Precise methods for distinguishing S-box faults in laser injection attacks

Fan Zhang, Yiran Zhang, Huilong Jiang, Xiang Zhu, Feng Lin✉ and Kui Ren

Fault attack is a type of active attack which retrieves the secret key by injecting computational faults. Laser is one of the most common fault injection methods. When attacking substitution-permutation networks-ciphers, S-box is often chosen as the laser injection target. However, the adversary has to know whether the S-box is corrupted or not after each injection, which was difficult in real-world attacks. Two analysing methods to distinguish the S-box faults for different ciphers is proposed in this Letter. A laser-based physical experiment is carried out to verify the authors' methods on Advanced Encryption Standard and PRESENT ciphers on an ATmega163L microcontroller. Experimental results show that their methods can precisely distinguish S-box faults merely using dozens of ciphertexts.

Introduction: In symmetric encryption, many block ciphers such as Advanced Encryption Standard (AES) [1] and PRESENT [2] use *substitution-permutation networks* (SPN) design. A general SPN cipher takes the plaintext and key as inputs, and iterates several rounds to produce the ciphertext. A typical round of operations will include: (i) a non-linear layer where a block of input (typically 4 or 8 bits) is substituted according to the so-called S-box; (ii) a linear layer which permutes all the bits of the input; (iii) an addition which XOR the round key with the input. Most designs of mainstream SPN ciphers are proved to be secure at the theoretical level. However, the security of their implementations is threatened by the recent fault attack (FA).

FA [3] is a type of active attack which retrieves the secret key by injecting computational faults. It is quite powerful and efficient when breaking the physical implementations of many ciphers. A typical FA consists of two phases. (i) *Fault Injection*. The adversary disturbs the operation of the target device in this phase, thus the produced ciphertext will become faulty. (ii) *Fault Analysis*. He analyses the faulty ciphertexts to recover the key. In the first phase, the laser is one of the most common injection methods due to its high precision [4]. It can modify a byte, or even precisely flip a bit of data stored in registers, RAM, flash, etc.

In practice, two problems need to be solved when conducting laser injection attacks. One is that most of the FAs need to modify an intermediate state, thus the shooting time of laser must be highly synchronised with the encryption. This problem can be solved by *persistent FA* (PFA) proposed in [5]. PFA explores the so-called persistent fault which is injected and persistent in S-box. The other problem is that: to disturb the encryption, the physical cell with target data must be focused by the laser. When injecting the faults to some special area that is targeted (e.g. S-box), this problem is getting much more challenging since the adversary cannot distinguish the S-box faults precisely, i.e. to identify those injections which do cause a fault in S-box. In this Letter, we propose two different methods for distinguishing S-box faults to solve this problem. Then a physical experiment is carried out to test our methods on both AES and PRESENT implementations. The result shows that we can distinguish all S-box faults with dozens of ciphertexts only.

Distinguishing S-box faults: At the very beginning, the adversary should roughly identify the physical area of the storage, which may contain the S-box. Then he shoots a laser inside it. A laser pulse may cause three different types of results: no effect, data-flip or latch-up. A laser with lower energy will not affect the encryption. A laser with higher energy will cause latch-up, where the device has to be reset before working again. Only the laser with suitable energy may cause data-flip. As to the influence on the output ciphertexts, if no faults are injected by the laser, the ciphertexts will always be correct. For data flips, only some areas, such as S-box, round constant etc. will lead to erroneous outputs. If latch-up happens, the communication will fail, thus no ciphertexts can be collected. The goal of this Letter is to distinguish S-box faults precisely, and meanwhile, to reduce the number of ciphertexts that need to be collected. Two efficient methods are proposed.

