Impacts of Technology Trends on Physical Attacks?

Philippe Maurine, Sylvain Guilley

Impacts of Technology Trends on Physical Attacks?

Philippe Maurine[1(✉)] and Sylvain Guilley[2,3]

[1] LIRMM, 161 Rue Ada, 34 090 Montpellier, France
philippe.maurine@lirmm.fr
[2] Secure-IC S.A.S., 15 Rue Claude Chappe, Bât. B, 35 510 Cesson-Sévigné, France
sylvain.guilley@secure-ic.com
[3] LTCI, Télécom ParisTech, Université Paris-Saclay, 75 013 Paris, France

Abstract. Chip fabrication technologies evolve at an explosive rate. Notwithstanding, we analyze that attacks on *smartcard* chips are almost not impacted: only the architecture which gets more complex (e.g., the devices transition from mono- to multi-core) and the advanced design solutions (adaptative voltage and frequency scaling, multiple clock domains, asynchronicity, etc.) somehow make attacks slightly more complex. The situation is different for chips tightly integrated in embedded devices, such as *smartphone* chips. Indeed, the chips size and complexity increase drastically, and thus attacks identification phase becomes extremely hard. In addition, the chip targetted by the attacks is usually stacked with other chips (like the memory), which makes access to leakages and injection of faults a challenging task. Therefore, we conclude that there is a clear gain of security in the future to use smartphones as secure elements. Attacks at printed circuit board level associated with signal processing and machine learning could question this conclusion. Also, as a perspective, we notice that new kinds of attacks become possible on smartphones. Those devices being intrinsically connected, the new side-channel and fault injection attacks are realized not physically, but in software (controlled from an external center attack process): such attacks are called microarchitectural cache timing attacks (regarding side-channels) and RowHammer attacks (regarding fault injections). We predict increasing progress in those *cyberattack* threats.

Keywords: CMOS (Complementary Metal-Oxide-Semiconductor) technology · Fabrication evolution · Physical attacks · Side-channel attacks (SCA) · Fault injection attacks (FIA) · Countermeasures · Smartcards · Smartphones

1 Introduction

Physical attacks on cryptographic implementations date back to 1996, i.e., more than twenty years ago. The first side-channel attacks were the *timing attack* [17] (1996) and the *differential power analysis* [18] (1999). Later on, other side-channels had been exploited, such as the electromagnetic (EM) field, which

allows to capture leakage non-invasively through the plastic packages, and also to narrow down the area of the captured signals. The first *fault injection* attack [4] (1997) consisted in the perturbation of a Rivest-Shamir-Adleman (RSA) computation using the Chinese Remainder Theorem (CRT).

We notice that the first side-channel attack (*timing attack* [17]) has been carried out on a Pentium chip, designed in 350 nm CMOS technology and clocked at 120 MHz. Today, the state-of-the-art processor of the same brand is the core `i7 7700`, designed in 14 nm CMOS technology and running at 4.20 GHz. This change is truly drastic[1]. This fantastic rate of innovation has been sustained accurately for 50 years[2]. Therefore, a natural question is thus to re-evaluate the potential of physical attacks given so many changes.

Physical attacks on integrated circuits proceed in two steps. First of all, some sensitive signals are either measured (case of *side-channel attacks*) or perturbed (case of *fault injection attacks*). Then, the traces and/or the effect of fault is analyzed, in a view to gain information on the secrets. The first step requires an access to the device. Clearly, the success of the attack depends on the reliability of the first phase, which in turns depends on the way the device is fabricated. As already mentioned, the fabrication technology evolves at a very high pace, for increased performance, cost, and integrability. Therefore, it is important to envision how the attacker potential will evolve. We make a difference between simple chips such as *smartcards* and integrated chips, such as *smartphones*.

In the rest of the paper, we first describe in Sect. 2 the various factors which allow for chip fabrication improvements. Then, in Sect. 3, we analyze how attacks are affected by these trends; our main result is summarized in Table 5 (c.f. Sect. 3.2.5). Emerging attacks for secure chips are discussed in Sect. 4. Finally, conclusions and perspectives are given in Sect. 5.

2 Integrated Circuits: Evolution and Trends

2.1 CMOS Technology Evolution

Gordon Moore is well known as co-founder of Fairchild Semiconductor and Intel corporation, but also owing to its famous "Moore law" [19]. This law predicts that the density of chips increases exponentially with time, namely that it doubles every eighteen months. Said differently, the minimum feature size, typically the transistors gate width, is multiplied by $1/\sqrt{2} \approx 0.707$ every eighteen months. Remarkably, this law has revealed true for more than 50 years. It is unclear today whether the law holds *per se* or whether it is self-realizing. Anyhow, this trend is a strong driver of the electronic industry, and has permitted many applications.

[1] Recall that, among all innovative technologies (health, biology, materials, etc.) developed worldwide, electronic chips are one where evolution is the largest and fastest: the number of patents filled every year is the most important (source: WIPO [30, Appendix B]), and the technology generation changes every eighteen months.

[2] Every one and a half year, a new *technological node* is released, where it is possible to integrate twice as more logic.

In practice, Moore law is merely an integration objective. However Dennard et al. [7] explain how to obtain an efficient scaling of MOSFETs (MOS Field Effect Transistors) in a view to integrate them in higher performance circuits.

It is all the more interesting as this explosive integration rate can even be sped up in practice, due to progress of related techniques: for instance, design methodologies and computer aided design (CAD) tools have allowed a better usage of the transistors for a given function.

