works only cover side channel attacks. This paper describes a formalism which is commun to both fault and side channel attacks. The short-term aim of this work is to define “new” attacks as new combinations of these concepts and algorithms. Then, we plan to provide efficient and modular implementation of these attacks. The long-term aim of this work is to merge the advantages of attack-specific protections to enable a more generic set of countermeasures.

In the first part of this article, we shortly describe side channel and fault attacks and give an informal explanation of the principles which are common in these two types of attack. The cryptographic algorithm used as example, \textit{AES}128 is also described. In the second part of the paper, we present a small set of definitions and the algorithms we use to extract information from physical attacks. Finally, we show that our proposed formalism easily fits with several representative examples.

1. SCOPE OF THE WORK

A. Physical cryptography

This paper deals with physical cryptanalysis of both software or hardware implementation of cryptographic algorithms (such as \textit{AES}128 described in I-C). These algorithms encrypt a plaintext with a key to obtain a ciphertext. Physical cryptanalysis is mainly based on the two following techniques:

1) Side channel: Side channel analysis exploits the fact that some physical values or “side channels” (such as power consumption, electromagnetic radiation or computation time) of a circuit depend on the key value [Koc96], [KJ99], [SMP07]. A more rough technique consists in measuring directly some key-related internal computations by using micro-probing techniques. There are two subcategories of key recovering techniques based on side channel measurements. The first one, called DPA-like, consists in building a set of mathematical models (i.e. mathematical formulate) from a priori knowledge about the circuit. Each model is associated with an hypothesis on the value of the key. Then, the models are compared with measurements. The model which matches the best with measurements is generally associated with the right key hypothesis. The second kind of side channel attacks needs a profiling step on another circuit. This circuit is supposed to be identical to the target and the attacker is supposed to be able to set the key value. In this case, the profiling step is used either to improve the model a priori (stochastic attacks) or to build a statistical model only based on measurements (template attacks).
2) Faults: Fault attacks consist in disrupting the circuit’s behavior. Their aim is to alter the correct progress of the algorithm. Faults are injected into the device using various means such as laser, clock glitches, spikes on the voltage supply or electromagnetic perturbations. A more rough technique consists in modifying the circuit’s operation by modifying internal computation through micro-probes, or even modifying the circuit itself by using focused ion beam. There are three subcategories of key recovering techniques that use the results of faults attacks. Algorithm modifications consist either in reducing the ciphering complexity of the cryptographic algorithm [CT05] or in bypassing hardware or software protections. Differential Fault Attack (DFA), originally described in [BDL97], [BS97] and in enhanced in [PQ03], [MSS06], [Muk09], [TM09], [Gir05], [RLK11], consists in retrieving the key by comparing the correct ciphertexts with faulty ones. A detailed comparison of DFA schemes against AES, for example, is given in [SLIO12]. To perform the third kind of fault attacks, called safe-error attacks [YJ00], [RM07], [LSG+] 10, the attacker does not necessarily need pairs of correct and faulty ciphertexts but only some information about the chip’s behavior.

B. Principle of the considered class of physical attacks

This paper does not deal yet with attacks which aim to modify the algorithms (by using fault or circuit modifications). It focuses on the different kinds of side channel attacks (DPA-like, stochastic or template), DFA and safe-error attacks.

For such attacks, physical access to the circuit gives non-intrusive information about the internal computations of the cryptographic algorithm. Thanks to these pieces of information, the attacker is able to apply a divide and conquer approach. Instead of retrieving the whole key in one step (typically constituted of one hundred to one thousand bits), the attacker performs the same elementary attack step several times. Every elementary attack enables him to retrieve a partial part of the key, called a partial key. At the end, the attacker is able to rebuild the whole key. We claim that the elementary steps of the attacks share the following method:

- Launch cryptographic operations on the target. Each launch is called an experiment. Different kinds of experimental data such as power, electromagnetic emission (EM), inputs or outputs, detailed in II-A, are recorded during these experiments.
- Make sure that some of these measurements are related to some others through the partial key. A more formal definition of “relations” between measurements and the key is provided in II-B.
- Build models that explain the relationship between these measurements. These models are parametrized with one value of partial key. These models are mathematical formule determined either by a priori knowledge on the chip or by using additional measurements. Examples of models used for attacks are described precisely in II-C.
- Compare the models to the measurements, by using the algorithms described in II-D. The best-fitted models with the measurements are generally models associated with the partial key.