Method A: For an S-box fault, the ciphertext will become faulty only if the S-box element is accessed during the encryption so that eventually only a certain part of all ciphertexts are faulty. Meanwhile, all of the ciphertexts will become faulty if other important data are corrupted, and none of the ciphertexts will become faulty if the injection has no effect or corrupted some irrelevant data. Considering an SPN cipher with L words and each word consists of B bits (i.e. the S-box has 2^B elements). For simplicity, we assume only one of those S-box elements is corrupted, and the input of S-box is uniformly distributed (which is promised by the property of block ciphers). Assuming the S-box is accessed T times during the encryption, P_c, the probability that faulty S-box element is not accessed, can be computed as

$$P_c = \left(1 - \frac{1}{2^B}\right)^T \tag{1}$$

It is equivalent to the probability that the ciphertext remains correct when S-box faults exist. For example, AES-128 uses an 8-bit S-box, which is accessed $16 \times 10 = 160$ times during encryption. So P_c for AES-128 is $(1 - \frac{1}{2^8})^{160} \approx 53.46\%$. With enough number of ciphertexts, a percentage of correct ciphertexts close to 53.46% will imply the existence of S-box faults. In the traditional method, the judgment condition relies on a ratio close to 50%, which requires thousands of ciphertexts. However, when the total number of ciphertexts is not enough, the traditional method will not work well.

In this Letter, a new method is proposed. The judgment condition is now termed as that there exist both correct and faulty ciphertexts. In this method, a non-S-box fault will never be misinterpreted into a 'real' one, in other words, the false positive rate P_{FP} is zero. And the probability of ignoring an S-box fault (i.e. the false negative rate P_{FN}) can be computed as $P_{\mathrm{FN}} = P_c^N + (1 - P_c)^N$ for N ciphertexts. For AES with the S-box of 256 elements, dozens of ciphertexts can make P_{FN} small enough. For example, when $N = 20$, P_{FN} will be less than 10^{-5}.

However, due to the large number of accesses to the same table for some lightweight block ciphers, P_c can be close to 0, which will bring the confusion of false negative. PRESENT is an example whose S-box has $B = 4$ bits and will be accessed $T = 496$ times. So $P_c = 1.25 \times 10^{-14}$. Even if trillions of ciphertexts are collected, P_{FN} is still closed to 1. In this case, a fault that did exist in S-box will be ignored and misinterpreted.

Method B: To cope with the case of lightweight block ciphers, a new method is proposed. Assuming an S-box element y is changed into another one (y'), the output of substitution will never contain y, and y' will appear twice. Considering the last round of an SPN cipher, the S-box output is linear permutated and XORed with round key k. For the sake of brevity, the linear permutation is ignored. The distribution of the ciphertext element (the XOR result of the output of S-box and the round key) also holds the property that one value is missing and another value appears more frequently. The distribution of jth word of ciphertext c_j follows:

$$\Pr(c_j = v) = \begin{cases} 0 & v = y \oplus k \\ 2 \times 2^{-B} & v = y' \oplus k \\ 2^{-B} & \text{otherwise} \end{cases} \tag{2}$$

To distinguish the S-box faults, our proposed method will collect N ciphertexts and check all of their L words. The judgment condition is termed as: if each of the words has at least one value that never appears, the fault will be judged as an S-box fault. In this method, an S-box fault will never be misjudged, i.e. the false negative rate is 0. However, with insufficient ciphertexts, a non-S-box fault could be misinterpreted. Supposing the words of ciphertexts follow uniform distribution, the false positive rate P_{FP} can be calculated as

$$P_{\mathrm{FP}} = \left(\sum_{i=0}^{2^B} (-1)^i \cdot C_{2^B}^i \cdot \left(1 - \frac{i}{2^B}\right)^N\right)^L \tag{3}$$

P_{FP} is the probability that each of the L words has at least one missing value in the N ciphertexts. It will converge to 0 if N is large enough, but the convergence speed depends on both B and L. Generally, when B is smaller, i.e. when the S-box is smaller, the speed is faster. The relation between P_{FP} and N for PRESENT ($B = 4$, $L = 16$) is shown in Fig. 1*a*. To make P_{FP} small enough, e.g. $P_{\mathrm{FP}} < 10^{-5}$, 51 ciphertexts are required for PRESENT. However, Method B does not work well for large S-box. As shown in Fig. 1*b*, thousands of ciphertexts are required for AES to make P_{FP} small enough.