Still, it is worth mentioning some peculiarities which occurred on the way of Moore/Dennard law. First of all, initially for a scaling $1/\kappa^2$ in density, we could observe an increase of κ of the maximum clock frequency, and a decrease of κ^2 of the power consumption (thence a constant power density ratio). However, starting from 2003 (with the 130 nm technological node), the clock frequency and the power consumption could not manage to scale at the same speed as that of density. This is due to the end of the supply (referred to as *vdd*) and of the threshold (referred to as *vth*) voltages shrinking. We recall that:

- *vdd* determines the power consumption of the chip (it varies as vdd^2), and
- *vth* determines at which voltage the CMOS gates switch; thus, the speed of the gates slows down when *vth* increases.

Their evolution with technological nodes is provided in Table 1, where the asymptotic limit $vdd \longrightarrow \approx 1$ V and $vth \longrightarrow \approx 0.3$ V can be clearly seen. Therefore, some problems arises, such as excessive power density. Second, the static power consumption started to become non-negligible compared to dynamic power consumption. Third, the feature size being so nanoscopic, variability issues arose. The reaction to make up for these issues were innovations at the architecture level:

- Power issues have been compensated by the use of clock gating, sleep modes, adaptative clock selection, and adaptative power. Indeed, playing with the *vbb*, for body bias voltage, it is possible to dynamically trade less speed for less power, and also to reduce static leakage currents.
- Multiplication of elements (e.g., multi-core circuits) allows to compensate for frequency limitation (the throughput is kept increasing at constant speed by increased parallelism).
- Variability due to process variation is mitigated by some redundancy in both the design redundancy (e.g., using spare resources) and adaptive design solutions.

Table 1. Indicative evolution of *vdd* and *vth* over 7 technological nodes

Node	250 nm	180 nm	130 nm	90 nm	65 nm	45 nm	28 nm
vdd	2.5 V	1.8 V	1.2 V	1.1 V	1.0 V	1.0 V	1.0 V
vth	0.5 V	0.4 V	0.3 V	0.3 V	0.3 V	0.3 V	0.3 V

Besides, even if the operation frequency is reaching a limit, it is not obvious to keep circuits operate so fast. Therefore, most circuits embed asynchronous clocks. For example the recent processors feature a main clock whose frequency is slightly modulated (thanks to a much slower clock), in order to avoid problems of resonance and electromagnetic compatibility (EMC).

As of today, the next node has a thinness of 7 nm to 5 nm (cf. Fig. 1), which is almost at the atomic scale. Therefore, it can be noticed another evolution of CMOS technologies, namely "More than Moore". This means that a variety of innovations allow to diversify what is feasible in CMOS logic. Example of such *CMOS helpers* are:

- new non-volatile memories (NVM) to make up for FLASH scaling limits and costs,
- 3D stacking of circuits, for a larger density, and also for the overall application to take advantage of various nodes at the same time. Indeed, it is expected an optimization in terms of cost and risk, e.g., due to potential yield issues of monolithic solutions. It is less risky to devise a system based on several chips proved to work in a robust way than with advanced heterogeneous technologies all implemented in a single chip. As a side-effect, the test complexity is also reduced. Eventually, costs are saved owing to a reuse of silicon-proven IPs.

  Currently, the technologies in production are *stacked dies* or *package on package*. However, tomorrow, the paradigm will shift from *3D IC packaging* to *3D IC integration* (which is still at research stage). This can consist in chips stacking, by exploiting the *through silicon vias* (TSV) process, or even in monolithic 3D solutions, where field effect transistors (FET) are stacked vertically.

Eventually, CMOS process itself might be questioned in a medium term future. For instance, CMOS might be traded by carbon nanotubes (CNT) or even quantum computing. However, those revolutions, should they occur, are considered out of the scope of this paper, because they would be so disruptive that it is hard to make accurate predictions.

2.2 Today and Tomorrow Secure ICs

We can observe that smartcards, which are single chip in a majority of the cases (since they must be low-cost), face the difficult problem to embed FLASH memory. Indeed FLASH technology requires the generation of voltages greater than *vdd* to write data in it, as it consists in a double gate, in which charges shall be injected permanently. Thus, charge pumps must be integrated, which is complex as they are laid out in analogue logic. Besides, there is an issue related to charge retention: advanced nodes for FLASH are thus less reliable, since data are saved on tiny floating gates. Eventually, FLASH process requires specific (hence more costly) manufacturing process, because double gate transistors need two polysilicon layers. Therefore, they are lagging about 5 to 7 technological nodes behind

the state-of-the-art circuits, such as smartphones. This is illustrated in Fig. 1, where the FLASH and pure logic technological specialties are highlighted in red ovals. It clearly appears that advanced technological nodes are most targeting large digital circuits, and not NVM such as FLASH. Indeed, those devices are characterized by the fact their NVM is off-chip, thereby solving the issue of common integration of CMOS logic and FLASH (called eFLASH). From a security point of view, embedded FLASH memories are also vulnerable targets as they become inoperative as soon as the charge pump is inhibited, which happens for instance when the attacker manages to illuminate it strongly with a focused LASER source. So-called *bumping attacks* are also a real-world threat [24].

