Memories: a survey of their secure uses in smart cards

Michael Neve, Eric Peeters, David Samyde, Jean-Jacques Quisquater

UCL Crypto Group
Catholic University of Louvain
Microelectronic Laboratory
Maxwell building, Place du Levant 3
B-1348 Louvain-la-Neuve, Belgium
http://www.dice.ucl.ac.be/crypto

***Abstract*— Smart cards are widely known for their tamper resistance, but only contain a small amount of memory. Though very small, this memory often contains highly valuable information (identification data, cryptographic key, etc). This is why it is often subject to many attacks, as the other parts of the smart card, and thus requires appropriately chosen protections.**

The use of memories in smart cards induces security problems, but also other more particular ones. The main constraint is naturally the limited physical expansion and integration, but fault level, aging and power consumption are not to be discarded.

Indeed, Kuhn [3] has proved that it is possible to read the content of a ROM by simply using a microscope. This is of course completely unacceptable because ROM contains sensitive information. The operating system and some particular applications, for example, are written in ROM. Skorobogatov [15], in his turn, has explained how the interaction between light and silicon can be used to create memory faults, while Quisquater *et al.* [12] have shown the utility of eddy current.

Nowadays, most customers refuse to go on using the once so promising Flash memories. The reason therefore is the high risk for attacks and software modifications, of which Pay TV cards have already been victim in the past.

The question arising now is what memory manufacturers can really do to solve these problems. They have indeed tried to increase the complexity of the memory points by the addition of more transistors. However, this technique increases the production costs, and consequently does not offer a long term solution. Efforts have also been made to reduce side channel leakage through the coupling of transistors or the use of dual rail logic. New technologies to counter these problems are appearing, as for instance FeRAM, but these do not solve the particular problem of intrinsic security.

This article gives a survey of all the existing techniques to counter these attacks.

***Index Terms*— Secure memories, Smart cards, Tamper resistance, Secure hardware, Side channels.**

I. Introduction and context

THE HISTORY of smart cards begins in the 1970s. At its origin the smart card was limited to a portable memory card (1974, Roland Moreno), permitting the decentralization of sensitive or personal information on a credit card sized card. Later, the card gained in performance by the addition of a microprocessor (1979, Michel Ugon) intended to meet security requirements by using cryptographic tools. Furthermore successive physical security systems upgraded the smart card and transformed it into a high-security chip. Such security levels are obviously needed to counter the always more efficient pirates, trying to break the privacy and motivated by various intentions.

With the passing years, the structure of a smart card included more and more components aimed at achieving better performance in terms of capabilities and capacity. The number and variety of its applications increased in the same time. The integration of the first smart cards in the 1980s was not without problems. At that time, there appeared to be big problems to make the regular (memory) and non regular (logical) part fit in on one single chip. Its security was tested. Problems revealed then to be of a different kind, namely in the connection between the two parts, which allowed attackers to damage the card.

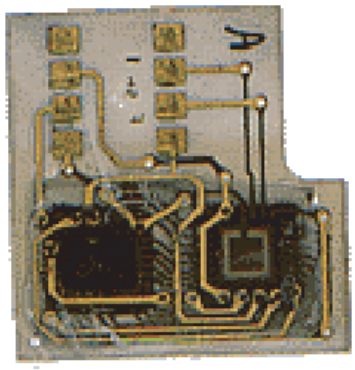

Fig. 1. A very old smart card with two pieces of silicon with available connections

It should be noticed that cards with two different silicon pieces for the memory and the processor still exist. An example thereof are the cards used by pay-TV pirates. A simple PIC (© Microchip) processor is inserted into the card with an I2C access memory (by the same manufacturers as for non pirate cards) in a way that respects the smart card's form factor. This type of card has a lower level of security and tamper resistance, but that does not matter because the form is the important factor in this case.

Current smart cards may include many components as 32-bit CPU core, memory (up to 224 kBytes of ROM, up to 66 kBytes of EEPROM, up to 8 kBytes of RAM, FeRAM,...), memory firewalls, cryptoprocessors (1088-bit modular arith-

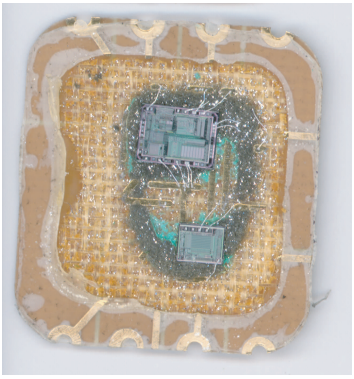

Fig. 2. A depackaged pirate card with an non secure serial memory

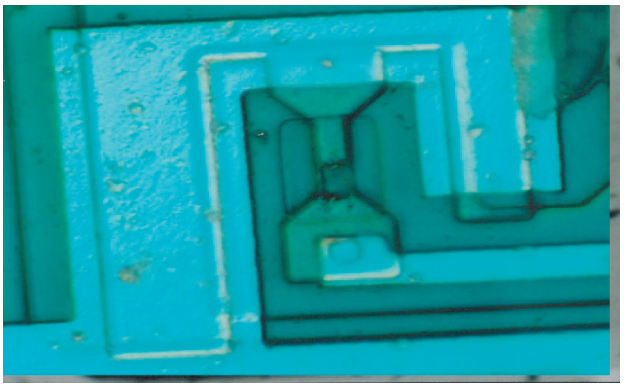

Fig. 3. A burnt fuse on a memory card

metic processors, hardware Data Encryption Standard accelerator, Advanced Encryption Standard,...), a random number generator (FIPS140-2 compliant), an internal clock generator (or re-generator), a serial I/O interface and so on.

Widely used throughout Asia and Europe, the applications of smart cards are manifold: from cell phones, payment systems, getting access to restricted area, to driving licenses, passports and ID cards. Overall, the leading advantage of smart cards is their intrinsic flexibility, security (resistance against tampering), and cost effectiveness. They constitute a simple solution for many security issues (access, payment,...).