To illustrate our formalism, we use as main example the AES crypto-algorithm shortly described above.

C. Example of cryptographic algorithm : AES_{128}

AES is a standard established by the NIST [NIS01] for symmetric key cryptography. In this paper, we focus on the 128-bit-key version of the AES (represented in Figure 1). This algorithm ciphers a 128-bit long data (called plaintext or input) by using a 128-bit long key to obtain a 128-bit long data (called ciphertext or output). The encryption consists first in transforming the input data into a two-dimensional array of bytes, called the state. Then, after a preliminary XOR between the input and the key, AES_{128} executes 10 times a function (called round) that operates on the state. The operations used during these rounds are the following ones:

- SubBytes is a non-linear transform working independently on individual bytes of the State. The result of this operation at round i is noted round[i].s.box
- ShiftRows is a rotation operation on each row of the State. The result of this operation at round i is noted round[i].s.row
- MixColumns is a linear matrix multiplication working on each column of the State. The result of this operation at round i is noted round[i].m.col
- AddRoundKey is a byte-wise XOR between the State and a value k.sch[i], computed from the key (according to a transform called Key Expansion which is not detailed in this paper). The result of this operation is the start of the next round noted Round[i + 1].start.

The rounds are identical, except for the last one in which the MixColumns operation is skipped. At the end of the ciphering operations, the state is copied to the output as depicted in Figure 1.

II. PROPOSED FRAMEWORK

A. Attacker’s environment

1) Observable data: As explained above, physical attacks are based on experiments, i.e. the launches of cryptographic operations.
operations on the target with or without modifying its behavior. During these experiments, the attacker is able to directly collect data such as digital values or analog ones. This set of values are further called observable data, or observables. Some observables, called stimuli, are used to launch the cryptographic operations and generally change at each experiment: plaintexts, whether there are perturbations or not, etc. Some observables also define the experimental setup and generally do not change during the experiments: kind of probes, gain and bandwidth of the amplifier, oscilloscope type, key value, etc. At last, some observables are the reactions or responses of the circuit according to the stimuli and the experimental setup. There are numerous reactions which can be recorded by the attacker:

- the EM, power traces or the signal provided by a micro-probe.
- the ciphertexts obtained without or with perturbations (faulty ciphertexts).
- the behavior (i.e. the normal or abnormal working) of the circuit in case of fault. This behavior may be determined, for example, by comparing the correct and faulty ciphering, by detecting the start of an alarm or not, the raise time of the alarm or by detecting a more or less premature stop in computation.
- the computation time. One way to measure this computation time consists in reducing the clock period step by step (on a specific clock cycle) and analyzing the circuit’s behavior. The step in which the behavior of the circuit switches from normal to abnormal is considered as an indicator for the computation time. This particular step is called “fault intensity” in [LSG+ 10].

Note that we also consider that any combination of observables or any observable obtained by applying any signal processing technique (filtering, temporal to frequency transformation, etc.) is also considered as an observable.

2) Hidden data: At the opposite of observables, the execution of a cryptographic algorithm involves data that cannot be directly measured by the attacker. This piece of data is called hidden data or internal data in the following paragraphs.

3) “Border” between observables and hidden data: The border between observables and hidden data depends on the target and the attacker’s means (i.e. laboratory equipment and threat scenario). In the following paragraphs, without loss of generality, the hidden data are the internal variables as defined in the standard (from round[1].start included to round[10].s_row and every variable computed in the key schedule).