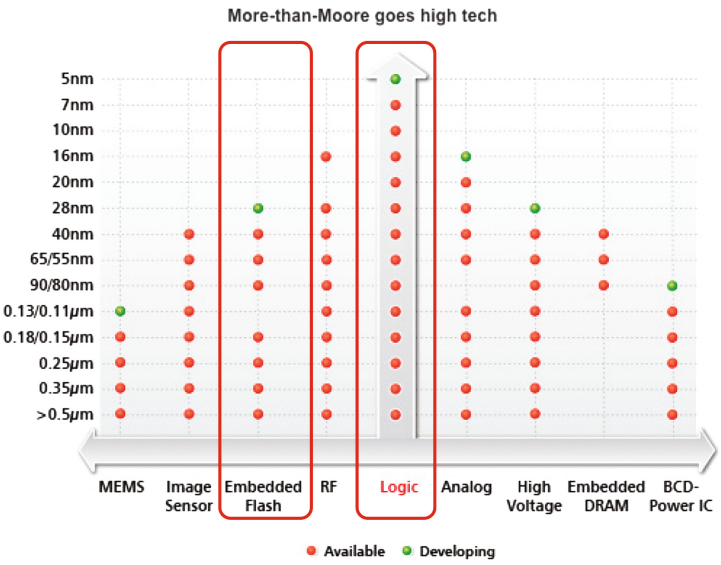

Fig. 1. Different nodes for different markets. Source: TSMC (courtesy of Ed Sperling, Fig. 3 of [25]—with our two "0" annotations)

Owing to the peculiarity of CMOS technologies evolutions (techniques to make up for *vdd* and *vth* limitation in terms of scaling, and "More than Moore" options), smartcards and smartphones secure chips do differ a lot. Typically, smartphone chips implement spacial parallelism (e.g., their architecture is *multi-core*) and 3D-stacking, which will have, as we shall see next, a positive impact on their security vis-à-vis physical attacks.

Still, objects that simple as single chip devices still have a usefulness in practice. The reason is the enormous growth in terms of internet-of-things (IoT) devices, for smart applications in retail, building, transportation, energy, health, etc. And let us mention that despite their apparent simplicity (being single chip), they remain all the same very powerful. For instance, a current smartcard microcontroller, such as an STMicroelectronics STM32F4 from 2013, has the same

Table 2. Comparison between a Pentium from year 1993 and a single-chip processor in year 2013.

Brand	Intel	STMicroelectronics
Model	Pentium	STM32F4
Year	1993	2013
Processing power	239 DMIPS @ 133 MHz	255 DMIPS @ 180 MHz
Power efficiency (P/MHz)	75 mW/MHz	40 µW/MHz
Size	3100 K transistors (≈775 K-gates)	1246 K-gates
Minimum feature size (technology node)	800 nm	90 nm
vdd	3.0 V	1.2 V

computing capabilities as a former Intel Pentium (top-class personal computer) processor had twenty years earlier, in 1993, while being much more efficient in terms of power consumption. Refer to Table 2 for quantitative details; the performance is expressed in terms of Dhrystone millions of instructions processed per second (DMIPS, or mega-instructions per second), is similar for both chips. The STM32F4 chip is much more power-efficient than the Pentium, which is a benefit of CMOS down-scaling. Actually, the STM32F4 chip with its on-chip eFLASH memory is 7 technological nodes behind state-of-the-art. Therefore, $20-7\times 1.5 = 12.5$ years of electronic fabrication progress separate the Intel Pentium and the STMicroelectronics STM32F4 chip, which coincides with Moore law:

$$\frac{\text{area}(t_0)}{\text{area}(t_0+12.5)} = \left(\frac{\text{size}(t_0)}{\text{size}(t_0+12.5)}\right)^2 = \left(\frac{800\text{ nm}}{90\text{ nm}}\right)^2 = 2^{12.5/1.5},$$

where $t_0 = 1993$ is the origin date of the oldest technology. Notice that neither Intel Pentium nor STMicroelectronics STM32F4 chips are secure; however, they are both representative of secure architectures (smartcards can be viewed as extremely secure microcontrollers). Eventually, we notice a final difference between smartcard and smartphone types of chips: as of today, a consumer is ready to spend $1 to buy a smart device to monitor its heart while jogging, but is less amenable to spend $1000 (`i7` Intel cost) for the similar purpose.

A comparison between features of today's smartcards and smartphones is given in Table 3. Basically, a smartphone processor consists in the assembly of several chips, whereas an archetype smartcard consists in general of only one. A smartphone has several processors, each within its own island, where *vdd*, *vbb* and *frequency* can be chosen independently, and changed dynamically depending on the load and/or the power policy. Eventually, the processors of smartphones are accelerated using cache memory to speed up the access to the main RAM, which is shared among cores.

Table 3. Comparison between features of today's smartcards and smartphones

Features	Smartcards	Smartphones
Number of chips	1	≥2 (processing + memory chips + MEMS, etc.)
Number of processors	1 with fixed *vdd*, *vbb* and *frequency*	≥4, each with its own configurable *vdd*, *vbb* and *frequency*
NVM	eFLASH	Stacked chips of external memory
Use of cache memory	No	Yes

3 Physical Attacks and Technology Trends

3.1 Current Practice of Physical Attacks

The environment in which state-of-the-art attacks are performed is described in Table 4. We notice that most of the secure chips tested, as of today, consist in single-chip cryptographic modules (as per jargon of [26, Sect. 4.5.2]) of "smartcard" type (e.g., trusted platform modules, secure ICs, etc.). This is mostly the result of a strict security regulation on those objects, for which the highest evaluation assurance levels are demanded (e.g., in terms of Common Criteria [5] certification).