Because they are delivered to individuals, they may be threatened by many different deliberate (or not) manipulations. An attacker can also exploit any physical leakage of such device and try to find out the secret key stored in the chip by analysis of these leakages. In 1996, Kocher *et al.* adapted the notion of *side-channel analysis* to smart cards [9], [10] and showed the importance for an implementation to be resistant against side-channel analysis and leakages from power consumption. Resistance against fault analysis [4], [5] is another issue: sensitive information may leak when the cryptosystem operates under unexpected conditions. More recently, in [6], [11] a new type of analysis has been presented, based on electromagnetic radiations of the processor when a crypto-algorithm is processed, (see also [1]), called *Electro-Magnetic Analysis* (EMA for short). In parallel, chips must be designed to counter also intrusive attacks techniques such as visual inspection (Kevlar © coating), probing (fuse), RF sniffing (inductive sensor),... In addition, to provide a secure physical environment for the chip, the module is designed to achieve a physical envelope that protects the chip from bending, scratching, mechanical stress, and static electricity encountered in normal operation.

The problem of fault insertion has also to be taken into account. Fault insertion techniques manipulate the environmental conditions of the system (voltage, clock, temperature, radiation, light, eddy current, etc.) to generate faults and observe the related behavior. Most of these attacks target the data during the computation of a sensitive algorithm, but some of them allow direct data corruption in the *memory*. Illuminate a transistor (*e.g.* with a laser beam) causes it to conduct,

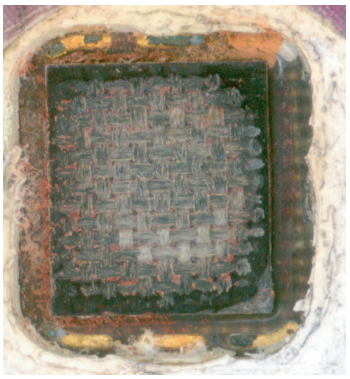

Fig. 4. A kevlar © protection

so a transient corrupted bit can appear introducing a fault during computation [3], [15]. But one must also be aware of fault attacks which consist in generating a malfunction (*e.g.* changing work frequency, or insert a glitch) during the computation that causes some bits to adopt the wrong value. One famous example of using an error generated during computation is, again, on the implementation of RSA using the CRT. A simple relation exists between the corrupted results and the secret key [3], [5], [8]. A transient fault can also be produced by using a very high magnetic field [12]. Some techniques, however, yield non-reversible faults. A technique given in [3] describes how to directly rewrite into the memory, or even how to destroy a transistor.

In the first part of this paper, we will discuss different problems that we are likely to encounter while using smart cards. In the second part, we will give a short overview of memory technologies. The third part will explain some possible attacks on memories. Another section will be dedicated to the possible countermeasures to such threats. And finally, we will give a conclusion.

II. Issues

The transmission with the card is public knowledge and the physical signal is easily accessible to hackers and eavesdroppers. A first problem comes from the physical interface. The physical access to the card can be achieved through a

serial communication with contacts (ISO 7816 or Universal Serial Bus) or through wireless communication at a normalized frequency. In the second case, the anticollision problem should be carefully taken into account, which means that there should not be any problem having two cards in the same field at the same time. Whatever the way data are accessed, the communication is always serial and therefore constitutes a bottleneck.

A consequence thereof is that the data access time limits the use of the cards. We will probably have to wait for a change of the standard before smart cards are able to receive a good quality video stream, compressing on the fly. The quantity of embedded data in memories is however small compared to the one that is nowadays commonly used in Personal Device Assistants (PDA) and in wireless telephones.

Unlike desktop computers that operate in a physically secure environment (Orange book), smart cards operate in a hostile environment. Anyway, smart cards must maintain privacy and integrity of data, and authenticity of all parties involved in a protocol. Confidentiality, user identification, secure software execution, secure storage, secure external access, secure content, secure data communications, and tamper resistance are the basic properties of smart cards.

Power consumption of smart cards should of course be reduced as much as possible. However, if the consumption is not sufficiently masked, it can constitute the source of very dangerous side channel leakages. The same remark applies to the electromagnetic radiation of the card. Some concepts should be embedded in the card in order to reach a better security than for a simple memory. This should however be done without a noticeable increase in power consumption.

There will be a new growth for smart cards thanks to the enforced standard in identity cards, and more specifically because of the recent events in the world. The fact that identity cards are administrative documents, linked to a particular state reenforces the existing requirements. A card has a lifetime of several years and this is reflected in the storage time and in the MTBF of the card. The more advanced cards will contain biometric elements and a fingerprint. This information requires a unitary memory size of 12Ko at least, which means that the demand for secure storage space will be higher than it is today.

In [13] we can read ”Additionally, hardware components such as secure RAM and secure ROM in conjunction with hardware based key storage and appropriate firmware can enable an optimized ’secure execution’ environment where only trusted code can be executed. A secure execution mode can be used for critical security operations such as key storage/management and run-time security to provide a strong security foundation for applications and services.”. This is particulary true.

So the issues with smart cards are completely different from those of other mobile systems like wireless telephones or personal device assistants (PDA). Indeed, as their name indicates, these cards are intelligent and especially known for their tamper resistance and their high cryptographic computation power. Smart cards are able to capture, record and communicate with external readers. In other words the smart card plays the role of a highly secure token that is able to perform secure computations.

Due to their form factor that does not include any embedded energy source (battery,. . .), these tiny cards are however easily lost or stolen. This implies that it is very important to make sure that they will not suffer from physical attacks.

These ”Personal Trusted Devices” will be components of tomorrow’s environment (European Project INSPIRED), which means that they need to have a low price and a considerable computing and communication capacity.