4) Dimension of observables: The measurements of observables may naturally be grouped in vectors or matrices of several dimensions. The power consumption, recorded with an oscilloscope, varies with the time. A power record (or trace) is thus a set of measurements which is grouped in a vector (dimension 1). An EM record, recorded with an EM probe moving in a plan above the chip, is naturally grouped in a 3-dimension matrix (time and X-Y location of the probe). For the sake of simplicity, we consider that a reaction is a scalar value (its dimension is equal to zero).

B. Relationships between variables

As they are computed by or emanate from the same circuit, hidden data and observables are physically related to each other. These relationships can be described by a mathematical expression (for example between the plaintext, the key and the ciphertext in a nominal environment), some others cannot (for example between the plaintext, the key and the power consumption). But even if there is no mathematical expression for these relationships, the authors in [MR03], [MOS09] have introduced the notion of leakage function for side channel attacks. These functions formalize the relationship between some internal data (which are generally binary values) and some analog signals (such as power, EM, etc.). To extend this notion to fault attacks, we also consider “error leakage” functions. These functions formalize the modification by a fault of an internal data. As the internal data considered in physical attacks are generally for AES 8-bit values, each endomorphism of a set of 8 bits can be considered as a possible error function.

An exploitable relationship is a relationship between observables. This kind of relationship can be used to perform an attack by using a divide and conquer approach. Thus, such a relationship needs to involve only a partial key. Note that a relationship may be exploitable even without a mathematical expression. Examples of widely used exploitable relationships are reported in Table I. The relations only involves each byte of the state but for clarity are written, in this Table, for the whole state. In the following paragraphs, we note $p = R(F, k, O)$ the relation that links the reactions $p$ and a set of observables $O$ according to the value $k$ of the key and a set of leakage functions $F$.

C. Relationship models

In the exploitable relationships described above, there is at least one unknown value: the partial key (or guess). So, the attacker has to create as many models as there distinct values of the guess key (i.e. 256 possibilities for an 8-bit guess key). For some relationships, other hidden data and one or several leakage functions also have to be guessed. It is impossible to test every possible leakage function because of their tremendous (even infinite in the case of side channel functions) number. The heuristics commonly used to choose one leakage function or a set of leakage functions are based on measurements or on a priori knowledge about the circuit and the measurement setup. Only the last case which concerns DPA-like, DFA and safe-error attacks, is described here. In this case, the models are built on the following small set of functions but other constructions are proposed for example in [DPRS11].

1) Elementary functions: A byte $B$ is a set of 8 bits noted \{b_7, b_6, b_5, b_4, b_3, b_2, b_1, b_0\}. The set with every bit set to one (resp. zero) is noted 1 (resp. 0). We match it with $B$ the binary value $B_b = b_7b_6b_5b_4b_3b_2b_1b_0$ but for the sake of simplicity, $B$ and $B_b$ will be used indifferently. The decimal value associated with $B$ is noted $B_d$ (for this conversion, $b_0$ is the less significant bit).
a) **Bit level functions:** The logical operations AND (symbol $\&$), OR (symbol $|$) and XOR (symbol $\oplus$) defined on a byte are considered bitwise.

b) **Weighted sum:** Let $B$ be a set of $N$ bits and $\Omega$ be a set of $N$ integer, real or binary numbers. The function $\sum_{\Omega} b$ is the sum

$$\sum_{\Omega}(B) = b_N \ast \omega_N + b_{N-1} \ast \omega_{N-1} + \ldots + b_1 \ast \omega_1 + b_0 \ast \omega_0$$

c) **Select subset (or restriction) of bits:** Let $B$ (resp. $\Omega$) be a set of bits $b_i$ (resp. $\omega_i$). The set of bits $b_i$ of such that $\omega_i$ of $\Omega$ is equal to “1”, is noted $R_2(B)$. Example: if $\Omega = \{0, 0, 1, 0, 0, 0, 0, 0\}$, we have $\Omega_d = 36$ and the restriction of $B$ by $R_2(B)$ is the set of bits $\{b_1, b_2\}$ of $B$. Note that if $B$ is constituted of $N$ values, there are $2^N - 1$ possible restrictions. A restriction is said monotib when only one bit of the byte is considered. So, the monotib restrictions are $R_1, R_2, R_4, \ldots, R_{128}$.