Table 4. List of conditions in which physical attacks are performed as of today

	Side-channel attacks	Fault injection attacks
List of conditions	Access to a leaking signal (power consumption, EM radiation, etc.)	Physical access to the device (laser, EM fault injection, etc.)
	Stability of the leaking signals, in space and time: – constant *vdd*, *vbb*, *frequency*, – constant location of the sensitive calculi	Stability of the targetted signals, in space and time: – constant *vdd*, *vbb*, *frequency*, – constant location of the sensitive calculi
	Moderated clock frequencies, few number of clock domains, asynchronicity	
	Moderated IC complexity (≈1 million gates equivalent)	
	Moderated computational noise	
	CMOS 90–65 nm technological nodes	

Clearly, the state-of-the-art of attacks practice (Table 4) is not meant to be rigidly interpreted: the attackers are smart, and adapt to new contexts. For instance, it is demonstrated in [21] an attack on a recent multi-chip system fabricated in 22 nm technology, probably made of hundreds of millions of transistors, and running at a frequency above the gigahertz.

3.2 Adversary's Challenges

We analyze here four factors in electronic circuits fabrication progress which impact the realization of attacks.

3.2.1 "CMOS Scaling" Factor

The effect of Moore's law in the reduction of size of transistors (gates) is clearly in the disadvantage of the attacker. As illustrated in Fig. 2, when the features in circuits shrink:

- **For side-channel attacks**:
  - the algorithmic noise increases,
  - unless the attacker is able to scale down the EM antennæ while conducting local measurements; some minor improvements can be made in this direction, as the radius of the probe (inductance) shall be at least 5 times the width of the metal ($\geqslant 10\,\mu m$). Another option to increase the signal-to-noise ratio (SNR) is merely to collect more traces. Indeed, when the noise is normal and independent from one trace to the other, the SNR increases linearly with the number of collected traces. Alternatively, new side-channels, such as photonic analysis [23] or voltage contrast microscopy [16], can also overcome the decreasing feature size of recent CMOS technologies.
- **For fault injection attacks**:
  - global faults [12] are less selective, since there are more signals (other than the sensitive ones) likely to be faulted,
  - whereas local faults require a scaling of the EM injection antennæ, or body bias probe tip [2], or laser spot diameter, etc. However, we reach here a limit as the section of a laser beam cannot be smaller than its wavelength, which is equal to $\approx 1\,\mu m$ for red light. Still, it is known that, for an attack to succeed, the attacker does not need to have the extremely strong capability to target one gate or one memory cell alone. Besides, it

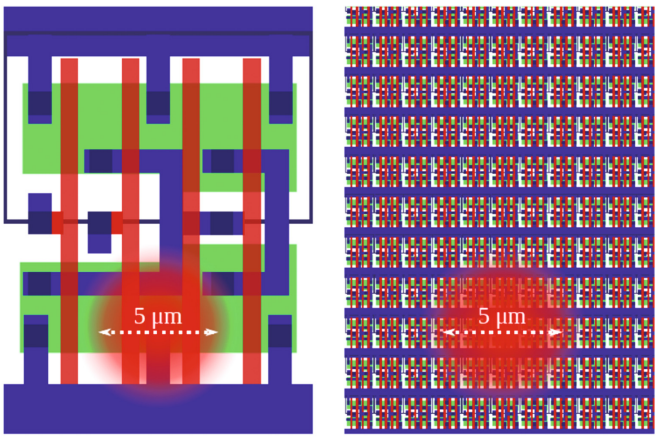

Fig. 2. Comparison of scale between a technology node where standard cells are layouted with a height Hstd_cell = 12.5 µm and a more recent technological node (which is seven nodes apart) where Hstd_cell = 1.2 µm

has been noticed adequately in [1] that the effect of a single *bit flip* can be obtained with a large injection area (wider than the gate carrying the bit to be flipped) all the same. Indeed, assuming a Gaussian profile for the laser beam, the attacker can reduce its intensity so that it effective area is not that at $1/e$ of the power, but at much higher threshold. Furthermore, if the attacker manages to setup an attack path where the bit to fault is surrounded by bits which are unused, then it suffices to fault very coarsely around the intended bit. The collateral effects have no consequence on the success of the attack.

On the contrary, one can notice that the effect of *vdd* and *vth* (recalled in Table 1) induces marginal changes in the current and voltages inside of the gates. Therefore, the *signal* an attacker is able to collect in a side-channel analysis does not weaken significantly. This means that his advantage is almost preserved. In a similar way, the propagation time in gates is also little affected, hence critical paths keep at the same order of magnitude. Thus, global fault attacks (e.g., clock tampering, underfeeding, etc. [12]) continue to work the same (provided the attacker manages to find an experimental fault injection procedure as targetted as possible, e.g., mostly the sensitive application runs, whilst the rest sleeps).

So, to conclude this analysis, one can say that CMOS scaling has either no impact on the physical attacks, or an impact which can be mitigated, typically by scaling down measurement and/or injection antennæ (or probe tips, laser beam focus, etc.).

3.2.2 "Physical Access to Device or Leaking Signals" Factor

Smartcards, by essence, are not concerned by 3D assembly. And even if some rare models implement this technology, one shall keep in mind that in a smartcard, attacks can be done both *frontside* and *backside*. Therefore, the attacker has more freedom to place its probe and/or injection tool. This is not the case for smartphones, since it is hardly possible to desolder the 3D integrated stack of chips and have them work standalone (because the attacker will have hard time to figure out how to plug the power supplies, the clocks, etc., but also because the system boot might be conditioned to the presence of other elements such as peripherals, which would stop the boot process unless connected). For the same reason, it is always possible for a side-channel attacker to monitor the current consumed by a smartcard (because it must be provided externally). However, this is not an option for smartphones, since it is very difficult to deport parts of the smartphone and still have them work (for reasons on signal integrity, in particular, and also because some models might implement anti-tampering techniques). We nonetheless attract the reader's attention to some recent trials of community building on this topic, e.g., through the organization of the http://www.hardwear.io conference. Thus, attention shall be kept on the topic of invasive attacks on complex smart devices such as smartphones.