III. Memory technologies

According to the requirements of a given application, there exist different types of electronic memories. Classical characteristics are the memory size, the time taken to access stored data, the access patterns, and so on.

Typical classification sorts the memories in three classes: Read Write Memories (RWM), Non Volatile Read Write Memories (NVRWM) and Read Only Memory (ROM).

A. ROM

ROM are mass produced because it is the simplest semiconductor memory. It is mainly used for the operating system or to store instructions or constants for smart cards. In a classical ROM only one word line can be high at a time. We see that R1 going high causes the column C1, C3, and C5 to be pulled low. The transistor in the superior part of the figure are long L pull-ups. Columns lines C2 and C4 are pulled high through the long L mosfet. If the information that is to be stored in the memory is unknown at the fabrication, each memory array is built with an n-channel mosfet at every intersection of a row and a column line. The memory is programmed by cutting the connection between the drain of the mosfet and the column line.

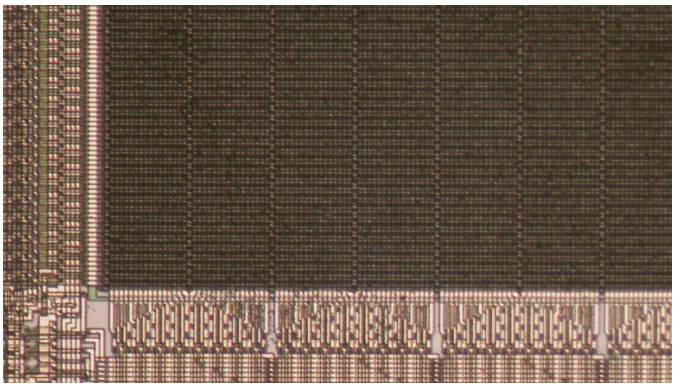

Fig. 5. A Read Only Memory

B. SRAM

SRAM does not need to be refreshed and the presence of the supply voltage is sufficient to retain its information. The generic Static RAM (SRAM) cell is introduced in Fig. 6 It requires six transistors. The *word line* enables the access to the cell by controlling the two pass-transistors *M5* and *M6*. In

contrast to the ROM cells, two bit lines transferring both the stored signal and its inverse are required. Although providing both polarities is not a necessity, doing so improves the noise margins during both read and write operations .

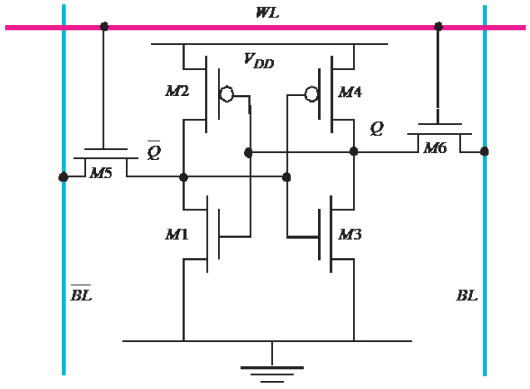

Fig. 6. A classical 6 transistors SRAM structure

A close lookup of the transient behavior shows that the read operation is the critical one. It requires (dis)charging the large bit-line capacitor by the small transistors of the selected cell. The *write* time depends on the propagation delay of the cross-coupled inverter pair; the drivers that set the desired value to BL and $\overline{BL}$ can be large.

The area consumption due to the six transistors, the two bit lines, the word line, both supply rails, the signal routing and connections are drawbacks of the SRAM.

C. DRAM

The concept of *Dynamic* RAM is charge storage on a capacitor. A three-transistor cell shown in Fig. 7 in enabled by the *write-word line* and the *read-word line*. The cell is written to by placing the appropriate data value on $BL1$ and on raising the *WWL*. When reading, $BL2$ is precharged to a load device to V_{DD} or $V_{DD} - V_T$. The series connection of M_2 and M_3 pulls *BL2* to low when a 1 is stored on the capacitance and remains high in the opposite case.

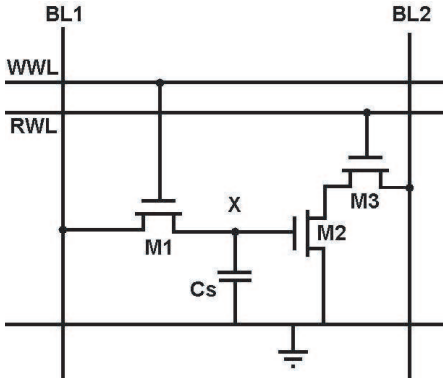

Fig. 7. A classical 3 transistors DRAM structure

A one-transistor version of DRAM exists by the sacrifice of some cell properties. Anyway, this kind of memory has been tested but is not presently used in smart cards.The RAM of smart cards is currently Static RAM (SRAM). The main reason is the possibility to use a power-saving mode when the CPU stays in sleep mode, the clock is permanently fixed to the high or the low level. Whereas a Dynamic RAM (DRAM) needs to be periodically refreshed [17].

D. EPROM

EPROMs make programming the ROM greatly easier. A modified n-channel mosfet is used at the intersection of the column and tow lines in the ROM memory array. A polysilicon layer is added directly above the original polysilicon layer (floating). The second layer is connected to the row lines. The result is a polysilicon capacitor with the bottom plate used in mosfet formation. Increasing the row line voltage turns the mosfet on and pulls the column line low. When the row line goes high, the mosfet turns on and pull the column line low. A large voltage is applied to the row line to be sure that the mosfet remains off when the line goes high. This voltage causes a large current to flow in the mosfet and at the same time an avalanche occurs in the substrate. When the large voltage is removed from a plate of the capacitor then the other plate will drop down to a negative voltage. The column line can stay high as the mosfet does not turns on under normal operation. The cell can be reprogrammed by illuminating the chip with ultraviolet light. As SiO_2 surrounds the gates, UV lights causes electron hole generation, and the immediate effect is to increase the insulator conductivity. Then the charges trapped on the polysilicon can leak off.