In summary, it can be claimed that Method A is better for large S-box (e.g. 8 bits) and B works for a small one (e.g. 4 bits). So the adversary can choose the proper method according to the size of S-box.

Experimental result: The two proposed methods are tested with physical experiments on ATmega163L microcontroller (μC). Both AES and PRESENT ciphers are implemented, and their S-boxes are stored in the same on-chip SRAM. The chip is first decapsulated. Then a through-substrate image of the μC is taken, as shown in the left part of Fig. 2. Our target, the 1KB SRAM, is located at the right-bottom corner and its size is about 2000μm × 1000 μm. The injection facility is composed of a pulsed laser generator, a microscope system, a three-dimensional motorised stage, and a computer. The μC is mounted on the 3D motorised stage. The pulsed laser beam is collimated and focused on the backside of μC by the microscope system, and the computer can control the facility by sending commands to the synchronisation control system via a serial port. The laser spot size is about 2 μ*m* which is precise enough to inject single bit faults into the μC, and each pulse of laser has a width about 17 *ps*. To pursue a higher rate of data-flips, the energy of the laser must be carefully selected. Repeated trials are conducted to check the effect. The data-flip threshold of SRAM is found to be about 100 pJ, and the latch-up threshold is 250–300 pJ. So the energy of 200 pJ is chosen.

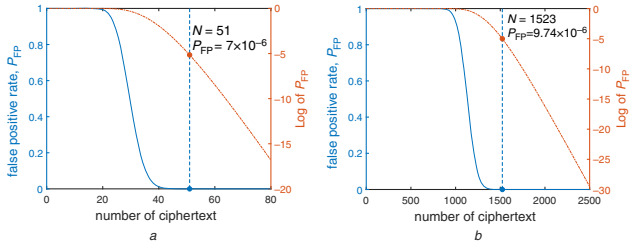

Fig. 1 *Theoretical false positive rate for Method B on different ciphers*
a PRESENT
b AES

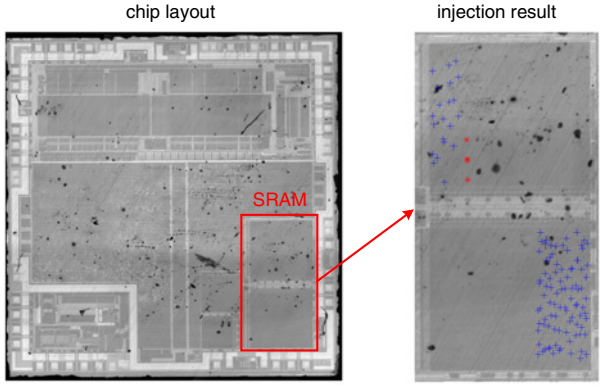

Fig. 2 *Through-substrate image of an ATmega163L (left) and the fault injection result in SRAM (right). The AES S-box faults are marked in blue and the PRESENT S-box faults are in red*

In our experiments, a total of 5000 laser pulses are randomly injected, which took about 30 min. After each laser pulse, 20 and 50 encryptions are conducted for AES and PRESENT, respectively. All the ciphertexts are collected for further fault analysis. According to the estimation aforementioned, such number of ciphertexts for each injection should make both misjudgments rates lower than 10^{-5} theoretically. With 5000 injections in the physical experiment, a total of 112 AES and 4 PRESENT faults are judged as S-box faults by Methods A and B, respectively.

To double check our results, the entire SRAM is dumped after each laser pulse. Comparing the SRAM data before and after the laser pulse, the ground truth of data flips can be read directly from the dumped file of SRAM contents. This comparison shows that a total of 458 data-flips are caused by the lasers among those 5000 injections. The S-box faults among the data-flips are identical to those found by our methods, which means that our proposals judged all the S-box faults very precisely.