Still, despite a 3D assembly, it can be imagined that leaking signals are conducted [27], hence can be measured even if the sensitive chip is placed in

sandwich between two unrelated chips. Therefore, methodology presented in ISO draft international standard 20085-1 [13] might apply.

The fault injection attacks will be more sensitive to the way the 3D integration is done in practice. As mentioned in Sect. 2.1, such integration is getting tighter and tighter (moving from *stacked dies* to *3D IC packaging/integration*). Therefore, the disassembly required to access the sensitive parts of the chips is getting very challenging. We thus rate such attack at maximum level. However, it shall not be forgotten that novel fault injection on chips assembly might show up (RowHammer, discussed latter in Sect. 4, is a testimony that innovative attacks might be revealed out of the blue). One research direction we would like to point out is the practical study of *conducted perturbation* fault attack [22].

3.2.3 "Architecture and Advanced Design Solutions" Factor

In a view to save energy and better address the tradeoff between power consumption and efficiency, new design strategies are emerging. They include adaptative voltage and frequency scaling (abridged AVFS) which can be activated dynamically. This implies independent clock domains to cooperate (some logic is even fully self-timed, i.e., asynchronous), and charge balance in multicore systems.

The main impact of this trend is that side-channel traces realignment will become difficult. Moreover, in the case of multicore systems, it will become hard to attribute such portion of code to that process (including the one under attack). Maybe side-channel analysis techniques can tolerate these experimental drawbacks (see for instance [6]). But we expect that more genericity will come with a price on the efficiency side. Typically, on mono-threaded systems where *vdd*, *vth* and the *frequency* are fixed, accurate leakage models, namely Hamming weight and Hamming distance are known to match reliably the reality. Such advantage might disappear in the more challenging setup of varying signal amplitude and pace.

Fault attacks can better tolerate asynchronicity. Indeed, a wide array of FIAs need only one single fault to be conclusive on the part of the secret to recover. On the contrary, SCAs, both for "*simple*" [18, Sect. 2] and "*differential*" [18, Sect. 4] analyses, need to accumulate many measurements to cancel out as much noise as possible (and indeed, the noise level is exacerbated in the context of our adaptative device). For instance, a fault attack which consists in skipping a test, can be repeated many times, until the test is eventually successfully skipped, which will statistically happen independently of the AVFS features which randomize the execution pattern of the targetted code. As another example, let us consider differential fault injection attacks on AES. Here, provided the injection is lucky enough (it shall attain only one byte in the antepenultimate round), one fault suffices to recover a full 128-bit key [29]. Obviously, still, FIAs that require many faults (which are not so widespread) encounter the same difficulties as SCA attacks. Concluding, FIAs are less threatened than SCAs in the dynamic environment changes due to smart power-optimizing behaviors of the device.

3.2.4 "Die Size and Complexity" Factor

Regarding smartcards, we notice that their size has been decreasing over time (from more than 10 mm^2 at their inception, to $\approx$1 mm^2 today). The reason is that technological design shrink is dominating, while those objects have become smarter and smarter, hence requiring more logic. From the attacker perspective, this means that the design is more complex. However, the basic building blocks of smartcards have not changed over time: for example, there is still the need for one CPU, however it moved from 8 to 16 bit and then from 16 to 32 bit, over time. Same situation happens for the eFLASH, RAM, ROM, EEPROM memories: they are always part and parcel of a smartcard, however over time, their capacity has been growing, so as to enable more interesting applications. Thus, the attacker can always identify the parts which he intends to measure or fault, and thus complexity does not hurt him.

Smartphones represent a different case. As mentioned in Sect. 3.2.2, owing to 3D stacking of chips, the side-channel measurements and the fault injections can no longer be made local and accurate, all the more so as the application itself is mobile within the chip (e.g., moving from one core to another, so as to balance the load). Therefore, side-channel traces now consist in a patchwork of execution of various unrelated processes, which is in practice very challenging if not impossible to unravel. Fault injection attacks suffer the same problem of sensitive activity "volatility". Tracking for the manipulation of a sensitive data or operation can thus be compared to *searching a needle in a haystack*, thereby making fault injection attacks almost impossible (though probably all the same easier than exhaustive key search).

3.2.5 Summary

Interestingly, CMOS evolution does not make the attacks easier. However, some attack paths are more impacted than others. In order to summarize in one chart the impact of the four considered factors, we provide a qualitative rating using this terminology:

-  : no impact at all (the attack remains robust despite technological evolution),
-  : small negative impact (the attack is still possible at the price of little more efforts),
-  : strong negative impact (the attack is becoming challenging—probably it is no longer a valid attack path),
-  : huge negative impact (the attacks becomes almost infeasible).

The impacts of the four factors discussed in Sects. 3.2.1, 3.2.2, 3.2.3 and 3.2.4 are summarized in Table 5, which constitutes the main result of this paper.