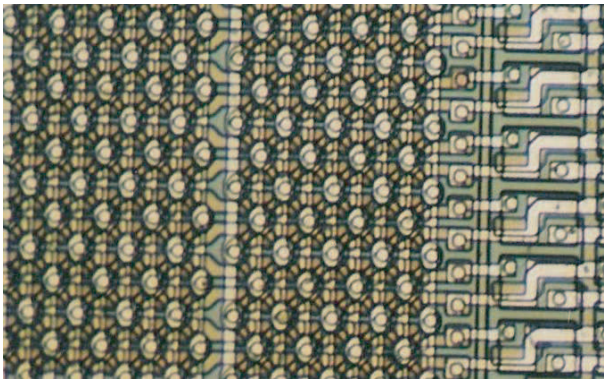

Fig. 8. An EPROM on a memory card

E. EEPROM and FLASH

With a on chip voltage generator [17] it is possible to create a large voltage needed to program the EEPROM memory cell. The gate oxide of an EEPROM is thinner than the one used in EPROM. The result is a tunnel effect (Fowler-Nordheim tunneling) between the substrate and polysilicon. This mechanism permits driving current in both directions. A FLASH memory is based on the two technologies. FLASH memory is programmed as EEPROM. Electrons are used to accumulate charges on polysilicon. The structure of a FLASH and an EEPROM are very closed except for the oxide thickness. The

main difference is that FLASH memory is programmed using hot electrons and erased using Fowler-Nordheim tunneling.

IV. Security Analysis

A. Dallas attack

Many applications require a big amount of RAM and bus encryption techniques are commonly used to protect the stored data and prevent eavesdropping accesses. In [2] a *protocol attack* is described against the Dallas Semiconductor DS5002FP microcontroller, which was widely used in numerous transaction terminals and pay-TV access control systems. The secret key is unique to each device and is protected by a self-destruct alarm process. Pseudo-random dummy accesses are performed when the CPU does not require external memory access. The chip uses two block ciphers : the first encrypts the addresses on 15-bit blocks while the last on 8-bit data blocks. The key of the second cipher is salted with the address of the byte being encrypted.

According to the IBM taxonomy, the attack is of class I (*clever outsiders*) since it simply requires a computer, a logic analyzer and a special read-out circuit. The basic idea of the attack is the feeding of the CPU with chosen encrypted instructions and the observation of their effects : *e.g.* the three-byte instruction `MOV 90h, #42h` encoded `75h 90h 42h` results in the output of the `42h` value on the parallel port (`90h`). Just before the CPU fetches, the read-out hardware replaces it and the control software observes the reactions. Through the 2^{16} combinations, one eventually discovers the ciphered instruction that sends the following byte to the parallel port. *This gives the data bus decryption function at the address from which the third instruction byte was fetched. By testing all 2^8 values for this byte, one can tabulate the data decryption for one address*.

Repeating the whole process, searching for the `NOP` instruction followed by the same instruction as before, *increases by one the address from which the third* `MOV` *instruction byte will be fetched*. This permits tabulating the encryption function for a consecutive list of addresses. It is then possible to send to the CPU core a sequence of machine instructions that dumps the memory to the parallel port.

The presented attack was performed against one of the 'top' commercial systems of that time. This example demonstrates that even bus encryption based systems can present unexpected flaws in their implementation and can be easily ruined.

B. Recovering data from semiconductors devices

Data remanence problems can occur in all semiconductor devices. They are quite difficult to counteract because they greatly depend on the intrinsic characteristics of semiconductors. Despite of all memory problems presented here, general guidelines are given in order to help reducing remanence issues. The danger of such attacks is that, if even part of the key is retrieved, the number of combinations remaining is dramatically decreased. This part is widely inspired from the paper written by Peter Gutmann in 1996 [7] and we refer the ready to this paper for further detailed information.

1) Semiconductor physics consideration: Modern semiconductor devices and integrated circuits (ICs) are based upon the conduction band, carrier transport and optical properties of semiconductor materials. Attacks using optical features of semiconductors are presented below. Most LSI (large-scale integrated circuit) devices and memories use the MOSFET (metal oxide semiconductor field-effect transistor). There are two types of MOSFET: n-channel and p-channel. In the first type, the current flow is dominated by electrons while holes dominate in p-channel type. In most common circuits, they are both usually combined in order to take advantage of their different characteristics in the form of complementary MOS(CMOS).

2) Effects related to the functioning of semiconductor: With various memory types, it is possible to analyze and retrieve data a long time after it has vanished. This is feasible by analyzing the circuit at the physical level.

i) **Electromigration** The high current densities ($> 5E5A/cm^2$) that are used in aluminum-based interconnection lines of integrated circuits induce a degradation mechanism known as Electromigration (EM). EM is a momentum exchange between moving electrons and the metal ions of the crystalline matrix. Due to EM, voids, whiskers and hillocks are formed in the metallization (See Figure 9).