d) **Subset of bits equal to:** Let $\Omega$ and $B$ be two sets of $N$ bits. Let $M$ be the number of the bits of the restriction of $R_1(B)$ and $\alpha$ a set of $M$ bits. We note $R_1(B) = \alpha$ the function that returns 1 when, for any bit $i$ of $R_1(B)$ and $\alpha$, we have $R_1(B) = \alpha$ and returns 0 otherwise.

d) **Bit flip:** A fault may invert a set of bits. The function which models the bit flip is the bitwise XOR between $B$ and $\Omega$ with $\Omega_d \in \{1, \ldots, 255\}$. When all the bits of the bitflip $\Omega$ are equal to 1 (resp. 0), all the bits are inverted (resp. unchanged).

b) **Set and reset:** A fault may set a set of bits to specific values. The function which models the reset (resp. set) is the bitwise AND (resp. OR) between $B$ and $\Omega$ with $\Omega_d \in \{1, \ldots, 255\}$. A reset with $\Omega_d = 0$ returns a set of 0. A set with $\Omega_d = 255$ returns a set of 1.

c) **Behavior:** When a fault sets a set of bits $R_2(B)$ of $B$ to some value $\alpha$, the circuit behavior is normal only when the perturbed result is expected to be equal to $\alpha$ and is abnormal otherwise. Such a behavior may be described by using the function $R_2(B) = \alpha$.

d) **Setup time violation:** When the faults are created by violating the setup time of latches, these faults may impact first the bytes which have the higher Hamming weight or may impact first only one bit. In the first case, chip’s behavior may be described by a Hamming weight and in the second case by the function that selects one bit $R_2$ with $i \in \{1, 2, 4, \ldots, 128\}$.

D. Key retrieving algorithm

In this part, we explain how to deal with exploitable relationships and measurements in order to discover a partial key. The first input of the algorithm 1 is a relationship between a set of observables and reactions. The models of this relationship is parameterized by the set of all the possible partial keys $K$ and a set of models $F$. This set of models formalizes the knowledge of the attacker about the circuit and the measurement setup. Two cases arise: in the first case, the attacker considers that one of these models is more probable than the others but he doesn’t known which one is the good one (case “dist(F)=several”). In the second case, he considers that all the models are possible and are equiprobable (case “dist(F)=one”). The second input of the algorithm is the method to select the good or the bad key from a comparison between the prediction and the measurements. The different comparison methods are described in the algorithm 2. Roughly, when predictions and measurements are both discrete variables and when the “noise” in the measurements is considered negligible (it is typically the case for DFA), a siev algorithm can be used. But when taking noise into account or when either predictions or measurements are continuous variables, some statistical distinguishers are preferred (a counting algorithm or a statistical similarity measurement). Mutual information (MIA) [GBT07], Pearson correlation [BC003], distance of means (DoM) [KJJ99] or principal component (PCA)[SNG+10] are possible examples of such statistical similarity measurement tools.

Once these parameters are chosen, the measurements are processed in an incremental way, i.e the attacker tries to
recover this partial key with a reduced set of measurements. If no partial key can be distinguished, then the attack is performed with other measurements. He proceeds in this way until the partial key is recovered.