Besides, it shall be noticed that smartcards favor the attackers, because the sensitive operations can be directly triggered by APDU (Application Protocol Data Unit, cf. ISO 7816-4 [28]) commands. This eases the attack, compared

Table 5. Summary of the effect of four evolutions in CMOS circuits on physical attacks

Factors	Smartcard		Smartphone	
	SCA	FIA	SCA	FIA
CMOS scaling (Sect. 3.2.1)	🙂	🙂	🙂	🙂
Physical access to device or leaking signals (Sect. 3.2.2)	🙂	🙂	😟	😖
Architecture & advanced design solutions (Sect. 3.2.3)	😟	😐	😟	😐
Die size and complexity (Sect. 3.2.4)	🙂	🙂	😖	😖

to smartphones, for which the synchronization is a real challenge. Indeed, on smartphones, the access to the API (Application Programming Interface) is less straightforward: there is no direct call from the user, hence it is difficult to master the manipulated data and to control the time/order of executions (which are often based on proprietary mechanisms).

The research efforts required by attackers to overcome the difficulties (denoted by 😟 and 😖 in Table 5) due to CMOS technologies evolutions are listed below:

- To make up for difficult access to the leakage or to the device itself, it is foreseen some advance in terms of *conducted leakage analysis* and *conducted perturbation*;
- Against dynamic behaviour of the chip (AVFS, existence of multiple cores operating in parallel, asynchronicity), we foresee the need for more advanced techniques of *signal processing* and of more flexible side-channel distinguishers;
- Same advances could definitely help advance the power of attacks despite increase of die size and complexity;
- In complement, investments in reverse-engineering (e.g., of 3D stacking structures) would clearly increase the success of fault injection attacks.

4 Logical Side-Channel and Fault Injection Attacks

4.1 Logical Attacks

As mentioned in Table 3, smartphone processors feature *cache* memories, all of them are shared (to some extend) among the cores. Therefore, *microarchitectural attacks* [9] appear to be a nice way to attack the device when other physical counterparts are made difficult due to the four factors presented in Sects. 3.2.1, 3.2.2, 3.2.3 and 3.2.4.

The advantage of such attacks, is that the attacker is a pure software piece of code, which is coming over the top (OTT). It will be "dropped", and then will "land" directly next to the program to attack (victim) where it will be executed. Hence, such attacks allow to circumvent the problem of physical identification

of the localization (in the $X-Y$ plane) where the targetted sensitive application runs: the operating system will directly install the attacker the most closely as possible to the victim, since in multitask systems, processors are close one from each other as they depend on the same cache memory.

In other architectures, the *logical* side-channel can arise from an abuse of some monitoring functions. For instance, the integrated sensor in field programmable gates array (FPGA) platforms can be diverted from its intended usage in terms of safety to spy on some IP [20]: it thus behaves as a Trojan horse.

RowHammer attacks [15] are the full cyber counterpart of physical FIA. They share the same advantage as cache attacks: there is no need for the attacker to know the physical layout of the chip(s), nor to have any physical access.

Therefore, we expect much research in those directions. This is illustrated in Fig. 3, regarding the growth of the *remote* threat and other (less successful) *local/physical* analyses.

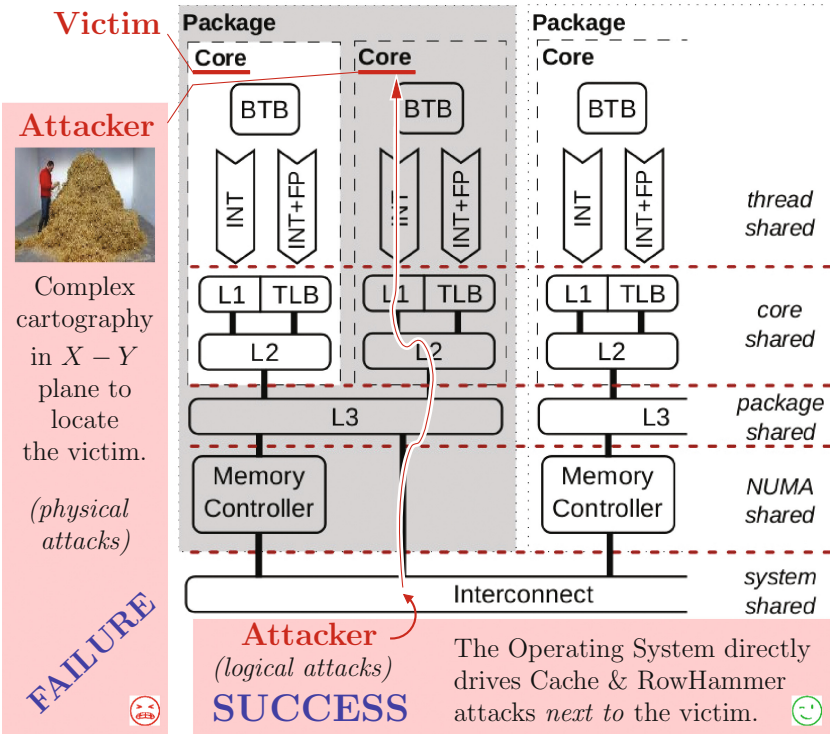

Fig. 3. Contented resources in a multiprocessor system, typical to smartphones, which allows to contrast *physical* cartography with *logical* cache attacks (background image courtesy of Qian Ge et al. [9, Fig. 1, p. 6].)