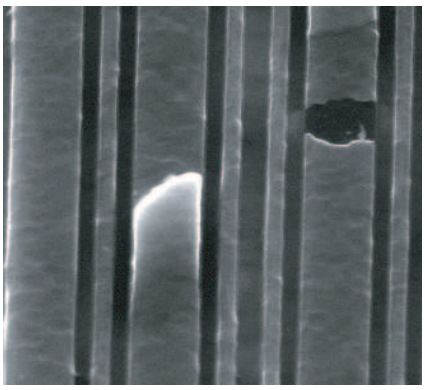

Fig. 9. Void and hillock due to electromigration © A.Scorzoni

Such effects can affect the operating characteristics in noticeable ways (*e.g.* wiring resistance, current leakage,...).

ii) **Hot Carriers** Electrons with high an energy level can in some cases cross the potential barrier of the gate oxide or even of the passivation layer by tunnel effect. So, some charges can be trapped in the insulator trench and the reverse process can require nanoseconds or even days. This excess of charge reduces the on-state current (n-MOS) or the off-state current (p-MOS). Depending on whether a 1 bit is written after a 0 bit or conversely, one can detect a drop or an increase in the threshold voltage.

iii) **Radiation-induced** Radiation can alter memory cell parameters such as voltage, level thresholds, timings, power supply and leakage current. This attack can allow locking out tamper-responding circuitry (*e.g.* erase-on-

tamper). That is the reason why high-end crypto devices include sensors to detect ionizing radiation.

iv) **EEPROM/Flash memory** EEPROM technology uses mainly the Fowler-Nordheim effect to tunnel electrons in a floating gate, to program (newer technologies use channel hot electron technic) and to erase. Some electrons trapped into the thin oxide; can produce a shift in threshold voltage or a change in program and erase times.

Many techniques are available to detect such effects in semiconductor devices. The most common ones are: measuring the power supply current, varying operating voltage and temperature to test for hot carrier effects, mechanical probing, focused ion beam for deep sub micron testing, reverse engineering,...

3) Recommendations:

- Do not store cryptovariables for long periods in the same location.
- Do not store cryptovariables in plaintext form in non-volatile memory .
- Cycle EEPROM/flash cells 10-100 times before using them.
- Do not assume that a key held in RAM has been destroyed when the RAM is cleared.
- Design devices to avoid repeatedly running the same signals over dedicated data lines.
- Beware of too-intelligent non-volatile memory devices that could leave copies of sensitive information in mapped-out memory blocks after the active copy has been erased.

C. Tampering with data in memory

The two following attacks presented target more precisely SRAM memory but they could be applied to all memory technologies using semiconductors. We describe here the general method used to carry out the experiment.

The transistors M1 and M2 create the CMOS inverter; together with the other similar pair, they create the flip-flop which is controlled by the transistors M3 and M6. (Figure 14.) If the transistor M1 could be opened for a very short time by an external stimulus, then it could cause the flip-flop to change state. By exposing the transistor M4, the state of the cell would be changed to the opposite. The main difficulties we might anticipate are focusing the ionizing P2 and choosing the proper intensity.

D. Eddy Current for Magnetic Analysis with Active Sensor

An alternating current in a coil near a conducting surface creates a magnetic field. This field induces eddy currents on the material surface. This effect was discovered by Lon Foucault and consequently called *Courant de Foucault*. The magnitude and the phase of the eddy currents will affect the loading and the impedance of the coil. This could be used to find some flaws and corrosion in metal items (propeller, vanes, fixing joint,...).

The attack presented here does not require any depackaging

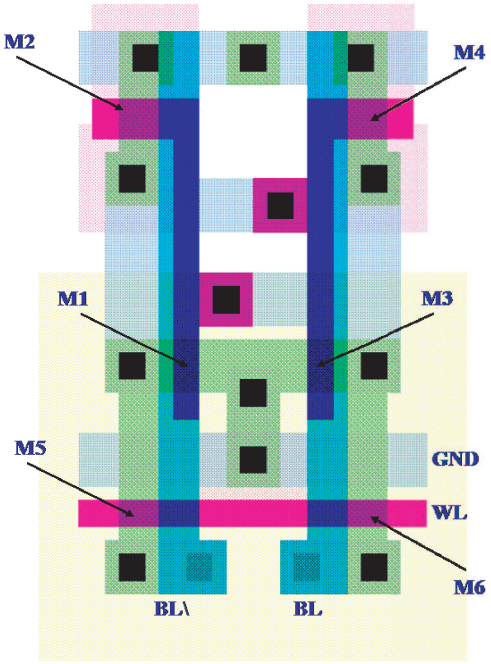

Fig. 10. A classical 6 transistors SRAM layout

of the processor. It is possible to induce faults in a smart card through the plastic. But, of course if you can see the chip, it is easier to inject a fault in a regular structure (memory or other).

So, the idea was to use this method to induce faults into a chip we want to attack. With various current intensities flowing through the coil, a fault can be transient or definitive, but it is also possible to destroy the silicon.

1) The principle of the attack: With adapted current intensities flowing through the coil (usually the goal is not to destroy the silicon), it is possible to change the energy level of electrons flowing trough transistors in the chip. With this increase of energy they are able to cross the gate oxide of transistors. This is called the Fowler-Nordheim effect and is well-known from the EEPROM manufacturers because this type of memory is programmed and erased by this tunnel effect. For a 6 or 7 nm oxide grid thickness (t_{ox}), with a 6MV/cm electric field, the Fowler-Nordheim effect appears. This current is linked to the surface connected to the grid. So it is possible to compare it to an antenna effect. In a regular structure such as a memory, very long lines are present. It is very easy to obtain a Fowler-Nordheim current, with important effects.

2) Practical consideration: The main advantage of this attack is that it does not require huge investments and facilities. Indeed, the attack was implemented from a simple camera flash gun and a needle. A wire was winded round the needle in order to obtain a coil containing hundreds or even thousands of whorls. When a current flows through the wire, it creates a magnetic field that is concentrated on the top of the needle. The current injected into the coil can be obtained starting from a simple camera flash gun. Once the bulb is withdrawn the coil

wires are connected across the contacts replacing the bulb. During the release of the flash, an important spark can appear when the coil and the wire are connected. The test probe is placed a few microns above the top of the processor or memory to be attacked (SRAM Sony 256 K, or KM41256-12 709). The field obtained at the end of the point creates an Eddy current in the chip. If the current in the coil is strong enough and the coil is very big, a permanent fault is inserted. But if the coil is smaller, or if a resistor is used, the Eddy current is reduced and the fault can be transient.