Then the influence of the number of ciphertexts is investigated. Fig. 3 shows how the number of S-box faults to be found corresponds to the different number of ciphertexts. The results of Method A for AES and Method B for PRESENT are shown in Figs. 3*a* and *b*, respectively. The blue curves show the number of S-box fault found by our methods. The orange lines show the ground truth. Comparing the results of our methods with the ground truth, with 11 ciphertexts for AES and 49 ciphertexts for PRESENT, no misjudgment will be made (i.e. all of the 112 AES S-box faults and 4 PRESENT faults can be found).

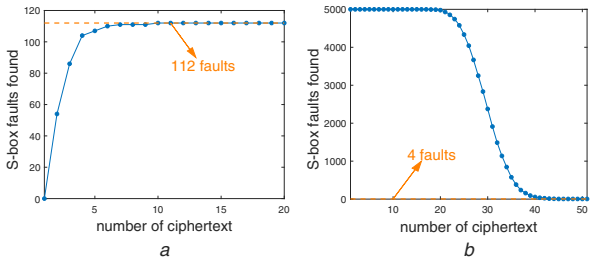

Fig. 3 *Number of S-box faults found versus the number of ciphertexts. The blue curves shows the number of faults found by our methods and the orange line shows the ground truth*
a AES
b PRESENT

Further, a closer look can be taken on the distribution of the coordinates of the 458 data flips. As shown in the right part of Fig. 2, all the data flips are represented as small black dots. Those AES and PRESENT S-box faults are highlighted as blue crosses and red stars, respectively. It can be found that both S-boxes are stored in a rectangular area with continuous address space, which can be utilised to estimate the physical area of S-boxes.

Conclusion: In this Letter, we propose two methods for distinguishing S-box faults of SPN ciphers within dozens of ciphertexts. Our methods are tested on both AES and PRESENT ciphers with a physical experiment on ATmega163L μC. Our result shows that these methods can distinguish the S-box faults with very high precision.

Acknowledgments: This work was supported in part by Open Fund of the State Key Laboratory of Cryptology (grant no. MMKFKT201805), by the Alibaba-Zhejiang University Joint Institute of Frontier Technologies, by Zhejiang Key R&D Plan (grant no. 2019C03133), by the Young Elite Scientists Sponsorship Program by CAST (grant no. 17-JCJQ-QT-045), and by the Major Scientific Research Project of Zhejiang Lab (grant no. 2018FD0ZX01).

Submitted: *24 August 2019* E-first: *21 October 2019*
doi: 10.1049/el.2019.2865
One or more of the Figures in this Letter are available in colour online.

Fan Zhang and Yiran Zhang (*College of Information Science and Electronic Engineering, Zhejiang University, People's Republic of China*)

Feng Lin and Kui Ren (*College of Computer Science and Technology, Zhejiang University, People's Republic of China*)

✉ E-mail: flin@zju.edu.cn

Huilong Jiang and Xiang Zhu (*National Space Science Center, Chinese Academy of Science, People's Republic of China*)

Fan Zhang: Also with State Key Laboratory of Cryptology, Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Zhejiang Lab, and School of Cyber Science and Technology, Zhejiang University, People's Republic of China

References

1 FIPS PUB 197. Advanced Encryption Standard. http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf
2 Bogdanov, A., Knudsen, L.R., Leander, G., *et al.*: 'PRESENT: an ultra-lightweight block cipher', *Lect. Notes Comput. Sci.*, 2007, **4727**, pp. 450–466
3 Dan, B., Demillo, R.A., and Lipton, R.J.: 'On the importance of checking cryptographic protocols for faults'. Int. Conf. on Theory and Application of Cryptographic Techniques, Konstanz, Germany, May 1997, pp. 37–51
4 Skorobogatov, S.P., and Anderson, R.J.: 'Optical fault induction attacks'. Int. workshop on cryptographic hardware and embedded systems, Redwood Shores, CA, USA, August 2002, pp. 2–12
5 Zhang, F., Lou, X., Zhao, X., *et al.*: 'Persistent fault analysis on block ciphers', *IACR Trans. Cryptographic Hardware Embedded Syst.*, 2018, **2018**, 150–172