4.2 Protections Against Logical Attacks

Cache timing and RowHammer attacks demand further focused studies, as there is, as of now, no regulatory incentive to avoid them. Indeed, they are explicitly out of scope of Global Platform TEE Protection Profile [10]. Besides, there is no systematic way to counter such attacks. As an example, cache timing attacks can be made more difficult by the application of some heuristics, such as:

- replacing tests (control flow irregularities) such as `d=c?a:b` by unconditional code such as `m=-(!!c), d=(a&m)^(b&~m)`,
- trading look-ups in a table `T` such as `y=T[x]` by address-independent code such as `y=0, for(i=0..#T-1) y^=(-(x==i))&T[i]`,
- memory access randomization based on oblivious RAM (also known as "ORAM") concept [11],
- obfuscation such as white box cryptography (WBC [3]), etc.

However, even those simple patterns are prone to implementation errors. The article ironically entitled *Make Sure DSA Signing Exponentiations Really are Constant-Time* [8] shows a mistake where the constant-time operations are coded but not called adequately, hence leaving the possibility for an attacker to exploit the code. RowHammer attacks also continue to work because it is possible to access DDR SRAM (Double Data Rate Synchronous Dynamic Random Access Memory) directly (i.e., bypassing the cache memories) at high rates thanks to legacy operations, such as `prefetch` and `clflush`. Those instructions allow fast access to the DDR; thereby, paradoxically enough, efficient processors are less secure. Besides, DDR is sensitive to faults because it is very integrated. Hence the practically, as of today, of *cyber-enabled RowHammer attacks*. Notice that error correcting codes (ECC) do not prevent those attacks because their error correction capability is very limited. For instance, 2-bit ECC reduces the success probability only by a factor $2^2 = 4$, hence attacks will require only 4 times more traces to succeed, which is negligible for a determined attacker.

5 Conclusion and Perspectives

We have analyzed the various factors which allow for chip fabrication improvements. Paradoxically, we derive that Moore's law (CMOS minimum feature size decreases over time) does not impact much state-of-the-art attacks. On the contrary, factors such as voltage scaling, designs with multicores and asynchronicity make attacks (slightly) more complex. In the case of smartphones, side-channel and fault injection attacks are very impeded due to the increase of complexity of the chip, and stacking makes it more difficult to access signals needed for side-channel attacks and particularly for fault injection attacks. We conclude that the security level actually increases over time for devices such as smartphones.

Nevertheless, we believe that the mere technological evolution is not going to eradicate the problem of physical attacks. The challenge in front of attackers is now to better process side-channel curves, perform horizontal analysis on

a single (or few) curves, develop side-channel specific pattern matching techniques, improve technology to resynchronize and interpret complex curves, etc. In particular, for smartphone devices, the resolution of the timing (required for timing-based side-channel attacks and fault injection triggering) can be enhanced by physical measurements directly on the printed circuit board. In parallel, new attack paths (re)appear, such as timing attacks, hence a paradigm shift in terms of security evaluation.

References

1. Agoyan, M., Dutertre, J.-M., Mirbaha, A.-P., Naccache, D., Ribotta, A.-L., Tria, A.: How to flip a bit? In: 2010 IEEE 16th International On-Line Testing Symposium (IOLTS), pp. 235–239, July 2010
2. Beringuier-Boher, N., Lacruche, M., El-Baze, D., Dutertre, J.-M., Rigaud, J.-B., Maurine, P.: Body biasing injection attacks in practice. In: Palkovic, M., Agosta, G., Barenghi, A., Koren, I., Pelosi, G. (eds.) Proceedings of the Third Workshop on Cryptography and Security in Computing Systems, CS2@HiPEAC, Prague, Czech Republic, 20 January 2016, pp. 49–54. ACM (2016)
3. Beunardeau, M., Connolly, A., Géraud, R., Naccache, D.: White-box cryptography: security in an insecure environment. IEEE Secur. Priv. **14**(5), 88–92 (2016)
4. Boneh, D., DeMillo, R.A., Lipton, R.J.: On the importance of checking cryptographic protocols for faults. In: Fumy, W. (ed.) EUROCRYPT 1997. LNCS, vol. 1233, pp. 37–51. Springer, Heidelberg (1997). doi:10.1007/3-540-69053-0_4
5. Common Criteria Consortium: Common Criteria (aka CC) for Information Technology Security Evaluation (ISO/IEC 15408) (2013). http://www.commoncriteriaportal.org/
6. de Chérisey, É., Guilley, S., Heuser, A., Rioul, O.: On the optimality and practicability of mutual information analysis in some scenarios. In: ArticCrypt (IACR Event), 17–22 July, Longyearbyen, Svalbard, Norway (2016)
7. Dennard, R.H., Gaensslen, F.H., Rideout, V.L., Bassous, E., LeBlanc, A.R.: Design of ion-implanted MOSFET's with very small physical dimensions. IEEE J. Solid-State Circ. **9**(5), 256–268 (1974)
8. García, C.P., Brumley, B.B., Yarom, Y.: Make sure DSA signing exponentiations really are constant-time. In: Weippl, E.R., Katzenbeisser, S., Kruegel, C., Myers, A.C., Halevi, S. (eds.) Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016, pp. 1639–1650. ACM (2016)
9. Ge, Q., Yarom, Y., Cock, D., Heiser, G.: A survey of microarchitectural timing attacks and countermeasures on contemporary hardware. Cryptology ePrint Archive, Report 2016/613 (2016). http://eprint.iacr.org/2016/613
10. GlobalPlatform Device Committee: TEE Protection Profile Version 1.2.1, GPD_SPE_021, December 2016. https://www.globalplatform.org/specificationform.asp?fid=7831
11. Goldreich, O.: Towards a theory of software protection and simulation by oblivious RAMs. In: Aho, A.V. (ed.) Proceedings of the 19th Annual ACM Symposium on Theory of Computing, 1987, New York, USA, pp. 182–194. ACM (1987)
12. Guilley, S., Danger, J.-L.: Global faults on cryptographic circuits. In: Joye, M., Tunstall, M. (eds.) Fault Analysis in Cryptography. Springer, Heidelberg (2012)