A patient attacker will be able to insert faults in memory and bind the geographical position of his sensor to the inserted fault. We succeeded in the case of all kinds of memories: RAM [16], EPROM, EEPROM, Flash to disturb entire columns of memory.

A memory is generally organized into rows and columns. For a RAM, each column contains several memory points and each of them is connected to its differential amplifier by two transistors. Columns can be easily disturbed. Indeed the reading amplifiers are based on a differential structure. This structure is very sensitive to external perturbations. When a whole column seems to be touched (results are shown on Figure 11), very often only the reading amplifier has been disturbed.

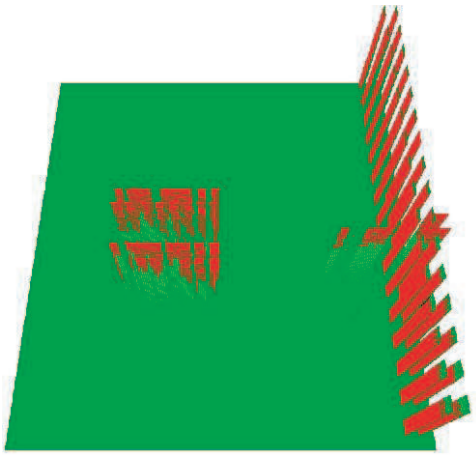

Fig. 11. Graphics of modified bits in a column of a memory.

When a RAM is switched on, the stored values are random. On a *faulty* static RAM, the great majority of amplifiers answer 1. So these values are not random anymore. When the field is relatively intense and brutal, the processor submitted to this field can be put in a *disrupted* mode. We noted cases where memories or microprocessors ceased functioning for several hours following a severe fault insertion. In a few cases after a time varying between five hours and several tens of hours, the processor or the memory turned back to a normal operation mode. We also heated the attacked components in order to see if the process was reversible. We showed that almost all sticky bits were maintained to their forced value [12]. These attacks permit inserting a fault on a bus or just changing the value of one bit. Of course, it is also possible to activate hidden opcodes or instructions, but we never managed to do that intentionally.

E. Insertion of faults using light

A well-known principle nowadays is the *photovoltaic effect* which consists of the production, as a result of the absorption, of a voltage difference across a pn junction.

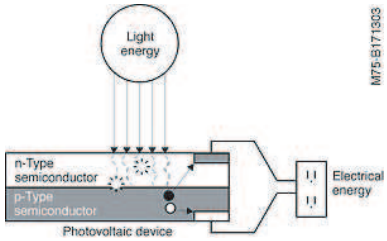

Fig. 12. Photovoltaic effect.

Smart card manufacturers know how to inject a fault into a chip using a flash and a microscope for a long time. Sergei Skoborogatov proposed in [15] to use the photovoltaic effect to create, as in the case of eddy current, a current in a transistor. Similar experiments were carried out to prove the interest of laser (until 1064 nm wavelength) in assessment of Integrated Circuits. In the paper related here, the author carried out his experiment using only a second-hand flash gun and a optical microscope to target a particular transistor and focus the light.

1) Background: Only photons with a certain level of energy are able to free electrons from their atomic bonds, to produce electric current. This level of energy, known as the band-gap energy, is defined as the amount of energy required to dislodge an electron from its covalent bond and allow it to become part of an electrical circuit. For example, crystalline silicon's band-gap energy is 1.1 eV. So the author decided to use a simple visible light source (a flashgun) and focus the light with appropriate magnification (it was set to the maximum: 1500x) of the microscope in order to reach the right level of ionization.

2) Experimental protocol: The author targeted to attack the SRAM memory of a microcontroller PIC16F84 (Microchip) that contains 68 bytes of SRAM. He depackaged the chip in order to get access to the layout and localize the SRAM. The microcontroller was programmed to upload and download its memory. By filling the whole memory with constant values, exposing it to the flash light, and downloading the result, he could observe which cells changed their state. By shielding the light from the flash with an aperture made from aluminum foil, he succeeded in changing the state of only one cell. So he proved that it was possible to change the content of an SRAM using a low cost semi-invasive attack.

V. Counter-Measures

Memory cards that contain only a few bits are the simplest smart cards after contactless cards, which are limited to giving their identifier. Memory cards were patented by Roland Moreno on March 17, 1975 and largely contributed to the vulgarization of smart card technology.

These cards contained a small amount of memory which was used to store telephone units. The first attacks on memories attempted to erase the used units and recharge the cards.

Therefore, attackers used to use EPROM erasers with UV light, having a wave length of 250nm. A first countermeasure against this attack consisted in the use of a burnt fuse during the personalization of the card, in order to refuse access to the protected zone of the card and thereby ensuring the uniqueness of the card. Packaging the card into resina or epoxy was another, rapidly introduced countermeasure that also aimed at countering these attacks and, in the mean time, served to favor the mechanical resistance of the card against torsion. Researchers from the University of Cambridge (Anderson and Kuhn) showed, however, how easily this resina can be removed, allowing so the access to the silicon chip without altering it.

Although card memories are small, both in storage capacity and in physical size, they can suffer from disruptions from the outside or even from an inattentive user. The question is to know what kind of disruptions and attacks the countermeasures should be able to stop. It should indeed be noticed that disruptions due to normal use of the card sometimes reveal to be very close to a deliberate attack. So, before being declared valid, the card is first submitted to a wide variety of tests. One of these tests consists in applying a high discharge of static electricity. These discharges may damage the processor or the memories of the card. Present smart cards are equipped in a way that prevents these troubles. Such a discharge may come from contact with an electrically charged person or may be deliberately caused by an attacker in order to disrupt the content of the memory.