Algorithm 1 Algorithm to distinguish the key (and the model) from the relationship $R$

Require: $p = R(K, F, O)$: A relationship with:

- $p$: A reaction
- $k \in K = \{k_0, \ldots, k_{2^n-1}\}$: a set of $2^n$ partial keys
- $f \in F = \{f_1, \ldots, f_m\}$: a set of $m$ leakage functions
- $o \in O = \{o_1, \ldots, o_n\}$: a set of $n$ observable
- $dist(F) \in \{\text{one, several}\}$ the hypothesis on the distribution of leakage functions
- $Max$: Maximum number of measurements
- $sim \in \{\text{SIEVE, COUNT, CORR}\}$: an algorithm that computes a similarity

Ensure: $k_r$: most “probable” key hypothesis

$O$: a matrix (size $Max \times n$) The observables

$p$: a vector (size $Max$) The reaction

$Pred_p$: a matrix (size $2^n \times m \times Max$) Prediction of the reaction $p$

$c$: a vector (size $2^n$)

Search ← TRUE; $e ← 0$; $k_r ← \text{NULL}$

while Search == TRUE and $e < Max$ do

- Perform the $e^{th}$ experiments
- Store the corresponding measurements in $o(e, 1:n)$
- Store the corresponding reaction in $p(e)$

for $i = 0$ to $2^n - 1$ do

for $j = 1$ to $m$ do

$Pred_p(i, j, e) = f_j(k_i, o(e, 1:n))$

end for

end for

for $i = 0$ to $2^n - 1$ do

if $dist(F) == \text{one}$ then

$ct ← 0$

for $j = 1$ to $m$ do

$ct(j) ← sim(p(1:e), Pred_p(i, j, 1:e))$

end for

$e(i) ← MAX(ct(j))$

else

$e(i) ← sim(p(1:e), Pred_p(i, 1:m, 1:e))$

end if

end for

if $\exists \alpha$ such that $\forall \beta$, $c(\alpha) > c(\beta)$ then

$k_r ← k_i$

Search ← FALSE

else

$e ← e + 1$

end if

end while

Algorithm 2 Algorithm to compute the similarity

Require: $Meas$ a vector of measurements (size $e$)

$Pred$ a matrix of measurements (size $m \times e$)

$sim \in \{\text{SIEVE, COUNT, CORR}\}$

$corr \in \{\text{DoM, CPA, MIA, PCA, \ldots statistical tool}\}$

Ensure: $e$

switch $sim$

case $\text{SIEVE}$

$c ← 1$

for $i = 1$ to $e$ do

$l ← 0$

for $j = 1$ to $m$ do

$l ← l + 1$

if $Pred(j, i) == Meas(i)$ then

$c ← c + 1$

end if

end for

end case

end for

case $\text{COUNT}$

$c ← 0$

for $i = 1$ to $e$ do

for $j = 1$ to $m$ do

if $Pred(j, i) == Meas(i)$ then

$c ← c + 1$

end if

end for

end case

case $\text{CORR}$

$index ← 1$

for $i = 1$ to $e$ do

for $j = 1$ to $m$ do

$ePred(index) ← Pred(j, i)$

$eMeas(index) ← Meas(i)$

$index ← index + 1$

end for

$c ← corr(ePred, eMeas)$

end for

end case

end for

end switch

III. APPLICATIONS

In order to show that our formalism covers a wide class of attacks, the relationships (described in II-B) and the parameters for the algorithms (described in II-D) are reported in Table 2. This paper presents a formalism which is common for a wide class of physical attacks. It highlights that, despite their differences in term of experimental setup, all these attacks share the main underlying set of concepts and algorithms. This paper also shows that this set is relatively small. It enable to

Conclusion

If for different attacks proposed in the literature. A simple but new attack is also proposed. It consists in probing the signal on one wire of a 8-bit data bus where the outputs of the Sboxes are supposed to pass through. Each leakage model corresponds to one bit of these outputs. If the attacker is able to temporally distinguish between the outputs of all the Sboxes, every method described above used to compute the similarity works. Otherwise, a method based on correlation is preferred. The attack scheme shows that the attacker may retrieve every bit of the key by probing only one single wire.
plan to define "new" attacks as new combinations of these concepts and algorithms. It will also enable to provide efficient and modular implementations of these attacks. Further work will consist in including in the proposed formalism template attacks, collision or algebraic attacks. Another future work will consist in integrating in our scheme the higher-order attacks.

REFERENCES