13. ISO/IEC JTC 1/SC 27/WG 3. ISO/IEC CD 20085–1:2017 (E): Information technology - Security techniques – Test tool requirements and test tool calibration methods for use in testing non-invasive attack mitigation techniques in cryptographic modules – Part 1: Test tools and techniques, 25 January 2017
14. Joye, M., Tunstall, M.: Fault Analysis in Cryptography. Springer, Heidelberg (2011). doi:10.1007/978-3-642-29656-7. ISBN 978-3-642-29655-0
15. Kim, Y., Daly, R., Kim, J., Fallin, C., Lee, J.-H., Lee, D., Wilkerson, C., Lai, K., Mutlu, O.: Flipping bits in memory without accessing them: an experimental study of DRAM disturbance errors. In: ACM/IEEE 41st International Symposium on Computer Architecture, ISCA 2014, Minneapolis, MN, USA, 14–18 June 2014, pp. 361–372. IEEE Computer Society (2014)
16. Kison, C., Frinken, J., Paar, C.: Finding the AES bits in the haystack: reverse engineering and SCA using voltage contrast. In: Güneysu, T., Handschuh, H. (eds.) CHES 2015. LNCS, vol. 9293, pp. 641–660. Springer, Heidelberg (2015). doi:10.1007/978-3-662-48324-4_32
17. Kocher, P.C.: Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In: Koblitz, N. (ed.) CRYPTO 1996. LNCS, vol. 1109, pp. 104–113. Springer, Heidelberg (1996). doi:10.1007/3-540-68697-5_9
18. Kocher, P., Jaffe, J., Jun, B.: Differential power analysis. In: Wiener, M. (ed.) CRYPTO 1999. LNCS, vol. 1666, pp. 388–397. Springer, Heidelberg (1999). doi:10.1007/3-540-48405-1_25
19. Moore, G.E.: Cramming more components onto integrated circuits. In: Readings in Computer Architecture, pp. 56–59. Morgan Kaufmann Publishers Inc., San Francisco (2000)
20. Ngo, X.T., Najm, Z., Bhasin, S., Roy, D.B., Danger, J.-L., Guilley, S., Sensor, I.: A backdoor for hardware trojan insertions? In: 2015 Euromicro Conference on Digital System Design, DSD 2015, Madeira, Portugal, 26–28 August 2015, pp. 415–422. IEEE Computer Society (2015)
21. Saab, S., Rohatgi, P., Hampel, C.: Side-channel protections for cryptographic instruction set extensions. Cryptology ePrint Archive, Report 2016/700 (2016). http://eprint.iacr.org/2016/700
22. Sauvage, L., Guilley, S., Danger, J.-L., Homma, N., Hayashi, Y.-I.: A fault model for conducted intentional electromagnetic interferences. In: 2012 IEEE International Symposium on Electromagnetic Compatibility (EMC), pp. 788–793, 5–10 August 2012, Pittsburgh, PA, USA (2012). http://2012emc.org/. doi:10.1109/ISEMC.2012.6351664
23. Schlösser, A., Nedospasov, D., Krämer, J., Orlic, S., Seifert, J.-P.: Simple photonic emission analysis of AES. J. Cryptogr. Eng. **3**(1), 3–15 (2013)
24. Skorobogatov, S.: Flash memory 'bumping' attacks. In: Mangard, S., Standaert, F.-X. (eds.) CHES 2010. LNCS, vol. 6225, pp. 158–172. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15031-9_11
25. Sperling, E.: Reworking Established Nodes, 24 May 2017. Blog article: https://semiengineering.com/reworking-established-nodes/. Accessed 1 June 2017
26. NIST FIPS (Federal Information Processing Standards): Security Requirements for Cryptographic Modules Publication 140-2, 25 May 2001. http://csrc.nist.gov/publications/fips/fips140-2/fips1402.pdf
27. Sugawara, T., Hayashi, Y., Homma, N., Mizuki, T., Aoki, T., Sone, H., Satoh, A.: Mechanism behind information leakage in electromagnetic analysis of cryptographic modules. In: Youm, H.Y., Yung, M. (eds.) WISA 2009. LNCS, vol. 5932, pp. 66–78. Springer, Heidelberg (2009). doi:10.1007/978-3-642-10838-9_6

28. ISO/IEC 7816 (Joint technical committee (JTC) 1/Sub-Committee (SC) 17): Identification cards - Integrated circuit cards. http://www.iso.org/iso/catalogue_detail.htm?csnumber=35168
29. Tunstall, M., Mukhopadhyay, D., Ali, S.: Differential fault analysis of the advanced encryption standard using a single fault. In: Ardagna, C.A., Zhou, J. (eds.) WISTP 2011. LNCS, vol. 6633, pp. 224–233. Springer, Heidelberg (2011). doi:10.1007/978-3-642-21040-2_15
30. WIPO (World Intellectual Property Organization): World Patent Report: A Statistical Review - 2008th edn. http://www.wipo.int/ipstats/en/statistics/patents/wipo_pub_931.html. Accessed 11 June 2017