Michel Ugon's idea of microprocessors on the one hand, and the massive use of smart cards on the other, incited industry to push both the integration density of memories in cards and the progressive use of different types of memories. This explains the introduction of volatile memories like ROM, intended to hold the operating system and volatile memories (pile or accumulator for a microprocessor) based on RAM memory, into the world of smart cards. Erasable storage memories like EEPROM have also showed up in smart cards and are intended to hold applications or patches of the operating system.

The first attack by fault insertion, called Bellcore attack, appeared in 1996. Based on this model, another type of cryptographic attack by fault insertion was introduced in 1997 by Dan Boneh *et al*. Although the model assumed the fault to appear during the execution of mathematical computations, it was rapidly adopted to fit in a memory fault. The corruption of stored data can easily be detected by the use of error correcting codes, which use a cyclic redundancy that is a number derived from a block of data in order to detect the corruption. By recalculation of the CRC and comparison with the block of data, the memory can detect some types of storage errors. Setting up error correcting codes is however not always a help. Indeed, the knowledge that data are corrupted is useful, but for cryptographic security reasons, it cannot always be accepted that the error is also corrected.

Fault tolerance and resistance to error apparition are problems that have been known for a long time in the scientific world handling silicon based memories. The effect of ionizing radiations has, for instance, been taken into account by engineers since the beginning of the conquest of space. Indeed, because the crystalline network of silicon is sensitive to ionizing radiations, it is possible that transistors switch state due to parasite effects. One needed also to develop electronics that were capable to function near active nuclear devices. The complexity of memory points has also highly increased, to the detriment of the cost price of the components. The presence of several transistors in the model for memory points or hardened processors were even at the origin of conferences (RADECS, ...) and highly interests prestigious laboratories, like Sandia, for applications in either the military or the civil nuclear domain. However, the present objective of growth for smart card and silicon manufacturers resides more in the integration of large quantities of memory having a low consumption, rather than in a sensible increase of the security of their memory points.

The work of Peter Gutman, in 1996, on the secure erasing and memory remanence of semiconductor circuits, draws the interest back that allow maintaining information in certain types of memory while the power supply is cut off. Skorobogatov *et al*. reported in a technical report of remanence times of several days for RAM, requiring only a cheap and easily usable equipment

The first articles published in scientific literature and mentioning memory attacks on smart cards, often reported only on simple visual inspection under microscope. In that way it was possible to read the ROM as easily as one would read a book. In order to limit this, silicon manufactures decided to scramble the memory location and to forbid two contiguous addresses from having a simple position link in the layout of the chip. From that time on, memories have voluntarily scrambled cell locations.

Soon after this, the power consumption of the memory showed out to reveal sensitive information to external attackers. Pirates used this observation and disrupted power supply from the moment they realized that the charge pump, which is necessary for writing into EEPROM, was triggered off. As a countermeasure against this attack one started to highly reduce and smoothen out the characteristics of the memory's power consumption and to apply software securities so that the interrupted memory could be writing anyway the next time the power supply would be turned on. Nowadays, it is still possible to detect the moment the charge pump is turned on, but a more dangerous thing is that the transported data sometimes modulate the local oscillator. This unexpected side channel is another leakage source that should be seriously taken into account, as it has already been pointed out by IBM [1].

In order to limit leakages due to power consumption measurements, silicon manufacturers have constructed cells that are able to conduct as many charges to the ground as to the positive side and this both during reading and writing operations on the memory. As this countermeasure showed out to be insufficient, silicon manufacturers have implemented the idea of Ross Anderson and his team, that consists in using two wires for every bit and to store the 0 and the 1 state in an opposite and, from a consumption point of view, balanced way. As the transitions were still visible, one decided to call upon precharged logic, forcing an identical relaxation state between every transition. This makes the analysis of the

memory content during reading and writing far more difficult, because the attack has now to be restricted to the moments where a state transition occurs.

Although a priori the consumption of these transitions is identical, this is not true for their radiation. In order to limit leakage from radiation and to inhibit from visual inspection, memory manufacturers have also placed metal layers on top of the memories. A simple laboratory device, available in every microelectronics laboratory should thus suffice to remove these metal layers. Smart cards are provided with sensors to prevent the use of these devices.

There are two types of sensors, namely active and passive ones. These sensors serve a large variety of functions. They are able to detect a temperature decrease, which makes them erase the memory, and thus; preventing the effect of data remanence. They can also set a flag due to light detection, a change in the environmental capacity or in the environmental resistance, in order to prevent an attacker from finding back and using recorded data.

A passive sensor may take the form of a randomly added or methodically arranged transistor inside the memory circuit, supposed to be able to detect a fault insertion attempt. As an active sensor, there is, for instance, the possibility of adding to wires, covering the entire card and connecting when an intrusion with microprobes is detected. Another countermeasure is to use a face to face interconnection or to use wafer scale packaging.

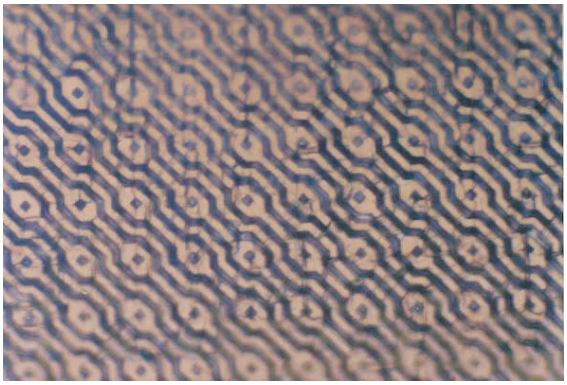

Fig. 13. A wire mesh

VI. Conclusion

Smart card security is an important trade-off between usability, cost, and electrical properties (power consumption, ...). A secure smart card design should contain memories and a lot of sensors, actuators, ..., which constitutes its added value compared to cheaper and larger memory. In the near future, there will be Trusted Personal Devices built on the smart card technology to ensure a high security level. Challenging the memory problem will not be trivial at all.

FeRAM and Magnetoresistive RAM are actually used by smart card manufacturers, and Battery Backup RAM (BRAM) will possibly take their place in the future. Anyway, the next generation of smart card will have an improved memory density and a high speed page access. To remain successful,

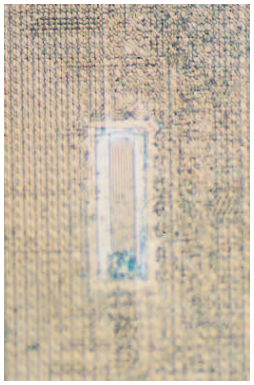

Fig. 14. A sensor

smart cards will need to have new, on chip micro batteries, micro sensors, permanent attack detection, and active countermeasures.

A new solution to this security paradigm could be Silicon On Insulator or Low Power Technology. For a long time now, people have known asynchronous logic, and smart card manufacturers are testing it. Maybe it will permit relaxing some constraints and improving memory management. Moore's law has an important advantage: attacking smart cards will in the future be more and more difficult as the size of the transistors is going down and down. Nowadays it is very difficult to use probes to defeat a card and within a few years time visual inspection under a microscope will be even simply impossible. At that time, attacks will cost more and will be more difficult to apply.

References

[1] Dakshi Agrawal, Bruce Archambeault, Josyula R. Rao and Pankaj Rohatgi *The EM side-channel(s).* In B.S. Kaliski Jr. and Ç.K. Koç, Ed., Cryptographic Hardware and Embedded Systems (CHES 2002), volume 2523 of Lecture Notes in Computer Science, pp. 29–45. Springer, 2002.

[2] Ross J. Anderson, Markus Kuhn. *Tamper resistance – a cautionary note*, Second USENIX Workshop on Electronic Commerce Proceedings, Oakland, California, pp. 1-11, 1996.

[3] Ross J. Anderson, Markus Khun, *Low Cost Attacks on Tamper Resistant Devices*, in proc. of 5th Security Protocols Workshop, LNCS 1361, pp. 125-136, Springer, 1997.

[4] Eli Biham and Adi Shamir. *Differential fault analysis of secret key cryptosystems.* In B.S. Kaliski Jr., Ed., Advances in Cryptology - CRYPTO '97, volume 1294 of Lecture Notes in Computer Science, pp. 513–525. Springer, 1997.

[5] Dan R. Boneh, Richard A. DeMillo, and Richard J. Lipton. *On the importance of eliminating errors in cryptographic computations for faults. Journal of Cryptology*, 14(2):101–119, 2001. An earlier version appears in EUROCRYPT '97 [**?**].

[6] Karine Gandolfi, Christophe Mourtel, and Francis Olivier. *Electromagnetic analysis: Concrete results.* In Ç.K. Koç, D. Naccache, and C. Paar, Ed., Cryptographic Hardware and Embedded Systems (CHES 2001),volume 2162 of Lecture Notes in Computer Science, pp. 251–261. Springer, 2001.

[7] Peter Gutmann, *Secure deletion of data from magnetic and solid-state memory*, 6th USENIX Security Symposium. San Jose, California, July 22-25, 1996.

[8] M. Joye and A. K. Lenstra and J.-J. Quisquater, *Chinese Remaindering Based Cryptosystems in the Presence of Faults*, Journal of Cryptology , 12(4): 241–245, 1999. citeseer.nj.nec.com/28125.html.

[9] Paul C. Kocher. *Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems.* In N. Koblitz, Ed., Advances in Cryptology - CRYPTO '96, volume 1109 of Lecture Notes in Computer Science, pp. 104–113. Springer, 1996.

[10] Paul C. Kocher, Joshua Jaffe, and Benjamin Jun. *Differential power analysis.* In M. Wiener, Ed., Advances in Cryptology - CRYPTO '99, volume 1666 of Lecture Notes in Computer Science, pp. 388–397. Springer, 1999.

[11] Jean-Jacques Quisquater and David Samyde. *Electromagnetic analysis (EMA): Measures and counter-measures for smart cards.* In I. Attali and T.P. Jensen, Ed., Smart Card Programming and Security (E-smart 2001), volume 2140 of Lecture Notes in Computer Science, pp. 200–210. Springer, 2001.

[12] Jean-Jacques Quisquater, D. Samyde, *Eddy current for Magnetic Analysis with Active Sensor* Proceedings of Esmart 2002 3rd edition, pp. 183–194. Nice, France. September 2002

[13] Anand Raghunathan, Srivaths Ravi and Jean-Jacques Quisquater. *Securing Mobile Appliances : New challenges for the system designer.* , IEEE , March 3-7, Munich, Germany, 2003.

[14] Sergei Skorobogatov, *Low temperature data remanence in static RAM*, presented 06-02-2001 at the Cambridge Computer Lab Security Seminar. `www.cl.cam.ac.uk/TechReports/UCAM-CL-TR-536.pdf`

[15] S.P. Skorobogatov, R.J. Anderson, *Optical Fault Induction Attacks*, Cryptographic Hardware and Embedded Systems (CHES 2002) volume 2523 of Lecture Notes in Computer Science, p. 2-12. Springer, 2002.

[16] Werner Schindler. *A timing attack against RSA with the chinese remainder theorem.* In Ç.K. Koç and C. Paar, Ed., Cryptographic Hardware and Embedded Systems (CHES 2000), volume 1965 of Lecture Notes in Computer Science, pages 109–124. Springer, 2000.

[17] Amaury Nve de Mevergnies, Denis Flandre, Jean-Jacques Quisquater. *Feasibility of Smart Cards in Silicon-On-Insulator (SOI) Technology*, USENIX workshop on Smartcard Technology, Chicago, USA, pp. 1-7, May 1999.